VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

METODY ZÍSKÁVÁNÍ ZNALOSTÍ Z DAT

DATA MINING TECHNIQUES

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **MONIKA KUBINCOVÁ**
AUTHOR

VEDOUCÍ PRÁCE **Ing. IVANA BURGETOVÁ, Ph.D.**
SUPERVISOR

BRNO 2023

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Zadání bakalářské práce

144245

Ústav: Ústav informačních systémů (UIFS)
Studentka: **Kubincová Monika**
Program: Informační technologie
Specializace: Informační technologie
Název: **Metody získávání znalostí z dat**
Kategorie: Data mining
Akademický rok: 2022/23

Zadání:

1. Prostudujte problematiku analýzy dat a získávání znalostí z dat.
2. Seznamte se s dostupnými daty zvolené obchodní společnosti.
3. Navrhněte několik analytických úloh nad dostupnými daty, jejichž výsledky by mohly být využitelné pro rozhodování.
4. Po dohodě s vedoucí zvolte některou/některé z těchto analytických úloh.
5. Navrhněte a v jazyce Python implementujte řešení zvolené úlohy.
6. Zhodnoťte dosažené výsledky.

Literatura:

- Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Third Edition. Morgan Kaufmann Publishers, 2012, 703 p., ISBN 978-0-12-381479-1. (Kapitola 1 a další informace relevantní ke zvolené úloze a použitým datům).
- Nielsen, F.A.: Data Mining with Python. 2015. 101 p. Dostupné na http://www.freetechbooks.com/data-mining-with-python-working-draft-t1159.html.

Při obhajobě semestrální části projektu je požadováno:

- Body 1 až 4.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce: **Burgetová Ivana, Ing., Ph.D.**
Vedoucí ústavu: Kolář Dušan, doc. Dr. Ing.
Datum zadání: 1.11.2022
Termín pro odevzdání: 10.5.2023
Datum schválení: 18.10.2022

Abstrakt

Bakalárska práca sa zaoberá spracovaním a analýzou dát obchodnej spoločnosti, s cieľom vytvoriť analytický nástroj na pravidelné získavanie poznatkov z dát, ktorý pomáha firme s dôležitými strategickými rozhodnutiami. V teoretickej časti práce sú popísané rôzne spôsoby získavania poznatkov a spracovania dát s výrazným zameraním na metódu zhlukovania. Práca ďalej opisuje dostupné dátové sady, ktoré boli využité na analýzu. Nasleduje implementácia navrhnutých úloh. V závere práce sú prezentované výsledky analýzy a využiteľnosť do budúcna, vrátane návrhov na zlepšenie.

Abstract

The Bachelor's thesis deals with the processing and analysis of data from a commercial company, aiming to create an analytical tool for regular knowledge extraction from data that assists the company with important strategic decisions. The theoretical part of the thesis describes various methods of data mining and data processing, with a significant focus on the clustering method. The thesis further describes the available datasets that were used for the analysis and implementation of the proposed tasks. The final part of the concludes results of the analysis and its future usability including suggestions for improvement.

Kľúčové slová

Dolovanie dát, Získavanie znalostí z dát, Zhlukovanie, Python, Zhluková analýza, Analýza dát, Hierarchické zhlukovanie

Keywords

Data mining, Data analysis, Python, Clustering, Hierarchical clustering

Citácia

KUBINCOVÁ, Monika. *METODY ZÍSKÁVÁNÍ ZNALOSTÍ Z DAT*. Brno, 2023. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Ivana Burgetová, Ph.D.

METODY ZÍSKÁVÁNÍ ZNALOSTÍ Z DAT

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracovala samostatne pod vedením pani Ing. Ivany Burgetovej, Ph.D. Uviedla som všetky literárne pramene, publikácie a ďalšie zdroje, z ktorých som čerpala.

.....
Monika Kubincová
9. mája 2023

Poďakovanie

Chcela by som poďakovať pani Ing. Ivany Burgetovej, Ph.D. za ochotu, trpezlivosť a odborné rady pri vedení tejto bakalárskej práce.

Obsah

1 Úvod **4**

2 Dolovanie dát **5**
2.1 Dáta . . . 5
2.1.1 Kvalita dát . . . 6
2.1.2 Klasifikácia štruktúry dát . . . 6
2.2 Znalosť . . . 7
2.3 Získavanie znalostí z dát . . . 8
2.3.1 Čistenie dát . . . 9
2.3.2 Integrácia dát . . . 9
2.3.3 Výber dát . . . 9
2.3.4 Transformácia dát . . . 9
2.3.5 Dolovanie dát . . . 9
2.3.6 Vyhodnotenie a interpretácia dát . . . 10
2.4 Typické úlohy získavania znalostí z dát . . . 10
2.4.1 Klasifikácia . . . 10
2.4.2 Asociačné pravidlá . . . 10
2.4.3 Predikcia . . . 11
2.4.4 Regresia . . . 11
2.4.5 Detekcia odľahlých objektov . . . 11
2.4.6 Sekvenčné vzory . . . 11

3 Zhlukovanie **12**
3.1 Využitie zhlukovej analýzy . . . 12
3.2 Delenie zhlukovacích metód . . . 13
3.2.1 Hierarchické metódy . . . 14
3.2.2 Metriky pre vzdialenosť medzi zhlukmi . . . 15
3.2.3 Kofenetický korelačný koeficient . . . 17
3.3 Vstupné a výstupné dáta . . . 18

4 Dáta a úlohy **19**
4.1 Opis dát . . . 19
4.2 Porozumenie dátam . . . 22
4.3 Posúdenie využiteľnosti rôznych metód . . . 23
4.4 Pracovné postupy . . . 24

5 Implementácia **25**
5.1 Predspracovanie dát . . . 26

5.2 Analýza dát 28
5.2.1 Analýza pozície ceny 28
5.2.2 Analýza predajnosti podľa kategórii 28
5.3 Zhlukovanie dát 28
5.4 Tvorba užívateľského rozhrania 30

6 Vyhodnotenie **31**
6.1 Vyhodnotenie spracovania a čistenia dát 31
6.2 Vyhodnotenie analýz dát 32
6.2.1 Predajnosť v čase 32
6.2.2 Predajnosť vzhľadom na konkurenčné ceny 35
6.3 Vyhodnotenie zhlukovania 36
6.3.1 Prínos vytvorenia užívateľského prostredia 37
6.4 Odporúčania firme v závislosti na ďalšie možnosti 38

7 Záver **40**

Literatúra **41**

A Manuál **43**

B Obsah priloženého pamäťového média **44**

Zoznam obrázkov

2.1 Štruktúra dát 7
2.2 Kroky procesu získavania znalostí z dát. 8

3.1 Hierarchické metódy. 14
3.2 Minimálna vzdialenosť. 15
3.3 Maximálna vzdialenosť. 16
3.4 Priemerná vzdialenosť. 16
3.5 Stredná vzdialenosť. 17
3.6 Dátová matica (prevzaté z [10]). 18
3.7 Vzdialenostná matica (prevzaté z [10]). 18

5.1 Matica pre hierarchické zhlukovanie (dendro_data) 29

6.1 Predajnosť podľa cenovej kategórie v internetovom obchode 33
6.2 Predajnosť podľa cenovej kategórie v kamennom obchode 34
6.3 Predajnosť podľa typu vozidla v internetovom obchode 35
6.4 Predajnosť produktov vzhľadom na pozíciu ceny v internetovom obchode . 36
6.5 Dendrogram zhlukovania predajov v kamennej predajni s metrikou maximálnej vzdialenosti 37
6.6 Dendrogram zhlukovania predajov v kamennej predajni s metrikou minimálnej vzdialenosti 38
6.7 Dendrogram zhlukovania predajov v kamennej predajni s metrikou priemernej vzdialenosti 38
6.8 Dendrogram zhlukovania predajov v kamennej predajni s metrikou strednej vzdialenosti 39

Kapitola 1

Úvod

Získavanie poznatkov z dát sa v posledných rokoch stalo dôležitým nástrojom na identifikáciu vzorov, trendov a kritických poznatkov pre riadenie a rozhodovanie v podnikoch. Avšak, v dnešnej dobe, kedy sa digitalizácia dotýka takmer každej činnosti, množstvo dostupných dát môže pôsobiť ako veľká nezrozumiteľná masa. Preto je dôležité vedieť túto masu dát rozlúštiť a identifikovať vzory, ktoré nám pomôžu pochopiť ich správanie. Existuje mnoho spôsobov získavania poznatkov z dát a ich využitie závisí od charakteristík dostupných datasetov alebo od požadovanej špecifikácie výsledkov. Zámerom tejto bakalárskej práce bolo preskúmanie metód dolovania dát a tvorba analytického nástroja, ktorý bude firma môcť pravidelne využívať aj v budúcnosti.

Na otestovanie týchto metód v praxi som sa rozhodla spolupracovať s malou lokálnou spoločnosťou, ktorá sa zaoberá predajom pneumatík, kde bola problematika spracovania a získavania poznatkov z dát vítaným prínosom pre ďalší strategický rozvoj. Na účely tejto práce som pracovala s dátovými sadami poskytnutými reálnou obchodnou spoločnosťou a tieto dáta obsahovali informácie o predaji, kategóriách produktov a ich pozícii v konkurenčnom prostredí. Predspracovanie a čistenie týchto dát si vyžiadalo značné množstvo času, kým som získala formu na ďalšie spracovanie. Z charakteru dostupných dát vyplynulo, že najlepšou metódou získavania poznatkov bude zhlukovanie. Dôležitou časťou spracovania dátových sád bolo aj analyzovanie a vizualizácia predajných dát a vytvorenie nástroja s intuitívnym používateľským rozhraním, ktoré môže pomôcť pri efektívnejšom získavaní a interpretácii poznatkov.

Práca sa začína teoretickou časťou - kapitola 2, kde je uvedený proces, najznámejšie metódy dolovania dát a ich využitie. Nasledujúca kapitola 3 je venovaná metóde zhlukovania, ktorú využívame pre analýzu. V kapitole 4 sa nachádza opis poskytnutých dát a návrhy úloh. Kapitola 5 sa zaoberá implementáciou zvolených úloh. Na ktorú priamo naväzuje kapitola 6 s vyhodnotením týchto úloh a interpretáciou ich výsledkov pre potreby obchodnej spoločnosti. Taktiež poskytuje návrhy na zlepšenie práce s dátami do budúcna.

Kapitola 2

Dolovanie dát

Dolovanie dát alebo data mining je proces hľadania vzorcov, vzťahov a ďalších trendov v dátach za účelom získania užitočných informácií a znalostí. Jeho cieľom je identifikácia dôležitých informácií, ktoré sa môžu využiť na optimalizáciu, zlepšenie rôznych procesov a predikciu budúcich trendov. V súčasnosti je data mining veľmi dôležitým procesom pri riešeniach mnohých problémov v oblasti vedy, obchodu a priemyslu.

Táto kapitola obsahuje informácie, ktoré čitateľa oboznámia s komplexnou témou získavania znalostí z dát. Konkrétne sa jedná o popis rôznych typov dát, kvalitu dát a klasifikáciu ich štruktúry. Ďalšie podsekcie sa zaoberajú procesom získavania znalostí z dát a typickými metódami dolovania.

2.1 Dáta

V dnešnej dobe sme zahltení veľkým množstvom dát, ale chýbajú nám znalosti. Tieto dáta pochádzajú z rôznych zdrojov, ako napríklad:

- databázy a dátové sklady
- webové stránky a sociálne médiá
- textové dokumenty a správy
- senzory a zariadenia pre zber dát
- obrazové a zvukové záznamy
- medicínske a biologické dáta
- finančné a ekonomické trhy.

Všetky tieto dáta môžu byť veľmi cenné, pokiaľ sú správne analyzované a interpretované. Získavanie znalostí z dát zahŕňa mnoho úloh, ako je výber a príprava dát, transformácia dát, modelovanie a evaluácia modelov a interpretácia výsledkov. Tieto úlohy sú veľmi podstatné pre efektívne využitie dát v rôznych oblastiach a majú veľký potenciál pre vývoj nových technológií a inovácií. Výsledky získavania znalostí z dát môžu mať rôzne formy, od jednoduchých vizualizácií a grafov po zložité modely strojového učenia. Závisí to od charakteru dát a cieľa analýzy.

2.1.1 Kvalita dát

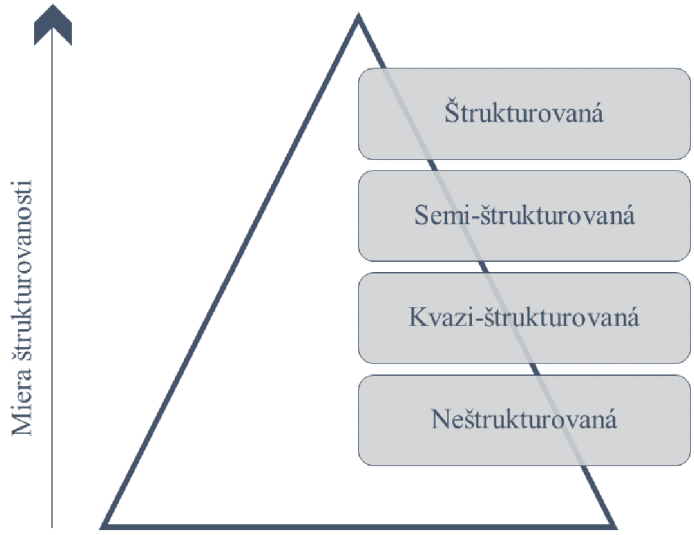
Reálne dáta, ktoré sa používajú ako zdroje pre dolovanie dát vo väčšine prípadov nie sú v takom stave, aby boli využiteľné bez akejkoľvek úpravy. Kvalita dát sa preto stala kľúčovým faktorom. Nízka kvalita môže viesť k nesprávnym alebo zavádzajúcim záverom. Existujú rôzne aspekty hodnotenia kvality dát, ako napríklad presnosť, úplnosť, konzistencia, aktuálnosť, interpretovateľnosť a dostupnosť [17].

