User Ratings Data Analysis

Marko Barišić, Lovro Matošević, Daniel Vusić

UVOD

Zbog razvoja tehnologije i sve veće umreženosti korisnika, danas je dostupno sve više podataka o korisničkim preferencama koje su prije svega važzan alat u sustavima za preporuke i online prodaji. Jedan takav slučcaj su i ocjene na servisu Google, na kojem korisnici ocjenjuju razlčiite tipove sadržzaja - od parkova i spomenika do pekara i restorana. Pri modeliranju ovakvih ocjena vrlo je bitno uzeti u obzir čiinjenicu da korisnici imaju različite preference i njihov ukus utječce na ocjene koje daju različitim sadržajima. Upravo zbog toga ovakve ocjene su vrlo korisne u modeliranju i predviđanju kakvu bi ocjenu mogao imati određeni sadržaj, odnosno kojem korisniku bi se kakav tip sadržaaja mogao svidjeti. U ovom projektu naglasak će biti na statističko zaključivanje vezano uz korisničke ocjene sadržaja, što je bitan korak u gradnji naprednih sustava za preporučiivanje kakvi se danas koriste u mnogim komercijalnim primjenama.

UČITAVANJE PODATAKA

```
pod = read.csv("google_review_ratings.csv", fill = TRUE,stringsAsFactors=FALSE)
head(pod)
```

```
##     User churches resorts beaches parks theatres museums malls  zoo restaurants
## 1 User 1        0     0.0    3.63  3.65        5    2.92     5 2.35        2.33
## 2 User 2        0     0.0    3.63  3.65        5    2.92     5 2.64        2.33
## 3 User 3        0     0.0    3.63  3.63        5    2.92     5 2.64        2.33
## 4 User 4        0     0.5    3.63  3.63        5    2.92     5 2.35        2.33
## 5 User 5        0     0.0    3.63  3.63        5    2.92     5 2.64        2.33
## 6 User 6        0     0.0    3.63  3.63        5    2.92     5 2.63        2.33
##   pubs.bars local.services burger.pizza.shops hotels.other.lodgings juice.bars
## 1      2.64            1.7               1.69                  1.70       1.72
## 2      2.65            1.7               1.69                  1.70       1.72
## 3      2.64            1.7               1.69                  1.70       1.72
## 4      2.64           1.73               1.69                  1.70       1.72
## 5      2.64            1.7               1.69                  1.70       1.72
## 6      2.65           1.71               1.69                  1.69       1.72
##   art.galleries dance.clubs swimming.pools gyms bakeries beauty...spas cafes
## 1          1.74        0.59            0.5    0      0.5             0     0
## 2          1.74        0.59            0.5    0      0.5             0     0
## 3          1.74        0.59            0.5    0      0.5             0     0
## 4          1.74        0.59            0.5    0      0.5             0     0
## 5          1.74        0.59            0.5    0      0.5             0     0
## 6          1.74        0.59            0.5    0      0.5             0     0
##   view.points monuments gardens
## 1           0         0       0
## 2           0         0       0
## 3           0         0       0
## 4           0         0       0
## 5           0         0       0
## 6           0         0       0
```

```
pod[pod==0] <- NA

head(pod)
```

```
##     User churches resorts beaches parks theatres museums malls  zoo restaurants
## 1 User 1       NA      NA    3.63  3.65        5    2.92     5 2.35        2.33
## 2 User 2       NA      NA    3.63  3.65        5    2.92     5 2.64        2.33
## 3 User 3       NA      NA    3.63  3.63        5    2.92     5 2.64        2.33
## 4 User 4       NA     0.5    3.63  3.63        5    2.92     5 2.35        2.33
## 5 User 5       NA      NA    3.63  3.63        5    2.92     5 2.64        2.33
## 6 User 6       NA      NA    3.63  3.63        5    2.92     5 2.63        2.33
##   pubs.bars local.services burger.pizza.shops hotels.other.lodgings juice.bars
## 1      2.64            1.7               1.69                  1.70       1.72
## 2      2.65            1.7               1.69                  1.70       1.72
## 3      2.64            1.7               1.69                  1.70       1.72
## 4      2.64           1.73               1.69                  1.70       1.72
## 5      2.64            1.7               1.69                  1.70       1.72
## 6      2.65           1.71               1.69                  1.69       1.72
##   art.galleries dance.clubs swimming.pools gyms bakeries beauty...spas cafes
## 1          1.74        0.59            0.5   NA      0.5            NA    NA
## 2          1.74        0.59            0.5   NA      0.5            NA    NA
## 3          1.74        0.59            0.5   NA      0.5            NA    NA
## 4          1.74        0.59            0.5   NA      0.5            NA    NA
## 5          1.74        0.59            0.5   NA      0.5            NA    NA
## 6          1.74        0.59            0.5   NA      0.5            NA    NA
##   view.points monuments gardens
## 1          NA        NA      NA
## 2          NA        NA      NA
## 3          NA        NA      NA
## 4          NA        NA      NA
## 5          NA        NA      NA
## 6          NA        NA      NA
```

```
pod[1] <- NULL

pod = transform(pod, local.services = as.numeric(local.services))
```

```
## Warning in eval(substitute(list(...)), `_data`, parent.frame()): NAs introduced
## by coercion
```

```
head(pod)
```

```
##   churches resorts beaches parks theatres museums malls  zoo restaurants
## 1       NA      NA    3.63  3.65        5    2.92     5 2.35        2.33
## 2       NA      NA    3.63  3.65        5    2.92     5 2.64        2.33
## 3       NA      NA    3.63  3.63        5    2.92     5 2.64        2.33
## 4       NA     0.5    3.63  3.63        5    2.92     5 2.35        2.33
## 5       NA      NA    3.63  3.63        5    2.92     5 2.64        2.33
## 6       NA      NA    3.63  3.63        5    2.92     5 2.63        2.33
##   pubs.bars local.services burger.pizza.shops hotels.other.lodgings juice.bars
## 1      2.64            1.70               1.69                  1.70       1.72
## 2      2.65            1.70               1.69                  1.70       1.72
## 3      2.64            1.70               1.69                  1.70       1.72
## 4      2.64            1.73               1.69                  1.70       1.72
## 5      2.64            1.70               1.69                  1.70       1.72
```

```
## 6      2.65           1.71               1.69                  1.69       1.72
##   art.galleries dance.clubs swimming.pools gyms bakeries beauty...spas cafes
## 1          1.74        0.59            0.5   NA      0.5            NA    NA
## 2          1.74        0.59            0.5   NA      0.5            NA    NA
## 3          1.74        0.59            0.5   NA      0.5            NA    NA
## 4          1.74        0.59            0.5   NA      0.5            NA    NA
## 5          1.74        0.59            0.5   NA      0.5            NA    NA
## 6          1.74        0.59            0.5   NA      0.5            NA    NA
##   view.points monuments gardens
## 1          NA         NA      NA
## 2          NA         NA      NA
## 3          NA         NA      NA
## 4          NA         NA      NA
## 5          NA         NA      NA
## 6          NA         NA      NA
```

```
# str(pod)
#tail(pod)

mean_data = colMeans(pod, na.rm = TRUE)

head(pod)
```

```
##   churches resorts beaches parks theatres museums malls  zoo restaurants
## 1       NA      NA    3.63  3.65        5    2.92     5 2.35        2.33
## 2       NA      NA    3.63  3.65        5    2.92     5 2.64        2.33
## 3       NA      NA    3.63  3.63        5    2.92     5 2.64        2.33
## 4       NA     0.5    3.63  3.63        5    2.92     5 2.35        2.33
## 5       NA      NA    3.63  3.63        5    2.92     5 2.64        2.33
## 6       NA      NA    3.63  3.63        5    2.92     5 2.63        2.33
##   pubs.bars local.services burger.pizza.shops hotels.other.lodgings juice.bars
## 1      2.64           1.70               1.69                  1.70       1.72
## 2      2.65           1.70               1.69                  1.70       1.72
## 3      2.64           1.70               1.69                  1.70       1.72
## 4      2.64           1.73               1.69                  1.70       1.72
## 5      2.64           1.70               1.69                  1.70       1.72
## 6      2.65           1.71               1.69                  1.69       1.72
##   art.galleries dance.clubs swimming.pools gyms bakeries beauty...spas cafes
## 1          1.74        0.59            0.5   NA      0.5            NA    NA
## 2          1.74        0.59            0.5   NA      0.5            NA    NA
## 3          1.74        0.59            0.5   NA      0.5            NA    NA
## 4          1.74        0.59            0.5   NA      0.5            NA    NA
## 5          1.74        0.59            0.5   NA      0.5            NA    NA
## 6          1.74        0.59            0.5   NA      0.5            NA    NA
##   view.points monuments gardens
## 1          NA         NA      NA
## 2          NA         NA      NA
## 3          NA         NA      NA
## 4          NA         NA      NA
## 5          NA         NA      NA
## 6          NA         NA      NA
```

```
variance_data = sapply(pod,var,na.rm = T)

sd_data = sqrt(variance_data)
```

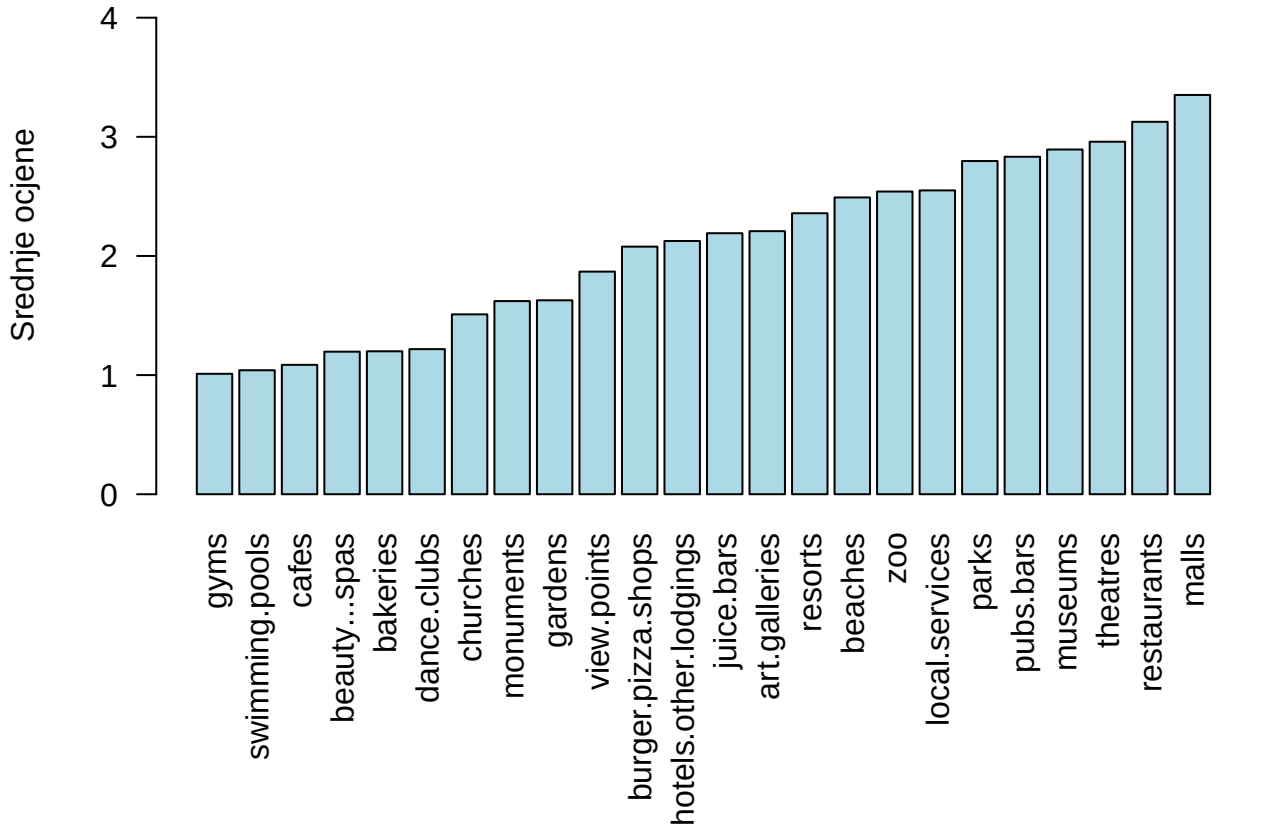
```
mean_data = sort(mean_data)
variance_data = sort(variance_data)
sd_data = sort(sd_data)


View(mean_data)
View(variance_data)
View(sd_data)

#
#
#
#
#
```

Bootstrap funkcija

```
bootstrapmeanpairedinterval <- function(data1, data2, alfa, n){

  data = c(data1 - data2, na.rm=TRUE)

  #dist1 = bootstrap(data1, n, mean)
  #dist2 = bootstrap(data2, n, mean)

  dist = bootstrap(data, n, mean, na.rm=TRUE)$thetastar

  lb = quantile(dist, alfa/2, na.rm = TRUE)
  ub = quantile(dist, 1 - alfa / 2, na.rm = TRUE)

  return(list(lb=lb,ub=ub,dist=dist))

}

bootstrapvariantpairedinterval <- function(data1, data2, alfa, n){

  dist1 = bootstrap(data1, n, var, na.rm=TRUE)$thetastar
  dist2 = bootstrap(data2, n, var, na.rm=TRUE)$thetastar


  dist = c(dist1/dist2)


  lb = quantile(dist, alfa/2, na.rm = TRUE)
  ub = quantile(dist, 1 - alfa / 2, na.rm = TRUE)

  return(list(lb=lb,ub=ub,dist=dist))


}
```

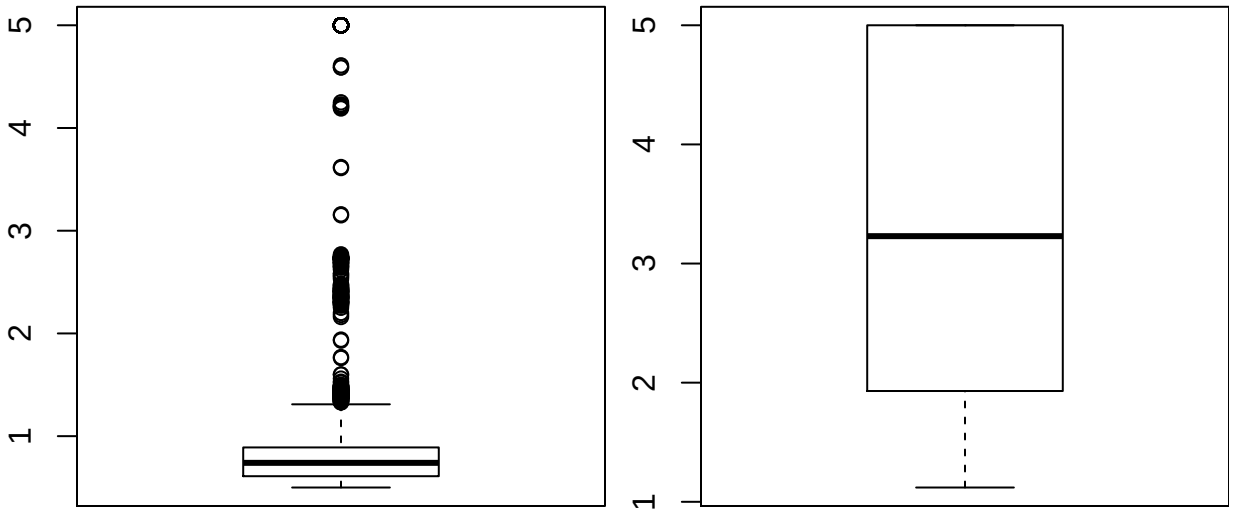
Pitanje: Usporedite odabrane kategorije po ocjenama - razlikuju li se znacajno po srednjoj ocjeni?

```
par(mar = c(9, 4, 0, 0))
barplot(mean_data, ylab="Srednje ocjene", ylim=c(0,1 + max(mean_data)) ,
col="lightblue", las=2)
```




```
# Najveca i najmanje srednja vrijednost
par(mfrow=c(1,2), mar=c(2.5,2.5,5,0), oma = c(0, 0, 2, 0))
boxplot(pod$gyms, na.rm=TRUE, names = "Gyms")
boxplot(pod$malls, na.rm=TRUE, names = "Malls")
mtext("Gyms & Malls",outer = TRUE,cex=1.5,font=2)
```

Gyms & Malls




```
# t testovi za slicne kategorije koje bi mogle imati iste srednje
# vrijednosti bez znacajne razlike
t.test(pod$beauty...spas,pod$bakeries,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$beauty...spas and pod$bakeries
## t = 2.7913, df = 4177, p-value = 0.005273
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01901374 0.10875075
## sample estimates:
## mean of the differences
##              0.06388224
```

```
t.test(pod$swimming.pools,pod$gyms,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$swimming.pools and pod$gyms
## t = 1.3895, df = 4369, p-value = 0.1647
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.007482266  0.043898742
## sample estimates:
## mean of the differences
##              0.01820824
```

```r
t.test(pod$monuments,pod$gardens,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$monuments and pod$gardens
## t = -0.98592, df = 5152, p-value = 0.3242
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05480992  0.01812837
## sample estimates:
## mean of the differences
##             -0.01834077
```

```r
t.test(pod$theatres,pod$museums,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$theatres and pod$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
```

```r
t.test(pod$dance.clubs,pod$beauty...spas,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$dance.clubs and pod$beauty...spas
## t = 0.23968, df = 4491, p-value = 0.8106
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04114049  0.05260086
## sample estimates:
## mean of the differences
##             0.005730187
```

```r
t.test(pod$juice.bars,pod$burger.pizza.shops,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$juice.bars and pod$burger.pizza.shops
## t = 5.0825, df = 5454, p-value = 3.85e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06880416 0.15521050
```

```
## sample estimates:
## mean of the differences
##               0.1120073
```

```
t.test(pod$gardens,pod$parks,alternative = "two.sided", paired = TRUE,
       na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$gardens and pod$parks
## t = -51.602, df = 5229, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.258655 -1.166519
## sample estimates:
## mean of the differences
##               -1.212587
```

```
# Iz testova mozemo zakljuciti da sljedece kategorije nemaju znacajnu
# razliku u srednjoj vrijednosti:
  # Swimming.pools i Gyms
  # Monuments i gardens
  # Dance.clubs i beauty spas

# Također ćemo napraviti par t trstova za kategorije koje smatramo da bi
# mogle biti nekakve suprotnosti
t.test(pod$burger.pizza.shops,pod$gyms,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$burger.pizza.shops and pod$gyms
## t = 47.244, df = 4437, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.113328 1.209727
## sample estimates:
## mean of the differences
##                1.161528
```

```
t.test(pod$art.galleries,pod$dance.clubs,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$art.galleries and pod$dance.clubs
## t = 37.347, df = 5339, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9407536 1.0449880
## sample estimates:
## mean of the differences
##               0.9928708
```

```r
t.test(pod$malls,pod$museums,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$malls and pod$museums
## t = 22.525, df = 5455, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4180518 0.4977584
## sample estimates:
## mean of the differences
##               0.4579051
```

```r
t.test(pod$pubs.bars,pod$gardens,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$pubs.bars and pod$gardens
## t = 43.691, df = 5229, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.133918 1.240457
## sample estimates:
## mean of the differences
##                1.187187
```

```r
t.test(pod$monuments,pod$beauty...spas,alternative = "two.sided",
       paired = TRUE, na.rm = TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  pod$monuments and pod$beauty...spas
## t = 19.251, df = 4423, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4571876 0.5608640
## sample estimates:
## mean of the differences
##               0.5090258
```

```r
# Iz testova možemo zaključiti da se testirane kategorije znatno
# razlikuju u srednjoj vrijednosti

# Iz razloga što podatci ne prate normalnu distribuciju moramo nad
# njima raditi neparametarske testove


#   Radimo wilcoxonov test predznačnih rangova
wilcox.test(pod$beauty...spas,pod$bakeries,alternative = "two.sided",
            paired = TRUE, na.rm = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  pod$beauty...spas and pod$bakeries
## V = 4463518, p-value = 4.035e-12
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(pod$swimming.pools,pod$gyms,alternative = "two.sided",
            paired = TRUE, na.rm = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  pod$swimming.pools and pod$gyms
## V = 3508729, p-value = 0.001133
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(pod$monuments,pod$gardens,alternative = "two.sided",
            paired = TRUE, na.rm = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  pod$monuments and pod$gardens
## V = 3289604, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(pod$theatres,pod$museums,alternative = "two.sided",
            paired = TRUE, na.rm = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  pod$theatres and pod$museums
## V = 5830466, p-value = 0.0008045
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(pod$dance.clubs,pod$beauty...spas,alternative = "two.sided",
            paired = TRUE, na.rm = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  pod$dance.clubs and pod$beauty...spas
## V = 5028066, p-value = 0.0371
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(pod$juice.bars,pod$burger.pizza.shops,alternative = "two.sided",
            paired = TRUE, na.rm = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  pod$juice.bars and pod$burger.pizza.shops
## V = 6182516, p-value = 0.01269
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(pod$gardens,pod$parks,alternative = "two.sided",
            paired = TRUE, na.rm = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  pod$gardens and pod$parks
## V = 1905119, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
#   Bootstrap testovi za provjeru imaju li navedene kategorije znacajne
# razlike u srednjoj vrijednosti

library(bootstrap)
print("Beauty...spas & Bakeries")
```

```
## [1] "Beauty...spas & Bakeries"
```

```r
spasBeakeries = bootstrapmeanpairedinterval(pod$beauty...spas,
                                            pod$bakeries, 0.05, 1000)
spasBeakeries$lb
```

```
##       2.5%
## 0.02187121
```

```r
spasBeakeries$ub
```

```
##     97.5%
## 0.1116126
```

```r
print("Swimming.pools & Gyms")
```

```
## [1] "Swimming.pools & Gyms"
```

```r
swimmingpoolsGyms = bootstrapmeanpairedinterval(pod$swimming.pools,
                                                pod$gyms , 0.05, 1000)
swimmingpoolsGyms$lb
```

```
##         2.5%
## -0.008323651
```

```r
swimmingpoolsGyms$ub
```

```
##      97.5%
## 0.04431904
```

```r
print("Monuments & Gardens")
```

```
## [1] "Monuments & Gardens"
```

```r
monumentsGardens = bootstrapmeanpairedinterval(pod$monuments,
                                               pod$gardens, 0.05, 1000)
monumentsGardens$lb
```

```
##        2.5%
## -0.05393221
```

```r
monumentsGardens$ub
```

```
##      97.5%
## 0.01966991
```

```r
print("Theatres & Museums")
```

```
## [1] "Theatres & Museums"
```

```r
theatresMuseums = bootstrapmeanpairedinterval(pod$theatres,
                                              pod$museums, 0.05, 1000)
theatresMuseums$lb
```

```
##       2.5%
## 0.03052167
```

```r
theatresMuseums$ub
```

```
##      97.5%
## 0.09901444
```

```r
print("Dance.Clubs & Beauty...spas")
```

```
## [1] "Dance.Clubs & Beauty...spas"
```

```r
danceBeauty = bootstrapmeanpairedinterval(pod$dance.clubs,
                                          pod$beauty...spas, 0.05, 1000)
danceBeauty$lb
```

```
##        2.5%
## -0.04423529
```

```r
danceBeauty$ub
```

```
##      97.5%
## 0.04996916
```

```r
print("Juice bars & Burger pizza shops")
```

```
## [1] "Juice bars & Burger pizza shops"
```

```r
juiceBurger = bootstrapmeanpairedinterval(pod$juice.bars,
                                          pod$burger.pizza.shops,
                                          0.05, 1000)
juiceBurger$lb
```

```
##       2.5%
## 0.07136728
```

```r
juiceBurger$ub
```

```
##     97.5%
## 0.1565551
```

```r
print("Gardens & parks")
```

```
## [1] "Gardens & parks"
```

