Jihočeská Univerzita v Českých Budějovicích

Ekonomická fakulta

Studijní program:	N6028 Ekonomika a management
Studijní obor:	Obchodní podnikání
Katedra:	Katedra aplikované matematiky a informatiky

Uplatnění moderní metody shlukové analýzy při segmentaci trhu

Vedoucí diplomové práce
Ing. Michael Rost, Ph.D.

Autor
Bc. Filip Joanidis, MBA

2012

Prohlášení

Prohlašuji, že svoji diplomovou práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s §47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své diplomové práce, a to v nezkrácené podobě, elektronickou cestou, ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích 16.4.2012 Filip Joanidis

Poděkování

Chtěl bych tímto poděkovat panu Ing. Michaelu Rostovi, Ph.D., za vedení diplomové práce, mnoho podnětných připomínek při jejím zpracování a pomoc při zvládání programového prostředí R. Dále bych chtěl poděkovat společnosti EON za možnost využít obchodní data pro účely zpracování této diplomové práce. Jsem přesvědčen, že právě tato možnost umožnila lépe posoudit vhodnost aplikovaných metod v reálném tržním prostředí. Poděkování náleží také manželce Ing. Lence Joanidisové za jazykovou korekturu a velikou trpělivost po dobu celého mého studia.

Abstrakt

Cílem této diplomové práce bylo popsat teoretické principy segmentace zákazníků pomocí moderních metod shlukové analýzy a jejich praktické ověření v komerčním prostředí utilitní společnosti.

Teoretická část práce vysvětluje principy shlukové analýzy prostřednictvím algoritmu K-means a algoritmů hierarchického shlukování. V rámci kapitoly SW podpory popisuje možnosti softwarových nástrojů R, Rapid Miner a KXEN. Pro účely zpracování dat jsou v teoretické části popsány metodiky SEMMA, 5A a metodika CRISP.

Praktická část práce se zaměřuje na ověření homogenity/heterogenity trhu s elektrickou energií v segmentu domácnosti v utilitní společnosti EON. Prostřednictvím dat ze zákaznického systému a dostupných externích dat, je pomocí programového prostředí R provedena hierarchická shluková analýza. Analýzou zjištěných shluků jsou identifikovány převažující charakteristiky jednotlivých shluků, které jsou následně podrobeny kritice z pohledu využitelnosti pro účely marketinku.

Klíčová slova:
Shluková analýza, CRISP, jazyk R, utilitní společnost, marketinkový management

Abstract

This thesis aims to describe the theoretical principles of customer segmentation using modern methods of cluster analysis and their practical verification in a commercial environment of utility company.

The theoretical part of the thesis describes the principles of clustering analysis using K-means algorithm and hierarchical clustering algorithms. The SW support chapter deals with the possibilities of R, Rapid Miner and KXEN software tools. For the purpose of data processing the theoretical part describes SEMMA, 5A and CRISP methodologies.

The practical part is focused on the analysis of homogeneity/heterogeneity of the energy market in the household segment in the utility company EON. Using data from company´s customer system, available external data and programming language R a hierarchical cluster analysis is processed. Through the analysis of gained clusters there are identified the main characteristics of particular clusters which are consequently examined from marketing usage perspectives.

Keywords:
Cluster analysis, CRISP, R language, utility industry, marketing management

Obsah

1 Úvod a cíl práce

Dnešní doba se vyznačuje velmi turbulentními změnami. Znalosti platné před několika desetiletími dnes pozvolna pozbývají na své hodnotě. Přístupy, které ještě v minulém desetiletí umožňovaly komerčním společnostem se držet na vrcholu trhu, je dnes sráží k zemi. Inovace, které dříve umožnili společnostem dlouhodobě vyniknout, jsou dnes napodobeny ve velmi krátké době, častokráte v měsících. Jednu z hlavních příčin lze pravděpodobně souhrnně pojmenovat termínem "globalizace".

Podíváme-li se na termín globalizace blíže, můžeme identifikovat jednotlivé směry, které mají na rozvoji globalizace zásadní podíl.

- **Rozvoj logistiky a mezinárodní dopravy** – díky vybudování poměrně husté sítě spočívající v letecké, lodní, vlakové a automobilové dopravě je možné zajistit přepravu zdrojů téměř z jakékoliv země do jiné. To umožňuje se jednak zaměřovat na efektivní využití zdrojů, které jsou v dané lokalitě k dispozici, zároveň to však umožňuje dopravovat na libovolné trhy nové produkty a zboží.
- **Rozvoj IT a komunikačních technologií** – díky technologiím jako jsou Internet, bezdrátové sítě, integrační technologie typu XML a EDI[1], je dnes možné daleko pružněji zajistit komunikaci a zpracování dat bez ohledu na vzdálenosti. Kromě zřetelného propojování jednotlivých společností v logistickém řetězci, vedou tyto technologie i k zásadním změnám v přístupu komunikace se zákazníky, ale i ke změně způsobu chování samotných zákazníků.
- **Politická stabilita** – i přes ekonomické problémy v posledních třech letech, můžeme definovat mezinárodní politické klima jako relativně stabilní pro ekonomickou spolupráci na úrovni celého světa. Současný stav je jednak dán vznikem mezinárodních uskupení zajišťující pravidla spolupráce v dané oblasti, ale i stále zlepšujícím se legislativním rámcem v jednotlivých zemích.
- **Globální využívání zdrojů** – oproti dřívějším dobám, je dnes snazší než kdykoliv dříve vyhledávat a využívat potřebné zdroje tam, kde jsou nejdostupnější a nejefektivnější. Zdrojem v tomto kontextu chápeme nejen nerostné a přírodní bohatství, ale i zemědělskou produkci, pracovní sílu či dokonce intelektuální kapitál.
- **Rozvoj v oblasti výzkumu** – možnosti uvedené výše zároveň přispívají k rozmachu nových objevů a to napříč všemi oblastmi – medicína, chemie, potravinářský průmysl, automobilový průmysl, informační technologie atd.

[1] XML, EDI - komunikační formáty, standardizující struktury pro výměnu dat. Používají se např. pro integraci informačních systémů mezi společnostmi

Globalizačních příčin a vlivů je možné samozřejmě nalézt i více. Smyslem tohoto výčtu bylo však pouze přiblížit některé změny, které vedly k tomu, že se dnešní společnosti potýkají se stále složitější situací na trhu. Kromě vlastního produktu, nebývale silné konkurence, růstu mezinárodních korporací a rychlosti inovací se dnešní komerční společnosti, které chtějí na trhu přežít, musejí zaměřit i na identifikaci potřeb zákazníka, porozumění jeho preferencím a rozpoznávání jednotlivých skupin zákazníků od sebe.

Z tohoto důvodu ve většině komerčních společností existuje útvar marketinku, který má vyhledávání nových zákazníků, jejich oslovování a zjišťování potřeb v náplni práce. Vlivem sílící konkurence si však nevystačí s metodami, které používali dříve, ale musejí vymýšlet stále nové a nové způsoby oslovování jednotlivých zákazníků. V dnešní době jistého informačního přehlcení, je však vyhledání těchto zákazníků a zjišťování jejich potřeb stále složitější a finančně nákladnější záležitostí.

Proto se jednotlivé společnosti obracejí k vědním disciplínám, které jim umožňují lépe nalézt danou cílovou skupinu zákazníků a na tu pak zacílit konkrétní marketinkové aktivity. Jednou z oblastí, která jim v tomto hledání může pomoci, je vícerozměrná explorační statistická analýza dat.

Statistika, jako jedna z oblastí aplikované matematiky, je vědou, zabývající se vysvětlováním určitých jevů prostřednictvím zkoumáním empirických dat. Na základě výsledků umožňují provedení rozhodování, a to nikoliv na bázi pocitů a intuice, ale na bázi konkrétních dat a faktů.

Aplikováno do oblasti marketinku, pomocí statistiky je možné např.

- Segmentovat zákazníky do vzájemně souvisejících skupin
- Hledat skryté vztahy v chování zákazníků (asociace)
- Identifikovat podvodné chování (fraud management)
- Predikovat budoucí vývoj poptávky
- Identifikovat hodnotu zákazníka (CLTV)
- Modelovat a simulovat chování zákazníků

Je naprosto zřejmé, že tyto znalosti, vhodně aplikované do praxe, umožní společnostem snížit náklady a zároveň lépe řídit vztahy se zákazníky a komunikaci s nimi.

Na druhé straně složitost statistických a matematických metod, nízká úroveň aplikace statistických znalostí do marketinkové praxe a v neposlední řadě i jistá osobnostní propast mezi lidmi zabývající se statistikou a lidmi z oblasti marketinku, vede k tomu, že i přes její značný potenciál, statistika v oblasti marketinku stále čeká na svou příležitost.

Jistým mostem je zde oblast, v komerčním světě nazývaná "business inteligence" s podoblastí nazývanou "data mining", která za pomoci softwarových nástrojů umožňuje převést statistické znalosti do komerční praxe, a to tak, aby se minimalizovaly potřebné matematické a statistické znalosti. Nicméně i přes více než dvacetiletou existenci těchto nástrojů, je jejich reálné využití s důrazem na získávání "business value[1]" stále diskutabilní.

Cílem této diplomové práce, je popsat využití statistických metod pro účely marketinku a následně je prakticky demonstrovat na vybraném příkladu. Vzhledem k poměrné šíři problematiky se zaměříme na možnosti segmentace zákazníků prostřednictvím algoritmů shlukové analýzy.

Práce je rozdělena do dvou hlavních částí. První část se zaměřuje na teoretický základ problematiky vícerozměrných statistických metod, aplikovaných metodik a SW nástrojů, druhá část pak aplikuje získané teoretické znalosti v oblasti segmentace zákazníků do oblasti utilitního trhu. Na závěr práce je provedeno zhodnocení získaných poznatků a uvedeno doporučení dalšího postupu.

[1] Hodnota, kterou daná společnost získává. Nemusí být nutně ekonomická, může to být např. i zvýšení prestiže, zlepšení kvality, apod.

2 Literární přehled

2.1 Marketingová segmentace trhu

Společnosti většinou nemohou oslovit veškeré zákazníky na trhu. Zákazníků je příliš mnoho, odlišují se svými potřebami a požadavky. Základní otázku, kterou si každý marketingový pracovník musí položit je, jak identifikovat segment, který má potenciál být novým trhem. Vzápětí je nutné si položit otázku, jaké kritéria zvolit, pro oslovení nejzajímavějšího segmentu. Proto společnosti musí nejprve provést segmentaci zákazníků se snahou provést nejlepší zacílení svých produktů na danou skupinu zákazníků.

Výchozím přístupem je tzv. **hromadný marketing (mass marketing).** V rámci tohoto přístupu společnosti v podstatě ignorují přání a požadavky jednotlivých zákazníků a snaží se zaměřit na masovou produkci, distribuci a propagaci svých produktů bez ohledu na geografické, kulturní, behaviorální či jiné odlišnosti zákazníků. Příkladem takového přístupu je např. společnost CocaCola nebo v minulosti H.Ford s modelem T Ford "v jakékoliv barvě jen když bude černá". Tento přístup je samozřejmě velmi nákladný a mohou si jej dovolit jen nadnárodní korporace. Z tohoto důvodu se většina společností zaměřuje na tvz. mikromarketing v jedné ze svých forem: segmentace, mikrosegmentace, lokální marketing a individualismus (Kotler, 2003).

Marketing segmentu je založen na oslovení zákazníků, kteří mají společné potřeby a požadavky. Takový přístup umožňuje se lépe zaměřit na tvorbu produktů, distribuční a komunikační strategie a tím i na efektivní využívání zdrojů. Tvorba segmentů je ale do jisté míry fikcí – každá společnost si identifikuje své segmenty a svou strategii jak se segmenty pracovat.

Marketinkový segment by neměl být zaměňován s pojmem sektoru. Sektor sice také tvoří homogenní skupinu zákazníků, ale nemusí nutně splňovat požadavek na stejné potřeby zákazníka. Jako příklad může být skupina zákazníků "studenti", kteří sice jsou dobře identifikovatelní, ale z hlediska jejich potřeb/požadavků nelze tvrdit, že jsou stejní.

Marketing mikrosegmentu se zaměřuje na ještě užší segment zákazníků. Hlavní filozofií je nabídnutí specializovaných vlastností produktů těm zákazníkům, kteří jsou za tyto vlastnosti ochotni zaplatit vyšší cenu. Zároveň úzkost segmentu by měla odradit ostatní konkurenci před vstupem na tento trh, takže společnost získává jistou exkluzivitu na trhu. Řada velkých firem tyto segmenty nechává zcela úmyslně stranou jejich zájmu (např. výroba náhradních dílů na stará auta).

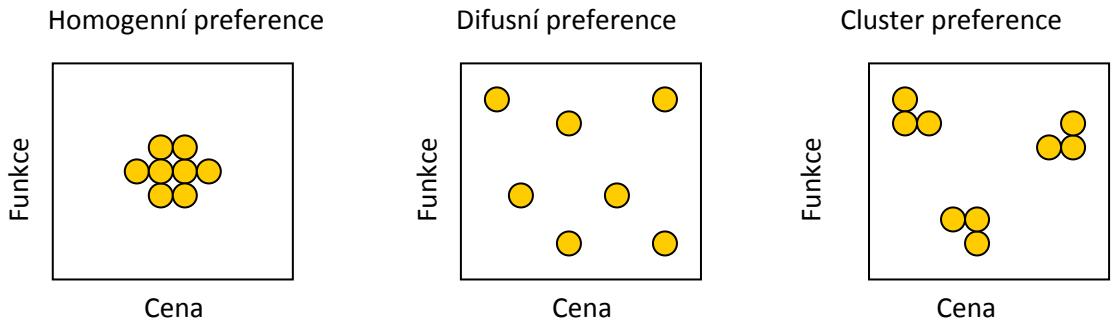
Lokální marketing se zaměřuje na individualizaci nabídky převážně dle geografických a demografických parametrů. Tímto způsobem se může odlišovat sortiment stejné společnosti v jedné čtvrti od sortimentu v jiné čtvrti. Tento přístup je však logisticky poměrně náročný. Tímto způsobem např. upravuje IKEA zboží a propagaci jednotlivých prodejen tak, aby odpovídali místní klientele. (Kotler, 2007)

Individuální marketing je poslední ze segmentačních strategií. Spočívá v respektování individuálních potřeb konkrétního zákazníka. Tento přístup byl dříve reprezentován v podstatě každým produktem, s příchodem průmyslové revoluce se však přesunul do oblasti masové standardizace. S rozvojem komunikačních a informačních technologií se tento přístup částečně navrací ve formě jakési "customizace" finálního produktu dle požadavků zákazníka.

2.2 Segmentační vzory

V této kapitole se zaměříme na situace, které z pohledu segmentace mohou nastat na jednotlivých trzích. Tyto situace mohou nabývat různých podob (Kotler 2003):

- **Homogenní preference** – trh obsahuje pouze zákazníky se stejnou preferencí. Neexistují zde žádné přirozené segmenty.
- **Difusní preference** – opačným extrémem je rozptýlení potřeb zákazníků indikující odlišnosti zákazníků napříč celým trhem. Tento vzor poukazuje na individualistické preference zákazníků (např. šperky)
- **Cluster preference** – je nejběžnějším vzorem rozdělujícím zákazníky do určitých skupin – segmentů, dle jejich individuálních požadavků. Pochopení těchto segmentů umožňuje firmám odlišovat své produkty a zaměřovat se na různé segmenty trhu a to i přesto že se pohybují v rámci jednoho sektoru.



Obrázek 1: Segmentační vzory (Kotler, 2003)

2.3 Segmentační kritéria

Jedním ze základních předpokladů pro provedení segmentace, je dostupnost vhodných kritérií, podle, kterých je segmentace prováděna. Čím více máme o dané skupině trhu informací, tím lepší můžeme provádět segmentaci. Pro účely segmentace můžeme rozdělit jednotlivá kritéria například do kategorií (Matula, 2012):

- **Spotřební trh:**
    - **Geografická kritéria** – teritorium, stát, kraj, město, velikost města, podnebí, vybavenost obce, morfologie krajiny
    - **Demografická kritéria** – věk, pohlaví, velikost rodiny, fáze životního cyklu rodiny
    - **Socio-ekonomická kritéria** – příjem, povolání, vzdělání
    - **Etnografická kritéria** – náboženství, rasa, národnost
    - **Fyziografická kritéria** – výška, váha, zdravotní stav, barva očí, vlasů
    - **Socio-psychologická kritéria** – sociální třída, životní styl, osobnost
    - **Behaviorální kritéria** – frekvence nákupu, objem nákupu, volba výrobků, preference vlastností výrobků, stupeň věrnosti, preference značky, komunikace případně jiná data získaná primárním výzkumem
- **Obchodní trhy -** pokud provádíme segmentaci pro obchodní trhy, je možné segmentaci doplnit o další kritéria:
    - **Charakter organizace** – právní forma, velikost organizace, obor, odvětví
    - **Provozní charakteristiky** – typ výroby, charakter spotřeby, požadavky na kvalitu a logistiku, obrat, profit
    - **Nákupní chování** – nákupní a platební politika, schéma rozhodování

2.4 Využití statistiky v marketinku

Pro účely marketinkových analýz nabízí statistika řadu metod. V rámci aplikované statistiky se v posledních letech prosazuje zejména koncept data miningu, který na základě analýzy dat, získaných z různých zdrojů, identifikuje vzory (patterny), které popisují určité vztahy a souvislosti mezi vstupními proměnnými (demografie, geografie atd.) a očekávaným výstupem (např. nákupní chování zákazníka). Při hledání vztahů, které by bylo možné zobecnit do použitelných pravidel, je nutné pracovat s řadou proměnných, definovat jejich vzájemné vazby, popř. provádět jejich redukci. Pro tyto účely lze efektivně využít metod vícerozměrné statistické analýzy. Mezi hlavní metody patří (Berka, 2003):

- **Regresní analýza** – pro zjišťování funkční závislosti jedné či více veličin na jiných vysvětlujících veličinách
- **Diskriminační analýza** – pro odlišení pozorování – objektů - patřících do různých tříd
- **Shluková analýza** – pro nalezení skupin (shluků) navzájem si podobných objektů
- **Korelační analýza** – pro posouzení závislosti mezi dvěma nebo více veličinami – proměnnými.
- **Analýza rozptylu** – posouzení rozdílu mezi průměry z různých výběrů
- **Faktorová analýza** – pro zjišťování závislosti jedné proměnné na tzv. faktorech vytvořených jako lineární kombinace jiných proměnných

Tyto metody se aplikují například na následující marketingové scénáře:

- **Analýza marketingových kampaní** – na základě analýzy výsledků historických marketingových kampaní se vytvoří model, prostřednictvím kterého se provádí odhady a modelování budoucích výsledků.
- **Predikce spotřeby** – analýzou historických dat je prostřednictvím predikčních modelu, založených např. na regresní analýze prováděn odhad budoucí spotřeby. Toto umožňuje např. obchodnímu oddělení lépe plánovat nákupy.
- **Detekce podvodů** – analýzou historických dat a následnou aplikací na nová data se provádí identifikace podezřelých transakcí či podezřelého chování zákazníků.
- **Segmentace zákazníků** – pomocí metod shlukových analýz jsou vytvořeny skupiny, které jsou si podobné svými potřebami, požadavky či nákupním chováním. Toto následně umožňuje lépe provádět marketingové kampaně, tvorbu nových produktů či přizpůsobení marketingového mixu
- **Analýza nákupního košíku** – na základě asociačních analýz jsou identifikovány produkty, které zákaznici kupují společně. Toto umožňuje optimalizovat rozložení produktů v obchodních jednotkách, nebo, typické pro internetové obchody, nabízet při nákupu zákazníkovi související produkty (fotoaparát + pouzdro + baterie)

- **Skóring, bonitace klienta** – na základě analýzy chování zákazníků provedené na historických datech, je vytvořena aplikace, která na základě zadaných parametrů provede klasifikaci zákazníka do dané bonitní kategorie. Na základě ní pak zákazník dostává/nedostává určité služby. Toto je typicky využíváno např. bankami při poskytování hypotečních úvěrů.
- **Analýza odchodů zákazníka (churn management)** – cílem je sestavit model identifikující zákazníky s největší tendencí k odchodu a následnou akcí toto riziko snížit.
- **Zlepšení procesů** – cílem úlohy je identifikovat signály/slabá místa vedoucí k neefektivitě či chybovosti procesů. Tímto způsobem je možné například indikovat vzrůstající pravděpodobnost přetížení sítě a informováním obsluhy zabránit jejímu výpadku.

