KAUNO TECHNOLOGIJOS UNIVERSITETAS
MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

Evaldas Ramoška

Pažangios analitikos priemonės apdorojant mažmeninės prekybos duomenis

Baigiamasis magistro projektas

Vadovai
doc. dr. Tomas Ruzgas
dr. Beata Šeinauskienė

KAUNAS, 2017

KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

Pažangios analitikos priemonės apdorojant mažmeninės prekybos duomenis

Baigiamasis magistro projektas

Didžiųjų verslo duomenų analitika (621G12002)

Vadovas
doc. dr. Tomas Ruzgas
doc. dr. Beata Šeinauskienė
Recenzentai
lekt. dr. Mantas Landauskas
doc. dr. Aušra Rūtelionė

Projektą atliko
Evaldas Ramoška

KAUNAS, 2017

KAUNO TECHNOLOGIJOS UNIVERSITETAS

Matematikos ir gamtos mokslų fakultetas
(Fakultetas)

Evaldas Ramoška
(Studento vardas, pavardė)

Didžiųjų verslo duomenų analitika, 621G12002
(Studijų programos pavadinimas, kodas)

„**Pažangios analitikos priemonės apdorojant mažmeninės prekybos duomenis**“

AKADEMINIO SĄŽININGUMO DEKLARACIJA

20 17 m. birželio 2 d.
Kaunas

Patvirtinu, kad mano, **Evaldo Ramoškos**, baigiamasis projektas tema „**Pažangios analitikos priemonės apdorojant mažmeninės prekybos duomenis**“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

(vardą ir pavardę įrašyti ranka)	*(parašas)*

Ramoška, Evaldas. Pažangios analitikos priemonės apdorojant mažmeninės prekybos duomenis. Magistro baigiamasis projektas / vadovai doc. dr. Tomas Ruzgas, doc. dr. Beata Šeinauskienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Mokslo kryptis ir sritis: Didžiųjų verslo duomenų analitika

Reikšminiai žodžiai: IBM, SAS, KNIME, mažmeninė prekyba, klasterizavimas.

Kaunas, 2017. 66 p.

SANTRAUKA

Vienas iš svarbiausių šių dienų įmonės uždavinių yra išsirinkti pažangią analitinę priemonę, kuri turėtų visus reikiamus įrankius, metodus ir algoritmus norimoms analizėms atlikti.

Tyrimo tikslas yra apžvelgti pažangiausias analitines priemones ir įvertinti jų funkcionalumą atliekant mažmeninės prekybos duomenų analizę. Vienas iš tyrimo uždavinių yra išrinkti ir apžvelgti duomenų mokslui skirtas pažangiausias analitines priemones. Atsižvelgiant į kompanijos Garner sudarytu reitingu buvo atrinktos trys analitinės platformos. Naudojantis tyrimui atrinktomis IBM, SAS ir KNIME analitinėmis platformomis darbe atlikta klasterinė ir laiko eilučių analizė, minėtos analizės yra orientuotos į mažmeninės prekybos duomenis. Klasterinė analizė buvo atlikta naudojantis k-vidurkių metodą, o kaip tyrimo objektas buvo naudotas RFM (Recency, Frequency, Monetary) modelis, tyrime su skirtingomis analitinėmis priemonėmis buvo surastos ir apžvelgtos tam tikros prasmingos klientų grupės. Kita analizės dalis buvo orientuota į laiko eilučių analizę, tiksliau pardavimo sumų prognozavimą į ateitį naudojant ARIMA modelį.

Atlikus analizę buvo įvertintas programinių priemonių funkcionalumas, t.y. greičio, metodų, algoritmų, įverčių ir statistikų įvairovė ir atitikimas šių dienų reikalavimams. Remiantis mažmeninės prekybos duomenų analize SAS buvo atrinkta, kaip pažangiausia analitinė priemonė.

Ramoška, Evaldas. Advanced Analytics Processing Retail Data: Master‘s thesis / supervisor assoc. prof. Tomas Ruzgas, assoc. prof. Beata Šeinauskienė. The Faculty of mathematics and nautral sciences, Kaunas University of Technology.

Research area and field: Business Big Data Analytics.

Key words: IBM, SAS, KNIME, retail, clustering.

Kaunas, 2017. 66 p.

SUMMARY

Nowadays, one of the main tasks that businesses have is to select an advanced analytical platform which would have all required tools, methods and algorithms to carry on wanted analysis.

The goal of the thesis is to overview the most advanced analytical tools and to evaluate their functionality while using retail sales data analysis. One of the thesis tasks is to choose and overview the tools that are being used in analytics. Based on Garner`s company rating there were three analytical platforms selected as being the best. Clustering analysis and time series where applied using selected analytical platforms IBM, SAS, KNIME. K – means method was used for clustering analysis and RFM (Recency, Frequency, Monetary) model was used with different analytical tools which helped finding meaningful client groups. The second part of analysis was focused on time series analysis, where sales sum forecasting to the future was done using ARIMA model.

The functionality of different analytical tools was evaluated based on analysis where speed, methods, algorithms, estimations and statistics variety of different tools was estimated. Based on analysis done on the retail sales data SAS was selected as the best analytical tool.

TURINYS

PAVEIKSLĖLIŲ SĄRAŠAS

LENTELIŲ SĄRAŠAS

TYRIMO TIKSLAS IR UŽDAVINIAI

Tyrimo tikslas

Apžvelgti pažangiausias analitines priemones ir įvertinti jų funkcionalumą atliekant mažmeninės prekybos duomenų analizę.

Uždaviniai

- Išanalizuoti šių dienų didžiųjų duomenų problemas.
- Apžvelgti ir išrinkti pažangiausias duomenų mokslui skirtas analitines priemones.
- Įvertinti klasterinės ir laiko eilučių analizės svarbą įmonių veikloje.
- Su pasirinktomis programinėmis priemonėmis atlikti klasterinę ir laiko eilučių analizę naudojant mažmeninės prekybos duomenis.
- Atsižvelgiant į rezultatus pateikti įžvalgas.
- Įvertinti pažangias analitines priemones ir jų funkcionalumą.

ĮVADAS

Šių laikų pasaulis pavirto į informacinę visuomenę, kuri vis labiau tampa priklausoma nuo duomenų. Kiekvieną dieną, kiekvieną sekundę informacinės sistemos generuoja milžiniškus duomenų kiekius. Visuomenė susiduria su duomenų pertekliaus problema. Akivaizdu, kad norint apdoroti milžiniškus ir vis didėjančius duomenų kiekius reikia milžiniškų duomenų talpyklų ir kompiuterio resursų. Tačiau programinių įrangų ir technologijų talpos didėjimas ribotas, o duomenų augimai beribiai.

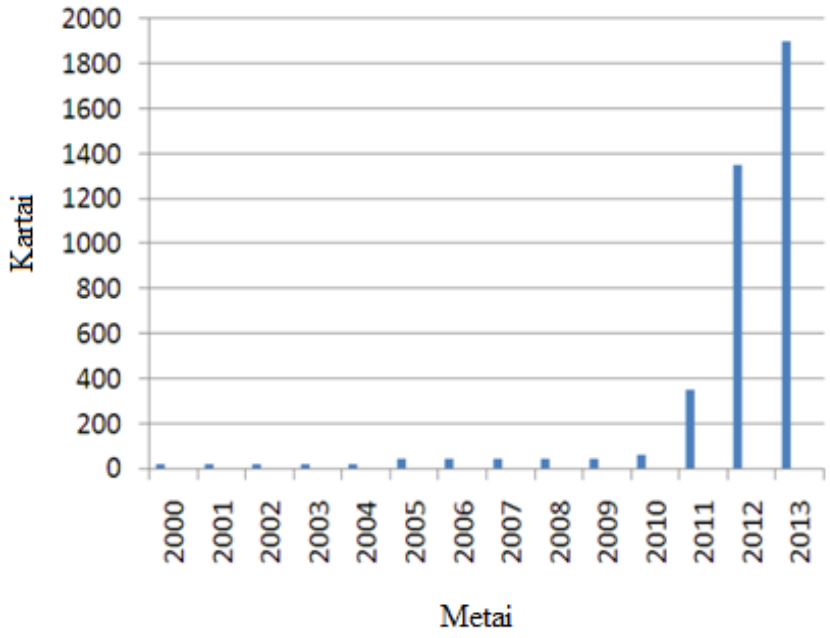
Organizacijos auga, generuojamų duomenų kiekiai auga dar labiau, taip įmonės susiduria su didelėmis problemomis apdorojant duomenis. Didžiausios organizacijos dirba su skirtingomis programinėmis platformomis taip gaunasi daug duomenų, kurie būna sutalpinti į skirtingų formatų failus. Kadangi šių laikų duomenys, būna skirtingų tipų, taip tampa labai sunku jiems priskirti tam tikras kategorijas su vienu algoritmu ar viena logika. Šiais laikais duomenys stipriai įtakoja įmonių procesus, informacija tampa įmonių pagrindu. Kuo duomenys tampa sudėtingesni, tuo svarbiau organizacijoms turėti įvairiapusiškesnes programines priemones teisingiems sprendimams priimti.

Viena iš pagrindinių užduočių įmonėms yra pasirinkti tinkamą analitinę priemonę, kuri leistų pažangiausius, novatoriškiausius, bet tuo pačiu paprasčiausius būdus reikiamoms analizėms atlikti ir jas atvaizduoti. Šiais laikais yra sukurta labai daug programinių priemonių, tačiau reta programinė priemonė yra pritaikytos šių dienų reikalavimams. Verslas reikalauja greičio, kokybės ir paprastumo, tačiau ne visos analitinės priemonės gali pasiūlyti greitas analizes, platų metodų bei algoritmų pasirinkimą ir aiškias bei pažangias vizualizacijos galimybes. Taip įmonės susiduria su programinės priemonės pasirinkimo iššūkiu.

LITERATŪROS APŽVALGA

1.1. Didžiųjų duomenų samprata

Nors šiandien terminas „didieji duomenys“ dažnai vartojamas, tačiau didžiųjų duomenų apibrėžimas yra vis dar besiformuojantis. Manoma, kad didžiųjų duomenų sąvoką ir idėją 1990 metais sugalvojo ir išplatino amerikietis Johnas Mashei. Nepaisant to, kad posakis buvo paminėtas gan seniai, tačiau kaip matome iš *1 pav*. plačiai jis buvo pradėtas naudoti tik maždaug po dvidešimties metų .



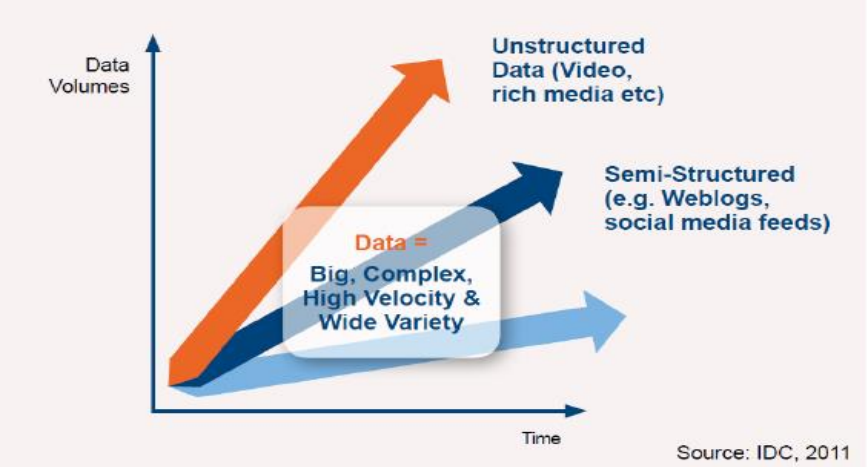
1 pav. ProQuest atradėjų bibliotekoje dokumentų skaičius, kuriuose buvo panaudotas terminas „Didieji duomenys“ (angl. „Big data“) pasiskirstymas [1]

Didžiųjų duomenų terminą daug kas suprasdavo ir interpretuodavo vis kitaip. 2001 metais Douglas Lanei pasiūlė duomenų valdymo iššūkius vadinti trimis dimensijomis tai dydžiu, įvairove ir greičiu (angl. Volume, Variety and Velocity), kitaip 3V.[2] . Vėliau 3V buvo pradėta vadinti bendru terminu, kuris apibūdina didžiuosius duomenis.[3],[4]. Nors visi didžiuosius duomenis apibūdino savaip, mokslininkai labiausiai vertina šiuos sąvokos „didieji duomenys“ apibrėžimus:

Didieji duomenys yra dideli kiekiai, didelis greitis ir didelė įvairovė. Informacinis turtas reikalauja efektyvių, naujoviškų apdorojimo formų siekiant patobulinti įžvalgas ir priimant tinkamus sprendimus. (Gartner.[5])

Panašiai TechAmerica įkūrėjas apibūdina didžiuosius duomenis :

Didieji duomenys yra sąvoka apibūdinanti didelių greičių didelius kiekius, sudėtingumą ir kintamus duomenis, kuriems reikalingi pažangūs metodai ir technologijos leidžiančios kaupti, valdyti, paskirstyti ir analizuoti informaciją.(TechAmerica Foundations Federal Big Data Commission, 2012. [6])



2 pav. Didžiųjų duomenų apibrėžimas

1.2. 3V didžiųjų duomenų problema

Duomenų mokslas buvo išsiskirstęs į tris dimensijas tai dydis, įvairumas ir greitis bei dar dvi vėliau prijungtas dimensijas – duomenų reikšmę ir tikrumą.

1.2.1. Dydis

Pirmoji dimensija – dydis, jis nusako duomenų kiekį, kuris yra analizuojamas ir apdorojamas siekiant gauti norimus rezultatus. Šio dydžio kitimą sunku nusakyti, bet jis kyla labai dideliu greičiu ir pagal prognozę duomenų kiekis kas metus kils po 40 procentų Tai tampa iššūkiu, nes norint valdyti ir analizuoti milžiniškus duomenų kiekius yra reikalingi milžiniški technologiniai resursai. Pavyzdžiui kompiuterių sistemos yra ribojamos technologijų dėl operacijos proceso greičio, o duomenų dydis, kuris turi būti apdorotas gali būti beribis, bet operacijos proceso greitis yra konstanta. Norint gauti didesnį proceso greitį reikia galingesnio kompiuterio, arba ieškoti kitokių išeičių, kaip pavyzdžiui duomenų atrinkimas ar suspaudimas.

1.2.2. Greitis

Antroji dimensija – greitis. Kartu su stipriai pagreitėjusiu duomenų kiekio augimu, ši dimensija atspindi padidėjusį duomenų atsiradimo, kitimo ir perdavimo greitį. Kai kurie duomenys tik atsiradę gali pasenti per kelias sekundes, o kai kurie gali būti dinamiškai apdoroti po sukūrimo, todėl išgaunant informaciją ar ją keičiant analizė turi būti atlikta nedelsiant. Tai reikalauja metodų, kuriais remiantis analizę būtų galima atlikti beveik realiu laiku.

1.2.3. Įvairovė

Trečioji didžiųjų duomenų dimensija – įvairovė. Ši dimensija reprezentuoja duomenų tipą, kuris yra kaupiamas, analizuojamas ir naudojamas. Duomenys, kurie būna kaupiami ir analizuojami būna įvairių tipų, kaip pavyzdžiui nuotraukos, vaizdo įrašai, simuliacijos, koordinatės ir t.t.. Čia prieiname prie iššūkio, kaip surūšiuoti, sutapatinti visus duomenis, kad jie taptų suprantami visiems, kurie jais naudosis.

1.2.4. 5V

Laikui bėgant atsirado ir daugiau idėjų kaip praplėsti 2001 metų Douglas Laney „3V‘s“ didžiųjų duomenų apibrėžimą. Buvo pasiūlyta ketvirtoji ir penktoji dimensija, tai reikšmė ir patikimumas.

1.2.5. Reikšmė

Ketvirtoji dimensija – reikšmė. Tai viskas apie duomenų, kurie kaupiami ir naudojami kokybę. Pagrindinė dimensijos idėja yra ta, kad nėra prasmės kaupti milžiniškų duomenų jeigu negalima jų tinkamai apdoroti, plėtoti bei panaudoti.

1.2.6. Patikimumas

Penktosios dimensijos esmė yra patikrinimas ar duomenys yra tinkami didžiųjų duomenų apibrėžimui, t.y. duomenų tikrumas ir nuoseklumas. Pavyzdžiui, jeigu A siunčia elektroninį laišką į B, B gaus laišką tokį, kokį išsiuntė A, priklausomai nuo to ar elektroninių laiškų bendrovė patikima ir niekas kitas tuo laišku nesinaudos. Jeigu didžiuosiuose duomenyse atsiranda duomenų nutekėjimas, tie duomenys gali tapti nebenaudingi ir nebepanaudojami.[7]

Apibendrinant didžiuosius duomenis, galima sakyti, kad jie masyvūs ir stipriai besiplečiantys, bet taip pat ir netvarkingi, chaotiški, staigiai besikeičiantys, daugybėje formatų ir nenaudingi be analizės ir vizualizacijos. Pasaulyje didieji duomenys ir analizė yra tarpusavyje susiję, vienas be kito jie yra praktiškai nereikalingi, bet kartu jų galimybės praktiškai beribės.

1.3. Pažangių analitinių platformų poreikis

Esamos analitinės platformos yra profesionalų tobulinamos tam, kad prisitaikytų prie vis didėjančių didžiųjų duomenų reikalavimų. Tobulinti esamas analitines priemones norint nuspėti ateities poreikius yra praktiškai neįmanoma, tačiau rinka verčia rasti naujų būdų kurti analitines platformas ir priemones atitinkančias šių dienų reikalavimus, o atsiradus naujoms analitinėms priemonėms atsiranda didelių pokyčių marketingo, pardavimų, procesų ir operacijų valdyme. Taip įmonės būna priverstos įdiegti pažangiausias analitines priemones į visus pagrindinius verslo procesus.

1.4. Pažangios analitinės priemonės

Šiais laikais egzistuoja labai daug analitinių priemonių, todėl renkantis pažangiausias, plačiausiai naudojamas ir geriausiai vertinamas duomenų mokslui pritaikytas pažangiąsias analitines platformas buvo remtasi Gartner, Inc.[8] sudarytas reitingas. Gartner, Inc tai Amerikos kompanija kuri verčiasi informacinių technologijų tyrimais ir konsultacijomis. Dažniausiai jų klientai būna didžiosios įmonės besinaudojančios informacinėmis technologijomis. Gartner Inc. šioje srityje yra lyderė pasaulyje.

Garner kompanija sukūrė „magiškąjį kvadrantą“ (*3 pav.*) norėdama suklasifikuoti duomenų mokslui skirtas programines priemones, taip įmonėms leidžiant atsirinkti pažangiausias, geriausiai pritaikytas pažangiai analizei priemones.

3 pav. Garner 2017 „magiškasis kvadrantas“ klasifikuojantis pažangiausias analitines priemones [8]

Magiškas kvadrantas analitines priemones išskiria į keturis tipus:

- Lyderiai – puikiai vykdančios ir vystančios savo esamą viziją, pasiruošusios rytojaus iššūkiams
- Turinčios viziją – suprantančios kaip vystosi rinka arba turi viziją kaip prie jos prisitaikyti, tačiau dar neatitinka šių dienų reikalavimų.

- Turinčios savo sritį – sėkmingai susikoncentravusios mažame segmente arba nesėkmingai besivystančios kituose segmentuose.

- Varžovai - dominuojantys dideliame segmente, tačiau nedemonstruoja supratimo apie šių dienų rinkos poreikių tendencijas.

2017 metų ataskaita parodo 16 duomenų mokslui skirtų analitinių priemonių, jos buvo rūšiuojamos pagal 15 kriterijų ir sudėliotos į ankščiau paminėtus keturis kvadrantus. Kompanija skirstė pažangias analitikos platformas, kurios gali pasiūlyti naujausius visų rūšių duomenų analizės sprendimus. Taip pat atsižvelgta kaip tuos sprendimus galima integruoti į verslo procesus.

Lyderiai : IBM, SAS, RapidMiner, Knime.

Turinčios viziją : Microsoft, H2O.aim Datalku, Domino Data Lab, Alpine Data.

Turinčios savo sritį : SAP, FICO, Teradata.

Varžovai : MathWorks, Quest, Alteryx, Angoss.

Žemiau apžvelgsiu jau trečius metus lyderiaujančias pažangias analitines platformas IBM, SAS ir KNIME

1.4.1. IBM

Atsižvelgiant į Garner magiškojo kvadranto ataskaitą IBM šiais metais pakilo į pirmąją vietą tarp pažangiausių analitinių platformų. IBM programinė platforma daugiausiai dėmesio skiria analizių vystyme ir turi labai stiprias programines priemones, kai kalbame apie analizavimą.

IBM duomenų mokslui labiausiai plėtoja šias programines priemones:

- IBM Watson

- IBM SPSS

Kompanija vysto „Watson“ programinę priemonę, kuri yra pažangi duomenų analizavimo priemonė prieinama ir taip vadinamame „debesyje“. Tai sistema, kuri turi pažangias duomenų analizavimo, prognozių analitikos, šviesleničių kūrimo ir vizualizacijos galimybes. Watson programinė priemonė turi ir daugiau pliusų. Vienas iš jų, kad ribotą versiją gali visi naudotis nemokamai. Ši versija leidžia importuoti duomenis, gauti vizualizacijas, kurti švieslentes. Taip pat ten yra keli pavyzdinių duomenų masyvai. Profesionali versija kainuoja 80 dolerių per mėnesį, tačiau ji suteikia daugiau galimybių ir leidžia programiniu paketu naudotis daugiau nei vienam vartotojui. Taip pat profesionalioji versija leidžia naudotis reliacinėmis duomenų bazėmis, suteikia prieigą prie IBM Cognos ataskaitų, pilną prieiga prie IBM Analytics Exchange duomenų.

