UNIVEZITA PARDUBICE

FAKULTA EKONOMICKO SPRÁVNÍ

Předzpracování ekonomických dat pomocí metod shlukové analýzy

Pavel Novák

Bakalářská práce

2009

Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky
Akademický rok: **2008/2009**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Pavel NOVÁK**

Studijní program: **B6209 Systémové inženýrství a informatika**

Studijní obor: **Informační a bezpečnostní systémy**

Název tématu: **Předzpracování ekonomických dat pomocí metod shlukové analýzy**

Zásady pro vypracování:

Metody shlukové analýzy
Předzpracování dat
Modelování ekonomických dat vybranými metodami shlukové analýzy
Analýza výsledků

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

MAŘÍK, V. a kol. Umělá inteligence (1). Praha : Academia, 1993.
JAIN, A. K., MURTY, M. N., FLYNN, P. J. Data Clustering: A Review. ACM Computer Surveys. 1999, vol.31., no.3, pp.264-323.
LUKASOVÁ, A., ŠARMANOVÁ, J. Metody shlukové analýzy. Praha : Státní nakladatelství technické literatury, 1985.
MELOUN, M., MILITKÝ, J. Statistická analýza experimentálních dat, Praha : Academia, 2004.
HEBÁK, P. Vícerozměrné statistické metody 3. Praha : Informatorium, 2007.

Vedoucí bakalářské práce: **Ing. Petr Hájek, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **6. října 2008**

Termín odevzdání bakalářské práce: **1. května 2009**

L.S.

doc. Ing. Renáta Myšková, Ph.D.
děkanka

doc. Ing. Jiří Křupka, Ph.D.
vedoucí ústavu

V Pardubicích dne 6. října 2008

Prohlašuji:

Tuto práci jsem vypracoval samostatně. Všechny literární prameny, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 15. 04. 2009

Pavel Novák

Rád bych touto cestou poděkoval panu Ing. Petru Hájkovi, Ph.D. za odborné vedení, konzultace a cenné rady při zpracování této bakalářské práce. Zároveň bych rád poděkoval všem, kteří se mnou v průběhu práce spolupracovali.

ANOTACE

Bakalářská práce pojednává o předzpracování dat pomocí shlukové analýzy. První část práce je zaměřena na samotné předzpracování dat a vysvětlení metod shlukové analýzy. Jedná se o metody hierarchického a nehierarchického shlukování.

V druhé části práce je uveden příklad, na kterém je předvedeno modelování ekonomických dat vybranými metodami shlukové analýzy. Použita je Wardova metoda pro hierarchické shlukování a metoda k-průměru pro nehierarchické shlukování. Dosažené výsledky pomocí metod shlukové analýzy jsou dále analyzovány a porovnávány.

KLÍČOVÁ SLOVA

Shluková analýza, předzpracování dat, úvěrová schopnost, metody shlukování

TITLE

Economic data pre-processing using clustering methods

ANNOTATION

Bachelor's thesis deals with the pre-data using cluster analysis. The first part of the work is focused on its own pre-data and an explanation of cluster analysis methods. That is the method of hierarchical and unhierarchical clustering.

In the second part of the work is an example, which presents modeling of economic data by selected cluster analysis methods. Ward´s method is used for hierarchical clustering and k-average method for unhierarchical clustering. The results achieved by using cluster analysis methods are analyzed and compared.

KEY WORDS

Cluster analysis, data preprocessing, credit capacity, clustering methods

OBSAH

Seznam tabulek

Seznam grafů

Seznam obrázků

Úvod

Jako téma bakalářské práce bylo vybráno předzpracování ekonomických dat pomocí metod shlukové analýzy. Shluková analýza (též clusterová analýza, anglicky cluster analysis) je vícerozměrná statistická metoda, která se používá ke klasifikaci objektů. Slouží k třídění jednotek do skupin (shluků) tak, aby si jednotky náležící do stejné skupiny byly podobnější než objekty ze skupin různých. Shlukovou analýzu je možné provádět jak na množině objektů, z nichž každý musí být popsán prostřednictvím stejného souboru znaků, které má smysl v dané množině sledovat, tak na množině znaků, které jsou charakterizovány prostřednictvím určitého souboru objektů, nositelů těchto znaků [9].

V první části bude popsán smysl shlukové analýzy. Bude kladen důraz na jednotlivé kroky, které jsou potřebné pro dosažení co nejpřesnějších výsledků shlukování. Mezi ně patří zejména předzpracování samotných dat. Do předzpracování je zařazeno na příklad hledání odlehlých hodnot a práce s nimi, dále nalezení chybějících hodnot a jejich nahrazení, standardizování a normalizování dat, atd. Dalším důležitým krokem je zjišťování závislostí pomocí korelačních koeficientů, což je důležité pro odhalení závislých znaků, které by ovlivňovaly výsledek shlukování. Dále budou popsány metody shlukové analýzy, které se dělí na hierarchické a nehierarchické. Z hierarchického shlukování, které se dělí na nomotetické a polytetické, bude vysvětlena metoda průměrné vazby pro mezishlukové vzdálenosti, metoda průměrné vazby pro vnitroshlukové vzdálenosti, metoda nejbližšího souseda, atd.

V další části je uveden příklad, na kterém je použito jak hierarchické, tak i nehierarchické shlukování. K dispozici jsou data německé banky, která popisují žadatele o poskytnutí úvěru různými znaky včetně rozdělení do tříd, což je výhodné pro srovnání dosažených výsledků s realizovanou klasifikací shluků. Hierarchické shlukování bude provedeno na základě Wardovy metody. Pro nehierarchické shlukování bude použita metoda k-průměru. Příklady budou realizovány v programovém prostředí STATISTICA 7.

Dílčím cílem práce je shrnout současné metody shlukové analýzy. Dále, na základě toho shrnutí, je cílem práce realizovat předzpracování dat o úvěrové schopnosti žadatelů o poskytnutí úvěru. Toto předzpracování bude realizováno v několika krocích. Nejprve bude zajištěno splnění požadavků na realizaci shlukové analýzy. Dále budou aplikovány vybrané metody shlukové analýzy. Výsledkem bude předzpracování uvedených dat v tom smyslu, že data budou připravena pro další modelování například pomocí metod učení s učitelem (např. diskriminační nebo regresní analýza) a navíc bude porovnáno, zdali nalezené shluky v rámci

shlukové analýzy jsou v souladu s třídami přiřazenými jednotlivým objektům. Toto se týká jednak počtu tříd v objektech a dále, zda-li klasifikace do tříd splňuje podmínky podobnosti objektů ve stejných třídách a naopak podmínku nepodobnosti objektů v různých třídách.

1 Shluková analýza a její metody

Základním cílem shlukové analýzy [7] je zařadit objekty do skupin (shluků), a to především tak, aby objekty stejného shluku si byly více podobné, než objekty z různých shluků. Přitom objekty mohou být různého charakteru. Lze shlukovat živočichy či rostliny, stejně jako textové dokumenty či webové stránky. Aby bylo dosaženo uvedeného cíle, je potřeba vyřešit celou řadu dílčích úkolů.

Prvním problémem je definování podobnosti dvou objektů. Aby mohla být podobnost kvantifikována, musí být každý objekt charakterizován pomocí svých vlastností. Například textový dokument je charakterizován klíčovými slovy, minerální voda koncentracemi určitých iontů, rostlina tvarem listů, barvou květů, délkou a šířkou okvětních lístků atd.

Pokud jde o aplikace metod shlukové analýzy, pak v posledních letech je pozornost ve velké míře soustředěna například na shlukování dokumentů, ať už klasických textových či webových, včetně jejich speciálních formátů. Se vzrůstajícím rozsahem informačních zdrojů roste potřeba jejich uspořádání, což je úloha těsně související právě se shlukováním. Význam shlukové analýzy spočívá v usnadnění vyhledávání informací, které jsou potřebné ve všech oblastech lidského života. Jsou důležité nejen pro výuku a vědecký výzkum, ale též pro běžné činnosti, jako je nakupování, cestování, kulturní využití a další.

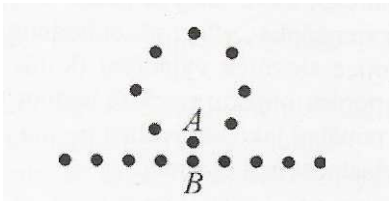
1.1 Konvenční a konceptuální shlukování

Konvenční shlukování je založené na měření podobnosti. Označme si dva objekty jako A a B. Symbolicky lze zapsat, že podobnost [7]

$$(A, B) = f\,(\text{vlastnosti}(A), \text{vlastnosti}(B)), \qquad (1)$$

tedy podobnost dvou objektů je funkcí jejich vlastností. Je však třeba poznamenat, že uvedené pojetí je značným zjednodušením reality a v některých případech může být interpretace výsledných skupin velmi obtížná.

V případě konceptuálního shlukování jsou vytvářené shluky založené na konceptuální soudržnosti, která je funkcí vlastností objektů, popisného jazyka a okolí. Popisný jazyk je způsob, jakým jsou popsány třídy (skupiny) objektů, a okolí je množina sousedících vzorů. Symbolicky můžeme zapsat, že konceptuální soudržnost (A,B)=f(vlastnost(A), vlastnosti(B), jazyk, okolí).



Obrázek 1 - Znázornění rozdílu blízkosti a konceptuální soudržnosti, zdroj [7]

Na obrázku 1 je uveden příklad, v němž objekty představují body v dvourozměrném prostoru. To znamená, že každý bod je charakterizován množinou dvou reálných čísel, z nichž první určuje hodnotu na ose X a druhé hodnotu na ose Y (bod se nachází na průsečíku přímek procházejících těmito místy na osách). Z těchto důvodů jsou vytvořeny dva obrazce, které připomínají kružnici a přímku. Prozkoumají-li se všechny dvojice bodů, pak nejblíže se nacházejí body A a B, jsou si tedy nejvíce podobné. Při konvenčním shlukování by dané body byly zařazeny do stejného shluku. Avšak konceptuální soudržnost těchto dvou bodů je malá, neboť patří do konfigurací reprezentujících různé koncepty. V praxi se konceptuální shlukování využívá například při analýze textových databází.

Konceptuální shlukování však předpokládá, že jsou k dispozici charakteristiky shluků, do kterých mohou být objekty zařazeny. Například maximální prediktivní klasifikace, která je založena na popisu objektů pomocí posloupností znamének „+“ a „-“ (v praxi se častěji používají hodnoty 1 a 0). Tato znaménka mohou vyjadřovat přítomnost, resp. nepřítomnost určité sledované vlastnosti (zda má objekt určitou barvu, chuť, vůni atd.). Každý možný shluk pak reprezentuje fiktivní objekt popsaný určitou posloupností těchto znamének. Objekt je zařazen do takového z předpokládaných shluků, s jehož fiktivním objektem má nejvíce společných vlastností (shoduje se nejvíce znamének). Vychází se tedy z předpokladu, že nejsou k dispozici žádné informace o třídách, do nichž mají být objekty zařazeny.

1.2 Typy proměnných dat

Při shlukování je každý objekt reprezentován množinou vlastností. Jsou tedy stanoveny znaky (proměnné, veličiny), které je třeba sledovat. Tyto proměnné mohou být různých typů:

- **Číselné**: jako délka a šířka okvětních lístků, koncentrace iontů v minerálních vodách apod.

- **Ordinální**: hodnoty lze uspořádat, nemusí se však přitom jednat o čísla. Intenzita osvětlení může být slabá, střední, silná, případně žádná. Obdobně lze slovně charakterizovat velikost zvířat (malé, střední, velké, případně velmi malé a velmi velké).
- **Nominální**: hodnoty uspořádat nelze. Bude se charakterizovat rostlinu pomocí listů, pak můžeme rozlišit listy celistvé, složené, případně zvláštní listové útvary. Podle potřeby lze popis více specifikovat (celistvý list okrouhlý, vejčitý, podlouhlý, klínovitý atd.)
- **Dichotomická**: nabývá pouze dvou různých hodnot. Příkladem jsou již výše uvedená znaménka „+“ a „-“ či hodnoty 1 a 0 (ve významu „ano“ a „ne“), případně jiné dvojice slovních vyjádření (kuřák a nekuřák, zaměstnaný a nezaměstnaný apod.) Při popisu objektu pomocí hodnot ve významu „ano“ a „ne“ se někdy termín vlastnost používá jako ekvivalent pojmu proměnná. Sleduje se, zda určitý objekt sledovanou vlastnost má, či nemá. V literatuře [7] je uvažován též popis objektů pomocí znaků, které jsou označovány jako symbolické. Příkladem je interval hodnot nebo pravděpodobnostní rozdělení s určitými parametry.
- **Fuzzy:** Jedním z typů fuzzy dat [6] je dvojice hodnot, z nichž jedna vyjadřuje střed intervalu a druhá rozpětí, pomocí něhož se získá dolní, resp. horní hranice intervalu. Jde o symetrická fuzzy data. Příkladem symetrických fuzzy dat je charakteristika pacientů založená na opakovaných měřeních krevního tlaku. Na základě měření prováděných v průběhu jednoho dne je zjištěna minimální a maximální hodnota (systolického a diastolického tlaku). Tyto hodnoty jsou pak základem pro vytvoření vstupního datového souboru pro analýzu, v němž je tlak charakterizován pomocí středu intervalu a rozpětí.V případě fuzzy dat se samozřejmě pro vyjádření podobnosti dvou objektů používají jiné funkce než pro reálná data.

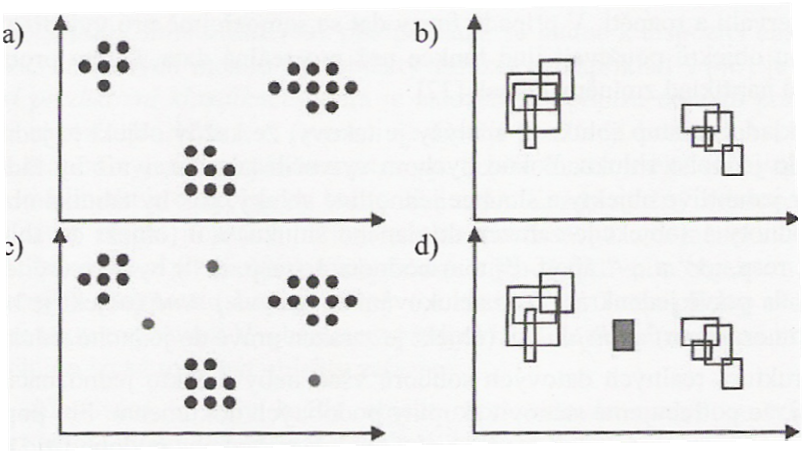
1.3 Základní přístupy ke shlukové analýze

Základní přístup shlukové analýzy [6] je takový, že každý objekt je jednoznačně zařazen do jednoho shluku. Pokud by se vytvořila tabulka, v níž by řádky představovaly jednotlivé objekty a sloupce jednotlivé shluky,pak by tabulka obsahovala pouze hodnoty 1 (objekt je zařazen do daného shluku) a 0 (objekt do shluku není zařazen), resp. „+“ a „-“ apod. Přitom hodnota 1 (resp. „+“) by se v určitém řádku vyskytovala právě jedenkrát. Toto shlukování se nazývá pevné (objekt je buď zařazen, nebo nezařazen) a disjunktní (objekt je zařazen právě do jednoho shluku).

Struktura reálných datových souborů však nebývá takto jednoznačná. Je na příklad potřeba stanovit skupiny podobných dokumentů. Pro popis dokumentů se vytvoří seznam slov, podle nichž má být zjišťována podobnost. Dokument pak bude popsán posloupností hodnot 0 (slovo se v dokumentu nevyskytuje) a 1 (slovo se vyskytuje). Některý dokument může pojednávat jak o aplikaci statistických metod, tak o aplikaci neuronových sítí. Pokud by výsledné shuky představovaly tématicky zaměřené dokumenty, měl by být zmíněný dokument zařazen do dvou shluků. Výsledkem jsou tedy překrývající se shluky.

Na základě typu dat a typu shlukování lze vymezit čtyři situace:

1. *pevná data a disjunktní shlukování* , viz obr. 2a (lze rozlišit 3 shluky bodů)
2. *fuzzy data a disjunktní shlukování*, viz obr. 2b (lze rozlišit 2 shluky obdélníků)
3. *pevná data a překrývající se shlukování*, viz obr. 2c (šedě zakreslené body mohou být přiřazeny současně ke dvěma shlukům)
4. *fuzzy data a překrývající se shlukování*, viz obr. 2d (šedě zakreslený obdélník může být přiřazen současně ke dvěma shlukům)



Obrázek 2 - Grafická reprezentace různých informačních situací, zdroj [7]

Analýza však může jít ještě hlouběji, kdy výsledná tabulka přiřazení objektů do shluků bude místo diskrétních hodnot 0 a 1 obsahovat reálná čísla z intervalu 0 a 1, vyjadřující stupeň příslušnosti objektu k danému shluku. Součet těchto hodnot pro každý jednotlivý objekt se přitom rovná hodnotě jedna. V tomto případě se hovoří o fuzzy shlukové analýze.

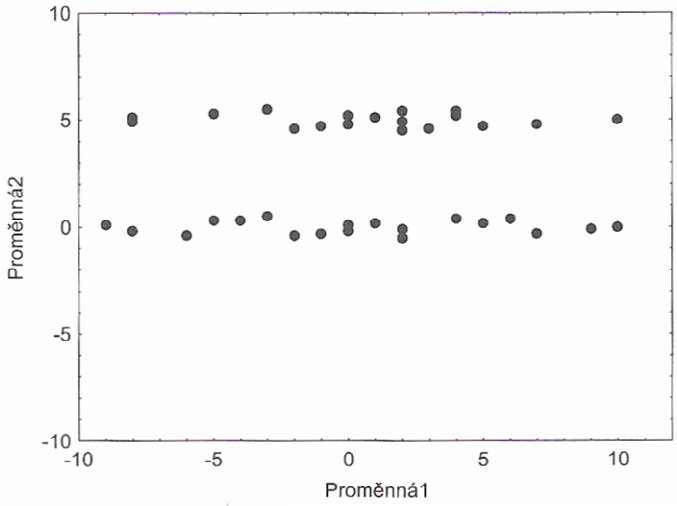
1.4 Stanovení optimálního počtu shluků

Ke stanovení počtu shluků se obvykle využívají dva základní přístupy, a to *heuristické procedury* [7]a *formální texty* [7]. Jako nejjednodušší příklad prvního přístupu lze uvést

navržení počtu shluků na základně dendrogramu, v němž mohou být v některých případech znázorněny výrazné shluky.