Marketinkových aplikací statistických metod je možné nalézt daleko více. V rámci dalšího textu této práce se však zaměříme především na analýzu a aplikaci metod shlukové analýzy.

2.5 Shluková analýza

2.5.1 Cíle shlukové analýzy

Pojem shluková analýza je netriviální proces, zahrnující celou řadu metod a přístupů, jejichž cílem je rozdělit pozorované objekty do skupin (shluků) vzájemně si blízkých případů a zároveň dosáhnout stavu, kde jednotlivé shluky jsou si podobné co nejméně. Zároveň platí, až na některé přístupy, že jednotlivé objekty náleží pouze do jednoho shluku[1].

Při provádění shlukové analýzy je klíčovou otázkou definice pojmu *podobnosti* jednotlivých objektů. Míry stanovení podobnosti se provádějí odlišně pro kvalitativní, kvantitativní a smíšené proměnné. Vlastní tvorbu shluků je možné provádět pomocí různých algoritmů, které v rámci klasické shlukové analýzy můžeme rozdělit na *hierarchické algoritmy* a *optimalizační algoritmy* (Hebák, 2005). Problematice výpočtu podobnosti i otázce shlukových algoritmů se budeme věnovat v rámci následujících kapitol.

Jak již bylo zmíněno v kapitole segmentačních vzorů, podmínkou smysluplných výsledků je samotná existence shluků v reálném světě. Aby bylo možné výstupy shlukové analýzy použít např. pro marketinkové kampaně, tvorbu produktů či segmentaci zákazníků, je nezbytné tyto výstupy následně správně interpretovat, ať už z pohledu obchodní logiky, znalosti daného trhu, či jiné hledané vazby. Samotné vytvořené shluky tedy ještě nemusejí obsahovat reálně použitelný výsledek.

Pro úplnost dodejme, že pro účely segmentace trhů je možné, kromě metod shlukové analýzy, použít i jiných statistických metod. Jedná se například o metody vícerozměrného škálování, faktorové analýzy, neuronové sítě, regresní metody či metody genetických algoritmů.

[1] Tento přístupu se nazývá „crisp clustering“. Výjimkou jsou „fuzzy cluster“ algoritmy

2.5.2 Míry vzdálenosti a podobnosti

Prvním krokem pro tvorbu shluků je výpočet měr vzdálenosti a podobnosti resp. nepodobnosti pozorovaných objektů. Zásadní roli pro určení těchto měr má typ dané proměnné

2.5.2.1 Míry vzdálenosti a nepodobnosti pro kvantitativní proměnné

Pokud jsou sledované objekty charakterizovány kvantitativními veličinami, pokud možno na stejných úrovních nebo vyjádřené ve stejných měrných jednotkách (Kč, MWH, počet), můžeme pro výpočet použít:

Euklidovskou vzdálenost – vychází z aplikace Pythagorovy věty aplikované do vícerozměrného prostoru:

$$D_E(x_i, x_{i'}) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}$$

[Vzorec 1]

Hemmingovu vzdálenost – nazývanou také Manhattan nebo city-block, která je definována jako prostý součet absolutních hodnot rozdílů mezi jednotlivými proměnnými:

$$D_H(x_i, x_{i'}) = \sum_{j=1}^{p} \left| x_{ij} - x_{i'j} \right|$$

[Vzorec 2]

V případě, že jednotlivé proměnné jsou reprezentovány odlišnými měrami, není možné provádět prostý součet jejich jednotlivých vzdáleností. Důvodem je jejich odlišný vliv na výsledek. Z tohoto důvodu je vhodné proměnné standardizovat.

2.5.2.2 Míry nepodobnosti pro kvalitativní proměnné

Marketinkové pozorování zřídka obsahuje pouze kvantitativní veličiny. Ve většině případů je nutné pracovat s odpověďmi z dotazníků, mající kvalitativní charakter, ať již nominální nebo ordinální. V případě takových proměnných je nutné pracovat s jiným vyjádřením míry nepodobnosti. Nejčastěji se pracuje s vyjádřením míry koeficientu prosté neshody, definovaném jako podíl počtu neshodných proměnných dvou objektů, vůči celkovému počtu proměnných (Hebák [3], 2007).

$$D_{SM}(x_i, x_{i'}) = \frac{\sum_{j=1}^{p} g_{ij}}{p}$$

[Vzorec 3]

Kde $g_{ij} = 1$ pokud $x_{ij} \neq x_{i'j}$ a $g_{ij} = 0$ v ostatních případech.

2.5.2.2.1 Konverze proměnných na binární čísla

Kvalitativní proměnné jsou ve většině případů reprezentovány textovým řetězcem. Tyto řetězce není možné ve většině případů použít pro matematicko-statistické vyhodnocení. Z tohoto důvodu je nezbytné provést konverzi jednotlivých proměnných na skupinu binárních čísel, popřípadě, pro ordinální data, na pořadová čísla. Po této konverzi je následně možné přistoupit k vlastnímu procesu vyhodnocení. Na následujících příkladech vysvětlíme, jak jsou jednotlivé typy kvalitativních proměnných konvertovány na binární čísla. Za zmínku stojí, že v rámci konverze dochází k zvětšování počtu proměnných a tedy i k zesložiťování celé úlohy (Hebák [3], 2007).

- **Binární proměnné**

Odpověď	X1
Ano	1
Ne	0

Tabulka 1: Převedení binomické proměnné na alternativní binární proměnnou

- **Nominální proměnné**

Barva očí	X1	X2	X3
Hnědá	1	0	0
Zelená	0	1	0
Modrá	0	0	1

Tabulka 2: Převedení nominální proměnné na skupinu alternativních proměnných se shodnou vzdáleností mezi proměnnými

- **Ordinální proměnné**

Spokojenost	X1	X2	X3
Nespokojen	0	0	0
Spíše nespokojen	1	0	0
Spíše spokojen	1	1	0
Velmi spokojen	1	1	1

Tabulka 3: Převedení ordinální proměnné na skupinu alternativních proměnných s rozdílnou vzdáleností mezi proměnnými

2.5.2.3 Míry nepodobnosti pro proměnné různých typů

V případě že výběrová statistika obsahuje soubor proměnných různých typů, doporučuje se počítat Gowerův koeficient nepodobnosti, představující vážený průměr dílčích měr nepodobnosti (Hebák [3], 2007):

$$D_{ii'} = \frac{\sum_{j=1}^{p} w_{ii'j} d_{ii'j}}{\sum_{j=1}^{p} w_{ii'j}} \qquad \text{[Vzorec 4]}$$

Kde:

Váha $w_{ii'j} = 0$ pokud hodnota x_{ij} nebo $x_{i'j}$ chybí, nebo jsou obě hodnoty rovny nule.

Hodnota $d_{ii'j}$ představuje míru nepodobnosti a závisí na typu proměnné:

- nominální nebo binomická data $d_{ii'j} = 0$ pro $x_{ij} = x_{i'j}$, v ostatních případech $d_{ii'j} = 1$
- ordinální proměnné nebo kvantitativní proměnné měřené na poměrové škále – je každá hodnota proměnné x_{ij} transformována na hodnotu z_{ij} a následně spočítána míra nepodobnosti jako pro proměnné měřené na intervalové škále.

$$z_{ij} = \frac{t_{ij} - 1}{M_j - 1} \qquad \text{[Vzorec 5]}$$

$t_{ij} \in \{1 .. M_j\}$, nahrazující proměnou X_j pořadovým číslem
M_j - maximální pořadí j-té proměnné

- kvantitativní proměnné měřené na intervalové škále

$$d_{ii'j} = \frac{\left| x_{ij} - x_{i'j} \right|}{\max x_p - \min x_p} \qquad \text{[Vzorec 6]}$$

2.5.3 Algoritmy shlukování

Základní metody shlukové analýzy je možné rozčlenit do dvou základních přístupů (Berka, 2003):

- Hierarchický přístup
  - Aglomerativní hierarchický přístup
  - Divizní monotetický přístup
  - Divizní polytetický přístup
  - Dvourozměrné aglomerativní shlukování

- Metody optimalizační např.:
  - Algoritmus K-průměrů
  - Fuzzy shluková analýza

V rámci této práce se zaměříme na aglomerativní hierarchický přístup a optimalizační algoritmus K-průměrů

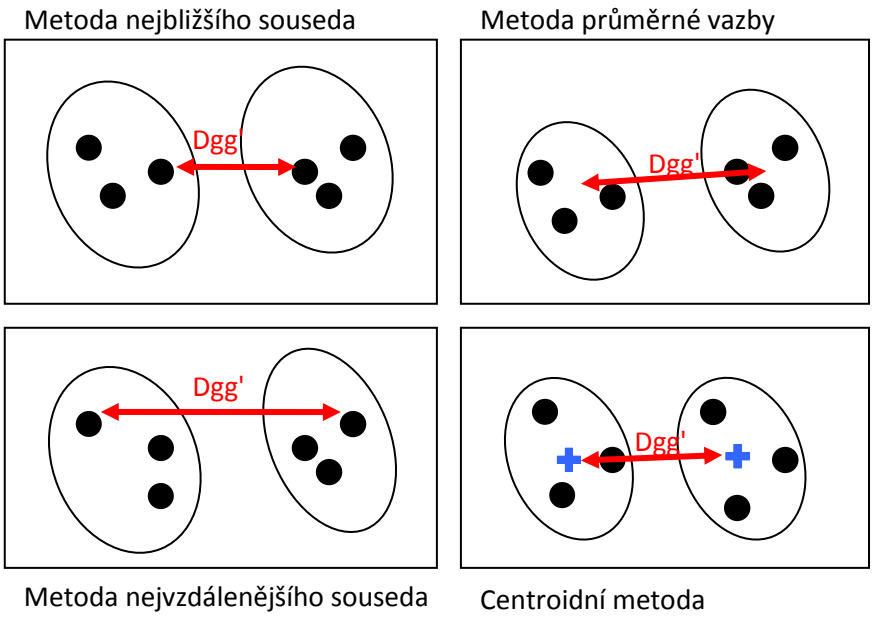
2.5.3.1 Aglomerativní hierarchický přístup

Algoritmus spočívá v postupném vytváření posloupnosti shluků zdola-nahoru, který lze popsat následujícím způsobem (Berka, 2003):

1) Inicializace
   a. Urči vzájemnou vzdálenost mezi jednotlivými objekty
   b. Zařaď každý objekt do samostatného shluku (tj. počet shluků = počet příkladů)
2) Hlavní cyklus
   a. Dokud je více než jeden shluk
      i. Najdi dva navzájem nejbližší shluky a spoj je
      ii. Spočítej pro tento shluk vzdálenost od ostatních shluků

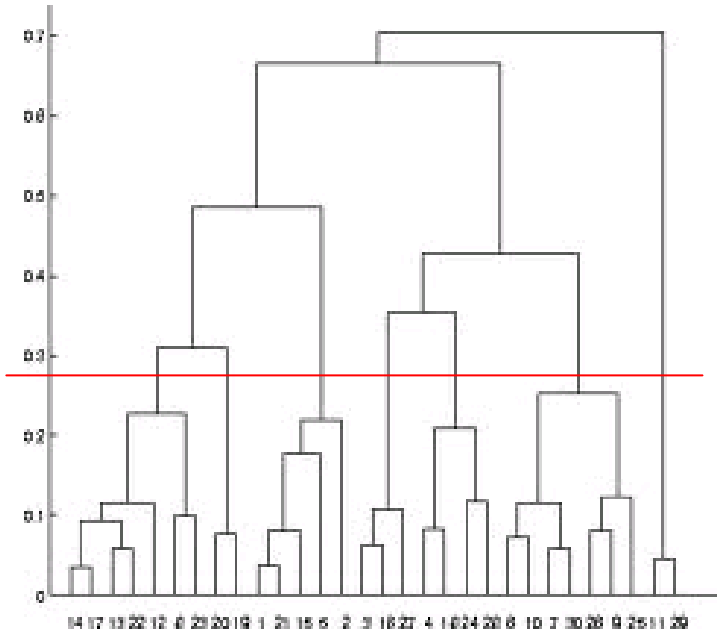
Výpočet nových vzdálenosti mezi novými shluky je možné určit různými způsoby:

- Metodou nejbližšího souseda – vzdálenost mezi shluky je dána minimem ze vzdáleností mezi jejich objekty
- Metodou nejvzdálenějšího souseda - vzdálenost mezi shluky je dána maximem ze vzdáleností mezi jejich objekty
- Metoda průměrné vazby (Sokalova-Snethova) - vzdálenost mezi shluky je dána průměrem ze vzdáleností mezi jejich objekty
- Centroidní metoda (Gowerova metoda) - vzdálenost mezi shluky je dána vzdáleností mezi středy shluků



Obrázek 2: Různé způsoby měření mezishlukových vzdáleností

Výsledek procesu shlukování je možné vizualizovat prostřednictvím tzv. dendrogramu. Ten zobrazuje proces postupného shlukování. Optimální počet shluků se odvozuje až v rámci analýzy výsledků tak, že proces aglomerativního shlukování rozdělíme na určité hladině



Obrázek 3: Příklad hierarchického dendrogramu

2.5.3.2 Algoritmus K-průměrů

Algoritmus k-means je založen na přesunování jednotlivých objektů mezi shluky. Základem je předpoklad, že víme, do kolika shluků chceme objekty rozdělit, tj. známe počet shluků a priory.

Popis algoritmu (Berka 2003, Hebák 2007):

1) Zvol počáteční rozklad objektů do *K* shluků (většinou náhodně)
2) Urči centroidy[1] pro všechny shluky v aktuálním rozkladu
3) Pro každý objekt
    a. Urči vzdálenost $D(x, c_k)$, kde c_k je centroid daného k-tého shluku
    b. Pokud má daný objekt nejblíže k vlastnímu centroidu, necháme jej v daném shluku; pokud má blíže k jinému centroidu, přesuneme jej do příslušného shluku
4) Nedochází-li již k žádnému přesunu, ukonči úlohu. Jinak se vracíme ke kroku 2.

Popsaný algoritmus je velmi efektivní, k suboptimálnímu výsledku dojdeme již po několika iteracích. Zároveň je výpočetně méně náročný než hierarchické shlukování a proto je možné jej použít i na větší data (Berka 2003, Hebák - Hustopecký 2007).

Uvedený algoritmu může mít i několik modifikací:

a) Proces shlukování můžeme zahájit s K vybranými objekty, které se tak stanou centroidy. Odpadá tak inicializace úlohy v krocích 1 a 2
b) Přepočet centroidů je možné provádět po každém přesunu (McQueenův algoritmus)
c) Zadáním několika parametrů lze dojít k rozkladu s vhodným počtem shluků. Toto je založeno na parametrech minimální nebo maximální přípustné vzdálenosti. Objekty, které dané parametry nesplňují, splynou s jiným shlukem, popřípadě vytvoří shluk nový.

Takto popsaný algoritmus je možné použít jen na kvantitativní data. V marketingu, kdy se setkáváme spíše s nominálními daty, je možné použít modifikovanou metodu tzv. k-modů, založenou na využití měr nepodobnosti.

[1] Centroidem je míněn aritmetický průměr všech objektů v daném shluku

2.5.4 Problematika kvality a velikosti dat

Metody popisované v předchozích kapitolách je z důvodu výpočetní náročnosti možné použít pouze pro malé datové soubory. Za velké soubory se již označují soubory mající více než 250 objektů. Má-li navíc soubor více než 16 proměnných, jsou již obtížně identifikovatelné rozdíly mezi jejich vzdálenostmi. (Hebák, 2003).

V marketinkové praxi se však často setkáváme s požadavky analyzovat soubory o několika tisících objektů, s desítkami až stovkami proměnných. V takových případech je nutné buď přistoupit k *redukci velikosti datového souboru*, nebo *využít nových metod*, které zpracování velkého množství objektů a proměnných umožňují.

Mezi postupy umožňující zmenšit velikost datového souboru můžeme řadit:

- **Redukce počtu proměnných** – ne veškeré proměnné mají význam pro určení výsledných shluků. Na základě určení hodnot vzájemné korelace proměnných je možné redukovat proměnné, které mají vzájemnou silnou korelaci. Jiný přístup je určení tzv. informačního zisku a vyloučení proměnných s malým informačním ziskem. Informační zisk vyjadřuje rozdíl entropie pro celá data (pro cílový atribut) a pro uvažovaný atribut (Berka, 2002). Místo informačního zisku, můžeme použít tzv. Giniho index, vyjadřující míru variability pro kategoriální proměnné.
- **Agregace kategorií dat** – spočívá v seskupení podobných kategorií jedné proměnné do nové skupiny. Tím dojde ke snížení variability proměnných a tím ke snížení požadavků na výpočetní výkon
- **Redukce velikosti prostřednictvím náhodného výběru** – analýza je provedena pouze na malém vzorku dat, reprezentující původní soubor. Pro tyto účely se používá tzv. stratifikovaný výběr, reprezentující zastoupení objektů z hlediska jednotlivých kategorií proměnných.
- **Využití principů z jiných metod, například z metod strojového učení**. V tomto přístupu se analýza provádí pouze na určité části dat (trénovací data), následně se výsledky ověřují na jiné množině dat představující testovací vzorek. Tímto způsobem je zajišťována tvorba kvalitního modelu.

Nové metody shlukové analýzy jsou založeny na těchto základních přístupech

- **Rozdělovací metody** – princip spočívá v rozdělení dat do p bloků (frakcí). V rámci těchto bloků je klasickými metody tvořeno *k* shluků. Tímto způsobem získáme pk objektů, které následně opět shlukujeme do *k* shluků. Zbývající objekty jsou pak rozděleny k těmto výsledným shlukům. Na této metodě je založena např. metoda CLARA, PAM
- **Metoda postupného shlukování** - přiděluje jednotlivé objekty krok za krokem do jednotlivých shluků. Pokud se k existujícím shlukům nehodí, je vytvořen shluk nový.

Tyto přístupy jsou pak aplikovány např. v metodách:

- **Hybridní klasifikace** – aplikace rozdělovacího přístupu na základní metodu k-means. Tento přístup je reprezentován např. algoritmem CLARA.
- **Frakcionalizace** – aplikace rozdělovacího přístupu na hierarchické algoritmy. Jedná se např. o algoritmy BIRCH, CURE, ROCK, Chameleon).
- **Metody založené na hustotě** – metody popisující shluky jako oblasti ve výběrovém prostoru, které se vyznačují značnou hustotou bodů. Objekty, které se nacházejí mimo tyto oblasti, jsou označovány za šum. Tyto metody jsou např. uplatněny v algoritmech DBSCAN, OPTICS, DENCLUE
- **Metody pro shlukování podprostorů** – metody určené pro datové soubory s velkým počtem proměnných (algoritmy CLIQUE, ENCLUS, MAFIA, OptiGrid).