IBM SPSS plėtoja dvi pažangias analitines priemones tai SPSS Modeler ir SPSS Statistics

SPSS Modeler yra analitinė platforma skirta kasdieninėms verslo problemoms spręsti. Priemonė turi daugybę pažangios analizės galimybių. SPSS Modeler plėtoja teksto, objektų, socialinių tinklų analizes taip pat turi automatizuoto modeliavimo, duomenų paruošimo, sprendimų valdymo ir optimizavimo galimybes. Programinė priemonė suteikia galimybę naudotis duomenimis, kad ir kur jie yra saugomi, nes duomenis galima išgauti iš duomenų saugyklų, duomenų bazių, Hadoop duomenų bazės. Taip pat priemonė apdoroja duomenis nesvarbu ar jie struktūrizuoti ar ne, kaip pavyzdžiui, tekstinius failus, elektroninius laiškus ar socialinių tinklų duomenis. SPSS Modeler tinka tiek paprastam verslininkui tiek duomenų tyrybos specialistui ar duomenų mokslininkui. Programinę priemonę trisdešimt dienų nemokamai galima išbandyti, tačiau pilna versija kainuoja 4 670$ metams.
SPSS Statistics yra pirmaujanti statistinė programinė priemonė naudojama sprendžiant mokslines ir verslo problemas. Galingas įrankis suteikia galimybę atlikti daug analizės rūšių, įskaitant situacijų analizę, hipotezių tikrinimus ir ataskaitų gamybą, taip leidžiant lengviau apdoroti, valdyti, atvaizduoti, analizuoti duomenis. IBM statistics suteikia nemokamą 15 dienų bandomąją versiją. Pilna, neribota versija kainuoja 99$ per mėnesį.[9]

1.4.2. SAS

Iš pirmosios vietos 2016 metais į antrąją 2017 metais nukritusi programinė platforma „SAS“ yra viena žinomiausių ir geriausią reputaciją turinčių pastarojo dešimtmečio programinių platformų, išleidusi daug geros kokybės pažangių analitinių priemonių.

SAS programinė platforma yra verslo analitikos lyderė rinkoje, taip pat SAS yra didžiausia nepriklausoma verslo įžvalgų prekiautoja rinkoje.

Programinė platforma siūlo duomenų tyrybai, statistinei analizei, prognozavimui, teksto analizei, optimizavimui ir simuliacijoms pritaikytas programines priemones.

Duomenų gavybos ir analizavimo kategorijai priskiriamos programinės priemonės yra SAS Enterprice Miner, Factory Miner, Scoring Accelerator ir Visual Analytics.

Statistinės analizės sričiai SAS siūlo Analytics Pro, ETS, In-Memory Statistics, Visual Data Discovery ir kitas analitines priemones.

SAS duomenų moksle:

- Lengva išmokti. SAS programinės kalbos ne tik lengva išmokti, bet SAS siūlo galimybę viduje platformos naudotis programine kalba SQL. Besimokantys

SAS programinės kalbos gali mokytis iš mokomųjų įrašų, kurie yra SAS internetinėje svetainėje ir daugelio universitetų internetiniuose puslapiuose.

- Puikus duomenų apdorojimas. SAS turi puikias duomenų apdorojimo galimybes. Programinė įranga gali apdoroti viską kas yra saugoma kompiuterio atmintyje ir tuo pačiu metu atlikti kitus skaičiavimus.

- Grafinės galimybes. SAS yra įdiegtos pažangios funkcinės grafikos galimybės. Su tam tikromis aplikacijomis, dizaineriai gali vystyti funkcionalumą. Suprasti grafinį SAS paketą nėra labai sudėtinga pasitelkus pagalbinius vaizdo įrašus, kurių yra daug ir juos surasti gana nesunku.

- Pažangūs SAS įrankiai. SAS yra įdiegti įrankiai ir funkcijos duomenų apdorojimui atsižvelgiant į naujausias technologijas. Programinės įrangos funkcionalumas visada atnaujinamas tik atsiradus naujovėms. Naujovės platformoje atsiranda gana greitai, nes paketai naudojami daugelyje mokslo akademijų.

- Globalus lyderis darbo rinkoje. Dėl plataus naudojimo, SAS vis dar yra lyderis darbo rinkoje. Dauguma komercinių įmonių dirba naudodamiesi SAS platformą. Tai paaiškinama tuo, kad įmonėms yra priimtinas platus SAS analitinių priemonių pasirinkimas kaip pavyzdžiui duomenų vizualizacija, kokybė, saugyklos ir ataskaitos.

SAS trūkumai:

- Brangumas. SAS programinis paketas vienam naudotojui kainuoja apie 9000 $ metams. Norint paketą gauti daugiau negu vienam vartotoju licencija gali kainuoti apie 100 000 $ metams.

- SAS programavimo kalba. Kitos programavimo kalbos turi panašumų, todėl jas mokytis yra lengviau. SAS turi savitą programavimo kalbą, ji neturi analogų.[10]

1.4.3. KNIME

Paskutinioji iš lyderių kvadranto tai nemokama analitinė platforma KNIME skirta greitam, nesunkiam ir pažangiam duomenų mokslui.

KNIME – Konstant Information Miner platforma, kuri buvo sukurta Konstanz universitete ir nuo įkūrimo buvo vystoma į daugiafunkcinę duomenų mokslui skirta platformą. Yra kelios KNIME versijos, kiekviena iš jų turi savas galimybes. Platformoje galima atlikti vienmatę ir daugiamatę statistinę analizę, duomenų gavybą, laiko eilučių analizę, vaizdų apdorojimą, tinklo analizę, teksto analizę ir socialinių tinklų analizę. Komerciniai plėtiniai, taip pat kaip ir atviri

plėtiniai, gali būti arba įsigyti arba parsisiųsti. KNIME suteikia aplikacijų programavimo sąsajos (API) galimybę ir yra pagrįstas Eclipse platforma, kurios praplečia priemonės funkcionalumą. Dar vienas didelis KNIME pliusas yra tai, kad platformą galima lengvai integruoti į populiarias programavimo kalbas, tokias kaip Perl, Python ar R. Tačiau KNIME turi ir trūkumų. Naujiems platformos vartotojams yra gan sunku priprasti prie šios, gan netradicinės, programinės priemonės, dažniausiai tai atima daug laiko. Dar vienas KNIME trūkumas yra informacijos apie programinę priemonę stoka. Didžiųjų duomenų apdorojimas nemokamoje versijoje nėra įtrauktas, tačiau galima nusipirkti KNIME didžiųjų duomenų plėtinį, kuris skirtas būtent didiesiems duomenims apdoroti ir analizuoti.[11]

1.5. Mažmeninės prekybos duomenų tyrimas

1.5.1. Klientų segmentavimas

Nuo klientų segmentavimo idėjos pradžios ankstyvais 1950, klientų segmentavimas tapo viena dažniausiai tyrinėjamų tematikų marketingu pagrįstoje literatūroje. Visai neseniai dauguma šios tematikos literatūros šaltinių pradėjo vystytis į technologijų ir modelių atradimui skirtą literatūrą.[12]

Beveik kiekviename marketingu pagrįstame vadovėlyje yra rašoma, kad pagrindinis sėkmingas marketingas yra pagrįstas segmentavimo strategija. Pardavėjas retai gali patenkinti kiekvieną pirkėją. Ne kiekvienas klientas mėgsta tuos pačius gėrimus, automobilius, universitetus ar filmus. Tuo tikslu pardavėjai pradėjo vystyti rinkos segmentavimą. [13]

Segmentavimas yra klasikinė marketingo strategija. Pagrindinė segmentavimo idėja yra suskirstyti klientus į tam tikras grupes, tuomet vystyti marketingą taip, kad jis būtų skirtas labiausiai patrauklioms grupėms. Patrauklumas šiuo atveju yra pelningumas ir lojalumas. Segmentavimo procesas keičiasi nuo metodo pasirinkimo, tačiau labai svarbu pabrėžti, kad nesvarbu, koks metodas bus naudojamas, galutinis modelis beveik niekada nebus automatizuotas. Daugelį sprendimų, kaip pavyzdžiui, kuriuos ir kiek segmentų rasti, kokie jų santykiniai dydžiai, nusprendžia verslo vadybininkai, nes segmentavimo specialistai retai supranta šią sritį.[14,15]

Segmentavimo rezultatai stipriai priklauso nuo nagrinėjamų kintamųjų, kurie gali būti demografiniai, geografiniai, gyvenimo būdo ir t.t.. Vis dėl to, kai nagrinėjam vartotojų elgsenos duomenys, norint sužinoti ką klientas pirks, kokius produktus jis mėgsta, jo išlaidas, pirkimo dažnį ir ar klientas reaguoja į siūlomas akcijas tuomet reikalingi labai ištobulinti segmentavimo metodai.[14,16]

Per pastaruosius metus pastebimas labai didelis internetinės mažmeninės prekybos augimas. Šis didelis virtualios prekybos augimas parodo, kad žmonių apsipirkinėjimo ir paslaugų pirkimo įpročiai stipriai pasikeitė.

Lyginant apsipirkimą įprastoje mažmeninės prekybos parduotuvėje nuo virtualaus apsipirkimo, virtualus apsipirkimas turi vieną unikalią savybę. Kiekvieno kliento apsipirkimo procesas ir pats apsipirkimas gali būti stebimas, kiekvieno pirkėjo užsakymas dažniausiai būna susietas su pirkėjo adresu, taip pat kiekvienas pirkėjas turi virtualios parduotuvės paskyrą su kontaktais ir apmokėjimo informacija. Šios vertingos virtualios parduotuvės saugomos charakteristikos leidžia virtualios mažmeninės prekybos parduotuvės pardavėjams suprasti individualių pirkėjų elgesį ir taip atlikti į vartotojus orientuotą analizę. [12,13]

Atsižvelgiant į vartotojų elgsenos analizę, virtualios parduotuvės pardavėjai sprendžia šiuos klausimus:

- Kurių produktų puslapiuose klientas lankėsi? Kiek ilgai klientas apžiūrinėjo tam tikrus produktus? Kuriuose prekių grupių internetiniuose puslapiuose klientas lankėsi?
- Kurie klientai pelningiausi internetinei parduotuvei?
- Kurie klientai labiausiai lojalūs?
- Koks yra klientų pirkimo elgesio pobūdis? Kurias prekes klientai dažniausiai pirkdavo kartu? Kokia seka tos kartu perkamos prekės buvo įsigytos?
- Kurie klientai dažniausiai atsako į elektroniniu laišku siunčiamus pasiūlymus?

Norint atsakyti į visus šiuos internetiniam verslui rūpimus klausimus, duomenų analizės sektorius pritaikė daugybę analizės modelių skirtų virtualios mažmeninės prekybos sektoriui. Vienos žinomiausių verslo analizių norint sužinoti pirkėjų pelningumą ir vertę yra kaip neseniai pirkėjas pirko, kaip dažnai pirkėjas pirko ir kiek pirkėjas išleido pirkdamas prekes (RMF) analizė ir customer lifetime value model – vartotojo gyvavimo ciklo vertės modelis (CLV). Daugeliui didžiausių virtualios mažmeninės prekybos pardavėjų duomenų analizė tapo kasdieninė veikla ir verslo proceso dalis bandant atlikti į klientus orientuotą verslo analizę. [14,15,16]

1.5.2. RFM modelis

RMF – apibrėžiama kaip:

Recency – kaip neseniai pirkėjas pirko;

Frequency – kaip dažnai pirkėjas pirko;

Monetary – kiek pirkėjas išleido pirkdamas prekes.

RFM modelio analizė yra marketingo technika naudojama pirkėjų elgsenai įvertinti. Šis metodas naudingas gerinant klientų segmentavimą, skirstant juos į tam tikras grupes pagal vartotojo pirkimo įpročius. [17]

1.5.3. Pardavimų prognozavimas

Pardavimų prognozavimas yra praeities stebėjimų ir rinkos sąlygų pagrįstas ateities spėjimas. Ateities prognozavimo atlikimas leidžia objektyviai pažvelgti į ateitį.[18]

Pardavimų prognozavimas yra tam tikros įmonės pardavimų kiekio ar kitų įmonės rodiklių ateities prognozė, pagrįsta rinkos ir praeities pardavimų informacija.[19]

Iš apibrėžimų galima teigti, kad pardavimų prognozavimas yra tam tikrų įmonės stebėjimų prognozė į ateitį atsižvelgiant į praeities ir rinkos rodiklius.

Pardavimų prognozavimas yra svarbi, bet kurios įmonės verslo dalis. Pardavimų prognozės dažniausiai naudojamos pardavimų apimčių ir pajamų sumai nustatyti taip pat ir laiko bei resursų planavimui. Ateities prognozavimo tikslumas yra įmonės ateities efektyvumo užtikrinimas. [20,22]

Įmonės, kurios įgyvendina tikslią pardavimų prognozę, gauna tokią naudą kaip:

- Gebėjimą žinoti ateities pinigines įplaukas;
- Žinojimą kada ir kiek pirkti;
- Galimybę planuoti produkcijos mastus.;
- Gebėjimą identifikuoti tam tikras pardavimų tendencijas;
- Gebėjimą apskaičiuoti investicijų grąžą.

Šių naudų visuma gali atnešti įmonei tokius rezultatus kaip:

- Didesnis pelnas;
- Išaugęs klientų lojalumas;
- Sumažėję kaštai;
- Padidėjęs įmonės efektyvumas.

Pardavimų prognozėje ypatingai svarbus yra prognozės tikslumas, nes netikslios prognozės įmonei gali atnešti nuostolį. Nustatant per mažus pardavimus, įmonė gali neįvykdyti visų įsipareigojimų klientams, taip prarasdama dalį rinkos. Esant per daug optimistiškai prognozei įmonė gali pernelyg daug nepagrįstai investuoti ne pagal savo turimus išteklius, kas gali atvesti iki bankroto. Tikslių prognozių atveju įmonė gali išvengti nenumatytų grynųjų pinigų trūkumo, taip pat efektyviau valdyti turimą ir įsigyjamą produkciją, personalą ir investicijas.[21]

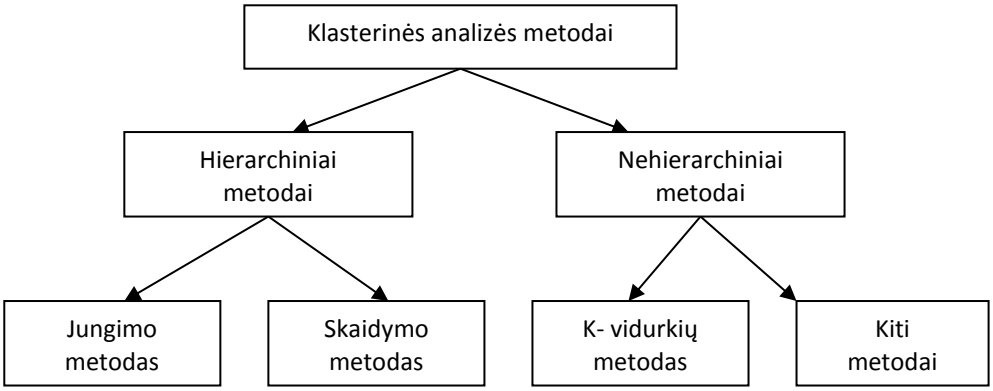
Apibendrinant galima teigti, kad vieni svarbiausių įmonės uždavinių yra klientų segmentavimas ir pardavimų prognozavimas. Klientų segmentavimas į prasmingas klientų grupes padeda įmonėms suprasti klientų elgseną ir padidinti įmonės efektyvumą. Pardavimų prognozavimas įmonei gali padėti ateities tikslams pasiekti.

2. MEDŽIAGOS IR TYRIMŲ METODAI.

2.1. Klasterinė analizė

Klasterinė analizė – statistinės analizės metodas, nusakantis tyrimo objektų panašumą ir suskirstantis juos į panašių objektų grupes, vadinamas klasteriais, taip, kad tos pačios grupės elementai butų kuo arčiau vienas kito, o elementai iš skirtingų grupių kuo toliau.

Skiriamos dvi metodų grupės: hierarchiniai ir nehierarchiniai metodai. Hierarchiniai metodai remiasi prielaida, kad visi duomenys laikomi vienu dideliu klasteriu, kurį sudaro mažesni klasteriai, įtraukiantys dar mažesnius ir t.t. Klasteriai gali būti randami skaidymo metodu, vienintelį klasterį skaidant į mažesnius, arba jungimo metodu, mažus klasterius jungiant į didesnius. Šiame darbe plačiau nagrinėjamas ir pristatomas tik Jungimo metodas. Nehierarchiniai metodai naudojami, kai iš anksto žinomas klasterių skaičius, bei kai tiriamų duomenų yra daugiau kaip 250. Objektai perskirstomi tol, kol panašumui klasteriuose tampa didžiausi lyginant su panašumas tarp klasterių.[23]



4 pav. Klasterizavimo analizės metodai [23]

2.1.1. K- vidurkių metodas

K – vidurkių metodo idėja yra nustatyti k centrus kiekvienam klasteriui. Tie centrai turėtų būti nustatyti teisingai, nes skirtingos centrų vietos įtakos skirtingus rezultatus, todėl centrai turėtų būti parinkti kuo toliau vienas nuo kito. Kitas žingsnis atlikti naują skaidymą į klasterius priskiriant artimiausiems centrams. Vėliau perskaičiuojami centrai ir žingsniai kartojami tol kol centrai nebesikeičia.

K – vidurkių algoritmas.

Tegul $X = \{x_1, x_2, x_3, \dots, x_n\}$ yra duomenų rinkinys, o $V = \{v_1, v_2, v_3, \dots, v_n\}$ centrų rinkinys.

1. Atsitiktinai parenkame c klasterių centrus
2. Apskaičiuojame atstumą tarp kiekvieno taško ir klasterio centro.
3. Priskirti duomenų taškus klasterių centrams, kad taškų atstumas nuo klasterio centro būtų mažiausias lyginant su likusiais klasterių centrais.
4. Perskaičiuoti naujus klasterių centrus naudojantis:

$$v_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c_i} x_i, \qquad (1)$$

kur c_i reprezentuoja skaičių taškų i – tajame klasteryje.

5. Perskaičiuoti atstumus tarp kiekvieno duomenų taško ir naujai suformuotų klasterių.
6. Jeigu nei vieno duomenų taško vieta nesikeičia, skaičiavimus stabdyti, kitu atveju kartoti veiksmus nuo 3. žingsnio. [24,25,26]

2.1.2. Kubinis klasterizavimo kriterijus

Kubinis klasterizavimo kriterijus naudojamas klasterių skaičiui nustatyti.

n = stebinių skaičius

n_k = k – tajame klasteryje esančių įrašų skaičius.

p = įrašų skaičius

q = klasterių skaičius

X = n ir p duomenų matrica

$\bar{X}$ = q ir p klasterių vidurkių matrica

Z = klasterių identifikavimo matrica su elementu $z_{ik} = 1$, jeigu i – tasis stebinys priklauso k – tajam klasteriui, kitu atveju $z_{ik} = 0$.

Tariame, kad be neapibrėžtumo praradimo kintamojo vidurkis lygus 0. Pažymime, kad $Z'Z$ yra įstrižainių matrica talpinanti n_k, gauname:

$$\bar{X} = (\mathrm{Z}'\mathrm{Z})^{-1}Z'X. \quad (2)$$

Viso pavyzdžio kvadratų suma ir vektorinės sandaugos matrica yra:

$$T = \bar{X}'Z'Z\bar{X}. \quad (3)$$

Vektorinė sandauga tarp klasterių užrašoma:

$$B =' Z'Z. \quad (4)$$

Vektorinė sandauga klasterių viduje:

$$W = (X - Z\bar{X})'(\mathrm{X} - \mathrm{Z}\bar{X}) = \mathrm{X}'\mathrm{X} - \bar{X}'\mathrm{Z}'\mathrm{Z}\bar{X} = \mathrm{T} - \mathrm{B}. \quad (5)$$

Keičiant sumų tvarką, galima parodyti, kad ženklas W yra lygus Euklido atstumui tarp stebinių ir klasterių vidurkių kvadratų sumai.

Kadangi T yra duotojo pavyzdžio konstanta, minimizuotas ženklas W yra ekvivalentus maksimizuotam:

$$R^2 = 1 - \frac{\check{\mathrm{Z}}enklas(W)}{\check{\mathrm{Z}}enklas(T)}. \quad (6)$$

2.1.3. Silueto koeficientas

Silueto koeficientas (angl. The silhouette Coefficient) yra populiarus metodas norint apibrėžti sąryšius ir išsiskaidymus. Silueto koeficientą galima apibrėžti trimis žingsniais:

1. Apskaičiuojame vidutinį i – tojo objekto atstumą nuo kitų objektų klasteryje ir pavadiname jį a_i.

2. i – tajam objektui ir bet kuriam klasteriui neturinčiam sąryšio su objektu apskaičiuojame vidutinį atstumą nuo kitų objektų duotajame klasteryje. Randame minimalią reikšmę atsižvelgiant į visus klasterius, tą reikšmę vadiname b_i.

3. i – tajam objektui silueto koeficientas yra:

$$s_i = (b_i - a_i)/\max(a_i, b_i). \quad (7)$$

Silueto koeficiento reikšmė gali būti intervale nuo 0 iki 1. Neigiama reikšmė yra netinkama, nes tada susiduriame su atveju kai a_i - vidutinis atstumas tarp klasterių yra didesnis nei b_i , minimalus vidutinis atstumas taškų kitame klasteryje.[27]

2.2. Laiko eilučių analizė

Laiko eilučių analizė – tai tyrimo metodas, kai naudojami nagrinėjamą reiškinį apibūdinantys laikotarpių duomenys. Analizuojant laiko eilutes galima išsiaiškinti pagrindines tendencijas ir kitus procesus, būdingus duomenų laiko eilutei. Kitaip tariant:

Analizuojamo atsitiktinio dydžio ξ_t stebėjimų, gautų laiko momentais $t = 1, \dots, T$ eilutė $Z_1, Z_2, \dots, Z_T$, vadinama laiko eilute.