2.1.2 Klasifikácia štruktúry dát

V oblasti získavania a analyzovania dát existujú rôzne typy dát, ktoré sa môžu líšiť svojou štruktúrou. Ako je vidieť na obrázku 2.1, medzi najbežnejšie typy dát patria [2]:

- **štruktúrované**
  - majú jasne definovanú štruktúru a organizáciu, kde každý atribút má presne určený typ a hodnoty.
  - relačné, transakčné, CSV atď.
- **semi-štruktúrované**
  - majú nepravidelnú štruktúru a organizáciu, ale obsahujú určitú formu štandardizovaných štítkov alebo značiek, ktoré pomáhajú s ich identifikáciou.
  - dokumenty XML, JSON, HTML.
- **kvázi-štruktúrované**
  - sú to dáta, ktoré nie sú plne normalizované alebo normalizované na minimálnej úrovni. Nemajú pravidelný formát.
  - záznam priechodu webom (clickstream).
- **neštruktúrované**
  - je to najväčšia skupina dát, nemajú žiadnu jasne definovanú štruktúru a organizáciu.
  - textové formáty, multimediálne súbory ako audio, obrázky a videá.

Každý typ dát má svoje vlastné metódy analýzy. Z toho dôvodu je dôležité ich správne identifikovať a vybrať vhodné metódy získavania znalostí a analýzy pre daný typ dát.



Obr. 2.1: Štruktúra dát

2.2 Znalosť

V oblasti data miningu sa pojem znalosť zvyčajne vzťahuje k informáciám, ktoré sú získané a použité k pochopeniu a vysvetleniu javov, ktoré dáta popisujú. Táto informácia by mala byť netriviálna (nedá sa logicky odvodiť alebo získať napr. jednoduchým SQL dotazom), nemusí byť na prvý pohľad zrejmá priamo z dát, ale je odhalená nejakou formou analytického procesu a mala by mať potenciál byť prakticky využitá [6].

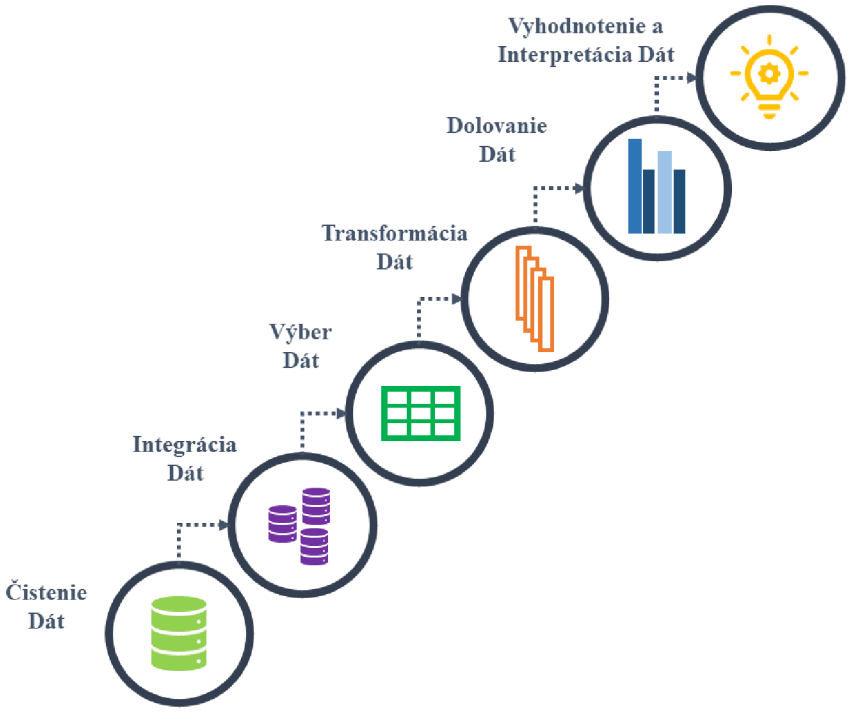
Tieto znalosti sa využívajú v rozličných odvetviach, napríklad v oblasti obchodu, financií, marketingu, výroby, medicíny a mnohých ďalších.

2.3 Získavanie znalostí z dát

Proces získavania znalostí z dát (z angl. Knowledge Discovery in Databases, skr. KDD) je proces, zložený z niekoľkých krokov, ktorých cieľom je identifikovať závislosti, vzorce a skryté informácie v dátach. Tento proces bol prvýkrát popísaný v roku 1989 Gregorym Piatetsky-Shapiro a Williamom Frawley v článku "Knowledge Discovery in Databases"[6]. Od tej doby sa stal dôležitým procesom v oblasti dolovania dát.

Hlavné kroky procesu dolovania dát boli prevzaté z [8] a vidíme ich aj na obrázku 2.2:

1. Čistenie dát
2. Integrácia dát
3. Výber dát
4. Transformácia dát
5. Dolovanie dát
6. Vyhodnotenie a interpretácia dát



Obr. 2.2: Kroky procesu získavania znalostí z dát.

2.3.1 Čistenie dát

Je to proces, ktorý sa zaoberá odstraňovaním chýb, nekompletných údajov a nepresností z databázy alebo datasetu. Účelom čistenia je zabezpečiť, kvalitu a relevantnosť pre modelovanie, analýzu alebo ďalšie spracovanie. Proces čistenia dát môže byť zdĺhavý a využívajú sa pri ňom rôzne nástroje a technológie.

Najčastejšie typy chýb sú duplicitné záznamy, chýbajúce alebo neúplne dáta, chybné hodnoty v dátach alebo nesprávne formátovanie. Tieto chyby môžu mať vplyv na spoľahlivosť a kvalitu analýz založených na daných dátach.

Vhodným spôsobom, ako riešiť duplicitné záznamy, môže byť odstránenie duplikátov alebo ich zoskupenie do jednej položky. Riešením chýbajúcich hodnôt, je odstránenie položiek s chýbajúcimi hodnotami alebo doplnenie chýbajúcich hodnôt, napríklad pomocou priemeru alebo mediánu. Chybné hodnoty v dátach je možné identifikovať pomocou pravidiel alebo algoritmov a nahradiť ich správnymi dátami [14].

2.3.2 Integrácia dát

Definuje sa ako proces zlučovania dát z rôznych zdrojov a ukladanie ich do jednotného celku, ktorý môže byť ďalej spracovaný. Za cieľ má vytvoriť spoľahlivú a ucelenú sadu dát, ktorá umožní lepšie a efektívnejšie vykonávanie analýz a modelovania. Problémy pri integrácií dát sú napríklad zdrojové dáta obsahujúce rozličné typy a formáty dát, odlišné hodnoty prislúchajúcich atribútov alebo môžu byť uložené v rôznych systémoch.

2.3.3 Výber dát

Výber dát je fáza, ktorá spočíva v identifikácii a výbere relevantných dát z rozsiahlej databázy alebo datasetu. Tento proces zahŕňa použitie rôznych techník na extrakciu dôležitých dát z množstva dát, kde niektoré časti môžu byť nezmyselné alebo nepotrebné v kontexte konkrétnej úlohy alebo problému. Cieľom výberu dát je zlepšiť kvalitu a efektivitu ďalších krokov v procese analýzy dát alebo modelovania. Výber dát je kritickou fázou v procese dolovania dát a má vplyv na výsledky a úspešnosť celého procesu. Existuje mnoho rôznych techník a algoritmov, ktoré sa používajú na výber dát, vrátane metód pre filtrovanie, transformáciu a selekciu dát.

2.3.4 Transformácia dát

Je to proces, ktorý sa zoberá zmenou štruktúry alebo formátu dát, na dosiahnutie lepšie interpretovateľných alebo spracovateľných dát. Môže to zhŕňať rôzne úpravy dát, ako napríklad agregáciu, normalizáciu, nahradenie chýbajúcich hodnôt alebo iné.

2.3.5 Dolovanie dát

Aplikovanie algoritmov a techník pre hľadanie vzorcov a získanie užitočných znalostí z dát. Dolovanie dát predstavuje jeden z najdôležitejších krokov v procese získavania znalostí z dát, pretože umožňuje odhaliť rôzne zaujímavé vzorce, závislosti a trendy v dátach, ktoré by boli inak ťažko rozpoznateľné. Podrobnejšie sa metódam dolovania dát venujem v sekcii 2.4.

2.3.6 Vyhodnotenie a interpretácia dát

V poslednej fáze procesu dolovania dát sa získané výsledky vyhodnocujú a interpretujú s cieľom odvodiť z nich relevantné poznatky. Využívajú sa rôzne techniky a nástroje, ako napríklad vizualizácia dát, štatistické metódy, strojové učenie a iné. Účelom je zistiť, čo získané výsledky znamenajú v súvislosti s riešením problému a či sú relevantné.

2.4 Typické úlohy získavania znalostí z dát

Existuje množstvo rôznych metód získavania znalostí z dát, ktoré sa líšia svojimi prístupmi, technikami a algoritmami. Každá z týchto metód má svoje výhody a obmedzenia, ktoré súvisia s charakteristikami dát a cieľmi získavania znalostí. Informácie v tejto sekcii sú prevzaté najmä z [8] a [16].

Typické úlohy dolovania dát zahŕňajú:

- Klasifikáciu
- Regresiu
- Asociačné pravidlá
- Predikciu
- Detekcia odľahlých objektov
- Sekvenčné vzory
- Zhlukovanie

V nasledujúcich podsekciách sú jednotlivé úlohy popísané podrobnejšie, s výnimkou zhlukovania, ktorému je venovaná samostatná kapitola 3, keďže táto metóda bola vybraná na analýzu získaných dát.

2.4.1 Klasifikácia

Klasifikácia je proces priraďovania dát do tried alebo kategórií na základe určitých charakteristík či vlastností. Ide o jednu z hlavných úloh v oblasti strojového učenia a data miningu, pričom sa využívajú rôzne algoritmy a techniky, ako napríklad rozhodovacie stromy, k-najbližších susedov, náhodný les a ďalšie. Proces dolovania dát pomocou klasifikácie určuje, čo charakterizuje dáta z jednotlivých tried a vymedzuje ich od ostatných. Jej výsledkom môže byť nový poznatok ako klasifikovať dáta [3].

Využíva sa napríklad pri predikcii správania zákazníkov, detekcii podvodov, diagnostike chorôb a podobne.

2.4.2 Asociačné pravidlá

Táto technika dolovania dát sa snaží nájsť vzorce a závislosti v dátach. Tieto vzorce predstavujú spojenia medzi rôznymi položkami alebo atribútmi dát a často nie sú viditeľné z pohľadu ľudského pozorovateľa. Asociačné pravidlá sa môžu používať napríklad na doporučenie produktov, ktoré sa často kupujú spolu (kúpa mobilného telefónu + kryt/ochranné sklo na telefón).

2.4.3 Predikcia

Predikcia je proces predpovedania budúceho vývoja na základe dát a informácií, ktoré boli natrénované na existujúcich dátach o minulom a súčasnom stave. V oblasti data miningu sa predikcia zameriava na vytváranie modelov a algoritmov, ktoré dokážu na základe existujúcich dát predpovedať budúce hodnoty. Je to dôležitý nástroj pri rozhodovaní a plánovaní budúcich krokov.

2.4.4 Regresia

Regresia je jednou z najpoužívanejších metód predikcie. Je to štatistická technika na skúmanie a modelovanie vzťahu medzi premennými. Pomáha predpovedať hodnoty nejakej premennej na základe hodnôt iných premenných. Používa sa v mnohých oblastiach, napríklad v ekonómii, marketingu, medicíne, biológii a iných. V regresii sa používajú rôzne typy modelov, ako napríklad lineárna regresia, nelineárna regresia, polynomiálna regresia a mnohé ďalšie. Výsledky regresnej analýzy sa často vyjadrujú pomocou regresných rovníc alebo grafických reprezentácií [11].

2.4.5 Detekcia odľahlých objektov

Identifikuje nezrovnalosti, odchýlky, anomálie a výrazne odlišné alebo neštandardné prvky v dátach, ktoré sa vymykajú od predpokladaných vzorcov z hľadiska podobnosti alebo správania. Tieto odchýlky často obsahujú užitočné informácie o nezvyčajných charakteristikách systémov a entít, ktoré ovplyvňujú proces tvorby dát. Rozpoznanie takýchto neobvyklých charakteristík poskytuje užitočné poznatky pre konkrétnu aplikáciu, napríklad môžu byť spôsobené chybami, podvodmi alebo inými neobvyklými situáciami [1].

2.4.6 Sekvenčné vzory

Slúžia na hľadanie vzorov a postupností v sekvenciách dát. Tieto vzory sú založené na časovej postupnosti, sú podobné asociačným pravidlám v tom, že sa hľadajú vzťahy medzi dátami (alebo udalosťami), ale vzťah je založený na čase [5]. Cieľom je identifikovať zákonitosti v sekvenciách, ktoré sa môžu vyskytnúť v určitom poradí alebo s určitou frekvenciou. Vyskytuje sa to napríklad v medicínskej diagnostike.

