TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Využití metod data miningu v marketingu

Diplomová práce

Studijní program:	N6208 Ekonomika a management
Studijní obor:	Podniková ekonomika – Marketing podniku
Autor práce:	**Bc. Stanislav Mašek**
Vedoucí práce:	Mgr. Tereza Semerádová, Ph.D. Katedra informatiky

Liberec 2021

TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Zadání diplomové práce

Využití metod data miningu v marketingu

Jméno a příjmení: **Bc. Stanislav Mašek**
Osobní číslo: E18000541
Studijní program: N6208 Ekonomika a management
Studijní obor: Podniková ekonomika – Marketing podniku
Zadávající katedra: Katedra informatiky
Akademický rok: **2020/2021**

Zásady pro vypracování:

1. Formulace cílů práce.
2. Úvod do problematiky data miningu.
3. Představení IT společnosti BROKEN MOUSE.
4. Demonstrace možností využití data miningu v oblastech marketingu na konkrétním příkladu z praxe.
5. Formulace závěrů a zhodnocení cílů.

Rozsah grafických prací:
Rozsah pracovní zprávy: 65 normostran
Forma zpracování práce: tištěná/elektronická
Jazyk práce: Čeština

Seznam odborné literatury:

- AGGARWAL, Charu C. 2015. *Data mining: the textbook*. Cham: Springer. ISBN 978-3-319-14141-1.
- HAN, Jiawei a Micheline KAMBER. 2013. *Data mining: concepts and techniques*. 3rd ed. San Francisco: Morgan Kaufmann Publishers. ISBN 1-55860-489-8.
- JANOUCH, Viktor. 2014. *Internetový marketing*. 2. vyd. V Brně: Computer Press. ISBN 978-80-251-4311-7.
- RUSSELL, Matthew A. 2014. *Mining the social web*. 2nd ed. Beijing: O'Reilly. ISBN 978-1-4493-6761-9.
- PROQUEST. 2018. *Databáze článků ProQuest [*online]. Ann Arbor, MI, USA: ProQuest. [cit. 2020-09-30]. Dostupné z: http://knihovna.tul.cz/

Konzultant diplomové práce: Ing. Tomáš Bartůněk, BROKEN MOUSE

Vedoucí práce: Mgr. Tereza Semerádová, Ph.D.
Katedra informatiky

Datum zadání práce: 31. října 2018
Předpokládaný termín odevzdání: 31. srpna 2021

L.S.

prof. Ing. Miroslav Žižka, Ph.D.
děkan

doc. Ing. Klára Antlová, Ph.D.
vedoucí katedry

V Liberci dne 31. října 2018

Prohlášení

Prohlašuji, že svou diplomovou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé diplomové práce a konzultantem.

Jsem si vědom toho, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu Technické univerzity v Liberci.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti Technickou univerzitu v Liberci; v tomto případě má Technická univerzita v Liberci právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS/STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má diplomová práce bude zveřejněna Technickou univerzitou v Liberci 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.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

25. dubna 2021 Bc. Stanislav Mašek

Na tomto místě bych rád poděkoval vedoucí své diplomové práce Ing. Tereze Semerádové, Ph.D. za odborné vedení a pomoc v průběhu zpracovávání této práce. Také bych rád poděkoval Ing. Lukášovi Václavkovi za cenné rady v oblasti marketingu a zpracování dat. Dále bych rád poděkoval Ing. Tomášovi Bartůňkovi za připomínky v oblasti data miningu.

Anotace

Tato diplomová práce se zabývá možnostmi využití data miningu v marketingu. Cílem práce je analyzovat rozsáhlá data internetového obchodu pomocí dataminingové metody CRISP-DM a z toho vyvodit závěry, které budou dále prezentovány e-shopu. Výpočty a grafické interpretace budou provedeny převážně v Excelu a MATLABu. Práce se věnuje datové problematice, data miningu, včetně možností jeho zneužití, dále pak dataminingovým metodám (CRISP-DM, SEMMA, 5A) a internetovému marketingu, který zahrnuje i CRM systémy. Výsledky data miningu jsou přehledně graficky interpretovány a následně i sumarizovány. V závěru práce jsou poskytnuta marketingová doporučení, kterými by se měl e-shop zabývat, pokud si chce nejenom udržet stávající klientelu, ale také posílit svoji pozici v zahraničí.

Klíčová slova

Data mining, Analýza dat, Databáze, Data, Online marketing, E-shop

Annotation

Application of Data Mining Methods in Marketing

This diploma thesis deals with the possibilities of using data mining in marketing. The aim of the work is to analyse extensive data of e-commerce using the data mining method CRISP-DM and to draw conclusions from it, which will be further presented in the e-shop. Calculations and graphical interpretations will be performed mainly in Excel and MATLAB. The work deals with data issues, data mining, including the possibility of its misuse, as well as data mining methods (CRISP-DM, SEMMA, 5A) and Internet marketing, which includes CRM systems. The results of the data mining are clearly graphically interpreted and subsequently summarized. At the end of the thesis, marketing recommendations are provided, which the e-shop should deal with if it wants to not only retain its existing clientele, but also strengthen its position abroad.

Key Words

Data mining, Data analysis, Databases, Data, Online marketing, E-shop

Obsah

Seznam obrázků ... 13

Seznam tabulek ... 14

Seznam zkratek ... 15

Úvod ... 16

Rešerše literatury ... 17

Cíle diplomové práce ... 20

1 Úvod do data miningu ... 21

1.1 Oblasti využití data miningu ... 21

1.2 Dataminingové úlohy ... 26

1.3 Dataminingové metody ... 28

1.3.1 Metoda CRISP-DM ... 29

1.3.2 Metoda SEMMA ... 31

1.3.3 Metoda 5A ... 32

1.4 Srovnání metod ... 34

2 Data ... 35

2.1 Atributy ... 36

2.2 Technologie a techniky používané v data miningu ... 38

2.2.1 Statistika ... 38

2.2.2 Strojové učení ... 39

2.2.3 Databázové systémy a datové sklady ... 42

3 Internetový marketing ... 44

3.1 Charakteristika marketingu na internetu ... 44

3.2 Diferencovaný přístup k zákazníkům ... 46

3.3 CRM systémy ... 48

3.4 Předpoklady úspěchu na internetu ... 51

3.4.1 Informační mlha 52
4 Možnosti zneužití data miningu 53
4.1 V business světě 54
4.2 V oblasti medicíny 55
4.3 V politice 56
5 Představení subjektu řešící DM problém – IT firma BROKEN MOUSE, s.r.o. 58
6 Výběr vhodného softwaru pro data mining 60
6.1 Excel 61
6.2 Statgraphics 62
6.3 MATLAB 63
7 Data pro analýzu a jejich příprava 65
8 Finální data mining 68
8.1 Dodací adresa 68
8.2 Nejčastěji prodávaný sortiment 70
8.3 Typ prodávané značky 72
8.4 Způsob přepravy 73
8.5 Závislost počtu objednávek na času 75
8.6 Počet nákupů 77
8.7 Vývoj počtu objednávek během týdne v průběhu dne 79
8.8 Celkový počet objednávek na město 81
8.9 Průměrná cena objednávky na oblast 83
8.10 Celková cena objednávky pro evropský trh 85
9 Souhrn poznatků 87
Závěr a marketingová doporučení 90
Citace 93
Bibliografie 95

Seznam obrázků

Obr. 1: Proces metody CRISP-DM a rekurzivní povaha dataminingového procesu..... 31
Obr. 2: Grafické znázornění metody SEMMA..... 32
Obr. 3: Data mining a mnoho oblastí jejichž techniky používá..... 38
Obr. 4: Kombinace učení s učitelem a bez učitele..... 41
Obr. 5: Model znázorňující rozpad datové kostky..... 43
Obr. 6: Oblasti řízení vztahů se zákazníky..... 50
Obr. 7: Názory odborníků na problematiku zneužití velkých dat a DM..... 53
Obr. 8: Dodaná data před úpravou po úvodním načtení do MATLABu..... 67
Obr. 9: Znázornění významnosti měst v závislosti na počtu provedených objednávek..... 69
Obr. 10: Graf nejčastěji prodávaného sortimentu po úpravě pomocí regulárních výrazů... 71
Obr. 11: Typy nejčastěji prodávaných značek..... 72
Obr. 12: Graf zvoleného způsobu přepravy..... 74
Obr. 13: Graf znázorňující závislost počtu objednávek na času..... 76
Obr. 14: Znázornění počtu nákupů podle cenových intervalů..... 78
Obr. 15: Vývoj počtu objednávek během týdne v průběhu dne..... 80
Obr. 16: Celkový počet objednávek na město..... 82
Obr. 17: Průměrná cena objednávky na oblast..... 84
Obr. 18: Celková cena objednávky pro evropský trh..... 86

Seznam tabulek

Tab. 1: Dataminingové úlohy a jejich příklady 28
Tab. 2: Srovnání dataminingových metod 34
Tab. 3: Nesprávný zápis názvu obce Albrechtice 66

Seznam zkratek

5A – Dataminingová metodika firmy SPSS

NASA – Národní úřad pro letectví a vesmír

B2B – Obchodní vztah mezi obchodními společnostmi

B2C – Obchodní vztah mezi obchodní společností a koncovým zákazníkem

BI – Business intelligence

CAD – Computer-Aided Design resp. Počítačem podporované projektování

CRISP–DM – Mezioborový standardní proces pro dolování dat

CRM – Řízení vztahů se zákazníky (Customer relationship management)

ČSOB – Československá obchodní banka

DHL – Přepravní společnost

DM – Data mining

DNA – Deoxyribonukleová kyselina

DPD – Direct Parcel Distribution (přepravní společnost)

HDD – Hard disk

OLAP – Online Analytical Processing (technologie uložení dat v databázi)

PPC – pay-per-click (platba za kliknutí)

PPL – Professional Parcel Logistic (přepravní společnost)

PSČ – poštovní směrovací číslo

RegEx – Regular Expression (regulární výraz)

SEMMA – Dataminingová metoda (Sample, Explore, Modify, Model, and Assess)

SEO – Optimalizace pro vyhledávače

SPSS – Softwarová společnost vyrábějící DM software

Úvod

Nyní se nacházíme ve věku často označovaném jako informační věk. V tomto věku věříme, že informace jsou zdrojem síly a úspěchu, kterého můžeme dosáhnout skrze počítače, satelitní systémy a další moderní zařízení, která sbírají obrovské množství informací. S příchodem počítačů a prostředků pro hromadné digitální ukládání jsme začali shromažďovat a ukládat všechny druhy dat, spoléháme se na sílu výpočetní techniky, která nám pomáhá třídit toto nepřeberné množství dat. Bohužel, obrovská sbírka dat, kterou lidstvo nyní disponuje, je natolik rozsáhlá, že není možné ji ani s tou nejmodernější technologií spolehlivě všechnu analyzovat. Efektivní systémy pro správu databází jsou velmi důležitým nástrojem pro správu velkého souboru údajů a zejména pro efektivní získávání konkrétních informací z velkého množství dat kdykoli je to nutné.

Dnes máme tedy daleko více informací, než je možné zpracovat. Sem můžeme zahrnout nejrůznější obchodní transakce a vědecké údaje, satelitní snímky, textové zprávy a vojenské zpravodajství. Nyní, když lidstvo disponuje takovým množstvím dat, byly vytvořeny nové potřeby. Snahou je analyzovat získaná data a učinit lepší manažerská rozhodnutí. Tohoto můžeme dosáhnout právě pomocí data miningu – tzv. dolování dat z databází. Pomocí této metody lze spolehlivě analyzovat velké množství nepřehledných dat, která lze poté srozumitelně interpretovat.

Autor pracuje v menší IT firmě BROKEN MOUSE, s.r.o., která svým zákazníkům nabízí nejrůznější služby jako např. servis PC a notebooků, správu sítí, stavbu PC, poradenství a školení nebo třeba tvorbu webů a s tím spojenou marketingovou kampaň. Nyní již delší dobu spolupracujeme se středně velkým internetovým obchodem, který má také kamennou pobočku. Z důvodu citlivosti poskytnutých dat nebude na žádost konzultanta práce jméno firmy uvedeno. Tato firma na svých stránkách nabízí nejrůznější oblečení a sportovní vybavení. Cílem této práce je analyzovat velkoobjemová data získaná z e-shopu a demonstrovat možnosti využití výsledků v marketingu. Jelikož se bude jednat o zpracovávání velkého množství dat, jeví se možnost použití data miningu jako ideální volba.

Rešerše literatury

V případě volby vhodných knih pro tak odborné téma, jakým je data mining je naprosto nezbytné „sáhnout“ po zahraniční literatuře v anglickém jazyce. Při výběru literatury byly jako hlavní faktory brány v potaz datum publikace, náročnost textu, znalosti a praxe autora a v neposlední řadě i dostupnost literatury. V průběhu literární rešerše byl kladen důraz na aktuálnost publikace (ne starší než 10 let), přičemž velmi cenné informace, jakým směrem se v oblasti literární rešerše ubírat získal autor na portálu Data Science Central® , který se zájmem navštěvují odborníci zabývající se statistikou, analytickými metodami, strojovým učením nebo třeba umělou inteligencí.

Protože firma, pro kterou budou dataminingové metody využívány má také své facebookové stránky (aktuálně s více než 10 000 odběrateli), bylo vhodné zařadit do seznamu i literaturu zabývající se také touto problematikou. Kniha **Mining the Social Web** od autora Matthew A. Russella z roku 2014 se zabývá nejenom problematikou dolování dat z Facebooku, ale i dalších sociálních sítí, jako jsou např. Twitter, LinkedIn případně Google+. I přes vysokou odbornost tématu je kniha psána velmi srozumitelně a čtivě. Ke každé kapitole je uvedený kód, přičemž autor využívá prostoru pro open source projekty na GitHub v prostředí IPython Notebook. Problémem zde pak může být neaktuálnost odkazů a nepříliš dobrá podpora uživatelů v případě, že se naskytne nějaký problém.

Pro uvedení čtenáře do problematiky data miningu je třeba stanovit pevné teoretické základy. Je potřeba objasnit, co si představujeme pod pojmy jako data, numerická a kategoriální proměnná, proces CRISP-DM a další. Této problematice se věnuje kniha **Data Mining and Analysis: Fundamental Concepts and Algorithms** (2014) od autorů Mohammed J. Zaki a Wagner Meira Jr.

Tato publikace je velmi přehledná a obsahuje nepřeberné množství matematických definic a grafů popisujících problematiku data miningu. Velikým plusem této publikace je fakt, že autoři uvádějí skoro ke každé definici konkrétní příklad, čímž se stává snadnější na pochopení a teorie se poté lépe převádí do praxe.

Další důležitou knihou, kterou bylo nutné pročíst pro pochopení probírané problematiky v této práci je také publikace s názvem **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management** (3. vydání z roku 2011) od autorů Gordon S. Linoff a Michael J. A. Berry. Jedná se o knihu zabývající se dolováním dat se zaměřením konkrétně na marketing a řízení podniku. Kniha obsahuje velký počet případových studií, díky kterým lze snáze pochopit, jak tyto metody aplikovat na skutečný svět. Dále je zde ukázáno, jak využít nejnovějších metod a techniky data miningu k řešení běžné obchodní problematiky.

Autoři se zabývají problematikou kampaně s přímým marketingem, identifikací nových zákaznických segmentů nebo třeba odhadováním úvěrového rizika. Dále pak jsou, pokryty pokročilejší témata, jako je příprava dat pro analýzu a vytvoření potřebné infrastruktury pro data mining u naší společnosti. Po teoretické stránce kniha pokrývá problematiku základních technik dolování dat, včetně rozhodovacích stromů, neuronových sítí, asociačních pravidel nebo třeba analýzy odkazů.

V češtině můžeme mezi top 2 knihy zařadit **Data mining a klasifikační modely** od prof. Hany Skalské (2010) a knihu přeloženou do českého jazyka **Data mining - Praktický průvodce dolováním dat** od Olivie Parr Rud (2002). Byť se v tomto případě jedná o publikaci z roku 2002, teoretické poznatky jsou stále platné.

Mezi další nalezenou zahraniční literaturu můžeme zařadit knihu od Charu C. Aggarwala z roku 2015 s názvem **Data mining: the textbook**, kde jsou srozumitelně popsány teoretické poznatky ohledně DM problematiky. Dále 3. vydání knihy od autorů Han Jiawei a Micheline Kamber z roku 2013, **Data mining: concepts and techniques**, která obsahuje teoretická východiska, podobně, jako kniha od autorů Jiawei a Kamber. Z oblasti marketingu v návaznosti na internet je dobrou volbou 2. vydání knihy **Internetový marketing** od Viktora Janoucha z roku 2014. Tato kniha obsahuje řadu pouček, jak rychle rozšířit povědomí o svých webových stránkách a další informace z marketingové praxe. Bude tedy nezbytnou součástí pro praktická východiska této práce.

Pro pochopení jeho fungování a i oblastí, kde se data mining vyskytuje a možností jeho implementace, je vhodná kniha s názvem **Data Mining and Learning Analytics: Applica-**

tions in Educational Research od autorů Samira ElAtia, Donald Ipperciel a Osmar R. Zaiane. Tato kniha pojednává o úvodu do problematiky DM, výzvách, problémech, očekáváních a praktické implementaci data miningu (DM) v oblasti vzdělávání. Úvodní kapitoly knihy provází čtenáře obecným přehledem modelů DM, Learning Analytics a sběru dat v kontextu pedagogického výzkumu. Zároveň jsou zde definovány a diskutovány čtyři hlavní principy dolování dat - predikce, shlukování, asociační pravidla a detekce odlehlých hodnot. Dále kniha řeší nově vznikající vzdělávací dolování dat, které urychluje pedagogický výzkum - od identifikace ohrožených studentů a odstraňování socioekonomických mezer, až po pomoc při hodnocení učitelů. Tato kniha obsahuje příspěvky od mezinárodních odborníků v různých oblastech, ale co se týče náročnosti, je psaná velmi odborně a určité pasáže je potřeba přečíst několikrát pro hlubší porozumění myšlence autorů.

Cíle diplomové práce

Cílem této diplomové práce je analyzovat data z e-shopu středně velké firmy. Jedná se o objednávky provedené v letech 2014-2019. Jako vhodná metoda pro zpracování dat byl zvolen data mining. V tomto bodě je nutné podotknout, že se nejedná o metody statistické analýzy, protože se stává, že jsou obě metody často zaměňovány. Oproti statistické analýze data mining neklade takový důraz na přesné porozumění použité metody. Stejně tak nebere v potaz žádné předpoklady o datech, a proto ani my žádné dělat nebudeme. Zároveň DM nepracuje pouze s výběrem hodnot, ale zabývá se velmi rozsáhlým množstvím dat. Problematika DM vyžaduje oproti statistické analýze porozumění problému z datového a business pohledu. Jsou tedy spíše vyžadovány znalosti z prostředí podniku a databázových systémů, než hluboké znalosti statistické analýzy.

Výstupem této práce by měly být relevantní informace, s nimiž bude moct e-shop dále nakládat dle svého uvážení a bude moct např. učinit další kroky k úspěšnému rozvoji online prodeje a naopak dostane také informace, čemu se případně vyvarovat, jaké zboží třeba stáhnout z nabídky apod. Právě z toho důvodu není cílem této práce ani tvorba marketingové kampaně, případně nějakých hlubokých business strategií. V závěru práce budou poskytnuta pouze stručná doporučení e-shopu, jelikož hlavním účelem této práce jsou dataminingové aktivity.

1 Úvod do data miningu

Spolu s rozvojem výpočetní techniky a počítačů a zavedení elektronického sběru dat vzniklo i množství velkých datových souborů. Společnost, a především podniky potřebovaly tato data nějakým způsobem analyzovat a použít informace v nich ukryté, aby zvýšily svůj výdělek, udržely krok s konkurencí nebo třeba optimalizovaly své produkty. Problém byl v tom, že pro takovéto objemy dat nebyly klasické statistické metody příliš vhodné. Bylo tedy potřeba vynaleznout takové metody, které dokážou nalézt i složité nelineární vztahy, a to navíc bez omezujících předpokladů. Bylo tedy potřeba využít výpočetní síly počítačů k nalezení určité struktury – pravidel, vzorů a asociací, namísto statistických parametrů, jako jsou střední hodnoty, váhy a uzly (StatSoft, 2014).

Termín data mining se poprvé objevuje v devadesátých letech. Roku 1991 byla napsána 1. definice data miningu Williamem J. Frawleym: „Data mining je netriviální získávání předtím neznámé a potenciálně užitečné informace ukryté v datech.“ Do českého jazyka se občas překládá jako „dolování“ či „vytěžování“ dat. Počátkem nového tisíciletí byl data mining osamostatněn jako nové odvětví statistiky (StatSoft, 2014).

1.1 Oblasti využití data miningu

Již delší dobu lidstvo shromažďuje nesčetné množství dat, od triviálních numerických měření a textových dokumentů, až ke složitějším informacím, jako jsou prostorová data, multimediální kanály a hypertextové dokumenty. Je zde nepřeberné množství různých informací shromážděných v digitální podobě v databázích a prostých databázových souborech.

1.1.1 Obchodní transakce

V dnešní době je skoro každá transakce v oblasti podnikání uložena a kdykoliv ji lze v historii dohledat. Takové transakce se obvykle vztahují k určitému času a zahrnují nejrůznější obchody, nákupy, výměny, bankovnictví, akcie atd. Případně se jedná o vnitropodnikové operace, jako je správa vlastních výrobků a majetku. Veliké obchodní domy díky rozšířenému používání čárových kódů uskuteční miliony transakcí denně, což představuje často terabajty dat. Úložný prostor ale není v tomto případě hlavním problémem (také proto, že cena HDD neustále klesá). Problémem je efektivně využít získaná data

v přiměřeném čase s ohledem na konkurenční boj. A právě v tomto může data mining významně pomoct. Data mining neslouží pouze firmám, které operují s nepřeberným množstvím dat. Nachází uplatnění také v menších a středních podnicích (Zaiane a Ipperciel, 2016).

Je všeobecně rozšířený omyl, že pro kvalitní data mining potřebujete obrovské množství dat. Faktem ale je, že i menší množství dat se dá pomocí data miningu analyzovat a získat na první pohled skryté informace, které by např. použití programu jako je Excel neodhalilo. V případě obchodních transakcí a nejrůznějších bankovních operací je v dnešní době data mining hojně využíván. A banky typu ČSOB, UniCredit Bank nebo MONETA využívají metod umělé inteligence a data miningu poměrně běžně. Stejně tak můžou banky využívat data miningu případně nástrojů umělé inteligence a fuzzy logiky k posouzení bonity klienta, jelikož se jedná o velmi náročné výpočty, kde je potřeba vzít v potaz velké množství informací.

Dále pak, co se týče obchodních transakcí, spousta firem má dnes tendence monitorovat své zákazníky a shromažďovat o nich velké množství dat, které by pak rádi analyzovali a výsledky použili pro navýšení zisku, případně jiné činnosti. Problém je v tom, že svým chování s neustálou snahou získávat od svých zákazníků informace, je mohou snadno odradit, nebo dokonce ztratit. A ztracená důvěra se zpět získává velmi těžko (Zaiane a Ipperciel, 2016).

1.1.2 Vědecké údaje

Naše společnost shromažďuje obrovské množství vědeckých údajů, které je potřeba analyzovat. Věda aktuálně sbírá nová množství dat rychleji a lidstvo nestíhá zpracovat ani ta stará, již nasbíraná. Můžeme sem zařadit vědní obory jako je např. bioinformatika, farmaceutická informatika, geoinformatika a další. Využití metod data miningu a ostatních pokročilých analytických nástrojů vědu v určitých oblastech nesmírně posunulo dopředu. Bylo tomu tak především díky novým možnostem snadného zpracování a interpretaci nasbíraných dat, což by člověku zabralo roky (Zaiane a Ipperciel, 2016).

Výhodou je, že DM software se dá ještě přizpůsobit, resp. „ušít na míru“ pro konkrétní vědecké oblasti. Lze říci, že v dnešní době je DM software nezbytnou součástí v určitých

oblastech vědeckého bádání. Přičemž použití DM technik v oblasti vědy a na akademické sféře dále napomáhá jeho rozvoji a popularizaci.

Při setkání s tímto tématem je většina laiků na první pohled zaskočena a nevěří tomu, že by mohli být sami schopni DM provádět. Dataminingové úlohy jsou veřejností vnímány jako vysoce odborné a určené práce pro sféru vědeckou, akademickou či podnikatelskou. Faktem ale je, že se jedná o metody, které jsou třeba oproti statistické analýze mnohokrát jednodušší a není zde vyžadováno vysoce odborných znalostí (Zaiane a Ipperciel, 2016).

1.1.3 Lékařské a osobní údaje

Neustále shromažďujeme rozsáhlé množství informací ohledně skupin i jednotlivců – od sčítání lidu až po osobní parametry zákazníků. Vlády, společnosti a organizace, jako jsou např. nemocnice, uchovávají velmi důležité množství osobních údajů, které jim pomáhají řídit lidské zdroje, lépe porozumět trhu nebo třeba pomoci klientům. Bez ohledu na problémy týkající se ochrany osobních údajů, které tento typ dat často odhaluje, jsou tyto informace sbírány, používány a dokonce i sdíleny mezi institucemi navzájem. Ve spojitosti s jinými údaji můžou tyto informace pomoci osvětlit chování zákazníků a dalších subjektů (Zaiane a Ipperciel, 2016).

Například, co se týče formulářů na sčítání lidu. Využití DM se jeví jako ideální postup. Jelikož se jedná o velmi rozsáhlá data s mnoha proměnnými. V tomto případě je DM velmi účinný analytický nástroj, díky kterému bude možné odhalit souvislosti, které by na první pohled nebyly vidět.
Co se týče lékařství a medicíny všeobecně, tak zde existuje také riziko zneužití data miningu v podobě rozklíčování genetické informace obsažené v DNA. Na základě čehož bude možné provádět úpravy na lidech ještě před tím, než se narodí. Tomuto tématu se ale budeme detailněji věnovat v sekci Možnosti zneužití data miningu, která se nachází dále v práci.