Jiný graf, který může být použit, je graf závislosti hodnot fúzních koeficientů na počtu shluků (fúzní koeficient pro k shluků je definován jako průměr maximálních vzdáleností uvnitř těchto k shluků). Je vybrán takový počet shluků, jemuž odpovídá určité zploštění v grafu.

Stanovení optimálního počtu shluků souvisí s ověřováním platnosti shluků, které se využívá při porovnání výsledků získaných různými metodami. V některých případech mohou existovat shluky, kdy vzdálenost dvou objektů patřících do stejného shluku je větší než vzdálenost dvou objektů patřících do různých shluků. Pokud jsou objekty charakterizovány pouze dvěma proměnnými, lze takový příklad názorně zobrazit graficky, viz obrázek 3.



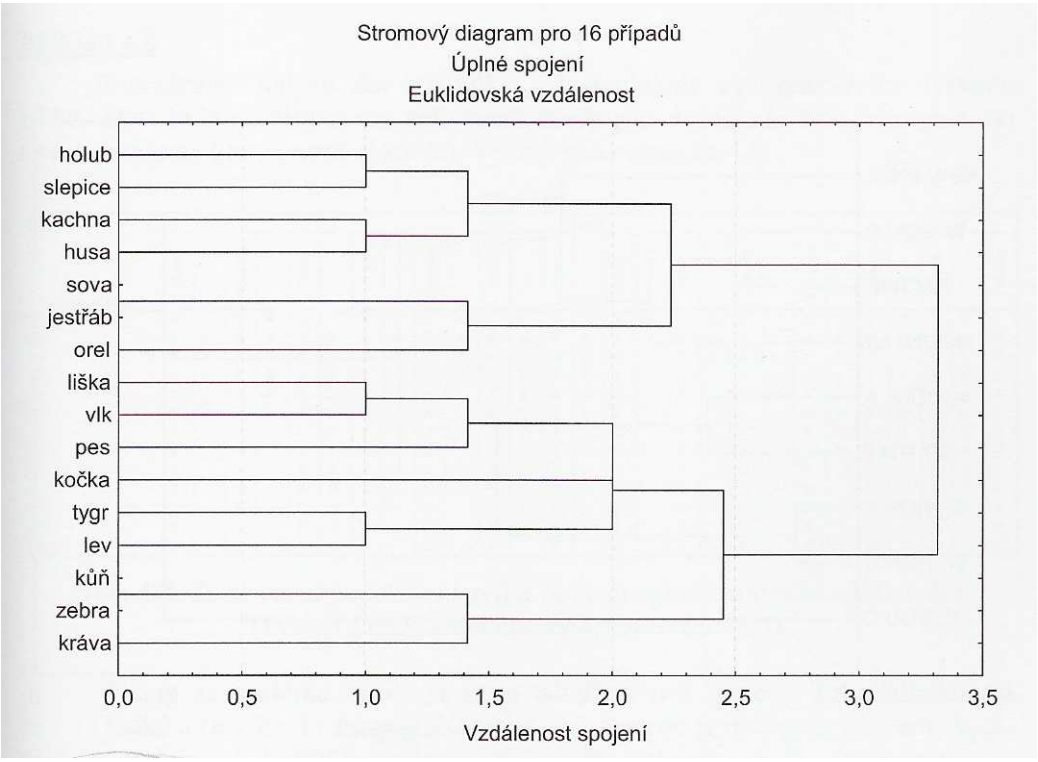
Obrázek 3 - Bodový graf dvou shluků, zdroj [7]

Z uvedeného poznatku vycházejí Qiu a Joe, kteří navrhují separační index [5]. Tento index stanovuje velikost mezery mezi dvěma shluky. Separační index je spočten pro všechny dvojice shluků stanovených na základě určitého způsobu shlukování. Pro porovnání výsledků získaných různými metodami, nebo pro zjištění optimálního počtu shluků, je pro každé shlukování vytvořena matice separačních indexů s hodnotami -1 na diagonále. Pro vhodně stanovenou množinu shluků by měly být všechny separační indexy charakterizující vztah dvou různých shluků kladné. Záporné hodnoty nebo hodnoty blízké nule mohou indikovat nevhodně specifikované shluky.

1.5 Grafická reprezentace shluků

V případě bodových grafů [7] teoreticky existují dvě varianty, tj. dvourozměrná a trojrozměrná, některé programové systémy však poskytují jen jednu. V bodových grafech jsou zakreslovány jednotlivé objekty jako body, přičemž se různé typy grafů liší reprezentací použitých os. Tyto osy mohou určovat hodnoty buď konkrétních proměnných, nebo hlavních komponent, případně dimenzí zjištěných pomocí vícerozměrného škálování.

Postup shlukování výstižně znázorňuje speciální graf, který se jmenuje dendrogram viz graf 1. Jde o stromový diagram, který znázorňuje postupné shlukování jinak jednotlivých objektů, jednak shluků vytvořených v předchozích krocích. Programové systémy ho vytváření buď v horizontální (objekty jsou uvedeny na ose Y) nebo vertikální (objekty jsou uvedeny na ose X) podobě. V prvním případě je na ose Y zakresleno n listů (jednotlivých objektů). Z těchto listů vycházejí větve. Nejprve se větve dvou objektů, mezi nimiž je nejmenší vzdálenost (resp. nepodobnost), spojí do jedné větve. Hladinou spojení, jejíž hodnota je zaznamenána na ose X, je právě tato vzdálenost. Další postup spojování větví odpovídá postupnému spojování shluků při aglomerativním shlukování.



Graf 1 – Dendogram, zdroj [7]

2 Předzpracování dat pro shlukovou analýzu

V kapitole jsou uvedeny základní kroky, které je třeba dodržet, aby bylo pomocí shlukové analýzy dosaženo co nejspolehlivějších výsledků. Jedná se zejména o problémy spojené s chybějícími a odlehlými hodnotami, objekty a závislosti mezi znaky. Pokud by totiž uvedené problémy nebyly ošetřeny, výsledky shlukování by mohly být zkreslené.

2.1 Náhrada chybějících hodnot a problematika odlehlých objektů

Datová matice nemusí být kompletní z nejrůznějších důvodů. Je zřejmé, že pokud u některého objektu chybí větší počet údajů, lze doporučit vypuštění tohoto objektu. Podobně pokud některý uvažovaný znak nebyl změřen anebo chybí údaje u většího počtu objektů, tak je vhodné vypuštění tohoto znaku. Naproti tomu hodnoty, jejichž nepřítomnost lze považovat za náhodnou (bez souvislosti se zkoumanými vlastnostmi objektů), se doplní uměle, čímž je alespoň částečně zmenšena ztráta informace.

Tabulkové kalkulátory automaticky identifikují chybějící hodnotu a označují ji v příslušném políčku tečkou. Statistické pakety při třech a více proměnných kromě uvedených možností umělé náhrady chybějících hodnot, nabízejí ještě jednu možnost. Místo nastavené varianty (listwise), při které výskyt chybějící hodnoty libovolného znaku má za následek automatické vyloučení celého řádku datové matice s alespoň jednou chybějící hodnotou z analýzy, je možné zvolit alternativní méně ztrátovou variantu (pairwice), která umožňuje při hodnocení dvojice proměnných vyloučit jen ty řádky, které se přímo týkají alespoň jedné z proměnných bez ohledu na to, že v jiných sloupcích použitých řádků nějaké údaje chybí.

Dále jsou uvedeny různé možnosti umělé náhrady [2] chybějící hodnoty v datové matici. Vesměs jde v podstatě o odhad, opírající se o zbývající data:

- *Náhrada průměrem* příslušné proměnné nebo příslušného objektu je nejjednodušší možnost, která však nerespektuje ani variabilitu, ani korelační strukturu dat.
- *Náhrada náhodným číslem* z rozdělení příslušné proměnné s parametry odhadnutými z výběru, její používání by již nevedlo k zkreslenému odhadu rozptylu, souvislosti s ostatními proměnnými však stále nejsou respektovány, tabulky náhodných čísel z rozdělení N(0,1), jinak se použije vhodný podprogram generující náhodná čísla.

- *Náhrada regresí*, tedy odhadem založeným na regresní rovnici charakterizující závislost příslušné proměnné na p–1 zbývajících proměnných (popř. jen na jedné nebo několika silně korelovaných proměnných) a vypočtené ze zbývajících n–1 objektů.

Je třeba, stejně jako u odlehlých pozorování, zvýšená opatrnost, protože možnost něco vypočítat ještě neznamená smysluplné výsledky a plnohodnotné použití dané techniky. Někdy ani jinak osvědčené náhrady chybějících pozorování nemusí stačit a přes nespornou ztrátu informace je jediným korektním řešením vypuštění všech řádků s chybějícími hodnotami některých proměnných.

Odlehlé objekty tvoří při použití základních algoritmů samostatné shluky. Ke zjištění těchto objektů se ovšem používají opět metody shlukové analýzy. Jsou-li odlehlé objekty identifikovány, pak by měly být ve vstupní matici vynechány, což je shodné s jedním z přístupů aplikovaným v případě výskytu chybějících údajů. V některých metodách je identifikace odlehlých objektů zahrnuta přímo do shlukovacího algoritmu.

2.2 Závislosti mezi znaky

Nejčastěji se závislosti mezi znaky zjišťují pomocí výběrového *korelačního koeficientu* [2]

$$A_r\left(x_j, x_{j'}\right) = \frac{s\left(x_j, x_{j'}\right)}{s\left(x_j\right)s\left(x_{j'}\right)} = \frac{\sum_{i=1}^{n}\left(x_j - \bar{x}_j\right)\left(x_{ij'} - \bar{x}_{j'}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2 \sum_{i=1}^{n}\left(x_{ij'} - \bar{x}_{j'}\right)^2}} \qquad (2)$$

Pro všechny páry objektů lze získat výběrovou korelační matici typu p x p, která má na diagonále jedničky.

Obvykle je třeba převést získanou matici na matici nepodobností. Existují dva přístupy, podle interpretace hodnoty -1. V případě, kdy hodnota -1 reprezentuje maximální nesouhlas, platí vztah D = 1 – A. Pokud jsou ovšem hodnoty -1 a 1 uvažovány jako maximální souhlas mezi proměnnými, pak můžeme použít buď $D = 1 - A^2$, nebo $D = 1 - |A|$.

Pro některé aplikace se doporučuje použít jako míru podobnosti kosinus úhlu mezi příslušnými dvěma vektory. Tato *kosinova míra* je spočtena podle vztahu [9]

$$A_c\left(x_j, x_{j'}\right) = \frac{\sum_{i=1}^{n} x_{ij} x_{ij'}}{\sqrt{\sum_{i=1}^{n} x_{ij}^{\ 2} \sum_{n=1}^{n} x_{ij'}^{\ 2}}}, \tag{3}$$

což je speciální případ výběrového korelačního koeficientu, kdy jsou výběrové průměry u obou sledovaných proměnných rovny hodnotě 0.

2.3 Standardizace a normalizace dat

Hodnoty jednotlivých znaků objektů jsou často v různých jednotkách. To může způsobovat, že se určité znaky jeví jako dominující a jiné znaky jen málo ovlivňují průběh shlukování [1]. Někdy je proto výhodné data upravit tak, aby byly všechny znaky souměřitelné. Jedním ze způsobů, jak toho docílit, je *standardizace dat*.

Nechť je dána matice dat $Z = (z_{ij})$ typu $n \times p$, jejíž řádky jsou p-rozměrné vektory čísel charakterizující n objektů. Standardizaci dat lze provést ve dvou krocích:

- Je vypočtena střední hodnota z_j j - tého znaku z_j a směrodatná odchylka s_j pro $j = 1,2\ldots p$, podle vzorců [9]:

$$\bar{z}_j = \frac{1}{n} \sum_{i=1}^{i} z_{ij}\ , \tag{4}$$

$$s_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(z_{ij} - \bar{z}_j \right)^2}\ . \tag{5}$$

- Původní hodnoty z_{ij} j-tého znaku i-tého objektu jsou přepočteny na tzv standardizované hodnoty [9]:

$$x_{ij} = \frac{\left(z_{ij} - \bar{z}_j \right)}{s_j}. \tag{6}$$

Tyto standardizované hodnoty znaků mají nyní střední hodnotu rovnu 0 a rozptyl 1.

Objekty pro shlukovou analýzu jsou určeny vektory představujících hodnoty vybraných p znaků. Normy těchto vektorů mohou nežádoucím způsobem ovlivňovat výsledky kvantitativního hodnocení podobnosti objektů [1]. V takových případech je vhodné normalizovat tyto vektory, aby měly stejnou normu (nejlépe jednotkovou).

3 Metody shlukové analýzy

Shluková analýza je zastřešující název pro skupinu metod, jejichž cílem je buď seskupit zadané objekty do shluků, nebo vytvořit hierarchii shluků objektů. Za zakladatele shlukové analýzy jsou považování Tryon, Ward a James [7]. V poslední době byla navržena řada nových algoritmů. Některé modifikují tradiční metody shlukové analýzy, jiné se vydávají novými směry. Protože nové metody jsou vyvíjeny často mimo sféru statistické analýzy dat (jde například o vědní obory z oblasti informatiky, jako je vyhledávání informací v rozsáhlých databázích či rozpoznávání vzorů), převažuje v literatuře [5] termín shlukovací techniky. Avšak někteří autoři používají pojem shluková analýza pro širší okruh metod, než jsou tradiční.

3.1 Hierarchické metody

Výsledkem hierarchických [1] metod je vytvoření hierarchie skupin objektů. Tyto metody lze rozdělit na aglomerativní (postupné shlukování objektů) a divizivní (postupné rozdělování množiny objektů na podmnožiny). Podle toho, zda se při vytváření shluků přihlíží pouze k dané vybrané proměnné nebo ke všem, rozlišuje se shlukování monotetické a polytetické. Méně často uváděnou klasifikací je členění metod na jednorozměrné a dvourozměrné, při kterém se shlukují současně objekty i proměnné, resp. současně kategorie dvou proměnných.

3.1.1 Monotetické shlukování

Monotetickou analýzou [7] se označuje speciální typ divizivního shlukování, aplikovaného na binární data. Na začátku se vychází z jediného shluku, který se má rozdělit do dvou. To lze učinit podle hodnot libovolné proměnné (jedna skupina bude obsahovat jedničky v této proměnné, druhá nuly). Vyjde-li se ze zavedeného značení počtu proměnných symbolem m, pak existuje m potenciálních rozdělení všech objektů do dvou skupin. Pro další dělení je k dispozici m-1 možností atd. Kritérium pro dělení je založeno na měření závislosti dvou proměnných. Monotetické shlukování má mimo jiné tu výhodu, že po provedení analýzy lze snadno zařadit nový objekt, který nebyl obsažen v původně analyzovaných datech. Každý shluk je definován nulami či jedničkami určitých proměnných, čímž jsou vytvořena alokační

pravidla pro zařazení nových objektů. Monotetická analýza je označována též jako jeden z přístupů konceptuálního shlukování.

3.1.2 Polytetické shlukování

U tohoto typu se rozlišuje shlukování aglomerativní a divizivní [3]. V literatuře [7] se v prvním případě lze setkat s názvem AGNES (AGlomerative NESting), v druhém případě s názvem DIANA (DIvisive ANAlysis), podle názvů programových prostředků. Při aglomerativním hierarchickém shlukování se vychází z toho, že na počátku je každý objekt samostatným shlukem. Postupuje se po krocích, přičemž se v každém kroku spojí dva nejpodobnější shluky. První shluk je vytvořen ze dvou objektů na základě matice (ne)podobnosti. V dalších krocích je (ne)podobnost shluků stanovována pomocí různých aglomerativních algoritmů (pro zjednodušení vysvětlení bude dále používán pouze pojem vzdálenost). Nejčastěji používané jsou:

- *Metoda nejbližšího souseda*. Postup je postaven na minimální vzdálenosti. Naleznou se dva objekty oddělené nejkratší vzdáleností a umístí se do shluku. Další shluk je vytvořen přidáním třetího nejbližšího objektu. Proces se opakuje, až jsou všechny objekty v jednom společném shluku. Vzdálenost mezi dvěma shluky je definována jako nejkratší vzdálenost libovolného bodu ve shluku vůči libovolnému bodu ve shluku jiném. Dva shluky jsou propojeny v libovolném stadiu nejkratší spojkou. Častou nevýhodou metody nejbližšího souseda je řetězový efekt, kdy se spojují shluky, jejichž dva objekty jsou sice nejbližší, ale vzhledem k většině ostatních objektů nejde o nejbližší shluky.
- *Metoda nejvzdálenějšího souseda*. Jde o metodu podobnou předchozí kromě toho, že kritérium je postaveno nikoliv na minimální, ale na maximální vzdálenosti. Nejdelší vzdálenost mezi objekty v každém shluku představuje nejmenší kouli, která obklopuje všechny objekty v obou shlucích. Metoda se také nazývá metodou úplného propojení, protože všechny objekty ve shluku jsou propojeny každý s každým při maximální vzdálenosti, čili minimální podobnosti. Může se tedy říci, že podobnost uvnitř shluku je rovna průměru shluku. Obě míry vystihují pouze jedno hledisko. Nejkratší vzdálenost postihuje pouze jednoduchý pár nejtěsnějších objektů a nejvzdálenějších postihuje také jediný pár, ale pár dvou extrémů.
- *Metoda průměrné vzdálenosti*. Kritériem vzniku shluků je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku. Takové techniky nezávisí na extrémních hodnotách, jako je tomu u nejbližšího souseda nebo u nejvzdálenějšího

souseda, ale vznik shluku závisí na všech objektech shluku, a ne jenom na jediném páru dvou extrémních objektů.

