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?

Počet ćemo s analizom deskriptivne statistike kako bismo stekli osnovni uvid u varijable poput starosti, 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")
head(participants)
```

```
##      id wave age                                    race gender
## 1 10000    1  21 Asian/Pacific Islander/Asian-American female
## 2 10001    1  24           European/Caucasian-American female
## 3 10002    1  25           European/Caucasian-American female
## 4 10003    1  23           European/Caucasian-American female
## 5 10004    1  21           European/Caucasian-American female
## 6 10005    1  23 Asian/Pacific Islander/Asian-American female
##   ambition_important attractive_important funny_important
## 1                 15                    15              15
## 2                  0                    45              20
## 3                 10                    35              10
## 4                 10                    20              20
```

```
## 5                 10                   20             25
## 6                  5                   10             25
##   intelligence_important shared_interests_important sincere_important     field
## 1                     20                         15                20       Law
## 2                     25                          5                 5       law
## 3                     35                          0                10 Economics
## 4                     20                         10                20       Law
## 5                     25                         15                 5       Law
## 6                     20                         15                25       law
##   importance_same_race importance_same_religion sports tvsports exercise dining
## 1                    2                        4      9        2        8      9
## 2                    2                        5      3        2        7     10
## 3                    8                        4      3        8        7      8
## 4                    1                        1      1        1        6      7
## 5                    8                        1      7        4        7      7
## 6                    1                        1     10        8        9      7
##   museums art hiking gaming clubbing reading tv theater movies concerts music
## 1       1   1      5      1        5       6  9       1     10       10     9
## 2       8   6      3      5        8      10  1       9      8        7     8
## 3       5   5      8      4        5       7  8       7      7        7     5
## 4       6   7      7      5        7       7  7       9      7        8     7
## 5       6   8      6      6        8       6  8       6      6        3     7
## 6       8   7      9      2        6       9  2       5      6        6     4
##   shopping yoga expected_happy_with_sd_people expected_num_interested_in_me
## 1        8    1                             3                            2
## 2        3    1                             4                            5
## 3        8    7                             4                            2
## 4        1    8                             1                            2
## 5        8    3                             7                           10
## 6        1    1                             4                            3
##   expected_num_matches attractive sincere intelligence funny ambition  X
## 1                    4          6       8            8     8        7 NA
## 2                    3          7       5           10     8        3 NA
## 3                   NA          8       9            8     9        8 NA
## 4                    2          7       8            9     7        8 NA
## 5                   NA          6       3            6    10        8 NA
## 6                    4          5       7            8     9        5 NA
```

```
head(dates)
```

```
##   date_id participant_id partner_id attractive_o sincere_o intelligence_o
## 1  100000          10000      10010            6         8              8
## 2  100001          10000      10011            7         8             10
## 3  100002          10000      10012           10        10             10
## 4  100003          10000      10013            7         8              9
## 5  100004          10000      10014            8         7              9
## 6  100005          10000      10015            7         7              8
##   funny_o ambitious_o shared_interests_o attractive_partner sincere_partner
## 1       8           8                 6                 6               9
## 2       7           7                 5                 7               8
## 3      10          10                10                 5               8
## 4       8           9                 8                 7               6
## 5       6           9                 7                 5               6
## 6       8           7                 7                 4               9
##   intelligence_partner funny_partner ambition_partner shared_interests_partner
```

```
## 1                    7            7                6                        5
## 2                    7            8                5                        6
## 3                    9            8                5                        7
## 4                    8            7                6                        8
## 5                    7            7                6                        6
## 6                    7            4                6                        4
##   like guess_prob_liked met decision
## 1    7                6   0        1
## 2    7                5   1        1
## 3    7               NA   1        1
## 4    7                6   0        1
## 5    6                6   0        1
## 6    6                5   0        0
```

Na temelju podataka vidimo da je set podataka za participante sadrži 40 stupaca dok za podatke o spojevima sadrži 19.

opis tablica i stupaca

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