2.6 Metodiky zpracování dat

Pouze znalost popsaných technik a matematických algoritmů pro praktické využití v marketinku nestačí. Potřebným propojením mezi statistickými znalostmi a marketinkovou praxí je dostupnost kvalitní metodiky. Její úlohou je zajisti konzistentní proces začínající definicí marketinkových cílů, výběrem vhodných postupů, validací předpokladů pro jejich použití, s následnou aplikací a interpretací výsledků do marketinkové praxe. K nejznámějším metodologiím patří (Berka 2003):

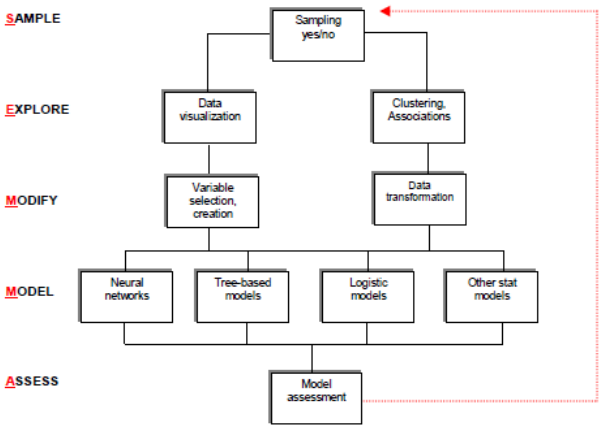
- SEMMA
- 5A
- CRISP-DM

2.6.1 Metodika SEMMA

Metodika SEMMA je dodávána firmou SAS, jednou z předních firem v oblasti Business Intelligence a Data Minig.

Metodika je rozdělena do pěti etap (CAIS 2002):

- **Sample** – v prvním kroku jsou výběrově vytvořeny jedna nebo více tabulek z dostupných dat. Analýza reprezentativního vzorku výrazně snižuje čas na zpracování.
- **Explore** – v dalším kroku se metodika zaměřuje na vizuální analýzu vybraného vzorku. Cílem je najít důležité vlastnosti souboru, shluky, odlehlé hodnoty, ověřit normalitu rozdělení apod. V rámci tohoto kroku se využívají techniky jako je korelační analýza, faktorová analýza či clustering.
- **Modify** - cílem třetího kroku je úprava vzorku dat pro následnou analýzu – odstranění chybných hodnot, vypořádání se s chybějícími údaji.
- **Model** – vlastní tvorba modelu pomocí data miningových technik
- **Assess** – poslední částí procesu je ověření robustnosti modelu na novém vzorku dat, vyhodnocení a interpretace výsledků.

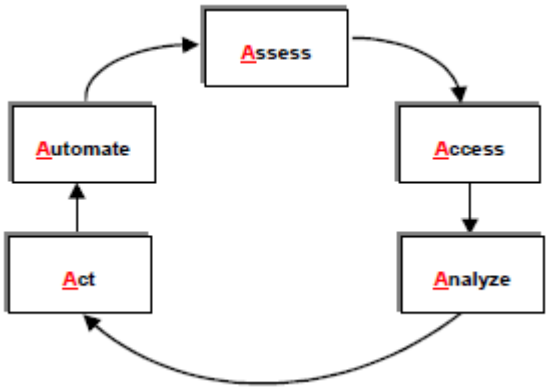


Obrázek 4: SEMMA Analytický proces (CAIS, 2002)

2.6.2 Metodika 5A

Metodika 5A je dodávána jakou součást produktu SPSS od firmy IBM. Metodika definuje kroky (CAIS 2002):

- Assess – posouzení potřeb projektu, definice strategie, cílů, analýza obchodního procesu
- Access – zajistění potřebných dat
- Analyze – analýza dat s aplikací vybraných dataminingových technik
- Act – interpretace získaných dat
- Automate – aplikace zjištěných výsledků do praxe



Obrázek 5: Metodika „5A“ (CAIS, 2002)

Metodika je v mnohém podobná metodice SEMMA a je nutno zmínit, že firma SPSS (dnes IBM) od rozvoje této metodiky upustila a zaměřuje se na metodiku CRISP-DM

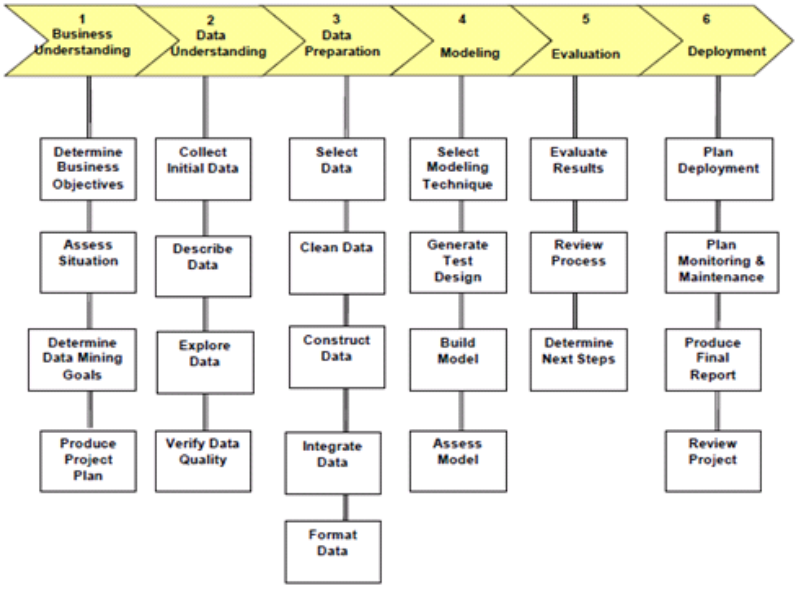
2.6.3 Metodika CRISP DM

Metodika CRISP DM, vyvinutá konsorciem firem NCR, Daimler Chrysler, OHRA a ISL je dnes jednou z nejvíce používaných metodik pro data mining. Zároveň je považována jako metodika nezávislá na daném SW prostředí či odvětví.

Metodika je rozdělena do etap (CRISP-DM, 2000)

- Porozumění problematice (Business Understanding)
- Porozumění datům (Data Understanding)
- Příprava dat (Data Preparation)
- Modelování (Modeling)
- Vyhodnocení výsledků (Evaluation)
- Využití výsledků (Deplyment)

Každá etapa je dekomponována do úloh a následně do sady výstupů.



Obrázek 6: CRISP DM Proces (CAIS, 2002)

- **Porozumění problematice (Business Understanding)** – první etapa se zabývá přípravou a organizací projektu. Mapují se obecné business cíle a požadavky. Následně se definují cíle dataminingové úlohy, stanovují se faktory úspěchu a provádí se hrubé mapování dostupných dat. Výstupem této etapy je připravený projektový plán.
- **Porozumění datům (Data Understanding)** - druhá etapa se zabývá zajištěním dostupných dat, jejich popisem (formáty, počty záznamů, zdroje), prozkoumáním a verifikací datové kvality.

- **Příprava dat (Data Preparation)** - v rámci třetí etapy se vybraný vzorek dat připravuje pro účely následného modelování. Data jsou očištěna o nežádoucí údaje, jsou doplňovány chybějící údaje, data jsou rozšiřována o derivované atributy. Zároveň se provádí integrace na další datové zdroje a data se formátují do požadovaného tvaru.
- **Modelování (Modeling)** - čtvrtá etapa se zabývá vlastním modelováním. V rámci etapy se vybere vhodná technika (rozhodovací stromy, shlukování atd.), připraví se scénář (typicky rozdělení do testovacího a validačního vzorku) a následně se spustí daný model. Tento postup je zpravidla nutné opakovat vícekrát, s modifikací vstupních parametrů, se snahou o dosažení nejlepších výsledků.
- **Vyhodnocení výsledků (Evaluation)** – další fáze se zaměřuje na vyhodnocení výsledků modelování, kontrolu správnosti celého procesu a definici následných kroků.
- **Využití výsledků (Deplyment)** – v rámci poslední fáze se výsledky převádějí do praxe. Etapa se zabývá plánováním aplikací získaných znalostí, přípravou a prezentací výsledků sponzorům, nastavení pravidelných činností a závěrečnou akceptací výsledků.

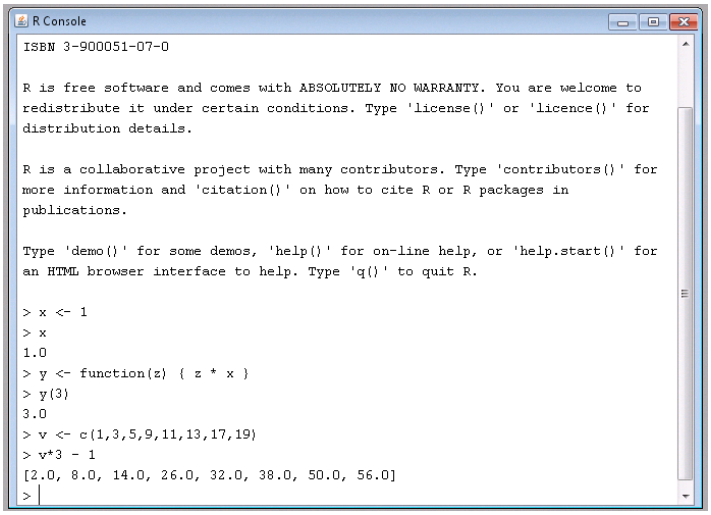
2.7 Softwarová podpora

Dalším důležitým rozhodnutím v rámci procesu získávání znalostí z dat je i volba vhodného SW nástroje. Na trhu existuje poměrně široká škála produktů, od volně šiřitelných (freeware) až po velmi komplexní a nákladná řešení významných SW firem. Rozdíly mezi jednotlivými produkty se, kromě ceny, projevují jednak v rozsahu podporovaných statistických metod, ale i ve způsobu podpory přípravy dat, podporované metodiky, grafické interpretace výsledků a schopnosti zpracovávat velké objemy dat. V neposlední řadě je i důležitým hlediskem uživatelská přívětivost a dostupnost technické podpory. V rámci kapitoly SW podpora se podíváme alespoň na některé z nich.

2.7.1 Statistické prostředí R

„R“ je programové prostředí určené pro statistiku a grafiku. Obsahuje velkou řadu statistických funkcí a algoritmů. Díky tomu, že program je zároveň šířen pod GNU[1] projektem, existují stovky nadšenců, kteří tento produkt ještě dále rozšiřují. R poskytuje řadu technik, od lineárních a nelineárních modelů, statistických testů, analýzu časových řad, až po klasifikaci a shlukovou analýzu. Kromě vlastních technik existuje řada specifických rozšíření (package) pro medicínu, biologii, finance a další.

Ačkoliv k prostředí R existuje řada grafických nadstaveb, samotné prostředí je v zásadě založeno na příkazovém řádku. Tato „uživatelská nepřívětivost“, i přes velice silný statistický základ, může mít za následek preferenci jiných nástrojů než R.



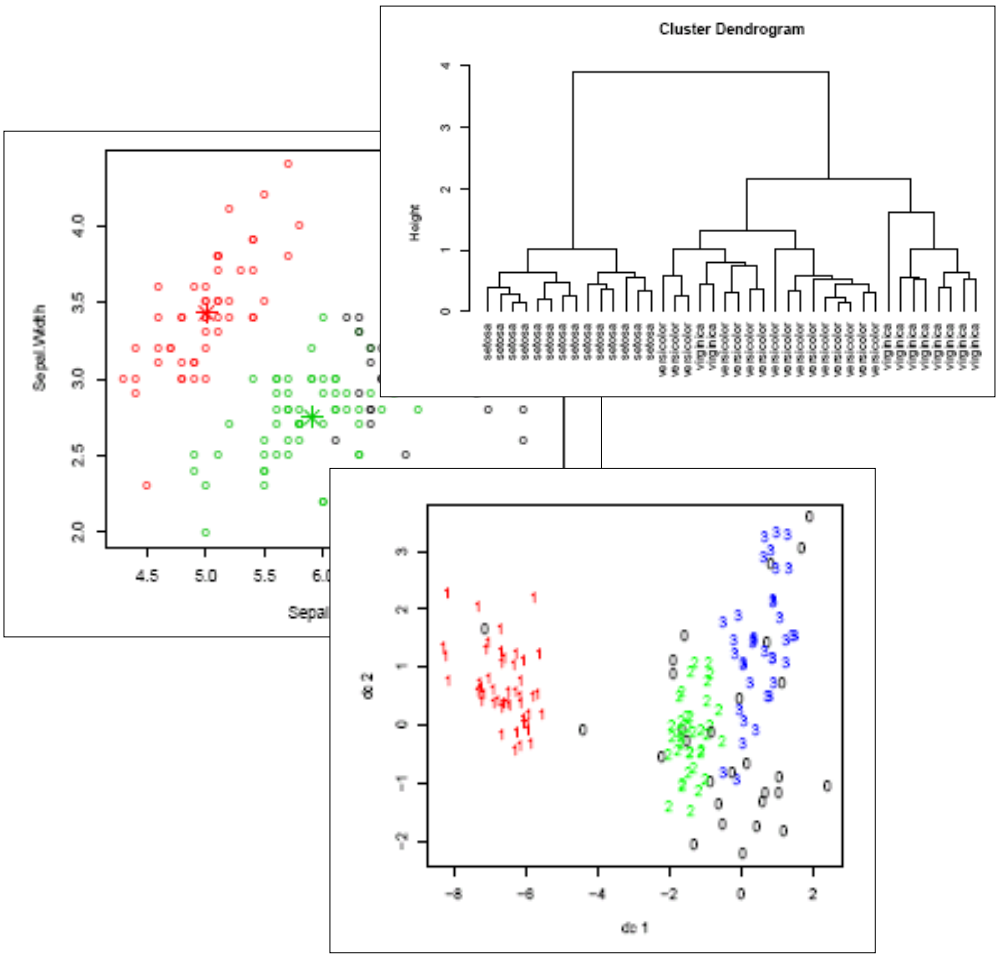
Obrázek 7: Dialogové prostředí jazyka R

[1] GNU – projekt, založený na filozofii volně šiřitelného software.

Pro účely shlukové analýzy je k dispozici řada algoritmů:

- K-means clustering (funkce *kmeans*)
- PAM – partitioning around medoids (funkce *pam*) – robustnější shluková analýza podobná K-means clusteringu
- Hierarchický clustering (funkce *hclust*)
- Divizní clustering (*diana,agnes,mona)* – monotetický a polytetický hierarchický přístup
- Dvourozměrný clustering (funkce *twins*)
- Density based clustering (funkce *dbscan*)
- Fuzzy Clustering (funkce *fanny*)
- Clustering Large Application (funkce *clara*) – shulková analýza určená pro rozsáhlá data

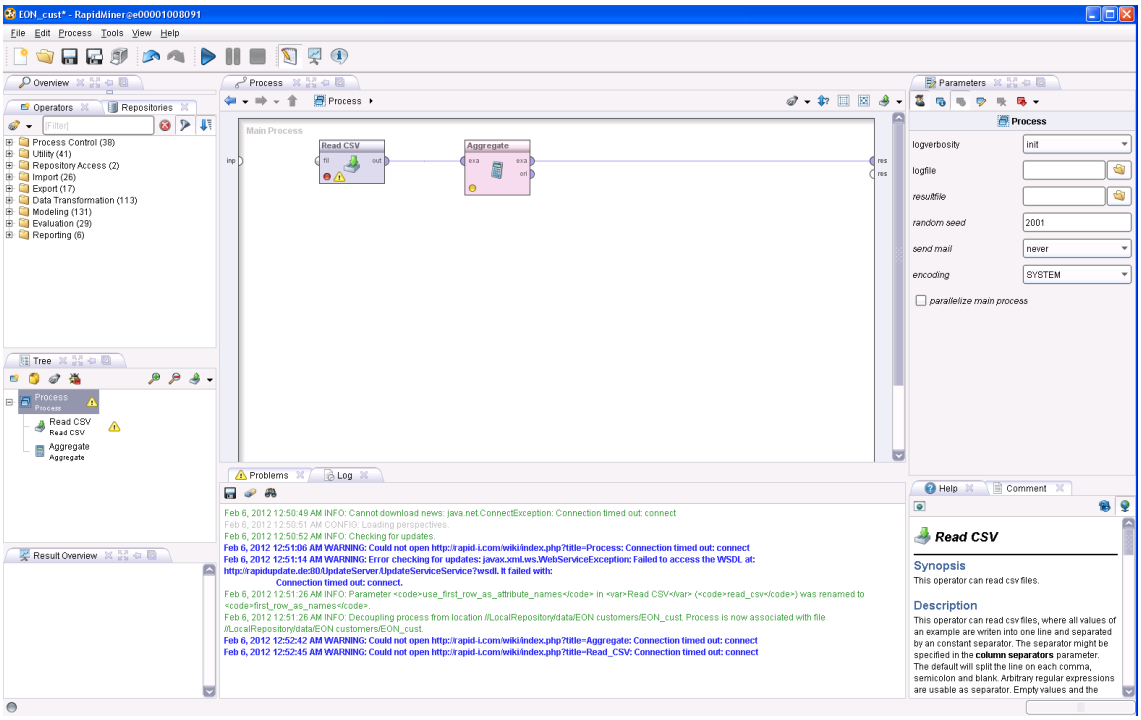
Pro výsledky je možné použít řadu grafických funkcí.



Obrázek 8: Příklad grafických výstupů jazyka R

2.7.2 Rapid miner

Rapid miner je dalším z nástrojů šířeného pod GNU licencí. Dle KDNudggets.com se jedná o jeden z nejrychleji se rozvíjejících SW pro data mining. Podobně jako jazyk R nabízí řadu modelovacích technik od regresních analýz, korelace, shlukových algoritmů až po metody strojového učení a algoritmů neuronových sítí. Na rozdíl o jazyka R je Rapid Miner založen na velice propracovaném grafickém rozhraní, umožňujícím provádět modelování i bez jakékoliv znalosti programování či jazykových syntaxí.



Obrázek 9: Grafické prostředí Rapid Miner

Pro účely shlukové analýzy Rapid Miner nabízí algoritmy

- K-Means včetně různých modifikací
- K-Medoids
- DBScan clustering
- EM-algoritmus
- Support Vector Clustering
- Random Clustering
- Agglomerative clustering (hierarchické metody)

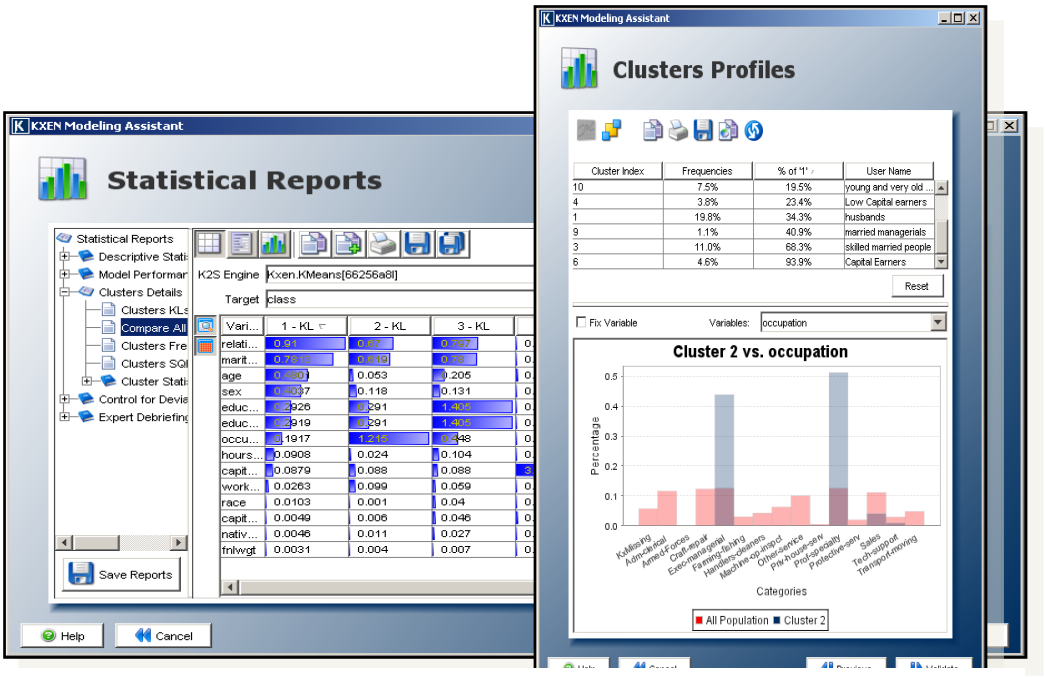
2.7.3 KXEN

KXEN je data miningový produkt francouzské společnosti KXEN. Produkt je rozdělen do několika modulů, které zajišťují problematiku vstupního zpracování dat, transformace, modelování až po funkcionalitu pro následnou interpretaci výsledků a aplikaci výsledků do praxe. Z hlediska modelování obsahuje funkce pro klasifikaci, regresní analýzu, segmentaci, modelování časových řad, predikci a asociační analýzu/analýza nákupního košíku.