Co se týče problematiky osobních údajů, tak aktuální problém se nyní vyskytuje v Číně, která používá metody data miningu k rozpoznávání obličejů na všudypřítomných kamerách. Své občany neustále monitoruje a v krátké době chce zavést systém, pomocí kterého jim bude přidělovat body na základě jejich chování. Toto lze tedy jednoznačně považovat za zneužití DM metod pro politické účely. Tomuto problému se budeme detailněji dále věnovat v sekci zabývající se možnostmi zneužití DM.

1.1.4 Satelitní snímky

Lidstvo disponuje obrovským počtem satelitů po celém světě: některé jsou pevně umístěny nad jednou oblastí a některé obíhají kolem Země, všichni ale vysílají nepřetržitý proud dat na povrch. NASA, která tyto satelity kontroluje, přijímá každou vteřinu více dat než je schopná vyhodnotit. Mnoho satelitních snímků a dat je zveřejněno ihned v okamžiku obdržení v naději, že ostatní výzkumní pracovníci pomohou s analýzou. Co se týče satelitních snímků a např. i vesmírných programů, DM není pouze výsadou USA. Na jeho vývoji se podílejí i další státy, jako např. Rusko a Indie. Především z oblasti Indie pochází spousta odborníků zabývající se tímto tématem (Zaiane a Ipperciel, 2016).

1.1.5 Hry

Naše společnost shromažďuje obrovské množství údajů a statistik her, hráčů a sportovců. Od hokeje, basketbalu, plavání a boxu až po nejrůznější šachové pozice, vše je ukládáno. Komentátoři a novináři tyto informace používají ve zpravodajství, zatímco trenéři a sportovci by chtěli tyto údaje využít, aby zlepšili výkonnost a lépe porozuměli svým soupeřům. Není překvapením, že metody data miningu se používají i k hazardním hrám (Zaiane a Ipperciel, 2016).

Pokud si např. chcete vsadit na tenisový zápas u Fortuny a na určitého hráče je vypsán kurz. Pak tento kurz velmi pravděpodobně není pouhý odhad bookmakera, ale jedná se o číslo, ke kterému se sázková kancelář dostala přes dlouhodobou analýzu kariéry hráče, včetně všech jeho úspěchů, neúspěchů, zranění, nových kontraktů, pauz ve hře, atd… A toto všechno je analyzováno před dataminingové či jiné analytické metody.

1.1.6 CAD/CAM a údaje ze softwarového inženýrství

Systémy CAD, generují obrovské množství dat, které lze dále ještě zpracovat. Softwarové inženýrství poskytuje nepřeberná množství zdrojových kódů, která lze analyzovat. Moderní výrobní systémy vybavené počítačovými systémy protokolování dat shromažďují velké objemy dat v reálném čase. Data mining pomáhá při analýze velkých databází a při objevování trendů, vzorů a znalostí. Data mohou obsahovat cenné informace, například i v aplikacích typu CAM - počítačem podporovaná výroba (Zaiane a Ipperciel, 2016).

Existuje mnoho aplikací ve strojírenském designu, výrobě a řízení provozu, které mohou využívat funkce data miningu. V dnešní době máme k dispozici nové metody těžby velkého množství obráběcích prvků, CAD modelů a výrobních dat. Tyto metody jsou založeny na technikách klasifikace learning-logic pro těžbu 3D CAD dat. Metody tohoto typu byly hodnoceny v řadě podobných aplikačních domén, konkrétně v klasifikačních úlohách. Experimentální výsledky prokázaly, že metody jsou účinné z hlediska přesnosti klasifikace, a proto je lze použít jako účinný nástroj pro dolování dat pro analýzu a klasifikaci CAD modelů (Zaiane a Ipperciel, 2016).

1.1.7 Textové zprávy a záznamy (e-maily)

Většina komunikace uvnitř i mezi společnostmi nebo výzkumnými organizacemi, případně ostatními probíhá lidmi na základě zpráv a poznámek v textových formulářích často vyměňovaných e-mailem. Tyto zprávy jsou pravidelně ukládány v digitální podobě pro budoucí použití a tvorbu referencí impozantní digitální knihovny (Zaiane a Ipperciel, 2016).

Zde opět existuje možnost zneužití dat, kdy je možné pomocí data miningu analyzovat proběhlé konverzace a vyhledat klíčová slova a na základě toho např. upravit obchodní strategii, případně data může firma dále prodat, byť se jedná o nelegální krok. Firmy si musí dávat pozor, aby nějakým nešikovným krokem nepřišly o své zákazníky, protože lidé jsou všeobecně velmi citliví na svá data.

Např. aplikace WhatsApp, která patří Facebooku přišla o miliony uživatelů, poté co se uživatelé dostali nové podmínky užívání k odsouhlasení. Tam se mohli dočíst, že v jejich aplikacích obchodníci, kteří komunikují se svými zákazníky na WhatsAppu, si nově mohou nechat ukládat své konverzace na serverech Facebooku. Díky tomu pak mohou tato data využívat pro své reklamní kampaně na Facebooku. Pro lepší zacílení reklamy budou moci být mezi WhatsAppem a Facebookem sdílena také některá data o platbách a transakcích.
Nové podmínky oproti minulé verzi také doslova říkají, že obchodníci nyní na WhatsAppu mohou dát třetím stranám (včetně Facebooku) přístup ke svým komunikacím. Tyto zprávy se zákazníky pak mohou spravovat. Třetí strana pak může jednotlivé komunikace posílat, ukládat nebo číst, což je samozřejmě uživatelům proti srsti.

1.1.8 World Wide Web úložiště

Od založení WWW roku 1993 byly dokumenty všech druhů formátů, obsahu a popisu shromážděné a propojené hypertextovými odkazy, což z nich činí největší repozitář, který kdy byl postaven. Navzdory své dynamické a nestrukturované povaze je World Wide Web je nejdůležitější sbírka dat, jakou lidstvo disponuje. Mnoho odborníků se domnívá, že World Wide Web se stane kompilátem lidských znalostí. Nyní se nachází na webu obrovské množství dat, které je možné pomocí data miningu analyzovat (Zaiane a Ipperciel, 2016).

1.2 Dataminingové úlohy

DM problematiku lze řešit pomocí statistických úloh. Tyto úlohy stojí nad samostatnými daty a zpravidla je lze rozdělit do několika skupin.

Klasifikační metody

Mají široké využití všude tam, kde je shromažďováno větší množství dat. Jedná se o proces zařazování objektů (klientů, pacientů, dlužníků) do tříd, přičemž třídou je myšleno např. splatí/nesplatí (nějaký dluh), zdravý/nemocný (pacient). Jde o nejčastější DM úlohu, kterou nad daty vykonáváme. V těchto úlohách se vyskytuje tzv. cílová proměnná (učitel), který definuje příslušnost konkrétního subjektu (např. zákazníka) do nějaké třídy – např. loajální (StatSoft, 2014)

Shlukování/Segmentace

Cílem tohoto typu úlohy je nalezení objektů, které jsou si podobné, případně skupiny vzájemně podobných objektů (např. zákazníků) bez znalosti či nějaké definice těchto skupin. Proto v této úloze neexistuje cílová proměnná. Tento typ analýzy nám umožní shlukovat objekty (zákazníky) do skupin podle jejich vzájemné podobnosti, která ale nemusí být na první pohled zřejmá (StatSoft, 2014).

Predikce

Sem patří úlohy, které se zaměřují na předpovědi vývoje určitého ukazatele v čase (objem poptávky, ceny a dalších ekonomických ukazatelů) pomocí netriviálních statistických technik – např. neuronové sítě (StatSoft, 2014).

Regrese

Tyto úlohy slouží k vysvětlení a předpovědi spojitých proměnných za pomoci dostupných informací z historických dat. Regresní úloha se liší od klasifikační především typem výsledku. V regresi je výsledkem spojitá číselná hodnota, nikoliv odhad dané kategorie (třídy). V některých oblastech se tyto metody nazývají úlohami typu: „Co se stane, když…“ (StatSoft, 2014).

Asociační pravidla

Pomocí asociačních pravidel lze z velkého počtu dat stanovit pravidlo, které např. říká, že pokud návštěvník klikne na záložku „Pro ženy“, tak s určitou pravděpodobností klikne také na „hubnutí a diety“. Snahou asociačních pravidel je zjistit mezi položkami takový vztah, že z přítomnosti jedné nebo více položek v transakci vyplyne výskyt jiných položek (StatSoft, 2014).

Text Mining

Text Mining se zabývá zpracováním nestrukturovaného textu. Můžeme ho definovat jako proces vytěžující cenné informace z textu. V textové proměnné jsou obvykle hledána klíčová slova, přičemž následně je provedena jejich frekvenční analýza. Případy (např. konkrétní klienti), kde se tato klíčová slova vyskytla, jsou indexována a následně vrácena do databáze jako nová číselná proměnná, která je pak využita v rámci klasifikačních metod. Mezi typické metody spadající do oblasti data miningu lze zařadit klasifikační a regresní stromy, neuronové sítě a metody strojového učení (StatSoft, 2014).

Pro přehlednost využití jednotlivých dataminingových úloh byla sestavena Tab. 1 na další stránce, kde jsou znázorněny konkrétní případy využití pro každou DM úlohu.

Tab. 1: Dataminingové úlohy a jejich příklady

Úloha	Příklad
Klasifikace	Detekce nevyžádané pošty (spam), Určování příčin závad
Shlukování	Rozčlenění klientů např. banky/pojišťovny – schopnost určovat třídy/skupiny klientů. Rozlišování na bonitní, firemní, s dluhy nebo vysokým investičním kapitálem
Predikce/Regrese	Předpověď vývoje kurzu měn, kryptoměn, Predikce spotřeby vody/elektrické energie
Asociační pravidla	Analýza změn – nákupního chování, poskytovatelů služeb (internet, operátoři, energie); Analýza webových stránek – kliky na reklamy, návštěvnost segmentů Analýza spotřebního koše
Text mining	Zápisy z call-centra (rozhovory) – extrahování atributů, Využití četnosti slov/písmen v dokumentech

Zdroj: Vlastní zpracování podle BERKA, P., 2003

1.3 Dataminingové metody

V dnešní době máme k dispozici několik data miningových metodologií. Za zmínku stojí uvést např. model SEMMA, 5A, velmi populární je také metoda SixSigma a v neposlední řadě také metodika CRISP-DM, která bude použita v praktické části této práce.
V průběhu doby začaly vznikat metodiky, jejichž cílem je poskytnout uživatelům jednotný rámec pro řešení různých úloh z oblasti dobývání znalostí z databází, jehož je data mining nedílnou součástí. Tyto metodiky umožňují sdílet a přenášet zkušenosti z úspěšných projektů.

Za některými metodikami stojí producenti programových systémů. Jmenovat můžeme metodiku „*5A*“ od firmy SPSS nebo metodiku *SEMMA* od firmy SAS. Jiné metodiky vznikají díky spolupráci výzkumných a komerčních institucí jako „softwarově nezávislé“. Právě sem můžeme zařadit metodiku *CRISP-DM*, která je z nich nejrozšířenější (Rauch a Šimůnek, 2014).

1.3.1 Metoda CRISP-DM

Metodika Cross-Industry Standard Process for Data mining, také označovaná jako CRISP-DM pokrývá kompletní proces data miningových úloh. Její výhodou je, že je nezávislá na odvětví společnosti zkoumající data, ani na použitých softwarových nástrojích, či na aplikaci konkrétní metody nebo algoritmu v oblasti data miningu (Rauch a Šimůnek, 2014).

Celý proces CRISP-DM lze rozdělit na 6 částí:

1. Pochopení obchodních souvislostí

Jak uvádí Rauch a Šimůnek (2014), tomto kroku je nezbytné porozumět obchodu. Zároveň je potřeba stanovit si podnikové cíle, kritéria úspěchu a cíle data miningu. Čeho chceme použitím data miningových nástrojů dosáhnout, co má být výstupem projektu?

2. Pochopení dat

Jednou z možných příčin neúspěchu v případě použití DM metod může být i selhání lidského faktoru. Je daleko pravděpodobnější, že chybu udělá člověk, jelikož výpočetní technika se v této oblasti dnes splete už málokdy (pokud sem tedy nezahrneme i metody umělé inteligence). V tomto kroku je tedy nezbytné porozumění získaným datům, přičemž následuje jejich průzkum a ověření kvality včetně nalezení odlehlých hodnot (Rauch a Šimůnek, 2014).

3. Příprava dat

Tento proces obvykle zabere přes 90 % celkového času věnovaného projektu. Lze sem zařadit úkony, jako sběr dat, jejich konsolidace a čištění – vazební tabulky, agregace a vypořádání se s chybějícími hodnotami. V procesu selekce je třeba zvážit, zda budou ignorována neužitečná data, jak naložit s odlehlými pozorováními, jakým způsobem data vybrat, použití vizualizačních nástrojů a transformační proces – vytváření nových odvozených proměnných (Rauch a Šimůnek, 2014).

4. Modelování

Výběr vhodných modelovacích technik je závislý na stanovených data miningových cílech (deskriptivní vs. prediktivní modelování). Rauch a Šimůnek, 2014 dále uvádí, že se většinou se jedná o iterační proces propojený s přípravou dat. Je potřeba rozlišovat přístup pro „supervised“ a „unsupervised learning“

Supervised learning – data (pozorování, měření, atd.) jsou označena předem definovanými/známými třídami. Nová testovací data jsou následně do těchto tříd rozřazena. Z pohledu kauzality daný model definuje vztah mezi vstupními daty a daty výstupními.

Unsupervised learning – předem nejsou definovány žádné třídy, přičemž, pro daná data je cílem prokázat existenci nějakých tříd. Z pohledu kauzality jsou pak všechna data chápána jako výstupní. Modelujeme závislost daných dat na určitých neznámých skrytých proměnných (Rauch a Šimůnek, 2014).

5. Vyhodnocení modelu

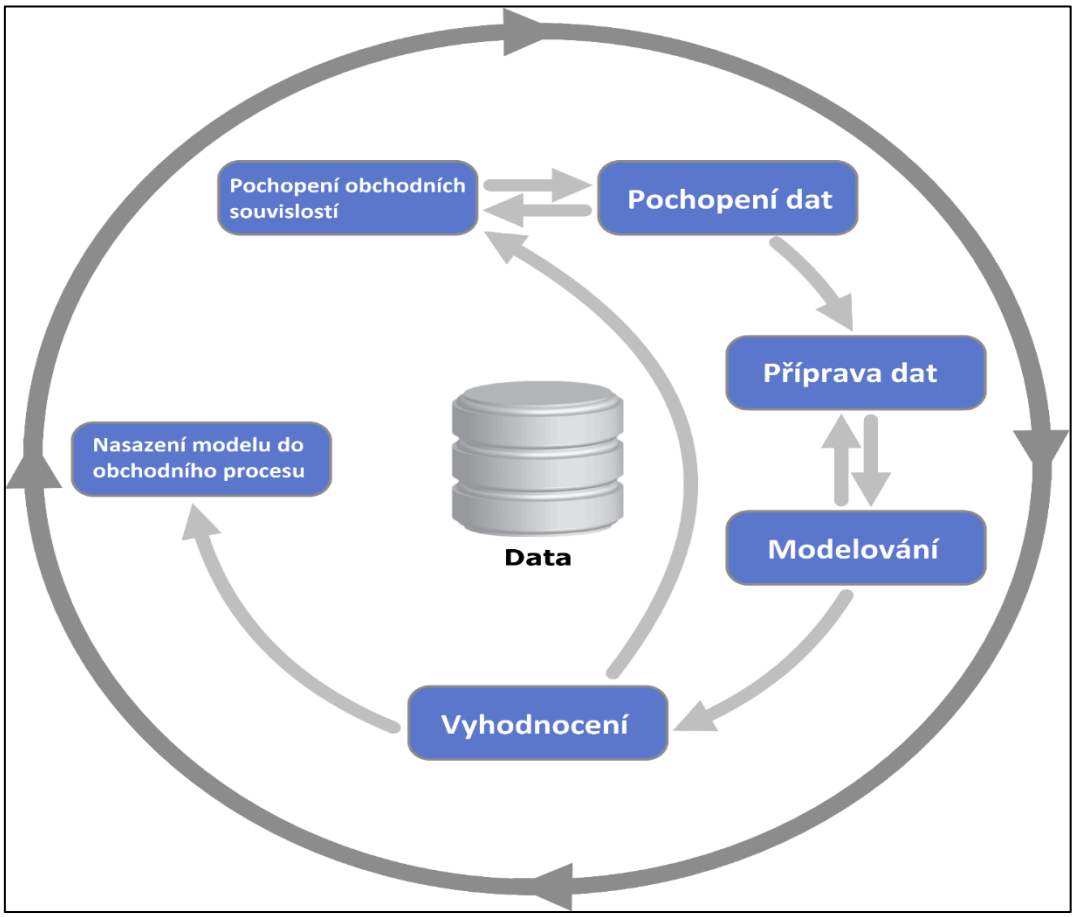
V tomto bodě jde o zjištění, jak se chová model na testovacích datech. Metody a kritéria jsou závislá na typu modelu (např. průměrná chyba pro regresní modely). Vyhodnocení modelu zahrnuje také jeho interpretaci, přičemž důležitost a obtížnost interpretace je do značné míry závislá na zvoleném modelovacím algoritmu (Rauch a Šimůnek, 2014).

6. Nasazení modelu do obchodního procesu

Zde je potřeba určit, jakým způsobem mají být výsledky využity, kdo je bude využívat a jak často je bude využívat? Po určité době bude možné vyhodnotit, zda bylo použití DM vhodné a přineslo požadovaný efekt (Rauch a Šimůnek, 2014).

Co se týče celého DM procesu, samotná analýza dat překvapivě zabere nejméně času. A v případě, že analýzu provádí profesionál, jedná se o otázku několika málo hodin. Co je ale nejtěžší část a zároveň ta část, kde může dojít k několika závažným chybám, je proces přípravy dat. Může se totiž stát, že získáte data, která byť analyzujete sebelíp, stejně budou nicneříkající. Také se může stát, že se bude jednat o data obsahující šum. V takovém případě budou nějakým způsobem zkreslena nebo jinak poškozena a bude potřeba je ještě dále upravovat. Této problematice se ale budeme věnovat dále v práci. Co se týče samotného data miningu, tak počítač chybu udělá málokdy. Spíše se stává, že uživatel analyzovaná data špatně interpretuje, nebo neumí s výsledky vhodně naložit. Vyhodnocení modelu tak nikdy nesmí být podceněno. Ohledně nasazení modelu do obchodního procesu, tak v tomto kroku dostane management vyhodnocena data a je jenom na něm, jakým způsobem výsledky využije.

Celý proces metody CRISP-DM je znázorněn na Obr. 1 níže. Nejprve si je potřeba stanovit cíle, proč vůbec DM provést? Co by mělo být jeho výstupem? Zda-li podnik požaduje pouze analyzovat data, nebo i odpovědět na předem zadané otázky?

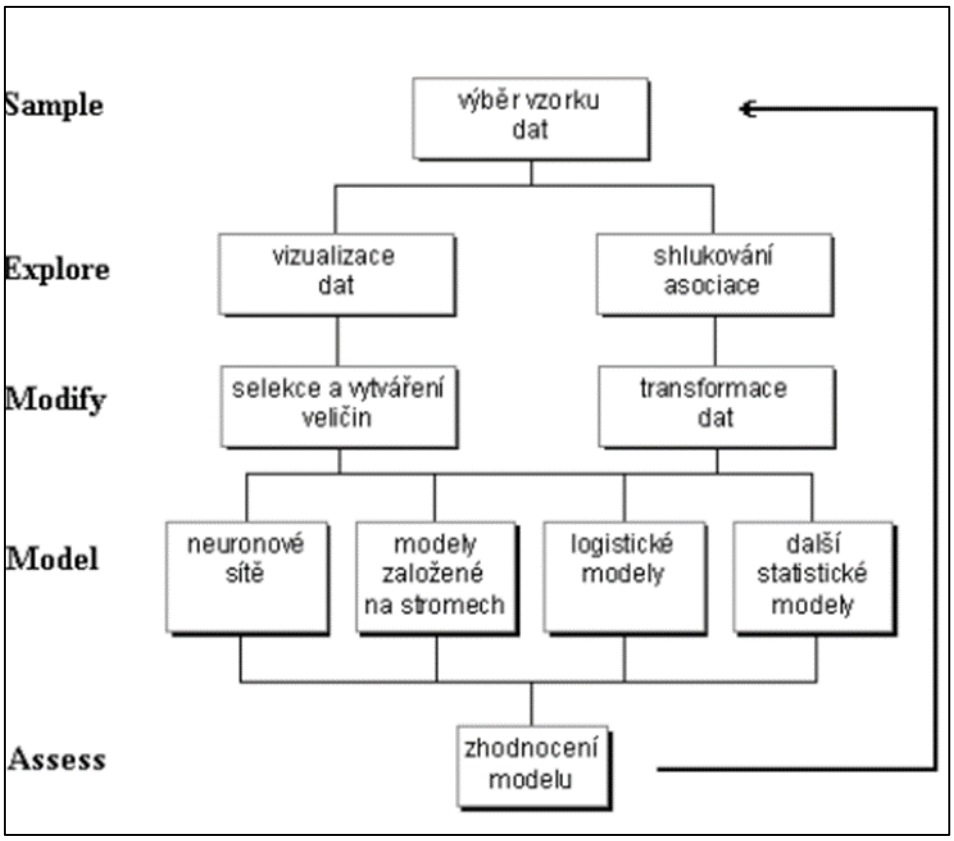


Obr. 1: Proces metody CRISP-DM a rekurzivní povaha dataminingového procesu
Zdroj: Vlastní zpracování podle Jiawei a Kamber, 2012

1.3.2 Metoda SEMMA

Jak popisují Jiawei a Kamber, 2012, jedná se o metodu těžby dat vyvinutou společností SAS, která se zabývá vývojem statistického a BI softwaru. Tato metodika popisuje proces, kterým musí člověk projít, aby získal vhled a znalosti ze zkoumaných údajů. Detailnější popis problematiky včetně konkrétních úkonů je rozebrán na Obr. 2 níže. V případě metody SEMMA je důraz kladen na snadnou interpretaci výstupů ve formě, která je srozumitelná lidem v businessu a obchodu celkově. Dle metody SEMMA můžeme dataminingový proces rozdělit na následující kroky:

- **S**ample - vybírání vhodných objektů
- **E**xplore - vizuální průzkum a redukce dat
- **M**odify - seskupování objektů a hodnot atributů, datové transformace
- **M**odel - analýza dat
- **A**ssess - porovnání modelů a jejich interpretace



Obr. 2: Grafické znázornění metody SEMMA

Zdroj: Berka, P, 2003

1.3.3 Metoda 5A

Jedná se o metodu, kterou nabízí firma SPSS pro dobývání znalostí z databází. Název této metody je akronymem pro jednotlivé prováděné kroky:

• Assess – posouzení potřeb projektu

• Access – shromáždění potřebných dat

• Analyze – provedení analýz

• Akt – přeměna znalostí na akční znalosti

• Automate – převedení výsledků analýzy do praxe

Berka, 2003 dále uvádí, že žádná data nemají význam, jestliže jsou oddělena od kontextu.

První krok v analytickém procesu se zabývá stanovením kontextu - cílů, strategií a procesů. Firma SPSS k tomu dodává, že je potřeba:

• Určit data, jejichž sběr, pořízení a skladování je nutné zajistit pro provedení takových analýz, které chceme realizovat
• Připravit se na své projekty a obory, v nichž se rozhodujeme - jejich porozuměním zabezpečíme ty analytické nástroje, které potřebujeme
• Vzdělávat se a trénovat všechny lidi, kteří myslí analyticky a používají efektivně software jako součást přemýšlení nad problémy a analýzu dat jako příslušnou složku rozhodovacího procesu

Druhým krokem v metodologii 5A je sběr a příprava dat. Je třeba získat vhodné soubory z podnikových datových skladů, datových bází, odkazových systémů a jiných interních zdrojů. Lze využít i data týkající se daného problému, která jsou nabízena veřejně (oficiální statistiky, rezortní data, demografické a psychografické údaje apod.). Data lze rovněž získat vlastními průzkumy nebo od výzkumné firmy (Berka, 2003).

Třetím krokem je používání různých analytických postupů k tomu, aby byly nalezeny odpovědi na otázky stanovené v prvním kroku. V tomto kroku se data přeměňují na informace a znalosti. Firma SPSS doporučuje širokou škálu nástrojů pro zkoumání a porozumění datům počínaje deskriptivní statistikou, přes metodu OLAP až po metody strojového učení (rozhodovací stromy, neuronové sítě). Doporučení je zřejmé: „Použijte více metod a porovnejte jejich výsledky a vhodnost, abyste získali nejlepší řešení a navíc rychle a jednoduše“ (Berka, 2003).

Čtvrtý krok procesu obsahuje doporučení, řadu dodatečných otázek a následné rozhodnutí. Znalosti nalezené v předcházejícím kroku se zde mění na znalosti akční. Nalezené výsledky by měly být předkládány v jasné a srozumitelné podobě (Berka, 2003).

Pátým krokem je převedení výsledků analýzy do praxe. Tento krok obsahuje všechny činnosti, kterými lze zajistit aplikaci učiněných rozhodnutí. Sem patří např. vytvoření praktického rozhraní k rozvinutí nalezených modelů do takového formátu, který je snadný

pro užívání a porozumění v běžné a opakované praxi organizace a monitorování výsledků (a důsledků) prováděných rozhodnutí. Další z doporučení zní „Automatizujte své analýzy tak, aby opakující se úlohy nezabíraly čas a abyste mohli snadno aktualizovat své modely s tím, jak přicházejí nové výsledky“ (Berka, 2003).

