

Big Five – velikih pet dimenzija ličnosti

Projekt iz predmeta Statistička analiza podataka, Fakultet elektrotehnike i računarstva

Filip Grebenar, David Konjevod, Lovre Mitrović, Mihaela Bakšić

14.1.2022.

Uvod

Ljudske osobine i njihovo ispoljavanje često su istraživan fenomen. Trenutno dominantni model za klasifikaciju osobina ličnosti je Big Five model. Inicialno je model bio razvijen s ciljem povezivanja osobina ličnosti i akademskih uspjeha i ponašanja. Barem četiri nezavisna skupa istraživača bavili su se problemom klasifikacija osobina ličnosti, te su svi diferencirali pet glavnih osobina ličnosti: ekstraverzija, ugodnost, savjesnost, neuroticizam i otvorenost. Podaci su prikupljeni kroz samoevaluacijske upitnike u kojima ispitanici odgovaraju na raznovrsna pitanja relevantna za analizu ličnosti. Za analizu podataka koristila se analiza faktora.

Ovaj projekt predstavit će generalni pregled podataka, pokušati pronaći pravilnosti u podacima i modelirati ih. Takoder, pozabavit će se ispitivanjem nekih uvriježenih ideja i mišljenja o osobinama ličnosti.

Skup podataka

```
bigfive = read.csv("./big_five_scores.csv")
head(bigfive)

##   case_id    country age sex agreeable_score extraversion_score openness_score
## 1      1 South Afri  24   1      0.7533333     0.4966667    0.8033333
## 2      3          UK  24   2      0.7333333     0.6800000    0.7866667
## 3      4          USA  36   2      0.8800000     0.7700000    0.8600000
## 4      5          UK  19   1      0.6900000     0.6166667    0.7166667
## 5      6          UK  17   1      0.6000000     0.7133333    0.6466667
## 6      7          USA  17   1      0.6033333     0.5866667    0.6533333
##   conscientiousness_score neuroticism_score
## 1            0.8866667        0.4266667
## 2            0.7466667        0.5900000
## 3            0.8966667        0.2966667
## 4            0.6366667        0.5633333
## 5            0.6333333        0.5133333
## 6            0.5966667        0.6233333

summary(bigfive)
```

```

##      case_id          country        age         sex
##  Min.   :    1   USA       :212625   Min.   :10.00   Min.   :1.000
##  1st Qu.: 83653  Canada    : 21798   1st Qu.:18.00   1st Qu.:1.000
##  Median :166286   UK        : 16489   Median :22.00   Median :2.000
##  Mean   :166682  Australia : 10400   Mean   :25.19   Mean   :1.602
##  3rd Qu.:249627  Netherland:  3469    3rd Qu.:29.00   3rd Qu.:2.000
##  Max.   :334161   India     : 2841    Max.   :99.00   Max.   :2.000
##              (Other)   :39691
##      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
##

```

Testiranje razlika u otvorenosti kod mladih i starih ispitanika

Često je prisutna pretpostavka kako su mladi ljudi otvoreniji novim iskustvima od starih. Stoga, provest će se testiranje te pretpostavke. Mladim ljudima smatrati će se svi stari 30 ili manje godina, dok će se starim ljudima smatrati oni stari 60 ili više godina.

```

cat('Medijan godina svih ispitanika je ', median(bigfive$age), '\n')

## Medijan godina svih ispitanika je 22

old_subjects = bigfive[bigfive$age >= 60 ,]
young_subjects = bigfive[bigfive$age <= 30 ,]

cat('Srednja vrijednost otvorenosti za mlade ( <= 30 godina ) ispitanike je ', mean(young_subjects$openne)

## Srednja vrijednost otvorenosti za mlađe ( <= 30 godina ) ispitanike je 0.7338046

cat('Srednja vrijednost otvorenosti za stare ( >= 60 godina ) ispitanike je ', mean(old_subjects$openne

## Srednja vrijednost otvorenosti za stare ( >= 60 godina ) ispitanike je 0.7265828

cat('Varijanca otvorenosti za mlade ( <= 30 godina ) ispitanike je ', var(young_subjects$openness_score)

## Varijanca otvorenosti za mlađe ( <= 30 godina ) ispitanike je 0.007601208

```

```

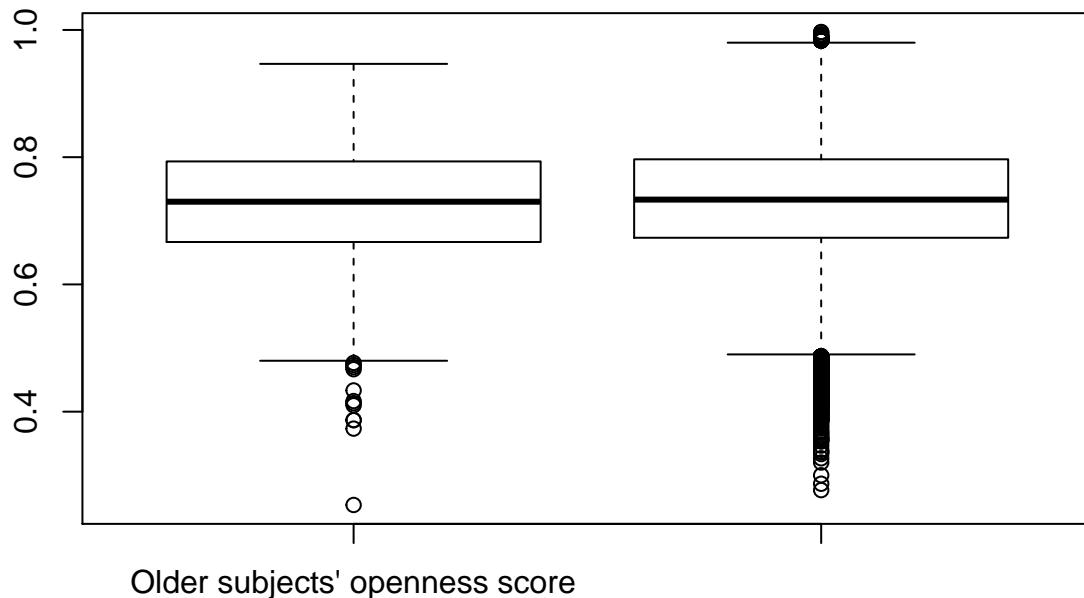
cat('Varijanca otvorenosti za stare ( >= 60 godina ) ispitanike je ', var(old_subjects$openness_score),

## Varijanca otvorenosti za stare ( >= 60 godina ) ispitanike je 0.008904973

boxplot(old_subjects$openness_score, young_subjects$openness_score,
        names = c('Older subjects\' openness score','Younger subjects\' openness score'),
        main='Boxplot of younger and older subjects\' openness score')

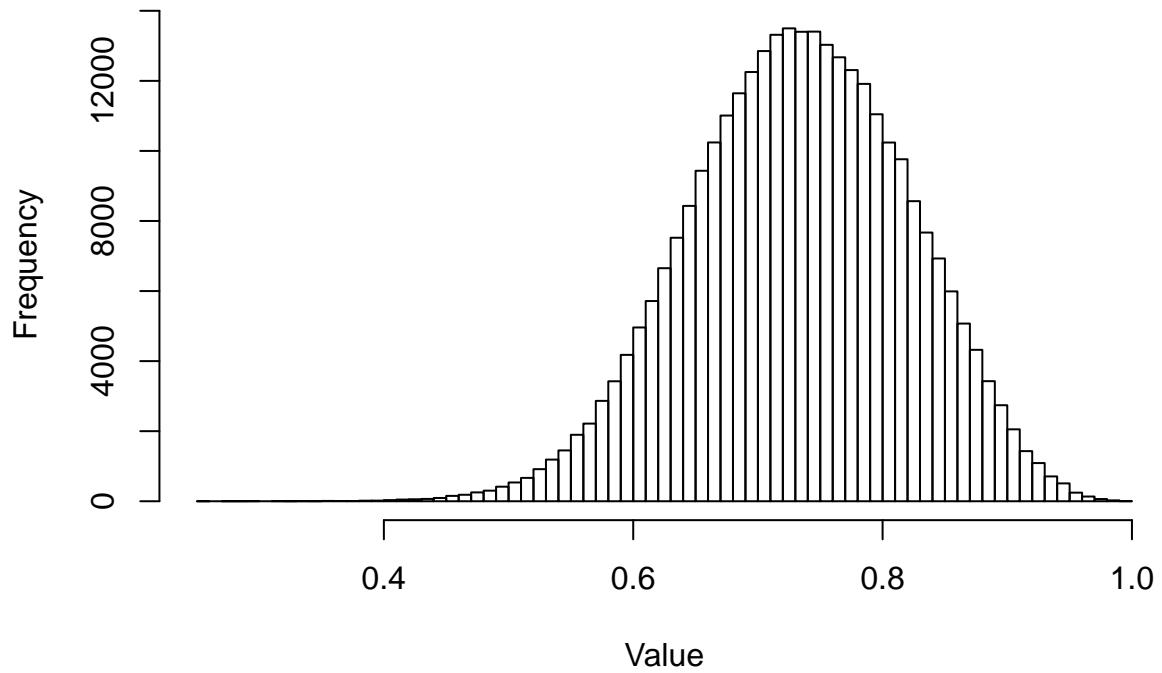
```

Boxplot of younger and older subjects' openness score



```
hist(bigfive$openness_score,main='Openness score histogram',xlab='Value',ylab='Frequency', breaks=100)
```

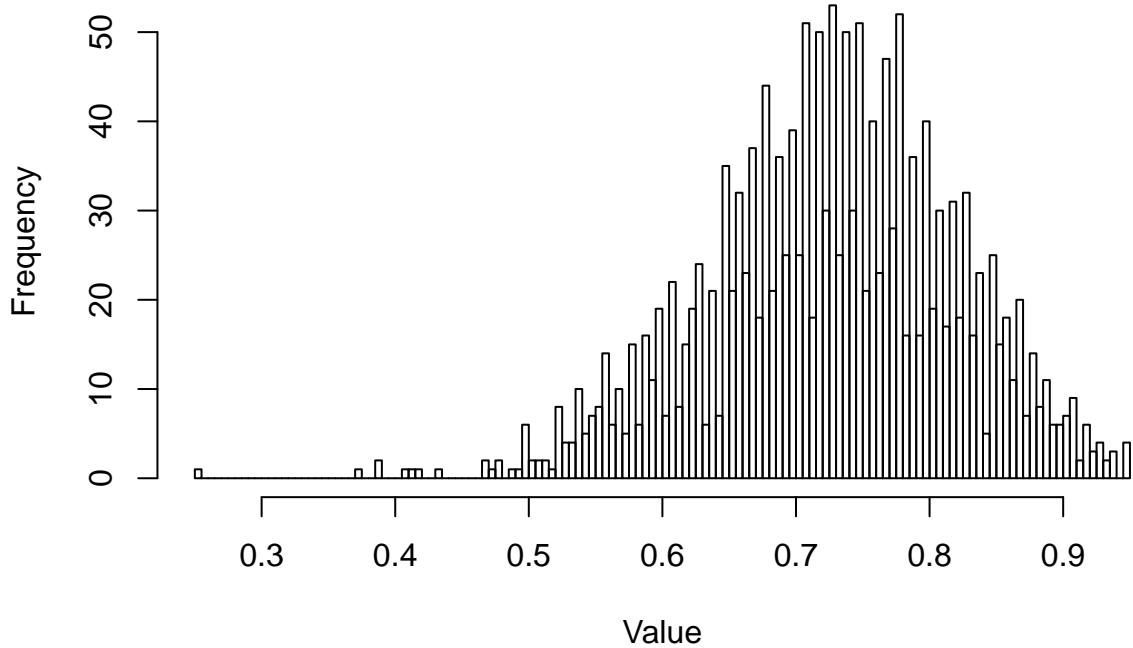
Openness score histogram



Normalnost podataka - grafovi

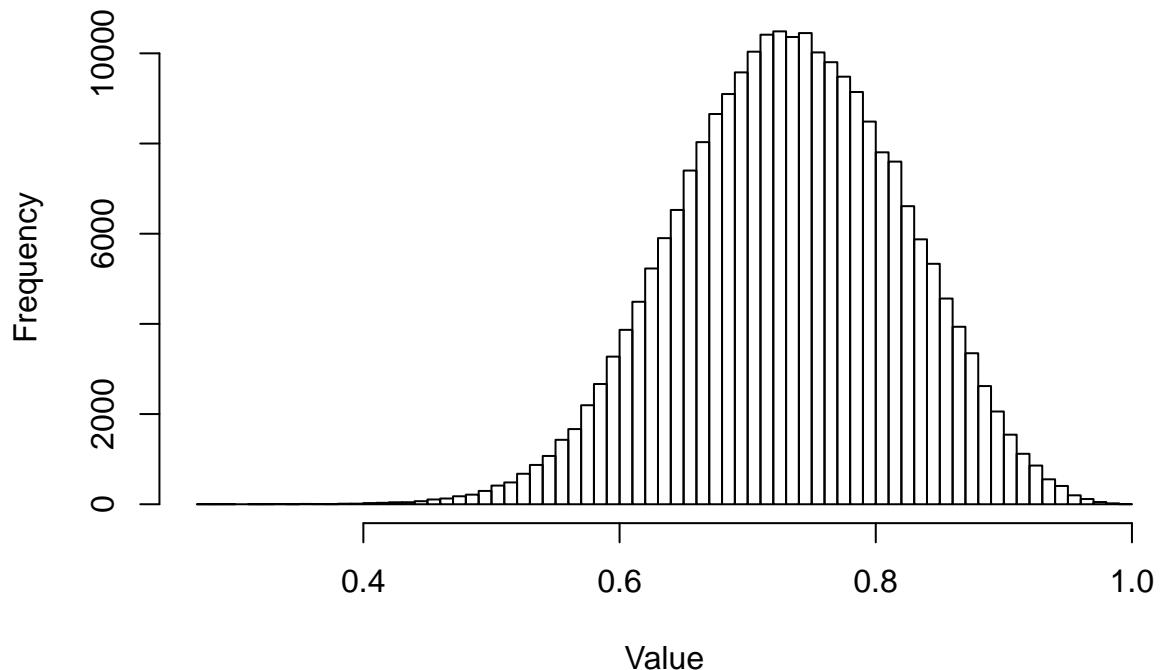
```
hist(old_subjects$openness_score, main='Openness score old subjects histogram', xlab='Value', ylab='Frequency')
```

Openness score old subjects histogram



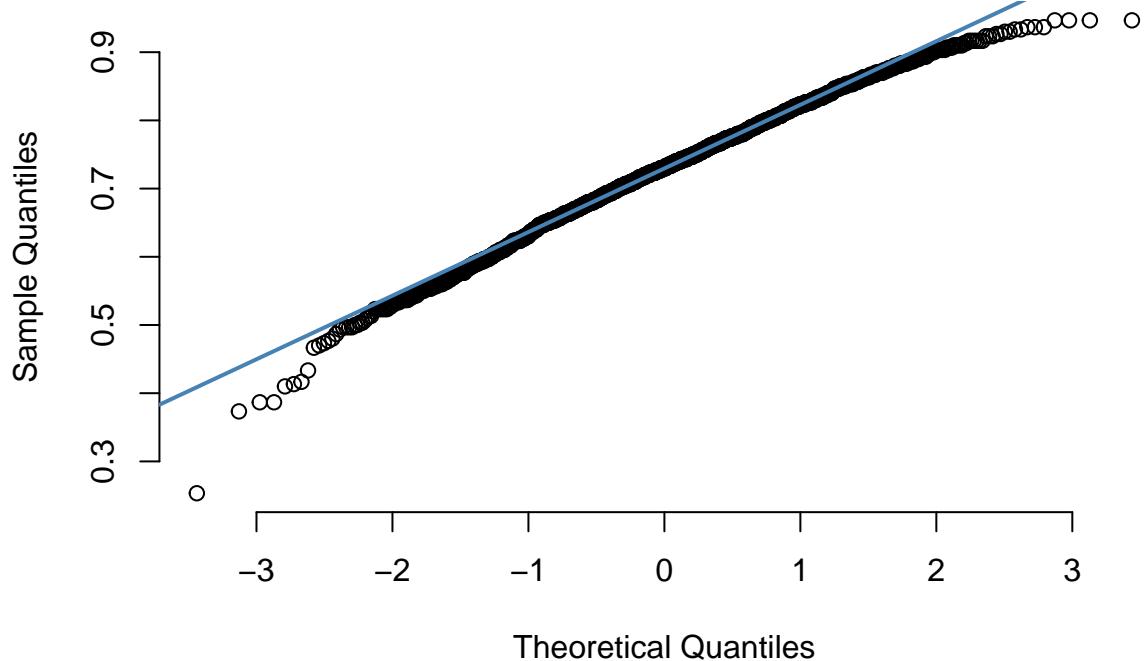
```
hist(young_subjects$openness_score,main='Openness score young subjects histogram',xlab='Value',ylab='Fr')
```

Openness score young subjects histogram

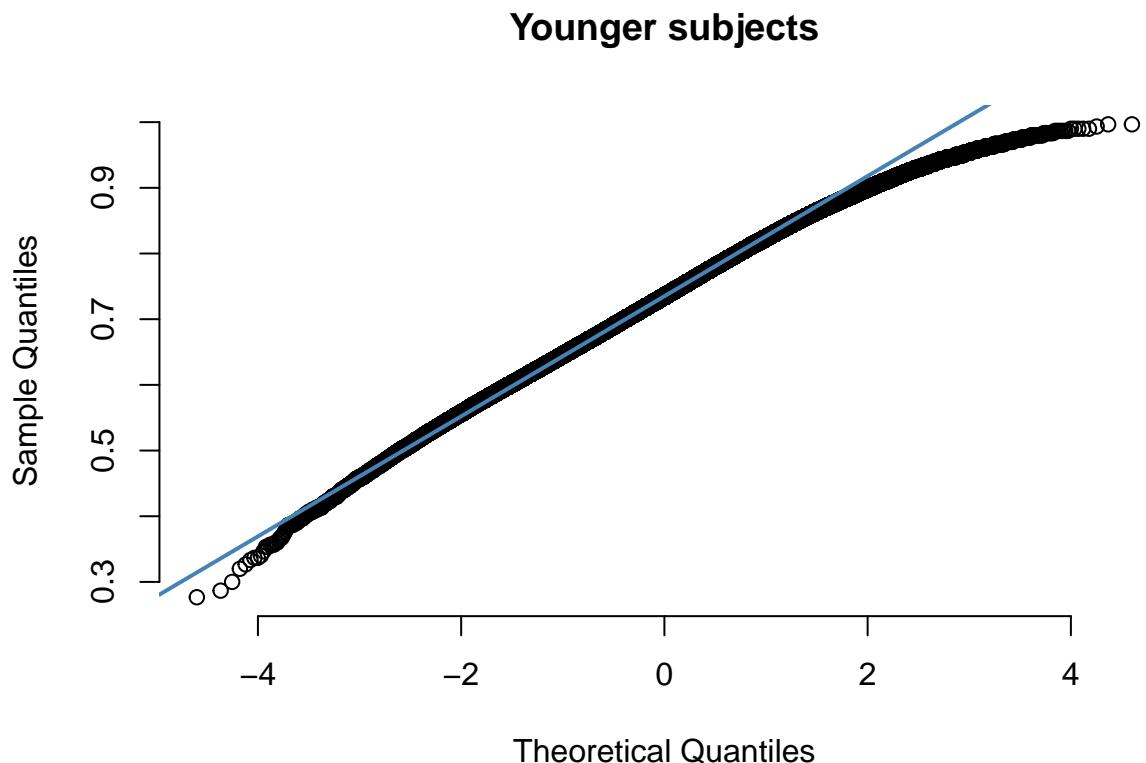


```
qqnorm(old_subjects$openness_score, pch = 1, frame = FALSE, main='Older subjects')
qqline(old_subjects$openness_score, col = "steelblue", lwd = 2)
```

Older subjects



```
qqnorm(young_subjects$openness_score, pch = 1, frame = FALSE, main='Younger subjects')
qqline(young_subjects$openness_score, col = "steelblue", lwd = 2)
```



Iz histograma i qq-plotova za openness_score mlađih i starih ispitanika može se naslutiti kako podaci pripadaju normalnoj razdiobi. Ta pretpostavka pokušat će se dodatno potvrditi provođenjem Lillieforsove inačice Kolmogorov-Smirnovljevog testa. Za sve testove koristi se razina značajnosti $\alpha = 0.05$.

