Projekt Speed Dating

Fran Canjuga, Gašpar Haramija, Leon Hegedić, Josipa Markić

2024-01-21

Uvod

U ovom projektnom zadatku, naziva Speed Dating, analizirat ćemo skup podataka prikupljen između 2002. i 2004. godine u sklopu kolegija Statistička analiza podataka na Fakultetu elektrotehnike i računarstva. Dostupni su dva skupa podataka, a analizirat ćemo ih pomoću deskriptivne statistike i inferencijalne analize kako bismo dobili dublji uvid u ponašanje sudionika i međuodnose varijabli.

Ciljevi ovog projekta su upoznavanje s dostupnim podacima o sudionicima i spojevima te izvlačenje relevantnih informacija pomoću deskriptivne statistike i inferencijalne analize. Paralelno s analizom, planiramo unaprijediti svoje razumijevanje osnovnih metodologija statističke analize podataka i praktičnu primjenu programskog jezika R što ćemo primijeniti na sljedećim hipotezama na temelju kojih ćemo izvući zaključke:

1. Je li inteligencija partnera ispitanicima važnija od izgleda?
2. Postoji li razlika u interesu za gaming prema zanimanju sudionika?
3. Možemo li temeljem drugih varijabli predvidjeti hoće li se sudioniku svidjeti partner?

U svakom od navedenih pitanja pristupit ćemo posebno te ćemo počet s analizom deskriptivne statistike kako bismo stekli osnovni uvid u potrebne varijable poput spola, zanimanja i ocjena partnera. Nakon toga, primijenit ćemo inferencijalnu analizu kako bismo istražili postavljene hipoteze, uključujući pitanja o važnosti inteligencije partnera u odnosu na izgled te razlike u interesu za gaming prema zanimanju sudionika.

Ovaj projekt ima značajnost iz perspektive studenata, poput nas samih, jer nam pruža priliku primijeniti stečeno znanje o statističkoj analizi podataka u stvarnom svijetu. Uvid u ovakav skup podataka može pomoći studentima u donošenju informiranih odluka u svakodnevnom životu, bilo da se radi o razumijevanju međuljudskih odnosa ili donošenju odluka temeljenih na vlastitim preferencijama.

Učitavanje i prikaz podataka

```
participants = read.csv("participant_data.csv")
dates = read.csv("speed_date_data.csv")
```

Tijekom analize možemo odlučiti izbaciti stupce koji nisu nužni jer ne pridonose zaključcima. Također, moguće je stvoriti pomoćne skupove podataka koji će biti korisni u određenim analizama, kao što je priprema podataka za svaki od zadataka.

Dodatno, potrebno je razmotriti klasau nekih varijabli koje nisu precizno definirane u izvornim skupovima podataka ukoliko će to biti potrebno za analizu zadatka. Primjerice, varijablu "spol" možemo pretvoriti u numeričku, a za varijablu "importance_same_race" koja trenutačno ima karakterističnu klasu (character), potrebno je razmotriti njenu pravu prirodu i prilagoditi klasifikaciju prema potrebi.

U skupu podataka "dates" primjećujemo značajan broj nedostajućih podataka označenih s NA. Prije daljnje analize, potrebno je donijeti odluku o postupanju s tim podacima, kao što je zamjena nedostajućih vrijednosti odgovarajućim podacima, brisanje redaka ili primjena drugih metoda obrade nedostajućih podataka, ovisno o kontekstu analize zadatka.

```r
# Osnovna deskriptivna statistika:
#summary(participants)
#summary(dates)
```

```r
cat("Dimenzija podataka o participantima: ", dim(participants))
```

```
## Dimenzija podataka o participantima:  540 40
```

```r
cat("Dimenzija podataka o spojevima: ", dim(dates))
```

```
## Dimenzija podataka o spojevima:  8086 19
```

Na temelju dostupnih podataka saznajemo da skup podataka za sudionike sadrži podatke o 540 sudionika te 40 stupaca, dok skup podataka o spojevima sadrži podatke o 8086 spojeva i ima 19 stupaca.

1. Je li inteligencija partnera ispitanicima važnija od izgleda?

Za rješavanje ovog pitanja ključni su nam podaci o važnosti izgleda i inteligencije među ispitanicima. Ovi podaci nalaze se u okviru varijabli "attractive_important" i "intelligence_important" unutar skupa podataka "participants".

Prvo ćemo predstaviti mjere centralne tendencije za navedene podatke te ćemo ih vizualizirati. Osim toga, podatke ćemo gledati za cijeli skup podataka, ali također ćemo podijeliti cijeli skup na dva dijela, muškarce i žene te ćemo ovom pitanju pristupiti iz više kuteva: je li inteligencija partnera ispitanicima važnija od izgleda, je li inteligencija partnera muškarcima važnija od izgleda u ovom skupu podataka te je li inteligencija partnera ženama važnija od izgleda u ovom skupu podataka.

Podatke ćemo prvo očistiti, vizualizirati te ćemo nastaviti s potrebnim testovima. Naposljetku, ćemo rezultate analizirati, usporediti te donijeti zaključke.

Prvo ćemo izvaditi sve null vrijednosti iz podataka.

```r
sum(is.na(participants$intelligence_important))
```

```
## [1] 0
```

```r
sum(is.na(participants$attractive_important))
```

```
## [1] 0
```

```r
df_cleaned_first <- participants[, c("intelligence_important", "attractive_important")]
df_cleaned_first_male <- participants[participants$gender == "male",
                                      c("intelligence_important", "attractive_important")]
df_cleaned_first_female <- participants[participants$gender == "female",
                                        c("intelligence_important", "attractive_important")]
summary(df_cleaned_first)
```

```
##  intelligence_important attractive_important
##  Min.   : 0.00          Min.   :  0.00
##  1st Qu.:17.29          1st Qu.: 15.00
##  Median :20.00          Median : 20.00
##  Mean   :20.17          Mean   : 22.52
##  3rd Qu.:23.02          3rd Qu.: 25.00
##  Max.   :50.00          Max.   :100.00
```

```r
summary(df_cleaned_first_male)
```

```
##  intelligence_important attractive_important
##  Min.   : 0.00          Min.   :  6.67
##  1st Qu.:16.00          1st Qu.: 19.57
```

```
##  Median :20.00          Median : 23.00
##  Mean   :19.42          Mean   : 27.01
##  3rd Qu.:22.22          3rd Qu.: 30.00
##  Max.   :42.86          Max.   :100.00
```

```
summary(df_cleaned_first_female)
```

```
##  intelligence_important attractive_important
##  Min.   : 2.00          Min.   : 0.00
##  1st Qu.:17.93          1st Qu.:12.12
##  Median :20.00          Median :15.09
##  Mean   :20.94          Mean   :17.93
##  3rd Qu.:25.00          3rd Qu.:20.00
##  Max.   :50.00          Max.   :90.00
```

```
sum(participants$gender == "male")
```

```
## [1] 273
```

```
sum(participants$gender == "female")
```

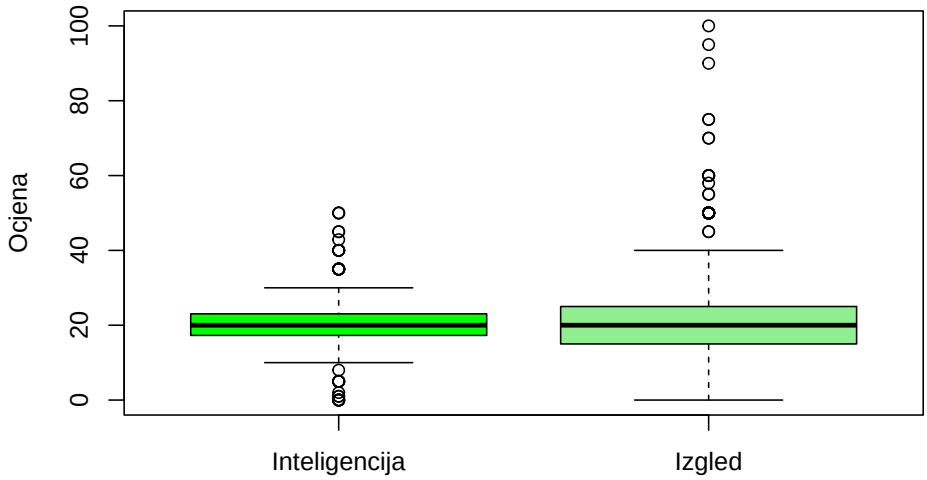
```
## [1] 267
```

Na temelju ispisa ovih podataka možemo vidjeti da su medijani za vrijednosti važnosti inteligencije i izgleda isti kada se gleda skup svih sudionika. Aritmetička sredina vrijednosti kod svih sudionika je veća za važnost izgleda nego za inteligenciju, redom, 22.52 i 20.17. Cijeli skup podataka podijelili smo u dva disjunktna podskupa, skup muškaraca veličine je 273, a žena 263. U podskupu muškaraca primjetna je veća razlika u vrijednostima medijana i aritmetičke sredine u korist važnosti izgleda za razliku od važnosti inteligencije. U podskupu žena je situacija suprotna, vrijednosti medijana i aritmetičke sredine ocjena veće su u korist važnosti inteligencije za razliku od važnosti izgleda. Ove podatke ćemo vizualizirati, a nakon toga i testirati jesu li razlike statistički značajne.

Prvo ćemo vizualizirati pomoću box-plota značajnost inteligencije i izgleda na tri načina, za sve sudionike, muškarce pa žene.

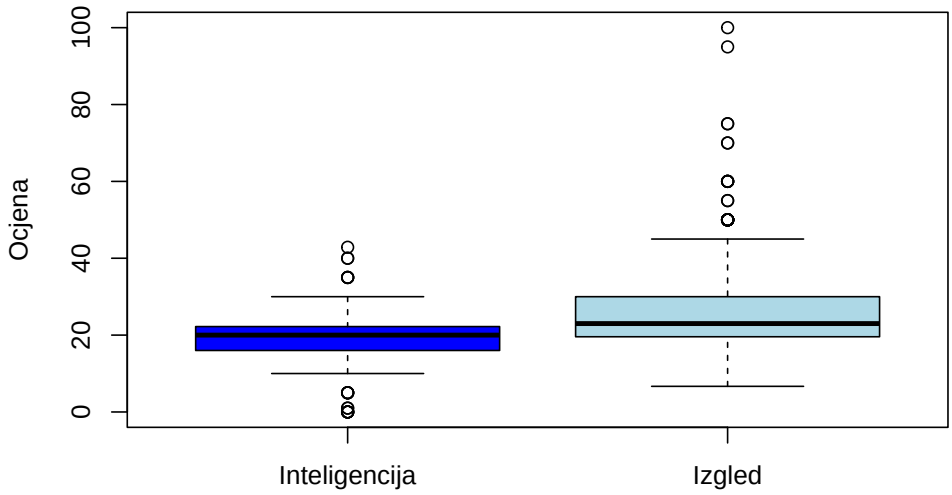
```
boxplot(cbind(df_cleaned_first$intelligence_important,
              df_cleaned_first$attractive_important),
        names = c("Inteligencija", "Izgled"),
        main = 'Značajnost inteligencije i izgleda kod svih sudionika - box-plot',
        ylab = 'Ocjena',
        col = c("green", "lightgreen"))
```

Znacajnost inteligencije i izgleda kod svih sudionika − box−plot

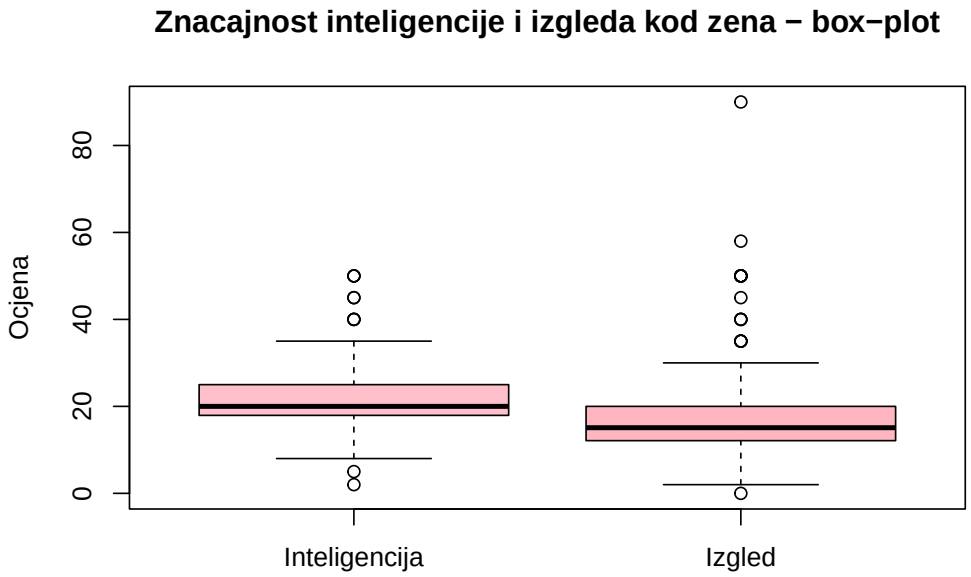



```
boxplot(cbind(df_cleaned_first_male$intelligence_important,
              df_cleaned_first_male$attractive_important),
        names = c("Inteligencija", "Izgled"),
        main = 'Značajnost inteligencije i izgleda kod muškaraca - box-plot',
        ylab = 'Ocjena',
        col = c("blue", "lightblue"))
```

Znacajnost inteligencije i izgleda kod muskaraca − box−plot




```
boxplot(cbind(df_cleaned_first_female$intelligence_important,
              df_cleaned_first_female$attractive_important),
        names = c("Inteligencija", "Izgled"),
        main = 'Značajnost inteligencije i izgleda kod žena - box-plot',
        ylab = 'Ocjena',
        col = c("pink", "lightpink"))
```



Na predočenim dijagramima vidljive su prije spomenute razlike kod muškaraca i žena te jednakost medijana kod svih sudionika. Sljedeći dijagrami prikazuju svaki od varijabli za sve tri spomenute kategorije od jedanput. Prvo za značajnost inteligencije, zatim za izgled.

