ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA STROJNÍ
ÚSTAV ŘÍZENÍ A EKONOMIKY PODNIKU

DIPLOMOVÁ PRÁCE

Analýza a segmentace zákazníků pomocí statistických metod

Customer analysis and segmentation using statistical methods

AUTOR: Bc. Nikita Zhitnikov

STUDIJNÍ PROGRAM: STROJNÍ INŽENÝRSTVÍ

VEDOUCÍ PRÁCE: Ing. Barbora Stieberová, Ph.D.

PRAHA 2019

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Zhitnikov** Jméno: **Nikita** Osobní číslo: **397738**

Fakulta/ústav: **Fakulta strojní**

Zadávající katedra/ústav: **Ústav řízení a ekonomiky podniku**

Studijní program: **Strojní inženýrství**

Studijní obor: **Řízení a ekonomika podniku**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Analýza a segmentace zákazníků pomocí statistických metod

Název diplomové práce anglicky:

Customer analysis and segmentation using statistical methods

Pokyny pro vypracování:

1. Úvod
2. Teoretická východiska – analýza zákazníků, segmentace trhu
3. Teoretická východiska – charakteristika statistických metod v oblasti analýzy zákazníků
4. Provedení analýzy zákazníků
5. Posouzení současného stavu segmentace
6. Návrh změn a posouzení jejich dopadů
7. Závěr

Seznam doporučené literatury:

1. Kotler, P., Keller K. Marketing management. New Jersey: Prentice-Hall, 2009. ISBN 978-0131457577
2. Hill, M., Meloun, M. Militký, J. Statistická analýza vícerozměrných dat v příkladech. Praha: Karolinum, 2017. 978-80-246-3618-4.
3. Brabenec, V., Šařecová, P. Statistické metody v marketingu a obchodu - vybrané přednášky a příklady. Praha: Česká zemědělská univerzita v Praze, 2011. ISBN: 978-80-213-0747-6.
4. Kožíšek, J., Stieberova, B. Statistika v příkladech. Praha: Verlag Dashöefer, 2012. 978-80-86897-48-6.
5. Chlebovský, V. Management zákaznických řešení. Praha: Grada, 2017. 978-80-271-0559-5.

Jméno a pracoviště vedoucí(ho) diplomové práce:

Ing. Barbora Stieberová, Ph.D., ústav řízení a ekonomiky podniku FS

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **24.10.2019** Termín odevzdání diplomové práce: **03.01.2020**

Platnost zadání diplomové práce: **28.02.2020**

Ing. Barbora Stieberová, Ph.D.	prof. Ing. František Freiberg, CSc.	prof. Ing. Michael Valášek, DrSc.
podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

29.10.19
Datum převzetí zadání

Podpis studenta

Prohlášení

Prohlašuji, že jsem tuto práci vypracoval samostatně a to výhradně s použitím pramenů a literatury, uvedených v seznamu citovaných zdrojů.

V Praze dne: 31.12.2019 ……

Podpis

Anotace

Předmětem diplomové práce je analýza portfolia zákazníku v byznys segmentu telekomunikačního trhu. První část je teoretická a popisuje statistickou analýzu vícerozměrných dat, obsahuje i zásady marketingu při analýze zákazníků. Druhá část je analytická, v ní se provádí shluková a korelační analýzy, posouzení jejích výstupu a doporučení pro management společnosti.

Klíčová slova

Regrese, korelace, shluk, analýza, segmentace, klient

Annotation

The subject of the thesis is an analysis of customers in a business segment of the telecomunication market. First part is theoretical and describes statistical methods of multidimensions data. Second part is analytical. It consists of analysis of the problem and it's solution. Last part is design and sums up the results.

Keywords

Regression, corelation, cluster, analysis, segmentation, client

Poděkování

Chtěl bych poděkovat své vedoucí diplomové práci Ing. Barboře Stieberové, Ph.D. za konzultace, cenné rady a odborné vedení při zpracování. Dále bych chtěl poděkovat vedení společnosti, které projevilo zájem o spolupráci, poskytnulo potřebné údaje a bylo otevřeno k následnému dialogu při hledání řešení.

Obsah

Úvod

Cílem moje práci provést analýzu portfolia zákazníků a jejich následnou segmentaci. Pro řešení úkolu budou použité vhodné softwarové nástroje jako MS Excel a SSPS Statistics od společnosti IMB. Data poskytla telekomunikační společnost působící na českém trhů. Je to dost specifická oblast podnikání jak z hlediska ekonomiky a marketingu tak i z pohledu technologii a legislativních předpisu.

Obsahově moje diplomová práce je rozdělená na několik částí. Do první spadá teorie, ve které popsují proces marketingové segmentace trhu, jeho postavení a pohledy na něj. Následně uvádím statistické metody, začínající popisnou statistikou a analýzou souboru dat, s pokračováním ve složitějších a náročných metodách jako hledání regrese, korelaci a vícerozměrné analýze údajů.

Druhá část je analytická. Začíná se popisem problematiky trhu a současného stavu portfolia zákazníku. Dále probíhá aplikace metod popsaných v teoretické části. Podrobné rozebrání souborů dat a provedení seskupení do segmentů na základě postupu clusterové analýzy. Taky budou uvedený grafické a hodnotové výstupy ze softwaru.

Obsahem poslední části je souhrn výsledků, jejích posouzení a okomentování. Závěrem jsou návrhy změn a opatření, doporučení k optimalizaci vnitřních systému či rady k volbě strategii.

1. Teoretická východiska – analýza zákazníku, segmentace trhu

1.1 Moderní pojetí zákazníka

Trh jako takový skládá se ze dvou hlavních složek: zákazník a prodávající. Proces směny tvoří mezi nimi vztah. Základem pro marketing první poloviny minulého století bylo 4P: výrobek (Product), propagace (Promotion), místo (Place) a cena (Price). Cílem bylo prodávat co nejvíc, protlačit produkt na trh bez ohledu na potřebu a poptávku o tom výrobku. Rozvoj komunikačních a informačních technologii ulehčil navázání ekonomických a sociálních vztahu, umožnilo se oslovení nových trhu a klientu, ale zároveň této změny vyvolali řadu problémů. Jak uvádí Chlebovský, 2005:

- Zpětná vazba, kterou lze zajistit relativně rychle v rámci regionu, v globálním prostředí má časové zpoždění které se nedá zanedbat;
- Produkt připravený pro globální trh špatně uspokojuje lokální požadavky;
- Místo, které si nemohou vybrat lokální marketingové manažeři ani jehož jsou přímou součástí.

Přes dlouhou historickou cestu marketingu a krize marketingového mixu 4P odbornici a filozofe marketingových škol přišli na to, že právě zákazník je tou hlavní složkou. A pokud dokážeme ho udržet, uděláme z něj stávajícího, loajálního kupujícího – dostaneme o hodně víc v dlouhodobě perspektivě. S každým rokem se zvyšuje integrita světa víc a víc, vstupují moderní technologie. Tím se zvyšuje i konkurence, jde o tak zvané „turbulentní prostředí“ ve kterém vzrůstává potřeba vybudování stabilních vztahu se zákazníkem.

Musíme vědět o klientovi co nejvíc – co preferuje, jeho životní cyklus, jaké má strachy a potřeby, co je pro něj důležité, na co dává pozor. A kdo ví nejvíc – vyhrává. Správné nastavené mechanizmy sběru dat o zákaznícovi, jejích následné zpracování a analýza je základem pro výběr strategie a přijetí rozhodnuti ku prospěchu podniku. Vhodným nástrojem k tomu slouží CRM.

1.2 CRM – podstata a detaile

CRM (Customer Relationship Management neboli Řízení vztahů se se zákazníky) je model interakce, vybudovaný na základě teorie, že na centrálním místě celé filozofie byznysu je klient a hlavními cíli jakékoliv podnikové činnosti jsou zajištění efektivního marketingu, prodeje a uspokojení potřeb zákazníku (Chlebovský, 2005). Podporou tomuto slouží sběr, ochrana a analýza dat klientu, dodavatelů, partneru a vnitřních procesu firmy.

Cílem pro zavádění systému je zvětšení uspokojenosti a loajality klientů, regulace tarifní politiky, kalibrace marketingových nástrojů. Osnovou je jediné skladiště informace, kam se zaznamenává veškeré interakce se zákazníky – tzv. klientská databáze. Její naplnění se provádí využitím různých kanálu oslovení: servis na prodejních místech, telefonické volání, elektronická pošta, akce a setkání, registrační formuláře na webových stránkách, reklamní odkazy, sociální sítě. Pomoci automatizované analýzy údajů lze efektivně a s minimálním využitím pracovníků kalkulovat individuální potřeby a přání. Zároveň kvůli rychlému procesu můžeme odhalovat rizika a potenciální příležitostí.

Faktický CRM představuje metodiku, převedenou do celopodnikové strategie, která rozdělena na procesy, zaměřené na vybudování ziskových vztahu se zákazníkem. Role informačních technologie v tomto systému pouze podpůrné a slouží k automatizace celkového procesu, který začíná získáním údajů, následnou analýzou potřeb a vzorců chování (přeměnou dat na informace), a využití těchto informací pro úspěšnou komunikaci s klienty:

- Operativní CRM – sběr primární informace, podpora pro front office. Veškeré interakce se zákazníkem sledované a zaznamenané v databázi, ke které uživatele mají jednoduchý přistup. Hlavními procesy jsou marketingové kampaně, automatizace prodeje a jejích sledování;
- Analytické CRM – analýza a reporting informací z různých pohledů: purchase funnel, výsledky marketingových kampaní, rentabilita a chování klientu, efektivita prodeje z pohledu produkce, segmentů, regionů atd;

- Kolaborativní CRM – komunikace podniku a jeho zákazníku prostřednictvím různých kanálů. Udává možnost ovlivnit procesy firmy: dotazování pro zlepšení kvality produktu či služby, změny pořadí obsluhování, monitorování své zakázky, interaktivní objednání)

1.3 Základní pohled – ziskovost zákazníka

Osnovou podnikaní je jednoduchá rovnice Zisk = Výnosy – Náklady. Podstatou firemní obchodní strategie je aplikace této rovnice na zákazníka nebo typy klientů. Některé z nich přináší firmě velké zisky a netvoři velké náklady, jiní naopak. Znalost ziskovosti umožňuje rozlišit přístup k jednotlivcům a tím minimalizovat náklady na neziskové klienty. Vypočítat příjmy od každého zákazníka není složité, kalkulovat náklady spojené s klientem je mnohém obtížněji. Zahrnuje to nejen přímé náklady na výrobu a dopravu zboží či služeb ale i veškeré náklady na podporu a realizace prodeje. Proto je nutné vymezit všechny činnosti spojené s konkrétním klientem a přiřadit podle toho náklady. Využívá se často metoda ABC (activity based costing).

Další důležitou charakteristikou customer value managementu je hodnota zákazníka – jeho současné a budoucí hodnoty diskontované na jeho čistou současnou hodnotu. Základem pro tento propočet slouží čtyři kvantifikovatelné veličiny: náklady, investice, obrat a riziko (Žáček, 2010).

Klient, který je ztrátový dnes muže přinášet velké zisky v budoucnu. Proto je mimo aktuální přínosy musíme brát v úvahu budoucí potenciál. Musíme porozumět jeho životnímu cyklu, jak se vyvíjí potřeby a jakým způsobem je firma muže uspokojit.

