Sem vložte zadání Vaší práce.

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra teoretické informatiky

Diplomová práce

Analytický framework nad interakčními daty z eshopů

Bc. Lukáš Dvořák

Vedoucí práce: Ing. Pavel Kordík, Ph.D.

1. května 2016

Poděkování

Rád bych touto cestou poděkoval všem, kteří se podíleli svými návrhy a připomínkami na podobě této práce - jmenovitě panu doktoru Kordíkovi za ochotu při jejím vedení a panu inženýru Řehořkovi za jeho pomoc a cenné rady.

Dále bych také rád poděkoval představitelům všech společností, které umožnily poskytnutí dat potřebných pro provedení případových studií, za vstřícnost a projevenou důvěru.

V neposlední řadě děkuji své rodině a přátelům za jejich podporu nejen během celého mého studia a Jaroslavu Hlinkovi za motivaci poskytnutou během psaní této práce.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů, zejména skutečnost, že České vysoké učení technické v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne 1. května 2016 .....

Odkaz na tuto práci

Dvořák, Lukáš. *Analytický framework nad interakčními daty z eshopů.* Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2016.

Abstrakt

Tato diplomová práce se zabývá specifiky e-shopů a průzkumem vhodných metod pro provádění analýz nad transakčními daty. Výsledkem je návrh analytického frameworku pro účely reportování majitelům transakčních a prodejních dat, který je možné nasadit pro využití v e-shopech libovolného typu, velikosti a nezávisle na implementaci. Součástí práce jsou dvě případové studie aplikující navrhovaný framework nad daty z reálných e-shopů.

Klíčová slova E-shopy, Segmentace, RFM analýza, CLV, Reportování

Abstract

This final thesis deals with identifying a specificity of e-shops and includes a research on suitable methods for analysing e-shop transactional data. The outcome is a design of an analytical framework for reporting transactional and sales data to the owners, that is able to be applied to e-shops of an arbitrary type, size and independently of implementation. Two real-world case studies using the framework design are part of the paper.

Keywords E-shops, Segmentation, RFM analyses, CLV, Reporting

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Význam oblasti internetového obchodování a elektronického poskytování služeb, která se souhrnně označuje jako *e-commerce*, každoročně stoupá a v současnosti už se jedná o přirozenou a pevně přijímanou součást trhu jak v obchodních vztazích typu *B2B* (*business to busines*), tak retailovém obchodování typu *B2C* (*business to customer*).

Stejně jako mezi klasickými obchody je v oblasti těch internetových velká rozmanitost z pohledu velikosti, zaměření, typu prodávaného zboží apod. Nicméně existují jistá specifika e-shopů z důvodů plynoucích právě ze způsobu nabízení zboží a komunikace se zákazníky internetovými kanály.

Tato práce má za cíl identifikovat specifika internetových obchodů a prozkoumat vhodné metody pro provádění analýz nad transakčními daty e-shopů. Výsledkem je návrh analytického frameworku pro účely reportování majitelům transakčních a prodejních dat, který je možné nasadit pro využití v e-shopech libovolného typu, velikosti a nezávisle na implementaci.

Kapitola 1 této práce představuje tzv. *RFM analýzu* a teoreticky popisuje možnosti jejího využití v kombinaci s dalšími postupy dolování dat a strojového učení, především v souvislosti s úlohou shlukování a asociačními pravidly.

Ve druhé kapitole je definován pojem *dlouhodobé hodnoty zákazníků* (*CLV*) a teoreticky popsány dva modely pro její odhadování.

Třetí kapitola se věnuje samotnému popisu navrženého frameworku na základě zkoumaných specifik týkajících se e-shopů, stanovuje požadavky nutné pro jeho nasazení a především popisuje jednotlivé kroky analýz, navržených na základě teoretických oblastí popsaných v kapitolách 1 a 2.

Poslední, nedílnou součástí této práce jsou dvě případové studie nasazení navrhovaného frameworku nad transakčními daty reálných internetových obchodů. Popisy konkrétního provádění analýz, interpretace výsledků a závěry včetně návrhů obchodních doporučení jsou součástí kapitoly 4.

Kapitola 1

RFM analýza

1.1 Definice a základní principy

Takzvané RFM – *Recency*, *Frequency* a *Monetary* – jsou po desítky let velice hojně využívané metriky pro kvantifikaci transakční historie zákazníků. Mezi hlavní výhody použití těchto metrik je vysoká vypovídací hodnota, snadná interpretace a v neposlední řadě vysoká dostupnost při extrakci dat z transakčních databází bez znalosti konkrétních dat o zákaznících.

Definice jednotlivých metrik podle (*Blattberg, 2008*)[1]:

- *Recency* – udává čas od poslední transakce provedené zákazníkem
- *Frequency* – značí počet transakcí za dané časové období
- *Monetary* – reprezentuje souhrnnou hodnotu transakcí provedenou zákazníkem

RFM analýza využívá model na základě výše uvedených metrik ke klasifikaci nebo scoringu zákaznických segmentů v rámci klientské základny [1].

Jedná se o velice populární framework, který byl v minulosti mnohokrát zkoumán a opakovaně byla potvrzována jeho velká použitelnost, ať už v oblasti direct mailingu nebo v kombinaci s pokročilými metodami pro klasifikaci a segmentaci zákazníků nebo predikci jejich tržní hodnoty. Viz např. (*Hughes, 1994*)[2], (*Bult, Wansbeek, 1995*)[3], (*McCarty, Hastak, 2006*)[4], (*Cheng, Chen, 2009*)[5] nebo (*Hu, Yeh, 2014*)[6].

V rámci RFM analýzy se nejprve provádí přiřazení diskrétních hodnot každému zákazníkovi z hlediska každé ze tří použitých metrik. Tímto způsobem každý zákazník obdrží trojici ohodnocení, která ho reprezentuje v třídimenzionálním prostoru, který model využívá.

(*Miglautsch, 2000*)[7] popisuje základní přístup k ohodnocení následujícím způsobem. Nejprve je pro každého zákazníka vypočítána hodnota jeho

recency jako počet měsíců od posledního nákupu, *frequency*, jako počet nákupů za dané období a *monetary*, jako součet tržeb z nákupů za dané období. Následně jsou zákazníci seřazeni vzestupně podle hodnoty *recency* a rozděleni do kvantilů s ohodnoceními 1 až 5 (5 pro zákazníky s nejnižšími hodnotami). Obdobně se postupuje pro zbývající dvě metriky, avšak zákazníci jsou řazeni sestupně a kvantily s ohodnocením pět obsahují zákazníky s nejvyššími hodnotami *frequency*, resp. *monetary*. Následně je výsledné RFM ohodnocení každého zákazníka určeno jako konkatenace jednotlivých ohodnocení R, F, M.

Výsledkem je 125 různých kombinací, které tvoří jednotlivé segmenty zákazníků se stejným ohodnocením v rámci každého segmentu. Obchodní významnost nebo například predikce respondibility je potom určována dále pro každý segment zvlášť.

1.2 RFM s expertním rozdělením

Nevýhoda výše zmíněného přístupu může být případný vysoký počet zákazníků v segmentech s podobným chováním, které zapříčiněno nerovnoměrným rozložením jednotlivých hodnot metrik. Například pokud převažuje skupina klientů s jediným nákupem nebo pokud pouze malá část klientely je pro obchod zajímavá viz tzv. *paretův princip*.

Tento obchodní princip, který se také často označuje jako pravidlo „20-80“, říká, že 80 % všech prodejů obvykle plyne z 20 % nejlukrativnějších klientů.

Další možností je tedy využití expertního rozdělení pro určení kategorií ohodnocení RFM [7]. Například:

Recency

5 - poslední transakce méně než před 1 měsícem

4 - poslední transakce méně než před 3 měsíci

3 - poslední transakce méně než před 6 měsíci

2 - poslední transakce méně než před 12 měsíci

1 - poslední transakce déle než před 12 měsíci

Frequency

5 - 6 a více transakcí za posledních 12 měsíců

4 - 4-5 transakce za posledních 12 měsíců

3 - 2-3 transakce za posledních 12 měsíců

2 - 1 transakce za posledních 12 měsíců

1 - žádná transakce za posledních 12 měsíců

Monetary

5 - transakce za více než 50 tis. Kč za posledních 12 měsíců

4 - transakce za 30-50 tis. Kč za posledních 12 měsíců

3 - transakce za 10-30 tis. Kč za posledních 12 měsíců

2 - transakce za 1-10 tis. Kč za posledních 12 měsíců

1 - transakce za méně než 1 tisíc Kč za posledních 12 měsíců

Takovéto rozdělení je nutné stanovit pro každou studii zvlášť společně s odborníky na danou oblast trhu, aby dané ohodnocení dobře popisovalo chování zákazníků. Pozitivním přínosem pak ale může být jednoduchá interpretovatelnost modelu, pakliže zvolenými experty budou přímo jeho uživatelé. Nicméně by mělo rozdělení směřovat k dostatečně rovnoměrnému zastoupení klientů ve výsledných segmentech.

1.3 Vážení RFM pro určení významnosti klienta

Výsledky ohodnocení RFM analýzy nemusí sloužit pouze k rozdělení zákazníků do segmentů, ale může být také vstupem pro odhad obchodní významnosti klienta.

Analýzou určování dlouhodobé hodnoty klienta, tzv. CLV, se zabývá následující kapitola. Zde je nastíněno využití RFM pro výpočet *RFM value*, které může být jedním z prediktorů pro předpovídání CLV, případně může posloužit jako jednoduchá aproximace.

(*Miglautsch, 2000*)[7] nabízí následující způsob výpočtu *RFM value* na základě původního ohodnocení z RFM analýzy [7].

1. Předpokládejme, že všechny tři metriky mají obchodně stejný význam. *RFM value* je potom vypočtena prostým součtem:

$$V_{RFM} = R + F + M$$

2. V obecnějším případě zvolme váhy W_R, W_F a W_M, které budou odpovídat významnosti jednotlivých metrik. *RFM value* pak vypočteme:

$$V_{RFM} = W_R \cdot R + W_F \cdot F + W_M \cdot M$$

Určení vah významnosti metrik závisí případ od případu a je nutné ho odvodit expertně na základě dané obchodní oblasti. (*Liu, Shih, 2004*)[8] navrhují využít metodiku AHP (*Analytical Hierarchy Process*) pro korektní určení významnosti každé z metrik RFM. Viz také (*Saaty, 2008*)[9].

Zde popsán proces AHP pro určení vah v následujících krocích dle [9]:

1. Několika experty je provedeno párové porovnání všech tří metrik na základě stanovené škály (1 – *stejná důležitost* až po 9 – *nezpochybnitelně větší důležitost první nad druhou*).

2. Ověření konzistence provedeného porovnání. Tento krok je nutno provést pro posouzení zaujatosti jednotlivých expertů během jejich vlastního porovnávání a také rozdílů jejich porovnání mezi sebou. Pakliže je porovnání označeno jako nekonzistentní, je třeba opakovat proces od začátku [10].

3. Výpočet relativních vah na základě provedených párových porovnání.

4. Stanovení konečných vah pomocí sloučení výsledků od každého z expertů za použití geometrického průměru.

WRFM, tedy vážení RFM, není důležité pouze pro výpočet *RFM value*, ale přináší také lepší výsledky při využití RFM pro shlukování [8].

1.4 Využití RFM pro segmentaci

Jak již bylo uvedeno, RFM analýza je nejčastěji používaná jako metoda pro snadné rozdělení zákazníků do segmentů se stejným chováním ve smyslu provádění transakcí s obchodníkem. Toto rozdělení může být poté využito pro zvolení odlišného přístupu obchodování s danými segmenty, například různé strategie oslovování nebo výběr vhodných segmentů pro direct mailing.

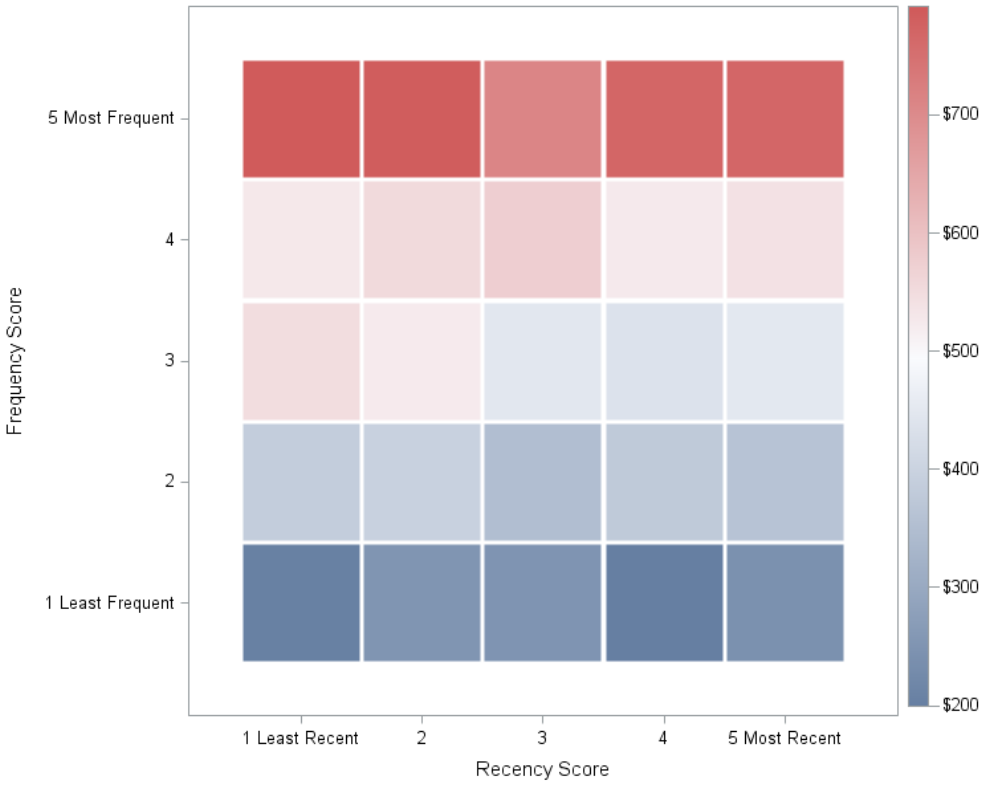
Dalším zajímavým využitím může být rychlá identifikace důležitých segmentů nebo sledování četnosti, případně změn četností v těchto segmentech, v rámci reportingu a vizualizace. Pro ilustraci viz obrázek 1.1.

Takovéto reporty mohou být díky své jednoduchosti a vysoké informační hodně velice vhodným nástrojem jako součást dashboardu reportovacího systému pro management obchodníka.

1.5 Využití RFM pro klasifikaci a regresi

Další možností je využití výstupu RFM analýzy v kombinaci s dalšími daminingovými metodami při klasifikaci zákazníků do předem daných skupin nebo predikci dalšího chování, jako je například pravděpodobnost respondibility na reklamní kampaň nebo doporučení produktů.

Rozdíl oproti samotné segmentaci je v tom, že cílem je na základě vstupních proměnných, v tomto případě ohodnocení z RFM analýzy, vypočítat nebo odhadnou hodnotu dané cílové proměnné. Dle typu výsledné proměnné se v diskrétním případě jedná o problém klasifikace, v případě spojitém o regresi. Z pohledu teorie strojového učení mluvíme o tzv. *učení s učitelem*.



Obrázek 1.1: Mapa segmentů RFM – barevnou škálou je vyznačena suma transakcí napříč segmenty podle hodnot *recency* (osa X) a *frequency* (osa Y).

Příkladem úspěšného využití pro predikci je studie (*Kim, Lee, 2012*)[11], kdy byly využity metriky RFM pro predikci vývoje technologií na základě dat z patentového úřadu. V prvním kroku byla provedena RFM analýza upravená pro transakce ve formě zveřejňování citací. Dále byly výstupy z této analýzy použity pro regresní model typu *CART – classification and regression tree*.

1.6 Kombinace RFM a metod shlukování

1.6.1 Definice pojmu shlukování

Obecně je úloha shlukování definována jako způsob rozdělení záznamů do skupin tak, že záznamy ve stejné skupině mají větší míru podobnosti než se záznamy v ostatních skupinách. Cílem je tedy najít přirozené nebo smysluplné skupiny – shluky – exitující v datech. Mírou podobnosti se myslí zvolená metrika vzdáleností záznamů v datovém prostoru. Cílem shlukování je minimalizovat vzdálenosti záznamů uvnitř společného shluku a zároveň maximalizovat

vzdálenosti mezi rozdílnými shluky (*Ricci, 2011*)[12].

Jedná se tak svým způsobem o klasifikaci do předem neznámých tříd a proto je shlukování řazeno mezi tzv. metody *učení bez učitele*.

1.6.2 Algoritmus *K*-means

Algoritmus K-means patří mezi základní metody shlukování a existuje mnoho jeho variant. Zde je popsána společná základní myšlenka algoritmu vycházející z (*Lloyd, 1982*)[13]:

1. V datovém prostoru je zvoleno a rozmíněno K centroidů.

2. Pro každý bod, který odpovídá danému záznamu, je spočtena vzdálenost ke každému z K centroidů.

3. Každý bod odpovídající danému záznamu je přiřazen k jednomu z K centroidů. Takto jsou záznamy rozděleny do K shluků.

4. Namísto původních je zvoleno K nových centroidů, které jsou zvoleny jako těžiště původních shluků.

5. Opakují se kroky 2 až 4 dokud již nedochází změně zařazení v kroku 3 nebo dokud není překročen stanovený počet iterací.

6. Záznamy jsou rozděleny do K shluků, které jsou reprezenotvány svými centroidy.

Počáteční rozmístění v kroku 1 lze provádět náhodně nebo zvolit z některých heuristických či aproximačních metod, viz (*Ostrovky et al., 2006*)[14].

Vzdáleností se myslí zvolená metrika vzdálenosti v datovém prostoru. Lze užít například prostou euklidovskou vzdálenost mezi body.

Míra kvality provedeného shlukování lze stanovit pomocí různých metrik, které berou v potaz vzdálenosti bodů od příslušných centroidů či ostatních bodů napříč shluky. Jedna z nečastěji používaných je metrik tzv. *TWSS* – (*total within sum of squares*), která je definována jako:

$$TWSS = \sum_{c=1}^{k} \sum_{i \in c} (x_i - \overline{c})^2$$

kde x_i je záznam náležící do jednoho z k shluků c a $\overline{c}$ je centroid příslučného shluku c.

1.6.3 Využití shlukování a RFM

Stejně jako v případě klasifikace a regrese, mohou být výstupy z RFM analýzy použity jako vstup pro další metody shlukování.

Práce (*Khajvand, 2010*)[15] popisuje dva způsoby využití RFM analýzy pro shlukování zákazníků jako vstup pro metodu K-means:

1. Pro shlukování jsou použity pouze metriky R, F, M. Pomocí tzv. *DUNN indexu* je odhadnuto optimální K - tedy množství výsledných shluků. Souřadnice centroidů těchto shluků jsou použity pro interpretaci chování zákazníků v rámci každého shluku. Je tedy takto nahrazeno rozdělení zákazníků do skupin, které vznikají při tradičním použití RFM analýzy. Výhodou je možnost zvolit menší počet výsledných shluků a snáze potom identifikovat důležité skupiny pro obchodní záměr.

   (*Birant, 2011*)[16] například navrhuje rozdělení do 8 shluků z toho důvodu, že existuje právě 8 kombinací ohodnocení menší/větší než průměr, které je posléze použito pro popis chování zákazníků v rámci každého shluku.

2. Pro shlukování jsou použity metriky R, F, M společně s dalšími nezávislými proměnnými, které popisují chování zákazníků. Tímto způsobem je možné odhalit složitější vztahy v datech a získat kvalitnější shluky.

(*Blattberg, 2008*)[1] popisuje 7 základních skupin proměnných vhodných pro účely segmentace:

Skupina proměnných	**Popis**
Preference benefitů	*Jaké benefity zákazník oceňuje (služby, citlivost na slevu)*
Psychografické údaje	P*opisy osobnosti, odhady na základě chování (fóra, sociální sítě atp.)*
Demografické údaje	*Věk, velikost příjmu, zaměstnání, rodinný status*
Geo-demografické údaje	*Regionální specifika dle místa bydliště*
Behaviorální údaje	*RFM metriky, zaujetí pro distribuční a komunikační kanály, zaujetí pro dané druhy zboží atp.*
Údaje o konkurenci	*Údaje popisující vliv konkurence na zákazníka, tzv. Wallet-share, preference konkurenčních služeb*
Hodnota zákazníka	*CLV, loajalita, dopady odchodu zákazníka, pravděpodobnost retence atp.*

Nevýhodou složitějšího modelu pro shlukování je však zpravidla obtížná zpětná interpretovatelnost zařazení zákazníka do daného shluku a popis společných rysů chování zákazníků v rámci výsledných shluků.

1.7 Kombinace RFM a asociačních pravidel

Asociační pravidla v nákupních koších odhalují vztahy mezi produkty. Jinými slovy lze takto analyzovat, které produkty jsou nejčastěji nakupovány pospolu, případně v čase ale stejnými klienty.

Využití asociačních pravidel bylo poprvé popsáno v (*Agrawal et al., 1993*)[17], o rok později byl publikován algoritmus *Apriori*, který slouží pro rychlé vyhledání asociačních pravidel ve velkých datasetech [18].

Kvalita zjištěných pravidel je posuzována na základě ukazatelů, mezi které se nejčastěji řadí hodnoty tzv. *suport* (kolik procent košů obsahuje asociované produkty zakoupené pospolu), *confidence* (jistotu, že obsahuje-li koš asociovaný produkt A, obsahuje také asoc. produkt B) a *lift* (míru nezávislosti asoc. produktů) (*Hahsler et al., 2005*)[19] .

(*Liu, Shih, 2004*)[8] ve své studii navrhují využít výstup z RFM analýzy pro nalezení silnějších asociačních pravidel v rámci modelu pro doporučování produktů. Byla porovnána úspěšnost doporučení vytvořených běžným způsobem z celé databáze transakcí s doporučeními odvozenými pouze v rámci jednotlivých shluků, které vznikly pomocí shlukovací metody *K-means* na základě výstupu předcházející vážené RFM analýzy.

U některých shluků, které reprezentovaly segmenty loajálních zákazníků, byla tímto způsobem dosažena lepší úspěšnost při doporučování než u metod pracujících s celou databází transakcí.

Kapitola 2

Odhad dlouhodobé hodnoty zákazníků – CLV

2.1 Definice a využití

S obchodního pohledu esenciální je také analýza zákaznické báze z hlediska odhadu očekávané profitability a dlouhodobé hodnoty zákazníků. Metrika, která se za tímto účelem využívá, je nejčastěji označována jako CLV (*Customer Lifetime Value*). Ekvivalentně se také v některé literatuře označuje jako *LTV*. Zde je uvedena definice CLV dle (*Blattberg, 2008*)[1]:

> *„Současná hodnota příjmů za celý životní cyklus určitého klienta poté, co byl získán, a po odečtení nákladů spojených s marketingem, prodejem, výrobou a obsluhou tohoto klienta.“*

Součástí procesu pro odhad hodnoty CLV jsou následující kroky [1]:

1. Predikce budoucích nákupů klienta
2. Výpočet nákladů související s klientem
3. Stanovení vhodné diskontní sazby pro výpočet současné hodnoty

Správně stanovená hodnota CLV klientů je důležitá nejen pro volbu vhodné obchodní nebo prodejní strategie, ale je také podstatná pro finanční plánování a výpočet tržní hodnoty společnosti.