Testiranje normalnosti - Lillieforsova inačica Kolmogorov-Smirnovljevog testa

Testiranje normalnosti otvorenosti za starije ispitanike

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

```
## 
##  Lilliefors (Kolmogorov-Smirnov) normality test
## 
##  data: old_subjects$openness_score
##  D = 0.031255, p-value = 0.000517
```

S obzirom da je $p\text{-vrijednost} = 0.000517 < \alpha$, nulta hipoteza o pripadnosti podataka o otvorenosti starih ispitanika normalnoj raspodjeli se odbacije u korist alternativne hipoteze.

Testiranje normalnosti faktora otvorenosti za mlađe ispitanike

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

```

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: young_subjects$openness_score
## D = 0.018819, p-value < 2.2e-16

```

S obzirom da je p-vrijednost $< 2.2\text{e-}16 < \alpha$, nulta hipoteza o pripadnosti podataka o otvorenosti mlađih ispitanika normalnoj raspodjeli se odbacije u korist alternativne hipoteze.

Testiranje jednakosti varijanci otvorenosti za stare i mlade ispitanike

Testiranje jednakosti varijanci provodi se F testom. Prepostavlja se nezavisnost uzorka.

Hipoteze:

$$H_0 : \sigma_{old}^2 = \sigma_{young}^2 \quad H_1 : \sigma_{old}^2 \neq \sigma_{young}^2$$

```
var.test(old_subjects$openness_score, young_subjects$openness_score)
```

```

## 
## F test to compare two variances
## 
## data: old_subjects$openness_score and young_subjects$openness_score
## F = 1.1715, num df = 1709, denom df = 238014, p-value = 2.084e-06
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.096503 1.254507
## sample estimates:
## ratio of variances
## 1.171521

```

S obzirom da je p-vrijednost $= 2.084\text{e-}06 < \alpha$ obzirujući da je $H_0 : \sigma_{old}^2 = \sigma_{young}^2$ u korist alternativne hipoteze $H_1 : \sigma_{old}^2 \neq \sigma_{young}^2$. Zaključujemo da podatci govore u prilog alternativnoj hipotezi i u testovima koji slijede koristit će se informacija o nejednakosti varijanci dobivena ovim testom.

Testiranje jednakosti srednjih vrijednosti otvorenosti za stare i mlade ispitanike

Testiranje jednakosti srednjih vrijednosti T testom zahtjeva da podaci dolaze iz normalne razdiobe. Prethodno provedenim Lillieforsovim testom normalnosti pokazalo se da to ne vrijedi niti za podatke o otvorenosti mlađih niti starih ispitanika. Ipak, s obzirom da histogrami i qq-plot grafovi govore u prilog normalnosti. Uvezši u obzir oblik grafova i veliku osjetljivost Lillieforstovog testa normalnosti, možemo nastaviti testiranje uz prepostavku normalnosti podataka.

Testiranje jednakosti srednjih vrijednosti otvorenosti za stare i mlade ispitanike provodi se kao T test za podatke s različitim varijancama, kako je pokazano prethodnim testom. Prepostavlja se nezavisnost uzorka.

Hipoteze:

$$H_0 : \mu_{old} = \mu_{young} \quad H_1 : \mu_{old} \neq \mu_{young}$$

```
t.test(old_subjects$openness_score, young_subjects$openness_score, alt = "two.sided", var.equal = FALSE)

##
## Welch Two Sample t-test
##
## data: old_subjects$openness_score and young_subjects$openness_score
## t = -3.155, df = 1730, p-value = 0.001633
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.011711273 -0.002732275
## sample estimates:
## mean of x mean of y
## 0.7265828 0.7338046
```

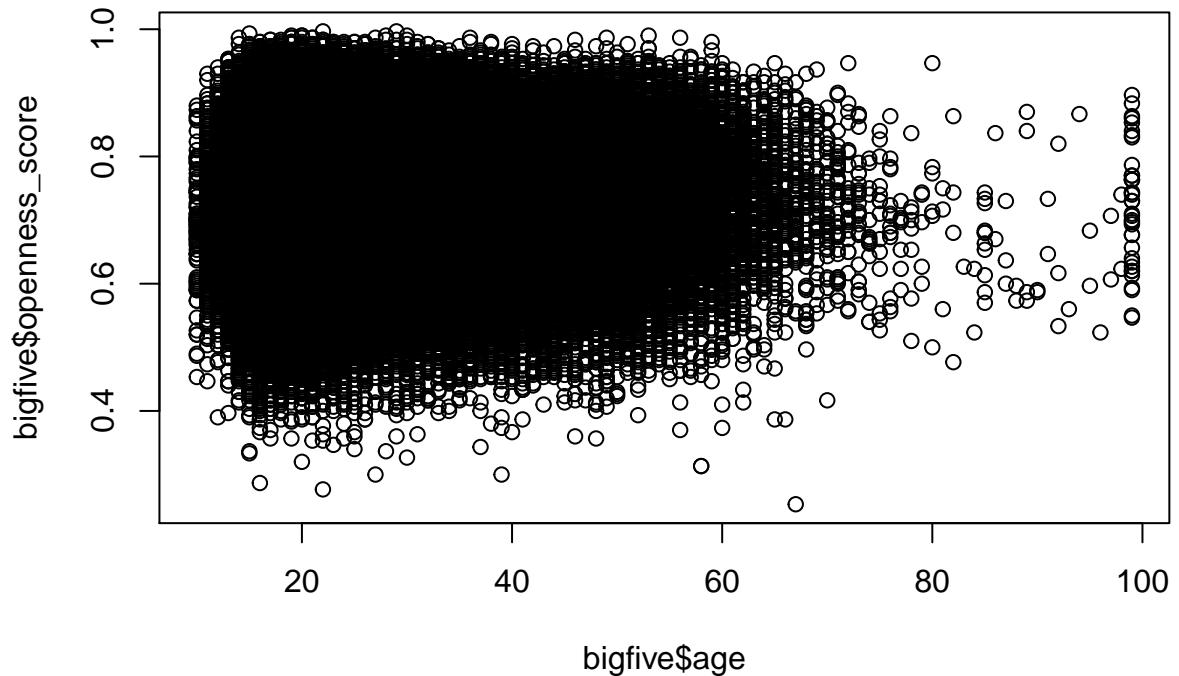
p-vrijednost = 0.00163 < α , stoga se hipoteza $H_0 : \mu_{old} = \mu_{young}$ odbacuje u korist alternativne hipoteze $H_1 : \mu_{old} \neq \mu_{young}$.

Statističkim testiranjem pokazano je kako postoji statistički značajna razlika između srednjih vrijednosti otvorenosti mladih i starih ispitanika.

Model linearne regresije za modeliranja odnosa godina i otvorenosti kod ispitanika

S obzirom na činjenicu da smo prethodnim testovima pokazali da je razlika između srednjih vrijednosti otvorenosti za mlade i stare ispitanike statistički značajna, vrijedilo bi se pozabaviti pitanjem odnosa između varijabli godina (age) i otvorenosti (openness_score). Za modeliranje ovisnosti koristi se model linearne regresije. Kako bi se dobio dojam o mogućoj zavisnosti varijable koja ocjenjuje otvorenost ispitanika i njihovih godina, prikazan je grafički prikaz u obliku scatter-plota.

```
plot(bigfive$age,bigfive$openness_score)
```



Ipak, na temelju izgleda grafa moglo bi se zaključiti da veza između varijabli ne postoji ili nije statistički značajna. Kako bi se moglo preciznije diskutirati o tom pitanju, nastaviti će se s modeliranjem njihovog odnosa modelom linearne regresije i provođenjem testova nad modelom, unatoč sugestiji grafičkog prikaza da je korelacija između varijabli vrlo slaba. Na temelju ocjene sposobnosti modela linearne regresije da modelira prikupljene podatke, ocjeniti će se i njihova linearna zavisnost. Izlazna varijabla u modelu je openness_score, dok je varijabla age regresor. S obzirom da je priutan samo jedan regresor, koristi se jednostavna linearna regresija.

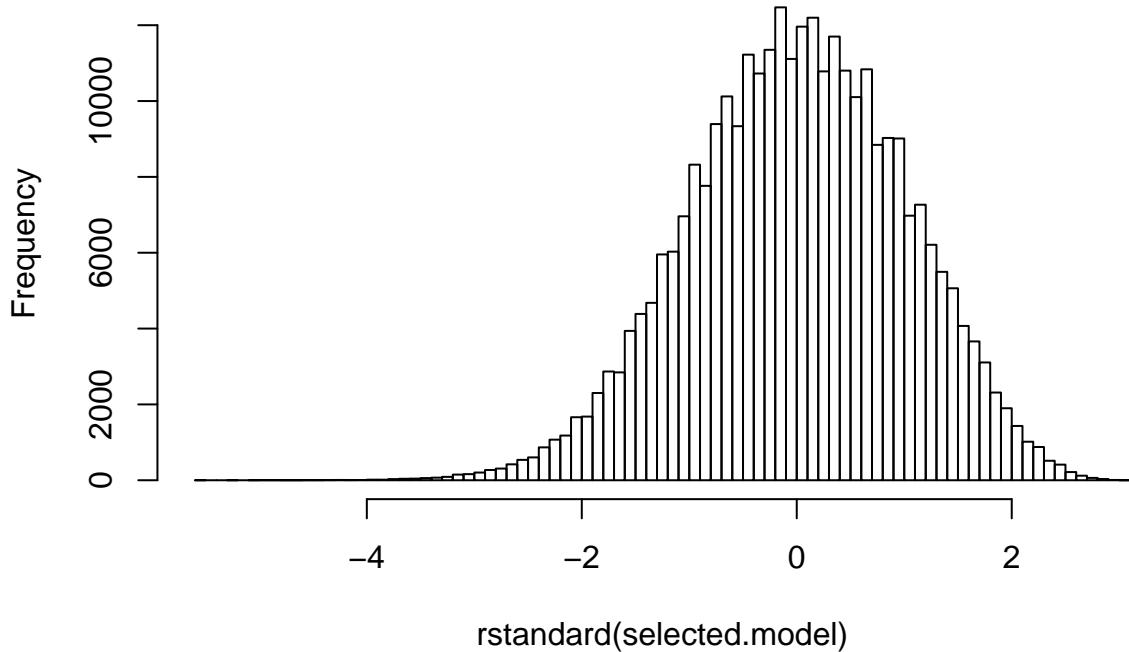
```
selected.model = lm(openness_score~age, data=bigfive)
```

Provjera pretpostavki modela

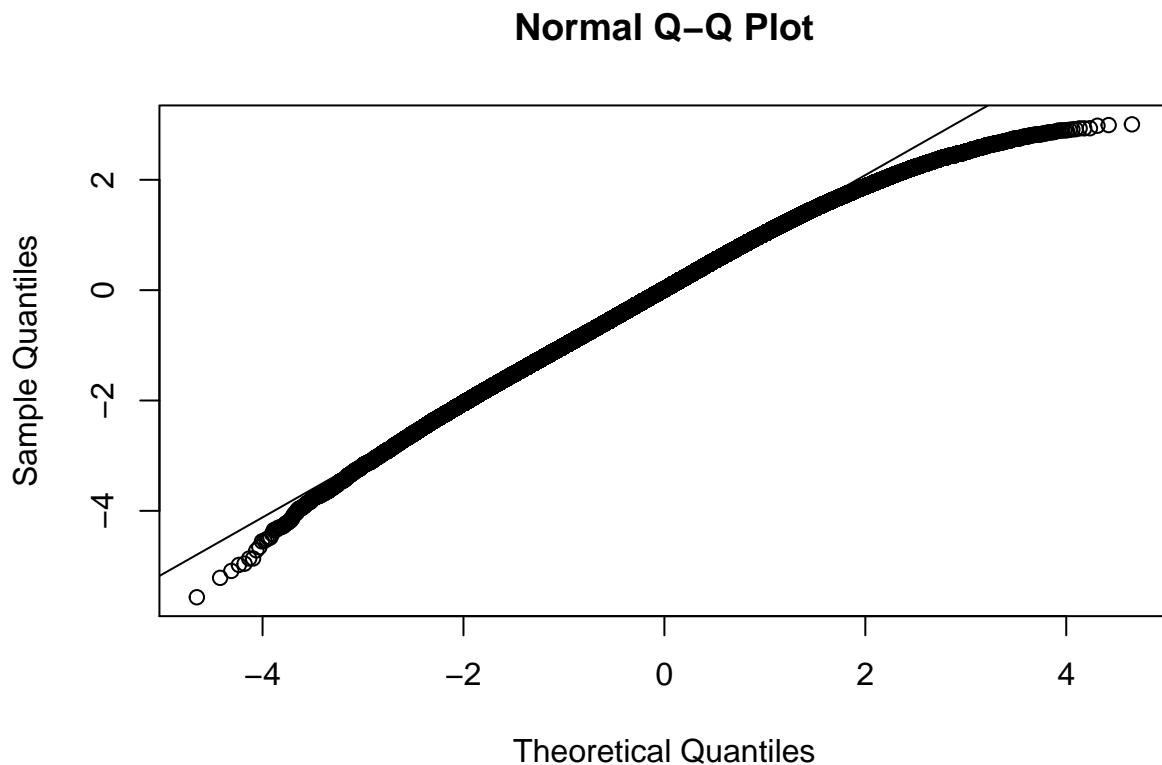
Prije svega treba provjeriti pretpostavke modela. Prva pretpostavka modela linearne regresije je da su reziduali normalno distribuirani i imaju homogenu varijancu. S obzirom da je korišten model jednostavne regresije, nije potrebno baviti se pretpostavkom o slaboj međusobnoj koreliranosti regresora.

```
hist(rstandard(selected.model), breaks=100)
```

Histogram of rstandard(selected.model)



```
qqnorm(rstandard(selected.model))  
qqline(rstandard(selected.model))
```



Histogram reziduala je vrlo obećavajuć te snažno podupire prepostavku o normalnosti reziduala. Ipak, potrebno je provesti i Lillieforsovu inačicu KS testa kako bi se normalnost insinuirana histogramom potvrdila.

```
ks.test(rstandard(selected.model), 'pnorm')

## Warning in ks.test(rstandard(selected.model), "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test

##
## One-sample Kolmogorov-Smirnov test
##
## data: rstandard(selected.model)
## D = 0.01445, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

S obzirom da je p-vrijednost $< \alpha$, nulta hipoteza o normalnosti standardiziranih reziduala se odbija u korist alternativnoj hipotezi. S obzirom na izgled histograma, izgled qq-plota i robusnost T-testa, ipak se može govoriti o dovoljnoj normalnosti reziduala za daljnja testiranja.

Za ocjenu kvalitete modela koristi se koeficijent determinacije R^2 , koji odgovara udjelu varijance zavisne varijable koju objašnjava dani model.

```
summary(selected.model)
```

```
##
```

```

## Call:
## lm(formula = openness_score ~ age, data = bigfive)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48752 -0.05958  0.00110  0.06259  0.26325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.7297757  0.0004282 1704.09 <2e-16 ***
## age         0.0001654  0.0000158   10.46 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08759 on 307311 degrees of freedom
## Multiple R-squared:  0.0003563, Adjusted R-squared:  0.000353
## F-statistic: 109.5 on 1 and 307311 DF, p-value: < 2.2e-16

```

R^2 za procjenjeni model iznosi 0.0003563, što je izrazito niska vrijednost. Također, p-vrijednost za F-test signifikantnosti modela je daleko ispod α . Uzimajući ove dvije ocjene u obzir, možemo sa velikom sigurnošću tvrditi kako je model linearne regresije neadekvatan odabir za modeliranje odnosa između godina i otvorenosti ispitanika. S obzirom na izgled scatter plota odnosa te dvije varijable, može se reći da je rezultat i očekivan.

#Savjesnost između regija ##Uvod

U ovom dijelu bavimo se pitanjem imaju li neke regije značajno različite rezultate u određenom faktoru. Primjerice je li opravдан mit o visokoj savjesnosti populacije istočne Azije naspram populacije drugih kontinenata.

Regije sam podijelio na osnovu UN-ove podjele (izvor:“<https://unstats.un.org/sdgs/indicators/regional-groups/>”).