1.4 Srovnání metod

Pro přehlednost můžeme tři výše zmíněné metody uspořádat do Tab.2, kde jsou podrobně vidět rozdíly v procesu zadání projektu a zjištění stavu, úpravy dat a modelování a nakonec evaluace výsledků a finální implementace.

Tab. 2: Srovnání dataminingových metod

<table>
<tr><th></th><th>5A</th><th>SEMMA</th><th>CRISP-DM</th></tr>
<tr><td rowspan="2">Zadání projektu zjištění stavu</td><td>Assess – posouzení potřeb projektu</td><td></td><td>Porozumění problematice</td></tr>
<tr><td rowspan="3">Access – shromáždění potřebných dat</td><td>Sample (vybrání vhodných objektů)</td><td rowspan="2">Porozumění datům</td></tr>
<tr><td rowspan="2">Úprava dat a modelování</td><td>Explore (vizuální explorace a redukce dat)</td></tr>
<tr><td>Modify (seskupování objektů a hodnot atributů, datové transformace),</td><td>Příprava dat</td></tr>
<tr><td>Analyze – provedení analýz</td><td>Model (analýza dat: neuronové sítě, rozhodovací stromy, statistické techniky, asociace a shlukování)</td><td>Modelování</td></tr>
<tr><td rowspan="2">Evaluace výsledků implementace</td><td>Akt – přeměna znalostí na akční znalosti</td><td>Assess (porovnání modelů a interpretace)</td><td>Hodnocení výsledků</td></tr>
<tr><td>Automate – převedení výsledků analýzy do praxe</td><td></td><td>Implementace vytvořeného modelu</td></tr>
</table>

Zdroj: Berka, dostupné z: https://sorry.vse.cz/~berka/4IZ450/

2 Data

Datová matice

Jak uvádí vzorec (1) data můžeme také znázornit jako $n \times d$ datovou matici s n řádky a d sloupci, kde řádky odpovídají entitám v datové sadě a sloupce představují atributy nebo zkoumané vlastnosti. Každý řádek v datové matici zaznamenává pozorované hodnoty atributů pro danou entitu. Datová matice $n \times d$ je definována jako:

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \tag{1}$$

kde xi označuje i-tý řádek, což je d-n-tice definovaná jako:

$$xi = (xi1, xi2, ..., xid) \tag{2}$$

a Xj označuje j-tý sloupec, což je n-tice definovaná jako:

$$Xj = (x1j, x2j, ..., xnj) \tag{3}$$

V závislosti na doméně aplikace mohou být řádky označovány také jako entity, instance, příklady, záznamy, transakce, objekty, body, vektory vektorů, n-tice atd. Podobně lze sloupce také nazývat atributy, vlastnosti, vlastnosti, rozměry, proměnné, pole atd. Počet instancí n se označuje jako velikost dat, zatímco počet atributů d se nazývá rozměrnost dat. Analýza jediného atributu se označuje jako jednorozměrná analýza, zatím co simultánní analýza dvou atributů se nazývá bivariantní analýza a simultánní analýza více než dvou atributů se nazývá vícerozměrná analýza (Zaki a Meira, 2014).

Ne všechny datové sady musí být ve formě datové matice. Složitější datové soubory se mohou vyskytovat i ve formě sekvencí (např. DNA a proteinové sekvence), textu, časových řad, obrázků, zvuků, videí atd. Tyto datové soubory pak mohou vyžadovat speciální analyzační techniky. I když prvotní data nejsou ve formě datové matice, v mnoha případech je možno, je do této formy transformovat. Například, pokud máme databázi obrázků, tak můžeme vytvořit datovou matici, kde řádky představují obrázky a sloupce představují funkce obrazu, jako je barva, struktura atd. Určité atributy mohou být spojeny speciální sémantiku vyžadující odlišné zacházení. Například s časovými nebo prostorovými atributy se často zachází odlišně. Je také třeba poznamenat, že tradiční analýza dat předpokládá, že každá entita nebo instance je nezávislá. Vzhledem k vzájemně propojené povaze světa, ve kterém žijeme, tento předpoklad nemusí vždy platit. Instance mohou být připojeny k jiné instanci prostřednictvím různých druhů vztahů, které vedou k datovému grafu, kde uzel představuje entitu a hrana představuje vztah mezi dvěma subjekty (Zaki a Meira, 2014).

2.1 Atributy

Atributy můžeme rozdělit do dvou hlavních typů v závislosti na jejich doméně, tj. v závislosti na typu hodnoty, kterou získávají.

Číselné atributy

Číselný atribut má doménu s reálnou hodnotou nebo s celočíselnou hodnotou. Například Věk s *doménou* (Věk) = N, kde N označuje množinu přirozených čísel (nezáporná celá čísla), je číselný atribut. Zaki a Meira, 2014 dále uvádí, že číselné atributy, které nabývají konečné nebo spočetně nekonečné množiny hodnot, se nazývají diskrétní, zatímco ty, které mohou nabývat jakékoli reálné hodnoty, se nazývají spojité. Zvláštním případem diskrétního číselného atributu je binární atribut. Je možné ho poznat tak, že má jako doménu množinu {0,1}. Číselné atributy lze dále rozdělit na dva typy:

• Intervalově škálované: U těchto druhů atributů mají smysl pouze rozdíly (sčítání nebo odčítání). Například atribut *teplota* měřená ve ° C je intervalově škálována. Pokud máme jeden den 20 ° C a následující den 10 ° C, má smysl tvrdit, že teplota poklesla o 10 ° C. Tvrdit ale že, teplota je dvakrát studenější než teplota předchozí den smysl nemá (Zaki a Meira, 2014).

• Poměrově škálované: U těchto atributů má smysl počítat jak rozdíly, tak i poměry mezi hodnotami. Například pro atribut *Věk* můžeme říct, že někdo, komu je 20 let, je dvakrát tak starší než někdo, komu je 10 let (Zaki a Meira, 2014).

Kategorické atributy

Kategorický atribut je takový atribut, který má doménu se stanovenou hodnotou složenou z množiny symbolů. Například Pohlaví a Vzdělání mohou být kategorickými atributy se svými doménami jako:

doména(Pohlaví) = {M, Ž} (4)

doména(Vzdělání) = {Střední škola, Bc., Ing., Ph.D.} (5)

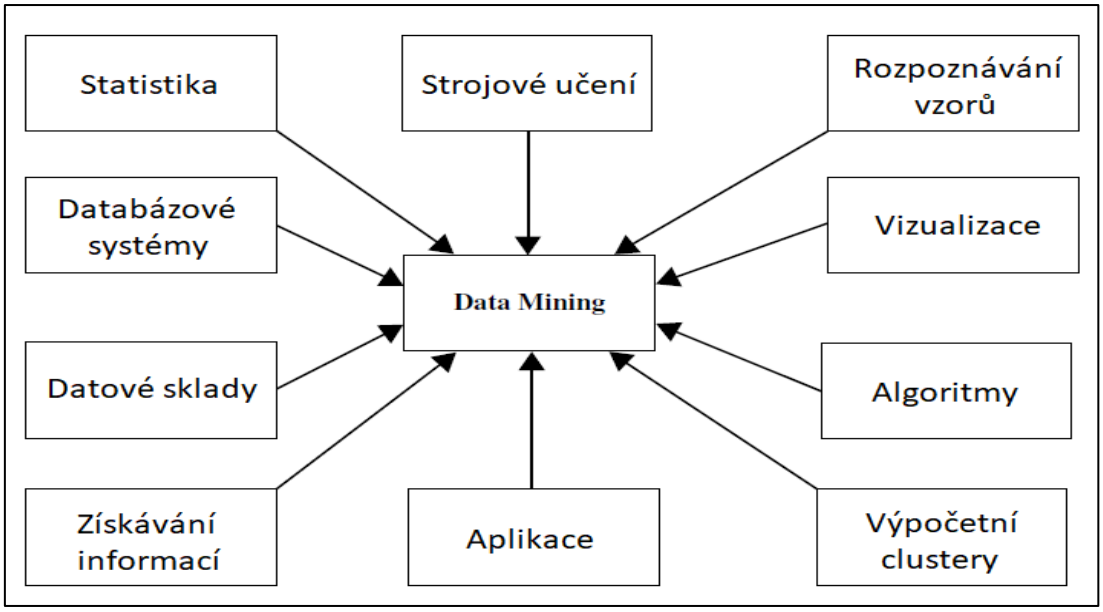
Dle Zaki a Meira, 2014 mohou být kategorické atributy dvou typů:

• *Nominální*: Hodnoty atributů v doméně jsou neuspořádány, a proto má smysl pouze srovnání rovnosti. To znamená, že můžeme zkontrolovat pouze to, zda je hodnota atributu pro dvě dané instance stejná nebo ne (například *Pohlaví*).

• *Pořadové*: Hodnoty atributů v doméně jsou seřazeny, a proto srovnání rovnosti (je jedna hodnota rovná jiné?) a srovnání nerovnosti (je jedna hodnota menší nebo větší než jiná?) jsou povoleny, i když nemusí být možné vyčíslit mezi hodnotami rozdíl. Například *Vzdělání* je pořadový atribut, protože hodnoty jeho domény jsou seřazeny od nižšího vzdělání k vyššímu.

2.2 Technologie a techniky používané v data miningu

Data mining jako metodologie, která je vysoce aplikačně řízená, má v sobě začleněno mnoho technik a technologií z ostatních oblastí. Sem můžeme zařadit statistiku, strojové učení, rozpoznávání vzorů, databáze a systémy datových skladů, získávání informací, vizualizace, algoritmy, výpočetní clustery a mnoho dalších (viz Obr. 3). Interdisciplinární povaha výzkumu a vývoje v oblasti data miningu významně přispívá k úspěchu dolování dat a jeho rozsáhlých aplikací. V této části práce jsou uvedeny příklady několika oborů, které značně ovlivňují vývoj dataminingových metod. (Jiawei a Kamber, 2012)



Obr. 3: Data mining a mnoho oblastí jejichž techniky používá
Zdroj: Vlastní zpracování podle (Jiawei a Kamber, 2012)

2.2.1 Statistika

Jak uvádí Zaki a Meira, 2014, statistika se zabývá sběrem dat, jejich analýzou, interpretací a prezentací. Data mining je na statistiku inherentně navázán. Statistický model je sada matematických funkcí, které popisují chování objektů v cílové třídě z hlediska náhodných proměnných a jejich přidružených rozdělení pravděpodobnosti. Statistické modely jsou používány k modelování dat a datových tříd. V úlohách DM, jako je charakterizace a klasifikace dat, lze sestavit statistické modely cílových tříd. Takové statistické modely pak mohou být výsledkem DM úloh. Případně mohou být DM úlohy postaveny nad statistické

modely. Můžeme použít statistiku k modelování šumu[1] a chybějících dat. Když je poté hledáme vzory ve velké datové sadě, pro DM proces lze použít model, který pomůže identifikovat a zpracovat šum nebo chybějící hodnoty v datech. Statistický výzkum vyvíjí nástroje pro predikci založenou na datech a statistických modelech. Statistické metody lze použít k souhrnu nebo popisu sbírky dat (Jiawei a Kamber, 2012).

Inferenční statistika modeluje data způsobem, který zohledňuje náhodnost a nejistotu v pozorovaných datech a používá se k vyvození závěrů o procesu nebo populaci u výzkumu. Statistické metody lze také použít pro ověření výsledků DM.
Po vytěžení klasifikačního nebo prediktivního modelu, by tento model měl být ověřen statisticky pomocí testování hypotéz. Testování statistických hypotéz (také nazývané jako konfirmační analýza) provádí statistická rozhodnutí za použití experimentálních dat. Výsledek můžeme nazvat statisticky významný, pokud je nepravděpodobné, že by k němu došlo náhodou. Pokud klasifikační nebo predikční model platí, pak popisná statistika modelu zvyšuje jeho spolehlivost (Jiawei a Kamber, 2012).

Využívaní statistických metod při data miningu je velmi náročnou procedurou. Často je poměrně obtížné přijít na způsob, jak rozšířit statistickou metodu na větší sadu dat. Mnoho statistických metod má vysokou výpočetní složitost. Při použití těchto metod na velké datové sady, je potřeba pečlivě navrhnout a vyladit algoritmy tak, aby se co nejvíce snížily výpočetní náklady. Tato problematika je ještě náročnější pro online aplikace, jako jsou návrhy pro online dotazy u webových vyhledávačů, kde je k nepřetržitému zpracování dat v reálném čase vyžadován DM pro datové toky. (Jiawei a Kamber, 2012).

2.2.2 Strojové učení

Strojové učení zkoumá, jak se počítače mohou učit (nebo zlepšit svůj výkon) na základě určitých dat. Hlavní oblastí výzkumu v případě strojového učení je naučit počítačové programy automaticky rozpoznávat složité vzorce a konat inteligentní rozhodnutí na základě poskytnutých dat. Typickým problémem strojového učení je např. naprogramování počítače tak, aby mohl automaticky rozpoznat ručně psané poštovní směrovací čísla v e-mailu poté, co se je naučí ze sady příkladů. Strojové učení je rychle rostoucí disciplína. Níže jsou

[1] Datový šum – jedná se o nesmyslné nebo nerelevantní informace, které poškozují nebo zkreslují data. Systém taková data správně nerozpozná a nedokáže je ani správně interpretovat.

uvedeny klasické oblasti strojové učení, které s data miningem úzce souvisí (Jiawei a Kamber, 2012).

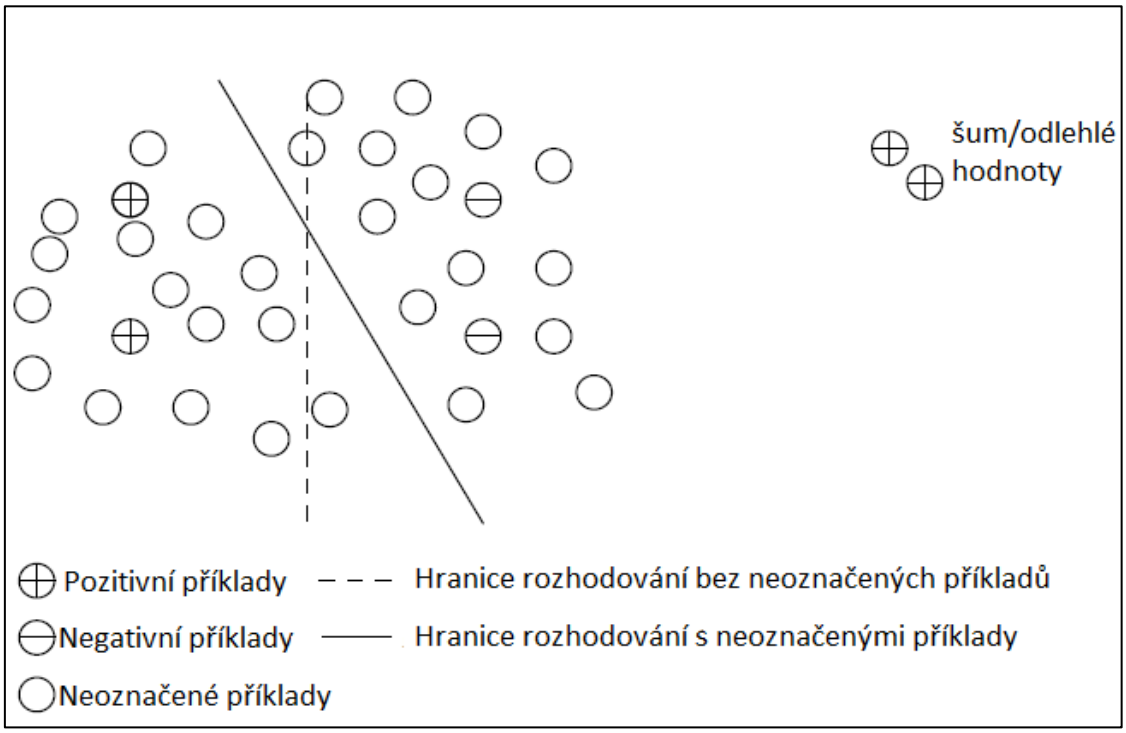
Učení s učitelem (anglicky *supervised learning*) je v zásadě synonymem pro klasifikační problém. Cílem zde je zařadit nový vzorek do určité kategorie pomocí množiny trénovacích dat, která obsahuje vzorky, jejichž kategorie už je nám známá. Například již zmiňovaný problém s rozpoznáním poštovního směrovacího čísla. Sada ručně psaných PSČ a jejich odpovídající strojově čitelné překlady lze použít jako zkušební příklady, které dohlížejí na učení u klasifikačního modelu (Jiawei a Kamber, 2012).

Učení bez učitele (anglicky *unsupervised learning*) je synonymem pro shlukování. Oproti předchozímu typu se jedná o nezávislejší přístup. Počítač se zde učí rozpoznat komplikované procesy a vzorce, aniž by mu člověk dal nějaké bližší vedení. Obvykle můžeme použít shlukování za účelem objevování tříd v datech. U metody učení bez učitele lze např. jako vstup pořídit sadu obrázků ručně psaných číslic. Předpokládejme, že je nalezeno 10 shluků dat. Tyto klastry mohou odpovídat 10 odlišným číslicím od 0 do 9. Jelikož však tréninková data nejsou označena, naučený model nám nemůže říct sémantický význam nalezených shluků (Jiawei a Kamber, 2012).

Kombinace učení s učitelem a bez učitele (anglicky semi-supervised learning) je technika strojového učení, při které je část vstupních dat k dispozici se známým výstupem. Pro další data ale takový výstup známý není. Analytické algoritmy se pak trénují na cvičné množině dat, kde jsou výstupy známé. Jako příklad můžeme uvést bankovní sektor. Algoritmus na podporu rozhodování, zda žadateli schválit, nebo neschválit bankovní úvěr, je natrénován na stovkách tisíc reálných záznamů z minulosti, kde je známý i výsledek, zda klient půjčku splatil, nebo ne (Jiawei a Kamber, 2012).

Nechť jsou zadány označené příklady, které jsou použity pro učení modelů tříd a neoznačené příklady jsou použity pro upřesnění hranic mezi třídami.
V případě problému dvou tříd můžeme množinu příkladů patřících do jedné třídy označit jako *pozitivní příklady* a tu patřící do druhé třídy jako *negativní příklady*.

Na Obr. 4 níže, pokud nebereme v potaz neoznačené příklady, tak přerušovaná čára je hranicí pro rozhodování, která odděluje pozitivní příklady od negativních. Pomocí neoznačených příkladů můžeme zpřesnit rozhodovací hranici na plnou čáru. Navíc to můžeme vypozorovat, že dva pozitivní příklady v pravém horním rohu, i když jsou označeny, jsou pravděpodobně šum nebo odlehlé hodnoty.



Obr. 4: Kombinace učení s učitelem a bez učitele
Zdroj: Vlastní zpracování podle (Jiawei a Kamber, 2012)

Aktivní učení (anglicky *active learning*) je přístup strojového učení, který umožňuje uživatelům hrát aktivní roli v procesu učení. Při aktivním přístupu k učení se lze zeptat uživatele (např. doménového experta), aby označil příklad, který může pocházet ze souboru neoznačených příkladů nebo byl syntetizován při strojovém učení. Cílem tohoto přístupu je optimalizovat kvalitu modelu pomocí aktivního získávání znalostí od lidských uživatelů, vzhledem k omezenému počtu příkladů, u kterých mohou být požádáni o označení (Jiawei a Kamber, 2012).

2.2.3 Databázové systémy a datové sklady

Výzkum databázových systémů se zaměřuje na tvorbu, údržbu a používání databází pro organizace i koncové uživatele. Databázové systémy jsou vysoce užívané v datových modelech, dotazovacích jazycích, zpracování dotazů a optimalizačních metodách, dále pak pro ukládání dat a metody indexování a přístupu. Tyto systémy jsou dobře známé pro svou vysokou škálovatelnost při zpracování objemných, relativně strukturovaných souborů datasetů. Mnoho dataminingových úkolů se musí umět vypořádat s velkými datasety nebo rychle streamovat v reálném čase. Z toho důvodu je DM dobře využitelný pro technologie škálování databází, kde je schopný dosáhnout vysoké účinnosti a škálovatelnosti na velkých souborech dat. DM úlohy lze také použít k rozšíření schopnosti stávajících databázových systémů uspokojit pokročilé uživatele se sofistikovanějšími požadavky na analýzu dat (Jiawei a Kamber, 2012).

Datové sklady

U novějších typů databázových systémů jsou do databáze zabudovány možnosti systematické analýzy dat využívající datové sklady a zařízení pro dolování dat. Datový sklad je v podstatě speciální typ relační databáze, kde je umožněno řešit úlohy zabývající se analytickým dotazováním nad objemným množstvím dat. Tento sklad integruje data pocházející z více zdrojů a různých časových rámců. Data jsou uspořádána ve vícerozměrném prostoru k pomocí datových kostek. Model datové kostky nejenže usnadňuje technologii OLAP[2] ve vícerozměrných databázích, ale také podporuje multidimenzionální dolování dat (Jiawei a Kamber, 2012).

2.2.3.1 Datová kostka (OLAP kostka)

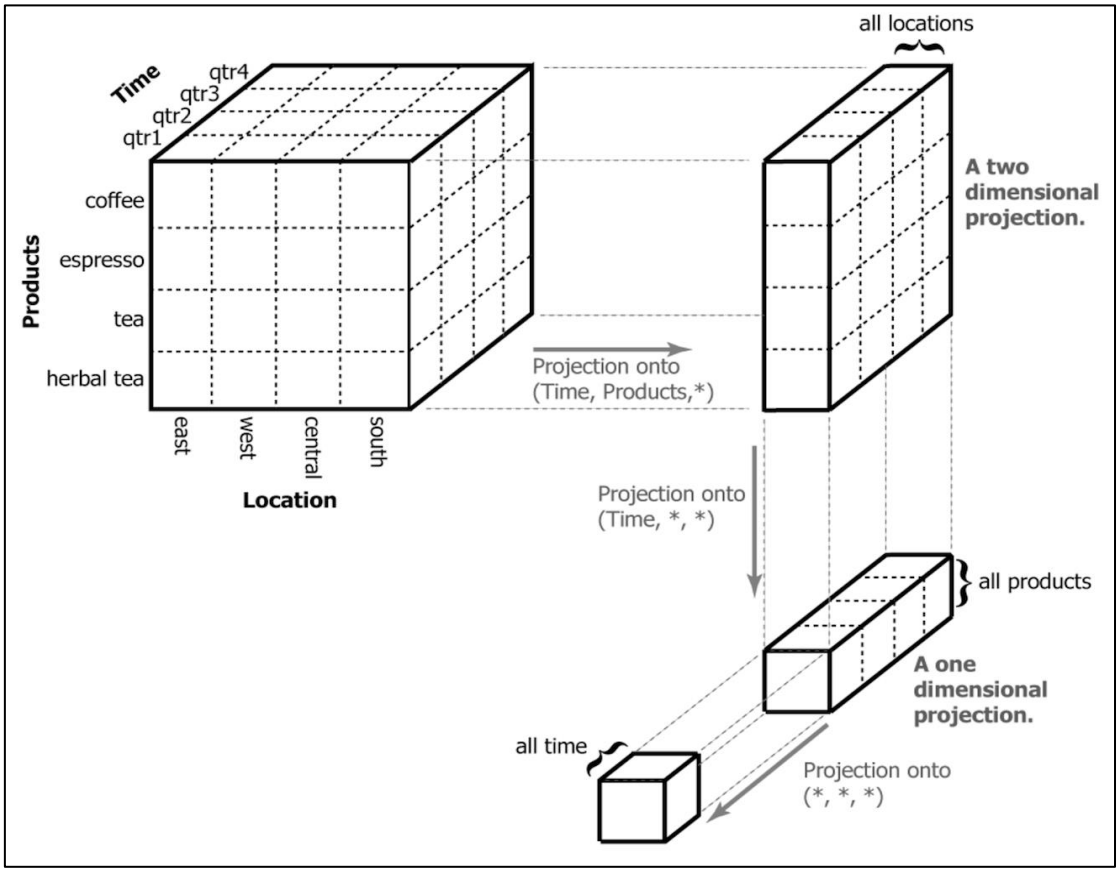
Jedná se o způsob uspořádání dat, který rozšiřuje dvourozměrný model tabulkového uspořádání takovým způsobem, že každá dimenze je uložena v jedné ose kostky (viz Obr. 5 na další straně). Krom toho, že jsou datové kostky široce používány, také poskytují výkonný mechanismus k abstrakci dat. Datové kostky uživatelům rychle poskytují souhrny dat s různou úrovní podrobností, spíše než shrnutí, jako je agregace každých dvou záznamů.

[2] OLAP (Online Analytical Processing) technologie uložení dat v databázi, která umožňuje uspořádat velké objemy dat tak, aby byla data přístupná a srozumitelná uživatelům zabývajícím se analýzou obchodních trendů a výsledků

Tohoto je dosaženo vytvořením mřížky datových kostek, které představují data na různých úrovních podrobností podle sémantické hierarchie a poskytnutých mechanismů pro souhrn každé kostky (Stolte a Tang, 2003).

Díky tomu, že jsou data uspořádána do vektorů kostek, lze k nim zpětně přistupovat z různých hledisek (dimenzí – místo, produkty, čas, viz níže). Není proto nutné spojovat mnoho tabulek relačních databází, což je pro systém zátěž. Toto má ale i své nevýhody, protože fyzické ukládání dat do kostek znemožňuje rychlou editaci. Pokud bychom editaci chtěli provést, pak musíme přepracovat celou kostku. Kostku tvoří hodnoty, které jsou zařazeny do dimenzí (Stolte a Tang, 2003).