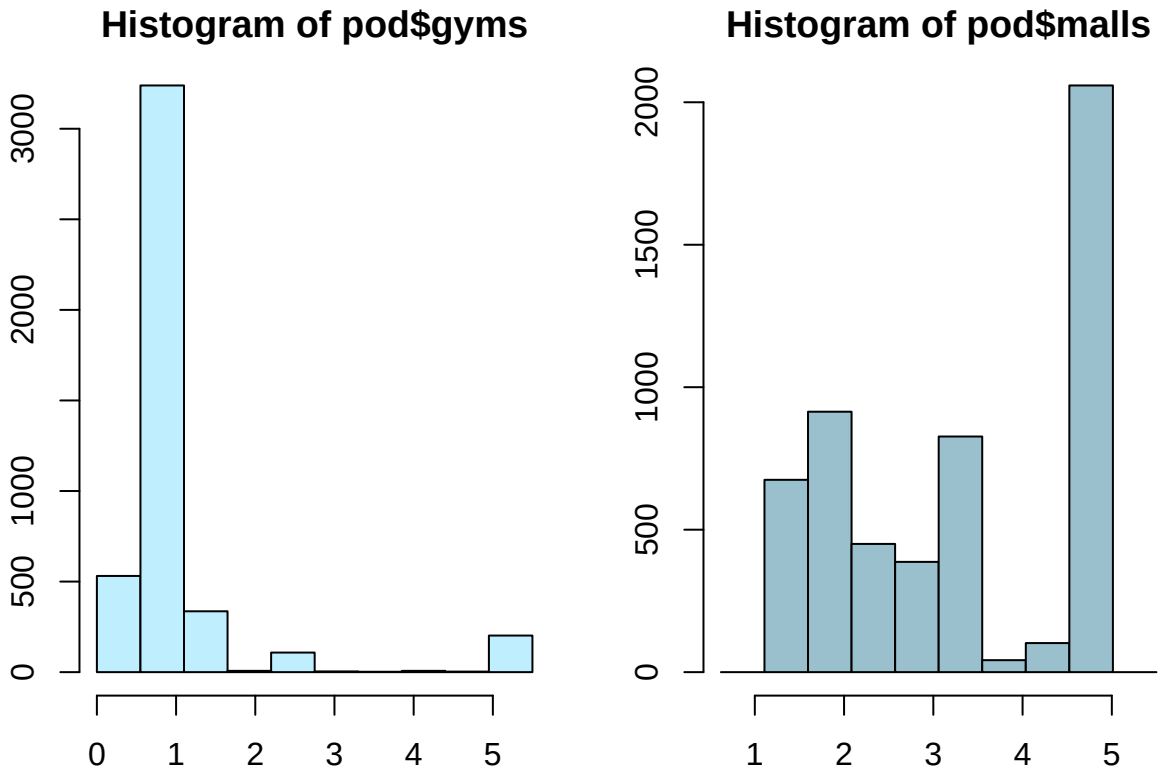
```r
gardensParks = bootstrapmeanpairedinterval(pod$gardens,
                                           pod$parks, 0.05, 1000)
gardensParks$lb
```

```
##      2.5%
## -1.259574
```

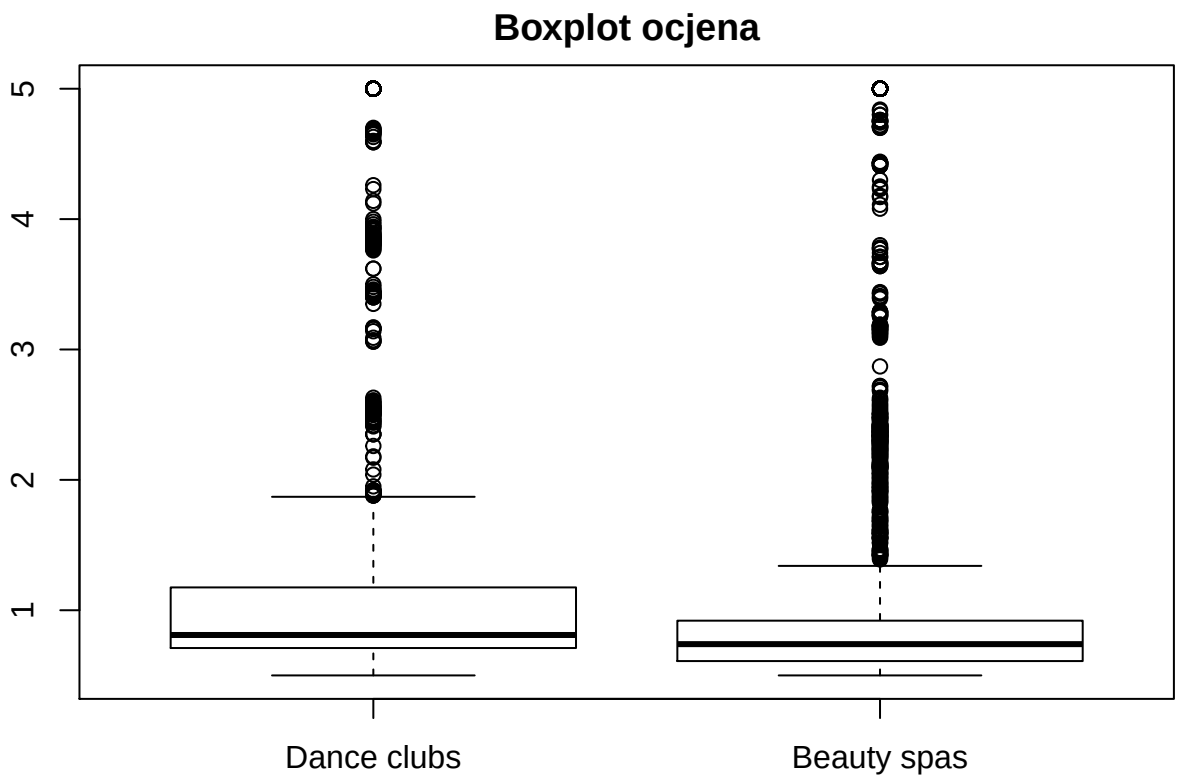
```r
gardensParks$ub
```

```
##    97.5%
## -1.16957
```

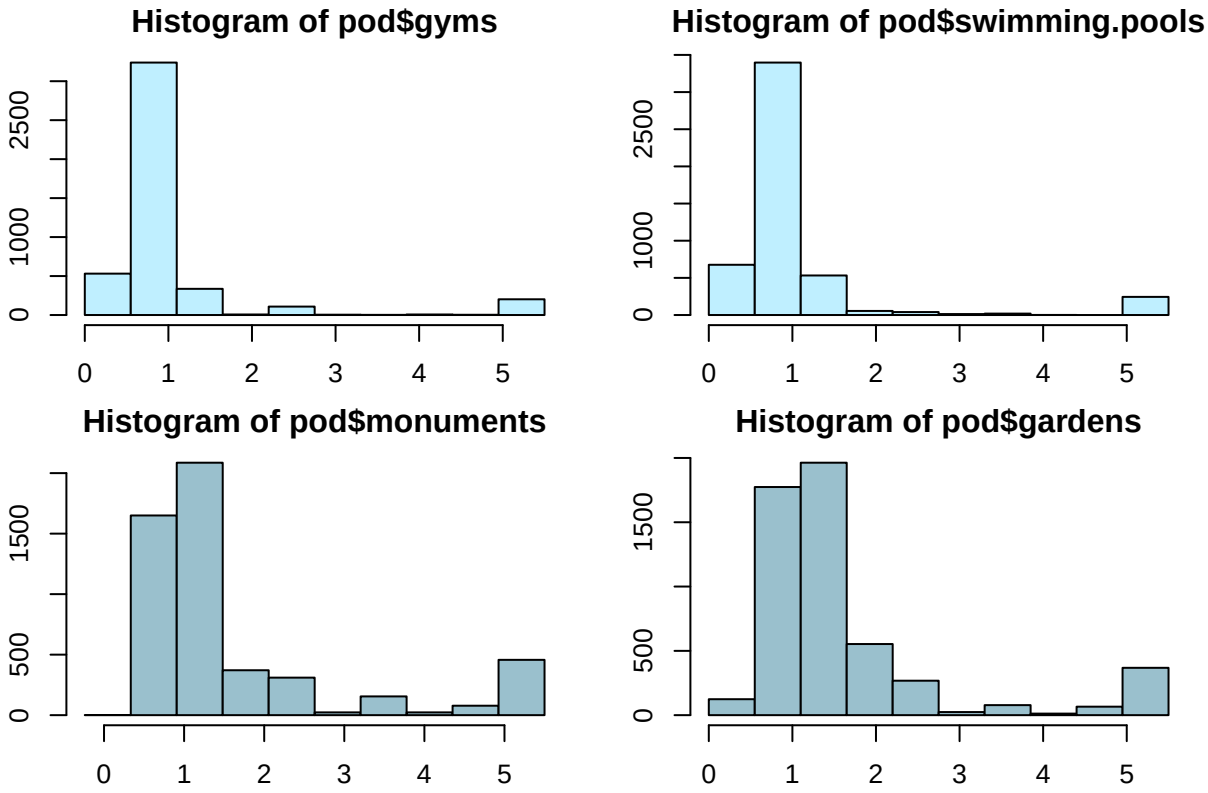
```
#   Iz navedenih bootstrap testova mozemo zakljuciti da se bootstrap
# testovi podudaraju sa izvedenim t testovima sto nam sa sigurnoscu
# ukazuje da kategorije sljedecih navedenih kategorija nemaju znacajnu
# razliku u srednjim vrijednostima:
#       Swimming.pools i Gyms
#       Monuments i gardens
#       Dance.clubs i beauty spas


# Prikazujemo kategorije najveće i najmanje srednje vrijednosti
par(mfrow=c(1,2),mar = c(2, 2, 2, 2))
hist(pod$gyms, col="lightblue1", breaks = seq(from = min(pod$gyms ,na.rm = TRUE)
                      - 0.5, to = max(pod$gyms, na.rm=TRUE) + 0.5, length = 11))
hist(pod$malls, col="lightblue3", breaks = seq(from = min(pod$malls ,na.rm = TRUE)
                    - 0.5, to = max(pod$malls, na.rm=TRUE) + 0.5, length = 11))
```




```
par(mfrow=c(1,1),mar = c(2, 2, 2, 2))
boxplot(pod$dance.clubs,pod$beauty...spas, names=c("Dance clubs", "Beauty spas"),
        na.rm=TRUE, main = "Boxplot ocjena")
```




```
# Kategorije sa najslicnijim srednjim ocjenama
par(mfrow=c(2,2),mar = c(2, 2, 2, 2))
hist(pod$gyms, col="lightblue1", breaks = seq(from = min(pod$gyms ,na.rm = TRUE)
                      - 0.5, to = max(pod$gyms, na.rm=TRUE) + 0.5, length = 11))
hist(pod$swimming.pools, col="lightblue1", breaks = seq(from =
    min(pod$swimming.pools ,na.rm = TRUE) - 0.5,
    to = max(pod$swimming.pools, na.rm=TRUE) + 0.5, length = 11))
hist(pod$monuments, col="lightblue3",
  breaks = seq(from = min(pod$monuments ,na.rm = TRUE) - 0.5,
    to = max(pod$monuments, na.rm=TRUE) + 0.5, length = 11))
hist(pod$gardens, col="lightblue3",
breaks = seq(from = min(pod$gardens ,na.rm = TRUE) - 0.5,
      to = max(pod$gardens, na.rm=TRUE) + 0.5, length = 11))
```




```
#Uklanjanje false usera --početak

#Uklanjanje false usera je provedeno u srvhu normalizacije podataka.
#False user se identificira ako zadovoljava neki od sljedećih uvjeta
# > sve ocjene usera su manje-jednako 1, jednake 5 ili NA
# > sredina (mean) ocjena je manje od 1.5
# > sredina (mean) ocjena je veća od 4.5

pod.no.false = pod
false.users = c()
false.sredine = c()
pod.no.false = pod
for(red in 1:nrow(pod)) {
  f = TRUE
  ocjene = c(pod[red,2],pod[red,3],pod[red,4],pod[red,5],pod[red,6],
             pod[red,7],pod[red,8],pod[red,9],pod[red,10],pod[red,11],
             pod[red,12],pod[red,13],pod[red,14],pod[red,15],pod[red,16],
             pod[red,17],pod[red,18],pod[red,19],pod[red,20],pod[red,21],
             pod[red,22],pod[red,23],pod[red,24],pod[red,25])
  for(x in ocjene){
    if(!is.na(x) && x > 1 && x < 5){
      f = FALSE
      break()
    }
  }
  sredina = mean(ocjene, na.rm = TRUE)
  if(f || sredina < 1.5 || sredina > 4.5) {
    false.users = c(false.users, red)
    false.sredine = c(false.sredine, sredina)
```

```
    pod.no.false=pod.no.false[-red,]
  }
}

#Ispis false usera i njihovih sredina, NAN  vrijednost sredine označava
# da su sve vrijednosti usera NA
print("Flase users: ")
```

```
## [1] "Flase users: "
```

```
false.users
```

```
##  [1] 1349 2388 2390 2424 2715 2807 3348 4201 4525 5179
```

```
false.sredine
```

```
##  [1]      NaN 1.482609 1.480000 1.462174      NaN 1.483478 1.449565 1.487826
##  [9] 1.495652 1.430952
```

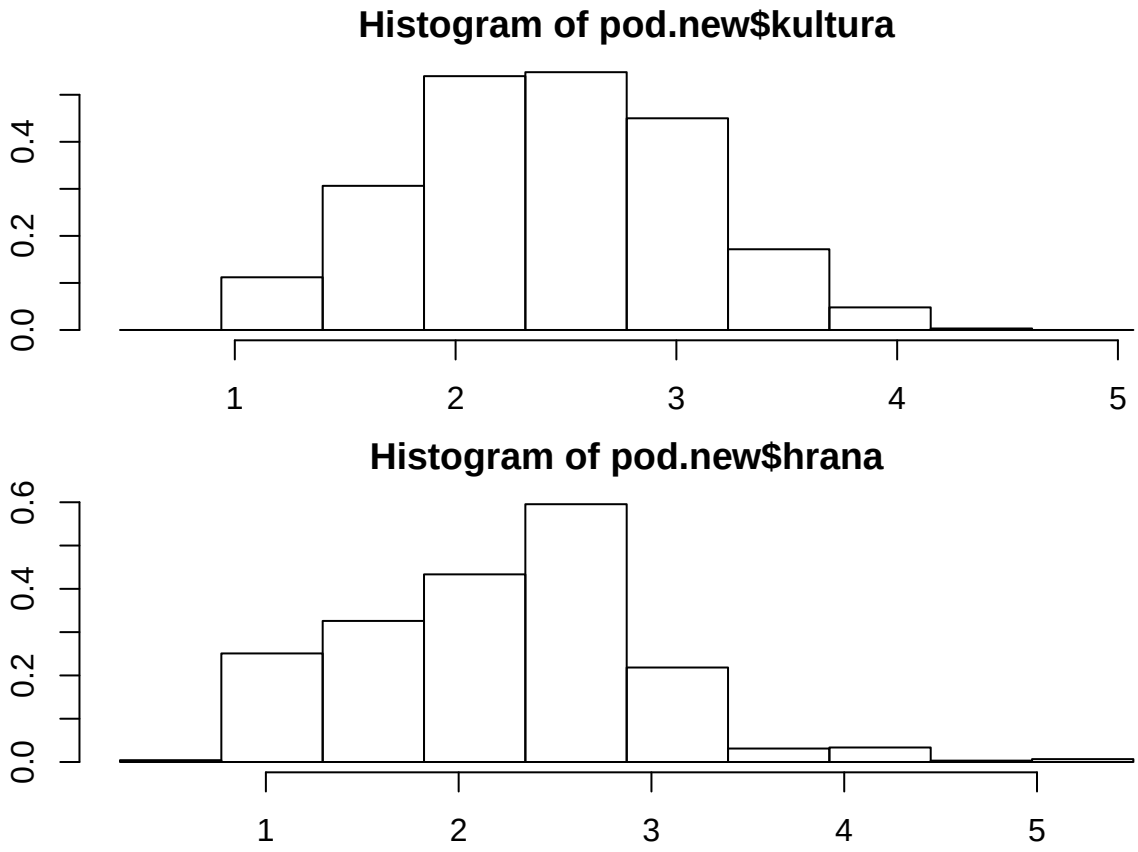
```
pod = pod.no.false
#Uklanjanje false usera --kraj



#Grupiranje podataka --početak

#Podaci su grupirani također u svrhu normalizacije.
#Podaci su grupirani u 9 klasa (grupa); kultura, hrana, pice, ugostiteljski,
# priroda, zabava, sport, religiozni i ostalo.

grupa.kultura = c("art.galleries","monuments","museums","theatres")
grupa.hrana = c("bakeries", "burger.pizza.shops", "restaurants")
grupa.pice = c("cafes","juice.bars","pubs.bars")
grupa.ugostiteljski = c("hotels.other.lodgings","resorts")
grupa.priroda  = c("beaches","gardens","parks","view.points")
grupa.zabava  = c("beauty...spas","dance.clubs","malls","zoo")
grupa.sport = c("gyms","swimming.pools")
grupa.religiozni = c("churches")
grupa.ostalo = c("local.services")

kultura = c()
hrana = c()
pice = c()
zabava = c()
ugostiteljski = c()
priroda = c()
sport = c()
religiozni = c()
ostalo = c()
User = c()

for (red in 1:nrow(pod)) {

  User = c(User, pod[red, "User"])

  pod.kultura = pod[red, grupa.kultura]
```

```
  vec.kultura = c(pod.kultura[1,1],pod.kultura[1,2],
                  pod.kultura[1,3],pod.kultura[1,4])
  kultura = c(kultura, mean(vec.kultura, na.rm=TRUE))

  pod.hrana = pod[red, grupa.hrana]
  vec.hrana = c(pod.hrana[1,1],pod.hrana[1,2],pod.hrana[1,3])
  hrana = c(hrana, mean(vec.hrana, na.rm=TRUE))

  pod.pice = pod[red, grupa.pice]
  vec.pice = c(pod.pice[1,1],pod.pice[1,2],pod.pice[1,3])
  pice = c(pice, mean(vec.pice, na.rm = TRUE))

  pod.ugostiteljski = pod[red, grupa.ugostiteljski]
  vec.ugostiteljski = c(pod.ugostiteljski[1,1], pod.ugostiteljski[1,2])
  ugostiteljski = c(ugostiteljski, mean(vec.ugostiteljski, na.rm=TRUE))

  pod.priroda = pod[red, grupa.priroda]
  vec.priroda = c(pod.priroda[1,1],pod.priroda[1,2],
                  pod.priroda[1,3],pod.priroda[1,4])
  priroda = c(priroda, mean(vec.priroda, na.rm=TRUE))

  pod.zabava = pod[red, grupa.zabava]
  vec.zabava = c(pod.zabava[1,1],pod.zabava[1,2],
                 pod.zabava[1,3],pod.zabava[1,4])
  zabava = c(zabava, mean(vec.zabava, na.rm = TRUE))

  pod.sport = pod[red, grupa.sport]
  vec.sport = c(pod.sport[1,1], pod.sport[1,2])
  sport = c(sport, mean(vec.sport, na.rm = TRUE))

  pod.religiozni = pod[red, grupa.religiozni]
  religiozni = c(religiozni, mean(pod.religiozni, na.rm=TRUE))

  pod.ostalo = pod[red, grupa.ostalo]
  ostalo = c(ostalo, mean(pod.ostalo, na.rm=TRUE))

}

kultura[is.nan(kultura)] = NA
hrana[is.nan(hrana)] = NA
pice[is.nan(pice)] = NA
zabava[is.nan(zabava)] = NA
sport[is.nan(sport)] = NA
ugostiteljski[is.nan(ugostiteljski)] = NA
priroda[is.nan(priroda)] = NA
religiozni[is.nan(religiozni)] = NA
ostalo[is.nan(ostalo)] = NA

pod.new = data.frame(kultura, hrana, pice, zabava, sport, priroda,
                     ugostiteljski, religiozni, ostalo)
head(pod.new)
```

