TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Inga Staršinova 183005IABM

KAUBAMAJA KLIENTIDE SEGMENTEERIMINE

Magistritöö

Juhendaja: Ants Torim

PhD

TALLINN 2021

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Inga Staršinova

10.05.2021

Annotatsioon

Käesoleva töö eesmärgiks on andmekaevanduse tehnikaid kasutades luua ettevõttele Kaubamaja AS kliendisegmendid, mis on kasutatavad turunduslikel eesmärkidel ja võimaldavad keskenduda müügile ning kliendisuhete juhtimisele.

Töö esimeses pooles kirjeldatakse varasemate uurimuste põhjal erinevaid andmekaevandusel põhinevaid klientide segmenteerimise viise jaekaubanduses. Järgnevalt valitakse välja meetodid, mille põhjal eesmärgini jõuda. Seejärel luuakse esinduslik testandmete kogum ning leitakse muutujate väärtused RFM- analüüsi jaoks. RFM-analüüsi ja klasteranalüüsi kombinatsioone kasutades luuakse 4 ja 5 klastrilised segmendid ning analüüsitakse mõlema eeliseid ja puuduseid. Dimensionaalsuse vähendamiseks kasutatakse klasterdamisel peakomponentide meetodit.

Lõputöö tulemuseks on Kaubamajas kasutatavad kliendisegmendid.

Lõputöö on kirjutatud eesti keeles ning sisaldab 55 lehekülge teksti. Töös on 6 peatükki, 19 joonist ja 8 tabelit.

Abstract

Customer Segmentation in Kaubamaja

The aim of this research is to create customer segments for Kaubamaja AS. The main purpose of this analysis is to help the business and marketers understand better its customers and therefore conduct customer-centric marketing more effectively. With a better understanding of customers behavior, retailers can integrate segmentation with market research to comprehend why certain customer segments behave the way they do. Are there any key insights that can be acquired through this discovery process that signal a need for changes to business strategy? The most important research questions is: Who are our main customers and to which customer segments they belong to? What is the optimal grouping of the customers?

In this master's thesis is used K- means clustering as an exploratory method to seek patterns in data and to detect customers segments as clusters. One of the major applications of K-means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company. To get a clearer picture from our customer segments author visualize the K-means clustering results using the PCA. PCA is the process of extracting the essence from a myriad of data, so the new, smaller dataset can represent the unique features of the original data without losing too much useful information.

In the last part of the work, the performance of the model was analyzed on the basis of the A / B test. Samples based on different methods have different conversion rates.

As opportunities for further development, it was considered that in addition to customer descriptive data, product and in particular brand purchase data should be included in the model. An analysis of which brands customers buy together or which brand offers should not be sent to customers could provide interesting results.

The main goal of the work was to create customer segments for Kaubamaja and to find out the features that characterize Kaubamaja's main customers. The resulting five-cluster model provides descriptions of the segments and the increase in conversion rate in the test sample suggests that the introduction of these segments in marketing and sales will help maintain and increase customer loyalty and conduct customer-centric marketing more effectively.

Lühendite ja mõistete sõnastik

CLC	Customer's Lifecycle – kliendi elutsükkel.
CR	Conversion Rate – konversioonimäär näitab, mitu protsenti külastajatest või valimi klientidest, on teinud soovitud tegevust (sooritanud ostu).
CRISP-DM	Cross Industry Standard Process for Data Mining – tegevusalast sõltumatu andmekaeve standardiseerimise protsess.
KPI	Key Performance Indicators - võtmenäitaja ehk tulemuslikkuse võtmemõõdik, mis aitab mõista, kuidas ettevõttel läheb.
PCA	Principal Component Analysis – meetod, millega leiatakse uued tunnused, mis on algsete tunnuste kombinatsioonid.
RFM	Recency, Frequency, Monetary - reeglistik, mis aitab jagada kliendid erinevatesse segmentidesse selle järgi, kui lojaalsed ja kasumlikud nad on. Hiljutisus näitab, millal klient viimati ostis. Sagedus näitab mitu korda klient perioodi jooksul oste on teinud. Rahaline väärtus näitab kliendi ostudele kulutatud summat.
SKU	Stock Keeping Unit - varudele määratud tähtnumbriline kood, mis lähtub kauba erinevatest omadustest.

Sisukord

Jooniste loetelu

Tabelite loetelu

Sissejuhatus

Käesoleva magistritöö teemaks on Kaubamaja klientide segmenteerimine andmekaevandamise tehnikate abil, mis on orienteeritud turundusele, müügile ja kliendisuhete juhtimisele.

Kliendilojaalsus on tänases päevas ettevõttele oluline konkurentsieelise looja. Ajal, kus suurem osa kaubandust on liikumas e-kaubanduse suunas ning innovaatilised väikeettevõtted suudavad pakkuda klientidele tooteid madalamate hindadega on kliendilojaalsus üks vähestest vahenditest, millega traditsioonilist kaubandust elus hoida. Tarbijate ostuotsuseid mõjutab järjest enam hea ostukogemus.

Lojaalsus ei ole enam juhus ega teiste valikute puudumine. Seetõttu tuleb õppida tundma oma klienti, aru saada teema eelistustest ning aktsepteeritavast hinnaskaalast. Kliendile soodustuse andmine ning lojaalsusprogrammi liitmine ei ole piisav, et klienti enda juures hoida. Hästi läheb ettevõttel, kes kliendisuhetest õpib. Aja jooksul saad oma klientidest üha rohkem teada ja neid teadmisi saab kasutada klientide paremaks teenindamiseks. Tulemuseks on õnnelik, rahulolev klient ja kasumlik äri [1]. Suurtel ettevõtetel, kellel on tuhandeid kliente, ei ole võimalust iga kliendiga isiklik suhe luua. Seetõttu peavad suured ettevõtted looma suheteks oma klientidega muid võimalusi. Eelkõige saab aga ära kasutada iga kliendisuhtluse käigus kogutud andmeid. Andmekaevandamisest on saanud äriprotsess suurte andmehulkade uurimiseks, et leida informatiivseid ja tähenduslikke mustreid ja reegleid.

Tulevikku vaatavad ettevõtted liiguvad selles suunas, et mõista oma klienti ja kasutada seda arusaamist, et kliendil oleks lihtsam osta neilt ja mitte pöörduda nende konkurendi poole. Need ettevõtted hindavad iga kliendi väärtust, et teada saada, kes on väärt, et neisse investeerida ja nende nimel pingutada, ning kellel lubada kliendisuhtest lahkuda. See tähendab, et fookus muutub laiematelt kliendisegmentidest üksikutele klientidele ning sellega peab muutuma kogu turundus, müük ja klienditugi.

Klientide hindamine omab suurt väärtust olukorras, kus kõikide klientideni jõudmiseks ei ole kas aega või eelarvet. Kui mõni klient tuleb kõrvale jätta, siis on otstarbekas jätta välja need, kes kõige vähemtõenäoliselt ettevõtte tegevustele reageerivad. Mõne ettevõtte turundusplaan võib näha ette pakkumisi kõikidele klientidele. Sel juhul ei oma klientide mudeldamine väärtust. Siiski võib ka sel juhul olla andmekaevandamisest kasu õigete sõnumite valimisel või

klientide käitumise prognoosimiseks. Tõenäolisem on aga stsenaarium, et turunduseelarve ei võimalda kõikide klientidega tegelemiseks samaväärseid tegevusi. Postitades uudiskirja 30% juhuslikult valitud kliendile, jõuate 30% sihtklientideni. Kui postitada uudiskiri mudeli abil leitud 30% parimale kliendile, jõuate 65% sihtklientideni [1]. Nende kahe valimi erinevus moodustab kasumi.

Individualiseeritud, asjakohane info, mis jõuab olemasolevate ja potentsiaalsete klientideni õigel hetkel, on muutunud üha tähtsamaks intensiivselt müüvate ja turundavate ettevõtete jaoks. Turunduse automatiseerimine ja klientide segmenteerimine on muutunud kasulikuks töövahendiks. Paljud osalejad, eelkõige e-kaubanduse valdkonnas, on võtnud kasutusele turunduse automatiseerimise põhimõtted ja töötavad süstemaatiliselt kliendi digitaalse jälje analüüsimisega. Seetõttu võib pidada suurandmetel põhinevat ärimudelit konkurentsieelise loojaks [7]. Ühendades kliendi ostukäitumise andmed lisateabega kliendi kui füüsilise isiku kohta, peaksime suutma oma kliente veelgi määratleda.

Et suunata oma tegevused õigetele klientidele, on loodud andmekaevandamise tehnikad, mis on mõeldud toime tulema suure andmemahuga. Andmete kaevandamine selgitab välja ja kinnitab seoseid erinevate kriteeriumite ja muutujate vahel [5]. Kuivõrd on need tehnikad efektiivsed, et neil põhinev segmenteerimine kasutusele võtta? Oluline on statistliste tööriistade rakendamine koostöös valdkonnateadmistega.

1. Magistritöö eesmärgid

Jaemüügi moevaldkond on otseselt seotud klientide elustiiliga. Mood on dünaamiline valdkond ning peab olema innovaatiline – uus kollektsioon peab klientideni jõudma vähemalt kaks korda aastas. Suurel määral on valdkond mõjutatud ajakirjadest, filmidest ja staar-isikutest. Edu tagab mitmekesisus, kiirus ja konkurentsivõimelisus ning vastavus klientide soovidele.

Kes on ettevõtte jaoks kõige väärtuslikumad või vähemväärtuslikumad kliendid? Mis on nende eristatavad omadused? Kes on kõige lojaalsemad või vähemlojaalsemad kliendid ja kuidas neid iseloomustada? Millised on klientide ostukäitumise mustrid? Millist tüüpi kliendid reageerivad tõenäolisemalt teatud reklaamimisele? Olgu eesmärgiks leida sarnaseid rühmi või tuvastada põhjus, mis hoiab kliendid sinu juures, saab andmekaevandustehnikaid kasutades need ootused täidetud. Andmekaevanduse eesmärgiks ei ole siinjuures leida mitte ükskõik milliseid mustreid, vaid mustreid, mis aitavad edendada äri.

Käesoleva töö eesmärgiks on leida vastused neile küsimustele ning luua klientide segmendid kasutades selleks andmekaevanduse tehnikaid ja reeglite kombinatsioone. Segmenteerimisel liigitatakse kliendid vastavalt nende vajadustele, tunnustele (sugu, vanus) ja ostukäitumisele. Klientide segmenteerimine aitab oma sihtgrupi vajadusi paremini mõista ja potentsiaali kasutada. See omakorda aitab püsida konkurentsis ning hoida ja kasvatada äri.

Kaubandus on üks tihedama konkurentsiga valdkond, seda nii füüsiliste poodide kui internetikaubanduse sektoris. Selleks, et pakkuda klientidele rohkem väärtust, teha paremaid personaliseeritud pakkumisi ja suurendada ettevõtte kasumit, kasutab töö autor võimalusi klientide segmenteerimiseks andmekaevanduse meetodite abil. Aluseks on tänapäevase ja moeteadliku inimese vajadused, mis on pidevas muutumises. Selle analüüsi peamine eesmärk on aidata ettevõttel oma kliente paremini mõista ja seeläbi tõhusamalt läbi viia kliendikeskset turundust. Töö käigus:

- Kaardistab autor erinevaid andmekaevandusel põhinevaid klientide segmenteerimise viise jaekaubanduses: Selgitab välja, millised on segmenteerimise tehnikad. Autor uurib teadusartiklites kirjeldatud alternatiivseid meetodeid klientide segmenteerimiseks. Viimasel kümnel aastal on loodud mitmeid mudeleid ja algoritme, et luua tulemuslikud kliendisegmendid.

- Viib läbi kuni 10 000 Kaubamaja kliendi andmetel põhineva eelanalüüsi klientide segmenteerimiseks ning loob Kaubamajale esmased kliendisegmendid.
- Testib esmaseid kliendisegmente uudiskirja valimi peal, koostades kolm erinevat valimit, kellest 2 saavad uudikirja ja 1 valim jääb kontrollgrupiks. Kõigi kolme grupi puhul hinnatakse nende reageerimist konversioonimäära kaudu.
- Selgitab välja tunnused, mis iseloomustavad Kaubamaja jaoks kasulikku klienti ning leiab seoseid, mis ilma antud töö läbiviimiseta ei pruugi välja paista.
- Hindab koos ettevõttega segmenteerimise tulemusi ning teeb eksperimendist kokkvõtted ja järeldused sh pakub välja tunnused, mis oleks mudeli täpsuse parendamiseks vajalik kirjeldada ja andmekogusse lisada.

Käesoleva töö ülesandeks ning lõppeesmärgiks on luua kliendisegmendid ettevõttele Kaubamaja. Varasemalt on Kaubamaja kasutanud demograafilistel näitajatel ja ostusummadel põhinevat segmenteerimist lähtuvalt konkreetsest kampaaniast. Uuringud on aga näidanud, et ainult ostusummadel põhinev klientide segmenteerimine ei ole piisav [3].

Töö lõpptulem võimaldab ettevõttel efektiivsemalt toimetada. Eesmärk on, et pakkumistes reklaamitakse õigeid tooteid õigetele klientidele läbi mille kasvab konversioonimäär (CR) ning väheneb kliendipakkumise kampaania ettevalmistusele kuluv aeg. Konversioonimäär näitab, mitu külastajat suudab jaekaupmees ostjaks muuta. Seda näitab % kui palju 100 külastajast sooritab ostu [14]. Konversioonimäär on erinevates valdkondades erinev ning pole kindlat mõõdupuud, millisest numbrist alates on näitaja hea või halb. Füüsilisel kauplusel, mis müüb tavalisi valmisriideid, jääb efektiivsuse määr 18–25% kanti. E-kaubanduses jääb efektiivsus 2–3% piiresse.

Kuigi andmekaevetehnikate kasutamisel turundusvaldkonnas on palju võimalusi, ei ole nende kasutamine Eestis jaekaubanduse valdkonnas täna tavapärane. Autorile teadaolevalt on ainus jaekaubandusettevõte, kes klientidele turunduslike pakkumiste saatmisel ostukorvianalüüsi kasutab Kaubamajaga samasse gruppi kuuluv tütarfirma Selver ning Kaubamajal on kasutusel e-poe soovitusmootor toodete osas.

Magistritöö eesmärk laiemalt on näidata andmekaevetehnikate kasutamise võimalusi ja tõsta teadlikkust nende kasutamise osas turundusvaldkonnas. Töö temaatika ja tulemused pakuvad huvi kontserni ettevõtetele, kes on huvitatud pikaajalise konkurentsieelise saavutamisest. Seda loodetakse saavutada kliendikeskse lähenemise ja kliendiandmete analüüsiga, mitte hinnakujundamisega, et lühiajaliselt käivet suurendada.

2. Varasemad uurimused ja teoreetilised lähtekohad

Selles peatükis antakse ülevaade jaekaubanduse kliente puudutavatest varasemastest uurimustest ja teoreetilistest kontseptsioonidest, mis haakuvad klientide segmenteerimisega ning põhinevad suurandmetel. Varasemate uurimuste põhjal tuuakse välja turunduses kasutavatavaid segmenteerimise võimalusi ning andmekaevandamise tehnikate kasutamise võimalusi turundusvaldkonnas. Käesolevas peatükis saab täpsema ülevaate, kuidas kasutada klientide segmenteerimiseks RFM analüüsi ja klasteranalüüsi.