Kapitola 3

Zhlukovanie

Zhlukovanie (clustering) je technika v oblasti data miningu, ktorá sa používa na rozdelenie množiny dát do skupín (zhlukov) na základe podobných charakteristík objektov v rámci skupín a zároveň na základe odlišností od objektov v iných skupinách. Zhlukovanie má za cieľ vytvoriť homogénne skupiny, ktoré nie sú explicitne deklarované a ktoré by mohli pomôcť pri lepšom pochopení dát a pri rozhodovaní. Zhlukovanie patrí do kategórie učenia bez učiteľa. Nevyžaduje žiadne preddefinované triedy ani trénovaciu množinu príkladov. Dobrá metóda zhlukovania dokáže vytvoriť zhluky s veľkou podobnosťou medzi objektami v rámci jednej triedy (high intra-class similarity) a s malou podobnosťou objektov medzi rôznymi triedami (low inter-class similarity).

Kvalita výsledkov zhlukovania závisí na implementácii zvolenej metódy alebo na vzdialenostnej funkcii. Táto funkcia je použitá na meranie podobnosti (odlišnosti) objektov a určuje, ako ďaleko sa nachádzajú objekty od seba a triedi ich podľa toho. Pre zhlukovanie sa často používajú nasledujúce vzdialenostné funkcie, ako Euclidean distance, Manhattan distance alebo Minkowski distance. Úspešnosť zhlukovania môže byť ovplyvnená voľbou správnej vzdialenostnej funkcie [8].

V tejto kapitole je popísané zhlukovanie, jeho využitie a delenie metód. Dôraz je kladený najmä na metódu hierarchického zhlukovania, ktoré je využívané pri analýze dát v praktickej časti tejto práce a preto je detailnejšie popísané v podkapitole 3.2.1.

3.1 Využitie zhlukovej analýzy

Zhluková analýza má mnoho rôznych využití, ako napríklad v bioinformatike, pri detekcii odľahlých hodnôt, rozpoznávaní vzorov, spracovaní obrazov alebo na klasifikáciu pri rozdelení objektov do tried. V praxi sa zhlukovanie taktiež často používa v oblasti ekonomiky a marketingu, kde sa snažíme analyzovať trh a identifikovať skupiny zákazníkov s podobným správaním alebo preferenciami, aby sme mohli zacieliť naše marketingové aktivity konkrétnym skupinám zákazníkov. Existuje viacero metód zhlukovania, ktoré sa líšia v závislosti od spôsobu, ako vypočítavajú podobnosť medzi jednotlivými objektmi a ako tieto objekty priraďujú do skupín. Medzi najpoužívanejšie metódy zhlukovania patria k-means, hierarchické zhlukovanie, DBSCAN a mnohé ďalšie.

3.2 Delenie zhlukovacích metód

Metódy zhlukovania delíme na [8]:

- **Metódy založené na rozdelení** (partitioning methods)
  - Používajú sa na vytvorenie k tried databáze, pričom $k<=n$ a n predstavuje počet objektov. Každý objekt patrí len do jednej triedy a každá trieda obsahuje aspoň jeden objekt. Na začiatku sa zvolí počiatočné rozdelenie dát do tried, ktoré sa následne postupne optimalizuje a za ideálnych okolností upravuje podľa vyhodnotenia všetkých možných rozdelení databázy. Avšak zvyčajne je týchto rozdelení veľa, tak sa začne s jedným a to sa následne optimalizuje. Typicky sa teda neprehľadajú všetky možné rozdelenia. Preto sa niekedy tieto metódy spúšťajú viackrát s rôznym počiatočným rozdelením prvkov.
  - K-means, K-medoids a CLARANS.
- **Metódy založené na hustote** (density-based methods)
  - Pri týchto metódach je zhluk definovaný ako oblasť v priestore, kde je vysoká hustota prvkov a je ohraničená oblasťou s nižšou hustotou. Tieto metódy sú schopné detekovať rôznorodé tvary a veľkosti zhlukov. Taktiež dokážu odfiltrovať objekty, ktoré sú považované za šum. Nevýhodou je potreba definovať parameter hustoty.
  - DenClue a DBSCAN.
- **Modelové metódy** (model-based methods)
  - tieto metódy využívajú určitý model alebo predpoklad o rozdelení dát a následne optimalizujú jeho parametre tak, aby najlepšie vystihovali skutočné rozdelenie dát.
  - Metoda Expectation-Maximization, Konceptuálne zhlukovanie, Metódy neurónových sietí (SOM).
- **Metódy založené na mriežke** (grid-based methods)
  - Tieto metódy delia priestor dát na konečný počet buniek, ktoré tvoria mriežku a následne vytvárajú zhluky na základe počtu objektov v bunkách mriežky. Tieto metódy často využívajú histogramy alebo rastrovú analýzu, ktorá umožňuje efektívne získanie štatistík pre jednotlivé bunky mriežky.
  - WaveCluster a Clique.
- **Hierarchické metódy** (hierarchical methods)
  - Sú metódy, ktoré vytvárajú hierarchiu zhlukov na základe ich podobnosti. Pri hierarchických metódach je výsledkom stromová štruktúra, kde listy predstavujú jednotlivé objekty a vrcholy reprezentujú zhluky.
  - Diana, Agnes, BIRCH, ROCK a Chameleon.

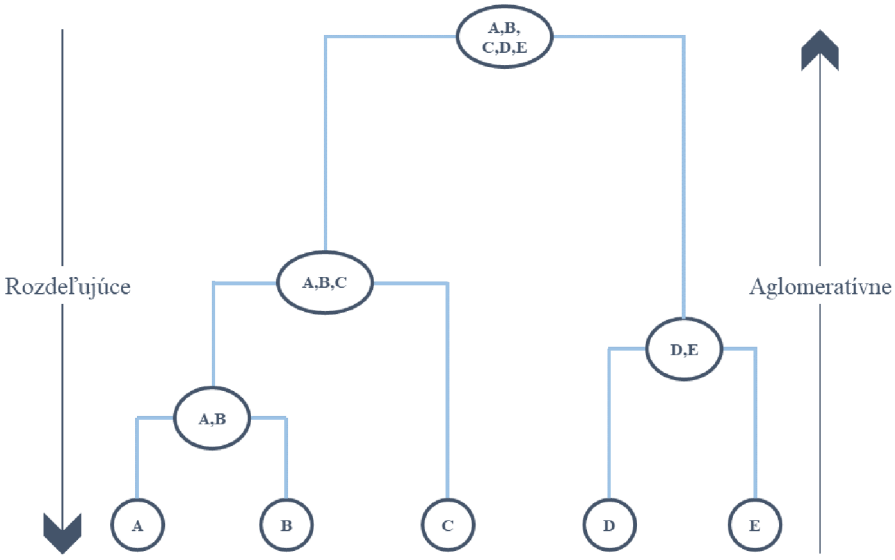
3.2.1 Hierarchické metódy

Hierarchické metódy zhlukovania produkujú hierarchické reprezentácie, v ktorých sú zhluky na každej úrovni hierarchie vytvárané zlúčením zhlukov na nižšej úrovni. Na najvyššej úrovni existuje len jeden zhluk obsahujúci všetky objekty [9].

Tieto metódy sa spoliehajú na koncept podobnosti alebo vzdialenosti medzi dátami [4]. Hierarchické metódy využívajú matice, ktoré sú znázornené v podsekcii 3.3. Nie je potrebné zadať počet tried, ale je vhodné zadať ukončujúcu podmienku.

Existujú dva prístupy znázornené na obrázku 3.1:

- **aglomeratívne hierarchické metódy**
  - Zhlukovanie zdola-nahor.
  - V začiatočnej fáze je každý objekt považovaný za samostatný zhluk. Postupne spája najbližšie zhluky a vytvára tak hierarchickú štruktúru zhlukov. Následne sa najbližšie zhluky spájajú do nového zhluku a algoritmus sa opakuje, až do splnenia ukončujúcej podmienky alebo kým všetky zhluky nie sú spojené do jedného veľkého zhluku.
- **rozdeľujúce hierarchické metódy**
  - Zhlukovanie zhora-nadol.
  - Začína s jediným zhlukom obsahujúcim všetky objekty a postupne ho delí. V každom kroku sa zhluk rozdelí na dva nové zhluky, ktoré sa ďalej delia na menšie. Tento proces sa opakuje, až kým sa nedosiahne požadovaný počet zhlukov alebo až kým každý zhluk neobsahuje len jeden objekt.



Obr. 3.1: Hierarchické metódy.

Pri hierarchických metódach je výsledkom stromová štruktúra, kde listy predstavujú jednotlivé objekty a vrcholy reprezentujú zhluky. Vrcholy na rovnakej úrovni stromu tvoria jednu úroveň, a podobne vrcholy na úrovni nadradenej tvoria nadradenú úroveň [9]. Tieto metódy sa využívajú napríklad v prípadoch, kde chceme poznať vzájomný vzťah jednotlivých zhlukov na rôznych úrovniach.

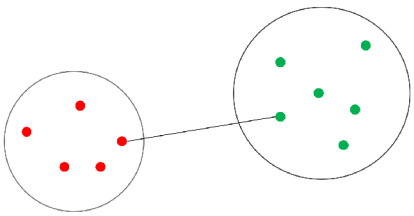
3.2.2 Metriky pre vzdialenosť medzi zhlukmi

Pri použití aglomeračnej alebo rozdeľujúcej metódy je kľúčové meranie vzdialenosti medzi dvoma zhlukmi. Štyri často využívané metriky pre túto vzdialenosť sú maximálna, minimálna, priemerná a stredná vzdialenosť, kde $\left|p - p'\right|$ predstavuje vzdialenosť medzi dvoma objektmi alebo bodmi, p a p'. Hodnota m_i je priemer pre zhluk c_i a n_i je počet objektov v zhluku c_i [8].

- **Minimálna vzdialenosť** (Single link)

  V tomto prípade sa podobnosť dvoch zhlukov meria na základe podobnosti medzi ich najbližšími členmi. Pozrieme sa na vzdialenosti všetkých dvojíc prvkov a ako vzdialenosť zhluku použijeme najmenšiu vzdialenosť. Táto metóda intuitívne priraďuje väčší význam oblastiam, kde sú zhluky najbližšie, a zanedbáva celkovú štruktúru zhluku. Preto táto metóda patrí medzi lokálne metódy založené na podobnosti. Jednou z hlavných nevýhod tejto metódy je však jej citlivosť na šum a odľahlé hodnoty v dátach. Vzorec na výpočet tejto vzdialenosti 3.1 vidíme nižšie.

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'| \tag{3.1}$$

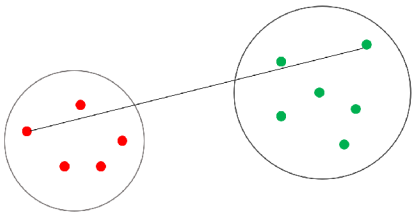


Obr. 3.2: Minimálna vzdialenosť.

- **Maximálna vzdialenosť** (Complete link)

  Táto metrika meria podobnosť dvoch zhlukov na základe najmenej podobných členov, berie teda maximálnu vzdialenosť. Táto metóda berie do úvahy celkovú štruktúru zhluku a má preto nelokálny charakter. Avšak, podobne ako v prípade jednoduchého spojenia, aj táto metóda je citlivá na odľahlé hodnoty. Nasledujúci vzorec 3.2 reprezentuje maximálnu vzdialenosť.

$$d_{max}(C_i, C_j) = max_{\ p \in C_i,\ p' \in C_j} \left| p - p' \right| \tag{3.2}$$

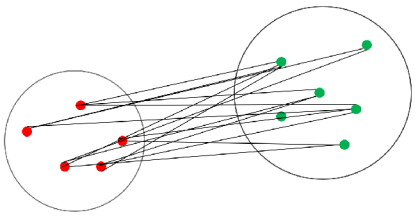


Obr. 3.3: Maximálna vzdialenosť.

- **Priemerná vzdialenosť** (Average link)

  Zohľadňuje podobnosť medzi všetkými bodmi v oboch zhlukoch. Vzdialenosť medzi dvomi zhlukmi je priemernou hodnotou všetkých dvojíc vzdialeností medzi bodmi v týchto dvoch zhlukoch. Na výpočet priemernej vzdialenosti sa využíva vzorec 3.3.

$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} \left| p - p' \right| \tag{3.3}$$

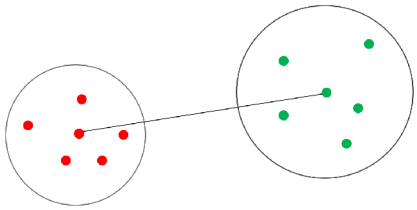


Obr. 3.4: Priemerná vzdialenosť.

- **Stredná vzdialenosť** (Centroid link)

  Táto metrika meria podobnosť medzi dvoma zhlukmi meraním vzdialenosti medzi centroidmi (stredovými bodmi) týchto zhlukov. Hlavný rozdiel medzi nimi je ten, že priemerná vzdialenosť berie do úvahy všetky dvojice dátových objektov pri výpočte priemerných párových podobností, zatiaľ čo zhlukovanie s centroidom používa iba centroid zhluku na výpočet podobnosti medzi dvoma rôznymi zhlukmi. Vzorec pre strednú vzdialenosť 3.4 je uvedený nižšie.

$$d_{mean}(C_i, C_j) = \left| m_i - m_j \right| \tag{3.4}$$



Obr. 3.5: Stredná vzdialenosť.