Pokud budeme sledovat ziskovost a hodnotu klienta během jeho životního cyklu budeme mocí rozdělit a segmentovat jejích v databází. Základní čtyři skupiny jsou: udržovací segment (minimalizovat náklady), dojný segment (osobní přístup, nejkvalitnější služby), útlumový segment (přenechat konkurenci), růstový segment (osobní přístup, nabízet ziskovější produkty).

1.4 Segmentace trhu

Segmentací rozumíme proces rozčlenění trhu do homogenních skupin, odlišných od sebe svými potřebami, charakteristikami a svým chováním při nakupování (Žáček, 2010). Při tom ne každá segmentace může být účelná. Marketingové specialisti podniku musí zvažovat kritéria a rozhodovat do jaké míry provádět proces tak aby segmentace měla přínos, nebyla zbytečná a ve výstupu skupiny klientů nebyli příliš malé.

Jsou různé rozlišovací úrovně, jak uvádí Kotler (2005). Na jedné stráně jsou velmi široké segmenty neboli masové, které nemají homogenitu, a můžeme je dále segmentovat. Na druhé straně vznikají malé segmenty se zvláštními, speciálními potřebami a přáními (**niche markets**).

Při klasickém marketingovém postupu se vybírá taková skupina zákazníků, která nabízí nejlepší příležitostí a s kterou se dá nejjednodušší pracovat. Provádí se zařazení klientů do tržních segmentu, které by měli být obsluhovány různými způsoby s ohledem na geografické, demografické, sociální a ekonomické faktory.

Tržní segment představuje skupinu klientů, kteří stejným způsobem reagují na ovlivnění marketingovými nástroji. Dvě základní podmínky skupin:

- podmínky homogenity – podobnost tržních projevů zákazníků uvnitř skupiny je co největší;
- podmínka heterogenity – vzájemná odlišnost tržních projevů mezi segmentu na daném trhu je co největší.

Kromě těchto základních podmínek by měli plnit odkryté segmenty ještě řádu dalších kritérií (Žáček, 2005):

1. dosažitelnost – segmenty trhu musí být pro podnik dosažitelné z hlediska distribučních cest a disponibilních zdrojů;
2. dostupnost – segment musí být dostupný k ovlivnění marketingovými nástrojami;

3. objektivita – vymezení segmentu musí být provedeno co nejvíce objektivně, bez vstupu subjektivních posudku a předpokladů;
4. stabilita – neměl by segment podléhat velkým a častým změnám, což se naopak projevuje u některých specifických segmentů;
5. velikost – musí být optimální pro efektivitu podniku a následnou možnost růstu.

1.5 Proces segmentace

Proces probíhá v několika na sebe navazujících fázích, které předchází marketingový průzkum trhu, sběr a analýza informace.

Vymezení trhu. Rozhodnutí jaký trh bude segmentován obecně se děla na základě dvou charakteristik a to jsou především produkt a geografická poloha zákazníka.

Stanovení rozhodujících kritérii. Stanovuji se jaké charakteristiky a tržní projevy vykazují výrazné rozdíly na daném trhu, v čem se to projevuje a jaká je jejich míra významnosti. V závislosti na tom je-li to spotřební nebo podnikový trh kritéria se liší hloubkou vymezení – buď do podrobná nebo ze širšího pohledu.

Rozpoznání segmentu. Ze stanoveného počtu kritérii se vybírá jejích kombinace, která se považuje za nejdůležitější a nejlíp odkrývá tržní segment z hlediska homogenity a heterogenity. Musí byt zvoleno minimálně jedno vymezující kritérium – popis jejích typu bude proveden v následující kapitole.

Rozvoj profilu segmentu. Tento proces zachycuje rozšíření marketingových charakteristik v závislosti na typu trhu: sledovanost masových médií, účast na veletrzích a výstavách atd.

1.6 Typy tržních kritérii

Každá firma sama stanovuje jakým způsobem a podle jakých kritérií segmentovat tržní prostředí. Volba se však zakládá na výrobků nebo službě, které podnik poskytuje. Na výběr ale má dvě velké skupiny (Kotler, Keller, 2007):

- **kritéria tržních projevů** – jsou vymezující proměnné, které udávají rozdíl mezi zákazníky ve vztahu k produktům a k danému tržnímu prostředí. Prověřují homogenitu a heterogenitu trhu. Dále se rozpadají na:
    - příčinná kritéria – jsou spojená s důvody chování zákazníka: očekávání od produktu, jak podle potřeb a přání vnímají konkrétní značky, jaké příležitosti přináší produktu a jeho užívání (sociální, časové, místní atd.);
    - kritéria užití – jsou založená na spotřebním projevu. Příkladem jsou: uživatelský status, který dělí zákazníky na dva široké segmenty – uživatele a neuživatele – s navazujícím ukazatelem potenciálu růstu trhu; míra užití dělicí klienty na silné a slabé uživatele produktu; kritérium věrnosti, které udává stálost užívání služeb a výrobků. Téhle tři údaje spolu tvoři strukturu pyramidy zákazníků v rámci CRM.
- **kritéria popisná** – jsou vysvětlující proměnné, vycházející z obecných charakteristik klienta. Vystupuji v roli nezávislých proměnných vůči kritériím tržních projevů. Rozpadají se na další složky jako:
    - tradiční kritéria – charakterizují zákazníky z geografických, demografických a etnografických pohledů. Jejich výhodou jsou dostupnost, měřitelnost, kvantifikace. Často mají velké vazby na změny ve spotřebním chování. Historicky vstoupily do marketingu nejdříve;

- psychografická kritéria – souhrn sociálních a psychických charakteristik. Vysvětluji odlišné chování v rámci stejné kategorii klientu.

Kritéria tržních projevů (vymezující proměnné)		Kritéria popisná (vysvětlující proměnné)	
Příčinná kritéria	Kritéria užití	Tradiční	Psychografická kritéria
Očekávaná hodnota	Uživatelský status	Demografická (v širším pojetí)	Sociální třída
Vnímaná hodnota	Míra užití	Etnografická	Životní styl
Příležitosti	Věrnost	Fyziologická	Osobnost
Prostoje, preference	Difúzní proces	Geografická	
	Způsob užití		

Tab. 1 Kritéria uplatňovaná při segmentaci spotřebních trhů (Žáček, 2010)

Při analýze B2B trhu se rámcově zachovává struktura typu kritérií, ale některá z nich v tomto případě mají větší váhu a tím vystupují do popředí, a to například rychlost dodávek.

Kritéria tržních projevů	Kritéria popisná - založená na charakteristikách
Očekávaná hodnota	Odvětví
Příležitosti	Velikost
Náročnost / Samostatnost	Geografie
Míra užití	Složení orgánu, který rozhoduje
Uživateský status	Nákupní politika
Zákaznická loajalita	Používané technologie
Rychlost dodávky	

Tab. 2 Kritéria uplatňovaná při segmentaci podnikových trhů (Žáček, 2010)

Segmentace na tomto typu trhu je méně komplikována v porovnání se spotřebními trhy z důvodu menšího počtu zákazníků, více racionálního a profesionálního jednáni, dostupnosti a objemu dat. Ale na druhé straně klienti mají složitější životní cyklus a systém rozhodování.

1.7 Přístupy k segmentaci a její metody

Postupy tržní segmentace lze rozdělit na dva směry. První založený na základě vlastních myšlenek, předpokladu a odhadu – intuitivní. Druhá varianta je systematická, kterou rozebereme podrobněji.

Systémový přistup má jako základní metodu pozorování okolí – dedukce. Provádí se odvození parametrů segmentů na základě záměru a politiky ostatních „hráčů“ v roli kterých mohou vystoupit konkurenční firmy, komplementární výrobce a další.

Druhou cestou je induktivní přístup, který vyjadřuje snahu vlastního odkrytí segmentu:

- post hoc – vybírá se několik kritérií, zkoumá se možnost jejich kombinací, provázanosti na chování zákazníků. Potom se vybírá jedno z kritérií a pomoci něho se odkrývá tržní segment;
- a priori – vybírá se jedno kritérium, podle kterého se vymezuje segment a bude se rozvíjet jeho profil:
    - forward – přístup založený na kritériích chování;
    - backward – přístup založený na popisných kritériích a zpětném prověření jejich vazby na kupním chování.

Při jakémkoliv přístupu je potřeba rozhodovat a volit strategie, což není možné bez vhodných informaci, které přináší jistotu a snižují riziko propadu. K tomu slouží metody sběru dat. Získání sekundárních údajů a jejich obsahová analýza jsou podporou deduktivního postupu segmentace. Pro induktivní přístup je

vhodné využit výzkumní agentury, dotazování a vyhodnocené prvotní údaje – data mining.

Po shromáždění potřebných informaci následuje jejích zpracování, k čemu slouží různé metody analýzy dat, které zachycují v sebe matematické výpočty a využiti statistických metod. K nim patři křížová, faktorová, shluková, diskriminační analýza, stromové strukturní metody, analýza regrese a korelace, statistické metody vícerozměrných dat. Některé z nich budou podrobně rozebrané v dalších kapitolách.

1.8 Tržní zacílení a umístění

Targeting je část cílového marketingu a začíná se po otevření tržních segmentů. V podstatě to je rozhodování, na které z nich by se měl podnik zaměřit. Vyhodnocování se provádí z hlediska atraktivity, musí se zvolit takový segment nebo segmenty, kterým podnik muže poskytnout největší hodnotu v co nejdelším časovém úseku. Každá firma má omezené zdroje a vlastní kapacity, proto by se měla počítat s tím při volbě počtu segmentu, na které chci vstoupit a působit. Takový přistup může omezit prodej ale zároveň snížit rizika i přesto podnik zůstane ziskový.

Většina podniků při vstupu na trh rozhoduje působit na jednom segmentu a až dosáhne na něm úspěchu, začne analyzovat další segmenty a rozhodovat o vstupu na něj. Ovládnutí celým trhem je legislativně omezené ve většině zemí. Ale velké podniky si mohou dovolit obsadit jeho velkou část při dostatečných ambicích a zdrojích.

Proces targetingu prochází následujícími kroky a měl by byt co nejvíc založen na objektivních podkladech než na hodnocení expertů (Žáček, 2010):

1. **Vymezení kritérií pro hodnocení vhodnosti tržního segmentu.** Vhodnost můžeme hodnotit z absolutního nebo relativního pohledů:
    a. Všeobecné platné charakteristiky (absolutní atraktivita): velikost, tempo růstu, kupní síla, náklady pro vstup do segmentu, současná a potenciální konkurence, hrozba substitutu, distribuční cesty, nákupní a platební zvyky;

b. Charakteristiky ve vztahu k podniku (relativní atraktivita): podnikové cíle, dostupné zdroje, know-how, distribuční kanály;

2. **Stanovení významnosti kritérií.** Při vícekriteriálním rozhodování stanovuje se váhy důležitosti kritérií, protože ne všechny z nich jsou stejně důležité v různých situacích;

3. **Stanovení stupnic jednotlivých kritérií.** Může byt použit jednotný rozsah pro všechny kritéria a hlediska, nebo stanovit specifické rozsahy jednotlivých kritérií. Výhodou stejného rozsahu je jednoduchost postupu, nevýhodou je ztráta rozlišení dat a riziko nivelizace;

4. **Hodnocení segmentu.** Přiřazení hodnot na stupnících u každého kritéria;

5. **Zvolení algoritmu.** Nejčastěji se používá aditivní algoritmus. Další možnost je multiplikační postup;

6. **Uplatnění algoritmu.** Výsledkem je pořadí segmentů podle jejích atraktivity.