- *Metoda těžiště.* U této metody jde o vzdálenost dvou těžišť shluků vyjádřených euklidovskou vzdáleností nebo čtvercem euklidovské vzdálenosti. Těžiště shluku má souřadnice odpovídající průměrným hodnotám objektů pro jednotlivé znaky. Po každém kroku shlukování se počítá nové těžiště. Poloha těžiště shluku poněkud migruje tak, jak se připojují nové objekty a vznikají větší shluky. Mohou se objevit také zmateční shluky, když vzdálenost mezi těžišti jednoho páru je menší než vzdálenost mezi těžišti jiného páru utvořeného v předešlém kroku. Výhodou této metody je menší ovlivnění odlehlými body, než je tomu u ostatních hierarchických metod.
- *Mediánová metoda*. Jde o jisté vylepšení metody těžiště, neboť se snaží odstranit rozdílné významnosti, které metoda těžiště dává různě velkým shlukům
- *Wardova metoda.* Principem není optimalizace vzdáleností mezi shluky, ale minimalizace heterogenity shluků podle kritéria minima přírustku vnitroskupinového součtu čtverců odchylek objektů od těžiště shluků. V každém kroku se pro všechny dvojice odchylek spočítá přírůstek součtu čtverců odchylek, vzniklý jejich sloučením a pak se spojí ty shluky, kterým odpovídá minimální hodnota tohoto přírůstku. V případě, že shluk tvoří *k* objektů, které jsou charakterizovány *m* znaky, je k dispozici matice *k* x *m* s prvky x_{ij} *(*hodnota *j*-tého znaku pro *k*-tý objekt. Vnitroshluková variabilita VSS je pak dána vztahem [5]:

$$VSS = \sum_{j=1}^{m} \sum_{i=1}^{k} \left(x_{ij} - \bar{x}_j \right)^2 , \qquad (7)$$

kde $\bar{x}_j = \frac{1}{k} \sum_{i=1}^{k} x_{ij}$. Přidáváním dalších shluků s k_1 objekty se zvětší počet řádků výchozí matice na $k+k_1$ a VSS se počítá pro větší počet objektů. Pokud se začíná od jednoprvkových shluků, bude VSS = 0. Tento postup má tendenci kombinovat shluky s malým počtem objektů.

3.2 Nehierarchické shlukování

Při tomto typu shlukování je vytvářen konkrétní počet shluků. Přiřazení ke shlukům je buď jednoznačné, nebo se počítá míra příslušnosti jednotlivých objektů ke shlukům. Jako příklad postupů, kterými lze získat jednoznačné přiřazení, lze uvést metody s konstantním počtem shluků, kterými jsou Forgyova a Janceyonova metoda, metoda k-průměrů a její modifikace (k-medoidů, k-modů, k-histogramů). Míru příslušnosti ke shlukům je možné zjistit pomocí fuzzy shlukové analýzy.

3.2.1 Forgyova a Janceyonova metoda

Obě metody využívají vzorových bodů, jimiž jsou těžiště shluků. Každý objekt pak patří do shluku, jehož vzorovému bodu je nejblíže. Algoritmy metod jsou založené na střídání dvou kroků a končí v případě, že se najde stabilní rozklad, který se dále nemění. V prvním kroku se přiřadí všechny objekty vzorovému bodu, kterému jsou nejpodobnější. V druhém kroku dojde k přepočítání souřadnic vzorových bodů. Právě ve způsobu výpočtu vzorových bodů se tyto dvě metody liší.

Forgyova metoda bere jako vzorové body těžiště jednotlivých shluků bodů. Janceyova umísťuje vzorové body do bodů souměrně sdružených s předcházejícími typickými body přes souřadnice nového těžiště skupiny.

Obě metody pak začínají zadáním, nebo vytvořením typických bodů, nebo rozkladem na k shluků. Pokračujeme buď výpočtem vzorových bodů, nebo v opačném případě přiřazením objektů do shluků se vzorovými body, k nimž mají nejblíže. Vznikne nový rozklad, ve kterém se znovu vypočtou vzorové body. Tak se postupuje dál, dokud dvě po sobě jdoucí iterace nemají stejné rozložení prvků.

Metody lokálně minimalizují funkcionál součtu čtverců chyb. Nevýhodou je závislost výsledků metod na počáteční volbě vzorových bodů, nebo rozkladu. Jednotlivé výsledky se dle některých zdrojů [8] liší v průměru o 7%. V závislosti na počáteční volbě bodů, může také dojít k situaci, kdy všechny objekty jsou přesunuty jinam. Tento shluk je pak zbytečný, protože je prázdný a výsledek nemusí být optimální.

Janceyova metoda by měla rychleji konvergovat k výsledku.

3.2.2 Metoda k-průměrů

Shlukování metodou k-průměrů [5] se používá v případě, že datový soubor obsahuje pouze kvantitativní proměnné. Jde o iterativní optimalizační metodu, která vychází z počátečního rozdělení objektů do *k* shluků (hodnotu *k* musí zadat analytik). Toto rozdělení je provedeno tak, že je nejprve určeno *k* počátečních centoridů, které mají tvořit „střed“ shluků. Pro stanovení počátečních centroidů existují různé přístupy, může to být například *k* prvních objektů souboru. Poté se postupně zkoumají vzdálenosti každého objektu od každého centroidu tak, že se pro každou takovou dvojici spočte euklidovská vzdálenost. Objekt je přiřazen k nejbližšímu centroidu (zjištěná vzdálenost od tohoto centroidu je menší než vzdálenosti od ostatních centroidů).

Pro každý shluk je spočten nových centroid, který je m-rozměrný vektor průměrných hodnot jednotlivých proměnných. Opět se postupně zkoumají vzdálenosti každého objektu od každého centroidu. V případě, že má objekt blíže k centroidu jiného shluku, je objekt do tohoto shluku přesunut. Celý postup je opakován tak dlouho, dokud dochází k přesunům.

Protože není potřeba pracovat s maticí vzdáleností, je tato metoda vhodná pro datové soubory s velkým počtem objektů. Získá se však pouze lokálně optimální řešení, které je závislé na pořadí objektů v datovém souboru.

3.2.3 Metoda k-medoidů

Pro tuto metodu se používá zkratka PAM (Partitioning Around Medoids) [4]. Stejně jako v předchozím případě je algoritmus určen pro kvantitativní proměnné a vychází z počátečního rozdělení objektů do *k* shluků. Pro každý vytvořený shluk je zajištěn medoid, což je konkrétní objekt ze shluku. Počáteční medoid je určen tak, aby součet vzdáleností jednotlivých objektů ve shluku od tohoto vybraného objektu byl minimální.

Poté se postupně zkoumají všechny objekty. Pokud má zkoumaný objekt nejblíže vlastnímu medoidu, je ponechán v původním shluku, v opačném případě je přemístěn do shluku, k jehož medoidu má nejblíže. To znamená, že objekt x_i je umístěn do shluku C_g, pokud se medoid m_g nachází blíže než kterýkoli jiný medoid m_u, tj. $D(x_i, m_g) \leq D(x_i, m_u)$ pro všechna u=1, 2, …, k.

V dalších iteracích jsou medoidy stanoveny minimalizací funkce, která je součtem vzdáleností jednotlivých objektů od medoidu ve „svém“ shluku. Pokud označíme medoid v *g*-tém shluku, k němuž je přiřazen i-tý objekt, jako $m_{g,i}$, pak jsou medoidy určeny tak, aby bylo

dosaženo minimum funkce $f = \Sigma^{n}_{i=1} D (x_i, m_{g,i})$. Celý postup je opakován tak dlouho, dokud klesá hodnota funkce.

3.2.4 Metoda k-modů a k-histogramů

V literatuře [3], [5], [7] jsou popsány návrhy na využití obdobného postupu, jako je algoritmus k-průměrů pro shlukování objektů charakterizovaných pomocí nominálních proměnných. Vychází se z fakt, že každá i-tá proměnná nabývá hodnot v_{lu} (u= 1, 2, …, K_l). Každý shluk je reprezentován m-rozměrným vektorem, který obsahuje buď modální (nejčetněji zastoupené) kategorie jednotlivých proměnných (v metodě k-modů, nebo údaje o četnostech kategorií jednotlivých proměnných (v metodě k–histogramů). Používají se přitom speciální míry nepodobnosti.

V případě algoritmu k-modů se používá koeficient prosté shody, resp. míra nepodobnosti z něho odvozená. Tehdy je m-rozměrný vektor modálních kategorií speciálním typem centroidu, obdobně jako vektor průměrů či mediánů. Stejně jako v případě shlukování algoritmem k-průměrů však získáváme pouze lokálně optimální řešení, které je závislé na pořadí objektů v datovém souboru.

4 Modelování ekonomických dat vybranými metodami shlukové analýzy

V kapitole bude nejprve popsána metoda Credit Scoring (úvěrový scoring), následně bude navržen model pro klasifikaci žadatelů o úvěr. Tento model bude aplikován na reálných ekonomických datech. Výsledky modelování budou na závěr analyzovány.

4.1 Credit Scoring

Credit Scoring je název metody, která se používá ve finančním odvětví na vyhodnocování žádostí o úvěr. Je založená na historických datech a statistických technikách. Metoda měří relativní stupeň nebezpečí nesplacení úvěru, které představuje dlužník pro věřitele.

Klient při podávání žádosti úvěr poskytuje bance informace, které se týkají jeho osoby a mohou ovlivnit splácení úvěru. Banka na základě těchto informací prostřednictvím Credit Scoringu přiřadí klientovi určitou hodnotu úvěruschopnosti, která vyjadřuje odhadovaný průběh úvěru. Čím vyšší je hodnota, tím vyšší je pravděpodobnost, že splácení úvěru bude bezproblémové. Hodnota je založená na informacích zákazníkovy úvěrové zprávy s přihlédnutím k průběhu úvěru klientů s podobnými vlastnostmi.

V praxi je metoda Credit Scoring aplikovaná pomocí různých metod, které si věřitel může vytvořit sám nebo použije model vyvinutý jinou společností, která se danou problematikou zabývá. U tvorby modelu je třeba si uvědomit skutečnost, že v případě dlužníků, kteří úvěr dostanou ale nesplácí, nastává výrazné poškození věřitele. V opačném případě u žadatelů, kteří úvěr nedostanou a byli by schopni ho splácet, nastává též poškození věřitele, ale ne tak výrazného jako u předcházející situace. Modely jsou navrhnuté všeobecně nebo specificky pro konkrétní typ úvěru.

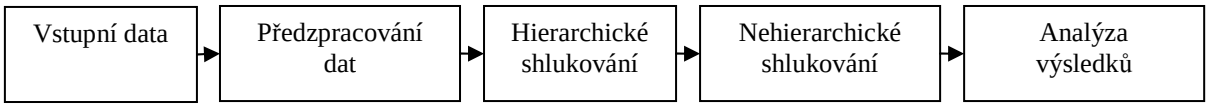
Je důležité si uvědomit, že Credit Scoring model nepoví věřitelovi s určitostí, jaký bude průběh úvěru, ale dobře navržený model je schopný rozdělit klienty na více a méně rizikové skupiny. Každý věřitel si zvolí výšku míry rizika, kterou už není ochotný podstoupit.

4.2 Postup shlukové analýzy

Shluková analýza poskytuje uživateli empirické a objektivní metody k provádění jedné z nejzákladnějších činností člověka – klasifikaci. Analýza shluků je vždy silnou analytickou pomůckou k účelům zjednodušení, průzkumu a potvrzení, která má širokou oblast použití. Když se analýzy shluků užije správně, může odhalit strukturu v datech, kterou by jinak nešlo nalézt.

4.3 Definování problému a návrh modelu

Základním cílem analýzy shluků je rozdělení použitých dat do dvou nebo více skupin, tříd či shluků, založených na podobnosti objektů. Řešený problém spočívá v analýze žadatelů o úvěr. K dispozici je datová sada, která již byla v minulosti realizována pomocí metod učení s učitelem [5]. Jedná se o data z německého bankovního institutu [8]. U daného příkladu bude použit model, který je uveden na obrázku 4. V první řadě budou data analyzována z hlediska potencionálních odlehlých objektů. Objekty identifikované jako odlehlé budou vyřazeny. Poté bude provedena korelační analýza, z důvodu zjištění závislosti mezi jednotlivými znaky. Následovat bude hierarchické a nehierarchické shlukování. Hierarchické shlukování bude provedeno Wardovou metodou. Pro výpočet vzdáleností mezi objekty bude použita euklidovská metrika. Z výsledku hierarchického shlukování bude zjištěno, jaká centra shluků a jaký počet shluků bude nastaven pro další krok, čímž je nehierarchické shlukování. Pomocí nehierarchického shlukování bude zjištěno, jaké konkrétní případy budou zařazeny do konkrétního shluku. Vyhodnocení bude spočívat v porovnání výsledku s třídami přiřazenými objektům bankou tj. 1 pro úvěruschopné a 2 pro neúvěruschopné klienty. Na závěr budou dosažené výsledky analyzovány.



Obrázek 4 - Postup modelování ekonomických dat, zdroj [vlastní]

4.4 Předpoklady shlukové analýzy

Analýza shluků není charakteru statistického testování, kdy jsou parametry výběru odhadovány jako představitelé celého souboru. Místo toho představuje analýza shluků metodu ke kvantifikaci strukturních vlastností souboru. Požadavky normality a linearity, které jsou

tolik důležité v ostatních vícerozměrných technikách, nemají ve shlukové analýze tak velký význam. Přesto existují dva důležité předpoklady: reprezentativnost dat a vliv multikolinearity.

4.5 Modelování úvěrové schopnosti žadatelů – Německo

Data použitá v této bakalářské práci se skládají z 1000 zájemců o spotřebitelský úvěr z Jihoněmecké banky a pocházejí ze skupinového výběru, což znamená, že 300 klientů bylo záměrně vybráno ze skupiny nedůvěryhodných a 700 ze skupiny důvěryhodných. Obsahují 1 výstupní proměnnou Y a 20 vstupních proměnných $X_1, \ldots . X_{20}$, u kterých se předpokládá, že úvěruschopnost klienta ovlivňují.

4.5.1 Vstupní data

Typ, struktura a popis jednotlivých dat je uveden níže:

- **Y** (alternativní proměnná – nabývá pouze dvou hodnot): Výstupní proměnná, která předvídá, či klient bude schopný splácet poskytnutý úvěr při daných podmínkách. Mohou nastat dvě možnosti, které jsou zachycené v následující tabulce.

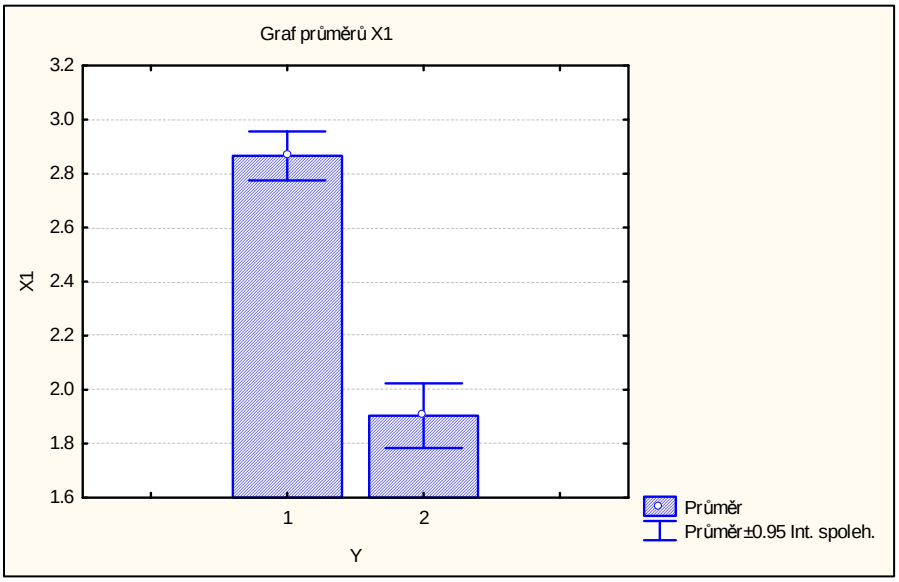
Tabulka 1- Vyhodnocení klienta

Y (úvěryschopnost)	**Hodnota**
Důvěryhodný	1
Nedůvěryhodný	2

- **X_1** (ordinální): Proměnná X_1 vyjadřuje, zda klient má v dané bance založený běžný účet a výši jeho zůstatku. Z grafu 2 vyplývá, že klienti, kterým byl úvěr poskytnut mají založený běžný účet se zůstatkem kolem 200 Euro.

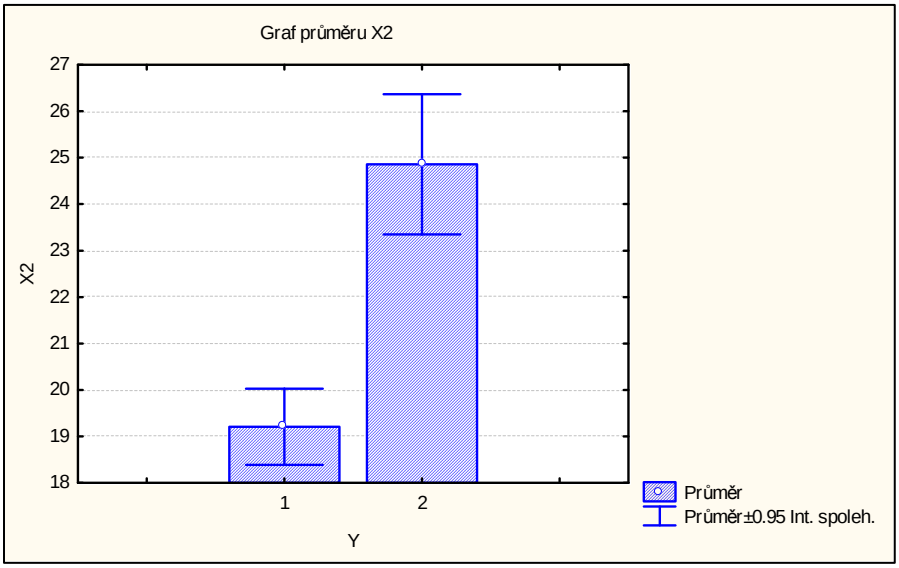
Tabulka 2 - Běžný účet

X_1 (běžný účet)	**Hodnota**
Bez běžného účtu	1
Účet v debetu	2
Do 200 Euro	3
Nad 200 Euro	4



Graf 2 - Běžný účet

- **X_2** (kardinální): Když věřitel poskytne klientovi úvěr, zajímá se za jakou dobu mu bude tato hotovost i s danými úroky vrácena. Z grafu 3 je patrné, že u klientů z první skupiny je tato doba kratší (20 měsíců), než u klientů, kterým nebyl úvěr poskytnut (25 měsíců).

