SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Josipa Vresk

Bruno Žitković

DOBRA KNJIGA

SEMINARSKI RAD

Varaždin, 2021.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Josipa Vresk, 0016130227

Bruno Žitković, 0016130386

Studij: Baze podataka i baze znanja

DOBRA KNJIGA

SEMINARSKI RAD

Mentorica:

Prof. dr. sc. Jasminka Dobša

Varaždin, lipanj 2021.

Sadržaj

1. Uvod

U ovom seminarskom radu bit će obrađen skup podataka koji sadrži zapise o knjigama i njihovim recenzijama. Skup podataka koji će biti obrađen preuzet je sa sustava Kaggle [1], a u svrhu obrade podataka bit će korišten alat R Studio.

Korišteni podaci sadrže zapise o 58 292 instance u kojima su sadržani podaci o naslovu knjige, godini izdanja, izdavaču, ocjeni knjige i drugo.

Kroz rad ćemo opisati varijable statističkog skupa i prikazati ih grafički, izračunati matricu korelacija između svih kvantitativnih varijabli, ispitati normalnost razdiobe određenih varijabli te definirati nekoliko novih varijabli. Uz to ispitat ćemo određene razlike i ovisnosti između pojedinih varijabli te ćemo definirati nekoliko modela regresije.

2. Zadatak a)

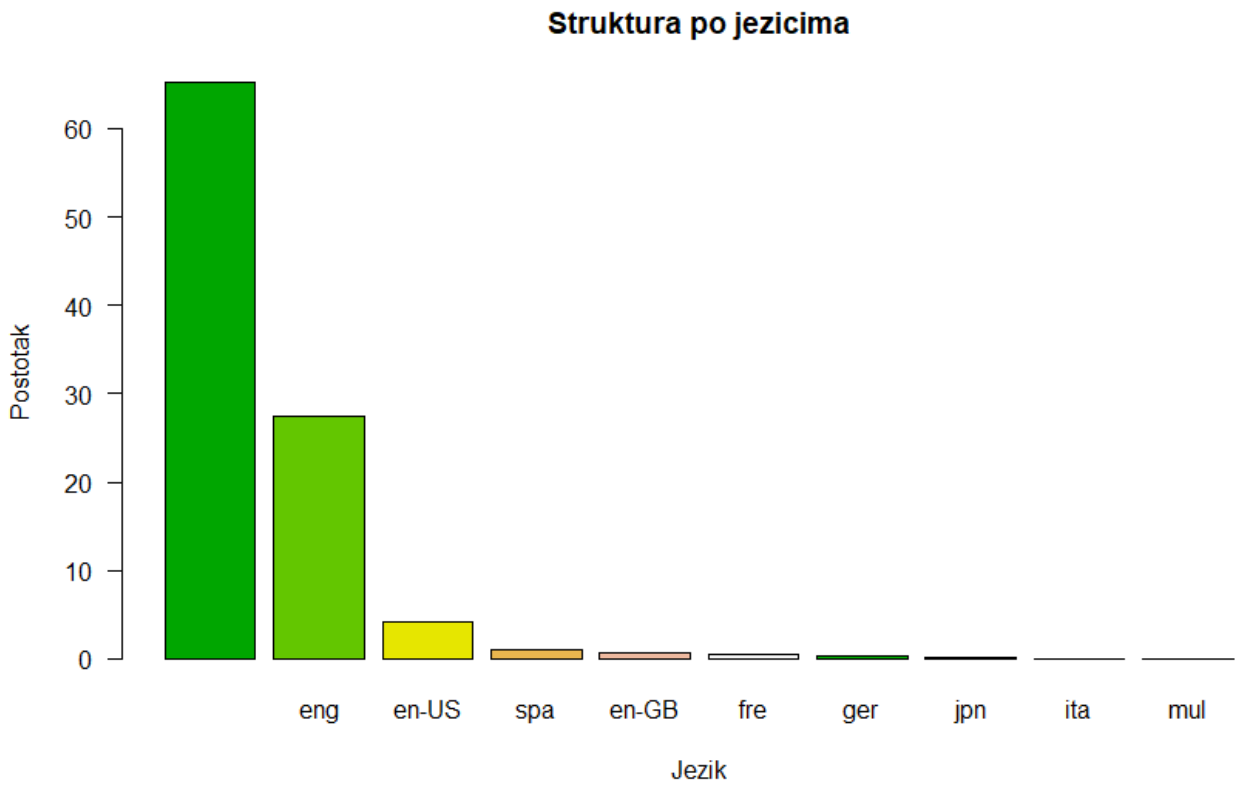
Opišite varijable statističkog skupa i grafički ih prikažite.

U statističkom skupu za dobru knjigu ima ukupno 18 varijabli koje zajedno s njihovim opisom možemo vidjeti u nastavku:

Id	Identifikator zapisa
Name	Naslov knjige
PageNumber	Broj stranica knjige
RatingDistTotal	Ukupni broj ocjena
PublishMonth	Mjesec u kojem je knjiga objavljena
PublishDay	Dan u kojem je knjiga objavljena
Publisher	Ime izdavača knjige
CountsOfReview	Broj recenzija za knjigu
PublishYear	Godina u kojoj je knjiga objavljena
Language	Jezik na kojem je knjiga objavljena
Authors	Autori knjige
Rating	Ocjena knjige
RatingDist1	Broj ocjene od 1 zvjezdice
RatingDist2	Broj ocjene od 2 zvjezdice
RatingDist3	Broj ocjene od 3 zvjezdice
RatingDist4	Broj ocjene od 4 zvjezdice
RatingDist5	Broj ocjene od 5 zvjezdice
ISBN	Međunarodni standardni knjižni broj

Tablica 1: opis statističkih varijabli

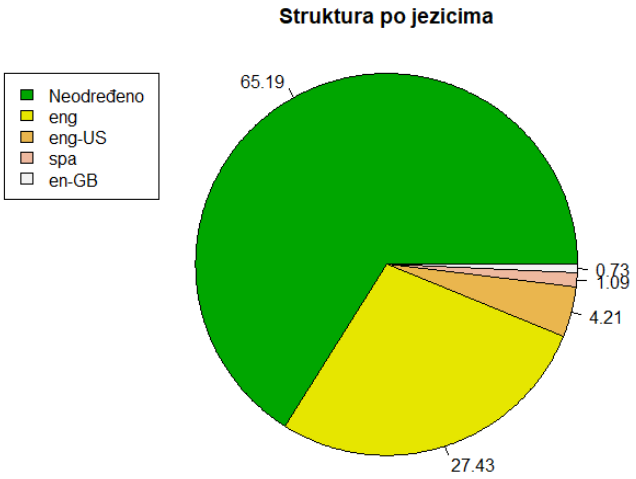
Kvalitativnu varijablu *Language* (Jezici) možemo prikazati pomoću barplota. Nakon filtriranja na prvih 10 najzastupljenijih jezika te sortiranja od većeg prema manjem dobivamo barplot kao na slici 1.



Slika 1: Barplot kvalitativne varijable *Language*

Na slici 1 možemo primijetiti kako najveći postotak knjiga nema definiran jezik, odnosno preko 60% naslova. Nakon toga slijedi engleski s nešto manje od 30%. Svi dalje navedeni jezici nalaze se unutar 5% u ukupnom broju jezika, a to su engleski s američkog govornog područja, španjolski jezik te ga slijedi engleski s britanskog govornog područja i ostali jezici.

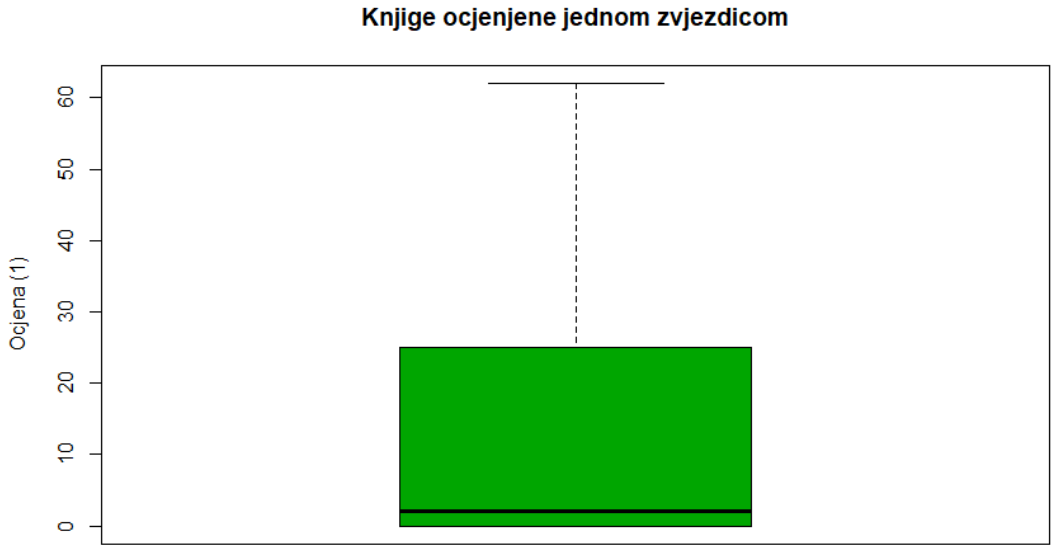
Jezike kao kvalitativnu varijablu možemo prikazati i pomoću pie chart-a odnosno kružnog grafa u kojem su jasno vidljivi postoci i udjeli jezika koji su definirani za knjige. Strukturu prema jezicima podatkovnog skupa možemo vidjeti na slici 2.



Slika 2: Pie chart kvalitativne varijable *Language*

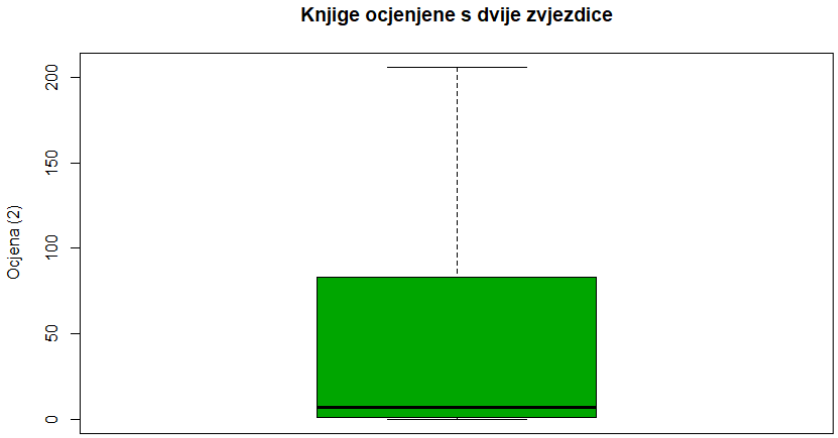
Na slici 2 možemo primijetiti postotke koji nam nisu vidljivi na slici 1 za isti skup podataka. Na taj način dolazimo do postotka kako 65.19% knjiga nema definiran jezik. Nakon čega slijedi engleski s 27.43%, engleski s američkog govornog područja s 4.21%, španjolski jezik s 1.09% te ga slijedi engleski s britanskog govornog područja s 0.73% i ostali jezici.

Kvantitativne varijable možemo prikazati pomoću boxplotova. Na slici 3-7 možemo vidjeti boxplot-ove za kvantitativne varijable vezane uz ocjenjivanje knjiga.



Slika 3: Boxplot kvantitativne varijable *RatingDist1*

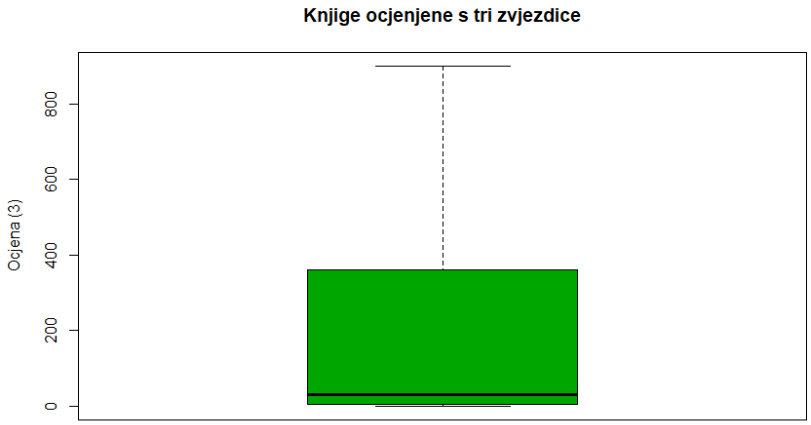
Na slici 3 možemo vidjeti minimum skupa koji iznosi 0 te maksimum skupa koji iznosi 62. Možemo zaključiti kako je 75% knjiga ocjenu 1 dobilo između 0 i 62 puta dok 25% knjiga dobilo ocjenu 1 nula puta. Zbog previše vrijednosti, a samim time i nepreglednosti iz grafičkog prikaza su isključeni outlieri.



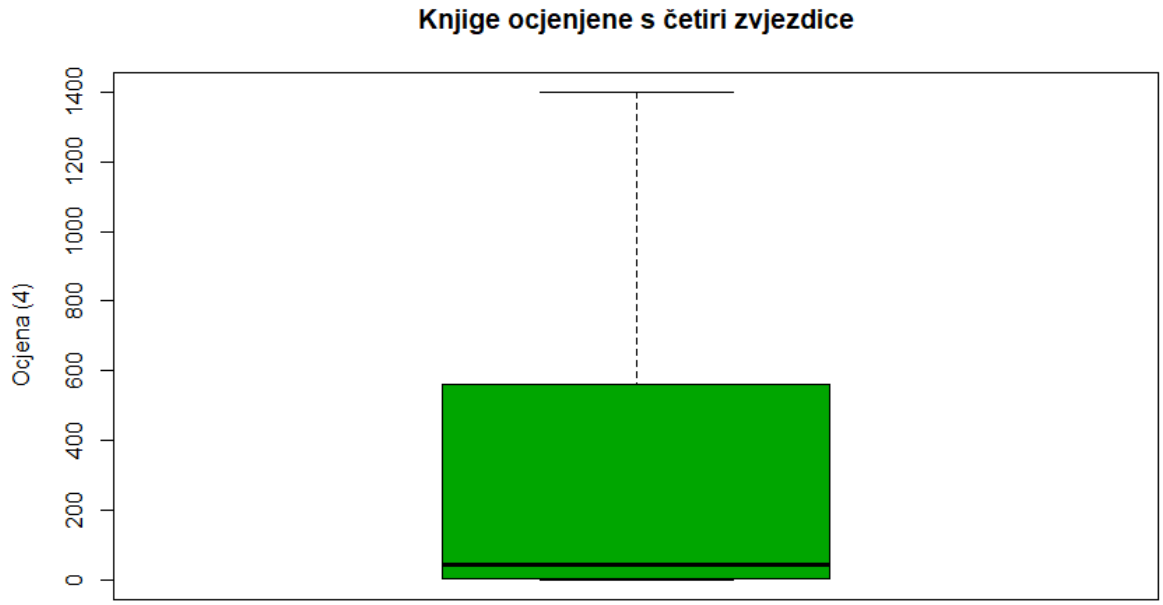
Slika 4: Boxplot kvantitativne varijable *RatingDist2*

Na slici 4 možemo vidjeti minimum skupa koji iznosi 0 te maksimum skupa koji iznosi 210 dok je medijan 5. Možemo zaključiti kako je 75% knjiga ocjenu 2 dobilo između 1 i 75 puta dok 25% knjiga je dobilo ocjenu 0 ili 1 puta. Zbog previše vrijednosti, a samim time i nepreglednosti iz grafičkog prikaza su isključeni outlieri.

