Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų tyrimo modelis

Baigiamasis magistro studijų projektas

Kotryna Nazarovaitė

Projekto autorė

Doc. Dr. Vytautas Janilionis

Vadovas

Doc. Dr. Aušra Rūtelionė

Vadovė

Kaunas, 2020

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų tyrimo modelis

Baigiamasis magistro studijų projektas

Didžiųjų verslo duomenų analitika (6213AX001)

Kotryna Nazarovaitė

Projekto autorė

Doc. Dr. Vytautas Janilionis

Vadovas

Doc. Dr. Aušra Rūtelionė

Vadovė

Doc. Dr. Tomas Ruzgas

Recenzentas

Doc. Dr. Beata Šeinauskienė

Recenzentė

Kaunas, 2020

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Kotryna Nazarovaitė

Apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų tyrimo modelis

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Kotrynos Nazarovaitės, baigiamasis projektas tema „Apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų tyrimo modelis“ 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.

Kotryna Nazarovaitė	
(vardą ir pavardę įrašyti ranka)	(parašas)

Nazarovaitė, Kotryna. Apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų tyrimo modelis. Magistro baigiamasis projektas.

Vadovas doc. dr. Vytautas Janilionis; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Vadovė doc. dr. Aušra Rūtelionė; Kauno technologijos universitetas, Ekonomikos ir verslo fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): pagrindinė kryptis – Taikomoji matematika (Matematikos mokslai), papildančiosios kryptys – Ekonomika (Socialiniai mokslai) ir Informatika (Informatikos mokslai)

Reikšminiai žodžiai: apgyvendinimo paslaugos, klientų patirtis ir pasitenkinimas, teksto tyryba, sentimentų analizė, faktorinė analizė, regresinė analizė

Kaunas, 2020. 78 p.

Santrauka

Darbo tyrimo objektas – apgyvendinimo paslaugų klientų atsiliepimų ir įvertinimų analizė. Pirmoje darbo dalyje atlikta mokslinės literatūros analizė bei praktika rodo, kad apgyvendinimo paslaugų klientų internetinių atsiliepimų reikšmė jų patirties tyrimui yra plačiai pripažinta įvairiuose literatūros šaltiniuose. Ankstesniuose tyrimuose daugiausia analizuota klientų patirtis arba pasitenkinimas, tačiau mažai tirtas jų ryšys. Daugelis mokslininkų teigia, jog vienas iš pagrindinių privalumų tiriant apgyvendinimo paslaugų klientų atsiliepimus ir įvertinimus yra tai, jog jie gali tiesiogiai atskleisti jų patirtį ir pasitenkinimą.

Išnagrinėjus mokslinę literatūrą, nustatyta pagrindinė darbo problema – kaip privataus būsto nuomos klientų patirties pobūdis, atskleidžiamas jų atsiliepimuose, gali būti panaudojamas jų pasitenkinimui identifikuoti? Antroje darbo dalyje sudaryta tyrimo metodika, siekiant sukurti modelį, kuris įvertintų sąsajas tarp apgyvendinimo paslaugas teikiančių vietų klientų patirties ir pasitenkinimo. Metodika realizuota programiškai panaudojus programines įrangas Python ir SAS. Šios priemonės automatizuoja apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų modelių sudarymą ir tyrimą.

Sukurtos priemonės pritaikytos realių duomenų analizei parodė, jog jos išsprendžia darbe suformuluotus uždavinius. Trečioje darbo dalyje tiriamos dvi privataus būsto nuomos atvejo analizės – Berlyno ir Miuncheno miestų bei Madrido ir Barselonos miestų. Berlyno ir Miuncheno miestų atveju išskirti 19 požymių, o remiantis sudarytu tiesinės regresijos modeliu gauta, kad stipriausią įtaką bendram būsto vertinimui daro naujai pasiūlytas požymis – šeimininko statusas. Modelis paaiškina 28,1 % bendro būsto vertinimo sklaidos apie vidurkį tiesine regresija išskirtų požymių atžvilgiu. Madrido ir Barselonos miestų atveju gautas modelis, kuris paaiškina 39,6 % bendro būsto vertinimo sklaidos apie vidurkį tiesine regresija išskirtų požymių atžvilgiu, o stipriausią įtaką bendram būsto vertinimui daro šeimininko statuso ir miesto požymiai.

Nazarovaite, Kotryna. A model for investigation relationship between accommodation services customers experience and satisfaction. Master's Final Degree Project.

Supervisor doc. Dr. Vytautas Janilionis; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Supervisor doc. Dr. Aušra Rūtelionė; Faculty of Economics and Business, Kaunas University of Technology.

Study field and area (study field group): main field – Applied Mathematics (Mathematics sciences), additional fields – Economics (Social sciences) and Informatics (Informatics sciences)

Keywords: accommodation services, customers experience and satisfaction, text mining, sentiment analysis, factor analysis, regression analysis

Kaunas, 2020. 78 p.

Summary

The main object of the thesis is accommodation services customers' comments and ratings analysis. Many studies and practical approaches analysis was performed in the first chapter of the thesis which show that accommodation services customers' internet reviews are highly recognized in various researches. In previous studies it was analysed either the customers' experience or the satisfaction, but the relationship between them was investigated rarely. Many researchers suggest that one of the main advantages in investigating customers' reviews and ratings is that they can directly reveal customers' experience and satisfaction.

After the analysis of previous studies, main problem of the thesis is identified – how could private sector accommodation customers' experience lying in their reviews be used to identify their satisfaction? Research technique is created in the second chapter of the thesis in order to build up the model for investigation relationship between accommodation services customers' experience and satisfaction. Created technique is implemented using Python and SAS software.

The research technique and tool were applied on real data and it showed that they solve tasks of the thesis. In the third chapter of the thesis the information from Berlin and Munich cities as well as Madrid and Barcelona cities was examined. 19 attributes were obtained in the case of Berlin and Munich and the regression analysis showed that newly proposed attribute host status has the main influence on customers' satisfaction. The obtained linear regression model explain 28,1 % of accommodation rating dispersion on the mean in respect of obtained attributes. In the case of Madrid and Barcelona, the linear regression model explain 39,6 % of accommodation rating dispersion on the mean in respect of obtained attributes, while the main attributes on customers' satisfaction are host status and city.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Įžanga

Darbo problema ir temos aktualumas. Klientų patirties ir pasitenkinimo valdymas yra svarbi paslaugų valdymo bei rinkodaros dalis, kuri lemia pakartotinį klientų pirkimą, rekomendacijas bei pardavimus [4]. Pastaruoju metu turizmo ir svetingumo šakose atsirandant vis daugiau dalijimosi ekonomikos platformų, stipriai pasikeitė apgyvendinimo vietos pasirinkimas. Anot Xu ir kt. [4], „Airbnb“ – kaip dalijimosi pramonės pradininkas – greitai įtvirtino internetinės apgyvendinimo paslaugų platformos lyderio poziciją bei tapo rimtu tradicinių viešbučių konkurentu. Dėl šios priežasties, pagrindiniai svečių patirties bei pasitenkinimo faktoriai gali padėti apgyvendinimo paslaugų teikėjams atpažinti klientų poreikius ir pagerinti paslaugų kokybę. Tai taip pat gali padėti tradiciniams viešbučiams identifikuoti konkurentų konkurencingumo pranašumą.

Pasak Xu ir kt. [4], kurie rėmėsi autoriais Callan‘u ir Bowman‘u, Knutson‘u, Qu, Ryan‘u ir Chu, Rhee ir Yang‘u, Shanka ir Taylor‘u bei Schmitt‘u, nors nemažai tyrimų orientavosi į patirties bei pasitenkinimo nagrinėjimą tradicinių viešbučių kontekste, pastaruoju metu itin išaugo susidomėjimas dalijimosi ekonomikos šakos, t. y., apgyvendinimo paslaugų, tyrimais. Ankstesniuose tyrimuose nagrinėtos dalijimosi ekonomikos srities – apgyvendinimo – problemos daugiausia orientavosi į klientų motyvaciją, patirtis, pasitenkinimą bei pasitikėjimą tarp šeimininkų ir svečių. Tokie tyrimai telkė dėmesį arba į patirtį, arba į pasitenkinimo pobūdį, tačiau neanalizavo ryšio tarp apgyvendinimo patirties bei klientų pasitenkinimo. Taigi, iki šiol atliktų tyrimų trūkumas yra tai, kad juose nebuvo bandoma paaiškinti, kokią įtaką klientų patirtį sudarantys faktoriai daro jų pasitenkinimui.

Remdamiesi Guttentag‘o ir kt., Möhlmann‘o, Tussyadiah‘o mintimis, Xu ir kt. [4] teigia, jog naujai atlikti moksliniai tyrimai apie klientų patirtį dalijimosi ekonomika daugiausia naudojo tradicinį duomenų rinkimo būdą – apklausas arba interviu. Palyginus duomenis, gautus iš tradicinių klausimynais grįstų tyrimų, internetiniai atsiliepimai yra praktiškesni, atspindi dabartį bei gali pateikti naujų įžvalgų, todėl, pritaikius tinkamą analizę, gali padėti vadovams bei apgyvendinimo savininkams pagerinti paslaugų kokybę.

Nors atlikta nemažai mokslinių tyrimų viešbučių klientų patirties ir pasitenkinimo tema [8, 9, 10, 11, 12, 13, 14, 15], tačiau privataus būsto nuomos atvejis, kai, pasitelkiant didžiuosius duomenis, analizuojamos klientų patirties bei pasitenkinimo sąsajos, nagrinėtas rečiau [4, 5, 6, 7, 16, 17, 18]. Šiame darbe bus tiriama tokio tipo **problema** – kaip apgyvendinimo paslaugų klientų patirties pobūdis, atskleidžiamas jų atsiliepimuose, gali būti panaudojamas jų pasitenkinimui identifikuoti?

Darbo tikslas – modelio, kuris įvertintų sąsajas tarp apgyvendinimo paslaugas teikiančių vietų klientų patirties ir pasitenkinimo, sukūrimas.

Darbo tikslui pasiekti yra išsikeliami tokie **uždaviniai**:

1. atlikti apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų tyrimo mokslinės literatūros analizę;
2. parinkti požymius ir metodus, kurie geriausiai tinka tokio pobūdžio tyrimams, sukurti metodiką apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų modelių kūrimui;
3. sukurtą metodiką ir programines priemones pritaikyti realių duomenų analizei ir pateikti išvadas.

1. Literatūros apžvalga

Šiame skyriuje apžvelgiama teorinė darbo pusė: aptariama apgyvendinimo paslaugų klientų patirties sąvoka, pasitenkinimas ir jo svarba bei patirties ir pasitenkinimo sąsaja, pasitelkiant didžiuosius duomenis. Taip pat aprašomi pagrindiniai metodai, naudojami darbo uždaviniams bei problemai išspręsti: teksto tyryba, faktorinė ir regresinė analizė.

1.1. Apgyvendinimo paslaugų klientų patirties samprata

Šiame poskyryje apžvelgiamos pradinės sąvokos: apgyvendinimas, klientas ir svečias. Taip pat pateikiamas apgyvendinimo paslaugų kliento patirties apibūdinimas.

Pagal Ekonominės veiklos rūšių klasifikatorių NACE (2 red.), sektorius I atitinka „Apgyvendinimo ir maitinimo paslaugų veiklą“ [1], o jo skyrius 55 atitinka „Apgyvendinimo veiklą“, kuri apima trumpalaikį apgyvendinimą viešbučiuose ir kitose tokio pobūdžio paslaugą teikiančiose vietose [2]. Žmonėms, kurie keliauja ar yra toliau nuo namų daugiau, nei vieną dieną, reikia apgyvendinimo vietos: miegui, poilsiui, maistui, saugumui, vietos savo bagažui ir galimybės naudotis kitomis namų ūkio funkcijomis. Apgyvendinimas yra viena iš dalijimosi ekonomika (angl. *Sharing economy*) formų.

Terminai „vartotojas“ ir „klientas“ yra beveik sinonimai ir dažnai yra vartojami pakaitomis. Vis dėlto, tarp šių sąvokų egzistuoja nežymus skirtumas. Vartotojas yra apibrėžiamas kaip asmuo arba verslas, kuris vartoja arba naudoja prekes bei paslaugas [3]. Klientas yra pirkėjas, kuris įsigyja prekes bei paslaugas [3].

Svečias – tai žmogus, kuriam yra suteikiamas svetingumas iš apgyvendinimo vietos šeimininko [3]. Svetingumas yra suprantamas kaip ryšis tarp svečio bei šeimininko, kai pastarasis suteikia geranorišką priėmimą bei įvairias pramogas ir paslaugas savo klientams ir svečiams. Taigi, svečias gali būti apibrėžiamas ir kaip klientas, perkantis apgyvendinimo paslaugą.

Cheng‘as ir Jin‘as [5], remdamiesi Bridges‘u ir Vásquez‘u, Cheng‘u, Zervas‘u ir kt., savo straipsnyje teigia, jog didžiulis vieno iš privataus sektorių paslaugų teikėjo „Airbnb“ svetainės vystymas ne tik siūlo alternatyvios nakvynės patirtį savo klientams, tačiau tuo pačiu kelia iššūkį išplėtotai tradicinės viešbučių rinkos teorijai bei praktikai. Internetinė svetainė yra apibūdinama kaip individuali internetinė platforma, kurioje kuriama „patikima bendruomenės rinka ieškoti, atrasti bei užsakyti unikalią nakvynės vietą visame pasaulyje“ bei kuri „sujungia žmones nepakartojamai keliavimo patirčiai“. Pasak Cheng‘o ir Jin‘o [5], pastaruoju metu įvairūs tyrėjai pradėjo nagrinėti požymius bei atributus, kurie formuoja patirtį privataus būsto nuomos sektoriuje.

Egzistuojantys empiriniai tyrimai pateikia platų panašių, bet dažnai prieštaringų, rezultatų spektrą. Taip pat skirtinguose šaltiniuose skiriasi ir privataus būsto klientų patirties požymių svarbumo tvarka. Ankstesni tyrimai socialinį (svečias-šeimininkas) bendravimą nagrinėjo kaip esminį patirties aspektą. Pavyzdžiui, remiantis Yannopoulou mintimis, Cheng‘as ir Jin‘as [5] teigia, kad privataus būsto paslaugas teikianti internetinė svetainė iš esmės suteikia „prasmingą gyvenimo praturtinimą, žmogišką kontaktą bei autentiškumą“. Festila ir Müller [6] su pastaruoju teiginiu nesutinka ir mano, jog kai kuriems svečiams privataus būsto nuomos patirtis tėra viešbučio patirtis žemesne kaina. Cheng‘as ir Jin‘as [5], remdamiesi Guttentag‘u, sako, kad klientai itin vertina praktines savybes ir iš dalies mažiau vertina su patirtimi susijusias savybes. Kalbant apie praktines savybes, išsiaiškinta, jog

„vieta“ nedaro statistiškai reikšmingos įtakos klientų pasitenkinimui, tuo tarpu „malonumai“, „patogumai“ bei „išlaidų sutaupymas“ yra vertinami teigiamai (reikšmingumo tvarka). Tokių prieštaringų rezultatų priežastys išlieka neaiškios, nors tyrinėtojai užsimena, kad to paaiškinimas gali būti priskirtinas:

- privataus būsto apgyvendinimo vietų standarto trūkumui;
- keliavimo pageidavimams ir svečių asmenybės tipui (introvertai – „keliauju dėl pojūčių“, ekstravertai – „keliauju pamatyti“).

Nepaisant įvairių diskusijų, bendrai nusistovėję atributai, formuojantys privataus būsto klientų patirtį yra „ekonominiai privalumai / mažesnė kaina“, „vieta“, „namų ūkio patogumai“, „švara“, „autentiška patirtis / šeimininko-svečio bendravimas“ bei laiko leidimas vietinėse apylinkėse.

Tuo tarpu atributai, lemiantys klientų viešbučių pasirinkimą, paslaugų pirkimą, patirties kokybę bei pasitenkinimą yra viena iš labiausiai nagrinėjamų svetingumą tiriančių mokslo darbų sričių. Išnagrinėję Albayrak‘o ir Caber‘io, Alcántara-Alcover‘io ir kt., Ariffin‘o ir Maghzi, Callan‘o ir Bowman‘o, Crnojevac‘o ir kt., Dolnicar‘o ir Otter‘io tyrimus, Cheng‘as ir Jin‘as [5] teigia, kad jie bandė sukurti viešbučio požymių sąrašą, kuris atspindėtų klientų lūkesčius bei apibrėžtų jų pirkimo sprendimą. Klientų viešbučių pasirinkimą lemiantys požymiai yra viešbučio prekinis ženklas bei įvaizdis, kaina, viešbučio fizinės ypatybės (dydis, dizainas, dekoracijos, švara, įrangos gausa, patogumai, plotas), kambarių savybės (dydis, baldų įvairovė ir jų gausa), paslaugos, saugumas, maisto ir gėrimų teikimas bei vieta. Pasak Cheng‘o ir Jin‘o [5], kurie rėmėsi Crnojevac‘u ir kt., Dolnicar‘u ir Otter‘iu, Liu ir kt., nors atributų sąrašas skiriasi jų išdėstymo tvarka, tačiau klientai viešbutį renkasi pagal paslaugas, vietą, kambarius, kainą arba kainos ir kokybės santykį, maisto bei gėrimų teikimą, viešbučio įvaizdį, saugumą ir viešbučio marketingo politiką.

Vis dėlto dėl privataus būsto nuomos, kaip alternatyvos apgyvendinimo pasirinkimui, augimo, viešbučių industrija siekia išsiaiškinti, kokie atributai lemia tokį jų populiarumą. Remiantis Belarmino ir kt. bei Mody ir kt. tyrimais, Cheng‘as ir Jin‘as [5] tvirtina, jog pagrindiniai skirtumai tarp privataus būsto ir tradicinių viešbučių slypi „šeimininko / viešbučio personalo“, „svečio“ bendravimo ir „atmosferos“ atributuose, o kiti požymiai yra gana panašūs.

1.2. Apgyvendinimo paslaugų klientų pasitenkinimo sąvoka

Šiame poskyryje apibūdinami skirtingi apgyvendinimo klientų pasitenkinimo sąvokos apibrėžimai, apžvelgiama, kokiais būdais jis matuojamas.

Mokslinėje literatūroje galima rasti nemažai modelių bei teorijų, aprašančių ir tiriančių klientų pasitenkinimą. Pasak Liang‘o, Choi‘o ir Joppe [7], viena iš pagrindinių teorijų yra Oliver‘io pasiūlyta lūkesčių nepatvirtinimo teorija (angl. *expectancy-disconfirmation theory*), vėliau Kristensen‘o ir kt. išplėsta į lūkesčių patvirtinimo teoriją (angl. *expectancy-confirmation theory*). Ja siekiama paaiškinti pasitenkinimą po pirkimo remiantis keturiais pagrindiniais elementais: lūkesčiais, suvokiamu rezultatu, nuomonės paneigimu bei pasitenkinimu. Remiantis Oh‘u ir Parks‘u, Liang‘as ir kt. [7] teigia, kad egzistuoja dar aštuonios kitos teorijos, tiriančios svečių pasitenkinimą. Pavyzdžiui, jie sako, kad Fang‘as ir kt. pasinaudojo Holmes‘o pasitenkinimo apibrėžimu ir suformulavo jį taip – pasitenkinimas yra praeityje gautos patirties įvertinimas. Pasak Liang‘o ir kt. [7], kurie rėmėsi Kim‘u, pasitenkinimas yra apibūdinamas paslaugų ir kokybės palyginimu, kurį klientas tikisi gauti po pirkimo.

Xiang‘as, Schwartz‘as, Gerdes‘as Jr. ir Uysal‘as [8] teigia, kad apgyvendinimo paslaugas teikiančių vietų klientų pasitenkinimo sąvoka yra sudėtinė žmogaus patirtis svetingumo paslaugų aplinkos atžvilgiu. Klientų pasitenkinimas tyrimuose pradėtas nagrinėti nuo 1970 metų, nuo tada atsirado daug skirtingų šios sąvokos apibrėžimų. Remiantis Hunt‘u, Xiang‘as ir kt. [8] sako, jog pasitenkinimas gali būti apibrėžiamas tada, kai klientų patirtis yra apibūdinama kaip gera bent jau tiek, kiek tokia turėtų būti. Tuo tarpu kiti autoriai klientų pasitenkinimą apibūdina kaip emocinį atsaką į produkto ar paslaugos naudojimą. Oh‘us ir Parks‘as, pasak Xiang‘o ir kt. [8], tvirtina, kad klientų pasitenkinimas apima pažintinius, emocinius, o taip pat ir psichologinius bei fiziologinius procesus. Plačiai naudojamas klientų pasitenkinimo apibrėžimas teigia, jog tai yra sąveikos tarp kliento lūkesčių prieš įsigyjant prekę / paslaugą bei jos įvertinimo po pirkimo rezultatas.

Iš vadovo perspektyvos, svarbiau suprasti kliento pasitenkinimo komponentus ar šios sąvokos pradmenis. Pavyzdžiui, viešbučio „produkto“ koncepciją sudaro keli lygiai. Tai paslaugos, vieta, kambarys, kaina, maistas ir gėrimai, įvaizdis, saugumas bei rinkodara. Dažnai minima dviejų faktorių teorija (angl. *Two Factor Theory*) teigia, jog:

- higienos faktoriai, tokie kaip viešbučio / kambario švara ir priežiūra, teigiamai neprisideda prie pasitenkinimo, nors jų nebuvimas ir lemia kliento nepasitenkinimą;
- motyvaciniai faktoriai, tokie kaip empiriniai viešnagės viešbutyje aspektai, lemia teigiamą pasitenkinimą.

Remiantis Chathoth‘u ir kt., Xiang‘as ir kt. [8] nurodo, jog pastaruoju metu mokslininkai teigia, kad klientų patirtis neturėtų būti apibrėžta tik apgyvendinimo vietos siūlomų paslaugų kiekiu / pobūdžiu, vietoje to, tai turėtų atspindėti paslaugų teikėją bei klientą bendrai. Taigi, kliento pasitenkinimas gali būti išreiškiamas kaip kliento patirties sąveikaujant su įvairių paslaugų sektoriaus sritimis įvertinimas.

Turint omenyje, kokia sudėtinga yra kliento patirties sąvoka, matuoti bei valdyti apgyvendinimo vietos kliento pasitenkinimą gali būti gana sudėtinga užduotis. Svetingumo srityje atlikti tyrimai rodo, kad egzistuoja atotrūkis tarp to, ką vadovai laiko svarbiu renkantis ir vertinant apgyvendinimą, bei to, kas iš tikrųjų yra svarbu klientams. Svečių apklausos, ypač klientų atsiliepimų kortelės, yra plačiai naudojamos viešbučių klientų pasitenkinimui matuoti. Nors šios apklausos yra efektyvios ir naudingos, tačiau dažnai susiduriama su prasta imties kokybės bei žemo atsakymų rodiklio problema, be to, šis metodas neretai pateikia neapibrėžtus klientų patirties vertinimus. Taip pat tokio tipo apklausos neatsižvelgia į individualių nakvynės vietos savybių klientui svarbą.

Kitas būdas, skirtas viešbučio klientų pasitenkinimui matuoti – „reikšmingumo-įvykdymo“ analizė (angl. *importance–performance analysis*) – gali sumažinti anksčiau minėtą problemą. Nepaisant to, tokia analizė reikalauja, kad vertinami viešbučio požymiai būtų iš anksto apibrėžti.

Pasak Xiang‘o ir kt. [8], kurie rėmėsi Crotts‘u ir kt., egzistuoja ir nelimituotų klausimų analizė (angl. *open-ended questions*), kuri gali sukurti gausių bei reikšmingų (asmeniškai) atsiliepimų imtį, tačiau kokybinis jos pobūdis gali būti sunkiai analizuojamas, o rezultatuose neretai trūksta apibendrinamumo. Be to, remiantis Oh‘u, teigiama, jog svarbu apsvarstyti naujus kintamuosius tobulinant apgyvendinimo vietos klientų pasitenkinimo teoriją. Tokia prielaida, kai siūloma įtraukti ir tirti naujus duomenų šaltinius geresniam klientų patirties ir pasitenkinimo supratimui, yra gana perspektyvi tolimesnių tyrimų kryptis.

Pastaruoju metu dedama vis daugiau pastangų į klientų sukurto turinio naudojimą matuojant klientų / turistų pasitenkinimą. Pavyzdžiui, išnagrinėjus Pan‘o ir kt. tyrimą, Xiang‘as ir kt. [8] sako, jog minėti autoriai nagrinėjo internetinių kelionių dienoraščių, kaip kokybinių duomenų rinkinio, naudą charakterizuojant, kas klientams patiko ir nepatiko jų pirkimo patirtyje. Kiti autoriai (pavyzdžiui, jau minėti Crotts‘as ir kt.) pritaikė kiekybinę padėties-kaitos analizę (angl. *quantitative stance-shift analysis*), norėdami išmatuoti viešbučių klientų pasitenkinimą besinaudodami internetinių dienoraščių pasakojimais, paskelbtais pačių klientų.

Be abejo, apgyvendinimo paslaugų svečių pasitenkinimą lemia įvairūs ir ne visada vienodi veiksniai. Nors skirtingi autoriai pasitenkinimo sąvoką apibrėžia savaip, tačiau Xiang‘as ir kt. [8] apibendrina, jog kliento pasitenkinimas gali būti apibūdinamas jo gautos patirties įvertinimu. Iki šiol atlikti tyrimai yra naudingi norint pagerinti klientų pasitenkinimo suvokimą, tačiau jie paremti gana nedidele duomenų imtimi, todėl yra riboti didžiųjų duomenų analitikos požiūriu.

1.3. Apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų tyrimai didžiųjų duomenų kontekste

Šiame poskyryje pateikta mokslinių tyrimų, kuriuose analizuojamos apgyvendinimo paslaugų klientų patirties bei pasitenkinimo sąsajos, taikant didžiųjų duomenų analizės metodus, apžvalga.

Hargreaves [9] tyrė Singapūro viešbučių svečių pasitenkinimo reitingus bei atsiliepimus. Jos tikslas – išanalizuoti klientų įvertinimus ir atsiliepimus naudojant statistinę analizę; įvertinti svečių paliktų reitingų požymius kiekvienam iš 5 Singapūre esančių viešbučių. Taip pat identifikuoti kiekvieną viešbutį atitinkančias savybes, veiksnius, darančius įtaką klientų pasitenkinimui bei veiksnius, kurie gali būti pagerinti.

Duomenų imtį sudaro 14175 atsiliepimai nuo 2005 rugsėjo mėn. iki 2014 rugsėjo mėn. Kiekvieną įrašą sudaro tekstinis svečio atsiliepimas apie jo individualią patirtį bei įvertinimai kiekvienam iš atributų: vietai, miego kokybei, kambariui, paslaugų kokybei, kainos ir kokybės santykiui, švarai. Pritaikyti analizės metodai: teksto analitika bei sentimentų analizė.

Autorė neišskiria tyrimo trūkumų bei rekomendacijų tolimesniems tyrimams. Nepaisant to, pateikia tokius rezultatus:

1. Remiantis žodžių dažniu ir įvertinimų įžvalgomis, gauta, kad kambario bei paslaugų kokybės kintamieji yra svarbiausi;
2. 5 pagrindiniai žodžiai: viešbutis, kambarys, personalas, vaizdas, puikus, geras, paslaugos, vieta.

Xiang‘as, Schwartz‘as, Gerdes‘as Jr. Ir Uysal‘as [8] savo tyrime kelia klausimą: ką didieji duomenys ir teksto tyryba gali pasakyti apie viešbučių svečių patirtį ir pasitenkinimą? Jų tikslas yra ištirti ir pademonstruoti didžiųjų duomenų analitikų naudą, bandant geriau suprasti svetingumo problemas, o tiksliau – ryšį tarp viešbučių svečių patirties bei pasitenkinimo.

Pagrindinis naudojamas metodas – teksto tyryba. Tyrime parodomas didžiųjų duomenų analitikos panaudojimas nustatant viešbučių svečių elgesio šablonus, naudojant jų sukurtą turinį, prieinamą internete. Tyrimo išvados parodė, kaip klientai „kalba“ apie savo patirtį atsiliepimuose, parašytuose internete.

Išskiriami gauti rezultatai:

1. Reikšminga higienos ir motyvacijos faktorių santykio diferenciacija;
2. Apibendrinantis šablonas – stipri asociacija tarp patirties ir pasitenkinimo;
3. Higienos veiksniai yra esminės viešbučių paslaugos, be kurių svečiai negali visiškai mėgautis savo patirtimi.

Nors buvo gautas šablonas, kuris apibendrino stiprų patirties ir pasitenkinimo ryšį, tačiau buvo ir analizės trūkumų, t. y.: 1) klientų šališkumo faktorius rašant atsiliepimus internete; 2) tyrime buvo nagrinėti tik TOP 100 miesto viešbučių Amerikoje; 3) naudoti duomenys buvo kelių metų senumo ir galėjo neatspindėti dabartinio klientų požiūrio.

Tyrimams šia tema ateityje rekomenduojama apsvarstyti trianguliacijos metodų taikymą daugialypiams duomenims tam, kad būtų galima pagrįsti svečių patirties semantinę struktūrą visapusiškam svečių pasitenkinimo tyrimui naudojant didžiųjų duomenų analitiką.

Kitame straipsnyje, Liu, Teichert‘as, Rossi, Li ir Hu [10] tiria teksto tyrybos įrankius svečių sukurtiems atsiliepimams apie pasitenkinimą viešbučiais. Jų tyrimo tikslas – maksimaliai išnaudoti klientų sukurtų atsiliepimų privalumus tam, kad būtų galima pasiūlyti naujų įžvalgų apie veiksnius, lemiančius viešbučių svečių pasitenkinimą, atskiriant klientus pagal kalbos grupę.

Tiriami 412784 klientų sukurti atsiliepimai internetiniame su kelionėmis susijusių paslaugų užsakymo puslapyje 10149 viešbučiams iš 5 didžiųjų Kinijos miestų. Šie atsiliepimai buvo surinkti naudojant PHP programavimo kalbą. Iš kiekvieno atsiliepimo buvo išgauti tokie požymiai: turinys, rašymo laikas, apibendrintas pasitenkinimo lygis bei kiekvieno iš penkių viešbučių savybių įvertinimas. Tiriamiems viešbučiams autoriai pridėjo du papildomus požymius: miestą bei viešbučio žvaigždučių skaičių. Kalbos identifikavimas buvo atliekamas naudojant MySQL bei R programines įrangas. Atlikus tyrimą, autoriai išskiria gautus rezultatus:

1. Identifikuota, jog kalbėjimas skirtinga kalba bei skirtingų kultūrinių šaknų turėjimas veikia klientų pirmenybės teikimą įvairioms viešbučių savybėms;
2. Lyginti Azijos šalių (konkrečiau Kinijos, Japonijos bei rusakalbių) klientų atsiliepimai. Toks lyginimas parodė, jog, pavyzdžiui, klientai iš Kinijos teikia didesnį įvertinimo svorį viešbučio kambariui, bet gana mažai dėmesio skiria viešbučio aptarnavimui;
3. Statistinė analizė patvirtino hipotezę apie tai, jog klientų pasitenkinime viešbučiais reikšmingą įtaką turi kainos ir aptarnavimo santykis.