3.2.3 Kofenetický korelačný koeficient

Výber správnej metriky pre vzdialenosť medzi zhlukmi je veľmi dôležitý a môže významne ovplyvniť výsledky zhlukovania. Na určenie ideálnej metriky môže byť využitý kofenetický korelačný koeficient, ktorý meria ako vierohodne dendrogram zachováva dvojice vzdialeností medzi pôvodnými nespracovanými dátovými body. Je to štatistická miera, ktorá sa pohybuje od -1 do 1, pričom čím je bližšie k 1, tým je lepší model zhlukov a vyššia vierohodnosť. Tieto informácie boli prevzaté z [15] a [7].

3.3 Vstupné a výstupné dáta

Vstupnými dátami pre zhlukovanie sú objekty, popísané rôznymi atribútmi, ktoré majú byť rozdelené do skupín.

Zvyčajne sa jedná o numerické dáta, ktoré sú obvykle reprezentované vo forme dátovej matice 3.6 alebo vzdialenostnej matice 3.7.

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1t} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{it} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{nt} \end{bmatrix}$$

Obr. 3.6: Dátová matica (prevzaté z [10]).

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ . & . & . & & \\ . & . & . & & \\ . & . & . & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

Obr. 3.7: Vzdialenostná matica (prevzaté z [10]).

Výstup hierarchického zhlukovania sa najčastejšie vizualizuje pomocou dendrogramu. Pojem dendrogram sa používa v troch rôznych významoch: ako matematický objekt, dátová štruktúra a grafické zobrazenie týchto dvoch[12]. V tejto práci sa odvolávam na tretí význam, teda konkrétne grafickú reprezentáciu hierarchických zhlukovacích binárnych stromov, ktorá zobrazuje postupnú organizáciu zhlukovania.

Na X-ovej osi sa nachádza množina pozorovaní a na Y-ovej osi je vzdialenosť medzi zhlukmi. V dendrogramoch je možnosť vizuálne identifikovať objekty zoskupené do zhlukov a aká je medzi nimi vzdialenosť. Každý zhluk tvorí samostatnú vetvu na strome [4].

Kapitola 4

Dáta a úlohy

Touto kapitolou sa začína praktická časť bakalárskej práce, kde začínam pracovať so získanými dátami od vybranej obchodnej spoločnosti. Obsahuje opis týchto dát a oboznámenie sa s nimi. Nasledujú návrhy úloh využitia metód dolovania dát a pracovné postupy.

4.1 Opis dát

V tejto časti práce opisujem získané dáta vybranej obchodnej spoločnosti, ktoré sa skladajú zo štyroch dátových súborov. Prvý dataset obsahuje report ponúkaného sortimentu zo slovenského e-commerce portálu a porovnávača cien Heureka.sk, ktorý umožňuje zákazníkom porovnávať ceny a parametre produktov a nakupovať online. Ďalšie dva datasety obsahujú záznamy jednotlivých predajov pneumatík.

Prvý dataset je generovaný z portálu Heureka.sk a obsahuje 26 440 záznamov ponúkaných produktov s 35 stĺpcami. Ten funguje na princípe porovnávania cien a parametrov produktov, kde si zákazníci môžu vybrať konkrétny produkt a porovnať jeho cenu a parametre na rôznych e-shopoch. Tento report sa skladá z kľúčových údajov pre prevádzkovateľov e-shopu, ktorí sú zodpovední za cenotvorbu a plánovanie sezónneho predzásobenia. Tieto kľúčové údaje sú napríklad počet konkurentov ponúkajúcich produkt alebo porovnanie ceny e-shopu s trhom.

Ako vidíme nižšie v príklade záznamu, atribúty tohto datasetu sú segment, kategória, názov produktu, identifikačné číslo produktu, výrobca, EAN, eshopID, URL, počet eshopov predávajúcich tento produkt, popularita produktu na trhu, naša cena, pozícia analyzovaného e-shopu podľa ceny a 10 najvyšších a najnižších cien za produkt.

Príklad záznamu s vynechanými stĺpcami cien:

Segment, Kategória, Názov produktu, ID Produktu, Výrobca, EAN, Item ID, Vaša URL, Heureka URL, E-shopov predávajúcich produkt, Popularita produktu na trhu, Vaša cena, Vaša pozícia podľa ceny, Vaša pozícia podľa biddingu, Cena za preklik, Najnižšia cena, Najvyššia cena

Auto-moto, Osobné pneumatiky, Pirelli SCORPION WINTER 2 235/50 R18 101V, 538562912, Pirelli, 8019227422702, 737111, https://www.pneuprofi.sk/235-50r18-101v-pirelli-scorpion-winter-2-xl-737111, https://pneumatiky.heureka.sk/pirelli-scorpion-winter-2-235-50-r18-101v, 41, 4, €191,26, 23, 8,4, €0,200, €176,00, €215,09

V druhom datasete sa nachádzajú záznamy predajov z kamennej predajne autoservisu z účtovno-informačného systému MRP, ktorý ponúka všetky informácie potrebné pre vedenie účtovníctva, prevádzku a efektívne riadenie firmy. Nachádza sa v ňom 5 907 záznamov a 37 atribútov. Tabuľka obsahuje dve identifikačné čísla, názov produktu, cenu za mernú jednotku (MJ), kód, mernú jednotku, poznámku, 6 druhov cien, výrobcu, 6 druhov cien s DPH, číslo pohľadávky, dátum, číslo objednávky, spolu (SP) počet kusov, cena za kus, zľava, spolu s dph, sadzba dph, skladová cena, suma, celková suma, suma s dph, kód podľa charakteristiky pneu.

Príklad záznamu:

IDR, CISLO, NAZOV, CENAMJ, KOD, MJ, POZNAMKA, CENA1, CENA2, CENA3, CENA4, CENA5, CENA0, KOD2, CENAMJSDPH, CENA1SDPH, CENA2SDPH, CENA3SDPH, CENA4SDPH, CENA5SDPH, CISLOPOH, DATUM, SP_POCETMJ, SP_CENAMJ, SP_ZLAVA, SP_DPH, SP_SADZBADPH, SP_SKLADCENA, SP_UCETCASTKA, SP_CELKOM, SP_CELKOMSDPH, KOD3

2141488, 10448, P-MOTO PIR 200/55 ZR17 (78W) M/C TL Diablo Rosso Corsa2 R, 142,24 , 2907500, ks, 269, 146,5083, 152,2, 163,575, 153,6167, 118,32, PIRELLI, 170,69, 322,8, 175,81, 182,64, 196,29, 184,34, 2211219, 15.07.2022, 1, 163,6417, 0, 32,7283, 20, 134,18, 134,18, 163,64, 196,37, 2005517

Dáta tretieho datasetu pochádzajú z internetového obchodu tejto firmy, jedná sa o záznamy uskutočnených predajov. Tento dataset sa skladá z 6 020 riadkov a 20 stĺpcov. Každý záznam tohto datasetu obsahuje číslo objednávky, dátum a čas objednávky, identifikačné číslo produktu, identifikačné číslo skladu v ktorom sa produkt nachádza, počet kusov, názov položky, cenu s DPH za kus, zvolený sklad, kód produktu, výrobcu, index nosnosti (LI), index rýchlosti (SI), šírku, priemer, výšku, segment, sezónu a hlučnosť. V prípade tohto datasetu sú ku každej objednávke priradené ďalšie záznamy s informáciami o doručení objednávky, ktoré je potrební neskôr odstrániť.

Príklad záznamov:

Číslo obj., Vytvorené v, export-o-product-id-product, export-o-product-id-product-storage, Počet ks., Názov položky, Predaj s DPH/ks, Zvolený sklad, Produkt: Kód, Názov výrobcu, LI, SI, Šírka, Priemer, Výška, Segment, Sezóna, Hlučnosť

202204390, 27.3.2023, 16:28, 674235, 139, 1, 215/65 R16 98V Fulda ECOCONTROL HP 2, 79,61, Jakub Nové Mesto, 542642, Fulda, 98, V, 215, 16, 65, osobne, letne, 70

202204390, 27.3.2023 16:28„,1, Osobne„„„„„„„

202204390, 27.3.2023 16:28„, 1, Osobný odber - Nové Mesto nad Váhom„„„„„„„

Dáta druhého a tretieho datasetu predstavujú predaje pneumatík v období od 01.07.2022 do 31.3.2023.

Dáta štvrtého poskytnutého súboru obsahujú len 3 stĺpce ale 254 739 riadkov. Z dôvodu nesúhlasných identifikačných čísel medzi predošlými tabuľkami bola poskytnutá tabuľka na pomoc pri mapovaní týchto identifikačných čísel. Táto tabuľka obsahuje názov výrobcu, EAN a kód.

Príklad záznamu:

Výrobca, EAN, KOD

Barum, 4024063616134, 15405850000

4.2 Porozumenie dátam

Z názvu pneumatiky dokážeme zistiť viaceré dôležité informácie a parametre. Napríklad pneumatika **Nexen Winguard Snow'G3 WH21 185/60 R14 82T** je určená pre osobné vozidlá a jej parametre znamenajú:

- Nexen - názov značky pneumatiky
- Winguard Snow'G3 WH21 - konkrétny typ pneumatiky
- 185/60 - Číslo 185 v tomto označení znamená šírku dezénu pneumatiky v milimetroch. To znamená, že dezén tejto pneumatiky má šírku 185 mm. Číslo 60 v označení znamená pomer výšky pneumatiky k jej šírke. To znamená, že výška pneumatiky je 60% z jej šírky, čo znamená 111 mm.
- R14 - Písmeno R označuje konštrukciu pneumatiky. R znamená radiálnu konštrukciu, čo znamená, že základ pneumatiky tvorí radiálna vrstva. Väčšina moderných pneumatík na trhu je práve s radiálnou konštrukciou, ktorá sa vyznačuje vysokou úrovňou stability a odolnosti. Číslo 14 v označení označuje priemer ráfika v palcoch. V tomto prípade je to 14 palcov.
- 82 - Číslo 82 označuje index nosnosti (LI - Load Index) pneumatiky. Index nosnosti udáva maximálnu hmotnosť, ktorou môže byť daná pneumatika zaťažená. Pneumatika Nexen Winguard Snow'G3 WH21 s indexom nosnosti 82 môže byť zaťažená maximálnou hmotnosťou 475 kg.
- T - Písmeno T je označenie rýchlostného indexu (SI - Speed Index). Značí, že táto pneumatika je konštrukčne určená na maximálnu rýchlosť 190 km/h. Rýchlostný index T je používaný pre vozidlá nižšej strednej kategórie. Písmená sa pohybujú od A do Y a každé písmeno označuje určitú maximálnu rýchlosť v km/h.

Je dôležité si uvedomiť, že pneumatika s nižším indexom nosnosti alebo rýchlosti, než je odporúčané pre dané vozidlo, môže znížiť bezpečnosť. Na druhej strane, pneumatika s vyšším indexom nosnosti alebo rýchlosti, než je potrebné, nie je nutná a môže byť zbytočne drahá. Preto je dôležité dbať na správnu voľbu pneumatík s ohľadom na index nosnosti a maximálnu povolenú rýchlosť.

4.3 Posúdenie využiteľnosti rôznych metód

Táto sekcia posudzuje jednotlivé metódy spomenuté v teoretickej časti 2.4 a ich využitie na poskytnuté dáta.

- **Klasifikácia**
  - Metóda klasifikácie by sa dala použiť napríklad na klasifikáciu zákazníkov do vopred určených skupín, na základe rôznych parametrov. Pre účel tejto práce neboli poskytnuté zákaznícke dáta, kvôli ich citlivosti. Avšak spoločnosť, s ktorou som spolupracovala, disponuje vhodnými dátami, ako sú typy zákazníkov a ich špecifické požiadavky a vlastnosti.

    Ak by spoločnosť mala záujem v budúcnosti prehĺbiť možnosti získavania znalostí z týchto dát, metóda klasifikácie by bola vhodnou cestou.
- **Asociačné pravidlá**
  - Asociačné pravidlá sa dajú využiť pri určovaní potenciálu predaja doplnkových produktov a služieb. Vybraná spoločnosť disponuje týmito dátami, avšak primárnym zameraním spoločnosti je predaj pneumatík a teda dáta poskytnuté pre účely tejto práce sa sústreďujú len na tento typ produktu.
- **Predikcia**
  - Používa sa na predpovedanie budúcich trendov a udalostí na základe histórie dát. Výhodou predikcie v oblasti obchodu a predaja je, že umožňuje efektívne riadenie zdrojov a nákladov a pomáha minimalizovať straty spôsobené nesprávnym zásobovaním. Tento druh analýzy by sa dal využiť v prípade, že by boli dostupné dáta z dlhšieho časového úseku. Poskytnuté dáta to neumožňujú, vzhľadom k tomu, že neobsahujú dostatok hodnôt pre stanovenie cyklických vplyvov.
- **Regresia**
  - V prípade uchovávania informácií o zákazníkoch a ich vzorcoch nákupného správania v minulosti, by mohla byť využitá metóda regresie. Výstupom tejto metódy je identifikácia kúpneho potenciálu zákazníkov v budúcnosti.
- **Detekcia odľahlých objektov**
  - V prípade predajných dát môže byť detekcia odľahlých objektov použitá napríklad na identifikáciu produktov, ktorých počty predajov sú výrazne odlišné od bežných hodnôt. Tieto odchýlky môžu byť dôsledkom sezónnych faktorov. Využitie tejto metódy pri získaných dátach nebolo opodstatnené z dôvodu obmedzeného objemu dát.