```
##        id             wave            age            race
##  Min.   :10000   Min.   : 1.00   Min.   :18.00   Length:540
##  1st Qu.:10141   1st Qu.: 7.00   1st Qu.:24.00   Class :character
##  Median :10276   Median :11.00   Median :26.00   Mode  :character
##  Mean   :10276   Mean   :11.11   Mean   :26.37
##  3rd Qu.:10413   3rd Qu.:15.25   3rd Qu.:28.00
##  Max.   :10550   Max.   :21.00   Max.   :55.00
##
##     gender          ambition_important attractive_important funny_important
##  Length:540         Min.   : 0.00      Min.   :  0.00       Min.   : 0.00
##  Class :character   1st Qu.: 5.00      1st Qu.: 15.00       1st Qu.:15.00
##  Mode  :character   Median :10.00      Median : 20.00       Median :18.00
##                     Mean   :10.82      Mean   : 22.52       Mean   :17.46
##                     3rd Qu.:15.00      3rd Qu.: 25.00       3rd Qu.:20.00
##                     Max.   :53.00      Max.   :100.00       Max.   :50.00
##
##  intelligence_important shared_interests_important sincere_important
##  Min.   : 0.00          Min.   : 0.000             Min.   : 0.00
##  1st Qu.:17.29          1st Qu.: 8.248             1st Qu.:15.00
##  Median :20.00          Median :10.935             Median :18.00
##  Mean   :20.17          Mean   :11.826             Mean   :17.31
##  3rd Qu.:23.02          3rd Qu.:16.000             3rd Qu.:20.00
##  Max.   :50.00          Max.   :30.000             Max.   :60.00
##
##     field           importance_same_race importance_same_religion
##  Length:540         Length:540           Min.   : 1.000
##  Class :character   Class :character     1st Qu.: 1.000
##  Mode  :character   Mode  :character     Median : 3.000
##                                          Mean   : 3.593
##                                          3rd Qu.: 6.000
##                                          Max.   :10.000
##
##      sports          tvsports        exercise          dining
##  Min.   : 1.000   Min.   : 1.00   Min.   : 1.000   Min.   : 1.000
```