Vzhledem k charakteru této práce se kapitola bude dále zabývat metodami predikování budoucích nákupů. Ostatní dva kroky jsou z obchodního hlediska také zcela zásadní a při jejich provádění je nutné provést pečlivou obchodní a marketingovou analýzu prodejního procesu.

Stanovení diskontní sazby potom určuje výnosovou míru, kterou je přepočítána budoucí peněžní hodnota na současnou (*Radová et al., 2013*)[20]. Pro detailnější výklad na toto téma viz [1],[20].

τ	r^τ	Očekávaný profit	Diskontní koef.	Diskontovaný profit
1	1	1000	1	1000
2	0,8	800	0,91	727
3	0,64	640	0,83	531
4	0,51	510	0,75	383
5	0,41	410	0,68	279
Sumární očekávaný profit za 5 období:				2920
Očekávaná hodnota CLV dle vzorce:				**3667**

Tabulka 2.1: Příklad užití retenčního modelu.

2.2 Retenční model

Základním a nejrozšířenějším zůsobem pro odhadování CLV je využití funkce přežití pro modelování retence zákazníků [1]. Tato funkce určuje, zdali zákazník „přežil“ až do nějakého časového bodu t.

Nechť T je náhodná veličina představující čas, kdy zákazník „umírá“ (přestává býti zákazníkem obchodu) s pravděpodobností danou funkcí hustoty pravděpodobnosti $f(t)$. Pravděpodobnost „úmrtí“ je $P(\tilde{T} < t) = F(t)$, kde F je distribuční funkce: $F(t) = \int_0^t f(x)dx$. Pravděpodobnost, že zákazník přežije uplynuvší čas t, je tedy

$$S_t = P(T \geq t) = 1 - F(t) = \int_t^\infty f(x)dx$$

(*Blattberg, 2008*)[1] navrhuje jednoduchý model v diskrétním případě s geometrickým rozdělením s parametrem h. V každé po sobě jdoucích obdobích (diskrétních krocích) je vždy konstantní hodnota pravděpodobnosti přežití $r = 1 - h$. Funkce přežití pro τ období od počátečního je r^τ.

V případě konstantní hodnoty pravděpodobnosti přežití r a konstantní hodnoty profitu zákazníka G (příjmy – náklady) v každé periodě je možné očekávanou hodnotu CLV vypočítat jako:

$$CLV = G \cdot \frac{1+d}{1+d-r}$$

kde d je zvolená diskontní sazba [1].

Pro ilustraci je uveden příklad výpočtu CLV v tabulce 2.1. Pravděpodobnost přežití je stanovena $r = 0,8$; zákazník v každé periodě generuje profit 1000 Kč. Hodnota diskontní sazby je zvolena $d = 0,1$.

Jednou z možností, jak zjistit příslušný parametr r, tedy pravděpodobnost přežití zákazníka v každém období, je provést analýzu historických dat přímým pozorováním. Pravděpodobnost může být vypočítána jako průměrná hodnota poměrů zákazníků, kteří byli aktivní v určitých dvou po sobě jdou-

cích obdobích vůči těm, kteří byli aktivní pouze v první z nich. Další možností může být stanovení budoucího r například pomocí regresních metod [1].

2.3 Migrační model

Dalším nejčastěji využívaným způsobem pro odhadování CLV je použití tzv. *migračního modelu*, který modeluje odhad CLV jako markovský řetězec. Označení „migrační“ je zvoleno proto, že model umožňuje přesouvání zákazníka mezi několika stavy. Nejobvyklejším způsobem je definování pomocí toho, kdy naposledy zákazník provedl nákup.

Na rozdíl od předchozího retenčního modelu tento způsob připouští situaci, kdy zákazník v určitém období není aktivní (například neprovede nákup), ale v některém následujícím období se znovu aktivuje. Z tohoto důvodu model lépe odpovídá situaci byznysu typu *B2C* („business to customer“), jako jsou katalogový prodejci nebo internetové obchody.

Hlavní idea migračního modelu spočívá v definici stavů životního cyklu, ve kterých se může daný zákazník nacházet. Nejdříve vždy dochází k akvizici zákazníka – zákazník provádí nákup a nachází se ve *stavu 1*. Pakliže v dalším období znovu provede nákup, zůstává ve *stavu 1*. V opačném případě putuje do *stavu 2*. Tento vzorec se dále opakuje. Pokud zákazník provede v dalším období nákup, vrací se do *stavu 1*, jinak se přesouvá do *stavu 3* atd.

Pakliže je možné odhadnout pravděpodobnost všech přechodů mezi definovanými stavy, lze sestavit markovský řetězec, který bude sloužit jako model pro výpočet odhadu CLV zákazníka nacházejícího se v každém ze stavů. [21]

(*Pfeiffer a Carraway, 2000*)[21] přišli s obecným postupem, jak ze sestaveného markovského řetězce odvodit CLV pro každý ze stavů pomocí následující formule:

$$\mathbf{V}^{\infty} = \lim_{T \to \infty} \mathbf{V}^{T} = [\mathbf{I} - (1+d)^{-1}\mathbf{P}]^{-1}\mathbf{G}$$

kde $\mathbf{P}$ je matice přechodů mezi stavy markovského řetězce, $\mathbf{I}$ je jednotková matice, d je zvolená diskontní sazba a $\mathbf{G}$ je vektor, jehož složky odpovídají hodnotám profitu zákazníka v odpovídajících stavech (příjmy – náklady). Složky výsledného vektoru $\mathbf{V}^{\infty}$ potom odpovídají hodnotám CLV, pokud se zákazník nachází na začátku v odpovídajícím stavu. Hodnoty profitu v případě nákupu je možné odvodit z průměrné hodnoty nákupu.

Pro ilustraci je uveden příklad migračního modelu se třemi stavy. Matice přechodů $\mathbf{P}$ má v tomto případě následující podobu:

$$\mathbf{P} = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{pmatrix} p_1 & 1-p_1 & 0 \\ p_2 & 0 & 1-p_2 \\ p_3 & 0 & 1-p_3 \end{pmatrix} \end{array}$$

Složky každého řádku matice **P** představují pravděpodobnosti, že dojde k migraci zákazníka z příslušného stavu do stavu prvního (představuje nákup), nebo do stavu, který odpovídá neaktivitě a posunu do dalšího období.

Například p_2 je pravděpodobnost, že zákazník provede nákup, pakliže v předcházejícím období nenakoupil (*stav 2*). Hodnota $1 - p_2$ naopak značí míru pravděpodobnosti, že klient nenakoupil ve dvou po sobě jdoucích obdobích a posune se do následujícího *stavu 3*.

V případě posledního stavu je třeba rozhodnout, zdali tento stav reprezentuje definitivní odchod zákazníka – potom je hodnota $p_3 = 0$. Ve smyslu teorie markovských řetězců se jedná o tzv. absorční stav. V opačném případě je zachována tranzitivita a stav v tomto konkrétním případě představuje neaktivitu po dvě a více předcházejících období.

Složky vektoru G odpovídají hodnotám:

$$\mathbf{G} = \begin{matrix} S_1 \\ S_2 \\ S_3 \end{matrix} \begin{pmatrix} P - N_1 \\ N_2 \\ N_3 \end{pmatrix}$$

Kde P je příjem z uskutečněného prodeje a N_i je součet nákladů na prodej, marketing, výrobu a obsluhu klienta v daném období i.

Kapitola 3

Analytický framework pro reportování

3.1 Popis a cíle při navrhování frameworku

Tato kapitola obsahuje návrh analytického frameworku pro účely reportování majitelům transakčních a prodejních dat e-shopů. Navrhovaný framework využívá obecné principy a postupy popsané v předcházejících kapitolách 1 a 2, s ohledem na specifické rysy e-shopů.

Cílem je vytvořit seznam a popis jednoznačných, po sobě jdoucích a částečně na sobě závislých kroků, které vedou k vytvoření přehledných a popisných reportů s ohledem na možnosti další interpretace z pohledu byznysových analytiků a nezávisle na konkrétní implementaci.

Důležitou součástí je výběr vhodných typů analýz, metod a způsobů reportování, s ohledem na jednoduchost ve smyslu interpretovatelnosti a využití majiteli dat a bez nutnosti komplexnějších znalostí v oblasti matematiky, statistiky nebo teorie strojového učení.

Navrhovaný framework definuje minimální požadavky na podobu a rozsah vstupních dat za účelem provedení základních analýz a vytvoření základních reportů a popisuje další možné analýzy a reportování včetně nutných předpokladů pro jejich provedení.

3.2 Specifické rysy e-shopů a popis minimální formy vstupních dat

Stejně jako mezi klasickými obchody je v oblasti těch internetových velká rozmanitost z pohledu velikosti, zaměření, typu prodávaného zboží apod. Nicméně existují jistá specifika e-shopů z důvodů plynoucích právě ze způsobů nabídky zboží a komunikace se zákazníky internetovými kanály, ať už se jedná o pozitiva, jako jsou minimální náklady na obsluhu zákazníků a pronájem prostor,

nebo negativa, jako zvýšená konkurence z pohledu ceny zboží a rizika spojená s expediční logistikou.

Tato práce se nebude zaobírat všemi obchodními a prodejními hledisky, nicméně popisuje několik spicifik internetových obchodů spojených s využitím dat o prodejích a zákaznické bázi, které je třeba brát v potaz jak během provádění samotných analýz a reportování výsledků, tak při jejich interpretaci.

Internetové obchody zpravidla:

1. mají centralizovaná data o zákaznících a prodejích uložená v transakčních databázích, které slouží pro realizování prodeje, případně v datových skladech
2. jsou svými zákazníky navštěvovány sporadicky a zákazníci provádějí nákupy v nepravidelných intervalech.
3. nejsou vždy schopny rozpoznat nákup od stejného zákazníka.
4. mají velký podíl zákazníků, kteří nakoupili pouze jedinkrát.
5. jsou ovlivněny sezónností prodejů (např. Vánoce, letní měsíce apod.).
6. mají velmi malé náklady spojené s obsluhou zákazníků, které je možné do jisté míry zanedbat.

Především výše zmíněné body 1 až 4 přímo souvisí s minimálními předpoklady pro nasazení analytického frameworku nad interakčními (prodejními) daty. Mimo samotnou existenci transakční databáze, což je přirozeně základním předpokladem, je nutné brát v potaz velikost zákaznické báze a frekvenci a objem prodejů každého jednotlivého obchodu, na který má být framework aplikován.

Navrhovaný analytický framework popisuje ve své minimální podobě prostředky pro analýzu zákaznických segmentů a reportování z transakcních dat takřka libovolného, tedy i menšího rozsahu, bez dalších znalostí o zákaznících nebo produktech. Tedy i případy, kdy databáze transakcí obsahuje relativně malé množství záznamů a nejsou k dispozici žádné další relevantní informace o charakteru jednotlivých zákazníků.

Zde jsou popsány požadavky na minimální formu vstupních dat pro účely nasazení frameworku v základní variantě:

1. Data o transakcích (prodejích) jsou uložena tak, že je možné je převést do vstupního datasetu ve formě databázové tabulky s následující strukturou:

Odvozený atribut	**Typ**	**Popis**
user_id	text	unikátní identifikátor zákazníka
recent_purchase	date	datum pořízení posledního nákupu
days_from_recent	numeric	počet dní uplynulých od posledního nákupu k datu provedení analýzy
num_of_purchases	numeric	počet provedených nákupů za pokryté období
sum_total	numeric	celková hodnota nákupů za pokryté období

2. Počet takto vzniklých záznamů v datasetu je dostatečně velký, aby měla data dostatečnou vypovídací hodnotu a bylo vyloučeno nahodilé chování z důvodu příliš malého vzorku.

3.3 Kroky základní analýzy a reportování

Základní podoba navrhovaného frameworku zahrnuje inspekci vstupních dat a vizualizaci vstupních dat v původní podobě a následně využívá aplikaci RFM analýzy pro vytvoření zákaznických segmentů, jejich vizualizace a interpretaci. Po stanovení vah významnosti složek *recency, frequency a monetary* je možné stanovit odhad významnosti klientů dle jednotlivých RFM segmentů a jejich četnosti v rámci segmentů využít jako indikátor úspěšnosti obchodu.

3.3.1 Inspekce dat a statistická analýza

V prvním kroku je provedena základní inspekce vstupních dat za účelem vytvoření přehledu o objemu transakcí a velikosti zákaznické báze za sledované období. Základní přehled obsahuje nápočty následujících položek:

- Vymezení sledovaného období
- Počet dnů ve sledovaném období
- Celkový počet transakcí
- Počet transakcí s nenulovou hodnotou nákupu
- Celkový počet nakupujících zákazníků
- Celkový počet nákupních košů*
- Celkový obrat všech transakcí

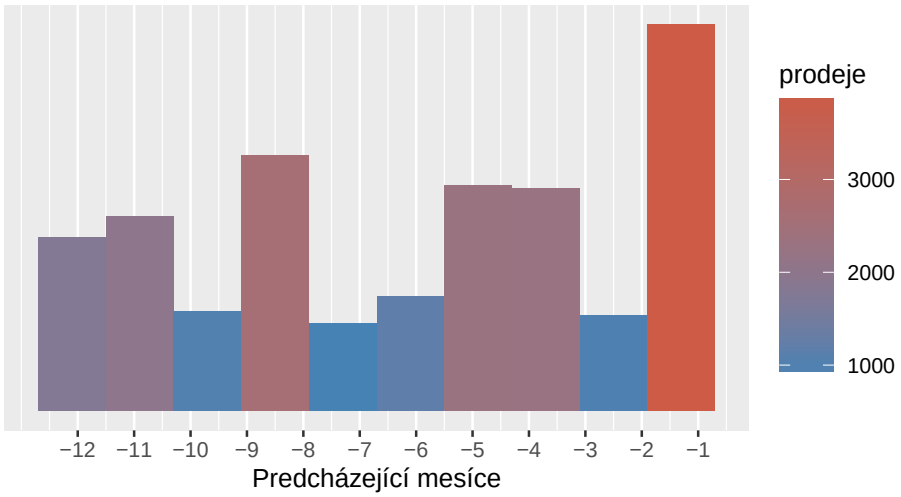
*Nákupní koš je definován jako množina produktů, kterou daný zakoupí najednou ve stejný čas v rámci jednoho nákupu.

Dále jsou vypočítány statistické míry - *aritmetický průměr*, *medián* a *směrodatná odchylka* pro vstupní data *days_from_recent*, *num_of_purchases* a *sum_total*.

Interpretace těchto měr je důležitá pro vytvoření představu o podobě běžného nákupu, pravidelnosti s jakou se zákazníci do obchodu vracejí apod.

Nástrojem pro odhalení sezónnosti je vizualizace ve formě histogramu nákupů. Jako vstup mohou být použita přímo původní transakční data nebo lze sezónnost odhadnout na základě *days_from_recent*. Výsledný histogram sestává z košů odpovídajících jednotlivým měsícům sledovaného období, seřazených od prvního měsíce sledovaného období až po současnost. Příklad takového histogramu je na obrázku 3.1.

Odhalení sezónnosti prodejů je důležité pro nadcházející RFM analýzu v souvislosti se segmentací dle *recency* s expertním rozdělením.



Obrázek 3.1: Histogram prodejů za předcházející měsíce.

3.3.2 RFM analýza

Základní podoba podoba RFM analýzy je při rozdělení do diskrétních hodnot 1-5 dle kvantilů po jednotlivých metrikách *recency, frequency a monetary* tak, jak je popsáno v kapitole 1.1.

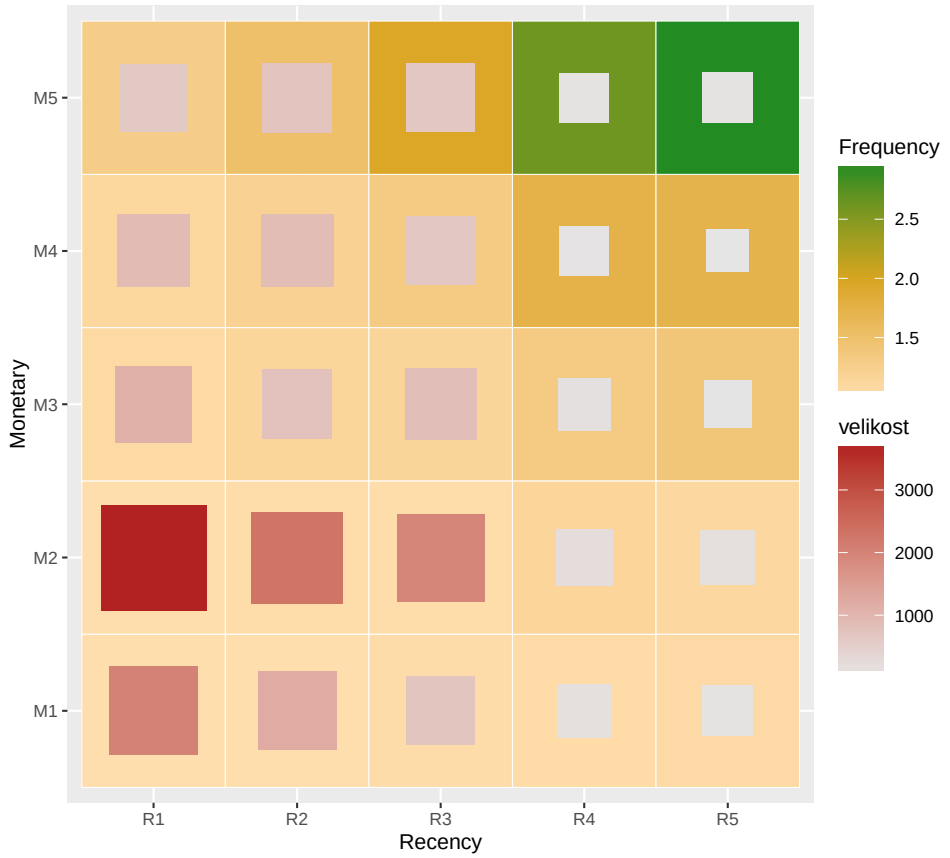
Pro tento účel jsou pro hodnoty atributů *days_from_recent*, *num_of_purchases* a *sum_safe* vypočítány hodnoty percentilů v rámci celé klientské báze a tyto percentily jsou při-řazeny k jednotlivým zákazníkům jako odvozené atributy *r_perc*, *f_perc* a *m_perc*.

U hodnot *recency* percentily v obráceném pořadí, tzn. hodnota 1 odpovídá zákazníkovi s nejnovějším nákupem.

Z těchto percentilů jsou poté odvozeny atributy *r_q*, *f_q* a *m_q*, které představují ohodnocení vzniklé na základě RFM analýzy:

Ohodnocení R (F, M resp.)	**Percentil**
1	$\leq$ 0,2
2	0,2–0,4
3	0,4–0,6
4	0,6–0,8
5	$\geq$ 0,8

Z hlediska interpretovatelnosti je vhodné použít variantu s expertním rozdělením dle kapitoly 1.2, pakliže je takové expertní rozdělení k dispozici. Jednotlivé hranice pro dělení do segmentů 1-5 dle *recency, frequency a monetary* mohou odpovídat obchodním záměrům. Například lze takto vytvořit segmentaci dle monetary na základě hranic stanovených podle cen různých produktových kategorií apod.



Obrázek 3.2: Možná vizualizace RFM segmentů. Barva pozadí a čtverce ukazuje průměrnou hodnotu *frequency*, resp. velikost segmentu.

Se způsoby interpretace je tedy možné počítat již při samotném definování expertního rozdělení a tím pádem je umožněn dobrý vhled do zákaznických segmentů, pakliže je toto rozdělení určeno vhodně.

Důležitým krokem je potom vhodná vizualizace výsledných RFM segmentů. Při provedení segmentace přes všechny tři dimenze RFM vždy vzniká 125 různých segmentů, což je pro prostou vizualizaci velmi vysoký počet. Pro snadnější orientaci ve výsledném prostorovém zobrazení je vhodné proto dimenzi s nejmenší vypovídací hodnotou vynechat. Pro výběr lze využít znalosti statistických měr směrodatné odchylky nebo byznysové požadavky. Pro zobrazení třetí dimenze a velikosti segmentů je vhodné využití barevné škály.

Obrázek 3.2 ukazuje možná způsob vizualizace segmentace ve dvou dimenzích dle *recency* a *monetary* společně s velikostmi jednotlivých segmentů a jejich průměrné hodnoty třetí dimenze *frequency*.

3.3.3 Odhad významnosti klientů pomocí WRFM

Rozdělení zákazníků nemusí sloužit pouze k segmentaci, ale může být využito také ke scoringu zákazníků v jednotlivých segmentech. Tak, jak je popsáno v kapitole 1.3, je možné přiřadit skóre každému z RFM segmentů v podobě hodnoty *RFM value*, která je dána součtem:

$$V_{RFM} = W_R \cdot R + W_F \cdot F + W_M \cdot M$$

kde W_R, W_F a W_M jsou váhy významnosti jednotlivých složek.

Navrhovaný framework pro stanovení vah využívá proces *AHP*, který je popsán v kapitole 1.3.

Výsledkem analýzy je nakonec odhad významnosti klientů zvlášť pro jednotlivé RFM segmenty z původní RFM analýzy.

Pro vizualizaci je vhodné použít dvoudimenzionální zobrazení RM segmentů společně s odpovídající průměrnou *RFM value* pro jednotlivé segmenty.

3.3.4 Segmentace pomocí metody shlukování

Provedená vizualizace RFM analýzy v dimenzích metrik RFM nabízí možný způsob segmentace zákazníků a při vhodné interpretaci naznačuje směry využitelné pro plánování a cílení obchodních kampaní.

Obecná nevýhoda při využití segmentů, které vznikli rozdělením na základě ohodnocení R, F a M je však velké množství výsledných segmentů, které však pravděpodobně obsahují zákazníky s podobným chováním.

Tento krok navrhovaného frameworku zabývá řešením problému snížení počtu zákaznických segmentů s využitím shlukování metodou K-means, která je popsána v kapitole 1.6.2. Cílem je využití výstupů s předešlé RFM analýzy a dosáhnoutí menšího počtu segmentů pro snadnější dělení zákaznické báze a jednodušší orientaci.

První variantou je využít přímo ohodnocení z výstupu RFM analýzy. Na vstupu pro algoritmus K-means jsou tedy tři dimenze – hodnoty R, F, M v rozsahu 1 až 5. Následující postup je proveden jak pro výsledky RFM analýzy kvantilového typu, tak pro RFM s expertním rozdělením.

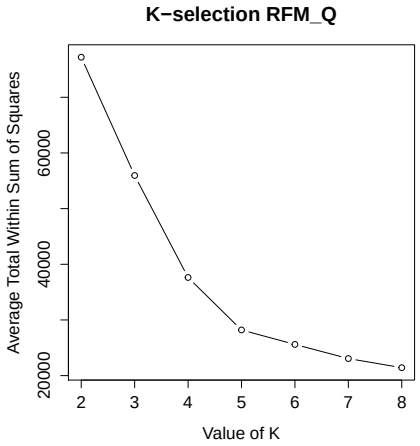
V prvním kroku je nejprve experimentálně zjištěna kvalita shlukování na základě sumární čtvercové vzdálenosti každého pozorování od příslušného centroidu určeného shluku – *TWSS* (*total within sum of squares, viz kapitola 1.6.2*). To je docíleno opakovaným měřením a průměrování pro zvolený rozsah k počtu shluků.

