Univerzitet u Novom Sadu
Prirodno-matematički fakultet
Departman za matematiku i informatiku

Dijana Krstić

Istraživačka analiza podataka (EDA) uz upotrebu statističkog softvera R

Master rad

Mentor

dr Zagorka Lozanov-Crvenković

Novi Sad, jun 2016

Predgovor

"Ne postoji grana matematike, koliko god bila apstraktna, da jednog dana ne bi mogla biti primijenjena u praksi"

Nikolay Ivanovich Lobachevsky

Matematička statistika je oblast za koju interesovanje primenjenih matematičara sve više raste. Podrazumeva prikupljanje, prikazivanje, analizu i korišćenje podataka u cilju da se izvedu zaključci i donesu odluke. Primenjuje se skoro u svim oblastima, za davanje procena, istraživanje tendencija, procenu rizika, analiziranje odnosa i faktora koji određuju posmatranu pojavu.
Istraživačka analiza podataka (eng. Exploratory data analysis - EDA) je interesantan alat matematičke statistike koji se bazira na grafičkom i kvantitativnom prikazu podataka. U ovom radu je opisano 12 tehnika (10 grafičkih i 2 kvantitativne) uključujući implementaciju uz pomoć statističkog softvera R.

Prvo poglavlje je posvećeno opisu pojedinačnih EDA tehnika i njihovoj matematičkoj osnovi. Pored toga, sve tehnike potkrepljene su primerima, grafičkim i kvantitativnim prikazima, kao i adekvatnom analizom dobijenih rezultata. Takođe, date su i naredbe korišćene prilikom implementacije svake od tehnika. Razlog tome je činjenica da R podržava sve EDA tehnike, što nije slučaj sa ostalim statističkim paketima. Podaci (excel tabele), koji se koriste u primerima nalaze se na CD-u, u prilogu master rada.
Drugo poglavlje je posvećeno R-u. Detaljno su opisane sve naredbe koje se koriste u implementaciji EDA tehnika. U radu osim tehnika, naglašavamo i statistički paket R jer je u širokoj upotrebi u praksi. Čitaoci će pored teorijske osnove biti upoznati i sa direktnom implementacijom.
Treće poglavlje posvećeno je kompletnom aparatu EDA tehnika. Jedan primer je obrađen kroz više tehnika i dobijena je velika količina informacija. Izvršena je analiza tržišta nekretnina i korišćene su skoro sve EDA tehnike koje su mogle da se primene na datim podacima. Dobijeni rezultati su protumačeni tako da mogu biti

korisni i kupcima i podavcima nekretnina na tržištu.

Želela bih, na kraju, da se zahvalim svom mentoru dr Zagorki Lozanov-Crvenković na angažovanju, podršci, savetima i za svo znanje preneto na osnovim i master studijama. Takođe, veliku zahvalnost dugujem i članovima komisije, dr Mirjani Ivanović na preporučenoj interesantnoj temi i nesebičnoj pomoći tokom izrade master rada kao i dr Ivani Štajner-Papuga na korisnim savetima i pomoći tokom dosadašnjeg studiranja.
Zahvaljujem se i svojim kolegama i dragim prijateljima koji su moje studentske dane učinili lepšim.
Najveću zahvalnost dugujem svojoj porodici, a posebno roditeljima i sestri na bezuslovnoj ljubavi i podrsci koju mi neprestano pružaju.

Djjana Krstic

Sadržaj

Glava 1

Istraživačka analiza podataka (EDA)

1.1 Pojam i razvoj EDA

Istraživačka analiza podataka se bavi procedurama i tehnikama za analizu podataka kao i interpretacijom dobijenih rezultata, prikazujući ih najčešće vizuelno. Začetnik ove analize je Džon Taki (John Tukey) koji se datom oblašću počeo baviti još 1961. godine. Takijeva prvobitna definicija analize podataka glasi: "Analiza podataka predstavlja procedure za analiziranje podataka, tehnike za interpretaciju rezultata tih procedura, načini planiranja skupova podataka koji bi analizu učinili lakšom, preciznijom i tačnijom uključujući sve aparate statistike koji se primenjuju za analizu podataka". Razvojem ove oblasti podstiče se i razvoj statističkih paketa koji bi podržali novonastale tehnike analize. Korišćenjem modernog statističko-programskog okruženja, EDA se unapređuje i sve češće koristi. Najpre dolazi do njenog razvoja korišćenjem programskih paketa S, S-plus a zatim i R. EDA u kombinaciji sa programskim paketima predstavlja olakšanje statističarima koji se bave naukom ali i inženjerima kojima su rezultati datih analiza neophodni za rad. U knjizi koju je Taki izdao 1977. godine pod nazivom "Istraživačka analiza podataka" (eng. "Exploratory data analysis") naglasio je da se previše ističu statističke hipoteze prilikom testiranja. Smatrao je da veći naglasak treba da bude na korišćenju podataka koji će ukazivati na hipoteze koje treba testirati.

U nastavku ćemo izvršiti poređenje nekoliko tipova analiza. Razlikuju se u nizu koraka koji se izvršavaju prilikom analize podataka.

- Klasična (problem → podaci → model → analiza → zaključak).
- Istraživačka (problem → podaci → analiza → model → zaključak).
- Bajesovska (problem → podaci → model → pred distribucija → analiza → zaključak).

Kao što vidimo, sve analize počinju na isti način. Imamo problem i podatke međutim, kod klasične analize prvo pravimo model pa tek onda analiziramo podatke.

Formiranje modela podrazumeva nametanje modela (normalnog, linearnog i sl.), formiranje ocenjivača, testiranje i praćenje parametara datog modela. Kod klasične obrade to formiranje modela ima prednost u odnosu na analizu i po Takiju to je glavna mana klasičnih statističkih analiza podataka. Vidimo da bajesovska analiza prati klasičnu do momenta formiranja modela, tada bajesovska formira pred distribuciju na osnovu modela, koju koristi prilikom analize. Bajesovska analiza teži da ukomponuje nauku i inženjerstvo tj. teoriju i praksu da bi definisala raspodelu. Ona koristi poznate distribucije sa parametrima izabranog modela i na osnovu toga formira novu distribuciju.

Što se tiče samih modela, klasična analiza koristi determinističke i probabilističke modele. Deterministički modeli podrazumevaju regresioni model i analizu varijanse (anova). Zajedničko probabilističkim modelima je to što pretpostavljaju da greška ima normalnu raspodelu. EDA pristup ne nameće modele nego omogućava da podaci sugerišu prihvatljive modele koji im najbolje odgovaraju.

Možemo zaključiti da postoji velika razlika između klasične i EDA analize. Cilj klasične analize je na modelu, ocenjivanju parametara modela, predviđanju na osnovu modela. EDA se fokusira na podatke, njihovu strukturu, autlajere (vrednosti koje iz nekog razloga odstupaju od ostalih vrednosti podataka) kao i modele koje su sugerisali podaci. U praksi, podaci se analiziraju kombinacijom ovih tipova analiza ali u ovom radu ćemo se baviti isključivo istraživačkom analizom podataka, (videti [2]).

Cilj razvoja EDA je da se:

- obezbedi maksimalni uvid u skup podataka,
- otkriju osnovne strukture podataka,
- detektuju važne promenljive,
- otkriju autlajeri i uticajne tacke,
- testiraju pretpostavke,
- razviju škrti modeli,
- utvrde uticajni faktori.

1.2 Tehnike EDA

EDA je karakteristična po tome što se prilikom analize podataka koriste dve vrste tehnika:

- grafičke i
- kvatitativne.

EDA u suštini, nije set tehnika nego pristup tome kako da se nosimo sa različitim tipovima podataka. Za razliku od statističkih grafika koji posmatraju podatke sa jednog aspekta, EDA ne koristi uobičajene pretpostavke nego koristi različite pristupe podacima. Fokusira se na informacije koje tražimo iz podataka, način na koji ih tražimo, na strukturu, model kao i način tumačenja podataka. EDA koristi tehnike prilikom ispitivanja ali one se znatno razlikuju od uobičajenih statističkih tehnika. Naime, grafičke tehnike su znatno bolje od kvantitativnih jer tako podaci najbolje otkrivaju svoje karakteristike. Grafički prikazi u kombinaciji sa našim mogućnostima za prepoznavanje obrazaca predstavljaju nesumnjivo najbolju tehniku prilikom tumačenja dobijenih informacija. U ovom radu biće detaljno opisane sledeće **grafičke tehnike EDA**:

- Histogram,
- Dijagram rasturanja,
- Boks dijagram,
- Dijagram protoka u vremenu koji prikazuje tok posmatranih podataka na vremenskoj osi.
- Pareto dijagram koji se koristi kada se iz velikog broja različitih činjenica želi na neki način identifikovati relativna važnost različitih pojedinosti procesa (ili grupe grešaka). Ovaj dijagram služi za identifikaciju najvažnijih problema, otkrivanje uzroka problema, i sl.
- Dijagram paralelnih koordinata je uobičajen način vizuelnog predstavljanja višedimenzionalnih promenljivih.
- OR količnik ili odnos šansi daje meru povezanosti dve nezavisne promenljive sa ishodom koji nas interesuje.
- Multidimenzionalno skaliranje grupa je metoda za procenu koordinata seta objekata iz podataka o udaljenosti između parova objekata.
- Dijagram stablo-list je prikaz koji se koristi za interpretaciju gustine i oblika podataka, detektovanje autlajera, prepoznavanje distribucije kao i pronalaženje medijane, modusa i sl.
- Violina dijagram vizuelno je sličan boks dijagramu. Jezgro violina dijagrama je rotirano.

Kada govorimo o **kvantitativnim tehnikama** tu se nameću

- Doterana sredina ispituje značaj različitih faktora višefaktorskih modela (robusniji od anove). Reprezentuje se postupkom koji otkriva uticaj medijane po vrstama i kolonama matrice (tabele) koja se sastoji od faktora.
- Usečena sredina koja se zasniva na aritmetičkoj sredini zbirova dvostruke medijane kompletnog uzorka i jednostrukim medijanama prvog i drugog kvartila uzorka.

1.2.1 Histogram

U statistici, histogram je grafički prikaz tabeliranih frekvencija podataka. Izumeo ga je Karl Pearson 1891. godine. Uopšteno, histogram je definisan kao način prikazivanja podataka raspoređenih u određene kategorije (klase) ili grupe. Prvi korak u kreiranju histograma je skupljanje podataka i razvrstavanje prikupljenih podataka u kategorije. Nadalje moramo odrediti koje su promenljive zavisne, a koje nezavisne. Karakteristika po kojoj smo grupisali podatke u kategorije predstavlja nezavisnu promenljivu, a broj prikupljenih podataka koji upadaju u određenu kategoriju predstavlja zavisnu promenljivu.
Histogram je zapravo stubičasti graf, koji na apscisnoj osi ima vrednosti nezavisne promenljive, a na ordinatnoj osi vrednosti zavisne promenljive. Oznake na osama treba da budu linearno raspoređene. Graf se crta tako da se prvo na apscisu nanesu vrednosti svih kategorija, čime dobijamo apscisu podeljenu na intervale. Zatim se broj podataka koji odgovaraju toj kategoriji (frekvencija) crta kao horizontalna linija iznad odgovarajućeg intervala. To je razlog zbog kojeg dobijamo stubičasti graf. Histogram je alat koji pomaže da se brzo uoči tip raspodele za uzorke koji sadrže veliki broj podataka. Izradi histograma prethodi: izručanavanje raspona populacije, određivanje intervala klasa, izrada tabela učestalosti, određivanje granica klasa, sračunavanje središta klasa i određivanje učestalosti prebrojavanjem uzorka. Na osnovu ovih podataka crta se histogram (videti [6]). Na osnovu izgleda histograma donose se zaključci o statističkoj prirodi populaciji. Da bismo bolje razumeli način prikazivanja podataka uradićemo primer uz upotrebu softvera.

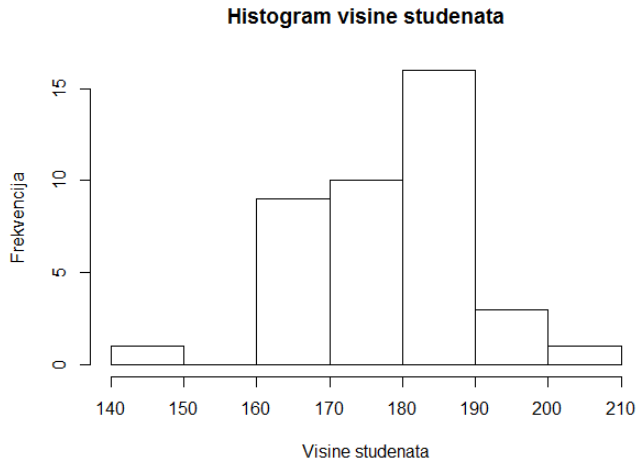
Primer 1.1. Na nekom fakultetu je odabran uzorak od 40 studenata i izmerene su im visine (podaci u prilogu pod nazivom visina.csv [1]). Nacrtaćemo histograme apsolutnih i relativnih frekvencija na osnovu datih podataka. Plavom bojom će biti označene naredbe za programski paket R, neophodne za crtanje grafika.

```
> visina<-read.table("visina.dat",header=FALSE)
> attach(visina)
> hist(V1, xlab="Visine studenata", ylab="Frekvencija", main="Histogram visine studenata")
```

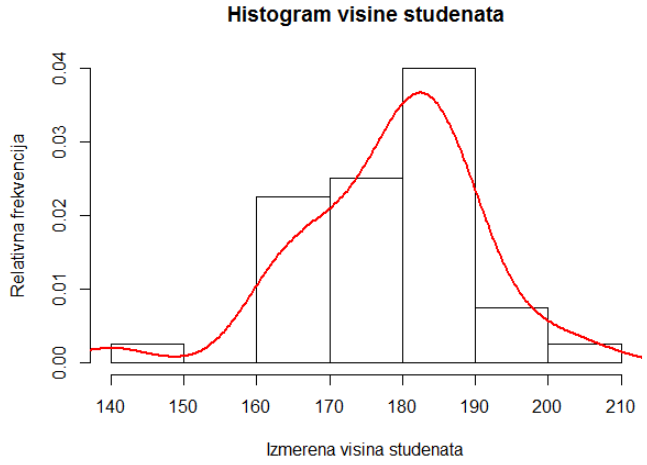
Ako umesto frekvencija f_i koristimo relativne frekvencije $r_i = \frac{f_i}{n}$, gde je $i = 1...n$ broj elemenata uzorka dobijamo drugačiji prikaz, (slika 1.2).

[1] podaci iz lične arhive

> hist(V1, probability=TRUE, ylab="Relativne frekvencije", xlab="Izmerena visina studenat", main="Histogram visine studenata")
> lines(density(V1))



Slika 1.1: Histogram na osnovu frekvencije podataka



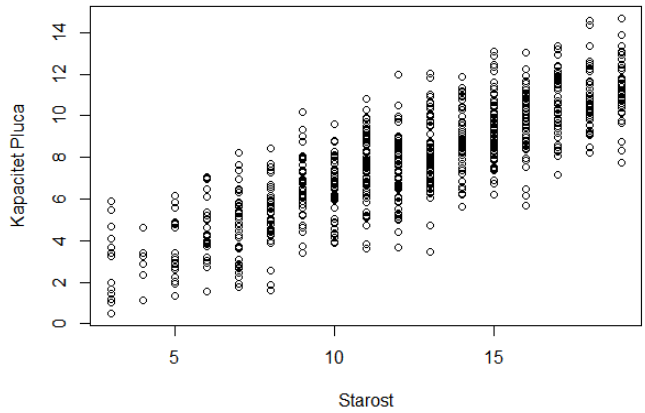
Slika 1.2: Histogram na osnovu relativne frekvencije podataka

1.2.2 Dijagram rasturanja

Dijagram rasturanja (eng. scatter plot) je tip matematičkog dijagrama koji se koristi da bi prikazao tipične vrednosti dve promenljive u koordinatnom sistemu. Kreirao ga je britanski statističar Francis Galton 1888. godine da bi prikazao vezu (korelaciju) između dve promenljive. Podaci su prikazani kao kolekcija tačaka u koordinatnom sistemu gdje su vrednosti promenljivih prikazane na x i y osi.

Primer 1.2. Posmatrajmo odnos vrednosti kapaciteta pluća kod osoba različite starosti, pola, visine i sklonosti ka pušenju. Koristeći podatke iz priloga pod nazivom KapacitetP.csv (videti [18]), nacrtaćemo dijagram rasturanja u kome ćemo prikazati kapacitet pluća u zavisnosti od starosti, (slika 1.3).

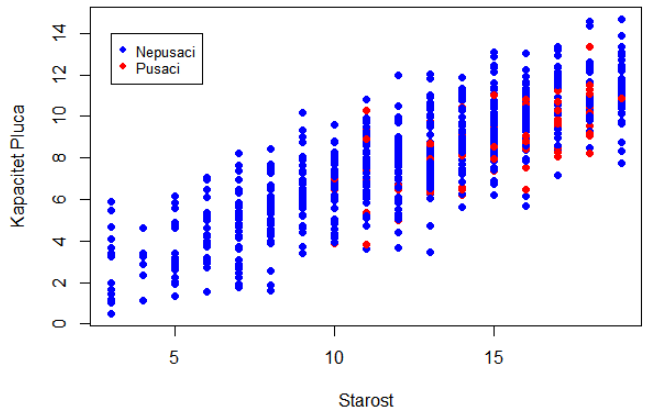
```
> plot(KapacitetP$Starost, KapacitetP$KapacitetPluca, pch=1, xlab= "Starost",
ylab= "Kapacitet Pluca")
```



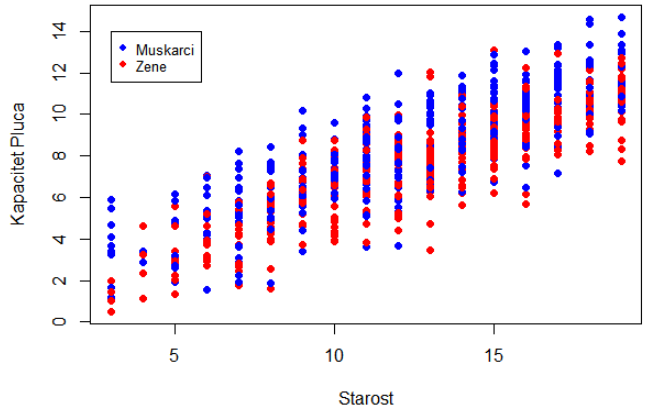
Slika 1.3: Dijagram rasturanja kapaciteta pluća u zavisnosti od starosti

Na slici 1.3 vidimo da se povećanjem broja godina (od rođenja do dvadesete godine) povećava i kapacitet pluća, linearno i zaključujemo da imamo međusobnu povezanost ove dve promenljive. Koeficijent korelacije je pozitivan ali je rasipanje podataka (disperzija) veliko.
Podatke o tome da li su posmatrane osobe muškarci ili žene kao i da li su pušači ili nepušači vidimo na slikama 1.4 i 1.5.

```
> plot(KapacitetP$Starost, KapacitetP$KapacitetPluca , pch=16, xlab= "Starost",
ylab= "Kapacitet Pluca", col= ifelse(KapacitetP$Pusaci=="no", "blue", "red"))
> legend(3,14, pch=c(16,16), col=c("blue", "red"), c("Nepusaci", "Pusaci"), bty="o",
box.col="black", cex=.8)
> plot(KapacitetP$Starost, KapacitetP$KapacitetPluca, pch=16, xlab= "Starost",
ylab= "Kapacitet Pluca", col= ifelse(KapacitetP$Pol == "male", "blue", "red"))
> legend(3,14, pch=c(16,16), col=c("blue", "red"), c("Muskarci", "Zene"), bty=
"o", box.col="black", cex=.8)
```



Slika 1.4: Dijagram rasturanja prema kriterijumu ”pušači”

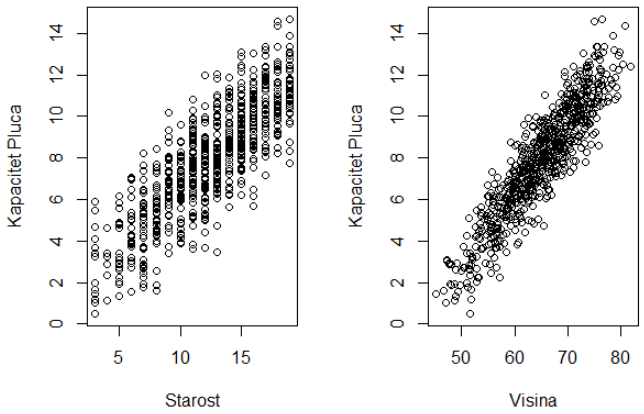


Slika 1.5: Dijagram rasturanja prema kriterijumu ”pol”

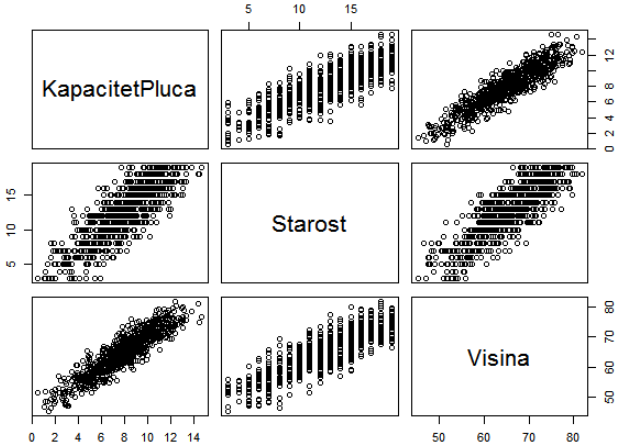
Ako u posmatranje uključimo i kategorijalnu promenljivu koja se odnosi na kategoriju ”pušači”, na slici 1.4 vidimo da osobe mlađe od 10 godina nikad nisu probale cigarete, dok se sa povećanjem broja godina povećava i broj pušača kao i da je njigov kapacitet pluća niži u odnosu na vršnjake koji spadaju u kategoriju nepušača. U kategoriji ”pol” (slika 1.5), vidimo da osobe ženskog pola imaju nešto manji kapacitet pluća u odnosu na osobe muškog pola istih godina.
Dijagram rasturanja nam pomaže i u slučaju kada želimo uporediti uticaj dve ili više nezavisnih promenljivih na zavisnu promenljivu (slike 1.6 i 1.7).

```
> par(mfrow=c(1,2))
> plot(KapacitetP$Starost, KapacitetP$KapacitetPluca, pch=1, xlab= ”Starost”, ylab= ”Kapacitet Pluca”)
> plot(KapacitetP$Visina, KapacitetP$KapacitetPluca, pch=1, xlab= ”Visina”, ylab= ”Kapacitet Pluca”)

> plot(KapacitetP[,c(1,2,3)])
```



Slika 1.6: Upredni dijagrami rasturanja (1)



Slika 1.7: Upredni dijagrami rasturanja (2)

Na osnovu dobijenih rezultata uporednog dijagrama zaključujemo da se kapacitet pluća verodostojnije prikazuje u zavisnosti od visine osoba. Povećanjem visine povećava se i kapacitet pluća, linearno, koeficijent korelacije između ove dve promenljive je pozitivan a rasipanje podataka je manje u odnosu na analizu kapaciteta pluća u zavisnosti od starosti.