Pagrindinė išvada – užsienio klientams svarbiausios viešbučio savybės yra aptarnavimo lygis, kambarys bei kaina, taip pat švara ir vieta. Nors buvo gauta reikšmingų rezultatų, neapsieita ir be tyrimo trūkumų: 1) kai kurių įvertinimų vidurkiai galėjo būti nulemti kultūrinių skirtumų; 2) turistai iš Kinijos gali turėti kitokių lūkesčių apgyvendinimui lyginant su kitomis šalimis; 3) neatsižvelgiama į faktą, jog, pavyzdžiui, anglų kalba gali kalbėti ne tik Didžiosios Britanijos, bet taip pat ir Amerikos ir kitų šalių piliečiai.

Panašios ar tokios pačios srities tyrimams ateityje rekomenduojama analizuoti daugiau, nei penkias viešbučių savybes. Taip pat įvertinti faktą, jog kelias kalbas mokantys klientai atsiliepimus gali rašyti nebūtinai savo gimtąja kalba bei viena konkreti kalba kaip pagrindinė gali būti naudojama ne vienoje šalyje.

Sthapit‘as ir Jiménez-Barreto‘as [16] savo straipsnyje tyrė privataus būsto atvejo turistams įsimintinas svetingumo patirtis. Jų analizės tikslas buvo išstudijuoti svarbiausius įsimintinos privataus būsto nuomos internetinės svetainės svetingumo patirties elementus, pasitelkiant grindžiamosios teorijos metodą. Duomenys tyrimui buvo surinkti pusiau struktūrizuotos apklausos metu, pritaikyta kokybinė jų analizė. Apklausos subjektai buvo asmenys, naudojęsi paslaugomis per paskutinius 12 mėnesių. Jie identifikuoti „sniego gniūžtės“ principu. Duomenų analizei naudotas grindžiamosios teorijos metodas.

Tyrimo metu gauti tokie rezultatai:

1. Respondentai savo patirtį su privataus būsto apgyvendinimu vertino teigiamai;
2. Pagrindinės tokio apgyvendinimo pasirinkimo priežastys buvo kaina ir vieta;
3. Socialinis bendravimas bei šeimininko požiūris buvo esminiai geros privataus būsto nuomos patirties veiksniai.

Straipsnio autoriai išskyrė ir keletą savo analizės trūkumų. Visų pirma, į atrankos kriterijus nebuvo įtrauktas apgyvendinimo tipas (privatūs kambariai ar visas namas / apartamentas) bei neatsižvelgta į tai, ar įsiregistravimas buvo savarankiškas, ar svečiai pasitikti šeimininko. Antra, dalyviai daugiausia buvo iš vakarų šalių bei buvo atlikta mažai apklausų. Taip pat duomenys buvo surinkti po viešnagės, todėl gali būti neatitikimų tarp patirties iš atsiminimų bei patirties, išreikštos išsiregistruojant.

Tolimesniems tyrimams pateikiamos rekomendacijos: 1) apklausos turėtų vykti iškart po viešnagės – pavyzdžiui, išsiregistruojant; 2) emocinės įtakos privataus būsto nuomos patirties įsimenamumui tyrimas gali pateikti tolimesnių įžvalgų numatomai vartojimo emocijų galiai; 3) tirti turistų privataus būsto nuomos patirties poveikį jų asmeninei gerovei.

Liang‘as, Choi‘us ir Joppe [7] atliko tyrimą, pavadinimu „Ryšio tarp pasitenkinimo bei pasitikėjimo ir perėjimo ketinimo bei pakartotinio pirkimo ketinimo tyrimas „Airbnb“ kontekste“. Straipsnyje pateikiamos tyrime keliamos hipotezės:

1. „Airbnb“ klientų pasitenkinimo lygis transakcijos metu turi teigiamą ryšį su: jų ketinimu įvykdyti pakartotinį pirkimą (H1a); jų pasitikėjimu „Airbnb“ (H1b); jų pasitikėjimu šeimininku (H1c); jų pasitenkinimo lygiu, gautu per patirtį (H1d), bet turi neigiamą ryšį su ketinimu pereiti pas konkurentus (H1e);
2. „Airbnb“ klientų pasitenkinimo lygis, gautas per patirtį, turi teigiamą ryšį su: jų pasitikėjimu „Airbnb“ (H2a); jų pasitikėjimu šeimininku (H2b); jų ketinimu įvykdyti pakartotinį pirkimą (H2d), bet turi neigiamą ryšį su jų ketinimu pereiti pas konkurentus (H2c);
3. „Airbnb“ klientų pasitikėjimas svetainės paslaugomis turi teigiamą ryšį su jų ketinimu įvykdyti pakartotinį pirkimą (H3a), bet turi neigiamą ryšį su jų ketinimu pereiti pas konkurentus (H3b).

Duomenų imtis buvo sudaryta iš 395 apklausų bei buvo vieno mėnesio laikotarpio (daugiausia 2015 metų sausio mėnesio). Populiacijos informacija buvo išskirta naudojantis SPSS programa. Pritaikyta patvirtinančioji faktorinė analizė bei struktūrinių lygčių metodas, įgyvendinti programa AMOS.

Autoriai pateikia tokius gautus rezultatus:

1. Gauti rezultatai stipriai paremia hipotezes, susijusias su dėl įvykusios transakcijos atsiradusiu klientų pasitenkinimu;

2. Patirtimi grįstas pasitenkinimas statistiškai reikšmingai neveikia klientų suvokiamo pasitikėjimo internetinės svetainės paslaugomis ar apgyvendinimo vietos šeimininku;
3. Pasitikėjimas privataus būsto nuomos internetine svetaine statistiškai reikšmingai neveikia svečių pasitikėjimo apgyvendinimo vietos šeimininku;
4. Nebuvo rasta jokio poveikio tarp pasitikėjimo privataus būsto nuomos internetine svetaine bei perėjimo pas konkurentus ketinimo.

Autoriai teigia, jog vienas iš jų tyrimo trūkumų yra tai, kad duomenų imtį sudarė tik tie svečiai, kurie buvo apsistoję Kanadoje ir Jungtinėse Amerikos Valstijose. Taip pat gauti rezultatai gali būti paveikti bendru taikyto metodo šališkumu. Be to, tyrime gali būti validumo problemų tarp transakcija paremtu pasitenkinimu bei pasitikėjimu apgyvendinimo vietos šeimininku (remiantis gauta reikšme bei AVE reikšme). Tolimesniems tyrimams rekomenduojama pabandyti atskirti pasitenkinimą „Airbnb“ paslaugomis bei šeimininku, taip pat pasitikėjimą tiek prieš, tiek po pirkimo. Įvertinus šiuos kintamuosius, vertėtų palyginti gautus rezultatus su straipsnyje įgyvendintu modeliu. Taipogi, siūloma plėsti ir geografinę tyrimo teritoriją.

Zhao, Xu ir Wang‘as [11] savo tyrime prognozavo klientų bendrą pasitenkinimo lygį, naudojant viešbučių internetinių tekstinių atsiliepimų didžiuosius duomenis. Tyrimo tikslas – prognozuoti bendrąjį klientų pasitenkinimo lygį naudojant tekstinių atsiliepimų internete požymius bei klientų dalyvavimą „atsiliepimų bendruomenėje“.

Buvo tirti 2017 m. birželio mėnesio 127629 atsiliepimai, surinkti iš kelionių planavimo ir su ja susijusių paslaugų užsakymo svetainės, modeliais, realizuotais programavimo kalba Python. Naudojamasi *Selenium* paketu sukurti kelioms Python programoms tam, kad išrinkti pagrindinį tyrimo turinį, pavyzdžiui, visų San Francisko viešbučių sąrašą, viešbučių reitingą, svečio profilį ir t. t. Vėlesniuose tyrimo etapuose taip pat naudojama MySQL bei SAS programos. Pasirenkama nagrinėti San Francisko duomenis dėl didelio viešbučių populiarumo bei jų internetinių atsiliepimų.

Atlikus analizę, pateikiami tokie rezultatai:

1. Tam tikri techniniai atsiliepimo požymiai, t. y., subjektyvumas, įskaitomumas bei ilgis, reikšmingai neigiamai veikia įvertinimus iš klientų;
2. Įvairovė bei priešingo požiūrio išreiškimas stipriai teigiamai veikia įvertinimus iš klientų;
3. Klientų įsipareigojimas palikti atsiliepimą teigiamai veikia įvertinimus.

Taip pat išskiriami tyrimo trūkumai: 1) duomenis sudaro tik vieno miesto viešbučių atsiliepimai, jie buvo surinkti naudojantis tik viena internetine svetaine; 2) internetinių tekstinių atsiliepimų techniniams požymiams įtaką galėjo daryti klientų naudojama kalba bei kultūra; 3) tam tikri požymiai atsiliepimuose bei viešbučių vertinimai bėgant laikui gali keistis.

Tolimesnei analizei rekomenduojama išplėsti šį tyrimą surenkant daugiau duomenų iš skirtingų miestų bei iš įvairių šaltinių. Taip pat rekomenduojama tirti ir lyginti internetinius tekstinius atsiliepimus, parašytus skirtingomis kalbomis bei skirtingos kultūros svečių. Tolimesnei analizei taip pat patariama bandyti nagrinėti atsiliepimus ne tik apie viešbučius, bet ir restoranus ar oro uostus.

Ahani‘as ir kt. [12] siekia ištirti Kanarų salų viešbučių svečių pasitenkinimą bei atskleisti pirmenybės teikimą, naudojant internetinių atsiliepimų analizę. Tyrimo tikslas susideda iš dviejų dalių:

- identifikuoti keliautojų pirmenybes bei segmentuoti jas remiantis internetiniais atsiliepimais bei reitingais iš kelionių planavimo ir užsakymo svetainės;
- sukurti sprendimų priėmimo sistemą, naudojantis klasterizavimo metodus bei daugiakriterinį sprendimų priėmimą, taip segmentuojant viešbučių svečius bei ranguojant jų pirmenybes.

Duomenų imtis – 1334 Kanarų salų 4 bei 5 žvaigždučių viešbučių svečių atsiliepimai bei reitingai. Tolimesni analizės žingsniai yra klientų reitingų klasterizavimas SOM (angl. *Self-organizing map*) metodu bei viešbučių savybių rangavimas TOPSIS (angl. *Technique for Order of Preference by Similarity to Ideal Solution*) metodu.

Analizė buvo atliekama MATLAB ir Excel programomis. Buvo gauti tokie rezultatai:

1. Identifikuoti 9 klasteriai;
2. Atsižvelgiant į klasterių informaciją, sudaryti 4 segmentai: ypač patenkintų klientų, patenkintų klientų, vidutiniškai patenkintų klientų ir nepatenkintų klientų.

Apibendrinant autoriai teigia, jog „TripAdvisor“ tinklalapis yra itin svarbus informacijos šaltinis, padedantis viešbučiams padidinti savo matomumą bei formuoti stipresnius ryšius su turistais per internetinius atsiliepimus bei reitingus.

Nors trūkumų straipsnyje išskirta nėra, tačiau pateikiamos rekomendacijos tolimesniems tyrimams: 1) tirti atsiliepimų įtaką bendram klientų paslaugų kokybės suvokimui; 2) nagrinėti, kaip internetiniai atsiliepimai gali būti susieti su reitingais, norint tiksliau ištirti klientų pasitenkinimą; 3) vystyti algoritmus, kurie tirtų internetinius atsiliepimus ir reitingus palaipsniui, kadangi bėgant laikui, atsiliepimai bei reitingai gali keistis.

Cheng‘as ir Jin‘as [5] analizuodami internetinius atsiliepimus bandė išsiaiškinti, kas yra svarbu privataus būsto nuomos klientams. Jų tyrimo tikslas – ištirti savybes, kurios daro įtaką tokio tipo apgyvendinimo klientų patirčiai, pasitelkiant didžiųjų duomenų rinkinį, sudarytą iš Sidnėjuje viešėjusių turistų internetinių atsiliepimų.

Duomenų rinkinį sudarė 181263 atsiliepimai. Jis buvo gautas iš „Inside Airbnb“ internetinės svetainės. Tyrime pritaikomi teksto tyrybos bei sentimentų analizės metodai, naudojantis Leximancer programa.

Atlikus analizę, gauti tokie rezultatai:

1. Privataus būsto nuomos klientai patirčiai išreikšti dažniausiai naudoja tokias pačias apgyvendinimo savybes, susijusias su viešnage, nors jų seka pagal svarbumą gali būti skirtinga;
2. Išskiriami trys pagrindiniai atributai: vieta, patogumai ir šeimininkas;
3. Privataus būsto nuomos klientų patirtis gali būti labiau susiskaldžiusi ir mažiau nuspėjama, nei viešbučių klientų patirtis, kuri dažnai yra labiau standartizuota;
4. Tinkama pradinė komunikacija vaidina svarbų vaidmenį kuriant pradinį pasitikėjimą, kadangi privataus būsto nuomos svetainė teikia nepažįstamasis-nepažįstamajam sandorį.

Nors mašininio mokymosi algoritmai yra naudingas įrankis, autoriai mano, jog jis vis dar reikalauja teorinio pagrindimo tikslinant ir aiškinant duomenis tarpdisciplininei auditorijai. Net jei programa Leximancer teikia darnų vizualinį didžiųjų duomenų reprezentavimą, tyrėjams vis tiek reikėjo nemažų laiko kaštų bei pastangų interpretuoti gautus rezultatus. Įvertinus tyrimo rezultatus bei

trūkumus, autoriai pateikia rekomendacijas panašaus pobūdžio analizei: 1) taikyti regresinę analizę įtraukiant kitus kintamuosius, pavyzdžiui, reitingus bei apgyvendinimo vietos aprašymus; 2) išsiaiškinti, kodėl didelis skaičius klientų savo atsiliepimus rašo ir anglų, ir gimtąja kalba; 3) kelių apgyvendinimo paslaugų užsakymo svetainių palyginimas tiriant daugiau šalių galėtų duoti papildomų, tarpkultūrinių įžvalgų; 4) nors tyrimo metu identifikuoti maži skirtumai tarp privataus būsto ir viešbučių, to nepakanka nustatyti, kiek Sidnėjaus miesto atvejo analizės rezultatai gali būti apibendrinami.

Joseph‘as bei Varghese‘as [17] naudojo teksto tyrybos metodus analizuojant privataus būsto nuomos klientų patirties atsiliepimus. Jų parašyto straipsnio tikslas buvo pateikti teksto tyrybos uždavinį – „Airbnb“ klientų atsiliepimų analizę – bei išsiaiškinti, kokios charakteristikos sužadina klientų pasitenkinimą.

Duomenų imtis analizei atlikti – atsiliepimai, palikti svetainėje privataus būsto nuomos svetainėje po viešnagės Londono mieste. Jų laikotarpis buvo nuo 2014-04-09 iki 2017-02-26. Analizei pritaikyta sentimentų analizė, atlikta RapidMiner programa.

Tyrimo metu gauti rezultatai:

1. Kambarių švara, paslaugų kokybė bei nusimanantys darbuotojai yra itin svarbūs matuojant klientų pasitenkinimą;
2. Nakvynės vieta gali pritraukti daug teigiamų atsiliepimų lyginant su kambarių patogumais, kurie sulaukė nemažai neigiamų atsiliepimų;
3. Klientų įsitraukimas bei aktyvus šeimininkų dalyvavimas priimant svečius gali sustiprinti šeimininko kompetentingumo poziciją.

Vienas iš tyrimo trūkumų, kurį įvardija autoriai, yra nepritaikyta prognozavimo analizė. Kitas tyrimo trūkumas – imtis yra tik iš vienos konkrečios vietos, t. y., Londono miesto. Be to, autoriai mini ir nepanaudotą klasterinę analizę. Atsižvelgus į šiuos trūkumus, rekomenduojama atlikti išsamesnį tyrimą, pasitelkiant didesnį atsiliepimų kiekį iš skirtingų lokacijų bei papildomai nagrinėti ir paliktus įvertinimus. Taip pat siūloma sukurti ir išplėtoti žodžių, pritaikytų privataus būsto nuomos svetainės duomenims, žodyną.

Padma‘as ir Ahn‘as [13] taiko didžiuosius duomenis tirdami klientų pasitenkinimą ir nepasitenkinimą prabangiuose viešbučiuose. Autoriai siekia išsiaiškinti, kurios prabangių viešbučių savybės labiausiai prisideda prie Malaizijos klientų pasitenkinimo bei nepasitenkinimo, naudojant didžiuosius duomenis.

Duomenys buvo surinkti 2019 metais iš internetinės kelionių užsakymo ir planavimo svetainės. Buvo pasirinkti keturi 5 žvaigždučių viešbučiai, remiantis atsiliepimų skaičiumi. Taip pat atsiliepimų turinio analizei atsitiktinai parinkti 800 klientų. Pagrindiniai analizės metodai: žodžių dažnumas, skirtas suklasifikuoti paslaugų kokybę apibūdinančias savybes, svarbiausių įvykių metodas, norint identifikuoti svarbiausias klientų požymius, t. y., pasitenkinimą ir nepasitenkinimą.

Tyrimo metu buvo gauti tokie rezultatai:

1. Identifikuoti keturi pagrindiniai paslaugų tipai, susiję su: viešbučiu, kambariu, personalu ir keliavimu. Jie padėjo suprasti svarbiausius tikėtinus įvykius, slypinčius klientų atsiliepimuose, norint prognozuoti tolimesnę elgseną;

2. Prabangių viešbučių klientai turi gerokai didesnius lūkesčius, nei paprastų viešbučių klientai;
3. Išryškinti kelionės ir viešbučių suvokimo skirtumai tarp prabangių bei įprastų viešbučių klientų;
4. Klientų bendravimas su prabangių viešbučių darbuotojais darė įtaką tolimesnėms rekomendacijoms potencialiems klientams.

Autoriai išskiria vieną tyrimo trūkumą – imtis buvo sudaryta tik iš prabangių viešbučių, esančių Malaizijoje, svečių atsiliepimų. Tolimesniems tyrimams patariama didinti geografinę tyrimo sritį, kad būtų galima geriau suvokti bendrą prabangių viešbučių klientų elgesį. Taip pat rekomenduojama kiekybiškai tirti ryšį tarp paslaugų kokybės, pasitenkinimo ir teigiamų rekomendacijų iš buvusių klientų būsimiems svečiams. Vėlesniems tyrimams patartina tirti skirtumus tarp vietinių bei kitų šalių svečių pasitenkinimą / nepasitenkinimą.

Baek'as, Choe'as ir Ok'as [14] tyrime „Veiksniai, lemiantys viešbučių svečių paslaugų patirtį: skirtumų tarp „gyvenimo būdo" ir įprastų viešbučių tyrimas" bando palyginti „gyvenimo būdo" ir tradicinių viešbučių semantinius faktorius, susijusius su svečių patirtimi, išskirtus iš socialinių tinklalapių atsiliepimų.

Iš pradžių, atrinkti 17 „gyvenimo būdo" viešbučių bei 18 iš 451 (kad būtų galima surinkti panašų duomenų kiekį) atsitiktinai parinktų tradicinių viešbučių Niujorko mieste iš kelionių planavimo ir užsakymo svetainės, iš viso buvo gauta 45490 viešbučių svečių atsiliepimų. Duomenys buvo tiriami teksto gavybos metodais, naudojantis programa R – pritaikyta tiriamoji faktorinė analizė (angl. *Exploratory factor analysis*), kad būtų galima išskirti semantinę viešbučių svečių patirties struktūrą bei įvertinti du tiesinės regresijos modeliai, norint išanalizuoti išskirtų semantinių subjektų svarbą ir sujungti juos su bendru reitingo balu.

Gauti rezultatai rodo, kad:

1. „Gyvenimo būdo" viešbučiai nuo tradicinių skiriasi šiais aspektais: svečių kambariais, darbuotojų ir klientų bendravimu bei svetingumo paslaugomis;
2. Asimetriniai negatyvių semantinių faktorių efektai bendram atsiliepimo įvertinimui teikia naudingas įžvalgas abiejų nagrinėjamų viešbučių tipų vadovams;
3. Internetiniai klientų atsiliepimai yra tinkamas duomenų šaltinis tokio tipo tyrinėjimams. Pastebima, kad tiriant klientų patirtis didelė duomenų apimtis yra tinkamesnė, nei maža populiacija.

Išskiriami tyrimo trūkumai: 1) naudojami duomenys galėjo būti paveikti šališkumo; 2) tiriami tik 35 viešbučių, esančių Niujorke, duomenys, naudota tik viena internetinė svetainė; 3) panaudoti klasikiniai statistinės analizės metodai galėjo nulemti tam tikras problemas skaičiavimuose.

Tolimesnei analizei patariama nagrinėti ne tik socialiniuose tinkluose esančius atsiliepimus, nes ne visi klientai juos rašo internete. Taip pat analizuoti daugiau viešbučių bei platinti analizuojamą geografinę padėtį. Tolimesniems tyrimams patariama analizei naudoti ir kitų socialinių tinklų informaciją bei taikyti pažangesnius teksto analitikos ir mašininio mokymosi metodus.

Li'as ir Ryan'as [15] atliko neįprastą tyrimą, pavadinimu „Vakarų klientų patirtis Pchenjano tarptautiniame viešbutyje, Šiaurės Korėjoje: pasitenkinimo tyrimas, esant priverstinio pasirinkimo aplinkybėms". Siekiama ištirti turistų iš vakarų patirtį tarptautiniame Yanggakdo viešbutyje ir išnagrinėti klientų patirties specifiką, esant riboto pasirinkimo aplinkybėms.

Analizei atlikti iš kelionių planavimo ir užsakymo svetainės surinkti 227 atsiliepimų duomenys, iš kurių unikalių žodžių kiekis – 46050. Pritaikyta SERVQUAL skalė. Atsiliepimams nagrinėti naudota teksto analizė, atlikta Leximancer ir QDA Miner programomis.

Tyrimo metu gauta ne tik tipinių rezultatų:

1. Vietos, fizinės aplinkos, panoramos, pramogų bei paslaugų kokybės aspektai yra pagrindiniai tarptautinių klientų patirties matmenys;
2. Dauguma klientų minėjo žodį „pabėgimas“, kadangi viešbutis turi nemažą, bet izoliuotą teritoriją, kurioje viešbučio svečiai gali vaikščioti be gido. Taip pat tokiu būdu nuo tarptautinių turistų yra „apsaugomi“ Šiaurės Korėjos gyventojai;
3. Dėl išėjimo iš viešbučio suvaržymų pramogos taip pat buvo vienas iš svarbių viešbučio savybių;
4. Sudarytas „gailėtis“ ir „džiaugtis“ savybių rinkinys:

    - viešbučio svečiai „gailėjosi“ laisvės trūkumu ir prabangaus viešbučio savybių pobūdžio;
    - svečiai „džiaugėsi“ turėję „ypatingą statusą“ – jie galėjo naudotis tomis viešbučio paslaugomis, kuriomis negalėjo naudotis Šiaurės Korėjos gyventojai.

Autoriai analizės trūkumų neišskiria, tačiau pateikia rekomendacijas – sudarytas „gailėtis“ ir „džiaugtis“ savybių rinkinys turėtų būti pagrindas tolimesniems tyrimams. Naudojantis šio rinkinio elementais, galėtų būti pritaikyta kiekybinė analizė. Taip pat tolimesniems tyrimams siūloma lyginti kinų bei vakarų turistų patirtį – taip būtų galima įžvelgti patirčiai daromą kultūrinį poveikį.

Jiao‘us ir Bai‘us [18] taiko empirinę privataus būsto nuomos klientų atsiliepimų sąrašo iš 40 Amerikos miestų analizę. Jie siekia ištirti, kaip demografiniai, socioekonominiai ir logistiniai aspektai gali paveikti atsiliepimų sąrašą iš 40 miestų Amerikoje.

Pradinį duomenų rinkinį sudarė 130097 privataus būsto nuomos svetainėje esantys įrašai iš 40 didžiųjų Amerikos miestų (duomenų laikotarpis 2017-01-01 – 2017-11-06), iš kurių atliekamai analizei panaudota 79198 įrašų. Atlikta aprašomoji duomenų analizė, taip pat daugialypis mišrių-efektų (angl. *multilevel mixed-effect model*) tiesinės regresijos modelis su programa STATA, pritaikyta ir svarbumo analizė (angl. *importance-performance analysis – IPA*).

Atlikus minėtus veiksmus, gauti tokie rezultatai:

1. Namų ūkio pajamos teigiamai susijusios su vidutine kaina (vienam žmogui vienai nakčiai);
2. Būsto vienetas (angl. *housing unit*) teigiamai koreliuoja su kaina, bet neigiamai su būstų tankumu teritorijoje;
3. Populiacijos tankis reikšmingai koreliuotas su vidutine kaina bei būstų tankumu, nors ryšys buvo gana silpnas;
4. Kuo būstas toliau nuo centro, tuo kaina bei būstų tankumas yra mažesni;
5. Geras susisiekimas bei daugiau restoranų ir barų, esančių netoliese, susijęs su aukštesne kaina bei didesniu šeimininkų skaičiumi teritorijoje.

Išnagrinėjus 2015 – 2020 m. laikotarpio mokslinius straipsnius viešbučių bei privataus būsto nuomos klientų patirties ir pasitenkinimo sąsajų didžiųjų duomenų kontekste tema, pastebėta, kad viešbučiai mokslinėje literatūroje pradėti nagrinėti gana seniai ir plačiai, o kitas apgyvendinimo – privataus būsto nuomos – atvejis ne taip seniai ir šioje srityje mokslinių straipsnių yra gerokai mažiau. Apibendrinantys požymiai, tinkantys abejoms sritims, yra pagrindiniai analizės įrankiai – dauguma

tyrimų naudoja teksto analizę, požymių dimensijos mažinimo metodus ir prognozavimo analitikos metodus. Vienas dažniausiai pasitaikiusių metodų yra tirti internetinėje svetainėje klientų paliktus įvertinimus bei atsiliepimus. Rekomendacijos tolimesniems tyrimams apima tokius pasiūlymus: naudoti daugiau skirtingų požymių; analizuoti daugiau nei vieną internetinę svetainę; pritaikyti daugiau / kitokius metodus bei tyrimo įrankius; didinti geografinę tyrimo teritoriją; bandyti sudaryti žodžių, pritaikytų informacijai iš atitinkamos internetinės svetainės, žodyną; plėtoti regresinę analizę įtraukiant naujus kintamuosius.

1.4. Tyrimuose naudojamų matematinių metodų ir programinės įrangos apžvalga

Mokslinės literatūros analizė viešbučių bei privataus būsto nuomos kontekste rodo, jog tinkamiausi bei rekomenduotini metodai, norint ištirti apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajas, yra teksto analizė, duomenų dimensijos mažinimo būdai bei prognozavimo analitikos metodai. Moksliniuose straipsniuose duomenys buvo vaizduojami įvairiais daugiamačių duomenų vizualizavimo metodais. Taip pat pastebėta, kad teksto analizės uždavinius galima spręsti naudojantis programomis SAS, R, SPSS, o viena dažniausiai analizei naudojamų programavimo kalbų yra Python. Šiame poskyryje minėti metodai ir programinė įranga aptariami plačiau.

1.4.1. Teksto analitikos metodai

Pastaruoju metu, teksto analitikos metodai taikomi gana dažnai bei yra itin populiarėjantys verslo aplinkoje, nes remiantis jais galima gauti inovatyvių rezultatų bei įžvalgų. Teksto analitiką sudaro septynios praktinės taikymo sritys [19]:

1. Paieškos ir informacijos gavyba (angl. *Search and information retrieval (IR)*) – tai tekstinių dokumentų tyryba ir kaupimas, įskaitant raktinių žodžių paiešką;
2. Dokumentų klasterizavimas (angl. *Document clustering*) – terminų, fragmentų, paragrafų ar dokumentų grupavimas ir kategorizavimas, naudojant duomenų tyrybos klasterizavimo metodus;
3. Dokumentų klasifikavimas (angl. *Document classification*) – terminų, fragmentų, paragrafų ar dokumentų grupavimas ir kategorizavimas, naudojant duomenų tyrybos klasifikavimo metodus, paremtus apmokytais pavyzdžiais;
4. Žiniatinklio tyryba (angl. *Web mining*) – internetinių duomenų ir teksto tyryba, ypatingą dėmesį skiriant svetainės dydžiui bei tarpusavio ryšiui;
5. Informacijos gavyba (angl. *Information extraction (IE)*) – nestruktūrizuoto teksto ryšių ir svarbių faktų identifikacija bei gavyba. Tai procesas, kai iš nestruktūrizuoto arba pusiau struktūrizuoto teksto gaunami struktūrizuoti duomenys;
6. NLP metodas (angl. *Natural Language Processing*)) – procesas, kai nestruktūrizuotam tekstui yra priskiriamos apibrėžtos žymos ar kategorijos, pavyzdžiui, kalbos aptikimas;
7. Pagrindinės temos gavyba (angl. *Concept extraction*) – žodžių ir jų junginių grupavimas į semantiškai panašias grupes.

Vienas iš teksto analitikos pogrupių yra teksto tyryba. Pastaroji yra naudinga norint iš tekstinio dokumento išgauti pagrindines temas ar žodžius. Ji skirstoma į dvi pagrindines grupes: tiriančiąją analizę, į kurią įeina temos gavyba, klasterinė analizė ir pan., bei sentimentų analizę, kuri gali būti apibūdinama kaip klasifikacijų analizė. Sentimentų analizė gali būti naudinga norint nustatyti bendrą tam tikro dokumento pobūdį. Jos rezultatas – dokumento įvertinimas pozityviu, negatyviu arba neutraliu.

Pagrindiniai teksto analizės etapai [19]:

1. **Duomenų surinkimas**. Tai duomenų, reikalingų analizei, rinkimo procesas;
2. **Teksto nagrinėjimas ir transformavimas**. Tai žodžių gavyba, jų valymas bei galutinio žodžių žodyno sudarymas naudojantis NLP metodais. Šis žingsnis apima nereikšminių žodžių (angl. *stop words*) pašalinimą, rašybos tikrinimą ir subjektų identifikavimą. Kitas itin svarbus žingsnis – teksto transformavimas. Jis paima tekstinio dokumento atvaizdavimą skaičiais, naudojantis tiesine algebra paremtais metodais (pavyzdžiui, latentinė semantinė analizė (angl. *LSA – latent semantic analysis*)) arba vektoriniu erdvės modeliu. Šio uždavinio rezultatas – dokumentų terminų matricos (angl. *term-by-document matrix*) sukūrimas;
3. **Teksto filtravimas**. Turint kelių tūkstančių dokumentų rinkinį, juose gali būti didelis kiekis nereikalingų terminų. Šių terminų ieškojimas ir šalinimas rankiniu būdu yra vienas iš svarbiausių ir ilgiausiai trunkančių žingsnių;
4. **Teksto tyryba**. Tai tradicinių teksto tyrybos metodų taikymas, pavyzdžiui, klasterizavimo, klasifikavimo, asociacijų analizės ir panašiai. Teksto tyrybą yra iteracinis procesas, susidedantis iš analizės kartojimo esant skirtingiems parametrams bei tam tikrų terminų įtraukimo / pašalinimo. Šio žingsnio rezultatas gali būti dokumentų klasteriai, vieno arba kelių terminų sąrašas, taip pat tam tikros taisyklės, atsakančios į klasifikavimo problemą.