Sada ćemo te podataka prikazati i uz pomoć histograma. Prvo za inteligenciju zatim za izgled.

Prije nego te podatke prikažemo uz pomoć histograma pogledat ćemo u kojem rasponu su ocjene važnosti inteligencije.

```
summary(participants$intelligence_important)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   17.29   20.00   20.17   23.02   50.00
```

Vidimo da je minimalna ocjena 0, a maksimalna 50.

Sada ćemo pogledati ocjene značajnosti privlačnosti kod partnera.

```
summary(participants$attractive_important)
```

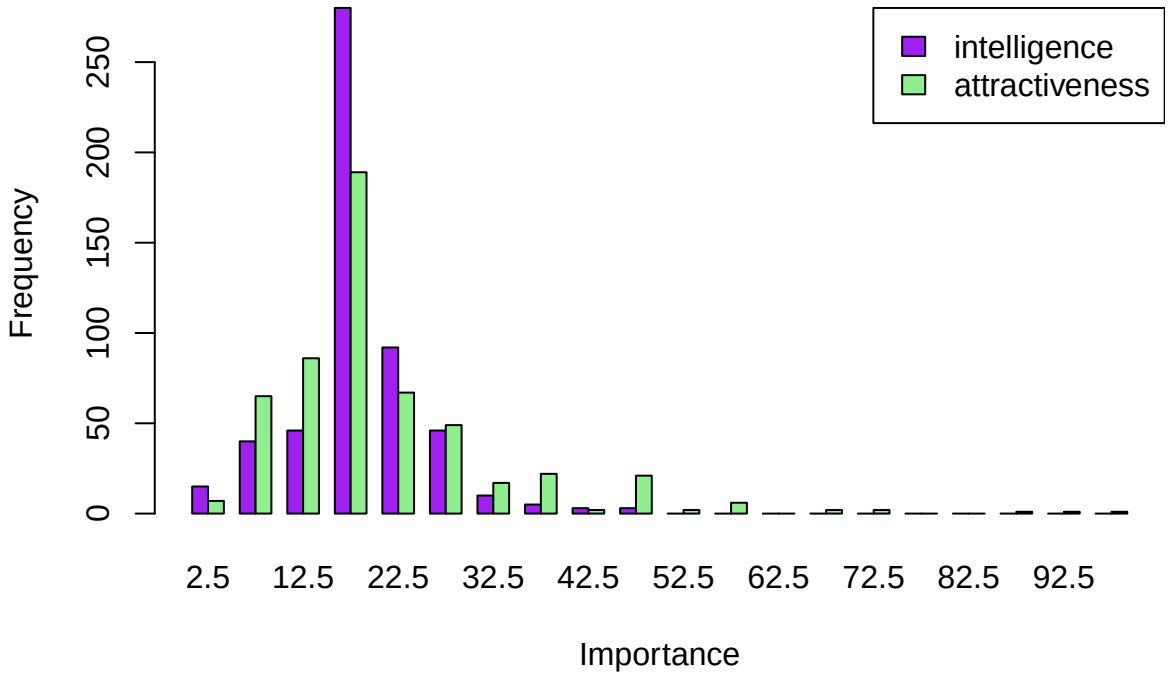
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   15.00   20.00   22.52   25.00  100.00
```

Vidimo da su ocjene u rasponu od 0 do 100.

Podatke o važnosti atraktivnosti i inteligencije ćemo usporediti u histogramu.

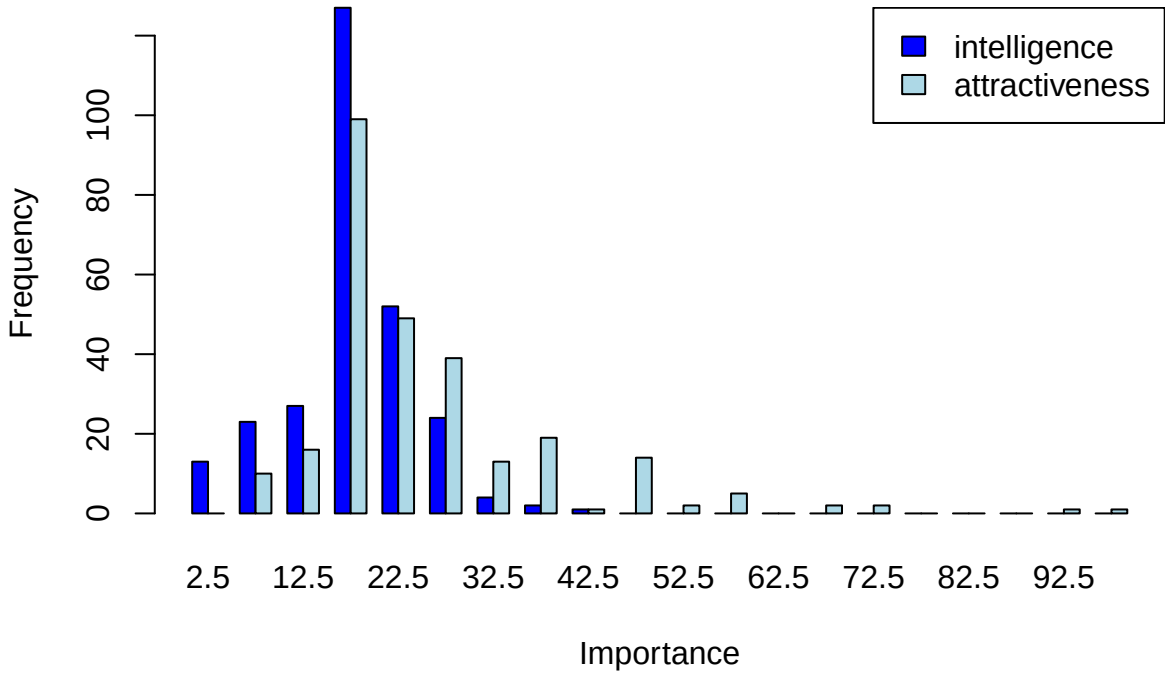
```
b <- seq(0, 100, by = 5)

h1 <- hist(df_cleaned_first$intelligence_important, breaks = b, plot = FALSE)
h2 <- hist(df_cleaned_first$attractive_important, breaks = b, plot = FALSE)

midpoints <- (head(b, -1) + tail(b, -1)) / 2

data <- t(cbind(h1$counts, h2$counts))

barplot(data, beside = TRUE,
        col = c("purple", "lightgreen"),
        xlab = "Importance",
        ylab = 'Frequency',
```

```
        names.arg = midpoints)
legend("topright",
       c("intelligence", "attractiveness"),
       fill = c("purple", "lightgreen"))
```

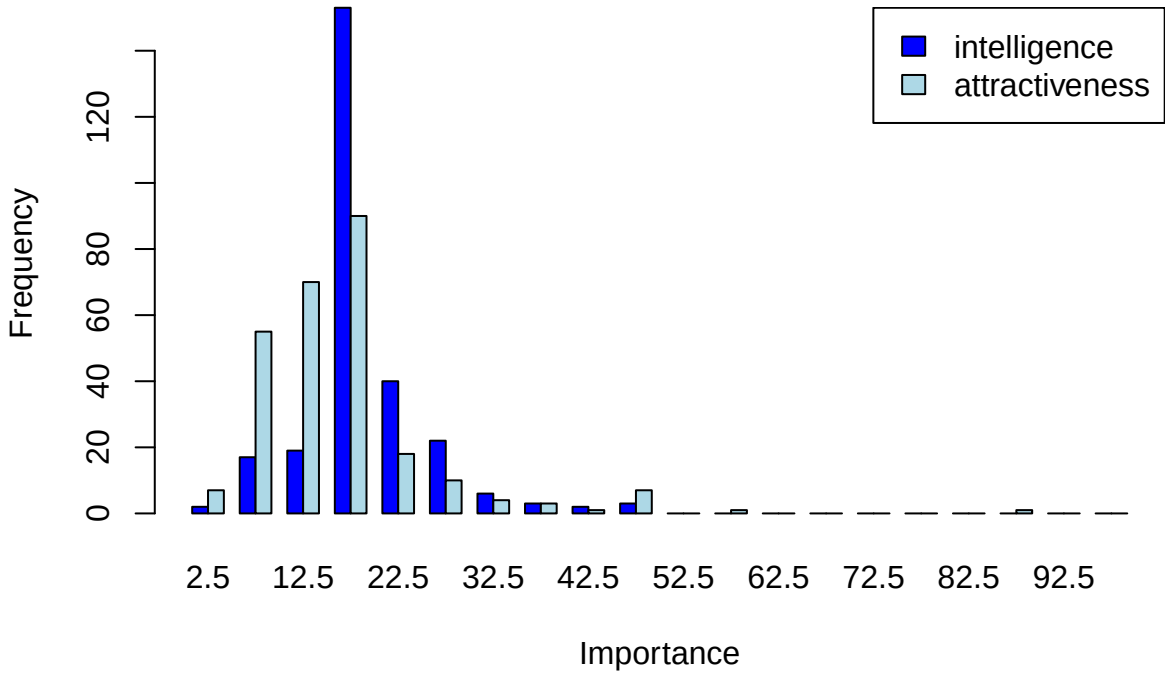


Iz grafa je vidljivo da većina ispitanika, njih preko 250, ocjenjuje značajnost inteligencije u rasponu između 15 i 20. Iz podataka vidimo da su najčešće vrijednosti za obje važnosti 17.5. Druga najčešća vrijednost važnosti inteligencije je 12.5, dok je za važnost izgleda 22.5.

```
b <- seq(0, 100, by = 5)

h3 <- hist(df_cleaned_first_male$intelligence_important, breaks = b, plot = FALSE)
h4 <- hist(df_cleaned_first_male$attractive_important, breaks = b, plot = FALSE)

midpoints <- (head(b, -1) + tail(b, -1)) / 2

data <- t(cbind(h3$counts, h4$counts))

barplot(data, beside = TRUE,
        col = c("blue", "lightblue"),
        xlab = "Importance",
        ylab = 'Frequency',
        names.arg = midpoints)
legend("topright",
       c("intelligence", "attractiveness"),
       fill = c("blue", "lightblue"))
```

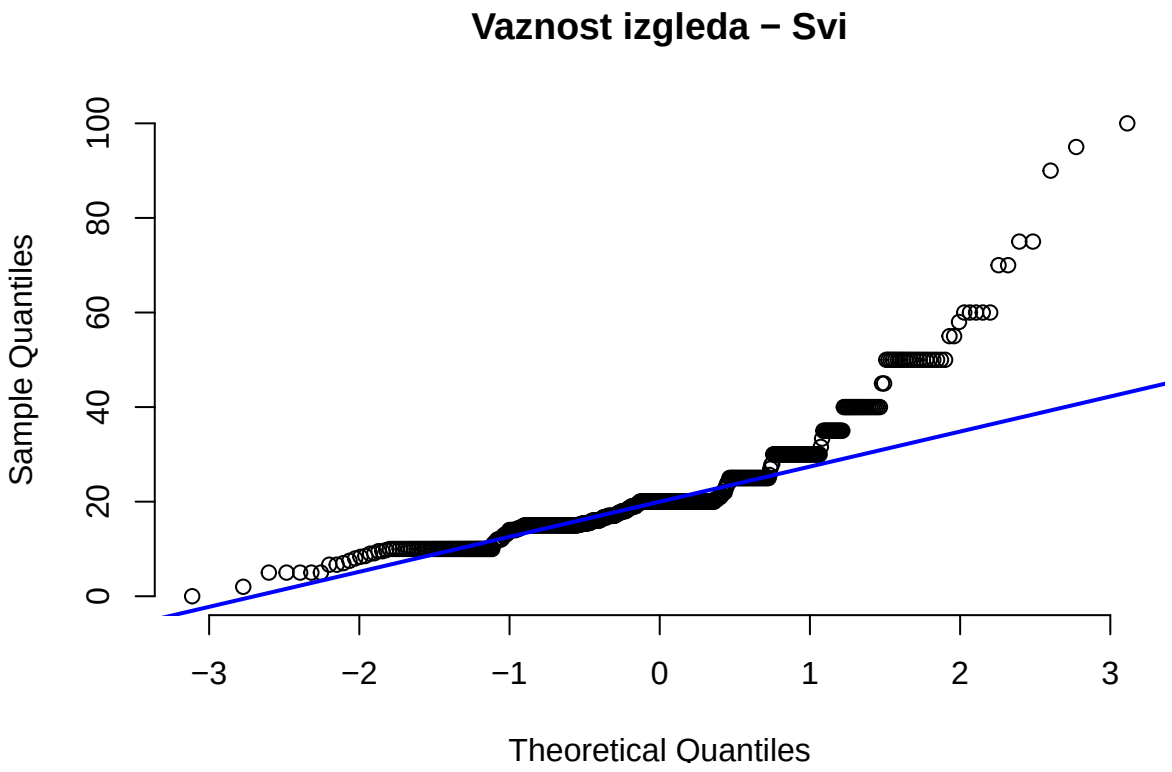