- **Sekvenčné vzory**
  - Sekvenčné vzory nám dokážu pomôcť identifikovať vzory a trendy v sekvenciách udalostí, ako napríklad kúpa jedného konkrétneho produktu môže spôsobiť potrebu nákupu naväzujúcich produktov v určitých časových intervaloch. Analýza takýchto minulých sekvenčných udalostí môže spoločnosti dopomôcť v tvorbe správne cielených marketingových a predajných stratégii. Sekvenčné vzory sa nedali využiť pri analýze dostupných dát z dôvodu absencie prepojenia predajov na jednotlivých zákazníkov.
- **Zhlukovanie**
  - Zhlukovacie metódy pomáhajú zoskupovať dáta a identifikovať skupiny a vzorce správania, na základe podobných vlastností. To môže pomôcť pri analýze dát a vytváraní predpovedí na základe podobnosti skupín. Táto metóda je vhodná na analýzu dostupných dát a bola zvolená pre implementáciu.

4.4 Pracovné postupy

Z viacerých možností dolovania dát, ktoré sú spomenuté v teoretickej časti, bola pre praktickú aplikáciu zvolená metóda zhlukovania.

Po oboznámení sa so získanými dátovými sadami, vyplynulo vzhľadom na vlastnosti dát, že najlepšou metódou bude hierarchické zhlukovanie. Ostatné metódy získavania znalostí z dát neboli vyhovujúce na naše účely z dôvodu nedostatku relevantných informácii, ako napríklad špecifické informácie o zákazníkoch.

Úvodom praktickej časti tejto bakalárskej práce je predspracovanie a príprava dát. Následne analýza dát a grafické vykreslenie predajnosti v danom časovom úseku podľa určených produktových kategórii a taktiež výsledky zhlukovania pomocou dendrogramov.

Jednou z analytických úloh je analýza vplyvu pozície ceny v konkurenčnom prostredí na objemy predajnosti daných produktov.

Výsledky vyššie spomenutých krokov sú následne spracované do jednoduchého a prehľadného užívateľského rozhrania pre efektívne využívanie znalostí získaných z dát. Takéto zjednodušenie umožňuje vedeniu a taktiež zamestnancom firmy prispôsobiť zobrazenia výsledkov analýz podľa potrieb jednotlivých funkcií.

Kapitola 5

Implementácia

V tejto časti bakalárskej práce by som rada predstavila konkrétne postupy implementácie praktickej časti. Hlavnými prvkami tejto implementácie sú skripty na spracovanie predajných dát do vhodnej formy pre následnú analýzu a tvorbu nástroja s užívateľským rozhraním pre vytváranie a vizualizáciu analýz s možnosťou prispôsobenia filtrov. Tento nástroj umožňuje analýzu a získavanie znalostí z predajných dát na pravidelnej báze. Užívatelia môžu jednoducho prispôsobiť výsledky analýzy a ich vizualizáciu podľa požadovaných kritérií, ako napríklad výber typu vozidla, značky alebo iných parametrov produktu.

Táto implementácia využíva technológie ako Python, MySQL, HTML a CSS. Ako hlavnú platformu na analýzu dát využívam Python, pretože poskytuje rýchle a efektívne nástroje na manipuláciu a vizualizáciu dát.

Využitie Pythonu na analýzu zahŕňa použitie niekoľkých knižníc a nástrojov ako napríklad:

- **Pandas**[1] - Táto knižnica sa využíva na manipuláciu s dátami a poskytuje možnosť načítavať dáta z rôznych zdrojov, ako napríklad CSV a spracovať ich do štruktúry DataFrame. Táto výkonná a flexibilná štruktúra umožňuje efektívne ukladanie, analýzu, vizualizáciu a spracovanie aj veľkých objemov dát [13].
- **NumPy**[2] - Ďalej sa využíva NumPy (Numeric Python) knižnica, ktorá poskytuje nástroje na prácu s numerickými dátami a výpočtami.
- **SciPy**[3] - Knižnica SciPy (Scientific Python) obsahuje niekoľko numerických algoritmov, ktoré pracujú s Numpy. SciPy obsahuje funkcie pre lineárnu algebru, optimalizácie, matice, algoritmy na diferenciálne rovnice, integráciu, Fourierove transformácie, štatistické funkcie a špeciálne matematické funkcie [13].
- **Plotly**[4] - Plotly je nástrojom na vizualizáciu dát v Pythone. Umožňuje tvoriť rôzne typy interaktívnych grafických vizualizácií. V implementácii sa taktiež využíva modul Plotly Express, ktorý je súčasťou knižnice Plotly a poskytuje ešte jednoduchšie a intuitívnejšie možnosti tvorby vizualizácií.
- **SQLAlchemy**[5] - SQLAlchemy je nástroj pre jazyk Python, ktorý umožňuje prácu s rôznymi relačnými databázovými systémami, ako napríklad MySQL.

[1] https://pandas.pydata.org
[2] https://numpy.org
[3] https://scipy.org
[4] https://plotly.com
[5] https://sqlalchemy.org

V postupe som sa zamerala na návrh a implementáciu užívateľského rozhrania, vytvorenie databázovej štruktúry a analýzy dát. Riešenie obsahuje 4 Python skripty, ktoré je možné spustiť z príkazového riadku. Pre spustenie stačí otvoriť príkazovú konzolu a zadať názov príkazu. Podrobnejší opis spustenia sa nachádza v prílohe A. V tejto kapitole podrobne popíšem postup implementácie každého zo skriptov, ako aj spôsob ich využitia na dosiahnutie stanovených cieľov tejto práce.

5.1 Predspracovanie dát

Spracovanie dát je kľúčovou časťou implementácie. V tejto sekcii podrobne opíšem postup a metódy, ktoré som využila na spracovanie vstupných dát a ich transformáciu do formátu vhodného pre ukladanie v databáze a následnú analýzu.

Na prípravu dát sú použité konkrétne dva skripty, kde prvý skript `eshopData.py` spracováva vstupné dáta predajov z internetového obchodu, ktoré sa nachádzajú v súbore `eshop_data.xlsx` a druhý `mrpData.py` z kamennej predajne s dátami v `mrp_data.xlsx`.

Súbory so vstupnými dátami sú vo formáte XLSX načítané pomocou `pandas.read_excel` do štruktúry `Pandas Dataframe`, ktorá poskytuje nástroje na manipuláciu s tabuľkovými dátami, ktoré sú pre náš účel vhodné. Cieľom bolo zabezpečiť, aby všetky vstupné dáta boli správne načítané. Keďže každý z týchto dátových súborov pochádza z iného zdroja, museli byť pri jednotlivých dátach použité trochu odlišné procesy, aby oba súbory boli upravené a prevedené do jednotného formátu, ktorý by umožnil ich ďalšie spracovanie.

Po vytvorení `DataFrame` sú aplikované procesy na čistenie dát, aby sa zaistilo, že dáta sú správne a relevantné pre potreby analýz. Tento proces zahŕňa napríklad odstraňovanie nepotrebných a nadbytočných dát, kontrolu dátových typov a formátov a odstraňovanie odľahlých hodnôt. Atribúty dát, ktoré sa využívajú, sú premenované na jednoznačné názvy.

Následne sú vo funkcii `setPriceClass` do stĺpca `priceClass` pridané cenové kategórie produktov na základe značiek.

Ďalej je vytvorená funkcia `setParameters`, ktorá vďaka regulárnemu výrazu `(\d{2,3})/(\d{2,3})\s+R(\d{2})\s+(\d{2,3})([A-Z])` získa z názvu produktu konkrétne parametre pneumatiky a uloží ich do príslušných stĺpcov. Tento regulárny výraz sa trochu líši pre každý súbor, vzhľadom k tomu, že názvy sú v iných formách. Pre jednoduchšie získavanie týchto parametrov v prípade dát z internetového obchodu, je potrebné namapovať produkty na dataset z Heureky a prepísať názvy produktov predajných dát z internetového obchodu na názvy z datasetu Heureky, ktoré majú štandardizovanú formu.

Vo funkcii `seType` sa do stĺpca `type` pridáva typ vozidla pre ktoré je určený konkrétny produkt. V prípade dát z kamennej predajne, zisťujem tento údaj z názvu, v ktorom sa nachádza aj typ vozidla. Keďže údaj o type vozidla nie je uvedený pri všetkých predajných dátach z internetového obchodu, získavam tento údaj z poslednej číslice parametru šírky pneumatiky. Ak toto číslo končí číslicou 5, je to pneumatika určená pre automobily a v prípade, že končí 0, jedná sa o motocyklovú pneumatiku.

Pri predajných dátach z kamennej predajne je potrebné pridať niektoré chýbajúce názvy značiek, ktoré sa dajú odvodiť a vybrať z poľa obsahujúceho názov produktu.

Ďalšou dôležitou úpravou je úprava dátumu do formátu `%Y-%m` , vďaka ktorej dokážeme napokon vyvodiť závery a vzorce predajnosti na mesačnej báze.

Posledným krokom prípravy dát je agregácia množstva predaných kusov konkrétnych produktov za určitý mesiac.

Vo výslednej tabuľke dát z kamennej predajne je potrebné ponechať aj stĺpec `KOD`, ktorý sa neskôr využije pri mapovaní na dataset z Heureky na analýzu pozície ceny 5.2.1.

Na záver sa spracované dáta v požadovanej forme, ktorú je možné vidieť v tabuľke 5.1, uložia do vopred pripravenej MySQL databázy `pneuProfi` na lokálnom počítači pre ďalšie analytické účely. Každý záznam týchto predpripravených predajných dát obsahuje dátum predaja, ID predanej položky, názov produktu, počet predaných kusov, značku, cenovú kategóriu značky, typ vozidla a parametre pneumatiky. Pri niektorých záznamoch boli navyše aj ďalšie atribúty, poznámky k produktom, rôzne ceny a iné kategórie, avšak tieto údaje boli odstránené pri predspracovaní dát.

Atribúty záznamov v datasetoch s predajnými dátami sú:

- **date** - dátum konkrétneho nákupu vo formáte `%Y-%m`
- **productID** - identifikačné číslo produktu, ktoré sa nachádza vo všetkých datasetoch
- **name** - celý názov produktu, ktorý obsahuje dôležité parametre
- **pieces** - počet predaných kusov tohto produktu
- **brand** - značka produktu
- **priceClass** - cenová kategória značky predaného produktu
- **type** - typ vozidla (automobil alebo motocykel), pre ktorý je pneumatika určená
- **width** - šírka dezénu pneumatiky
- **height** - pomer výšky pneumatiky k jej šírke
- **R** - priemer ráfika v palcoch
- **LI** - index nosnosti
- **SI** - index rýchlosti

productID	date	pieces	name	brand	priceClass	type	width	height	R	LI	SI
543817	2022-09	1	Pirelli ANGEL GT 180/55 R17 73W	PIRELLI	1	moto	180	550	R17	73	W
543608	2022-10	4	Sava ESKIMO S3 + 195/65 R15 91T	Sava	3	auto	195	65	R15	91	T
510269	2023-01	2	Nexen N'blue HD Plus 155/70 R13 75T	Nexen	2	auto	155	70	R13	75	T

Tabuľka 5.1: Príklad záznamov dát

5.2 Analýza dát

V tejto sekcii sa budeme venovať implementácii analýzy dát. Po zbere a predspracovaní dát je analýza ďalším krokom v procese získavania užitočných informácií z datasetov.

5.2.1 Analýza pozície ceny

Prvou analytickou úlohou je zisťovanie vplyvu pozície ceny v konkurenčnom prostredí na predajnosť produktov. Riešenie tejto úlohy sa nachádza v skripte `position.py`. Pri tejto úlohe využívam pripravené dáta v databáze, tabuľku štatistík z portálu Heureka `heureka.xlsx`, ktorá je opísaná v sekcii 4.1 a pomocnú tabuľku `codes.xlsx` na mapovanie identifikačných čísel. Keďže každý vstupný súbor využíva iné identifikačné čísla produktov, je potrebné ich pre túto analýzu namapovať na tabuľku dát z Heureky, aby sme mohli priradiť do záznamov predajov stĺpec s konkrétnou pozíciou ceny. V prípade dát z internetového obchodu sa identifikačné čísla `productID` mapujú na `itemID` v tabuľke Heureky.

Pri dátach z kamennej predajne využívame identifikačné číslo `KOD` na namapovanie EAN kódu pomocou tabuľky `codes.xlsx` a následne vďaka `EAN` priradíme príslušnú pozíciu ceny.

Následne sa zoskupia dáta podľa pozície ceny a spočíta sa celkový počet predaných kusov produktov.

Nakoniec výsledky vizualizujeme pomocou grafov a knižnice Plotly.

5.2.2 Analýza predajnosti podľa kategórii

Ďalšie analýzy sa nachádzajú v Python súbore `app.py`, ktorý spúšťa užívateľské rozhranie a uskutočňuje požadované analýzy. V tejto časti je možné analyzovať predajné dáta v časovom období od júla 2022 do apríla 2023 podľa mesiacov. Dáta sa tu zoskupujú podľa potreby vybranej analytickej kategórie a mesiaca v roku. Výsledky týchto analýz sú taktiež vizualizované pomocou grafov, pre prehľadnú orientáciu v záveroch. Možné analytické kategórie sú napríklad cenové kategórie značiek, ich výrobcovia alebo rôzne parametre a vlastnosti pneumatík, ako šírka dezénu pneumatiky, radiálna konštrukcia, index nosnosti a index rýchlosti. Tieto kategórie boli stanovené po dohode so zastupiteľmi vybranej obchodnej spoločnosti. Rovnako je možné tieto analýzy prispôsobiť podľa požiadaviek jednotlivca, ktorý si jednoducho označí požadované vstupy vo vytvorenom užívateľskom prostredí.