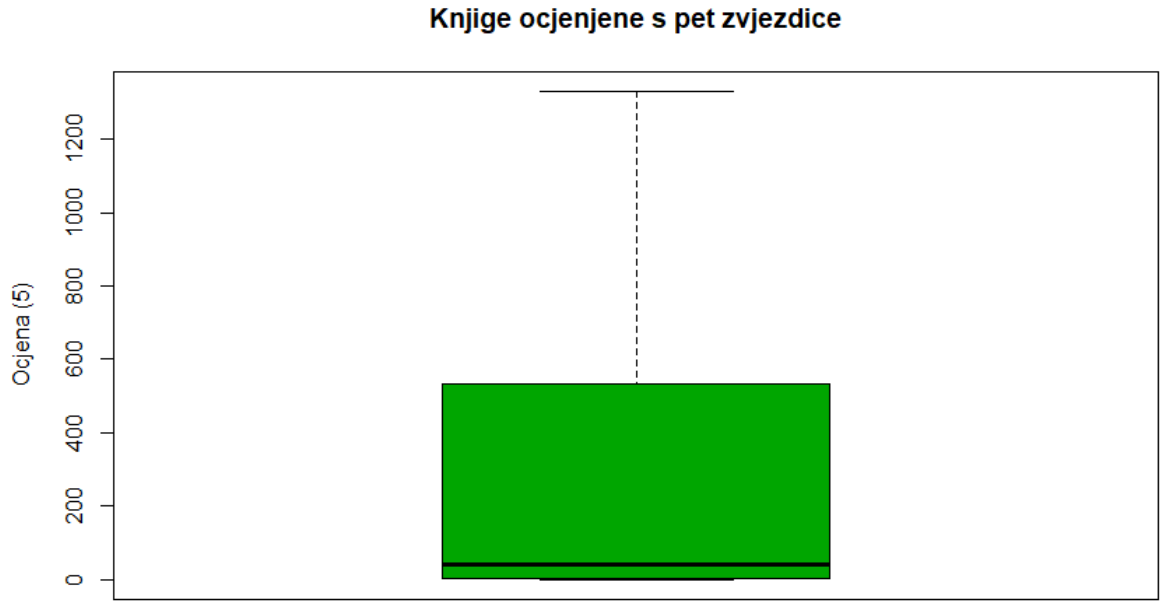
Na jednake načine interpretiraju se boxplotovi na slikama od 5 do 7 za vrijednosti ocjena 3, 4 i 5.



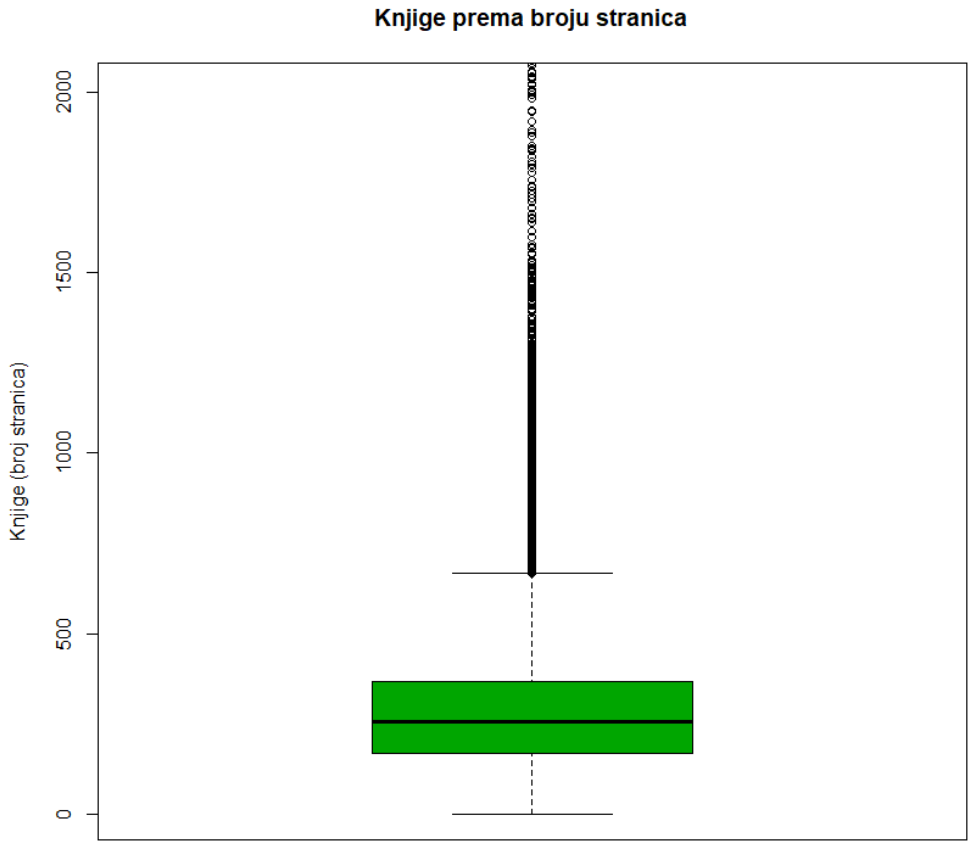
Slika 5:Boxplot kvantitativne varijable *RatingDist3*



Slika 6:Boxplot kvantitativne varijable *RatingDist4*



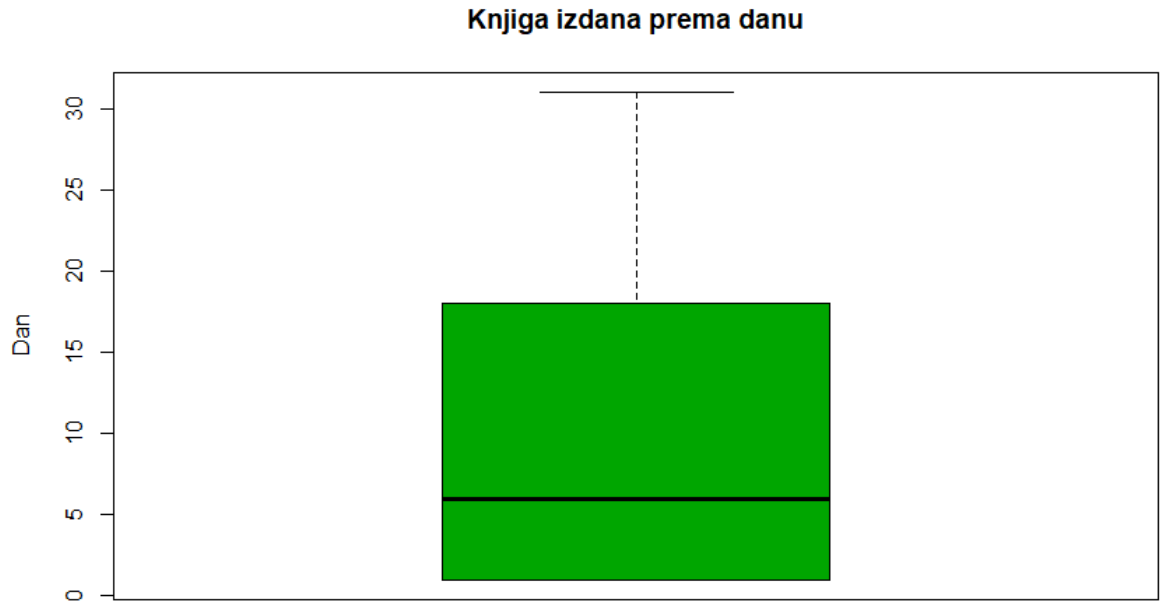
Slika 7:Boxplot kvantitativne varijable *RatingDist5*



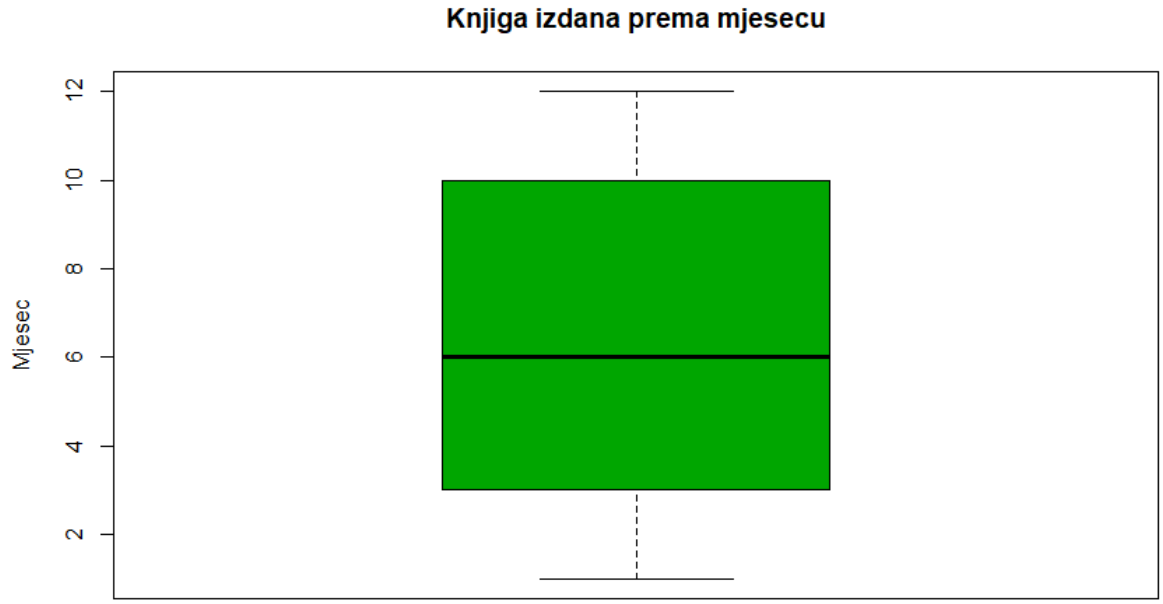
Slika 8: Boxplot kvantitativne varijable *PageNumbers*

Na slici 8 možemo vidjeti minimum skupa koji iznosi 1 te maksimum skupa koji iznosi 700 dok je medijan 300. Možemo zaključiti kako 75% knjiga ima broj stranica između 200 i 700 dok 25% knjiga ima između 1 i 200 stranica. Iznad maksimuma skupa vidljivi su outlieri koji predstavljaju abnormalne vrijednosti tj. ekstreme.

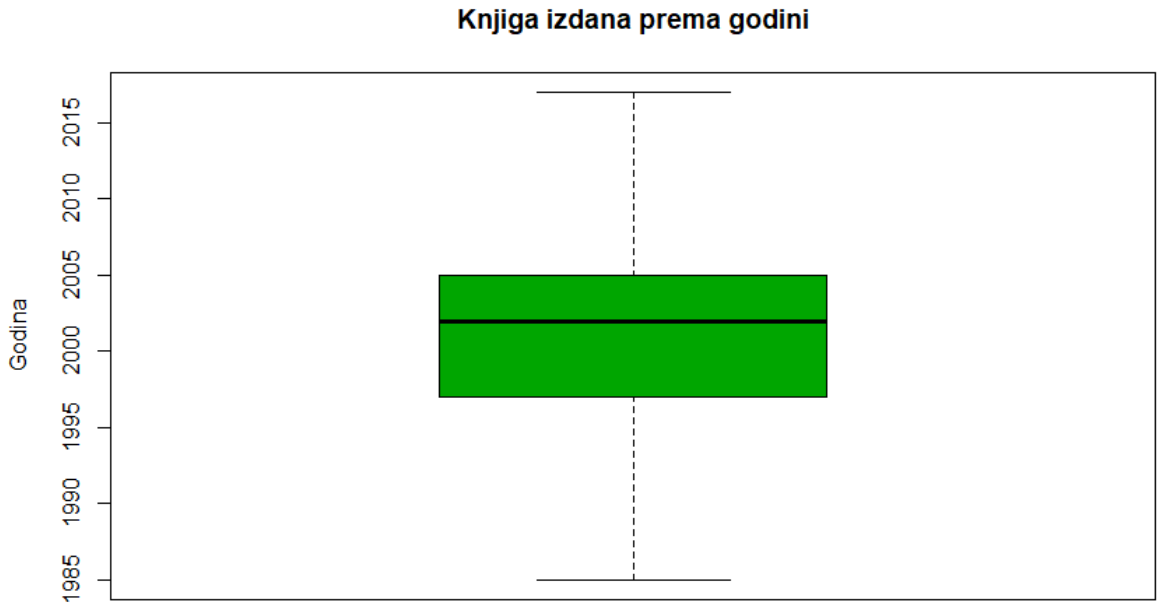
Na jednake načine interpretiraju se boxplotovi na slikama od 9 do 13 za vrijednosti varijabli *PublishDay, PublishMonth, PublishYear, Rating* za koje su isključeni outlieri zbog previše vrijednosti, a samim time i nepreglednosti grafičkog prikaza te *CountsofReview* za koje su outlieri prikazani.