2.1 Andmekaevandamise tehnikate võimalused turundusvaldkonnas

Traditsiooniliselt on turunduse eesmärkideks: teha kindlaks uusi väärtusvõimalusi ja luua uusi väärtuspakkumisi ning rakendada võimalusi uute väärtuspakkumiste turustamiseks. Konkurentsi tihenedes on turusegmendid ajaga väiksemaks ning fragmenteeritumaks muutunud. Tehnoloogiad aga võimaldavad täna ettevõttel oma klientuuri sobivateks segmentideks jagada [9]. Turundusvaldkonnas on mitmeid erinevaid analüütilisi meetodeid klientide segmenteerimiseks. Kõige traditsioonilisem segmenteerimise viis on demograafiline segmenteerimine. Uuemad analüüsid võtavad segmenteerimisel arvesse ka klientide käitumist, motivatsiooni ning käitumise ja kasutamise mustreid [4]. Tegelema peaks ettevõte aga põhiklientidega, kellelt saadav tulu väheneb. Kui klient toob plaanitust alla 20 protsendi vähem raha sisse, tuleb hakata temaga tegelema [10]. Kui ettevõttel on kogum andmeid: andmed klientide kohta ja miljoneid ridu ostuandmeid, on tegemist keerulise andmekoguga. Automaatsed tehnikad aitavad nende andmete põhjal luua sarnaseid hulkasid.

Mõned turundusteadlased leiavad, et klientide segmenteerimise puhul on oluline kliendikülastuse eesmärk. Räägitakse erinevatest eesmärkidest nagu kiire ost, hommikusöögi-külastus või iganädalase kauba täiendamine. Eesmärk on mõista külastuse kavatsusi ja selle põhjal genereerida kliendikülastuse segmendid [8]. Missioonipõhist segmenteerimist soovitatakse poe ümberkujundamise puhuks.

Teised teadlased toovad välja ennustava analüüsi olulisuse kui suureneva rolliga analüüsi jaekaubanduses. Ennustavat analüüsi toetavad eelkõige korrelatsioonitehnikad [17]. Näiteks teeb Amazon koostööd kohalike jaemüüjatega, mis võimaldab paljude SKUde tarnimist samal päeval just ennustava analüüsi tõttu [13]. Kliendispetsiifilised ja asukohapõhised andmed võimaldavad hulgaliselt ennustavaid analüüse ja keerukamaid ennustusmudeleid. Kliendi varasem ostuajalugu koos reklaamivastuse ajaloo ja klikkimisvooga võib aidata kujundada isikupäraseid reklaame.

Tehnoloogia arenedes leitakse võimalusi siduda kliente asukohaga ning kasutada neid andmeid ära ostude suurendamiseks. Kui sinu klient asub hetkel konkureeriva jaemüüja juures ning on määratletud geokoodiga, tuleb sel hetkel saata neile omapoolne pakkumine, et neid konkurendi juurest eemale meelitada [1].

Äri edendamiseks soovitavad mõned teadlased aga assotsiatsiooni reeglitel põhinevat ostukorvi analüüsi, mis leiab sagedasi mustreid ning annab infot selle kohta, milliseid esemeid kliendid sageli koos ostavad. Nii on loodud klientidele üksikute toodete soovitused, mida kliendid tõenäoliselt ostaksid [19]. Neid reegleid kasutatakse omakorda profiilide loomiseks, mis põhinevad kui-siis reeglitel.

Paljud ettevõtted leiavad, et segmenteerimine on vajalik erinevate turunduseesmärkide korral: ristmüügipakkumiste puhul, klientide mitte kaotamiseks, fokuseeritud sõnumite edastamiseks. Eeldatakse, et keskendudes sarnastele kliendirühmadele, on need toimingud tõhusamad kui oleksid juhul kui kõikidele klientidele ühtviisi lähenedes [1]. Ükskõik millist tehnikat sarnaste kliendirühmade leidmiseks kasutatakse, tuleb klastreid hinnata ja paljude rakenduste puhul tõlgendada. Klastrite hindamiseks ja tõlgendamiseks on kirjeldatud mitmeid lähenemisviise. Mõnikord toob klastrite tõlgendamine uusi teadmisi, millel on äriline väärtus.

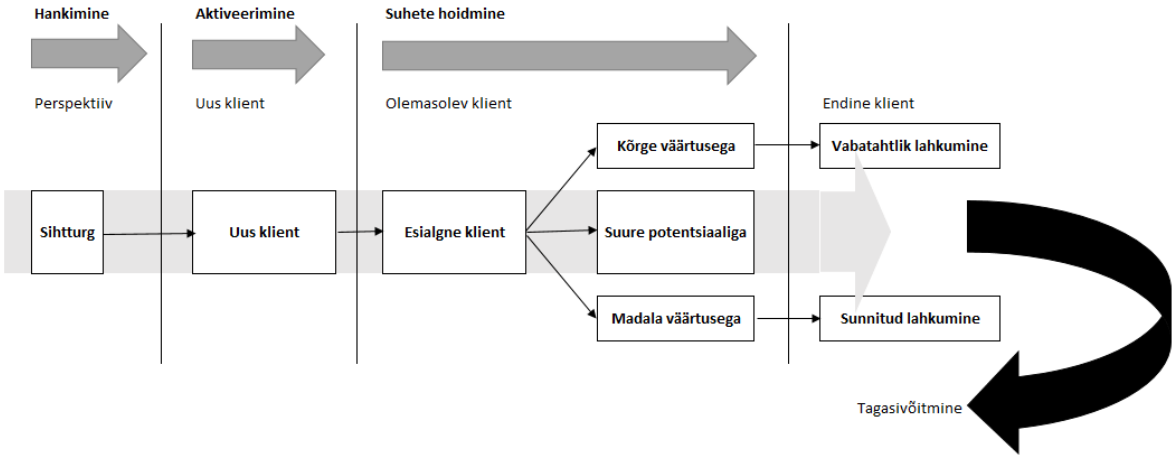
Klastrite tuvastamine annab võimaluse õppida tundma keerukate andmete struktuuri, et jagada andmed, mis üksteisega konkureerida võivad, lihtsamateks komponentideks. Suunates fookuse konkreetselt määratletud omadustega klastrile, saab ka küsimustele kergemini vastata. Järgnevalt kirjeldatakse võimalusi, kuidas nende sarnaste klientideni jõuda.

Personaliseerimine on võime pakkuda klientidele kohandatud sisu ja teenuseid, mis põhinevad teadmistel nende klientide käitumise ja eelistuste kohta. Personaliseerimisel kasutatakse tehnoloogiat ja infot klientide kohta. Personaliseerimine seisneb kliendi lojaalsuse

suurendamises luues mõtestatud suhe kliendi ja ettevõtte vahel [11]. Mõistes kliendi vajadusi, saame täita eesmärki. Dogan jt (2018) on kasutanud oma uurimises RFM mudelit koos klasteranalüüsi algoritmidega, et segmenteerida 700 000 spordikaupade jaemüügiklienti. Analüüsi eesmärgiks oli luua kliendikaardi tasemed: pronks, kuld ja Premium. Analüüsi tulemusel loodi kolm erinevat klastrit, mis erinesid oluliselt ettevõtte varasemast segmenteerimise praktikast. Dogani jt poolt läbi viidud analüüsi spordikaupade klient omab mitmeid sarnasusi Kaubamaja kliendiga sh on ettevõtetel samal määral kliente, kes ostavad vaid 1 korra aastas. Dogan jt tõid analüüsi tulemusel välja, et suutsid rühmitada kliente, kellel on sarnased vajadused, soovid ja käitumismustrid [3]

2.1.1 Kliendi elutsükkel

Andmekaevandamise tehnikad ei eksisteeri eraldiseisvana, vaid need eksisteerivad ettevõtluse kontekstis. Kuigi need tehnikad võivad olla iseenesest huvitavad, on need vahendid eesmärgi saavutamiseks. Andmekaevandamise seisukohalt on kliente käsitledes oluline kliendi elutsükkel (CLC). Ärisuhe klientidega areneb aja jooksul ja kuigi iga ettevõtte on erinev jagatakse kliendisuhe viide erinevasse etappi [1] , mis on näha joonisel 1.



Joonis 1. Kliendi elutsükkel.

Perspektiiv – potentsiaalsed kliendid on sihtturul, kuid pole veel kliendid

Reageerijad – potentsiaalsed kliendid, kes on teatud huvi üles näidanud, näiteks täitnud kliendikaardi avalduse või loonud e-poe konto.

Uued kliendid – reageerijad, kes on kas võtnud kohustuse, teinud maksekokkuleppe või esimese ostu.

Olemasolevad kliendid – uued kliendid, kes tulevad tagasi ja kelle kliendisuhe süveneb.

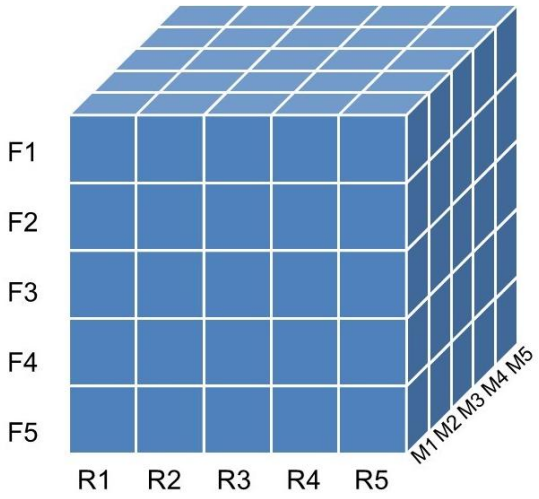
Endised kliendid – on kliendid, kes on kliendisuhtest lahkunud. See võib toimuda vabatahtlikult, kui nad hakkavad konkurendi kliendiks või ei oma meie toode enam nende jaoks väärtust. Lahkumine võib olla sunnitud, kui klient jätab maksmata arved või nad ei ole enam sihtkliendiks, sest on piirkonnast ära kolinud.

Äriprotsessid viivad kliendi ühest elutsüklietapist teise. Need protsessid on olulised, kuna muudavad kliendi aja jooksul väärtuslikumaks. Ettevõtetel, kes saadavad klientidele uudiskirju või otseposti, seavad olemasolevad andmed andmekaevandusele piirid. Andmekaevandust kasutatakse kliendi reageerimise maksimeerimiseks näiteks otsepostituste puhul. Eesmärk on piirata kontakte selliselt, et sisse jääksid need kliendid, kes suurema tõenäosusega reageerivad ja saavad seeläbi headeks klientideks [1]. Mudel võib põhineda näiteks kliendi demograafilistel andmetel, kui need andmed on ettevõttel olemas. Teine viis, kuidas kliente profileerida on mõõta kliendi ja profiili vahelist sarnasust, mida nimetatakse ka kauguseks. Andmed saadakse mingi kliendi hulga pealt, kes esindavad kliendibaasi konkreetsel ajal. Kommunikatsioon kliendiga peab lähtuma sellest, millises etapis klient parajasti on. Selleks, et kliendi elutsüklit hallata, tuleb alustada kliendibaasi segmenteerimisest. Igal ettevõttel ja selle klientidel on oma eripära, seega on erinev ka elutsükli juhtimise strateegia.

2.1.2 RFM analüüs

RFM analüüsimeetod on reeglistik, mis aitab jagada kliendid erinevatesse segmentidesse selle järgi, kui lojaalsed ja kasumlikud nad on. RFM analüüs võtab arvesse müügitehingute ajalugu ning jagab kliendid segmentidesse nende ostukäitumise alusel. Analüüsi tulemusena tekivad klientide rühmad väärtuslikematest ja vähemväärtuslikematest klientidest [2]. RFM analüüs

võtab kliendi puhul arvesse hiljutisust (recency) ehk millal klient viimati ostis. Hiljutisus näitab, kas tegemist on aktiivse kliendiga. Sagedus (frequency) ehk kliendi külastuse sagedus, kui tihti klient meil käib ja rahalist väärtust (monetary) ehk kui ostuvõimeline on klient. Rahalist väärtust mõõdab keskmise ostu suurus. Analüüsi kasutatakse kui turunduse tööriista, mille alusel hinnata klienti. Analüüsi tulemusel arvutatakse klientidele skoor, mille põhjal saab teada, kes on kasumlikumad kliendid ja kui suur on nende osakaal, ning kes on kliendid, keda kaotada ei tohi [3]. Linoff ja Berry on jaotanud kõik kolm RFM-dimensiooni kvintiilideks, et moodustada RFM-kuup 125 lahtriga. Nende mudel on toodud joonisel 2 [1].

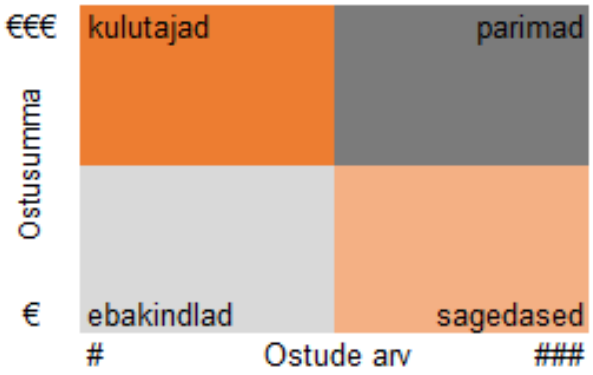


Joonis 2. RFM mudel.

RFM analüüs on tihti kasutusel otseturunduse maailmas. Tegemist on kolmemõõtmelise mudeliga, mida kasutatakse, et jõuda klientideni, kes tõenäolisemalt pakkumisele reageerivad. RFM-i loogika on lihtne: hiljuti ostu teinud kliendid teevad lähitulevikus suurema tõenäosusega ostu. Varem palju oste sooritanud kliendid sooritavad lähitulevikus suurema tõenäosusega uue ostu ning varem palju raha kulutanud kliendid kulutavad tulevikus tõenäoliselt rohkem raha.

RFM ei ole meetod, mida kasutada uute klientide puhul. Ainult olemasolevate klientide kohta on olemas hiljutisuse, sageduse ja rahalise väärtuse info. Seega kasutatakse meetodit olemasolevate klientide hindamiseks.

Turunduses on kasutatud tihtipeale kolmemõõtmelise mudeli asemel kahemõõtmelist modifitseeritud mudelit. Nii on Marcus oma uuringus välistanud hiljutisuse näitaja madalama väärtusega kliendid ning loonud ülejäänud klientide hindamiseks kahemõõtmelise mudeli ainult sageduse ja rahalise väärtuse põhjal [4]. Marcuse loodud mudelit ilmestab joonis 3.



Joonis 3. RFM mudelist tulenevad kliendisegmendid.

RFM mudelit on kasutanud oma veebipoodi hõlmavas uurimistöös Chen, Sain ja Guo (Chen jt 2012), jaotades mudeli alusel ettevõtte kliendid tähendust omavatesse rühmadesse, kus tarbijate peamised omadused said selgelt määratletud. Uuritav veebipood oli suhteliselt uus siseneja veebimüügisektorisse ning omab selgeid paralleele Kaubamaja veebipoega. Chen jt analüüsi eesmärgiks oli aidata ettevõttel kliente paremini mõista ning viia läbi kliendikeskset turundust. RFM-i puhul eeldatakse, et konkreetsed kliendid, mudeli osades reageerivad järgmisele kampaaniale samamoodi nagu nad reageerisid viimasele [6].

2.1.3 RFM mudeli hindamine

RFM-i kasutatakse sageli otseturunduse kontekstis, kus sarnaseid kampaaniaid korraldatakse aastaringselt. Näiteks moele või kodukaupadele spetsialiseerunud kataloogil on selles valdkonnas huvitatud klientide kogum. Kataloogide eesmärk on nendele klientidele tooteid müüa. Kuigi tooted võivad hooajati erineda, on kampaaniad siiski üsna sarnased. RFM meetodi puhul on lisaks üldise reageerimise mõõtmisele ka võimalus mõõta inkrementaalset vastust. Luues näitajatel põhinevale valimile, kellega kontakteerutakse ka kontrollgrupi valim, mis põhineb RFM analüüsil, aga kellega kontakti ei astuta ja pakkumisi/ katalooge ei saadeta, kuid kelle käitumist sellegipoolest samal perioodil jälgitakse.

