Big Five – velikih pet dimenzija ličnosti

Projekt - SAP

Grupa Sapunaši - Tin Popović , Martina Galić , Ivan Hajpek, Petra Dunja Grujić Ostojić

14.1.2022.

Uvod

Big Five je dominantan model klasifikacije osobina ličnosti kojim je predložena podjela ljudskih karakteristika u sljedećih pet glavnih faktora:

- *Ekstraverzija* razlikuje ljude po njihovoj interakciji s okolinom. Ona obuhvaća osobine poput društvenosti i poduzetnosti. Na drugom se polu nalazi introverzija povezana s karakteristikama poput povučenosti i mirnoće.
- *Ugodnost* mjeri empatiju i kooperativnost nasuprot sumnjičavosti i animozitetu.
- *Savjesnost* obuhvaća osobine vezane uz odnos prema obavezama i poslu. Osobe s niskom razinom savjesnosti smatraju se nepouzdanima te sklonijima spontanom ponašanju.
- *Neuroticizam* je dimenzija ličnosti koja promatra emocionalnu (ne)stabilnost.
- *Otvorenost* vrednuje sklonost prema novim idejama i konceptima, razlikuje kreativne i konvencionalne osobe.

Svima nam su dobro poznate društvene predrasude koje opravdavaju karakteristike neke osobe putem spola, starosti ili kulture (države podrijetla). Kroz ovaj projekt pozabavit ćemo se s takvim predrasudama, te statistički zaključiti postoji li ikakva istina u njima.

Deskriptivna analiza skupa podataka

Započetak moramo učitati potrebne pakete:

```
library(dplyr)
```

Zatim učitajmo podatke , i pregledajmo dimenzije našeg skupa.

```
BigFive = read.csv("datasets/big_five_scores.csv")
dim(BigFive)
```

```
## [1] 307313      9
```

Skup se sastoji od 307313 zapisa i 9 njihovi opisa(varijabli).Svaki zapis predstavlja jednog ispitanika . Varijable koje opisuju svakog ispitanika su:

```
data.frame(names(BigFive))
```

```
##            names.BigFive.
## 1                 case_id
## 2                 country
## 3                     age
## 4                     sex
## 5         agreeable_score
## 6      extraversion_score
## 7          openness_score
```

```
## 8 conscientiousness_score
## 9       neuroticism_score
```

Vidimo da su osobe opisane osnovnim informacijama (kao što su starost, spol i država podrijetla), te varijablama osobnosti (ekstraverzija, ugodnost, savjesnost, neuroticizam, otvorenost).

```
View(BigFive)
```

Nakon pregleda čitavog podatkovnog skupa, zasigurno možemo potvrditi preglednost ovog skupa. Osam od devet varijabli nam predstavljaju ključne podatke koje ćemo nadalje koristiti u statističkoj analizi. Prva varijabla skupa ("case_id"), ne predstavlja nikakvu važnost za naše nadolazeće analize, ali numeracija (i identifikacija) zapisa u podatkovnom skupu je uvijek poželjna. Zbog toga nećemo raditi nikakve modifikacije nad podatkovnim skupom.

Radimo s izrazito velikom količinom podataka, te u takvim slučajevima pojava NA vrijednosti nije malena.Ukoliko one postoje moramo s njima pažljivo rukovati, jer mogu dovesti do pogrešnih statističkih odluka. Zbog toga moramo provjeriti njihovu prisutnost u našem skupu.

```
check = 1
for (col_name in names(BigFive)){
  if (sum(is.na(BigFive[,col_name])) > 0){
    check=0
    cat('Ukupno nedostajućih vrijednosti za varijablu ',col_name, ': ',
        sum(is.na(BigFive[,col_name])),'\n')
  }
}
if(check==1)
  print("Podatkovni skup ne sadrži nedostajuće vrijednosti")
```

```
## [1] "Podatkovni skup ne sadrži nedostajuce vrijednosti"
```

Kao što vidimo ovaj podatkovni skup ne sadrži nedostajuće vrijednosti, što je odlično jer nije potrebno ni izbacivati opservacije ni osmišljati način za nadopunu nedostajućih podataka.

Nadalje, želimo proučiti tip i ponašanje naših varijabla.

```
summary(BigFive)
```

```
##     case_id         country                age             sex
##  Min.   :     1   Length:307313      Min.   :10.00   Min.   :1.000
##  1st Qu.: 83653   Class :character   1st Qu.:18.00   1st Qu.:1.000
##  Median :166286   Mode  :character   Median :22.00   Median :2.000
##  Mean   :166683                      Mean   :25.19   Mean   :1.602
##  3rd Qu.:249627                      3rd Qu.:29.00   3rd Qu.:2.000
##  Max.   :334161                      Max.   :99.00   Max.   :2.000
##  agreeable_score  extraversion_score openness_score   conscientiousness_score
##  Min.   :0.2000   Min.   :0.2000     Min.   :0.2533   Min.   :0.2067
##  1st Qu.:0.6400   1st Qu.:0.6000     1st Qu.:0.6733   1st Qu.:0.6300
##  Median :0.7033   Median :0.6800     Median :0.7367   Median :0.7067
##  Mean   :0.6968   Mean   :0.6723     Mean   :0.7339   Mean   :0.7020
##  3rd Qu.:0.7633   3rd Qu.:0.7500     3rd Qu.:0.7967   3rd Qu.:0.7767
##  Max.   :1.0000   Max.   :0.9933     Max.   :0.9967   Max.   :1.0000
##  neuroticism_score
##  Min.   :0.1967
##  1st Qu.:0.4867
##  Median :0.5700
##  Mean   :0.5744
##  3rd Qu.:0.6600
```

```
##  Max.   :0.9967
```

```
sapply(BigFive, class)
```

```
##                 case_id                 country                     age
##               "integer"             "character"               "integer"
##                     sex         agreeable_score      extraversion_score
##               "integer"               "numeric"               "numeric"
##          openness_score conscientiousness_score       neuroticism_score
##               "numeric"               "numeric"               "numeric"
```

Vidimo da su brojčane vrijednosti dominante u ovom podatkovnom skupu. Spol je također brojčana vrijednost gdje 1 predstavlja muški , a 2 ženski spol. Ključne varijable ovog podatkovnog skupa su varijable ličnosti (agreeable_score, extraversion_score, openness_score, conscientiousness_score i neuroticism_score). One predstavljaju rezultate testa osobnosti. Rezultati za pojedinu varijablu mogu biti maksimalno 1, a minimalno 0. Zbog toga ćemo se detaljnije upoznati s njima te prikazati putem pravokutnog dijagrama, koji daje dobar uvid u ponašanje tih podataka.

```
par(mfrow=c(3,2))

boxplot(BigFive$agreeable_score,
        main='Pravokutni dijagram: faktor ugodnosti',
        ylab='ugodnost',col = cm.colors(1))
boxplot(BigFive$extraversion_score,
        main='Pravokutni dijagram: faktor ekstraverzije',
        ylab='ekstraverzija', col = "lavender")
boxplot(BigFive$openness_score,
        main='Pravokutni dijagram: faktor otvorenosti',
        ylab='otvorenost', col = "lightblue")
boxplot(BigFive$conscientiousness_score,
        main='Pravokutni dijagram: faktor savjesnosti',
        ylab='savjesnost', col = "lightyellow")
boxplot(BigFive$neuroticism_score,
        main='Pravokutni dijagram: faktor neuroticizma',
        ylab='neuroticizam', col = "purple")
```

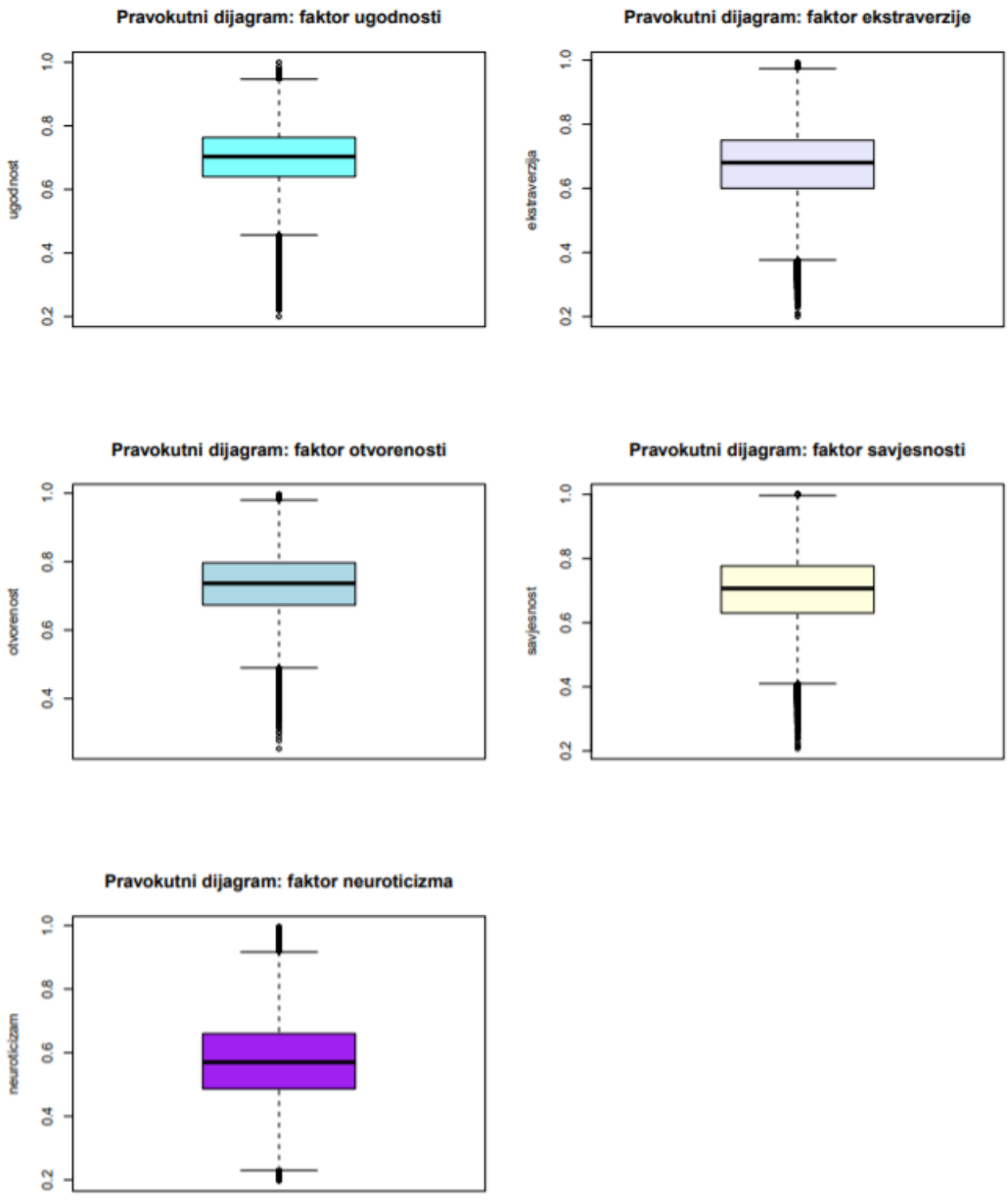

Figure 1: Grafički prikaz faktora

Kod svih faktora vidljiva je prisutnost ekstremnih vrijednosti (vrijednosti za više od 1.5*IQR (interkvartilni rang) udaljene od 1.odnosno 3. kvartila). Također vidimo da u svim faktorima (osim neuroticizma) rezultati ispitanika teže većim vrijednostima. Iste te varijable ćemo sada prikazati putem histograma, kako bi bolje prikazali numeričku raspodjelu podataka.

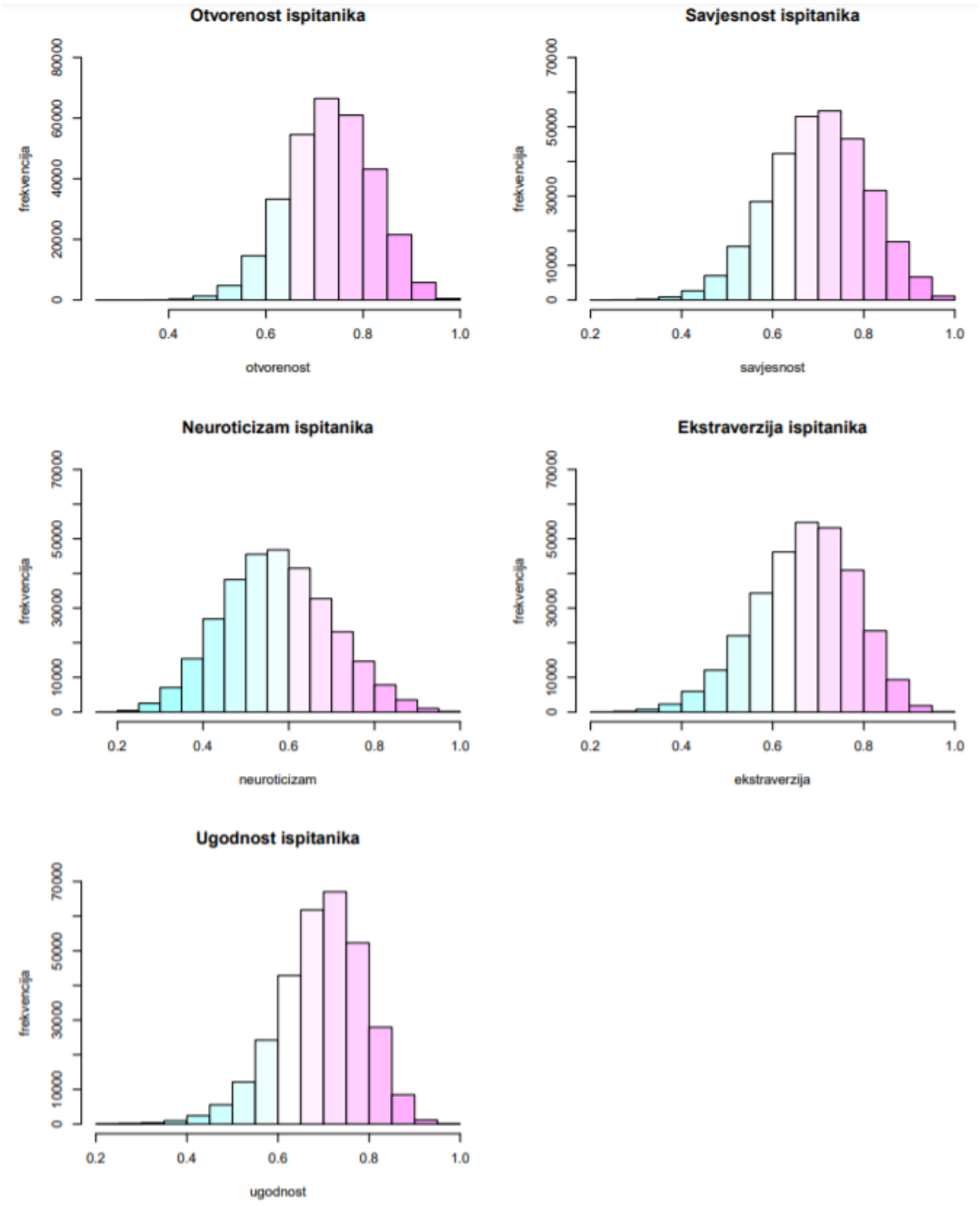

Figure 2: Graficki prikaz frekvencija faktora

Pri pogledu na histograme faktora osobnosti uočavamo da su otvorenost, ekstraverzija, savjesnost i ugodnost (više ili manje) lijevo zakrivljene distribucije, a neuroticizam je približno simetrično (normalno) distribuiran faktor. Tako već pri vizualizaciji podataka pomoću histograma vidimo da je više otvorenih, savjesnih, ekstrovertiranih i "ugodnih" ljudi među ispitanicima.

Pored varijabla ličnosti , pokazali smo da naš skup sadrži informacije o spolu, starosti i državi podrijetla pojedinog ispitanika. Ponašanje i proporcije tih podataka nećemo odjednom sada prikazati, kao što smo to učinili s varijablama ličnosti. Zbog preglednosti njih ćemo kratko predstaviti kada budemo donosili statističke zaključke u kojima oni imaju veliku ulogu.