```

`%!in%` <- Negate(`%in%`)

region_east_asian =c('Brunei', 'Cambodia', 'China', 'Hong Kong', 'Indonesia', 'Japan', 'Macau', 'Malaysia', 'Philippines', 'Singapore', 'Thailand')

east_asia = bigfive[bigfive$country %in% region_east_asian,]

rest_regions = bigfive[bigfive$country %!in% region_east_asian,]

summary(bigfive$conscientiousness_score)

##      Min. 1st Qu.  Median  Mean 3rd Qu.  Max.
##  0.2067  0.6300  0.7067  0.7020  0.7767  1.0000

summary(east_asia$conscientiousness_score)

##      Min. 1st Qu.  Median  Mean 3rd Qu.  Max.
##  0.3167  0.6333  0.6967  0.6946  0.7567  1.0000

summary(rest_regions$conscientiousness_score)

##      Min. 1st Qu.  Median  Mean 3rd Qu.  Max.
##  0.2067  0.6300  0.7067  0.7023  0.7800  1.0000

```

```

cat('Varijanca savjesnosti za istočnu aziju ', var(east_asia$conscientiousness_score), '\n')

## Varijanca savjesnosti za istočnu aziju  0.009029865

cat('Varijanca savjesnosti za ostatak svijeta ', var(rest_regions$conscientiousness_score), '\n')

## Varijanca savjesnosti za ostatak svijeta  0.01162203

boxplot(east_asia$conscientiousness_score, rest_regions$conscientiousness_score,
        names = c('Istočna Azija\ savjesnost','Ostatak svijeta savjesnost'),
        main='Boxplot za istočnu Aziju i ostatak svijeta savjesnost')

## Warning in axis(side = 1, at = 1:2, labels = c("Istočna Azija savjesnost", :
## conversion failure on 'Istočna Azija savjesnost' in 'mbcsToSbcs': dot
## substituted for <c4>

## Warning in axis(side = 1, at = 1:2, labels = c("Istočna Azija savjesnost", :
## conversion failure on 'Istočna Azija savjesnost' in 'mbcsToSbcs': dot
## substituted for <8d>

## Warning in axis(side = 1, at = 1:2, labels = c("Istočna Azija savjesnost", :
## conversion failure on 'Istočna Azija savjesnost' in 'mbcsToSbcs': dot
## substituted for <c4>

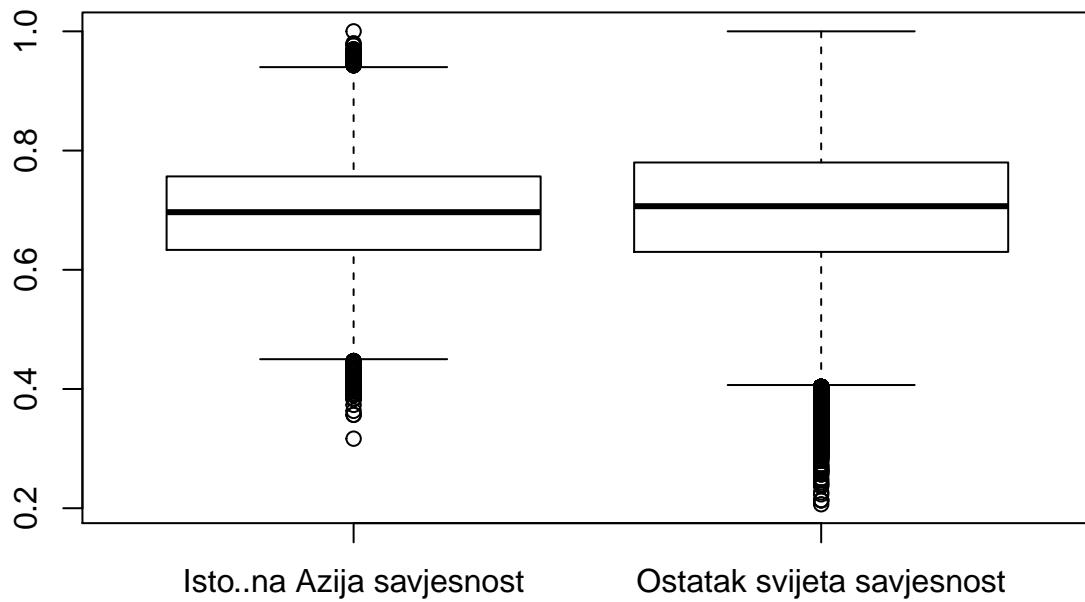
## Warning in axis(side = 1, at = 1:2, labels = c("Istočna Azija savjesnost", :
## conversion failure on 'Istočna Azija savjesnost' in 'mbcsToSbcs': dot
## substituted for <8d>

## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'Boxplot za istočnu Aziju i ostatak svijeta savjesnost' in
## 'mbcsToSbcs': dot substituted for <c4>

## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'Boxplot za istočnu Aziju i ostatak svijeta savjesnost' in
## 'mbcsToSbcs': dot substituted for <8d>

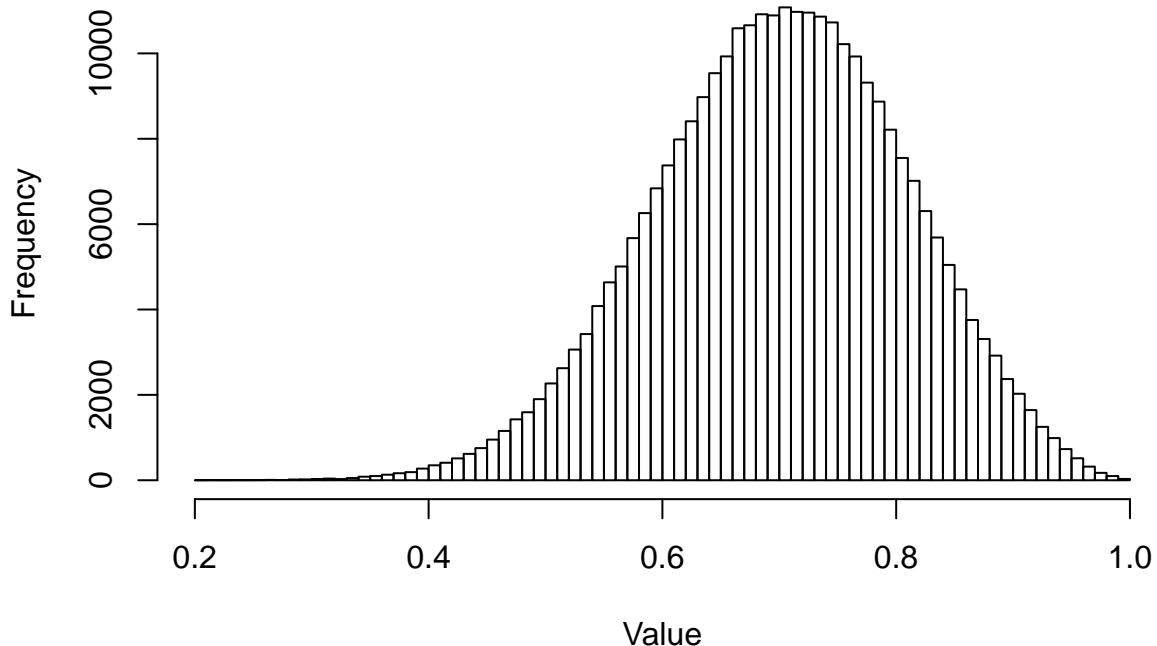
```

Boxplot za isto..nu Aziju i ostatak svijeta savjesnost



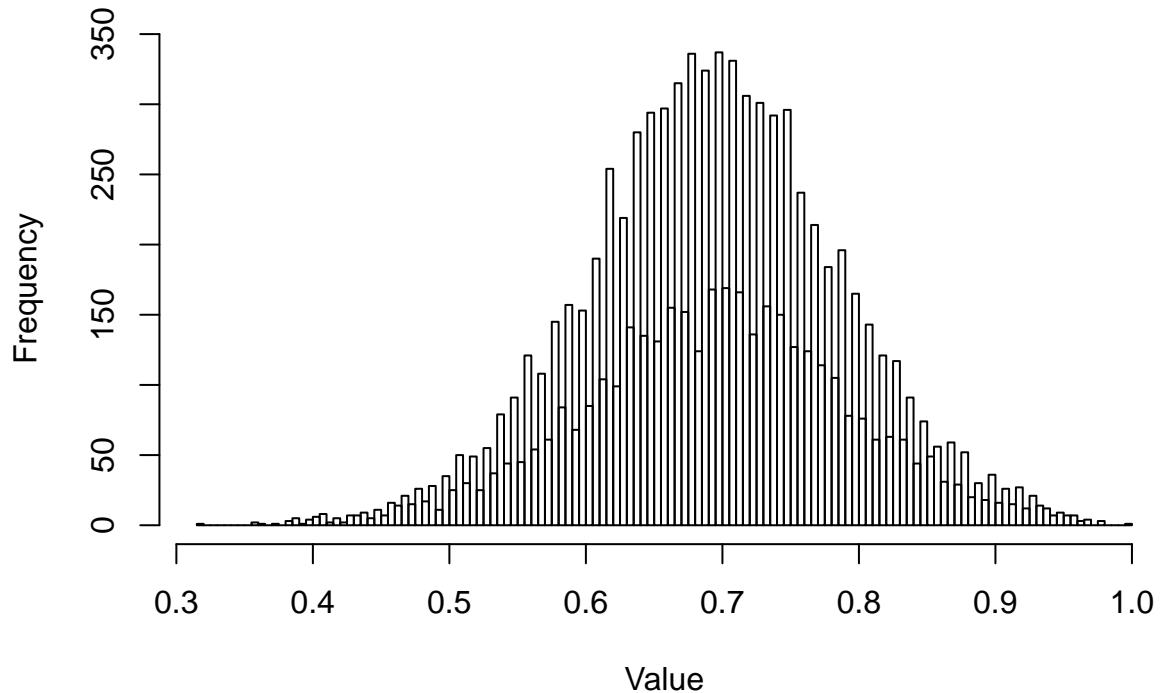
```
hist(bigfive$conscientiousness_score, main='Savjesnost histogram', xlab='Value', ylab='Frequency', breaks=
```

Savjesnost histogram



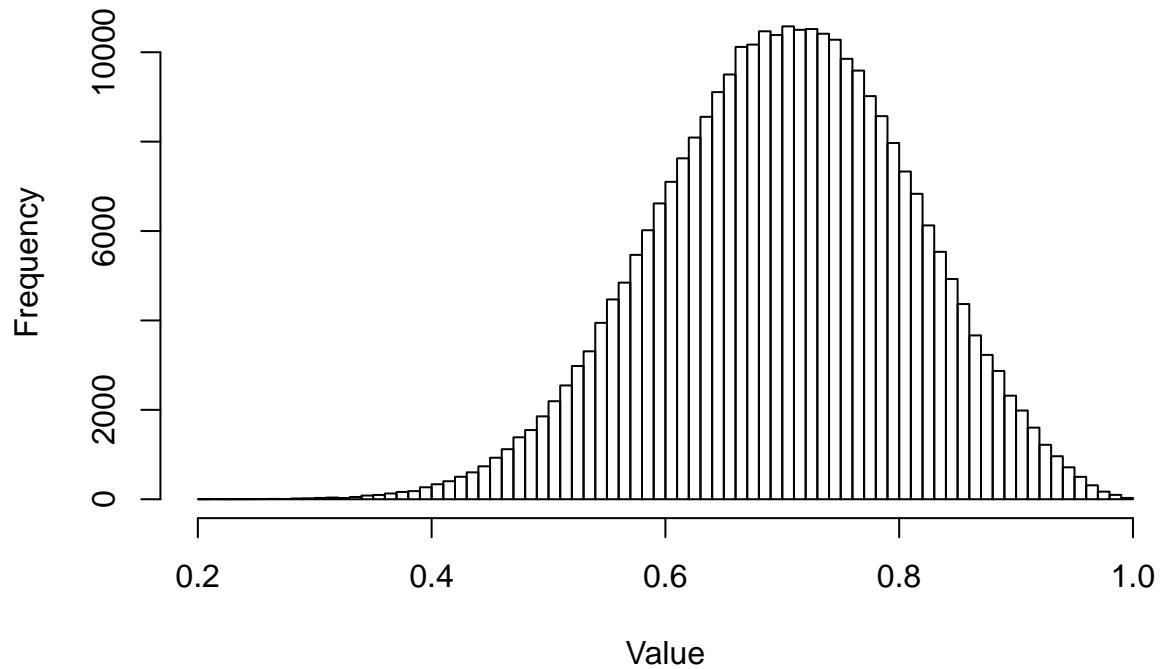
```
hist(east_asia$conscientiousness_score,main='Histogram savjesnosti istočne Azije',xlab='Value',ylab='Fr...  
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):  
## conversion failure on 'Histogram savjesnosti istočne Azije' in 'mbcsToSbcs': dot  
## substituted for <c4>  
  
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):  
## conversion failure on 'Histogram savjesnosti istočne Azije' in 'mbcsToSbcs': dot  
## substituted for <8d>
```

Histogram savjesnosti isto..ne Azije



```
hist(rest_regions$conscientiousness_score,main='Histogram savjesnosti ostatka svijeta',xlab='Value',ylab='Frequency')
```

Histogram savjesnosti ostatka svijeta



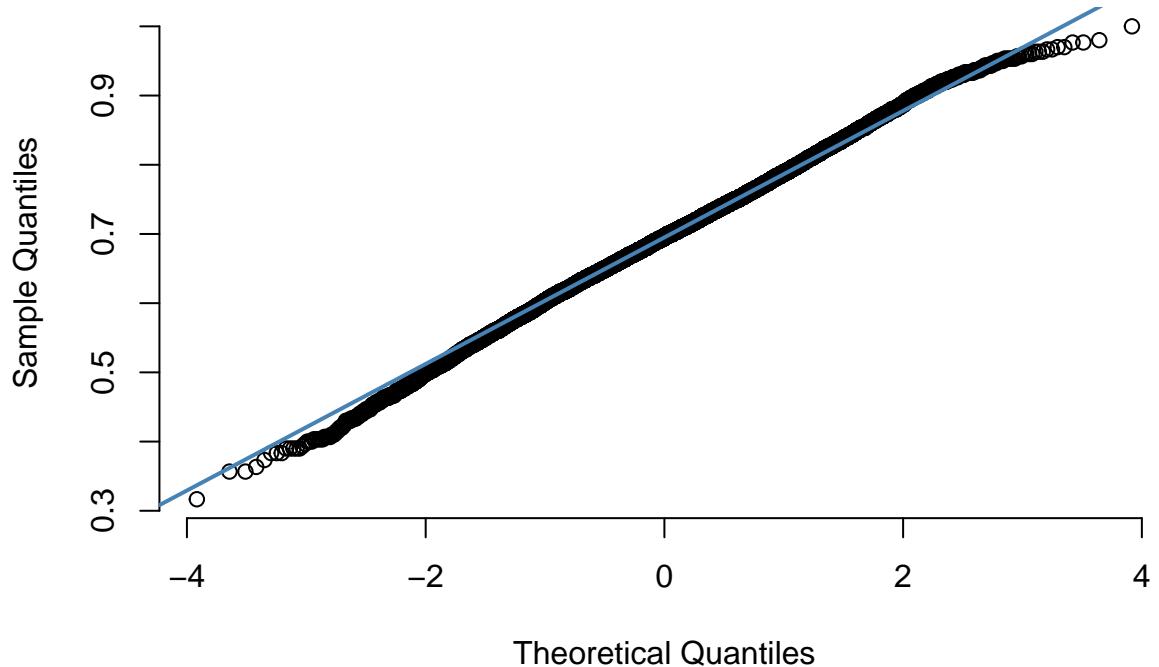
```
qqnorm(east_asia$conscientiousness_score, pch = 1, frame = FALSE, main='Istočna Azija')
```

```
## Warning in title(...): conversion failure on 'Istočna Azija' in 'mbcsToSbcs':  
## dot substituted for <c4>
```

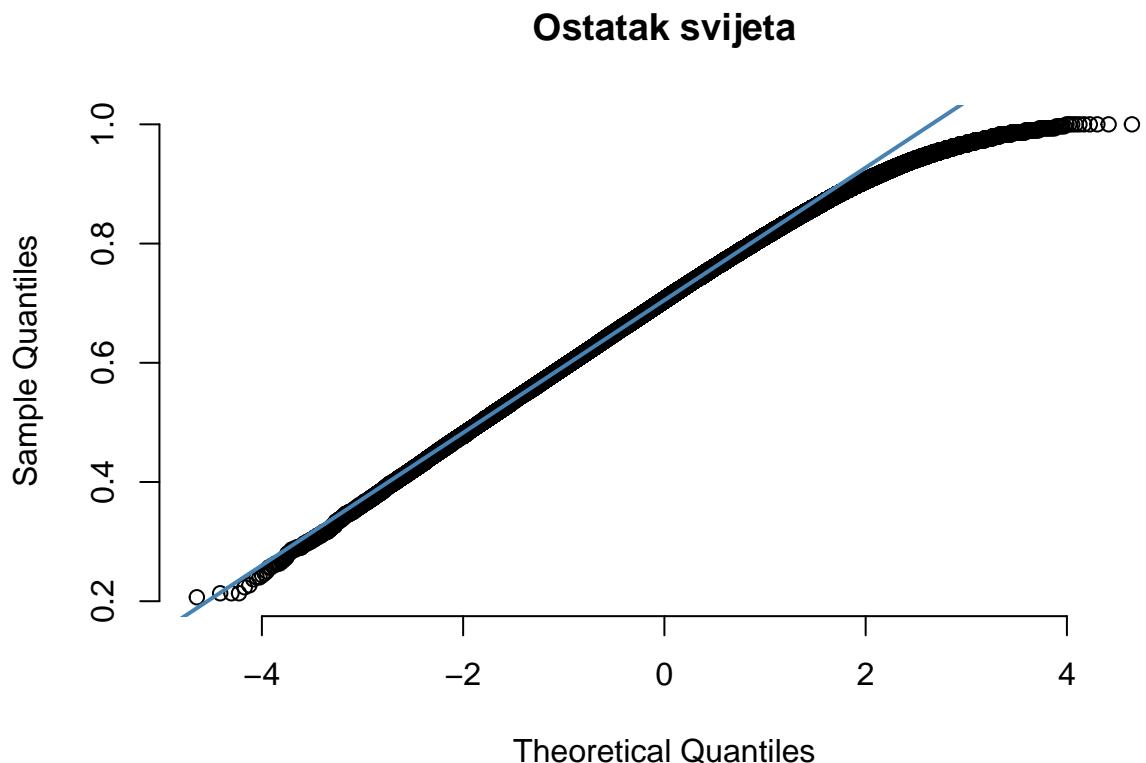
```
## Warning in title(...): conversion failure on 'Istočna Azija' in 'mbcsToSbcs':  
## dot substituted for <8d>
```

```
qqline(east_asia$conscientiousness_score, col = "steelblue", lwd = 2)
```

Isto..na Azija