1.4.2. Požymių erdvės dimensijos mažinimo metodai

Turint didelius tekstinių duomenų rinkinius, norima didelį kintamųjų kiekį sumažinti, juos išreiškiant mažesniais, kintamuosius apibendrinančiais elementais. Tokiu atveju naudojami požymių erdvės dimensijos mažinimo metodai – faktorinė arba pagrindinių komponenčių analizė.

Faktorinėje analizėje kintamieji $X_1, X_2, \dots, X_d$ yra atvaizduojami kelių atsitiktinių kintamųjų tiesinėmis kombinacijomis $F_1, F_2, \dots, F_m$ $(m < d)$, vadinamomis faktoriais, bendraisiais faktoriais arba latentiniais (paslėptaisiais) kintamaisiais [20]. Faktorinės analizės tikslas – sumažinti kintamųjų perteklių, išreiškiant juos mažesniu faktorių skaičiumi. Egzistuoja du faktorinės analizės tipai: tiriančioji ir patvirtinančioji [30]. Tiriančioji faktorinė analizė (angl. EFA – *exploratory factor analysis*) dažniausiai naudojama tiriant daugiamačius duomenis, norint nustatyti jų latentinę, arba paslėptąją, struktūrą. Pavirtinančioji faktorinė analizė (angl. CFA – *confirmatory factor analysis*) leidžia tyrėjui pačiam nuspėti latentinių kintamųjų skaičių bei jų struktūrą ir patikrinti suformuluotą hipotezę. Toliau apibūdinamas tiriančiosios faktorinės analizės metodas.

Faktorinė analizė susideda iš šių žingsnių [22]: 1) Duomenų tinkamumo faktorinei analizei tikrinimas; 2) Skaičiavimo metodo bei faktorių skaičiaus parinkimas; 3) Faktorių sukimas bei jų interpretavimas; 4) Faktorių reikšmių įverčių skaičiavimas.

Norint įsitikinti, kad faktorinė analizė yra tinkama turimiems duomenims, reikia apskaičiuoti du kriterijus: Bartlett‘o sferiškumo kriterijų (angl. *Bartlett‘s test of sphericity*), Kaiser-Meyer-Olkin (KMO) tinkamumo matą MSA_i (angl. *Measure of Sampling Adequacy*). Vėliau skaičiuojami faktorių svoriai (angl. *loadings*) bei bendrumo įverčiai (angl. *communality*), kuriuos galima gauti skirtingais faktorių išskyrimo būdais. Faktoriai bei juos atitinkantys svoriai ir bendrumo įverčiai gali būti apskaičiuojami pagrindinių komponenčių metodu, pagrindinių faktorių metodu, didžiausio tikėtinumo metodu, vaizdo faktorių išskyrimas ir kt.

Taikant faktorinę analizę, faktorių skaičius nėra iš anksto apibrėžiamas. Tyrėjas šį skaičių gali nustatyti analizės eigoje. Išskyrus kintamuosius apibūdinančius faktorius bei apskaičiavus kintamųjų svorius, neretai sunku gautus faktorius interpretuoti, dažnai juos sudaro daugiau, nei keli kintamieji. Tokiu atveju yra naudojamas faktorių sukimas, kuriuo siekiama faktorių svorius padaryti artimus 0, 1 arba −1. Nors po sukimo faktorių interpretacija pasidaro paprastesnė, bet sukimu yra pažeidžiama jų nepriklausomumo prielaida. Faktorių sukimai gali būti ortogonalieji (angl. *orthogonal rotation*), kai faktorių ašys išlaiko geometrinį statmenumą (*varimax, quartimax, equamax*), arba neortogonalieji (angl. *oblique rotation*), arba įstriži, kai gaunami tarpusavyje priklausomi faktoriai (*promax, oblimin, quartimin*).

Pagrindinių komponenčių analizė (toliau – PKA) yra tiesinė duomenų transformacija, naudojama daugiamačių duomenų sumažinimui iki kelių dimensijų [44]. Pagrindinė metodo idėja yra tai, kad dideli duomenų rinkiniai yra sudaryti iš koreliacijų tarp dimensijų, todėl dalis duomenų imties yra nereikalinga. PKA transformuoja duomenis taip, kad kiek įmanoma daugiau pradinių duomenų dispersijos būtų atvaizduota kuo mažesniu dimensijų kiekiu. Tai leidžia sumažinti duomenų kiekį ignoruojant kitas jos dimensijas.

PKA atliekama iš pradžių ieškant krypties, pagal kurią dispersija yra didžiausia [42]. Ši kryptis vadinama pirmąja pagrindine komponente. Pirmoji pagrindinė komponentė eina per duomenų centrinį tašką ir ji yra kiek įmanoma arčiau visų duomenų imties taškų. Antroji pagrindinė komponentė taip pat eina per centrinį duomenų tašką, tačiau ji turi būti statmena pirmajai komponentei. Taikant PKA, pagrindinių komponenčių skaičius nėra iš anksto apibrėžiamas, jis yra pasirenkamas tyrėjo analizės eigoje.

Nors faktorinės analizės metodas yra panašus į pagrindinių komponenčių analizės metodą, kadangi abu yra naudojami paprastesnei kintamųjų struktūrai gauti, tačiau jie skiriasi ne vienu aspektu. Keletas pagrindinių yra šie:

1. Pagrindinių komponenčių metodas išreiškia pagrindines komponentes tiesinėmis stebėjimų kombinacijomis. Faktorinės analizės metode stebėjimai yra išreiškiami tiesinėmis faktorių kombinacijomis;
2. Taikant pagrindinių komponenčių analizę, bandoma rasti didžiausias dispersijas turinčias duomenų projekcijas, o faktorinėje analizėje yra nusakomas stebėjimų modelis.

1.4.3. Prognozavimo analitikos metodai

Prognozavimo analitikos metodai skirti nustatyti ryšius tarp priklausomų ir nepriklausomų kintamųjų bei leidžia prognozuoti priklausomą kintamąjį [21]. Tokie metodai gali būti:

1. Reikšmių prognozavimas. Tai metodas, kai pagal nepriklausomus kintamuosius yra prognozuojama priklausomo kintamojo reikšmė;
2. Klasifikavimas. Naudojant šį metodą, stebiniai priskiriami tam tikroms, iš anksto nustatytoms kategorijoms. Šis metodas naujam įrašui nustato klasę ir tikimybę, kad konkretus įrašas priklauso tai klasei;
3. Laiko sekų nuoseklumų paieška. Šiuo metodu ieškoma laiko dėsningumų, taip pat jis apima prognozavimą laike ir prognostinių scenarijų tyrimą.

Tipinis uždavinys, sprendžiantis problemą, kai norima prognozuoti priklausomo kintamojo reikšmę bei nepriklausomais kintamaisiais bandoma paaiškinti priklausomojo kintamojo priklausomybę, yra tiesinė regresinė analizė. Kai nagrinėjamame modelyje yra vienas priklausomas bei daug nepriklausomų kintamųjų, tokia analizė vadinama daugialype tiesine regresine analize. Jos modelis apibrėžiamas taip [14, 18, 21, 23, 24]:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \varepsilon, \quad (1)$$

čia Y – priklausomas kintamasis, $X_1, \dots, X_K$ – nepriklausomi kintamieji (arba regresoriai), ε – liekamoji paklaida, t. y., tai, nuo ko dar gali priklausyti Y. Koeficientai $\beta_0, \dots, \beta_K$ nėra žinomi. Įverčiai $\hat{\beta}_0, \dots, \hat{\beta}_K$ gaunami panaudojant imties duomenis. Regresijos lygtis apytikslei $\hat{Y}$ reikšmei yra

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_K X_K. \quad (2)$$

Koeficientų $\hat{\beta}_0, \dots, \hat{\beta}_K$ ženklai aprašo, ar nepriklausomiems kintamiesiems didėjant $\hat{Y}$ didės, ar ne. Jei $\hat{\beta}_K > 0$, tai didėjant X_K, didėja ir $\hat{Y}$, o jei $\hat{\beta}_K < 0$, tai didėjant X_K, $\hat{Y}$ mažėja. Kiek pasikeis $\hat{Y}$ reikšmė, jei vienu vienetu padidintume X_1, o kitus regresorius fiksuotume, parodytų koeficientas $\hat{\beta}_1$.

Taikant tiesinės regresijos analizę, skaičiuojami ne tik modelio, bet ir standartizuotieji beta koeficientai. Modelį taikant skirtingoms regresorių matavimo skalėms, jos suvienodinamos skaičiuojant standartizuotąsias z reikšmes. Jos naudojamos nepriklausomų kintamųjų santykinės įtakos priklausomam kintamajam palyginimui. Kuo standartizuotasis beta koeficientas yra didesnis, tuo nagrinėjamame modelyje atitinkamo nepriklausomojo kintamojo įtaka didesnė.

Pagrindiniai daugialypės tiesinės regresinės analizės uždaviniai yra: 1) regresijos funkcijos analizinės išraiškos radimas; 2) regresijos funkcijos koeficientų taškinių ir intervalinių įverčių radimas; 3) hipotezių apie regresijos funkcijos koeficientus tikrinimas; 4) optimalios regresijos lygties sudarymas; 5) regresijos modelio prielaidų tikrinimas; 6) prognozavimo paklaidų vertinimas.

Taikant daugialypės tiesinės regresijos modelį, duomenys turi atitikti tam tikrus reikalavimus, o modelis – prielaidas. Norint patikrinti, ar modelio prielaidos yra tenkinamos, atsižvelgiama į [23] dispersijos mažėjimo daugiklį VIF (angl. *Variance Inflation Factor*), įtakos matą DfBetas_{ji}, Kuko įtakos matą (CooksD_i), Durbino-Vatsono statistiką, standartizuotąsios liekamąsios paklaidas, Šapiro-Vilko kriterijaus *p* reikšmę, liekamųjų paklaidų grafikus, standartizuotųjų prognozuojamų reikšmių ir liekamųjų paklaidų grafiką. Jei visos prielaidos yra tenkinamos, tai negarantuoja, jog modelis bus sudarytas tinkamai. Modelio tinkamumą aprašo šie rodikliai: 1) apibrėžtumo koeficientas R^2 (parodo, kokią kintamojo Y sklaidos apie vidurkį $\bar{Y}$ dalį galima paaiškinti tiesine regresija); 2) Koreguotas apibrėžtumo koeficientas R^2 (naudojamas daugialypėje tiesinėje regresinėje analizėje, nes, skaičiuojant jį, yra atsižvelgiama į imties dydį n ir regresorių skaičių K); 3) *ANOVA p* reikšmė (nurodo, ar yra regresorių, susijusių su priklausomu kintamuoju); 4) T (*Stjudento*) kriterijus atskiriems nepriklausomiems kintamiesiems (naudojamas norint nuspręsti, ar regresorius modelyje yra tinkamas, ar vis dėlto jis turėtų būtų pašalinamas).

Neretai praktikoje atsitinka taip, kad nėra tenkinamos visos, arba kai kurios, prielaidos. Tokiu atveju, galima alternatyva – neparametrinė ir netiesinė regresija. Ji skirstoma į:

1. Stabilizuotųjų liekamųjų paklaidų regresiją. Ji taikoma tada, kai yra pažeista homoskedastiškumo prielaida;
2. Atspariąją regresiją. Taikoma tada, kai duomenys turi išskirčių;
3. Kvantilių regresiją. Ji taikoma tada, kai normalumo prielaida yra pažeista;
4. Netiesinę regresiją. Taikoma tada, kai tarp kintamųjų yra stebimos netiesinės priklausomybės.

1.4.4. Daugiamačių duomenų vizualizavimo metodai

Duomenų vizualizavimas yra informacijos pateikimas grafiškai [42]. Pagrindinis jo tikslas – duomenis pateikti paprastesne, bet kuriam vartotojui labiau suprantama forma. Grafinis informacijos pateikimas gali palengvinti išvadų apie duomenis formulavimą, padėti pastebėti duomenų aibes bei poaibius. Tuo tarpu daugiamačių duomenų vizualizavimas gali padėti nustatyti duomenų klasterius, t. y., tam tikras duomenų grupes, pastebėti išskirtis, t. y., itin išsiskiriančius taškus, įvertinti panašumus tarp duomenų objektų, jų grupių ir t. t.

Daugiamačiai duomenys gali būti vaizduojami tiesioginio vizualizavimo metodais arba projekcijos (matmenų skaičiaus mažinimo) metodais.

1. Tiesioginio vizualizavimo metodais kiekvienas daugiamačio kintamojo parametras išreiškiamas tam tikru vizualiu pavidalu. Tokie metodai skirstomi į geometrinius metodus, simbolinius metodus bei hierarchinius metodus;
2. Projekcijos, arba duomenų mažinimo, metodai leidžia daugiamačius duomenis atvaizduoti mažesnio skaičiaus matmenų erdvėje. Jie skirstomi į tiesinės projekcijos metodus bei netiesinės projekcijos metodus.

1.4.5. Programinės įrangos apžvalga

Vienos dažniausiai naudojamų programų mokslinėje literatūroje yra SAS, R, SPSS, o programavimo kalba – Python. 1 lentelėje pateikiama minėtų programų ir programavimo kalbos palyginimas.

1 lentelė. Programinių įrangų palyginimas

Programa / programavimo kalba	**Paskirtis**	**Privalumai**
SAS	Duomenų analitikos programinė įranga, skirta duomenų valdymui, sudėtingesnei analitikai, daugiamatei analizei bei prognozavimo analitikai [28]. SAS yra viena iš plačiausiai naudojamų statistinių programinių įrangų.	Didžiulis statistikos ir duomenų tyrybos metodų ir algoritmų spektras, ypač orientuotas į sudėtingų duomenų analitikos metodų sritį. Didelis analizės bei išvesties pasirinkimų kiekis. Aukšta grafikos kokybė, tinkama publikacijoms. Plačiai naudojama tiek verslo, tiek medicinos ir kitose srityse; Gausi ir aktyvi internetinė bendruomenė [29].
R	Programavimo kalba bei aplinka, skirta statistiniam skaičiavimui bei grafiniam vaizdavimui [45]. R suteikia platų	Efektyvus duomenų apdorojimas ir saugojimas.

	statistinių (tiesinis ir netiesinis modeliavimas, klasikiniai statistiniai testai, laiko eilučių analizė, klasifikavimas, klasterizavimas ir t. t.) ir grafinio vaizdavimo technikų spektrą.	Operatorių, skirtų darbui su masyvais, rinkinys. Didelis įrankių rinkinys, skirtas duomenų analizei. Grafinės duomenų analizės technikos, rezultatų demonstravimas ekrane arba jų išvedimas į failą. Gerai išplėtota, paprasta ir efektyvi programavimo kalba, į kurią įeina sąlygos, ciklo ar naudotojo aprašytos funkcijos bei įvedimo ir išvedimo įrankiai [45].
SPSS	Programinė įranga, teikianti pažangią statistinę analizę, platų mašininio mokymo algoritmų kiekį, teksto analizę, atviro kodo išplečiamumą, didžiųjų duomenų integravimą ir sklandų programos diegimą [46].	Paprastas naudojimas ir lankstumas suteikia galimybę naudotis bet kokio pasirengimo vartotojams. Tinkamas įvairių dydžių ir sudėtingumų projektams, efektyvumo didinimui bei rizikų minimizavimui.
Python	Lengvai interpretuojama, į tikslą orientuota, aukšto lygio programavimo kalba su dinamiška semantika [27]. Python pasižymi paprasta, lengvai išmokstama sintakse, pabrėžiančia skaitomumą ir sumažinančia programos palaikymo kaštus. Įvairių programos paketų palaikymas didina programinio kodo pakartotinį panaudojamumą.	Dėl lengvos ir paprastos sintaksės, pradėti naudotis programavimo kalba Python nesunku tiek pradedančiajam, tiek jau pažengusiam programuotojui. Oficialioje Python svetainėje galima rasti daug naudojimosi vadovų, skirtų mokymuisi. Python bendruomenė rengia įvairius susitikimus, konferencijas ir bendradarbiauja tobulinant programinį kodą. Be to, visą informaciją galima rasti ir Python dokumentacijoje. „Python Package Index (PyPI)“ teikia tūkstančius trečiųjų šalių paketų. Tiek standartinė Python biblioteka, tiek bendruomenės kurti paketai suteikia be galo daug galimybių. Python yra sukurta pagal „OSI“ patvirtinto, laisvai prieinamo šaltinio licenciją, leidžiant ja nemokamai naudotis bei platinti net komerciniais tikslais. Python licencija yra administruojama „Python Software Foundation“ [26].

1.5. Darbo tikslo ir uždavinių pagrindimas

Išnagrinėjus mokslinę literatūrą, galima patvirtinti, jog apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų analizė yra svarbi sritis, suteikianti galimybę paslaugų teikėjams geriau suprasti klientų lūkesčius bei tobulinti aptarnavimo kokybę. Nors viešbučių sektorius pradėtas nagrinėti gana seniai, privataus būsto nuomos atvejis dar nėra taip plačiai išanalizuotas. Apskritai, modelių, paaiškinančių apgyvendinimo paslaugų klientų patirties ir pasitenkinimo ryšį, mokslinėje literatūroje dar trūksta. Iš to išplaukia pagrindinis šio darbo tikslas – išanalizavus privataus būsto svečių patirtį ir pasitenkinimą, pasiūlyti modelį, tinkantį jų sąsajoms įvertinti. Tikslui pasiekti reikia išspręsti įžangoje suformuluotus uždavinius.

2. Tyrimų metodai

Šiame skyriuje pateikta darbe pasiūlyta apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų modelio sudarymo metodika, apimanti duomenų paruošimą analizei, požymių atranką, naujų požymių kūrimą, požymių erdvės dimensijos mažinimą, apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų regresijos modelio sudarymą bei jo tikimo duomenims tyrimą. Metodika realizuota programiškai, panaudojus Python ir SAS programines įrangas.

2.1. Duomenų paruošimas ir požymių atranka

Duomenų paruošimo ir požymių atrankos etapai:

1. Duomenų rinkinio pasirinkimas. Pasirenkamas ir apibūdinamas analizuojamas duomenų rinkinys ir atliekama tiriamoji analizė;
2. Požymių atranka. Atrenkami ir apibrėžiami analizei reikalingi požymiai. Pagrindiniai iš jų yra klientų pateikti atsiliepimai bei įvertinimai. Atlikus literatūros analizę pastebėta, kad būsto vertinimas yra laikomas požymiu, rodančiu kliento pasitenkinimą. Kiti požymiai gali būti nuomos data, nuomos trukmė, būstą apibūdinantys kintamieji (pavyzdžiui, apartamento ar kambario tipas, tiksli jo vieta, kaina) bei šeimininką apibūdinantys kintamieji (pavyzdžiui, šeimininko vardas ar jo aktyvumą svetainėje nusakantys požymiai). Su apgyvendinimo paslaugų kliento patirtimi ir pasitenkinimu susiję požymiai apjungiami į vektorių:

$$v_i = (a_{i1}, \dots, a_{iK}, \omega_i), i = \overline{1,N}, j = \overline{1,K} \quad (3)$$

   čia a_{ij} – kiekybiniai arba kokybiniai požymiai, apibūdinantys klientą, nuomojamą būstą, būsto šeimininką ir t. t., ω_i – kliento atsiliepimas (tekstas, nestruktūrizuoti duomenys), N – analizuojamas klientų atsiliepimų skaičius, K – požymių skaičius. Iš vektorių v_i sudaroma pradinė stebinių matrica:

$$U_1 = \begin{pmatrix} a_{11} & \dots & a_{1K}\,\omega_1 \\ \vdots & \vdots & \vdots \;\; \vdots \\ a_{N1} & \dots & a_{NK}\omega_N \end{pmatrix}, i = \overline{1,N}, j = \overline{1,K} \quad (4)$$

3. Duomenų tvarkymas. Gali būti, kad skaitines reikšmes turintys požymiai yra priskiriami ne skaitiniam tipui (pavyzdžiui, kainą atitinkantis požymis gali būti sudarytas iš skaičiaus ir valiutą nurodančio ženklo, todėl priskiriamas simbolinis tipas). Tokie požymiai turi būti perkoduojami į reikiamą tipą. Taip pat dažnai kiekybinių požymių nulinėms reikšmėms turi būti priskirtas trūkstamos reikšmės kodas. Analizuojant požymių reikšmes, kainos ar būsto vertinimo požymio reikšmė lygi nuliui. Vienu atveju nulinės reikšmės gali reikšti įvedimo klaidą, kitu atveju – trūkstamą informaciją ir turi būti pašalintos. Suteikus trūkstamas reikšmes, iš stebinių matricos šalinami visi stebiniai, kur bent vienas požymis įgyja trūkstamą reikšmę. Taip pat atliekamas kokybinių kintamųjų tekstinių reikšmių perkodavimas į skaitines.
4. Naujų požymių kūrimas. Priklausomai nuo turimos informacijos, į analizę gali būti įtraukiami nauji požymiai, tikintis, kad jie pagerins apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų tyrimo modelį. Pavyzdžiui, turint kliento atsiliepimą, galima įvesti naują požymį, nurodantį atsiliepimo žodžių skaičių arba turint informaciją apie būsto vietą, t. y., geografinę ilgumą ϕ bei platumą ψ, ją galima panaudoti suskaičiuojant tikslų atstumą iki miesto, kuriame nuomojamas būstas, centro. Šį atstumą galima apskaičiuoti pagal Haversino formulę

[36], kuri leidžia apskaičiuoti atstumą tarp dviejų taškų sferoje, turint tų taškų platumos ir ilgumos koordinates:

$$l = 2r \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\psi_2 - \psi_1}{2}\right)}\right), \quad (5)$$

čia l – atstumas tarp dviejų taškų su platumos ir ilgumos koordinatėmis (ψ, ϕ), r – Žemės spindulys. Šiame etape sudaroma stebinių matrica U_2 su naujai įvestais požymiais b_{NS}:

$$U_2 = \begin{pmatrix} a_{11} & \dots & a_{1K} & b_{11} \dots & b_{1S} & \omega_1 \\ \vdots & \vdots & \vdots & \vdots \vdots & \vdots & \vdots \\ a_{N1} & \dots & a_{NK} & b_{N1} \dots & b_{NS} & \omega_N \end{pmatrix}, i = \overline{1,N}, j = \overline{1,K}, s = \overline{1,S} \quad (6)$$

čia a_{ij} – pradiniai požymiai, b_{is} – naujai sukurti požymiai, ω_i – nestruktūrizuotas tekstinis požymis.

5. Stebinių matricos U_2 pjūvio formavimas. Šiame žingsnyje vykdomas filtravimas pagal atitinkamus požymius. Tarkime, pasirinktas duomenų rinkinys yra sukauptas per ilgą laikotarpį, o norima analizuoti tik naujausią informaciją, tokiu atveju, filtruojama pagal datos požymį.
6. Požymių tiriamoji analizė. Šiame žingsnyje grafiškai analizuojama bendro būsto vertinimo priklausomybė nuo atitinkamų požymių. Tai reikalinga tam, kad pradiniame analizės etape būtų galima pastebėti galimas išskirtis ar nelogiškas reikšmes. Pavyzdžiui, jei kainos požymis įgyja reikšmes, mažesnes už 10 arba kitą nustatytą reikšmę, laikoma, kad požymio reikšmės įvestos neteisingai, nes šio požymio reikšmės parodo konkretaus būsto nuomos kainą už naktį.

2.2. Teksto tyryba

Teksto tyrybos metodai taikomi analizuojant klientų atsiliepimus. Analizės tikslas – sudaryti klientų patirtį atspindinčių žodžių sąrašą, kuris kitame etape naudojamas apgyvendinimo paslaugų klientų patirties ir pasitenkinimo modelio naujų požymių sudarymui. Darbe analizuojami tik anglų kalba parašyti atsiliepimai.

1. Nustatoma atsiliepimo kalba. Nustačius atsiliepimo kalbą, sukuriamas naujas požymis *atsiliepimo kalba* ir pašalinamos visos stebinių matricos eilutės, kuriose kalba yra ne anglų;
2. Atsiliepimo požymio paruošimas. Iš atsiliepimo pašalinami įvairūs tyrimui nereikšmingi tarnybiniai simboliai. Pavyzdžiui, klaustukas, šauktukas, kablelis, taškas ir t. t., taip pat sakinių kėlimas į kitą eilutę, dvigubi (ar didesni) tarpai tarp žodžių yra keičiami į vieną;
3. Nereikšminių žodžių sąrašas. Sudaromas nereikšminių žodžių sąrašas, kuris susideda iš: nereikšminių angliškų žodžių sąrašo [34] (tai gali būti, pavyzdžiui, sąrašas, pateiktas literatūros šaltinyje [33]), kitų papildytų žodžių (tai gali būti, pavyzdžiui, tinklalapio pavadinimas ir t. t.). Jei turima informacija apie atitinkamo būsto šeimininką, būsto pavadinimą ar pan., šie žodžiai taip pat įtraukiami į sąrašą. Visi žodžiai, įtraukti į nereikšminių žodžių sąrašą, yra pašalinami iš atsiliepimo;
4. Pritaikoma sentimentų analizė. Apskaičiuojamas atsiliepimo sentimento įvertis, įgyjantis reikšmes nuo −1 (labai neigiamas atsiliepimas) iki 1 (labai teigiamas atsiliepimas);
5. Ne angliškų žodžių paieška. Sudaromas atskirų žodžių masyvas ir jame darkart ieškoma ne angliškų žodžių – tai gali būti įvairūs klaidingai įvesti žodžiai. Pavyzdžiui, kelios raidės sukeistos

vietomis (vietoje žodžio „night“, parašytas žodis „nihgt“), rašybos klaida (vietoj žodžio „little“, parašytas žodis „litle“) ir t. t. Tokie žodžiai yra pašalinami;
6. Prasminių žodžių sąrašo tvarkymas. Gavus atskirų žodžių sąrašą, skaičiuojami jų dažniai;
7. Kliento patirtį apibūdinančių žodžių sąrašo sudarymas. Sudaromas išvalytas atskirų žodžių ir jų dažnių sąrašas, kuris turi būti peržiūrimas ir paliekami tik žodžiai, kurie atspindi kliento patirtį. Šiame etape būtina remtis moksline literatūra [4, 8]. Žingsnyje nenagrinėjami retai pasikartojantys žodžiai, kurie pasikartoja retai, t. y., sudaro pasirinktą procentinę visų žodžių dalį, pavyzdžiui, 5 proc. Tada iš tolimesnės analizės pašalinami stebiniai, kuriuose nėra nei vieno žodžio iš sudaryto, klientų patirtį atspindinčio sąrašo. Tokiu būdu nenagrinėjami atsiliepimai, kuriuose „nekalbama“ apie nuomos metu gautą patirtį.

Vienas iš teksto transformavimo į kiekybinę matricą „atsiliepimai-požymiai“ būdų yra vektorinės erdvės modelis (angl. *vector space model*) [35]. Statistiškai grįstame vektorinės erdvės modelyje dokumentas yra vaizduojamas vektoriumi, kurį sudaryto raktiniai žodžiai, išgauti iš dokumento, ir jų svoriais, atitinkančiais svarbą tiek dokumente, tiek visame dokumentų rinkinyje. Raktinio žodžio svoris dokumente gali būti apibrėžiamas keletu būdų. Vienas dažniausių yra $tf \times idf$ metodas, kuris žodžio svorį apibrėžia dvejais daugikliais:

- Žodžio j dažnis i-tame atsiliepime (tf_{ij});
- Žodžio j dažnis visame atsiliepimų rinkinyje (df_j).

Žodžio j svoris i atsiliepime yra išreiškiamas formule

$$c_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log\frac{N}{df_j}\ , \tag{7}$$

čia N – atsiliepimų skaičius, idf_j – atvirkštinis dažnis.

Atlikus klientų atsiliepimų rinkinio analizės ir transformavimo žingsnius, gaunamas rezultatas:

$$D = \begin{pmatrix} c_{11} & \dots & c_{1L} \\ \vdots & \vdots & \vdots \\ c_{N1} & \dots & c_{NL} \end{pmatrix}, i = \overline{1,N}, j = \overline{1,L} \tag{8}$$

čia D – atsiliepimų-požymių matrica, N – atsiliepimų skaičius, L – požymių (žodžių $C_1, C_2, \dots, C_L$) skaičius.

2.3. Požymių dimensijos mažinimas

Faktorinė analizė taikoma duomenų dimensijos mažinimui, išskiriant mažesnį, tiesiogiai nestebimų (latentinių) faktorių kiekį. Faktorinės analizės modelis aprašo tiesinę priklausomybę tarp faktorių bei stebimų kintamųjų [20, 22, 30]:

$$\begin{matrix} C_{11} = \lambda_{11}F_1 + \dots + \lambda_{1\mathrm{M}}F_M + \varepsilon_1, \\ \vdots \\ C_{LM} = \lambda_{\mathrm{L}1}F_1 + \dots + \lambda_{LM}F_M + \varepsilon_L. \end{matrix} \quad i = \overline{1,L}, j = \overline{1,M} \tag{9}$$

Daugikliai λ_{ij} yra vadinami faktorių svoriais (angl. *loadings*), o kintamieji ε_i – charakteringaisiais faktoriais. Apibrėžiame faktorinės analizės modelio kintamųjų dispersijas ir kovariacijas:

$$cov(C_i, C_j) = \lambda_{i1}\lambda_{j1} + \cdots + \lambda_{iM}\lambda_{jM}, \qquad i \neq j \tag{10}$$

$$\mathbb{D}C_i = \lambda_{i1}^2 + \cdots + \lambda_{iM}^2 + \varphi_i \tag{11}$$

$$cov(C_i, F_j) = \lambda_{ij}. \tag{12}$$

Taikant faktorinės analizės metodą, reikia atlikti duomenų tinkamumo tikrinimą.

1. Apskaičiuojami tinkamumo kriterijai. Skaičiuojamas bendras KMO rodiklis visiems duomenims, MSA_i kiekvienam požymiui (žodžiui) atskirai bei Bartlett'o sferiškumo kriterijus.

   - KMO matas. Šiuo matu yra apskaičiuojamas empirinių koreliacijos koeficientų ir dalinių koreliacijos koeficientų palyginimas indeksas. Kuo šis indeksas yra artimesnis vienetui, tuo geriau. Faktorinė analizė yra nepriimtina, jei KMO yra mažesnis už 0,5. Norint išmatuoti kiekvieno kintamojo stebėjimų tinkamumo matą MSA_i, taikoma formulė:

$$MSA_i = \frac{\sum_{j \neq i} r_{ij}}{\sum_{j \neq i} r_{ij} + \sum_{j \neq i} \tilde{r}_{ij}}, \tag{13}$$

     čia r_{ij} – empirinis požymių C_i, C_j koreliacijos koeficientas, o $\tilde{r}_{ij}$ – dalinės koreliacijos koeficientas. Požymis laikomas netinkamu faktorinei analizei, jei jo MSA_i < 0,5. Tokius poyžmius (žodžius) rekomenduojama šalinti. Taigi, tolimesnėje analizėje turi būti paliekami tik tie požymiai, kurių MSA_i rodiklis yra nemažesnis nei 0,5.
   - Bartlett'o sferiškumo kriterijus. Faktorinė analizė yra tinkama tik tada, kai stebimi kintamieji yra tarpusavyje koreliuoti. Bartlett'o sferiškumo kriterijumi yra tikrinama hipotezė, jog kintamųjų koreliacijų matrica yra vienetinė.