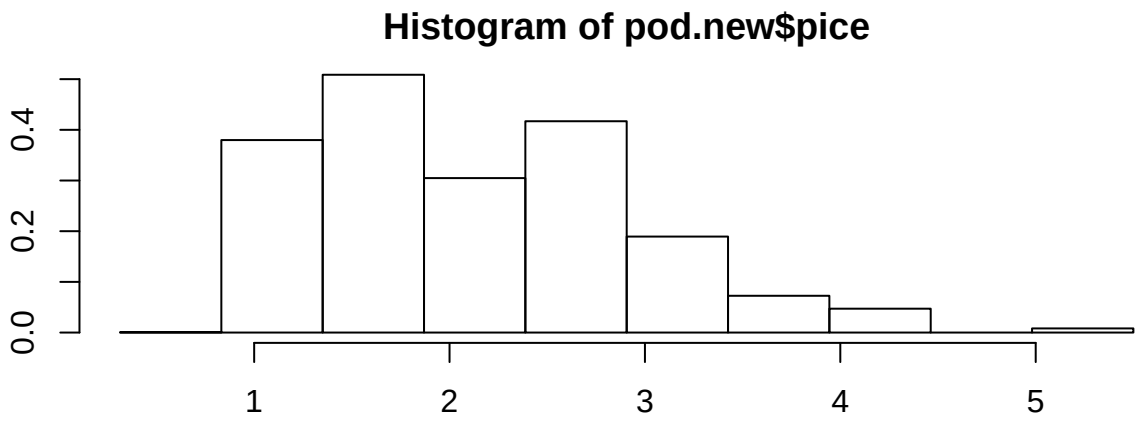
```
##   kultura    hrana  pice   zabava sport priroda ugostiteljski religiozni ostalo
## 1    3.22 1.506667 2.180 2.646667   0.5    3.64          1.70         NA   1.70
```

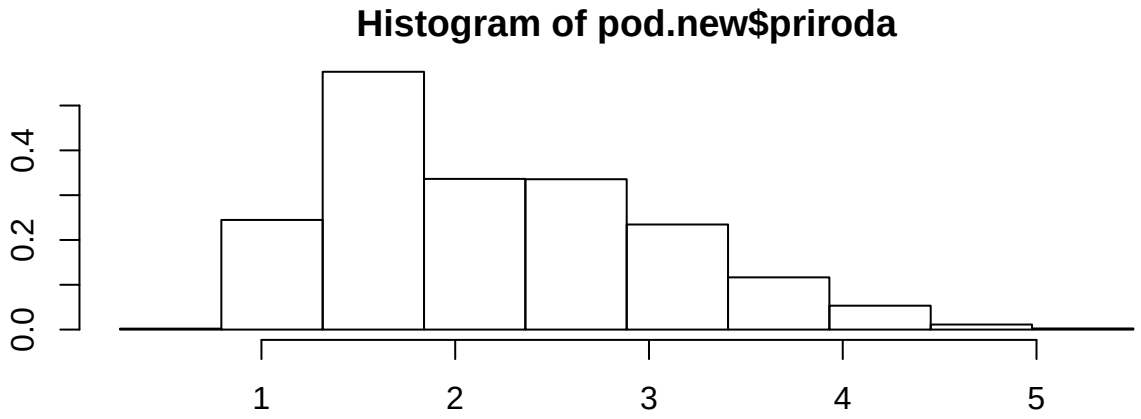
```
## 2    3.22 1.506667 2.185 2.743333   0.5    3.64          1.70         NA   1.70
## 3    3.22 1.506667 2.180 2.743333   0.5    3.63          1.70         NA   1.70
## 4    3.22 1.506667 2.180 2.646667   0.5    3.63          1.10         NA   1.73
## 5    3.22 1.506667 2.180 2.743333   0.5    3.63          1.70         NA   1.70
## 6    3.22 1.506667 2.185 2.740000   0.5    3.63          1.69         NA   1.71
```

```
#Prikazani su histogrami grupiranih podataka. Jasno vidimo da je
# postupak grupiranja pomogao pri normalizaciji podataka
par(mfrow=c(2,1),mar = c(2, 2, 2, 2))
hist(pod.new$kultura, freq=FALSE,
  breaks = seq(from = min(pod.new$kultura, na.rm = TRUE) - 0.5,
  to = max(pod.new$kultura, na.rm = TRUE) + 0.5, length = 11))
hist(pod.new$hrana, freq=FALSE,
  breaks = seq(from = min(pod.new$hrana, na.rm = TRUE) - 0.5,
  to = max(pod.new$hrana, na.rm = TRUE) + 0.5, length = 11))
```

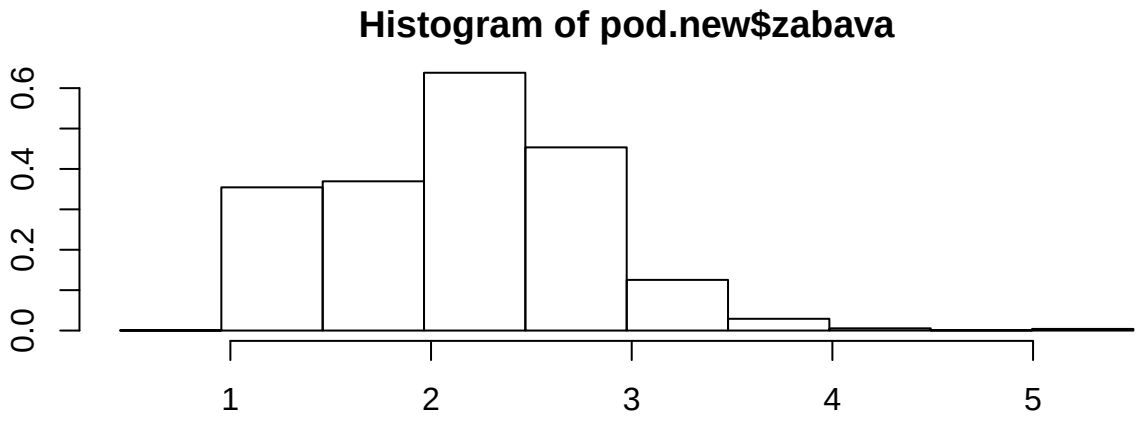


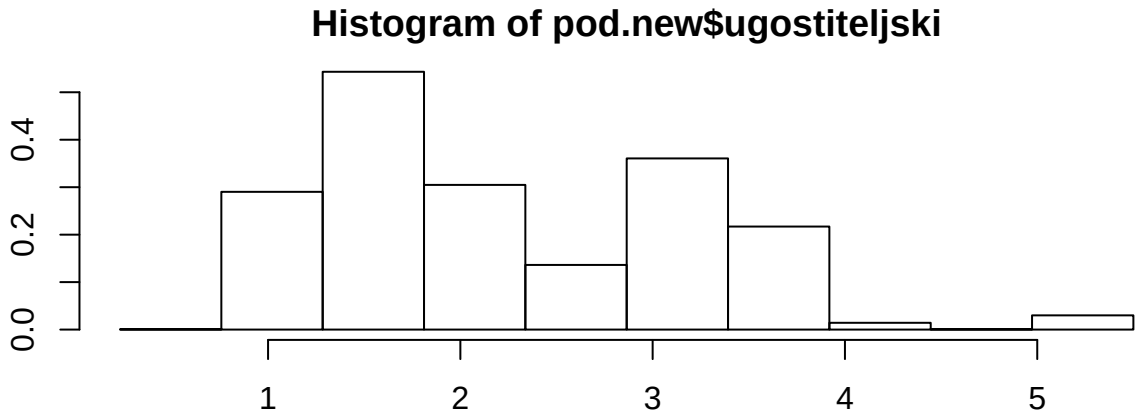

```
hist(pod.new$pice, freq=FALSE,
  breaks = seq(from = min(pod.new$pice, na.rm = TRUE) - 0.5,
  to = max(pod.new$pice, na.rm = TRUE) + 0.5, length = 11))
hist(pod.new$priroda, freq=FALSE,
  breaks = seq(from = min(pod.new$priroda, na.rm = TRUE) - 0.5,
  to = max(pod.new$priroda, na.rm = TRUE) + 0.5, length = 11))
```



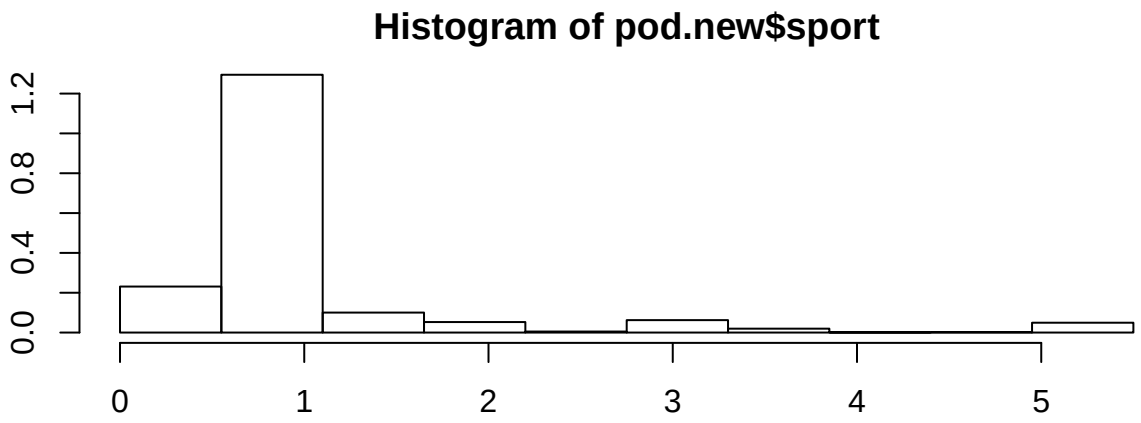


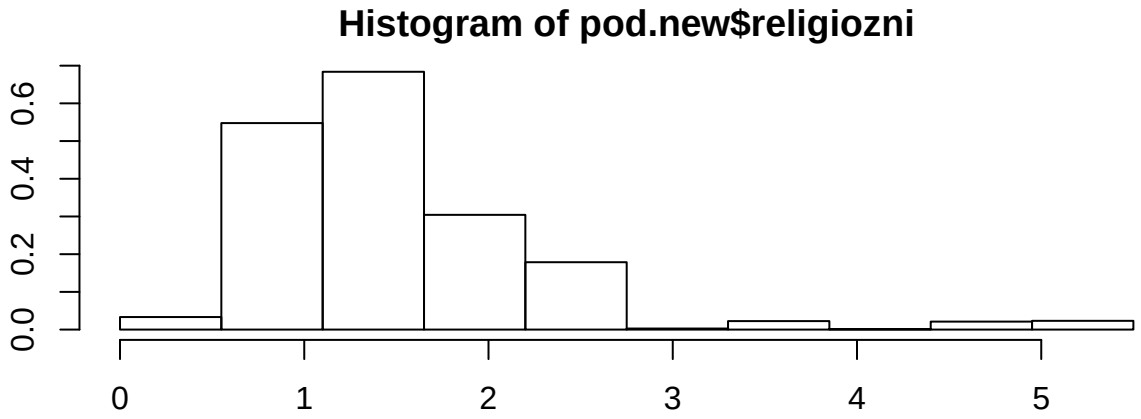

```r
hist(pod.new$zabava, freq=FALSE,
  breaks = seq(from = min(pod.new$zabava, na.rm = TRUE) - 0.5,
  to = max(pod.new$zabava, na.rm = TRUE) + 0.5, length = 11))
hist(pod.new$ugostiteljski, freq=FALSE,
  breaks = seq(from = min(pod.new$ugostiteljski, na.rm = TRUE) - 0.5,
  to = max(pod.new$ugostiteljski, na.rm = TRUE) + 0.5, length = 11))
```



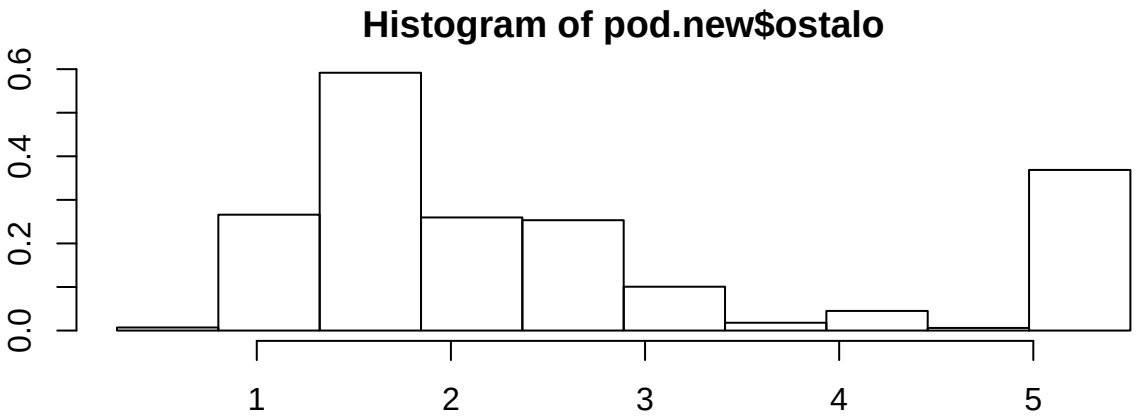



```
hist(pod.new$sport, freq=FALSE,
  breaks = seq(from = min(pod.new$sport, na.rm = TRUE) - 0.5,
  to = max(pod.new$sport, na.rm = TRUE) + 0.5, length = 11))
hist(pod.new$religiozni, freq=FALSE,
  breaks = seq(from = min(pod.new$religiozni, na.rm = TRUE) - 0.5,
  to = max(pod.new$religiozni, na.rm = TRUE) + 0.5, length = 11))
```

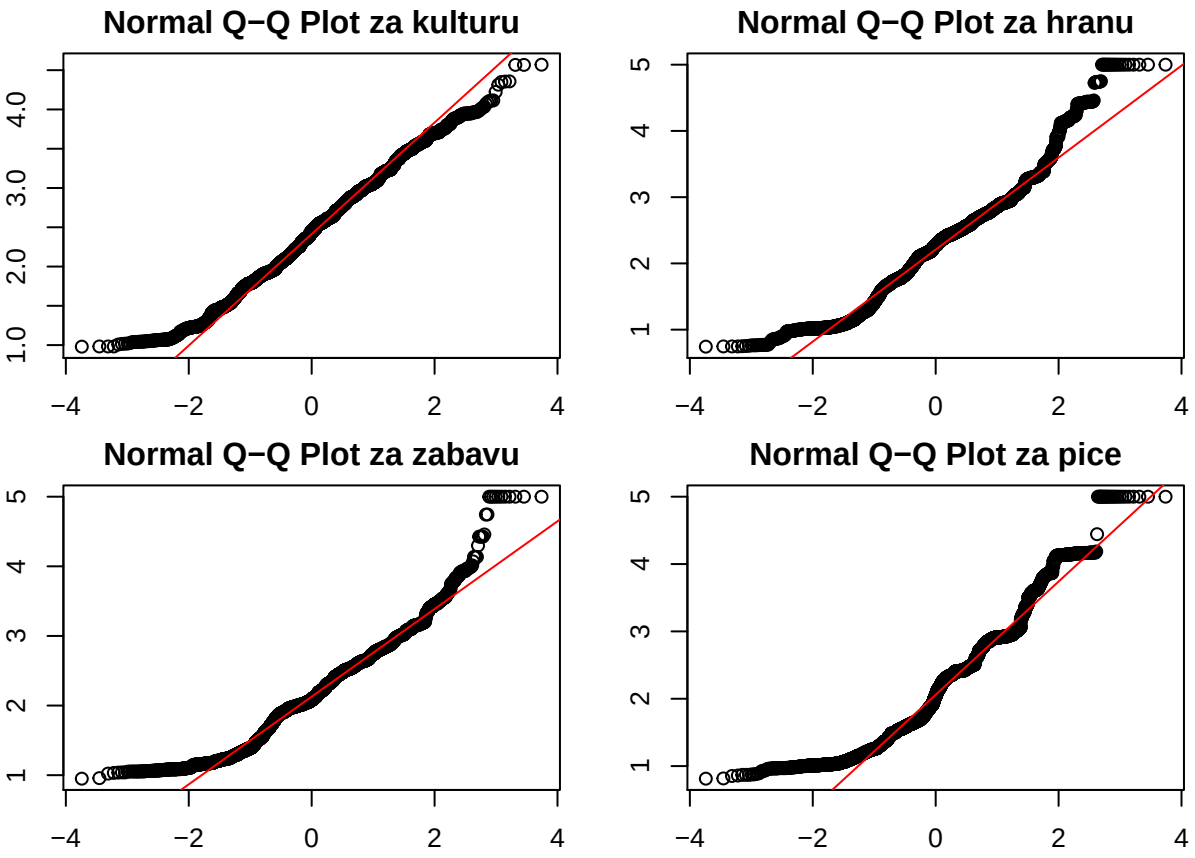





```r
hist(pod.new$ostalo, freq=FALSE,
  breaks = seq(from = min(pod.new$ostalo, na.rm = TRUE) - 0.5,
  to = max(pod.new$ostalo, na.rm = TRUE) + 0.5, length = 11))

#Grupiranje podataka --kraj



#T-testovi --početak
#Kako bi se dobio bolji uvid u podatke pri odabiru kategorija
# koje će ući u T-testove.
#Prema grupama, istim kao u dijelu gdje su podaci grupirani,
# stvorene su tablice srednjih vrijednosti kategorija unutar grupe.
#Iz takvihh tablica će se lakše odlučiti o kategorijama koje
# će ući u T-testove.

par(mfrow=c(2,2),mar = c(2, 2, 2, 2))
```




```
qqnorm(pod.new$kultura, main="Normal Q-Q Plot za kulturu")
qqline(pod.new$kultura, col='red')
qqnorm(pod.new$hrana, main="Normal Q-Q Plot za hranu")
qqline(pod.new$hrana, col='red')
qqnorm(pod.new$zabava, main="Normal Q-Q Plot za zabavu")
qqline(pod.new$zabava, col='red')
qqnorm(pod.new$pice, main="Normal Q-Q Plot za pice")
qqline(pod.new$pice, col='red')
```




```
#qqnorm(pod.new$zabava)
#qqline(pod.new$zabava, col='red')

kategorija = grupa.kultura
srednja.ocjena = c(mean(pod$art.galleries, na.rm = TRUE),
```

```
                  mean(pod$monuments, na.rm = TRUE),
                  mean(pod$museums, na.rm = TRUE),
                  mean(pod$theatres, na.rm = TRUE))
kultura = data.frame(kategorija, srednja.ocjena)



kategorija = grupa.hrana
srednja.ocjena = c(mean(pod$bakeries, na.rm = TRUE),
                   mean(pod$burger.pizza.shops, na.rm = TRUE),
                   mean(pod$restaurants, na.rm = TRUE))
hrana = data.frame(kategorija, srednja.ocjena)

kategorija = grupa.pice
srednja.ocjena = c(mean(pod$cafes, na.rm = TRUE),
                   mean(pod$juice.bars, na.rm = TRUE),
                   mean(pod$pubs.bars, na.rm = TRUE))
pice = data.frame(kategorija, srednja.ocjena)

kategorija = grupa.priroda
srednja.ocjena = c(mean(pod$beaches, na.rm = TRUE),
                   mean(pod$gardens, na.rm = TRUE),
                   mean(pod$parks, na.rm = TRUE),
                   mean(pod$view.points, na.rm = TRUE))
priroda = data.frame(kategorija, srednja.ocjena)

kategorija = grupa.ugostiteljski
srednja.ocjena = c(mean(pod$hotels.other.lodgings, na.rm = TRUE),
                   mean(pod$resorts, na.rm = TRUE))
ugostiteljski = data.frame(kategorija, srednja.ocjena)

kategorija = grupa.religiozni
srednja.ocjena = c(mean(pod$churches, na.rm = TRUE))
religiozni = data.frame(kategorija, srednja.ocjena)

kategorija = grupa.sport
srednja.ocjena = c(mean(pod$gyms, na.rm = TRUE),
                   mean(pod$swimming.pools, na.rm = TRUE))
sport = data.frame(kategorija, srednja.ocjena)

kategorija = grupa.zabava
srednja.ocjena = c(mean(pod$beauty...spas, na.rm = TRUE),
                   mean(pod$dance.clubs, na.rm = TRUE),
                   mean(pod$malls, na.rm = TRUE),
                   mean(pod$zoo, na.rm = TRUE))
zabava = data.frame(kategorija, srednja.ocjena)

kategorija = grupa.ostalo
srednja.ocjena = c(mean(pod$local.services, na.rm = TRUE))
ostalo = data.frame(kategorija, srednja.ocjena)

#Tablice grupiranih objekata
#KULTURA
```

```
#kultura
#-maksimalna srednja ocjena
#kultura[which(kultura$srednja.ocjena == max(kultura$srednja.ocjena)),]
#-minimalna srednja ocjena
#kultura[which(kultura$srednja.ocjena == min(kultura$srednja.ocjena)),]

#HRANA
#hrana
#-maksimalna srednja ocjena
#hrana[which(hrana$srednja.ocjena == max(hrana$srednja.ocjena)),]
#-minimalna srednja ocjena
#hrana[which(hrana$srednja.ocjena == min(hrana$srednja.ocjena)),]

#PICE
#pice
#-maksimalna srednja ocjena
#pice[which(pice$srednja.ocjena == max(pice$srednja.ocjena)),]
#-minimalna srednja ocjena
#pice[which(pice$srednja.ocjena == min(pice$srednja.ocjena)),]

#ZABAVA
#zabava
#-maksimalna srednja ocjena
#zabava[which(zabava$srednja.ocjena == max(zabava$srednja.ocjena)),]
#-minimalna srednja ocjena
#zabava[which(zabava$srednja.ocjena == min(zabava$srednja.ocjena)),]

#SPORT
#sport
#-maksimalna srednja ocjena
#sport[which(sport$srednja.ocjena == max(sport$srednja.ocjena)),]
#-minimalna srednja ocjena
#sport[which(sport$srednja.ocjena == min(sport$srednja.ocjena)),]

#UGOSTITELJSTVO
#ugostiteljski
#-maksimalna srednja ocjena
#ugostiteljski[which(ugostiteljski$srednja.ocjena ==
#max(ugostiteljski$srednja.ocjena)),]
#-minimalna srednja ocjena
#ugostiteljski[which(ugostiteljski$srednja.ocjena ==
#min(ugostiteljski$srednja.ocjena)),]

#PRIRODA
#priroda
#-maksimalna srednja ocjena
#priroda[which(priroda$srednja.ocjena == max(priroda$srednja.ocjena)),]
#-minimalna srednja ocjena
#priroda[which(priroda$srednja.ocjena == min(priroda$srednja.ocjena)),]

#RELIGIOZNI
#religiozni
#-maksimalna srednja ocjena
```

```r
#religiozni[which(religiozni$srednja.ocjena ==
#max(religiozni$srednja.ocjena)),]
#-minimalna srednja ocjena
#religiozni[which(religiozni$srednja.ocjena ==
#min(religiozni$srednja.ocjena)),]

#OSTALO
#ostalo
#-maksimalna srednja ocjena
#ostalo[which(ostalo$srednja.ocjena == max(ostalo$srednja.ocjena)),]
#-minimalna srednja ocjena
#ostalo[which(ostalo$srednja.ocjena == min(ostalo$srednja.ocjena)),]

#T-testovi u koje ulaze kategorije maksimalne i minimalne
# srednje ocjene za svaku grupu.
t.test(pod$theatres,pod$monuments, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod$theatres and pod$monuments
## t = 56.357, df = 5144, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.355733 1.453453
## sample estimates:
## mean of the differences
##                1.404593
```

```r
t.test(pod$restaurants,pod$bakeries, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod$restaurants and pod$bakeries
## t = 59.276, df = 4401, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.795447 1.918274
## sample estimates:
## mean of the differences
##                1.856861
```

```r
t.test(pod$pubs.bars,pod$cafes, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod$pubs.bars and pod$cafes
## t = 66.811, df = 4843, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  1.613767 1.711336
## sample estimates:
## mean of the differences
##                1.662552
```

```
t.test(pod$malls,pod$beauty...spas, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod$malls and pod$beauty...spas
## t = 68.411, df = 4551, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.080559 2.203325
## sample estimates:
## mean of the differences
##                2.141942
```

```
t.test(pod$swimming.pools,pod$gyms, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod$swimming.pools and pod$gyms
## t = 1.375, df = 4361, p-value = 0.1692
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.007685361  0.043783481
## sample estimates:
## mean of the differences
##              0.01804906
```

```
t.test(pod$resorts,pod$hotels.other.lodgings, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod$resorts and pod$hotels.other.lodgings
## t = 9.369, df = 5356, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2158081 0.3300198
## sample estimates:
## mean of the differences
##               0.2729139
```

```
t.test(pod$parks,pod$gardens, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod$parks and pod$gardens
```

```
## t = 51.586, df = 5220, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.167043 1.259250
## sample estimates:
## mean of the differences
##                1.213147
```

```
#Kako su grupe religiozni i ostalo jednočlane obje će ući u jedan T-test.
t.test(pod$churches,pod$local.services, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod$churches and pod$local.services
## t = -42.651, df = 5250, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0539920 -0.9613574
## sample estimates:
## mean of the differences
##               -1.007675
```

```
#Još neki testovi
#t.test(pod$art.galleries,pod$malls, alternative = "two.sided",
#paired = TRUE, na.rm = TRUE)
#t.test(pod$museums,pod$theatres, alternative = "two.sided",
#paired = TRUE, na.rm = TRUE)
#t.test(pod$gyms,pod$beauty...spas, alternative = "two.sided",
#paired = TRUE, na.rm = TRUE)
#t.test(pod$dance.clubs,pod$parks, alternative = "two.sided",
#paired = TRUE, na.rm = TRUE)
#t.test(pod$churches,pod$gyms, alternative = "two.sided",
#paired = TRUE, na.rm = TRUE)
#t.test(pod$view.points,pod$museums, alternative = "two.sided",
#paired = TRUE, na.rm = TRUE)
#T-testovi --kraj
```