2.2 Peakomponentanalüüs

Peakomponentanalüüs (PCA) on meetod muutujate arvu vähendamiseks analüüsis. Peakomponentidel on teatud optimaalsed omadused, mis muudavad need võimsaks ja võimaldavad andmete visualiseerimist väheste muutujate põhjal. Dimensionaalsuse vähendamiseks defineeritakse uued kunstlikud atribuudid. Eesmärgiks on atribuutide arvu vähendamine asendades iga tugevas korrelatsioonis olevate atribuutide grupi uue atribuudiga [1]. Peakomponentanalüüs leiab uued dimensioonid ehk komponendid olemasoleva d dimensioonide lineaarse kombinatsioonina. j -ndale peakomponendile vastab kaaluvektor $\mathbf{w}^{(j)} = (w_1, \ldots, w_d)^{(j)}$

Rea i peakomponendi j väärtuseks on kaaluvektori $\mathbf{w}^{(j)}$ skalaarkorrutis reavektoriga $\mathbf{x}^{(i)}$:

$$t_{ij} = \mathbf{x}^{(i)} \cdot \mathbf{w}^{(j)} = x_1^{(i)} \cdot w_1^{(j)} + \ldots + x_d^{(i)} \cdot w_d^{(j)}$$

Esitades algse andmestiku $n \times d$ maatriksina X ja peakomponentide kaaluvektorid $d \times k$ maatriksina W leiatakse nende maatrikskorrutisena uue teisendatud $n \times k$ maatriksi T , kus n on ridade arv ja k on peakomponentide arv:

$T=XW$ [20].

Peakomponentanalüüsi puhul moodustatakse algtunnustest lineaarsed kombinatsioonid nii, et esimese kombinatsiooni ehk peakomponendi hajuvus oleks nii suur kui võimalik. See tähendab, et esimene peakomponent kirjeldab võimalikult suure osa kõigi algtunnuste variatiivsusest. Seda põhimõtet jätkatakse ja moodustatakse sama palju peakomponente kui on algtunnuseid. Peakomponendid üheskoos kirjeldavad ära kogu algtunnuste variatiivsuse. Oluline kirjeldusvõime on aga esimestel peakomponentidel ning viimased peakomponendid on väga väikese hajuvusega, seega väikese kirjeldusvõimega.

Peakomponendid on atribuutide kovariatsioonimaatriksi omavektorid. Nende peakomponentide olulisuse määravad vastavate atribuutide kovariatsioonimaatriksi omaväärtused [1].

Peakomponentide leidmise sammud:

- Standardiseerida d -dimensionaalne andmestik X
- Konstrueerida $d \times d$ atribuutide kovariatsioonimaatriks Σ

- Leiada $\boldsymbol{\Sigma}$ omavektorid ja omaväärtused
- Valida $\boldsymbol{k}$ suurimale omaväärtusele vastavat omavektorit ja konstrueerida nendest projektisoonimaatriks $\boldsymbol{W}$
- Tulemuseks on kovariatsioonimaatriks $\boldsymbol{\Sigma}$

2.3 Klasteranalüüs

Klasteranalüüsi käigus toimub andmete grupeerimine klassidesse ehk klastritesse, nii et iga klastri objektid on omavahel väga sarnased ning erinevad teiste klastrite objektidest. Objektidevahelised erinevused põhinevad atribuutide väärtustel, mida saab avaldada kauguste erinevuste läbi. Andmekaevanduses on uurimiste teemaks selliste klasterdamismeetodite leidmine, mis oleksid efektiivsed suurtel andmebaasidel [12].

Andmemaatriks klasteranalüüsis koosneb objektidest ***(n)*** näiteks isikutest ning nende objektidele iseloomulikest atribuutidest ehk tunnustest ***(p)*** näiteks vanus, sugu, elukoht. Andmemaatriksi struktuuri esitatakse relatsioonitabeliga või ***n x p*** maatriksiga. Kauguste maatriks on objektide ***(n)*** omavaheliste kauguste maatriks, mida esitatakse ***n x n*** tabeliga. Kui andmed on esitatud andmemaatriksina, peab olema see teisendatud kauguste maatriksiks, et klasterdamist rakendada [12].

Automaatne klastrite tuvastamine on suunamata andmekaevandamise tehnika, mida saab kasutada keerukate andmete struktuuri sisekaemuseks. Klasterdamine ei vasta otseselt ühelegi küsimusele, kuid klastrite uurimine võib anda väärtuslikku infot. Klasterdamise üks oluline väljund on klientide segmenteerimine. Klientide klastrid koosnevad inimestest, kes on sarnased ja kellel võivad olla sarnased vajadused ja huvid [1]. Klasteranalüüsi kasutatakse segmenteerimiseks turunduses üsna sageli.

2.3.1 Klasterdamismeetodid

Nagu ka teiste modelleerimistehnikate puhul, leiavad klasterdamise algoritmid mudelikomplektist mustreid, mida saab väljendada reeglite või valemitena. Neid saab

omakorda rakendada teistele andmekogumitele skooride saamiseks. Klastermudelite skoorid on klastrite määramiseks. Skoorid on ühe klastri aluseks või võivad anda hinnangu liikmelisuse tõenäosuse kohta igas klastris.

Klasterdamise algoritme on mitmeid erinevaid. Vajalik algoritm sõltub andmete iseloomust ja püstitatud ülesandest. Klasterdamise algoritmide kõige tuntum klassifikatsioon on nende jaotamine eraldusmeetoditeks ja hierarhilisteks meetoditeks. Eraldusmeetodite tulemuseks on üks jaotus ning hierarhiliste meetodite tulemuseks mitu hierarhilist andmete jaotust klastriteks [12].

Eraldusmeetod jagab andmebaasi objektid (n) k grupiks ($k \leq n$) nii, et iga grupp sisaldab vähemalt ühte objekti ning iga objekt kuulub ainult ühte gruppi. Etteantud klastrite arvu k jaoks tekitab algoritm esialgse andmete jaotuse. Seejärel püüab algoritm jaotust parandada vahetades punktid klastrite vahel. Hea jaotuse näitaja on see, et ühe klastri objektid on üksteisele lähemal ning erinevate klastrite objektid on üksteisest kaugemal. Kõige populaarsema eraldusmeetod on k-keskmise meetod [12]. K-keskmised toetuvad andmete geomeetrilisele tõlgendamisele ruumipunktidena. Kahe andmepunkti vaheline kaugus sõltub nende esitusest, nii et klastri tuvastamisel on andmete ettevalmistamise nõuded sarnased.

- Algoritm: k-keskmise eraldusalgoritm
- Sisend: klastrite arv *k*, andmestik *n* objektiga
- Väljund: klastrid (k tükki), mis minimeerivad ruutvea kriteeriumi.
- Meetod:
    1. Valida juhuslikult k objekti
    2. Korrata
    3. Panna iga objekti kõige lähedasema tsentroidiga (keskpunkt) klastrisse
    4. Arvutada klastrite tsentroidide väärtused ümber
    5. Kuni ruutvea kriteerium koondub

K-keskmise meetodil klasterdamine on kiire ka suure objektide arvu korral. Meetod põhineb klastri keskmistel ja ei vaja kõigi objektipaaride kaugusi. Mitu klastrit peaks olema? Paljudel juhtudel annab sellele küsimusele vastuse äriline eesmärk. Kui klientide segmenteerimise eesmärk on kujundada iga segmendi jaoks erinevad pakkumised, määratakse k vastavalt segmentide arvule, mida ettevõte saab mõistlikult lubada. Optimaalse klastri arvu leidmiseks on aga mitmeid meetodeid. Kõige tuntumaks peetakse elbow meetodit, kus klastrite arv määratakse kindlaks visuaalselt. Sellisel viisil leitav klastrite arv ei pruugi olla täpne. Alternatiivsed lahendused on silueti meetod [15], mis põhineb numbrilisel siluetikordajal ja Davies-Bouldini meetod [16].

Hierarhiline meetod tekitab mitu teineteisest saadud objektide jaotust. Hierarhilised meetodid jagunevad jagavateks ja ühendavateks. Jagav meetod alustab ühest klastrist, kus on kõik andmebaasi objektid. Igal järgneval sammul jaotatakse klaster kaheks väiksemaks klastriks, nii et igal sammul tekib juurde üks klaster. Protsess jätkub kuni klastris on ainult üks objekt või kuni peatumiskriteeriumini. Ühendav meetod alustab jaotusega, kus iga objekt on iseseisvas klastris ning igal sammul ühendatakse kaks lähedasemat klastrit kuni peatumisriteeriumini või üheainsa klastrini. Hierarhiline meetod on hästi kasutatav siis, kui meil on suhteliselt vähe objekte või kui on oodata, et klastrid suhteliselt selgelt üksteisest eristuvad [12].

Tunnuste grupeerimine on hierarhiline klasteranalüüs, mille eesmärgiks on leida omavahel kõige enam seotud tunnused ja moodustada selle põhjal ligilähedast fenomeni peegeldavate tunnuste grupid. Klasteranalüüsi kõrval saab siin kasutada ka faktoranalüüsi, mis on keerukam ja täpsem.

2.3.2 Klasteranalüüsi põhiste segmentide hindamine

Igas klasterdamise projektis on üks esimesi küsimusi see, milliseid muutujaid tuleks klastrite määratlemiseks kasutada. Tihti saab kasutada muutujaid, mida on sama kliendibaasi mudeldamise puhul varem kasutatud, ja mis on osutunud varasemalt produktiivseteks. Need on muutujate komplektid klientide käitumisest ja demograafilistest näitajatest. Eesmärgiks on luua olemasolevate tunnuste asemel fiktiivsed tunnused, mida saab klasterdamise puhul kasutada. Fiktiivsed tunnused esindavad neid tunnuseid ja sisaldavad kompaktsel kujul infot algsete

tunnuste asemel. Mudelisse kaasatakse sel viisil väike osa kõikidest tunnustest. Jaekaubanduse puhul on ootus, et klient kauplust lühikese aja järel taas külastaks. Sel juhul on tegemist suunamata andmete kaevandamisega suunatud eesmärgi saavutamiseks. Klastrid peaksid lähtuma kliendi järgmisest ostust. Kui sihtmuutuja on võimalik kindlaks määrata, peaks loodama segmendid erinevuste maksimeerimiseks. Klasteranalüüsi põhiseid segmente tuleb hinnata lähtuvalt probleemipüstitusest [1]. Klastrite sobivust näitab nende kasulikkus ärieesmärkide täitmisel. See on kõige olulisem sobivuse näitaja, teisalt on seda raske kvantitatiivselt mõõta.

Klasteranalüüsi tulemust iseloomustab alati subjektiivsus. Ülesande püstitamisel on suur valikuvõimalus ja lahendamisel ja lahendi tõlgendamisel on uurijal väga suur omavoli. Klasteranalüüsil ei ole olemas ühte ja õiget tulemust. Klasteranalüüs on aga heaks allikaks oma klientide tundmaõppimiseks uurides konkreetse klastri karakteristikuid. Selle põhjal saab turundusanalüütik luua ülevaate klastrisse kuuluvate klientide näitajatest. Klastreid saab iseloomustada sellega, mis on neisse kuuluvatel liikmetel ühist ja mis eraldab liikmeid kliendipopulatsioonist tervikuna. Neid omadusi arevesse võttes, saab luua uue pakkumise, mis võiks anda parema reageerimistulemuse. Klastrite tulemuslikkust saab hinnata turunduskampaaniale, kliendipakkumisele või uudiskirjale reageerimise alusel.

Enamasti peaks kalsteranalüüsi rakendamiseks, sealhulgas klientide segmenteerimise jaoks, olema klastris ligikaudu sama arv liikmeid. Erandiks on see, kui klastreid kasutatakse pettuste või muude anomaaliate tuvastamiseks. Võib olla väike arv, kuid paljude liikmetega klastreid, mis esindavad kõige tavalisemaid juhtumeid, ja mõned vähemate liikmetega rühmad ebatavalistest juhtumitest, mis väärivad täiendavat uurimist [1].

3. Metoodika

Selles peatükis kirjeldatakse töö metoodika ning töö protsess samm-sammult. Teiseks antakse ülevaade ettevõttest ning kasutatavatest andmetest. Peatükist leiab andmeid kirjeldava statistika ning antakse selgitused analüüsis kasutatavatele atribuutidele. Lisaks leitakse kasutatavate atribuutide omavahelised seosed.

Otseturunduses ja personaliseeritud turunduses ei ole kliendiprofiilide leidmine ning segmenteerimine uus - mida täpsemalt ja võimalikult väikese kuluga potentsiaalsed ostjad ära tabada, seda suurem on kasum. Andmekaevandamine pakub võimaluse töödelda suuremat kliendibaasi rohkema ja kaudsema informatsiooniga, kus tunnusteks on peale traditsioonilise demograafilise bloki ka ostukäitumised ja –harjumused [3]. Käesolevas töös kasutab autor RFM- analüüsi ja klasteranalüüsi kombinatsioone ning dimensionaalsuse vähendamise meetodit PCA.

Segmentide moodustamiseks ja klasterdamiseks kasutatakse käesolevas töös järgmist protsessi:

1. Andmed on mõõdikute kujul – kasutatavatel andmetel on numbriline tähendus

2. Andmete skaleerimine – tehnika mitmemõõtmeliste andmete struktuuri võimalikult täpseks graafiliseks esitamiseks ühe-, kahe- või kolmemõõtmelisel kujul. Mõni erineva skaalaga muutuja olemasolu võib tekitada tulemuse, mis on tingitud selle konkreetse muutuja suurest väärtusest. Sel juhul tuleb kaaluda andmete standardiseerimist nii, et algsete tunnuste keskmine on 0 standardhälve 1.

3. Segmenteerimismuutujate valimine – eeldab valdkonna tundmist. Näiteks klientide segmenteerimiseks võib kasutada klientide hiljutisuse, ostusageduse ja rahalise väärtuse andmeid ning seejärel võib kasutada demograafilisi või käitumispõhiseid andmeid leitud segmentide profileerimiseks.

4. Sarnasuse mõõt – eesmärk on rühmitada kliendid selle põhjal, kui sarnased nad on. Sarnasust vaadatakse objektide vahelise distantsi kaudu. Levinuim klasterdamise algoritm on k-keskmiste algoritm, mida kutsutakse ka Lloydi algoritmiks. Iga objekt, kuulub selle alusel klastrisse, mille keskpunktile see kõige lähemal on.

5. Kauguste visualiseerimine – üksikute atribuutide paarikaupa visualiseerimine.

6. Segmenteerimise meetod ja segmentide arv – segmenteerimiseks on mitmeid klasterdamismeetodeid. Enim kasutatud meetoditeks on kas k-keskmise või hierarhiline klasterdamise meetod. K-keskmise meetodi puhul on vaja määratleda ka segmentide arv. Selleks on levinuim viis elbow meetod ehk küünarnuki meetod. Lõplik valik tuleks aga teha nii statistilisi kui kvalitatiivseid kriteeriume näiteks ärieesmärki arvesse võttes.

7. Tulemuste tõlgendamine – klastri klientide tõlgendamine. Kes on need kliendid, kes klastrisse kuuluvad. Kuna tulemusi kasutatakse otsuste langetamiseks, on vaja neid tulemusi eelkõige mõista. Näiteks võib vaadata erinevate segmentide erinevate atribuutide keskmisi.

8. Tulemuste hindamine – mudeli tulemuslikkuse hindamine A/B testimise läbi. Uuritakse erinevatel tehnikatel loodud valimite konversioonimäärasid.

Analüüsi läbiviimiseks kasutab autor vabavaralist R keelt ja RStudio keskkonda [18].