2.2.1. ARIMA

Autoregresinis integruotas slenkančio vidurkio metodas (AutoRegressive Integrated Moving Average) – ARIMA yra plačiai naudojamas laiko eilučių analizei. Jo esmė – sujungti autoregresijos, diferencijavimo ir slenkančiųjų vidurkių metodo galimybes. Visos trys sudėtinės dalys yra paremtos atsitiktinio reikšmių išsibarstymo („triukšmo“), iškreipiančio laiko eilutės sisteminę komponentę, samprata ir turi savo būdingą reakcijos į šį atsitiktinį išsibarstymą aprašymo būdą. Bendriausias ARIMA modelis apima visas tris paminėtas dalis ir yra užrašomas taip:

$$ARIMA(p, d, q)$$

ARIMA metodas išskiriamas į nesezoninius $ARIMA(p, d, q)$ ir sezoninius $ARIMA(p, d, q)(P, D, Q)_s$ modelius.

Nesezoninis ARIMA modelis klasifikuojamas kaip ARIMA(p,d,q), kur parametrai p,d,q:

p - autoregresijos eilė (prieš tai buvusios reikšmės (atliekų kiekio) eilė ((t – 1), (t – 2) ir t.t.), kuri yra laiko eilutės reikšmės t momentu funkcija)

d - diferencijavimo eilė (pritaikytų diferencijavimo procedūrų skaičius, reikalingo proceso suvedimui į stacionarųjį pavidalą)

q – slenkančiųjų vidurkių narių skaičius (prieš tai buvusios „triukšmo“ reikšmės eilė ((t – 1), (t – 2) ir t.t.), kuri yra laiko eilutės „triukšmo“ reikšmės t momentu vidurkis)

Bendroji nesezoninio ARIMA(p,d,q) modelio išraiška:

$$\phi_p(B)(1 - B)^d y_t = \mu + \theta_d(B)e_t \qquad (8)$$

y_t – kintamojo reikšmė laiko momentu

μ – laiko eilutės vidurkis

B – poslinkio atgal per vieną laiko vienetą operatorius.

Atliekant prognozavimą į ateitį stebimos reikšmės keičiamos įverčiais $\hat{y}_t$, gaunasi :

$$B\hat{y}_t = \hat{y}_{t-1} \tag{9}$$

Pakeitus įverčiais galima sudaryti nesezoninio modelio prognozės lygtį:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} - \theta_1 e_{t-1} \ldots - \theta_q e_{t-q}, \tag{10}$$

kur μ – laiko eilutės vidurkis, $\phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}$ – stacionariųjų eilučių poslinkių sąlygos (y poslinkių reikšmės),o $-\theta_1 e_{t-1} \ldots - -\theta_q e_{t-q}$ judančio vidurkio sąlygos (poslinkių klaidos). [28,29]

Skirtingoms parametrų reikšmėms egzistuoja skirtingi ARIMA modeliai, žemiau aprašyti dažniausiai pasitaikantys ARIMA modeliai prognozavimui:

ARIMA(1,0,0) su konstanta – pirmos eilės autoregresinis modelis. Naudojamas jeigu laiko eilutė stacionari ir autokorealiuota. Prognozės lygtis šiuo atveju yra :

$$\hat{Y}_t = \mu + \phi_1 Y_{t-1}, \tag{11}$$

kur Y autokorealiacijos koeficientas, ϕ - gradientinis nusileidimas. Jeigu Y vidurkis yra nulis, tada konstanta į modelį nebeįtraukiama.

ARIMA(2,0,0) – antrosios eilės autoregresinis modelis. Šio modelio lygtis prognozei:

$$\hat{Y}_t = \mu + \phi_1 Y_{t-2} \tag{12}$$

ARIMA(0,1,0) – „atsitiktinis vaikščiojimas“. Šis modelis naudojamas jeigu eilutė nestacionari. Prognozės lygtis šiam modeliui gali būti aprašyta taip :

$$\hat{Y}_t - Y_{t-1} = \mu \tag{13}$$

arba

$$\hat{Y}_t = \mu + Y_{t-1} \tag{14}$$

ARIMA(0,1,0) su konstanta – „atsitiktinio vaikščiojimo“ su poslinkiu modelis. Šio metodo esmė yra ta, kad kiekviename laiko taške laiko eilutė mažai pakeičia kryptį nuo paskutinio taško

su žingsniais, kurių vidurkis yra nulis, jeigu žingsnių vidurkis yra kažkokia tai nenulinė reikšmė α. Šio modelio prognozės lygtis yra :

$$\hat{Y}_t = \alpha + Y_{t-1} \tag{15}$$

ARIMA(1,1,0) – Skirtumo pirmosios eilės autoregresinis modelis. Šis modelis naudojamas jeigu „atsitiktinio vaikščiojimo“ paklaidos yra autokorealiuotos. Modelio prognozės lygtis yra:

$$\hat{Y}_t - Y_{t-1} = \mu + \phi_1(Y_{t-1} + Y_{t-2}) \tag{16}$$

$$\hat{Y}_t - Y_{t-1} = \mu\ , \tag{17}$$

kuri gali būti perrašyta taip

$$\hat{Y}_t = \mu + Y_{t-1} + \phi_1(Y_{t-1} + Y_{t-2}). \tag{18}$$

ARIMA(0,1,1) be konstantos – paprastasis eksponentinio glotninimo modelis. Tai kita strategija taisant atsitiktinio „vaikščiojimo sudarytas“ sudarytas autokorealiuotas paklaidas. Prognozės lygtis gali būti aprašyta taip:

$$\hat{Y}_t = \hat{Y}_{t-1} + \alpha e_{t-1}, \tag{19}$$

pagal apibrėžimą :

$$e_{t-1} = Y_{t-1} - \hat{Y}_{t-1}, \tag{20}$$

todėl lygtį galima perrašyti taip:

$$\hat{Y}_t = Y_{t-1} - (1-\alpha)e_{t-1} =$$

$$= Y_{t-1} - \theta_1 e_{t-1}, \tag{21}$$

ARIMA(0,1,1) su konstanta – paprastasis eksponentinio glodinimo modelis su augimu. Papildžius ARIMA(0,1,1) be konstantos modelį konstanta, modelis įgauna lankstumo. Pirma MA(1) koeficientas gali būti neigiamas taip gaunasi, kad glodumo indeksas gali būti didesnis už 1, kas paprastame eksponentiniame glodumo modelyje nėra galima. Antra, Šio modelio prognozės lygtis yra tokia:

$$\hat{Y}_t = \mu + Y_{t-1} - \theta_1 e_{t-1} \tag{22}$$

ARIMA(0,2,1) arba ARIMA(0,2,2) be konstantos. Tiesinio eksponentinio glodinimo modeliai. Šie modeliai naudoja du nesezoninius skirtumus kartu su kintančio vidurkio sąlygomis. ARIMA(0,2,2) modelis be konstantos prognozuoja, kad antrasis laiko eilutės skirtumas lygus dviejų paskutinių tiesinės funkcijos prognozės klaidoms :

$$\hat{Y}_t - 2Y_{t-1} + Y_{t-2} = \theta_1 e_{t-1} - \theta_2 e_{t-2} \quad (23)$$

lygtis gali būti perrašyta taip:

$$\hat{Y}_t = 2Y_{t-1} - Y_{t-2} - \theta_1 e_{t-1} - \theta_2 e_{t-2}, \quad (24)$$

kur θ_1 ir θ_2 yra slenkančių vidurkių skalės koeficientai. Šis tiesinis eksponentinis glodnumo modelis yra toks pats kaip Holto modelis ir kai kuriais atvejais toks pats kaip ir Browno modelis. [30]

ARIMA(1,1,2) be konstantos – amortizuojančios tendencijos tiesinio eksponentinio glodinimo modelis. Šio modelio prognozės lygtis:

$$\hat{Y}_t = Y_{t-1} + \phi_1(Y_{t-1} - Y_{t-2}) - \theta_1 e_{t-1} - \theta_2 e_{t-2} \quad (25)$$

Jei nagrinėjama laiko eilutė kinta sezoniškai, modelis užrašomas ARIMA(p,d,q)(P,D,Q), kur antra parametrų grupė aprašo sezoninę modelio komponentę. Sezoninis ARIMA modelis klasifikuojamas kaip ARIMA(p,d,q)x(P,D,Q)S, kur parametrai P,D,Q,S :

P – sezoninių autoregresinių sąlygų skaičius

D – sezoninių skirtumų skaičius

Q – sezoninių judančio vidurkio sąlygų skaičius,

S – pasikartojančio sezoniškumo skaičius.

Bendra sezoninio ARIMA modelio išraiška yra tokia:

$$\phi(B^s)\varphi(B)(1-B)^d(1-B^s)^D(y_t - \mu) = \theta(B)\Theta(B^s)e^t, \quad (26)$$

kur $\phi(B)= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ ir $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ – Box – Jenkins modelio daugianariai;

$\varphi(B^s) = 1 - \varphi_1 B^s - \varphi_2 B^{2s} - \cdots - \varphi_p B^{Ps}$ – sezoninis autoregresijos (AR) komponentas;

$\Theta(B^s) = 1 - \Theta B^s - \Theta_2 B^{2s} - \cdots - \Theta_p B^{Qs}$ - sezoninis slenkančio vidurkio (MA) kompnentas;

s – sezono (ciklo) ilgis.

y_t – kintamojo reikšmė laiko momentu t,

μ – laiko eilutės vidurkis.

B – poslinkio atgal per vieną laiko vienetą operatorius.

Taikant prognozavimą į ateitį stebimos reikšmės y_t keičiamos jų įverčiais $\hat{y}_t$, taip keičiant gauname:

$$B\hat{y}_t = \hat{y}_{t-1}. \tag{27}$$

Skirtingoms parametrų reikšmėms egzistuoja skirtingi sezoniniai ARIMA modeliai, žemiau aprašyti dažniausiai pasitaikantys sezoniniai modeliai prognozavimui:

ARIMA(0,0,0)x(0,1,0) – sezoninis atsitiktinis vaikščiojimo modelis. Šis modelis vadinamas atsitiktinio vaikščiojimo modeliu, nes jis parodo, kad kiekvieno sezono reikšmių forma yra atsitiktinis vaikščiojimas. Pavyzdžiui, modelis parodo, kad šių metų rugsėjo reikšmė yra atsitiktinio žingsnio atstumu nuo praėjusių metų rugsėjo reikšmės.

Jeigu laiko eilučių sezoninis skirtumas panašus į atsitiktinį triukšmą, tai prognozuojamo modelio vidurkis turi būti lygus sezonų periodų skirtumui tam tikrame laiko momente t. Pritaikius vidurkio modelį šioms eilutėms gauname:

$$\hat{Y}_t - Y_{t-p} = \mu \tag{28}$$

perdarant lygtį galima gauti modelio prognozės lygtį

$$\hat{Y}_t = Y_{t-p} + \mu \tag{29}$$

μ – sezonų skirtumų vidurkis,

p – sezonų skaičius.

ARIMA(0,1,0)x(0,1,0) – sezoninis atsitiktinės tendencijos modelis. Dažnai laiko eilutės, kurios turi stiprią sezoniškumo tendenciją, nėra pakankamai stacionarios, todėl atsitiktinio vaikščiojimo modelis, kuris prognozuoja, kad sezoninis skirtumas yra konstanta, nerodo gerų rezultatų. Šiuo atveju naudojamas sezoninis atsitiktinės tendencijos modelis. Pirmasis sezoninio periodo t skirtumas turi tokią išraišką:

$$(Y_t - Y_{t-p}) - (Y_{t-1} - Y_{t-p+1}) \tag{30}$$

Pritaikant šioms eilutėms nulinio vidurkio prognozavimo modelį, gauname prognozės lygtį:

$$(\hat{Y}_t - Y_{t-p}) - (Y_{t-1} - Y_{t-p+1}) = 0 \tag{31}$$

Pertvarkant lygtį gauname:

$$\hat{Y}_t = Y_{t-p} + Y_{t-1} - Y_{t-p+1} \tag{32}$$

ARIMA(0,1,1)(0,1,1) - sezoninės atsitiktinės tendencijos modelis su pridėtomis nesezoninėmis ir sezoninėmis slenkančio vidurkio sąlygomis. Modelis yra patobulintas sezoninės atsitiktinės tendencijos modelis pridedant poslinkių klaidas, šios klaidos atsiranda dėl judančio vidurkio komponenčių. Šis modelis yra panašus „Winters“ modeliui, tačiau efektyviau pritaiko eksponentinį glodumą tendencijoms ir sezoniškumui. Modelį lygtimi prognozei galima išreikšti taip:

$$\hat{Y}_t = Y_{t-p} + Y_{t-1} - Y_{t-p+1} - \theta_1 e_{t-1} - \Theta_1 e_{t-p} + \theta_1 \Theta_1 e_{t-p+1} \quad (33)$$

θ_1 – judančio vidurkio nesezoninis koeficientas

Θ_1 - judančio vidurkio sezoninis koeficientas. [31]

2.2.2. Apibrėžtumo koeficientas

Dimensijos T vektoriai:

$SST = \sum_{t=1}^{T}(y_t - \bar{y})^2 = (y - i\bar{y})'(y - i\bar{y})$ - kvadratų suma;

$SSR = \sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 = (\hat{y} - i\bar{y})'(\hat{y} - i\bar{y})$ - regresijos kvadratų suma;

$SSE = \sum_{t=1}^{T}(y_t - \hat{y}_t)^2 = (y - \hat{y})'(y - \hat{y}) = \hat{\varepsilon}'\,\hat{\varepsilon}$ - kvadratų sumos paklaida.

Apibrėžėme nukrypimo įverčius:

$$\hat{Y}(y) = SST/T \quad (34)$$

$$\hat{Y}(\hat{y}) = SSR/T \quad (35)$$

$$\hat{Y}(\varepsilon) = SSE/T \quad (36)$$

Iš lygčių gauname apibrėžtumo koeficientą:

$$R^2 = 1 - \left(\frac{\hat{Y}(\varepsilon)}{\hat{Y}(y)}\right) = 1 - \left(\frac{SSE}{SST}\right) \quad (37)$$

[33]

3. TYRIMŲ REZULTATAI IR JŲ APTARIMAS.

Viena iš pagrindinių prekybos, šiuo atveju mažmeninės prekybos, sėkmės veiksnių yra tinkamos analitinės priemonės pasirinkimas. Šiais laikais yra sukurta labai daug skirtingų analitinių įrankių skirtų duomenų mokslui ir verslo analitikai. Analitinių platformų kūrėjai siūlo daugybę įvairių analizių, metodų, algoritmų, tačiau ne visos analitinės priemonės turi pakankamai išplėtotus įrankius norint visapusiškai atlikti duomenų ir verslo įžvalgų analizes. Net jeigu yra

galimybė atlikti analizes ir gauti norimus rezultatus, ne visos analitinės priemonės turi įrankius tų rezultatų tikslumui ir patikimumui nustatyti.

Šio tyrimo tikslas yra apžvelgti ir nustatyti, kurios pasirinktos pažangios analitinės priemonės turi geriausiai visapusiškai išplėtotus analitinius įrankius skirtus klasterinei ir laiko eilučių analizei nagrinėjant mažmeninės prekybos duomenis.

Programinių priemonių vertinimas buvo suskirstytas į tris dalis. Pirmoji dalis - bendrieji programinės priemonės kriterijai, antroji – klasterinės analizės atlikimo kriterijai, trečioji – laiko eilučių analizės atlikimo kriterijai.

Programinės priemonės bendrieji kriterijai:

- Dydis. Programinės priemonės galimybė apdoroti didelius duomenų masyvus.
- Įvairumas. Programinės priemonės galimybė apdoroti įvairių tipų failus.
- Integracijos. Kitų programinių priemonių integracijos į tiriamą programinę priemonę galimybė.
- Valdoma kodu. Programinės priemonės galimybė būti valdomai kodu.

Klasterinės analizės atlikimo kriterijai:

- Greitis. Kaip greitai programinė priemonė sugebėjo agreguoti 370278 eilučių ir 8 stulpelių duomenų masyvą į klasterinei analizei reikalingą duomenų masyvą.
- Metodų skaičius. Programinėje priemonėje integruotų metodų, skirtų klasterinei analizei, skaičius.
- Išskirčių šalinimas. Programinės priemonės galimybė surasti ir pašalinti išskirtis.
- Skalės problema. Programinės priemonės galimybė standartizuoti duomenis.
- Klasterių skaičiaus nustatymas. Programinės priemonės galimybė nustatyti klasterių skaičių.
- Klasterizavimo kokybės nustatymas. Klasterinės analizės kokybės įvertinimo galimybė programinėje priemonėje.
- Vizualizacijos. Programinės priemonės galimybė atvaizduoti klasterinei analizei reikiamas vizualizacijas.
- Statistikos. Reikiamų statistikų atvaizdavimo galimybė.

Laiko eilučių analizės atlikimo kriterijai:

- Greitis. Kaip greitai programinė priemonė sugebėjo agreguoti 370278 eilučių ir 8 stulpelių duomenų masyvą į klasterinei analizei reikalingą duomenų masyvą.
- Metodų skaičius. Programinėje priemonėje integruotų metodų, skirtų laiko eilučių analizei, skaičius.
- Stacionarumo nustatymas. Programinės priemonės galimybė įvertinti laiko eilutės stacionarumą.
- ARIMA parametrų radimas. Programinės priemonės galimybė automatiškai surasti ARIMA parametrus (p,d,q).
- Vizualizacijos. Programinės priemonės galimybė atvaizduoti laiko eilučių analizei reikalingas vizualizacijas.
- Statistikos. Reikiamų statistikų atvaizdavimo galimybė.

Tyrimui buvo pasirinktos lyderių pozicijas užimančios, Garnerio reitinguotos, analitinės platformos. Buvo atrinktos IBM, SAS ir KNIME sukurtos tinkamiausios analitinės priemonės analizuojant ir įvertinant mažmeninės prekybos duomenis.

3.1. Klasterinė ir laiko eilučių analizė

Pirmoji tyrimo dalis yra pažangiųjų analitinių priemonių galimybės atliekant klasterinę analizę, taip rekomenduojant pažangiausią, iš tyrimui pasirinktų, analitinę priemonę atliekant tam tikrą klientų segmentavimą. Norint giliau pažvelgti į analitinių priemonių galimybes, detaliai žingsnis po žingsnio, buvo atlikta klasterinė analizė naudojantis RFM modeliu, atliekant k-vidurkių klasterizavimo metodą pasirinktomis analitinėmis priemonėmis. RFM modelio pagrindu buvo atlikta klasterinė analizė norint sužinoti ar remiantis tai kaip klientas neseniai, dažnai pirko prekes ir kiek toms prekėms išleido apsipirkdamas yra galima atlikti reikšmingą ir prasmingą klientų segmentavimą.

Antroje tyrimo dalyje buvo atlikta laiko eilučių analizė, o tiksliau pardavimų prognozavimas į ateitį. Kaip buvo minėta pardavimų, atsargų ar kitų rodiklių prognozavimas yra viena iš pagrindinių įmonės analizių, nes tai gali užtikrinti įmonės sėkmę. Šiuo atveju buvo prognozuoti internetinės parduotuvės ateities gautinų pinigų suma tam tikromis dienomis. Internetinei parduotuvei ši analizė galėtų būti naudinga investicijų, kaip pavyzdžiui į prekes, atsargas, patalpas ar žmogiškuosius resursus planavimui.

3.2. Aprašomoji duomenų analizė

Tyrimui naudojami internetinės parduotuvės duomenys. Internetinė parduotuvė daugiausiai pardavinėja unikalias, progines dovanas. Duomenys susideda iš 541909 eilutės ir 8 stulpelių. Trumpai apžvelgsime stulpelių informacija:

1. Pirkimo numeris. 6 skaičių unikalus kiekvienam pirkimui numeris.
2. Prekės kodas. 5 skaičių unikalus prekės kodas. Kiekviena prekė turi savo kodą.
3. Prekės pavadinimas.
4. Prekės kiekis. Vienu pirkimu nupirktas prekės kiekis.
5. Pirkimo data. Pirkimo datoje nurodyti pirkimo metai, mėnesis, diena, valandos ir minutės. Laiko intervalas yra nuo 2011.09.20 11:05:00 iki 2011.12.09 12:00:00.
6. Prekės kaina. Prekės vieneto kaina.
7. Kliento identifikatorius. 5 skaičių unikalus kliento identifikatorius.
8. Šalis. Šalis iš kurios buvo pirktos prekės.

Duomenyse egzistavo nepilnos informacijos ir nelogiškų įrašų, kaip pavyzdžiui pirktas prekės kiekis - neigiama reikšmė. Tokie įrašai buvo pašalinti iš duomenų masyvo. Pašalinus įrašus duomenų masyve liko 370278 eilutės. Pirkimo data buvo išskaidytą į stulpelius diena ir laikas. Stulpelyje diena nurodyti pirkimo metai, mėnesis ir diena, o stulpelyje laikas – pirkimo valanda ir minutės. Taip pat buvo sudarytas dar vienas stulpelis – pardavimo suma. Šį stulpelį sudaro pirktas prekės kiekis padaugintas iš prekės kainos. Taip gaunama vieno pirkimo pinigų suma.