Na Obr. 5 níže je znázorněn rozpad datové kostky, kde můžeme vidět v levém horním rohu, že kostka je složena z dimenze místa, která se rozpadá ještě na jih, východ, západ a střed. Dále zde máme dimenzi produktů, kterou tvoří kafe, čaj, espresso a bylinkový čaj. Třetí dimenzi pak tvoří čas, který se dále rozpadá na čtyři kvartály.



Obr. 5: Model znázorňující rozpad datové kostky

Zdroj: Stolte a Tang, 2003

3 Internetový marketing

Internet značnou měrou přispěl k rozvoji marketingu. Díky tomu, že informace jsou dnes snadno dostupné, lidé mají možnost porovnat nabídku včetně cen, můžou si vyměňovat si názory na produkty na diskuzních fórech, hodnotit je a také je prostřednictvím internetu nakupovat. Na celou věc lze nahlížet jako na obrovskou tržnici, kde se vyskytuje téměř neomezená nabídka produktů s velmi dobře informovanými zákazníky. Prostor na internetu je otevřený všem a na velikosti firem nezáleží. V dnešní době, pokud je jednotlivec nebo malá firma schopná a umí dobře využívat pestré možnosti reklamy, podpory prodeje, public relations, případně přímého marketingu, může mít velký vliv na zákazníky (Janouch, 2014).

V dnešní době je internetový marketing mnohem významnější než klasický marketing především tam, kde lidé používají vyspělé technologie. Klasický marketing a ten internetový od sebe ale nelze oddělovat. Marketing je jen jeden a dokonce i firmy, které prodávají pouze přes e-shop, komunikují se zákazníky přes offline média, řeší přepravu nebo třeba cenovou politiku. Oproti tomu mnohé jiné firmy zcela opustily offline prostředí, co se týče marketingu. V určitých případech to může být odůvodněné, ale obecně je to chyba (Janouch, 2014).

3.1 Charakteristika marketingu na internetu

Od klasického marketingu se ten internetový v několika aspektech liší. Internetový marketing je způsob, jakým lze dosáhnout vytyčených marketingových cílů skrze internet. Podobně jako klasický marketing zahrnuje celou řadu aktivit spojených s ovlivňováním, přesvědčováním nebo udržováním vztahů se zákazníky. Internetový marketing se soustředí hlavně na komunikaci, ale často je spjat také s tvorbou cen (Janouch, 2014).

Internetový marketing je často označován jako e-marketing, web-marketing nebo taky online marketing. Často se setkáváme taky s pojmem digitální marketing. V online marketingu a digitálním marketingu je zahrnut, kromě internetového marketingu, také marketing prostřednictvím mobilních zařízeni (Janouch, 2014).

Dnešní moderní marketing vyžaduje osobní přístup, péči o každého jednotlivého zákazníka a možnost individualizace určité služby nebo produktu. A musí být komplexní. Pokud se aktivity realizují jednotlivě, pak ztrácí smysl. Tento komplexní přístup má několik složek – vztahovou, integrovanou, interní a společensky zodpovědný marketing. Internetový marketing je navíc kontinuální činnost, jelikož podmínky jsou nepřetržitě měněny (Janouch, 2014).

Janouch, 2014 pak dále uvádí, že kvůli novým technologickým možnostem firmy musely začít hledat jiné způsoby oslovení zákazníků. Komunikační prostředky a formy marketingové komunikace se tak zásadně rozšířily a zákazníci se dostali do zcela jiného postavení, než tomu bylo dříve. Internetový marketing je z tohoto důvodu charakterizován právě ve vztahu k nim. Internetový marketing obsahuje:

- konverzaci
- posílení pozice zákazníka
- spoluúčast

Konverzace

Konverzace na internetu je v podstatě jakýmsi trhem. Lidé spolu komunikují naprosto bez zábran, o čemkoliv, s mimořádnou rychlostí. Na tržištích se lidé potkávali už ve starověku. Nejenom aby kupovali a prodávali, ale hlavně aby se setkali a mluvili spolu. Část této konverzace byla o obchodě a produktech, část byly novinky, názory nebo drby. Poté ale nastala doba velkovýroby a masmédií a došlo k odcizení prodávajícího a kupujícího v obrovském měřítku. Internet tuto situaci zachraňuje a vrací zpět. V dnešní době internet ke konverzaci přímo vyzývá a pokud chtějí firmy přežít, tak se musí přizpůsobit (Janouch, 2014).

Posílení pozice zákazníka

I přesto, že žijeme v moderní době a máme k dispozici nejrůznější sociální média a komunikační zařízení, komunikace mnoha firem se zákazníky stále ještě probíhá formálně. Zvláštním jazykem jsou sdělovány naprosté samozřejmosti jako např. slogany – „jsme tu pro vás“, „řídíme se vašimi požadavky“ atd. Tyto informace jsou sdělovány bez humoru, bez nápadu, nebo zatajováním věcí, případně se lhaním (Janouch, 2014).

Pozice zákazníka je čím dále silnější. Je tomu tak, protože má kolem sebe síť, která je několikrát násobně větší, než si bylo možné kdykoliv v minulosti představit. Každý uživatel internetu může najít řešení svého problému nejen na sociálních sítích, ale také vyhledáváním ve vyhledávačích, procházením diskusních fór, položením svého dotazu na nejrůznějších portálech nebo třeba ve Wikipedii
(Janouch, 2014).

Propojení lidí může firmu velmi rychle zlikvidovat, např. pokud má špatné recenze nebo také posunout mezi nejvýznamnější hráče na trhu. Trh na síti nemá žádné slitování s firmami, které jsou neochotné nebo neschopné se přizpůsobit dnešním vysokým požadavkům zákazníků a rychlým změnám. Podstatou businessu se opět stává člověk (Janouch, 2014).

Spoluúčast

Problém tkví v tom, že v dnešní době se manažeři většinou odcizili produktu – už nerozumí tomu, co firma vyrábí a jak se co dělá – proto je nutné začít využívat lidí spojených s produktem (lidi z výroby, vývoje, servisu atd.) a propojit je se zákazníky. Byť to může znít jako klišé, je také nutné nazývat věci pravými jmény. To, že použijete frázi jako „máme řešení“, když prodáváte počítače, je úplně zcestné. Protože nemáte řešení, ale prostě prodáváte počítače! Firmy, které jsou poučené a chytré se snaží své zákazníky zapojit do procesu vývoje nebo třeba přizpůsobování produktů. Jedině tímto způsobem si lze zajistit loajalitu zákazníků a své budoucí zisky (Janouch, 2014).

3.2 Diferencovaný přístup k zákazníkům

Internet nutí firmy měnit své koncepce marketingové strategie. V počátcích vývoje marketingu firmy uplatňovaly strategii masového marketingu. V případě tohoto typu marketingu je poskytována zákazníkům hodnota na základě jejich převažujících charakteristik na konkrétním trhu. Takový přístup můžeme označit jako one-size-fits-all (jedna velikost padne všem). Problém je v tom, že, taková strategie má dnes již velmi malou šanci na úspěch, i přesto určitě mohou existovat výjimky (Janouch, 2014).

Nyní se uplatňuje mnohem úspěšnější strategie. Tou je strategie cílení na určité segmenty trhu, které mohou být jak větší, tak menší. I přesto ale tato strategie stále neodráží skutečnost, že některý zákazník je pro firmu mnohem důležitější než jiný. Tomuto zákazníkovi je

potřeba vyjít vstříc v nabídce a péči o něj. Firma musí také vytvářet vyšší hodnotu pro zákazníky. Zde přichází na řadu uplatňování strategie CRM (Customer Relationship Management), resp. řízení vztahů se zákazníky (Janouch, 2014).

Vývoj CRM strategie zahrnoval tři stadia. Dle Janoucha, 2014 v počátcích byla uplatňována strategie *masové personalizace*, která je determinována nabídkou standardních produktů, ale firma již se zákazníkem komunikuje individuálně na základě jeho rozpoznání (zná jeho jméno, nebo nákupní chování). Dalším stadiem je *masová kastomizace*. Ta přináší zákazníkům vyšší hodnotu tím, že firma zákazníkovi přizpůsobí produkt podle jeho požadavků.

Stále jde však o standardní produkty, které však mohou mít různé vlastnosti a tím odlišný užitek pro zákazníka. Podle Janoucha, 2014 je typická je možnost sestavení finálního produktu z různých komponent. Případně výběrem požadovaných funkcí (nehmotné produkty). Rozdílné požadavky zákazníků však reflektuje až třetí stadium označované jako *diferencovaná kastomizace*. V případě uplatňování této strategie jsou produkty vytvářeny přesně pro konkrétního zákazníka. A s tím se pojí také individualizovaná marketingová komunikace.

Jak uvádí Janouch, 2014, uplatnění všech těchto uvedených typů strategie je podporováno internetem. Rozpoznání zákazníka je zajištěno nástroji pro analýzu návštěvnosti, informacemi sbíranými při nákupech v elektronických obchodech, informacemi ze sociálních médií, vlastními webovými stránkami a marketingovou komunikací se zákazníky obecně. Internetové technologie umožňují sestavování finálního produktu zákazníky samými. Marketingová komunikace se zákazníky se tak prostřednictvím internetu významně posouvá od jednostranné komunikace ke komunikaci oboustranné. Firmy sice mají k dispozici nejrůznější informace o zákaznících, ale na druhou stranu si zákazníci mohou opatřit prakticky jakékoliv informace o firmě nebo jejích produktech. Oboustranná komunikace se zákazníky v dnešní době není jen možnost, ale naprostá nutnost.

Marketing současnosti se vyznačuje krom jiného i tím, že zkoumá hodnotu, jakou přináší produkt zákazníkovi, a zároveň jaký zisk zákazník přináší firmě. Na základě tohoto pak firmy přizpůsobují produkt, cenu i komunikaci konkrétním zákazníkům. Zdánlivě je to možné a využívané pouze v rámci B2B, avšak rozvoj informačních technologií postupně

přináší diferencovaný přístup také do B2C oblasti (výhodou e-shopů je neanonymní zákazník). Zde stojí za zmínku, že zejména v B2B mohou být rozdíly mezi zákazníky poměrně značné a od cílení na segmenty se proto přechází k individualizovanému marketingu (Janouch, 2014).

S využitím informací od zákazníků získaných v rámci marketingové komunikace může firma přizpůsobit daný produkt jejich požadavkům, preferencím a přáním. Spokojený a vracející se zákazník znamená pro firmu zisk, a proto je třeba takové zákazníky doslova hýčkat. K tomu je Marketing na Internetu ovšem nutné mít přehled a zpracovávat velké množství informací. Z řízení vztahů se zákazníky (CRM) se stává strategická záležitost (Janouch, 2014).

Řada firem dělá v přístupu k zákazníkům jednu závažnou chybu. Pro nové zákazníky nabízí všemožné výhody, zato stávající zákazníky považuje za jisté, a dokud budou platit, tak se o ně nebude zajímat. Tento přístup vidíme například u mobilních operátorů, dodavatelů energií, bank apod. Převážně tedy na trzích, kde mohou vzhledem k omezenému počtu hráčů na trhu existovat kartelové dohody nebo jde o odvětí s velmi obtížným vstupem (vysoké investiční náklady, regulace ze strany státu, právní překážky). Takže v žádném případě neignorujte vaše stávající zákazníky. Za jejich věrnost jim naopak nabídněte nějaké výhody (Janouch, 2014).

3.3 CRM systémy

S procesem přizpůsobování produktů a celého marketingového mixu zákazníkům souvisí využívání systémů CRM. Strategii diferencovaného přístupu lze jen těžko uplatnit bez systému, který umožňuje shromažďovat data o zákaznících a další významné údaje – požadavky, preference, názory na produkty atd. Také je třeba monitorovat každý čin ve vztahu se zákazníkem – telefonní rozhovory, e-mailová komunikace, osobní setkání, nákupy, reklamace a další. Veškerá tato data jsou automaticky nebo ručně zapisována do systému CRM. S daty ze systému se pak dále pracuje. Informační systémy pomáhají data identifikovat, integrovat a následně analyzovat. Bez analýzy dat nelze zjistit, jak zákazníci vnímají hodnotu produktu. Bez analýzy dat také nelze produkty přizpůsobit a tím ani zajistit firmě zákazníky s maximálním přínosem (Janouch, 2014).

Payne, 2005 pak ještě tvrdí, že systém CRM, lze rozdělit do několika klíčových oblastí, které do sebe komplexně zapadají a mají návaznost na fungování celku, jako takového. Dále lze však tyto procesy rozdělit ještě do dvou základních skupin, které fungují kooperativně:

Operativní CRM – zahrnuje samotnou podporu obchodu a celkovou komunikaci se zákazníkem, tyto údaje jsou předávány k analýze, která vyhodnocuje klíčové aktivity, slabá místa, potenciální hrozby a příležitosti. Při této úrovni systému jsou nastaveny otázky, která systém musí pevně dodržovat:

- Jak si udržet stávajícího zákazníka?
- Jak porozumět zákazníkům?
- Jak správně naslouchat zákazníkovi?
- Jak správně rozpoznat klíčové procesy?
- Jak zvýšit spokojenost zákazníka při zdokonalování procesů?
- Jaká by měla být marketingová strategie pro nové zákazník a udržet stávající?
- Jaké schopnosti jsou potřeba k oslovení nových zákazníků?

Jak dále uvádí Payne, 2005, cílem **analytického CRM** je analýza zákaznických dat, pomocí které bude firma schopna dosáhnout různorodých cílů:

- Optimalizace efektivnosti marketingových kampaní a jejích vyhodnocení
- Hledání potenciálních prodejních kanálů (cross-selling, up-selling a udržení zákazníka)
- Analýza chování zákazníků – tvorba cen a vývoj nových výrobků
- Podpora v rozhodování – předpovídání a analyzování zákaznické rentability

Kolaborativní CRM

Kolaborativní řízení vztahů se zákazníky obsahuje speciální funkcionalitu, která umožňuje společnosti a jejím zákazníkům komunikovat prostřednictvím nejrůznějších kanálů za účelem dosažení vyšší kvality interakce se zákazníky. Payne, 2005 dále uvádí, že tento typ CRM nabízí užitečné informace, které vznikají např. při jednání se zákazníkem, nebo také při interakci s obchodním oddělením, jako je prodej, technická podpora a marketing.

Může se jednat například o poskytnutí informací o specifických zákaznických požadavcích, případně dotazy ohledně nové služby z technické podpory prodeje marketingu. Cílem tohoto CRM je sdílení uvedených informací získaných ze všech oddělení, díky čemuž lze zvýšit kvalitu poskytovaných služeb pro zákazníky (Payne, 2005).

Jak je uvedeno na Obr. 6 níže, řízení vztahů se zákazníky dává do spojitosti oblasti strategického plánování, marketingu, prodejů, doručovacích možností, podpory a reportů a analýz. Pokud je byť jedna z oblastí CRM narušena, může dojít ke ztrátám v podniku, a proto by ani jedna z těchto složek neměla být podceňována.

Obr. 6: Oblasti řízení vztahů se zákazníky
Zdroj: https://www.jakzacitpodnikani.cz/proc-pouzivat-crm-system

Budování důvěry

Jak uvádí Janouch, 2014, v průběhu konverzace mezi firmami dochází k vzájemnému poznávání dodavatele a zákazníka. Důvěra je budována pomalu, ale pak má většinou dlouhodobý účinek. Pokud se lidé se mezi sebou dobře znají, dokážou pak řešit vzniklé problémy nekonfliktním způsobem. Důvěru je ale nutno budovat již ve fázi, kdy zákazníka teprve získáváme. K tomuto účelu slouží především:

- kvalitní webové stránky
- články, zprávy, recenze, výukové materiály
- vertikální odkazy (mezi firmami v dodavatelském řetězci)

Se stávajícími zákazníky je nutné vztahy dále rozvíjet například pomocí:

- uzavřených blogů a diskusí (jen pro zákazníky)
- posilováním sociálních aspektů zejména během offline aktivit

Cílem je vytvářet vztahy, které vyústí v prodej a také v získání cenných informací od zákazníků směrem k firmě (Janouch, 2014).

3.4 Předpoklady úspěchu na internetu

Stejně jako při jakékoliv činnosti, tak také v marketingu na internetu platí, že musíme vědět, co chceme. Bez strategie, resp. *definování cíle*, nelze provádět žádný marketing, a to ani v 21. století.

Pokud neznáme cíl, nemůžeme volit taktiku, tedy způsob, jak toho cíle dosáhnout. Na začátku je nutné získat co nejvíce informací o zákaznících a konkurenci, protože poznání zákazníka je klíčem k úspěchu. Protože se lidé snaží v záplavě informací vše filtrovat, vyhnout se tomu, co nechtějí slyšet (vidět, vědět), je třeba zaměřit úsilí na komunikaci s takovými zákazníky, které vaše sdělení skutečně zajímá (Janouch, 2014).

Lidé věnují daleko více času komunikaci než obsahu samotnému. Také je známo, že si na internetu předávají informace mnohem více mezi sebou, než aby se je pokoušeli získat z oficiálního zdroje. To je způsobeno tím, že oficiálním zdrojům informací moc nevěří. V takové situaci je pro firmy čím dál těžší se do omezeného vnímání prosadit přímo. Oproti tomu ale důvěra v informace od jiných lidí je poměrně vysoká. Jde hlavně o zkušenosti těch, co už nějaký produkt či službu zakoupili. Marketéři by se měli snažit takové lidi si získat a taky udržet. Pro všechny aktivity je ale nezbytné mít především kvalitní produkt. Jako příklad můžeme uvést prohlášení „když nebudete spokojeni, vrátíme vám peníze“ . Toto už předem vyvolává pocit, že si zákazník kupuje kvalitní produkt. Faktem je, že lidé jsou velmi

skeptičtí, ale když je něco opravdu dobré, pak propadnou velkému nadšení a rádi a ochotně tyto informace předávají dál (Janouch, 2014).

Z faktů uvedených výše vyplívá, že uspět na internetu znamená hlavně:

.

- mít kvalitní produkt
- stanovit si reálné cíle
- umět poznat zákazníka
- dobře komunikovat

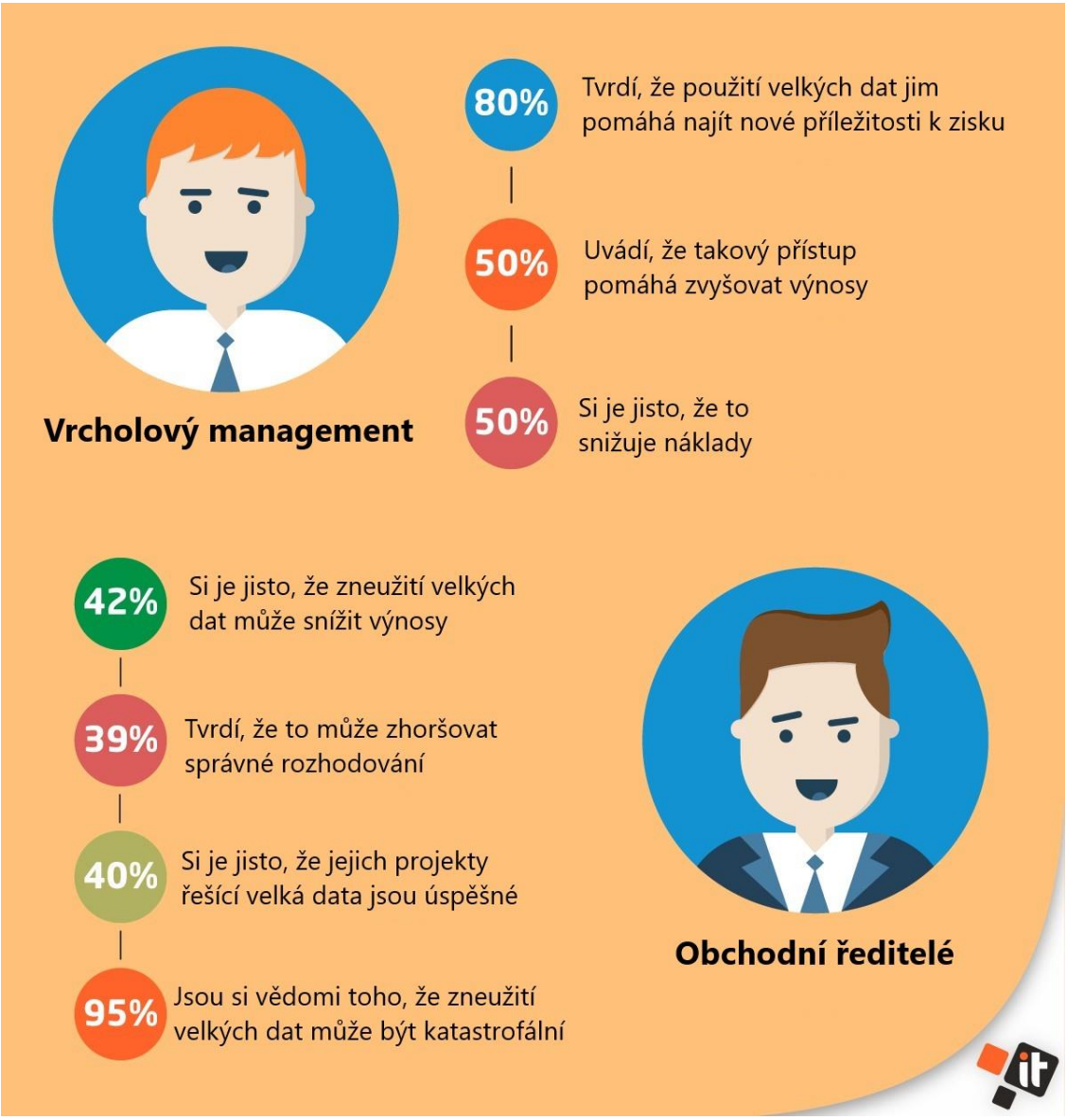
3.4.1 Informační mlha

V literatuře se můžeme setkat také s pojmem *informační mlha*. Jedná se o situaci, kdy je spotřebitel neustále obklopen takovým množstvím informací, že je pro něj velmi obtížné rozlišit, co je důležité a co ne. A proto na něj klasická reklama už nemá takový vliv, jak by za normálních okolností měla. Informační mlha není situací vyskytující se pouze mimo internet, kdy např. jedete po dálnici a míjíte billboardy, které už mozek z důvodu přehlcení ignoruje. S informační mlhou se můžeme setkat i na internetu, kde jsme obklopeni nepřeberným množstvím nabídek a personalizovaného obsahu.

Reklamy se v dnešní době stávají agresivnějšími a sofistikovanějšími. Sice zde existují nejrůznější rozšíření webových prohlížečů, které jsou poměrně účinné (např. AdBlock), ale problém je v tom, že webová stránka vám neumožní stáhnout požadovaný soubor, případně přečíst si článek, za předpokladu, že máte rozšíření aktivované. Reklamám se tak nevyhnete.

4 Možnosti zneužití data miningu

Potenciální hrozbu data miningu a velkých dat si uvědomuje i vedení podniků (viz Obr. 7). Je otázka, zda je této problematice věnována dostatečná pozornost, protože, jak někteří odborníci předpovídají, umělá inteligence používající DM techniky je pro lidstvo nevyhnutelná. V následujících 3 kapitolách budou rozebrány oblasti, kde se tento problém už bohužel vyskytuje.



Obr. 7: Názory odborníků z oblasti businessu na problematiku zneužití velkých dat a DM

Zdroj: Vlastní zpracování podle: https://towardsdatascience.com/big-data-misuse-can-break-your-business-ef6432dfd188

4.1 V business světě

Waxer, 2013 uvádí, že big data [3]nyní představují obchodní příležitost za několik miliard dolarů. Organizace, které toho umí využít, dosahují obrovských zisků. Je jedno, zda se jedná o maloobchodníky nebo velké výrobce. Ti, kteří pochopili, jak cenná jsou např. poštovní směrovací čísla spotřebitelů a historie již realizovaných nákupů, velmi snadno zvyšují svůj celkový zisk.

Odborníci odhadují, že big data mohou zvýšit zisky v maloobchodě až o jen těžko představitelných 60 %. Stejně tak se předpokládá, že osobní údaje mohou společnostem pomoct dosáhnout vyšší obchodní efektivity a přizpůsobit nové produkty (Waxer, 2013).

I když je využití síly datové analýzy nepochybně konkurenční výhodou, horlivý data mining může snadno selhat. Společnosti se sice stávají odborníky datovou analýzu a jsou schopny z dat vyčíst podrobnosti tak osobní, jako jsou třeba možnosti selhání splacení hypotéky u klienta, rizika infarktu a dalších mimořádně citlivých dat, tak narůstá hrozba závažného narušení soukromí (Waxer, 2013).

Prodejce moderního oblečení Urban Outfitters čelil hromadné žalobě za údajné porušení zákonů na ochranu spotřebitele tím, že kupujícím, kteří platili kreditní kartou, sdělil, že musí poskytnout své PSČ. Což ale nebyla pravda - a poté tyto informace použil k získání adresy nakupujících
(Waxer, 2013).

Naprostou marketingovou legendou se ale stala kauza z Minnesoty, kdy obchodní řetězec Target věděl o těhotenství náctileté dívky dříve než její otec. Na základě dat z její zákaznické karty (nákup volného oblečení, tělového mléka a kyseliny listové) jí začal zasílat slevové kupony na pleny, což vedlo k prozrazení do té doby utajeného těhotenství (Waxer, 2013).

[3] Big data - jde o takové soubory dat, jejichž velikost je mimo schopnosti zachycovat, spravovat a zpracovávat data běžně používanými softwarovými prostředky v rozumném čase

4.2 V oblasti medicíny

To, co znepokojuje vědce a další odborníky z celého světa jsou možnosti genetického inženýrství, které byly až do nedávna vnímány jako pouhé sci-fi. Snaha o dosažení co nejlepšího genetického fondu člověka je tu s námi už poměrně dlouhou dobu. Touto problematikou se zabývá eugenika.