3.1 Ettevõtte ja andmete kirjeldus

3.1.1 Tallinna Kaubamaja

Kaubamaja on üks esimesi ja suuremaid Tallinna kaubanduskeskusi ning omab väärikat ajalugu alates 1960. aastast. Kaubamaja missiooniks on olla Eesti kaubanduse rätsepaülikond masstootmise ajastul – üks ja ainus Kaubamaja linnasüdames. Visiooniks on Kaubamajal saada Eestist Põhjamaade parimaks. Kuuekümne tegevusaasta jooksul on Kaubamaja olnud alati novaator ning paljusi kaubanduses toimunud uuenduste juuri leidub Kaubamajas. Nii käidi Kaubamajast snitti võtmas selle avamisest alates, sest üleminek iseteenindusele oli tol ajal ennekuulmatu. Kaubamaja oli esimene kaubandusettevõte Eestis, kus teenindajad hakkasid kliente tervitama, pälvides esialgu kundede üllatunud reaktsioone. Hea teenindus on üks Kaubamaja tugevusi ja konkurentsieeliseid. Pidev areng ja muutused toimuvad täna andmete tasemel, et olla kaubanduses teistest sammuke ees.

Kaubamaja on olnud suunanäitajaks ka reklaami valdkonnas. 1960. aastatel puudusid jaemüüjatel kasumi teenimise motiivid ning reklaami võis kasutada tarbijate harimiseks. Tallinna Kaubamaja lõi hulga reklaamiprogramme ning sel viisil nähti võimalust lahti saada

kaubaartiklitest, milles valitses ülejääk. Eesti Reklaamfilmi kultuslikuks muutunud looming oli suurel määral seotud Kaubamajaga. Ettevõtte pikaealisuse edu võti on paindlikkus ning klientide soovidega sammu pidamine.

Börsiettevõte Tallinna Kaubamaja Grupp on kasvanud kaubanduskontserniks, mis annab täna tööd üle 4 200 inimesele. Grupi ettevõtted moodustavad enam kui kümnendiku kogu Eesti jaekaubandusest. Grupi lojaalsusprogramm Partnerkaart on üle 670 000 püsikliendiga suurim Eestis. Partnerkaardi klientidele pakub Kaubamaja 5% allahindlust ning ostusumma pealt kogutavat boonusraha. Ettevõtte 2019. aasta müügitulu oli 717,2 miljonit eurot. Kaubamajade ärisegmendi 2019. aasta müügitulu oli 102,8 miljonit eurot. Kaubamaja Grupi põhiväärtusteks on ausus, hoolivus, usaldusväärsus, uuendusmeelsus ja keskkonnateadlikkus.

Kaubamaja Guppi kuuluv Selver AS kasutab täna klientidele personaalsete pakkumiste tegemisel ostukorvianalüüsi. Ostukorvianalüüs näitab milliseid tooteid või tootegruppe ostetakse sageli koos. Ostukorvianalüüsi kasutatakse läbimüügi suurendamiseks. Supemarketis võib üks ost sisaldada keskmiselt 20 erinevat toodet ning seeläbi saab teha analüüse milliseid kaupu või kaubagruppe koos ostetakse. Moekaupu ostetakse enamasti korraga vaid üks või kaks toodet. Seetõttu kasutab Kaubamaja täna ostukorvi analüüsi oma e-poe soovitusmootori ühe osana. Küll aga kasutab Kaubamaja andmete analüüsil põhinevat kaupade paigutamist külastussegmentide järgi.

3.1.2 Andmete kirjeldus

Töös kasutatakse Kaubamaja andmebaasidesse salvestatud andmeid, mis klientide konfidentsiaalsuse tagamiseks ei sisalda kliente tuvastavaid andmeid ning andmetöötlus toimub vastavalt Euroopa Liidu isikuandmete kaitse üldmäärusega. Kaubamaja Grupp kasutab isikuandmete kogumisel, säilitamisel ja töötlemisel asjakohaseid ja piisavaid tehnilisi ja korralduslikke turvameetmeid, mis tagavad isikuandmete järjepideva korrektse ja turvalise töötlemise.

Mudeli testimiseks on koostatud klientide andmetel põhinev testandmete kogu jälgides, et testandmed oleksid esinduslikud klientide soo, vanuse, suhtluskeele ja elukoha osas.

Kliendi ja ostuandmed saadakse pärituna andmeaidast, kus andmed on juba eelnevalt puhastatud ja struktureeritud. Andmed on andmekaevanduse lahutamatu osa ning andmeait on protsess, mis andmekogumisele ja andmete analüüsimisele palju kasu toob.

Töös kasutatakse tegevusalast sõltumatut andmekaeve standardiseeritud protsessi ehk CRISP-DMi kuna töö eesmärgiks on lahendada äriline probleem ning CRISP-DM mudeli puhul on arvestatud ka ärilist mõõdet.

CRISP-DM protsessi etappide kirjeldus:

1. Äriprotsessi mõistmine – aru saadakse ärieesmärkidest ning äripoole ülesehitusest
2. Andmete mõistmine – info vajalike andmete kohta: millisel kujul on andmed ja kus nad asuvad
3. Andmete ettevalmistus – andmete korrastamine, andmete atribuutide korrelatsioonide uurimine, andmete transformatsioon
4. Modelleerimine – sobivate andmekaeve meetodite ja algoritmide valimine ning nende rakendus andmetel. Vajadusel saab sellest etapist pöörduda tagasi andmete ettevalmistamise juurde kui see peaks lähtuvalt andmekaevemeetodist vajalik olema
5. Hindamine – konstrueeritud on väärtuslikud mudelid ning neid hinnatakse lähtuvalt probleemi püstitusest.
6. Juurutamine – saadud teadmisi rakendatakse äriliste otsuste vastuvõtmisel, määratletakse vajalikud sammud mudelite elluviimiseks.

Et võrrelda mudeli andmestiku ulatuses tunnuste toimet, kasutatakse töös andmete standardiseerimist. Standardiseerimise käigus luuakse uus näitaja, kus tunnused on ühetaolisel skaalal keskmisega 0 ja standardhälbega 1 ehk kasutatakse tsentreerimist tunnuse keskmise suhtes. Standardiseerimist kasutatakse andmete kergemini mõistetavateks muutmiseks. Tunnuste standardiseerimise eesmärk on anda tunnustele võrdsed kaalud.

Testandmed sisaldavad sotsiaal-demograafilisi andmeid ning ostuandmeid. Sotsiaal-demograafilised andmed on esitatud andmestikus veergudes 1-5 ning ostuandmetel põhinevad andmed veergudes 6-19. Andmefaili on lisatud RFM analüüsi jaoks vajalikud väärtused ja skoorid. Testandmed sisaldavad 9 486 kliendi andmeid, mis on arvestades uuritavate koguarvu piisavalt esinduslik valim. Testandmed on koostatud 2-aastase perioodi kohta ning hõlmavad

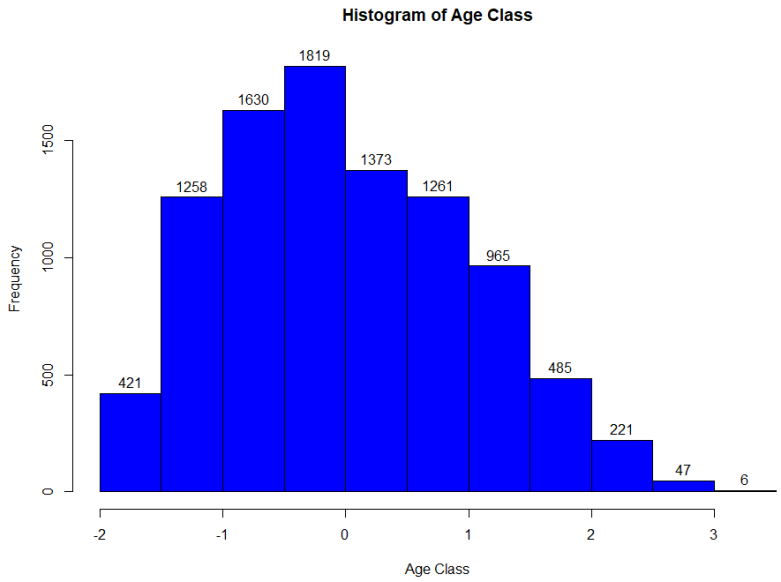
andmeid 2019-2020 aasta kohta. Alla 18-aastased kliendid on testandmete hulgast välja jäetud kuna tegemist ei ole Kaubamaja põhikliendiga.

Tabelis 1 on toodud testandmete kirjeldav statistika. Vaadeldud muutujate väärtused on üsna suure hajususega ning on mõistetavamaks muudetud standardiseerimise käigus. Hiljutisus näitab, millal klient viimati ostis. Selle põhjal saab otsustada, kui aktiivse kliendiga on tegemist. Mida väiksem on hiljutisuse näit, seda vähem on möödunud päevi kliendi viimasest ostust. Sagedus näitab kui tihti klient kaupluses käib. Sageduse suurem number viitab püsikliendile ja sageduse väiksem number juhukliendile. Rahaline väärtus näitab, kui palju klient perioodis on kokku kulutanud.

Tabel 1. Testandmete kirjeldav statistika.

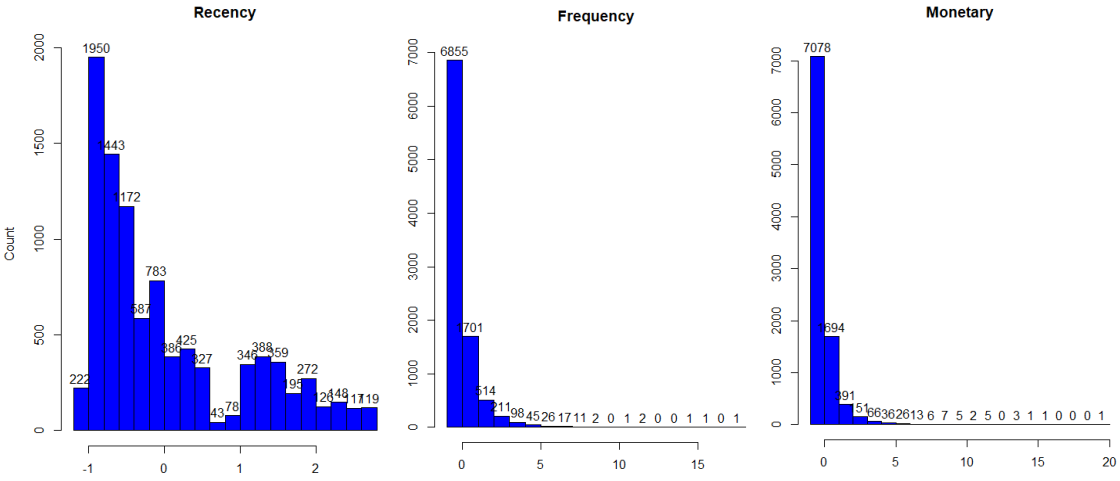
	Keskmine	Mediaan	Standardhälve	Min	Max
Vanus (age)	0	-0,09	1	-1,81	3,35
Hiljutisus (recency)	0	-0,42	1	-1,17	2,76
Sagedus (frequency)	0	-0,34	1	-0,59	17,94
Rahaline väärtus (monetary)	0	-0,31	1	-0,53	19,26
Keskmine ostusumma (avg monetary)	0	-0,27	1	-0,82	34,91

Klientide vanuselist jaotust näitab histogramm joonisel 4. Histogramm näitab klientide jaotumist sageduse järgi ning antud testandmete põhjal kuulub kõige enam kliente vanuserühma, kus on kliendid, kes on keskmisest kliendist veidi nooremad. Vähem on kliente kõige nooremas ja kõige vanemas vanusegrupis. Mediaanvanus erineb keskmisest vanusest üsna vähe.



Joonis 4. Klientide vanuseline jaotus.

Klientide hiljutisuse, sageduse ja rahalise väärtuse jaotumised on toodud histogrammil joonisel 5. Hiljutisuse osas paistab joonisel silma, et jaotumine ei toimu ühtlaselt. Väärtustega, mis langevad 0 ja 1 vahele toimub järsk kliendiarvu vähenemine. See on nii seetõttu, et andmed on 2019 ja 2020 aastate kohta, ning 2020. aasta varakevadel olid kauplused covid-19 viiruse leviku tõttu suletud. Sageduse ja rahalise väärtuse jaotumised on oodatavate tulemustega. Suure rahalise väärtuse skoori ja suure sageduse näitajaga kliente on vähe, kuid nende mõju keskmisele on suur.



Joonis 5. Hiljutisuse, sageduse ja rahalise väärtuse jaotumine sageduse järgi.

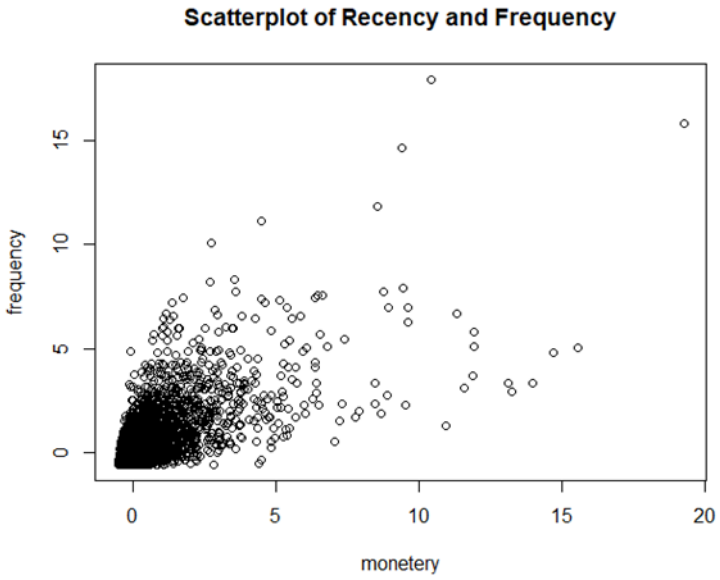
Keskmine ostusumma on sagedane KPI väärtus, mida jaekaubanduses jälgitakse. See näitab keskmist ostusummat ühe ostukorra kohta. Keskmise ostusumma jaotumise graafik on sarnane rahalise väärtuse jaotumisele. Kliente, kelle keskmine ostusumma on standardiseeritud väärtuses suurem kui 0 on vähem kui neid, kelle keskmine ostusumma jääb miinuspoolele. Hajuvus on aga selle näitaja puhul veel suurem ulatudes minimaalsest näitajast -0,82 maksimaalse näitajani, milleks on 34,91.

Mõnede valitud muutujate vahel esineb vastastikune seos. Tabelis 2 on toodud kõikide muutujate omavahelised korrelatsioonid. Muutujate omavahelistest korrelatsioonidest nähtub, et vanus ei oma märkimisväärset seost ühegi kasutatud muutujaga. Üsna nõrk on seos rahalise väärtuse ja keskmise ostusumma vahel.

Tabel 2. Segmenteerimiseks kasutatavate muutujate korrelatsioonimaatriks.