Od předchozích produktů se tento nástroj liší naprosto odlišným přístupem k problematice zpracování dat. Filozofie vychází z předpokladu, že většina uživatelů nemá potřebné znalosti statistiky pro správný výběr a zpracování dat. Proto je veškerá problematika statistického zpracování dat dána na pozadí, uživatel je pomocí asistovaných průvodců proveden pouze jednoduchými scénáři bez nutnosti znalostí jednotlivých algoritmů a jejich předpokladů použití. Systém sám na pozadí provádí veškerou přípravu dat, jejich případnou standardizaci, výběr nejvhodnější algoritmů, uplatnění modelu a výslednou validaci. Výsledky analýzy jsou zpřístupněny ve formě reportů a grafických analýz. Velký důraz je kladen na reálné využití v marketinkové praxi.

Jako nevýhoda tohoto nástroje může být vnímána právě tato uzavřenost s nemožností ovlivnit způsob statistického zpracování a algoritmizace úlohy.

Z pohledu shlukové analýzy systém obsahuje pouze analýzu založenou na algoritmu K-means, který je navíc provázán s dalšími algoritmy pro zpracování velkých objemů dat, interpretací výsledků a aplikací výsledků do praxe. Jedním z výsledků zpracování je kromě vytvoření vlastních shluků i prezentace analýzy profilu získaných shluků, která umožňuje zkoumat zastoupení jednotlivých proměnných v rámci daného shluku.



Obrázek 10: Analýza zastoupení proměnných v jednotlivých clusterech

3 Materiál a metodika

Trh s utilitami (elektřina, plyn) se během posledních deseti let výrazně změnil. Díky liberalizaci trhu, došlo k oddělení původně monopolního procesu na části výroby, distribuce a prodeje el. energie. Zejména v oblasti prodeje toto oddělení mělo obrovský vliv na situaci na trhu. Kromě původních monopolních společností se na trhu objevila řada nových společností, nabízejících zákazníkům své produkty. V dnešní době existuje v České Republice asi 110 společností s licencí na prodej el. energie, z toho přibližně 10 jich je velmi aktivních.

Tato změna vyžaduje i změnu přístupu k zákazníkovi. Cílem jednotlivých společností je získání co největšího počtu zákazníků a prostřednictvím vhodně strukturovaných produktů zajistit jejich co možná největší loajalitu - retenci. Identifikace jednotlivých skupin zákazníků, pochopení jejich potřeb a chování je jednou z priorit pro zvolení vhodné produktové strategie a oslovení správného segmentu zákazníků.

Cílem této diplomové práce je provést segmentaci zákazníků ve vybrané společnosti s uplatněním moderních statistických metod. Jako společnost, na které budou statistické metody ověřovány, byla vybrána společnost EON České Republika, člen mezinárodní skupiny EON, jeden z největších utilitních dodavatelů v Evropě. Základem je předpoklad, že existují přirozené skupiny zákazníků, které je možné prostřednictvím shlukové analýzy identifikovat. Rozborem těchto skupin (shluků) budou následně navrženy praktické možnosti využití v marketinkové praxi.

Jako metodika zpracování bude použita metodika CRISP, popsaná v rámci oddílu „literární přehled“. Správné dodržení této metodiky by mělo zaručit, že žádný z kroků zpracování dat nezůstane opominut. Tímto by měl být zajištěn i co nejvěrohodnější výsledek a zároveň by měly být eliminovány možné chyby zapříčiněné nesprávným postupem.

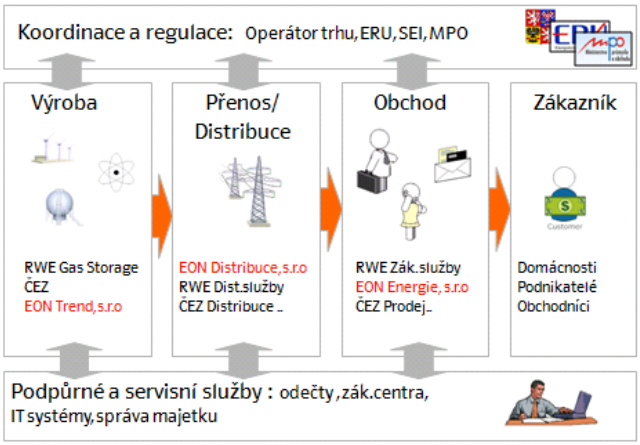
Základní data pro analýzu budou pořízena z informačního systému EON ČR, a to zejména ze zákaznického systému SAP ISU, systému pro řízení vztahů se zákazníky SAP CRM a z datawarehouse systému SAP BW. Z důvodu velkého rozsahu dat, bude nezbytné provést redukci datového souboru na reprezentativní vzorek. Toto bude provedeno jednak redukcí množství proměnných pomocí korelační analýzy, jednak výběrem vzorku se zachováním poměru četností u hlavních proměnných.

Z hlediska statistických metod bude analýza provedena aplikací algoritmu aglomerativního hierarchického shlukování. Jako programové prostředí bude použito statistické prostředí R.

4 Praktická část

4.1 Prostředí utilitního trhu

Ačkoliv řada spotřebitelů stále ještě považuje trh s utility (elektřina, plyn) jako monopol, ve skutečnosti již od roku 2005 mají zákazníci možnost volby výběru svého dodavatele. V rámci tzv. procesu unbundlingu došlo k rozdělení původních monopolních společností do jednotlivých samostatných částí – výroba, distribuce, obchod.



Obrázek 11: Princip utilitního trhu

Výroba – předmětem je výroba elektrické energie popř. těžba plynu. Z pohledu změn na trhu nejsou výrobci jen provozovatelé tepelných, jaderných či vodních elektráren, ale v principu kdokoli, kdo má na střeše fotovoltaický panel či majitelé malých vodních elektráren. K 1.1.2012 bylo registrováno 14 598 výrobců energie a 21 výrobců plynu (zdroj ERU).

Distribuce – jedná se o jedinou část, která si zachovala svůj přirozený monopolní charakter (nemá smysl budovat např. dvoje el. vedení do jedné obce). Cílem je zajistit rozvoj, provoz a údržbu distribuční soustavy (zjednodušeně el.sítě a plynové potrubí), a tím umožnit přenos energie od výrobce k zákazníkovi. Za své služby si distributor účtuje státem regulované poplatky.

Obchod – jedná se o část procesu nejvíce vystavenou konkurenčnímu tlaku. Obchodníkem může být jakýkoliv subjekt, který získá licenci na nákup a prodej energie/plynu. Cílem obchodníka je získat dostatečný počet zákazníků, smluvně si zajistit výrobu energie u výrobce (nebo resp. na burze) a pomocí distributora ji dodat zákazníkovi. To, že se nejedná pouze o teoretickou konkurenci potvrzuje nejvíce skutečnost že k 1.1.2012 bylo na ERU registrováno 353 obchodníků s el. energií a 143 obchodníků s plynem (zdroj ERU).

Abychom demonstrovali dynamičnost změny v oblasti trhu s elektřinou a plynem, uvádíme statistiku vývoje počtu zákazníků jednotlivých dodavatelů

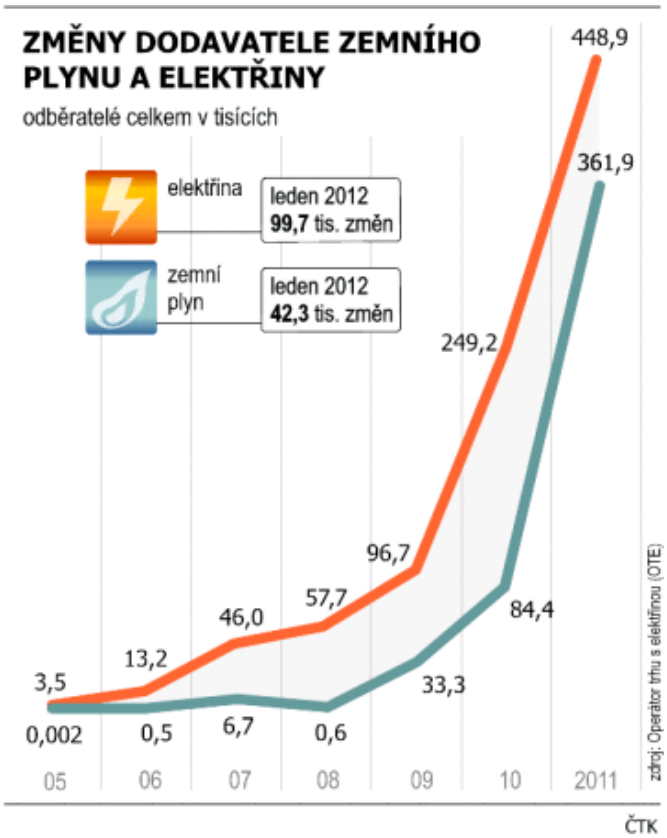
Název dodavatele	I.11	XII.11	Změna
ČEZ Prodej, s.r.o.	3 500 000	3 400 000	-3%
E.ON Energie, a.s.	1 300 000	1 200 000	-8%
Pražská energetika, a.s.	588 000	560 000	-5%
BOHEMIA ENERGY entity s.r.o.	147 191	237 911	62%
CENTROPOL ENERGY, a.s.	103 363	208 218	101%
BICORN s.r.o.	42 710	51 123	20%
RWE Energie, a.s.	12 846	47 563	270%
České Energetické Centrum a.s.	11 105	47 260	326%
United Energy Trading, a.s.	6 939	10 476	51%
VEMEX Energie a.s.	2 420	13 404	454%
Lumen Energy a.s.	2 706	2 982	10%
ARMEX ENERGY, a.s.	1 087	4 165	283%
Nano Energies Trade s.r.o.	1 483	2 129	44%
Europe Easy Energy a.s.	-	2 022	
Optimum Trading, s.r.o.	177	2 522	1325%

Tabulka 4: Vývoj počtu zákazníků jednotlivých dodavatelů el.energie (zdroj OTE)

Název dodavatele	01.2011	12.2011	Změna
RWE Energie, a.s.	2 300 000	2 100 000	-9%
Pražská plynárenská, a.s.	432 790	430 000	-1%
E.ON Energie, a.s.	110 000	120 000	9%
BOHEMIA ENERGY entity s.r.o.	74 720	138 632	86%
ČEZ Prodej, s.r.o.	35 850	131 852	268%
CENTROPOL ENERGY, a.s.	10 432	38 094	265%
České Energetické Centrum a.s.	5 890	29 743	405%
LAMA energy a.s. (Levnyplyn.cz)	5 049	16 704	231%
BICORN s.r.o.		7 548	
VEMEX Energie a.s.	1 457	6 140	321%
České Energetické Centrum Jih s.r.o.		2 877	
X Energie, s.r.o.		2 414	
United Energy Trading, a.s.	1 115	2 069	86%
Gas International s.r.o.		2 068	
ELIMON a.s.		1 699	
HALIMEDES, a.s. (Plynule.cz)	408	1 555	281%
GLOBAL ENERGY, a.s.		1 437	
Optimum Energy, s.r.o.		1 201	
SPP CZ, a.s.	419	779	86%
ARMEX ENERGY, a.s.		521	
Europe Easy Energy a.s.		198	
ENRA SERVICES s.r.o.		160	

Tabulka 5: Vývoj počtu zákazníků jednotlivých dodavatelů plynu (zdroj OTE)

Uvedené tabulky pouze zobrazují stavy na začátku a na konci roku. Dle informací od operátora trhu (OTE) se počet změn dodavatelů energie za rok 2011 blížil 450 tis a počet změn dodavatelů plynu 362 tis zákazníkům. Vzhledem k celkovému počtu všech zákazníků se jedná téměř 6,4 resp. 12 procent.



Obrázek 12: Vývoj změn dodavatelů plynu a elektřiny (Zdroj ČTK)

Z výše uvedeného je patrné, že se utilitní trh výrazně mění. **Základním cílem každé společnosti zabývající se obchodem, je získání a udržení zákazníka.** Pomineme-li absolutní výši spotřeby, není nutné rozlišovat jednotlivé zákazníky – tj. i zákazník s nízkou spotřebou, pokud řádně platí zálohy a faktury, je pro společnost stejně žádoucí jako zákazník velký[1]. V rámci této diplomové práce se budeme nadále zaměřovat na trh s elektrickou energií v segmentu domácností.

[1] Nezabýváme se velkoodběrateli, kde platí jiné podmínky

4.2 Definování cílů projektu

Jak bylo definováno v předchozí kapitole, základním cílem každé utilitní obchodní společnosti je **získání zákazníka** a co **nejdelší udržení zákazníka**. Nyní se na tento cíl zaměříme trochu podrobněji. Samotný trh s elektrickou energií je v podstatě trhem komoditním. Jako komoditu definujeme zboží, které je na trhu obchodováno bez rozdílů v kvalitě. Dodávky různých dodavatelů jsou na trhu vzájemně zastupitelné. Pomineme-li výpadky elektřiny, které jsou stejně v zodpovědnosti distribuce a ne obchodu, je pro každého zákazníka elektřina jednotným produktem. Jakým způsobem lze tedy působit na zákazníka, abychom zajistili jeho získání a následně jeho loajalitu?

První věcí, která každého pravděpodobně napadne, je odlišit se prostřednictvím **ceny**. Je naprosto zřejmé, že cena bude jedním z hlavních faktorů, podle kterého se zákazník bude rozhodovat. Konkurovat samotnou cenou však není tak jednoduché – smlouvy s výrobci se uzavírají na dlouhou dobu dopředu, je nutné správně odhadnou budoucí spotřebu zákazníka, samotná cena energie se každý den mění dle cen vstupů a poptávky na trhu. Operativní změny cen nejsou v principu možné. Nicméně jak lze např. vypozorovat na příkladu mobilních operátorů, cena není jediným faktorem podle kterého se zákazníci orientují. Podobně jako se snaží mobilní operátoři získat své zákazníky pomocí různých dodatečných služeb (mobil, internet zdarma, volání zdarma, atd.), snaží se jednotliví obchodníci v rámci otevřeného trhu s elektřinou získat zákazníky nabídkou různě **konfigurovatelných cenových produktů** a prostřednictvím **dodatečných služeb**.

V rámci společnosti EON Energie, na které je tato analýza prováděna, se společnost snaží získat/udržet zákazníky pomocí:

- Nabídky diferenciovaných produktových řad, kdy kromě vlastní ceny za elektřinu volí zákazník produkt dle svých preferencí
    - Preference zachovat si nezávislost
    - Sleva při určité délce kontraktu
    - Body do bonusových programů
    - Preference ekologické energie
    - Garance ceny na určitou délku období
- Prezentace EONu jako spolehlivého partnera
    - Komunikace se zákazníkem prostřednictvím různých komunikačních kanálů
    - Samoobslužný zákaznický portál → úspora času
    - Sledování kvality služeb (reklamace, spokojenost, atd.)
    - Společenská odpovědnost a veřejné aktivity (vzdělávací programy, sponzoring, podpora energeticky úsporných projektů, rodinné soutěže)
- Nabídka dodatečných služeb
    - Bonusový program
    - Slevy na partnerských obchodních sítích (kasa.cz)

- Příspěvek na stěhování
- Partnerské programy VZP a Tesco Club Card
- Poradenství v oblasti el. energie prostřednictvím aliančních partnerů
- Příspěvky na tepelné čerpadlo
- Servisní služby

Důvody preferencí, které daný zákazník má, závisí na řadě faktorů – příjem, velikost spotřeby energie, geografická kritéria, kulturní a sociální prostředí ve kterém žije, vzdělání apod.

Motivace projektu: Pochopení vzájemných souvislostí uvedených výše, pomůže k lepšímu zacílení trhu při přípravě budoucích marketinkových kampaní či přípravě nových produktů. Ekonomickým výsledkem pak může být jednak snížení nákladů na marketinkové kampaně, jednak zvýšení „response rate“, tedy počtu zákazníků, které daná kampaň zaujme

Cíl projektu: Shromáždit dostupná data o zákaznících společnosti EON Energie a pomocí shlukové analýzy provést jejich rozdělení do jednotlivých shluků. Následnou analýzou identifikovat odlišnosti jednotlivých shluků.

Rizika: Jedním z hlavních rizik je, že se pomocí dostupných dat nepodaří identifikovat smysluplné shluky. Jak bylo zmíněno v teoretické části, předpokladem využitelnosti výsledků shlukové analýzy je existence přirozených shluků v reálném světě. Pokud tento předpoklad není splněn, výstupy shlukové analýzy nemusejí mít žádné komerční uplatnění.

Druhé riziko projektu je spojeno s požadavky na výkon počítačového systémy a software R. Z hlediska vysokého počtu pozorování (1,5 mil) a vysokého počtu proměnných (10-100) bude nezbytné provést výběry z celkového statistického souboru.

4.3 Porozumění datům

V dalším kroku se, dle metodiky CRISP, zaměříme na analýzu dostupných datových zdrojů, identifikaci jednotlivých proměnných, definování strategie pro jejich výběr a ověření datové kvality.

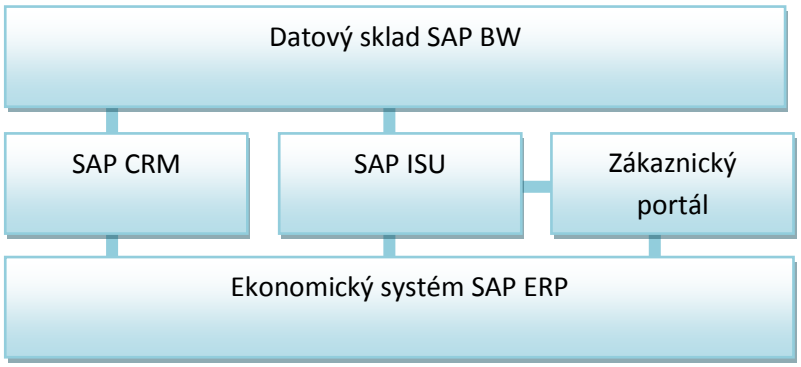
Dostupné datové zdroje je možné rozdělit do dvou základních kategorií:

- Interní data – data z informačního systému EON Energie. Data lze rozdělit jednak z technologického hlediska jejich dostupnosti v jednotlivých částech informačního systému, jednak z hlediska jejich procesně-obchodní problematiky.
- Externí data – data z externích zdrojů, o které je možné rozšířit dostupná data z interního systému. V našem případě jsou dostupná pouze data z ČSU, kterými budou interní data obohacena o geografické a demografické údaje.