TESTOVI ZA GRUPIRANE KATEGORIJE

```
#meanKult = mean(pod.new$kultura, na.rm = TRUE)
#meanHrana = mean(pod.new$hrana, na.rm = TRUE)
#meanPice = mean(pod.new$pice, na.rm = TRUE)
#meanZabava = mean(pod.new$zabava, na.rm = TRUE)
#meanSport = mean(pod.new$sport, na.rm = TRUE)
#meanPriroda = mean(pod.new$priroda, na.rm = TRUE)
#meanUgost = mean(pod.new$ugostiteljski, na.rm = TRUE)
#meanRelig = mean(pod.new$religiozni, na.rm = TRUE)
#cat("Srednja vrijednost grupe kultura: " , meanKult, "\n")
#cat("Srednja vrijednost grupe hrana: " , meanHrana, "\n")
#cat("Srednja vrijednost grupe pice: " , meanPice, "\n")
#cat("Srednja vrijednost grupe zabava: " , meanZabava, "\n")
#cat("Srednja vrijednost grupe sport: " , meanSport, "\n")
#cat("Srednja vrijednost grupe priroda: " , meanPriroda, "\n")
#cat("Srednja vrijednost grupe ugostiteljski: " , meanUgost, "\n")
```

```r
#cat("Srednja vrijednost grupe religiozni: " , meanRelig, "\n")




# Provjera ima li značajne razlike kod najvećih srednjih
#vrijednosti i najmanjih među grupama
t.test(pod.new$sport,pod.new$religiozni, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod.new$sport and pod.new$religiozni
## t = -31.179, df = 4857, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5255121 -0.4633352
## sample estimates:
## mean of the differences
##              -0.4944236
```

```r
t.test(pod.new$kultura,pod.new$ugostiteljski, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod.new$kultura and pod.new$ugostiteljski
## t = 10.908, df = 5446, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1361601 0.1958245
## sample estimates:
## mean of the differences
##               0.1659923
```

```r
#   Zaključujemo da je grupa s najmanjom srednjom vrijednošću
#sport a grupa s najvećom je kultura

# T testovi za neke grupe slicnih srednjih vrijednosti
t.test(pod.new$hrana,pod.new$ugostiteljski, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod.new$hrana and pod.new$ugostiteljski
## t = -2.3182, df = 5446, p-value = 0.02047
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.064078866 -0.005358745
## sample estimates:
## mean of the differences
##             -0.03471881
```

```r
t.test(pod.new$pice,pod.new$zabava, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod.new$pice and pod.new$zabava
## t = -2.4072, df = 5446, p-value = 0.01611
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.053582491 -0.005481214
## sample estimates:
## mean of the differences
##             -0.02953185
```

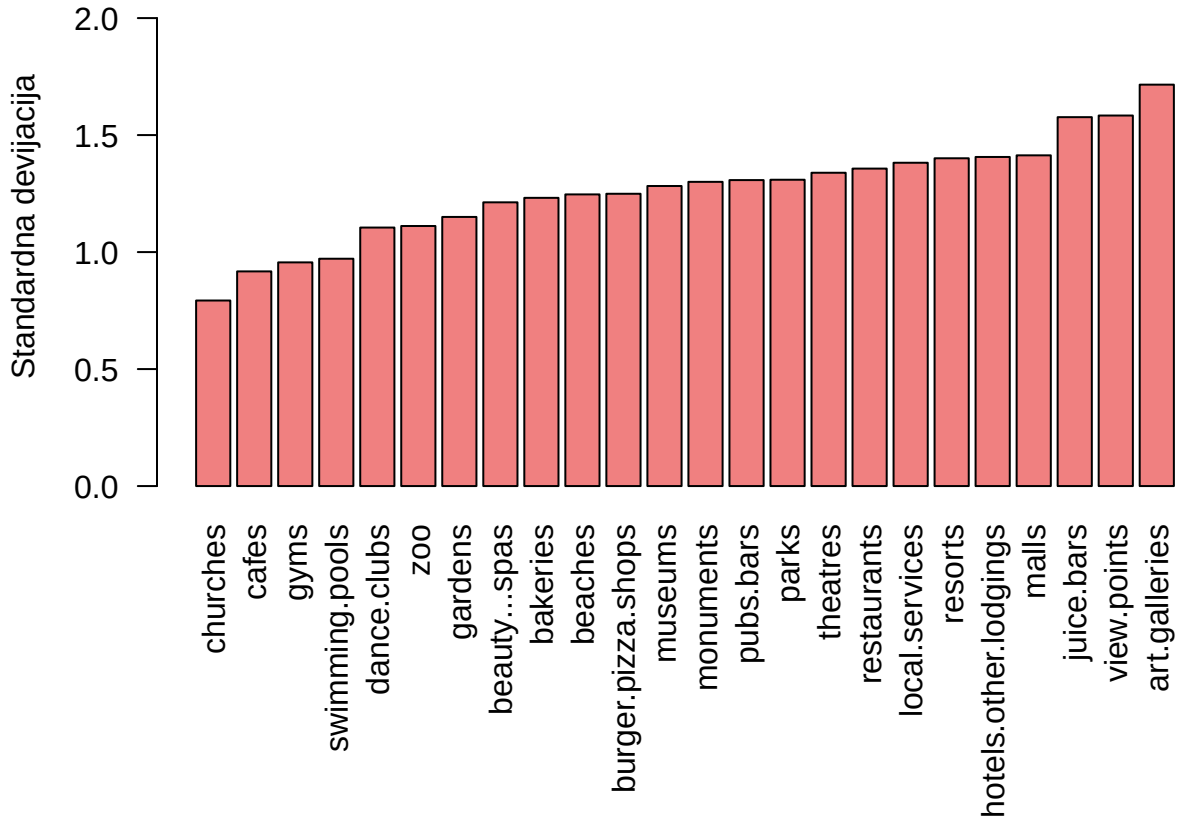
```r
t.test(pod.new$priroda,pod.new$hrana, alternative = "two.sided",
       paired = TRUE, na.rm = TRUE)
```

```
##
##  Paired t-test
##
## data:  pod.new$priroda and pod.new$hrana
## t = -1.5609, df = 5446, p-value = 0.1186
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.064075756  0.007269563
## sample estimates:
## mean of the differences
##              -0.0284031
```

```r
# Od ovih testova samo za priroda i hrana nemaju znacajne razlike
#u srednjim vrijednostima
```

Pitanje: Koja kategorije bi mogle biti najvise ”polarizirajuce”, a oko kojih se ljudi najvise slazu? Usporedite rasprsenja ocjena po odabranim kategorijama.

```r
sd_data = sort(sd_data)
par(mar = c(9, 5, 0, 0))
barplot(sd_data, ylab="Standardna devijacija", col="lightcoral",
        ylim=c(0,0.5 + max(sd_data)), las=2)
```




```
sd_data = sapply(pod,sd,na.rm=TRUE)
sd_mean = mean(sd_data)
cat("Srednja vrijednost standardnih devijacija svih kategorija: ",
    sd_mean, "\n")
```

```
## Srednja vrijednost standardnih devijacija svih kategorija:  1.263309
```

```
print("Polarizirajuce kategorije: ", quote=F)
```

```
## [1] Polarizirajuce kategorije:
```

```
sd_art_galleries = sd(pod$art.galleries, na.rm=TRUE)
cat("  Standardna devijacija kategorije \"art.galleries\"",
    sd_art_galleries, "\n")
```

```
##   Standardna devijacija kategorije "art.galleries" 1.715967
```

```
sd_view_points = sd(pod$view.points, na.rm=TRUE)
cat("  Standardna devijacija kategorije \"view.points\"",
    sd_view_points, "\n")
```

```
##   Standardna devijacija kategorije "view.points" 1.583232
```

```
print("Kategorije oko kojih se korisnici najvise slazu:",
      quote=F)
```

```
## [1] Kategorije oko kojih se korisnici najvise slazu:
```

```
sd_churches = sd(pod$churches, na.rm=TRUE)
sd_cafes = sd(pod$cafes, na.rm=TRUE)
cat("  Standardna devijacija kategorije \"churches\"",
```

```
    sd_churches, "\n")
```

```
##   Standardna devijacija kategorije "churches" 0.792771
```

```
cat("  Standardna devijacija kategorije \"cefes\"",
    sd_cafes, "\n")
```

```
##   Standardna devijacija kategorije "cefes" 0.9173273
```

```
# Radimo f-testove kako bi provjerili jeli stvarno postoji
# signifikantna razlika između kategorija
var.test(pod$churches, pod$art.galleries)
```

```
##
##  F test to compare two variances
##
## data:  pod$churches and pod$art.galleries
## F = 0.21344, num df = 5251, denom df = 5442, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2022995 0.2252019
## sample estimates:
## ratio of variances
##           0.213441
```

```
var.test(pod$view.points, pod$cafes)
```

```
##
##  F test to compare two variances
##
## data:  pod$view.points and pod$cafes
## F = 2.9788, num df = 5101, denom df = 4843, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.817588 3.149096
## sample estimates:
## ratio of variances
##           2.978794
```

```
# Radimo f-testove kako bi provjerili postoji li signifikantna
# razlika kod dvije najvise i najmanje polarizirajuce
# kategorije

var.test(pod$churches, pod$cafes)
```

```
##
##  F test to compare two variances
##
## data:  pod$churches and pod$cafes
## F = 0.74687, num df = 5251, denom df = 4843, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7067206 0.7892578
## sample estimates:
## ratio of variances
##          0.7468732
```

```r
var.test(pod$view.points, pod$art.galleries)
```

```
##
##  F test to compare two variances
##
## data:  pod$view.points and pod$art.galleries
## F = 0.85128, num df = 5101, denom df = 5442, p-value = 5.428e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8065282 0.8985537
## sample estimates:
## ratio of variances
##           0.851278
```

```r
# f-testovi za neke kategorije slicne varijance

var.test(pod$parks, pod$theatres)
```

```
##
##  F test to compare two variances
##
## data:  pod$parks and pod$theatres
## F = 0.95572, num df = 5446, denom df = 5446, p-value = 0.09472
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9062733 1.0078630
## sample estimates:
## ratio of variances
##          0.9557193
```

```r
var.test(pod$pubs.bars, pod$monuments)
```

```
##
##  F test to compare two variances
##
## data:  pod$pubs.bars and pod$monuments
## F = 1.0093, num df = 5446, denom df = 5144, p-value = 0.736
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9563403 1.0651904
## sample estimates:
## ratio of variances
##            1.00932
```

```r
var.test(pod$juice.bars, pod$view.points)
```

```
##
##  F test to compare two variances
##
## data:  pod$juice.bars and pod$view.points
## F = 0.99186, num df = 5446, denom df = 5101, p-value = 0.7666
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9396873 1.0468863
## sample estimates:
## ratio of variances
```

```
##          0.9918635
```

```
#var.test(pod$museums, pod$theatres)

#Koristimo bootsrap testove iz razloga što su podatci
# neparametarski
library(bootstrap)
print("Parks & Theatres")
```

```
## [1] "Parks & Theatres"
```

```
parksTheatres = bootstrapvariantpairedinterval(pod$parks,
                                    pod$theatres, 0.05, 1000)
parksTheatres$lb
```

```
##      2.5%
## 0.9227891
```

```
parksTheatres$ub
```

```
##     97.5%
## 0.9906637
```

```
print("Pubs.bars & Monuments")
```

```
## [1] "Pubs.bars & Monuments"
```

```
pubsMonuments = bootstrapvariantpairedinterval(pod$pubs.bars,
                                    pod$monuments , 0.05, 1000)
pubsMonuments$lb
```

```
##      2.5%
## 0.9528404
```

```
pubsMonuments$ub
```

```
##   97.5%
## 1.06926
```

```
print("Juice.bars & View.Points")
```

```
## [1] "Juice.bars & View.Points"
```

```
juiceView = bootstrapvariantpairedinterval(pod$juice.bars,
                                   pod$view.points, 0.05, 1000)
juiceView$lb
```

```
##      2.5%
## 0.9466142
```

```
juiceView$ub
```

```
##    97.5%
## 1.041033
```

```
print("Museums & Theatres")
```

```
## [1] "Museums & Theatres"
```

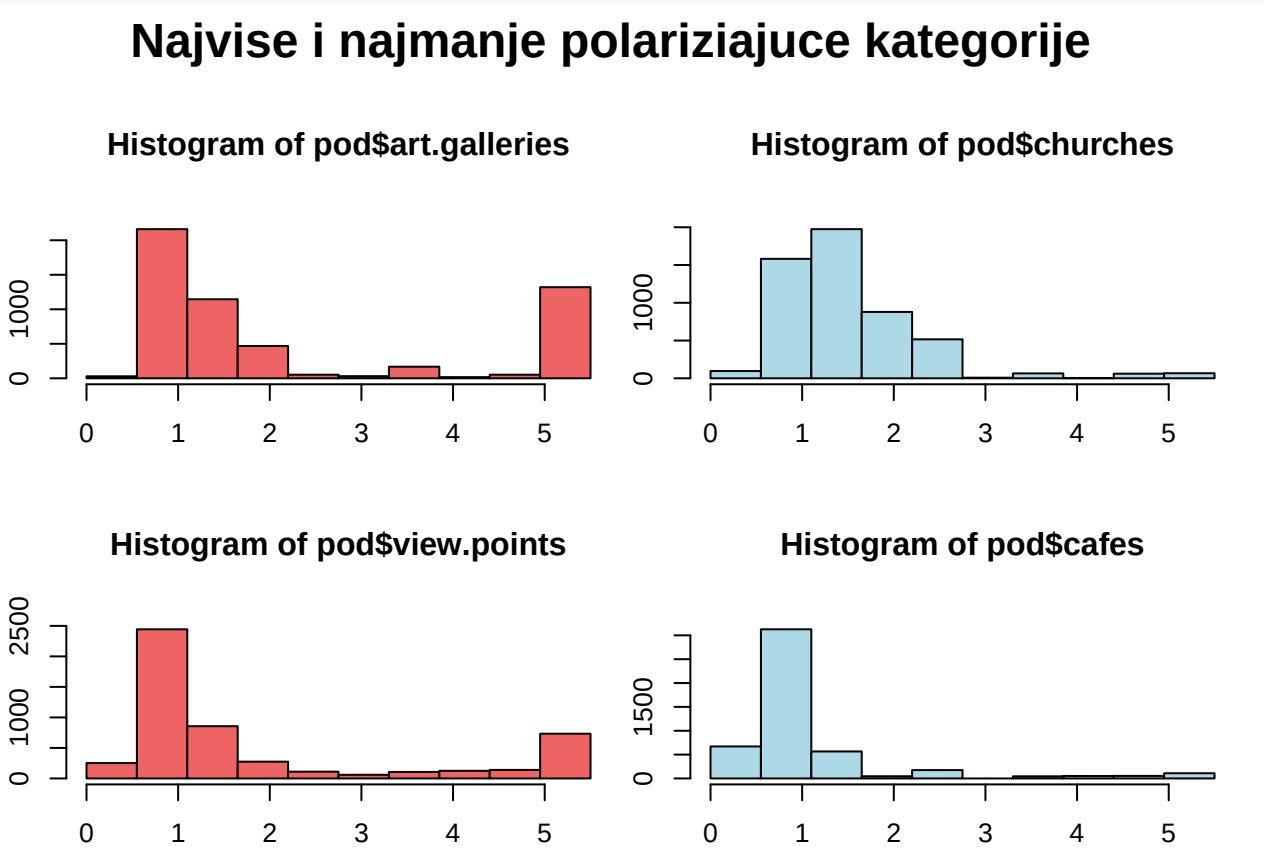
```
museumsTheatres = bootstrapvariantpairedinterval(pod$museums,
                                    pod$theatres, 0.05, 1000)
museumsTheatres$lb
```

```
##      2.5%
## 0.8865156
```

```
museumsTheatres$ub
```