```
##  1st Qu.: 4.000   1st Qu.: 2.00   1st Qu.: 5.000   1st Qu.: 7.000
##  Median : 7.000   Median : 4.00   Median : 7.000   Median : 8.000
##  Mean   : 6.409   Mean   : 4.55   Mean   : 6.278   Mean   : 7.778
##  3rd Qu.: 9.000   3rd Qu.: 7.00   3rd Qu.: 8.000   3rd Qu.: 9.000
##  Max.   :10.000   Max.   :10.00   Max.   :10.000   Max.   :10.000
##
##     museums            art             hiking           gaming
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 6.000   1st Qu.: 5.000   1st Qu.: 4.000   1st Qu.: 1.000
##  Median : 7.000   Median : 7.000   Median : 6.000   Median : 3.000
##  Mean   : 6.981   Mean   : 6.711   Mean   : 5.759   Mean   : 3.874
##  3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 6.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :14.000
##
##     clubbing         reading             tv           theater
##  Min.   : 0.000   Min.   : 1.000   Min.   : 1.00   Min.   : 0.000
##  1st Qu.: 4.000   1st Qu.: 7.000   1st Qu.: 3.00   1st Qu.: 5.000
##  Median : 6.000   Median : 8.000   Median : 6.00   Median : 7.000
##  Mean   : 5.717   Mean   : 7.631   Mean   : 5.37   Mean   : 6.752
##  3rd Qu.: 8.000   3rd Qu.: 9.000   3rd Qu.: 7.00   3rd Qu.: 9.000
##  Max.   :10.000   Max.   :13.000   Max.   :10.00   Max.   :10.000
##
##      movies          concerts          music          shopping
##  Min.   : 0.000   Min.   : 0.000   Min.   : 1.00   Min.   : 1.000
##  1st Qu.: 7.000   1st Qu.: 6.000   1st Qu.: 7.00   1st Qu.: 4.000
##  Median : 8.000   Median : 7.000   Median : 8.00   Median : 6.000
##  Mean   : 7.906   Mean   : 6.865   Mean   : 7.88   Mean   : 5.656
##  3rd Qu.: 9.000   3rd Qu.: 8.000   3rd Qu.: 9.00   3rd Qu.: 8.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.00   Max.   :10.000
##
##       yoga        expected_happy_with_sd_people expected_num_interested_in_me
##  Min.   : 0.000   Min.   : 1.000                Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.: 5.000                1st Qu.: 2.000
##  Median : 4.000   Median : 6.000                Median : 4.000
##  Mean   : 4.433   Mean   : 5.491                Mean   : 5.869
##  3rd Qu.: 7.000   3rd Qu.: 7.000                3rd Qu.: 8.750
##  Max.   :10.000   Max.   :10.000                Max.   :20.000
##                                                 NA's   :410
##  expected_num_matches   attractive        sincere        intelligence
##  Min.   : 0.000       Min.   : 1.000   Min.   : 2.000   Min.   : 2.000
##  1st Qu.: 2.000       1st Qu.: 6.000   1st Qu.: 8.000   1st Qu.: 7.000
##  Median : 3.000       Median : 7.000   Median : 8.000   Median : 8.000
##  Mean   : 3.043       Mean   : 7.052   Mean   : 8.275   Mean   : 7.714
##  3rd Qu.: 4.000       3rd Qu.: 8.000   3rd Qu.: 9.000   3rd Qu.: 9.000
##  Max.   :18.000       Max.   :10.000   Max.   :10.000   Max.   :10.000
##  NA's   :73           NA's   :5        NA's   :2        NA's   :2
##      funny           ambition            X
##  Min.   : 3.000   Min.   : 2.000   Min.   :4.000
##  1st Qu.: 8.000   1st Qu.: 7.000   1st Qu.:7.000
##  Median : 8.000   Median : 8.000   Median :8.000
##  Mean   : 8.387   Mean   : 7.589   Mean   :7.571
##  3rd Qu.: 9.000   3rd Qu.: 9.000   3rd Qu.:9.000
##  Max.   :10.000   Max.   :10.000   Max.   :9.000
##  NA's   :2        NA's   :2        NA's   :533
```