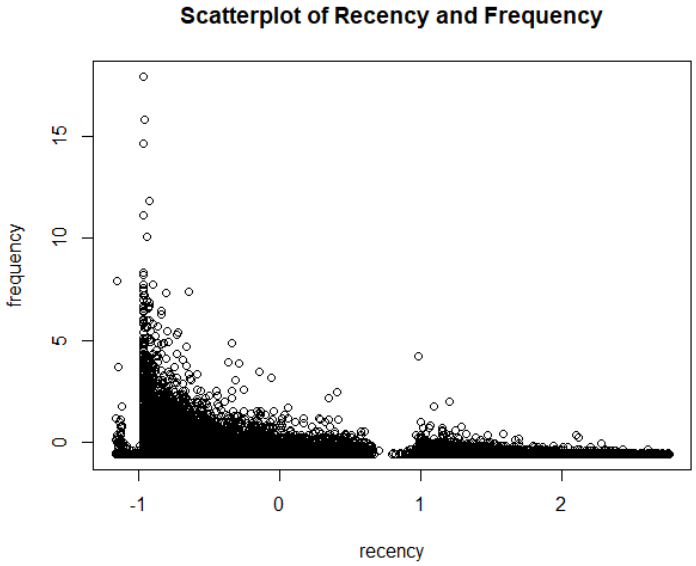
	age	recency	frequency	monetary	avg_mon
age	1,00	0,10	-0,01	0,03	0,07
recency	0,10	1,00	-0,37	-0,27	0,10
frequency	-0,01	-0,37	1,00	0,72	-0,08
monetary	0,03	-0,27	0,72	1,00	0,27
avg_mon	0,07	0,10	-0,08	0,27	1,00

Rahalise väärtuse ja ostusageduse vahel esineb aga tugev positiivne seos (r = 0,72). Mida sagedamini klient ostab, seda suurem on kulutatud ostusumma. Joonis 6 näitab graafiliselt rahalise väärtuse ja ostusageduse tugevat positiivset korrelatsiooni. Testandmestiku RFM väärtused on paarikaupa visualiseeritud hajuvusdiagrammidel joonistel 6, 7 ja 8.



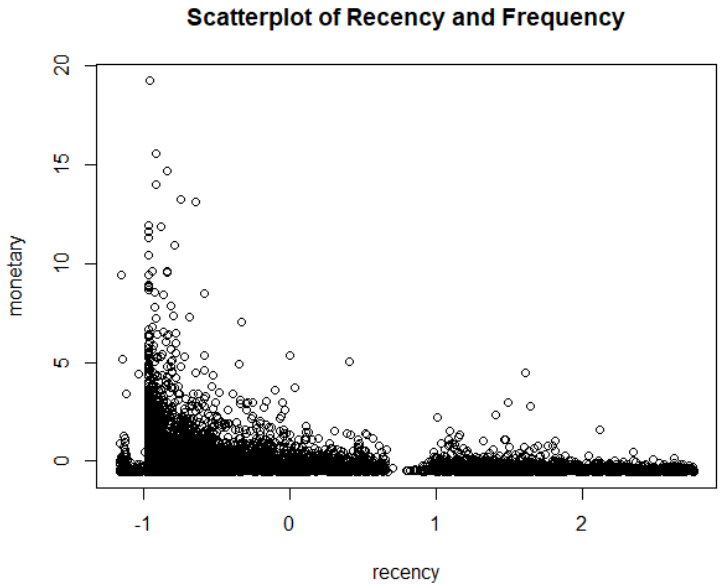
Joonis 6. Rahalise väärtuse ja ostusageduse hajuvusdiagramm.

Joonisel 7 on toodud graafiliselt hiljutisuse ja ostusageduse näitajad. Nende näitajate vahel valitseb nõrk seos (r = -0,37). Tähtsaim faktor RFM analüüsis on see, millal klient viimati ostu sooritas. Mida rohkem aega on möödunud viimasest ostust, seda väiksemaks muutub võimalus, et see klient tagasi pöördub. Jooniselt on näha, et sagedasemad ostjad on teinud oste ka hiljuti.



Joonis 7. Hiljutisuse ja ostusageduse hajuvusdiagramm.

Joonisel 8 on graafiliselt kujutatud hiljutisuse ja rahalise väärtuse seosed. Nende vahelised seosed on kõige nõrgemad (r = -0,27). Ostusageduse faktor aitab eristada püsikliente uutest klientidest. Madal ostusageduse näitaja tähendab seda, et klient on kas juhuklient või kui samaaegselt on ka hiljutisuse näitaja väike, siis uus klient.



Joonis 8. Hiljutisuse ja rahalise väärtuse hajuvusdiagramm.

4. Meetodi rakendamine ettevõtte andmetel

Peatükis kirjeldatakse RFM – analüüsi kriteeriumide leidmist ning analüüsi läbiviimist. Luuakse maatriks RFM mudeli skooride hindamiseks. Kasutatakse peakomponentanalüüsi ja selle tulemusel valitud peakomponente. Leitakse elbow meetodil sobiv klastrite arv ja viiakse läbi klasteranalüüs.

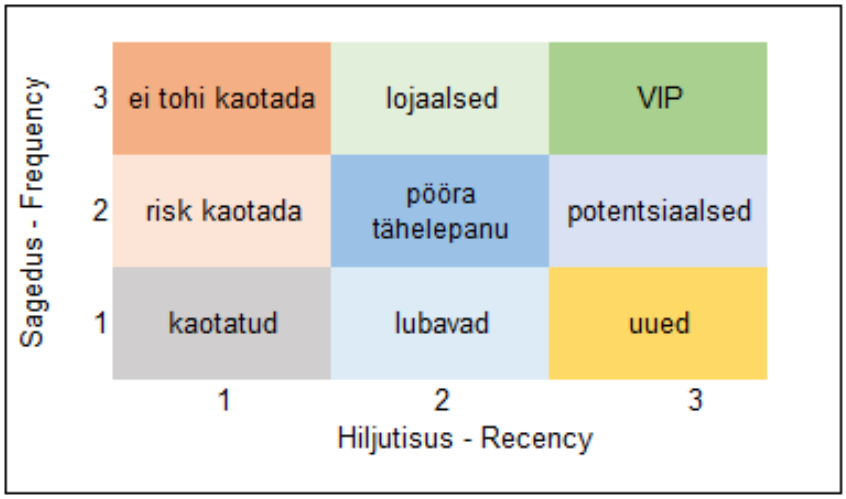
4.1 RFM analüüs Kaubamaja andmetel

Selleks, et selgitada välja, kes on uuritavas ettevõttes kasumlikumad kliendid ja kui suur on nende osakaal ning, kes on kliendid, keda kaotada ei tohi, viiakse käesolevas töös läbi RFM analüüs. RFM analüüsi jaoks on autor loonud lihtsad ja praktilised kriteeriumid RFM skoori arvutamiseks igale kliendile. Hiljutisus leitakse viimase ostu kuupäeva järgi ja arvutatakse ümber päevade arvuks konkreetsest kuupäevast tagasi. Sageduse arvutamiseks loetakse iga kliendi ostupäevade arv ehk ridade arv faktitabelis. Rahaline väärtus on kliendi ostusumma ehk iga kliendi kokku liidetud käive. Need kriteeriumid on toodud tabelis 3 ning väärtused 1 ja 3 on saadud variatsioonirea väärtuste alumise ja ülemise kvartiili põhjal.

Tabel 3. RFM analüüsi kriteeriumid.

Muutuja	Väärtus 1	Väärtus 2	Väärtus 3
Hiljutisus R	>0,47	0,47 kuni -0,77	< -0,77
Sagedus F	< -0,51	-0,51 kuni 0,08	> 0,08
Rahaline väärtus M	< -0,44	-0,44 kuni 0,01	> 0,01

RFM mudeli skooride hindamiseks on autor koostanud 3 x 3 maatriksi sageduse ja hiljutisuse põhjal. Maatriks on ära toodud joonisel 9 ning selle põhjal toimub edasine kliendi analüüs nende klientide osas, kelle rahalise väärtuse skoor on kas 2 või 3.



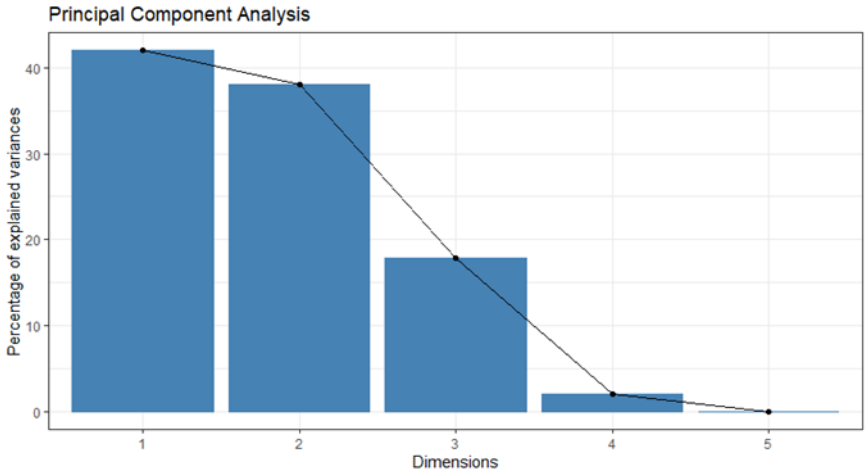
Joonis 9. RFM mudeli skooride hindamise maatriks.

1. VIP kliendid on kõige olulisemad kliendid, keda tuleb hoida.
2. Lojaalsed kliendid on muutumas passiivsemaks ning vajavad tähelepanu, et neid mitte kaotada.
3. Ei tohi kaotada kliendid on kunagised VIP kliendid ning kohe tegutsedes on neid võimalik veel tagasi saada.
4. Potentsiaalsed ja uued kliendid on potentsiaalsed tulevased VIP kliendid, kellele tuleks tutvustada VIP kliendi eeliseid ning teha neist oma lojaalsed kliendid.
5. Pööra tähelepanu ja lubavad kliendid on tavakliendid, kellega regulaarset kommunikatsiooni hoida.
6. Kaotatud ja risk kaotada kliendid on kliendid, kellele panustamine ei too suurt efekti.

Kaardistades kliendid RFM meetodil, saab edaspidi analüüsida, mille poolest erinevad ühe segmendi ostuharjumused teise segmendi omadest. Sellest tulenevalt on võimalik teha muutuseid, et tuua rohkem kliente sellesse segmenti, mis on meile kasulikumad.

4.2 Peakomponentide analüüs

Klientide kirjeldamiseks on meil andmebaasis väga palju tunnuseid. Need tunnused on olulised kliendi kirjeldamiseks, samas teevad aga analüüsi keeruliseks ning raskendavad tulemuste tõlgendamist. Alternatiivne meetod nende kõikide muutujate vähendamiseks on peakomponentide meetod. Peakomponentidel on teatud optimaalsed omadused, mis muudavad need võimsaks. See võimaldab suurest hulgast andmetest leida kasulik informatsioon, mille abil saada klientide iseloomustus, kaotamata sealjuures olulist informatsiooni. Peakomponentide meetod leiab uued tunnused, mis on algsete tunnuste kombinatsioonid. Need lineaarkombinatsioonid moodustatakse nii, et esimene peakomponent kirjeldab ära võimalikult suure osa algtunnuste koguvarieeruvusest. Teine peakomponent kirjeldab ära võimalikult suure osa alles jäänud varieeruvusest jne [1]. Joonisel 10 on toodud analüüsis kasutatavad peakomponendid ning osatähtsused, mille iga peakomponent ära kirjeldab.



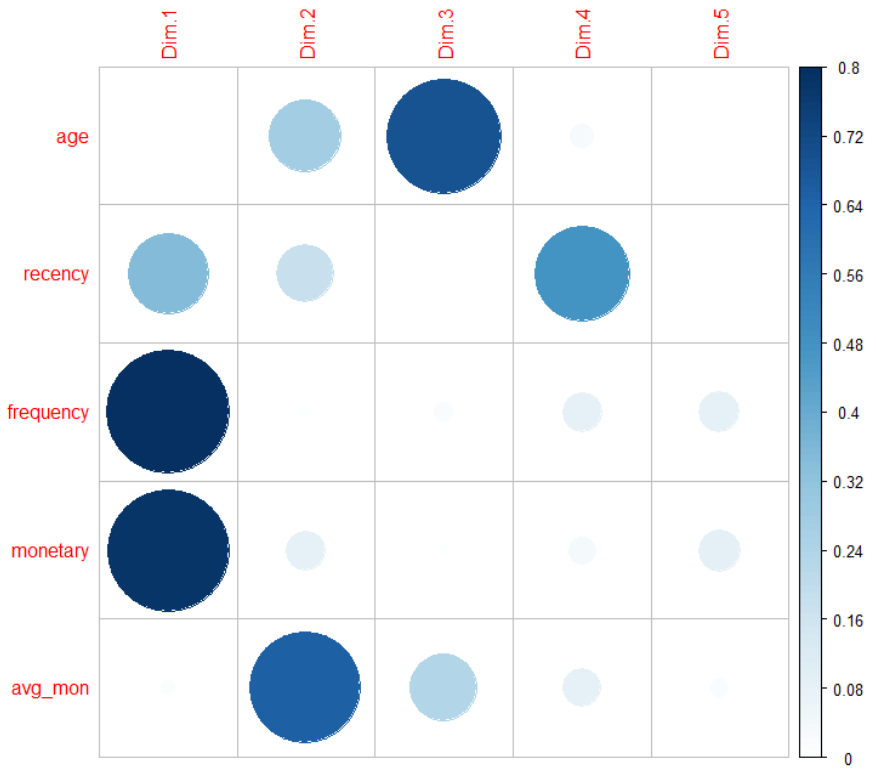
Joonis 10. Peakomponendid ning nende osatähtsused.

Peakomponendi väärtus mingile objektile leitakse summeerides algsete atribuutide väärtuste ja vastavate peakomponentide kaalude korrutised. Selleks luuakse kovariatsioonimaatriks ja iga PCA kaaluvektor on maatriksi kovariatsioonimaatriksi omavektor. Seejärel arvutatakse kovariatsioonimaatriksile omaväärtused ja omavektorid.

Tabel 4. Peakomponendid ja atribuudid.

	age	recency	frequency	monetary	avg_mon
1	0.11469493	-0.8536231	3.25756927	3.19760865	0.19100051
2	0.14645114	0.3422908	-0.41818491	0.52553534	3.33442527
3	0.07871193	1.6295296	-0.49867012	-0.39116799	-0.05119533
4	-0.74624333	-0.5027745	-0.03412650	-0.13744399	-0.20865261
5	0.99932609	-0.4023490	-0.04475227	-0.09180902	-0.13522450

Peakomponendid ja atribuudid on toodud tabelis 4 ning peakomponentide leidmine on visuaalselt kujutatud joonisel 11. Nii on näha, et esimene peakomponent koosneb sagedusest, rahalisest väärtusest ja hiljutisusest ning vanuse ja keskmise ostu osa on väikese osatähtsusega. Teine peakomponent aga põhineb rohkem vanusele ja keskmisele ostule, hiljutisus ja rahaline väärtus on samuti esindatud, kuid sageduse osa väiksem.



Joonis 11. Peakomponendid ja atribuudid

Kuna algtunnused on standardiseeritud, siis on nende hajuvuse väärtus 1. Kuna peakomponendid olid moodustatud algtunnustest nii, et esimeste hajuvus oleks võimalikult suur ja viimastel võimalikult väike, siis on esimeste peakomponentide dispersioon suurem kui 1, mis on suurem kui üksikutel algtunnustel ning viimastel väiksem kui 1, mis on väiksem kui

üksikutel algtunnustel. Tavaliselt võetakse mudelisse need peakomponendid, mille kirjeldusvõime on suurem kui üksikutel algtunnustel st mille omaväärtus on suurem kui 1. Peakomponentide olulisused on toodud tabelis 5. Viimaste peakomponentide mudelist välja jätmisega on vähendatud algtunnuste eripärast tingitud variatiivsust. Heaks peetakse tavaliselt mudelit, mis kirjeldab üle 60% algtunnuste variatiivsusest.

Tabel 5. Peakomponentide olulisus.