Razlike u osobnosti ovisno o državi podrijetla

Svaka kultura, okolina, tradicija neke države oblikuje pojedinca i za očekivati je da primjerice Latinoamerikanci i Azijati neće imati iste rezultate na testovima ličnosti. Prije nego što postavimo hipoteze o ličnostima na temelju države podrijetla, pogledajmo koje su države prisutne u skupu.

```
print(paste('Broj država iz kojih ima ispitanika: ',dim(table(BigFive$country))))
```

```
## [1] "Broj država iz kojih ima ispitanika:  236"
```

```
barplot(table(BigFive$country),las = 2, cex.names=.5,main='Država podrijetla ispitanika',
        col = cm.colors(dim(table(BigFive$country))))
```

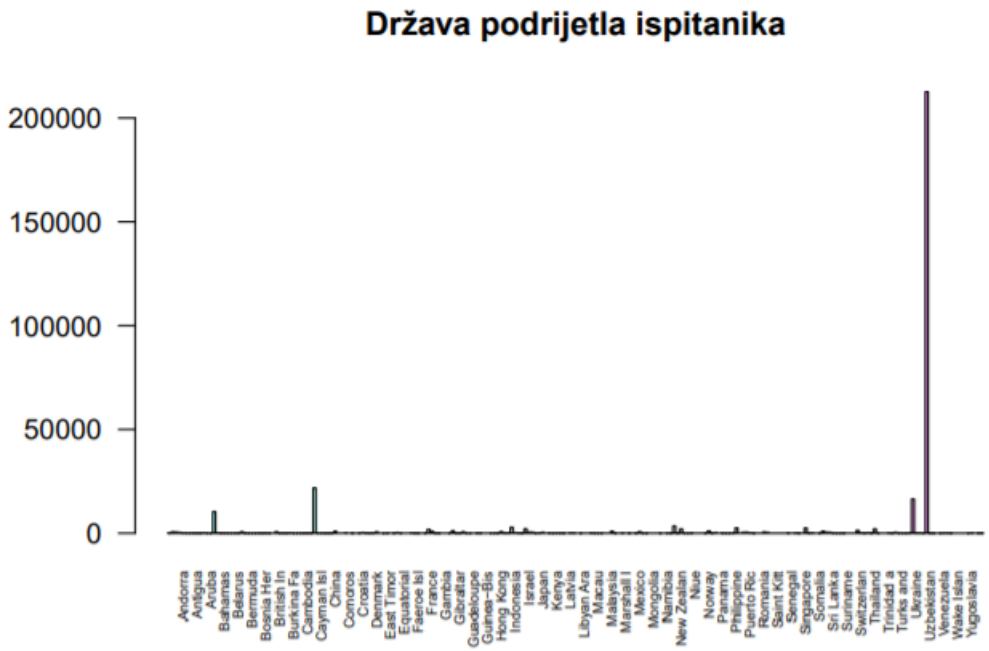

Figure 3: Graficki prikaz frekvencija podrijetla ispitanika

```
print(paste('Iz Hrvatske je ispitanika: ',nrow(BigFive[BigFive$country == c("Croatia"),])))
```

```
## [1] "Iz Hrvatske je ispitanika:  307"
```

Najviše ispitanika imamo iz Sjedinjenih Američkih Država. Također uočavamo da imamo 307 ispitanika iz Hrvatske. Histogram kojeg smo prikazali je nepregledan, te ćemo radi preglednosti daljnje vizualizacije vezane uz države prikazivati na karti svijeta. Za prikaz podataka na karti svijeta koristit ćemo R pakete "rworldmap" i "countrycode".

```
library(rworldmap,verbose = FALSE) ## učitavanje paketa za prikaz podataka na mapi svijeta
library(countrycode, verbose = FALSE) ## konverzija imena države u kod
library(RColorBrewer, verbose = FALSE) ## različite palete boja



BigFive$countrycode <- countrycode(BigFive$country, 'country.name', 'iso3c', warn=TRUE)
```

Putem R paketa "cuntrycode" modificirali smo naš podatkovni skup. Dodali smo mu još jednu varijablu: "countrycode", koja predstavlja kraticu te države. Ovaj dodatak je bio nužan da bi predstavili naše podatke na karti. Ako država nije ispravno napisana , ili više ne postoji pojavit će se NA vrijednost. Na početku smo

naglasili koju opasnost NA vrijednosti nose, stoga s njima treba oprezno postupati. Pogledajmo prisutnost nepostojećih ili neispravno napisanih država.

```
BigFive.remove = BigFive[!complete.cases(BigFive),]
BigFive.remove %>% group_by(country) %>% summarise(
          count = n())
```

```
## # A tibble: 34 x 2
##    country      count
##    <chr>        <int>
##  1 ""             172
##  2 "Arabian Gu"    32
##  3 "Borneo"        11
##  4 "British In"    21
##  5 "British Vi"    60
##  6 "Central Af"    19
##  7 "Columbia"     209
##  8 "Dominican"     75
##  9 "El Salvado"    47
## 10 "Equatorial"     1
## # ... with 24 more rows
```

Otkrili smo 34 neispravnosti kod država, to moramo ispraviti.Neke su pogrešno napisane, dok neke (npr. Jugoslavija) više ne postoje.Također uočili smo 172 zapisa koji imaju prazan zapis umjesto imena države. Prvo ćemo ispraviti značajne tipfelere, te izbrisati ostale vrijednosti jer zbog količine nisu od velike važnosti za nadolazeće testove u ovom projektu..

```
#ispravili smo nekoliko tipfelera
BigFive$country[BigFive$country == "New Zealan"] <- "New Zealand"
BigFive$country[BigFive$country == "Netherland"] <- "Netherlands"
BigFive$country[BigFive$country == "Philippine"] <- "Philippines"
BigFive$country[BigFive$country == "Puerto Ric"] <- "Puerto Rico"
BigFive$country[BigFive$country == "South Kore"] <- "South Korea"
BigFive$country[BigFive$country == "South Afri"] <- "South Africa"
BigFive$country[BigFive$country == "British Vi"] <- "British Virgin Islands"
BigFive$country[BigFive$country == "Columbia"] <- "Colombia"
BigFive$country[BigFive$country == "Saudi Arab"] <- "Saudi Arabia"
BigFive$country[BigFive$country == "Saint Kitt"] <- "Saint Kitts"
BigFive$country[BigFive$country == "United Ara"] <- "United Arab Emirates"
BigFive$country[BigFive$country == "Dominican"] <- "Dominican Republic"
BigFive$country[BigFive$country == "El Salvado"] <- "El Salvador"
BigFive$country[BigFive$country == "Arabian Gu"] <- "Arabian Gulf"
BigFive$country[BigFive$country == "North Kore"] <- "North Korea"
BigFive$country[BigFive$country == "Virgin Isl"] <- "Virgin Islands"
BigFive$country[BigFive$country == "New Caledo"] <- "New Caledonia"

BigFive <- na.omit(BigFive)

BigFive$countrycode <- countrycode(BigFive$country, 'country.name', 'iso3c', warn=TRUE)
```

Napravili smo korekcije nad našim skupom, sada ćemo prikazati kartu svijeta,i “rasprostranjenost” faktora ekstrovertizma na njoj.

```
BigFive_medijan_drzave <- aggregate(BigFive[names
                          (BigFive) %in%
                          c("agreeable_score",
```

```
                            "extraversion_score",
                            "openness_score",
                            "conscientiousness_score",
                            "neuroticism_score")  ],
                          list(BigFive$country,
                               BigFive$countrycode),
                          median)

BigFive_mean_drzave <- aggregate(BigFive[names
                          (BigFive) %in%
                          c("agreeable_score",
                            "extraversion_score",
                            "openness_score",
                            "conscientiousness_score",
                            "neuroticism_score")  ],
                          list(BigFive$country,
                               BigFive$countrycode),
                          median)



sPDF <- joinCountryData2Map(BigFive_medijan_drzave
                            ,joinCode = "ISO3"
                            ,nameJoinColumn = "Group.2"
                            ,mapResolution = "coarse"
                            , verbose = F)
```

```
## 193 codes from your data successfully matched countries in the map
## 9 codes from your data failed to match with a country code in the map
## 54 codes from the map weren't represented in your data
```

```
sPDF2 <- joinCountryData2Map(BigFive_mean_drzave
                            ,joinCode = "ISO3"
                            ,nameJoinColumn = "Group.2"
                            ,mapResolution = "coarse"
                            , verbose = F)
```

```
## 193 codes from your data successfully matched countries in the map
## 9 codes from your data failed to match with a country code in the map
## 54 codes from the map weren't represented in your data
```

```
colourPalette <- brewer.pal(7,'GnBu')



mapParams <- mapCountryData(sPDF,
                            nameColumnToPlot="extraversion_score",
                            addLegend=FALSE,
                            colourPalette=colourPalette,
                            mapTitle = "Ekstraverzija ispitanika
                            diljem svijeta (medijan)" )

do.call(addMapLegend
        ,c(mapParams
           ,legendLabels="all"
```

```
           ,legendWidth=0.5
           ,legendIntervals="data"
           ,legendMar = 2))
```

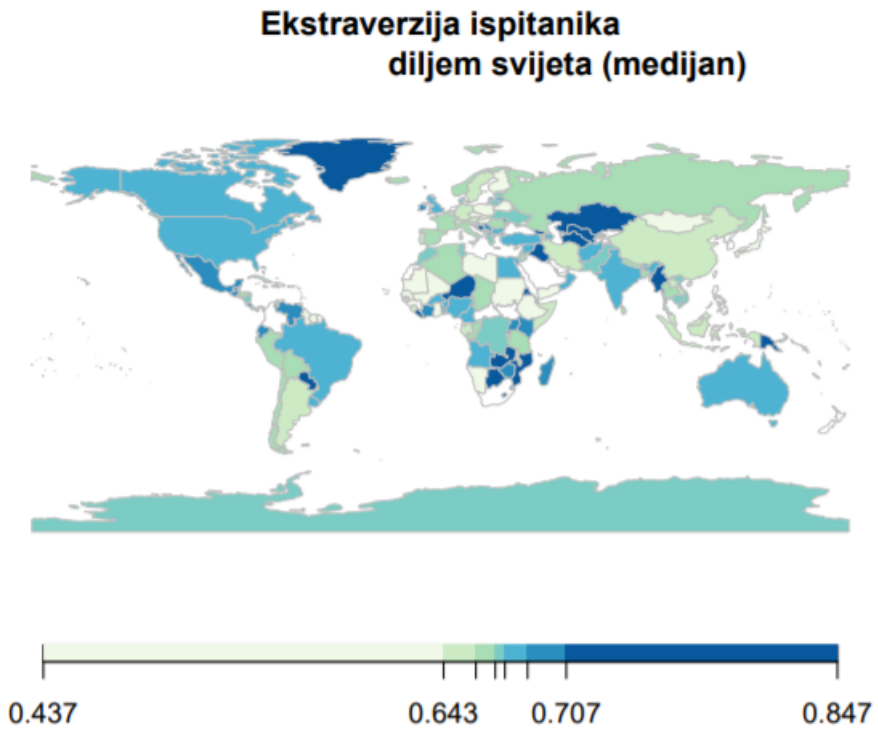

Figure 4: Geografski prikaz frekvencija medijana ekstraverzije diljem svijeta

```
mapParams2 <- mapCountryData(sPDF2,
                             nameColumnToPlot="extraversion_score",
                             addLegend=FALSE,
                             colourPalette=colourPalette,
                             mapTitle = "Ekstraverzija ispitanika
                             diljem svijeta (aritmetička sredina)" )

do.call(addMapLegend
        ,c(mapParams2
           ,legendLabels="all"
           ,legendWidth=0.5
           ,legendIntervals="data"
           ,legendMar = 2))
```

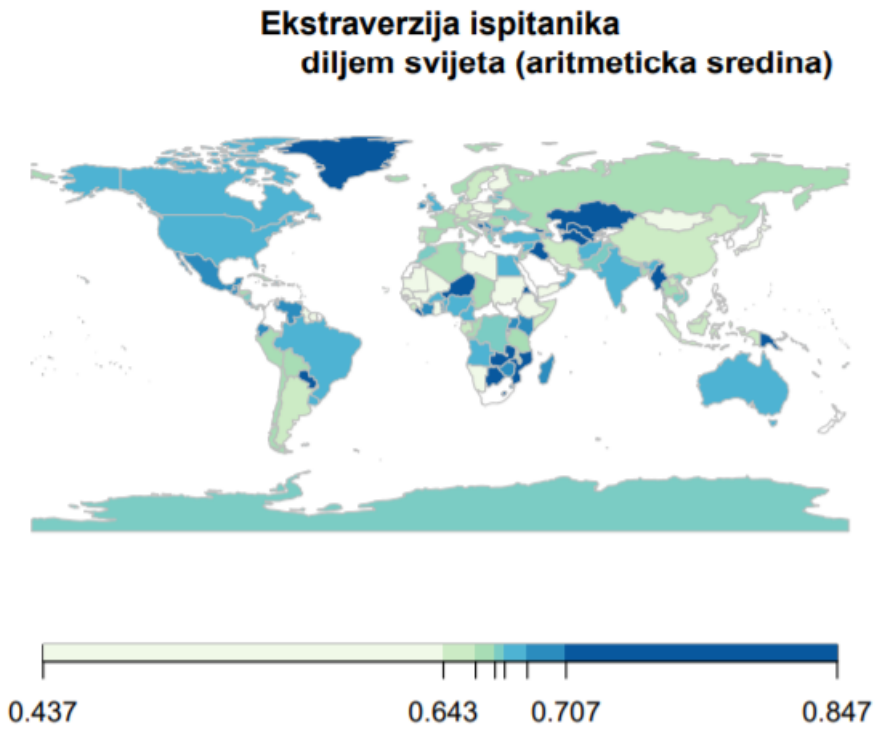

Figure 5: Geografski prikaz frekvencija aritemticke sredine faktora ekstraverzije diljem svijeta

Vizualizacije potvrđuju zdravorazumsku pretpostavku da faktor ekstraverzije nije jednak u svim državama diljem svijeta. I unutar kontinenata možemo uočiti razlike, ali ipak i po medijanu (robusnija mjera) i po aritmetičkoj sredini mogli bismo reći da je Azija introvertiranija od Amerike (i Sjeverne i Južne), a za Europu bismo mogli reći da je između. Čini se da su pojedinci iz Afrike i Australije ekstrovertiraniji od Europljana. Jasno, navedene opservacije nisu nikakvi dokazi već samo naš komentar na gore prikazanu vizualizaciju.

Postoji li razlika ekstrovertiranosti među pojedincima iz različitih država?

Na našoj karti primijetili smo razlike između Japana, Rusije i Hrvatske u faktoru ekstraverzije. Izabrali smo udaljene države različitih kultura pa pretpostavljamo da se radi i o ljudima različitih karakteristika. Pojedine kulture (države) potiču otvorenost i pristupačnost više od drugih. Svakako treba imati na umu da je svaki pojedinac neovisna i složena ličnost i da sigurno u sve 3 države ima i onih ekstremno ekstrovertiranih i onih ekstremno introvertiranih. Također njihova (ne)ekstrovertiranost nije samo produkt države iz koje su. Ipak, očekujemo da bi se srednje vrijednosti rezultata mogle razlikovati.

Mi smo odlučili ispitati hipotezu o jednakosti srednjih vrijednosti ekstrovertiranosti ovih država. Za to testiranje moramo primijeniti ANOVA test.

ANOVA (engl. *ANalysis Of VAriance*) je metoda kojom testiramo sredine više populacija. U analizi varijance pretpostavlja se 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 upravo je ustanoviti jesu li te razlike između grupa samo posljedica slučajnosti ili su statistički značajne.

ANOVA zahtijeva poznavanje sljedećih varijabli:

$$
\begin{aligned}
SST &= \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = \text{total sum of squares},\\
SSA &= n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2 = \text{treatment sum of squares},\\
SSE &= \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 = \text{error sum of squares}.
\end{aligned}
$$

$$SST = SSA + SSE$$

Figure 6: Razlicite varijabilnosti

Pri testiranju značajnosti razlike srednjih vrijednosti tretmana gledamo omjer SSA-a i SST-a. Što je taj omjer veći vjerojatnije je da razlika srednjih vrijednosti uzoraka nije slučajna.

Pretpostavke ANOVA-e su:

- nezavisnost pojedinih podataka u uzorcima
- normalna razdioba podataka
- homogenost varijanci među populacijama.

Snaga testa je bolja ako su uzorci podjednake veličine. Stoga pored želje za ispitivanjem ekstrovertiranosti različitih kultura, odlučili smo se za države koje imaju barem donekle sličan broj ispitanika, iako jednakost uzoraka nije preduvjet za ANOVU.

Mi ćemo primijeniti jednofaktorsku ANOVA-u s 3 razine, gdje je faktor država iz koje je ispitanik. Ispitajmo pretpostavke ANOVE:

Budući da se radi o različitim ispitanicima iz različitih država, uzorci su jasno nezavisni. Normalnost smijemo provjeriti qq-plotom jer je ANOVA dovoljno robusna za korištenje i kad podatci nisu potpuno normalno distribuirani.

```
par(mfrow=c(1,3))

Japan = BigFive[BigFive$country == "Japan",]
print(paste("Broj ispitanika iz Japana :",nrow(Japan)))
```

```
## [1] "Broj ispitanika iz Japana : 398"
```

```
Hrvatska = BigFive[BigFive$country == "Croatia",]
print(paste("Broj ispitanika iz Hrvatske :",nrow(Hrvatska)))
```

```
## [1] "Broj ispitanika iz Hrvatske : 307"
```

```
Rusija = BigFive[BigFive$country == "Russian Fe",]
print(paste("Broj ispitanika iz Rusije :",nrow(Rusija)))
```

```
## [1] "Broj ispitanika iz Rusije : 366"
```

```
uzorak = rbind(Japan,Rusija, Hrvatska)
uzorak1 = uzorak[, names(uzorak) %in% c("extraversion_score","country")]
uzorak1$country <- as.factor(uzorak1$country)

qqnorm(Japan$extraversion_score , pch = 1, frame=FALSE, main = 'Ekstraverzija: Japan')
qqline(Japan$extraversion_score, col ="red", lwd= 2)
```

```
qqnorm(Rusija$extraversion_score , pch = 1, frame=FALSE, main = 'Ekstraverzija: Rusija')
qqline(Rusija$extraversion_score, col ="red", lwd= 2)

qqnorm(Hrvatska$extraversion_score , pch = 1, frame=FALSE, main = 'Ekstraverzija: Hrvatska')
qqline(Hrvatska$extraversion_score, col ="red", lwd= 2)
```

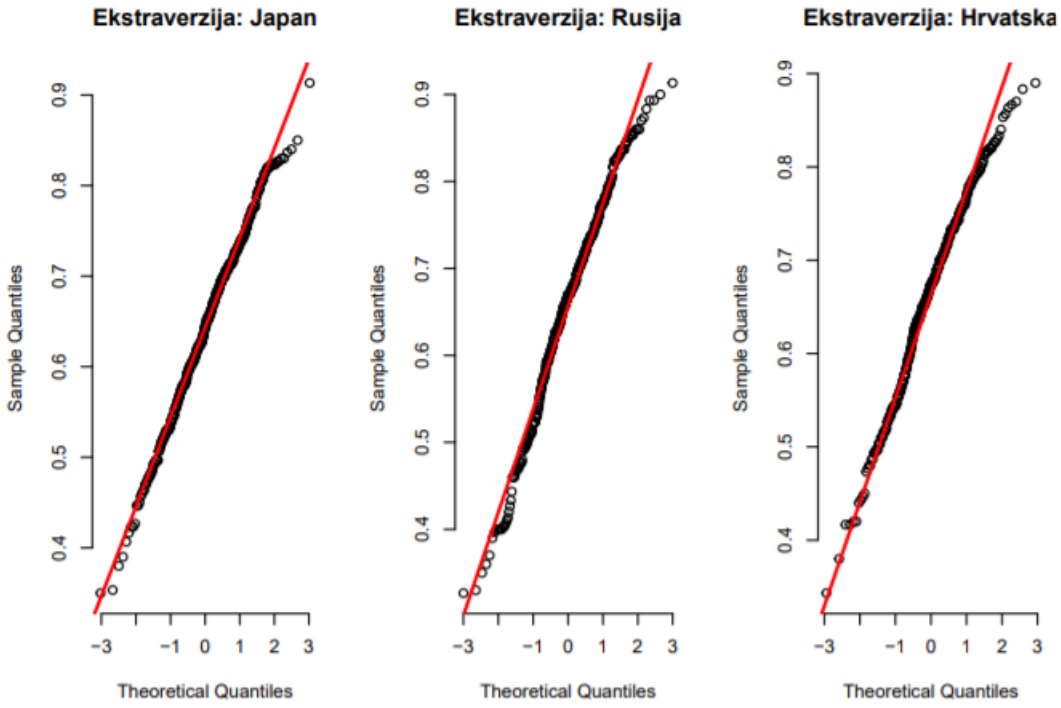

Figure 7: Graficki prikaz normalnosti faktora ekstraverzije

Kao što možemo vidjeti podatci unutar grupa ne odstupaju (pretjerano) od normalne razdiobe.