Vzhledem k požadované aplikaci shlukování, tedy významnému snížení počtu segmentů při zachování kvality segmentaci, bylo K zvoleno v rozsahu 2 až 8. Pruměrnou hodnotu *TWSS* je nutné pro každé volené k vypočítat z většího počtu nezávislých běhů algoritmu.

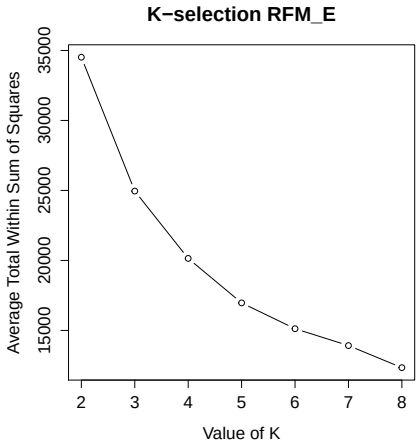
Grafy 3.3 a 3.4 ukazují příklad výsledků měření běhu algoritmu K-means na základě RFM analýzy kvantilového typu (RFM_Q) i s expertním rozdělením (RFM_E). Grafy ukazují závislost TWSS na hodnotě zvoleného k.

Při volbě vhodného k je třeba brát v potaz nejen absolutní hodnotu *TWSS*, která se stoupajícím k zpravidla klesá, ale především jakési body zlomu, kdy dochází k výtazným skokovým slepšením nebo se dále *TWSS* s přibývajícími středy již tolik významně nezlepšuje.

Na obrázcích 3.3 a 3.4 těmto bodům zlomu odpovídají hodnoty k 4 nebo 5. Z pohledu na oba grafy lze také vyčíst, že v varianta *RFM_E* v tomto případě dosahuje lepších hodnot *TWSS* a je tedy pro využití shlukování vhodnější.



Obrázek 3.3: Příklad - Graf závislosti TWSS na K



Obrázek 3.4: Příklad - Graf 2 závislosti TWSS na K

Druhá varianta shlukování nevyužívá vstup z předchozí RFM analýzy a na vstupu pro algoritmus K-means vstupují hodnoty *days_from_recent*, *num_of_purchases* a *sum_total*. V rámci této varianty jsou nabídnuta dvě řešení:

1. Jako vstup pro shlukování použít přímo původní nediskretizované atributy d*ays_from_recent*, *num_of_purchases* a *sum_total*.

2. Využít percentily *r_perc*, *f_perc* a *m_perc*, odvozené z původních hodnot. Zde se vlastně jedná o využití velice jemné diskretizace na výsledném prostoru 100 x 100 x 100.

Postup je shlukování je stejný jako v předchozím případě s tím, že aby bylo možné poměřovat kvalitu shlukování, musí být použit stejný rozměr jako v předchozím případě. Hodnoty dimenzí je tedy nutno v obou případech normalizovat do intervalu 1 až 5.

V případě využití percentilů stačí původní rozsahy s_{0-1} 0 až 1 upravit dle vzorce:

$$s_{1-5} = s_{0-1} \cdot 4 + 1$$

V případě využití původních hodnot je nutné také myslet na velký vliv outliers a hodnoty nejprve standardizovat ve statistickém smyslu s využitím metody *z-score* a potom aplikovat převod na rozsah hodnot 0 až 1:

$$s_{0-1} = s_z - \frac{min(s_z)}{max(s_z) - min(s_z)}$$

$$s_z = \frac{(x - u)}{o}$$

Interpretace shluků, vzniklých jakoukoliv z popsaných metod, je nejdůležitějším bodem této části analýzy. Pro indikaci toho, jaký typ klientů segmenty obsahují lze využít:

1. Velikost výsledného segmentů.

2. Pozice centroidu v dimenzích RFM (na škále 1-5).

3. Znalosti vyplívající z expertního rozdělení.

K dalšímu zpřesnění interpretace shluků je možno využít dodatečnou *analýzu nákupních košů* z následujícího oddílu.

3.4 Dodatečné analýzy a reportování

Tato část frameworku popisuje dvě dodatečné analýzy, které jsou vhodné k nasazení pro e-shopy s větším objemem prodejů a větší zákaznickou bází. Nutnou podmínkou pro úspěšné provedení je také významnější podíl klientů, kteří daný e-shop navštěvují opakovaně.

3.4.1 Interpretace segmentů pomocí analýzy nákupních košů

Další možností, jak obohatit interpretaci provedené segmentace pomocí základních analýz, je segmenty vyšetřit také z hlediska produktového. Pro tento účel je využívána tzv. *analýza nákupních košů*, která zkoumá zastoupení jednotlivých typů produktů v nákupech a také vztahy a spojitosti jednotlivých produktů během nákupů.

Předpokladem pro provedení analýzy nákupních košů je existence dat o produktech, která jsou uložena takovým způsobem, že je lze převést do vstupního datasetu ve formě databázové tabulky s následující strukturou:

Odvozený atribut	**Typ**	**Popis**
item_id	text	unikátní identifikátor produktu
name	text	název produktu
category	enum	produktová skupina

Z důvodu velkého množství různých produktů, které ale mohou být z obchodního hlediska podobné, je vhodné na místo s jednotlivými produkty raději pracovat s menším množstvím produktových skupin - kategorií. Ty mohou být odvozeny z katalogizace, kterou internetové obchody obvykle používají pro usnadnění orientace při nakupování svým zákazníkům, nebo stanoveny expertně.

Navrhovaný framework popisuje dva kroky analýzy nákupních košů:

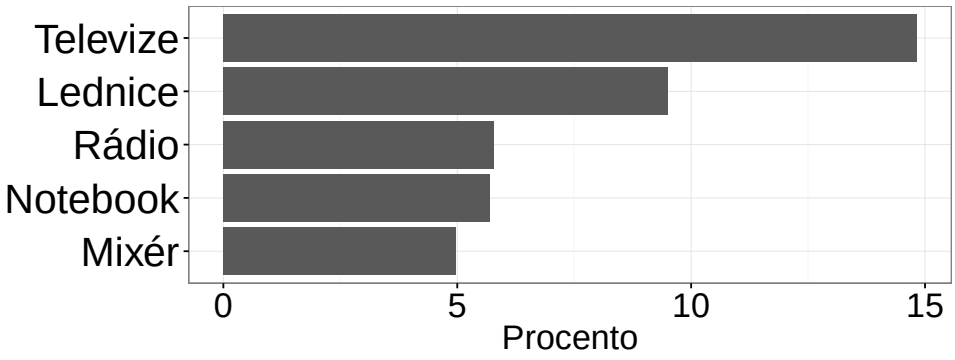
1. Určení nejprodávanějších produktů celého obchodu a po jednotlivých segmentech.

2. Analýza asociačních pravidel mezi produkty v nákupních koších celého obchodu a dle jednotlivých segmentů pomocí algoritmu *apriori* (viz kapitola 1.7).

V rámci prvního kroku jsou vypočítány četnosti prodejů všech produktů jak napříč celou zákaznickou bází, tak zvlášť pro klienty spadající do jednotlivých segmentů (či shluků) z předchozí segmentace.

Výsledky je vhodné vizualizovat jako seznam top produktů procentuálně zastoupených v nákupních koších. Pro příklad viz obrázek 3.5.

Interpretace výsledků je přímočará a vede k přiřazení typických produktů pro dané segmenty a ověření, jakým způsobem se od sebe zjištěné segmenty



Obrázek 3.5: Top 5 nejprodávanějších produktů procentuálně v koších.

liší. Lze takto také potvrdit nebo vyvrátit některé indikace na základě znalosti expertního rozdělení.

Ještě hlubší vhled do zákaznického chování v rámci jednotlivých segmentů může přinést analýza asociačních pravidel mezi nákupy různých produktů. Při extrahování pravidel je však důležité brát v potaz především hodnotu *support* (viz kapitola 1.7) každého jednotlivého pravidla a využít pouze pravidla s dostatečnou statistickou významností. Na základě velikosti datasetu je nutné zvolit vhodnou hranici ve formě minmální přípustné hodnoty *support*.

Pro vizualizaci nalezených asoc. pravidel lze využít mnoho způsobů, nicméně při větším množství pravidel se jedná o poměrně netriviální problém. Tento framework navrhuje využít postupy popsané v (*Hahsler a Chelluboina, 2010*)[22]. Jeden z příkladů možné vizualizace nalezených pravidel je na obrázku 3.6.

3.4.2 Odhad očekávané hodnoty klientů - CLV

Poslední část se zaměřuje na analýzu zákaznické báze z hlediska odhadu očekávané profitability a hodnoty zákazníků - CLV. Navrhovaný framework využívá migrační model (viz kapitola 2.3), který je vhodný pro odhadování CLV pro obchody typu B2C (*business to customer*), tedy i pro oblast e-shopů. Migrační model je použit v kombinaci výsledné segmentace z předchozí RFM analýzy.

Sestavení modelu spočívá v definici stavů, ve kterých se zákazník může nacházet, v odhadu pravděpodobností přechodů mezi těmito stavy a v určení hodnot profitu zákazníka v odpovídajících stavech (příjmy – náklady).

Pakliže je možné odhadnout pravděpodobnost všech přechodů mezi definovanými stavy, lze sestavit markovský řetězec, který bude sloužit jako model pro výpočet odhadu CLV zákazníka nacházejícího se v každém ze stavů. Pro definici stavů navrhovaný framework využívá segmentaci zákaznické báze dle recency z RFM analýzy.

Odhady pravděpodobností přechodů mezi segmenty mohou být stanoveny průměrným poměrem mezi klienty, kteří provedli opakovaný nákup v odpoví-

dajícím období, a klienty, kteří nenakoupili a posunuli v následujícím období do segmentu s nižší *recency*.

Hodnoty profitu v případě nákupu je možné odvodit z průměrné hodnoty nákupu. Navrhovaný framework navíc používá rozdílné průměrné hodnoty nákupu po jednotlivých segmentech dle *monetary*.

Ke stanovení odhadu CLV pro jednotlivé segmenty je použit postup a formule z kapitoly 2.3.

Na vstupu pro výpočet pomocí uvedené formule je $\mathbf{P}$ matice odhadnutých přechodů mezi stavy markovského řetězce, d zvolená diskontní sazba a $\mathbf{G}_{1-5}$ vektory, jejichž složky odpovídají hodnotám profitu zákazníka v odpovídajících stavech (příjmy – náklady). Složky výsledného vektoru potom odpovídají CLV, pokud se zákazník nachází na začátku v odpovídajícím stavu.

Uvedený postup je postupně opakován pro případy $\mathbf{G}_i, i = 1 \ldots 5$, ve kterých se hodnota příjmu z nákupu odvíjí od rozdílných průměrných hodnot nákupu po jednotlivých segmentech dle *monetary*. Výsledkem analýzy je tedy nakonec odhad průměrné hodnoty CLV zvlášť pro jednotlivé RM segmenty z původní RFM analýzy.

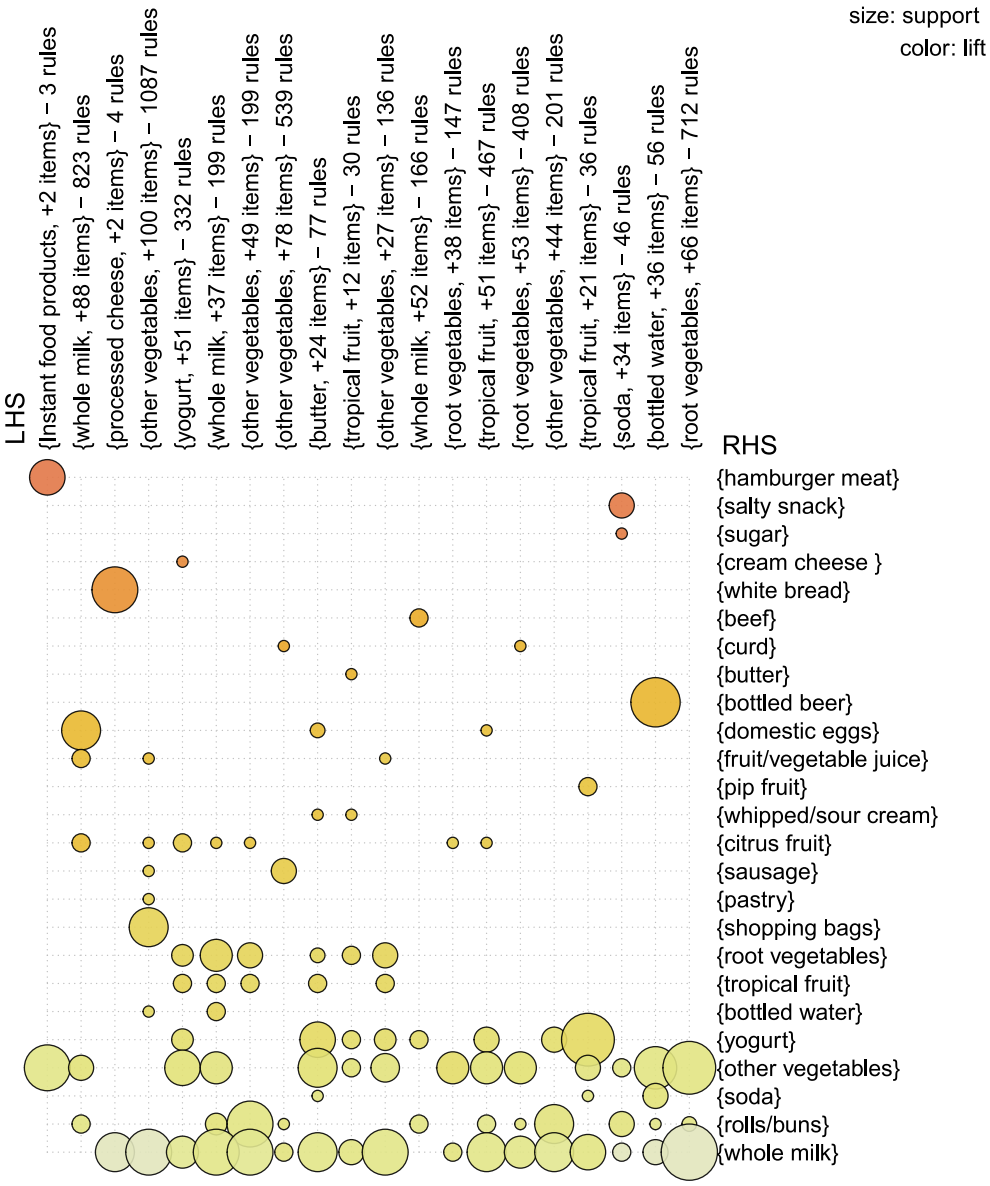
Pro vizualizaci je vhodné použít dvoudimenzionální zobrazení RM segmentů společně s odpovídajícím odhadem CLV pro jednotlivé segmenty. Viz příklad na obrázku 3.7.

3.5 Indikátory úspěšnosti e-shopů

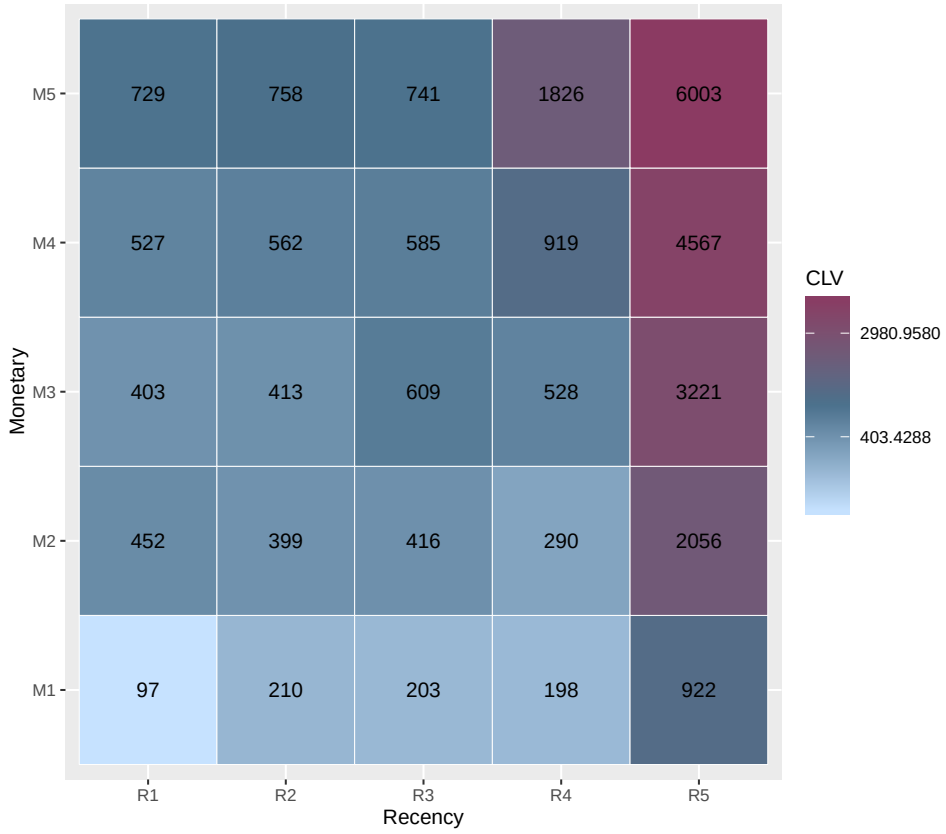
Navrhovaný framework stanovuje následující indikátory úspěšnosti daného e-shopu na základě výsledků výše popsaných analýz:

1. Velikost zákaznické báze a celkový obrat transakcí k datu provádění studie.
2. Historický vývoj počtu transakcí s nenulovou hodnotou v čase po měsících a vliv sezónnosti.
3. Podíl prvonákupců vůči celé zákaznické bázi.
4. Podoba rozvržení zákaznické báze do RFM segmentů.
5. Odhad významnosti zákazníků dle *RFM value* po jednotlivých RFM segmentech.
6. Odhad CLV zákazníků po jednotlivých RFM segmentech.

Tyto indikátory v kombinaci s vhodnou interpretací slouží jako zdroj pro doporučení zavedení obchodních strategií pro vytěžení stávající zákaznické báze.



Obrázek 3.6: Příklad vizualizace. Graf vztahů mezi asociovanými produkty. Na ose X jsou vyneseny produkty na pravé a na ose Y produkty na levé straně asoc. pravidel. Velikost bubliny odpovídá hodnotě *support* nalezeného pravidla a barevná škála (bílá-červená) hodnotě *lift*.



Obrázek 3.7: Možná vizualizace CLV pro RM segmenty. Barevná škála pozadí zobrazuje odhad CLV pro jednotlivé segmenty.

Kapitola 4

Případové studie

4.1 Případová studie 1 – Kursport.cz

4.1.1 Popis a charakteristika

KURsport s.r.o. je obchod se sportovním zbožím, který se zaměřuje především na prodej cyklistických kol a příslušenství. Mimo několika kamenných prodejen provozuje také internetový prodej na adrese *www.kursport.cz*. Data zákazníků internetového obchodu (dále jen obchod) jsou předmětem této studie. Transakční databáze obchodu je spravována v rámci systému společnosti *CloudSailor*. Data pro analýzu byla dále upravena, anonymizována a poskytnuta pro potřeby analytického systému společnosti *Recombee*, která pak data ve specifikované podobě poskytla pro tuto studii.

4.1.2 Cíle prováděné studie

Cílem studie je vyšetřit transakční (prodejní) data obchodu pro možnosti analýzy zákaznických segmentů a využít postupy navrhovaného frameworku (viz kapitola 3) pro reportování užitečných informací majitelům dat. Dále si potom studie klade za cíl využít zjištěné znalosti a navrhnout doporučení pro zavedení obchodních strategií vhodných pro vytěžení stávající klientské báze.

Z důvodu menší velikosti obchodu je zvolena základní podoba frameworku dle kapitoly 3.3.

4.1.3 Použité nástroje

Pro extrakci dat byly využity nástroje pro zálohování a obnovu databázového stroje *PostgreSQL 9.5* a tento tento databázový systém byl dále využit pro datové transformace a nápočty v sekcích 4.1.4 a 4.1.5 pomocí dotazovacího nástroje *SELECT*.

Atribut	Typ	Popis	Poznámka
userid	text	identifikátor zákazníka	prim. klíč
itemid	text	identifikátor zakoupeného produktu	prim. klíč
timestamp	timestamp	časová značka	prim. klíč
income	numeric	hodnota nákupu v Kč	

Tabulka 4.1: Struktura tabulky *purchases*

RFM analýza v sekci 4.1.6 a následné shlukování 4.1.7 byla provedena pomocí statistického softwaru *R (verze 3.2.3)* s použitím vývojového prostředí *RStudio* a s podporou tabulkového procesoru *MS Excel 2007*.

Všechny vizualizace byly rovněž vygenerovány v softwaru *R*, konkrétně pomocí balíčků *ggplot2* a *arulesViz*.

4.1.4 Extrakce dat

Extrakce dat byla provedena pomocí zálohovacích nástrojů databázového systému *Postgres* a importována na databázový stroj *PostgreSQL 9.2*. Přejaté datové schéma obsahuje mimo jiné následující databázové tabulky, které odpovídají definovaným analytickým potřebám pro tuto studii:

items - seznam nabízených produktů

users - seznam uživatelů (zákazníků) obchodu

purchases - seznam uskutečněných prodejů

Jako zdroj dat o provedených prodejích, který slouží jako vstup pro analýzu interakcí zákazníků, v tomto případě poslouží tabulka *purchases*. Obsahuje záznamy o jednotlivých prodejích, které jsou definovány čtveřicí (zákazník, zakoupený produkt, hodnota nákupu, čas uskutečnění nákupu).

Kompletní struktura databázové tabulky *purchases* je uvedena v tabulce 4.1

4.1.5 Inspekce a předzpracování dat pro RFM analýzu

Navržený analytický framework využívá RFM analýzu jako základní nástroj pro analýzu zákaznických segmentů a reportování z transakčních dat libovolného, tedy i menšího rozsahu bez dalších znalostí o zákaznících nebo produktech. To je případ této studie, kdy databáze transakcí obsahuje relativně malé množství záznamů a nejsou k dispozici žádné další relevantní informace o charakteru jednotlivých zákazníků. Tabulka 4.2 obsahuje shrnutí z prvotní inspekce transakční databáze.

Z hlediska analýzy RFM je pro frekvenci nákupu určující počet provedených nákupů ve smyslu návštěv zákazníků v obchodě, při kterých došlo

Data pokrývají období	27. 11. 2014 – 22. 12. 2015
Počet dnů v pokrytém období	390
Celkový počet transakcí	3 294
Počet transakcí s nenulovou hodnotou nákupu	2 335
Celkový počet nakupujících zákazníků	987
Celkový počet nákupních košů	1 048
Celkový obrat všech transakcí	5 086 361

Tabulka 4.2: Shrnutí inspekce transakční databáze

Odvozený atribut	**Typ**	**Popis**
user_id	text	identifikátor zákazníka (nyní jako unikátní záznam)
recent_purchase	date	datum pořízení posledního nákupního koše
days_from_recent	numeric	počet dní uplynulých od posledního nákupu k 22. 12. 2015
num_of_purchases	numeric	počet pořízených nákupních košů za pokryté období
sum_total	numeric	celková hodnota nákupů za pokryté období

Tabulka 4.3: Seznam odvozených atributů

k zakoupení jednoho nebo více kusů libovolného zboží. Proto byla data z jednotlivých transakcí agregována do podoby tzv. *nákupních košů*.