Tento pojem použil poprvé v 19. století britský matematik a vědec Francis Galton. Eugeniku můžeme rozdělit na pozitivní a negativní. Pozitivní eugenikou se chápe snaha o rozšíření žádoucích znaků v populaci. Jako negativní eugenika je pak označována snaha o vymýcení znaků nežádoucích z populace (Spektorowski a Saban-Ireni, 2016).

Ačkoli byla původní myšlenka eugeniky ušlechtilá, bohužel se do lidské historie zapsala velmi negativně. Bylo tomu tak především kvůli programům nacistického Německa, které se snažilo vytvořit „nadčlověka“. Eugenické programy také používali např. v USA v letech 1902 až 1964, přičemž se jednalo především o nedobrovolné sterilizace mentálně postižených, epileptiků a vězňů coby nositelů „defektních“ a „nežádoucích“ dědičných znaků (Spektorowski a Saban-Ireni, 2016).

Jak uvádí Spektorowski a Saban-Ireni, 2016, kritikům eugeniky vadí především představa cílevědomého šlechtění lidského genofondu. Je tomu tak hlavně z politických a etických důvodů. Také není jasné, kdo by měl rozhodnout, které lidské vlastnosti jsou ty požadované. A ve spojitosti s tím kritici upozorňují na to, kam eugenické postupy vedly v minulosti (rasistické, antisemitistické, homofobní postoje historické většiny eugeniků, jež vyústily v podobě nacismu).

Co se týče aktuální eugenické praxe, tak v současnosti se objevují etické otázky v souvislosti s umělým oplodněním, preimplantačním genetickém testování embryí a možnost rozvoje genové terapie (Spektorowski a Saban-Ireni, 2016).

Co má s tím vším společného data mining?

Pomocí data miningu bude možné např. dekódovat lidský genom, což nevyhnutelně povede k eugenickým procesům, byť postaveným na vědeckém základě. Může tak dojít k situaci, že si bude možné zaplatit za zvýšenou imunitu či inteligenci vašeho dítěte, případně modrou

barvu jeho očí. Před podobnou myšlenkou varoval i prof. Hawking, který tvrdil, že v budoucnosti se lidé budou dělit na ty „vylepšené“ a „nevylepšené“ a že rasa nadlidí bude jednou koncem pro lidstvo, jak ho dnes známe.

4.3 V politice

Dalším velikým rizikem zneužití data miningu jsou politické účely, byť možná z určitého úhlu pohledu dobře míněné. Představte si společnost, ve které vás vláda hodnotí jako důvěryhodné nebo nedůvěryhodné občany. Je vám přiděleno skóre, které se neustále mění na základě toho, jak se chováte. Za „vhodné“ chování získáváte plusové body a za to „nevhodné“ vám body naopak sráží. To se na první pohled nemusí nutně jevit negativně. Problém je ale v tom, co daný stát označuje za nevhodné chování.

Pokud okradete člověka, pak je sražení bodů celkem logickým krokem. Jenže v případě Číny a jeho Systému sociálního kreditu se za „nevhodné“ chování považuje např. také zveřejnění politických příspěvků online bez povolení, nebo pochybování a odporování oficiálnímu příběhu vlády o aktuálních událostech a jakékoli další kroky proti režimu. Oproti tomu vysoké skóre vám umožní přístup k rychlejším internetovým službám nebo zrychlenému vízu do Evropy.

Vaše „skóre občana“ vás sleduje, ať jste kdekoli. Pro výpočet skóre musí soukromé společnosti spolupracující s vládou neustále procházet obrovským množstvím dat ze sociálních médií a online nakupování a dalších oblastí. Vzhledem k objemu dat, který je nutno zpracovat, je jasné, že je pro tyto účely používán data mining.

Jakmile opustíte svůj byt, vaše akce ve fyzickém světě jsou ihned zaznamenány do sítě. Vláda shromažďuje obrovské množství informací prostřednictvím videokamer umístěných na ulicích a po celém městě. Pokud spácháte trestný čin, kam lze mimochodem zařadit i nepoužití přechodu pro chodce, když přecházíte ulici, algoritmy rozpoznávání obličeje spojí videozáznamy vaší tváře s vaší fotografií v národní databázi ID. Nebude to dlouho trvat a u vašich dveří se objeví policie (Mitchell a Diamond, 2018).

Tato společnost se může zdát dystopická, určitě se ale nejedná o nějakou vzdálenou budoucnost. Čína plánuje spuštění Systému sociálního kreditu ještě tento rok. Usiluje o to, aby jako první zavedla všudypřítomný systém algoritmického sledování. Čínská komunistická strana využívá pokroky v oblasti umělé inteligence a data miningu a ukládání dat k vytvoření podrobných profilů všech občanů. Rozvíjí „skóre občanů“, aby stimulovalo „dobré“ chování. Obrovská doprovodná síť monitorovacích kamer bude neustále sledovat pohyby občanů, údajně s cílem omezit kriminalitu a terorismus (Mitchell a Diamond, 2018).

Čínský vyvíjející se algoritmický sledovací systém se bude spoléhat na to, že bezpečnostní orgány státu komunistické strany budou filtrovat, shromažďovat a analyzovat ohromující objemy dat proudících přes internet. Byť máme dnes k dispozici možnosti umělé inteligence, tyto systémy ještě nejsou natolik dokonalé, aby mohly samy neomylně rozhodovat o tak komplikovaném jevu, jako lidské chování (Mitchell a Diamond, 2018).

Čína původně plánovala vyvinout systém dohledu „Zlatý štít“ umožňující snadný přístup k místním, národním a regionálním záznamům o každém občanovi. Tento projekt byl ale až doposud omezen na filtrování obsahu Great Firewall, který zakazuje zahraniční internetové stránky včetně Google, Facebook a The New York Times. Podle organizace Freedom House, jejímž cílem je hájení demokracie tvrdí, úroveň svobody na internetu v Číně je již nejhorší na planetě. Komunistická strana Číny nyní díky data miningu konečně buduje rozsáhlý víceúrovňový systém sběru dat, o kterém po celá desetiletí tak snila (Mitchell a Diamond, 2018).

5 Představení subjektu řešící DM problém – IT firma BROKEN MOUSE, s.r.o.

Autor této práce je členem menší IT firmy, sídlící v Jablonci nad Nisou, která se krom jiného zabývá řešením v oblasti e-shopů, webů, marketingu a IT. Tým je složen z nadšenců do moderních technologií, vystupujících pod hlavičkou Broken Mouse s.r.o. Firma si zakládá na dobré komunikaci, dlouhodobosti a výsledcích. V oboru IT se tato firma pohybuje již 10 let a od roku 2014 se začíná rozrůstat v Jablonci nad Nisou. Na svém webu se prezentuje následujícím sloganem:

„Prošli jsme si mateřskou školkou i univerzitou. Propojujeme poznatky akademického světa s praxí. Úzce spolupracujeme se vzdělávacími institucemi i firmami napříč obory. Máme rádi osobní přístup, férovost, dobrou náladu a dlouhodobé rostoucí partnerství“.

Firma má poměrně široké pole působnosti a zaměřuje se na následující oblasti:

- tvorba nového / úprava stávajícího e-shopu na platformě Shoptet
- napojení e-shopu na nejrůznější systémy, vč. zaškolení
- web design, tvorba webů, UX
- grafika, fotografie, bannery, loga, katalogy, letáčky + tisk
- kompletní marketingová strategie
- PPC, SEO
- zbožáky (Heuréka, Zboží, Google nákupy…)
- emailing, copywriting, analýzy
- správa a inzerce na sociálních sítí
- správa a servis IT (servery, počítače, sítě, kamery, SW…)
- expanze (kolegové rodilí mluvčí)
- poradenství a školení (angličtina, IT, MS Office…)

Krom výše zmíněného firma nabízí poradenství v oblasti marketingu, např. se zaměřením na rozvoj e-shopů. Již delší dobu probíhá spolupráce s firmou, která krom kamenného

obchodu provozuje i e-shop. Jeden z kolegů a konzultantů této práce je expertem na online marketing a již delší dobu má na starosti marketingové aktivity internetového obchodu.

Jelikož má e-shop k dispozici velké množství dat ohledně uskutečněných objednávek, jeví se myšlenka použití data miningových metod jako velmi vhodná.

Pro účely data miningu e-shop dodal potřebná data ve formátu .xlsx, která obsahují několik domén. V tomto kroku je ale potřeba vzít v potaz relevantnost poskytnutých dat, abychom se vyhnuli jednomu z častých dataminingových problémů. Tímto problémem je použití nevhodných nebo nic neříkajících dat. Pro naše účely je tedy logické některé domény zanedbat (např. nás nebude zajímat číslo dokladu nebo číslo objednávky samotné). Zajímat nás budou následující domény:

- Cena s DPH
- Čas vytvoření objednávky
- Datum vytvoření objednávky
- Dobírka ANO/NE
- Dodací adresa
- Množství
- Název produktu
- Poštovní směrovací číslo
- Velikost
- Značka
- Způsob dopravy
- Způsob platby

6 Výběr vhodného softwaru pro data mining

V dnešní době je již k dispozici poměrně velké množství nástrojů pro dataminingové účely. Je tedy otázkou, jaký nástroj vybrat? V procesu rozhodování byly vzaty v úvahu 3 programy. Těmi jsou: Excel, Statgraphics a MATLAB. Jako nejjednodušší možnost se z počátku jevilo použití Excelu, resp. jeho doplňku Data Mining Add-in. Dále bylo potřeba vzít v potaz, že disponujeme plnými verzemi programů Statgraphics a MATLAB díky školní licenci.

Od kvalitního nástroje pro data mining očekáváme kromě dalšího i odpovídající uživatelskou podporu. Případně dostupná internetová fóra, kde je podobná problematika řešena více uživateli. Dále pak, výstup modelování by měl být snadno pochopitelný a použitelný v reálném světě. To vše by mělo bez nutnosti převádět data do nějakého specifického prostředí.

Předpokládá se, že právě v marketingových aplikacích nemá moc význam urputně lpět na přesnosti vyvinutého modelu. Mnohokrát je lepší nějaký robustní a lépe srozumitelnější model.

Jak dále uvádí Karpíšková, 2010, pokud se bavíme o "lepším modelu", je potřeba si ujasnit, co tím vlastně myslíme? V případě řešení praktických situací nemusí být vhodné mít "technicky dokonalý" model. Důležitým kritériem pro hodnocení analytického prostředí softwaru jsou např. požadavky na míru automatizace při jeho využívání. Zde je potřeba vzít v potaz odbornost osoby, která bude program ovládat. Určitý program může být vhodnější pro méně zkušeného nebo méně snaživého uživatele, jiný typ programu může být vhodnější pro kvalitního modeláře, který má navíc větší množstvím času, který tomu může věnovat.

Např. naučit se s MATLABem včetně jeho pokročilejších funkcí určitě není otázka několika dní. Oproti tomu, Excel, který používáte víceméně už od základní školy je oproti tomuto ve výhodě. Co se týče MATLABu, na internetu je k dispozici dostatek materiálů, včetně YouTube návodů a diskuzních fór, kde se řeší nejrůznější problematika, a tak není těžké se dobrat odpovědi, případně si podobný, již probíraný problém včetně jeho řešení upravit pro svoje potřeby.

Dále Karpíšková, 2010 uvádí, že při výběru dodavatele data miningových ale i jiných programů je vhodné zkusit odhadnout, jaké jsou perspektivy vzhledem k podpoře programu a jeho vývoji včetně např. oprav chyb. Uznává ale, že není zcela jasné, jestli je vhodnější produkt velkého komerčního dodavatele, nebo úspěšný a hojně využívaný open source nástroj.

Pokud bychom měli celou tuto problematiku shrnout, tak při hodnocení, případně výběru softwarového produktu je vhodné co nejpřesněji definovat požadavky a zvážit technické, personální a finanční možnosti prostředí, ve kterém by řešení mělo fungovat (Karpíšková, 2010).

Z open source nástrojů připadá v úvahu také R. Jedná se o prostředí pro výpočty a statistickou analýzu, případně tvorbu kvalitních grafických výstupů. Je to příjemný vysokoúrovňový programovací jazyk, ale k industrializovanému data miningu se R podle autorů tolik nehodí. Karpíšková, 2010 dále uvádí, že z komerčních nástrojů pro data mining je zajímavou volbou KXEN a jeho InfiniteInsight, který je po akvizici nyní pod taktovkou německé softwarové firmy SAP. Autorka dále uvádí, že ve větších společnostech se běžně pracuje s více analytickými programy najednou.

Byť se z počátku jevilo nelogické používat v této práci více nástrojů společně, v průběhu tvorby vyšlo najevo, že kombinace Excelu a MATLABu není úplně špatná volba. Bylo tomu tak hlavně z toho důvodu, že určité procedury je prostě jednodušší provést v Excelu a na něco už Excel po výpočetní a interpretační stránce nestačí, a proto musel přijít na řadu MATLAB.

6.1 Excel

Ne každý ví, že Excel v kombinaci s dataminingovým doplňkem se stává solidním nástrojem pro data mining. Jedná se o velmi užitečný a intuitivní nástroj, který by měla být bez problému schopna obsluhovat po krátkém zaškolení i laická veřejnost.

Poté, co uživatel nainstaluje doplněk, vytvoří v Excelu tabulku a nechá ji analyzovat. Excel si ji vezme a pošle do Analysis Services. Zde se data zpracují, přičemž do Excelu se pošlou výsledky. Ty jsou pak v Excelu přehledně prezentovány. Nevýhodou tedy je, že doplněk

nemůže fungovat bez Analysis Services. Jelikož jsou Analysis Services součást MS SQL Serveru, musíte ho mít dostupný. Konkrétně v některé z dražších verzí, která Analysis Services obsahuje.

Fakt, že jde o „doplněk“, by nemělo uživatele vést názoru, že se jedná pouze o nějaké drobné vylepšení nebo malou funkci navíc. Jde o zásadní věc, která posouvá Excel na vyšší úroveň. Je otázkou, zda lze Excel v kombinaci s doplňkem označit za plnohodnotný nástroj pro DM. Faktem je, že i přes jeho rozšíření a nesporné výhody je mnohými odborníky považován za nedostatečný a nevhodný pro akademické, či výzkumné účely. Právě toto je jeden z důvodů, proč Excel nebude v této práci použit.

Pro data, která je potřeba v této práci zpracovat by Excel s doplňkem bohatě stačil. Grafické zpracování dat by možná nebylo tak reprezentativní a v určitých sekcích by nebylo jít možné až tak do hloubky, ale to stejně není cílem této práce. Další překážkou je neschopnost Microsoftu poskytnout tento doplněk pro Office 2016 a vyšší. I přestože existují nějaké způsoby, jak toto obejít přes editor registru, bude lepší poohlédnout se po nástroji, kde se uživatel nesetká s problémy už při instalaci produktu.

V Excelu bylo původně zamýšleno data pro procedury v MATLABu připravit. Problém ale nastal např. v případě, když bylo potřeba odebrat duplicitní hodnoty a vymazat prázdné řádky. Soubor obsahuje +- 140 tisíc údajů a vzhledem k tomu, že PC na kterém byly analýzy prováděny je po hardwarové stránce dobře vybaven, bylo celkem překvapující, že Excel neustále zamrzal a nakonec bylo potřeba pro určité typy příprav použít MATLAB.

6.2 Statgraphics

Dalším vhodným nástrojem pro data mining je program Statgraphics. Tento program v jednoduchém jazyce provádí a vysvětluje základní i vysoce pokročilé statistické funkce. Jedná se o užitečný nástroj pro podnikatelskou komunitu. Jednu dobu byl dokonce nejprodávanějším analytickým programem na světě. Statgraphics je velmi oblíbený software používaný profesionály, výzkumnými pracovníky, akademiky a průmyslovými podniky.

Oproti Excelu s doplňkem nahrává použití tohoto programu i fakt, že je k dispozici zdarma jeho plná verze díky školní licenci. V porovnání s dalším a nakonec i zvoleným programem Statgraphics ale není designově až tak daleko.

Dále je potřeba vzít v potaz uživatelskou podporu a informace dostupné na nejrůznějších diskuzních fórech a portálech. A v tomto případě Statgraphics musí ještě hodně dohánět. Dále pak, možná je to subjektivní pocit autora, ale grafické uživatelské rozhraní, celkové ovládání programu a následná prezentace výsledků je minimálně o stupínek níže, než nakonec zvolený program. Tím je MATLAB.

6.3 MATLAB

Jedná se o interaktivní programové prostředí a taky skriptovací programovací jazyk. MATLAB umožňuje počítání s maticemi, vykreslování 2D a 3D grafů funkcí, implementaci algoritmů, počítačovou simulaci, analýzu a prezentaci dat nebo třeba vytváření aplikací včetně uživatelského rozhraní. Tento program má uživatele především z řad vědeckotechnických pracovníků, studentů a zaměstnanců vysokých škol. MATLAB je využíván pro vědecké, ale i výzkumné účely a to jak v soukromém sektoru, tak i v akademických řadách. Mezi hlavní oblastí využití patří technické obory a ekonomie.

Název MATLAB je vlastně zkratka slov MATrix LABoratory („maticová laboratoř“), což odpovídá faktu, že hlavní datovou strukturou při výpočtech v MATLABu jsou matice.

Výběru tohoto programu nahrává několik faktů. Tím prvním je, že díky školní licenci je opět možné používat jeho plnou verzi včetně všech jeho toolboxů[4]. Dále pak jeho grafické rozhraní je uživatelsky velmi příjemné, oproti Statgraphicsu je minimálně o třídu výše. A byť MATLAB klade na uživatele nepochybně vyšší nároky než Statgraphics, práce s ním byla pro autora jednodušší, jelikož MATLAB použil již dříve pro účely své bakalářské práce.

[4] Toolbox – jedná se o knihovny funkcí/balíky vědomostí z nejrůznějších oborů (např. Financial Toolbox, který obsahuje funkce pro matematické modelování a statistickou analýzu finančních dat)

Co mělo ale opravdu zásadní vliv na použití MATLABu pro tuto práci a co musí i autor vyzdvihnout, je uživatelská podpora a dostupnost informací na internetu. Diskuzí fóra na internetu jsou plná nepřeberných informací řešící problémy jednotlivých uživatelů. Například problematika regulárních výrazů a klouzavých průměrů je na internetu řešena v diskuzních fórech hned několikrát. Což je ale nejspíše dáno celosvětovou rozšířeností MATLABu.

7 Data pro analýzu a jejich příprava

E-shop pro který je analýza prováděna dodal kompletní data za roky 2014-2019. Data byla dodána ve formátu .xlsx, přičemž data bylo nutné před zpracováním ještě upravit, jako např. zanedbání duplicitních čísel objednávky v případě odběru více kusů zboží. Nebo třeba zavedení jednoznačné identifikace názvu produktu, aby MATLAB uměl data vhodně interpretovat. Pro program může být totiž velmi náročné srozumitelně graficky znázornit data následujícího formátu:

Peněženka typ A Peněženka typ B Peněženka typ C Peněženka typ D

Za předpokladu, že e-shop nabízí opravdu široké spektrum sortimentu a krom výše zmíněných peněženek ještě např. kalhoty, trička, mikiny, boxerky, trenky, brýle, ponožky, opasky atd… a každá z těchto domén má ještě svůj typ, tak je pak velmi náročné data graficky rozumně interpretovat. Proto, pokud budeme na data nahlížet jako na celek, pak pro lepší srozumitelnost zavedeme označení takto: Peněženka = {typ A; typ B; typ C; typ D}

MATLAB tedy zanedbá konkrétní typy peněženek a v případě, že při vyhledávání najde konkrétní typ peněženky, označí ji prostě jako „Peněženka“. Abychom tohoto v MATLABu docílili, je třeba hodnoty domény Název produktu upravit pomocí použití regulárních výrazů. Pomocí regexu pak dostáváme pouze základní informace o typu produktu jako: peněženka, boty, triko atd.

Dále je potřeba data očistit o duplicitní hodnoty, protože by jinak došlo ke zkreslení výsledných hodnot. Např. pokud zákazník koupil jednu mikinu, dvoje tepláky a čepici, pak tomu v tabulce dodaných dat odpovídá 4x platba bankovním převodem a stejně tak 4x zvolený způsob dopravy Českou poštou, tím pádem by došlo k nepřesným výpočtům, a proto je třeba tyto hodnoty očistit, tak aby pro každé identifikační číslo objednávky byl přiřazen pouze 1x způsob zvolené dopravy a 1x způsob platby.

Jak už bylo řečeno výše, data je potřeba před jednotlivými procedurami vhodně upravit. Otázka je, zda pro úpravu dat použít Excel, kde to může být o něco jednodušší než za pomocí MATLABu. Problém nastal v času dokončení procedury, kdy Excel vyžadoval několik hodin

a nakonec program ještě „zamrznul“. Přitom výpočty probíhaly na hardwarově poměrně solidně vybaveném PC. Pro představu, v počáteční fázi, před jakýmikoliv úpravami, bylo potřeba zpracovat 141 386 řádků o 22 sloupcích. Některé úpravy proto musely být provedeny přímo v MATLABu.

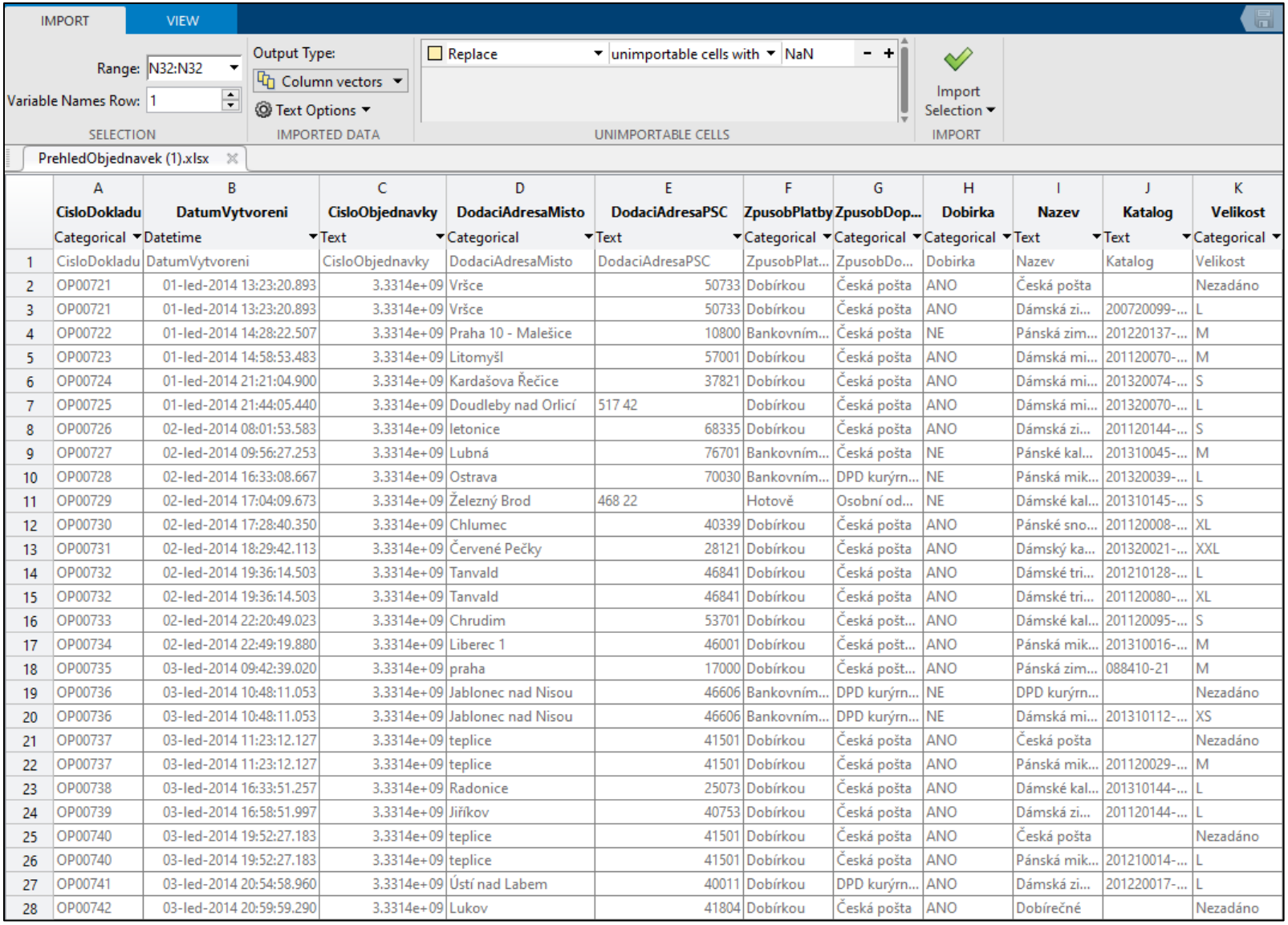
Dodaná data ještě před úpravou znázorňuje Obr. 8 na další straně. Za povšimnutí stojí např. smíšená data ve sloupci *I*, kde máme název produktu, ale vyskytuje se nám zde i způsob dopravy Česká pošta a DPD kurýrní služba. Dále ve sloupci *B* můžeme vidět *Datum vytvoření objednávky*. Zde se ale vyskytuje i čas který navíc obsahuje hodiny, minuty, sekundy a milisekundy. Časové údaje včetně data vytvoření objednávky bude určitě potřeba nějakým způsobem zploštit, ideálně klouzavým průměrem. Jak uvádí Tab. 3 níže, dalším problémem, se kterým bylo možné se v průběhu přípravy dat setkat, byl například rozdílný zápis názvu u obce Albrechtice v Jizerských horách.