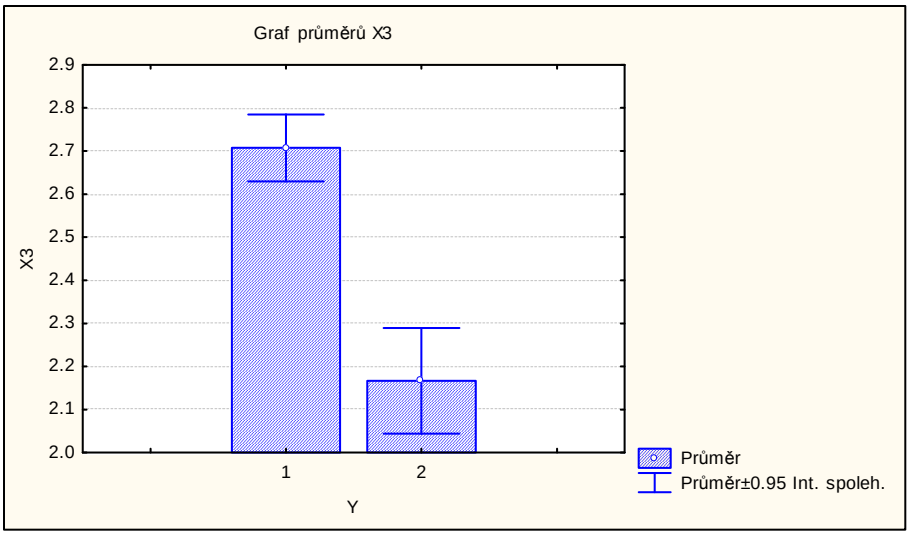


Graf 3 - Splatnost úvěru v měsících

- **X_3** (ordinální): při vyhodnocování žádostí o úvěr banku zajímá historie dat a chování klienta. Zajímá se, zda klient už dostal úvěr, jak ho splácel, zdali u splátek nastali nějaké problémy, atd. X_3 tedy zaznamenává chování určitého klienta v podobných situacích. Klienti z první skupiny patří mezi ty, kteří mají všechny úvěry splacené, tudíž jsou pro banku bezproblémovými.

Tabulka 3 - Chování klienta v případě ostatních úvěrů

X_3 (předcházející úvěry)	Hodnota
Problémové platby předcházejícího úvěru	0
Další současné úvěry u jiné banky / problémový běžný účet	1
Bez předcházejících úvěrů / splacené všechny předcházející úvěry	2
Bezproblémový současný úvěr v této bance	3
Splacené předcházející úvěry v této bance	4

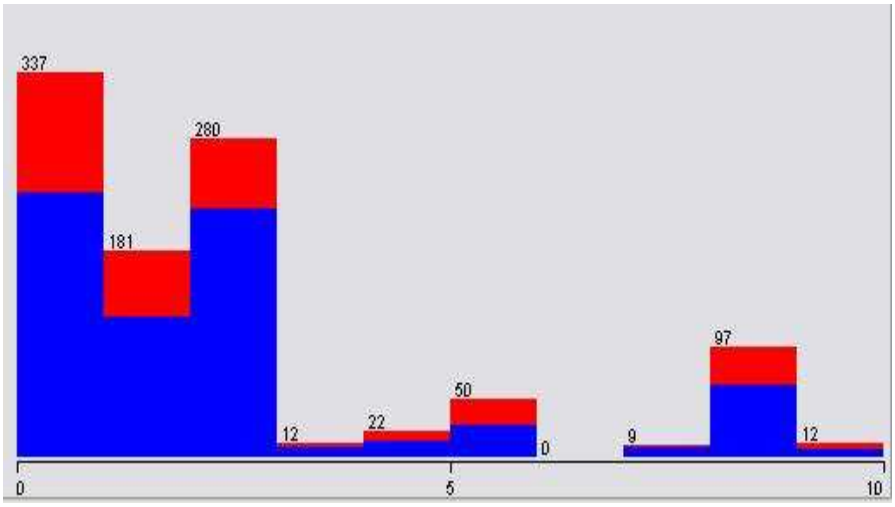


Graf 4 - Chování klienta v případě ostatních úvěrů

- X_4 (nominální): Použití finanční hotovosti, o kterou klient žádá je zachyceno v tabulce 4. Většina klientů, kteří úspěšně žádali o úvěr, chtěli hotovost použít na podnikání, dále koupi automobilu. I přes skutečnost, že znak je nominální, hodnoty byly seřazeny od nejvyšších částek po nejnižší tak, aby se mohly dál použít ve shlukové analýze.

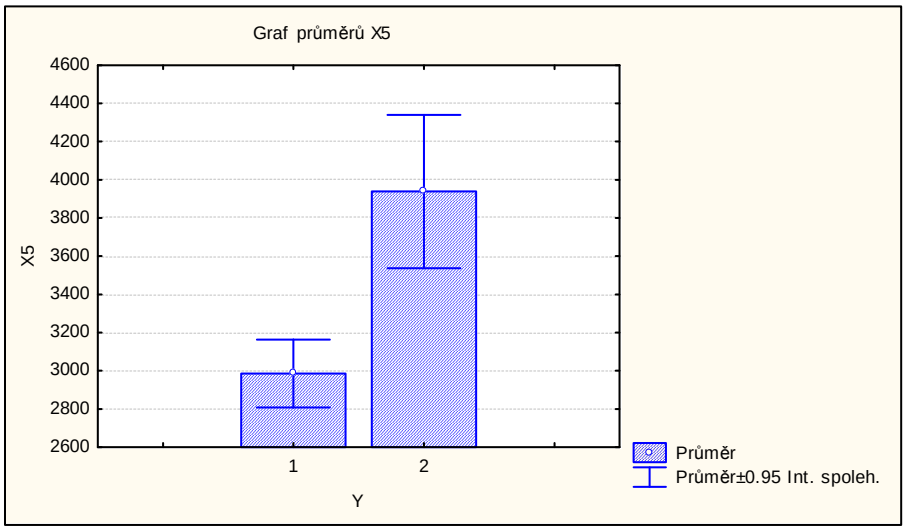
Tabulka 4 - Účel úvěru

X_4 (použití úvěru)	Hodnota
Podnikání	1
Nové auto	2
Použité auto	3
Nábytek	4
Spotřebiče	5
Opravy	6
Vzdělání	7
Dovolená	8
Přeškolení	9
Televize, radio	10



Graf 5 - Účel úvěru

- **X_5** (kardinální): Jednou z nejvýznamnějších veličin je výše požadovaného úvěru. Z této proměnné se určuje výše měsíční splátky a schopnost klienta jednotlivé částky splácet. Z grafu 6 je patrné, že klienti, kterým byl úvěr poskytnut žádali většinou o částky kolem 3000 Euro, zatímco klienti, kterým nebyl úvěr poskytnut žádali o částku kolem 4000 Euro.

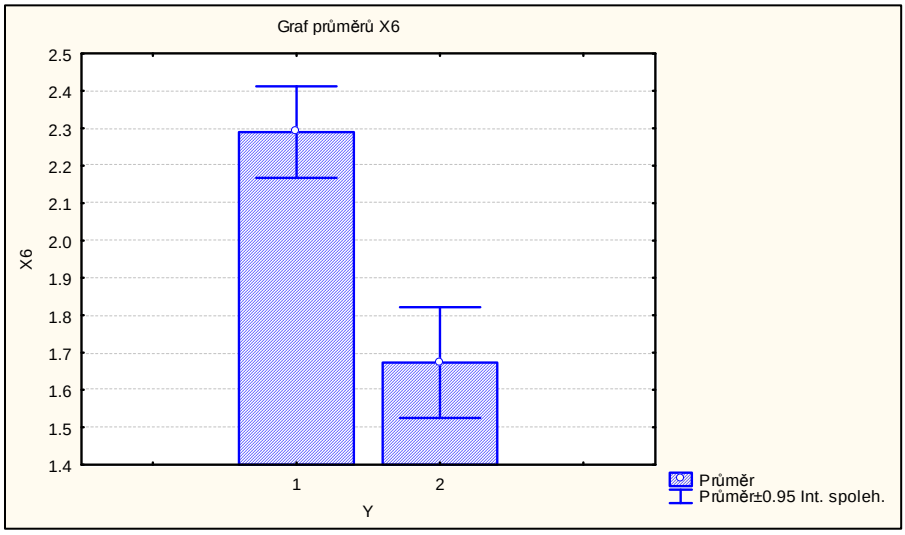


Graf 6 - Výše úvěru

- **X_6** (ordinální): Při splácení úvěru se věřitel zajímá o včasné splacení jednotlivých splátek. Když má dlužník k dispozici větší finanční hotovost, předpokládá se, že u jeho splátek nenastanou problémy. Klienti z první skupiny měli dostupnou finanční hotovost v průměru kolem 500 Euro, oproti klientům z druhé skupiny, kteří disponovali v průměru pouze do 100 Euro.

Tabulka 5 - Hodnota úspor a cenných papírů

X_6 (dostupná finanční hotovost)	Hodnota
Bez úspor / nedostupné úspory	1
Do 100 Euro	2
101 až 500 Euro	3
501 až 1000 Euro	4
1000 a více Euro	5

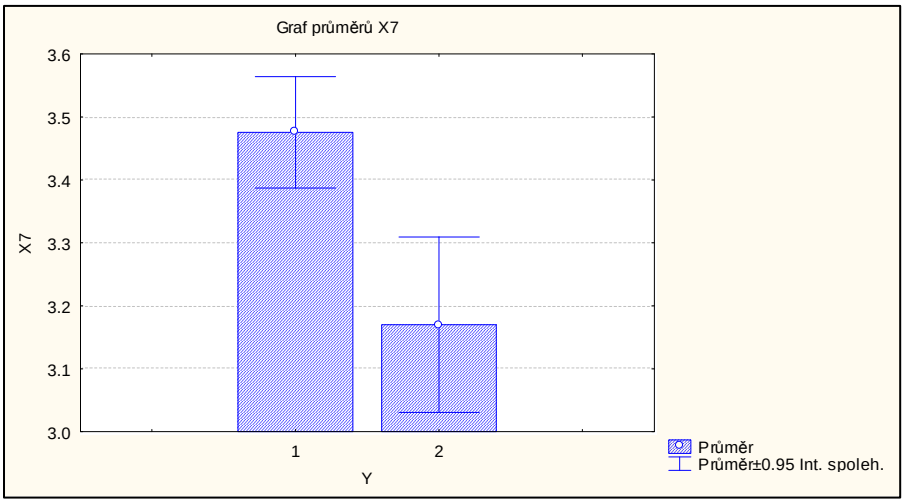


Graf 7 - Hodnota úspor a cenných papírů

- X_7 (ordinální): Jak dlouho klient pracuje na současném místě zachycuje tabulka 6. Klienti z první skupiny jsou zaměstnaní většinou 5 a více let, oproti klientům z druhé skupiny, kteří jsou zaměstnání nejvíce 4 roky.

Tabulka 6 - Doba zaměstnání v rocích

X_7 (Délka zaměstnání)	Hodnota
Nezaměstnaný	1
Do 1	2
1 až 4	3
4 až 7	4
7 a více	5

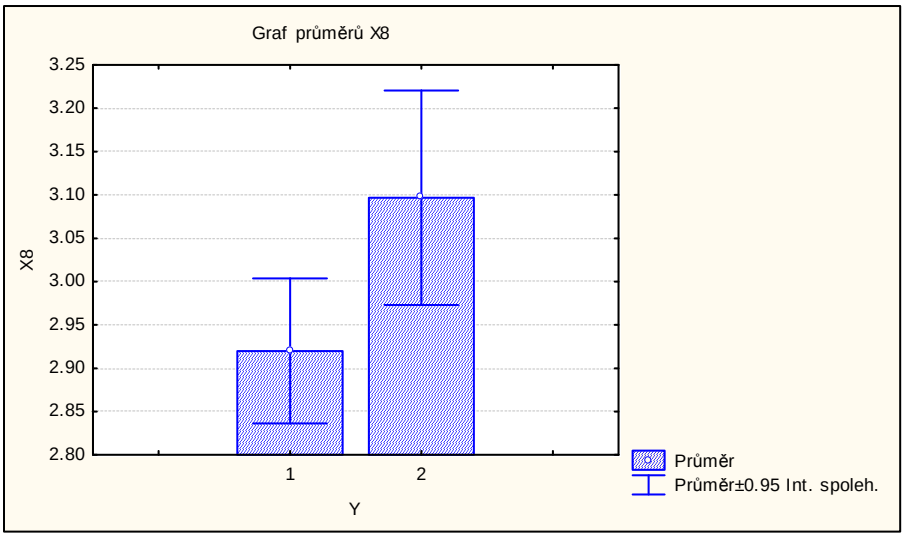


Graf 8 - Doba zaměstnání v rocích

- **X_8** (kardinální): Dalším faktorem, který ovlivňuje výstupní veličinu je velikost částky z příjmu vynaložené na splátku úvěru. Čím větší část tvoří splátka, tím víc jednotlivec pociťuje tíhu tohoto břemene. Z grafu 9 je zřejmé, že klienti, kteří nesplnili podmínky pro schválení úvěru tvořila splátka 35 % a více procent z příjmu, zatímco u klientů, kterým byl úvěr poskytnut tvořila splátka v průměru kolem 20% z příjmu.

Tabulka 7 - Výška splátky v % z dostupného příjmu

X_8 (% splátky z příjmu)	Hodnota
Do 20	1
21 až 25	2
26 až 35	3
36 a více	4

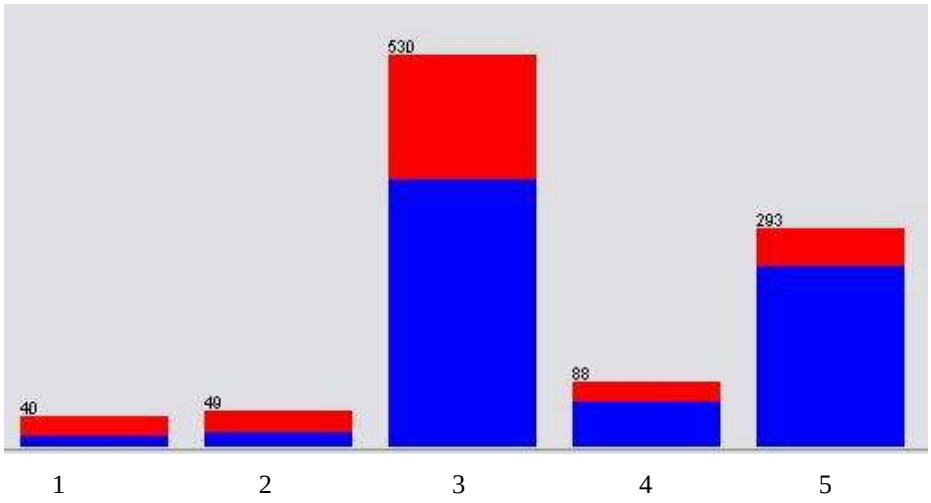


Graf 9 - Výška splátky v % z dostupného příjmu

- **X_9** (nominální): Na průběh úvěru může působit pohlaví klienta a jeho rodinné zázemí. Dle výsledků z grafu 10 vyplývá, že klienti, kteří nejvíce žádali o úvěr, byli většinou ženatí muži nebo vdané ženy. I přes skutečnost, že znak je nominální, hodnoty byly seřazeny podle spolehlivosti od nejnižší po nejvyšší tak, aby se daly využít při shlukové analýze.

Tabulka 8 - Stav / pohlaví

X_9 (Manželský stav / pohlaví)	**Hodnota**
Muž - rozvedený / žijící odděleně od manželky	1
Muž - svobodný	2
Muž – ženatý / vdovec	3
Žena – rozvedená / žijící samostatně	4
Žena - vdaná	5

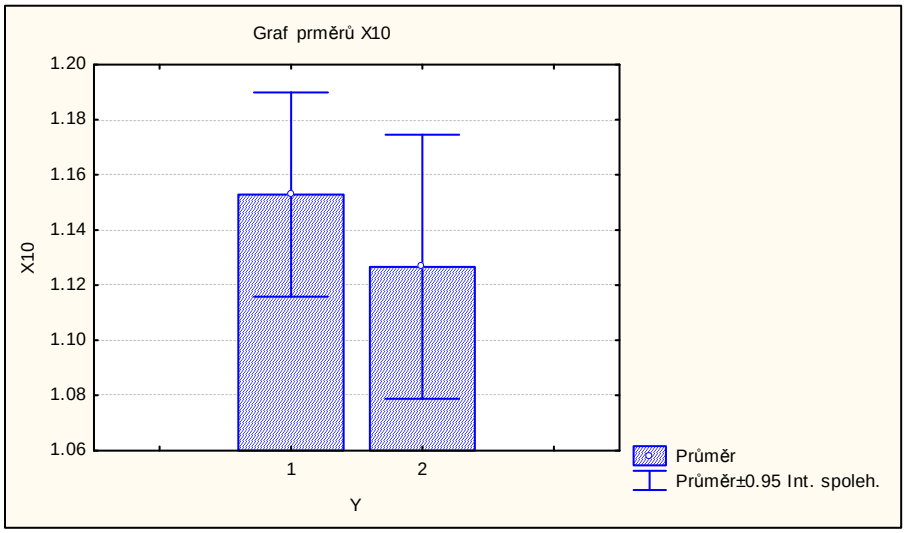


Graf 10 - Stav / pohlaví

- **X_{10}** (ordinální): Když klient má další závazky nebo je ručitelem na určitou finanční hotovost, část jeho příjmu je určená na splátku možných výdajů. Tzn. sníží se peněžní obnos, který je klient schopný vynaložit na splácení úvěru. Z grafu 11 vyplývá, že tato proměnná nemá rozhodující vliv při určování, zdali má klient na úvěr nárok či nikoliv.