Kiekvienam tyrimui buvo atlikta duomenų agregavimas. Visose programinėse priemonėse yra integruota agregavimo galimybė. Duomenys buvo neagreguoti tik atliekant klasterinę analizę naudojantis SAS Enterprise Miner. Bandomoji SAS Enterprise Miner analitinė programa nesuteikė galimybės nuskaityti tokio dydžio failo, todėl agregavimas buvo atliktas naudojantis kita analitine platformą, taip į SAS Enterprise Miner integruojant sumažintą failą. Taip buvo išsiaiškinta, kad visos programinės priemonės sugeba apdoroti didelius duomenų masyvus. Agregavimo metu buvo įvertinti programinių priemonių greičiai. Agreguojant nagrinėjamą, duomenų masyvą, programinės priemonės užtruko skirtingą laiko intervalą. Agreguojant pradinį duomenų masyvą į duomenų masyvą, skirtą klasterinei analizei, SPSS Modeler užtruko 2 min. 26 s., laiko eilučių analizei – 2 min.. SAS – atitinkamai 20 s. ir 15 s., o KNIME – 7 s. ir 4 s. Šių dienų versle egzistuoja daug didesni, nei nagrinėjami duomenų

masyvai, todėl SPSS Modeler nėra pritaikyta greitam duomenų apdorojimui. Šiuo atveju, greičiausia programinė priemonė yra KNIME.

Klasterinei ir laiko eilučių analizei duomenys buvo skirtingai agreguojami. Klasterinei analizei buvo sukurtas papildomas stulpelis „Recency“ – kaip seniai klientas pirko. Taip pat duomenys buvo suagreguoti pagal Kliento numerį, transformuojant duomenis taip, kad kiekvienam klientui būtų priskirta kiek kartų per nagrinėjamą laikotarpį jis apsipirko internetinėje parduotuvėje ir suminė klientų išleidžiamų pingų reikšmė. Šia reikšmę vadinsime Frequency ir Monetary. Taip gauname RFM analizės kintamuosius – Recency, Frequency ir Monetary. Laiko eilučių analizei duomenys buvo agreguoti naudojantis „diena“ stulpelį. Pagal dieną buvo susumuoti pardavimai, t.y. gautos kiekvienos dienos pajamos.

Prieš atliekant analizę, buvo patikrintos programinės priemonės galimybės nuskaityti įvairių tipų failus ir galimybė kitų programinių priemonių integracija į nagrinėjamą priemonę. Visos programinės priemonės turi integruotus įrankius įvairių tipų failams nuskaityti. Kiekvienoje programinėje priemonėje jų galima suskaičiuoti daugiau nei 10. Kitų programinių priemonių integracijos egzistuoja visose nagrinėjamose priemonėse. Taip pat kiekvienoje programinėje priemonėje yra integruotos kitų programinių priemonių, kaip pvz. „R“, „Python“ ir kt., naudojimosi galimybės.

3.3. Klasterinė analizė naudojantis IBM analitine platforma.

Pažangiam duomenų mokslui IBM plėtoja dvi analitines priemones – Watson ir SPSS. IBM kaip pažangiausią analitinę priemonę atlikti klasterinę ir laiko eilučių analizę nurodo IBM SPSS. IBM SPSS plėtoja dvi analitines priemones – Statistics ir Modeler. Pabandžius abejas pažangias analitines priemones ir apžvelgus abiejų analitinių priemonių įvertinimus buvo pasirinkta SPSS Modeler. SPSS Modeler analitinė priemonė turi pažangesnes vizualizacijos galimybes, yra paprastesnė. Tarp analitinių galimybių daug skirtumų nebuvo pastebėta. Atliekant klasterinę analizę buvo naudota SPSS Modeler 18. bandomąja versija. Analitinė priemonė SPSS Modeler yra valdoma mazgais.

3.3.1. Išskirčių radimas ir šalinimas

Išskirtys – netipinės ir retos reikšmės, kurios yra žymiai nukrypusios nuo kitų duomenų pasiskirstymo.

Žinant, kad k – vidurkių metodas yra labai jautrus esančioms išskirtims, reikalingas išskirčių radimas ir šalinimas.

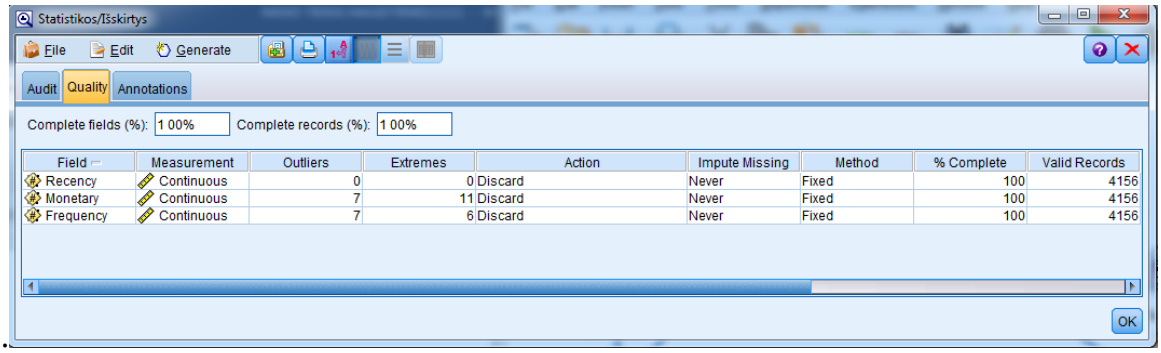
„Data Audit“ mazgas leidžia pasirinkti išskirčių ir ekstremalių reikšmių ieškojimo metodus:

- Atrenkamos reikšmės, kurių standartinis nuokrypis yra didesnis už pasirinktą reikšmę.

- Atrenkamos reikšmės, kurios yra mažesnės reikšmės negu apatinis kvartilis ir didesnės reikšmės negu viršutinis kvartilis.

Pasirinkus metodą ir reikšmes programa atrenka reikšmes, kurios potencialiai yra išskirtys.

Nenorint stipriai sumažinti duomenų kiekio, buvo šalinamos tik ryškios išskirtys, todėl buvo parinktas išskirčių, kurių standartinis nuokrypis daugiau nei 7 šalinimas.



5 pav. Atrastos ir pašalintos išskirtys

3.3.2. Skalės problema

Norint tinkamai atlikti klasterinę analizę reikia pašalinti galimai esamą skalės problemą, atliekant kintamųjų standartizavimą. Kintamųjų standartizavimas reikalingas tam, kad kintamųjų matavimo skalė būtų suvienodinta, nes skaičiuojant atstumus tarp atveju didesnės skalės kintamieji turės didesnę įtaką negu mažesnės skalės.[34] Naudojamo k – vidurkių metodo mazgas „K-Means“ atliekant klasterinę analizę automatiškai standartizuoja kintamuosius, paversdamas skalę taip, kad minimali galima reikšmė butų 0, o maksimali 1.

3.3.3. Duomenų dalinimas į apmokymo ir testavimo imtis.

Atliekant duomenų dalinimą į apmokymo ir testavimo imtis buvo panaudotas „Partition“ mazgas. Nagrinėjami duomenys buvo padalinti atitinkamai 80% duomenų arba 3343 klientai buvo priskirta apmokymo imčiai, o 20% duomenų arba 788 klientai testavimo imčiai.

3.3.4. Modelio sudarymas

SPSS Modeler pateikia galimybes atlikti klasterinę analizę keliais metodais:

- K – vidurkių
- Kohonen
- Dviejų žingsnių

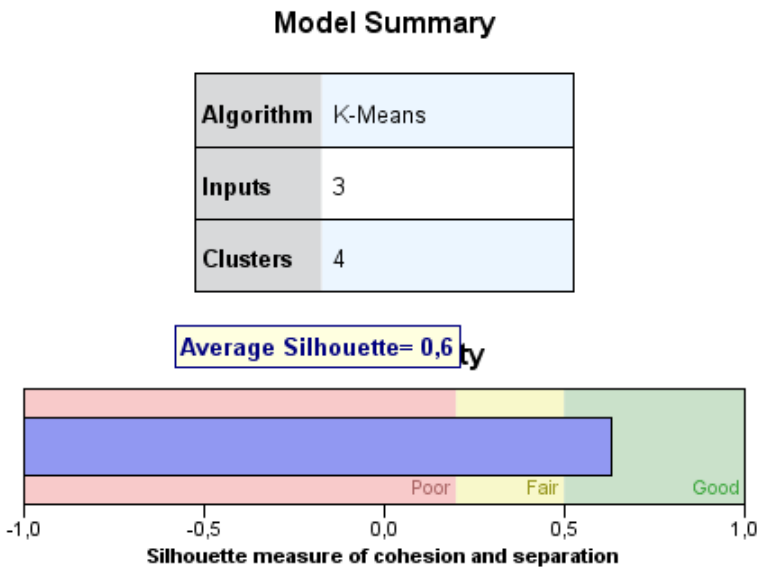
Kaip buvo minėtą, analizei buvo pasirinktas k – vidurkių metodas, sudarant modelį naudosime apmokymo duomenis.

Iš teorijos žinoma, kad naudojantis k – vidurkių metodu reikia iš anksto žinoti klasterių kiekį. Nustatant klasterių kiekį buvo atliktas silueto koeficiento(angl. The Silhouette Coefficient) apskaičiavimas. Šis koeficientas nurodo klasterių kokybę. Koeficientas gaunamas naudojantis „K-Means“ mazgu, nurodžius, kintamuosius pagal kuriuos buvo klasterizuota. Nurodžius nagrinėjamus kintamuosius – Recency, Frequency ir Monetary ir atlikus klasterizavimą keičiant klasterių kiekį buvo gauti tokie silueto koeficientai:

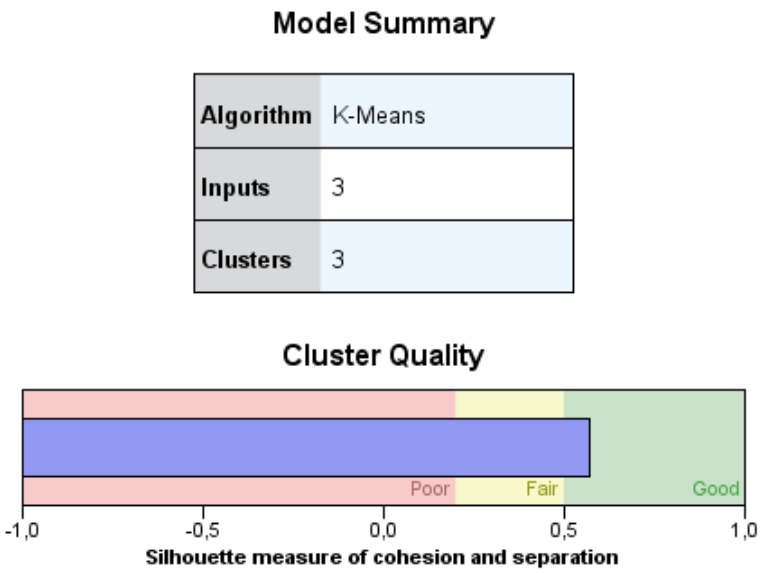
1 lent. Silueto kriterijaus reikšmės besikeičiant klasterių skaičiui

Klasterių skaičius	*3*	*4*	*5*	*6*	*7*
Silueto koeficientas	0,6	0,6	0,5	0,5	0,5

Iš 1 lent. matoma, kad geriausia modelio kokybė kai klasterių skaičius yra 3 arba 4. Tačiau vizualiai iš *6* ir *7 pav*. matoma, kad tinkamiausias klasterių skaičius yra 4.



6 pav. Klasterizavimo kokybė „K-Means“ mazge naudojantis silueto kriterijumi, kai analizei naudojami 4 klasteriai

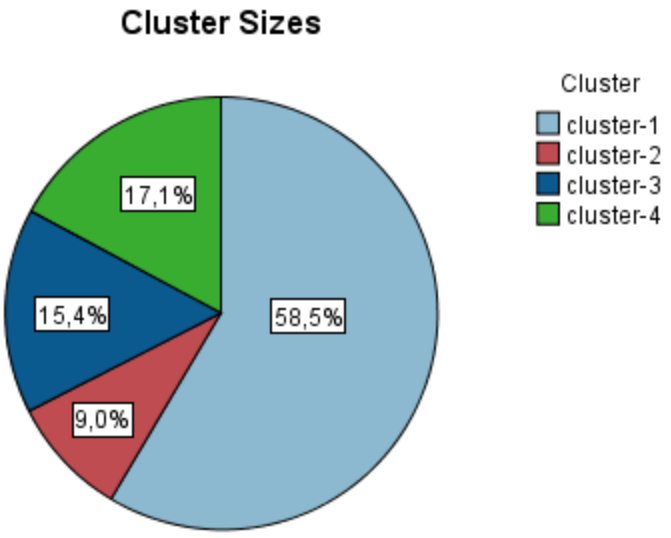


7 pav. Klasterizavimo kokybė „K-Means“ mazge naudojantis silueto kriterijumi, kai analizei naudojami 3 klasteriai

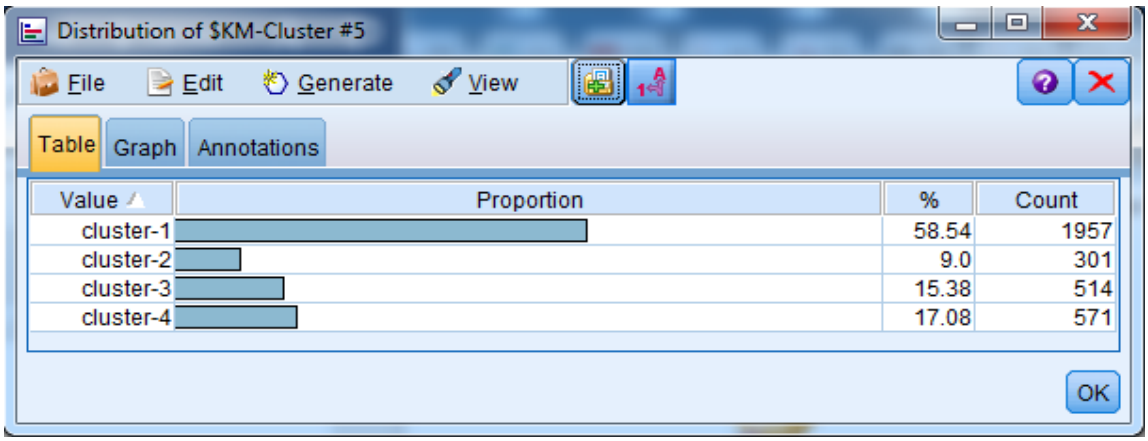
Kitas žingsnis nustatyti iteracijų skaičių. „K-Means“ mazgas leidžia pasirinkti maksimalų iteracijų skaičių. Buvo nustatytas maksimalus iteracijų skaičius – 50, tačiau modeliui prireikė tik 28 iteracijų, vykdant 29 iteraciją klaidų skaičius buvo lygus 0.

Nustačius visus parametrus atliekama klasterinė analizė.

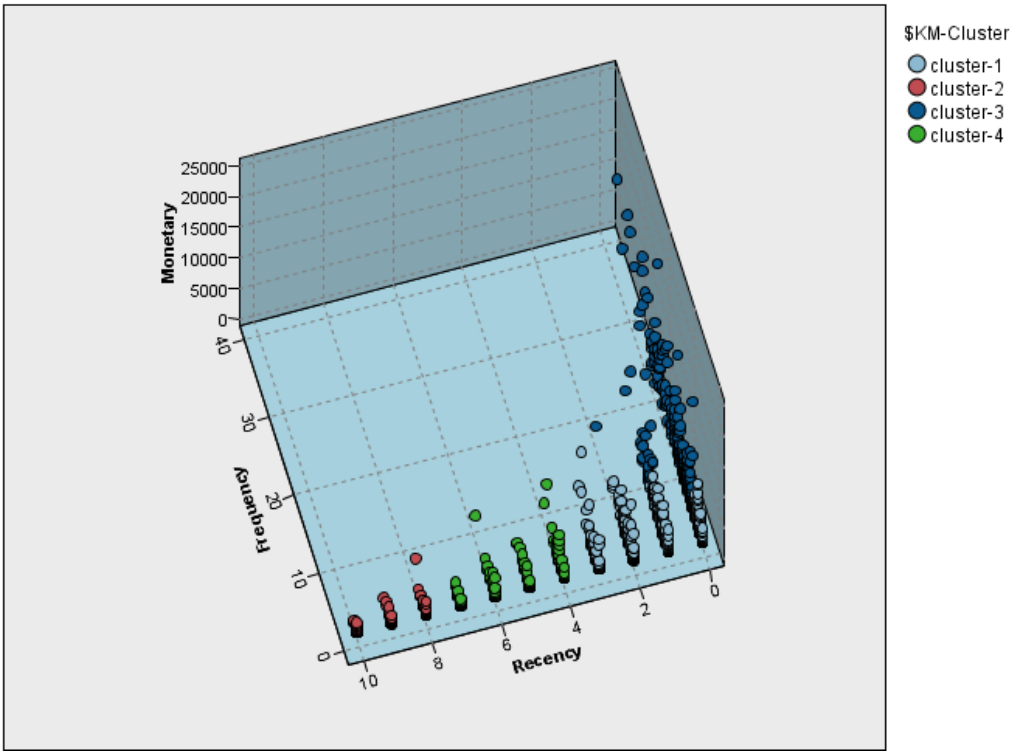
„K-Means“ mazgas sugeneruotų klasterių dydžius gali atvaizduoti keliais būdais, vienas iš jų - grafikas, kur vizualiai nurodyti klasterių dydžiai su procentais. (*8 pav.*). Naudojant „Distribution“ mazgą lentele buvo atvaizduotos klasterių proporcijos (*9 pav.*). Įvairioms vizualizacijoms atlikti SPSS Modeler aplinkoje yra naudojamas mazgas „Graphboard“. Panaudojus šį mazgą, taškine diagrama buvo atvaizduoti klasterių ir klientų pasiskirstymai.(*10* ir *11 pav.*).



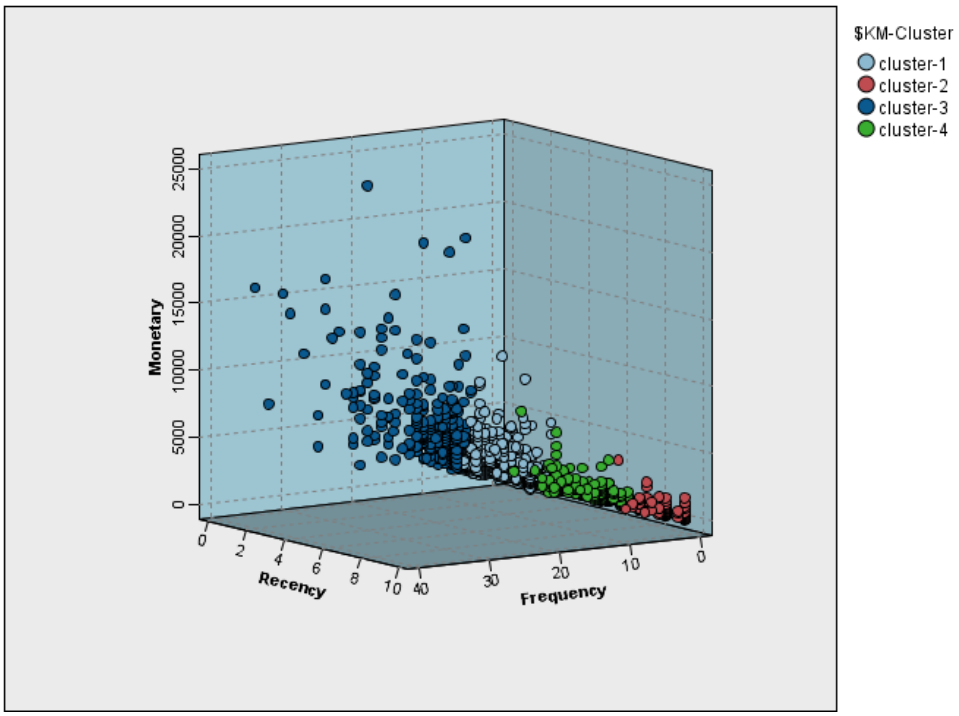
8 pav. Klasterių dydžių grafikas



9 pav. Klasterių proporcijos



10 pav. Klasterių ir klientų klasteriuose išsidėstymo grafikas (1)



11 pav. Klasterių ir klientų klasteriuose išsidėstymo grafikas (2)

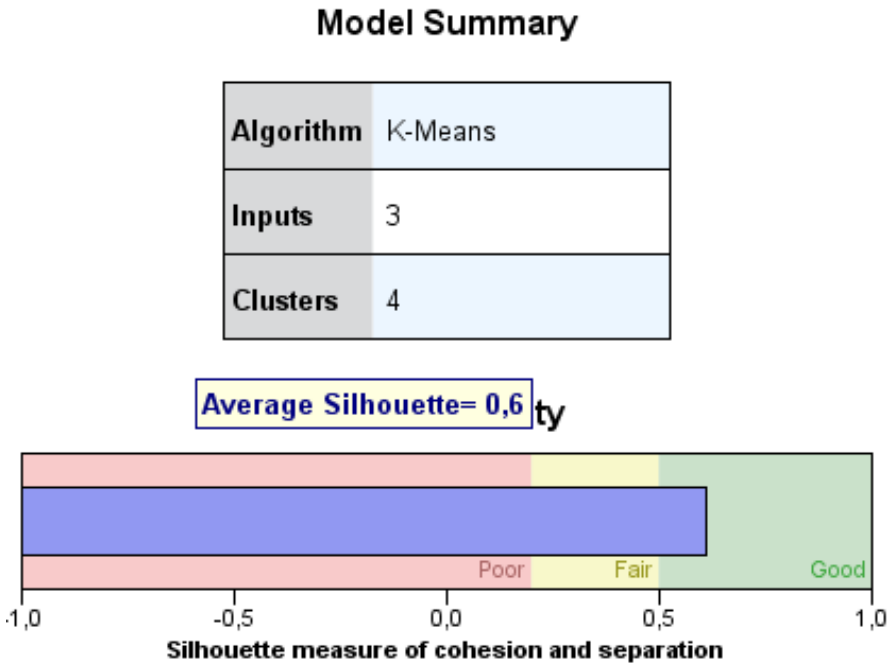
2 lent. Klasterių statistikos

	Minimali reikšmė	Vidurkis	Maksimali reikšmė
Cluster 1			
Recency	0	0.943	3
Frequency	1	2.5	9
Monetary	201.12	4188.404	22558.74
Cluster 2			
Recency	8	8.784	10
Frequency	1	1.203	5
Monetary	17.55	364.165	4036.96
Cluster 3			
Recency	0	0.084	2
Frequency	4	10.37	35
Monetary	201.12	4188.404	22558.74
Cluster 4			
Recency	4	5.361	7
Frequency	1	1.732	10
Monetary	2,9	557.165	9864.26

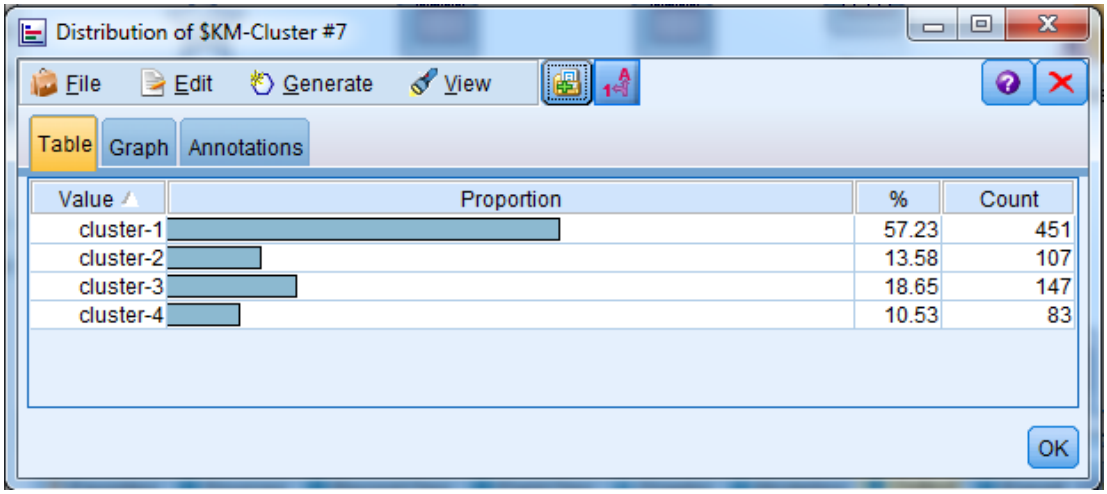
3.3.5. Modelio tinkamumo įvertinimas

Prieš atliekant klasterių apibendrinimą reikia patikrinti modelio tinkamumą. Vienas iš būdų patikrinti klasterinės analizės modelio teisingumą yra duomenų suskaidymas į poimčius ir stebėjimas ar atliekant tą patį modelį abiem poaibiams klasterių proporcijos reikšmingai nesiskiria.