```r
b <- seq(0, 100, by = 5)

h5 <- hist(df_cleaned_first_female$intelligence_important, breaks = b, plot = FALSE)
h6 <- hist(df_cleaned_first_female$attractive_important, breaks = b, plot = FALSE)

midpoints <- (head(b, -1) + tail(b, -1)) / 2

data <- t(cbind(h5$counts, h6$counts))

barplot(data, beside = TRUE,
        col = c("blue", "lightblue"),
        xlab = "Importance",
        ylab = 'Frequency',
        names.arg = midpoints)
legend("topright",
       c("intelligence", "attractiveness"),
       fill = c("blue", "lightblue"))
```

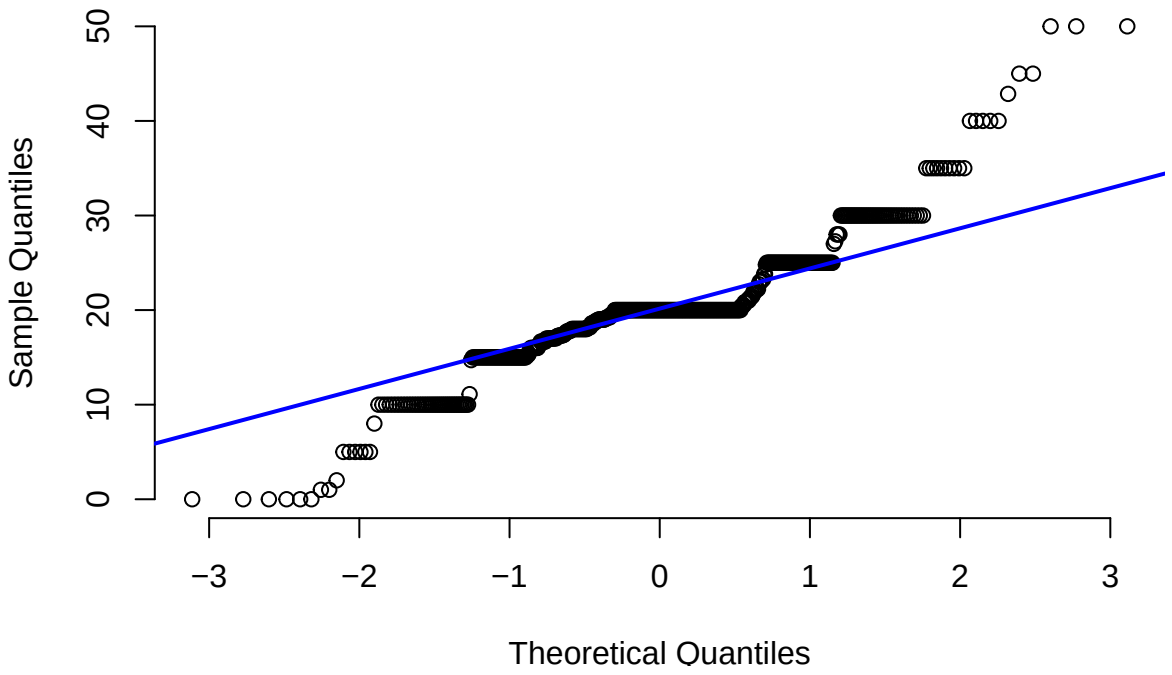


Za provjeru normalnosti podataka planiramo koristiti qq-plotove i test o jednakosti varijanci za provjeru jednakosti varijanci.

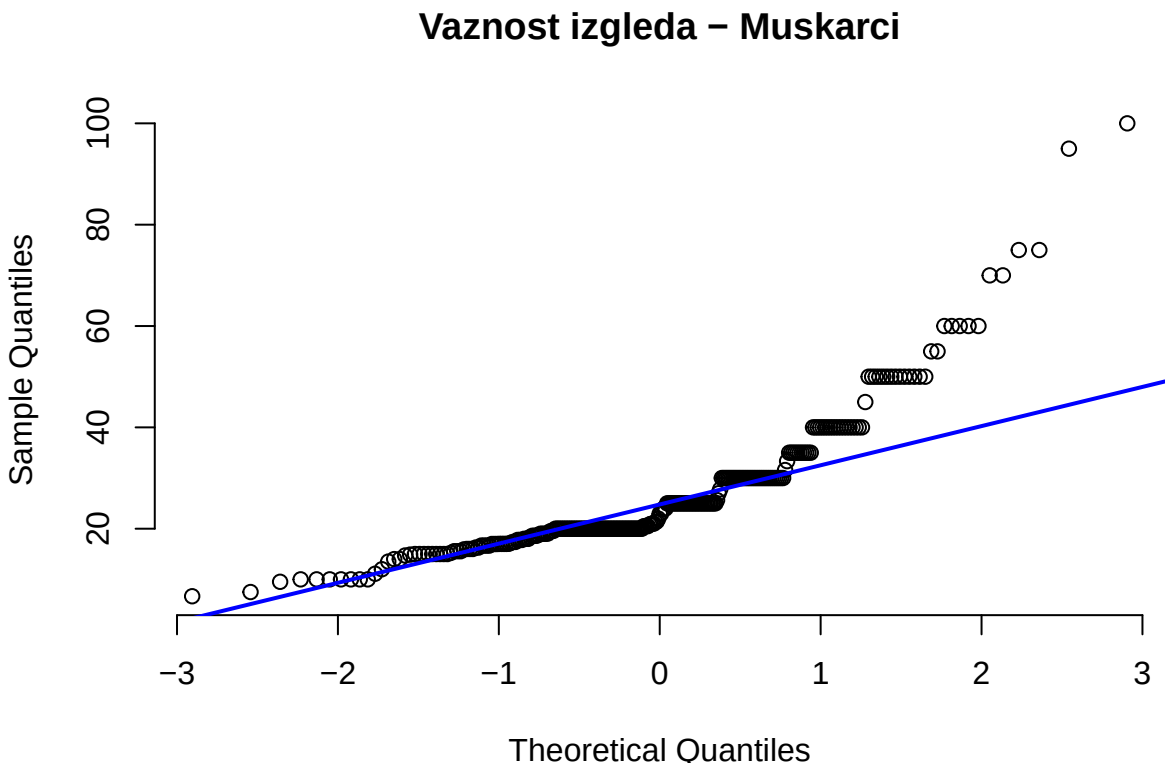
```
qqnorm(df_cleaned_first$attractive_important,
       pch = 1,
       frame = FALSE,
       main = "Važnost izgleda - Svi")
qqline(df_cleaned_first$attractive_important,
       col = "blue",
       lwd = "2")
```



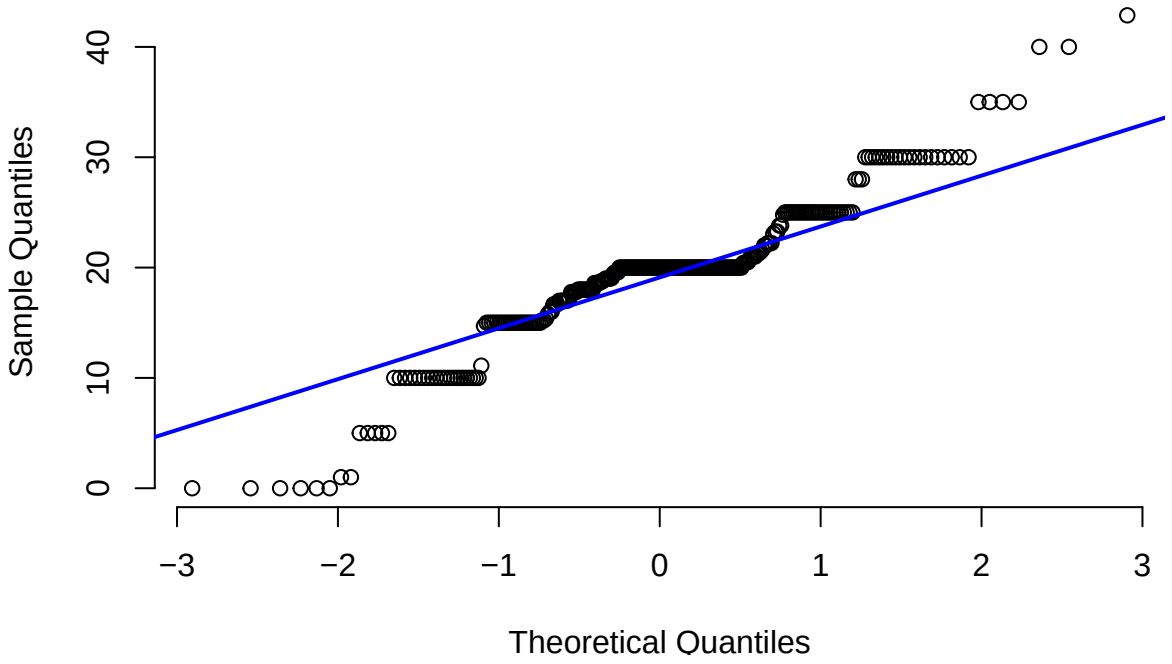

```
qqnorm(df_cleaned_first$intelligence_important,
       pch = 1,
       frame = FALSE,
       main = "Važnost inteligencije - Svi")
qqline(df_cleaned_first$intelligence_important,
       col = "blue",
       lwd = "2")
```



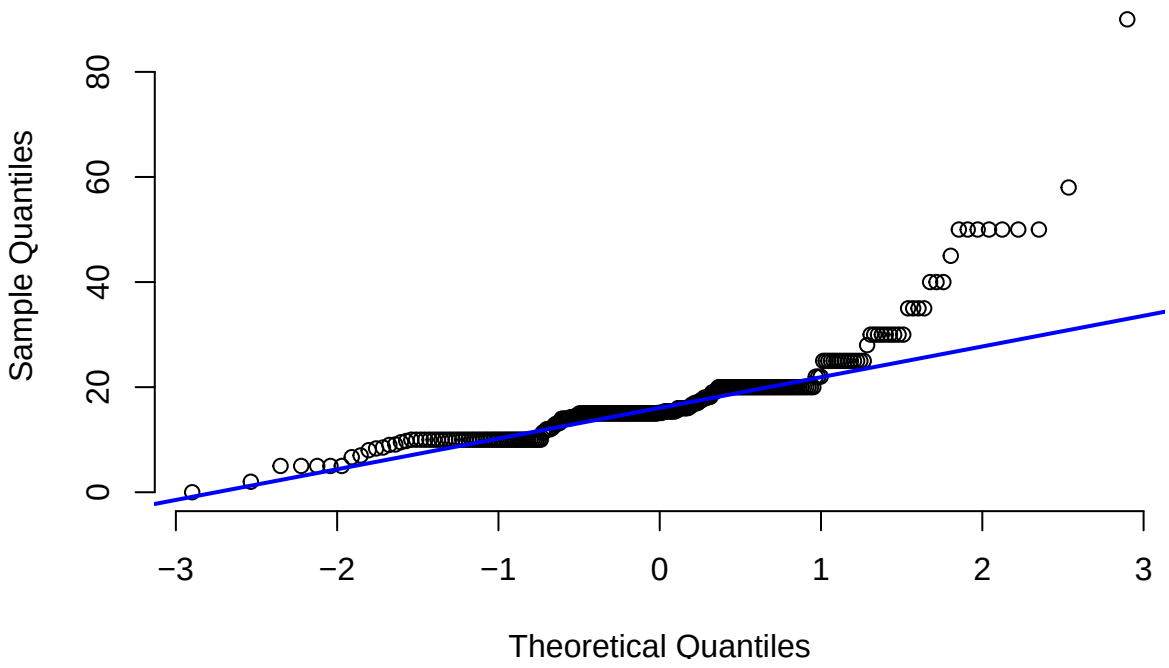

```
qqnorm(df_cleaned_first_male$attractive_important,
       pch = 1,
       frame = FALSE,
       main = "Važnost izgleda - Muškarci")
qqline(df_cleaned_first_male$attractive_important,
       col = "blue",
       lwd = "2")
```



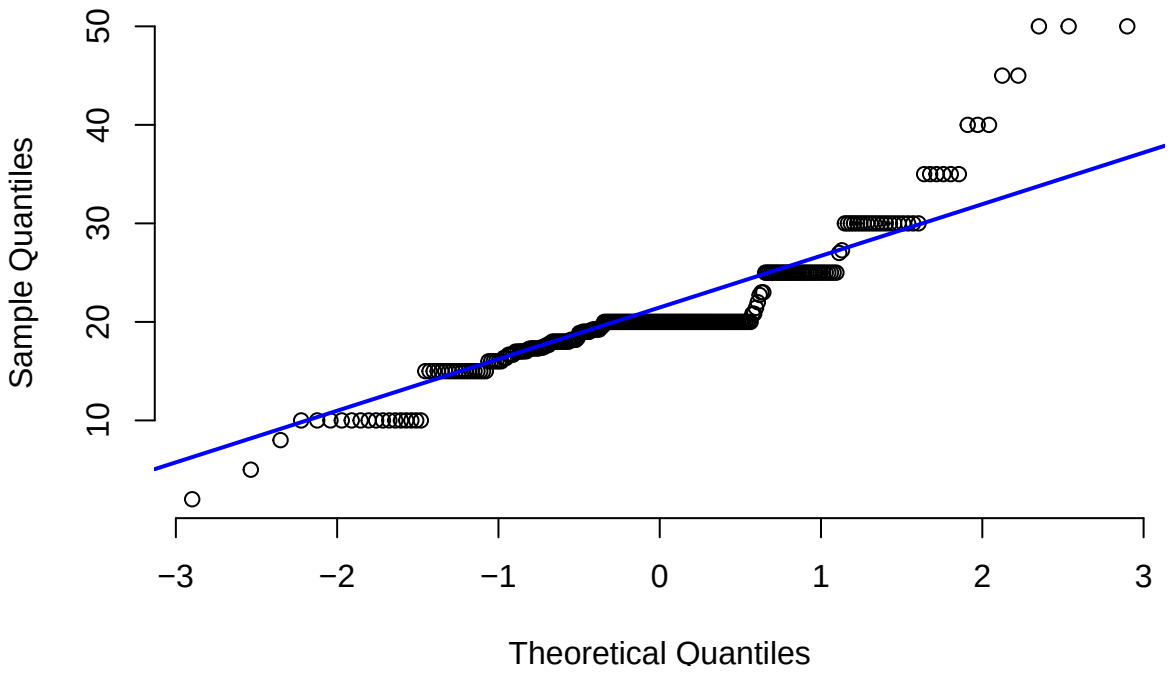