Tabulka 9 - Další závazky / ručení

X_{10} (závazky / ručení)	**Hodnota**
Žádné	1
Spoluručitel	2
Ručitel	3

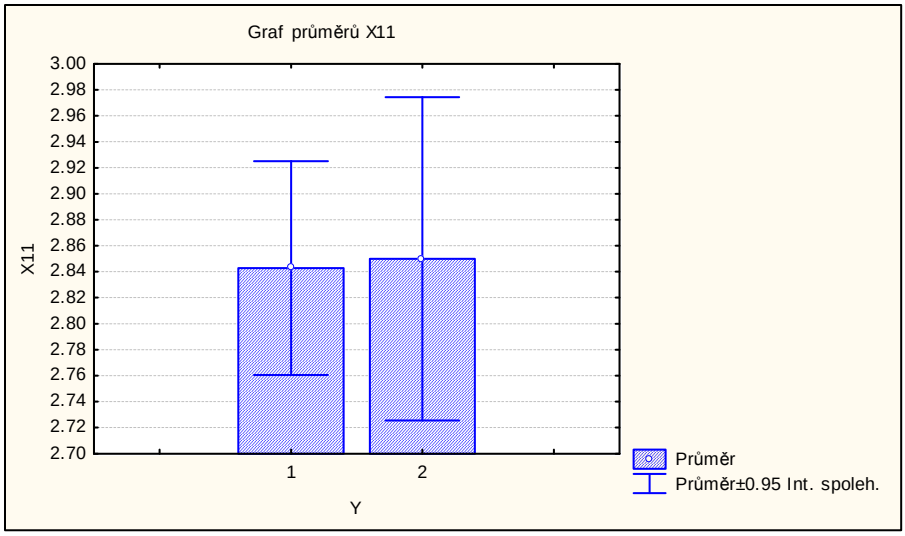


Graf 11 - Další závazky / ručení

- **X_{11}** (ordinální):Lidé, kteří často mění místo bydliště mají tendenci hůř splácet úvěry. Jak je zřejmé v grafu 12, nemá tento faktor vliv na rozhodnutí banky, zdali poskytne úvěr. Hodnoty v obou skupinách jsou téměř vyrovnané.

Tabulka 10 - Doba bydlení na současném bydlišti

X_{11} (Doba bydlení)	**Hodnota**
Do 1 rok	1
1 až 4 roky	2
4 až 7 let	3
7 a více	4



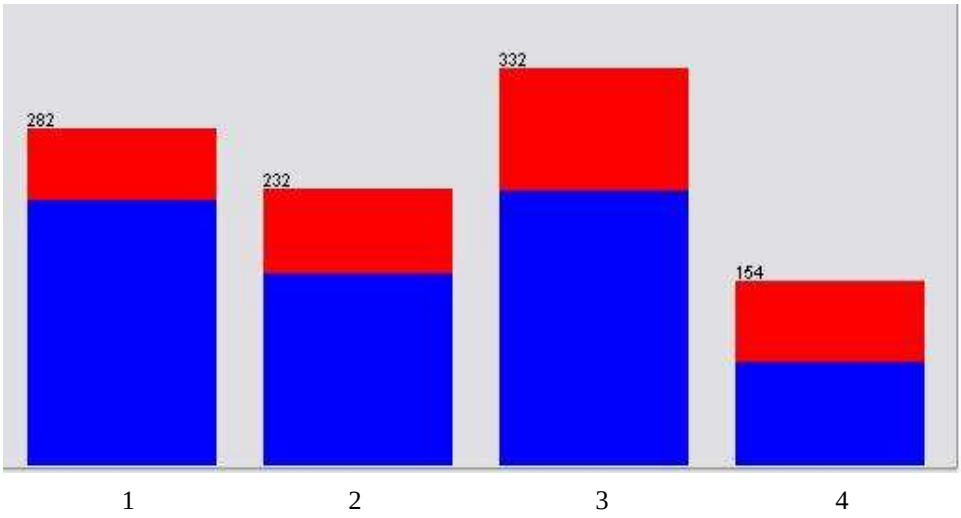
Graf 12 - Doba bydlení na současném bydlišti v rocích

- **X_{12}** (ordinální): Možnost ručení určitým aktivem zobrazuje tabulka 11. Schopnost ručení je proměnná, která má vysokou váhu při rozhodovaní, zdali banka úvěr schválí.

Z grafu 13 je viditelné, že není rozhodující, čím klient ručí, protože jsou hodnoty takřka vyrovnané, pouze ve čtvrtém případě, který označuje skutečnost, že klient neručí ničím, je důvěryhodnost banky menší.

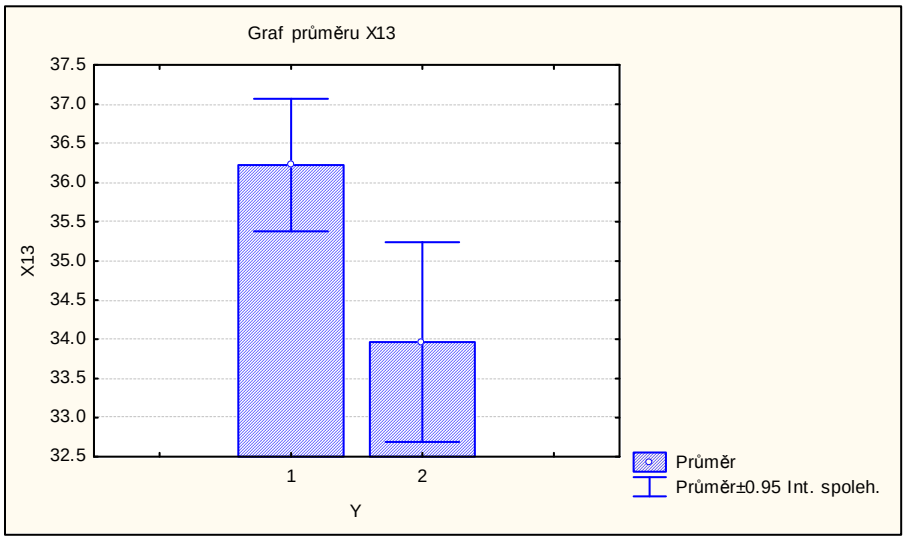
Tabulka 11 - Dostupná aktiva

X_{12} (Dostupné aktivum)	Hodnota
Vlastní dům / pozemek	1
Stavební spoření	2
Auto / ostatní	3
Bez vlastních aktiv	4



Graf 13 - Dostupná aktiva

- **X_{13}** (kardinální): Banka se zajímá o věk z hlediska doby splácení úvěru. Je důležité, zda se klient nachází v předproduktivním, produktivním, nebo neproduktivním období svého života. Všichni žadatelé jsou v produktivním věku, proto tento faktor nemá velký vliv na rozhodnutí banky, což je vyobrazené v grafu 14..

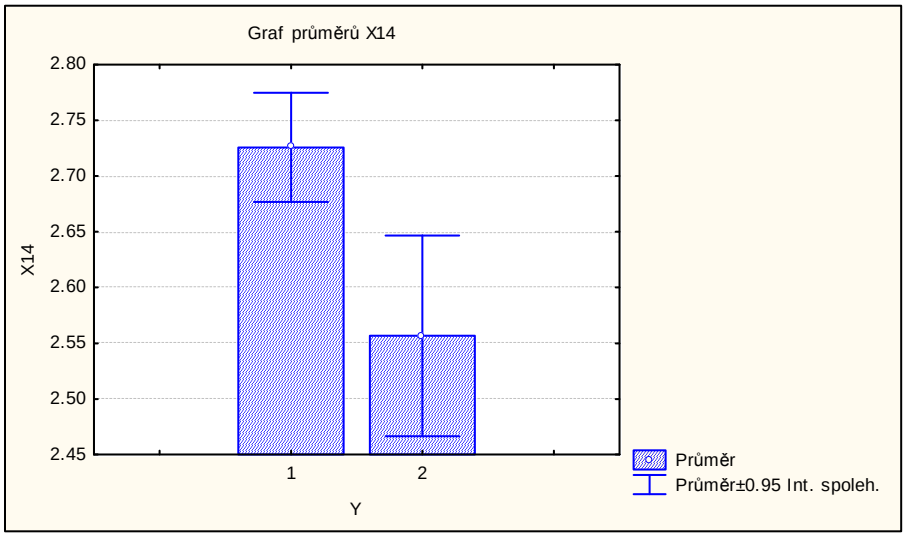


Graf 14 - Věk klientů

- **X_{14}** (ordinální): Následující tabulka zaznamenává, zda daný klient v současnosti splácí nějaký jiný úvěr. Pokud ano, tak v jaké instituci. Z grafu 15 je patrné, že klienti, kteří splnili kritéria banky nemají jiný úvěr u dané banky ani v jiných institucích. Klientům, kteří mají úvěr u dané banky nebo v jiných institucích se výrazně snižuje možnost poskytnutí úvěru z důvodu finančního zatížení splácením již poskytnutého úvěru.

Tabulka 12 - Další úvěry

X_{14} (Současné úvěry)	Hodnota
U ostatních finančních institucí	1
U dané banky	2
Bez jiných úvěrů	3

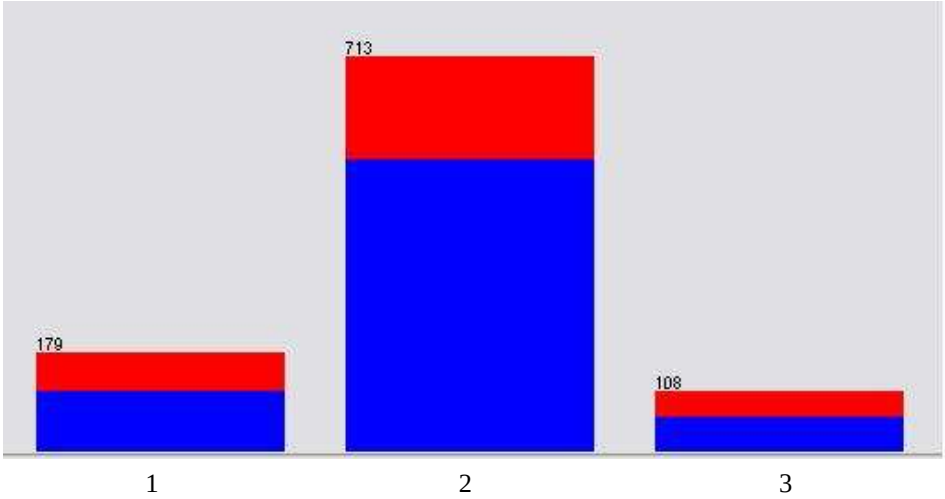


Graf 15 - Další úvěry

- $\mathbf{X_{15}}$ (ordinální): Způsob bydlení vypovídá o klientově zázemí a možnosti ručit daným aktivem. Dle grafu 16 jsou klienti z první i druhé skupiny převážně ubytovaní v pronajatém domě nebo bytě.

Tabulka 13 - Způsob bydlení

$\mathbf{X_{15}}$ (Bydlení)	**Hodnota**
Vlastní byt / dům	1
Pronajatý byt / dům	2
Ostatní	3

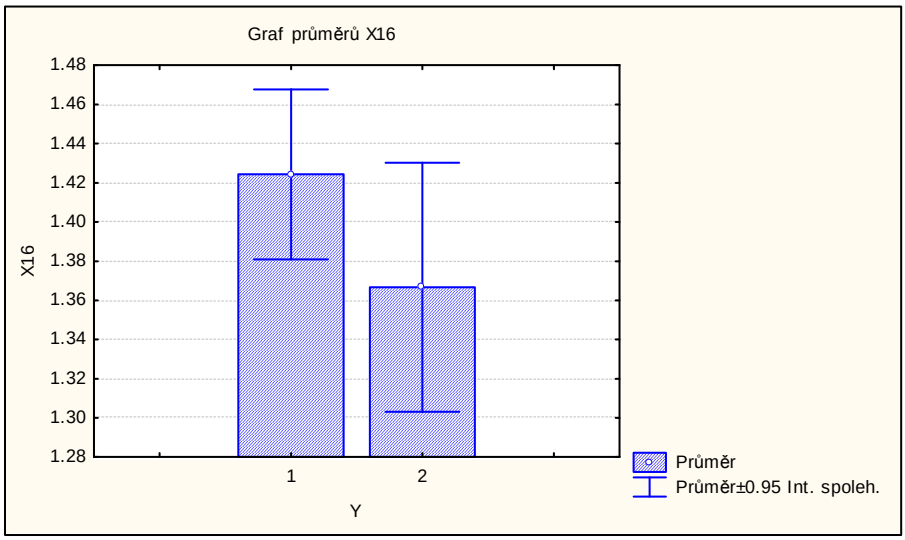


Graf 16 - Způsob bydlení

- $\mathbf{X_{16}}$ (ordinální): Žadatel, který už klientem banky v minulosti byl a v průběhu úvěru nenastal žádný problém má větší šanci na znovuzískání dalšího úvěru. Vyšší množství dobře splacených závazků k dané bance zvyšuje důvěryhodnost klienta. Tuto skutečnost ukazuje graf 17. V první skupině jsou obsažení klienti, kteří měli v minulosti u této banky bez problému splacených 2 a více úvěrů.

Tabulka 14 - Úvěry v této bance

$\mathbf{X_{16}}$ (úvěry v dané bance)	**Hodnota**
1	1
2 až 3	2
4 až 5	3
6 a více	4

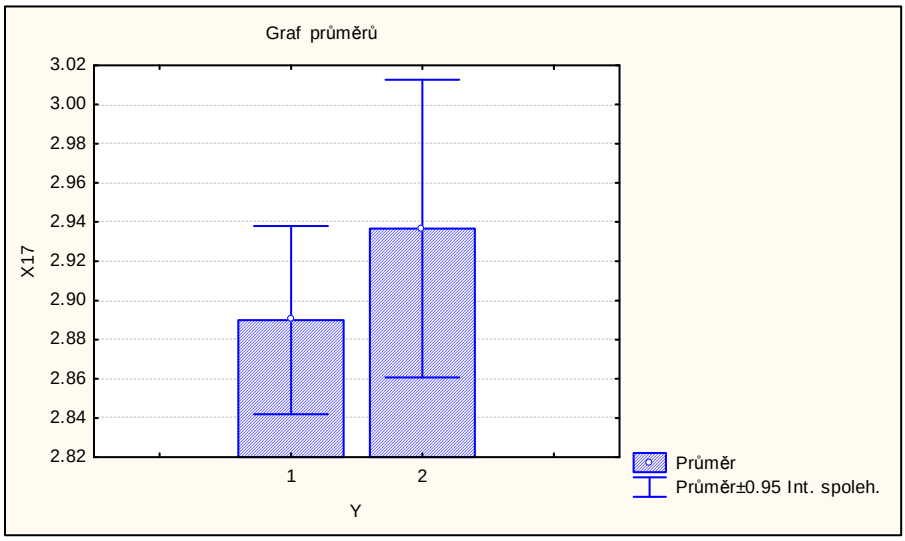


Graf 17 - Úvěry v dané bance

- **X_{17}** (ordinální): Nepřímo zaznamenává sociální postavení ve společnosti. Z grafu 18 vyplývá, že tato proměnná není důležitým rozhodujícím faktorem pro banku.

Tabulka 15 - Zaměstnání

X_{17} (Zaměstnání)	**Hodnota**
Nezaměstnaný	1
Bez vyučení	2
Vyučený dělník / zaměstnanec / státní úředník	3
Vedoucí zaměstnanec / podnikatel / vyšší státní úředník	4

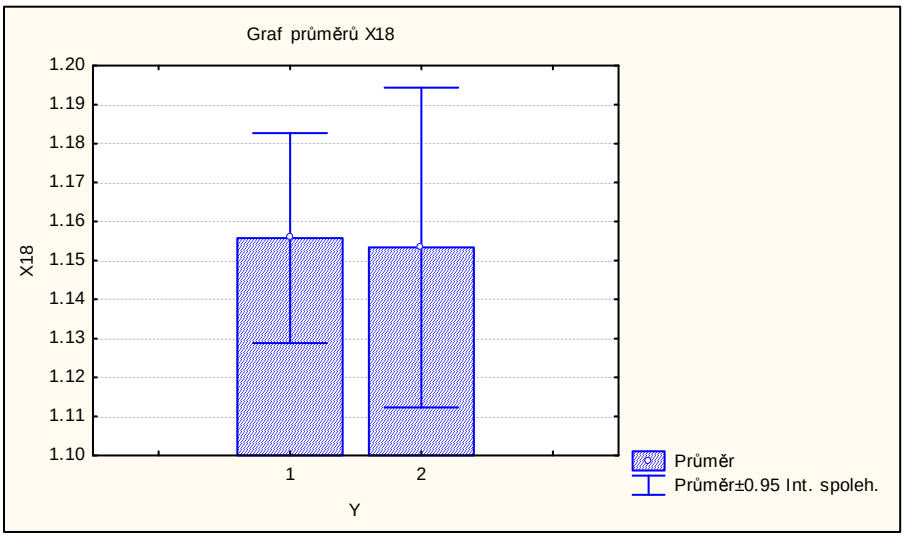


Graf 18 – Zaměstnání

- $\mathbf{X_{18}}$ (alternativní): Zohledňuje množství lidí dělících se o příjem domácnosti.

