Úvod

Cílem našeho projektu je kategorizovat data *data.txt* podle nakupovacích návyků (proměnné *q34#11–q34#29*). Proměnné *q34#11–q34#29* jsou kategoriální nabývající hodnot 0 nebo 1, kde hodnota 0 nám říká, že daná osoba nenakupuje v daném řetězci *q34#??*. Dále je naším úkolem výslednou kategorizaci interpretovat a popsat, jak se mezi sebou dané skupiny liší.

Na začátek bude třeba data vyčistit, ať se nám s daty snadno pracuje a nevyskytují se žádné problémy při analýze a vytváření modelů. Dále budeme muset provést exploratorní analýzu a zjistit, zda se v datech vyskytují případná odlehlá pozorování či zda některé proměnné nekorelují. Ve třetí části zkusíme vytvořit různé modely a vybereme ten, který naše data popisuje nejlépe. Nakonec provedeme analýzu našeho výsledného modelu a pokusíme se jej interpretovat.

Kapitola 1

Vypracování

1.1 Čištění dat

K dispozici jsme měli data o 38 proměnných a 2889 pozorování. Všechny proměnné byly kategoriální. Data byla velmi kvalitní, nemuseli jsme ani odstraňovat mnoho pozorování.

Data vznikla na základě výzkumu populace ČR kolem roku 2000. Jedná se o reprezentativní vzorek občanů ČR té doby. Data byla sbírána za účelem velkého průzkumu chování obyvatel ČR. Takže se nejedná jen o chovatele zvířat.

Prvně jsme si potvrdili, že proměnná *id* je vskutku unikátní pro všechna pozorování a že se rovná počtu všech pozorování. Poté jsme postupně určili, jakých hodnot nabývá každá proměnná, tím jsme odhalili jednak, že každá proměnná nabývá jen předem určených hodnot a dále, jestli proměnná obsahuje nějaká prázdná pozorování. Zjistili jsme, že proměnné nabývají pouze správných hodnot. Dále jsme zjistili, že chybějící hodnoty byly v zásadě pouze v proměnných *pinc* a *hint*. Nakonec jsme objevili 4 pozorování id: 913, 2028, 2377, 2725, u kterých chybělo více proměnných. Tato pozorování nešla nijak doplnit, a tak jsme je byli nuceni odebrat z našich dat.

Proměnná *hint* popisuje, jak často se daná domácnost připojuje k internetu. Chybějící hodnoty této kategoriální proměnné jsme se rozhodli doplnit modem. Modus nám vyšel roven jedné. Hodnotu jsme tedy doplnili namísto chybějících hodnot této proměnné.

Proměnná *pinc* obsahuje platové kategorie pro jedince (nabývá hodnot 11–31). Rozhodli jsme se tedy chybějící pozorování doplnit pomocí proměnné *hinc*, která popisuje příjem domácnosti (a u které žádná pozorování nechyběla). Prvně jsme si ověřili, že příjem domácnosti není nižší než příjem jedince. Dále jsme si vytvořili pomocnou proměnnou s názvem *cenova_kategorie*, která vznikla z proměnné *pinc* následujícím způsobem: pokud pro dané pozorování nabývala proměnná *pinc* hodnoty 11, potom byla proměnné *cenova_kategorie* přiřazena hodnota 80000. Pokud pro dané pozorování nabývala proměnná *pinc* hodnoty 31, potom byla proměnné *cenova_kategorie* přiřazena hodnota 3000. Jinak byla proměnné *cenova_kategorie* přiřazena průměrná hodnota intervalu, který je re-

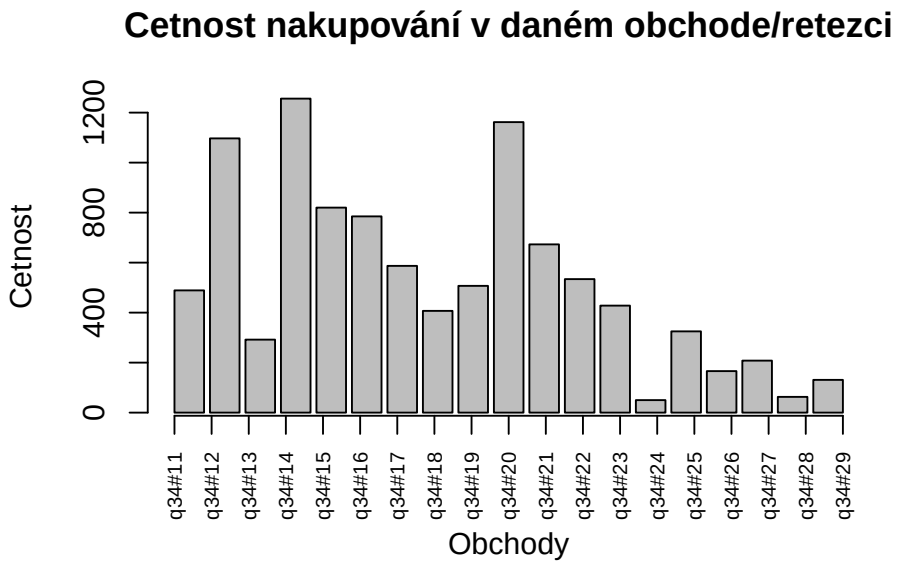
prezentován kategoriální proměnnou *pinc*. Tj. například pro $pinc = 19$ byla proměnné *cenova_kategorie* přiřazena hodnota 11000.