5.3 Zhlukovanie dát

Ako už bolo vyššie spomenuté, metóda hierarchického zhlukovania sa na základe charakteru dostupných dát ukázala ako najvhodnejší spôsob získavania znalostí. Keďže táto metóda je založená na zoskupovaní dát do zhlukov podľa podobnosti, rozhodli sme sa ju aplikovať a pomocou nej identifikovať opakujúce sa vzory v správaní zákazníkov v rôznych časových obdobiach.

Ako vidíme v ukážke kódu 5.1, ktorá sa taktiež nachádza v súbore `app.py`, v prvom kroku bolo nutné zoskupiť dáta podľa indexu pomocou funkcie `groupby()` z knižnice Pandas. Výsledkom tejto operácie sú zoskupené dáta podľa hodnôt `priceClass` a `date` a ich súčet v stĺpci `pieces`. Ďalej sa vytvára matica 5.1 pre dendrogram pomocou funkcie `pivot_table()` s hodnotami `pieces`, indexom `priceClass` a stĺpcami `date`.

Ďalej sa využijú funkcie `pdist`, `linkage` a `cophenet` z knižnice `SciPy`. Vytvorí sa vzdialenostná matica, vďaka ktorej sa vypočíta kofenetický korelačný koeficient. Funkcia `cophenet` vracia hodnotu kofenetickej korelačnej vzdialenosti a vzdialenostnú maticu.

Nakoniec sa pomocou funkcie `create_dendrogram()` z knižnice `Plotly Express` vygeneruje dendrogram z dát v matici `dendro_data`. Tento dendrogram je následne zobrazený pomocou funkcie `show()`.

V ukážke kódu 5.1 sa využíva ako metrika pre vzdialenosť medzi zhlukmi maximálna vzdialenosť (complete), avšak pri zhlukovaní je možnosť si vybrať aj inú z metód opísaných v sekcii 3.2.2.

```
# zoskupenie podla indexu
df_grouped = salesby_priceclass.groupby(['priceClass', 'date'])
       ['pieces'].sum().reset_index()

# vytvorenie tabulky pre zhlukovanie
dendro_data = df_grouped.pivot_table(values='pieces',
       index='priceClass', columns='date')

# vytvorenie matice vzdialenosti a vypocet kofenetickeho koeficientu
distances = pdist(dendro_data.transpose(), metric="euclidean")
linkage_matrix = linkage(distances, method="average")
c, coph = cophenet(linkage_matrix, distances)


# vytvorenie dendrogramu
fig = ff.create_dendrogram(dendro_data.T, orientation='bottom',
       labels=dendro_data.columns,
       linkagefun=lambda x: linkage(x, method='average'))

fig.update_layout(title='Dendrogram predajnosti v mesiacoch ')

fig.show()
```

Výpis 5.1: Príklad vizualizácie hierarchického zhlukovania

date	2022-07	2022-08	2022-09	2022-10	2022-11	2022-12	2023-01	2023-02	2023-03
priceClass									
1	590	517	543	865	684	361	328	487	1133
2	136	118	253	816	866	418	167	251	342
3	553	457	746	920	1476	676	325	295	557
4	9	17	31	55	70	75	18	18	14

Obr. 5.1: Matica pre hierarchické zhlukovanie (dendro_data)

5.4 Tvorba užívateľského rozhrania

Vytváranie užívateľského rozhrania sa nachádza v Python súbore `app.py`, ktorým sa dá spustiť nástroj na analýzu. Pri tvorbe som využívala Dash[6], open-source framework na tvorbu webových rozhraní pre vizualizáciu dát. Dash kombinuje niekoľko knižníc, ako je napríklad Plotly pre tvorbu grafov a vizualizácií a Flask pre vytváranie webových stránok.

Dash sa skladá z dvoch hlavných častí: layoutu a callbackov. Layout definuje, ako sa aplikácia zobrazuje. Layout obsahuje rôzne prvky ako napríklad grafy, tabuľky, polia, tlačidlá atď., ktoré sa zobrazujú na stránke. Callbacky sú funkcie, ktoré sa spúšťajú, keď sa zmení hodnota niektorého vstupného prvku v aplikácii.

Aplikácia obsahuje dropdown menu kde je na výber analyzovaná kategória a na boku stránky sú možnosti na prispôsobenie analýzy. Táto aplikácia má tiež definovaný callback, ktorý mení zobrazovaný graf na základe výberu z menu.

[6] https://dash.plotly.com

Kapitola 6

Vyhodnotenie

V tejto sekcii identifikujem získané poznatky z dát a hodnotím ich význam a prínos pre vybranú obchodnú spoločnosť. Pri vyhodnocovaní získaných poznatkov z dát je dôležité zohľadniť aj ich praktickú aplikovateľnosť. Zistenie súvislostí a vzorcov v dátach môže byť užitočné najmä vtedy, ak dokážeme tieto poznatky preniesť do praktických opatrení a zlepšiť tak výkonnosť alebo dosiahnuť ciele v konkrétnej oblasti.

6.1 Vyhodnotenie spracovania a čistenia dát

Najzdĺhavejšou časťou spracovania dát bolo pri praktickej časti oboznámenie sa s dostupnými dátami, nájdenie súvislostí a vzťahov medzi dátami a ich úprava do použiteľnej formy. Príprava dát je základom pre analýzu a hľadanie poznatkov z dát. So správnym spracovaním a čistením je spojená kvalita a efektívnosť všetkých nasledujúcich krokov. Z tohto dôvodu je kľúčové venovať dostatočnú pozornosť ich pochopeniu a príprave a vyhnúť sa tak možným chybám alebo nedostatkom.

Kvalita výstupov znamená, že výsledky analýzy a vizualizácie dát sú porovnávané s očakávanými výsledkami a sú posudzované na základe ich presnosti a relevancie.

Vizualizácia dát napomáha lepšiemu porozumeniu získaných znalostí.

V prípade dát, ktoré boli poskytnuté pre účely tejto práce bolo nevyhnutných mnoho úprav. Medzi hlavné problémy daného datasetu boli:

- **Neprehľadnosť** - pri viacerých atribútoch nebolo možné určiť význam obsiahnutých informácií a ich samotné pochopenie vyžadovalo viaceré konzultácie so zastupiteľmi obchodnej spoločnosti, pre ktorých bola taktiež náročná orientácia v dátach.
- **Nepresnosť a neúplnosť** - záznamy boli v mnohých prípadoch nepresné alebo úplne absentujúce, chýbali kľúčové informácie, ktoré sa museli dohľadať v iných datasetoch pre poskytnutie celistvého obrazu o dátach.
- **Dáta nezodpovedajúce atribútom** - bolo očividné, že v niektorých prípadoch nie je jasne definovaný požadovaný vstup atribútu a vkladané údaje boli rôznorodého charakteru a nemali medzi sebou žiadne súvislosti.
- **Nadbytočné dáta** - viaceré stĺpce obsahovali rovnaké dáta alebo z nich vyplývali rovnaké informácie.
- **Nekonzistentnosť** - taktiež boli zaznamenané prípady vstupov, ktorých obsah bol síce správny, avšak spôsob ich formulácie bol odlišný. Dalo by sa predpokladať, že

zamestnanci nemajú jednotný spôsob zadávania údajov do systémov, čo vyžaduje značné množstvo času na ich zjednotenie pre využitie na následnú analýzu.

- **Zbytočné atribúty** - z dostupných dát vyplynulo, že mnohé atribúty nie sú vôbec využívané, keďže neobsahujú žiadne údaje.

6.2 Vyhodnotenie analýz dát

Nasledujúcim krokom po príprave dát je analýza, ktorá má za úlohu odhaliť:

1. zákaznícke správanie v určitých časovým obdobiach a jeho koreláciu so stanovenými kategóriami produktov
2. cenovú senzitivitu zákazníka

Cieľom týchto konkrétnych analýz je získať užitočné informácie o predajnosti produktov a poskytnúť ich pre rozhodovanie a zlepšenie stratégie spoločnosti.

6.2.1 Predajnosť v čase

Z možných analytických kategórií, ktoré sú spomenuté v podsekcii implementácie 5.2.2, bola vybraná na účel priblíženia zvoleného analytického procesu predajnosť produktov v závislosti na vopred určených cenových kategóriách značiek.

Do cenových kategórií boli roztriedené predávané značky produktov, podľa ich kvalitatívnych a cenových vlastností:

1. **Premium (1)** - tieto značky pneumatík patria do najvyššej kategórie. Ich ceny sú najvyššie a sú známe svojou vysokou kvalitou, výkonom a odolnosťou.
2. **Medium (2)** - tieto značky pneumatík patria medzi strednú až vyššiu kategóriu. Tieto pneumatiky ponúkajú kvalitné materiály a vysokú úroveň výkonu a bezpečnosti.
3. **Budget (3)** - tieto značky pneumatík sú vhodné pre bežné vozidlá a ponúkajú dobrý pomer ceny a kvality.
4. **Low-cost (4)** - tieto značky pneumatík sú najlacnejšie na trhu. Tieto pneumatiky sú vhodné pre bežné vozidlá, ale majú nižšiu kvalitu a výkon ako pneumatiky v iných kategóriách.

Toto rozdelenie sa vzťahuje hlavne na automobilové pneumatiky, vzhľadom k tomu, že vybraná spoločnosť ponúka svojim zákazníkom len prémiovú kategóriu pneumatík určených pre motocykle.

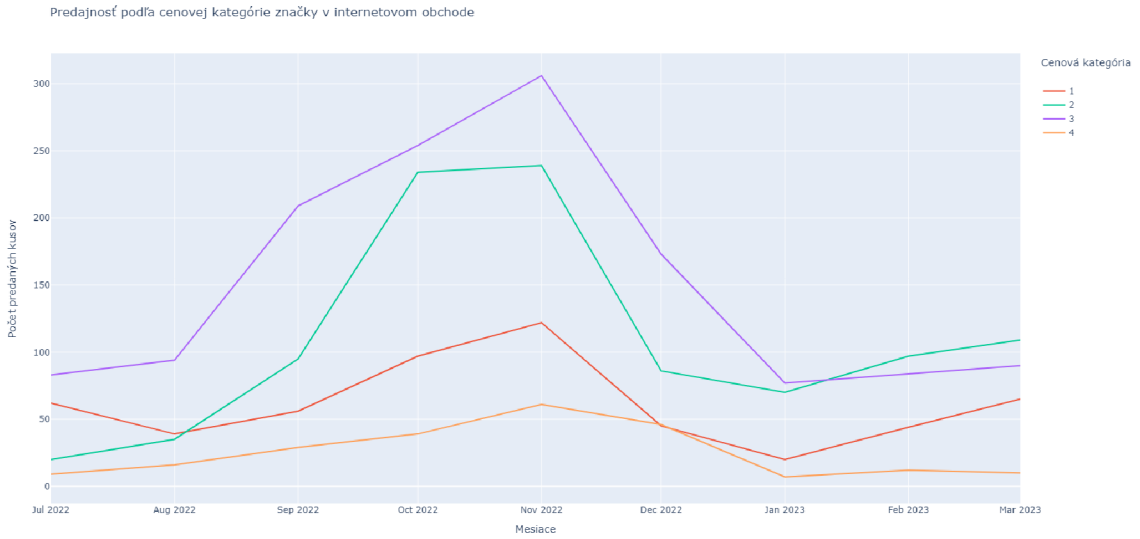
Na prehľadnú vizualizáciu výsledkov som využila zobrazenie pomocou grafu, ktorý vykresľuje počty predajov produktov v jednotlivých cenových kategóriách v danom časovom horizonte. Pre vytvorenie relevantných záverov z tejto analýzy, sme odfiltrovali pneumatiky určené pre motocykle. V opačnom prípade by skresľovali pohľad na prémiovú kategóriu pneumatík.

Ako je možné vidieť na grafe 6.1 predajov internetového obchodu, Low-cost je najmenej žiadanou kategóriu, z čoho vyplýva, že ľuďom záleží na kvalite a výkone pneumatík a cena nie je primárnym faktorom pri rozhodovaní sa o kúpe.

Pneumatiky z kategórií low-cost a prémium nie sú až tak výrazne ovplyvnené a nepodliehajú sezónnym vplyvom ako zvyšné dve kategórie. Ako môžeme jednoznačne vidieť pri

kategóriách Medium a Budget, ktoré sa nachádzajú uprostred škály, tieto pneumatiky majú výraznejšie výkyvy predajnosti v hlavných prezúvacích sezónach, ako je jeseň a jar.

Najvýraznejší pokles predajov pneumatík v internetovom obchode je v mesiacoch december a január, čo súvisí samozrejme s koncom prezúvacej sezóny. Ďalším faktorom, ktorý môže vplývať na tento značný pokles môže byť, že zákazník prioritizuje iné výdavky v tomto období.