Importance of components:

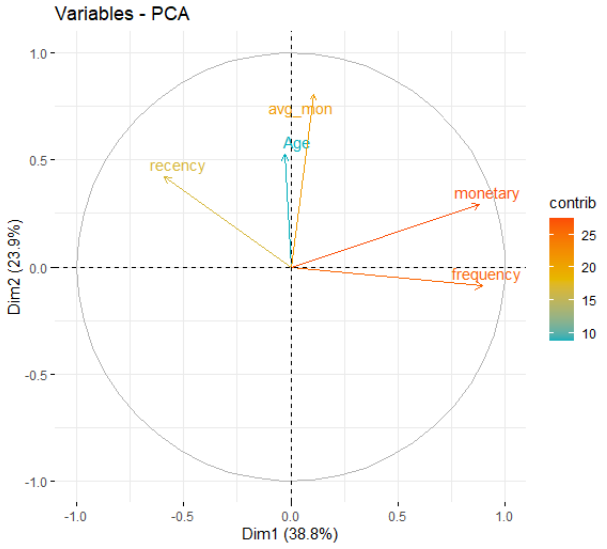
	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.3936	1.0943	0.9771	0.8409	0.44563
Proportion of Variance	0.3884	0.2395	0.1909	0.1414	0.03972
Cumulative Proportion	0.3884	0.6279	0.8189	0.9603	1.00000

Peakomponentide meetodi puhul nagu mitmedimensionaalses analüüsis, kus üritatakse suurt hulka andmeid selgemalt ja lihtsustatult esitada, läheb mingi hulk andmeid kaduma. Seetõttu on oluline määrata piir, milleni on andmeid mõtet lihtsustada, et liiga palju infot kaduma ei läheks, aga et andmed oleksid samal ajal selgemalt esitatud. Tabelis 6 on toodud esimesed kaks peakomponenti, mis kirjeldavad rohkem kui 60% algtunnuste variatiivsusest (62,8%).

Tabel 6. Esimesed 2 peakomponenti.

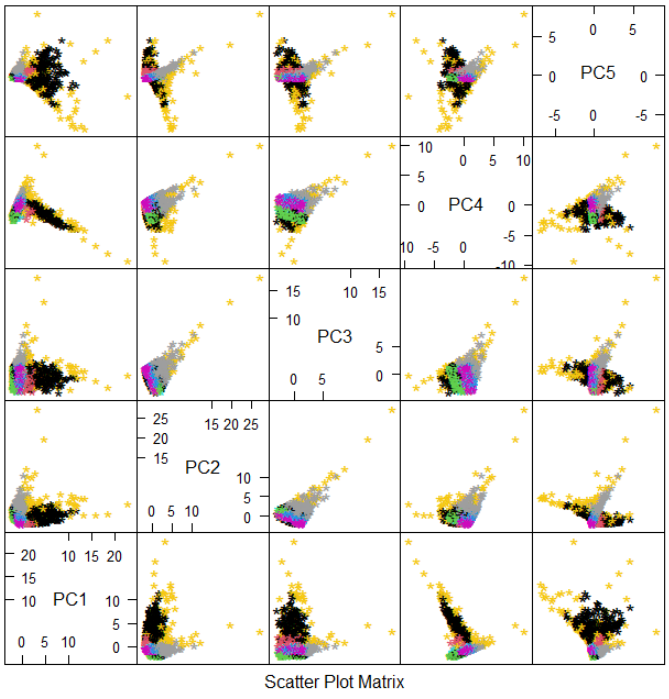
	PC1	PC2
age	-0.02023340	0.47936537
recency	-0.42309470	0.38457636
frequency	0.64301387	-0.08172118
monetary	0.63334246	0.26492824
avg_mon	0.07740785	0.73854217

Analüüsi atribuudid on visuaalselt kujutatud kahe peakomponendi põhjal joonisel 12.



Joonis 12. Esimesed kaks peakomponenti ning atribuutide väärtused.

Hajuvusdiagrammid joonisel 13 näitavad viie peakomponendi hajuvusi paarikaupa võrrelduna. Nii on vasakus alumises nurgas hajuvusgraafikul kaks esimest peakomponenti.



Joonis 13. Viie peakomponendi hajuvusdiagrammid

Kuigi peakomponentide puhul ei ole arvesse võetud kõiki mõõtmeid ja punktid, mis tunduvad teineteisele lähedal, võivad olla üksteisest kaugel, on peakomponendid parim viis paljude mõõdete esitamiseks ning nende uurimine võib anda uusi tähendusi.

4.3 Klasteranalüüs

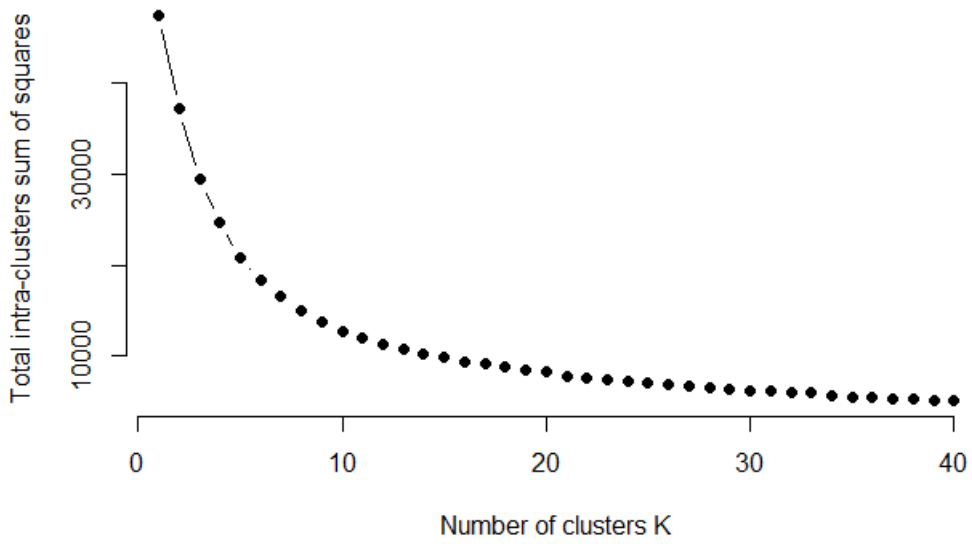
Andmetest mustrite otsimiseks ja tuvastamiseks kasutatakse käesolevas töös k-keskmise eraldusalgoritmi. K-keskmise algoritm on lihtne, kiire ning töötab hästi olukorras, kus segmenteeritavaid objekte on palju ning andmestik suur. Hierarhiline ja mudelipõhine algoritm eeldavad täielikku kauguste maatriksi arvutamist ning suurt mälumahtu. Need sobivad pigem tunnuste kui klientide klasterdamiseks. Rühmitamisel algoritmi kasutamise eesmärk on, et andmepunktide vaheline kaugus klastri sees on väike võrreldes kahe klastri vahelise kaugusega [1]. See tähendab, et rühma liikmed on sarnased ja eri rühmade liikmete vahelised erinevused suured. K-keskmise algoritmi kasutamise eesmärk antud töös on klientide parem mõistmine, mida saaks kasutada omakorda turundustegevuste efektiivsemateks tulemusteks.

4.4 Klastrite arvu leidmine

Ärilise eesmärgi püstitamisel ei ole kindel klastrite arv paika pandud. Tingimuseks on antud vaid, et klastreid/ segmente ei oleks rohkem kui 5 kuna eesmärk on luua igale segmendile sobivad pakkumised. Üle viie erineva pakkumise ei ole äriliselt mõistlik luua.

Optimaalsete klastrite arvu leidmiseks on mitmeid erinevaid meetodeid. Selleks, et leida sobiv klastrite arv, kasutatakse antud töös kõige tuntumat ehk elbow meetodit. Elbow meetod ehk küünarnuki meetod aitab välja selgitada optimaalse klastrite arvu, sobitades mudeli ***k*** väärtuste vahemikuga. Selle rakendamisel saab jooniselt välja lugeda sobivaima klastrite arvu. Klastrite arv leitakse visuaalselt. Elbow meetodi puhul otsitakse punkti, kus klastsrite arvu ***k*** korral väheneb klastrisiseste hajuvuste summa ja edasine vähenemine on ühtlane. Elbow meetodit ei peeta kõige täpsemaks optimaalsete klastrite leidmise meetodiks, kui punkt, millest alates hajuvus väheneb pole leitav. Käesolevas töös jäädakse Elbow meetodi juurde ja peetakse

silmas, et klastreid ei saaks olema rohkem kui viis. Visuaalselt võiks luua klastreid kas neli või viis. Elbow meetodil leitud optimaalsete klastrite arvu saab lugeda jooniselt 14.



Joonis 14. Elbow meetodil leitud optimaalsete klastrite arv.

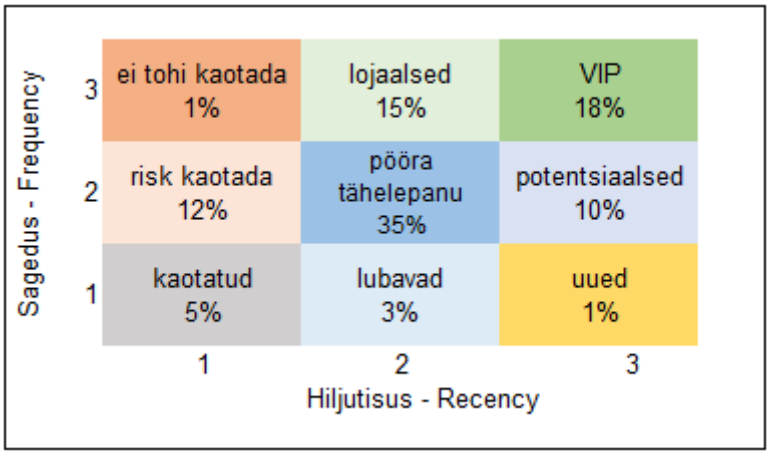
Mudeli olulisus nelja klastri põhjal on 47,9% ja viie klastri põhjal 56,1% (between_SS / total_SS = 56,1%), mis võiks olla veidi kõrgem.

5. Rakendatud meetodi tulemused

Selles peatükis kirjeldatakse RFM - analüüsi tulemusi. Leitakse klientide osakaalud sageduse ja hiljutisuse alusel. Teiseks kirjeldatakse klasteranalüüsi tulemusi nelja ja viie klastri põhjal ning tehakse ettepanek võtta kasutusele viiel klastril põhinev mudel.

5.1 RFM analüüsimeetodi rakendamise tulemused

FRM meetodi puhul kasutati müügitehingute ajaloo andmeid, et jagada kliendid ostukäitumise alusel väiksematesse segmentidesse. Meetodi reeglistiku alusel on kliendid jagatud erinevatesse segmentidesse selle järgi, kui lojaalsed ja kasumlikud nad on. RFM analüüsi tulemusena jaotuvad kliendid üheksaks rühmaks, mis on välja toodud joonisel 15.



Joonis 15. Klientide osakaalud RFM mudeli skooride hindamise maatriksil.

RFM analüüsi tulemusel selguvad kliendid, keda kaotada ei tohi. Lisaks saab analüüsi tulemusel hinnata, millal klient viimati ostis ning kui suur on tema ostude sagedus. Analüüsi tulemusel saab hinnata kliendibaasiga seotud riske. Mida rohkem aega on möödunud viimasest ostust, seda väiksem on võimalus, et klient tagasi pöördub. Kliendid, kes ostavad tihti ostavad tõenäolisemalt uuesti, kui need, kelle ostud on harvad.

Analüüsis on nende klientide andmed, kelle rahalise väärtuse skoor oli 2 ja 3 ehk rahalises väärtuses on tegemist heade klientidega. Kui varasem hea klient ei osta enam nii tihti või samas

väärtuses, mis varem, on seda võimalik veel parandada. Kui loobujateks on parimad ehk VIP kliendid, siis on see ettevõtte jaoks juba probleem.

Kliendibaasi suurima osa 35% moodustavad ’pööra tähelepanu’ kliendid, kellele tuleb oma kliendisuhtlus koondada. Koos ’ei tohi kaotada’ ja ’risk kaotada’ klientidega moodustavad need kliendid 48% headest klientidest, kelle rahalise väärtuse skoor on hea. Need kliendid vajavad ekstra tähelepanu, et neid mitte kaotada. Nende klientide puhul tasub mõelda personaalse pakkumise peale, mis annab lisasoodustust ning tuletab neile meelde, mis on Kaubamaja eelised ja miks see neile ostukohana meeldib. ’Risk kaotada’ ja ’ei tohi kaotada’ kliendid on kunagised VIP kliendid ja kohe tegutsedes on viimane võimalus neid tagasi saada. 11% klientidest ehk ’uued’ ja ’potentsiaalsed’ kliendid on need, kellele tuleb tutvustada VIP kliendi eeliseid ja teha neist oma fännid. ’Kaotatud’ klientidele panustamine enam suurt mõju ei avalda. Neid on kliendibaasis 5%. ’Lojaalsed’ ja ’VIP’ kliendid peavad aga kindlasti saama infot uue hooaja toodete ja pakkumiste kohta.

5.2 Klasteranalüüsi tulemused

Ärieesmärki silmas pidades ning Elbow meetodit ehk küünarnuki meetodit kasutades selgub, et optimaalne klastrite arv on 4 või 5. Seega vaadatakse käesolevas töös nii 4 kui 5 klastri põhist segmenteerimist ning otsustatakse peale klastrite visualiseerimist, kumb variant on ärieesmärkide saavutamiseks efektiivsem.

5.2.1 Klasterdamine nelja klastri põhjal

K-keskmise klastedamise tulemusel nelja klastri põhjal on kliendid jaotatud neljaks rühmaks, mille suurused on 467, 2195, 4030, 2794. Suurim rühm moodustab klastri 3 ning sinna kuulub 4030 valimi klienti ehk 42% valimist. Väikseim rühm on klastris 1, kuhu kuulub 467 valimi klienti ehk 5% klientidest. Ülejäänud kahte klastrisse kuulub vastavalt 23% ja 29% klientidest. Tabelis 7 on toodud k-keskmise klasterdamise tulemused.

Tabel 7. K-keskmise klasterdamise tulemused nelja klastri põhjal.

	age	recency	frequency	monetary	avg_mon
1	0.10523950	-0.8309709	3.08117689	3.16603828	0.457516568
2	0.08483214	1.6009933	-0.49943205	-0.34640813	0.221952673
3	-0.74403538	-0.5015257	-0.04046826	-0.13265165	-0.171589323
4	0.98894387	-0.3954790	-0.06426957	-0.06570791	-0.003343373

1	2	3	4
467	2195	4030	2794

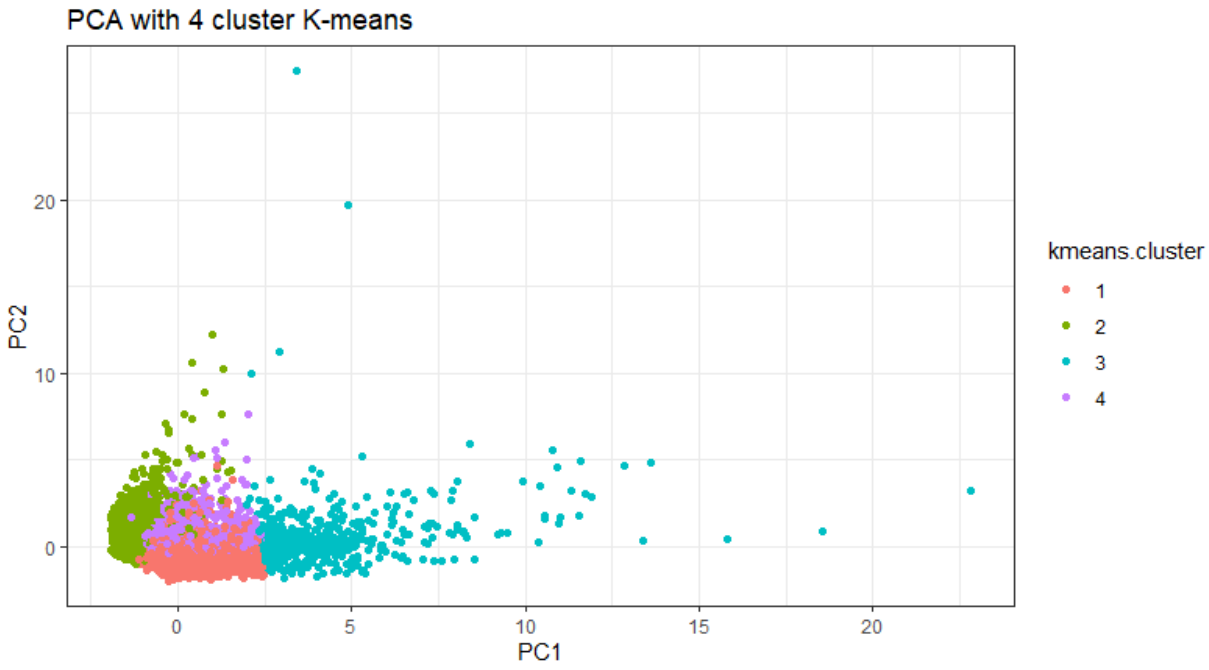
Kõige väiksema klastri, klastri nr 1 klientide kulutused on kõige suuremad. Samuti on selle klastri keskmine ostusagedus teiste klastritega võrreldes kõrgem ning hiljutisus kõige väiksem. See tähendab, et nad on poodi külastanud üsna hiljuti. Neid kliente võib nimetada VIP-klientideks. Selle klastri klientide keskmine ostusumma on oluliselt kõrgem kui teistes klastrites. Vanuse poolest on kõige väiksema klastri kliendid keskmisest veidi vanemad.