Pomocnou proměnnou *cenova_kategorie2* jsme vytvořili obdobným způsobem z kategoriální proměnné *hinc*.

Vznesli jsme hypotézu, že *cenova_kategorie2* je dvojnásobkem proměnné *cenova_kategorie*. Hypotézu jsme si ověřili na našich datech. Vydělili jsme tedy u každého pozorování proměnnou *cenova_kategorie* pomocí proměnné *cenova_kategorie2* a vypočetli průměrnou hodnotu tohoto dělence. Výsledná hodnota nám vyšla jako 0.498965, tedy se nám hypotéza potvrzuje. Rozhodli jsme se tedy nahradit chybějící hodnoty u proměnné *cenova_kategorie* pomocí 0.5-násobku proměnné *cenova_kategorie2*. Nakonec pouze stačilo doplnit chybějící hodnoty u proměnné *pinc* kategorizováním proměnné *cenova_kategorie*. Výsledný počet pozorování po vyčištění dat je tedy 2885.

1.2 EDA

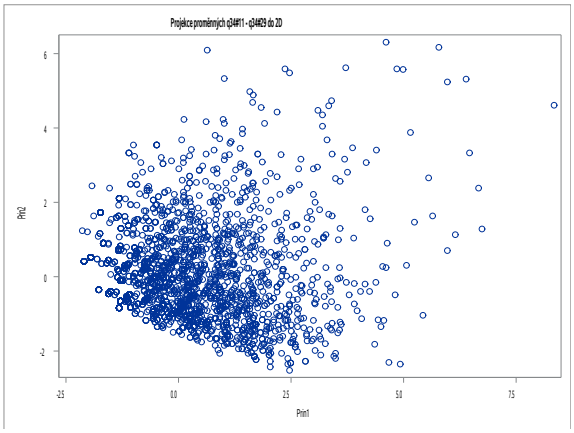
Jelikož budeme vytvářet model pouze pro proměnné *q34#11* – *q34#29*, bude nám stačit provést exploratorní analýzu pouze pro tyto proměnné. Proměnné jsou kategoriální, bez chybějících hodnot. Nabývají hodnot 0 a 1. Četnost nakupování v jednotlivých obchodech si znázorníme pomocí sloupcového grafu 1.1.



Obrázek 1.1: Barchart

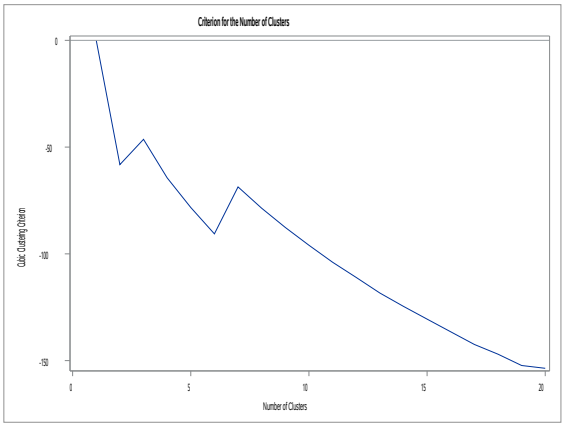
1.3 Shluková analýza

Ze začátku jsme chtěli vytvořit shluky pomocí Wardovy metody, avšak číselná charakteristika CCC (cubic clustering criterion) nám říkala, že nejnižší vhodný počet shluků je buďto 2 anebo 17. Obě tyto možnosti jsou velmi nevhodné, jelikož dva shluky by se dosti prolínaly, protože naše data vytvářejí jeden velký shluk a nejsou žádným zřetelným způsobem rozdělena na dva, vizte obrázek 1.2. A pro shluků 17 by se nám hledala těžko interpretace.



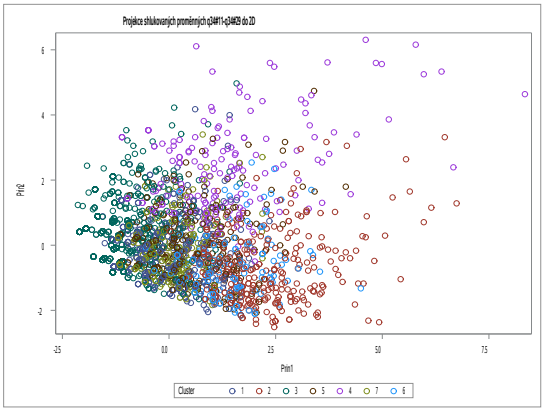
Obrázek 1.2: Shluk

Dále jsme tedy zkusili použít funkci *proc fastclus* v softwaru SAS. Funkce *proc fastclus* využívá metody „k medoids“. Pro tuto metodu nám hodnota CCC vracela počet shluků 7, vizte obrázek 1.3. Stejný výsledek vycházel i pomocí námi udělaného makra. Sedm shluků je mnohem snáze interpretovatelný výsledek než-li shluků 17, a tedy volíme metodu „k medoids“ k vytvoření shluků.