Obr. 6.1: Predajnosť podľa cenovej kategórie v internetovom obchode

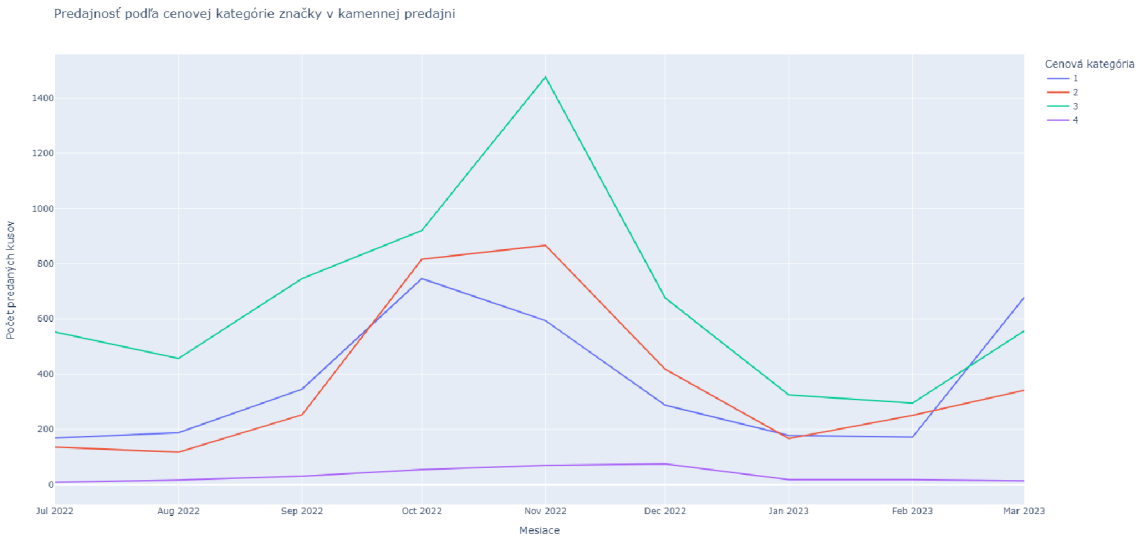
Nasledujúci graf 6.2 zobrazuje rovnaké faktory ako v predošlej analýze, ale jedná sa o predajnosť pneumatík z kamenného obchodu.

Môžeme vidieť veľmi výrazný vplyv prezúvacej sezóny na objem predaných produktov. Toto je pravdepodobne spôsobené aj možnosťou využitia služby prezutia pneumatík v kamennej predajni. Naznačuje to, že predaje na začiatku zimnej sezóny sú intenzívnejšie v porovnaní so začiatkom letnej sezóny prezúvania. V zime neprezutie na zimné pneumatiky môže spôsobiť výrazné zhoršenie jazdných vlastností vozidiel, zníženie bezpečnosti a priame ohrozenie účastníkov cestnej premávky. Taktiež je uzákonená povinnosť používania zimných pneumatík v období od 15.11. do 31.3. v prípade súvislej vrstvy snehu alebo ľadu na vozovke.

Zatiaľ čo obdobie prezúvania pneumatík na letnú sezónu je skôr pozvoľné, vzhľadom na to, že vodiči nemajú legislatívne stanovenú podmienku prezutia a taktiež sa neprezutím na letné pneumatiky nevystavujú takému nebezpečenstvu, ako je to v prípade zimnej sezóny.

Prémiové, a teda najdrahšie pneumatiky majú približne stabilný objem predajov počas celého roka, aj keď je vidieť nárast počas sezóny prezúvania. Nakoľko zákazníci, ktorí nakupujú túto kategóriu pneumatík, uprednostňujú kvalitu a výkon kvôli svojej bezpečnosti a pohodliu, sú ochotní si priplatiť za tieto faktory kedykoľvek. Sú to napríklad zákazníci s vysokým kilometrovým nájazdom, ktorí z dôvodu rýchlejšieho opotrebenia, vyžadujú výmenu pneumatík za nové v častejších intervaloch, ktoré nemusia nutne podliehať sezónnym trendom.

Cenovo senzitívnejší zákazníci aj v prípade nákupu v kamennom obchode radšej volia cenové kategórie ako Medium a Budget, ktoré poskytujú dobrý pomer ceny a kvality. Je tu



Obr. 6.2: Predajnosť podľa cenovej kategórie v kamennom obchode

vidieť štandardný trend sezónneho prezúvania, kedy sa pri prezutí zistí opotrebenie starých pneumatík a vyplynie potreba kúpy nových.

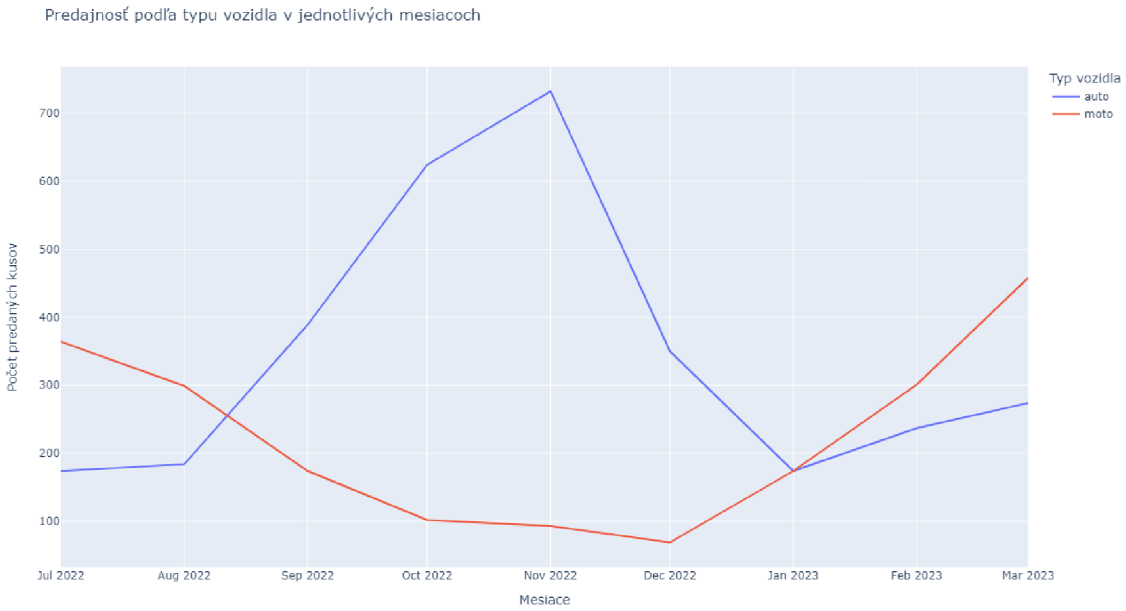
Rozdiely medzi predajmi internetového obchodu a kamennej predajne:

Na prvý pohľad sa môže zdať, že vzorce predaja z oboch zdrojov sú podobné, avšak pri detailnejšom skúmaní, sme identifikovali následné odlišnosti.

V prípade internetového obchodu vidíme väčšie objemy predaja v kategórii Low-cost, čo je ovplyvnené napríklad predajnou stratégiou v kamennom obchode a faktom, že zákazníci, ktorí si zvolia nákup v kamennej predajni často radi využijú poradenstvo personálu, ktoré dokáže objasniť výhody nákupu pneumatík z vyšších kategórií.. Avšak Low-cost stále zostáva najmenej populárnou kategóriu aj v prípade internetového obchodu.

Napriek tomu, že obe stredové kategórie sú najviac ovplyvnené sezónnosťou, vidíme, že v prípade internetového obchodu sú objemy predajov v týchto kategóriách podobnejšie ako v kamenných predajnich.

Ako môžeme vidieť v grafe 6.3, ktorý zobrazuje predajnosť pneumatík podľa typu vozidla, pre ktoré sú určené (motocykle a automobily), hlavné sezóny predaja sa líšia. Pri automobilových pneumatikách sa jedná o dve sezóny, teda jesenná a jarná. Motocyklové pneumatiky majú len jednu, ale dlhšiu predajnú sezónu počas teplejších mesiacov roka, ktorá trvá približne od jari po jeseň.



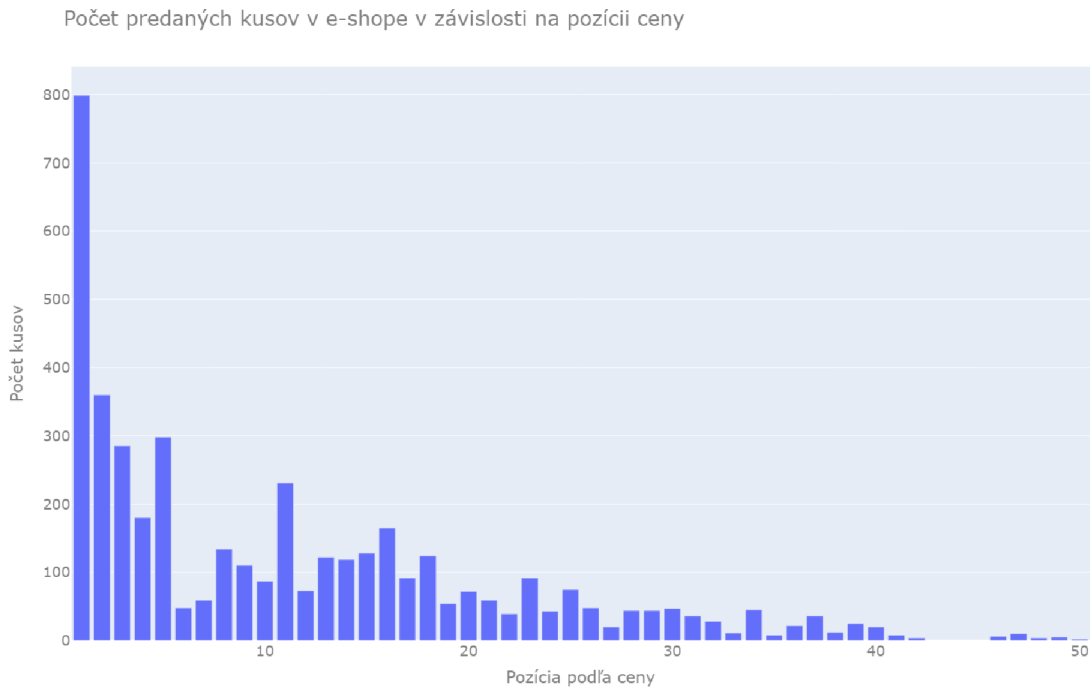
Obr. 6.3: Predajnosť podľa typu vozidla v internetovom obchode

6.2.2 Predajnosť vzhľadom na konkurenčné ceny

Graf 6.4 zobrazuje porovnanie pozície ceny produktov vybranej spoločnosti s ostatnými predajcami figurujúcimi na portáli Heureka.sk. Skupiny pneumatík sa nachádzajú na 1. až 50. pozícii, ktorú určuje cena v porovnaní s konkurenciou. Na Y-ovej osi vidíme celkový počet predaných kusov produktov v časovom období od 01.07.2022 do 31.3.2023.

Ak je vybraná spoločnosť na popredných pozíciách podľa ceny v porovnaní s konkurenciou, vidíme najvyššie predajné objemy. Avšak v grafe môžeme vidieť značné výkyvy, ako v prípade šiestej alebo siedmej pozície, kde firma ponúka výhodnejšiu cenu a napriek tomu, predaje nie sú veľmi vysoké. Tu musíme však zohľadniť veľmi dôležitý faktor, akého druhu produktu sa týka táto pozícia ceny. Ak je nízka cena z dôvodu nízkej predajnosti produktu, čo môže byť napríklad atypický rozmer, menej žiadaný produkt alebo značka, ani najnižšia cena na trhu dostatočne nemotivuje zákazníkov ku kúpe.

Z vyššie spomenutej analýzy vyplynul dôležitý záver, že cena je rozhodne jedným z najdôležitejších faktorov pri rozhodovaní sa o kúpe pneumatík, avšak zákazníka stále zaujímajú aj iné faktory a parametre, ako sú výkon, bezpečnosť, kvalita a jazdné vlastnosti.



Obr. 6.4: Predajnosť produktov vzhľadom na pozíciu ceny v internetovom obchode

Pri pokuse analyzovať vplyv pozície na predaje v kamennej predajni, som opäť narazila na problém s kvalitou dát a chýbajúcimi hodnotami, takže nebolo možné namapovať dostatočné množstvo záznamov, pre vytvorenie relevantných poznatkov z dát. Dostupný dataset o predajoch z kamennej predajne som sa snažila spojiť prostredníctvom súboru `codes.xlsx`, ktorý bol poskytnutý priamo na účel mapovania dát z kamennej predajne na dáta z Heureky. Pre potreby mapovania bolo potrebné očistiť dáta súboru `codes.xlsx` od produktov, ktoré nie sú zahrnuté v našej analýze. Tento krok bol nutný najmä pre zvýšenie efektivity mapovacieho procesu, keďže súbor s identifikačnými číslami obsahoval viac ako 250 000 záznamov. Z tohto počtu bolo konkrétne možné napárovať len 89 predajných záznamov, čo nie je dostatočne relevantná vzorka na celkový objem dostupných predajných záznamov z kamennej predajne, ktorých bolo viac ako 6 000.

6.3 Vyhodnotenie zhlukovania

Zhlukovacia metóda nám ukázala vzorce predajnosti v časovom horizonte a pomôže firme prispôsobiť stratégiu predaja v závislosti na správaní zákazníkov na trhu v rôznych časových obdobiach. Dôležitou súčasťou je aj interpretácia získaných znalostí a ich zasadenie do reálneho kontextu.

Pri tvorbe dendrogramov som skúšala aplikovať štyri metriky, ktoré sú opísané v sekcii 3.2.2. Na dendrogramoch 6.5, 6.6, 6.7 a 6.8 z dátovej matice 5.1 s predajmi skupín pneumatík podľa cenových kategórií v jednotlivých mesiacoch, je vidieť použitie odlišných vzdialenostných metód. Ich vierohodnosť pri daných dátach som overila pomocou kofenetického korelačného koeficientu, ktorý je opísaný v podkapitole 3.2.3. Na základe týchto

korelácii, ktoré sú ukázané v tabuľke 6.1, som vybrala priemernú vzdialenosť (Average link) ako najlepšiu metriku pre vzdialenosť medzi zhlukmi.