```
summary(dates)
```

```
##     date_id       participant_id    partner_id     attractive_o
##  Min.   :100000   Min.   :10000   Min.   :10000   Min.   : 0.000
##  1st Qu.:102149   1st Qu.:10155   1st Qu.:10155   1st Qu.: 5.000
##  Median :104170   Median :10278   Median :10278   Median : 6.000
##  Mean   :104187   Mean   :10282   Mean   :10282   Mean   : 6.194
##  3rd Qu.:106268   3rd Qu.:10405   3rd Qu.:10405   3rd Qu.: 8.000
##  Max.   :108377   Max.   :10550   Max.   :10550   Max.   :10.500
##                                                   NA's   :176
##    sincere_o      intelligence_o      funny_o        ambitious_o
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 6.000   1st Qu.: 6.000   1st Qu.: 5.000   1st Qu.: 6.00
##  Median : 7.000   Median : 7.000   Median : 7.000   Median : 7.00
##  Mean   : 7.185   Mean   : 7.381   Mean   : 6.409   Mean   : 6.79
##  3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 8.00
##  Max.   :10.000   Max.   :10.000   Max.   :11.000   Max.   :10.00
##  NA's   :247      NA's   :268      NA's   :321      NA's   :678
##  shared_interests_o attractive_partner sincere_partner  intelligence_partner
##  Min.   : 0.000     Min.   : 0.000     Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 4.000     1st Qu.: 5.000     1st Qu.: 6.000   1st Qu.: 6.000
##  Median : 6.000     Median : 6.000     Median : 7.000   Median : 7.000
##  Mean   : 5.489     Mean   : 6.194     Mean   : 7.185   Mean   : 7.381
##  3rd Qu.: 7.000     3rd Qu.: 8.000     3rd Qu.: 8.000   3rd Qu.: 8.000
##  Max.   :10.000     Max.   :10.000     Max.   :10.000   Max.   :10.000
##  NA's   :1029       NA's   :176        NA's   :247      NA's   :268
##  funny_partner    ambition_partner shared_interests_partner      like
##  Min.   : 0.000   Min.   : 0.00    Min.   : 0.000           Min.   : 0.000
##  1st Qu.: 5.000   1st Qu.: 6.00    1st Qu.: 4.000           1st Qu.: 5.000
##  Median : 7.000   Median : 7.00    Median : 6.000           Median : 6.000
##  Mean   : 6.409   Mean   : 6.79    Mean   : 5.489           Mean   : 6.142
##  3rd Qu.: 8.000   3rd Qu.: 8.00    3rd Qu.: 7.000           3rd Qu.: 7.000
##  Max.   :10.000   Max.   :10.00    Max.   :10.000           Max.   :10.000
##  NA's   :321      NA's   :678      NA's   :1029             NA's   :213
##  guess_prob_liked      met            decision
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 4.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 5.000   Median :0.0000   Median :0.0000
##  Mean   : 5.216   Mean   :0.0501   Mean   :0.4221
##  3rd Qu.: 7.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :10.000   Max.   :8.0000   Max.   :1.0000
##  NA's   :278      NA's   :343
```

Na temelju predefiniranih hipoteza možemo zaključiti da svi stupci u podatcima nisu jednako relevantni. U slučaju analize je li inteligencija partnera važnija od fizičkog izgleda od najveće su važnosti podaci vezani uz inteligenciju i fizički izgled. Analogno za analizu povezanosti gaminga i zanimanja participanta najvažniji su podaci vezani uz gaming i samo područje zanimanja participanta.

Promotrimo relevantne varijable koje ćemo kasnije koristiti u istraživačkim pitanjima. U vezi prvog pitanja bitni su nam inteligencija i fizički izgled partnera.

Je li inteligencija partnera ispitanicima važnija od izgleda?

Za ovo pitanje su nam bitni podaci o važnosti izgleda, odnosno važnosti inteligencije kod ispitanika.Te podatke pronalazimo u participants pod varijablama naziva attractive_important tj. intelligence_important.

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 sa potrebnim testovima. Naposljetku, ćemo rezultate analizirati, usporediti te donijeti zaključke.

Prvo ćemo izvaditi sve null vrijednosti iz podataka.

```
sum(is.na(participants$intelligence_important)) # ukupno nepostojecih vrijednosti za podatak inteligenc
```

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

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

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

```
df_cleaned_first <- participants[, c("intelligence_important", "attractive_important")] # potrebni poda
df_cleaned_first_male <- participants[participants$gender == "male", c("intelligence_important", "attra
df_cleaned_first_female <- participants[participants$gender == "female", c("intelligence_important", "a
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
```