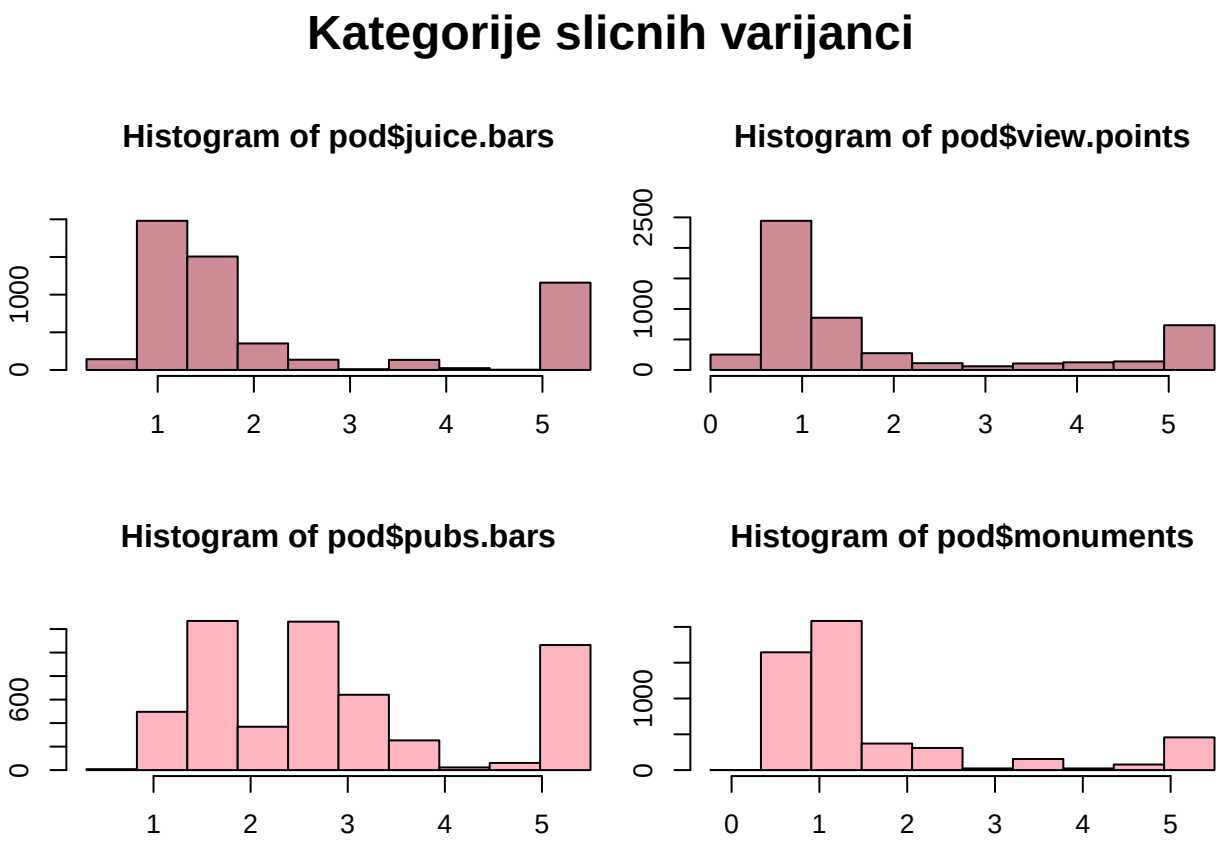
```
##     97.5%
## 0.9460874
```

```
# Prikazat cemo neke kategorije pomocu histograma

#   Prvo pokazujemo kategorije velikih razlika
par(mfrow=c(2,2), mar=c(2.5,2.5,5,0), oma = c(0, 0, 2, 0))
hist(pod$art.galleries,col="indianred2",
  breaks = seq(from = min(pod$art.galleries ,na.rm = TRUE) - 0.5,
  to = max(pod$art.galleries, na.rm=TRUE) + 0.5,  length = 11))
hist(pod$churches,col="lightblue",
  breaks = seq(from = min(pod$churches ,na.rm = TRUE) - 0.5,
  to = max(pod$churches, na.rm=TRUE) + 0.5, length = 11))
hist(pod$view.points,col="indianred2",
  breaks = seq(from = min(pod$view.points ,na.rm = TRUE) - 0.5,
  to = max(pod$view.points, na.rm=TRUE) + 0.5, length = 11))
hist(pod$cafes,col="lightblue",
  breaks = seq(from = min(pod$cafes ,na.rm = TRUE) - 0.5,
  to = max(pod$cafes, na.rm=TRUE) + 0.5, length = 11))
mtext("Najvise i najmanje polariziajuce kategorije",
      outer = TRUE,cex=1.5,font=2)
```



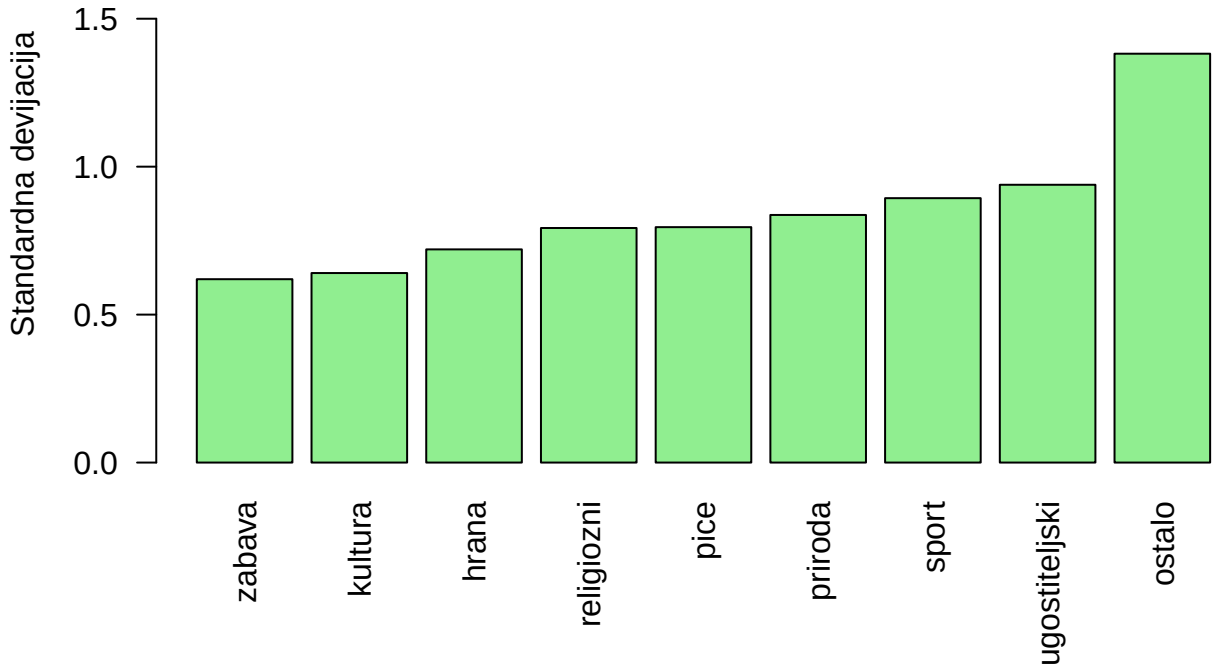

```
#   Zatim pokazujemo kategorije slicnih varijanci
par(mfrow=c(2,2), mar=c(2.5,2.5,5,0), oma = c(0, 0, 2, 0))
```

```r
hist(pod$juice.bars, col="lightpink3",
  breaks = seq(from = min(pod$juice.bars ,na.rm = TRUE) - 0.5,
  to = max(pod$juice.bars, na.rm=TRUE) + 0.5, length = 11))
hist(pod$view.points,  col="lightpink3",
  breaks = seq(from = min(pod$view.points ,na.rm = TRUE) - 0.5,
  to = max(pod$view.points, na.rm=TRUE) + 0.5, length = 11))
hist(pod$pubs.bars,  col="lightpink",
  breaks = seq(from = min(pod$pubs.bars ,na.rm = TRUE) - 0.5,
  to = max(pod$pubs.bars, na.rm=TRUE) + 0.5, length = 11))
hist(pod$monuments,  col="lightpink",
  breaks = seq(from = min(pod$monuments ,na.rm = TRUE) - 0.5,
  to = max(pod$monuments, na.rm=TRUE) + 0.5, length = 11))
mtext("Kategorije slicnih varijanci",outer = TRUE,cex=1.5,font=2)
```




```r
######## GRUPE ###########3


#varKult = var(pod.new$kultura, na.rm = TRUE)
#varHrana = var(pod.new$hrana, na.rm = TRUE)
#varPice = var(pod.new$pice, na.rm = TRUE)
#varZabava = var(pod.new$zabava, na.rm = TRUE)
#varSport = var(pod.new$sport, na.rm = TRUE)
#varPriroda = var(pod.new$priroda, na.rm = TRUE)
#varUgost = var(pod.new$ugostiteljski, na.rm = TRUE)
#varRelig = var(pod.new$religiozni, na.rm = TRUE)
#cat("Varijanca grupe kultura: " , varKult, "\n")
#cat("Varijanca grupe hrana: " , varHrana, "\n")
```

```r
#cat("Varijanca grupe pice: " , varPice, "\n")
#cat("Varijanca grupe zabava: " , varZabava, "\n")
#cat("Varijanca grupe sport: " , varSport, "\n")
#cat("Varijanca grupe priroda: " , varPriroda, "\n")
#cat("Varijanca grupe ugostiteljski: " , varUgost, "\n")
#cat("Varijanca grupe religiozni: " , varRelig, "\n")

variance_group_data = sapply(pod.new,var,na.rm = T)
sd_group_data = sqrt(variance_group_data)
sd_group_data = sort(sd_group_data)

par(mfrow=c(1,1),mar = c(6, 4, 0, 0))
barplot(sd_group_data, ylab="Standardna devijacija",
    ylim=c(0,0.5 + max(sd_group_data)) , col="lightgreen", las=2)
```




```r
#   Testiramo koje su grupe najvise polarizirajuce a koje najmanje
var.test(pod.new$zabava, pod.new$kultura)
```

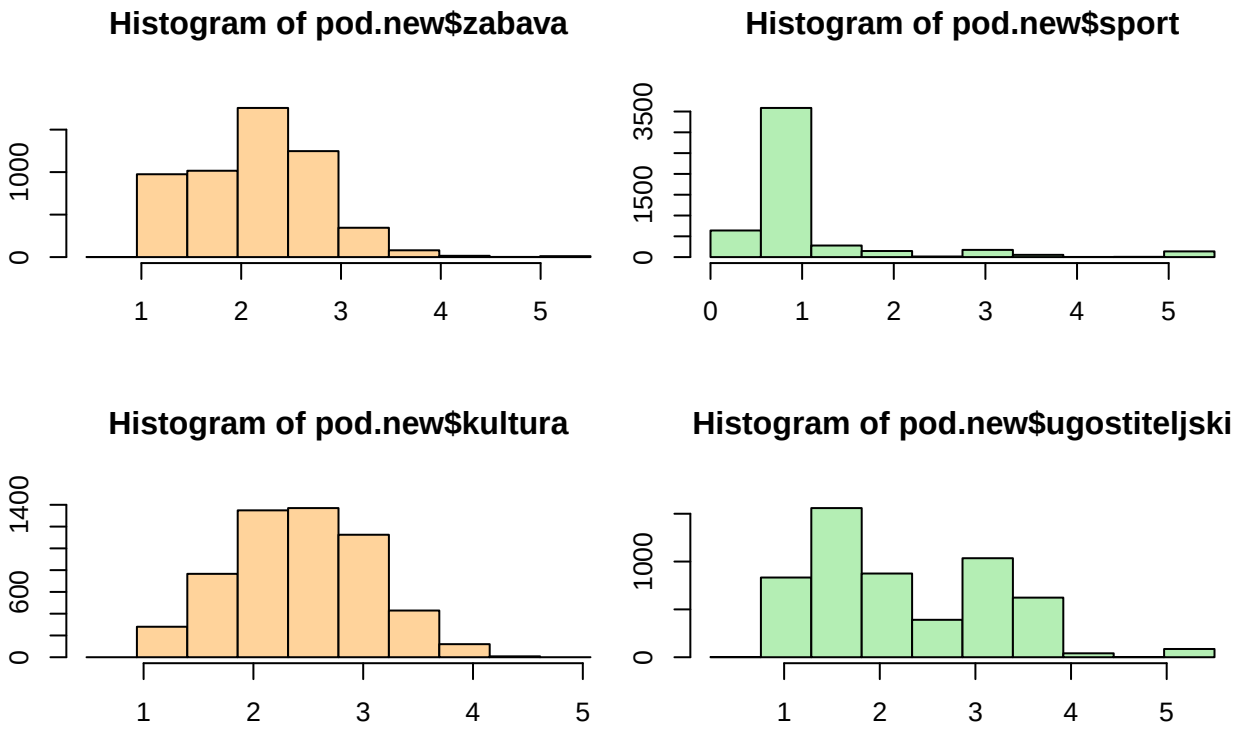
```
##
##  F test to compare two variances
##
## data:  pod.new$zabava and pod.new$kultura
## F = 0.93614, num df = 5446, denom df = 5446, p-value = 0.01491
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8877088 0.9872174
## sample estimates:
## ratio of variances
##          0.9361419
```

```r
var.test(pod.new$sport, pod.new$ugostiteljski)
```

```
##
```

```
##  F test to compare two variances
##
## data:  pod.new$sport and pod.new$ugostiteljski
## F = 0.90599, num df = 5036, denom df = 5446, p-value = 0.0003604
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8582191 0.9564680
## sample estimates:
## ratio of variances
##          0.9059864
```

```
par(mfrow=c(2,2),mar=c(2.5,2.5,5,0), oma = c(0, 0, 2, 0))
hist(pod.new$zabava,  col="burlywood1",
    breaks = seq(from = min(pod.new$zabava ,na.rm = TRUE) - 0.5,
    to = max(pod.new$zabava, na.rm=TRUE) + 0.5, length = 11))
hist(pod.new$sport,col="darkseagreen2",
    breaks = seq(from = min(pod.new$sport ,na.rm = TRUE) - 0.5,
    to = max(pod.new$sport, na.rm=TRUE) + 0.5, length = 11))
hist(pod.new$kultura, col="burlywood1",
    breaks = seq(from = min(pod.new$kultura ,na.rm = TRUE) - 0.5,
    to = max(pod.new$kultura, na.rm=TRUE) + 0.5, length = 11))
hist(pod.new$ugostiteljski, col="darkseagreen2",
    breaks = seq(from = min(pod.new$ugostiteljski ,na.rm = TRUE) - 0.5,
    to = max(pod.new$ugostiteljski, na.rm=TRUE) + 0.5, length = 11))
```




```
# Provjera var testom za grupe slicnih varijanci

var.test(pod.new$religiozni, pod.new$pice)
```

```
##
##  F test to compare two variances
##
```

```
## data:  pod.new$religiozni and pod.new$pice
## F = 0.99294, num df = 5251, denom df = 5446, p-value = 0.7958
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9411201 1.0476442
## sample estimates:
## ratio of variances
##          0.9929415
```

```
var.test(pod.new$priroda, pod.new$sport)
```

```
##
##  F test to compare two variances
##
## data:  pod.new$priroda and pod.new$sport
## F = 0.87702, num df = 5446, denom df = 5036, p-value = 2.045e-06
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8307298 0.9258318
## sample estimates:
## ratio of variances
##          0.8770182
```

```
#Računanje korelacijske matrice. Funkcija cor() računa
#korelacijski koefcijent, a funkcija rcorr() osim korelacijskog
#koeficijenta vraća i p-vrijednost korelacije. Na prvi graf idu
#rezultati iz prve korelacijske matrice, a na drugi idu rezultati druge,
#s tim da se u drugom grafu ne prikazuju vrijednosti čija je
#p-vrijednost ispod 0.01.

#Corrplot je jako zgodan za prikaz korelacija, može se dosta toga
#modificirati u ispisu. Ostavio sam obje metode s nekim različitim
#parametrima da pogledate. Order hclust je po meni najprikladniji
#jer se time postiže hijerarhijski prikaz vrijednosti, probajte
#umjesto hclust staviti alphabet, dosta je ružnije.


require(ggpubr)
```

```
## Loading required package: ggpubr
```

```
## Loading required package: ggplot2
```

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
require(Hmisc)
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
require(corrplot)
```

```
## Loading required package: corrplot
```

```
## corrplot 0.84 loaded
```

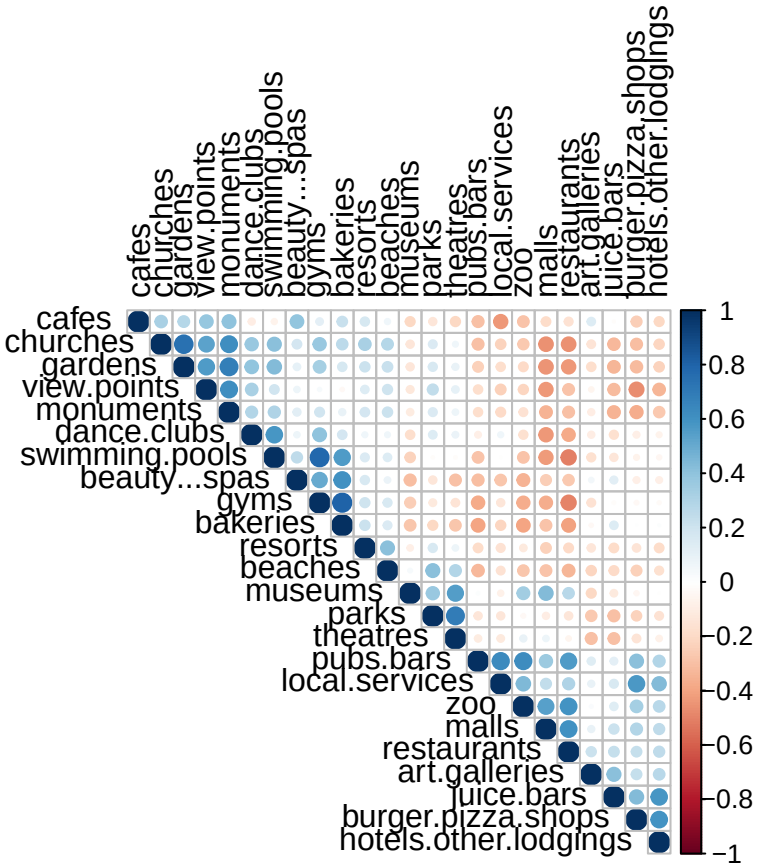
```
res_cor <- cor(pod,use="complete.obs",method="spearman")

res_rcorr <- rcorr(as.matrix(pod),type = c("spearman"))