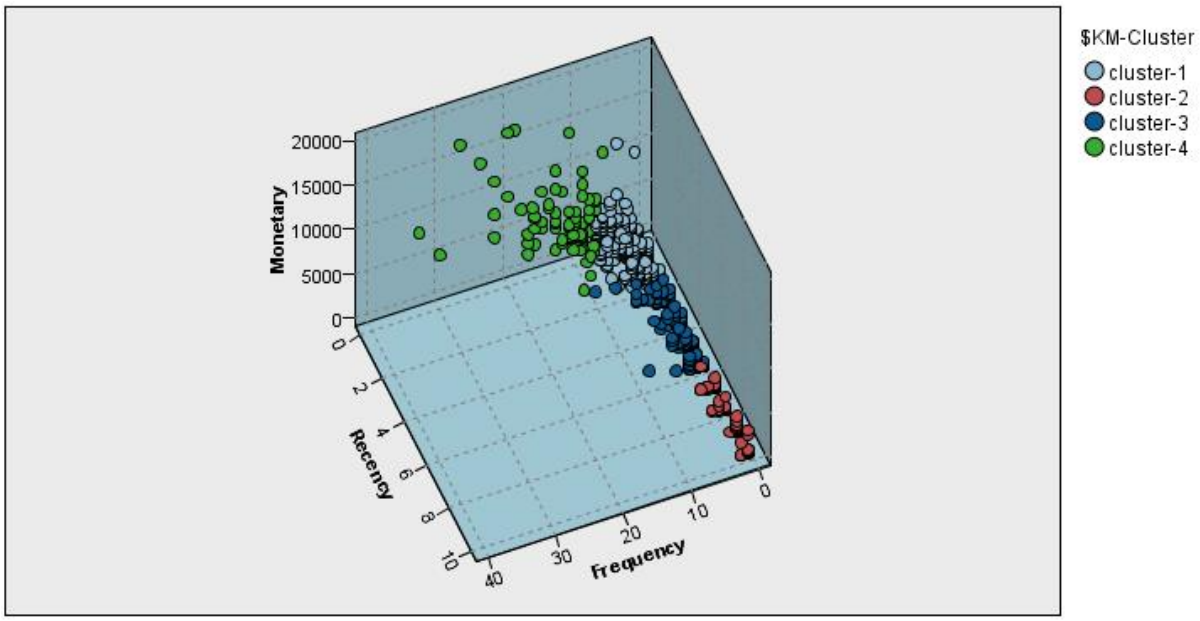
Kaip buvo minėta, duomenys buvo suskaidyti į apmokymo ir testavimo imtis. Nekeičiant parametrų atlikus klasterinę analizei testavimo imčiai buvo gauti tokie rezultatai:



12 pav. Silueto koeficientas



13 **pav.** Klasterių proporcijos



14 pav. Klasterių ir klientų pasiskirstymas

Iš *12, 13 ir 14 pav*. galima matyti, kad klasterizavimo kokybė išliko nepakitus. Klasterių proporcijos ir pasiskirstymas nuo su apmokymo duomenimis atlikto modelio skiriasi nereikšmingai, tad galima teigti, kad modelis teisingas

3.3.6. Klasterių apibendrinimas.

Suprasti ir interpretuoti atrastus klasterius yra vienas iš svarbiausių į klientus orientuotų verslo įžvalgos analizių.

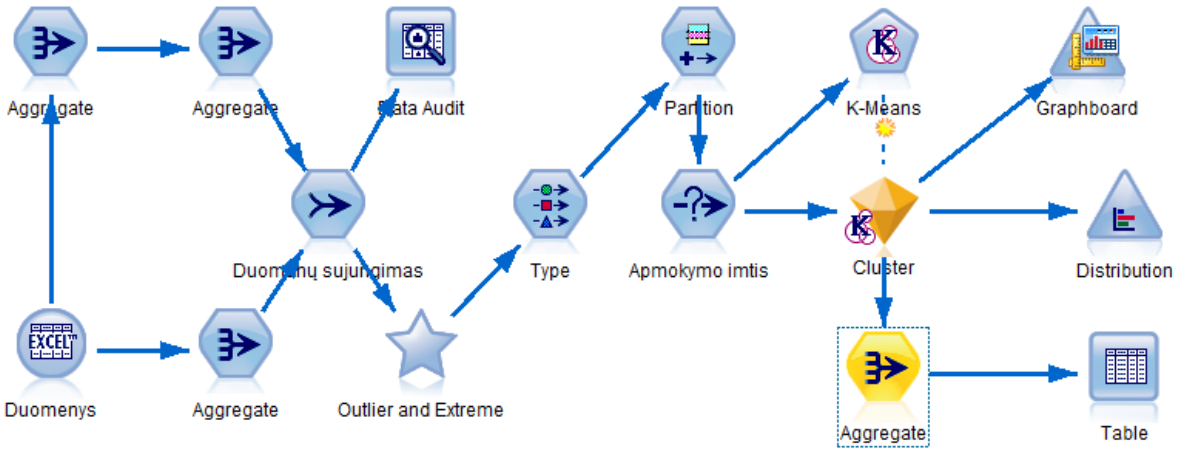
Atliktą klasterinę analizę pritaikytą RFM modeliui analizuosime atsižvelgdami į *8, 9, 10* ir *11 pav*. bei į 2 lent.

Pirmasis klasteris susideda iš 1957 klientų, kurie sudaro 58,5% visos populiacijos. Šiame didžiausiame klasteryje yra įtraukti pirkėjai, kurie paskutinį kartą buvo apsipirkę per paskutiniuosius 3 mėnesius. Atsižvelgiant į pardavimų apimtis tai yra klientai, kurie retai apsiperka internetinėje parduotuvėje, nes rodiklio Frequency (kaip seniai klientas pirko) vidurkis yra 2.5 karto. Tai yra antras pagal vidutinį kliento pelningumą klasteris, nes vidutiniškai klientas per nagrinėjamą laikotarpį pirkimams išleisdavo 916,512$. Šis klasteris yra labai svarbus internetinei parduotuvei, nes šio klasterio klientai sugeneruoja daugiau nei trečdalį internetinės parduotuvės pajamų.

Antrąjį ir ketvirtąjį klasterį atitinkamai sudaro 301 ir 571 klientas ir tai yra 9 ir 17,1 procentai populiacijos. Šie klasteriai yra mažiausiai pelno atnešančios klientų grupės, kartu šių klasterių klientai nesudaro dešimtadalio internetinės parduotuvės pardavimų sumos. Šiai grupei priklausantys klientai paskutinius keturis mėnesius internetinėje parduotuvėje nėra pirkę prekių,

be to jie vidutiniškai per nagrinėjamą laikotarpį pirko atitinkamai tik 1,203 ir 1,732 kartus. Šiuos klientus galima interpretuoti kaip specifines prekes perkančius klientus arba kaip klientus nusivylusius šia internetine parduotuve.

Trečioji klientų grupė susideda iš 514 klientų ir tai sudaro 15,4% populiacijos. Paskutinį kartą šios grupės klientai internetinėje parduotuvėje apsipirko per paskutiniuosius du nagrinėjamo laikotarpio mėnesiu ir tai galimai yra nuolatiniai klientai, nes vidutinis apsipirkimo dažnis – Frequency (kiek kartų klientas pirko) yra 10,37 kartai. Tai yra pati pelningiausia klientų grupė, kurioje vidutiniškai klientas per nagrinėjamą laikotarpį išleido 4188,404$. Šie klientai yra patys svarbiausi internetinei parduotuvei, nes jie generuoja beveik pusę šios parduotuvės pelno. Internetinės parduotuvės savininkas turėtų didžiausią dėmesį skirti šių lojaliausių ir didžiausią pelną atnešančių klientų išlaikymui.



15 pav. Modelio realizavimas SPSS Modeler aplinkoje

3.4. Laiko eilučių analizė su IBM analitine platforma.

Kaip ir klasterinei analizei, taip ir laiko eilučių analizei naudojama IBM SPSS Modeler 18 bandomoji versiją. Bandomoji versija atliekant laiko eilučių analize suteikė laiko taškų skaičiaus ribojimus, todėl analizė buvo atlikta su dienos lygyje agreguotais duomenimis.

Prieš pradedant modelio sudarymą, naudojantis „Partition“ mazgu duomenys buvo padalinti į apmokymo ir testavimo imtis. Testavimo imtis susideda iš paskutinių dviejų savaičių pardavimų, o likę duomenys priklauso apmokymo imčiai.

3.4.1. Modelio sudarymas

Norint sudaryti laiko eilutę SPSS Modeler aplinkoje reikia nurodyti laiko intervalus, o tai programinėje priemonėje atlieka „Time Intervals“ mazgas. Nurodžius laiko intervalus, kas mūsų

atveju buvo iš anksto suagreguota, galima atlikti laiko eilutės analizę. Laiko eilutės sudarymo funkciją SPSS Modeler programinėje priemonėje atlieka „Time Series“ mazgas.

„Time Series“ mazgas suteikia duomenų specifikavimo galimybes, kaip pavyzdžiui nustatymas, kokie laiko intervalai norimi prognozavimui, duomenų agregavimas, skirtingais metodais trūkstamų reikšmių užpildymas. Laiko intervalas nagrinėjamame modelyje buvo pasirinktas iš anksto, tai suagreguotos 271 dienos. Trūkstamos reikšmės buvo pakeistos į eilutės vidurkio reikšmes. Sudarant laiko eilutės ir prognozavimo modelį „Time Series“ series mazgas suteikia tokius pasirinkimus:

- ARIMA modeliavimas;
- Eksponentinio glodinimo modeliavimas;
- Ekspertinis modelis.

ARIMA modeliavime galima pasirinkti parametrus tiek sezoniniam, tiek nesezoniniam ARIMA modeliui sudaryti.

ARIMA nesezoniniam modeliui leidžia pasirinkti norimą p - autoregresijos eilę, d-diferencijavimo eilę ir q – slenkančiųjų vidurkių narių skaičių. Sezoniniam ARIMA modeliui papildomai galima nurodyti P – sezoninių autoregresinių sąlygų skaičių, D – sezoninių skirtumų skaičių, Q – sezoninių judančio vidurkio sąlygų skaičių.

Renkantis eksponentinio glodinimo modeliavimą, galima rinkti vieną iš kelių modelių tipų. Paprastasis, Holto tiesinės tendencijos, Brauno tiesinės tendencijos, slopinimo tendencijos, paprastasis sezoninis, „Winters“ adityvusis ir „Winters dauginamasis modelis yra tarp eksponentinio glodinimo modelio tipo pasirinkimų.

„Time series“ mazgas suteikia galimybę pasirinkti ekspertinį modeliavimą. Šis modelio tipas automatiškai identifikuoja ir apskaičiuoja geriausiai tinkantį ARIMA ar eksponentinio glodinimo modelį nagrinėjamų duomenų laiko eilutės sudarymui.

Sudarant laiko eilutę buvo pasirinkta ekspertinio modeliavimo galimybė.

Nustačius laiko eilutės modelio parametrus, mazgas suteikia galimybę prognozuoti norimus įrašus į ateitį. Modelio nustatymuose galima nustatyti norimų prognozuojamų į ateitį įrašų skaičių. Tyrime prognozuojama 14 įrašų į ateitį, kas mūsų duomenyse atitinka dvi savaites.

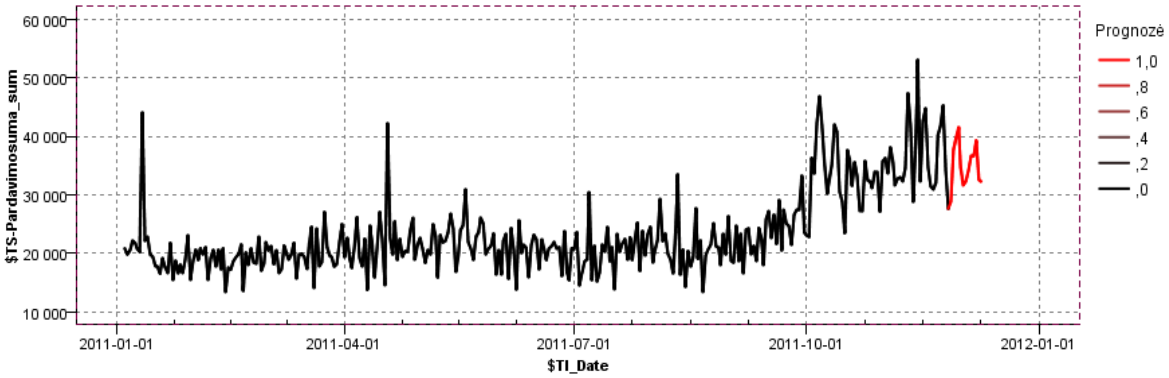
3.4.2. Rezultatai

Naudojant apmokymo imtį buvo paleistas modelis ir gauti rezultatai:

Model Information

Model Building Method		ARIMA
		Non-seasonal p=0,d=0,q=14; Seasonal p=0,d=0,q=0
Number of Predictors		0
Model Fit	MSE	4,795E+007
	RMSE	6 924,750
	RMSPE	5,040
	MAE	5 190,144
	MAPE	29,719
	MAXAE	23 374,953
	MAXAPE	350,078
	AIC	5 774,416
	BIC	5 808,498
	R Square	0,528
	Stationary R Square	0,528
Ljung-Box Q(#)	Statistic	20,100
	df	14,0
	Significance	0,1

16 pav. Apmokymo duomenų statistikos

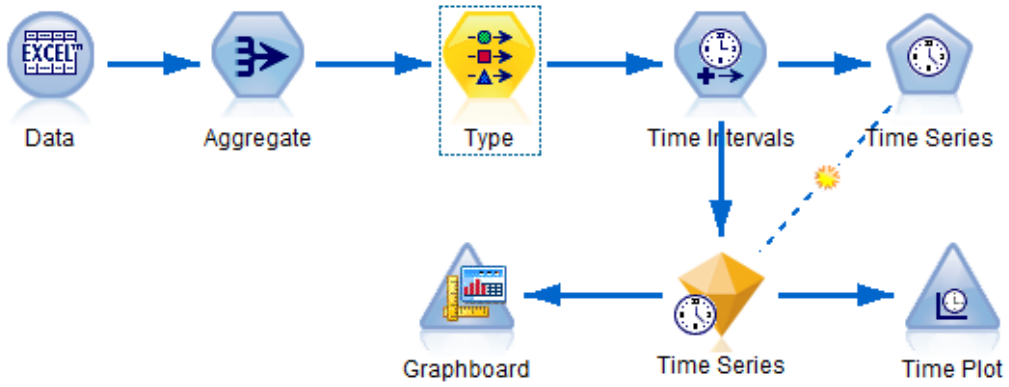


17 pav. Sudaryto modelio ir prognozės grafikas

16 pav. atvaizdavo ekspertinio modeliavimo parinktą modelį, iš čia galima pasakyti, kad laiko eilutė stacionari, nes ekspertinio modelio su sudarytos laiko eilutės diferencijavimo eilės parametro reikšmė lygi 0. Kaip matome geriausiai modeliui tiko nesezoninis ARIMA(0,0,14) modelis. Iš *17 pav.* galime matyti prognozuoto modelio grafiką ir raudonai pažymėtas prognozuotas reikšmes į ateitį. *16 pav*. taip pat atvaizdavo daugybę statistikų modelio įvertinimui, tačiau šiuo atveju, atsižvelgiant į persimokymo galimybę, negalima sakyti, kad prognozuojamų duomenų statistikos atitinka *16 pav*. nurodytas statistikas. Patikrinti modelio tikslumą, palyginsime prognozuotas reikšmes su testavimo imties reikšmėmis. Palyginus prognozuotus įrašus su testavimo imties įrašais apibrėžtumo koeficientas (angl. R-Squared) yra lygus 0.378. Iš pirmo žvilgsnio atrodo, kad tai mažas tikslumas, tačiau žinant, kad tai stacionari, nesezoninė laiko eilutė su nedideliu kiekiu įrašų galima daryti išvadą, kad tikslumas patenkinamas.

Nors tokia pardavimo sumos kiekio prognoze nebūtų galima visiškai pasitikėti, tačiau turint daugiau duomenų ir tarkime antrų metų pardavimus, tikslumas būtų tikrai geresnis ir tai galėtų labai pagelbėti verslui prognozuojant pardavimus ir atsargas. SPSS Modeler programinė priemonė turėjo visus įrankius norint atlikti pažangų pardavimų prognozavimą, tačiau pastebimas testų ir statistikų atvaizdavimo trūkumas.

SPSS Modeler turi visus reikiamus įrankius atliekant klasterinę analizę, atvaizduojant vizualizacijas ir gaunant norimas statistikas. Ši pažangi analitinė programa galėtų būti įmonės pagalbinė priemonė atliekant verslo įžvalgas pagrįstas klasterine analize. Dėl programinės priemonės paprastumo primityvią klasterinę analizę gali atlikti ir pradedantysis, tačiau norint atrasti norimus modelius, rezultatus ir statistikas reikalingas įdirbis.



18 pav. Laiko eilučių modelio realizavimas SPSS Modeler aplinkoje

3.5. Programinės priemonės apibendrinimas

3 lent. Bendrieji SPSS Modeler kriterijai

SPSS Modeler	
Dydis	+
Įvairumas	+
Integracijos	+
Valdoma kodu	-

4 lent. Klasterinės analizės naudojantis SPSS Modeler kriterijai

SPSS Modeler	Klasterinė analizė
Greitis	2 min. 26 s.
Metodų skaičius	3
Išskirčių šalinimas	+
Skalės problema	+

Klasterių skaičiaus nustatymas	-
Klasterizavimo kokybės nustatymas	+
Vizualizacijos	+
Statistikos	+

5 lent. Laiko eilučių analizės naudojantis SPSS Modeler kriterijai

SPSS Modeler	Laiko eilučių analizė
Greitis	2 min.
Metodų skaičius	2
Stacionarumo nustatymas	Automatinis
ARIMA parametrų radimas	Automatinis
Vizualizacijos	+
Statistikos	+

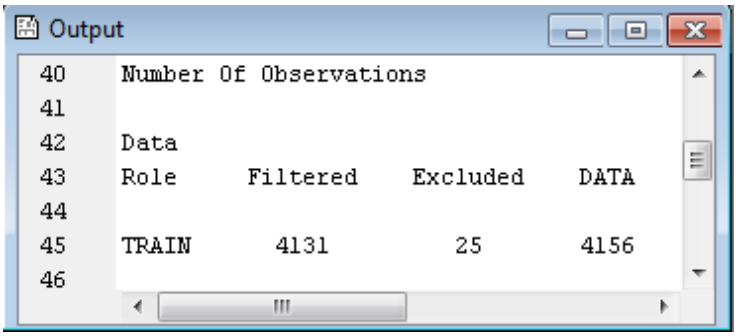
3.6. Klasterinė analizė naudojantis SAS analitine platforma

SAS analitinė platforma plėtoja daug pažangių analitinių priemonių. Renkantis programinę priemonę atliekant klasterinę analizę buvo remtasi SAS instituto darbuotojo patarimu. Buvo nurodyta, kad pažangiausia SAS plėtojama analitinė priemonė klasterizavimui yra SAS Enterprise Miner. Ši programinė priemonė yra valdoma mazgais, tačiau naudojantis mazgais yra generuojami programinių kalbų SAS, C, Java ir PMML kodai.

