Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky

Lidé v digitálním světě

Bc. Jan Barva

Diplomová práce
2017

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: **2016/2017**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Jan Barva**
Osobní číslo:	**E15679**
Studijní program:	**N6209 Systémové inženýrství a informatika**
Studijní obor:	**Informatika ve veřejné správě**
Název tématu:	**Lidé v digitálním světě**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

Diplomová práce se zabývá využitím Internetu, počítačů, získávání e-dovedností a dalších aktivit na Internetu mladými lidmi; je zaměřena na porovnání ČR a členských států EU. Obsahem je analýza dostupných dat, návrh a tvorba modelu pomocí metod data miningu.

Zásady pro vypracování:
- zhodnocení současného stavu ve zvolené oblasti, formulace problému;
- sběr dat z dostupných datových zdrojů a jejich analýza;
- návrh a tvorba modelu pomocí metod data miningu;
- interpretace získaných výsledků a formulace závěrů.

Rozsah grafických prací:

Rozsah pracovní zprávy: **cca 50 stran**

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

BERKA, P. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, 366 s. ISBN 80-200- 1062-9.
HAN, J., KAMBER, M. Data Mining: Concepts and Techniques. Vyd. 2. San Frencisco: Morgan Kaufmann, 2006. ISBN 1-55860-901-6.
MELOUN, M., MILITKÝ, J. Kompendium statistického zpracování dat: metody a řešené úlohy včetně CD. Vyd. 1. Praha: Academia Praha, 2002. ISBN 80-200-1008-4.
PETR, P. Data Mining. Díl 1. Pardubice: Univerzita Pardubice, 2008. 139 s. ISBN 978-80- 7395-098- 9.
RUD, O. P. Data Mining. Praha: Computer Press, 2001. 329 s. ISBN 80-7226-577-6.
Zdroje Internetu.

Vedoucí diplomové práce: **Ing. Miloslava Kašparová, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání diplomové práce: **4. září 2016**

Termín odevzdání diplomové práce: **28. dubna 2017**

L.S.

doc. Ing. Romana Provazníková, Ph.D.
děkanka

doc. Ing. Pavel Petr, Ph.D.
vedoucí ústavu

V Pardubicích dne 4. září 2016

PROHLÁŠENÍ

Prohlašuji, že jsem tuto práci 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., autorský zákon, 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. 9/2012, bude práce zveřejněna v Univerzitní knihovně a prostřednictvím Digitální knihovny Univerzity Pardubice.

V Pardubicích dne 28. 6. 2017 Bc. Jan Barva

PODĚKOVÁNÍ:

Rád bych touto cestou poděkoval vedoucí své diplomové práce paní Ing. Miloslavě Kašparové, Ph.D., za odborné vedení, poskytnuté materiály, cenné připomínky, náměty, rady a čas věnovaný konzultacím.

ANOTACE

Diplomová práce zkoumá chování v digitálním světě mladými lidmi, porovnává Českou republiku a jednotlivé členské státy Evropské unie. Samotné modelování je provedeno metodou shlukové analýzy. Vyhodnocení jednotlivých modelů je provedeno graficky i popisně.

KLÍČOVÁ SLOVA

Digitální svět, shluková analýza, modelování, příprava dat.

TITLE

People in the digital world.

ANNOTATION

Thesis deals with a behavior of young people in the digital world, compares the Czech Republic and individual member states of the European Union. Cluster analysis is used for modeling. An evaluation of individual models is done graphically and descriptively.

KEYWORDS

Digital world, cluster analysis, modeling, data preparation.

OBSAH

SEZNAM TABULEK

SEZNAM OBRÁZKŮ

SEZNAM ZKRATEK

CERN	Conseil Européen pour la recherche nucléaire
CRISP-DM	Cross Industry Standard Process for Data Mining
ČR	Česká republika
ČSÚ	Český statistický úřad
EHS	Evropské hospodářské společenství
EU	Evropská unie
HDP	Hrubý domácí produkt
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IP	Internet Protocol
TCP	Transmission Control Protocol
URL	Uniform Resource Locator

ÚVOD

Mladí lidé se obecně stávají přímými aktéry digitálního pokroku. Je to generace, která má nové moderní technologie tak říkajíc v sobě, vyrůstá v digitálním světě. Internet, tablety, smartphony jsou naprosto samozřejmé věci, bez kterých si již dnešní život neumí představit. Dostupnost veškerých informací na Internetu, vyhledávání potřebných informací, nakupování či prodávání prostřednictví e-shopů, placení v bankingových aplikacích, jsou naprosté samozřejmosti dnešního světa. A pro generaci mladých lidí, kteří vlastně ani nepoznali jiný než digitální svět, je to velká výzva, nutnost sledovat neustálý vývoj digitálních technologií, snažit se udržet v kontaktu, zvládat digitální dovednosti, která jsou následně výborně uplatnitelné jak v soukromém, tak zejména v pracovním životě.

Již samotný proces vzdělávání je přizpůsobován digitální éře. Dnešní mladá generace byla „při tom“, když běžel v České republice projekt Internet do škol. Kdy se proces vzdělávání snažil přizpůsobit nejnovějším informačním trendům. V oblasti digitálních dovedností se dnes již nevzdělávají jen studenti, kteří si vybrali obory se zaměřením na oblast informačních a komunikačních technologií, ale vzděláváni musí být v této oblasti všichni. Vždyť digitální Evropa potřebuje digitální dovednosti od všech zúčastněných. Česká republika se stala součástí Evropské unie, a tak se musí zvolna přizpůsobit trendům běžným v některých členských státech Evropské unie.

Cílem této diplomové práce je porovnat současný stav vnímání, vztahu a interpretace digitálního světa mladými lidmi žijícími v České republice a v dalších členských zemích Evropské unie. Pomocí navržených a následně vytvořených modelů zhodnotit využívání počítačů, Internetu, sociálních sítí a dalších vybraných atributů vztahujících se k digitálnímu světu.

1 POPIS SOUČASNÉHO STAVU

V této kapitole jsou vymezeny základní pojmy z oblasti informačních technologií, digitálního světa, popsány vybrané charakteristiky jednotlivých zemí Evropské unie (EU), vymezeny základní atributy a popisné charakteristiky.

1.1 Základní pojmy

Internet

Globální internetová síť zvaná Internet má přes 200 milionů účastníků. Internet poskytuje přístup k hypertextovým dokumentům, elektronickou poštu, audiovizuální přenos, přenos datových souborů a programů. Internet nikdo neřídí ani nevlastní, je to volně organizovaná mezinárodní spolupráce propojených sítí, které dobrovolně přijaly a dodržují standardní protokoly a procedury. [21]

Internet lze definovat jako globální informační systém, který je logicky propojen do jednoho celku prostřednictví adresního prostoru založeném na protokolu IP a je schopen podporovat komunikaci prostřednictvím rodiny protokolů TCP. [29] [32]

Internet je nepředvídaný výsledek projektu ARPANET, který měl za cíl vytvořit komunikační síť. ARPANET síť neměla žádný centrální řídící člen a byla zcela decentralizovaná. Tento koncept komunikačního systému navrhl Paul Baran v roce 1962. Celý projekt agentury ARPA je financován Pentagonem. Samotný vznik slova Internet je spojován s přijmutím standardního protokolu TCP/IP, který umožnil komunikaci počítačů bez ohledu na jejich operační systém. [21] [32]

Základními články Internetu jsou síť, server a klient. Síť se skládá ze serverů, pracovních stanic, přenosových linek a programů umožňujících přenos různých typů elektronických informací. Pod serverem si lze představit počítač v síti, který odesílá soubory nebo spouští aplikace pro jiné počítače v síti. Klientským počítačem je obvykle pracovní stanice, která umožňuje připojení k serveru a správu získaných informací. [21]

Samotná síť propojených hypertextových dokumentů byla vynalezena koncem roku 1990 v CERN. Dále byla definována internetová adresa URL, hypertextový protokol HTTP a programovací jazyk HTML. Velký boom Internetu začal vývojem nových prohlížečů v roce 1995, ke kterému přispěly zejména firmy Netscape a Microsoft. Většina uživatelů Internetu se sídlem v České republice má národní doménu cz. [21]

Domnívat se, že Internet již prodělal ty nejpodstatnější změny, by bylo obrovskou chybou. Internet se mění neustále, podobně jako se vyvíjí podobně jako počítačový průmysl. Na Internetu jsou neustále nabízeny nové služby jako například přenos v reálném čase, přenos audia a videa umožňuje sledovat díky Internetu rozhlasové či televizní stanice. Dostupností nových možností připojení na Internet z přenosných zařízení se Internet skutečně stal globálním komunikačním prostředím. Vývoj přináší i nové aplikace nebo důležitou roli jako prostor pro elektronické obchodování. [29]

Informační prostor Internetu roste každý den a zároveň přibývají nové online dostupné informační zdroje. Internet představuje rozlehlý prostor, a proto jsou dnes stále populárnější vyhledávací služby, které jsou využívány k online získávání informací. [29]

Na Internet lze nahlížet z různých úhlů pohledu. Můžeme si ho představit jako gigantické seskupení počítačů, soubor programů, online zdrojů nebo jako celosvětovou elektronickou knihovnu. Pohled jednotlivého uživatele vždy závisí na tom, jak Internet používá, co potřebuje dělat. V počátcích Internetu se lidé dívali na Internet jako síť tvořenou počítači. Při hledání informací bylo potřeba znát přesnou adresu počítače, kde jsou uloženy potřebné soubory. Což je zřejmou nevýhodou, protože aby uživatel mohl efektivně pracovat s informacemi, musel znát množství jmen počítačů. A tak vznikla doménová jména počítačů, pro snadnější zapamatování. [29]

Internet je na jedné straně zdrojem informací, ale na straně druhé také zdrojem dezinformací. Tudíž vše, co objevíte, není zaručeně pravdivé. Materiály dnes na Internetu publikují jednotlivci, firmy a organizace, jejich názory jsou přirozeně různé. Většina dostupných materiálů neprochází žádným vydavatelským procesem a lze tedy narazit na množství chyb, opomenutí, lží i podjatých textů. Při posuzování zdroje významně pomůže jeho identifikaci či dřívější zkušenosti uživatele s daným zdrojem. [29]

Dnes jsme na Internet připojeni přes počítač, mobil, tablet, ale také třeba různými domácími zařízeními od kávovaru po celý dům. Ukládání dat a čerpání služeb z Internetu je dnes extrémně rozšířené. Proto množství zařízení a počet uživatelů připojených k Internetu stále roste. [31]

Dnes má připojení k Internetu zhruba polovina světové populace. Počet zařízení připojených k Internetu se odhaduje přibližně na 10 miliard a stále stoupá. Současně s tím roste i globální objem datové výměny. Internet se stal jedním z nejvýraznějších fenoménů přelomu druhého a třetího tisíciletí a jeho role se dá v historii s největší pravděpodobností přirovnat k vynálezům knihtisku, parního stroje nebo dynama na výrobu elektřiny. Internetu,

na rozdíl od většiny významných technologických milníků, nemůžeme jednoznačně přiřadit jeho autora v podobě konkrétní osoby nebo skupiny osob. Nejpodstatnější vlastností Internetu je obrovský informační dopad. Cokoliv je zveřejněno na Internetu, to se okamžitě stává veřejným. [31]

Webová úložiště umožňují uživatelů ukládat a sdílet soubory a složky s ostatními uživateli Internetu. Mezi nejznámější globální úložiště patří například Apple iCloud, Dropbox, Google Drive nebo Microsoft OneDrive. V českých podmínkách jsou známé servery Uschovna.cz či Uloz.to. Dnes se hovoří spíše o cloudových službách na místo serverů. Cloud computing je to, že uživatel využívá službu ze vzdáleného centra vlastněného a spravovaného poskytovatelem cloudu prostřednictvím Internetu. [30] [31]

Internet by se technologicky nejvhodněji definoval jako celosvětový systém navzájem propojených sítí, ve kterých dochází ke komunikaci počítačů pomocí rodiny protokolů TCP/IP. Oprávnění uživatelé mohou využívat přenosových kapacit a zdrojů této sítě. Všechny činnosti prováděné na Internetu spočívají v získávání, přenosu, zpracování a šíření informací nebo dat. [31]

Data a informace

Pojmy data a informace jsou velmi rozšířené, tuto dvojici ještě doplňuje pojem znalost, který se používá v souvislosti s umělou inteligencí. [29]

V oblasti informatiky se pojem data vždy používal jako označení například pro čísla, text, zvuk, obraz. Rozlišujeme dva typy dat. Strukturovaná data zachycují fakta, atributy, objekty, apod. Typickým příkladem strukturovaných dat je ukládání dat pomocí relačních databázových systémů, kde se obvykle používá hierarchie pole-záznam-relace-databáze. Nestrukturovaná data jsou vyjádřena jako tok bytů bez dalšího rozlišení. Data slouží pro vyjádření faktů, atributů, odrazů dějů a věcí. [29]

Pojem informace je používán v mnoha oborech, a proto existuje mnoho definic. Obecně lze informace chápat jako množinu poznatků, která je někým použita v konkrétní situaci pro řešení problémů. Jelikož nejsou informace mnohdy k dispozici, jsou vyhledávány v externích zdrojích. Proces transformace od poznatků k informacím obstarávají informační systémy. Pro získání úplné informace je nutno čerpat z různých informačních zdrojů a skládat jako mozaiku. [29]

Další pohled na pojetí informací přinesla umělá inteligence. Znalostí se rozumí vzájemně provázané struktury souvisejících poznatků. Za znalost lze považovat i reprezentaci v podobě

kognitivního modelu, včetně schopnosti s ním provázet řadu kognitivních operací a tím tak předvídat reálný svět. [29]

Informační gramotnost je schopnost porozumět a užívat získané informace. Jednou z nejdůležitějších věcí při práci s informačními zdroji Internetu je nutnost kritického myšlení. Jeho ignorování nebo podcenění může vést k tomu, že Internet se pro nás stane nebezpečnou a klamavou zónou. Internet nabízí nové možnosti pro tradiční média, současně však vytváří interaktivní obsah, který vychází vstříc požadavkům uživatelů. [29]

V širším úhlu pohledu se nedá říci, že bychom trpěli nedostatkem informací. Informace jsou všude kolem nás v různých podobách. Často se lidé musejí vyrovnávat s problémem přehlcení informacemi. V dnešní době tištěné materiály všeho druhu představují méně než procento celkového objemu informací, i když psané slovo je velmi efektivní cestou ke sdělování informací. Velká část celkového objemu je dnes vytvářena v digitální formě. [29]

Kyberprostor

Kyberprostor je pátou dimenzí života, se všemi rysy každodenních společenských aktivit. A stejně jako reálný svět, tak nabývá i kyberprostor všech společenských atributů, náboženských, kulturních, emocionálních, obchodních a politických. Život v kyberprostoru si ale formuluje svoje vlastní pravidla, která se často vymykají přirozenému řádu. Chcete-li v kyberprostoru přežít, nezbývá nic jiného, než přizpůsobit stará pravidla chování anebo vytvořit nová pravidla. Příkladem kyberprostoru mohou být systémy virtuální reality, počítačem simulované prostředí, počítačové hry a Internet. [16]

Digitální svět

Digitální svět je vše kolem nás, digitální informační technologie nás obklopují a ovlivňují náš život. Rozšířením smartphonů, které máme stále u sebe, ještě více narostla míra využívání digitálních informačních technologií. Už to není jen televize, počítač, Internet či tablet, přibyly i již zmíněné smartphony a možnosti jejich využívání se rozšířily. Na druhou stranu se stále někde registrujeme, ukládáme svoje data, a tím pádem má využívání digitálního světa i svoje negativní stránky. A to nejen v možnosti zneužití dat, ale i o vlivu digitálního světa na lidský intelekt a duševní a tělesné zdraví. [33]

Hrubý domácí produkt

Hrubý domácí produkt (HDP) je dle [35] tokovou veličinou měřenou za určité časové období a je peněžním vyjádřením celkové hodnoty statků a služeb nově vytvořených v daném

období na určitém území, používá se pro stanovení výkonnosti ekonomiky. Může být definován, resp. spočten, třemi způsoby: produkční metodou, výdajovou metodou a důchodovou metodou.

Inflace

Obecně inflace dle [35] znamená nepřetržitý růst cenové hladiny v čase. Statistické vyjadřování inflace vychází z měření čistých cenových změn pomocí indexů spotřebitelských cen. Cenové indexy poměřují úroveň cen vybraného koše reprezentativních výrobků a služeb ve dvou srovnávaných obdobích, přičemž váha, která je jednotlivým cenovým reprezentantům ve spotřebním koši přisouzena, odpovídá podílu daného druhu spotřeby, který zastupují, na celkové spotřebě domácností. Do spotřebního koše je zařazeno potravinářské zboží, nepotravinářské zboží a služby.

Nezaměstnanost

Jak uvádí [35], základním ukazatelem měření nezaměstnanosti je míra nezaměstnanosti, která vyjadřuje podíl nezaměstnaných na ekonomicky aktivním obyvatelstvu v procentech.

Počet obyvatel

Dle [4] je počet obyvatel k určitému okamžiku, tedy stav obyvatelstva, jednou ze základních charakteristik, kterou sleduje demografická statistika. Veškeré údaje se přitom týkají všech obyvatel, kteří mají v České republice (ČR) trvalé bydliště, a to bez ohledu na státní občanství. Od roku 2001 údaje zahrnují také cizince s vízy nad 90 dnů a cizince s přiznaným azylem. Od 1. 5. 2004, v návaznosti na tzv. euronovelu zákona č. 326/1999 Sb., o pobytu cizinců, se údaje týkají i občanů zemí EU s přechodným pobytem na území ČR a občanů třetích zemí s dlouhodobým pobytem.