Tab. 3: Nesprávný zápis názvu obce Albrechtice

Albrechtice v j.h
Albrechtice v j.h.
Albrechtice v jiz. horách
Albrechtice v jiz horach
Albrechtice v jizerských horách
Albrechtice v jizerských horách

Zdroj: Vlastní zpracování

Vzhledem k tomu, že data byla dodána s poškozenými PSČ, bylo nutné identifikaci provádět přes názvy měst a obcí, což proceduru identifikace značně zkomplikovalo. Toto je jen několik málo věcí z mnoha, které bude potřeba ošetřit.



	A	B	C	D	E	F	G	H	I	J	K
	CisloDokladu	**DatumVytvoreni**	**CisloObjednavky**	**DodaciAdresaMisto**	**DodaciAdresaPSC**	**ZpusobPlatby**	**ZpusobDop...**	**Dobirka**	**Nazev**	**Katalog**	**Velikost**
	Categorical	Datetime	Text	Categorical	Text	Categorical	Categorical	Categorical	Text	Text	Categorical
1	CisloDokladu	DatumVytvoreni	CisloObjednavky	DodaciAdresaMisto	DodaciAdresaPSC	ZpusobPlat...	ZpusobDo...	Dobirka	Nazev	Katalog	Velikost
2	OP00721	01-led-2014 13:23:20.893	3.3314e+09	Vršce	50733	Dobírkou	Česká pošta	ANO	Česká pošta		Nezadáno
3	OP00721	01-led-2014 13:23:20.893	3.3314e+09	Vršce	50733	Dobírkou	Česká pošta	ANO	Dámská zi...	200720099-...	L
4	OP00722	01-led-2014 14:28:22.507	3.3314e+09	Praha 10 - Malešice	10800	Bankovním...	Česká pošta	NE	Pánská zim...	201220137-...	M
5	OP00723	01-led-2014 14:58:53.483	3.3314e+09	Litomyšl	57001	Dobírkou	Česká pošta	ANO	Dámská mi...	201120070-...	M
6	OP00724	01-led-2014 21:21:04.900	3.3314e+09	Kardašova Řečice	37821	Dobírkou	Česká pošta	ANO	Dámská mi...	201320074-...	S
7	OP00725	01-led-2014 21:44:05.440	3.3314e+09	Doudleby nad Orlicí	517 42	Dobírkou	Česká pošta	ANO	Dámská mi...	201320070-...	L
8	OP00726	02-led-2014 08:01:53.583	3.3314e+09	letonice	68335	Dobírkou	Česká pošta	ANO	Dámská zi...	201120144-...	S
9	OP00727	02-led-2014 09:56:27.253	3.3314e+09	Lubná	76701	Bankovním...	Česká pošta	NE	Pánské kal...	201310045-...	M
10	OP00728	02-led-2014 16:33:08.667	3.3314e+09	Ostrava	70030	Bankovním...	DPD kurýrn...	NE	Pánská mik...	201320039-...	L
11	OP00729	02-led-2014 17:04:09.673	3.3314e+09	Železný Brod	468 22	Hotově	Osobní od...	NE	Dámské kal...	201310145-...	S
12	OP00730	02-led-2014 17:28:40.350	3.3314e+09	Chlumec	40339	Dobírkou	Česká pošta	ANO	Pánské sno...	201120008-...	XL
13	OP00731	02-led-2014 18:29:42.113	3.3314e+09	Červené Pečky	28121	Dobírkou	Česká pošta	ANO	Dámský ka...	201320021-...	XXL
14	OP00732	02-led-2014 19:36:14.503	3.3314e+09	Tanvald	46841	Dobírkou	Česká pošta	ANO	Dámské tri...	201210128-...	L
15	OP00732	02-led-2014 19:36:14.503	3.3314e+09	Tanvald	46841	Dobírkou	Česká pošta	ANO	Dámské tri...	201120080-...	XL
16	OP00733	02-led-2014 22:20:49.023	3.3314e+09	Chrudim	53701	Dobírkou	Česká pošt...	ANO	Dámské kal...	201120095-...	S
17	OP00734	02-led-2014 22:49:19.880	3.3314e+09	Liberec 1	46001	Dobírkou	Česká pošt...	ANO	Pánská mik...	201310016-...	M
18	OP00735	03-led-2014 09:42:39.020	3.3314e+09	praha	17000	Dobírkou	Česká pošt...	ANO	Pánská zim...	088410-21	M
19	OP00736	03-led-2014 10:48:11.053	3.3314e+09	Jablonec nad Nisou	46606	Bankovním...	DPD kurýrn...	NE	DPD kurýrn...		Nezadáno
20	OP00736	03-led-2014 10:48:11.053	3.3314e+09	Jablonec nad Nisou	46606	Bankovním...	DPD kurýrn...	NE	Dámská mi...	201310112-...	XS
21	OP00737	03-led-2014 11:23:12.127	3.3314e+09	teplice	41501	Dobírkou	Česká pošta	ANO	Česká pošta		Nezadáno
22	OP00737	03-led-2014 11:23:12.127	3.3314e+09	teplice	41501	Dobírkou	Česká pošta	ANO	Pánská mik...	201120029-...	M
23	OP00738	03-led-2014 16:33:51.257	3.3314e+09	Radonice	25073	Dobírkou	Česká pošta	ANO	Dámské kal...	201310144-...	L
24	OP00739	03-led-2014 16:58:51.997	3.3314e+09	Jiříkov	40753	Dobírkou	Česká pošta	ANO	Dámská zi...	201120144-...	L
25	OP00740	03-led-2014 19:52:27.183	3.3314e+09	teplice	41501	Dobírkou	Česká pošta	ANO	Česká pošta		Nezadáno
26	OP00740	03-led-2014 19:52:27.183	3.3314e+09	teplice	41501	Dobírkou	Česká pošta	ANO	Pánská mik...	201210014-...	L
27	OP00741	03-led-2014 20:54:58.960	3.3314e+09	Ústí nad Labem	40011	Dobírkou	DPD kurýrn...	ANO	Dámská zi...	201220017-...	L
28	OP00742	03-led-2014 20:59:59.290	3.3314e+09	Lukov	41804	Dobírkou	Česká pošta	ANO	Dobírečné		Nezadáno

Obr. 8:Dodaná data před úpravou po úvodním načtení do MATLABu

Zdroj: Vlastní zpracování

8 Finální data mining

Nyní, když už máme všechna data očištěná a odebrané duplicitní hodnoty tam, kde je to potřeba, můžeme se pustit do procesu samotného data miningu. Zajímat nás budou *dodací adresy*, upravené na města s největším počtem objednávek. Dále pak nejčastěji a zároveň nejméně *prodávaný sortiment*, který jsme pomocí regexu zúžili pouze na názvy kategorií. Zajímat nás bude také typ *prodané značky*, přičemž, na základě požadavku konzultanta nebudou uvedeny značky úzce spjaté s e-shopem. Konzultant se domnívá, že po zveřejnění této DP by jejich prozrazení mohlo poskytnout konkurentům cenná data. Jedná se o dvě značky, které budou v grafu nahrazeny výrazy *značka1* a *značka2*. Co se týče základních analýz, bude zde analyzován ještě *způsob přepravy*, kde se budeme věnovat jednotlivým přepravcům.

Co se týče hlubších analýz, pak zde bude graficky znázorněna závislost počtu objednávek na času, přičemž zde budeme moct porovnat vývoj bez klouzavého průměru nebo s ním. Dále se v práci budeme zabývat analýzou počet realizovaných nákupů podle cenových intervalů.

Na tomto místě je také potřeba upozornit, že grafy jsou kolikrát velmi detailní a je velmi obtížné je čitelně znázornit na formátu A4, tak, aby byly vidět podstatné detaily včetně např. legendy. A právě z tohoto důvodu se v určitých případech nebudou vyskytovat popisy jednotlivých os grafů.

8.1 Dodací adresa

Co se týče dodací adresy, i když v případě data miningu nestanovujeme žádné počáteční hypotézy, lze předpokládat, že na základě počtu obyvatel bude mezi města s největším počtem objednávek patřit Praha, Brno, Ostrava, Plzeň, Olomouc, případně Liberec.

V případe grafického znázornění měst s nejčastější dodací adresou je potřeba vzít v potaz fakt, že se jedná o relativně velké množství měst, a proto je potřeba použít vhodný graf. Standardní grafy jsou vzhledem k množství měst nevhodné. Použít by šly pouze za předpokladu, že bychom počet měst redukovali, nebo nějakým způsobem sloučily do krajů

nebo okresů. Jenže nám jde v tomto případě o detailnější data, a proto je vhodné použít graf, který se nazývá Word Cloud, resp. „oblak slov“. Tento graf je vhodný pro znázornění většího množství dat, přičemž největším písmem jsou znázorněny ta slova, která se vyskytují v datech nejčastěji.

Co se týče měst, jak už jsme si řekli, dá se předpokládat, že největší zastoupení budou mít města s největším počtem obyvatel. Obr. 9 tuto domněnku potvrzuje. Zároveň se zde však vyskytují i města, která bychom zde nečekali v takové intenzitě oproti ostatním. Za zmínku stojí např. Jablonec nad Nisou, který se svými 45 773 obyvateli má poměrně vysoké umístění. Dalším nečekaným objevem je město Třemošná, které s 5 082 obyvateli má poměrně vysoký počet objednávek (180). Co je ale opravdu nečekané a překvapivé i pro konzultanta práce je velmi vysoký počet objednávek u obce Lubná, která má pouze 947 obyvatel ale objednávek bylo realizováno 900.

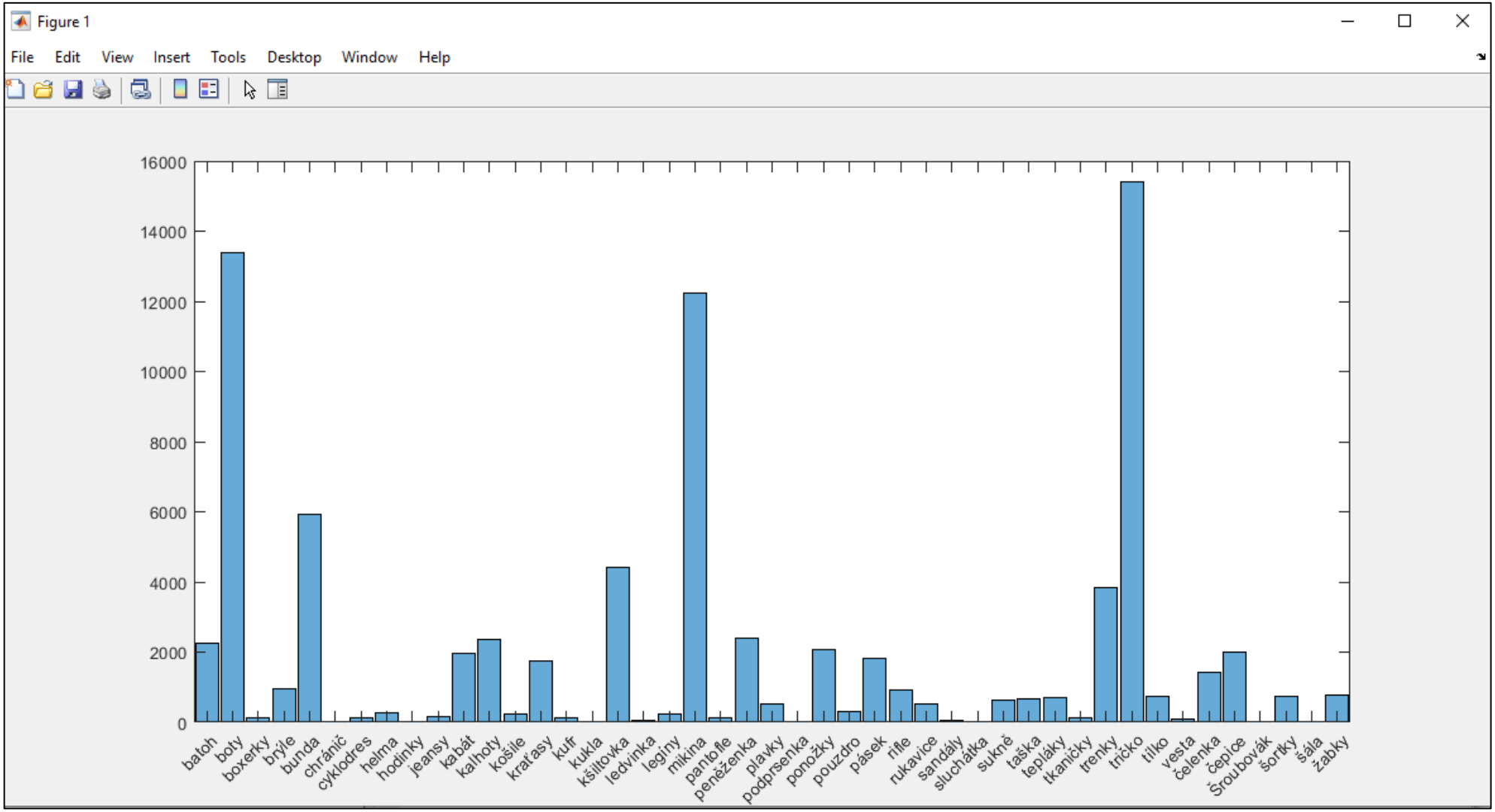
Obr. 9: Znázornění významnosti měst v závislosti na počtu provedených objednávek
Zdroj: Vlastní zpracování

8.2 Nejčastěji prodávaný sortiment

Co se týče nejčastěji prodávaného sortimentu, jak je uvedeno na Obr. 10, dle výsledků z MATLABu se jedná o *trička* na prvním místě, *mikiny* na druhém a hned v závěsu jsou *boty*. Což není překvapující zjištění, jelikož se jedná o běžně nakupovaný sortiment. Co ale bylo už méně očekávané, jsou počty prodaných kusů, kterých dosáhly kšiltovky. Zde se prodané kusy pohybují okolo 5000 kusů, což je poměrně nečekané a původně bylo i ze strany konzultanta práce očekáváno číslo menší.

Co se týče sortimentu, jehož prodané množství se v grafu pohybuje u spodní hranice a není skoro ani vidět (*chrániče, hodinky, kufr, ledvinka, podprsenka, sluchátka a šroubovák*), tak zde je potřeba vzít v potaz fakt, že náš e-shop se primárně nezabývá prodejem ochranných pomůcek jako jsou chrániče, případně se ani nejedná o e-shop s potřebami pro kutily, a proto se dalo i očekávat, že tento typ sortimentu nebude dosahovat vysokých hodnot prodejů.

Největším překvapením z oblasti prodávaného sortimentu je ale poměr prodejů mezi položkami *boxerky* a *trenky*. Studie prokazují, že za několik posledních let se trend vyvíjí tak, že jsou boxerky upřednostňovány většinou spotřebitelů před volnějšími trenkami. My máme k dispozici data z let 2014-2019, které pro náš e-shop vypovídají naopak. Prodeje trenek se mezi roky 2014-2019 pohybují kolem 4000, kdežto u boxerek se jedná o řádově o několik málo stovek. To je ale možná dáno tím, že se jedná o e-shop se sportovněji zaměřeným sortimentem a typický zákazník u našeho e-shopu může mít odlišné preference, co se týče oblékání např. oproti typickému zákazníkovi e-shopů, jako jsou ABOUT YOU nebo třeba ZALANDO, které jsou nepochybně veřejnosti známější a mají i širší okruh zákazníků.



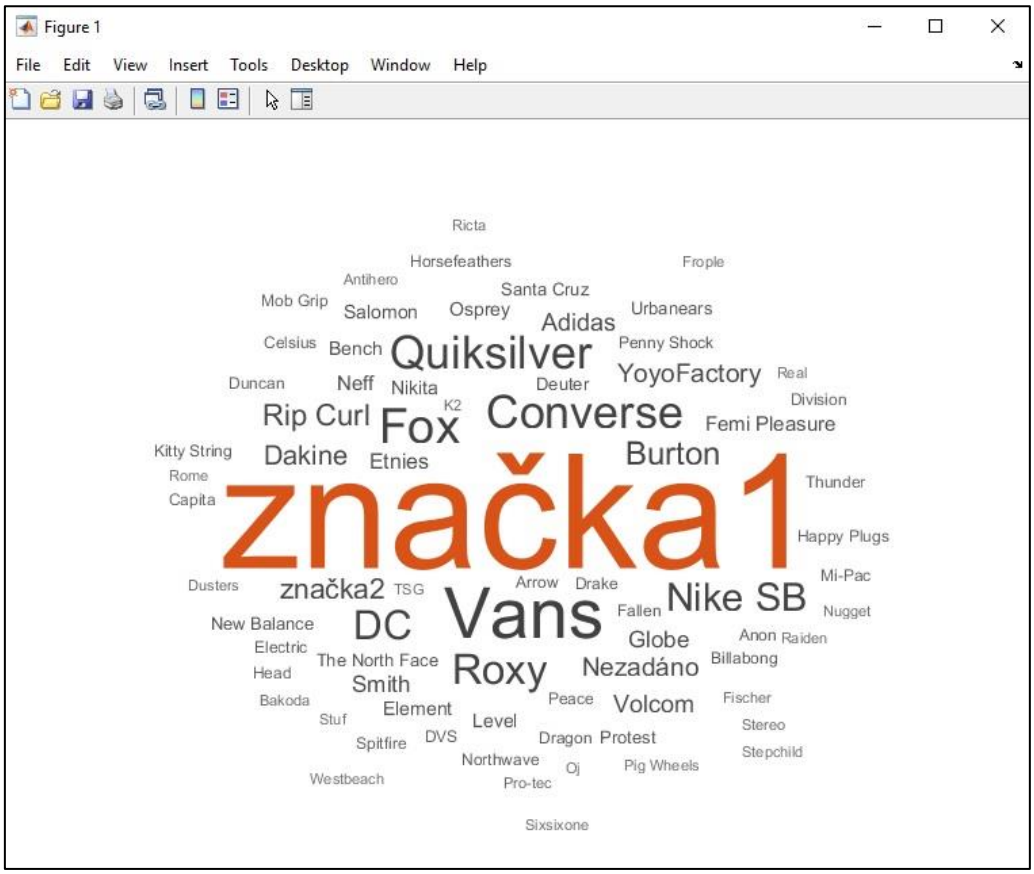
Obr. 10: Graf nejčastěji prodávaného sortimentu po úpravě pomocí regulárních výrazů

Zdroj: Vlastní zpracování

8.3 Typ prodávané značky

V případě grafického znázornění frekvence prodávaných značek bylo potřeba vzít v potaz prosbu konzultanta a zároveň i jednu z hlavních podmínek poskytnutí dat pro tuto práci. Tou podmínkou je anonymizace určitých typů značek z důvodu ochrany/konkurenčního ohrožení e-shopu. Jedná o se dvě značky, které mají velmi úzkou spojitost s e-shopem. Tou první značkou je privátní značka e-shopu, kterou budeme v grafu označovat jako „značka1“. Druhou značku mající také úzkou spojitost s e-shopem z důvodu jejího vlastnictví budeme označovat jako „značka2“.

Pro grafické znázornění typů prodávaných značek opět nemáme z důvodu množství lepšího grafu než je Word Cloud. Jak vyplývá z Obr.11 níže, největší zastoupení má právě privátní značka e-shopu, což byl i původní předpoklad konzultanta. Největší množství produktů na e-shopu nese právě tuto značku, a proto se dalo předpokládat takové výsledky. Další značka s e-shopem spjatá je co se týče prodejů spíše upozaděna, což je dáno tím, že i množství produktů nesoucí její název není nijak markantní.



Obr. 11: Typy nejčastěji prodávaných značek

Zdroj: Vlastní zpracování

8.4 Způsob přepravy

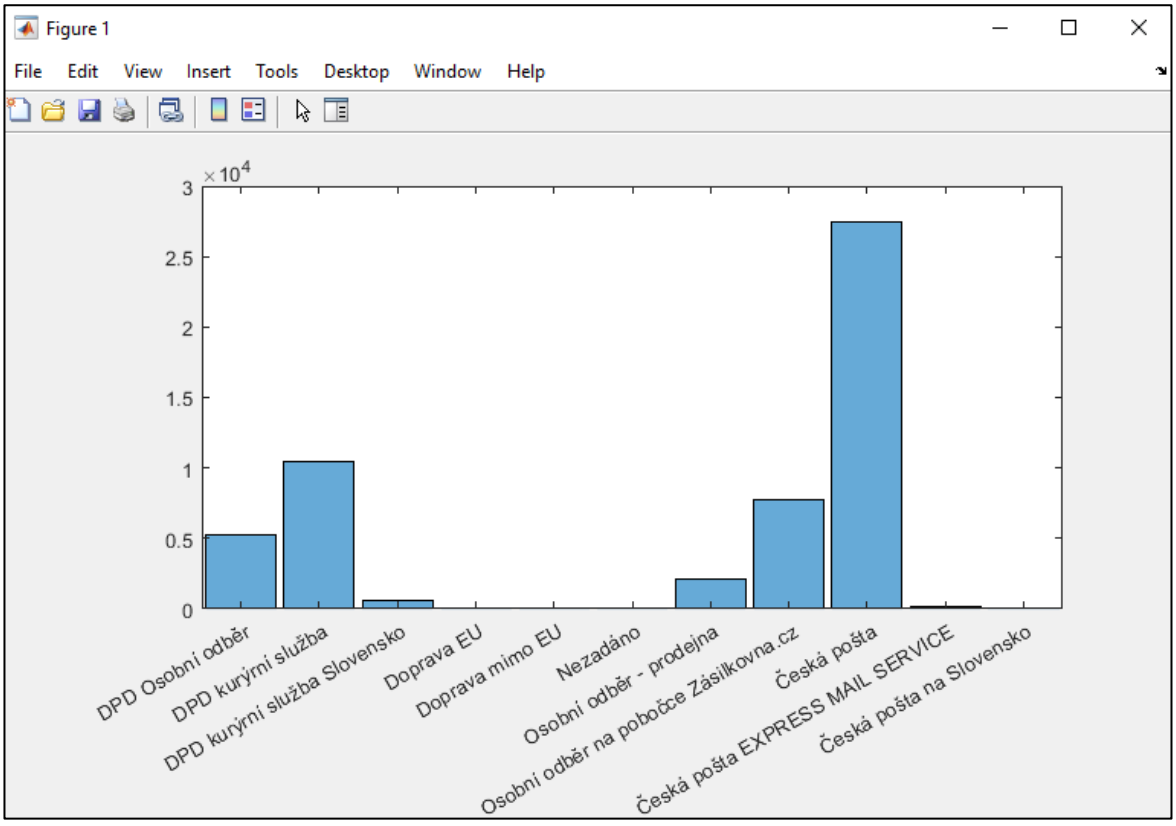
Další analyzovanou částí byl způsob zvolené dopravy, přičemž zákazníci měli na výběr z mnoha variant. Kromě České pošty nabízí e-shop ještě přepravu přes přepravní společnost DPD, nově možnost osobního odběru přes Zásilkovnu, případně osobní odběr na prodejně. Další možnosti, kam můžeme zařadit např. dopravu mimo EU a do EU nebo speciální služby České pošty nás moc nezajímají. Jak je uvedeno na Obr. 12 na další straně, dalo se očekávat, že nejvyšší podíl bude zabírat Česká pošta. Byť se jedná o zkostnatělou instituci, která se snaží udržet krok s moderní dobou, faktem je, že na přepravní společnosti typu DHL, PPL a UPS už přestává stačit a není to dáno pouze nízkou ochotou zaměstnanců nebo špatným managementem. Jedná se o celkový přístup organizace. Sice se jedná o subjektivní pocit, ale bez pomoci státu by bylo obtížné pro tuto instituci udržet krok s konkurenty.

To, že je Česká pošta na špici ve způsobu zvolené přepravy, je dáno především tím, že se jedná o veřejně známou instituci. Na druhém místě se umístila kurýrní služba DPD, která je v ČR poměrně rozšířená, ale i přesto, pokud má určitá sorta lidí na výběr (typicky starší občané), pak díky její dlouhodobé tradici stejně upřednostní Českou poštu. Co se ale týče pohledu mladší generace, pokud nenastane nějaká razantní změna v přístupu této organizace, bude to mít do budoucna velmi těžké.

Dalším zjištěním, a ne překvapivým je poměrně vysoký počet realizovaných doručení pomocí Zásilkovny. Jedná se o ryze českou přepravní společnost, která funguje už od roku 2010 a její popularita neustále roste. Z vlastní zkušenosti může autor říci, že se jedná o velmi dobře fungující způsob přepravy. Stát se v dnešní době výdejním místem Zásilkovny je poměrně jednoduchý proces. Není proto divu, že tyto místa přibývají velmi rychlým tempem. Co se týče procesu vrácení zboží, tak přepravní společnosti typu DHL či PPL jsou oproti české poště o krok dále i co se týče plynulosti procesu. Např. v případě vrácení zboží u e- shopu ABOUT YOU prostě stačí odnést balíček zpět na výdejní místo a pouze sdělit své telefonní číslo, kam vám poté bude zaslána SMS s potvrzením převzetí a to je vše. U české pošty je potřeba mít speciální formulář suplující podací lístek, bez kterého se neobejdete a který když ztratíte, tak musíte pátrat na webu a dožadovat se nového.

Co se týče osobního odběru na prodejně, jedná se o situaci, která je pochopitelná. Zákazník si objedná na prodejnu, která se nachází nejblíže jeho bydlišti zboží, které je dostupné třeba jen v prodejně na druhé straně republiky. To si přímo na prodejně vyzkouší a když nebude spokojen, dostane peníze za provedenou platbu okamžitě zpět.

V dnešní době, když přihlédneme ještě navíc ke koronavirové krizi, kde je většina prodejen uzavřena, situace taková, že je běžnou praxí si objednat více kusů na jednou. Vznikají tak situace, kdy si např. zákazník není jistý velikostí, tak objedná velikost Medium i Large, to si doma vyzkouší a nepadnoucí oblečení prostě zašle zpět, nebo ho vrátí přímo na prodejně, pokud obchod tuto variantu nabízí. V případě výše zmíněné situace s objednáním všeho ve dvou velikostech objednávky jednotlivce kolikrát dosahují řádově tisíců korun. A v případě, že bylo oblečení pouze vyzkoušeno a stále obsahuje cenovky, nemají e-shopy sebemenší problém reklamace uznat.