round(res_cor,2)
```

```
##                       churches resorts beaches parks theatres museums malls
## churches                  1.00    0.33    0.29  0.15     0.06   -0.13 -0.45
## resorts                   0.33    1.00    0.41  0.17     0.08   -0.08 -0.24
## beaches                   0.29    0.41    1.00  0.41     0.30    0.04 -0.28
## parks                     0.15    0.17    0.41  1.00     0.70    0.37 -0.10
## theatres                  0.06    0.08    0.30  0.70     1.00    0.56  0.08
## museums                  -0.13   -0.08    0.04  0.37     0.56    1.00  0.43
## malls                    -0.45   -0.24   -0.28 -0.10     0.08    0.43  1.00
## zoo                      -0.26   -0.11   -0.27 -0.04     0.10    0.35  0.54
## restaurants              -0.45   -0.19   -0.34 -0.14    -0.05    0.27  0.60
## pubs.bars                -0.29   -0.19   -0.32 -0.14    -0.11    0.02  0.37
## local.services           -0.22   -0.16   -0.15 -0.13    -0.11   -0.07  0.25
## burger.pizza.shops       -0.30   -0.14   -0.24 -0.22    -0.15   -0.05  0.30
## hotels.other.lodgings    -0.21   -0.19   -0.16 -0.15    -0.08    0.00  0.25
## juice.bars               -0.32   -0.19   -0.20 -0.30    -0.29   -0.11  0.22
## art.galleries            -0.15   -0.14   -0.22 -0.27    -0.30   -0.20  0.10
## dance.clubs               0.37    0.06    0.06  0.15     0.08   -0.18 -0.44
## swimming.pools            0.42    0.16    0.14 -0.02    -0.02   -0.23 -0.43
## gyms                      0.37    0.21    0.16 -0.12    -0.14   -0.25 -0.36
## bakeries                  0.27    0.22    0.15 -0.20    -0.28   -0.28 -0.28
## beauty...spas             0.19    0.17    0.08 -0.13    -0.29   -0.30 -0.25
## cafes                     0.32    0.17    0.07 -0.13    -0.21   -0.20 -0.19
## view.points               0.53    0.14    0.21  0.25     0.11   -0.12 -0.43
## monuments                 0.61    0.18    0.22  0.16     0.08   -0.11 -0.34
## gardens                   0.74    0.22    0.23  0.17     0.10   -0.13 -0.45
```

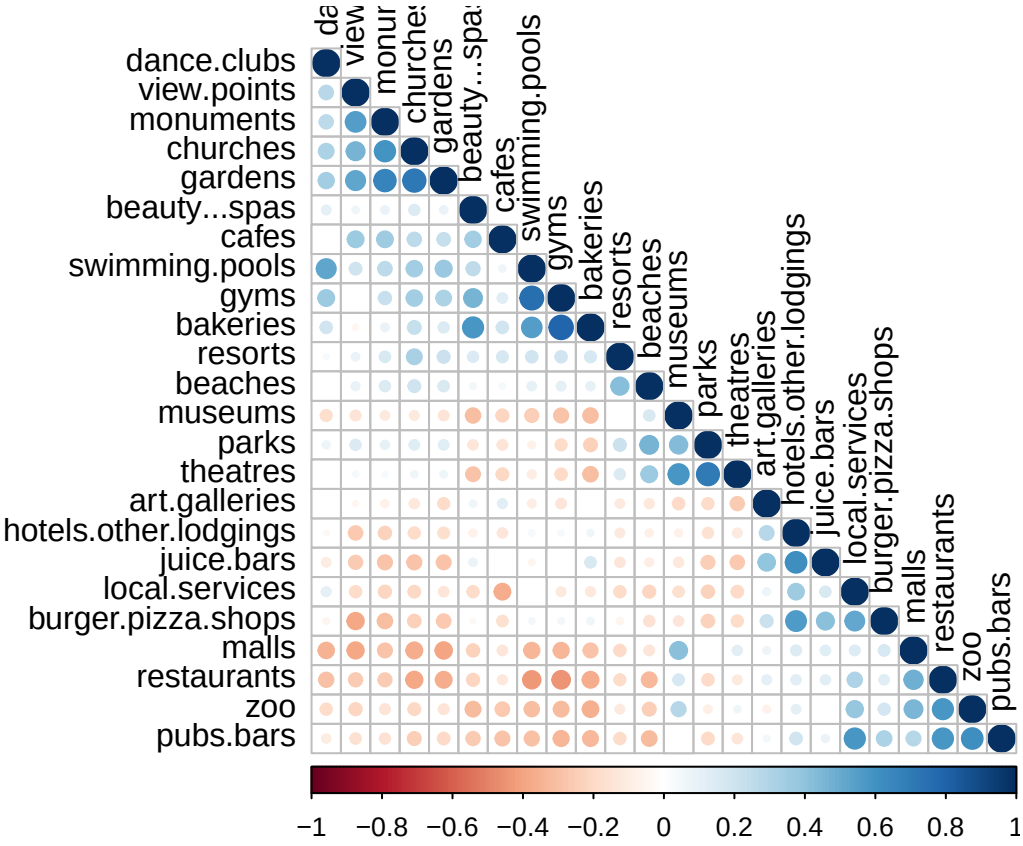
```
##                           zoo restaurants pubs.bars local.services
## churches              -0.26       -0.45     -0.29          -0.22
## resorts               -0.11       -0.19     -0.19          -0.16
## beaches               -0.27       -0.34     -0.32          -0.15
## parks                 -0.04       -0.14     -0.14          -0.13
## theatres               0.10       -0.05     -0.11          -0.11
## museums                0.35        0.27      0.02          -0.07
## malls                  0.54        0.60      0.37           0.25
## zoo                    1.00        0.60      0.63           0.45
## restaurants            0.60        1.00      0.56           0.31
## pubs.bars              0.63        0.56      1.00           0.64
## local.services         0.45        0.31      0.64           1.00
## burger.pizza.shops     0.33        0.23      0.42           0.58
## hotels.other.lodgings  0.28        0.25      0.30           0.43
## juice.bars             0.14        0.23      0.12           0.16
## art.galleries          0.02        0.21      0.13           0.10
## dance.clubs           -0.17       -0.37     -0.09           0.07
## swimming.pools        -0.28       -0.50     -0.28           0.00
## gyms                  -0.37       -0.50     -0.37          -0.13
## bakeries              -0.39       -0.41     -0.40          -0.21
## beauty...spas         -0.35       -0.26     -0.30          -0.28
## cafes                 -0.28       -0.16     -0.30          -0.43
## view.points           -0.21       -0.29     -0.17          -0.24
## monuments             -0.16       -0.29     -0.17          -0.20
## gardens               -0.19       -0.44     -0.24          -0.17
##                       burger.pizza.shops hotels.other.lodgings juice.bars
## churches                           -0.30                -0.21      -0.32
## resorts                            -0.14                -0.19      -0.19
## beaches                            -0.24                -0.16      -0.20
## parks                              -0.22                -0.15      -0.30
## theatres                           -0.15                -0.08      -0.29
## museums                            -0.05                 0.00      -0.11
## malls                               0.30                 0.25       0.22
## zoo                                 0.33                 0.28       0.14
## restaurants                         0.23                 0.25       0.23
## pubs.bars                           0.42                 0.30       0.12
## local.services                      0.58                 0.43       0.16
## burger.pizza.shops                  1.00                 0.60       0.43
## hotels.other.lodgings               0.60                 1.00       0.59
## juice.bars                          0.43                 0.59       1.00
## art.galleries                       0.23                 0.28       0.42
## dance.clubs                        -0.09                -0.06      -0.17
## swimming.pools                     -0.06                -0.05      -0.13
## gyms                               -0.04                -0.03      -0.02
## bakeries                            0.00                 0.01       0.14
## beauty...spas                      -0.09                -0.09       0.12
## cafes                              -0.25                -0.20      -0.01
## view.points                        -0.47                -0.33      -0.31
## monuments                          -0.36                -0.27      -0.35
## gardens                            -0.32                -0.22      -0.34
##                       art.galleries dance.clubs swimming.pools  gyms bakeries
## churches                      -0.15        0.37           0.42  0.37     0.27
## resorts                       -0.14        0.06           0.16  0.21     0.22
## beaches                       -0.22        0.06           0.14  0.16     0.15
```

```
## parks                          -0.27        0.15          -0.02 -0.12    -0.20
## theatres                       -0.30        0.08          -0.02 -0.14    -0.28
## museums                        -0.20       -0.18          -0.23 -0.25    -0.28
## malls                           0.10       -0.44          -0.43 -0.36    -0.28
## zoo                             0.02       -0.17          -0.28 -0.37    -0.39
## restaurants                     0.21       -0.37          -0.50 -0.50    -0.41
## pubs.bars                       0.13       -0.09          -0.28 -0.37    -0.40
## local.services                  0.10        0.07           0.00 -0.13    -0.21
## burger.pizza.shops              0.23       -0.09          -0.06 -0.04     0.00
## hotels.other.lodgings           0.28       -0.06          -0.05 -0.03     0.01
## juice.bars                      0.42       -0.17          -0.13 -0.02     0.14
## art.galleries                   1.00       -0.10          -0.16 -0.15    -0.03
## dance.clubs                    -0.10        1.00           0.59  0.41     0.18
## swimming.pools                 -0.16        0.59           1.00  0.78     0.57
## gyms                           -0.15        0.41           0.78  1.00     0.80
## bakeries                       -0.03        0.18           0.57  0.80     1.00
## beauty...spas                   0.05        0.07           0.26  0.50     0.60
## cafes                           0.14       -0.10          -0.07  0.11     0.23
## view.points                    -0.05        0.33           0.19  0.03    -0.03
## monuments                      -0.10        0.30           0.31  0.20     0.10
## gardens                        -0.18        0.39           0.44  0.34     0.18
##                          beauty...spas cafes view.points monuments gardens
## churches                          0.19  0.32        0.53      0.61    0.74
## resorts                           0.17  0.17        0.14      0.18    0.22
## beaches                           0.08  0.07        0.21      0.22    0.23
## parks                            -0.13 -0.13        0.25      0.16    0.17
## theatres                         -0.29 -0.21        0.11      0.08    0.10
## museums                          -0.30 -0.20       -0.12     -0.11   -0.13
## malls                            -0.25 -0.19       -0.43     -0.34   -0.45
## zoo                              -0.35 -0.28       -0.21     -0.16   -0.19
## restaurants                      -0.26 -0.16       -0.29     -0.29   -0.44
## pubs.bars                        -0.30 -0.30       -0.17     -0.17   -0.24
## local.services                   -0.28 -0.43       -0.24     -0.20   -0.17
## burger.pizza.shops               -0.09 -0.25       -0.47     -0.36   -0.32
## hotels.other.lodgings            -0.09 -0.20       -0.33     -0.27   -0.22
## juice.bars                        0.12 -0.01       -0.31     -0.35   -0.34
## art.galleries                     0.05  0.14       -0.05     -0.10   -0.18
## dance.clubs                       0.07 -0.10        0.33      0.30    0.39
## swimming.pools                    0.26 -0.07        0.19      0.31    0.44
## gyms                              0.50  0.11        0.03      0.20    0.34
## bakeries                          0.60  0.23       -0.03      0.10    0.18
## beauty...spas                     1.00  0.39        0.07      0.13    0.11
## cafes                             0.39  1.00        0.39      0.40    0.29
## view.points                       0.07  0.39        1.00      0.61    0.57
## monuments                         0.13  0.40        0.61      1.00    0.69
## gardens                           0.11  0.29        0.57      0.69    1.00
```

```
######    SVUGDJE U CORRPLOT MOZE SE DODATI method=""
#a moguce vrijednosti su  "circle", "square", "ellipse",
#"number", "shade", "color", "pie"  #######

corrplot(res_cor, type = "upper",method="circle", order = "hclust",
         tl.col = "black")
```




```
corrplot(res_rcorr$r, type="lower",method= "circle", order="hclust",
         tl.col = "black",
         p.mat = res_rcorr$P, sig.level = 0.01, insig = "blank")
```

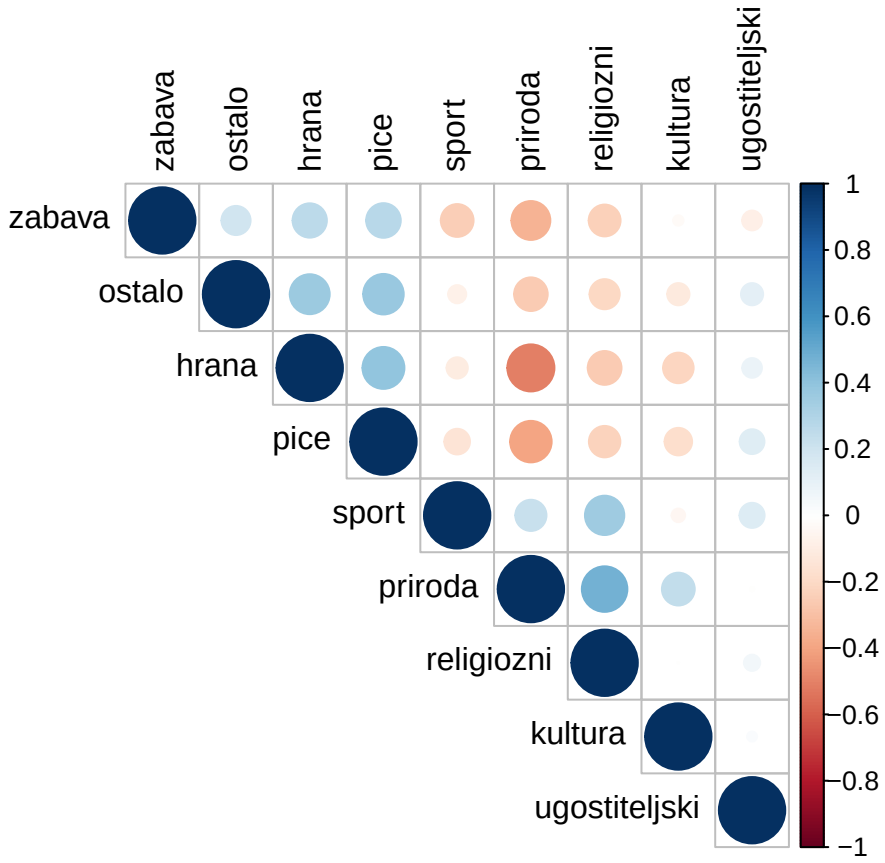



```
#metodom cor.test() možemo ispitati korelaciju između para
#uzoraka ako nam zatreba

#korelacije za markov dio

res_cor <- cor(pod.new,use="complete.obs",method="spearman")

res_rcorr <- rcorr(as.matrix(pod.new), type = c("spearman"))

round(res_cor,2)
```

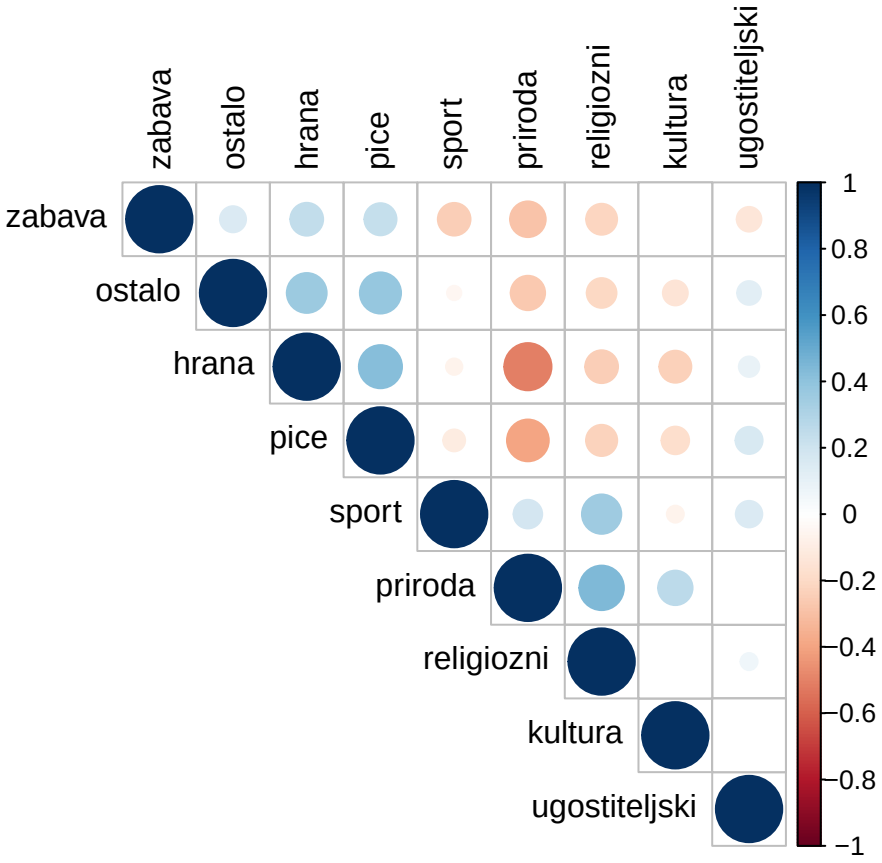
```
##               kultura hrana  pice zabava sport priroda ugostiteljski religiozni
## kultura          1.00 -0.21 -0.17  -0.02 -0.04    0.24          0.02       0.00
## hrana           -0.21  1.00  0.40   0.27 -0.10   -0.50          0.09      -0.26
## pice            -0.17  0.40  1.00   0.27 -0.15   -0.39          0.14      -0.22
## zabava          -0.02  0.27  0.27   1.00 -0.24   -0.34         -0.09      -0.23
## sport           -0.04 -0.10 -0.15  -0.24  1.00    0.22          0.14       0.35
## priroda          0.24 -0.50 -0.39  -0.34  0.22    1.00          0.00       0.47
## ugostiteljski    0.02  0.09  0.14  -0.09  0.14    0.00          1.00       0.06
## religiozni       0.00 -0.26 -0.22  -0.23  0.35    0.47          0.06       1.00
## ostalo          -0.11  0.36  0.37   0.19 -0.07   -0.26          0.11      -0.21
##               ostalo
## kultura        -0.11
## hrana           0.36
## pice            0.37
## zabava          0.19
## sport          -0.07
```

```
## priroda        -0.26
## ugostiteljski   0.11
## religiozni     -0.21
## ostalo          1.00
```

```
corrplot(res_cor, type = "upper" ,method="circle", order = "hclust",
         tl.col = "black")
```




```
corrplot(res_rcorr$r, type = "upper",method="circle", order="hclust",
tl.col = "black", p.mat = res_rcorr$P, sig.level = 0.01, insig = "blank")
```




```
#Linearna regresija --početak

#Stvorene su nove tablice gdje je uklonjen samo stupac "users".
pod.cor = pod
pod.new.cor = pod.new

#Prvo je fokus na običnim podacima, tj. podacima koji nisu grupirani.
#Iz korelacijskog dijagrama vidi se kako je kategorija restaurants korelirana s možda
#najviše drugih kategorija.

fit.restaurants = lm(restaurants ~ ., data = pod.cor)
summary(fit.restaurants)
```

```
##
## Call:
## lm(formula = restaurants ~ ., data = pod.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1523 -0.5695 -0.1802  0.4330  3.8471
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.26894    0.14344  15.818  < 2e-16 ***
## churches              -0.11464    0.02237  -5.124 3.14e-07 ***
## resorts                0.03606    0.01267   2.845 0.004466 **
## beaches               -0.06007    0.01577  -3.809 0.000142 ***
```

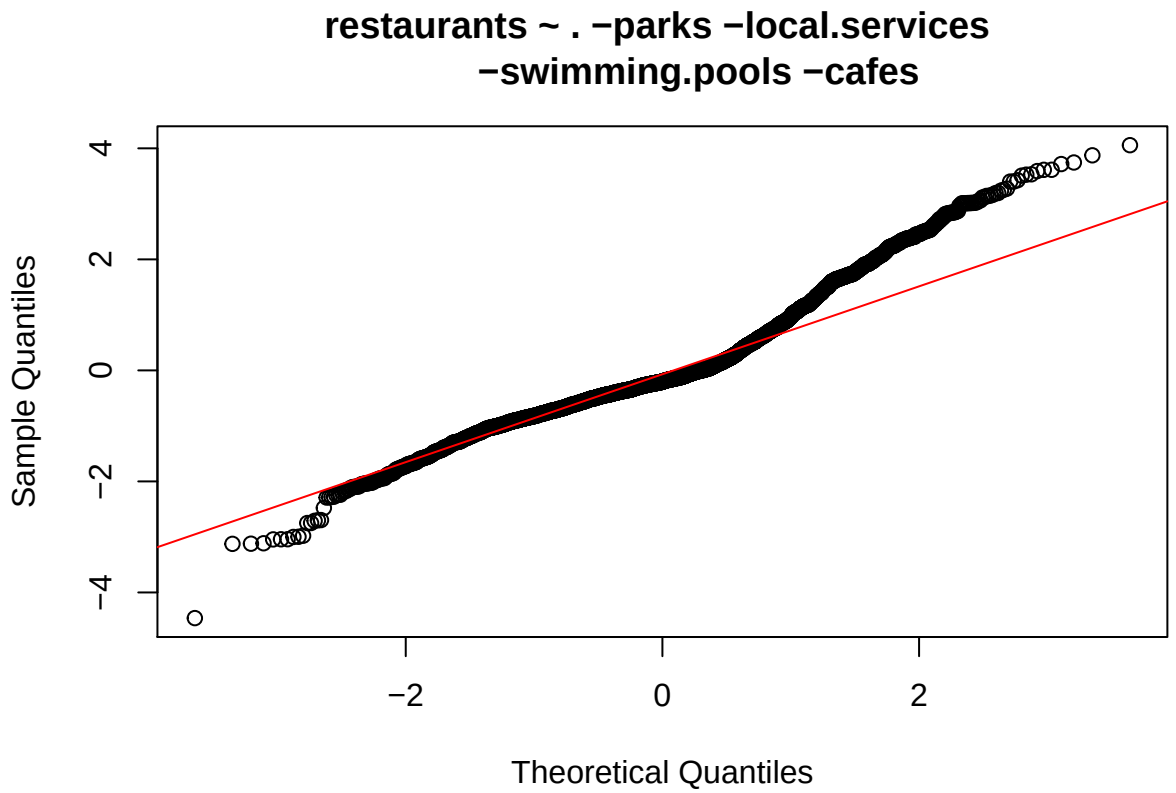
```
## parks                   0.01199    0.01641   0.730 0.465171
## theatres               -0.14883    0.01699  -8.761  < 2e-16 ***
## museums                 0.07717    0.01650   4.678 3.00e-06 ***
## malls                   0.18071    0.01540  11.735  < 2e-16 ***
## zoo                     0.26113    0.01956  13.348  < 2e-16 ***
## pubs.bars               0.30920    0.01693  18.265  < 2e-16 ***
## local.services          0.01412    0.01520   0.929 0.352781
## burger.pizza.shops     -0.16802    0.01588 -10.583  < 2e-16 ***
## hotels.other.lodgings   0.04511    0.01354   3.333 0.000868 ***
## juice.bars             -0.06583    0.01248  -5.275 1.41e-07 ***
## art.galleries           0.05341    0.01010   5.288 1.31e-07 ***
## dance.clubs            -0.05997    0.01581  -3.794 0.000151 ***
## swimming.pools         -0.02836    0.02316  -1.225 0.220798
## gyms                   -0.10538    0.02259  -4.664 3.21e-06 ***
## bakeries               -0.06668    0.01527  -4.367 1.29e-05 ***
## beauty...spas           0.03780    0.01352   2.795 0.005213 **
## cafes                   0.02057    0.01998   1.029 0.303472
## view.points            -0.03011    0.01238  -2.432 0.015082 *
## monuments              -0.05615    0.01325  -4.237 2.32e-05 ***
## gardens                -0.14609    0.01551  -9.416  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9413 on 3692 degrees of freedom
##   (1732 observations deleted due to missingness)
## Multiple R-squared:  0.5669, Adjusted R-squared:  0.5642
## F-statistic: 210.1 on 23 and 3692 DF,  p-value: < 2.2e-16
```

```
#Vidi se da je vrijednost Adjusted R-squared u redu, ali vidimo kako neke kategorije
#nisu signifikantne za linearnu regresiju,
#tj. kada bi faktor uz njih bio 0, linearna regresija se nebi signifikantno
#razlikovala od trenutne.

fit.restaurants.edit = lm(restaurants ~ . -parks -local.services -swimming.pools -cafes,
                          data = pod.cor)
summary(fit.restaurants.edit)
```