Nákupní koš vznikl agregací transakcí za následujících podmínek:

1. Byly sdruženy všechny transakce se společnými hodnotami atributů *userid* a *timestamp*.

2. Byly započítány pouze transakce s nenulovou hodnotou nákupu, tzn. atribut $income > 0$.

Pro účely následující RFM analýzy byly z původních dat odvozeny další atributy. Viz tabulka 4.6.

Výsledná matice po předzpracování, které bylo provedeno pomocí agregačních funkcí dotazovacího nástroje *SELECT* jazyka SQL, obsahuje 987 záznamů – jeden pro každého klienta, který zakoupil alespoň jeden nákupní koš za pokryté období od 27. 11. 2014 do 22. 12. 2015. Ve sloupcích výsledné matice jsou odvozené atributy, které vznikly agregováním původních atributů.

Takto vzniklá výsledná matice slouží jako vstup pro RFM analýzu a reportování za pokryté období k datu 22. 12. 2015.

4.1.6 RFM analýza

4.1.6.1 Statistická analýza

Nejprve byla provedena statistická analýza hodnot atributů ve vstupní matici:

Atribut	Průměr	Medián	Sm. odchylka
days_from_recent	174,42	165	115
num_of_purchases	1,06	1	0,28
sum_total	5 150,61	915	14 820,42

Následuje diskuze závěrů vytvořených na základě zjištěných hodnot.

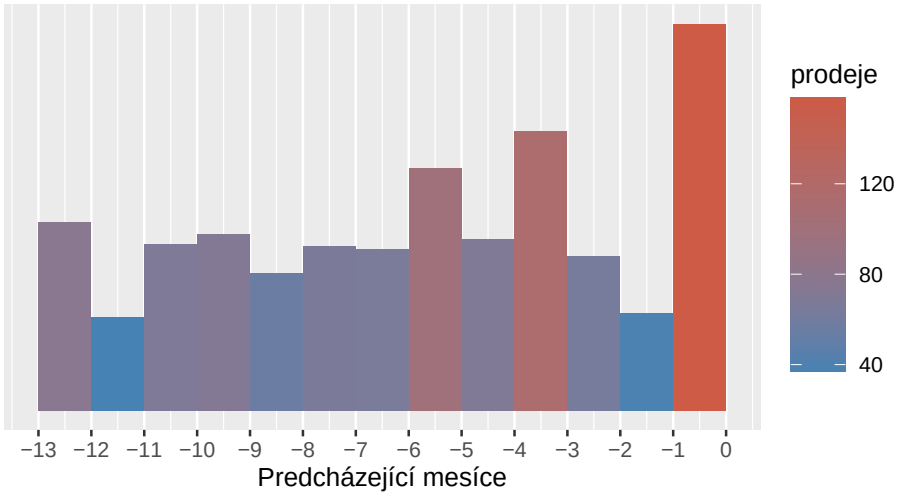
Je pouze malé množství zákazníků, kteří provedli více než jeden nákup. Typický zákazník obchodu je tedy charakterizován jako prvonákupce, který nákup provedl nahodile a nemá žádnou loajalitu k obchodu.

Frekvenční analýza ukazuje, že celkem je pouze 5,17% podíl zákazníků s opakovanými nákupy. Z celkového množství 987 zákazníků pouze 42 nakoupilo právě dvakrát a jen 9 provedlo tři a více opakovaných nákupů.

Pro RFM analýzu to v tomto případě znamená, že proměnná typu *frequency* bude mít minimální význam pro segmentaci zákazníků.

Důsledkem předchozího jevu je skutečnost, že atribut *days_from_recent*, tedy počet dní od posledního nákupu pro jednotlivé zákazníky, u drtivé většiny zároveň představuje období jediného provedeného nákupu.

Při vytvoření histogramu, kdy jsou jednotlivé hodnoty *days_from_recent* sdruženy do košů o velikosti odpovídajících délce jednoho měsíce, lze sledovat sezónnost prodejů. Viz graf 4.1.



Obrázek 4.1: Graf počtu prodejů za předcházející měsíce. Čas 0 na ose X představuje konec měsíce prosince 2015.

Z histogramu je patrný vliv předvánočního období v prosinci 2015, kdy je frekvence prodejů zhruba dvojnásobná oproti ostatním měsícům, kdy jsou četnosti prodejů víceméně rovnoměrné. Výjimku tvoří pouze měsíce červenec a září 2015.

4.1.6.2 RFM kvantilovou metodou

Jako první varianta byla zvolena základní podoba RFM analýzy při rozdělení do diskrétních hodnot 1-5 dle kvantilů po jednotlivých metrikách *recency, frequency a monetary* tak, jak je popsáno v kapitole 1.1.

Pro tento účel byly pro hodnoty atributů *days_from_recent*, *num_of_purchases* a *sum_safe* vypočítány hodnoty percentilů v rámci celé klientské báze a tyto percentily přiřazeny k jednotlivým zákazníkům jako odvozené atributy *r_perc*, *f_perc* a *m_perc*.

U hodnot *recency* percentily v obráceném pořadí, tzn. hodnota 1 odpovídá zákazníkovi s nejnovějším nákupem.

Z těchto percentilů byly poté odvozeny atributy *r_q*, *f_q* a *m_q*, které představují ohodnocení vzniklé na základě RFM analýzy:

Ohodnocení R (F, M resp.)	**Percentil**
1	$\leq$ 0,2
2	0,2–0,4
3	0,4–0,6
4	0,6–0,8
5	$\geq$ 0,8

4.1.6.3 RFM s expertním rozdělením

Provedena byla i druhá varianta ohodnocení klientů pomocí RFM analýzy na základě expertního rozdělení. Expertní rozdělení bylo provedeno obchodním analytikem:

Ohodnocení R	***days_from_recent***
1	$>$ 365
2	180–365
3	60–179
4	30–59
5	0–29

Ohodnocení F	***num_of_purchases***
1	1
2	2
3	3
4	4
5	≥ 5

Ohodnocení M	***sum_total***
1	$< 1\ 000$
2	1 000–5 000
3	5 000–10 000
4	10 000–20 000
5	$\geq 20\ 000$

Expertní rozdělení pro ohodnocení *monetary* bylo stanoveno na základě cen sortimentu, nabízeného v obchodě.

Vhodnost takto stanoveného expertního rozdělení bude diskutována dále.

4.1.6.4 Výsledky RFM analýzy

Vzhledem ke zjištění, že 95 % klientské základny provedlo za pokryté období pouze jediný nákup, bylo rozhodnuto, že při vizualizaci bude vynechána metrika *frequency* a zobrazení výsledků bude provedeno ve dvou dimenzích RM s tím, že klienti s opakovanými nákupy budou vyhodnoceni zvlášť.

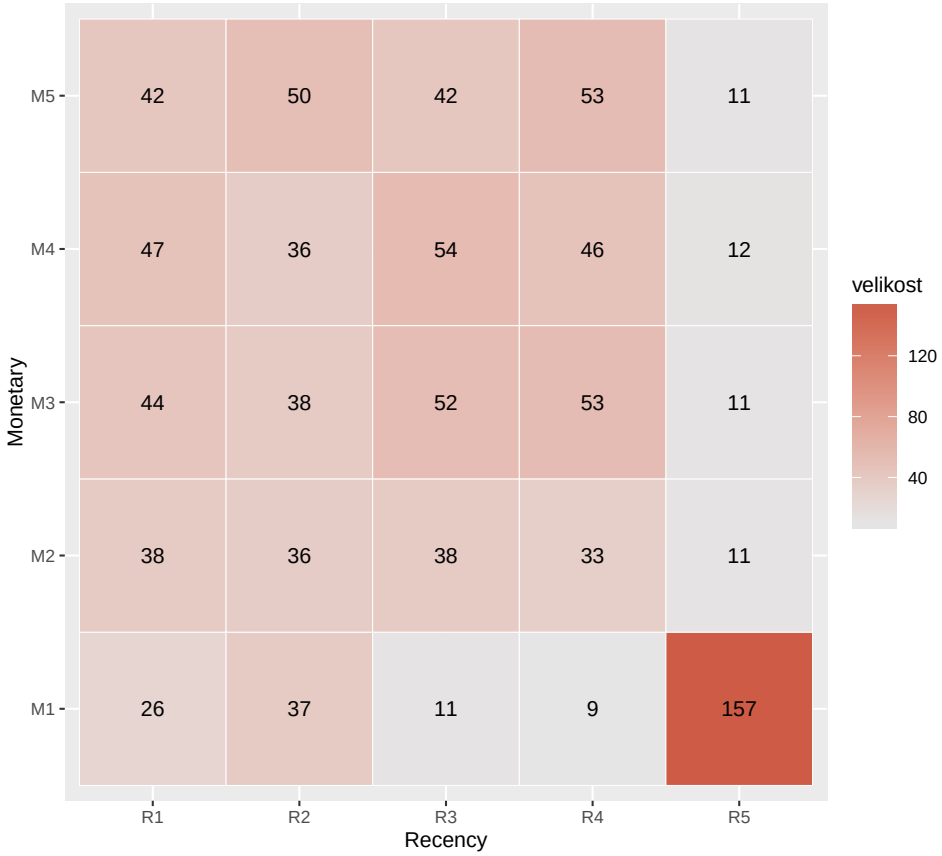
Následující grafy ukazují rozdělení četnosti zastoupení klientů v jednotlivých segmentech, které jsou dány příslušnými ohodnoceními R a M. Graf 4.2 ukazuje výsledky pro analýzu s kvantilovým rozdělením a graf 4.3 variantu s expertním rozdělením. Graf 4.4 potom ukazuje analýzu pouze pro klienty s opakovanými nákupy.

Obě metody shodně odhalili segment klientů (na grafech v pravém spodním rohu), kteří zřejmě navštívili ochod v období před Vánoci a provedli drobný nákup.

Zatímco kvantilová metoda ze své podstaty rozdělila klienty mezi segmenty rovnoměrně, expertní rozdělení umožňuje vizuálně jednodušeji separovat klienty dle obchodní logiky, ze které vychází. Interpretaci výsledků, která následuje, doprovází obrázek 4.5.

Například v horních dvou řádcích jsou klienti, kteří zakoupili produkty v hodnotě vyšší, než deset, respektive dvacet tisíc korun. Tyto hodnoty byly vybrány záměrně, protože odpovídají cenovým hladinám nabízených jízdních kol v nižší, respektive vyšší cenové kategorii. V prvních dvou řádcích jsou tedy klienti, kteří si s vysokou pravděpodobností zakoupili některé jízdní kolo.

Dále je patrné, že většina prodaných nákupních košů má hodnotu do 5 000 korun s převládajícími menšími nákupy do jednoho tisíce.



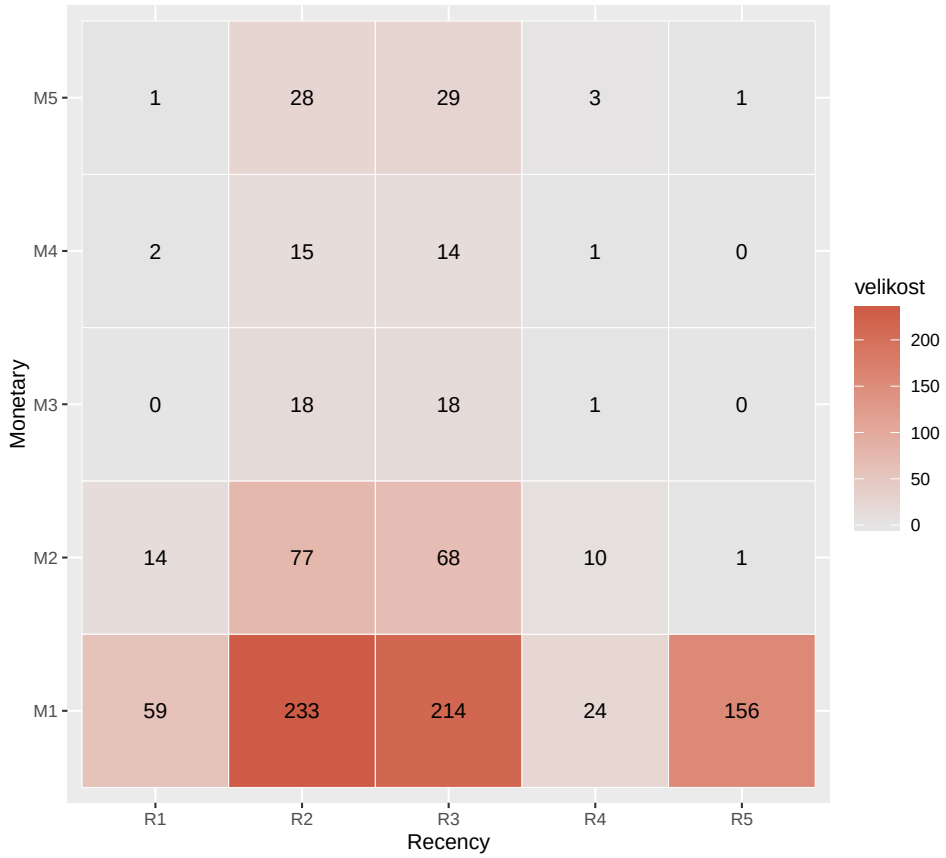
Obrázek 4.2: Rozdělení do RM segmentů kvantilovou metodou.

Největší skupinu představují klienti, kteří dříve v minulosti provedli drobný nákup. Zastoupení klientů s největší obchodní významností, kterou lze pomocí RFM ohodnocení aproximovat (viz. kapitola 1.3), je bohužel velice malé.

4.1.7 Další segmentace pomocí metody shlukování

Provedená vizualizace RFM analýzy v dimenzích metrik R a M nabídla možný způsob segmentace zákazníků a naznačila směry využitelné pro plánování a cílení obchodních kampaní.

Tato část studie se zabývá řešením problému snížení počtu zákaznických segmentů v souladu s postupy navrženého frameworku pro reportování, v tomto případě s využitím shlukování metodou K-means. Cílem je využít výstupů s předešlé RFM analýzy a dosáhnout menšího počtu segmentů pro snadnější dělení zákaznické báze a jednodušší orientaci.



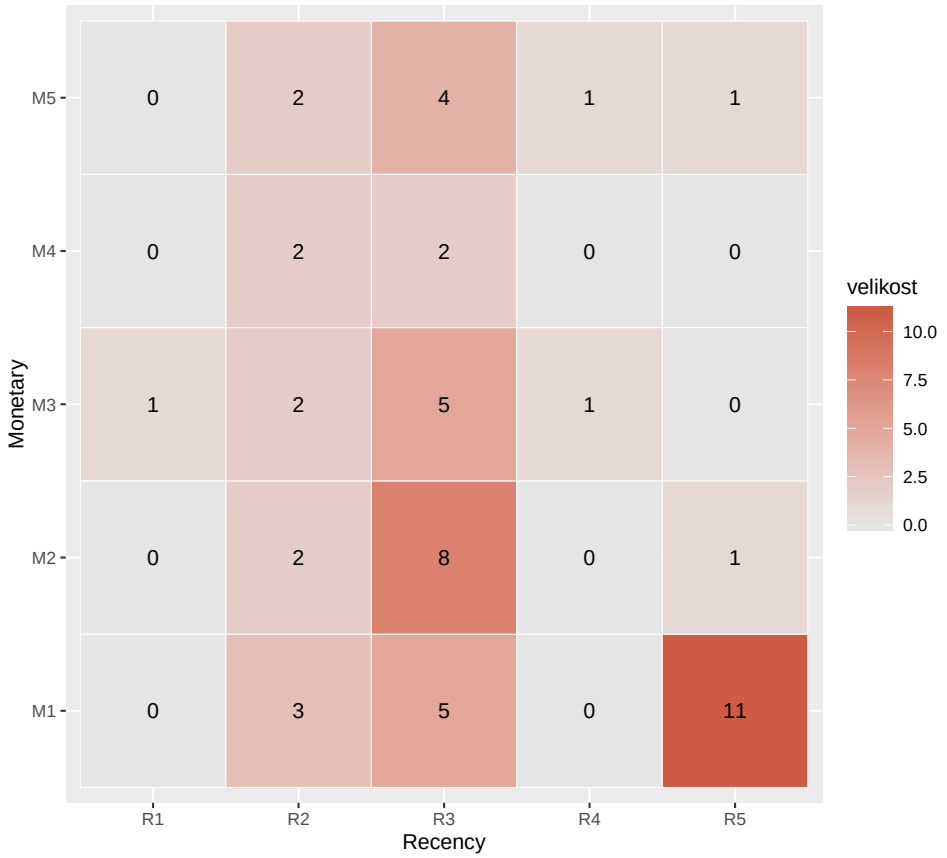
Obrázek 4.3: Rozdělení do RM segmentů dle expertního rozdělení.

4.1.7.1 Shlukování na základě ohodnocení RFM

V prvním kroku je nejprve experimentálně zjištěna kvalita shlukování na základě sumární čtvercové vzdálenosti každého pozorování od příslušného centroidu určeného shluku – *TWSS* (*total within sum of squares*). To je docíleno opakovaným měřením a průměrování pro zvolený rozsah k počtu shluků.

Vzhledem k požadované aplikaci shlukování, tedy významnému snížení počtu segmentů při zachování kvality segmentaci, bylo K zvoleno v rozsahu 2 až 8. Pruměrná hodnota total within sum of squares byla pro každé volené k počítána ze 100 nezávislých běhů algoritmu. Grafy 4.6 a 4.7 ukazují výsledky pro K-means na základě RFM analýzy kvantilového typu (RFM_Q) i s expertním rozdělením (RFM_E).

Vzhledem ke stejnému rozměru v rámci obou metod RFM lze poměřit kvalitu shlukování pro oba případy. Z grafů je pro shlukování K-means jednoznačně viditelná lepší použitelnost vstupu z RFM analýzy s expertním rozdělením. Pro další kroky bude tedy zvolena pouze tato varianta.



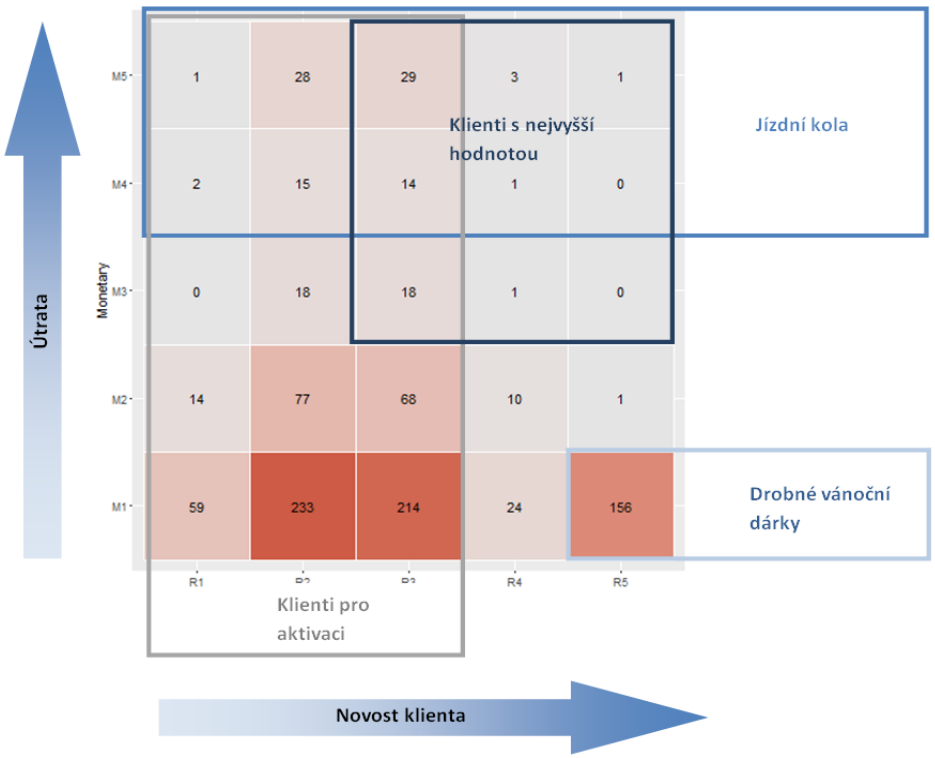
Obrázek 4.4: Rozdělení do RM segmentů dle expertního rozdělení pro zákazníky s opakovanými nákupy.

V dalším kroku je zvolena vhodná hodnota K. Z grafu 4.7 jsou viditelné významnější skoky ve kvalitě shlukování u K rovno 3, respektive 4. Při větším počtu shluků se již kvalita dále významně nezlepšuje.

V posledním kroku je provedeno samotné shlukování pro zvolená K, v tomto případě $K = 3$ a $K = 4$. Z pozic centroidů v dimenzích R, F a M a z velikostí výsledných shluků je nakonec možné interpretovat společné rysy zákazníků v rámci každého shluku. Viz tabulky 4.8 a 4.9.

Při analýze výsledných shluků lze snadno identifikovat segment zákazníků, kteří nakupovali v předvánočním období, stejně tak segment zákazníků, kteří utratili nejvíce peněz – nákupců kol, dražšího zboží nebo klientů, kteří se do obchodu vrátili pro další nákupy.

Shlukování se čtyřmi segmenty navíc rozděluje zbytek zákazníků mezi segment klientů, kteří poslední nákup provedli během uplynulých 6 měsíců a jsou tedy vhodnější pro nějakou aktivační kampaň, a segment klientů, kteří nakoupili už před poměrně dávnou dobou a v souvislosti s tzv. *customer attri-*



Obrázek 4.5: Rozdělení do RM segmentů dle expertního rozdělení – interpretace.

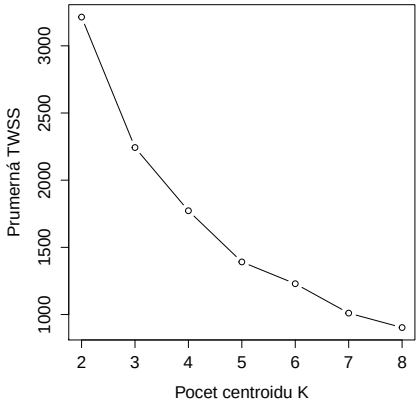
tion jsou označování slangově jako „mrtví“. Obchodní potenciál u takovýchto klientů bývá spíše nízký.

4.1.7.2 Shlukování na základě původních atributů

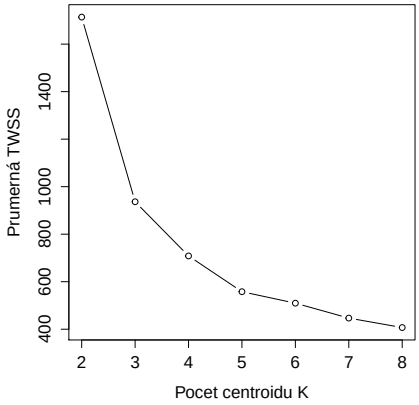
Předchozí varianta shlukování na základě ohodnocení z RFM analýzy umožňuje snadnou interpretaci shluků, alespoň tedy v případě menšího počtu zvolených segmentů. V další variantě nasazení shlukování bude zkoumáno, jaký vliv bude mít, pokud se vynechá krok samotné RFM analýzy a na vstupu pro algoritmus K-means bude nediskretizovaný vstup.