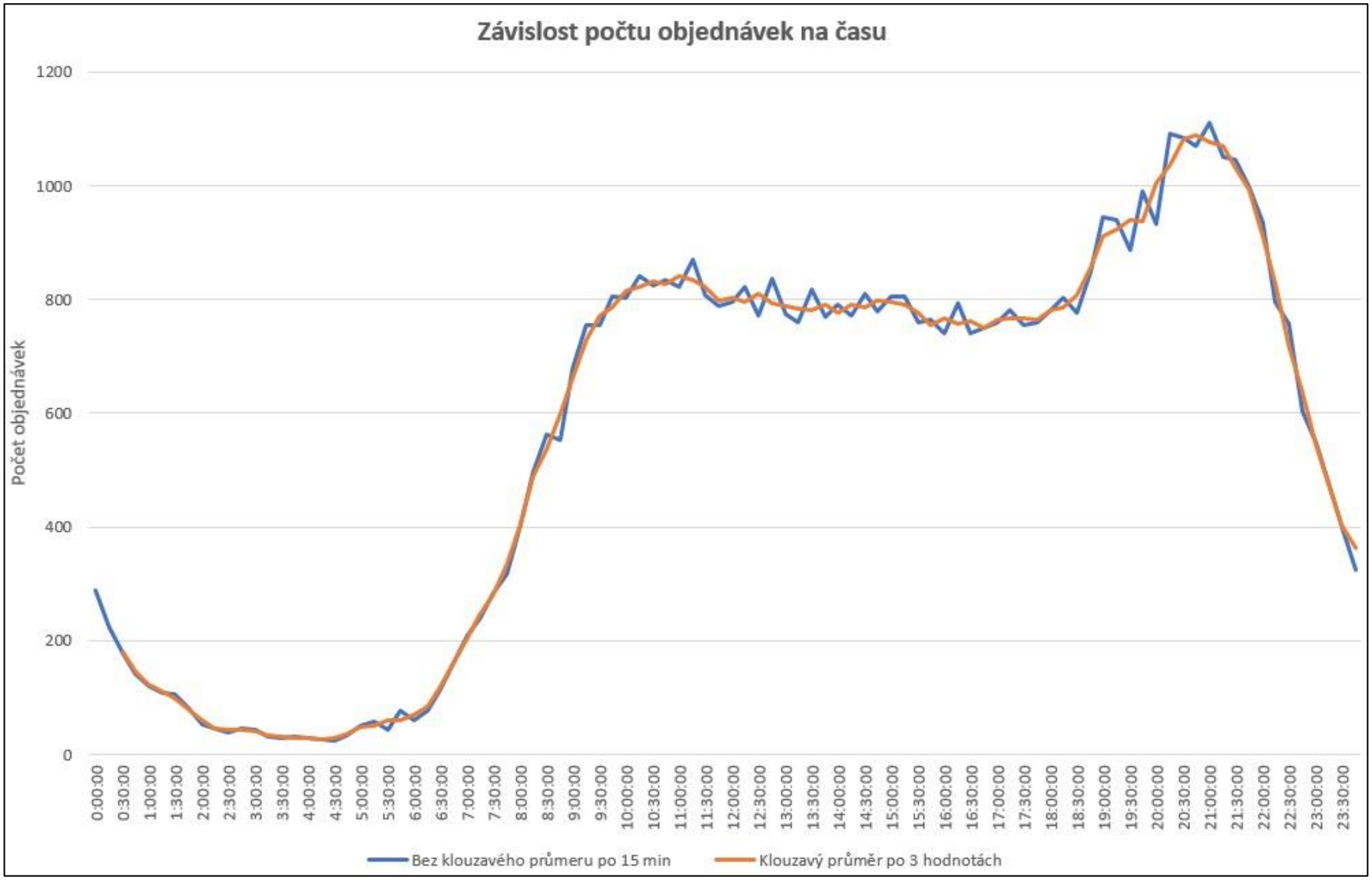
Obr. 12: Graf zvoleného způsobu přepravy

Zdroj: Vlastní zpracování

8.5 Závislost počtu objednávek na času

Co se týče počtu objednávek, chtěli jsme zjistit, jak se počet objednávek vyvíjí v času během dne pro souhrn veškerých realizovaných objednávek. Na Obr. 13 na další stránce máme graf, kde na ose *X* máme časové intervaly rozdělené po 30 minutách. Na ose *Y* máme počet objednávek. Jelikož objednávky byly realizovány v nejrůznějších časech během dne a my jsme chtěli odhalit trend, bylo potřeba použít klouzavý průměr k vyhlazení časových údajů. Graf obsahuje jak „ostřejší“ data, která jsou měřena po 15 minutách bez klouzavého průměru (modrá křivka) a pak pro srovnání i plošší data, na která byl použit klouzavý průměr po třech hodnotách (oranžová křivka).

Graf reprezentující nákupní chování zákazníků na e-shopu začíná od půlnoci, kde se počet objednávek pohybuje okolo 300. Dále následuje pozvolný propad na zhruba 50 objednávek až do cca 4:30, a poté začíná frekvence nákupů prudce růst až do 10:00. Zde se pohybuje na hranici 800 objednávek s odchylkou přibližně 50. V tomto pásmu se až na menší výkyvy ustálí. Toto ustálení vydrží až do cca 18:30, od kdy opět následuje prudký růst, který dosahuje denního maxima přibližně ve 21 hodin, kde se dostáváme na nějakých 1 100 objednávek. Poté následuje prudký pokles až na +- 370 objednávek ve 23:30. Toto nákupní chování potvrzuje všeobecný trend, že nejčastěji lidé nakupují ve večerních hodinách, kdy už jsou pravděpodobně doma z práce a v pohodlí domova. Z grafu lze dále vyčíst, že mezi 10:00-18:30 nedochází k žádným větším výkyvům a nákupní chování je v tomto intervalu ustálené, což může být dáno faktem, že touto dobou je většina lidí v práci a nákup provede, až dorazí domů.



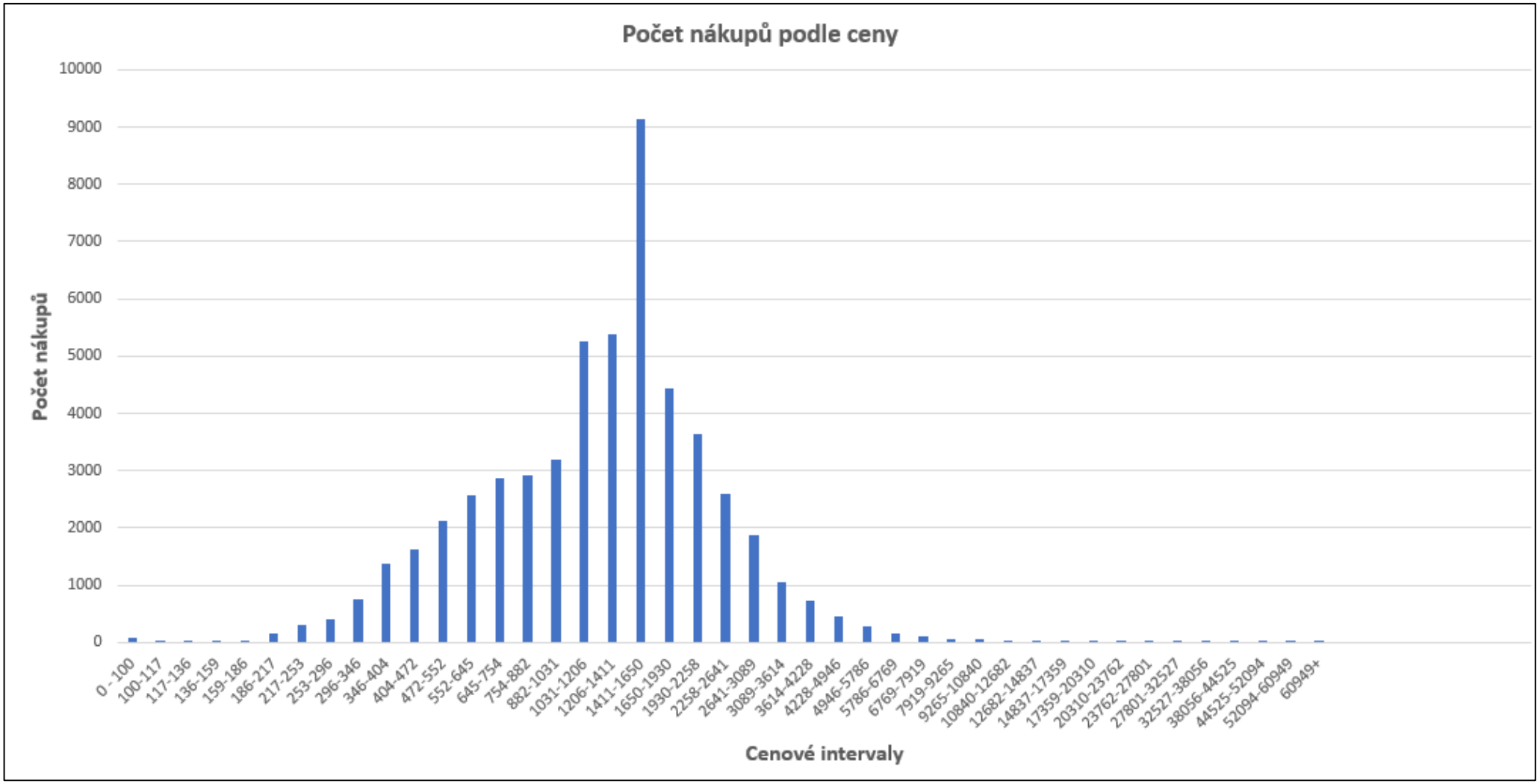
Obr. 13: Graf znázorňující závislost počtu objednávek na času

Zdroj: Vlastní zpracování

8.6 Počet nákupů

Dále jsme zkoumali počet realizovaných nákupů dle ceny, přičemž pro grafickou interpretaci byl použit histogram. Vzhledem k tomu, že ceny byly opět různorodé, bylo potřeba je uspořádat do intervalů. Z Obr. 14 vyplývá, že největší počet realizovaných nákupů (něco málo nad 9000) se pohybuje v intervalu 1411-1650 Kč.

Kolem čísla 5 500 objednávek jsou nákupy pohybující se v intervalu 1131-1411 Kč. Za povšimnutí stojí také interval 0-100 Kč, kdy jsou lidé ochotni provést nákup i za předpokladu, že za dopravu zaplatí minimálně zhruba polovinu částky, kterou dají za zboží.



Obr. 14: Znázornění počtu nákupů podle cenových intervalů

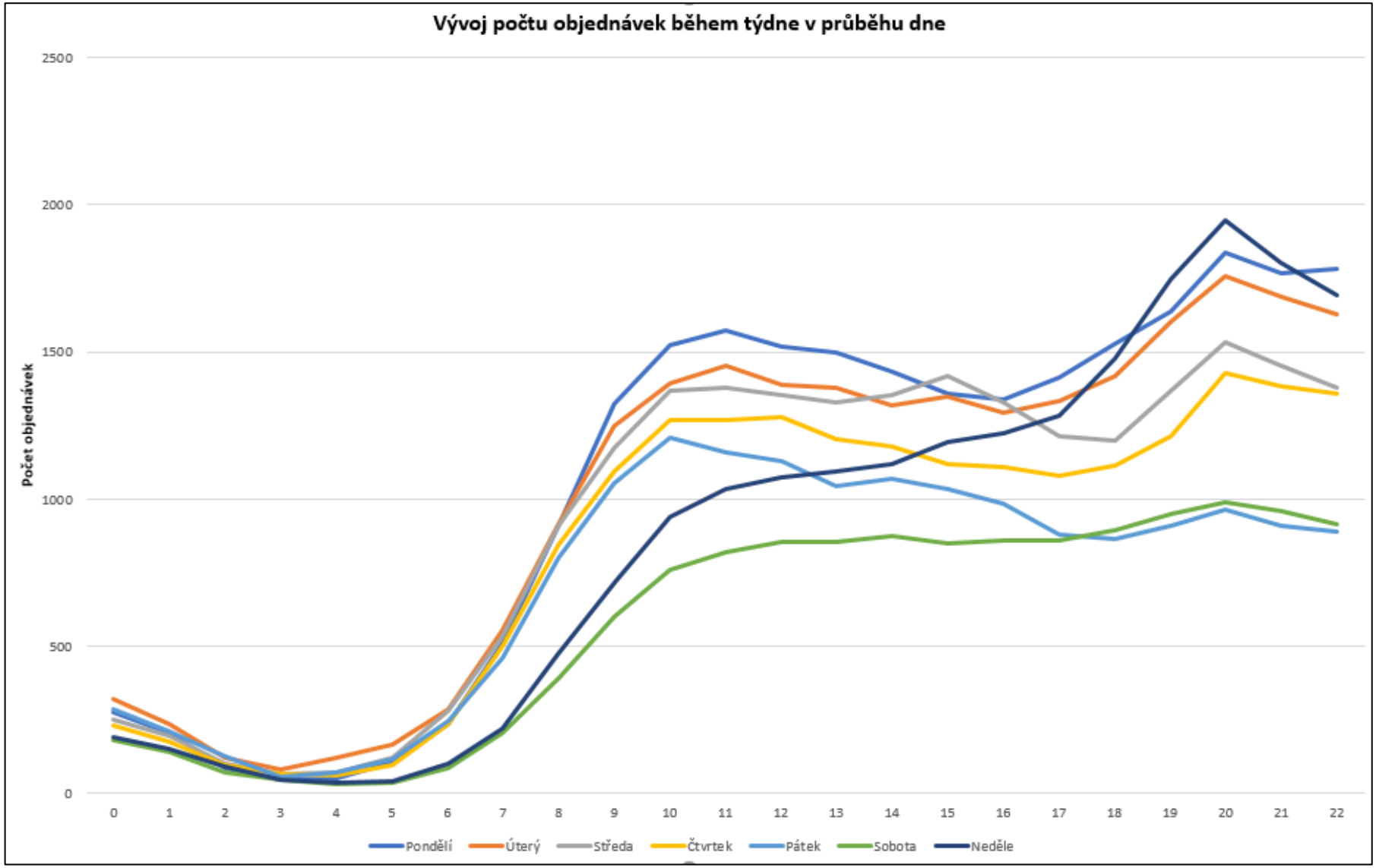
Zdroj: Vlastní zpracování

8.7 Vývoj počtu objednávek během týdne v průběhu dne

Dále nás zajímalo, jak se mění nákupní chování zákazníků v průběhu týdne. Zda např. nakupují více o víkendu, nebo všední dny. Aby byl graf dobře čitelný, opět bylo potřeba využít klouzavého průměru pro vyhlazení časových hodnot, stejně tak z důvodu čitelnosti legendy zde není uveden popisek osy *X*, kde se nachází časové údaje. Osa Y pak reprezentuje průměrný počet objednávek. Jak je uvedeno na Obr. 15, nejnižší průměrný počet objednávek ze všech dní je realizován v sobotu, kdy nepřesáhne ani hranici 1000 objednávek a vrcholu dosahuje přibližně ve 20 hodin.

Zajímavý průběh objednávek v času má neděle, která sice dosahuje nejvyššího počtu objednávek skoro 2000 ve 20 hodin, ale oproti všedním dnům spolu se sobotou má pomalejší nástup (viz sobota a neděle a interval 04-10h). Předpokládá se, že to je dáno především faktem, že o víkendu lidé všeobecně vstávají později než ve všední dny, a proto je právě o víkendu počáteční růst v ranních hodinách pozvolnější. Z grafu dále vyplívá, že vrcholu dosahují všechny dny přibližně ve stejnou dobu, a to kolem 20. hodiny, přičemž poté následuje pro všední dny strmý pokles, kromě pondělí, kde můžeme vidět po 21 hodině mírnější nárůst. O víkendu není pokles po 20. hodině tak dramatický.

Co se týče dopoledních hodin, tak nejvyššího průměrného počtu objednávek dosahuje pondělí a za ním překvapivě následují ostatní všední dny přesně tak, jak jdou za sebou v kalendáři. Zde stojí za zmínku středa, která přibližně kolem oproti ostatním menšího nárůstu.



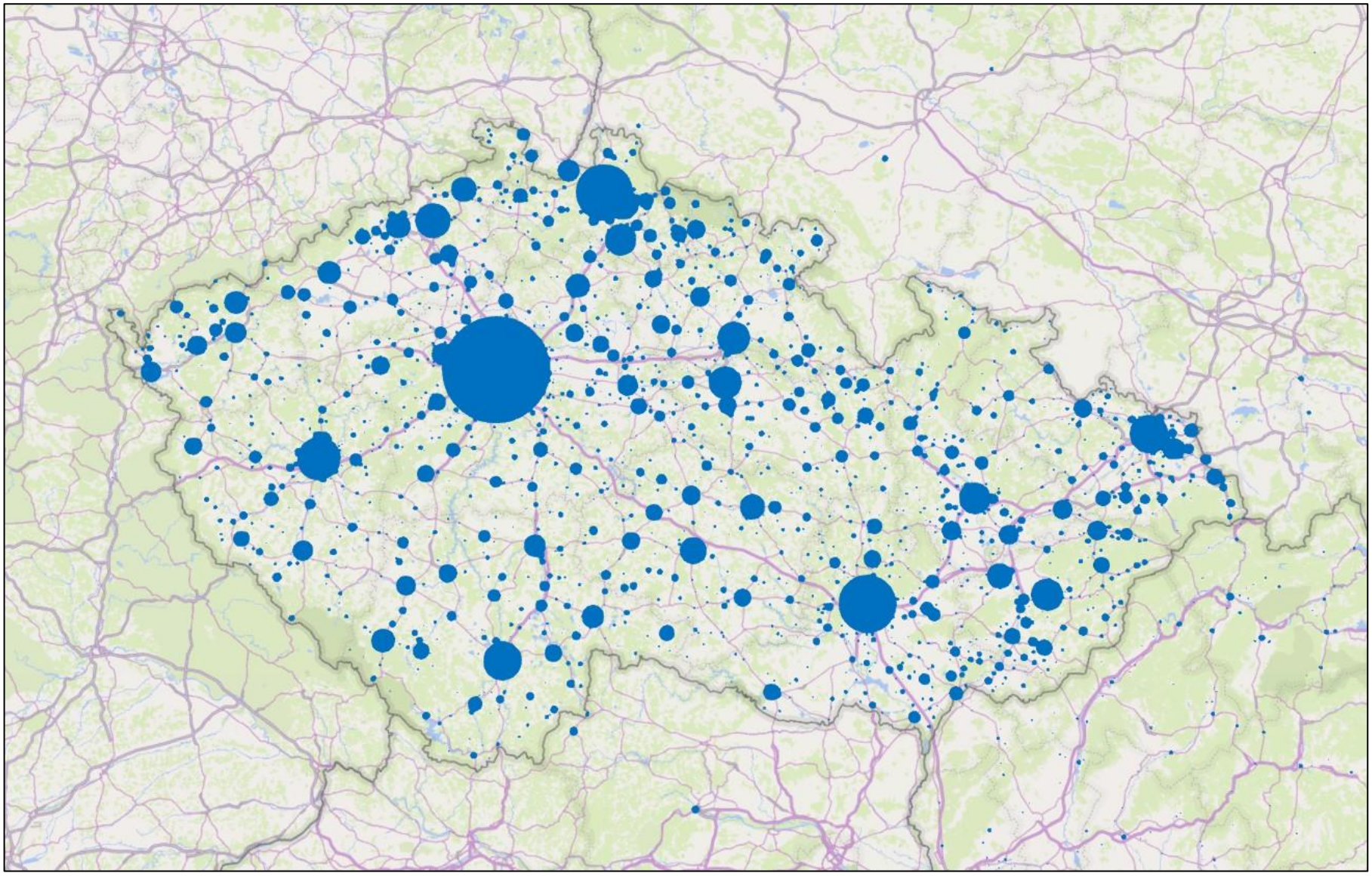
Obr. 15: Vývoj počtu objednávek během týdne v průběhu dne

Zdroj: Vlastní zpracování

8.8 Celkový počet objednávek na město

Obr.16 na následující straně, suma počtu objednávek na město za roky 2014 – 2019, zobrazuje místa v republice, která jsou nejvíce vytížena objednávkami. Jedná se o absolutní počet, nebereme v potaz cenu. Této statistice dominuje Praha přibližně s počtem 6 000 objednávek v tomto období. V nejvýše dosažených pozicích v tomto seznamu si můžeme všimnout vysoké intenzity ve větších městech, konkrétně se jedná o Plzeň, Olomouc, Ostrava, České Budějovice, Pardubice, Hradec Králové nebo Zlín. Zde se počet objednávek pohybuje v rozmezí od 300 – 800, což je pochopitelné, jelikož ve větších městech je silnější sorta obyvatelstva a více nakupují specifické potřeby a oblečení přes internet. Konkrétně značkové oblečení, které je vyhledávané náročnějším zákazníkem s pevnou vazbou na daný produkt nebo na konkrétní značku. Tito lidé vědí, co očekávat od produktu a mají plnou důvěru v objednávání sortimentu z e-shopu.

Největším vychýlením z tohoto trendu velkých měst je obec Lubná, která již byla zmíněna v předchozím textu, a ve zkoumaném období bylo realizováno 906 objednávek sortimentu, přičemž populace zde dosahuje pouhých 947 obyvatel. Je tedy otázkou, zda má tato obec v okrese Svitavy nějakou návaznost na určitou aktivitu nebo subjekt, který objednává věci za účelem autorovi neznámým. Toto může být předmět dalšího zjišťování informací v nadcházející marketingové strategii a zjištění bližších podrobností s návazností k tomuto místu. Lze si všimnout, že pokud jde o pominutí těchto vysokých hodnot, objednávky jsou rozprostřeny téměř rovnoměrně po celém území České republiky, větší koncentrace je v okolí měst. Naopak nejméně lze zaznamenat aktivitu v Jeseníkách a na jihu Vysočiny, kde nejsou soustředěny významné metropole České republiky, avšak tyto oblasti nelze opomíjet i při dalším postupu v expanzi pro e-shop.



Obr. 16: Celkový počet objednávek na město

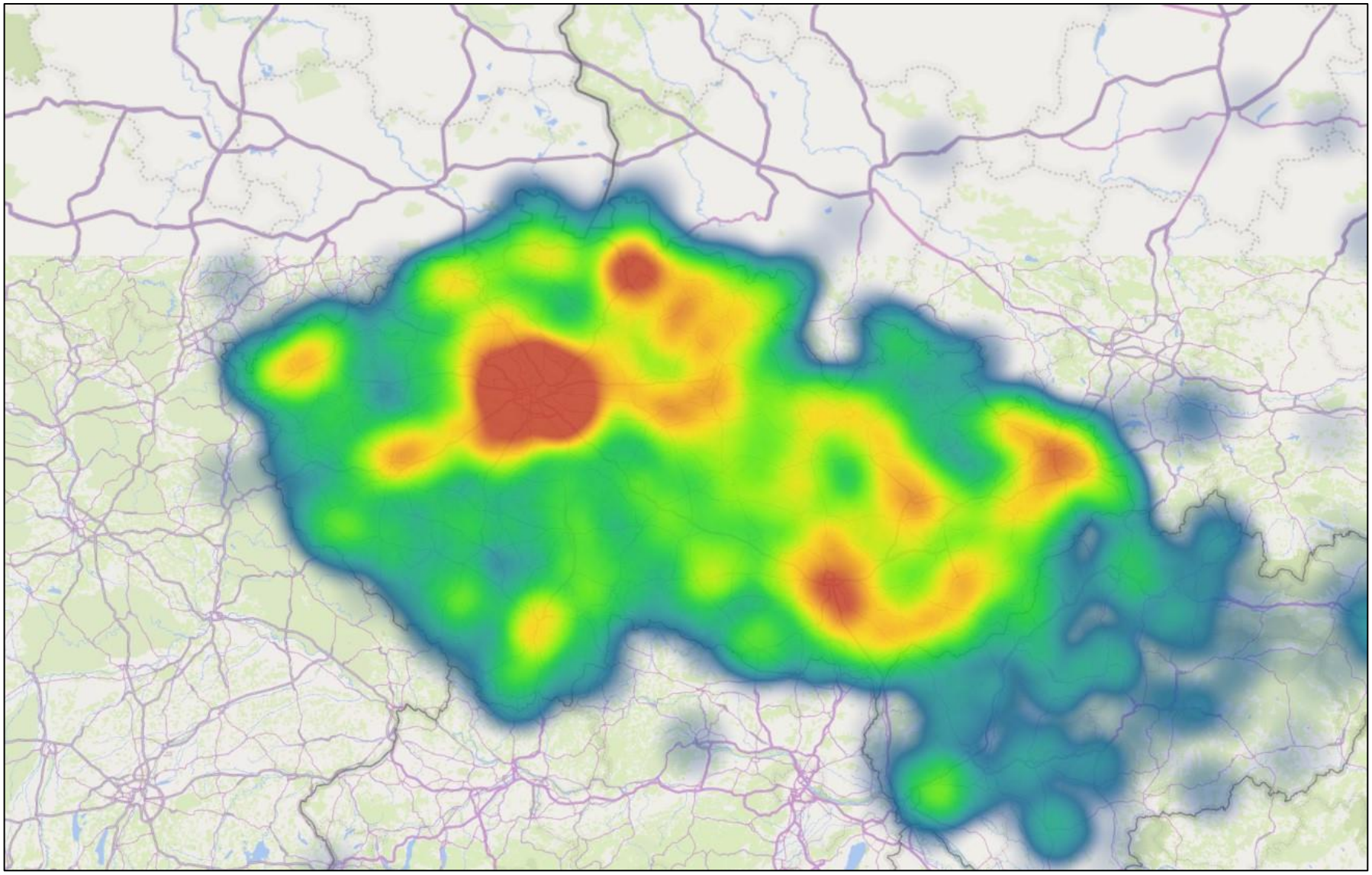
Zdroj: Vlastní zpracování

8.9 Průměrná cena objednávky na oblast

Na Obr. 17 na další stránce máme pomocí heat mapy znázorněnou průměrnou cenu objednávky na oblast za roky 2014 – 2019. I v tomto případě si lze opět povšimnout vysoké cenové hladiny v okolí Prahy. Tato lokalita je tedy dominantní nejenom v počtu objednávek za sledované období, ale i průměrnou cenou na jednu objednávku. Cena je zvýrazněna tudíž i intenzitou, která je zdůrazněna faktorem počtu. Průměrná cena jedné objednávky v Praze se pohybuje kolem 1 600 Kč. Tento trend v poměru vyšší průměrné ceny a vysokého počtu objednávek dodržují již pouze další 3 velká města, a to Liberec, Brno a Ostrava, jak je na mapě zřejmé i podle zvýrazněné intenzity. Pokud však budeme brát nejvyšší průměrnou cenu na menší počet objednávek, vyloučíme pro tyto případy lokality pouze s jednou objednávku.