```
##
## Call:
## lm(formula = restaurants ~ . - parks - local.services - swimming.pools -
##     cafes, data = pod.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1866 -0.5663 -0.1912  0.4360  3.8008
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.31326    0.13359  17.317  < 2e-16 ***
## churches              -0.11539    0.02233  -5.169 2.48e-07 ***
## resorts                0.03598    0.01257   2.863 0.004222 **
## beaches               -0.05829    0.01525  -3.822 0.000134 ***
## theatres              -0.14525    0.01502  -9.673  < 2e-16 ***
## museums                0.07576    0.01626   4.660 3.28e-06 ***
## malls                  0.18060    0.01515  11.920  < 2e-16 ***
```

```
## zoo                     0.25922    0.01929  13.439  < 2e-16 ***
## pubs.bars               0.31655    0.01611  19.646  < 2e-16 ***
## burger.pizza.shops     -0.16821    0.01547 -10.874  < 2e-16 ***
## hotels.other.lodgings   0.04703    0.01310   3.590 0.000335 ***
## juice.bars             -0.06764    0.01235  -5.478 4.58e-08 ***
## art.galleries           0.05330    0.00999   5.336 1.01e-07 ***
## dance.clubs            -0.05972    0.01560  -3.829 0.000131 ***
## gyms                   -0.11609    0.02092  -5.548 3.10e-08 ***
## bakeries               -0.06946    0.01518  -4.577 4.87e-06 ***
## beauty...spas           0.03750    0.01333   2.814 0.004922 **
## view.points            -0.02585    0.01197  -2.160 0.030844 *
## monuments              -0.05504    0.01309  -4.204 2.69e-05 ***
## gardens                -0.14695    0.01537  -9.560  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9412 on 3696 degrees of freedom
##   (1732 observations deleted due to missingness)
## Multiple R-squared:  0.5665, Adjusted R-squared:  0.5643
## F-statistic: 254.2 on 19 and 3696 DF,  p-value: < 2.2e-16
```

```
#Tu vidimo kako uklanjanjem nesignifikatnih atributa iz jednadžbe linearne regresije u
#ovom slučaju ne dobivamo znatno različitu vrijednost Adjusted R-squared

qqnorm(rstandard(fit.restaurants.edit),  main = "restaurants ~ . -parks -local.services
       -swimming.pools -cafes")
qqline(rstandard(fit.restaurants.edit), col = "red")
```




```
#Danas su teretane jako popularne, pa stoga pokušajmo vidjeti može li se iz ostalih
#kategorija predvidjeti ocjena teretane za usera
```

```r
fit.gyms = lm(gyms ~ ., data = pod.cor)
summary(fit.gyms)
```

```
##
## Call:
## lm(formula = gyms ~ ., data = pod.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7299 -0.2137 -0.0370  0.0825  3.9927
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.842312   0.106754   7.890 3.94e-15 ***
## churches                 0.006760   0.016306   0.415 0.678489
## resorts                 -0.005054   0.009214  -0.549 0.583338
## beaches                 -0.026176   0.011469  -2.282 0.022526 *
## parks                   -0.039198   0.011900  -3.294 0.000997 ***
## theatres                -0.042649   0.012445  -3.427 0.000617 ***
## museums                 -0.015891   0.012012  -1.323 0.185961
## malls                    0.010652   0.011389   0.935 0.349719
## zoo                     -0.005257   0.014547  -0.361 0.717841
## restaurants             -0.055581   0.011918  -4.664 3.21e-06 ***
## pubs.bars               -0.029057   0.012829  -2.265 0.023576 *
## local.services          -0.001533   0.011038  -0.139 0.889524
## burger.pizza.shops      -0.006021   0.011704  -0.514 0.606958
## hotels.other.lodgings    0.027681   0.009834   2.815 0.004908 **
## juice.bars              -0.002665   0.009099  -0.293 0.769639
## art.galleries           -0.026840   0.007350  -3.652 0.000264 ***
## dance.clubs              0.072014   0.011442   6.294 3.45e-10 ***
## swimming.pools           0.374570   0.015650  23.934  < 2e-16 ***
## bakeries                 0.168031   0.010768  15.604  < 2e-16 ***
## beauty...spas            0.007755   0.009830   0.789 0.430205
## cafes                    0.029043   0.014508   2.002 0.045383 *
## view.points             -0.024644   0.008991  -2.741 0.006157 **
## monuments                0.005385   0.009647   0.558 0.576729
## gardens                  0.036488   0.011386   3.205 0.001364 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6836 on 3692 degrees of freedom
##   (1732 observations deleted due to missingness)
## Multiple R-squared:  0.3841, Adjusted R-squared:  0.3802
## F-statistic: 100.1 on 23 and 3692 DF,  p-value: < 2.2e-16
```

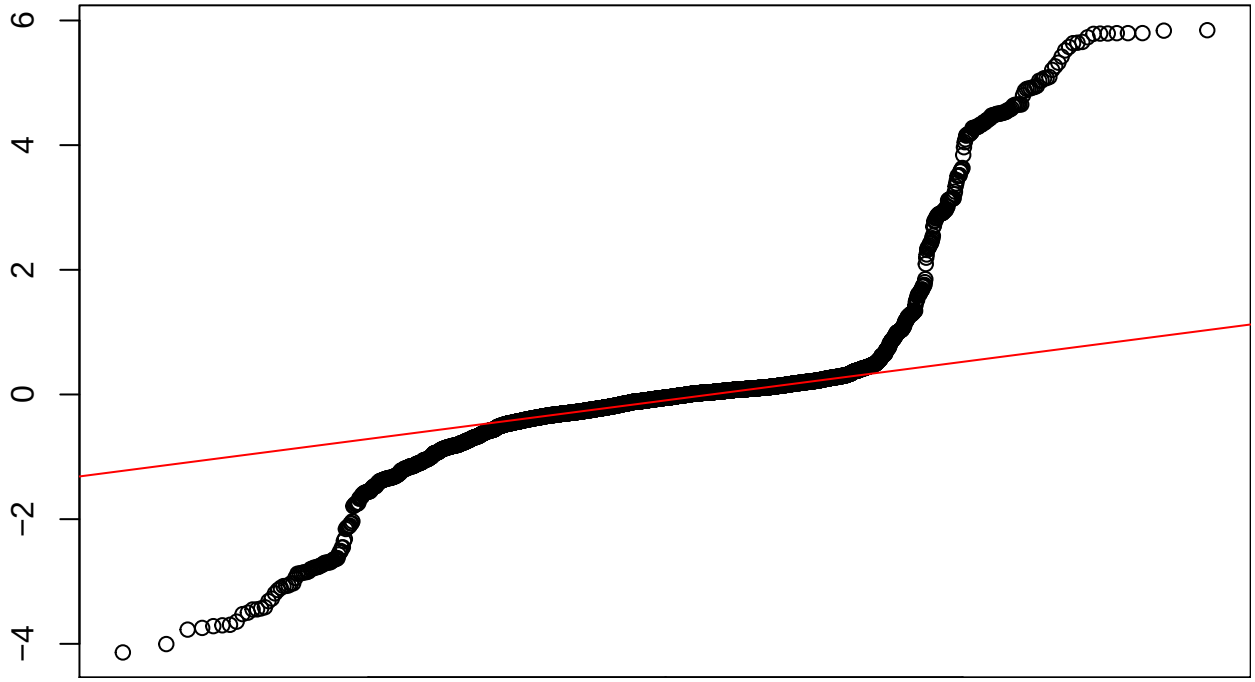
```r
#Adjusted R-squared je u redu vrijednost, no opet vidimo puno nesignifikantnih kategorija,
#možemo li poboljšati adjusted R-squared vrijednost micanjem tih kategorija.
#Napravimo linearnu regresiju za gyms koristeći sve signifikatne kategorije.
fit.gyms.edit = lm(gyms~beaches+parks+theatres+restaurants+pubs.bars+hotels.other.lodgings+
                     art.galleries+dance.clubs+swimming.pools+bakeries+
                     cafes+view.points+gardens, data=pod.cor)
summary(fit.gyms.edit)
```

```
##
```

```
## Call:
## lm(formula = gyms ~ beaches + parks + theatres + restaurants +
##     pubs.bars + hotels.other.lodgings + art.galleries + dance.clubs +
##     swimming.pools + bakeries + cafes + view.points + gardens,
##     data = pod.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8270 -0.2089 -0.0358  0.0773  4.0116
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.792192   0.078717  10.064  < 2e-16 ***
## beaches                -0.022745   0.010516  -2.163 0.030613 *
## parks                  -0.036873   0.011506  -3.205 0.001364 **
## theatres               -0.052408   0.011337  -4.623 3.91e-06 ***
## restaurants            -0.056286   0.010673  -5.274 1.41e-07 ***
## pubs.bars              -0.027866   0.011041  -2.524 0.011648 *
## hotels.other.lodgings   0.026475   0.008171   3.240 0.001205 **
## art.galleries          -0.023063   0.007085  -3.255 0.001144 **
## dance.clubs             0.074460   0.011301   6.589 5.04e-11 ***
## swimming.pools          0.401779   0.014695  27.340  < 2e-16 ***
## bakeries                0.166625   0.010428  15.978  < 2e-16 ***
## cafes                   0.030197   0.013531   2.232 0.025692 *
## view.points            -0.022735   0.007980  -2.849 0.004412 **
## gardens                 0.036677   0.010440   3.513 0.000448 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6872 on 3813 degrees of freedom
##   (1621 observations deleted due to missingness)
## Multiple R-squared:  0.4065, Adjusted R-squared:  0.4044
## F-statistic: 200.9 on 13 and 3813 DF,  p-value: < 2.2e-16
```

```r
#Ovaj put smo ipak dobili različitu vrijednost za adjusted r-squared i to veću.

#Pogledajmo reziduale
par(mfrow=c(1,1),mar = c(0, 2, 5,0))
qqnorm(rstandard(fit.gyms.edit), main ="gyms~beaches+parks+theatres+restaurants+pubs.bars+
hotels.other.lodgings+art.galleries+dance.clubs+swimming.pools+
bakeries+cafes+view.points+gardens")
qqline(rstandard(fit.gyms.edit), col = "red")
```

gyms~beaches+parks+theatres+restaurants+pubs.bars+ hotels.other.lodgings+art.galleries+dance.clubs+swimming.pools+ bakeries+cafes+view.points+gardens




```
#Pekare i kafići bi u stvarnome svijjetu mogli biti viđeni kao negativno korelirani s
#teretanama
#Možemo li iz ocjena pekara i kafića predvidjeti ocjenu teretane?

fit.gym.spec1 = lm(gyms~bakeries+cafes, data = pod.cor)
summary(fit.gym.spec1)
```

```
##
## Call:
## lm(formula = gyms ~ bakeries + cafes, data = pod.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7736 -0.1974 -0.1073 -0.0118  4.2710
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.49640    0.02142  23.174  < 2e-16 ***
## bakeries     0.30596    0.01036  29.524  < 2e-16 ***
## cafes        0.09947    0.01309   7.599  3.7e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.78 on 3951 degrees of freedom
##   (1494 observations deleted due to missingness)
## Multiple R-squared:  0.2089, Adjusted R-squared:  0.2085
## F-statistic: 521.7 on 2 and 3951 DF,  p-value: < 2.2e-16
```

```r
#Vrijednost adjusted r-squared je ponovno uredna, no možemo li poboljjšati našu regresiju?

#Što ako dodamo swimming.pools?

fit.gym.spec2 = lm(gyms~bakeries+cafes+swimming.pools, data = pod.cor)
summary(fit.gym.spec2)
```

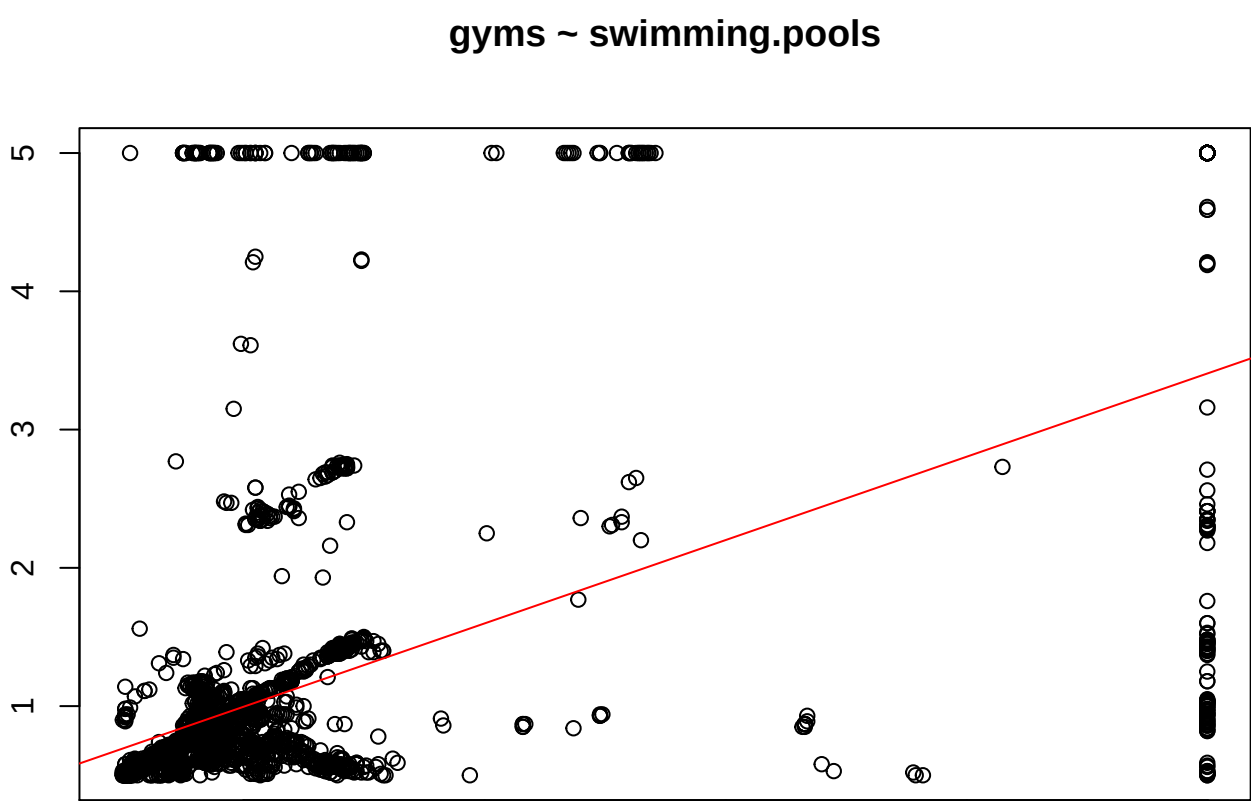
```
##
## Call:
## lm(formula = gyms ~ bakeries + cafes + swimming.pools, data = pod.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8938 -0.1139 -0.0570  0.0019  4.2117
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.212532   0.021275   9.990  < 2e-16 ***
## bakeries       0.203836   0.009833  20.730  < 2e-16 ***
## cafes          0.059161   0.011815   5.007 5.76e-07 ***
## swimming.pools 0.455261   0.014204  32.051  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6997 on 3881 degrees of freedom
##   (1563 observations deleted due to missingness)
## Multiple R-squared:  0.3741, Adjusted R-squared:  0.3736
## F-statistic: 773.3 on 3 and 3881 DF,  p-value: < 2.2e-16
```

```r
#Adjusted r-squared vrijednost je poskočila, izgleda da su teretane i bazeni dobro
#korelirani.
#To također možemo vidjeti iz vrijednosti faktora uz swimming.pools.
#Kako radimo regresiju jedne kategorije sa samo još jednom drugom ovo je sjajna prilika
#za crtanje točaka.

plot(x = pod.cor$swimming.pools, y = pod.cor$gyms,
     main = "gyms ~ swimming.pools", xlab = "swimming.pools", ylab = "gyms" )
fit.gym.swim = lm(gyms~swimming.pools, data = pod.cor)
summary(fit.gym.swim)
```

```
##
## Call:
## lm(formula = gyms ~ swimming.pools, data = pod.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9054 -0.1717 -0.1164 -0.0483  4.2884
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.39220    0.01775   22.10   <2e-16 ***
## swimming.pools  0.60263    0.01277   47.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.7844 on 4360 degrees of freedom
##   (1086 observations deleted due to missingness)
## Multiple R-squared:  0.338,  Adjusted R-squared:  0.3378
## F-statistic:  2226 on 1 and 4360 DF,  p-value: < 2.2e-16
```

```
#Multiple R-squared vrijednost je opala, no ne previše, što je bilo za očekivati.
#Ipak i dalje je u redu.
#Iscrtajmo i dobivenu liniju regersiju
abline(fit.gym.swim, col="red")
```




```
#Utjecaj religiju u kulturi je vidljiv kroz cijelu povijest, no očituje li se i u
#našim podacima.
#Pogledajmo možemo li ocjenu crkava predvidjeti ocjenama kulturnih kategorija.

fit.religija.kultura = lm(churches ~ museums + monuments + art.galleries + theatres,
                          data = pod.cor)
summary(fit.religija.kultura)
```

```
##
## Call:
## lm(formula = churches ~ museums + monuments + art.galleries +
##     theatres, data = pod.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6494 -0.3860 -0.1239  0.2374  3.9631
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.649132   0.040627  40.592  < 2e-16 ***
```

```
## museums       -0.090433   0.009161  -9.871  < 2e-16 ***
## monuments      0.188656   0.008065  23.392  < 2e-16 ***
## art.galleries -0.051668   0.006383  -8.095 7.11e-16 ***
## theatres      -0.016478   0.009012  -1.828   0.0675 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7246 on 5136 degrees of freedom
##   (307 observations deleted due to missingness)
## Multiple R-squared:  0.1519, Adjusted R-squared:  0.1512
## F-statistic: 229.9 on 4 and 5136 DF,  p-value: < 2.2e-16
```

```
#Vidimo da kazališta možemo izbaciti iz regresije
fit.religija.kultura.edit = lm(churches ~ museums + monuments + art.galleries,
                               data = pod.cor)
summary(fit.religija.kultura.edit)
```

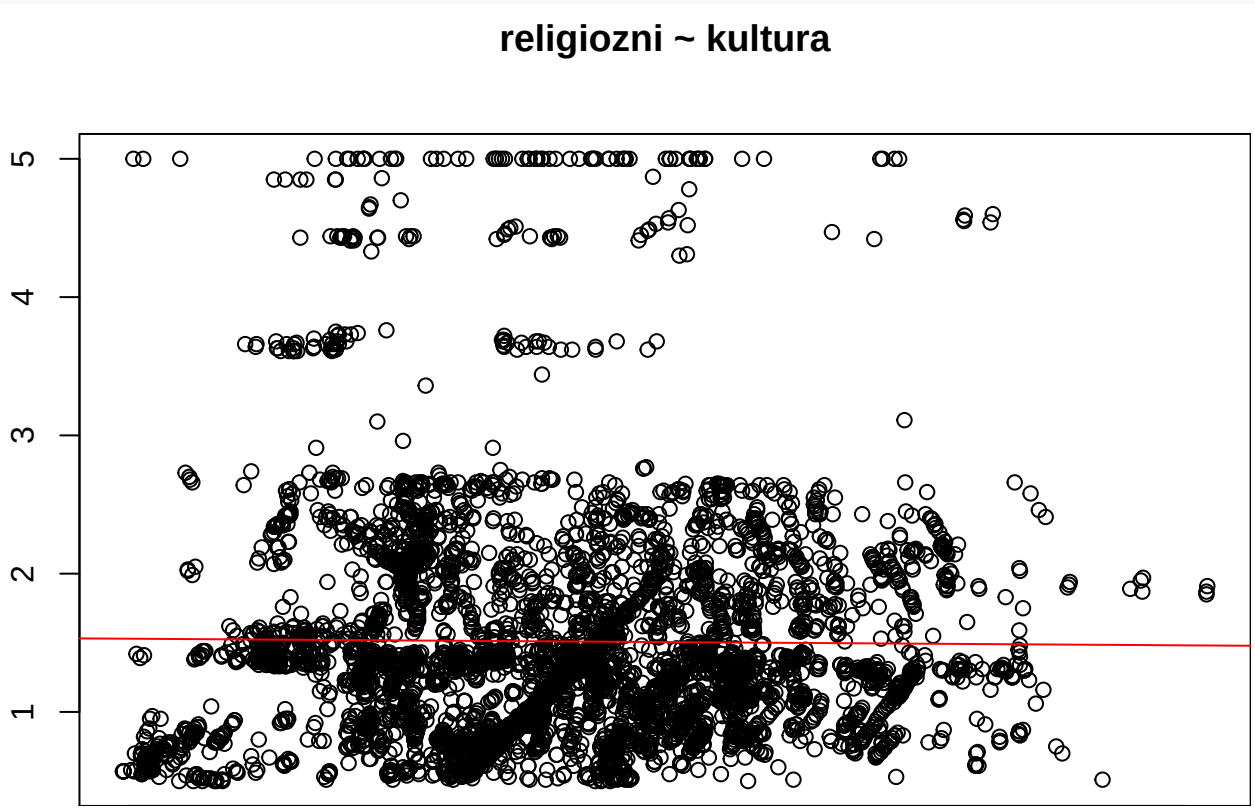
```
##
## Call:
## lm(formula = churches ~ museums + monuments + art.galleries,
##     data = pod.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6289 -0.3918 -0.1224  0.2336  3.9956
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.617822   0.036851  43.902  < 2e-16 ***
## museums       -0.097955   0.008188 -11.964  < 2e-16 ***
## monuments      0.186907   0.008010  23.335  < 2e-16 ***
## art.galleries -0.048684   0.006172  -7.888 3.74e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7247 on 5137 degrees of freedom
##   (307 observations deleted due to missingness)
## Multiple R-squared:  0.1513, Adjusted R-squared:  0.1508
## F-statistic: 305.3 on 3 and 5137 DF,  p-value: < 2.2e-16
```

```
#Izbacivanjem kazališta nismo postigli ništa, no adjusted R-squared vrijednost
#je i dalje uredna.