```
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
```

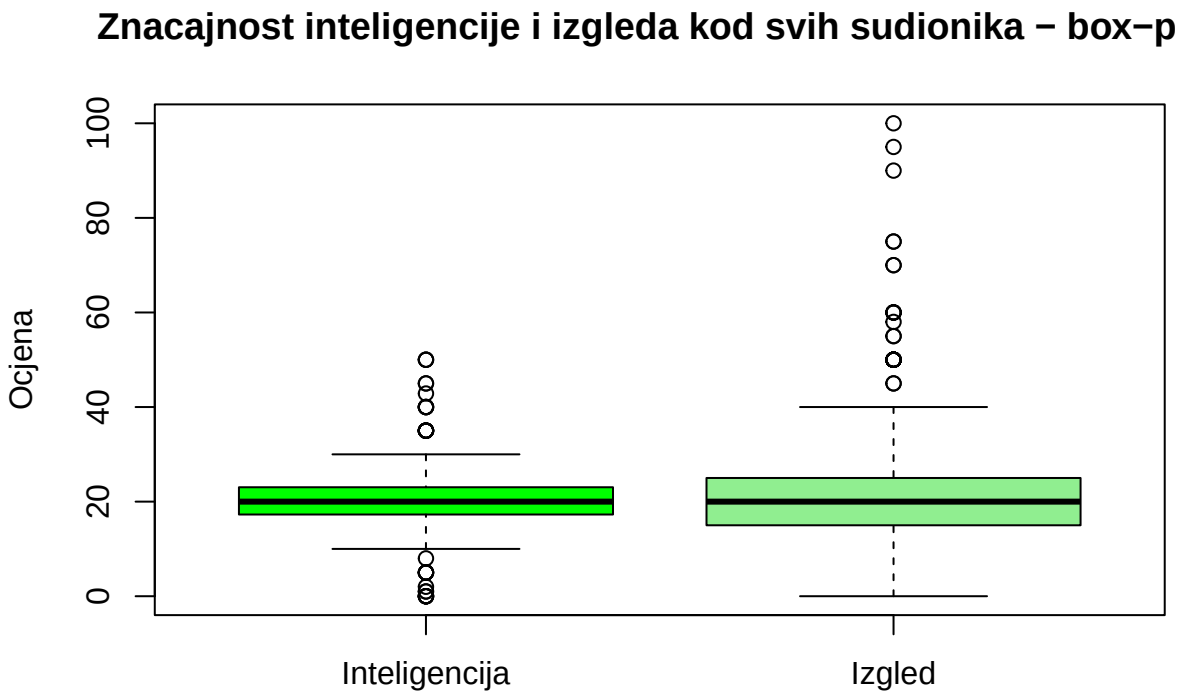
```
sum(participants$gender == "female")
```

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

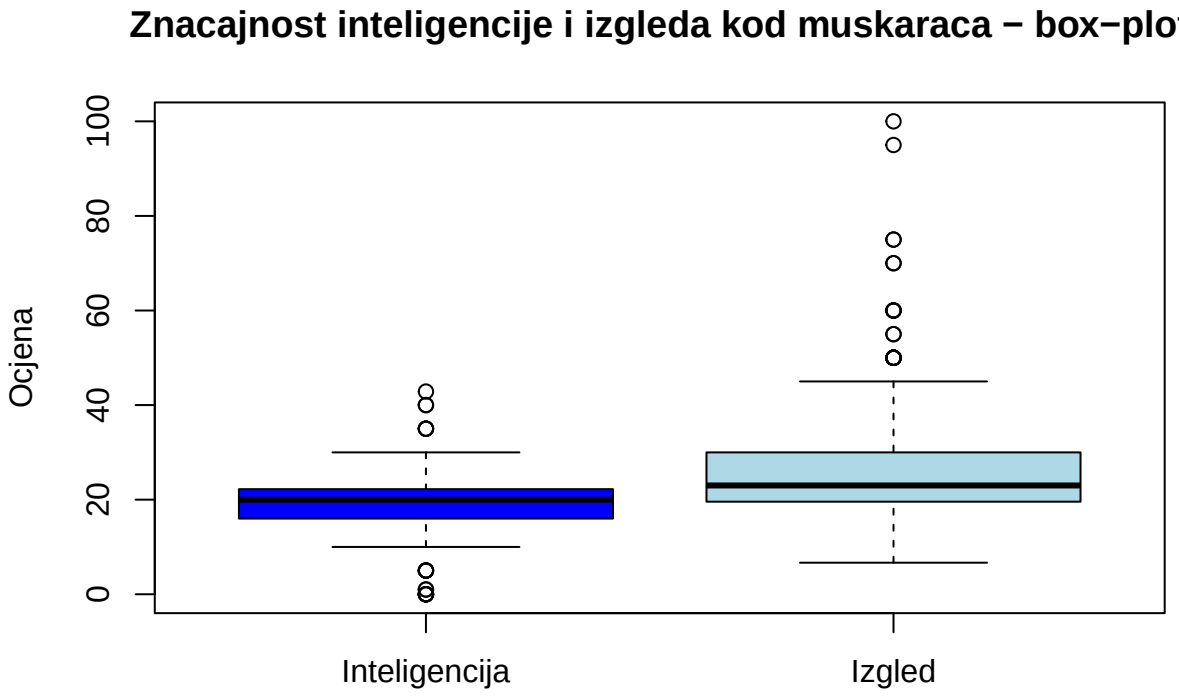
Na temelju ispisa ovih podataka možemo vidjeti da su medijani za vrijednosti važnosti inteligencije i izgleda isti kada se gleda skup svih sudionika. Aritmetička sredina vrijednosti kod svih sudionika je veća za važnost izgleda nego za inteligenciju, redom, 22.52 i 20.17. Cijeli skup podataka podijelili 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 vizalizirati, 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.