Grafy 4.10 a 4.11 ukazují porovnání kvality shlukování pro obě řešení – vstup s percentily (RFM_P) a vstup s originálními – „raw“ hodnotami (RFM_R).

Ani jedna z metod nedokázala dosáhnout stejné kvality segmentace pro $K = 3$, jako shlukování typu RFM_E z předchozí varianty. Pro vzrůstající K bylo dosaženo obdobné kvality jako RFM_E. V případě RFM_P pro $K = 5$, v případě RFM_R pro $K = 4$. Tabulky 4.12 a 4.13 popisují shluky pro zvolená



Obrázek 4.6: Kursport - RFM_Q Graf závislosti TWSS na K



Obrázek 4.7: Kursport - RFM_E Graf závislosti TWSS na K

K = 3		Souřadnice centroidů			Popis shluku
Shluk	Velikost	R	F	M	
1	188	4.83	1.08	1.07	*Předvánoční drobné nákupy*
2	647	2.32	1.03	1.46	*Mainstream, prvonákupčí, k aktivaci*
3	152	2.53	1.18	4.09	*Hodnotní zákazníci, opakované nákupy, kola*

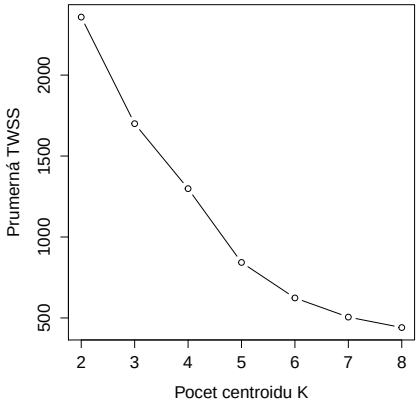
Obrázek 4.8: Tabulka segmentů pro $K = 3$, RFM analýza s expertním rozdělením

K = 4		Souřadnice centroidů			Popis shluku
Shluk	Velikost	R	F	M	
1	128	2.66	1.17	4.28	*Hodnotní zákazníci, opakované nákupy, kola*
2	454	2.61	1.04	1.31	*Mainstream, prvonákupčí, k aktivaci*
3	188	4.84	1.08	1.07	*Předvánoční drobné nákupy*
4	217	1.66	1.03	1.94	*Mrtví klienti*

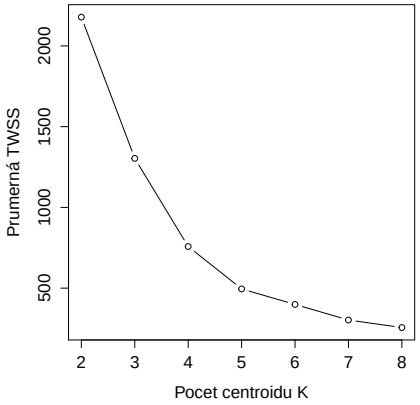
Obrázek 4.9: Tabulka segmentů pro $K = 4$, RFM analýza s expertním rozdělením

K.

Shlukování na základě percentilů dokázalo dobře identifikovat segment zákazníků s opakovanými nákupy. Podobně jako shlukování RFM_E z předchozí varianty zahrnuje segment klientů, kteří prováděli drobné předvánoční nákupy a segment neperspektvních „mrtvých“ klientů. Zbytek rozděluje „lepší“ a „horší“ segment z hlediska jak rozměru *recency*, tak *monetary*. Segment opravdu hodnotných klientů není zvlášť vymezen, ale je částečně reprezento-



Obrázek 4.10: Kursport - RFM_P Graf závislosti TWSS na K



Obrázek 4.11: Kursport - RFM_R Graf závislosti TWSS na K

RFM_P K = 5		Souřadnice centroidů			Popis shluku
Shluk	Velikost	R*	F*	M*	
1	229	1.83	1.00	3.99	*Mrtví klienti*
2	301	3.56	1.00	3.66	*Lepší mainstream, prvonákupčí, k aktivaci*
3	223	2.10	1.00	1.94	*Horší mainstream, prvonákupčí, neaktivita*
4	183	4.53	1.00	1.17	*Předvánoční drobné nákupy*
5	51	3.56	4.83	3.49	*Klienti s opakovanými nákupy*

Obrázek 4.12: Tabulka segmentů pro $K = 5$, shlukování s percentilovým vstupem. *Souřadnice v prostoru 5x5x5, který vychází z původního prostoru RFM

ván segmentem zákazníků s opakovanými nákupy.

Řešení se vstupem ve formě původních hodnot má z hlediska kvality shlukování stejně dobré, jako RFM_E z předchozí varianty, nicméně interpretovatelnost samotných segmentů je podstatně složitější.

To komplikuje přímé použití pro plánování obchodní strategie, nicméně se může ukázat užitečné po tom, co bude na jeho základě provedeno nějaké sondování zákaznické báze a vyhodnocena úspěšnost respondibility pro výsledné shluky.

Toto platí ostatně pro všechny segmenty vzniklé pomocí různých metod, ať už z těch uvedených v této studii, nebo dalších jiných. Nicméně segmentace s využitím metody K-means na základě ohodnocení z předchozí RFM analýzy umožňuje kvalitní shlukování při nízkém počtu shluků a především jednoduchou interpretovatelnost vzniklých segmentů na základě vhodně zvoleného expertního rozdělení.

RFM_R K = 4		Souřadnice centroidů			Popis shluku
Shluk	Velikost	R*	F*	M*	
1	349	1.00	4.88	4.55	
2	67	1.39	5.00	1.65	
3	106	2.10	1.29	4.97	
4	465	5.00	1.42	1.16	

Obrázek 4.13: Tabulka segmentů pro $K = 4$, shlukování se vstupem původních hodnot.

4.1.8 Závěry analýzy

Případová studie se zabývala analýzou transakčních dat obchodu za období mezi 27. 11. 2014 a 22. 12. 2015. Toto období zahrnuje přes 2 300 transakcí s nenulovou hodnotou nákupu, které byly provedeny v rámci více než tisíce nákupních košů. Bylo rozpoznáno bezmála tisíc nakupujících klientů.

Statistická analýza mezi nimi odhalila zhruba 5 % klientů, kteří nakupovali opakovaně a byla zjištěna výrazná sezónnost prodejů, především v souvislosti s předvánočními obdobími.

Následně byla provedena RFM analýza a klientská báze rozdělena do 25 RM segmentů dle variant RFM kvantilovou metodou a s expertním rozdělením. Výstupy z analýzy sloužily také jako vstup pro další shlukování, kdy bylo v několika variantách navrženo rozdělení zákaznické báze do 3 až 5 segmentů. Segmentace na základě RFM analýzy a shlukování byla doplněna interpretací s využitím znalosti expertního rozdělení.

Vhodně zvolené expertní rozdělení přináší do procesu jistou apriorní informaci, která se ukazuje jako podstatná pro kvalitu shlukování a vyvažuje ztrátu informace, ke které nutně dochází při předcházející RFM analýze. Výsledkem je v této studii výhodný trade-off mezi přesností modelu a jeho použitelnosti ve smyslu interpretovatelnosti z pohledu obchodu.

Podoba rozvržení zákaznické báze po RM segmentech tvoří indikátor úspěšnosti obchodu a v kombinaci s vhodnou interpretací slouží jako zdroj pro doporučení zavedení obchodních strategií pro vytěžení stávající zákaznické báze.

Vizualizace výsledků dílčích kroků analýzy slouží jako prostředek pro reportování majitelům dat.

4.2 Případová studie 2 – Huskycz.cz

4.2.1 Popis a charakteristika

Husky cz, s.r.o. je obchod s outdoorovým zbožím, který se zaměřuje především na prodej oblečení, turistického vybavení a příslušenství. Mimo desítek kamenných prodejen provozuje také internetový prodej na adrese *www.huskycz.cz*. Tato studie se zabývá analýzou prodejních dat právě tohoto internetového obchodu (dále jen obchodu). Stejně jako v předchozím případě je transakční databáze spravována v rámci systému společnosti *CloudSailor* a data pro analýzu byla dále dodána pro potřeby analytického systému společnosti *Recombee*, která pak data ve specifikované podobě poskytla pro tuto studii.

4.2.2 Cíle prováděné studie

Cílem studie je vyšetřit transakční (prodejní) data obchodu pro možnosti analýzy zákaznických segmentů a využít postupy navrhovaného frameworku (viz kapitola 3) pro reportování užitečných informací majitelům dat. Dále si potom studie klade za cíl využít zjištěné znalosti a navrhnout doporučení pro zavedení obchodních strategií vhodných pro vytěžení stávající klientské báze.

Vzhledem k obsáhlejší transakční databázi a větší klientské bázi obchodu bylo zvoleno nasazení navrhovaného frameworku včetně dodatečných analýz dle kapitol 3.3 a 3.4.

4.2.3 Použité nástroje

Pro extrakci dat byly využity nástroje pro zálohování a obnovu databázového stroje *PostgreSQL 9.5* a tento tento databázový systém byl dále využit pro datové transformace a nápočty v sekcích 4.2.4, 4.2.5, 4.2.8 a 4.2.9 pomocí dotazovacího nástroje *SELECT*.

RFM analýza v sekci 4.2.6 a následné shlukování 4.2.7 byla provedena pomocí statistického softwaru *R (verze 3.2.3)* s použitím vývojového prostředí *RStudio* a s podporou tabulkového procesoru *MS Excel 2007*, stejně tak jako analýza nákupních košů 4.2.8 a odhadování CLV v sekci 4.2.9.

Všechny vizualizace byly rovněž vygenerovány v softwaru *R*, konkrétně pomocí balíčků *ggplot2* a *arulesViz*.

4.2.4 Extrakce dat

Extrakce dat byla provedena stejným způsobem jako v případě předcházející studie, viz kapitola 4.1.3. Přejaté datové schéma je ve stejném formátu:

items - seznam nabízených produktů

users - seznam uživatelů (zákazníků) obchodu

Atribut	Typ	Popis	Poznámka
itemid	text	identifikátor nabízeného produktu	prim. klíč
name	text	název produktu a kategorie	kategorii lze parsovat
annotation	text	popis produktu	
priceVat	numeric	cena produktu v Kč	

Tabulka 4.4: Struktura tabulky *items*

purchases - seznam uskutečněných prodejů

Mimo tabulky *purchases*, která obsahuje záznamy o jednotlivých prodejích, je v rámci této studie pro účely analýzy nákupních košu využita také tabulka *items*. Ta obsahuje informace o nabízených produktech.

Struktura databázové tabulky *items* je uvedena v tabulce 4.4

4.2.5 Inspekce a předzpracování dat pro RFM analýzu

Navržený analytický framework využívá RFM analýzu jako základní nástroj pro analýzu zákaznických segmentů a reportování z transakčních dat libovolného, tedy i menšího rozsahu bez dalších znalostí o zákaznících nebo produktech. To je případ této studie, kdy nejsou k dispozici žádné další relevantní informace o charakteru jednotlivých zákazníků. Tabulka 4.5 obsahuje shrnutí z prvotní inspekce transakční databáze.

Data pokrývají období	13. 5. 2014 – 9. 3. 2016
Počet dnů v pokrytém období	667
Celkový počet transakcí	68 614
Počet transakcí s nenulovou hodnotou nákupu	65 968
Celkový počet nakupujících zákazníků	20 411
Celkový počet nákupních košů	24 395
Celkový obrat všech transakcí	48 901 677

Tabulka 4.5: Shrnutí inspekce transakční databáze

Stejně jako v případě předcházející studie byla data z jednotlivých transakcí agregována do podoby tzv. *nákupních košů*. Viz kapitola 4.1.5.

Pro účely následující RFM analýzy byly z původních dat odvozeny další atributy. Viz tabulka 4.6.

Výsledná matice po předzpracování, které bylo provedeno pomocí agregačních funkcí dotazovacího nástroje *SELECT* jazyka *SQL*, obsahuje 20 411 záznamů – jeden pro každého klienta, který zakoupil alespoň jeden nákupní koš za pokryté období od 13. 5. 2014 do 9. 3. 2016. Ve sloupcích výsledné matice jsou odvozené atributy, které vznikly agregováním původních atributů.

Odvozený atribut	Typ	Popis
user_id	text	identifikátor zákazníka (nyní jako unikátní záznam)
recent_purchase	date	datum pořízení posledního nákupního koše
days_from_recent	numeric	počet dní uplynulých od posledního nákupu k 9. 3. 2016
num_of_purchases	numeric	počet pořízených nákupních košů za pokryté období
sum_total	numeric	celková hodnota nákupů za pokryté období

Tabulka 4.6: Seznam odvozených atributů

Takto vzniklá výsledná matice slouží jako vstup pro RFM analýzu a reportování za pokryté období k datu 9. 3. 2016.

Graf 4.14 je vizualizací výsledné matice. Každý bod představuje jednoho zákazníka, jeho pozice na ose X a Y značí hodnotu *days_from_recent*, respektive *sum_total*. Barevnou škálou jsou vyznačeni klienti se dvěma a více nákupy – *num_of_purchases*.

4.2.6 RFM analýza

4.2.6.1 Statistická analýza

Nejprve byla provedena statistická analýza hodnot atributů ve vstupní matici:

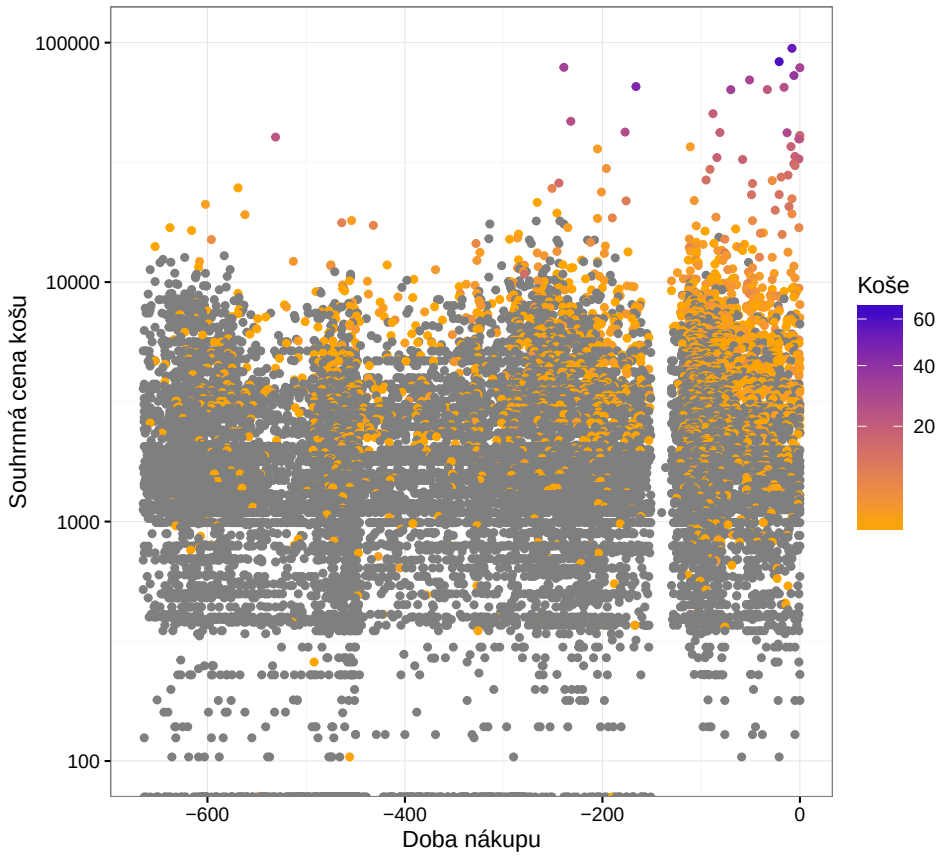
Atribut	Průměr	Medián	Sm. odchylka
days_from_recent	314,57	283	194,2
num_of_purchases	1,2	1	1,46
sum_total	2 395,85	1780	3 425,31

Následuje diskuze závěrů vytvořených na základě zjištěných hodnot.

1. Je pouze malé množství zákazníků, kteří provedli více než jeden nákup. Typický zákazník obchodu je tedy charakterizován jako prvonákupce, který nákup provedl nahodile a nemá žádnou loajalitu k obchodu.

2. Frekvenční analýza ukazuje, že celkem je pouze 10,45% podíl zákazníků s opakovanými nákupy. Z celkového množství 20 411 zákazníků pouze 7,8 % nakoupilo právě dvakrát a jen 2,7 % provedlo tři a více opakovaných nákupů.

Pro RFM analýzu to v tomto případě znamená, že proměnná typu *frequency* bude mít relativně malý význam pro segmentaci zákazníků. Zákazníci s opa-



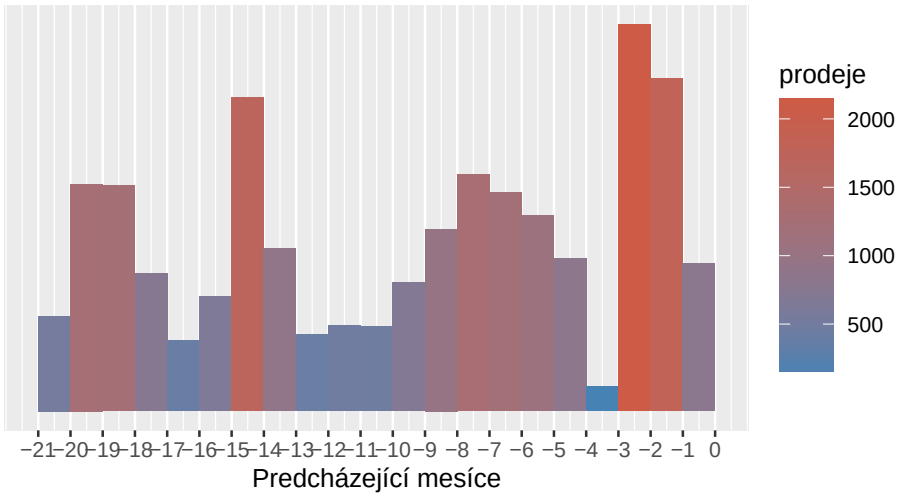
Obrázek 4.14: Graf zákaznické báze – osa X sleduje počet dní uplynulých od posledního nákupu a osa Y souhrnnou cenu všech zakoupených košů. Barevně jsou zobrazeni klienti se dvěma a více nakoupenými koši.

kovanými nákupy budou v rámci segmentace přirozeně tvořit specifickou skupinu.

Důsledkem předchozího jevu je skutečnost, že atribut *days_from_recent*, tedy počet dní od posledního nákupu pro jednotlivé zákazníky, u drtivé většiny zároveň představuje období jediného provedeného nákupu.

Při vytvoření histogramu, kdy jsou jednotlivé hodnoty *days_from_recent* sdruženy do košů o velikosti odpovídajících délce jednoho měsíce, lze sledovat sezónnost prodejů. Viz graf 4.15.

Z histogramu je patrná sezónnost předvánočního a povánočního období v prosinci a lednu 2015/2016 a předvánočního období v prosinci 2014. Pozitivní vliv letní sezóny je možné pozorovat mezi červnem a zářím 2015, respektive červencem a srpnem 2014. Jako prodejně neúspěšné se potom jeví období podzimních měsíců v roce 2014 a především jarní měsíce v roce 2015.



Obrázek 4.15: Graf počtu prodejů za předcházející měsíce. Čas 0 na ose X představuje konec měsíce února 2016.

4.2.6.2 RFM kvantilovou metodou

Jako první varianta byla zvolena základní podoba RFM analýzy při rozdělení do diskrétních hodnot 1-5 dle kvantilů po jednotlivých metrikách *recency, frequency a monetary* tak, jak je popsáno v kapitole 1.1.

Pro tento účel byly pro hodnoty atributů *days_from_recent*, *num_of_purchases* a *sum_safe* vypočítány hodnoty percentilů v rámci celé klientské báze a tyto percentily přiřazeny k jednotlivým zákazníkům jako odvozené atributy *r_perc*, *f_perc* a *m_perc*.

U hodnot *recency* percentily v obráceném pořadí, tzn. hodnota 1 odpovídá zákazníkovi s nejnovějším nákupem.

Z těchto percentilů byly poté odvozeny atributy *r_q*, *f_q* a *m_q*, které představují ohodnocení vzniklé na základě RFM analýzy:

Ohodnocení R (F, M resp.)	Percentil
1	$\leq$ 0,2
2	0,2–0,4
3	0,4–0,6
4	0,6–0,8
5	$\geq$ 0,8

4.2.6.3 RFM s expertním rozdělením

Provedena byla i druhá varianta ohodnocení klientů pomocí RFM analýzy na základě expertního rozdělení. Expertní rozdělení bylo provedeno obchodním

analytikem:

Ohodnocení R	***days_from_recent***
1	> 365
2	180–365
3	60–179
4	30–59
5	0–29

Ohodnocení F	***num_of_purchases***
1	1
2	2
3	3
4	4
5	≥ 5

Ohodnocení M	***sum_total***
1	$< 1\,000$
2	1 000–2 000
3	2 000–3 000
4	3 000–4 500
5	$\geq 4\,500$

Expertní rozdělení pro ohodnocení *monetary* bylo stanoveno na základě cen sortimentu, nabízeného v obchodě.

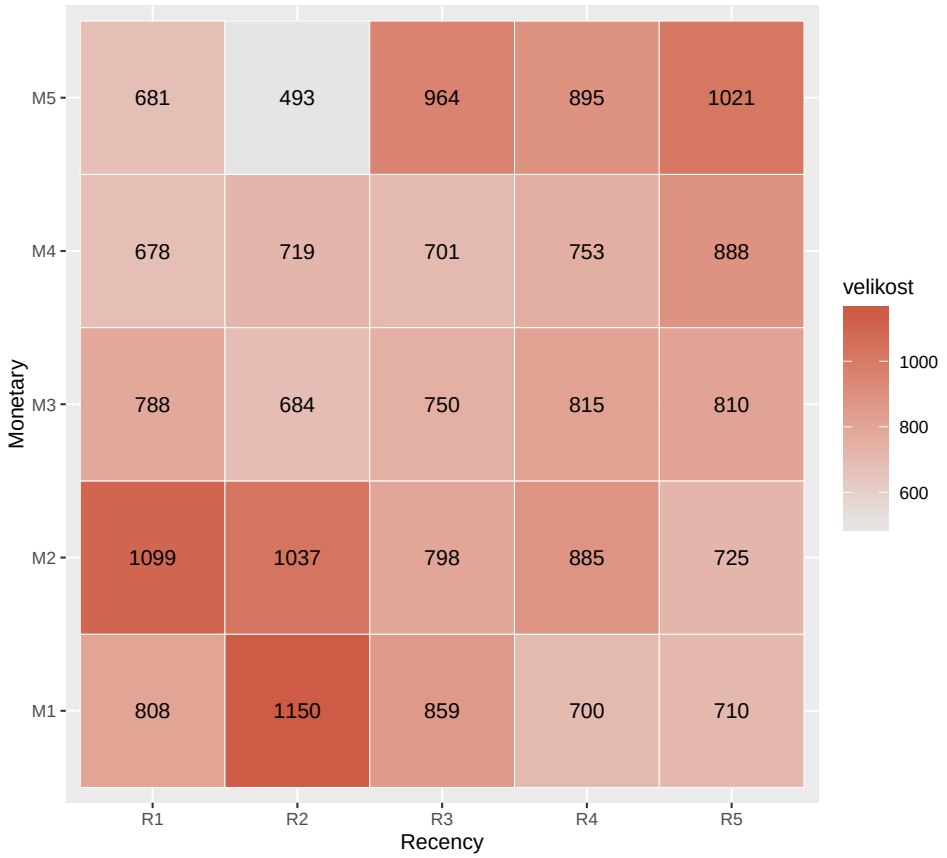
Vhodnost takto stanoveného expertního rozdělení bude dále diskutována.