```
qqnorm(df_cleaned_first_male$intelligence_important,
       pch = 1,
       frame = FALSE,
       main = "Važnost inteligencije - Muškarci")
qqline(df_cleaned_first_male$intelligence_important,
       col = "blue",
       lwd = "2")
```




```
qqnorm(df_cleaned_first_female$attractive_important,
       pch = 1,
       frame = FALSE,
       main = "Važnost izgleda - Žene")
qqline(df_cleaned_first_female$attractive_important,
       col = "blue",
       lwd = "2")
```




```
qqnorm(df_cleaned_first_female$intelligence_important,
       pch = 1,
       frame = FALSE,
       main = "Važnost inteligencije - Žene")
qqline(df_cleaned_first_female$intelligence_important,
       col = "blue",
       lwd = "2")
```



Uz pretpostavku normalnosti podataka možemo nastaviti testiranje varijanca.

Provest ćemo test o jednakosti varijanci uz pomoć F-testa.

$$F = \frac{S_{X_1}^2/\sigma_1^2}{S_{X_2}^2/\sigma_2^2}$$

Pri ćemu za statistike s1 (intelligence_important) i s2 (attractive_important) vrijedi :

$$S_{X_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_1^i - \bar{X}_1)^2, \quad S_{X_2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_2^i - \bar{X}_2)^2.$$

Uz $(n_1 - 1, n_2 - 1)$ stupnjeva slobode. Hipoteze testa o jednakosti varijanca glase :

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Varijanca važnosti izgleda:

```
var_attractive <- var(df_cleaned_first$attractive_important)
cat("Varijanca od attractive_important za sve sudionike:", var_attractive, "\n")
```

```
## Varijanca od attractive_important za sve sudionike: 162.2717
```

```
var_attractive_male <- var(df_cleaned_first_male$attractive_important)
cat("Varijanca od attractive_important za muškarce:", var_attractive_male, "\n")
```

```
## Varijanca od attractive_important za muškarce: 185.3076
```

```
var_attractive_female <- var(df_cleaned_first_female$attractive_important)
cat("Varijanca od attractive_important za žene:", var_attractive_female, "\n")
```

```
## Varijanca od attractive_important za žene: 97.50458
```

Varijanca važnosti inteligencije:

```
var_intelligence <- var(df_cleaned_first$intelligence_important)
cat("Varijanca od intelligence_important za sve sudionike:", var_intelligence, "\n")
```

```
## Varijanca od intelligence_important za sve sudionike: 47.30045
```

```
var_intelligence_male <- var(df_cleaned_first$intelligence_important)
cat("Varijanca od intelligence_important za muškarce:", var_intelligence_male, "\n")
```

```
## Varijanca od intelligence_important za muškarce: 47.30045
```

```
var_intelligence_female <- var(df_cleaned_first$intelligence_important)
cat("Varijanca od intelligence_important za žene:", var_intelligence_female, "\n")
```

```
## Varijanca od intelligence_important za žene: 47.30045
```

Te radimo F-test:

```
var.test(df_cleaned_first$attractive_important,
         df_cleaned_first$intelligence_important)
```

```
##
##  F test to compare two variances
##
## data:  df_cleaned_first$attractive_important and df_cleaned_first$intelligence_important
## F = 3.4307, num df = 539, denom df = 539, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.897158 4.062401
## sample estimates:
## ratio of variances
##           3.430659
```

```
var.test(df_cleaned_first_male$attractive_important,
         df_cleaned_first_male$intelligence_important)
```

```
##
##  F test to compare two variances
##
## data:  df_cleaned_first_male$attractive_important and df_cleaned_first_male$intelligence_important
## F = 3.9273, num df = 272, denom df = 272, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  3.094976 4.983532
## sample estimates:
## ratio of variances
##           3.927329
```

```
var.test(df_cleaned_first_female$attractive_important,
         df_cleaned_first_female$intelligence_important)
```

```
##
##  F test to compare two variances
##
## data:  df_cleaned_first_female$attractive_important and df_cleaned_first_female$intelligence_importa
## F = 2.1004, num df = 266, denom df = 266, p-value = 2.248e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
```

```
##  1.650777 2.672383
## sample estimates:
## ratio of variances
##           2.100359
```

Zaključujemo uz p-vrijednosti koje su manje od predodređene razine značajnosti od 0.05 da su te varijance različite. Odbacujemo H0 hipotezu u koristi H1 te ćemo koristiti T-test za dva uzorka uz nejednake varijance.

T-test za dva uzorka uz nepoznate i nejednake varijance

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2 \quad , \quad \mu_1 > \mu_2 \quad , \quad \mu_1 \neq \mu_2$$

```
t_test_result <- t.test(df_cleaned_first$intelligence_important,
                        df_cleaned_first$attractive_important,
                        var.equal = FALSE)

print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  df_cleaned_first$intelligence_important and df_cleaned_first$attractive_important
## t = -3.7697, df = 828.62, p-value = 0.000175
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.571237 -1.125652
## sample estimates:
## mean of x mean of y
##  20.17174  22.52019
```

```
t_test_result_male <- t.test(df_cleaned_first_male$intelligence_important,
                        df_cleaned_first_male$attractive_important,
                        var.equal = FALSE)

print(t_test_result_male)
```

```
##
##  Welch Two Sample t-test
##
## data:  df_cleaned_first_male$intelligence_important and df_cleaned_first_male$attractive_important
## t = -8.2239, df = 402.08, p-value = 2.754e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.403482 -5.775126
## sample estimates:
## mean of x mean of y
##  19.41956  27.00886
```

```
t_test_result_female <- t.test(df_cleaned_first_female$intelligence_important,
                        df_cleaned_first_female$attractive_important,
                        var.equal = FALSE)

print(t_test_result_female)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  df_cleaned_first_female$intelligence_important and df_cleaned_first_female$attractive_importa
## t = 4.0999, df = 472.48, p-value = 4.865e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.567480 4.452894
## sample estimates:
## mean of x mean of y
##  20.94082  17.93064
```

Zaključak

Uz male p-vrijednosti odbacujemo hipotezu H0 u korist H1 za sva tri testa - za sve sudionike, za muškarce i za žene. Na temelju ovih testova zaključujemo da kod muškaraca postoji veća razlika između važnosti izgleda i inteligencije u korist izgleda, te se na temelju testa zaključuje da muškim ispitanicima važnost inteligencije nije veća od važnosti izgleda. Kod žena u ovom skupu podataka situacija je suprotna, također nisu jednako važni inteligencija i izgled, ali u ovom slučaju prednost ima inteligencija partnera to jest drugim riječima, izgled ima manju ulogu.

Kod svih ispitanika zaključujemo da važnost inteligencije i izgleda kod odabira partnera nisu jednaki, izgled je važniji što smo pokazali i ovim testom. Paralelnim prikazom međuodnosa inteligencije i izgleda kod muškaraca i žena prikazali smo da u slučaju drukčijeg skupa podataka u kojem je veći disbalans između broja muškaraca i žena (ovdje, 273 i 267) mogli doći do drukčijeg zaključka za sve ispitanike.

2. Postoji li razlika u interesu za gaming prema zanimanju sudionika?

Za rješavanje ovog pitanja, ključni su nam podaci o zanimanju sudionika i njihovom interesu za igranje video igara. Prije nego što krenemo s analizom, potrebno je očistiti podatke kako bismo ih pripremili za daljnje korištenje.

Prvo, fokusirat ćemo se na podatke o interesu za video igre. Nepostojeće ocjene zamijenit ćemo s medijanom svih ocjena svih sudionika, osiguravajući time konzistentnost i relevantnost podataka.

Zatim, obratit ćemo pažnju na podatke o zanimanju sudionika. Uklonit ćemo sudionike čije zanimanje nije definirano.

Kako bismo pojednostavnili ovaj proces, prethodno ćemo kreirati novi skup podataka koji će sadržavati samo relevantne stupce iz tablice "participants" - gaming i field stupce.

```
df_field_gaming <- participants[, c("gaming", "field")]
names(df_field_gaming) # ispisuje imena stupaca
```

```
## [1] "gaming" "field"
```

```
dim(df_field_gaming) # dimenzije novo stvorenog data frame-a participants_data
```

```
## [1] 540   2
```

```
class(df_field_gaming$gaming)
```

```
## [1] "integer"
```

```
class(df_field_gaming$field)
```

```
## [1] "character"
```

Ovaj ispis otkriva da naš skup podataka sadrži 540 sudionika s informacijama o njihovom interesu za igranje video igara i zanimanju. Proučavamo dva ključna stupca: "gaming" koji sadrži cijele brojeve na intervalnoj

skali te "field" koji predstavlja kategorijsku varijablu s informacijama o zanimanju sudionika. Ova analiza pomaže nam razumjeti prirodu podataka prije daljnje obrade i analize.

```
summary(df_field_gaming$gaming)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   3.000   3.874   6.000  14.000
```

```
sum(df_field_gaming$gaming > 10)
```

```
## [1] 5
```

```
sum(is.na(df_field_gaming$gaming)) # nepostojeće vrijednosti za gaming
```

```
## [1] 0
```

```
sum(is.na(df_field_gaming$field)) # nepostojeće vrijenosti za field
```

```
## [1] 0
```

```
length(unique(df_field_gaming$field)) # različite vrijednosti u field
```

```
## [1] 252
```

```
df_field_gaming$gaming[df_field_gaming$gaming > 10] <- 10
df_field_gaming$gaming[df_field_gaming$gaming < 1] <- 1
```

Za naše pitanje odnosa interesa za gaming i zanimanja sudionika relevantne su kategorijska varijabla zanimanja i numerička varijabla ocjene na intervalnoj skali. Iz tog razloga daljnjoj analizi pristupit ćemo na dva različita načina:

1. način Kategorizacija ocjena za gaming kako bismo mogli testirati nezavisnost između dvije kategorijske varijable. Test nezavisnosti u programskom jeziku R kao ulaz prima kontigencijsku tablicu te treba imati na umu da je pretpostavka testa da očekivana frekvencija pojedinog razreda mora biti veća ili jednaka 5.
2. način Provjera primjenom jednofaktorske ANOVE, metoda kojom testiramo sredine više populacija te joj je cilj odrediti jesu li razlike između grupa statistički značajne.

Iz ispisa ocjena za gaming zaključujemo da ima 5 vrijednosti ocjena koje su veće od 10, 2 vrijednosti koje su manje od 1 te da nema nepostojećih vrijednosti. U procesu čišćenja podataka smo vrijednosti koje su veće od 10 zamijenili s 10, a vrijednosti koje su manje od 1 zamijeniti s 1. Također nema nepostojećih vrijednosti u stupcu field te stoga možemo koristiti ovaj data frame u daljnjoj analizi.

Za oba navedena načina potrebno je dodatno kategorizirati zanimanja sudionika.

```
levels_field <- levels(factor(df_field_gaming$field))
print(levels_field[1:20])
```

```
##  [1] "Acting"                              "African-American Studies/History"
##  [3] "American Studies"                    "American Studies [Masters]"
##  [5] "anthropology"                        "Anthropology"
##  [7] "Anthropology/Education"              "Applied Maths/Econs"
##  [9] "Applied Physiology & Nutrition"      "Architecture"
## [11] "art education"                       "Art Education"
## [13] "art history"                         "Art History"
## [15] "Art History/medicine"                "Arts Administration"
## [17] "bilingual education"                 "Bilingual Education"
## [19] "Biochemistry"                        "Biochemistry & Molecular Biophysics"
```

```r
length(unique(df_field_gaming$field))
```

```
## [1] 252
```

Ovdje smo ispisali samo dio različitih vrijednosti u stupcu field kojih ima 252. Kao što smo napomenuli, zanimanje sudionika je kategorijska varijabla i mnoge od tih vrijednosti imaju zajedničke karakteristike koje ih svrstavaju u istu kategoriju. Prije nastavka analize potrebno je zanimanja sudionika grupirati u manji broj kategorija. U procesu grupiranja sva srodna zanimanja ćemo staviti u istu kategoriju tako da kategorije budu nezavisne.

Postoje mnogi načini za svrstavanje zanimanja sudionika, u ovom slučaju na različite podskupove. Cilj podjele na manji broj kategorija bio je dobiti nezavisne podskupove zanimanja sa zajedničkim karakteristikama kako bi kategorije bile relevantne za predmet našeg istraživanja - razlika u interesu za gaming prema zanimanju korisnika. Osim koherencije kategorija cilj je bio i smanjenje dimenzionalnosti uz, ako je moguće, održavanje podjednake raspodjele sudionika po kategorijama. Ovim smjernicama stvarali smo kategorije srodnih zanimanja te smo metodom pokušaja i pogreške došli do sljedeće kategorizacije - Business&Economics, STEM, Social Studies&Humanities, Law&International Affairs.