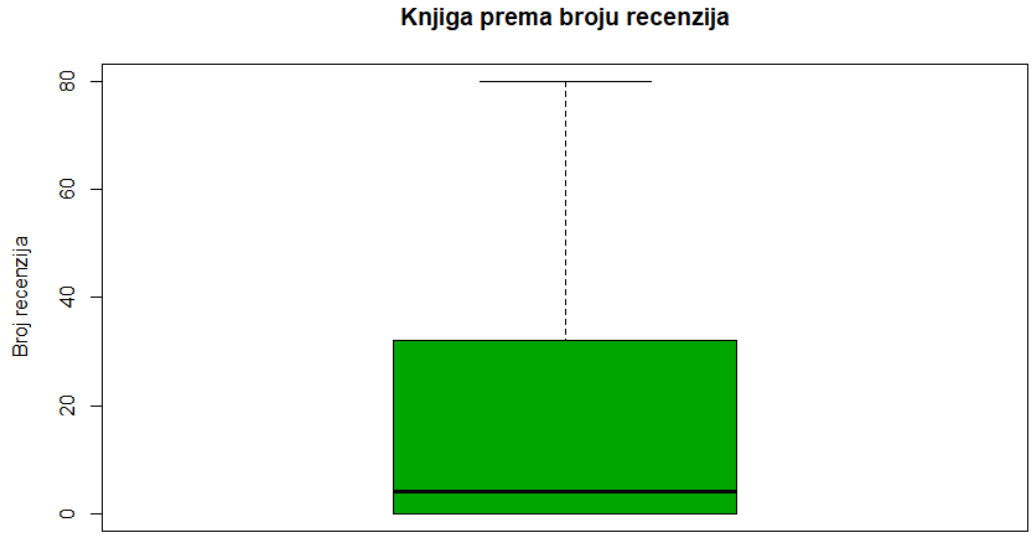
Slika 9: Boxplot kvantitativne varijable *PublishDay*



Slika 10: Boxplot kvantitativne varijable *PublishMonth*

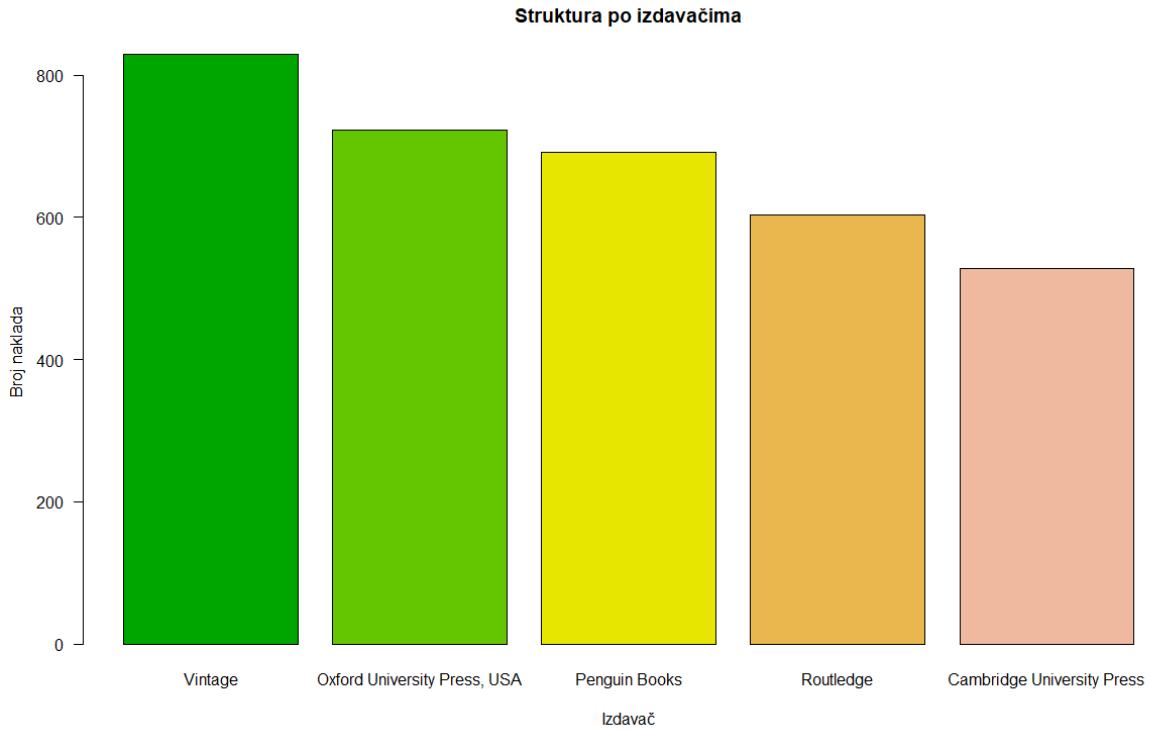


Slika 11: Boxplot kvantitativne varijable *PublishYear*



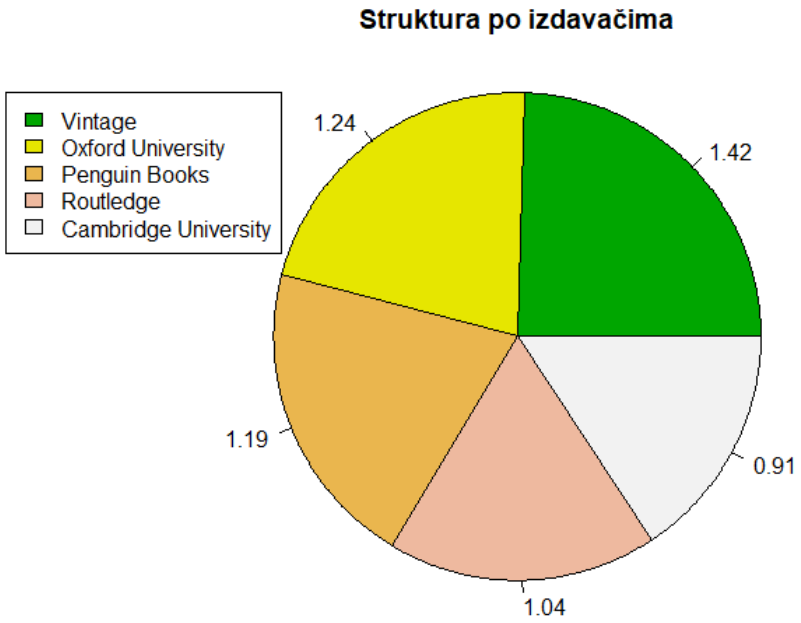
Slika 12: Boxplot kvantitativne varijable *Rating*

Slika 13: Boxplot kvantitativne varijable *CountsOfReview*



Slika 14: Barplot kvalitativne varijable *Publisher*

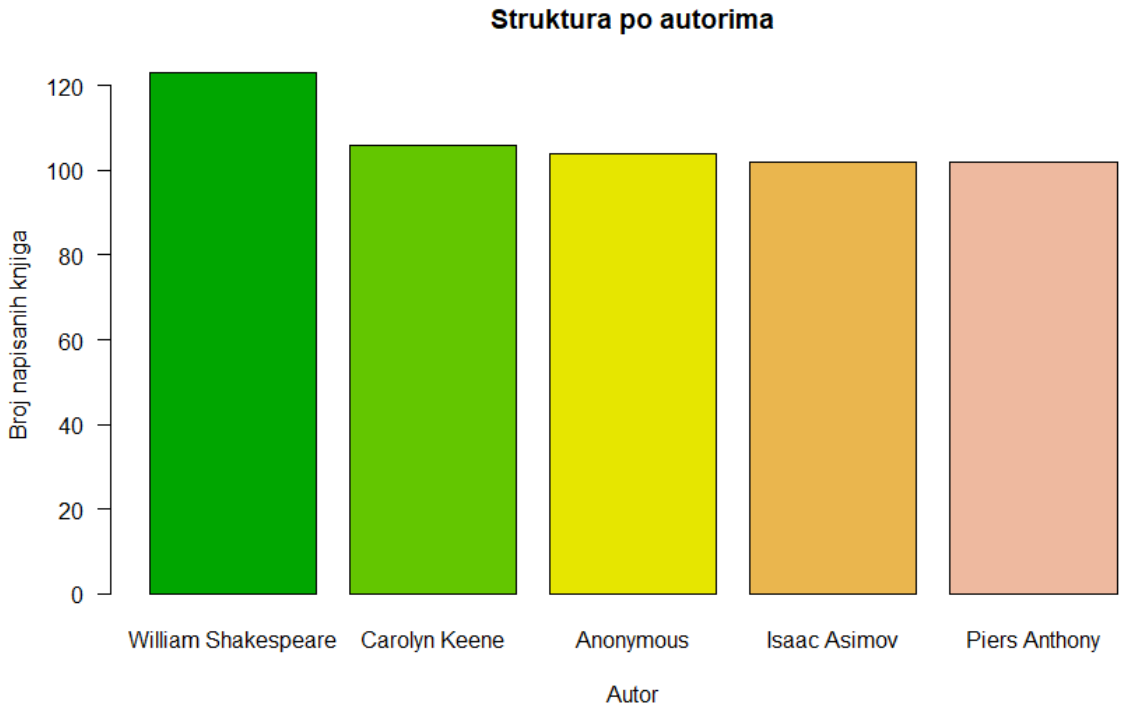
Na slici 14 možemo vidjeti barplot kvalitativne varijable *Publisher* koji nam govori kako je izdavač s najvišim brojem naklada Vintage ima otprilike 900 naklada, a slijede ga Oxford University Press, USA sa 700 naklada, Penguin Books sa 680 naklada, Routledge sa 600 naklada i Cambridge University Press sa 500 naklada. Zbog velikog broja izdavača izdvojeno je 5 njih s najviše naklada.



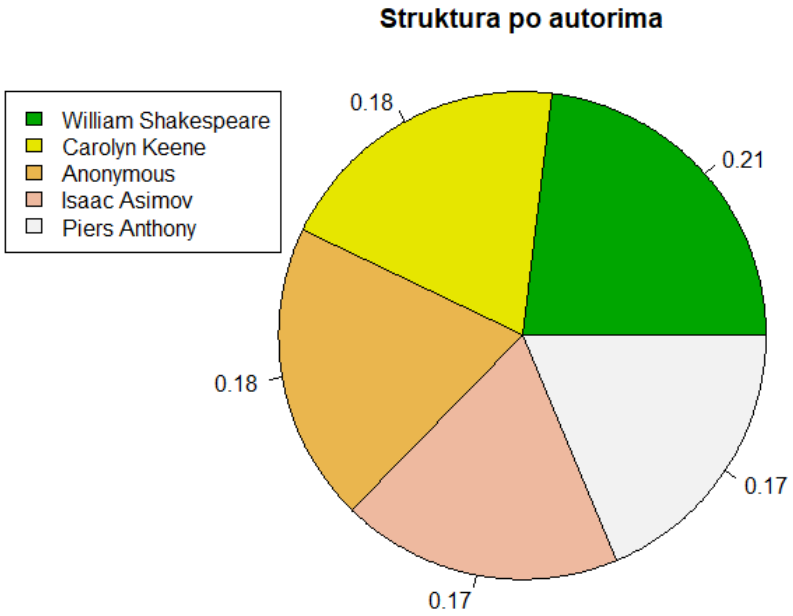
Slika 15: Pie chart kvalitativne varijable *Publisher*

Na slici 15 možemo vidjeti kružni grafikon kvalitativne varijable *Publisher* koji nam govori kako je izdavač s najvišim brojem naklada Vintage ima otprilike ima udio od 1.42% u ukupnom skupu, a slijede ga Oxford University Press, USA s udjelom od 1.24%, Penguin Books s udjelom od 1.19%, Routledge s udjelom 1.04% i Cambridge University Press s udjelom od 0.91%. Zbog velikog broja izdavača izdvojeno je 5 njih s najviše naklada.

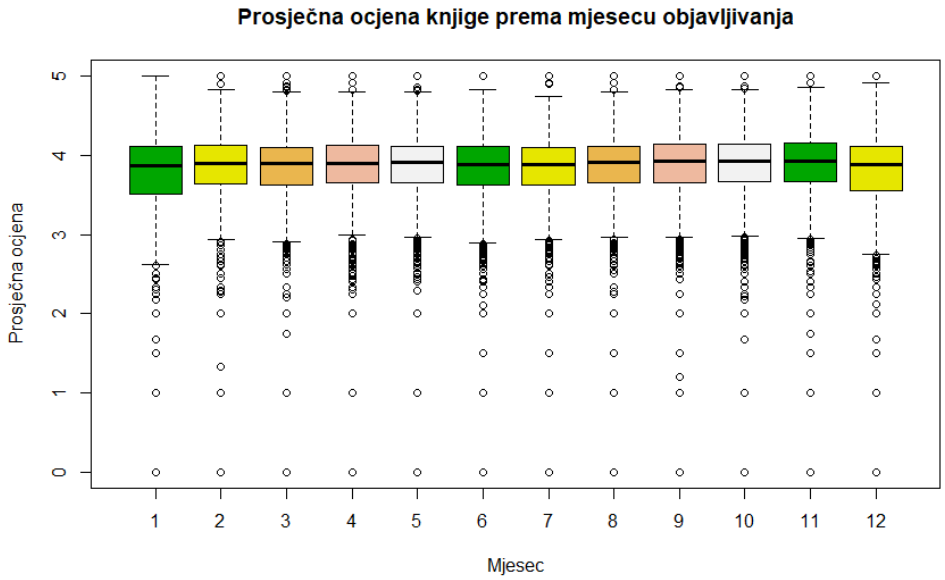
Na isti način interpretiraju se grafički prikazi na slikama 16 i 17 koje prikazuju broj napisanih knjiga za 5 autora s najvišim brojem napisanih knjiga.