Sada ćemo provjeriti jednakost varijanci različitih populacija. Dakle, želimo testirati:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 \quad H_1 : \neg H_0$$

U našem je slučaju k=3 jer imamo 3 države. Navedenu hipotezu možemo testirati Bartletovim testom koji je u R-u implementiran naredbom `bartlett.test()`.

```
print(paste("Varijanca Rusije: ",var(Rusija$extraversion_score)))
```

```
## [1] "Varijanca Rusije:  0.0145279164275436"
```

```
print(paste("Varijanca Japana: ",var(Japan$extraversion_score)))
```

```
## [1] "Varijanca Japana:  0.0096338982907822"
```

```
print(paste("Varijanca Hrvatske: ",var(Hrvatska$extraversion_score)))
```

```
## [1] "Varijanca Hrvatske:  0.0108476636884697"
```

```
bartlett.test(uzorak1$extraversion_score ~ uzorak1$country)
```

```
##
##  Bartlett test of homogeneity of variances
##
```

```
## data:  uzorak1$extraversion_score by uzorak1$country
## Bartlett's K-squared = 17.06, df = 2, p-value = 0.0001974
```

Prema testu, odbacili bismo nultu hipotezu o jednakosti varijanci među populacijama, ali vidimo da razlika i nije tako velika. S obzirom na to da je ANOVA robusna metoda, a podatci su približno normalno distribuirani i broj elemenata u uzorcima je (više/manje) podjednak, možemo nastaviti s testiranjem.

Jednofaktorski ANOVA model glasi:

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

gdje je μ_j sredina svake populacije $j = 1, .., k$.

ANOVA-om testiramo:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k, \quad H_1 : \neg H_0.$$

Prije postavljanja gornjih hipoteza pretpostavili bismo da se rezultat Japanaca razlikuje i od Rusije i od Hrvatske. Kulture Istočne Azije cijene introspekciju, provođenje vremena u tišini i prirodi pa bismo očekivali da je faktor ekstroverzije kod njih prosječno manjih vrijednosti. Pogledajmo pravokutne dijagrame za ekstrovertiranost ovisno o faktoru - državi podrijetla.

```
boxplot(extraversion_score ~ country ,data = uzorak,
        col = c( "honeydew","lightblue","tomato1"),
        xlab = "država", ylab = "ekstraverzija",
        main ="Pravokutni dijagrami: ekstraverzija ~ država",
        names = c("Hrvatska","Japan","Rusija"))

means <- tapply(uzorak$extraversion_score,uzorak$country,mean)
m <- tapply(uzorak$extraversion_score,uzorak$country,median)

points(means,col="green",pch=18)
```

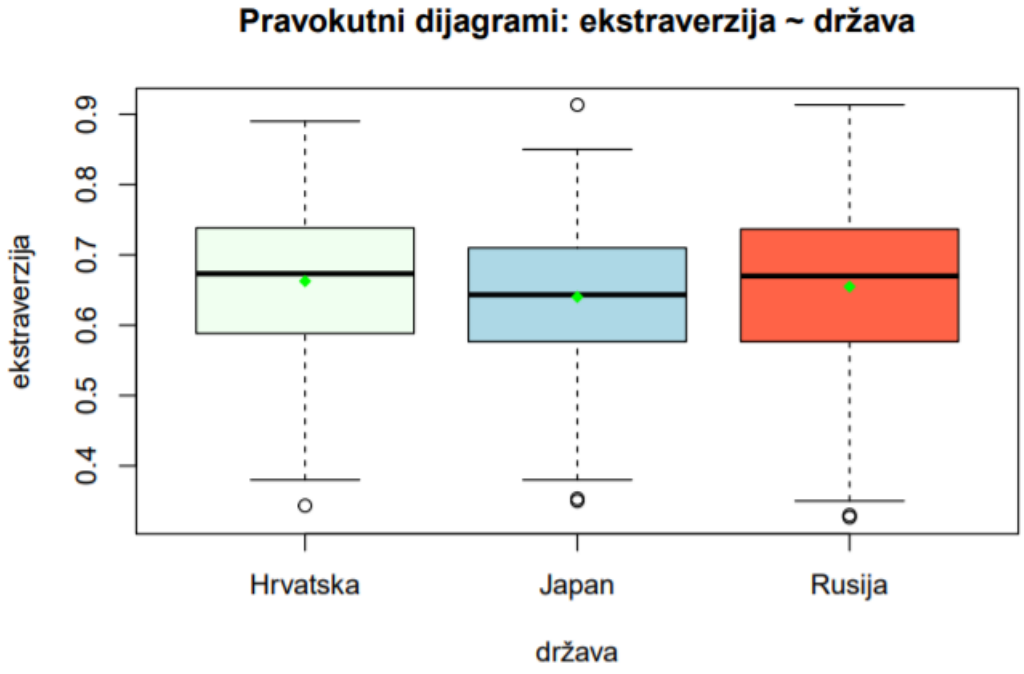

Figure 8: Graficki prikaz odnosa države prema faktoru ekstraverzije

Već na prvi pogled vidimo razliku u ekstraverziji ove tri države. Sada moramo provesti ANOVA-u kako bi sa sigurnošću prihvatili našu pretpostavku.

```
a = aov(uzorak1$extraversion_score ~ uzorak1$country)
anova(a)
```

```
## Analysis of Variance Table
##
## Response: uzorak1$extraversion_score
##                    Df  Sum Sq  Mean Sq F value  Pr(>F)
## uzorak1$country     2  0.0933 0.046652   4.003 0.01853 *
## Residuals        1068 12.4467 0.011654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Što je f-statistika (omjer sume kvadrata tretmana i sume kvadrata greške) veća to je veća vjerojatnost da razlika u rezultatu ekstrovertizma nije slučajna već je prisutna zbog različitosti faktora (država). Na razini značajnosti 5% možemo odbaciti H_0 hipotezu o jednakosti ekstrovertiranosti između država. Ipak na razini značajnosti 1% to ne bismo mogli zato s ovim zaključcima treba biti oprezan.

Na temelju rezultata ANOVA-e zaključili smo da prosječni rezultati za ove 3 države nisu jednaki, ali ne možemo reći koja država odstupa od koje. To je glavni nedostatak ANOVA testa. Sada ćemo odrediti da li se Hrvatska bitno razlikuje od Japana i Rusije po pitanju ekstraverzije.

Za usporedbu faktora možemo koristiti kontraste u sredinama tretmana, linearne funkcije oblika

$$\omega = \sum_{i=1}^{k} c_i \mu_i$$

za koju vrijedi:

$$\omega = \sum_{i=1}^{k} c_i = 0$$

.

Imamo 3 faktora pa je za usporedbu moguće uzeti 2 ortogonalna kontrasta:Usporedimo li prvu državu s preostale dvije i drugu državu sa trećom ,mi smo indirektno usporedili i treću s preostalima.

U R-u je kao temeljeni (*engl. baseline*) kontrast postavljena Hrvatska i R radi prvo usporedbu Hrvatske i Japana,a onda Hrvatske i Rusije.

```
contrasts(uzorak1$country)
```

```
##            Japan Russian Fe
## Croatia        0          0
## Japan          1          0
## Russian Fe     0          1
```

```
summary.lm(aov(uzorak1$extraversion_score ~ uzorak1$country))
```

```
##
## Call:
## aov(formula = uzorak1$extraversion_score ~ uzorak1$country)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32812 -0.06973  0.00985  0.07521  0.27318
##
## Coefficients:
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.662638   0.006161 107.548   <2e-16 ***
## uzorak1$countryJapan        -0.022488   0.008200  -2.742   0.0062 **
## uzorak1$countryRussian Fe -0.007848   0.008355  -0.939   0.3478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.108 on 1068 degrees of freedom
## Multiple R-squared:  0.007441,   Adjusted R-squared:  0.005582
## F-statistic: 4.003 on 2 and 1068 DF,  p-value: 0.01853
```

Slobodan član (*engl. intercept*) označava srednju vrijednost ekstrovertizma hrvatskih ispitanika, a faktori uz Japan i Rusiju razliku između tih zemalja i Hrvatske. Vidimo da su hrvatski ispitanici (statistički) značajno ekstrovertiraniji od Japanaca, a između Rusa i Hrvata nema statistički značajne razlike.

Želimo li usporediti Japnace s preostalim tretmanima(državama, točnije sa Rusijom i Hrvatskom), promijenimo matricu kontrasta i pogledajmo ima li značajnih razlika u faktoru ekstrovertizma.

```
contrastmatrix<-cbind(c(1,0,0),c(0,0,1))
contrasts(uzorak1$country) <- contrastmatrix
contrasts(uzorak1$country)
```

```
##            [,1] [,2]
## Croatia       1    0
## Japan         0    0
## Russian Fe    0    1
```

```
summary.lm(aov(uzorak1$extraversion_score ~ uzorak1$country))
```

```
##
## Call:
## aov(formula = uzorak1$extraversion_score ~ uzorak1$country)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32812 -0.06973  0.00985  0.07521  0.27318
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.640151   0.005411 118.299   <2e-16 ***
## uzorak1$country1 0.022488   0.008200   2.742   0.0062 **
## uzorak1$country2 0.014640   0.007818   1.873   0.0614 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.108 on 1068 degrees of freedom
## Multiple R-squared:  0.007441,   Adjusted R-squared:  0.005582
## F-statistic: 4.003 on 2 and 1068 DF,  p-value: 0.01853
```

Kao što smo već vidjeli Hrvati(uzorak1$country1) su značajno ekstrovertiraniji od Japanaca, a razlika između Japanaca i Rusa je statistički značajna tek uz vjerojatnosti pogreške 5 %.

Možemo zaključiti da su hrvatski ispitanici značajno ekstrovertirani u odnosu na Japance, a da među Rusima i Japancima nema jako značajne razlike (iako manja razlika u podatcima je primjetna). Također među Hrvatima i Rusima nema značajne razlike u faktoru ekstrovertizma.

Postoji li razlika ekstrovertizma ovisno o spolu i državi iz kojih potječe pojedinac?

Pokazali smo razliku ekstrovertiranosti između Hrvatske, Rusije i Japana. Sada ćemo ispitati ekstrovertiranost preko faktora narodnosti i spola ispitanika. Za ovo testiranje provodimo dvofaktorsku ANOV-u

Potrebno je testirati hipoteze:

- H_0': prvi faktor (država podrijetla) je beznačajan
- H_0'': drugi faktor (spol) je beznačajan
- H_0''': nema interakcije između države i spola osobe

Normalnost ekstrovertizma po državama je već provjerena, stoga nam sada ostaje provjeriti samo normalnost distribucije faktora ekstrovertiranosti po spolovima u uzorku (populaciji).

```
par(mfrow=c(1,2))

qqnorm(uzorak[uzorak$sex == 1,]$extraversion_score , pch = 1, frame=FALSE, main = 'Ekstraverzija: muški
qqline(uzorak[uzorak$sex == 1,]$extraversion_score, col ="red", lwd= 2)

qqnorm(uzorak[uzorak$sex == 2,]$extraversion_score , pch = 1, frame=FALSE, main = 'Ekstraverzija: ženski
qqline(uzorak[uzorak$sex == 2,]$extraversion_score, col ="red", lwd= 2)
```

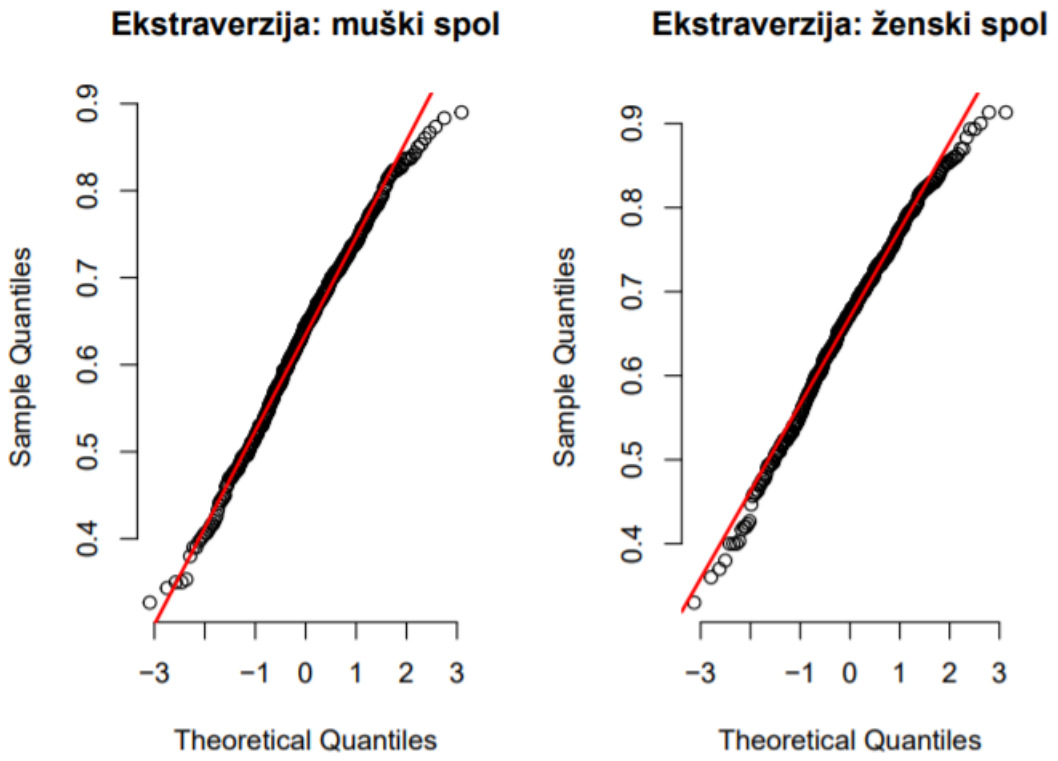

Figure 9: Graficka provjera normalnosti varijabli

Možemo reći da je odstupanje od normalne razdiobe zanemarivo.

```
#par(mfrow=c(2,3))
Japan = BigFive[BigFive$country == "Japan",]
Hrvatska = BigFive[BigFive$country == "Croatia",]
Rusija = BigFive[BigFive$country == "Russian Fe",]
```

```
uz = rbind(Japan, Rusija, Hrvatska)
uz$sex = factor(uz$sex,levels = c(1,2),labels = c('muški','ženski'))
uz$country <- as.factor(uz$country)
```

Upoznali smo se sa raspodijelom ekstraverzija po državama, sada pogledajmo ima li razlike ekstraverzije među muškarcima i ženama.