4.3.1 Informační systém

Architektura informačního systému EON Energie je tvořena převážně SW produkty společností SAP. Z hlediska jednotlivých komponent je tvořena systémy:

- Systém pro řízení vztahu se zákazníky SAP CRM
- Zákaznický systém SAP ISU
- Datový sklad SAP BW
- Ekonomickým systémem SAP ERP
- Zákaznickým portálem



Obrázek 13: Architektura inf. systému

Systém pro řízení vztahu se zákazníky SAP CRM – slouží jako tzv.“ frontend“ systém, tj. systém ve kterém se zaznamenává jakákoliv přímá iterace se zákazníkem. Z hlediska funkcionalit obsahuje informace o zákaznících, smlouvách, produktech, komunikaci se zákazníkem – žádosti, stížnosti, nabídky, poskytnuté služby. Systém zároveň podporuje různé procesy – marketinkové kampaně, připojení odběratele, odpojení odběratele.

Zákaznický systém SAP ISU – je srdcem obchodního zákaznického systému. V rámci tohoto systému jsou evidovány odběrná místa, spotřeba zákazníků, řízení zálohování, fakturace, upomínkování.

Zákaznický portál – je systém rozšiřující základní služby o funkcionalitu samoobslužného portálu a nabídku různých bonusových programů.

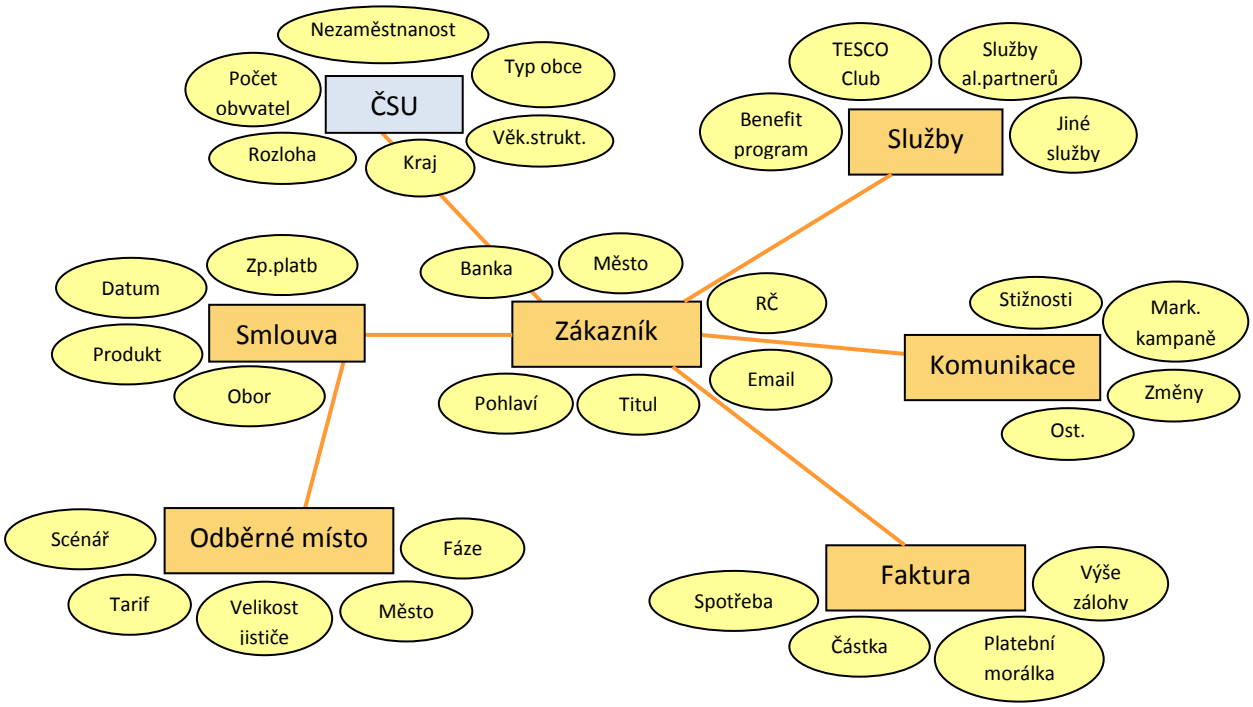
Datový sklad SAP BW – datový sklad slouží pro konsolidaci dat pocházejících z různých datových zdrojů a jejich dlouhodobé uchovávání. Nad daty se provádí řada analýz od statického reportingu, ad-hoc analýz, přípravy dat pro marketinkové kampaně, až po úlohy pro podporu manažerského rozhodování.

4.3.2 Externí datové zdroje

Z hlediska možného doplnění interních dat o externí data, byla identifikována pouze možnost rozšířit data o informace z ČSU. Na základě adresních informací z interního inf. systému předpokládáme propojení na data ČSU, pro doplnění zejména geografických a socioekonomických informací – velikost města, věková struktura, nezaměstnanost, vybavenost apod.

4.3.3 Analýza datových domén

Nyní se zaměříme na identifikaci hlavních datových domén, které budeme následně analyzovat a vyhodnocovat z hlediska jejich použitelnosti pro provedení shlukové analýzy.



Obrázek 14: Struktura datových domén

V rámci analýzy byly identifikovány tyto hlavní datové domény

- Zákazník – představuje stávajícího nebo bývalého odběratele el. energie.
- Smlouva – reprezentuje vztah mezi zákazníkem a odběrným místem. Prostřednictvím smlouvy dochází k uzavření obchodního vztahu mezi zákazníkem a obchodníkem. Zároveň obsahuje i definici obchodních podmínek, kterými lze částečně určit preference daného zákazníka
- Odběrné místo – reprezentuje technický objekt pro dodávku energie. Obsahuje zejména technické údaje, umožňující vytvoření určité představy o způsobu vybavení domácnosti – velikost jističe, vybavení, předpokládaná spotřeba apod.
- Služby – nereprezentují v pravém slova smyslu samostatnou existující doménu. V rámci projektu budou znalosti o zákazníkovi doplněny informacemi o dodatečných službách, které daný zákazník využívá. Jedná se zejména o členství v bonusových programech, Tesco clubu, VZP apod.
- Komunikace – rozsah a způsob komunikace se zákazníkem může představovat jednu z doplňkových informací o aktivitě/pasivitě daného zákazníka. Jedná se např. o informace o počtu stížností, mark. nabídek, odpovědí na nabídky apod.
- Data ČSU – interní informace o zákazníkovi lze rozšířit o externí informace doplňující znalost o zákazníkovi a o dílčí znalost prostředí, ze kterého zákazník pochází.

4.3.3.1 Zákazník

Objekt „Zákazník“ (nebo také „Business Partner“), představuje základní vazební entitu na ostatní datové objekty. Objekt je možné charakterizovat atributy:

- Číslo zákazníka, identifikuje daného zákazníka
- Kód obce, jeden z identifikačních údajů adresy
- Titul zákazníka
- Rodné číslo zákazníka
- E-mail adresa pro komunikaci
- Telefonní číslo zákazníka
- Kód bankovního spojení

4.3.3.2 Smlouva

Objekt *„Smlouva“* je vazebním prvkem vytvářející obchodní vztah zákazníka a společnosti a propojující technický prvek reprezentovaný odběrným místem a zákazníkem. Z hlediska atributů objektu *„Smlouva“* rozlišujeme:

- Identifikace smlouvy
- Účetní okruh – identifikuje danou společnost. V rámci projektu bude použit kód „4013-EON Energie“
- Produkt CRM.
- Datum přihlášení smlouvy – datum uzavření smlouvy.
- Datum odhlášení smlouvy. Pokud pole není vyplněné vyplňuje se hodnotou 00.00.0000. Pokud je vyplněné, jedná se o ztraceného zákazníka
- Obor reprezentuje komoditu – 01 – Elektřina, 02 – plyn,… V rámci naší analýzy budou vybrána data za elektřinu, tj. s typem oboru 01
- Způsob platby - převod z účtu, složenkou, SIPO, Inkaso

4.3.3.3 Odběrné místo

Objekt „Odběrné místo“ představuje technické místo dodávky elektřiny. V rámci analýzy je popsáno atributy

- Identifikace odběrného místa
- Scénář - udává typ deregulovaného scénáře – získaný zákazník, zákazník se sdruženou smlouvou, ztracený zák..
- Typ sazby – obsahuje jednak informaci o produktu, jednak o typu spotřeby
- Počet fází
- Velikost jističe
- Adresa odb. místa - může si lišit od adresy zákazníka (např. chata)
- Instalovaný příkon
- Historická spotřeba

4.3.3.4 Faktura

Objekt „Faktura“ obsahuje časově proměnlivé informace obsahující ekonomicko-technické informace o chování zákazníka. V rámci analýzy bude seznam zákazníků fakturovaných v období 06.2011 použit jako výchozí vzorek pro analýzu dat. Z hlediska dostupných údajů bude využito:

- Spotřeba el. energie
- Poplatek za obchod
- Poplatek za distribuční část
- Výše zaplacené zálohy

4.3.3.5 Komunikace

Dalším ze zdrojů informací o chování a požadavcích zákazníka jsou data ze zákaznického centra CRM. V rámci dostupných údajů je možné pracovat s atributy:

- Identifikace obchodního partnera
- Kategorie činnosti
- Počet kontaktů

4.3.3.6 Služby

Objekt *„Služby“* představuje souhrnný objekt agregující data o zákaznících z různých zákaznických programů. V rámci analýzy se pokusíme využít data z programů VZP, E24, TescoClub. Z pohledu atributů budou data transformována do atributů:

- Identifikace zákazníka
- Kód příslušného zákaznického programu
- Čítač představující logickou proměnnou Ano/Ne

4.3.3.7 Data ČSU

Data z Českého Statistického Úřadu představují možnost jak interní data obohatit o externí údaje. V rámci zpracování se zaměříme především na údaje:

- Kód obce - jednoznačný identifikátor obce dle číselníku ČSU
- Počet obyvatel
- Nezaměstnanost v %

4.3.4 Průzkum dat

V rámci prvotního průzkumu dat jsme se zaměřili na explorační a deskriptivní analýzu dat. Cílem je identifikovat charakter jednotlivých dat, počty kategorií, vnitřní strukturu dat, identifikace anomálií a úvodní zkoumání datové kvality. Předmětem zájmu jsou zejména prázdné hodnoty, hodnoty vyplněné specifickým znakem (např. #, 9999, xxx) a hodnoty, které z hlediska obchodní logiky nepatří do požadované kategorie.

4.3.4.1 Proměnné s vnitřní strukturou dat

Typ sazby

Typ sazby je jedním z hlavních identifikátorů potřeb a preferencí zákazníka. Proměnná je tvořena 9-ti místným kódem, který obsahuje:

Pozice znaku	Význam
1	Obor (E – elektřina, P – Plyn)
2	Společnost (O – Obchod, D – Distribuce)
3-4	Segment (DO – domácnosti, PO – Podnikatelé)
5	Oddělovač – pomlčka
6 – 7	Produkt • 01.. Klasik – základní nabídka bez závazku doby kontraktu • 02 .. Jistota • 03 .. Trend – zákazníci preferující stabilní ceny na fix.období • 04 .. Eko – zákazníci preferující ekologické zdroje energie • 05 .. @ produkt – zákazníci komunikující přes Interent
8 – 9	Charakter sazby • 01 .. Klasik – běžná spotřeba • 02 .. Aku – pro zákazníky využívající akumulační vytápění/ohřev vody • 03 .. Kombi – zákazníci s hybridními systémy pro vytápění a ohřev vody • 04 .. Přímotop – pro zákazníky využívající přímotopné vytápění • 05 .. Víkend – pro zákazníky s víkendovou spotřebou (např. chaty)

Tabulka 6: Struktura identifikátoru "Typ sazby"

Scénář

Scénář zařazuje daného zákazníka do určité kategorie.

Kód	Význam
101	Stálí zákazní – aktuální zákazníci na území EON
102	Ztracený zákazník – zákazníci, kteří ukončili smlouvu s EON
103	Získaný zákazník – získaní zákazníci z jiných distribučních území

Tabulka 7: Význam kódu "Scénář"

Způsob došlé platby

Způsob jakým zákazníci hradí platby za elektrickou energii

Kód	Význam
B	převod z účtu
E	SIPO
I	Inkaso
P	poštovní poukázkou

Tabulka 8: Význam kódu "Způsob došlé platby"

Rodné číslo

Jednoznačný 11-ti místný kód, identifikující daného zákazníka. V rámci naší analýzy bude použita vnitřní struktura rodného čísla pro odvození věku a pohlaví zákazníka

Pozice znaku	Význam
1-2	Rok narození
3-4	Měsíc narození, v případě žen se k měsíci přidává +50. Př.: 07 – červenec, muž, 57 – červenec, žena

Tabulka 9: Vnitřní struktura "Rodného čísla"

4.3.4.2 Analýza kvality dat

Vstupním datovým souborem, byl soubor o rozsahu 108 729 záznamů. Tento soubor byl podroben prvotní datové analýze s cílem identifikace problematických záznamů. Na základě této analýzy vyplynuly tyto závěry:

Proměnná	Anomálie	Detail	Počet výskytů
Typ sazby	Výskyt typů sazeb jiného segmentu	EOPO* EX01	16232
Phone	Chybné číslo	Pouze +4200	824
	Nevyplněná hodnota	‘ ’	35886
Rodné číslo	Nevyplněná hodnota	‘ ’	25054
	Chybná hodnota	DRXXXIV 28	22
Email	Nevyplněná hodnota	‘ ’	91 094
SIPO	Nevyplněná hodnota	‘ ’	40 939
Titul	Nevyplněná hodnota	‘ ’	98 093
Inst.příkon	Nevyplněná hodnota	‘ ’	73 732
Hist. NT	Nevyplněná hodnota	‘ ’	76 090
Hist VT	Nevyplněná hodnota	‘ ’	37 540
Fáze	Nevyplněná hodnota	‘ ’	1
Kód obce	Nevyplněná hodnota	‘ ’	14 896
Jistič	Nevyplněná hodnota		1
Datum do	Hodnota reprezentující platnou smlouvu	00-00-0000	
Scénář	Nevyplněná hodnota		1

Tabulka 10: Výsledky analýzy kvality dat

4.4 Příprava dat

Na základě výsledků identifikace dostupných datových zdrojů, znalostí vnitřní struktury dat a výsledků datové kvality se v rámci dalšího zpracování zaměříme na přípravu vhodného datového souboru, na kterém bude následně provedena analýza. Příprava dle metodiky CRISP obsahuje výběr dat, čištění, doplnění, integraci a konsolidaci a formátování do podoby vhodné pro následné zpracování.

4.4.1 Výběr dat

Za základní datové soubory byly zvoleny

- **Data ze zákaznického systému ISU** – informace o zákaznících, smlouvách, spotřebě fakturaci. Vzorek byl omezen na fakturaci za období 06.2011 a segment „Maloodběr – domácnosti (MOO)“. Vzorek obsahuje 108 729 záznamů. Z hlediska celkového objemu zákazníků za segment MOO, vzorek představuje cca. 1/12 celkového objemu zákazníků.
- **Data z kontaktního centra CRM** – informace o aktivitách a kontaktech zákazníku. Vzorek obsahuje 147 849 záznamů. Vzorek není časově ohraničen.
- **Data z ČSÚ** – informace o městech a obcích. Vzorek obsahuje 6 250 záznamů
- **Vazební tabulka SAP - ČSU**- pro určení vazby mezi kódy obcí SAP a ČSU byla ze systému ISU vytvořena vazební tabulka.
- **Data ze zákaznických programů** – data z programů Energie24, TescoClub, VZP

Vyjma dat ČSU byla veškeré data pořízena ze systému SAP BW prostřednictvím reportů vytvořených v SAP BW BEX Analyzeru a přímo ze systému SAP ISU. Výsledné datové soubory byly importovány do databáze MySQL.

4.4.2 Čištění dat

Na základě analýzy datové kvality provedené v předchozí etapě, byly provedeny následující úpravy

1) **Odstranění chybných nebo nekompletních záznamů** – bylo provedeno u záznamů, kdy by absence nebo chybná hodnota měla zásadní vliv na výsledky analýzy. Takto byly odstraněny záznamy s chybným typem sazby, chybějícím rodným číslem, chybějícím kódem obce, chybnou hodnotou fáze a jističe a chybějícím způsobem platby.
2) **Vyloučení vybraných proměnných z analýzy** – proměnné s vysokou četností nevyplněných hodnot nebo proměnné s nevěrohodnými údaji byly vyloučeny z dalšího zpracování dat. Jedná se o proměnné – historie VT, historie NT, email, titul, instalovaný příkon.
3) **Konverze anomálních hodnot** – proměnné obsahující nekompletní datové záznamy popř. jiné anomální záznamy byly zkonvertovány na iniciální hodnotu. Jednalo se konverzi tel. čísla, rodného čísla.

Po výše provedených úpravách byl **původního vzorek 108 729** záznamů **zredukován** na **62 625 záznamů**. Tato poměrně významná redukce byla naprosto nezbytná, abychom zabezpečili požadovanou kvalitou dat.

4.4.3 Doplnění dat

V rámci dalšího kroku jsme se zaměřili na doplnění datového souboru o proměnné, které byly možné odvodit na základě původních dat.

Proměnná	Zdroj dat	Popis
Produkt	Typ sazby, 6-7 znak	Informace o produktu zákazníka
Charakter spotřeby	Typ sazby, 8-9 znak	Informace o typu domácnosti
Věk	Rodné číslo, 1-2 znak. Provedena konverze na věk	Věk zákazníka
Pohlaví	Rodné číslo, 3-4 znak, hodnoty < 50 → muž, hodnoty > 50 → žena	Pohlaví zákazníka
Příznak – akt. zákazník	Pokud Datum do = 00.00.0000 příznak 1 jinak 0	Příznak platného zákazníka
Příznak – telefon	Pokud Phone = ‘ ’ příznak 0 jinak 1	Příznak zákazníka, který uvedl tel. Kontakt
Příznak – platba SIPO	Pokud SIPO = ‘ ’ příznak 0 jinak 1	Příznak zákazníka platícího prostřednictvím SIPO
Příznak – CRM aktivní	Pokud Počet > 0 příznak 1 jinak 0	Příznak „aktivního“ zákazníka, tj. zákazníka, který byl za poslední rok v kontaktu s call centrem EON

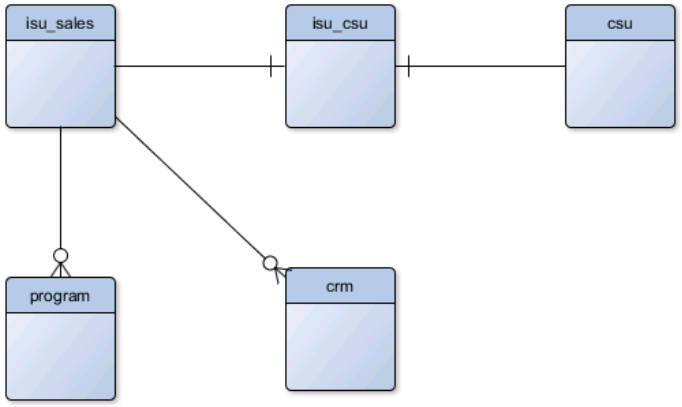
Tabulka 11: Odvozené proměnné

4.4.4 Integrace a konsolidace

V dalším kroku jsme provedli integraci datového souboru ISU na ostatní zdroje dat – data ČSU data z CRM a data z partnerských programů. Pro integraci byly použity tabulky:

- isu_sales – kmenová data a fakturační data zákaznického systému
- isu_csu – vazební tabulka mezi číselníkem měst zák. systému a daty ČSU
- csu – data z Českého Statistického Úřadu o velikostech měst a obcí
- program – data o zákaznících partnerských a bonusových programů
- crm – data ze systému CRM o komunikaci zákazníků

Propojením těchto předpřipravených dat jsme vytvořili vlastní obchodní model analyzované oblasti, představující relační vztahy mezi jednotlivými zdroji dat:



Obrázek 15: Relační datový model analyzovaných dat

Integrace byla provedena pomocí SQL jazyka v prostředí databáze MySQL. Výsledek dotazu byl uložen do csv souboru jako základ pro další práci v jazyce R.