   Jei duomenys yra tinkami, toliau taikoma faktorinė analizė.
2. Nustatomas faktorių skaičius. Faktorių skaičius gali būti parenkamas pagal tokius kriterijus:

   - Faktorių skaičius yra parenkamas toks, kad paaiškintų iš anksto apsibrėžiamą dispersijos dalį, pavyzdžiui, 80 %;
   - Faktorių skaičius gali būti lygus tikrinių reikšmių, didesnių už jų vidurkį, skaičiui;
   - Faktorių skaičius parenkamas pagal tikrinių reikšmių grafiką. Nustatomas grafiko lūžio taškas („alkūnės metodas").

   Šie būdai padeda nustatant faktorių skaičių, tačiau šiame žingsnyje galutinį jų skaičių parenka pats tyrėjas, atsižvelgdamas į gautų faktorių interpretavimo galimybes.
3. Faktorių išskyrimo metodas. Nusprendus, kokį faktorių skaičių išskirti, pasirenkamas faktorių išskyrimo metodas.
4. Nustatomas požymių priskyrimo faktoriams svorio slenkstis. Teorijoje apibrėžiama, kad faktorius laikomas susijusiu su požymiais tada, kai jų svorių įverčiai absoliučiu didumu yra ne mažesni nei 0,4. Tuo tarpu mokslinėje literatūroje aprašytuose tyrimuose rasta, jog autoriai, taikydami faktorinės analizės metodą klientų patirties tyrimo srityje, nurodo, kad požymių svoris turėtų būti daugiau už 0,2 arba 0,3. Toks slenkstis pasirenkamas tam, kad būtų įtraukta kuo daugiau požymių.
5. Faktorių interpretacija. Gauti faktoriai $F_1, \ldots, F_M$ ir juos atitinkantys požymiai $C_1, \ldots, C_L$. Tikrinama, ar galima faktorius interpretuoti. Jei rezultatai netenkina, galima bandyti pasirinkti kitą faktorių išskyrimo metodą. Taip pat jei gaunama, kad tam tikras požymis priskiriamas keliems faktoriams, naudojamas faktorių sukimas – jis turi būti apibrėžiamas faktorių išskyrimo žingsnyje. Dėl to, jog faktorinės analizės rezultatas vėliau bus naudojamas regresinėje analizėje,

reikėtų rinktis vieną iš ortogonaliųjų sukimo metodų – tokiu būdu bus tenkinama viena iš regresijos prielaidų, t. y., požymiai yra nepriklausomi.

6. Paaiškinama požymių dispersijos dalis. Gavus interpretuojamus faktorius, skaičiuojama, kokią dispersijos dalį paaiškina gauti faktoriai. Teorijoje apibrėžiama, kad paaiškinama dispersijos dalis turėtų būti ne mažiau nei 60-70 %, tačiau mokslinės literatūros analizė parodė, kad straipsniuose pateikiami tyrimai, kuriuose faktoriai paaiškina apie 20-25 % dispersijos dalies.
7. Faktorių svorių skaičiavimas. Atlikus visus žingsnius, kiekvienam atsiliepimui apskaičiuojami faktorių svoriai, naudojami kitame žingsnyje, t. y., sudarant regresijos modelius.

Po šio žingsnio gaunama stebinių matrica U_3:

$$U_3 = \begin{pmatrix} a_{11} & \dots & a_{1K} & b_{11} & \dots & b_{1S} & F_{11} & \dots & F_{1M} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{N1} & \dots & a_{NK} & b_{N1} & \dots & b_{NS} & F_{N1} & \dots & F_{NM} \end{pmatrix} \quad (14)$$
$$i = \overline{1,N}, j = \overline{1,K}, s = \overline{1,S}, m = \overline{1,M}$$

čia U_3 – stebinių matrica, a_{ij} – pradiniai požymiai, b_{is} – naujai įvesti požymiai, F_{im} – požymiai (faktoriai), gauti atlikus faktorinę analizę teksto tyrybos žingsnyje išskirtiems požymiams (žodžiams).

2.4. Klientų patirties ir pasitenkinimo sąsajų tyrimo modelis

Šiame poskyryje pateikta apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų regresijos modelio sudarymo ir jo tikimo analizuojamiems duomenims tyrimo metodika, kuri realizuota programiškai, panaudojus Python ir SAS programines įrangas.

2.4.1. Tiesinės regresijos taikymas

Atsižvelgiant į mokslinės literatūros analizę, pritaikomas vienas iš gana dažnai rekomenduojamų metodų apgyvendinimo paslaugų klientų patirties bei pasitenkinimo ryšio tyrimui – regresinė analizė. Jos nepriklausomi kintamieji bus ankstesniuose žingsniuose išskirti požymiai (kintamieji), o priklausomas kintamasis Y – klientų nuomojamo bendras būsto vertinimas.

Sudaromas pradinis daugialypės tiesinės regresinės analizės modelis:

$$Y = \beta_0 + \beta_1 a_1 + \dots + \beta_K a_K + \beta_{K+1} b_1 + \dots + \beta_{K+S} b_S + \beta_{K+S+1} F_1 + \dots + \beta_{K+S+M} F_M + \varepsilon, \quad (15)$$

čia Y – priklausomas kintamasis, $a_1, b_1, F_1, \dots, a_K, b_S, F_M$ – nepriklausomi kintamieji, ε – liekamoji paklaida. Daugialypes tiesinės regresijos lygtis apytikslei $\hat{Y}$ reikšmei:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 a_1 + \dots + \hat{\beta}_K a_K + \hat{\beta}_{K+1} b_1 + \dots + \hat{\beta}_{K+S} b_S + \hat{\beta}_3 F_1 + \dots + \hat{\beta}_{K+S+M} F_M. \quad (16)$$

Taikant daugialypės tiesinės regresijos modelį, duomenys turi atitikti tam tikrus reikalavimus, o modelis – prielaidas [23]:

1. Duomenys turi būti kiekybiniai, išmatuoti intervalų arba santykių skalėje, o jei yra kokybinių kintamųjų – jie turi būti dvireikšmiai;
2. Priklausomas kintamasis ir nepriklausomi kintamieji, išskyrus dvireikšmius, yra intervaliniai, o liekamosios paklaidos – normaliai pasiskirsčiusios;

3. Neretai, norint kintamuosius (tiek priklausomą, tiek nepriklausomus) naudoti regresinės analizės modeliuose, jie transformuojami taikant ištiesinimo metodą, pavyzdžiui, juos logaritmuojant;
4. Siekiant pagerinti regresijos modelio tikslumą, į modelį gali būti įtraukiami dvireikšmiai kategoriniai kintamieji. Tokie kintamieji dar vadinami pseudokintamaisiais. Kategoriniai kintamieji, dažnai įtraukiami į regresijos modelį, yra lytis, gyvenamoji vieta, tautybė, politinės pažiūros ir t. t. Jei kategorinis kintamasis turi $n > 2$ kategorijų, jis keičiamas į $n - 1$ dvireikšmių (pseudo) kintamųjų;
5. Kintamųjų liekamosios paklaidos ε_i turi būti nekoreliuotos;
6. Negali būti stiprios koreliacijos tarp regresorių. Jei yra priešingai, sakoma, kad kintamieji yra multikolinearūs. Kuo imtis didesnė, tuo multikolinearumas gali daryti mažiau įtakos, tačiau ryšiai tarp regresorių gali tapti nelogiški;
7. Duomenyse nėra išskirčių. Tai reikšmės, stipriai besiskiriančios nuo kitų tiriamo kintamojo reikšmių. Sudarytas modelis, kai duomenyse yra išskirčių, tampa nepatikimu;
8. Duomenys yra homoskedastiški. Tai reiškia, kad liekamosios paklaidos ε_i sąlyginė dispersija yra pastovi. Jeigu yra priešingai, sakoma, kad duomenys yra heteroskedastiški. Modelis, kai duomenys yra heteroskedastiški – nepatikimas, tada reikia taikyti heteroskedastiškumui atsparius regresijos modelius.

Apibrėžus nepriklausomus kintamuosius, reikalavimus duomenims ir prielaidas modeliui, toliau sudarinėjamo daugialypės tiesinės regresijos modelio eiga yra nurodyta literatūros šaltinyje [23].

Analizuojant labai didelius duomenų masyvus, regresinės analizės prielaidų tikrinimui tikslinga atrinkti atsitiktinę imtį, nes tikrinat hipotezes jos bus atmetamos ir esant labai mažiems jų pažeidimams. Prielaidos tikrinamos naudojant šiuos kriterijus:

- dispersijos mažėjimo daugiklis VIF (angl. *Variance Inflation Factor*). Juo matuojama, ar yra multikolinearumo problema.
- įtakos matas $\mathrm{DfBetas}_{ji}$. Jis parodo, kokią įtaką *i*-tasis stebėjimas daro regresijos koeficientui β_j. Norint nustatyti išskirtis, naudojami standartizuotieji $\mathrm{DfBetas}_{ji}$.
- Kuko įtakos matas (CooksD_i). Pašalinus *i*-tąjį stebėjimą, šis matas parodo prognozės pokytį.
- standartizuotosios liekamosios paklaidos. Šios paklaidos naudojamos priklausomojo kintamojo Y normalumo tikrinimui.
- Šapiro-Vilko kriterijaus p reikšmė. Šis kriterijus yra naudojamas liekanų normalumo tikrinimui. Jei nagrinėjama imtis yra didelė, šis kriterijus gali nepagrįstai atmesti hipotezę apie normalumą, todėl tokiu atveju geriau ją tikrinti grafiškai.
- liekamųjų paklaidų grafikai. Kiekvienam regresoriui yra braižomi liekamųjų paklaidų grafikai, parodantys regresorių reikalingumą modelyje.
- standartizuotųjų prognozuojamų reikšmių ir liekamųjų paklaidų grafikas. Braižomas grafikas, kai *y* ašyje yra atidėtos standartizuotosios liekamosios paklaidos, pagal kurias yra tikrinamas homoskedastiškumas.

Taikant prognozavimo analitikos metodus praktikoje, neretai pasitaiko, jog ne visos prielaidos yra tenkinamos. Tokiu atveju, taikomos tiesinės regresijos alternatyvos. 2.4.2 skyrelyje plačiau aptariama atsparioji regresija.

2.4.2. Tiesinės regresijos alternatyvos taikymas

Atsparioji regresija (angl. *robust regression*) yra tinkama tada, kai tiriami duomenys turi išskirčių. Dažniausiai ji taikoma, kai [23]:

- taikant tiesinės regresijos modelį, nustatomos išskirtys (pavyzdžiui, pagal Kuko, $DfBetas_{ji}$ matus);
- manoma, kad išskirtys yra atsitiktinės – t. y., nežinoma, dėl kokios priežasties jos atsirado;
- rankiniu būdu šalinti išskirtis yra labai sudėtinga (pavyzdžiui, analizuojant didelius duomenų kiekius).

Pagrindinis atspariosios regresijos taikymo tikslas yra aptikti išskirtis ir pateikti atsparius išskirtims (t. y., stabilius) rezultatus [37]. Prieš atliekant regresinę analizę, atsparioji regresija nustato:

- išskirtis priklausomo kintamojo *Y* kryptimi;
- daugiamates išskirtis nepriklausomų kintamųjų *X* erdvėje (dar vadinamos įtakos (angl. *leverage*) taškais);
- išskirtis tiek *Y* kryptimi, tiek *X* erdvėje.

Šioms problemoms spręsti siūloma nemažai metodų, tačiau išskirčių nustatymo bei atspariosios regresijos taikymo statistiniuose tyrimuose dažniausiai taikomi šie metodai:

1. M įvertinys, kurį pasiūlė Huber‘as [38], yra vienas paprasčiausių metodų. Nors jis nėra „atsparus“ įtakos taškams, tačiau plačiai taikomas, kai daroma prielaida, jog išskirtys egzistuoja priklausomo kintamojo *Y* kryptimi;
2. Mažiausių apkarpytų kvadratų metodo įvertinys (angl. *least trimmed squares – LTS*) pasiūlytas Rousseeuw‘o [39];
3. S įvertinys, kurį pasiūlė Rousseeuw‘as ir Yohai [40]. S įvertinys dažnai yra efektyvesnis už mažiausių apkarpytų kvadratų įvertinį;
4. MM įvertinys, pasiūlytas Yohai [41]. Jis apjungia mažiausią apkarpytą kvadratų įvertinį bei M įvertinius ir yra už juos efektyvesnis.

Atsparios išskirtims regresijos taikymo etapai:

1. Sudaromas atspariosios regresijos modelis. Sudarinėjant modelį, taikomas mažiausių apkarpytų kvadratų metodas (LTS), tam, kad būtų prarandama kuo mažiau informacijos;
2. Matricai U_3 gaunamas išskirčių ir įtakos taškų sąrašas, apskaičiuojamas Mahalanobio atstumas (angl. *Mahalanobis distance*) bei standartizuota atsparioji liekana (angl. *Standardized Robust Residual*);
3. Grubiausių pažeidimų identifikavimas. Pagal stebėjimo numerį sujungiamos išskirčių ir įtakos taškų bei pradinių duomenų matricos. Nustatomos ir pašalinamos grubiausios išskirtys;
4. Pakarotinai sudaromas tiesinės regresijos modelis.

2.5. Metodikos programinė realizacija

Šiame poskyryje aprašyta pasiūlytos metodikos programinė realizacija. Kiekviename žingsnyje nurodoma jo įvestis bei išvestis.

Pirmi keturi žingsniai realizuoti sukuriant Python kodą ir naudojant „Jupyter Notebook“ (laisvai prieinama internetinė programa, leidžianti jos vartotojui kurti ir dalintis dokumentais, sudarytais iš tiesiogiai koreguojamo kodo, lygčių, vizualizacijų ir aprašomojo teksto [25], buvo naudojama 6.0.3 „Jupyter Notebook“ versija). 1 priede pateikiamas visų reikalingų paketų bei naudojamų formulių diegimo programos kodas.

Penktas žingsnis realizuotas naudojant programinę įrangą SAS 9.4.

1. Duomenų paruošimas. Duomenų failo nuskaitymo ir apdorojimo žingsnis. Šio žingsnio įvestis – duomenų matrica U_1, kurioje vienas iš požymių yra klientų atsiliepimai, kitas – jų įvertinimai bei kiti požymiai, susiję su nuomojamu būstu, jo šeimininku ir t. t. Atliekama požymių atranka, naujų požymių kūrimas, požymių tipų perkodavimas, filtravimas, trūkstamų reikšmių šalinimas. Išvestis – duomenų matrica U_2 su pasirinktais bei naujai sukurtais požymiais, naudojamais tolimesniuose žingsniuose. Žingsnis atliekamas sukurtu Python kodu, kuris yra pateiktas 2 priede.
2. Duomenų matricoje U_2 esančių klientų atsiliepimų analizė ir transformavimas. Šalinami stebiniai, kuriuose atsiliepimas parašytas kirilicos, graikų, arabų abėcėlėmis arba hieroglifais, šalinami įvairūs nereikalingi simboliai, taip pat nustatoma kiekvieno iš atsiliepimų kalba, sudaromas nereikšminių žodžių sąrašas, pritaikoma sentimentų analizė. Atliekamas atsiliepimų analizė, sudaromas kliento patirtį atspindinčių žodžių sąrašas. Žingsnio įvestis yra duomenų matrica U_2 su pasirinktais ir naujai sukurtais požymiais, išvestis – atsiliepimų-požymių matrica *D*. Žingsnis atliekamas sukurtu Python kodu, kuris yra pateiktas 3 priede.
3. Matricos *D* dimensijos mažinimas. Naudojant atsiliepimų-požymių matricą *D* ir taikant faktorinę analizę, mažinama matricos *D* dimensija išskiriant apibendrintus faktorius. Žingsnio išvestis – duomenų matrica U_3, papildyta naujais požymiais. Žingsnis atliekamas sukurtu Python kodu, kuris yra pateiktas 4 priede.
4. Tiesinės regresijos ir atspariosios regresijos taikymas. Standartizuotiems pradiniams bei naujai įvestiems požymiams taikoma tiesinė regresinė analizė. Tikrinant prielaidas, atrenkama atsitiktinė imtis. Pritaikomas atspariosios regresijos metodas, žingsnio išvestis yra išskirčių ir įtakos taškų sąrašas. Jis vėliau naudojamas tiesinės regresijos modeliui pagerinti. Tiesinės regresinės modelio analizė kartojama po atspariosios regresijos taikymo ir grubiausių išskirčių šalinimo. Šis žingsnis atliekamas SAS programos kodu, kuris yra pateiktas 5 priede.

3. Tyrimų rezultatai

Sukurta metodika ir programinės priemonės šiame skyriuje pritaikoma realiems duomenims. Tiriami du „Airbnb“ atvejai – duomenų rinkiniai yra sudaryti iš klientų, kurie buvo apsistoję tam tikruose miestuose ir po viešnagės privataus būsto nuomos tinklalapyje „Airbnb“ parašė atsiliepimą. „Airbnb“ yra viena didžiausių prekyviečių pasaulyje, siūlanti daugiau nei 7 milijonus unikalių bei autentiškų apgyvendinimo vietų, nuomojamų vietinių šalies gyventojų. „Airbnb“ skatina žmonių tarpusavio ryšį, bendruomeniškumą ir pasitikėjimą bei yra prieinama 62 kalbomis per daugiau nei 220 šalių bei regionų [47]. Naudojantis „Airbnb“ tinklalapiu galima tiek nuomotis privatų būstą, tiek siūlyti savo būstą nuomai.

Analizei pasirinkti Europos žemyno miestai, nes mokslinėje literatūroje daugiausia sutinkamų tyrimų nagrinėja JAV, Kinijos bei Australijos regionus. Duomenys paimti iš tinklalapio „InsideAirbnb“ [31] – jame galima rasti skirtingų miestų informaciją, gautą iš viešai prieinamo „Airbnb“ tinklalapio. Analizuojami du duomenų rinkiniai, sudaryti iš klientų atsiliepimų, kurie buvo apsistoję Vokietijos miestuose (Berlyno ir Miuncheno) bei Ispanijos miestuose (Madrido ir Barselonos). Abu atvejai tiriami, kai duomenų laikotarpis yra nuo 2015-01-01 iki 2020-01-01. Analizėje nagrinėjami tik apartamento tipo būstai, kai būstas nuomojamas ne daugiau nei 5 naktų (trumpalaikė nuoma). Tyrimas atliktas remiantis 2 skyriuje pateikta metodika ir jos programine realizacija.

3.1. Vokietijos „Airbnb“ atvejo analizė

Pasirinkti Berlyno bei Miuncheno miestuose nuomojamų būstų klientų atsiliepimų sąrašai – *listings.csv.gz* ir *reviews.csv.gz* aplankai:

- aplanke *listings.csv.gz* yra *csv* formato failas su nuomojamų būstų sąrašu bei su juo susijusi informacija (šeimininko informacija, būsto informacija, nuomos sąlygos ir pan.);
- aplanke *reviews.csv.gz* yra *csv* formato failas su atsiliepimų sąrašu bei jų autoriais.

Turint šiuos keturis duomenų failus pasirinktiems miestams (po du *csv* failus konkrečiam miestui), jie apjungiami. *Listings.csv* ir *reviews.csv* sujungiami į vieną failą pagal *id* (iš *listings.csv*) bei *listings_id* (iš *reviews.csv*) konkretiems miestams. Gaunama, kad Berlyno duomenų imtį sudaro 545703 stebinių, o Miuncheno – 175413 stebinių. Po apjungimo duomenų rinkinį sudaro 721116 stebinių.

Toliau pasirenkami tolimesnėje analizėje naudojami požymiai. Juos sudaro požymiai apie atsiliepimą ir svečią, būstą bei šeimininką. Požymiai ir informacija apie juos pateikiama 2 lentelėje.

2 lentelė. Atrinkti požymiai

Požymio pavadinimas programoje	**Požymis**
Atsiliepimo požymiai	
date	Atsiliepimo data
comments	Atsiliepimo tekstas
guests_included	Svečių skaičius (nuomojosi būstą su atsiliepimo autoriumi)
Būsto požymiai	
review_scores_rating	Bendras būsto vertinimas

(vėliau pervadinamas į *org_rating*)	
latitude	Būsto geografinė koordinatė (platuma)
longitude	Būsto geografinė koordinatė (ilguma)
market	Miestas (kuriame yra būstas)
property_type	Būsto tipas
room_type	Kambario tipas
bathrooms	Vonių skaičius
bedrooms	Miegamųjų skaičius
price	Kaina (už naktį)
minimum_nights	Minimalus viešnagės naktų skaičius
Šeimininko požymiai	
host_name	Būsto šeimininko vardas
host_is_superhost	Būsto šeimininko statusas (Jei šeimininkas yra „*superhost*“, tai reiškia, kad jis savo būstą bent 10 kartų yra nuomojęs arba turėjęs bent 3 rezervacijas mažiausiai 100 naktų laikotarpiui. Taip pat šeimininkas turi išlaikyti ne mažesnį negu 90 % atsakymų dažnį, ne didesnį negu 1 % atšaukimų dažnį (tai reiškia 1 atšaukimą, tenkantį 100 rezervacijų) bei 4.8 bendrą reitingą (skaičiuojamas už paskutines 365 dienas, remiantis data, kai svečias paliko atsiliepimą).)

1 pav. pateikta informacija apie požymius: numeris, pavadinimas, stebinių skaičius, tipas. Kainos požymis buvo nurodytas su dolerio ženklu priekyje, todėl požymis yra „object“ tipo, nors turėtų būti „float64“, t. y., skaitinis tipas. Šio požymio tipas yra perkoduojamas į skaitinį.

```
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   date                 721116 non-null  object
 1   comments             720716 non-null  object
 2   guests_included      721116 non-null  int64
 3   review_scores_rating 720270 non-null  float64
 4   latitude             721116 non-null  float64
 5   longitude            721116 non-null  float64
 6   market               720001 non-null  object
 7   property_type        721116 non-null  object
 8   room_type            721116 non-null  object
 9   bathrooms            719730 non-null  float64
 10  bedrooms             720725 non-null  float64
 11  price                721116 non-null  object
 12  minimum_nights       721116 non-null  int64
 13  host_is_superhost    721072 non-null  object
 14  host_name            721072 non-null  object
```

1 pav. Informacija apie požymius (Berlyno ir Miuncheno miestų duomenys)

Taip pat iš 1 pav. galima pastebėti, jog ne visų požymių stebėjimų skaičius yra vienodas. Tai reiškia, kad požymiai turi skirtingą kiekį trūkstamų reikšmių, kurios tolimesnėje analizėje neturėtų būti nagrinėjamos. Prieš šalinant trūkstamas reikšmes, atliekamas papildomas žingsnis – visų skaitinių požymių, kurių bent viena reikšmė yra 0, jai priskiriama trūkstama reikšmė, nes bet kurio skaitinio požymio nulinė reikšmė šiuo atveju reiškia įvedimo klaidą ir neteikia naudingos informacijos. Po to pašalinami visi stebiniai, kurių bent vienas iš požymių įgyja trūkstamą reikšmę.

Toliau įvykdomas kokybinių požymių tekstinių reikšmių perkodavimas. Tai pritaikoma trims požymiams: miestui, kambario tipui bei būsto šeimininko statusui.

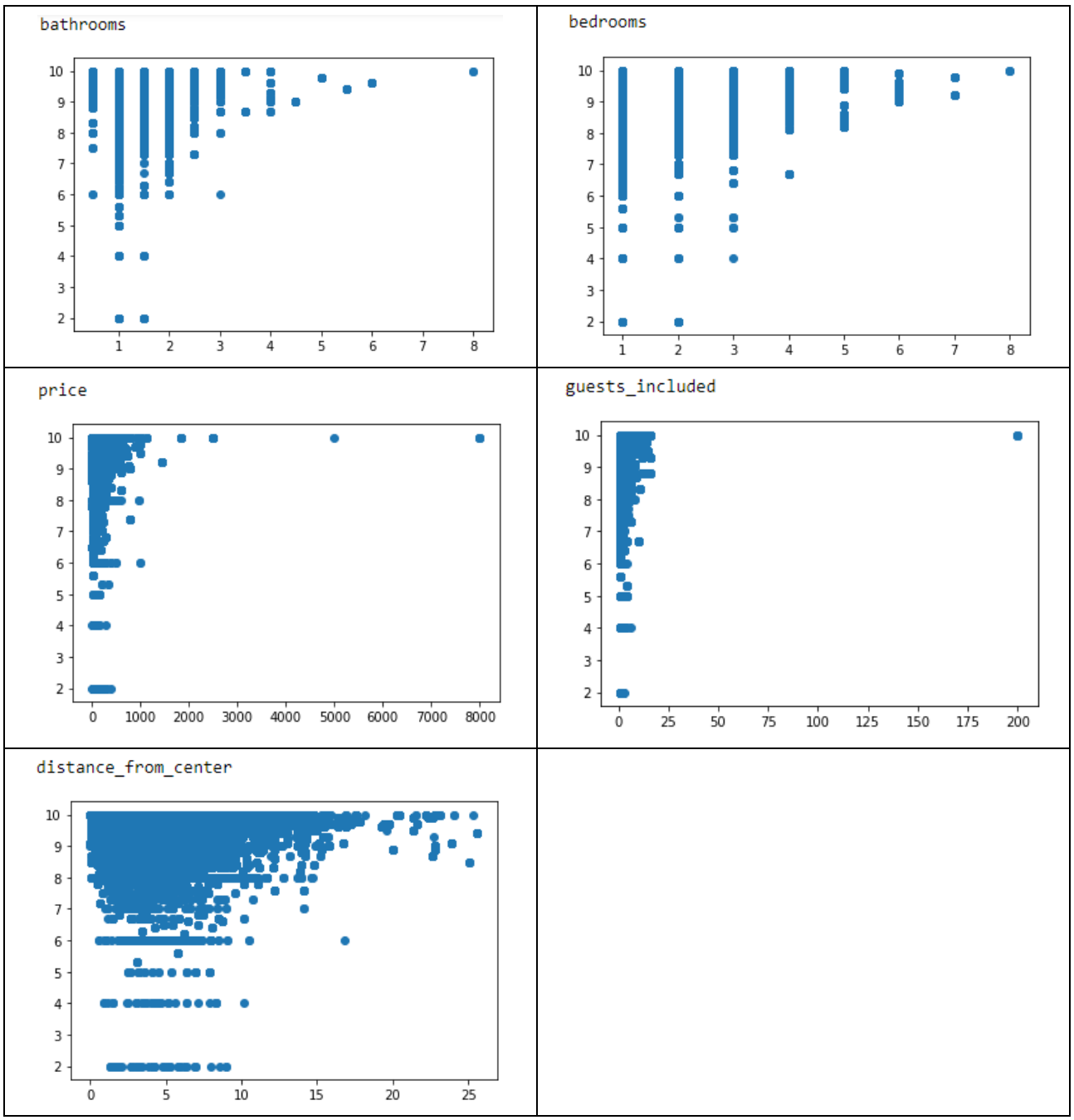
3 lentelė. Kategorinių požymių santykiniai dažniai (Berlyno ir Miuncheno miestų duomenys)

Požymis	**Požymio vardas programoje**	**Požymio kategorija**	**Dažnis**
Miestas	market	Berlin (Berlynas)	76 %
		Munich (Miunchenas)	24 %
Kambario tipas	room_type	Entire home / apt (visas apartamentas)	53 %
		Private room (privatus kambarys) arba Shared room (bendras kambarys)	47 %
Šeimininko statusas	host_is_superhost	f (false – nėra „superhost“)	58 %
		t (true – yra „superhost“)	42 %

Darbe iš turimų požymių pasiūlyta įvesti naujų požymių. Vienas iš jų yra atstumas iki miesto centro (požymio pavadinimas programoje – „distance_from_center“). Apibrėžiamos Berlyno bei Miuncheno centro koordinatės ir pasinaudojus turima būsto geografinės ilgumos ir geografinės platumos informacija, pagal Haversino formulę [36] sukuriamas naujas požymis, nurodantis, kiek atitinkamas būstas yra nutolęs miesto, kuriame jis yra nuomojamas, centro. Šis požymis gali turėti reikšmingą įtaką bendram būsto vertinimui – keliama hipotezė, kad kuo labiau būstas yra nutolęs nuo centro, tuo vietos vertinimas yra žemesnis, o tuo pačiu mažėja ir bendras būsto vertinimas. Antras pasiūlytas požymis yra atsiliepimo ilgis (požymio pavadinimas programoje – „comment_length“), t. y., žodžių, iš kurių sudarytas atsiliepimas, skaičių. Tolimesnėje analizėje nenagrinėjami atsiliepimai, kuriuos sudaro mažiau nei du žodžiai.

Tada formuojamas duomenų pjūvis. Pagal datos požymį filtruojami tik 2015-01-01 – 2020-01-01 klientų palikti atsiliepimai, pagal būsto tipo požymį filtruojami tie klientų atsiliepimai, kad požymio reikšmės atitiktų apartamento tipą, nes jie sudaro didžiąją stebėjimų dalį, o pagal minimalų viešnagės naktų skaičių filtruojami tik tie atsiliepimai, kai minimalus viešnagės naktų skaičius yra ne daugiau nei 5.

Po duomenų pjūvio formavimo, braižomi bendro būsto vertinimo nuo kitų požymių priklausomybės grafikai. Tokiu būdu galima pastebėti išsiskiriančias arba nelogiškas nagrinėjamų požymių reikšmes. Grafikuose *X* ašyse pavaizduotas atitinkamas požymis, o *Y* – bendras būsto vertinimas.



2 pav. Požymių priklausomybės nuo bendro būsto vertinimo grafikai (Berlyno ir Miuncheno miestų duomenys)

Matricos U_3 stebiniai yra filtruojami pagal šią sąlygą – (0 < „bathrooms“ ≤ 3) | (0 < „bedrooms“ ≤ 4) | (20 ≤ „price“ ≤ 400) | (1 ≤ „guests_included“ ≤ 6) | („distance_from_center“ ≤ 18). Po visų minėtų pakeitimų bei filtravimų, duomenų kiekis sumažėjo nuo 721116 iki 460687 stebėjimų.

Tada nustatoma kiekvieno iš atsiliepimų kalba. Taikomas kalbos aptikimo paketas – sukuriamas naujas požymis, kuris žymi, kokia kalba parašytas atitinkamas atsiliepimas. Rezultato fragmentas pateikiamas 3 pav.

	comments	language
471	Can's apartment is very well situated in the P...	en
472	Great place and location	en
474	AMAZING Location.\n\nFrankly, the location cou...	en
476	Great place to stay nice and central but still...	en
478	Can's place was large and spacious. It is conv...	en
479	Perfect host & appartment, as described. But a...	en
481	very nice, quiet apartment in the center of P-...	en

3 pav. Atsiliepimų ir kalbos požymių fragmentas (Berlyno ir Miuncheno miestų duomenys)

Šis kalbos aptikimo paketas palaiko 55 skirtingas kalbas pagal ISO 639-1 kodus [43]. Nustačius atsiliepimo kalbą, apskaičiuojama kiek stebinių kokia kalba yra parašyti. 4 lentelėje pateikiamos tik 5 didžiausią duomenų dalį užimančios kalbos.