```
boxplot(extraversion_score ~ sex, data=uz,
        col = c("red","blue"), xlab = "spol",
        ylab = "faktor ekstraverzije",
        main ="Ovisnost ekstraverzije o spolu",
        names = c("muškarci","žene"))
means <- tapply(uz$extraversion_score,uz$sex,mean)
points(means,col="green",pch=18)
```

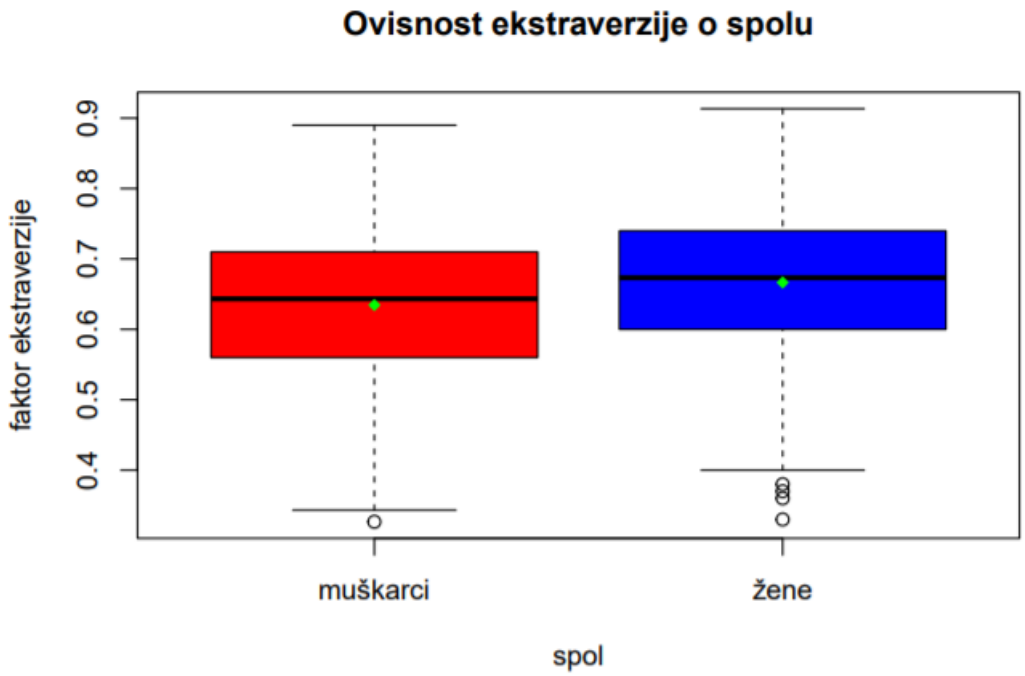

Figure 10: Graficki prikaz odnosa spola prema faktoru ekstraverzije

Vidimo da postoji razlika u veličini faktora ekstraverzije osoba različitih spolova, žene se čine ekstrovertiranijima.

Box-plot dijagramom ćemo prikazati ovisnost ekstraverzije o interakciji države i spola.

```
interakcija = interaction(uz$country,uz$sex)

boxplot(uz$extraversion_score ~ interakcija, cex.axis=0.5,
        col = c("brown2","yellow","lightgreen", "darkred", "gold4", "darkgreen"),
        xlab = "interakcija", ylab = "faktor ekstraverzije",
        main ="Ovisnost ekstraverzija o interakciji države i spola",
        names = c("Hrvatska-muškarci","Japan-muškarci","Rusija-muškarci",
                  "Hrvatska-žene","Japan-žene","Rusija-žene"))
```

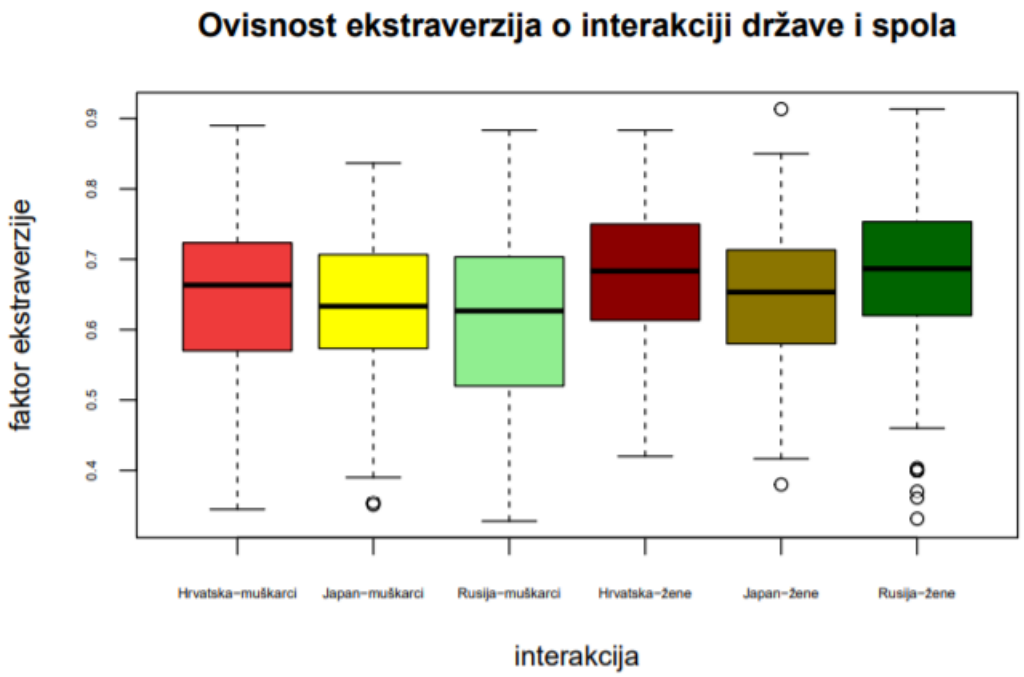

Figure 11: Graficki prikaz ovisnosti ekstraverzije o interakciji spola i države

Vidljive su razlike među novim podijeljenim skupovima, primjerice Japanke su manje ekstrovertirane od Ruskinja i Hrvatica. Sada ćemo putem Bartlettova testom provjeriti homogenost varijanci.

```
bartlett.test(uz$extraversion_score ~ interakcija)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  uz$extraversion_score by interakcija
## Bartlett's K-squared = 15.052, df = 5, p-value = 0.01014
```

Na razini značajnosti 5 % možemo reći da varijance prema testu nisu jednake, no ako računamo na razini značajnosti 1 %, varijance ispadaju jednake. Stoga, možemo reći da se varijance razlikuju, ali ta je razlika dovoljno malena da ne naruši pretpostavke modela ANOVA.

Dvofaktorski ANOVA test ima iste pretpostavke kao jednofaktorski, uz zahtjev na jednake veličine uzoraka pojedinih grupa (populacija). To u praksi najčešće nije slučaj, pa se koriste verzije s otežanim srednjim vrijednostima. U R-u je upravo takav pristup defaultni u funkciji `aov()` pa možemo nastaviti s našim testiranjem.

```
uz = rbind(Japan[Japan$sex == "1",],Japan[Japan$sex == "2",], Rusija[Rusija$sex == "1",],
           Rusija[Rusija$sex == "2",], Hrvatska[Hrvatska$sex == "1",],Hrvatska[Hrvatska$sex == "2",])
spolJapan = factor(rep(c("M","Ž"), c(185, 213)))
spolRusija = factor(rep(c("M","Ž"), c(154, 212)))
spolHrvatska = factor(rep(c("M","Ž"), c(161, 146)))
faktor <- levels(spolJapan)[c(spolJapan, spolRusija, spolHrvatska)]
spolFaktor <- factor(x = faktor, levels = levels(spolJapan))
drzava = factor(rep(c('Japan', 'Rusija', 'Hrvatska'), c(398, 366, 307)))
means = tapply(uz$extraversion_score, INDEX = list(spolFaktor, drzava) , FUN = mean)
means
```

```
##    Hrvatska     Japan    Rusija
```

```
## M 0.6482195 0.6355676 0.6189610
## Ž 0.6785388 0.6441315 0.6808176
```

```
matplot(means, type = "b", xlab = "Spol", ylab = "Koeficijent ekstraverznosti",
        main = "Grafički prikaz interakcije države i spola")
```

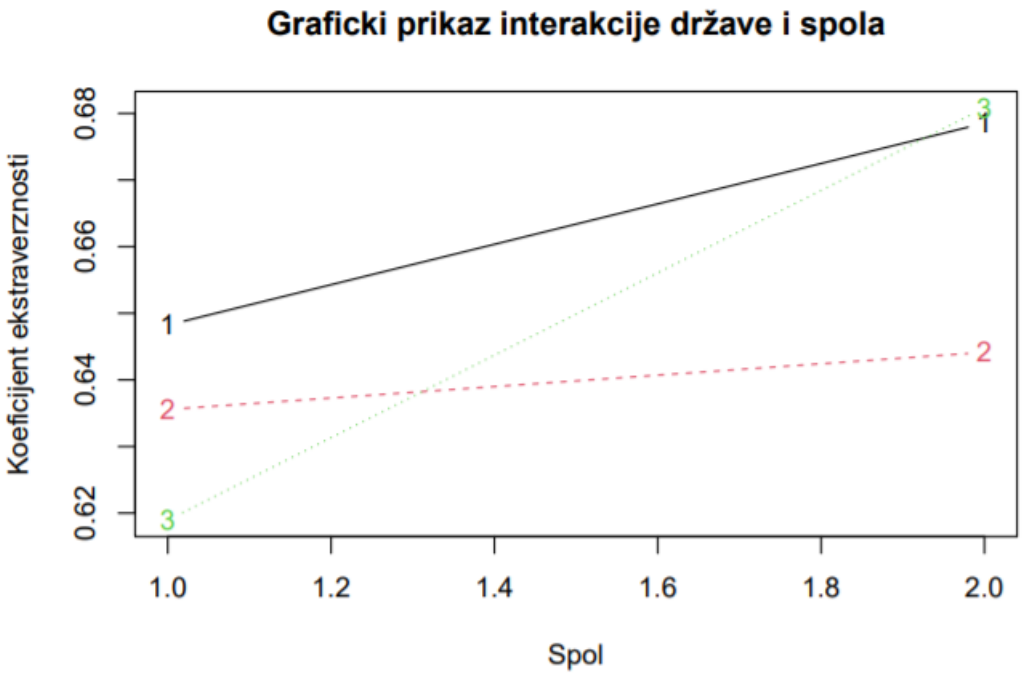

Figure 12: Graficki prikaz interakcije države i spola

Na temelju grafa zaključujemo da je prisutna interakcija spola i dobi, linije nisu paralelne. Ipak, u sve tri je države veći faktor ekstraverzije za žene (2), razlika je najveća za Rusiju (država 3), a najmanja za Japan (država 2). Rusi su manje ekstrovertirani od Japanaca i Hrvata, a Ruskinje su ekstrovertiranije i od Japanki i od Hrvatica.

```
matplot(t(means), type = "b", xlab = "Država", ylab = "Koeficijent ekstraverznosti",
        main = "Grafički prikaz interakcije spolova s državama")
```

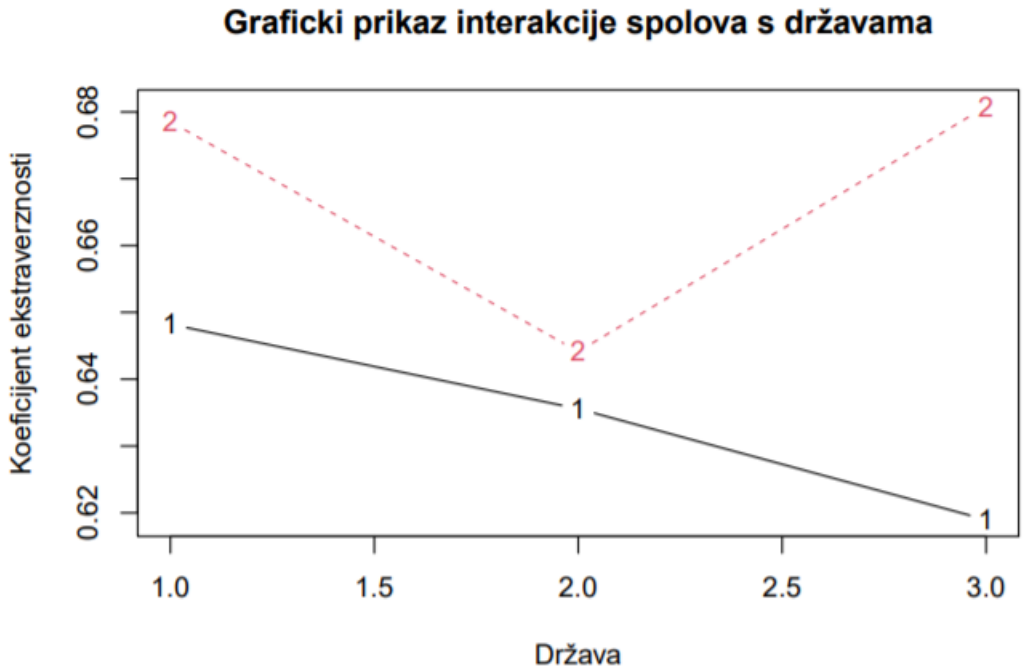

Figure 13: Graficki prikaz interakcije spolova s državama

Iz grafa je jasno vidljivo zbog nakrivljenosti linija na grafu da interackija spola i države podrijetla osobe utječe na razliku ekstraverzija pojedinaca. U Japanu (država 2) je ekstrovertiranost žena (2) značajno manja nego u druge dvije države, a u Rusiji (država 3) je ekstrovertiranost muškaraca (1) značajno manja.

```
model = lm(uz$extraversion_score ~ spolFaktor * drzava)
summary(model)
```

```
##
## Call:
## lm(formula = uz$extraversion_score ~ spolFaktor * drzava)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35082 -0.07081  0.00777  0.07253  0.26920
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.648219   0.008375  77.396   <2e-16 ***
## spolFaktorŽ                0.030319   0.012145   2.496   0.0127 *
## drzavaJapan               -0.012652   0.011454  -1.105   0.2696
## drzavaRusija              -0.029258   0.011978  -2.443   0.0147 *
## spolFaktorŽ:drzavaJapan   -0.021755   0.016173  -1.345   0.1789
## spolFaktorŽ:drzavaRusija   0.031537   0.016556   1.905   0.0571 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1063 on 1065 degrees of freedom
## Multiple R-squared:  0.04085,    Adjusted R-squared:  0.03635
## F-statistic: 9.072 on 5 and 1065 DF,  p-value: 1.85e-08
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: uz$extraversion_score
##                      Df  Sum Sq  Mean Sq F value    Pr(>F)
## spolFaktor            1  0.2734 0.273368 24.2054 1.003e-06 ***
## drzava                2  0.1050 0.052476  4.6465  0.009790 **
## spolFaktor:drzava     2  0.1339 0.066970  5.9298  0.002748 **
## Residuals          1065 12.0278 0.011294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rezultati analize varijance (engl. *Analysis of Variance Table*) pokazuju nam da faktori spola i države (i pojedinačno i u interakciji) značajno utječu na vrijednosti ekstrovertiranosti ljudi.

Odnos starijih i mlađih ispitanika

Donijeli smo par zaključaka na temelju države podrijetla nekog ispitanika. Sada ćemo se osvrnuti na dob naših ispitanika, te pokušati na temelju te varijable donijeti par zaključaka. Započetak pogledajmo kako je naš podatkovni skup raspoređen prema starosti ispitanika.

```
hist(BigFive$age,main='Dob ispitanika', ylim = c(0,150000),xlim = c(0,100),xlab='Dob',
     ylab='Frekvencija',col = blues9)
```

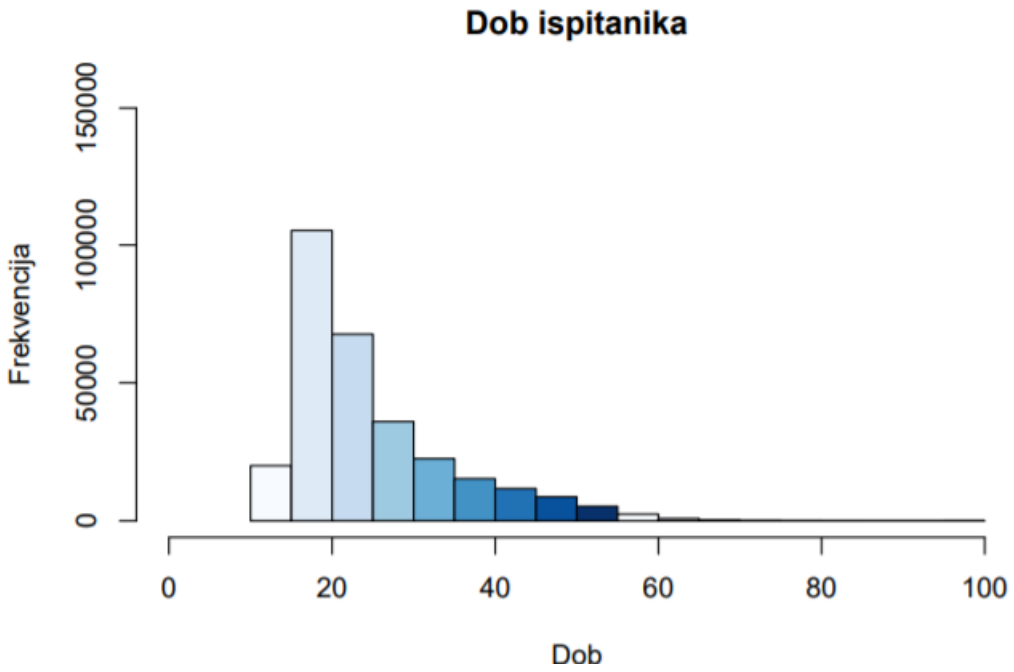

Figure 14: Grafički prikaz dobi ispitanika

```
boxplot(BigFive$age,
        main='Pravokutni dijagram: Dob Ispitanika',
        ylab='Frekvencija',col = cm.colors(1))
```

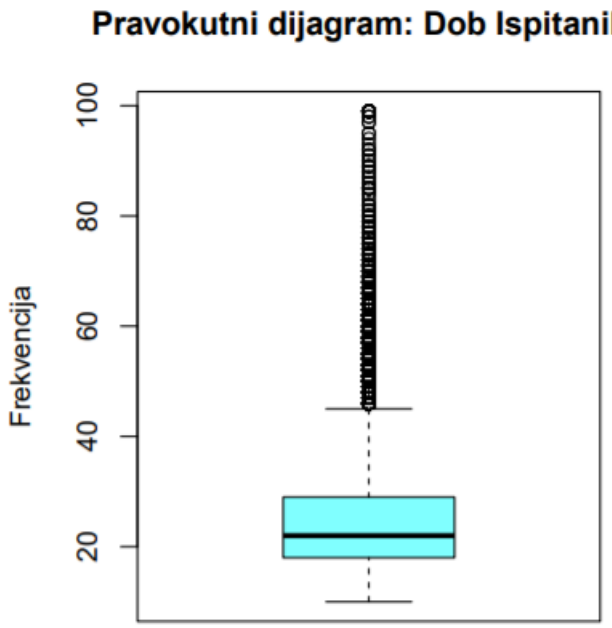

Figure 15: Pravokutni dijarama: Dob ispitanika

Vidljivo je da je dominantan broj mladih ispitanika. Osim toga, na pravokutnom dijagramu vidimo da je velika prisutnost outliera. Ova činjenica je od velike važnosti, i ubrzo ćemo se osloniti na nju. Zbog nadolazećih ispitivanja, podijelili smo naš skup na dva dijela : mladi (ispod 35 godina) i stariji (35 godina i iznad) Pogledajmo zastupljenost mladih i starijih u našem skupu

```
## [1] "Broj ispitanika koji su ispod 35 godina starosti : 247825"
```

```
## [1] "Broj ispitanika koji su iznad 35 godina starosti : 48719"
```

Broj mladih ispitanika čini približno 84% svih naših zapisa.