Obrázek 1.3: CCC

Provedeme shlukování proměnných *q34#11–q34#29* pomocí funkce *proc fastclus* v SASu a metodou hlavních komponent si tento výsledek promítněme do 2D. Znázorněme si tedy promítnutí pouze samotných shlukovaných proměnných *q34#11–q34#29*; obrázek 1.4.



Obrázek 1.4: Proměnné *q34#11–q34#29* ve 2D

Znázorněme si také absolutní a relativní četnosti zastoupení jednotlivých shluků v našich datech na obrázku 1.5. Vidíme, že žádný shluk striktně nedominuje v počtu pozorování.

Velikost shluků

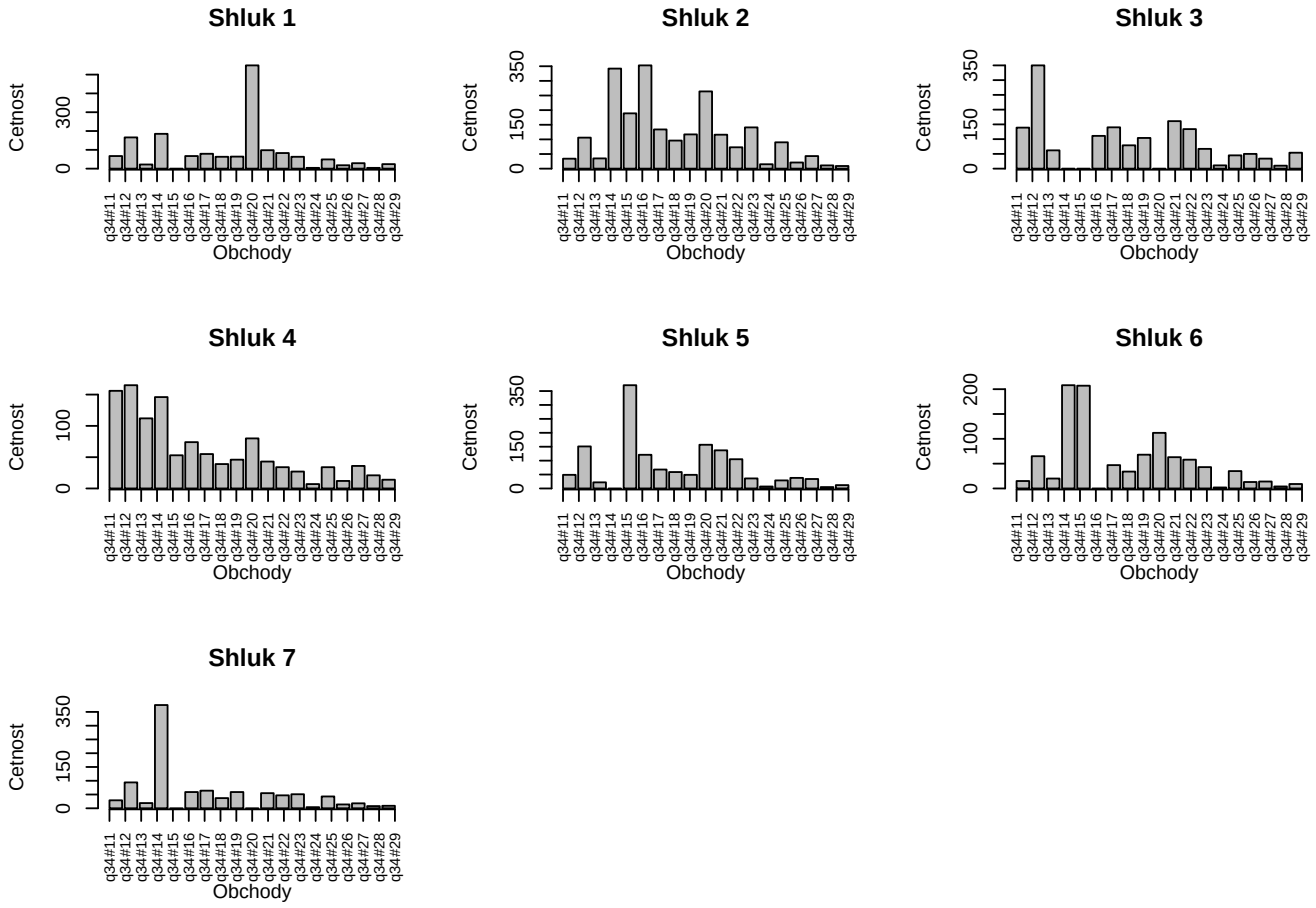
The FREQ Procedure

Cluster		
CLUSTER	Frequency	Percent
1	549	19.03
2	362	12.55
3	840	29.12
4	180	6.24
5	371	12.86
6	208	7.21
7	375	13.00

Obrázek 1.5: Četnost shluků

1.4 Interpretace shlukové analýzy

Znázorněme si nyní nakupovací návyky jednotlivých shluků na obrázku 1.6.