1.2 Členské státy Evropské unie

Evropská unie je svého druhu ojedinělý hospodářský a politický celek 28 evropských zemí, do něhož náleží podstatná část evropského kontinentu. Patří sem tyto země: Belgie, Bulharsko, ČR, Dánsko, Estonsko, Finsko, Francie, Chorvatsko, Irsko, Itálie, Kypr, Litva, Lotyšsko, Lucembursko, Maďarsko, Malta, Německo, Nizozemsko, Polsko, Portugalsko, Rakousko, Rumunsko, Řecko, Slovensko, Slovinsko, Spojené království, Španělsko, Švédsko. [9] [19]

Základy EU byly položeny jen několik let po doznění druhé světové války. První krok evropské integrace spočíval v upevnění hospodářské spolupráce. Vycházelo se

z jednoduchého předpokladu: Je méně pravděpodobné, že státy, které spolu obchodují a jsou tak na sobě ekonomicky závislé, vyvolají ozbrojený konflikt. [9]

V roce 1958 bylo založeno Evropské hospodářské společenství (EHS), které ve svých začátcích rozvíjelo hospodářskou spolupráci šesti zemí: Belgie, Francie, Německa, Itálie, Lucemburska a Nizozemska. Od té doby se vytvořil rozsáhlý jednotný trh, který se neustále rozvíjí v zájmu plného využití svého potenciálu. [19]

Postupně se připojovaly další státy, jak je uvedeno v následující tabulce 1.

Tabulka 1: Datum přistoupení do EU

Datum přistoupení	**Stát**
1. ledna 1958	Belgie
	Francie
	Itálie
	Lucembursko
	Německo
	Nizozemsko
1. ledna 1973	Dánsko
	Irsko
	Spojené království
1. ledna 1981	Řecko
1. ledna 1986	Portugalsko
	Španělsko
1. ledna 1995	Finsko
	Rakousko
	Švédsko
1. května 2004	Česká republika
	Estonsko
	Kypr
	Litva
	Lotyšsko
	Maďarsko
	Malta
	Polsko
	Slovensko
	Slovinsko
1. ledna 2007	Bulharsko
	Rumunsko
1. července 2013	Chorvatsko

Zdroj: upraveno dle [9]

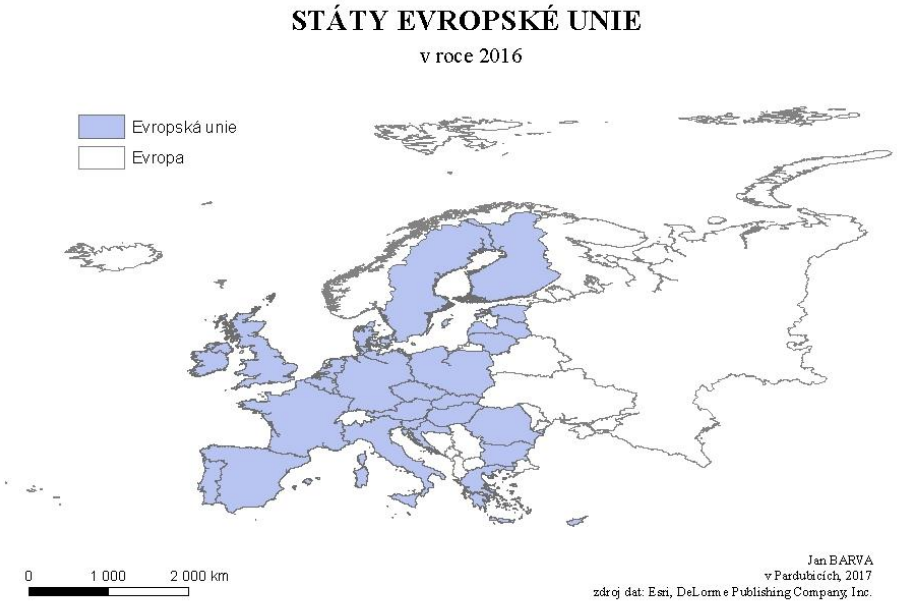
Jednotlivé členské státy EU jsou v krátkosti dle [9] [10] [19] stručně charakterizovány v příloze A. Statistické srovnání ČR a průměru EU u jednotlivých zkoumaných atributů je uvedeno v tabulce 2.

Tabulka 2: Srovnání ČR a EU u zvolených atributů ve věkové skupině 16 – 29 let

Zkoumaný atribut	EU	ČR
Používání Internetu (v %)	91,0	91,0
Používání počítače (v %)	79,0	88,0
Elektronická komunikace (v %)	86,0	91,0
Aktivita na sociální sítě (v %)	83,0	87,0
Prodej zboží a služeb na Internetu (v %)	20,0	21,0
Vyhledávání zboží a služeb na Internetu (v %)	75,0	84,0
Internetové bankovnictví (v %)	50,0	51,0
Vyhledávání informací o zdraví (v %)	51,0	40,0
Hovory a videohovory na Internetu (v %)	50,0	53,0
Vyhledávání informací o dopravě a ubytování (v %)	42,0	64,0
Reálný hrubý produkt na obyvatele (v €)	26 900	16 400
Hustota zalidnění (počet obyvatel / km^2)	117	137
Hodnota inflace (v %)	0,3	0,6
Dlouhodobá zaměstnanost (v %)	71,0	77,0
Nezaměstnanost (v %)	4,0	2,0

Zdroj: upraveno dle [8]

Grafické rozložení členských státu EU v rámci celé Evropy bylo vytvořeno v programu ArcMap a je zobrazeno na obrázku 1.



Obrázek 1: Mapa členských zemí EU

Zdroj: vlastní zpracování

1.3 Zkoumaná věková skupina

Definování zkoumané cílové skupiny, tj. mladých lidí, nejlépe vystihuje [34]. Pro pojem mládež uvádí: *„Věkově i sociálně jde o mezivrstvu mezi dětmi a dospělými, která má své specifické zájmy, aspirace, postoje, společenské postavení a svou roli, prestiž. Zvláštní status mládeže vyplývá z toho, že je ve stádiu neúplné nebo odložené ekonomické aktivity a profesionální přípravy, a že je sociálně a ekonomicky závislá na světě dospělých. Mládež je dynamická a vnitřně variabilní kategorie. V našich podmínkách se o mládeži uvažuje většinou od 14–15 let, kdy končí povinná školní docházka, do 30 let, kdy jsou završeny dílčí procesy sociálního zrání u všech skupin mládeže.“* [34, s. 635]

Naproti tomu dle statistik a databází ČSÚ [5] je rozložení věkové skupiny mladých lidí na pětileté kategorie, a to 15 – 19 let, 20 – 24 let a 25 - 29 let.

Databáze EUROSTAT [8] je rovněž zřejmě s ohledem na různé statistické výzkumy rozdělena na kategorie 16 – 19 let, 20 – 24 let a 25 - 29 let, ev. kategorii spojenou 16 – 29 let.

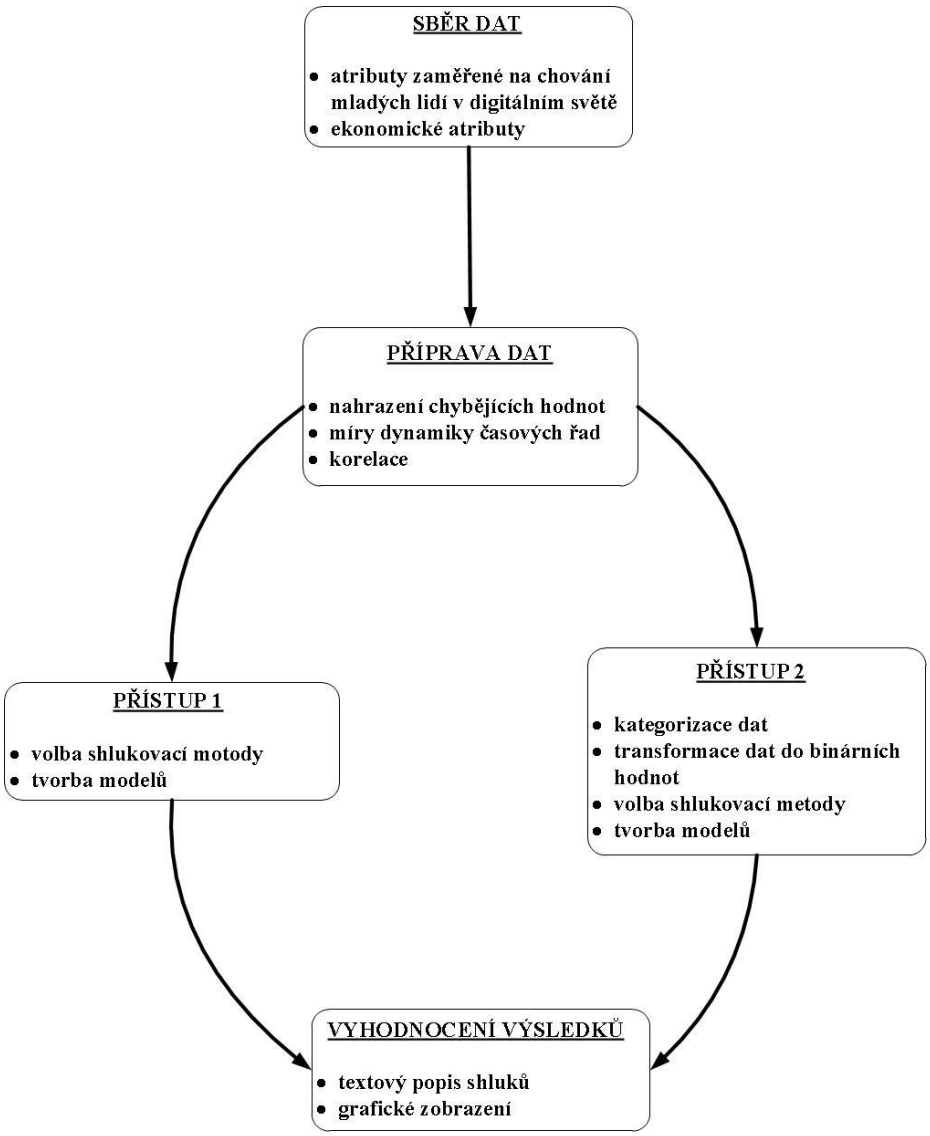
Věkové rozložení je tedy hodně podrobné. Pro zpracování diplomové práce bylo zvoleno věkové rozložení 16 – 29 let dle [8], a to hlavně s ohledem na potřebu získání dat o jednotlivých členských zemích EU.

2 FORMULACE PROBLÉMU

Cílem práce je získání dat o chování vybrané skupiny obyvatelstva, tj. mladých lidí, v digitálním světě. Data by měla obsahovat například údaje o využívání Internetu, aktivní účasti na sociálních sítích, nákupech a prodejích prostřednictvím Internetu, aby v následných výstupech z modelování mohly být získané výsledky porovnány a zhodnoceny.

Postupováno bude podle vybraných etap metodiky CRISP-DM, jejíž základní obecné charakteristiky jsou uvedeny v příloze B.

Schéma jednotlivých konkrétních kroků zpracování uvádí následující obrázek 2.



Obrázek 2: Schéma postupu zpracování

Zdroj: upraveno dle [22]

Metodou, která bude využita v rámci zpracování diplomové práce, je shluková analýza. Základním cílem shlukové analýzy dle [6] [17] [26] je zařadit objekty do skupin, které nazýváme shluky, a to tak, aby dva objekty stejného shluku si byly co nejvíce podobné. Přitom objekty mohou být různého charakteru. Shlukování patří mezi jeden ze základních typů získávání znalostí. Vstupem pro shlukování je datová matice. Výstupem je identifikace shluků. Objekty jsou do shluků shlukovány podle míry podobnosti či nepodobnosti.

Datová matice

Základem vícerozměrné statistické analýzy jsou m-rozměrná pozorování objektů. Počet těchto objektů je obvykle označován písmenem n. Vstupní matice je tedy rozměru n x m. Dalšími typy vstupní matice jsou dvourozměrná tabulka sdružených četností a matice vzdáleností. [17] [25]

Shluky

V dnešní době je nejčastější shlukování objektů i proměnných. Rozlišujeme shlukování disjunktivní, kdy objekty jsou jednoznačně přiřazeny do tříd, a překrývající, ve kterém lze objekt zařadit do více shluků. [20] [26]

Míra podobnosti

Cílem shlukování je vytvořit skupiny objektů tak, aby si objekty v jednom shluku byly co nejvíce podobné a co nejméně podobné objektům v ostatním shlucích. Míry podobnosti nabývají hodnot od nuly pro maximální rozdílnost až do jedničky pro totožnost. [17] [26]

Míra nepodobnosti

Metody shlukové analýzy jsou obvykle založeny na míře nepodobnosti nebo na vzdálenosti. Míra nepodobnosti je tedy stejný jev měřený v opačném směru. Dva body představující objekty jsou například charakterizovány délkou úsečky spojující tyto dva body. Čím je vzdálenost menší, tím jsou si body, které představují dané objekty, podobnější. [17] [26]

Eukleidovská vzdálenost je nejčastější vzdálenostní mírou. Tato míra vzdálenosti bývá též nazývána geometrická metrika. Výpočet této metriky je založen na Pythagorově větě. Platí, že vzdálenost (1):

$$d_E(x_k, x_l) = \sqrt{\sum_{j=1}^{m} \left(x_{kj} - x_{lj}\right)^2} . \qquad (1)$$

představuje standardní typ vzdálenosti. Vedle Eukleidovské vzdálenosti se užívá také čtverec Eukleidovské vzdálenosti. [20] [26]

Často užívanou vzdáleností je **Manhattanská vzdálenost**, zvaná také Hammingova metrika. Lze ji definovat vztahem (2):

$$d_H(x_k, x_l) = \sum_{j=1}^{m} \left| x_{kj} - x_{lj} \right| . \tag{2}$$

Před použitím této metriky se musíme ujistit, že žádné znaky spolu nekorelují. Když by znaky spolu korelovaly, shluky by byly nesprávné. [20] [26]

Hierarchické shlukování

U hierarchického shlukování rozlišujeme způsoby definice nepodobnosti shluků metodou nejbližšího souseda, metodou nejvzdálenějšího souseda a centroidní metodou. Na počátku tvoří každý objekt dané množiny objektů vlastní shluk, tento stav je označován jako nultý rozklad. Dále je nutné definování způsobu kvantitativního hodnocení podobnosti vztahů mezi shluky. Při prvním kroku shlukování pak vybereme dva shluky, které si jsou podle definice podobnosti shluků nejvíce podobné. Tyto dva shluky sloučíme a vytvoříme nový shluk. Nově vytvořený shluk se zbývajícími shluky, které obsahují vždy jeden objekt, tvoří první rozklad množiny objektů. Při dalších rozkladech se postupuje naprosto stejným způsobem. Postupně se nám tedy snižuje počet shluků. Tento postup tedy opakujeme tak dlouho, až dospějeme k poslednímu rozkladu o jediném shluku, obsahujícím všechny shlukované objekty. [17] [20] [36]

Tento typ shlukování rozlišuje dva přístupy. Přístup monetický a přístup polytetický. Monetický přístup je charakteristický tím, že se shluky na určité úrovni vytvářejí vždy pouze podle jedné proměnné. Všechny proměnné vždy zároveň se berou v úvahu naopak u polytetického přístupu. [17] [20] [26]

Dalším možným kritériem pro členění je to, zda se jedná o analýzu podobnosti či nepodobnosti. Analýza podobnosti, neboli aglomerativní přístup, vychází ze stavu, v němž každý objekt je samotným shlukem. Postupně se k sobě spojí dvojice shluků od nejvíce podobných k nejméně podobným. Výsledkem je poté jeden shluk. [11] [20] [26]

Divizní přístup, jak je nazývána analýza nepodobnosti, je založena na předpokladu, že jeden shluk je tvořen všemi objekty. Tento shluk je poté postupně rozdělován až do stavu, v němž je samostatným shlukem každý objekt. [17] [20] [26]

Při aglomerativním hierarchickém shlukování se vychází z toho, že na počátku každý objekt tvoří samostatný shluk. Postupuje se po krocích. V každém kroku se spojují dva nejpodobnější shluky podle matice podobnosti či nepodobnosti, která je stanovena pomocí různých algoritmů.