1.2.3 Boks dijagram

Neka je $x = (x_1, ..., x_n)$ uzorak neke numeričke promenljive u rastućem, varijacionom nizu.
Medijana m je mera srednje vrednosti i to je vrednost koja elemente uzorka u rastućem, variacionom nizu, deli na dva jednaka dela (videti [7]).

$$m = \begin{cases} x_{(n+1)/2}; & \text{za } n \text{ neparno,} \\ \frac{x_{n/2}+x_{n/2+1}}{2}; & \text{za } n \text{ parno} \end{cases}$$

Raspon uzorka

$$R = x_n - x_1$$

Interkvartil

$$IQR = q_u - q_l$$

gde su q_l i q_u donji i gornji kvartil

$$q_l = x_{(n+1)/4}$$
$$q_u = x_{3(n+1)/4}$$

Kvantil
Za $p \in (0, 1)$ definišemo p kvantil kao

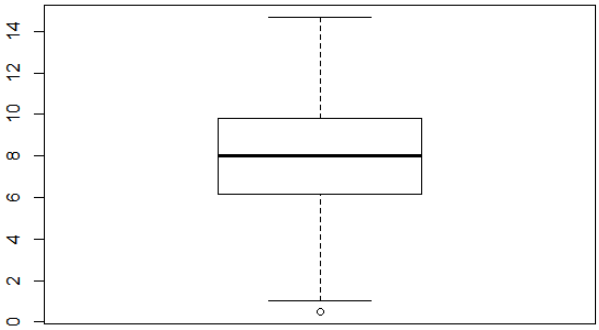
$$q_p = x_{(p(n+1))}$$

Donji kvartil je onda 0.25-kvantil, a gornji kvartil je 0.75-kvantil. Medijana je 0.5-kvantil.

Kutijasti ili boks dijagram (eng. box and whisker plot) je jednostavan graf koji prikazuje karakterističnu petorku (x_1, q_l, m, q_u, x_n) (eng. five-number summary). Boks dijagram se sastoji od pravougaonika koji prikazuje podatke od donjeg do gornjeg kvartila. Horizontalna linija po pravougaoniku označava medijanu. Donje i gornje horizontalne linije se nazivaju "whisker". Mogu se različito definisati, ali najčešće predstavljaju najmanji i najveći podatak koji se nalazi unutar 1.5 puta interkvartilni raspon gledajući od donjeg, odnosno gornjeg kvartila. Sve tačke izvan te granice se crtaju posebno i smatraju autlajerima. Izgled boks dijagrama ukazuje na stepen raspršenosti i asimetričnosti, te može pokazati autlajere među podacima.

Primer 1.3. Posmatraćemo ponovo podatke KapacitetP.csv, (slika 1.8). Na slici su prikazane sve pomenute karakteristike, uključujući i autlajer.

```
> boxplot(KapacitetPluca)
```



Slika 1.8: Boks dijagram kapaciteta pluća

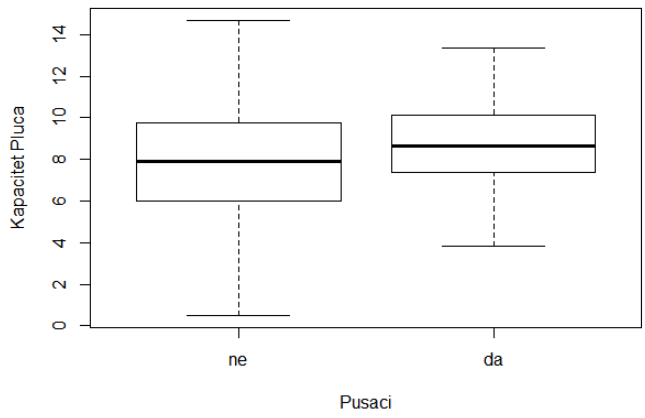
Ukoliko želimo i tačne podatke o kvartilima, minumumu, maksimumu, medijani i aritmetičkoj sredini koristićemo naredbu summary.

```
> summary(KapacitetPluca)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.507	6.150	8.000	7.863	9.800	14.680

Primer 1.4. Posmatrajmo sada kapacitet pluća u zavisnosti od toga da li je osoba pušač ili ne, (slika 1.9).

```
> boxplot(KapacitetPluca ~ Pusaci, data=KapacitetP, xlab="Pusaci",
ylab="Kapacitet Pluca", names=c("ne","da"), las=1)
```



Slika 1.9: Boks dijagram kapaciteta pluća u odnosu na kategoriju pušači-nepušači

Na slici 1.9 primećujemo manji raspon uzorka između posmatranih kategorija. Razlog tome je što su mlađa deca nepušači pa u tom uzrastu nemamo podatke o kapacitetu pluća pušača a samim tim ni elemente boks dijagrama. Kada posmatramo gornju granicu ovde primećujemo manji kapacitet pluća kod pušača u odnosu na nepušače. Primetimo još i da je vrednost medijane kod pušača veća. Odavde se lako može izvući pogrešan zaključak da pušači u proseku imaju veći kapacitet pluća od nepušača, što nije istina. Razlog veće vrednosti medijane kod pušača je činjenica da je prosečna starosna dob kod pušača veća od prosečne starosne dobi nepušača u posmatranom uzorku (deca do 10 godina su uglavnom nepušači).

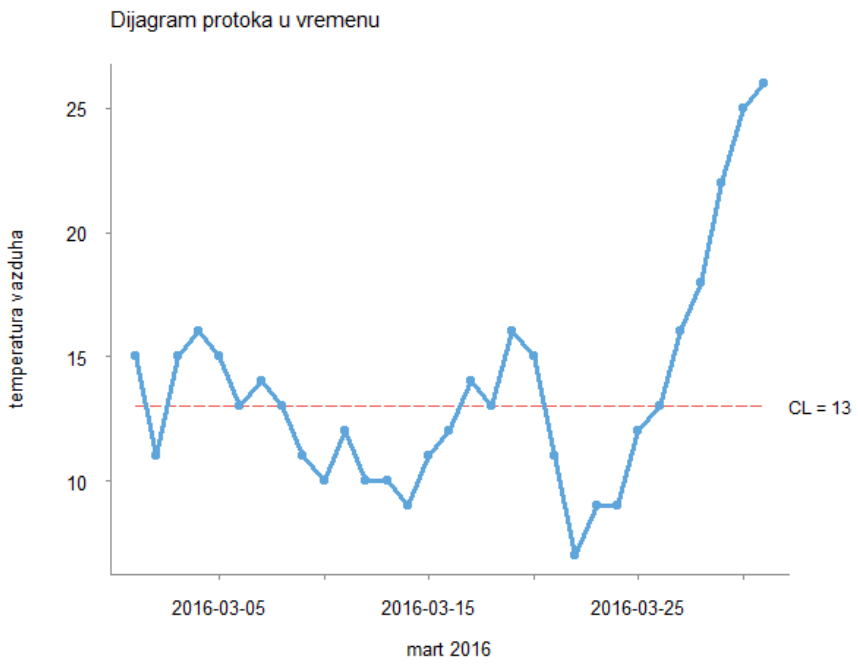
1.2.4 Dijagram protoka u vremenu

Dijagram protoka u vremenu (eng. run chart) je grafički prikaz promenljive zavisne od vremena. Horizontalna x-osa predstavlja vremensku osu (dane, mesece, godine, kvartale). Vertikalna y-osa predstavlja vrstu promenljive (indikatora) koju posmatramo. Centralna linija na grafiku predstavlja medijanu (videti [11]).
Karakteristike dijagrama protoka u vremenu koje otkrivaju da prikazani podaci imaju obrasce su:

- **Promena (eng. shift)** - odnosi se na šest ili više uzastopnih tačaka koje se nalaze iznad ili ispod medijane. Grafički se naglašava koje uzastopne vrednosti su manje ili veće od medijane.
- **Trendovi (eng. trends)** - podrazumevaju slučaj da pet i više uzastopnih tačaka formiraju rast ili pad funkcije na grafiku. Ako su vrednosti dve ili više tačaka jednake uzimamo samo vrednost prve a ostale zanemarujemo da ponovljene vrednosti ne bi pokvarile trend strogog rasta ili pada.
- **Tok (eng. run)** - uzorci koji nisu slučajni signalizuju nekoliko (ili više) tokova, odnosno prelazaka preko linije medijane. Tok je niz povezanih tačaka koje se nalaze na jednoj strani u odnosu na liniju medijane. Kada linija grafika preseče medijanu formira se novi tok. Neke tačke mogu biti na liniji medijane i onda je teško odrediti kom toku one pripadaju.
- **Ekstremne tačke (eng. astronomical point)** - detektuju se neobično visoke ili neobično niske vrednosti podataka.

Primer 1.5. Posmatrajmo maksimalne dnevne vrednosti temperature vazduha u Srbiji, u martu 2016., videti [21], podaci martT.csv, (slika 1.10).

```
> install.packages("qicharts")
> library(qicharts)
> attach(martT)
> y<-temperatura
> x<- seq.Date(as.Date('2016-03-01'), by='day', length=31)
> qic(y,x=x, ylab="temperatura vazduha", xlab="mart 2016", main= "Dijagram
protoka u vremenu")
```



Slika 1.10: Dijagram protoka u vremenu (maksimalne dnevne tremperature u martu 2016.)

Na slici 1.10 primećujemo da je medijana 13°. Promene uočavamo od 9. do 16. marta kao i od 20. do 26. marta kada su zabeležene temperature koje su ispod srednjih vrednosti za to doba godine. Trend primećujemo od 22. do 31. marta i zaključujemo da je to trend rasta temperature sve do kraja meseca. Takođe, primećujemo i pet tokova (pet kontinualnih promena temperature u odnosu na medijanu). Ekstremnih tačaka u ovom slučaju nema (iako primećujemo visoku vrednost temperature na kraju meseca ona je narasla kontinualno- pratila je trend rasta).

1.2.5 Pareto dijagram

Pareto analiza (dijagram), nazvana prema italijanskom ekonomisti Vilfredu Paretu, razvijena je kao dijagramska metoda za grupisanje uzroka problema prema njihovom relativnom značaju. Predstavlja postupak odabira prioritetnih problema za rešavanje i jedan je od dobrih i jednostavnih načina za diferencijaciju najvažnijih problema. Pareto dijagram je vrsta grafikona koja prikazuje frekvenciju (sumu) događaja koji pripadaju različitim kategorijama, od najveće frekvencije sa leve, do najmanje frekvencije sa desne strane grafika, sa linijom koja prekriva dijagram i koja prikazuje kumulativni procenat događaja. Vertikalna osa sa leve strane prikazuje frekvenciju (sumu) dok vertikalna osa sa desne strane prikazuje kumulativni procenat. Polje primene ove metode je široko: otkrivaju se osnovne vrste nedostataka, najčešći razlozi za reklamaciju od strane kupaca, istražuju se mogućnosti da se smanje gubici, smanje zastoji u proizvodnom procesu, racionalizuje potrošnja materijala, u istraživanju rentabilnosti proizvodnog programa, pri proučavanju rada, da bi se prikazao skup nekih promenljivih (npr. novac, energija, vreme) koje mogu biti klasifikovane prema određenoj kategoriji i sl.

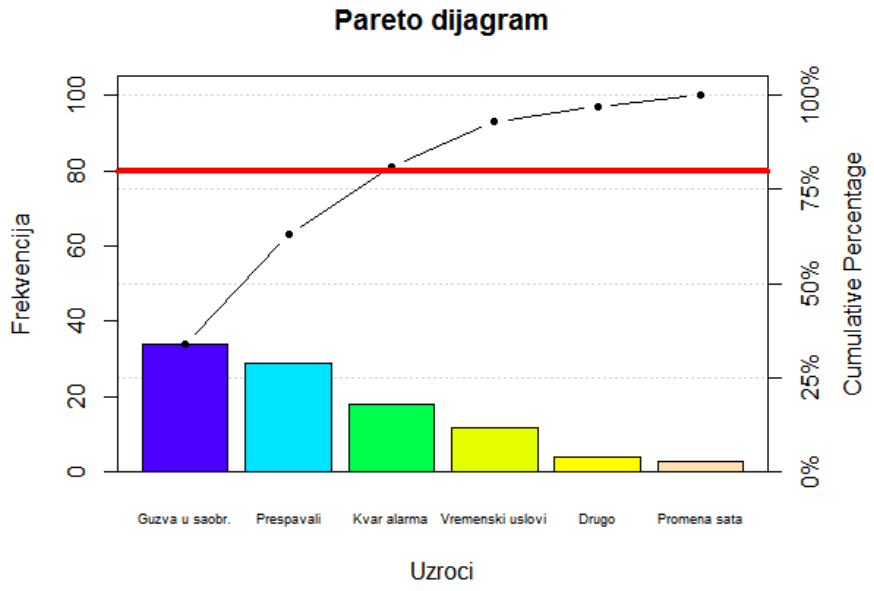
Osnovu metode čini ideja nekoliko značajnih i mnogo beznačajnih faktora koji utiču na pojavu koja se ispituje. Vrlo često se događa da je više od polovine svojstava jednog problema rezultat jednog istog uzroka. U takvoj situaciji mnogo bolji prilaz predstavlja lokalizacija i eliminisanje najznačajnijih, nego pokušaj eliminisanja svih uzroka istovremeno. Eliminisanje jednog značajnog uzroka će rezultovati drastičnim poboljšanjem kvaliteta uz minimalni uloženi napor. Obično se koristi da razdvoji značajnu manjinu od neznačajne većine. Koristeći Pareto princip koji se zove i "Pravilo $80-20$" koje tvrdi da 80 % efekata potiče od 20 % uzroka za mnoge sisteme. Na primer 80 % problema javlja se kod 20 % uređaja ili kod 20 % zaposlenih ili, 80 % blagostanja društva koncentrisano je na 20 % stanovništva. Pareto dijagram se može koristiti bilo sa kvantitativnim, bilo sa atributivnim podacima, ali najčešce sa atributivnim. Obično su ovi podaci izraženi u procentima (videti [13]). Na primer, raspoloživi podaci se mogu grupisati tako da se stekne uvid da je mnoštvo neispravnosti proizvoda (usluge) izazvano nekim uzrocima.

Primer 1.6. Zaposleni često kasne na posao i želimo saznati uzroke njihovog kašnjenja. Anketirano je 100 ispitanika i dobijeni su rezultati prikazani u tabeli 1.1 (videti [13]).

Uzrok kašnjenja	Ukupno
Loše vreme	12
Kvar alarma	18
Saobraćaj	34
Prespavali	29
Promena sata	3
Drugo	4

Tabela 1.1: Uzroci kašnjenja

```
> uzroci <- c(12,29,18,3,34,4)
> Uzroci<-data.frame(uzroci)
> names(uzroci) <- c("Vremenski uslovi", "Prespavali", "Kvar alarma", "Promena
sata", "Gužva u saobraćaju", "Drugo")
> install.packages("qcc")
> library(qcc)
> pareto.chart(uzroci, main= "Pareto dijagram", xlab= "Uzroci", ylab=
"Frekvencija", cex.names= 0.5, las=1, col= topo.colors(6))
> abline(h=(sum(Uzroci)*.8),col="red",lwd=4)
```



Slika 1.11: Pareto dijagram

Pareto chart analiza

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
Guzva u saobracaju	34	34	34	34
Prespavali	29	63	29	63
Kvar alarma	18	81	18	81
Vremenski uslovi	12	93	12	93
Drugo	4	97	4	97
Promena sata	3	100	3	100

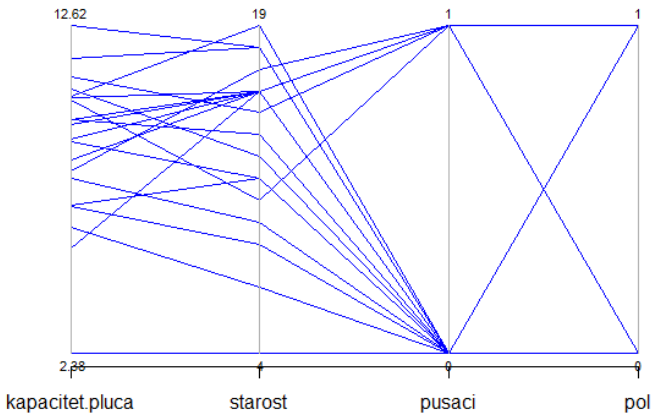
Na osnovu dobijenih rezultata vidimo da gužva u saobraćaju i zaposlenu koji se nisu probudili na vreme (prespavali su) čine 63% uzroka kašnjenja. Gužva u saobraćaju, prespali i kvar alarma čine značajnu većinu od 81% dok vremenski uslovi, promena sata i drugi uzroci čine beznačajnu manjinu od 19%. Dakle, ukoliko bi zaposleni preduzeli neke mere, kao što su npr. promene ruta kojima se kreću do odredišta, redovno i pažljivo navijanje nekoliko alarma sprečili bi značajnu većinu kašnjenja a samim tim i smanjenje zarade zbog kašnjenja.

1.2.6 Dijagram paralelnih koordinata

Dijagram paralelnih koordinata (eng. parallel coordinate chart) predstavlja vizuelizaciju podataka pomoću vertikalnih osa. Svaka osa predstavlja jedan skup podataka. Linije koje seku vertikalne ose predstavljaju vezu između pomenutih skupova podataka.
Ideja o korišćenju paralelnih koordinata potiče od 1885. godine kada je Maurice D' Ocagne kreirao nomograf [2]. Pojam paralelne koordinate popularizovan je tek 1959. godine kada je Alfred Inselberg osmislio koordinatni sistem sa paralelnim koordinatama.
Paralelne koordinate su dobar način vizualizacije multivarijantnih podataka, pri čemu je moguće uočiti prisustvo autlajera, postojanje i smer korelacije između pojedinih promenljivih kao i svojevrstan multivarijantni trag svake ispitivane observacije (videti [22]).

Primer 1.7. Posmatraćemo podatke o vrednosti kapaciteta pluća (pluca.csv, [18]) u odnosu na starost i kategorije u koje spadaju posmatrani ispitanici. Kategorija "pušači" uzimaće vrednost 0 za ispitanike koji su nepušači a vrednost 1 za ispitanike koji su pušači. Kategodija "pol", vrednost 0 za muškarce, vrednost 1 za žene. Na slici 1.12 prikazani su dobijeni rezultati.

