

Analiza korisničkih ocjena sadržaja

Slučajni Šumari

Sadržaj

1	Uvod	2
2	Učitavanje podataka i deskriptivna statistika	3
3	Komparativna analiza kategorija ocjenjivanja i inferencijalna statistika	10
4	Linearna regresija	36
5	Komprativna analiza korisnika	43
6	Zaključak	46

1 Uvod

U današnje vrijeme, kada su ljudi potpuno umreženi i kada tehnologija napreduje svakodnevno, informacija je postala najjače oružje u svijetu. Velikim korporacijama poput Amazona, Googlea, Applea i sličnih cilj je što više doprijeti do korisnika i što uspješnije predvidjeti njihove potrebe i preferencije. Kako bi što uspješnije predvidjeli hoće li se određeni proizvod svidjeti određenom dijelu populacije te kako bi odredili koji dio populacije bi mogao biti potencijalni kupac njihovih usluga, navedene kompanije rade analize ponašanja korisnika na internetu na temelju stranica i sadržaja koje oni gledaju, kupuju, ocjenjuju itd. U našem projektu, analizirali smo skup korisničkih ocjena raznog sadržaja. U našem se podatkovnom skupu nalaze ocjene 5456 korisnika tražilice Google. Svaki korisnik ocjenjivao je najviše 25 kategorija. Ocjene za pojedine kategorije poprimaju vrijednost iz intervala $[0, 5]$ pri čemu ocjena 0 označava da korisnik nije ocijenio tu kategoriju. U projektu smo usporedivali odabранe kategorije po ocjenama, analizirali koliko se razlikuju, odredili koje su kategorije najviše polarizirajuće te na kraju pokušali predvidjeti korisničke ocjene na temelju ocjena drugih kategorija, ali i na temelju ocjena drugih korisnika. Zbog čega nam je uopće zanimljiva ovakva analiza? Jasno je da ovakva analiza ne bi imala smisla kad bismo svi bili jednaki. Različitost korisničkih preferencija u kojoj je ipak moguće pronaći neke pravilnosti temelj je sustava preporučivanja i personaliziranog oglašavanja. Važno je naglasiti da ovo nije napredna analiza te je cilj ovog projekta upoznati se s metodama statističkog zaključivanja. Naravno, ovakva analiza nije nimalo beskorisna jer je upravo ona bitan korak u gradnji naprednih algoritama i sustava koji se koriste za preporučivanje u komercijalne svrhe.

2 Učitavanje podataka i deskriptivna statistika

```
df <- read_csv("data.csv")

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   User = col_character()
## )

## See spec(...) for full column specifications.
df_original <- df

Ocjene korisnika koje iznose 0 potrebno je elminirati s obzirom na to da one označavaju da korisnik nije ocjenjivao tu kategoriju. Zbog toga su te ocjene zamijenjene s NA.

df[df == 0] <- NA
head(df)

## # A tibble: 6 x 25
##   User  churches resorts beaches parks theatres museums malls   zoo
##   <chr>    <dbl>   <dbl>   <dbl> <dbl>    <dbl>   <dbl> <dbl> <dbl>
## 1 User~     NA     NA     3.63  3.65      5    2.92     5  2.35
## 2 User~     NA     NA     3.63  3.65      5    2.92     5  2.64
## 3 User~     NA     NA     3.63  3.63      5    2.92     5  2.64
## 4 User~     NA     0.5    3.63  3.63      5    2.92     5  2.35
## 5 User~     NA     NA     3.63  3.63      5    2.92     5  2.64
## 6 User~     NA     NA     3.63  3.63      5    2.92     5  2.63
## # ... with 16 more variables: restaurants <dbl>, `pubs/bars` <dbl>, `local
## #   services` <dbl>, `burger/pizza shops` <dbl>, `hotels/other`
## #   lodgings` <dbl>, `juice bars` <dbl>, `art galleries` <dbl>, `dance
## #   clubs` <dbl>, `swimming pools` <dbl>, gyms <dbl>, bakeries <dbl>,
## #   `beauty & spas` <dbl>, cafes <dbl>, `view points` <dbl>,
## #   monuments <dbl>, gardens <dbl>
glimpse(df)

## Observations: 5,456
## Variables: 25
## $ User          <chr> "User 1", "User 2", "User 3", "User 4"...
## $ churches      <dbl> NA, NA, NA, NA, NA, NA, NA, NA...
## $ resorts        <dbl> NA, NA, NA, 0.50, NA, NA, 5.00, 5.00, ...
## $ beaches        <dbl> 3.63, 3.63, 3.63, 3.63, 3.63, 3.63, 3....
## $ parks          <dbl> 3.65, 3.65, 3.63, 3.63, 3.63, 3.63, 3....
## $ theatres       <dbl> 5.00, 5.00, 5.00, 5.00, 5.00, 5.00, 5....
## $ museums        <dbl> 2.92, 2.92, 2.92, 2.92, 2.92, 2.92, 2....
## $ malls          <dbl> 5.00, 5.00, 5.00, 5.00, 5.00, 5.00, 3....
## $ zoo            <dbl> 2.35, 2.64, 2.64, 2.35, 2.64, 2.63, 2....
## $ restaurants     <dbl> 2.33, 2.33, 2.33, 2.33, 2.33, 2.33, 2....
## $ `pubs/bars`    <dbl> 2.64, 2.65, 2.64, 2.64, 2.64, 2.65, 2....
## $ `local services` <dbl> 1.70, 1.70, 1.70, 1.73, 1.70, 1.71, 1....
## $ `burger/pizza shops` <dbl> 1.69, 1.69, 1.69, 1.69, 1.69, 1.69, 1....
## $ `hotels/other lodgings` <dbl> 1.70, 1.70, 1.70, 1.70, 1.70, 1.69, 1....
## $ `juice bars`   <dbl> 1.72, 1.72, 1.72, 1.72, 1.72, 1.72, 1....
## $ `art galleries` <dbl> 1.74, 1.74, 1.74, 1.74, 1.74, 1.74, 1....
## $ `dance clubs`  <dbl> 0.59, 0.59, 0.59, 0.59, 0.59, 0.59, 0....
```

```

## $ `swimming pools`      <dbl> 0.50, 0.50, 0.50, 0.50, 0.50, 0....
## $ gyms                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ bakeries               <dbl> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5...
## $ `beauty & spas`        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ cafes                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ `view points`          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ monuments              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ gardens                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...

```

Deskriptivna statistika nam daje informacije o srednjim vrijednostima ocjena po kategorijama, njihovoj raspršenosti i općenitoj prirodi raspodjele opservacija u uzorku. Spoznaje dobivene naredbom `summary(df)` vizualizirane su boxplotovima, histogramima i density plotovima prikazanima na slikama 2.1, 2.2, 2.3.

```

ggplot(melt(df), aes(x="", y=value)) +
  stat_boxplot(geom='errorbar', linetype = 1, width = 0.5) +
  geom_boxplot(fill = "steelblue", color = "steelblue4") +
  facet_wrap(~variable, ncol = 4, nrow = 6) +
  theme_minimal() +
  theme(strip.text.x = element_text(size = 12)) +
  xlab("") +
  ylab("")

## Using User as id variables

ggplot(melt(df), aes(x = value)) +
  geom_histogram(binwidth = 0.25, fill = "steelblue",
                 color = "steelblue4", aes(label=..count..)) +
  facet_wrap(~variable, ncol = 4, nrow = 6) + theme_minimal() +
  theme(strip.text.x = element_text(size = 12))

## Using User as id variables

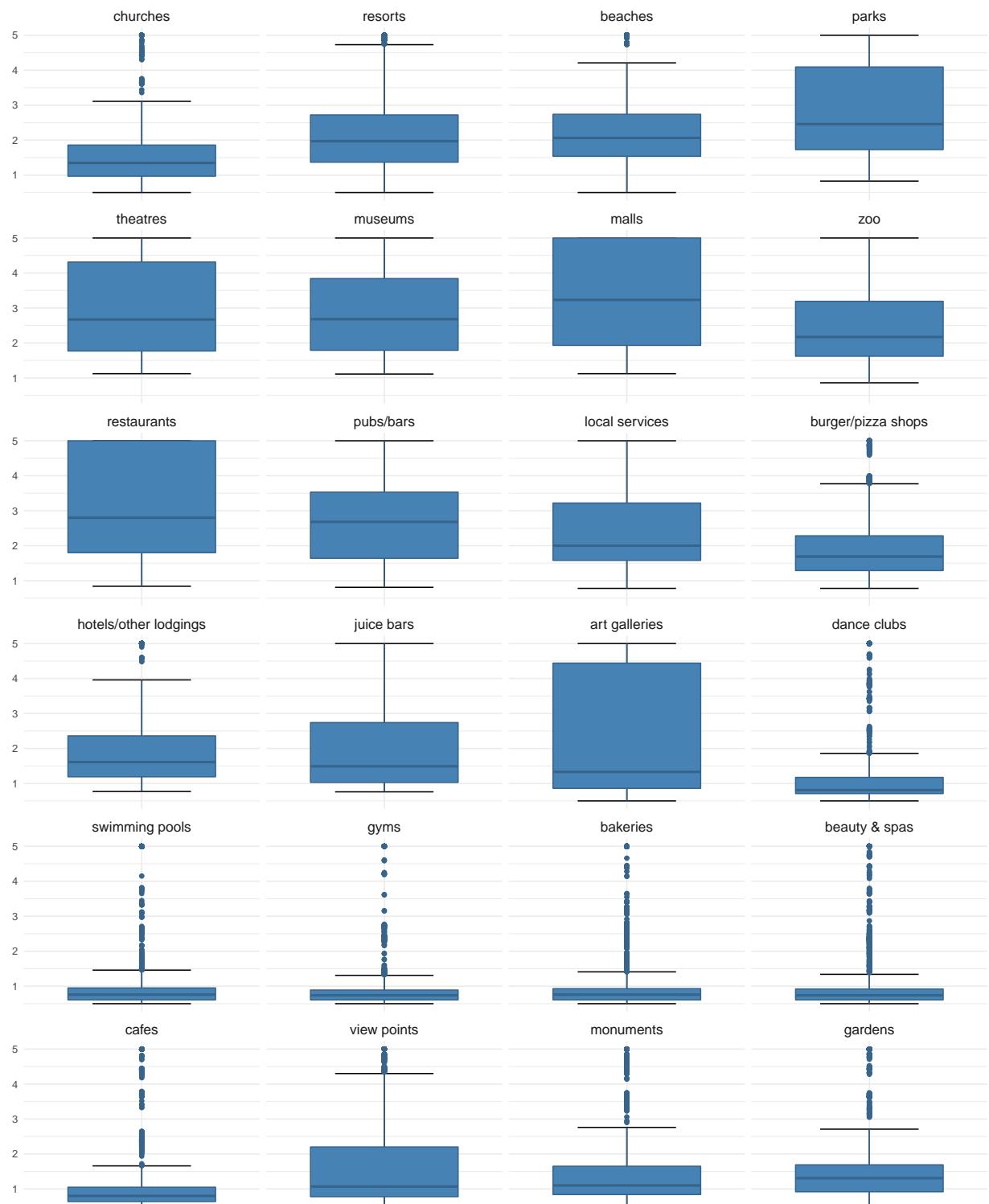
ggplot(melt(df), aes(x = value)) +
  geom_density(fill = "steelblue", color = "steelblue4", alpha = 0.5) +
  facet_wrap(~variable, ncol = 4, nrow = 6) + theme_minimal() +
  theme(strip.text.x = element_text(size = 12))

## Using User as id variables

ggplot(melt(df), aes(sample = value)) +
  stat_qq(color = "steelblue", shape=1) +
  facet_wrap(~variable, ncol = 4, nrow = 6) + theme_minimal() +
  theme(strip.text.x = element_text(size = 12))

## Using User as id variables

```



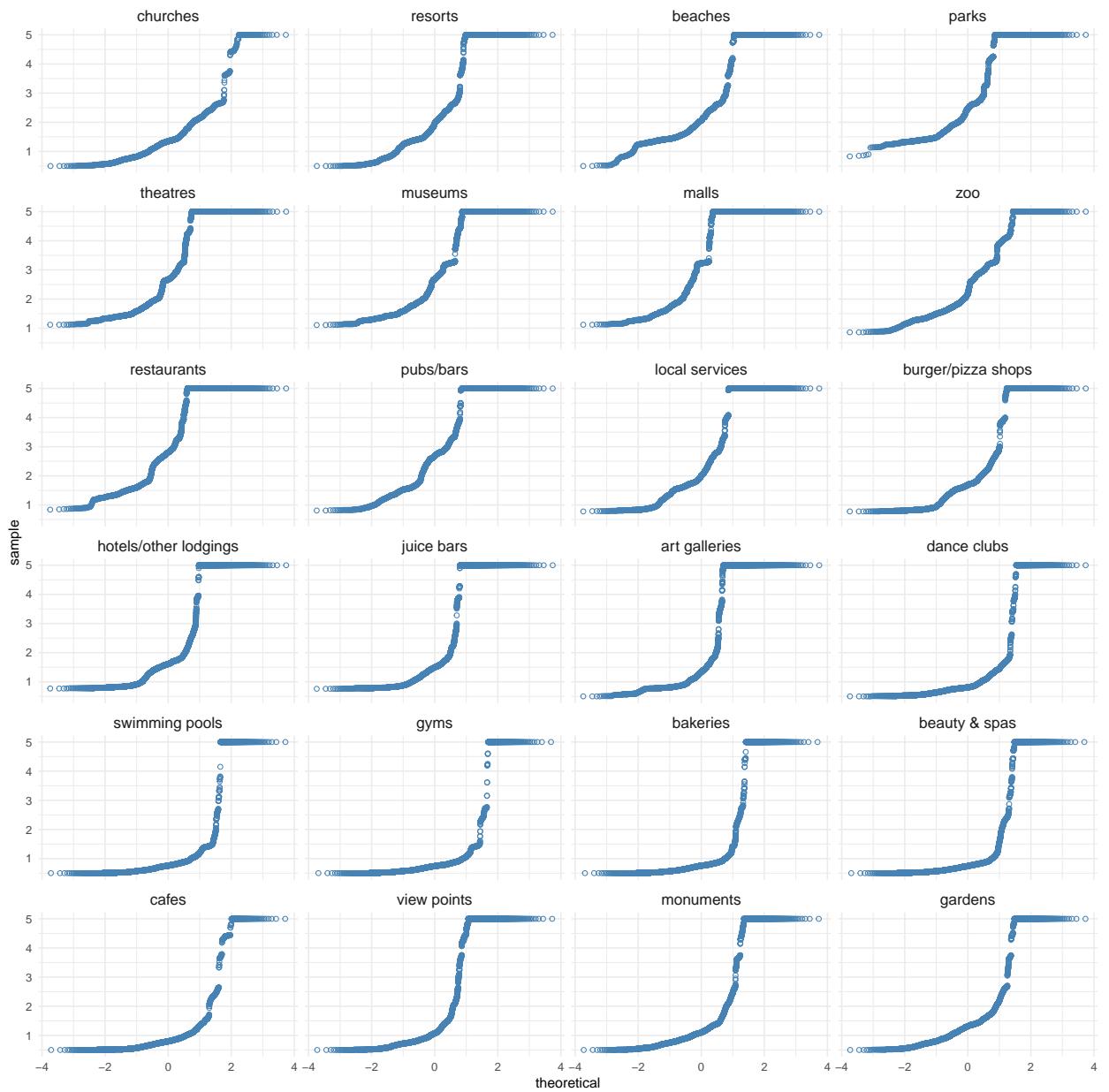
Slika 2.1: Boxplotovi ocjena u svakoj kategoriji



Slika 2.2: Histogrami ocjena u svakoj kategoriji



Slika 2.3: Density plotovi ocjena u svakoj kategoriji



Slika 2.4: QQ-plotovi ocjena u svakoj kategoriji

Iz QQ-plotova prikazanih slikom 2.4 vidimo da niti jedna kategorija nema normalnu razdiobu svojih ocjena jer niti jedan QQ-plot nije ni približno linearne funkcije. Ovo nam može biti problem pri provođenju neparametarskih statističkih testova, međutim ovaj problem umanjuje činjenica da se u podatkovnom skupu nalazi dovoljno velik broj podataka.

