Univerzita Pardubice

Fakulta elektrotechniky a informatiky

Data Science a vizualizace dat

Bc. Tomáš Prudký

Diplomová práce

2023

Univerzita Pardubice
Fakulta elektrotechniky a informatiky
Akademický rok: 2022/2023

ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Bc. Tomáš Prudký**
Osobní číslo:	**I21284**
Studijní program:	**N0613A140007 Informační technologie**
Téma práce:	**Data Science a vizualizace dat**
Zadávající katedra:	**Katedra softwarových technologií**

Zásady pro vypracování

Cílem práce je popsat možnosti a způsoby využití Data Science a vizualizace dat a demonstrace formou ukázek analýzy totožných dat a jejich vizualizace pomocí jazyků a nástrojů využívaných v této oblasti, a to minimálně pomocí jazyka Python s využitím DS knihoven (NumPy, Pandas, Plotly), R a Scala a příslušných nástrojů, a to nejméně využití JupyterLab nebo Jupyter notebooku, Apache Zeppelin ve spojení s Apache Spark a dále ELK Stack. Práce bude dodržovat metodiku Data Science, zohledňovat Data Lifecycle Management, procesní model CRoss Industry Standard Process for Data Mining (CRISP-DM) a model vyspělosti Data Science Maturity Models. V rámci práce budou využity historická data odpovídajícího rozsahu.

Rozsah pracovní zprávy: **cca 60 stran**
Rozsah grafických prací: **–**
Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam doporučené literatury:

HILL, Raymond R. Becoming a Data Head: How to Think, Speak and Understand Data Science, Statistics and Machine Learning. 2021.

VANDERPLAS, Jake. Python data science handbook: Essential tools for working with data. „ O'Reilly Media, Inc.".

MCKINNEY, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. „ O'Reilly Media, Inc.".

Vedoucí diplomové práce: **Ing. Monika Borkovcová, Ph.D.**
Katedra informačních technologií

Datum zadání diplomové práce: **8. listopadu 2022**
Termín odevzdání diplomové práce: **19. května 2023**

L.S.

Ing. Zdeněk Němec, Ph.D. v.r.
děkan

prof. Ing. Antonín Kavička, Ph.D. v.r.
vedoucí katedry

V Pardubicích dne 30. listopadu 2022

Prohlášení autora

Práci s názvem Data Science a vizualizace dat jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Beru na vědomí, že v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a směrnicí Univerzity Pardubice č. 7/2019 Pravidla pro odevzdávání, zveřejňování a formální úpravu závěrečných prací, ve znění pozdějších dodatků, bude práce zveřejněna prostřednictvím Digitální knihovny Univerzity Pardubice.

V Pardubicích dne 25. 8. 2023

Bc. Tomáš Prudký

PODĚKOVÁNÍ

Na tomto místě bych rád poděkoval paní Ing. Monice Borkovcové, Ph.D., vedoucí mé diplomové práce, za podporu, důvěru, trpělivost a čas, který mi v průběhu psaní věnovala. Za cenné informace a rady, které mi poskytla, kdykoli jich bylo potřeba a za její pozitivní energii, kterou mi po celou dobu dodávala.

ANOTACE

Cílem této diplomové práce je představit čtenářům možnosti a využiti Data Science a vizualizace dat. V první části bude představena historie oboru Data Science, možnosti využití a popis vybraných typů vizualizací. Diplomová práce představuje procesní model CRISP-DM, model vyspělosti Data Science Maturity a Data Life Cycle Management. Poslední teoretická část se zaměřuje na nástroje používané datovými vědci, jako je Jupyter, Python, R, Scala, MATLAB, Julia a Elastic Stack. Praktická část je věnována vybraným nástrojům z teoretické části, pomocí kterých bude provedena vizualizace a demonstrace analýzy totožných dat. Praktická část bude dodržovat metodiku Data Science a zohledňovat metodiky popisované v teoretické části.

KLÍČOVÁ SLOVA

data science, DSMM, DLM, CRISP-DM, Jupyter, Python, R, Scala, MATLAB, Julia, Elastic stack, vizualizace

TITLE

Data Science and Data Visualization

ANNOTATION

The aim of this master's thesis is to introduce readers to the possibilities and utilization of Data Science and data visualization. The first part will present the history of the Data Science field, its potential applications, and describe selected types of data visualizations. The thesis introduces the CRISP-DM process model, the Data Science Maturity model, and Data Life Cycle Management. The final theoretical section focuses on tools commonly used by data scientists, such as Jupyter, Python, R, Scala, MATLAB, Julia, and the Elastic Stack. The practical portion will focus on the selected tools from the theoretical part, through which data visualization and the demonstration analysis of identical data will be performed. The practical section will adhere to the methodology of Data Science and incorporate the methodologies described in the theoretical part.

KEYWORDS

data Science, DSMM, DLM, CRISP-DM, Jupyter, Python, R, Scala, MATLAB, Julia, Elastic stack, visualization

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM ZKRATEK A ZNAČEK

2D	Dvoudimenzionální
3D	Trojdimenzionální
ACM	Association for Computing Machinery
SIGKDD	Special Interest Group on Knowledge Discovery in Data
API	Application Programming Interface
CMM	Capability Maturity Model
CPU	Central Processing Unit
CRISP-DM	Cross-Industry Standard Process for Data mining
CSV	Comma-Separated Values
DBMS	Database Management System
DLM	Data Life Cycle Management
DSMM	Data Security Maturity Model
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HDF5	Hierarchical Data Format version 5
HTML	Hypertext Markup Language
IASC	Inter-Agency Standing Committee
IFIP	International Federation for Information Processing
IoT	Internet of Things
JDBC	Java Database Connectivity
JSON	JavaScript Object Notation
KDD	Knowledge Discovery and Data Mining

LSA	Latent Semantic Analysis
LSD	Latent Structure Discovery
LSTM	Long Short-Term Memory
ML	Machine Learning
PNG	Portable Network Graphics
REST	Representational State Transfer
SQL	Structured Query Language
SVG	Scalable Vector Graphics
TPU	Tensor Processing Unit
TSV	Tab-Separated Values
USA	United States of America
USD	United States Dollar
XML	Extensible Markup Language
ČR	Česká republika

ÚVOD

V dnešním digitálním světě jsou hnacím motorem všeho data. Každou chvíli je generováno nepředstavitelné množství informací. Tyto informace mohou sloužit jako surový materiál k odhalení skrytých vzorů, získání znalostí a poskytnutí cenných informací pro podporu rozhodování. Právě zde vstupuje na scénu obor nazývaná „Data Science“. Svět se stále více spoléhá na data jako zdroj informací pro rozhodování ve všech oblastech, od podnikání po vědecké výzkumy. Také jednotlivá města a státy řídí své činnosti právě ze získaných dat, které mají k dispozici a na jejich základě konají různá rozhodnutí a stanovují strategické cíle. Možnost pracovat s daty a extrahovat z nich užitečné informace, na jejichž základě budou vytvořeny predikční modely, které mají obrovský potenciál přinést nové inovace a zlepšit efektivitu v různých oborech. Data Science, česky datová věda, je stále velmi rozvíjejícím se oborem, díky novým technologiím, které každoročně přichází, a proto je stále možnost přispět s novými myšlenkami a přístupy, jež je možné začlenit do životního cyklu tohoto oboru.

Diplomová práce je rozdělena na teoretickou a praktickou část. V teoretické části se práce nejdříve zaměřuje na samotné vysvětlení pojmu Data Science. Dále je věnována podkapitola vývoji datové vědy, kde je stručně popsána historie vývoje tohoto oboru. Jsou zmíněny také možnosti využití, se kterými se lze setkat v běžném světě. Poměrně velká část teoretické práce je věnována možnosti vizualizace dat, kde jsou popsány různé typy grafů, které se nejčastěji využívají pro vizualizaci dat v praxi. Další důležitou částí teoretické práce je popis životního cyklu, metodiky a procesů datové vědy, jež se běžně používají v tomto oboru. V poslední části teoretické části jsou představeny jedny z nejčastěji používaných nástrojů při práci datového vědce. Pro každou technologii jsou popsány různé balíčky, tyto balíčky jsou určeny a často využívány právě ve spojení s datovou vědou.

Praktická část si klade za cíl provést analýzu totožných historických dat a jejich následnou vizualizaci. Po celou dobu by měla praktická část dodržovat metodiky, které byly představeny v teoretické části. V úvodu praktické části bude představen úvod do řešené ukázkové problematiky, kde bude definován problém, kterým se bude analýza poté zabývat. Dále se bude soustředit na technologie využívané pro analýzu dat. V další části pak budou uvedeny a popsány vybrané datové sady, na kterých budou analýzy prováděny. Součástí úvodu do analýzy je i představení výběru výzkumných otázek. V navazující části bude uveden popis postupu zpracování dat pro jednotlivé výzkumné otázky. Demonstrace výsledků analýzy bude formou jednotlivých vizualizací. V závěru práce bude uvedeno vyhodnocení výsledků plynoucích

z provedených analýz. Praktická část bude uzavřena celkovým zhodnocením praktické části a možnými doporučeními.

1 DATA SCIENCE

Pojem Data Science může být pro běžného čtenáře naprosto neznámým pojmem. Často se můžeme setkat také s českým překladem „datová věda“. Přesto je často tento pojem běžnému čtenáři, nepohybujícímu se ve světě informačních technologií, datové analytiky, byznysu či jiných blízkých odvětví, nic neříkající. V následujících podkapitolách bude tento pojem vysvětlen jako vstupní úvod do problematiky.

Jak uvádí Jan Hendl (2021), vědu o datech pojímáme jako interdisciplinární oblast, která využívá vědecké metody, procesy a algoritmy k extrahování znalostí a získání vhledu do daného problému pomocí mnoha strukturovaných a nestrukturovaných dat. Data Science kombinuje znalosti z matematiky, statistiky, programování, pokročilé analytiky, umělé inteligence a strojového učení se specifickými znalostmi z daného oboru (Hendl, 2021). Získané informace pak mohou být klíčové v chodu organizace. Díky nim lze činit kvalitnější rozhodnutí pro chod společnosti, zvyšovat provozní efektivitu, identifikovat nové obchodní příležitosti nebo zlepšovat marketingové a prodejní programy. Mezi hlavní benefity pak patří konkurenční výhoda nebo úspora financí, pokud jsou vhodně využity získané znalosti (Stedman, 2018).

Častou definicí pojmu Data Science bývá Vennův diagram od Drew Conwaye (viz. Obrázek 1), který se často vyskytuje v mnoha publikacích. Tento diagram se skládá ze tří hlavních částí, kterými jsou Hacking Skills, Math & Statistics Knowledge a Substantive Expertise. Dalšími částmi jsou Machine Learning, Traditional Research a Danger Zone. Poslední částí je Data Science, která je kombinací tří hlavních složek (Stedman, 2018).

Pojem Hacking Skills může být zprvu velice zavádějící. Ve skutečnosti se nejedná o žádné hackování nebo schopnosti napadnout systém či jiného zneužití systému. Hlavními dovednostmi jsou manipulace s datovými soubory, algoritmizace a zájem o učení nových technologií a nástrojů (Stedman, 2018).

Math & Statisctics Knowledge, zde již lze naopak z názvu odhadnout o co se jedná. Používají se matematické a statistické vzorce, postupy k vyčištění dat, analýze a následnému zpracování (Stedman, 2018).

Třetí hlavní část Substantive Expertise se zaměřuje na znalosti z daného oboru. Aby bylo možné objevovat nové věci a znalosti, je potřeba definovat vhodné hypotézy. A dále je nutné, aby hypotézy byly dostatečně kvalitní, je třeba mít dostatečné povědomí o oboru, který je zkoumán.

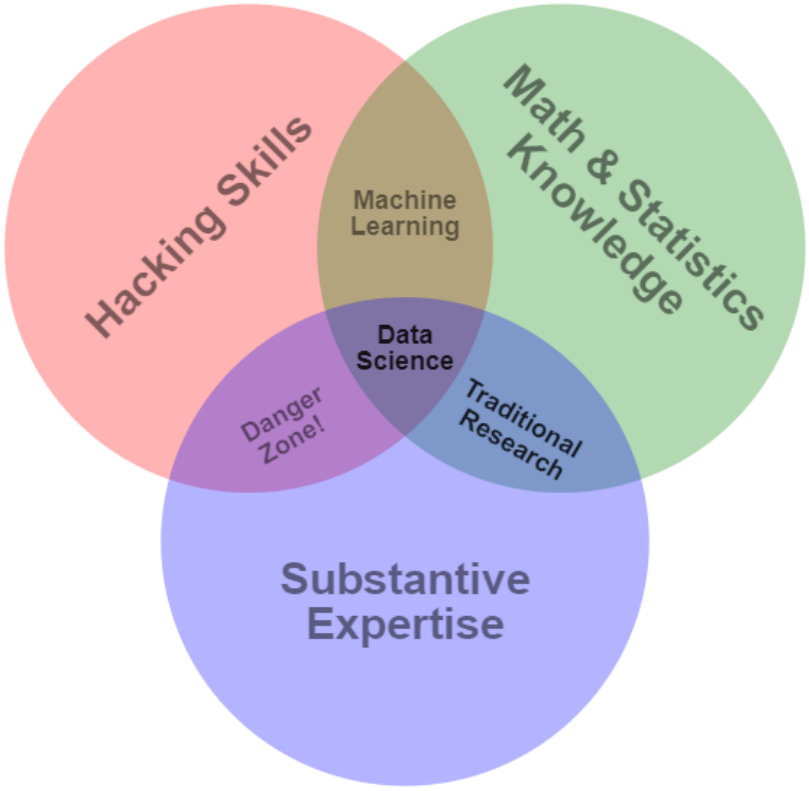
Jakmile jsou hypotézy stanoveny, lze začít provádět testování statistickými metodami (Stedman, 2018).

Machine Learning neboli strojové učení využívá algoritmů a dat, pomocí kterých se napodobuje učení, jako probíhá u lidí. Tato disciplína se stává důležitou součástí Data Science (IBM, nedatováno).

Traditional Research se týká vědeckých pracovníků, kteří pro své výzkumy nepoužívají nové technologie. Většinou jsou omezeni pouze svými znalostmi z oboru, kterému se věnují.

Předposlední částí je Danger Zone. Pojmenování této části by se dalo brát za nešťastné, protože se ve skutečnosti nejedná o žádné nebezpečí, pouze autor ho tak vnímá. Podle něj sem patří lidé, kteří vědí dost na to, aby byli nebezpeční a jedná se o nejproblematičtější část tohoto diagramu. Dovednosti těchto lidí umožňují vytvářet analýzy, které se mohou jevit jako kvalitní. Protože ale neovládají matematické a statistické dovednosti, tak nejspíše dostatečně nerozumí tomu, co vlastně vytvořili a jak se k těmto výsledkům dopracovali (Conway, 2010).

Všechny tyto jednotlivé části společně tvoří oblast Data Science.



Obrázek 1: Data Science Venn Diagram (Conway, 2010)

1.1 Vývoj v oblasti datové vědy

Data Science je poměrně mladým oborem, který se začíná dostávat do podvědomí v posledních dvou dekádách. První zmínky je možné dohledat již v 60. letech 20. století, kdy vzniká nová profese, která by měla za úkol porozumět a interpretovat velké množství dat. Mnoho zdrojů uvádí, že se jedná o disciplínu, která vzniká spojením dvou oborů, a to statistiky a informatiky.

V roce 1962 napsal statistik a matematik John Tukey publikaci „The Future of Data Analysis“, ve kterém uvádí, že začíná mít dojem, že jeho hlavním zájmem se stává datová analýza. Datovou analýzou má namysli spojení statistiky a informatiky, kdy se pro matematické výpočty začínají používat počítače. Také uvádí, že počítače nejsou pro tyto operace příliš nutné, ale jsou důležité.

O dvanáct let později v roce 1974 Peter Naur publikoval knihu „Concise Survey of Computer Methods“, ve které se často objevuje spojení Data Science. V předmluvě knihy je zmíněno, že na kongresu IFIP v roce 1968 byl představen kurz s názvem „Datalogie, věda o datech a datových procesech a její místo ve vzdělání“. Naur v knize poskytuje také vlastní definici, jako „Užitečnost dat a datových procesů vyplývá z jejich použití při vytváření a zpracování modelů reality“.

V roce 1977 byla založena IASC, která oznámila, že jejich posláním je propojit tradiční statistickou metodologii a moderní počítačové technologie a znalosti odborníků v dané oblasti, s cílem přeměnit data na informace a znalosti. Téhož roku vydal John Tukey druhý článek s názvem „Exploratory Data Analysis“, ve kterém uvádí, že je třeba klást větší důraz na výběr dat k navrhování hypotéz. Také tvrdí, že explorační analýza dat a konfirmační faktorová analýza by měly probíhat zároveň.

Roku 1989 Gregory Piatetsky-Shapiro organizuje a předsedá prvnímu workshopu KDD. O pár let později se z KDD stává každoroční konference ACM SIGKDD. Tato konference slouží jako fórum pro pokrok, vzdělání a přijetí vědy o objevování znalostí a dolování dat ze všech typů dat uložených v počítačích a počítačových sítí (KDD, 2022).

V září 1994 časopis Businessweek vydal článek, že společnosti začaly shromažďovat velké množství osobních údajů a plánují odstartovat nové podivné marketingové kampaně. Pro mnoho manažerů byla data nic neříkající a nevěděli, jak z nich čerpat užitečné informace (Press, 2013).

O 5 let později roku 1999 upozornil Jacob Zahavi ve své publikaci, „Mining Data for Nuggets of Knowledge“ na potřebu nových nástrojů pro zpracování obrovského a neustále rostoucího množství dat. Zmiňuje, že bude nutný vývoj modelů, které budou schopny lépe analyzovat data, odhalovat nelineární vztahy a interakce mezi jednotlivými prvky. Dále také píše, že možná bude nutné vyvinout speciální nástroje pro dolování dat, které se budou zabývat rozhodováním na webových stránkách (Press, 2013).

Mezi rokem 1999 až 2005 vyšla spousta publikací, studií nebo časopisů, které se zabývaly problematikou Data Science. Větší pokrok však přišel v roce 2006, kdy byl představen nerelační databáze Hadoop 0.1.0, která byla vydána jako otevřený software (Press, 2013).

V roce 2008 DJ Patil a Jeff Hammerbacher ze společností LinkedIn a Facebook zpopularizovali pojem „Data Scientist“, kdy pojmenovali skupinu na LinkedIn právě „Data Scientist“ (Gorelik, 2019).

O tři roky později v roce 2011 počet pracovních nabídek pro datové vědce zvýšil o 15 000 %. Důsledkem toho se zvyšovaly semináře, konference týkající se právě Data Science. O rok později vydala Harvardova Univerzita článek, ve kterém uvádí, že datoví vědci mají nejvíce atraktivní práci 21. století (Foote, 2021).

Většina těchto událostí měla zásadní vliv na vznik a vývoj Data Science. Stále se však jedná o mladý obor, který se bude v následujících letech ještě vyvíjet a formovat.

V dnešní době se dostává do popředí umělá inteligence, která dosáhla v současnosti mnoha významných pokroků. Stává se stále sofistikovanější a schopnější řešit komplexní úkoly. Díky tomu mohou datový vědci automatizovat různé pracovní postupy, jako je například čištění dat, trénování modelů nebo interpretace výsledků. Umělá inteligence prokazuje svůj potenciál při analýze nestrukturovaných dat, zpracování přirozeného jazyka, generování dat nebo generování počítačového kódu. Díky možnostem, které umělá inteligence přináší se mohou datový vědci soustředit na složitější úkoly, kdy je potřeba se rozhodovat nejenom na základě dat. Mohou se věnovat složitějším úkonům, které umělá inteligence nemusí tak dobře zvládat, jako je například vytváření přesnějších prediktivních modelů nebo vylepšení vizualizací. Datový vědci v budoucnu budou muset umět dobře ovládat nástroje umělé inteligence, které jim usnadní velkou část jejich práce (Hassani, 2023).

1.2 Využití Data Science

Většina velkých společností má v dnešní době přístup k velkému množství dat. Pokud jsou tyto data dobře zpracována a využita, může to mít zásadní vliv na jejich fungování. Může se jednat například o různé měření metrik, analyzování trendů nebo přijímat zásadní rozhodnutí, která budou mít vliv na budoucí chod podniku. Společnost Walmart, která se zaměřuje na maloobchod využívá již roky Data Science, aby zvýšila své příjmy. Díky rozsáhlému množství dat o preferencích svých zákazníků jsou schopni nabízet zákazníkům produkty o které mají zájem. Jsou schopni na základě těchto dat, také optimalizovat zásoby v jejich skladech. Walmart zaznamenal nárůst příjmů o 10 až 15 % (Kelleher, 2018).

Velice podobně cílí také Netflix na své zákazníky. Kdy je schopen dle preferencí daného uživatele a preferencí dalších uživatelů, kteří mají podobný vkus doporučovat seriály nebo filmy, které jsou pro ostatní zajímavé.

Data Science prorazila také do sportu. V dnešní době, má každý profesionální tým svého datového analytika, který zodpovídá za různé statistiky týkající se jednotlivých hráčů, jejich postů nebo statistik celého týmu a ostatních týmů v soutěži. Zajímavostí je, že vznikl také film dle skutečných událostí s Bradem Pittem, který se jmenuje Moneyball. Film pojednává o tom, jak baseballový tým v USA začal využívat Data Science pro nábor hráčů. Stanovili si kritéria, která jsou pro ně nejdůležitější a na základě těchto výsledků přivedli do týmu podceňované hráče, kteří byli úspěšní (Kelleher, 2018).

Může se jednat o byznys, který se zabývá vývojem nových produktů nebo jejich vylepšením a řízením dodavatelských řetězců. V zábavném průmyslu, jako byl zmiňovaný Netflix, kdy lze cílit obsah na základě preferencí, měřit různé metriky o sledování anebo vytvářet nový obsah. Ve finanční sféře, kde lze předcházet různým podvodům, porušení zabezpečení nebo lze řídit rizika investičních portfolií. Ve zdravotnictví můžeme například sledovat výskyt nemocí v reálném čase, léčit pacienty dle podobnosti symptomů či vyvíjet nové léky. Obecně by se dalo tedy říct, že Data Science lze využívat v dnešní době úplně všude, kde je potřeba činit nějaká rozhodnutí, která budou mít zásadní vliv pro chod organizace (Microsoft, © 2023).

Za těmito všemi rozhodnutími stojí osoby, které nazýváme Data Scientists neboli datoví vědci. Datový vědec by měl být schopen analyzovat a následně smysluplně interpretovat data, které má k dispozici. Tato data jsou většinou uložena v datových skladech nebo datových centrech a následně využita pro řešení obchodních problému, optimalizace výkonu a shromažďování informací, které poté využívají datový vědci (Rouse, 2020). V těchto datech by měl být schopen

dobře číst, odhalovat v nich vzorce, poznatky a vytvářet z nich predikce, které budou pro cílovou skupinu užitečné. Při své práci se setkávají jak se strukturovanými, tak s nestrukturovanými daty. V případě strukturovaných dat se může jednat například o jména, kalendářní data, platební údaje či jiná tabulková data, která jsou rozdělena do řádků a sloupců. V případě dat nestrukturovaných může jít o celé texty z dokumentů, data ze sociálních sítí, informace z mobilních zařízení, obsah internetových stránek či videa. Za účelem efektivního zpracování velkého množství dat a získání z nich užitečných informací, potřebuje datový vědec širokou paletu znalostí a dovedností. Mezi tyto dovednosti můžeme zařadit schopnost programování, matematické a statistické znalosti, a také schopnost pracovat s pravděpodobností. Dále by měl mít znalosti z oblasti, ze které pocházejí data, aby byl schopen je správně interpretovat. Vizuální prezentace dat je důležitá, a proto by měl datový vědec data správně a efektivně vizualizovat. Nakonec by měl dobře komunikovat a veřejně prezentovat své poznatky a výsledky (Microsoft, © 2023).

1.3 Vizualizace dat

Tato kapitola se zabývá jedním z klíčových prvků Data Science, a to je vizualizace dat. Firmy pravidelně získávají velké množství nových dat, které jsou často ukládány v surovém stavu a bez dalšího zpracování z nich nelze vyčíst důležité informace. Vizualizace dat je klíčovým nástrojem, který umožňuje lépe a rychleji porozumět těmto datům (Microsoft, © 2023).

Před vytváření samotných vizualizací, je důležité provést řadu určitých kroků, které tomu předchází. Často se lze setkat s definicí, která rozděluje životní cyklus do pěti fází. Dle Kalifornské Univerzity Barkley je prvním krokem „Capture“ neboli zachycení. Tento krok se zaměřuje na sběr nezpracovaných strukturovaných a nestrukturovaných dat z různých zdrojů. Druhým krokem je „Maintain“ což znamená údržba. Zde jsou data sestavena a zpřístupněna v konzistentním formátu pro analýzu, strojové učení nebo modely hlubokého učení. V tomto kroku se také provádí čištění dat, odstranění duplicit a přeformátování. Třetím krokem je „Process“ neboli zpracování. V tomto kroku se datoví vědci zabývají zkoumáním vzorů, rozsahu a zkreslení dat. Předposlední fází je „Analyze“ neboli analýza. V této fázi se provádí analýza dat, přičemž datoví vědci využívají statistickou analýzu, prediktivní analýzu, regresi, strojové učení a algoritmy hlubokého učení. Posledním krokem je „Communicate“ neboli komunikace. Zde se vyskytuje i samotná vizualizace dat (Daley, 2023). Datoví vědci mají za úkol prezentovat své zjištění pomocí tabulek, grafů, zpráv a všech možných vizualizací. Právě samotná vizualizace usnadňuje lepší a snadnější pochopení jejich výsledků (Berkley, © 2023).

Vizualizace dat je proces, kdy jsou vytvářeny grafické reprezentace z konkrétních informací, které jsou k dispozici. Tento proces pomáhá ke snadnějšímu pochopení prezentovaných dat. Samotná vizualizace může být statická nebo dynamická. K tomuto účelu existují různé nástroje a techniky, pomocí kterých lze data vizualizovat. Mezi základní a nejdůležitější techniky patří Pie Chart, Bar Chart, Histogram, Gantt Chart, Heat Map, Box and Whisker Plot, Waterfall Chart, Area Chart, Scatter Plot, Pictogram Chart, Timeline, Higlight Table, Bullet Graph, Choropleth Map, Word Cloud, Network Diagram a Correlation Matrices. Je důležité, aby datový vědec ovládal tyto nejzákladnější techniky. V důsledku pak může prezentovat data v co nejjasnější a nejpřehlednější podobě. Na internetu nebo v praxi je možné se setkat i s jinými techniky vizualizace, které jsou však méně používané nebo se používají pro specifické účely. V následující části budou krátce představeny vybrané techniky vizualizace.

Pie chart (Koláčový graf), také známý jako Circle graph (Kruhový graf) je jedním z nejběžnějších a nejzákladnějších typů vizualizace dat, který se dá použít v mnoha situacích. Tento graf je velice vhodný právě pro zobrazení poměrů nebo porovnání jednotlivých částí. Tyto typy grafů jsou velice jednoduché, a proto není většinou potřeba znát širší souvislosti dat a bude zcela jasné, co nám mají říct. Pokud by data měla vysvětlovat složitější souvislosti, je vhodnější zvážit použití jiného typu grafu (Miller, 2019).

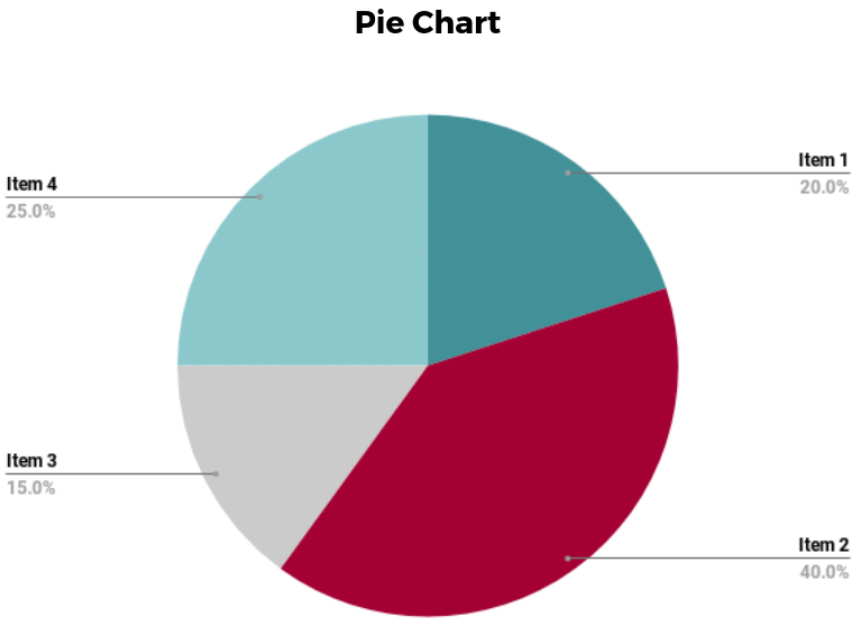
Výhody

- Jednoduchý a intuitivní formát grafu, který je snadno srozumitelný pro většinu lidí.
- Zobrazuje data jako zlomkové části z celku, což může být účinný komunikační nástroj i pro neinformované publikum.
- Okamžitě umožňuje pochopení a srovnání dat v rámci celkového souhrnu (Rajasekhar, © 2023).
- Může vizualizovat velké množství dat a poskytnout přehled o celkovém rozložení kategorií (Data Presentation: Pie Charts, 2022).

Nevýhody

- Není vždy snadné získat přesné hodnoty pro každou kategorii z grafu, což může vést k nesprávné interpretaci dat.
- Pokud chceme vizualizovat změny v čase, je třeba použít více samostatných grafů.
- Nemusí vystihovat klíčové informace o datech zejména v případech, kdy se jsou data složitější (Data Presentation: Pie Charts, 2022).

- Při velkém počtu kategorií se Pie Chart stává méně čitelným a méně efektivním nástrojem pro vizualizaci dat.
- Záporné hodnoty mohou být obtížné na vizualizaci a mohou vést k nejasnému nebo nesprávnému pochopení dat (Rajasekhar, © 2023).



Obrázek 2: Ukázková vizualizace Pie chart (Miller, 2019)

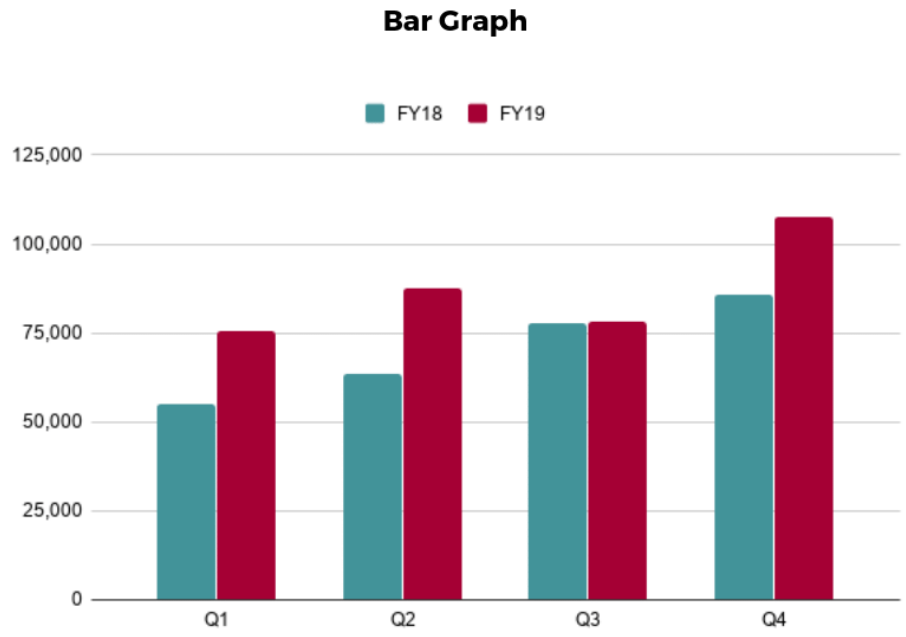
Bar chart (Sloupcový graf) je dalším často používaným typem grafu. Tento graf se používá k porovnání relativních velikostí různých kategorií a zobrazuje je pomocí sloupců, které mohou být buď svislé nebo vodorovné. Bar chart také obsahuje dvě osy, kdy osa x zobrazuje porovnávané kategorie a osa y hodnoty nebo četnost dané kategorie. Sloupcový graf se často používá k porovnání dat mezi různými skupinami, jako jsou různé produkty, regiony nebo časová období (Miller, 2019).

Výhody

- Sloupcové grafy dobře ukazují trendy a pomáhají porovnávat data.
- Jsou snadno srozumitelné i pro ty, kteří nejsou odborníky na data.
- Shrnuje velká data a umožňuje odhalit klíčové hodnoty na první pohled.
- Umožňuje dobře vizuálně kontrolovat přesnost jednotlivých sloupců v grafu.
- Dobře zobrazuje relativní počty nebo podíly více kategorií (Advantages and Disadvantages of Bar Graphs, 2022).

Nevýhody

- Je potřeba dodatečných informací, které vysvětlí, co data v grafu znamenají.
- Pokud jsou nesprávně vytvořené, mohou být grafy zavádějící a mohou být snadno manipulující.
- Není vhodný, pokud data obsahují velké množství kategorií.
- Neodhalí širší souvislosti, jako jsou příčiny či důsledky (Advantages and Disadvantages of Bar Graphs, 2022).



Obrázek 3: Ukázková vizualizace Bar chart (Miller, 2019)

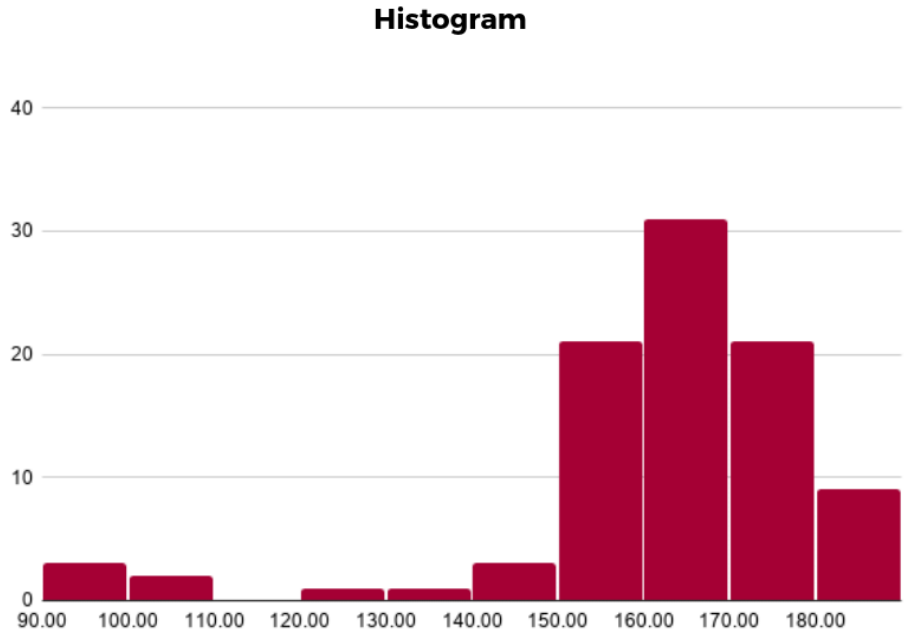
Histogram je grafický způsob zobrazení dat, který zobrazuje, jak jsou data rozdělena do určitých intervalů pomocí sloupců. Tyto intervaly mohou být různé velké, ale jejich hodnoty by měly být spojité. Graf má dvě osy x a y. Jedna osa ukazuje rozsah hodnot a druhá osa ukazuje četnost hodnot v tomto rozsahu. Velikost jednoho sloupce představuje počet hodnot v daném rozsahu. Jednotlivé sloupce na histogramu by měly být bez mezer a měly by se navzájem dotýkat. Pokud se v histogramu nachází mezera mezi sloupci, znamená to, že v daném rozsahu nebyla zaznamenána žádná hodnota (Bar Chart Vs Histogram: What Are The Key Differences, 2023).

Výhody

- Jedná se o jeden z nejoblíbenějších a nejčastěji používaných technik pro vizualizaci dat.
- Na rozdíl od sloupcového diagramu je šířka i délka histogramu důležitou informací o datech.
- Oproti jiným grafům může zobrazovat velký počet sloupců neboli kategorií (Advantages and Disadvantages of Histogram, 2022).

Nevýhody

- Histogram je omezený pouze na specifický typ vizualizací.
- Je vhodný pouze pro zobrazení spojitých rozdělení a nelze ho použít pro diskrétní rozdělení.
- Na rozdíl od jiných typů grafů, jako například Box Plot nám neposkytne informace jako medián, horní kvantil nebo dolní kvantil.
- Nelze ho použít pro porovnání dvou různých souborů dat.
- Neumožňuje vypočítat průměr ani medián jen modus (Advantages and Disadvantages of Histogram, 2022).



Obrázek 4: Ukázková vizualizace histogramu (Miller, 2019)

Gantt chart je druh diagramu, který byl pojmenovaný po Henry Ganttovi, který tento graf zavedl, aby zlepšil plánování, rozvrhování a sledování projektů. V dnešní dobře je často využívaný projektovými manažery pří řízení projektu (Gantt Chart: Definition and Examples,

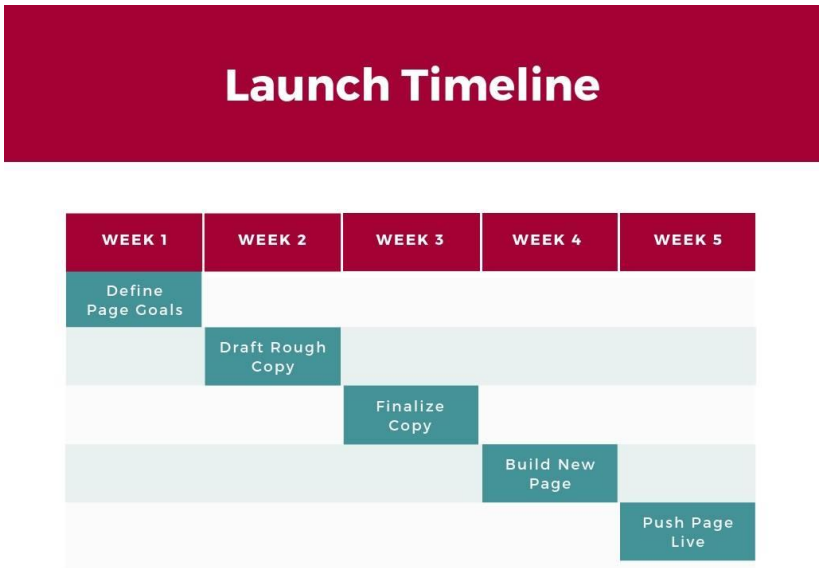
© 2023). Na svislé ose grafu jsou zobrazeny úkoly, které jsou potřeba provést a na vodorovné ose jsou zobrazeny časové intervaly. Vodorovné sloupce v grafu představují dobu trvání jednotlivých činností, které jsou naplánovány. I když není člověk odborník v řízení projektu, může Ganttův graf pomoci zvýšit produktivitu a udržet si přehled o činnostech (Miller, 2019).