Slika 16: Barplot kvalitativne varijable *Authors*

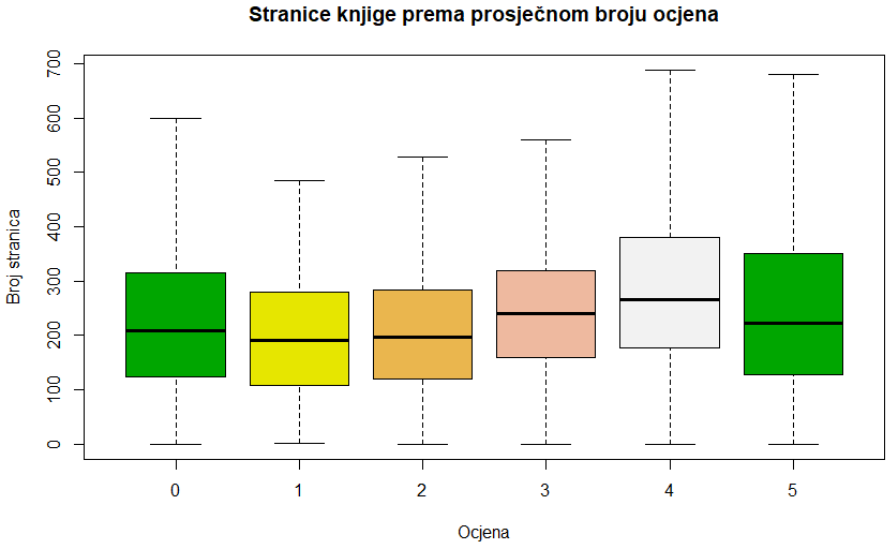


Slika 17: Pie chart kvalitativne varijable *Authors*



Slika 18: Boxplot odnosa varijabli *PublishedMonth* i *Rating*

Boxplot na slici 18 prikazuje odnos varijable *Rating* prema *PublishedMonth* na kojem možemo vidjeti kako mjesec izdavanja knjige nema prevelik utjecaj na ocjenu koja je dodijeljena knjizi.



Slika 19: Boxplot odnosa varijabli *Rating* i *PagesNumber*

Boxplot na slici 19 pokazuje kakav je odnos varijabli *PagesNumber* i *Rating-a* na kojem možemo vidjeti kako knjige s većim brojem stranica dobivaju veće ocjene.

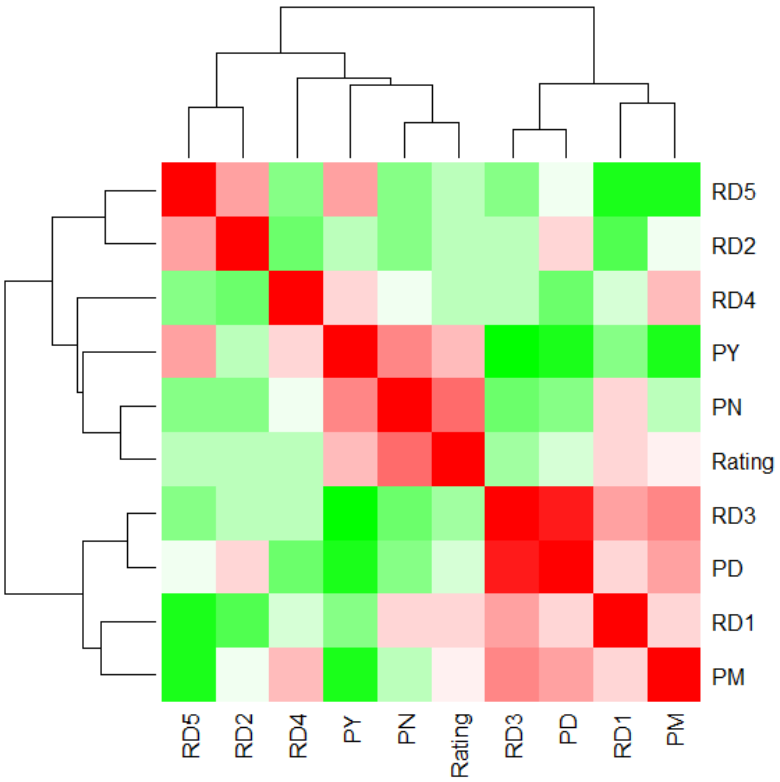
3. Zadatak b)

Izračunajte matricu korelacija između svih kvantitativnih varijabli, grafički prikažite korelacije i interpretirajte rezultate.

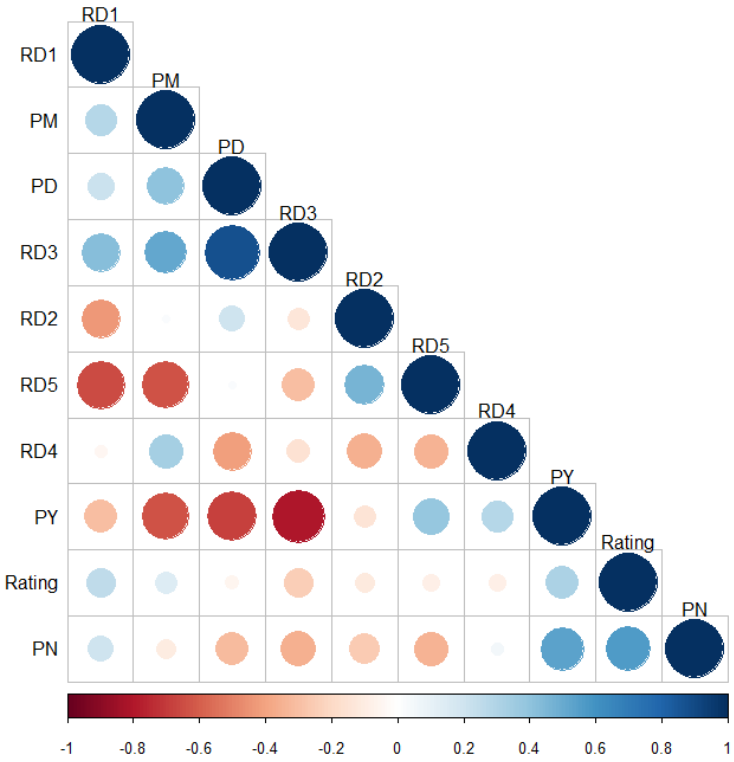
	Publish Day	Publish Month	Rating	Pages Number	Publish Year	Rating Dist1	Rating Dist2	Rating Dist3	Rating Dist4	Rating Dist5
Publish Day	1,00	0,40	-0,06	-0,32	-0,69	0,21	0,20	0,87	-0,42	0,02
Publish Month	0,40	1,00	0,15	-0,11	-0,64	0,28	0,02	0,51	0,33	-0,64
Rating	-0,06	0,15	1,00	0,56	0,31	0,25	-0,12	-0,25	-0,09	-0,09
Pages Number	-0,32	-0,11	0,56	1,00	0,53	0,20	-0,26	-0,36	0,05	-0,34
Publish Year	-0,69	-0,64	0,31	0,53	1,00	-0,31	-0,15	-0,80	0,28	0,38
Rating Dist1	0,21	0,28	0,25	0,20	-0,31	1,00	-0,44	0,42	-0,05	-0,66
Rating Dist2	0,20	0,02	-0,12	-0,26	-0,15	-0,44	1,00	-0,14	-0,36	0,46
Rating Dist3	0,87	0,51	-0,25	-0,36	-0,80	0,42	-0,14	1,00	-0,16	-0,31
Rating Dist4	-0,42	0,33	-0,09	0,05	0,28	-0,05	-0,36	-0,16	1,00	-0,34
Rating Dist5	0,02	-0,64	-0,09	-0,34	0,38	-0,66	0,46	-0,31	-0,34	1,00

Tablica 2: tablični prikaz matrice korelacija

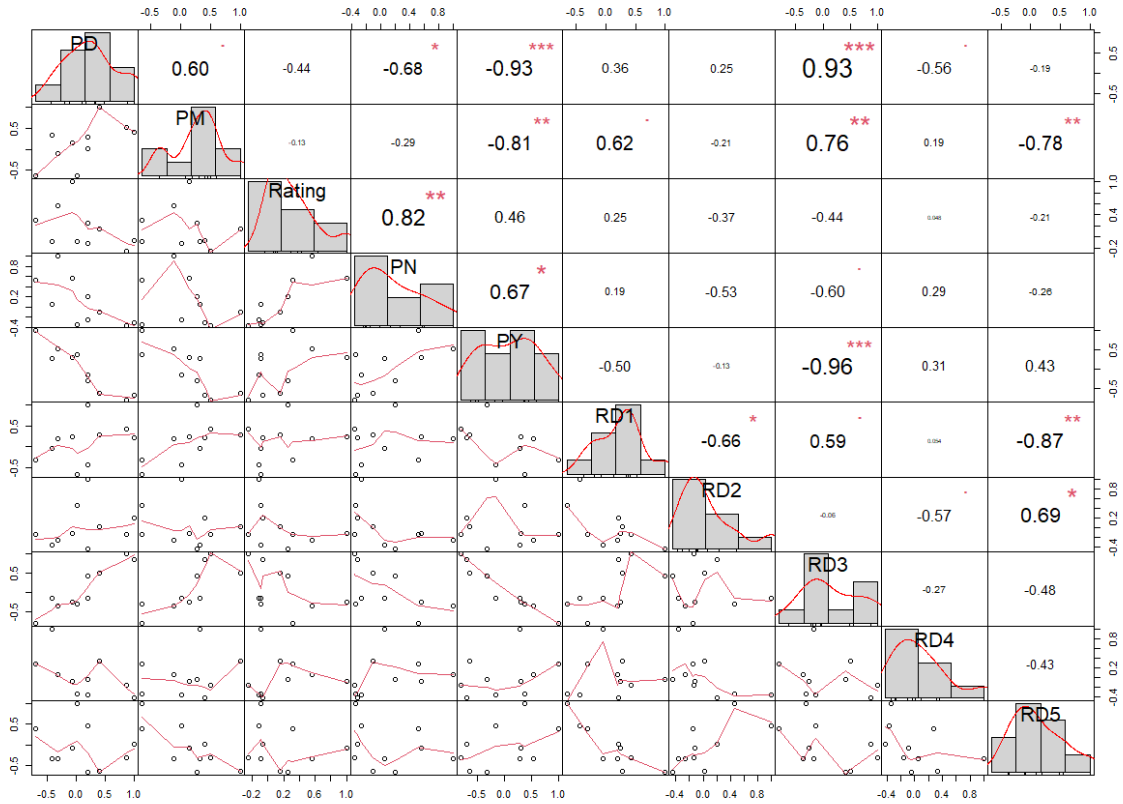
Tablicom 1 prikazan je odnos varijabli odnosno njihova matrica korelacije. Snažnu pozitivnu korelaciju imaju *RatingDist3* i *PublishDay* koja iznosi 0.87 dok snažnu negativnu korelaciju imaju varijable *PublishYear* i *RatingDist3* koja iznosi -0.80. Grafičke prikaze matrice korelacije možemo vidjeti na slikama 20 , 21 i 22.



Slika 20: Heat map matrice korelacija



21: Corrplot matrice korelacije

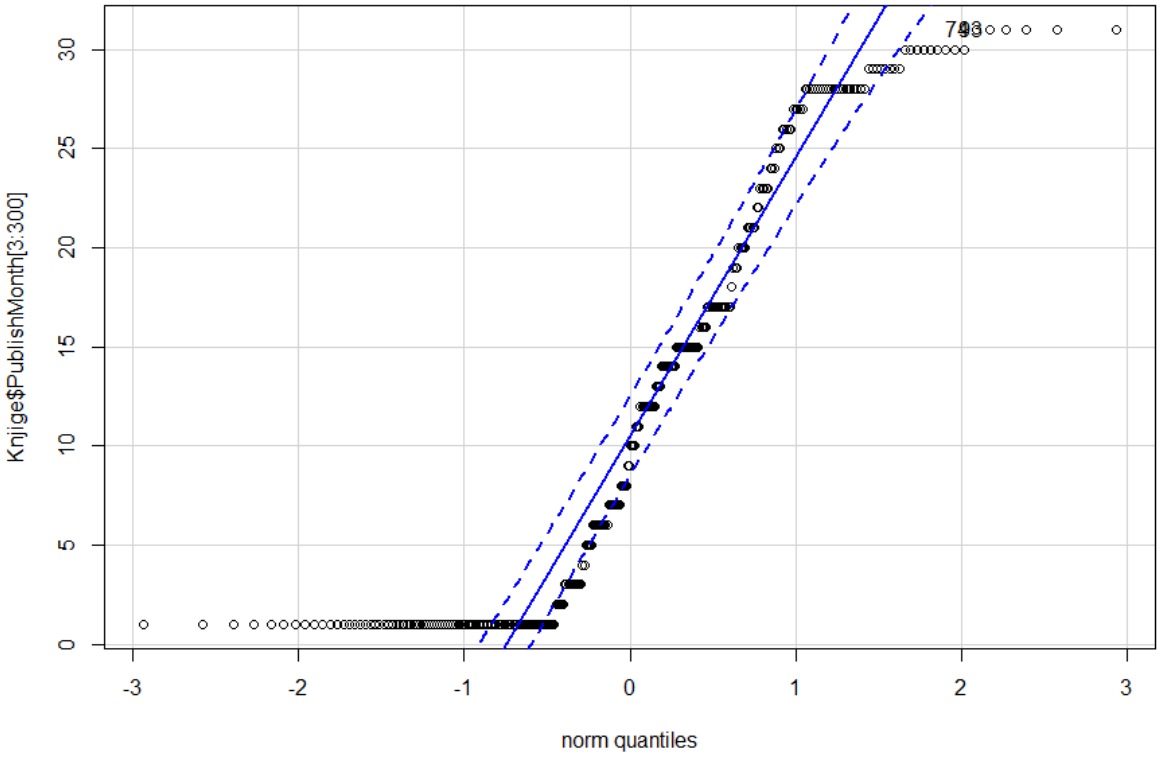


Slika 22: Grafički prikaz analitike matrice korelacije

4. Zadatak c)

Ispitajte normalnost razdiobe za varijable iz b) dijela zadatka.

```
1. shapiro.test(Knjige$PublishMonth[1:300])
2. Shapiro-Wilk normality test
3. data:  Knjige$PublishMonth[1:300]
4. W = 0.8527, p-value = 3.431e-16
```