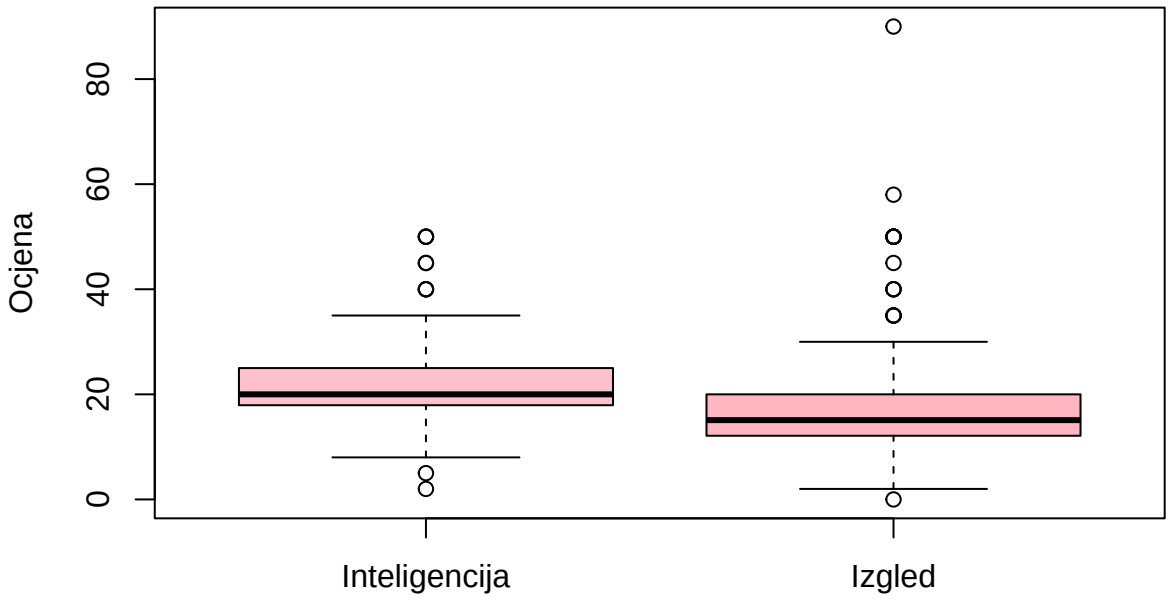
```
boxplot(cbind(df_cleaned_first$intelligence_important, df_cleaned_first$attractive_important),
        names = c("Inteligencija", "Izgled"),
        main = 'Značajnost inteligencije i izgleda kod svih sudionika - box-plot',
        ylab = 'Ocjena',
        col = c("green", "lightgreen"))
```




```
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"))
```




```
boxplot(cbind(df_cleaned_first_female$intelligence_important, df_cleaned_first_female$attractive_importa
        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.