Ze stanoveného pořadí se zvolí segment, na kterém bude podnik působit a oslovovat klienti. Musí zvolit pozice výrobku v představení zákazníka, kterým se ten bude lišit od konkurenční produkce. Positioning je nástroj propagace značky, vymezuje výrobek v myslích klientu, charakterizuje, čím se liší od konkurenčních služeb a výrobků s cílem dosáhnout strategické výhody.

Klasickým příkladem konkurenční výhody je poměr ceny a kvality. Podnik muže mít stejnou a popřípadě vyšší cenu než ostatní na tom trhu ale odůvodnit to lepší kvalitou svých výrobků či služeb. Další možnost je se odlišit nižší cenou nebo větší hodnotou pro zákazníka. Jakmile se firma rozhodne, čím zaujme vybraný segment, musí se soustředit na tom jak informovat zákazníka o své pozici.

1.9 Kritéria pro analýzu zákazníků

Stanovení a sledování hodnoty klienta pro firmu je klíčovou aktivitou pro úspěšný a efektivní prodej. Analýza portfolia zákazníků, se kterými firma momentálně pracuje, dává možnost včas přizpůsobovat obchodní politiku. Zároveň umožňuje cílově působit na proces vybudování u udržení vztahu se stálými klienty.

K porovnání zákazníků je potřeba vybrat vhodná kritéria a stanovit škály hodnocení ke každému z nich. Cílem takového vícekriteriálního srovnání je rozsegmentovat klientskou bázi a následně zvolit optimální strategii a činnosti pro každý úsek. Nejčastěji se k tomu používa následující kritéria:

- objem prodeje za časový period – ukazuje jaké služby či výrobky nejvíc nakupují klienti, pomáhá odhalovat potřeby a přání;
- vyšší tržeb za stejné období. Smyslem obchodu je zisk, proto je potřeba soustředit se na klienty, které přinášejí nejvíc;
- doba splacení – je důležitým ukazatelem z toho důvodu že na ní je závislý cash-flow firmy. Často můžeme vidět v praxi, že klient dostává lepší obchodní nabídky, pokud stále dodržuje platební podmínky;
- růst objemu nákupu – demonstruje zájem klienta o výrobek a jeho rostoucí závislost na něm, což slouží k vybudování dlouholetého a výhodného partnerství;
- kupní síla. Pokud existuje řada potenciálních odběratelů – několik menších a jeden velký, musí se firma věnovat víc času a úsilí tomu největšímu z nich aby dosáhla většího obratu produktu a vydělala co nejvíc;
- nákladovost – má být sledovaná spolu s tržbami, aby se dalo posoudit kolik reálně přináší klient, zda není ztrátový kvůli časovým, finančním a lidským zdrojům, potřebným pro jeho údržbu.

2. Teoretická východiska - charakteristika statistických metod v oblasti analýzy zákazníků

Jak bylo zmíněno v předchozích kapitolách - přistupuji podrobněji k popisu statistických metod, podporujících analýzu a segmentaci zákazníků. Počet takových nástrojů a modelu je velký a nemá smysl popisovat všechny z nich jak z důvodu obsahu, jinak z hlediska vhodnosti – každá z těch metod použitelná pro stanovení řešení v rámci omezujících podmínek určitých problémů, proto uvedu jen té z nich, které budou následně použité v praktické části.

2.1 Popisné statistiky analýzy dat

Před zpracováním získaných dat se obyčejně provádí analýza souboru jako takového. Jde o statistický charakteristiky, které vyjadřuji velikost sledovaných znaků (polohy) a nesouměrnosti v rozdělení (měnlivosti): výkyvy, špičatosti, a koncentrace hodnot. Takové předběžné posouzení umožňuje pochopit zda-li soubor dat je vhodný pro následné zpracování a zároveň z výstupu už můžeme říct jednoduché závěry.

K tomu využíváme míry polohy, které ukazuji, kde se data nacházejí, která hodnota je středem souboru. K ním patří: průměr – střední hodnota souboru:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{X} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}$$

Medián je prostřední hodnota souboru, která ho rozděluje na dvě poloviny. Má smysle pouze u reálných jednorozměrných veličin jako hmotnost, výška atd. Při lichém počtu hodnot medián je prostřední hodnotou. Při sudem počtu dat za medián se označuje aritmetický průměr dvou prostředních hodnot (n/2; n/2+1). Modus – nejčastěji se opakující hodnota:

$$\hat{x} = x_D + \frac{V_1}{V_1+V_2}.h, \; kde \; h = x_H - x_D$$

kde V_1 … četnost předcházejícího intervalu,

V_2 … četnost následujícího intervalu

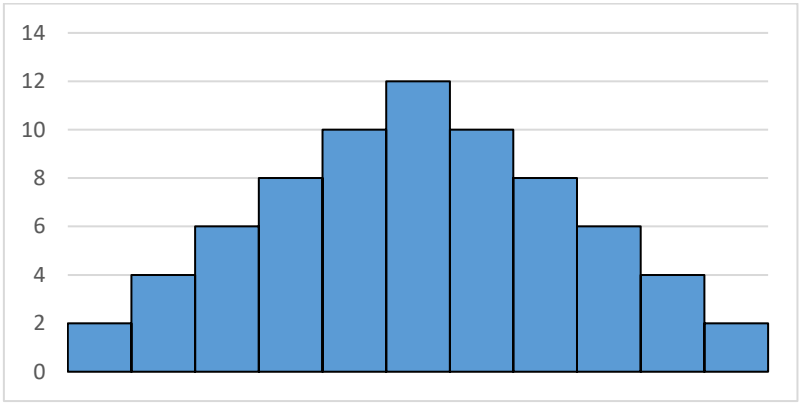
Míry variability udávají informaci o tom jak jsou hodnoty souboru navzájem blízké nebo vzdálené. K ním patří: rozptyl – rozmezí, ve kterém se hodnoty pohybuji:

$$s_x^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \quad s_x^2 = \frac{\sum_{i=1}^{k}(X_i - \bar{X})^2 n_i}{\sum_{i=1}^{k} n_i}$$

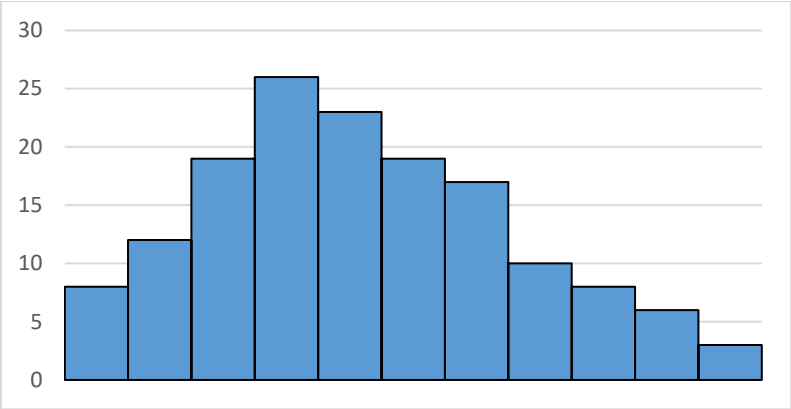
Směrodatná odchylka, která ukazuji, jak se hodnoty souboru navzájem liší:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}} \quad s_x = \sqrt{\frac{\sum_{i=1}^{k}(X_i - \bar{X})^2 n_i}{\sum_{i=1}^{k} n_i}}$$

Výstupem jsou histogramy rozdělení, na kterých grafický vidíme, do kterých skupin se rozpadají hodnoty nejvíc a nejmíň:



Graf č. 1 – Symetrický histogram

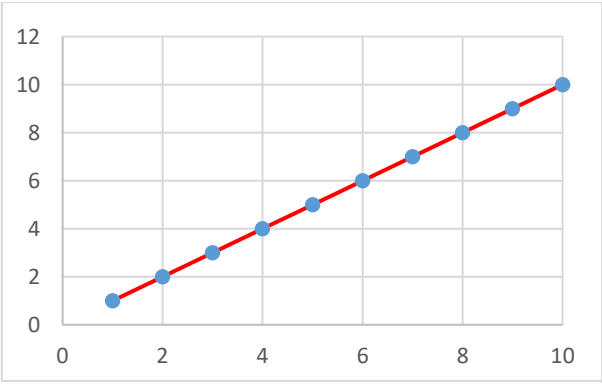


Graf č. 2 – Nesymetrický histogram

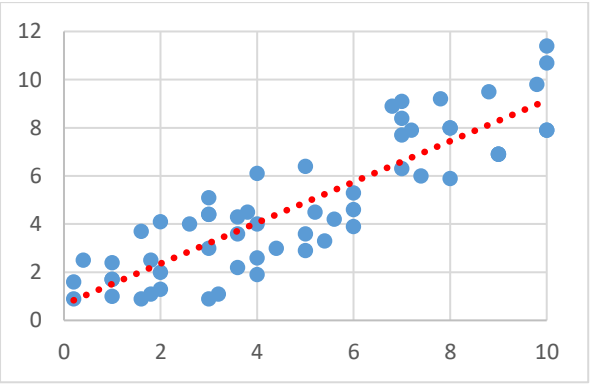
2.2 Úvod do regresní a korelační analýzy

Specialisty z různých oblasti – od ekonomiky do řízení jakosti, sleduji závislosti a vztahy mezi jednotlivými jevy. Odbornici se snaží zjistit jaký faktor má vliv na sledované ukazatele a stanovit vzorec kterým se řídi. Znalost vztahu mezi závislou a nezávislou proměnnou dává možnost ovlivňovat celý systém. Však důležité vědět i těsnost tohoto vztahu.

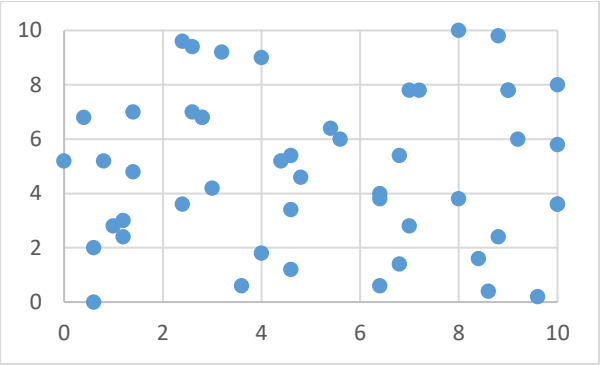
Závislost může byt představena matematicky nebo korelačně. U funkční (matematické) závislosti určité hodnotě nezávislé proměnné odpovídá jedná hodnota závislé. Je znázorněna na grafu č. 1. U korelační závislostí každé hodnotě nezávislé proměnné odpovídá rozdělení četností závislé proměnné, je zobrazená na grafu č. 2. Kromě toho existuje opačná charakteristika – nezávislost proměnných, kterou se dá taky znázornit graficky. Na grafu č. 3 vidíme, že hodnoty nevykazují žádnou tendenci.



Graf č. 3 – Matematická závislost



Graf č. 4 – Korelační závislost



Graf č. 5 – Nezávislé hodnoty

Abychom mohli popsat a analyzovat vztah mezi hodnotami musíme vyřešit regresní úkol – stanovit hlavní tendenci vztahu a následně nahradit korelační pole matematickou funkci pomoci metody nejmenších čtverců. Podstatou metody je výstižnost regresní čáry, která daná podmínkou:

$$s_{y,x}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min.$$

kde y_i … empirické hodnoty

$\hat{y}_i$ … hodnoty vyrovnané

$s_{y,x}^2$ … reziduální (zbytkový) rozptyl

Součet čtverců odchylek empirických hodnot od hodnot vyrovnaných musí byt minimální. Za regresní funkce volíme funkcionální regresi typu:

$$\hat{y}_i = b_0 + b_1F_1(x_i) + b_2F_2(x_i) + ...b_kF_k(x_i)$$

Jak už bylo zmíněno druhým důležitým úkolem je stanovit stupeň těsnosti korelačních vztahů neboli spolehlivost regresního odhadu. K tomu se využívá korelačního koeficientu u lineární závislosti a korelačního indexu u nelineární závislosti, jak uvidíme později.