Obrázek 1.6: Nakupovací návyky jednotlivých shluků

Poté jsme provedli analýzu jednotlivých proměnných v jednotlivých shlucích a dále uvedeme pouze proměnné, ve kterých se shluky výrazně liší. Odlišnosti daných proměnných jednotlivých shluků znázorníme pomocí sloupcových grafů. Těmito kategoriálními proměnnými, ve kterých se jednotlivé shluky liší, jsou:

1. proměnná *reg* znázorněná na obrázku 1.7, která říká, ve kterém regionu má daný člověk trvalé bydliště, nabývá těchto hodnot

11	21	31	32	41	42	51
Praha	Středočeský	Jihočeský	Plzeňský	Karlovarský	Ústecký	Liberecký

52	53	61	62	71	72	81
Královehradecký	Pardubický	Vysočina	Jihomoravský	Olomoucký	Zlínský	Moravskoslezský

2. proměnná *vb* znázorněná na obrázku 1.8, která nám říká o velikosti města, ve kterém žije daný člověk

1	2	3	4	5	6
pod tisíc obyvatel	1–5 tisíc obyvatel	5–20 tisíc obyvatel	20–100 tisíc obyvatel	nad 100 tisíc obyvatel	Praha

3. proměnná *sex* znázorněná na obrázku 1.9 nám říká o pohlaví nakupujícího

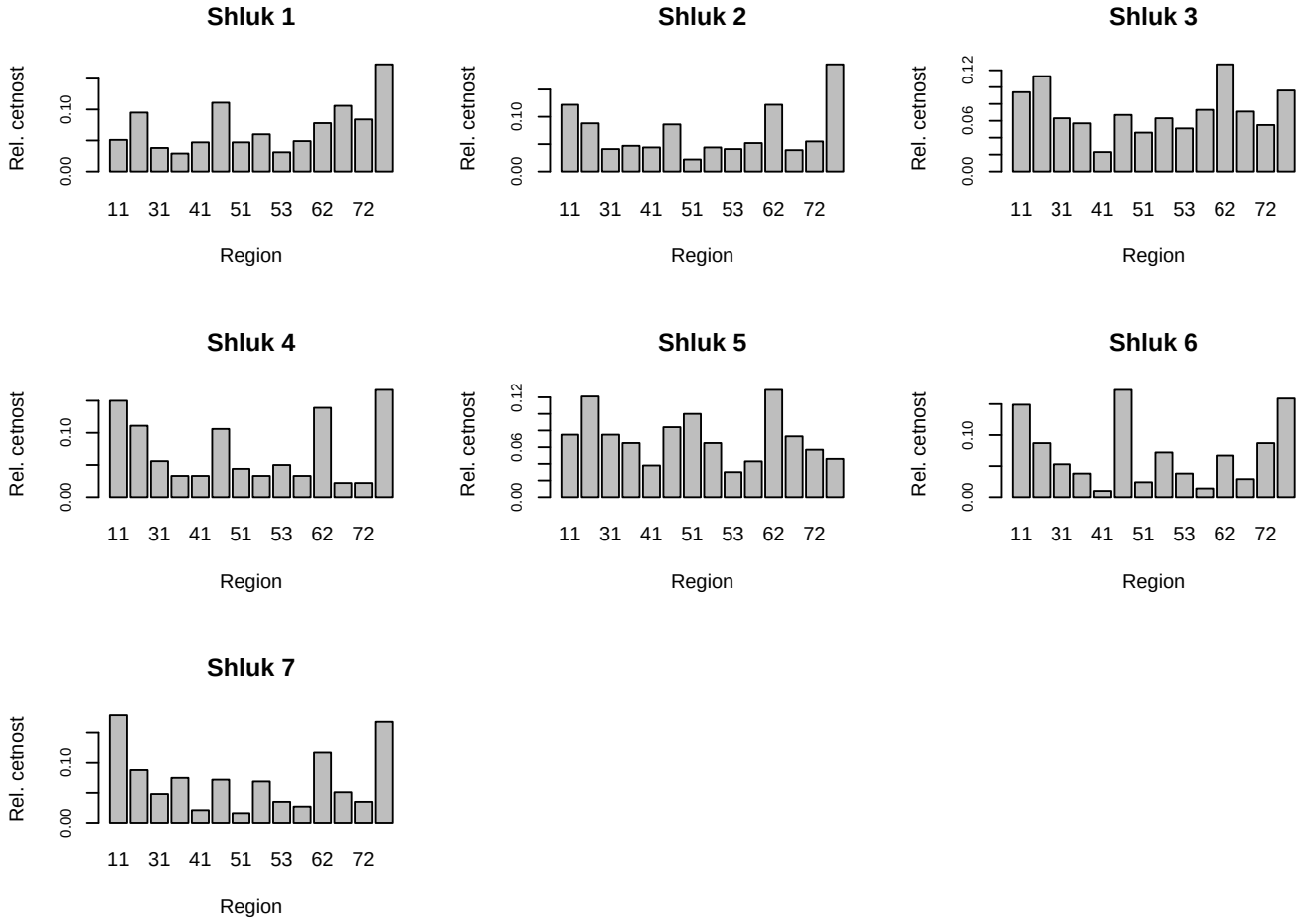
1	2
muž	žena

4. proměnná *agecat* znázorněná na obrázku 1.10, která nám zařazuje nakupujícího do jaké věkové kategorie

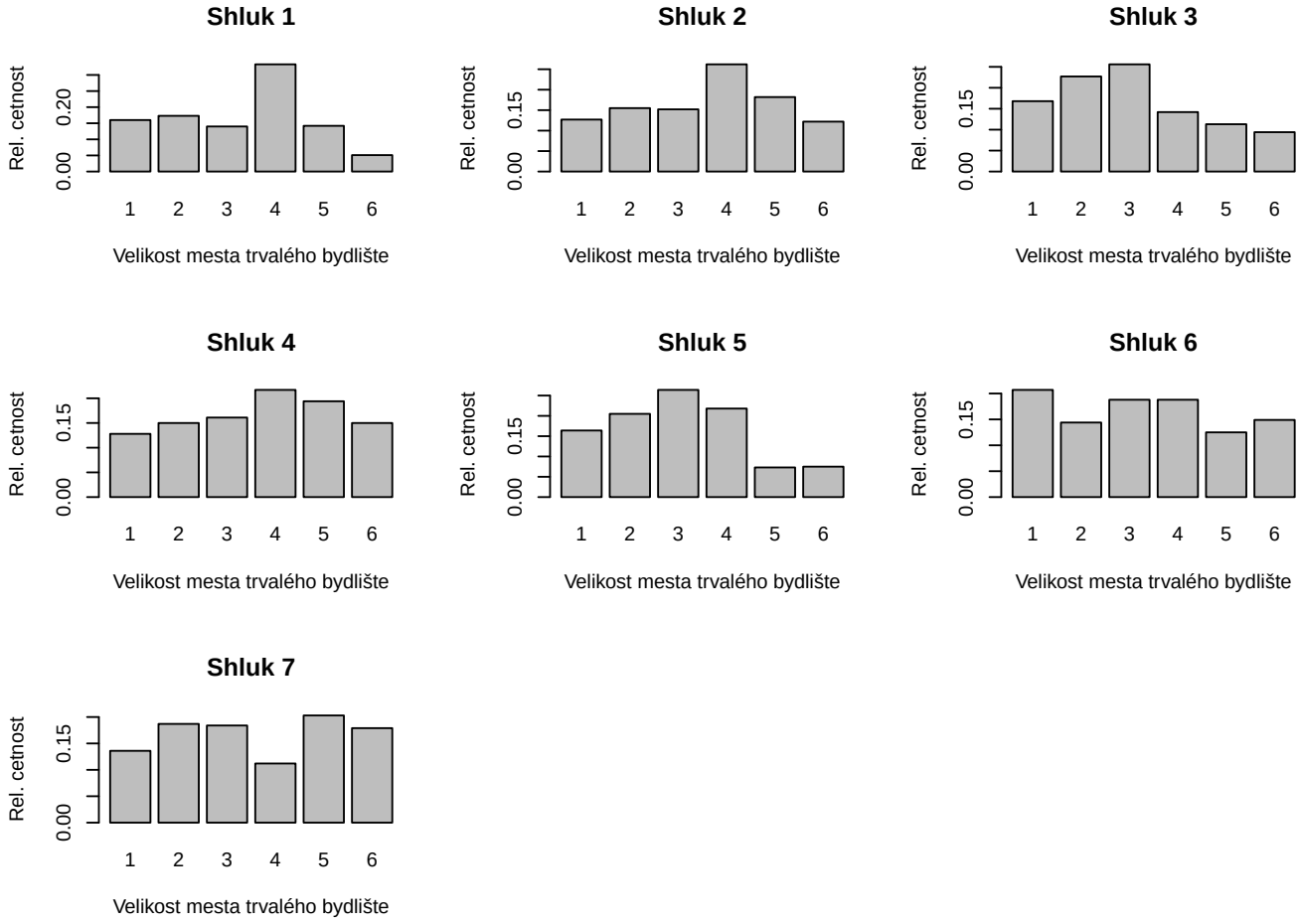
1	2	3	4	5	6
15–20	21–24	25–30	31–40	41–50	51–65