Slika 23: QQ-plot *PublishMonth*

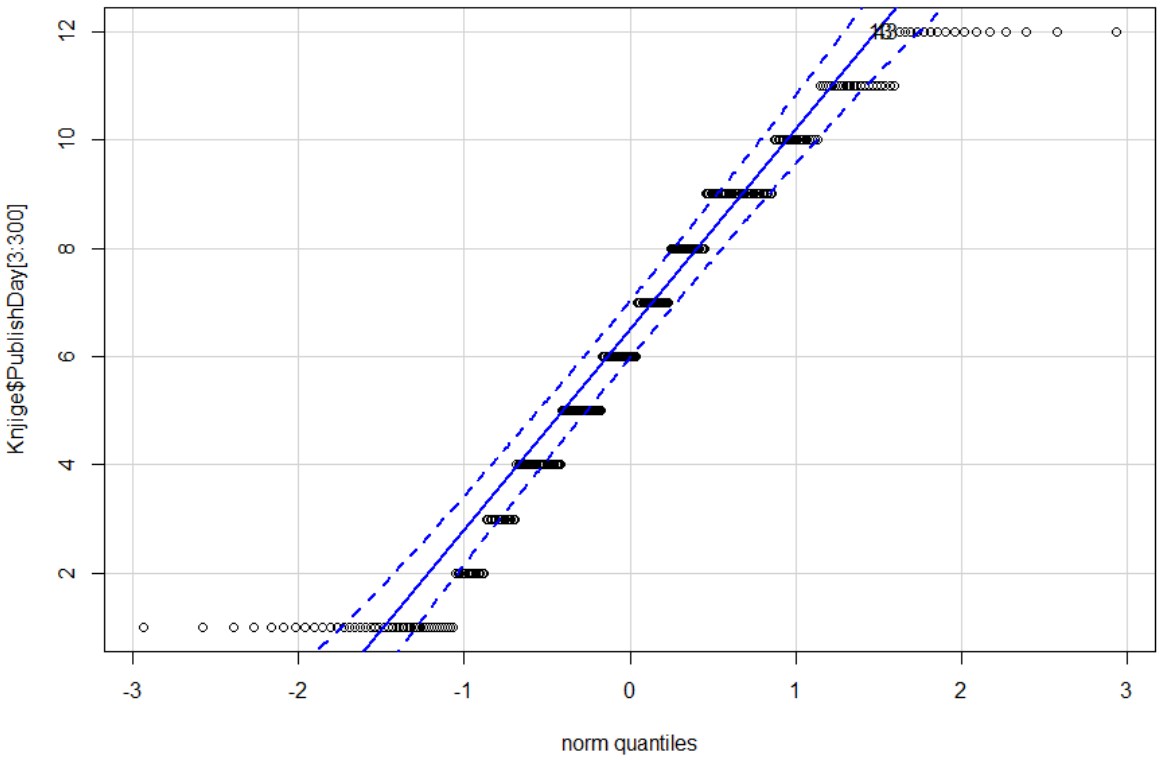
Normalnost razdiobe testirali smo Shapiro-Wilk testom s razinom signifikantnosti od 95%. Postavili smo dvije hipoteze:

- H0 – vrijednosti varijable dolaze iz normalno distribuiranog skupa
- H1 – vrijednosti varijable ne dolaze iz normalno distribuiranog skupa

Provodeći test za varijablu *PublishMonth*, dobili smo rezultat da *p-vrijednost iznosi* $3.431e^{-16}$ što je manje od 0.005, a to znači da odbacujemo H0 hipotezu i prihvaćamo H1, tj. da vrijednosti varijable ne dolaze iz normalno distribuiranog skupa, a to možemo i vidjeti na slici 23.

Provodeći isti test za ostale varijable iz b) dijela zadatka utvrdili smo kako kod svih odbacujemo H0 hipotezu i prihvaćamo H1 hipotezu, a to možemo vidjeti na slikama od 24 do 32.

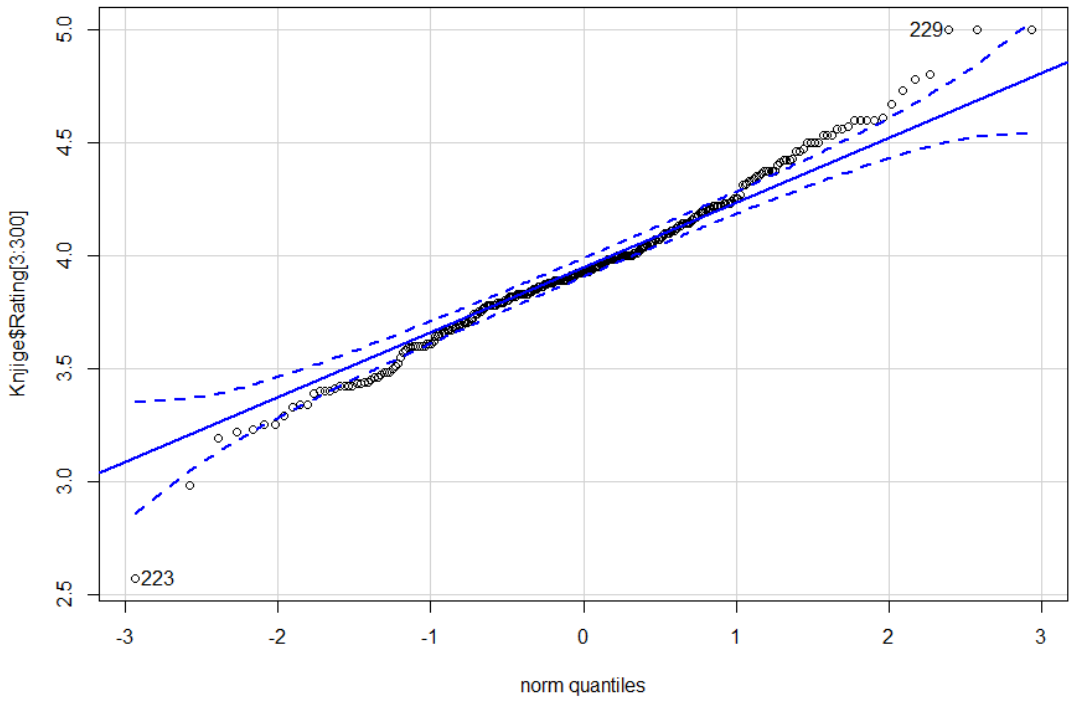
```
1. shapiro.test(Knjige$PublishDay[1:300])
2. Shapiro-Wilk normality test
3. data:  Knjige$PublishDay[1:300]
4. W = 0.93882, p-value = 8.346e-10
```



Slika 24: QQ-plot *PublishDay*

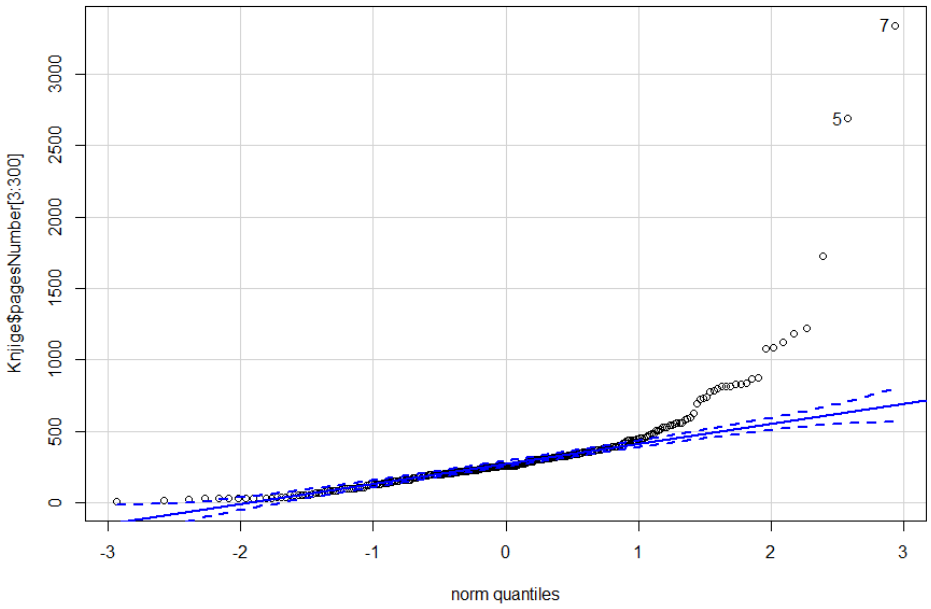
```
1. shapiro.test(Knjige$Rating[1:300])
2. Shapiro-Wilk normality test
```

```
3. data:  Knjige$Rating[1:300]
4. W = 0.98779, p-value = 0.01239
```



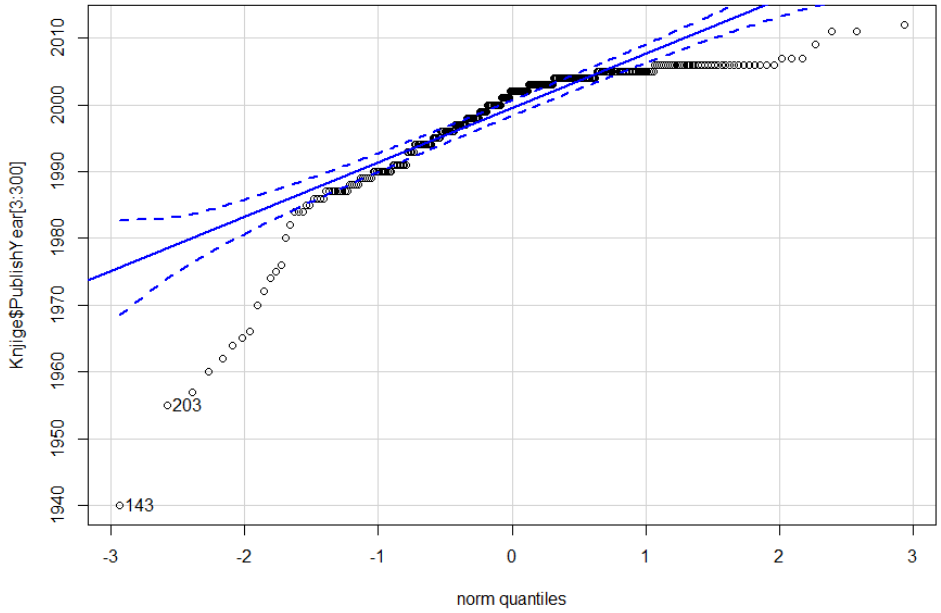
Slika 25: QQ-plot *Rating*

```
1. shapiro.test(Knjige$pagesNumber[1:300])
2. Shapiro-Wilk normality test
3. data:  Knjige$pagesNumber[1:300]
4. W = 0.60884, p-value < 2.2e-16
```



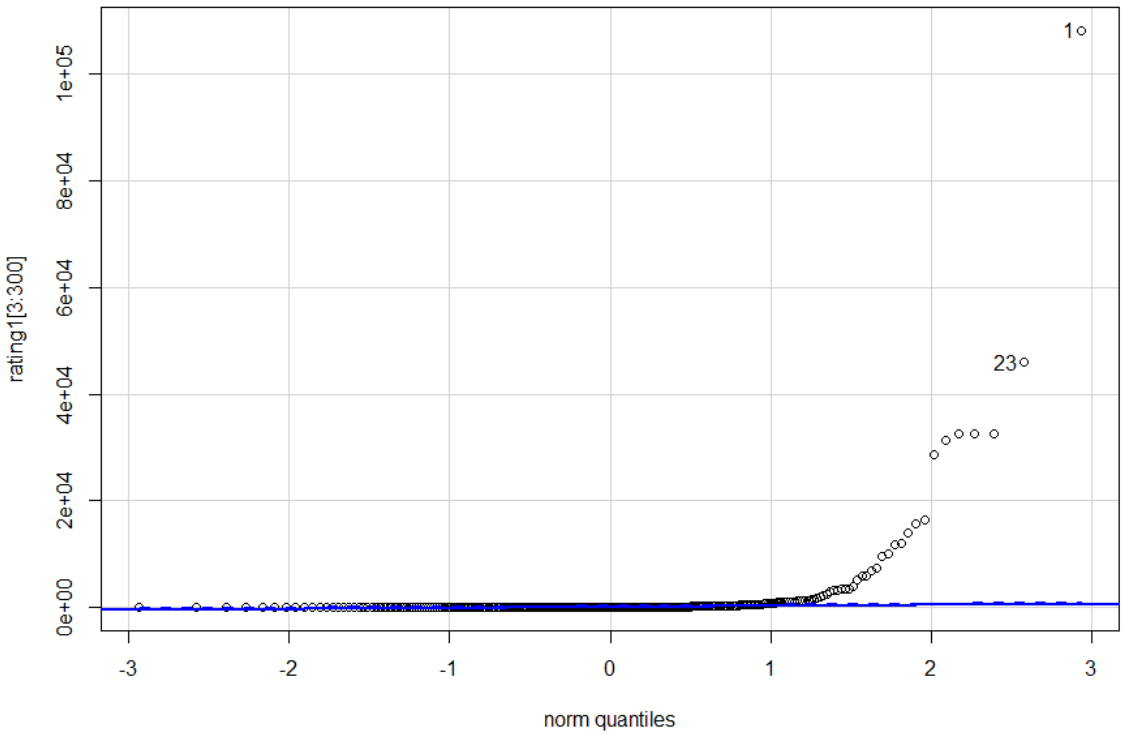
Slika 26: QQ-plot *pagesNumber*

```
1. shapiro.test(Knjige$PublishYear[1:300])
2. Shapiro-Wilk normality test
3. data:  Knjige$PublishYear[1:300]
4. W = 0.78183, p-value < 2.2e-16
```



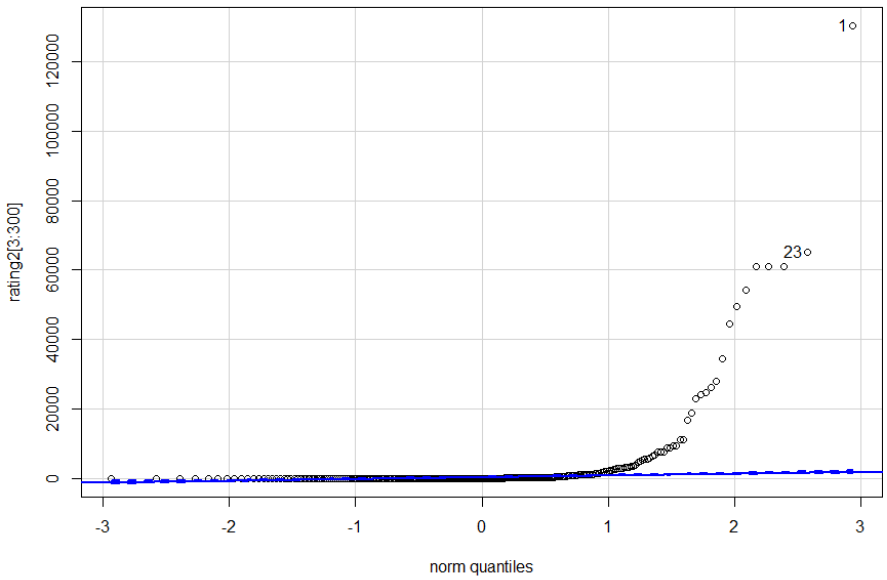
Slika 27: QQ-plot *PublishYear*

```
1. shapiro.test(rating1[1:300])
2. Shapiro-Wilk normality test
3. data:  rating1[1:300]
4. W = 0.21313, p-value < 2.2e-16
```