2.3 Lineární regrese a korelace dvou proměnných

Jedním z typů korelace je lineární. Při němž předpokládáme že závislost y na x lze vyjádřit přímkou. Postup je takový že soubor bodu v grafu prokládá přímkou. Mezi ypsilonovými hodnotami měřených bodů a ypsilonovými hodnotami ležícími na přímce bude odchylka. Podstatou metody lineární regrese je nalezení takové přímky, aby součet druhých mocnin těchto odchylek byl co nejmenší. Dosáhnout cíle umožňuje aproximace daných hodnot přímkou pomoci metody nejmenších čtverců.

Lineární závislost dvou proměnných je daná vztahem ŷi = ayx + byxxi a podmínka pro vyrovnání je vyjádřená takto:

$$F(a_{yx}, b_{yx}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a_{yx} - b_{yx}x_i)^2 = \min.$$

Musíme stanovit vyhovující regresní parametry a_{yx} a b_{yx} proto bereme vztah podmínky za funkci těchto parametrů a budeme hledat její minimum. Provedeme parciální derivace funkci podle obou parametru a položíme rovnými nule:

$$\frac{\partial F(a_{yx}, b_{yx})}{\partial a_{yx}} = \frac{2}{n}\sum_{i=1}^{n}(y_i - a_{yx} - b_{yx}x_i)(-1) = 0$$

$$\frac{\partial F(a_{yx}, b_{yx})}{\partial b_{yx}} = \frac{2}{n}\sum(y_i - a_{yx} - b_{yx}x_i)(-x_i) = 0$$

Vztahy upravíme na normální rovnice:

$$\sum_{i=1}^{n} y_i = n.a_{yx} + b_{yx}\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i x_i = a_{yx}\sum_{i=1}^{n} x_i + b_{yx}\sum_{i=1}^{n} x_i^2$$

Soustavu normálních rovnic vyřešíme pomocí determinantu a dostaneme regresní koeficient b_{yx}, který vyjadřuje průměrnou změnu funkci při jednotkové změně proměnné x:

$$b_{yx} = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Tak že z vyřešené soustavy dostáváme parametr a_{yx}:

$$a_{yx} = \frac{\sum_{i=1}^{n} y_i}{n} - b_{yx}\frac{\sum_{i=1}^{n} x_i}{n} = \bar{Y} - b_{yx}\bar{X}$$

Další užitnou charakteristikou je kovariance, která popisuje jak se závislost chová. Je-li kovariance záporná jde o nepřímou lineární korelační závislost. To znamená, že pokud se zvětšují hodnoty závisle proměnné, budou se zmenšovat hodnoty nezávislé. Je-li kovariance vychází kladná jde o přímou závislost – zvětšení hodnoty závisle proměnné vyvolává zvětšení hodnoty nezávislé. Pokud ukazatel kovariance vyjde nulový budeme uvažovat o lineárně nekorelovaných veličinách nebo o nezávislosti veličin v případě jejích normálního rozdělení. Vztah pro vypočet hodnoty kovariance je:

$$s_{yx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) = \mathrm{cov}(x, y)$$

Z předchozího vzorce se odvozuje korelační koeficient, který nám udává stupeň těsnosti závislosti neboli spolehlivost regresního odhadu a pohybuje se

v mezích <-1;+1>. Vztah je velmi těsný, pokud je koeficient větší nebo roven 0,8; je střední v mezích od 0,3 do 0,8 a považuje se za slabý do 0,3. Vzorec pro výpočet:

$$r_{yx} = \frac{s_{yx}}{s_y s_x} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2\right].\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2\right]}}} =$$

$$\frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

2.4 Vícenásobná lineární korelace

V reálném světě však málo kdy máme systém v němž nějaký děj je závislý pouze na jednom faktoru, většinou závislé proměnné jsou ovlivněný celou sadou nezávislých faktorů. Proto se často používá vícenásobná korelace, která se zabývá prozkoumáním činitelů působících na závislou proměnnou a vyhledáním mezi nimi nejpodstatnějších. Přičemž i nelineární vztahy se da převést na lineární pomoci vhodné zvolené transformaci.

Regresní model vyžaduje splnění předpokladu, že závislá proměnná musí mít lineární korelaci s každou nezávislou proměnnou, které se zároveň navzájem nesmí korelovat:

$$\hat{y}_i = a_{y.12...k} + b_{y1.23...k}.x_1 + ... + b_{yk.12...k-1}.x_k$$

Odvození potřebných parametrů se provádí metodou nejmenších čtverců podobně jako u dvojnásobné lineární regresi ale s odpovídajícím počtem parciálních derivaci pro jednotlivé proměnné.

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \min$$

$$F(a_{y.12}; b_{y1.2}; b_{y2.1}) = \sum_{i=1}^{n}(y_i - a_{y.12} - b_{y1.2}x_1 - b_{y2.1}x_2)^2$$

$$\frac{\partial F}{\partial a_{y.12}} = \frac{\partial F}{\partial b_{y1.2}} = \frac{\partial F}{\partial b_{y2.1}} = 0$$

Po algebraických výpočtech dostáváme soustavu normálních rovnic:

$$\sum_{i=1}^{n} y = n.a_{y.12} + b_{y1.2}\sum_{i=1}^{n} x_1 + b_{y2.1}\sum_{i=1}^{n} x_2$$

$$\sum_{i=1}^{n} y.x_1 = a_{y.12}\sum_{i=1}^{n} x_1 + b_{y1.2}\sum_{i=1}^{n} x_1^2 + b_{y2.1}\sum_{i=1}^{n} x_1 x_2$$

$$\sum_{i=1}^{n} y.x_2 = a_{y.12}\sum_{i=1}^{n} x_2 + b_{y1.2}\sum_{i=1}^{n} x_1 x_2 + b_{y2.1}\sum_{i=1}^{n} x_2^2$$

Po vyřešení soustavy dostáváme dílčí regresní koeficienty, které udávají o kolik se změní závislá proměnná při jednotkově změně nezávislé. Koeficient $b_{y1.2}$ ukazuje změnu y při změně hodnoty x_1 a zároveň pozastavení hodnoty x_2:

$$b_{y1.2} = \frac{s_y}{s_1}.\frac{r_{y1} - r_{y2}.r_{12}}{1 - r_{12}^2}$$

$$b_{y2.1} = \frac{s_y}{s_2}.\frac{r_{y2} - r_{y1}.r_{12}}{1 - r_{12}^2}$$

Parametr $a_{y.12}$ se vypočítá vydělením první normální rovnice n:

$$a_{y.12} = \bar{y} - b_{y1.2}.\bar{x}_1 - b_{y2.1}.\bar{x}_2$$

Těsnost lineární závislosti y na ostatních dvou nezávisle proměnných se měří pomocí mnohonásobného korelačního koeficientu (Kožíšek, Stieberová, 2014):

$$I_{y.12} = r_{y.12} = \sqrt{1 - \frac{s_{y.12}^2}{s_y^2}} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2.r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}}$$

Při analýze vícerozměrného souboru dat je taky vhodné ověřit, jsou-li hodnoty na sebe závislé. Vhodným nástrojem k tomu slouží korelační matice, která má 1 na diagonále a korelační koeficienty mezi jednotlivými parametry. Koeficienty pohybují v rozmezích od -1 do 1. Střední závislost charakterizovaná pásmem od 0,3 do 0,7; silná závislost od 0,8; zbytek hodnot ukazuje na slabou či neexistující závislost hodnot. Výsledek je základem pro volbu správné metod vícerozměrné analýzy dat. Příklad matice:

	x1	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*
x1	1								
x2	-0,02396	1							
x3	-0,11957	0,019215	1						
x4	-0,12049	0,156509	0,326214	1					
x5	0,043213	0,000882	-0,0093	-0,00718	1				
x6	0,500983	-0,04338	-0,11319	-0,10385	0,070107	1			
x7	0,274278	-0,0154	-0,03527	-0,0325	0,030482	0,268635	1		
x8	0,134664	-0,00025	-0,00785	-0,00446	0,00598	0,142105	0,060105	1	
x9	0,032473	0,00767	0,014226	0,015104	-0,00444	0,026855	0,015014	0,005176	1

Tab. 3 Korelační matice

2.5 Analýza shluků

Analýza clusteru nebo CLU je metoda zabývající podrobným zkoumáním a tříděním objektu s velkým počtem naměřených hodnot. Metoda je například podrobně popsána v: Meloun a kol, 2017. Příslušnost objektu do tříd není známá. Od začátku počet shluků obvykle taky není známý, ale často se dá předpokládat. Nejvíc se používá v oblastech, kde objekty projevují přirozenou tendenci se seskupovat. Například se v biologie využívá ke klasifikaci živých organizmů. Takový postup se nazývá numerická taxonomie. V medicíně identifikuje nemoci a jejích stadia.

Cíle analýzy lze popsat v třech bodech:

- zjednodušení dat;
- popis systematiky – empirická klasifikace objektů;
- identifikaci vztahu – odhalení shluků i struktury objektu umožňuje stanovit vztahy mezi objekty.

Pro snadný proces a dosažení cílů musí byt stanovené vhodné znaky, na jejích základě se provádí charakterizování shlukových objektů. Znaky se voli z různých pohledů – teoretických, pojmových a z praktického hlediska. Nesprávná sada znaků může mít závažný dopad na výsledek celé analýzy.

Analýza clusteru je velmi citlivá na nevýznamné znaky a zároveň na extrémy mezi objekty. Za extrémy se považuji objekty, které se silně odlišuji od ostatních. Mohou to byt buď velké odchylky nebo patologické objekty (outlier), které nereprezentuji celek, ale jsou zvláštním případem. Oba dva typy zhoršuji strukturu dat a vyvolává chybnost nalezených shluku a jejích nereprezentativnost skutečné struktury. Přitom vyloučení extrémních objektu nese další riziko – jejích odstranění taky muže poškodit strukturu. Takový krok musí být velmi zvážen.

Osnovou metody však je podobnost objektů. Prvním krokem je stanovení znaku podobnosti, které se kombinuji do podobnostních měr. Potom můžeme objekty mezi sebou srovnávat a měřit jejích podobnost. K tomu se hodí tři základní skupiny metod: míry korelace, míry vzdálenosti, míry asociace. Každá skupina reprezentuje zvláštní pohled na podobnost, který je založen na typu objektu a dat. První dvě se využívá v případě metrických dat, poslední skupina metod je určená pro nemetrická data.