5. proměnná *marits* znázorněná na obrázku 1.11, která nám říká o partnerském statutu daného nakupujícího

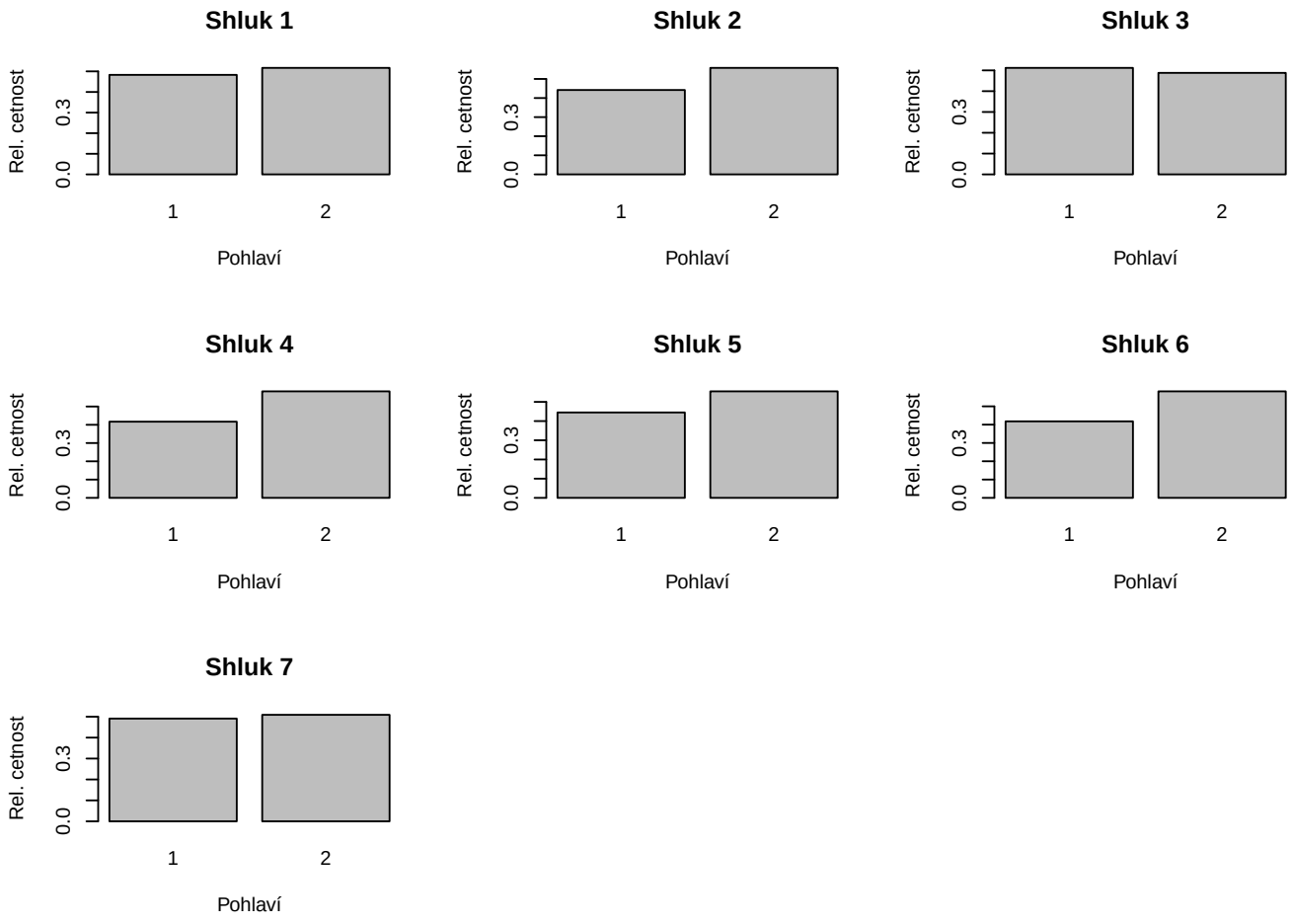
1	2	3	4	5
svobodný(á)	ženatý/vdaná	společná domácnost s partnerem	rozvedený(á)	vdovec/vdova