Analizei atlikti buvo naudojama akademiniams tikslams skirtas SAS Enterprise Miner 14.1 įrankis.

3.6.1. Išskirčių radimas ir šalinimas

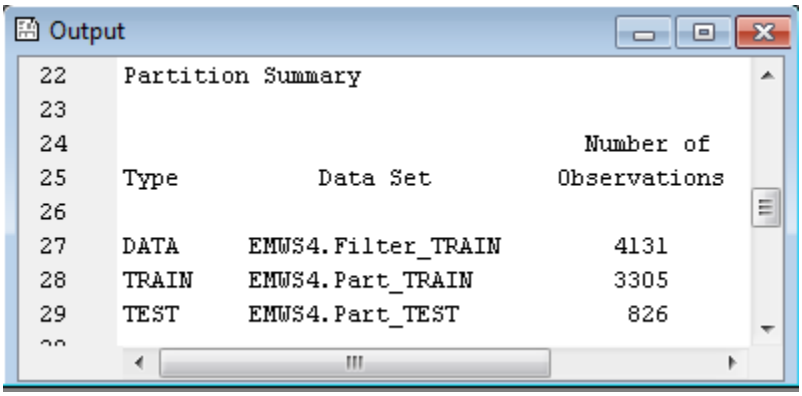
Programinėje priemonėje SAS Enterprise Miner išskirčių radimą ir šalinimą atlieka mazgas „Filter“. Išskirtys buvo pašalintos tuo pačiu būdu, t.y. pašalinti įrašai, kurių reikšmės nukrypusios daugiau nei 7 standartinio nuokrypio vidurkiai. *19 pav*. parodo duomenų skaičių prieš ir po išskirčių pašalinimo, bei rastų išskirčių kiekį.



19 pav. Atrastos ir pašalintos išskirtys

3.6.2. Duomenų dalinimas į apmokymo ir testavimo imtis

Duomenų dalinimui SAS Enterprise Miner aplinkoje atliekamas su „Data Partition“ mazgu. Duomenys buvo padalinti į dvi dalis. Apmokymo imčiai buvo skirta 80% įrašų, testavimo – 20%. Iš *20 pav*. matome kaip duomenys pasiskirstė į apmokymo ir testavimo imtis.



20 pav. Duomenų imtys prieš ir po duomenų padalinimo

3.6.3. Modelio sudarymas

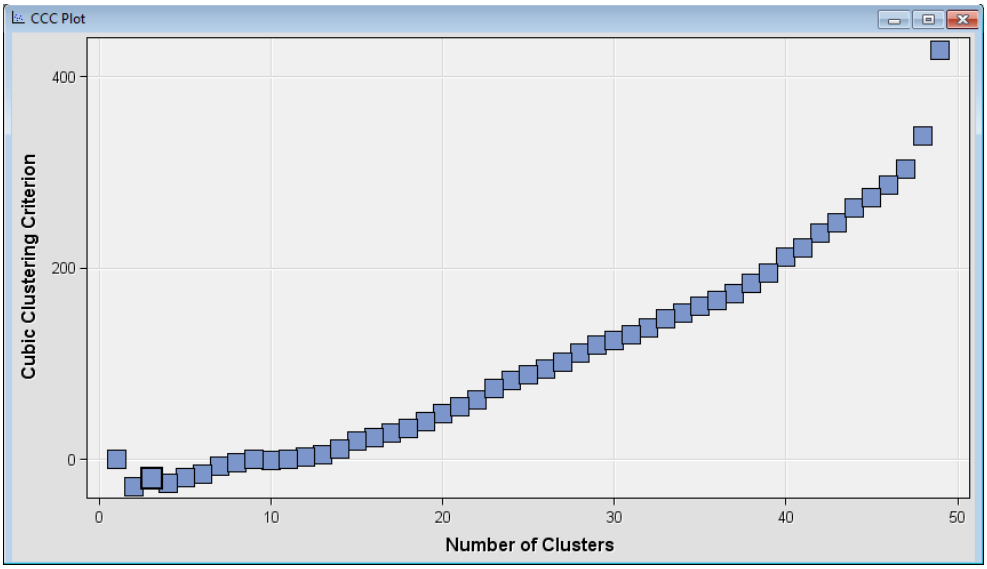
SAS Enterprise Miner aplinkoje klasterinę analizę atlieka trys mazgai:

- Cluster
- Variable Clustering
- SOM/Kohonen

SAS Enterprise Miner aplinkoje klasterinę analizę k- vidurkių metodu atlieka „Cluster“ mazgas. Šis mazgas savyje turi stacionarizavimo funkciją, kuri išsprendžia skalės problemą. Kaip žinoma k-vidurkių metodui reikia iš anksto nustatyti klasterių kiekį. Klasterių kiekiui nustatyti programinėje priemonėje yra integruota kubinio klasterizavimo kriterijaus atvaizdavimo galimybė. Prieš tikrinat klasterių skaičių, pirmiausia reikia pasirinkti atstumo apskaičiavimo metodą, pagal kurį bus grupuojami klasteriai. Programinėje priemonėje yra pateikti 3 grupavimo metodai:

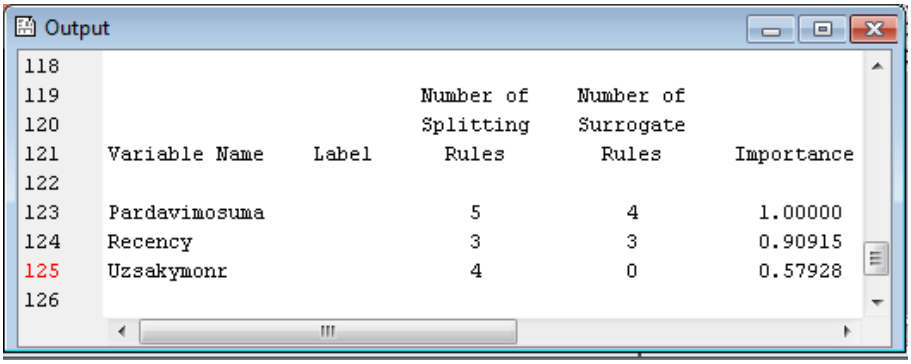
- Ward;
- Centroid;
- Vidurkio.

Klasterinei analizei atlikti buvo pasirinktas Ward metodas. Atliekant klasterinę analizę, kaip atstumo matą pasirinkus Ward metodą, gautas klasterių skaičius pagal kubinio klasterizavimo metodą yra 3.(*21 pav*.)



21 pav. Kubinio klasterizavimo kriterijaus grafikas

Nustačius visus parametrus buvo atliktas klasterizavimas naudojantis apmokymo imtimi ir stebimi rezultatai. Norint gauti įvairias statistikas ir vizualizacijas reikalingi „Segment Profile“ ir „StateExplore“ mazgai. Iš *23 pav*. matoma kaip klientai pasiskirstė klasteriuose, o iš *22 pav*. galima pamatyti kokią įtaką modelio sudarymui darė nagrinėjami kintamieji. Galima teigti, kad klasterizavimas su SAS Enterprise Miner skiriasi nuo klasterizavimo su SPSS Modeler, tuo, kad klasteriai buvo sudaromi remiantis kitomis taisyklėmis, nes skiriasi kintamųjų įtaka modeliui nuo nagrinėto modelio su SPSS Modeler, galimai tai ir įtakojo, kad modeliuose skyrėsi klasterių skaičius.

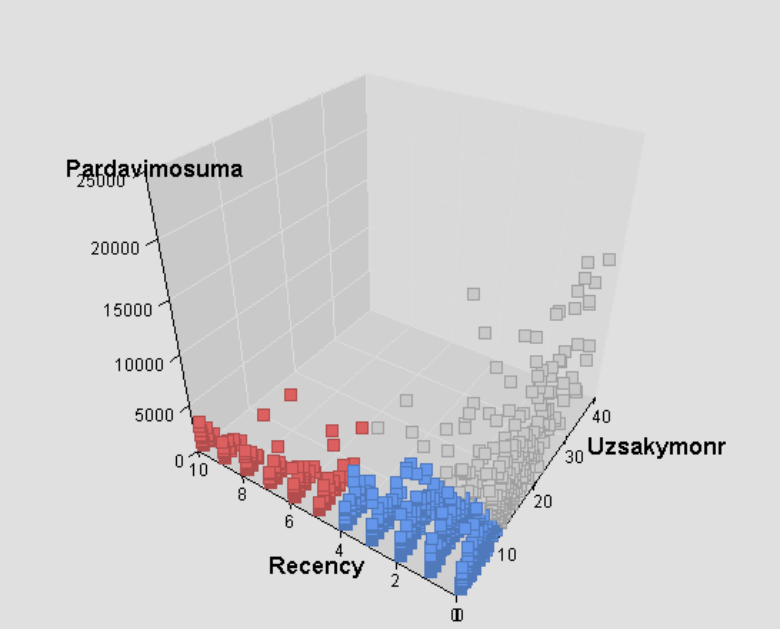


22 pav. Kintamųjų svarba modelio sudarymui

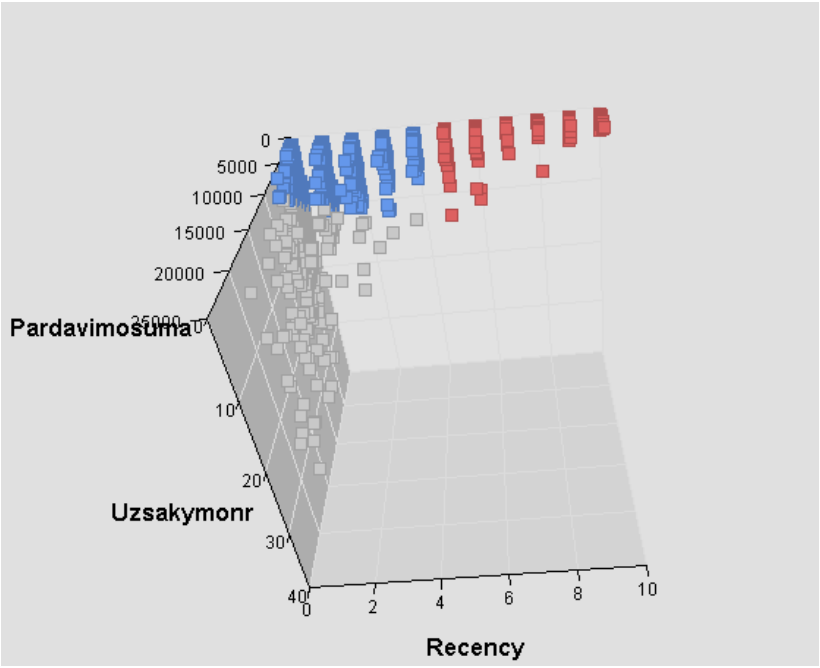
Output

```
81
82    Data Role=TRAIN
83
84    Data      Variable                              Frequency
85    Role        Name          Role       Level         Count     Percent
86
87    TRAIN     _SEGMENT_     SEGMENT       1            2291     69.3192
88    TRAIN     _SEGMENT_     SEGMENT       3             717     21.6944
89    TRAIN     _SEGMENT_     SEGMENT       2             297      8.9864
90
91
```

23 pav. Apmokymo imties klasterių sudėtis



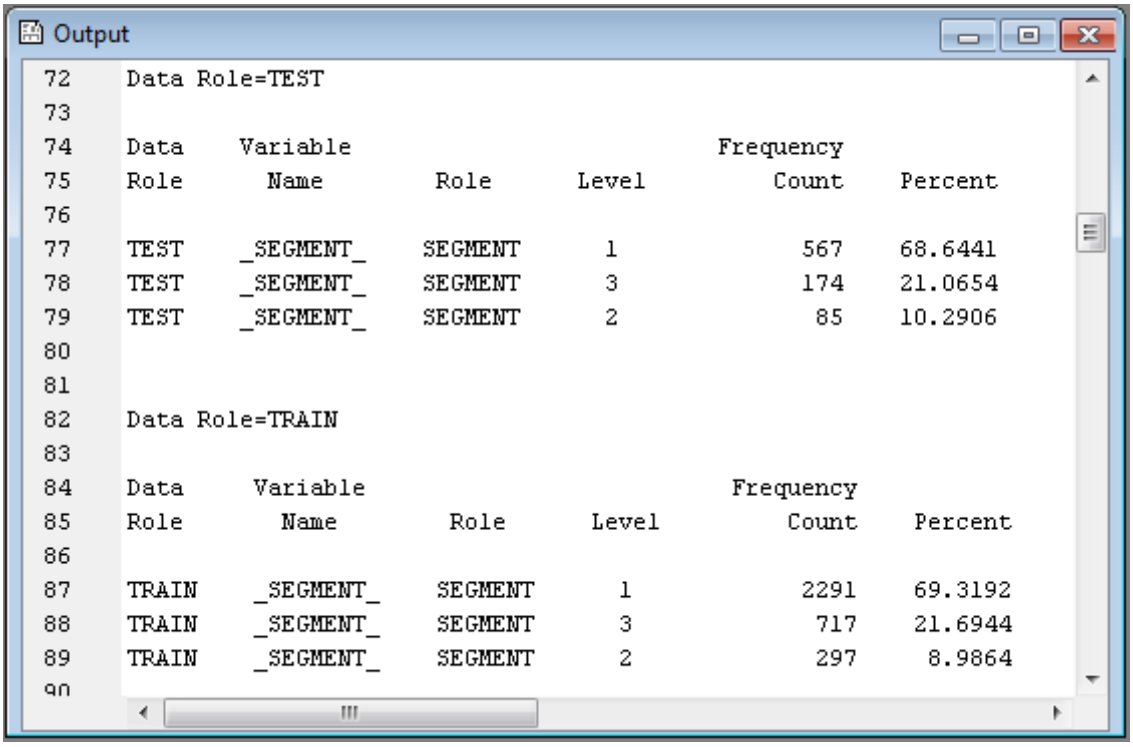
24 pav. Klasterių ir klientų klasteriuose išsidėstymo grafikas (1)



25 pav. Klasterių ir klientų klasteriuose išsidėstymo grafikas (1)

6 lent. Klasterių statistikos

	Minimali reikšmė	Vidurkis	Maksimali reikšmė
Cluster 1			
Recency	0	1,079	4
Frequency	1	2,893	10
Monetary	6,9	1014,251	7829,89
Cluster 2			
Recency	0	0,175	4
Frequency	3	13,037	35
Monetary	1303,04	5770,515	21279,29
Cluster 3			
Recency	5	7,159	10
Frequency	1	1,448	10
Monetary	2,9	433,3861	9864,26



26 pav. Apmokymo ir testavimo imties sudėtis

Prieš aptariant modelio rezultatus, reikia įvertinti modelio tinkamumą. Iš *26 pav.* matoma, kad klasterių proporcijos su apmokymo ir testavimo duomenimis atlikus klasterizavimą skiriasi nereikšmingai, todėl galima teigti, kad modelis teisingas.

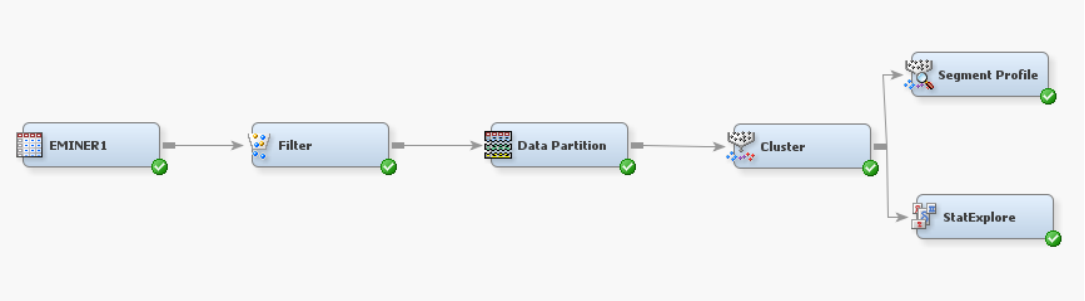
3.6.4. Rezultatai

Atsižvelgiant į *23, 24, 25 pav.* ir 3 lent. buvo trumpai apibendrinta klasterių sudėtis ir klientų pasiskirstymas klasteriuose.

Pirmąjį klasterį sudaro 2291 klientai arba 69,3% visos populiacijos, tai didžiausia klientų grupė. Šie klientai yra apsipirkę per paskutiniuosius keturis mėnesius ir atsižvelgiant į pardavimus tai yra nedažnai apsiperkantys klientai. Nagrinėjama klientų grupė sudaro daugiau nei pusę internetinės parduotuvės gaunamų pinigų sumos, todėl tai yra labai svarbi klientų grupė.

Antrasis klasteris susideda iš 297 klientų ir tai yra tik nepilni 9% visos populiacijos, tačiau tai pati svarbiausia lojalių klientų grupė, kurios klientai vidutiniškai per mėnesį bent kartą apsiperka internetinėje parduotuvėje. Šios grupės klientai yra daugiausiai pelno generuojantys klientai, vidutiniškai vienas klientas per nagrinėjamą laikotarpį išleido 5770,515$ ir nors tai yra tik 9% visų klientų, jie sudaro 40% visų internetinės parduotuvės gaunamų pajamų . Šių asmenų kaip klientų išlaikymas internetinei parduotuvei yra svarbiausia užduotis.

Trečiąjį klasterį sudaro 717 klientų, tai yra 21% visos populiacijos. Šie klientai paskutinį kartą apsipirko internetinėje parduotuvėje prieš 5 mėnesius arba seniau. Tai patys internetinei parduotuvei nepelningiausi klientai. Galimai tai atsitiktinai internetinėje parduotuvėje apsilankę ir apsipirkę, arba specifines prekes perkantys klientai. Ši klientų grupė internetinės parduotuvės savininkui galėtų būti įdomi, tik aiškinantis, kodėl klientai nustojo apsipirkinėti internetinėje parduotuvėje.



27 pav. Modelio realizacija SAS Enterprise Miner aplinkoje

3.7. Laiko eilučių analizė su SAS analitine platforma

Iš SAS anlitinės platformos įrankių, laiko eilutės analizei buvo pasirinktas SAS Enterprise Guide įrankis. SAS instituto darbuotojas nurodė, kad pats pažangiausias įrankis sudarant laiko eilutes yra SAS Forecast Studio, tačiau nebuvo galimybės pasinaudoti šia programine priemone. SAS Enterprise Guide programinė priemonė yra viena iš pažangiausių SAS platformos įrankių laiko eilučių analizei. Ši programinė priemonė taip pat yra valdoma mazgais, tačiau yra galimybė programą valdyti ir kodo pagalba.

Prieš pradedant modelio sudarymą, duomenys buvo padalinti į apmokymo ir testavimo imtis. Testavimo imtis susidaro iš paskutinių dviejų savaičių pardavimų, o likę duomenys priklauso apmokymo imčiai.

3.7.1. Modelio sudarymas

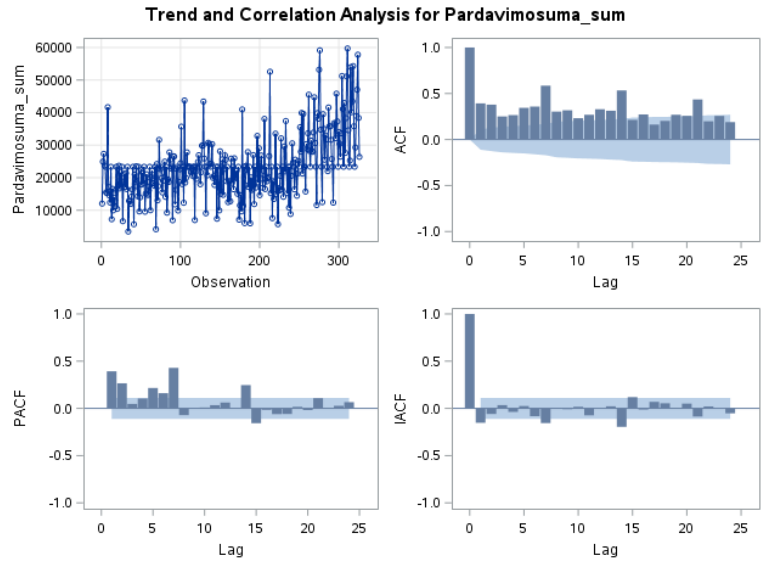
Norint atlikti laiko eilučių analizę reikia sudaryti laiko eilutę. Šią funkciją SAS Enterprise Guide programinėje priemonėje atlieka „Create Time Series Data“ mazgas. Mazge buvo sudaryta laiko eilutė ir trūkstami įrašai pakeisti laiko eilutės vidurkio reikšme.