4.2.6.4 Výsledky z RFM analýzy

Vzhledem ke zjištění, že 90 % klientské základny provedlo za pokryté období pouze jediný nákup, bylo rozhodnuto, že zobrazení výsledků bude provedeno ve dvou dimenzích RM na osách X a Y. Třetí dimenze pro *frequency* bude znázorněna barevnou škálou.

Následující grafy ukazují rozdělení četnosti zastoupení klientů v jednotlivých segmentech, které jsou dány příslušnými ohodnoceními R a M. Grafy 4.16 a 4.17 ukazují výsledky pro analýzu s kvantilovým rozdělením a grafy 4.18 a 4.19 variantu s expertním rozdělením. Grafy 4.20 a 4.21 potom sdružují naráz všechny 4 sledované dimenze: pozice na X, Y čtverci odpovídá danému segmentu dle *recency* a *monetary*, barva pozadí pole je ve škále dle průměrné hodnoty *frequency* příslušného RM segmentu a velikost a barva malých obdelníků zobrazuje velikost příslušných RM segmentů.

Zatímco kvantilová metoda ze své podstaty rozdělila klienty mezi segmenty rovnoměrně, expertní rozdělení umožňuje vizuálně jednodušeji separovat kli-



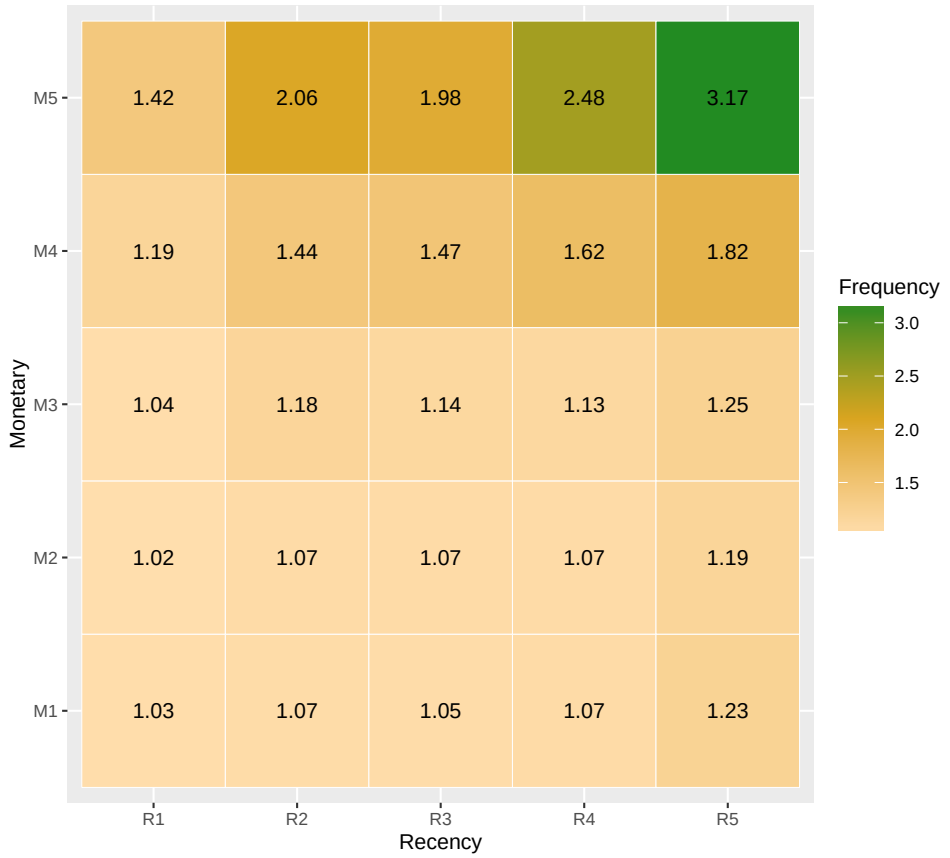
Obrázek 4.16: Rozdělení do RM segmentů kvantilovou metodou (RM_q).

enty dle obchodní logiky, ze které vychází. Interpretaci výsledků, která následuje, doprovází obrázek 4.22.

Například v horním řádku tabulky jsou klienti, kteří zakoupili produkty v hodnotě vyšší než 4,5 tisíce korun. Tato hodnota byla vybrána záměrně, protože odpovídá cenové hladině prémiových výrobků nejvyšších tříd pro sportovce a náročné klienty. V prvním řádku jsou tedy lukrativní klienti, kteří si v minulosti nakoupili prémiové zboží, případně provedli hned několik nákupů.

Dále je patrné, že velká část nákupních košů má hodnotu do 2 000 korun, kde převládají nákupu mezi 1 a 2 tisíci nad nejdrobnějšími nákupy do 1 tisíce.

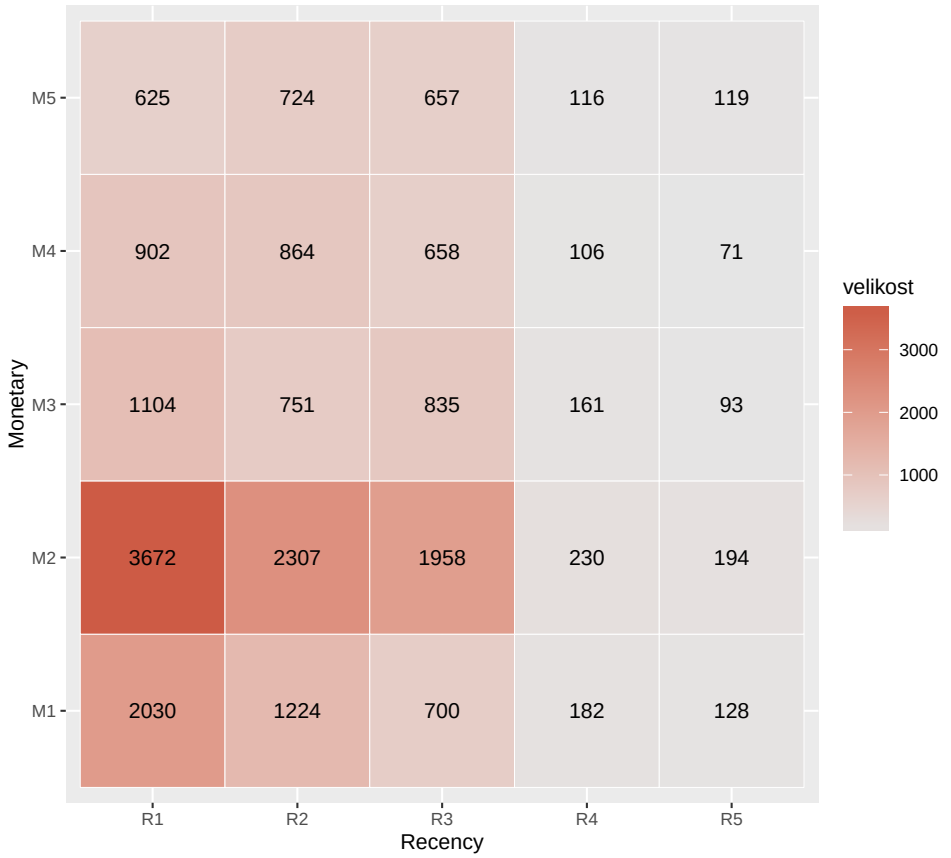
Početně významnou skupinu tvoří klienti v segmentech s ohodnocením *recency* 1 až 3, tedy klienti, kteří nakoupili naposledy dříve než před 2 měsíci. Klienty patřící do segmentů v prvním sloupci lze označit z obchodního hlediska spíše za „mrtvé“ a pravděpodobnost jejich respondibility v rámci reklamních kampaní je spíše nízká. Podstatnou část klientské základny však tvoří klienti s posledním nákupem mezi 60 a 180 dny, kteří jsou vhodní kandidáti pro reklamní kampaně vytvořené za účelem aktivace.



Obrázek 4.17: Rozvržení průměrných hodnot Frequency pro segmenty *RM_q* kvantilovou metodou.

Z grafu 4.19 lze také vyčíst pozitivní korelace mezi *frequency*, která představuje loajalitu a návratovost zákazníků, a *recency*, respektive *monetary*. Jinými slovy, zákazníky s vyšší loajalitou ve smyslu opakovaného nákupu lze najít v segmentech klientů s vyššími hodnotami útraty a s nákupy v nedávné době.

Tyto segmenty jsou na grafu zastoupeny v pravém horním rohu. Jedná se však o relativně malé segmenty a dohromady obsahují jen malý díl zákaznické báze. Cílem aktivačních a prodejních kampaní je přesunout klienty právě do těchto segmentů a zároveň tyto nejcennější klienty udržet.



Obrázek 4.18: Rozdělení do RM segmentů dle expertního rozdělení (RM_e).

4.2.7 Odhad významnosti klientů pomocí WRFM

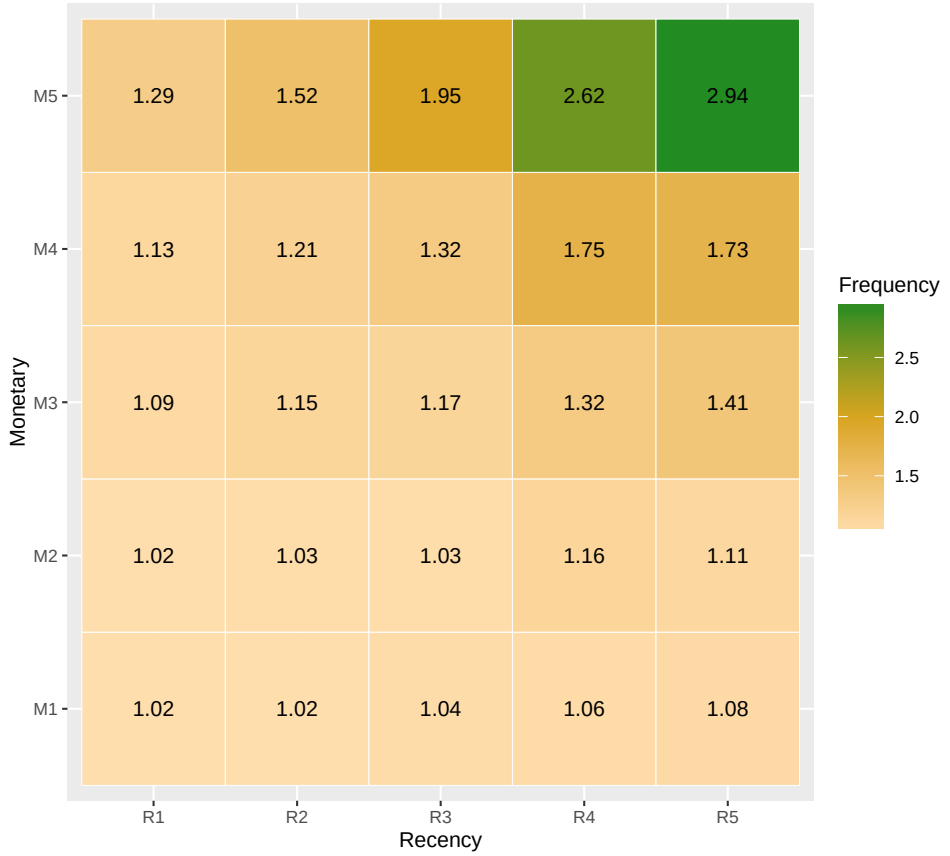
Pro účely zjištění vah jednotlivých složek RFM byla aplikována metoda AHP. Čtyři obchodem vybraní experti poskytnuli vyplněné dotazníky pro porovnání složek z hlediska významnosti po dvojicích.

Tři odpovědi, které prošli testem na konzistenci tvořily vstup pro vyhodnocení určení vah. Výsledky od jednotlivých expertů byly sloučeny a byly stanoveny následující váhy:

Váha	**Hodnota**
W_R	0,136
W_F	0,166
W_M	0,694

Z výsledků vyplívá, že expertní názor významně upřednostňuje podíl složky *monetary* nad ostatními.

Na grafu 4.23 je v barevné škále zobrazena průměrná hodnota



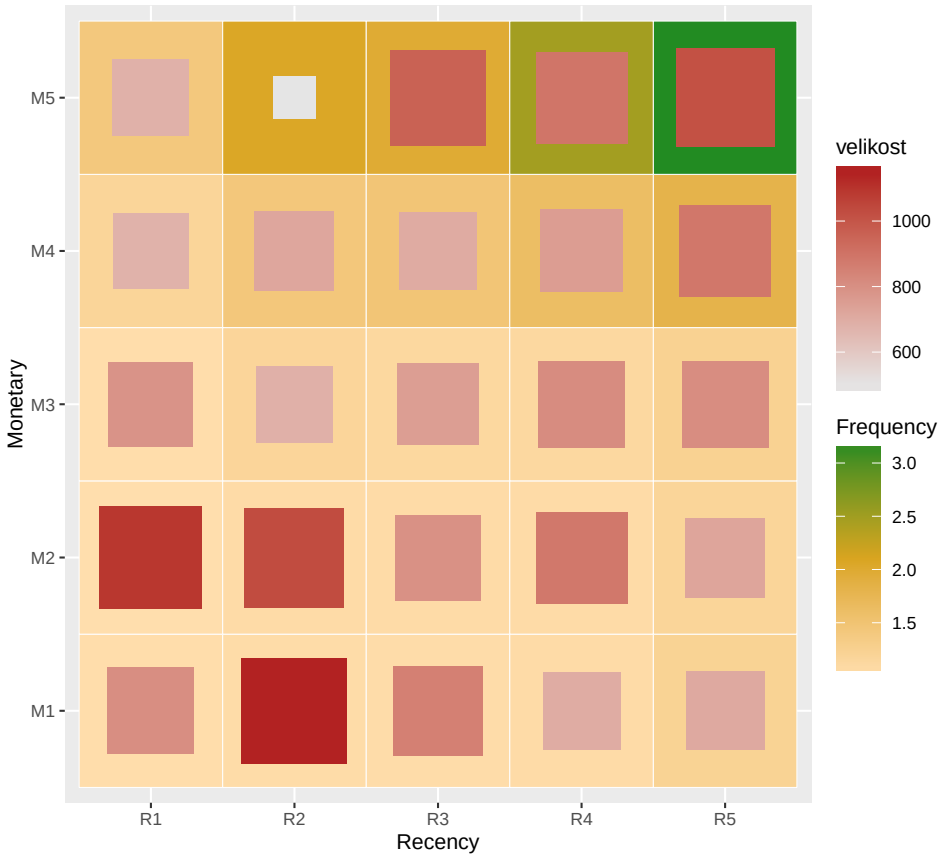
Obrázek 4.19: Rozvržení průměrných hodnot Frequency pro segmenty *RM_e*

$score = V_{RFM} \cdot 100$, která představuje odhad významnosti klientů v jednotlivých RM segmentech.

4.2.8 Další segmentace pomocí metody shlukování

Provedená vizualizace RFM analýzy v dimenzích metrik R a M nabídla možný způsob segmentace zákazníků a naznačila směry využitelné pro plánování a cílení obchodních kampaní. Obecná nevýhoda při využití segmentů, které vznikli rozdělením na základě ohodnocení R, F a M je však velké množství výsledných segmentů, které však pravděpodobně obsahují zákazníky s podobným chováním.

Tato část studie se zabývá řešením problému snížení počtu zákaznických segmentů v souladu s postupy navrženého frameworku pro reportování, v tomto případě s využitím shlukování metodou K-means. Cílem je využít výstupů s předešlé RFM analýzy a dosáhnout menšího počtu segmentů pro snadnější dělení zákaznické báze a jednodušší orientaci.



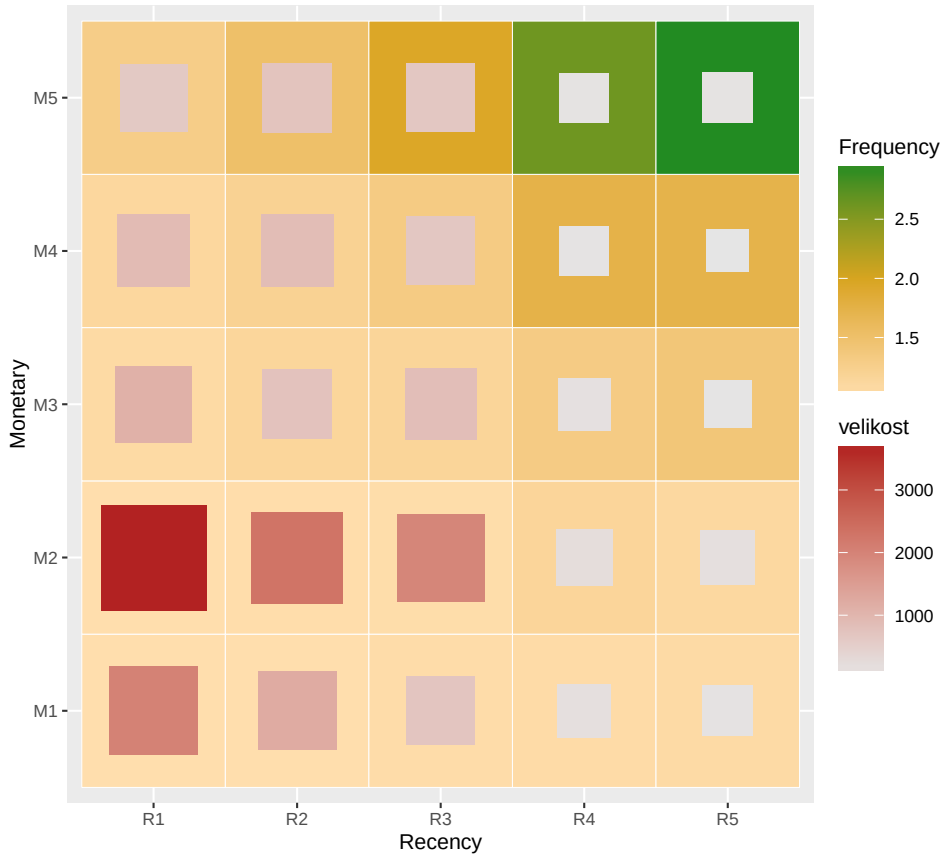
Obrázek 4.20: *RFM_q*. Barva pozadí a čtverce ukazuje průměrnou hodnotu *frequency*, resp. velikost segmentu.

4.2.8.1 Shlukování na základě ohodnocení RFM

V prvním kroku je nejprve experimentálně zjištěna kvalita shlukování na základě sumární čtvercové vzdálenosti každého pozorování od příslušného centroidu určeného shluku – *TWSS* (*total within sum of squares*). To je docíleno opakovaným měřením a průměrování pro zvolený rozsah k počtu shluků.

Vzhledem k požadované aplikaci shlukování, tedy významnému snížení počtu segmentů při zachování kvality segmentaci, bylo K zvoleno v rozsahu 2 až 8. Pruměrná hodnota total within sum of squares byla pro každé volené k počítána ze 100 nezávislých běhů algoritmu. Grafy 4.24 a 4.25 ukazují výsledky pro K-means na základě RFM analýzy kvantilového typu (RFM_Q) i s expertním rozdělením (RFM_E).

Vzhledem ke stejnému rozměru v rámci obou metod RFM lze poměřit kvalitu shlukování pro oba případy. Z grafů je pro shlukování K-means jednoznačně viditelná lepší použitelnost vstupu z RFM analýzy s expertním rozdě-



Obrázek 4.21: *RFM_e*. Barva pozadí a čtverce ukazuje průměrnou hodnotu *frequency*, resp. velikost segmentu.

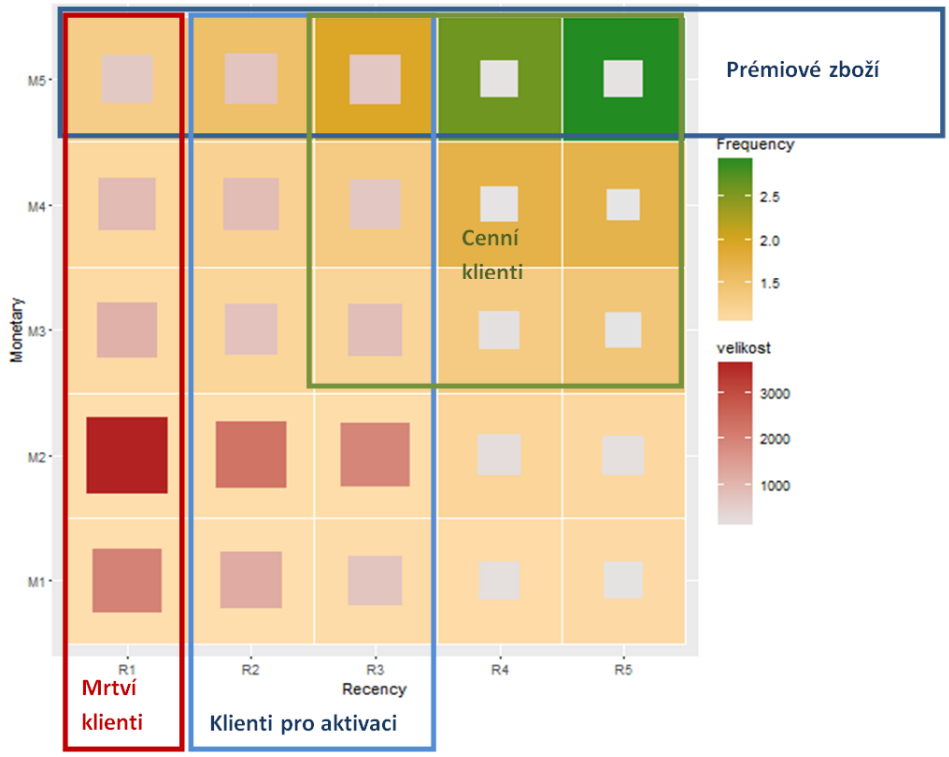
lením. Pro další kroky bude tedy zvolena pouze tato varianta.

V dalším kroku je zvolena vhodná hodnota K. Z grafu 4.7 jsou viditelné významnější skoky ve kvalitě shlukování u K rovno 3, respektive 4. Při větším počtu shluků se již kvalita dále významně nezlepšuje.

V posledním kroku je provedeno samotné shlukování pro zvolená K, v tomto případě $K = 3$ a $K = 4$. Z pozic centroidů v dimenzích R, F a M a z velikostí výsledných shluků je nakonec možné interpretovat společné rysy zákazníků v rámci každého shluku. Viz tabulky 4.26 a 4.27.

Při analýze výsledných shluků lze snadno identifikovat segment zákazníků, kteří prováděli větší, případně opakované nákupy a utratili větší množství peněz. Dále je zde viditelný segment s nízkou hodnotou průměrnou *recency*, tedy obsahující významné množství zákazníků, kteří nakoupili dříve než před 1 rokem a z hlediska *customer attrition* je možné je označit za tzv. „mrtvé“ klienty. Obchodní potenciál u takovýchto klientů bývá spíše nízký.