```
> install.packages("parcor")
> library(parcor)
> p<-data.frame(pluca)
> parcoord(p, col=4, lty=1, var.label=TRUE)
```



Slika 1.12: Dijagram paralelnih koordinata

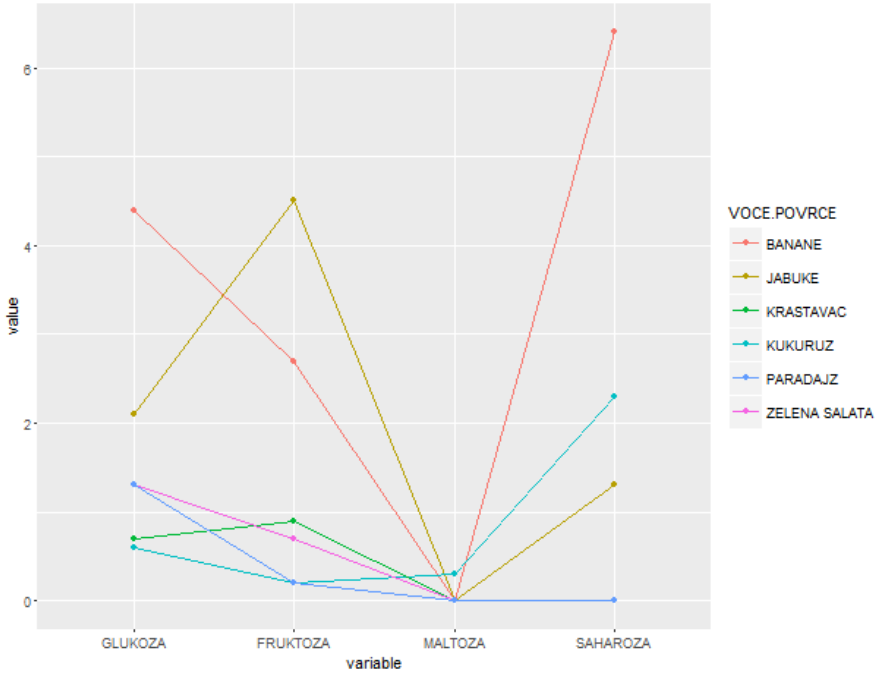
Na slici 1.12 vidimo linearnu vezu između vrednosti kapaciteta pluća i starosti, što su ispitanici stariji to imaju veći kapacitet pluća. Takođe, vidimo da su uglavnom

[2] Klasični primer nomografa je termometar koji sa jedne strane meri temperaturu vazduha u Faradima a sa druge strane temperaturu u Celzijusovim stepenima.

stariji ispitanici pušači. Primetimo da dvoje ispitanika koji spadaju u kategoriju pušača imaju i manji kapacitet pluća (kapacitet pluća je u negativnoj korelaciji sa starosnom dobi). Primećujemo, takođe, da su i većina ispitanika nepušači, oba pola.

Primer 1.8. Posmatrajmo količine četiri vrste šećera u voću i povrću: glukoza[3], fruktoza[4], saharoza[5] i maltoza[6], (podaci u prilogu pod nazivom voce.csv).

```
> install.packages("GGally")
> library(GGally)
> attach(voce)
> ggparcoord(voce, columns=2:5, groupColumn=1, showPoints=TRUE,
scale="globalminmax")
```



Slika 1.13: Dijagram paralelnih koordinata

Na slici 1.13 primećujemo drugačiji izgled dijagrama paralelnih koordinata u odnosu na sliku 1.12 pa je samim tim i tumačimo na drugačiji način. Osoba koja npr. iz zdravstvenih razloga mora da brine o tome koliku količinu koje vrste šećera će uneti u organizam, na osnovu ovog dijagrama lako može odlučiti koje voće (povrće) će da konzumira. Primećujemo da banane imaju veliku količinu saharoze pa je to uzrok težeg razgrađivanja u organizmu. Banane i jabuke imaju veliku količinu fruktoze (šećera koji je po ukusu najslađi) pa je to uzrok slatkastog ukusa ovog voća. Svo voće i povrće ima određene količine glukoze koja povećava nivo šećera u krvi. Maltoza je jedini šećer koji nije prisutan u velikim količinama ni u jednom voću (povrću) osim u kukuruzu (žitaricama).

[3]Glukoza-najvažniji šećer za živu ćeliju organizama koji se brzo razgrađuje;
[4]Fruktoza-brzo razgradiv voćni šećer, mnogo slađi od glukoze;
[5]Sahazoza-beli šećer koji se teže razgrađuje;
[6]Maltoza-teže razgradiv šećer, najzastupljeniji u žitaricama;

1.2.7 OR količnik

Statistička analiza sa kojom smo se upoznali u toku školovanja uglavnom se bavila analizom numeričkih registrovanih podataka. Međutim, ekspanzija razvoja metoda za analiziranje kategorijalnih (atributivnih) podataka koja je započela 1960. godine, nastavila je da se razvija ubrzano.
U bilo kom regresionom modelu ključno je odrediti očekivanu vrednost zavisne promenljive Y za datu vrednost nezavisne promenljive X, u oznaci $E(Y|X)$. Kako je zavisna promenljiva Y dihotomna i uzima vrednosti 0 i 1, uzećemo da uzima vrednost 1 sa verovatnoćom π, a vrednost 0 sa verovatnoćom $1-\pi$. Slučajna promenljiva $Y|X$ će takođe uzimati vrednosti 1 i 0, sa verovatnoćama $\pi(x)$ i $1-\pi(x)$, tj.

$$Y|X = \begin{pmatrix} 1 & 0 \\ \pi(x) & 1-\pi(x) \end{pmatrix}$$

Slučaj kada je i nezavisna promenljiva u logističkom regresionom modelu dihotomna predstavlja osnovu za druge slučajeve i podrazumeva da nezavisna promenljiva može uzeti dve vrednosti. U našem slučaju neka je nezavisna promenljiva kodirana sa 0 i 1. Uvešćemo pojam odnos šansi (unakrsni odnos šansi, odds ratio) koji daje meru povezanosti nezavisne promenljive sa ishodom od interesa. **Šansa** je odnos verovatnoća da se događaj desi prema verovatnoći da se događaj ne desi (videti [8]). Šansa da je zavisna promenljiva uzela vrednost 1, kada nezavisna promenljiva uzme vrednost 1 je:

$$\frac{P(Y=1|X=1)}{P(Y=0|X=0)} = \frac{\pi(1)}{1-\pi(1)}$$

Kada nezavisna promenljiva uzme vrednost 0, šansa da je zavisna promenljiva uzela vrednost 1 je:

$$\frac{P(Y=1|X=0)}{P(Y=0|X=0)} = \frac{\pi(0)}{1-\pi(0)}$$

Odnos šansi (unakrsni odnos šansi), u oznaci OR, je definisan kao odnos ove dve šanse, tj.

$$OR = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} \tag{1.1}$$

Primer 1.9. U tabeli 1.2 dati su rezultati jednog od prvih istraživanja veze između karcinoma pluća i pušenja, sprovedena od strane Ričarda Dola i Bredforda Hila. U 20 bolnica u Londonu, pacijenti primljeni sa rakom pluća ispitivani su o svojim pušačkim navikama proteklih godina. Za svakog od 709 pacijenata sa rakom, istraživači su ispitivali i pušačke navike pacijenata bez karcinoma, istog pola i približno istih godina.

Pušači	Oboleli	Nisu oboleli	Ukupno
Da	688	650	1338
Ne	21	59	59
Ukupno	709	709	1418

Tabela 1.2: Rezultati istraživanja

Na osnovu ovih podataka vidimo da je verovatnije da oboli pušač nego nepušač pa ćemo obolele pušače uzeti kao referentan ishod. Dakle, tražićemo koliko veću šansu pušači imaju da obole u odnosu na nepušače.
Šansa da pušač oboli je:

$$\frac{\frac{688}{709}}{\frac{21}{709}} = 32.76$$

Šansa da pušač ne oboli je:

$$\frac{\frac{650}{709}}{\frac{59}{709}} = 8.23$$

Odnos šansi je:

$$OR = \frac{32.76}{8.23} = 3.98$$

Dakle, pušači imaju skoro 4 puta veće šanse da obole u odnosu na nepušače.
Ukoliko bismo želeli da odnos šansi izračunamo koristeći R, to ćemo uraditi na sledeći način:

```
> mymatrix <- matrix(c(688,650,21,59),nrow=2,byrow=TRUE)
> colnames(mymatrix) <- c("Oboleli","NisuOboleli")
> rownames(mymatrix) <- c("Pusaci","Nepusaci")
> print(mymatrix)

          Oboleli NisuOboleli
Pusaci        688         650
Nepusaci       21          59
> calcOddsRatio <-function(mymatrix,alpha=0.05,referencerow=2,quiet=FALSE)
+  numrow <- nrow(mymatrix)
+ myrownames <- rownames(mymatrix)
+ for (i in 1:numrow)
+  rowname <- myrownames[i]
+ OboleliNepusaci <- mymatrix[referencerow,1]
```

```
+ NisuOboleliNepusaci <- mymatrix[referencerow,2]
+ if (i != referencerow)
+  OboleliPusaci <- mymatrix[i,1]
+ NisuOboleliPusaci <- mymatrix[i,2]
+ + totPusaci <- OboleliPusaci + NisuOboleliPusaci
+ totNepusaci<- OboleliNepusaci + NisuOboleliNepusaci
+ + sansaOboleliPusaci<- OboleliPusaci/totPusaci
+ sansaOboleliNepusaci <- OboleliNepusaci/totNepusaci
+ sansaNisuOboleliPusaci <- NisuOboleliPusaci/totPusaci
+ sansaNisuOboleliNepusaci <- NisuOboleliNepusaci/totNepusaci
+ oddsRatio <- (sansaOboleliPusaci*sansaNisuOboleliNepusaci)/
+ (sansaNisuOboleliPusaci*sansaOboleliNepusaci)
+ if (quiet == FALSE)
+  print(paste("category =", rowname, ", odds ratio = ",oddsRatio)) +
confidenceLevel <- (1 - alpha)*100
+ sigma <- sqrt((1/OboleliPusaci)+(1/NisuOboleliPusaci)+
+ (1/OboleliNepusaci)+(1/NisuOboleliNepusaci))
+ z <- qnorm(1-(alpha/2))
+ lowervalue <- oddsRatio * exp(-z * sigma)
+ uppervalue <- oddsRatio * exp( z * sigma)
+ if (quiet == FALSE)
+ print(paste("category =", rowname, ", ", confidenceLevel,
+ "% confidence interval=[",lowervalue,",",uppervalue,"]"))
+ if (quiet == TRUE numrow == 2)
+ return(oddsRatio)
> calcOddsRatio(mymatrix,alpha=0.05)
(1) "category = Pusaci , odds ratio = 2.97377289377289"
(1) "category = Pusaci , 95% confidence interval = [ 1.78673703018007 ,
4.9494274055803 ]"
```

Odnos šansi je parametar od interesa u logističkoj regresiji između ostalog i zbog jednostavne interpretacije. Primetimo još i da je analitička analiza jednostavnija nego softverska u smislu vremena koje ćemo utrošiti.

1.2.8 Multidimenzionalno skaliranje

Multidimenzionalno skaliranje, (eng. multidimensional scaling - MDS) grupa je metoda za procenu koordinata seta objekata iz podataka o udaljenosti između parova objekata (Manly, 1986). Vrlo često za ovu metodu autori koriste naziv analiza glavnih koordinata - Principal Coordinate Analysis (Digby i Kempton, 1987). Različite su metode računanja udaljenosti kao i funkcija koje određuju odnos između tih udaljenosti i stvarnih podataka. Ulazni podaci mogu dakle, biti različite matrice udaljenosti, a rezultat je "mapa" odnosa između njih. "Mapa" može biti u jednoj dimenziji (ako objekti leže na pravoj), u dve dimenzije (ako objekti leže u ravni), u tri dimenzije (ako su objekti tačke u prostoru) ili u većem broju dimenzija (u kom slučaju više nije moguć neposredan grafički prikaz). Multidimenzionalno skaliranje je metoda poznata kao perceptualno mapiranje, tj. metoda koja pomaže analitičaru u određivanju relativnog odnosa između objekata nekog seta u prostoru. Upotrebljava se kao oblik pregrupisanja objekata na način koji može najbolje aproksimirati uočene udaljenosti. Mapu čini prostorna konfiguracija definisana vrednostima i brojem dimenzija koje se dobiju iterativnim postupcima. Formiranje vrednosti dimenzija rezultat je algoritma minimizacije funkcije, koji testira različite modele sa ciljem maksimizacije mere valjanosti (eng. goodness-of-fit) (ili minimizacije lackof-fit tj. badness-of-fit). Kruskalov *stress* merilo je valjanosti podudaranja, kojim se vrednuje koliko dobro određeni model predstavlja (ili koliko se dobro uklapa) matricu rastojanja (videti [12]) Definisan je jednačinom:

$$stress = \sqrt{\frac{\sum\limits_{i<j} (d_{ij} - \hat{d_{ij}})^2}{\sum\limits_{i<j} d_{ij}^2}} \tag{1.2}$$

Vrednost $stress - a$ se smanjuje približavanjem procenjenog rastojanja $\hat{d_{ij}}$ pravom (poznatom) rastojanju d_{ij}. *Stress* je najmanji kada se udaljenosti između objekata na "mapi" najbolje poklapaju sa poznatim udaljenostima. Kruskal je između ostalog definisao i interpretaciju dobijene vrednosti $stress - a$ kao meru valjanosti podudaranja, (tabela 1.3).

stress	mera valjanosti podudaranja
>0.20	slaba
0.10	dovoljna
0.05	dobra
0.01	odlična
0.00	perfektna

Tabela 1.3: Mera valjanosti podudaranja

Klasični MDS algoritam počiva na koordinatama matrice $\mathbf{X}$ koja se izvodi preko EVD (eigenvalue decomposition) gde je matrica $\mathbf{B} = \mathbf{X}\mathbf{X}'$. Matrica $\mathbf{B}$ se konstruiše tako što se elementi matrice poznatih rastojanja $\mathbf{P}$ kvadriraju, matrica $\mathbf{P}$ pomnoži

matricom $\mathbf{J} = \mathbf{I} - \mathbf{n^{-1}A'}$, gde je $\mathbf{I}$ jedinična matrica i $\mathbf{A}$ matrica čije su sve komponente jedinice. Ova procedura se naziva dvostruko centriranje ([14]). Naredni koraci nas dovode do matrice koordinata $\mathbf{X}$:

1. Formirati matricu poznatih rastojanja (lokacija u blizini) $\mathbf{P^2} = [\mathbf{p^2}]$.
2. Primeniti dvostruko centriranje: $\mathbf{B} = -\frac{\mathbf{1}}{\mathbf{2}}\mathbf{JP^2J}$, koristeći $\mathbf{J} = \mathbf{I} - \mathbf{n^{-1}A'}$, gde je n broj elemenata.
3. Pronaći najveće pozitivne karakteristične korene $\lambda_1, ..., \lambda_m$ matrice $\mathbf{B}$ koji odgovaraju karakterističnim vektorima $e_1, ..., e_m$.
4. m-dimenzionalna prostorna konfiguracija n elemenata izvodi se iz koordinata matrice $\mathbf{X} = \mathbf{E_m}\mathbf{\Lambda_m^{\frac{1}{2}}}$, gde je $\mathbf{E_m}$ matrica karakterističnih vektora matrice $\mathbf{B}$ a $\mathbf{\Lambda_m}$ dijagonalna matrica karakterističnih korena koji odgovaraju karakterističnim vektorima, resprektivno.

Kada smo pronašli koordinate matrice $\mathbf{X}$ možemo komentarisati valjanost dobijenih rezultata (*stress*) i to sa dva aspekta. Postoji:

- Metričko multidimenzionalno skaliranje i
- Nemetričko multidimenzionalno skaliranje.

Metričko multidimenzionalno skaliranje podrazumeva korišćenje Euklidske norme pri izračunavanju rastojanja na mapi.

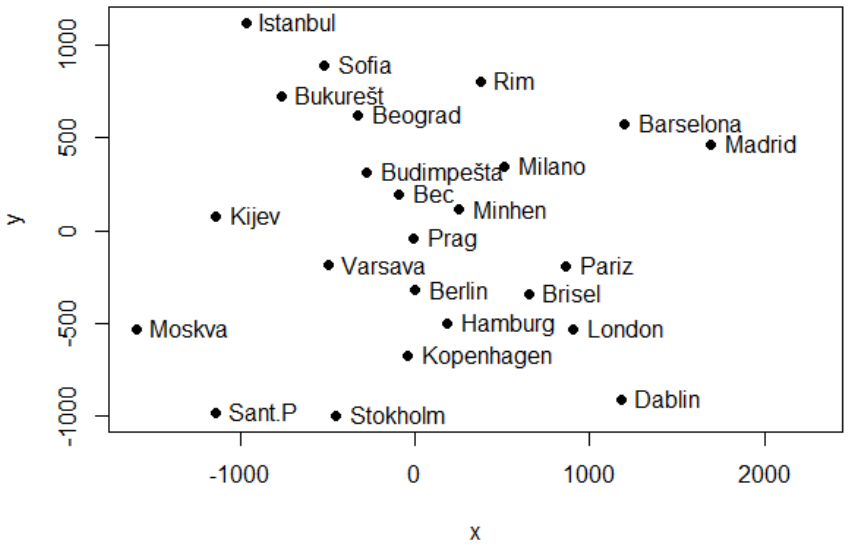
$$stress = \sqrt{\frac{\sum\limits_{i<j}(d_{ij} - ||x_i - x_j||)^2}{\sum\limits_{i<j} d_{ij}^2}} \tag{1.3}$$

Nemetričko multidimenzionalno skaliranje nalazi neparametarsku monotonu vezu između sličnosti elemenata matrice i Euklidsnog rastojanja između elemenata. Veza se dobija izotoničnom regresijom gde x predstavlja vektor poznatih rastojanja a $f_{(x)}$ monotona transformacija od x.

$$stress = \sqrt{\frac{\sum\limits_{i<j} f_{(x)-(d_{ij})}{}^2}{\sum\limits_{i<j} d_{ij}^2}} \tag{1.4}$$

Primer 1.10. Posmatrajmo sada rastojanja između 24 evropska grada (rastojanje.csv), videti [16]. Koristeći naredbu cmdscale dobićemo grafički prikaz posmatranih gradova, (slika 1.14).

```
> attach(rastojanje)
> row.names(rastojanje) <- rastojanje[, 1]
> rastojanje <- rastojanje[, -1]
> fit <- cmdscale(rastojanje, eig = TRUE, k = 2)
> x <- fit$points[, 1]
> y <- fit$points[, 2]
> plot(x, y, pch = 19, xlim = range(x) + c(0, 600))
> city.names <- c("Barselona", "Beograd", "Berlin", "Brisel", "Bukurešt",
"Budimpešta", "Kopenhagen", "Dablin", "Hamburg", "Istanbul", "Kijev",
"London", "Madrid", "Milano", "Moskva", "Minhen", "Pariz", "Prag", "Rim",
"Sant.P", "Sofia", "Stokholm", "Bec", "Varsava")
> text(x, y, pos = 4, labels = city.names)
```

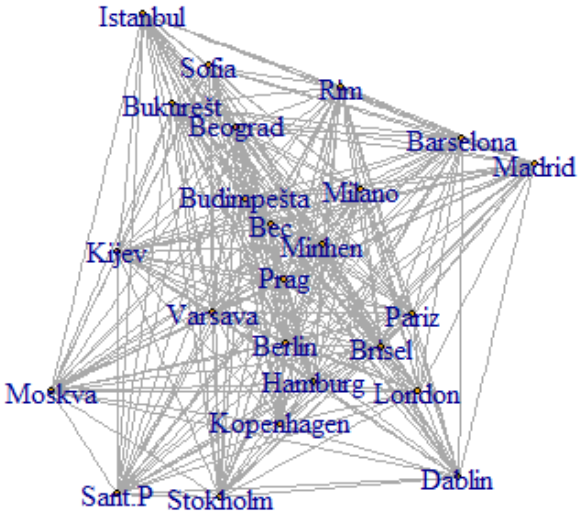


Slika 1.14: Grafički prikaz evropskih gradova dobijen postupkom multidimenzionalnog skaliranja

Mapa se dobija nešto drugačijom implementacijom i prikazana je na slici 1.15.