```r
bussines_economics_keywords <- c("economy", "economics", "econs",
                                 "finance", "finanace", "financial",
                                 "business", "consulting", "fundraising",
                                 "marketing", "mba", "money")

stem_keywords <- c("architecture", "engineering", "operations research",
                   "engg.", "computer science", "science", "scientific",
                   "chemistry", "biology", "biotechnology", "biotechnology",
                   "biochemsistry", "epidemiology", "genetics", "physics",
                   "marine geophysics", "mathematics", "math", "statistics",
                   "stats", "applied maths", "climate", "ecology",
                   "urban planning", "nutrition", "physiology",
                   "nutritiron", "health", "medical informatics",
                   "speech", "medicine", "biomedicine", "biomedical")

social_studies_humanities_keyword <- c("american studies", "history",
                                       "anthropology", "qmss", "psychology",
                                       "sociology", "social", "ma in quantitative methods",
                                       "philosophy", "religion", "literature",
                                       "english", "french", "poetry", "writing",
                                       "polish", "classics", "education",
                                       "teaching", "instructional ", "tesol",
                                       "art", "arts", "acting", "film",
                                       "theater", "theatre", "communications",
                                       "journalism")

law_keywords <- c("law", "human rights",
                  "international", "intrernational affairs",
                  "public", "relations")

categorize_field <- function(field) {
  matches <- c(
    any(grepl(paste(bussines_economics_keywords, collapse = "|"), field, ignore.case = TRUE)),
    any(grepl(paste(stem_keywords, collapse = "|"), field, ignore.case = TRUE)),
    any(grepl(paste(social_studies_humanities_keyword, collapse = "|"), field, ignore.case = TRUE)),
    any(grepl(paste(law_keywords, collapse = "|"), field, ignore.case = TRUE))
  )
```

```
  if (sum(matches) > 1) {
    return(field)
  } else {
    if (matches[1]) return("Business&Economics")
    if (matches[2]) return("STEM")
    if (matches[3]) return("Social Studies&Humanities")
    if (matches[4]) return("Law & International Affairs")
    return(field)
  }
}

df_field_gaming$field = sapply(df_field_gaming$field, categorize_field)
```

```
zadane_kategorije <- c("Business&Economics",
                       "Law & International Affairs",
                       "STEM", "Social Studies&Humanities")

redovi_bez_zadanih_kategorija <- subset(df_field_gaming, !(field %in% zadane_kategorije))

head(redovi_bez_zadanih_kategorija)
```

```
##    gaming                                field
## 33      4                  Applied Maths/Econs
## 47      5                       Undergrad - GS
## 48      8                  Mathematical Finance
## 63      6      Business & International Affairs
## 89      2 International Educational Development
## 94      1    Climate-Earth and Environ. Science
```

Pregledavanjem nesvrstanih zanimanja koja se ne mogu dodati u kategorije kako bi se zadržala nezavisnost, za neka smo zaključili da se mogu uključiti.Također postoje i podaci koji nisu iskoristivi jer ne daju zanimanje sudionika. Sve navedene podatke smo očistili i neće se koristit u testovima.

```
if ("Climate-Earth and Environ. Science" %in% df_field_gaming$field) {
  df_field_gaming$field[participants$field == "Climate-Earth and Environ. Science"] <- "STEM"
}
if ("international finance and business" %in% df_field_gaming$field) {
  df_field_gaming$field[participants$field == "international finance and business"] <- "Business&Econom
}
if ("International Business" %in% df_field_gaming$field) {
  df_field_gaming$field[participants$field == "International Business"] <- "Business&Economics"
}
if ("International Finance" %in% df_field_gaming$field) {
  df_field_gaming$field[participants$field == "International Finance"] <- "Business&Economics"
}
if ("Financial Engineering" %in% df_field_gaming$field) {
  df_field_gaming$field[participants$field == "Financial Engineering"] <- "Business&Economics"
}
if ("Law and Social Work" %in% df_field_gaming$field) {
  df_field_gaming$field[participants$field == "Law and Social Work"] <- "Law & International Affairs"
}
if ("Public Health" %in% df_field_gaming$field) {
  df_field_gaming$field[participants$field == "Public Health"] <- "STEM"
}
```

```
df_field_gaming <- subset(df_field_gaming, field %in% zadane_kategorije)
```

```
table(df_field_gaming$field)
```

```
##
##          Business&Economics Law & International Affairs
##                         120                          81
##   Social Studies&Humanities                        STEM
##                         163                         144
```

```
length(unique(df_field_gaming$field))
```

```
## [1] 4
```

```
dim(df_field_gaming)
```

```
## [1] 508   2
```

Nakon kategorizacije zanimanja sudionika imamo četiri kategorije zanimanja: - Business&Economics, - Law & International Affairs, - Social Studies&Humanities, - STEM U nastavku slijede mjere centralne tendencije za navedena zanimanja te vizualizacija podataka, također i za gaming.

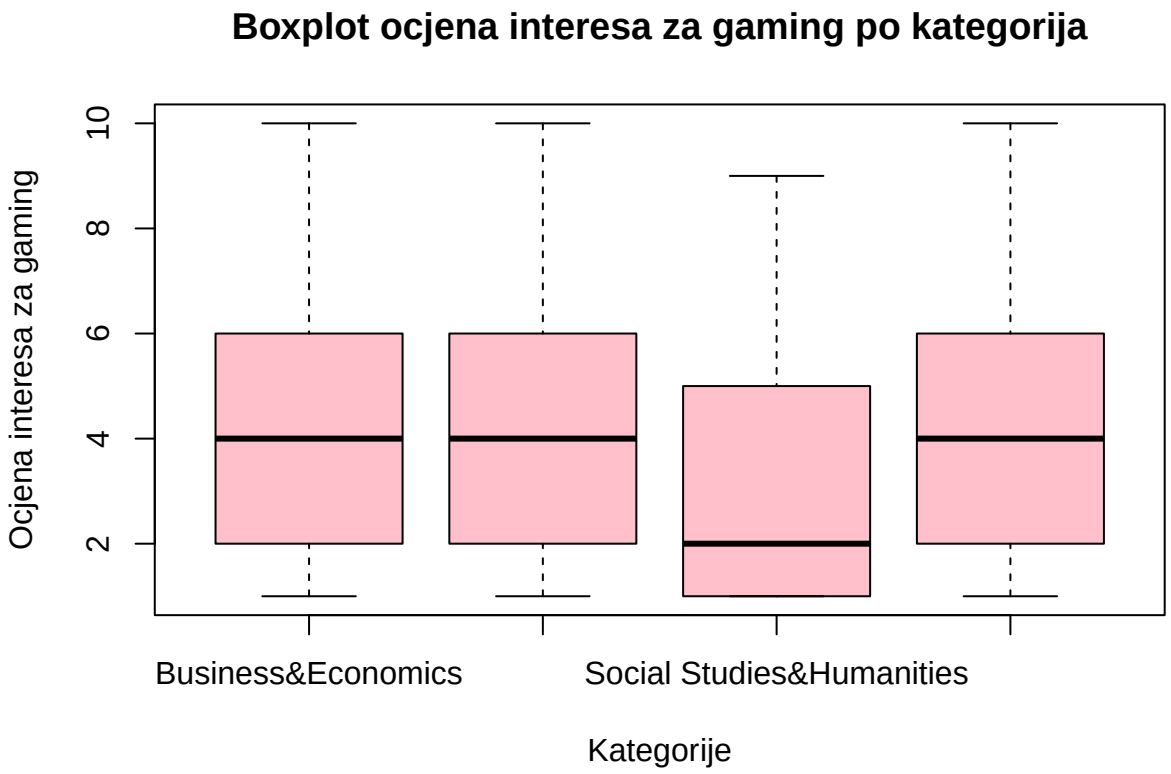
Vizualizacija i prikaz podataka

```
aggregate(df_field_gaming["gaming"], list(df_field_gaming$field), mean)
```

```
##                       Group.1   gaming
## 1          Business&Economics 4.125000
## 2 Law & International Affairs 4.197531
## 3   Social Studies&Humanities 3.263804
## 4                        STEM 4.000000
```

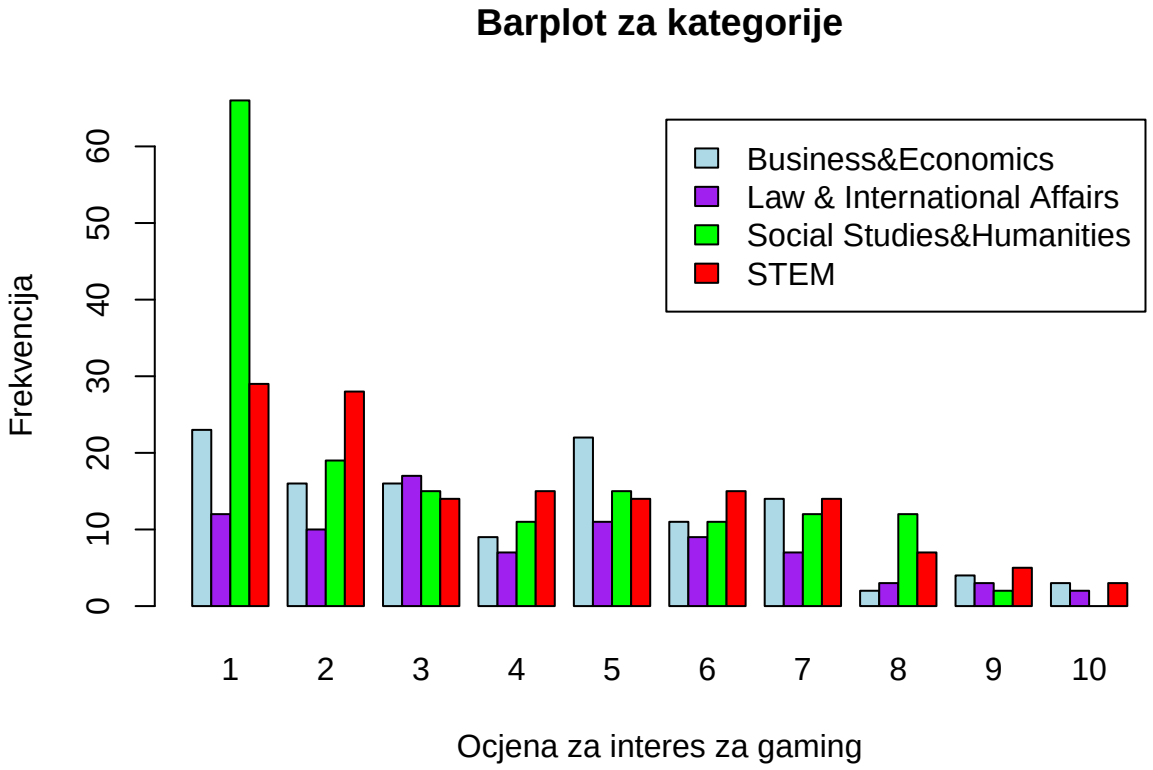
Na temeljuaritmetičke sredine ocjena za gaming grupiranih po prije određenim kategorijama na prvi pogled možemo primijetiti da je vrijednost aritmetičke sredine manja za jednu ocjenu.

```
#boxplot(df_field_gaming$gaming ~ df_field_gaming$field)

boxplot(gaming ~ field, df_field_gaming,
        main = "Boxplot ocjena interesa za gaming po kategorija",
        xlab = "Kategorije",
        ylab = "Ocjena interesa za gaming",
        col = "pink"  # Boja boxplot-a
)
```



Manja vrijednost medijana vidljiva je i na prikazanom boxplotu ocjena interesa za gaming po kategorijama.

```
barplot(
  t(table(df_field_gaming$gaming, df_field_gaming$field)),
  col = c('lightblue', 'purple', 'green', 'red'),
  beside = TRUE,
  legend.text = TRUE,
  xlab = 'Ocjena za interes za gaming',
  ylab = 'Frekvencija',
  main = 'Barplot za kategorije'
)
```



Koristeći različite vrste prikaza podataka za vrstu kojom raspolažemo, kategorijska varijabla područja zanimanja i numerička varijabla ocjene, prikazali smo raspodjelu podataka po kategorijama i ocjenama unutar svake kategorije te međusobno između kategorija.

Nakon vizualizacije podataka nastavit ćemo s prije spomenutim prvim načinom pristupa ovom pitanju odnosa interesa za gaming prema zanimanju sudionika. To podrazumijeva kategorizaciju ocjena za gaming kako bismo testirali nezavisnost između dvije tada kategorijske varijable - područje zanimanje, ocjena interesa za gaming. Kategorija ocjene varijabla je na ordinalnoj skali dok je kategorija područja zanimanja kategorijska varijabla nominalne skale - područja zanimanja ne mogu se rangirati dok se kategorije ocjena mogu.

Test nezavisnosti kategorijskih varijabli

Test nezavisnosti, χ^2 test, u programskom paketu R implementiran je u funkciji `chisq.test()` koja kao ulaz prima kontingencijsku tablicu podataka koje testiramo na nezavisnost. Također,test nezavisnosti koristi se kako bi se provjerilo postoji li statistički značajna veza između dvije kategorijske varijable, u našem slučaju kategorije ocjena i područja zanimanja. Stoga, hipoteze za navedeni test su sljedeće: Nulta hipoteza - nema statistički značajne veze između zanimanja sudionika i interesa za gaming.

H_0 : Područje zanimanja sudionika i interes za gaming su nezavisne varijable.

Alternativna hipoteza - Postoji statistički značajna veza između zanimanja sudionika i interesa za gaming.

H_1 : Područje animanja sudionika i interes za gaming nisu nezavisne varijable..

Ako p-vrijednost dobivena testom bude dovoljno mala - manja od 0.05, odbit ćemo nultu hipotezu i zaključiti da postoji statistički značajna veza.

Test nezavisnosti u jeziku R prima kontingencijsku tablicu kao ulaz te treba imati na umu pretpostavku da je frekvencija očekivanog razreda veća ili jednaka 5.

Prvi korak za navedeno ispitivanje je kategoriziranje ocjena interesa za gaming. To ćemo napraviti na sljedeći način, ocjene koje su cijeli brojevi od 1 do 10 ćemo podijeliti u tri nezavisne kategorije: Niska, Srednja i

Visoka. Gdje će u grupi "Niska" biti ocjene od 1 do 2, u grupi "Srednja ocjene" od 3 do 6 i u grupi "Visoka ocjene" od 7 do 10.

```
df_field_gaming <- df_field_gaming %>%
  mutate(kat_gaming = case_when(
    gaming %in% 0:2 ~ "Niska",
    gaming %in% 3:6 ~ "Srednja",
    TRUE ~ "Visoka"
  ))
```

```
kat_gaming_count <- table(df_field_gaming$kat_gaming)
print(kat_gaming_count)
```

```
##
##   Niska Srednja  Visoka
##     203     212      93
```

Nakon kategorizacije ocjena interesa za gaming možemo nastaviti s prikazom kontingencijske tablice, zatim dodavanjem marginalnih suma i provjerom pretpostavke testa nezavisnosti.

Kontingencijska tablica

```
kontigencijska_tabl <- table(df_field_gaming$field, df_field_gaming$kat_gaming)
margine_tabl = addmargins(kontigencijska_tabl)
print(margine_tabl)
```

```
##
##                               Niska Srednja Visoka Sum
##   Business&Economics             39      58     23 120
##   Law & International Affairs    22      44     15  81
##   Social Studies&Humanities      85      52     26 163
##   STEM                           57      58     29 144
##   Sum                           203     212     93 508
```

Provjera pretpostavke testa nezavisnosti - očekivane frekvencije moraju biti veće ili jednake 5.

```
for (col_names in colnames(margine_tabl)){
  for (row_names in rownames(margine_tabl)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za razred ',col_names,'-',row_names,': ',(margine_tabl[row_names,'Sum'
    }
  }
}
```

```
## Očekivane frekvencije za razred  Niska - Business&Economics :  47.95276
## Očekivane frekvencije za razred  Niska - Law & International Affairs :  32.36811
## Očekivane frekvencije za razred  Niska - Social Studies&Humanities :  65.13583
## Očekivane frekvencije za razred  Niska - STEM :  57.54331
## Očekivane frekvencije za razred  Srednja - Business&Economics :  50.07874
## Očekivane frekvencije za razred  Srednja - Law & International Affairs :  33.80315
## Očekivane frekvencije za razred  Srednja - Social Studies&Humanities :  68.02362
## Očekivane frekvencije za razred  Srednja - STEM :  60.09449
## Očekivane frekvencije za razred  Visoka - Business&Economics :  21.9685
## Očekivane frekvencije za razred  Visoka - Law & International Affairs :  14.82874
## Očekivane frekvencije za razred  Visoka - Social Studies&Humanities :  29.84055
## Očekivane frekvencije za razred  Visoka - STEM :  26.3622
```

Sve očekivane frekvencije su veće od 5. Možemo nastaviti sa χ^2 testom.

```r
chisq.test(kontigencijska_tabl,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  kontigencijska_tabl
## X-squared = 20.041, df = 6, p-value = 0.002724
```

P-vrijednost manje je od prije definirane vrijednosti 0.05, ona iznosi 0.002724 te iz tog razloga odbacujemo nultu hipotezu u korist alternativne. Postoji statistički značajna veza između zanimanja sudionika i interesa za gaming to jest područje zanimanja sudionika i njegov interes za gaming zavisne su varijable.

Provjera primjenom jednofaktorske ANOVE

Sada ćemo nastaviti s drugim pristupom ovom pitanju, primjenom jednofaktorske ANOVE gdje ne gledamo ocjene kao kategorijsku varijablu nego numeričku.

ANOVA (engl. *ANalysis Of VAriance*) metoda je kojom testiramo sredine više populacija, a koja pretpostavlja da je ukupna varijabilnost u podatcima posljedica varijabilnosti podataka unutar svake pojedine grupe (populacije) i varijabilnosti između različitih grupa. Varijabilnost unutar pojedinog uzorka je rezultat slučajnosti, a ako postoje razlike u sredinama populacija, one će biti odražene u varijabilnosti među grupama. Jedan od glavnih ciljeva analize varijance je ustanoviti jesu li upravo te razlike između grupa samo posljedica slučajnosti ili su statistički značajne.

U jednofaktorskom ANOVA modelu razmatra se utjecaj jednog faktora koji ima k razina. Neka su:

$$\begin{aligned} X_{11}, X_{12}, \ldots, X_{1n_1} &\sim N(\mu_1, \sigma^2) \\ X_{21}, X_{22}, \ldots, X_{2n_2} &\sim N(\mu_2, \sigma^2) \\ &\vdots \\ X_{k1}, X_{k2}, \ldots, X_{kn_k} &\sim N(\mu_k, \sigma^2) \end{aligned}$$

nezavisni uzorci iz k različitih populacija (populacije se razlikuju upravo po razini faktora od interesa). Jednofaktorski ANOVA model glasi:

$$X_{ij} = \mu_i + \epsilon_{ij},$$

gdje je μ_i sredina svake populacije $i = 1, .., k$. Analizom varijance testiramo:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \ldots = \mu_k \\ H_1 &: \text{barem dvije sredine nisu iste.} \end{aligned}$$

Kako bi krenuli s analizom ANOVE moramo ispitati njene pretpostavke, a to su: 1. nezavisnost pojedinih podataka u uzorcima 2. normalna razdioba podataka 3. homogenost varijanci među populacijama

Nezavisnost podataka u populacijama smo osigurali kategorizacijom područja zanimanja iz tog razloga je taj uvjet ispunjen. Nastavljamo s uvjetom homogenosti varijanci među populacijama. Ova pretpostavka je važna kako bi ANOVA bila pouzdana, a njezino kršenje može dovesti do netočnih rezultata. Što se tiče homogenosti varijanci različitih populacija, potrebno je testirati:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 \\ H_1 &: \text{barem dvije varijance nisu iste.} \end{aligned}$$

Navedenu hipotezu možemo testirati Bartlettovim testom. Bartlettov test u R-u implementiran je naredbom `bartlett.test()`.

```r
# Testiranje homogenosti varijance uzoraka Bartlettovim testom
bartlett.test(df_field_gaming$gaming ~ df_field_gaming$field)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  df_field_gaming$gaming by df_field_gaming$field
## Bartlett's K-squared = 0.3714, df = 3, p-value = 0.9461
```

```r
var((df_field_gaming$gaming[df_field_gaming$field=='Business&Economics']))
```

```
## [1] 6.026261
```

```r
var((df_field_gaming$gaming[df_field_gaming$field=='Law & International Affairs']))
```

```
## [1] 5.835494
```

```r
var((df_field_gaming$gaming[df_field_gaming$field=='Social Studies&Humanities']))
```

```
## [1] 6.207756
```

```r
var((df_field_gaming$gaming[df_field_gaming$field=='STEM']))
```

```
## [1] 6.517483
```

Dobivena p-vrijednost Barlettovim testom, 0.9461, ukazuje na to da nema dovoljno statističkih dokaza da varijance nisu jednake. Stoga, zaključujemo i da nema dovoljno dokaza da su varijance različite među populacijama koje predstavljaju različita područja zanimanja sudionika. Ovo podržava homogenost varijanci - na temelju rezultata ne odbacujemo nultu hipotezu.

Potrebno je još provjeriti pretpostavku normalnosti. Provjera normalnosti može se za svaku pojedinu grupu napraviti Kolmogorov-Smirnov testom ili Lillieforsovom inačicom Kolmogorov-Smirnov testa.

```r
require(nortest)
```

```
## Loading required package: nortest
```

```r
lillie.test(df_field_gaming$gaming)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_field_gaming$gaming
## D = 0.16581, p-value < 2.2e-16
```

```r
lillie.test(df_field_gaming$gaming[df_field_gaming$field=='Business&Economics'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_field_gaming$gaming[df_field_gaming$field == "Business&Economics"]
## D = 0.13496, p-value = 1.339e-05
```

```r
lillie.test(df_field_gaming$gaming[df_field_gaming$field=='Law & International Affairs'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_field_gaming$gaming[df_field_gaming$field == "Law & International Affairs"]
## D = 0.17144, p-value = 3.51e-06
```

```
lillie.test(df_field_gaming$gaming[df_field_gaming$field=='Social Studies&Humanities'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_field_gaming$gaming[df_field_gaming$field == "Social Studies&Humanities"]
## D = 0.22313, p-value < 2.2e-16
```

```
lillie.test(df_field_gaming$gaming[df_field_gaming$field=='STEM'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_field_gaming$gaming[df_field_gaming$field == "STEM"]
## D = 0.17914, p-value = 2.896e-12
```

Testiranjem normalnosti zaključujemo da nije zadovoljen uvjet normalnosti jer je p-vrijednost testa manja od odabrane razine značajnosti, 0.05. Kao sljedeći korak pokušat ćemo transformirati podatke kako bismo ih približili normalnoj distribuciji - koristit više transformacijskih tehnika te odabrati najbolju. Ovdje je prikazana logaritamska transformacija. Nakon transformacije ćemo provesti test normalnosti.

```
df_field_gaming$log_gaming <- log(df_field_gaming$gaming)

require(nortest)
lillie.test(df_field_gaming$log_gaming)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_field_gaming$log_gaming
## D = 0.1776, p-value < 2.2e-16
```

```
lillie.test(df_field_gaming$log_gaming[df_field_gaming$field=='Business&Economics'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_field_gaming$log_gaming[df_field_gaming$field == "Business&Economics"]
## D = 0.18357, p-value = 1.077e-10
```

```
lillie.test(df_field_gaming$log_gaming[df_field_gaming$field=='Law & International Affairs'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_field_gaming$log_gaming[df_field_gaming$field == "Law & International Affairs"]
## D = 0.14753, p-value = 0.0001609
```

```
lillie.test(df_field_gaming$log_gaming[df_field_gaming$field=='Social Studies&Humanities'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_field_gaming$log_gaming[df_field_gaming$field == "Social Studies&Humanities"]
## D = 0.26234, p-value < 2.2e-16
```

```
lillie.test(df_field_gaming$log_gaming[df_field_gaming$field=='STEM'])
```

```
##
```

```
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df_field_gaming$log_gaming[df_field_gaming$field == "STEM"]
## D = 0.13994, p-value = 2.756e-07
```

Eksperimentiranjem s transformacijskim tehnika nismo uspjeli transformirati ocjene interesa za gaming kako bismo potvrdili pretpostavku normalnosti podataka. Iz tog razloga ne možemo nastaviti s provedbom ANOVA testa jer nisu ispunjene sve pretpostavke.

KRUSKAL-WALLISOV TEST

Nastavit ćemo s neparametarskom alternativom jednofaktorske analize varijance koja se koristi kada nisu zadovoljene pretpostavke parametarske anove kao u našem slučaju.

Hipotete Kruskal-Wallisovog testa su:

$$H_0 : M_1 = M_2 = \ldots = M_k$$
$$H_1 : \text{barem dva medijana nisu ista.}$$

M_k su medijani uzoraka

```
require(nortest)
kruskal.test(df_field_gaming$gaming ~ df_field_gaming$field, data = df_field_gaming)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df_field_gaming$gaming by df_field_gaming$field
## Kruskal-Wallis chi-squared = 15.92, df = 3, p-value = 0.001177
```

Testiranjem smo dobili da je p-vrijednost 0.001117 te zaključujemo da možemo odbaciti nultu hipotezu u korist alternative - barem dva medijana nisu jednaka.

Rezultat Kruskal-Wallisovog testa vidljiv je također na prije prikazanom box plotu gdje se vidi razlika u medijanu između kategorija. U drugom načinu pristupa pitanju postoji li razlika za interes za gaming prema području zanimanja sudionika došli smo do zaključka da razlika postoji.

Drugim riječima, postoji statistički značajna razlika između ocjena interesa za gaming prema području zanimanja sudionika.

3.Možemo li temeljem drugih varijabli predvidjeti hoće li se sudioniku svidjeti partner?

Prije same analize podataka potrebno je očistiti skup podataka odnosno ukloniti sve nepostojeće vrijednosti.

```
na_count_per_column <- sapply(dates, function(x) sum(is.na(x)))
data.frame(Column = names(na_count_per_column), NA_Count = na_count_per_column)
```

```
##                                      Column NA_Count
## date_id                             date_id        0
## participant_id                 participant_id        0
## partner_id                         partner_id        0
## attractive_o                     attractive_o      176
## sincere_o                           sincere_o      247
## intelligence_o                 intelligence_o      268
## funny_o                               funny_o      321
## ambitious_o                       ambitious_o      678
## shared_interests_o           shared_interests_o     1029
```

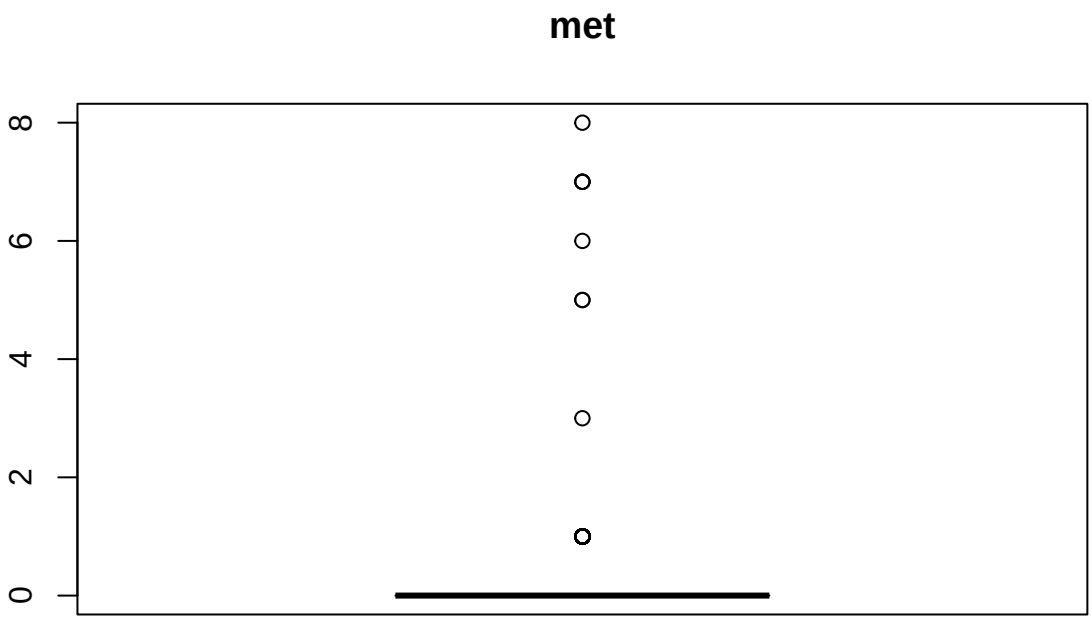
```
## attractive_partner             attractive_partner      176
## sincere_partner                   sincere_partner      247
## intelligence_partner         intelligence_partner      268
## funny_partner                       funny_partner      321
## ambition_partner                 ambition_partner      678
## shared_interests_partner shared_interests_partner     1029
## like                                         like      213
## guess_prob_liked                 guess_prob_liked      278
## met                                           met      343
## decision                                 decision        0
```

```
columns_of_interest <- c("attractive_o", "sincere_o", "intelligence_o",
                         "funny_o", "ambitious_o", "shared_interests_o",
                         "attractive_partner", "sincere_partner",
                         "intelligence_partner", "funny_partner",
                         "ambition_partner", "shared_interests_partner",
                         "like", "guess_prob_liked", "met")
threshold <- length(columns_of_interest) / 2
partially_filtered_dates <- dates[
  rowSums(is.na(dates[, columns_of_interest])) <= threshold,
  ]
```

```
partially_filtered_dates[, columns_of_interest] <-
  lapply(partially_filtered_dates
         [, columns_of_interest, drop = FALSE], function(x) {
  median_value <- median(x, na.rm = TRUE)
  x[is.na(x)] <- median_value
  return(x)
})

filtered_dates <- partially_filtered_dates
#filtered_dates
```

Definiramo stršeće vrijednosti:

```
#boxplots <- lapply(columns_of_interest, function(column_name) {
# boxplot(filtered_dates[[column_name]], main = column_name)
#})

boxplot(filtered_dates[["met"]], main = "met")
```



Met ima stršeće vrijednosti koje ćemo maknuti na način da ćemo vrijednosti veće od 1 smanjiti na 1.