Tabulka 16 - Počet členů v domácnosti

$\mathbf{X_{18}}$ (Členů domácnosti)	**Hodnota**
0 až 2	1
3 a více	2

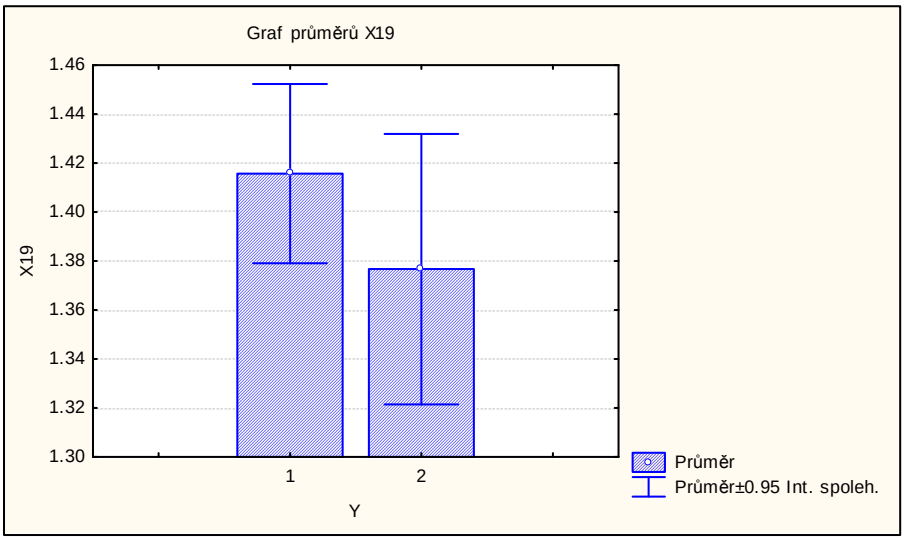


Graf 19 - Počet členů v domácnosti

- $\mathbf{X_{19}}$ (alternativní): V grafu 20 je zobrazen počet klientů, kteří vlastní telefon, což zvyšuje dostupnost klienta a též může znamenat větší movitost.

Tabulka 17 - Vlastnictví telefonu

$\mathbf{X_{19}}$ (telefon)	**Hodnota**
Ne	1
Ano	2

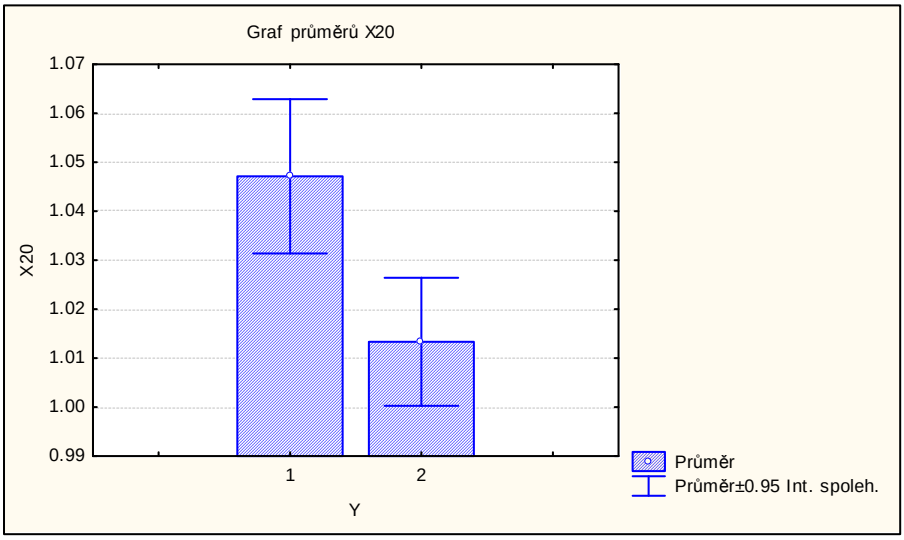


Graf 20 - Vlastnictví telefonu

- $\mathbf{X_{20}}$ (alternativní): U osoby cizí národnosti může nastat složitější vymáhání pohledávek banky. Z grafu 21 vyplývá, že klienti, kteří žádali o úvěr nejsou cizí národnosti.

Tabulka 18 - Pracující cizinec

X_{20} (cizinec)	Hodnota
Ne	1
Ano	2

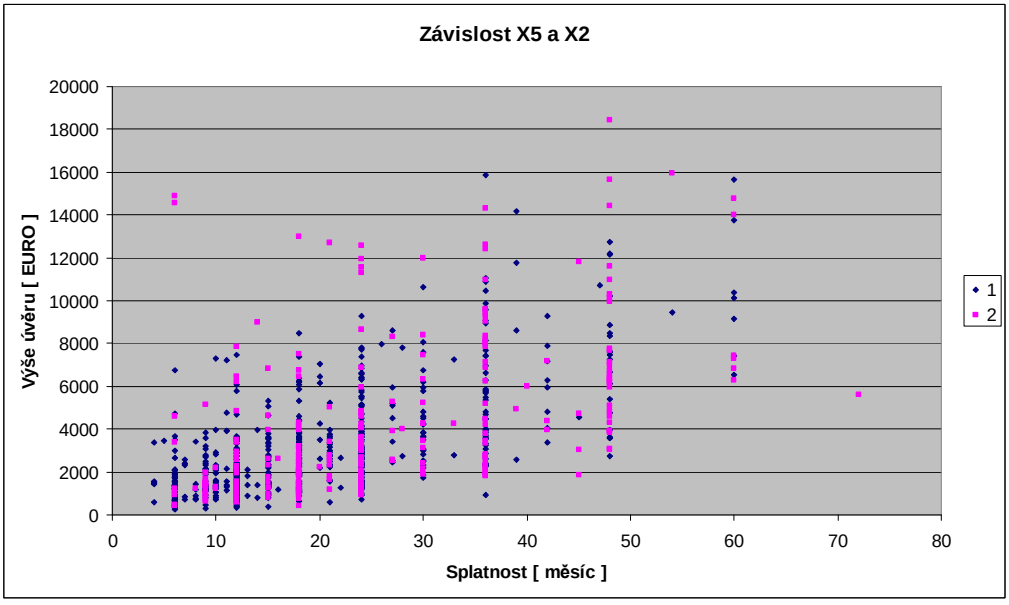


Graf 21 - Pracující cizinec

4.5.2 Analýza závislostí

V této kapitole budou vybrány znaky, které nejvíce ovlivňují, zda úvěr klientovi bude poskytnut či nikoliv. Pro lepší viditelnost ovlivnění budou závislosti vyobrazeny v grafech.

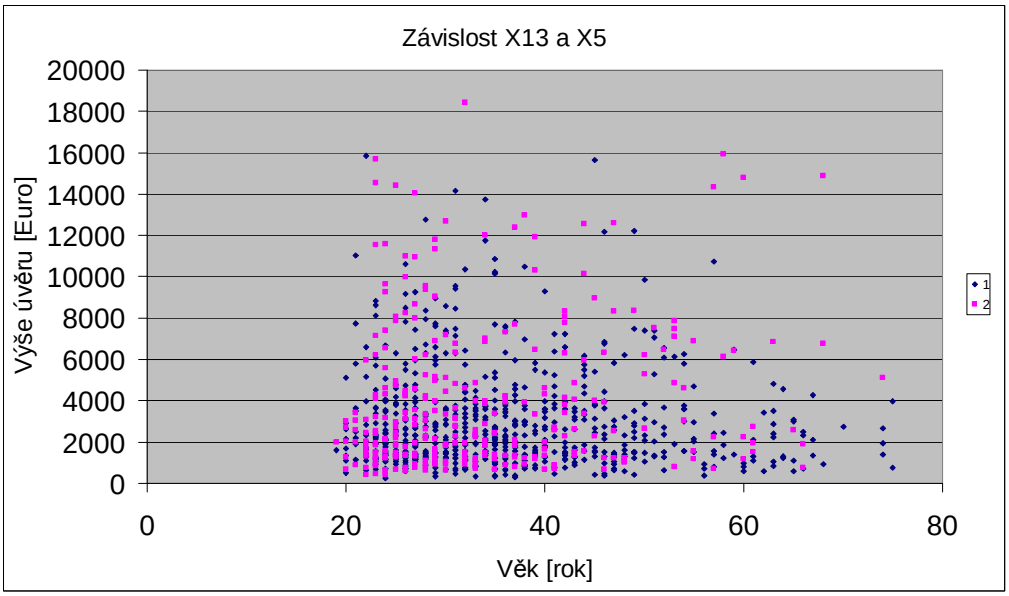
Jako první příklad byl zkoumán znak X2 (splatnost) spolu se znakem X5 (výše požadovaného úvěru). Výsledek je zobrazen v grafu 22, kde je jasně viditelné, že nejvíce schválených úvěrů se pohybuje do 5000 Euro při splatnosti do 25 měsíců.



Graf 22 - Závislost X2, X5

Jako další příklad byla vybrána závislost mezi X5 (výše věru) a X13 (věk). Předpoklad je takový, že mladší člověk nebude mít nárok na úvěr s vyšší částkou. Vzhledem k věku kolem 20 let, se předpokládá, že klient ještě nebude mít schopnost ručit vyšším aktivem, než je automobil. Oproti klientům, kteří jsou ve věku mezi 40 a 60 let, mají již např. vlastní dům nebo byt.

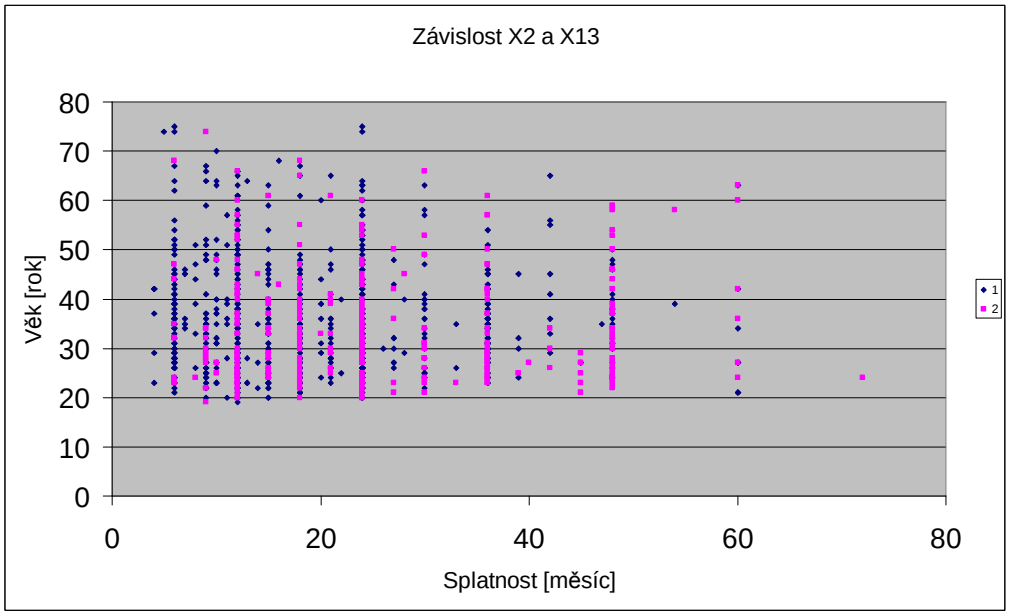
Dle grafu 23 je zřejmé, že předpoklad byl správný. Kolem hranice 20 let, je mnoho úvěrů, které nebyly schváleny, zatímco postupem věku se objevují klienti, kterým byl úvěr povolen. Z grafu 23 je též patrné, že nejvyšší úvěry banka schvaluje klientům ve věku kolem 40 let. Je to způsobené tím, že v tomto věku má klient většinou stabilní zaměstnání a možnost ručení aktivem o vyšší hodnotě.



Graf 23 - Závislost X13, X5

Dalším příkladem porovnání je závislost mezi X2 (splatnost) a X13 (věk). Předpoklad je takový, že mladší klienti nemají nižší příjem než starší klienti, a proto budou muset mladší klienti déle splácet úvěr, což je nevýhodné pro banku.

Z grafu 24 vyplývá, že předpoklad byl správný. Většina klientů, kteří jsou schopni úvěr splatit do 20 měsíců jsou ve věku mezi 35 a 55 let.



Graf 24 - Závislost X2, X13

Popisná statistika

Vzhledem k velice rozsáhlému základnímu souboru dat, byla použita popisná statistika, díky níž se získá přehlednější a úspornější popis dat.

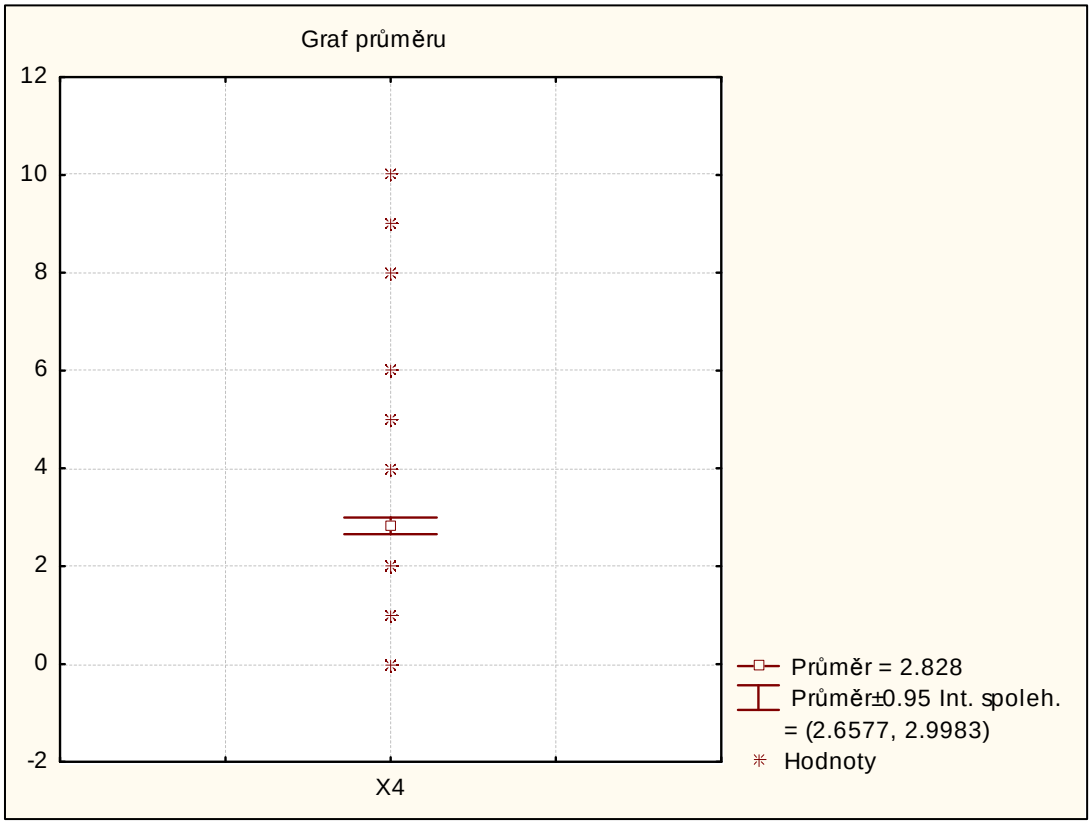
Tabulka 19 - Popisná statistika

Proměnná	Popisné statistiky											
	Průměr	Int. spolehl. -95.000%	Int. spolehl. +95.000%	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Rozptyl	Sm. odch.	Šikmost	Špičatost
X1	2.577	2.499	2.655	2.000	1.0000	4.00	1.000	4.000	2	1.258	0.00696	-1.66370
X2	20.903	20.155	21.651	18.000	4.0000	72.00	12.000	24.000	145	12.059	1.09418	0.91978
X3	2.545	2.478	2.612	2.000	0.0000	4.00	2.000	4.000	1	1.083	-0.01189	-0.57906
X4	2.828	2.658	2.998	2.000	0.0000	10.00	1.000	3.000	8	2.744	1.17889	0.55408
X5	3271.258	3096.094	3446.422	2319.500	250.0000	18424.00	1365.000	3972.500	7967843	2822.737	1.94963	4.29259
X6	2.105	2.007	2.203	1.000	1.0000	5.00	1.000	3.000	2	1.580	1.01668	-0.68022
X7	3.384	3.309	3.459	3.000	1.0000	5.00	3.000	5.000	1	1.208	-0.11761	-0.93433
X8	2.973	2.904	3.042	3.000	1.0000	4.00	2.000	4.000	1	1.119	-0.53135	-1.21047
X9	2.682	2.638	2.726	3.000	1.0000	4.00	2.000	3.000	1	0.708	-0.30515	-0.00257
X10	1.145	1.115	1.175	1.000	1.0000	3.00	1.000	1.000	0	0.478	3.26425	9.32876
X11	2.845	2.777	2.913	3.000	1.0000	4.00	2.000	4.000	1	1.104	-0.27257	-1.38145
X12	2.358	2.293	2.423	2.000	1.0000	4.00	1.000	3.000	1	1.050	0.04567	-1.23852
X13	35.546	34.840	36.252	33.000	19.0000	75.00	27.000	42.000	129	11.375	1.02074	0.59578
X14	2.675	2.631	2.719	3.000	1.0000	3.00	3.000	3.000	0	0.706	-1.82652	1.51259
X15	1.929	1.896	1.962	2.000	1.0000	3.00	2.000	2.000	0	0.531	-0.07080	0.47298
X16	1.407	1.371	1.443	1.000	1.0000	4.00	1.000	2.000	0	0.578	1.27258	1.60444
X17	2.904	2.863	2.945	3.000	1.0000	4.00	3.000	3.000	0	0.654	-0.37429	0.50189
X18	1.155	1.133	1.177	1.000	1.0000	2.00	1.000	1.000	0	0.362	1.90944	1.64927
X19	1.404	1.374	1.434	1.000	1.0000	2.00	1.000	2.000	0	0.491	0.39187	-1.85014
X20	1.037	1.025	1.049	1.000	1.0000	2.00	1.000	1.000	0	0.189	4.91303	22.18220

4.5.3 Předzpracování dat

Identifikace odlehlých objektů

Vzhledem k citlivosti shlukové analýzy na přítomnost odlehlých hodnot, byla použita v programu STATISTICA 7 funkce grafu odlehlých hodnot. Vzhledem ke skutečnosti, že všech 17 proměnných z 20 jsou celočíselná ordinální data s rozsahem nevýše 10, bylo hledání odlehlých hodnot mnohem jednodušší, než u hodnot číselných, kde byl nastaven koeficient, který vyjadřoval násobek překročení průměru v dané proměnné. Analyzován byl celý datový soubor s negativním výsledkem. Jako příklad byl použit graf proměnné X4, z kterého je patrné, že neobsahuje žádnou odlehlou hodnotu.