```
>library(igraph)
> g <- graph.full(nrow(rastojanje))
> V(g)$label <- city.names
> layout <- layout.mds(g, dist = as.matrix(rastojanje))
> plot(g, layout = layout, vertex.size = 3)
```

Na osnovu posmatranih podataka o rastojanjima postupak MDS nam je dao mapu gradova. Primetimo, dobijena mapa liči na (obrnutu) kartu Evrope jer postupak ne razlikuje pravce sever-jug. Ova primena je razlog što se navedeni postupak intenzivno koristi u geostatistici kao i u mnogim drugim disciplinama.



Slika 1.15: Mapa evropskih gradova dobijena postupkom multidimenzionalnog skaliranja

1.2.9 Dijagram stablo-list

Dijagram stablo-list (eng. stem and leaf) predstavlja grafičku reprezentaciju numeričkih podataka. Izumeo ga je Artur Bovli (Arthur Bowley), 1900. godine. Koristan je za vizuelizaciju podataka, prepoznavanje raspodele, grafičko pronalaženje medijane, modusa i autlajera. Za razliku od histograma, koji prikazuje samo interval podataka, dijagram stablo-list prikazuje sve podatke koji se nalaze u posmatranom intervalu. List čini poslednju cifru posmatranog broja dok stablo predstavljaju sve cifre broja osim poslednje (videti [6]). Pronaći ćemo stablo-list dijagram za sledeći niz brojeva: 5, 12, 23, 13, 123.
5 - stablo=0, list=5;
12 - stablo=1, list=2;
23 - stablo=2, list=3;
12 - stablo=1, list=3;
123 - stablo=12, list=3;
Dijagram formiramo tako što rezultate pišemo u dve kolone koje su odovjene vertikalnom linijom. Sa leve strane u odnosu na vertikalu zapisujemo stablo a sa desne strane list (tabela 1.4). U ovom primeru najmanji broj stabla je 0, najveći 12. Dakle, stablo će biti numerisano brojevima 0,1,2 i 12. Elemente lista zapisujemo sa desne strane, pored stabla kome pripadaju. Stablu 0 pripada broj 5, stablu 1 pripadaju

brojevi 2 i 3, stablu 2 broj 3 dok stablu 12 pripada broj 3. Kada bismo imali de-

stablo	list
0	5
1	23
2	3
12	3

Tabela 1.4: Dijagram stablo list

cimalni broj, prva cifra broja predstavlja stablo dok druga po redu predstavlja list (vrednost druge cifre zavisi od toga da li je treća cifra broja manja ili veća od 5, u slučaju da je veća vrednost lista se zaokružuje na veći broj). Napravimo dijagram za niz decimalnih i negativnih brojeva: -23.678758, -12.45, -3.4, 4.43, 5.5, 5.678, 16.87, 24.7, 56.8 (tabela1.5).

stablo	list
-2	4
-1	2
-0	3
0	4 6 6
1	7
2	5
3	
4	
5	7

Tabela 1.5: Dijagram stablo list za decimalne i negativne brojeve

Primer 1.11. Posmatrajmo broj bodova studenata osvojenih na nekom ispitu. Na osnovu ovih podataka napravićemo dijagram stablo-list (u prilogu podaci NIZBR.csv).

```
> attach(NIZBR)
> stem(niz)
```

0	17
1	238
2	577899
3	1267779
4	12234468
5	3889
6	13
7	28
8	0
9	9

Na osnovu dobijenog dijagrama zaključujemo da dati podaci imaju χ^2 raspodelu (posmatramo ih tako što ceo grafik posmatramo horizontalno, tako da je list paralelan sa x osom). Medijanu dobijamo tako što vršimo eliminaciju prvog i poslednjeg broja u listu, naizmenično. Eliminišemo 1 i 9, zatim 7 i 0, onda 2 i 8, 3 i 2 i tako redom, sve dok ne stignemo do poslednjeg, u ovom slučaju 1 i 9, kada poslednja dva spojimo sa njihovim stablima dobijemo 39 i 41, sredinu ova dva čini broj 40 koji nam predstavlja medijanu ovog uzorka. Modus je broj koji se najviše puta ponavalja u uzorku. Posmatrajući list vidimo da se sedmica ponavlja tri puta, dakle modus uzorka je 37. Primetimo da poslednji element tabele, broj 99, znatno odstupa od prikazanih podataka pa ga možemo smatrati autlajerom.

Ukoliko bismo želeli da smanjimo interval u kome posmatramo uzorak, jer na primer imamo preveliki broj podataka, to možemo učiniti sledećom naredbom:

```
> stem(niz, scale=2)
```

```
  0 | 1
  0 | 7
  1 | 23
  1 | 8
  2 |
  2 | 577899
  3 | 12
  3 | 67779
  4 | 122344
  4 | 68
  5 | 3
  5 | 889
  6 | 13
  6 |
  7 | 2
  7 | 8
  8 | 0
  8 |
  9 |
  9 | 9
```

Naredba stem je najjednostavnija naredba za prikaz stablo-list dijagrama. Hans Peter Wolf (University of Bielefeld) je izumeo nešto složenije naredbe koje su dostupne u paketu *aplpack* i koje daju više informacija o podacima u odnosu na naredbu stem.

```
> install.packages("aplpack")
> library(aplpack)
> stem.leaf(niz)
1 | 2: represents 12
```

leaf unit: 1
n: 36

2	0	17
5	1	238
11	2	577899
(7)	3	1267779
18	4	12234468
10	5	3889
6	6	13
4	7	28
2	8	0

HI: 99

Primer 1.12. Posmatrajmo broj bodova studenata osvojenih na nekom ispitu u prvom i drugom ispitnom roku (niz i niz1, podaci NIZBR.csv [7]). Na osnovu ovih podataka napravićemo uporedni dijagram stablo-list.

>stem.leaf.backback(niz, niz1)
1 | 2: represents 12, leaf unit: 1

	niz		niz1	
2	71	0	1	1
5	832	1		
11	998775	2	235	4
(7)	9777621	3	499	7
18	86443221	4	1559	11
10	9883	5	11569	16
6	31	6	13689	(5)
4	82	7	223457	15
2	0	8	15678	9
1	9	9	568	4
		10	0	1

Na osnovu dobijenih podataka vidimo da je ova naredba korisna prilikom upoređivanja dve vrste podataka. Na jednom dijagramu primećujemo razlike u bodovima u dva različita ispitna roka. Primećujemo da su studenti u drugom ispitnom roku bili bolji, raspodela ima drugačiji oblik, aritmetička sredina se pomerila na više. Primećujemo da su skoro svi studenti osvojili više od 20 bodova, osim jednog (autlajer) za kog se da zaključiti da uopšte nije učio za ispit.

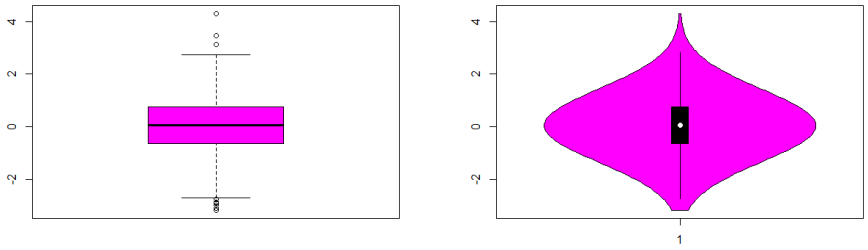
[7] Podaci iz lične arhive

1.2.10 Violina dijagram

Violina dijagram (eng. violin plot) je nastao modifikacijom originalnog Takijevog boks dijagrama (Jerry L. Hintze, Ray D.Nelson, 1998.). Dobija se kombinacijom boks dijagrama i gustine podataka. Prva razlika između violina i boks dijagrama je u tome što je sredina na boks dijagramu prikazana horizontalnom linijom a na violina dijagramu krugom. Ta promena na violina dijagramu olakšava poređenje kada imamo više grupa podataka. Druga razlika je u prikazu autlajera. Na boks dijagramu tačke izvan boksa predstavljaju autlajere dok kod violina dijagrama oni nisu istaknuti simbolima. Drugi i treći kvartil na violina dijagramu prikazani su uzanim pravougaonikom a prvi i četvrti vertikalnom linijom, isto kao i na boks dijagramu. Površina omeđena zatvorenom linijom predstavlja gustinu. Gustina je dodatak koji nam govori o distribuciji i karakteristikama podataka kao što to radi i histogram (videti [5]).

Primer 1.13. Pomenute karakteristike uočićemo posmatrajući boks i violina dijagrame za normalnu i uniformnu raspodelu (slike 1.16 i 1.17).

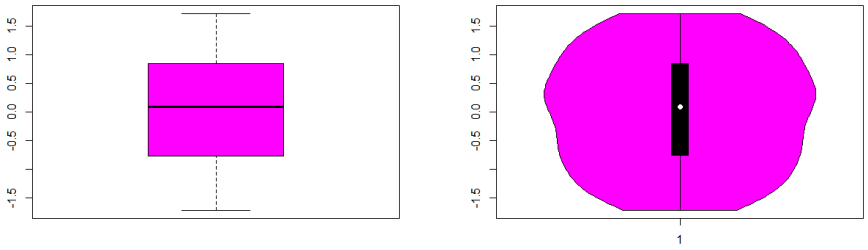
```
> install.packages("vioplot")
> library(vioplot)
> install.packages("sm")
> library(sm)
> Normal=rnorm(200, mean=0, sd=1)
> boxplot(Normal, col="magenta")
> vioplot(Normal)
```



Slika 1.16: Normalna raspodela

```
> Uniform=runif(200, min=-1.73, max=1.73)
> boxplot(Uniform, col="magenta")
> vioplot(Uniform, col="magenta")
```

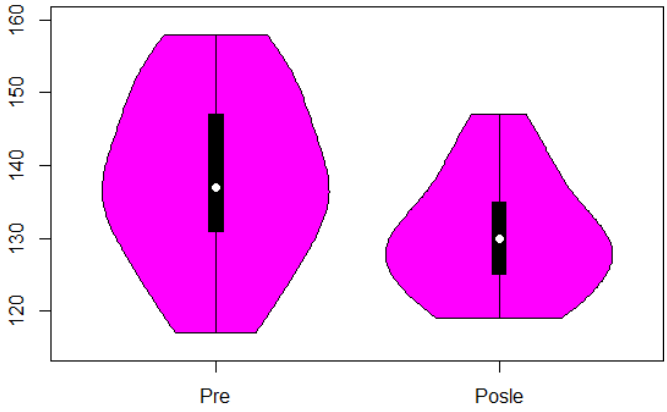
Primećujemo da je boks dijagram isti kod obe raspodele, srednja vrednost je 0, disperzija 1 (isto nam govori i violina dijagram). Violina dijagram se razlikuje zbog gustine podataka i rasporele koja se formira a koja je na grafiku prikazana simetrično u odnosu na vertikalnu osu. Dakle, violina dijagram nam daje više podataka od boks dijagrama.



Slika 1.17: Uniformna raspodela

Primer 1.14. Vršeno je ispitivanje 24 pacijenta koji pate od poremećaja sistolnog pritiska. Merene su vrednosti pritiska pre i posle terapije u trajanju od mesec dana (SPritisak.csv), (podaci iz 18). Rezultati su prikazani na slici 1.18.

```
> attach(SPritisak)
> vioplot(pre, posle, names=c("Pre","Posle"), ylim=c(115,160))
```



Slika 1.18: Violina dijagram

Na slici 1.18 vidimo da se srednja vrednost pritiska smanjila nakon terapije kao i interval u kome su se kretale vrednosti pritiska. Vrednosti podataka pre terapije formiraju grafik normalne rasporele sa velikim rasipanjem (debelim krajevima). Razlog tome je činjenica da mnogo pacijenata ima ili ekstemno visok ili ekstemno nizak pritisak. Nakon terapije, čiji cilj je bio da reguliše sistolni pritisak kod pacijenata, vidimo da podaci formiraju grafik χ^2 raspodele. Dakle, smanjila se ukupna srednja vrednost izmerenog pritiska i smanjile su se eksremno visoke vrednosti pritiska. Ekstremno niske vrednosti su sada porasle, nisu više ekstremne jer su bliže prosečnom pritisku (odstupaju za manje od 10 jedinica mere) i imamo mnogo pacijenata čiji pritisak je nešto niži od prosečnog.

1.2.11 Doterana sredina

Doterana sredina (eng. median polish) je EDA tehnika koja se koristi pri ispitivanju značajnosti faktora u jednofaktorskim i višefaktorskim modelima. Znamo da i anova (analiza varijanse) ispituje uticaj faktora u modelima ali doterana sredina je mnogo moćnija tehnika. Da bismo shvatili tehniku doterane sredine i njen značaj, moramo se prvo podsetiti anove. Naime, analiza varijanse predstavlja matematičko-statistički postupak pomoću kojeg se testira značajnost razlike između aritmetičkih sredina različitih uzoraka. Koristi se u ispitivanjima kako jedan ili više kontrolisanih faktora utiče na formiranje vrednosti posmatranog obeležja. Ova tenhika je pogodna za analiziranje podataka u kojima nemamo autlajera tj. vrednosti podataka koji značajno odstupaju od uobičajenih vrednosti za uzorke. Autlajeri su vrednsti koje obično nastaju kao poslednica nepredviđenih situacija, grešaka u uzorkovanju a ako se na vreme primete obično se isključuju iz ispitivanja. Postojanje autlajera utiče na aritmetičku sredinu posmatranih podataka a pošto anova koristi aritmetičku sredinu prilikom analize, postoji mogućnost da izvedeni zaključci neće biti tačni. Iz tog razloga, razvila se tehnika doterana sredina. Naime, doterana sredina je tehnika za ispitivanje efekata po vrstama i kolonama matrice (tabele sa podacima) i koristi medijanu umesto aritmetičke sredine. Kada koristimo ovu tehniku i u podacima nam se pojavi autlajer, on neće imati značajan uticaj na krajnji zaključak.

Pretpostavimo da posmatramo podatke koji sadrže dva kategorijalna faktora (two way functional median polish) i interesuju nas efekti koji proizvode ti faktori. Zvaćemo ih efekti po vrstama i efekti po kolonama. Ukoliko imamo samo jedan faktor onda posmatramo samo efekat vrste (one way functional median polish). Nakon obrade podataka, rezultati će biti predstavljeni na sledeći način:

$$y_{ijk} = m(x) + a_i(x) + b_j(x) + e_{ijk}(x) \tag{1.5}$$

gde je $i = 1, ..., r$, $j = 1, ..., c$, $k = 1, ...s_{ij}$ i $r \geq 2$, broj kolona, $c \geq 2$ broj vrsta i s_{ij} broj ponavljanja u ćelijama i-te vrste i j-te kolone. Ovde m predstavlja značajan funkcionalni efekat, $a_i(x)$ funkcionalni efekat vrsta dok je $b_j(x)$ funkcionalni efekat kolona. Ograničenja iterativnog postupka su da je $medijana_i\{a_i\} = 0$, $medijana_j\{b_j\} = 0$ i $medijana_i\{e_{ijk}\} = medijana_j\{e_{ijk}\} = 0$ za svako k, (videti [3]) Da bismo pronašli date efekte pratimo sledeći algoritam:

- Pronađemo medijanu uzorka (svih podataka).
- Pronađemo medijane za svaku vrstu i zapišemo ih pored vrsta kojima pripadaju. Zatim od vrednosti podataka po vrstama oduzmemo vrednosti dobijenih medijana.
- Pronađemo medijane za svaku kolonu i zapišemo pored kolona kojima pripadaju. Od vrednosti podataka iz kolona oduzmemo vrednosti dobijenih medijana.
- Postupke pronalaska medijana po vrstama i kolonama ponavljamo sve dok bar jedna vrednost medijana vrsta i kolona ne bude nula.

Da bi nam postupak bio jasniji potkrepićemo ga primerom.

Primer 1.15. Meri se vrednost niacina (vitamina B) u hlebu u laboratorijama A, B i C (videti [23]). Prilikom pravljenja hleba dodaju se doze od 0 mg niacina na 100 grama smese, 2 mg/100g i 4 mg/100g (dodatak doprinosi svežini i kvalitetu hleba). Pošto svako brašno, samo po sebi sadrži niacin ali nam nije poznato u kojim dozama, potrebno je da izmerimo krajnje vrednosti niacina u hlebu jer je prevelika vrednost štetna po zdravlje. Izvršena su merenja i odabran je uzorak prikazan u tabeli 1.6. Na ovom primeru prikazaćemo iterativni postupak kojim se dobijaju pomenuti efekti vrsta, kolona kao i greške (reziduali).

	Laboratorije		
Niacin	A	B	C
0	3.6	3.8	3.9
2	5.3	5.6	5.5
4	6.8	7.6	7.3

Tabela 1.6: Uzorak

Za početak tražimo medijanu uzorka. Sortiraćemo uzorak u rastući niz:
3.6, 3.8, 3.9, 5.3, 5.5, 5.6, 6.8, 7.3, 7.6
S obzirom da imamo neparan broj elemenata u uzorku, vrednost medijane nam je vrednost u sredini uzorka tj 5.5mg/100g. U narednom koraku, pravimo novu tabelu (tabela 1.7) sa vrednostima uzorka uzimajući u obzir i medijanu koju pišemo u gornji levi ugao. Sada, tražimo medijane po vrstama (ne uzimajući u obzir prvu

5.5	0	0	0
0	3.6	3.8	3.9
0	5.3	5.6	5.5
0	6.8	7.6	7.3

Tabela 1.7: Izračunavanje medijane uzorka

kolonu) i zapisujemo ih sa desne strane (tabela 1.8), pored vrsta kojima pripadaju. Dobijene vrednosti medijane dodajemo prvoj koloni. Nakon toga pravimo razliku

5.5	0	0	0	**0**
0	3.6	3.8	3.9	**3.8**
0	5.3	5.6	5.5	**5.5**
0	6.8	7.6	7.3	**7.3**

Tabela 1.8: Izračunavanje medijane po vrstama

svih elemenata iz vrste i njihove medijane i to nam zapravo daje reziduale. Govori

nam koliko se izmerene vrednosti niacina razlikuju od medijane uzorka. Element druge vrste i druge kolone dobijamo tako što od vrednosti 3.6 oduzmemo vrednost medijane, dakle 3.6-3.8=0.2. Druga vrsta, treća kolona iznosi 3.8-3.8=0, ostale vrednosti dobijamo analogno i prikazane su u tabeli 1.9. Sada želimo uticaje kolona,

5.5	0	0	0
3.8	-0.2	0	0.1
5.5	-0.2	0.1	0
7.3	-0.5	0.3	0

Tabela 1.9: Izračunavanje reziduala

tj. kakav uticaj na rezultate ima laboratorija u kojoj se vrše merenja. Na sličan način, tražimo medijane ali sada po kolonama tabele, izuzimajući sada prvu vrstu. Rezultati su prikazani u tabeli 1.10. Dobijene rezultate iz poslednje prepisujemo

5.5	0	0	0
3.8	-0.2	0	0.1
5.5	-0.2	0.1	0
7.3	-0.5	0.3	0
5.5	**-0.2**	**0.1**	**0**

Tabela 1.10: Izračunavanje medijane po kolonama

u prvu vrstu a od ostalih vrsta ih oduzimamo praveći tablicu reziduala. Analogno prethodnom postupku za uticaj vrsta dobijamo rezultate (tabela 1.11).

5.5	**-0.2**	**0.1**	**0**
-1.7	0	-0.1	0.1
0	0	0	0
1.8	-0.3	0.2	0

Tabela 1.11: Izračunavanje reziduala

U tabeli 1.11 vidimo efekte vrsta, kolona i reziduale.

Efekti vrsta: (-1.7, 0. 1.8)

Efekti kolona: (-0.2, 0.1, 0)

0	-0.1	0.1
0	0	0
-0.3	0.2	0

Tabela 1.12: Matrica reziduala

Program ima ugrađenu funkciju i daje nam iste rezultate. Koristimo podatke niacin.r.csv.