3 Komparativna analiza kategorija ocjenjivanja i inferencijalna statistika

U ovom potpoglavlju pozabavit ćemo se komparativnom analizom kategorija ocjenjivanja. Drugim riječima pogledat ćemo postoji li povezanost između ocjena različitih kategorija. Sve pretpostavke bit će potvrđene, odnosno opovrgnute primjenom metoda inferencijalne statistike, odnosno odgovarajućim statističkim testovima.

```
corr <- cor(df[-1], use="pairwise.complete.obs")
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
ggcorrplot(corr, outline.color = "white", lab = T,
           colors = c("violetred4", "white", "steelblue4"),
           lab_size = 7,
           hc.order = T,
           tl.cex = 25) +
  theme(legend.key.size = unit(1.5, "cm"),
        legend.text = element_text(size = 20),
        legend.title = element_text(size = 20))
```

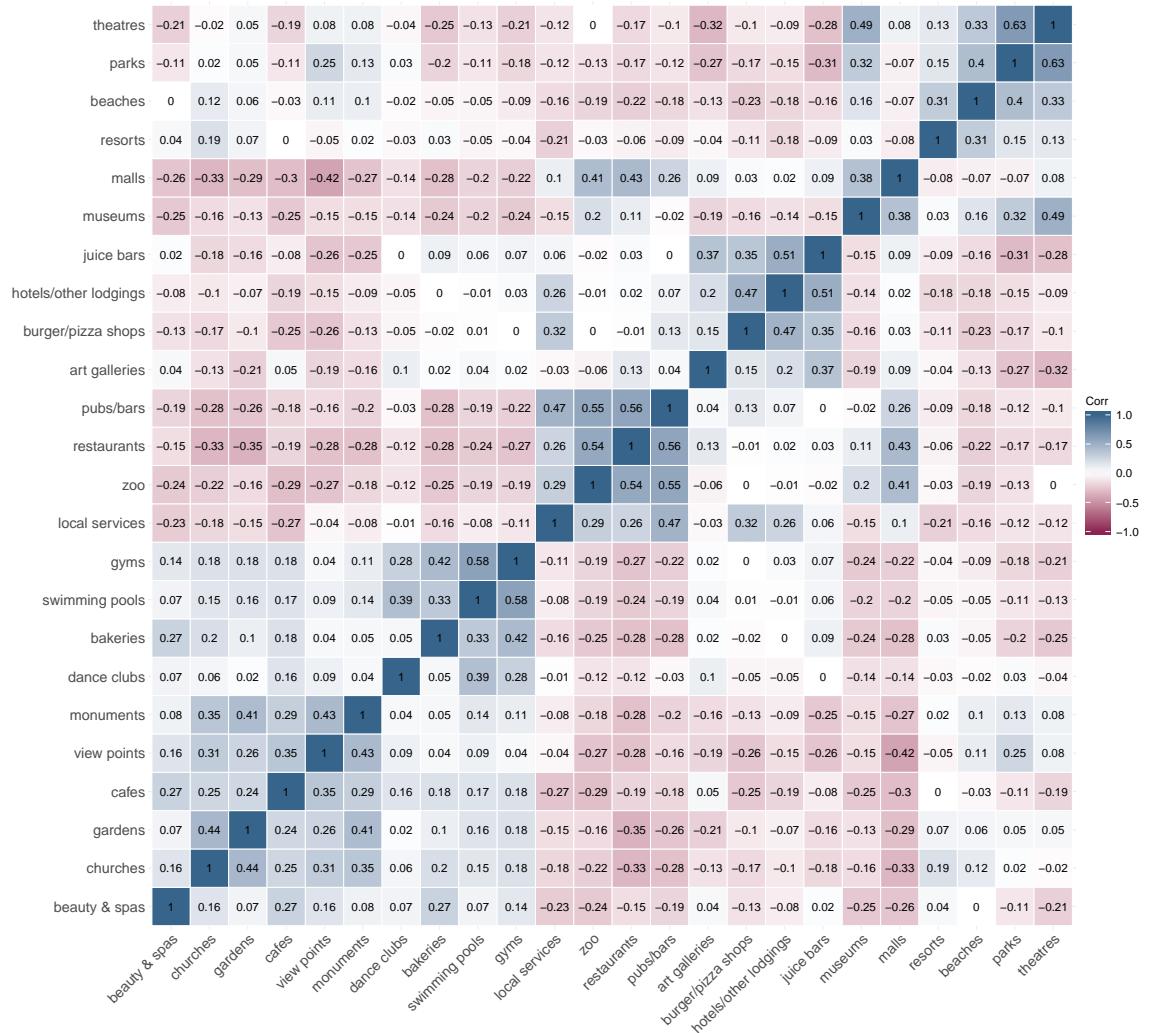
Na slici 3.1 prikazana je matrica korelacija svih parova kategorija ocjenjivanja. Za trenutak ćemo se suzdržati komentiranja korelacija između kategorija budući da u nastavku slijedi vizualno nešto atraktivniji i pregledniji prikaz korelacija parova kategorija: u obliku grafa.

```
corr %>%
  as.data.frame() %>%
  mutate(var1 = rownames(.)) %>%
  gather(var2, value, -var1) %>%
  arrange(desc(value)) %>%
  group_by(value) %>%
  dplyr::filter(row_number()==1 & var1 != var2) -> corrs

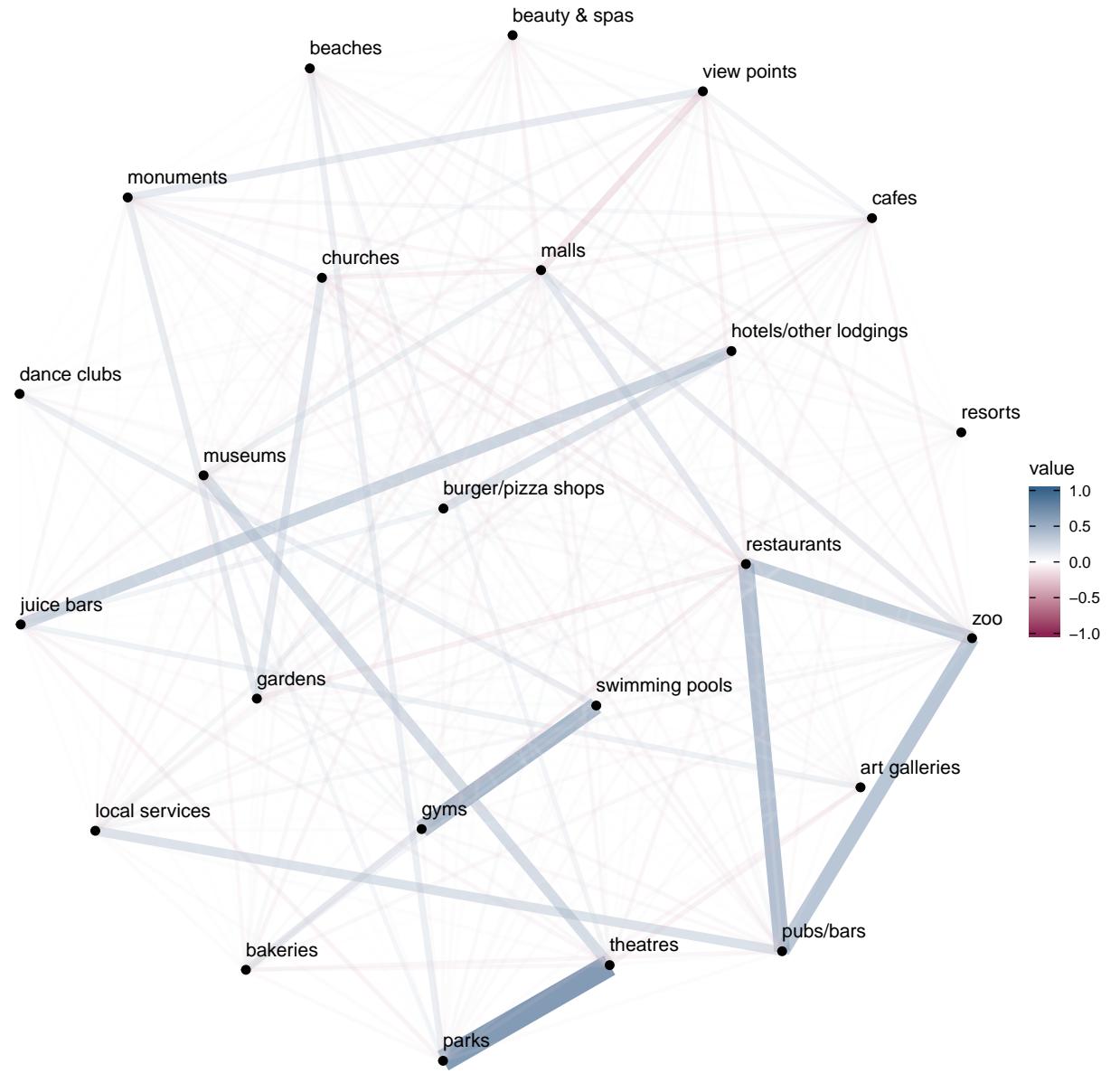
corrs %>% graph_from_data_frame -> corrs_graph

set.seed(100)
graph(corrs_graph, "nicely") +
  geom_edge_link(aes(edge_alpha = value^4, edge_width = value^4, color = value)) +
  geom_node_point(size = 2) +
  geom_node_text(aes(label = name), vjust = -1, hjust = 0) +
  guides(edge_alpha = "none", edge_width = "none") +
  scale_edge_colour_gradientn(limits = c(-1, 1),
                               colors = c("violetred4", "#FFFFFF", "steelblue4")) +
  theme_void()
```

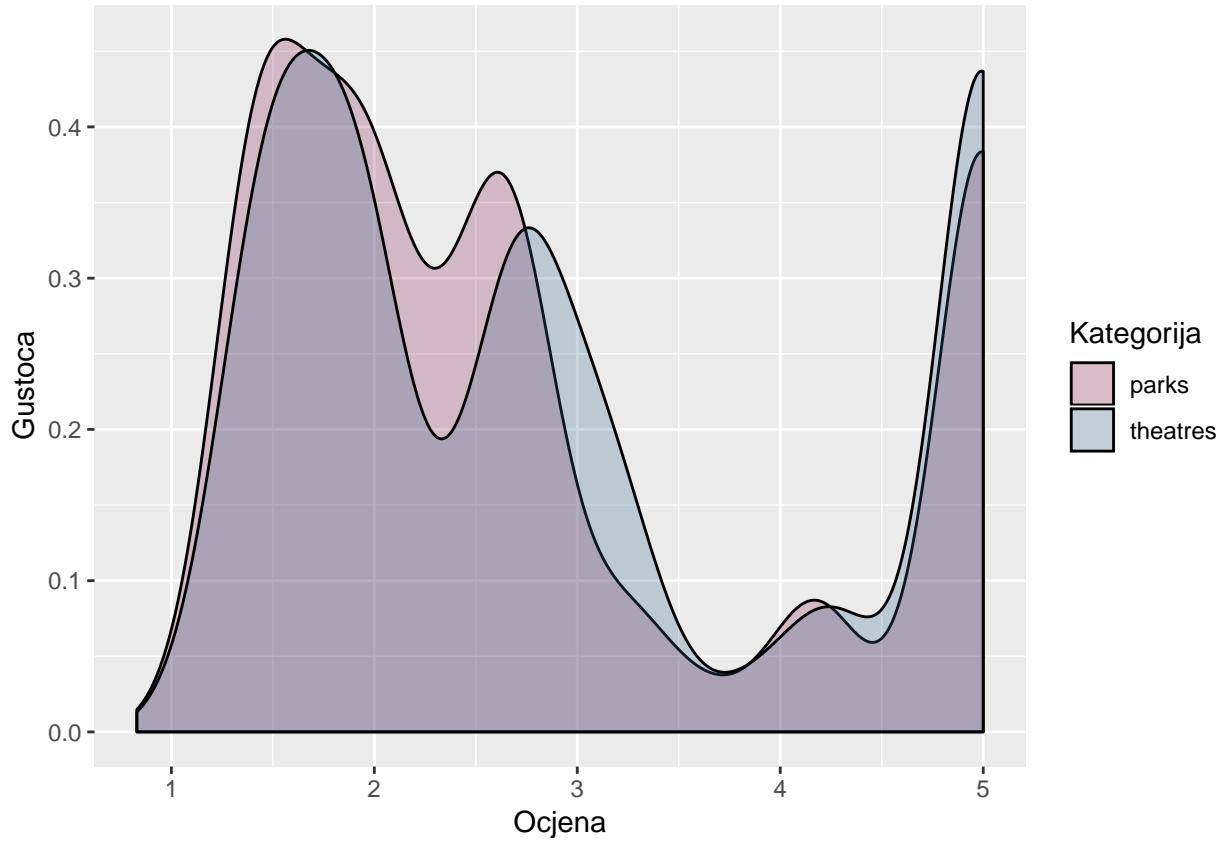
Gledajući graf na slici 3.2 odmah u oči upada poveznica između kategorija `parks` i `theatres` čija korelacija kao što možemo iščitati iz korelacijske matrice sa slike 3.1 iznosi 0.63. Vrlo je upadljiv i trokut čiji su vrhovi kategorije `restaurants`, `zoo` i `pubs/bars`. Korelacije parova `restaurants` i `zoo`, `zoo` i `pubs/bars` te `pubs/bars` i `restaurants` iznose redom: 0.54, 0.55 i 0.56. Relativno veliku pozitivnu korelaciju imaju i `juice bars` i `hotels/other lodgings`, `gyms` i `swimming pools` te `theatres` i `museum`. Pozitivne je korelacije lako uočiti, međutim, morat ćemo se malo bolje potruditi kako bismo uočili i negativne korelacije koje su na slici 3.2 prikazane crvenim linijama. Najveću negativnu korelaciju imaju kategorije `malls` i `view points` i ona iznosi -0.42, ostali parovi kategorija koji su negativno korelirani nemaju značajno veliku absolutnu vrijednost koeficijenta korelacija. Visoke korelacije navedenih parova kategorija ocjena u skladu su s pretpostavkama do kojih bismo mogli doći zdravim razumom i nema potrebe objašnjavati ih detaljno. U nastavku su za neke parove kategorija statističkim testovima provjerene jednakosti njihovih srednjih vrijednosti i podudarnosti njihovih razdioba. Prije statističkih testova uvijek će biti prikazani density plotovi, histogrami i boxplotovi kategorija koje se uspoređuju.