Laiko eilučių analizei SAS Enterprise Guide yra integruoti keli mazgai:

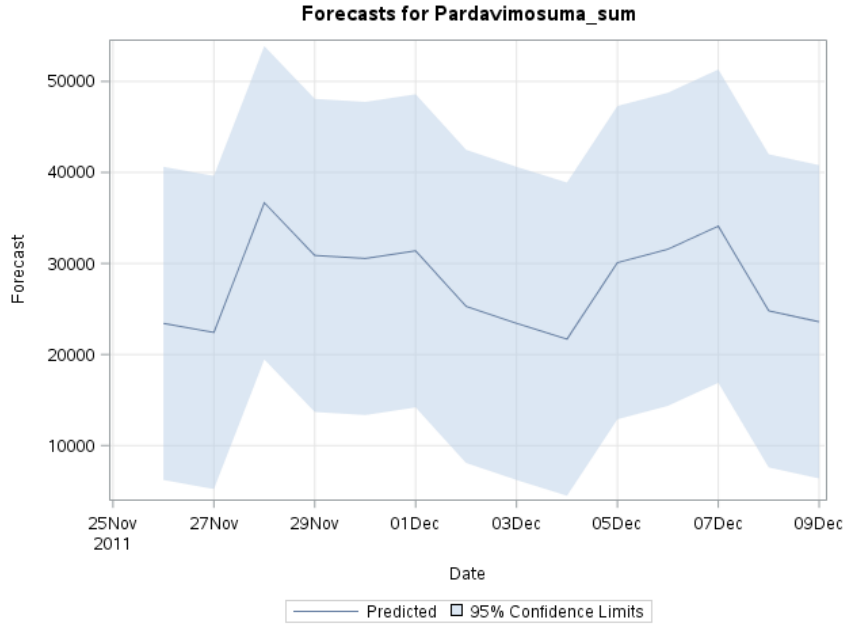
- Basic Forecasting
- ARIMA Modeling and Forecasting
- Regression Analysis

Pardavimų prognozavimui buvo pasirinktas „ARIMA Modeling and Forecasting“ mazgas.

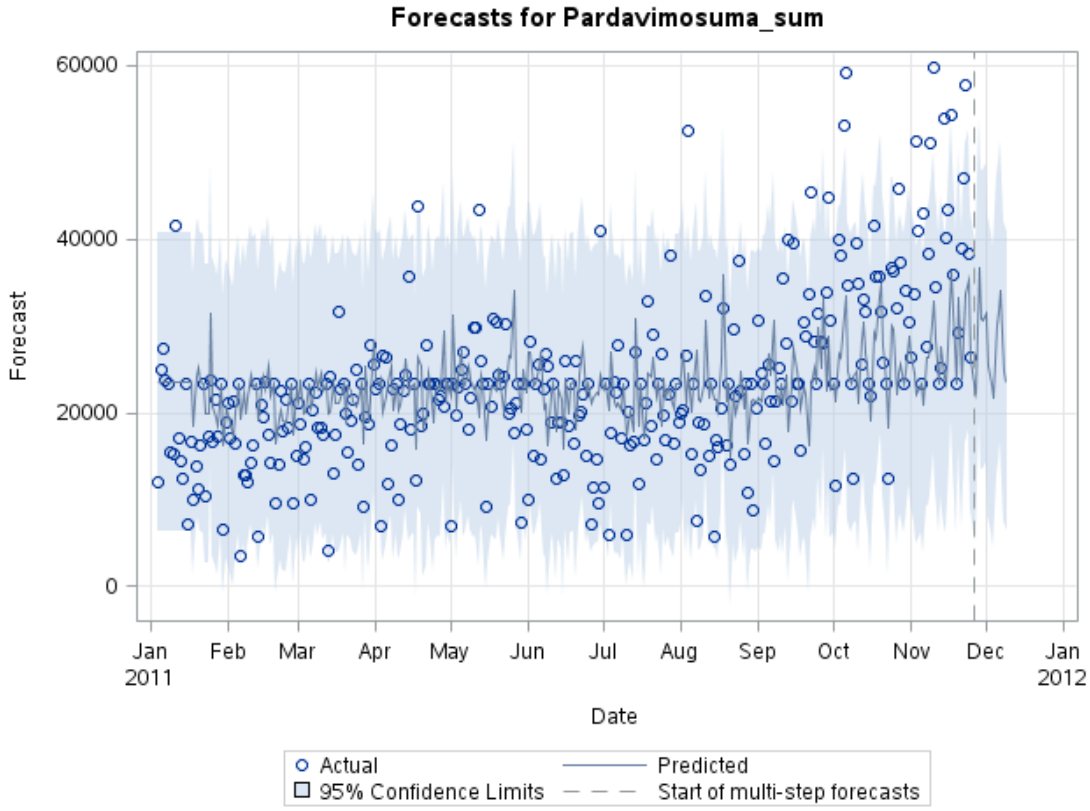
Sudarant laiko eilutės modelį, pirmiausia reikia patikrinti ar tai stacionari eilutė, jeigu eilutė nestacionari tuomet reikia atlikti diferencijavimą. SAS Enterprise Guide neturi specialaus įrankio laiko eilutės stacionarumui nustatyti, tačiau, iš sudaryto modelio su SPSS Modeler ir *28 pav.* pavaizduoto autokoreliacijos grafiko (ACF) galime teigti, kad laiko eilutė stacionari. Sudarant laiko eilutę su SAS Enterprise Guide reikia iš anksto nurodyti autoregresijos ir judančio vidurkio parametrus. Programinės priemonės aplinkoje nėra specialaus įrankio parametrams nustatyti, tačiau, iš tyrimo su SPSS Modeler žinome, kad autoregresijos parametras yra 0, o judančio vidurkio parametras yra 14.



28 pav. Tendencijos ir autokoreliacijos grafikai



29 pav. Pardavimo sumos prognozė

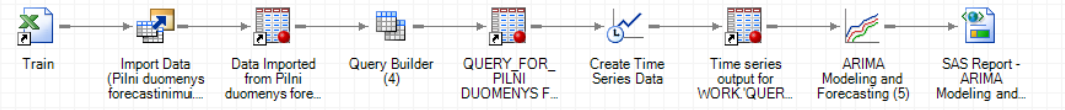


30 pav. Realios ir prognozuotos pardavimų sumos

3.7.2. Rezultatai

Programinė priemonė, tiksliau „Arima Modeling and Forecasting“ mazgas, turi daugybę integruotų statistikų, tokių kaip Akaike informacinis kriterijus, Shwartzo Bayeso kriterijus ir kt. Taip pat mazgas gali atvaizduoti autokoreliacijos grafikus, esamas bei prognozuotas reikšmes (*28, 29* ir *30 pav*.). Patikrinti modelio tikslumą, buvo palygintos prognozuotos reikšmes su testavimo

imties reikšmėmis. Palyginus prognozuotus įrašus su testavimo imties įrašais, buvo naudojamas apibrėžtumo koeficientas (angl. R-Squared), kuris buvo lygus 0.5. Prognozės tikslumas didesnis nei prognozuojant su SPSS Modeler. Taip yra todėl, kad nėra žinoma kokie įvertinimo kriterijai ir koks iteracijų skaičius buvo pasirinktas. Modelio tikslumas taip pat nėra didelis, tačiau tai galėtų padėti internetinei parduotuvei apsibrėžti tam tikras pinigų sumas, kurių galimai gali reikėti ateityje.



31 pav. Modelio realizacija SAS Enterprise Guide aplinkoje

3.8. Programinės priemonės apibendrinimas

7 lent. Bendrieji SAS kriterijai

	SAS
Dydis	+
Įvairumas	+
Integracijos	+
Valdoma kodu	+

8 lent. Klasterinės analizės naudojantis SAS Enterprise Miner kriterijai

SAS Enterprise Miner	Klasterinė analizė
Greitis	20 s.
Metodų skaičius	3
Išskirčių šalinimas	+
Skalės problema	+
Klasterių skaičiaus nustatymas	-
Klasterizavimo kokybės nustatymas	+
Vizualizacijos	+
Statistikos	+

9 lent. Laiko eilučių analizės naudojantis SAS Enterprise Guide kriterijai

SAS Enterprise Guide	Laiko eilučių analizė
Greitis	15 s.
Metodų skaičius	3
Stacionarumo nustatymas	3 testai
ARIMA parametrų radimas	-
Vizualizacijos	+
Statistikos	+

3.9. Klasterinė analizė naudojantis KNIME analitine platforma

Klasterinė analizė buvo atlikta naudojantis KNIME analytics platform. Ši programinė priemonė yra valdoma mazgais

3.9.1. Išskirčių radimas iš šalinimas

Šalinti išskirtis KNIME analitinėje platformoje yra integruotas „Outlier Removal“ mazgas. Šis mazgas suteikia galimybę pašalinti išskirtis dviem būdas. Pirmasis yra šalinti įrašus, kurie nutolę nuo duomenų pasiskirstymo per nurodytą standartinio nuokrypio vidurkio kiekį. Antrasis, naudotis „Boxplot“ grafiku. Kaip ir prieš tai atliktose analizėse išskirtis šalinsime tuo pačiu metodu, t.y. pašalinsime įrašus, kurie yra nutolę daugiau nei 7 standartinio nuokrypio vidurkiai.

3.9.2. Duomenų dalinimas į apmokymo ir testavimo imtis

Naudojantis analitinėje priemonėje integruotu mazgu „Partition“ duomenys buvo padalinti į apmokymo ir testavimo imtis. Apmokymo dalį sudaro 80% duomenų, o testavimo įmtį – 20%.

3.9.3. Skalės problema

Sprendžiant skalės problemą, KNIME analitinė platforma turi integruotą „Normalizer“ mazgą. Norint pašalinti skalės problemą duomenys buvo supresuoti į skalę nuo 0 iki 1.

3.9.4. Modelio sudarymas

KNIME analitinė platforma pateikia galimybes atlikti klasterinę analizę trimis metodais:

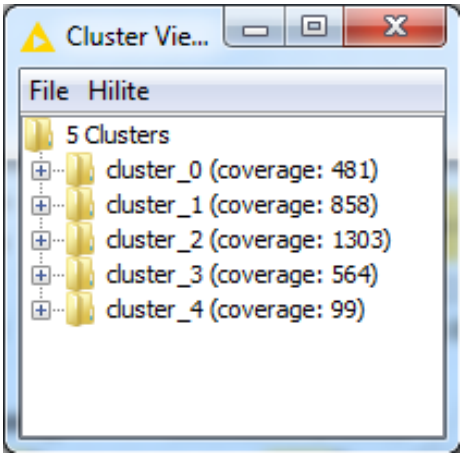
- K – vidurkių ;
- K-medoidų;
- Hierarchinio klasterizavimo.

Kaip buvo minėtą, analizei buvo pasirinktas k – vidurkių metodas, modelis buvo sudarytas naudojant apmokymo duomenis.

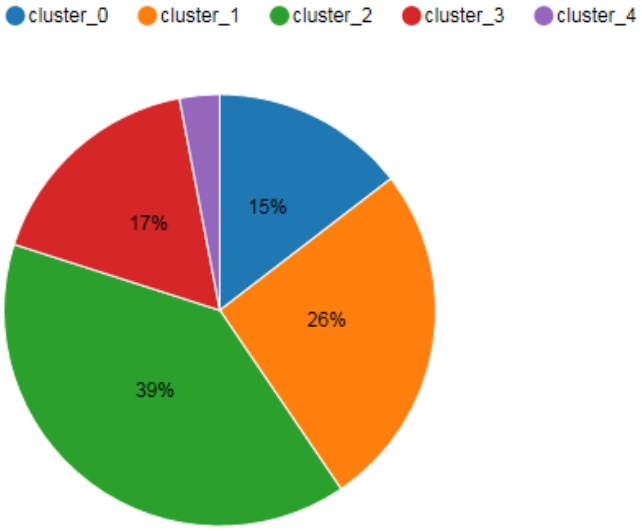
KNIME analitinė platforma neturi specialaus mazgo ar parametro optimaliam klasterių kiekiui nustatyti. Šiuo atveju klasterių skaičius buvo pasirinktas, norint gauti skirtingus rezultatus nei prieš tai atliktose analizėse. Nustačius klasterių skaičių 3 ir 4 buvo gauti beveik identiški rezultatai kaip sudarant modelį su IBM ir SAS platformomis, todėl analizei buvo pasirinktas 5 klasterių skaičius.

Klasterizavimą k-vidurkių metodu KNIME aplinkoje atlieka mazgas „k-Means“. Mazge galima pasirinkti klasterių skaičių. Kaip buvo minėta, buvo pasirinktas 5 klasterių modelis.

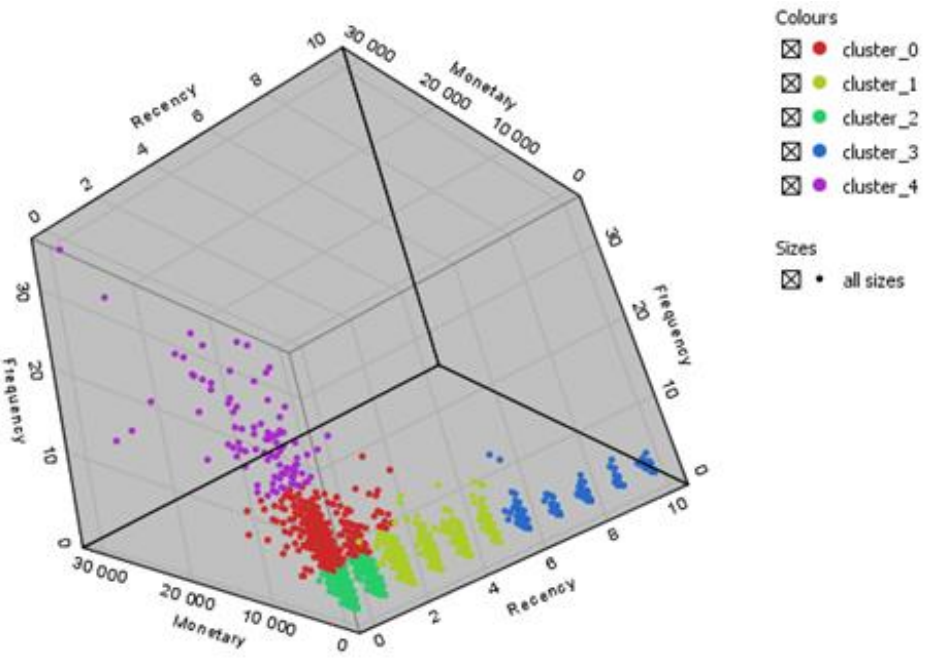
KNIME analitinė platforma turi daug vizualizacijos pasirinkimų. Klasterinei analizei geriausias būdas atvaizduoti klasterius ir įrašus klasteriuose yra taškinis grafikas. Programinė priemonė tam turi integruotą „2D/3D Scatterplot“ mazgą (*34* ir *35 pav*.)



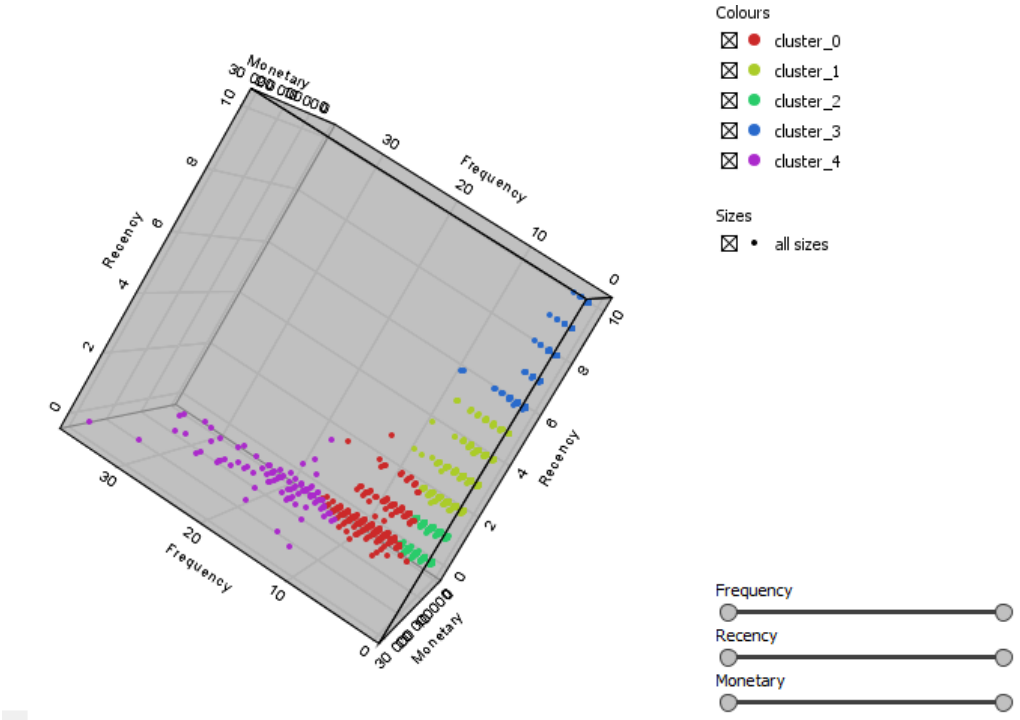
32 pav. Klientų skaičius klasteriuose skaičius



33 pav. Klasterių dydis



34 pav. Klasterių ir klientų klasteriuose pasiskirstymas (1)



35 pav. Klasterių ir klientų klasteriuose pasiskirstymas (2)

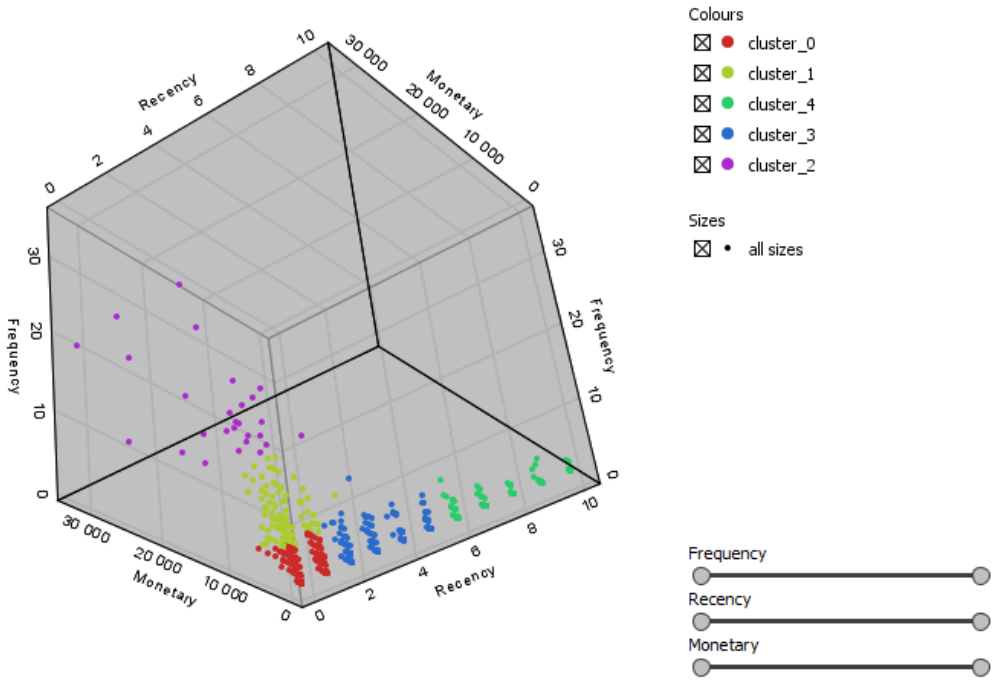
10 lent. Klasterių statistikos

	Minimali reikšmė	Vidurkis	Maksimali reikšmė
Cluster 1			
Recency	0	0,2536	3
Frequency	3	8,1393	16
Monetary	201,12	3181,7019	13 219,74

Cluster 2			
Recency	2	3,1503	5
Frequency	1	2,1993	9
Monetary	13,3	767,3556	6 617,65
Cluster 3			
Recency	0	0,3784	1
Frequency	1	2,3876	5
Monetary	6,2	823,5785	4 932,2
Cluster 4			
Recency	6	7,7358	10
Frequency	1	1,3191	9
Monetary	3,75	410,779	3 251,07
Cluster 5			
Recency	0	0,0505	2
Frequency	12	19,6162	35
Monetary	10151,2145	9042,8506	30 687,62

3.9.5. Modelio tikrinimas

Prieš klasterių apibendrinimą reikalingas modelio tinkamumo patikrinimas. Naudodami testavimo imties duomenis, su tokiais pačiais parametrais, buvo sudarytas modelis. Lyginant *36* ir *35 pav*. matoma, kad klasterių ir klientų pasiskirstymai reikšmingai nesiskiria, todėl galime teigti, kad modelis yra teisingas.



36 pav. Testavimo imties modelio sudaryti klasteriai ir klientų pasiskirstymas juose

3.9.6. Rezultatai

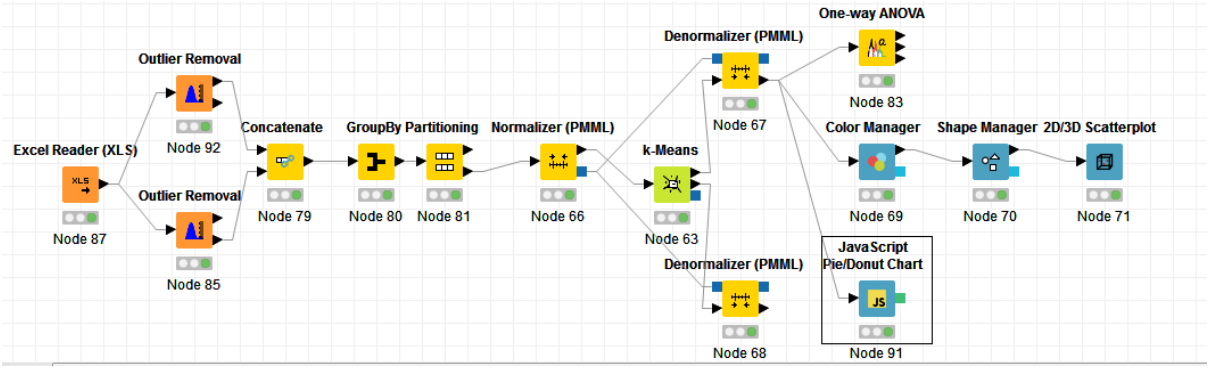
Atsižvelgiant į *32, 33, 34* ir *35 pav*. ir 4 lent. buvo apžvelgti gauti klasteriai ir klientų pasiskirstymas klasteriuose.