```
filtered_dates$met[filtered_dates$met > 1] <- 1

unique(filtered_dates$met)
```

```
## [1] 0 1
```

Da bismo procijenili može li se na temelju ocjena sudionika o njihovim partnerima predvidjeti je li se sudioniku partner svidio, prvo ćemo definirati što smatramo "sviđanjem partnera". U našoj analizi, smatrati će se da se sudioniku svidio partner ako sudionik svoje cjelokupno iskustvo ocijeni s 1, a neuspješnim ako je ocijenjeno s 0. Ova definicija uzima u obzir subjektivnost ocjena sudionika.

Budući da je odluka o tome je li se sudioniku partner svidio binarna (1 za "svidio mu se partner", 0 za "nije mu se svidio partner"), odlučili smo primijeniti model logističke regresije. Logistička regresija je statistički model koji se koristi za analizu odnosa između binarne zavisne promjenljive i više nezavisnih promjenljivih. Oblik modela je sljedeći:

$$F(x'\beta) = \Lambda(x'\beta) = \frac{1}{1 + e^{-x'\beta}}$$

Cilj je predviđanje vjerojatnosti da će zavisna promjenljiva varijabla imati vrijednost 1 (uspjeh) ili 0 (neuspjeh) na osnovu linearnih kombinacija nezavisnih promjenljivih varijabli ugurane u sigmoidalnu funkciju. Model logističke regresije ima funkciju vjerodostojnosti:

$$L(\beta) = \prod_{i=1}^{N} (\Lambda(x_i'\beta))^{y_i} (1 - \Lambda(x_i'\beta))^{1-y_i}$$

ili u log obliku:

$$l(\beta) = \sum_{i=1}^{N} y_i log(\Lambda(x_i'\beta)) + \sum_{i=1}^{N} (1 - y_i) log(1 - \Lambda(x_i'\beta))$$

Ovaj problem nema konkretno rješenje, ali može se riješiti ili iterativno metodom gradijentnog spusta ili numerički Newton-Raphsonovom metodom. Logistička regresija omogućava interpretaciju utjecaja svake

nezavisne varijable primjenjive na log-odds vjerojatnosti, pružajući također mogućnost procjene vjerojatnosti klasifikacije.

Kako bismo vidjeli snagu ovog modela odlučili smo primijeniti unakrsnu provjeru, koja uvodne podatke na trening_set:testing_set u omjeru 70:30:

```
set.seed(123)

train_indices <- sample(1:nrow(filtered_dates), 0.7 * nrow(filtered_dates))

training_set <- filtered_dates[train_indices, ]

testing_set <- filtered_dates[-train_indices, ]
```

Kod za učenje i rezultati učenja su vidljivi ovdje:

```
logistic_model <- glm(decision ~ attractive_partner +
                        sincere_partner + intelligence_partner +
                        funny_partner + ambition_partner +
                        shared_interests_partner + like +
                        guess_prob_liked + met,
                      data = training_set, family = "binomial")

summary(logistic_model)
```

```
##
## Call:
## glm(formula = decision ~ attractive_partner + sincere_partner +
##     intelligence_partner + funny_partner + ambition_partner +
##     shared_interests_partner + like + guess_prob_liked + met,
##     family = "binomial", data = training_set)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.31329    0.23808 -26.517  < 2e-16 ***
## attractive_partner        0.45971    0.02627  17.502  < 2e-16 ***
## sincere_partner          -0.20030    0.02984  -6.713 1.90e-11 ***
## intelligence_partner     -0.02509    0.03624  -0.692    0.489
## funny_partner             0.14737    0.02856   5.160 2.47e-07 ***
## ambition_partner         -0.16493    0.02844  -5.798 6.70e-09 ***
## shared_interests_partner  0.09507    0.02379   3.997 6.42e-05 ***
## like                      0.54269    0.03557  15.257  < 2e-16 ***
## guess_prob_liked          0.16753    0.01997   8.389  < 2e-16 ***
## met                      -0.13486    0.17494  -0.771    0.441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7569.8  on 5543  degrees of freedom
## Residual deviance: 5254.1  on 5534  degrees of freedom
## AIC: 5274.1
##
## Number of Fisher Scoring iterations: 5
```

Kako bismo odredili preciznost modela, prvo smo odlučili implementirati ROC krivulju. ROC(Receiver

Operating Characteristic) je graf koji ilustrira sposobnost dijagnostičkog modela u binarnoj klasifikaciji na različitim pragovima odlučivanja. Prikazuje odnos između stvarno pozitivnih rezultata (osjetljivosti) i lažno pozitivnih rezultata (1 - specifičnost), pomažući vizualizaciji kompromisa između osjetljivosti i specifičnosti na različitim pragovima. AUC (površina ispod krivulje ROC) mjera je ukupne performanse modela. Što je AUC bliže 1, to je model bolji.

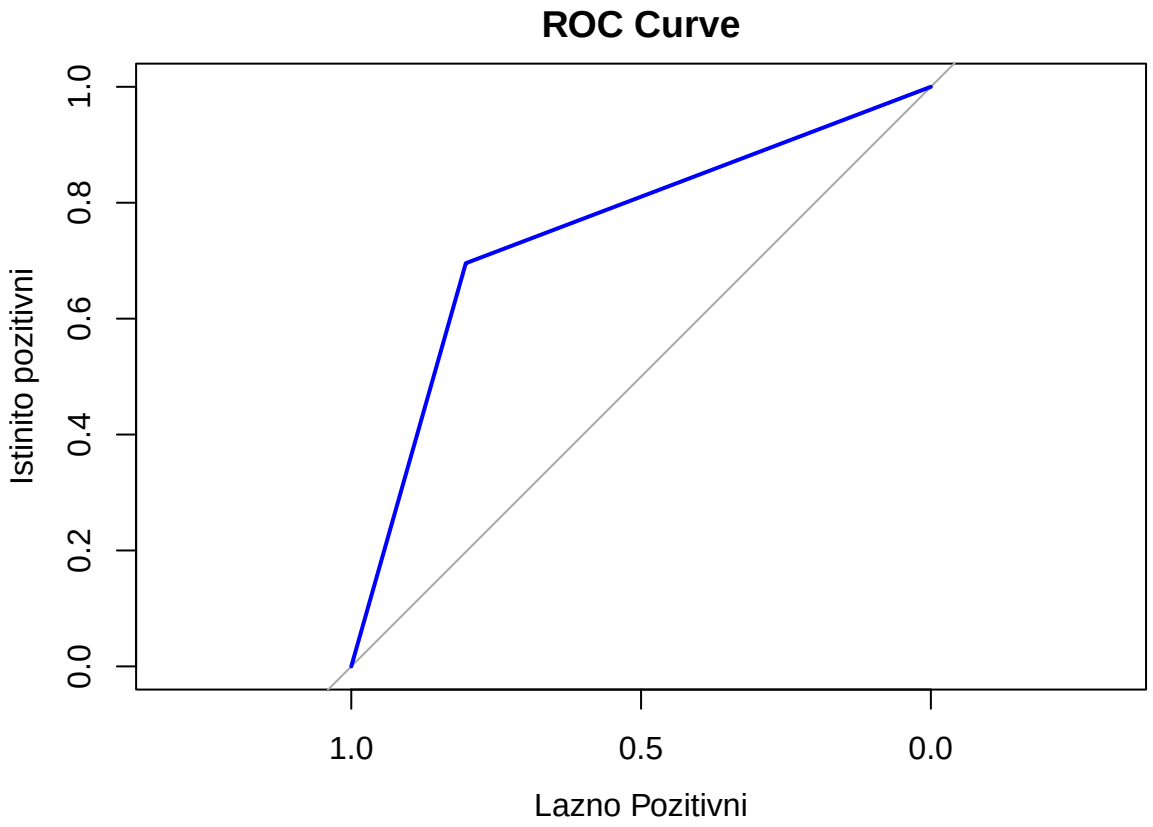
Graf ROC je ovdje:

```
library(pROC)
library(pheatmap)

predictions <- predict(logistic_model, testing_set, type = "response")

# Convert predicted probabilities to binary predictions (0 or 1)
predicted_class <- ifelse(predictions > 0.5, 1, 0)

# Assuming 'response_variable' is the actual response variable in your data
actual_class <- testing_set$decision

# Confusion matrix
conf_matrix <- table(actual_class, predicted_class)

roc_curve <- roc(actual_class, predicted_class)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(roc_curve, main = "ROC Curve",
     col = "blue",
     lwd = 2,
     xlab = "Lažno Pozitivni",
     ylab="Istinito pozitivni")
```



S AUC vrijednosti ispod grafa:

```
# Area under the ROC curve (AUC)
auc_value <- auc(roc_curve)
cat("AUC:", auc_value, "\n")
```

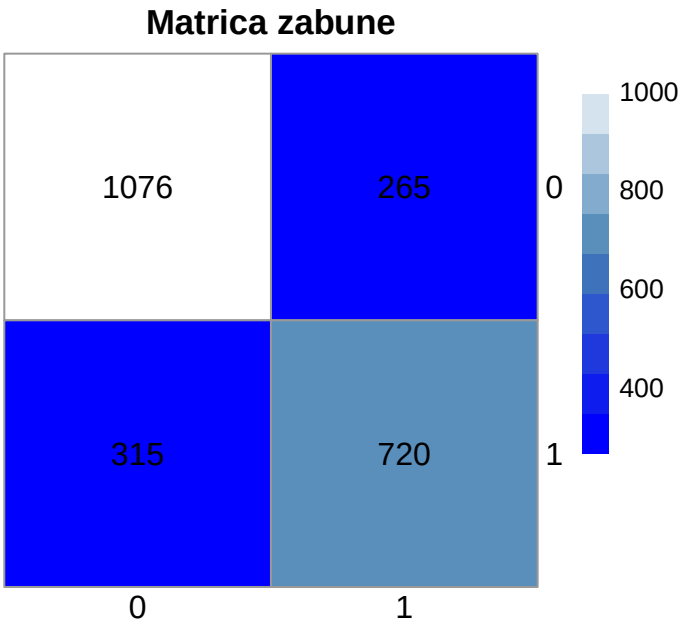
```
## AUC: 0.7490192
```

Sljedeća performansa koju smo htjeli pokazati je matrica zabune. Matrica zabune je tablica koja prikazuje broj stvarno pozitivnih, stvarno negativnih, lažno pozitivnih i lažno negativnih klasifikacija u kontekstu evaluacije binarnog klasifikacijskog modela, pružajući uvid u performanse modela i vrste pogrešaka koje čini. Ova matrica često služi kao osnova za izračunavanje različitih evaluacijskih mjera, poput preciznosti, osjetljivosti i specifičnosti.

Naša matrica zabune:

```
# Stylish Confusion Matrix using pheatmap
pheatmap(
  conf_matrix,
  main = "Matrica zabune",
  fontsize_row = 12,  # Adjust font size for row labels
  fontsize_col = 12,  # Adjust font size for column labels
  cellwidth = 100,     # Adjust cell width
  cellheight = 100,    # Adjust cell height
  cluster_cols = FALSE,
  cluster_rows = FALSE,
  display_numbers=TRUE,
  color = colorRampPalette(c("blue", "steelblue", "white"))(10),
  show_rownames = TRUE,  # Show row names (actual labels)
  show_colnames = TRUE,  # Show column names (predicted labels)
```