Zjistíme, že lokality, které realizovali více než 50 objednávek – Kyjov, Třinec, Litomyšl, Říčany a Jablonné nad Orlicí se pohybovaly od 4 000 do 8 000 Kč. Lze tedy usoudit, že tento trend disponuje určitým parametrem a záměrem konkrétních subjektů, které vyžadují sortiment ve větším měřítku než pro osobní spotřebu včetně např. B2B obchody. V obrázku lze vysledovat, že určité procento objednávek směřuje i na zahraniční obchod. Největšími odběrateli jsou zákazníci na hranicích se Slovenskem, kde se však útraty nepohybují ve významných částkách nebo objemech, avšak je potřeba s tím faktem počítat a brát v potaz pro případnou expanzi firmy.

Pouze minimální odběry jsou zaznamenávány v okolních zemích (Polsko, Německo, Rakousko), kde se objednávky z e-shopu pohybují pouze v jednotkách kusů a tudíž v aktuální situaci nejsou pro tento výzkum až tak významné, avšak tento případ bude vyvrácen na dalším obrázku, který se bude věnovat evropskému formát prodejů. Pokud se zaměří autor na nejmenší odběry z e-shopu, jedná se pouze o jednotky kusů, maximálně do 200 Kč na jednu objednávku. Z tohoto lze konstatovat, že se jedná pouze o „nahodilé“ nákupy s cíleným motivem pro koupi jedné konkrétní věci.



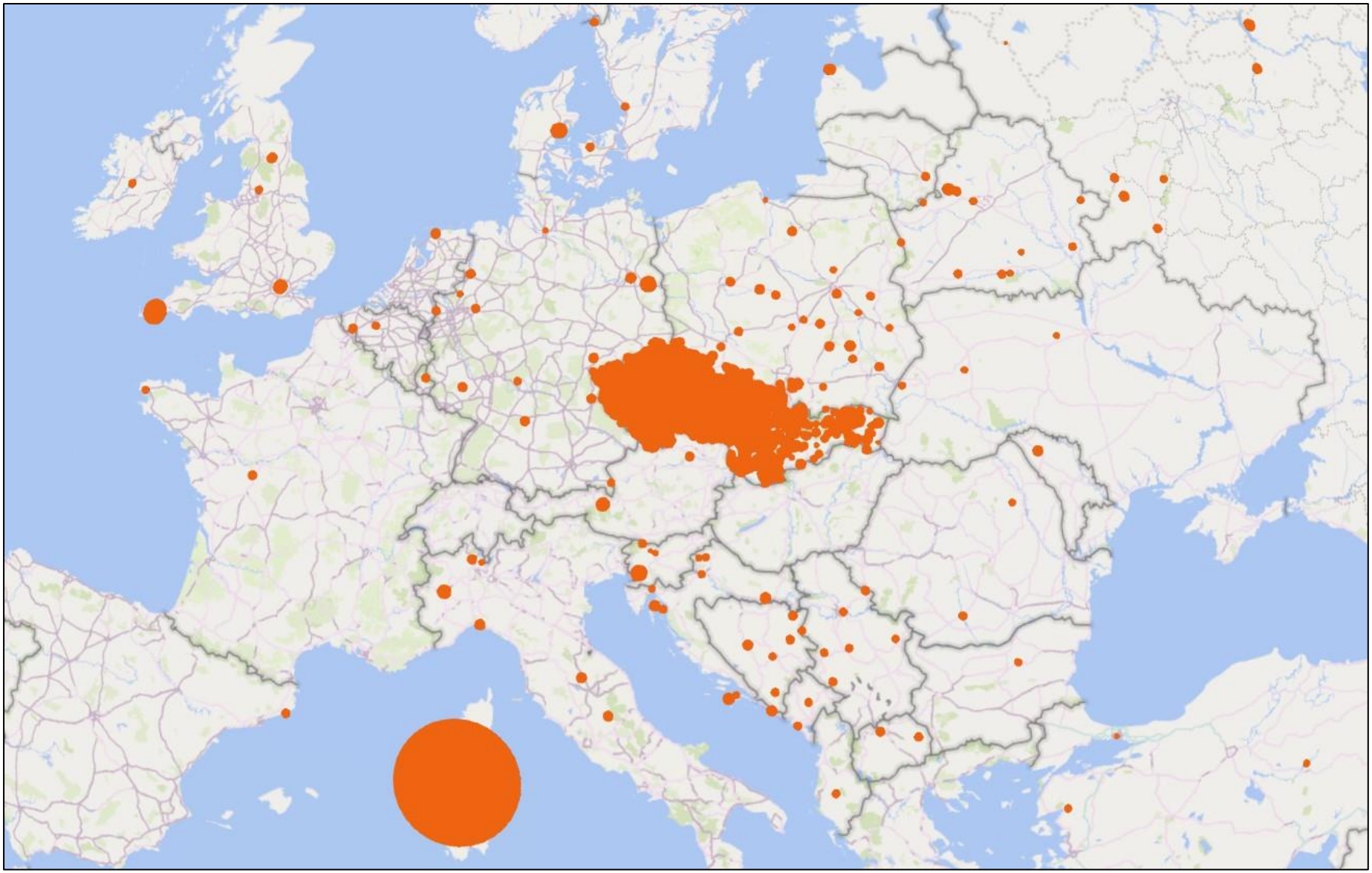
Obr. 17: Průměrná cena objednávky na oblast

Zdroj: Vlastní zpracování

8.10 Celková cena objednávky pro evropský trh

Na posledním Obr. 18 na další stránce je zobrazena celková cena objednávek pro evropský trh. Jsou zde již zobrazeny veškeré evropské lokality, které e-shop v minulosti kontaktoval s obchodní nabídkou a skutečně byl obchod realizován. Na první pohled je důležité zmínit, že koncentrace na Střední Evropu je zřejmá, jelikož se jedná o český e-shop, je jisté, že především nejvíce objednávek bude z České republiky, potažmo ze Slovenska (i kvůli komunikaci). Slovensko jako takové bylo na předchozím grafu průměrné ceny objednávky zvýrazněno pouze částečně, zde již vidíme skutečný rozsah aktivity, které můžeme vyhodnotit přes celé území Slovenska.

Menšími lokalitami po celé Evropě lze vytipovat jednotlivé nákupy bez opakovaného podnětu, zde se jedná opět o jednotky artiklů, které jsou pravděpodobně v daný okamžik dostupné nebo jsou velmi limitované v omezeném množství. Největší zmínkou na mapě je však místo zvané Sassari, ve kterém byla realizována nejvyšší průměrná cena objednávky, a to konkrétně necelých 280 000 Kč za jediný artikl, což tuto destinaci řadí na první místo v této statistice. Dalšími významnými lokacemi jsou například britský Cornwall s objednávkami za 40 000 Kč nebo Ask v Norsku a Fürstenwalde v Německu (oboje shodně po 5 000 Kč).



Obr. 18: Celková cena objednávky pro evropský trh

Zdroj: Vlastní zpracování

9 Souhrn poznatků

Cílem této práce bylo analyzovat data za roky 2014-2019, které budou poté prezentovány zástupci e-shopu a bude možné je využít pro další marketingové účely. V této práci jsme zkoumali dodací adresy, kde jsme předpokládali, že na základě počtu obyvatel přední příčky obsadí města jako Praha, Brno, Ostrava, Plzeň, Olomouc, případně Liberec. To se nakonec i potvrdilo. Praha byla s 6073 objednávkami a průměrnou cenou objednávky 1612,94 Kč největším městem. Následoval Liberec (1741; 1472,85), Brno (1698; 146,56), Plzeň (964; 1483, 96), ale pak nastalo obrovské překvapení, když na další příčce se objevila obec Lubná nacházející se v Pardubickém kraji, která má pouze +- 950 obyvatel. Obec Lubná realizovala ale 906 objednávek s průměrnou cenou 1833,93 Kč, a tak v žebříčku předběhla i Ostravu (772; 1312,31), České Budějovice (753; 1591,95) a Ústí nad Labem (659; 1448,42) nebo Hradec Králové (567; 1354,21). Dále pak nás vzhledem k počtu obyvatel, případně počtu provedených objednávek zaujal Třinec (137; 7917,03) a Říčany (83; 8797,84).

Pokud se bavíme o extrémních nebo nečekaných hodnotách, pak za zmínku stojí i obchody realizované mimo ČR. Sem můžeme zařadit italské město Sassari ležící v oblasti Sardinie, které realizovalo pouhou jednu objednávku, ale o hodnotě 279 443,45 Kč a dále pak Cornwall (3; 13 070,53).

Poté jsem se zabývali nejčastěji prodávaným sortimentem, kde nebylo žádným překvapením, že první příčky obsadily trička, mikiny a boty. Jediné, co bylo trochu překvapující, bylo umístění trenek před boxerkami, což neodpovídá současnému trendu, kdy jsou většinou boxerky upřednostňovány.

Co se týče typu prodávané značky, zde jsme museli anonymizovat značky navázané na e-shop. Proto jsme použili označení *značka1* a *značka2*. První příčku dle očekávání obsadila privátní značka e-shopu *značka1*. Poté následovaly značky Vans, Quiksilver, Fox, Converse, Nike, Roxy a další.

Dále jsme řešili také způsob přepravy. Zde byla dominantní Česká pošta, byť kvalitou poskytovaných služeb třeba nedosahuje na konkurenty. Výsledek je dán pravděpodobně všeobecnou rozšířeností a známostí České pošty. Za poštou se umístili DPD kurýrní služba a poslední dobou i velmi oblíbená Zásilkovna.

Dále jsme přešli do hlubší analýzy a zkoumali jsme závislost počtu objednávek na času. Zde jsme zjistili, že frekvence nákupů prudce roste cca od 6:00 až do 10:00, poté se ustálí až do cca 18:30, od kdy opět následuje prudký růst, který dosahuje denního maxima přibližně ve 21 hodin. Poté následuje prudký pokles. Jak již bylo řečeno toto nákupní chování potvrzuje všeobecný trend, že nejčastěji lidé nakupují ve večerních hodinách, kdy už jsou pravděpodobně doma z práce a v pohodlí domova.

Poté jsme se zaobírali počtem nákupů dle ceny, které jsme uspořádali do intervalů. Zde jsme zjistili, že největší počet realizovaných nákupů (něco málo nad 9000) se pohybuje v intervalu 1411-1650 Kč.

Další velmi zajímavou statistiku nám poskytnul graf zabývající se vývojem počtu objednávek během týdne v průběhu dne. Zjistili jsme, že neděle, která sice dosahuje nejvyššího počtu objednávek (skoro 2000 ve 20 hodin) má oproti všedním dnům spolu se sobotou pomalejší nástup. Předpokládáme, že to je dáno především faktem, že o víkendu lidé všeobecně vstávají později než ve všední dny, a proto je právě o víkendu počáteční růst v ranních hodinách pozvolnější. Z grafu dále vyplívá, že vrcholu dosahují všechny dny přibližně ve stejnou dobu, a to kolem 20. hodiny, přičemž poté následuje pro všední dny strmý pokles, kromě pondělí, které má po 21 hodině mírnější nárůst.

Celkový počet objednávek na město nepřekvapivě vyhrává Praha s cca 6 000 objednávek, jak již bylo zmíněno výše. Není překvapivá ani vysoké intenzita ve větších městech jako jsou Plzeň, Olomouc, Ostrava, České Budějovice, Pardubice, Hradec Králové nebo Zlín. Zde se počet objednávek pohybuje v rozmezí od 300 – 800, což je pochopitelné, jelikož ve větších městech je silnější sorta obyvatelstva a více nakupují specifické potřeby a oblečení přes internet.

Další statistiku – průměrnou cenu objednávky na oblast jsme zobrazili pomocí heat mapy. I v tomto případě jsme si mohli povšimnout vysoké cenové hladiny v okolí Prahy. Tuto lokalitu jsme tedy označili jako dominantní nejenom v počtu objednávek za sledované období, ale i průměrnou cenou na jednu objednávku. Průměrná cena jedné objednávky v Praze se pohybuje kolem 1 600 Kč. Stejný trend dodržovala pouze další 3 velká města, a to Liberec, Brno a Ostrava.

Nakonec jsme se zabývali celkovou cenou objednávek pro evropský trh. Vzhledem k tomu, že se jedná o český e-shop, nejvíce objednávek bylo z České republiky, potažmo ze Slovenska. Dále se zde vyskytovaly menšími lokality po celé Evropě.

Překvapením bylo italské město Sassari, ve kterém byla realizována nejvyšší průměrná cena objednávky, (cca 280 000 Kč za jedinou objednávku), což tuto destinaci řadí na první místo v této statistice. Dalšími významnými lokacemi jsou například britský Cornwall s objednávkami za 40 000 Kč nebo Ask v Norsku a Fürstenwalde v Německu (oboje shodně po 5 000 Kč).

Tyto oblasti mohou být klíčovým rozvojem samotné firmy, je proto potřeba tyto ojedinělé hodnoty nadále sledovat a přizpůsobovat jim chování na zahraničním trhu, případně získávat důležité kontakty pro případnou expanzi výrobků a zboží.

Pokud se jedná o aktivity samotného e-shopu, dokáže je, jak je patrné z mapy realizovat po celé Evropě i ve větších částkách a pravděpodobně i ve větším objemu jednotlivého zboží. Prioritou zůstává český a slovenský trh, který je již z grafu plně rozvinut a je potřeba ustálit úspěšně veškeré aktivity a prodeje na těchto územích.

Závěr a marketingová doporučení

Pro e-shop, který se zaměřuje především na textilní sortiment je marketingové prostředí velmi flexibilní a je často řízeno externími vlivy – trendy, roční období, slevové akce, preference zákazníka nebo aktuální vývoj potřeb. Pandemie koronaviru celému lidstvu ukázala, jak dokáže měnit uspořádání trhu a preference zákazníků. V této době bude tedy marketingová reakce o poznání složitější, jelikož je tento aspekt podnícen poklesem tržeb, úbytkem zákazníků a zůstatkem starého zboží na skladech (myšleno z minulé sezony). Samotná doporučení se opírají o získaná data, která byla prostřednictvím data miningu zjištěna.

Sortiment společnosti je velmi široký, převládá dominance v prodejích vlastní značky. Dle vypozorovaných dat, že kvantitativně dosáhly prodeje číslovky více než 5000 kusů jednotlivých položek, což považuje sám konzultant za velmi příjemné překvapení. Důležitou stránkou je však složení prodejů, které mají mezi sebou vysoké výkyvy. Na jedné straně jsou položky, které jsou ve vysokých číslech – boty, trička, mikiny, kraťasy nebo kšiltovky. Na druhém protipólu jsou však produkty, které při výzkumu dosahovaly minimálních hodnot. V tomto případě je změna struktury sortimentu na e–shopu žádoucí. Marketingově se jedná o ztrátové položky, které vyžadují pozornost vlastníka, zda již v sortimentu nenabízet, prodat je za sníženou cenu či vytvářet speciální kombinované nabídky, které zahrnou i tyto produkty. Doporučení je tedy koncentrace soustředění na stávající profitabilní produkty, vyřazení produktů neatraktivních pro zákazníka a nahradit produkty, které splňují trendy, preference a dokáží uspokojit stávající poptávku. K tomuto úkonu slouží průzkum trhu pomocí jednoduchého dotazníku skrze e-shopu a sbírání podnětů směrem od zákazníků.

Důležitost vlastních značek, které jsou v textu označeny jako „značka1“ a „značka2“ je pro oslovenou firmu klíčovou aktivitou. Oblíbenost první značky vystupuje především v prodejnosti e-shopu a lze si tak uvědomit, že se jedná o způsob „lovebrandu“, kteří klienti upřednostňují před známějšími. Péče o značku je nedílnou součástí celého fungování, proto je důležité inovovat prvky, především cílit reklamou a rozšiřovat v oblasti sortimentu tyto atributy. Tvořit speciální akce, přesvědčit zákazníka, že značka je výjimečná a osobitá prostřednictvím předvádění produktů a správné interpretace výhod jednotlivých produktů. Známé značky (Roxy, Converse nebo Burton) v tomto případě těžko měnit, zákazník je

přesvědčen o celosvětové image těchto produktů a sází na ověřenou kvalitu a cenu, zná velmi dobře tyto produkty. Zkoumaná „značka2“ je slabším místem v prodejích, jak je naznačeno ve zkoumaném grafu, zde se nabízí otázka, zda je sortiment pro zákazníka dostatečně atraktivní. Opatrnějšími prostředky, jako jsou speciální nabídky, zdůraznění kvality materiálu a kvality výrobku nebo zaměření se na zážitek se značkou u zákazníka jsou prvky strategie značky, které lze realizovat v tento okamžik. Otázkou je i priorita samotného vlastníka, zda chce mít obě vlastní značky na pomyslném čele prodejů.

Distribuce produktů je převážně realizována společností Česká pošta. Způsob distribuce zahrnuje pro konečného zákazníka hlavní atributy – rychlost, kvalita a cena. V tomto okamžiku je potřeba si uvědomit, zda není možné vyjednat lukrativnější podmínky s přepravními společnostmi, které mohou uzavírat dohodu a redukci ceny při distribuci určitého počtu zakázek. Lze taktéž nabízet zákazníkům výhodnější ceny například prostřednictvím víkendových akcí, slevových kuponů či výhod, plynoucích z předchozích nákupů. Upřednostnění nového distribučního způsobu je podstatné pro doplňkovou kvalitu služeb a pro plynulý chod celé objednávky.

Faktor počtu objednávek na času je klíčový z mnoha hledisek pro průběh a zkoumání celé objednávky, je důležité vědět, odkud zákazník přišel na webové stránky. Zda se jednalo o prokliky ze sociálních sítí, šel přímo na stránky či vyhledal cenu například ze srovnávače cen. K zjištění těchto údajů by bylo vhodné využít data propojená z Google analytics, která nám sdělí procentní podíly prokliků z PPC reklam nebo návštěvnosti jiných sítí. Díky těmto atributům majitel zjistí, zda je potřeba vytvářet cílenou reklamu na určité oblasti, zda zvyšovat počty objednávek v méně oblíbené časy nebo naopak podporovat stabilitu silných prodejních čas – zde je otázka odolání systému objednávek v jeden okamžik. Podpora reklam a sociálních sítí musí sloužit k vhodné prezentaci a především k přesvědčení o nákupu konkrétního produktu. Tato návaznost do sebe zahrnuje i průběh každého prodejního týdne, který byl zjištěn klouzavým průměrem již v analýze.

Poslední zkoumaný faktor je demografické rozložení objednávek. Jak již bylo zjištěno, objednávky se soustředí do velkých měst, kde jsou lidé zvyklí nakupovat přes internet různé oblečení a doplňky, zde tento trend by se měl udržovat a podporovat připomínajícími faktory reklamy a upozorňovat na nové a limitované zboží. Důležitá je v tomto případě otázka evropského trhu, ve kterém byly zaznamenány objednávky s daleko hodnotnějším

sortimentem, než tomu bylo v Česku. Pokud se jedná o komunikaci těchto „výkyvů“ mezi běžným průměrem, je zapotřebí si definovat analýzu, která zkoumá podnět a příčiny na území tuzemském i zahraničním. Je tedy otázkou, zda budovat například i B2B, který v řádech statisíců až milionů by mohl být efektivní. K této komunikaci se stanoví obchodní zástupce, který uskuteční analýzu z těchto dat, zjistí důvody, které vedly k těmto objednávkám a tímto způsobem může ustanovit i velkoobjemové obchodní partnery, kteří by mohli odebírat produkty ve větším měřítku. Zde se naskytuje možnost rozšíření služeb a dalších aktivit, které mohou vést i ke sponzoringu v evropském měřítku, jelikož se jedná o specializovaný e-shop zaměřený na užší specializace a uvést se do povědomí v rámci expanze by tak mohlo být klíčové.

Výsledky této práce budou přes konzultanta prezentovány vedení firmy a mohou posloužit jako velmi cenný zdroj pro zahájení nových marketingových aktivit. Autor doufá, že data najdou v podniku uplatnění a firmě přeje mnoho úspěchů do budoucna.

Citace

AGGARWAL, Charu C. 2015. *Data mining: the textbook*. Cham: Springer. ISBN 978-3-319-14141-1.

BERKA, Petr. 2003. *Dobývání znalostí z databází*. Praha: Academia. ISBN 80-200-1062-9. Dostupné také z: http://sorry.vse.cz/~berka/4IZ450

CAPOCCI, Mauro. 2018. *New eugenics, genomics and human big data: A perspective on the marketing and the use of genes in society* [online]. 487-498. Societa editrice il Mulino, Bologna [cit. 2021-03-13]. Dostupné z: doi:10.1409/90644

DEAN, Jared. 2014. Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners [online]. 289: Canada: John Wiley & Sons, Inc., Hoboken, New Jersey [cit. 2020-12-27]. ISBN 978-1-118-92069-5. Dostupné z: https://www.pdfdrive.com/big-data-data-mining-and-machine-learning-e33454197.html

HAN, Jiawei, Micheline KAMBER a Jian PEI. 2012. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier/Morgan Kaufmann. The Morgan Kaufmann series in data management systems. ISBN 978-0-12-381479-1.

KARPÍŠKOVÁ, Nikola. 2010. OPEN SOURCE NÁSTROJE A DATA MINING. *Data-mind.cz* [online]. [cit. 2021-02-20]. Dostupné z: https://www.datamind.cz/cz/blog/Open-source-nastroje-a-data-mining

MITCHELL, ANNA a LARRY DIAMOND. 2018. *China's Surveillance State Should Scare Everyone* [online]. [cit. 2021-02-14]. Dostupné z: https://www.theatlantic.com/international/archive/2018/02/china-surveillance/552203/

PAYNE, Adrian, 2005. *HANDBOOK OF CRM: Achieving Excellence in Customer Management*. Butterworth-Heinemann. ISBN 978-07506-6437-0.

RAUCH, Jan a Milan ŠIMŮNEK. 2014. *Dobývání znalostí z databází, LISp-Miner a GUHA*. Praha: Oeconomica, nakladatelství VŠE. Odborná kniha s vědeckou redakcí. ISBN 978-802-4520-339.

SPEKTOROWSKI, Alberto a Liza IRENI-SABAN. 2016. *Politics of Eugenics: Productionism, Population, and National Welfare*. Routledge. ISBN 9781138676244.
StatSoft: Úvod do data miningu [online], 2014. [cit. 2021-03-28]. Dostupné z: http://www.statsoft.cz/file1/PDF/newsletter/2014_02_26_StatSoft_Uvod_do_data_miningu.pdf

STOLTE, Chris a Diane TANG. 2013. *Multiscale visualization using data cubes* [online] [cit. 2021-04-15]. Dostupné z: doi:10.1109/TVCG.2003.1196005

WAXER, Cindy. 2013. *Big data blues: The dangers of data mining* [online] [cit. 2021-03-11]. Dostupné z: https://www.computerworld.com/article/2485493/enterprise-applications-big-data-blues-the-dangers-of-data-mining.html

ZAIANE, Osmar R., Donald IPPERCIEL a Samira ELATIA. 2016. *Data Mining and Learning Analytics: Applications in Educational Research (Wiley Series on Methods and Applications in Data Mining)*. Wiley. ISBN 978-1118998236.

ZAKI, Mohammed J. a Wagner MEIRA JR. 2014. *Data mining and analysis: fundamental concepts and algorithms*. II. Title. Rensselaer Polytechnic Institute, Troy, New York. ISBN 978-0-521-76633-3.

Bibliografie

LINOFF, Gordon a Michael J. A. BERRY. 2011. *Data mining techniques /: Gordon S. Linoff, Michael J. A. Berry*. 3rd ed. Indiana: Wiley. ISBN 978-0-470-65093-6.

RAUCH, Jan a Milan ŠIMŮNEK. 2014. Dobývání znalostí z databází, LISp-Miner a GUHA. Praha: Oeconomica, nakladatelství VŠE. Odborná kniha s vědeckou redakcí. ISBN 9788024520339.

RUSSELL, Matthew A. 2014. *Mining the social web: data mining Facebook, Twitter, LinkedIn, Google , Github, and more*. 2nd ed. Sebastopol: O´Reilly. ISBN 978-1-449-36761-9.

SKALSKÁ, Hana. 2010. Data mining a klasifikační modely. Hradec Králové: Gaudeamus. Recenzované monografie. ISBN 9788074350887.

ZAÏANE, Osmar R. 1999. *Principles of Knowledge Discovery in Databases* [online].s. 15 [cit. 2021-03-28].
Dostupné z:http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf