Vilniaus Universitetas

Tiesioginio duomenų vizualizavimo metodai

Laboratorinis darbas

Darbą atliko:

Antanas Užpelkis, Roland Gulbinovič, Matas Kamarauskas, Matas Amšiejus

Vilnius, 2021 03 11

TURINYS

DUOMENYS

Laboratorianiam darbui atlikti buvo paimti duomenys apie nekilnojamo turto įsigyjimą Sinbėjaus rajone, esančiame šiaurės Taivane 2012 metų pabaigoje, 2013 metais. Duomenyse yra pateikta informacija apie 414 nekilnojamo turto (namai\butai) įsigyjimą. Duomenų atributai yra pirkimo data, pastato amžius (metais), atstumas iki artimiausio metro (metrais), parduotuvių kiekis gyvenamoje vietoje (vienetais), pastato koordinatės: platuma ir ilguma, pastato kaina. Duomenyse kaina yra nurodoma 10,000 naujųjų Taivano dolerių (apytiksliai 296 eurai) vienam Ping (1 Ping = 3,3 kv. metrai). Duomenys yra paimti iš UCI Machine Learning Repository svetainės (tiksli nuoroda į duomenis yra aprašo pabaigoje).

PIRMA UŽDUOTIS

Tikslai ir metodai

Tyrimo tikslas buvo nustatyti namų kainų tendecijas pagal turimus duomenis pasitelkiant duomenų vizualizavimo metodus. Pasigilinę į duomenis nustatėme, ką konkrečiai norėtume ištirti ir kokiais vizualizavimo metodais tai pavaizduoti.

Tikslai ir vizualizavimo metodai:

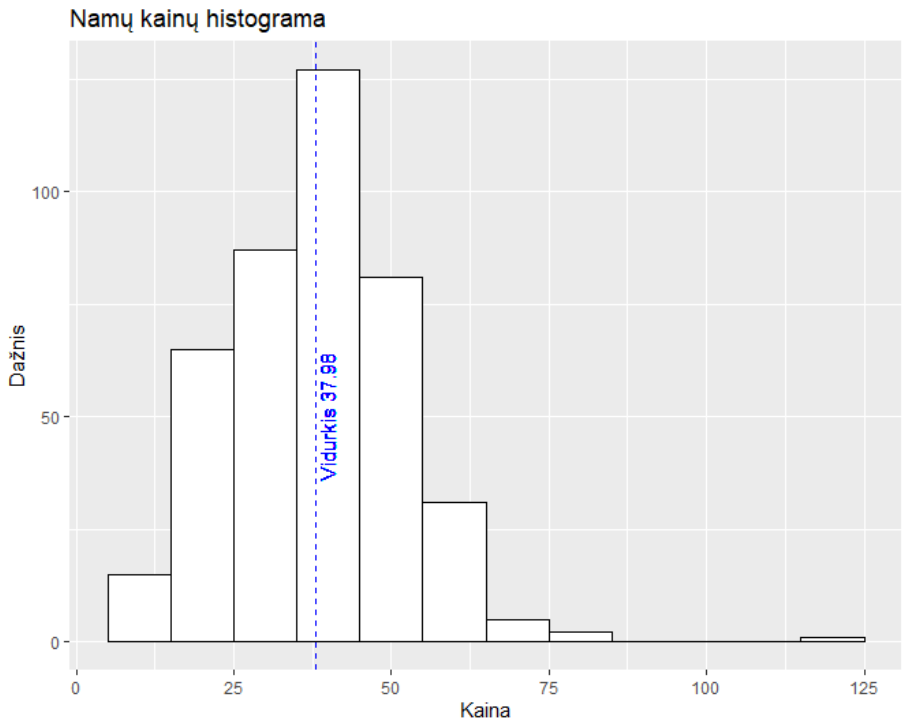
1. Paskaičiuoti kainų pasiskirstymą ir rasti vidutinę būsto kainą. Metodas – histograma.
2. Rasti kaip priklauso būstų pasiskirstymas nuo parduotuvių skaičiaus ir atstumo iki metro skirtingų kainų kategorijose (pigūs, vidutiniški, brangūs). Metodas – sklaidos diagramos sudėtos į groteles.
3. Rasti kaip būstų pirkimai priklauso nuo būsto amžiaus ir kainos. Metodas – šilumos žemėlapis (angl. heatmap).
4. Atrasti kaip geografiškai pasiskirstę įsigyti namai. Metodas – sklaidos diagrama.

Tikslų realizavimas

1. Kainų pasiskirstymas

Norėjome pamatyti, kaip yra pasiskirsčiusios namų kainos, koks yra vidurkis. Taip pat buvo naudinga pastebėti ar nėra išskirčių.

Duomenų nereikėjo apdoroti, pasinaudojus R paketu „ggplot2“ nubrėžėme histogramą. Taip pat pridėjome vertikalią punktyrinę liniją, kuri rodo kainų vidurkį.



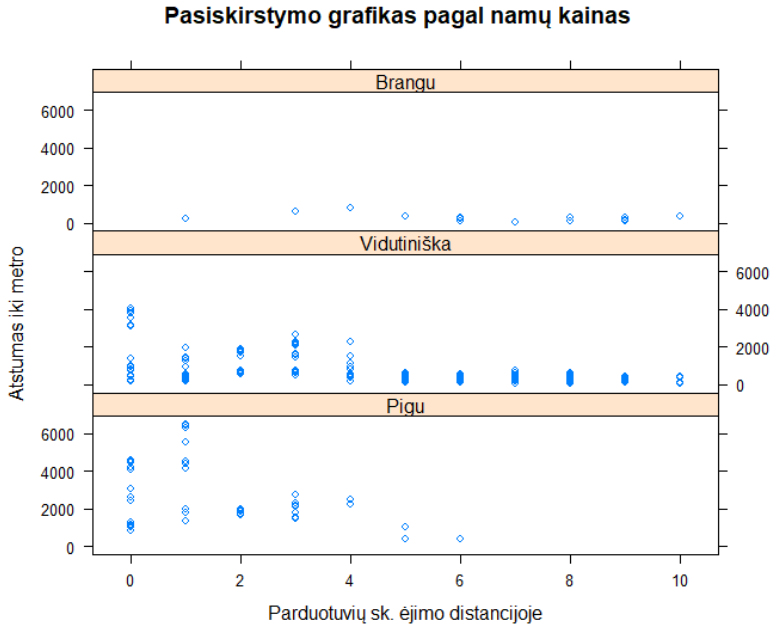
pav. 1 Namų kainų pasiskirstymas. Histograma

Išvados: matome kad vidutiniškai būstas kainuoja 37,98 (pastaba, čia ir toliau kainos vienetai bus 10tūkst.NTD/Ping (NTD – new Taiwan dollar)), didžioji dauguma namų kainuoja nuo 20 iki 60, taip pat matome, kad yra išskirčių – namų, kurie kainuoja itin daug, tačiau patikrinus duomenis nustatėme, kad išskirtis tik viena. *R kodas: 1.*

2. Priklausomybė nuo parduotuvių skaičiaus ir atstumo iki metro

Pagal iškeltą hipotezę galvojome, kad skirtingų kainų namai turi tam tikrus panašumus tarpusavyje. Tarkime, kad tie namai, kurie bus netoli metro stočių ar netoli parduotuvių bus brangūs, nes tai yra papildomi patogumai, o namai, kurie bus toli nuo metro ir nebus arti parduotuvių bus pigesni dėl nepatogesnių sąlygų.

Duomenys buvo suskirstyti pagal kainą į tris lygius: pigūs (kaina iki 25 10tūkst.NTD/Ping), vidutiniški (nuo 25 iki 60), brangūs (nuo 60). Naudojant „lattice“ paketą buvo nubrėžtos trys sklaidos diagramos (kiekvienam kainos lygiui po vieną), kur ordinatėje buvo atstumas iki metro, o abscisėje - parduotuvių skaičius netoliese.



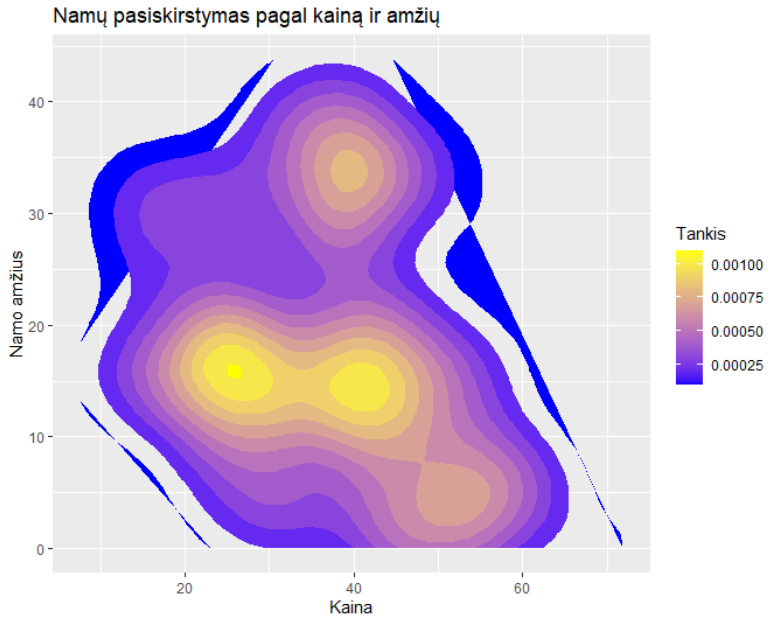
pav. 2 Atstumo iki metro ir parduotuvių skaičiaus pasiskirstymas pagal namo kainą. Sklaidos diagrama

Išvados: didžioji dalis brangių namų yra netoli metro stotelės, tačiau parduotuvių skaičius nėra pastovus. Dauguma pigesnių namų yra vietose, kur nėra daug parduotuvių ir atstumas iki artimiausios metro stoties gana didelis (galima tikėtis, kad namai yra nuošaliau nuo centro). Dauguma namų, kurie yra vidutiniškos kainos, pasiskirstę nevienodai pagal parduotuves, tačiau yra įsikūrę arčiau metro stočių. Kai parduotuvių skaičius siekia daugiau nei keturias, matome, kad beveik visi namai yra visai šalia metro. *R kodas: 2.*

3. Būstų pasiskirstymas priklausomai nuo amžiaus ir kainos

Norėjome pamatyti ar yra kokia nors sąsaja tarp pastato amžiaus ir jo kainos. Tarkime, seni namai turėtų kainuoti pigiau, nes galbūt yra prastesnės kokybės, o naujai pastayti yra brangesni, nes jiems statant buvo naudojamos pažangesnės technologijos (efektyvesnis šildymas, dizainas) ir dėl to jie yra geresnės kokybės.

Pasinaudojus „ggplot2“ paketu buvo nubrėžtas šilumos žemėlapis, kuriame ordinatėje yra būsto amžius, o abscisėje – kaina. Skirtingos spalvos vaizduoja pastatų kiekį (tankumą), kuo spalva mėlynesnė, tuo mažiau yra pastatų, kuo geltonesnė – tuo daugiau.



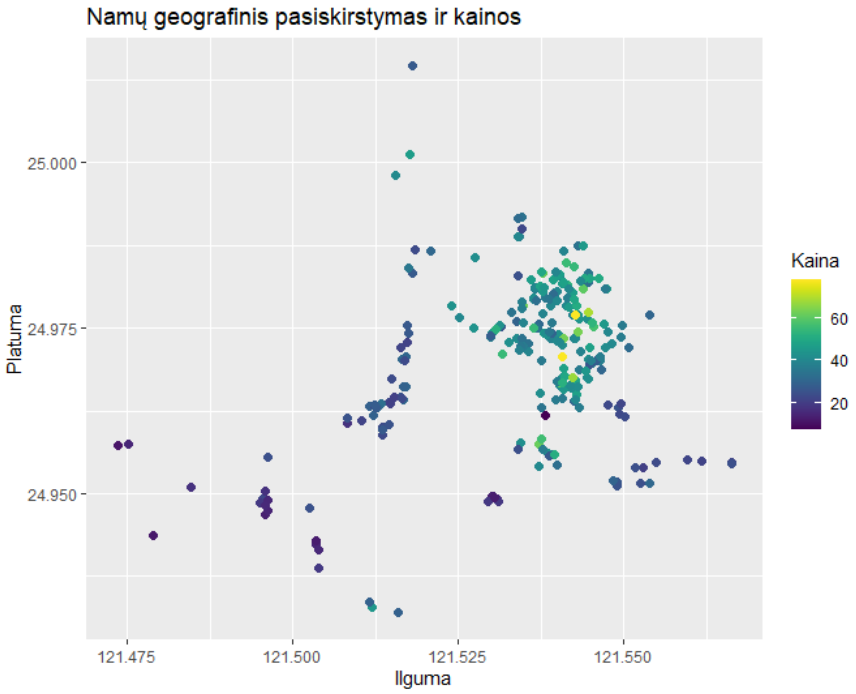
pav. 3 Būstų pirkimo dažnis pagal kainą ir amžių. Šilumos diagrama

Išvados: galima pastebėti, kad daugiausia nupirkta namų, kurių amžius yra nuo 10 iki 20 metų, o jų kainos yra apytiksliai nuo 20 iki 48 10tūkst.NTD\Ping. Šis požymis leidžia suprasti, kad dauguma nori sąlyginai naujų (arti šio tūkstantmečio) namų už vidutinę kainą. Taip pat matome, kad yra nemažai namų, kuriems yra apie 35 metai ir jų kaina yra apie 40 10tūkst.NTD\Ping. Galima bandyti nuspėti, kad tai seni, tačiau kadaisę buvę prabangūs namai. Dar viena hipotezė gali būti, kad tai nuo seno patogioje geografinėje lokacijoje statyti namai, kurių kaina su laiku dėl vietos tik brangsta. *R kodas: 3.*

4. Geografinis namų pasiskirstymas

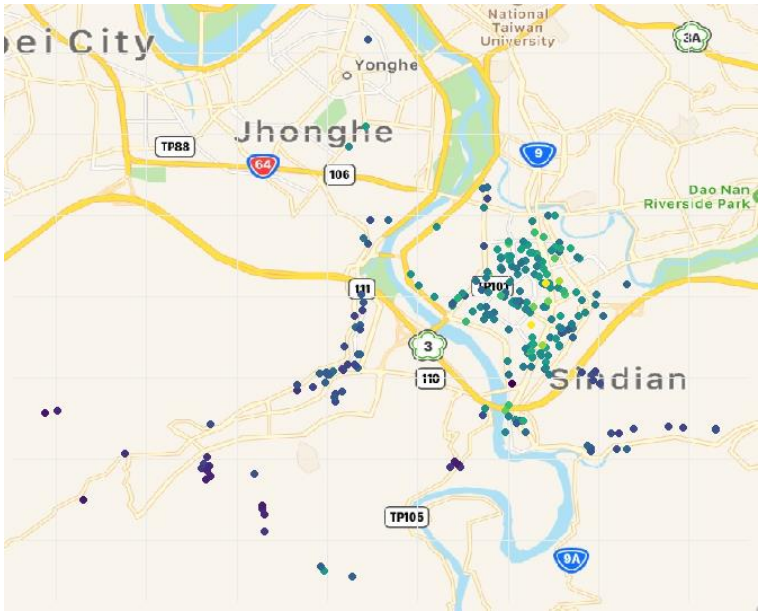
Kadangi turėjome namų koordinates, galėjome susidaryti žemėlapį, kaip jie yra išsidėstę. Panaudojus namų kainų duomenis, galėjome nustatyti, kur yra brangesni/pigesni namai ir iš to susidaryti įvaizdį apie gyvenamą rajoną.

Norėdami pamatyti tikslesnį kainų pasiskirstymą, mes atrinkome tik tuos namus, kurių kainos mažesnės nei 80 (buvo tik vienas namas, kuris kainavo daugiau, jis kainavo 117). Atrinktus duomenis vaizdavome „ggplot2“ paketo pagalba, sklaidos diagrama, kur y ašyje – platuma, x ašyje – ilguma. Skirtingos taškų kainos indikuoja namų kainas – kuo brangesnis, tuo taškas geltonesnis, kuo pigesnis, tuo labiau mėlynas.



pav. 4 Namų geografinis pasiskirstymas ir jų kainos. Sklaidos diagrama (be fono)

Norėdami pamatyti tikslų šio rajono vaizdą pasinaudojome „Google žemėlapiais“. Didžiausios ir mažiausios ilgumos ir platumos reikšmės, buvo mūsų žemėlapio kontūrai. „Išsikirpę“ rastą žemėlapį uždėjome ant mūsų nubrėžtos sklaidos diagramos kaip foną. Taip gavome žemėlapį, kuriame buvo pavaizduoti nupirkti namai ir kokia jų kaina. (Pastaba: kadangi žemėlapis buvo nustatytas rankiniu būdu, duomenys nėra visiškai tikslūs. Tačiau paklaida yra pakankamai gera, kad būtų galima padaryti išvadas). *R kodas: 4.*

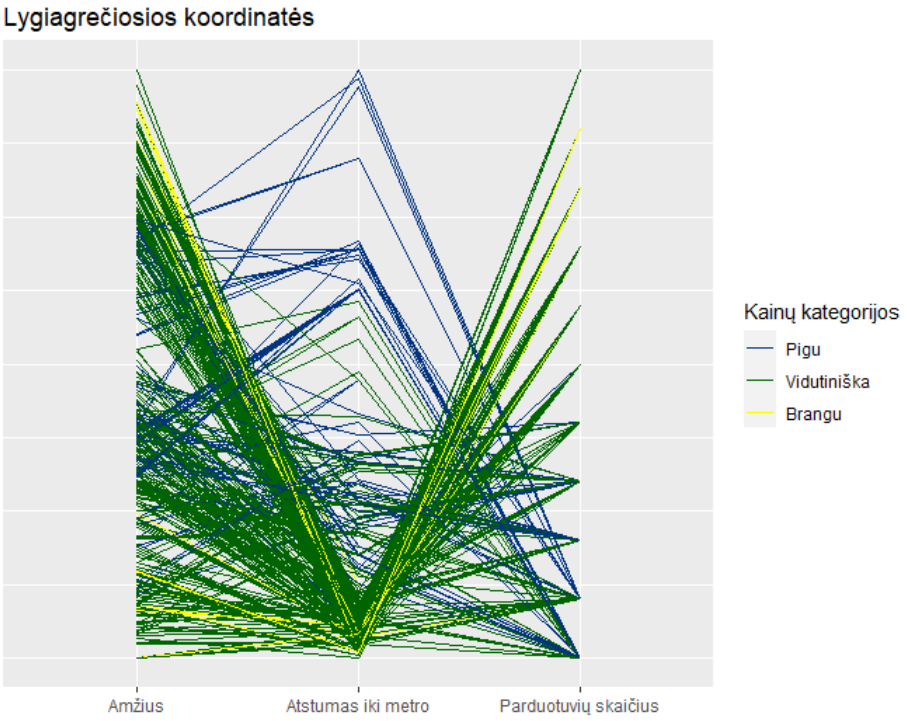


pav. 5 Namų geografinis pasiskirstymas ir jų kainos. Sklaidos diagrama (su fonu)

Išvados: grafike matosi, kaip yra išsidėstę pastatai ir kurie rajonai yra brangesni/pigesni. Didžioji dalis namų yra įsikūrę rajono centre, ten figūruoja ir didesnės kainos. Taip pat nemaža dalis gyventojų pirko namus netoli pagrindinių gatvių dėl patogesnio susisiekimo. *R kodas: 4 (su nuotrauka)*

ANTRA UŽDUOTIS

Reikėjo pasirinkti vieną geometrinio tiesioginio duomenų vizualizavimo metodą ir pavyzdžiu jį iliustruoti. Tyrimui pasilikome tuos pačius duomenis. Pasirinkome vizualizuoti lygiagrečiųjų koordinačių metodu. Kadangi buvome suskirstę duomenis pagal jų kainą, turėjome kategorinių duomenų. Taip pat turėjome 5 skaitinius atributus, kuriuos teoriškai galėjome naudoti (amžius, atstumas iki metro, parduotuvių skaičius, platuma, ilguma), tačiau nusprendėme atmesti geografines koordinates ir analizuoti tik namo amžių, atstumą iki metro ir parduotuvių skaičių netoliese. Taigi trijose vertikaliose ašyse buvo išdėstyti atrinkti kintamieji, mažiausia kintamojo reikšmė buvo apatinė riba, didžiausia – viršutinė. Skirtingų lygių kainos buvo vaizduojamos skirtingomis spalvomis. Šiam grafikui naudojome „ggally“ paketą.



pav. 6 Lygiagrečios koordinatės pagal būstų amžių, atstumą iki metro ir parduotuvių skaičių

Išvados: pagal namų amžių sunku išskirti aiškią grupės tendenciją. Tačiau dauguma vidutiniškų ir brangių namų yra netoli metro stočių, o pigių –toliau nei vidurkis. Parduotuvių skaičiaus pasiskirstymą yra galima išskirti tik pigesnių namų grupei, kurie telkiasi prie mažesnio parduotuvių skaičiaus. *Antros užduoties kodas*

TREČIA UŽDUOTIS

Reikėjo pavyzdžiu iliustruoti vieną simbolinį tiesioginio vizualizavimo metodą. Šiam uždaviniui taip pat naudojome duomenis apie būstus. Pasirinkome duomenis vaizduoti Černovo veidų metodu. Kadangi turėjome 414 atskirų namų ir juos visus atvaizdavus būtų sunku atskirti skirtingus veidus, mes atsirinkome atvaizduoti tik pirmus 25 būstus. Pasinaudojus paketu „aplpack“ pavaizdavome turimus duomenis.

pav. 7 Černovo veidų metodu atvaizduojami būstų duomenys

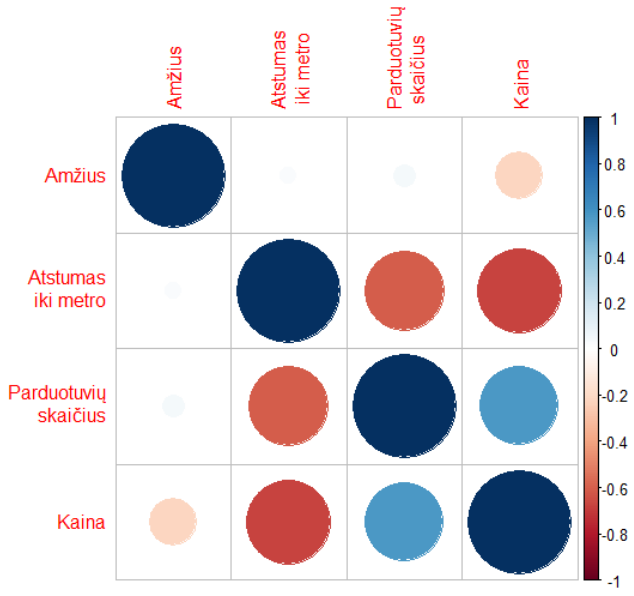
Atributai ir kokį viedo bruožą keičia:

- Pastato amžius: veido aukštis; akių aukštis; nosies plotis.
- Atstumas iki metro: veido plotis; akių plotis; ausų plotis.
- Parduotuvių skaičius: veido struktūra; plaukų aukštis; ausų aukštis.
- Platuma: burnos aukštis; plaukų plotis.
- Ilguma: burnos plotis; plaukų stilius (šukuosena).
- Kaina: šypsena; nosies aukštis.

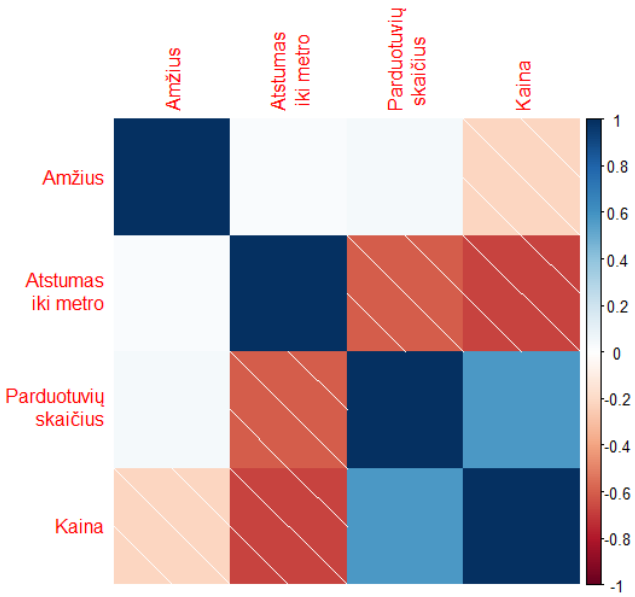
Trečios užduoties kodas

KETVIRTA UŽDUOTIS

Pavyzdžiu iliustruoti koreliacijos koeficientų vizualizavimą, naudojant „corrplot“ paketą. Šiam uždaviniui taip pat naudojome tuos pačius duomenis. Pasirinkome nustatyti kainos, būsto amžiaus, atstumo iki metro ir parduotuvių skaičiaus netolimoje distancijoje koreliacijas.



pav. 8 Koreliacijos lentelė („circle“)



pav. 9 Koreliacijos lentelė („shade“)

Išvados: galime pastebėti, kad yra stipri atvirkštinė koreliacija (raudona, arti -1) tarp namo kainos ir atumo iki metro stotelės, tai yra kuo didesnė kaina, tuo mažesnis atstumas. Vidutiniškai stiprūs sąryšiai tarp parduotuvių skaičiaus ir atstumo iki metro (atvirkštinė koreliacija, apie -0,6) bei kainos ir parduotuvių skaičiaus (tiesioginė koreliacija, apie 0,5). Taip pat koreliacijos beveik nėra tarp parduotuvių skaičiaus ir namo amžiaus arba atstumo iki metro ir namo amžiaus. Namo kaina neturi stipraus sąryšio su būsto amžiumi. *Ketvirtos užduoties kodas*

GALUTINĖS IŠVADOS

Apibendrinant galime teigti, kad didžioji dauguma namų telkiasi rajono centre, kur gausu parduotuvių ir metro stotys yra arti. Ten figūruoja aukštesnė kaina. Kainai didelės įtakos neturi namo amžius, duomenys pasiskirstę nepastoviai. Brangesni namai telkiasi prie susisiekimui patogesnių, tačiau nebūtinai smarkiai urbanizuotų vietų. Pigesni namai yra nuošaliau nuo centro, jų amžius didesnis, o jų pastatymo vieta yra mažiau patogi susisiekimo ir parduotuvių atžvilgiu. Vidutiniško brangumo namai pasiskirstę įvairiai, tačiau telkiasi prie susisiekimui patogesnių vietų.

PRIEDAI

Pirmos užduoties kodas

1.

```
ggplot(data=duomenys, aes(kaina)) + geom_histogram(binwidth=10, color = "black", fill = "white") + labs(x="Kaina",y="Dažnis")+ ggtitle("Namų kainų histograma") + geom_vline(aes(xintercept = mean(kaina)), linetype="dashed",col = "blue")+ geom_text(aes(x=mean(kaina), label=paste("Vidurkis",round(mean(kaina),2)), y=50), colour="blue", angle=90, vjust = 1.2)
```

2.

```
xyplot(duomenys$atstum_metro~duomenys$pard_sk|duomenys$kainos_iv, data=duomenys, main="Pasiskirstymo grafikas pagal namų kainas", ylab="Atstumas iki metro", xlab="Parduotuvių sk. ėjimo distancijoje", layout = c(1,3))
```

3.

```
ggplot(data=duomenys, aes(kaina, amzius)) + stat_density2d(aes(fill=..level..), geom="polygon") +scale_fill_gradient(low="blue", high="yellow") + labs(fill="Tankis")+ggtitle("Namų pasiskirstymas pagal kainą ir amžių") + labs(x="Kaina", y="Namo amžius")
```

4.

```
ggplot(data = temp, aes(ilguma, platuma, color=kaina)) + geom_point(size=2.3) + scale_color_viridis("Kaina") + ggtitle("Namų geografinis pasiskirstymas ir kainos") + labs(x="Ilguma", y="Platuma")
```

4 (su nuotrauka)

```
library(ggimage)

img="gg.jpg"

bandymas<-ggplot(data = temp, aes(ilguma, platuma, color=kaina)) + geom_point(size=2.3) + scale_color_viridis() +  theme(legend.position = "none", axis.title.x=element_blank(), axis.text.x=element_blank(), axis.title.y=element_blank(), axis.text.y=element_blank(), axis.ticks.y=element_blank(), axis.ticks.x=element_blank())

ggbackground(bandymas, img)
```

Antros užduoties kodas

```
install.packages("GGally")

library("GGally")
```

```
ggparcoord(duomenys, columns = 1:3, groupColumn = 7, scale="uniminmax") +
ggtitle("Lygiagrečiosios koordinatės") + theme(axis.title.y=element_blank(),
axis.text.y=element_blank(), axis.ticks.y=element_blank())+ labs(x="")+
scale_color_manual("Kainų kategorijos", values=c("#003380", "dark green", "yellow")) +
scale_x_discrete(labels=c("Amžius", "Atstumas iki metro", "Parduotuvių skaičius"))
```

Trečios užduoties kodas

```
library(aplpack)

faces(duomenys[1:25,1:6])
```

Ketvirtos užduoties kodas

```
library(corrplot)

duom4 <- subset(duomenys, select=c(amzius, atstum_metro, pard_sk, kaina))

M <-cor(duom4)

colnames(M) <- c("Amžius", "Atstumas\n iki metro", "Parduotuvių\n skaičius", "Kaina")

rownames(M) <- c("Amžius", "Atstumas\n iki metro", "Parduotuvių\n skaičius", "Kaina")

corrplot(M, method = "circle")

corrplot(M, method = "shade")
```

ŠALTINIAI

Duomenys: https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set