Shlukování se čtyřmi segmenty navíc odhaluje segment klientů, kteří po-



Obrázek 4.22: Rozdělení do RM segmentů dle expertního rozdělení – interpretace.

slední nákup provedli během uplynulých 6 měsíců a jsou tedy vhodnější pro aktivační kampaň. Tento segment se také vyznačuje spíše drobnějšími nákupy.

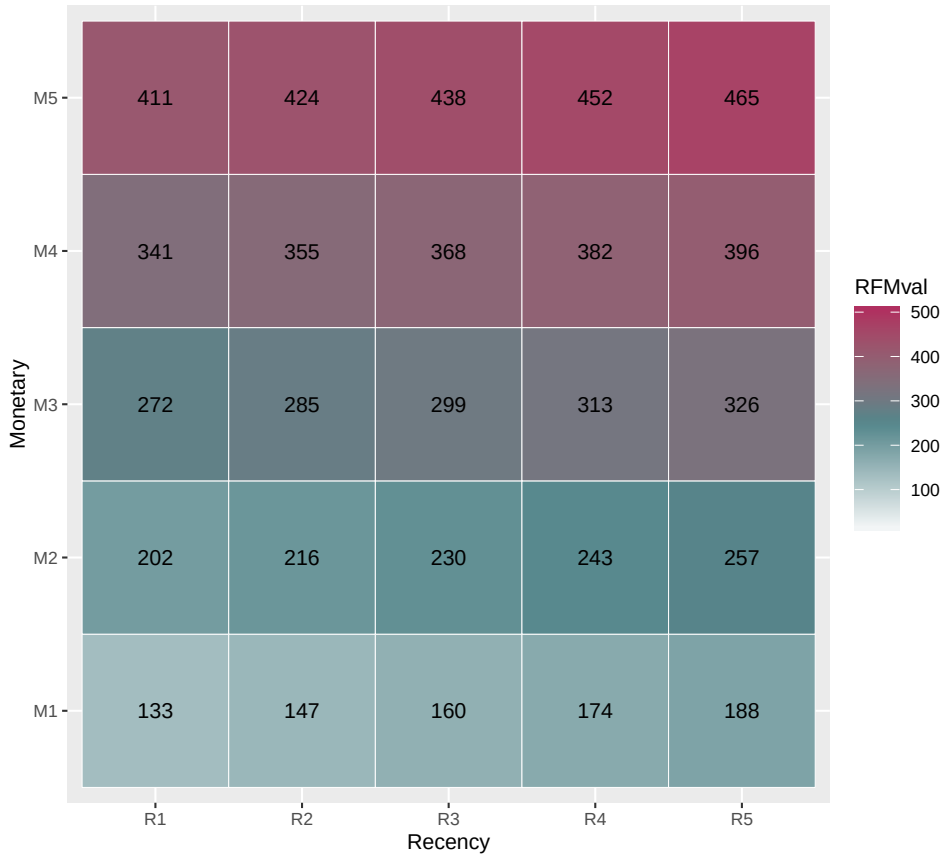
4.2.8.2 Shlukování na základě původních atributů

V další variantě nasazení shlukování bude zkoumáno, jaký vliv bude mít, pokud se vynechá krok samotné RFM analýzy a na vstupu pro algoritmus K-means bude nediskretizovaný vstup.

Grafy 4.28 a 4.29 ukazují porovnání kvality shlukování pro obě řešení – vstup s percentily (RFM_P) a vstup s originálními – „raw“ hodnotami (RFM_R).

Ani jedna z metod nedokázala dosáhnout stejné kvality segmentace pro $K = 3$, jako shlukování typu RFM_E z předchozí varianty. Pro vzrůstající K bylo dosaženo obdobné kvality jako RFM_E. Tabulky 4.30 a 4.31 popisují shluky pro zvolená K.

Shlukování na základě percentilů dokázalo dobře identifikovat segment cenných zákazníků s většími a opakovanými nákupy. Podobně jako shlukování RFM_E z předchozí varianty zahrnuje segment klientů, kteří jsou vhodní



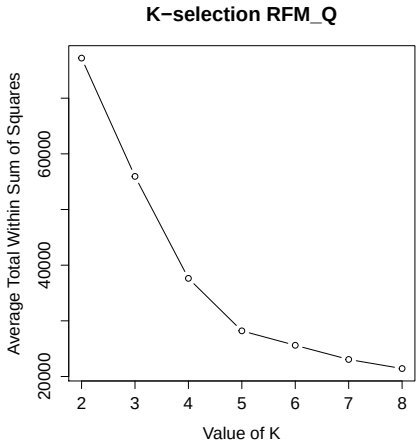
Obrázek 4.23: Huskycz – průměrná hodnota $score = V_{RFM} \cdot 100$ dle RM segmentů.

k aktivaci, respektive segment drobnějších nákupů. Zastoupen je také segment méně perspektvních „mrtvých“ klientů.

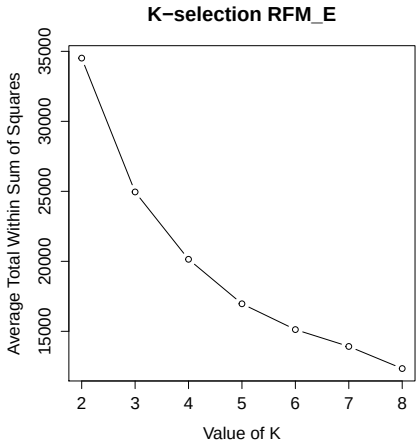
Řešení se vstupem ve formě původních hodnot má z hlediska kvality shlukování stejně dobré, jako RFM_E z předchozí varianty, jenomže interpretovatelnost samotných segmentů je podstatně složitější.

To komplikuje přímé použití pro plánování obchodní strategie, nicméně se může ukázat užitečné po tom, co bude na jeho základě provedeno nějaké sondování zákaznické báze a vyhodnocena úspěšnost respondibility pro výsledné shluky.

Toto platí ostatně pro všechny segmenty vzniklé pomocí různých metod, ať už z těch uvedených v této studii, nebo dalších jiných. Nicméně segmentace s využitím metody K-means na základě ohodnocení z předchozí RFM analýzy umožňuje kvalitní shlukování při nízkém počtu shluků a především jednoduchou interpretovatelnost vzniklých segmentů na základě vhodně zvoleného expertního rozdělení.



Obrázek 4.24: Huskycz - RFM_Q Graf závislosti TWSS na K



Obrázek 4.25: Huskycz - RFM_E Graf závislosti TWSS na K

K = 3		Souřadnice centroidů			Popis shluku
Shluk	Velikost	R	F	M	
1	6806	1.00	1.03	1.86	*Mrtví klienti*
2	8662	2.68	1.05	1.94	*Mainstream, prvonákupčí, klienti k aktivaci*
3	4943	2.16	1.48	4.43	*Větší a opakované nákupy*

Obrázek 4.26: Tabulka segmentů pro $K = 3$, RFM analýza s expertním rozdělením

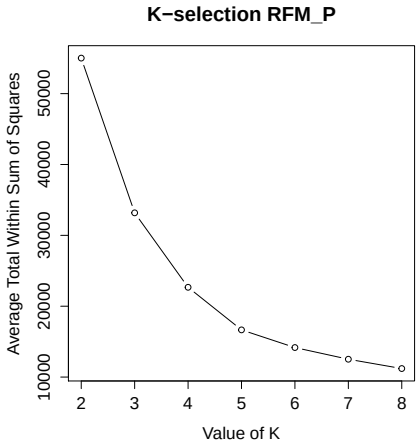
K = 4		Souřadnice centroidů			Popis shluku
Shluk	Velikost	R	F	M	
1	1540	3.22	2.25	4.72	*Hodnotní zákazníci, větší a opakované nákupy*
2	5182	1.59	1.11	3.86	*Mainstream, prvonákupčí*
3	4456	3.31	1.08	2.01	*Perspektivní klenti k aktivaci, drobné nákupy*
4	9233	1.38	1.02	1.65	*Mrtví klienti*

Obrázek 4.27: Tabulka segmentů pro $K = 4$, RFM analýza s expertním rozdělením

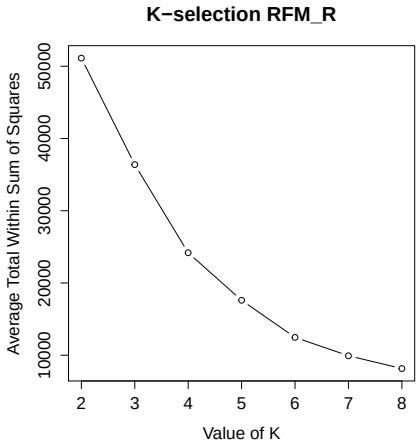
4.2.9 Interpretace segmentů pomocí analýzy nákupních košů

Další možností, jak obohatit interpretaci výše provedené segmentace je vyšetřit jednotlivé segmenty nejen pouze z hlediska zákazníků, čímž se zabývaly analýzy výše, ale také z hlediska produktového. Pro tento účel je využívána tzv. analýza nákupních košů, která zkoumá zastoupení jednotlivých typů produktů v nákupech a také vztahy a spojitosti jednotlivých produktů během nákupů.

Samotné analýze nákupních košů obchodu nejprve předcházelo předzpra-



Obrázek 4.28: Huskycz - RFM_P Graf závislosti TWSS na K



Obrázek 4.29: Huskycz - RFM_R Graf závislosti TWSS na K

RFM_P K = 5		Souřadnice centroidů			Popis shluku
Shluk	Velikost	R*	F*	M*	
1	4203	3.96	1.00	1.99	*Nedávno nakupivší, drobné nákupy*
2	4082	1.91	1.00	3.79	*Méně perspektivní klienti s většími nákupy*
3	5284	2.00	1.00	1.85	*Méně perspektivní klienti s drobnými nákupy*
4	2133	3.69	4.67	4.15	*Cenní klienti s většími a opakovanými nákupy*
5	4709	3.92	1.00	3.86	*Nedávno nakupivší, větší nákupy*

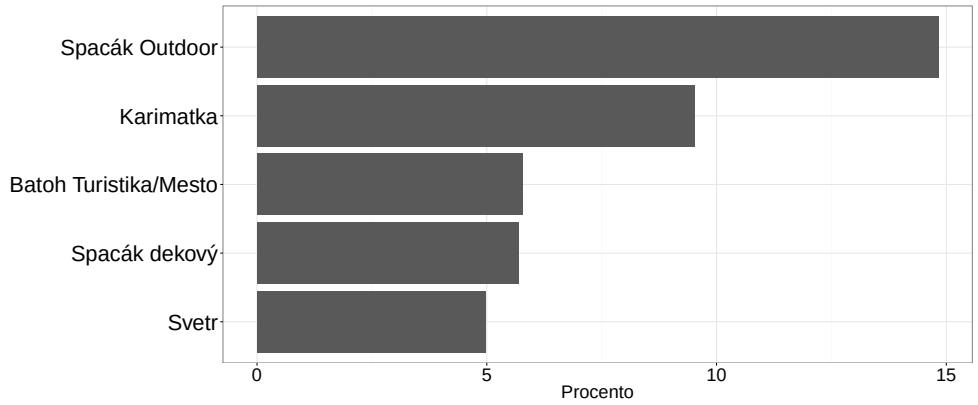
Obrázek 4.30: Tabulka segmentů pro $K = 5$, shlukování s percentilovým vstupem. *Souřadnice v prostoru 5x5x5, který vychází z původního prostoru RFM

RFM_R K = 4		Souřadnice centroidů			Popis shluku
Shluk	Velikost	R*	F*	M*	
1	6442	2.19	1.00	1.85	*Méně perspektivní klienti*
2	2133	3.69	4.67	4.15	*Cenní klienti s většími a opakovanými nákupy*
3	5257	2.27	1.00	3.95	*Dražší zboží, klienti k aktivaci*
4	6579	4.17	1.00	2.91	*Nedávno nakupivší*

Obrázek 4.31: Tabulka segmentů pro $K = 4$, shlukování se vstupem původních hodnot. *Souřadnice v prostoru 5x5x5, který vychází z původního prostoru RFM

cování databáze nabízených produktů, především spojování jednotlivých produktových položek do produktových skupin. Extrahovány přitom byly informace o produktové kategorii pro účely nakupování v e-shopu pomocí parsování tabulky *items*. Dále byly také sdruženy produkty lišící se v přízvisku „pánské/dámské“ apod.

Takto vzniklo celkem 118 produktových skupin, na kterých byla posléze



Obrázek 4.32: Top 5 nejprodávanějších produktů procentuálně v koších.

provedena samotná analýza.

V této studii byly v návaznosti na předchozí segmentaci z kapitoly 4.2.7.1, konkrétně shlukování na základě RFM s expertním rozdělením do 4 segmentů (RFM_E, $K = 4$) provedeny tyto dva druhy analýzy nákupního koše:

1. Určení nejprodávanějších produktů celého obchodu a jednotlivých segmentů.

2. Analýza asociačních pravidel mezi produkty v nákupních koších celého obchodu a dle jednotlivých segmentů.

4.2.9.1 Určení nejprodávanějších produktů

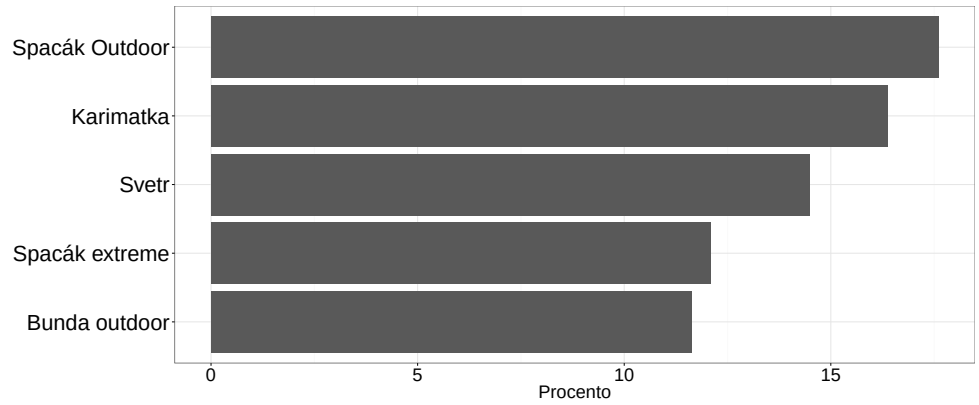
Nejprve byly zjištěny nejprodávanější produkty v rámci prodejů celé zákaznické báze. Graf 4.32 ukazuje procentuální zastoupení pěti nejprodávanějších produktů v nákupních koších. První místo obsadil produkt „Spacák Outdoor“, který se vyskytuje v bezmála 15 % všech provedených nákupech.

Dále byly zjištěny nejprodávanější produkty pro jednotlivé segmenty dle výsledků shlukování (RFM_E, $K = 4$) z kapitoly 4.3.6.1., viz tabulka 4.27.

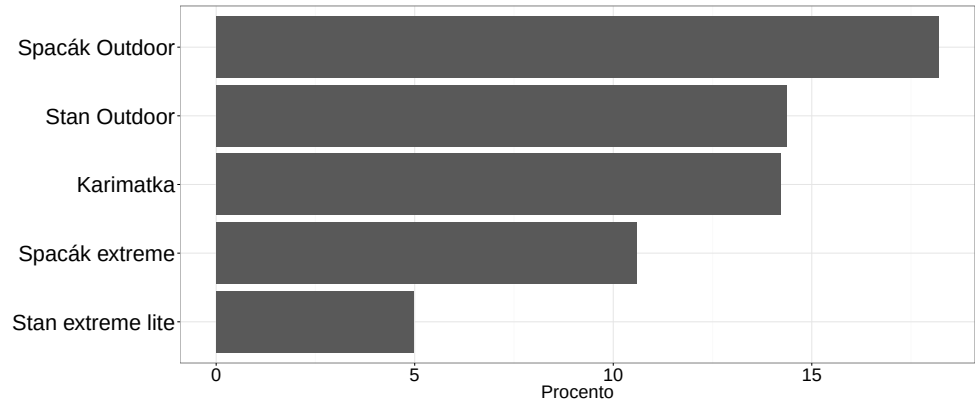
Přestože se nejprodávanější produkty v rámci určených segmentů příliš neliší od nejprodávanějších produktů v rámci celé báze ani mezi sebou, je možné pozorovat některé výjimky, které přidávají do segmentace další informaci. Viz grafy 4.33, 4.34, 4.35 a 4.36.

V rámci *segmentu 2* („prvonákupci, mainstream“) lze pozorovat, že mezi nejprodávanější zboží patří výhradně campingové a stanovací vybavení. Vzhledem k charakteru obchodu a jeho specializaci na tento typ produktů se lze domnívat, že zákazníci provádějící svůj úplně první nákup navštíví obchod právě za účelem nákupu tohoto specializovaného zboží.

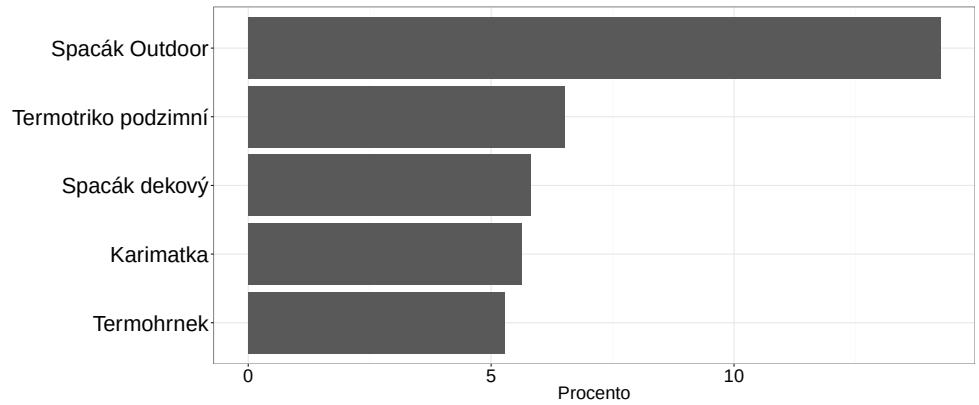
Typický zákazník obchodu – prvonákupce (pozn.: zákazníci s jediným nákupem tvoří cca 90 % celé zákaznické báze) tedy nakupují specializované,



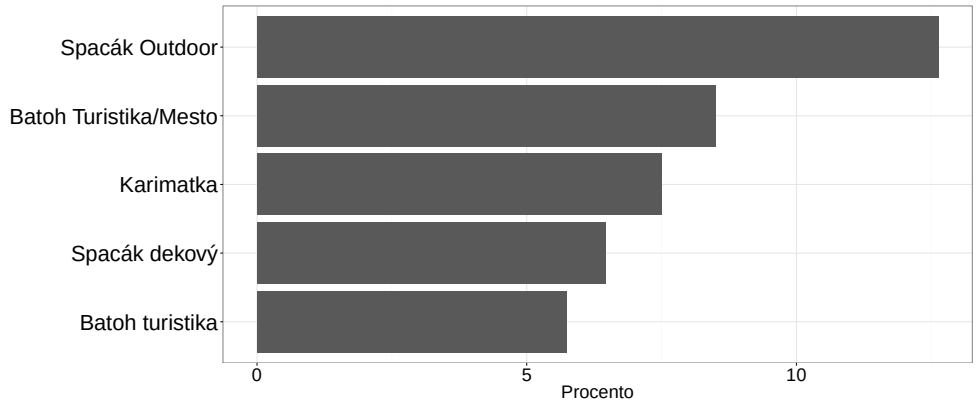
Obrázek 4.33: Segment 1 – Top 5 nejprodávanějších produktů procentuálně v koších.



Obrázek 4.34: Segment 2 – Top 5 nejprodávanějších produktů procentuálně v koších.



Obrázek 4.35: Segment 3 – Top 5 nejprodávanějších produktů procentuálně v koších.



Obrázek 4.36: Segment 4 – Top 5 nejprodávanějších produktů procentuálně v koších.

spíše dražší zboží. V rámci obchodních kampaní je na tento segment vhodné cílit především za účelem aktivace pomocí *cross-sellingu/up-sellingu* dalších typů zboží a nabídnout těmto zákazníkům i jiné typy sortimentu, případně příslušenství ke specializovanému zboží.

Rozvržení nejprodávanějších produktů ve třetím segmentu („drobné nákupy, perspektivní klienti“) potvrzuje příslušnost klientů, které provádějí drobnější nákupy levnějšího zboží – „termotriko“, „termohrnek“.

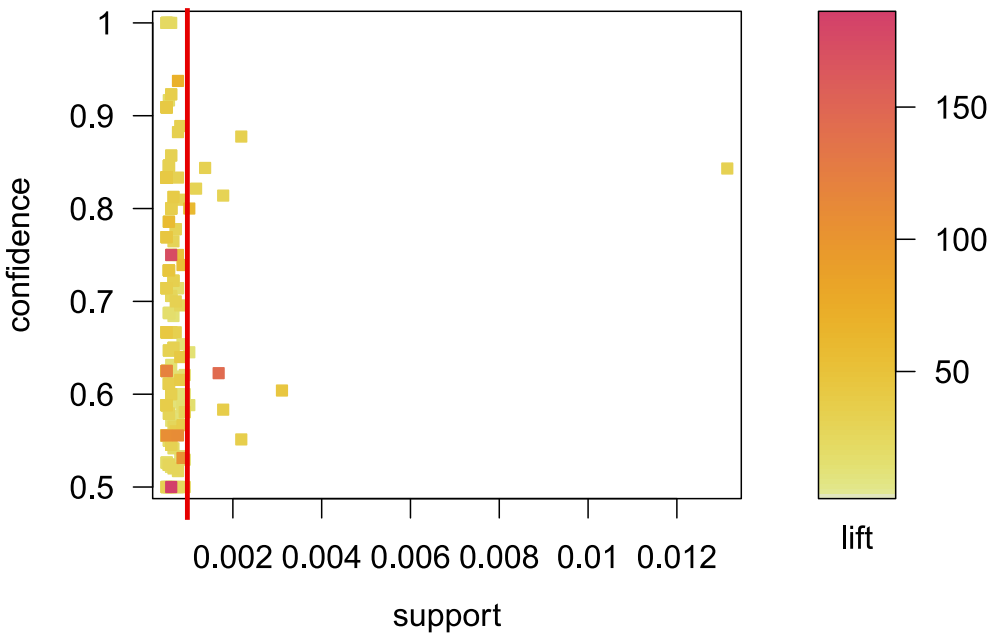
Lze se domnívat, že zákazníci v rámci tohoto segmentu se budou ochotni zajímat se o rozličné tipy zboží za nízkou cenu. Dle vysoké průměrné hodnoty *recency* se jedná o klienty, kteří nakoupili v průběhu posledních 180 dní a jsou tedy také vhodní kandidáti pro aktivační kampaň.

4.2.9.2 Analýza pomocí asociačních pravidel

Vzhledem k tomu, že jen pouze 10 % zákazníků provedlo opakovaný nákup, nebylo nalezeno příliš velké množství pravidel, které by splňovalo požadavky minimální hodnoty *support* ve smyslu statistické významnosti. Jinými slovy, nalezená pravidla mohou být s vysokou pravděpodobností pouze náhodné výskyty, jejich vzorek v datech je příliš malý. Hranice pro hodnotu *support* byla nastavena 0.001. Graf 4.37 ukazuje parametry nalezených asociačních pravidel a znázorňuje hranici pro výběr relevantních z nich.