2.6 Míry podobnosti

V této kapitole rozebereme podrobněji jednotlivé skupiny metod pro měření podobnosti. První skupinu tvoří korelační míry popisující vztah mezi dvěma objekty či znaky x_i a y_i. V kardinální škále jejích podobnost představuje Pearsonův párový korelační koeficient R pro náhodné veličiny X a Y. Pohybuje se v mezích <-1,1>. Čím je koeficient větší a přibližuje se k jedničce tím je objektová podobnost silnější.

$$R(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

Analogickým koeficientem v ordinální škále je Spearmanův korelační koeficient, který vychází z transponované matice dat X^T. Matice je tvořena z objekty ve sloupcích a znaky v řádcích. Korelaci mezi dvojicí objektů představují napočtené

korelační koeficienty mezi dvěma sloupci. Hodnota koeficientů pohybuje v mezích <-1,1>. Pokud existuji dvě náhodné veličiny X a Y u kterých není, známe pravděpodobnostní rozdělení, můžeme uspořádat jejich hodnoty podle velikosti a pak hodnota Spearmánův koeficientu se rovna:

$$\rho = 1 - \frac{6\sum_{i}(p_i - q_i)^2}{n(n^2 - 1)}$$

kde n ... počet uspořádaných hodnot x_i, y_i

p_i, q_i ... pořadová čísla uspořádaných hodnot

Další skupinou jsou míry vzdálenosti a jsou nejčastěji využívanou. Objekty se nacházejí v prostoru souřadnic, které jsou tvořené jednotlivými znaky. Jak uvádí Milan Meloun a Jiří Militký jestli požadavky symetrie d(x, y) = d(y, x) a trojúhelníková nerovnost d(x, y) ⩽ d(x, z) + d(y, z) jsou splněné mírou vzdálenosti můžeme uvažovat o tzv. metriky. Nejčastější metrikou je geometrická neboli euklidovská vzdálenost, která je daná délkou přepony pravoúhlého trojúhelníku vypočtené pomoci Pythagorové věty:

$$d_E(x_k, x_l) = \sqrt{\sum_{j=1}^{n}(x_{ki} - \bar{X}_{lj})^2}$$

Ale samozřejmě má i svoje nevýhody – je citlivá k patologickým objektům v souboru dat. Zároveň se projevuje i citlivost ke škále měření u jednotlivých znaků. Například pokud máme znak Revenue v milionech, Počet zaměstnanců v rozmezí od 1 do 10 a ROA v desetinách je nejdřív potřeba provést normalizaci dat.

Existuji i jiné míry vzdáleností. Například se často užívá manhattanská vzdálenost (vzdálenost městských bloků) neboli Hammingová metrika. Podmínkou použití je to že shluky spolu nekorelují. Metrika je daná vztahem:

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

Metrika Minkovského u které zaleží na proměnné z. S navýšením parametru zvětšuje se rozdíl mezi vzdálenými objekty. Pokud je z = 1 jde o Hammingovou metriku, z = 2 o Euklidovou:

$$d_M(x_k, x_l) = \sqrt[z]{\sum_{j=1}^{m} | x_{kj} - x_{lj} |^z}$$

Taky se občas používá tětivová vzdálenost, zejména v oblasti ekologických průzkumu:

$$d_{CH}(x_k, x_l) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{m} x_{kj} x_{lj}}{\sum_{j=1}^{m} x_{kj}^2 \sum_{j=1}^{m} x_{lj}^2}\right)}$$

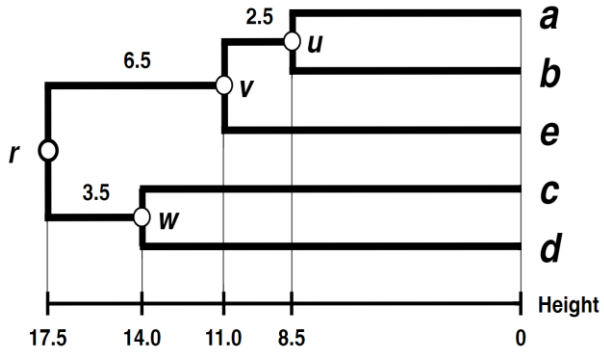
Poslední skupinou jsou míry asociace, které se používají při specifickém typu dat. Pokud máme soubor nemetrických dat, to znamená binárních, můžeme aplikovat metody posouzení asociace. Klasickým příkladem je dotazník typů Ano/Ne. Porovnání se provede procentem souhlasu nebo zamítnuti. Při složitějším úkolu s porovnáním více kategorii jak nominálních tak i ordinálních znaků se používají různé typy koeficientu asociace.

2.7 Metody shlukování

Clustery tvoří skupiny, ve kterých vzdálenost mezi jednotlivými elementy v nich je menší než do objektů, které do shluků nepatří. Postupy jak provádět shlukování se liší. Rozděluji se na hierarchické, které má podskupiny aglomeračního a divizního, a nehierarchické. Dále popíšu některé z nich a jejích aplikaci bude následovat v praktické části.

Nejdřív rozebereme hierarchické shlukováni. Má výhodu v tom, že nemusíme vědět od začátků optimální počet shluků a můžeme stanovit potřebné množství až na konci. Jak už bylo uvedeno hierarchické shlukování má dva odlišné postupy. První je aglomerační, který je postaven na myšlence „od jednotlivců k celku“. Objekty s nejmenší vzdálenosti se spojují do shluku, které teď budou považovány

za objekty, a vypočte se nová matice vzdáleností. Znova se provede sjednocení do shluku a bude se postup opakovat, dokud nezůstane buď předem požadovaný počet clusterů, nebo jeden velký shluk, obsahující všechny objekty. Druhý postup je divizní shlukování a provádí se obráceně – „od celků k jednotlivcům“. Na začátku je jeden shluk se všemi objekty a postupně se děli na menší clustery, dokud nedojdeme do samotných objektů. S provedením procedury shlukování souvisí i volba metody metriky, existuje celá řada takových: metoda nejbližšího souseda, nejvzdálenějšího souseda, průměrné vzdálenosti, Wardova metoda, metoda těžiště, mediánová. Grafickým výstupem je dendrogram (vývojový strom), který má na jedné ose jednotlivé objekty a na druhé vzdálenost mezi nimi, zároveň graf ukazuje, jak se objekty postupně seskupují do clusterů.



Obr. č. 1 – Dendrogram

Nehierarchické shlukováni je postaveno na volbě příkladových objektů. Uživatel na základě svých požadavku a znalosti musí předem stanovit řadu takových objektů, které budou tvořit zaklad nových shluku. Ostatní objekty budou rozdělené do shluku metodou podle jejích vzdálenosti od řady „typických“. Existuje několik postupů toho rozdělení (Meloun a kol, 2017):

- sekvenční práh – provádí se postupné shlukování. Zvoli se jeden typický objekt a vzdálenost, ostatní objekty kolem v jejím dosahu budou zařazený do jednoho shluku a dále s nimi už nebudeme počítat. Pak se zvolí nový typický objekt a vzdálenost;
- paralelní práh – odlišuje se od sekvenčního tím, že hned se stanovuje několik typických objektu a provádí se shlukování. Mohou zůstat nezařazené objekty;

- optimalizace – má stejný postup jako u předchozích dvou, ale dovoluje znovuzařazení objektů. Pokud by se objevilo, že už zařazený do shluku objekt má menší vzdálenost do jiného shluku, bude tam přemístěn.

3. Praktická část. Provedení analýzy zákazníků.

3.1 Popis firmy

Tato diplomová práce byla vytvořena ve spolupráci s telekomunikační společnosti působící na českém trhů. Do širokého portfolia služeb, které firma nabízí, patři: hlasové mobilní služby, mobilní data, pevné internet připojení, služby TV a celou řadu softwarových řešení.

Okolí podniku tvoří malá řada konkurenčních firem, což je standartní situace na trzích telekomunikačního odvětví. Dynamika trhu je více méně stabilní a spočívá v tom, že zákazníci často mění dodavatele. Důvodem je skoro stejná technická a technologická úroveň firem, podobnost služeb a cenových nabídek.

Trh je segmentován na dvě velké skupiny – firemní a nefiremní zákazníky. Dalším bodem odlišnosti klientů je typ smlouvy – se závazkem nebo bez závazků. A podsegmenty se tvoří výší revenue, které zákazník přináší. Oslovení potenciálních a stávajících klientů se provádí telefonický (call-centra), online přes internet a osobně (sales representatives).

3.2 Úvod do problematiky trhu telekomunikaci

Telekomunikace je základem pro digitální ekonomiku moderního světa. Zabezpečuje informační tok, potřebný pro přijetí rozhodnutí jak v byznysu, tak i na úrovni státu. Je osnovou pro rozvoj klíčových oblastí v ekonomice jako například: obchod, finance, média, pojištění, vzdělávání atd.

Díky významnému technologickému pokroku posledních par desítek let trh telekomunikaci je jedním z nejrychleji se rozvijících sektorů světové ekonomiky. V roce 2019 jeho objem byl kolem 1,6 trilionu dolarů. Největší položkou trhů je

mobilní sítě – 50%. Prognóza růstu mobilního segmentu v nejbližší době je 2% ročně. Zároveň se předpokládá snížení ceny hlasových služeb a SMS ale narůst objemu dat, počtu M2M, FinTech a MFS servisů (IDC, 2018).

Tempo růstu abonentů mobilních sítí se však snižuje v posledních letech, což znamená přesycení segmentů. Ale zvětšuje se spotřeba dat, která je způsobena rozvojem technologii 3G, 4G a 5G. Zároveň roste počet domácností, které využívají pevný internet. Během posledních deseti let se jejích počet zvětšil dva krát (ITU World Telecommunication, 2018).

Existuje i řada negativních faktorů ovlivňujících trh telekomunikace, které snižuje rozvoj (Sinitsa, 2019):

- oligopol – ve většině zemí existuje jen par firem, působících v danem segmentu. Snižuje se konkurenceschopnost;
- legislativa – velmi přísná regulace státem;
- loajálnost – uživatele rychle a jednoduše mění operátora, obzvlášť kvůli sezonním nabídkám a slevám;
- nepřesnost ekonomických ukazatelů – chyby v propočtech kvůli neaktivním abonentům zhoršuje výsledky;
- vysoké náklady – zavedení nových technologii jako 5G a rozšíření pokrytí stoji obrovské peníze;
- ekonomické a sociální faktory – zpomalení ekonomického růstu rozvojových zemí, nízká úroveň životu, klimatické podmínky, nepřítomnost potřebných zákonů.

3.3 Struktura a analýza souboru dat

Data poskytnutá ke zpracování obsahují anonymizované údaje 80 057 firemních zákazníků. Obsah je tvořen jenom aktivními zákazníky. Pro analýzu byli zvolený následující osm charakteristik klientů:

- počet SIM karet
- průměrné revenue – v korunách za rok;
- průměrné náklady – v korunách za rok. Primárně jsou tvořené voláním přes konkurenční sítě;
- průměrné výnosy z konkurence – v korunách za rok. Vytváří se využitím našich sítí konkurenčními firmami;
- počet stížností – za rok. Každý klient má právo nahlásit problémy se službami pomocí online supportu nebo přes volání na speciální linků;
- součet využitých hlasových služeb – v hodinách za rok. Normálně se uvádí v sekundách, ale hodnoty byli převedené pro usnadnění využití shlukové analýzy;
- součet spotřebovaných dat – v gigabajtech za rok;
- platební morálka – ukazuje kolikrát měl klient fakturu po splatnosti za poslední rok.