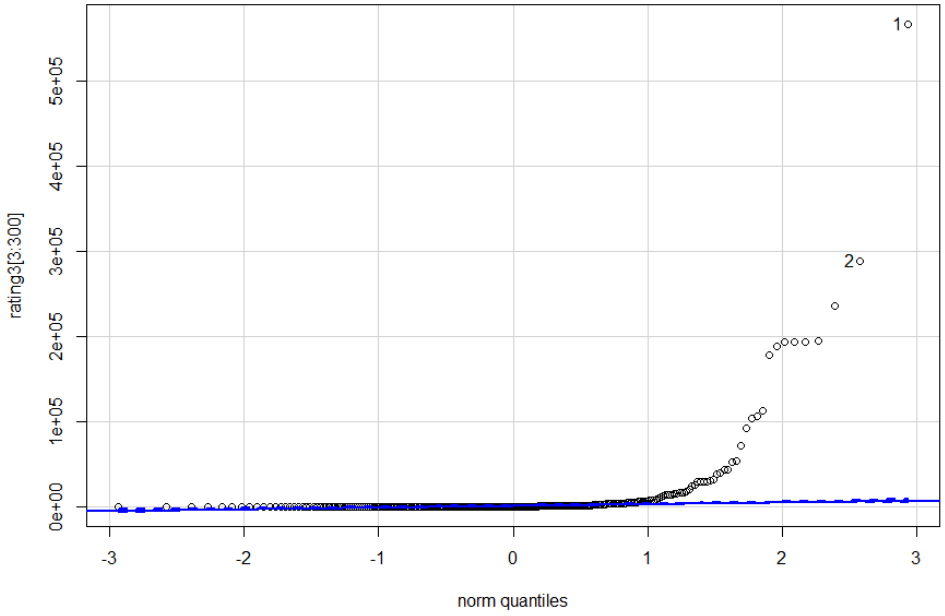
Slika 28: QQ-plot *RatingDist1*

```
1. shapiro.test(rating2[1:300])
2. Shapiro-Wilk normality test
3. data:  rating2[1:300]
4. W = 0.29685, p-value < 2.2e-16
```



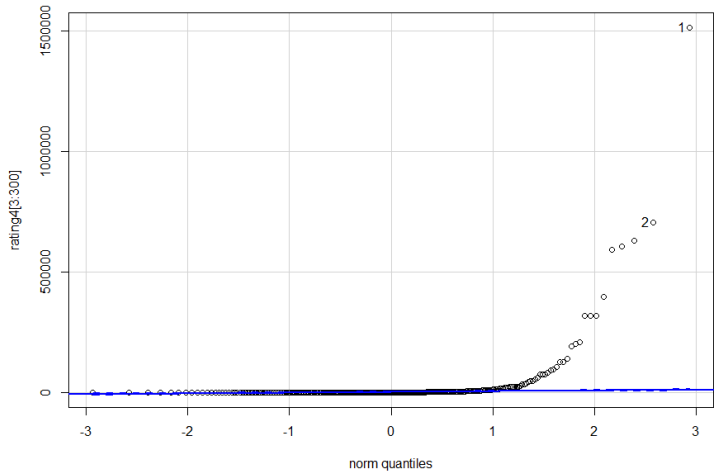
Slika 29: QQ-plot *RatingDist2*

```
1. shapiro.test(rating3[1:300])
2. Shapiro-Wilk normality test
3. data:  rating3[1:300]
4. W = 0.2843, p-value < 2.2e-16
```



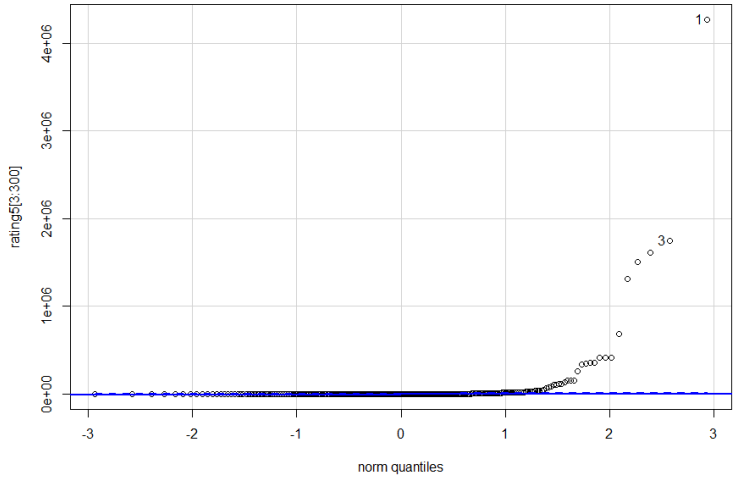
Slika 30: QQ-plot *RatingDist3*

```
1. shapiro.test(rating4[1:300])
2. Shapiro-Wilk normality test
3. data:  rating4[1:300]
4. W = 0.24589, p-value < 2.2e-16
```



Slika 31: QQ-plot *RatingDist4*

```
1. shapiro.test(rating5[1:300])
2. Shapiro-Wilk normality test
3. data:  rating5[1:300]
4. W = 0.18894, p-value < 2.2e-16
```



Slika 32: QQ-plot *RatingDist5*

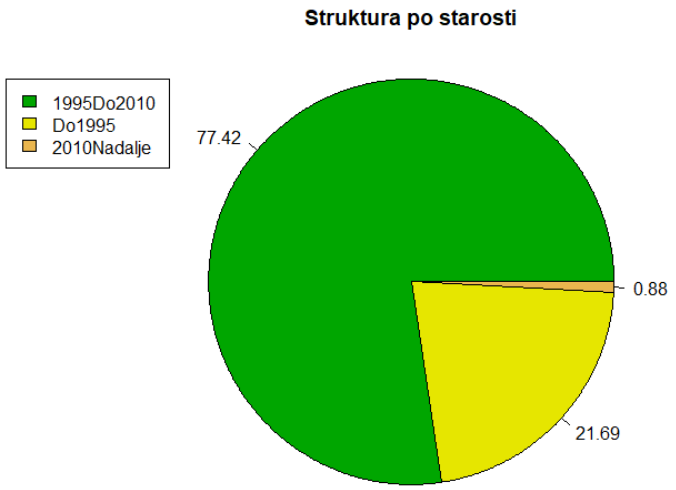
5. Zadatak d)

Definirajte novu varijablu za definiranu varijablu korištenjem varijable PublishYear na sljedeći način:

$$Starost = \begin{cases} Do1995 & Ako\ je\ PublisYear \leq 1995 \\ 1995Do2010 & Ako\ je\ 1995 < PublishYear \leq 2010 \\ 2010Nadalje & Ako\ je\ PublishYear > 2010 \end{cases}$$

```
196 # Dodavanje nove varijable Starost
197 Knjige$Starost <- cut(Knjige$PublishYear,breaks=c(162,1995,2010,Inf),labels=c("Do1995", "1995Do2010", "2010Nadalje"),
198                    as.factor.result=TRUE)
199
200 count<-table(Knjige$Starost)
201 count<-sort(count, decreasing = TRUE )
202 count
203
204 postoci<-round(100*count/dim(Knjige)[1],2)
205 pie(postoci, labels=postoci, edges = 900, radius = 1, col = terrain.colors(5), main="Struktura po starosti")
206 legend(-2, 1.0, c("1995Do2010", "Do1995", "2010Nadalje"), NULL, fill=terrain.colors(5))
```

Slika 33: Prikaz koda definiranja nove varijable *Starost*



Slika 34: Pie chart varjable *Starost*

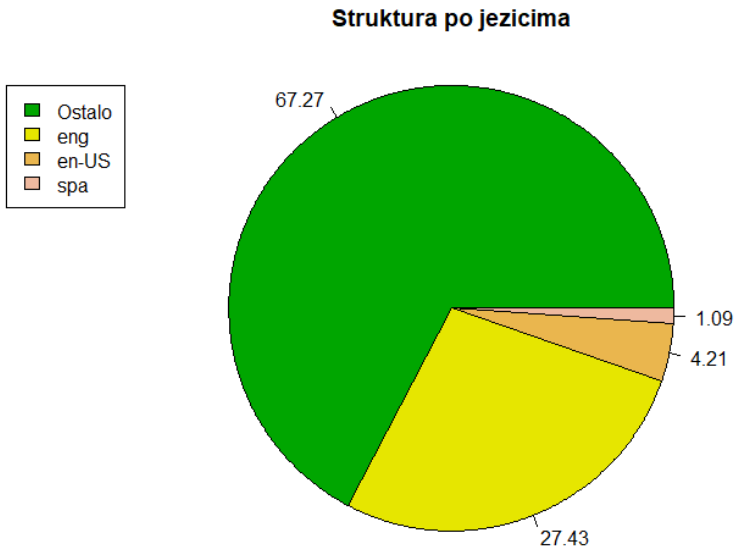
Nakon što smo definirali novu varijablu *Starost* i podijelili zapise prema zadanim parametrima na knjige napisane do 1995., od 1995. do 2010. i od 2010. nadalje, prikazali smo strukturu te podjele pomoću kružnog dijagrama na slici 34.

6. Zadatak e)

Definirajte novu varijablu *Language1* korištenjem varijable *Language* tako da podaci ostanu isti za tri najčešća jezika, a za ostale jezike stavite oznaku *Ostali*.

```
208 # Dodavanje nove varijable Languge1
209 count<-table(Knjige$Language1)
210 count<-sort(count, decreasing = TRUE )
211
212 Knjige$Language1 <- Knjige$Language
213
214 Knjige$Language1[Knjige$Language1 == "eng"] <- "eng"
215 Knjige$Language1[Knjige$Language1 == "en-US"] <- "en-US"
216 Knjige$Language1[Knjige$Language1 == "spa"] <- "spa"
217 Knjige$Language1[Knjige$Language1 == ""] <- "Ostalo"
218 Knjige$Language1[Knjige$Language1 != "eng" & Knjige$Language1 != "en-US" & Knjige$Language1 != "spa"] <- "Ostalo"
219
220 postoci<-round(100*count/dim(Knjige)[1],2)
221 pie(postoci, labels=postoci, edges = 900, radius = 1, col = terrain.colors(5), main="Struktura po jezicima")
222 legend(-2, 1.0, c("Ostalo", "eng", "en-US", "spa"), NULL, fill=terrain.colors(5))
223
```

Slika 35: Prikaz koda definiranja nove varijable *Language1*



Slika 36: Pie chart varjable *Language1*

Nakon što smo definirali novu varijablu *Language1* i podijelili zapise prema tri najčešće korištena jezika te na ostale jezike dobivamo strukturu jezika prikazanu na slici 36.

view	PublishYear	Language	Authors	Rating	RatingDist2	RatingDist5	ISBN	RatingDist3	Language1	Starost
1	2002	fre	Paul Auster	4.32	2:1	5:10	2742741461	3:1	Ostalo	1995Do2010
0	1999		Bernd Herzogenrath	4.00	2:0	5:2	9042004533	3:2	Ostalo	1995Do2010
0	2003		Harold Bloom	3.82	2:1	5:2	0791076628	3:2	Ostalo	1995Do2010
11	1997		Paul Auster	3.87	2:23	5:121	0140267506	3:134	Ostalo	1995Do2010
0	2001		Aliki Varvogli	3.60	2:0	5:1	0853236879	3:5	Ostalo	1995Do2010
0	1995		Dennis Barone	4.00	2:0	5:2	0812215567	3:2	Ostalo	Do1995
2806	2005	eng	Jared Diamond	3.93	2:2859	5:17546	0143036556	3:12582	eng	1995Do2010
26	2006	en-US	Stephen Leeb	3.39	2:28	5:40	0446579785	3:104	en-US	1995Do2010
88	1990	eng	Joseph A. Tainter	4.16	2:25	5:341	052138673X	3:127	eng	Do1995
518	2001	eng	Robert D. Putnam	3.80	2:329	5:1293	0743203046	3:1406	eng	1995Do2010
5	1997		Scott M. Huse	4.10	2:4	5:31	0801057744	3:13	Ostalo	1995Do2010
0	1988		Vincent Dunn	4.83	2:0	5:5	0878149031	3:0	Ostalo	Do1995
4	2002	eng	Philip K. Howard	3.91	2:3	5:24	034543871X	3:18	eng	1995Do2010
54	2003		Stephen Kotkin	3.73	2:30	5:117	0195168941	3:198	Ostalo	1995Do2010
20	1998	eng	Joy Harjo	4.39	2:4	5:163	0393316281	3:29	eng	1995Do2010
6	2001		Marianne Mithun	4.53	2:0	5:21	052129875X	3:3	Ostalo	1995Do2010
1	2003		Elizabeth Seay	3.89	2:1	5:3	1592281958	3:2	Ostalo	1995Do2010
2	1969		William Tomkins	3.95	2:5	5:15	048622029X	3:7	Ostalo	Do1995
21	2014		Nikola Tesla	4.05	2:15	5:107	0932813194	3:42	Ostalo	2010Nadalje

Slika 37: Tablični prikaz promjene statističkog skupa uvođenjem novih varijabli *Starost* i *Language1*

Nakon uvođenja novih varijabli možemo na slici 37 vidjeti u nekoliko zapisa kakva je razlika između izvornih varijabli i novokreiranih varijabli.

7. Zadatak f)

Ispitajte postoji li razlika u varijabli Rating u ovisnosti o varijabli Starost. Koristite odgovarajući parametarski test. Ispitajte pretpostavke za njegovu primjenu. Grafički prikažite sredine korištenjem box-plota. Ukoliko postoje razlike između grupa provedite i odgovarajuće post hoc testove. Testirajte i pomoću odgovarajućeg neparametarskog testa.