Graf 25 – Identifikace odlehlých objektů

Korelační koeficienty

Pro zjištění závislosti a podobnosti objektů byla použita korelační tabulka 20 vytvořena pomocí programového prostředí STATISTICA 7 podle vzorce 2 na straně 19. Je známo, že objekty jsou si tím podobnější, čím je jejich párový koeficient větší a bližší jedné. Jako hranice ovlivnitelnosti výsledků shlukování byla nastavena velikost koeficientu 0.5. Po přezkoumání hodnot koeficientů byla zjištěna silná korelace mezi znaky X2 (doba splatnosti) a X5 (výše úvěru). Vzhledem ke skutečnosti, že faktory které spolu silně korelují mohou zkreslovat samotné shlukování byl znak X2 odstraněn a bylo počítáno se s tím, že dojde k lepšímu semknutí shluků a znak X5, který byl v datovém souboru ponechán, bude reprezentovat i vyloučený znak X2.

Tabulka 20 - Korelační koeficienty

	Korelace N=1000																			
Proměnná	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
X1	1.00	-0.07	0.19	0.03	-0.04	0.22	0.11	-0.01	0.04	-0.13	-0.04	-0.03	0.06	0.05	0.02	0.08	0.04	-0.01	0.07	-0.03
X2	-0.07	1.00	-0.08	0.15	**0.62**	0.05	0.06	0.07	0.01	-0.02	0.03	0.30	-0.04	-0.05	0.16	-0.01	0.21	-0.02	0.16	-0.14
X3	0.19	-0.08	1.00	-0.09	-0.06	0.04	0.14	0.04	0.04	-0.04	0.06	-0.05	0.15	0.12	0.06	0.44	0.01	0.01	0.05	0.01
X4	0.03	0.15	-0.09	1.00	0.07	-0.02	0.02	0.05	0.00	-0.02	-0.04	0.01	0.00	-0.10	0.02	0.05	0.01	-0.03	0.08	-0.10
X5	-0.04	**0.62**	-0.06	0.07	1.00	0.06	-0.01	-0.27	-0.02	-0.03	0.03	0.31	0.03	-0.05	0.14	0.02	0.29	0.02	0.28	-0.05
X6	0.22	0.05	0.04	-0.02	0.06	1.00	0.12	0.02	0.02	-0.11	0.09	0.02	0.08	0.00	0.01	-0.02	0.01	0.03	0.09	0.01
X7	0.11	0.06	0.14	0.02	-0.01	0.12	1.00	0.13	0.11	-0.01	0.25	0.09	0.26	-0.04	0.11	0.13	0.10	0.10	0.06	-0.03
X8	-0.01	0.07	0.04	0.05	-0.27	0.02	0.13	1.00	0.12	-0.01	0.05	0.05	0.06	-0.00	0.09	0.02	0.10	-0.07	0.01	-0.09
X9	0.04	0.01	0.04	0.00	-0.02	0.02	0.11	0.12	1.00	0.05	-0.03	-0.01	0.01	-0.04	0.10	0.06	-0.01	0.12	0.03	0.07
X10	-0.13	-0.02	-0.04	-0.02	-0.03	-0.11	-0.01	-0.01	0.05	1.00	-0.03	-0.16	-0.03	-0.06	-0.07	-0.03	-0.06	0.02	-0.08	0.12
X11	-0.04	0.03	0.06	-0.04	0.03	0.09	0.25	0.05	-0.03	-0.03	1.00	0.15	0.27	0.00	0.01	0.09	0.01	0.04	0.10	-0.05
X12	-0.03	0.30	-0.05	0.01	0.31	0.02	0.09	0.05	-0.01	-0.16	0.15	1.00	0.07	-0.09	0.35	-0.01	0.28	0.01	0.20	-0.13
X13	0.06	-0.04	0.15	0.00	0.03	0.08	0.26	0.06	0.01	-0.03	0.27	0.07	1.00	-0.04	0.30	0.15	0.02	0.12	0.15	-0.01
X14	0.05	-0.05	0.12	-0.10	-0.05	0.00	-0.04	-0.00	-0.04	-0.06	0.00	-0.09	-0.04	1.00	-0.07	-0.05	-0.00	-0.08	-0.02	0.02
X15	0.02	0.16	0.06	0.02	0.14	0.01	0.11	0.09	0.10	-0.07	0.01	0.35	0.30	-0.07	1.00	0.05	0.11	0.11	0.10	-0.06
X16	0.08	-0.01	0.44	0.05	0.02	-0.02	0.13	0.02	0.06	-0.03	0.09	-0.01	0.15	-0.05	0.05	1.00	-0.03	0.11	0.07	-0.01
X17	0.04	0.21	0.01	0.01	0.29	0.01	0.10	0.10	-0.01	-0.06	0.01	0.28	0.02	-0.00	0.11	-0.03	1.00	-0.09	0.38	-0.10
X18	-0.01	-0.02	0.01	-0.03	0.02	0.03	0.10	-0.07	0.12	0.02	0.04	0.01	0.12	-0.08	0.11	0.11	-0.09	1.00	-0.01	0.08
X19	0.07	0.16	0.05	0.08	0.28	0.09	0.06	0.01	0.03	-0.08	0.10	0.20	0.15	-0.02	0.10	0.07	0.38	-0.01	1.00	-0.11
X20	-0.03	-0.14	0.01	-0.10	-0.05	0.01	-0.03	-0.09	0.07	0.12	-0.05	-0.13	-0.01	0.02	-0.06	-0.01	-0.10	0.08	-0.11	1.00

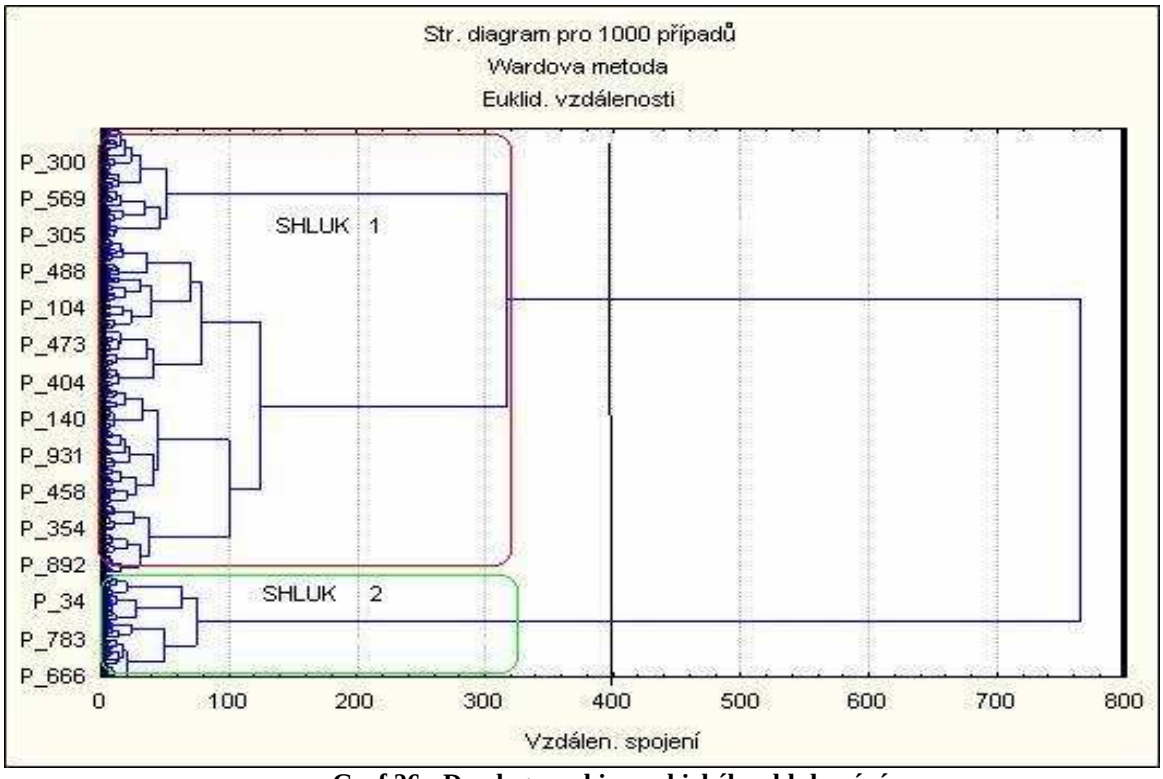
4.5.4 Hierarchické shlukování

Před samotným shlukováním byla na kardinálních znacích (X5, X13) provedena standardizace. Pro hierarchické shlukování byla vybrána v programovém prostředí STATISTICA 7 Wardova metoda, protože jejím principem není optimalizace vzdáleností mezi shluky, ale spočívá v minimalizaci heterogenity shluků podle kritéria minima přírůstku vnitroskupinového součtu čtverců odchylek objektů od těžiště shluků. Důležitou vlastností hierarchických procedur je skutečnost, že výsledky předešlého kroku jsou vždy přidány k výsledkům v následujícím kroku a vytváří tak strukturu ve formě stromu nazývaného dendogram, graf 3.

Pro výpočet vzdáleností mezi jednotlivými znaky byla použita čtverec euklidovská vzdálenost, která tvoří základ Wardovy metody a je definována vztahem [5]

$$d(x_k, x_l) = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2} \tag{7}$$

Po proložení kolmice dendogramem ve vzdálenosti spojení 400 bylo zjištěno, že se data rozdělily do dvou základních shluků. Z nichž první shluk obsahuje cca 70% objektů a druhý sluk cca 30% objektů, což je velice shodné s výsledky obsaženými v datovém souboru.



Graf 26 - Dendogram hierarchického shlukování

4.5.5 Nehierarchické shlukování

Pro další analýzu dat bylo použito nehierarchické shlukování. Byla vybrána metoda k–průměru, protože tato metoda poskytuje pouze jediné řešení pro zadaný počet požadovaných shluků. Jak vyplynulo z hierarchického shlukovaní, byly případy rozděleny do dvou shluků. Tato skutečnost byla zohledněna při volbě počtu shluků.

Postup je založen na nejbližším těžišti, kdy je objekt zařazen do shluku s nejmenší vzdáleností mezi objektem a těžištěm shluku. Vzhledem k tomu, že počáteční těžiště shluků lze získat pomocí hierarchického shlukování [5] realizovaného v předchozí části práce, byla zvolena možnost určování těžišť pomocí prvních *n* pozorování, kde tato pozorování byla získána jako nejbližší hodnoty těžišť shluků získaných hierarchickým shlukováním.

Řešení, které rozdělilo datový soubor obsahující 19 znaků a1000 proměnných do dvou shluků, jak je zobrazeno v tabulce 21, bylo dosaženo již po první iteraci.

Tabulka 21 - Popisná statistika výsledku shlukování

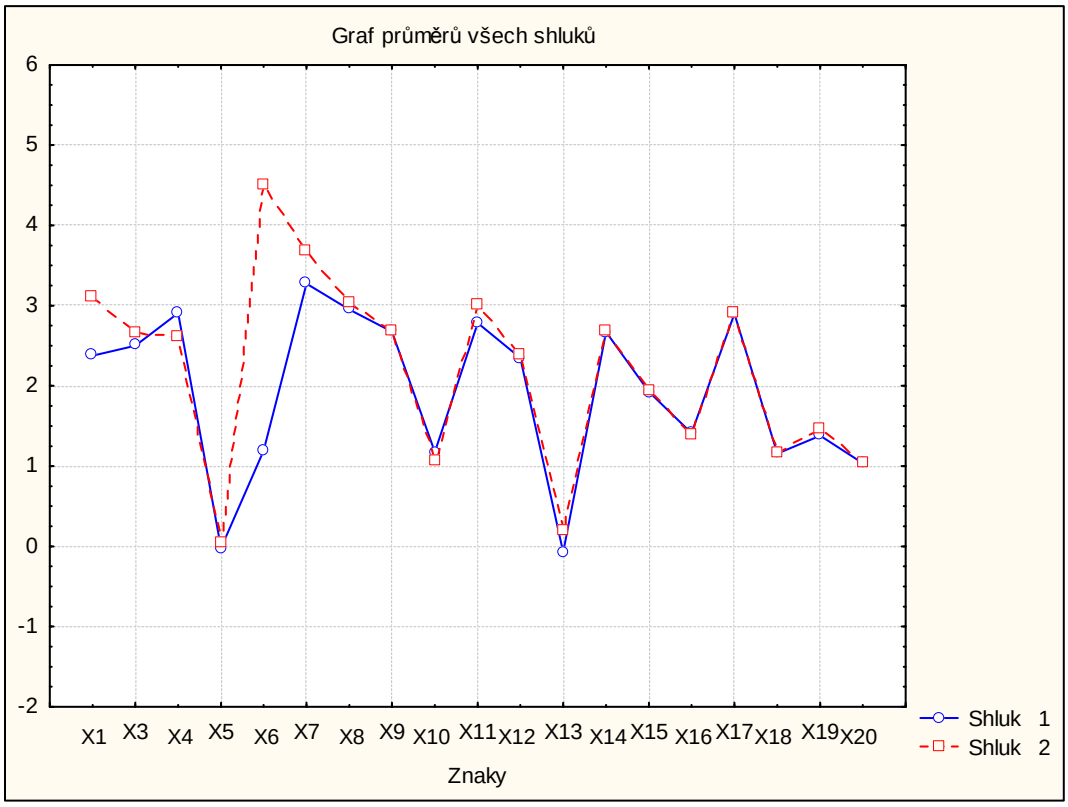
Proměnná	Prům. shluků	
	Shluk 1 (726)	Shluk 2 (274)
X1	2.378453	3.097826
X3	2.501381	2.659420
X4	2.907459	2.619565
X5	-0.021380	0.056084
X6	1.191989	4.500000
X7	3.273481	3.673913
X8	2.950276	3.032609
X9	2.676795	2.695652
X10	1.176796	1.061594
X11	2.784530	3.003623
X12	2.348066	2.384058
X13	-0.073982	0.194069
X14	2.674033	2.677536
X15	1.924033	1.942029
X16	1.412983	1.391304
X17	2.897790	2.920290
X18	1.153315	1.159420
X19	1.378453	1.471014
X20	1.035912	1.039855

4.5.6 Analýza výsledků

Výsledky dosažené pomocí shlukování metodou k-průměru jsou zobrazeny v tabulce 21, která znázorňuje porovnání průměrů hodnot jednotlivých znaků v prvním shluku s hodnotami s druhého shluku. Z tabulky je zřejmé, že se jednotlivé případy rozdělily opět do

dvou shluků v poměru 72.6% případů v prvním shluku a 27,4% případů v druhém shluku. Vzhledem ke skutečnosti, že bylo shlukováno 1000 případů se může konstatovat vysoká shoda s již přiloženým rozdělením v základním datovém souboru, který je rozdělen v poměru 70% případů v prvním shluku a 30% případů v druhém sluku.

Pro přesnější vizualizaci výsledků byl zvolen spojnicový graf 27, který znázorňuje průměry jednotlivých znaků v prvním shluku a zároveň průměry jednotlivých znaků ve druhém shluku.

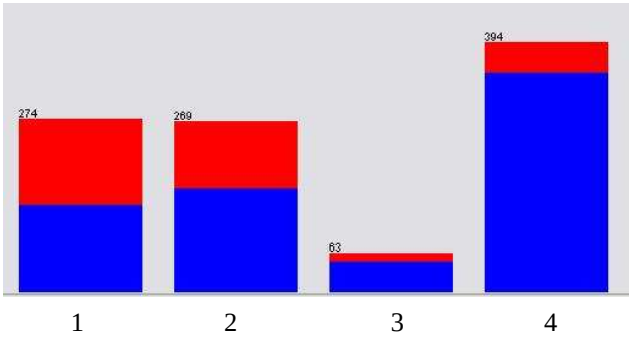


Graf 27 - Průměry shluků

Z grafu 27 vyplývá, že se jednotlivé shluky liší zejména průměry znaků X1, X3, X4, X6, X7, X11,X13. Proto je jisté, že již zmíněné znaky budou rozhodujícími faktory pro zjištění, zdali žadateli o úvěr bude úvěr poskytnut či nikoliv. Z grafu 27 je patrné, že klienti, kterým byl úvěr poskytnut měli na svém běžném účtu zůstatek kolem 200 Euro, což zahrnuje znak X1. Dalším rozhodovacím faktorem je skutečnost, že klienti z prvního shluku mají splacené všechny předešlé úvěry (X3). Poskytnutý úvěr byl dle průměru znaku X4 použit zejména na podnikání a koupi automobilu. Zřejmě největší rozdíl průměrů je u znaku X6, který vyjadřuje dostupnou finanční hotovost klienta. Dle výsledků byla nejmenší výše pro schválení úvěru 100 Euro. Dalším rozhodujícím faktorem byla doba trvání současného zaměstnání (X7), která byla kolem 4 let. Mezi důležité znaky, díky kterým se banka rozhoduje patří i doba působení na adrese trvalého bydliště (X11), která byla opět kolem 4 let a dále věk

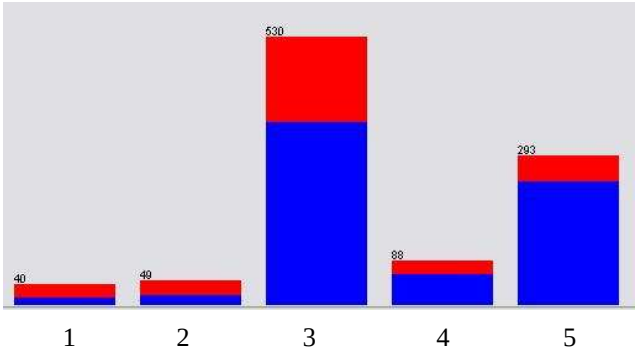
klienta (X13), který byl kolem 30 let. Z tabulky 27 a grafu 27 dále vyplývá, že ostatní znaky nemají výrazně rozhodující vliv na rozdělení případů do jednotlivých shluků. Spíše jsou shodné.