4 lentelė. Atsiliepimų kalba ir jos duomenų dalis (Berlyno ir Miuncheno miestų duomenys)

Kalba	**Duomenų dalis**
Anglų k.	88,85 %
Vokiečių k.	6,44 %
Olandų k.	1,02 %
Ispanų k.	0,82 %
Prancūzų k.	0,59 %

Taip pat atsiliepimai buvo parašyti ir kitomis kalbomis: norvegų, lenkų ir t. t., tačiau jie sudarė itin mažą duomenų dalį. Pagal 4 lentelės duomenis matome, kad didžiąją dalį, net 88,85 %, sudaro anglų kalba parašyti atsiliepimai. Tolimesnėje analizėje paliekami tik jie ir duomenų imtis sumažėja iki 266628 stebėjimų.

Pradedamas atsiliepimo požymio paruošimas. Iš šio požymio stebinių šalinami įvairūs simboliai (apostrofai, klaustukai, šauktukai, kableliai, taškai, skaičiai ir t. t.), sakinių kėlimai į kitą eilutę, dvigubi (ar didesni) tarpai tarp žodžių pakeičiami į vieną, pašalinami žodžiai, sudaryti tik iš vieno simbolio. Taip pat pašalinami nereikšminiai žodžiai, kuriuos sudaro:

- nereikšminiai žodžiai, kuriuos sudaro Glazgo universiteto svetainėje [33] pateiktas nereikšminių angliškų žodžių sąrašas [34] bei kiti žodžiai, papildomi rankiniu būdu: miestų pavadinimai, internetinės svetainės, iš kurios gaunama informacija, pavadinimas ir pan. Bendras nereikšminių žodžių kiekis yra 405;
- šeimininkų vardai, kurių informaciją galima gauti iš pradinių duomenų imties. Šeimininkų vardų kiekis – 4992.

Po šio valymo tikrinama, ar yra pasikartojančių atsiliepimų ir pastebėta, kad yra tokių atsiliepimų, kurie, kaip manoma, buvo sugeneruoti automatiškai pačios internetinės svetainės. Tokie stebėjimai yra pašalinami iš tolimesnės analizės. Jų duomenų fragmentas pateiktas 4 pav.

	comments	guests_included	org_rating	market	room_type	bathrooms	bedrooms	price	minimum_nights	host_is_superhost	distance_from_center
1627	canceled reservation arrival automated posting	2	9.4	0	0	2.5	2.0	90.0	4	0	1.5
1628	canceled reservation arrival automated posting	2	9.4	0	0	2.5	2.0	90.0	4	0	1.5
1750	canceled reservation arrival automated posting	6	9.2	0	0	2.0	2.0	269.0	3	1	2.9

4 pav. Automatiškai sugeneruoti pasikartojantys atsiliepimai (Berlyno ir Miuncheno miestų duomenys)

Kitas žingsnis – sentimentų analizė – norima nustatyti, ar atsiliepimas yra teigiamas, neigiamas ar neutralus ir apskaičiuoti bendrą sentimento įvertį. Gaunamas naujas požymis, t. y. sentimentų įvertis (požymio pavadinimas programoje – „compound“). Kuo jo įverčio reikšmė arčiau −1, tuo labiau jis atitinka neigiamą atsiliepimą, o kuo arčiau 1, tuo labiau jis atitinka teigiamą atsiliepimą. 5 pav. pateikiama šio požymio aprašomoji statistika.

	compound
count	254955.000000
mean	0.814104
std	0.217821
min	-0.990500
25%	0.771700
50%	0.891000
75%	0.946000
max	0.999600

5 pav. Sentimentų įverčio aprašomoji statistika (Berlyno ir Miuncheno miestų duomenys)

Pastebime, kad sentimentų įverčio vidurkis yra 0,81 – tai reiškia, kad vertinant visą atsiliepimų imtį, dauguma atsiliepimų yra teigiami.

Atlikus visus žingsnius, susijusius su visu duomenų imties transformavimu ir mažinimu, pradedamas tik atsiliepimų požymio nagrinėjimas. Pirmiausia, visi atsiliepimų požymio stebėjimai sujungiami į vieną žodžių sąrašą. Tolimesnėje analizėje nagrinėjami tik žodžiai, sudaryti iš nemažiau nei dviejų simbolių. Taip pat darkart ieškoma, ar po kalbos nustatymo liko kokių nors ne angliškų žodžių, pavyzdžiui, rašybos klaidų, kai vietoje „little“, parašytas žodis „litle“. Gauta, kad tokių liko 23917, jie taip pat yra pašalinami. Gaunamas rezultatas pateiktas 6 pav.

```
Unique word count before removing Non-English words:  43754
Unique word count after removing Non-English words:  19837
Difference:  23917
```

6 pav. Atskirų žodžių ir bendras visų žodžių skaičius (Berlyno ir Miuncheno miestų duomenys)

Gauta, kad unikalių žodžių kiekis prieš valymą yra 43754, po valymo – 19837. Iš viso žodžių kiekis yra 3,7 milijono. 5 lentelėje pateikiamas 20 dažniausių žodžių ir jų dažnumo sąrašas.

5 lentelė. 20 dažniausių žodžių sąrašas (Berlyno ir Miuncheno miestų duomenys)

Nr.	Žodis	Dažnumas	Nr.	Žodis	Dažnumas
1.	great	138232	11.	comfortable	38680
2.	place	111921	12.	perfect	37781
3.	stay	98104	13.	helpful	30397
4.	nice	89372	14.	friendly	28090
5.	location	86329	15.	city	27713
6.	clean	77562	16.	quiet	27090
7.	good	56584	17.	lovely	26975
8.	room	54354	18.	restaurants	26382
9.	recommend	52370	19.	located	25266
10.	close	44260	20.	spacious	23642

Pagal 5 lentelę pastebima, kad gauti rezultatai yra labai panašūs į rezultatus, gautus mokslinėje literatūroje [4, 8].

6 lentelė. Žodžių dažnio pasiskirstymas visoje žodžių imtyje (Berlyno ir Miuncheno miestų duomenys)

Žodžių sąrašo Nr.	Dažnumas	Procentinė dalis	Suminė procentinė dalis
1-100	1932992	54,32 %	54,32 %
101-200	375835	10,56 %	64,88 %
201-300	224834	6,32 %	71,20 %
301-400	158400	4,45 %	75,65 %
401-500	120812	3,39 %	79,05 %
501-19837	745682	20,95 %	100 %

6 lentelėje galima pastebėti, kad pirmi 100 atskirų žodžių sudaro daugiau nei pusę visų žodžių imties. Taip pat pirmi 500 atskirų žodžių sudaro net ~79 % visų žodžių imties.

Toliau šis atskirų žodžių ir jų dažnių sąrašas yra peržiūrimas ir rankiniu būdu sumažinamas iki tokių žodžių sąrašo, kurie apibūdina kliento patirtį. Šiame žingsnyje remiamasi ir moksline literatūra, ir savo nuožiūra. Nenagrinėjant mažai pasikartojančių žodžių, kurių dažnis yra mažiau, nei 100, gaunamas patirties žodžių žodynas iš 308 žodžių (visas sąrašas pateiktas 6 priede). Patirtį atspindinčių žodžių debesis pavaizduotas 7 pav.

7 pav. Patirtį atspindinčių žodžių debesis

Sukuriama nauja atsiliepimų-požymių matrica *D*. Svarbu paminėti, kad tolimesnėje analizėje nagrinėjami tik tie atsiliepimai, kuriuose yra bent vienas žodis iš patirties žodžių sąrašo, tokiu būdu nenagrinėjant atsiliepimų, kuriuose apie patirtį nekalbama. Šie atsiliepimai sudaro 99,7 % visų atsiliepimų imties.

Norint toliau taikyti duomenų dimensijos mažinimo metodus, matrica turi būti įvertinama skaitinėmis reikšmėmis. Šiame žingsnyje naudojamasi vektorinės erdvės modeliu įvertinti, kokį svorį turi kiekvienas žodis atitinkamame atsiliepime.

	bars	basic	bathroom	beautiful	bed	bedroom	best	big
1	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
2	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
3	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.026742
4	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
5	0.041051	0.0	0.0	0.0	0.0	0.0	0.0	0.000000

8 pav. Patirties žodžių svorių atsiliepimuose matricos fragmentas (Berlyno ir Miuncheno miestų duomenys)

Įvertinus patirtį atspindinčių žodžių įverčius kiekviename stebėjime, ši matrica toliau naudojama duomenų dimensijos mažinimo analizėje.

Pirmiausia, atliekami keli paruošiamieji žingsniai. Apskaičiuojamas bendras KMO rodiklis, parodantis, ar faktorinės analizės taikymas yra priimtinas. Rezultatas pateikiamas programinio kodo fragmentu ir gauta KMO reikšme.

```
kmo_all, kmo_model = calculate_kmo(df4)
round(kmo_model,2)
```

```
0.5
```

9 pav. KMO mato rezultatas (Berlyno ir Miuncheno miestų duomenys)

9 pav. pateiktas rezultatas rodo, jog duomenys faktorinei analizei yra tinkami tik iš dalies. Taip pat apskaičiuojamas MSA_i rodiklis kiekvienam požymiui ir, atlikus keletą bandymų, pastebėta, kad geriausias rezultatas gaunamas tada, kai pasirenkami tik tie požymiai, kurių $MSA_i > 0{,}57$. Po šio filtravimo, bendras KMO matas padidėja iki 0,59. Taip pat skaičiuojamas ir Bartlett‘o sferiškumo kriterijus.

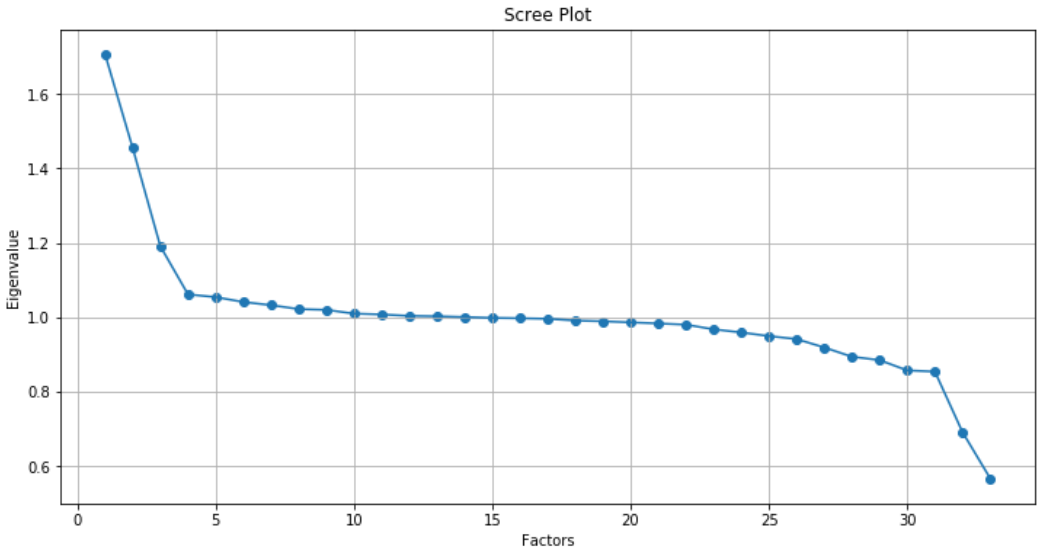
```
chi_square_value, p_value = calculate_bartlett_sphericity(df4)
round(chi_square_value, 1), round(p_value,3)
```

```
(130897.2, 0.0)
```

10 pav. Bartlett‘o sferiškumo kriterijus (Berlyno ir Miuncheno miestų duomenys)

Iš 10 pav. matome, kad gauta *p* reikšmė yra mažesnė už reikšmingumo lygmenį α, t. y., 0,0 < 0,5. Tai reiškia, kad duomenys pagal šį kriterijų taip pat yra tinkami faktorinei analizei taikyti.

Patvirtinus, kad duomenys yra tinkami, nustatinėjamas optimalus faktorių skaičius. Braižomas tikrinių reikšmių priklausomybės nuo faktorių skaičiaus grafikas.



11 pav. Tikrinių reikšmių priklausomybės nuo faktorių skaičiaus grafikas (Berlyno ir Miuncheno miestų duomenys)

Pagal 11 pav. matome, kad grafiko linija pradeda tiesėti maždaug nuo 6 faktoriaus. Po kelių bandymų nustatyta, jog optimaliausias faktorių skaičius yra 8, o metodas jų išskyrimui – pagrindinių komponenčių metodas. Taip pat pasirenkamas ortogonalusis faktorių sukimas *varimax*, nes tokiu būdu gauti faktoriai tenkins vieną iš regresinės analizės prielaidų – kintamieji yra nepriklausomi. Nustatoma, kad būtų išvedami tik požymiai, kurių svoris faktoriuje yra daugiau nei 0,35, kadangi norima gauti kuo didesnį požymių skaičių.

	0	1	2	3	4	5	6	7
shops	0.52							
cafes	0.54							
bars	0.67							
restaurants	0.78							
kitchen		0.42						
private		0.46						
room		0.61						
bathroom		0.74						
heating			0.36					
shower			0.54					
water			0.63					
fridge				0.36				
fresh				0.6				
towels				0.66				
trendy					-0.36			
metro					0.44			
various					0.44			
stops					0.52			
large						0.51		
desk						0.57		
clubs							0.55	
nightlife							0.59	
uncomfortable								0.48
dirty								0.62

12 pav. Faktorių išskyrimo rezultatas (Berlyno ir Miuncheno miestų duomenys)

Pagal 12 pav. pastebime, kad faktorius sudaro 2-4 požymiai. 7 lentelėje kiekvienas iš faktorių yra apibūdinamas bendru, jį nusakančiu pavadinimu.

7 lentelė. Faktorių apibūdinimas (Berlyno ir Miuncheno miestų duomenys)

Faktoriaus Nr.	**Faktoriaus pavadinimas**	**Požymiai**
1.	Paslaugos netoliese	Parduotuvės Kavinės Barai Restoranai
2.	Pagrindinis produktas	Virtuvė Privatus Kambarys Vonios kambarys
3.	Apartamento privalumai	Šildymas Dušas Vanduo
4.	Papildomi apartamento privalumai	Šaldytuvas Naujas Rankšluosčiai
5.	Susisiekimas	Modernus Metro

		Įvairus Stotelės
6.	Būsto patogumai	Didelis Stalas
7.	Pramogos suaugusiems	Klubai Naktinis gyvenimas
8.	Neigiama patirtis	Nepatogus Nešvarus

Pastebėta, jog nėra nei vieno faktoriaus, apibūdinančio būsto šeimininką, nors patirties žodžių žodyne būdvardžių šeimininkui yra nemažai. Toliau skaičiuojama, kokią dispersijos dalį paaiškina gauti faktoriai. Rezultatai pateikiami 13 pav., kur kintamasis „cum_var“ nurodo kaupiamąją dispersijos dalį, o kintamasis „var“ – kiekvieno faktoriaus dispersijos dalį.

	0	1	2	3	4	5	6	7
cum_var	0.050708	0.092953	0.128086	0.161025	0.192538	0.223979	0.255249	0.286499
var	0.050708	0.042245	0.035133	0.032939	0.031514	0.031441	0.031270	0.031250

13 pav. Paaiškinama faktorių dispersijos dalis (Berlyno ir Miuncheno miestų duomenys)

Gauta, kad 8 faktoriai apibūdina 28,65 % visų požymių dispersijos. Toks rezultatas tenkina – gauta reikšmė yra didesnė nei nagrinėtoje mokslinėje literatūroje [4, 8].

Po visų įvykdytų žingsnių, turima informacija papildoma gautais faktoriais. Toliau duomenys ruošiami regresinės analizės metodo taikymui. Tai atliekama tokia eiga:

- kiekvienam atsiliepimui apskaičiuojami faktorių svoriai (fragmentas pateiktas 7 priede);
- sudaroma duomenų matrica iš pradinių bei naujai sukurtų požymių ir gautų faktorių U_3;
- požymiai yra standartizuojami, kad jų reikšmės būtų pasiskirsčiusios panašioje skalėje;
- po standartizavimo, darkart braižomi bendro būsto vertinimo ir požymių priklausomybės grafikai. Pastebėjus aiškiai išsiskiriančias reikšmes, požymiai yra filtruojami.

Duomenys yra paruošti ir toliau naudojami taikant regresinės analizės metodą.

Norint nustatyti, kuriuos požymius iš tolimesnės analizės pašalinti, o kuriuos – palikti, sudaromas tiesinės regresijos modelis ir atsižvelgiama į koreguotą apibrėžtumo koeficientą bei nepriklausomumo hipotezės *p* reikšmę (tikrinama, ar atskiras požymis yra reikšmingas regresijos lygtyje). Multikolinearumas tikrinamas pagal *VIF*. Gaunama, kad koreguotas apibrėžtumo koeficientas yra 27,39 % bei yra kintamųjų, kurie tiesinės regresijos lygtyje yra nereikšmingi ($p > 0,05$), o koeficientas *VIF* visiems požymiams rodo, kad multikolinearumo problemos nėra. Toliau po vieną šalinami nereikšmingi požymiai ir stebima, kaip kinta koreguotas apibrėžtumo koeficientas bei kitų požymių reikšmingumas. Pašalinus požymį FA7, 14 pav. pavaizduoti rezultatai, kai lygtyje lieka tik statistiškai reikšmingi kintamieji.

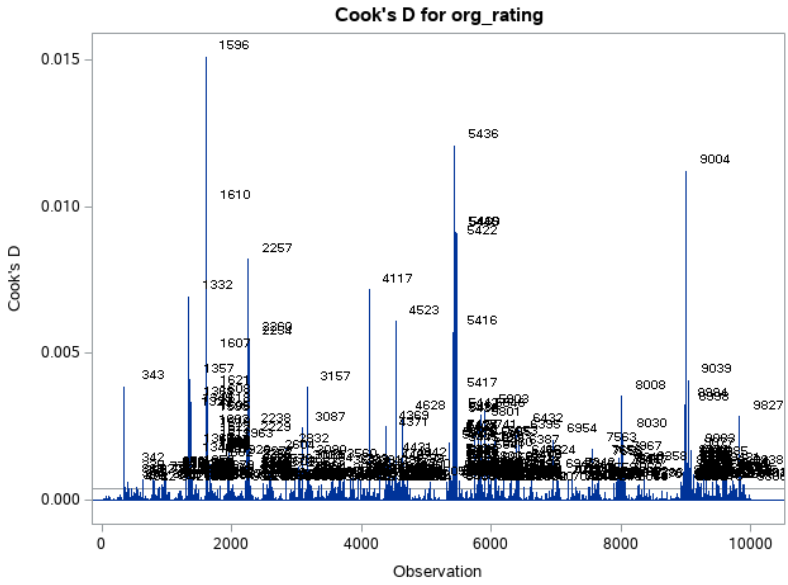
Root MSE	0.80966	R-Square	0.2739
Dependent Mean	0.01256	Adj R-Sq	0.2739
Coeff Var	6447.74747		

Parameter Estimates

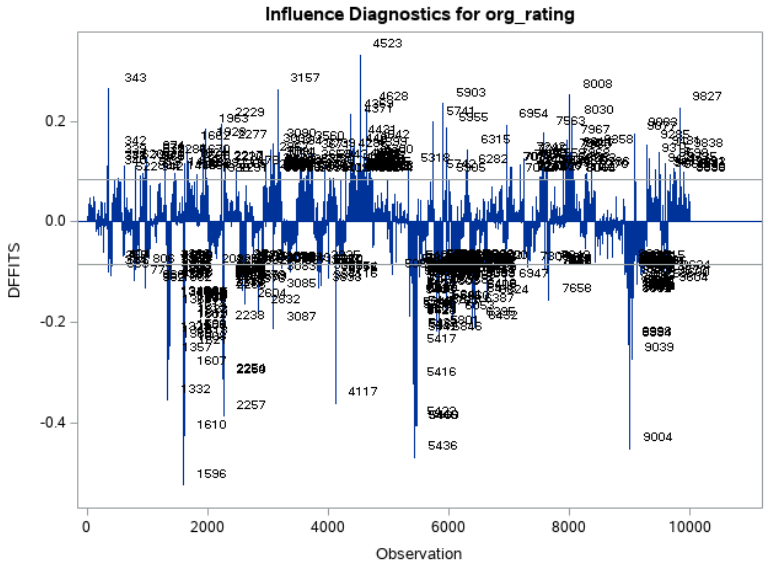
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	1	-0.48183	0.00914	-52.74	<.0001	0	0	-0.49974	-0.46392
FA1	1	0.00621	0.00177	3.50	0.0005	0.00604	1.02891	0.00273	0.00968
FA2	1	-0.02691	0.00177	-15.18	<.0001	-0.02635	1.04116	-0.03038	-0.02343
FA3	1	-0.02058	0.00181	-11.36	<.0001	-0.01961	1.03027	-0.02414	-0.01703
FA4	1	-0.00482	0.00192	-2.52	0.0118	-0.00431	1.01360	-0.00858	-0.00107
FA5	1	-0.00453	0.00191	-2.37	0.0178	-0.00406	1.01193	-0.00828	-0.00078454
FA6	1	0.00553	0.00196	2.83	0.0047	0.00483	1.01038	0.00170	0.00937
FA8	1	-0.02644	0.00225	-11.76	<.0001	-0.02019	1.01882	-0.03084	-0.02203
guests_included	1	-0.08330	0.00218	-38.29	<.0001	-0.08355	1.64598	-0.08756	-0.07903
market	1	0.12465	0.00397	31.43	<.0001	0.05614	1.10317	0.11688	0.13243
room_type	1	0.25271	0.00418	60.48	<.0001	0.13279	1.66679	0.24452	0.26090
bathrooms	1	0.04520	0.00635	7.12	<.0001	0.01327	1.20027	0.03276	0.05764
bedrooms	1	-0.13175	0.00385	-34.19	<.0001	-0.08101	1.94113	-0.13930	-0.12420
price	1	0.13101	0.00268	48.84	<.0001	0.12950	2.43027	0.12575	0.13626
minimum_nights	1	0.10686	0.00158	67.55	<.0001	0.11859	1.06564	0.10376	0.10996
host_is_superhost	1	0.82448	0.00329	250.68	<.0001	0.43088	1.02138	0.81803	0.83093
distance_from_center	1	0.03872	0.00169	22.84	<.0001	0.04032	1.07695	0.03540	0.04204
comment_length	1	-0.01921	0.00172	-11.16	<.0001	-0.02008	1.11836	-0.02258	-0.01584
compound	1	0.13174	0.00172	76.81	<.0001	0.13690	1.09818	0.12838	0.13510

14 pav. Tiesinės regresijos rezultatai po požymių šalinimo (Berlyno ir Miuncheno miestų duomenys)

Koreguoto apibrėžtumo koeficiento reikšmė nepasikeitė. Taip pat pastebima, kad stipriausią įtaką modelyje turi požymis, nurodantis būsto šeimininko statusą (pagal standartizuotą įvertį iš „Standardized Estimate“). Kitas žingsnis – modelio išskirčių tikrinimas.

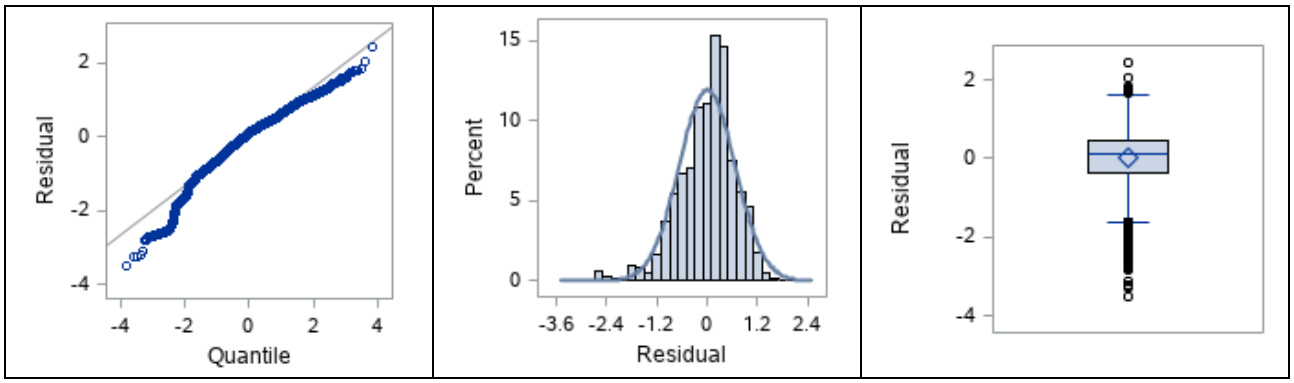


15 pav. Kuko mato reikšmių grafikas (Berlyno ir Miuncheno miestų duomenys)



16 pav. DFFITS reikšmių grafikas (Berlyno ir Miuncheno miestų duomenys)

Iš 15 ir 16 pav. pastebima, kad modelyje stebima nemažai išskirčių. Patikrinamos ir modelio liekanos.



17 pav. Tiesinės regresijos modelio prielaidų tikrinimo grafikai (Berlyno ir Miuncheno miestų duomenys)

17 pav. galima pastebėti, kad modelio liekanos netenkina normalumo prielaidos bei rodo, kad modelyje yra išskirčių. Visas išskirtis šalinti rankiniu būdu būtų per daug sudėtinga, todėl tam taikomas atspariosios regresijos metodas.

Pirmiausia, duomenų imtis sumažinama iki atsitiktinai parinktų 10000 stebėjimų, nes, turint didelius duomenų rinkinius, modelio prielaidų tikrinimas atliekamas ne su visa imtimi, o su 5-10 % imties, priklausomai nuo pradinių duomenų kiekio, nes rezultatai su didele duomenų imtimi gali būti ne tokie informatyvūs. Tada taikomas atspariosios regresijos modelis su visais pradiniais požymiais. Gaunamas išskirčių ir įtakos taškų sąrašas (fragmentas pateikiamas 8 priede), kuris vėliau sujungiamas su pradiniais duomenimis ir atliekami tokie pakeitimai – pašalinami visi stebėjimai, jei „PRobustDist“ > 300 arba „Outlier“ = 1, tokiu būdu pašalinant grubiausias išskirtis. Kintamasis „PRobustDist“ atitinka Mahalanobio atstumą, nurodantį, kokiu atstumu tam tikras taškas nutolęs nuo modelio ašies. Kitas kintamasis, „Outlier“, atitinka išskirtį *Y* ašies kryptimi. Po šio pašalinimo, 10000 stebėjimų sumažėja 9238. Iš naujo kartojamas tiesinės regresijos modelis – 18 pav. pateikiami rezultatai po nereikšmingų kintamųjų šalinimo.

Root MSE	0.67985	R-Square	0.2819
Dependent Mean	0.08382	Adj R-Sq	0.2809
Coeff Var	811.07484		

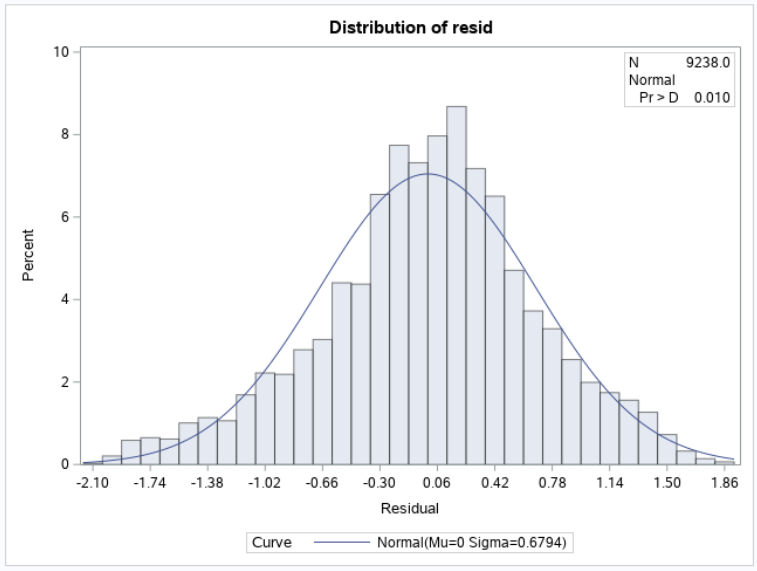
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	1	-0.37044	0.03135	-11.82	<.0001	0	0	-0.43189	-0.30898
FA2	1	-0.02460	0.00809	-3.04	0.0024	-0.02739	1.04165	-0.04045	-0.00874
FA3	1	-0.02385	0.01009	-2.36	0.0181	-0.02118	1.03061	-0.04362	-0.00408
FA5	1	-0.02035	0.01070	-1.90	0.0572	-0.01694	1.01903	-0.04132	0.00062106
guests_included	1	-0.04477	0.00950	-4.71	<.0001	-0.05205	1.56558	-0.06339	-0.02616
market	1	0.15810	0.01717	9.21	<.0001	0.08504	1.09514	0.12445	0.19175
room_type	1	0.26093	0.01804	14.47	<.0001	0.16251	1.62083	0.22558	0.29629
bedrooms	1	-0.11025	0.01639	-6.73	<.0001	-0.08040	1.83424	-0.14237	-0.07813
price	1	0.12367	0.01119	11.06	<.0001	0.14579	2.23381	0.10174	0.14560
minimum_nights	1	0.08749	0.00696	12.56	<.0001	0.11451	1.06725	0.07384	0.10114
host_is_superhost	1	0.69983	0.01438	48.68	<.0001	0.43385	1.02041	0.67165	0.72801
distance_from_center	1	0.04024	0.00726	5.54	<.0001	0.05067	1.07265	0.02602	0.05447
compound	1	0.11649	0.00731	15.94	<.0001	0.14219	1.02195	0.10217	0.13082

18 pav. Tiesinės regresijos rezultatai po išskirčių šalinimo (Berlyno ir Miuncheno miestų duomenys)

Pastebime, kad koreguotas apibrėžtumo koeficientas padidėjo iki 28,1 %. Lyginant su moksline literatūra [4], gautas rezultatas yra didesnis. Toliau tikrinamos modelio prielaidos.

1. Liekamosios paklaidos pasiskirsčiusios pagal normalųjį skirstinį. Ši prielaida yra tikrinama grafiškai. Kaip matome 19 pav., liekamosios paklaidos nėra visiškai pasiskirsčiusios pagal normalųjį skirstinį, tačiau grubių pažeidimų nėra. Tai patvirtina ir eksceso koeficiento reikšmė (20 pav.), lygi 0,15, rodanti šiek tiek smailesnę viršūnę lyginant su normaliuoju skirstiniu bei asimetrijos koeficientas (20 pav.), kurio reikšmė yra −0,25, rodanti, kad grafikas turi nedidelę kairinę asimetriją. Prielaida yra tenkinama iš dalies.