SQL kód propojující zdrojová data:

```
SELECT isu_sales.*, csu.*, crm.*, program.*
FROM isu_sales
JOIN sap_csu ON isu_sales.BP_CSA = sap_csu.SAPID
JOIN csu on ON _csu.CSUID = csu.kodob
LEFT JOIN crm ON isu_sales.BP = crm.BP
LEFT JOIN program ON isu_sales.BP = program.BP
```

Výsledné propojení dat neposkytlo žádné hodnoty z tabulky zákaznického programu (tabulka „program“, obr. 15), proto se hodnoty ze zákaznického programu v dalším zpracování nevyskytují.

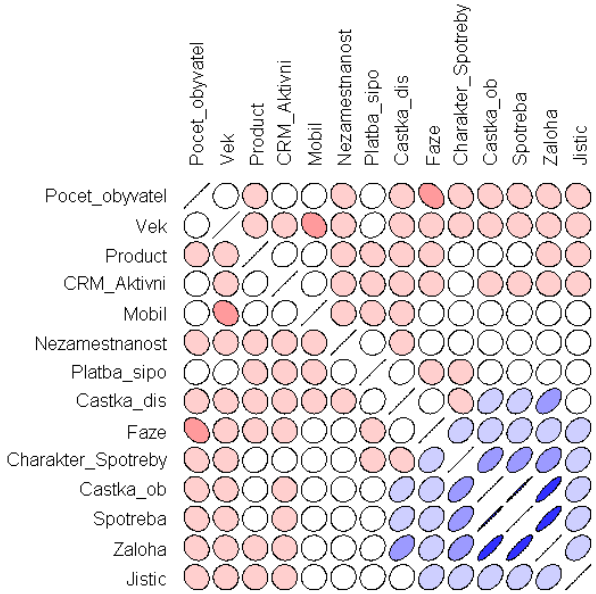
Vzhledem k stále velkému rozsahu datového souboru, byl následně v jazyce R, prostřednictvím metody *Stratified Sampling*, vybrán vzorek dat odpovídající poměrnému zastoupení četnosti vybraných proměnných v původním datovém souboru.

Za vybrané proměnné byly zvoleny proměnné „Produkt“, „Charakter spotřeby“ a „Počet obyvatel“. Prostřednictvím metody byly postupně zpracovány tři výsledné soubory dat, které byly následně spojeny do vzorku o 1827 záznamů.

```
>Dataset  <-  read.table("C:/Data/KXEN/DP/result_all_62k_v2.csv",
header=TRUE,
 +   sep=";", na.strings="NA", dec=".", strip.white=TRUE)
> d1<-stratified(Dataset,1,9,.01) # Vzorek Produkt
> d2<-stratified(Dataset,1,10,.01) #Vzorek „Charakter spotřeby“
> d3<-stratified(Dataset,1,16,.01)  # Vzorek „Počet obyvatel“
> Result<-rbind(d1,d2,d3)
> nrow(Result)  # Výsledný počet záznamů vzorkování dat:
  [1] 1827
```

4.4.5 Ověření korelace proměnných

Aby nedošlo k přílišnému zkreslení výsledků analýzy, je důležité ověřit, zda nejsou jednotlivé proměnné příliš korelované. Případnou eliminací korelovaných proměnných je zároveň možné snížit požadavky na výpočetní kapacitu. Závislost resp. nezávislost zvolených proměnných jsme ověřili prostřednictvím korelační analýzy. Výsledek analýzy zobrazuje následující graf. Úroveň korelace je zobrazena elipsou, červená barva zobrazuje negativní korelaci, modrá pozitivní. Z výsledků je patrná pozitivní korelace mezi spotřebou, částkou faktury a výši zálohy.



Obrázek 16: Výsledky korelační analýzy

4.4.6 Formátování dat

Pro účely dalšího zpracování dat jsme provedli přeformátování proměnných Fáze, Jistič, Mobil, Aktivita CRM a Platba_SIPO na kategoriální proměnné.

Data byla přeformátována přímo v jazyku R, pomocní příkazu „as.fakctor“:

```
> Dataset <- read.table("C:/Data/KXEN/DP/result_all_62k_v2.csv", header=TRUE,
+   sep=";", na.strings="NA", dec=".", strip.white=TRUE)
> Dataset$CRM_Aktivni <- as.factor(Dataset$CRM_Aktivni)
> Dataset$Faze <- as.factor(Dataset$Faze)
> Dataset$Charakter_Spotreby <- as.factor(Dataset$Charakter_Spotreby)
> Dataset$Jistic <- as.factor(Dataset$Jistic)
> Dataset$Mobil <- as.factor(Dataset$Mobil)
> Dataset$Platba_sipo <- as.factor(Dataset$Platba_sipo)
> Dataset$Product <- as.factor(Dataset$Product)
> Dataset$Vek <- as.factor(Dataset$Vek)
```

Výsledná data jsou dostupná v prostředí R v datovém objektu s názvem „Dataframe“ a budou použita pro další zpracování.

4.5 Modelování a analýza dat

4.5.1 Shluková analýza pomocí hierarchické metody

Pro analýzu získaného vzorku jsme použili hierarchickou shlukovou metodu. Jako způsob výpočtu vzdálenosti jsme použili Gowerův koeficient výpočtu vzdáleností, který umožňuje použití i kvalitativních dat. Pro samotné shlukování byl použit Wardův algoritmus, který v rámci shlukové analýzy poskytoval nejlepší výsledky.

Výpočet vzdálenosti pomocí Gowerova koeficientu

```
> vzdalenost <-daisy(Dataset,stand=TRUE)
```

Provedení shlukové analýzy

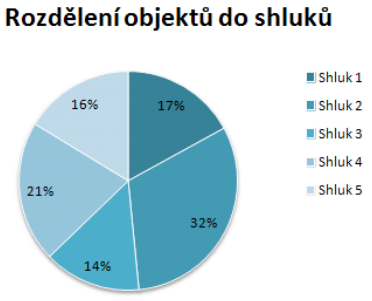
```
> model<-hclust(vzdalenost,method="ward")
```

Spojení výsledků s původním objektem *„Dataset“* a výpis četností

```
> Result<-cbind(Dataset, Cluster=cutree(model,k=5))
> table(Result$Cluster)
  1   2   3   4   5                    # Výsledky četností jednotlivých shluků
311 574 263 381 298
```

Výsledkem modelu bylo vytvoření pěti shluků s následujícím rozdělením do jednotlivých objektů:

Shluk	Počet	Rel.
Shluk 1	311	17%
Shluk 2	574	31%
Shluk 3	263	14%
Shluk 4	381	21%
Shluk 5	298	16%
Celkem	**1827**	**100%**



Obrázek 17: Tabulka a graf rozdělení četností v jednotlivých shlucích

Z tabulky je patrné poměrně rovnoměrné rozdělení jednotlivých objektů mezi výsledné shluky s určitou převahou ve druhém shluku.

Rozložení výsledných shluků můžeme analyzovat i graficky prostřednictvím zobrazení v hierarchickém dendogramu. Z něj je patrné symetrické rozdělené jednotlivých objektů do výsledných shluků.

Vykreslení dendogramu provedeme opět pomocí jazyka R s rozdělením na úrovni pěti shluků:

```
> plot(model,xlab="objekty",sub="",main="Hierarchické shlukování")
> rect.hclust(model, k=5, border="red")
```

Obrázek 18: Hierarchický dendogram

Na závěr byl výsledek shlukové analýzy uložen do souboru *Result_hclust.csv*

```
> write.csv(Result, file="Result_hclust.csv")
```

Pro účely detailního rozboru výsledků jednotlivých proměnných již budeme pracovat pouze s tímto souborem výsledků.

4.5.2 Analýza výsledků

Po vytvoření modelu se dále budeme zabývat výsledky hierarchického shlukování z hlediska rozdělení jednotlivých proměnných v rámci daného shluku.

4.5.2.1 Produkt

Proměnná produkt umožňuje členit zákazníky dle jejich osobních preferencí. Zákazníci prostřednictvím výběru daného produktu mohou zvolit mezi možností kdykoliv změnit dodavatele elektřiny (smlouva bez časového závazku), jistotou fixní ceny bez ohledu na vývoj trhu (smlouva se závazkem na 2 roky), preferencí ekologické energie z obnovitelných zdrojů (dražší energie) nebo slevou při změně komunikace se společností na komunikaci prostřednictvím internetu.

Pomocí R jsme provedli výpis základního rozdělení četností proměnné v daném shluku[1]:

```
> by(Dataset$Product,Dataset$Cluster,table)          #kód v jazyce R
```

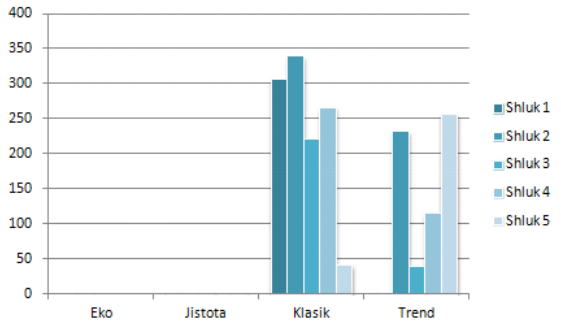
Následující obrázek již zobrazuje upravené tabulkové a grafické zobrazení výsledků v rámci jednotlivých shluků:

Produkt	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
Eko		0,2%			
Jistota	1,0%	0,3%	0,8%		0,3%
Klasik	98,4%	59,1%	84,0%	69,8%	13,8%
Trend	0,6%	40,4%	15,2%	30,2%	85,9%
Celkem	100,0%	100,0%	100,0%	100,0%	100,0%

% zastoupení produktu v daném shluku

Produkt	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
Eko		100,0%			
Jistota	37,5%	25,0%	25,0%		12,5%
Klasik	26,1%	28,9%	18,8%	22,7%	3,5%
Trend	0,3%	36,0%	6,2%	17,8%	39,7%
Celkem	17,0%	31,4%	14,4%	20,9%	16,3%

% zastoupení shluku v daném produktu



Obrázek 19: Výsledky analýzy proměnné „Produkt“

Z analýzy výsledků je patrné, že i přes rozmanitou produktovou řadu zákazníci dávají přednost zejména produktům z řad Klasik a Trend. Jedná se o dvě základní skupiny rozdělující zákazníky na skupinu se závazkem trvání smlouvy (Trend) a na zákazníky s možností okamžité změny dodavatele (Klasik)

[1] Vzhledem k analogii kódu při analýze dalších proměnných již v dalším textu kód neuvádíme

Z hlediska odlišností je možné pozorovat:

- Produkty Eko a Jistota mají velmi nízkou četnost, produkt @ se pro nízkou četnost do stratifikovaného vzorku ani nedostal
- Shluk 1 a 3 je zejména reprezentován zákazníky produktu Klasik
- Shluk 5 je reprezentován zákazníky produktu Trend

4.5.2.2 Charakter spotřeby

Charakter spotřeby vypovídá o způsobu použití el. energie v dané domácnosti. Obsahuje informaci, zda domácnost používá elektřinu jako primární zdroj vytápění a ohřevu vody (tarify Aku, Kombi, Přímotop), nebo zda elektřina slouží k běžnému provozu (Klasik). Zároveň je možné odvodit, jakým způsobem je elektřina používána (akumulační ohřev, vytápění formou přímotopu, nárazově apod.).

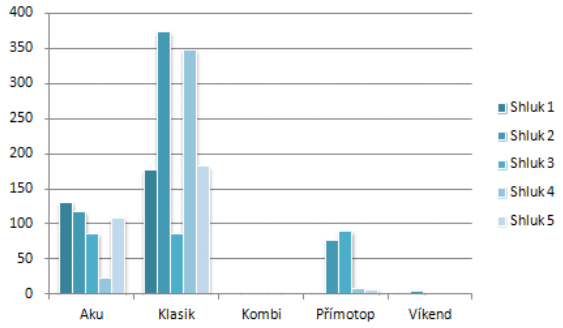
Zjištěné údaje uvádíme ve dvou tabulkách, zobrazujících jak rozdělení dat v daném shluku, tak zastoupení daného shluku v rámci jednotlivých typů spotřeby.

Charakter spotř.	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
Aku	42%	21%	33%	6%	37%
Klasik	57%	65%	32%	91%	61%
Kombi		0%	1%	1%	
Přímotop	0%	13%	34%	2%	2%
Víkend	1%	1%			
Celkem	100%	100%	100%	100%	100%

% zastoupení charakt.spotřeby v daném shluku

Charakter spotř.	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
Aku	28,1%	25,3%	18,5%	4,7%	23,4%
Klasik	15,2%	32,0%	7,3%	29,9%	15,6%
Kombi		16,7%	33,3%	50,0%	
Přímotop	0,5%	42,1%	49,2%	4,4%	3,8%
Víkend	28,6%	71,4%			
Celkem	17,0%	31,4%	14,4%	20,9%	16,3%

% zastoupení shluku dle charakteru spotřeby



Obrázek 20: Výsledky analýzy proměnné "Charakter spotřeby"

Z analýzy vyplývá:

- Většina domácností používá el. energii pouze k běžnému provozu bez vytápění (68%). Z hlediska způsobu vytápění dominují spotřebiče s akumulační sazbou (71%) oproti přímotopům (28%)
- Shluk 2 a 4 je převážně zastoupen zákazníky s typem spotřeby Klasik
- Typ spotřeby „Přímotop“ je převážně reprezentován ve shlucích 2 a 3
- Typ spotřeby „Víkend“ je převážně reprezentován ve shluku 2

4.5.2.3 Jistič

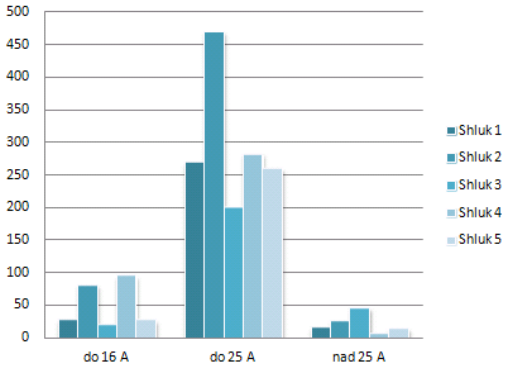
Vzhledem k velikému počtu kategorií a malé četnosti v jednotlivých kategoriích u proměnné Jistič, jsme nejprve provedli agregaci výsledků do skupin „do 16A“, „do 25A“ a „nad 25A“. Následně byly vypočteny podíly kategorií jističů v jednotlivých skupinách (viz obr. 21):

Jistič	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
do 16 A	11,0%	32,5%	7,7%	38,2%	10,6%
do 25 A	18,2%	31,7%	13,5%	19,1%	17,5%
nad 25 A	14,6%	24,3%	43,7%	4,9%	12,6%
Celkem	17,0%	31,4%	14,4%	20,9%	16,3%

% zastoupení shluků v dané kategorii jističů

Jistič	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
do 16 A	8,7%	13,9%	7,2%	24,7%	8,7%
do 25 A	86,5%	81,7%	75,7%	74,0%	86,9%
nad 25 A	4,8%	4,4%	17,1%	1,3%	4,4%
Celkem	100,0%	100,0%	100,0%	100,0%	100,0%

% zastoupení kategorie jističů v daném shluku



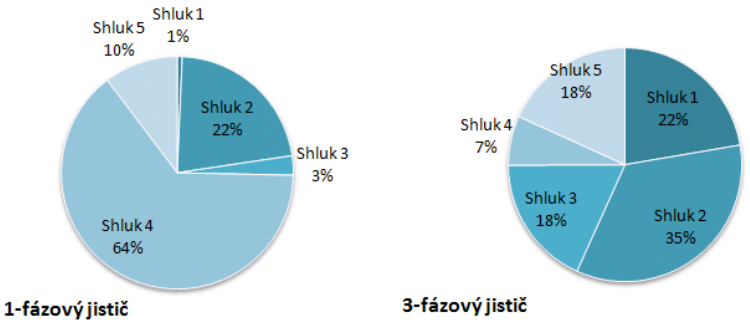
Obrázek 21: Výsledky analýzy proměnné "Jistič"

Z rozdělení četnosti je patrné, že ve všech vzorcích převládají převážně domácnosti s velikostí jističe do 25A. Z hlediska rozdělení do jednotlivých shluků, je možné pozorovat odlišnosti:

- v segmentu do 16A převládají shluky 2 a 4
- v segmentu do 25A převládá shluk 2,
- v segmentu nad 25A převažují zákazníci ve shluku 3
- shluky 1 a 5 jsou zastoupeny takřka výhradně v segmentu do 25A

4.5.2.4 Fáze

Proměnné fáze je tvořena především domácnostmi s tří-fázovým jističem. Výjimku tvoří kategorie jedno-fázových odběrů, které jsou reprezentovány převážně ve 4. shluku:



Počet fází	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5	Celkem za fázi
1	3	98	12	287	46	446
3	308	476	251	94	252	1381
Celkem	311	574	263	381	298	1827

Obrázek 22: Výsledky analýzy proměnné "Fáze"

Na základě výsledku lze tedy usuzovat, že shluk 4 budou tvořit převážně domácnosti v panelových bytech městských domů s nízkou spotřebou.

4.5.2.5 Spotřeba

Proměnná „Spotřeba“ představuje kvantitativní proměnnou. Na rozdíl od kvalitativních proměnných, které jsme doposud analyzovali prostřednictvím četnosti dané kategorie v jednotlivých shlucích, kvantitativní proměnnou můžeme analyzovat prostřednictvím popisných statistik vyjádřených minimální a maximální hodnotou, průměrem, mediánem a 25% a 75% kvantily. Na základě průměrných hodnot můžeme stanovit, zda se spotřeba v daném shluku pohybuje nad úrovní průměru celého zkoumaného vzorku. Pomocí hodnoty mediánu je možné si udělat představu, zda rozložení hodnot ve shluku je symetrické (hodnota mediánu je podobná hodnotě průměru) nebo asymetrické (hodnota mediánu je odlišná od průměru). Hranice kvantilu umožňují vytvoření představy o rozptylu hodnot v jednotlivých shlucích.

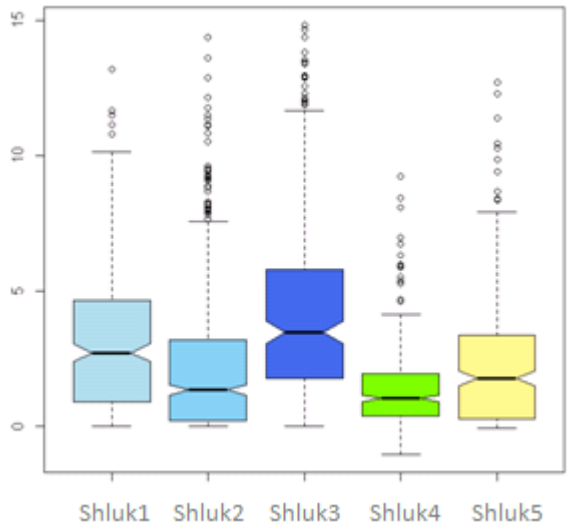
Shluk	Min	1.kvantil	Median	Průměr	3.kvantil	Max
Shluk 1	0	0,9455	2,741	3,17	4,676	18,68
Shluk 2	0	0,2362	1,41	2,531	3,291	25,46
Shluk 3	0	1,862	3,916	5,757	7,338	32
Shluk 4	-1,035	0,405	1,049	1,412	1,964	9,227
Shluk 5	-0,048	0,2785	1,822	2,473	3,408	15,7
Celkem	-1,035	0,463	1,81	2,862	3,777	32

Obrázek 23: Tabulka s výsledky analýzy proměnné "Spotřeba"

Popisnou statistiku můžeme zobrazit i pomocí box plot grafu, ze kterého je zřejmá existence odlehlých hodnot a nadprůměrná spotřeby ve shlucích 1 a 3.