Pirmasis klasteris susideda iš 481 klientų ir tai sudaro 14,5% populiacijos. Iš kintamojo Recency (kaip seniai klientas pirko) matome, kad šios grupės klientai yra pirkę prekes iš internetinės parduotuvės per pastaruosius 3 mėnesius. Atsižvelgiant į pardavimus galima teigti, kad šie klientai yra labai dažnai šioje parduotuvėje apsiperkantys klientai ir šie klientai generuoja daugiau nei trečdalį internetinės parduotuvės gaunamų pajamų. Tai viena iš svarbiausių klientų grupių, todėl jos išlaikymas internetinėje parduotuvėje turėtų būti vienas iš prioritetų.

Antrasis ir trečiasis klasteriai yra panašūs. Juos sudaro atitinkamai 858 ir 1303 klientai, kas atitinka 26% ir 39,5% populiacijos. Šiuose klasteriuose esantys klientai yra apsipirkę per pastaruosius 5 mėnesius, tačiau tai retai apsiperkantys klientai. Negalima teigti, kad tai nereikšmingos grupės, nes kartu sudėjus šios klientų grupės generuoja daugiau nei trečdalį internetinės parduotuvės pajamų, tačiau šie klientai, atsižvelgiant į jų apsipirkimo dažnį, galimai parduotuvėje perka tik tam tikras specifines prekes.

Ketvirtoji grupė susideda iš 564 klientų ir tai sudaro 17% visos populiacijos. Ši klientų nėra pirkusi prekių per pastarąjį pusmetį. Šie klientai vidutiniškai per metus nagrinėjamoje internetinėje parduotuvėje apsipirko tik 1,3191 karto. Galima daryti hipotezę, kad tai klientai, kurie ieškojo tik tam tikros prekės ir ją nusipirko arba klientai pirkę internetinėje parduotuvėje, tačiau yra nusivylę arba pristatymu arba prekės kokybe. Galima sakyti, kad ši klientų grupė yra nereikšminga, nes ji sudaro tik šiek tiek virš 5% visų gautinų pajamų per nagrinėjamą laikotarpį, tačiau išsiaiškinus, kodėl klientai nustojo pirkti šioje parduotuvėje, ateityje būtų galima daryti išvadas kaip išlaikyti klientus lojaliais.

Penktoji klientų grupė susideda iš 99 klientų ir tai sudaro tik 3% visos populiacijos. Nors ši grupė sudaro tik 3% populiacijos, tačiau jos generuojamas pelnas sudaro virš 20% viso pelno. Šiuos klientus galima traktuoti kaip nuolatinius, nes jie galimai dažnai apsipirkinėjo per visą nagrinėjamą laikotarpį. Tai patys internetinei parduotuvei pelningiausi klientai ir jų išlaikymas internetinės parduotuvės savininkui turėtų būti viena iš svarbiausių užduočių.



37 pav. modelio realizavimas KNIME analitinėje priemonėje

3.10. Laiko eilučių analizė su KNIME analitine platforma

Prieš pradedant modelio sudarymą, duomenys buvo padalinti į apmokymo ir testavimo imtis. Testavimo imtį sudaro paskutinių dviejų savaičių pardavimai, o likę duomenys priklauso apmokymo imčiai.

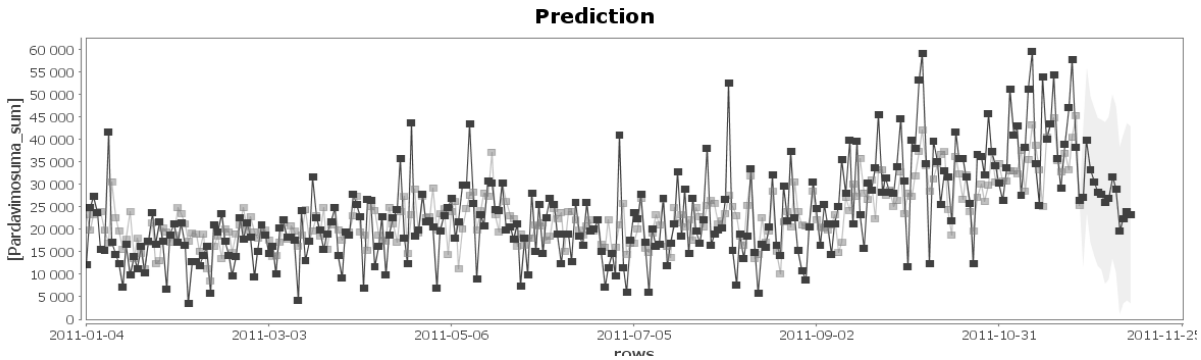
3.10.1. Modelio sudarymas

Laiko eilučių analizei KNIME yra integruoti keli mazgai:

- ARIMA Learner
- Regression Predictor

Pardavimų prognozavimui buvo pasirinktas „ARIMA Learner“ mazgas.

Sudarant laiko eilutės modelį, pirmiausia reikia patikrinti ar tai stacionari eilutė, jeigu eilutė nestacionari tuomet reikia atlikti diferencijavimą. Programinė priemonė neturi stacionarumui nustatyti integruoto įrankio, tačiau iš sudaryto modelio su SPSS Modeler galima teigti, kad laiko eilutė stacionari. Sudarant laiko eilutę su KNIME reikia iš anksto nurodyti autoregresijos ir judančio vidurkio parametrus. Iš tyrimo su SPSS Modeler žinome, kad autoregresijos parametras yra 0, o judančio vidurkio parametras 14.



38 pav. Realios ir prognozuotos pardavimų sumos

3.10.2. Rezultatai

Programinė priemonė neturėjo patikimų apibrėžtumo koeficiento ar kitų statistikų apskaičiavimo galimybių. Tikrinant prognozės tikslumą buvo panaudota kita programinė priemonė. Kaip ir prieš tai nagrinėtais atvejais buvo palygintos prognozuotos reikšmės su testavimo imties reikšmėmis. Palyginus prognozuotus įrašus su testavimo imties įrašais apibrėžtumo koeficientas (angl. R-Squared) yra lygus 0.31. Prognozės tikslumas mažesnis nei prognozuojant su SPSS Modeler ir SAS Enterprise Guide. Taip yra todėl, kad programinė priemonė neturėjo įrankio laiko intervalams sudaryti, dėl to tuščios reikšmės tarp datų buvo traktuojamos kaip tuščios vietos su nulinėmis reikšmėmis, taip pat nebuvo nurodyti apskaičiavimo metodai.

3.11. Programinės priemonės apibendrinimas

11 lent. Bendrieji KNIME kriterijai

KNIME	
Dydis	+
Įvairumas	+
Integracijos	+
Valdoma kodu	-

12 lent. Klasterinės analizės naudojantis KNIME kriterijai

KNIME	Klasterinė analizė
Greitis	7 s.
Metodų skaičius	3
Išskirčių šalinimas	+
Skalės problema	+
Klasterių skaičiaus nustatymas	-
Klasterizavimo kokybės nustatymas	-
Vizualizacijos	+
Statistikos	+

13 lent. Laiko eilučių analizės naudojantis KNIME kriterijai

KNIME	Laiko eilučių analizė
Greitis	4 s.
Metodų skaičius	2
Stacionarumo nustatymas	-
ARIMA parametrų radimas	-
Vizualizacijos	+
Statistikos	-

3.12. Programinių priemonių apibendrinimas.

Visos programinės priemonės gali apdoroti didelius duomenų masyvus, kas yra labai svarbu šiuo dienų duomenų apimčių augimui, ypač mažmeninės prekybos duomenims. Priemonės gali analizuoti įvairių tipų duomenų failus, o šių dienų duomenys yra talpinami įvairių tipų failuose, taip programinės priemonės tampa įvairiapusiškesnės. Kitų programinių priemonių integracijos paprastiems vadybininkams, atliekantiems primityvias analizes gal ir nereikalingas, tačiau integracijos yra labai svarbios duomenų mokslininkams dėl tam tikrų įrankių nagrinėjamose priemonėse nebūvimo, dėl integracijų programinės priemonės tampa funkcionalesnės. Vienintelė programinė priemonė SAS yra valdoma tiek mazgais, tiek kodu, tai programinei priemonei suteikia įvairiapusiškumo, t.y. galimybę integruoti ir koreguoti metodus, parametrus ar funkcijas, kurios neegzistuoja mazguose.

Agreguojant duomenis, buvo nustatyti programinių priemonių greičiai transformuojant sąlyginai didelį duomenų masyvą į mažesnį. Programinė priemonė SPSS Modeler atliko ilgiausią transformavimą, galima teigti, kad SPSS Modeler nėra pritaikyta analizėms su didžiaisiais duomenimis. Nagrinėjamas duomenų masyvas yra tik sąlyginai didelis, nes nagrinėjama mažmeninės prekybos internetinė parduotuvė generuoja sąlyginai mažus pardavimus, todėl nagrinėjant didesnes apimtis turinčią mažmeninės prekybos parduotuvę SPSS Modeler nesugebėtų greitai apdoroti informacijos, kas šių dienų prekybai yra labai svarbu. Šiuo aspektu, derėtų rinktis SAS arba KNIME analitinę platformą.

Metodų atžvilgiu, visos programinės priemonės atliekant klasterinę ir laiko eilučių analizes turi panašų skaičių metodų, tačiau metodų atlikimo galimybės stipriai skiriasi.

Skalės ir išskirčių radimo ir šalinimo problemas gali išspręsti visos programinės priemonės.

Klasterių skaičiaus nustatymo galimybė egzistuoja tik SAS Enterprise Miner analitinėje platformoje naudojant kubinio klasterizavimo kriterijų, tačiau naudojantis SPSS Modeler, klasterių skaičių galima nustatyti tikrinant klasterizavimo kokybę naudojantis Silueto koeficientą, šios funkcijos SAS Enterprise Miner neturi. KNIME analitinė platformoje nėra integruota nei klasterio skaičiaus, nei klasterizavimo kokybės nustatymo funkcijos.

Vizualizacijos galimybės, analitinėse priemonėse, atliekant klasterinę analizę reikšmingai nesiskiria. Nors programinėse priemonėse egzistuoja daugybė vizualizacijos galimybių, tačiau klasterinei analizei reikalingiausias yra taškinis grafikas, kuris geriausiai atvaizduoja klasterių pasiskirstymą ir klientų pasiskirstymą klasteriuose. Taip pat kiekviena programinė priemonė atvaizdavo įvairias statistikas, kaip pvz. klasterių svarbą modeliams, klientų pasiskirstymą klasteriuose ir kt.

Atliekant laiko eilučių analizę pirmiausia reikia išsiaiškinti laiko eilutės stacionarumą. Vienintelė programinė priemonė SPSS Modeler gali automatiškai įvertinti laiko eilutės stacionarumą, tačiau SAS Enterprise Guide suteikia galimybę atlikti stacionarumo testus, nubrėžti autokoreliacijos grafikus, iš kurių galima nuspręsti laiko eilutės stacionarumą. KNIME programinė priemonė neturi galimybių patikrinti laiko eilutės stacionarumą. Atliekant laiko eilučių analizę naudojantis ARIMA modeliu reikalingas parametrų p,d,q nustatymas. SPSS Modeler parametrus nustato automatiškai parinkus ekspertinį ARIMA modeliavimą. SAS Enterprise Guide suteikia galimybę parametrus nustatyti arba iš autokoreliacijos grafikų, arba keičiant parametrus ieškant geriausio modelio tikslumo. KNIME programinėje priemonėje nustatant tinkamus ARIMA suteikia vienintelę galimybę - keičiant parametrus ieškoti geriausio modelio tikslumo.

Visos programinės priemonės turėjo reikiamas vizualizacijas atvaizduojant tiek esamą, tiek prognozuotą laiko eilutę, tačiau grafiškai palyginti jų SPSS Modeler nesuteikė galimybės.

SPSS Modeler ir SAS Enterprise Guide turėjo galimybę atvaizduoti įvairias statistikas. Modelio tinkamumui – apibrėžtumo koeficientas, modelių palyginimui – Akaikė informacijos kriterijus, vidutinė kvadratinė paklaida ir t.t. KNIME programinė priemonė neturi patikimų galimybių atvaizduoti modelio tinkamumui ir palyginimui skirtų statistikų.

Galima daryti išvadą, kad naujam, besikuriančiam mažmeninės prekybos verslui, turint nedidelį duomenų kiekį ir neprofesionalius darbuotojus, perkant programinę priemonę, būtų patartina rinktis SPSS Modeler. SPSS Modeler yra nesunkiai suprantama, pažangi, automatizuota programinė priemonė, kuri suteikia reikiamas vizualizacijas ir statistikas. Greičio trūkumas yra pagrindinis programinės priemonės minusas. Atsižvelgiant į visus pliusus ir minusus, galima teigti, kad SAS analitinė platforma yra pažangiausia priemonė nagrinėjant mažmeninės prekybos duomenis. Esant didesniam verslui, samdantis duomenų mokslininkus, reikėtų rinktis SAS analitinę platformą, dėl greitos analizės, pažangių įrankių turinčių daugybę modelių, statistikų ir vizualizacijos pasirinkimų. Taip pat programinės priemonės didelis pliusas yra jos valdymas kodu. KNIME programinė priemonė negalėtų būti naudojama be kitos programinės priemonės integracijos. Nors KNIME yra greita, tačiau jaučiamas įrankių trūkumas. Norint naudotis programinės priemonės integracijomis reikalingas duomenų moksle pažengusio vartotojo, o mažmeninės prekybos vadovai ne visuomet ieško patyrusių duomenų specialistų.

IŠVADOS

Atlikus literatūros apžvalgą, buvo išanalizuotos šių dienų didžiųjų duomenų problemos. Viena iš įmonių, generuojančių didžiulius duomenų masyvus problemų yra programinės įrangos pasirinkimas, todėl darbe buvo apžvelgtos pažangiausios duomenų mokslui skirtos analitinės priemonės. Atsižvelgiant Gartner, Inc kompanijos, kuri verčiasi informacinių technologijų tyrimais, sudarytu reitingu, buvo pasirinktos trys pažangios analitinės priemonės analizėms atlikti.

Naudojantis IBM, SAS ir KNIME analitinėmis platformomis tyrime buvo atlikta klasterinė ir laiko eilučių analizė orientuota į mažmeninės prekybos duomenis. Atliekant klasterinę analizę klientai buvo susegmentuoti į tam tikras reikšmingas grupes, taip atskleidžiant tam tikrą klientų elgesį. Laiko eilučių analizės metu buvo prognozuotos pardavimų sumos į ateitį.

Atliekant klasterinę ir laiko eilučių analizę nagrinėjant mažmeninės prekybos duomenis buvo stebimos pasirinktų analitinių priemonių galimybės. Apžvelgus visus analitinių priemonių pliusus ir minusus buvo prieita išvados, kad SAS analitinė platforma yra pažangesnė ir geriausiai atitinka šių dienų reikalavimams.

LITERATŪROS SĄRAŠAS

[1] KAULESHWAR, P., R. ARPANA. Statistical Survey on Big Data Analytics. Computer and IT Sci.. Research Journal of Computer and Information Technology Sciences, 2016, vol. 4(9), 22-24. E-ISSN 2320 – 6527.

[2] LANEY D. 3D Managment: Controlling Data Volume, Velocity, and Variety [interaktyvus]. Stamford, META Group, 2001. Prieiga per: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[3] OHBYUNG K, L. NAMYEON, S. BONGSIK . Data quality management, data usage experience and acquisition intention of big data analytics. International Journal of Information Managemen, 2014, vol. 34, 387–394.

[4] HSINCHUN Ch, ROGER H. L. CHIANG, VEDA C. Storey. Business intelligence and analytics: from big data to big impact. U.S.A: University of Arizona, University of Cincinnati, Georgia State University, 2012.

[5] Big Data [interaktyvus]. 2016. Prieiga per: http://www.gartner.com/it-glossary/big-data/

[6] TechAmerica Foundation. Demystifying big data: defining big data & buiseness mission [interaktyvus]. Prieiga per: https://www.attain.com/sites/default/files/take-aways-pdf/Solutions_Demystifying%20Big%20Data%20-%20A%20Practical%20Guide%20To%20Transforming%20The%20Business%20Of%20Government.pdf

[7] GANDOMI A, HAIDER M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management [interaktyvus]. Toronto: Ted Rogers School of Management, Ryerson University, 2015, 35, 137–144. Prieiga per: Science Direct.

[8] Gartner [interaktyvus]. 2017. Prieiga per: http://www.gartner.com/technology/home.jsp

[9] IBM internetinis svetainė [interaktyvus]. Prieiga per: https://www.ibm.com/us-en/

[10] SAS internetinė svetainė [interaktyvus]. https://www.sas.com/en_us/home.html

[11] KNIME internetinė svetainė [interaktyvus]. https://www.knime.org/

[12] WEDER M. and W.KAMAKURA. MARKET SEGMENTATION: Conceptual and Methodological Foundations Second Edition. Kluwer Academic Publishers, 2000, ISBM 978-1-4613-7104-5.

[13] KOTLER PH. Marketing Managment, Millenium Edition: Tenth Edition. Prentice-Hall,Inc, 2000.

[14] MCDONALD M., I. DUNBAR, Market Segmentation: How to Do It, How to Profit From It, Goodfellow Publishers Ltd., 1995.

[15] STEENKAMP J.B.E.M., F.T. Hofstede, International market segmentation: issues and perspectives, Intern. J. Res. Mark. 19 (2002) 185–213.

[16] KAMAKURA W.A., G.J.A. RUSSELL, Probabilistic choice model for market segmentation and elasticity structure, J. Mark. Res. 26 (1989) 379–390.

[17] KUMAR V. and WERNER J. R. Customer Relationship Managment A Databased Approach: instructor‘s Presentation Slides [interaktyvus]. Prieiga per: http://web.nchu.edu.tw/~jodytsao/CRM/40471825-CRM-A-Database-Approach-Kumar-Reinartz-Ch05.pdf

[18] LARSEN N, Market Segmentation. Aarhus School of Business, 2010. Prieiga per: http://pure.au.dk/portal/files/11462/ba.pdf

[19] RFM Analysis For Successfull Customer Segmentation [interaktyvus]. Prieiga per: http://www.putler.com/rfm-analysis/

[20] ROUSE M. RFM analysis (recency, frequency, monetary)[interaktyvus]. Prieiga per: http://searchdatamanagement.techtarget.com/definition/RFM-analysis

[21] Retail Customer Analysis: Predictive Customer Lifetime Value analysis [interaktyvus]. Prieiga prie: https://custora.com/tour/feature_predictive_customer_lifetime_value_clv_retail/

[22] BIRANT D. Data Mining Using RFM Analysis. InTech China, 2011, ISBN: 978-953-307-154-1.

[23] GHOSE S. Sales Forecasting: Meaning, Factors, imortance and Limitations [interakyvus] prieiga per: http://www.yourarticlelibrary.com/sales/sales-forecasting-meaning-factors-importance-and-limitations/50997/

[24] HOFFMAN J. Sales Forecasts: A Question of Method, Not Magic [interaktyvus]. Prieiga per: http://labs.openviewpartners.com/files/2012/07/Forecasting-eBook-FINAL.pdf

[25] WEEDMARK D. Objectives of sales forecasting [interaktyvus]. Prieiga per: http://smallbusiness.chron.com/objectives-sales-forecasting-60407.html

[26] LEEDS B. University. Forecast and plan you sales. Prieiga per: http://www.leedsbeckett.ac.uk/assets/docs/employability/resourcecentre/How%20to%20forecast%20and%20plan%20your%20sales.doc

[27] THOMAS H. Realizing the Potential of Retail Analytics [interaktyvus]. 2009. Prieiga per: http://analytics.typepad.com/files/retailanalytics.pdf

[28] ČEKANAVIČIUS V., MURAUSKAS G. Statistika ir jos taikymai II: vadovėlis. V.:TEV,2004. ISBN 995 – 491 -16 -7

[29] JAIN K. A. and DUBES C. R. Algorithms for Clustering Data. Englewood, New Jersey, 1988, ISBN 0-13-022278-X.

[30] HANS H. B. Probabilistic models in cluster analysis. Computational Statistics and Data Analysis, 1996, 23, 5-28.

[31] K-means clustering algorithm [interaktyvus]. Prieiga per: https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

[32] Cluste Analysis: Basic Concepts and Algorithms [interaktyvus]. Prieiga per: https://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

[33] MADSEN H. Time Series Analysis. London, 2008, ISBN 978-1-4200-5967-0.

[34] STABINGIENĖ L. Laiko eilučių analize: laiko eilučių samprata [interaktyvus]. 2014. Prieiga per: http://www.ilab.lt/stabingiene/sk8_2.html

[35] NAU Robert Staistical forecasting: Introduction to ARIMA [interaktyvus]. Fuqua School of Business Duke University. 2017. Prieiga per: https://people.duke.edu/~rnau/411arim.htm

[36] NAU Robert Staistical forecasting: Seasonal differencing in ARIMA models [interaktyvus]. Fuqua School of Business Duke University. 2017. Prieiga per: https://people.duke.edu/~rnau/411sdif.htm

[37] The Pennsylvania State University. The Cofficient of Determination r-squared [interaktyvus]. 2017. Prieiga per: https://onlinecourses.science.psu.edu/stat501/node/255

[38] MORKEVIČIUS V. Statistinė kiekybinių duomenų analizė su spss ir stata: statistinės analizės pavyzdžių naudojant pavyzdinę skaitmeninę duomenų bazę medžiaga [interaktyvus]. Prieiga per: http://www.lidata.eu/index.php?file=files/mokymai/stat/stat.html&course_file=stat_III_8_1_1.html

[39] IT Central Station [interaktyvus]. 2017. Prieiga per: https://www.itcentralstation.com/products/comparisons/ibm-spss-modeler_vs_ibm-spss-statistics