#Možemo li dobiti drukčije rezutate ako koristimo grupe religiozni i kultura iz
#grupirane tablice?
fit.religija.kultura.grupe = lm(religiozni ~ kultura, data = pod.new.cor)
summary(fit.religija.kultura.grupe)
```

```
##
## Call:
## lm(formula = religiozni ~ kultura, data = pod.new.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0272 -0.5415 -0.1630  0.3554  3.5055
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.54326    0.04351  35.468   <2e-16 ***
## kultura     -0.01373    0.01725  -0.796    0.426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7928 on 5250 degrees of freedom
##   (196 observations deleted due to missingness)
## Multiple R-squared:  0.0001206,  Adjusted R-squared:  -6.982e-05
## F-statistic: 0.6334 on 1 and 5250 DF,  p-value: 0.4262
```

```r
#Izgleda da ova regresija nije dobra što je očito iz jako male vrijednosti
#multiple R-squared vrijednosti
#Pogledajmo kako izgledaju podaci na grafu
plot(x = pod.new.cor$kultura,y =  pod.new.cor$religiozni,
     main = "religiozni ~ kultura",
     xlab = "kultura", ylab = "raligiozni")
abline(fit.religija.kultura.grupe, col = "red")
```




```r
#Kao što smo i mogli pročitati iz summary funkcije ove linearne regresije,
#p-vrijednost F-statistike
#nam ukazuje da linija regresije koju povlačimo nije značajno bolja u predikciji
#od linije y = sredina y vrijednosti na grafu, tj. y-intercepta

#Nastavimo istraživati tablicu grupiranihh podataka.
#Hrana se čini kao grupa s najboljim korelacijama s ostalim grupama.
fit.hrana = lm(hrana ~ ., data = pod.new.cor)
summary(fit.hrana)
```

```
##
## Call:
## lm(formula = hrana ~ ., data = pod.new.cor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58074 -0.39863 -0.06014  0.29784  2.59807
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.065904   0.066261  31.178  < 2e-16 ***
## kultura       -0.062720   0.013613  -4.607 4.18e-06 ***
## pice           0.148259   0.012348  12.006  < 2e-16 ***
## zabava         0.094197   0.015168   6.210 5.73e-10 ***
## sport          0.035067   0.009436   3.716 0.000204 ***
## priroda       -0.277570   0.011477 -24.184  < 2e-16 ***
## ugostiteljski  0.062951   0.009428   6.677 2.71e-11 ***
## religiozni    -0.045700   0.011109  -4.114 3.96e-05 ***
## ostalo         0.108237   0.006333  17.090  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5691 on 4848 degrees of freedom
##   (591 observations deleted due to missingness)
## Multiple R-squared:  0.3488, Adjusted R-squared:  0.3477
## F-statistic: 324.6 on 8 and 4848 DF,  p-value: < 2.2e-16
```

```
#Adjusted R-squared vrijednost je u redu, te niti jedna grupa nije nesignifikatna
#u ovoj linearnoj regresiji

#Razmislimo na trenutak o stvarnome svijetu, mogli bismo reći da ima smisla
#pretpostavka da
#korisnici koji vole boravak u prirodi nisu baš nasretniji kada vrijeme provode u
#zatvorenim prostorima
#npr. u kafićima, trgovačkim centrima te restoranima i slično
#Ispitajmo tu pretpostaku, pogledajmo možemo li iz ocjena grupa zabava, hrana i
#pice predvidjeti ocjenu grupe priroda.
fit.priroda.pretp = lm(priroda ~ zabava + hrana + pice, data = pod.new.cor)
summary(fit.priroda.pretp)
```

```
##
## Call:
## lm(formula = priroda ~ zabava + hrana + pice, data = pod.new.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8186 -0.5212 -0.1374  0.4581  2.5317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.99085    0.04254   93.81   <2e-16 ***
## zabava      -0.20441    0.01607  -12.72   <2e-16 ***
## hrana       -0.43640    0.01484  -29.40   <2e-16 ***
## pice        -0.18198    0.01329  -13.70   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7072 on 5443 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2862, Adjusted R-squared:  0.2858
## F-statistic: 727.4 on 3 and 5443 DF,  p-value: < 2.2e-16
```

```
#Dobili smo urednu vrijednost za adjusted R-squared, pogledajmo koliko bolje možemo
#predvidjeti prirodu ako uključimo
#ostale grupe u regresiju
fit.priroda.sve = lm(priroda ~ ., data = pod.new.cor)
summary(fit.priroda.sve)
```

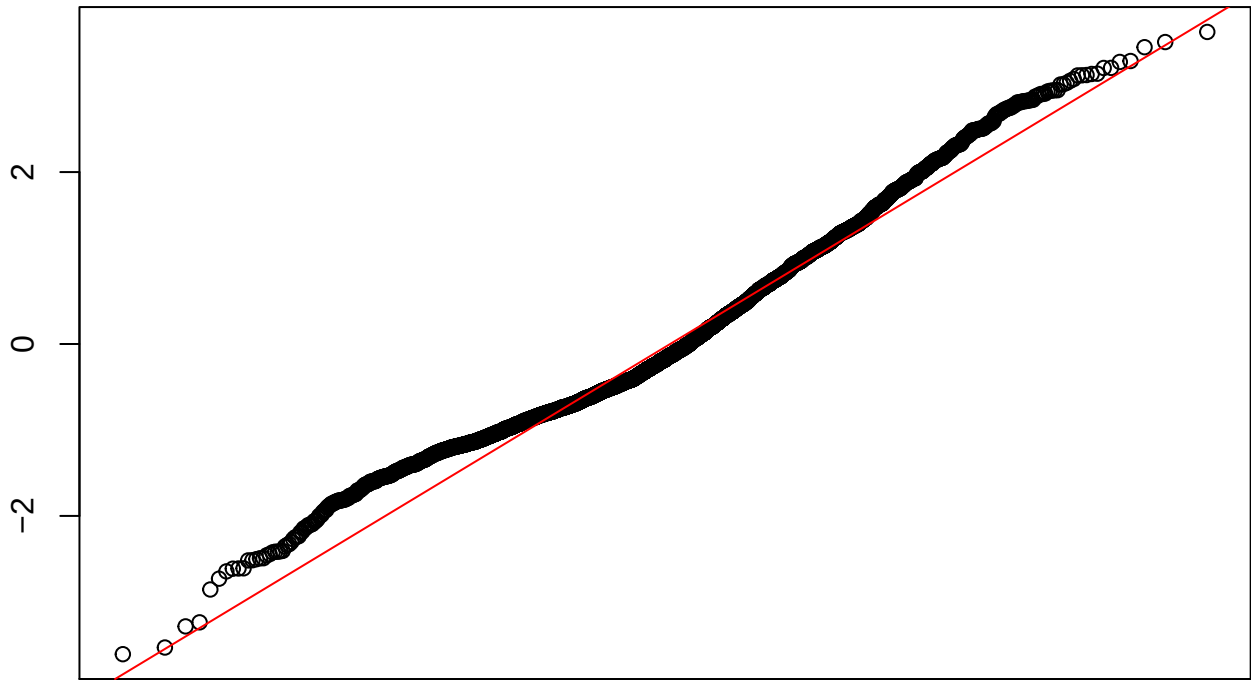
```
##
## Call:
## lm(formula = priroda ~ ., data = pod.new.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3998 -0.4872 -0.1144  0.4440  2.4337
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.904020   0.075003  38.719  < 2e-16 ***
## kultura        0.233437   0.015774  14.799  < 2e-16 ***
## hrana         -0.387843   0.016037 -24.184  < 2e-16 ***
## pice          -0.142930   0.014669  -9.744  < 2e-16 ***
## zabava        -0.257921   0.017616 -14.642  < 2e-16 ***
## sport          0.004460   0.011169   0.399   0.6897
## ugostiteljski -0.021985   0.011191  -1.965   0.0495 *
## religiozni     0.229254   0.012736  18.000  < 2e-16 ***
## ostalo         0.047753   0.007678   6.219 5.41e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6727 on 4848 degrees of freedom
##   (591 observations deleted due to missingness)
## Multiple R-squared:  0.3709, Adjusted R-squared:  0.3698
## F-statistic: 357.2 on 8 and 4848 DF,  p-value: < 2.2e-16
```

```
#Izbacimo nesignifikantne grupe sport i ugostiteljski
fit.priroda.sve = lm(priroda ~ . -sport -ugostiteljski, data = pod.new.cor)
summary(fit.priroda.sve)
```

```
##
## Call:
## lm(formula = priroda ~ . - sport - ugostiteljski, data = pod.new.cor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4227 -0.4896 -0.1172  0.4411  2.4406
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.877202   0.072480  39.696  < 2e-16 ***
## kultura      0.230356   0.015651  14.718  < 2e-16 ***
```

```
## hrana       -0.390953   0.015921 -24.555  < 2e-16 ***
## pice        -0.146864   0.014526 -10.111  < 2e-16 ***
## zabava      -0.253653   0.017485 -14.507  < 2e-16 ***
## religiozni   0.227452   0.012490  18.211  < 2e-16 ***
## ostalo       0.047215   0.007608   6.206  5.9e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6728 on 4850 degrees of freedom
##   (591 observations deleted due to missingness)
## Multiple R-squared:  0.3703, Adjusted R-squared:  0.3696
## F-statistic: 475.4 on 6 and 4850 DF,  p-value: < 2.2e-16
```

```
#Regresija nam je dobra, pogledajmo reziduale, očekivano je da će biti bolji nego
#u regresijama
#s kategorijama običnihh podataka, ipak su ove grupe normalnije distribuirane
qqnorm(rstandard(fit.priroda.sve), main="priroda ~ . -sport -ugostiteljski")
qqline(rstandard(fit.priroda.sve), col = "red")
```

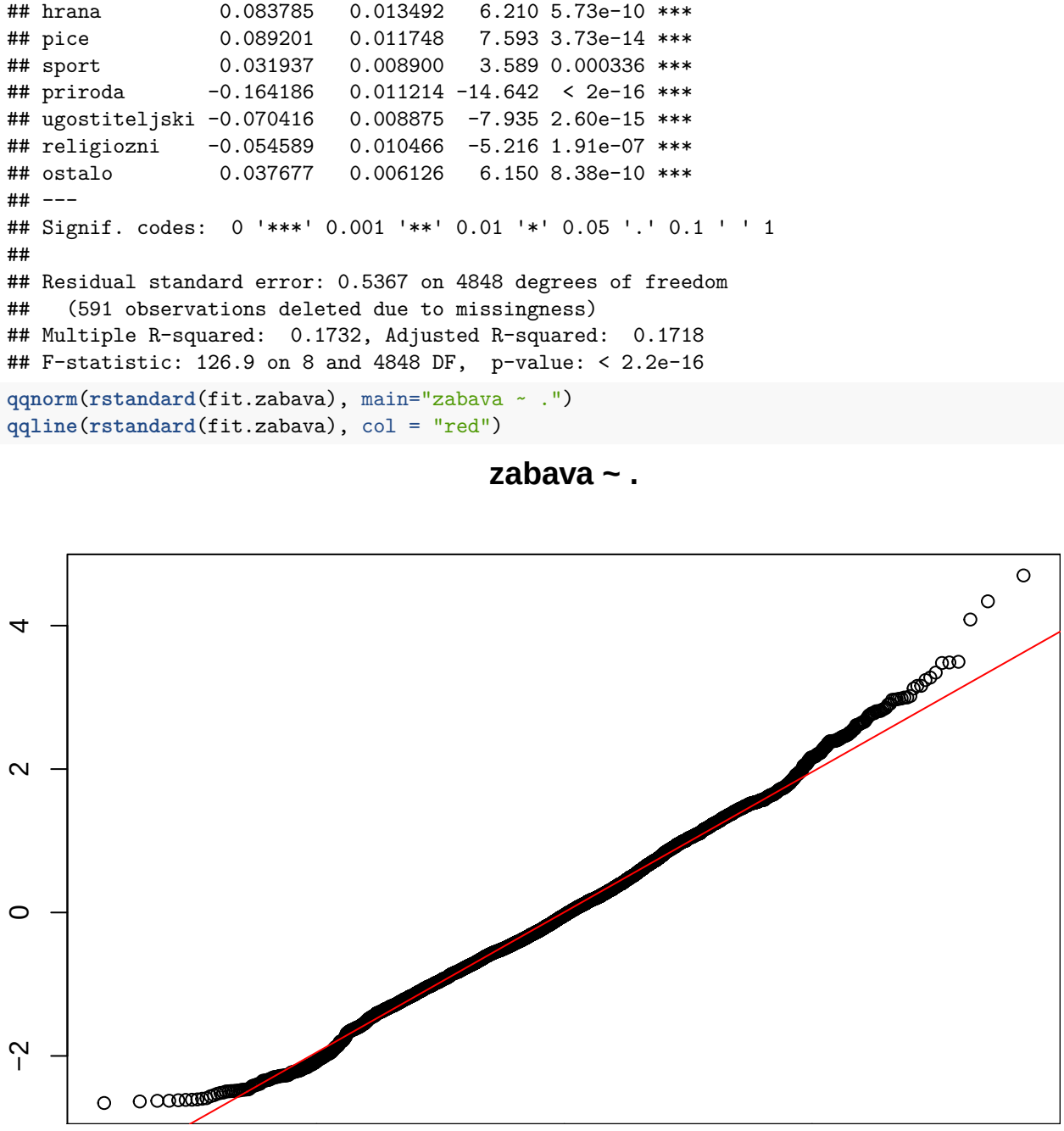



```
#Linearna regresija za grupu zabava prema svim grupama daje dobre reziduale,
#pogledajmo.
fit.zabava = lm(zabava ~ ., data = pod.new.cor)
summary(fit.zabava)
```

```
##
## Call:
## lm(formula = zabava ~ ., data = pod.new.cor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.42321 -0.35321 -0.02359  0.35359  2.52101
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.005900   0.062116  32.293  < 2e-16 ***
## kultura        0.092262   0.012798   7.209 6.50e-13 ***
## hrana          0.083785   0.013492   6.210 5.73e-10 ***
## pice           0.089201   0.011748   7.593 3.73e-14 ***
## sport          0.031937   0.008900   3.589 0.000336 ***
## priroda       -0.164186   0.011214 -14.642  < 2e-16 ***
## ugostiteljski -0.070416   0.008875  -7.935 2.60e-15 ***
## religiozni    -0.054589   0.010466  -5.216 1.91e-07 ***
## ostalo         0.037677   0.006126   6.150 8.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5367 on 4848 degrees of freedom
##   (591 observations deleted due to missingness)
## Multiple R-squared:  0.1732, Adjusted R-squared:  0.1718
## F-statistic: 126.9 on 8 and 4848 DF,  p-value: < 2.2e-16
```

```
qqnorm(rstandard(fit.zabava), main="zabava ~ .")
qqline(rstandard(fit.zabava), col = "red")
```

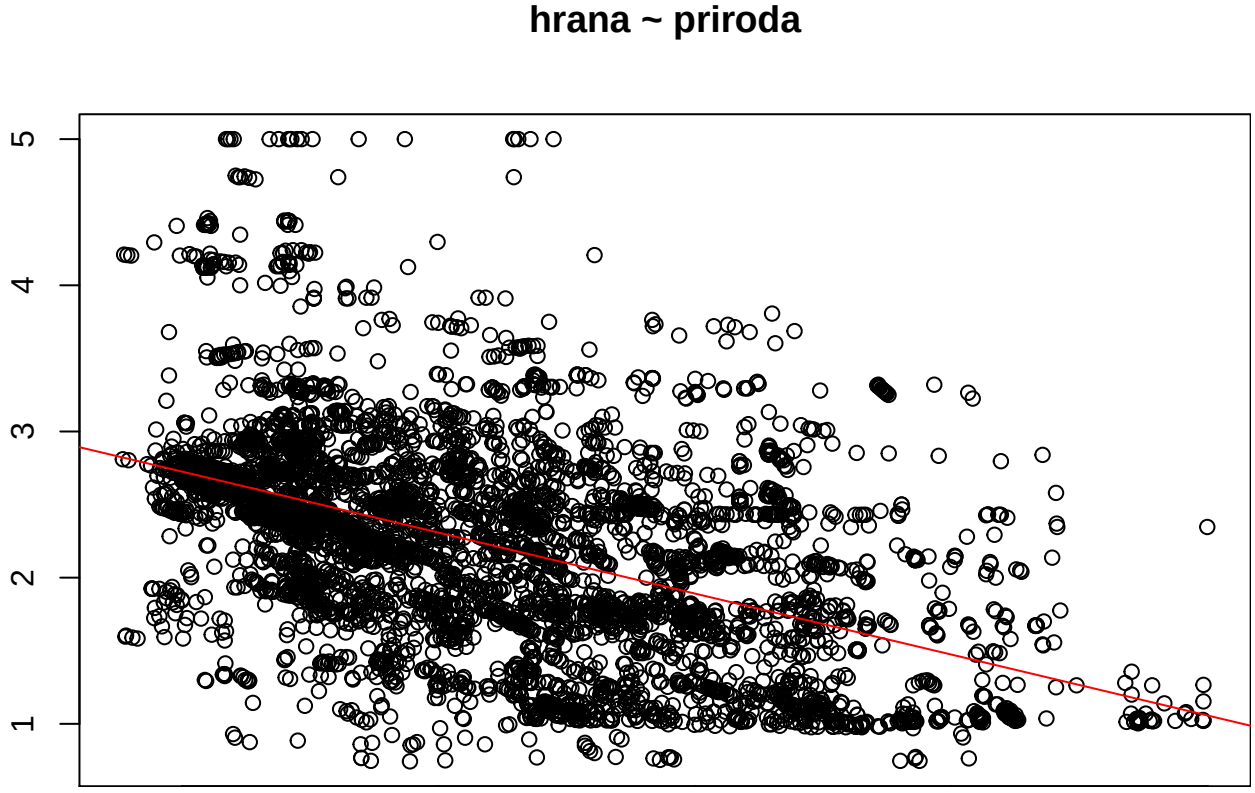



```
#Korelacija grupa hrana i priroda dobro izgleda na grafu korelacija, isprobajmo
fit.hrana.priroda = lm(hrana ~ priroda, data = pod.new.cor)
summary(fit.hrana.priroda)
```

```
##
## Call:
```

```
## lm(formula = hrana ~ priroda, data = pod.new.cor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74038 -0.40235 -0.06486  0.34274  2.87888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.14300    0.02401  130.93   <2e-16 ***
## priroda     -0.41709    0.01021  -40.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6304 on 5445 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2347, Adjusted R-squared:  0.2345
## F-statistic:  1670 on 1 and 5445 DF,  p-value: < 2.2e-16
```

```
plot(x = pod.new.cor$priroda, y = pod.new.cor$hrana, main = "hrana ~ priroda",
     xlab = "priroda", ylab = "hrana")
abline(fit.hrana.priroda, col = "red")
```




```
#Vidi se da je riječ o negativnoj korelaciji, regresija se čini uredna.

#Dodajmo sada grupu piće i ostalo u regersiju i pogledajmo koliko
#blizu možemo doći vrijednosti adjusted R-squared regresije hrane sa svim
#atributima
fit.hrana.edit = lm(hrana ~ priroda + pice + ostalo, data = pod.new.cor)
summary(fit.hrana.edit)
```

```
##
## Call:
## lm(formula = hrana ~ priroda + pice + ostalo, data = pod.new.cor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78502 -0.39206 -0.04809  0.29541  2.58746
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.227113   0.039437   56.47   <2e-16 ***
## priroda     -0.317528   0.010175  -31.21   <2e-16 ***
## pice         0.192951   0.011020   17.51   <2e-16 ***
## ostalo       0.113507   0.006025   18.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5847 on 5442 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3417, Adjusted R-squared:  0.3413
## F-statistic: 941.6 on 3 and 5442 DF,  p-value: < 2.2e-16
```

```
fit.hrana.sve = lm(hrana ~ ., data = pod.new.cor)
summary(fit.hrana.sve)
```

```
##
## Call:
## lm(formula = hrana ~ ., data = pod.new.cor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58074 -0.39863 -0.06014  0.29784  2.59807
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.065904   0.066261  31.178  < 2e-16 ***
## kultura       -0.062720   0.013613  -4.607 4.18e-06 ***
## pice           0.148259   0.012348  12.006  < 2e-16 ***
## zabava         0.094197   0.015168   6.210 5.73e-10 ***
## sport          0.035067   0.009436   3.716 0.000204 ***
## priroda       -0.277570   0.011477 -24.184  < 2e-16 ***
## ugostiteljski  0.062951   0.009428   6.677 2.71e-11 ***
## religiozni    -0.045700   0.011109  -4.114 3.96e-05 ***
## ostalo         0.108237   0.006333  17.090  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5691 on 4848 degrees of freedom
##   (591 observations deleted due to missingness)
## Multiple R-squared:  0.3488, Adjusted R-squared:  0.3477
## F-statistic: 324.6 on 8 and 4848 DF,  p-value: < 2.2e-16
```

```
#Vidimo da razlika u adjusted R-squared vrijednostima nije velika, dakle umjesto
#regresije sa svim grupama,
#za grupu hrana možemo koristiti regresiju samo s tri grupe; priroda, pice i ostalo.
```

```
#Linearna regresija --kraj
```