Svima nama je dobro poznat sukob stariji i mladi ljudi prvenstveno oko životnih vrijednosti. Nije rijetko vidjeti starije osobe kako kritiziraju mlade zbog zatvorenosti i nesklonosti novim iskustvima. Tom sukobu ćemo stati na kraj, jer u nastavku ćemo provesti statističke testove i dati konačan odgovor (na temelju ovog podatkovnog skupa). Pretpostavit ćemo da su mladi podjednako otvoreni novim iskustvima kao i stariji, protivnom da su mladi više otvorenije osobe. Odnosno:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

, gdje μ_1 predstavlja srednju vrijednost otvorenosti mladih , a μ_2 starijih ispitanika. Oslanjajući se na činjenicu da imamo dva uzorka iz dvije različite dobne skupine, možemo prihvatiti nezavisnost naših skupova. Sada moramo provjeriti normalnost skupova. To ćemo učiniti putem histograma ,qq-plota te KS-testom (ili Lillieforsovom inačicom tog testa))

```
m<-mean(youngerPeople$openness_score);std<-sqrt(var(youngerPeople$openness_score))

H= hist(youngerPeople$openness_score,
     main = 'Histogram otvorenosti za mlade osobe',
    xlab='Mjera otvorenosti', ylab="Frekvencija" ,prob = TRUE)
curve(dnorm(x, mean=m, sd=std), col="darkblue", lwd=2, add=TRUE)
```

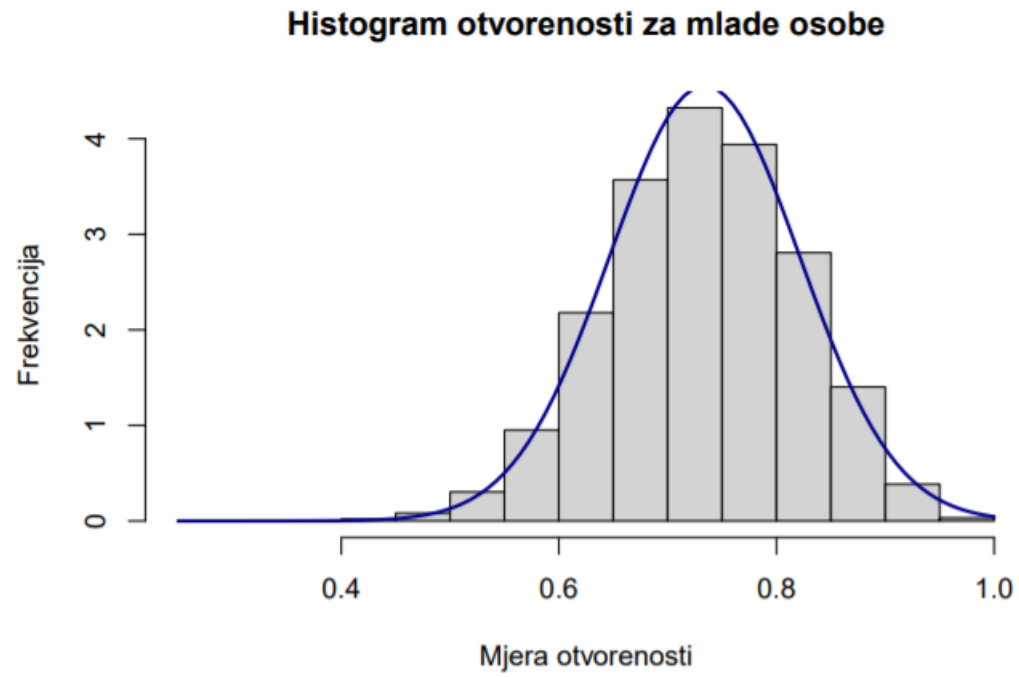

Figure 16: Histogram otvorenosti za mlade ljude

```
m<-mean(olderPeople$openness_score);std<-sqrt(var(olderPeople$openness_score))
hist(olderPeople$openness_score,
     main = 'Histogram otvorenosti za starije osobe',
    xlab='Mjera otvorenosti',ylab="Frekvencija",prob = TRUE)
curve(dnorm(x, mean=m, sd=std), col="darkblue", lwd=2, add=TRUE)
```

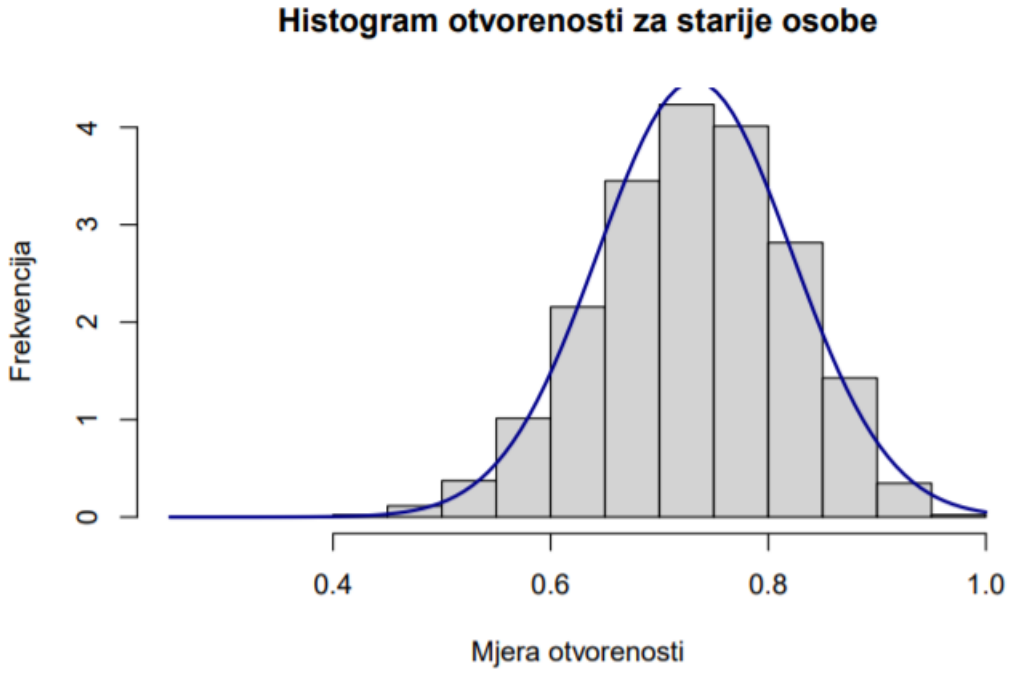

Figure 17: Histogram otvorenosti za starije ljude

Histogram upućuje na normalnost podataka (premda ima ekstremnih vrijednosti), ali on nije dovoljan da prihvatimo normalnost podataka. Stoga ćemo se osloniti na qq-plot te Lilliefors test.

```
qqnorm(youngerPeople$openness_score , pch = 1, frame=FALSE, main = 'Mlade osobe')
qqline(youngerPeople$openness_score, col ="red", lwd= 2)
```

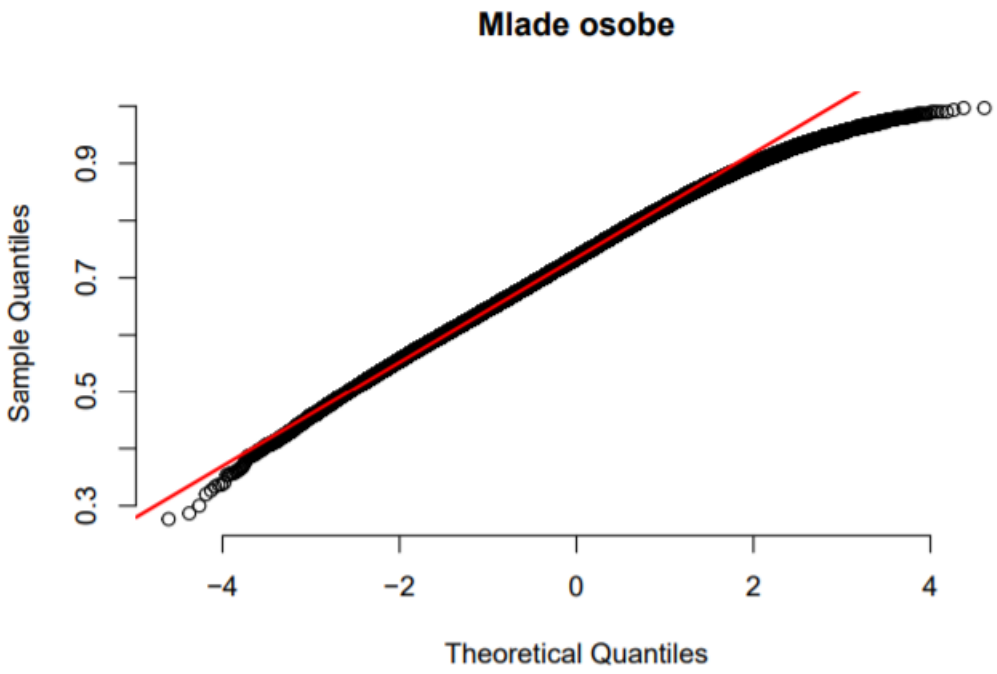

Figure 18: Graficki prikaz normalnosti podataka za mlade ljude

```
qqnorm(olderPeople$openness_score , pch = 1, frame=FALSE, main = 'Starije osobe')
qqline(olderPeople$openness_score, col ="red", lwd= 2)
```

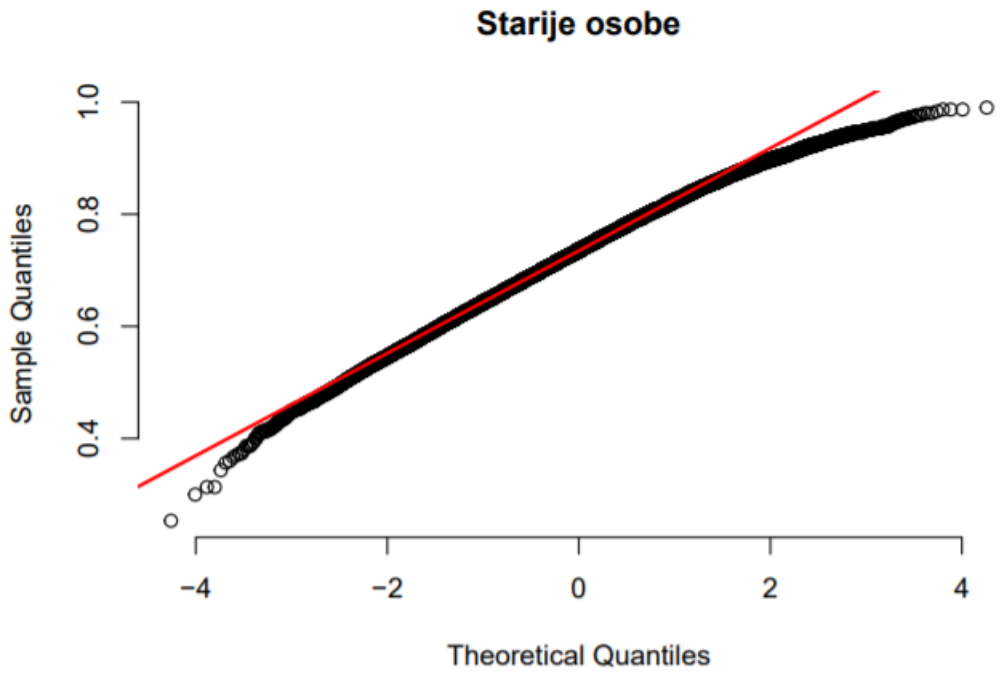

Figure 19: Graficki prikaz normalnosti podataka za starije ljude

```
require(nortest)

lillie.test(youngerPeople$openness_score )
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  youngerPeople$openness_score
## D = 0.019414, p-value < 2.2e-16
```

```
lillie.test(olderPeople$openness_score)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  olderPeople$openness_score
## D = 0.024188, p-value < 2.2e-16
```

Lilliefors-test koristimo kako bismo provjerili je li distribucija podataka normalna. Ovaj test je korekcija KS-testa te se može koristiti kada nije poznata varijanca i očekivanje populacije pa se procjenjuje iz uzorka. Naši podaci ne zadovoljavaju kriteriji ovog testa, ali prisjetimo se da smo putem box-plota pokazali prisustvo outliera. Naime, u podatkovnom skupu je puno više mlađih ispitanika pa ne čudi da su stariji outlieri. Na temelju toga možemo objasniti zašto naši podaci ne ispunjavaju kriterije ovog testa. Budući da je T-test robustan na pretpostavku normalnosti ipak ga smijemo provesti.

Hipoteze smo uspješno postavili, a sada moramo detaljno predstaviti koji test koristimo i zašto.

Imamo dva nezavisna slučajna uzorka, točnije X_1, X_2 koji dolaze iz normalnih distribucija s očekivanjima μ_1 i μ_2 , te s nepoznatim varijancama σ. Ukoliko su varijance jednake , krećemo s računanjem zajedničke disperzije uzorka. Ona se računa kao težinska sredina disperzija S_{X_1} i S_{X_2}:

$$S_X^2 = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2].$$

Pokazali smo da radimo s velikom veličinom podataka, zbog toga moramo koristiti t distribuciju. U tu svrhu koristimo sljedeću statistiku :

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_X \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

T statistika ima t distribuciju s $n_1 + n_2 - 2$ stupnja slobode.

Ukoliko imamo 2 nezavisno normalo distribuirana uzorka, ali ovoga puta sa različitim varijancama, tada koristimo testnu statistiku

$$T' = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2_{X_1}}{n_1} + \frac{s^2_{X_2}}{n_2}}}$$

koja ima aproksimativnu t-distribuciju sa stupnjevima slobode

$$v = \frac{(s^2_{X_1}/n_1 + s^2_{X_2}/n_2)^2}{(s^2_{X_1}/n_1)^2/(n_1 - 1) + (s^2_{X_2}/n_2)^2/(n_2 - 1)}$$

gdje je

$$s^2_{X_i} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_i^j - \bar{X}_i)^2$$

za $i = 1, 2$. Za testiranje jednakosti varijanci koristimo slučajnu varijablu:

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

koja ima Fisherovu distribuciju s $(n_1 - 1, n_2 - 1)$ stupnjeva slobode, pri čemu vrijedi:

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

U programskom paketu R test o jednakosti varijanci je implementiran u funkciji `var.test()`, koja prima uzorke iz dvije populacije čije varijance uspoređujemo. Test o jednakosti srednjih vrijednosti dvije populacije u R-u je implementiran u funkciji `t.test()`.

Upoznali smo se s prirodom našeg T i F testa. Provjerili pretpostavke testa, sada ispitajmo jednakost varijanci. Hipoteze testa jednakosti varijanci glase:

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

Već smo opisali prirodu testa za jednakost varijanci, sada ćemo provesti taj test putem funkcije `var.test()`, koja prima uzorke iz dvije populacije čije varijance uspoređujemo.

```
##
##  F test to compare two variances
##
## data:  youngerPeople$openness_score and olderPeople$openness_score
## F = 0.96457, num df = 247824, denom df = 48718, p-value = 2.354e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9513809 0.9778817
## sample estimates:
## ratio of variances
##          0.9645713
```

P-vrijednost je približno jednaka nuli što znači da odbacujemo H_0 , odnosno varijance naših uzoraka nisu jednake.

Napočetku smo postavili hipoteze o rezultatu otvorenosti maldih i starijih, i upoznali se s testom kojeg ćemo koristiti. Sada provodimo taj test preko funkcije t.test() uz pretpostavku nejednakosti varijanci.

```
# Bitan je poredak kojim funkciji 't.test()' prosljedjujemo uzorke!
t.test(youngerPeople$openness_score, olderPeople$openness_score, alt = "greater",
       var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  youngerPeople$openness_score and olderPeople$openness_score
## t = 3.1853, df = 68462, p-value = 0.0007233
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.0006790285          Inf
## sample estimates:
## mean of x mean of y
## 0.7339531 0.7325490
```

P-vrijednost iznosi svega : 0.009592, te s obzirom na takvo malu vrijednost možemo odbaciti H_0 hipotezu o jednakosti prosječnih vrijednosti u korist H_1.