Suurima klastri, klastri nr 3 kliendid on kõige nooremad. Nende viimasest poekülastusest ei ole möödunud väga palju aega, kuid nende ostusummad ei ole kõige suuremad. Ühe ostukorra keskmine summa on selles klastris võrreldes teiste klastri klientidega oluliselt väiksem. Nende hiljutisuse näitajat arvestades võib üsna kindel olla, et nad poodi tagasi tulevad. Selle klastri kliendid on hiljutisuse ja sageduse näitajatele tuginedes uued kliendid. Õige kommunikatsiooni korral aga lootustandvad tuleviku VIP kliendid.

Teistest klastritest oluliselt kõrgema vanusega klastrisse, klastrisse nr 4, kuulub kolmandik kliente. Nende klientide kulutused on üsna suured, jäädes siiski alla VIP klientide ostudele. Siiski on vanima vanusega klientide klastrisse kuuluvate klientide näol tegemist Kaubamaja lojaalsete klientidega, kes külastavad poodi regulaarselt ning sama regulaarselt tuleb nendega ka kommunikatsiooni hoida.

Viimase klastri, klastri nr 2, klientide viimasest poekülastusest on möödunud kõige enam aega. Selle klastri kliendid külastavad poodi harva, kuid nende keskmine ostusumma on suurem kui klastritel 3 ja 4. Nende kulutused ostudele on võrreldes teiste klastritega kõige väiksemad, mis on samas seletatav sellega, et nad jõuavad kauplusesse kõige harvem. Kuigi hiljutisuse näitaja põhjal võib arvata, et siia klastrisse kuuluvad kliendid, kelle oleme kaotanud, esineb võimalus, et kohe tegutsedes on võimalik neid veel tagasi saada.

Kõik neli klastrit on kujutatud joonisel 15, mis näitab klastreid kahe peakomponendi põhjal.

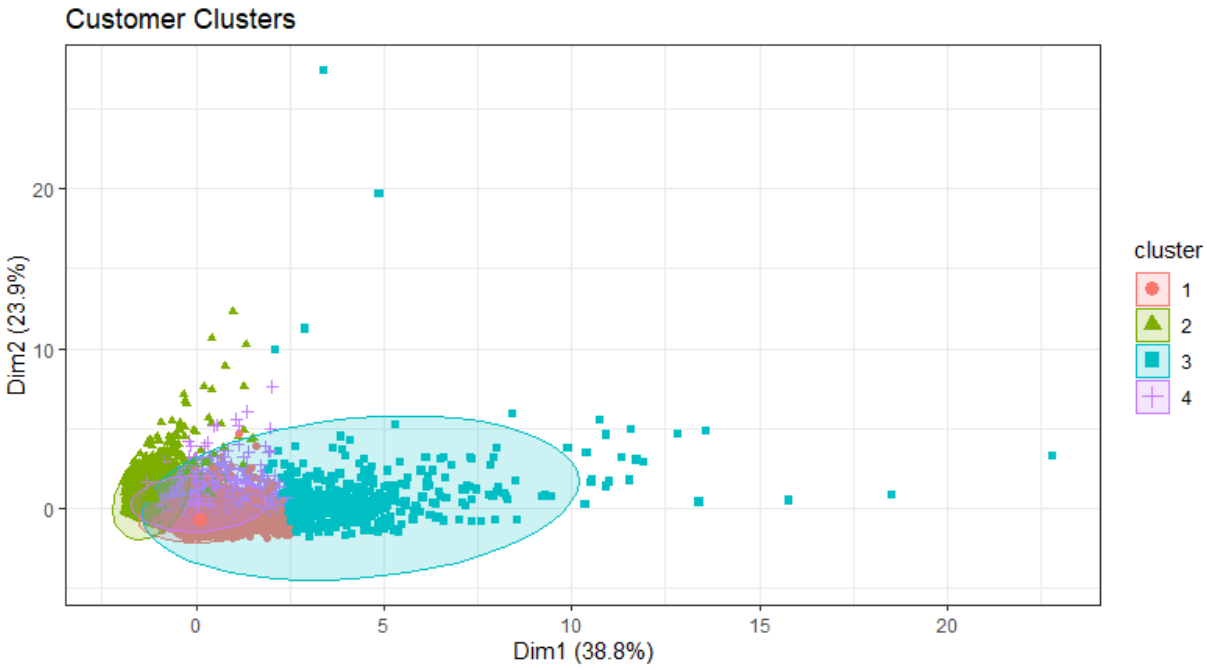


Joonis 16. K-keskmise klasterdamise tulemused 4 klastri põhjal koos PCA-ga.

Joonis 16 põhjal on näha, et esimese peakomponendi PC1 kirjeldusvõime on suurem. See tuleneb sellest, et esimene peakomponent kirjeldab ära võimalikult suure osa algtunnuste koguvarieeruvusest.

Kõige väiksem klaster on joonisel klaster 4, millesse kuuluvad VIP kliendid. Suurim on klaster 3, mille kliendid olid kõige nooremad. Klaster, mille kliendid ei ole viimasel ajal poodi jõudnud, on joonisel klaster 1. Suurima keskmise vanusega klaster on joonisel klaster 2.

Kui püüda nelja klastrit visuaalselt rohkem eraldada nagu on seda tehtud joonisel 17, on näha, et klastrid mõnel määral kattuvad ja nende piirid on hägusad. Selgelt eristuvad klastrid 2 ja 3 ning klastrid 1 ja 4 on omavahel lähedasemad.



Joonis 17. K-keskmise klastrite eraldamine nelja klastri põhjal.

5.2.2 Klasterdamine viie klastri põhjal

K-keskmise klasterdamise tulemusel valimi andmete põhjal saab öelda, et kuigi kliendi keskmisel vanusel on teatav mõju kliendi ostusuurusele ja ostusagedusele, ei ole vanus klastritesse jaotamise põhitunnus. Olulisemalt suurem mõju on hiljutisusel ja sagedusel. Tabelis 8 on klastrite suurused ja tulemused viie klastri põhjal. Viies klaster on saadud eelmise joonise teise ja neljanda klastri muutmisel (VIP klastri ja suurima keskmise vanusega klastri muutmisel).

Tabel 8. K-keskmise klasterdamise tulemused viie klastri põhjal.

	age	recency	frequency	monetary	avg_mon
1	0.11469493	-0.8536231	3.25756927	3.19760865	0.19100051
2	0.14645114	0.3422908	-0.41818491	0.52553534	3.33442527
3	0.07871193	1.6295296	-0.49867012	-0.39116799	-0.05119533
4	-0.74624333	-0.5027745	-0.03412650	-0.13744399	-0.20865261
5	0.99932609	-0.4023490	-0.04475227	-0.09180902	-0.13522450

1	2	3	4	5
437	363	2038	3957	2691

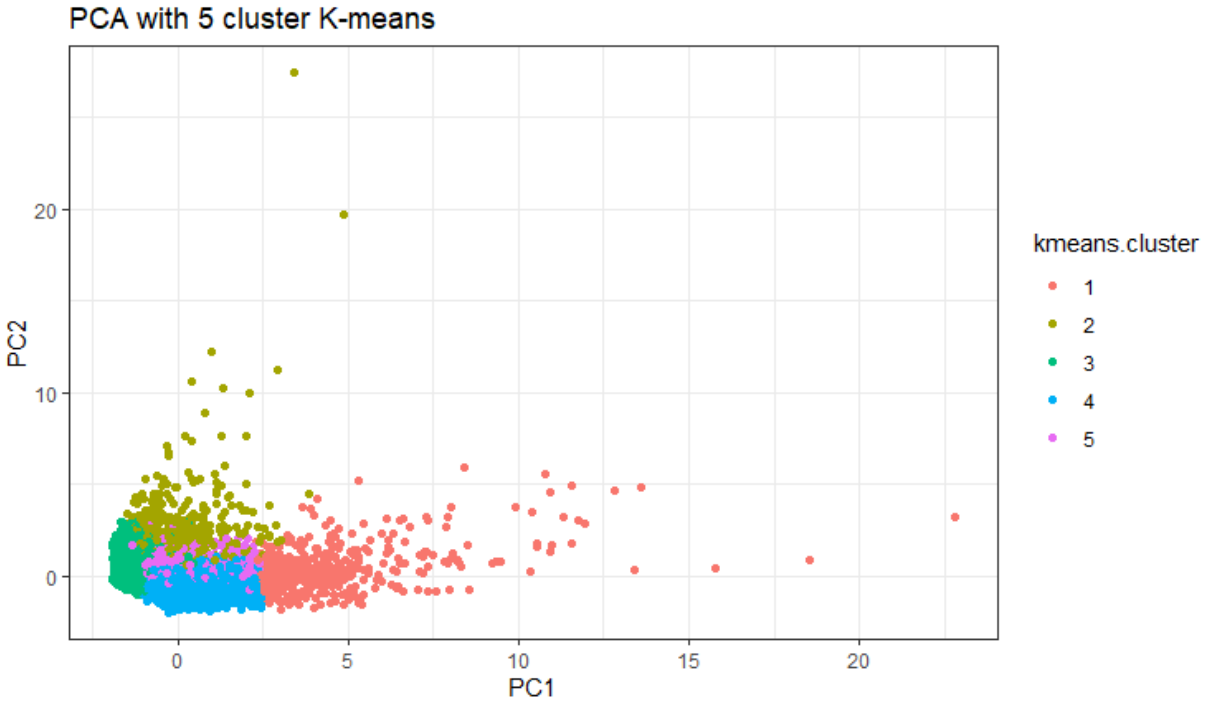
Suurim klaster, klaster 4, moodustab endiselt 42% valimist ning selle klastri kliendid on keskmisest nooremad. Nad on külastanud poodi üsna hiljuti ja teevad seda üsna sagedasti, jäädes siiski hiljutisuse ja sageduse poolest alla VIP klientidele. Nende kulutused ostudele on keskmisest veidi väiksemad ning korraga nad suuri oste ei tee, ühe ostu keskmine summa on teiste klastritega võrreldes oluliselt väiksem. See on klaster, kellele tasub suunata peamine tähelepanu ja kes on mõjutatav kommunikatsiooniga. Ilmselt kuuluvad siia klastrisse ka uued kliendid. Nende hiljutisus on suhteliselt väike, aga samal ajal on väike ka nende sagedus.

Kõige väiksem klaster, klaster 2, moodustab 4% valimist. Vanuse poolest kuuluvad siia kliendid, kes on sarnased klastrile 1. Nende poe külastuse hiljutisus, sagedus, rahaline väärtus ja keskmine ostusumma erineb aga klastrist 1 suurel määral. Kõige väiksema klastri kliendid ei ole hiljuti poodi külastanud. Nende poe külastuse sagedus on keskmisest väiksem. Küll aga on keskmisest oluliselt suurem nende keskmine ostusumma ja ka kogu perioodi ostud on neil keskmisest suuremad. Need kliendid külastavad poodi harva, aga teevad sel juhul suuri oste korraga.

5% moodustavad valimist ka klastri 1 kliendid. Need kliendid on keskmisest veidi vanemad. Kauplust on nad külastanud ilmselt eile, nende hiljutisus on teiste klastritega võrreldes oluliselt väiksem. Ostusagedus on sellel klastril oluliselt suurem ning perioodi ostusumma on teiste klastrite omast oluliselt kõrgem. Keskmine ostusumma jääb alla küll klastile 2, kuid klastri 1 kliendid käivad kaupluses lihtsalt tihedamini. Selle klastri kliendid on lojaalsed VIP kliendid.

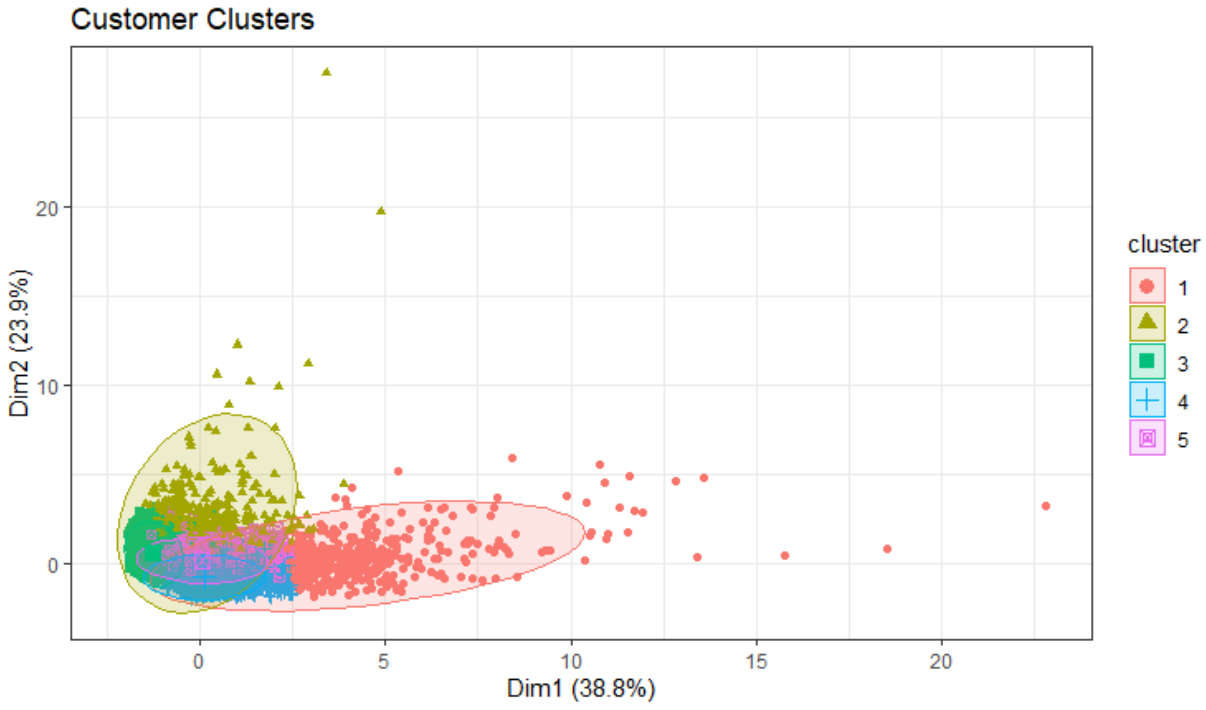
Klastrisse number 3 kuuluvad kliendid on keskmise vanusega. Nende viimasest poekülastusest on möödunud kõige enam aega. Kauplusesse satuvad nad harva ning nende rahalised kulutused on teiste klastritega võrreldes kõige väiksemad. Ka on väike nende keskmine ostusumma. Ilmselt on selle klastri klientide näol tegemist kaotatud klientidega, kellele kommunikatsioon väga suurt tulemuslikkust ei anna.