Mezi nejčastěji používané metody v praxi dle [17] [20] [26] patří např.:

- Metoda průměrné vazby (3):

$$D_{g<h,h'>} = \frac{n_h}{n_h + n_{h'}} D_{gh} + \frac{n_{h'}}{n_h + n_{h'}} D_{gh'} \ . \tag{3}$$

- Při metodě nejbližšího souseda je vzdálenost shluků dána minimální vzdáleností objektů z různých shluků (4):

$$D_{g<h,h'>} = \frac{1}{2}\left(D_{gh} + D_{gh'} - \left|D_{gh} - D_{gh'}\right|\right). \tag{4}$$

- U metody nejvzdálenějšího souseda je určující maximální vzdálenost objektů (5):

$$D_{g<h,h'>} = \frac{1}{2}\left(D_{gh} + D_{gh'} + \left|D_{gh} - D_{gh'}\right|\right). \tag{5}$$

- Centroidní metoda počítá vzdálenost mezi shluky jako euklidovskou vzdálenost mezi jejich centroidy, což jsou vektory aritmetických průměrů počítaných pro všechny objekty obsažené ve shluku (6):

$$D_{g<h,h'>} = \frac{n_h}{n_h+n_{h'}} D_{gh} + \frac{n_{h'}}{n_h+n_{h'}} D_{gh'} - \frac{n_h n_{h'}}{\left(n_h+n_{h'}\right)^2} D_{hh'} \ . \tag{6}$$

- Medianová metoda má postup podobný jako centroidní metoda. Rozdíl je v tom, že jsou brány v úvahu velikosti shluků neboli počet jejich prvků (7):

$$D_{g<h,h'>} = \frac{1}{2} D_{gh} + \frac{1}{2} D_{gh'} - \frac{1}{4} D_{hh'} \ . \tag{7}$$

- Wardova metoda spojuje shluky, u nichž je přírůstek celkového vnitroskupinového nového součtu čtverců odchylek jednotlivých hodnot od shlukového průměru minimální (8):

$$D_{g<h,h'>} = \frac{\left(n_h+n_g\right)D_{gh} + \left(n_{h'}+n_g\right)D_{gh'} - n_g D_{hh'}}{n_h+n_{h'}+n_g} \ . \tag{8}$$

Shlukovací postup znázorňuje graf zvaný dendrogram. Dendrogram je dle [17] [18] [20] stromový diagram znázorňující postup shlukování jednotlivých objektů a shluků vytvořených v předchozích krocích.

Horizontální dendrogram má objekty uvedeny na ose Y, vertikální má objekty umístěny naopak na ose X. Objekty tvoří listy, z kterých vycházejí větve. Do jedné větve se spojí vždy větve, které mají mezi sebou nejmenší vzdálenost. Hladinou spojení je právě vzdálenost mezi dvěma shluky. Další postup spojování větví odpovídá postupu spojování shluků při aglomerativním shlukování. [17] [20] [26]

Snad nejvíce matoucí je dosažení konečného počtu shluků, jež se také nazývá terminační kritérium. Neexistuje objektivní způsob určení tohoto terminačního kritéria. Proto byla vyvinuta řada pomocných kritérií, které slouží k řešení tohoto problému. Jedná se například o relativně jednoduché vyšetření měr podobnosti mezi shluky v každém kroku, pokud míra podobnosti překročí předem definovanou velikost, nebo když následné hodnoty se skokově změní. [17] [18] [20]

3 PŘÍPRAVA DAT

Tato kapitola uvádí postup zpracování získaných dat až po výslednou datovou matici, která je následně vstupem do datového modelu.

Použitá data pochází z databáze EUROSTATU [8]. Atributy popisují život mladých lidí jednotlivých zemí EU (celkem 28 států) v digitálním světě na základě vybraných ukazatelů, které přibližuje datový slovník, jehož ukázka je na obrázku 3. Celý datový slovník je uveden v příloze C. Mezi sledované atributy jednotlivých států patří například používání Internetu a počítače nebo jaké služby na Internetu využívají. Data byla doplněna o významné ekonomické ukazatele jednotlivých zemí EU, jakými jsou hrubý domácí produkt, hustota zalidnění, inflace, zaměstnanost a nezaměstnanost. Hodnoty všech daných ukazatelů jsou zachyceny v letech 2011 až 2016. Při kontrole kvality získaných dat byla zjištěna skutečnost, že data nejsou úplná a je tedy vhodné chybějící hodnoty doplnit.

Sloupec	Název atributu	Typ	Rozsah	Popis atributu
A	Stát	set	[Objekt 1, …, Objekt 29]	Stát EU
B	A1	range	< 47 ; 99 >	Jednotlivci ve věku 16 až 29 let (v %), kteří používají Internet denně v letech 2011 až 2016.
C	A2	range	< 49 ; 94 >	Jednotlivci ve věku 16 až 29 let (v %), kteří používají počítač denně v letech 2011 až 2015.
D	A3	range	< 65 ; 99 >	Jednotlivci ve věku 16 až 29 let (v %), kteří používají přijímají/odesílají elektronickou poštu v letech 2012 až 2016.
E	A4	range	< 52 ; 95 >	Jednotlivci ve věku 16 až 29 let (v %), kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.
F	A5	range	< 2 ; 57 >	Jednotlivci ve věku 16 až 29 let (v %), kteří prodávají produkty nebo služby na Internetu v letech 2011 až 2016.
G	A6	range	< 36 ; 95 >	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu produkty nebo služby v letech 2011 až 2016.
H	A7	range	< 3 ; 92 >	Jednotlivci ve věku 16 až 29 let (v %), kteří používají internetové bankovnictví v letech 2011 až 2016.
I	A8	range	< 24 ; 80 >	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu informace o zdraví v letech 2011, 2013, 2015 a 2016.
J	A9	range	< 18 ; 85 >	Jednotlivci ve věku 16 až 29 let (v %), kteří uskutečňují na Internetu hovory a videohovory v letech 2011 až 2016.
K	A10	range	< 13 ; 69 >	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu informace o cestování a ubytování v letech 2011 až 2016.
L	A11	range	< 5300 ; 83700 >	Reálný hrubý domácí produkt na obyvatele v letech 2011 až 2016.
M	A12	range	< 17,1 ; 1369,5 >	Počet obyvatel na km2 v letech 2011 až 2016.

Obrázek 3: Ukázka datového slovníku původních dat

Zdroj: vlastní zpracování

3.1 Nahrazení chybějících hodnot

V časové řadě se může stát, že nějaké pozorování chybí. Před zahájením dalších výpočtů je vhodné tyto hodnoty doplnit. Doplněné údaje nejsou plnohodnotné.

Podle účelu transformace lze postupovat několika způsoby dle [2] [26] [27]:

- Nahradit chybějící hodnoty nulami. Používá se v případě, že o řadě nic nevíme anebo víme pouze to, že průměrný člen by měl být nulový.
- Využitím některé centrální charakteristiky souboru naměřených hodnot. Například aritmetickým průměrem nebo mediánem. Lze použít i centrální charakteristiku okolních bodů namísto celého souboru.
- Pro řady, které vykazují výraznou setrvačnost, se doporučuje nahrazení chybějící hodnoty lineární interpolací mezi sousedními body.
- Nahradit chybějící hodnoty regresí a trendem vhodné křivky.
- Nahrazení odhadem založeným na známém či odhadnutém modelu chování.

Dlouhodobé změny v průměrném chování časové řady odráží trend. Jedná se o obecnou tendenci vývoje zkoumaného jevu za dlouhé období. Je to výsledek dlouhodobě působících faktorů, které působí ve stejném směru. Trend může být rostoucí, klesající nebo v čase neměnný. Jednou z možností, jak kvantifikovat trend dle [14] [20], je model, který se označuje jako model lineárního deterministického trendu (9):

$$X_t = \alpha + \beta_t + u_t \,. \tag{9}$$

Parametr β charakterizuje přírůstek řady X_t, když se čas t změní o jednotku.

3.2 Dynamiky časové řady

Dynamiku časové řady lze kvantifikovat dle [1] [2] také pomocí měr dynamiky. Absolutní přírůstek je definován jako (10):

$$\Delta X_t = X_t - X_{t-1} \tag{10}$$

a udává změnu hodnoty časové řady v čase *t* ve srovnání s hodnot1ou v čase *t-1*. Po odečtení modelu lineárního trendu v čase *t-1* od modelu lineárního trendu v čase t získáme model ve tvaru (11):

$$X_t - X_{t-1} = \beta + e_t \quad , \mathrm{t} = 2,3, \dots , \mathrm{T}, \text{kde } e_t = (u_t - u_t - 1). \tag{11}$$

Odhad parametru β se získá metodou nejmenších čtverců. Je interpretován také jako průměrný absolutní přírůstek. V případě, že model není lineární, je vhodné model linearizovat například logaritmickou transformací. [1] [2]

Další možností dle [1] [2], jak charakterizovat dynamiku časové řady, jsou koeficienty růstu a relativní přírůstek. Koeficient růstu je definován jako (12):

$$k_t = \frac{X_t}{X_{t-1}} . \tag{12}$$

Geometrický průměr koeficientů růstu interpretujeme dle [1] [2] jako průměrný koeficient růstu. Relativní přírůstek je definován jako (13):

$$\delta_t = \frac{\Delta X_t}{X_{t-1}} = \frac{X_t - X_{t-1}}{X_{t-1}} = \frac{X_t}{X_{t-1}} - 1 . \tag{13}$$

Chybějící hodnoty uprostřed časové řady byly nahrazeny dle [1] [2] [20] jednou z možností – a to aritmetickým průměrem předcházejícího období a následujícího období.

Chybějící hodnota na začátku nebo na konci řady byla predikována pomocí trendové křivky dané časové řady. Byla volena taková trendová křivka, kde hodnota reziduálního součtu čtverců R^2 byla maximální. [1] [2] [20]

Ukázka datového slovníku po nahrazení chybějících hodnot je na obrázku 4 a celý datový slovník je uveden v příloze D.

Sloupec	Název atributu	Typ	Rozsah	Popis atributu
A	Stát	set	[Objekt 1, …, Objekt 29]	Stát EU.
B	A1	range	< 47 ; 99 >	Jednotlivci ve věku 16 až 29 let (v %), kteří používají Internet denně v letech 2011 až 2016.
C	A2	range	< 49 ; 94 >	Jednotlivci ve věku 16 až 29 let (v %), kteří používají počítač denně v letech 2011 až 2016.
D	A3	range	< 65 ; 99 >	Jednotlivci ve věku 16 až 29 let (v %), kteří používají přijímají/odesílají elektronickou poštu v letech 2011 až 2016.
E	A4	range	< 52 ; 95 >	Jednotlivci ve věku 16 až 29 let (v %), kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.
F	A5	range	< 2 ; 57 >	Jednotlivci ve věku 16 až 29 let (v %), kteří prodávají produkty nebo služby na Internetu v letech 2011 až 2016.
G	A6	range	< 36 ; 95 >	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu produkty nebo služby v letech 2011 až 2016.
H	A7	range	< 3 ; 92 >	Jednotlivci ve věku 16 až 29 let (v %), kteří používají internetové bankovnictví v letech 2011 až 2016.
I	A8	range	< 24 ; 80 >	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu informace o zdraví v letech 2011 až 2016.
J	A9	range	< 18 ; 85 >	Jednotlivci ve věku 16 až 29 let (v %), kteří uskutečňují na Internetu hovory a videohovory v letech 2011 až 2016.
K	A10	range	< 13 ; 69 >	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu informace o cestování a ubytování v letech 2011 až 2016.
L	A11	range	< 5300 ; 83700 >	Reálný hrubý domácí produkt na obyvatele v letech 2011 až 2016.
M	A12	range	< 17,1 ; 1372,2 >	Počet obyvatel na km2 v letech 2011 až 2016.

Obrázek 4: Ukázka datového slovníku dat po nahrazení chybějících hodnot

Zdroj: vlastní zpracování

3.3 Výpočet odvozených proměnných

V datech se již nevyskytují chybějící hodnoty. Byla zkoumána dynamika časových řad jednotlivých atributů dle [20] a interpretována pomocí průměrného koeficientu růstu. Průměrné koeficienty růstu byly vypočteny podle vzorce (14):

$$\bar{k} = \sqrt[n-1]{k_2 k_3 \dots k_n} = \sqrt[n-1]{\frac{y_n}{y_1}} \ . \tag{14}$$

Následně byl vytvořen nový datový slovník. Ukázka datového slovníku koeficientů průměrných hodnot je zobrazena na obrázku 5 a celý datový slovník je uveden v příloze E.

Sloupec	Název atributu	Typ	Rozsah	Popis atributu
A	Stát	set	[Objekt 1, …, Objekt 29]	Stát EU.
B	A1	range	< 0,99125 ; 1,09206 >	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří používají Internet denně v letech 2011 až 2016.
C	A2	range	< 0,95479 ; 1,08388 >	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří používají počítač denně v letech 2011 až 2016.
D	A3	range	< 0,97501 ; 1,0296 >	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří používají přijímají/odesílají elektronickou poštu v letech 2011 až 2016.
E	A4	range	< 0,98876 ; 1,0702 >	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.
F	A5	range	< 0,83255 ; 1,2686 >	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří prodávají produkty nebo služby na Internetu v letech 2011 až 2016.
G	A6	range	< 0,94294 ; 1,06608 >	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu produkty nebo služby v letech 2011 až 2016.
H	A7	range	< 0,99041 ; 1,12888 >	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří používají internetové bankovnictví v letech 2011 až 2016.
I	A8	range	< 0,93053 ; 1,1487 >	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu informace o zdraví v letech 2011 až 2016.
J	A9	range	< 0,94409 ; 1,22176 >	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří uskutečňují na Internetu hovory a videohovory v letech 2011 až 2016.
K	A10	range	< 0,91967 ; 1,05436 >	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu informace o cestování a ubytování v letech 2011 až 2016.
L	A11	range	< 0,96901 ; 1,03112 >	Průměrný koeficient růstu reálného hrubé domácího produktu na obyvatele v letech 2011 až 2016.

Obrázek 5: Ukázka datového slovníku s odvozenými proměnnými

Zdroj: vlastní zpracování

3.3.1 Korelace

Před samotným zahájením shlukové analýzy se dle [20] doporučuje provést korelační analýzu a ověřit, zda neexistují mezi atributy závislosti, které by shlukování mohly ovlivnit.

Základní mírou podobnosti dvou objektů či znaků vyjádřených jako kvantitativní data může být Pearsonův korelační koeficient (korelační koeficient). Čím je jejich korelační koeficient větší a bližší jedničce, tím jsou si objekty podobnější. V případě pořadových čísel je mírou podobnosti Spearmanův korelační koeficient. [20] [23]

Vzájemné korelace atributů přibližuje obrázek 6.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
A1	1														
A2	0,817753	1													
A3	0,278222	0,193766	1												
A4	0,752112	0,693979	0,383474	1											
A5	0,048197	0,018286	0,124683	0,310209	1										
A6	0,227741	0,085074	0,45926	0,339303	0,24427	1									
A7	0,470316	0,461926	0,045435	0,463338	0,008052	0,342325	1								
A8	-0,21398	-0,20653	0,245936	0,227898	0,492182	0,211456	-0,23145	1							
A9	0,20446	-0,0461	0,1719	0,143322	0,010547	0,173895	-0,03021	-0,25447	1						
A10	-0,0589	-0,11869	-0,0322	0,271208	0,085063	0,27699	-0,07188	0,652539	-0,01488	1					
A11	0,080528	0,253381	-0,4069	0,056984	-0,10305	0,203565	0,546081	-0,18897	-0,1697	0,001324	1				
A12	-0,46986	-0,65643	0,05151	-0,23616	-0,10024	0,089622	-0,29724	0,413934	0,173251	0,426225	-0,28656	1			
A13	-0,29694	-0,26083	-0,27251	-0,19308	-0,11859	-0,21241	-0,47755	0,404096	-0,1933	0,647217	-0,06046	0,211126	1		
A14	-0,07058	0,142575	-0,37809	-0,01139	0,003732	0,138895	0,447101	-0,0465	-0,2864	0,12895	0,800866	-0,29444	0,166974	1	
A15	0,194037	-0,02792	0,247989	0,067898	-0,02585	-0,23891	-0,35713	-0,12873	0,286411	-0,20377	-0,72024	0,302631	-0,24264	-0,85552	1

Obrázek 6: Korelace atributů

Zdroj: vlastní zpracování

Zvýrazněny byly hodnoty, kdy korelační koeficient přesahuje hodnotu 0,8 v absolutní hodnotě. Následně byla vytvořena tabulka korelovaných atributů. Vzhledem k hodnotě korelačního koeficientu a faktické podobnosti byly vyloučeny atributy používání počítače (A2) a dlouhodobá zaměstnanost (A14). Hodnoty korelačního koeficientu u nejvíce korelovaných jsou znázorněny na obrázku 7.

Atribut 1	Atribut 2	Korelační koeficient
A1	A2	0,818
A14	A11	0,801
A14	A15	-0,809

Obrázek 7: Korelované atributy

Zdroj: vlastní zpracování

Současně byl z datové matice odstraněn případ průměrných hodnot EU, který bude dále sloužit pouze jako případ pro srovnání při vyhodnocování jednotlivých modelů. Výsledný rozměr datové matice je 14x28 (proměnné x případy).

4 MODELOVÁNÍ

Tato kapitola obsahuje výběr modelovacích technik, nastavení základních parametrů modelů, vytvoření modelů, jejich popis a vyhodnocení vybraných. Jako modelovací technika bylo zvoleno hierarchické shlukování.

4.1 Tvorba modelů

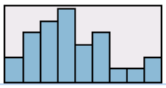
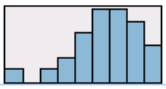
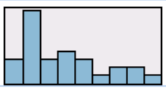
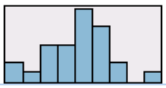
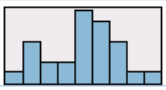
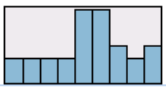
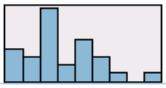
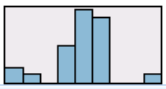
Shluková analýza byla provedena v softwaru IBM SPSS Statistics 19. Obrázek 8 níže nám zobrazuje ukázku nahraných dat do prostředí softwaru.

Stat	A1	A2	A3	A4	A5
Belgie	1,0208030	1,00213370	1,00000000	1,0525193	1,08018520
Bulharsko	1,0325882	1,02261570	,99570800	1,0376769	1,03713730
Česká republika	1,0508884	1,04833870	1,00535990	1,0701028	1,02021840
Dánsko	1,0106407	,98270900	1,00124170	1,0086406	1,04841320
Estonsko	1,0240977	1,01024260	1,00908460	1,0554676	1,09565420
Finsko	1,0127159	,97146030	1,00146470	1,0138942	1,01924490
Francie	1,0290337	,98979380	,98995690	1,0027249	,96285890
Chorvatsko	1,0276916	,99324260	1,02760040	1,0417063	1,26860370
Irsko	1,0362377	1,01054450	,99674100	1,0380827	,93985380
Itálie	1,0270661	,99475040	,99399847	1,0393944	,92909440
Kypr	1,0490091	1,03439350	1,00774430	1,0439612	1,18466450
Litva	1,0230125	1,01185820	,98197020	1,0335671	1,08447180
Lotyšsko	1,0201054	1,01047590	,99049914	,9912459	,94888010
Lucembursko	1,0215251	,96140300	1,00448060	1,0327795	,92626170
Maďarsko	1,0283468	1,00933620	,99889260	1,0232807	1,00000000
Malta	1,0349675	1,01509300	,98949593	1,0343935	,99346350
Německo	1,0177494	,98803150	1,00106730	1,0023420	,99324260

Obrázek 8: Ukázka dat v IBM Statistics 14.1

Zdroj: vlastní zpracování

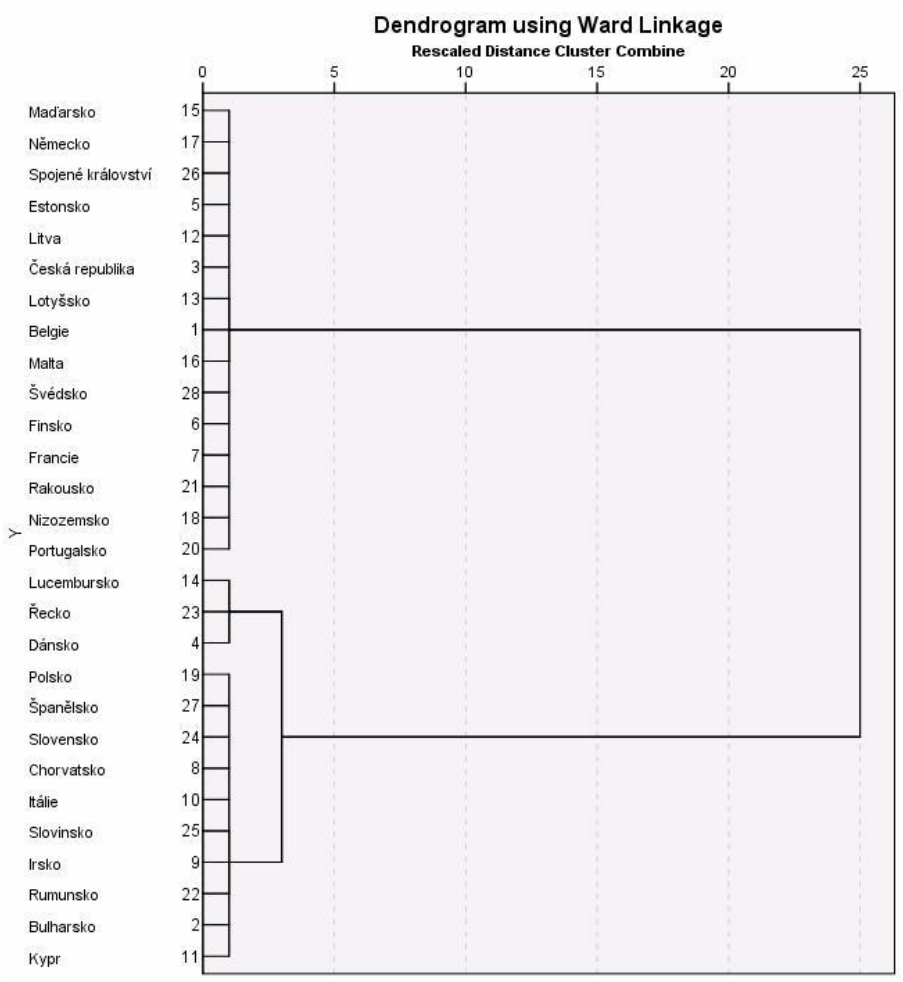
Dále byla data zobrazena v IBM SPSS Modeler 14.1, kde byla zpracována popisná statistika. Výstup z uzlu Data Audit nám přibližuje obrázek 9.

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
A5		Continuous	0.833	1.269	1.020	0.107	0.665	--	28
A6		Continuous	0.943	1.066	1.023	0.026	-0.960	--	28
A7		Continuous	0.990	1.129	1.042	0.037	0.776	--	28
A8		Continuous	0.931	1.149	1.031	0.046	0.091	--	28
A9		Continuous	0.944	1.222	1.080	0.065	-0.186	--	28
A10		Continuous	0.920	1.054	0.993	0.034	-0.245	--	28
A11		Continuous	0.984	1.071	1.014	0.019	0.979	--	28
A12		Continuous	0.981	1.023	1.001	0.008	-0.232	--	28

Obrázek 9: Ukázka datového auditu

Zdroj: vlastní zpracování

Postupně byly voleny jednotlivé shlukovací algoritmy (např. metoda nejbližšího souseda, metoda, nejvzdálenějšího souseda, centroidní metoda, mediánová metoda) a k nim příslušné metriky (např. Euklidovská metrika, Euklidovská čtvercová metrika) dle [8]. Pomocí dendrogramu byl určen jako nejlepší výstup model (model 0) s nastavením Euklidovské čtvercové metriky a Wardovy metody shlukování. Na obrázku 10 je dendrogram modelu 0.



Obrázek 10: Dendrogram s inflací (model 0)

Zdroj: vlastní zpracování

Z obrázku 10 je patrné, že toto nastavení by rozdělilo data do tří shluků. Při zkoumání typických vlastností jednotlivých shluků byla zjištěna skutečnost, že státy byly zařazeny do shluků především na základě inflace (A13). Jeden shluk obsahoval státy s kladnou hodnotou inflace, druhý s nulovou inflací a třetí shluk s inflací zápornou, jak je uvedeno v tabulce 3.

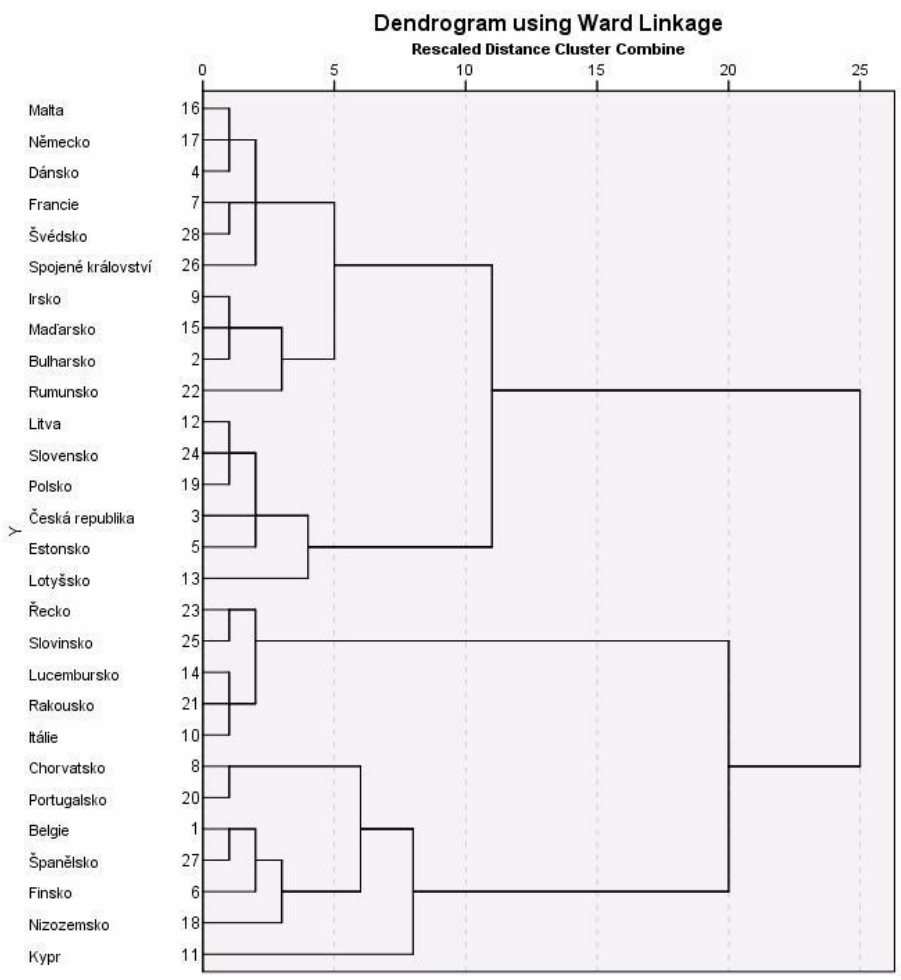
Tabulka 3: Inflace v jednotlivých členských státech EU

	Stát	Inflace v roce 2016 (v %)
Shluk 1	Bulharsko	-1,3
	Kypr	-1,2
	Chorvatsko	-0,6
	Rumunsko	-1,1
	Irsko	-0,2
	Slovensko	-0,5
	Španělsko	-0,3
	Slovinsko	-0,2
	Polsko	-0,2
	Itálie	-0,1
Shluk 2	Dánsko	0
	Lucembursko	0
	Řecko	0
Shluk 3	Lotyšsko	0,1
	Nizozemsko	0,1
	EU	0,3
	Maďarsko	0,4
	Finsko	0,4
	Francie	0,3
	Spojené království	0,7
	Estonsko	0,8
	Německo	0,4
	Portugalsko	0,6
	Litva	0,7
	Česká republika	0,6
	Rakousko	1,0
	Malta	0,9
	Belgie	1,8
	Švédsko	1,1

Zdroj: upraveno dle [18]

Inflace (A13) byla ze vstupů do shlukovací analýzy vyřazena a byly vytvořeny nové modely.

Opět byly postupně voleny jednotlivé shlukovací metody a k nim příslušné metriky. Dle rozložení dendrogramu vycházely nejlépe dva modely (model 1, model 2). Model 1 byl s nastavením Euklidovské čtvercové metriky a Wardovy shlukovací metody. Obrázek 11 zobrazuje dendrogram modelu 1 a následující obrázek 12 pak výpis jednotlivých kroků při spojování objektů. [13]



Obrázek 11: Model 1 (dendrogram)

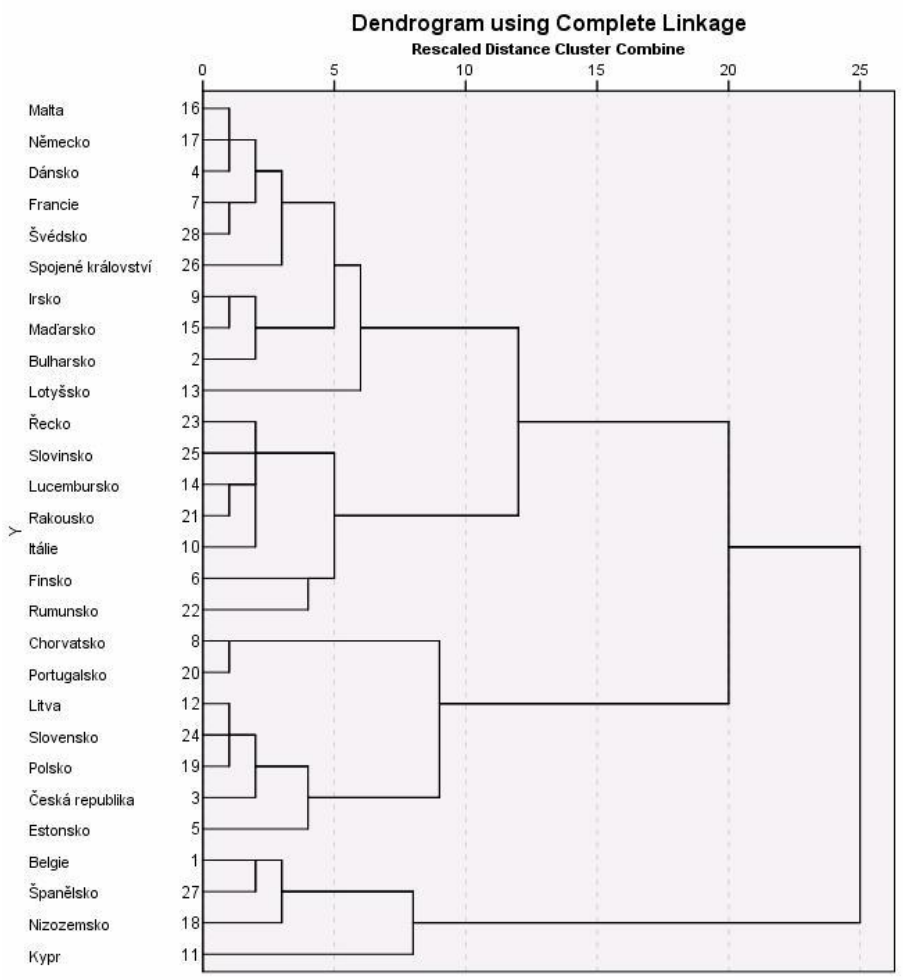
Zdroj: vlastní zpracování

Agglomeration Schedule						
	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	16	17	,003	0	0	4
2	7	28	,007	0	0	14
3	14	21	,012	0	0	12
4	4	16	,018	0	1	14
5	9	15	,024	0	0	11
6	8	20	,031	0	0	23
7	12	24	,037	0	0	8
8	12	19	,045	7	0	13
9	1	27	,055	0	0	17
10	23	25	,066	0	0	15
11	2	9	,077	0	5	19
12	10	14	,088	0	3	15
13	3	12	,102	0	8	18
14	4	7	,116	4	2	16
15	10	23	,133	12	10	26
16	4	26	,149	14	0	22
17	1	6	,168	9	0	20
18	3	5	,189	13	0	21
19	2	22	,213	11	0	22
20	1	18	,238	17	0	23
21	3	13	,274	18	0	25
22	2	4	,317	19	16	25
23	1	8	,374	20	6	24
24	1	11	,443	23	0	26
25	2	3	,540	22	21	27
26	1	10	,718	24	15	27
27	1	2	,946	26	25	0

Obrázek 12: Výpis jednotlivých kroků při spojování objektů (model 1)

Zdroj: vlastní zpracování

Model 2 byl vytvořen na základě metody nejvzdálenějšího souseda a Euklidovské metriky. Dendrogram shlukování při tomto nastavení modelu 2 znázorňuje obrázek 13. Výpis jednotlivých kroků při spojování objektů modelu 2 je na obrázku 14.



Obrázek 13: Model 2 (dendrogram)

Zdroj: vlastní zpracování

Agglomeration Schedule						
	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	16	17	,006	0	0	3
2	7	28	,008	0	0	10
3	4	16	,010	0	1	10
4	14	21	,010	0	0	13
5	9	15	,012	0	0	12
6	8	20	,013	0	0	24
7	12	24	,014	0	0	8
8	12	19	,015	7	0	14
9	1	27	,019	0	0	17
10	4	7	,021	3	2	16
11	23	25	,022	0	0	15
12	2	9	,023	0	5	20
13	10	14	,026	0	4	15
14	3	12	,028	0	8	18
15	10	23	,028	13	11	21
16	4	26	,039	10	0	20
17	1	18	,039	9	0	23
18	3	5	,046	14	0	24
19	6	22	,046	0	0	21
20	2	4	,057	12	16	22
21	6	10	,062	19	15	25
22	2	13	,071	20	0	25
23	1	11	,097	17	0	27
24	3	8	,110	18	6	26
25	2	6	,138	22	21	26
26	2	3	,238	25	24	27
27	1	2	,301	23	26	0

Obrázek 14: Výpis jednotlivých kroků při spojování objektů (model 2)

Zdroj: vlastní zpracování

4.2 Tvorba modelů na základě kategorizace dat

Následovala transformace dat do binárních proměnných. Transformace byla provedena v programu IBM SPSS Modeler 14.1 za pomocí uzlů Binning a SetToFlag. Pro převedení číselných hodnot na kategorie byl využit uzel Binning (znázorněno na obrázku 15), kde byla nastavena kategorizace do tří stejně početných skupin. Při převáděném počtu 28 proměnných bylo rozdělení 9, 9, 10. [15] [22]

Obrázek 15: Kategorizace hodnot

Zdroj: vlastní zpracování

Následovalo převedení do tvaru binárních hodnot. Byl použit uzel SetToFlag (znázorněno na obrázku 16), který slouží pro převod dat typu množina na typ příznak. Ze vstupních 13 atributů máme tedy najednou 39 atributů, kdy každý atribut je nyní binární. [22]

Obrázek 16: Převod do binárních hodnot

Zdroj: vlastní zpracování

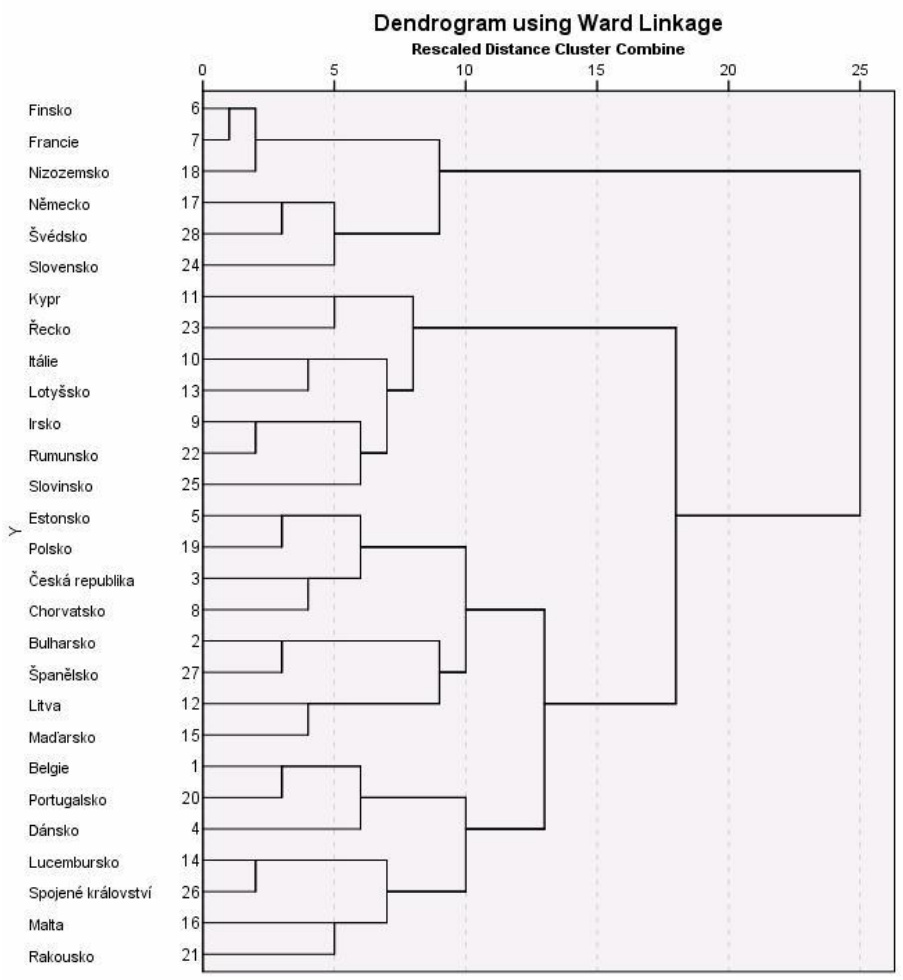
Po převedení dat do tvaru binárních hodnot byl vytvořen nový datový slovník, který je uveden v příloze F. Ukázku tohoto datového slovníku znázorňuje obrázek 17.