Postavljamo hipoteze:

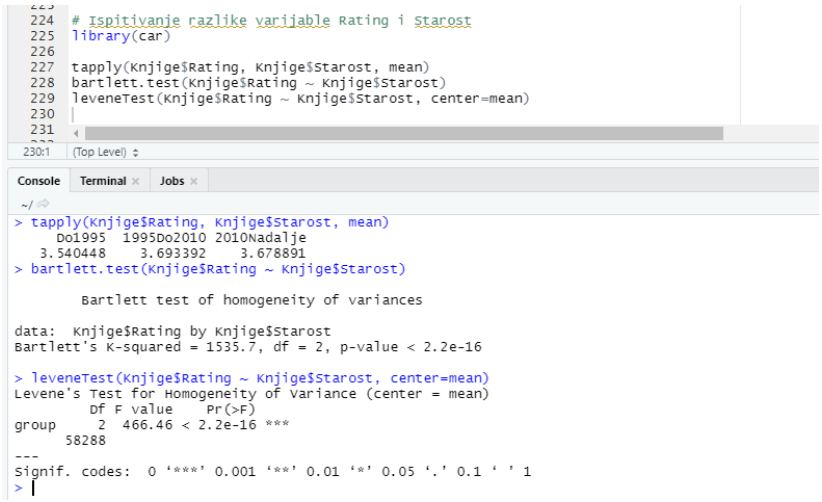
- H0 – ne postoji razlika u varijabli *Rating* u ovisnosti o varijabli *Starosti*
- H1 – postoji razlika u varijabli *Rating* u ovisnosti o varijabli *Starosti*

Prvo smo izvršili funkciju *tapply* u ovisnosti *Rating-a* o *Starosti* i utvrdili smo kako je prosječan *Rating* za knjigu iz kategorije Do1995 3.54, za knjigu iz kategorije 1995Do2010 iznosi 3.69, a za knjigu iz kategorije 2010Nadalje 3.68.

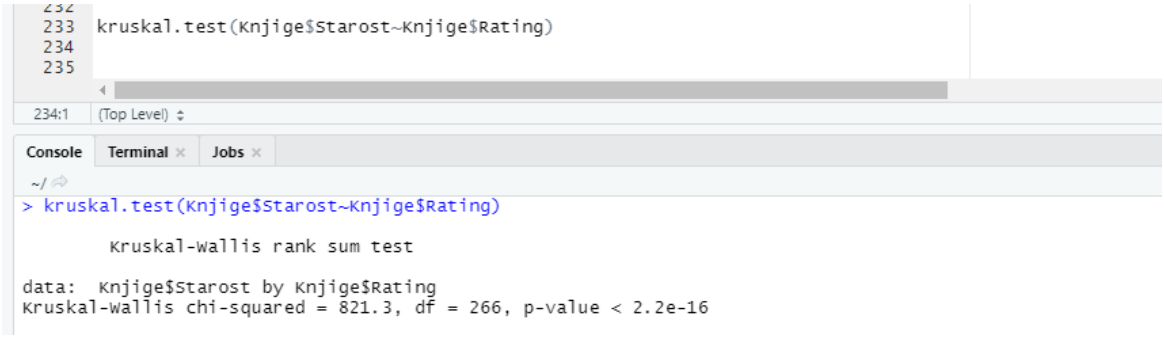
Kako smo već radnije utvrdili da nam varijable ne dolaze iz normalno distribuiranog skupa trebamo koristiti parametarske testove koji nisu osjetljivi na nenormalnosti u uzorcima. Iako je Bartlett-ov test poznat po tome da je osjetljiv na nenormalnost ipak je proveden kako bi se vidjeli rezultati da ih se kasnije može usporediti s ostalima.

Izvršavanjem Bartlett-ovog test dobili smo kako je *p-vrijednost* manja od 0.05 što dovodi do zaključka kako podaci odbacujemo H0 i prihvaćamo H1.

Nakon Bartlett-ovog testa proveden je parametarski Levene-ov test te neparametarski Kruskal-Wallis-ov test kojima smo utvrdili da je *p-vrijednost* manja od 0.05 čime odbacujemo hipotezu H0 i prihvaćamo H1.



Slika 38: Parameremetarski testovi varijabli *Rating* i *Starost*



Slika 39: Prikaz provedbe Kruskal-Wallis testna nad varijablama *Rating* i *Starost*

8. Zadatak g)

Ispitajte postoji li razlika u varijabli CountOfReview u ovisnosti o varijabli Language1. Koristite odgovarajući parametarski test. Ispitajte pretpostavke za njegovu primjenu. Grafički prikažite sredine korištenjem box-plota. Ukoliko postoje razlike između grupa provedite i odgovarajuće post hoc testove. Testirajte i pomoću odgovarajućeg neparametarskog testa.

Postavljamo hipoteze:

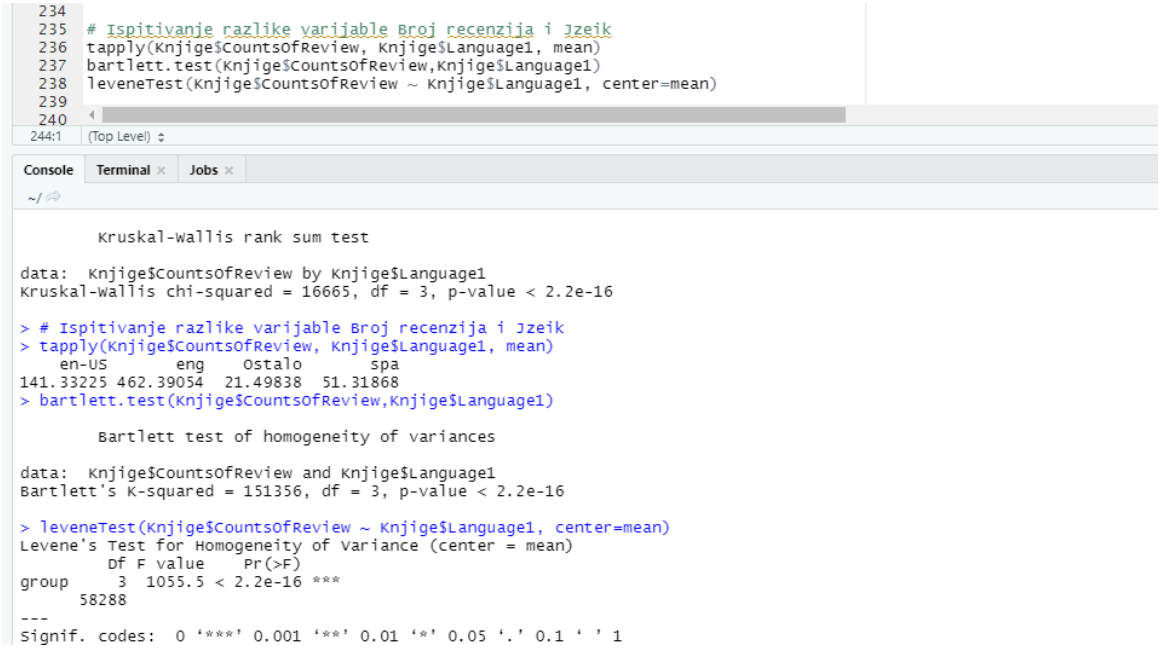
- H0 – ne postoji razlika u varijabli *CountOfReview* u ovisnosti o varijabli *Language1*
- H1 – postoji razlika u varijabli *CountOfReview* u ovisnosti o varijabli *Language1*

Prvo smo izvršili funkciju *tapply* u ovisnosti *CountOfReview-a* o *Languag1* i utvrdili smo kako je prosječan *CountOfReview-a* za knjigu napisanu na en-US 141.33, za knjigu napisanu na eng 462.39, za knjigu napisanu na spa 51.32, dok je za knjige na svim ostalim jezicima 21.50.

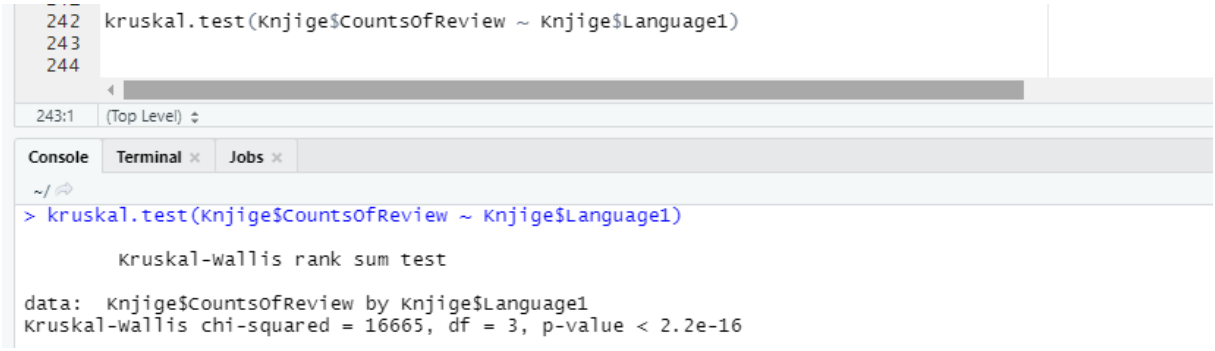
Kako smo već radnije utvrdili da nam varijable ne dolaze iz normalno distribuiranog skupa trebamo koristiti parametarske testove koji nisu osjetljivi na nenormalnosti u uzorcima. Iako je Bartlett-ov test poznat po tome da je osjetljiv na nenromalnost ipak je proveden kako bi se vidjeli rezultati da ih se kasnije može usporediti s ostalima.

Izvršavanjem Bartlett-ovog test dobili smo kako je *p-vrijednost* manja od 0.05 što dovodi do zaključka kako podaci odbacujemo H0 i prihvaćamo H1.

Nakon Bartlett-ovog testa proveden je parametarski Levene-ov test te neparametarski Kruskal-Wallis-ov test kojima smo utvrdili da je *p-vrijednost* manja od 0.05 čime odbacujemo hipotezu H0 i prihvaćamo H1.



Slika 40: Paramerametarski testovi varijabli *CountOfReview* i *Language1*



Slika 41: Prikaz provedbe Kruskal-Wallis testna nad varijablama *CountOfReview* i *Language1*

9. Zadatak h)

Definirajte model regresije kod kojeg će zavisna varijabla biti Rating, a nezavisne kvantitativne varijable po izboru te varijable Starost i Language. Komentirajte parametre regresije: koeficijent determinacije i korigirani koeficijent determinacije. Interpretirajte skupni i pojedinačne testove signifikantnosti regresije za svaku od promatranih nezavisnih varijabli. Interpretirajte koeficijente u jednadžbi regresije. Provedite izbor varijabli koristeći neku od metoda za izbor varijabli. Nacrtajte normalni prikaz rezidualnih vrijednosti za definirani regresijski model i interpretirajte ga.

Provođenjem regresijske analize (skup je ograničen na 500 slogova zbog velike količine podataka) dobivamo rezultate prikazane na slici 42. Koeficijent determinacije iznosi 0.04109, a korigirani koeficijent determinacije iznosi 0.02148. To nam govori kako je model nije dobar. Da bi model proglasili dobrim, koeficijent determinacije bi trebao biti preko 0.72.

Nezavisne varijable analiziraju se preko koeficijenta *p*. Što je vrijednost *p* manje, to je utjecaj veći. U slučaju nezavisnih varijabli prikazanih na slici 42, jedino pagesNumber i Languge(eng) značajnije utječu na zavisnu varijablu Rating, dok je CountOfReview dosta blizu, a ostale nezavisne varijable nemaju prevelik utjecaj.

```
Call:
lm(formula = Rating[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500] +
    Starost[1:500] + Language[1:500])

Residuals:
    Min      1Q  Median      3Q     Max
-3.8000 -0.1300  0.0519  0.2639  1.2537

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.793e+00  6.849e-02  55.383  < 2e-16 ***
pagesNumber[1:500]         6.913e-05  2.594e-05   2.665  0.00795 **
CountsOfReview[1:500]      1.144e-05  6.279e-06   1.821  0.06916 .
Starost[1:500]1995Do2010  -5.171e-02  6.804e-02  -0.760  0.44764
Starost[1:500]2010Nadalje  2.358e-01  2.887e-01   0.817  0.41454
Language[1:500]en-GB      -1.347e-01  4.489e-01  -0.300  0.76428
Language[1:500]en-US       2.086e-02  1.258e-01   0.166  0.86843
Language[1:500]eng         1.392e-01  6.070e-02   2.293  0.02228 *
Language[1:500]fre         3.095e-01  4.490e-01   0.689  0.49086
Language[1:500]mul         5.828e-02  6.330e-01   0.092  0.92668
Language[1:500]spa         2.323e-01  3.683e-01   0.631  0.52843
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6316 on 489 degrees of freedom
Multiple R-squared:  0.04109,   Adjusted R-squared:  0.02148
F-statistic: 2.095 on 10 and 489 DF,  p-value: 0.02341
```

Slika 42: Prikaz regresijske analize

```
> step1 <- stepAIC(RegModelRating, direction="both")
Start:  AIC=-456
Rating[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500]

                       Df Sum of Sq    RSS     AIC
<none>                              198.46 -456.00
- CountsOfReview[1:500]  1    2.0418 200.51 -452.88
- pagesNumber[1:500]     1    2.8253 201.29 -450.93
> summary(step1)

Call:
lm(formula = Rating[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500])

Residuals:
    Min      1Q  Median      3Q     Max
-3.8677 -0.1341  0.0769  0.2685  1.1793

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.816e+00  3.013e-02 126.656  < 2e-16 ***
pagesNumber[1:500]   6.891e-05  2.591e-05   2.660  0.00807 **
CountsOfReview[1:500] 1.388e-05  6.139e-06   2.261  0.02418 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6319 on 497 degrees of freedom
Multiple R-squared:  0.0245,    Adjusted R-squared:  0.02058
F-statistic: 6.242 on 2 and 497 DF,  p-value: 0.002103
```