19 pav. Liekamųjų paklaidų histogramos lyginimas su normalaus skirstinio tankio kreive (Berlyno ir Miuncheno miestų duomenys)

Moments			
N	9238	Sum Weights	9238
Mean	0	Sum Observations	0
Std Deviation	0.67940604	Variance	0.46159257
Skewness	-0.2570161	Kurtosis	0.14612455
Uncorrected SS	4263.73058	Corrected SS	4263.73058
Coeff Variation	.	Std Error Mean	0.00706871

20 pav. Liekamųjų paklaidų skaitinės charakteristikos (Berlyno ir Miuncheno miestų duomenys)

2. Liekamųjų paklaidų vidurkis yra lygus 0. Patikrinus hipotezę apie populiacijos liekamųjų paklaidų vidurkio lygybę 0, ji neatmetama ir prielaida yra tenkinama.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000

21 pav. Liekamųjų paklaidų vidurkio lygybės 0 tikrinimas (Berlyno ir Miuncheno miestų duomenys)

3. Liekamųjų paklaidų dispersijos yra lygios – homoskedastiškumo prielaida. Patikrinus hipotezę apie populiacijos liekamųjų paklaidų dispersijų lygybę, ji atmetama ir prielaida yra netenkinama.

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
87	1478.87	<.0001

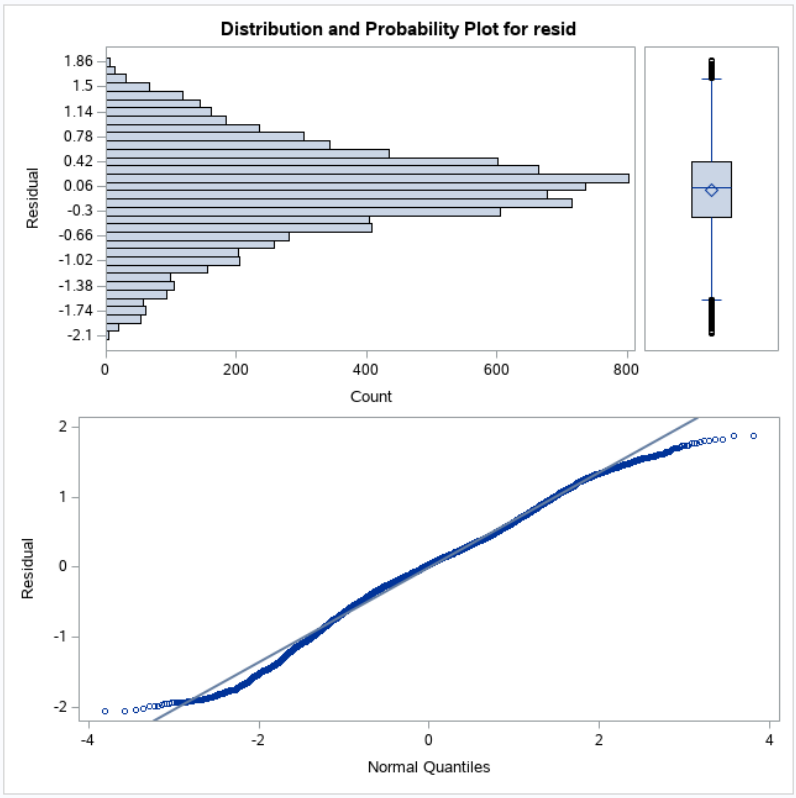
22 pav. Liekamųjų paklaidų dispersijų lygybės tikrinimas (Berlyno ir Miuncheno miestų duomenys)

4. Visos atsitiktinės paklaidos yra nepriklausomos, t. y., nėra autokoreliacijos. Prielaida tikrinama pagal Durbin‘o-Watson‘o statistiką. Gauta statistikos *p* reikšmė yra 1,169 (>0,05), tai reiškia, kad autokoreliacijos modelyje nėra. Taip pat pastebėta, kad atsižvelgiant į kairę kritinę sritį, nulinė hipotezė yra neatmetama, atsižvelgiant į dešinę – turėtų būti atmetama. Ketvirtoji prielaida yra iš dalies tenkinama.

Durbin-Watson D	1.169
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	9238
1st Order Autocorrelation	0.416

23 pav. Autokoreliacijos tikrinimas (Berlyno ir Miuncheno miestų duomenys)

5. Duomenyse nėra išskirčių. Atlikus pradinę išskirčių analizę, buvo pastebėta, kad duomenyse yra nemažai išskirčių. Pritaikius atspariosios regresijos metodą ir pašalinus grubiausias išskirtis, tikrinami liekanų grafikai. 24 pav. matome, jog ne visos išskirtys yra pašalintos, tačiau stipriai išsiskiriančių stebėjimų, grubių išskirčių nebėra. Prielaida yra iš dalies tenkinama.



24 pav. Išskirčių pagal liekanų grafikus tikrinimas (Berlyno ir Miuncheno miestų duomenys)

6. Tarp nepriklausomų kintamųjų nėra stiprios koreliacijos. Tarp likusių požymių yra stebima vidutinė arba silpna koreliacija. Vadinasi, ši prielaida yra iš dalies tenkinama.

Gauta, kad netenkinama 3 prielaida, todėl tikrinami heteroskedastiškumo pasikliautinieji intervalai.

Parameter Estimates												
						Heteroscedasticity Consistent						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits		Heteroscedasticity Consistent 95% Confidence Limits	
Intercept	1	-0.37044	0.03135	-11.82	<.0001	0.03270	-11.33	<.0001	0	-0.43189	-0.30898	-0.43453	-0.30635
FA2	1	-0.02460	0.00809	-3.04	0.0024	0.00767	-3.20	0.0014	1.04165	-0.04045	-0.00874	-0.03964	-0.00955
FA3	1	-0.02385	0.01009	-2.36	0.0181	0.00973	-2.45	0.0143	1.03061	-0.04362	-0.00408	-0.04293	-0.00477
FA5	1	-0.02035	0.01070	-1.90	0.0572	0.01161	-1.75	0.0798	1.01903	-0.04132	0.00062106	-0.04312	0.00242
guests_included	1	-0.04477	0.00950	-4.71	<.0001	0.00996	-4.50	<.0001	1.56558	-0.06339	-0.02616	-0.06429	-0.02525
market	1	0.15810	0.01717	9.21	<.0001	0.01686	9.38	<.0001	1.09514	0.12445	0.19175	0.12506	0.19114
room_type	1	0.26093	0.01804	14.47	<.0001	0.01876	13.91	<.0001	1.62083	0.22558	0.29629	0.22416	0.29771
bedrooms	1	-0.11025	0.01639	-6.73	<.0001	0.01747	-6.31	<.0001	1.83424	-0.14237	-0.07813	-0.14448	-0.07601
price	1	0.12367	0.01119	11.06	<.0001	0.01176	10.51	<.0001	2.23381	0.10174	0.14560	0.10061	0.14673
minimum_nights	1	0.08749	0.00696	12.56	<.0001	0.00712	12.30	<.0001	1.06725	0.07384	0.10114	0.07354	0.10143
host_is_superhost	1	0.69983	0.01438	48.68	<.0001	0.01361	51.44	<.0001	1.02041	0.67165	0.72801	0.67316	0.72650
distance_from_center	1	0.04024	0.00726	5.54	<.0001	0.00708	5.68	<.0001	1.07265	0.02602	0.05447	0.02636	0.05413
compound	1	0.11649	0.00731	15.94	<.0001	0.00799	14.59	<.0001	1.02195	0.10217	0.13082	0.10084	0.13215

25 pav. Tiesinės regresijos rezultatai su heteroskedastiškumui atspariais pasikliautinaisiais intervalais (Berlyno ir Miuncheno miestų duomenys)

Sudarytas galutinis modelis ir gauta, jog prielaidos modelio tinkamumui yra arba tenkinamos, arba tenkinamos tik iš dalies – 4 prielaidos tenkinamos iš dalies, o 1 tenkinama pilnai, 1 – netenkinama. Gautas koreguotas apibrėžtumo koeficientas yra 0,2809, vadinasi 28,1 % bendro reitingo sklaidos apie vidurkį galima paaiškinti tiesine regresija požymių „FA2“, „FA3“, „FA5“, „guests_included“, „market“, „room_type“, „bedrooms“, „price“, „minimum_nights“, „host_is_superhost“,

„distance_from_center“, „compound“ atžvilgiu. Aprašoma gauta analizinė tiesinės regresijos išraiška pagal regresijos lygties koeficientus:

$$\textbf{org\_rating} = -0{,}37 - 0{,}02*\textbf{FA2} - 0{,}02*\textbf{FA3} - 0{,}02*\textbf{FA5} - 0{,}04*\textbf{guests\_included} + 0{,}16*\textbf{market} + 0{,}26*\textbf{room\_type} - 0{,}11*\textbf{bedrooms} + 0{,}12*\textbf{price} + 0{,}09*\textbf{minimum\_nights} + 0{,}70*\textbf{host\_is\_superhost} + 0{,}04*\textbf{distance\_from\_center} + 0{,}12*\textbf{compound}$$

Iš regresijos lygties matome, kad, pavyzdžiui, padidinus kainą 1 doleriu, bendras reitingas vidutiniškai padidėja per 0,12 balo. Taip pat padidinus miegamųjų skaičių vienu vienetu, bendras reitingas sumažėja vidutiniškai 0,11 balo. Pagal 25 pav. su 95 % garantija galima teigti, kad, pavyzdžiui, kainą padidinus 1 doleriu, bendras būsto vertinimo vidurkio padidėjimas bus intervale nuo 0,10 balo iki 0,15 balo.

Aprašoma gauta standartizuota tiesinės regresijos lygtis:

$$\textbf{org\_rating*} = -0{,}03*\textbf{FA2} - 0{,}02*\textbf{FA3} - 0{,}02*\textbf{FA5} - 0{,}05*\textbf{guests\_included} + 0{,}09*\textbf{market} + 0{,}16*\textbf{room\_type} - 0{,}08*\textbf{bedrooms} + 0{,}15*\textbf{price} + 0{,}11*\textbf{minimum\_nights} + 0{,}43*\textbf{host\_is\_superhost} + 0{,}05*\textbf{distance\_from\_center} + 0{,}14*\textbf{compound}$$

Analizuojant gautą išraišką galima pastebėti, kad stipriausią teigiamą įtaką bendram būsto vertinimui turi šeimininko statusas – vadinasi, klientai linkę palikti aukštą vertinimą tada, kai šeimininko statusas atitinka „superhost“. Taip pat stipriausią neigiamą įtaką bendram būsto vertinimui turi miegamųjų skaičius ir svečių skaičius – taip gali atsitikti dėl to, jog klientai keliauja didesnėmis grupėmis ir apartamentai neturi pakankamai atskirų miegamųjų visoms keliautojų poroms ir / arba šeimos nariams. Pastebėta, kad visi iš atsiliepimo tekstų sudaryti požymiai: „Pagrindinis produktas“, „Apartamento privalumai“, „Susisiekimas“ – turi neigiamą įtaką bendram būsto vertinimui. Nors faktoriuose „Pagrindiniai produkto požymiai“ bei „Apartamento privalumai“ neigiamų žodžių ar būdvardžių nėra, bet dėl to, jog šie faktoriai turi neigiamą įtaką, galima daryti išvadą, jog jie ne visada buvo aprašomi teigiamai.

3.2. Ispanijos „Airbnb“ atvejo analizė

Ispanijos atvejo analizė atliekama analogiškai Vokietijos atvejui, tačiau šiuo atveju pasirenkami Ispanijos miestai, t. y., Madridas ir Barselona. Po duomenų nuskaitymo gauta, kad Madrido duomenų imtį sudaro 804306 stebinių, o Barselonos – 722841 stebinių. Po apjungimo duomenų rinkinį sudaro 1527147 stebinių.

Tolimesnėje analizėje nagrinėjami tie patys požymiai, kaip ir Vokietijos atveju (pavaizduota 8 lentelėje). Po trūkstamų reikšmių šalinimo duomenų imtį sudaro 1442116 stebinių.

Tada įvykdomas požymių tekstinių reikšmių perkodavimas.

8 lentelė. Kategorinių požymių santykiniai dažniai (Madrido ir Barselonos duomenys)

Požymis	Požymio vardas programoje	Požymio kategorija	Dažnis
Miestas	market	Madrid (Madridas)	51 %
		Barcelona (Barselona)	49 %

Kambario tipas	room_type	Entire home / apt (visas apartamentas)	62 %
		Private room (privatus kambarys), Shared room (bendras kambarys) arba Hotel room (viešbučio tipo kambarys)	38 %
Šeimininko statusas	host_is_superhost	f (false – nėra „superhost“)	62 %
		t (true – yra „superhost“)	38 %

Įvedami tokie patys, kaip Vokietijos atveju, požymiai: atstumas iki miesto centro (požymio pavadinimas programoje – „distance_from_center“) ir atsiliepimo ilgis (požymio pavadinimas programoje – „comment_length“).

Formuojamas duomenų pjūvis (analogiškai Vokietijos atvejui) ir braižomi bendro būsto vertinimo priklausomybės nuo kitų požymių grafikai. Atlikus kelis filtravimo bandymus, tolimesnei analizei požymiai filtruojami – (0 < „bathrooms“ ≤ 4) | (0 < „bedrooms“ ≤ 6) | (20 ≤ „price“ ≤ 400) | (1 ≤ „guests_included“ ≤ 8) | („distance_from_center“ ≤ 12). Po visų pakeitimų duomenų kiekis nuo 1442116 stebinių sumažėja iki 1021376.

Atliekant duomenų matricos pagal atsiliepimų požymį transformavimą, nustatoma kiekvieno iš atsiliepimų kalba. Po to apskaičiuojama kuria kalba kiek atsiliepimų yra parašyta. 9 lentelėje pateikiamos 5 didžiausią duomenų dalį užimančios kalbos.

9 lentelė. Atsiliepimų kalba ir jos duomenų dalis (Madrido ir Barselonos duomenys)

Kalba	**Duomenų dalis**
Anglų k.	85,28 %
Ispanų k.	9,28 %
Italų k.	1,2 %
Olandų k.	0,89 %
Prancūzų k.	0,65 %

Tolimesnėje analizėje paliekami tik anglų kalba parašyti komentarai ir duomenų imtis sumažėja iki 490966 stebėjimų.

Pradėjus atsiliepimo požymio paruošimą, šalinami įvairūs simboliai, sakinių kėlimai, dvigubi (ar didesni) tarpai tarp žodžių keičiami į vieną, šalinami žodžiai, sudaryti iš vieno simbolio. Šalinami nereikšminių žodžių sąrašas, kurį sudaro:

- nereikšminiai žodžiai, kurių kiekis yra 404;
- šeimininkų vardai, kurių kiekis – 4179.

Po šio valymo pašalinami visi besidubliuojantys atsiliepimai ir atliekama sentimentų analizė. 26 pav. pateikiama sentimentų įverčio aprašomoji statistika.

	compound
count	477653.000000
mean	0.814649
std	0.224906
min	-0.996300
25%	0.778300
50%	0.893400
75%	0.947700
max	0.999300

26 pav. Sentimentų įverčio aprašomoji statistika (Madrido ir Barselonos duomenys)

Sentimentų įverčio vidurkis yra 0,82, tai reiškia, kad dauguma atsiliepimų yra teigiami.

Pradedamas atsiliepimų požymio nagrinėjimas. Visus atsiliepimų požymio stebėjimus sujungus į vieną duomenų masyvą, atliekamas jo transformavimas ir tikrinami ne angliški žodžiai. Gaunama, kad ne angliškų žodžių yra 61956, jie yra pašalinami. Gautas rezultatas pateikiamas 27 pav.

```
Unique word count before removing Non-English words:  61956
Unique word count after removing Non-English words:  25048
Difference:  36908
```

27 pav. Atskirų žodžių ir bendras visų žodžių skaičius (Madrido ir Barselonos duomenys)

Gauta, kad unikalių žodžių kiekis prieš ne angliškų žodžių išvalymą yra 61956, po – 25048. Iš viso gautas žodžių kiekis yra 7,9 milijono.

10 lentelė. Žodžių dažnio pasiskirstymas visoje žodžių imtyje (Madrido ir Barselonos duomenys)

Žodžių sąrašo Nr.	**Dažnumas**	**Procentinė dalis**	**Suminė procentinė dalis**
1-100	3800604	50,06%	50,06%
101-200	855761	11,27%	61,33%
201-300	527640	6,95%	68,28%
301-400	373720	4,92%	73,21%
401-500	279113	3,68%	76,88%
501-25048	1755071	23,12%	100 %

Pastebima, kad pirmi 100 atskirų žodžių sudaro maždaug pusę visų žodžių imties, o pirmi 500 atskirų žodžių sudaro ~77 % visų žodžių imties.

Sukuriama nauja atsiliepimų-požymių matrica, sudaryta iš Ispanijos atvejo atsiliepimų bei jau sudarytu klientų patirtį atspindinčiu žodžių sąrašu. Ši matrica užpildoma svorių, gautų taikant vektorinės erdvės modelį, reikšmėmis. Taikomas faktorinės analizės metodas. Pirma, apskaičiuojamas bendras KMO rodiklis, gaunama reikšmė – 0,5, tai reiškia, kad duomenys faktorinei analizei yra tinkami tik iš dalies. Apskaičiuojamas MSA_i rodiklis kiekvienam požymiui ir po analizės pasirenkami tik tie požymiai, kurių $MSA_i > 0{,}59$. Atlikus šį žingsnį, bendras KMO matas padidėja iki 0,58. Skaičiuojamas Bartlett‘o sferiškumo kriterijus ir gaunama p reikšmė yra mažesnė už reikšmingumo lygmenį α, t. y., $0{,}0 < 0{,}5$ (28 pav.)

```
chi_square_value, p_value = calculate_bartlett_sphericity(df4)
round(chi_square_value, 1), round(p_value,3)
```

```
(363992.3, 0.0)
```

28 pav. Bartlett‘o sferiškumo kriterijus (Madrido ir Barselonos duomenys)

Atlikus optimalaus faktorių skaičiaus nustatymą, gaunama, kad optimaliausias faktorių skaičius yra 7, metodas jų išskyrimui – pagrindinių komponenčių metodas, sukimas – ortogonalusis faktorių sukimas *varimax*. Apibrėžus, kad būtų išvedami tik požymiai, kurių svoris faktoriuje yra daugiau nei 0,35, gaunamas rezultatas pateiktas 29 pav.

	0	1	2	3	4	5	6
cafes	0.38						
shops	0.54						
bars	0.71						
restaurants	0.78						
private		0.44					
room		0.71					
bathroom		0.77					
near			0.45				
station			0.76				
metro			0.79				
machine				0.43			
fridge				0.57			
microwave				0.65			
blankets					0.56		
towels					0.56		
various						0.37	
dirty						0.39	
broken						0.53	
sleep							-0.39
free							0.42
garage							0.55

29 pav. Faktorių išskyrimo rezultatas (Madrido ir Barselonos duomenys)

Kaip ir Vokietijos atveju gaunama, kad faktorius sudaro 2-4 požymiai. 11 lentelėje faktoriai apibūdinami bendru pavadinimu.

11 lentelė. Faktorių apibūdinimas (Madrido ir Barselonos duomenys)

Faktoriaus Nr.	**Faktoriaus pavadinimas**	**Požymiai**
1.	Paslaugos netoliese	Kavinės Parduotuvės Barai Restoranai
2.	Pagrindinis produktas	Privatus Kambarys Vonios kambarys

3.	Susisiekimas	Šalia Stotis Metro
4.	Būsto patogumai	Mašina Šaldytuvas Mikrobangų krosnelė
5.	Papildomi apartamento privalumai	Patalynė Rankšluosčiai
6.	Neigiama patirtis	Įvairus Nešvarus Sulūžęs
7.	Mišrus faktorius	Miegas Nemokamas Garažas

Skaičiuojama, kokią dispersijos dalį paaiškina 7 gauti faktoriai.

	0	1	2	3	4	5	6
cum_var	0.051044	0.098846	0.145672	0.181933	0.215281	0.247317	0.279203
var	0.051044	0.047802	0.046826	0.036261	0.033348	0.032036	0.031886

30 pav. Paaiškinama faktorių dispersijos dalis (Madrido ir Barselonos duomenys)

Gaunama, kad 7 faktoriai apibūdina 27,92 % visų požymių dispersijos.

Duomenys ruošiami regresinei analizei tokia eiga:

- apskaičiuojami faktorių svoriai (fragmentas pateiktas 9 priede);
- sudaroma duomenų matrica iš pradinių, naujai sukurtų požymių bei gautų faktorių;
- požymiai standartizuojami;
- nubraižius bendro būsto vertinimo ir požymių priklausomybės grafikus bei pastebėjus aiškiai išsiskiriančias reikšmes, požymiai yra filtruojami.

Sudarius tiesinės regresijos modelį su visais pradiniais kintamaisiais, gaunama, kad koreguotas apibrėžtumo koeficientas yra 40,12 %, bet yra kintamųjų, kurie tiesinės regresijos lygtyje yra nereikšmingi ($p > 0,05$), o VIF parodė, jog multikolinearumo problemos nėra. Pašalinus nereikšmingus požymius (FA3 ir FA4), 31 pav. pavaizduoti gaunami rezultatai.

Root MSE	0.74862	R-Square	0.4013
Dependent Mean	-0.00408	Adj R-Sq	0.4012
Coeff Var	-18330		

Parameter Estimates									
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	1	-0.31535	0.00526	-60.01	<.0001	0	0	-0.32565	-0.30505
FA1	1	0.00473	0.00122	3.87	0.0001	0.00453	1.01303	0.00233	0.00712
FA2	1	-0.03843	0.00129	-29.73	<.0001	-0.03563	1.06325	-0.04097	-0.03590
FA5	1	-0.03254	0.00166	-19.58	<.0001	-0.02321	1.04017	-0.03580	-0.02929
FA6	1	-0.02292	0.00172	-13.30	<.0001	-0.01586	1.05352	-0.02630	-0.01954
FA7	1	0.03519	0.00159	22.10	<.0001	0.02609	1.03156	0.03207	0.03831
guests_included	1	-0.04013	0.00111	-36.16	<.0001	-0.05560	1.75090	-0.04230	-0.03795
market	1	-0.29411	0.00262	-112.36	<.0001	-0.15061	1.33065	-0.29924	-0.28898
room_type	1	0.21632	0.00352	61.51	<.0001	0.10680	2.23244	0.20943	0.22322
bathrooms	1	0.03908	0.00293	13.32	<.0001	0.01806	1.36127	0.03333	0.04483
bedrooms	1	-0.07974	0.00209	-38.12	<.0001	-0.06652	2.25596	-0.08385	-0.07564
price	1	0.22446	0.00232	96.80	<.0001	0.16561	2.16786	0.21992	0.22901
minimum_nights	1	0.07059	0.00117	60.42	<.0001	0.07223	1.05831	0.06830	0.07288
host_is_superhost	1	1.09002	0.00234	465.71	<.0001	0.55386	1.04748	1.08543	1.09460
distance_from_center	1	0.05544	0.00141	39.28	<.0001	0.04856	1.13175	0.05267	0.05820
comment_length	1	-0.00618	0.00121	-5.12	<.0001	-0.00625	1.10128	-0.00854	-0.00381
compound	1	0.11139	0.00120	93.13	<.0001	0.11281	1.08648	0.10905	0.11374

31 pav. Tiesinės regresijos rezultatai po požymių šalinimo (Madrido ir Barselonos duomenys)

Šiuo atveju koreguoto apibrėžtumo koeficiento reikšmė taip pat nepasikeitė. Būsto šeimininko statusas (pagal standartizuotą įvertį iš „Standardized Estimate“) išlieka darančiu stipriausią įtaką.

Patikrinus sudaryto modelio išskirtis ir liekanas, gauta, jog duomenyse taip pat yra nemažai išskirčių. Sudaroma duomenų imtis iš 10000 atsitiktinai atrinktų stebinių ir pritaikius atspariosios regresijos metodą, gaunamas išskirčių ir įtakos taškų sąrašas (fragmentas pateikiamas 10 priede), sujungiamas su pradiniais duomenimis. Tada šalinami stebiniai, jei „PRobustDist“ > 400 arba „Outlier“ = 1, pašalinant grubiausias išskirtis. Duomenų imtis sumažėja iki 9418 stebinių. Kartojamas tiesinės regresijos modelis ir pašalinus nereikšmingus kintamuosius 32 pav. pateikiami gauti rezultatai.

Root MSE	0.76491	R-Square	0.3970
Dependent Mean	-0.01622	Adj R-Sq	0.3961
Coeff Var	-4715.94149		

Parameter Estimates									
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	1	-0.28296	0.03247	-8.72	<.0001	0	0	-0.34660	-0.21932
FA2	1	-0.04383	0.00905	-4.84	<.0001	-0.03987	1.05733	-0.06157	-0.02608
FA5	1	-0.06015	0.01150	-5.23	<.0001	-0.04279	1.04438	-0.08270	-0.03760
FA6	1	-0.03896	0.01257	-3.10	0.0019	-0.02529	1.03818	-0.06359	-0.01432
FA7	1	0.03021	0.01066	2.83	0.0046	0.02308	1.03432	0.00931	0.05111
guests_included	1	-0.02612	0.00768	-3.40	0.0007	-0.03591	1.74051	-0.04118	-0.01106
market	1	-0.32974	0.01810	-18.22	<.0001	-0.16610	1.29630	-0.36522	-0.29426
room_type	1	0.20081	0.02361	8.51	<.0001	0.09767	2.05564	0.15454	0.24708
bedrooms	1	-0.05529	0.01411	-3.92	<.0001	-0.04539	2.09283	-0.08294	-0.02763
price	1	0.18802	0.01589	11.83	<.0001	0.13659	2.07922	0.15686	0.21918
minimum_nights	1	0.04136	0.00812	5.09	<.0001	0.04190	1.05499	0.02545	0.05728
host_is_superhost	1	1.12117	0.01635	68.58	<.0001	0.56092	1.04328	1.08912	1.15322
distance_from_center	1	0.07088	0.01004	7.06	<.0001	0.05999	1.12670	0.05120	0.09057
compound	1	0.11334	0.00809	14.01	<.0001	0.11349	1.02327	0.09748	0.12920

32 pav. Tiesinės regresijos rezultatai po išskirčių šalinimo (Madrido ir Barselonos duomenys)

Gauta, kad koreguotas apibrėžtumo koeficientas sumažėjo iki 39,61 %, tačiau tai vis tiek patenkinamas rezultatas. Tikrinamos modelio prielaidos.

1. Liekamosios paklaidos pasiskirsčiusios pagal normalųjį skirstinį. Patikrinus prielaidą grafiškai pastebėta, kad liekamosios paklaidos nėra visiškai pasiskirsčiusios pagal normalųjį skirstinį. Tai parodė ir eksceso koeficiento reikšmė, lygi 3,01, rodanti smailesnę viršūnę lyginant su normaliuoju skirstiniu bei asimetrijos koeficientas, kurio reikšmė yra −1,11, rodanti, kad grafikas turi kairinę asimetriją. Prielaida yra tenkinama iš dalies.
2. Liekamųjų paklaidų vidurkis yra lygus 0. Tikrinama hipotezė, kad populiacijos liekamųjų paklaidų vidurkis yra lygus 0 – ji neatmetama ir prielaida yra tenkinama.
3. Liekamųjų paklaidų dispersijos yra lygios – homoskedastiškumo prielaida. Tikrinama hipotezė, kad populiacijos liekamųjų paklaidų dispersijos yra lygios – ji atmetama ir prielaida yra netenkinama.
4. Visos atsitiktinės paklaidos yra nepriklausomos, t. y., nėra autokoreliacijos. Prielaida tikrinama pagal Durbin‘o-Watson‘o statistiką ir gaunama, kad statistikos *p* reikšmė yra 1,130 (>0,05), tai reiškia, kad autokoreliacijos modelyje nėra, tačiau atsižvelgus į dešinę kritinę sritį – nulinė hipotezė turėtų būti atmetama. Prielaida yra tenkinama iš dalies.
5. Duomenyse nėra išskirčių. Pradinė analizė parodė, jog duomenyse yra nemažai išskirčių. Pritaikius atspariosios regresijos metodą pastebėta, kad ne visos išskirtys yra pašalintos, tačiau grubių išskirčių nebėra. Taigi, prielaida yra iš dalies tenkinama.
6. Tarp nepriklausomų kintamųjų nėra stiprios koreliacijos. Tarp likusių požymių yra stebima vidutinė arba silpna koreliacija. Prielaida yra tenkinama iš dalies.

Tikrinami heteroskedastiškumo pasikliautinieji intervalai, nes 3 prielaida buvo netenkinama.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Heteroscedasticity Consistent Standard Error	Heteroscedasticity Consistent t Value	Heteroscedasticity Consistent Pr > |t|	Variance Inflation	95% Confidence Limits		Heteroscedasticity Consistent 95% Confidence Limits	
Intercept	1	-0.28296	0.03247	-8.72	<.0001	0.03438	-8.23	<.0001	0	-0.34660	-0.21932	-0.35035	-0.21557
FA2	1	-0.04383	0.00905	-4.84	<.0001	0.00894	-4.90	<.0001	1.05733	-0.06157	-0.02608	-0.06136	-0.02630
FA5	1	-0.06015	0.01150	-5.23	<.0001	0.01236	-4.87	<.0001	1.04438	-0.08270	-0.03760	-0.08438	-0.03592
FA6	1	-0.03896	0.01257	-3.10	0.0019	0.01394	-2.80	0.0052	1.03818	-0.06359	-0.01432	-0.06628	-0.01164
FA7	1	0.03021	0.01066	2.83	0.0046	0.01171	2.58	0.0099	1.03432	0.00931	0.05111	0.00726	0.05316
guests_included	1	-0.02612	0.00768	-3.40	0.0007	0.00840	-3.11	0.0019	1.74051	-0.04118	-0.01106	-0.04258	-0.00966
market	1	-0.32974	0.01810	-18.22	<.0001	0.01774	-18.59	<.0001	1.29630	-0.36522	-0.29426	-0.36450	-0.29497
room_type	1	0.20081	0.02361	8.51	<.0001	0.02407	8.34	<.0001	2.05564	0.15454	0.24708	0.15362	0.24799
bedrooms	1	-0.05529	0.01411	-3.92	<.0001	0.01444	-3.83	0.0001	2.09283	-0.08294	-0.02763	-0.08358	-0.02699
price	1	0.18802	0.01589	11.83	<.0001	0.01775	10.59	<.0001	2.07922	0.15686	0.21918	0.15322	0.22282
minimum_nights	1	0.04136	0.00812	5.09	<.0001	0.00823	5.02	<.0001	1.05499	0.02545	0.05728	0.02522	0.05751
host_is_superhost	1	1.12117	0.01635	68.58	<.0001	0.01464	76.61	<.0001	1.04328	1.08912	1.15322	1.09248	1.14986
distance_from_center	1	0.07088	0.01004	7.06	<.0001	0.01012	7.01	<.0001	1.12670	0.05120	0.09057	0.05105	0.09072
compound	1	0.11334	0.00809	14.01	<.0001	0.00952	11.90	<.0001	1.02327	0.09748	0.12920	0.09467	0.13201

33 pav. Tiesinės regresijos rezultatai su heteroskedastiškumui atspariais pasikliautinaisiais intervalais (Madrido ir Barselonos duomenys)

Sudarytas galutinis tiesinės regresijos modelis ir gaunama, jog 4 prielaidos modelio tinkamumui yra tenkinamos iš dalies, 1 – pilnai, 1 – netenkinama. Gaunama, jog koreguotas apibrėžtumo koeficientas yra 0,3961, tai reiškia, kad 39,61 % bendro reitingo sklaidos apie vidurkį galima paaiškinti tiesine regresija požymių „FA2“, „FA5“, „FA6“, „FA7“ „guests_included“, „market“, „room_type“, „bedrooms“, „price“, „minimum_nights“, „host_is_superhost“, „distance_from_center“, „compound“ atžvilgiu. Gaunama analizinė tiesinės regresijos išraiška pagal regresijos lygties koeficientus:

org_rating = −0,28 – 0,04*__FA2__ – 0,06*__FA5__ – 0,04*__FA6__ + 0,03*__FA7__ – 0,03*__guests_included__ – 0,33*__market__ + 0,20*__room_type__ – 0,06*__bedrooms__ + 0,19*__price__ + 0,04*__minimum_nights__ + 1,12*__host_is_superhost__ + 0,07*__distance_from_center__ + 0,11*__compound__

Atsižvelgus į analizinę modelio išraišką galima teigti, kad, pavyzdžiui, padidinus svečių skaičiaus požymį vienu vienetu, bendras būsto vertinimas vidutiniškai sumažėja 0,03 balo. Pagal 33 pav. su 95 % garantija galima teigti, kad, pavyzdžiui, svečių skaičių padidinus vienu vienetu, bendras būsto vertinimo vidurkio sumažėjimas bus intervale nuo –0,04 balo iki –0,01 balo.