```
>niacin.r=as.data.frame(niacin.r)
>names(data)=c("niacin", "lab", "level")
>a=aggregate(niacin ~ lab+level, data=data, median)
>m=matrix(a[,3], nrow=3, ncol=3, byrow=T)
>m
```

	(,1)	(,2)	(,3)
(,1)	3.6	3.8	3.9
(,2)	5.3	5.6	5.5
(,3)	6.8	7.6	7.3

```
>medpolish(m)
1 : 7
Final: 7
Median Polish Results (Dataset: "m")
Overall: 5.5
Row Effects:
(1) -1.7, 0, 1.8
Column Effects:
(1) -0.2, 0.1, 0
Residuals:
```

	(,1)	(,2)	(,3)
(1,)	0	-0.1	1
(2,)	0	0	0
(3,)	-0.3	0.2	0

Ako u identitet (1.5) zamenimo vrednosti efekata iz tabele 1.12 dobijamo baš početnu vrednost iz tabele 1.6.
$5.5 - 1.7 - 0.2 + 0 = 3.6$
$5.5 + 0 + 0.1 + 0 = 5.6$

Dolazimo do zaključka da niacina u proseku imamo 5.5mg/100g. Za merenje koje je izvršeno na uzorcima u koje prvobitno nismo dodavali niacin imamo 5.5-1.7= 3.6 mg/100g tj. vrednost ispod proseka. Za uzorke u kojima smo prvobitno dodavali 2mg/100g imamo prosečnu vrednost dok za uzorke u kojima je upotrebljeno 4mg/100g imamo vrednost iznad proseka. Kada analiziramo laboratorije u kojima se vrše ispitivanja, na osnovu efekata po kolonama vidimo da laboratorija B daje najveće vredsnoti a laboratorija A najmanje. Da bismo postupak doterane sredine

uporedili sa anova postupkom potrebno je da izračunamo koliko procenata varijacija je objašnjeno datim podacima. Za ovu analizu koristićemo veličnu R^* koja se dobija iz formule (1.6) .

$$R^* = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}|y_{ij} - m(x)| - \sum_{i=1}^{r}\sum_{j=1}^{c}|e_{ij}|}{\sum_{i=1}^{r}\sum_{j=1}^{c}|y_{ij} - m(x)|} \tag{1.6}$$

Gde $\sum_{i=1}^{r}\sum_{j=1}^{c}|y_{ij} - m(x)|$ predstavlja sumu apsolutnih vrednosti odstupanja vrednosti podataka od zajedničke sredine (podrazumeva zbir odstupanja nastalih pod dejstvom faktora-između nivoa i odstupanja nastala pod dejstvom slučajne greške-unutar nivoa).$\sum_{i=1}^{r}\sum_{j=1}^{c}|e_{ij}|$ predstavlja sumu apsolutnih vrednosti odstupanja nastalih dejstvom slučajnih grešaka (unutar nivoa, unutar tretmana) tj. sumu apsolutnih vrednosti reziduala. R^* nam govori o odnosu odstupanja nastalih pod dejstvom faktora (u nivou) i odstupanju svih podataka od zajedničke sredine. Što je R^* bliže 1 to je slučajna greška (nastala unutar nivoa) manja, što nam je i cilj.

U prethodnom primeru $R^* = 0.9346$.

Ako prethodni postupak anologno ponovimo tako što umesto medijane koristimo aritmetičku sredinu dobićemo postupak dvofaktorske analize varijanse (two-way anova). Impelementiraćemo ga na istom uzorku (tabela 1.6). Rezultati su prikazani u tabeli 1.13.

5.49	**-0.26**	**0.18**	**0.08**
-1.72	0.08	-0.15	0.05
0.02	0.09	-0.05	-0.05
1.74	-0.17	0.19	-0.01

Tabela 1.13: Rezultati dobijeni korišćenjem aritmetičke sredine

Koristeći identitet (1.6) i prethodne rezultatate dobijamo $R^* = 0.922$. U prvom slučaju kada smo koristili medijanu dobili smo da je $R^* = 0.9345$. Vidimo da je u slučaju kada smo koristili medijanu slučajna greška nastala unutar nivoa manja tj. R^* je bliže jedinici a procenat varijacije objašnjene uticajem faktora je veći. Na osnovu toga zaključujemo da je korišćenje postupka doterane sredine tačnije od anova postupka.
Uporedimo sada efekte vrsta i efekte kolona (tabela 1.14). Prvo ćemo posmatrati efekte kada ne dodajemo niacin i kada dodamo 2mg/100g, zatim efekat kada ne dodamo niacin i dodamo 4mg/100g i na kraju razliku efekta kada dodamo 2mg/100g i 4mg/100g.

Na osnovu dobijenih rezultata vidimo da se vrednosti dobijene postupcima doterana sredina i anova ne razlikuju mnogo. Obe govore da je najveća razlika u rezultatima baš u slučaju kada ne dodajemo niacin i kada dodamo 4mg/100g. Preostale dve kombinacije ne prave razlike u efektima. Kada gledamo pojedinačne vrednosti i uporedimo postupke vidimo da doterana sredina u dva od tri slučaja daje veće

	2-0	4-0	4-2
median polish	1.7-0=1.7	3.5	1.8
anova	1.74	3.46	1.72

Tabela 1.14: Upoređivanje efekata prema količini niacina

vrednosti. S obzirom da u podacima sa kojima smo radili nije bilo autlajera, očekivalo se da će oba postupka dati slične vrednosti tj. da će krajnji zaključak biti isti. Statistički značajna razlika se vidi u slučaju kada ne dodamo niacin u hleb i kada dodamo 4mg/100g.
Posmatrajmo sada efekte kolona. Uzmimo razlike između laboratorija A i C, zatim razlike u merenju između laboratorija A i B i na kraju razlike između rezultata dobijenih u B i C laboratorijama (tabela 1.15).

	C-A	B-A	C-B
doterana sredina	0.2-0=0.2	0.3	-0.1
anova	0.34	0.44	-0.1

Tabela 1.15: Upoređivanje efekata prema laboratoriji

Ovo je bila analitička implementacija s ciljem da shvatimo kako postupak funkcioniše. Program nam olakšava postupak obrade i u narednim koracima ćemo prikazati implementaciju ovog postupka u programskom paketu R. Dobijamo iste rezultate.

```
> lab=c(3.6, 5.3, 6.8, 3.8, 5.6, 7.6, 3.9, 5.5, 7.3)
> niacin=c(3.6, 5.3, 6.8, 3.8, 5.6, 7.6, 3.9, 5.5, 7.3)
> lab=c(rep("A",3), rep("B",3), rep("C",3))
> class(lab)
(1) "character"
> lab<- factor(lab, levels=c("A", "B", "C"))
> class(lab)
(1) "factor"
> uticaj=data.frame(niacin, lab)
> rezultati=aov(niacin ~ lab, data=uticaj)
> summary(rezultati)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lab	2	0.309	0.1544	0.051	0.951
Residuals	6 18.140	3.0233			

```
> TukeyHSD(rezultati, conf.level=0.95)
Tukey multiple comparisons of means
95% family-wise confidence level
```

lab	diff	lwr	upr	p adj
B-A	0.4333333	-3.922704	4.789371	0.9503581
C-A	0.3333333	-4.022704	4.689371	0.9702171
C-B	-0.1000000	-4.456037	4.256037	0.9972696

Fit: aov(formula = niacin lab, data = uticaj)
> niacin=c(3.6, 3.8, 3.9, 5.3, 5.6, 5.5, 6.8, 7.6, 7.3)
> level=c(rep("0",3), rep("2", 3), rep("4",3))
> class(level)
(1) "character"
> level<- factor(level, levels=c("0","2","4"))
> class(level)
(1) "factor"
> uticajL=data.frame(niacin,level)
> rezL=aov(niacin ∼ level, data=uticajL)
> summary(rezL)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
level	2	18.03	9.014	128.8	1.18e-05
Residuals	6	0.42	0.070		

Residuals
—
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(rezL, conf.level=0.95)
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = niacin level, data = uticajL)

level	diff	lwr	upr	p adj
2-0	1.700000	1.037177	2.362823	0.0005460
4-0	3.466667	2.803844	4.129490	0.0000087
4-2	1.766667	1.103844	2.429490	0.0004415

Primer 1.16. Posmatrajmo podatke koji sadrže autlajer (vrednost koja iz nekog razloga odstupa od uobičajenih vrednosti podataka). Neka su dati podaci (tabela 1.16). Vidimo da je u tabeli 1.16 autlajer vrednost 0.5.

Ponovimo sada postupak traženja efekata po vrstama i kolonama postupkom doterane sredine. Rezultati su prikazani u tabeli 1.17.

	Laboratorije		
Niacin	A	B	C
0	**0.5**	3.8	3.9
2	5.3	5.6	5.5
4	6.8	7.6	7.3

Tabela 1.16: Uzorak sa autlajerom

5.5	**-0.5**	**0.1**	**0**
-1.7	-2.8	-0.1	0.1
0	0.3	0	0
1.8	0	0.2	0

Tabela 1.17: Rezultati dobijeni postupkom doterane sredine

Kada isti postupak ponovimo koristeći anova postupak dobijemo nove rezultate (tabela 1.18).

5.14	**-0.94**	**0.52**	**1.27**
-2.41	-1.29	0.55	-0.01
0.33	0.77	-0.39	-1.24
2.09	0.51	-0.15	-1.42

Tabela 1.18: Rezultati dobijeni anova postupkom

Na osnovu prethodnih rezultata vidimo da je autlajer napravio značajne promene. Napravićemo ponovo uporednu analizu rezultata. Najpre izračunajmo R^*. Koristeći identitet (1.5) i rezultate dobijene anova postupkom dobijamo $R^* = 0.564$. U prvom slučaju kada smo koristili postupak doterane sredine dobili smo da je $R^* = 0.7463$. Dakle, kada smo koristili medijanu slučajna greška nastala unutar nivoa je manja tj. R^* je bliže jedinici pa je procenat varijacije objašnjene uticajem faktora veći. Autlajer je napravio promene u smislu da tačnost ni u jednom od ova dva slučaja sa autlajerom nije tako dobra kao kada imamo podatke bez autlajera ali upotreba medijane daje tačnije rezultate. Na osnovu toga zaključujemo da je korišćenje postupka doterana sredina tačnije od anova postupka.

Uporedimo sada efekte vrsta i efekte kolona. Analogno prethodnoj analizi bez autlajera dobijamo sledeće rezultate (tabela 1.19).

Na osnovu dobijenih rezultata vidimo da se vrednosti dobijene postupcima doterana sredina i anova razlikuju baš u slučajevima kada ne dodajemo niacin i kada dodajemo 2mg/100g. Kao posledica se javlja povećanje razlike u slučaju kada ne dodajemo niacin i kada dodamo 4mg/100g. U slučaju kada smo dodali 2mg/100g i 4mg/100g nema velike razlike. Zaključujemo da je autlajer izvršio uticaj i to se

	0-2	0-4	2-4
doterana sredina	-1.7	-3.5	-1.8
anova	-2.74	-4.5	-1.76

Tabela 1.19: Upoređivanje efekata prema količini niacina (sa autlajerom)

osetilo prilikom upotrebe anova postupka.
Kada uporedimo sa tabelom rezultata u kojoj nismo imali autlajer (tabela 1.14) vidimo značajan uticaj autlajera koji se manifestuje kada upotrebljavamo anova postupak. Postupak doterane sredine apsorbuje dejstvo autlajera i rezultati su isti kao i rezultati dobijeni u slučaju kada nismo imali autlajer.
Posmatrajmo sada efekte na osnovu laboratorija u kojima su vršena ispitivanja (efekte kolona). U tabeli 1.20 prikazani su rezultati.

	A-C	A-B	B-C
median polish	-0.5	-0.6	0.1
anova	-2.21	-1.46	-0.75

Tabela 1.20: Upoređivanje efekata prema laboratorijama (sa autlajerom)

U slučaju kada nismo imali autlajer (tabela 1.15) rezultati su bili nešto drugačiji. Kada posmatramo anova postupak primećujemo značajne razlike kako u poređenju sa postupkom doterane sredine tako i u poređenju uticaja koje nam daju laboratorije. Tačnije, u slučaju kada imamo autlajer naglašava se razlika između laboratorija A-C i laboratorija A-B što se i očekivalo jer je autlajer nastao prilikom merenja u laboratoriji A. Rezultati dobijeni postupkom doterane sredine su slični i kada imamo autlajer i kada posmatramo rezultate bez autlajera, zaključak je isti. Dakle i u ovom slučaju postupak doterane sredine apsorbuje uticaj autlajera.

1.2.12 Usečena sredina

Usečena sredina je mera centralne tendencije posmatranih podataka, kao i atitmetička sredina, medijana i modus. Podrazumeva ponderisani prosek medijane, prvog i poslednjeg kvartila (videti [17]). Ideju je prvi plasirao britanski statističar i ekonomista Artur Bowly. Matematički se usečena sredina (TM) zapisuje na sledeći način:

$$TM = \frac{q_l + 2m + q_u}{4} = \frac{1}{2}(m + \frac{q_l + q_u}{2}) \tag{1.7}$$

gde su q_l, q_u i m donji, gornji kvartil i medijana, respektivno[8].
Primetimo da usečena sredina podrazumeva polovinu vrednosti medijane i po četvrtinu vrednosti kvartila. Dakle, potpuno drugačija mera centralne tendencije koja uzima u obzir distribuciju podataka.

Primer 1.17. Posmatrajmo visine dve grupe studenata [9] (izraženih u cm):
I grupa: 155, 158, 161, 162, 166, 170, 171, 174 i 179
II grupa: 162, 162, 163, 165, 166, 175, 181, 186 i 192

Medijana za obe grupe iznosi $m = 166cm$. Kvartili kod prve grupe studenata iznose $q_l = 161$ i $q_u = 171$. Usečena sredina $TM = 166cm$ i poklapa se sa medijanom. Kod druge grupe studenata situacija je nešto drugačija, $q_l = 163$, $q_u = 181$. Usečena sredina $TM = 170cm$. Kod prve grupe studenata vidimo da se medijana i usečena sredina poklapaju što nam govori da imamo uravnoteženu distribuciju tj. da su većina tačaka na istoj udaljenosti od medijane, sa obe strane. Kod druge grupe studenata usečena sredina je veća od medijane i ovde je distrubicija neuravnotežena. Gornji (treći) kvartiil je udaljeniji od medijane u odnosu na prvi kvartil. Vrednost usečene sredine reflektuje sklonost podataka. Koristeći paket pracma dobijamo:

```
> install.packages("pracma")
> library(pracma)
> x<- c(155, 158, 161, 162, 166, 170, 171, 174, 179)
> trimmean(x,50)
(1) 166
> y<-c( 162, 162, 163, 165, 166, 175, 181, 186, 192)
> trimmean(y,50)
(1) 170
```

Federalna banka Dalasa (USA) koristi usečenu sredinu da bi prikazala stopu inflacije. Usečena sredina je alternativna mera stope inflacije prikazana preko izdataka za ličnu potrošnju (PCA). Naime, analizirajući vremensku seriju stope inflacije od 1977. do 2009. godine uočili su da su vrednosti bazne inflacije [10] slične vrednostima inflacije koje se dobiju upotrebom usečene sredine. Od 2009. godine pa sve do danas koriste metodu usečene sredine prilikom izračunavanja stope inflacije na mesečnom, polugodišnjem i godišnjem nivou.

[8] Kvartili i medijana prethodno definisani u pododeljku 1.2.3
[9] Podaci iz lične arhive
[10] Bazna inflacija isključuje hranu i energente prilikom izračunavanja inflacije i koristi se jer je indikator dugoročnog trenda ukupne inflacije.

Glava 2

Statistički softver R

2.1 Pojam i razvoj R-a

se može posmatrati kao implementacija S jezika koji je razvijen 1976. godine u Bell laboratorijima od strane Ricka Beckera, Johna Chambersa i Allana Wilksa. S je okruženje specijalizovano za računanje i vizulizaciju odgovora na statistička pitanja. Ros Ihaka i Robert Džentlmen, profesori Univerziteta u Oklandu (Novi Zeland), 1993. godine kreću sa razvojem novog S i S-PLUS jezika, koji ubrzo postaje popularan među statističarima širom sveta. Da li zbog imena autora (Ros, Robert), ili zbog sličnosti sa S jezikom, novi programski jezik za statističku analizu intuitivno je nazvan R. Postoje samo male razlike između R i S-PLUS-a, tako da većina kodova iz R radi i u S-PLUS i obrnuto. Sve do 1997. godine, R je razvijan od strane Ihake i Gentlmen-a (uz pomoć Martina Mehlera sa Tehnološkog instituta u Cirihu), nakon čega se formira veća grupa statističara odgovornih za njegov dalji razvoj - R Development Core Team (videti [24] i[9]).

R predstavlja integrisano programsko okruženje za upravljanje podacima (videti [10]). Ovaj softver je popularan, besplatan, programski alat - programski jezik za statističku i drugu matematičku upotrebu, računanje i grafički prikaz. R programsko okruženje između ostalog poseduje:

- Niz operatora za računanje sa poljima podataka a posebno matricama;
- Veliku integrisanu zbirku programskih paketa za analizu podataka;
- Širok izbor statističkih mogućnosti za anlizu podataka (linearnih i nelinearnih modela, klasičnih statističkih testova, analiza vremenskih serija, klasifikacija, klasteri, grafičkih tehnika za analizu podataka i sl.);
- Dobro razvijen, jednostavan i korisan programski jezik koji uključuje petlje, rekurzivne funkcije definisane od strane korisnika, te postupke za učitavanje i čuvanje podataka.

Upotreba se zasniva na učitavanju komandi koje su organizovane u različitim paketima. Broj razvijenih paketa eksponencijalno raste, prema podacima iz juna 2015.

godine registrovano je 6789 paketa dok taj broj trenutno dostiže 8000. Svi su dostupni na CRAN internet stranici (http://cran.r-project.org). Na primer, paket "ggplot2" je paket za kreiranje kako jednostavnih tako i komplikovanih grafičkih prikaza. Neki jednostavniji paketi su ugrađeni u R dok se većina njih dodaje ručno. Program R se može preuzeti besplatno sa adrese: http://cran.r-project.org/bin/windows/base/, [24].

2.2 Osnove R

Kako smo u prethodnoj sekciji dali osnovne podatke o programskom paketu R u nastavku će biti date osnovne instrukcija za njegovo korišćenje. Da bismo nesmetano mogli da se bavimo EDA tehnikama u R-u neophodno je dati određena uputstva za početnike (videti [4]). Naime, kao što je već pomenuto, paketi koji nisu ugrađeni u R, instaliraju se ručno, jednostavnim učitavanjem:

```
>install.packages("ggplot2")
>library(ggplot2)
```

Lista instaliranih paketa može se videti na installed.packages(). Lista komandi koje sadrži paket "xyz" može biti viđena sa library(help="xyz ") ili samo help(xyz). Komande available.packages(), download.packages(), update.packages() mogu pomoći pri instalaciji i ažuriranju podataka. Po ugledu na svaki program i R ima svoj "help" i naredne naredbe će nam biti od pomoći u radu:

```
> help(solve) # Prikazuje stranicu za pomoć za komandu "solve".
> ?solve # Isto kao help(solve).
> ??solve # Prikazuje listu komandi koje mogu biti povezane sa stringom "solve".
```

Prilikom unošenja određenih vrsta podataka koristićemo zapise vektora, matrica, gotove podatke iz excel tabela ili formiranjem okvira podataka.

```
> c(2,5,3,7) # poveže elemente u vektor
(1) 2 5 3 7
> seq(from=1,to=10,by=3) # kreira niz od 1 do 10 sa korakom 3
(1) 1 4 7 10
> m <- matrix( data = 1:8, nrow=4, ncol=2) # generisanje matrice brojeva od 1
do 8 u 4 vrste i 2 kolone
> m <- c(10,4,5,1,18,8,12,3,9,4,7,2,8,1,3,2,1) # ručno unošenje elemenata matrice
> dim(m)<- c(4,4) # definišemo dimenziju matrice
> m # prikaz matrice
```

	(,1)	(,2)	(,3)	(,4)
(,1)	10	18	9	1
(,2)	4	8	4	3
(,3)	5	12	7	2
(,4)	1	3	2	1

Podatke koji su formirani u excel tabelama unosimo nizom sledećih akcija: Import data sets → From text file → filename.txt (file.name.csv). Dakle, neophodno je da excel tabela ima ekstenzije .txt ili .csv. To isto postižemo i naredbom read.table ("filename.txt", header=TRUE). Okviri podataka su tipične reprezentacije skupova podataka u R-u. Okviri podataka su liste sa ograničenjem da su svi elementi vektori iste dužine. Komandom data.frame() kreiramo okvir podataka.