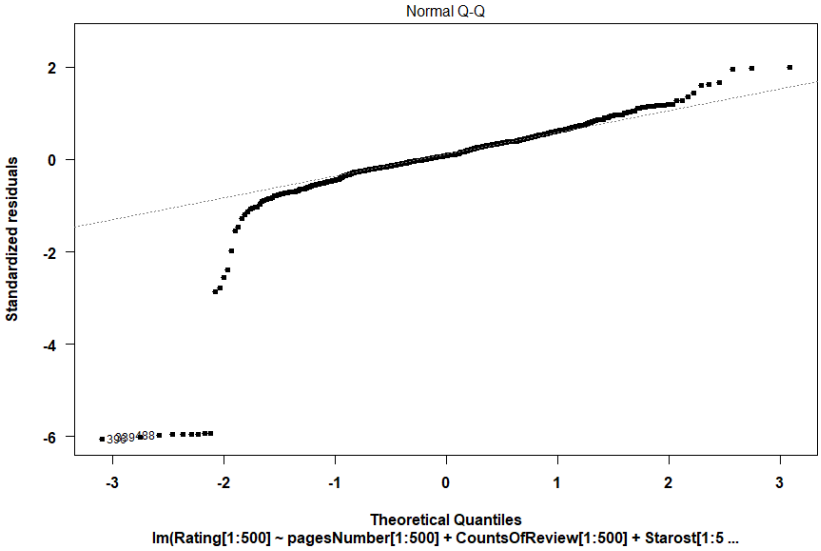
Slika 43: Izbor varijabli direction=“both“

```
> step2 <- stepAIC(RegModelRating, direction="backward")
Start:  AIC=-456
Rating[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500]

                       Df Sum of Sq    RSS     AIC
<none>                              198.46 -456.00
- CountsOfReview[1:500]  1    2.0418 200.51 -452.88
- pagesNumber[1:500]     1    2.8253 201.29 -450.93
>
```

Slika 44: Izbor varijabli direction=“backward“

```
> step3 <- stepAIC(RegModelRating, direction="forward")
Start:  AIC=-456
Rating[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500]

>
```

Slika 45: Izbor varijabli direction=“forward“



Slika 46: Prikaz rezidualnih vrijednosti

```
> shapiro.test(residuals(RegModelRating))

        Shapiro-Wilk normality test

data:  residuals(RegModelRating)
W = 0.63876, p-value < 2.2e-16
```

Slika 47: Provedba Shapiro-Wilk testa

Slika 46 prikazuje rezidualne vrijednosti prikazane na grafu za ranije navedeni regresijski model, te iz grafa možemo zaključiti kako nemamo normalnost reziduala. Isto tako, slikom 47, gdje je proveden Shapiro-Wilk test i gdje je p-vrijednost manja od 0.05 vidimo kako nema normalnosti reziduala.

10. Zadatak i)

Definirajte model regresije kod kojeg će zavisna varijabla biti RatingDistTotal, a nezavisne varijable izaberite po svom izboru. Komentirajte parametre regresije: koeficijent determinacije i korigirani koeficijent determinacije. Interpretirajte skupni i pojedinačne testove signifikantnosti regresije za svaku od promatranih nezavisnih varijabli. Interpretirajte koeficijente u jednadžbi regresije. Provedite izbor varijabli koristeći neku od metoda za izbor varijabli. Nacrtajte normalni prikaz rezidualnih vrijednosti za definirani regresijski model i interpretirajte ga.

Provođenjem regresijske analize (skup je ograničen na 500 slogova zbog velike količine podataka) dobivamo rezultate prikazane na slici 48. Koeficijent determinacije iznosi 0.7362, a korigirani koeficijent determinacije iznosi 0.7308. To nam govori kako je model dobar. Da bi model proglasili dobrim, koeficijent determinacije bi trebao biti preko 0.72.

Nezavisne varijable analiziraju se preko koeficijenta *p*. Što je vrijednost *p* manje, to je utjecaj veći. U slučaju nezavisnih varijabli prikazanih na slici 48, jedino CountOfReview i Languge(spa) značajnije utječu na zavisnu varijablu RatingDistTotal, a ostale nezavisne varijable nemaju prevelik utjecaj.

```
Call:
lm(formula = ratingTotal[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500] +
    Starost[1:500] + Language[1:500])

Residuals:
    Min      1Q  Median      3Q     Max
-854938  -35853  -10757   -9952 2505782

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -19484.304  24163.214  -0.806    0.420
pagesNumber[1:500]         -1.895      9.151  -0.207    0.836
CountsOfReview[1:500]      78.563      2.215  35.467  < 2e-16 ***
Starost[1:500]1995Do2010 30306.094  24004.877   1.262    0.207
Starost[1:500]2010Nadalje  -796.434 101863.990  -0.008    0.994
Language[1:500]en-GB      -7326.183 158368.972  -0.046    0.963
Language[1:500]en-US       7737.656  44392.953   0.174    0.862
Language[1:500]eng        25551.057  21413.075   1.193    0.233
Language[1:500]fre         3798.963 158392.323   0.024    0.981
Language[1:500]mul       -11833.420 223318.468  -0.053    0.958
Language[1:500]spa       679810.916 129928.912   5.232 2.49e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 222800 on 489 degrees of freedom
Multiple R-squared:  0.7362,    Adjusted R-squared:  0.7308
F-statistic: 136.5 on 10 and 489 DF,  p-value: < 2.2e-16
```

Slika 48: Prikaz regresijske analize

```
> step1 <- stepAIC(RegModelRatingDistTotal, direction="both")
Start:  AIC=12325.05
ratingTotal[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500] +
    Starost[1:500] + Language[1:500]

                        Df  Sum of Sq        RSS   AIC
- Starost[1:500]         2 8.1868e+10 2.4363e+13 12323
- pagesNumber[1:500]     1 2.1282e+09 2.4283e+13 12323
<none>                               2.4281e+13 12325
- Language[1:500]        6 1.3924e+12 2.5673e+13 12341
- CountsOfReview[1:500]  1 6.2460e+13 8.6741e+13 12960

Step:  AIC=12322.74
ratingTotal[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500] +
    Language[1:500]

                        Df  Sum of Sq        RSS   AIC
- pagesNumber[1:500]     1 1.0036e+09 2.4364e+13 12321
<none>                               2.4363e+13 12323
+ Starost[1:500]         2 8.1868e+10 2.4281e+13 12325
- Language[1:500]        6 1.3384e+12 2.5701e+13 12338
- CountsOfReview[1:500]  1 6.3304e+13 8.7667e+13 12961

Step:  AIC=12320.76
ratingTotal[1:500] ~ CountsOfReview[1:500] + Language[1:500]

                        Df  Sum of Sq        RSS   AIC
<none>                               2.4364e+13 12321
+ pagesNumber[1:500]     1 1.0036e+09 2.4363e+13 12323
+ Starost[1:500]         2 8.0744e+10 2.4283e+13 12323
- Language[1:500]        6 1.3386e+12 2.5702e+13 12336
- CountsOfReview[1:500]  1 6.3325e+13 8.7688e+13 12959
> summary(step1)

Call:
lm(formula = ratingTotal[1:500] ~ CountsOfReview[1:500] + Language[1:500])

Residuals:
    Min      1Q  Median      3Q     Max
-850299  -27442   -5790   -4464 2515391

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            4463.715  13908.227   0.321    0.748
CountsOfReview[1:500]    78.798      2.204  35.760  < 2e-16 ***
Language[1:500]en-GB  -16713.391 157967.128  -0.106    0.916
Language[1:500]en-US    5813.126  44294.908   0.131    0.896
Language[1:500]eng     21575.509  21160.299   1.020    0.308
Language[1:500]fre     -6172.262 157966.529  -0.039    0.969
Language[1:500]mul     -6572.252 222965.045  -0.029    0.976
Language[1:500]spa    665380.463 129228.948   5.149  3.8e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 222500 on 492 degrees of freedom
Multiple R-squared:  0.7353,    Adjusted R-squared:  0.7315
F-statistic: 195.3 on 7 and 492 DF,  p-value: < 2.2e-16
```

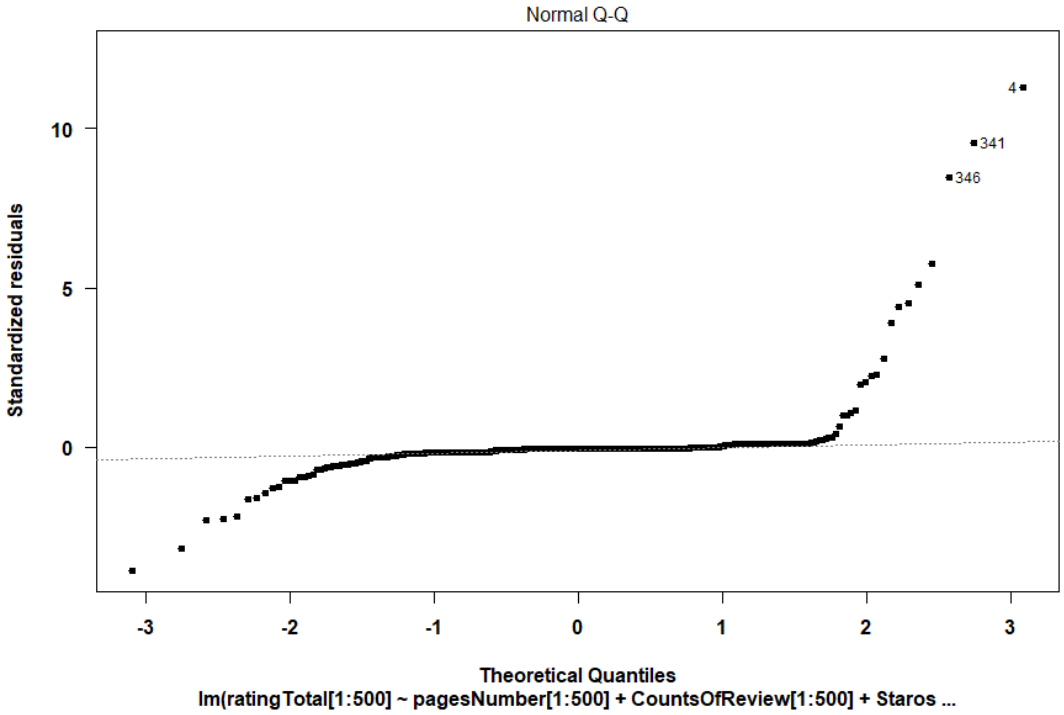
Slika 49: Izbor varijabli direction=“both“

```
> step2 <- stepAIC(RegModelRatingDistTotal, direction="backward")
Start:  AIC=12325.05
ratingTotal[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500] +
    Starost[1:500] + Language[1:500]

                        Df  Sum of Sq        RSS   AIC
- Starost[1:500]         2 8.1868e+10 2.4363e+13 12323
- pagesNumber[1:500]     1 2.1282e+09 2.4283e+13 12323
<none>                               2.4281e+13 12325
- Language[1:500]        6 1.3924e+12 2.5673e+13 12341
- CountsOfReview[1:500]  1 6.2460e+13 8.6741e+13 12960

Step:  AIC=12322.74
ratingTotal[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500] +
    Language[1:500]

                        Df  Sum of Sq        RSS   AIC
- pagesNumber[1:500]     1 1.0036e+09 2.4364e+13 12321
<none>                               2.4363e+13 12323
- Language[1:500]        6 1.3384e+12 2.5701e+13 12338
- CountsOfReview[1:500]  1 6.3304e+13 8.7667e+13 12961

Step:  AIC=12320.76
ratingTotal[1:500] ~ CountsOfReview[1:500] + Language[1:500]

                        Df  Sum of Sq        RSS   AIC
<none>                               2.4364e+13 12321
- Language[1:500]        6 1.3386e+12 2.5702e+13 12336
- CountsOfReview[1:500]  1 6.3325e+13 8.7688e+13 12959
```

Slika 50: Izbor varijabli direction=“backward“

```
> step3 <- stepAIC(RegModelRatingDistTotal, direction="forward")
Start:  AIC=12325.05
ratingTotal[1:500] ~ pagesNumber[1:500] + CountsOfReview[1:500] +
    Starost[1:500] + Language[1:500]
```

Slika 51: Izbor varijabli direction="forward"



Slika 52: Prikaz rezidualnih vrijednosti

```
> shapiro.test(residuals(RegModelRatingDistTotal))

        Shapiro-Wilk normality test

data:  residuals(RegModelRatingDistTotal)
W = 0.31801, p-value < 2.2e-16
```

Slika 53: Provedba Shapiro-Wilk testa

Slika 52 prikazuje rezidualne vrijednosti prikazane na grafu za ranije navedeni regresijski model, te iz grafa možemo zaključiti kako nemamo normalnost reziduala. Isto tako, slikom 53, gdje je proveden Shapiro-Wilk test i gdje je p-vrijednost manja od 0.05 vidimo kako nema normalnosti reziduala.

11. Zaključak

Ovaj seminarski rad bavio se temom analizom skupa podataka o knjigama i podacima vezanih za njihove ocjene, recenzije, vrijeme izdanja i drugi. Analiza skupa podataka obrađena je u programskom alatu R. U početku opisane su varijable statističkog skupa i grafički su prikazane, nadalje, izračunata je matrica korelacija između svih kvantitativnih varijabli, ispitana je normalnost razdiobe određenih varijabli te su definirane nove varijable Language1 i Starost.

Isto tako ispitane su određene razlike i ovisnosti između pojedinih varijabli su definirana dva modela regresije, gdje je u prvom zavisna varijabla bila Rating, a u drugom RatingDistTotal.

Nakon kreiranja nove varijable Starost, zaključujemo kako je najviše knjiga izdano u vremenskom periodu od 1995 do 2010, zatim do 1995, a od 2010 pa nadalje ima najmanji broj izdanih primjeraka. Zatim, kreiranjem nove varijable Languge1 utvrđujemo kako su knjige najviše napisane na ostalim jezicima, a nakon toga se ističu eng, eng-US i spa.

Izračunom matrice korelacije zaključujemo da snažnu pozitivnu korelaciju imaju *RatingDist3* i *PublishDay* koja iznosi 0.87 dok snažnu negativnu korelaciju imaju varijable *PublishYear* i *RatingDist3* koja iznosi -0.80.

Na kraju, možemo zaključiti kako skup podataka s kojim smo radili nije normalno distribuiran, a regresijskim modelom utvrdili smo kako je za zavisnu varijablu Rating model bio dobar, dok za RatingDistTotal nije, ali za oba modela nema normalnosti reziduala.

Popis literature

[1] Jannesar, B., Goodreads Book Datasets With User Rating 10M, 2021. [Na internetu]. Dostupno: https://www.kaggle.com/bahramjannesarr/goodreads-book-datasets-10m [pristupano 03.06.2021.]

Popis slika

Popis tablica

Prilozi

[1] GitHub repozitorij projekta: https://github.com/bzitkovic/obrada-podataka-R