Taip pat aprašoma gauta standartizuota tiesinės regresijos lygtis:

org_rating* = – 0,04*__FA2__ – 0,04*__FA5__ – 0,03*__FA6__ + 0,02*__FA7__ – 0,04*__guests_included__ – 0,16*__market__ + 0,10*__room_type__ – 0,05*__bedrooms__ + 0,14*__price__ + 0,04*__minimum_nights__ + 0,56*__host_is_superhost__ + 0,06*__distance_from_center__ + 0,11*__compound__

Šiuo atveju pastebima, kad, kaip ir Vokietijos atveju, stipriausią teigiamą įtaką bendram būsto vertinimui turi šeimininko statusas, o stipriausią neigiamą – miegamųjų skaičius. Galutiniame modelyje likę faktoriai „Pagrindinis produktas“, „Papildomi apartamento privalumai“ bei „Neigiama patirtis“, turi neigiamą įtaką bendram būsto vertinimui, o faktorius „Mišrus faktorius“ – teigiamą.

Išvados

1. Literatūros analizė parodė, kad apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajoms tirti naudojami jų internetiniai atsiliepimai ir įvertinimai. Pagrindiniai naudoti analizės metodai yra teksto tyryba, požymių dimensijos mažinimas ir prognozavimo analitika. Taip pat nustatyta, kad apgyvendinimas privačiame būste nagrinėtas mažiau, nei viešbučiuose.

2. Darbe pasiūlyta apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų modelių metodika ir jos programinė realizacija, kuri apima: duomenų tvarkymą ir paruošimą analizei, požymių atranką, naujų požymių kūrimą taikant teksto tyrybos ir kitus metodus, požymių erdvės dimensijos mažinimą taikant faktorinės analizės metodą, apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų regresijos modelio sudarymą ir jo tikimo analizuojamiems duomenims tyrimą. Sukurtos priemonės palengvina ir automatizuoja apgyvendinimo paslaugų klientų patirties ir pasitenkinimo sąsajų modelių sudarymą ir tyrimą.

3. Sukurtų priemonių taikymas realių duomenų analizei, parodė, kad jos sprendžia darbe suformuluotus uždavinius. Išanalizavus Berlyno ir Miuncheno miestų duomenis, išskirti 19 požymių. Sudarius regresinės analizės modelį gauta, kad stipriausią įtaką bendram būsto vertinimui daro naujai pasiūlytas požymis – šeimininko statusas. Modelis tenkina prielaidas ir paaiškina 28,1 % bendro būsto vertinimo sklaidos apie vidurkį tiesine regresija atstumo iki centro, kainos, sentimentų įverčio, minimalaus nakvynių skaičiaus, kambario tipo, miegamųjų skaičiaus, svečių skaičiaus, miesto, šeimininko statuso ir faktorių pagrindiniai produkto požymiai, apartamento požymiai ir susisiekimas atžvilgiu. Atlikus Madrido ir Barselonos miestų privataus būsto nuomos analizę, gautas tiesinės regresijos modelis, kuris paaiškina 39,6 % bendro būsto vertinimo sklaidos apie vidurkį tiesine regresija išskirtų požymių atžvilgiu. Stipriausią įtaką bendram būsto vertinimui daro šeimininko statuso ir miesto požymiai.

Literatūros sąrašas

1. *REGISTRŲ CENTRAS* [interaktyvus] [žiūrėta 2020-03-10]. Prieiga per: https://www.registrucentras.lt/jar/fa/klasif/v_rusys.php?kla_nr=2
2. EUROSTAT. *Accommodation and food service statistics - NACE Rev. 2* [interaktyvus]. ISSN 2443-8219. [žiūrėta 2020-03-20]. Prieiga per: https://ec.europa.eu/eurostat/statistics-explained/index.php/Accommodation_and_food_service_statistics_-_NACE_Rev._2#Sectoral_analysis
3. *CAMBRIDGE DICTIONARY* [interaktyvus] [žiūrėta 2020-03-10]. Prieiga per: https://dictionary.cambridge.org/dictionary/english/
4. XU F., LA L., ZHEN F., LOBSANG T., HUANG C. A data-driven approach to guest experiences and satisfaction in sharing. *Journal of Travel & Tourism Marketing* [interaktyvus]. 2019, vol. **36**(4), 484-496 [žiūrėta 2020-03-15]. ISSN 1540-7306. doi: https://doi.org/10.1080/10548408.2019.1570420
5. CHENG M., JIN X. What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management* [interaktyvus]. Elsevier, 2019, vol. **76**(A), 58-70 [žiūrėta 2020-03-15]. ISSN 0278-4319. Prieiga per: Science Direct; doi: https://doi.org/10.1016/j.ijhm.2018.04.004
6. Festila M., Müller S. THE IMPACT OF TECHNOLOGY-MEDIATED CONSUMPTION ON IDENTITY: THE CASE OF AIRBNB. *Proceedings of the 50th Hawaii International Conference on System Sciences* [interaktyvus]. 2017, [žiūrėta 2020-03-15]. ISBN 978-0-9981331-0-2. Prieiga per: http://hdl.handle.net/10125/41157; doi: doi.org/10.24251/HICSS.2017.007
7. LIANG L. J., CHOI H. C., JOPPE M. Exploring the relationship between satisfaction, trust and switching intention, repurchase intention in the context of Airbnb. *International Journal of Hospitality Management* [interaktyvus]. Elsevier, 2018, vol. **69**, 41-48 [žiūrėta 2020-03-16]. ISSN 0278-4319. Prieiga per: Science Direct; doi: http://dx.doi.org/10.1016/j.ijhm.2017.10.015
8. XIANG Z., SCHWARTZ Z., GERDES JR. J. H., UYSAL M. What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management* [interaktyvus]. Elsevier, 2015, vol. **44**, 120-130 [žiūrėta 2020-03-08]. ISSN 0278-4319. Prieiga per: Science Direct; doi: https://doi.org/10.1016/j.ijhm.2014.10.013
9. HARGREAVES C. A. Analysis of Hotel Guest Satisfaction Ratings and Reviews: An Application in Singapore. *American Journal of Marketing Research* [interaktyvus]. 2015, vol. **1**(4), 208-214 [žiūrėta 2020-03-08]. Prieiga per: Research Gate: https://www.researchgate.net/publication/291832800_Analysis_of_Hotel_Guest_Satisfaction_Ratings_and_Reviews_An_Application_in_Singapore
10. LIU Y., TEICHERT T., ROSSI M., LI H., HU F. Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews. *Tourism Management* [interaktyvus]. Elsevier, 2017, vol. **59**, 554-563 [žiūrėta 2020-03-08]. ISSN 0261-5177. Prieiga per: Science Direct; doi: http://doi.org/10.1016/j.tourman.2016.08.012
11. ZHAO Y., XU X., WANG M. Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management* [interaktyvus]. Elsevier, 2019, vol. **76**(A), 111-121 [žiūrėta 2020-03-08]. ISSN 0278-4319. Prieiga per: Science Direct; doi: https://doi.org/10.1016/j.ijhm.2018.03.017
12. AHANI A., NILASHI M., YADEGARIDEHKORDI E., SANZOGNI L., TARIK A. R., KNOX K., SAMAD S., IBRAHIM O. Revealing customers‘ satisfaction and preferences through online review analysis: The case of Canary Islands hotels. *Journal of Retailing and Consumer Services*

[interaktyvus]. Elsevier, 2019, vol. **51**, 331-343 [žiūrėta 2020-03-08]. ISSN 0969-6989. Prieiga per: Science Direct; doi: https://doi.org/10.1016/j.jretconser.2019.06.014
13. PADMA P., AHN J. Guest satisfaction & dissatisfaction in luxury hotels: An application of big data. *International Journal of Hospitality Management* [interaktyvus]. Elsevier, 2020, vol. **84** [žiūrėta 2020-03-30]. ISSN 0278-4319. Prieiga per: Science Direct; doi: https://doi.org/10.1016/j.ijhm.2019.102318
14. BAEK J., CHOE Y., OK C. M. Determinants of hotel guests‘ service experiences: an examination of differences between lifestyle and traditional hotels. *Journal of Hospitality Marketing & Management* [interaktyvus]. 2020, vol. **29**, 88-105 [žiūrėta 2020-03-30]. ISSN 1936-8631. Prieiga per: doi: https://doi.org/10.1080/19368623.2019.1580173
15. LI F. S., RYAN C. Western guest experiences of a Pyongyang international hotel, North Korea: Satisfaction under conditions of constrained choice. *Tourism Management* [interaktyvus]. Elsevier, 2020, vol. **76** [žiūrėta 2020-03-30]. ISSN 0261-5177. Prieiga per: Science Direct; doi: https://doi.org/10.1016/j.tourman.2019.07.001
16. STHAPIT E., JIMÉNEZ-BARRETO J. Exploring tourists‘ memorable hospitality experiences: An Airbnb perspective. *Tourism Management Perspectives* [interaktyvus]. Elsevier, 2018, vol. **28**, 83-92 [žiūrėta 2020-04-01]. ISSN 2211-9736. Prieiga per: Science Direct; doi: https://doi.org/10.1016/j.tmp.2018.08.006
17. JOSEPH, George and Vinu VARGHESE. Analyzing Airbnb Customer Experience Feedback Using Text Mining [žiūrėta 2020-04-01]. Iš: SIGALA, M., et al. *Big Data and Innovation in Tourism, Travel, and Hospitality*. Singapore (ZG): Springer, 2019, pp. 147–162. ISBN 9789811363382.
18. JIAO J., BAI S. An empirical analysis of Airbnb listings in forty American cities. *Cities* [interaktyvus]. Elsevier, 2020, vol. **99** [žiūrėta 2020-04-01]. ISSN 0264-2751. Prieiga per: Science Direct; doi: https://doi.org/10.1016/j.cities.2020.102618
19. JANILIONIS V. Išklausyto modulio „Didžiųjų duomenų rinkinių tyrybos metodai“ paskaitų medžiaga (2018 m.).
20. KAVALIAUSKAS M. Išklausyto modulio „Daugiamatės statistinės analizės modeliai“ paskaitų medžiaga (2019 m.).
21. JANILIONIS V. Išklausyto modulio „Daugiamatės statistinės analizės modeliai“ paskaitų medžiaga (2019 m.).
22. VAITKEVIČIUS, Raimundas ir Aušra SAUDARGIENĖ. *Psichologinių tyrimų duomenų analizė. Praktikos darbai*. Kaunas: VDU leidykla, 2010. ISBN 9789955125617.
23. ČEKANAVIČIUS, Vydas ir Gediminas MURAUSKAS. *Taikomoji regresinė analizė socialiniuose tyrimuose*. Vilnius: Vilniaus universitetas, 2014. ISBN 9786094593000.
24. JANILIONIS V. Mokomoji medžiaga „Mokymai apie kiekybinių ir kokybinių HSM tyrimų duomenų analizės metodus“: Daugialypės regresinės analizės taikymas socialiniuose tyrimuose [interaktyvus] [žiūrėta 2020-04-10]. Prieiga per: LiDA: http://www.lidata.eu/index.php?file=files/apie.html
25. *JUPYTER*. [interaktyvus] [žiūrėta 2020-04-15]. Prieiga per: https://jupyter.org/
26. *WHAT IS PYTHON? EXECUTIVE SUMMARY*. [interaktyvus] [žiūrėta 2020-04-15]. Prieiga per: https://www.python.org/doc/essays/blurb/
27. *PYTHON. ABOUT*. [interaktyvus] [žiūrėta 2020-04-15]. Prieiga per: https://www.python.org/about/

28. *ABOUT SAS*. [interaktyvus] [žiūrėta 2020-04-15]. Prieiga per: https://www.sas.com/en_us/company-information.html
29. *Kent State University. STATISTICAL & QUALITATIVE DATA ANALYSIS SOFTWARE: ABOUT SAS*. [interaktyvus] [žiūrėta 2020-04-15]. Prieiga per: https://libguides.library.kent.edu/statconsulting/SAS
30. RENCHER, Alvin C. and William F. CHRISTENSEN. Chapter 13: Exploratory factor analysis. Iš: *Methods of Multivariate Analysis*. 2012, pp. 435-478. ISBN 9781118391655. Prieiga per: ProQuest Ebook Central: https://ebookcentral.proquest.com/lib/ktu-ebooks/home.action
31. *INSIDE AIRBNB. GET THE DATA*. [interaktyvus]. [žiūrėta 2020-02-05] Prieiga per: http://insideairbnb.com/get-the-data.html
32. MACKENZIE Charles E. *Coded Character Sets, History and Development*. United States of America: Library of Congress, 1980. ISBN 0201144603.
33. *UNIVERSITY OF GLASGOW. SCHOOL OF COMPUTING SCIENCE*. [interaktyvus] [žiūrėta 2020-02-25]. Prieiga per: https://www.gla.ac.uk/schools/computing/research/researchsections/ida-section/informationretrieval/
34. *STOP WORD LIST*. [interaktyvus] [žiūrėta 2020-02-25]. Prieiga per: http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words?fbclid=IwAR0g1QgPGvIRJ7dKHiPcSEHUjcVN3ShXXFju7uqoH4nAneg85L9DzO7wHe4
35. LEE D.L., CHUANG H., SEAMONS K. Document Ranking and the Vector-Space Model. *IEEE Software* [interaktyvus]. 1997, vol. **14**(2) [žiūrėta 2020-04-23]. ISSN 0740-7459. Prieiga per: doi: https://doi.org/10.1109/52.582976
36. CHOPDE N. R., NICHAT M. K. Landmark Based Shortest Path Detection by Using A* and Haversine Formula. *International Journal of Innovative Research in Computer and Communication Engineering* [interaktyvus]. 2013, vol. **1**(2) [žiūrėta 2020-05-05]. ISSN 2320-9798. Prieiga per: Research Gate: https://www.researchgate.net/publication/282314348_Landmark_based_shortest_path_detection_by_using_A_Algorithm_and_Haversine_Formula
37. *THE ROBUSTREG PROCEDURE* [interaktyvus] [žiūrėta 2020-05-06]. Prieiga per: https://documentation.sas.com/?docsetId=statug&docsetVersion=15.1&docsetTarget=statug_rreg_overview.htm&locale=en
38. HUBER, P. J. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics* [interaktyvus]. 1973, vol. **1**(5), 799-821 [žiūrėta 2020-05-06]. Prieiga per: doi: https://doi.org/10.1214/aos/1176342503
39. ROUSSEEUW P. Least Median of Squares Regression. *Journal of the American Statistical Association* [interaktyvus]. 1984, vol. **79**(388), 871-880 [žiūrėta 2020-05-06]. Prieiga per: doi: https://doi.org/10.2307/2288718
40. ROUSSEEUW P., YOHAI V. Robust Regression by Means of S-Estimators [interaktyvus]. 1984 [žiūrėta 2020-05-06]. Prieiga per: doi: https://doi.org/10.1007/978-1-4615-7821-5_15
41. YOHAI V. J. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics* [interaktyvus]. 1987, vol. **15**(2), 642-656 [žiūrėta 2020-05-06]. Prieiga per: doi: https://doi.org/10.1214/aos/1176350366
42. DZEMYDA, Gintautas, KURASOVA Olga, ŽILINSKAS Julius. *Daugiamačių duomenų vizualizavimo metodai*. Vilnius: Mokslo aidai, 2008. ISBN 9789986680420.

43. *ISO 639 LANGUAGE CODES.* [interaktyvus] [žiūrėta 2020-03-02]. Prieiga per: https://www.iso.org/iso-639-language-codes.html
44. WALLISCH P. Chapter 19 - Principal Components Analysis. *MATLAB for Neuroscientists (Second Edition)* [interaktyvus]. Elsevier, 2014, 305-315 [žiūrėta 2020-03-30]. Prieiga per: doi: https://doi.org/10.1016/B978-0-12-383836-0.00017-5
45. *WHAT IS R?* [interaktyvus] [žiūrėta 2020-04-15]. Prieiga per: https://www.r-project.org/about.html
46. *IBM SPSS SOFTWARE* [interaktyvus] [žiūrėta 2020-04-15]. Prieiga per: https://www.ibm.com/analytics/spss-statistics-software
47. *AIRBNB. ABOUT US* [interaktyvus] [žiūrėta 2020-05-12]. Prieiga per: https://news.airbnb.com/about-us/

Priedai

1 priedas. Modelių programinei realizacijai reikalingų Python paketų diegimas

```
###Modelių programinei realizacijai reikalingų Python paketų diegimas
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import string
import re
from langdetect import detect
import collections as cl
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from wordcloud import WordCloud
import time
import swifter
from nltk.corpus import wordnet
from factor_analyzer import FactorAnalyzer
import seaborn as sns
import math
from factor_analyzer.factor_analyzer import calculate_kmo, calculate_bartlett_sphericity
from sklearn import preprocessing
from statsmodels.formula.api import ols
import statsmodels.api as sm
impot warnings
warnings.filterwarnings("ignore")
###Pagrindinių Python funkcijų apibrėžimas
def show(dataframe, x): #duomenų matricos sudarymas
    with pd.option_context("display.max_rows",None, "display.max_columns",None):
        display(dataframe.head(x))
def charts(df, metric): #duomenų reikšmių skaičiavimas, vaizdavimas
    fig, (ax1, ax2) = plt.subplots(1, 2, figsize = (20, 5))
    ax1.set_title('Share of ' + metric)
    ax2.set_title('Count of ' + metric)
    df[metric].value_counts(normalize = True).plot(kind = 'bar', ax = ax1)
    df[metric].value_counts(normalize = False).plot(kind = 'bar', ax = ax2);
def distance_from_center(lat1, lon1, lat2, lon2): #atstumo skaičiavimas
    radius = 6371 # km
    dlat = math.radians(lat2-lat1)
    dlon = math.radians(lon2-lon1)
    a = math.sin(dlat/2) * math.sin(dlat/2) + math.cos(math.radians(lat1)) \
        * math.cos(math.radians(lat2)) * math.sin(dlon/2) * math.sin(dlon/2)
    c = 2 * math.atan2(math.sqrt(a), math.sqrt(1-a))
    d = round(radius * c, 1)
    return d
def nltk_sentiment(sentence): #sentimentų analizė
    nltk_sentiment = SentimentIntensityAnalyzer()
    score = nltk_sentiment.polarity_scores(sentence)
    return score
```

2 priedas. Duomenų paruošimas

```
###Berlyno miesto būstų ir atsiliepimų duomenys
berlin_listings = pd.read_csv('listings_berlin.csv', engine='python')
print(berlin_listings.shape) #išvedamas duomenų matricos dydis
berlin_listings.head(3) #išvedami pirmi 3 įrašai
berlin_reviews = pd.read_csv('reviews_berlin.csv', engine='python')
berlin_reviews.drop(columns = 'id', inplace = True)
berlin_reviews.rename(columns = {'listing_id':'id'}, inplace = True)
print(berlin_reviews.shape)
berlin_reviews.head(3)
```

```
###Miuncheno miesto būstų ir atsiliepimų duomenys
munich_listings = pd.read_csv('listings_munich.csv', engine='python')
print(munich_listings.shape)
munich_listings.head(3)
munich_reviews = pd.read_csv('reviews_munich.csv', engine='python')
munich_reviews.drop(columns = 'id', inplace = True) #požymio pašalinimas
munich_reviews.rename(columns = {'listing_id':'id'}, inplace = True) #požymio pervadinimas
print(munich_reviews.shape)
munich_reviews.head()
###Berlyno ir Miuncheno duomenų sujungimas
berlin = berlin_reviews.merge(berlin_listings, on='id')
munich = munich_reviews.merge(munich_listings, on='id')
data = berlin.append(munich)
data.reset_index(drop=True, inplace = True)
print(berlin.shape, munich.shape)
show(data, 3)
###Požymių atranka
relevant_col = ['date', 'comments', 'guests_included', 'review_scores_rating', 'latitude', 'longitude', 'market', 'property_type', 'room_type', 'bathrooms', 'bedrooms', 'price', 'minimum_nights', 'host_is_superhost', 'host_name']
data = data[relevant_col]
data.rename(columns = {'review_scores_rating': 'org_rating'}, inplace = True)
data.info() #informacija apie požymius: pavadinimas, netrūkstamų reikšmių kiekis, tipas
data['price'] = data['price'].str.replace('$', '').str.replace(',', '').astype('float64') #požymio tipo pakeitimas
%%time #funkcija, po jos įvykdymo grąžinanti laiką, per kurį buvo įvykdyta
for col in data.select_dtypes(exclude=['object']).columns:
    data[col] = data[col].apply(lambda x: None if x == 0 else x) #nulinės reikšmės keičiamos į trūkstamas
data.dropna(inplace = True) #trinamos trūkstamos reikšmės
data.shape
data['org_rating'] = data['org_rating'] / 10 #skalės pakeitimas iš 100 į 10
data['room_type'].value_counts(normalize = True).round(2) #išvedamos požymio reikšmių kiekio procentinės dalys
data['room_type'] = [0 if x == 'Entire home/apt' else 1 for x in data['room_type']] #reikšmių perkodavimas
data['room_type'].value_counts(normalize = True).round(2)
data['host_is_superhost'].value_counts(normalize = True).round(2)
data['host_is_superhost'] = [1 if x == "t" else 0 for x in data['host_is_superhost']]
data['host_is_superhost'].value_counts(normalize = True).round(2)
data['market'].value_counts() #išvedamas požymio reikšmių kiekis
data = data[(data['market'] == 'Berlin') | (data['market'] == 'Munich')] #požymio filtravimas paliekant konkrečias reikšmes
data.shape
data['market'] = [0 if x == 'Berlin' else 1 for x in data['market']]
data['market'].value_counts(normalize = True).round(2)
###Kuriamas naujas požymis – atstumas iki centro
lat_berlin, lon_berlin = (52.519991, 13.404902)
lat_munich, lon_munich = (48.135303, 11.581940)
%time data['distance_from_center'] = [distance_from_center(lat_berlin, lon_berlin, x, y) if z == 0 else distance_from_center(lat_munich, lon_munich, x, y) for x, y, z in zip(data['latitude'], data['longitude'], data['market']) ]
###Kuriamas naujas požymis – atsiliepimo ilgis
%%time
data['comment_length'] = data['comments'].str.replace('[^A-Za-z0-9]+',' ')
data['comment_length'] = data['comment_length'].str.replace(' +', ' ')
data['comment_length'] = data['comment_length'].apply(lambda x: len(str(x).split()))
print("Word count test (symbols are not counted): \n", data['comment_length'].loc[720571], "\n\n", data['comments'].loc[720571])
###Formuojamas duomenų pjūvis
print('\nMin: ', pd.to_datetime(data['date']).min(),
      '\nMax: ', pd.to_datetime(data['date']).max()) #išvedamos mažiausia ir didžiausia požymio reikšmės
data = data[(data['date'] > '2015-01-01') & (data['date'] < '2020-01-01')]
data.shape
plt.figure(figsize = (15, 5)) #braižoma požymio reikšmių stulpelinė diagrama
data['property_type'].value_counts(normalize = True).plot(kind = 'bar')
plt.xlabel("property_type")
plt.ylabel("count");
data = data[data['property_type'] == 'Apartment']
data.shape
```

```
plt.figure(figsize = (15, 5))
data['minimum_nights'].value_counts(normalize = True).plot(kind = 'bar')
plt.xlabel("minimum_nights")
plt.ylabel("count");
data = data[data['minimum_nights'] <= 5]
print(data.shape)
```
###Požymių ir bendro būsto vertinimo priklausomybės grafinis vaizdavimas
```
x = ['bathrooms','bedrooms','price','guests_included','distance_from_center']
for i in x:
   print(i)
   plt.scatter(x=i, y="org_rating", data=data)
   plt.show()
   continue
```
###Požymių filtravimas
```
data = data[(data['bathrooms'] > 0) & (data['bathrooms'] <= 3)]
data = data[(data['bedrooms'] > 0) & (data['bedrooms'] <= 4)]
data = data[(data['price'] >= 20) & (data['price'] <= 400)]
data = data[(data['guests_included'] >= 1) & (data['guests_included'] <= 6)]
data = data[data['distance_from_center'] <= 18]
print(data.shape)
for i in x:
   print(i)
   plt.scatter(x=i, y="org_rating", data=data)
   plt.show()
   continue
```
###Formuojama duomenų matrica šalinant analizėje nenaudojamus požymius
```
data.drop(columns = ['latitude','longitude', 'property_type', 'date'], inplace = True)
data.head(3)
```