Pošto smo dodali podatke, naredbu attach(file.name) koristimo da ih učitamo tj. uputimo program na mesto gde treba da traži promenljive.

R programski jezik omogućava rad sa 21 raspodelom. Normalna raspodela se najčešće koristi, uglavnom zbog njene veze sa centralnom graničnom teoremom. U nastavku su prikazane osnovne naredbe za ispitivanje raspodele zadatog skupa podataka:
>rnorm(n) # generiše slučajan uzorak obima n standardizovane normalne raspodele
>mean(v) # računa uzoračku srednju vrednost podataka v
>var(v) # računa uzoračku varijansu
>sd(v) # računa standardnu devijaciju uzorka
>cov(v,w) # računa uzoračku kovarijansu
>cor(v,w) # računa uzorački koeficijent korelacije za vektore v i w
>median(v) # ispisuje vrednost medijane uzorka
>quantile(v) # ispisuje kvartile uzorka
>summary(v) # ispisuje srednju vrednost, varijansu, medijanu i kvartile uzorka

2.3 R i EDA tehnike

Pošto smo se upoznali sa EDA tehnikama, u ovom odeljku će biti detaljno opisani paketi i naredbe programskog paketa R koje su korištene za implementaciju EDA tehnika (korisno je videti [19]). Svi paketi i njihove funkcionalnosti dostupni su na internetu (videti [15]).

2.3.1 Histogram u R-u

Da bismo u programskom paketu R generisali histogram potrebno je primeniti sledeću naredbu:

>hist(x, probability = TRUE, main = " ", xlab = " ", ylab= " ", ...)
gde je:
hist # crta histogram na osnovu podataka x
probability = TRUE # naredba za crtanje histograma relativnih frekvencija
main # dodaje naziv histogramu
xlab/ylab # naziv x/y osa

Da bismo nacrtali liniju koja prati raspodelu podataka na histogramima koristimo komandu:

>lines(dencity(x)) # crta krivu liniju koja prati raspodelu podataka na histogramu

2.3.2 Dijagram rasturanja u R-u

Dijagram rasturanja u programskom paketu R se generše se naredbom:

>plot(podaci$x, podaci$y, pch=number, xlab=" ", ylab=" ", col=ifelse (podaci$kategorija==" ", "color1", "color2"))
gde je:
plot # naredba za crtanje dijagrama rasturanja
podaci$x/y # naziv podataka$podaci koji se crtaju na x/y osu
pch=number # broj koji definiše oblik tačke na grafiku (pch=16 daje zvezdice)
xlab/ylab # naziv x/y osa
col=ifelse(podaci$kategorija==" " # definišemo kategoriju koja uzima referentnu vrednost)
"color1" # boja1 ako je vrednost kategorijalne promenljive referentna vrednost
"color2" # boja2 ako vrednost kategorijalne promenljive nije referentna vrednost

Ukoliko želimo da na grafik unesemo legendu koja nam daje nešto više podataka o samom grafiku koristićemo naredbu:

>legend(3,14, pch=c(16,16), col=c(”color1”, ”color2”), c(”Nepusaci”, ”Pusaci”), bty=”o”, box.col=”black”, cex=.8)
gdje je:
legend # naredba za crtanje legende na grafiku
3,14 # položaj legende na grafiku
pch=c(number, number) # broj koji definiše oblik tačke na grafiku za obe kategorije (pch=16 daje zvezdice)
color=c(”color1”, ”color2”) # boje koje govore o vrednostima kategorijalne promenljive na grafuku
c(”category1”, ”category2”) # kategorije koje odgovaraju bojama, respektivno
bty=”o” # tip boksa legende
box.col=”black” # boja ivica boksa legende
cex=.8 # veličina slova

Za upoređivanje dve promenljive koristimo naredbu:

>par(mfrow=c(1,2)

Za upoređivanje tri promenljive koristimo naredbu:

>plot(x[,c(1,2,3)] # naredba za upoređivanje dve ili više promenljivih, gde x predstavlja podatke a brojevi 1, 2 i 3 redne brojeve kolona tabele čije podatke crtamo.

2.3.3 Boks dijagram u R-u

Da bismo u programskom paketu R nacrtali boks dijagram podataka x koristimo naredbu:

>boxplot(x)

U slučaju da želimo podatke o kvartilima, minumumu, maksimumu, medijani i aritmetičkoj sredini podataka x koristimo naredbu:

> summary(x)

Ukoliko želimo boks dijagram sa određenim dodacima i pod određenim uslovima, koristimo naredbu:

> boxplot(x ∼ y, data=DATA, xlab=" ", ylab=" ", names=c("category1", "category2"), las=number)
gde je:
x ∼ y # označavanje zavisne i nezavisne promenljive, gde nezavisna promenljiva mora biti kategorijalna
data=DATA # DATA predstavlja naziv podataka
xlab/ylab # naziv x/y osa
names=c("category1","category2") # naziv kategorija nezavisne promenljive
las=number # numerička vrednost koja određuje način upisivanja teksta na grafiku (0=uobičajeno, 1=horizontalno, 2=vertikalno u odnosu na osu i 3=vertikalno)

2.3.4 Dijagram protoka u vremenu u R-u

Da bismo nacrtali dijagram protoka u vremenu u programskom paketu R najpre treba da instaliramo i učitamo sledeći paket:

>install.packages("qicharts")

>library(qicharts) # učitavanje paketa

Pre upotrebe naredbe za crtanje dijagrama, u narednim koracima formiraćemo vrednosti na osama i učitaćemo potrebne podatke.

>y <- data # na y osu se učitavaju podaci

>x <- seq.Date(as.Date('date'), by=day, length=31) # formiranje vremenske ose na x osi, počev od datuma 'date'
by=day # korak "dan" ("week", "month" ili "year")
length=31 # dužina vremenske ose sa unapred definisanim korakom

Tek sada možemo upotrebiti naredbu za crtanje dijagrama protoka u vremenu. Ona se formira na sledeći način:

> qic(y, x=x, ylab=" ", xlab=" ", main=" ") # naredba za crtanje dijagrama protoka u vremenu
gde :
main # dodaje naziv dijagrama
xlab/ylab # dodaje naziv x/y osa

2.3.5 Pareto dijagram u R-u

Pre nego što upotrebimo naredbu za crtanje pareto dijagrama potrebno je da unesemo podatke. U ovom primeru unosili smo tako što formiramo vektor podataka i dajemo imena podacima.

> x <- c( ) # formiramo vektor podataka

> names(x) <- c(" ", " ", ...) # dajemo imena podacima

> X <- data.frame(x) # učitavamo podatke

Paket koji je neophodan za crtanje pareto dijagrama je paket qcc.

> install.packages("qcc") # instaliramo paket za crtanje pareto dijagrama

> library(qcc) # učitavamo paket

U učitanom paketu imamo naredbu za crtanje pareto dijagrama:

>pareto.chart(x, main=" ", xlab=" ", ylab=" ", cex.names=number, las=number, col=topo.colors(6))
gde:
pareto.chart(x) # naredba za crtanje pareto dijagrama podataka x
main # dodaje naziv dijagrama
xlab/ylab # naziv x/y osa
cex.names=number # faktor koji definiše veličinu naziva osa
las=number # numerička vrednost koja određuje način upisivanja teksta na grafiku (0=uobičajeno, 1=horizontalno, 2=vertikalno u odnosu na osu i 3=vertikalno)
col=topo.colors(n) # naredba za kreiranje n različitih boja

Na kraju, ukoliko na grafiku želimo liniju koja odvaja 80% značajnih faktora od preostalih 20% beznačajnih korisitćemo naredbu:

> abline(h=(sum(defect.counts)*.8),col="red",lwd=4)

2.3.6 Dijagram paralelnih koordinata u R-u

Da bismo nacrtali dijagram paralelnih koordinata potrebno je najpre da instaliramo odgovarajući paket.

> install.packages("parcor") # instaliranje paketa za crtanje dijagrama paralelnih koordinata

> library(parcor) # učitavanje paketa

> p<-data.frame(x) # učitavanje podataka

Sada smo u mogućnosti da upotrebimo naredbu za crtanje dijagrama:

> parcoord(p, col=4, lty=1, var.label=TRUE)
parcoord(p) # naredba za crtanje dijagrama paralelnih koordinata
col=number # broj koji definiše boju linija na dijagramu
lty=number # broj koji definiše tip linije na dijagramu (puna, isprekidana i sl.)
var.label=TRUE # ako je vrednost TRUE, svaka osa je označena maksimalnom i minimalnom vrednošću promenljive

Drugi način crtanja dijagrama paralelnih koordinata postižemo naredbom iz paketa "GGally".

> install.packages("GGally") # instaliranje paketa za crtanje dijagrama paralelnih koordinata

> library(GGally) # učitavanje paketa

> attach(y) # učitavanje podataka

Koristeći dati paket upotrebićemo drugu naredbu za crtanje dijagrama paralelnih koordinata.

> ggparcoord(y, columns=2:5, groupColumn = 1, showPoints = TRUE, scale= "globalminmax")
gde su:
ggparcoord # naredba za crtanje dijagrama paralelnih koordinata
columns=m:n # m:n definiše kolone podataka koje se uzimaju u obzir, od kolone m do kolone n
groupColumn=1 # naredba za grupisanje svake kolone pojedinačno
showPoints=TRUE # prikazuje tačke u kojima se seku koordinate i linije koje povezuju podatke
scale="globalminmax" # skala na dijagramu je definisana u odnosu na vrednosti globalnog minimuma i maksimuma podataka

2.3.7 OR količnik u R-u

Da bismo iznačunali vrednost odnosa šansi potrebno je da prvo striktno definišemo podatke. Neophodno je da se definišu nazivi vrsta i kolona kako bismo mogli da ih upotrebimo u narednim koracima.

> mymatrix <- matrix(c(688,650,21,59),nrow=2,byrow=TRUE) # definisanje elemenata matrice

> colnames(mymatrix) <- c("Oboleli","NisuOboleli") # definisanje naziva kolona

> rownames(mymatrix) <- c("Pusaci","Nepusaci") # definisanje naziva vrsta

> print(mymatrix) # prikaz matrice

Sada ćemo definisati sve što je neophodno da se izračuna odnos šansi:

> calcOddsRatio <- function(mymatrix,alpha=0.05, # učitavamo matricu i nivo značajnosti od 5%
+ referencerow=2,quiet=FALSE) # označavamo referentnu vrednost (u odnosu na koju posmatramo ishod)
+ numrow <- nrow(mymatrix) # broj vrsta
+ myrownames <- rownames(mymatrix) # broj kolona
+ for (i in 1:numrow) # for petlja po vrstama traži referentnu vrednst
+ rowname <- myrownames[i]
+ OboleliNepusaci <- mymatrix[referencerow,1]
+ NisuOboleliNepusaci <- mymatrix[referencerow,2]
+ if (i != referencerow)
+ OboleliPusaci <- mymatrix[i,1]
+ NisuOboleliPusaci <- mymatrix[i,2]
+ totPusaci <- OboleliPusaci + NisuOboleliPusaci # sabira vrednosti referentnog ishoda, koje će kasnije da koristi za odnos šansi
+ totNepusaci<- OboleliNepusaci + NisuOboleliNepusaci
+ sansaOboleliPusaci<- OboleliPusaci/totPusaci
+ sansaOboleliNepusaci <- OboleliNepusaci/totNepusaci
+ sansaNisuOboleliPusaci <- NisuOboleliPusaci/totPusaci
+ sansaNisuOboleliNepusaci <- NisuOboleliNepusaci/totNepusaci # računa šanse da se dogode pojedinačni ishodi
+ oddsRatio <- (sansaOboleliPusaci*sansaNisuOboleliNepusaci)
+ (sansaNisuOboleliPusaci*sansaOboleliNepusaci) # formira formulu za računanje odnosa šansi
+ if (quiet == FALSE)
+ print(paste("category =", rowname, ", odds ratio = ",oddsRatio))
+ confidenceLevel <- (1 - alpha)*100 # u narednim koracima generišemo interval poverenja za vrednost odnosa šansi
+ sigma <- sqrt((1/OboleliPusaci)+(1/NisuOboleliPusaci)

```
+ (1/OboleliNepusaci)+(1/NisuOboleliNepusaci))
+ z <- qnorm(1-(alpha/2))
+ lowervalue <- oddsRatio * exp(-z * sigma)
+ uppervalue <- oddsRatio * exp( z * sigma)
+ if (quiet == FALSE)
+ print(paste("category =", rowname, ", ", confidenceLevel,
+ "% confidence interval=[",lowervalue,",",uppervalue,"]"))
+ if (quiet == TRUE numrow == 2)
+ return(oddsRatio)
```

u prethodnim koracima izvršeno je definisanje funkcije za izračunavanje OR količnika za zadatu matricu podataka (definisanje nivoa značajnosti, referentne vrednosti, identiteta koji računa odnos šansi i intervala poverenja)

> calcOddsRatio(mymatrix,alpha=0.05) # računa vrednost OR količnika na osnovu prethodno isprogramirane funkcije

2.3.8 Multidimenzionalno skaliranje u R-u

Da bismo primenili multidimenzionalno skaliranje i nacrtali mapu neophodno je da najpre učitamo podatke, zatim damo imena vrstama (kolonama) jer je to simetrična matrica.

> attach(rastojanje) # učitavanje podataka

> row.names(rastojanje) <- rastojanje[, 1] # nazivi vrsta

> rastojanje <- rastojanje[, -1] # simetrična matrica (nazivi vrsta su isti kao i nazivi kolona

Sada, na učitane podatke primenimo naredbu:

> fit <- cmdscale(rastojanje, eig = TRUE, k = 2) # definisanje funkcije fit
gde je: cmdscale # naredba za multidimenzionalno skaliranje
eig=TRUE # računa karakteristične korene
k=2 # definiše dvodimenzionalan prostor

Sada, potrebno je da definišemo koordinate na mapi, na osnovu prethodnog postupka:

> x <- fit$points[, 1] # definisanje x koordinate tačke na osnovu prethodno definisane funkcije fit

> y <- fit$points[, 2] # definisanje y koordinate tačke na osnovu prethodno definisane funkcije fit

Najzad, crtamo mapu na osnovu prethodno formiranog postupka.
> plot(x, y, pch = number, xlim = range(x) + c(0, 600)) # naredba za crtanje definisanih vrednosti
gde je:
pch=number # broj koji definiše oblik tačke na grafiku
xlim= ” ” # oblast definisanosti x ose

Sada, dajemo imena tačkama na mapi:

> point.names <- c(” ”, ” ”, ...) # definišemo imena tačaka na grafiku

> text(x, y, pos = 4, labels = point.names) # definišemo nazive tačaka u određenoj poziciji na grafiku za zadate koordinate

2.3.9 Dijagram stablo-list u R-u

Da bismo u generisali dijagram stablo list neophodno je najpre da unesemo niz brojeva:

>x=c( ) # unosimo niz brojeva

Naredba za formiranje stablo-list dijagrama zadatog niza je:

>stem(x, scale=2)
gde je:
stem # naredba za formiranje stablo-list dijagrama
scale=2 # naredba koja pravi 2 elementa stabla od jednog a elemente lista deli u interval [0,4] za prvo stablo i [5,9] za drugo stablo

Za dodatne mogućnosti instaliramo sledeći paket:

> install.packages(”aplpack”) # paket koji sadrži dodatne naredbe za stablo-list dijagram

> library(aplpack) # učitavanje paketa

Novi paket nam daje i novu naredbe:

> stem.leaf(x) # prikazuje frekvenciju podataka, najmanju, najveću vrednost, autlajer i broj elemenata u nizu

>stem.leaf.backback(x, x1) # naredba za upoređivanje 2 niza podataka

2.3.10 Violina dijagram u R-u

Da bismo u programskom paketu R generisali violina dijagram potrebno je najpre da instaliramo i učitamo odgovarajuće pakete.

> install.packages("vioplot") # paket neophodan za crtanje violina dijagrama

> library(vioplot) # učitavanje paketa

> install.packages("sm") # paket neophodan za crtanje violina dijagrama

> library(sm) # učitavanje paketa

S obzirom na činjenicu da su nam bile potrebne normalna i uniformna raspodela sledeće naredbe će ih generisati.

> Normal=rnorm(200, mean=0, sd=1) # generisanje standardizovane normalne rasporele

> boxplot(Normal, col="magenta") # naredba za crtanje boks dijagrama u određenoj boji

> vioplot(Normal) # naredba za crtanje violina dijagrama normalne raspodele

> Uniform=runif(200, min=-1.73, max=1.73) generisanje uniformne raspodele sa očekivanjem 0 i disperzijom 1, kao kod normalne raspodele

Sada ćemo nacrtati boks i violina dijagram za zadate raspodele:

> boxplot(Uniform, col="magenta")

> vioplot(Uniform) # naredba za crtanje violina dijagrama uniformne raspodele

> attach(P) # učitavanje podataka sa 2 promenljive

Sada ćemo na osnovu učitanih podataka sa dve promenljive nacrtati violina dijagram:

> vioplot(x, y, names=c(" "," "), ylim=c(,))
gde je:
vioplot(x,y) # naredba za crtanje 2 violina dijagrama koji odgovaraju dvema promenljivima (naredba se može uopštiti za više promenljivih)
names=c(" "," ") # davanje naziva promenljivima na dijagramu
ylim=c(,) # oblast definisanosti y ose

2.3.11 Doterana sredina u R-u

U programskom paketu R doterana sredina se dobija naredbom medpolish. Međutim, pre toga potrebno je generisati podatke sledećim naredbama:

>niacin.r=as.data.frame(niacin.r) # učitavamo podatke

>names(data)=c("niacin", "lab", "level") # dajemo imena podacima (kolonama)

>a=aggregate(niacin ∼ lab+level, data=data, median) # generišemo funkciju koja će posmatrati uticaj laboratorije i količine na vrednost niacina za datu tabelu podataka i koristiti medijanu

>m=matrix(a[,3], nrow=3, ncol=3, byrow=T) # generišemo matricu sa 3 vrste i tri kolone koje će sadržati vrednost trećeg merenja u laboratoriji A kada ne dodajemo niacin, vrednost trećeg merenja za laboratoriju B kada ne dodajemo niacin, vrednost trećeg merenja za laboratoriju C kada ne dodajemo niacin i tako ponovo za svaku laboratoriju kada dodajemo niacin.

>m # prikaz dobijene matrice

>medpolish(m) # naredba za dobijanje karakterističnih rezultata doterane sredine

Sada želimo ispitati anova efekte da bismo ih mogli upotrebiti sa rezultatima doterane sredine. Generisaćemo podatke i primeniti naredbu aov

> lab=c(3.6, 5.3, 6.8, 3.8, 5.6, 7.6, 3.9, 5.5, 7.3) # sami generišemo uzorak ispitivanja

> niacin=c(3.6, 5.3, 6.8, 3.8, 5.6, 7.6, 3.9, 5.5, 7.3)

> lab=c(rep("A",3), rep("B",3), rep("C",3)) # kolone nam predstavljaju laboratorije

> class(lab) # ispitujemo klasu podataka laboratorije

> lab<- factor(lab, levels=c("A", "B", "C")) # menjamo klasu podataka u faktor

> class(lab)

> uticaj=data.frame(niacin,lab) # generišemo podatke koje ćemo posmatrati

> rezultati=aov(niacin ∼ lab, data=uticaj) # formiramo funkciju uticaja vrsta i kolona koristeći anovu

> summary(rezultati) # prikaz rezultata definisane funkcije

> TukeyHSD(rezultati, conf.level=0.95) # Takijev test sa intervalom poverenja 95%

2.3.12 Usečena sredina u R-u

Da bismo izračunali usečenu sredinu u programskom paketu R neophodno je prvo da instaliramo paket koji sadrži neophodne naredbe.