```
qqnorm(rest_regions$conscientiousness_score, pch = 1, frame = FALSE, main='Ostatak svijeta')
qqline(rest_regions$conscientiousness_score, col = "steelblue", lwd = 2)
```



Već iz priloženih grafova mogu naslutiti da su obje distribucije normalne, da je rasipanje veće kod populacije istočne Azije nego u ostatku svijeta (prepostavljam da je to najviše zbog činjenice da je set podataka za ostatak svijeta mnogo veći) te suprotno početnom pitanju izgleda da je savjesnost ispitanika istočne Azije manja nego u ostatku svijeta.

##Testiranje normalnosti distribucije

Hipoteze:

H_0 : distribucija $east_asiaconscientiousness_score$ pripada normalnoj razdiobi
 H_1 : distribucija $east_asiaconscientiousness_score$ NE pripada normalnoj razdiobi

```
lillie.test(east_asia$conscientiousness_score)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
##  data:  east_asia$conscientiousness_score
##  D = 0.017328, p-value = 5.76e-08
```

Odbacujem nultu hipotezu u korist alternative da se savjesnost ispitanika istočne Azije ne ravna po normalnoj distribuciji.

Hipoteze:

H_0 : distribucija $rest_regionsconscientiousness_score$ pripada normalnoj razdiobi
 H_1 : distribucija $rest_regionsconscientiousness_score$ NE pripada normalnoj razdiobi

```
lillie.test(rest_regions$conscientiousness_score)
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: rest_regions$conscientiousness_score  
## D = 0.020199, p-value < 2.2e-16
```

Odbacujem nultu hipotezu u korist alternative da se savjesnost ispitanika iz ostalih regija ne ravna po normalnoj distibuciji.

Uzevši u obzir preveliku osjetljivost ovog testa i činjenicu da je iz histograma vidljiva normalnost u nastavku prepostavljam normalnost za obje populacije.

##Testiranje jednakosti varijance savjesnosti

Prepostavljam nezavisnost uzoraka i izvodim F test.

Hipoteze:

$$H_0 : \sigma_{eastAsia}^2 = \sigma_{rest}^2 \quad H_1 : \sigma_{eastAsia}^2 \neq \sigma_{rest}^2$$

```
var.test(east_asia$conscientiousness_score, rest_regions$conscientiousness_score)
```

```
##  
## F test to compare two variances  
##  
## data: east_asia$conscientiousness_score and rest_regions$conscientiousness_score  
## F = 0.77696, num df = 11152, denom df = 296159, p-value < 2.2e-16  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.7565887 0.7981497  
## sample estimates:  
## ratio of variances  
## 0.7769608
```

Na osnovu male p vrijednosti odbacujem nultu hipotezu u korist alternative da su varijance distribucije različite

##Testiranje jednakosti razine savjesnosti

Koristim T test i prepostavljam nezavisnost varijabli i različitost varijanci te normalnu razdiobu.

Hipoteze:

$$H_0 : \mu_{eastAsia} = \mu_{rest} \quad H_1 : \mu_{eastAsia} > \mu_{rest}$$

```
t.test(east_asia$conscientiousness_score, rest_regions$conscientiousness_score, alt = "greater", var.equal = TRUE)
```

```
##  
## Welch Two Sample t-test  
##  
## data: east_asia$conscientiousness_score and rest_regions$conscientiousness_score  
## t = -8.3476, df = 12258, p-value = 1
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## -0.009206648      Inf
## sample estimates:
## mean of x mean of y
## 0.6945865 0.7022776
```

Uz veliku p vrijednost nisam uspio odbaciti nullu hipotezu i zaključujem nije opravdan mit o visokoj savjesnosti populacije istočne Azije.

##Savjesnost između više regija ###Uvod

Pošto se nisam uvjero u ispravnost mita da su stanovnici istočne Azije savjesniji od ostatka svijeta pitam postoji li uopće razlike u savjesnosti između stavnika različitih regija. Za primjer ću uzeti pet regija za koje imam sličan broj podataka. Moje pitanje glasi: "Postoji li razlika u savjesnosti između stavnovnika Zapada, latinske Amerike, južne Afrike, sjeverne Afrike i zapadne Azije(arapskog svijeta) i centralne Azije"

`summary(west$orange)` shows:

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 0.2000 0.6800 0.7067 0.7010 0.7767 1.0000
```

summary(latin_america\$consciousness_scores)

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 0.2167 0.6467 0.7267 0.7234 0.8067 0.9200
```

```
summary(africa$conscientiousness_score)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 0.2967 0.6400 0.7167 0.7139 0.7900 0.9867
```

```
summary(arabic_world$conscientiousness_score)

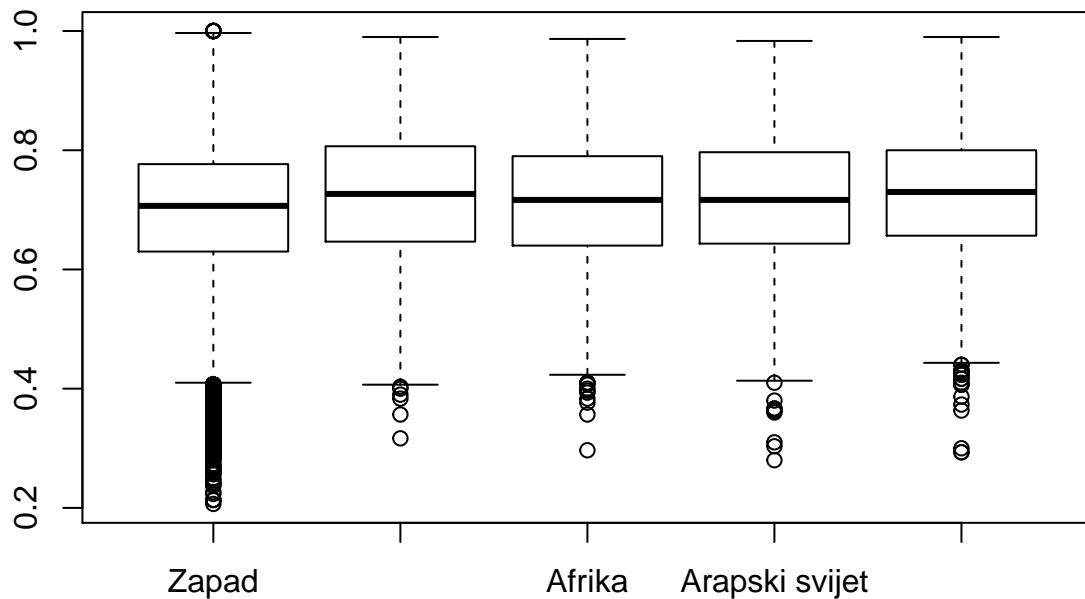
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 0.2800  0.6433  0.7167  0.7157  0.7958  0.9833
```