Výhody

- Ganttův graf poskytuje přehledné informace o projektu a jeho časovém harmonogramu, což může být užitečné pro rychlé pochopení celkového průběhu projektu.
- Může zlepšit odpovědnost a efektivitu komunikace mezi členy týmu a dalšími zainteresovanými stranami projektu.
- Graf poskytuje důmyslnější, realističtější a praktičtější možnosti plánování projektu, které může vést k vylepšení plánování organizace projektu.
- Umožňuje sledovat pokrok projektu a určit, zda se pracuje podle plánu nebo zda jsou nutné úpravy (Freeman, © 2023).

Nevýhody

- Aktualizace Ganttova grafu může být časově náročná a vyžaduje pravidelné úpravy, aby byl v souladu s aktuálním stavem projektu.
- Poskytuje pouze vysokou úroveň pohledu na projekt, a to může být problém, pokud projekt obsahuje mnoho složitých úkolů a podúkolů.
- Slouží primárně k plánování projektu a sledování jeho pokroku. Pokud jsou potřeba další funkce, může být nutné použít další nástroje (Freeman, © 2023).



Obrázek 5: Ukázková vizualizace Gantt chart (Miller, 2019)

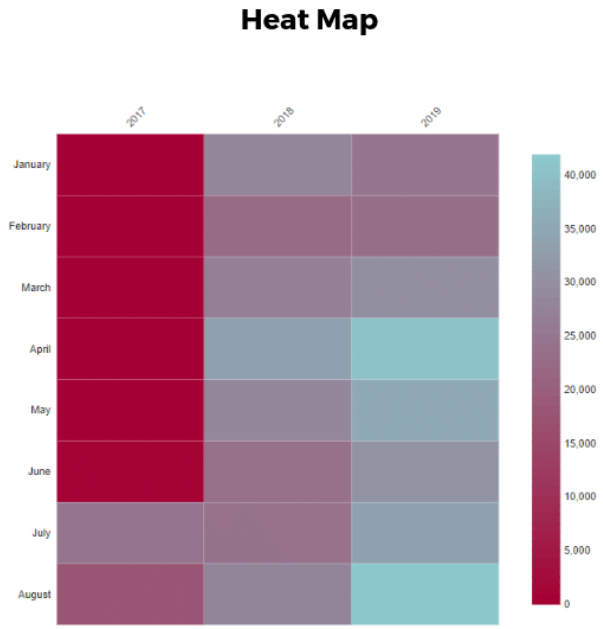
Heat map (Tepelná mapa) je typ vizualizace, která se používá k zobrazení rozdílů v datech pomocí různých odstínů barev. Rozdílné hodnoty jsou zobrazovány různými odstíny barev, jako například odstíny zelené pro příznivé hodnoty a červené pro méně příznivé hodnoty. Barva pro hodnoty mezi těmito dvěma extrémy se může lišit podle autora a většinou je popsána v legendě, která je součástí grafu (Miller, 2019). Například lze použít Heat map pro analýzu, kdy má maloobchodní prodejna největší tržby v závislosti na dni v týdnu a čase v daný den. Podle barvy pak je možné přesně odhalit, v jaké hodiny má prodejna největší tržby. V dnešní době se také často tepelné mapy používají na internetu, kdy je možné s jejich pomocí zjistit s jakou částí stránky uživatelé nejčastěji interagují (Morris, 2021)

Výhody

- Dokáže dobře optimalizovat umístění potřebných prvků na webové stránce nebo v aplikaci.
- Lze díky Heat mapám činit rychlá a efektivní rozhodnutí.

Nevýhody

- Porovnávat sytost jednotlivých barev může být složité. Každá osoba navíc může barvy vnímat trochu jinak.
- Graf ovšem neposkytne mnoho podrobností. (Grover, 2022)



Obrázek 6: Ukázková vizualizace Heat map (Miller, 2019)

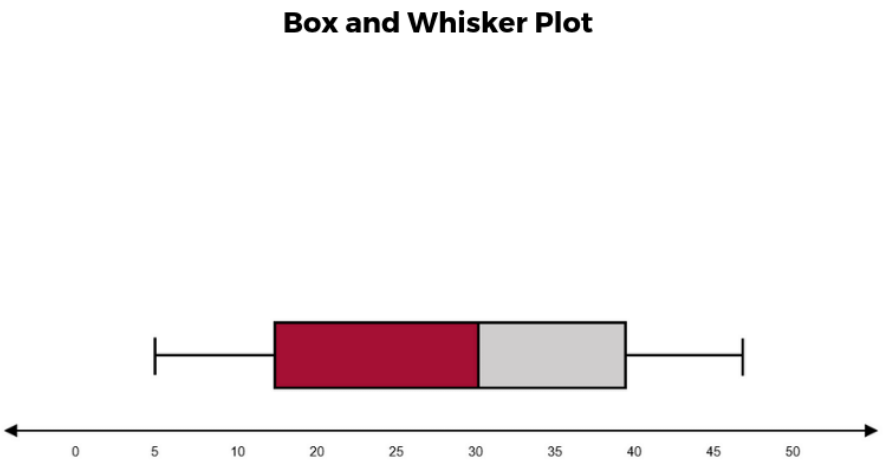
Box plot neboli **Box and Whisker plot** (Krabicový graf) slouží k vizuálnímu shrnutí dat prostřednictvím kvartilů. Tento graf umožňuje zobrazit velké množství informací v malé grafické podobě. Nejprve se vytvoří krabice, která představuje 50 % hodnot v daném souboru dat. Vertikální čára uprostřed představuje medián. Hranice krabice znázorňují první a třetí kvartil. Poté se nakreslí čáry, které se nazývají „fousky“, od prvního kvartilu a od třetího kvartilu souboru dat, které se táhnout k minimu (dolnímu extrému) a maximu (hornímu extrému). Odlehlé hodnoty jsou reprezentovány jednotlivými body, které jsou v linii s „fousky“. Celkově Box plot poskytuje přehlednou a srozumitelnou vizualizaci rozložení dat v daném souboru (Miller, 2019).

Výhody

- Krabicový graf je užitečný pro porovnání rozsahu a rozdělení číselných dat do skupin.
- Dobře organizuje velké objemy dat a zobrazuje odlehlé hodnoty, což umožňuje identifikovat extrémní hodnoty v datech (Box plot, 2023).

Nevýhody

- Není vhodný pro podrobnou analýzu dat, protože se zabývá pouze shrnutím rozložení dat (Box plot, 2023).
- V grafu nejsou zobrazeny původní data, takže některé informace jako průměr, modus nebo rozptyl.
- Může se skládat pouze z číselných hodnot a nelze ho použít pro vizualizace jiných druhů dat (Advantages & Disadvantages of Box Plots, nedatováno).



Obrázek 7: Ukázková vizualizace Box plot (Miller, 2019)

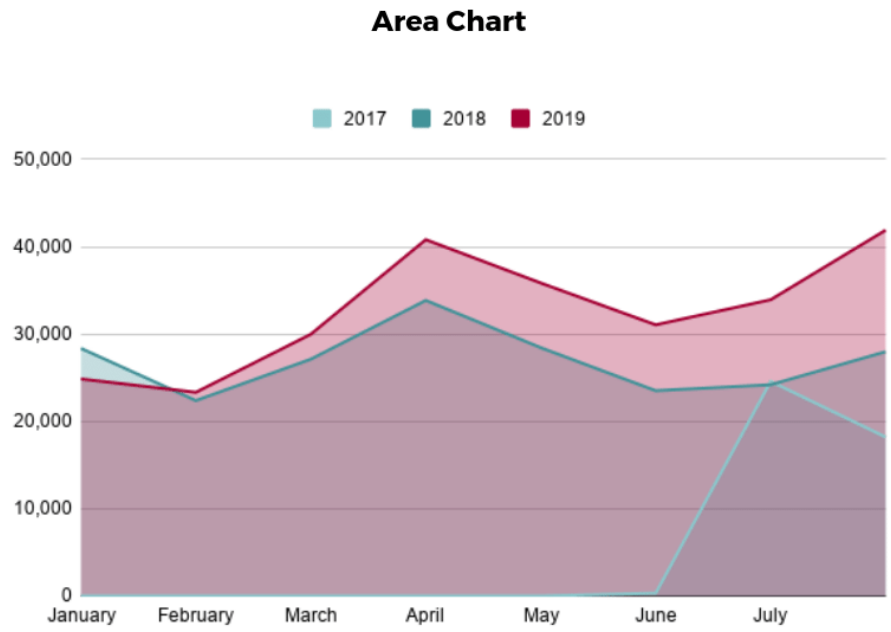
Area map, chart nebo graf (Plošný graf) s těmito názvy se můžeme v praxi setkat. Jedná se o variantu jednoho z nejběžnějších typů grafu, kterým je Line chart (Spojnicový graf). Hlavní rozdíl mezi spojnicovým grafem a plošným grafem spočívá v tom, že plošný graf má pod čarou stínování (Freeman, © 2023). Toto stínování pomáhá při porovnávání řad dat, protože umožňuje lépe si představit rozdíly mezi nimi. Pokud je potřeba porovnat více datových řad v jednom grafu, používá se tzv. skládání plošných grafů. Plošné grafy jsou často používány pro zobrazení změn jedné nebo více veličin v čase a zobrazení trendů (Miller, 2019).

Výhody

- Data jsou velice jednoduchá na pochopení a na sledování.
- Skvělé pro zobrazení trendů v čase a analýzu statistik.
- Díky velikosti plochy rychle rozpoznáme, jaká soubor dat v čase převládá (Freeman, © 2023).

Nevýhody

- Nelze sledovat přesné hodnoty.
- Neefektivní při porovnávání velkých skupin dat, protože plochy se mohou překrývat a je obtížné rozeznat, kde končí a začínají nové jednotlivé plochy (Freeman, © 2023).



Obrázek 8: Ukázková vizualizace Area chart (Miller, 2019)

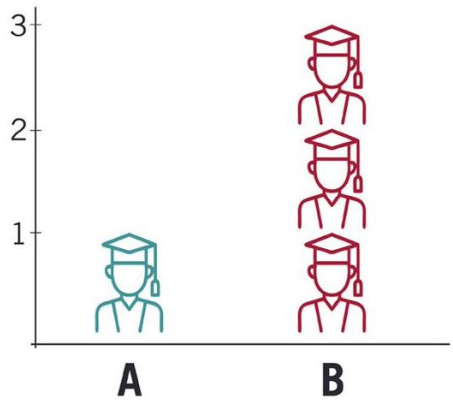
Pictogram chart (Piktogram) je typ vizualizace, který se používá pro jednoduchá data. Díky svému velice poutavému vzhledu a správné volbě ikon je snadné si představit, co přesně data znamenají. Každá ikona může zastupovat jednu jednotku nebo libovolný počet jednotek. Například jedna ikona může představovat desítky či stovky dat. Soubory dat se typicky porovnávají vedle sebe buď ve sloupcích nebo v řádcích, aby bylo možné porovnat jednotlivé kategorie mezi sebou. Použití vhodných ikon může pomoci překonat jazykové, kulturní a vzdělanostní rozdíly (Miller, 2019), (Cetiner, nedatováno).

Výhody

- Piktogramy jsou snadno čitelné a intuitivní, protože používají obrázky pro znázornění čísel.
- Jedná se o velmi univerzální graf, kdy mu budou rozumět všichni nehledě na jazyk nebo například inteligenci.
- Jsou to velice efektivní výukové nástroje, protože obrázky jsou nenáročné na pochopení, a proto se nejčastěji používají pří výuce dětí.
- Dodatečné informace mohou být získány z velikosti, barvy a dalších vlastností ikon (All Things Statistics, 2022).

Nevýhody

- Velký počet kategorií v datech by mohl vést k nepřehlednosti skrze nutnost použít velký počet různých ikon.
- Špatně zvolené ikony mohou lehce vést k nejasnosti grafu.
- V piktogramech nelze zobrazit zlomkové hodnoty ani záporné hodnoty.
- Nejsou vhodné pro obchodní účely, protože mohou působit příliš jednoduše (All Things Statistics, 2022).



Obrázek 9: Ukázková vizualizace piktogramu (Miller, 2019)

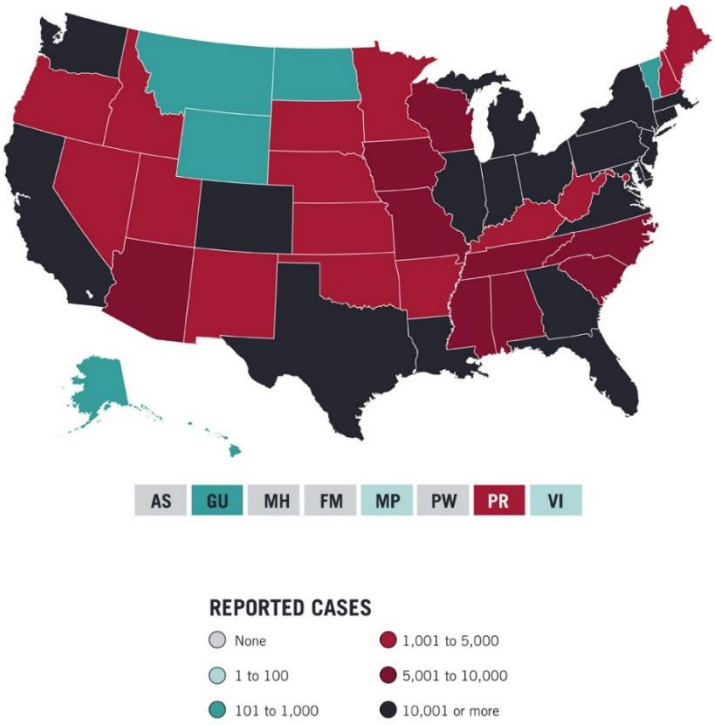
Choropleth maps (Kartogram) zobrazují rozdělené zeměpisné oblasti nebo regiony, které jsou vybarvené, stínované nebo vzorované v závislosti na vstupních datech. Tyto mapy umožňují vizualizovat číselné hodnoty v různých geografických oblastech a ukázat rozdíly, souvislosti nebo vzorce na základě těchto dat. Často se s kartogramy používají interaktivní prvky, které umožňují získat rozšířené informace o konkrétních oblastech, kdy například po najetí myší na oblast lze získat data týkající se dané oblasti (Miller, 2019).

Výhody

- Jsou vizuálně velice efektivní a lze z nich vyčíst velké množství informací a obecných vzorců.
- Používají se pro reprezentaci map, zemí, regionů nebo městských částí.
- Jedná se o jednu z nejsnáze pochopitelných možností vizualizace a lehce přitáhne pozornost lidí (Civil, 2022).

Nevýhody

- Jednou z nevýhod je to, že z barev nelze přesně vyčíst konkrétní hodnoty.
- Dalším problém je to, že větší oblasti se mohou jevit více významné než menší oblasti.
- Častou chybou je kódování hodnot surových dat namísto použití normalizovaných hodnot.
- V některých případech mohou hranice hrát velkou roli a nemusí být přesně jasné, z jaké části státu se data vyskytují (Civil, 2022).



Obrázek 10: Ukázková vizualizace Choropleth map (Miller, 2019)

Word cloud (Slovní mrak), také známý jako Tag Cloud, je způsob vizuálního zobrazení textových dat. Textová data jsou zobrazena jako slovní mrak, kde každému slovu je přiřazena jeho významnost na základě toho, jak často se vyskytuje v textu (Betterevaluation, © 2022). Častěji se vyskytující slova jsou v grafu zobrazena větší, tlustší nebo jinak výraznější formou než méně častá slova. Tento typ vizualizace se používá například k reprezentaci slov, kterými zákazníci popsali daný produkt (Miller, 2019).

Výhody

- Word Cloud umožňuje snadné získání podstatných informací z velkého množství dat.
- Tento typ vizualizace je obvykle poutavější a může přitáhnout zájem publika.
- Může podporovat myšlenky, odhalovat vzorce a pomáhat publiku lépe porozumět datům (Bush, 2021).

Nevýhody

- Vizualizace je navržena tak, aby slova byla zvýrazněna podle své velikosti, takže délka slova může zkreslovat důležitost slova.

- Některé generátory pro Word Cloud nemusí dobře rozumět textu a mohou přiřazovat barvy náhodně, což může vést ke špatnému porozumění důležitosti dat.
- Použití různých fontů může snížit čitelnost vizualizace a snížit význam některých slov.
- Nastavení generátoru je důležité, aby zohlednil pouze významná slova, a ne například předložky, spojky a další obecné slova, které by mohly ovlivnit výslednou vizualizaci (Alida, 2012).

Obrázek 11: Ukázková vizualizace Word cloud (Miller, 2019)

Network diagram (Síťový diagram), často označovaný pouze jako graph, zobrazuje vztahy mezi jednotlivými entitami, které jsou v diagramu znázorněny jako uzly spojené hranami. Každý uzel reprezentuje jednu datovou položku a hrany znázorňují vztahy mezi těmito položkami. Uzly se obvykle kreslí jako tečky nebo kolečka a hrany jako jednoduché čáry. Existují dva hlavní typy síťových diagramů: orientované a neorientované. Rozdíl mezi nimi je, že orientovaný diagram používá šipky na reprezentaci směru vztahů mezi uzly (Miller, 2019).

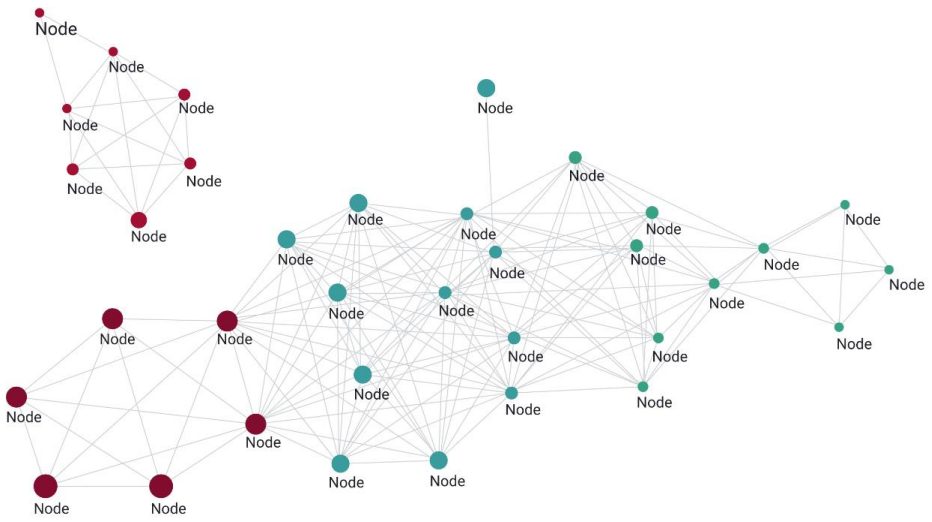
Výhody

- Reprezentace složitých dat, zejména pokud vztahy mezi jednotlivými daty nejsou jednoduché.
- Pomocí síťového grafu lze odhalit vzorce, trendy a informace, které jiné vizualizace nezobrazují.

- Síťové grafy jsou často používány při síťové analýze ke studiu vztahů mezi jednotlivci nebo organizacemi.
- Umožňují identifikovat klíčové prvky a vztahy v síti, což je užitečné zejména v oblastech sociálních věd, obchodu a marketingu (Geeksforgeeks, nedatováno).

Nevýhody

- Omezený počet informací, které mohou být reprezentovány v síťovém grafu. Uzly a hrany nemohou nést příliš mnoho informací, což může vyžadovat další informace k pochopení celého grafu.
- Při velké rozsáhlosti dat, může být složité síťový diagram sestavit a může to vyžadovat analytické nebo odborné znalosti z dané oblasti.
- Existuje mnoho různých typů grafů, každý s vlastními specifickými vlastnostmi. To může být problém při srovnávání a interpretaci různých grafů.
- Síťové grafy mohou odhalovat citlivé informace, jako například v případě analýzy sociálních sítí nebo marketingu (Geeksforgeeks, nedatováno).



Obrázek 12: Ukázková vizualizace Network diagram (Miller, 2019)

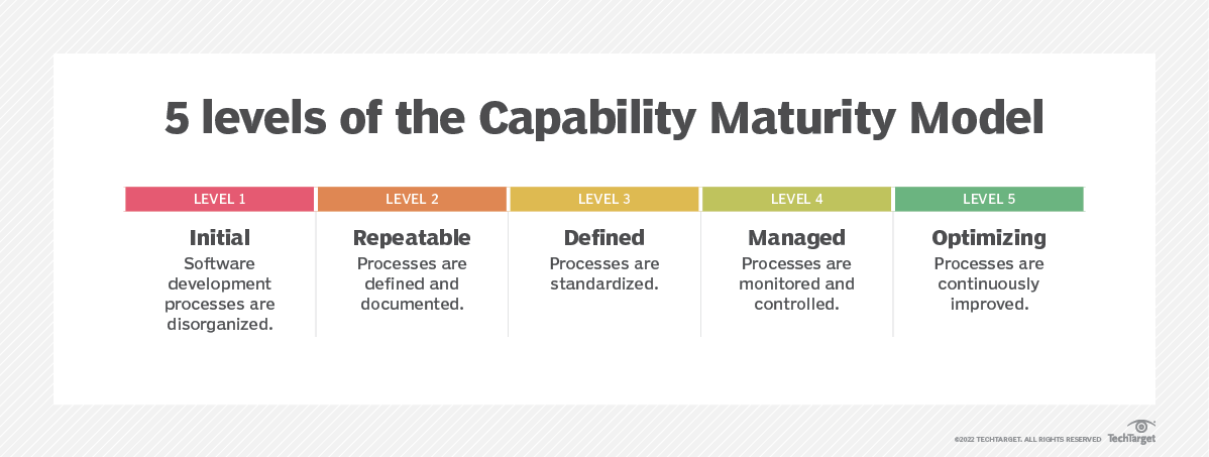
2 ŽIVOTNÍ CYKLUS, METODIKY A PROCESY

Předchozí poměrně velká část práce byla věnována samotnému úvodu do datové vědy a vhodnou prezentací výsledků v podobě vizualizace dat. Ovšem Data Science samozřejmě není jen o samotné vizualizaci, ale je to souhrn více úkonů, které jsou potřeba správně vykonat, aby výstupy analýzy byly srozumitelné a hlavně odpovídající. Tato kapitola bude věnována právě životnímu cyklu dat, metodikám a procesům používaných v Data Science. Na úvod bude zmíněn Data Science Maturity Model, který se používá pro zhodnocení zralostí organizací v oblasti Data Science. Poté se kapitola bude soustředit na Data Life Cycle Management, což je proces správy dat, který zahrnuje všechny fáze životního cyklu dat. A nakonec budou představeny jedny z hlavních metodik, které se často používají pro řízení procesů spojených s prací s daty a zefektivnění celého projektu.

2.1 Data Science Maturity Model

V praxi se setkáváme s různými typy modelů vyspělosti. Tyto modely slouží jako nástroj, který organizace využívají k posouzení výkonnosti a neustálému zlepšování svého byznysu nebo projektu. Maturity model se odlišuje od jiných nástrojů tím, že dokáže vyhodnocovat kvalitativní data a stanovit dlouhodobý směr a výkonost společnosti. Jeho cílem je zjistit, zda organizace dosahují zralosti, tedy jestli neustále testuje, roste a zlepšuje se. Modely definují různé úrovně efektivity a umožňují určit aktuální pozici jednotlivých osob, týmů, projektů nebo celé společnosti v rámci modelu (Eads, 2023).

Data Science Maturity Model vychází ze známého CMM – Capability Maturity Model (viz. Obrázek 13) (Arcalea, © 2023). CMM je metodika používaná k hodnocení zralosti procesů nebo organizací. Jedná se o nástroj, který pomáhá měřit různé aspekty procesů nebo organizací. Jeho hlavním cílem je podpořit organizace při snaze dosáhnout vyšší úrovně organizovanosti a systematického přístupu k podnikání (Proenca, 2016). Tento model se využívá především při hodnocení úrovně nebo vyspělosti jednotlivých procesů. Maturity model se obvykle skládá z pěti úrovní vyspělosti, ale některé zdroje uvádí i nultou fázi, která znamená, že řízení prakticky neexistuje a je označováno za chaotické (Carnegie, 2019).



Obrázek 13: Capability Maturity Model

Jak bylo již zmíněno, tak Data Science Maturity Model vychází z Capability Maturity Model (viz. Obrázek 13). DSMM je však pouze obecný rámec, který je určen odborníkům a vedoucím pracovníkům v oblasti datové vědy. Jeho hlavním účelem je pomoci identifikovat současný a cílový stav organizace, odhalit nedostatky a nasměrovat jejich budoucí investice v oblasti datové vědy. Je však důležité mít na paměti, že každá společnost má flexibilitu ve svém vlastním definování DSMM. To znamená, že model vyspělosti může být trochu jiný pro každou společnost, aby lépe odpovídal jejím potřebám (Steele, 2016). Příkladem může být společnost Domino Data Lab. Ta rozděluje model do čtyř úrovní vyspělosti a pěti dimenzí, které se poté vztahují na všechny různé organizace. To znamená, že tento model lze uplatnit ve všech odvětvích, jak pro pojišťovny, pro výrobce a další, kde datová věda může hrát důležitou roli (Steele, 2016).

Dalším příkladem je organizace Oracle, která vydala dokument s názvem „A Data Science Maturity Model for Enterprise Assessment“. Tento model má sloužit také jako nástroj pro hodnocení zralosti organizace, ale je komplexnější než většina. Skládá se z pěti úrovní vyspělosti, přičemž level 1 znamená, že organizace je nejméně vyspělá a level 5 je nejvíce vyspělá. Oproti modelu od Domino Data Lab poskytuje model od Oracle více dimenzí jako jsou Stragegy, Roles, Collaboration, Methodology, Data Awareness, Data Access, Scalability, Asset Management, Tools, Deployment (Hornick, 2020).

Následující tabulka popisuje Data Science Maturity Model (viz. Tabulka 1), tak jak ho shrnuje společnost Oracle. Na internetu se můžeme setkat s něčím podobným, co je nazýváno Data Science Maturity Test. Jedná se o otázky, které si může daná organizace položit, aby zjistila, v jaké situaci se aktuálně nachází. Je možné, že se bude pohybovat na rozmezí více levelů anebo je tabulka nedostačující a je potřeba zavést level 6. Tabulka obsahuje na každém řádku

jednotlivé dimenze, první sloupec obsahuje otázku, kterou by si daná společnost měla položit a další sloupce obsahují jednotlivé úrovně na kterých se mohou organizace nacházet (Hornick, 2020).

	OTÁZKY	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
Strategie	Jaká je strategie podniku pro datovou vědu?	Podnik nemá žádnou řídící strategii pro aplikaci datové vědy.	Podnik zkoumá hodnotu datové vědy jako klíčovou vlastnost.	Podnik uznává datovou vědu jako klíčovou vlastnost pro získání konkurenční výhody.	Podnik přijímá postup rozhodování založený na datech.	Data jsou vnímána jako důležitý podniková kapitál, například jako kapitál v podobě dat.
Role	Jaké role jsou v podniku definovány a vytvořeny pro podporu aktivit datové vědy?	Datoví analytici zkoumají a shrnují data pomocí deduktivních technik.	Vzniká role "datového vědce", který využívá ML a další pokročilé techniky.	Je zavedena role hlavního úředníka pro data (Chief Data Officer – CDO), která pomáhá spravovat data jako podnikové aktivum.	Kariérní cesta datového vědce je uzákoněna a standardizována v rámci celého podniku.	Je zavedena role hlavního úředníka pro datovou vědu (Chief Data Science Officer – CDSO)
Spolupráce	Jak spolupracují datoví vědci mezi sebou a s ostatními týmy datových vědců, aby se vyvíjely a předávaly si výsledky jejich práce v oblasti datové vědy?	Datoví analytici a/nebo datoví vědci pracují nezávisle na sobě a ukládají data a výsledky své práce do lokálních prostřední.	Existuje větší spolupráce mezi IT oddělením a organizačními jednotkami odpovědnými za provozní činnosti.	Je uznáváno, že je potřeba větší spolupráce při sdílení, modifikaci a předávání výsledků práce v oblasti datové vědy uvnitř týmů datových vědců.	Jsou zavedeny nástroje, které umožňují sdílení, modifikaci, sledování a předávání výsledků práce v oblasti datové vědy.	Jsou zavedeny standardizované nástroje napříč celým podnikem, které umožňují bezproblémovou spolupráci.
Metodologie	Jaký je způsob přístupu nebo metodologie podniku k datové vědě?	Analytika dat se zaměřuje na podnikovou inteligenci a vizualizaci dat pomocí ad hoc metodologie.	Analytika dat se rozšiřuje o využití strojového učení pro řešení podnikových problémů, avšak stále s použitím ad hoc metodologie.	Jednotlivé organizace začínají definovat a pravidelně používat metodologii datových věd.	Jsou definovány základní osvědčené postupy metodologie datových věd.	Osvědčené postupy metodologie datových věd jsou formálně zavedeny v rámci celého podniku.

Povědomí o datech	Jak je pro datové vědce snadné se seznámit s podnikovými datovými zdroji?	Uživatelé dat nejsou obeznámeni s širšími datovými prostředky, které jsou k dispozici v rámci podniku	Datoví analytici a datoví vědci hledají další zdroje dat prostřednictvím "klíčových osob".	Podnikové datové zdroje jsou katalogizovány a hodnoceny z hlediska kvality a užitečnosti pro řešení podnikových problémů.	Podnik zavádí nástroje pro správu metadat a katalogizaci dat.	Podnik se standardizuje na nástroj(e) pro katalogizaci dat/správu metadat a institucionalizuje jejich používání pro všechny datové prostředky.
Přístup k datům	Jak datoví analytici a datoví vědci žádají o přístup k datům a jakým způsobem je tento přístup kontrolován, spravován a monitorován?	Data se obvykle získávají prostřednictvím plochých souborů (flat file), které jsou získávány od IT oddělení nebo jiných zdrojů.	Přístup k datům je možný pomocí přímého programového přístupu k databázi.	Datoví vědci mají oprávněný programový přístup k velkému objemu dat, ale správci databází se potýkají s obtížemi při správě životního cyklu přístupu k datům.	Přístup k datům je pečlivěji kontrolován a spravován pomocí nástrojů pro správu identity.	Sledování původu přístupu k datům umožňuje jednoznačné odvození dat a identifikaci zdroje – úplná správa životního cyklu.
Škálovatelnost	Jsou nástroje dostatečně škálovatelné a výkonné pro průzkum dat, přípravu dat, modelování, hodnocení a nasazení? S růstem dat, datových projektů a týmu datové vědy, je podnik schopen tyto potřeby dostatečně podporovat?	Objemy dat jsou obvykle "malé" a omezené hardwarovými a softwarovými prostředky pro desktopové nasazení, přičemž analýzy provádějí jednotlivci pomocí jednoduchých pracovních postupů.	Projekty v oblasti datové vědy se stávají stále složitějšími a využívají větší objemy dat.	Jednotlivé skupiny přijímají různé škálovatelné nástroje pro datovou vědu a poskytují datovým vědcům větší hardwarové prostředky.	Podnik standardizuje integrovanou sadu škálovatelných automatizovaných nástrojů pro datovou vědu a vyčleňuje dostatečnou hardwarovou kapacitu pro projekty datové vědy.	Datoví vědci mají na vyžádání přístup k elastickým výpočetním zdrojům v budově i v cloudu s vysoce škálovatelnými algoritmy a infrastrukturou.

Správa aktiv	Jak jsou spravovány a kontrolovány prostředky datové vědy?	Analytické pracovní produkty jsou vlastněny, organizovány a udržovány jednotlivými členy týmu datové vědy.	Probíhají snahy o zajištění bezpečnosti, zálohování a obnovu pracovních produktů datové vědy.	Správa pracovních produktů datové vědy je systematicky řešena.	Správa pracovních produktů datové vědy je pevně zakotvena na podnikové úrovni s rostoucí podporou správy modelů.	Systematická správa všech pracovních produktů datové vědy se používá s plnou podporou správy modelů.
Nástroje	Jaké nástroje jsou v podniku používány pro cíle datové vědy? Mohou datoví vědci využívat open source nástroje v kombinaci s kvalitní produkční infrastrukturou?	Většinou se využívá nekoordinovaná řada neškálovatelných nástrojů pro izolovanou analýzu dat na desktopových zařízeních.	Data jsou spravována prostřednictvím správců databází (DBMS) a týmy se spoléhají na rozsáhlé open-source knihovny spolu se specializovanými komerčními nástroji.	Podnik hledá škálovatelné nástroje, které by podporovaly projekty datové vědy zahrnující velké objemy dat.	Podnik standardizuje sadu nástrojů, které splňují cíle projektů datové vědy.	Podnik pravidelně vyhodnocuje nejmodernější algoritmy, metodiky a nástroje pro zlepšení přesnosti, přehledu, výkonu a produktivity řešení.
Nasazení	Jak snadno lze výsledky datové vědy uvést do provozu, aby byly včas splněny obchodní cíle?	Výsledky datové vědy mají omezený dosah a tím pádem poskytují omezenou obchodní hodnotu.	Nasazení modelů do produkce je považováno za cenné, ale často vyžaduje opakovanou tvorbu infrastruktury pro každý projekt.	Podnik začíná využívat nástroje, které poskytují zjednodušené, automatizované nasazování modelů, včetně otevřeného softwaru a prostředí.	Heterogenní systémy v podniku vyžadují nasazování modelů mezi různými platformami a roste potřeba aplikací pro proudové zpracování dat.	Podnik si uvědomuje výhody okamžitého nasazení (případně přenastavení) modelů v rámci různorodých prostředí pomocí standardizované sady nástrojů s efektivním monitorováním obchodního dopadu.

Tabulka 1: Data Science Maturity model dle Oracle (Hornick, 2020)

2.2 Data Life Cycle

Data Life Cycle v oblasti Data Science se zaměřuje na celkový proces práce s daty v rámci datových vědeckých projektů. Tento cyklus se skládá z několika fází, které zahrnují sběr dat, jejich přípravu a čištění, analýzu a modelování, interpretaci výsledků a prezentaci. Tento cyklus je často znázorňován jako kruhový diagram, který ukazuje postupný tok dat a souvisejících aktivit s jejich zpracováním (Stobierski, 2021).

2.3 Data Life Cycle Management (DLM)

Data jsou pro podniky cenným aktivem, a proto je nezbytné je řádně spravovat. S narůstajícím objemem dat může být obtížné udržet přehled o všech datech, ať už jde o jejich původ, použití anebo co se s nimi má provést. V této chvíli přichází na řadu řízení životního cyklu dat (IBM, nedatováno).

Porozumění řízení životního cyklu dat je pro podniky klíčové, pokud chtějí data efektivně využívat. Řízení životního cyklu dat představuje proces správy dat od jejich vzniku až po odstranění. Cílem řízení životního cyklu dat je zajištění bezpečného a efektivního uchování dat, jejich dostupnosti v době, kdy jsou potřeba a správného odstranění, když nejsou potřeba. Data jsou rozdělena do fázi na základě různých kritérií. Během procesu mohou data přecházet mezi fázemi na základě splnění určitých úkolů nebo požadavků. To přispívá k dosažení klíčových cílů, jako je zajištění bezpečnosti a dostupnosti dat (IBM, nedatováno).

Řízení životního cyklu dat umožňuje podnikům připravit se na možné katastrofické scénáře, jako je ztráta dat nebo selhání systému. Kvalitní strategie DLM klade důraz na ochranu dat a obnovu po havárii, zejména v době rychlého růstu počtu hrozeb. Díky tomu je již předem vypracován efektivní plán obnovy dat v případě havárie, což minimalizuje některé negativní dopady na zisk a pověst značky (IBM, nedatováno).

Obrázek 14: Data Life Cycle Management (Data Lifecycle Management (DLM) : A New Way of Managing Data, 2021)

Životní cyklus dat se skládá většinou z pěti fází, lze se setkat i s více fázemi, kdy některá z pěti fází je rozdělena na více. Například se může jednat o fázi „Usage“, která bývá rozdělena na dvě fáze „Usage“ a „Share“. Všechny fáze se řídí souborem zásad, které maximalizují hodnotu dat v každé fázi životního cyklu. Přestože existuje mnoho různých výkladů řízení životního cyklu dat, tak jej můžeme shrnout následovně (IBM, nedatováno).

Fáze 1: Data Creation

Počáteční fází životního cyklu je sběr dat. Organizace obvykle vytváří data jedním ze tří způsobů. Prvním způsobem je získávání dat z již existujících zdrojů, která byla vytvořena mimo organizace. Druhým způsobem je ruční zadávání nových dat zaměstnanci v rámci organizace. Posledním způsobem je zachycení dat, která jsou generována zařízeními používanými v různých procesech v rámci organizace (DataWorks, nedatováno). Mezi zdroje dat mohou patřit webové a mobilní aplikace, zařízení internetu věcí (IoT), formuláře, průzkumy, data z databází a další zdroje (IBM, nedatováno).

Fáze 2: Data Storage

Jakmile jsou data v organizaci vytvořena, je důležité je uložit a chránit s použitím odpovídající úrovně zabezpečení. Data se mohou lišit svou strukturou, což může ovlivnit volbu datového úložiště. Strukturovaná data se nejčastěji ukládají v relačních databázích, zatímco nestrukturovaná data obvykle v NoSQL databázích. Po vybrání vhodného typu úložiště je důležité posoudit infrastrukturu z hlediska bezpečnosti. Organizace musí zajistit ochranu před

nebezpečnými subjekty a zároveň dodržovat všechny zásady, jako je například GDPR. Další vlastností ochrany dat může být redundance. Vytvoření kopií uložených dat může sloužit jako záloha, aby byla zajištěna ochrana před jejich neúmyslným smazáním, poškozením nebo jiným nežádoucím incidentem (IBM, nedatováno).

Fáze 3: Data Usage

V této fázi jsou data použita k podpoře činnosti organizace. Životní cyklus definuje, kdo má oprávnění data používat a k jakému účelu. Data lze prohlížet, zpracovávat, upravovat a ukládat. Pro veškerá kritická data by měla být udržována auditní stopa, aby bylo zajištěno, že všechny změny jsou dohledatelné (DataWorks, nedatováno). Použití dat nemusí být omezeno pouze pro interní potřeby organizace, ale mohou být využita i pro externí účely. Například se může jednat o marketingovou analýzu, reklamu, prezentaci a další jiné využití (IBM, nedatováno).

Fáze 4: Data Archival

Ve čtvrté fázi, nazvané „Archivace dat“, se data přesunou do archivu, když přestanou být pro každodenní provoz užitečná. Je důležité jasně definovat, kdy, kde a jak budou archivována. Data by měla být uložena na místě, kde nedochází k žádné údržbě ani běžnému používání. Pokud by v budoucnu byla potřeba data znovu použít, například v případě soudních sporů nebo jiných potřeb, budou data přenesena do produkčního prostřední nebo jiného vhodného místa, kde mohou být použita (DataWorks, nedatováno).

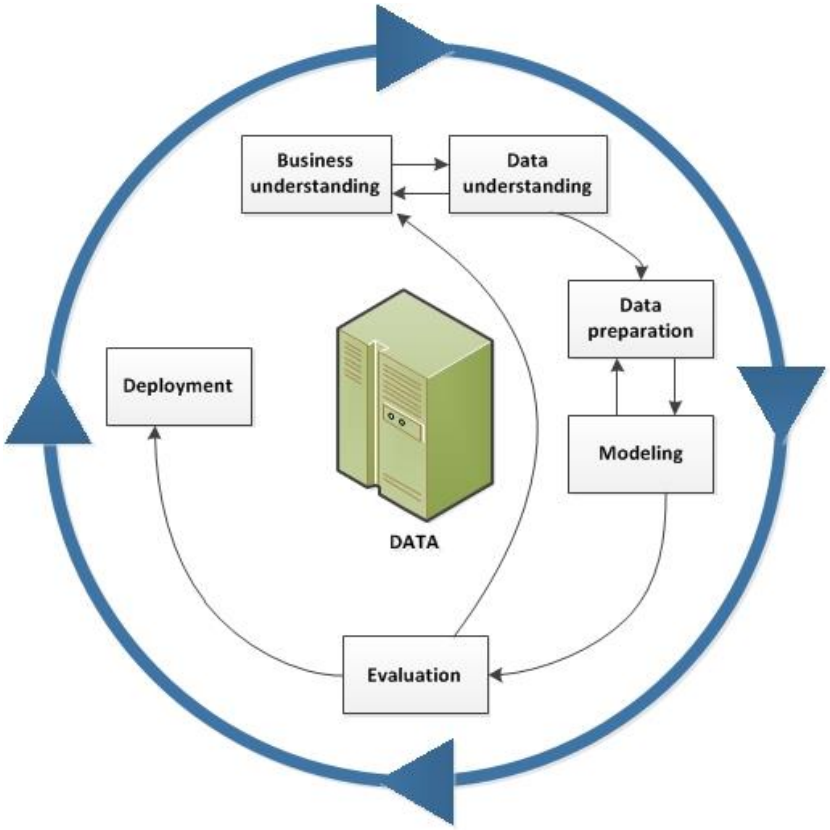
Fáze 5: Data Destruction

Poslední fází je zničení dat. Jakmile data přesáhnou zákonem požadovanou dobu pro jejich uchování, není nutné je dále uchovávat. Dalším důvodem může být potřeba vytvořit volné místo pro ukládání nových dat, která jsou pro organizaci hodnotnější a aktuálnější (Creative, © 2023), (IBM, nedatováno).

2.4 Cross-Industry Standard Process for Data Mining (CRISP-DM)

Rozdíl mezi předchozí metodikou DLM a CRISP-DM spočívá v jejich zaměření. DLM se zabývá celkovým řízením a správou dat po celou dobu jejich životního cyklu. Na druhou stranu se CRISP-DM soustředí na proces těžby dat a dolování znalostí v rámci projektů analýzy dat. První verze CRISP-DM 1.0 byla koncipována již v roce 1996 (Chapman, 2000). Dle průzkumů se jedná již mnoho let o jednu z nejpoužívanějších metodik. Její popularita spočívá v tom, že byla navržena tak, aby byla nezávislá na konkrétním softwaru, dodavateli nebo technologiích použitých pro analýzu dat.

Model životního cyklu (viz. Obrázek 15) se skládá ze šesti fází, mezi které řadíme: business understanding, data understanding, data preparation, modeling, evaluation a deployment. Tyto fáze se poté dělí na další etapy, které budou popsány v další části. Všechny hlavní fáze se točí kolem samotných dat, která se nacházejí uprostřed modelu. Běžně se proces pohybuje ve směru šipek, ale jelikož je proces flexibilní, tak datový vědec nemusí procházet proces vždy lineárně. Například, pokud výsledek aktuální fáze není očekávaný, může se vrátit zpět nebo opakovat předchozí fáze (Chapman, 2000).



Obrázek 15: Diagram jednotlivých fází CRISP-DM (CRISP-DM Help Overview, 2021)

Business understanding, z počátku je důležité si uvědomit v jakém oboru společnost podniká, v jakém odvětví, jak funguje a v podstatě vše, co s podnikáním souvisí. Pokud datový vědec nedisponuje těmito informacemi, může nepřesně specifikovat aktuální cíle podniku. To by mohlo vést k nevhodně zvolené strategii nebo řešení některých situací. Někteří datoví vědci občas přehlížejí informace o své firmě a snaží se z dat vytěžit co nejvíce informací, což může vést k závěrům, které nejsou klíčové nebo relevantní pro podnikové cíle (Chapman, 2000).

Business understanding se skládá ze čtyř dílčích úloh:

- **Determine business objectives** (určení obchodního cíle), z obchodního hlediska se jedná o pochopení, čeho klient nebo společnost chtějí dosáhnout. Cílem analytika je ihned odhalit důležité faktory a stanovit kritéria úspěchu (Chapman, 2000).
- **Assess situation** (vyhodnocení situace), tato část zahrnuje sepsání všech zdrojů, které má organizace k dispozici. Dalším úkolem je vytvořit výstup všech požadavků projektu včetně harmonogramu dokončení, srozumitelnosti a kvality výsledků, zajištění bezpečnosti. Měli by se stanovit předpoklady, které by se daly ověřit během těžby dat, ale také nekontrolovatelné předpoklady o podnikání. Například se může jednat o různé omezení týkající se dat. V neposlední řadě by tato část měla obsahovat seznam rizik, slovníček pojmů, analýzu nákladů a přínos projektu (Chapman, 2000)
- **Determine data science goals** (určení cílů datové vědy) vyjadřuje cíle v obchodní terminologii. Může se jednat například o zvýšení prodejů produktů nebo předpovědi, kolik produktů se prodá. Cílem této etapy by mělo být popsat výstupy, které budou umožňovat dosáhnout obchodních cílů a také by se měla definovat kritéria podle kterých se pozná úspěch projektu (Chapman, 2000).
- **Produce project plan** (vypracování plánu projektu), v tomto bodě se vybuduje plán celého projektu, ať už jde o dobu trvání, vstupy, výstupy, závislosti a další body. Na konci první fáze se provede posouzení použitých nástrojů a technik a vyberou se nástroje pro dolování dat (Chapman, 2000).

Data understanding, fáze porozumění začíná sběrem dat a následně je třeba se s daty seznámit, identifikovat jejich problémy, zjistit jejich kvalitu a jaká jsou jejich rizika. Na základě těchto prvotních informací je třeba odhalit zajímavé podskupiny, které budou sloužit pro vytvoření hypotéz pro odhalení informací (Chapman, 2000).

- **Collect initial data** (shromažďování počátečních dat), výstupem této úlohy by měla být zpráva, která bude obsahovat výpis datových souborů, metody použité k jejich získání a případné problémy, které se vyskytly. Zaznamenání problémů a jejich řešení usnadní práci, pokud by v budoucnu došlo u podobných projektů ke stejným nebo alespoň částečně stejným problémům (Chapman, 2000).

- **Describe data** (popis dat), úkolem je vytvořit zprávu, která bude obsahovat popis dat. Zpráva bude obsahovat informace jako je formát dat, množství dat a další možné informace, které je možné z prvního pohledu na data získat (Chapman, 2000).
- **Explore data** (zkoumání dat), tato etapa se soustředí na otázky ohledně dolování dat, které lze řešit pomocí dotazování, vizualizací a reportů. Analyzuje se rozložení klíčových atributů, vztahy mezi atributy, výsledky agregací, vlastnosti podmnožin a provádějí se statistické analýzy. Výsledky jsou zaznamenány a je vypracována zpráva z průzkumu.
- **Verify data quality** (ověřování kvality dat), zahrnuje ověření kompletnosti dat, jejich správnosti a výskyt chyb. Je také důležití identifikovat chybějící hodnoty a jejich četnost. Ve výsledné zprávě je vhodné uvést výsledky této kontroly a pokud jsou zjištěny problémy s kvalitou, tak navrhnout možná řešení (Chapman, 2000).

Data preparation, třetí fáze, která se zabývá finální přípravou dat z dat původních. Fáze je rozdělena do pěti samostatných úloh. Ty se mohou opakovaně provádět a neplatí pro ně žádné předepsané pořadí v jakém by se měly provádět (Chapman, 2000).

- **Select data** (výběr dat), mělo by padnout rozhodnutí o tom, která data se použijí pro analýzu na základě zvolených kritérií. Data, která se nepoužijí by měla být zaznamenána společně s důvody vyloučení (Chapman, 2000).
- **Clean data** (čištění dat), jedná se o zvýšení kvality dat na požadovanou úroveň. Může se jednat o úpravu hodnot, vložení výchozích hodnot pro chybějící hodnoty nebo se mohou použít metody, které chybějící hodnoty odhadnou pomocí modelování. Tyto kroky musí být zdokumentovány, aby bylo možné zpětně kontrolovat možné dopady na výsledky (Chapman, 2000).
- **Construct data** (vytvoření dat), v této části se dochází k vytváření nových dat nebo k transformaci již existujících. Může se jednat například o vytvoření nového atributu plocha = délka × šířka. Tyto operace by měly být zaznamenány (Chapman, 2000).
- **Integrate data** (integrace dat), v tomto kroku se kombinují informace z více zdrojů, většinou z více tabulek za účelem vytvoření nových záznamů. Může se jednat o jednoduché sloučení dat nebo agregaci dat (Chapman, 2000).
- **Format data** (formátování dat), tento úkon se týká hlavně syntaktických změn, které nezmění význam dat, ale je nutné je provést kvůli používanému nástroji pro modelování. Například některé nástroje pracují s desetinou čárkou a některé s desetinou tečkou. Také se může jednat o ořezání hodnot na maximální délku nebo odstranění nepotřebných znaků. Některé nástroje mohou požadovat specifické řazení dat, kdy první záznam mohou

vyžadovat unikátní identifikátor. Pro neuronové sítě je vhodné použít náhodné řazení sloupců (Chapman, 2000).

Modeling, obvykle je nedostačující vytvořit jeden model s jedním scénářem, který by přinesl uspokojivé výsledky. Proto je většinou třeba vytvořit více modelů, které budou spuštěny s různými výchozími parametry. Často se může stát, že bude potřeba se vrátit do předchozích fází (Chapman, 2000).

- **Select modeling technique** (výběr modelovací techniky), v této fázi nedochází k výběru nástrojů ty byly vybrány předchozích fázích. Prvním krokem při modelování je zvolit modelovací techniku. Může se jednat například o rozhodovací stromy, klasifikační metody, neuronové sítě nebo jiné techniky. Až je technika vybrána je proveden její popis, který je zaznamenán. Pokud se používá více technik, je tato část provedena pro každou techniku samostatně (Chapman, 2000).
- **Generate test design** (generování návrhu testu), než bude vytvořen model je potřeba vytvořit postup nebo mechanismus, který bude sloužit pro kontrolu jeho kvality a platnosti. U některých typů dolování dat se můžeme setkat s klasifikací, pro kterou je typické používat jako měřítko kvality chybovost. Pro tyto metody je vytvořena trénovací sada, na které je hodnocena kvalita. Výsledkem by měl být plán, který bude obsahovat, jak rozdělit dostupnou sadu dat na trénovací data, testovací data a validační data (Chapman, 2000).
- **Build model** (vytvoření modelu), tato část má za úkol s pomocí vybraného modelovacího nástroje vytvořit jeden nebo více modelů. Často lze zvolit různé nastavení parametrů pro tvorbu modelu. Výstupem jsou poté konkrétní modely s podrobným popisem (Chapman, 2000).
- **Assess model** (posouzení modelu), poslední krok před finálním vytvořením cílového modelu. Do této doby si vystačil odborník na dolování dat sám se svými odbornými znalostmi, stanovenými kritérii a s návrhem testu. V tuto chvíli jsou potřeba znalosti obchodních analytiků a odborníků v daném oboru, aby diskutovali o výsledcích z obchodního hlediska. Odborník pro dolování dat se poté snaží ohodnotit modely nezávisle na ostatních výsledcích. Na základě hodnocení může dojít k úpravě parametrů a celý proces se může opakovat. Pokud má odborník více modelů, lze je mezi sebou porovnat (Chapman, 2000).

Evaluation, v předposlední fázi by už měly být vytvořeny dostatečně kvalitní modely pro analýzu dat. Před nasazením modelu je nutné ho vyhodnotit a zkontrolovat, zda odpovídá podnikovým cílům. Důležitým krokem je zajistit, aby nedošlo k opomenutí klíčových

problému, které jsou potřeba řešit. Na konci fáze by se mělo vyhodnotit, jestli organizace bude moci dostatečně využít výsledky z modelu.

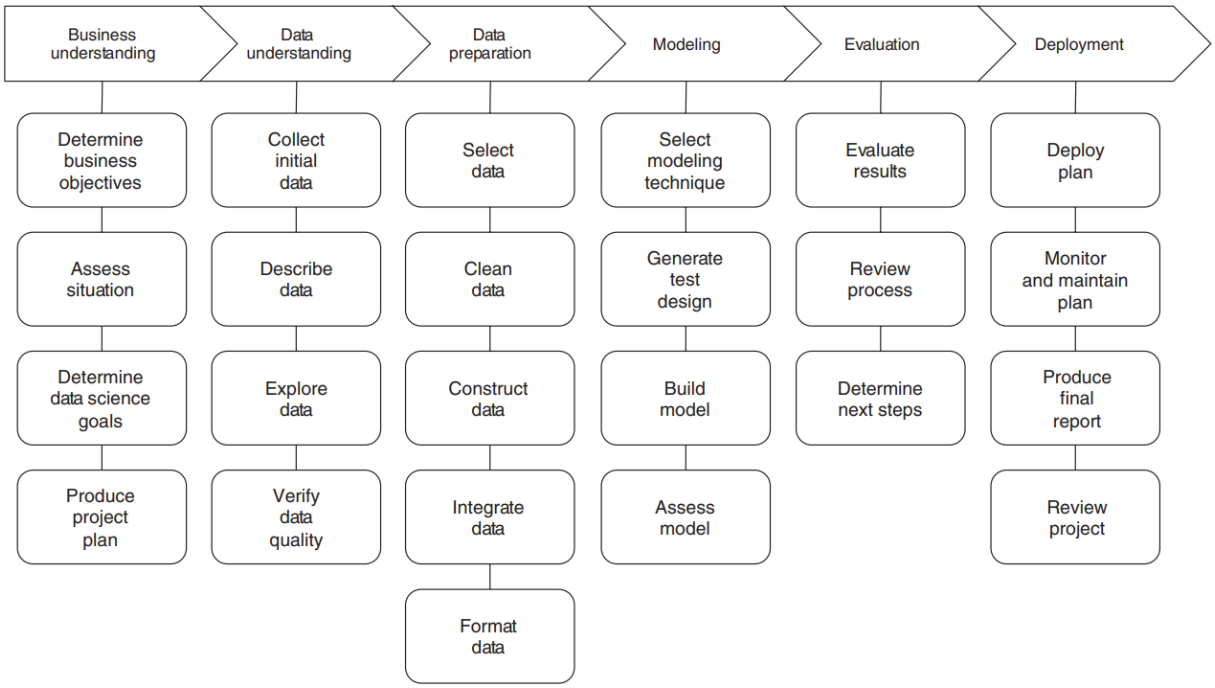
- **Evaluate results** (hodnocení výsledků), v předešlých fázích se věnovalo modelu spíše z hlediska návrhu. V této části je zkoumáno, zda model splňuje podnikové cíle a jestli existuje nějaký důvod proč by mohl být model nedostačující. Pokud to čas a rozpočet dovoluje, lze nasadit model na testovací prostředí. Hodnocení výsledků může odhalit cíle, které nebyly plánované, ale mohou přinést nové cíle do budoucna. Jakmile jsou modely v souladu se všemi kritérii stávají se schválenými modely a lze pokračovat (Chapman, 2000).
- **Review process** (posouzení procesu), tato část slouží jako jakási kontrola. Model by měl být připraven k použití a nemělo by být potřeba jej měnit. Hlavní činností v téhle fázi je přezkoumáni modelu a jeho atributů. Přezkoumání by mělo být zdokumentováno a pokud došlo k opomenutí některých činností, měly by být zaznamenány (Chapman, 2000).
- **Determine next steps** (určení dalších kroků), jakmile je model vyhodnocen dochází k rozhodování, co se má stát v dalších krocích. Může dojít k nasazení projektu, zahájení dalších iterací nebo začít nové projekty. Měly by se brát v úvahu také finance, které mohou zásadně ovlivnit toto rozhodnutí. Výstupem by měl být seznam možných akcí s důvody pro a proti společně s rozhodnutím a odůvodněním (Chapman, 2000).

Deployment, poslední fází je nasazení. Získané znalosti z datového modelování se uplatňují v organizaci. Cílem je získat data, která jsou potřeba dobře a správě předávat. Může se jednat o triviální prezentace, reporty nebo může být potřeba aplikovat živé modely, které budou reagovat na aktuální situaci, například na webové stránce. Většinou se model nasazuje na stranu klienta, proto je potřeba, aby zákazník dostatečně porozuměl, co musí udělat, aby mohl vytvořené modely využívat pro své podnikání (Chapman, 2000).

- **Deploy plan** (plán nasazení), v této fázi se na základě výsledků hodnocení vypracovává strategie pro nasazení. Pokud byl identifikován obecný postup pro vytvoření příslušných modelů, je zde zdokumentován pro pozdější nasazení. Plán nasazení shrnuje nezbytné kroky a jejich provedení (Chapman, 2000).
- **Monitor and maintain plan** (plán monitorování a údržby), tento proces je důležitý, pokud se výsledky dolování dat stávají součástí každodenního provozu podniku. Cílem je předejít dlouhodobému nesprávnému využívání těchto výsledků. Plán monitorování zahrnuje detailní strategii pro sledování nasazených modelů, která je přizpůsobena konkrétnímu

typu nasazení. Je důležité plánovat pravidelnou údržbu a opravy v případě potřeby (Chapman, 2000).

- **Produce final report** (vypracování závěrečné zprávy), na konci projektu je vypsána závěrečná zpráva. To je důležitou součástí dolování dat, kde projektový tým shrnuje všechny předchozí výstupy a výsledky. Zpráva poté slouží k prezentaci projektu a jeho zkušeností. Výstupem může být také závěrečná prezentace, který umožňuje prezentaci výsledků zákazníkovi (Chapman, 2000).
- **Review project** (přehled projektu), zahrnuje posouzení úspěchu, neúspěchu a identifikaci oblastí, které je potřeba zlepšit. Obsahuje dokumentaci zkušeností z projektu s důležitými poznatky, jako jsou rizika, zavádějící přístupy nebo rady pro výběr nejvhodnějších technik dolování dat v podobných situací. V nejlepším případě zahrnuje dokumentace také zprávy a zápisy jednotlivých členů týmu o jejich činnosti (Chapman, 2000).



Obrázek 16: Jednotlivé úlohy modelu CRISP-DM (Chapman, 2000).

3 NÁSTROJE URČENÉ K PRÁCI S DATA SCIENCE

Následující kapitola je zaměřena na představení klíčových technologií, nástrojů a knihoven, které jsou běžně využívány datovými vědci. Tyto technologie a nástroje jsou rozděleny do jednotlivých fází, ve kterých se lze nejčastěji setkat s jejich použitím. Je však důležité si uvědomit, že tyto technologie mohou být využívány i v jiných fázích nebo v různém pořadí, neboť každý datový vědec může preferovat odlišné nástroje a přístupy.

Rozdělení technologií do fází v této kapitole není nutně totožné s reálným cyklem, kterým se datoví vědci musí řídit. Přesto představené nástroje patří mezi jedny z nejrozšířenějších a nejčastěji používaných v oblasti Data Science.

Cílem této kapitoly je tak poskytnout čtenářům ucelený přehled a povědomí o klíčových technologiích, nástrojích a knihovnách, které mohou využít při práci s daty. Tím je umožněno jim lépe porozumět a vybrat si ty nejvhodnější nástroje pro své konkrétní potřeby v rámci Data Science projektů.

3.1 Jupyter

Jupyter je projekt, který zahrnuje několik softwarových produktů, z nichž nejznámější jsou Jupyter Notebook a JupyterLab. Projekt Jupyter se zaměřuje na vytváření nástrojů, které umožňují uživatelům provádět interaktivní výpočty pomocí zápisníků. Tyto zápisníky slouží jako sdílené dokumenty, které mohou obsahovat počítačový kód, různé popisky, data, 3D modely, grafy, obrázky a ovládací prvky. Zápisník nabízí rychlé interaktivní prostředí pro vytváření prototypů a vysvětlování kódu, zkoumání dat, vizualizaci dat a sdílení nápadů s ostatními (Jupyter, 2015).

Jupyter Notebook je dostupný online a podporuje jazyky jako Python, Julia, R, C++ nebo Ruby. Existuje také mnoho verzí vytvořených komunitou uživatelů, které je možné použít pro své potřeby (Jupyter, © 2023). Hlavní výhody Jupyter Norebooku zahrnují možnost editace kódu přímo v prohlížeči, zvýrazňování syntaxe, automatické doplňování kódu, spouštění kódu přímo v prohlížeči a zobrazení výsledků. Zápisníky jsou uloženy ve formátu ipynb a mohou být převedeny do formátu HTML, LaTeX, PNG, SVG. V zápisnících může být použit běžný text nebo značkovací jazyk Markdown k popisu jednotlivých částí (Jupyter, 2015). Zápisníky obsahují samostatné spustitelné buňky, které mohou být opakovaně spouštěny v libovolném pořadí.

Rozdíl mezi Jupyter Notebook a JupyterLab spočívá v tom, že v JupyterLabu jsou již předem nainstalovány podpory pro různé programovací jazyky včetně Pythonu, Julia, Scala a R. Jupyter Notebook poskytuje jednodušší rozhraní oproti propracovanějšímu JupyterLabu, který umožňuje otevírat nejenom poznámkové bloky, terminály a textové soubory, ale také konzole, editory CSV, editory Markdown, interaktivní mapy a další (Domino, © 2023). JupyterLab umožňuje také spolupracovat více uživatelům na stejném zápisníku, což může usnadnit práci celému týmu. Jedná se o modernější, všestrannější verzi, která uživatelům nabízí podporu pro větší škálu formátů (SaturnCloud, 2023).

3.2 Apache Zeppelin

Jedná se o nástroj, který slouží jako webový notebook, který umožňuje interaktivní analýzu dat. Podporuje velké množství programovacích jazyků a frameworku pro zpracování dat. V současné době podporuje například Apache Spark, Apache Flink, Python, R, JDBC, Markdown a Shell. Pokud nemá potřebný interpreter podporu, tak existuje lehký návod na to, jak přidat nový. Apache Zeppelin poskytuje integrovanou podporu pro Apache Spark, takže není potřeba vytvářet samostatné moduly, pluginy nebo knihovny (Apache Zeppelin, nedatováno).

Apache Zeppelin s integrovaným Sparkem poskytuje automatické vkládání SparkContextu, SQLContextu a SparkSession. Načítání závislostí z jar souborů za běhu z lokálního souborového systému nebo repozitáře Maven. Lze zrušit úlohu a zobrazit její průběh. Podporuje interpreter Livy, což je REST rozhraní pro interakci se Spark odkudkoli. Podporuje spouštění úryvků kódu nebo programů v SparkContext, který běží lokálně nebo v YARN. Lze vizualizovat Spark Dataset nebo DataFrame pomocí knihoven pro vykreslování v Pythonu nebo R. Je také možné, aby v jedné instanci Zeppelin pracovalo více uživatelů bez vzájemného ovlivňování. Dále je možné využívat rozhraní Zeppelin notebooku jako REST API (Apache Zeppelin, nedatováno).

3.3 Python

Python je oblíbeným a všestranným programovacím jazykem s elegantní syntaxí a velkým množstvím knihoven. Často je srovnáván s jazyky jako Perl, Ruby nebo Java. Python je objektově orientovaný a může být spuštěn na různých operačních systémech, jako jsou Mac OS X, Windows, Linux a Unix (Python, 2022). Pro datovou vědu je Python ideální, protože zvládá složité matematické a vědecké výpočty. Jeho efektivní správa paměti umožňuje pracovat

s velkými objemy dat. Python také poskytuje balíčky pro integraci s jinými jazyky, jako je Java nebo C (Data Science Python - Getting Started, © 2023).

Je třeba zdůraznit, že Python je open-source platformou s velkou komunitou vývojářů, kteří pravidelně přispívají různými nástroji a knihovnami pro datovou vědu. Mezi nejznámější knihovny patří NumPy, pandas, Matplotlib, SciPy, TensorFlow a Keras, které budou dále zmíněny v práci (Chandan, 2023), (Luna, 2023).

3.3.1 NumPy

Numerical Python neboli NumPy je jednou z klíčových knihoven pro vědecké výpočty, která má uplatnění téměř ve všech oblastech datové vědy a techniky. Používá se především k ukládání a manipulaci s daty. Uživatelům poskytuje vícerozměrná pole, maticové struktury, ndarray, což je homogenní objekt n-rozměrného pole společně s metodami pro práci s ním. Dále je možné provádět různé matematické operace, Fourierovy transformace, generování náhodných čísel nebo provádět čtení a zápis na disk (NumPy, © 2022). Obsahuje také velkou sbírku matematických funkcí, které jsou schopné pracovat s maticemi a poli (McKinney, 2022).

3.3.2 Pandas

Pandas je moderní balíček postavený na knihovně NumPy, který nabízí efektivní implementaci datové struktury nazývané DataFrame. DataFrame obsahuje integrované indexování a lze si datovou strukturu představit jako vícedimenzionální pole s popisky pro řádky a sloupce. Může obsahovat různé typy dat nebo dokonce nemusí obsahovat žádná data (Kniha: Python Data Science Handbook, Jake VanderPlas). Pandas poskytuje nástroje pro snadné čtení a zápis datových struktur, které mohou být uloženy v paměti, textových nebo CSV souborech, SQL databázích nebo ve formátu HDF5. Tato knihovna umožňuje datovým vědcům provádět širokou škálu operací nad daty. Mezi nejčastěji používané operace patří agregace a transformace dat, rozdělování dat na základě kategorií, slučování a spojování datových sad, stejně jako vkládání nebo odstraňování sloupců z datových struktur. Pandas obsahuje mnoho dalších funkcionalit, které usnadňují analýzu a manipulaci s daty (Pandas, © 2023).

3.3.3 Plotly

Plotly je grafická knihovna, která slouží pro vizualizace. Jedná se o open-source knihovnu, ale je možné získat také komerční verze Dash Enterprise a Chart Studio Enterprise. Knihovna podporuje více než 40 jedinečných typů grafů, které mohou složit pro vizualizaci statistických, finančních, vědeckých nebo trojrozměrných dat. Plotly bylo postaveno na Javascriptové knihovně plotly.js. Uživatelé mohou vytvářet interaktivní vizualizace, které je možné zobrazit

v zápisnících Jupyter, ukládat do HTML souborů nebo je vložit jako součást webových aplikací vytvořených v čistém jazyce Python pomocí nástroje Dash. Výstupy se také mohou vkládat do reportů ve formě vysoce kvalitních vektorových obrázků (Plotly, © 2023).

3.3.4 Matplotlib

Knihovna Matplotlib slouží pro tvorbu grafů a dvourozměrných vizualizací. Jedná se o jednu z nejoblíbenějších knihoven pro vizualizace, a to hlavně díky dobré integraci s dalšími systémy a grafickými nástroji (McKinney, 2022). V dnešní době může působit zastarale oproti nástrojům jako ggplot a ggvis v jazyce R, které využívají nástroje pro vizualizace založené na JavaScriptové knihovně D3js a HTML5 canvasu. Lidé stále vyvíjí různé balíčky, které se snaží přidat do grafů moderní prvky. Novější knihovny, které jsou založeny na knihovně Matplotlib jsou například Seaborn, ggplot, HoloViews a Altair (Vanderplas, 2017).

3.3.5 Scipy

Scipy je knihovna, která je založena na knihovně NumPy. Poskytuje různé balíčky, které jsou organizovány do kategorií a pokrývají širokou škálu oblastí pro vědecké výpočty. Obsahuje balíčky s funkcemi pro shlukování, fyzikální a matematické konstanty, rychlou Fourierovu transformaci, řešení rovnic, práci se vstupy a výstupy a mnoho dalších (SciPy, 2023). Společně s knihovnou NumPy tvoří Scipy dostatečně komplexní nástroj pro provádění různých vědeckých operací (McKinney, 2022).

3.3.6 Scikit-learn

Scikit-learn nebo v některých případech Sklearn je jedna z hlavních a univerzálních knihoven pro strojové učení. Byla postavena nad knihovnami NumPy, SciPy a Matplotlib. Nabízí velké množství dobře zpracovaných algoritmů a vyznačuje se čistým, jednotným a zjednodušeným API a také velmi užitečnou online dokumentací. Mezi nejoblíbenější podmoduly patří algoritmy učení s učitelem, kde lze najít skoro všechny populární algoritmy jako například lineární regrese, SVM nebo rozhodovací strom. Patří tam také algoritmy učení bez učitele, jako například shlukování, faktorová analýza nebo neuronové sítě. V neposlední řadě sem patří shlukování, křížové ověřování, redukce dimenzionality, ansámblové metody, extrakce rysů a výběr rysů (Tutorialspoint, nedatováno).

3.3.7 Statsmodels

Jedná se o balík především pro statistickou analýzu. Obsahuje hlavně algoritmy pro klasickou statistiku a ekonometrii. Obsahuje například regresní modely, kam patří například lineární regrese, zobecněné lineární modely nebo robustní lineární modely. Dále moduly pro analýzu

rozptylu, analýzu časových čas, neparametrické metody, jako například odhad hustoty jádra nebo jádrová regrese. Výsledky je možné zobrazit pomocí modulu pro vizualizace. Statsmodels poskytuje pouze odhady, protože pracuje se statistikou. Scikit-lear je oproti tomuto balíčku zaměřen více na predikci (Vanderplas, 2017).

3.3.8 Keras a TensorFlow 2

V této části se podíváme na dvě knihovny. Tou první je Keras, která je od verze TenserFlow 2 její součástí. Keras je knihovna zaměřená na neuronové sítě na vysoké úrovni. Jedná se o knihovnu, která nabízí jednoduchost, flexibilitu a vysoký výkon. Provádění úloh můžeme spustit na CPU, TPU nebo na velkých clusterech GPU. Podporuje konvoluční, rekurentní sítě a také jejich kombinace (Keras, nedatováno). Keras pokryje všechny dílčí kroky pracovního postupu strojového učení, od zpracování dat přes ladění hyperparametrů až po nasazení. Je možné přidávat a odebírat neuronové vrstvy dle potřeby, díky tomu je možné vytvářet jednoduché, tak složití architektury. Je také možné definovat modely s více vstupy, orientované acyklické grafy nebo modely se sdílenými vrstvami (Salazar, 2022). Výsledné modely lze obsluhovat skrze webové rozhraní nebo je model extrahovat pro spuštění v prohlížeči nebo mobilních zařízeních (TensorFlowe, 2023).

TensorFlow 2 je knihovna pro strojové učení, která se snaží zaměřit více na jednoduchost a snadné používání oproti předchozí verzi. Jak již bylo zmíněno, od novější verze je Keras součástí TenserFlow. Umožňuje programátorům zaměřit se na logiku aplikace místo na detaily potřebné k průchodu neuronové sítě. TensorFlow umožňuje používat nejenom vysokoúrovňové operace, ale také nízkoúrovňové operace. Další výhodou je použití TPU od společnosti Google, které jsou speciálně přizpůsobeny k práci s TensorFlow pro strojové učení. Za zmínku stojí, že výpočty na GPU jsou možné pouze na grafických kartách značky NVIDIA (Salazar, 2022).

3.3.9 Scrapy

Scrapy je framework, který se zaměřuje na sběr dat. Podporuje dvě techniky „web scraping“ a „web crawling“, které se používají k procházení webů a získávání dat z nich. Scrapy se dá použít také k monitorování webů nebo testování webových aplikací. Data získaná z internetu lze uložit jako strukturovaná data a zároveň je před ukládáním očistit. Scrapy využívá asynchronní mechanismus, díky čemuž je možné paralelně zpracovávat více úloh zároveň. Často se lze setkat i s termínem „spider“, jedná se o třídu, která definuje set instrukcí pro sběr dát. Dále framework obsahuje selektory, pomocí kterých lze vybrat konkrétní HTML tagy nebo regulární výrazy. Získaná data se vracejí jako položky, které je možné následně ověřit, očistit

a uložit. Lze extrahovat data jak z CSV, XML nebo JSON souborů. Nevýhodou by mohla být práce s Javasriptem, který je na webech velmi rozsáhlý (Azram, 2021). Vhodnou alternativou může být například framework BeatifulSoup, který je vhodnější pro menší projekty nebo například Selenium (PyCoach, 2020).

3.4 R

Hlavním konkurentem Pythonu je jazyk R, který sice v posledních letech ztrácí na popularitě, ale stále zůstává jednou z nejlepších voleb pro datové vědce. R je specificky navržen pro práci s daty a nachází široké uplatnění v oblastech jako finance, statistika nebo akademická sféra. Jde o volně dostupný software, který je kompatibilní s operačními systémy UNIX, Windows, MacOS, FreeBSD a Linux. Jazyk R nabízí rozsáhlou kolekci softwarových nástrojů pro práci s daty, výpočty a grafické zobrazení. Obsahuje více než 6400 balíčků, které lze využít pro různé úkoly. Jednou z výhod je snadné vytváření kvalitních grafů, včetně matematických symbolů a vzorců, vhodných pro publikace (Pimpler, © 2017). Pro složitější výpočty je také možné propojit R s jazyky jako C, C++ nebo Fortran (R Foundation, nedatováno).

Existuje řada vývojových prostředí, která se používají pro vývoj v R. Mezi nejznámější patří RStudio, které nyní spadá pod značku Posit. Jedná se o otevřený software pro datovou vědu, vědecký výzkum a technickou komunikaci. Mezi další často používané nástroje patří JupyterLab, Eclipse StatET nebo PyCharm, který byl původně vyvinut pro Python, ale lze jej také používat pro vývoj v jazyku R (Tribune Trust, 2023).

Mezi nejpopulárnější balíčky v jazyce R patří sbírka Tidyverse, která je zaměřena právě na datovou vědu. Dalšími známými balíčky jsou ggplot2, dplyr, Shiny, TensorFlow a tidymodels (Posit, © 2023).

3.4.1 Tidyverse

Jedná se o kolekci balíčku pro jazyk R, které jsou určeny pro datovou vědu. Balíčky slouží pro manipulaci s daty, jejich průzkum a vizualizaci. Všechny balíčky sdílejí společnou filozofii návrhu. Jejich cílem je zvýšit produktivitu statistiků a datových vědců. Do kolekce Tidyverse patří balíčky jako readr, tibble, tidyr, dplyr, ggplot2, purrr a další. Jednotlivé balíčky mají dobře zpracované dokumentace.

3.4.2 Readr

Readr má za úkol poskytnout rychlý a přívětivý způsob pro čtení tabulkových dat. Může se jednat například o soubory CSV nebo TSV. Je navržen tak, aby dokázal analyzovat různé typy dat, se kterými se lze setkat a poskytoval informativní hlášení o problémech, pokud by vedl parsování dat k neočekávaným výsledkům (Wickham, 2023). Je možné použít jiné balíky pro čtení dat, ale použití balíku readr umožňuje převod datového souboru na datový rámec tibble. Datový rámec tibble usnadňuje práci při odhalování anomálií v souborech. Tibble nemění názvy ani datové typy proměnných, nevyhazuje chyby, pokud proměnná neexistuje nebo chybí její hodnota (Vidhya, 2019).

3.4.3 Tibble

Tibble je další balíček, který poskytuje již zmiňovaný datový rámec tibble. Většina hlavních funkcí byla zmíněna v předchozí části. Ve zkratce tedy balíček nabízí vylepšenou strukturu pro uložení dat. Struktura nemění informace o datech ani samotná data a tím uživatele nutí čistit problémy v samotném počátku. Samotný balíček neobsahuje velké množství funkcí. Za zmínku stojí funkce print(), která je vylepšená oproti klasické funkci pro výpis a usnadňuje výpis velkých datových sad se složitými objekty (Müller, 2023).

3.4.4 Tidyr

Tidyr je balíček, který usnadňuje práci při vytváření tzv. „tidy dat“. To jsou taková data, kde každý sloupec je proměnná, každý řádek je pozorování a každá buňka obsahuje jednu hodnotu. S tímto typem dat je možné nadále pracovat v celé sadě tidyverse, což značně usnadňuje práci, pokud používáme nástroje z této sady (Wickham, nedatováno). Balíček obsahuje nástroje pro změnu a hierarchii datové sady.