3 priedas. Klientų atsiliepimų teksto tyryba

###Šalinami atsiliepimai, parašyti ne ASCII simboliais [34]
```
print('Count of rows before cleaning: ', len(data))
data = data[~data['comments'].str.contains(r'[^\x00-\x7F]+')]
print('Count of rows after non-ASCII comments deleted: ', len(data))
data = data[data['comments'].str.contains('[A-Za-z]')] #šalinami atsiliepimai, sudaryti iš skaičių
```
###Taikomas kalbos atpažinimo paketas
```
%time data['language'] = data['comments'].swifter.allow_dask_on_strings().apply(detect)
data[['comments', 'language']][80:95]
data['language'].value_counts(normalize = True).round(5)
data2 = data[data.language == 'en'] #paliekami tik anglų kalba parašyti atsiliepimai
data2.drop(columns = ['language'], inplace = True)
print('Count of rows with all languages: ', len(data),
     '\nCount of rows only EN language: ', len(data2),
    '\nDifference: ', len(data) - len(data2))
```
###Duomenų matricos transformavimas
```
%%time
data2['comments'] = data2['comments'].str.replace('[^A-Za-z0-9]+',' ') #iš atsiliepimų šalinami bet kokie ne raidiniai simboliai
data2['comments'] = data2['comments'].str.replace("'",' ').replace('"',' ') # iš atsiliepimų šalinami apostrofai
data2['comments'] = data2['comments'].str.replace('\\n', ' ').replace('\\r', ' '#iš atsiliepimų šalinami kėlimai į kitą eilutę
data2 = data2[data2['comments'].notnull()]
data2['comments'] = data2['comments'].apply(lambda x: ' '.join(word.lower() for word in x.split() if not any(i.isdigit() for i in word))) #šalinami skaičiai
data2['comments'] = list(map(lambda x: ' '.join(word for word in x.split() if len(word) >= 2), data2['comments'])) #šalinami tik vieną simbolį sudarantys žodžiai
data2['comments'] = list(map(lambda x: ' '.join(word.replace(' +', ' ') for word in x.split()), data2['comments'])) #atsiliepimų kintamajame dvigubi (ar didesni) tarpai keičiami į vieną
data2.dropna(inplace = True)
for i in data['comments'][0:3]:
   print(i, '\n', "-" * 50)
print('\n')
for i in data2['comments'][0:3]:
```

```
    print(i, '\n', "-" * 50)
data2.shape
```
###Nereikšminių žodžių sąrašas
```
stopwords = ['a', 'about', 'above', 'across', 'after', 'afterwards','again', 'against', 'all', 'almost', 'alone', 'along', 'already', 'aren',
'also', 'although', 'always', 'am', 'among', 'amongst', 'amoungst', 'amount', 'an', 'and', 'another', 'any', 'anyhow', 'anyone',
'anything', 'anyway', 'anywhere', 'are', 'arent', 'around', 'as', 'at', 'back', 'be', 'became', 'because', 'become', 'becomes',
'becoming', 'been', 'before', 'beforehand', 'behind', 'being', 'below', 'beside', 'besides', 'between', 'beyond', 'bill', 'both',
'bottom', 'but', 'by', 'call', 'can', 'cannot', 'cant', 'co', 'computer', 'con', 'could', 'couldnt', 'cry', 'de', 'dont', 'day', 'days',
'describe', 'detail', 'did', 'didn',  'do', 'done', 'down', 'due', 'during', 'each', 'eg', 'eight', 'either', 'eleven', 'else', 'elsewhere',
'empty', 'enough', 'etc', 'even', 'ever', 'every', 'everyone', 'everything', 'haven', 'hadn', 'everywhere', 'except', 'few', 'fifteen',
'fifty', 'fill', 'find', 'fire', 'first', 'five', 'for', 'former', 'formerly', 'forty', 'found', 'four', 'from', 'front', 'full', 'further', 'get', 'give',
'go', 'had', 'has', 'hasnt', 'have', 'he', 'hence', 'her', 'here', 'hereafter', 'hereby', 'herein', 'hereupon', 'hers', 'herself', 'him',
'himself', 'his', 'how', 'however', 'hundred', 'i', 'ie', 'if', 'in', 'isn', 'inc', 'indeed', 'interest', 'into', 'is', "isnt", 'it', 'its', 'itself',
'keep', 'last', 'latter', 'latterly', 'least', 'less', 'ltd', 'made', 'many', 'may', 'me', 'meanwhile', 'might', 'mill', 'mine', 'min', 'max',
'more', 'moreover', 'most', 'mostly', 'move', 'much', 'must', 'my', 'myself', 'name', 'namely', 'neither', 'never', 'nevertheless',
'next', 'nine', 'no', 'nobody', 'none', 'noone', 'nor', 'not', 'nothing', 'now', 'nowhere', 'of', 'off', 'often', 'on','once', 'one', 'only',
'onto', 'or', 'other', 'others', 'otherwise', 'our', 'ours', 'ourselves', 'out', 'over', 'own', 'part', 'per', 'perhaps', 'please', 'put', 'rather',
're', 's', 'same', 'see', 'seem', 'seemed', 'seeming', 'seems', 'serious', 'several', 'she', 'should', 'show', 'side', 'since', 'sincere',
'six', 'sixty', 'so', 'some', 'somehow', 'someone', 'something', 'sometime', 'sometimes', 'somewhere', 'still', 'such', 'system',
'take', 'ten', 'than', 'that', 'the', 'their', 'them', 'themselves', 'then', 'thence', 'there', 'thereafter', 'thereby', 'therefore', 'therein',
'thereupon', 'these', 'they', 'thick', 'thin', 'third', 'this', 'those', 'though', 'three', 'three', 'through', 'throughout', 'thru', 'thus', 'to',
'together', 'too', 'top', 'toward', 'towards', 'twelve', 'twenty', 'two', 'un', 'under', 'until', 'up', 'upon', 'u', 'us', 'very', 'via', 'was',
'we', 'well', 'were', 'what', 'whatever', 'when', 'whence', 'whenever', 'where', 'whereafter', 'whereas', 'whereby', 'wherein',
'whereupon', 'wherever', 'whether', 'which', 'while', 'whither', 'who', 'whoever', 'whole', 'whom', 'whose', 'why', 'will', 'with',
'within', 'without', 'would', 'yet', 'you', 'your', 'yours', 'yourself', 'yourselves', 'airbnb', 'airbnbs', 'apartment', 'appartment',
'flat', 'berlin', 'really', 'just','highly', 'recommended', 'lot', 'lots', 'need', 'needed', 'time', 'area', 'like', 'definitely', 'feel', 'hosts',
'host', 'gave', 'stayed', 'ubahn', 'felt', 'places', 'minutes', 'exactly', 'thanks', 'thank', 'away', 'pictures', 'people', 'right', 'tips',
'train', 'didnt', 'local', 'transportation', 'met', 'kreuzberg', 'make', 'information', 'night', 'meet', 'way', 'went', 'minute', 'eat',
'coffee', 'late', 'come', 'person', 'nights', 'absolutely', 'welcome', 'want', 'staying', 'walking', 'neighbourhood', 'check', 'bit',
'things', 'building', 'help', 'main', 'munich', 'munchen']
stopwords = list(dict.fromkeys(stopwords))
stopwords = [x.lower() for x in stopwords]
print("Amount of unique stop words: ", len(stopwords))
```
###Šeimininkų vardų sąrašas
```
host_names = data2['host_name'][data2['host_name'].notnull()]
host_names = set(host_names)
host_names = [x.lower() for x in host_names]
data2.drop(columns = ['host_name'], inplace = True)
print("Amount of unique host names: ", len(host_names))
```
###Šalinami nereikšminiai žodžiai ir šeimininkų vardai
```
%%time
data2['comments'] = list(map(lambda x: ' '.join(word for word in x.split() if word not in stopwords), data2['comments']))
data2['comments'] = list(map(lambda x: ' '.join(word for word in x.split() if word not in host_names), data2['comments']))
data2.shape
data2[data2.comments.duplicated()] #ieškoma dublikatų
data2.comments.loc[1627] #konkretaus atsiliepimo išvedimas
data2 = data2[~data2.comments.str.contains('automated posting')] #trinami atsiliepimai, kuriuose yra nurodyti žodžiai
data2.shape
data2['cleaned_word_count'] = list(map(lambda x: len(x.split()), data2['comments']))
data2 = data2[data2['cleaned_word_count'] > 1] #paliekami tik tie atsiliepimai, kai juos sudaro daugiau nei vienas žodis
data2.drop(columns = 'cleaned_word_count', inplace = True)
data2.shape
```
###Taikoma sentimentų analizė
```
%time nltk_results = data2['comments'].swifter.allow_dask_on_strings().apply(lambda x: nltk_sentiment(x))
nltk_results
nltk_results = pd.DataFrame(nltk_results.astype(str).str.split(', ').tolist(), columns=['neg','neu', 'pos', 'compound'])
nltk_results.drop(columns = ['neg','neu', 'pos'], inplace = True)
nltk_results.set_index(data2.index, inplace = True)
for i in nltk_results.columns:
    nltk_results[i] = nltk_results[i].str.split(' ').str[1]
    nltk_results[i] = nltk_results[i].str.split('}').str[0]
    nltk_results[i] = nltk_results[i].astype(float)
nltk_results.describe()
```

```
data2 = data2.join(nltk_results)
data2.head(5)
```
###Atsiliepimų transformavimas
```
full_word_list = str(list(data2['comments'].values)).split()
word_list = [re.sub(r'[^\w\s_]+', ' ', x) for x in full_word_list]
word_list = [x.rstrip().lstrip() for x in word_list]
word_list = [x for x in word_list if len(x) > 2]
print('Unique word count: ', len(set(word_list)))
print('Total word count: ', len(word_list), '\n')
print(word_list[:100])
#ieškomi ne angliški žodžiai
not_en_words = [w for w in set(word_list) if not wordnet.synsets(w)]
print('Unique Non-English word count: ', len(not_en_words))
#šalinami ne angliški žodžiai
cleaned_word_list = [w for w in set(word_list) if w not in not_en_words]
print('Unique word count before removing Non-English words: ', len(set(word_list)))
print('Unique word count after removing Non-English words: ', len(set(cleaned_word_list)))
print('Difference: ', len(set(word_list)) - len(set(cleaned_word_list)))
#skaičiuojamas atskirų žodžių skaičius
wordfreq = cl.Counter(word_list)
wordfreq = [(wordfreq[p], p) for p in cleaned_word_list]
#sudaromas atskirų žodžių ir jų dažnių sąrašas
sorted_dict = sorted(wordfreq, reverse = True)
print(len(sorted_dict))
sorted_dict
pd.DataFrame(sorted_dict, columns=['count', 'word']).to_excel('Final_Word_Count_list_Germany.xlsx', index = False)
#atksirų žodžių ir jų dažnių sąrašas išvedamas į „Excel“ failą ir peržiūrimas rankiniu būdu
exp_words = pd.read_excel('experience.xlsx') #importuojamas klientų patirtį atspindinčių žodžių sąrašas
exp_words.drop(columns = 'count', inplace = True)
wc_words = " ".join(word for word in exp_words.word)
#žodžių debesies generavimas
wordcloud = WordCloud(background_color="white", colormap="Dark2").generate(wc_words)
#žodžių debesies vaizdavimas
plt.figure(figsize=(10, 8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off");
word_df = pd.DataFrame(sorted_dict, columns = ('count', "word"))
word_df.drop(columns = 'count', inplace = True)
word_df = pd.merge(exp_words, word_df, on=['word'], how='left')
word_df = word_df.set_index('word').T
word_df
```
###Vektorinės erdvės modelio taikymas
```
#sudaroma atsiliepimų-požymių matrica su tuščiomis reikšmėmis, kurios vėliau yra užpildomos
%%time
frames = [word_df, data2['comments'].apply(lambda x: x.lower())
df3 = pd.concat(frames)
df3 = df3.rename(columns = {0: "comments"})
print(df3.shape)
show(df3, 3)
#skaičiuojamas tf įvertis
%%time
word_tf = df3.copy()
for column in word_tf.columns:
    if column != 'comments':
        word_tf[column] = word_tf['comments'].str.contains(str(column)).astype(int)
word_tf['Col_Total'] = word_tf.sum(axis = 1)
print(word_tf['Col_Total'][word_tf['Col_Total'] == 0].shape, word_tf['Col_Total'][word_tf['Col_Total'] > 0].shape)
word_tf = word_tf[word_tf['Col_Total'] > 0] #šalinami atsiliepimai, kuriuose nėra nei vieno žodžio iš sudaryto, klientų patirtį atspindinčių žodžių sąrašo
word_tf.drop(columns = "Col_Total", inplace = True)
word_tf['word_count'] = word_tf['comments'].apply(lambda x: len(str(x).split()))
for column in word_tf.columns:
    if column not in ('word_count', 'comments'):
```

```
        word_tf[column] = word_tf[column] / word_tf['word_count']
word_tf.drop(columns = 'word_count', inplace = True)
show(word_tf, 3)
#skaičiuojamas idf įvertis
%%time
word_idf = df3.copy()
for column in word_idf.columns:
    if column != 'comments':
        word_idf[column] = word_idf['comments'].str.contains(str(column)).astype(int)
word_idf.drop(columns = 'comments', inplace = True)
word_idf['Col_Total'] = word_idf.sum(axis = 1)
print(word_idf['Col_Total'][word_idf['Col_Total']==0].shape, word_idf['Col_Total'][word_idf['Col_Total']!=0].shape)
word_idf = word_idf[word_idf['Col_Total']>0]
word_idf.drop(columns = "Col_Total", inplace = True)
show(word_idf, 3)
%%time
word_idf = word_idf.sum()
word_idf = pd.DataFrame(word_idf)
word_idf = word_idf.reset_index()
word_idf['idf'] = word_idf[0].apply(lambda x: math.log(len(word_tf['comments']) / x, 10) if x != 0 else 0)
print(word_idf.shape)
word_idf.head(5)
#skaičiuojamas tf-idf įvertis
%%time
tf_idf = word_tf.drop(columns = 'comments').copy()
for column in tf_idf.columns:
    tf_idf[column] = tf_idf[column].values * word_idf['idf'][word_idf['index'] == column].values
print(tf_idf.shape)
show(tf_idf, 5)
```

4 priedas. Faktorinės analizės modelis

```
###Faktorinės analizės modelis
df4 = tf_idf.copy()
print(df4.shape)
#bendro KMO skaičiavimas
kmo_all, kmo_model = calculate_kmo(df4)
round(kmo_model,2)
#KMO skaičiavimas atskiriems požymiams ir jo filtravimas
kmo_results = pd.DataFrame(data = kmo_all, index = df4.columns)
kmo_results = kmo_results[kmo_results[0] > 0.57]
kmo_col = kmo_results.reset_index().set_index('index').T.columns
print(kmo_results.shape)
kmo_results.sort_values(by = 0, ascending = False).head(10)
#bendro KMO perskaičiavimas
df4 = df4[kmo_col]
kmo_all, kmo_model = calculate_kmo(df4)
round(kmo_model, 2)
#tikrinama Bartlett’o sferiškumo kriterijaus reikšmė
chi_square_value, p_value = calculate_bartlett_sphericity(df4)
round(chi_square_value, 1), round(p_value,3)
#taikomas faktorinės analizės metodas
fa = FactorAnalyzer()
fa.fit(df4)
#braižomas tikrinių reikšmių ir faktorių skaičiaus grafikas
plt.figure(figsize = (12, 6))
plt.scatter(range(1, df4.shape[1] + 1), ev)
plt.plot(range(1, df4.shape[1] + 1), ev)
plt.title('Scree Plot')
plt.xlabel('Factors')
plt.ylabel('Eigenvalue')
plt.grid();
#faktorinės analizės vykdymas – apibrėžiamas faktorių išskyrimo metodas, jų skaičius, sukimo metodas
```

```
fa = FactorAnalyzer(method = 'principal', n_factors = 8, rotation = 'varimax')
fa.fit(df4)
factor_data = fa.loadings_
factor_data.shape, fa
#išvedami gauti faktoriai ir juos atitinkantys požymiai su svoriais
factor_analysis2 = pd.DataFrame(data = factor_data[0:,:], index = df4.columns)
factor_analysis = factor_analysis2.copy()
for column in factor_analysis.columns:
    factor_analysis[column] = factor_analysis[column].apply(lambda x: round(x, 2) if abs(x) > 0.35 else None)
factor_analysis.dropna(how = 'all', inplace = True)
factor_analysis.fillna("", inplace = True)
factor_analysis.sort_values(by = list(factor_analysis.columns), inplace = True)
with pd.option_context('display.max_rows', None, 'display.max_columns', None):
    display(factor_analysis)
len(factor_analysis) #skaičiuojamas faktorius sudarančių požymių kiekis
#skaičiuojama, kokią dispersijos dalį sudaro gauti faktoriai
fac_var = fa.get_factor_variance()
cum_var = pd.DataFrame(data = fac_var[2], index = factor_analysis.columns, columns = ['cum_var'])
fa_var = pd.DataFrame(data = fac_var[1], index = factor_analysis.columns, columns = ['var'])
frames = [cum_var, fa_var]
varian = pd.concat(frames, axis = 1)
varian.T
#skaičiuojami faktorių įverčiai
fa_scores = fa.transform(df4)
fa_scores_pd = pd.DataFrame(data = fa_scores[0:,:], index = word_tf.index, columns = ["FA1", "FA2", "FA3", "FA4",
"FA5", "FA6", "FA7", "FA8"])
fa_scores_pd.round(3).head(5)
```

###Duomenų paruošimas regresinei analizei

```
final_df = data2[data2.comments.isin(word_tf.comments)]
final_df = final_df.merge(fa_scores_pd, left_index = True, right_index = True) #sujungiamos duomenų matricos
print(final_df.shape)
show(final_df, 3)
final_df.dtypes
final_df['org_rating'].value_counts() #bendro būsto vertinimo dinamika
#bendro būsto vertinimo dinamikos vaizdavimas
plt.figure(figsize = (15, 5))
final_df['org_rating'].value_counts(normalize = True).plot(kind = 'bar')
plt.xlabel("org_rating")
plt.ylabel("count");
final_df = final_df.select_dtypes(exclude=['object']) #iš tolimesnės analizės šalinami nereikalingi požymiai
final_df.head(5)
#požymių standartizavimas – skalių suvienodinimas
scaled_features = final_df.copy()
col_names = ['price', 'FA1' , 'FA2' , 'FA3' , 'FA4' , 'FA5' , 'FA6' , 'FA7', 'FA8', 'distance_from_center', 'org_rating',
'comment_length', 'compound']
features = scaled_features[col_names]
scaler = preprocessing.StandardScaler().fit(features.values)
features = scaler.transform(features.values)
scaled_features[col_names] = features
scaled_features.head(5)
#standartizuotų požymių ir bendro būsto vertinimo priklausomybės grafinis vaizdavimas
for i in scaled_features.columns:
    if i in('org_rating', 'market', 'host_is_superhost','room_type', 'bedrooms', 'guests_included', 'minimum_nights',
'bathrooms'):
        continue
    else:
        print(i)
        plt.scatter(x = i, y = "org_rating", data = scaled_features)
        plt.show()
        continue
#standartizuotų požymių filtravimas dėl aiškiai išsiskiriančių stebėjimų
scaled_features_2 = scaled_features[(scaled_features['org_rating'] >= -5)
                                    & (scaled_features['comment_length'] < 13)
```

```
                & (scaled_features['distance_from_center'] < 6)
                & (scaled_features['price'] < 4.6)
                & (scaled_features['FA1'] > -2) & (scaled_features['FA1'] < 7)
                & (scaled_features['FA2'] > -2.5) & (scaled_features['FA2'] < 10)
                & (scaled_features['FA3'] > -5) & (scaled_features['FA3'] < 15)
                & (scaled_features['FA4'] > -5) & (scaled_features['FA4'] < 15)
                & (scaled_features['FA5'] > -8) & (scaled_features['FA5'] < 13)
                & (scaled_features['FA6'] > -8) & (scaled_features['FA6'] < 15)
                & (scaled_features['FA7'] > -10) & (scaled_features['FA7'] < 20)
                & (scaled_features['FA8'] > -10) & (scaled_features['FA8'] < 14)
                ].copy()
#po filtravimo pakartojamas standartizuotų požymių ir bendro būsto vertinimo priklausomybės grafinis vaizdavimas
for i in scaled_features_2.columns:
   if i in('org_rating', 'market', 'host_is_superhost','room_type', 'bedrooms', 'guests_included', 'minimum_nights', 'bathrooms'):
      continue
   else:
      print(i)
      plt.scatter(x = i, y = "org_rating", data = scaled_features_2)
      plt.show()
      continue
scaled_features_2.to_csv('germany.csv', index=False) #galutiniai duomenys: pradiniai ir naujai sukurti požymiai bei faktoriai išvedami į csv failą, toliau naudojamą regresinei analizei
```

5 priedas. Regresinės analizės modelis

```
/*duomenų failo importavimas*/
proc import datafile = '/home/nazarovaitek0/Regresija/germany.csv'
 out = duom
 dbms = CSV
 ;
run;
/*koreliacijos tikrinimas*/
proc corr data=duom Pearson Fisher plots=MATRIX (Histogram);
var org_rating guests_included market room_type bathrooms bedrooms price minimum_nights host_is_superhost distance_from_center comment_length compound FA1 FA2 FA3 FA4 FA5 FA6 FA7 FA8;
run;
ods graphics on;
/*modelio sudarymas su visais pradiniais kintamaisiais*/
proc reg data=duom;
model org_rating = FA1 FA2 FA3 FA4 FA5 FA6 FA7 FA8 guests_included market room_type bathrooms bedrooms price minimum_nights host_is_superhost distance_from_center comment_length compound/stb clb vif;
run;
/*modelio sudarymas su atrinktais, statistiškai reikšmingais kintamaisiais*/
proc reg data=duom;
model org_rating = FA1 FA2 FA3 FA4 FA5 FA6 FA8 guests_included market room_type bathrooms bedrooms price minimum_nights host_is_superhost distance_from_center comment_length compound/stb clb vif;
run;
/*koreliacijos tikrinimas po kintamųjų šalinimo*/
proc corr data = duom Pearson Fisher plots = matrix(Histogram);
var org_rating FA1 FA2 FA3 FA4 FA5 FA6 FA8 guests_included market room_type bathrooms bedrooms price minimum_nights host_is_superhost distance_from_center comment_length compound;
run;
ods graphics on;
/*tiesinės regresijos modelio prielaidų tikrinimas*/
proc reg data=duom(obs=10000) plots(maxpoints=10000)=(diagnostics(stats=none) RStudentByLeverage(label) residuals(smooth) CooksD(label) DFFITS(label) DFBETAS);
model org_rating = FA1 FA2 FA3 FA4 FA5 FA6 FA8 guests_included market room_type bathrooms bedrooms price minimum_nights host_is_superhost distance_from_center comment_length compound / clb influence spec hccmethod=1 white dwprob;
output out=resdat r=resid;
run;
/*liekanų normalumo tikrinimas*/
```

```
proc univariate data = resdat normal plot;
var resid;
run;
/*atsitiktinės imties iš turimų duomenų parinkimas (kiekvienąkart paleidus procedūra, atsitiktinai parinkta duomenų imtis keičiasi)*/
proc surveyselect data=duom
out=germany_reduced
method=srs
sampsize=10000;
run;
/*atsitiktinės imties pagrindinių charakteristikų išvedimas*/
proc means data = germany_reduced;
run;
/*atsitiktinai parinkta duomenų imtis įrašoma į papildomą failą tam, kad išvengti duomenų pasikeitimo*/
proc export data = germany_reduced
outfile = "/home/nazarovaitek0/Regresija/germany_reduced.csv"
dbms = dlm replace;
delimiter = ',';
run;
/*atspariosios regresijos modelio sudarymas ir išskirčių išvedimas*/
proc robustreg data=germany_reduced method=lts plots=all;
  model org_rating = FA1 FA2 FA3 FA4 FA5 FA6 FA7 FA8 guests_included market room_type bathrooms bedrooms price minimum_nights host_is_superhost distance_from_center comment_length compound/ diagnostics leverage;
  ods select MCDCenter MCDCov Diagnostics;
  ods output diagnostics=Diagnostics;
run;
/*atspariosios regresijos rezultatų failo sujungimas su pradiniu duomenų failu*/
proc sql;
create table residual as
select Obs, RResidual, Leverage, Outlier, PRobustDist
from work.Diagnostics;
run;
data new;
set germany_reduced;
Obs=_n_;
run;
data final;
merge new residual;
by Obs;
run;
data final1;
set final;
if RResidual = . then RResidual=0;
if Leverage = . then Leverage=0;
if Outlier = . then Outlier=0;
if PRobustDist = . then PRobustDist=0;
run;
/*duomenų matrica įrašoma į papildomą failą, jei prireiktų jį naudoti darkart, bet nenorima kartoti pradinių žingsnių*/
proc export data = final1
outfile = "/home/nazarovaitek0/Regresija/germany_final.csv"
dbms = dlm replace;
delimiter = ',';
run;
/*šalinamos grubiausios išskirtys*/
data germany_final;
set final1;
if PRobustDist > 300 or Outlier = 1 then delete;
run;
/*sudaromas tiesinės regresijos modelis su visais pradiniais kintamaisiais po grubiausių išskirčių šalinimo*/
proc reg data=germany_final;
model org_rating = FA1 FA2 FA3 FA4 FA5 FA6 FA7 FA8 guests_included market room_type bathrooms bedrooms price minimum_nights host_is_superhost distance_from_center comment_length compound/stb clb vif;
run;
```

```
/*sudaromas tiesinės regresijos modelis su atrinktais kintamaisiais po grubiausių išskirčių šalinimo*/
proc reg data=germany_final;
model org_rating = FA2 FA3 FA5 guests_included market room_type bedrooms price minimum_nights host_is_superhost distance_from_center compound/stb clb vif;
run;
/*tikrinamos sudaryto modelio prielaidos*/
proc reg data=germany_final plots(maxpoints=10000)=(diagnostics(stats=none) RStudentByLeverage(label) residuals(smooth) CooksD(label) DFFITS(label) DFBETAS ObservedByPredicted(label));
model org_rating = FA2 FA3 FA5 guests_included market room_type bedrooms price minimum_nights host_is_superhost distance_from_center compound /clb partial vif spec hccmethod=1 white dwprob;
output out=resdat r=resid;
run;
/*liekanų normalumo tikrinimas*/
proc univariate data = resdat normal plot;
var resid;
run;
/*liekanų histogramos su normaliojo skirstinio kreive braižymas*/
ods select Histogram;
proc univariate data = resdat normal plot;
histogram resid / normal(percents = 20 40 60 80 midpercents);
  inset n normal(ksdpval) / pos = ne format = 6.3;
var resid;
run;
/*koreliacijos tikriniamas po kintamųjų atrinkimo ir grubiausių išskirčių šalinimo*/
proc corr data = final2 Pearson Fisher plots = matrix(Histogram);
var org_rating FA2 FA3 FA5 guests_included market room_type bedrooms price minimum_nights host_is_superhost distance_from_center compound;
run;
ods graphics on;
```

6 priedas. Klientų patirtį atspindinčių žodžių sąrašas

Nr.	Žodis	Nr.	Žodis	Nr.	Žodis
1.	great	104.	return	207.	respectful
2.	place	105.	hospitable	208.	secure
3.	stay	106.	towels	209.	reachable
4.	location	107.	accommodation	210.	various
5.	nice	108.	appreciated	211.	satisfied
6.	clean	109.	accessible	212.	renovated
7.	good	110.	machine	213.	reasonable
8.	room	111.	parking	214.	reliable
9.	recommend	112.	noise	215.	affordable
10.	close	113.	connected	216.	informative
11.	comfortable	114.	advice	217.	quirky
12.	perfect	115.	sweet	218.	answering
13.	helpful	116.	simple	219.	authentic
14.	city	117.	garden	220.	condition
15.	friendly	118.	furnished	221.	uncomplicated
16.	lovely	119.	thoughtful	222.	broken
17.	restaurants	120.	charming	223.	professional
18.	quiet	121.	smooth	224.	artistic
19.	station	122.	offered	225.	surprised
20.	spacious	123.	extra	226.	essentials
21.	walk	124.	useful	227.	negative
22.	super	125.	suggestions	228.	dishwasher

23.	home	126.	facilities	229.	microwave
24.	public	127.	communicative	230.	shampoo
25.	beautiful	128.	peaceful	231.	reservation
26.	amazing	129.	issue	232.	exceptional
27.	neighborhood	130.	free	233.	delightful
28.	wonderful	131.	pleasure	234.	disappointed
29.	bed	132.	real	235.	grateful
30.	communication	133.	options	236.	tiny
31.	kitchen	134.	problems	237.	patient
32.	transport	135.	basic	238.	absolute
33.	kind	136.	fridge	239.	expensive
34.	near	137.	sofa	240.	soap
35.	big	138.	worth	241.	immaculate
36.	enjoyed	139.	atmosphere	242.	silent
37.	excellent	140.	attentive	243.	knowledgeable
38.	bathroom	141.	lively	244.	necessities
39.	experience	142.	incredible	245.	activities
40.	welcoming	143.	courtyard	246.	standard
41.	center	144.	internet	247.	oven
42.	best	145.	fresh	248.	additional
43.	cozy	146.	bike	249.	famous
44.	metro	147.	brilliant	250.	careful
45.	bars	148.	advertised	251.	precise
46.	fantastic	149.	allowed	252.	easygoing
47.	convenient	150.	organized	253.	elegant
48.	little	151.	gorgeous	254.	fancy
49.	shops	152.	bad	255.	memorable
50.	distance	153.	accurate	256.	backyard
51.	small	154.	special	257.	creative
52.	warm	155.	relaxing	258.	luxurious
53.	access	156.	terrace	259.	blankets
54.	stylish	157.	fabulous	260.	adequate
55.	quick	158.	hard	261.	original
56.	cool	159.	generous	262.	sensitive
57.	equipped	160.	personal	263.	advantage
58.	large	161.	polite	264.	refrigerator
59.	cafes	162.	sightseeing	265.	gentle
60.	provided	163.	greeted	266.	iron
61.	family	164.	handy	267.	playground
62.	accommodating	165.	design	268.	organic
63.	arrival	166.	delicious	269.	courteous
64.	balcony	167.	service	270.	supportive
65.	food	168.	airy	271.	zoo
66.	shower	169.	museum	272.	thankful
67.	responsive	170.	quality	273.	traditional
68.	amenities	171.	prompt	274.	sufficient
69.	pleasant	172.	clubs	275.	alternative
70.	easily	173.	interior	276.	galleries

71.	awesome	174.	dryer	277.	cheerful
72.	airport	175.	cheap	278.	honest
73.	far	176.	difficult	279.	impeccable
74.	supermarket	177.	functional	280.	strange
75.	light	178.	washer	281.	assistance
76.	available	179.	outstanding	282.	minimalist
77.	bedroom	180.	detailed	283.	regular
78.	described	181.	unique	284.	impressive
79.	decorated	182.	dirty	285.	unbeatable
80.	bright	183.	gracious	286.	unforgettable
81.	safe	184.	trendy	287.	approachable
82.	flexible	185.	caring	288.	international
83.	tidy	186.	positive	289.	inconvenience
84.	value	187.	natural	290.	independent
85.	hospitality	188.	nightlife	291.	terrible
86.	breakfast	189.	bonus	292.	pool
87.	pretty	190.	pillows	293.	punctual
88.	park	191.	heating	294.	specific
89.	fast	192.	sheets	295.	responsible
90.	expected	193.	efficient	296.	effective
91.	view	194.	smell	297.	unusual
92.	price	195.	adorable	298.	garage
93.	huge	196.	considerate	299.	bohemian
94.	sleep	197.	uncomfortable	300.	refund
95.	stops	198.	desk	301.	enthusiastic
96.	cute	199.	terrific	302.	social
97.	private	200.	checkout	303.	weird
98.	explore	201.	suitable	304.	classic
99.	ideal	202.	equipment	305.	suite
100.	attractions	203.	maintained	306.	active
101.	water	204.	replied	307.	powerful
102.	new	205.	stunning	308.	extraordinary
103.	welcomed	206.	practical		

7 priedas. Faktorių svoriai kiekvienam stebiniui (fragmentas Berlyno ir Miuncheno miestų duomenims)

	FA1	FA2	FA3	FA4	FA5	FA6	FA7	FA8
371	-0.336	-0.393	-0.095	-0.066	-0.125	-0.157	0.007	0.044
373	-0.336	-0.393	-0.095	-0.066	-0.125	-0.157	0.007	0.044
375	0.857	-0.214	-0.073	0.102	-0.080	-0.065	-0.165	0.064
376	-0.336	-0.393	-0.095	-0.066	-0.125	-0.157	0.007	0.044
377	2.049	-0.007	-0.194	-0.239	-0.049	-0.030	-0.084	-0.031

8 priedas. Modelio išskirčių ir įtakos taškų sąrašas (fragmentas Berlyno ir Miuncheno miestų duomenims)

The ROBUSTREG Procedure

Diagnostics

Obs	Projected Distance			Leverage	Standardized Robust Residual	Outlier
	Mahalanobis	Robust	Off-Plane			
1	3.1643	22.8646	1.6316	*	0.0340	
3	4.0596	45.9392	1.1284	*	-0.0481	
4	5.9561	51.6842	2.4522	*	0.2453	
5	4.2457	10.5211	0.0000	*	0.0720	
7	3.6256	13.0979	3.8675	*	-0.5564	
10	3.2822	12.0848	9.6552	*	-0.6106	
11	4.9292	7.4559	0.0000	*	-0.8081	
12	4.6666	24.7737	3.8432	*	-0.7824	
13	5.4390	48.3219	2.0061	*	-0.6693	
15	4.6267	31.3408	0.0000	*	-0.7391	
16	7.1042	22.5450	4.4561	*	-2.1885	
19	7.1706	80.9961	0.6352	*	-0.4020	
21	6.0321	58.1073	2.7658	*	-0.3313	
22	4.6589	37.5902	5.5155	*	-0.6992	
24	4.3880	12.6060	0.3785	*	-0.5815	
28	4.1596	37.9233	6.0896	*	-1.5992	
29	5.7333	14.6758	0.4319	*	-1.5518	
31	3.8590	10.2047	0.7805	*	-2.0636	
32	3.3727	24.4670	1.3885	*	0.3525	
37	2.9474	27.1380	0.7446	*	-0.4446	
38	3.0157	6.3456	0.0000	*	-0.3787	
39	2.9912	6.3272	0.0000	*	-0.3593	
40	8.4622	89.2831	5.7698	*	0.0441	
41	5.3817	8.5843	0.0000	*	-0.0427	
43	4.0712	43.4605	1.8054	*	-0.2357	
44	5.0370	51.3959	1.8092	*	-0.4710	
48	3.1575	19.0757	0.0000	*	-0.1698	
51	6.1737	74.3766	1.6702	*	-0.4027	
52	2.4936	11.2959	0.9324	*	-0.4220	
54	2.2116	2.3239	0.0000		-3.0003	*
55	2.2864	2.4843	0.0000		-3.0169	*

9 priedas. Faktorių svoriai kiekvienam stebiniui (fragmentas Madrido ir Barselonos miestų duomenims)

	FA1	FA2	FA3	FA4	FA5	FA6	FA7
0	-0.312	-0.398	-0.394	-0.136	-0.087	-0.091	0.030
1	-0.312	-0.398	-0.394	-0.136	-0.087	-0.091	0.030
4	-0.372	1.186	-0.222	1.487	0.666	-0.539	0.665
5	-0.351	-0.394	0.455	-0.100	-0.044	-0.023	0.015
6	0.140	-0.034	-0.895	5.358	-0.530	-1.842	1.083

10 priedas. Modelio išskirčių ir įtakos taškų sąrašas (fragmentas Madrido ir Barselonos miestų duomenims)

The ROBUSTREG Procedure

Diagnostics					
Obs	Mahalanobis Distance	Robust MCD Distance	Leverage	Standardized Robust Residual	Outlier
1	4.5679	7.6733	*	-0.3943	
3	3.8126	24.6018	*	-1.0281	
4	5.4581	584.8812	*	-0.3264	
5	5.9780	15.2868	*	0.0568	
6	3.3067	11.7738	*	-0.2318	
9	5.4358	1151.532	*	-0.7874	
12	8.9504	2338.271	*	-1.1446	
13	3.8949	61.9576	*	-1.3838	
16	2.8012	3.0668		-3.6970	*
17	3.7560	158.4740	*	-3.6250	*
18	2.9462	130.8057	*	-3.4089	*
19	2.5064	2.5350		-3.8211	*
20	3.1750	7.2972	*	-3.8030	*