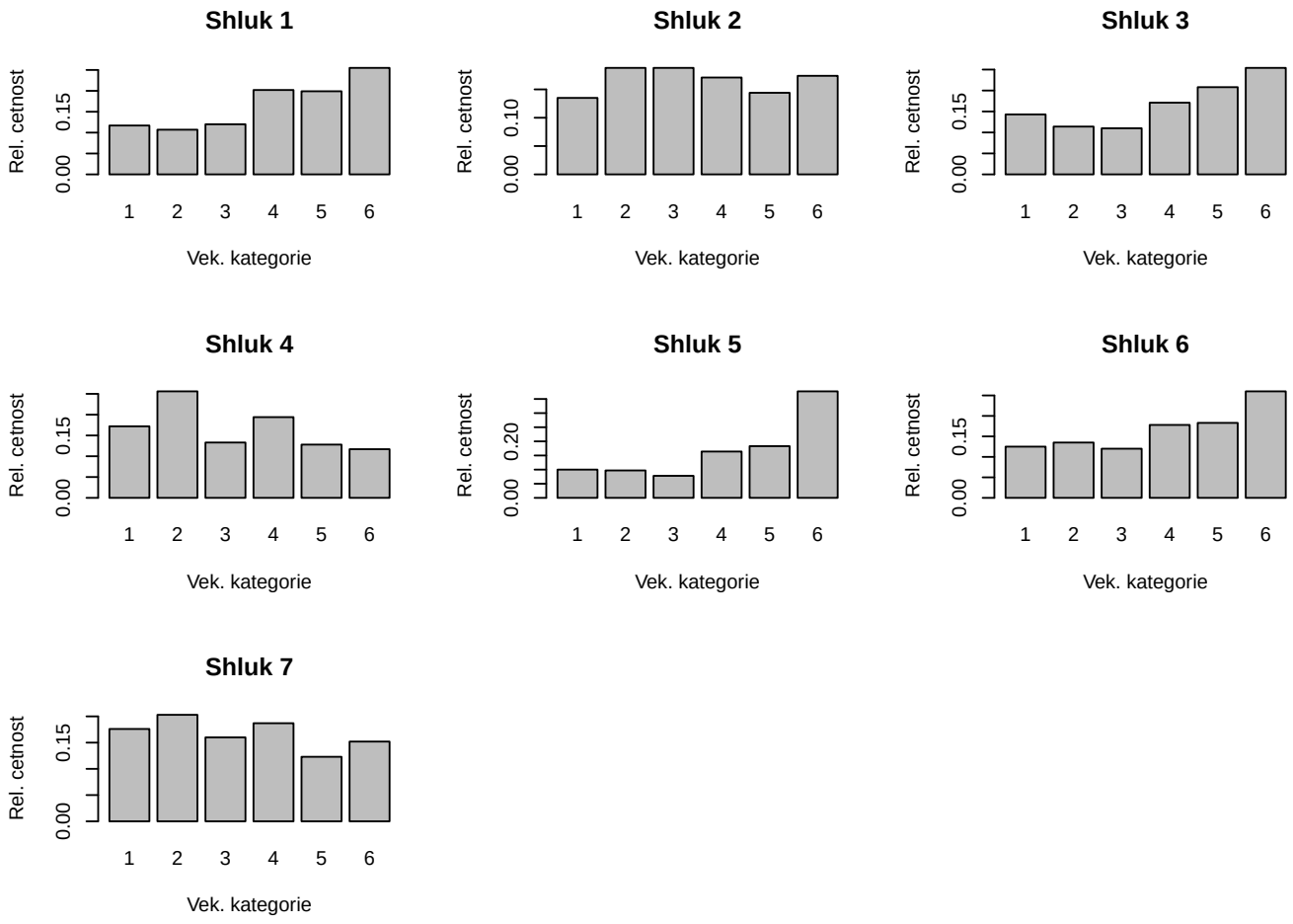
Obrázek 1.7: Četnost regionu trvalého bydliště jednotlivých shluků



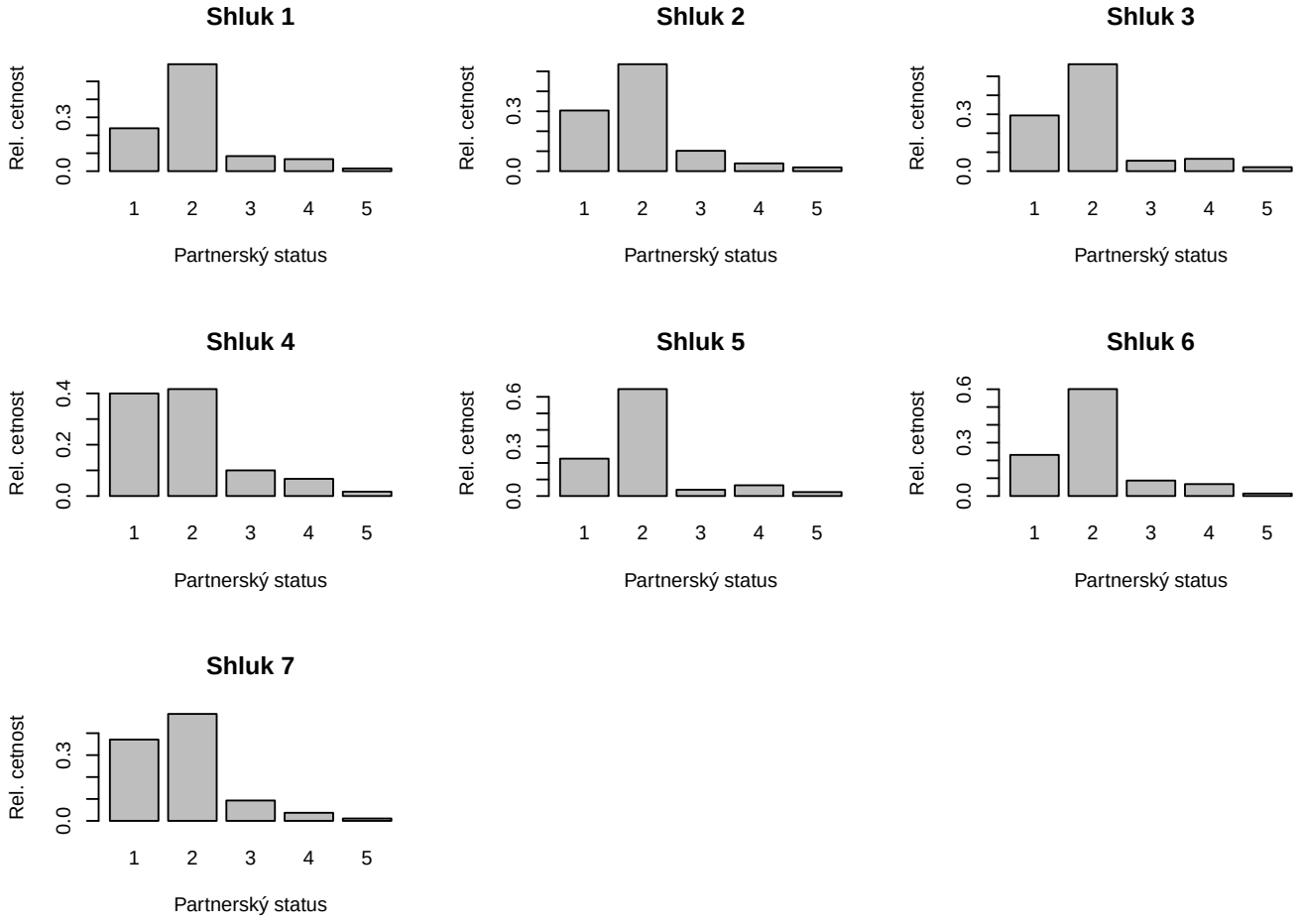
Obrázek 1.8: Četnost velikosti města trvalého bydliště jednotlivých shluků



Obrázek 1.9: Četnost pohlaví jednotlivých shluků



Obrázek 1.10: Četnost věkových kategorií jednotlivých shluků



Obrázek 1.11: Četnost partnerského statutu jednotlivých shluků

Nyní si interpretujeme jednotlivé shluky:

1. Shluk 1 můžeme označit jako shluk starších (31+ věku) ženatých lidí z Moravskoslezského kraje, který nejčastěji nakupuje v Kauflandu (proměnná *q34#20*). Velikost města trvalého bydliště je 21–100 tisíc obyvatel.

2. Shluk 2 označíme jako skupinu, ve které jsou zastoupeny téměř rovným dílem všechny věkové kategorie vdaných žen z Moravskoslezského kraje, které žijí ve městech s 20–100 tisíci obyvateli. Tyto ženy nakupují převážně v Tescu a Hypernově (proměnné *q34#14* a *q34#16*).

3. Shluk 3 označíme jako shluk starších (31+ věku) ženatých lidí z Jihomoravského a Středočeského kraje, kteří žijí v menších městech (20 tisíc obyvatel a méně). Tito lidé nejčastěji nakupují v malých samoobsluhách (proměnná *q34#12*).

4. Shluk 4 označíme jako shluk mladších žen (21–24 věku), které jsou buďto single či vdané. Žijí v Praze, Jihomoravském a Moravskoslezském kraji a nakupují nejčastěji v Malých pultových obchodech, Malých samoobsluhách, Večerkách a Tescu (proměnné *q34#11–q34#14*).

5. Shluk 5 označíme jako shluk vdaných starších žen (věku 51–65), které žijí ve městech s 5–100 tisíci obyvateli a nakupují v Lidlu (proměnná *q34#15*).

6. Shluk 6 označíme jako shluk vdaných starších žen z Ústeckého, Moravskoslezského kraje a Prahy, který nakupuje v Tescu a Lidlu (proměnné *q34#14–q34#15*).

7. Shluk 7 označíme jako shluk vdaných a single lidí z Prahy a Moravskoslezského kraje, který nejčastěji nakupuje v Tescu (proměnná *q34#14*).

Závěr

Dostali jsme za úkol shlukovat data *data.txt* podle nakupovacích návyků (proměnné *q34#11–q34#29*). Abychom odpověděli na otázky, tak: „Ano, v datech lze identifikovat skupiny osob, které mají podobné nakupovací návyky. Přesněji je to 7 skupin. A skupiny se mezi sebou odlišují několika faktory: regionem, ve kterém žijí; velikostí města, ve kterém žijí; pohlavím; věkovou kategorií; partnerským vztahem a v neposlední řadě samotnými nakupovacími návyky.“