Sloupec	Název atributu	Typ	Rozsah	Popis atributu
A	Stát	set	[Objekt 1, …, Objekt 28]	Stát EU
B	A1_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají Internet denně v letech 2011 až 2016.
C	A1_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají Internet denně v letech 2011 až 2016.
D	A1_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají Internet denně v letech 2011 až 2016.
E	A2_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají počítač denně v letech 2011 až 2016.
F	A2_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají počítač denně v letech 2011 až 2016.
G	A2_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají počítač denně v letech 2011 až 2016.
H	A3_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají přijímají/odesílají elektronickou poštu v letech 2011 až 2016.
I	A3_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají přijímají/odesílají elektronickou poštu v letech 2011 až 2016.
J	A3_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají přijímají/odesílají elektronickou poštu v letech 2011 až 2016.
K	A4_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.
L	A4_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.

Obrázek 17: Ukázka datového slovníku dat převedených do binárních hodnot

Zdroj: vlastní zpracování

Takto vytvořená data byla nahrána do softwaru IBM SPSS Statistics 19 a bylo provedeno hierarchické shlukování. Z vytvořených modelů nejlepšího výsledku dosáhla opět Wardova metoda s čtvercovou Euklidovskou metrikou (model 3), jak znázorňuje obrázek 18 a výpis jednotlivých kroků při spojování objektů tohoto modelu 3, která je znázorněna na obrázku 19.



Obrázek 18: Model 3 (dendrogram)

Zdroj: vlastní zpracování

Agglomeration Schedule						
	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	6	7	2,000	0	0	2
2	6	18	5,333	1	0	22
3	14	26	9,333	0	0	18
4	9	22	13,333	0	0	16
5	17	28	18,333	0	0	12
6	2	27	23,333	0	0	21
7	1	20	28,333	0	0	15
8	5	19	33,333	0	0	17
9	12	15	39,333	0	0	21
10	10	13	45,333	0	0	19
11	3	8	51,333	0	0	17
12	17	24	58,333	5	0	22
13	11	23	65,333	0	0	20
14	16	21	72,333	0	0	18
15	1	4	80,000	7	0	24
16	9	25	88,000	4	0	19
17	3	5	96,000	11	8	23
18	14	16	105,500	3	14	24
19	9	10	115,100	16	10	20
20	9	11	125,643	19	13	26
21	2	12	137,143	6	9	23
22	6	17	149,143	2	12	27
23	2	3	161,893	21	17	25
24	1	14	175,012	15	18	25
25	1	2	190,743	24	23	26
26	1	9	212,333	25	20	27
27	1	6	242,357	26	22	0

Obrázek 19: Výpis jednotlivých kroků při spojování objektů (model 3)

Zdroj: vlastní zpracování

4.3 Vyhodnocení výsledků a jejich využití

Interpretace dendrogramu podobnosti jednotlivých států byla cílem shlukovací analýzy. Státy byly rozděleny do 3 shluků. Shluky jsou níže popsány a charakterizovány jejich typické vlastnosti. Nejdůležitější vlastnosti jednotlivých shluků přibližuje tabulka 4.

Tabulka 4: Typické vlastnosti shluků

Průměrný koeficient růstu - Wardova metoda (model 1)		
Shluk 1	Shluk 2	Shluk 3
HDP roste	roste oblíbenost elektronické komunikace	roste oblíbenost nabízení zboží a služeb na Internetu
nezaměstnanost klesá	nezaměstnanost roste	HDP klesá
Průměrný koeficient růstu - metoda nejvzdálenějšího souseda (model 2)		
Shluk 1	Shluk 2	Shluk 3
klesá vyhledávání informací o zdraví	roste využívání služeb internetového bankovnictví	HDP roste
oblíbenost nabízení zboží a služeb na Internetu má malé procentuální zastoupení	roste oblíbenost aktivit na sociálních sítích	nezaměstnanost klesá
Binární data - Wardova metoda (model 3)		
Shluk 1	Shluk 2	Shluk 3
nezvyšuje se počet uživatelů sociálních sítí	klesá hladina inflace	roste vyhledávání informací z oblasti cestovního ruchu
nezvyšuje se využívání internetového bankovnictví	klesá vyhledávání informací z oblasti cestovního ruchu	roste vyhledávání informací o zdraví

Zdroj: vlastní zpracování

4.3.1 Model 1

Z výsledného dendrogramu modelu 1 (obrázek 11) bylo usouzeno, že státy můžeme rozdělit do tří shluků.

Shluk 1

Do shluku byly zařazeny tyto státy: Bulharsko, Česká republika, Dánsko, Estonsko, Francie, Irsko, Litva, Lotyšsko, Maďarsko, Malta, Německo, Polsko, Rumunsko, Slovensko, Spojené království a Švédsko.

Shluk číslo 1 byl vytvořen především na základě podobnosti ekonomických atributů. Jsou zde zastoupeny především státy s nižší absolutní hodnotou HDP na jednoho obyvatele, než je průměr EU, který činí 26 900 EUR na obyvatele. Naproti tomu je koeficient růstu HDP u těchto států vysoký. V těchto zemích se nejvíce snižuje procento nezaměstnanosti ve sledovaném období, což vypovídá o celkovém ekonomickém růstu a zvyšování

produkčních možností ekonomik těchto států. Z geografického hlediska se nejedná o státy jednoznačně zařaditelné, naopak, rozptyl je po celém území EU.

Shluk 2

Do shluku byly zařazeny tyto státy: Itálie, Lucembursko, Rakousko, Řecko a Slovinsko.

Shluk je typický zvyšující se nezaměstnaností, ke které dochází zejména z důvodu zvýšení odchodu věku do důchodu a současně i přílivem pracovních sil ze zahraničí. Hodnota inflace se u těchto států pohybuje nejčastěji okolo nulové hodnoty. V hodnotě ukazatele HDP na obyvatele patří do průměru EU, který činí 26 900 EUR na obyvatele, jeho hodnota u států zařazených do tohoto shluku je konstantní nebo mírně klesá.

Z atributů charakterizujících internetové aktivity je pak nejvíce zastoupena zvyšující se hodnota atributu komunikace prostřednictvím emailové schránky a vyhledávání informací o zdraví na Internetu.

Shluk 3

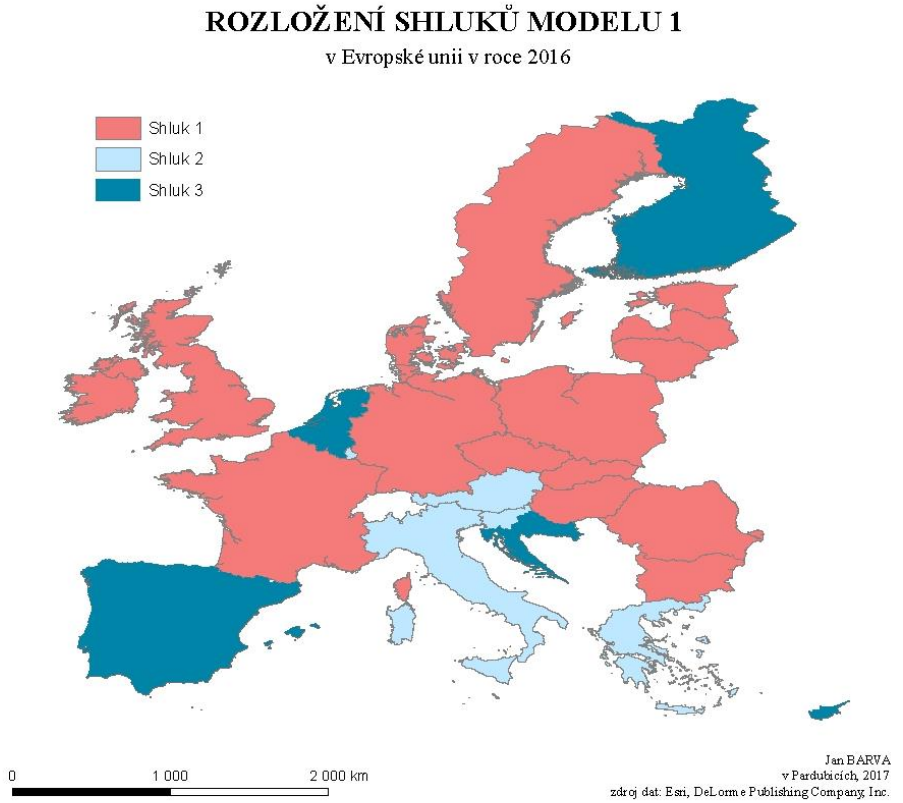
Do shluku byly zařazeny tyto státy: Belgie, Finsko, Chorvatsko, Kypr, Nizozemsko, Portugalsko a Španělsko.

Shluk 3 představují státy, kde se v obecném hledisku nechá říci, že mladí lidé čím dál více objevují výhody internetových technologií, které přestavuje mailová a elektronická komunikace, využívání sociálních sítí a i prodávání věcí a služeb na Internetu, stejně tak vyhledávání informací o zdravém životním stylu. O čím dál větší dostupnosti internetových služeb svědčí i zvyšující se procento mladých lidí uskutečňujících hovory či videohovory prostřednictvím Internetu.

Z ekonomických ukazatelů je pro tento shluk typický zvyšující se průměrný koeficient růstu u nezaměstnanosti, avšak v hodnotě ukazatele HDP na obyvatele patří do průměru až nadprůměru EU, jeho hodnota je konstantní nebo mírně klesá.

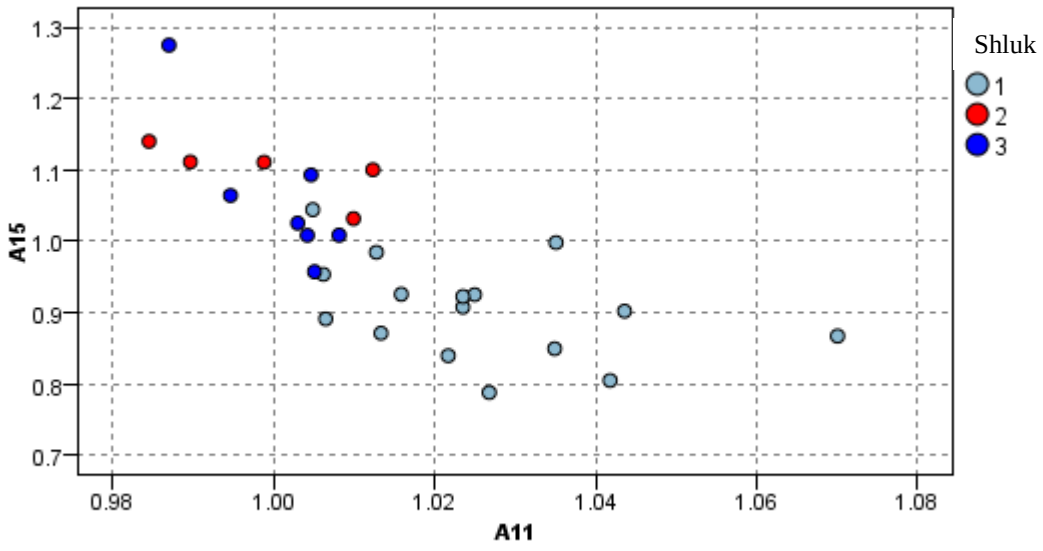
Rozložení do jednotlivých shluků v geografickém vyjádření zpracovaném v programu ArcMap je na obrázku 20.

Na následujícím obrázku 21 je ukázána závislost vlivů proměnné nezaměstnanost (A15) a HDP (A11) v rámci jednotlivých států. Jako další parametr pro zobrazení je příslušnost objektu ke shluku v barevném provedení bodů.



Obrázek 20: Rozložení shluků modelu 1

Zdroj: vlastní zpracování



Obrázek 21: Grafické zobrazení modelu 1

Zdroj: vlastní zpracování

V rámci tohoto modelu je ČR zařazena do shluku 1. Typickými vlastnostmi tohoto shluku je růst HDP a pokles nezaměstnanosti. V porovnání s průměrnými hodnotami EU dosahuje však ČR v případě HDP pouze 60% hodnoty průměru EU, ale naproti tomu u dlouhodobé nezaměstnanosti nedosahuje ani poloviční hodnoty průměru EU. HDP v ČR zvolna každoročně roste.

4.3.2 Model 2

Dendrogram modelu 2 (obrázek 13) nám přibližuje, že existují tři skupiny států s podobnými vlastnostmi.

Shluk 1

Do shluku byly zařazeny tyto státy: Bulharsko, Dánsko, Finsko, Francie, Irsko, Itálie, Lotyšsko, Lucembursko, Maďarsko, Malta, Německo, Rakousko, Rumunsko, Řecko, Slovinsko, Spojené království a Švédsko.

Tento shluk není typický ekonomickými atributy. Z geografického hlediska se rovněž jedná o shluk bez typického zastoupení určením geografické polohy jednotlivých států. Z atributů digitálního světa je typický atribut nabízení produktů a služeb na Internetu mladými lidmi, jehož hodnota výrazně klesá a má velmi malé procentuální zastoupení. Mladí lidé v tomto shluku zdaleka tak nevyhledávají informace cílené na otázku zdraví.

Shluk 2

Do shluku byly zařazeny tyto státy: Česká republika, Estonsko, Chorvatsko, Litva, Polsko, Portugalsko a Slovensko.

Shluk číslo 2 zpracovaný metodou nejvzdálenějšího souseda je typický svým rostoucím trendem počtu mladých lidí téměř u všech sledovaných atributů charakterizujících internetové aktivity. Růst je zaznamenán u používání počítače, používání emailové schránky, aktivity na sociálních sítích, nabízení zboží a služeb, využívání služeb internetového bankovnictví, vyhledávání informací v oblasti zdravotnictví a cestovního ruchu. Naopak uskutečňování hovorů přes Internet je na ústupu, což může být ovlivněno dostupností hlasových tarifů jednotlivý mobilních operátorů.

Státy zařazené do této skupiny zvyšují svojí ekonomickou výkonost, ale jejich hodnota HDP stále zůstává pod průměrem EU. Dochází zde k mírnému úbytku obyvatelstva ovlivněné pozdějším věkem pro početí dětí. Hodnota nezaměstnanosti se snižuje, což svědčí o ekonomickém růstu při současné malé změně hodnoty inflace.

Shluk 3

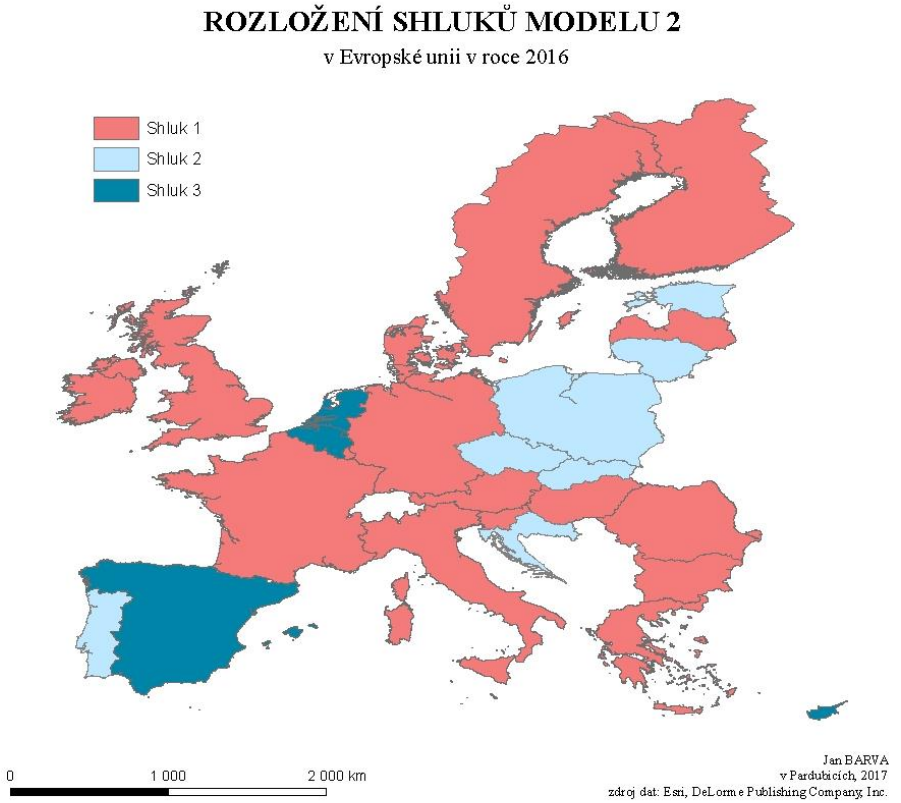
Do shluku byly zařazeny tyto státy: Belgie, Kypr, Nizozemsko a Španělsko.

Státy zařazeny v tomto shluku jsou v jednom typické, byť se nejednalo o zkoumaný atribut. Přelidněnost se vyjadřuje jako nedostatek prostoru na osobu. Jedná se tedy o míru přelidnění domácností, která je v těchto státech nejnižší. Rovněž jsou zde zastoupeny státy s nejnižšími hodnotami zalidnění. I to může mít vliv na velmi populární atribut, v této skupině států atribut s rostoucí tendencí, který představuje nabízení a prodej zboží a služeb přes digitální svět Internetu. Výkonnost ekonomik v těchto státech stagnuje či mírně klesá, zároveň dochází ke zvyšování míry nezaměstnanosti ve sledovaném období.

Model 2 zařadil ČR do shluku, kde typickou vlastností byl růst využívání služeb internetového bankovnictví a růst oblíbenosti aktivit na sociálních sítích. V posledních letech v ČR došlo k rozšíření služeb poskytovaných bankovními společnostmi, jehož důsledkem je právě větší možnost a rozšíření oblíbenosti internetového bankovnictví. V porovnání s EU se ČR pohybuje v jejím průměru. V případě aktivit na sociálních sítích je průměrná hodnota EU 83%, ČR pak má dokonce 87%. Jedná se o oblíbenost u mladých lidí, proto jsou procentuální vyjádření takto vysoká.

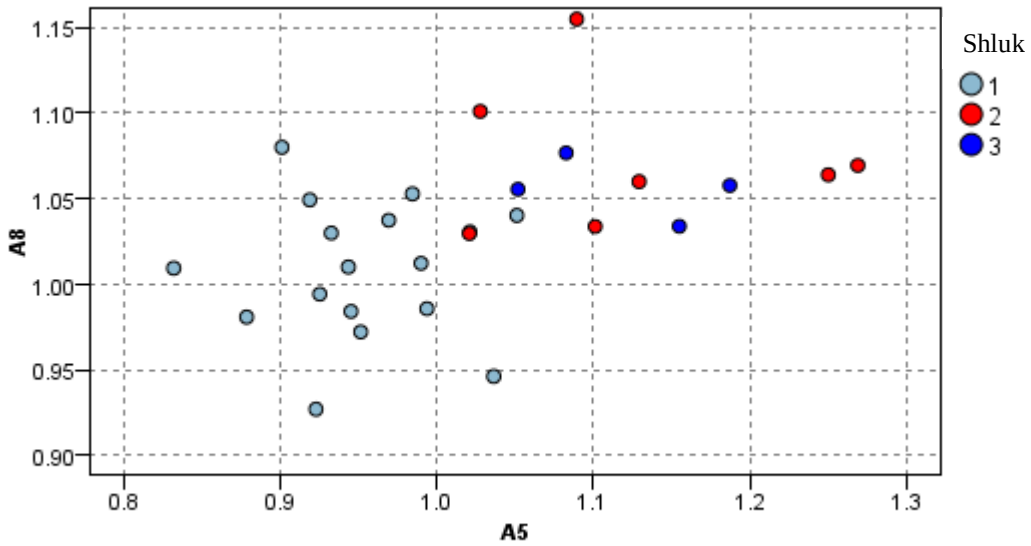
Výstižné grafické zobrazení rozložení shluků v rámci států EU je znázorněno na obrázku 22.

Na následujícím obrázku 23 je ukázána závislost vlivů vyhledávání informací na Internetu o zdraví (A8) a nabízení a prodej zboží a služeb na Internetu (A5) v rámci jednotlivých států. Jako další parametr pro zobrazení je příslušnost objektu ke shluku v barevném provedení bodů.



Obrázek 22: Rozložení shluků modelu 2

Zdroj: vlastní zpracování



Obrázek 23: Grafické zobrazení modelu 2

Zdroj: vlastní zpracování

4.3.3 Model 3

Tento model pracoval na vstupu s binárními proměnnými. I model 3 podle dendrogramu (obrázek 18) rozdělil státy do tří shluků.

Shluk 1

Do shluku byly zařazeny tyto státy: Finsko, Francie, Německo, Nizozemsko, Slovensko a Švédsko.

Do shluku 1 zpracovaného z binárních vstupních dat byly zařazeny státy s rostoucí zalidněností, tj. s přírůstkem obyvatel, což může být ovlivněno například i migrací. Používání Internetu již není na vzestupu, zůstává spíše na konstantní úrovni. Mladí lidé již tolik nevyužívají počítač a komunikaci elektronickou poštou. Rovněž tak používání sociálních sítí, vyhledávání informací o zdraví či informací spojených s cestovním ruchem a internetové bankovnictví zůstává na stejné úrovni, na průměrných hodnotách EU jako celku. Stagnaci či mírný pokles evidujeme u počtu hledajících i prodávajících na Internetu zboží a služby.

Shluk 2

Do shluku byly zařazeny tyto státy: Irsko, Itálie, Kypr, Lotyšsko, Rumunsko, Řecko a Slovinsko.

Ve shluku číslo 2 jsou zastoupeny státy se zápornou inflací a stoupající nezaměstnaností, což rozhodně neprospívá ekonomikám těchto států. V případě těchto jevů z makroekonomického hlediska mluvíme o hospodářské krizi. Hodnota zalidnění je konstantní. Naopak používání počítače a Internetu je na vzestupu. Boom v těchto zemích zažívají sociální sítě, kde se registruje a aktivně účastní čím dál více mladých lidí. Na podprůměr EU poklesem se dostává návštěvnost stránek s tématikou zdraví, cestování a prodejní portály. Což opět je zřejmě důsledkem finanční nestability těchto států.

Shluk 3

Do shluku byly zařazeny tyto státy: Belgie, Bulharsko, Česká republika, Dánsko, Estonsko, Chorvatsko, Litva, Lucembursko, Maďarsko, Malta, Polsko, Portugalsko, Rakousko, Spojené království a Španělsko.

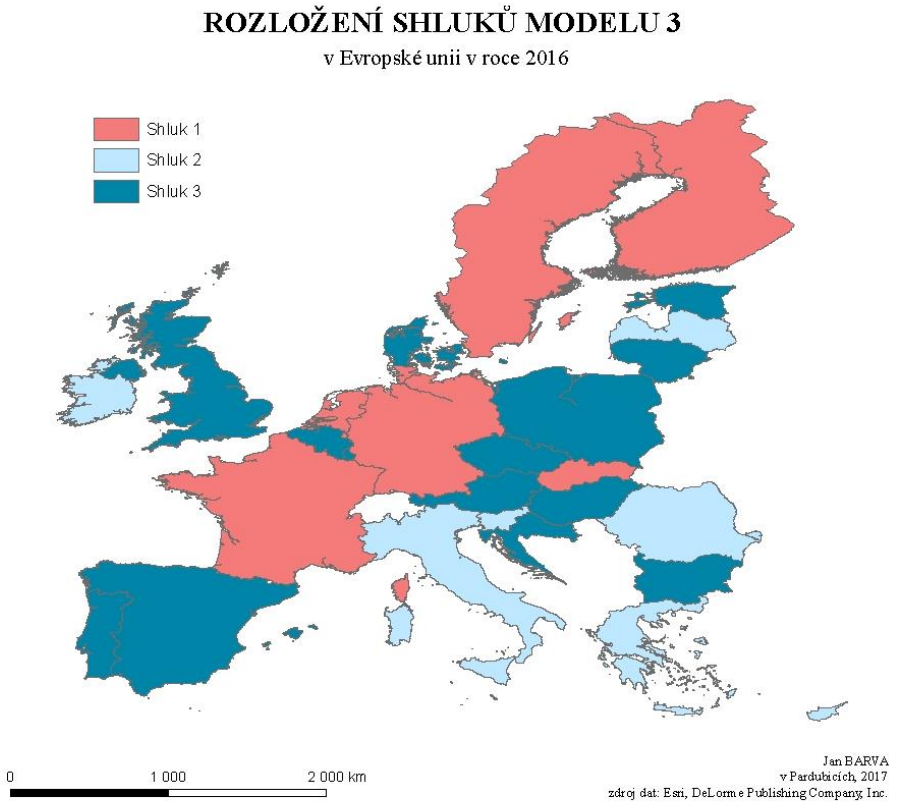
Státy, které Wardova metoda přiřadila do třetího shluku, mají vyšší hodnotu inflace než je průměr EU, ale se zvolna klesajícím trendem. Z hlediska ekonomického postavení je v těchto státech vykazován hospodářský růst. Nad průměr EU se u mladých lidí dostává v prostředí digitálního světa vyhledávané informace o zdraví a cestovní ruch. Zvyšuje se procento

zastoupení vyhledávajících informace o zboží a službách, nebo je prodávají. Aktivní účast na sociálních sítích je na průměrné hodnotě EU.

Tento model zařadil ČR do shluku 3. Typickými vlastnostmi zde jsou růst vyhledávání informací z oblasti cestovního ruchu a informací o zdraví. V případě mladých lidí je zájem o cestování a vyhledávání informací o cestování jeden a půl krát vyšší než průměr v EU. Vliv na tento stav má jistě i stoupající životní úroveň v ČR.

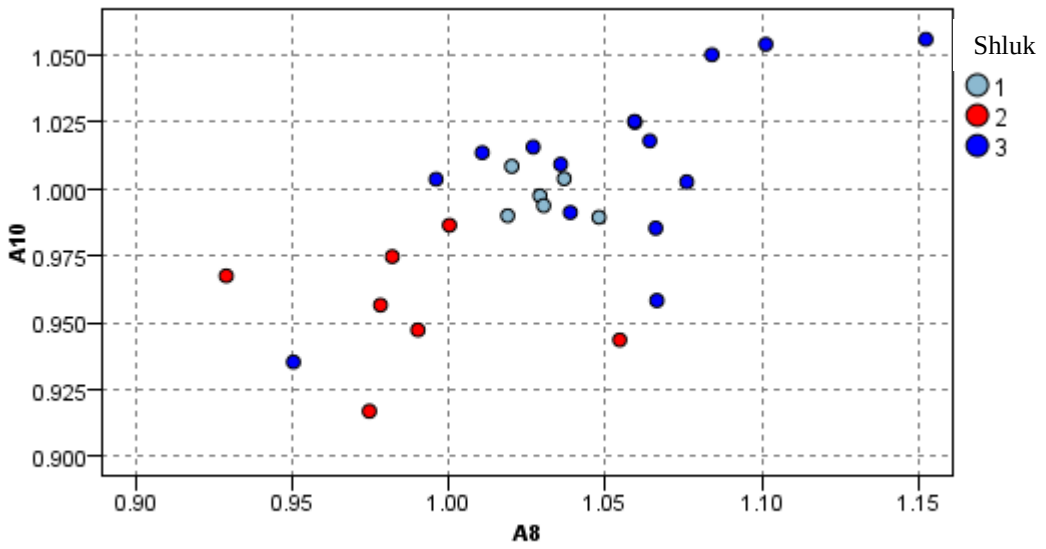
Výstižné grafické zobrazení rozložení shluků v rámci států EU je znázorněno na obrázku 24.

Na následujícím obrázku 25 je ukázána závislost vlivů vyhledávání informací na Internetu o cestování, ubytování (A10) a vyhledávání informací o zdraví (A8) v rámci jednotlivých států. Jako další parametr pro zobrazení je příslušnost objektu ke shluku v barevném provedení bodů.



Obrázek 24: Rozložení shluků modelu 3

Zdroj: vlastní zpracování



Obrázek 25: Grafické zobrazení modelu 3

Zdroj: vlastní zpracování

5 ZÁVĚR

Cíl diplomové práce byl naplněn, byly vytvořeny modely porovnávající život v digitálním světě mladých lidí žijících v České republice a v jednotlivých státech Evropské unie. Hodnoty jednotlivých vstupních atributů se vztahovaly k letům 2011 až 2016.

V rámci zpracování diplomové práce proběhl sběr dat. Data byla po důkladné analýze předzpracována, tj. došlo k nahrazení chybějících hodnot, kategorizaci dat a k transformaci do binárních hodnot. Tato část zpracování diplomové práce byla co do časového rozsahu nejnáročnější, takto nakonec uvádí i metodika CRISP-DM. Samotné modelování a závěrečné porovnání jednotlivých výstupů z modelů, jak je uvedeno v kapitole 3.3, je provedeno jak popisnou, tak grafickou interpretací.

Dle výstupů z modelování se Česká republika nikterak nevymyká z průměrných hodnot zjištěných v Evropské unii jako celku, naopak je možné s potěšením konstatovat, že mladí lidé ve věku 16 až 29 let žijící v naší zemi jsou na základě svého chování v digitálním světě řazeni mezi vyspělejší státy. Je to zřejmě odraz dřívějších finančních injekcí do základního školství. Tyto finanční prostředky, většinou z různých dotačních titulů hrazených ze zdrojů Evropské unie, podpořily rozvoj technické a informační infrastruktury základních a středních škol, což se projevuje v dnešní době právě na skupině definované jako mladí lidé vyšší digitální gramotností a schopností vnímat stále se rozvíjející moderní informační technologie.

Toto pozitivní zjištění by však nemělo být jediným měřítkem. Na druhou stranu v některých státech Evropské unie (Finsko, Francie, Německo, Nizozemsko a Švédsko) již dochází k drobnému poklesu množství a frekvence času stráveného mladými lidmi na Internetu. Je to bez pochyby důsledek i toho, že mladí lidé dnes mají touhu žít svůj život i jinak, realizovat se i v osobním životě.

Celkové vnímání digitálního světa vede nejen mladou generaci k nutnosti stále se vzdělávat, pracovat na sobě, být adaptabilní, kreativní, vždyť takový je i svět informačních technologií, stále nový, mladý, rozvíjející se, rychle se měnící. Dnešní mladí lidé mají stále „při ruce“ veškeré informace, mobilní internetové připojení se stává téměř samozřejmostí u zkoumané skupiny lidí.

POUŽITÁ LITERATURA

[1] ARLT, Josef a Markéta ARLTOVÁ. *Ekonomické časové řady: [vlastnosti, metody modelování, příklady a aplikace]*. Praha: Grada, 2007. ISBN 978-80-247-1319-9.

[2] ARLT, Josef. *Moderní metody modelování ekonomických časových řad*. Praha: Grada Publishing, 1999. ISBN 978-80-7169-539-4.

[3] BERKA, Petr. *Dobývání znalostí z databází*. Praha: Academia, 2003. ISBN 80-200-1062-9.

[4] *Český statistický úřad: Počet obyvatel - Metodika* [online]. [cit. 2017-06-04]. Dostupné z: https://www.czso.cz/csu/czso/pocet_obyvatel_m

[5] *Český statistický úřad: Sestavení vlastní tabulky* [online]. [cit. 2017-06-04]. Dostupné z: https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=uziv-dotaz#k=5

[6] *Data mining with SPSS modeler: theory, exercises and solutions*. ISBN 978-3-319-28707-2.

[7] EuroMISE [online]. 2002 [cit. 2017-02-16]. Proces dobývání znalostí. Dostupné z:http://euromise.vse.cz/kdd/index.php?page=metody

[8] *Eurostat: Database* [online]. [cit. 2017-06-04]. Dostupné z: http://ec.europa.eu/eurostat/web/youth/data/database

[9] *Evropská unie: Jednotlivé země* [online]. [cit. 2017-06-04]. Dostupné z: https://europa.eu/european-union/about-eu/countries_cs#tab-0-1

[10] FIALA, Petr a Markéta PITROVÁ. *Evropská unie*. 2., dopl. a aktualiz. vyd. Brno: Centrum pro studium demokracie a kultury, 2009. ISBN 978-80-7325-223-6.

[11] HAN, Jiawei. a Micheline. KAMBER. *Data mining: concepts and techniques*. 2nd ed. San Francisco, CA: Morgan Kaufmann, c2006. ISBN 978-1-55860-901-3.

[12] HENDL, Jan. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Praha: Portál, 2004. ISBN 80-7178-820-1.

[13] CHEN, Chun-houh., Wolfgang. HÄRDLE a Antony. UNWIN. *Handbook of data visualization*. 2008. Berlin: Springer, c2008. ISBN 978-3-540-33036-3.

[14] CHRISTIAN ALBRIGHT, WAYNE L. WINSTON, CHRISTOPHER ZAPPE, S. Christian Albright, Wayne L. Winston, Christopher Zappe a WITH CASES BY

MARK BROADIE .. [ET AL.]. *Data analysis & decision making with Microsoft Excel*. 3rd ed., International student ed. Mason, Ohio: Thomson/South-Western, 2005. ISBN 0324400861.

[15] *IBM Knowledge Center* [online]. [cit. 2017-06-04]. Dostupné z: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/base/idh_webhelp_scatter_options_palette.html

[16] JIROVSKÝ, Václav. *Kybernetická kriminalita: nejen o hackingu, crackingu, virech a trojských koních bez tajemství*. Praha: Grada, 2007. ISBN 978-80-247-1561-2.

[17] LUKASOVÁ, Alena a Jana ŠARMANOVÁ. *Metody shlukové analýzy.* 1. vyd. Praha: SNTL, 1985. 210 s.

[18] MAIMON, Oded. a Lior. ROKACH. *Decomposition methodology for knowledge discovery and data mining: theory and applications*. London: World Scientific, c2005. ISBN 981-256-079-3.

[19] MAREK, Dan a Michael J. BAUN. *Česká republika a Evropská unie*. Brno: Barrister & Principal, 2010. ISBN 978-80-87029-89-3.

[20] MELOUN, Milan a Jiří MILITKÝ. *Kompendium statistického zpracování dat: metody a řešené úlohy včetně CD*. Praha: Academia, 2002. ISBN 80-200-1008-4.

[21] NONDEK, Lubomír a Lenka ŘENČOVÁ. *Internet a jeho komerční využití*. Praha: Grada, 2000. Manažer. ISBN 80-7169-933-0.

[22] PETR, Pavel. *Data Mining*. Vyd. 2. Pardubice: Univerzita Pardubice, 2008. ISBN 978-80-7395-098-9.

[23] PETR, Pavel. *Metody Data Miningu*. Pardubice: Univerzita Pardubice, 2014. ISBN 978-80-7395-872-5.

[24] RUD, Olivia Parr. *Data Mining: praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM)*. Praha: Computer Press, 2001. ISBN 80-7226-577-6.

[25] ŘEZANKOVÁ, Hana a Stanislava HRONOVÁ. *Statistická data*. Praha: Vysoká škola ekonomická, 2000. ISBN 80-245-0021-3.

[26] ŘEZANKOVÁ, Hana, Dušan HÚSEK a Václav SNÁŠEL. *Shluková analýza dat*. 2., rozš. vyd. Praha: Professional Publishing, 2009. ISBN 978-80-86946-81-8.

[27] ŘEZANKOVÁ, Hana. *Analýza dat z dotazníkových šetření*. 2. vyd. Praha: Professional Publishing, 2010. ISBN 978-80-7431-019-5.

[28] SKALSKÁ, Hana. *Data mining a klasifikační modely*. Hradec Králové: Gaudeamus, 2010. Recenzované monografie. ISBN 978-80-7435-088-7.

[29] SKLENÁK, Vilém. *Data, informace, znalosti a Internet*. Praha: C.H. Beck, 2001. C.H. Beck pro praxi. ISBN 80-7179-409-0.

[30] SMEJKAL, Vladimír. *Internet a §§§*. Praha: Grada, 2001. ISBN 80-247-0058-1.

[31] SMEJKAL, Vladimír. *Kybernetická kriminalita*. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, 2015. Pro praxi. ISBN 978-80-7380-501-2.

[32] SMEJKAL, Vladimír. *Právo informačních a telekomunikačních systémů*. Praha: C.H. Beck, 2001. Právo a hospodářství (C.H. Beck). ISBN 80-7179-552-6.

[33] SPITZER, Manfred. *Kybernemoc!: Jak nám digitalizovaný život ničí zdraví*. Přeložila Iva KRATOCHVÍLOVÁ. Brno: Host - vydavatelství, 2016. ISBN 978-80-7491-792-9.

[34] *Velký sociologický slovník*. Praha: Karolinum, 1996. ISBN 80-7184-164-1.

[35] VOLEJNÍKOVÁ, Jolana. *Ekonomie I pro SII*. Pardubice: Univerzita Pardubice, 2014. ISBN 978-80-7395-831-2.

[36] WITTEN, I. H. a Eibe. FRANK. *Data mining: practical machine learning tools and techniques*. 2nd ed. Boston, MA: Morgan Kaufman, 2005. ISBN 0-12-088407-0.

SEZNAM PŘÍLOH

Příloha A – Charakteristika členských států EU

Česká republika

ČR je vnitrozemský stát ve střední Evropě, který vznikl v roce 1993 rozdělením Československa na dvě země.

Nejdůležitějším odvětvím českého hospodářství je průmysl. Hlavními vývozními trhy ČR jsou Německo, Slovensko a Polsko. Česká republika nejvíce dováží z Německa, Polska a Číny.

Belgie

Belgie je zakládajícím členem a byla u prvních integračních uskupení. Patří k vyspělým státům Unie v oblasti práva, demokracie i ekonomiky. V Belgii, resp. v Bruselu, sídlí některé hlavní organizace EU jako například Evropský parlament a Evropská komise. Belgie je federální stát, který tvoří tři regiony: na severu nizozemsky mluvící Vlámsko, na jihu francouzsky mluvící Valonsko a hlavní město, kde mají statut úředního jazyka oba uvedené jazyky.

Nejdůležitějšími odvětvími belgického hospodářství jsou veřejná správa, obrana, vzdělávání, zdravotní, sociální péče. Hlavními vývozními trhy Belgie jsou Německo, Francie a Nizozemsko. Belgie dováží nejvíce z Nizozemska, Německa a Francie.

Bulharsko

Bulharsko leží v jihovýchodní části Balkánského poloostrova a má rozmanitý povrch. Severu země dominují rozlehlé nížiny. Jih země je naopak formován pohořími a náhorními plošinami. Na východě je černomořské pobřeží, které láká turisty.

Nejdůležitějším odvětvím bulharského hospodářství je průmysl. Hlavními vývozními partnery Bulharska jsou Německo, Itálie a Turecko. Bulharsko dováží nejvíce z Německa, Ruska a Itálie.

Dánsko

Dánsko je konstituční monarchií od roku 1849. Dánsko leží na Jutském poloostrově a na více jak čtyři sta pojmenovaných ostrovech. K Dánskému státu můžeme za jistých okolností počítat také Grónsko či Faerské ostrovy. Na jihu sousedí s Německem, se Švédskem jej spojuje dvoupatrový most pro auta a vlaky.

Nejdůležitějšími odvětvími dánského hospodářství jsou veřejná správa, obrana, vzdělávání, zdravotní a sociální péče. Hlavními vývozními trhy Dánska jsou Německo, Švédsko a Spojené státy. Dánsko nejvíce dováží z Německa, Švédska a Nizozemska.

Estonsko

Estonsko patří mezi státy severní Evropy a představuje nejsevernější stát takzvaných pobaltských republik, které mají společnou minulost v podobě "vazalských" států SSSR. Estonsko, respektive Estonci, mají velmi blízko ke skandinávským národům - nejen historicky, ale i kulturně a mentalitou. Historie Estonska sahá přibližně do 12. století n.l.

Nejdůležitějšími odvětvími estonského hospodářství jsou velkoobchod, maloobchod, doprava, ubytovací a stravovací služby. Hlavními vývozními partnery Estonska jsou Švédsko, Finsko a Lotyšsko. Estonsko dováží nejvíce z Finska, Německa a Litvy.

Finsko

Finsko je jednou z pěti severských zemí. Je nejsevernější zemí EU. Finsko je jednou z nejřidčeji osídlených zemí EU.

Nejdůležitějšími odvětvími finského hospodářství jsou veřejná správa, obrana, vzdělávání, zdravotní a sociální péče. Hlavními vývozními trhy Finska jsou Německo, Švédsko a Spojené království. Finsko nejvíce dováží z Německa, Švédska a Ruska.

Francie

Francie je největší zemí EU a rozprostírá se od Severního až po Středozemní moře. Její krajina je různorodá. Na východě a na jihu jsou pohoří – nachází se tam i alpský štít Mont Blanc, který je nejvyšší horou Evropy.

Nejdůležitějšími odvětvími francouzského hospodářství jsou veřejná správa, obrana, vzdělávání, zdravotní a sociální péče. Hlavními vývozními trhy Francie jsou Německo, Španělsko a Spojené státy. Francie dováží nejvíce z Německa, Belgie a Itálie.

Chorvatsko

Chorvatsko je nezávislým státem od roku 1991. Má dlouhé a členité jaderské pobřeží, které lemuje více než 1 000 ostrovů a ostrůvků, z nichž jen 48 je trvale obýváno.

Nejdůležitějšími odvětvími chorvatského hospodářství jsou velkoobchod a maloobchod, doprava, pohostinství. Hlavními vývozními partnery Chorvatska jsou Itálie, Slovinsko a Německo. Chorvatsko dováží nejvíce z Německa, Itálie a Slovinska.

Irsko

Irsko zaujímá pět šestin stejnojmenného ostrova. Severovýchodní část ostrova tvoří Severní Irsko, které je součástí Spojeného království. Na západě ostrov omývá severní Atlantik a na jihu Keltské moře. Na východě je Irsko odděleno od Velké Británie Irským mořem.

Nejdůležitějším odvětvím irského hospodářství je průmysl. Hlavními vývozními trhy Irska jsou Spojené státy, Spojené království a Belgie. Irsko dováží nejvíce ze Spojeného království, Spojených států a z Francie.

Itálie

Itálie hraničí na severu s Francií, Švýcarskem, Rakouskem a Slovinskem a tyto hranice jsou do značné míry přirozeně vymezeny hřebenem Alp. Na jihu zaujímá Itálie celý Apeninský poloostrov, Sicílii, Sardinii, a také přibližně 68 menších ostrovů. Uvnitř italského území najdeme i dva malé nezávislé státy: Vatikán v Římě a Republiku San Marino.

Nejdůležitějšími odvětvími italského hospodářství jsou velkoobchod a maloobchod, doprava, ubytovací a stravovací služby. Hlavními vývozními trhy Itálie jsou Německo, Francie a Spojené státy. Itálie dováží nejvíce z Německa, Francie a Číny.

Kypr

Kypr je třetí nejmenší zemí EU, hned za Maltou a Lucemburskem. Kypr vstoupil do EU jako rozdělený na dvě části. Územím EU je však celý ostrov. Kyperští Turci jsou občany EU, neboť jsou státními příslušníky jednoho z členských států EU, tj. Kyperské republiky, a to přesto, že žijí v části ostrova, nad níž nevykonává kyperská vláda kontrolu.

Nejdůležitějšími odvětvími kyperského hospodářství jsou velkoobchod a maloobchod, doprava, ubytovací a stravovací služby. Hlavními vývozními trhy Kypru jsou Řecko, Irsko a Spojené království. Kypr dováží nejvíce z Řecka, Spojeného království a Itálie.

Litva

Litva leží ze všech tří pobaltských států nejjižněji. Zároveň je z těchto tří zemí největší a nejlidnatější. Kromě zvlněného území na západě a kopců na východě jde o převážně rovinatou zemi. Lesy pokrývají pouze přibližně třicet procent území.

Nejdůležitějšími odvětvími litevského hospodářství jsou velkoobchod a maloobchod, doprava, ubytovací a stravovací služby. Hlavními vývozními trhy Litvy jsou Rusko, Lotyšsko a Polsko. Litva dováží nejvíce z Ruska, Německa a Polska.

Lotyšsko

Lotyšské pobřeží Baltského moře je dlouhé přes pět set kilometrů. Pozemní hranici má Lotyšsko s Estonskem, Litvou, Ruskem a Běloruskem. Lesy pokrývají více než čtyřicet procent rozlohy této nížinaté země. V Lotyšsku se nachází více než tři tisíce jezer a dvanáct tisíc řek.

Nejdůležitějšími odvětvími lotyšského hospodářství jsou velkoobchod a maloobchod, doprava, ubytovací a stravovací služby. Hlavními vývozními partnery Lotyšska jsou Litva, Rusko a Estonsko. Lotyšsko dováží nejvíce z Litvy, Německa a Polska.

Lucembursko

Lucemburské velkovévodství je vnitrozemská země. Jde o jednu z nejmenších zemí EU, nicméně také nejbohatší – má nejvyšší příjem na obyvatele z celé EU. Území Lucemburska je převážně zalesněné a tvoří jej většinou pahorkatiny.

Nejdůležitějšími odvětvími lucemburského hospodářství jsou finančnictví a pojišťovnictví. Hlavními vývozními partnery Lucemburska jsou Německo, Francie a Belgie. Lucembursko dováží nejvíce z Belgie, Německa a Číny.

Maďarsko

Maďarsko je vnitrozemský stát ve střední Evropě, který sousedí dokonce se sedmi zeměmi. Povrch Maďarska je převážně rovinatý, na severu se nachází několik nižších pohoří.

Nejdůležitějším odvětvím maďarského hospodářství je průmysl. Hlavními vývozními trhy Maďarska jsou Německo, Rumunsko a Slovensko. Maďarsko dováží nejvíce z Německa, Číny a Rakouska.

Malta

Malta se rozkládá na pěti ostrovech ve Středozemním moři, na jih od italského ostrova Sicílie, východně od Tuniska a na sever od Libye. Malta je jednou z nejmenších a nejhustěji osídlených zemí světa.

Nejdůležitějšími odvětvími maltského hospodářství jsou velkoobchod a maloobchod, doprava, ubytovací a stravovací služby. Hlavními vývozními trhy Malty jsou Německo, Francie a Hong Kong. Malta dováží nejvíce z Itálie, Nizozemska a Spojeného království.

Německo

Německo se rozprostírá od Severního a Baltského moře až po hřebeny Alp na jihu a má ze všech zemí EU nejvíce obyvatel. Na severu hraničí s Dánskem, na východě s Polskem

a ČR, na jihu s Rakouskem a Švýcarskem, na jihozápadě s Francií a Lucemburskem a na severozápadě s Belgií a Nizozemskem.

Nejdůležitějším odvětvím německé ekonomiky je průmysl. Hlavními vývozními trhy Německa jsou Francie, Spojené státy a Spojené království. Německo dováží nejvíce z Nizozemska, Francie a Číny.

Nizozemsko

Jak již napovídá název státu, povrch této země tvoří nížiny. Přibližně čtvrtina jeho území leží pod úrovní hladiny moře. Řadu oblastí Nizozemska chrání před povodněmi mořské a vnitrozemské hráze. Velká část jeho území byla získána odčerpáváním vody a velkoplošným vysušením. Nizozemsko má dlouhé pobřeží Severního moře, na jihu země sousedí s Belgií a na východě s Německem.

Nejdůležitější odvětvími nizozemského hospodářství byly v roce 2015 veřejná správa, obrana, vzdělávání, zdravotní a sociální péče. Hlavními vývozními trhy Nizozemska jsou Německo, Belgie a Spojené království. Nizozemsko nejvíce dováží z Německa, Číny a Belgie.

Polsko

Polsko je jednou ze zemí střední Evropy. Polsku patří dlouhá část pobřeží Baltského moře. Jeho povrch tvoří převážně nížiny a pahorkatiny. Jeho jižní hranice se sestává z několika pohoří. Karpaty, které tvoří hranici se Slovenskem.

Nejdůležitějším odvětvím polského hospodářství je průmysl. Hlavními vývozními trhy Polska jsou Německo, Spojené království a Česká republika. Polsko dováží nejvíce z Německa, Ruska a Číny.

Portugalsko

Portugalsko najdeme v západní části Pyrenejského poloostrova. Ze zemí kontinentální Evropy se nachází nejzápadněji. Kromě pevninské části zahrnuje Portugalsko také souostroví Azory a Madeira, které jsou autonomními regiony.

Nejdůležitějšími odvětvími portugalského hospodářství jsou velkoobchod a maloobchod, doprava, ubytovací a stravovací služby. Hlavními vývozními partnery Portugalska jsou Španělsko, Francie a Německo. Portugalsko dováží nejvíce ze Španělska, Německa a Francie.

Rakousko

Rakousko je velice hornatá země díky své poloze na východním okraji Alp. Toto pohoří vévodí západní a jižní části Rakouska, zatímco východní spolkové země leží v podunajské nížině.

Nejdůležitějšími odvětvími rakouského hospodářství jsou velkoobchod a maloobchod, doprava, ubytovací a stravovací služby. Hlavními vývozními trhy Rakouska jsou Německo, Spojené státy a Itálie. Rakousko dováží nejvíce z Německa, Itálie a Švýcarska.

Rumunsko

Rumunsko se nachází v jihovýchodní Evropě. Severní a střední části dominují Karpaty, zatímco na jihu se rozkládá rozlehlé podunajské údolí, které nakonec přejde v deltu, v níž se Dunaj vlévá do Černého moře.

Nejdůležitějším odvětvím rumunského hospodářství je průmysl. Hlavními vývozními trhy Rumunska jsou Německo, Itálie a Francie. Rumunsko dováží nejvíce z Německa, Itálie a Maďarska.

Slovensko

Slovensko leží ve východní části střední Evropy. Na západě sousedí s ČR a Rakouskem, na severu s Polskem, na východě s Ukrajinou a na jihu s Maďarskem. Severní polovina země je hornatá. Přírodní hranici mezi Slovenskem a Polskem tvoří Vysoké Tatry a další karpatská pohoří. V jižní polovině země se rozprostírají nížiny.