Zpracování prostřednictvím jazyka R:

```
> legenda<-c("Shluk1","Shluk2","Shluk3","Shluk4","Shluk5")
> barvicky<-c('lightblue', 'skyblue', 'RoyalBlue', 'LawnGreen', 'Yellow')
> boxplot(Spotreba~Cluster,data=Dataset, col=barvicky, notch=TRUE, xlab="Spotřeba", names=labels)
```



Obrázek 24: Boxplot diagram proměnné "Spotřeba"

Na základě analýzy můžeme definovat:

- Shluky 1 a 3 obsahují nadprůměrné spotřeby
- Shluk 4 obsahuje mírně podprůměrné spotřeby
- V rámci datového souboru se vyskytuje řada odlehlých hodnot, vybočující z průměrné spotřeby běžných domácností.

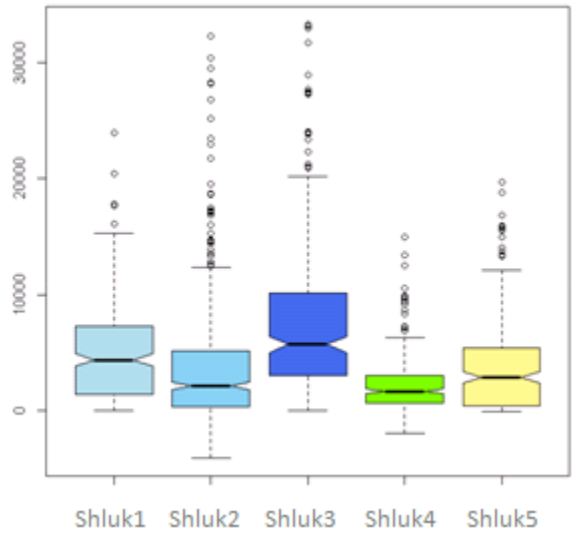
4.5.2.6 Platba obchod

Platba za silovou složku energie je částečně korelovaná s proměnnou spotřeba. Rozdělení platby prezentuje následující tabulka s popisnými statistikami.

Shluk	Min	1.kvantil	Median	Průměr	3.kvantil	Max
Shluk 1	0	1 443	4 356	4 855	7 276	23 930
Shluk 2	-4 089	365	2 187	3 893	5 196	39 880
Shluk 3	0	3 103	5 756	8 936	11 600	51 990
Shluk 4	-1 978	633	1 659	2 189	3 064	14 960
Shluk 5	-78	416	2 873	3 779	5 403	19 700
Celkem	-4 089	710,9	2852	4409	5760	51 990

Obrázek 25: Tabulka s výsledky analýzy proměnné "Platba obchod"

Z box-plot grafu můžeme pozorovat i existenci odlehlých plateb reprezentující zákazníky s vyšší spotřebou:



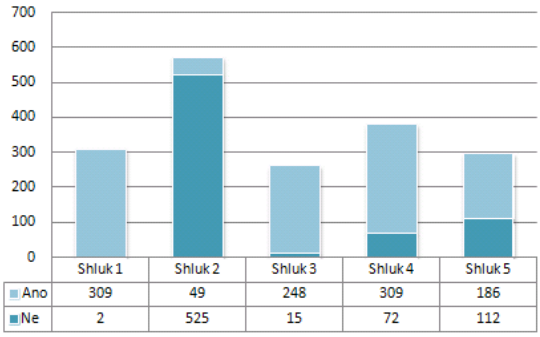
Obrázek 26: Box plot diagram proměnné "Platba obchod"

Závěry, které lze vyvodit z popisných statistik jednotlivých shluků, jsou v podstatě shodné se závěry definované v rámci analýzy shluků předchozí proměnné.

- Shluky 1 a 3 obsahují platby vyšší než celkový průměr
- Shluk 4 obsahuje platby nižší než celkový průměr plateb
- V rámci datového souboru se vyskytuje řada odlehlých hodnot, vybočující z průměru plateb běžných domácností.

4.5.2.7 SIPO

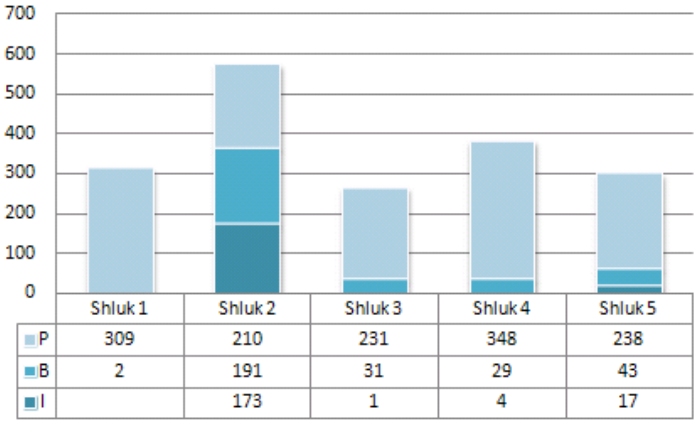
Proměnná SIPO reprezentuje zákazníky, kteří pro platbu účtů používají konsolidovaný způsob pravidelných plateb pomocí SIPO. Tím dochází k úspoře bankovních poplatků. Z grafu je zřejmé, že platba SIPO výrazně převládá ve všech shlucích s výjimkou shluku 2. Převaha jiného způsobu platby ve shluku 2 je natolik výrazná, že může např. reprezentovat vymezenou skupinu zákazníků, které velikost bankovních poplatků „nezajímá“ – např. sociálně silné vrstvy.



Obrázek 27: Graf rozdělení četnosti proměnné "SIPO"

4.5.2.8 Způsob platby

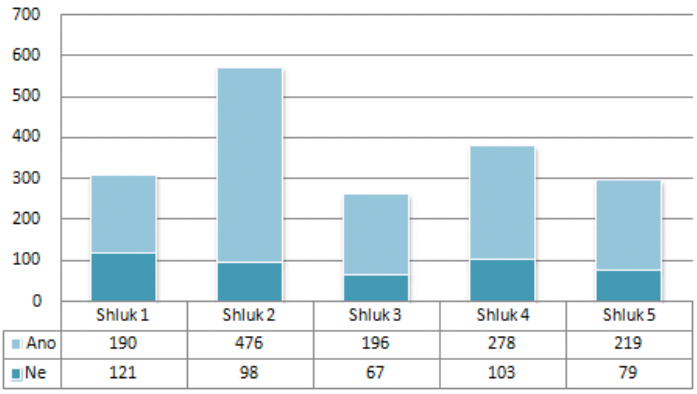
Podobně jako u platby SIPO vykazoval jistou odchylku Shluku 2, tak i v rámci způsobu platby se tento shluk odlišuje od ostatních shluků, kde převládá platba poštovní poukázkou.



Obrázek 28: Graf rozdělení četnosti proměnné "Způsob platby"

4.5.2.9 Mobil

Proměnná mobil analyzuje datový soubor z hlediska ochoty zákazníka uvést i nepovinný osobní údaj – své telefonní číslo. Z analýzy však nevyplývá z hlediska rozdělení shluků žádná odlišnost četností od ostatních.

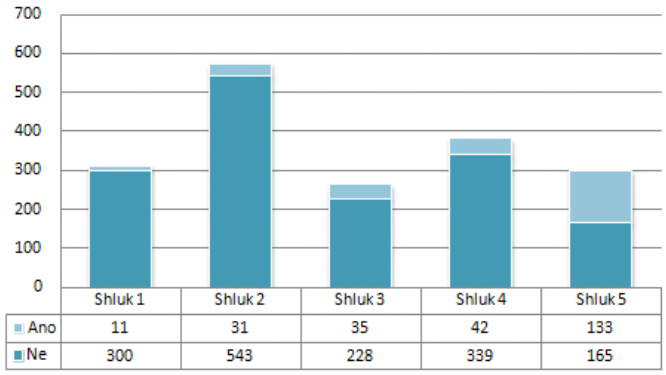


	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
Ano	190	476	196	278	219
Ne	121	98	67	103	79

Obrázek 29: Graf rozdělení četnosti proměnné "Mobil"

4.5.2.10 CRM Aktivita

Aktivita zákazníka je jedním z ukazatelů, který lze použít pro hodnocení intenzity komunikace zákazníka s dodavatelem. Nemusí se jednat pouze o reklamace, ale i o poskytování dodatečných služeb, informací či pomoc při řešení problémů. Celkově četnost aktivních zákazníků vzhledem k neaktivním je pouze 14%. Ačkoliv shluk 5 vykazuje vyšší podíl aktivních zákazníků, může se jednat jen o aktivitu způsobenou přepisem z produktů Klasik do produktu Trend.

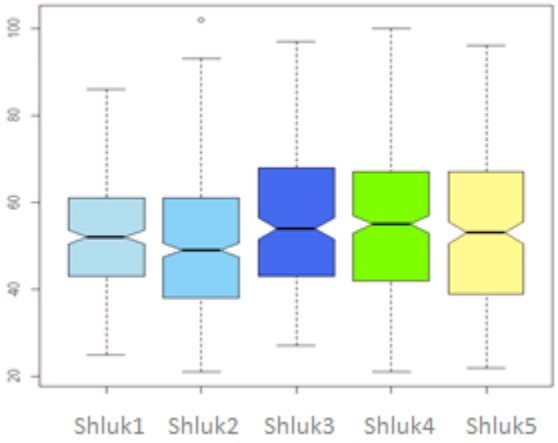


	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
Ano	11	31	35	42	133
Ne	300	543	228	339	165

Obrázek 30: Graf rozdělení četnosti proměnné "CRM Aktivita"

4.5.2.11 Věk

Analýza kvantitativní proměnné věk nezobrazuje žádné výrazné odlišnosti mezi jednotlivými shluky. Rozdíly v průměru jsou natolik minoritní, že z nich nelze vyvodit žádné závěry. Určitým překvapením je poměrně vysoká hodnota průměrného věku zákazníků. Toto může být způsobeno např. těžko dostupným bydlením pro mladé lidi, kteří pak žijí v pronájmech.



Shluk	Min	1.kvantil	Median	Průměr	3.kvantil	Max	Celkem
Shluk 1	25	43	52	52,58	61	86	311
Shluk 2	21	38	49	50,27	61	102	574
Shluk 3	27	43	54	55,87	68	97	263
Shluk 4	21	42	55	55,64	67	100	381
Shluk 5	22	39	53	53,83	67	96	298

Obrázek 31: Popisná statistika a box-plot diagram proměnné "Věk"

4.5.2.12 Nezaměstnanost

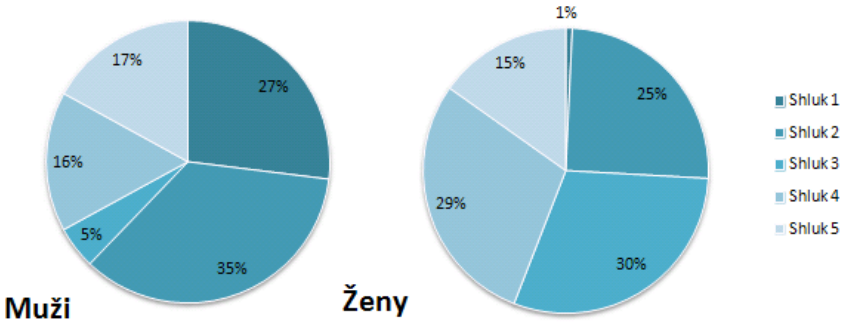
Nezaměstnanost představuje ukazatel pocházející z dat z ČSÚ. Z uvedené tabulky vyplývá, že vliv nezaměstnanosti na jednotlivé shluky je víceméně zanedbatelný s malou výjimkou ve shluku 4, kde je nezaměstnanost nižší než průměrná. Vzhledem k výsledkům předchozí analýzy je toto dáno skutečností, že shluk 4 obsahuje zákazníky převážně z velkých měst, kde je nezaměstnanost přirozeně nižší.

Shluk	Min	1.kvantil	Median	Průměr	3.kvantil	Max
Shluk 1	2,13	7,92	10,84	11,68	15,52	33,33
Shluk 2	2	7,92	10,1	11,4	14,55	33,33
Shluk 3		8,56	11,16	12,21	15,92	27,06
Shluk 4	3,66	9,86	9,86	10,07	9,86	26,39
Shluk 5	3,66	7,92	10,46	11,63	14,85	28,46
Celkem		8,56	9,86	11,32	13,64	33,33

Obrázek 32: Popisná statistika proměnné "Nezaměstnanost"

4.5.2.13 Pohlaví

Kvantitativní proměnná „pohlaví“ je proměnná odvozená na základě rodného čísla. Analýzou můžeme pozorovat, že hodnota proměnné „muži“ je výrazně zastoupena ve shluku 1, a naopak proměnná typu „ženy“ ve shluku 3. V ostatních shlucích je rozdíl v četnostech méně patrný.



Pohlaví	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
Muži	306	400	55	181	193
Ženy	5	174	208	200	105
Celkem	**311**	**574**	**263**	**381**	**298**

Obrázek 33: Zastoupení mužů a žen v jednotlivých shlucích

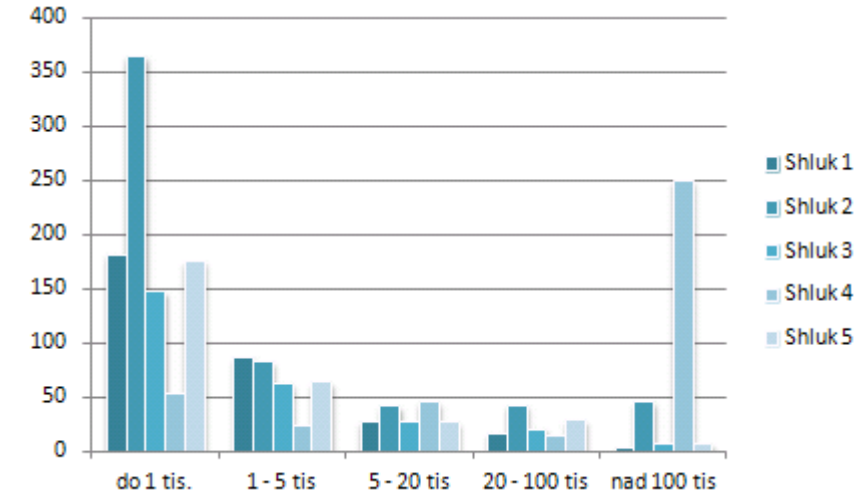
4.5.2.14 Počet obyvatel

Informace o velikosti počtu obyvatel v daném odběrném místě, byla získána na základě propojení obchodních dat zákaznického systému ISU s daty z ČSÚ. Vzhledem k tomu že se jedná o kvantitativní proměnnou, prvotní posouzení provedeme pomocí popisných statistik minima, maxima, průměru, mediánu a 1. a 3. kvantilu.

Shluk	Min	1.kvantil	Median	Průměr	3.kvantil	Max
Shluk 1	42	345	718	4290	1833	371400
Shluk 2	30	347,2	606	32140	3939	371400
Shluk 3	47	396	791	13980	3092	371400
Shluk 4	92	7423	371400	245400	371400	371400
Shluk 5	30	318,2	679	12440	3186	371400
Celkem	-4 089	710,9	2852	4409	5760	51 990

Obrázek 34: Popisná statistiky proměnné "Počet obyvatel"

Pro lepší posouzení rozdělení proměnné „počet obyvatel“ jsme provedli seskupení velikostí měst a obcí do 5-ti skupin:



.

Skupina	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5	Celkem
do 1 tis.	181	364	147	52	175	919
1 - 5 tis	86	82	62	22	63	315
5 - 20 tis	27	41	27	45	26	166
20 - 100 tis	16	42	20	13	28	119
nad 100 tis	1	45	7	249	6	308
Celkem	311	574	263	381	298	1827

Obrázek 35: Graf a tabulka rozdělení četností dle "Počtu obyvatel"

Na základě analýzy popisných statistik můžeme odvodit:

- Rozdělení četnosti počtu obyvatel není symetrické. Toto lze pozorovat jak na grafickém zobrazení, tak na rozdílu medianu a průměru (mean) v popisných statistikách
- Shluky 1,2,3,5 reprezentují převážně vesnice či menší města do 5 tis. obyvatel
- Shluk 2 představují zejména zákazníci v obcích do 1 tis obyvatel
- Shluk 4 je reprezentován převážně obyvateli velkých měst nad 100 tis obyvatel

5 Hodnocení výsledků a závěr

Předchozí kapitoly se zabývaly analýzou odchylek mezi shluky v rámci jednotlivých proměnných. Nyní se zaměříme na interpretaci výsledků dílčích analýz pro účely použitelné v rámci dalších marketinkových aktivit. Následující doporučení uvedená u jednotlivých shluků nelze brát jako 100% platná nebo jako doporučení platná pro celý shluk. Cílem je identifikovat převažující charakteristiky daného shluku, které by mohly vést ke zvýšení prodejů a k lepšímu cílení marketinkových kampaní. V rámci poslední kapitoly uvedeme i doporučení pro implementaci metod shlukové analýzy do reálného prostředí.

5.1 Celkové hodnocení zkoumaného vzorku

Z hlediska celého datového souboru můžeme vyvodit následující závěry:

1) Ačkoliv společnost nabízí pět produktových řad, použitý statistický vzorek, představující jednoměsíční fakturaci, obsahuje převážně produktové řady Klasik a Trend. Řady Jistota, EKO a @ se v podstatě nevyužívají. Je tedy otázkou zda mají tyto produkty z marketinkového pohledu smysl, popř. zda je jejich propagace dělána dostatečně cíleně.
2) Ve vzorku převládá produkt Klasik (64%), což představuje zákazníky, kteří nejsou vázáni žádnou smlouvou a mohou kdykoli změnit dodavatele. Vzhledem k tomu, že nejdůležitějším parametrem je délka kontraktu, převod těchto zákazníků pod jiný produkt by se měl stát jeden z hlavních cílů, na který by se měla společnost dále zaměřit.
3) Ve vzorku se vyskytují zákazníci s odlehlými (extrémními) hodnotami spotřeby, kteří se výrazně odlišují od průměrné spotřeby zkoumaného vzorku. V rámci procesů marketinku by mohlo být přínosné se samostatně zaměřit na tuto skupinu zákazníků.
4) Nabídka zákaznických programů Energie24, VZP a Tesco Club se v rámci použité analýzy neprojevila vůbec. Toto může být zapříčiněno následujícími důvody:
    a. Krátká doba od jejich uvedení – malá četnost dat
    b. Použitý vzorek základních dat (fakturace) je příliš malý
    c. Zákaznické programy se míjejí s požadavky zákazníků

    Aby bylo možné potvrdit nebo vyvrátit jednu z uvedených variant, bylo by nutné provést analýzu cíleně zaměřenou na chování zákazníků, kteří využívají tyto zákaznické programy.

Nyní se zaměříme na interpretaci výsledků vyplývající z rozboru proměnných v rámci jednotlivých shluků.

5.2 Hodnocení jednotlivých shluků

5.2.1 Hodnocení - shluk 1

- Velikost shluku 1 reprezentuje 17% z celého analyzovaného datového souboru.
- Shluk se vyznačuje zákazníky především z ***produktové řady KLASIK (98,4%)***, tedy zákazníky, kteří nevyužili žádnou z možností fixace ceny el. energie.
- Z pohledu charakteru využití energie jsou zde rovnoměrně zastoupeni zákazníci s typem spotřeby Klasik a s typem spotřeby Aku. Podle majoritní kombinace fáze a jističe 3x25A lze usuzovat, že zákazníci využívají elektřinu jako hlavní zdroj pro vytápění.
- Předchozí tvrzení podporuje i proměnná Spotřeba, představující ***nadprůměrnou spotřebu 3,17 MWh***.
- Shluk je tvořen ***převážně mužskou populací,*** žijící převážně v obcích či malých městech ***do 5000 obyvatel*** s mírně nadprůměrnou nezaměstnaností.
- Z hlediska platby převládá platba prostřednictvím SIPO

Doporučení: zákazníci z tohoto segmentu nejsou vázáni dlouhodobou smlouvu se společností – produkt Klasik. Zároveň se jedná o zákazníky s nadprůměrnou výší spotřeby pocházející z ekonomicky slabšího prostředí malých měst a obcí. Můžeme vyslovit hypotézu, že zákazníci této skupiny budou mít tendenci volit takovou dodavatelskou společnost, která jim nabídne nejvýhodnější cenu. Fakt, že nezvolili žádný z produktů nabízející slevy nebo alespoň fixaci ceny, může být dán nedostatečnou informovaností. Toto lze ověřit oslovením zákazníků např. pomocí „door2door[1]“ kampaní.

5.2.2 Hodnocení - shluk 2

- Velikost shluku 2 reprezentuje **největší segment** zákazníků, obsahující 31% z celého analyzovaného souboru.
- Shluk je tvořen zákazníky produktové řady Trend a Klasik, s převahou typu spotřeby „Klasik“(68% shluku 2). Zároveň však **obsahuje i většinu ze segmentu** zákazníků s **typem spotřeby „Víkend“ a „Přímotop“** (71% ze všech zákazníků s typem spotřeby Víkend, 42,1% ze všech zákazníků s typem spotřeby Přímotop).
- Z hlediska spotřeby se jedná o průměrnou spotřebu, čemu odpovídá i poloviční podíl jističů v kategorii 3x16A.
- Z demografického hlediska se jedná **o smíšenou skupinu** zákazníků, **s mírně podprůměrným věkem** žijící **převážně v obcích do 1000 obyvatel**.
- Jistou odchylku představuje silná **preference skupiny k jiným platbám než SIPO (91%)**, což může znamenat ekonomicky silnější zákazníky, kteří se nezajímají o možnosti úspor z bankovních transakcí. Tomuto by odpovídala i četnost sazby Víkend – majitelé chat.

[1] Kampaně, kdy obchodník navštěvuje zákazníka v místě jeho bydliště, doslovně „ode dveří ke dveřím“

Doporučení: představitelé tohoto segmentu budou pravděpodobně ekonomicky stabilní zákazníci v produktivním věku. Největším problémem této skupiny je většinou nedostatek času. Jejich stabilitu proto může např. ovlivnit nabídka dodatečných služeb, která jim ušetří starosti a tedy i čas (např. revize el. zařízení, pojištění, profylaktická údržba kotle apod.).

5.2.3 Hodnocení – shluk 3

- Shluk 3 představuje **nejmenší skupinu** z celého vzorku analyzovaných dat (14%)
- Je tvořen převážně zákazníky produktové řady **Klasik (84%)**, tj. bez fixace smlouvy
- Skupina obsahuje rovnoměrně rozložené zákazníky s typem spotřeby Klasik, Aku a Přímotop;
- Za zmínku stojí, že z hlediska typu spotřeby shluk 3 představuje **49,2%** všech zákazníků s charakterem spotřeby **Přímotop**
- **Spotřeba** zákazníků, s ohledem na celý zkoumaný vzorek, je **nadprůměrná**, z hlediska srovnání s ostatními shluky je spotřeba dokonce **nejvyšší**. Průměrná spotřeba v rámci shluku je 5,75 MWh. Z hlediska variability shluk obsahuje i řadu odlehlých hodnot s vysokou spotřebou.
- Z demografického hlediska shluk tvoří především **ženy (80%)**, žijící spíše v městech do 5tis. obyvatel, s mírně vyšší mírou nezaměstnanosti.

Doporučení: shluk převážně obsahuje skupinu zákazníků v produktové řadě Klasik, kteří nejsou vázáni žádnou dlouhodobou smlouvou, a tedy mohou kdykoli změnit svého dodavatele. Zároveň shluk reprezentuje zákazníky s vysokou spotřebou s významnou částí zákazníků používající přímotopné vytápění. Shluk je navíc tvořen převážně ženskou populací. Pokud připustíme jejich menší technickou zdatnost, nemusí být zvolený způsob vytápění nejvhodnější. Nabízí se tedy možnost ověřit, zda např. nabídkou poradenských služeb v oblasti spotřeby el.energie by nebylo možné zajistit jejich loajalitu.

5.2.4 Hodnocení – shluk 4

- Velikost shluku 4 představuje 21% podíl z celého analyzovaného souboru dat
- Shluk obsahuje především zákazníky z **produktové řady Klasik (70%)**
- Z hlediska spotřeby jednoznačně převažuje **běžná spotřeba (Klasik, 91%)** s **převahou malého příkonu** s jističem **1x16A.** Celkově lze spotřebu charakterizovat jako podprůměrnou.
- Výrazná odchylku oproti ostatním shlukům představují proměnné „Počet obyvatel“ a „Nezaměstnanost“, kde shluk představuje především **zákazníky z velkých měst s nižší mírou nezaměstnanosti** (80% z celého zkoumaného vzorku).

Doporučení: uvedený segment poměrně jednoznačně identifikuje skupinu zákazníků. Jedná se o zákazníky z větších měst, s relativně nízkou spotřebou a se smlouvou umožňující kdykoliv provést změnu dodavatele. Pravděpodobně se jedná o zákazníky žijící v panelových bytech. Bohužel právě z důvodu nízké spotřeby není úplně zřejmé, jakým vhodným způsobem zajistit jejich loajalitu.

5.2.5 Hodnocení – shluk 5

- Shluk 5 představuje 16% podíl z celkového zkoumaného vzorku.
- Skupina obsahuje převážně zákazníky **produktové řady Trend (89%)**
- Z hlediska typu spotřeby převažuje běžná spotřeba (Klasik, 61%) následovaná akumulační sazbou (Aku, 37%)
- Spotřeba patří mezi průměrné, z hlediska příkonu se jedná převážně o jističe 3x25A
- Z hlediska velikosti obce jsou zákazníci zastoupeni ve všech sledovaných segmentech, s vyšší četností obcí do 1tis.obyvatel
- Ve srovnání s ostatními shluky je zde nejčetnější aktivita reprezentovaná kontakty s call centrem. Tato aktivita však může být spojena s přepisem produktu na řadu Trend.
- Z hlediska platby převažuje platba SIPO

Doporučení: shluk představuje poměrně konzistentní skupinu zákazníků se středně velkou spotřebou, představující do určité míry stabilizované zákazníky. Informace z tohoto shluku je možné použít pro vytipování a oslovení zákazníků patřící do jiného shluku (pravděpodobně ze shluku 4) s nabídkou Trend.

5.2.6 Určení hlavních charakteristik zjištěných shluků

Na závěr této kapitoly můžeme provést odhad hlavních charakteristik daného shluku, které můžeme použít jako východisko pro další marketinkové analýzy.

Shluk	Odhadovaná charakteristika shluku
Shluk 1	Zákazníci s nadprůměrnou spotřebou z menších měst s preferencí ceny
Shluk 2	Ekonomicky silní zákazníci s preferencí času nad penězi
Shluk 3	Ženy, s domácnostmi s nadprůměrnou spotřebou tvořenou přímotopy
Shluk 4	Zákazníci z velkých měst s nízkou spotřebou v produktové řadě Klasik
Shluk 5	Relativně stabilizovaní zákazníci s průměrnou spotřebou v produktové řadě Trend

Tabulka 12: Hlavní charakteristiky jednotlivých shluků

5.3 Doporučení pro implementaci shlukové analýzy

Aby bylo možné zjištěné poznatky využít i v rámci reálného marketinkového prostředí, je vhodné zmínit i několik předpokladů, které musejí být splněny a které považujeme za nezbytné uvést:

1) **Znalost obchodní problematiky** - důkladná znalost vlastního podnikatelského prostředí, obsahový význam jednotlivých informací, znalost podnikatelských procesů a správné určení marketinkových cílů, je nezbytná pro zajištění správných výsledků. Bez přímého vstupu lidí s požadovanými znalostmi do analýzy, je hledání smysluplného výsledku velmi obtížné popř. nereálné
2) **Příprava a kvalita dat** – velkou pozornost je nutné věnovat přípravě dat a zajištění datové kvality. V průběhu práce bylo nutno data očistit a přizpůsobit, aby byla zajištěna jejich dostatečná kvalita. Z pohledu času je příprava dat jednou z nejnáročnějších činností v průběhu statistické analýzy. Vlastní modelování již tolik času nezabere
3) **Metodika** – jak se v průběhu práce potvrdilo, volba vhodné metodiky – v našem případě metodiky CRISP, je jedním ze způsobů jak předejít řadě problémů a chyb. Metodika obsahuje řadu doporučení a postupů, jejichž opomenutí by mohlo přinést znehodnocení práce a nutnosti opakovat některé činnosti. Vzhledem k předchozímu bodu zmiňujícímu pracnost přípravy dat, by se mohlo jednat o velmi nákladný omyl.
4) **SW podpora** – jedním z důvodů, proč je využívání pokročilých statistických algoritmů v oblasti marketinku stále na nízké úrovni jsou vysoké nároky na požadované znalosti. Analytik se v rámci přípravy dat setkává s mnoha náročnými úkony, které je nezbytné provést pro zajištění nejlepšího výsledku. Namátkou zmiňme výběr vzorku dat se

správným rozdělením, transformaci dat a proměnných, ověření normality rozdělení, aplikaci vhodných metod a následnou verifikaci a analýzu výsledků. Toto jsou všechno znalosti, kterými běžný pracovník marketinku nedisponuje. Volba vhodného SW nástroje, která mu umožní jednoduchým způsobem provádět požadované analýzy, je jeden ze způsobů jak zvýšit využívání těchto stat. metod. Z hlediska zkušeností získaných v průběhu práce, můžeme potvrdit, že SW R je neuvěřitelně silným a robustním statistickým nástrojem, nicméně požadovaná úroveň znalostí a nízký uživatelský komfort jej směřují spíše do rukou experta než pracovníka marketinku.

5) **Časová náročnost** – zpracování dat pomocí statistických algoritmů je velmi časově náročná činnost. I když vlastní modelování je poměrně rychlé, příprava dat, analýza a interpretace výsledků popř. opakování experimentů jsou velmi časově náročné činnosti. Pokud společnost plánuje využití statistiky v oblasti marketinku provádět průběžně, systematická příprava datového prostředí, budování pracovních teamů a standardizace postupů mohou tuto pracnost významně snížit.
6) **Skromnost v očekávání** – ačkoliv by se zdálo, že po aplikaci složitých, pracných a nákladných statistických metod získá společnost významnou znalost, je třeba uvést, že tomu tak nemusí být vždy. Statistika je složitá věda, analytici pracují s nepřebernou řadou různých modelů, kdy každý z nich může poskytovat odlišné výsledky. I přes důkladné testování a validace výsledků je nutné být připraven, že získaný model neodpovídá realitě na marketinkovém trhu. Nicméně i přes tato rizika, příklady z řad společností ukazují, že v případě úspěchu ekonomické výsledky opodstatňují všechny případné neúspěchy.
7) **Uplatnění shluků** – výstupy shlukové analýzy jsou pouze vstupním podkladem pro práci marketinkových pracovníků. Samotný shluk pouze obsahuje sadu společných atributů pro definovanou skupinu zákazníků; praktické využití této znalosti však vyžaduje dobrou znalost obchodní problematiky. Se shluky se většinou nepracuje samostatně, výsledky se kombinují s dalšími znalostmi např. s výstupy analýzy konkrétních produktů, analýzy zákaznické loajality apod. Zavedení používání shlukové analýzy si zároveň může vyžádat i změnu přístupu k plánování marketinkových kampaní. Od relativně přímočaré a technicky jednoduché propagace vybraných produktů přistupujeme k individuálnímu plánování pro vybraný shluk. Tento přístup však požaduje dodatečné úsilí marketinkových pracovníků a společnost na to musí být připravena.

5.4 Závěrečné zhodnocení

Vytyčeným cílem této diplomové práce bylo provést segmentaci zákazníků ve vybrané společnosti s uplatněním metod shlukové analýzy. Pro účely analýzy dat byl požadavek použít programové prostředí R.

V rámci teoretické části práce byly rozebrány marketinkové cíle a důvody segmentace trhu, teoretická východiska shlukové analýzy, metodiky pro přípravu dat a SW nástroje, které je možné využít pro účely segmentace.

Praktická část se zaměřila na analýzu části zákazníků společnosti EON ze segmentu domácnosti - elektřina. V úvodní části byla nejprve popsána obchodní problematika utilitního trhu s elektrickou energií. Praktická část byla zpracována podle metodiky CRISP, jejíž použití se zároveň ukázalo jako vhodným vodítkem při zpracování úlohy segmentace.

Prostřednictvím analýzy vnitropodnikového informačního systému byla identifikována dostupná data, která byla následně exportována z jednotlivých zdrojových systémů do samostatné databáze MySQL. Již samotný rozbor dostupných dat poodhalil řadu zajímavých informací, zejména z hlediska kvality dat a z hlediska rozložení četnosti jednotlivých proměnných v datech. Data byla následně upravena, očištěna o chybné záznamy a rozšířena o další informace. Z důvodu velikosti datového souboru byl připraven výběrový vzorek, na který byly následně aplikovány metody shlukové analýzy.

Samotná aplikace shlukových algoritmů se ukázala jako poměrně jednoduchý způsob jak přistoupit k automatizované segmentaci zákazníků. Již rozborem jednotlivých proměnných byla odhalena řada zajímavých skutečností, které samy o sobě poskytly řadu zajímavých podkladů pro další marketinkové činnosti. Můžeme tedy říci, že shluková analýza umožnila jednoduchým a rychlým způsobem napomoci k odhalování skrytých informací.

Další interpretací výsledků jednotlivých proměnných v kontextu daného shluku jsme se následně pokusili identifikovat možné hlavní charakteristiky daného shluku a jeho využití pro marketinkové účely. Z pohledu marketinku je klíčovou činností nalezení vhodné interpretace výsledků a nalezení reálného využití získaných shluků pro marketinkové účely. Nezbytným předpokladem je zde dobrá znalost obchodní problematiky a znalost daného trhu.

Souhrnně lze prohlásit, že shluková analýza nabízí poměrně jednoduchý způsob jak převést problematiku zákaznické segmentace na automatizovaný způsob zpracování dat s následnou analýzou získaných výsledků. Využití programového prostředí R pak umožňuje efektivním způsobem provedení vlastní analýzy s možností využití širokých možností tohoto statistického jazyka.

6 Přehled zdrojů a použité literatury

6.1 Tištěné dokumenty

BERKA, Petr. *Dobývání znalostí z databází*: *analýza a metaanalýza dat*. Vyd. 1. Praha: Academia, 2003, 366 s. ISBN 80-200-1062-9.

BURNS, Alvin C a Ronald F BUSH. PEARSON INTERNATIONAL EDITION. *Basic marketing research*: *using Microsoft Excel data analysis*. 2nd ed. Upper Saddle River, N.J.: Pearson Prentice Hall, c2008, xxix, 4510. ISBN 978-0-13-135421-0.

HEBÁK, Petr. *Vícerozměrné statistické metody [1]*. 2., přeprac. vyd. Praha: Informatorium, 2005, 239 s. ISBN 978-80-7333-056-9.

HEBÁK, Petr. *Vícerozměrné statistické metody [2].* 1. vyd. Praha: Informatorium, 2005, 366 s. ISBN 80-733-3036-9.

HEBÁK, Petr. *Vícerozměrné statistické metody [3]*. Vyd. 1. Praha: Informatorium. ISBN 80-733-3039-3.

HENDL, Jan. *Přehled statistických metod zpracování dat*: *analýza a metaanalýza dat*. 1. vyd. Praha: Portál, 2004, 583 s. ISBN 80-717-8820-1.

HENDL, Jan. *Kvalitativní výzkum*: *základní teorie, metody a aplikace*. 2., aktualiz. vyd. Praha: Portál, 2008, 407 s. ISBN 978-807-3674-854.

JOHANSON A. C. a KOKOCINSKI T. M. *Not (Just) Another Stats Book*. 2nd ed. Paradox, Wiley 2000 ISBN 04-714-1027-6

KOTLER, Philip. *Marketing management*. 11th ed. Upper Saddle River: Prentice Hall, c2003, 706 s. ISBN 01-303-3629-7.

KOTLER, Philip. *Moderní marketing*: *4. evropské vydání*. 1. vyd. Praha: Grada, 2007, 1041 s. ISBN 978-802-4715-452.

MACHKOVÁ, Hana. *Mezinárodní marketing*: *nové trendy a reflexe změn ve světě*. 3., aktualiz. a přeprac. vyd. Praha: Grada, c2009, 196 s. Expert (Grada). ISBN 978-802-4729-862.

RUD, Olivia. *Data Mining*. Vyd. 1. Praha: Computer Press, 2001, 329 s. ISBN 80-722-6577-6.

SEGER, Jan. *Statistické metody v tržním hospodářství*: *základní teorie, metody a aplikace*. 1. vyd. Praha: Victoria Publishing, 1995, 435 s. ISBN 80-718-7058-7.

6.2 Elektronické dokumenty

CRISP-DM CONSORCIUM. *CRISP-DM 1.0*. 2000. Dostupné z: CRISP-DM [online]. [cit. 2012-04-15]. Dostupné z: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/

JACKSON, Joyce. Data Mining: A Conceptual Overview. *Communications of the Association for Information Systems*. 2002, č. 8, s. 31. Dostupné z: http://faculty.wiu.edu/C-Amaravadi/is524/res/dm_c_ov.pdf

MATULA, Vladimír. Segmentace trhu, segmentace zákazníků. [online]. [cit. 2012-01-17]. Dostupné z: http://www.vladimirmatula.zjihlavy.cz/segmentace-trhu.php

R: Cluster analysis. In: *Wikipedia*: *the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2012-04-10]. Dostupné z: http://wiki.math.yorku.ca/index.php?title=R:_Cluster_analysis&redirect=no

Segmentace trhu. [online]. [cit. 2012-01-17]. Dostupné z: http://managementmania.com/segmentace-trhu

Souhrnný přehled vydávání licencí pro podnikání v energetických odvětvích. [online]. ERU, 1.1.2012 [cit. 2012-04-10]. Dostupné z: http://www.eru.cz/user_data/files/licence/info_o_drzitelich/souhrn_12_01.pdf

Srovnání dodavatelů energie podle počtu zákazníků za rok 2011. In: [online]. [cit. 2012-04-10]. Dostupné z: http://www.cenyenergie.cz/nejnovejsi-clanky/srovnani-dodavatelu-energie-podle-poctu-zakazniku-za-rok-2011.aspx

VÍT, David. *Využití shlukové analýzy v marketingu* [online]. [cit. 2012-04-15]. Dostupné z: www.lemonway.com/research/dp-final.pdf

7 Rejstřík a seznamy

7.1 Seznam obrázků

7.2 Seznam tabulek