> install.packages("pracma") # paket neophodan za izračunavanje usečene sredine

> library(pracma) # učitavanje paketa

Podaci na osnovu kojih tražimo usečenu sredinu generišemo kao vektor:

> x<- c( ) # definišemo podatke kao vektor

Naredba potrebna za pronalazak usečene sredina na osnovu generisanih podataka je sledeća:

> trimmean(x,number) # naredba za usečenu sredinu podataka x
gde je:
number # predstavalja broj između 0 i 100, gdje je 50 uobičajena vrednost koja nam daje prethodno definisanu usečenu sredinu (50% podataka uključuje u izračunavanje medijane a preostalih 50% za kvartile)

Glava 3

EDA tehnike u primeni

3.1 Objedinjene EDA tehnike

Ovo poglavlje biće posvećeno implementaciji EDA tehnika. Uz adekvatno tumačenje dobijenih rezultata videćemo koliko može biti moćan kompletan aparat EDA tehnika. Jedan primer biće obrađen kroz više tehnika te ćemo videti količinu informacija koje se mogu dobiti prostim tumačenjem grafika.

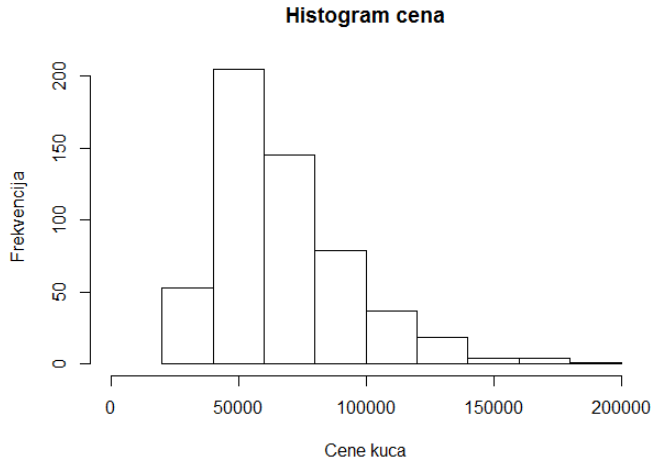
Primer 3.1. U prilogu su dati podaci cena 547 kuća (sa placevima) u Kanadi (podaci iz baze podataka [20]). Osim cena imamo podatke i o površini placeva na kojima se te kuće nalaze, broju spavaćih soba, kupatila, spratova, garaža. Zatim, da li kuća poseduje odgovarajući put, bazen, podrum, klimu, da li se kuća greje na gas i da li je na poželjnoj lokaciji.

Na osnovu datih podataka i stečenog znanja o EDA tehnikama pokušaćemo da izvučemo što više informacija o njihovoj povezanosti, mogućnostima prodaje i stanju na tržištu.

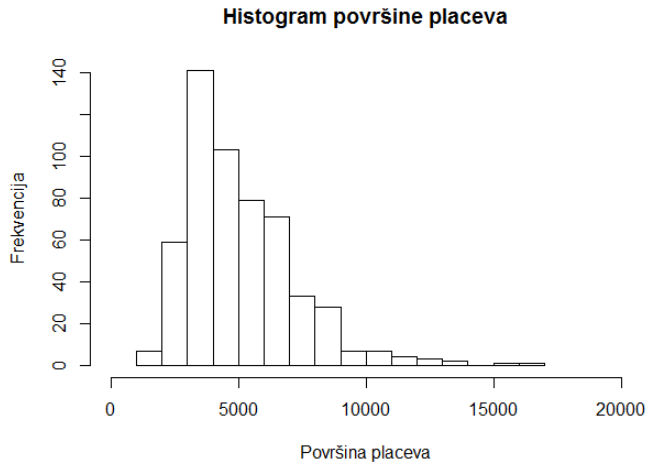
Histogram

Analizu započinjemo histogramima. Na slikama 3.1 i 3.2 su prikazana kretanja cena kuća i površina njihovih placeva. Primećujemo da na tržištu Kanade preovladavaju cene od 50000$ (200 kuća sa sličnom cenom) a kreću se u intervalu od 10000$ do 200000$ dolara.

Najčešće površine placeva su oko $3000m^2$ (140 placeva) a kreću se od $10000m^2$ do $170000m^2$.
Takođe, na oba histogrma primećujemo χ^2 raspodelu. Ona nam govori da su kuće sa manjim placevima i nižim cenama prisutnije na tržištu. Tačnije, vidimo da na tržištu ima najviše kuća sa cenom do 100000$ i površinom do $9000m^2$. Kuće sa većom cenom i površinom postoje ali su retke.



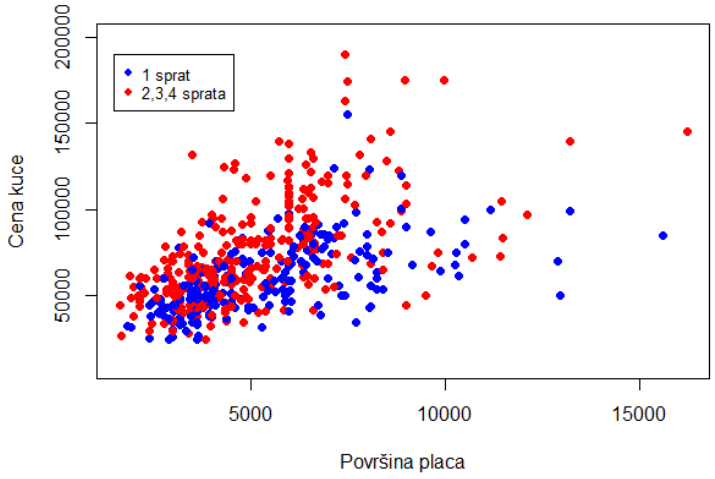
Slika 3.1: Histogram cena kuća



Slika 3.2: Histogram površine placeva

Dijagram rasturanja

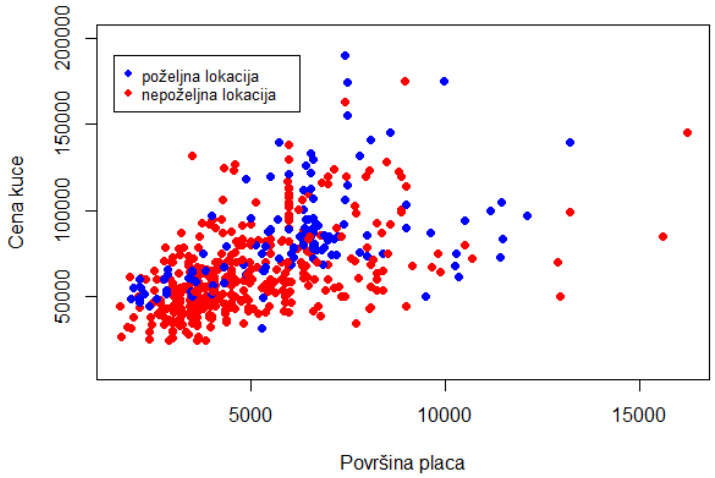
Na slici 3.3 primećujemo blagu linearnu povezanost između površine placeva i cena kuća (ako ne uzimamo u obzir broj spratova). Linearna povezanost je prisutna kod kuća koje se nalaze na manjim placevima. Kada je reč o velikim placevima tu primećujemo da kuće nemaju najviše cene, kao što se kod linearne povezanosti očekuje. Razlog tome je činjenica da veličina placa očigledno nije presudan faktor kod kupovine kuće, u narednim ispitivanjima saznaćemo i zašto. Međutim, primećujemo i veliko rasipanje podataka koje je posledica kategorijalnih faktora. Primećujemo, takođe da su najzastupljenije kuće sa površinom placa do $6000m^2$ i cenom do 100000$.



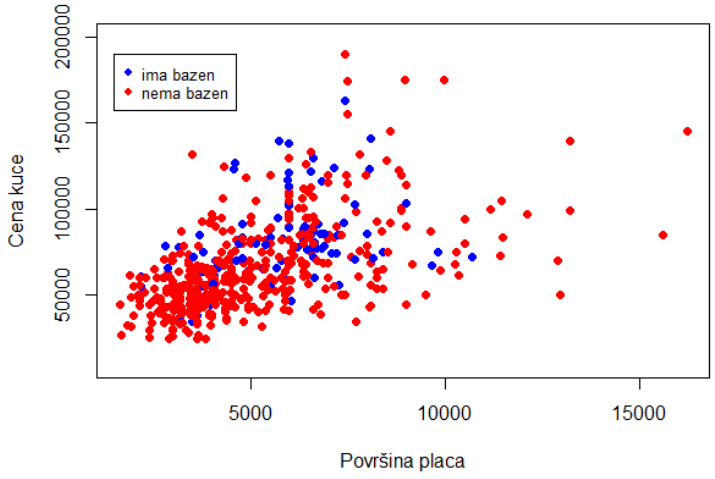
Slika 3.3: Dijagram rasturanja cene kuće u zavisnosti od površine placa i broja spratova

Sada ćemo razmotriti uticaj broja spratova, bazena i lokacije na cenu. Na slici 3.3 vidimo da kuće sa 2 i više spratova imaju i veću cenu u odnosu na jednospratnice sa istom površinom placa i tu imamo jaču linearnu povezanost, što se i očekivalo. Jednospratnice imaju nižu cenu i vidimo da se uglavnom one nalaze na velikim placevima a slika 3.4 nam govori da su to nepoželjne lokacije. Pretpostavljamo da su to poljoprivredna zemljišta koja se eksploatišu. Takođe, slika 3.4 nam govori i da kuće na poželjnijim lokacijama imaju i veću cenu i obrnuto. Višespratnice su nam zapravo i najskuplje kuće.

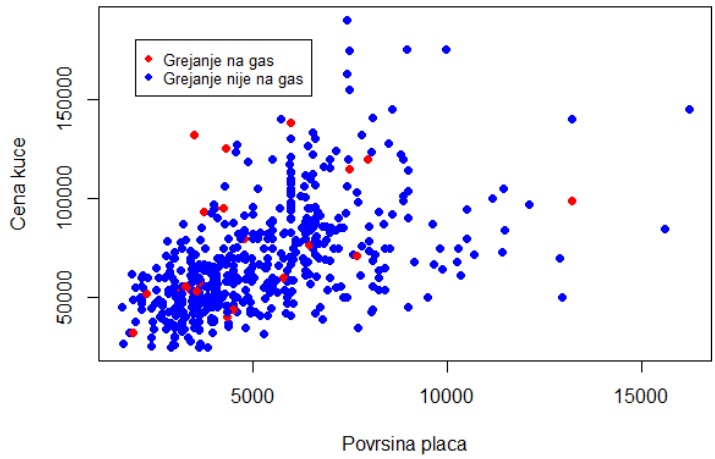
Kada je reč o kategoriji "bazen" na slici 3.5 vidimo da kuće koje poseduju bazen spadaju u grupu skupljih kuća, pa bi kupac za kuću sa bazenom morao da izdvoji preko 70000$. Takođe, da se primetiti da kuće sa bazenima nemaju prevelike placeve, najviše do $10000m^2$. Kategorija "grejanje na gas", (slika 3.6) nam govori da većina kuća ne koristi grejanje na gas kao i da cena onih koje ga koriste nije veća od onih koje ga ne koriste.



Slika 3.4: Dijagram rasturanja cene kuće u zavisnosti od površine placa i kategorije ”lokacija”



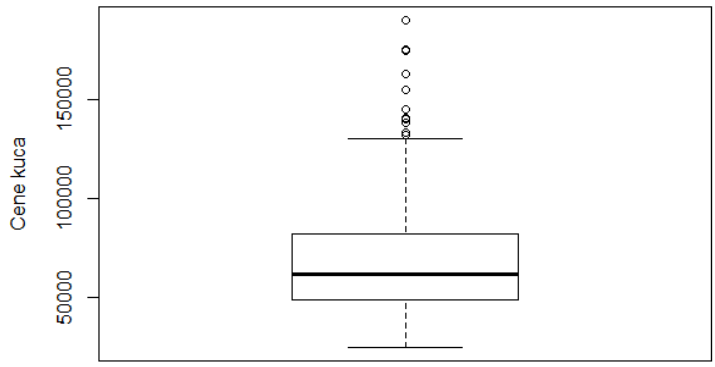
Slika 3.5: Dijagram rasturanja cene kuće u zavisnosti od površine placa i kategorije ”bazen”



Slika 3.6: Dijagram rasturanja cene kuće u zavisnosti od površine placa i kategorije "grejanje na gas"

Violina i boks dijagram

Videćemo sada kretanja cena kuća u odnosu na to koje pogodnosti ima kuća. Posmatrajmo prvo uopštene cene kuća, boks dijagram (slika 3.7). Osim intervala u kome se cene najčešće kreću vidimo da su prisutne izuzetno visoke cene pojedinih kuća koje se detektuju kao autlajeri.



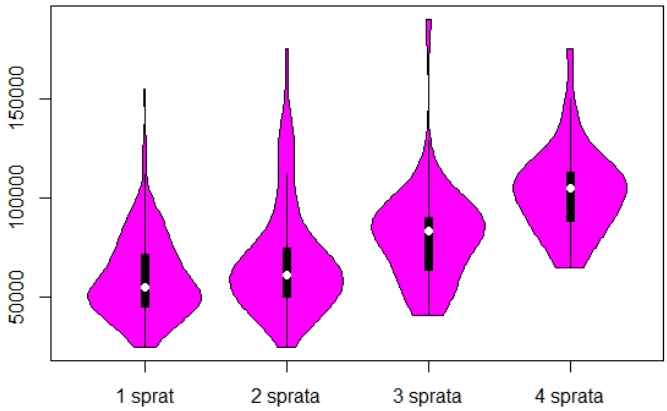
Slika 3.7: Boks dijagram koji prikazuje cene kuća

Na slici 3.8 prikazana je zavisnost cena kuća u odnosu na broj spratova. Vidimo da je najčešća cena jednospratnica 50000$ i da su one u proseku najjeftinije. Dvospratnice su nešto skuplje sa prosečnom cenom oko 60000$ dok je prosečna cena trospratnica oko 80000$. Cena trospratnica ne pada ispod 30000$ a primećujemo da pojedine dvospratnice i trospratnice dostižu i izuzetno visoke cene. Četvorospratnice se ne mogu kupiti bez 80000$ a njihova prosečna cena je 100000$.

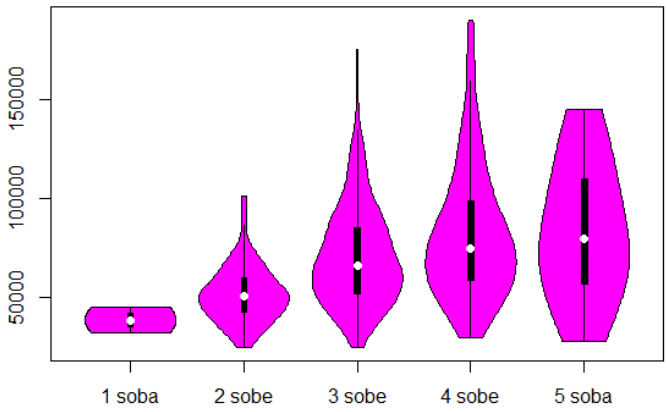
Na slici 3.9 prikazana je zavisnost cena kuća u odnosu na broj spavaćih soba. Primećujemo da su kuće sa 1 spavaćom sobom retke i izuzetno jeftine. Kuće sa 2 spavaće sobe nisu skuplje od 100000$ i većina njih košta oko 50000$. Kada su u pitanju kuće sa tri ili četiri spavaće sobe one mogu dostizati vrtoglavo visoke cene a interkvatilni raspon nam govori da većina košta između 50000$ i 100000$. Kuće sa 5 spavaćih soba takođe imaju širok interkvartilni raspon, veliku disperziju i malu gustinu što govori o velikim razlikama u ceni.

Na slici 3.10 vidimo značajan uticaj lokacije na cenu kuće. Većina kuće sa poželjnom lokacijom ima znatno više cene od kuća koje su na nepoželjnoj lokaciji.

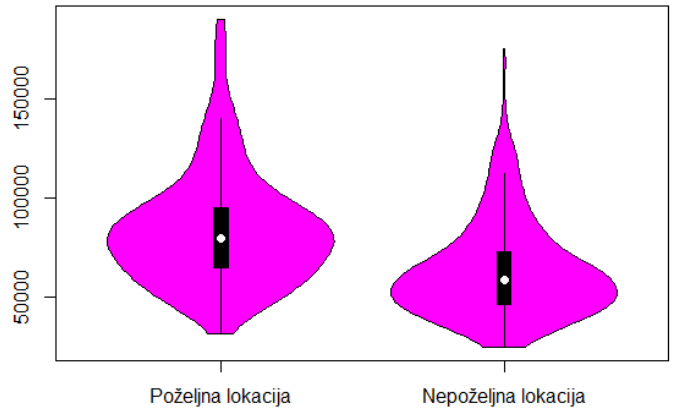
Slika 3.11 govori nam da je put potreban uslov za visoku cenu kuće. Kuće za koje ne postoji put do njih su izuzetno jeftine i retke.



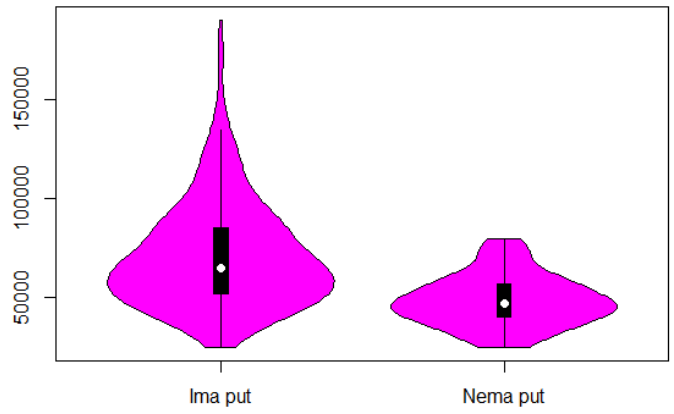
Slika 3.8: Violina dijagram cene u odnosu na broj spratova



Slika 3.9: Violina dijagram cene u odnosu na broj soba



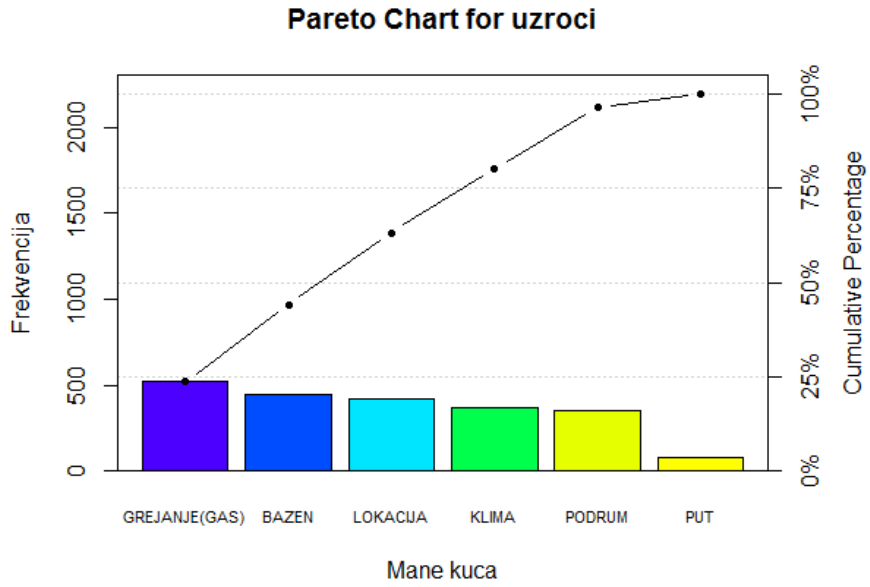
Slika 3.10: Violina dijagram cene u odnosu na lokaciju



Slika 3.11: Violina dijagram cene u odnosu na kategoriju ”put”

Pareto dijagram

Za formiranje pareto dijagrama koristićemo kategorijalne promenljive, uzećemo broj kuća koje nemaju određene pogodnosti. Najznačajnije kategorije za formiranje visokih cena su one koje većina kuća nema. Dakle, one pogodnosti koje su retke najviše će uticati na visinu cene kuća. Na slici 3.12 prikazan je dobijeni pareto dijagram. Znamo da pareto dijagram daje 80% značajnih i 20% beznačajnih. U ovom slučaju dobijamo da grejanje na gas, bazen, klima i lokacija čine značajnu većinu koja utiče na visinu cene kuće, dok podrum i put poseduje većina kuća pa nam one čine beznačajnu manjinu od 20%. Takođe, na osnovu dijagrama rasturanja smo zaključili da grejanje na gas nije u korelaciji sa cenom i da ga većina kuća nema pa ga ne možemo posmatrati kao značajan faktor.