```
  angle_col = 0,        # Rotate column names for better visibility
  number_color = "black",  # Color of the text inside cells
  number_format = "%.0f", # Format for the numbers (adjust as needed)
  fontsize_number = 12    # Font size of the numbers inside cells
)
```



Reci matrice prikazuju istinite vrijednosti, dok stupci prikazuju predviđene vrijednosti za model. Polja na dijagonali prikazuju broj točno previđenih primjera, dok polja izvan dijagonale prikazuju broj krivo predviđenih primjera. Kao što se vidi iz našeg modela, ima mnogo točno predviđenih modela, ali i mnogo krivo predviđenih modela, koji će se isto vidjeti u kasnijim metrikama.

Još neke mjere preciznosti:

Točnost: Glavna interpretacija koliko je dobar model. Formula:

$$1) Točnost(Accuracy) = \frac{TP + TN}{FP + FN + TP + TN}$$

Preciznost: Pokazatelj lažno pozitivnih rezulata. Ključna je metrika kada je pogreška lažne klasifikacija velika. Formula:

$$2) Preciznost(Precision) = \frac{TP}{FP + TP}$$

Osjetljivost: važan kada su troškovi propuštanja pozitivnih slučajeva visoki. Cilj mu je minimizirati lažno negativne rezultate. Formula:

$$3) Osjetljivost(Recall) = \frac{TP}{TP + FN}$$

F1: Koristan je kada želimo jednu metriku koja uzima u obzir i lažno pozitivne i lažno negativne rezultate. Formula:

$$4) F1 = \frac{2 * Precisnost * Osjeltjivost}{Preciznost + Osjetljivost}$$

gdje su TP(točno pozitivni), FP(lažno pozitivni), TN(točno negativni) i FN(lažno negativni). Vrijednosti ovih metrika za naš model su prikazane ovdje:

```
# Calculate accuracy
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
```

```r
# Calculate precision
precision <- conf_matrix[2, 2] / sum(conf_matrix[, 2])

# Calculate recall
recall <- conf_matrix[2, 2] / sum(conf_matrix[2, ])

# Calculate F1 score
f1_score <- 2 * (precision * recall) / (precision + recall)

# ROC curve and AUC-ROC
library(pROC)
roc_curve <- roc(actual_class, predictions)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
auc_roc <- auc(roc_curve)

cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.7558923
```

```r
cat("Precision:", precision, "\n")
```

```
## Precision: 0.7309645
```

```r
cat("Recall:", recall, "\n")
```

```
## Recall: 0.6956522
```

```r
cat("F1 Score:", f1_score, "\n")
```

```
## F1 Score: 0.7128713
```

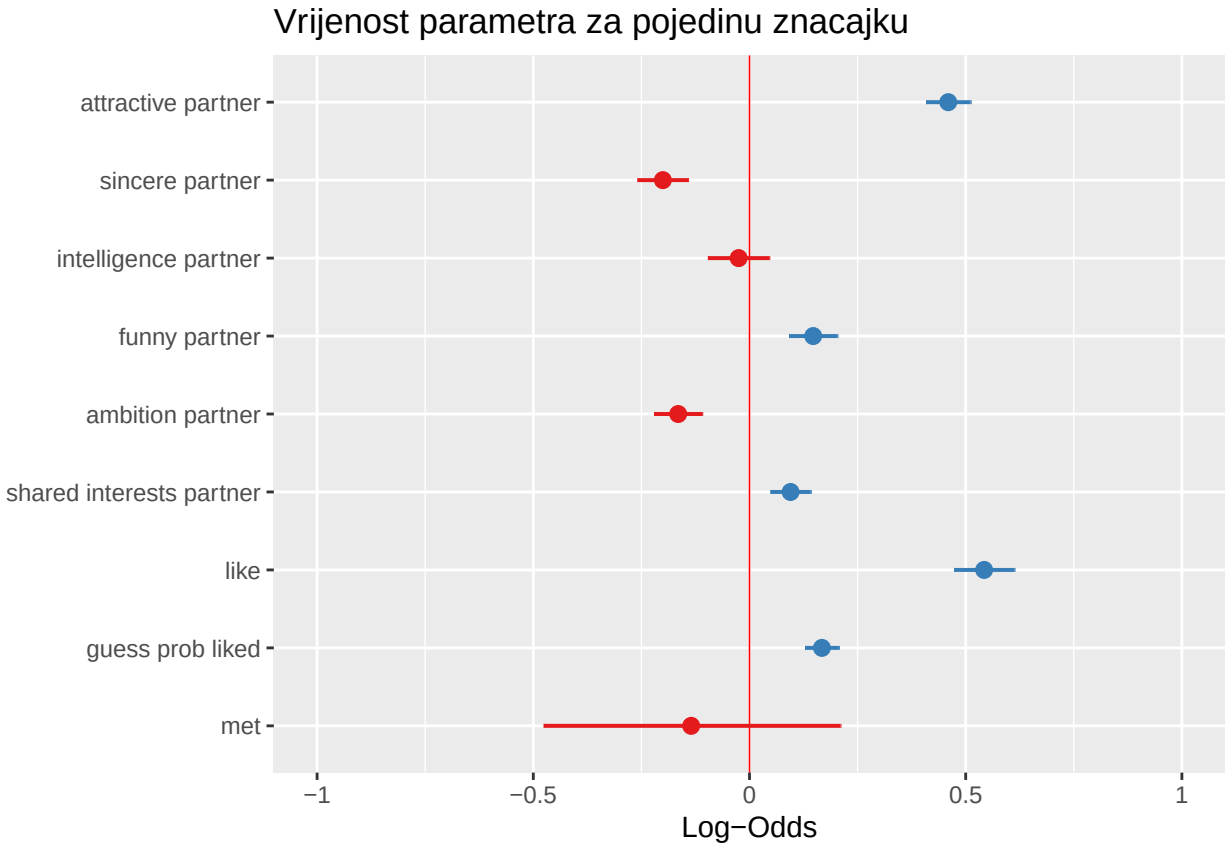
```r
#cat("AUC-ROC:", auc_roc, "\n")
```

Vidimo da su rezultati u redu. Preciznost modela je veća od 75% i većina ostalih parametara je iznad 70%.

Pitanje kojim se nadovezujemo na prethodno: Koje od značajki utječu na model najviše, a koje najmanje?

Gdje vrijednost pojedinog parametra možemo i vidjeti na sljedećem grafu:

```r
plot_model(logistic_model, vline.color = "red", transform = NULL, title="Vrijenost parametra za pojedinu
```



Jasno se vidi da parametri "like" i "atractive_partner" najviše utječu na konačnu mogućnost da model predviđa. Ali mi bismo htjeli vidjeti kojeg regresora bismo potencijalno mogli isključiti iz modela, a kojeg ne, a da a ne utječemo previše na izlaz modela. Za to ćemo iskoristiti LR test.

LR test (Likelihood Ratio test) je statistički test koji se koristi u kontekstu procjene značajnosti razlika između dvaju modela koji su ugniježđeni jedan u drugi. Test se temelji na usporedbi logaritma vjerodostojnosti (likelihood) modela koji je potpuno specifičan (nula restrikcija) s logaritmom vjerodostojnosti modela koji ima neke restrikcije (npr. postavljanje određenih parametara na nulu). Statistička testna statistika LR testa slijedi distribuciju χ-kvadrat, a p-vrijednost testa pomaže u odlučivanju o odbacivanju nulte hipoteze o tome da su modeli ekvivalentni.

Prvo što ćemo napraviti je hipoteze za LR test:

$$H_0 : \beta_g = 0$$
$$H_1 : \beta_g \neq 0$$

Sada ćemo ići po svim regresorima u originalnom modelu, istrenirati broj modela koji je jednak originalnom broju regresora, ali svaki od tih modela neće imati regresor koji želimo testirati. I vidjeti koliko je li neki regresor redundantan s nivoom značajnosti 0.05. Rezultati su vidljivi u donjoj tablici:

```
columns_of_interest <- c("attractive_partner", "sincere_partner",
                         "intelligence_partner", "funny_partner",
                         "ambition_partner", "shared_interests_partner",
                         "like", "guess_prob_liked", "met")
models <- list()

for (col in columns_of_interest) {
  formula <- as.formula(paste("decision ~",
                              paste(columns_of_interest[!columns_of_interest %in% col],
                                    collapse = "+")))
```

```
  model <- glm(formula, data = training_set, family = "binomial")
  models[[col]] <- model
}

custom_format <- function(x) {
  if (abs(x) >= 0.001) {
    return(sprintf("%.6f", x))
  } else {
    return(format(x, scientific = TRUE))
  }
}

p_values <- numeric(length(columns_of_interest))
i = 1
for(col in columns_of_interest) {
  p_value <- lrtest(logistic_model, models[[col]])$`Pr(>Chisq)`[2]
  #print(p_value)
  p_values[i] <- p_value
  i <- i+1
}

p_values <- data.frame(column = columns_of_interest, p_value = p_values)

p_values <- p_values %>% mutate(p_value = sapply(p_value, custom_format))

p_values
```

```
##                     column      p_value
## 1       attractive_partner 9.492624e-76
## 2          sincere_partner 1.345313e-11
## 3     intelligence_partner     0.488660
## 4            funny_partner 2.152523e-07
## 5         ambition_partner 5.292164e-09
## 6 shared_interests_partner 6.143274e-05
## 7                     like 1.921095e-57
## 8         guess_prob_liked 2.655156e-17
## 9                      met     0.441962
```

Vidimo po rezultatima da su "attractive_partner" i "like" najznačajniji parametri, dok "intelligance_partner" najneznačajniji za izlaz modela. Što je i očekivano po vrijednostima parametara u originalnom modelu.

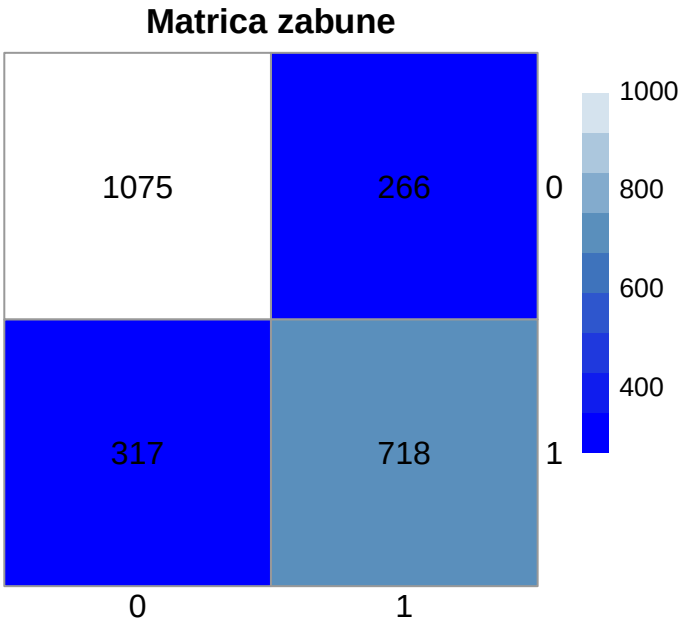
Za konačni eksperiment u ovom radu, htjeli smo vidjeti je li moguće poboljšati izlaz modela ako izbacimo regresore koji su zadovoljili LR test. Konačni rezultati su ovdje:

```
logistic_model_better <- glm(decision ~ attractive_partner +
                               sincere_partner + funny_partner +
                               ambition_partner + shared_interests_partner +
                               like + guess_prob_liked,
                             data = training_set, family = "binomial")

logistic_model_better
```

```
##
## Call:  glm(formula = decision ~ attractive_partner + sincere_partner +
##     funny_partner + ambition_partner + shared_interests_partner +
```

```
##     like + guess_prob_liked, family = "binomial", data = training_set)
##
## Coefficients:
##              (Intercept)        attractive_partner           sincere_partner
##                  -6.34963                   0.45945                   -0.20837
##            funny_partner          ambition_partner  shared_interests_partner
##                  0.14395                  -0.17238                   0.09453
##                     like          guess_prob_liked
##                  0.53886                   0.16870
##
## Degrees of Freedom: 5543 Total (i.e. Null);  5536 Residual
## Null Deviance:       7570
## Residual Deviance: 5255  AIC: 5271
```

```r
predictions <- predict(logistic_model_better, testing_set, type = "response")

predicted_class <- ifelse(predictions > 0.5, 1, 0)

# Assuming 'response_variable' is the actual response variable in your data
actual_class <- testing_set$decision

# Confusion matrix
conf_matrix <- table(actual_class, predicted_class)

pheatmap(
  conf_matrix,
  main = "Matrica zabune",
  fontsize_row = 12,  # Adjust font size for row labels
  fontsize_col = 12,  # Adjust font size for column labels
  cellwidth = 100,     # Adjust cell width
  cellheight = 100,    # Adjust cell height
  cluster_cols = FALSE,
  cluster_rows = FALSE,
  display_numbers=TRUE,
  color = colorRampPalette(c("blue", "steelblue", "white"))(10),
  show_rownames = TRUE,  # Show row names (actual labels)
  show_colnames = TRUE,  # Show column names (predicted labels)
  angle_col = 0,       # Rotate column names for better visibility
  number_color = "black",  # Color of the text inside cells
  number_format = "%.0f", # Format for the numbers (adjust as needed)
  fontsize_number = 12    # Font size of the numbers inside cells
)
```




```
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.7546296
```

Na temelju ovog modela odgovorili smo na pitanje može li se predvidjeti hoće li se sudioniki svidjeti partner ili neće. U procesu izgradnje modela odredili smo koje varijable pozitivno utječu na odgovor na ovo pitanje te smo uz korištenje tih varijabli izgradili model. Varijable su sljedeće:

attractive_partner - Ocjena sudionika o privlačnosti njihovog partnera. sincere_partner - Ocjena sudionika o iskrenosti njihovog partnera. funny_partner - Ocjena sudionika o smislu za humor njihovog partnera. ambition_partner - Ocjena sudionika o ambiciji njihovog partnera. shared_interests_partner - Ocjena koju partner daje za zajedničke interese. like - Koliko je sudioniku bio sviđao njihov partner. guess_prob_liked - Pretpostavka sudionika o tome koliko je vjerojatno da se njihovom partneru svidjeli.

S time da smo odredili kako varijable attractive_partner i like najviše utječu na izlaz modela.