3.4.5 Dplyr

Dplyr se snaží nabídnout jednoduché nástroje pro úlohy, které vyžadují manipulaci s daty. Balíček se inspiroval balíčkem plyr, který trpěl v některých případech svou rychlostí. Dplyr tento problém řeší prováděním velké části výpočtu v C++. Další výhodou je, že je možné pracovat s daty uloženými přímo v externí databázi. Lze tedy provádět dotazy na databázi a vracet pouze výsledky na tyto dotazy. Díky tomu částečně odpadá problém R, že se všechny operace provádějí v paměti (Datacarpentry, 2017). Jedná se o výkonný balík pro transformaci a sumarizaci tabulkových dat s řádky a sloupci. Obsahuje sadu funkcí, které provádějí operace jako filtrování řádků, výběr konkrétních sloupců, změnu pořadí řádků a přidávání nových sloupců (Irizarry, nedatováno).

3.4.6 Ggplot2

Ggplot2 je jedna z nejrozšířenějších alternativ pro vizualizaci dat oproti základním možnostem jazyka R a je velice flexibilní. Vychází z knihy „The Grammar of Graphics“ od L. Wilkinson. Vizualizace jsou vytvářeny a přizpůsobovány na základě přidávání vrstev, kdy každá vrstva představuje určitou součást grafu. Právě díky možnosti přidávat vrstvy dle potřeby je možné vytvářet jakákoli druh statické vizualizace dat (An Introduction to g g p l o t 2, nedatováno). Je možné přizpůsobit vše, co se týká grafu od barev, typů čar, písma, zarovnání, velikosti a mnohé další s funkcemi balíčku. Existují i funkce, pomocí kterých lze přidat nadpisy, podnadpisy, čáry, šipky nebo celé texty (Soage, 2023).

Za zmínku stojí určitě dva další balíčky z rodiny tidyverse. První z nich je Purrr, který rozšiřuje sadu nástrojů pro funkcionální programování a o sadu nástrojů pro práci s funkcemi a vektory. Obsahuje například funkci map, pomocí které lze nahradit „for loops“. Obsahuje také funkce pro filtraci, indexaci nebo modifikaci (Wickham, 2023). Druhý balíček se jmenuje stringr. Ten slouží hlavně pro úpravu řetězců. Je navržen tak, aby usnadnil práci s řetězci. Je postaven nad balíkem stringi, který poskytuje komplexní sadu funkcí, které pokrývají téměř vše, co s řetězci lze dělat. Samotný balíček stringr zase obsahuje nejdůležitější a nejčastěji používané funkce (Wickham, 2022).

3.5 Scala

Scala, i když v současnosti ustupuje v popularitě ve prospěch Pythonu, stále má své místo v oblasti datové vědy. Název Scala je zkratkou anglického výrazu "Scalable language", což znamená škálovatelný jazyk. Byla vytvořena s cílem překonat nedostatky jazyka Java, na které měli někteří vývojáři mnoho připomínek. Scala je malý, rychlý a efektivní multiparadigmatický programovací jazyk, který kombinuje principy objektově orientovaného a funkcionálního programování. Využívá JVM (Java Virtual Machine), což znamená, že kód je nejprve přeložen do bajtkódu a poté spuštěn na JVM, kde je generován výstup z tohoto bajtkódu. Vývojáři Scaly čerpali inspiraci z oblasti vědeckých a statistických výpočtu, které se často používají v oblasti zpracování dat (Boudreau, 2021).

Scala využívá různé knihovny pro zpracování datových proudů v reálném čase, například Spark Framework. K dispozici jsou také knihovny jako Apache Spark MLlib a ML pro strojové učení, ScalaNLP, Epic a Puck pro zpracování přirozeného jazyka a DeepLearning.scala pro hluboké učení. Pro analýzu dat jsou dostupné nástroje jako Breeze, Saddle, Scalalab a pro vizualizaci dat Breeze-viz nebo Vegas.

V současné době je otázka, zda má smysl používat Scala, když Python má mnohem širší použití. Scala má však několik výhod oproti Pythonu, například je téměř desetkrát rychlejší. Má podobnou syntaxi a některé funkce jako jiné populární jazyky, například Java nebo Ruby. Obsahuje vylepšené funkce pro porovnávání řetězců a vzorů. Nevýhodou může být menší komunita a obtížnost porozumění kombinaci funkcionálních a objektově orientovaných vlastností (Hamid, 2023).

3.5.1 ScalaNLP

Jedná se o sadu knihoven pro strojové učení, vědecké výpočty a zpracování přirozeného jazyka. Projekt ScalaNLP zastřešuje několik projektů včetně Breeze, Epic a Puck (Hall, © 2023).

Breeze je jedna z nejpopulárnějších a nejvýkonnějších knihoven pro lineární algebru. Je to čistá a výkonná numerická knihovna pro zpracování číselných dat. Vychází z knihovny NumPy, Matlabu a R. Byla vytvořena jako nástupce knihovny Scalala. Breeze dále nabízí sadu operátorů, které jsou podobné těm v Matlabu nebo Numpy, operace pro matice, pravděpodobnostní funkce, optimalizační balík a experimentální sadu pro vizualizace (Hall, 2021).

Epic je sada nástrojů, která slouží ke zpracování přirozeného jazyka a nástroje pro strukturovanou predikci. Obsahuje také tokenizaci, segmentaci vět, syntaktický rozbor, rozpoznávání pojmenovaných entit a označování částí řeči. Epic se dá používat buď programově z příkazového řádku, a to buď pomocí přetrénovaných modelů, nebo pomocí modelů, které si uživatel sám natrénuje. Aktuálně má Epic tři druhy modelů, a to jsou parsery, značkovače sekvencí a segmentátory. Parsery vytvářejí syntaktické reprezentace vět. Značkovače sekvencí slouží pro značkování částí řeči. Mohou například identifikovat podstatná jména, slovesa atd. Segmentátory zase rozdělují větu na posloupnost polí. Je možné pak identifikovat všechny osoby, místa nebo věci ve větě (Hall, 2014).

Puck je vysokorychlostní a vysoce přesný syntaktický parser, který používá ke zpracování grafické procesory (Hall, 2014). Ve většině případů se jedná o grafické karty značky Nvidia. Na hardwaru střední kvality dokáže analyzovat více než půl milionu slov za minutu. Aktuálně podporuje pouze angličtinu (Hall, 2023).

3.5.2 Smile

Jedná se o rychlý a komplexní systém pro strojové učení, zpracování přirozeného jazyka, lineární algebru, grafy, interpolaci a vizualizace. Má datové struktury a algoritmy díky kterým je schopen dosahovat vysoké výkonosti. Knihovna pokrývá každou část pro strojové učení

včetně klasifikace, regrese, shlukování, dolování asociačních pravidel, výběru příznaků, učení v nízko rozměrném prostoru, multidimenzionálního škálování, genetických algoritmů, doplňování chybějících hodnot, efektivního vyhledávání nejbližších sousedů a dalších. Smile v jednom benchmarktu výrazně porazil například R, Python, Spark, H2O, xgboost (Haifeng, 2014).

3.5.3 ScalaPy

V případě balíčku ScalaPy se nejedná o balík zaměřený na datovou vědu, ale byl vytvořen proto, aby usnadnil komunitě datové vědy implementaci aplikací pro datovou vědu a zvýšil výkon Scaly (ScalaPy, © 2022). Funkcí tohoto balíku je umožnit používat knihovny jazyka Python ve Scale, jako by se jednalo o její vlastní knihovny. To umožňuje využívat velké množství knihoven jazyka Python pro strojové učení a umělou inteligenci, a přitom využívat silné stránky Scaly, jako například zpracování velkých objemů dat (Baeldung, 2023).

3.6 MATLAB

MATLAB je jeden z nejstarších softwarových nástrojů na trhu a zároveň programovací jazyk, který pravidelně dostává aktualizace a nové verze. Jeho název "MATLAB" vychází z anglického termínu "matrix laboratory" což znamená maticová laboratoř. Je dostupný pro operační systémy Linux, Windows a Mac OS X. Jedná se o interaktivní programové prostředí a skriptovací programovací jazyk. Původně byl vytvořen pro studenty, aby nemuseli učit složitý programovací jazyk Fortran pro matematické výpočty (CIMSS, © 2023). Nicméně v současnosti má Matlab mnohem širší využití než pouze matematické výpočty. Je hojně využíván inženýry a vědci po celém světě v průmyslu, akademickém prostředí a ve výzkumech. Matlab se používá v oblastech jako je hluboké učení, strojové učení, zpracování signálů a komunikace, zpracování obrazu a videa, řídicí systémy, testování a měření, finance a biologie.

Jednou z výhod Matlabu je bohatá dokumentace, která obsahuje mnoho ukázkových příkladů, které mohou být velmi užitečné. Dokumentace poskytuje návody pro základy umělé inteligence, modelování AI, simulace a nasazení, což může být pro datovou vědu velmi cenné (Mathworks, © 2023).

3.6.1 Statistics and Machine Learning Toolbox

Jedna z mnoha sad nástrojů, které Matlab obsahuje je „Statistics and Machine Learning Toolbox“. Ta poskytuje funkce a aplikace pro popis, analýzu a modelování dat. Nástroje lze použít pro deskriptivní statistiku, vizualizace a shlukování pro průzkumnou analýzu dat, přizpůsobení rozdělení pravděpodobností dat, generovat náhodná čísla pro simulace Monte

Carlo a provádět testy hypotéz. Poskytuje pro analýzu vícerozměrných dat a extrakci prvků sadu nástrojů pro analýzu hlavních komponent, regularizaci, redukci rozměrů a metody výběru prvků. Obsahuje také algoritmy pro učení s učitelem, bez učitele nebo jejich kombinaci včetně metody podpůrných vektorů, vylepšených rozhodovacích stromů a dalších metod (Mathworks, © 2023).

3.6.2 Deep Learning Toolbox

Sada nástrojů pro hluboké učení obsahuje nástroje pro návrh a implementaci hlubokých neuronových sítí s algoritmy, předem natrénované modely a aplikace. Je možné využívat konvoluční neuronové sítě, LSTM sítě pro klasifikaci a regresi na obrázcích, časových řadách a textových datech. Pomocí automatické diferenciace, vlastních trénovacích smyček a sdílených vah můžete vytvářet architektury sítí, jako jsou generativní protikladné sítě (GAN) a siamské neuronové sítě. Pomocí aplikace Deep Network Designer lze graficky navrhovat, analyzovat a trénovat sítě. Aplikace Experiment Manager napomáhá spravovat experimenty v oblasti hlubokého učení, sledovat parametry trénování, analyzovat výsledky a porovnávat kód z různých experimentů. Je možné také vizualizovat aktivace vrstev a graficky monitorovat průběh trénování. MATLAB dokáže naimportovat jiné neuronové sítě a grafy vrstev z jiných knihoven jako je například TensorFlow2, TensorFlow-Keras nebo PyTorch. Je možné sítě také exportovat. Obsahuje také nástroje, které umožňují urychlit trénování na jednom nebo více grafických procesorů nebo používat pro výpočty cloud (Mathworks, © 2023).

3.6.3 Curve Fitting Toolbox

V této sadě jsou aplikace a funkce pro přizpůsobení křivek a povrchů dat. Nástroje umožňují provádět průzkumovou analýzu dat, předzpracovávat a zpracovávat data, porovnávat kandidátní modely a odstraňovat odlehlé hodnoty. Za použití knihovny lineárních a nelineárních modelů je možné provádět regresní analýzu nebo specifikovat vlastní rovnice. Knihovna poskytuje optimalizované parametry řešiče a počáteční podmínky, které slouží ke zlepšení kvality přizpůsobení. Sada obsahuje také nástroje pro neparametrické techniky modelování, splajny, interpolace a vyhlazování. Jakmile jsou nastaveny všechny přizpůsobení je možné aplikovat různé postprocessingové metody pro vykreslování, interpolaci, extrapolaci, odhadování intervalů spolehlivosti a výpočet integrálů a derivací (Mathworks, © 2023).

3.6.4 Text Analytics Toolbox

Tato sady nástrojů poskytuje algoritmy a vizualizace pro předběžné zpracování, analýzu a modelování textových dat. Modely, které jsou vytvořeny za pomoci této sady nástrojů je možné použít v aplikacích jako je analýza sentimentu, prediktivní údržba a modelování témat. Do této sady patří také nástroje pro zpracování surového textu z různých zdrojů, jako jsou záznamy z různých zařízení, zpravodajské kanály, průzkumy, zprávy operátorů a sociálních sítí. Je možné extrahovat text z populárních typů souborů, předzpracovávat surový text, extrahovat jednotlivá slova, převádět text do číselných reprezentací a vytvářet statistické modely. Lze využít techniky strojového učení LSA, LSD a vnoření slov pro hledání shluků a vytváření atributů z vysoko rozměrných textových datových souborů. Charakteristiky, které jsou vytvořeny za pomocí této sady je možné poté kombinovat s charakteristikami z jiných zdrojů dat a vytvářet modely strojového učení, které využívají textová, číselná a jiná data (MathWorks, © 2023).

3.7 Julia

Poslední jazyk, který zde budeme zmiňovat, je open-source jazyk Julia. Byl představen v roce 2012 a je relativně nový. Je vhodný pro různé oblasti, jako je strojové učení, vědecké výpočty a datová věda. Julia je vysokoúrovňový a dynamický jazyk, který se svým výkonem srovnává s tradičními staticky typovanými jazyky. Správně napsaný a optimalizovaný kód by měl být téměř stejně rychlý jako kód napsaný v jazyce C. Julia je multiparadigmatický jazyk, který kombinuje různé programovací paradigma, jako je imperativní, funkcionální, objektově orientované programování a metaprogramování. Lze v ní snadno a rychle provádět numerické výpočty na vysoké úrovni, podobně jako v jazycích Python, R nebo MATLAB, ale umožňuje i tradiční programování (Julialang, 2022). Julia přináší výhody z jiných jazyků, ale její nevýhodou je menší rozšířenost a podpora ve srovnání s jinými jazyky. To může znamenat, že některé problémy nemusí být snadno řešitelné a může být nutné, že bude potřeba vytvořit vlastní knihovny (Darina, 2022).

3.7.1 DataFrames.jl

Jedná se o knihovnu, která je podobná knihovnám pandas v Pythonu a data.frame, data.table a dplyr v jazyce R. To z této knihovny dělá výborný nástroj pro všeobecné účely v oblasti datové vědy. Knihovna poskytuje sadu nástrojů pro práci s tabulkovými daty a hraje důležitou roli v celém Julia data ekosystému. Pro Julia existuje mnoho knihoven, které jsou určeny pro práci s tabulkovými daty, ale DataFrames.jl má dobře propracované nástroje, se kterými se pracuje obdobně jako v jiných jazycích (DataFrames.jl, 2023).

Existuje také knihovna DataFramesMeta.jl, která je určena pro manipulaci s objekty DataFrames.jl. Obsahuje makra, která zlepšují výkon a poskytují pohodlnější syntaxi. Knihovna je inspirována knihovnou dplyr z jazyka R a LINQ z jazyka C# (Introduction, 2023).

3.7.2 Makie.jl

Makie je vysoce výkonný, rozšiřitelný a multiplatformní ekosystém pro vizualizace dat. Tato knihovna umožňuje práci s poli, jako jsou vektory a matice a díky tomu je schopna zpracovat různá tabulková data a datové struktury DataFrame.

Používá speciální typy bodů Point2f a Point3f, které se používají pro definování vektorů bodů ve 2D nebo 3D prostoru. Balíček Makie.jl je frontend, který definuje funkce pro vykreslování, které jsou potřeba pro vytvoření objektů k vykreslení. Tyto objekty ukládají potřebné informace k vykreslení a jsou vykresleny pomocí některého backendu.

Pro Makie.jl existují čtyři hlavní backendy. CairoMakie.jl pro neinteraktivní 2D vektorovou grafiku. GLMakie.jl pro interaktivní 2D a 3D vykreslování v samostatných oknech GLFW.jl. WGLMakie.jl poskytuje interaktivní 2D a 3D vykreslování v prohlížeči založené na WebGL. RPRMakie.jl je experimentální backend pro ray tracing, který využívá AMD RadeonProRender (Storopoli, 2021).

3.7.3 ScikitLearn.jl

Tato knihovna implementuje populární rozhraní scikit-learn, které je primárně určeno pro Python. Obsahuje jednoduché a efektivní nástroje pro prediktivní analýzu. Knihovna má velkou oblibu u výzkumníků strojového učení a datových vědců.

Knihovna poskytuje jednotné rozhraní pro trénování a používání modelů a také sadu nástrojů pro zřetězené zpracování úloh, vyhodnocování a ladění hyperparametrů modelů. Díky stejnému rozhraní, jako má Python je možné přistupovat k více než sto padesáti již existujícím modelům.

Podporuje také křížovou validaci, DataFrames a FeatureUnion. FeatureUnion slouží k propojení více objektů do nového, který kombinuje jejich výstupy. Jednotlivé transformace jsou na tyto objekty aplikovány samostatně a jsou spojeny až výsledky jednotlivých operací (Cstjean/ScikitLearn.js, 2023).

3.8 Elastic Stack

Jedná se o jednu z nejpopulárnějších platforem pro správu a analýzu logů. Původní název ELK Stack se odvíjí od tří klíčových open-source technologií: Elasticsearch, Logstash a Kibana. V dnešní době se k této čtveřici přidal nástroj Beats a proto nalezneme na oficiálním webu

název Elastic Stack, který se skládá z technologií Elasticsearch, Kibana a pojmu „Integrations“, který zastřešuje ostatní technologie, které je možné použít pro shromažďování dat. Tato sada nástrojů je rychlá a vysoce škálovatelná, umožňuje získat data z jakéhokoli zdroje v jakémkoli formátu. Následně lze v těchto datech vyhledávat, analyzovat je a vytvářet vizualizace (Elasticsearch Platform — Find real-time answers at scale, © 2023).

3.8.1 Logstash a Beats

Logstash je nástroj používaný pro zpracování dat na straně serveru. Přijímá data z různých zdrojů a následně je transformuje a odesílá dále, aniž by se zaměřoval na jejich formát nebo složitost. S využitím pluginu grok dokáže rozpoznat strukturu dat, dešifrovat zeměpisné souřadnice z IP adres, anonymizovat nebo vynechat citlivá pole. Logstash disponuje širokou paletou více než dvou set pluginů, které je možné vzájemně kombinovat pro řízení vstupů, výstupů a filtrů. (Logstash: Collect, Parse, Transform Logs, © 2023)

Beats funguje jako platforma pro jednoúčelové nástroje. Instaluje se jako agent na server, kde běží a poté odesílá data do nástrojů Logstash nebo Elasticsearch. Rodina Beats je rozdělena do různých modulů, z nichž každý se specifikuje na určitý typ dat. Tímto způsobem je možné odesílat logové soubory, metriky, data o síťovém provozu, protokoly událostí systému Windows, auditní data, žurnály a jiná data. (Beats: Data Shippers for Elasticsearch, © 2023)

3.8.2 Elasticsearch

Jeden z nejdůležitějších nástrojů celého Elastic Stacku je distribuovaný vyhledávací a analytický nástroj Elasticsearch. Jeho hlavní funkcí je ukládání, vyhledávání a analýza velkého množství dat téměř v reálném čase. Elastichsearch je schopný pracovat s různými typy dat, včetně strukturovaného nebo nestrukturovaného textu, číselných dat a geoprostorových dat.

Je schopen efektivně data ukládat a indexovat způsobem, který podporuje velmi rychlé vyhledávání (What is Elasticsearch?, © 2023). Data jsou uložena ve formě dokumentů, které jsou serializované JSON dokumenty. Elasticsearch využívá datovou strukturu invertovaný index, která podporuje velmi rychle fulltextové vyhledávání. Invertovaný index obsahuje seznam všech jedinečných slov, který se vyskytují v jakémkoli dokumentu a identifikuje všechny dokumenty, ve kterých se slova nachází.

Velkou výhodou je škálovatelnost, což může přispět k rozdělení výkonu mezi ostatní dostupné uzly (Data in: documents and indices, © 2023). Elasticsearch poskytuje také možnost vytvářet agregace dat, ze kterých lze získat klíčové metriky, vzorce a trendy z analyzovaných dat (Information out: search and analyze, © 2023). Nástroj automaticky distribuuje uložená data

a tím rozděluje zatížení mezi všechny dostupné uzly (Scalability and resilience: clusters, nodes, and shards, © 2023). Díky těmto funkcím je Elasticsearch mocným nástrojem pro správu a analýzu velkých objemů dat.

3.8.3 Kibana

Kibana je poslední klíčovou aplikací z celého Elastic Stacku a možná jednou z nejdůležitějších. Jejím hlavním účelem je umožnit tvorbu vizualizací, avšak nabízí i další užitečné funkce. Kromě vizualizací je možné využívat Kibanu i k vyhledávání a prohlížení indexovaných dat. Díky několika vestavěným aplikacím, jako jsou Lens, Canvas nebo Maps, je možné vytvářet různé typy grafů, tabulky, histrogramy, mapy a jiné vizualizace.

Důležitým prvkem Kibany je také schopnost sdílet vizualizace přes prohlížeč a zobrazovat aktuální data přímo z Elasticsearch. Tímto způsobem mohou uživatelé pracovat s reálnými daty a provádět analýzy z jednoho či více indexů. K častému využití patří tzv. Dashboard, který umožňuje zobrazovat více vizualizací nebo jiných dat na jednom místě. Krom toho je možné rozšířit funkčnost Kibany pomocí různých bezplatných pluginů, což dává uživatelům větší volnost a možnost přizpůsobit si aplikaci podle svých potřeb (What is Kibana?, © 2023).

4 PRAKTICKÁ ČÁST

Tato kapitola bude věnována praktickému výstupu práce, který si klade za úkol představit vybrané technologie, které byly popisovány v teoretické části. Hlavním cílem této části je provést ukázku analýzy dat a jejich následnou vizualizaci pomocí jazyků a nástrojů využívaných v této oblasti, které byly popisovány v teoretické části práce. Praktická část bude částečně dodržovat jednotlivé metodiky, které byly popisovány v teoretické části a přistupovat k nim dle potřeby práce.

4.1 Úvod do problematiky

V dnešní době se lze setkat s vysokým znehodnocováním osobních financí. To je zejména způsobené vysokou inflací, která je v Evropě, a především v České republice nejvyšší za poslední roky. Inflace má za následek pokles kupní síly peněz, což znamená že za stejnou částku si lze koupit méně zboží a služeb, než tomu bylo dříve. Příčin vzniku inflace může být více. Jednou z hlavních příčin je růst objemu peněz v ekonomice. Větší množství peněz vyvíjí tlak na růst cen. Za množství peněz v ekonomice zodpovídá centrální banka, která je zodpovědná za vydávání peněz do oběhu, tudíž výše inflace závisí na ní. Dalším druhem může být nabídková inflace, která je způsobena hlavně zdražováním vstupů potřebných na výrobu produktů. Stejný vliv může ale mít také oslabení měny, protože vstupy dovážené ze zahraničí jsou dražší. V neposlední řadě mohou způsobit nabídkovou inflaci také požadavky na vyšší mzdy zaměstnanců, kdy produktivita jejich práce neodpovídá výši jejich mzdě. Pokud firma zvedne mzdy zaměstnancům, tak to znamená, že se zvyšují její náklady, a tím vyvíjí tlak na růst cen produktu (Peníze, © 2023).

Jednou z možností, jak částečně ochránit své peníze jsou spořící účty. Spořící účty slouží k dlouhodobému ukládání peněz. Peníze na spořícím účtu jsou kdykoli k dispozici a oproti běžnému účtu nabízí vyšší úrok. Banky nabízí takový úrok, který je schopen peníze částečně chránit před inflací, která je znehodnocuje. Většinou má každá banka své podmínky, které mohou být pro každého jinak výhodné. Může se jednat například o různé velikosti úroku pro rozdílné částky peněz uložených na účtu. Úroky se poté mohou připisovat denně, měsíčně nebo čtvrtletně. Banka může průběžně měnit poskytovanou výši úroku dle sjednaných podmínek (Peníze, © 2023).

Další možností a v dnešní době velice populární může být investice do kryptoměn. Jedná se o riskantnější volbu, než jsou již zmiňované spořící účty, ale o to v zásadě se zajímavějším ziskem. Kryptoměny mohou být oproti spořícím účtům výhodné také pro krátkodobé investice

díky své vysoké volatilitě (Midjourney, 2023). Hlavním rozdílem mezi klasickou měnou a kryptoměnou je ten, že není ovlivňována žádnou centrální bankou nebo státní autoritou, která by mohla ovlivňovat její cenu. Většina kryptoměn je navíc omezena svým množstvím a může být tedy podobná vzácným komoditám jako je například zlato (Peníze, © 2023).

K řešení této problematiky jsou využita reálná data, která jsou volně dostupná na internetu. Data pochází z ověřených zdrojů jako je Český statistický úřad, World Bank, Yahoo Finance a Penize.cz. Data jsou očištěna a uložena jako soubory typu csv. Ty budou uchovány v systému, ve kterém se budou provádět analýzy a vizualizace daných dat. Pro tuto problematiku jsou zvoleny data o inflaci v Evropě a České republice, data o spořících účtech a data o vývoji cen kryptoměn Bitcoin, Ethereum a Tether USDt. Tyto kryptoměny byly vybrány proto, jelikož se jedná o nejrozšířenější aktuálně využívané kryptoměny. V následujících odstavcích budou krátce představeny uvedené kryptoměny pro jejich lepší pochopení.

Bitcoin je decentralizovaná kryptoměna, která má největší peněžní zastoupení na trhu kryptoměn. Jedná se o online měnu typu peer-to-peer. To znamená, že všechny provedené transakce probíhají přímo mezi účastníky dané transakce a není zapotřebí žádná autorita, která by musela povolovat nebo jinak usnadňovat tuto transakci. Podle slov tvůrce Bitcoinu vznikl právě proto, aby online platby nemusely procházet žádnou finanční institucí. Tato kryptoměna se může pyšnit hlavně tím, že se začala skutečně využívat jako jedna z prvních kryptoměn. Kryptoměna má kolem sebe obrovskou komunitu nadšenců, kteří vytvářejí nové kryptoměny, investují do nich, obchodují s nimi a používají je v každodenním životě. Bitcoin stále zůstává i po více než deseti letech na vrcholu trhu co se týče tržní kapitalizace. Jednou z jeho nejdůležitějších funkcí je, že se používá jako decentralizované úložiště hodnoty. Díky své špičkové kryptografii je právě považován za uchovatele hodnoty jako například zlato. Právě proto mnoho uživatelů vnímá Bitcoin jako investici pro uchování hodnoty namísto běžného platidla (CoinMarketCup, © 2023).

Ethereum je decentralizovaný open-source blockchainový systém, který obsahuje vlastní kryptoměnu Ether. Jedná se o druhou kryptoměnu s největší tržní hodnotou mezi kryptoměnami. Funguje jako platforma pro řadu dalších kryptoměn, ale také pro provádění tzv. decentralizovaných chytrých kontraktů, čím se nejvíce odlišuje například od Bitcoinu. Právě Ethereum je průkopníkem chytrých kontraktů pomocí blockchainu. Chytré kontrakty jsou počítačové programy, které automaticky provádějí nezbytné operace, které jsou potřeba ke splnění dohody mezi několika stranami, typicky mezi prodávajícím a kupujícím. Hlavním cílem

Etherea je stát se přední platformou pro decentralizované aplikace, které budou sloužit uživatelům psát a provozovat software odolný vůči cenzuře, výpadkům a podvodům (CoinMarketCup, © 2023).

Tether je platforma, které využívá blockchain a jejím hlavním úkolem je usnadnit používání klasických měn digitálním způsobem. Má sloužit jako digitální americký dolar, který je v poměru 1:1 s klasickým americkým dolarem. Oproti klasickému dolaru nabízí decentralizovaný způsob výměny hodnoty za pomocí známe účetní jednotky. Jeho unikátnost spočívá v tom, že jeho hodnota je garantována společností Tether, tak aby odpovídala právě americkému dolaru. Pokud společnost Tether přidá do oběhu nové jednotky USDt, tak zároveň přidá stejné množství amerických dolarů do svých rezerv a tím zajistí krytí kryptoměny hotovostí. Díky tomu se nabízí jako možnost pro investory v dobách vysoké volatility a není potřeba své prostředky proměnit ve skutečné dolary (CoinMarketCup, © 2023).

4.2 Vývojové prostředí

Tato kapitola je inspirována modelem Data Science Maturity Model, kdy je zapotřebí definovat nástroje, které budou použity pro cíle datové vědy. Pro praktickou část byly zvoleny nástroje Docker, Jupyter, Python, R a nástroje z rodiny Elastic Stack. Nástroje byly zvoleny tak, aby dokázaly vhodně řešit jednotlivé problémy, které budou řešeny v praktické části. Nástroje Jupyter, Python a R budou spouštěny v rámci jednoho společného balíčku, který obsahuje všechny tyto technologie. Dále budou použity jednotlivé Docker images pro nástroje z rodiny Elastic, kterými jsou Logstash, Elasticsearch a Kibana. V následující kapitole bude popsáno nastavení jednotlivých nástrojů, aby bylo možné lehce replikovat jejich nastavení pro spuštění za stejných podmínek.

4.2.1 Docker Desktop

Docker byl zvolen pro svou kontejnerizaci aplikací. Použití kontejnerů usnadňuje vývoj, zavádění a spouštění aplikací. Kontejner je zapouzdřená aplikace, která je spouštěna ve vlastním operačním systému. Docker se postará o všechny závislosti, jako jsou knihovny a balíčky, které jsou nutné pro běh aplikace. Docker nabízí především výbornou přenositelnost aplikace. Právě díky této přenositelnosti je možné spustit kontejner v jakémkoli operačním systému.

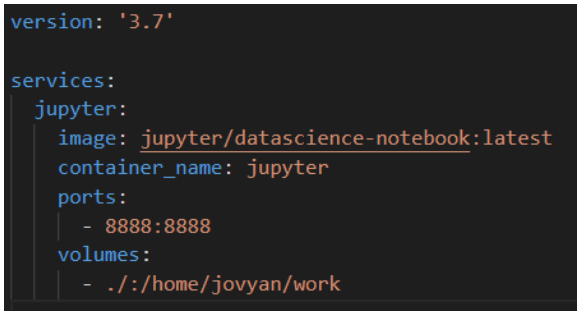
Nástroje JupyterLab, Python a R jsou zavedeny pomocí jedno společného docker image, který obsahuje samotný JupyterLab, Python, R, jazyk Julia, nástroje pro analýzu dat a mnoho dalších

nástrojů. Pro nástroje z rodina Elastic jsou použity oficiální obrazy pro nástroj Logstash, Elasticsearch a Kibana.

4.2.2 JupyterLab, Python a R

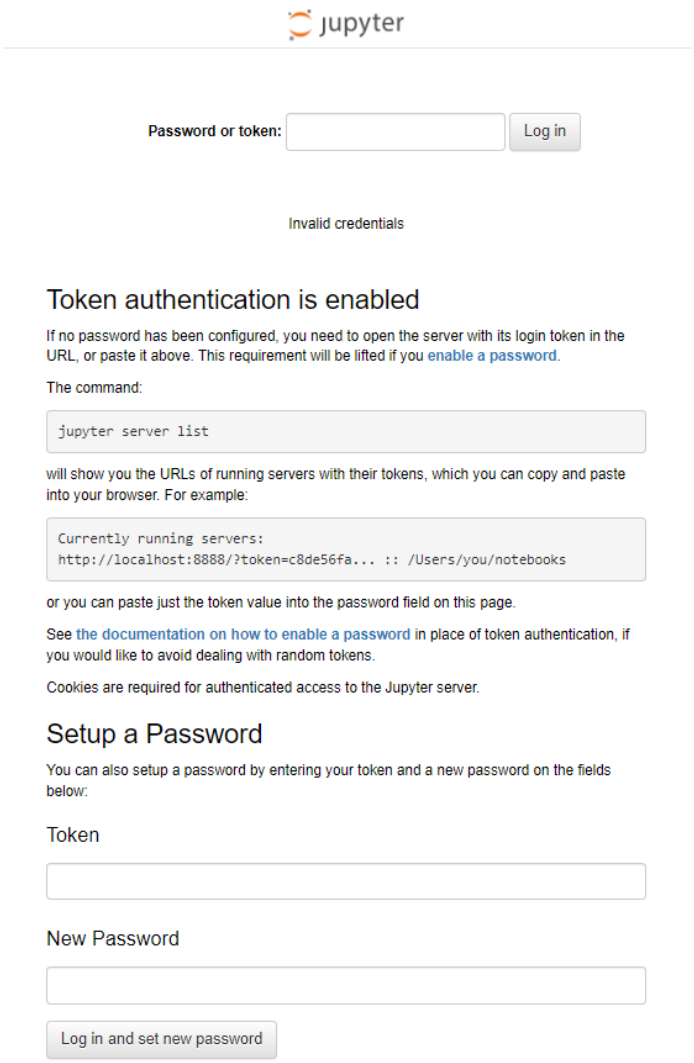
Nástroj JupyterLab byl zvolen hlavně pro své interaktivní prostředí, kdy je možné psát do notebooku kód, který je možné následně spustit. Velkou výhodou je, že existují docker images, které již obsahují nástroje pro datové vědce a tím usnadňují počáteční nastavení vývojového prostředí.

Pomocí příkazu „docker-compose up“ se provede spuštění souboru dokcer-copose.yaml. Pokud se Docker image nenachází v zařízení, bude stažen a poté spuštěn. Soubor obsahuje konfiguraci pro kontejner (viz. Obrázek 17), který bude sloužit pro Jupyter, Python a R. V konfiguraci, která je na následujícím obrázku lze vidět používaný Docker image „jupyter/datascience-notebook:latest“, jedná se o poslední dostupnou verzi tohoto nástroje. Vytvořený kontejner bude pojmenován „jupyter“ a bude naslouchat na portu 8888. Poslední část konfigurace slouží k propojení pracovního adresáře k propojení s virtuálním strojem.



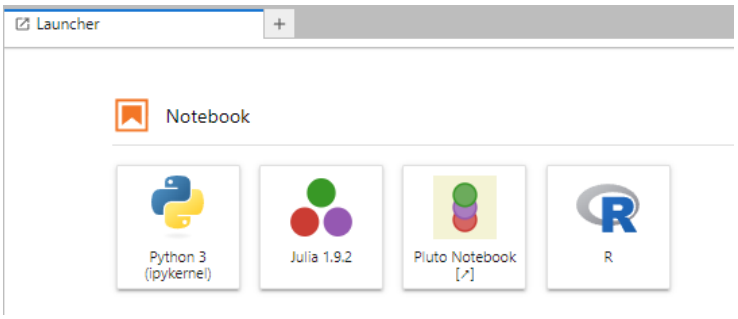
Obrázek 17: Kód ze souboru docker-compose.yaml pro nástroje Jupyter, Python a R

Při prvním spuštění je třeba se přihlásit pomocí tokenu, který lze najít v konzoli, kde je spuštěna aplikace. Pokud se nechceme vždy přihlašovat pomocí tokenu, tak je možné si vytvořit heslo, kterým se poté bude možné autentizovat.



Obrázek 18: Autentizace do aplikace Jupyter

Po přihlášení máme možnost si vybrat na kartě „Launch“ (viz. Obrázek 19) jeden z požadovaných Notebooku. Na výběr jsou ale také různé konzole, terminály nebo jiné soubory, které jsou v tomto Docker image podporovány.



Obrázek 19: Domovská karta v nástroji Jupyter

Python byl zvolen hlavně pro svou popularitu u datových vědců. Pokud chceme používat Jupyter Notebook, tak musíme mít nainstalovaný také Python. Proto je v kombinaci s Jupyterem ideální kombinací pro vývoj. Při

Trošku odlišným jazykem je R, který byl původně vyvinut pro statistiku a analýzu dat. Díky tomu nabízí různé specializované knihovny, které mohou pomoci při výpočtech, modelování dat a vizualizacích.

4.2.3 Elastic Stack

Logstash byl zvolen jako nástroj pro nahrání potřebných datových sad do databáze Elasticsearch. Následující část kódu, která je na obrázku je ve společném souboru „docker-compose.yaml“ společně s natavením pro Elasticsearch a Kibanu.

V tomto nastavení (viz. Obrázek 20) je použit oficiální Docker image od Elasticu ve verzi 7.9.3, i když aktuálně nejnovější verze je 8.8.1. Tato volba je motivována dobrou zkušeností s verzí 7.9.3, která má osvědčenou funkčnost a stabilitu.

Pro z nástroje Logstash je definován kontejner s názvem logstash, který bude naslouchat na portu 5044. Závislost na kontejneru elasticsearch zajišťuje, že Logstash bude spuštěn až poté, co je databáze Elasticsearch inicializována a připravena komunikovat. Tímto nastavením je zajištěno, že aplikace se spustí v pořadí, které zajišťuje jejich vzájemnou dostupnost a komunikaci.

Nakonec je nastaveno sdílení složek pomocí volumes. Složka ./logstash/config/ na fyzickém zařízení bude dostupná v kontejneru v cestě /usr/share/logstash/pipeline/, což umožní přenos konfiguračních souborů. Složka ./logstash/data/ na fyzickém zařízení bude sdílena s kontejnerem v cestě /data, což umožní přenos dat zpracovaných Logstashem do databáze Elasticsearch.

```
logstash:
  image: docker.elastic.co/logstash/logstash:7.9.3
  container_name: logstash
  ports:
    - "5044:5044"
  depends_on:
    - elasticsearch
  volumes:
    - ./logstash/config/:/usr/share/logstash/pipeline/
    - ./logstash/data/:/data
```

Obrázek 20: Kód ze souboru docker-compose.yaml pro nástroj Logstash

Následující tři obrázky (viz. Obrázky 21-23) se týkají konfiguračního souboru logstash.conf, který se stará o přenos dat. Následující část definuje soubor, který se bude posílat do Elasticsearch. Je definováno umístění souboru a odkud se soubor má začít procházet. Další část sincedb_path => „NULL“ je v souboru proto, aby se ignorovalo sledování stavu změn zpracování souboru. Tímto řádkem se také lze vyhnout nežádoucím hlášením při startu kontejneru. Poslední část označí data tagem „btc-usd“ a díky tomu poté lze rozeznat jednotlivé datové sady.

```
input {

  file {
    path => "/data/Crypto/BTC-USD.csv"
    start_position => "beginning"
    sincedb_path => "NULL"
    tags => ["btc-usd"]
  }
```

Obrázek 21: Nastavení konfiguračního souboru pro čtení souboru BTC-USD.csv

Nastavení filter obsahuje vlastnosti, kterými se definují vlastnosti souboru, jako jsou jeho sloupce, oddělovač sloupců, formát datumu a datové typy jednotlivých sloupců.

```
filter {

  if "btc-usd" in [tags] {
    csv {
      separator => ","
      columns => ["Date","Open","High","Low","Close","Adj Close","Volume"]
    }

    date {
      match => ["Date", "yyyy-MM-dd"]
      target => "Date"
      add_field => { "Date" => "%{Date}"}
    }

    mutate {
      convert => {
        "Open" => "float"
        "High" => "float"
        "Low" => "float"
        "Close" => "float"
        "Adj Close" => "float"
        "Volume" => "integer"
      }
    }
  }
```

Obrázek 22: Nastavení konfiguračního souboru pro filtraci atributů ze souboru BTC-USD.csv

Poslední část obsahuje informace o tom, kam se data mají poslat v našem případě do kontejneru elasticsearch s portem 9200. Data lze vyhledat pod indexem s názvem „indexelk“.

```
output {
  elasticsearch {
    hosts => ["elasticsearch:9200"]
    index => "indexelk"
  }
```

Obrázek 23: Nastavení pro výstup dat v konfiguračním souboru logstash.conf

Další nástroj z rodiny Elastic je Elasticsearch. Nastavení (viz. Obrázek 24) pro tento nástroj se nachází také v souboru „docker-compose.yaml“. Je použit oficiální Docker image ve stejné verzi 7.9.3 jako předchozí nástroj Logstash kvůli zaručení kompatibility. Vytvořený kontejner se jmenuje elasticsearch. Je použita proměnná dicovery.type s hodnotou single-node, která zjednodušuje konfiguraci pro lokální vývojové prostředí a není potřeba vyhledávat další uzly v síti. Elasticsearch bude naslouchat na portu 9200, pokud by byl tento port obsazen použije se port 9300. Poslední část slouží pro ukládání dat mimo kontejner elasticsearch.

```
elasticsearch:
  image: docker.elastic.co/elasticsearch/elasticsearch:7.9.3
  container_name: elasticsearch
  environment:
    - discovery.type=single-node
  ports:
    - "9200:9200"
    - "9300:9300"
  volumes:
    - esdata:/usr/share/elasticsearch/data
```

Obrázek 24: Kód ze souboru docker-compose.yaml pro nástroj Elasticsearch

Posledním nástrojem z Elastic Stacku je Kibana. Opět je použit oficiální Docker image od Elasticu ve verzi 7.9.3. Po spuštění je vytvořen kontejner, který se jmenuje kibana a naslouchá na portu 5601. Také Kibana potřebuje pro svůj běh Elasticsearch proto je nutné nastavit závislost na kontejneru elasticsearch.

```
kibana:
  image: docker.elastic.co/kibana/kibana:7.9.3
  container_name: kibana
  ports:
    - "5601:5601"
  depends_on:
    - elasticsearch
```

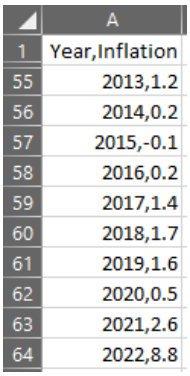
Obrázek 25: Kód ze souboru docker-compose.yaml pro nástroj Kibana

4.3 Datové sady

V této kapitole budou podrobně popsány datové sady, které byly zvoleny pro praktickou část práce. Bude zmíněn zdroj, odkud data pochází a struktura dat.

4.3.1 Inflace v Evropské unii

Datový soubor InflationEU.csv (viz. Obrázek 26) byl vytvořen na základě dat, která pochází z webu The World Bank, kde jsou uvedena data o inflaci v Evropské unii. Na webové stránce jsou uvedeny tyto údaje, a to rok a míra inflace v tento rok. Na základě těchto dat byl vytvořen soubor InflationEU.csv.

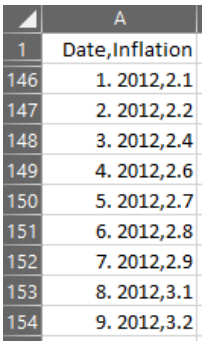


	A
1	Year,Inflation
55	2013,1.2
56	2014,0.2
57	2015,-0.1
58	2016,0.2
59	2017,1.4
60	2018,1.7
61	2019,1.6
62	2020,0.5
63	2021,2.6
64	2022,8.8

Obrázek 26: Ukázka datového souboru InflationEU.csv

4.3.2 Inflace v České republice

Datový soubor InflationCZ.csv (viz. Obrázek 27) byl vytvořen z dat, která poskytuje Český statistický úřad. Jde o údaje o inflaci, kde míra inflace je vyjádřena jako procentní změna průměrné ceny spotřebitelských výrobků za posledních 12 měsíců ve srovnání s průměrem cen za předcházejících 12 měsíců. Vytvořený datový soubor InflationCZ.csv obsahuje datové údaje jako je datum a inflace v daném roce.



	A
1	Date,Inflation
146	1. 2012,2.1
147	2. 2012,2.2
148	3. 2012,2.4
149	4. 2012,2.6
150	5. 2012,2.7
151	6. 2012,2.8
152	7. 2012,2.9
153	8. 2012,3.1
154	9. 2012,3.2

Obrázek 27: Ukázka datového souboru InflationCZ.csv

4.3.3 Spořící účty

Datový soubor SavingAccounts.csv (viz. Obrázek 28) obsahuje údaje o spořících účtech od 25. 1. 2018. Data, která byla použita pro vytvoření datového souboru pochází z webu Peníze.cz, kde jsou k nalezení srovnání spořících účtů pro různé časové období. Datový soubor obsahuje sloupce:

- Date – Datum, ze kterého záznam o spořícím účtu pochází.
- Banka – Název banky, která poskytuje spořící účet.
- Název – Název konkrétního produktu, který banka nabízí jako spořící účet.
- Maximální úrok (p.a.) – Maximální úrok (p.a.) znamená nejvyšší možnou míru úroku za rok.
- Omezení – Jednotlivé produkty mohou obsahovat různé omezující podmínky.

	A	B	C	D	E
2	17.07.2023	VÚB	Spoření bez limitů	6.15	Bez podmínek. ale: Jde o pobočku slovenské banky. Pokud klient nedodá potvrzení o daňovém domicilu z finančního úřadu. strhne mu banka 19% daň z příjmu (úrokového výnosu). Kdo potvrzení odevzdá. musí si sám vyřešit daňovou povinnost za kalendářní rok (15% sazba v Česku). Kdo nepřekročí limity pro zdanitelné příjmy. nemusí přiznání podávat.
3	17.07.2023	Max banka	Spořicí účet	6.01	Bez podmínek.
4	17.07.2023	UniCredit Bank	Akce k novému běžnému účtu	6	Pouze pro nové klienty. kteří si založí běžný účet. Úročí se vklady do půl milionu. Banka garantuje. že úrok nesníží do konce července. Dřívějším klientům dává standardně jen 2.5 %.
5	17.07.2023	Trinity Bank	Výhoda+ Dobrý klient	5.68	Pro vklady do 500 000 Kč. Pro část nad tento limit je úrok 4.08 %. Na účtech zřízených od 22. 3. 2023 při splnění podmínek (zejména nového vkladu) garantuje. že úrok 5.68 % neklesne do konce roku.
6	17.07.2023	Banka Creditas	Spořicí účet+	5.6	Pro vklady do 500 000 Kč. Pro část nad tento limit je úrok 3.1 %. S výpovědní lhůtou jeden měsíc je úrok 6 % (do 500 000) a 4 % (nad 500 000).
7	17.07.2023	Fio banka	Fio konto	5.5	Pro vklady do 200 000 Kč. Pro pásmo od 200 000 do milionu je úrok 0.1 %. od jednoho do deseti milionů 0.15 %. nad deset milionů 0.2 %. Minimální zůstatek je 100 Kč. vrací se až při zrušení účtu.
8	17.07.2023	mBank	Spoření	5.5	Podmínkou je (bezplatný) běžný účet a volba jedné ze čtyř variant pravidelného spoření. To probíhá formou cílů. v každém lze naspořit max. 100 000 Kč. Překročení hranice znamená pád celého vkladu do prakticky nulového úročení. Limit lze zvýšit až na 800 000 Kč díky tomu. že je možné založit až osm cílů.

Obrázek 28: Ukázka datového souboru SavingAccounts.csv

4.3.4 Bitcoin

Data o Bitcoinu pochází ze stránky Yahoo Finance, která přebírá data z webu CoinMarketCap. Jedná se o jednu z nejdůvěryhodnějších stránek, která obsahuje informace o kryptoměnách. Má sloužit pro zpřístupnění a zefektivnění objevování kryptoměn všem uživatelům. Poskytuje nestranné, vysoce kvalitní a přesné informace pro vyvození informovaných závěrů (CoinMarketCup, © 2023).

Soubor BTC-USD.csv (viz. Obrázek 29) obsahuje data od 17. 9. 2014 do 3. 8. 2023. Jedná se o více než 3 tisíce záznamů. Soubor obsahuje následující datové sloupce:

- Date – Tento datový sloupce obsahuje záznam o datumu, z kterého kdy byl záznam zaznamenán.

- Open – Jedná se o cenu, za kterou byl daný finanční nástroj obchodován jako první transakce v daném časovém období, obvykle na začátku dne.
- High – Nejvyšší cena, za kterou se v daný den finanční nástroj obchodoval.
- Low – Nejnižší cena, za kterou se v daný den finanční nástroj obchodoval.
- Close – Označuje cenu finančního instrumentu, za kterou se provedla poslední transakce v daném časovém období.
- Adj Close – Jedná se o upravenou konečnou cenu, která je upravena o příslušná rozdělení, výplaty dividend nebo jiné události, které by mohly ovlivnit cenu finančního nástroje.
- Volume – Celkový počet transakcí, které se provedly v daný den. Jedná se o prodej a koupi (Smigel, 2023).

	A	B	C	D	E	F	G	H
1	Date,Open,High,Low,Close,Adj Close,Volume							
2	2014-09-17,465.864014,468.174011,452.421997,457.334015,457.334015,21056800							
3	2014-09-18,456.859985,456.859985,413.104004,424.440002,424.440002,34483200							
4	2014-09-19,424.102997,427.834991,384.532013,394.795990,394.795990,37919700							
5	2014-09-20,394.673004,423.295990,389.882996,408.903992,408.903992,36863600							
6	2014-09-21,408.084991,412.425995,393.181000,398.821014,398.821014,26580100							
7	2014-09-22,399.100006,406.915985,397.130005,402.152008,402.152008,24127600							
8	2014-09-23,402.092010,441.557007,396.196991,435.790985,435.790985,45099500							
9	2014-09-24,435.751007,436.112000,421.131989,423.204987,423.204987,30627700							
10	2014-09-25,423.156006,423.519989,409.467987,411.574005,411.574005,26814400							

Obrázek 29: Ukázka datového souboru BTC-USD.csv

4.3.5 Ethereum

Jako v případě Bitcoinu, tak i data o Ethereum pochází ze stránky Yahoo Finance, která Yahoo Finance přebírá z webu CoinMarketCap. Datová sada ETH-USD.csv (viz. Obrázek 30) obsahuje stejné datové sloupce jako předchozí datová sada. Obsahuje záznamy od 9. 11. 2017 do 3. 8. 2023, kterých je přes 2 tisíce. Datové sloupce jsou:

- Date – Tento datový sloupce obsahuje záznam o datumu, z kterého kdy byl záznam zaznamenán.
- Open – Jedná se o cenu, za kterou byl daný finanční nástroj obchodován jako první transakce v daném časovém období, obvykle na začátku dne.
- High – Nejvyšší cena, za kterou se v daný den finanční nástroj obchodoval.
- Low – Nejnižší cena, za kterou se v daný den finanční nástroj obchodoval.
- Close – Označuje cenu finančního instrumentu, za kterou se provedla poslední transakce v daném časovém období.

- Adj Close – Jedná se o upravenou konečnou cenu, která je upravena o příslušná rozdělení, výplaty dividend nebo jiné události, které by mohly ovlivnit cenu finančního nástroje.
- Volume – Celkový počet transakcí, které se provedly v daný den. Jedná se o prodej a koupi (Smigel, 2023).

	A	B	C	D	E	F	G	H
1	Date,Open,High,Low,Close,Adj Close,Volume							
2	2017-11-09,308.644989,329.451996,307.056000,320.884003,320.884003,893249984							
3	2017-11-10,320.670990,324.717987,294.541992,299.252991,299.252991,885985984							
4	2017-11-11,298.585999,319.453003,298.191986,314.681000,314.681000,842300992							
5	2017-11-12,314.690002,319.153015,298.513000,307.907990,307.907990,1613479936							
6	2017-11-13,307.024994,328.415009,307.024994,316.716003,316.716003,1041889984							
7	2017-11-14,316.763000,340.177002,316.763000,337.631012,337.631012,1069680000							
8	2017-11-15,337.963989,340.911987,329.812988,333.356995,333.356995,722665984							
9	2017-11-16,333.442993,336.158997,323.605988,330.924011,330.924011,797254016							
10	2017-11-17,330.166992,334.963989,327.523010,332.394012,332.394012,621732992							

Obrázek 30: Ukázka datového souboru ETH-USD.csv

4.3.6 Tether USDt

Data jsou stažena ze stránky Yahoo Finance, která data přebírá z webu CoinMarketCap. Poslední datový soubor USDT-USD.csv (viz. Obrázek 31) není jiný oproti předchozím souborům obsahujících záznamy o kryptoměnách. Obsahuje záznamy od 9. 11. 2017 do 3. 8. 2023, kde je více než 2 tisíce záznamů. Datové sloupce jsou také:

- Date – Tento datový sloupce obsahuje záznam o datumu, z kterého kdy byl záznam zaznamenán.
- Open – Jedná se o cenu, za kterou byl daný finanční nástroj obchodován jako první transakce v daném časovém období, obvykle na začátku dne.
- High – Nejvyšší cena, za kterou se v daný den finanční nástroj obchodoval.
- Low – Nejnižší cena, za kterou se v daný den finanční nástroj obchodoval.
- Close – Označuje cenu finančního instrumentu, za kterou se provedla poslední transakce v daném časovém období.
- Adj Close – Jedná se o upravenou konečnou cenu, která je upravena o příslušná rozdělení, výplaty dividend nebo jiné události, které by mohly ovlivnit cenu finančního nástroje.
- Volume – Celkový počet transakcí, které se provedly v daný den. Jedná se o prodej a koupi (Smigel, 2023).

	A	B	C	D	E	F	G
1	Date,Open,High,Low,Close,Adj Close,Volume						
2	2017-11-09,1.010870,1.013270,0.996515,1.008180,1.008180,358188000						
3	2017-11-10,1.006500,1.024230,0.995486,1.006010,1.006010,756446016						
4	2017-11-11,1.005980,1.026210,0.995799,1.008990,1.008990,746227968						
5	2017-11-12,1.006020,1.105910,0.967601,1.012470,1.012470,1466060032						
6	2017-11-13,1.004480,1.029290,0.975103,1.009350,1.009350,767884032						
7	2017-11-14,1.005240,1.013430,0.996898,1.006830,1.006830,429857984						
8	2017-11-15,1.004580,1.011630,1.000250,1.003180,1.003180,449671008						
9	2017-11-16,1.005820,1.010890,0.993232,1.002120,1.002120,650278976						
10	2017-11-17,0.995758,1.011810,0.995758,1.001390,1.001390,639398016						

Obrázek 31: Ukázka datového souboru USDT-USD.csv

4.4 Data

Tato kapitola přebírá z modelu CRISP-DM dva hlavní rámce. Prvním rámcem je „Data Understanding“. Konkrétně se jedná o ověřování kvality dat, kdy je potřeba zaručit, že zdroje odkud data pochází jsou dostatečně kvalitní. Dále se k tomuto rámci vztahuje také shromažďování počátečních dat. Druhý rámec, který se vyskytuje v této kapitole je „Data preparation“. Sem spadají úkoly jako příprava, očištění a transformace dat.

4.4.1 Popis zdrojů dat.

První tři datové sady (BTC-USD.csv, ETH-USD.csv a USDT-USD.csv), které jsou použity pochází z webu Yahoo finance. Jedná se o kryptoměny Bitcoin, Ethereum a Tether USDt. Jedná se o web, který poskytuje komplexní informace o finančních trzích, akciích, komoditách, měnových kurzech, kryptoměnách a mnoho dalších informací ze světa financí. Jedná se o velmi populární zdroj, který je využíván mnoha odborníky a dá se považovat za kvalitní zdroj. Data ohledně kryptoměn nepochází přímo z Yahoo finance, ale jsou přebírány ze stránky CoinMarketCap. Jedná se o webovou platformu, která se specializuje na sledování, analýzu a prezentací informací o kryptoměnách. Tato služba je jedna z nejrozšířenějších poskytovatelů informací o kryptoměnách a čerpá z ní mnoho jiných zdrojů. Vzhledem k tomu, že jsou kryptoměny více anonymní a není u nich taková regulace a dohled jako u tradičních akcií, tak by data měla být z jakéhokoli zdroje brána s rezervou.

Další datová sada inflace v Evropské unii pochází z webu The World Bank, což je mezinárodní instituce, která se zaměřuje na poskytnutí finančních a technických prostředků ve snaze podpořit rozvojové země a snížit chudobu po celém světě. Mezi klíčové vlastnosti Světové banky patří sběr, analýza a publikování ekonomických dat a indikátorů, které mají za cíl poskytnout informace o ekonomické situaci zemí a regionů. The World Bank se dá považovat za velice spolehlivý zdroj (World Bank Group, © 2023).

Data o inflaci v České republice pochází z českého zdroje Český statistický úřad. Jedná se o ústřední orgán státní správy České republiky. Hlavními úkoly statistického úřadu je bezpečně získávat a zpracovávat údaje pro statistické účely a poskytnout statistické informace státním orgánům, orgánům uzemní samosprávy, veřejnosti a do zahraničí. Český statistický úřad je obecně vnímán jako velice kvalitní, ověřený zdroj a data, která poskytuje by měla být dostatečně relevantní (ČSO, 2023).

Poslední data pochází z webu Peníze.cz. Jedná se o internetový magazín, který se vyskytuje na internetu více než 20 let. Jeho úkolem je poskytnout čtenářům informace z oblasti osobních a rodinných financí co nejvíce srozumitelnou formou. Informace o spořících účtech pochází z článků, které se vyskytují na webu. Jedná se o články, které obsahují srovnání spořících účtů v tabulce. Tabulky v jednotlivých článcích uvádí jako zdroj oficiální dokumenty bank. Z tohoto důvodu je možné na data pohlížet tak, že pocházejí z ověřeného a kvalitního zdroje.

4.4.2 Sběr dat

Yahoo finance nabízí možnost zobrazeních historických dat. Lze si zvolit časové období ze kterého budou data zobrazena. Dále je možné si vybrat jednu ze tří frekvencí datových záznamů a to denní, týdenní a měsíční. Pro Bitcoin, Ethereum a Tether bylo zvoleno maximální možné časové období s frekvencí denních záznamů. Následně je možné si data stáhnout pomocí nabízení možnosti „Download“, která stáhne datový soubor ve formátu csv. Výsledné datové soubory jsou BTC-USD.csv, ETH-USD.csv a USDT-USD.csv.

V ostatních případech jsou data sbírána manuálně a tvořena dle potřeby. The World Bank nabízí také možnost stažení souboru, který obsahuje mnoho informací nejenom o inflaci v Evropské unii. Při pokusu o stažení souboru se data zdála neúplná, odlišná. Proto byl zvolen manuální přepis z vizualizace, která je k dispozici na webu. Data byla sesbírána z maximálního možného roku, který byl k dispozici. Data jsou tedy v rozsahu let 1960-2020. Výsledný získaný soubor se jmenuje InflationEU.csv

Český statistický úřad nabízí data o inflaci v přehledné tabulce, která je k dispozici na jejich webu. Data byla manuálně překopírovaná do souboru v excelu a uložena do formátu csv. Tabulka je ve formátu, když na řádcích jsou roky 2000-2023 a ve sloupcích jednotlivé měsíce 1-12. Vytvořený soubor s daty se jmenuje InflationCZ.csv

Poslední data byla sesbírána z různých článků o srovnání spořících účtu na webu Peníze.cz. Článek vždy obsahoval tabulku, která obsahovala potřebné informace. Data byla nutná překopírovat z jednotlivých článků do výsledného souboru, který byl pojmenován

SavingAccounts.csv. Data neobsahovala žádné časové údaje, a proto bylo nutné jim přidělit datumy, které se nacházely buď v titulku tabulky nebo byla získána z doby vydání příslušného článku či byl k dataci použit obsah a datum doplněno.

4.4.3 Příprava dat

V případě relevantnosti atributů vyskytujících se v datových souborech kryptoměn se dají považovat všechny atributy za relevantní. Bylo by možné uvažovat o vynechání atributu „Adj Close“, který se na první pohled neliší od atributu „Close“. Ceny by se mohly lišit, pokud by u dané kryptoměny například docházelo k vyplácení dividendy, docházelo k dělení nebo slučování kryptoměn nebo by došlo k jiné technické anomálii. Protože je možné, že během doby, ze které historická data pochází mohlo dojít ke korekci, tak bude atribut ponechán.

Pro data týkající se inflace nedochází k žádné změně jak pro Evropskou unii, tak pro Českou republiku, protože data poskytována The World Bank a Českým statistickým úřadem obsahovala pouze časový údaj a hodnotu inflace.

Poslední data se týkají spořících účtů. Také v tomhle případě se data již nachází v konečné formě. Pro analýzu by bylo možné vynechat atribut „Omezení“, který můžeme mít svůj význam při zkoumání jednotlivých produktů, které banky nabízejí.

4.4.4 Očištění a transformace

V datových souborech BTC-USD.csv, ETH-USD.csv a USDT-USD.csv chybí záznamy k datu 2. 8. 2023. Tento záznam je možné doplnit z jiného zdroje, odstranit v nástroji ve kterém bude probíhat datová analýza nebo odstranit již v datovém souboru, kde jsou data uložena. V našem případě bude záznam s neplatnými hodnotami odstraněn ve všech datových souborech, aby se předešlo výskytu možných chyb při načítání nebo jiném zpracování souboru. Žádné transformace dat nejsou nutné vzhledem k datovému formátu všech sloupců. Bylo by možné odstranit některé datové sloupce, jako například „Adj Close“ nebo „Volume“, ale vzhledem k malé velikosti souboru, je lepší ponechat soubor kompletní, pokud by bylo nutné zkoumat nové hypotézy.

Datový soubor InflationEU.csv nebyl potřeba očišťovat ani transformovat vzhledem k tomu, že byl tvořen manuálně. Soubor InflationCZ.csv bylo potřeba upravit i transformovat. Původní data byla zkopírována z webu, kde byly rozdělena do dvanácti sloupců. Nově byla data transformována do dvou sloupců „Date“ a „Inflation“. Hodnota inflace obsahovala desetinnou čárku, která byla nahrazena desetinnou tečkou.

Pro soubor SavingAccounts.csv došlo ke změně desetinné čárky ve sloupci „Maximální úrok (p.a.)“ a ve sloupci „Omezení“ za desetinnou tečku. Ve sloupci „Maximální úrok (p.a.)“ bylo potřeba odstranit znak procenta, aby bylo možné provádět numerické operace.

4.5 Úvod do analýzy

V této kapitole budou představeny výzkumné otázky, kterým se budě věnovat praktická část. Kapitola je inspirována procesním modelem CRISP-DM konkrétně rámcem „Business Understanding“ do kterého spadá určení cílů datové vědy.

4.5.1 Výzkumné otázky

VO1: *Jak se budou vyvíjet jednotlivé kryptoměny, inflace a nabízené úroky na spořících účtech v čase?*

První otázka si klade za cíl sledování dlouhodobého vývoje kryptoměn, inflace a úroku na spořících účtech. Pro získání odpovědí na tuto otázku lze provést časovou analýzu historických dat, vytvořit grafy a trendy a vyhodnotit, jakým způsobem se jednotlivé ukazatele vyvíjejí.

VO2: *Jaké jsou základní statistiky kryptoměn a inflace a jaký vliv má inflace na hodnotu kryptoměn?*

Tato otázka si klade za cíl zjistit, zda existuje vztah mezi inflací a hodnotou kryptoměn. Je možné provést korelační analýzu mezi inflací a cenami kryptoměn, porovnat vzory změn a zkoumat, zda existuje nějaký náznak vlivu inflace na hodnotu kryptoměn.

VO3: *Jaký je rozdíl mezi inflací a nabízenými úroky na spořících účtech za posledních 5 let?*

Třetí výzkumná otázka se zabývá možným vztahem mezi inflací a úrokovými sazbami na spořících účtech. Zde je také možné provést analýzu historických dat úrokových sazeb a inflace. Lze zkoumat, zda se úrokové sazby na spořících účtech mění v reakci na inflační změny.

VO4: *Jak se proměňují historické trendy cen kryptoměn v různých časových obdobích?*

Ve výzkumné otázce číslo čtyři je za úkol sestrojit podrobnou vizuální analýzu historických trendů cen kryptoměn v různých časových obdobích. Cílem je identifikovat a zkoumat možné vzory, výkyvy a závislosti mezi vývojem cen kryptoměn.

4.6 Postup zpracování dat

VO1: *Jak se budou vyvíjet jednotlivé kryptoměny, inflace a nabízené úroky na spořících účtech v čase?*

Pro zpracování výzkumné otázky č. 1 je vybrán nástroj Kibana, který nabízí využití strojového učení pro predikci budoucího chování. Konkrétně se jedná o nástroj pro detekci anomálií. Tento nástroj provede analýzu minulosti, přítomnosti a zároveň detekuje anomálie v datech. Funkce strojového učení automatizuje analýzu časových řad dat tím, že vytváří přesné základní úrovně normálního chování v datech a identifikuje anomální vzory v těchto datech. Funkce pro detekce anomálií využívají unikátní kombinace různých technik, jako je shlukování, různé typy rozkladu časových řad, modelování Bayesovských distribucí a korelační analýzu (Elasticsearch, © 2023).

Jakmile funkce strojového učení vytvoří základní model normálního chování pro data, lze přejít k předpovědi budoucího chování. Zde je možné nastavit délku předpovědi pro příštích 3650 dní. V našem případě je zvoleno 1825 dní pro předpověď budoucnosti. Díky této předpovědi je možné odhadnout například budoucí cenu Bitcoinu v libovolný den.

Kibana byla zvolena pro analýzu této otázky hlavně z důvodu, že je schopná rychle zanalyzovat historická data, odhalit anomálie a je schopná pomocí svých algoritmů předpovědět budoucnost. Pokud bychom chtěli použít Python nebo R, tak by bylo nutné pomocí dat natrénovat, otestovat a vytvořit vlastní model pro predikci budoucího chování, což by zabralo více času. Samozřejmě by taková možnost umožňovala si vytvořit model dle svých potřeb, ale to ovšem představuje nutnost určitých znalostí z oblasti strojového učení.

VO2: *Jaké jsou základní statistiky kryptoměn a inflace a jaký vliv má inflace na hodnotu kryptoměn?*

Pro zpracování výzkumné otázky č. 2 byl zvolen nástroj R. Jazyk R nabízí širokou škálu pokročilých statistických nástrojů a balíčků, které umožňují provádět analýzu dat. Díky schopnosti manipulovat s daty, vytvářet vizualizace a provádět statistické testy je vhodným nástrojem pro tuto analýzu.

Jazyk R nabízí funkci „summary“, která slouží pro zobrazení souhrnných statistik různých objektů. Poskytuje základní statistiky jako je minimum, maximum, průměr, medián, kvartily a počet chybějících hodnot.

Pro zjištění závislosti mezi inflací a jednotlivými kryptoměny bude sestrojena korelační matice, která umožňuje analyzovat vztahy mezi různými proměnnými v datech. Korelační matice identifikuje vzájemné vztahy a určí sílu těchto vztahů. Bude použita funkce „cor“, která využívá Pearsonův korelační koeficient k měření síly lineární závislosti mezi dvěma veličinami.

V jazyku R byl vytvořen společný datový rámec, ve kterém jsou informace o dané kryptoměně společně s inflací v České republice. Data o inflaci jsou za každý měsíc v roce, proto byla tato hodnota přiřazena vždy odpovídajícímu záznamu dané kryptoměny z téhož měsíce a roku.

VO3: *Jaký je rozdíl mezi inflací a nabízenými úroky na spořících účtech za posledních 5 let?*

Třetí otázka si klade za cíl zjistit, jak se liší nabízené úroky bank na spořících účtech oproti výši inflace za posledních 5 let. Pro tuto otázku by bylo možné zvolit, kterýkoli ze tří používaných nástrojů, avšak Python díky své jednoduchosti, rychlosti a knihovnám nabízí ideální řešení.

Pro zpracování této úlohy bude vytvořena vizualizace pomocí balíčku matplotlib, která bude porovnávat vývoj inflace a vývoj průměrného úroku od roku 2018. Je potřeba provést výpočty průměrné inflace a průměrného úroku pro každý rok. Jako druhý výstup pro tuto otázku bude tabulka, která bude zobrazovat roční rozdíl mezi průměrnou roční inflací a průměrným maximálním úrokem, který banky nabízejí.

VO4: *Jak se proměňují historické trendy cen kryptoměn v různých časových obdobích?*

Poslední výzkumná otázka se snaží odhalit trendy kryptoměn v různých časových obdobích. Bude provedena analýza cen jednotlivých kryptoměn v rámci týdnů, měsíců a roků. Pro tuto analýzu bude použit nástroj Python, který nabízí funkce, které umožňují rozdělit datovou sadu na potřebné části, které je možné poté vizualizovat. Pro vizualizaci dat bude použita knihovna matplotlib, pro práci s daty knihovna pandas. Při práci je použita například funkce „pct_change“, která vypočítá procentuální změnu aktuálního prvku oproti předchozímu. Dále je použita funkce „resample“, která zase převede na základě časového údaje data na požadované časové frekvence, jako například týden, měsíc, rok. Dále jsou použity funkce pro tvorbu vizualizací a funkce pro deskriptivní statistiku.

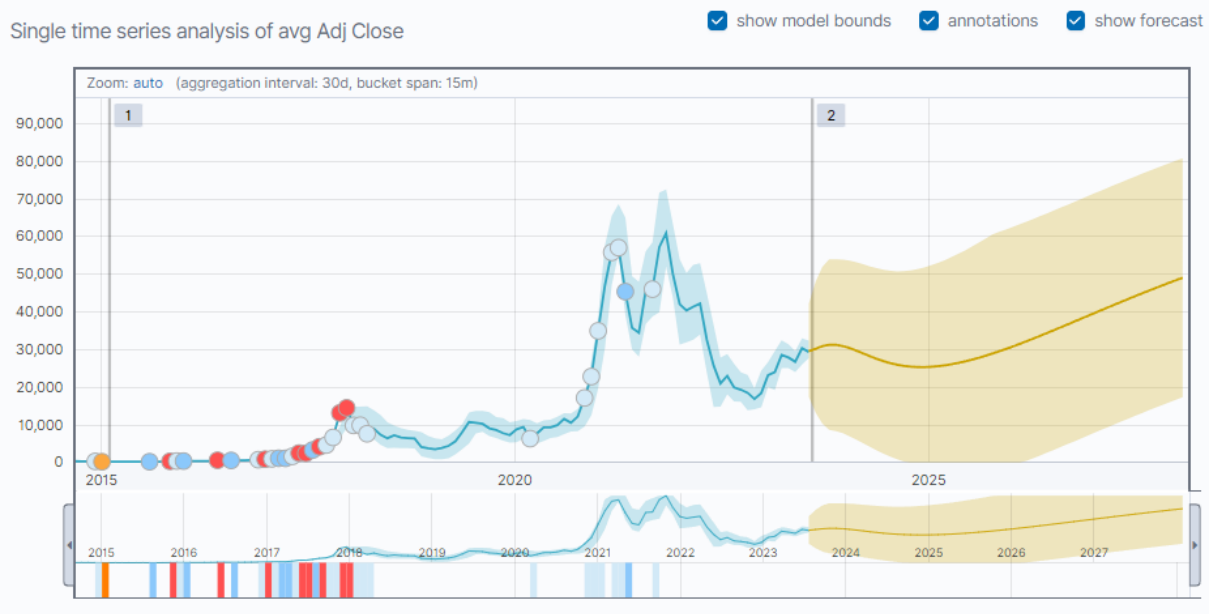
4.7 Vizualizace dat

VO1: *Jak se budou vyvíjet jednotlivé kryptoměny, inflace a nabízené úroky na spořících účtech v čase?*

Všechny vizualizace byly provedeny pro příštích 1825 dní od posledního záznamu, který byl k dispozici. Po vytvoření jednotlivých vizualizací bylo zobrazeno varování, že nelze přesně předpovědět budoucnost pro celý požadovaný časový interval, proto se v některých případech nejedná přesně o 1825 dnů. Zároveň vždy byly použity průměrné hodnoty atributů, které byly použity pro vytvoření modelů a následných vizualizací.

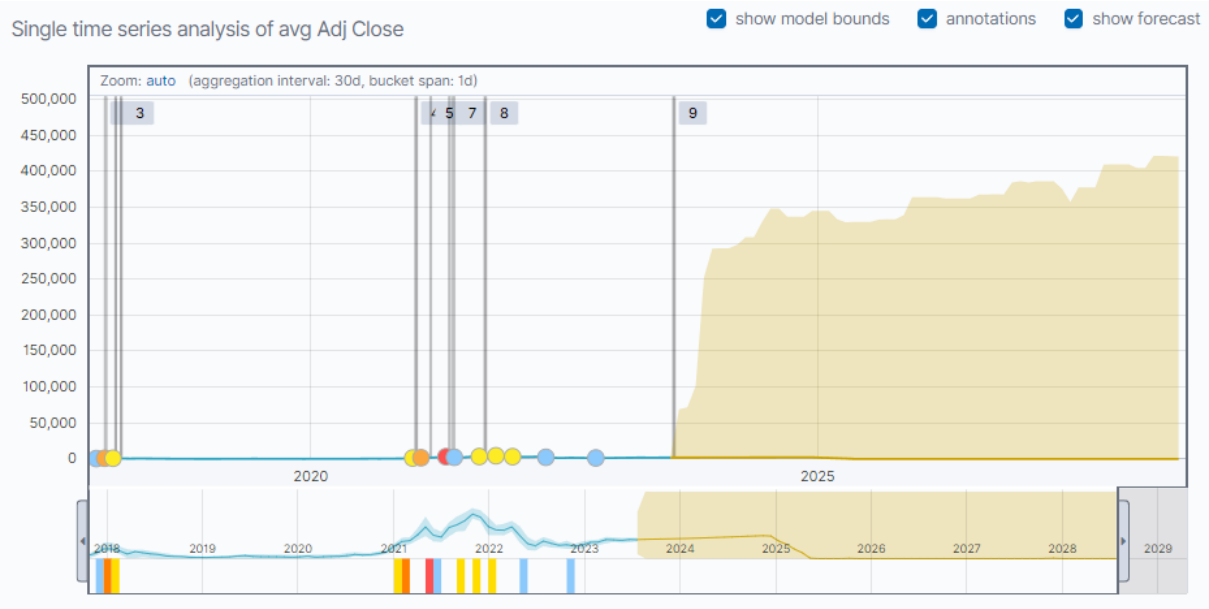
Každá vizualizace se skládá z analýzy minulosti, detekce anomálií a předpovědi budoucnosti. Modrá křivka je zobrazení konkrétních hodnot, které pochází z konkrétního datového souboru. Modré stínování kolem této křivky zobrazuje předpokládané odlehlé hodnoty. Dále je na grafu možné vidět anomálie v podobě barevných koleček. Ty značí, že se v daném místě hodnota odchýlila od předpokládaných hodnot a je vyšší nebo nižší. Poslední žlutá část grafu se týká požadovaného budoucího chování. Žlutá křivka značí predikovanou hodnotu v čase. Žluté stínování označuje rozptyl možných odlehlých hodnot. Čím delší období, tím vzniká větší možný rozptyl dat.

Jako první vizualizace (viz. Obrázek 32) byla provedena predikce pro Bitcoin konkrétně pro atribut „Adj Close“, který byl zvolen z důvodu možných úprav hodnoty Bitcoinu oproti atributu „Close“. Na následujícím obrázku můžeme vidět, že cena by měla v budoucnosti klesnout, ale poté by měla začít opět stoupat. Na konci predikovaného období 22. 2. 2028 by měla být hodnota atributu „Adj Close“ 49,343.419 USD. Horní hranice by se měla pohybovat na hodnotě 80,343.419 a dolní hranice na hodnotě 17,953.3 USD. Tato hodnota byla také nejvyšší vyskytující se v celé předpovědi.



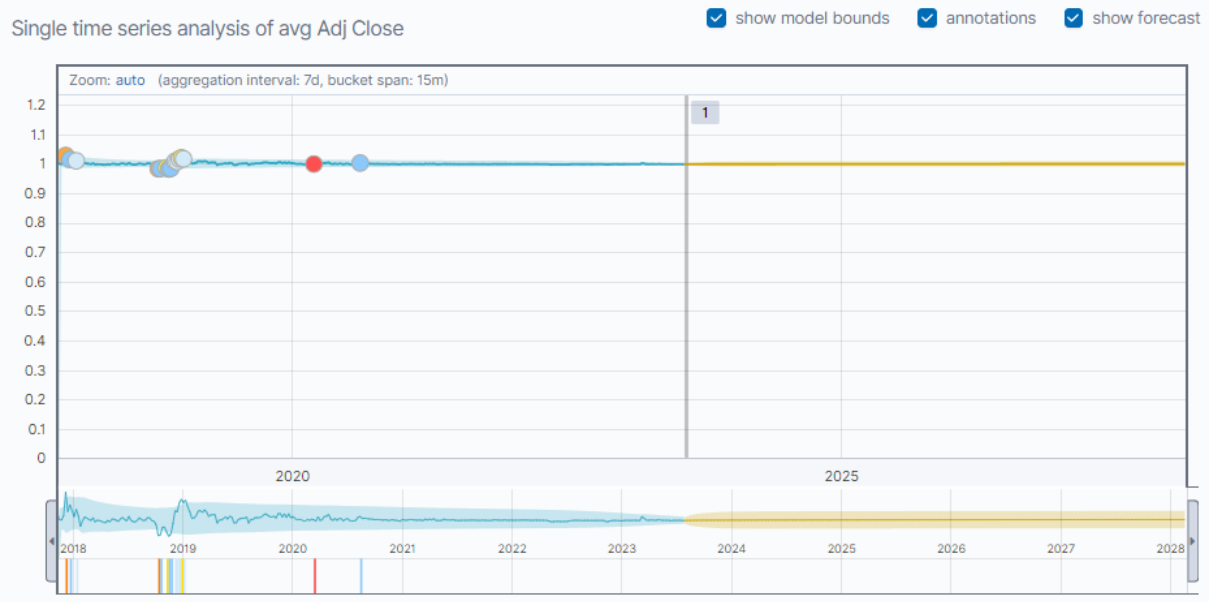
Obrázek 32: Predikce vývoje ceny Bitcoinu

Na následujícím obrázku je vizualizace vývoje kryptoměny Ethereum (viz. Obrázek 33). Oproti předchozí vizualizaci je na první pohled vidět velký rozptyl horní hranice a téměř splývající křivku predikce s hodnotou 0 na svislé ose. Také pro tuto kryptoměnu byl využit atribut „Adj Close“. Na konci predikce 28. 6. 2028 byla hodnota atributu 0 USD. Horní hranice dosahovala hodnoty 420,623.783 USD a spodní hranice byla totožná s predikcí 0 USD. Nejvyšší hodnota kryptoměny Ethereum byla dosažena 1. 2. 2025 v hodnotě 2,311.28 USD. Horní hranice v tento den dosahovala hodnoty 344,736.177 USD a spodní hranice hodnoty 0 USD.



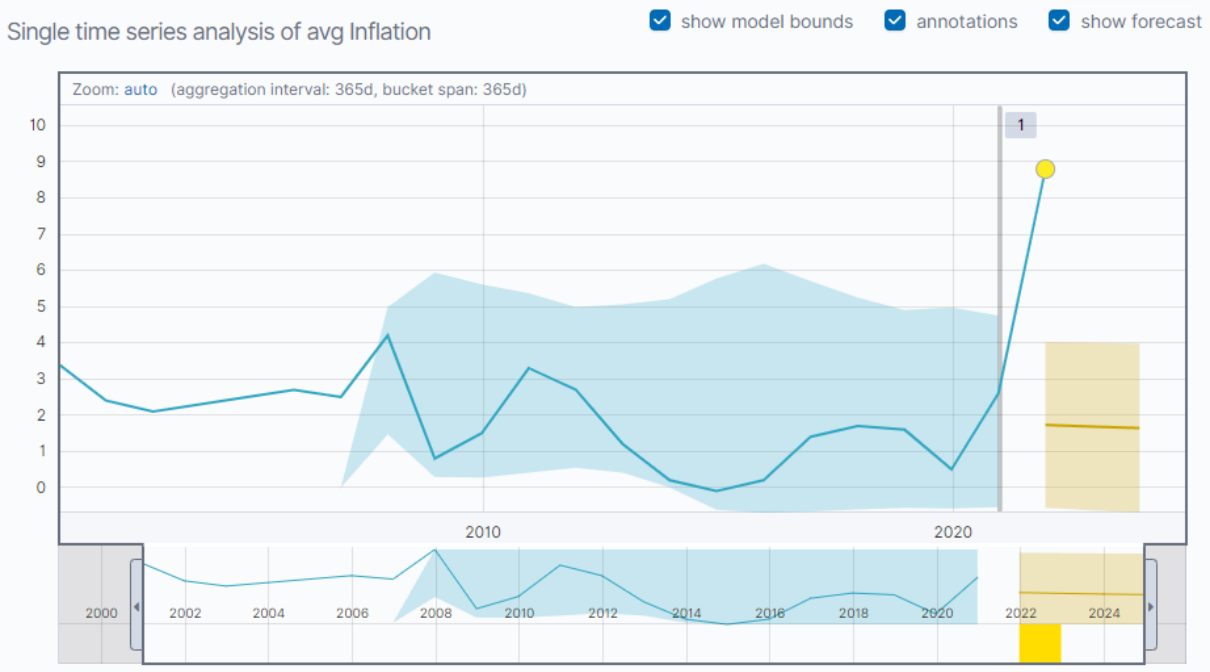
Obrázek 33: Predikce vývoje ceny Etherea

Poslední kryptoměnou k vizualizaci je Tether. Z vizualizace (viz. Obrázek 34) je vidno, že se jedná o téměř rovnou přímku, která leží v úrovní hodnoty 1 na svislé ose. Na konci predikce 21. 2. 2028 byl atribut „Adj Close“ na hodnotě 1.001 USD, horní hranice na hodnotě 1.01 USD a spodní hranice na hodnotě 0.992 USD. Nutno poznamenat, že se tato hodnota ustálila již v půlce roku 2024 a vyšších hodnot již nedosáhla.



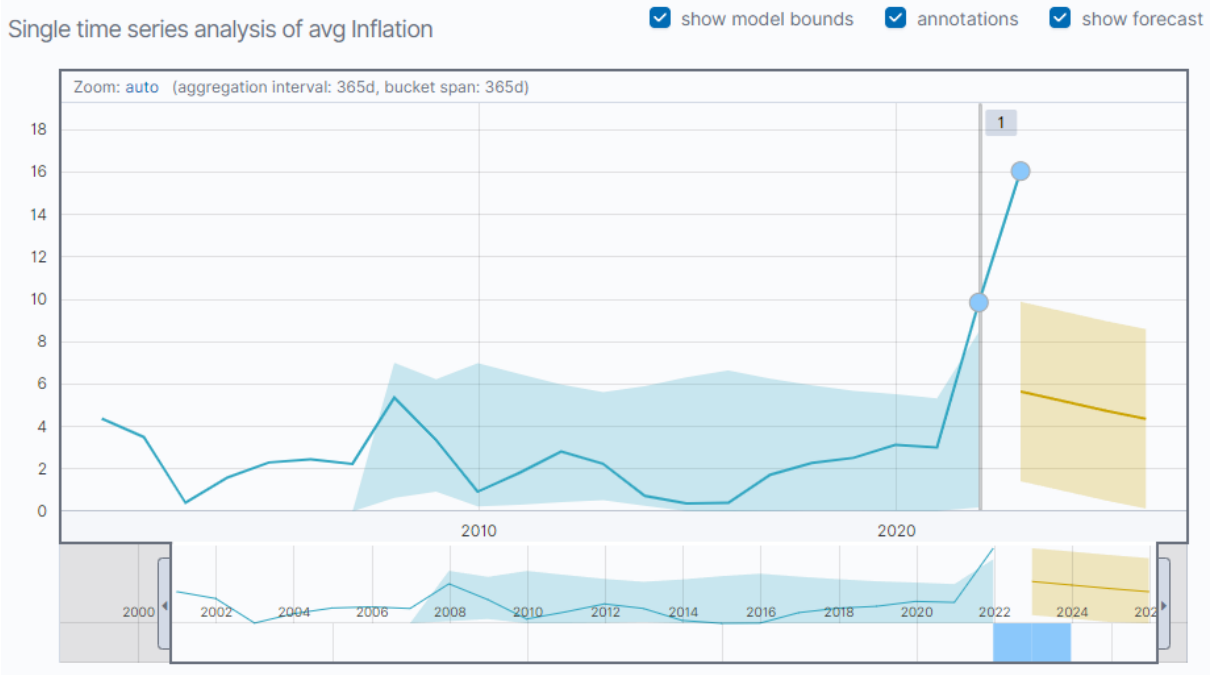
Obrázek 34: Predikce vývoje ceny Tetheru

Další vizualizací je vývoj inflace v Evropské unii (viz. Obrázek 35). V tomto případě není použitý celý rozsah datového souboru, aby bylo možné srovnat predikovaný vývoj inflace pro Českou republiku a Evropskou unii. Vzhledem k tomu, že byl použit poměrně malý vzorek dat, tak nástroj Kibana byl schopen predikovat vývoj inflace pouze do 18. 12. 2024. Predikovaná hodnota inflace na tuto dobu byla 1.611 %, horní hranice 3.959 % a spodní hranice -0.737 %. Předpokládaný budoucí vývoj inflace je predikován na základě historických hodnot a nezohledňují aktuální ekonomické ukazatele, které mají vliv na její hodnotu.



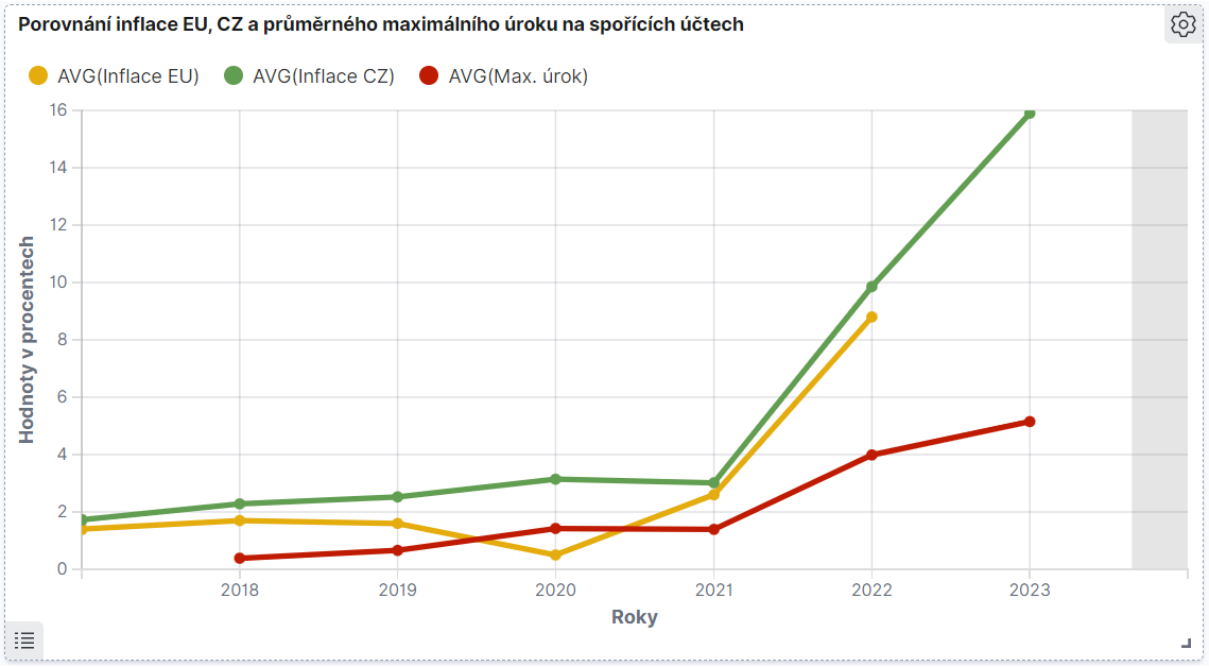
Obrázek 35: Predikce vývoje inflace v Evropské unii

Druhý zkoumaný vývoj inflace se týká budoucí predikce inflace v České republice (viz. Obrázek 36). Také pro tuto predikci byl zvolen poměrně malý datový soubor s tím rozdílem, že datový soubor o inflace v České republice obsahuje měsíční záznamy, které byly agregovaný a použity jako roční, aby odpovídaly stejnému rozptylu jako data o inflaci v EU. Kibana byla schopna provést predikci do 18. 12. 2025 a předpokládaná hodnota inflace má být 4.366 %. Horní hranice potom 8.6 % a spodní hranice 0.132 %. Pro srovnání s inflací v EU má inflace v ČR 18. 12. 2024 hodnotu predikované inflace 4.767 %, horní hranice 9.001 % a spodní hranice 0.533 %. Předpokládaný budoucí vývoj inflace je predikován na základě historických hodnot a nezohledňují aktuální ekonomické ukazatele, které mají vliv na její hodnotu.



Obrázek 36: Predikce vývoje inflace v České republice

Poslední vizualizací (viz. Obrázek 37), která se týká první výzkumné otázky by měla být predikce budoucího vývoje maximálního úroku na spořících účtech. Vzhledem k datům, které byly sesbírány z internetu není možné predikovat budoucí vývoj maximálního úroku na spořících účtech. Jedná se o nespojitá data, dle kterých nelze vytvořit predikci budoucího chování. Namísto toho vizualizace obsahuje srovnání vývoje historických dat inflace v EU, ČR a také nabízeného maximálního úroku na spořících účtech. Co je na vizualizaci zajímavé je, že v roce 2020 nabízely banky v České republice větší průměrný maximální úrok, než byla inflace v Evropské unii.



Obrázek 37: Porovnání historických dat inflace EU, ČR a průměrného maximálního úroku

Součástí předchozí vizualizace je také tabulka (viz. Obrázek 38) obsahující maximální, minimální a průměrnou hodnotu maximálního úroku na spořících účtech pro jednotlivé roky. Vizualizace byly rozděleny z důvodu lepší čitelnosti. Minimální průměrný maximální úrok na spořících účtech, který byl za posledních 5 let v České republice nabízen dosahoval hodnot 0.15 %. Naopak maximální nabízel v roce 2022 a 2023 6.15%

Tabulka min, max, avg úrok

Date per year	Max Maximalni_urok(p_a_)	Min Maximalni_urok(p_a_)	Average Maximalni_urok(p_a_)
2018	0.8	0.15	0.388
2019	1.5	0.2	0.66
2020	5	0.5	1.425
2021	3.08	0.4	1.393
2022	6.15	2.5	3.989
2023	6.15	3.5	5.151

Obrázek 38: Tabulka obsahující srovnání maximálního, minimálního a průměrného úroku

VO2: *Jaké jsou základní statistiky kryptoměn a inflace a jaký vliv má inflace na hodnotu kryptoměn?*

Na následujícím obrázku jsou zobrazeny základní statistiky týkající se kryptoměny Bitcoin (viz. Obrázek 38). Lze vidět minimální, maximální nebo střední hodnoty, ale také první a třetí kvartil. První kvartil znamená, že 75 % dat je větších než první kvartil. Druhý kvartil představuje medián a třetí kvartil znamená, že 75 % hodnot je menších než třetí kvartil. Díky těmto hodnotám si lze představit distribuci a rozložení dat v datovém souboru.

Data pochází od 9. 17. 2014 do 3. 8. 2023. Lze pozorovat, že cena Bitcoinu se pohybovala od 171.5 USD až po maximální hodnotu 68,789.6 USD. Průměrná hodnota Bitcoinu při zavírací ceně je 13,740.6 USD.

Z informací o inflaci můžeme zjistit, že minimální inflace byla na hodnotě 0 %. Průměrná inflace v České republice je 2.908 % a maximální inflace byla 16.4 %.

```
    Date.x                Open              High              Low         
Min.   :2014-09-17   Min.   :  176.9   Min.   :  211.7   Min.   :  171.5  
1st Qu.:2016-12-05   1st Qu.:  768.5   1st Qu.:  774.7   1st Qu.:  758.7  
Median :2019-02-23   Median : 7765.0   Median : 7985.1   Median : 7559.7  
Mean   :2019-02-23   Mean   :13733.0   Mean   :14065.6   Mean   :13369.5  
3rd Qu.:2021-05-14   3rd Qu.:20573.2   3rd Qu.:20993.8   3rd Qu.:20178.4  
Max.   :2023-08-03   Max.   :67549.7   Max.   :68789.6   Max.   :66382.1  
    Close           Adj Close           Volume             Inflation      
Min.   :  178.1   Min.   :  178.1   Min.   :     5914570   Min.   : 0.000  
1st Qu.:  769.7   1st Qu.:  769.7   1st Qu.:   131570000   1st Qu.: 1.400  
Median : 7774.8   Median : 7774.8   Median : 10325509442   Median : 2.400  
Mean   :13740.6   Mean   :13740.6   Mean   : 16545347674   Mean   : 2.908  
3rd Qu.:20595.3   3rd Qu.:20595.3   3rd Qu.: 27326943244   3rd Qu.: 3.200  
Max.   :67566.8   Max.   :67566.8   Max.   :350967941479   Max.   :16.400  
```

Obrázek 39: Deskriptivní statistika pro Bitcoin a inflaci v ČR

Tabulka na dalším obrázku je výsledek korelační analýzy pro Bitcoin (viz. Obrázek 39). Jedná se čistě o textovou verzi. Čísla na korelační matici vyjadřují míru lineárního vztahu mezi dvěma proměnnými. Hodnoty na hlavní diagonále jsou vždy rovny jedné, protože pokud bychom vzali dvě stejné hodnoty, tak budou vždy mít perfektní lineární vztah. Pokud se hodnota blíží k -1, tak to znamená, že se jedná o perfektní negativní lineární korelaci. Což znamená, že pokud jedna proměnná roste, tak druhá klesá. Na druhou stranu, pokud se hodnota blíží k 1, tak to znamená, že obě hodnoty rostou ve stejném poměru.

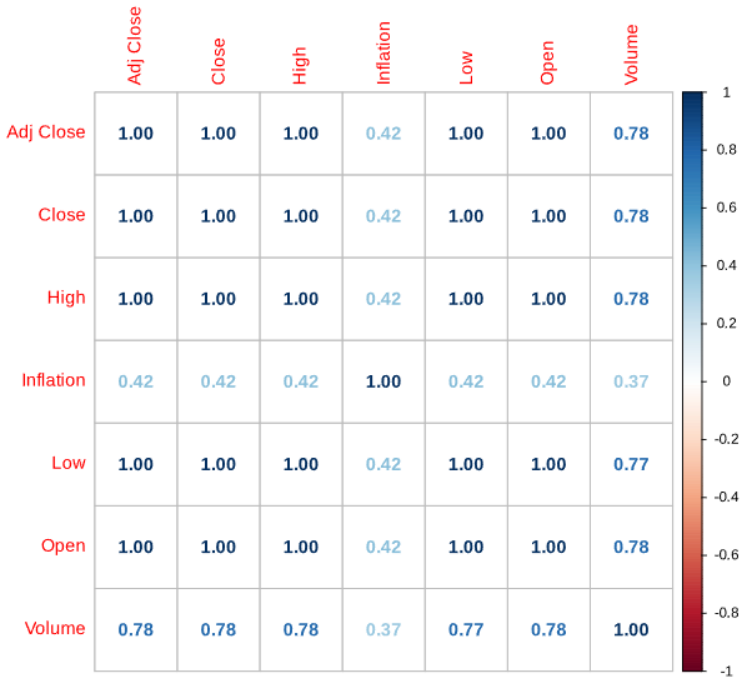
U Bitcoinu můžeme vypozorovat, že obsahuje pouze pozitivní hodnoty mezi atributy. To znamená, že mezi jednotlivými proměnnými je pozitivní lineární korelace a pokud roste jedna proměnná poroste i druhá. Za pozitivní a silnou lineární korelaci se dají považovat hodnoty větší než 0.5. Hodnoty mezi 0.3 až 0.5 mají mírně pozitivní lineární vazbu. Vzhledem k tomu,

že je mezi inflací a cenou mírná pozitivní vazba, tak by se dalo očekávat, že jsou na sebe nějak vázané. Může se jednat například o společný faktor, který tyto hodnoty ovlivňuje.

	Open	High	Low	Close	Adj Close	Volume	Inflation
Open	1.0000000	0.9994308	0.9996825	0.9989540	0.9989540	0.7751868	0.4246964
High	0.9994308	1.0000000	0.9994045	0.9997307	0.9997307	0.7769869	0.4163557
Low	0.9996825	0.9994045	1.0000000	0.9994052	0.9994052	0.7735271	0.4241804
Close	0.9989540	0.9997307	0.9994052	1.0000000	1.0000000	0.7758198	0.4166445
Adj Close	0.9989540	0.9997307	0.9994052	1.0000000	1.0000000	0.7758198	0.4166445
Volume	0.7751868	0.7769869	0.7735271	0.7758198	0.7758198	1.0000000	0.3715743
Inflation	0.4246964	0.4163557	0.4241804	0.4166445	0.4166445	0.3715743	1.0000000

Obrázek 40: Korelační matice pro Bitcoin a inflaci v ČR

Na následujícím obrázku (viz. Obrázek 40), je předchozí korelační matice jako vizualizace. Tento typ vizualizace obsahuje mnoho různých nastavení a je možné si vizualizace nastavit dle svých potřeb. Toto nastavení obsahuje například řazení dle abecedy a metodu pro zobrazení „number“. Pravá část vizualizace obsahuje barvy od modré k červené, což značí sílu korelace od 1 do -1.



Obrázek 41: Vizualizace korelační matice pro Bitcoin a inflaci v ČR

Další obrázek obsahuje tabulku se základními statistickými výpočty kryptoměny Etherea (viz. Obrázek 41). Lze například vidět, že maximální hodnota atributu „High“ byla 4891.70 USD. Minimální zavírací cena byla 84.31 USD a nejvyšší 4812.09 USD.

```
    Date.x                Open             High             Low
Min.   :2017-11-09   Min.   :  84.28   Min.   :  85.34   Min.   :  82.83
1st Qu.:2019-04-16   1st Qu.: 224.56   1st Qu.: 229.19   1st Qu.: 217.28
Median :2020-09-20   Median : 698.15   Median : 729.16   Median : 669.83
Mean   :2020-09-20   Mean   :1192.81   Mean   :1228.65   Mean   :1152.85
3rd Qu.:2022-02-25   3rd Qu.:1842.49   3rd Qu.:1876.38   3rd Qu.:1801.40
Max.   :2023-08-03   Max.   :4810.07   Max.   :4891.70   Max.   :4718.04
    Close           Adj Close           Volume             Inflation
Min.   :  84.31   Min.   :  84.31   Min.   :  621732992   Min.   : 0.000
1st Qu.: 224.48   1st Qu.: 224.48   1st Qu.: 4796651246   1st Qu.: 1.400
Median : 697.95   Median : 697.95   Median : 9723646871   Median : 2.400
Mean   :1193.34   Mean   :1193.34   Mean   :12358432904   Mean   : 2.908
3rd Qu.:1843.53   3rd Qu.:1843.53   3rd Qu.:17157714562   3rd Qu.: 3.200
Max.   :4812.09   Max.   :4812.09   Max.   :84482912776   Max.   :16.400
```

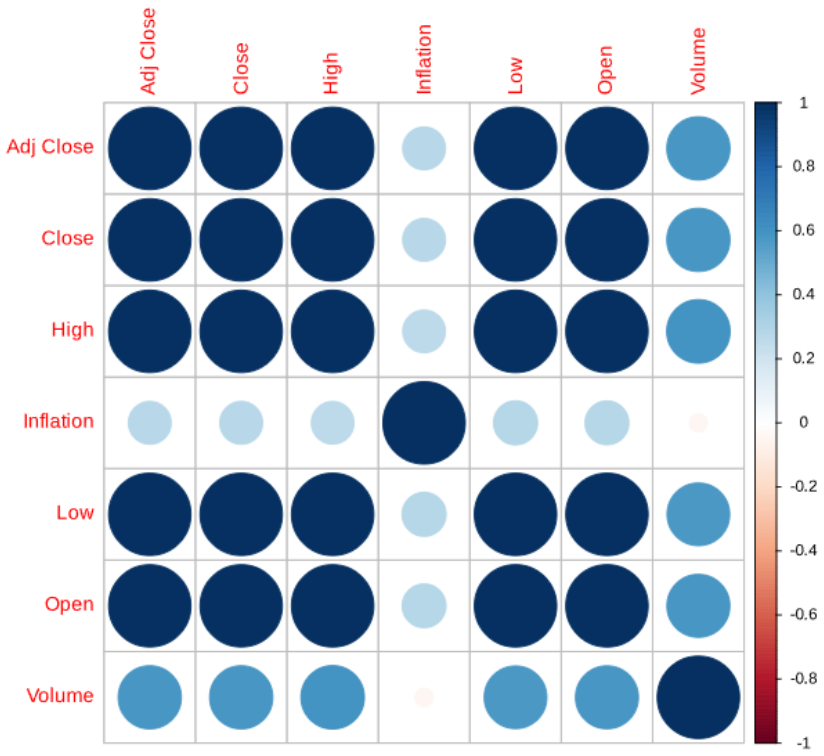
Obrázek 42: Deskriptivní statistika pro Ethereum a inflaci v ČR

Následující tabulka na obrázku obsahuje korelační matici Ethereum a inflace v České republice (viz. Obrázek 42). Je možné pozorovat silnou pozitivní lineární korelaci mezi atributy „Open“, „High“, „Low“, „Close“, „Adj Close“ a „Volume“. Pozitivní vztah se dá vypozorovat také pro atribut „Inflation“ ve spojení s předchozími atributy, krom atributu „Volume“, kde je tato hodnota záporná. Atributy „Volume“ a „Inflation“ mezi sebou velice slabou negativní lineární korelaci, která se blíží spíše k 0, což naznačuje, že mezi těmito atributy není lineární korelace. To samé se dá uvažovat i pro ostatní atributy ve spojení s inflací, poněvadž se zde vyskytují hodnoty menší než 0.3.

```
                Open      High       Low     Close Adj Close     Volume  Inflation
Open       1.0000000 0.9990346 0.9994631 0.9980937 0.9980937  0.5840968  0.2827081
High       0.9990346 1.0000000 0.9990618 0.9995459 0.9995459  0.5936365  0.2679741
Low        0.9994631 0.9990618 1.0000000 0.9988534 0.9988534  0.5786063  0.2854072
Close      0.9980937 0.9995459 0.9988534 1.0000000 1.0000000  0.5884086  0.2702050
Adj Close  0.9980937 0.9995459 0.9988534 1.0000000 1.0000000  0.5884086  0.2702050
Volume     0.5840968 0.5936365 0.5786063 0.5884086 0.5884086  1.0000000 -0.0477545
Inflation  0.2827081 0.2679741 0.2854072 0.2702050 0.2702050 -0.0477545  1.0000000
```

Obrázek 43: Korelační matice pro Etherum a inflaci v ČR

Poslední obrázek týkající se Ethereum je vizualizace korelační matice (viz. Obrázek 43). Pro tuto vizualizace byl zvolen jiný typ zobrazení hodnot než u předchozí vizualizace Bitcoinu. Jedná o základní zobrazení, které využívá pouze abecední řazení. Síla vztahu mezi atributy je dána opět barevných přechodem, ale zároveň velikosti kolečka. Je možné vidět, že atributy „Inflation“ a „Volume“ mají nejmenší kolečko, které má téměř bílou barvu. To znamená, že mezi těmito atributy neexistuje korelace.



Obrázek 44: Vizualizace korelační matice pro Ethereum a inflaci v ČR

Poslední zkoumanou kryptoměnou je Tether. Jak je na následujícím obrázku vidět (viz. Obrázek 44), tak Tether nedosahuje takových hodnot jako předchozí kryptoměny. Pokud se zaměříme na minimální hodnotu atributu „Low“ a na maximální hodnotu atributu „High“ zjistíme, že se liší přibližně o 23 %. Minimální hodnota pro atribut „Close“ činila 0.9666 USD a nejvyšší 1.0779 USD.

```
    Date.x                Open             High             Low        
Min.   :2017-11-09   Min.   :0.9725   Min.   :0.9787   Min.   :0.8995  
1st Qu.:2019-04-16   1st Qu.:1.0000   1st Qu.:1.0006   1st Qu.:0.9951  
Median :2020-09-20   Median :1.0004   Median :1.0024   Median :0.9993  
Mean   :2020-09-20   Mean   :1.0015   Mean   :1.0068   Mean   :0.9968  
3rd Qu.:2022-02-25   3rd Qu.:1.0022   3rd Qu.:1.0106   3rd Qu.:1.0000  
Max.   :2023-08-03   Max.   :1.0810   Max.   :1.1059   Max.   :1.0218  
     Close           Adj Close          Volume             Inflation      
 Min.   :0.9666   Min.   :0.9666   Min.   :   358188000   Min.   : 0.000  
 1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.: 12376436010   1st Qu.: 1.400  
 Median :1.0004   Median :1.0004   Median : 32423781475   Median : 2.400  
 Mean   :1.0015   Mean   :1.0015   Mean   : 39896336338   Mean   : 2.908  
 3rd Qu.:1.0021   3rd Qu.:1.0021   3rd Qu.: 57257598779   3rd Qu.: 3.200  
 Max.   :1.0779   Max.   :1.0779   Max.   :279067455600   Max.   :16.400
```

Obrázek 45: Deskriptivní statistika pro Tether a inflaci v ČR

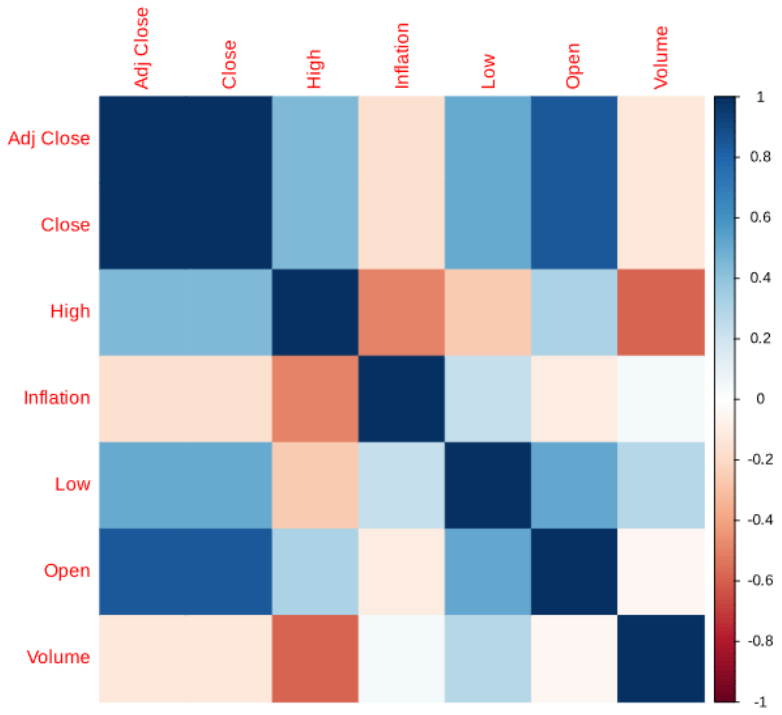
Následující obrázek obsahuje tabulku (viz. Obrázek 45), na které je korelační matice pro Tether. Na první pohled je vidět, že většina atributů má mezi sebou pozitivní lineární korelaci. Nejsilnější poziviní vazba mezi dvoumi rozdílnými atributy je mezi atributy „Open“ a „Close“ respektive také „Adj Close“ a je to hodnota 0.844. Nejsilnější negativní lineární vazba je mezi

atributy „High“ a „Volume“, která se rovná -0.580. Druhá nejsilnější negativní vazba je mezi atributy „High“ a „Inflation“, který činí -0.497. Je možné tedy uvažovat, že pokud by inflace rostla atribut „High“ by klesal. To samé by platilo mezi atributy „High“ a „Volume“.

	Open	High	Low	Close	Adj Close	Volume	Inflation
Open	1.00000000	0.3137125	0.5162677	0.8442093	0.8442093	-0.04724939	-0.10609374
High	0.31371251	1.0000000	-0.2514077	0.4427322	0.4427322	-0.58020014	-0.49772114
Low	0.51626775	-0.2514077	1.0000000	0.5055375	0.5055375	0.28996531	0.23847214
Close	0.84420934	0.4427322	0.5055375	1.0000000	1.0000000	-0.12069041	-0.16481783
Adj Close	0.84420934	0.4427322	0.5055375	1.0000000	1.0000000	-0.12069041	-0.16481783
Volume	-0.04724939	-0.5802001	0.2899653	-0.1206904	-0.1206904	1.00000000	0.04809335
Inflation	-0.10609374	-0.4977211	0.2384721	-0.1648178	-0.1648178	0.04809335	1.00000000

Obrázek 46: Korelační matice pro Tether a inflaci v ČR

Poslední vizualizace zobrazuje další možnost vizualizace korelační matice oproti předchozím, které byly použity pro Bitcoin a Ethereum (viz. Obrázek 46). Je použito nastavení „color“, které místo čísel použije pouze barvy. Vizualizace pak vypadá jako heat map, kdy jsou hodnoty prezentovány pouze pomocí barev. To může pomoci se ve vizualizaci lépe orientovat. Z obrázku je viditelné, že nejsilnější pozitivní korelace je mezi atributy „Adj Close“ a „Close“ a nejsilnější záporná korelace mezi atributy „Volume“ a „High“.

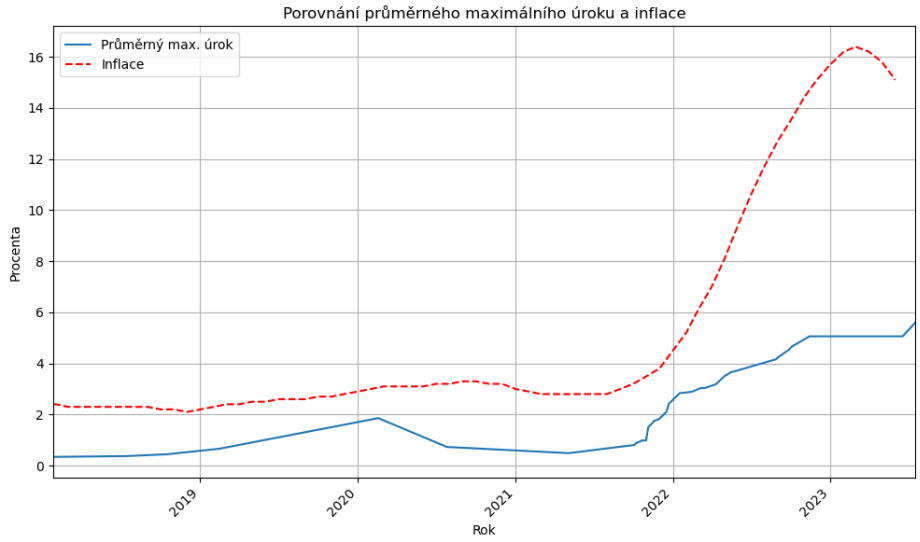


Obrázek 47: Vizualizace korelační matice pro Tether a inflaci v ČR

VO3: *Jaký je rozdíl mezi inflací a nabízenými úroky na spořících účtech za posledních 5 let?*

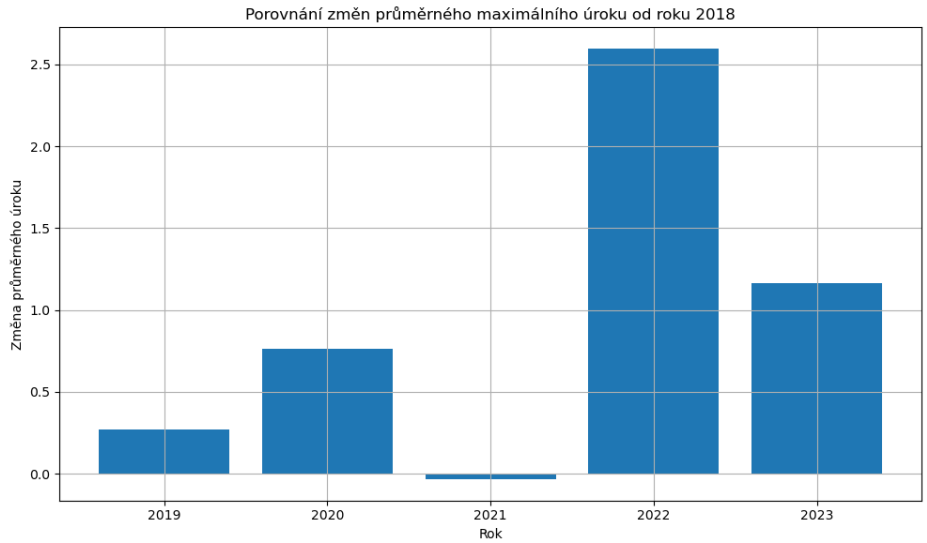
První vizualizace obsahuje liniový graf (viz. Obrázek 47), na kterém je možné vidět srovnání průměrné roční inflace s průměrným maximálním úrokem, který banky nabízejí. Červená

čárkovaná křivka představuje vývoj inflace mezi roky 2018 až 2023. Modrá plná křivka zase nabízený maximální úrok. Na první pohled je možné vysledovat, že nabízený úrok ve sledovaném období byl vždy menší než inflace. Od roku 2022 je možné vypozorovat, že nabízený úrok je mnohonásobně menší než inflace. V předchozích letech je možné vidět, že se křivky částečně kopírovaly s tím rozdílem, že nabízený úrok na spořícím účtu byl přibližně o 2 % méně.



Obrázek 48: Porovnání historického vývoje inflace s průměrným maximálním úrokem

Na následujícím obrázku je porovnání o kolik procent se každý rok změnila průměrná výše maximálního nabízeného úroku na spořících účtech (viz. Obrázek 48) oproti předchozímu roku. Největší skok byl zaznamenám mezi roky 2021 a 2022, kdy nabízený úrok vyrostl o více než 2.5 %. V roce 2021 je možné vidět, že nabízený úrok klesl oproti roku 2020.



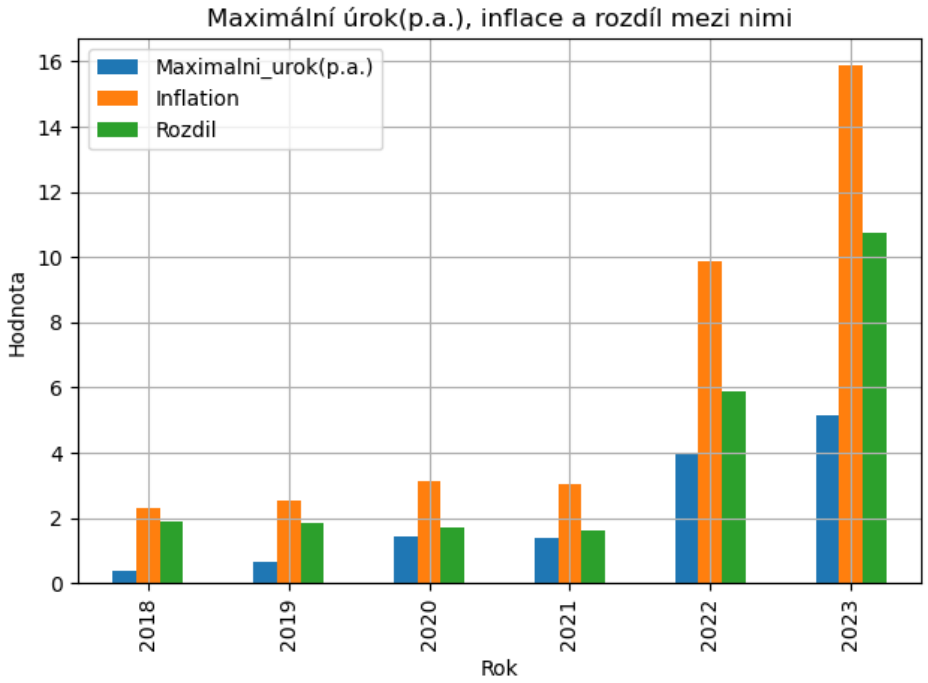
Obrázek 49: Porovnání změn průměrného maximálního úroku

Tabulka na následujícím obrázku obsahuje průměrné hodnoty maximálního nabízeného úroku, průměrnou hodnotu inflace a jejich rozdíl pro každý rok (viz. Obrázek 49). Z tabulky je možné vypozorovat, že v letech 2018-2021 se rozdíl mezi inflací a maximálním úrokem lišil přibližně o 1.6 % až 1.9 %. V roce 20222 tento rozdíl stoupl až na 5.8 % a v roce 2023 téměř dvojnásobně na 10.74 %.

Year	Maximalni_urok(p.a.)	Inflation	Rozdil
2018	0.387600	2.283333	1.895733
2019	0.660000	2.525000	1.865000
2020	1.424615	3.141667	1.717051
2021	1.392846	3.016667	1.623821
2022	3.988534	9.858333	5.869799
2023	5.150615	15.900000	10.749385

Obrázek 50: Porovnání inflace a průměrného maximálního úroku za jednotlivé roky

Následující vizualizace zobrazuje předchozí tabulku jako vizualizaci pomocí grafu typu box plot (viz. Obrázek 50). Modrá barva představuje průměrný maximální úrok, oranžová průměrnou míru inflace a zelená rozdíl mezi nimi.

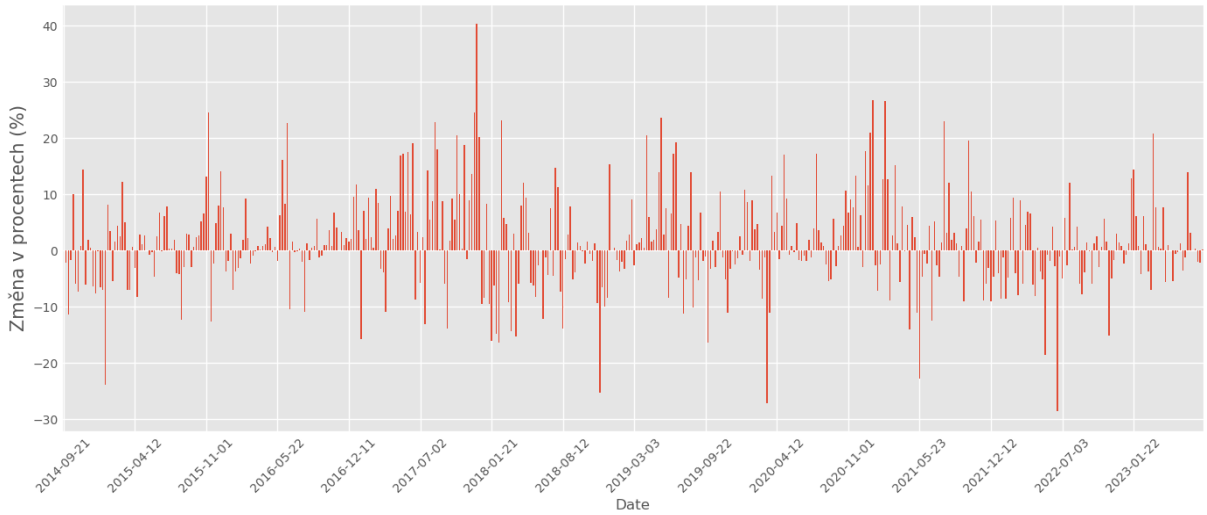


Obrázek 51: Vizualizace porovnání inflace a průměrného maximálního úroku za jednotlivé roky

VO4: *Jak se proměňují historické trendy cen kryptoměn v různých časových obdobích?*

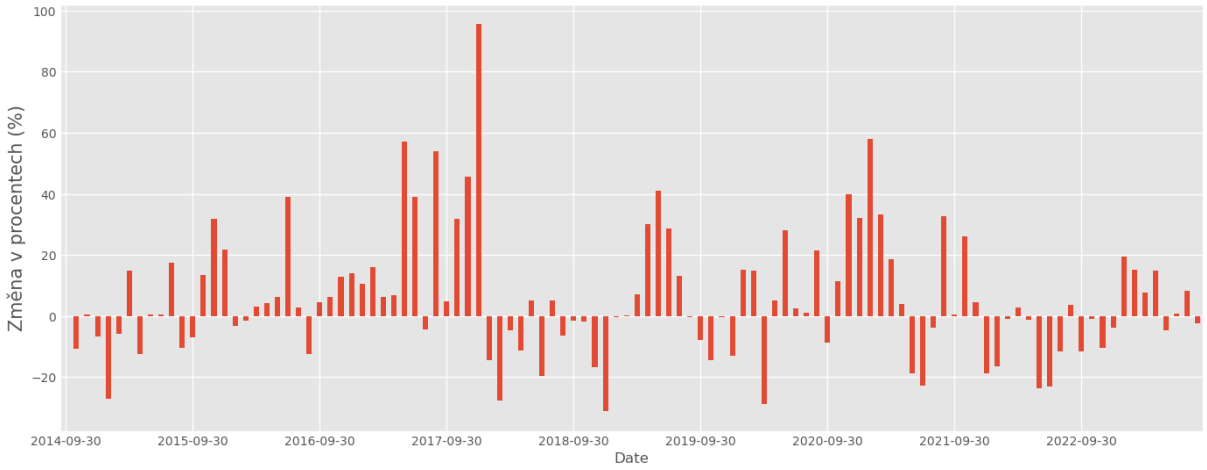
Následující obrázek obsahuje vizualizace týdenních procentuálních změn ceny Bitcoinu (viz. Obrázek 51). Rozdíly týdenních cen se pohybují v rozmezí -30 % až +40 %. Z grafu lze vypozorovat, že u Bitcoinu se objevuje v rozmezí jednotlivých týdnů vysoká volatilita. Největší

kladná změna ceny Bitcoinu mezi jednotlivými týdny byla 40.33 %. Zároveň největší cenový propad byl a to -28.67 %. Průměrná denní změna ceny Bitcoinu je 1.27 %.



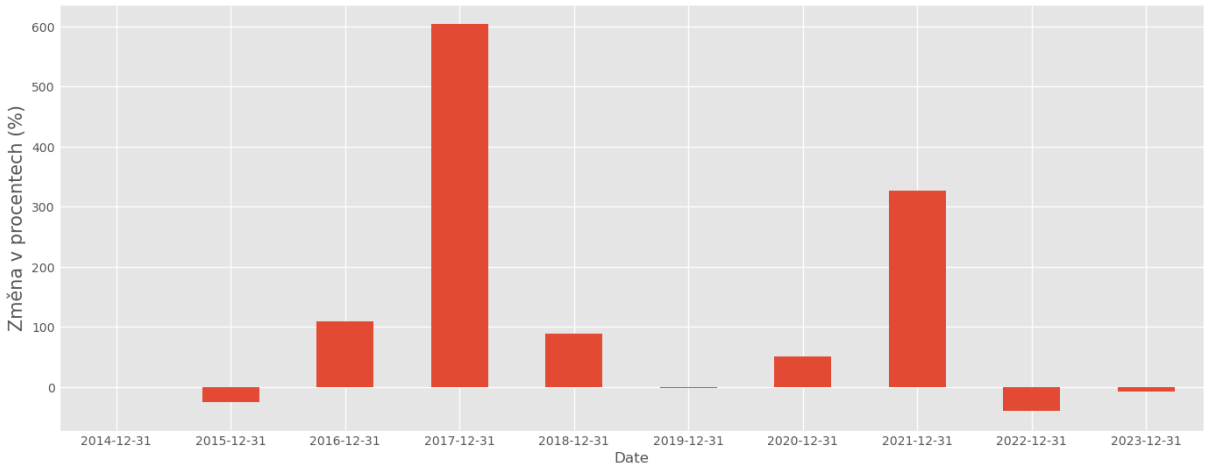
Obrázek 52: Porovnání týdenních změn ceny Bitcoinu

Další zkoumané časové období Bitcoinu bylo provedeno pro jednotlivé měsíce vždy ke konci jednotlivého měsíce (viz. Obrázek 52). Je možné vypozorovat, že se procentuální rozdíly dosahují plusových hodnot častěji než ty záporné. Největší meziměsíční změna byla 95.75 % a největší záporná změna -31.21 %. Průměrná měsíční změna Bitcoinu činila 5.91 %.



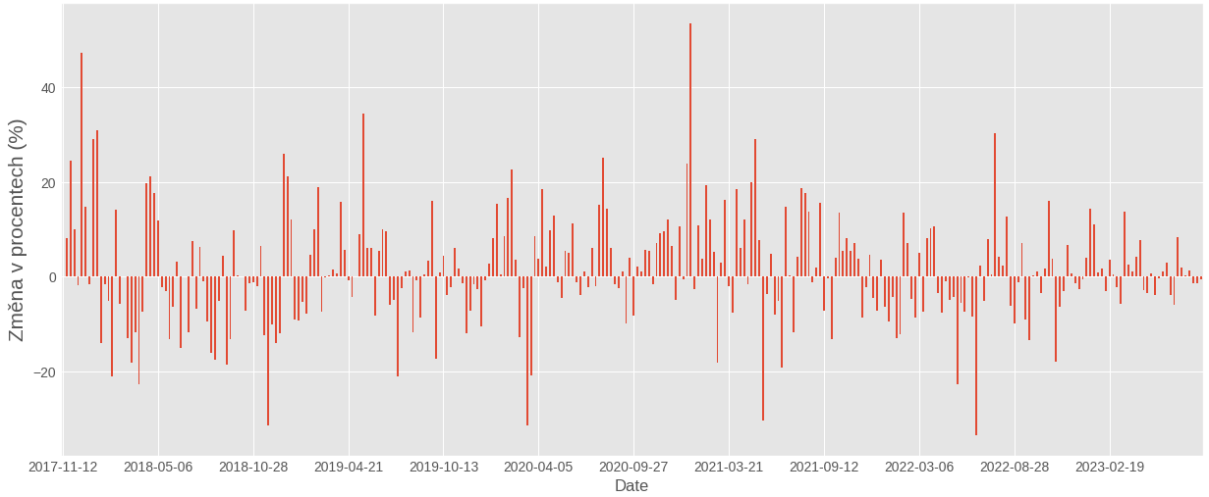
Obrázek 53: Porovnání měsíčních změn ceny Bitcoinu

Poslední zkoumané období se týkalo jednotlivých roků (viz. Obrázek 53). Rozdíly jsou zaznamenány vždy k poslednímu dni v roce. Z vizualizace je možné vypozorovat největší změnu v roce 2017, kdy Bitcoin dosáhl za rok +604.67 % oproti hodnotě na konci roku 2016. Na delším časovém období je možné vidět, že se Bitcoin může těšit spíše z většího růstu oproti rokům, kdy zaznamenal ztrátu. Největší ztráta byla zaznamenána v roce 2022 a to -40.55 %.



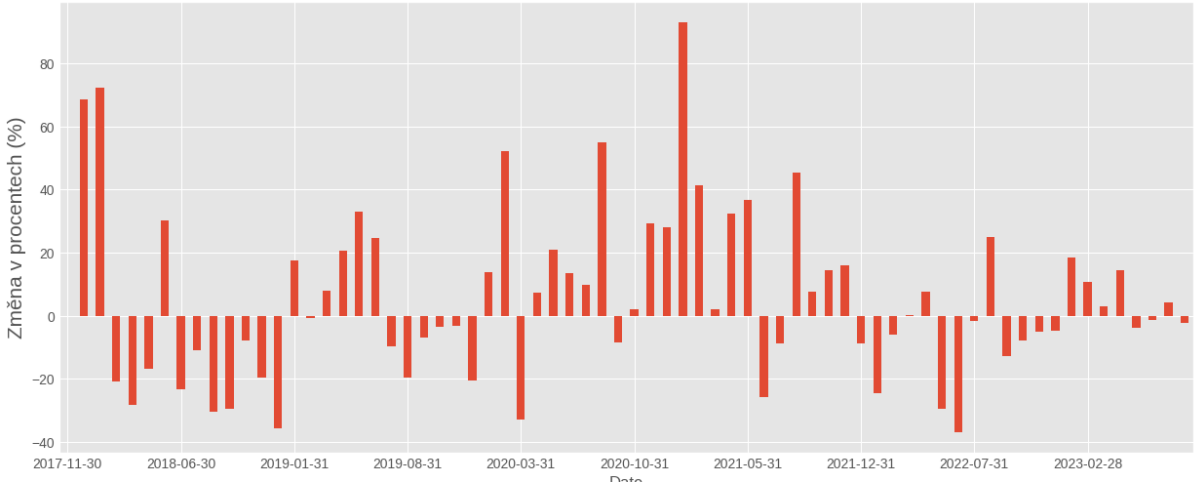
Obrázek 54: Porovnání ročních změn ceny Bitcoinu

Další zkoumanou kryptoměnou je Ethereum. Opět jako u předchozí kryptoměny je možné na vizualizaci (viz. Obrázek 54) vypozorovat poměrně velkou volatilitu během jednotlivých týdnů. Nejvyšší navýšení během týdne bylo o 53.55 %. Největší propad byl zaznamenán o -33.38 %. Průměrná týdenní změna v ceně Etherea byla 1.24 %.



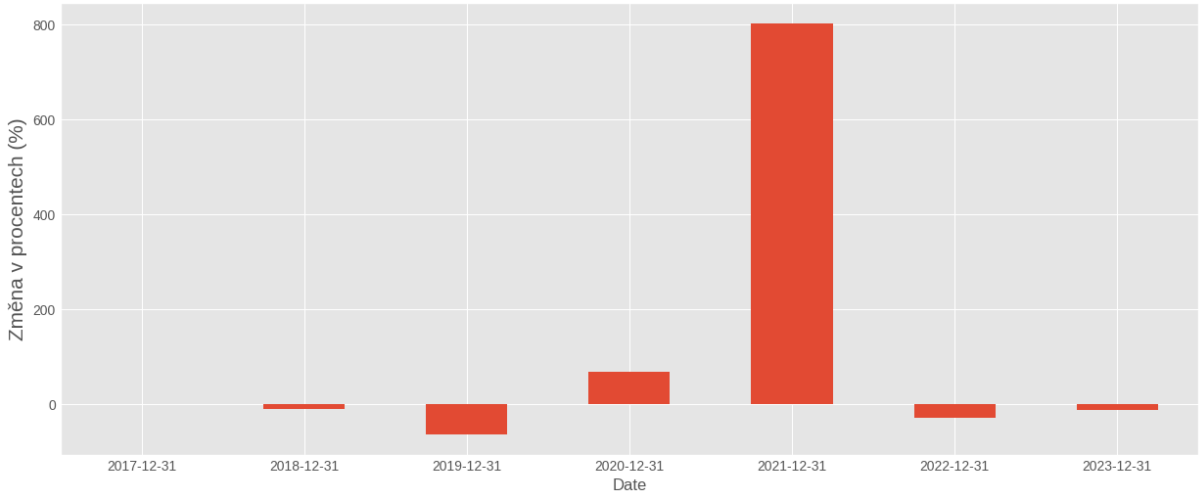
Obrázek 55: Porovnání týdenních změn cen Etherea

Na následujícím snímku je provedena vizualizace pro jednotlivé měsíce (viz. Obrázek 55). Při pohledu na vizualizaci týdenních a měsíčních změn je možné vypozorovat shodné chování. Rozdíl je ale v dosažených změnách. Kdy měsíční změny dosahují až přes +90 %, konkrétně nejvyšší dosažené maximum je 93.17 %. Záporné hodnoty se již tolik neliší oproti týdenním změnám. Maximální měsíční změna je -36.78 %. Průměrná měsíční změna hodnoty Etherea dosahuje 5.40 %.



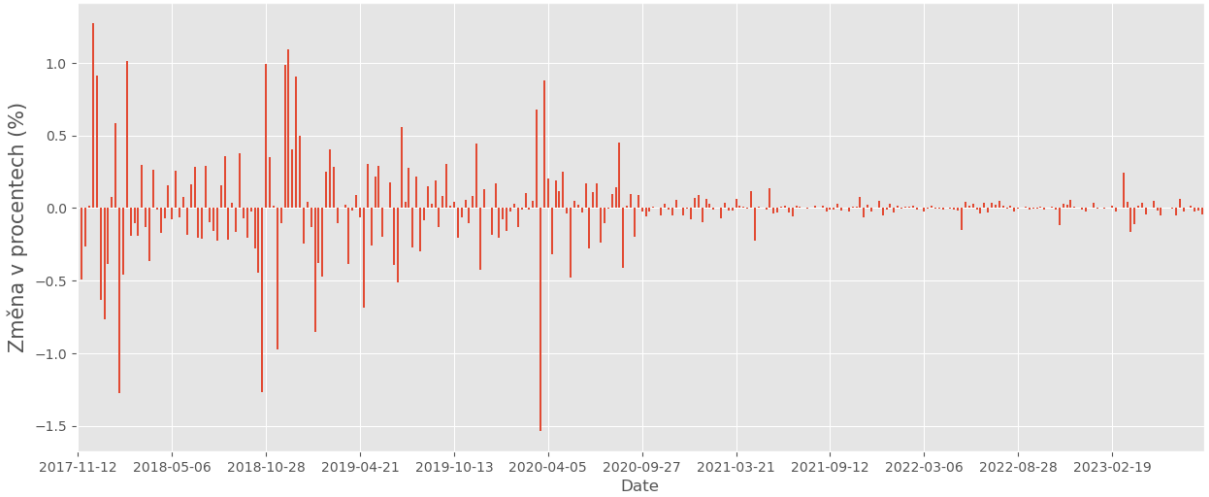
Obrázek 56: Porovnání měsíčních změn ceny Etherea

Při vizualizaci ročních změn Etherea (viz. Obrázek 56), je na první pohled viditelný konec roku 2021, kdy byl meziroční růst kryptoměny 803.40 %. Krom roku 2020 a 2021 všechny ostatní roky zaznamenaly záporný meziroční růst. Avšak v žádném roce, kdy byl růst záporný, tak nikdy nepřesáhla meziroční ztráta 100 %. Největší meziroční ztráta byla v roce 2019 a to -62.40 %. Průměrný meziroční růst Bitcoinu je 126.77 %.



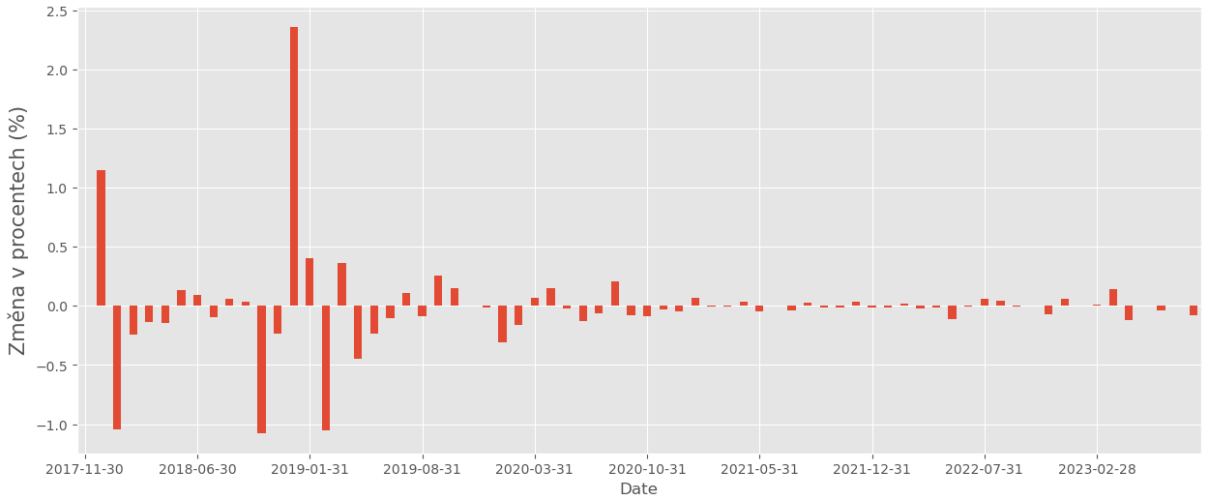
Obrázek 57: Porovnání ročních změn ceny Etherea

Poslední sledovanou kryptoměnou je Tether. Na následující vizualizace (viz. Obrázek 57) je vidět, jak se z počátku vyskytovala vysoká volatilita, ale přibližně v polovině zkoumaného souboru se data ustalují a výkyvy nejsou tak velké. Největší kladný růst ceny byl o 1.27 %. Naopak největší záporný růst ceny byl o -1.53 %. Průměrná měsíční změna ceny byla -0.002 %.



Obrázek 58: Porovnání týdenních změn ceny Tetheru

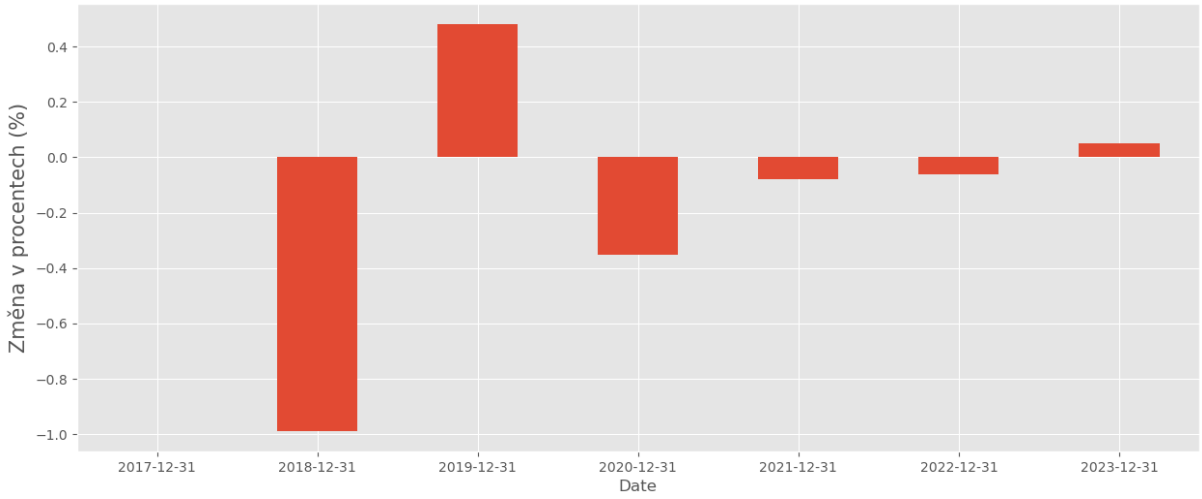
Další vizualizace se opět týká měsíčních procentuálních změn ceny Tetheru (viz. Obrázek 58). Na obrázku je možné vidět, jak v druhé polovině roku 2019 přestává být Tether více volatilní oproti začátku vizualizace. Nejvyšší kladná změny ceny Tetheru byla o 2.36 % a největší propad ceny byl zaznamenán o -1.07 %. Průměrná měsíční změna hodnoty činila -0.004 %.



Obrázek 59: Porovnání měsíčních změn ceny Tetheru

Na poslední vizualizaci týkající se měny Tetheru jsou zobrazeny roční procentuální změny (viz. Obrázek 59). Největší propad měny nastal již koncem druhého zkoumaného roku. Propad byl

o -0.98 % a nejvyšší kladná změna byla koncem příštího roku 2019 o 0.48 %. Roční průměrná změna ceny Tetheru byla -0.15 %.



Obrázek 60: Porovnání ročních změn ceny Tetheru

4.8 Vyhodnocení

VO1: *Jak se budou vyvíjet jednotlivé kryptoměny, inflace a nabízené úroky na spořících účtech v čase?*

Jako první byla provedena budoucí předpověď pro kryptoměnu Bitcoin, konkrétně pro atribut „Adj Close“. Poslední získaná hodnota z dat je 29,016.945313 USD. Predikovaná hodnota podle nástroje Kibana je na 49,343.419 USD. Rozdíl mezi těmito hodnotami je 70.05 %. Dle této predikce by bylo možné tedy odhadnout, že v následujících letech by měla cena Bitcoinu růst. Je však důležité brát v potaz, že horní a spodní hranice predikce byly poměrně rozptýlené.

Další zkoumaná kryptoměna byla Ethereum. V případě této kryptoměny docházelo k extrémně velkému rozptylu horní hranice. Kdy dosahovala rozdílu o více než 22,000 % oproti poslední hodnotě z dat. Predikovaná hodnota na konci simulace byla 0 USD. Největší predikovaná hodnota byla 2,311.28 USD a byla o 26.17 % větší oproti poslední zaznamenané hodnotě. V tomto případě je velké riziko hlavně z možnosti budoucí hodnoty na 0 USD.

Tether oproti předchozím kryptoměnám dosáhl velmi malých výsledků. V průběhu predikce se jeho cena rychle stabilizovala a pokračovala beze změny. Na konci dosahoval hodnoty 1.001 USD, což je o pouhých 0.19 % více než měl na konci zaznamenaného období. Výsledek predikce byl takový, že cena by se měla v budoucnu stabilizovat a neměnit.

Predikce vývoje inflace v Evropské unii nebylo možné simulovat na delší období z důvodu malého datového souboru a kvůli rozsahu datového souboru o inflaci v České republice, aby

jejich srovnání a budoucí predikce byly relevantní. Kibana byla schopna provést předpověď pouze do konce roku 2024. Pro tu dobu byla predikována výše inflace na 1.611 %. Taková výše inflace je poměrně nízká ve srovnání s poslední zaznamenanou ve výši 8.8 %. Je důležité si uvědomit, že inflaci se nedá předpovědět pouze z historických údajů, jelikož na inflaci má vliv vícero faktorů.

Jako další byla zkoumána inflace pro Českou republiku, kde aktuálně dosahovala velmi vysokých hodnot. Kibana byla schopna provést předpověď do konce roku 2025. Výše předpovězené inflace byla stanovena na 4.366 %. V tomto případě se inflace poměrně razantně snížila oproti posledním údajům z datového souboru. Predikovaná inflace na konci roku 2024 zůstává o přibližně 3 % vyšší než v Evropské unii.

Poslední vizualizací pro tuto otázku byla předpověď pro vývoj nabízeného maximálního úroku na spořících účtech. Tu nebylo možné vytvořit na základě nespojitosti dat, která byla nasbírána. Vizualizace byla alespoň nahrazena srovnáním průběhu inflace v Evropské unii a České republice společně s nabízeným průměrným maximálním úrokem za posledních 5 let. Z vizualizace je vidět, že banky nabízejí menší úrok na spořících účtech, než je inflace. Výjimečně v roce 2020 nabízely banky v ČR větší průměrný úrok na spořících účtech, než byla inflace v EU. Součástí vizualizace byla také tabulka, která zobrazuje roční minima, průměry a maxima pro úroky na spořících účtech.

VO2: *Jaké jsou základní statistiky kryptoměn a inflace a jaký vliv má inflace na hodnotu kryptoměn?*

Tato výzkumná otázka si kladla za cíl zjistit deskriptivní statistiky jednotlivých kryptoměn a inflace v České republice a jejich vzájemný vztah. Jako první byla provedena analýza pro kryptoměnu Bitcoin a inflaci. Z výsledků deskriptivní statistiky je možné vidět minimum, maximum, průměr, medián, první a třetí kvartál. Průměrná cena Bitcoinu pro atribut „Adj Close“ je 13,740.6 USD to je rozdíl oproti poslední zaznamenané hodnotě o -52.65 %. Poslední zaznamenaná hodnota inflace se od průměrné liší o 12.19 %. V další části je potřeba vyhodnotit korelační analýzu. U cen Bitcoinu si lze všimnout, že všechny jeho atributy, krom atributu „Volume“ jsou v silné pozitivní korelaci. Atribut „Volume“ nemá tak silnou pozitivní korelaci, ale přesto lze předpokládat, že pokud poroste nebo klesne hodnota ostatních atributů, tak bude růst nebo klesat hodnota atributu „Volume“. Při pohledu na korelaci mezi inflací a Bitcoinem, je možné vypozorovat, také slabší pozitivní lineární korelaci. Je možné předpokládat, že pokud

poroste cena Bitcoinu, tak je možné, že bude také růst inflace v České republice ovšem ne takovým tempem.

Další zkoumaný soubor byl Ethereum s inflací. Z deskriptivní statistiky vychází, že průměrná hodnota atributu „Adj Close“ se oproti poslední zaznamenané hodnotě liší o -34.84 %. Inflace v tomhle případě obsahuje stejné výsledky, jelikož se jedná o statistiku samostatného souboru. V případě korelační analýzy vycházejí vztahy mezi atributy Etherea velice podobně, jako u Bitocinu. Mezi Etherem a inflací je menší kladná pozitivní korelace, která se už více blíží k nule. Dalo by se tedy uvažovat, že výše inflace nemá takový vliv na cenu Etherea. Hlavní rozdíl oproti Bitcoinu je, že se zde vyskytuje záporná lineární korelace mezi atributy „Volume“ a „Inflation“. Ta je ovšem tak malá, že se dá považovat nulová korelace mezi těmito atributy. Celkově je vztah mezi inflací a Ethereem menší a je možné uvažovat, že neexistuje mezi nimi závislost.

Poslední analýza se týkala Tetheru a inflace. V případě deskriptivní statistiky se poslední zaznamenaná hodnota liší oproti průměrné o -0.24 %. Z korelační analýzy vychází, že jednotlivé atributy již nemají mezi sebou tak silnou vazbu. Nejsilnější pozitivní korelace je mezi atributy „Open“, „Close“ a „Adj Close“. Mezi inflací a atributy Tetheru je většina atributů v záporném vztahu. Za zmínku stojí hlavně vztah mezi atributem „High“ a „Inflation“, kde nabývá hodnoty -0.49. Pokud by hodnota inflace rostla, tak by měla hodnota atributu „High“ klesat. Ostatní atributy mají velice slabou lineární závislost a závislost mezi nimi by neměla existovat.

VO3: *Jaký je rozdíl mezi inflací a nabízenými úroky na spořících účtech za posledních 5 let?*

Další analýza se týkala maximálního nabízeného úroku na spořících účtech a inflace v České republice. Z první vizualizace je dobře vidět, že banky vždy nabízejí úrok pod aktuální inflací. V posledních dvou letech je navíc nabízený úrok vysoko pod hodnotou inflace. Z další analýzy bylo zjištěno, že v roce 2022 reagovaly banky na vysokou inflaci navýšením průměrného maximálního úroku o 2.5 %. V roce 2023 se růst zpomalil a úroky stouply o necelé 1.5 %. Rozdíl mezi inflací a průměrným nabízeným úrokem se v roce 2022 vyšplhal na 5.86 % a v roce 2023 na 10.74 %.

VO4: *Jak se proměňují historické trendy cen kryptoměn v různých časových obdobích?*

Jako první bylo provedeno zkoumání historických změn ceny Bitcoinu během jednotlivých týdnů. Na vizualizaci těchto změn je dobře vidět, že ceny se během týdne hojně mění. Cena

Bitcoinu během týdne se v historii měnila až o 40 %. Celkově je z vizualizace týdenních změn vidět vysoká volatilita ceny Bitcoinu. Při pohledu na měsíční změny ceny Bitcoinu je možné vidět, že se poměrně shodují s týdenními změnami. Změny v tomto případě, ale dosahují větších změn. Nejvyšší dosažená měsíční změna byla 95 %. Poslední zkoumané historické období bylo pro jednotlivé roky. Zde je z grafu vidět, že cena Bitcoinu měla největší meziroční růst v roce 2017 a 2021.

Další zkoumanou kryptoměnou bylo Ethereum. Tato kryptoměna vykazovala při týdenních změnách ceny také vysokou volatilitu, kdy dosahovala zm ěn až o 53 %. Měsíční změny se opět velice podobaly těm týdenním s tím rozdílem, že opět změny dosahovaly vyšších hodnot. Největší změna hodnoty byla o 93 % za jeden měsíc. Roční změny ceny se pohybovaly častěji v záporných hodnotách než v kladných. Největší změna v meziroční ceně Etherea byla v roce 2021 o 803 %.

Poslední zkoumanou kryptoměnou byl Tether. Ten na týdenní vizualizaci změn v ceně dosahuje od poloviny vizualizace velmi malé rozptyly. Celkově Tether dosahoval velice nízkých procentuálních změn, kdy největší procentuální změna během týdne byla -1.53 %. To samé platí také pro měsíční změny, kdy se změny v ceně začaly ustalovat již po roce 2019. Největší změna v ceně dosahovala 2.36 % za jediný měsíc. Na poslední vizualizace jsou opět meziroční změny v ceně Tetheru. Největší změna nastala již ve druhém sledovaném roce, kdy byla meziroční změna -0.98 %.

4.9 Zhodnocení

Úvod praktické části byl věnován problému vysoké inflace a snaze ochránit před ní osobní finance. Ze zkoumaných výsledků v praktické části by se dalo uvažovat, že ze zvolených kryptoměn by byla nejlepší investice kryptoměna Bitcoin. Pro tuto kryptoměnu byla předpovězena nejvyšší cena předpovědi pro následujících 1825 dnů. Na základě historických dat a předpokladu podobného vývoje ceny Bitcoinu i v dalších letech by mohly být vhodné i krátkodobé investice v rámci týdnů nebo měsíců.

Pokud by se jednalo o krátkodobé investice, bylo by možné uvažovat také o kryptoměně Ethereum, která by dle predikce v příštích pár letech nemusela dosahovat nulové hodnoty, ale při predikci ceny v celkové délce 1825 dnů byla hodnota Etherea 0 USD. Nutno podotknout, že predikce horní hranice ceny Etherea by mohla být dobrou investiční příležitostí, kdyby se predikce naplnila.

Poslední kryptoměna Tether by dle zkoumaných skutečností nejspíše nebyla vhodnou investicí pro ochránění osobních financí. V případě, že by česká koruna slábla oproti americkému dolaru, tak by bylo možné pokusit se ochránit na krátkou dobu své finance směněním peněz do kryptoměny Tether.

Poslední možností pro ochranu financí před inflací bylo zkoumání spořících účtů. Zde nebylo možné predikovat budoucí vývoj, ale dle dat za posledních 5 let nejsou plně schopny spořící účty ochránit finance před inflací. Je také potřeba brát v potaz, že jednotlivé produkty mají své omezení, které je potřeba brát pří výběru spořícího účtu v potaz, protože může dojít k úpravě úroku během spoření. Největší jejich výhodou je však to, že se u nich nevyskytuje volatilita, jako u kryptoměn a jsou zde celkově menší rizika s investicemi.

Určitě by nebylo doporučeno se rozhodovat, jak investovat dle výsledků, které byly zjištěny během analýzy. Bylo by potřeba provést další analýzy, které by obsahovaly mnohem více atributů, které mají vliv na jednotlivé ceny, popřípadě výši inflace. Pro predikci ceny by bylo možné dále sestavit vlastní modely, které by mohly být přesnější než model používaný nástrojem Kibana pro predikci. Dále by bylo možné provést výzkum jednotlivých kryptoměn a pokusit se najít tu, která se hodí pro investice nejvíce, oproti zvoleným nejvíce zastoupeným na trhu. Kryptoměna Etherum totiž slouží nejenom jako investiční aktivum, ale díky svým vlastnostem může sloužit pro více účelů. Třetí použitá kryptoměna Tether má zase sloužit jako ekvivalent pro skutečný americký dolar a nedá se tedy předpokládat velké zhodnocení.

5 SHRNUTÍ METODIKY PRAKTICKÉ ČÁSTI

Praktická část se inspirovala popisovanými metodikami a procesy uvedené v kapitole dvě, kdy po celou dobu tvorby praktické části, bylo přihlíženo k Data Science Maturity Model (DSMM), Data Life Cycle, Data Life Cycle Management (DLM) a Cross-Industry Standard Process for Data Mining (CRISP-DM). Na jejich základě byla vypracována celá praktická část. V některých případech se jednotlivé metodiky prolínaly mezi sebou. Na následujícím obrázku (viz. Obrázek 60) je zpracováno, která část byla použita nebo která odpovídá konkrétní kapitole z praktické části. Obrázek je rozdělen do čtyř sloupců, kdy v prvním sloupci je obsah z praktické části, ve druhém sloupci jsou úlohy z modelu CRISP-DM, ve třetím se vyskytují kategorie z Data Science Maturity Model dle Oracle a v posledním sloupci jsou úlohy z Data Life Cycle Managementu.