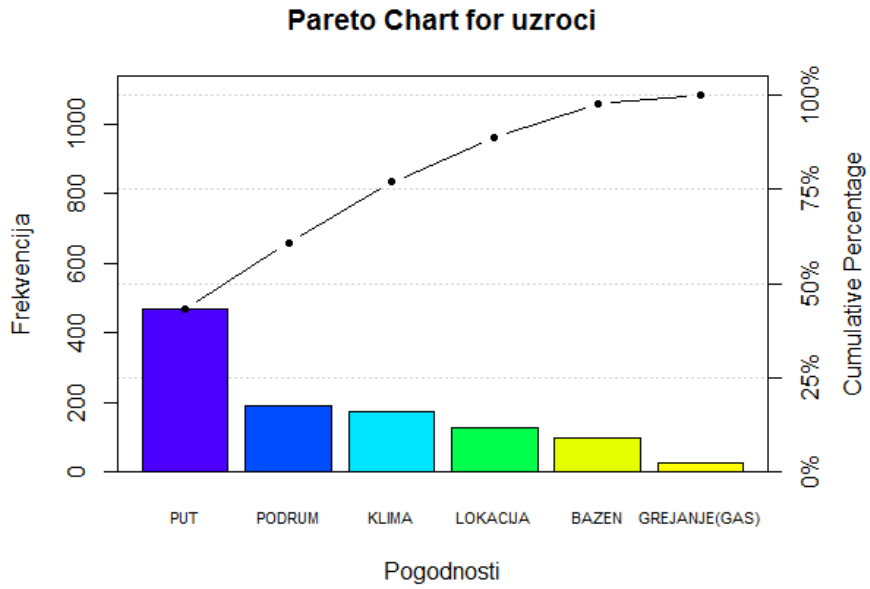
Slika 3.12: Pareto dijagram sa aspekta mana kuća

Pareto chart analiza

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
Grejanje(gas)	521	521	23.757410	23.75741
Bazen	449	970	20.474236	44.23165
Lokacija	418	1388	19.060648	63.29229
Klima	373	1761	17.008664	80.30096
Podrum	355	2116	16.187870	96.48883
Put	77	2193	3.511172	100.00000

Ako posmatramo sa drugog aspekta, (slika 3.13) i fokusiramo na ono što kuće imaju vidimo da većina kuća, 80% njih ima put i podrum i da to ima značajna većina kuća koje su na tržištu. Dakle, da bi kuća imala bilo kakvu vrednost neophodno je da na prvom mestu ima put koji vodi do nje kao i podrum. To je zapravo osnova

za formiranje viših cena koje generiše neznačajna manjina. Imali smo priliku da na violina dijagramu vidimo da kuće koje nemaju put imaju izuzetno niske cene, (slika 3.11).



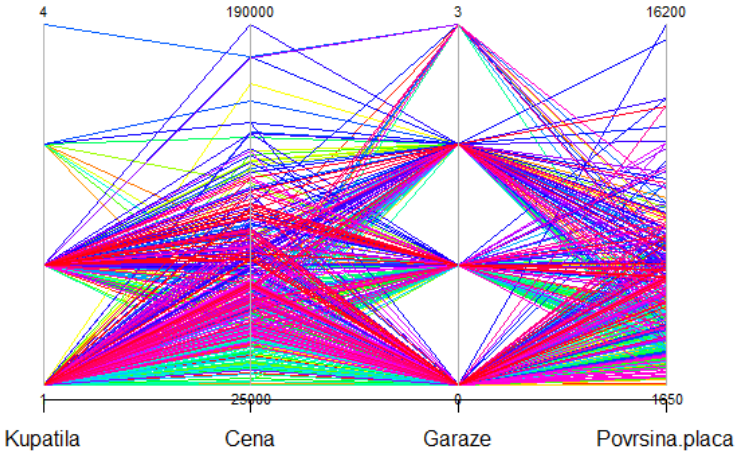
Slika 3.13: Pareto dijagram sa aspekta pogodnosti kuća

Pareto chart analiza

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
Put	469	469	43.305633	43.30563
Podrum	191	660	17.636196	60.94183
Klima	173	833	15.974146	76.91597
Lokacija	128	961	11.819021	88.73500
Bazen	97	1058	8.956602	97.69160
Grejanje(gas)	25	1083	2.308403	100.00000

Dijagram paralelnih koordinata

U prvom poglavlju bilo je nešto više reči o dijagramu paralelnih koordinata. Sada ćemo to znanje primeniti na zadati primer. S obzirom da smo već analizirali većinu kategorijalnih promenljivih i znamo njihove veze sa cenama kuća, preostalo nam je još da analiziramo broj kupatila i garaža koje kuće poseduju i njihovu povezanost sa cenama kuća i veličinama placeva. Posmatraćemo kakav uticaj na cenu ima broj kupatila, zatim vezu između cene i broja garaža kao i vezu između broja garaža i površine placa. Rezultati su prikazani na slici 3.14. Slika 3.14 nam govori da kuće



Slika 3.14: Dijagram paralelnih koordinata

sa jednim kupatilom obično imaju niže cene, jednu ili ni jednu garažu. Primećujemo da na tržištu preovladavaju kuće sa jednim i dva kupatila. Kuće sa 2 kupatila imaju i višu cenu. Što se tiče kuća sa 3 kupatila, retke su i cene im variraju dok sa 4 imamo samo jednu kuću čija cena je prilično visoka.
Ako posmatramo broj garaža primećujemo da su kuće bez garaža obično jeftinije i nalaze se na manjim placevima. Kuće sa jednom i dve garaže imaju i više cene i nalaze se na većim placevima dok su kuće sa 3 garaže retke, nalaze se na manjim placevima i uglavnom su jeftine.

Dijagram stablo-list i usečena sredina

Histogram je pokazao raspodelu podataka, međutim, bolji uvid u podatke daje nam stablo-list dijagram, (tabela 3.1).

2	555567778
3	000122333344444
3	55555566666777788888888999
4	00111111112222222233333333444444
4	55555555555555566666777777777888888888889999999999
5	00000000000000000000011111222222222233333333334444444444
5	55555555566666666677777778888888899999999
6	00000000000000000000011111111122222222333333334444444444
6	55555555555556666666677777778888899999
7	00000000000000001111222222333334 4
7	55555555555555667788888999999
8	00000000001222222333334444 4
8	55555555566677777788899999
9	00000002222333334
9	555555555667778999
10	01112334
10	5555566677888
11	0023344
11	55679
12	00000123444
12	5778
13	00223
13	8
14	001
14	55
15	
15	5
16	3
16	
17	
17	555
18	
18	
19	0

Tabela 3.1: Dijagram stablo-list koji prikazuje cene kuća

Vidmo da kuće najčešće imaju cenu od 50 do 54 hiljade dolara. Modus uzorka je 60000$, medijana 62000$. Vidimo da su cene koncentrisane u prvoj polovini tabele što nam govori o χ^2 raspodeli.
Takođe, primećujemo i 6 kuća koje koštaju više od 150000$ i koje znatno odstupaju od uobičajenih cena kuća pa ih možemo tretirati kao autlajere. Posmatrajmo naj-

višu od 190000$, koja najviše odstupa. Ako pogledamo prethodno dobijene rezultate vidimo da ta kuća ima 2 kupatila, 2 garaže, 4 spavaće sobe, 3 sprata, na poželjnoj je lokaciji, površine oko $7500m^2$ i nema bazen. S obzirom da smo pareto dijagramom ustanovili da je bazen jedan od ključnih faktora koji utuče na cenu kuće, a ova kuća ga nema možemo zaključiti da je ovaj podatak o ceni autlajer ili da kuća poseduje neku posebnu pogodnost koja u ovim podacima nije iskazana. Obično kuće sa sličnim karakteristikama imaju niže cene. Grejanje na gas isključujemo iz ispitivanja jer ono nije faktor koji povećava cenu.

Usečena sredina za cenu kuća nam je 63305$ a za površinu placa $4778m^2$.

Odnos šansi

Iskoristićemo odnos šansi da ispitamo koje su šanse da kuće koje su na poželjnoj lokaciji imaju cenu veću od 100000$, (tabela 3.3).

	Poželjna lokacija	Nepoželjna lokacija	Ukupno
$\geq$ 100000$	26	39	65
$<$ 100000$	103	379	482
Ukupno	129	418	547

Tabela 3.2: Podaci u odnosu na lokaciju i cenu kuće

Odnos šansi je:

$$OR = 2.359$$

Zaključujemo da je 2.539 puta veća šansa da kuća na poželjnoj lokaciji ima cenu veću od 100000$ u odnosu na kuće na nepoželjnoj lokaciji, tj. 2.539 puta je veća šansa da kuće na nepoželjnoj lokaciji imaju cenu manju od 100000$.

Ispitaćemo sada koje su šanse da kuće koje imaju cenu veću od 100000$ imaju bazen, (tabela 3.3).

	Ima bazen	Nema bazen	Ukupno
$\geq$ 100000$	25	41	66
$<$ 100000$	167	314	481
Ukupno	192	355	547

Tabela 3.3: Podaci u odnosu katogoriju "bazen" i cenu kuće

Odnos šansi je:

$$OR = 1.147$$

Zaključujemo da je 1.05467 puta veća šansa da kuća koja ima bazen košta više od 100000$ u odnosu na kuću koja nema bazen.

Na osnovu dva dobijena odnosa šansi vidimo da je lokacija ipak ključan faktor za visoku cenu kuće.

Zaključak

Rezultat istraživanja prezentovan ovim master radom je sistematizacija modernih metoda za analizu podataka. Koristeći skup EDA tehnika i znanja u oblasti statistike na posmatranim podacima dobijamo velike količine informacija. Dobijene informacije mogu biti korisne u svim oblastima, za davanje procena, istraživanje tendencija, procenu rizika, formiranje mapa, analiziranje odnosa i faktora koji određuju posmatranu pojavu.

U prvom poglavlju smo videli korisnost pojedinačnih EDA tehnika. Svaka tehnika nam daje drugačije informacije i koristimo ih u zavisnosti od informacija koje želimo da dobijemo iz posmatranih podataka.
Takođe, videli smo i značaj statističkog paketa R koji nam olakšava crtanje grafika i izvačenje informacija iz velike grupe podataka. Kada je reč o ogromnim količinama podataka, znanje o EDA tehnikama bi nam bilo bezvredno da ne koristimo neki statistički paket. R se u ovom slučaju pokazao kao merodavan jer poseduje sve pakete koji su nam neophodni.
Na kraju, treće poglavlje nam je zapravo sistematizacija svega naučenog o EDA tehnikama. Vidimo da se one dopunjavaju, da jedna tehnika nije dovoljna za donošenje zaključaka o prirodi cele populaciji. Upotrebom više tehnika možemo čak dobiti i različite rezultate koji nam signaliziraju da treba da razmislimo o prirodi problema i ne zaključujemo samo na osnovu rezultata koje nam je računar dao.

Možemo zaključiti da je aparat EDA tehnika vrlo značajan i primenljiv. Istraživačka analiza podataka je mnogo slobodnija tehnika u odnosu na bajesovsku i klasičnu analizu podataka. Za razliku od bajesovske i klasične, EDA nam sama formira model, fokusira se na podatke, njihovu strukturu, autlajere kao i modele koje sugerišu podaci. Videli smo da EDA nije samo set tehnika nego pristup tome kako da se nosimo sa različitim tipovima podataka.

Literatura

[1] Brillinger D., Fernholz L., Morgenthaler S.: *The practice of data analysis: Essays in Honor of John W. Tukey. (eBook, Paperback and Hardcover)*. Princeton legacy library, (1998)

[2] Croarkin C., Filliben J., Guthrie W., Heckert N., Hembree B., Prinz J., Tobias P.: *E- Handbook od Statistical Methods*. National istitute of standards and technology's, (2002)

[3] Genton M., Sun Y.: *Functional median polish*. International Biometric Society (2012)

[4] Hautzenthaler M.: *R course*. February, (2011)

[5] Hintze J., Nelson R.: *Violin plots: A box plot-density trace synergism*. The American Statistician, (1998), 181-184.

[6] Hoaglin D., Velleman P.: *Applications, basics and computing of Exploratory data analysis*. Duxbury Press, Boston, Massachusetts, (1981)

[7] Lozanov-Crvenkovic Z.: *Statistika*. Prirodno matematički fakultet, Novi Sad (2012)

[8] Lozanov-Crvenkovic Z.: *Statističko modeliranje-skripta*. Prirodno matematički fakultet, Novi Sad (2015)

[9] Lumley T.: *R fundamentals and programming techniques*. E core development team and UW Dept of Biostatistics, Birmingham (2006), 27-28.

[10] Maindonald J.H.: *Using R for Data Analysis and Graphics Introduction, Code and Commentary*. Centre for Mathematics and Its Applications, Australian National University, (2008)

[11] Murray S., Perla R., Provost L.,: *The run chart: a simple analytical tool for learning from variation in healthcare processes*. BMJ Qual Saf, (2011)

[12] Pecina M.: *Metode multivarijantne analize-interna skripta*. Agronomski fakultet, Zagreb (2006)

[13] Radziwill N.: *Pareto charts*. Quality and innovation, (2012), 75-925

[14] Wickelmaier F.: *An Introduction to MDS*. Sound Quality Research Unit, Aalborg University, Denmark (2003)

[15] Web sajt: Available CRAN Packages By Date of Publication,
https://cran.r-project.org/

[16] Web sajt: Estimated Travel Distance between European Cities,
http://www.mapcrow.info/europeantraveldistance.html

[17] Web sajt: Explorable (trimean),
https://explorable.com/trimean

[18] Web sajt: MarinStatsLectures,
http://www.statslectures.com/index.php/r-stats-datasets

[19] Web sajt: R bloggers,
https://http://www.r-bloggers.com/

[20] Web sajt: R Datasets,
https://vincentarelbundock.github.io/Rdatasets/datasets

[21] Web sajt: Republički hidrometeorološki zavod,
http://www.hidmet.gov.rs/

[22] Web sajt: Safari,
https://www.safaribooksonline.com/blog/2014/03/31/
mastering-parallel-coordinate-charts-r/

[23] Web sajt: STAT 464, applied nonparametric statistics,
https://onlinecourses.science.psu.edu/stat464/node/65

[24] Web sajt: The R Project for Statistical Computing,
https://www.r-project.org/

Biografija

Dijana Krstić rođena je 6.11.1992. godine u Bijeljini. Završila je Osnovnu školu "Dositej Obradović" 2007. godine kao vukovac i učenik generacije. Iste godine upisuje Srednju tehničku školu "Mihajlo Pupin" u Bijeljini i završava je takođe kao vukovac i učenik generacije. Nakon toga, 2011. godine upisuje osnovne studije Primenjene matematike na Prirodno-matematičkom fakultetu u Novom Sadu koje završava u septembru 2014. godine.
Iste godine upisuje master studije Primenjene matematike. Položila je sve ispite predviđene planom i programom master studija uključujući i pedagošku grupu predmeta zaključno sa junskim rokom 2016. godine i prosekom 8,72.

Bila je stipendista grada Bijeljine kao srednjoškolac od 2007. do 2011. i kao student u periodu od 2011. do 2014. godine.

Novi Sad, jun 2016

Dijana Krstić

UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATICKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
RBR

Identifikacioni broj:
IBR

Tip dokumentacije: Monografska dokumentacija
TD

Tip zapisa: Tekstualni štampani materijal
TZ

Vrsta rada: Master rad
VR

Autor: Dijana Krstić
AU

Mentor: dr Zagorka Lozanov-Crvenković
MN

Naslov rada: Istraživačka analiza podataka uz upotrebu statističkog softvera R
NR

Jezik publikacije: srpski (latinica)
JP

Jezik izvoda: srpski/engleski
JI

Zemlja publikovanja: Srbija
ZP

Uže geografsko područje: Vojvodina
UGP

Godina: 2016
GO

Izdavač: Autorski reprint
IZ

Mesto i adresa: Novi Sad, Departman za matematiku i informatiku, Prirodno-matematički fakultet, Univerzitet u Novom Sadu, Trg Dositeja Obradovića 4
MA

Fizički opis rada: (3/81/21/32/0/10)
(broj poglavlja/strana/lit./tabela/slika/grafika/priloga)
FO

Naučna oblast: Matematika
NO

Naučna disciplina: Statistika
ND

Predmetna odrednica/Ključne reči: Statistika, Istraživačka analiza podataka, EDA, Histogram, Boks dijagram, Dijagram rasturanja, Dijagram protoka u vremenu, Pareto dijagram, Dijagram paralelnih koordinata, OR količnih, Multidimenzionalno skaliranje, Stablo-list dijagram, Violina dijagram, Doterana sredina, Usečena sredina, Statistički softver R
PO
UKR

Čuva se: Biblioteka Departmana za matematiku i informatiku Prirodno
-matematičkog fakulteta Univerziteta u Novom Sadu
ČU

Važna napomena:
VN

Izvod:

Istraživačka analiza podataka je pristup za analizu podataka kao i interpretacijom dobijenih rezultata, prikazujući ih najčešće vizuelno. Začetnik ove analize je Džon Taki (John Tukey) koji se datom oblašću počeo baviti još 1961. godine. Karakteristična je po tome što se prilikom analize podataka koriste posebne grafičke i kvantitativne tehnike. Kvantitivne tehnike predtavljaju skup procedura koji daje numeričke rezultate. Sa druge strane, grafičkih tehnika ima više tako da ćemo njima posvetiti više pažnje. Prvo poglavlje je posvećeno opisu pojedinačnih EDA tehnika i njihovoj matematičkoj osnovi. Izlazne vrednosti i grafici su dobijeni korišćenjem statističkog softvera R.
Drugo poglavlje je posvećeno R-u. Detaljno su opisane sve naredbe koje se koriste u implementaciji EDA tehnika.
Treće poglavlje posvećeno je kompletnom aparatu EDA tehnika. Jedan primer je obrađen kroz više tehnika i dobijena je velika količina informacija.
IZ

Datum prihvatanja teme od strane NN veća: 03.03.2016.
DP

Datum odbrane: jun 2016.
DO

Članovi komisije:
KO

Predsednik: dr Mirjana Ivanović, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Član: dr Zagorka Lozanov-Crvenković, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu, mentor

Član: dr Ivana Štajner-Papuga, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

UNIVERSITY OF NOVI SAD
FACULTY OF NATURAL SCIENCE AND MATHEMATICS
KEY WORDS DOCUMENTATION

Accession number:
ANO

Identification number:
INO

Dokument type: Monograph documentation
DT

Type of record: Textual printed material
TR

Contents code: Master thesis
CC

Author: Dijana Krstić
AU

Mentor: dr Zagorka Lozanov-Crvenković
MN

Title: Exploratory data analysis (EDA) using software for statistical analysis R
TI

Language of text: Serbian (latin)
LT

Language of abstract: Serbian/ English
LA

Country of publication: Serbia
CP

Locality of publication: Vojvodina
LP

Publication year: 2016.
PY

Publisher: Author's reprint
PU

Publ. place: Faculty of Natural Science and Mathematics, Novi Sad, Trg Dositeja Obradovića 4
PP

Physical deskription: (3/81/21/32/0/10)
PD

Scientific field: Mathematics
SF

Scientific discipline: Statistics
SD

Subject Key words:: Statistics, Exploratory data analysis, EDA, Histogram, Box plot, Scatter plot, Run chart, Pareto chart, Parallel coodrinate chart, OR ratio, Multidimensional scaling, Stem and leaf , Violin plot, Median polish, Tremmean, Software for statistical analysis R
SKW
UC

Holding data: Library of the Department of Mathematics and Computer Sciences
HD

Note: None
N

Abstract:

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. Exploratory data analysis was promoted by John Tukey, 1961. Exploratory data analysis procedures can broadly be split into two parts: quantitative and graphical. Quantitative techniques are the set of statistical procedures that yield numeric or tabular output. On the other hand, there is a large collection of statistical tools that we generally refer to as graphical techniques. First section describes many techniques that are commonly used in exploratory data analysis. The sample plots and output in this section were generated with the statistical software R. Second section describes stastistical software R and R with EDA techniques. Last section concerned one example and we use all EDA techniques to describe them.
AB

Accepted on the Scientific board on: 03.03.2016.
AS

Defended: 2016.
DE

Thesis Defend board:
D

President: dr Mirjana Ivanović, full professor, Faculty of Science, University of Novi Sad

Mentor: dr Zagorka Lozanov-Crvenković, full professor, Faculty of Science, University of Novi Sad

Member: dr Ivana Štajner-Papuga, full professor, Faculty of Science, University of Novi Sad