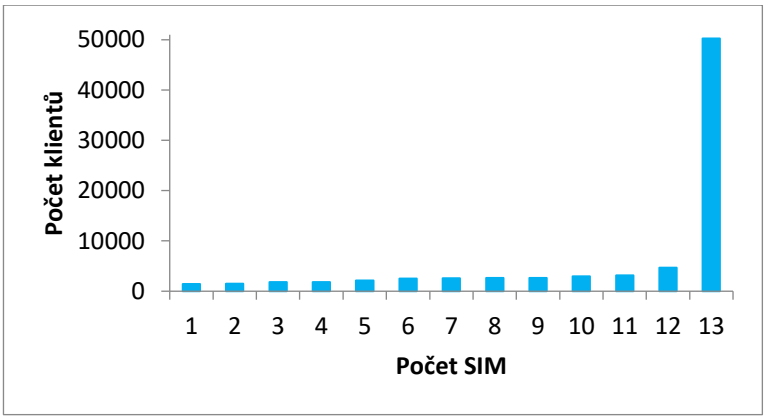
	Počet SIM	Prům. revenue	Prům. náklady	Prům. výnosy z konkurence
Průměr	10,98	1 325,56	224,65	190,29
Median	13,00	838,09	110,35	106,90
Směrodatná odchylka	3,34	1 900,19	659,01	291,72

Tab. 4 Popisné statistiky I

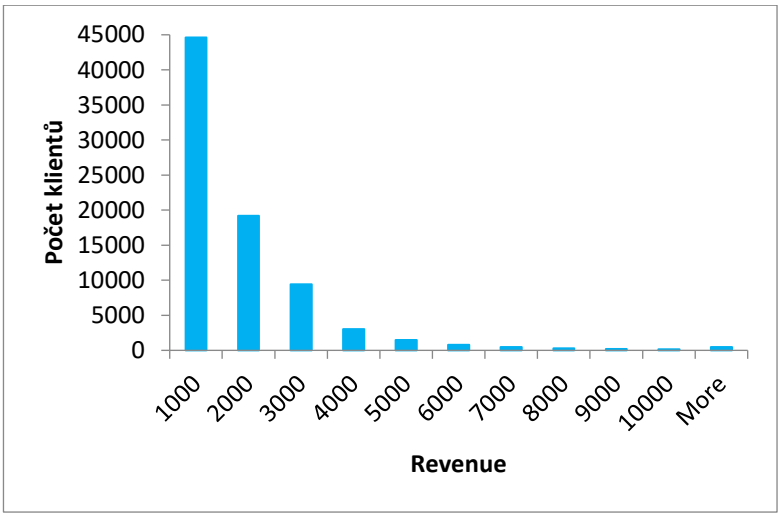
	Počet stižnosti	Hlasové služby	Data	Platební morálka
Průměr	0,03	266,40	168,42	0,87
Median	-	138,39	23,34	-
Směrodatná odchylka	0,21	422,28	547,17	6,69

Tab. 5 Popisné statistiky II

Z grafu č. 6 plyne, že zákazníci většinou mají vetší počet SIM karet což na první pohled potvrzuje i graf č. 7 – 55% klientů přináší revenue do 1 000 korun. V další kapitole ověřím jak je výše výnosů závisle na počtu SIM.

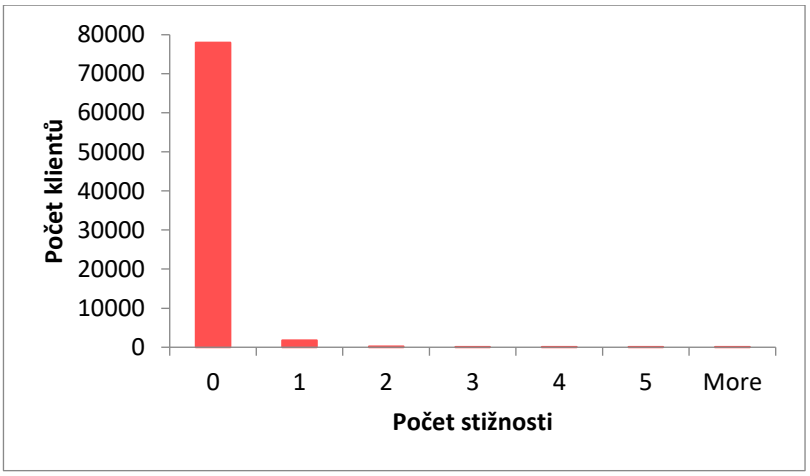


Graf č. 6 – Rozdělení zákazníků podle počtu SIM karet



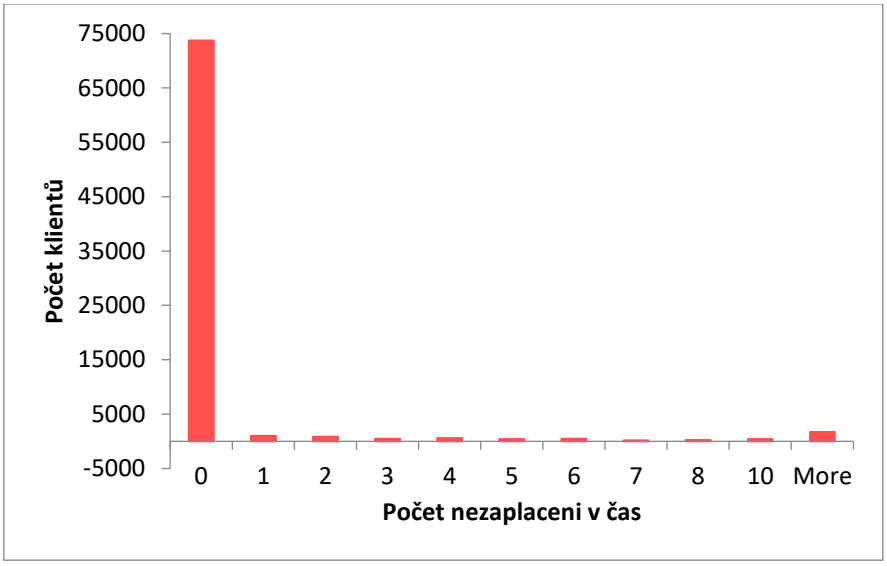
Graf č. 7 – Rozdělení zákazníků podle velikosti revenue

Z histogramu č. 8 je patrné že drtivá většina klientů nemá stížností po dobu roku. Vidím tři varianty proč to tak je: buď zákazníci nevědí jak sdílet svoje názory ohledně služeb a produktů, nebo neprovádí se pořádné monitorování, anebo nabízený servis je natolik kvalitní že většina uživatelů se s nim úplně spokojená. Poslední varianta je podle mého názoru je příliš pozitivní a proto bych spíš doporučil ověření prvních dvou nápadů. Zpětná vazba a znalost proč zákazník nespokojen jsou základem pro modernizaci a optimalizaci.



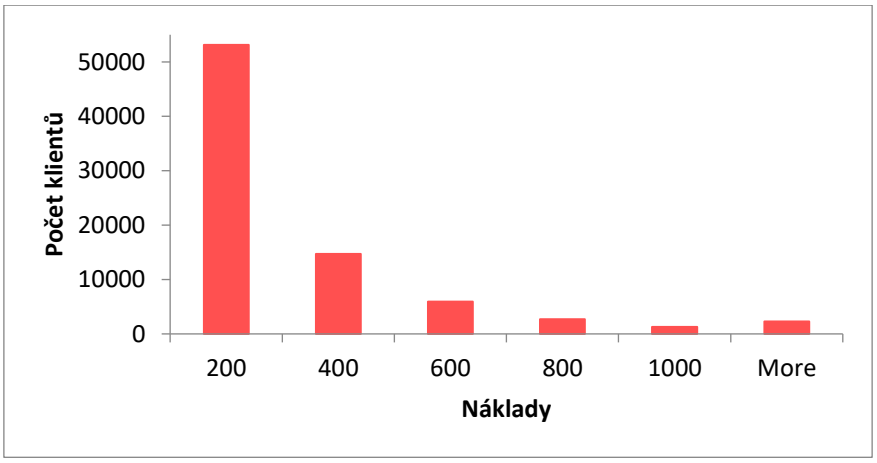
Graf č. 8 – Rozdělení zákazníků podle jejich stížnosti

Graf č. 9 udává počet faktur nezaplacených včas, ze kterého lze vidět, že platební morálka klientů je na dost vysoké úrovni. Jen 3% z celé databázi byli po splatnosti aspoň jednou během roku. Ukazuje to správnost nastavení platebních podmínek. Má tohle i ekonomický význam – neměl by podnik mít problém s Cash-flow.

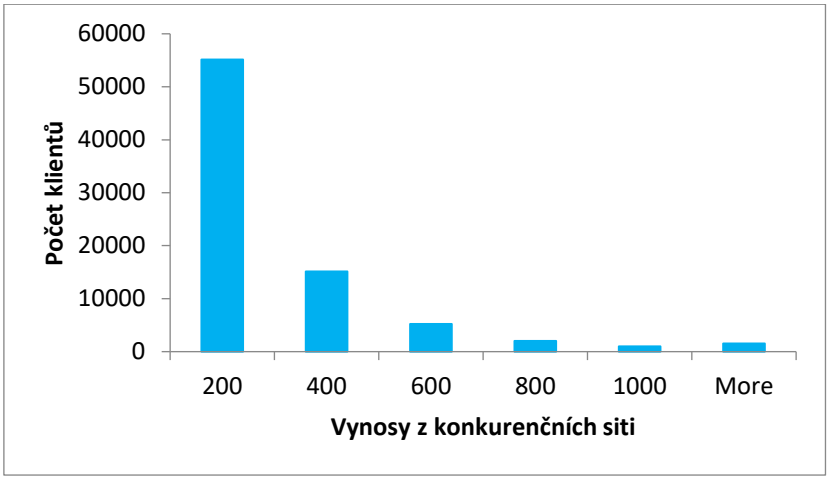


Graf č. 9 – Rozdělení zákazníků podle jejich platební morálky

Z následujících dvou grafů lze vidět, že náklady na klienta v telekomunikační sféře jsou minimální, ale nejsou úplně zanedbatelný s důvodu velkého počtu klientů. Graf č. 10 udává informaci o tom jak moc klienti provolávají přes konkurenční sítě a tím vytvářejí náklad ve smyslů poplatků. Další histogram ukazuje obracený poplatek – voláni konkurence do našich sítí.

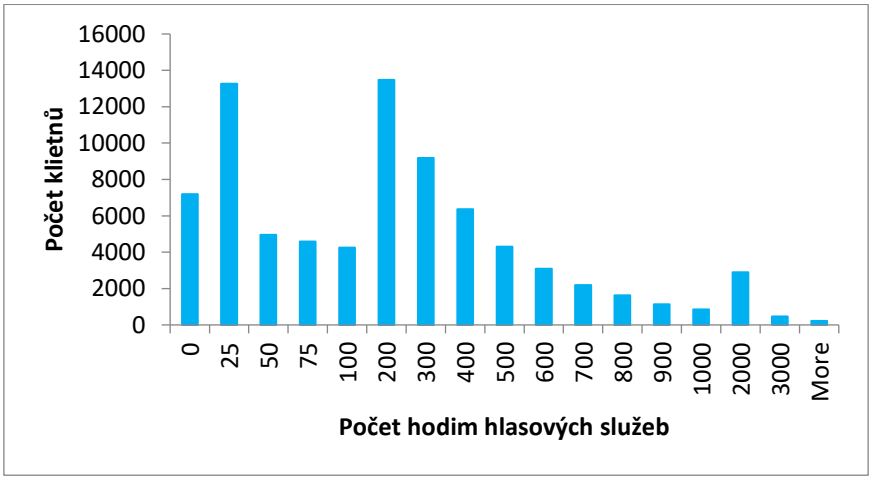


Graf č. 10 – Rozdělení zákazníků podle nákladů

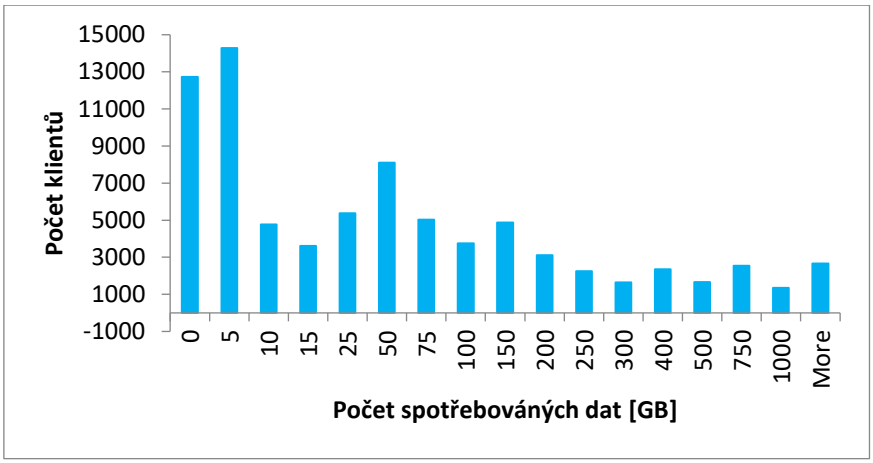


Graf č. 11 – Rozdělení zákazníků podle poplatků od konkurence

Poslední dva histogramy zobrazují spotřebu hlasových a datových služeb za rok.