Drugim riječima možemo reći da su mladi ljudi u prosjeku više otvoreni prema novim iskustvima od starijih osoba.

Razlike osobnosti ovisno o spolu ispitanika

Pokazali smo razliku u otvorenosti među dobnim skupinama, sada želimo pokazati utjecaj spola na pojedine varijable ličnosti.

Kao što je to bio slučaj između mladih i stariji ispitanika, postoje mnoge predrasude o ljudima koje se temelje isključivo na spolu. Kroz par testova probat ćemo ustanoviti da li zapravo postoji povezanost između pojedinih karakternih crta i spola ispitanika. Započetak moramo podijeliti naš skup prema spolu ispitanika, te pregledati odnos ta dva skupa.

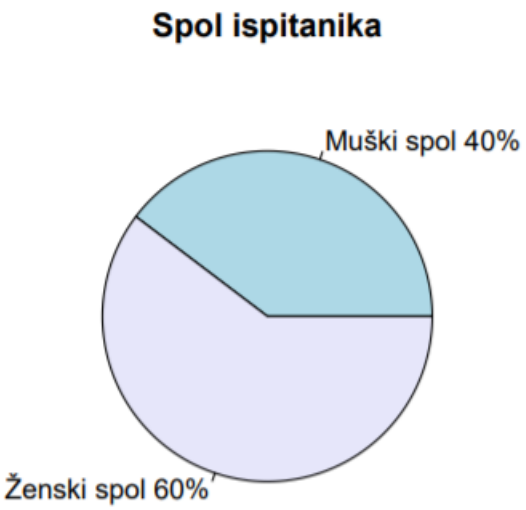

Figure 20: Graficki prikaz odnosa muških i ženskih ispitanika

Vidimo da naš dataset ima više ženskih ispitanika. Sada pogledajmo kako se ponašaju sve varijable ličnosti, ako podijelimo naš skup po spolu. Usporedit ćemo naše podatke pomoću medijana , aritmetičke sredine i podrezane aritmetičke sredine.

```
BF_copy = data.frame(BigFive) #kopiramo podatke kako ne bi mijenjali podatke originalnom
                              #datasetu
BF_copy[BF_copy$sex == 1,]$sex <- "muški spol"
BF_copy[BF_copy$sex == 2,]$sex <- "ženski spol"
aggregate(BF_copy[names(BF_copy) %in%
  c("agreeable_score","extraversion_score","openness_score","conscientiousness_score",
    "neuroticism_score")]
  , list(BF_copy$sex), mean)
```

```
##       Group.1 agreeable_score extraversion_score openness_score
## 1  muški spol       0.6705354          0.6617084      0.7229063
## 2 ženski spol       0.7142377          0.6792915      0.7408418
##   conscientiousness_score neuroticism_score
## 1               0.6960056         0.5443898
## 2               0.7060118         0.5948017
```

```
aggregate(BF_copy[names(BF_copy) %in%
  c("agreeable_score","extraversion_score","openness_score","conscientiousness_score",
    "neuroticism_score")  ]
  , list(BF_copy$sex), median)
```

```
##       Group.1 agreeable_score extraversion_score openness_score
## 1  muški spol       0.6766667          0.6700000      0.7233333
## 2 ženski spol       0.7200000          0.6866667      0.7433333
##   conscientiousness_score neuroticism_score
## 1               0.6966667         0.5366667
## 2               0.7100000         0.5900000
```

```
aggregate(BF_copy[names(BF_copy) %in%
    c("agreeable_score","extraversion_score","openness_score","conscientiousness_score",
      "neuroticism_score")],
    list(BF_copy$sex), mean,trim=0.1)
```

```
##       Group.1 agreeable_score extraversion_score openness_score
## 1  muški spol       0.6745997          0.6652697      0.7241769
## 2 ženski spol       0.7185825          0.6827667      0.7420128
##   conscientiousness_score neuroticism_score
## 1               0.6970492         0.5413303
## 2               0.7084016         0.5933174
```

Na prvi pogled ne uočavamo jako veliku razliku u rezultatima muškaraca i žena . Najveća razlika je u faktorima: ugodnost i neuroticizam što nije veliko iznenađenje uzevši u obzir predrasudu da se muškarci bolje nose sa stresom i da su žene empatičniji spol. Prije nego što ispitamo zavisnost spola te faktora ugodnosti i neuroticizma , pogledajmo bolje ponašanje ovih vrijednosti.

```
par(mfrow=c(1,2))
boxplot(neuroticism_score ~ sex,data=BigFive, col = c("lightblue", "lavender"),
        names = c("muški","ženski"), xlab = "spol", ylab = "neuroticizam")
boxplot(agreeable_score ~ sex,data=BigFive, col = c("lightblue", "lavender"),
        names = c("muški","ženski"), xlab = "spol", ylab = "ugodnost")
```

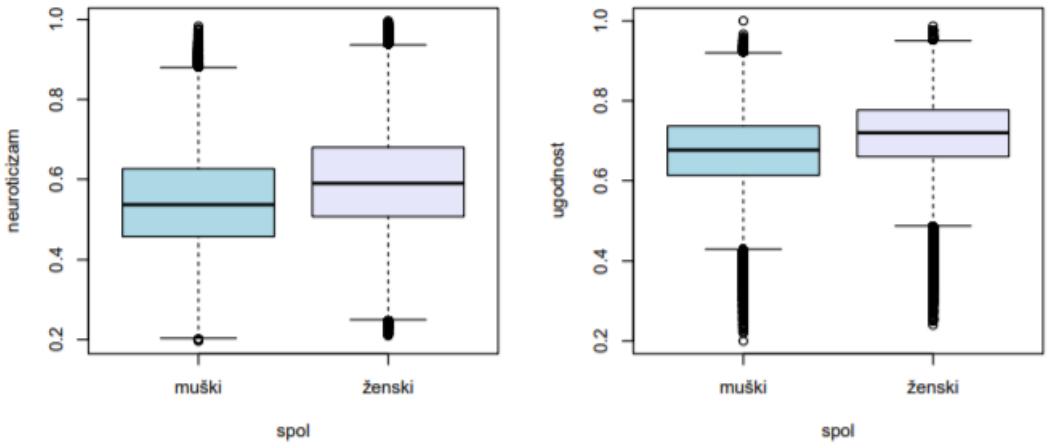

Figure 21: Razlicitost utjecaja spolova na faktore licnosti

```
par(mfrow=c(1,2))

# Ako grupiramo podatke i radimo histogram:
b = seq(min(BigFive$agreeable_score) - 0.1,max(BigFive$agreeable_score) + 0.1,0.2)

h1 = hist(BigFive[BigFive["sex"] == 1,]$agreeable_score,
          breaks=b,
          plot=FALSE)

h2 = hist(BigFive[BigFive["sex"] == 2,]$agreeable_score,
          breaks=b,
          plot=FALSE)


data <- t(cbind(h1$counts,h2$counts))
#data
barplot(data,beside=TRUE, col=c("lightblue", "lavender"), ylim = c(0,130000)
        ,xlab="ugodnost", ylab='frekvencija',)
legend("topleft",c("muški spol","zenski spol"),fill = c("lightblue", "lavender"))



h1 = hist(BigFive[BigFive["sex"] == 1,]$neuroticism_score,
          breaks=b,
          plot=FALSE)

h2 = hist(BigFive[BigFive["sex"] == 2,]$neuroticism_score,
          breaks=b,
          plot=FALSE)


data <- t(cbind(h1$counts,h2$counts))
#data
barplot(data,beside=TRUE, col=c("lightblue", "lavender"), ylim = c(0,130000)
        ,xlab="neuroticizam", ylab='frekvencija',)
legend("topleft",c("muški spol","zenski spol"),fill = c("lightblue", "lavender"))
```

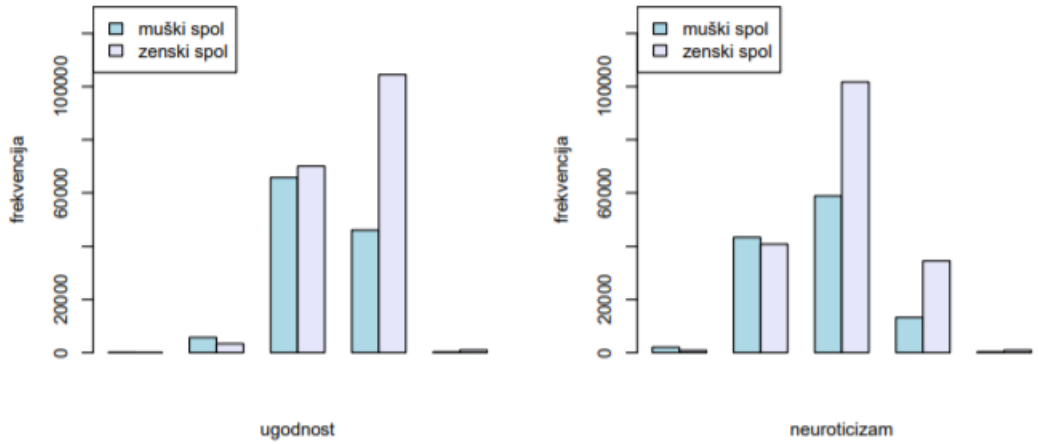

Figure 22: Razlicitost utjecaja spolova na osobnosti

Jasno se vidi da žene (u ovom uzorku) naginju većim vrijednostima rezultata za faktor neuroticizma i ugodnosti, no naš je cilj to statistički potvrditi. U nastavku ćemo provesti odgovarajuće statističke testove, te zaključiti postoji li utjecaj spola na faktor neuroticizma i ugodnosti.

Nezavisnost faktora (neuroticizam i ugodnost) i spola

Za ovakav oblik testiranja, potrebna nam je kontingencijska tablica. To je tablica kojom prikazujemo frekvencije uzorka dobivenog mjerenjem dvodimenzionalnog kategorijskog ili diskretnog numeričkog obilježja.

Prvo ispitujemo odnos spola i neuroticizma, te započinjemo sa stvaranjem kontingencijske tablice. Podijelit ćemo rezultate ispitanika u 4 kategorije:

- Y1 -> jako slabi neuroticizam [0, 0.25>
- Y2 -> slabi neuroticizam [0.25, 0.5>
- Y3 -> neuroticizam [0.5, 0.75>
- Y4 -> jaki neuroticizam [0.75, 1]

Pogledajmo sadržaj naše kontingencijske tablice.

```
BF_copy = data.frame(BigFive) #kopiramo podatke kako ne bi mijenjali
                              #podatke originalnom datasetu
BF_copy[BF_copy$sex == 1,]$sex <- "muški spol"
BF_copy[BF_copy$sex == 2,]$sex <- "ženski spol"

BF_copy[BigFive$neuroticism_score >= 0 &
          BigFive$neuroticism_score < 0.25,]$neuroticism_score <- "Y1"
BF_copy[BigFive$neuroticism_score >= 0.25 &
          BigFive$neuroticism_score < 0.5,]$neuroticism_score <- "Y2"
BF_copy[BigFive$neuroticism_score >= 0.5 &
          BigFive$neuroticism_score < 0.75,]$neuroticism_score <- "Y3"
BF_copy[BigFive$neuroticism_score >= 0.75 &
          BigFive$neuroticism_score <= 1,]$neuroticism_score <- "Y4"


tbl = table(BF_copy[BF_copy$neuroticism_score == "Y1" |
                      BF_copy$neuroticism_score == "Y2" |
                    BF_copy$neuroticism_score == "Y3" |
                      BF_copy$neuroticism_score == "Y4" ,]$sex,
```

```
            BF_copy[BF_copy$neuroticism_score == "Y1" |
                      BF_copy$neuroticism_score == "Y2"
                    |BF_copy$neuroticism_score == "Y3" |
                      BF_copy$neuroticism_score == "Y4" ,]$neuroticism_score)

added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##                    Y1     Y2     Y3     Y4    Sum
##   muški spol      282  43916  66348   7166 117712
##   ženski spol      69  40122 118159  20482 178832
##   Sum             351  84038 184507  27648 296544
```

Imamo spremnu kontingencijsku tablicu, sada postavimo hipoteze:

$$H_0 : o_i = e_i$$
$$H_1 : o_i \neq e_i$$

,gdje o_i predstavlja opaženu vrijednost, a e_i očekivanu vrijednost.

Za ovakvo testiranje koristimo statistiku :

$$X = \sum_{i,j} \frac{(o_i j - e_i j)^2}{e_i j}$$

,koja ima χ^2 distribuciju s $(r-1)(c-1)$ stupnjeva slobode. Broj stupaca je c , a broj redaka r.

Ovaj test u programskom paketu R implementiran je u funkciji `chisq.test()` ,koja kao ulaz prima kontingencijsku tablicu podataka koje testiramo na nezavisnost.

Pretpostavka testa je da očekivana frekvencija pojedinog razreda mora biti veća ili jednaka 5 (`chisq.test()` pretpostavlja da je ovaj uvjet zadovoljen stoga je prije provođenja testa potrebno to provjeriti).

```
for (col_names in colnames(added_margins_tbl)){
  for (row_names in rownames(added_margins_tbl)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za razred ',col_names,
      '-',row_names,': ',(added_margins_tbl[row_names,'Sum']
      * added_margins_tbl['Sum',col_names]) / added_margins_tbl['Sum','Sum'],'\n')
    }
  }
}
```

```
## Ocekivane frekvencije za razred  Y1 - muški spol :  139.3281
## Ocekivane frekvencije za razred  Y1 - ženski spol :  211.6719
## Ocekivane frekvencije za razred  Y2 - muški spol :  33358.56
## Ocekivane frekvencije za razred  Y2 - ženski spol :  50679.44
## Ocekivane frekvencije za razred  Y3 - muški spol :  73239.34
## Ocekivane frekvencije za razred  Y3 - ženski spol :  111267.7
## Ocekivane frekvencije za razred  Y4 - muški spol :  10974.77
## Ocekivane frekvencije za razred  Y4 - ženski spol :  16673.23
```

Uvjet je zadovoljen - možemo nastaviti s testiranjem. Nulta hipoteza testa nezavisnosti je da je spol neovisan o kategoriji neuroticizma ispitanika.

```
chisq.test(tbl,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 9049.9, df = 3, p-value < 2.2e-16
```

S obzirom na malu p vrijednost, možemo odbaciti H0 u korist alternative: kategorija neuroticizma zavisna je o spolu ispitanika.

Provedimo analogan test nezavisnosti za faktor ugodnosti. Podijelit ćemo rezultate ispitanika u 4 kategorije u istom omjeru kao i u prijašnjem testu :

```
BF_copy[BigFive$agreeable_score >= 0 &
          BigFive$agreeable_score < 0.25,]$agreeable_score <- "Y1"
BF_copy[BigFive$agreeable_score >= 0.25 &
          BigFive$agreeable_score < 0.5,]$agreeable_score <- "Y2"
BF_copy[BigFive$agreeable_score >= 0.5 &
          BigFive$agreeable_score < 0.75,]$agreeable_score <- "Y3"
BF_copy[BigFive$agreeable_score >= 0.75 &
          BigFive$agreeable_score <= 1,]$agreeable_score <- "Y4"


tbl = table(BF_copy[BF_copy$agreeable_score == "Y1"
                    |BF_copy$agreeable_score == "Y2"
                    |BF_copy$agreeable_score == "Y3"
                    |  BF_copy$agreeable_score == "Y4"
                    ,]$sex,
            BF_copy[BF_copy$agreeable_score == "Y1"
                    |BF_copy$agreeable_score == "Y2"
                    |BF_copy$agreeable_score == "Y3"
                    |  BF_copy$agreeable_score == "Y4"
                    ,]$agreeable_score)

added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##                   Y1     Y2     Y3     Y4    Sum
##   muški spol      14   5487  87955  24256 117712
##   ženski spol      1   3160 108761  66910 178832
##   Sum             15   8647 196716  91166 296544
```

```
for (col_names in colnames(added_margins_tbl)){
  for (row_names in rownames(added_margins_tbl)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za razred ',col_names,
          '-',row_names,': ',(added_margins_tbl
        [row_names,'Sum'] * added_margins_tbl['Sum',col_names])
        / added_margins_tbl['Sum','Sum'],'\n')
    }
  }
}
```

```
## Ocekivane frekvencije za razred  Y1 - muški spol :  5.954192
```

```
## Ocekivane frekvencije za razred  Y1 - ženski spol :  9.045808
## Ocekivane frekvencije za razred  Y2 - muški spol :  3432.393
## Ocekivane frekvencije za razred  Y2 - ženski spol :  5214.607
## Ocekivane frekvencije za razred  Y3 - muški spol :  78085.66
## Ocekivane frekvencije za razred  Y3 - ženski spol :  118630.3
## Ocekivane frekvencije za razred  Y4 - muški spol :  36187.99
## Ocekivane frekvencije za razred  Y4 - ženski spol :  54978.01
```

```
chisq.test(tbl,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 10650, df = 3, p-value < 2.2e-16
```

S obzirom na malu p vrijednost, možemo odbaciti H0 u korist alternative: kategorija ugodnosti zavisna je o spolu ispitanika.

U kontingencijskim tablicama možemo primijetiti da žene prevladavaju u kategoriji jake ugodnosti, ali i jakog neuroticizma.

Utječu li varijable ličnosti međusobno jedna na drugu?

Do sad smo provjeravali utjecaj drugih varijablih na varijable ličnosti. Jasno je zaključiti da varijable ličnost nisu nezavisne. Logično je povezati visoku vrijednosti ekstraverzije i otvorenosti. Ipak, za model koji opisuje dimenzije ličnosti ne bi bilo dobro da su varijable međusobno jako korelirane jer je cilj svakog modela smanjiti redundanciju, a istovremeno što bolje opisati stvarnost. U ovom dijelu ćemo ispitati međusobni utjecaj varijabli ličnosti koristeći regresiju. Proučavat ćemo utjecaj ostalih varijabli podatkovnog skupa na faktor neuroticizma.

Utjecaj jedne varijable na drugu se predstavlja putem linearne regresije. Pri analizi odnosa između regresora i reakcije razlikujemo:

1. jednostavnu regresiju (jedan regresor)
2. višestruku regresiju (više regresora)

Vrlo često imamo specifičan odnos među varijablama takav da neku varijablu možemo smatrati slučajnom reakcijom na neku drugu nezavisnu varijablu (regresor). Ta veza nije deterministička, a model je često simplifikacija stvarnog stanja.

Model linearne regresije pretpostavlja linearnu vezu između ulaznih i izlaznih varijabli:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon$$

Procjena modela iz podataka:

$$\hat{Y} = b_0 + \sum_{j=1}^{p} b_j x_j + e,$$

Za procjenu koristimo metodu najmanjih kvadrata:

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = (\mathbf{y} - \mathbf{Xb})^T (\mathbf{y} - \mathbf{Xb})$$

Kad promatramo utjecaj samo jedne nezavisne varijable X na neku zavisnu varijablu Y, grafički posmatrano najbolji prikaz ćemo ostvariti uporabom scatter plot-a.

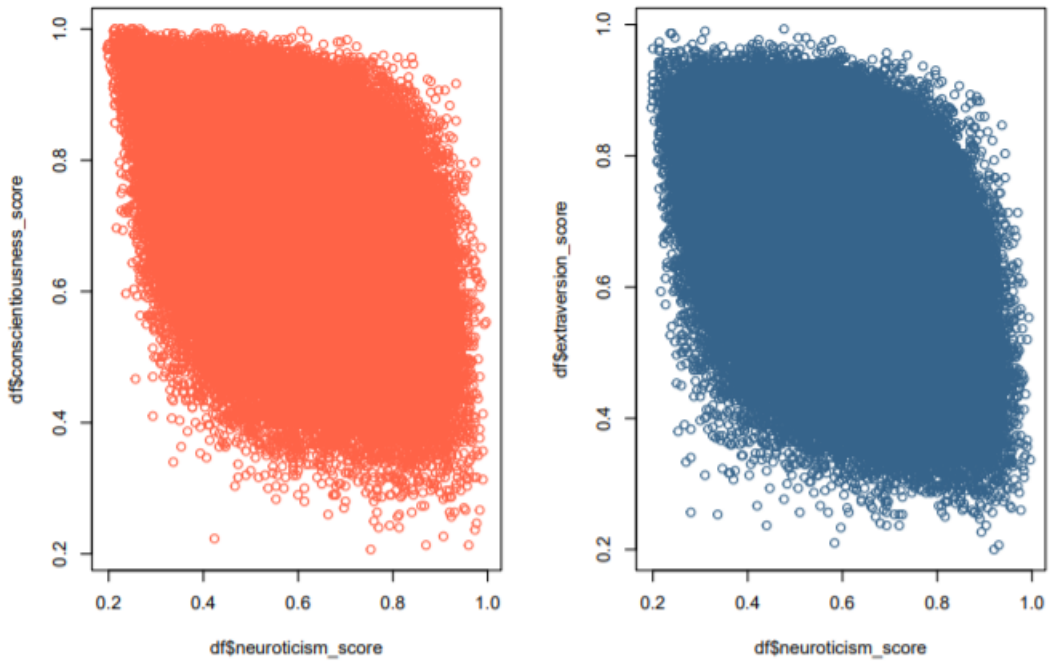

Figure 23: Graficki prikaz ovisnosti ekstraverzije/savjesnosti o neuroticizmu

Primjećujemo da conscientiousness_score i extraversion_score imaju utjecaj na izlaznu varijablu. Konkretno, smanjenjem conscientiousness_score-a i extraversion_score-a, povećava se neuroticism_score. Dakle, prosječno su neurotične osobe manje savjesne i zatvorenije. Barem na prvi pogled.

Kako bi ispitali pojedinačni utjecaj ovih varijabli, procijenit ćemo model jednostavne regresije - po jedan za svaku nezavisnu varijablu.

Koristit ćemo funkciju lm() za procjenu regresijskog modela. Dodat ćemo i pravce linearne regresije na graf. Što je prilagodba pravcu bolja, to je R^2 bliže 1.

```
par(mfrow=c(1,2))
fit.neurocon = lm(neuroticism_score~conscientiousness_score,data=df)
plot(df$conscientiousness_score,df$neuroticism_score)
lines(df$conscientiousness_score,fit.neurocon$fitted.values,col='red')

fit.neuroext = lm(neuroticism_score~extraversion_score,data=df)
plot(df$extraversion_score,df$neuroticism_score)
lines(df$extraversion_score,fit.neuroext$fitted.values,col='red')
```

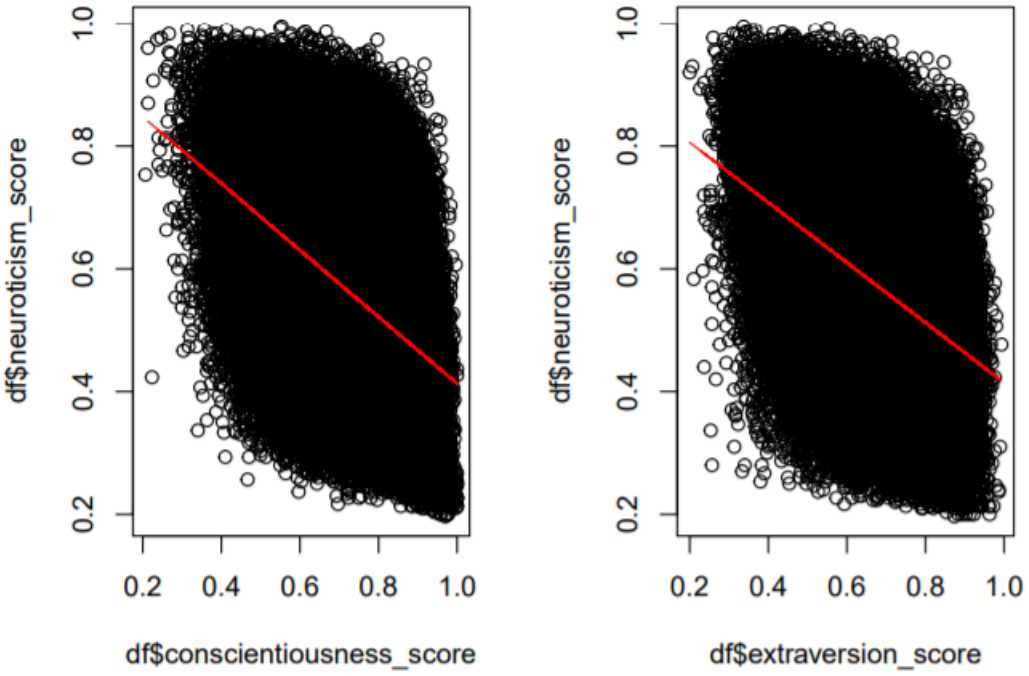

Figure 24: Graficki prikaz ovisnosti neuroticizma o ekstraverziji/savjesnosti

Nagibi pravaca linearne regresije potvrđuju tvrdnje o efektima pojedinih razmatranih varijabli na izlaznu varijablu. Prije daljne analize, prvo ćemo provjeriti pretpostavke modela. Ukoliko nisu jako narušene, možemo nastaviti. Provjerit ćemo pretpostavke o regresorima i rezidualima. U višestrukoj regresiji, regresori ne smiju biti međusobno jako korelirani, a reziduali trebaju pokazivati normalnu razdiobu.

Provjera normalnost reziduala i homogenosti varijance

Prvo provjeravamo normalnost reziduala i to grafički, pomoću kvantil-kvantil plota, te statističkim testovima. Koristit ćemo Kolmogorov-Smirnovljev test (Lilliefors).

```
par(mfrow=c(3,2))

hist((fit.neurocon$residuals))
hist(rstandard(fit.neurocon))

#q-q plot reziduala s linijom normalne distribucije
qqnorm(rstandard(fit.neurocon))
qqline(rstandard(fit.neurocon))

plot(fit.neurocon$fitted.values,fit.neurocon$residuals) #reziduale je dobro
                                  #prikazati u ovisnosti o procjenama modela


#KS test na normalnost
ks.test(rstandard(fit.neurocon),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
```

```
## data:  rstandard(fit.neurocon)
## D = 0.017324, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
require(nortest)
lillie.test(rstandard(fit.neurocon))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.neurocon)
## D = 0.017323, p-value < 2.2e-16
```

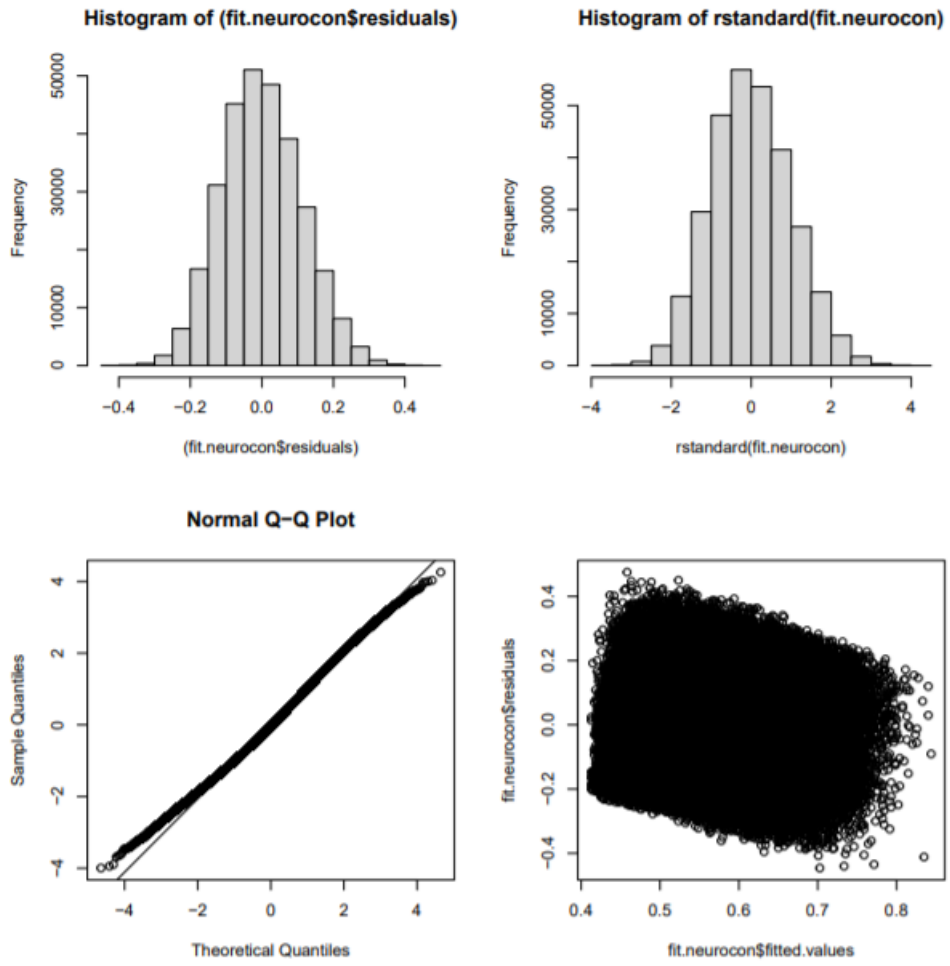

Figure 25: Grafički prikaz normalnosti i razdiobe reziduala

Korišteni histogram pokazuje razdiobu (standardiziranih) reziduala. U našem slučaju zaključujemo da distribucija jako nalikuje normalnoj.

Kolmogorov-Smirnovljev test je neparametarski test te služi za provjeru dolaze li podaci iz neke točno određene distribucije. Lillieforceva inačica se koristi ako želimo testirati dolaze li podaci iz normalne distribucije, a ne poznajemo iznos očekivanja i varijance populacije. Iako se rezultati testova razlikuju, reziduali ne pokazuju preveliko odstupanje od normalnosti te je poznato da je t-test robustan na normalnost pa možemo zaključiti

da se regresijski model može koristiti i dalje za statističko razmatranje.

Nadalje, koristimo funkciju summary() nad objektom koji vraća lm() kako bismo dobili parametre poput SSE-a, koeficijent determinacije, prilagođeni koeficijent determinacije, f-statistika. Koeficijent determinacije R^2 opisuje koji postotak varijance u izlaznoj varijabli Y je estimirani linearni model objasnio/opisao, a f-statistika pokazuje je li linearni model adekvatan.

```
summary(lm(neuroticism_score~extraversion_score,data=df))
```

```
##
## Call:
## lm(formula = neuroticism_score ~ extraversion_score, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49810 -0.07965 -0.00197  0.07846  0.44716
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.903643   0.001311   689.5   <2e-16 ***
## extraversion_score -0.489137   0.001924  -254.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1142 on 296542 degrees of freedom
## Multiple R-squared:  0.1789, Adjusted R-squared:  0.1789
## F-statistic: 6.461e+04 on 1 and 296542 DF,  p-value: < 2.2e-16
```

```
summary(lm(neuroticism_score~conscientiousness_score,data=df))
```

```
##
## Call:
## lm(formula = neuroticism_score ~ conscientiousness_score, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44601 -0.07879 -0.00438  0.07536  0.47516
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.956237   0.001353   706.7   <2e-16 ***
## conscientiousness_score -0.543339   0.001905  -285.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1116 on 296542 degrees of freedom
## Multiple R-squared:  0.2152, Adjusted R-squared:  0.2152
## F-statistic: 8.133e+04 on 1 and 296542 DF,  p-value: < 2.2e-16
```

```
summary(lm(neuroticism_score~openness_score,data=df))
```

```
##
## Call:
## lm(formula = neuroticism_score ~ openness_score, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.37686 -0.08951 -0.00413  0.08603  0.42407
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.612345   0.001944  314.93   <2e-16 ***
## openness_score -0.051183   0.002631  -19.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1259 on 296542 degrees of freedom
## Multiple R-squared:  0.001274,   Adjusted R-squared:  0.001271
## F-statistic: 378.4 on 1 and 296542 DF,  p-value: < 2.2e-16
```

```
summary(lm(neuroticism_score~agreeable_score,data=df))
```

```
##
## Call:
## lm(formula = neuroticism_score ~ agreeable_score, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41425 -0.08802 -0.00416  0.08500  0.44304
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.733216   0.001708  429.36   <2e-16 ***
## agreeable_score -0.227332   0.002429  -93.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1242 on 296542 degrees of freedom
## Multiple R-squared:  0.0287, Adjusted R-squared:  0.0287
## F-statistic:  8763 on 1 and 296542 DF,  p-value: < 2.2e-16
```

```
summary(lm(neuroticism_score~sex,data=df))
```

```
##
## Call:
## lm(formula = neuroticism_score ~ sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38480 -0.08814 -0.00480  0.08520  0.43894
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4939779  0.0007773   635.5   <2e-16 ***
## sex         0.0504119  0.0004637   108.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1236 on 296542 degrees of freedom
## Multiple R-squared:  0.03832,    Adjusted R-squared:  0.03832
## F-statistic: 1.182e+04 on 1 and 296542 DF,  p-value: < 2.2e-16
```

Kao što je bilo vidljivo iz inicijalnih grafičkih prikaza, extraversion_score i conscientiousness_score kao

varijable imaju najveći efekt na neuroticism_score i objašnjava najveće vrijednosti parametra R^2.

Utječe li skup više određenih varijabli ličnosti na jednu varijablu?

Da bismo odredili utječe li više varijabli ličnosti na izlaznu varijablu, koristit ćemo višestruku regresiju. Regresija s jako koreliranim ulaznim varijablama će uglavnom dati neke rezultate, ali na temelju njih je otežano donositi zaključke. Korelacijski koeficijent je povezan s linearnom regresijom i koeficijentom determinacije R^2 i iznosi $r = \sqrt{R^2}$.