Slika 3.1: Matrica korelacija svih parova kategorija ocjenjivanja



Slika 3.2: Prikaz korelacija parova kategorija ocjenjivanja u obliku grafa



Slika 3.3: Density plotovi kategorija parks i theatres

```

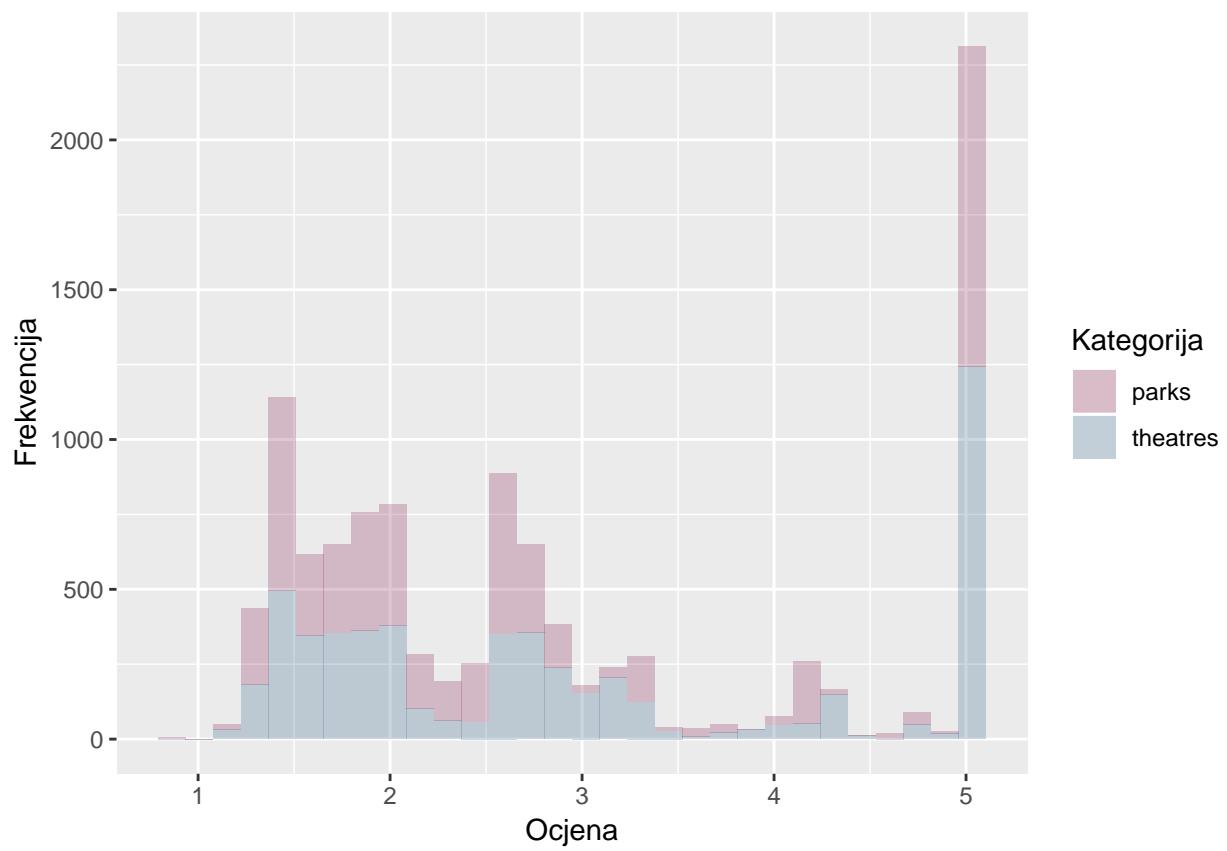
data <- melt(df[c("parks", "theatres")])

## No id variables; using all as measure variables
ggplot(data, aes(x=value, fill=variable)) +
  geom_density(alpha=0.25) +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Ocjena") + ylab("Gustoća")

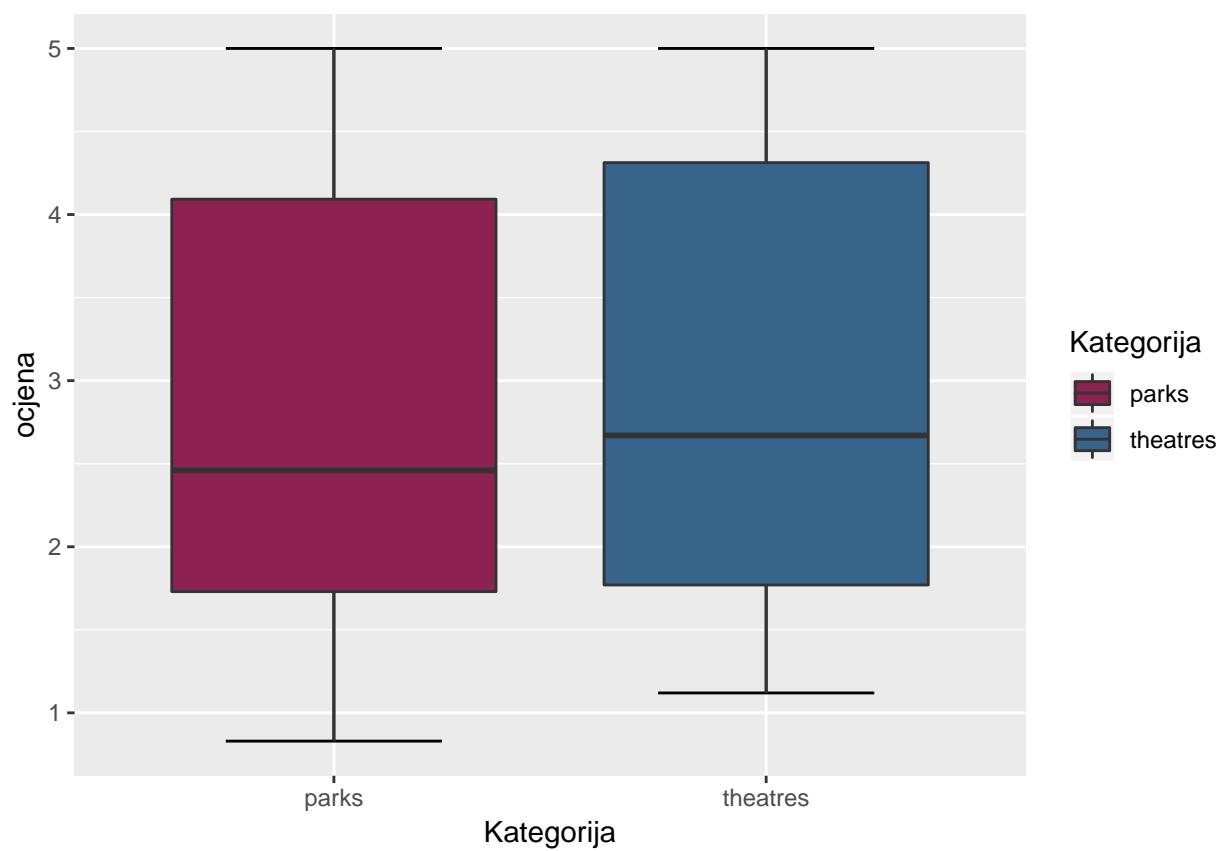
ggplot(data,aes(x=value, fill=variable)) +
  geom_histogram(alpha=0.25) +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Ocjena") + ylab("Frekvencija")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
ggplot(data,aes(x=variable, y=value, fill=variable)) +
  stat_boxplot(geom='errorbar', linetype = 1, width = 0.5) +
  geom_boxplot() +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Kategorija") + ylab("ocjena")

```



Slika 3.4: Histogrami kategorija parks i theatres



Slika 3.5: Boxplotovi kategorija parks i theatres

Prvi test koji provodimo bit će F test, kojim ćemo provjeravati jesu li varijance kategorija `parks` i `theatres` jednake. Ta informacija će nam koristiti prilikom provođenja t-testa jer ćemo znati možemo li koristiti pretpostavku da su varijance ovih kategorija jednake.

```
var.test(df$parks, df$theatres, conf.level = 0.95, paired = T)

##
##  F test to compare two variances
##
## data: df$parks and df$theatres
## F = 0.95584, num df = 5455, denom df = 5455, p-value = 0.09539
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9064308 1.0079496
## sample estimates:
## ratio of variances
##                 0.9558434
```

Iz provedenog F testa vidimo da se hipoteza o jednakosti varijanci ne može odbaciti na razini značajnosti od 5%, stoga ćemo u t-testu postaviti parametar `var.equal` na `True`.

```
t.test(df$parks, df$theatres, conf.level = 0.95, var.equal = T, paired = T)
```

```
##
##  Paired t-test
##
## data: df$parks and df$theatres
## t = -10.462, df = 5455, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.1924213 -0.1316879
## sample estimates:
## mean of the differences
##                  -0.1620546
```

Iz provedenog t-testa uočavamo da je p vrijedost izuzetno malena, stoga hipotezu da su aritmetičke sredine kategorija `parks` i `theatres` jednake odbacujemo na razini značajnosti od 5%. U nastavku ćemo provjeriti imaju li te dvije kategorije jednaku razdiobu, što bismo mogli pretpostaviti gledajući sliku 3.3.

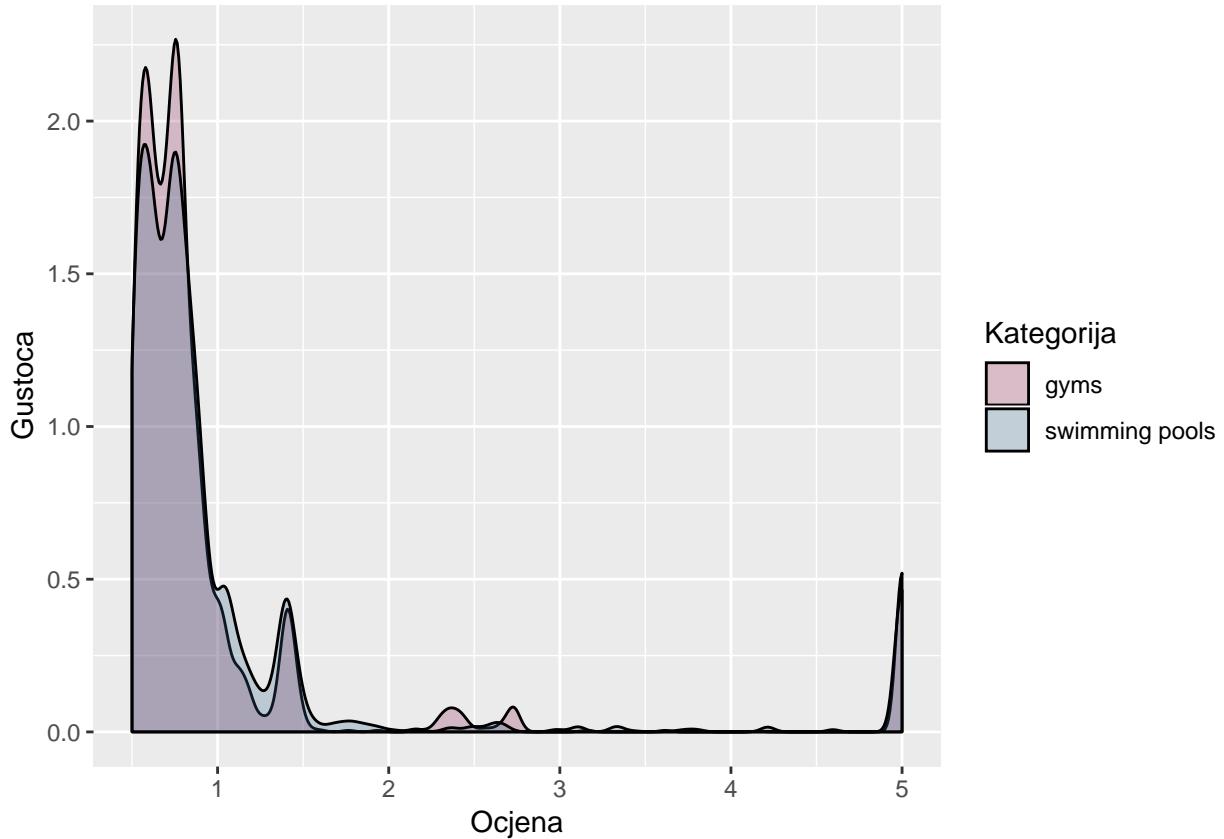
```
wilcox.test(df$parks, df$theatres, conf.level = 0.95, var.equal = T, paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data: df$parks and df$theatres
## V = 4317700, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Nakon provođenja Wilcoxonova testa dobivamo jednake rezultate kao i u t-testu, dakle odbacujemo nultu hipotezu.

```
ks.test(df$parks, df$theatres, conf.level = 0.95, var.equal = T, paired = T)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data: df$parks and df$theatres
## D = 0.1283, p-value < 2.2e-16
```



Slika 3.6: Density plotovi kategorija gyms i swimming pools

```
## alternative hypothesis: two-sided
```

Rezultati provedenog Kolmogorov-Smirnovljeva idu u prilog odbacivanju hipoteze da kategorije parks i theatres imaju jednaku razdiobu.

```
data <- melt(df[c("gyms", "swimming pools")])
```

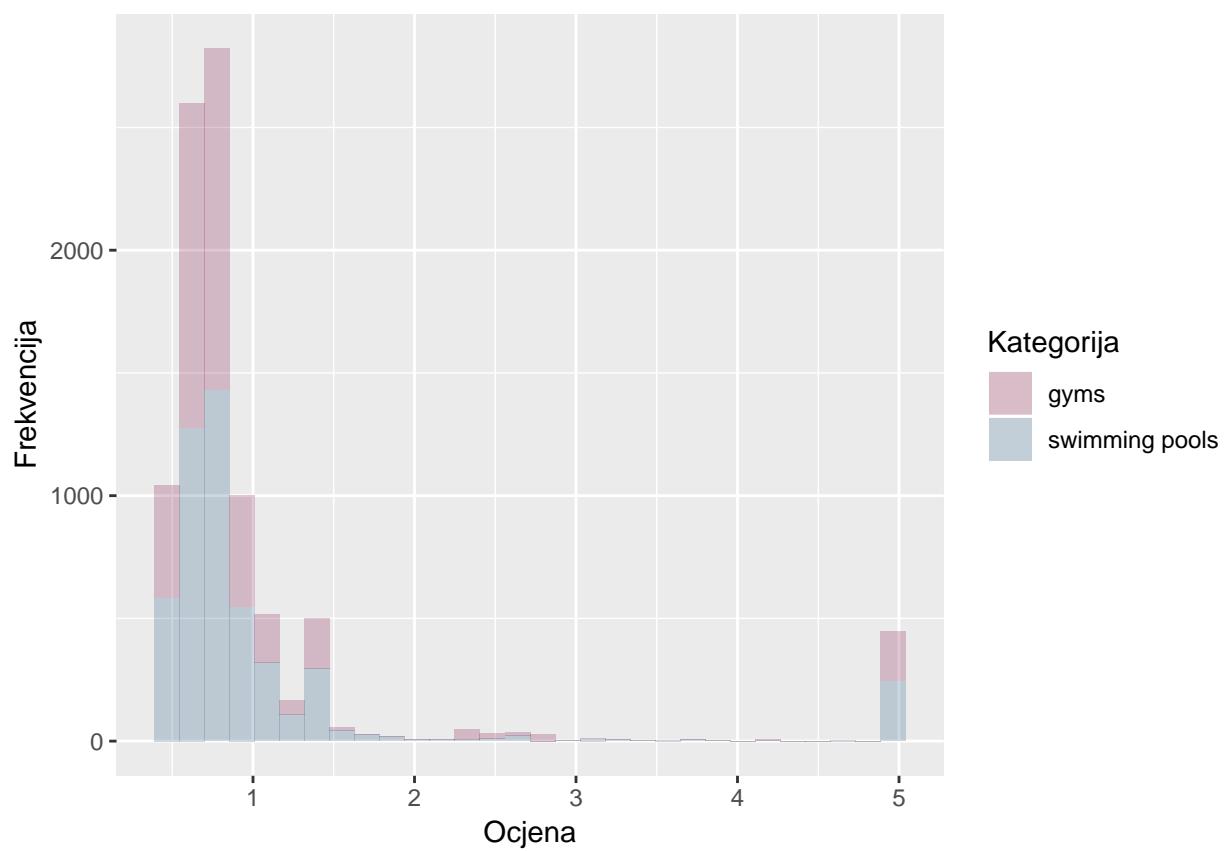
```
## No id variables; using all as measure variables
```

```
ggplot(data, aes(x=value, fill=variable)) +
  geom_density(alpha=0.25) +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Ocjena") + ylab("Gustoća")
```

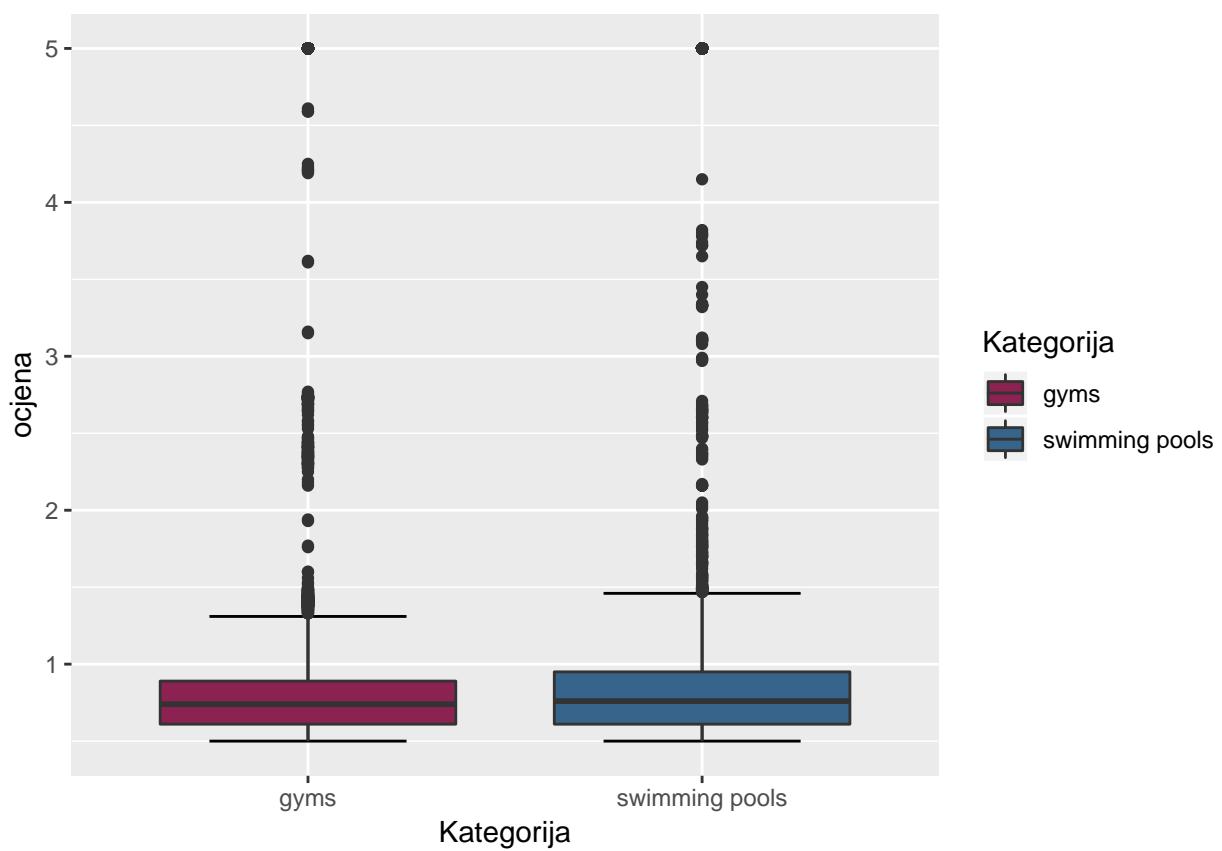
```
ggplot(data,aes(x=value, fill=variable)) +
  geom_histogram(alpha=0.25) +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Ocjena") + ylab("Frekvencija")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
ggplot(data,aes(x=variable, y=value, fill=variable)) +
  stat_boxplot(geom='errorbar', linetype = 1, width = 0.5) +
  geom_boxplot() +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Kategorija") + ylab("ocjena")
```



Slika 3.7: Histogrami kategorija gyms i swimming pools



Slika 3.8: Boxplotovi kategorija gyms i swimming pools

```

var.test(df$gyms, df$`swimming pools`, paired = T)

##
## F test to compare two variances
##
## data: df$gym and df$`swimming pools`
## F = 0.97203, num df = 4438, denom df = 4976, p-value = 0.3316
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.9179965 1.0293322
## sample estimates:
## ratio of variances
## 0.9720256

```

I u ovom slučaju možemo koristiti prepostavku da su varijance kategorija `gym` i `swimming pools` jednake.

```
t.test(df$gym, df$`swimming pools`, paired = T, var.equal = T)
```

```

##
## Paired t-test
##
## data: df$gym and df$`swimming pools`
## t = -1.3161, df = 4369, p-value = 0.1882
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.043040745 0.008464086
## sample estimates:
## mean of the differences
## -0.01728833

```

Nakon provođenja t-testa zaključujemo da na razini značajnosti od 5% ne možemo odbaciti nultu hipotezu, stoga zaključujemo da su aritmetičke sredine ove dvije kategorije jednake.

```
wilcox.test(df$gym, df$`swimming pools`, paired = T, var.equal = T)
```

```

##
## Wilcoxon signed rank test with continuity correction
##
## data: df$gym and df$`swimming pools`
## V = 3102800, p-value = 0.00135
## alternative hypothesis: true location shift is not equal to 0

```

No, kada koristimo Wilcoxonov test, odbacujemo nultu hipotezu, te za razliku od t-testa ne možemo zaključiti da su aritmetičke sredine ocjena razmatranih kategorija jednake.

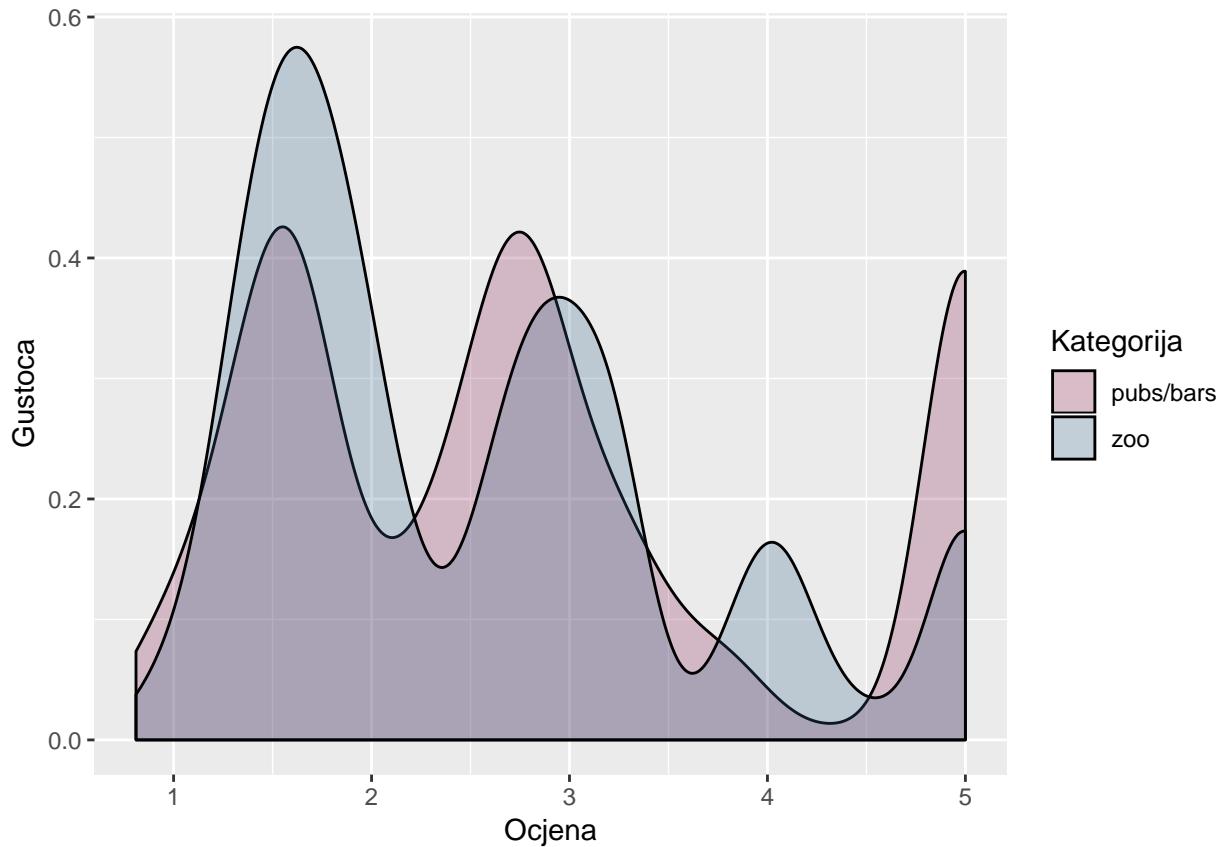
```
ks.test(df$gym, df$`swimming pools`, paired = T)
```

```

##
## Two-sample Kolmogorov-Smirnov test
##
## data: df$gym and df$`swimming pools`
## D = 0.06483, p-value = 5.441e-09
## alternative hypothesis: two-sided

```

Nakon provođenja Kolmogorov-Smirnovljeva testa možemo odbaciti nultu hipotezu na nivou značajnosti 5%, čime zaključujemo da njihove razdiobe ne podudaraju.



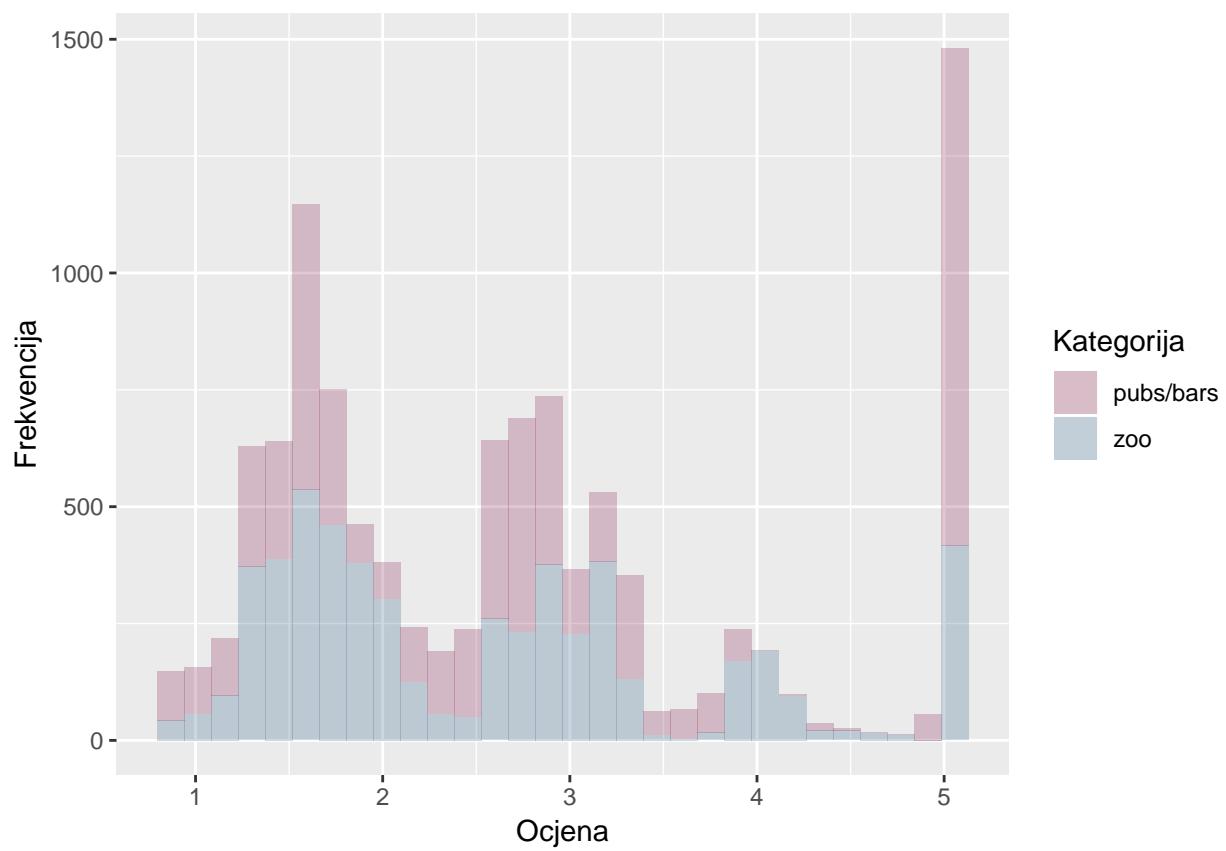
Slika 3.9: Density plotovi kategorija pubs/bars i zoo

```
data <- melt(df[c("pubs/bars", "zoo")])

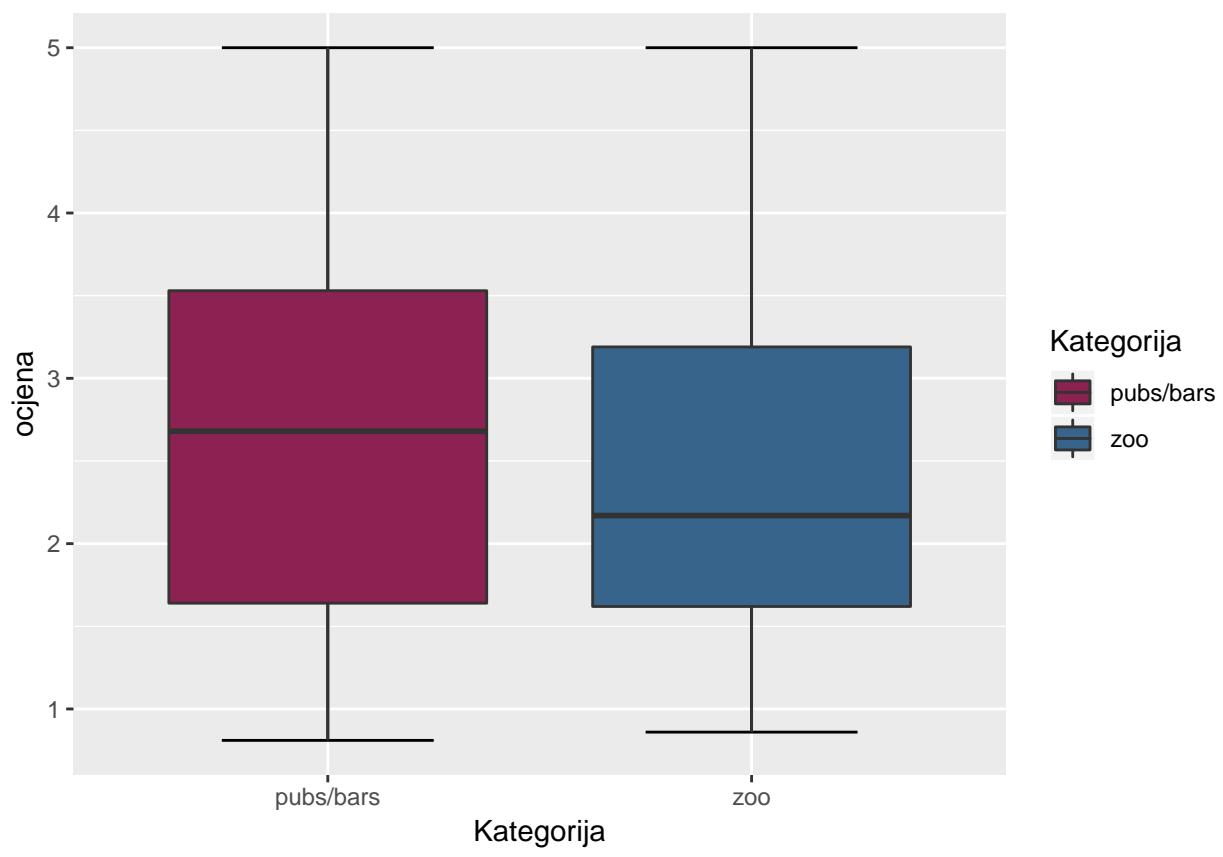
## No id variables; using all as measure variables
ggplot(data, aes(x=value, fill=variable)) + geom_density(alpha=0.25) + scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4"))

ggplot(data,aes(x=value, fill=variable)) + geom_histogram(alpha=0.25) + scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4"))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
ggplot(data,aes(x=variable, y=value, fill=variable)) + stat_boxplot(geom='errorbar', linetype = 1, width = 0.5) + geom_boxplot() + scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) + xlab("Kategorija")
```



Slika 3.10: Histogrami kategorija pubs/bars i zoo



Slika 3.11: Boxplotovi kategorija pubs/bars i zoo

```

var.test(df$`pubs/bars`, df$zoo)

##
## F test to compare two variances
##
## data: df$`pubs/bars` and df$zoo
## F = 1.3844, num df = 5455, denom df = 5455, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.312825 1.459860
## sample estimates:
## ratio of variances
## 1.384392

```

U ovom slučaju ne možemo pretpostaviti da su varijance kategorija bars/pubs i zoo jednake, tako da ta pretpostavka neće biti korištena u t-testu.

```
t.test(df$`pubs/bars`, df$zoo, paired = T)
```

```

##
## Paired t-test
##
## data: df$`pubs/bars` and df$zoo
## t = 18.611, df = 5455, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.2611832 0.3226841
## sample estimates:
## mean of the differences
## 0.2919337

```

Nakon provedenog t-testa odbacujemo nultu hipotezu na nivou značajnosti od 5%, što znači da ove dvije kategorije nemaju jednaku aritmetičku sredinu.

```
wilcox.test(df$`pubs/bars`, df$zoo, paired = T)
```

```

##
## Wilcoxon signed rank test with continuity correction
##
## data: df$`pubs/bars` and df$zoo
## V = 8974900, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0

```

Kao i u t-testu, u Wilcoxonovom testu odbacujemo nultu hipotezu s vrlo visokom sigurnošću.

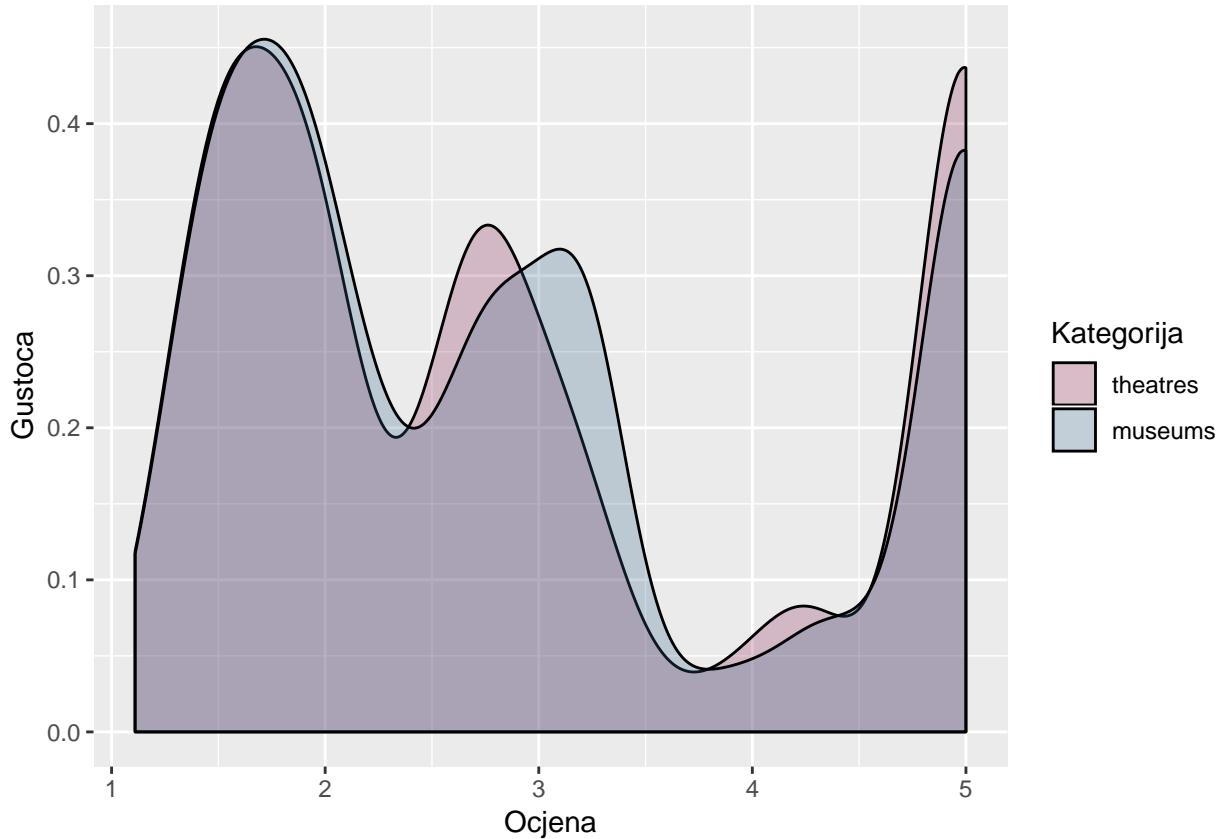
```
ks.test(df$`pubs/bars`, df$zoo, paired = T)
```

```

##
## Two-sample Kolmogorov-Smirnov test
##
## data: df$`pubs/bars` and df$zoo
## D = 0.14168, p-value < 2.2e-16
## alternative hypothesis: two-sided

```

Kao što se moglo očekivati iz rezultata prethodna dva testa, i prema Kolmogorov-Smirnovljevu testu odbacujemo nultu hipotezu te zaključujemo da se ove dvije kategorije ne podudaraju po distribuciji.



Slika 3.12: Density plotovi kategorija theatres i museums

```

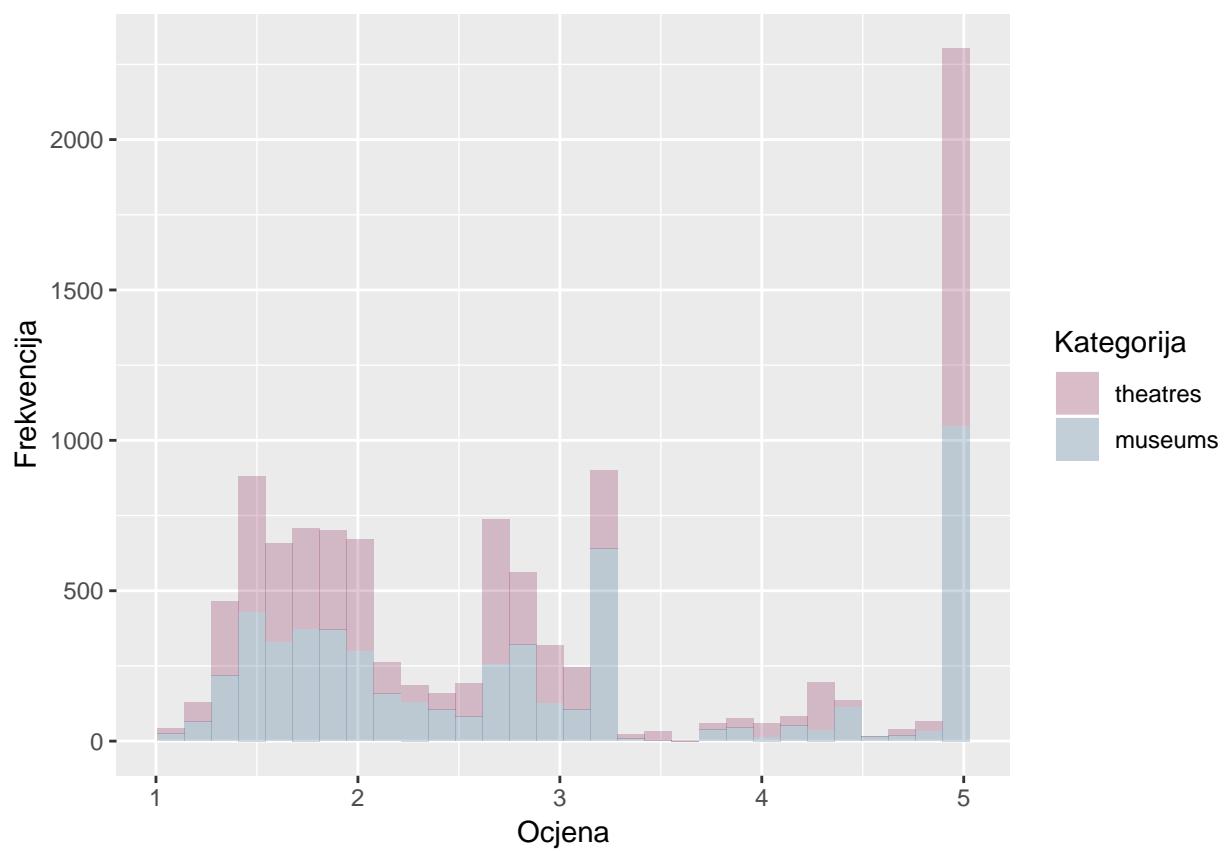
data <- melt(df[c("theatres", "museums")])

## No id variables; using all as measure variables
ggplot(data, aes(x=value, fill=variable)) +
  geom_density(alpha=0.25) +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Ocjena") + ylab("Gustoća")

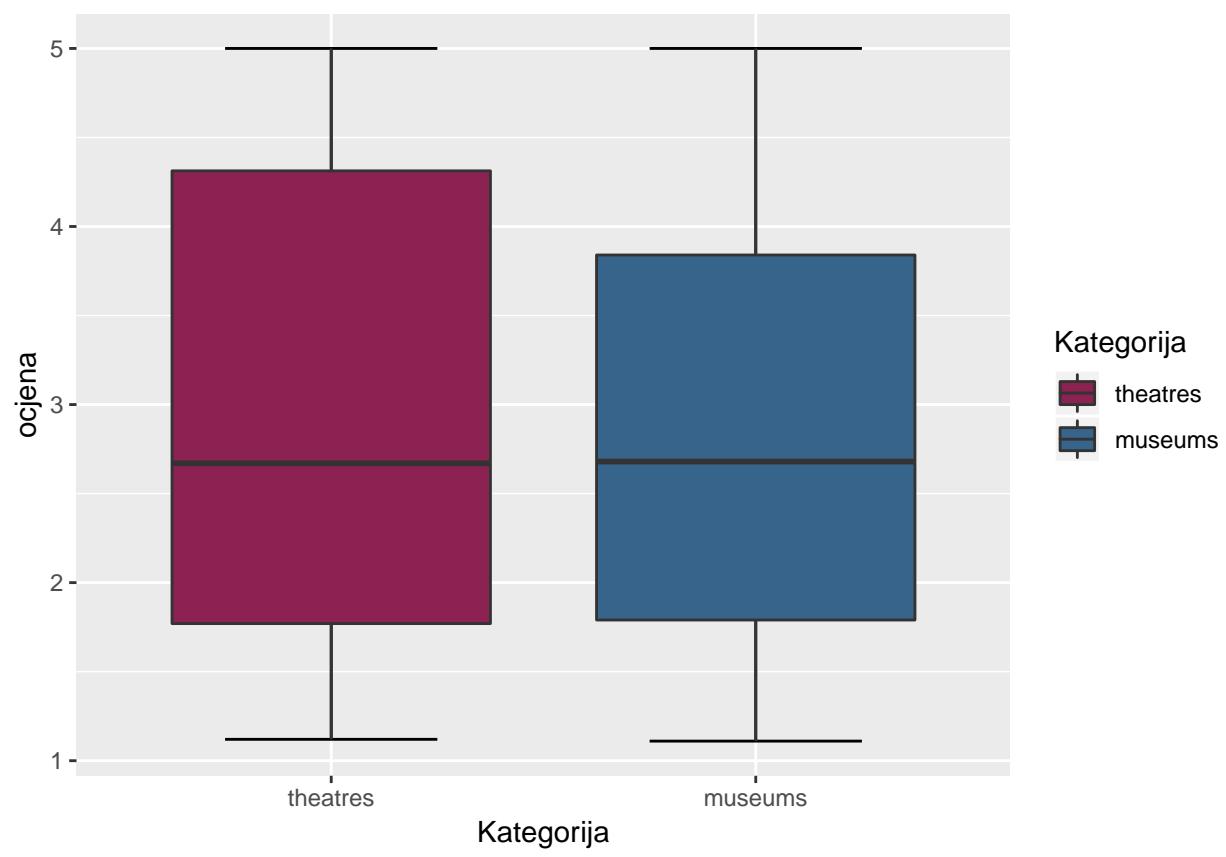
ggplot(data,aes(x=value, fill=variable)) +
  geom_histogram(alpha=0.25) +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Ocjena") + ylab("Frekvencija")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
ggplot(data,aes(x=variable, y=value, fill=variable)) +
  stat_boxplot(geom='errorbar', linetype = 1, width = 0.5) +
  geom_boxplot() +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Kategorija") + ylab("ocjena")

```



Slika 3.13: Histogrami kategorija theatres i museums



Slika 3.14: Boxplotovi kategorija theatres i museums

```

var.test(df$theatres, df$museums)

##
## F test to compare two variances
##
## data: df$theatres and df$museums
## F = 1.0903, num df = 5455, denom df = 5455, p-value = 0.001411
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.033947 1.149748
## sample estimates:
## ratio of variances
## 1.090311

```

Iz prethodnog testa možemo vidjeti da se varijance kategorija `theatres` i `museums` razlikuju. Stoga u t-testu nećemo koristiti pretpostavku da su one jednake.

```
t.test(df$theatres, df$museums, paired = T)
```

```

##
## Paired t-test
##
## data: df$theatres and df$museums
## t = 3.6494, df = 5455, p-value = 0.0002654
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.03029135 0.10061041
## sample estimates:
## mean of the differences
## 0.06545088

```

Nakon provođenja t-testa uočavamo da se nulta hipoteza mora odbaciti, iako bismo iz boxplota na slici 3.14 mogli zaključiti da kategorije `theatres` i `museums` imaju jednake aritmetičke sredine. Međutim taj boxplot ukazuje na jednakost njihovih medijana, što možemo provjeriti Moodovim testom.

```
mood.test(df$theatres, df$museums, paired = T)
```

```

##
## Mood two-sample test of scale
##
## data: df$theatres and df$museums
## Z = 3.0113, p-value = 0.002602
## alternative hypothesis: two.sided

```

Ipak, provedeni Moodov test rezultirao je relativno malom p-vrijednost, što znači da nultu hipotezu možemo odbaciti.

```
wilcox.test(df$theatres, df$museums, paired = T)
```

```

##
## Wilcoxon signed rank test with continuity correction
##
## data: df$theatres and df$museums
## V = 5830500, p-value = 0.0008045
## alternative hypothesis: true location shift is not equal to 0

```

Rezultati Wilcoxonova testa su jednaki kao i za t-test, odbacujemo nultu hipotezu.

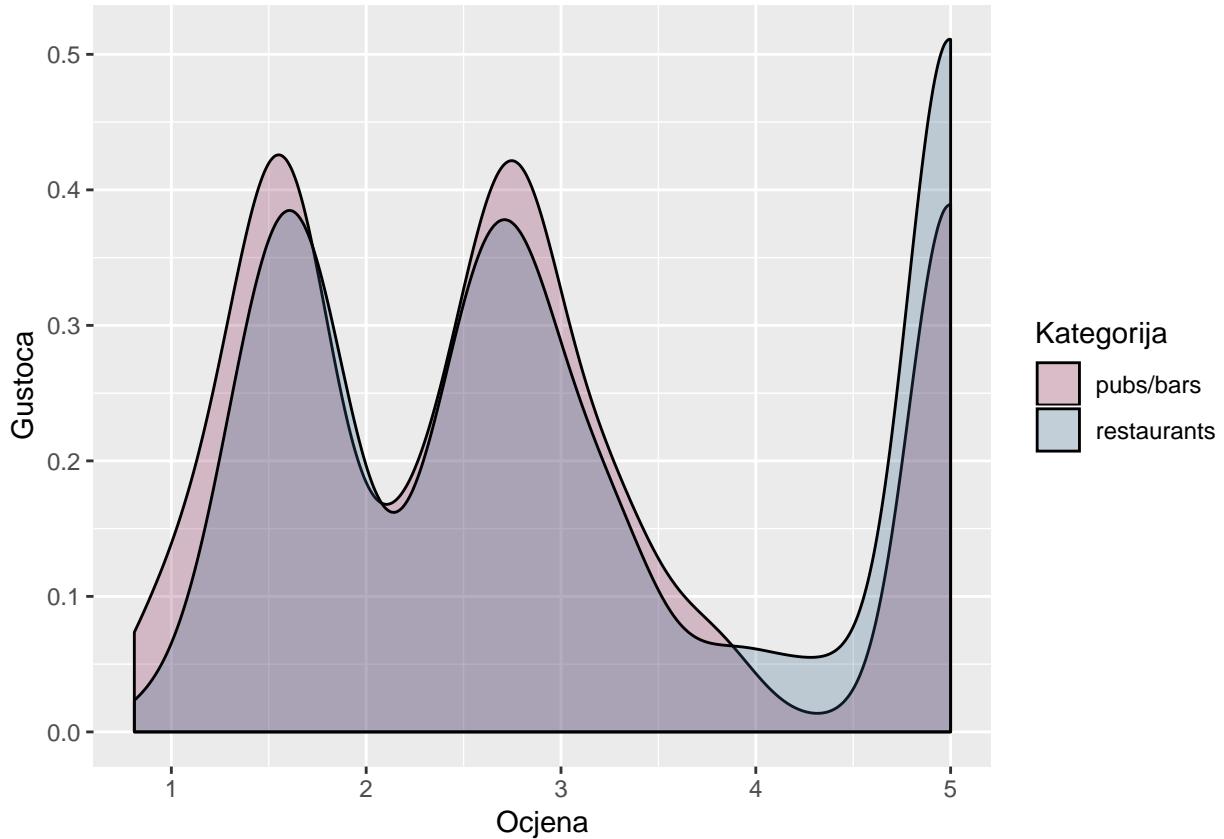
```
ks.test(df$theatres, df$museums, paired = T)

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  df$theatres and df$museums
## D = 0.045455, p-value = 2.543e-05
## alternative hypothesis: two-sided
```

Iz rezultata Kolmogorov-Smirnovljeva testa vidimo da se nulta hipoteza odbacuje, tako da možemo zaključiti da se ni distribucije ove dvije kategorije ne podudaraju.

```
data <- melt(df[c("pubs/bars", "restaurants")])

## No id variables; using all as measure variables
```



Slika 3.15: Density plotovi kategorija pubs/bars i restaurants

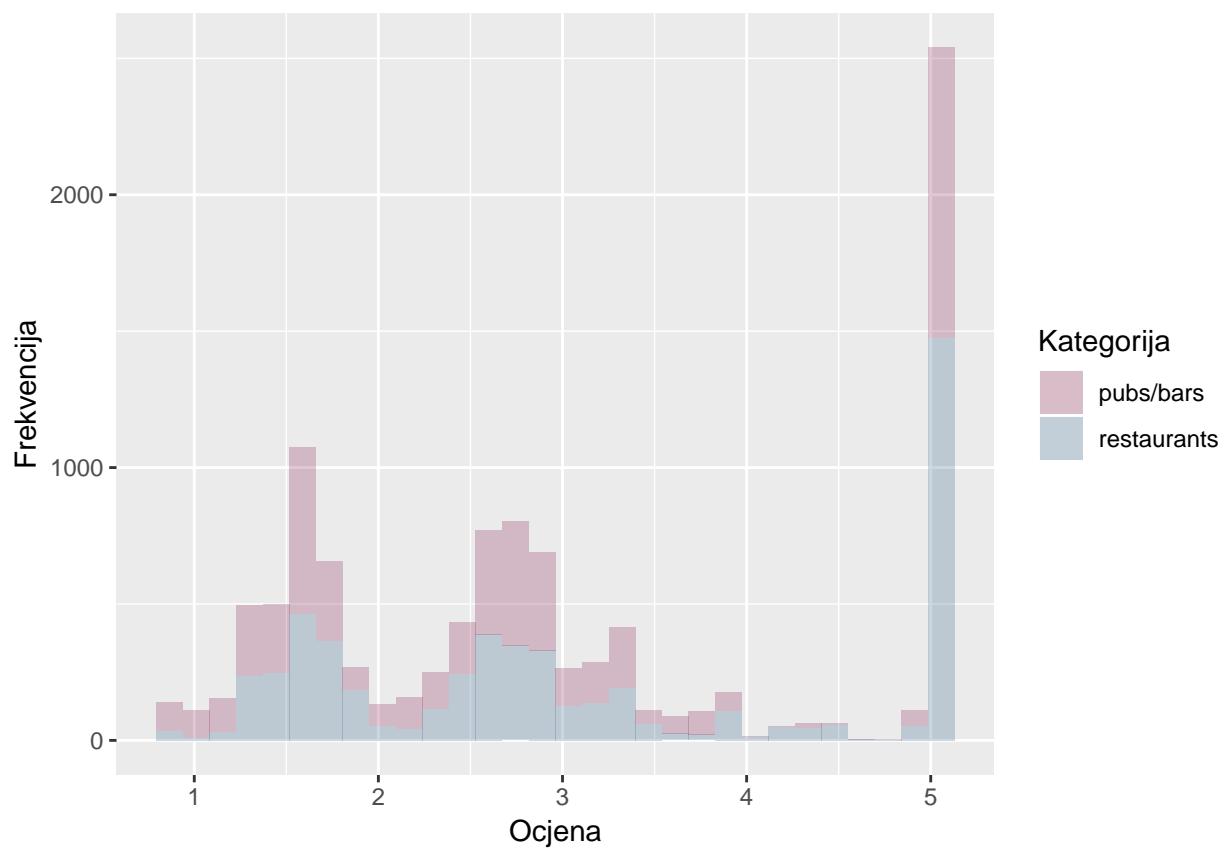
```

ggplot(data, aes(x=value, fill=variable)) +
  geom_density(alpha=0.25) +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Ocjena") + ylab("Gustoća")

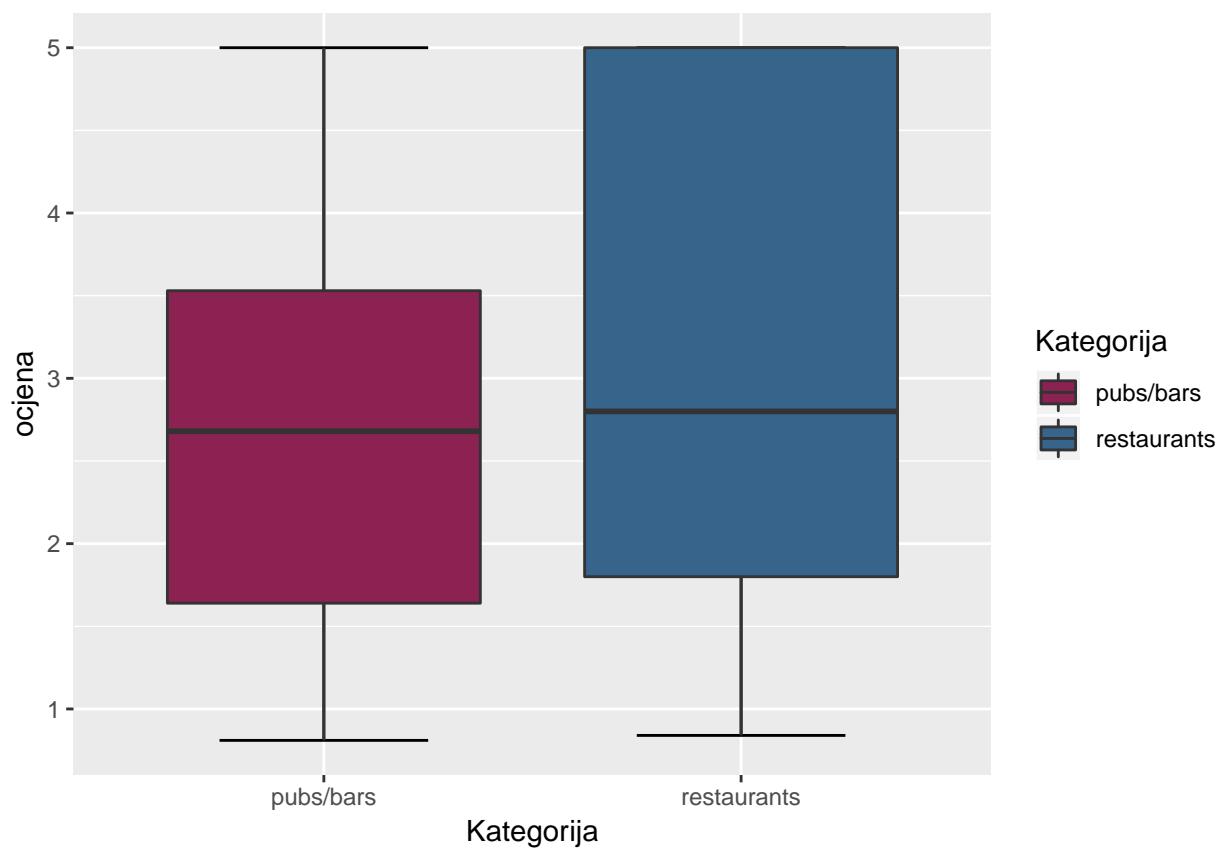
ggplot(data, aes(x=value, fill=variable)) +
  geom_histogram(alpha=0.25) +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Ocjena") + ylab("Frekvencija")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
ggplot(data,aes(x=variable, y=value, fill=variable)) +
  stat_boxplot(geom='errorbar', linetype = 1, width = 0.5) +
  geom_boxplot() +
  scale_fill_manual("Kategorija", values = c("violetred4", "steelblue4")) +
  xlab("Kategorija") + ylab("ocjena")

```



Slika 3.16: Histogrami kategorija pubs/bars i restaurants



Slika 3.17: Boxplotovi kategorija pubs/bars i restaurants

Provjerit ćemo još i podudarnost kategorija `pubs/bars` i `restaurants`.

```
var.test(df$`pubs/bars`, df$restaurants)

##
## F test to compare two variances
##
## data: df$`pubs/bars` and df$restaurants
## F = 0.92888, num df = 5455, denom df = 5455, p-value = 0.006449
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.8808617 0.9795168
## sample estimates:
## ratio of variances
## 0.9288804
```

Mala p-vrijednost dobivena provedenim F-testom ukazuje na odbacivanje hipoteze o jednakosti varijanci razmatranih kategorija.

```
t.test(df$`pubs/bars`, df$restaurants, paired=T)

##
## Paired t-test
##
## data: df$`pubs/bars` and df$restaurants
## t = -17.405, df = 5455, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.3263244 -0.2602555
## sample estimates:
## mean of the differences
## -0.29329

wilcox.test(df$`pubs/bars`, df$restaurants, paired=T)

##
## Wilcoxon signed rank test with continuity correction
##
## data: df$`pubs/bars` and df$restaurants
## V = 3722100, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Rezultati provedenog t-testa i Wilcoxonova testa ukazuju nam na odbacivanje nulte hipoteze, što znači da kategorije `pubs/bars` i `restaurants` neamju jednake aritmetičke sredine. Moodovim testom provjerit ćemo jednakost njihovih medijana.

```
mood.test(df$`pubs/bars`, df$restaurants, paired=T)

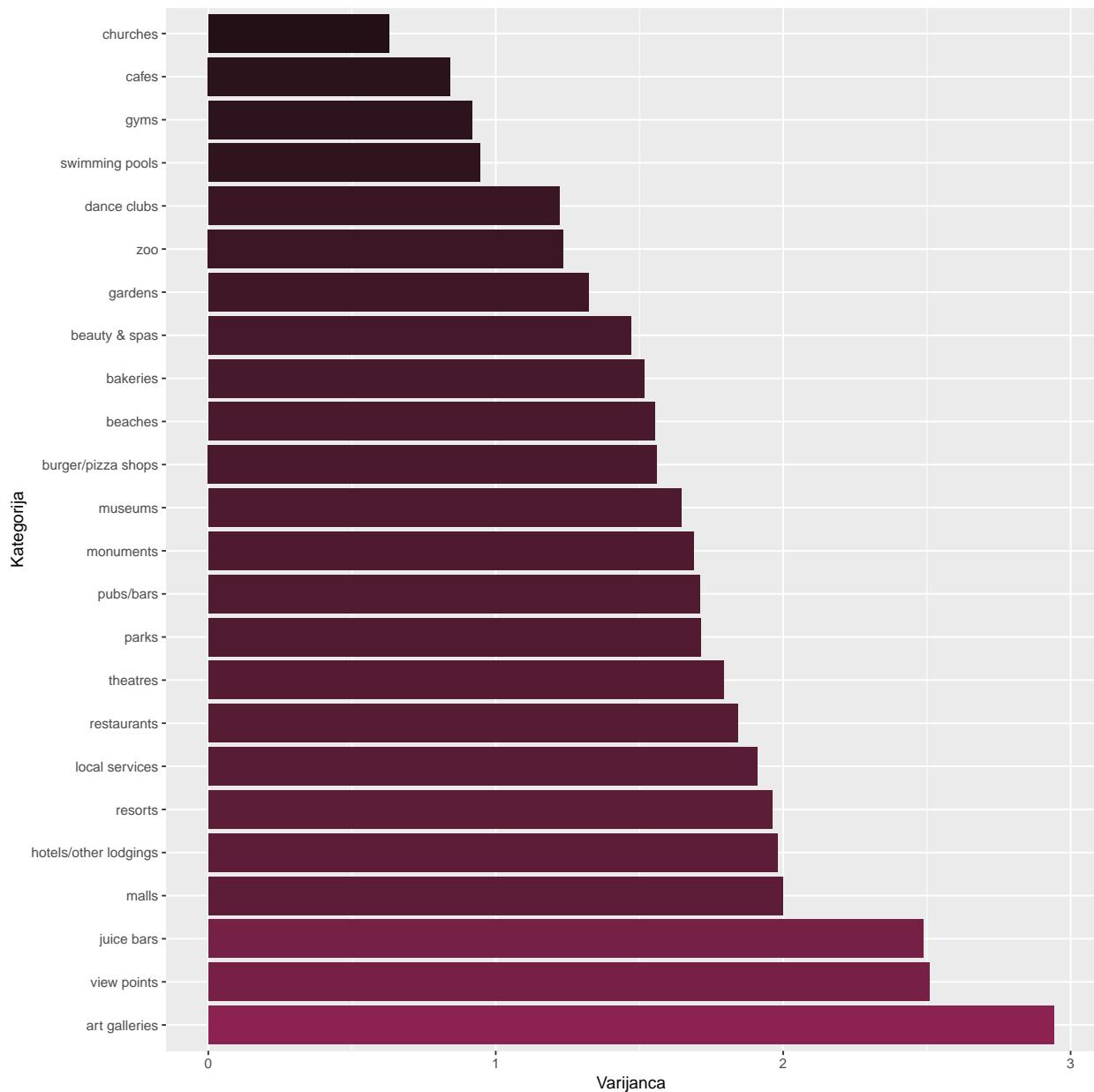
##
## Mood two-sample test of scale
##
## data: df$`pubs/bars` and df$restaurants
## Z = 1.108, p-value = 0.2679
## alternative hypothesis: two.sided
```

Rezultat provedenog Moodova testa je p-vrijednost koja iznosi 0.2679, što znači da hipotezu o jednakosti medijana razmatranih kategorija ne možemo odbaciti.

Nakon usporedbi odabranih parova kategorija pokušat ćemo utvrditi postoje li kategorije koje su posebno

polarizirajuće ili nepolarizirajuće, odnosno one kategorije oko kojih se korisnici najviše ili najmanje slažu. Za mjeru polariziranosti koristit ćemo varijancu. Na slici 3.18 uočavamo da je kategorija `churches` najviše polarizirajuća, što zaključujemo iz činjenice da varijanca te kategorije manja od varijanci svih ostalih kategorija. Jednako tako, možemo vidjeti da je kategorija `art galleries` najmanje polarizirajuća, jer ima najveću varijancu.

```
k <- lapply(df[-1], function(x) {var(x[!is.na(x)])})  
ggplot(melt(k), aes(x = reorder(L1, -value), y = value, fill = value)) +  
  geom_bar(stat = "identity") +  
  coord_flip() + xlab("Kategorija") + ylab("Varijanca") +  
  scale_fill_gradient2(mid = "black", high = "violetred4") +  
  theme(legend.position="none")
```



Slika 3.18: Stupčasti dijagram varijanci kategorija

4 Linearna regresija

U ovom poglavlju napravljena je regresijska analiza odabranih parova kategorija. Prva linearna regresija napravljena je nad parom `gyms` i `swimming pools`.

```
linmod <- lm(gyms ~ `swimming pools`, df[-1])  
  
summary(linmod)  
  
##  
## Call:  
## lm(formula = gyms ~ `swimming pools`, data = df[-1])  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -2.9058 -0.1724 -0.1169 -0.0489  4.2878  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.39287   0.01777  22.11  <2e-16 ***  
## `swimming pools` 0.60258   0.01280  47.09  <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.786 on 4368 degrees of freedom  
##   (1086 observations deleted due to missingness)  
## Multiple R-squared:  0.3367, Adjusted R-squared:  0.3365  
## F-statistic: 2217 on 1 and 4368 DF, p-value: < 2.2e-16  
ggplot(df, aes(x = gyms, y = `swimming pools`)) +  
  geom_point(shape = 1) + geom_smooth(method = lm)
```

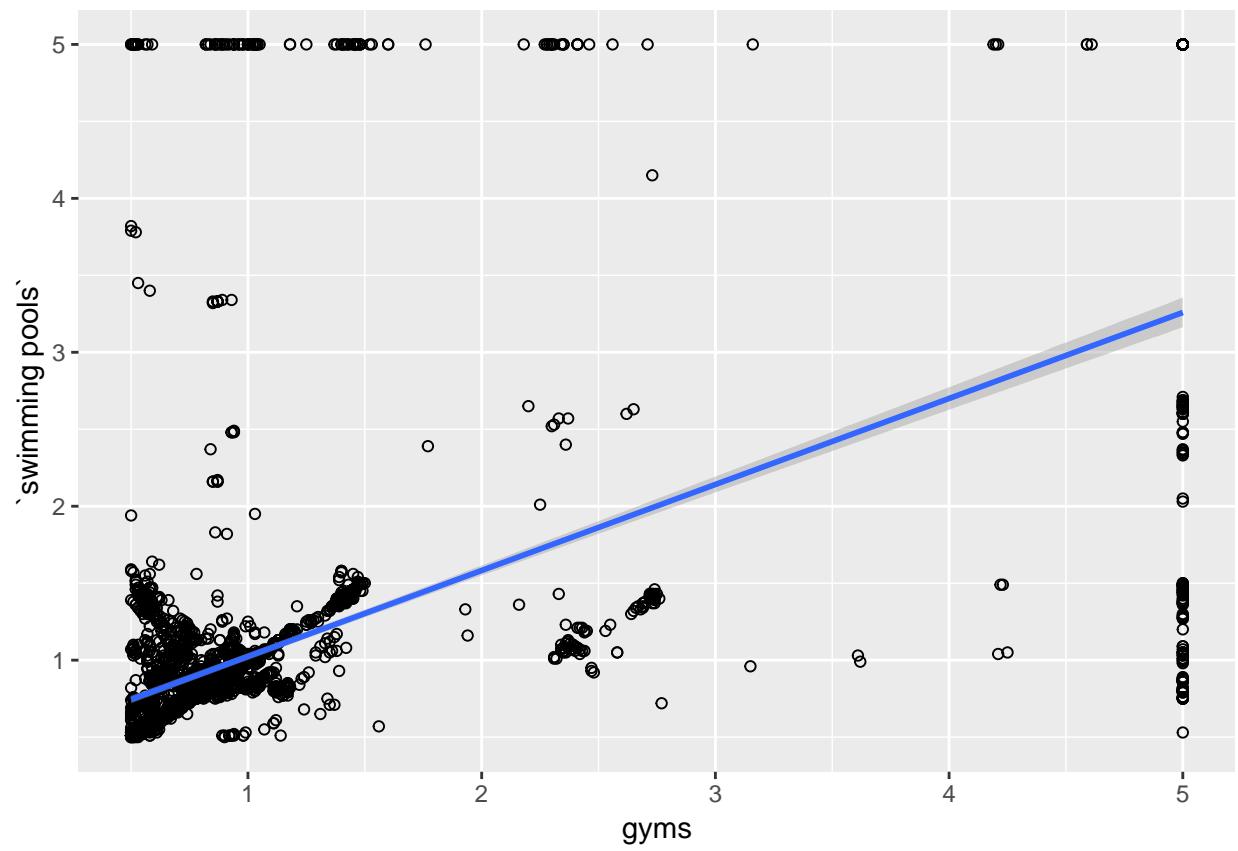
Usprkos tome što se gledajući u graf na slici 4.1 čini da su točke razbacane po grafu bez nekog posebnog reda, pravac dobiven linearnom regresijom postiže ukupnu pogrešku od 0.786, što znači da pravac u prosjeku pogriješi za jednu ocjenu pri predikciji ocjene kategorije `swimming pools` na temelju kategorije `gyms`. U nastavku ćemo provjeriti je li razdioba reziduala regresijskog pravca normalna.

```
ggqqplot(rstandard(linmod), shape=1) +  
  ggtitle("") + xlab("Teoretski kvantil") +  
  ylab("Standardizirani rezidual")
```

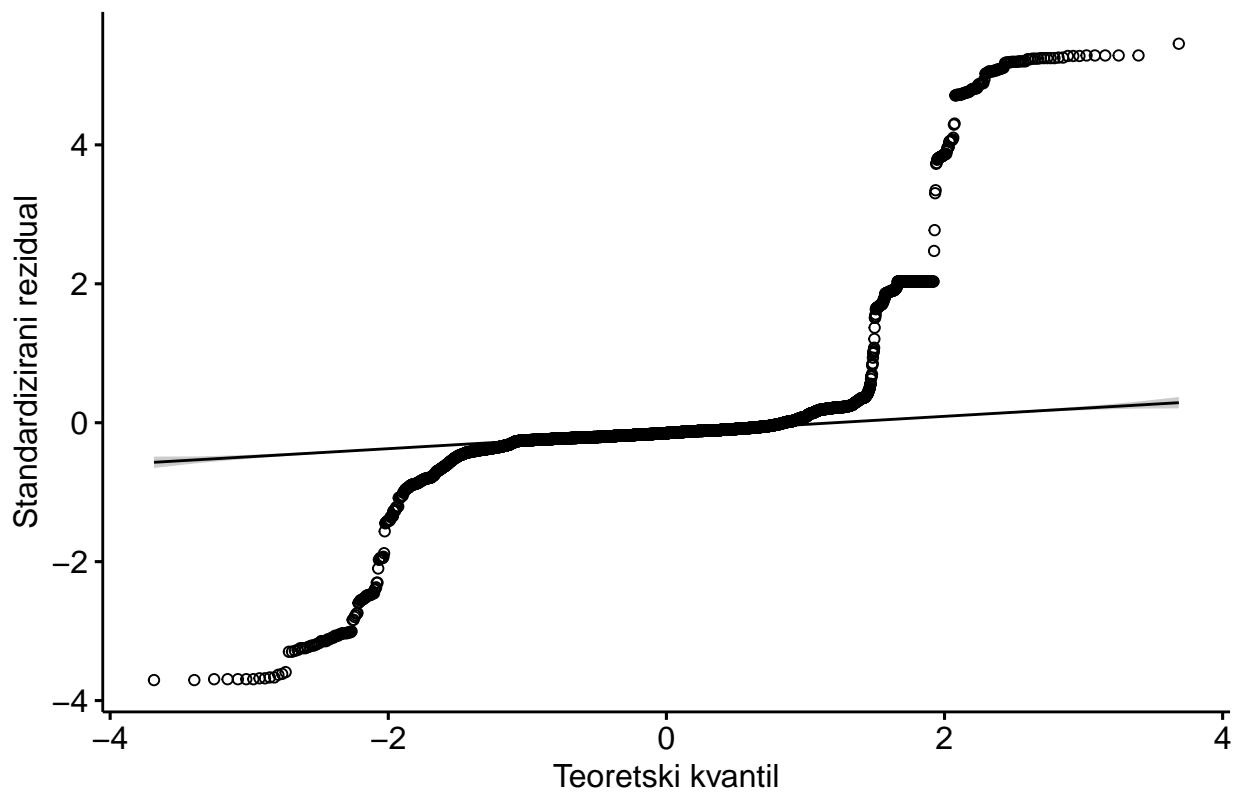
```
ks.test(rstandard(linmod), 'pnorm')  
  
##  
## One-sample Kolmogorov-Smirnov test  
##  
## data: rstandard(linmod)  
## D = 0.31075, p-value < 2.2e-16  
## alternative hypothesis: two-sided
```

Kao što iz qqplota na slici 4.2 i provedenog Kolmogorov-Smirnovljeva testa možemo zaključiti, standardizirani reziduali odstupaju od normalne razdiobe.

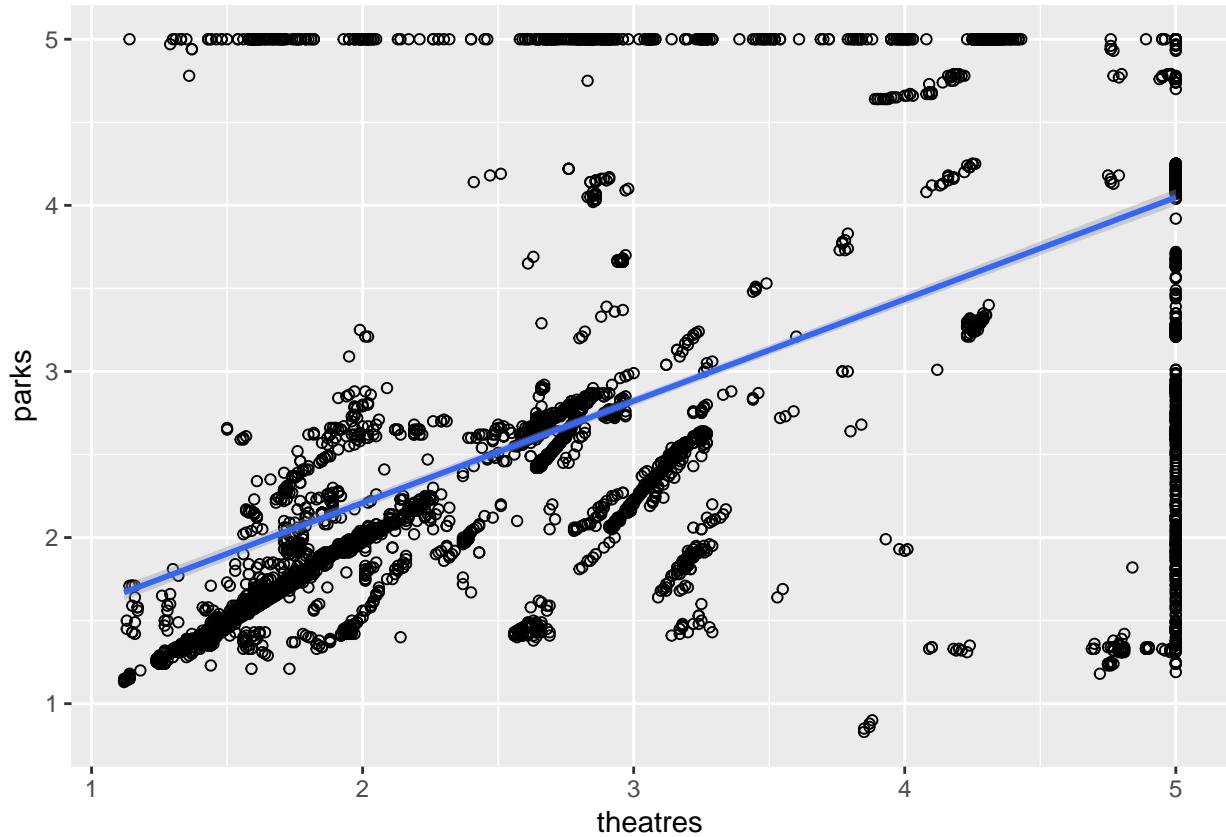
```
linmod <- lm(`theatres` ~ parks, df[-1])  
  
summary(linmod)  
  
##
```



Slika 4.1: Vizualizacija linearne regresije kategorija gyms i swimming pools



Slika 4.2: QQ-plot reziduala



Slika 4.3: Vizualizacija linearne regresije kategorija theatres i parks

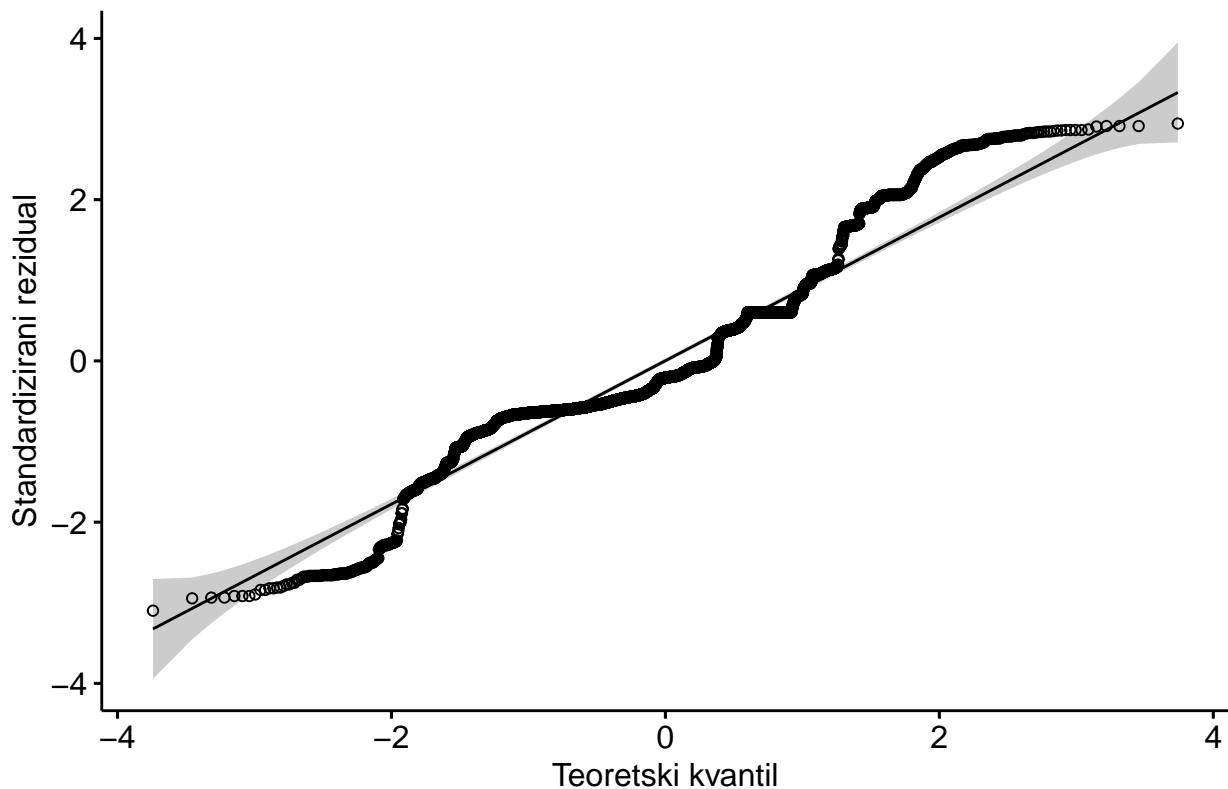
```

## Call:
## lm(formula = theatres ~ parks, data = df[-1])
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -3.2315 -0.6230 -0.2182  0.6285  3.0714 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.16562   0.03332  34.98   <2e-16 ***
## parks       0.64118   0.01079  59.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.043 on 5454 degrees of freedom
## Multiple R-squared:  0.393, Adjusted R-squared:  0.3929 
## F-statistic: 3531 on 1 and 5454 DF, p-value: < 2.2e-16

ggplot(df, aes(x = theatres, y = parks)) +
  geom_point(shape = 1) + geom_smooth(method = lm)

```

Iako bismo, gledajući u točke na grafu 4.3, mogli pretpostaviti da će ukupna pogreška predikcije biti manja nego u prethodnom primjeru, dobivena je pogreška od 1.043. Dobiveni regresijski pravac odstupa od pravca koji bismo povući intuitivno, što nimalo ne čudi zbog velikog broja stršećih vrijednosti. Ponovno ćemo



Slika 4.4: QQ-plot reziduala

provjeriti pripadaju li reziduali regresijskog pravca normalnoj razdiobi.

```
ggqqplot(rstandard(linmod), shape=1) +
  ggtitle("") + xlab("Teoretski kvantil") +
  ylab("Standardizirani rezidual")
```

```
ks.test(rstandard(linmod), 'pnorm')
```

```
##
## One-sample Kolmogorov-Smirnov test
##
## data: rstandard(linmod)
## D = 0.13947, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Reziduali, dakle, ponovno odstupaju od normalne razdiobe.

Sada ćemo pokušati predvidjeti kategoriju `malls` na temelju kategorija `view points`, `zoo` i `restaurants`.