Nejdůležitějším odvětvím slovenského hospodářství je průmysl. Hlavními vývozními trhy Slovenska jsou Německo, ČR a Polsko. Slovensko dováží nejvíce z Německa, ČR a Rakouska.

Slovinsko

Slovinsko leží v jižní části střední Evropy. Na severu země dominují Alpy. Na jihozápadě se nachází plošina Kras s několika horskými masívy, vápencovými jeskyněmi a soutěskami.

Nejdůležitějším odvětvím slovinského hospodářství je průmysl. Hlavními vývozními a dovozními partnery Slovinska jsou Německo, Itálie a Rakousko.

Španělsko

Španělsko se rozprostírá na Pyrenejském poloostrově a dominují mu vysoké náhorní plošiny a pohoří, jako jsou Pyreneje na severu a Sierra Nevada na jihu. Ke Španělsku patří

také Baleárské ostrovy ve Středozemním moři, Kanárské ostrovy v Atlantském oceánu a dvě autonomní enklávy v severní Africe: Ceuta a Melilla.

Nejdůležitějšími odvětvími španělského hospodářství jsou velkoobchod a maloobchod, doprava, ubytovací a stravovací služby. Hlavními vývozními trhy Španělska jsou Francie, Německo a Spojené království. Španělsko dováží nejvíce z Německa, Francie a Číny.

Švédsko

Švédsko má největší počet obyvatel ze severských zemí a je třetí největší zemí Evropské unie, pokud jde o rozlohu. Západní hranici s Norskem tvoří Skandinávské pohoří. S Dánskem na jihu je Švédsko spojeno silničním a železničním mostem.

Nejdůležitějšími odvětvími švédského hospodářství jsou veřejná správa, obrana, vzdělávání, zdravotní a sociální péče. Hlavními vývozními trhy Švédska jsou Norsko, Německo a Spojené státy. Švédsko nejvíce dováží z Německa, Norska a Nizozemska.

Spojené království

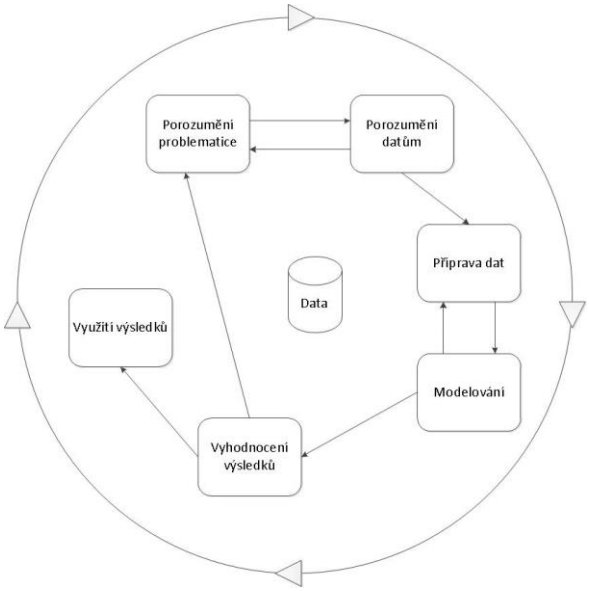
Spojené království se skládá z Anglie, Walesu, Skotska a ze Severního Irska. Krajina Spojeného království je rozmanitá, od útesů na některých částech pobřeží, přes vysočiny a nížiny až po mnoho ostrovů, zejména na severu u skotského pobřeží.

Nejdůležitějšími odvětvími britského hospodářství jsou velkoobchod a maloobchod, doprava, ubytovací a stravovací služby. Hlavními vývozními trhy Spojeného království jsou Spojené státy, Německo a Švýcarsko. Spojené království dováží nejvíce z Německa, Číny a Spojených států.

Příloha B – Metodika CRISP-DM

Tato metodika vznikla v rámci Evropského výzkumného projektu, jehož cílem bylo navrhnout univerzální a standardní model procesu dobývání znalostí z databází. Metodika CRISP-DM umožňuje řešit rozsáhlé úlohy dobývání znalostí rychleji, efektivněji, spolehlivěji a s nižšími náklady. [22]

Metodika CRISP-DM je dle [3] a [11] blíže graficky znázorněna na obrázku 1. Z obrázku jsou na první pohled patrné jednotlivé fáze a vazby mezi nimi.



Obrázek: Fáze CRISP-DM

Zdroj: upraveno dle [3]

Porozumění problematice

Úvodní fáze se zaměřuje na pochopení cílů úlohy a požadavků na řešení formulovaných ve srozumitelné podobě z manažerského hlediska a jejich transformace do úlohy vhodné k řešení z pohledu dobývání z dat v databázích. Již v této fázi je vhodné stanovit předběžný plán a soupis prací. [3] [22]

Porozumění datům

Samotný název této fáze je dostatečně výstižný. Nejprve je potřeba data získat a následně si o nich utvořit představu z pohledu jejich kvality, případně zjistit i jejich popisné charakteristiky. Je velmi vhodné i nahlédnutí do dat či pouhého vzorku dat. [7] [28]

Příprava dat

Samotná příprava dat, jež je třetí fází metodologie CRISP-DM, zahrnuje činnosti vedoucí k vytvoření datového souboru, který bude připraven pro zpracování některou z modelovacích technik. Je vhodné selekcí z dat vybrat pouze ta data, která jsou potřebná. Následně provést jejich čištění, transformaci, integraci, vytvořit nově odvozené datové proměnné a následně celý datový soubor zformátovat a upravit do stavu pro další využití. Časově je tato fáze nejnáročnější ze všech. Některé úkony je třeba provádět opakovaně. [24]

Modelování

V oblasti Data Miningu jsou různé modelovací metody. Pro každý konkrétní typ je třeba vybrat tu nejvhodnější, případně použít více různých metod a jejich výsledky kombinovat, a vhodně nastavit jejich parametry. Mezi nejčastěji používané analytické metody patří např. rozhodovací stromy, asociační pravidla, rozhodovací pravidla, neuronové sítě, či některá ze statistických metod, jako např. regresní analýza, shluková analýza či diskriminační analýza. [28]

Vyhodnocení výsledků

Ve fázi vyhodnocení výsledků je prováděno hodnocení dosažených výsledků dobývání z dat na základě cílů, které byly stanoveny v první fázi. Pokud cíle nebylo uspokojujícím způsobem z manažerského pohledu dosaženo, celý proces se opakuje od první fáze. Pokud bylo dosaženo cíle, mělo by být rozhodnuto o využití výsledků. [3] [22]

Využití výsledků

Na rozhodnutí, zda bylo dosaženo cíle, navazuje i celá tato fáze metodologie CRISP-DM. Proces využití výsledků znamená, že je sepsána projektová zpráva a celý projekt je zhodnocen. Podoba této zprávy by měla být čitelná pro zadavatele úlohy, aby na první pohled bylo zřejmé, jaké kroky je potřeba učinit, aby dosažené výsledky byly využity co možná nejefektivněji. [3] [22]

Příloha C - Datový slovník pro statistický sběr dat

Název atributu	Typ	Rozsah	Popis atributu
Stát	set	[Objekt 1, …, Objekt 29]	Stát EU
A1	range	<47; 99>	Jednotlivci ve věku 16 až 29 let (v %), kteří používají Internet denně v letech 2011 až 2016.
A2	range	<49; 94>	Jednotlivci ve věku 16 až 29 let (v %), kteří používají počítač denně v letech 2011 až 2015.
A3	range	<65; 99>	Jednotlivci ve věku 16 až 29 let (v %), kteří používají přijímají/odesílají elektronickou poštu v letech 2012 až 2016.
A4	range	<52; 95>	Jednotlivci ve věku 16 až 29 let (v %), kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.
A5	range	<2; 57>	Jednotlivci ve věku 16 až 29 let (v %), kteří prodávají produkty nebo služby na Internetu v letech 2011 až 2016.
A6	range	<36; 95>	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu produkty nebo služby v letech 2011 až 2016.
A7	range	<3; 92>	Jednotlivci ve věku 16 až 29 let (v %), kteří používají internetové bankovnictví v letech 2011 až 2016.
A8	range	<24; 80>	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu informace o zdraví v letech 2011, 2013, 2015 a 2016.
A9	range	<18; 85>	Jednotlivci ve věku 16 až 29 let (v %), kteří uskutečňují na Internetu hovory a videohovory v letech 2011 až 2016.
A10	range	<13; 69>	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu informace o cestování a ubytování v letech 2011 až 2016.
A11	range	<5300; 83700>	Reálný hrubý domácí produkt na obyvatele v letech 2011 až 2016.
A12	range	<17,1; 1369,5>	Počet obyvatel na km2 v letech 2011 až 2016.

Název atributu	Typ	Rozsah	Popis atributu
A13	range	<-1,6; 5,8>	Hodnota inflace v letech 2011 až 2016.
A14	range	<52,9; 81,2>	Dlouhodobě zaměstnaní (v %) v letech 2011 až 2016.
A15	range	<1,2; 19,5>	Dlouhodobě nezaměstnaní (v %) v letech 2011 až 2016.

Příloha D - Datový slovník pro statistický sběr dat po nahrazení chybějících hodnot

Název atributu	Typ	Rozsah	Popis atributu
Stát	set	[Objekt 1, …, Objekt 29]	Stát EU.
A1	range	<47; 99>	Jednotlivci ve věku 16 až 29 let (v %), kteří používají Internet denně v letech 2011 až 2016.
A2	range	<49; 94>	Jednotlivci ve věku 16 až 29 let (v %), kteří používají počítač denně v letech 2011 až 2016.
A3	range	<65; 99>	Jednotlivci ve věku 16 až 29 let (v %), kteří používají přijímají/odesílají elektronickou poštu v letech 2011 až 2016.
A4	range	<52; 95>	Jednotlivci ve věku 16 až 29 let (v %), kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.
A5	range	<2; 57>	Jednotlivci ve věku 16 až 29 let (v %), kteří prodávají produkty nebo služby na Internetu v letech 2011 až 2016.
A6	range	<36; 95>	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu produkty nebo služby v letech 2011 až 2016.
A7	range	<3; 92>	Jednotlivci ve věku 16 až 29 let (v %), kteří používají internetové bankovnictví v letech 2011 až 2016.
A8	range	<24; 80>	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu informace o zdraví v letech 2011 až 2016.
A9	range	<18; 85>	Jednotlivci ve věku 16 až 29 let (v %), kteří uskutečňují na Internetu hovory a videohovory v letech 2011 až 2016.
A10	range	<13; 69>	Jednotlivci ve věku 16 až 29 let (v %), kteří vyhledávají na Internetu informace o cestování a ubytování v letech 2011 až 2016.
A11	range	<5300; 83700>	Reálný hrubý domácí produkt na obyvatele v letech 2011 až 2016.
A12	range	<17,1; 1372,2>	Počet obyvatel na km2 v letech 2011 až 2016.

Název atributu	Typ	Rozsah	Popis atributu
A13	range	<-1,6; 5,8>	Hodnota inflace v letech 2011 až 2016.
A14	range	<52,9; 81,2>	Dlouhodobě zaměstnaní (v %) v letech 2011 až 2016.
A15	range	<1,2; 19,5>	Dlouhodobě nezaměstnaní (v %) v letech 2011 až 2016.

Příloha E - Datový slovník pro převedené hodnoty na průměrný koeficient růstu

Název atributu	Typ	Rozsah	Popis atributu
Stát	set	[Objekt 1, …, Objekt 29]	Stát EU.
A1	range	<0,99125; 1,09206>	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří používají Internet denně v letech 2011 až 2016.
A2	range	<0,95479; 1,08388>	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří používají počítač denně v letech 2011 až 2016.
A3	range	<0,97501; 1,0296>	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří používají přijímají/odesílají elektronickou poštu v letech 2011 až 2016.
A4	range	<0,98876; 1,0702>	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.
A5	range	<0,83255; 1,2686>	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří prodávají produkty nebo služby na Internetu v letech 2011 až 2016.
A6	range	<0,94294; 1,06608>	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu produkty nebo služby v letech 2011 až 2016.
A7	range	<0,99041; 1,12888>	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří používají internetové bankovnictví v letech 2011 až 2016.
A8	range	<0,93053; 1,1487>	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu informace o zdraví v letech 2011 až 2016.
A9	range	<0,94409; 1,22176>	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří uskutečňují na Internetu hovory a videohovory v letech 2011 až 2016.
A10	range	<0,91967; 1,05436>	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu informace o cestování a ubytování v letech 2011 až 2016.
A11	range	<0,96901; 1,03112>	Průměrný koeficient růstu reálného hrubého domácího produktu na obyvatele v letech 2011 až 2016.

Název atributu	Typ	Rozsah	Popis atributu
A12	range	<0,98473; 1,02124>	Průměrný koeficient růstu počtu obyvatel na km2 v letech 2011 až 2016.
A13	range	<-0,93487; 1,02832>	Průměrný koeficient růstu inflace v letech 2011 až 2016.
A14	range	<0,98263; 1,0193>	Průměrný koeficient růstu jednotlivců dlouhodobě zaměstnaných v letech 2011 až 2016.
A15	range	<0,88903; 1,26484>	Průměrný koeficient růstu jednotlivců dlouhodobě nezaměstnaných v letech 2011 až 2016.

Příloha F - Datový slovník binárního modelu

Název atributu	Typ	Rozsah	Popis atributu
Stát	set	[Objekt 1, …, Objekt 28]	Stát EU
A1_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají Internet denně v letech 2011 až 2016.
A1_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají Internet denně v letech 2011 až 2016.
A1_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají Internet denně v letech 2011 až 2016.
A2_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají počítač denně v letech 2011 až 2016.
A2_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají počítač denně v letech 2011 až 2016.
A2_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají počítač denně v letech 2011 až 2016.
A3_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají přijímají/odesílají elektronickou poštu v letech 2011 až 2016.
A3_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají přijímají/odesílají elektronickou poštu v letech 2011 až 2016.
A3_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají přijímají/odesílají elektronickou poštu v letech 2011 až 2016.
A4_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.
A4_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.

Název atributu	Typ	Rozsah	Popis atributu
A4_plus	flag	[0; 1]	Průměrný koeficient růstu jednotlivců ve věku 16 až 29 let, kteří jsou aktivní na sociálních sítích v letech 2011 až 2016.
A5_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří prodávají produkty nebo služby na Internetu v letech 2011 až 2016.
A5_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří prodávají produkty nebo služby na Internetu v letech 2011 až 2016.
A5_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří prodávají produkty nebo služby na Internetu v letech 2011 až 2016.
A6_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu produkty nebo služby v letech 2011 až 2016.
A6_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu produkty nebo služby v letech 2011 až 2016.
A6_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu produkty nebo služby v letech 2011 až 2016.
A7_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají internetové bankovnictví v letech 2011 až 2016.
A7_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají internetové bankovnictví v letech 2011 až 2016.
A7_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří používají internetové bankovnictví v letech 2011 až 2016.
A8_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu informace o zdraví v letech 2011 až 2016.
A8_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu informace o zdraví v letech 2011 až 2016.

Název atributu	Typ	Rozsah	Popis atributu
A8_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu informace o zdraví v letech 2011 až 2016.
A9_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří uskutečňují na Internetu hovory a videohovory v letech 2011 až 2016.
A9_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří uskutečňují na Internetu hovory a videohovory v letech 2011 až 2016.
A9_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří uskutečňují na Internetu hovory a videohovory v letech 2011 až 2016.
A10_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu informace o cestování a ubytování v letech 2011 až 2016.
A10_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu informace o cestování a ubytování v letech 2011 až 2016.
A10_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců ve věku 16 až 29 let, kteří vyhledávají na Internetu informace o cestování a ubytování v letech 2011 až 2016.
A11_minus	flag	[0; 1]	Nízká hodnota reálného hrubého domácího produktu na obyvatele v letech 2011 až 2016.
A11_nula	flag	[0; 1]	Střední hodnota reálného hrubého domácího produktu na obyvatele v letech 2011 až 2016.
A11_plus	flag	[0; 1]	Vysoká hodnota reálného hrubého domácího produktu na obyvatele v letech 2011 až 2016.
A12_minus	flag	[0; 1]	Nízká hodnota počtu obyvatel na km2 v letech 2011 až 2016.
A12_nula	flag	[0; 1]	Střední hodnota počtu obyvatel na km2 v letech 2011 až 2016.
A12_plus	flag	[0; 1]	Vysoká hodnota počtu obyvatel na km2 v letech 2011 až 2016.
A13_minus	flag	[0; 1]	Nízká hladina inflace v letech 2011 až 2016.

Název atributu	Typ	Rozsah	Popis atributu
A13_nula	flag	[0; 1]	Střední hladina inflace v letech 2011 až 2016.
A13_plus	flag	[0; 1]	Vysoká hladina inflace v letech 2011 až 2016.
A14_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců dlouhodobě zaměstnaných v letech 2011 až 2016.
A14_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců dlouhodobě zaměstnaných v letech 2011 až 2016.
A14_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců dlouhodobě zaměstnaných v letech 2011 až 2016.
A15_minus	flag	[0; 1]	Nízká hodnota počtu jednotlivců dlouhodobě nezaměstnaných v letech 2011 až 2016.
A15_nula	flag	[0; 1]	Střední hodnota počtu jednotlivců dlouhodobě nezaměstnaných v letech 2011 až 2016.
A15_plus	flag	[0; 1]	Vysoká hodnota počtu jednotlivců dlouhodobě nezaměstnaných v letech 2011 až 2016.