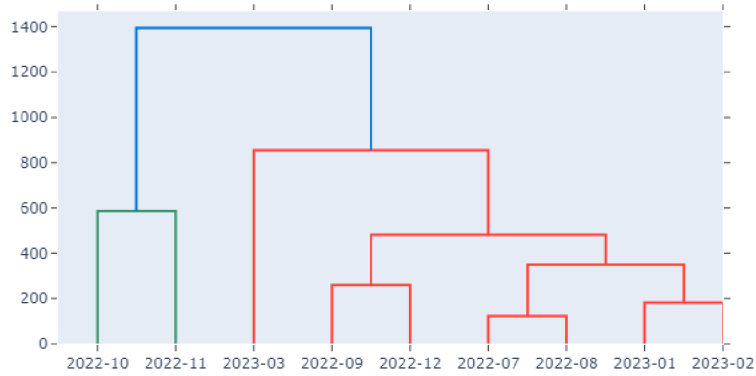
Minimálna vzdialenosť	0.8592
Maximálna vzdialenosť	0.8866
Priemerná vzdialenosť	0.8916
Stredná vzdialenosť	0.8907

Tabuľka 6.1: Porovnanie kofenetického korelačného koeficientu pre vybrané metódy zhlukovania

Výsledkom hierarchického zhlukovania vybraných dát na ukážkach, sú zhluky mesiacov roku s podobnými charakteristikami predajov.

Ako vidíme na vizualizácii zhlukovania, najväčšie podobnosti boli zistené v mesiacoch júl a august, keďže sú spojené do zhluku ako prvé. Mesiace január a február majú najnižšie predaje a preto sú taktiež skoro spojené do jedného zhluku. Najviac sa líšia mesiace október a november, čo môže byť spôsobené začiatkom prezúvacej sezóny a výrazne zvýšeným objemom predajov. Marec sa pripája k zvyšným mesiacom až ako posledný, predtým ako sa pripoja október s decembrom. Môže to byť tým, že sa predaje v prvej cenovej kategórii v tomto mesiaci výrazne líšia od ostatných kategórií.

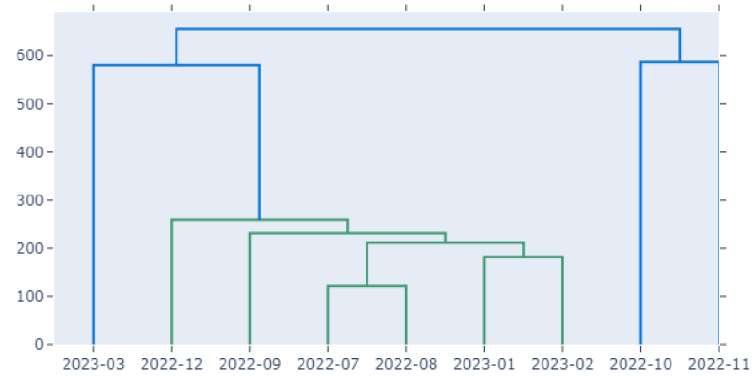
Ďalšie možnosti zhlukovania, ako napríklad zhluky značiek alebo cenových kategórií s podobnými predajnosťami, sú užívateľom dostupné pri využití nástroja na analýzu.



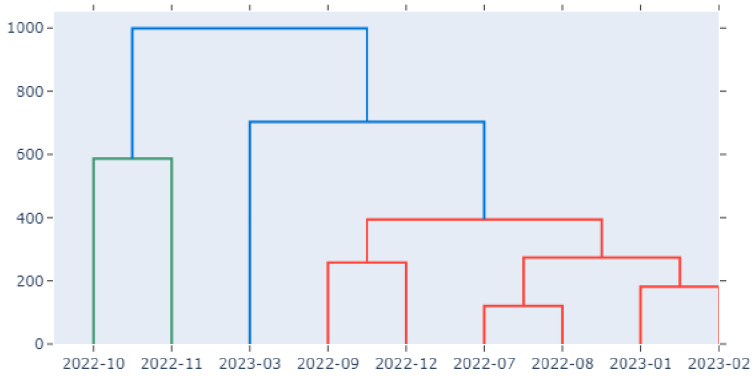
Obr. 6.5: Dendrogram zhlukovania predajov v kamennej predajni s metrikou maximálnej vzdialenosti

6.3.1 Prínos vytvorenia užívateľského prostredia

Priaznivé užívateľské prostredie bolo vytvorené podľa požiadaviek, ktoré vyplynuli z konzultácie s klientom. Vyžadované parametre boli prispôsobiteľnosť rozličným potrebám a náhľadom jednotlivých užívateľov.



Obr. 6.6: Dendrogram zhlukovania predajov v kamennej predajni s metrikou minimálnej vzdialenosti

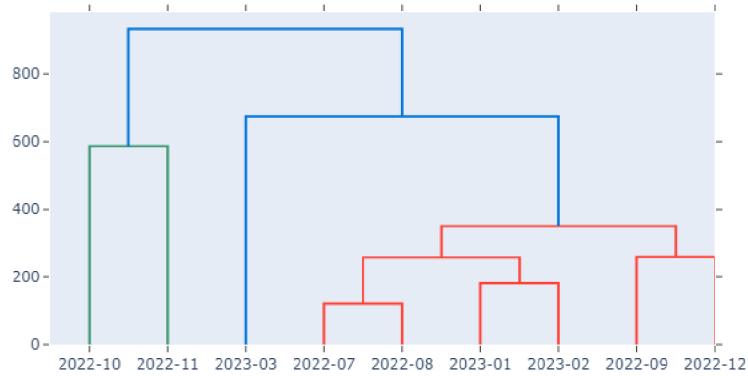


Obr. 6.7: Dendrogram zhlukovania predajov v kamennej predajni s metrikou priemernej vzdialenosti

6.4 Odporúčania firme v závislosti na ďalšie možnosti

Jedno zo základných odporúčaní je zabezpečenie kvality dátových vstupov. Zodpovednosť pre každý typ uchovávaných dát by mala byť jasne definovaná a komunikovaná všetkým zamestnancom, ktorí prichádzajú do styku s dátami. Zamestnanci by mali byť zodpovední za udržanie dát v čo najlepšej kvalite a definovanej forme.

Rovnako dôležitá je presná definícia a opis pre každý typ údajov, ktoré sa uchovávajú v systéme, aby sa zabezpečila jednotnosť zadávaných vstupných dát. Taktiež je potrebné rozlišovať a ukladať správne typy a formáty dát pre umožnenie lepšej manipulácie, jednoduchého vyhľadávania a spracovania dát. Môže to pomôcť zamestnancom zabezpečiť, že



Obr. 6.8: Dendrogram zhlukovania predajov v kamennej predajni s metrikou strednej vzdialenosti

údaje, ktoré sa používajú, sú jednoznačné a jasné. Zlepšenie kvality na základe týchto odporúčaní umožní zrýchliť proces predspracovania a čistenia dát. Vďaka týmto opatreniam bude možné jednoduchšie a pravidelnejšie analyzovať zákaznícke správanie a priebežne reagovať na dynamický vývoj trhu a prispôsobiť mu rozhodovanie o strategických krokoch firmy.

Podobne by bolo veľmi prospešné zaviesť konzistentný spôsob uchovávania dát z kamennej predajne a internetového obchodu tým, že sa vytvoria jednotné atribúty a stanoví sa ich presný formát pre účely párovania dát z týchto obchodných kanálov.

Mali by byť ukladané všetky dáta, ktoré sú vhodné pre analýzu a využitie v rámci obchodných procesov. Ako je spomínané v sekcii 4.3, pre rozšírenie možností získavania znalostí z dát, by bolo pre obchodnú spoločnosť výhodné rozšíriť databázu o informácie o jednotlivých zákazníkoch, ich nákupnú históriu a detaily objednávok. Niektoré užitočné informácie o zákazníkovi by boli napríklad typ vozidla, demografické údaje, ako napríklad vek, pohlavie alebo miesto bydliska.

Pre získavanie takéhoto typu údajov je vhodný spôsob napríklad vernostný program, ktorý prináša zákazníkovi rôzne benefity za výmenu poskytnutia týchto informácií.

V prípade, že firma implementuje spomenuté návrhy, bude možné využiť aj ďalšie metódy získavanie znalostí z dát na detailnejšie porozumenie zákazníckeho správania, ako napríklad klasifikáciu, regresiu, asociačné pravidlá alebo sekvenčné vzory.

Kapitola 7

Záver

Táto bakalárska práca poskytuje prehľad o spôsoboch získavania znalostí z dát a ich prenesenie do praxe na dátach reálnej obchodnej spoločnosti. Pri snahe o implementáciu jednotlivých metód som narazila na limitácie dostupných dát v podobe nepresností, chýbajúcich kľúčových informácií a možností párovania dát. Tieto faktory boli tak isto dôvodom k tomu, že predspracovanie a čistenie dát vyžadovalo výrazné množstvo času a práce. Výsledné dátové sady boli následne podrobené zvolenej metóde hierarchického zhlukovania, ktorá vyplynula z charakteristiky dostupných dát ako najvhodnejšia možnosť. Získané znalosti boli spracované v jazyku Python do prehľadných vizualizácií, a taktiež som vytvorila užívateľské prostredie, ktoré umožňuje opakovať analýzu a vizualizáciu získaných výsledkov a jej prispôsobenie podľa špecifických potrieb jednotlivých užívateľov.

Na základe takto dostupných informácií je spoločnosť schopná rozhodovať o strategických krokoch na základe poznatkov o správaní zákazníkov na trhu a ich senzitivite na jednotlivé faktory. Na záver by som zdôraznila dôležitosť prinášania teoretických postupov do praxe, ktoré mne samej a rovnako spoločnosti s ktorou som spolupracovala priniesli množstvo dôležitých poznatkov o získavaní znalostí z dát. Na jednej strane išlo o kvalitatívne požiadavky týchto metód na dátové vstupy a na druhej strane ako môže firma zlepšiť narábanie a uchovávanie dát pre ich lepšiu využiteľnosť pri tvorbe budúcej obchodnej stratégie. Do budúcna by bola zaujímavá možnosť rozšíriť databázu uchovávaných informácií o dodatočné atribúty, ktoré by nám dovolili využiť aj iné metódy na porozumenie dát.

Literatúra

[1] Aggarwal, C. C. *Data Mining: The Textbook*. 1st. Springer, 2015. ISBN 978-3-319-14141-1.

[2] Association, I. R. M. *Big Data: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2016. Premier reference scource. ISBN 9781466698406.

[3] Bramer, M. *Principles of Data Mining*. 3. vyd. Springer London, 2013. ISBN 978-1-4471-7306-9.

[4] Cios, K. J., Pedrycz, W. a Swiniarski, R. W. *Data mining methods for knowledge discovery*. Springer Science & Business Media, 2012.

[5] Dunham, M. *Data Mining: Introductory And Advanced Topics*. Pearson Education, 2006. ISBN 9788177587852.

[6] Fayyad U., . S. P. From Data Mining to Knowledge Discovery in Databases. 1996, zv. 17, s. 37–50. DOI: 10.1609/aimag.v17i3.1230. Dostupné z: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230.

[7] Gregorová, V. *Zhluková analýza v systémoch STATISTICA a MATLAB*. Brno, CZ, 2008. Diplomová práca. Masarykova univerzita, Přírodovědecká fakulta. Dostupné z: https://is.muni.cz/th/qpzjr/diplomova_praca.pdf.

[8] Han, J., Kamber, M. a Pei, J. *Data Mining: Concepts and Techniques*. 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790.

[9] Hastie, T., Tibshirani, R. a Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Február 2009. ISBN 0387848576.

[10] Mishra, S., Mohsin, N. a Sahu, N. Binary Data Matrix Model for Document Clustering. *International Journal of Current Trends in Engineering Research e-ISSN 2455–1392*. Máj 2016, zv. 2, s. 2455–1392.

[11] Montgomery, D., Peck, E. a Vining, G. *Introduction to Linear Regression Analysis*. Wiley, 2012. Wiley Series in Probability and Statistics. ISBN 9780470542811.

[12] Müllner, D. Modern hierarchical, agglomerative clustering algorithms. September 2011.

[13] Nielsen, F. Årup. *Data Mining with Python*. November 2017. Dostupné z: http://www.freetechbooks.com/data-mining-with-python-working-draft-t1159.html.

[14] Rahm, E. a Do, H. Data Cleaning: Problems and Current Approaches. *IEEE Data Eng. Bull.* Január 2000, zv. 23, s. 2–12.

[15] Saraçli, S., Dogan, N. a Dogan, I. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications.* December 2013. DOI: 10.1186/1029-242X-2013-203.

[16] Witten, I. H., Frank, E. a Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques.* 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123748569.

[17] Zendulka, J., Bartík, V., Lukáš, R. a Rudolfová, I. *Získávání znalostí z databází, Studijní opora.* 2006.

Príloha A

Manuál

Táto príloha obsahuje manuál k spusteniu riešenia praktickej časti tejto bakalárskej práce.

Spustenie

Na stiahnutie potrebných knižníc použite príkaz: `pip install -r requirements.txt`.
Skripty sa spúšťajú pomocou príkazu: `python <subor.py>`.

Najskôr je potrebné spustiť skripty v súbore s dátovými sadami na predspracovanie dát:

- `python eshopData.py`
- `python mrpData.py`

Následne je možné spustiť nástroj na analýzu pomocou:

- `python app.py`

Analýza predajnosti podľa pozície ceny v konkurenčnom prostredí sa spúšťa pomocou:

- `python position.py`

Príloha B

Obsah priloženého pamäťového média

- Zdrojové kódy v adresári **src**
- Dátové sady použité v praktickej časti v adresári **data**
- Zdrojové súbory pre technickú správu v adresári **doc**
- Technická správa bakalárskej práce **xkubin24.pdf**