```
linmod <- lm(malls ~ `view points` + zoo + restaurants , df[-1])
```

```
summary(linmod)
```

```
##
## Call:
## lm(formula = malls ~ `view points` + zoo + restaurants, data = df[-1])
```

```

## 
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.1032 -0.9446 -0.2661  1.0684  3.2569
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.46965   0.05699  43.33 <2e-16 ***
## `view points` -0.26338   0.01092 -24.12 <2e-16 ***
## zoo          0.24155   0.01761  13.72 <2e-16 ***
## restaurants  0.25455   0.01431  17.79 <2e-16 ***
## ---    
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.174 on 5107 degrees of freedom
##   (345 observations deleted due to missingness)
## Multiple R-squared:  0.3092, Adjusted R-squared:  0.3088 
## F-statistic: 761.8 on 3 and 5107 DF,  p-value: < 2.2e-16

```

Rezultat ove linearne regresije je linearni model čija standardna pogreška iznosi 1.174, što je zadovoljavajuć rezultat ako uzmemo u obzir relativno male korelacije tih kategorija.

```
ggqqplot(rstandard(linmod), shape=1) +
  ggtitle("") + xlab("Teoretski kvantil") +
  ylab("Standardizirani rezidual")
```

```
ks.test(rstandard(linmod), 'pnorm')
```

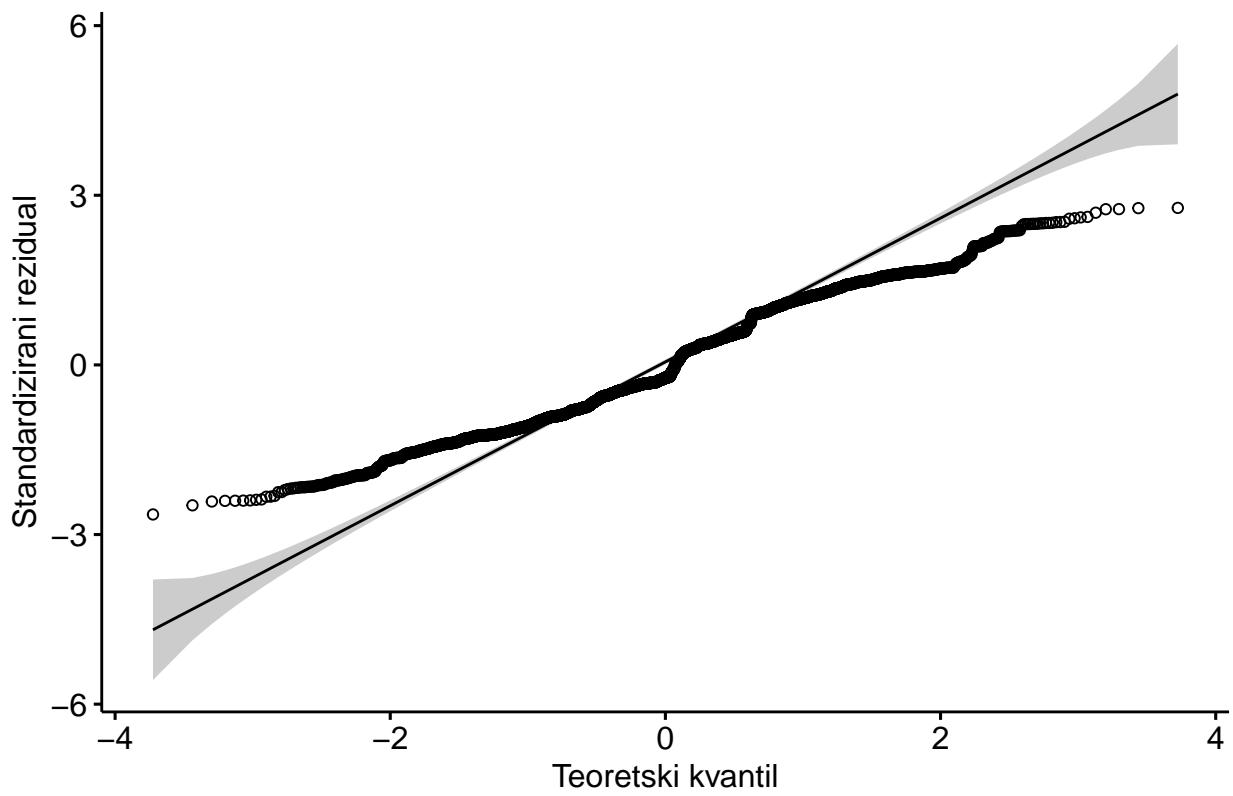
```

## 
## One-sample Kolmogorov-Smirnov test
## 
## data: rstandard(linmod)
## D = 0.095274, p-value < 2.2e-16
## alternative hypothesis: two-sided
rstandard(linmod) %>% as_tibble
```

```

## # A tibble: 5,111 x 1
##       value
##   <dbl>
## 1 -0.376
## 2 -0.374
## 3 -0.180
## 4 -0.443
## 5 -0.636
## 6 -0.638
## 7 -0.632
## 8 -0.630
## 9 -0.625
## 10 -0.627
## # ... with 5,101 more rows
```

Reziduali regresijskog pravca ponovno ne pripadaju normalnoj razdiobi.



Slika 4.5: QQ-plot reziduala

5 Komparativna analiza korisnika

U nastavku će biti prikazana implementacija algoritma kolaborativnog filtriranja. To je u suštini algoritam koji čini automatizirane predikcije o interesima korisnika na način da prikuplja informacije o interesima velikog broja korisnika. Temeljna pretpostavka algoritma je da ako dva korisnika imaju jednak interes za određeno područje, vjerojatnije je da će imati slična mišljenja na ostalim područjima, u odnosu na neku treću osobu koja nema slične interese kao ove dvije. Algoritam je u vrlo širokoj uporabi u praksi te se najčešće koristi u sustavu preporuka.

```
sim <- function(x, y){
  sum(x * y) / (norm(x, type = "2") * norm(y, type = "2"))
}

n_max_idx <- function(row, index, n){
  row <- row[-1]
  coscor <- apply(df_numerical, 1, function(x){sim(x, row)})
  coscor[index] <- 0
  maxes <- c()
  for(i in 1:n){
    idx <- which.max(coscor)
    maxes <- c(maxes, idx)
    coscor[idx] <- 0
  }
  maxes
}

df_numerical <- df_original[-1]
sim(df_original[1, -1], df_original[2, -1])

## [1] 0.9996665
```

U kodu iznad implementirane su dvije funkcije: `sim` i `n_max_idx`. Prva funkcija izračunava kosinusnu sličnost između dva ulazna vektora prema formuli:

$$\begin{aligned} \text{similarity} &= \cos(\theta) \\ &= \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \\ &= \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \end{aligned}$$

gdje su **A** i **B** ulazni vektori. Ona predstavlja kosinus kuta između dva normirana višedimenzionalna vektora te kada su što su oni sličniji vrijednost će biti bliža 1.

Druga funkcija vraća indekse n vektora iz podatkovnog skupa najsličnijih vektoru koji je dobiven na ulazu kao varijabla `row`.

```
row_index <- 500
test_row <- df_original[row_index, ]
test_row

## # A tibble: 1 x 25
##   User  churches resorts beaches parks theatres museums malls   zoo
##   <chr>    <dbl>   <dbl>   <dbl> <dbl>    <dbl>   <dbl> <dbl> <dbl>
```

```

## 1 User~      1.62      1.64      1.65      5       5       5  1.74  1.73
## # ... with 16 more variables: restaurants <dbl>, `pubs/bars` <dbl>, `local
## #   services` <dbl>, `burger/pizza shops` <dbl>, `hotels/other`
## #   lodgings` <dbl>, `juice bars` <dbl>, `art galleries` <dbl>, `dance
## #   clubs` <dbl>, `swimming pools` <dbl>, gyms <dbl>, bakeries <dbl>,
## #   `beauty & spas` <dbl>, cafes <dbl>, `view points` <dbl>,
## #   monuments <dbl>, gardens <dbl>
idx <- n_max_idx(test_row, row_index, 20)

df_original[idx, ]

## # A tibble: 20 x 25
##   User  churches resorts beaches parks theatres museums malls   zoo
##   <chr>    <dbl>   <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 User~     1.59     1.6     1.61     5       5       5  1.75  1.74
## 2 User~     1.56     1.58     1.59     5       5       5  4.51  1.76
## 3 User~     1.58     1.59     1.61     5       5       5  1.72  1.71
## 4 User~     1.48     2.2      1.52     5       5       5  4.45  1.78
## 5 User~     1.56     1.57     1.6      5       5       5  1.73  1.72
## 6 User~     1.46     2.19     1.5      3.57    5       5  4.44  1.79
## 7 User~     1.55     1.8      1.81     5       5       5  3.25  3.15
## 8 User~     1.44     1.47     1.49     5       5       5  4.43  1.79
## 9 User~     1.53     1.55     1.58     5       5       5  4.47  1.71
## 10 User~    1.66     2.38     2.87     5       5       5  2.06  1.72
## 11 User~    1.44     1.46     1.49     5       5       5  4.42  1.74
## 12 User~    1.46     1.48     1.5      5       5       5  4.43  1.74
## 13 User~    1.43     1.45     1.47     3.56    5       5  4.43  1.8
## 14 User~    1.47     2.2      1.5      5       5       5  4.44  1.79
## 15 User~    1.49     1.51     1.54     5       5       5  4.45  1.72
## 16 User~    1.47     1.5      1.52     5       5       5  4.44  1.73
## 17 User~    1.61     1.63     1.66     5       5       5  1.68  1.67
## 18 User~    1.76     1.78     1.8      5       5       5  1.63  1.62
## 19 User~    1.78     2.03     2.05     5       5       5  3.25  3.08
## 20 User~    2.6      3.08     5       5       5       5  1.63  1.57
## # ... with 16 more variables: restaurants <dbl>, `pubs/bars` <dbl>, `local
## #   services` <dbl>, `burger/pizza shops` <dbl>, `hotels/other`
## #   lodgings` <dbl>, `juice bars` <dbl>, `art galleries` <dbl>, `dance
## #   clubs` <dbl>, `swimming pools` <dbl>, gyms <dbl>, bakeries <dbl>,
## #   `beauty & spas` <dbl>, cafes <dbl>, `view points` <dbl>,
## #   monuments <dbl>, gardens <dbl>
```

U ovom kodu uzeli smo 500. unos iz podatkovnog skupa te smo pozvali funkciju `\texttt{n_max_idx}` kako bismo pronašli 20 najsličnijih korisnika tome korisniku.

```

df_neighbors <- df_numerical[idx, ]
corvalues <- apply(df_neighbors, 1, function(x) sim(x, test_row[-1]))

test_gyms <- test_row$`gyms` 

df_neighbor_gyms <- df_neighbors[ , c("gyms")]

prediction <- sum(corvalues * df_neighbor_gyms) / sum(abs(corvalues))
```

```
"Predikcija:"  
## [1] "Predikcija."  
test_gyms  
  
## [1] 0.57  
"Stvarna vrijednost:"  
## [1] "Stvarna vrijednost."  
prediction  
  
## [1] 0.5843514
```

Kao demonstraciju rada sustava preporuke uzeli smo tih 20 najsličnijih korisnika, izračunali njihovu sličnost s našim testnim vektorom te smo dali predikciju koliko bi iznosila vrijednost kategorije `gyms` ako bi je izračunali pomoću sličnosti tih 20 korisnika.

6 Zaključak

Primjenom statističkih i vizualizacijskih metoda uspješno smo napravili analizu skupa korisničkih ocjena sadržaja. Na početku smo uočili da ocjene niti jedne kategorije ne pripadaju normalnoj razdiobi. Također, kod promotrimo li histograme kategorija vidimo da su ekstremne ocjene prilično zastupljene u većini kategorija. Dakle, za većinu kategorija imamo velik broj ocjena 5 i veliku koncentraciju ocjena oko 1. Ovo se možemo objasniti pretpostavkom da korisnici ocjenjuju sadržaj koji im je ili jako dobar ili jako loš, dok rijetko daju manje ekstremne ocijene jer o onome što na njih ostavi snažniji dojam, bio on pozitivan ili negativan, više razmišljaju. Kako bi se ova pretpostavka potvrdila potrebno je provesti iscrpnije istraživanje. Provedenom analizom korelacija ocjena parova kategorija i statističkim testovima kojima su testirane njihove srednje vrijednosti dobili smo značajne informacije o preferencijama pojedinih korisnika koje se svakako mogu iskoristiti u dalnjim istraživanjima. U analizi smo koristili F-test, T-test, Kolmogorov-Smirnovljev test, Wilcoxonov test i Moodov test. Razmatrajući varijance ocjena u pojedinim kategorijama utvrđili smo da se korisnici najviše slažu oko ocjena crkvi te da su najviše podijeljeni oko umjetničkih galerija. Činjenica da su crkve najviše polarizirajuća kategorija zanimljiva je s obzirom na to da je religija često predmet rasprave pa je za očekivati da će crkve prouzrokovati podjelu među korisnicima. Zanimljivo je i da crkve imaju vrlo nisku prosječnu ocjenu: 1.51. S druge strane, neslaganje korisnika oko umjetničkih galerija je očekivano zbog toga što je umjetnost, kao i religija stvar osobnih preferencija. Provedenom linearnom regresijom utvrđili smo da, iako su korelacije između pojedinih kategorija relativno male te su njihove ocjene poprilično raspršene, linearni modeli koje smo istrenirali u prosjeku nisu u predikcijama promašivali više od jedne ocjene, što je zadovoljavajuće s obzirom na spomenute spoznaje o ocjenama. Važno je naglasiti da je dobivena pogreška izmjerena na istom podatkovnom skupu koji je korišten za trening modela pa bi se ona mogla razlikovati kada bismo dobivene modele iskoristili za predikciju na podatkovnim skupovima s još nevidenim podatcima. Razmatrajući QQ-plotove standardiziranih reziduala linearног modela te rezultate provedenih Kolmogorov-Smirnovljevih testova, zaključili smo da reziduali ne pripadaju normalnoj razdiobi, što znači da dobivenim linearnim modelima ne možemo u potpunosti vjerovati. To nimalo ne čudi s obzirom na velik broj stršećih vrijednosti ocjena. Ipak, predikcije korisničkih ocjena možemo vršiti s velikom pouzdanošću, ali ne na temelju ocjena neke drugih kategorija, već na temelju korisničkih ocjena drugih korisnika koji su slični korisniku čije ocjene želimo predvidjeti. Predviđanje korisničkih ocjena na temelju ocjena drugih korisnika proveli smo primjenom metode collaborative filteringa koja se zasniva na kosinusnoj sličnosti vektora čiji su elementi korisničke ocjene pojedinih kategorija. Pokazali smo da ovom matematički relativno jednostavnom metodom možemo dobiti izvrsne rezultate na problemu predviđanja korisničkih ocjena. Ova je metoda, stoga, temelj sustava preporučivanja, što nije čudno s obzirkom na njezinu prikazanu moć i jednostavnost.