Graf č. 12 – Rozdělení zákazníků podle hlasové spotřeby



Graf č. 13 – Rozdělení zákazníků podlé datové spotřeby

3.4 Regresní a korelační analýza

Na začátku jsem vypočetl korelační matici pro všechny hodnoty pomocí MS Excel. Pro všechny hodnoty se počítala lineární závislost. Z výstupu lze vidět, že hodnoty mezi sebou korelují. Například můžeme vidět, jaký vztah má počet SIM karet s využitím hlasových služeb a velikostí výnosů z konkurence (koeficienty jsou > 0,2). Čím víc klient má SIM karet tím víc provolává. Zároveň revenue má silnou závislost na hlasových, datových službách, nákladech a výnosech z provolání – čím víc volá a spotřebovává data, tím větší náklady tvoři a zároveň se zvětšuje pravděpodobnost volání k němu přes konkurenční sítě, což se projevuje ve výnosech z konkurence.

	Počet SIM	*Prům Revenue*	*Prům náklady*	*Prům vynosy z konkurenci*	*Stižnosti*	*Hlasové služby*	*Datové služby*	*Platební morálka*
Počet SIM	1							
Prům Revenue	0,196409962	1						
Prům náklady	0,119339815	0,667720988	1					
Prům vynosy z konkurenci	0,223209649	0,622872816	0,509662477	1				
Stižnosti	0,050545119	0,059901827	0,031876183	0,062169385	1			
Hlasové služby	0,296637123	0,6477515	0,406196618	0,776514955	0,071319064	1		
Datové služby	0,142030124	0,305775739	0,131107193	0,237570521	0,058166043	0,270607214	1	
Platební morálka	0,017343528	0,209238178	0,070418386	0,139587585	0,034754081	0,137204502	0,097329445	1

Tabulka 6 – Korelační matice souboru

K podrobnému zkoumání závislostí proměnných jsem zvolil 2 případy. Oba dva jsou spojené s revenue, protože je to nejzajímavější položkou pro jakýkoliv podnik. V první variantě byla ověřena závislost mezi revenue a velikosti spotřeby datových a hlasových služeb. Prvním výstupem je korelační matice a regresní

tabulka. Z nich patrné že závislost revenue na spotřebě služeb je. Důležitou informací je taky že významnost hlasových služeb je dvakrát větší než datových – to znamená že „hlas“ tvoří větší výnos:

	Average of revn_amt	Sum of voice_dur	Sum of data_vol_GB
Average of revn_amt	1,000	,648	,306
Sum of voice_dur	,648	1,000	,271
Sum of data_vol_GB	,306	,271	1,000

Obrázek 2 – Korelační matice Revenue-Služby

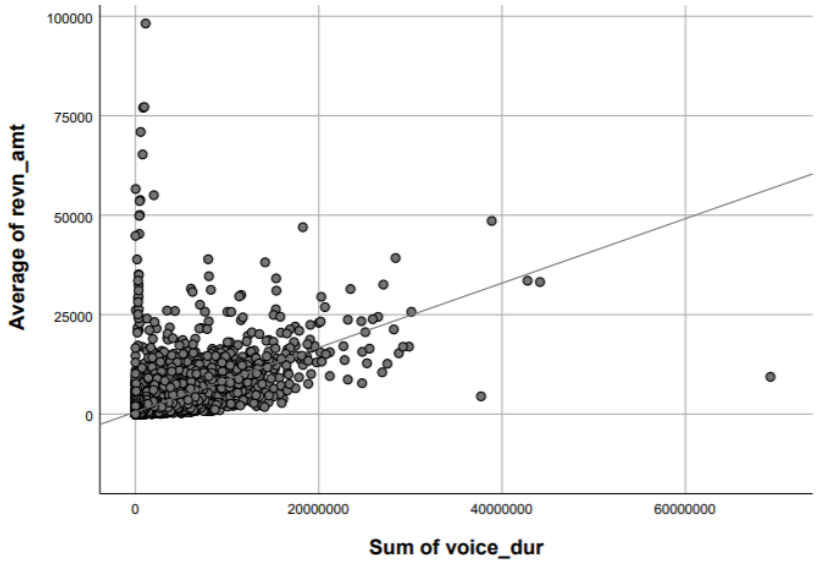
Koeficienty[a]

Model		Nestandartizované koeficienty		Standartizované koeficienty			Statistika kolineárnosti	
		B	Chyba	Beta	t	Význam	Odchylka	VIF
1	(Константа)	512,369	5,996		85,449	0,000		
	Sum of voice_dur	0,001	0,000	0,610	221,498	0,000	0,927	1,079
	Sum of data_vol_GB	0,489	0,010	0,141	51,156	0,000	0,927	1,079

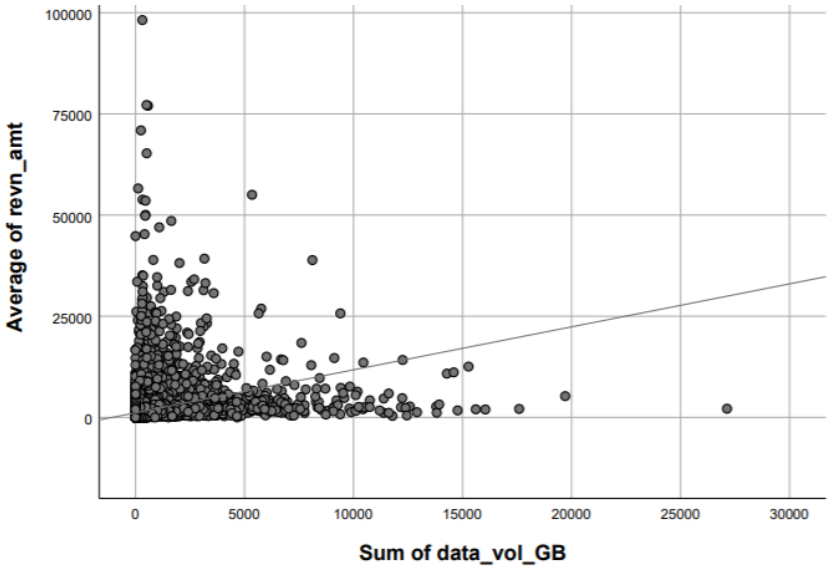
a. Závislá proměnná: Average of revn_amt

Obrázek 3 – Regresní tabulka Revenue-Služby

Druhým výstupem jsou grafy závislosti proložené přímkou:



Obrázek 4 – Regrese a korelace Revenue-Hlas



Obrázek 5 – Regrese a korelace Revenue-Data

Druhou variantou, kterou jsem ověřoval je závislost Revenue na počtu SIM karet, kterou potvrzuje vypočtený koeficient Beta:

	Average of revn_amt	Count of cnt_act_subsc_per_cust
Average of revn_amt	1,000	0,196
Count of cnt_act_subsc_per_cust	0,196	1,000

Obrázek 6 – Korelační matice Revenue-SIM

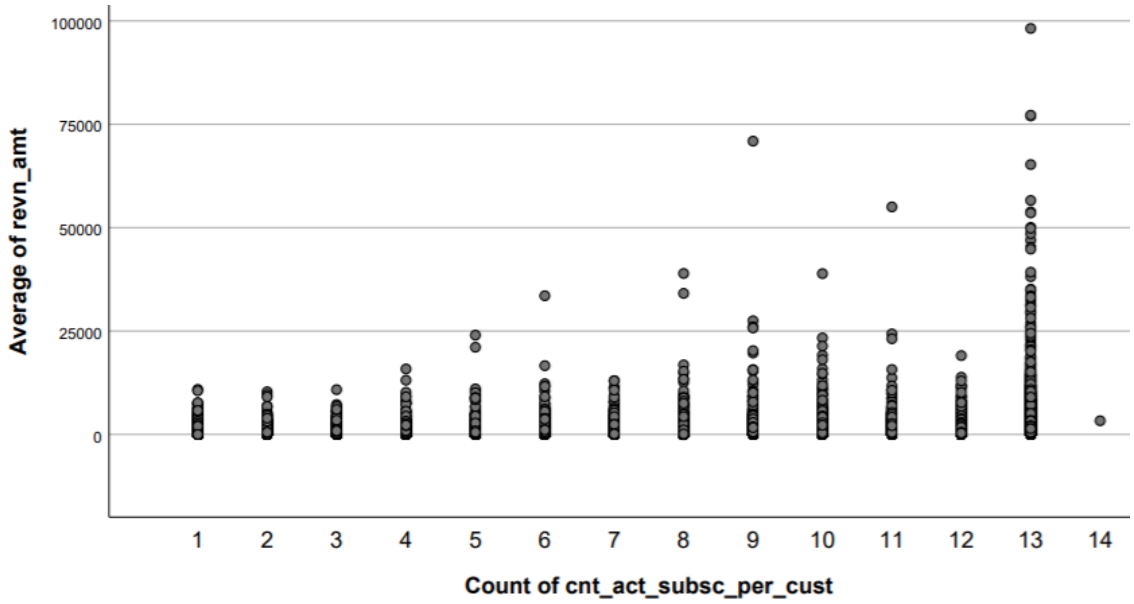
Koeficienty[a]

Model		Nestandartizované koeficienty		Standartizované koeficienty			Statistika kolineárnosti	
		B	Chyby	Beta	t	Významnost	Odchylka	VIF
1	(Константа)	97,851	22,641		4,322	0,000		
	Count of cnt_act_subsc_per_cust	111,811	1,973	0,196	56,676	0,000	1,000	1,000

a. Závislá proměnná: Average of revn_amt

Obrázek 7 – Regresní tabulka Revenue-SIM

Mám k dispozici i grafický výstup této analýzy, na kterém lze vidět že s rostoucím počtem SIM u klienta vzrůstává i revenue:



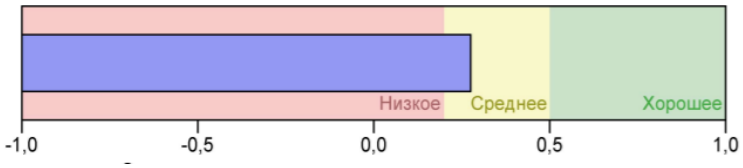
Graf č. 16 – Regrese a korelace Revenue-SIM

Z předchozích grafů a histogramů lze vidět, že soubor dat obsahuje spousta extrémních hodnot, o kterých bohužel nevíme, jestli jde o chybu nebo o nestandartní výkyvy. Takové údaje skreslují výstupy analýzy. Mnou byl vyzkoušen oříznutý soubor a v něm se lip projevují korelační závislostí, přesněji se stanovuje významnost kritérií. Myslím si že ve velkých firmách s dlouholetou historií datové systémy nesou v sebe větší riziko výskytu chyb a musíme se s tím počítat. Nicméně jsem rozhodl nechat soubor v původní podobě, protože jistota toho že extrémy jsou chybné nebyla dostatečně velká.