Znaky, které byly vyhodnoceny jako důležité faktory pro rozhodnutí banky byly dál zkoumány pomocí grafů. Z grafu 28, který znázorňuje znak X1 vyplývá, že nejvíce schválených úvěrů bylo u klientů s průměrným zůstatkem 200 Euro a více, což je shodné s výsledkem dosaženým bankou.



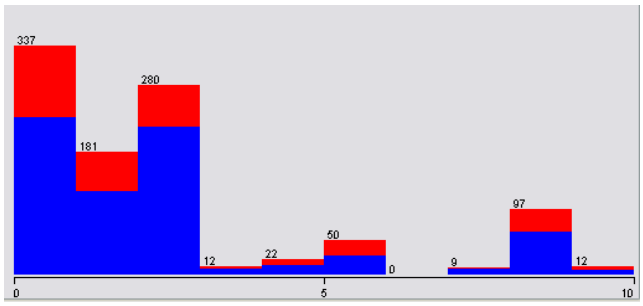
Graf 28 - Znak X1

Jako další zkoumaný znak byl znak X3, který znázorňuje graf 29. Z grafu je viditelné, že nejvíce schválených úvěrů je ve třetím sloupci, který znamená, že klient neměl v minulosti žádný úvěr nebo má všechny úvěry bez problému splacené.



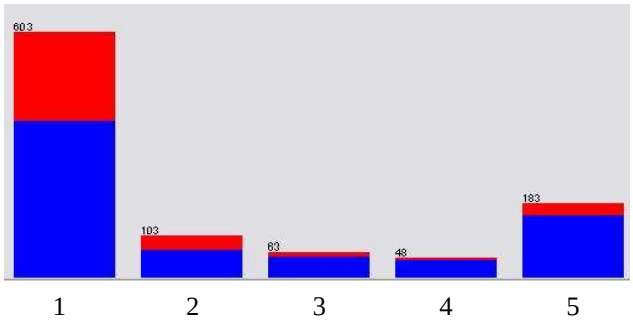
Graf 29 - Znak X3

Zkoumaný znak X4, který vyjadřuje, proč klient o úvěr žádá, je zobrazen v grafu 30. Jak již bylo zmíněno v popisu vstupních dat, tak i ve výsledku shlukování pomocí algoritmu k-průměru, byl úvěr nejvíce schvalován na podnikaní a koupi automobilu.



Graf 30 - Znak X4

V grafu 31 je znázorněn znak X6, který vyjadřuje jakou má klient dostupnou finanční hotovost nebo cenné papíry. V prvním sloupci jsou obsaženi klienti, kteří mají velice nízkou hotovost, proto žádají o úvěr ve znatelně větším počtu, než klienti, kteří mají hotovost kolem 1000 Euro, kterým je úvěr téměř vždy schválen.



Graf 31 - Znak X6

4.5.7 Porovnání výsledků

V závěru příkladu byly porovnány výsledky dosažené pomocí programového rozhraní STATISTICA 7 s výsledky obsaženými v základním datovém souboru.

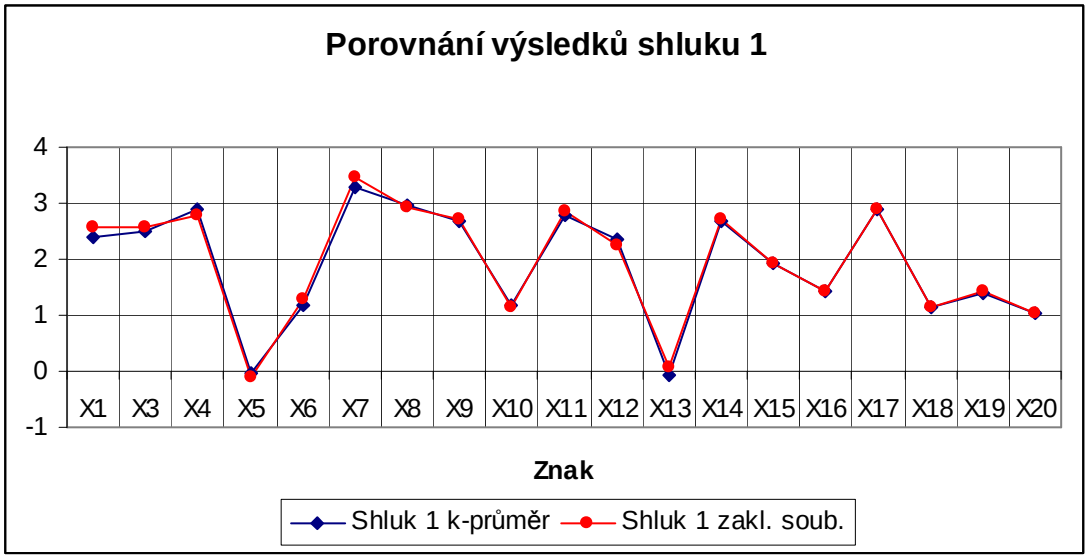
Data rozdělená do jednotlivých shluků pomocí programového rozhraní STATISTICA 7, byla porovnána s klasifikací obsaženou v základním datovém souboru.

V tabulce 22 jsou znázorněny průměry prvního shluku dosaženého pomocí metody k-průměru v porovnání s již obsaženou klasifikací v základním datovém souboru. Tabulka 22 popisuje klienty, kteří splnili kritéria požadované německou bankou. Tudíž jim byl úvěr poskytnut.

Tabulka 22 – Průměry shluku 1

Proměnná	Prům. shluků	
	Shluk 1 k-průměr	Shluk 1 zákl. soub.
X1	2.378453	2.865714
X3	2.501381	2.707143
X4	2.907459	2.795714
X5	-0.021380	-0.101250
X6	1.191989	2.290000
X7	3.273481	3.475714
X8	2.950276	2.920000
X9	2.676795	2.722857
X10	1.176796	1.152857
X11	2.784530	2.842857
X12	2.348066	2.260000
X13	-0.073982	0.059627
X14	2.674033	2.725714
X15	1.924033	1.935714
X16	1.412983	1.424286
X17	2.897790	2.890000
X18	1.153315	1.155714
X19	1.378453	1.415714
X20	1.035912	1.047143

Z grafu 32, který je grafickým znázorněním tabulky 22 vyplývá, že klasifikace dosažena nehierarchickým shlukováním pomocí metody k-průměru je velice shodná s klasifikací přiloženou v základním datovém souboru. Viditelně se liší pouze ve znaku X1 (stav účtu) v průměru o 0,49, který vyjadřuje stav účtu. Dále se liší ve znaku X3 (předcházející úvěry) v průměru o 0,2 a ve znaku X7 (délka zaměstnání) v průměru o 0,2.



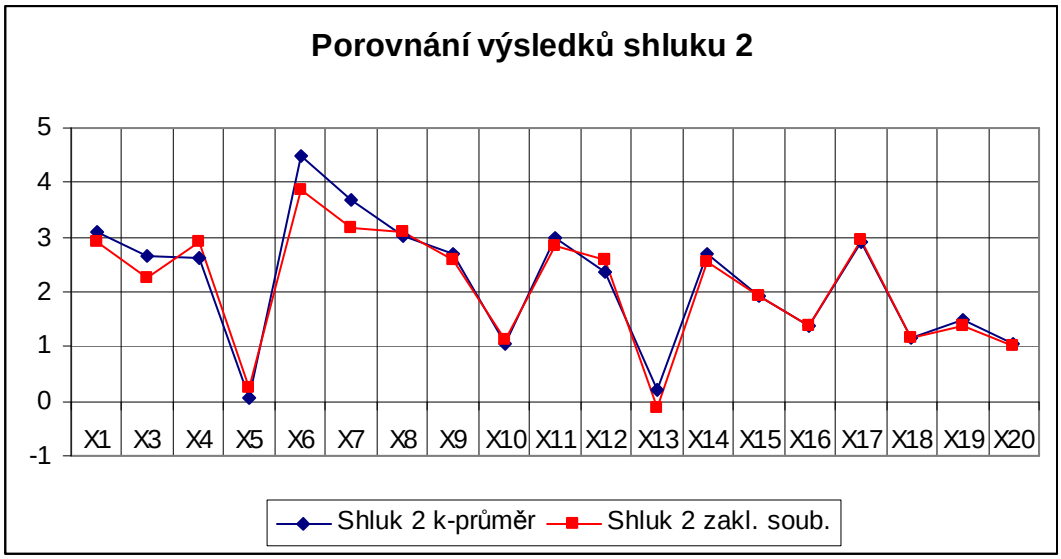
Graf 32 - Porovnání shluk 1

Naopak tabulka 23 znázorňuje průměry shluků 2, který obsahuje klienty, kteří kritéria stanovené bankou nesplnili a z tohoto důvodu jim nebyl úvěr schválen. V tabulce 23 jsou opět porovnány výsledky dosažené nehierarchickým shlukováním pomocí algoritmu metody k-průměru s výsledky obsaženými v základním datovém souboru.

Tabulka 23 - Popisná statistika 2

Proměnná	Prům. shluků	
	Shluk 2 k-průměr	Shluk 2 zákl. soub.
X1	3.097826	2.90333333
X3	2.65942	2.26666667
X4	2.619565	2.90333333
X5	0.05608416	0.236248965
X6	4.5457767	3.87333333
X7	3.673913	3.176565
X8	3.032609	3.09666667
X9	2.695652	2.58666667
X10	1.061594	1.12666667
X11	3.003623	2.8587766
X12	2.384058	2.58666667
X13	0.1940688	-0.1391298
X14	2.677536	2.55666667
X15	1.942029	1.91333333
X16	1.391304	1.36666667
X17	2.92029	2.93666667
X18	1.15942	1.15333333
X19	1.471014	1.37666667
X20	1.039855	1.01333333

V grafu 33 je znázorněna tabulka 23, z které je patrné, že se výsledky výrazněji odlišují pouze v pěti znacích. Konkrétně jsou to znaky X3, X4, X6, X7, X12. Znak X3 (předcházející úvěry) se liší v průměru o 0,39, znak X4 (účel použití úvěru) o 0,29, znak X6 (dostupná finanční hotovost) o 0,67, znak X7 (doba zaměstnání) o 0,5 a znak X12 (dostupné aktiva) se liší v průměru o 0,2.



Graf 33 - Porovnání shluku 2

Znak 1 (úvěr schválen) a znak 2 (úvěr neschválen), obsaženy v tabulce 24, značí počet případů v jednotlivých shlucích dosažených pomocí algoritmu k-průměru vzhledem k původním hodnotám obsažené v základním datovém souboru. Z tabulky 24 vyplývá, že pomocí algoritmu k-průměru bylo zařazeno do prvního shluku o 26 případů více, než bylo zařazeno v původní klasifikaci a do druhého sluku bylo algoritmem k-průměru zařazeno o 26 případů méně než v původní klasifikaci.. V převedení na procentuální vyjádření bylo zjištěno, že algoritmus k-průměru zařadil 97,4% případů do jednotlivých shluků shodně s již provedenou klasifikací obsaženou v základním datovém souboru.

Tabulka 24 - Matice porovnání

Matice porovnání		K - průměr		
		1	2	**Celkem**
Zákl. soubor	1	700	0	**700**
	2	26	274	**300**
	Celkem	**726**	**274**	**1000**

5 Závěr

Práce je soustředěna na aplikaci shlukové analýzy. První část je rozdělena na tří základní body, které popisují shlukovou analýzu. V prvním bodě je popsán samotný účel shlukové analýzy, proč se využívá a k čemu slouží. Je zde vysvětlen rozdíl mezi konvenčním a konceptuálním shlukováním. Dále jaké typy dat se používají pro shlukování. Druhý bod této části je věnován problematice předpokladů shlukové analýzy, do kterých patří zejména nalezení a ošetření chybějících a odlehlých hodnot, dále zjišťování závislosti mezi hodnotami pomocí korelačních koeficientů. Do předpokladů shlukové analýzy byla též zahrnuta problematika standardizace a normalizace. V kapitole 3 jsou vymezeny metody shlukové analýzy. Jsou zde vysvětleny metody hierarchického shlukování, jako je například metoda průměrné vazby, metoda nejbližšího souseda, metoda nejvzdálenějšího souseda, centroidní metoda, mediánová metoda a Wardova metoda. V případě nehierarchického shlukování je popsána metoda k–průměru, metoda k–medoidů, metoda k-modů a k–histogramů.

Druhá část práce je věnována modelování ekonomických dat pomocí metod shlukové analýzy. V uvedeném případě jsou k dispozici data od německé banky, která se na základě provedení analýzy těchto dat rozhoduje, zdali žadateli o poskytnutí úvěru žádost schválí či neschválí. Bankou byly stanoveny kritéria, podle kterých budou žádosti posuzovány. Je to například stav účtu žadatele, podle úvěru spláceného v minulosti, pokud žadatel už nějaký úvěr splácel, podle výše úvěru o jaký klient žádá, podle majetku klienta, věku, zadlužení, zaměstnání a další. Předpoklad tohoto příkladu byl, že se klienti rozdělí do dvou shluků, z nichž v jednom budou obsaženi klienti, kterým bude úvěr schválen, zatímco ve druhém shluku se budou nacházet klienti, kteří na požadovaný úvěr nemají nárok. Vstupní data byla před samotným shlukováním ošetřena od odlehlých hodnot a byla zjištěna závislost mezi jednotlivými znaky. Poté bylo na datech provedeno shlukování a to jak hierarchické, tak i nehierarchické. Pro hierarchické shlukování byla vybrána Wardova metoda a pro nehierarchické shlukování metoda k-průměrů. Po analyzování výsledků bylo zjištěno, že předpoklady rozdělení byly správné. Klienti byli rozděleni do dvou tříd z nichž v první byli obsaženi ti klienti, kteří měli dle kriterií stanovených bankou nárok na poskytnutí úvěru. Z celkového počtu klientů, jich 726 bylo způsobilých pro poskytnutí úvěru, kdežto 274 nikoliv.

V daném příkladu bylo k datům přiloženo i skutečné rozdělení klientů. Při porovnání skutečných výsledků s výsledky shlukové analýzy bylo zjištěno, že algoritmem k-průměru bylo dosaženo správnosti klasifikace 97,4%.

Cíle, které byly stanoveny v úvodu, byly splněny. V první části práce byly shrnuty metody shlukové analýzy, které se v současnosti nejčastěji využívají. V druhé časti práce bylo provedeno předzpracování ekonomických dat z oboru bankovnictví. Jednalo se o data používaná pro schválení či zamítnutí úvěrů. Bylo použito jak hierarchické, tak i nehierarchické shlukování. Hierarchické shlukování bylo realizováno pro ověření počtu shluků v datech a pro zjištění počátečních center pro nehierarchické shlukování. V daném případě se potvrdila existence dvou shluků, tj. úvěruschopných a neúvěruschopných klientů. Dále byly objekty shlukovány algoritmem nehierarchického shlukování, tj. algoritmem k-průměru. Výsledky byly analyzovány tak, že byly porovnány reprezentanti shluků s původními třídami získanými z banky. Ukázalo se, že takové předzpracování dat může být použito pro další metody učení s učitelem, neboť byla analyzována struktura v datech, a navíc, pomocí samotné shlukové analýzy bylo možno správně klasifikovat více než 97,4% klientů.

Použitá literatura

[1] HEBÁK, Petr, et al. *Vícerozměrné statistické metody 1*. 2. přeprac. vyd. Praha : INFORMATORIUM, spol. s. r. o., 2007. 245 s. ISBN 978-80-7333-056-9.

[2] HEBÁK, Petr, et al. *Vícerozměrné statistické metody 3*. 2. přeprac. vyd. Praha : INFORMATORIUM, spol. s. r. o., 2007. 262 s. ISBN 978-80-7333-001-9.

[3] LUKASOVÁ, Alena, ŠARMOVÁ, Jana. *Metody shlukové analýzy*. Praha : Státní nakladatelství technické literatury, 1985. 276 s.

[4] MAŘÍK, V., et al. *Umělá inteligence (1)*. Praha : Academia, 1993. 253 s.

[5] MELOUN, Milan, MILITKÝ, Jiří, HILL, Martin. *Počítačová analýza vícerozměrných dat v příkladech*. 1. vyd. Praha : Academia, 2005. 445 s. ISBN 80-200-1335-0.

[6] MELOUN, Milan, MILITKÝ, Jiří. *Statistická analýza experimentálních dat*. Praha : Academia, 2004. 953 s. ISBN 80-200-1254-0.

[7] ŘEZÁNKOVÁ, Hana, DUŠAN, Ondřej, VÁCLAV, Snášel. *Shluková analýzy dat*. 1. vyd. Příbram : Profesional Publishing, 2007. 196 s. ISBN 978-80-86946-26-9.

[8] *UCI Machine Learning Repository* [online]. 1995 [cit. 2009-03-21]. Dostupný z WWW: <http://www.ics.uci.edu/>.

[9] *Wikipedie, otevřená encyklopedie* [online]. 2002-2003 [cit. 2009-03-23]. Dostupný z WWW: <http://cs.wikipedia.org>.