Pogledajmo korelacijske koeficijente za parove regresora.

```
library(corrplot,verbose = FALSE)
podatci = BigFive[, names(BigFive) %in% c("agreeable_score",
                  "extraversion_score","conscientiousness_score",
                  "openness_score","neuroticism_score","age")]

M <- cor(podatci)
corrplot(M, method = 'number')
```

Figure 26: Graficki prikaz korelacije koeficijenata osobnosti

Prije procjene modela višestruke regresije potrebno je provjeriti da pojedini parovi varijabli nisu (previše) korelirani. U principu je određena korelacija između varijabli neizbježna, ali varijable s vrlo visokom korelacijom će uzrokovati probleme u interpretaciji regresijskih rezultata. Budući da nisu jako korelirani, možemo odabrati extraversion_score i conscientiousness_score (faktor između kojeg je linearna veza s neuroticizmom najjača, koeficijent determinacije najveći) za naš model.

```
fit.agreeneuro = lm(neuroticism_score  ~ extraversion_score +
                      conscientiousness_score  , BigFive)
summary(fit.agreeneuro)
```

```
##
## Call:
## lm(formula = neuroticism_score ~ extraversion_score + conscientiousness_score,
##     data = BigFive)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50831 -0.07078 -0.00247  0.06928  0.42156
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.189578   0.001580   752.8   <2e-16 ***
## extraversion_score       -0.415594   0.001745  -238.2   <2e-16 ***
## conscientiousness_score -0.477721   0.001767  -270.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1023 on 296541 degrees of freedom
## Multiple R-squared:  0.3413, Adjusted R-squared:  0.3413
## F-statistic: 7.681e+04 on 2 and 296541 DF,  p-value: < 2.2e-16
```

```
fit.agreeneuro = lm(neuroticism_score  ~ extraversion_score +
                  conscientiousness_score + agreeable_score +
                    age + openness_score , BigFive)
summary(fit.agreeneuro)
```

```
##
## Call:
## lm(formula = neuroticism_score ~ extraversion_score + conscientiousness_score +
##     agreeable_score + age + openness_score, data = BigFive)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51901 -0.07010 -0.00217  0.06870  0.42512
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.179e+00  2.266e-03  520.17   <2e-16 ***
## extraversion_score       -4.432e-01  1.832e-03 -241.98   <2e-16 ***
## conscientiousness_score -4.406e-01  1.935e-03 -227.73   <2e-16 ***
## agreeable_score          -2.877e-02  2.143e-03  -13.42   <2e-16 ***
## age                      -9.015e-04  1.978e-05  -45.57   <2e-16 ***
## openness_score            6.277e-02  2.244e-03   27.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1018 on 296538 degrees of freedom
## Multiple R-squared:  0.3476, Adjusted R-squared:  0.3476
## F-statistic: 3.16e+04 on 5 and 296538 DF,  p-value: < 2.2e-16
```

Dodavanjem dodatnih varijabli u model, prilagođeni koeficijent determinacije je samo za 1% bolji od modela koji uključuje ekstraverziju i savjesnost pa možemo reći da je model s manje varijabli preferiran jer podjednako dobro objašnjava podatke, a jednostavniji je.

Prethodno smo vidjeli da žene imaju veće rezultate na faktoru neuroticizma od muškaraca pa očekujemo da će uključivanje spola u model poboljšati model.

Za predstavljanje kategorijskih varijabli kao ulaz regresijskog modela postoje različite tehnike, a jedna od najjednostavnijih i najčešće korištenih su tzv. dummy varijable. Svaka kategorija u kategorijskoj varijabli predstavljena je svojom vlastitom indikatorskom varijablom koja poprima vrijednost 1 u slučaju da originalna kategorijska varijabla poprima vrijednost te kategorije, a 0 inače. U našem slučaju sex_1 označava muški spol, a sex_2 ženski. Korištenje kategorijskih varijabli s više od dvije kategorije kao int vrijednosti u regresiji se ne preporuča za nominalne varijable, ali za spol koji ima dvije kategorije ova je pretvorba implicitno napravljena pa nisu potrebne dummy varijable.

```
require(fastDummies, quietly = TRUE)
#BigFive.d = dummy_cols(BigFive,select_columns='sex')

#BigFive.d
#fit.multi.d = lm(neuroticism_score ~ extraversion_score
#+ conscientiousness_score + sex_1, BigFive.d)
#summary(fit.multi.d)


fit.multi = lm(neuroticism_score ~ extraversion_score
               + conscientiousness_score + sex, BigFive)
summary(fit.multi)
```

```
##
## Call:
## lm(formula = neuroticism_score ~ extraversion_score + conscientiousness_score +
##     sex, data = BigFive)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51872 -0.06718 -0.00253  0.06550  0.43757
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.1094855  0.0015781   703.1   <2e-16 ***
## extraversion_score       -0.4364188  0.0016689  -261.5   <2e-16 ***
## conscientiousness_score -0.4874628  0.0016866  -289.0   <2e-16 ***
## sex                       0.0629631  0.0003675   171.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09755 on 296540 degrees of freedom
## Multiple R-squared:  0.4006, Adjusted R-squared:  0.4006
## F-statistic: 6.606e+04 on 3 and 296540 DF,  p-value: < 2.2e-16
```

Primjetimo da je vrijednost R^2 pri višestrukoj regresiji extraversion_score-a, conscientiousness_score-a i spola iznosi skoro 40%. Teško je odrediti koliki R^2 je potreban za podatke poput ovih. Za neke studije je dovoljno da bude 30%, dok za neke neće biti dovoljno ni 80%. Obzriom da se ova studija vezana za ljudsko ponašanje, rezultat od skoro 40% je zadovoljavajući.

Možemo li odrediti spol osobe koristeći ostale dostupne podatke o osobi?

Želimo li koristeći postojeće podatke o osobi odrediti spol te osobe, možemo pokušati procijeniti regresijski model s podacima o osobi kao nezavisnim varijablama. Kada zavisna varijabla nije kontinuirana smisla ima koristiti logističku regresiju.

Imamo na raspolaganju skup podataka $D = \{X_1, ..., X_N\}$ gdje je svaki X_i vektor vrijednosti prediktorskih varijabli, one mogu biti diskretne (uz prikladno dummy-kodiranje) ili kontinuirane. Imamo i skup očekivanih

izlaza $\{y_1, ..., y_n\}$ gdje je svaki y_i binarna varijabla tj. 0 ili 1. Želimo dobiti kao izlaz modela skup izlaza $\{\hat{y_1}, ..., \hat{y_N}\}$. Idealno bismo od dobrog modela očekivali da bude $\hat{y_i} = y_i$, tj. da radi dobre predikcije.

Koristit ćemo `summary` naredbu kako bismo imali uvid ponajprije u devijancu (i to null deviance i residual deviance). Devijanca nam govori koliko je model dobar(veći broj znači da je prilagodba gora). Koristeći te dvije veličine, moguće je i izračunati R^2 danog modela kao:

$$R^2 = 1 - \frac{D_{mdl}}{D_0}.$$

Pomoću R^2 možemo odrediti koliko je procijenjeni model blizu null modelu.

```
require(caret, quietly = TRUE)


logreg.mdl = glm(as.factor(sex) ~ openness_score + neuroticism_score
+ conscientiousness_score + agreeable_score + age + extraversion_score,
data = BigFive, family = binomial())
summary(logreg.mdl)
```

```
##
## Call:
## glm(formula = as.factor(sex) ~ openness_score + neuroticism_score +
##     conscientiousness_score + agreeable_score + age + extraversion_score,
##     family = binomial(), data = BigFive)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8719  -1.0877   0.6123   0.9299   2.9632
##
## Coefficients:
##                           Estimate Std. Error  z value Pr(>|z|)
## (Intercept)             -1.339e+01  7.794e-02 -171.748   <2e-16 ***
## openness_score           9.060e-01  4.907e-02   18.462   <2e-16 ***
## neuroticism_score        7.250e+00  4.493e-02  161.380   <2e-16 ***
## conscientiousness_score  2.631e+00  4.666e-02   56.383   <2e-16 ***
## agreeable_score          6.088e+00  4.967e-02  122.557   <2e-16 ***
## age                      2.473e-04  4.319e-04    0.573    0.567
## extraversion_score       4.381e+00  4.561e-02   96.044   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 398409  on 296543  degrees of freedom
## Residual deviance: 349759  on 296537  degrees of freedom
## AIC: 349773
##
## Number of Fisher Scoring iterations: 3
```

```
Rsq = 1 - logreg.mdl$deviance/logreg.mdl$null.deviance
Rsq
```

```
## [1] 0.1221107
```

Naknadno ćemo dobivene devijance usporediti s devijancama nešto drugačijih modela te doći do zaključka koji model je najbolji. Pošto omjer oznaka u izlaznoj varijabli može utjecati na neke mjere kvalitete modela, koristimo matricu zabune tj. *confusion matrix*.

Mjere koje mogu biti od interesa su:

- točnost (eng. *accuracy*): jedan od bitnijih pokazatelja kvalitete modela jer nam upravo on izravno govori o postotku testnih primjera koji su ispravno klasificirani ($\frac{TP+TN}{TP+FP+TN+FN}$)
- preciznost (eng. *precision*): predstavlja udio točno klasificiranih razreda u svim situacijama kada je određeni podatak bio klasificiran kao TRUE ($\frac{TP}{TP+FP}$)
- odziv (eng. *recall*): daje informaciju o tome koliko je puta podatak bio točno klasificiran u određeni razred u odnosu na ukupan broj pojava tog pojedinog razreda (ukupan broj primjera u skupu koji su stvarno TRUE) ($\frac{TP}{TP+FN}$)
- specifičnost (eng. *specificity*): udio točno klasificiranih "negativnih" primjera u odnosu na sve negativne primjere iz skupa ($\frac{TN}{TN+FP}$)

```
test<- ifelse(BigFive$sex==1, "muški", "ženski")
yHat <- logreg.mdl$fitted.values > 0.5
tab <- table(test, yHat)

tab
```

```
##         yHat
## test      FALSE   TRUE
##   muški   55534  62178
##   ženski  30086 148746
```

```
accuracy = sum(diag(tab)) / sum(tab)
precision = tab[2,2] / sum(tab[,2])
recall = tab[2,2] / sum(tab[2,])
specificity = tab[1,1] / sum(tab[,1])

accuracy
```

```
## [1] 0.6888691
```

```
precision
```

```
## [1] 0.7052114
```

```
recall
```

```
## [1] 0.8317639
```

```
specificity
```

```
## [1] 0.6486101
```

U matrici zabune FALSE predstavlja muški spol, dok TRUE pak ženski spol. Osnovne mjere su poprilično dobre, naročito ako uzimamo u obzir da se radi o ljudskim osobinama kao prediktorima. Vidimo da mjera odziva doseže vrijednost 83,17% što znači da je velik udio žena u uzorku točno klasificiran. Najslabija mjera ovog modela je specifičnost (udio točnih primjera u svim koji su klasificirani kao FALSE) vezana uz točnost predikcije muškog spola.

Test omjera izglednosti (likelihood ratio test)

Nadalje testiramo model s dodatnim interakcijskim članom I(extraversion_score/neuroticism_score) te provjeravamo je li novi model bolji od originalnog. To možemo provjeriti tako što usporedimo devijance oba modela i ukoliko novi model ima značajno manju devijancu, možemo ga prihvatiti. Odgovor na to pitanje nam daje test omjera izglednosti.

```
logreg.mdl = glm(as.factor(sex) ~ openness_score +neuroticism_score+
conscientiousness_score+agreeable_score+age+extraversion_score,
data = BigFive, family = binomial())
summary(logreg.mdl)
```

```
##
## Call:
## glm(formula = as.factor(sex) ~ openness_score + neuroticism_score +
##     conscientiousness_score + agreeable_score + age + extraversion_score,
##     family = binomial(), data = BigFive)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8719  -1.0877   0.6123   0.9299   2.9632
##
## Coefficients:
##                            Estimate Std. Error  z value Pr(>|z|)
## (Intercept)             -1.339e+01  7.794e-02 -171.748   <2e-16 ***
## openness_score           9.060e-01  4.907e-02   18.462   <2e-16 ***
## neuroticism_score        7.250e+00  4.493e-02  161.380   <2e-16 ***
## conscientiousness_score  2.631e+00  4.666e-02   56.383   <2e-16 ***
## agreeable_score          6.088e+00  4.967e-02  122.557   <2e-16 ***
## age                      2.473e-04  4.319e-04    0.573    0.567
## extraversion_score       4.381e+00  4.561e-02   96.044   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 398409  on 296543  degrees of freedom
## Residual deviance: 349759  on 296537  degrees of freedom
## AIC: 349773
##
## Number of Fisher Scoring iterations: 3
```

```
logreg.mdl.2 =glm(as.factor(sex) ~  openness_score +
neuroticism_score+extraversion_score+ conscientiousness_score
+agreeable_score+age+I(extraversion_score/neuroticism_score),
data = BigFive, family = binomial())
summary(logreg.mdl.2)
```

```
##
## Call:
## glm(formula = as.factor(sex) ~ openness_score + neuroticism_score +
##     extraversion_score + conscientiousness_score + agreeable_score +
##     age + I(extraversion_score/neuroticism_score), family = binomial(),
##     data = BigFive)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8669  -1.0830   0.6110   0.9248   2.9959
##
## Coefficients:
##                                          Estimate Std. Error  z value Pr(>|z|)
```

```
## (Intercept)                                -1.226e+01  8.700e-02 -140.930   <2e-16
## openness_score                              9.732e-01  4.920e-02   19.782   <2e-16
## neuroticism_score                           5.170e+00  8.499e-02   60.828   <2e-16
## extraversion_score                          5.972e+00  7.234e-02   82.561   <2e-16
## conscientiousness_score                     2.690e+00  4.676e-02   57.514   <2e-16
## agreeable_score                             6.107e+00  4.974e-02  122.782   <2e-16
## age                                         7.556e-04  4.333e-04    1.744   0.0812
## I(extraversion_score/neuroticism_score) -8.929e-01  3.144e-02  -28.398   <2e-16
##
## (Intercept)                                ***
## openness_score                             ***
## neuroticism_score                          ***
## extraversion_score                         ***
## conscientiousness_score                    ***
## agreeable_score                            ***
## age                                        .
## I(extraversion_score/neuroticism_score) ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 398409  on 296543  degrees of freedom
## Residual deviance: 348933  on 296536  degrees of freedom
## AIC: 348949
##
## Number of Fisher Scoring iterations: 3
```

```
Rsq = 1 - logreg.mdl.2$deviance/logreg.mdl.2$null.deviance
Rsq
```

```
## [1] 0.1241849
```

```
anova(logreg.mdl, logreg.mdl.2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: as.factor(sex) ~ openness_score + neuroticism_score + conscientiousness_score +
##     agreeable_score + age + extraversion_score
## Model 2: as.factor(sex) ~ openness_score + neuroticism_score + extraversion_score +
##     conscientiousness_score + agreeable_score + age + I(extraversion_score/neuroticism_score)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1    296537     349759
## 2    296536     348933  1   826.38 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Također možemo testirati i razliku originalnog modela i smanjenog modela koji ne sadrži neke nesignifikantne regresore. U tom slučaju ćemo prihvatiti smanjeni model ukoliko devijanca nije značajno veća. Novi model neće sadržati regresor age.

```
logreg.mdl.3 = glm(as.factor(sex) ~ openness_score +
neuroticism_score+ conscientiousness_score+
agreeable_score+extraversion_score, data =BigFive,
family = binomial())
```

```
summary(logreg.mdl.3)
```

```
##
## Call:
## glm(formula = as.factor(sex) ~ openness_score + neuroticism_score +
##     conscientiousness_score + agreeable_score + extraversion_score,
##     family = binomial(), data = BigFive)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8727  -1.0877   0.6123   0.9299   2.9636
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -13.38177    0.07759 -172.46   <2e-16 ***
## openness_score            0.90808    0.04894   18.56   <2e-16 ***
## neuroticism_score         7.24797    0.04478  161.86   <2e-16 ***
## conscientiousness_score   2.63638    0.04562   57.79   <2e-16 ***
## agreeable_score           6.09002    0.04952  122.97   <2e-16 ***
## extraversion_score        4.37527    0.04456   98.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 398409  on 296543  degrees of freedom
## Residual deviance: 349760  on 296538  degrees of freedom
## AIC: 349772
##
## Number of Fisher Scoring iterations: 3
```

```
Rsq1 = 1 - logreg.mdl$deviance/logreg.mdl$null.deviance
Rsq = 1 - logreg.mdl.3$deviance/logreg.mdl.3$null.deviance
Rsq1
```

```
## [1] 0.1221107
```

```
Rsq
```

```
## [1] 0.1221099
```

```
anova(logreg.mdl, logreg.mdl.3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: as.factor(sex) ~ openness_score + neuroticism_score + conscientiousness_score +
##     agreeable_score + age + extraversion_score
## Model 2: as.factor(sex) ~ openness_score + neuroticism_score + conscientiousness_score +
##     agreeable_score + extraversion_score
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1    296537     349759
## 2    296538     349760 -1 -0.32787   0.5669
```

Obzirom na minimalnu razliku u devijancama, možemo ovaj model prihvatiti kao bolji model. Za kraj ćemo napraviti analizu konačnog modela:

```r
test<- ifelse(BigFive$sex==1, "muški", "ženski")
yHat <- logreg.mdl.3$fitted.values > 0.5
tab <- table(test, yHat)
tab
```

```
##         yHat
## test      FALSE   TRUE
##   muški   55528  62184
##   ženski  30102 148730
```

```r
accuracy = sum(diag(tab)) / sum(tab)
precision = tab[2,2] / sum(tab[,2])
recall = tab[2,2] / sum(tab[2,])
specificity = tab[1,1] / sum(tab[,1])

accuracy
```

```
## [1] 0.6887949
```

```r
precision
```

```
## [1] 0.7051689
```

```r
recall
```

```
## [1] 0.8316744
```

```r
specificity
```

```
## [1] 0.6484643
```

Vidimo da je i ovaj model dovoljno dobar nakon detaljne analize. Accuracy, precision, recall i specificity su se minimalno smanjile obzirom na originalni model te su i dalje jako visoke. Za čak 83% žena je točno određen spol što vidimo iz mjere recall, dok je nešto slabija karakteristika specificity, odnosno od svih predikcija muškog spola, 65% ih je bilo ispravno.

Zaključak

U ovom projektu pozabavili smo se mnogim društvenim predrasudama. Na temelju adekvatnih statističkih testova donijeli smo par zaključaka. Mladi su otvorenije osobe od starijih, Japanci su manje ekstrovertirani od Rusa i Hrvata, žene prevladavaju u kategoriji ugodnosti, neurotične osobe su manje savijesne , itd.... Neki od naših zaključaka idu u korist predrasudama, dok drugi ne. Iako nije ugodno opravdavati svoje tvrdnje preko predrasuda, mi smo zaključili da je nekad upravo to i ispravno. Čitav projekt je implementiran u okruženju R-Studio. Korišteni su nastavni materijali dostupni na stranici predmeta "Statistička analiza podataka".