3.5 Shluková analýza

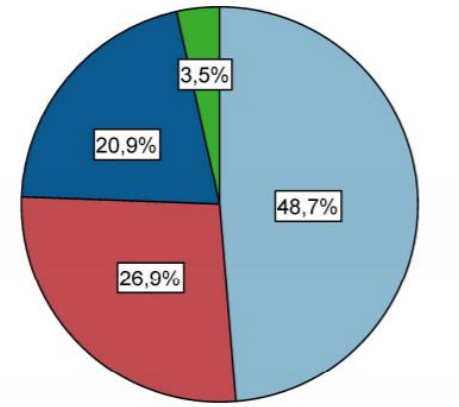
Měl jsem k dispozici data 80 000 klientů. To jsou elementy, které mají tendenci seskupovat sami o sobě. Proto jsem použil clusterovou analýzu pro jejích segmentaci. Výpočet se prováděl pomoci softwarového nástroje SPSS Statistics od firmy IBM. Shlukování bylo provedeno dvoustupňovou metodou z toho důvodu, že hierarchické shlukování nelze provést při takovém počtů záznamů z technických důvodů.

Vhledem k velikosti souboru a výskytu extrémních hodnot, analýza dostala do žlutého pásma – střední kvality modelu:



Obrázek 8 – Kvalita modelu (červená – nízká, žlutá – střední, zelená – vysoká)

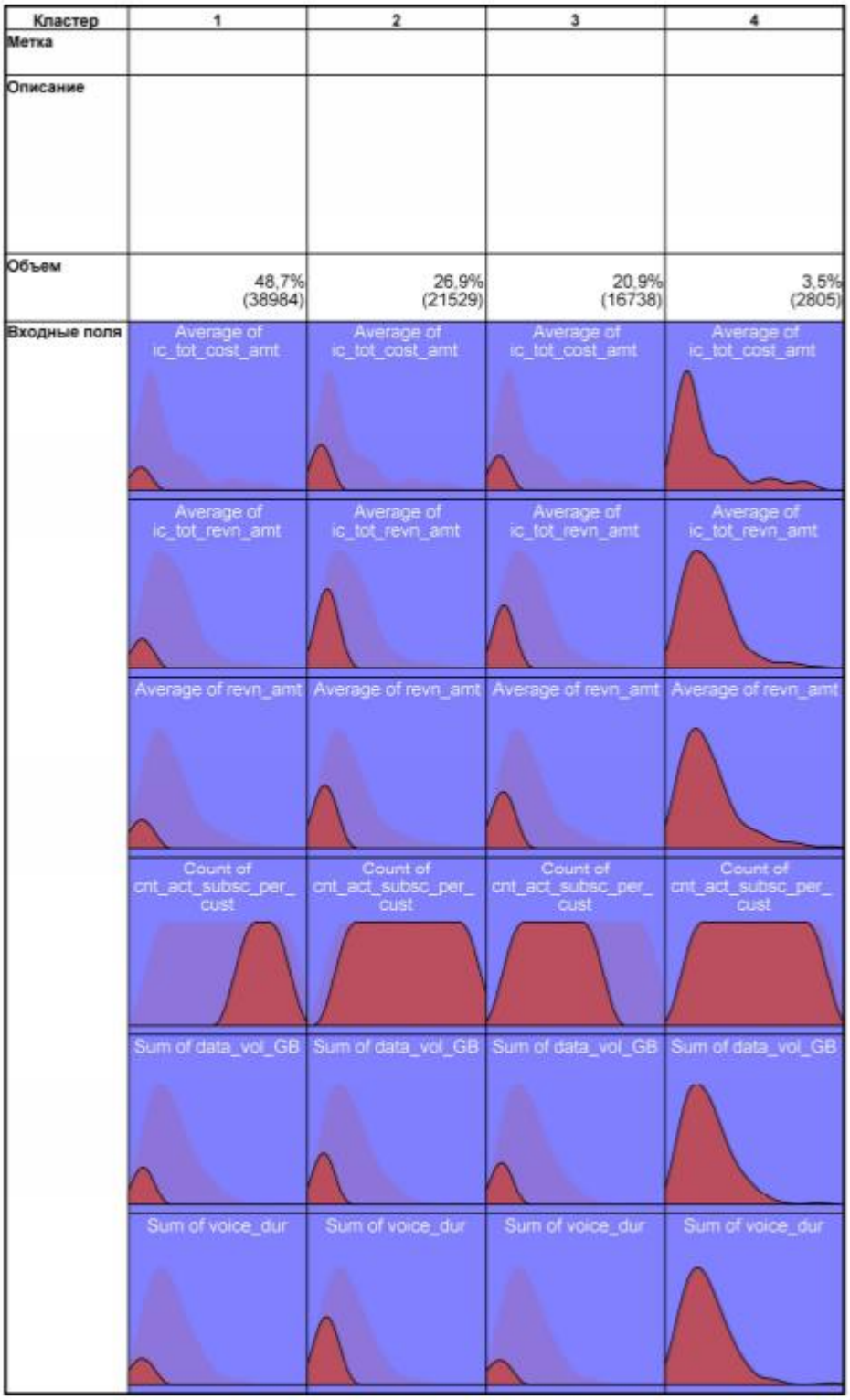
Na výstup jsem dostal čtyři clustery. Velikost každého segmentu považuju za dostatečnou, jsou to významné části celku:



Obrázek 9 – Poměr clusteru

Кластер	1	2	3	4
Метка				
Описание				
Объем	48,7% (38984)	26,9% (21529)	20,9% (16738)	3,5% (2805)
Входные поля	Average of ic_tot_cost_amt 96,82	Average of ic_tot_cost_amt 418,63	Average of ic_tot_cost_amt 57,94	Average of ic_tot_cost_amt 1 507,24
	Average of ic_tot_revn_amt 90,72	Average of ic_tot_revn_amt 367,54	Average of ic_tot_revn_amt 53,82	Average of ic_tot_revn_amt 1 027,98
	Average of revn_amt 734,83	Average of revn_amt 2 324,77	Average of revn_amt 539,25	Average of revn_amt 6 558,31
	Count of cnt_act_subsc_per_cust 12,43	Count of cnt_act_subsc_per_cust 12,69	Count of cnt_act_subsc_per_cust 5,15	Count of cnt_act_subsc_per_cust 12,48
	Sum of data_vol_GB 53,66	Sum of data_vol_GB 298,64	Sum of data_vol_GB 21,95	Sum of data_vol_GB 1 637,94
	Sum of voice_dur 453 510,14	Sum of voice_dur 1 973 591,10	Sum of voice_dur 116 300,92	Sum of voice_dur 5 227 056,69

Obrázek 10 – Charakteristiky clusteru



Obrázek 10 – Absolutní rozdělení hodnot v jednotlivých clusterech

Popis charakteristik jednotlivých segmentů od nejmenšího k největšímu:

- Cluster č. 4 – „otazníky“. Objem je 3,5% z celkového souboru. Z jeho středních hodnot vidíme, že obsahuje všechny zákazníky s extrémními hodnotami. Tento cluster bych doporučil k prověření, zda jsou hodnoty reálné nebo v systému se začínají hromadit chyby;

- Cluster č. 3 – „malý segment“. Má v sebe 20,9% od celkového souboru. Má nejmenší střední hodnoty ze všech. To znamená, že jeho obsahem jsou menší zákaznici. Vzhledem k velikosti doporučuji stanovení zvláštní strategii nacílené na malé firmy;

- Cluster č. 2 – „velký segment“. Činí 26,9% od celkového souboru. Jeho střední hodnoty neprůměrné vysoké, ale realistické. Pravděpodobně jsou to největší klienty z portfolia. Pro tento segment doporučuji provést analýzu podobnosti – co preferují, v čem jsou stejné. Na její základě musí být stanovená strategie na udržení stávajících zákazníků a zároveň obchodní podmínky pro získání nových velkých klientů;

- Cluster č. 1 – „střední segment“. Je největší ze všech. Zahrnuje v sebe skoro polovinu portfolia – 48,7%. Vykazuje střední hodnoty a vzhledem k velikosti můžeme říct, že je to segment „běžných“ zákazníků. Z jeho podrobnější analýzy by šlo zjistit, které služby jsou nejvíc chtěny a pokusit se odvést trendy budoucích požadavků.

4. Závěr

Závěrem své diplomové prací bych rád sjednotil jednotlivé kapitoly a výsledky provedených výpočtů.

V první, teoretické části byla rozebrána marketingová problematika segmentace trhu a řízení vztahu se zákazníkem. Následně byli rozebrané statistické metody stanovení závislosti a vícerozměrné analýzy dat pomoci různých metrik.

V praktické části této metody byli použité. Nejprve byl analyzován vstupní soubor dat pomoci měr polohy a měnlivosti. Následným krokem bylo stanovení závislosti mezi parametry a dospělo k prvním výsledkům – které služby tvoří největší revenue, byli stanovené závislé a nezávislé proměnné. Na konci byla provedená shluková analýza, kterou byli stanovené a popsané včetně doporučení čtyři segmenty.

Výsledná řešení, modely, komentáře a doporučení byly předložené telekomunikační společnosti k posouzení a dalšímu využití v interních procesech.

5. Seznam literatury

ZAHRADNÍK JAROSLAV. *Management podniku.* 2003. Praha : ČVUT, ISBN 80-01-02724-4

FOTR J., ŠVECOVÁ L. A KOL. *Manažerské rozhodování.* 2010. 2, Praha: Ekopress, ISBN 978-80-86929-59-II.

MILAN MELOUN, JIŘÍ MILITKÝ, MARTIN HILL. *Statistická analýza vícerozměrných dat v příkladech.* 2017. Univerzita Karlová, ISBN 978-80-246-3618-4

KOŽÍŠEK JAN, STIEBEROVÁ BARBORA, VANIŠ LADISLAV. *Statistická a rozhodovací analýza*. 2008. Praha: Česká technika.

CHLEBOVSÝ VÍT. *CRM řízení vztahů se zákazníky.* 2005. Brno: Computer Press, a.s. ISBN 80-251-0798-1

Hodnota klienta. *Clientbrige.* [online] Dostupné z: *https://www.clientbridge.ru/blog/opredelite-realnuyu-tsennost-klienta/*

Worldwide Spending on Telecommunications Services and Pay TV to Speed Up Slightly in 2018, According to IDC. *International Data Corporation's*. [online] Dostupné z: *https://www.idc.com/getdoc.jsp?containerId=prUS43809518*

ICT Indicators database. *ITU World Telecommunication*. [online] Dostupné z: *http://www.itu.int/en/ITU-D/Statistics/Pages/definitions/regions.aspx*

SINITSA S.A. *Analysis of trends in the global telecommunications services market*, 2019. The Eurasian Scientific Journal. [online] Dostupné z: *https://esj.today/PDF/27ECVN119.pdf*

6. Seznam obrázků

7. Seznam tabulek

8. Seznam grafů