```
summary(central_asia$conscientiousness_score)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 0.2933  0.6567  0.7300  0.7253  0.8000  0.9900
```

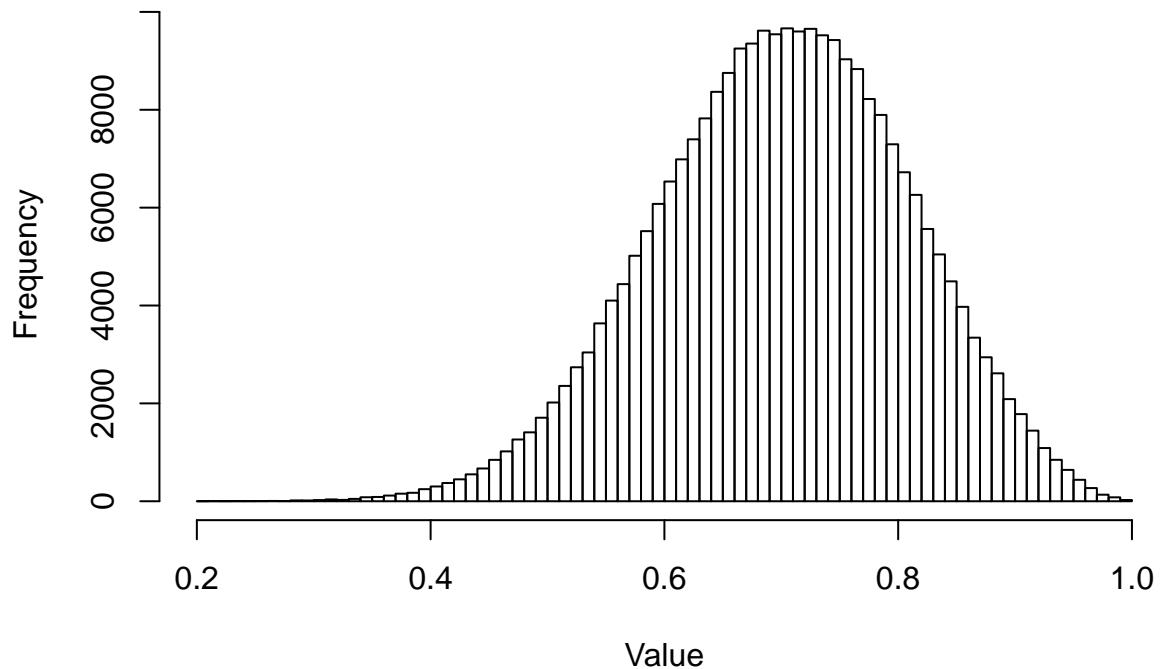
```
boxplot(west$conscientiousness_score,latin_america$conscientiousness_score,africa$conscientiousness_score,
        names = c('Zapad','Latinska amerika','Afrika','Arapski svijet','Centralna\n Azija'),
        main='Boxplot savjesnosti za zadane regije')
```

Boxplot savjesnosti za zadane regije



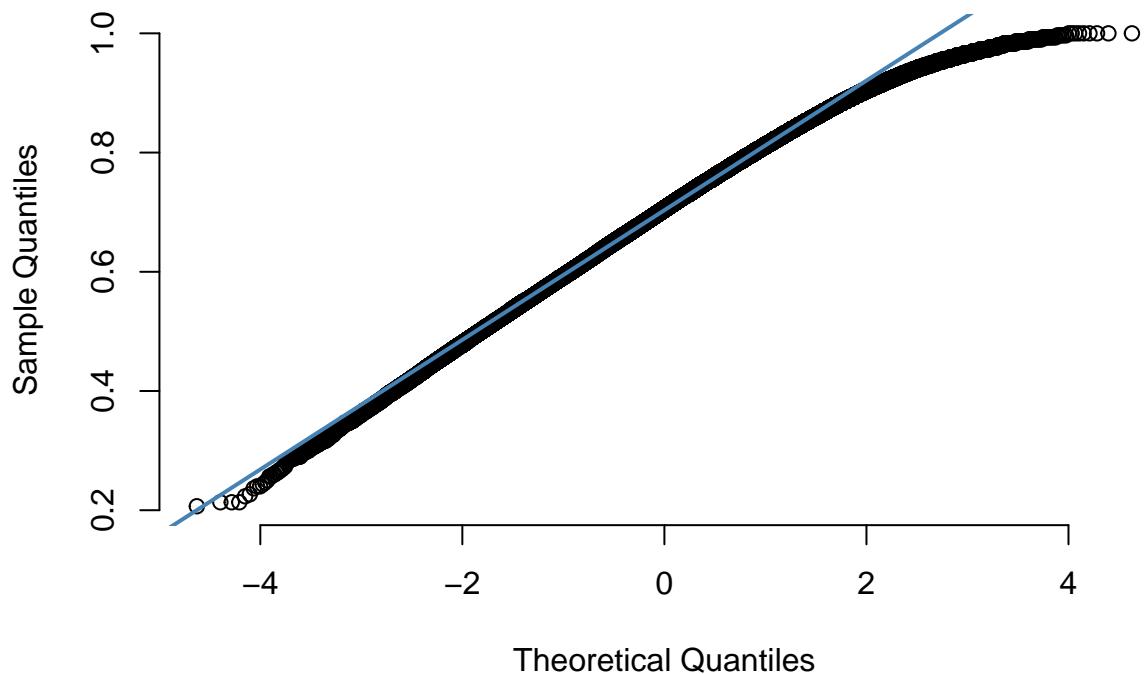
```
hist(west$conscientiousness_score,main='Histogram savjesnosti Zapada',xlab='Value',ylab='Frequency', br
```

Histogram savjesnosti Zapada



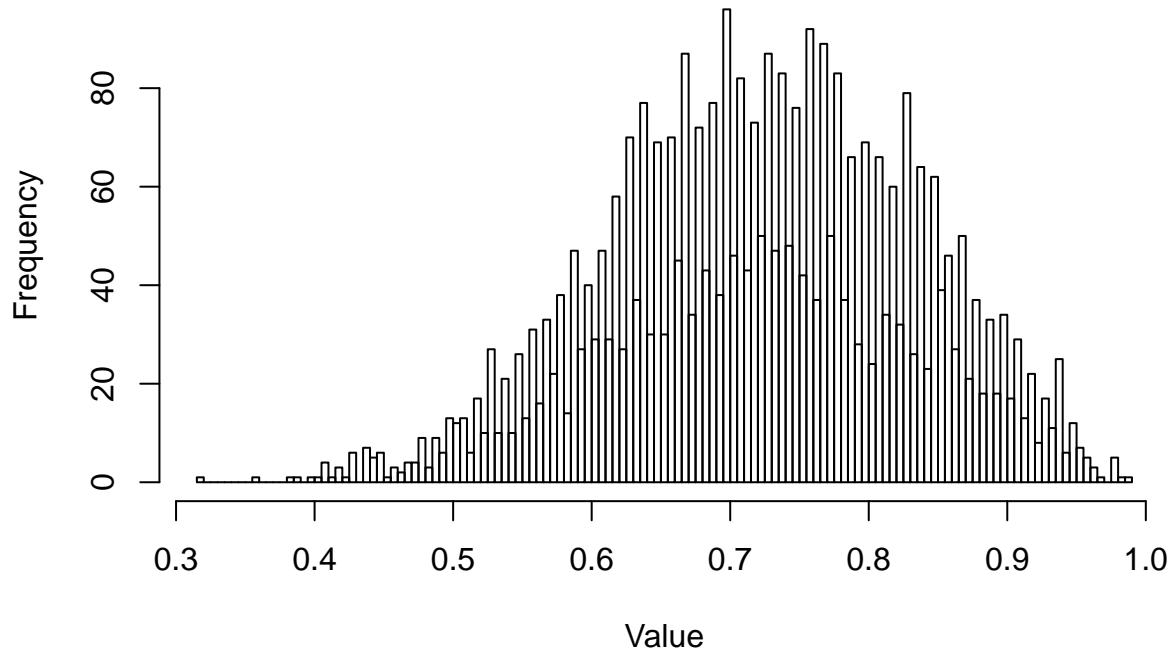
```
qqnorm(west$conscientiousness_score, pch = 1, frame = FALSE, main='Savjesnost Zapada')
qqline(west$conscientiousness_score, col = "steelblue", lwd = 2)
```

Savjesnost Zapada



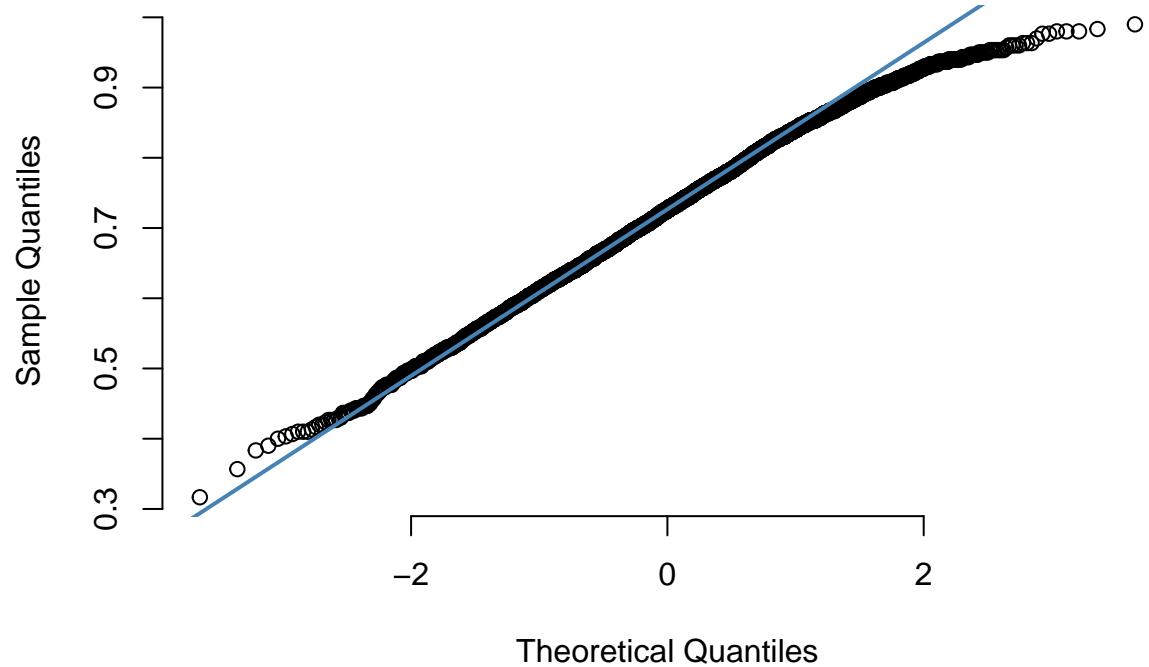
```
hist(latin_america$conscientiousness_score,main='Histogram savjesnosti Latinske Amerike',xlab='Value',y
```

Histogram savjesnosti Latinske Amerike



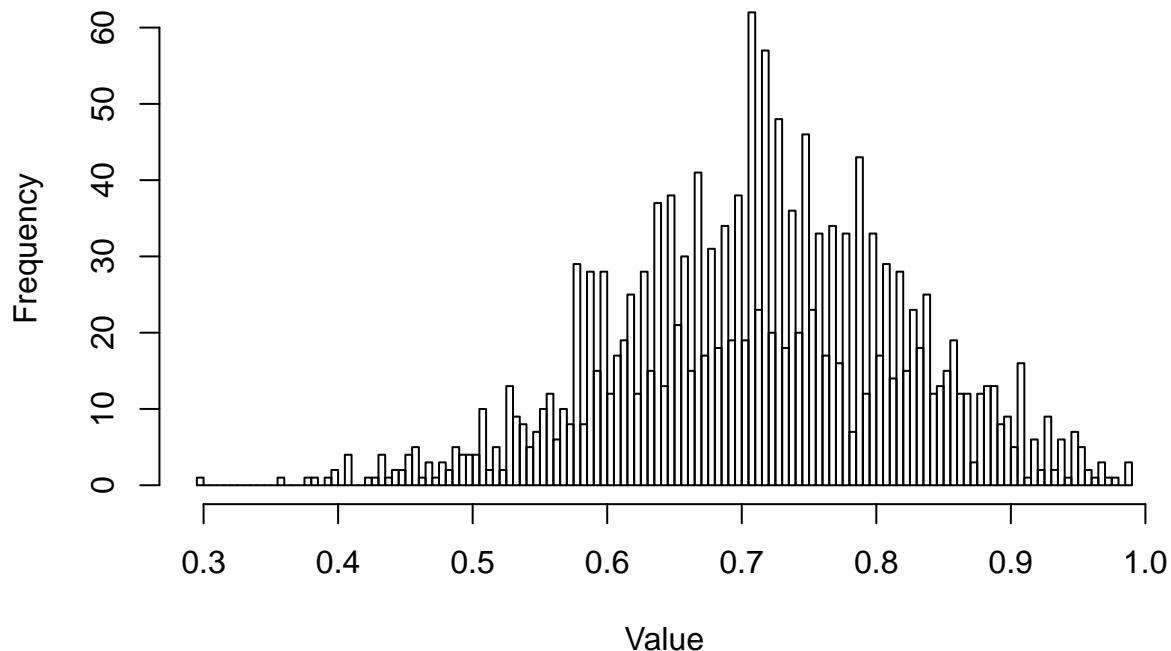
```
qqnorm(latin_america$conscientiousness_score, pch = 1, frame = FALSE, main='Savjesnost Latinske Amerike')
qqline(latin_america$conscientiousness_score, col = "steelblue", lwd = 2)
```

Savjesnost Latinske Amerike



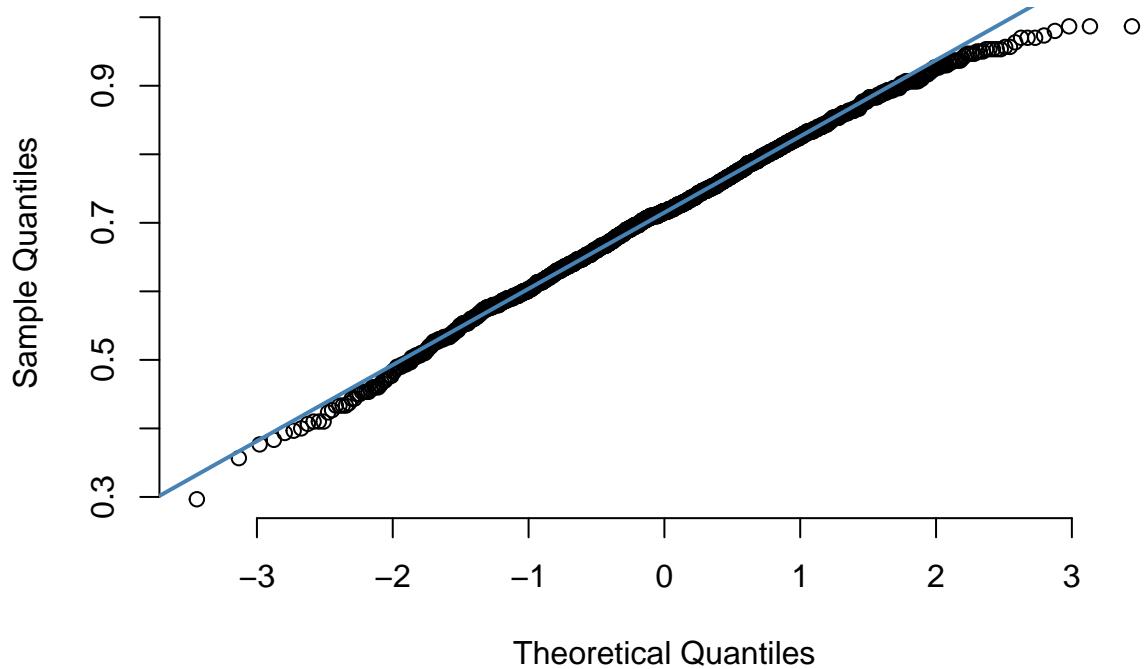
```
hist(africa$conscientiousness_score,main='Histogram savjesnosti Afrike',xlab='Value',ylab='Frequency',
```

Histogram savjesnosti Afrike



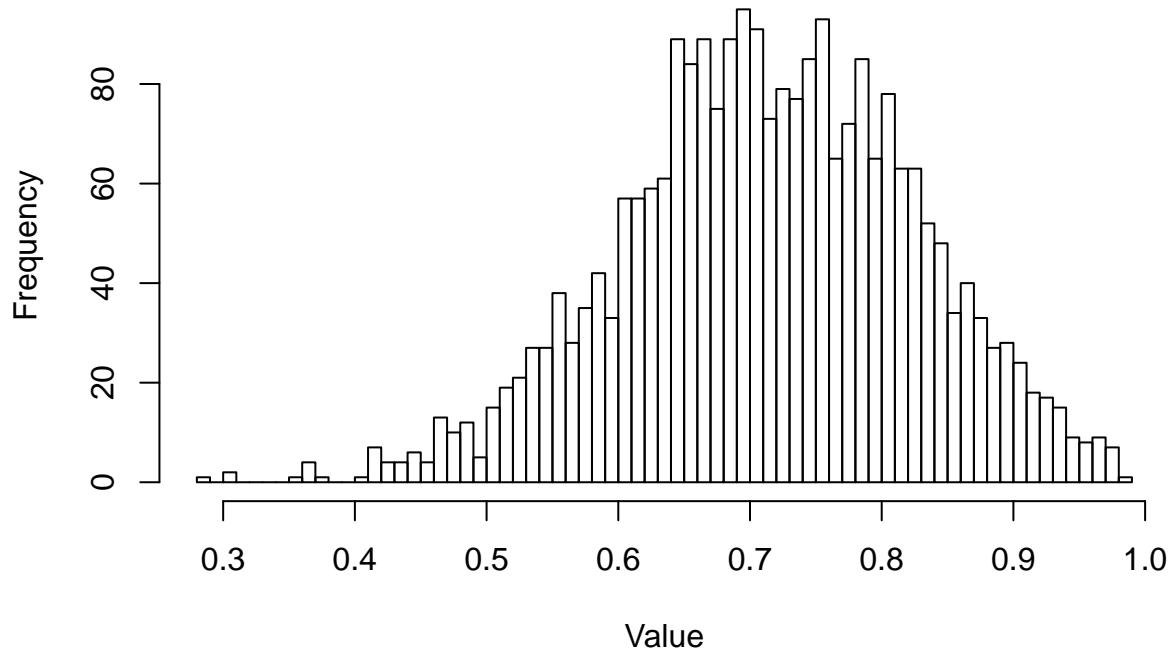
```
qqnorm(africa$conscientiousness_score, pch = 1, frame = FALSE, main='Savjesnost Afrike')
qqline(africa$conscientiousness_score, col = "steelblue", lwd = 2)
```

Savjesnost Afrike



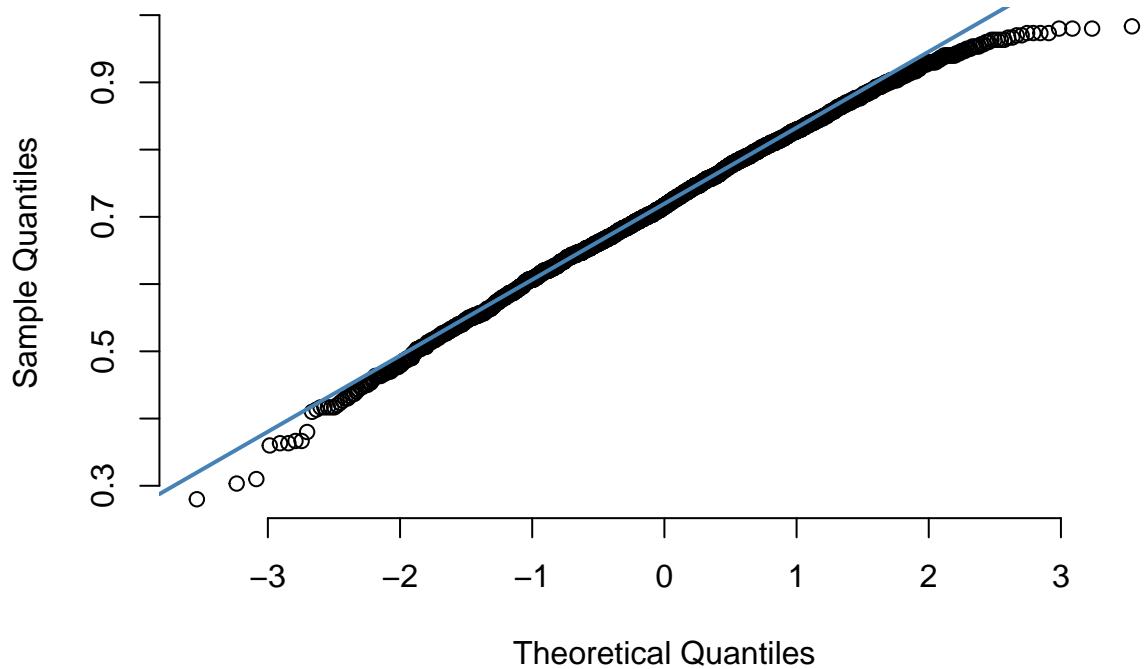
```
hist(arabic_world$conscientiousness_score,main='Histogram savjesnosti Arapskog svijeta',xlab='Value',ylab='Frequency')
```

Histogram savjesnosti Arapskog svijeta



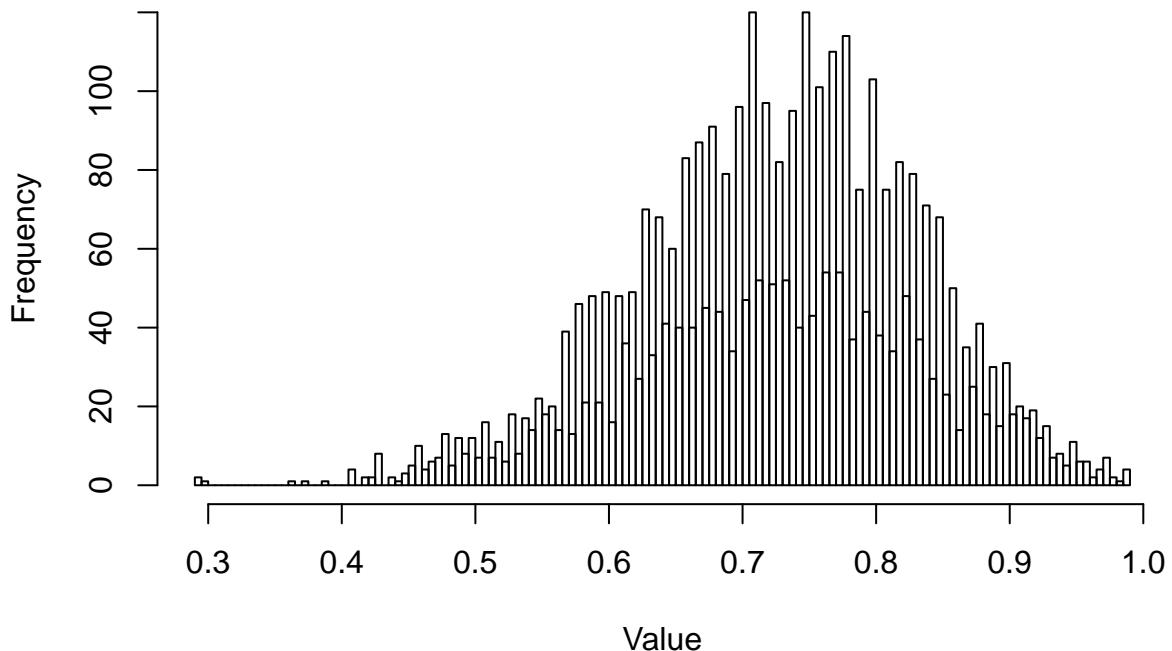
```
qqnorm(arabic_world$conscientiousness_score, pch = 1, frame = FALSE, main='Savjesnost Arapski svijet')
qqline(arabic_world$conscientiousness_score, col = "steelblue", lwd = 2)
```

Savjesnost Arapski svijet



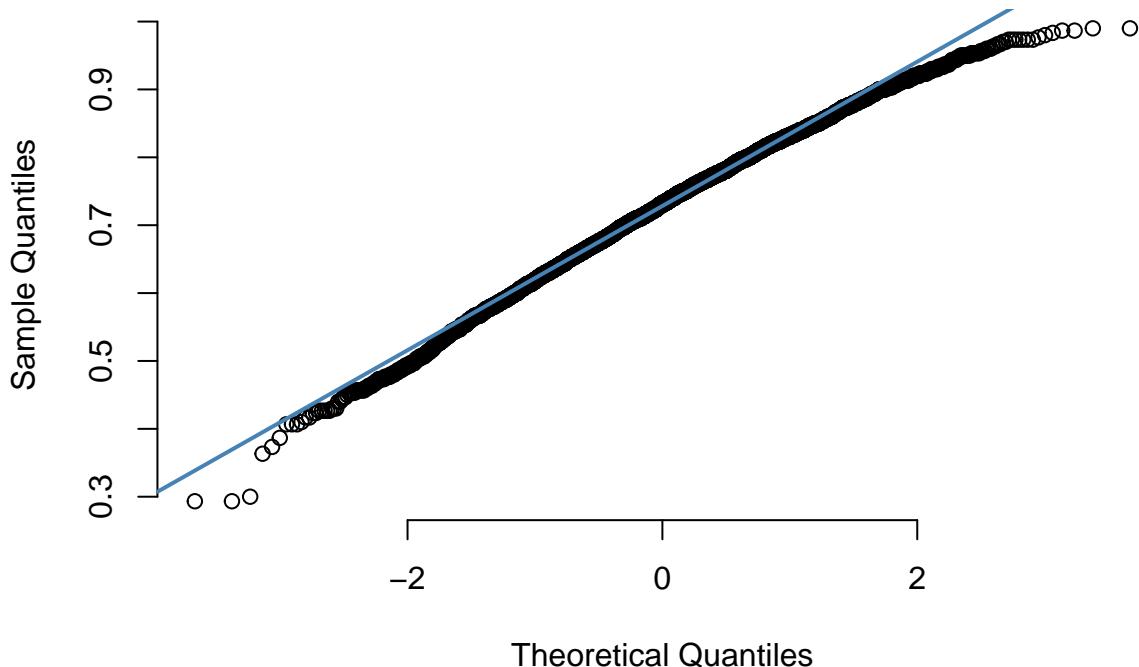
```
hist(central_asia$conscientiousness_score,main='Histogram savjesnosti centralne Azije',xlab='Value',ylab='Frequency')
```

Histogram savjesnosti centralne Azije



```
qqnorm(central_asia$conscientiousness_score, pch = 1, frame = FALSE, main='Savjesnost centralne Azije')
qqline(central_asia$conscientiousness_score, col = "steelblue", lwd = 2)
```

Savjesnost centralne Azije



Iz histograma i qq plota je vidljivo da se savjesnost svih regija ravna po normalnoj razdiobi.

```
###Testiranje normalnosti distribucije
```

Hipoteze:

H_0 : distribucija regije pripada normalnoj razdiobi H_1 : distribucija regije NE pripada normalnoj razdiobi

```
lillie.test(west$conscientiousness_score)
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: west$conscientiousness_score  
## D = 0.020534, p-value < 2.2e-16
```

```
lillie.test(latin_america$conscientiousness_score)
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: latin_america$conscientiousness_score  
## D = 0.026175, p-value = 3.018e-06
```

```
lillie.test(africa$conscientiousness_score)
```

```

##  

## Lilliefors (Kolmogorov-Smirnov) normality test  

##  

## data: africa$conscientiousness_score  

## D = 0.028427, p-value = 0.002351

lillie.test(arabic_world$conscientiousness_score)

##  

## Lilliefors (Kolmogorov-Smirnov) normality test  

##  

## data: arabic_world$conscientiousness_score  

## D = 0.022587, p-value = 0.005327

lillie.test(central_asia$conscientiousness_score)

##  

## Lilliefors (Kolmogorov-Smirnov) normality test  

##  

## data: central_asia$conscientiousness_score  

## D = 0.032478, p-value = 1.122e-10

```

Za savjesnost svake regije na Lillieforovom testu smo dobili izuzetno malenu p vrijednost te odbacujemo nultu hipotezu za svaku regiju. Zaključak na osnovu Lillieforova testa jest da se nijedna distribucija savjesnosti ne ravna po normalnoj razdiobi. Ali na osnovu histograma i qq plota u nastavku pretpostavljamo da se svaka distibucija savjesnosti ravna po normalnoj razdiobi.

###Testiranje homogenosti varijance između populacija

Nulta hipoteza je da su varijance savjesnosti svih regija jednake. Alternativa je da nisu odnosno da se barem jedan par varijanci nije jednak.

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 \\ H_1 : \neg H_0.$$

Za testiranje ove hipoteze koristimo Bartlettov test.

```

score = c(latin_america$conscientiousness_score,arabic_world$conscientiousness_score,central_asia$consci

west_vec = rep(c("west"),length(west$conscientiousness_score))
la_vec = rep(c("latin_america"),length(latin_america$conscientiousness_score))
afr_vec = rep(c("africa"),length(africa$conscientiousness_score))
arb_vec = rep(c("arabic_world"),length(arabic_world$conscientiousness_score))
asi_vec = rep(c("central_asia"),length(central_asia$conscientiousness_score))

region = c(la_vec,arb_vec,asi_vec)

df = data.frame(region,score)

bartlett.test(df$score ~ df$region)

```

```

##  

##  Bartlett test of homogeneity of variances  

##  

## data: df$score by df$region  

## Bartlett's K-squared = 9.0166, df = 2, p-value = 0.01102

var(west$conscientiousness_score)

## [1] 0.01157589

var(latin_america$conscientiousness_score)

## [1] 0.01216558

var(africa$conscientiousness_score)

## [1] 0.01200678

var(arabic_world$conscientiousness_score)

## [1] 0.01235618

var(central_asia$conscientiousness_score)

## [1] 0.01124714

```

U nastavku prepostavljam homegenost varijanci.

###Testiranje jednakosti savjesnosti

Provjeravamo postoje li razlike u savjesnosti.

$$H_0 : \mu_1^2 = \mu_2^2 = \mu_3^2 = \mu_4^2 = \mu_5^2$$

$$H_1 : \neg H_0.$$

```

# Test
a = aov(df$score ~ df$region)
summary(a)

##           Df Sum Sq Mean Sq F value    Pr(>F)
## df$region      2   0.15  0.07578   6.396 0.00167 ***
## Residuals  10374 122.91  0.01185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

nrow(west)

## [1] 271137

```

```

nrow(latin_america)

## [1] 3795

nrow(africa)

## [1] 1734

nrow(arabic_world)

## [1] 2474

nrow(central_asia)

## [1] 4108

nrow(oceania)

## [1] 12456

nrow(east_asia)

## [1] 11153

#karakteristike prema spolu ispitanika ##1=male, 2=female ##članak na ovu temu(https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00178/full)
Neuroticism Neuroticism describes the tendency to experience negative emotion and related processes in response to perceived threat and punishment; these include anxiety, depression, anger, self-consciousness, and emotional lability. Women have been found to score higher than men on Neuroticism as measured at the Big Five trait level, as well as on most facets of Neuroticism included in a common measure of the Big Five, the NEO-PI-R (Costa et al., 2001). Additionally, women also score higher than men on related measures not designed specifically to measure the Big Five, such as indices of anxiety (Feingold, 1994) and low self-esteem (Kling et al., 1999). The one facet of Neuroticism in which women do not always exhibit higher scores than men is Anger, or Angry Hostility (Costa et al., 2001)

male_subjects = bigfive[bigfive$sex == 1 ,]
female_subjects = bigfive[bigfive$sex == 2 ,]

cat('Srednja vrijednost neurocizma za pripadnice ženskog spola je ', mean(female_subjects$neuroticism_)

## Srednja vrijednost neurocizma za pripadnice ženskog spola je 0.5944653

cat('Srednja vrijednost otvorenosti za pripadnike muškog spola je ', mean(male_subjects$neuroticism_sco

## Srednja vrijednost otvorenosti za pripadnike muškog spola je 0.5439874

```

```

cat('Varijanca vrijednost neurocizma za pripadnice ženskog spola je ', var(female_subjects$neuroticism))

## Varijanca vrijednost neurocizma za pripadnice ženskog spola je 0.01498152

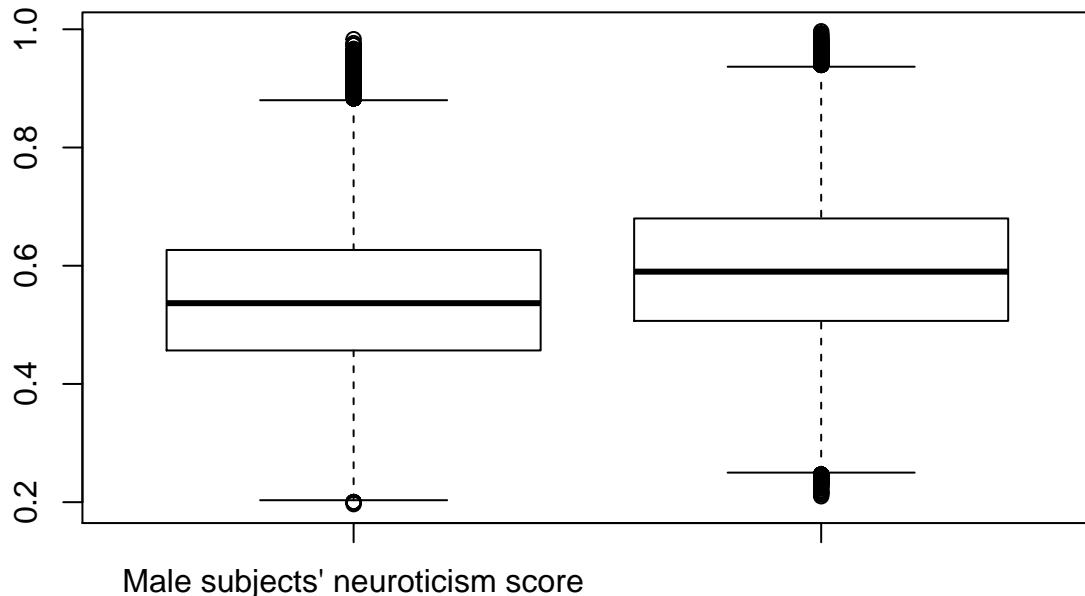
cat('Varijanca vrijednost otvorenosti za pripadnike muškog spola je ', var(male_subjects$neuroticism_score))

## Varijanca vrijednost otvorenosti za pripadnike muškog spola je 0.01553764

boxplot(male_subjects$neuroticism_score, female_subjects$neuroticism_score,
        names = c('Male subjects\' neuroticism score','Female subjects\' neuroticism score'),
        main='Boxplot of male and female subjects\' neuroticism_score')

```

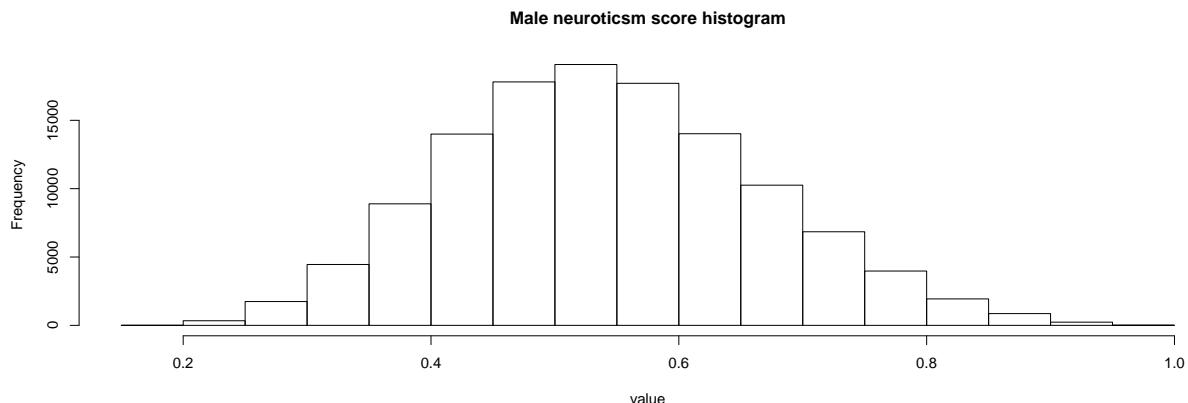
Boxplot of male and female subjects' neuroticism_score



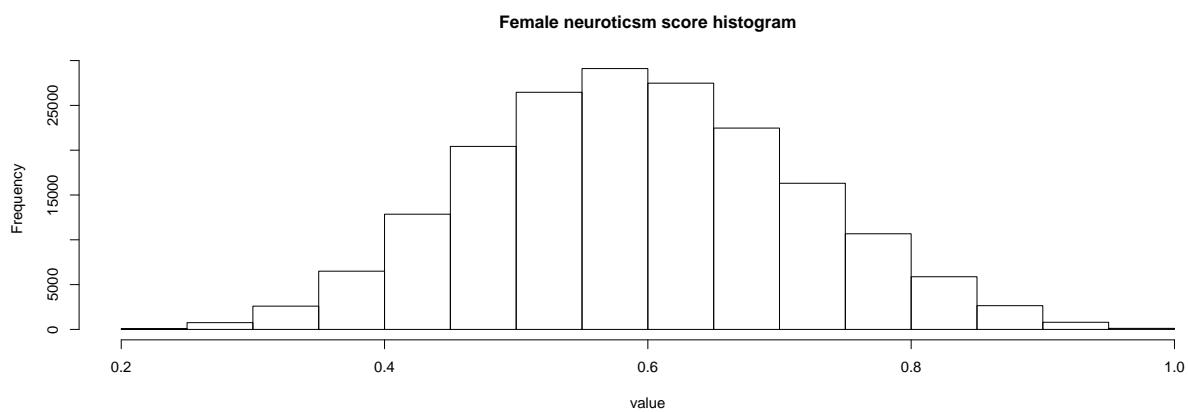
Na ovom skupo podataka možemo uočiti razliku u srednjoj vrijednosti neurocizma između pripadnika ženskog i muškog spola. Stoga postoje indikacije da bi neuroticism_score trebao biti viši kod žena nego muškaraca.

Sljedeći korak je provjeriti normalnost podataka koju najčešće provjeravamo: histogramom, qq-plotom te KS-testom (kojim provjeravamo pripadnost podataka distribuciji).

```
hist(male_subjects$neuroticism_score,main='Male neuroticism score histogram', xlab='value', ylab='Frequency')
```

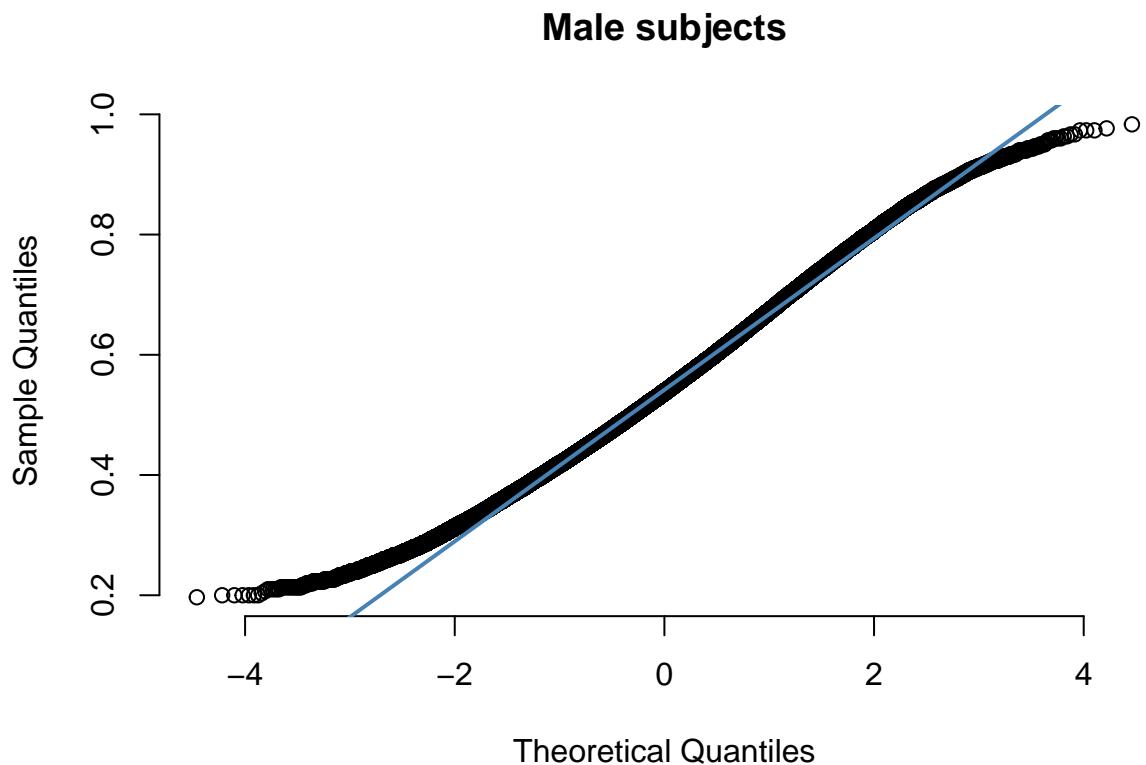


```
hist(female_subjects$neuroticism_score, main='Female neuroticism score histogram', xlab='value', ylab='Frequency')
```

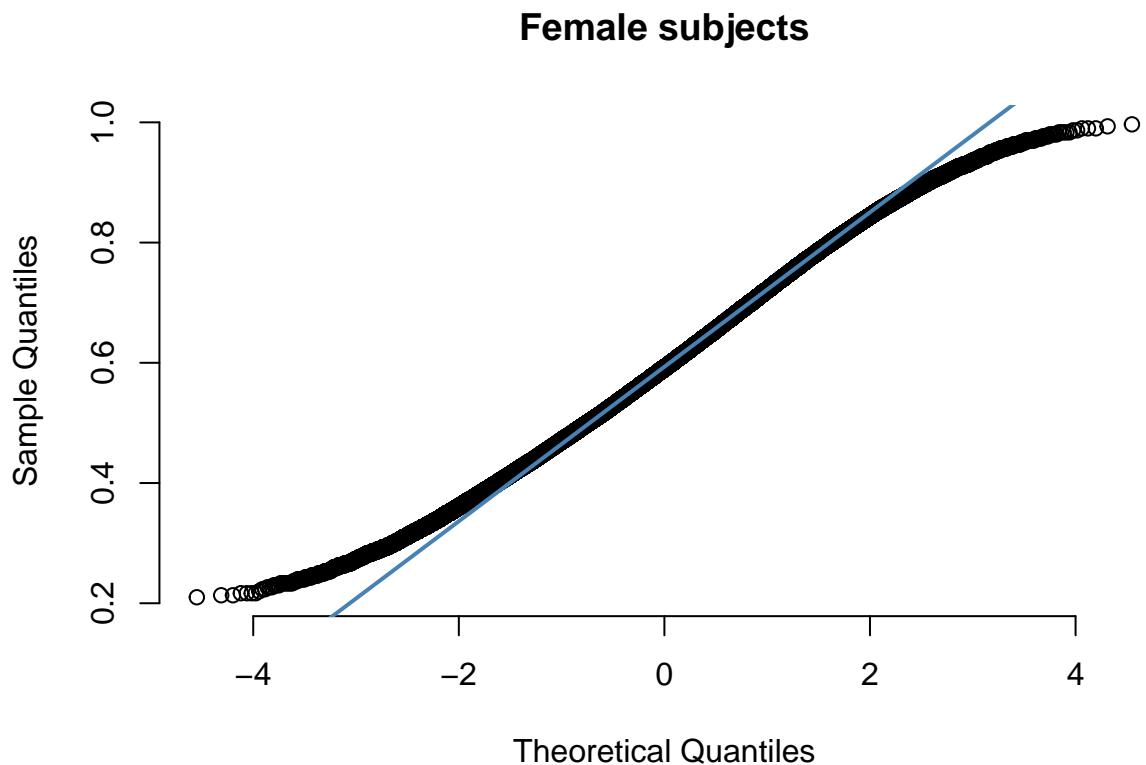


Histogrami upućuju na normalnost podataka. Normalnost možemo još provjeriti i qqplot-ovima ili testom koji ispituje normalnost.

```
qqnorm(male_subjects$neuroticism_score, pch = 1, frame = FALSE, main='Male subjects')
qqline(male_subjects$neuroticism_score, col = "steelblue", lwd = 2)
```



```
qqnorm(female_subjects$neuroticism_score, pch = 1, frame = FALSE, main='Female subjects')
qqline(female_subjects$neuroticism_score, col = "steelblue", lwd = 2)
```



```
var(male_subjects$neuroticism_score)
```

```
## [1] 0.01553764
```

```
var(female_subjects$neuroticism_score)
```

```
## [1] 0.01498152
```

#Testiranje normalnosti distribucije neuroticizma

Hipoteze:

H_0 : distribucija male_subjects\$neuroticism_score pripada normalnoj razdobi H_1 : distribucija male_subjects\$neuroticism_score NE pripada normalnoj razdobi

```
lillie.test(male_subjects$neuroticism_score)
```

```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: male_subjects$neuroticism_score
## D = 0.025863, p-value < 2.2e-16
```

Odbacujem nullu hipotezu u korist alternative da se neuroticizam muških ispitanika ravna po normalnoj distribuciji.

Hipoteze:

H_0 : distribucija female_subjects\$neuroticism_score pripada normalnoj razdiobi H_1 : distribucija female_subjects\$neuroticism_score NE pripada normalnoj razdiobi

```
lillie.test(female_subjects$neuroticism_score)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: female_subjects$neuroticism_score
## D = 0.018533, p-value < 2.2e-16
```

Odbacujem nultu hipotezu u korist alternative da se neuroticizam ženskih ispitanica ravna po normalnoj distibuciji.

#Testiranje jednakosti varijance neuroticizma

Prepostavljam nezavisnost uzorka i izvodim F test.

Hipoteze:

$$H_0 : \sigma_{Male}^2 = \sigma_{Female}^2 H_1 : \sigma_{Male}^2 \neq \sigma_{Female}^2$$

```
var.test(male_subjects$neuroticism_score, female_subjects$neuroticism_score)
```

```
##
## F test to compare two variances
##
## data: male_subjects$neuroticism_score and female_subjects$neuroticism_score
## F = 1.0371, num df = 122163, denom df = 185148, p-value = 2.546e-12
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.026583 1.047776
## sample estimates:
## ratio of variances
## 1.03712
```

Na osnovu male p vrijednosti odbacujem nultu hipotezu u korist alternative da su varijance distribucije različite

#Testiranje jednakosti razine savjesnosti

Koristim T test i prepostavljam nezavisnost varijabli i različitost varijanci.

Hipoteze:

$$H_0 : \mu_{Female} = \mu_{Male} H_1 : \mu_{Female} > \mu_{Male}$$

```
t.test(female_subjects$neuroticism_score ,male_subjects$neuroticism_score, alt = "greater", var.equal =
```

```

## Welch Two Sample t-test
##
## data: female_subjects$neuroticism_score and male_subjects$neuroticism_score
## t = 110.65, df = 258116, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.04972757      Inf
## sample estimates:
## mean of x mean of y
## 0.5944653 0.5439874

```

p-vrijednost je manja od 0.05, stoga se hipoteza $H_0 : \mu_{Female} = \mu_{Male}$ odbacuje u korist alternativne hipoteze $H_1 : \mu_{Female} > \mu_{Male}$

##Motivacija: Možemo li zaključiti nešto o jednoj osobini uzimajući kao ulazne varijable jednu ili više ostalih osobina?

Ono što pritom očekujemo je da ne bi smjela postojati visoka korelacija između pojedinih osobina ličnosti pošto svaka od njih na jedinstven način opisuje određeni element ljudske osobnosti. Uz ovih 5, poznat je i velik broj drugih osobina, ali ovih 5 je izabранo na način da u idealnom slučaju u potpunosti definiraju osobnost neke osobe i razlikuju je od osobnosti drugih pojedinaca.

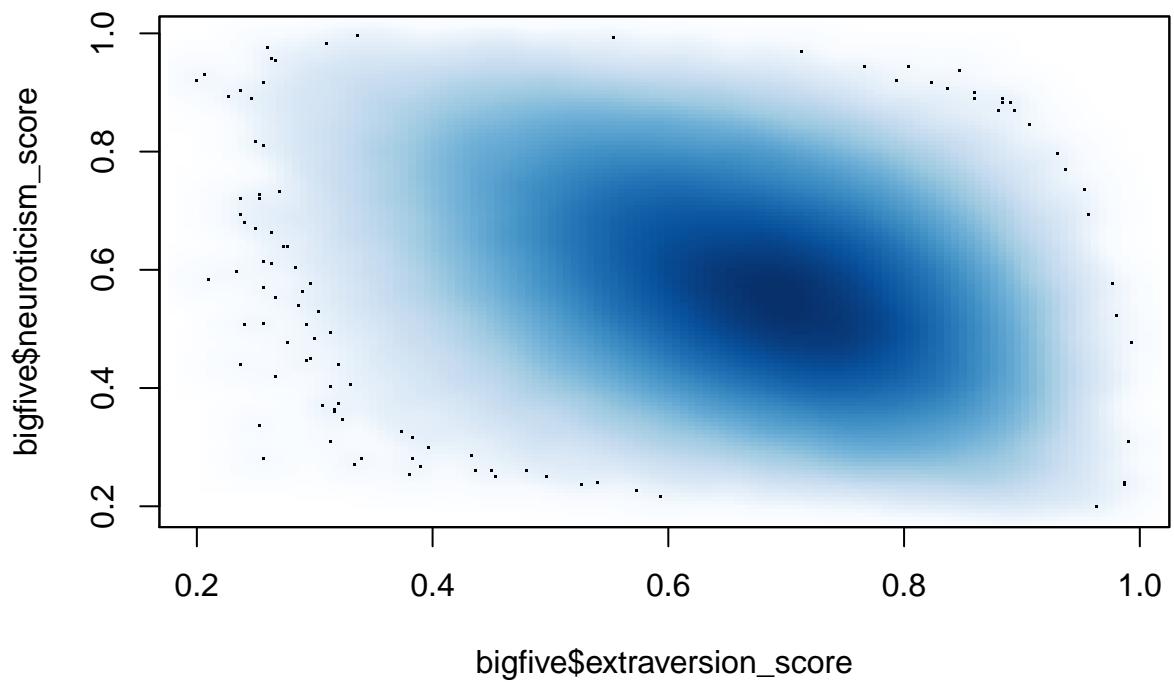
Kao izlaznu varijablu proizvoljno smo odabrali neuroticizam, a pregledavajući znanstvenu literaturu došli smo do zaključka da bi ekstraverzija mogla u najvećoj mjeri objasniti varijaciju u neuroticizmu što i ima smisla jer osobe koje su introverti teže artikuliraju svoje emocije te su skloniji anksioznosti i depresiji.

Kako bismo mogli bolje uočiti potencijalnu vezu dviju varijabli, uzet ćemo poduzorak našeg ulaznog skupa.

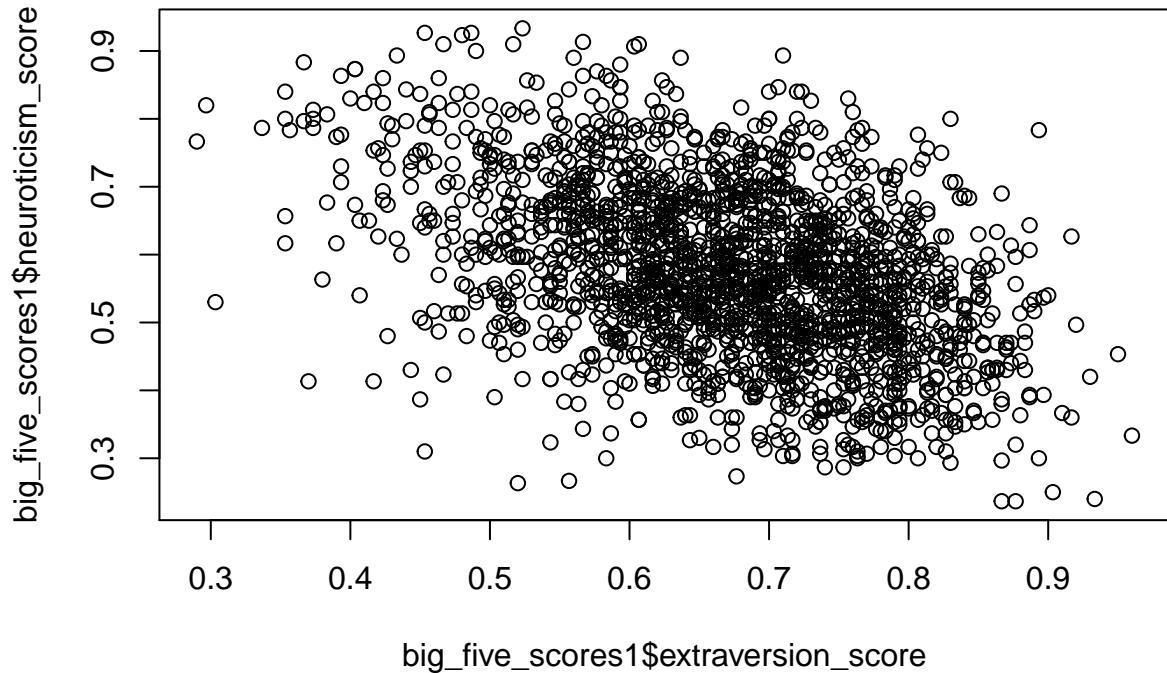
```

big_five_scores1 = bigfive[sample(nrow(bigfive), 2000),]
smoothScatter(bigfive$extraversion_score, bigfive$neuroticism_score)

```



```
plot(big_five_scores1$extraversion_score, big_five_scores1$neuroticism_score)
```

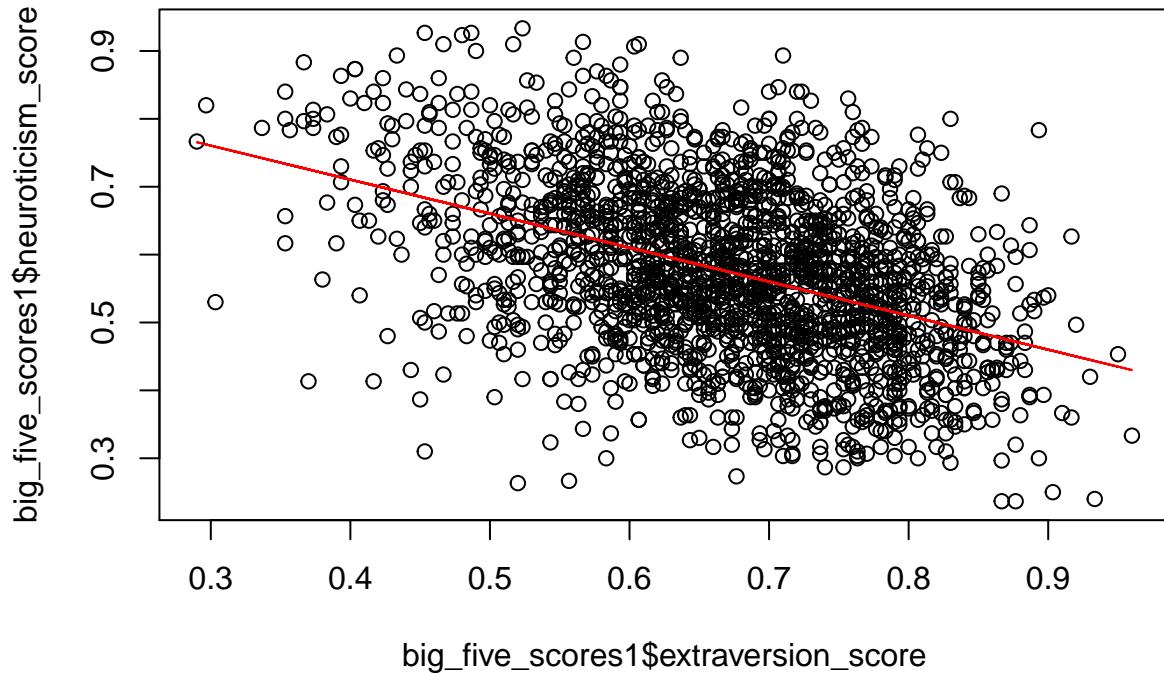


Scatter plot pokazuje da postoji negativan efekt varijable extraversion_score na izlaznu varijablu neuroticism_score. Zato ćemo pokušati kreirati model jednostavne regresije koji uključuje spomenute varijable:

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

Pritom pretostavljamo da je epsilon normalno distribuirana slučajna varijabla (pogreška) s homogenom i nezavisnom varijancom.

```
fit.neuroticism = lm(neuroticism_score ~ extraversion_score, data = big_five_scores1)
plot(big_five_scores1$extraversion_score, big_five_scores1$neuroticism_score)
lines(big_five_scores1$extraversion_score, fit.neuroticism$fitted.values, col = "red")
```



##Provjera prepostavki

Sada ćemo provjeriti prepostavke o normalnosti reziduala i homogenosti varijance. To ćemo učiniti grafički pomoću histograma i q-q plota te statističkim testovima - Kolmogorov-Smirnovljev test i Lillieforsova korekcija.

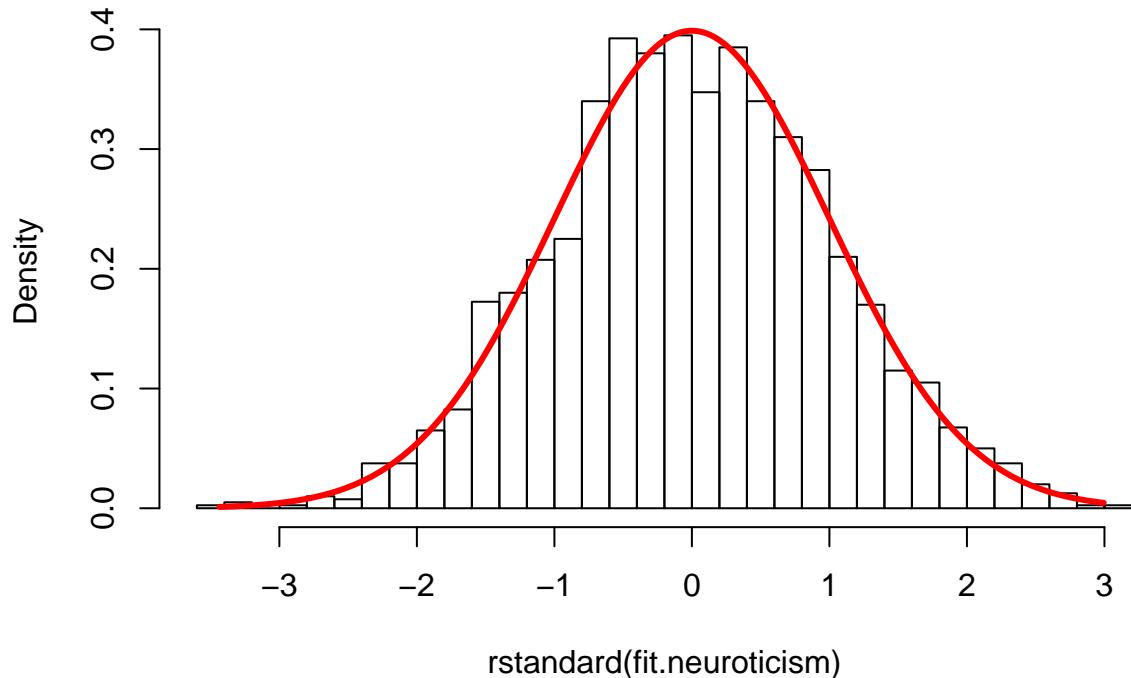
#grafički prikazi

```
hist(rstandard(fit.neuroticism), probability = TRUE, breaks = 30)
```

```
y <- seq(min(rstandard(fit.neuroticism)), max(rstandard(fit.neuroticism)), 0.01)
```

```
lines(y, dnorm(y, mean(rstandard(fit.neuroticism)), sd(rstandard(fit.neuroticism))), col = "red", lwd =
```

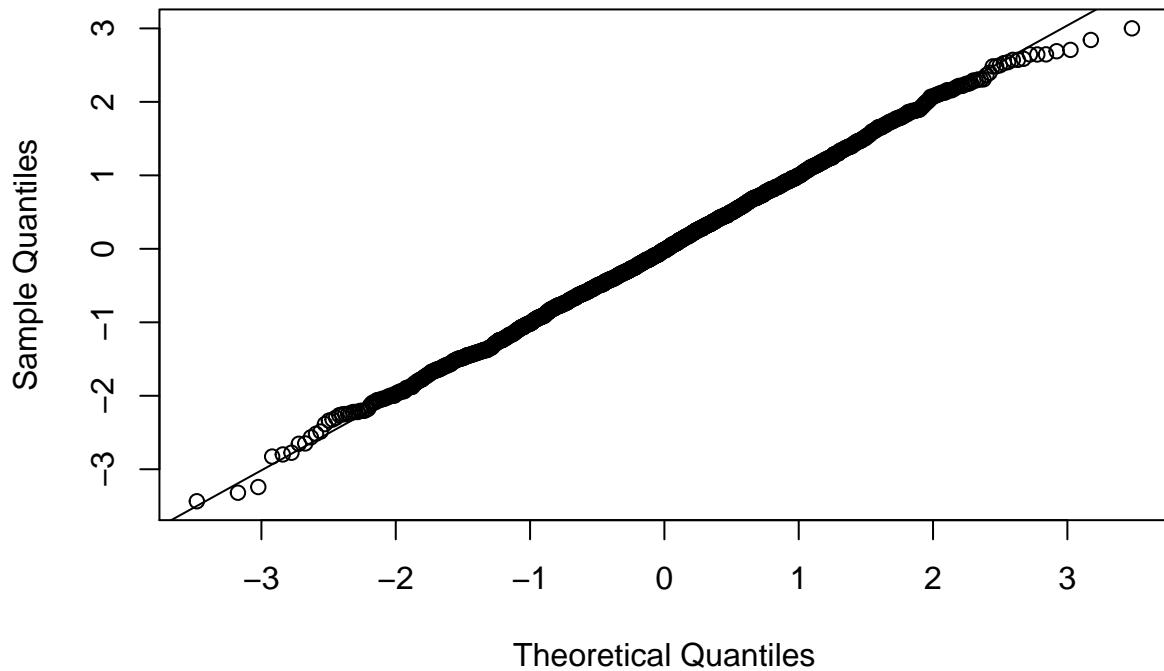
Histogram of rstandard(fit.neuroticism)



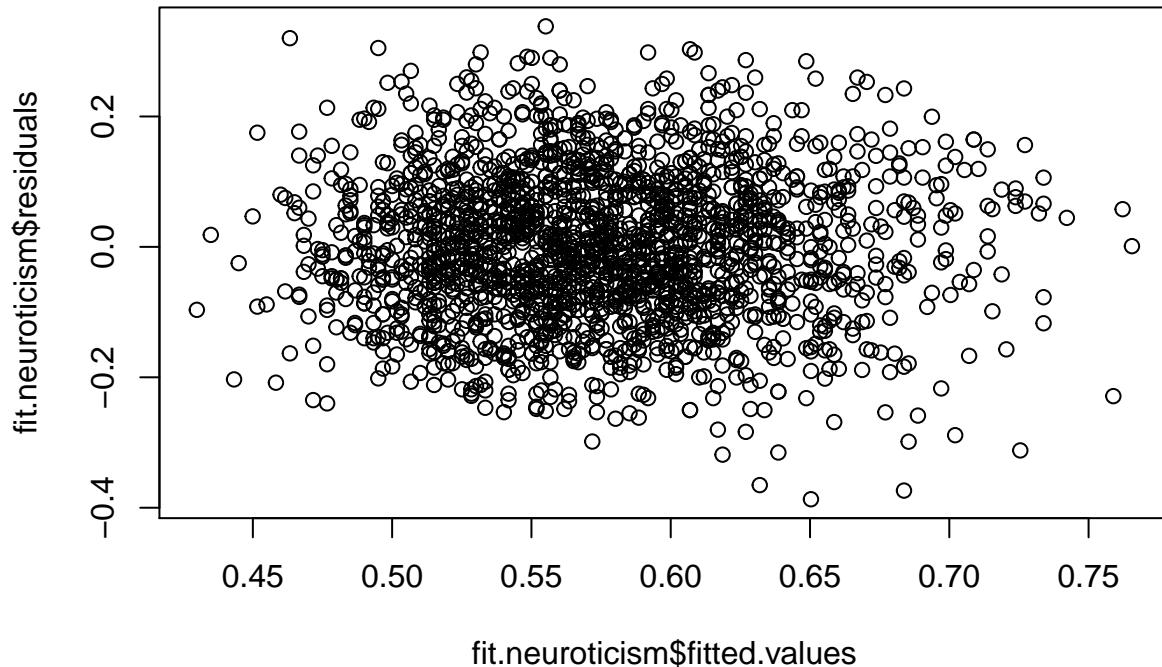
```
qqnorm(rstandard(fit.neuroticism))
```

```
qqline(rstandard(fit.neuroticism)) # q-q plot pokazuje da se kvantili uzoracke distribucije u velikoj m
```

Normal Q-Q Plot



```
plot(fit.neuroticism$fitted.values,fit.neuroticism$residuals) #nema naznake nehomogenosti pogreške
```



```
#normalnost reziduala - statistički testovi

ks.test(rstandard(fit.neuroticism), "pnorm")

## Warning in ks.test(rstandard(fit.neuroticism), "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test

## 
## One-sample Kolmogorov-Smirnov test
##
## data: rstandard(fit.neuroticism)
## D = 0.014378, p-value = 0.8027
## alternative hypothesis: two-sided

require(nortest)
lillie.test(rstandard(fit.neuroticism))

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fit.neuroticism)
## D = 0.014352, p-value = 0.407
```

```
shapiro.test(rstandard(fit.neuroticism))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: rstandard(fit.neuroticism)  
## W = 0.99907, p-value = 0.4051
```

Iz p vrijednost KS-testa možemo zaključiti da uz nivo značajnosti od 5% ne možemo odbaciti nullu hipotezu o normalnosti, dok iz Lillieforsove inačice možemo zaključiti da uz isti nivo značajnosti možemo odbaciti nullu hipotezu. U ovom slučaju, pošto testiramo populaciju nepoznatog očekivanja i varijance, veću važnost pridajemo drugom testu koji je ujedno i stroži od KS testa. Pošto su t-testovi koji se koriste u analizi regresijskog modela robusni na nenormalnost, a i pošto p vrijednost koju testom dobijemo nije daleko od razine značajnosti možemo nastaviti dalje s regresijskom analizom.

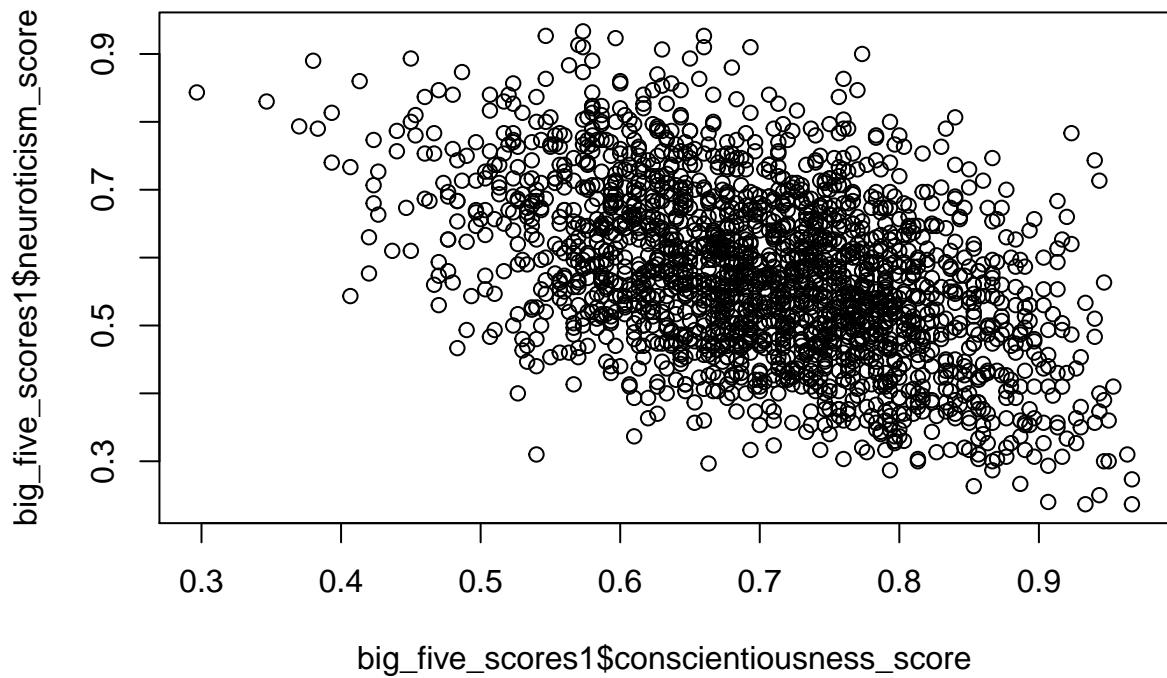
```
summary(fit.neuroticism)
```

```
##  
## Call:  
## lm(formula = neuroticism_score ~ extraversion_score, data = big_five_scores1)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -0.38700 -0.07527 -0.00183  0.07814  0.33819  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  0.91085   0.01591  57.24  <2e-16 ***  
## extraversion_score -0.50100   0.02341 -21.40  <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.1127 on 1998 degrees of freedom  
## Multiple R-squared:  0.1865, Adjusted R-squared:  0.1861  
## F-statistic:  458 on 1 and 1998 DF,  p-value: < 2.2e-16
```

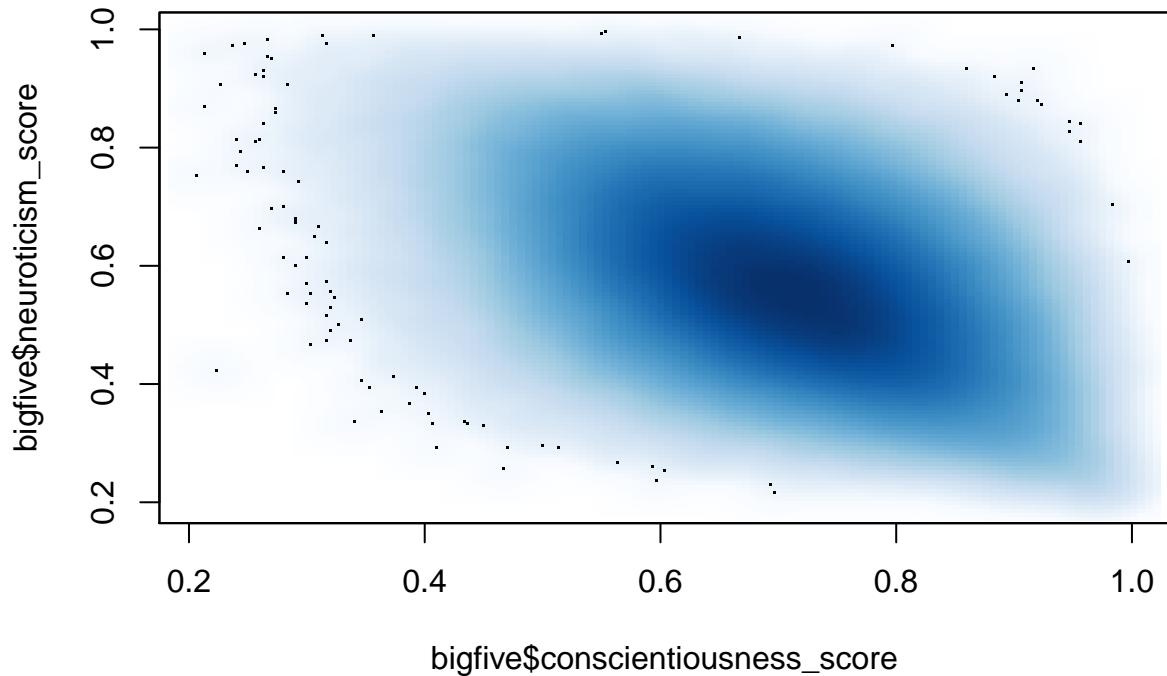
```
# t-test --> koeficijenti su signifikantni  
# f-test --> Cijeli model je signifikantan  
# zaključak o R2?
```

Osim ekstraverzije kao potencijalne osobine iz koje bismo mogli nešto zaključiti o neuroticizmu osobe, pokazalo se da bi još jedna osobina u Big Five modelu mogla objašnjavati rezultat koji osoba postiže na neuroticizmu, a to je savjesnost. To možemo interpretirati time da osobe koje su nediscplinirane, neuredne i nemaju rutinu mogu biti sklonije neurotičnom ponašanju od savjesnih i organiziranih ljudi.

```
plot(big_five_scores1$conscientiousness_score, big_five_scores1$neuroticism_score)
```



```
smoothScatter(bigfive$conscientiousness_score, bigfive$neuroticism_score)
```



Iz scatter plota možemo uočiti naznaku negativne korelacije, odnosno zavisnosti dviju varijabli što odgovara našoj početnoj tezi. Kako bismo kvantificirali navedenu korelaciju, možemo izračunati Pearsonov korelacijski koeficijent za ove dvije varijable, a dodatno možemo pogledati korelacijsku tablicu svih 5 faktora što će nam pomoći i u kreiranju modela višestruke regresije i provjere njenih prepostavki.

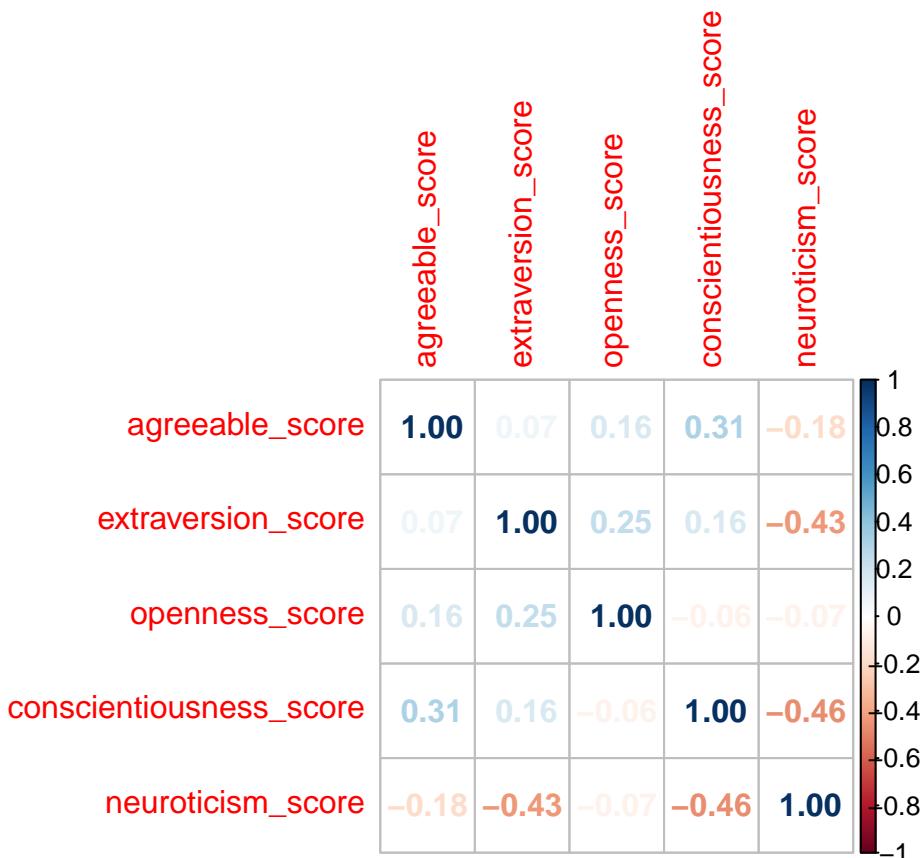
```
correlations <- cor(cbind(big_five_scores1$agreeable_score,big_five_scores1$extraversion_score, big_five_scores1$openness_score, big_five_scores1$conscientiousness_score))

colnames(correlations) <- c("agreeable_score", "extraversion_score", "openness_score", "conscientiousness_score")
rownames(correlations) <- c("agreeable_score", "extraversion_score", "openness_score", "conscientiousness_score")

library(corrplot)

## corrplot 0.92 loaded

corrplot(correlations, method = "number")
```

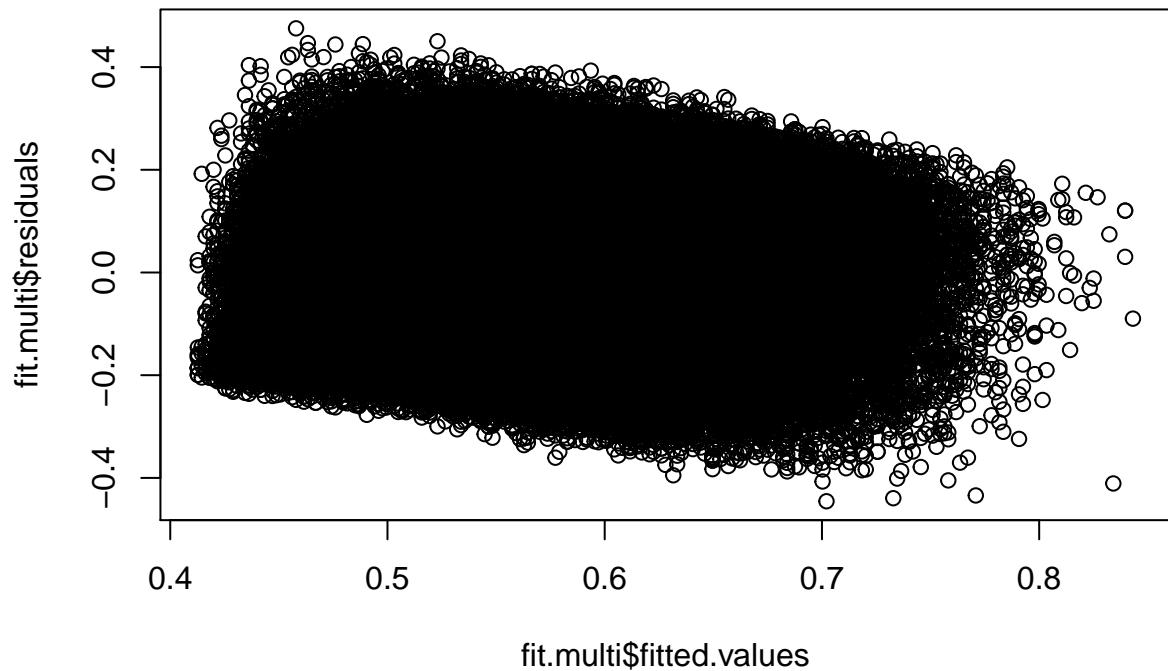


```
cor.test(big_five_scores1$extraversion_score, big_five_scores1$neuroticism_score)
```

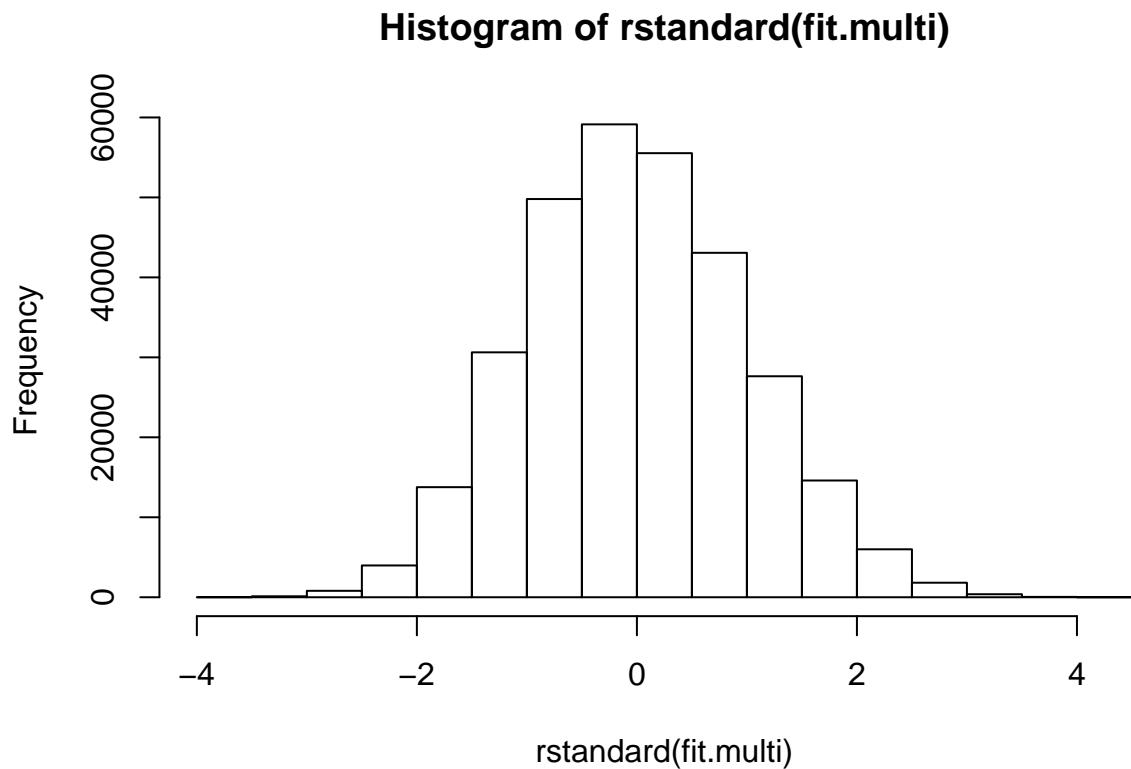
```
##
## Pearson's product-moment correlation
##
## data: big_five_scores1$extraversion_score and big_five_scores1$neuroticism_score
## t = -21.401, df = 1998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.4668393 -0.3955000
## sample estimates:
##       cor
## -0.4318448
```

Iz korelacijske tablice naslućujemo da postoji određena veza između varijabli neuroticism_score i conscientiousness_score. Ono što također možemo uočiti je da nema kršenja prepostavki, tj. da nema značajne korelacije između dviju varijabli extraversion_score i conscientiousness_score što nam ujedno pokazuje i t-test.

```
fit.multi = lm(neuroticism_score ~ conscientiousness_score, data=bigfive)
plot(fit.multi$fitted.values, fit.multi$residuals)
```



```
hist(rstandard(fit.multi))
```



```
ks.test(rstandard(fit.multi), "pnorm")
```

```
## Warning in ks.test(rstandard(fit.multi), "pnorm"): ties should not be present
## for the Kolmogorov-Smirnov test
```

```
##
## One-sample Kolmogorov-Smirnov test
##
## data: rstandard(fit.multi)
## D = 0.01722, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
lillie.test(rstandard(fit.multi))
```

```
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fit.multi)
## D = 0.017219, p-value < 2.2e-16
```

Iz grafičkog prikaza i testova zaključujemo da nema kršenja prepostavke normalnosti reziduala.

```

fit.neuroticism_m = lm(neuroticism_score ~ extraversion_score + conscientiousness_score, bigfive)
summary(fit.neuroticism_m)

##
## Call:
## lm(formula = neuroticism_score ~ extraversion_score + conscientiousness_score,
##      data = bigfive)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -0.50744 -0.07071 -0.00254  0.06919  0.42177 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.187790  0.001553  764.8 <2e-16 ***
## extraversion_score -0.414690  0.001717 -241.5 <2e-16 ***
## conscientiousness_score -0.476616  0.001738 -274.2 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 0.1022 on 307310 degrees of freedom
## Multiple R-squared:  0.3399, Adjusted R-squared:  0.3399 
## F-statistic: 7.914e+04 on 2 and 307310 DF, p-value: < 2.2e-16

```

Iz modela multivarijantne regresije možemo uočiti da su svi koeficijenti signifikantni uz bilo koju razinu značajnosti. Također, koeficijent determinacije se povećao u odnosu na model jednostavne regresije što znači da je conscientiousness_score doprinio poboljšanju funkcije ovisnosti izlazne varijable o regresorima, odnosno da je veća proporcija varijacije varijable Y objašnjena ovim modelom u odnosu na prošli.

Možemo probati uključiti i preostala dva faktora u model.

```

fit.multi = lm(neuroticism_score ~ extraversion_score + conscientiousness_score + agreeable_score + openness_score, bigfive[bigfive$age > 20, ])
summary(fit.multi)

##
## Call:
## lm(formula = neuroticism_score ~ extraversion_score + conscientiousness_score +
##      agreeable_score + openness_score, data = bigfive[bigfive$age >
##      20, ])
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -0.49825 -0.07158 -0.00331  0.06911  0.43462 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.224961  0.003030  404.27 <2e-16 ***
## extraversion_score -0.444337  0.002438 -182.24 <2e-16 ***
## conscientiousness_score -0.491877  0.002551 -192.78 <2e-16 ***
## agreeable_score -0.056987  0.002831 -20.13 <2e-16 ***
## openness_score  0.041358  0.002964  13.96 <2e-16 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

```

```
##  
## Residual standard error: 0.1027 on 177131 degrees of freedom  
## Multiple R-squared:  0.3698, Adjusted R-squared:  0.3698  
## F-statistic: 2.599e+04 on 4 and 177131 DF,  p-value: < 2.2e-16
```

Vidimo da su opet svi koeficijenti signifikantni do na bilo koju razinu značajnosti. Povećanje koeficijenta determinacije postoji, ali je zanemarivo pa možemo zaključiti da dodani faktori ne objašnjavaju u značajnijoj mjeri neobjašnjenu varijaciju modela s postojeća dva faktora. Najveći efekt na izlaznu varijablu imaju regresori extraversion_score i conscientiousness_score.

Zaključak