Klastrisse number 5 kuuluvad kõige vanemad kliendid. Selle klastri keskmine vanus on teiste klastritega võrreldes oluliselt kõrgem. Tegemist on lojaalsete klientidega, kes külastavad Kaubamaja üsna regulaarselt, kuid nende ostudele kulutatud summa jääb teistele klastritele oluliselt alla. Võib eeldada, et need kliendid ei ole kõige kallimate brändide ostjad, sest madal on ka nende keskmine ostusumma. Visuaalselt on 5 klastrit kujutatud joonisel 18 kahe peakomponendi põhjal.



Joonis 18. K-keskmise klasterdamise tulemused viie klastri põhjal koos PCA-ga.

Viie klastri visuaalsel eraldamisel joonisel 19 on näha, et klastrid 1,4 ja 5 on pigem põhinevad esimesele peakomponendile ja klastrid 2 ja 3 on kirjeldatud rohkem teise peakomponendi põhjal.



Joonis 19. K-keskmise klastrite eraldamine viie klastri põhjal.

Olles klasterdamise tulemusi analüüsinud nelja ja viie klastri põhjal, nähtub, et viis klastrit selgitavad kliendibaasi paremini lahti. Ka on viiel klastril põhineva mudeli olulisus kõrgem. Kui eesmärk on välja töötada kommunikatsiooni strateegia, siis võttes aluseks viie klastri mudeli, piisab kui luua strateegia neljale klastrile. Kaotatud klientide klastrile, kes moodustasid valimist 20% ei anna kommunikatsioon ilmselt tulemuslikkust.

6. Tulemuste hindamine

Peatükis testitakse mudeli tulemuslikkust A/B testimise läbi. Luuakse valimid uudiskirja saatmiseks ja uuritakse erinevate valimite konversioonimäärasid. Selgub, et valimite konversioonimäärad on üsna erinevad.

Mõõtmise ja pideva parendamise väärtust tunnustatakse küll laialdaselt, kuid ometi on see siiski vähem tähelepanu saanud, kui see väärt oleks kuna sellel puudub kohene investeeringutasuvus. Paljusi ärijuhtumeid rakendatakse, ilma et keegi läheks tagasi vaatama, kui hästi reaalsus plaanidega sobis. Andmetel põhinevad otsused peaksid andma paremaid tulemusi. Turunduslik kasu peaks avalduma finantsandmetes ning mudeli kasulikkust peab seega saama mõõta.

Kaubamaja saadab regulaarselt oma klientidele e-posti teel uudiskirju ning jälgib, kui palju nendest kirjadest avatakse ning, kes uudiskirja saajatest ka ostu sooritab. Seega mõõdetakse CR-i näitajat. E-postiturundust iseloomustab vähene tähelepanu määr, kuid selle väärtus seisneb automatiseerimises ning suures personaliseerituse määras. Erinevad uuringud leiavad, et tänu personaliseerimisele on tulemus efektiivsem, sest klikkimise määr võib olla kuni 2 korda kõrgem [11].

6.1 A/B testimine

Kui uudiskirja väljasaatmise järgselt müügid kasvavad, jääb ikkagi küsimus, kas oleksime saanud teha midagi paremini. Selleks, et hinnata uudiskirja konversiooni, saab kasutada A/B testimist. Tegemist on kvantitatiivse testimismeetodiga, mis antud juhul aitab hinnata loodud mudeli efektiivsust. Läbi A/B testimise saab proovida kahte erinevat strateegiat või lähenemist ning uurida, kumb lahendus paremini töötab. A/B testimine on turunduses laialt kasutatav meetod, mis testib kõige sagedamini tarkvara ja veebirakendusi, kuid selle kasutamise valdkond on väga lai.

Uudiskirja eesmärk on tõsta müüke ning tutvustada klientidele uue kollektsiooni kaupu. See eesmärk on hõlpsalt mõõdetav läbi konversioonimäära, mis näitab veebilehe külastajate hulka, kes on läbinud ostuotsustusprotsessi ning muutunud veebilehe külastajast ettevõtte toote tarbijaks.

A/B testimise käigus võrreldakse kahte versiooni (A ja B), mis on suures osas identsed, kuid erinevad ühe potentsiaalse kasutaja tegutsemist mõjutava detaili poolest. Versioon A on tihti see, mis on hetkel kasutusel, ning versiooni B on mingil moel modifitseeritud. Antud töö käigus loodud mudeli testimiseks kasutatakse kahte uudiskirja sihtgruppi, et veenduda kas loodud mudel töötab. Lisaks kasutatakse kontrollgruppi, kes uudiskirja ei saa, kuid kelle ostukäitumist samal perioodil hinnatakse.

A/B testimise grupid:

- Grupp A – Kaubamajas praegu kasutusel olev demograafilistel näitajatel põhinev uudiskirja valim.
- Gupp B – Klasteranalüüsi tulemusel tegelemist vajavasse klastrisse kuuluvad kliendid, kellele Kaubamaja turundustegevused peamiselt peaks suunatud olema.
- Kontrollgrupp – Klasteranalüüsi tulemusel tegelemist vajavasse klastrisse kuuluvad kliendid, kellele uudiskirja ei saadeta, kelle ostukäitumist samal perioodil jälgitakse.

Grupile A ja B saadetakse samal päeval täpselt ühesugune uudiskiri ning kõigi kolme grupi ostutegevusi jälgitakse kahe päeva vältel.

6.2 Mudeli hindamine

Uut mudelit ei tohiks rakendada enne, kui see näitab paremaid tulemusi kui vana. A/B testimise käigus on saadud konversioonimäärad kõigile kolmele grupile, et hinnata klientide käitumist. Välja tuleb tuua, et see on see osa töö tsüklist, mida kasutamise käigus tuleks aegajalt üle vaadata ja vajadusel segmenteerimises muudatusi teha. Mis tõstab ühe kliendi tulemuslikkust, on kasulik kogu organisatsioonile. Valimid moodustati klasterdamise tulemusel tegelemist vajavate klientide klastrite põhjal ja lähtuti kõige tõenäolisemalt reageerivatest klientidest (kelle hiljutisus ei ole madal): keskmisest kõrgem vanus, väikese hiljutisuse, suure sageduse, suure rahalise väärtuse ning keskmisest kõrgema keskostu põhjal.

Reageerimist hinnati uudiskirja spetsiifilisuse alusel. Uudiskiri on lisatud käesolevale tööle lisas 1. Tegemist oli moele suunatud kirjaga, seega jälgiti oste moe-osakondades.

Gruppide A, B ja kontrollgrupi konversioonimäärad:

- Grupp A – valimi suurus 176 000 klienti, CR 3%
- Grupp B – valimi suurus 6 000 klienti, CR 8%
- Kontrollgrupp - valimi suurus 6 000 klienti, CR 6%

Konversioonimäärade võrdluse tulemusel on näha, et klasterdamise tulemusel kõige tõenäolisemalt reageerivateks klientideks määratletud kliendid tegid uudiskirja tulemusel üle 2,5 korra rohkem oste. Teisalt, kui vaadata kliente, kes on väärtuslikumad, võib nende reageerimine või mitte reageerimine olla seotud muude põhjustega, millel pole mingit seost uudiskirjaga. Nii tuleb testimise käigus selgelt välja, et kliendid, kes uudiskirja ei saanud on teinud konversioonimäära alusel poole rohkem oste kui grupis A demograafilistel näitajatel põhineva valimi kliendid. Grupi B klientidel on sellegipoolest kolmandiku võrra rohkem oste tehtud, mis viitab mudeli tulemuslikkusele.

Kokkuvõte

Käesoleva magistritöö eesmärgiks oli Kaubamaja klientide segmenteerimine andmekaevanduse tehnikate ning reeglite kombinatsioonide abil ning loodud segmentide analüüsimine uudiskirja konversioonimäära suurenemise põhjal.

Töö esimese pooles tutvuti varasemate uurimuste ja teoreetiliste kontseptsioonidega, mis haakuvad klientide segmenteerimisega ja põhinevad suurandmetel. Varasemate uurimuste põhjal toodi välja turunduses kasutatavad segmenteerimise võimalused. Selgus, et mitte kõik kliendid ei ole valmis ostu sooritama. Kui klient on alles brändi ja tooteid tundma õppimas, ei ole mõtet talle sooduspakkumistega uudiskirja postitada. Analüüsida tuleb kliendi elutsüklit ja vastavalt sellele ka oma kommunikatsioon kujundada. Põhjalikumalt uuriti RFM analüüsi ja klasteranalüüsi, mis on varasemalt andnud klientide segmenteerimisel efektiivseid tulemusi ja mis aitab määratleda, millises elutsükli etapis klient parajasti asub.

Töö teises pooles on kirjeldatud magistritöö metoodika ning antud ülevaade ettevõttest ja kasutatavatest andmetest. Lisaks leiti kriteeriumid ja atribuudid, mida töös kasutada. Leitud kriteeriumitele ja atribuutidele anti selgitused ja leiti atribuutide omavahelised seosed. Meetodi rakendamise käigus sai kirjeldatud peakomponentanalüüs ja leitud sobiv klastrite arv. Seejärel kirjeldati klasteranalüüsi tulemusi nelja ja viie klastri põhjal ja leiti iga klastri kliente iseloomustavad tunnused. Leiti, et kasutusele tuleks võtta viiel klastril põhinev mudel, mis kliendisegmente täpsemalt kirjeldab.

Töö viimases osas analüüsiti mudeli tulemuslikkust A/B testi põhjal. Selleks loodi erinevad valimid, kellele saadeti e-posti teel uudiskiri ja uuriti nende valimite konversioonimäärasid. E-posti turundust iseloomustab vähene tähelepanu määr. E-kaubanduses jääb efektiivsus enamasti valdkonnas 2-3% piiresse. Varasemastel tingimustel loodud valimi konversioonimäär just sellist tulemust näitas ning jäi 3% juurde. Kõige tõenäolisemalt reageerivateks klientideks määratletud kliendid tegid uudiskirja tulemusel üle 2,5 korra rohkem oste ning nende konversioonimäär oli 8%. Kõrge oli konversioonimäär ka kontrollgrupis. Leiti, et erinevatel meetoditel põhinevad valimid, on üsna erinevate konversioonimääradega.

Edasiarendamise võimalustena leiti, et lisaks kliente kirjeldavatele andmetele, tuleks mudelisse kaasata ka toodete ja eelkõige brändide ostuandmeid. Huvitavaid tulemusi võiks anda analüüs

selle kohta, milliseid brände kliendid koos ostavad või, milliste brändide pakkumisi klientidele saatma ei peaks.

Töö peamiseks eesmärgiks oli luua Kaubamajale kliendisegmendid ning selgitada välja tunnused, mis iseloomustavad Kaubamaja põhikliente. Töö tulemusena loodud viiel klastril põhinev mudel, annab kirjeldused segmentidele ning konversioonimäära kasv test-valimi osas lubab arvata, et nende segmentide kasutusele võtmine turunduses ja müügis aitab hoida ja suurendada klientide lojaalsust ning tõhusamalt läbi viia kliendikeskset turundust.

Kasutatud kirjandus

[1] Berry, M., J., A., Linoff, G., S. (2004). Data mining techniques: for marketing, sales and customer relationship management. Indianapolis: Wiley

[2] Chang, H., Tsai, H (2011) Group RFM analysis as a novel framework to discover better customer consuption behavior, Expert Systems with Applications, Elsevier https://www.sciencedirect.com/science/article/pii/S0957417411008189?casa_token=uFw7XUxXiqUAAAAA:Myou1JavgrvQVDaGgqitKSa1j4uHkpL4_F4GmZFKrilFy-oYkRS2bI7KHxqqS5peqyLfXRsUnQ

[3] Dogan, O., Aycin, E., Bulut, Z (2018) Customer Segmentation by Using RFM Model and Clustering Methods: A Case Study In Retail Industry, International Journal of Contemporary Economics and Administrative Sciences, Volume: 8, pp 1-19

[4] Marcus, C (1998) A practical yet meaningful approach to customer segmentation, Journal of Consumer Marketing, vol 15 no 5, MCB University Press, pp 494-504

[5] Zhang, G., Zhou, F., Wang, F., Luo, Jian (2008) Knowledge Creation in Marketing Based on Data Mining, International Conference on Intelligent Computation Technology and Automation (ICICTA)

[6] Chen, D., Sain, S.L., Guo, K (2012) Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing & Customer Strategy Management volume 19, pp 197–208

[7] Sorescu, A. (2017) Data-Driven Business Model Innovation. Journal of Product Innovation Management. 34, pp 691-696

[8] A. Griva, C. Bardaki, K. Pramatari ja D. Papakyriakopoulos, (2018) „Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data,“ Expert Systems with Applications, kd. 100

[9] Kotler, P., Jain, D., Maesincee, S. (2003) Muutuv turundus, Pegasus lk 103-112

[10] Wilson, C (2003) Tulusad kliendid: Kuidas neid ära tunda, arendada ja hoida. Eesti Ekspressi Kirjastuse AS lk 73-75

[11] Adomavicius, G. And Tuzhilin, A. (2005) Personalization technologies: A processoriented perspective. Communications of the ACM, October, 2005. Vol. 48. No. 10, 83-90

[12] Juhkam, M. (2004) Klasterdamine andmekaevanduses. Andmekaevandamise uurimisseminar, Tartu Ülikool

[13] Stone, B. (2017) Pood, kust saab kõike: Amazoni ja Jeff Bezose lugu. Rahva Raamat AS lk 209-225, 335-338

[14] Aden. A (2020) What is Conversion Rate? How to Calculate and Improve Your Conversion Rate, retrieved from: https://www.disruptiveadvertising.com/conversion-rate-optimization/conversion-rate/ 14.05.2020

[15] Amorim, R.C., Hennig, C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. (2015) Information Sciences, 324, pp. 126-145. doi:10.1016/j.ins.2015.06.039

[16] Al-Anazi, S., AlMahmoud, H., Al-Turaiki, I. Finding similar documents using different clustering techniques. (2016) Procedia Computer Science 82, pp. 28-34. doi:10.1016/j.procs.2016.04.00

[17] E. T. Bradlow, M. Gangwar, P. K. Kopalle ja S. Voleti, (2017) „The Role of Big Data and Predictive Analytics in Retailing," Journal of Retailing, kd. 93, nr 1, pp. 79-95

[18] Kassambara, A. (2017) Practical Guide To Cluster Analysis in R: UnsupervisedMachine Learning. STHDA. Retrieved from: https://www.datanovia.com/en/wp-content/uploads/dn-tutorials/book-preview/clustering_en_preview.pdf 25.04.2021

[19] C. G. Yee, M. F. A. Aziz and S. S. Hasan, Applying Instant Business Intelligence in Marketing Campaign Automation. (2010) Second International Conference on Computer Research and Development, 2010, pp. 643-646, doi: 10.1109/ICCRD.2010.180.

[20] Raschka, S (2019) Python Machine Learning. Chapter 5 – Compressing Data via Dimensionality Reduction. Retrieved from: https://github.com/rasbt/python-machine-learning-book/blob/master/code/ch05/ch05.ipynb