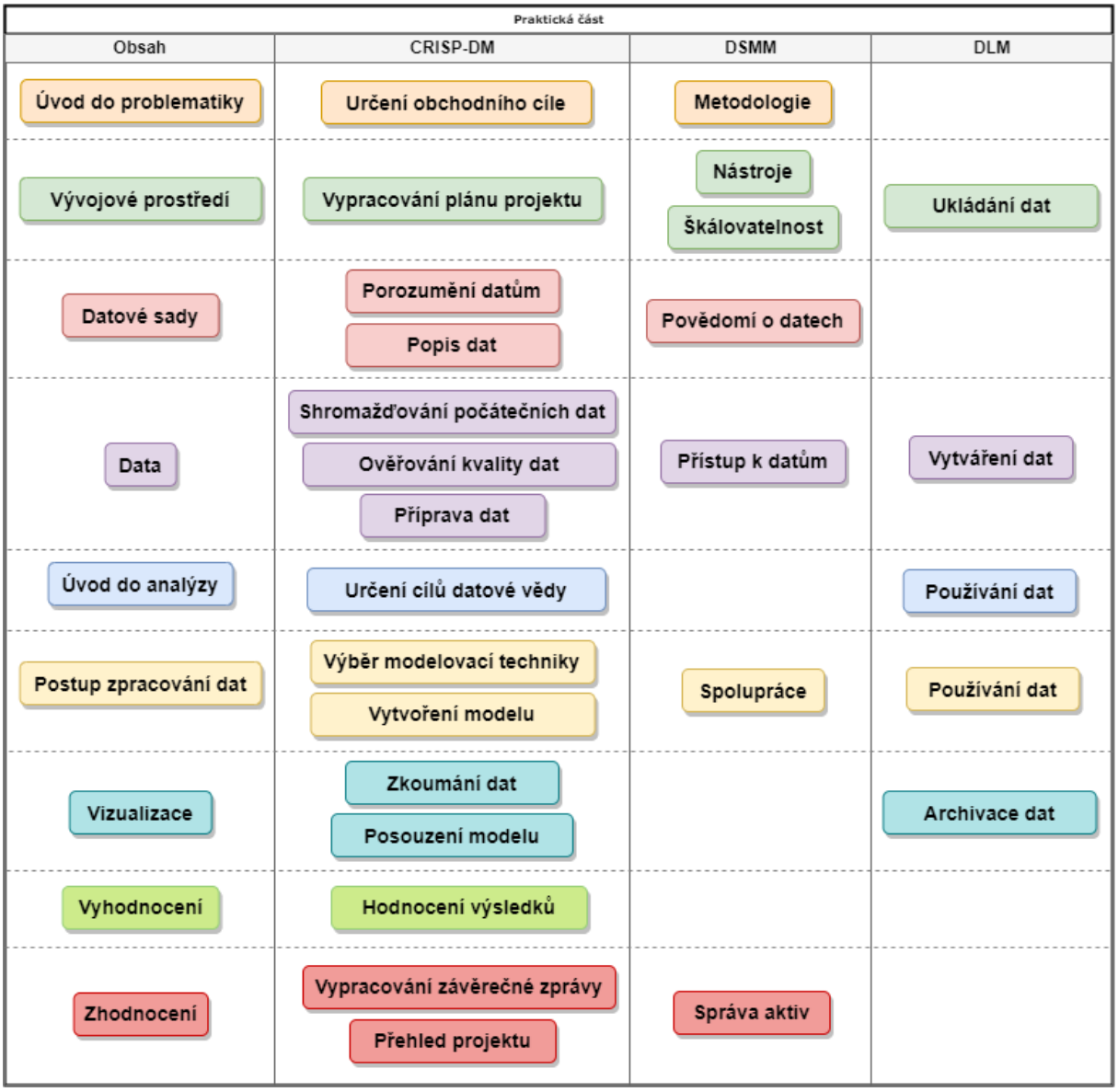
První a nejčastěji prolínající model s praktickou částí je CRISP-DM. Kapitola „Úvod do problematiky“ byla vytvořena na základě rámce určení obchodního cíle z „Business understanding“. Zde je popisován problém a co je cílem praktické části. Další kapitola „Vývojové prostředí“ odpovídá časti rámce vypracování plánu projektu, kde se vybírají použité nástroje pro práci. Kapitola „Datové sady“ je v souladu s rámcem popis dat, který má za úkol sestrojit zprávu s informacemi o datech. Další kapitola „Data“ je vytvořena ve shodě s větším počtem rámců. Jako první rámec je na obrázku (viz. Obrázek 60) shromažďování počátečních dat. Tomuto kroku odpovídá podkapitola „Sběr dat, ve které jsou pospány metody pro získání dat. Další rámec je ověřování kvality dat, na který reaguje podkapitola „Popis zdrojů dat“. Zde jsou popsány zdroje, ze kterých data pochází. Poslední rámec, který se stále týká kapitoly „Data“ je příprava dat. Příprava dat je třetí fází v modelu CRISP-DM a obsahuje rámce jako výběr dat, čištění dat, vytvoření dat, integrace dat a formátování dat. Výběr dat odpovídá podkapitole „příprava dat“ ve které je psáno o důležitosti jednotlivých atributů. Podkapitola „Očištění a transformace“ souvisí s rámci čištění dat a transformace, kde jsou prováděny úpravy dat získaných z internetu. Kapitola „Úvod do analýzy“ byla vytvořena na základě určení cílů datové vědy. V této kapitole jsou určeny cíle projektu na základě stanovených výzkumných otázek. „Postup zpracování dat“ souvisí s výběrem modelovací techniky a vytvoření modelu. Zde se nachází postup, který byl použit pro provedení jednotlivých analýz. Nejedná se přímo o popis výběru modelovací techniky, ale již o konkrétní popis postupu při realizaci analýz. Vytvoření modelu je, že za pomoci již vybraných nástrojů jsou provedeny konkrétní výstupy. Další kapitolou jsou „Vizualizace“ v té jsou již konkrétní výstupy, které vznikly na základě výzkumných otázek. Sem je možné zařadit rámec zkoumání dat, které má odhalit vztahy mezi

atributy. Tomu vyhovuje například korelační tabulka. Další rámec, který je v souladu této kapitoly posouzení modelu. Zde dochází již k ohodnocení modelu dle odborných znalostí v tomto případě k posouzení jednotlivých vizualizací. Kapitola „Vyhodnocení“ souvisí s rámcem hodnocení výsledků, kde dochází k vyhodnocení získaných výsledků na základě vizualizací. Pokud by výsledky nebyly vhodné, muselo by dojít k navrácení do předchozích částí. Poslední kapitola „Zhodnocení“ nejvíce odpovídá dvěma rámcům, a to vypracování závěrečné zprávy a přehled projektu. Jako závěrečná zpráva může sloužit celá praktická část, která obsahuje popis jednotlivých vstupů, výstupů nebo procesů. Přehled projektu by odpovídal kapitole textu v kapitole „Zhodnocení“, která zhodnocuje praktickou část.

Rámce, které se prolínají s praktickou částí a DLM jsou vytváření dat, ukládání dat, používání dat a archivace dat. Konkrétně ukládání dat je součástí kaptioly „Vývojové prostředí“, kdy jsou data uložena buď v databázi Elasticsearch nebo pro kontejner jupyter lokálně na zařízení kam je možnost přistoupit. Vytváření dat je pak směřováno do kapitoly „Data“ ve které je popsáno, jak probíhal sběr a vytváření dat. Používání dat je možné zařadit do kapitoly „Úvod do analýzy“. Zde jsou definovány výzkumné otázky na základě dat, které jsou k dispozici a nelze vytvářet otázky, pokud není známo, jaké atributy data obsahují. Ten samý rámec je možné zařadit do kapitoly „Postup zpracování dat“, protože tento rámec již odpovídá postupu, který byl použit pro vytváření vizualizací a zodpovězení výzkumných otázek. Rámec archivace dat se dá zařadit do kapitoly „Vizualizace“, protože již jsou analýzy a vizualizace provedeny a uložená data nejsou potřeba a je možné je archivovat.

Pro zjištění vyspělosti Data Science praktické části by bylo možné si v jednotlivých částech práce položit některé otázky z tabulky (viz. Tabulka 1) týkající se DSMM, které by mohly vést ke zlepšení práce. Pro kapitolu „Úvod do problematiky“ by se mohlo jednat o otázku zaměřující se na rámec metodologie. Práce by se mohla vyskytovat mezi úrovněmi 3 až 4, protože využívá osvědčené metodologie datových věd. Další kapitola „Vývojové prostředí by mohla spadat do rámců nástroje a škálovatelnost. Nástroje je možné zařadit do úrovně 3, vzhledem k tomu, že jsou využity ověřené nástroje, které se běžně používají pro práci s datovou vědou. Pokud jde o škálovatelnost, tady by bylo vhodné zvolit úroveň 1, protože praktická část byla prováděna na desktopovém počítači, který má omezený hardware a software a analýzy byly prováděny jednotlivcem. Kapitola „Datové sady“ odpovídá rámci povědomí o datech. Tento rámec lze ohodnotit úrovní 3, protože datové zdroje jsou popsány a ohodnoceny z hlediska kvality. Další rámec přístup k datům odpovídá kapitole „Data“, které by bylo vhodné přiřadit úroveň 1 z důvodu, že jsou data uložena v souborech typu csv. Další kapitola, které by bylo možné

přiřadit některý rámec z DSMM je kapitola „Postup zpracování dat“. Pro tuto kapitolu by byl přiřazen rámec spolupráce. Zde by byla práce ohodnocena úrovní 1, protože práci zpracovává pouze jedna osoba a všechny výsledky jsou ukládány lokálně. Poslední rámec k přiřazení je správa aktiv, která by mohla odpovídat kapitole „Zhodnocení“. Tento rámec by se mohl vyskytovat v různých kapitolách. Do poslední kapitoly byl zařazen z toho důvodu, že otázka k tomuto rámci se dotazuje, jak jsou spravovány a kontrolovány prostředky datové vědy. Tomu odpovídá úroveň 1, která říká, že analytické pracovaní produkty jsou vlastněny, organizovány a udržovány jednotlivými členy týmu datové vědy.



Obrázek 61: Metodiky použité v praktické části

ZÁVĚR

Diplomová práce se zabývala představení pojmu Data Science a jeho využití v praxi. Hlavním cíle práce bylo seznámit čtenáře s pojmem Data Science a jeho využití v praxi s možnými prostředky, které formalizují práci datového vědce. Dále si práce kladla za cíl představení možných vizualizací, které se používají pro demonstraci analýz, které se provádí při práci v tomto oboru. Cílem práce bylo také představit jednotlivé metodiky, které se běžně používají v podnicích, kde se provádí datová věda. Tyto metodiky měly být dodržovány v praktické části, která si kladla za cíl provést analýzu totožných historických dat a jejich následnou vizualizací ve vybraných technologiích.

Práce byla rozdělena na teoretickou a praktickou část. Kdy teoretická část obsahovala již zmiňované představení pojmu Data Science a jeho historii. Dále byly v této části představeny možné využití Data Science. Velká část práce se věnovala představení možných typů vizualizací, kde byly popsány grafy jako koláčový graf, sloupcový graf, krabicový graf a další. Společně s popisem vizualizací byly popsány krátce jejich výhody a nevýhody. V další části se práce zabývala metodiky jako Data Science Maturity Model, Data Life Cycle Management a CRISP-DM. Tyto metodiky byly popsány tak, aby jejich popis byl co nejsrozumitelnější s cílem rychlého pochopení použití v případě potřeb. V poslední části teoretické práce byly popsány jednotlivé nástroje, které často využívají datový vědci při své práci. V práci byly popsány nástroje Jupyter, Apache Zeppelin, Python, R, Scala, MATLAB, Julia a Elastic Stack. Pro jednotlivé nástroje byly vybrány knihovny nebo nástroje, které se nejčastěji používají při práci s datovou vědou.

Zvolená problematika v praktické části se zabývala problémem vysoké inflace a ochranou osobních financí. Pro tuto problematiku byly zvoleny jako možné ochranné prostředky před inflací tři nejpoužívanější kryptoměny na trhu Bitcoin, Ethereum a Tether společně se spořícími účty v České republice. Z internetu byla sesbírána příslušná data, která odpovídají zvoleným prostředkům. Byly provedeny přípravy dat, očištění a transformace, aby byla data vhodná k použití jednotlivými nástroji. V následujících částí byly stanoveny čtyři výzkumné otázky, které se snažili odpovědět na definovaný problém, byl popsán stručný popis zpracování dat, aby mohly vzniknout vizualizace, které pomohou se zodpovězením výzkumných otázek a prezentací výsledků. V poslední části praktického výstupu práce je celkové vyhodnocení, jak pro jednotlivé výzkumné otázky, tak celkové shrnutí a doporučení.

Diplomová práce může sloužit jako úvod pro začátečníky v oboru datové vědy, ale i pro zkušenější čtenáře, kteří mohou hlavně využít jednotlivých metodik pro zlepšení své práce nebo práce v rámci celého podniku. Čtenář by měl získat především pochopení, co to Data Science je, jak tento obor vznikl a jakým směrem se může v budoucnu vyvíjet. Dále by měl být schopen zvolit vhodné vizualizace pro specifické analýzy, aby publikum dobře pochopilo, co má daná vizualizace sdělovat. Měl by také získat přehled o nástrojích, které se dají použít pro datovou vědu a o konkrétních knihovnách, které jsou specifické pro řešení určitých úloh. Díky praktické části by měl být schopen pochopit některé kroky, které odpovídají použitým metodikám a následně je schopen replikovat při svém výzkumu. Data Science a vizualizace dat jsou velice rozsáhlá téma, kterým se dá věnovat mnohem podrobněji. Tato práce se snažila představit, alespoň základní charakteristiky všech důležitých částí datové vědy pro celkové pochopení. Práce by mohla být rozšířena o zvolení konkrétního nástroje či metodiky, které by byly použity pro řešení jiných problémů. Dále by bylo vhodné takovou práci rozšířit o více členů, kteří by měli vědomosti nejenom z oboru informačních technologií, ale byli by odborníci i ve zkoumaných oborech, jako jen například ekonomie či produkty finančního trhu.

POUŽITÁ LITERATURA

ALL THINGS STATISTICS, 2022. Advantages & Disadvantages of Pictographs. *All Things Statistics* [online]. [cit. 2023-08-21]. Dostupné z: https://allthingsstatistics.com/miscellaneous/advantages-disadvantages-pictographs/

APACHE ZEPPELIN, ed., nedatováno. WHAT IS APACHE ZEPPELIN. *Apache Zeppelin* [online]. [cit. 2023-08-16]. Dostupné z: https://zeppelin.apache.org/

APACHE ZEPPELIN, ed., nedatováno. Spark interpreter for Apache Zeppelin. *Zeppelin Apache* [online]. [cit. 2023-08-18]. Dostupné z: https://zeppelin.apache.org/docs/latest/interpreter/spark.html

ARCALEA, ed., © 2023. *Data Science Maturity Models* [online]. [cit. 2023-08-08]. Dostupné z: https://www.arcalea.com/blog/data-science-maturity-models

AZRAM, Komal, 2021. Selenium vs Scrapy: Which One Should You Choose for Web Scraping? *BlazeMeter* [online]. [cit. 2023-08-19]. Dostupné z: https://www.blazemeter.com/blog/scrapy-vs-selenium

BAELDUNG, 2023. Introduction to ScalaPy. *Baeldung* [online]. [cit. 2023-08-18]. Dostupné z: https://www.baeldung.com/scala/scalapy-intro

BERKLEY, ed., © 2023. *What is Data Science?* [online]. [cit. 2023-08-07]. Dostupné z: https://ischoolonline.berkeley.edu/data-science/what-is-data-science/

BETTEREVALUATION, © 2022. Word cloud. *Betterevaluation* [online]. [cit. 2023-08-21]. Dostupné z: https://www.betterevaluation.org/methods-approaches/methods/word-cloud

BOUDREAU, Emma, 2021. Should You Learn Scala For Data Science In 2021? *Towards data science* [online]. [cit. 2023-08-20]. Dostupné z: https://towardsdatascience.com/should-you-learn-scala-for-data-science-in-2021-cf7810be7bfc

BUSH, Marela, 2021. What is a Word Cloud and What are the Benefits? *Chartattack* [online]. [cit. 2023-08-21]. Dostupné z: https://www.chartattack.com/what-is-word-cloud/

CARNEGIE, ed., 2019. *Capability Maturity Model* [online]. [cit. 2023-05-25]. Dostupné z: https://managementmania.com/en/cmm-capability-maturity-model

CETINER, Serif, ed., nedatováno. Pictogram Chart. *Dataforvisualization* [online]. [cit. 2023-08-21]. Dostupné z: https://dataforvisualization.com/charts/pictogram-chart/

CIMSS, © 2023. What Is MATLAB? *CIMSS* [online]. Wisconsin [cit. 2023-08-18]. Dostupné z: https://cimss.ssec.wisc.edu/wxwise/class/aos340/spr00/whatismatlab.htm

CIVIL, 2022. What Is A Choropleth Map? Choropleth Map Advantages And Disadvantages. *Civilstuff* [online]. [cit. 2023-08-21]. Dostupné z: https://civilstuff.com/what-is-a-choropleth-map

COINMARKETCAP, © 2023. About CoinMarketCap. *CoinMarketCap* [online]. [cit. 2023-08-10]. Dostupné z: https://coinmarketcap.com/about/

COINMARKETCUP, © 2023. Tether. *CoinMarketCap* [online]. [cit. 2023-08-10]. Dostupné z: https://coinmarketcap.com/currencies/tether/

COINMARKETCUP, © 2023. About Bitcoin. *CoinMarketCap* [online]. [cit. 2023-08-10]. Dostupné z: https://coinmarketcap.com/currencies/bitcoin/

COINMARKETCUP, © 2023. Ethereum. *CoinMarketCap* [online]. [cit. 2023-08-10]. Dostupné z: https://coinmarketcap.com/currencies/ethereum/

CONWAY, Drew, 2010. *The Data Science Venn Diagram* [online]. [cit. 2023-08-07]. Dostupné z: https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html

CREATIVE, ed., © 2023. *Data lifecycle management (DLM) – A guide for businesses* [online]. [cit. 2023-08-08]. Dostupné z: https://www.creative.onl/data-lifecycle-management/

ČSO, 2023. Působnost Českého statistického úřadu. *Český statistický úřad* [online]. [cit. 2023-08-18]. Dostupné z: https://www.czso.cz/csu/czso/4_pusobnost_ceskeho_statistickeho_uradu

DALEY, Sam, 2023. *What Is Data Science?* [online]. [cit. 2023-08-07]. Dostupné z: https://builtin.com/data-science

DARINA, 2022. The Julia Programming Language: The History and Uses. *Leftronic* [online]. [cit. 2023-08-18]. Dostupné z: https://leftronic.com/blog/julia-programming-language/

DATACARPENTRY, 2017. Aggregating and analyzing data with dplyr. *Datacarpentry* [online]. [cit. 2023-08-19]. Dostupné z: https://datacarpentry.org/R-genomics/04-dplyr.html

DATAWORKS, ed., nedatováno. *The 5 Stages of Data LifeCycle Management* [online]. [cit. 2023-08-08]. Dostupné z: https://www.dataworks.ie/5-stages-in-the-data-management-lifecycle-process/

DOMINO, ed., © 2023. What is Jupyter Notebook? *Domino.ai* [online]. [cit. 2023-08-08]. Dostupné z: https://domino.ai/data-science-dictionary/jupyter-notebook

EADS, Audrey a Jocelyne GAFNER, 2023. *What Are Maturity Models?* [online]. [cit. 2023-05-25]. Dostupné z: https://www.indeed.com/career-advice/career-development/what-are-maturity-models

ELASTICSEARCH, ed., © 2023. Analyzing the past and present. *Elasticsearch* [online]. United States [cit. 2023-08-18]. Dostupné z: https://www.elastic.co/guide/en/machine-learning/7.17/ml-overview.html#ml-forecasting)

FOOTE, Keith, 2021. *A Brief History of Data Science* [online]. [cit. 2023-08-07]. Dostupné z: https://www.dataversity.net/brief-history-data-science/#

FREEMAN, James, © 2023. Gantt Charts' Advantages and Disadvantages for Project Management. *Edraw* [online]. [cit. 2023-08-21]. Dostupné z: https://www.edrawsoft.com/project/advantages-disadvantages-gantt-chart.html?gclid=CjwKCAjw0N6hBhAUEiwAXab-TWLcFex80Z9ItVH73Q5gGgqCREZjldoZfi2gHKKcAnRyAZdiZ5MuRRoCTKwQAvD_BwE

FREEMAN, James, © 2023. Advantages and Disadvantages of Area Charts. *Edrawsoft* [online]. [cit. 2023-08-21]. Dostupné z: https://www.edrawsoft.com/basic-areachart-knowledge.html

FREEMAN, James, © 2023. *Area Chart* [online]. [cit. 2023-08-21]. Dostupné z: https://www.edrawsoft.com/basic-areachart-knowledge.html

GEEKSFORGEEKS, nedatováno. Applications, Advantages and Disadvantages of Graph. *Geeksforgeeks* [online]. [cit. 2023-08-21]. Dostupné z: https://www.geeksforgeeks.org/applications-advantages-and-disadvantages-of-graph/

GORELIK, Alex, 2019. *The enterprise big data lake: delivering the promise of big data and data science*. Sebastopol, California: iO'Reilly Media. ISBN 14-919-3155-8.

GROVER, Shubham, 2022. What are Heat Maps: Types, Benefits & How to Use Them? *Adpushup* [online]. [cit. 2023-08-21]. Dostupné z: https://www.adpushup.com/blog/what-are-heat-maps-adpushup/

HAIFENG, Li, 2014. Smile. *Smile* [online]. [cit. 2023-08-21]. Dostupné z: https://haifengl.github.io/

HALL, David, 2023. ScalaNLP. *ScalaNLP* [online]. [cit. 2023-08-21]. Dostupné z: http://www.scalanlp.org/about/

HALL, David, 2021. Quickstart · scalanlp/breeze Wiki · GitHub. *Quickstart · scalanlp/breeze Wiki · GitHub* [online]. [cit. 2023-08-21]. Dostupné z: https://github.com/scalanlp/breeze/wiki/Quickstart

HALL, David, 2014. GitHub - dlwh/epic. *GitHub - dlwh/epic* [online]. [cit. 2023-08-21]. Dostupné z: https://github.com/dlwh/epic

HALL, David, 2014. GitHub - dlwh/puck. *GitHub - dlwh/puck* [online]. [cit. 2023-08-21]. Dostupné z: https://github.com/dlwh/puck

HALL, David, © 2023. ScalaNLP. *ScalaNLP* [online]. [cit. 2023-08-20]. Dostupné z: www.scalanlp.org

HAMID, Kaiser, 2023. How To Use Scala for Data Science. *Knowledgehut* [online]. [cit. 2023-08-20]. Dostupné z: https://www.knowledgehut.com/blog/data-science/how-to-use-scala-for-data-science

HASSANI, Hossein a Emmanuel Sirmal SILVA, 2023. The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field. *Big Data and Cognitive Computing* [online]. **7**(2) [cit. 2023-08-10]. ISSN 2504-2289. Dostupné z: doi:10.3390/bdcc7020062

HENDL, Jan, 2021. *Big data: věda o datech - základy a aplikace*. Praha: Grada Publishing. Průvodce (Grada). ISBN 978-80-271-3031-3.

HORNICK, Mark, 2020. *A Data Science Maturity Model for Enterprise Assessment*. Oracle.

CHANDAN, 2023. Python for Data Science Tutorial. *Python for Data Science Tutorial* [online]. [cit. 2023-08-18]. Dostupné z: https://www.geeksforgeeks.org/python-for-data-science/

CHAPMAN, Pete, Julian CLINTON a Randy KERBER, 2000. *CRISP-DM: Step-by-step data mining guide*.

IBM, ed., nedatováno. *What is data lifecycle management?* [online]. [cit. 2023-08-08]. Dostupné z: https://www.ibm.com/topics/data-lifecycle-management

IRIZARRY, Rafael, nedatováno. Dplyr tutorial. *Github* [online]. [cit. 2023-08-19]. Dostupné z: https://genomicsclass.github.io/book/pages/dplyr_tutorial.html

JULIALANG, ed., 2022. Julia 1.9 Documentation. *Julialang* [online]. [cit. 2023-08-18]. Dostupné z: https://docs.julialang.org/en/v1/#man-introduction,

JUPYTER, ed., 2015. *Project Jupyter Documentation* [online]. [cit. 2023-07-10]. Dostupné z: https://docs.jupyter.org/en/latest/

JUPYTER, ed., © 2023. *Try Jupyter* [online]. [cit. 2023-07-10]. Dostupné z: https://jupyter.org/try

JUPYTER, ed., 2015. *The Jupyter Notebook* [online]. [cit. 2023-07-10]. Dostupné z: https://jupyter-notebook.readthedocs.io/en/latest/notebook.html

KDD, 2022. *SIGKDD* [online]. [cit. 2023-08-07]. Dostupné z: https://www.kdd.org/

KELLEHER, John D. a Brendan TIERNEY, 2018. *Data science*. Cambridge, Massachusetts: The MIT Press. MIT Press essential knowledge series. ISBN 978-0262535434.

KERAS, nedatováno. About Keras. *Keras* [online]. [cit. 2023-08-18]. Dostupné z: https://keras.io/about

LUNA, Javier, 2023. Top programming languages for data scientists in 2023. *DataCamp* [online]. [cit. 2023-08-18]. Dostupné z: https://www.datacamp.com/blog/top-programming-languages-for-data-scientists-in-2022

MATHWORKS, © 2023. Deep Learning Toolbox. *Mathworks* [online]. [cit. 2023-08-18]. Dostupné z: https://www.mathworks.com/help/deeplearning/index.html?s_tid=CRUX_lftnav

MATHWORKS, © 2023. AI, Data Science, and Statistics. *Mathworks* [online]. [cit. 2023-08-18]. Dostupné z: https://www.mathworks.com/help/overview/ai-data-science-and-statistics.html?s_tid=CRUX_lftnav,

MATHWORKS, © 2023. Statistics and Machine Learning Toolbox. *Mathworks* [online]. [cit. 2023-08-18]. Dostupné z: https://www.mathworks.com/help/stats/index.html?s_tid=CRUX_lftnav

MATHWORKS, © 2023. Curve Fitting Toolbox. *MathWorks* [online]. [cit. 2023-08-18]. Dostupné z: (https://www.mathworks.com/help/curvefit/index.html?s_tid=hc_product_card

MATHWORKS, © 2023. Text Analytics Toolbox. *Mathworks* [online]. [cit. 2023-08-18]. Dostupné z: https://www.mathworks.com/help/textanalytics/index.html?s_tid=CRUX_lftnav

MCKINNEY, Wes, 2022. *Python for data analysis: data wrangling with pandas, NumPy, and Jupyter*. Third edition. Beijing: O'Reilly. ISBN 978-1098104030.

MICROSOFT, ed., © 2023. *Co je vizualizace dat* [online]. [cit. 2023-08-07]. Dostupné z: https://powerbi.microsoft.com/cs-cz/data-visualization/

MICROSOFT, © 2023. *Co je to datová věda?* [online]. [cit. 2023-08-07]. Dostupné z: https://azure.microsoft.com/cs-cz/resources/cloud-computing-dictionary/what-is-data-science/#introduction-science

MIDJOURNEY, 2023. Jak rozmnožit své kryptoměnové jmění. Někdy digitální mince padají z nebe. *E15* [online]. [cit. 2023-08-10]. Dostupné z: https://www.e15.cz/kryptomeny-investice

MILLER, Kelsey, 2019. *17 DATA VISUALIZATION TECHNIQUES ALL PROFESSIONALS SHOULD KNOW* [online]. [cit. 2023-08-07]. Dostupné z: https://online.hbs.edu/blog/post/data-visualization-techniques

MORRIS, Andy, 2021. Heat Maps: Types, Benefits & How to Use Them. *Netsuite* [online]. [cit. 2023-08-21]. Dostupné z: https://www.netsuite.com/portal/resource/articles/erp/heat-map.shtml

MÜLLER, Kirill, 2023. Tibble: Simple Data Frames. *Tibble* [online]. [cit. 2023-08-19]. Dostupné z: https://tibble.tidyverse.org/

NUMPY, ed., © 2022. What is NumPy? *NumPy* [online]. [cit. 2023-08-18]. Dostupné z: https://numpy.org/doc/stable/user/whatisnumpy.html

PANDAS, © 2023. About pandas. *Pandas* [online]. [cit. 2023-08-18]. Dostupné z: https://pandas.pydata.org/about/

PIMPLER, Eric, © 2017. *Data Visualization and Exploration with R* [online]. GeoSpatial. ISBN 978-1727588484.

PLOTLY, © 2023. Getting Started with Plotly in Python. *Plotly* [online]. [cit. 2023-08-18]. Dostupné z: https://plotly.com/python/getting-started/

POSIT, © 2023. Open source for a better world. *Posit* [online]. [cit. 2023-08-19]. Dostupné z: https://posit.co/about/

PROENCA, Diogo a José BORBINHA, 2016. *Maturity Models for Information Systems - A State of the Art* [online]. [cit. 2023-08-08]. Dostupné z: doi:https://doi.org/10.1016/j.procs.2016.09.279

PYCOACH, 2020. Web Scraping with Beautiful Soup, Selenium, or Scrapy? *Towardsdatascience* [online]. [cit. 2023-08-19]. Dostupné z: https://towardsdatascience.com/web-scraping-with-beautiful-soup-selenium-or-scrapy-62c6f3545de7

PYTHON, 2022. Python Guide. *Wiki Python* [online]. [cit. 2023-08-18]. Dostupné z: https://wiki.python.org/moin/BeginnersGuide/Overview

RAJASEKHAR, © 2023. Advantages and disadvantages of pie charts. *ExcelR* [online]. [cit. 2023-08-21]. Dostupné z: https://www.excelr.com/advantages-and-disadvantages-of-pie-charts

R FOUNDATION, nedatováno. What is R? *R project* [online]. [cit. 2023-08-19]. Dostupné z: https://www.r-project.org/about.html

ROUSE, Margaret, 2020. *Data Scientist* [online]. [cit. 2023-08-07]. Dostupné z: https://www.techopedia.com/definition/28177/data-scientist

SALAZAR, Rogel, 2022. Tensorflow, PyTorch or Keras for Deep Learning. *Domino* [online]. [cit. 2023-08-19]. Dostupné z: (https://www.dominodatalab.com/blog/tensorflow-pytorch-or-keras-for-deep-learning?_ga=2.161136954.644948688.1688507525-688597834.1687871382

SATURNCLOUD, ed., 2023. Jupyter Notebook vs JupyterLab. *Saturncloud* [online]. [cit. 2023-08-08]. Dostupné z: https://saturncloud.io/glossary/jupyter-notebook-vs-jupyterlab/

SCALAPY, © 2022. Getting Started with ScalaPy. *ScalaPy* [online]. [cit. 2023-08-18]. Dostupné z: https://scalapy.dev/docs/

SMIGEL, Leo, 2023. What Is Open High Low Close in Stocks? *Analyzingalpha* [online]. [cit. 2023-08-10]. Dostupné z: https://analyzingalpha.com/open-high-low-close-stocks

SOAGE, José, 2023. The ggplot2 package. *The ggplot2 package* [online]. [cit. 2023-08-21]. Dostupné z: https://r-charts.com/ggplot2/

STEELE, Mac, 2016. Introducing the Data Science Maturity Model. *Domino* [online]. [cit. 2023-08-21]. Dostupné z: https://domino.ai/blog/introducing-the-data-science-maturity-model.

STEEL, Mac, 2016. *Introducing the Data Science Maturity Model* [online]. [cit. 2023-08-08]. Dostupné z: https://domino.ai/blog/introducing-the-data-science-maturity-model

STOBIERSKI, Tim, 2021. *8 STEPS IN THE DATA LIFE CYCLE* [online]. [cit. 2023-06-21]. Dostupné z: https://online.hbs.edu/blog/post/data-life-cycle

STOROPOLI, Jose, Rik HUIJZER a Lazaro ALONSO, 2021. *Julia Data Science*. ISBN 9798489859165.

TENSORFLOW, 2023. Keras: The high-level API for TensorFlow. *TensorFlow* [online]. [cit. 2023-08-19]. Dostupné z: https://www.tensorflow.org/guide/keras

TRIBUNE TRUST, 2023. 5 Best R IDEs & Editors in 2023. *Tribune Trust* [online]. [cit. 2023-08-19]. Dostupné z: https://www.tribuneindia.com/news/brand-connect/5-best-r-ides-editors-in-2023-481023

TUTORIALSPOINT, nedatováno. Scikit Learn. *Tutorialspoint* [online]. [cit. 2023-08-18]. Dostupné z: https://www.tutorialspoint.com/scikit_learn/scikit_learn_introduction.htm

VIDHYA, 2019. A Beginner's Guide to Tidyverse – The Most Powerful Collection of R Packages for Data Science. *Analyticsvidhya* [online]. [cit. 2023-08-19]. Dostupné z: https://www.analyticsvidhya.com/blog/2019/05/beginner-guide-tidyverse-most-powerful-collection-r-packages-data-science/,

WICKHAM, Hadley, 2022. Simple, Consistent Wrappers for Common String Operations. *Simple, Consistent Wrappers for Common String Operations* [online]. [cit. 2023-08-21]. Dostupné z: https://stringr.tidyverse.org/

WICKHAM, Hadley, 2023. Readr: Read Rectangular Text Data. *Readr: Read Rectangular Text Data* [online]. [cit. 2023-08-19]. Dostupné z: https://readr.tidyverse.org/

WICKHAM, Hadley, 2022. Simple, Consistent Wrappers for Common String Operations. *Simple, Consistent Wrappers for Common String Operations* [online]. [cit. 2023-08-21]. Dostupné z: https://stringr.tidyverse.org/

WICKHAM, Hadley a Lionel HENRY, 2023. Purrr: Functional Programming Tools. *Purrr: Functional Programming Tools* [online]. [cit. 2023-08-21]. Dostupné z: https://purrr.tidyverse.org

WORLD BANK GROUP, ed., © 2023. The World Bank Group about. *The World Bank Group* [online]. [cit. 2023-08-18]. Dostupné z: https://data.worldbank.org/about

Data Presentation: Pie Charts, 2022. *Barcelona Field Studies Centre* [online]. Barcelona [cit. 2023-08-21]. Dostupné z: https://geographyfieldwork.com/DataPresentationPieCharts.htm

Peníze: Spořící účty [online], © 2023. NextPage Media [cit. 2023-08-10]. ISSN 1213-2217. Dostupné z: www.penize.cz

Peníze: Inflace [online], © 2023. NextPage Media [cit. 2023-08-10]. ISSN 1213-2217. Dostupné z: www.penize.cz

Advantages and Disadvantages of Bar Graphs, 2022. *All Things Statistics* [online]. [cit. 2023-08-21]. Dostupné z: https://allthingsstatistics.com/miscellaneous/bar-graphs-advantages-disadvantages/

Advantages and Disadvantages of Histogram, 2022. *All Things Statistics* [online]. [cit. 2023-08-21]. Dostupné z: https://allthingsstatistics.com/miscellaneous/histogram-advantages-disadvantages/

Advantages & Disadvantages of Box Plots, nedatováno. *StudyLib* [online]. [cit. 2023-08-21]. Dostupné z: https://studylib.net/doc/5346518/advantages-and-disadvantages-of-box-plots

Peníze: Kryptoměny [online], © 2023. NextPage Media [cit. 2023-08-10]. ISSN 1213-2217. Dostupné z: https://www.penize.cz/kryptomeny

What is Kibana? [online], © 2023. Mountain View: Elastic [cit. 2023-07-18]. Dostupné z: https://www.elastic.co/what-is/kibana

Beats: Data Shippers for Elasticsearch [online], © 2023. Mountain View: Elastic [cit. 2023-07-18]. Dostupné z: https://www.elastic.co/beats/

Box plot, 2023. *QlikTech* [online]. King of Prussia [cit. 2023-08-21]. Dostupné z: https://help.qlik.com/en-US/sense/February2023/Subsystems/Hub/Content/Sense_Hub/Visualizations/BoxPlot/box-plot.htm

Gantt Chart: Definition and Examples, © 2023. *ProjectManager* [online]. Austin: ProjectManager [cit. 2023-08-13]. Dostupné z: https://www.projectmanager.com/guides/gantt-chart

Bar Chart Vs Histogram: What Are The Key Differences, 2023. *Software Testing Help* [online]. Software Testing Help [cit. 2023-08-13]. Dostupné z: https://www.softwaretestinghelp.com/bar-chart-vs-histogram/#What_Is_A_Histogram

Logstash: Collect, Parse, Transform Logs [online], © 2023. Mountain View: Elastic [cit. 2023-07-18]. Dostupné z: https://www.elastic.co/logstash/

Data in: documents and indices, © 2023. *Elasticsearch Platform — Find real-time answers at scale | Elastic* [online]. Mountain View: Elastic [cit. 2023-07-18]. Dostupné z: https://www.elastic.co/guide/en/elasticsearch/reference/8.8/documents-indices.html

Data Lifecycle Management (DLM) : A New Way of Managing Data, 2021. *Techno-PM* [online]. Sydney [cit. 2023-08-21]. Dostupné z: https://www.techno-pm.com/2021/08/data-lifecycle-management-dlm.html

Scalability and resilience: clusters, nodes, and shards, © 2023. *Elasticsearch Platform — Find real-time answers at scale | Elastic* [online]. Mountain View: Elastic [cit. 2023-07-18]. Dostupné z: (https://www.elastic.co/guide/en/elasticsearch/reference/8.8/scalability.html)

Information out: search and analyze, © 2023. *Elasticsearch Platform — Find real-time answers at scale | Elastic* [online]. Mountain View: Elastic [cit. 2023-07-18]. Dostupné z: https://www.elastic.co/guide/en/elasticsearch/reference/8.8/search-analyze.html

What is Elasticsearch?, © 2023. *Elasticsearch Platform — Find real-time answers at scale | Elastic* [online]. Mountain View: Elastic [cit. 2023-07-18]. Dostupné z: https://www.elastic.co/guide/en/elasticsearch/reference/8.8/elasticsearch-intro.html

Cstjean/ScikitLearn.js, 2023. *GitHub* [online]. San Francisco: GitHub [cit. 2023-07-19]. Dostupné z: https://github.com/cstjean/ScikitLearn.jl#readme

CRISP-DM Help Overview, 2021. *CRISP-DM Help Overview* [online]. Armonk [cit. 2023-08-21]. Dostupné z: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

An Introduction to g g p l o t 2, nedatováno. *An Introduction to g g p l o t 2* [online]. Cincinnati [cit. 2023-08-21]. Dostupné z: https://uc-r.github.io/ggplot_intro

DataFrames.jl [online], 2023. Introduction · DataFrames.jl: Introduction · DataFrames.jl [cit. 2023-07-19]. Dostupné z: https://dataframes.juliadata.org/stable/

Data Science Python - Getting Started, © 2023. *Tutorials Point* [online]. [cit. 2023-08-18]. Dostupné z: https://www.tutorialspoint.com/python_data_science/python_data_science_introduction.htm

Introduction [online], 2023. DataFramesMeta Documentation: DataFramesMeta Documentation [cit. 2023-07-19]. Dostupné z: https://juliadata.github.io/DataFramesMeta.jl/stable/