Bublinový graf 4.38 popisuje asociace jednotlivých produktů na levé, resp. pravé straně nalezených pravidel. Na ose X jsou vyneseny produkty na pravé a na ose Y produkty na levé straně asoc. pravidel. Velikost každé bubliny představuje hodnotu *support* odpovídajícího pravidla a barevná škála hodnotu *lift*. Například nejvyšší hodnotu *lift* mezi nalezenými pravidly má asoc. pravidlo {„*Plyn k vařiči*“} $\longrightarrow$ {„*Vařič*“}.

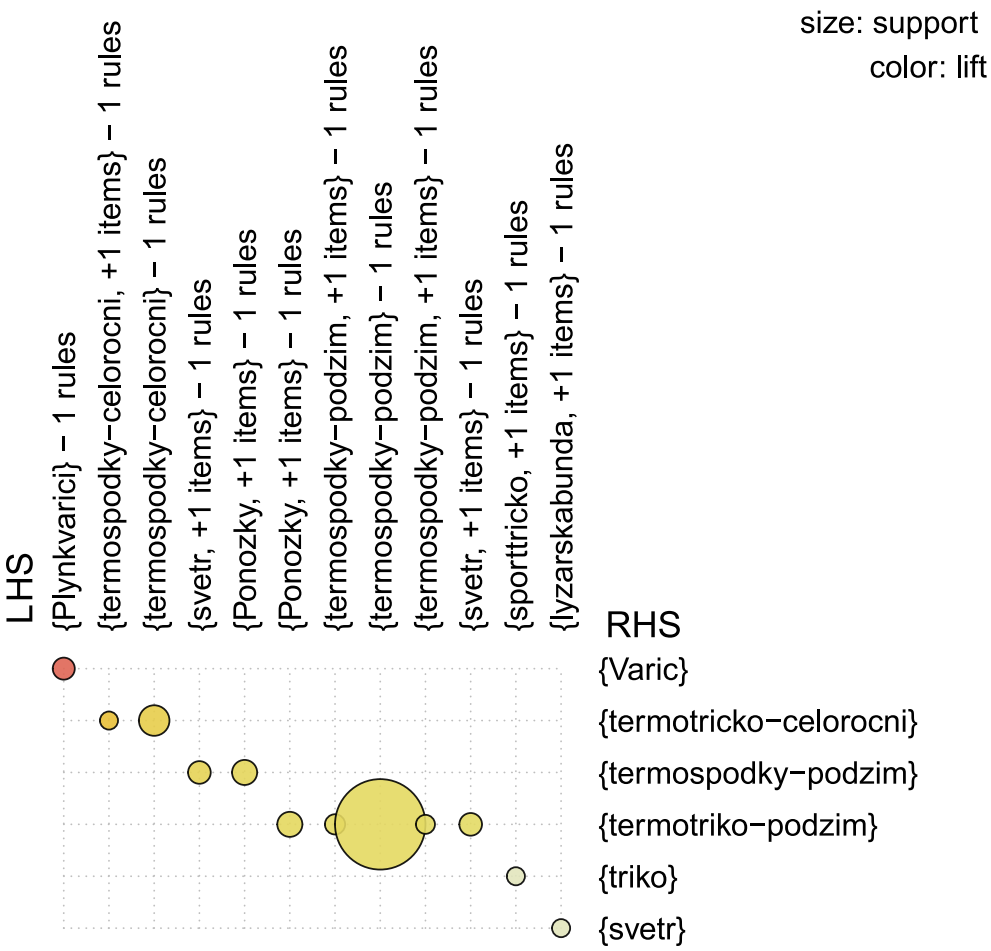
Dále byla provedena analýza asoc. pravidel v rámci jednotlivých segmentů.



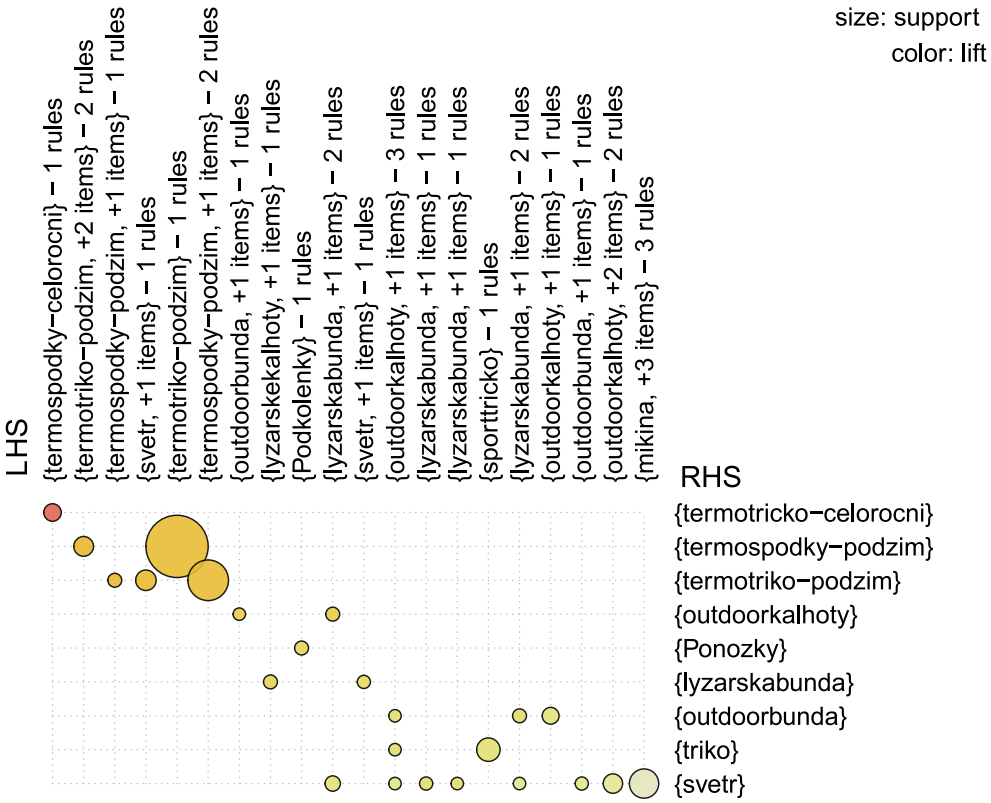
Obrázek 4.37: Graf parametrů nalezených asoc. pravidel pro celou datovou bázi. Červená čára znázorňuje hranici pro statistickou významnost na hodnotě $support = 0,001$.

Došlo k situaci, že nebylo nalezeno dostatečné množství pravidel, které by splňovalo požadavky minimální hodnoty *support* ve smyslu statistické významnosti.

To se však netýká pravidel v rámci segmentu 1 („větší a opakované nákupy“), který tvoří právě zákazníci, již nakupovali několikrát nebo několik produktů najednou. Nalezená pravidla mají dostatečnou statistickou významnost a vysoké hodnoty *confidence* ($> 0,5$) a *lift* (> 2) značí dobrou použitelnost těchto pravidel například pro tvorbu cílených nabídek zákazníkům tohoto segmentu nebo doporučování vhodných produktů při nakupování. Viz bublinový graf 4.39 a graf 4.40, který rozvíjí vizualizaci vzájemných asociací mezi produkty. Na rozdíl od bublinového grafu jsou zobrazeny všechny vztahy mezi produkty (nejen po dvojicích). Velikost a barva uzlů, spojující produkty, odpovídá hodnotám *support*, resp. *lift*.



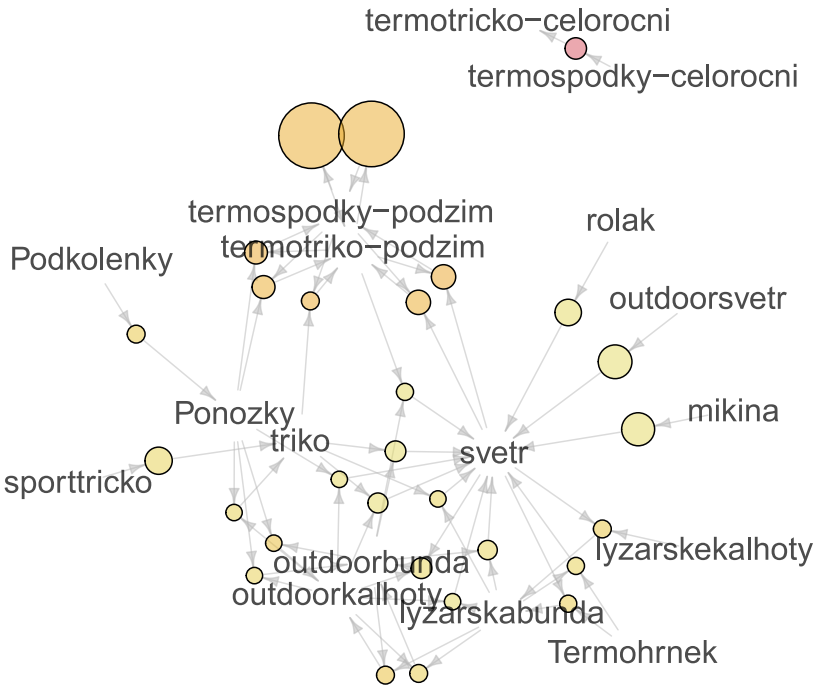
Obrázek 4.38: Graf vztahů mezi asociovanými produkty. Na ose X jsou vyneseny produkty na pravé a na ose Y produkty na levé straně asoc. pravidel. Velikost bubliny odpovídá hodnotě *support* nalezeného pravidla a barevná škála (bílá-červená) hodnotě *lift*.



Obrázek 4.39: Segment 1 – Graf vztahů mezi asociovanými produkty. Na ose X jsou vyneseny produkty na pravé a na ose Y produkty na levé straně asoc. pravidel. Velikost bubliny odpovídá hodnotě *support* nalezeného pravidla a barevná škála (bílá-červená) hodnotě *lift*.

Zde jsou vybraná asociační pravidla pro segment 1:

Antecedent	**Konsekvent**	***Support***	***Confidence***	***Lift***
{Termospodky podzimní}	{Termotriko podzimní}	0,05	0,87	8,94
{Lyžařská bunda; Triko}	{Svetr}	0,01	0,70	4,80
{Outdoor bunda; Ponožky}	{Outdoor kalhoty}	0,01	0,59	7,13



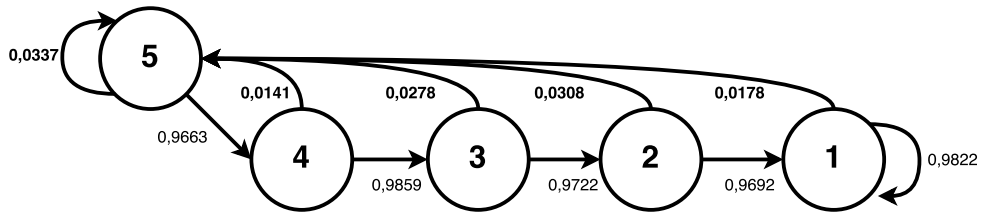
Obrázek 4.40: Segment 1 – Orientovaný graf vztahů mezi asociovanými produkty. Barevné uzly odpovídají asociacím mezi produkty. Velikost uzlů odpovídá hodnotě *support* nalezeného pravidla a barevná škála (bílá-červená) hodnotě *lift*.

4.2.10 Odhad očekávané hodnoty klienta - CLV

Pro CLV analýzu zákazníků obchodu z této studie byly využity výsledky RFM analýzy s expertním rozdělením. Pravděpodobnosti přechodů mezi segmenty byly stanoveny průměrným poměrem mezi klienty, kteří provedli opakovaný nákup v odpovídajícím období, a klienty, kteří nenakoupili a posunuli v následujícím období do segmentu s nižší *recency*.

V určitém měsíci byli například zjištěni klienty, kteří naposledy nakoupili před 30 a 60 dny a mezi nimi zjištěn poměr těch, kteří v tomto období provedli nákup (a zároveň ne dříve) a těmi, kteří neprovedli. To odpovídá pravděpodobnosti přechodu klientů s hodnotou *recency* = 4 (*stav 4*) do *stavu 5*, resp. *stavu 3*. Toto bylo zjištěno pro všechny stavy a zprůměrováno přes všechny měsíce sledovaného období.

Obrázek 4.41 zobrazuje sestavený markovský řetězec včetně zjištěný pravděpodobností přechodů mezi stavy. Stavy odpovídají příslušnosti zákazníků k segmentům dle *recency*.



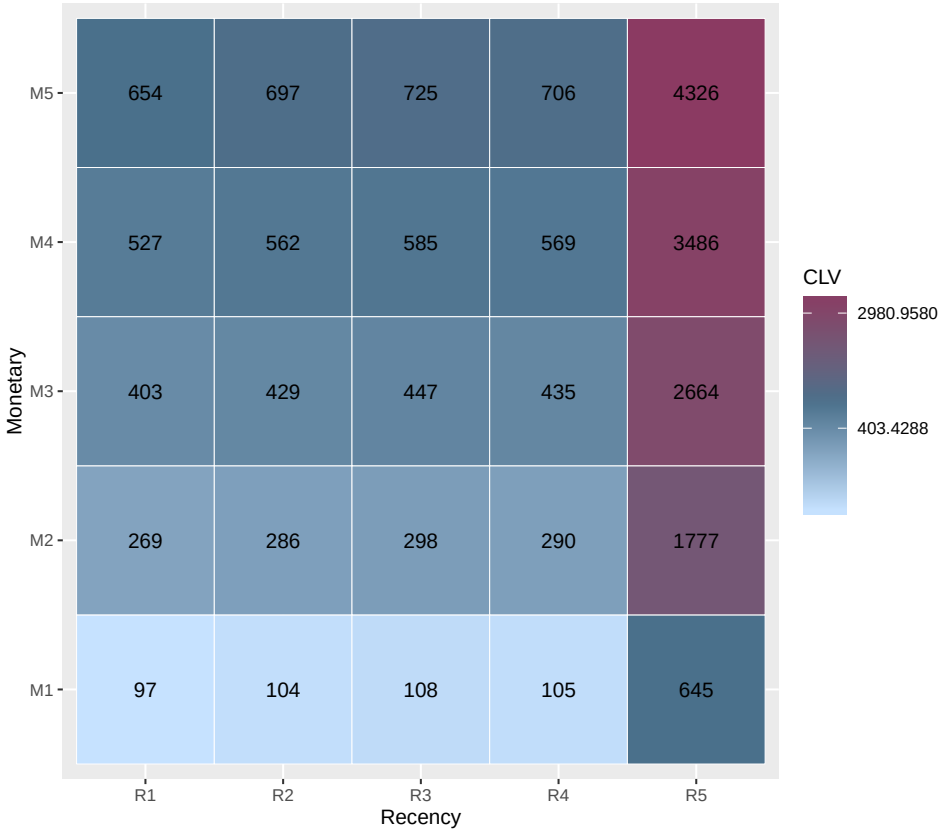
Obrázek 4.41: Markovský řetězec s ohodnocením hran dle odhadnuté pravděpodobnosti přechodů. Stavy odpovídají příslušnosti zákazníků k segmentům dle *recency*.

Profit zákazníka byl odvozen z hodnot průměrného nákupu po jednotlivých segmentech dle *monetary*. Je nutno zdůraznit, že hodnota nákupu se rovná tržbě, ne zisku obchodu z prodeje daného produktu. Tato data nebyla pro studii k dispozici a je nutné tedy k výsledkům přistupovat s odpovídající interpretací. Ostatní náklady na údržbu klienta byly s ohledem na charakter internetového obchodu zanedbány. Hodnota diskontní sazby d byla zvolena $0,1$.

Následující graf 4.42 ukazuje odhad hodnot CLV pro jednotlivé RM segmenty z předcházející RFM analýzy. V pravém sloupci jsou zastoupeni klienti, kteří právě provedli nákup. Směrem doleva je potom viditelný pokles tak, jak klesá pravděpodobnost opakovaného nákupu.

4.2.11 Závěry analýzy

Případová studie se zabývala analýzou transakčních dat obchodu za období mezi 13. 5. 2014 a 9. 3. 2016. Toto období zahrnuje bezmála 66 tisíc transakcí



Obrázek 4.42: CLV pro RM segmenty. Barevná škála pozadí zobrazuje odhad CLV pro jednotlivé segmenty.

s nenulovou hodnotou nákupu, které byly provedeny v rámci více než 24 tisíc nákupních košů. Bylo rozpoznáno přes 20 tisíc nakupujících klientů.

Statistická analýza mezi nimi odhalila zhruba 10 % klientů, kteří nakupovali opakovaně a byla zjištěna výrazná sezónnost prodejů, především v souvislosti s předvánočními obdobími.

Následně byla provedena RFM analýza a klientská báze rozdělena do 25 RM segmentů dle variant RFM kvantilovou metodou a s expertním rozdělením. Výstupy z analýzy sloužily také jako vstup pro další shlukování, kdy bylo v několika variantách navrženo rozdělení zákaznické báze do 4 nebo 5 segmentů. Segmentace na základě RFM analýzy a shlukování byla doplněna interpretací s využitím znalosti expertního rozdělení.

Vhodně zvolené expertní rozdělení přináší do procesu jistou apriorní informaci, která se ukazuje jako podstatná pro kvalitu shlukování a vyvažuje ztrátu informace, ke které nutně dochází při předcházející RFM analýze. Výsledkem je v této studii výhodný trade-off mezi přesností modelu a jeho použitelnosti ve smyslu interpretovatelnosti z pohledu obchodu.

Další interpretace byla doplněna za pomoci provedené analýzy nákupních košů, v rámci níž byly odhaleny i některé vztahy mezi nakupováním určitých produktů.

Nakonec byl stanoven odhad očekávané dlouhodobé hodnoty zákazníků CLV v rámci jednotlivých RM segmentů. Ten, společně s podobou rozvržení zákaznické báze po RM segmentech a odhadem významnosti klientů dle *WRFM*, tvoří indikátor úspěšnosti obchodu a v kombinaci s vhodnou interpretací slouží jako zdroj pro doporučení zavedení obchodních strategií pro vytěžení stávající zákaznické báze.

Vizualizace výsledků dílčích kroků analýzy slouží jako prostředek pro reportování majitelům dat.

Závěr

Na začátku této práce byly nastudovány různé techniky dolování dat vhodné pro analýzy nad interakčními daty. Především byla prozkoumána použitelnost *RFM analýzy* v kombinaci s dalšími metodami strojového učení, zejména pro úlohu segmentace zákaznické báze v souvislosti s úlohou shlukování a využití asociačních pravidel. Dále byly popsány modely pro odhad *dlouhodobé hodnoty klienta – CLV*.

Na základě prozkoumaných metod byl navržen analytický framework pro účely reportování majitelům transakčních a prodejních dat, který respektuje identifikovaná specifika e-shopů. Součástí navrhovaného frameworku je určení požadavků pro jeho nasazení, popis jednotlivých kroků analýz a stanovení indikátorů úspěšnosti e-shopů na základě popsaných analýz.

Navrhovaný framework je popsán v základní podobě, která zahrnuje inspekci vstupních dat a vizualizaci vstupních dat v původní podobě a následně využívá aplikaci *RFM analýzy* pro vytvoření zákaznických segmentů, jejich vizualizace a interpretaci. Dále jsou popsány dodatečné analýzy, které jsou vhodné k nasazení pro e-shopy s větším objemem prodejů a větší zákaznickou bází.

Nakonec jsou uvedeny dvě případové studie nad daty z reálných e-shopů, které aplikují popstupy navrhovaného frameworku. Součástí studií jsou mimo popisu konkrétního postupu a použitých nástrojů také interpretace výsledků a závěry včetně návrhů obchodních doporučení.

Literatura

[1] Blattberg, R. C.; Kim, B. D.; Neslin, S. A.: *Database Marketing: Analyzing and Managing Customers*. Springer, 2008.

[2] Hughes, A. M.: *Strategic Database Marketing*. Probus Publishing, 1994.

[3] Bult, J. R.; Wansbeek, T.: Optimal selection for direct mail. *Marketing Science*, , č. 14, 1995: s. 378–395.

[4] McCarty, J. A.; Hastak, M.: Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, , č. 6, 2007: s. 656–662.

[5] Cheng, C.-H.; Chen, Y.-S.: Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, , č. 3, 2009: s. 4176–4184.

[6] Hu, Y.-H.; Yeh, T.-W.: Discovering valuable frequent patterns based on RFM analysis without customer identification information. *Knowledge-Based Systems*, , č. 61, 2014.

[7] Miglautsch, J.: Thoughts on RFM scoring. *Journal of Database Marketing*, , č. 8, 2000: s. 67–72.

[8] Liu, D.-R.; Shih, Y.-Y.: Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information and Management*, , č. 42, 2005: s. 387–400.

[9] Saaty, T. L.: Decision making with the analytic hierarchy process. *Int. J. Services Sciences*, , č. 1, 2008: s. 83–98.

[10] D. Ssebuggwawo, H. P., S. Hoppenbrouwers: Evaluating Modeling Sessions Using the Analytic Hierarchy Process. In *Persson, A., Stirna, J. (eds.) PoEM 2009*, Springer, 2009, s. 69–83.

[11] Kim, D.; Lee, J.; Ahn, S.; aj.: RFM analysis for detecting future core technology. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, RACS, ACM Press, 2012, s. 55–59.

[12] Ricci, F.: *Recommender systems handbook*. Springer, 2011.

[13] Lloyd, S.: Least Squares Quantization in PCM. *IEEE Trans. Information Theory*, , č. 28, 1982: s. 129–137.

[14] Ostrovsky, R.; Rabani, Y.; Schulman, L.; aj.: The Effectiveness of Lloyd-Type Methods for the k-Means Problem. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, IEEE, 2006, s. 165–174.

[15] Khajvand, M.; Zolfaghar, K.; Ashoori, S.; aj.: Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. In *World Conference on Information Technology*, WCIT, Procedia Computer Science, 2011, s. 57–63.

[16] Birant, D.: Data Mining Using RFM Analysis. In *Knowledge-oriented applications in data mining*, Rijeka: In-Tech, 2011.

[17] Agrawal, R.; Imielinski, T.; Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ACM Press, 1993, s. 207–216.

[18] Agrawal, R.; Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In *VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers, 1994, s. 487–499.

[19] Hahsler, M.; Grun, B.; Hornik, K.: A computational environment for mining association rules and frequent item sets. *Journal of statistical software*, 2005.

[20] J. Radová, J. M., R. Dvořák: *Finanční matematika pro kazdého*. Grada, 2013.

[21] Pfeifer, P. E.; Carraway, R. L.: Modeling customer relationships as Markov chains. *Journal of Interactive Marketing*, , č. 14, 2000: s. 43–55.

[22] Hahsler, M.; Chelluboina, S.: Visualizing association rules: Introduction to the R-extension package arulesViz. 2010.

Příloha **A**

Seznam použitých zkratek

CART Classification and regression tree

CLV Customer lifetime value

LTV (Customer) lifetime value

RFM Recency, frequency a monetary

RFM_Q RFM analýza kvantilovou metodou

RFM_E RFM analýza s expertním rozdělením

RFM_P Shlukování dle RFM - vstup s percentily

RFM_R Shlukování dle RFM - vstup s originálními (raw) hodnotami

TWSS Total within sum of squares

WRFM Weighted recency, frequency a monetary

Příloha **B**

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src`
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF