MBTI dataset transformation and analysis

Vedran Moškov, Lucija Runjić, Borna Josipović, Lana Bartolović

2024-01-21

Motivacija i opis problema

MBTI je test ličnosti koji kategorizira pojedince prema četiri dimenzije (Ekstroverzija/Introverzija, Senzornost/Intuicija, Razmišljanje/Osjećanje, Prosudba/Spontanost), dodjeljujući im jedan od 16 osobnosnih tipova. Istražiti ćemo povezanost između osobnosti, dobivene kroz MBTI test i fizičkih karakteristika poput držanja tijela, težine i visine.

Učitavanje i uređivanje podatkovnog skupa

Učitavanje i proučavanje podatkovnog skupa

Učitavamo podatkovni skup u varijablu "dataset".

```
dataset <- read_csv("../data/MBTI.csv")
```

Proučavamo podatkovni skup kako bi ga znali urediti na način da nam je lakše raditi s njim kasnije.

```
head(dataset)
```

```
## # A tibble: 6 x 21
##    ...1 ‘S No‘   AGE HEIGHT WEIGHT SEX    ‘ACTIVITY LEVEL‘ ‘PAIN 1‘ ‘PAIN 2‘
##   <dbl>  <dbl> <dbl>  <dbl>  <dbl> <chr>  <chr>               <dbl>    <dbl>
## 1     0      1    53     62    125 Female Low                   0        0
## 2     1      2    52     69    157 Male   High                  7        8
## 3     2      3    30     69    200 Male   High                  0        0
## 4     3      4    51     66    175 Male   Moderate              9.5      9.5
## 5     4      5    45     63    199 Female Moderate              4        5
## 6     5      6    68     74    182 Male   Low                   0        2.5
## # i 12 more variables: ‘PAIN 3‘ <dbl>, ‘PAIN 4‘ <dbl>, MBTI <chr>, E <dbl>,
## #   I <dbl>, S <dbl>, N <dbl>, T <dbl>, F <dbl>, J <dbl>, P <dbl>,
## #   POSTURE <chr>
```

```r
glimpse(dataset)
```

```
## Rows: 97
## Columns: 21
## $ ...1             <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
## $ `S No`           <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16~
## $ AGE              <dbl> 53, 52, 30, 51, 45, 68, 62, 65, 66, 58, 61, 33, 48, 5~
## $ HEIGHT           <dbl> 62, 69, 69, 66, 63, 74, 68, 61, 67, 69, 67, 62, 64, 6~
## $ WEIGHT           <dbl> 125, 157, 200, 175, 199, 182, 263, 143, 180, 165, 210~
## $ SEX              <chr> "Female", "Male", "Male", "Male", "Female", "Male", "~
## $ `ACTIVITY LEVEL` <chr> "Low", "High", "High", "Moderate", "Moderate", "Low",~
## $ `PAIN 1`         <dbl> 0.0, 7.0, 0.0, 9.5, 4.0, 0.0, 7.0, 0.0, 0.5, 0.0, 5.0~
## $ `PAIN 2`         <dbl> 0.0, 8.0, 0.0, 9.5, 5.0, 2.5, 10.0, 9.0, 3.5, 7.5, 0.~
## $ `PAIN 3`         <dbl> 0.0, 5.0, 0.0, 9.5, 2.0, 1.5, 10.0, 5.0, 0.5, 7.0, 0.~
## $ `PAIN 4`         <dbl> 0.0, 3.0, 0.0, 1.5, 2.0, 0.0, 10.0, 10.0, 9.5, 3.0, 9~
## $ MBTI             <chr> "ESFJ", "ISTJ", "ESTJ", "ISTJ", "ENFJ", "ISFP", "ISTP~
## $ E                <dbl> 0.9084579, -0.6045853, 0.4727891, -0.6045853, 0.34875~
## $ I                <dbl> -1.0968036, 0.4727891, -0.6045853, 0.4727891, -0.4727~
## $ S                <dbl> -0.06968492, -0.28221615, -0.13971030, 0.21042839, 0.~
## $ N                <dbl> -0.6744898, -0.4307273, -0.5894558, -1.0853249, -0.96~
## $ T                <dbl> -0.3186394, 1.1503494, 0.3186394, 0.1046335, -0.31863~
## $ F                <dbl> 0.1046335, -1.1503494, -0.3186394, -0.1046335, 0.3186~
## $ J                <dbl> 0.78103381, 0.16421078, 0.05451891, 0.93881432, 0.511~
## $ P                <dbl> -0.93881432, -0.27592106, -0.16421078, -1.12433823, -~
## $ POSTURE          <chr> "A", "B", "A", "D", "A", "D", "B", "D", "C", "D", "B"~
```

Uređivanje podataka podatkovnog skupa

Faktoriziramo i modificiramo stupce "SEX", "ACTIVITY LEVEL", "MBTI", "POSTURE" kako bismo kasnije mogli lakše grupirati podatke i bolje ih analizirati.

```
dataset$SEX <- as.factor(dataset$SEX)
dataset$`ACTIVITY LEVEL` <- as.factor(dataset$`ACTIVITY LEVEL`)
dataset$`ACTIVITY LEVEL` <- factor(
  dataset$`ACTIVITY LEVEL`, levels = c("Low", "Moderate", "High")
  )
dataset$MBTI <- as.factor(dataset$MBTI)
dataset$POSTURE <- as.factor(dataset$POSTURE)
dataset$POSTURE <- factor(dataset$POSTURE, levels = c("A", "B", "C", "D"),
                          labels = c("idealno", "kifoza/lordoza", "ravna leđa", "nagnuto"))
```

Uklonit ćemo prva dva stupca podatkovnog skupa obzirom da su jedinstveni identifikatori stoga nam ne pomažu u analizi.

```
dataset$...1 <- NULL
dataset$`S No` <- NULL
```

Preimenovat ćemo stupce "ACTIVITY LEVEL", "PAIN 1", "PAIN 2", "PAIN 3" i "PAIN 4" radi jednostavnosti.

```
colnames(dataset)[5] <- "ACTIVITY_LEVEL"
colnames(dataset)[6] <- "PAIN_1"
colnames(dataset)[7] <- "PAIN_2"
colnames(dataset)[8] <- "PAIN_3"
colnames(dataset)[9] <- "PAIN_4"
```

Pretvorit ćemo podatke u stupcima "HEIGHT" i "WEIGHT" u centimentre i kilograme.

```
dataset$HEIGHT <- round(dataset$HEIGHT * 2.54, 1)
dataset$WEIGHT <- round(dataset$WEIGHT * 0.45359237, 1)
```

Dodat ćemo neke nove stupce pomoću kojih ćemo grupirati podatke u manje grupe kako bismo ih mogli bolje analizirati. Dodajemo stupac "IS_ACTIVE" (na temelju stupca "ACTIVE" grupira osobe u one aktivne i neaktivne) te stupce "IE", "SN", "TF", "JP" podatke uzimamo iz rezultata MBTI testa iz stupca "MBTI").

```
dataset$GROUP <- as.factor(color(dataset$MBTI))
dataset$IS_ACTIVE <- as.factor(
  ifelse(dataset$ACTIVITY_LEVEL == "Low", "Inactive", "Active")
  )
dataset$IE <- as.factor(substring(dataset$MBTI, 1, 1))
dataset$SN <- as.factor(substr(dataset$MBTI, 2, 2))
dataset$TF <- as.factor(substr(dataset$MBTI, 3, 3))
dataset$JP <- as.factor(substr(dataset$MBTI, 4, 4))
```

Ovako naš podatkovni skup izgleda nakon uređivanja njegovih podataka.

```r
head(dataset)
```

```
## # A tibble: 6 x 25
##     AGE HEIGHT WEIGHT SEX    ACTIVITY_LEVEL PAIN_1 PAIN_2 PAIN_3 PAIN_4 MBTI
##   <dbl>  <dbl>  <dbl> <fct>  <fct>           <dbl>  <dbl>  <dbl>  <dbl> <fct>
## 1    53   158.   56.7 Female Low               0      0      0      0   ESFJ
## 2    52   175.   71.2 Male   High              7      8      5      3   ISTJ
## 3    30   175.   90.7 Male   High              0      0      0      0   ESTJ
## 4    51   168.   79.4 Male   Moderate          9.5    9.5    9.5    1.5 ISTJ
## 5    45   160    90.3 Female Moderate          4      5      2      2   ENFJ
## 6    68   188    82.6 Male   Low               0      2.5    1.5    0   ISFP
## # i 15 more variables: E <dbl>, I <dbl>, S <dbl>, N <dbl>, T <dbl>, F <dbl>,
## #   J <dbl>, P <dbl>, POSTURE <fct>, GROUP <fct>, IS_ACTIVE <fct>, IE <fct>,
## #   SN <fct>, TF <fct>, JP <fct>
```

```r
glimpse(dataset)
```

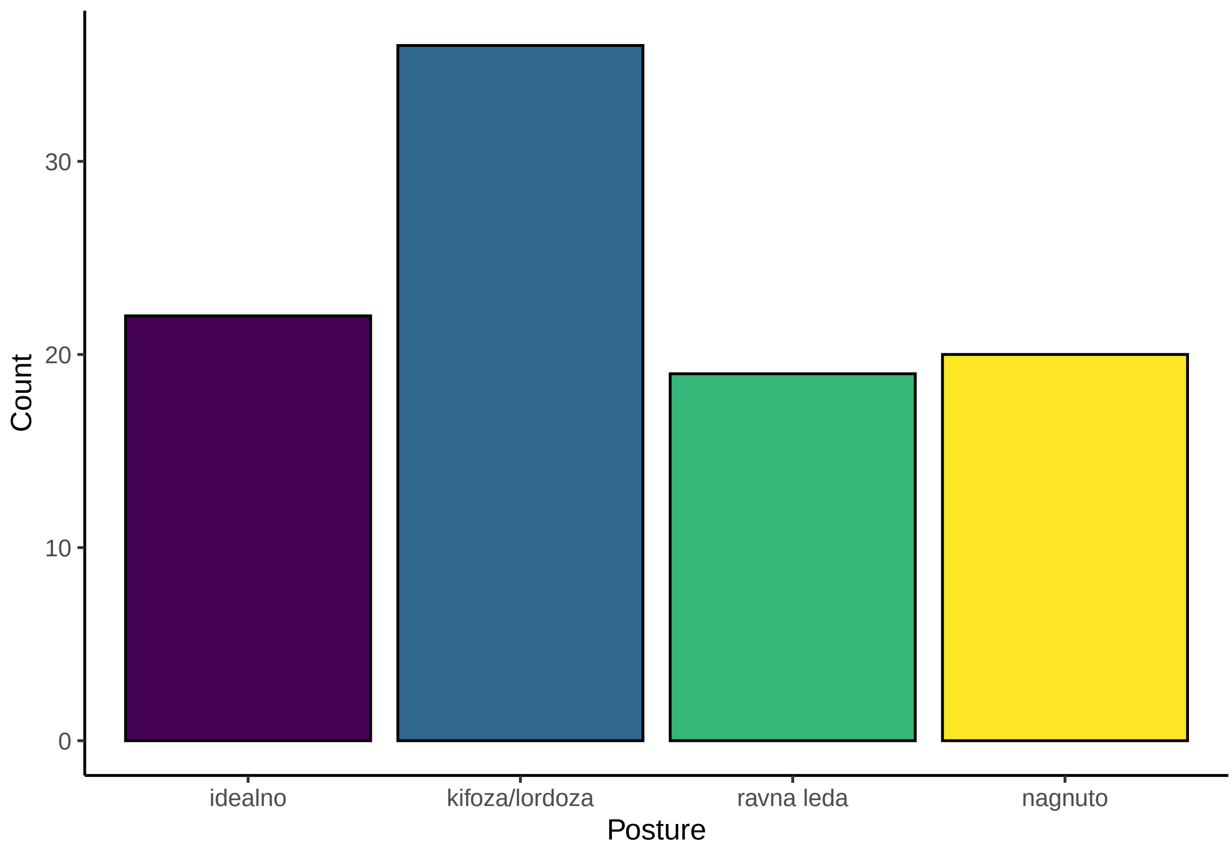
```
## Rows: 97
## Columns: 25
## $ AGE            <dbl> 53, 52, 30, 51, 45, 68, 62, 65, 66, 58, 61, 33, 48, 57,~
## $ HEIGHT         <dbl> 157.5, 175.3, 175.3, 167.6, 160.0, 188.0, 172.7, 154.9,~
## $ WEIGHT         <dbl> 56.7, 71.2, 90.7, 79.4, 90.3, 82.6, 119.3, 64.9, 81.6, ~
## $ SEX            <fct> Female, Male, Male, Male, Female, Male, Male, Female, M~
## $ ACTIVITY_LEVEL <fct> Low, High, High, Moderate, Moderate, Low, Low, Low, Low~
## $ PAIN_1         <dbl> 0.0, 7.0, 0.0, 9.5, 4.0, 0.0, 7.0, 0.0, 0.5, 0.0, 5.0, ~
## $ PAIN_2         <dbl> 0.0, 8.0, 0.0, 9.5, 5.0, 2.5, 10.0, 9.0, 3.5, 7.5, 0.0,~
## $ PAIN_3         <dbl> 0.0, 5.0, 0.0, 9.5, 2.0, 1.5, 10.0, 5.0, 0.5, 7.0, 0.0,~
## $ PAIN_4         <dbl> 0.0, 3.0, 0.0, 1.5, 2.0, 0.0, 10.0, 10.0, 9.5, 3.0, 9.0~
## $ MBTI           <fct> ESFJ, ISTJ, ESTJ, ISTJ, ENFJ, ISFP, ISTP, ESTJ, ESFJ, I~
## $ E              <dbl> 0.9084579, -0.6045853, 0.4727891, -0.6045853, 0.3487557~
## $ I              <dbl> -1.0968036, 0.4727891, -0.6045853, 0.4727891, -0.472789~
## $ S              <dbl> -0.06968492, -0.28221615, -0.13971030, 0.21042839, 0.13~
## $ N              <dbl> -0.6744898, -0.4307273, -0.5894558, -1.0853249, -0.9674~
## $ T              <dbl> -0.3186394, 1.1503494, 0.3186394, 0.1046335, -0.3186394~
## $ F              <dbl> 0.1046335, -1.1503494, -0.3186394, -0.1046335, 0.318639~
## $ J              <dbl> 0.78103381, 0.16421078, 0.05451891, 0.93881432, 0.51193~
## $ P              <dbl> -0.93881432, -0.27592106, -0.16421078, -1.12433823, -0.~
## $ POSTURE        <fct> idealno, kifoza/lordoza, idealno, nagnuto, idealno, nag~
## $ GROUP          <fct> Sentinels, Sentinels, Sentinels, Sentinels, Diplomats, ~
## $ IS_ACTIVE      <fct> Inactive, Active, Active, Active, Active, Inactive, Ina~
## $ IE             <fct> E, I, E, I, E, I, I, E, E, I, E, I, E, E, E, I, E, E, E~
## $ SN             <fct> S, S, S, S, N, S, S, S, S, N, N, S, S, N, S, S, S, N, S~
## $ TF             <fct> F, T, T, T, F, F, T, T, F, F, T, F, F, T, T, T, T, F, F~
## $ JP             <fct> J, J, J, J, J, P, P, J, J, J, P, J, J, J, P, J, J, P, P~
```

Analiza podatkovnog skupa

Veza između tipa ličnosti i načina držanja

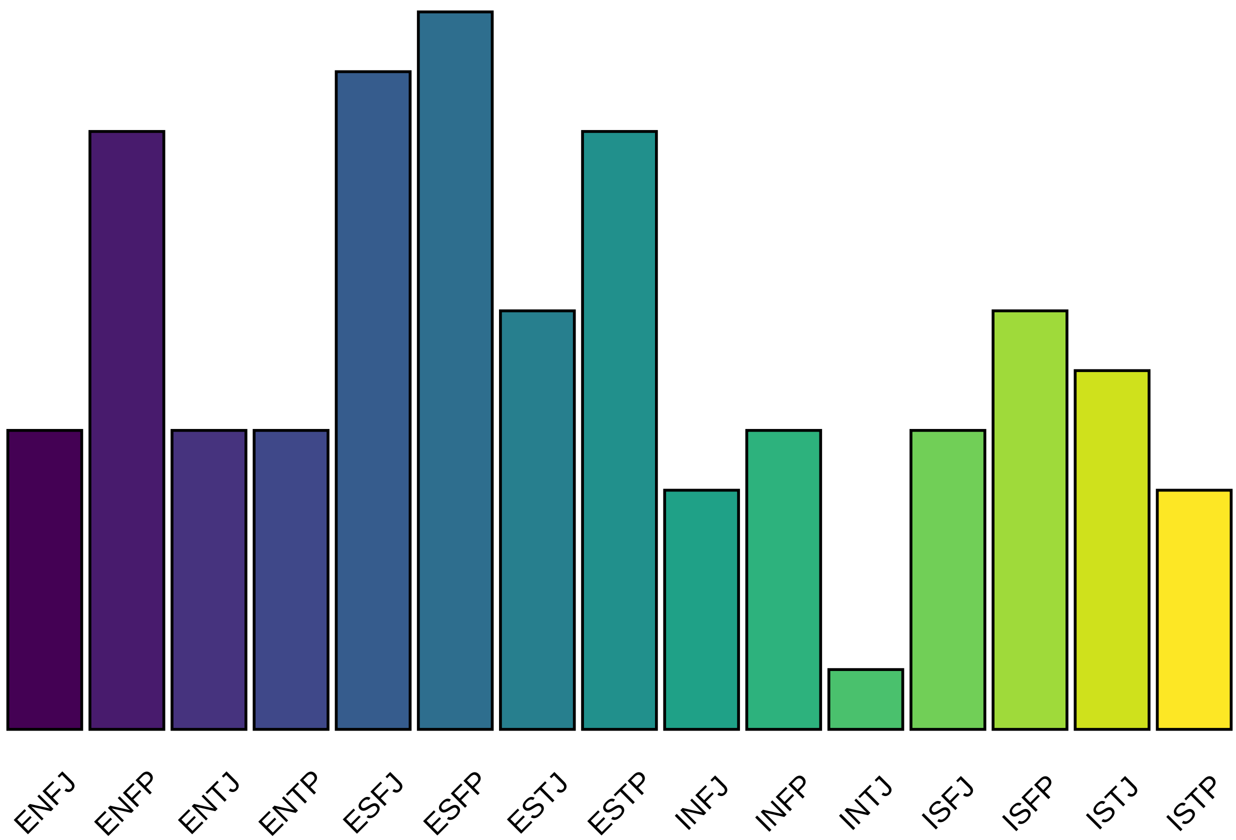
U našem podatkovnom skupu imamo stupac "POSTURE" koji predstavlja kategorije načina držanja i poprima vrijednosti "idealno", "kifoza/lordoza", "ravna leđa" i "nagnuto".

```
ggplot(dataset, aes(x = POSTURE, fill = POSTURE)) +
  geom_bar(color = "black") +
  scale_fill_ordinal() +
  labs(x = "Posture", fill = "Posture", y = "Count") +
  theme_classic() +
  theme(legend.position = "none")
```



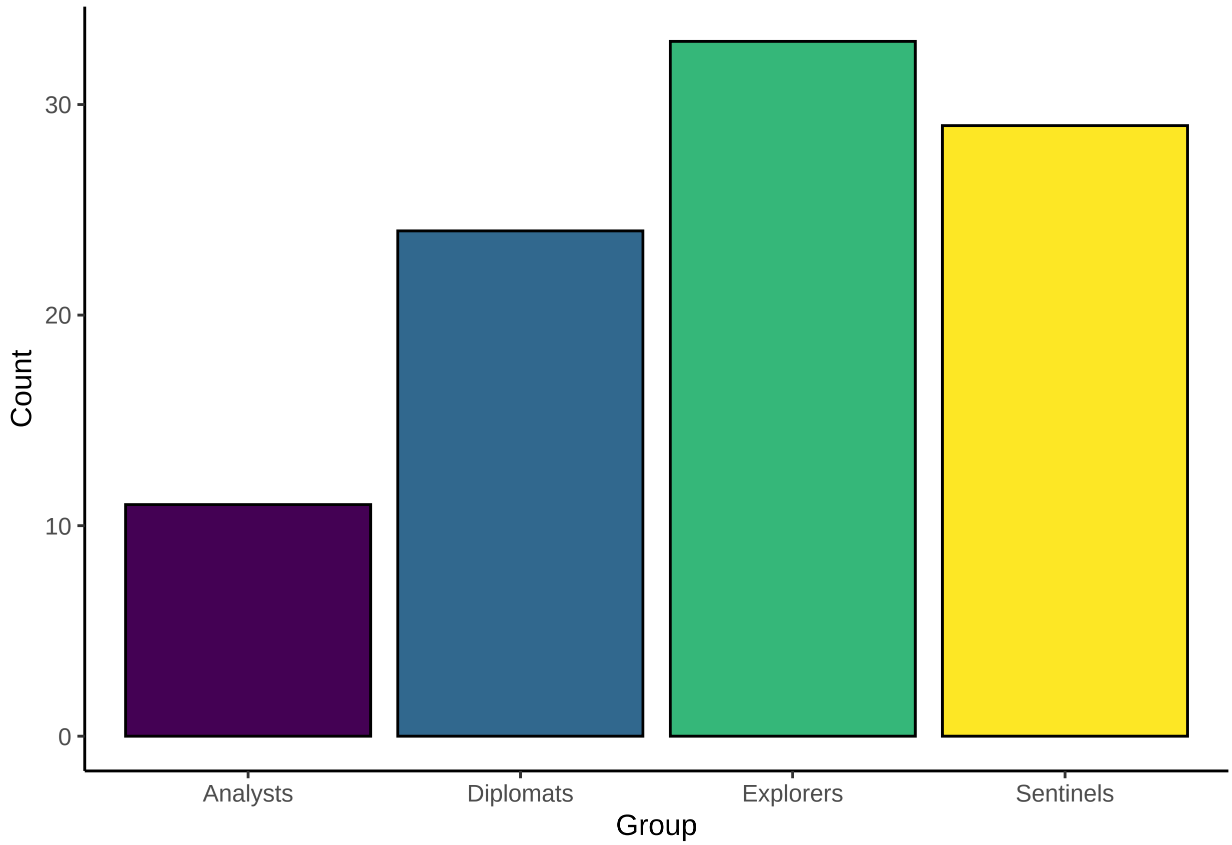
Također imamo stupac "MBTI" koji predstavlja tipove ličnosti i razlikujemo 16 vrsta tipova osobnosti.

```
ggplot(dataset, aes(x = MBTI, fill = MBTI)) +
  geom_bar(color = "black") +
  scale_fill_ordinal() +
  theme_void() +
  theme(axis.text.x = element_text(angle = 45), legend.position = "none") +
  labs(y = "Count")
```



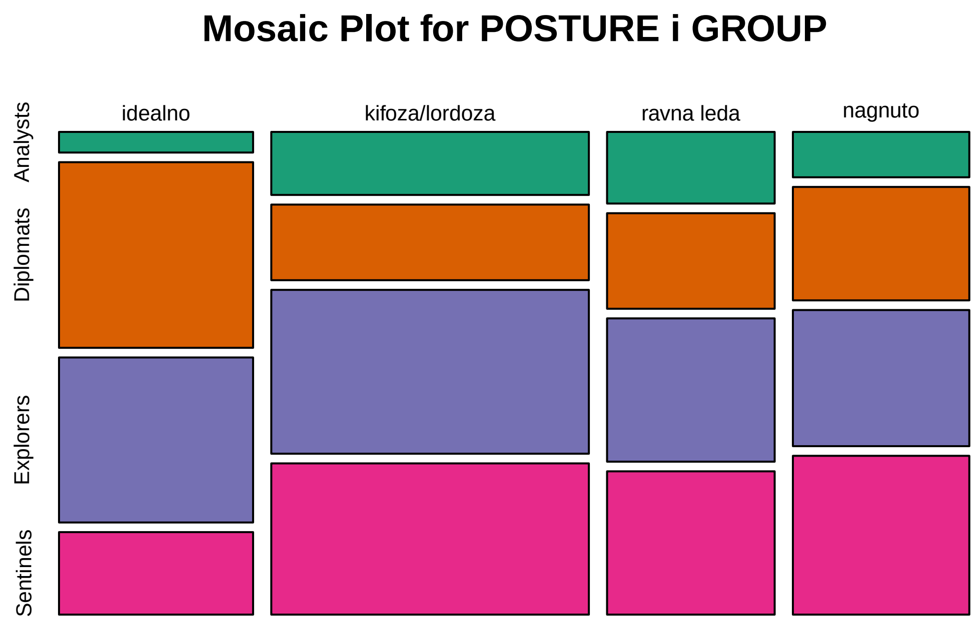
Tih 16 tipova osobnosti radi lakšeg prikaza i analize grupiramo u 4 podskupine u skladu s Myers-Briggsovim modelom (analiziramo samo prvo slovo tipa osobnosti):

```
ggplot(dataset, aes(x = GROUP, fill = GROUP)) +
  geom_bar(color = "black") +
  scale_fill_ordinal() +
  labs(x = "Group", fill = "Group", y = "Count") +
  theme_classic() +
  theme(legend.position = "none")
```



Vizualizirajmo podatke iz stupca "POSTURE" u odnosu na podatke iz stupca "GROUP".

```
mosaicplot(table(dataset$POSTURE, dataset$GROUP),
           main = "Mosaic Plot for POSTURE i GROUP",
           color = brewer.pal(4, "Dark2"))
```



Provest ćemo χ^2 test za dvije kategorijske varijable gdje ćemo proučavati stupce "POSTURE" i "IE" (introvert/ekstrovert) podatkovnog okvira.

Prije nego krenemo s testom, moramo provjeriti imamo li uvjete za njegovu provedbu. Najmanja očekivana vrijednost svake ćelije mora biti veća ili jednaka 5, a to provjeravamo pomoću funkcije `check_expected`.

Nulta hipoza testa je da su varijable nezavisne, a alternativna hipoteza je da su varijable zavisne.

$H0: \ \ POSTURE \ \ i \ \ IE \ \ su \ \ nezavisne \ \ varijable$

$H1: \ \ POSTURE \ \ i \ \ IE \ \ su \ \ zavisne \ \ varijable$

```r
contingency_table <- table(dataset$POSTURE, dataset$IE)
kable(contingency_table, caption = "Kontingencijska tablica za POSTURE i IE", align = "r")
```

Table 1: Kontingencijska tablica za POSTURE i IE

	E	I
idealno	21	1
kifoza/lordoza	30	6
ravna leđa	8	11
nagnuto	6	14

```r
check_expected(contingency_table)
```

```
## Table meets the expected values criteria
```

```r
chisq <- chisq.test(contingency_table)
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 30.114, df = 3, p-value = 1.306e-06
```

Provedbom χ^2 testa dobivamo iznimno maleni p-value, iz čega slijedi da odbacujemo nultu hipotezu uz razinu značajnosti od 0.05 te zaključujemo da su ekstrovertnost i način držanja zavisne varijable

Isti postupak ćemo ponoviti i za raspoznavanje/intuiciju.

$H0: \ POSTURE \ i \ SN \ su \ nezavisne \ varijable$

$H1: \ POSTURE \ i \ SN \ su \ zavisne \ varijable$

```r
contingency_table <- table(dataset$POSTURE, dataset$SN)
kable(contingency_table, caption = "Kontingencijska tablica za POSTURE i SN", align = "r")
```

Table 2: Kontingencijska tablica za POSTURE i SN

	N	S
idealno	10	12
kifoza/lordoza	11	25
ravna leđa	7	12
nagnuto	7	13

```r
check_expected(contingency_table)
```

```
## Table meets the expected values criteria
```

```r
chisq <- chisq.test(contingency_table)
chisq
```

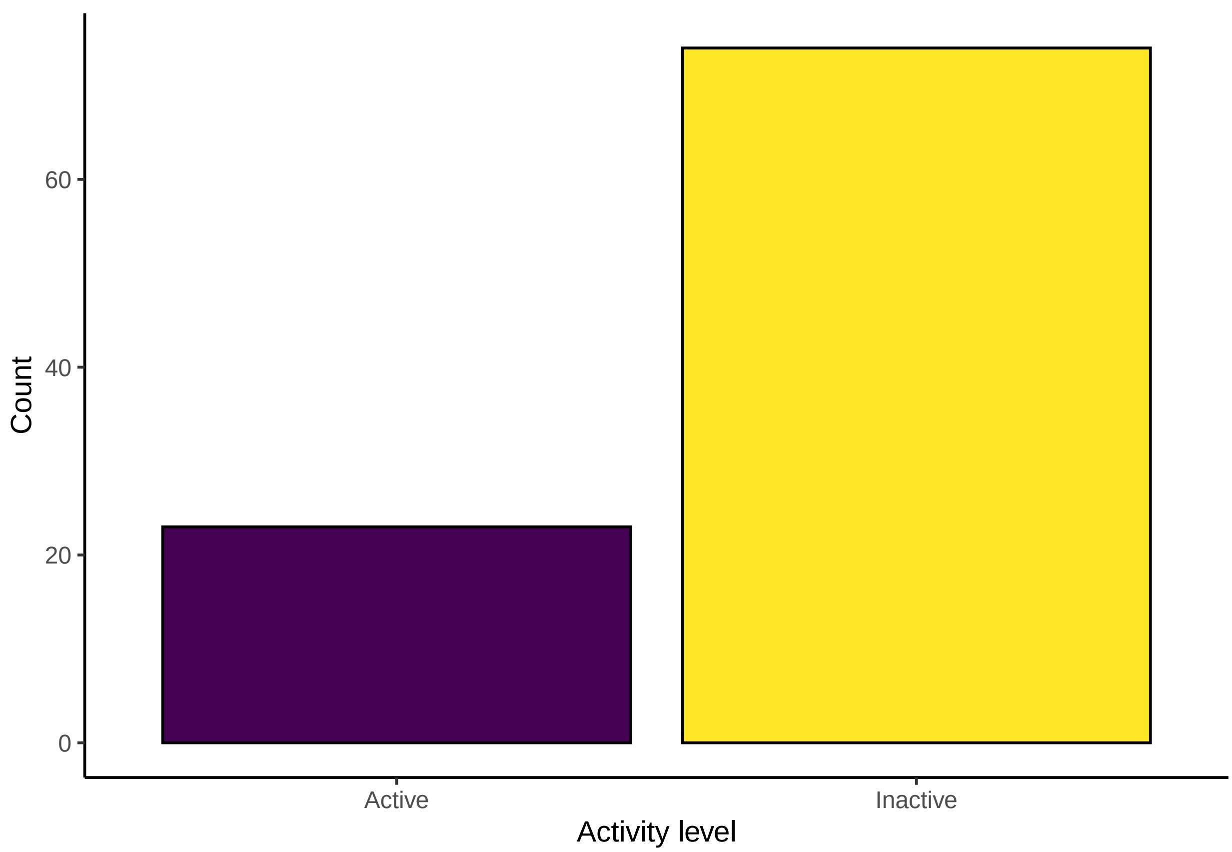
```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 1.3296, df = 3, p-value = 0.7221
```

χ^2 test nam u ovom slučaju daje p-value od 0.7, iz čega slijedi da ne možemo odbaciti nultu hipotezu uz razinu značajnosti od 0.05 te zaključujemo da raspoznavanje/intuicija i način držanja ne ovise jedno o drugome.

Veza između fizičke aktivnosti i razine ekstrovertiranosti

Fizičku aktivnost nam predstavlja stupac "IS_ACTIVE":

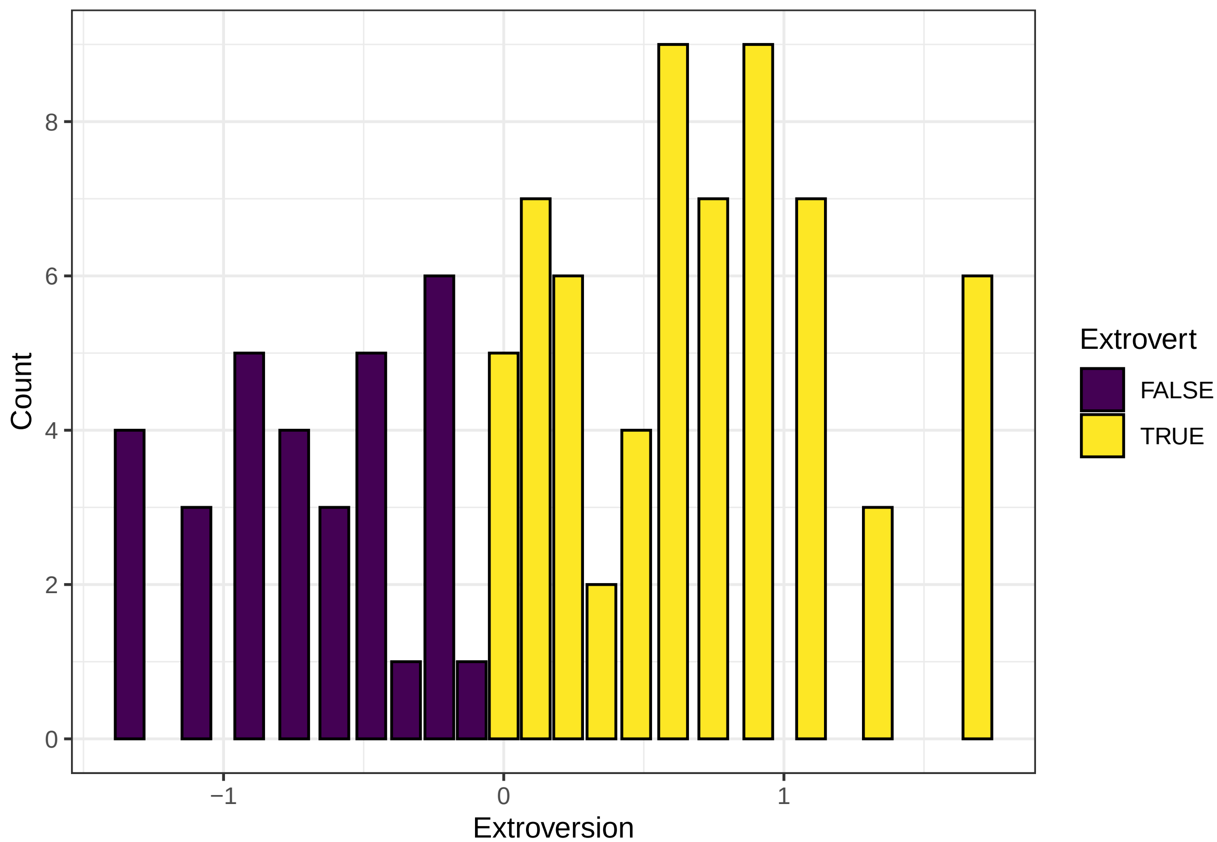
```r
ggplot(dataset, aes(x = IS_ACTIVE, fill = IS_ACTIVE)) +
  geom_bar(color = "black") +
  scale_fill_ordinal() +
  labs(x = "Activity level", fill = "Activity level", y = "Count") +
  theme_classic() +
  theme(legend.position = "none")
```



Prema ovom grafu ljude dijelimo u fizički aktivne i neaktivne.

Za koeficijent ekstrovertnosti imamo stupac s nazivom “E”:

```
ggplot(dataset, aes(x = E, fill = E > I)) +
  geom_bar(color = "black") +
  scale_fill_ordinal() +
  labs(x = "Extroversion", fill = "Extrovert", y = "Count") +
  scale_y_continuous(breaks = seq(0, 10, 2)) +
  theme_bw()
```



Na grafu vidimo raspodjelu ljudi koji imaju veći koeficijent ekstrovertnosti od koeficijenta introvertnosti.

Nezavisnost razine aktivnosti i ekstrovertnosti

Provest ćemo χ^2 test za kategorijske podatke gdje ćemo proučavati gore navedene stupce.

```
contingency_table <- table(dataset$IS_ACTIVE, dataset$IE)
kable(contingency_table, caption = "Kontingencijska tablica za IS_ACTIVE i IE", align = "r")
```

Table 3: Kontingencijska tablica za IS_ACTIVE i IE

	E	I
Active	17	6
Inactive	48	26

```
check_expected(contingency_table)
```

```
## Table meets the expected values criteria
```

```
chisq <- chisq.test(contingency_table)
chisq
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 0.30497, df = 1, p-value = 0.5808
```

Na temelju rezultata testa, na razini značajnosti 0.05, ne odbacujemo nultu hipotezu te zaključujemo da ne postoji veza između fizičke aktivnosti i ekstrovertnosti.

Razlika srednje vrijednosti ekstrovertnosti kod aktivnih i neaktivnih ljudi

Sada ćemo provesti t-test gdje ćemo usporediti srednje vrijednosti stupca "E" (koeficijent ekstrovertnosti) za fizički aktivne i neaktivne osobe, kako bi usporedili rezultate t-testa s dobivenim rezultatima χ^2 testa.

Prije provedbe t-testa provest ćemo f-test kako bismo provjerili jednakost varijanci budući da su one nepoznate.

$H0: \quad \sigma_{\text{active}} = \sigma_{\text{inactive}}$

$H1: \quad \sigma_{\text{active}} \neq \sigma_{\text{inactive}}$

```
f_test <- var.test(E ~ IS_ACTIVE, data = dataset)
f_test
```

```
##
##  F test to compare two variances
##
## data:  E by IS_ACTIVE
## F = 1.0246, num df = 22, denom df = 73, p-value = 0.8948
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
##  0.5487212 2.1727909
## sample estimates:
## ratio of variances
##           1.024608
```

Provedbom f-testa dobili smo p-value od 0.8948 na temelju kojeg uz razinu značajnosti 0.05 ne odbacujemo nultu hipotezu te zaključujemo da su varijance jednake.

Sad ćemo provesti t-test:

```r
t_test <- t.test(E ~ IS_ACTIVE, data = dataset, var.equal = TRUE)
t_test
```

```
##
##  Two Sample t-test
##
## data:  E by IS_ACTIVE
## t = 1.0445, df = 95, p-value = 0.2989
## alternative hypothesis: true difference in means between group Active and group Inactive is not equal
## 95 percent confidence interval:
##  -0.1793685  0.5776305
## sample estimates:
##   mean in group Active mean in group Inactive
##              0.3917264              0.1925954
```

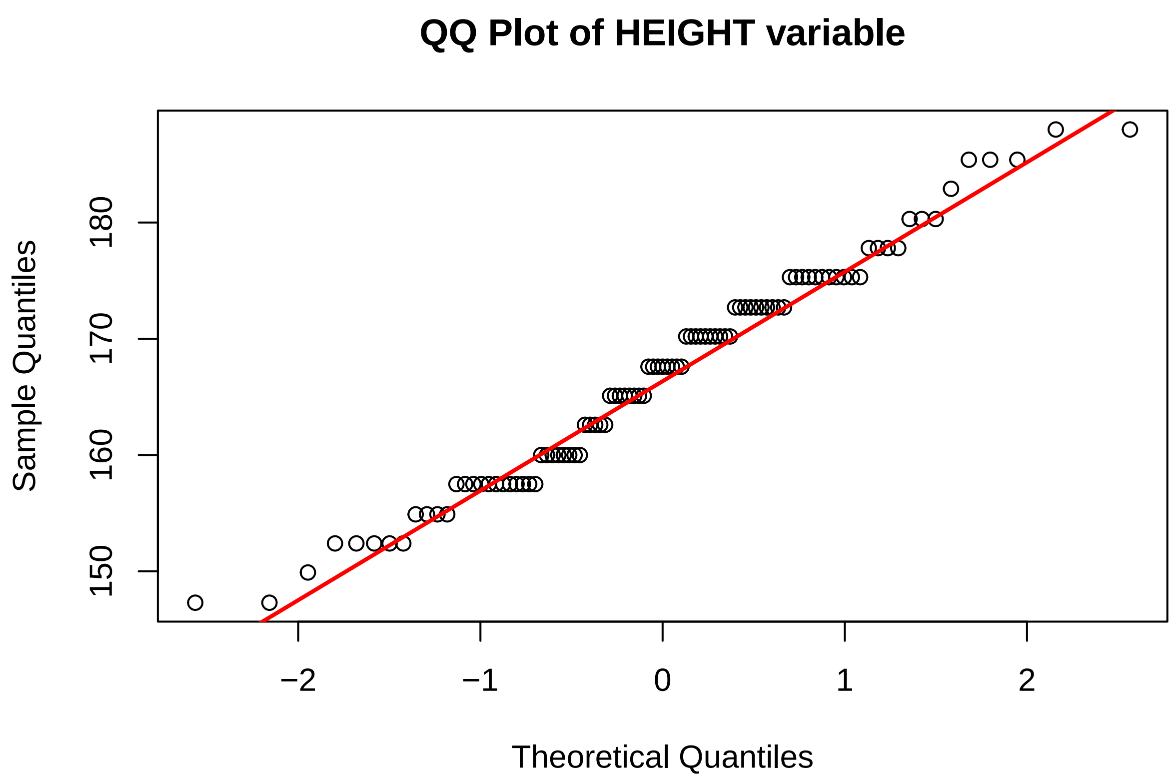
Na temelju rezultata t-testa ne odbacujemo nultu hipotezu te zaključujemo da ne postoji statistički značajna razlika između srednjih vrijednosti koeficijenata ekstrovertnosti za fizički aktivne i neaktivne osobe, što je u skladu s rezultatima χ^2 testa koji nam kaže da ne postoji veza između fizičke aktivnosti i ekstrovertnosti.

Razlika u visini/težini obzirom na tip ličnosti

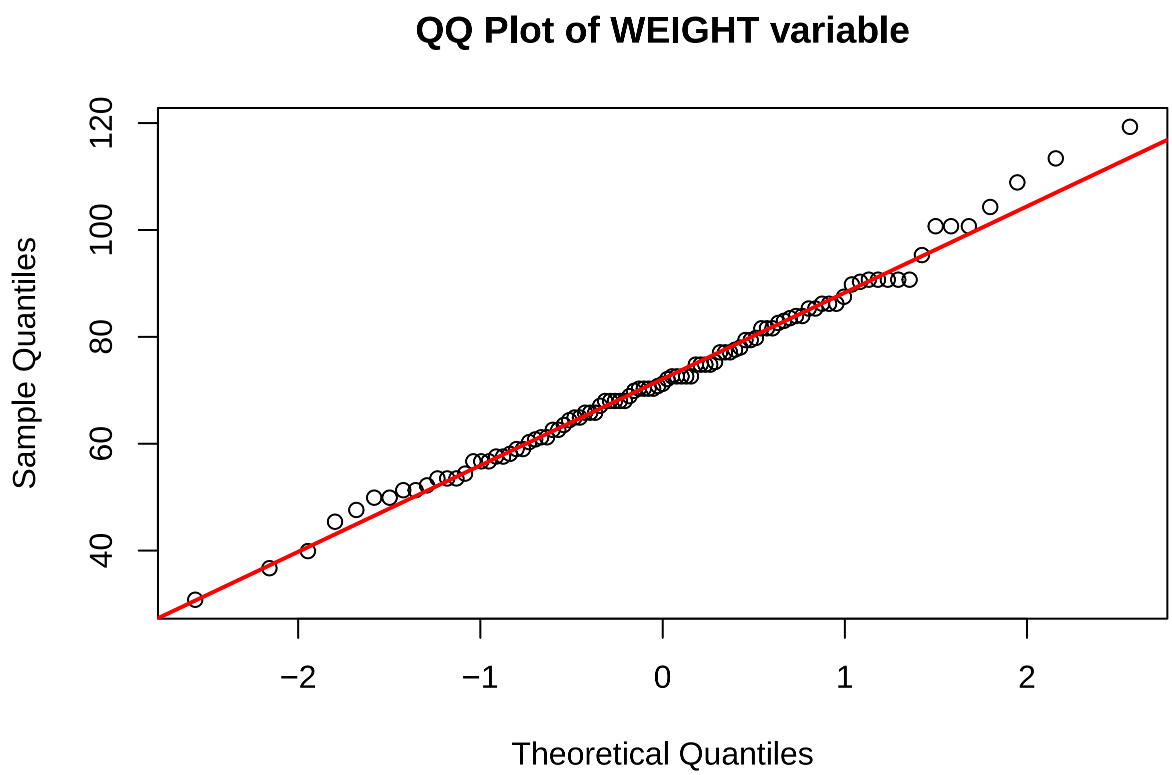
U našem podatkovnom skupu imamo stupce "WEIGHT" (težina), "HEIGHT" (visina), "GROUP" (stupac u kojem je 16 tipova osobnosti raspoređeno u 4 grupe) te stupac "IE" (u kojem je sadržan podatak o tome je li osoba introvert ili ekstrovert) koje ćemo koristiti u analiziranju visine i težine obzirom na tip ličnosti.

Prvo ćemo provjeriti normalnost numeričkih varijabli "HEIGHT" i "WEIGHT koje ćemo koristiti tokom analize.

```
qqnorm(dataset$HEIGHT, main = "QQ Plot of HEIGHT variable")
qqline(dataset$HEIGHT, col = "red", lwd = 2)
```



```r
qqnorm(dataset$WEIGHT, main = "QQ Plot of WEIGHT variable")
qqline(dataset$WEIGHT, col = "red", lwd = 2)
```



Iz grafova možemo naslutiti kako se radi o naizgled normalno distribuiranim varijablama.

Pogledajmo sada konkretne p-vrijednosti Shapiro-Wilkovog testa normalnosti kako bi ustanovili pripadaju li ove varijable uistinu normalnoj distribuciji.

Provjerimo prvo varijablu HEIGHT:

$H0: \ \ Varijabla \ \ je \ \ normalno \ \ distribuirana$

$H1: \ \ Varijabla \ \ nije \ \ normalno \ \ distribuirana$

```r
shapiro.test(dataset$HEIGHT)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset$HEIGHT
## W = 0.978, p-value = 0.1028
```

Shapiro-Wilkov test normalnosti nam daje p-vrijednost 0.1028 uz razinu značajnosti 0.05, što znači da ne možemo odbaciti nultu hipotezu i zaključujemo kako je varijabla HEIGHT normalno distribuirana.

Provedimo isti test za varijablu WEIGHT:

$H0: \ Varijabla \ \ je \ \ normalno \ \ distribuirana$

$H1: \ Varijabla \ \ nije \ \ normalno \ \ distribuirana$

```
shapiro.test(dataset$WEIGHT)
```

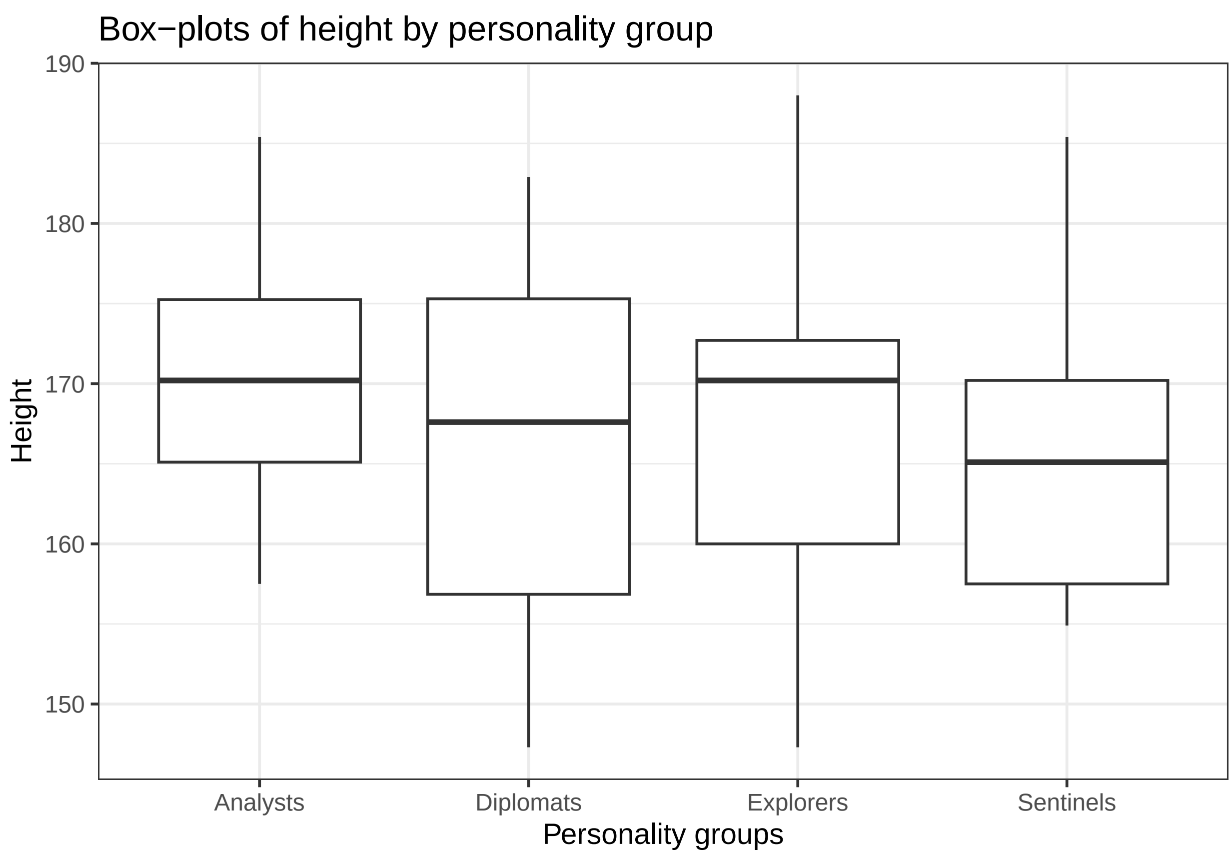
```
##
##  Shapiro-Wilk normality test
##
## data:  dataset$WEIGHT
## W = 0.99172, p-value = 0.8154
```

Za varijablu WEIGHT Shapiro-Wilkov test normalnosti nam daje p-vrijednost od 0.81 iz ćega slijedi da uz razinu značajnosti 0.05 ne možemo odbaciti nultu hipotezu i zaključujemo da je varijabla WEIGHT normalno distribuirana.

Usporedba srednjih visina za četiri grupe osobnosti

Za usporedbu srednjih visina za četiri grupe osobnosti primijenit ćemo analizu varijance (ANOVA) na naš podatkovni skup kako bismo usporedili srednje vrijednosti visine za četiri grupe osobnosti.

```
ggplot(dataset, aes(x = GROUP,y = HEIGHT)) +
  geom_boxplot() +
  labs(x = "Personality groups",
       y = "Height",
       title = "Box-plots of height by personality group") +
  theme_bw()
```



Iz ovog grafa možemo vidjeti distribucije pojedinih grupa osobnosti u obliku box-plota iz kojih dobijemo vizualnu reprezentaciju njihovih medijana i interkvartilnih rangova.

Možemo vidjeti kako su medijani grupa "Analysts" i "Explorers" na rubu granica interkvartilnog ranga grupe "Sentinels", stoga ovdje možemo očekivati neke probleme oko jednakosti sredina i standardnih devijacija ovih grupa.

Postavljamo hipoteze, neka je μ_{i} prosječna vrijednost visina i-te skupine. Osobnosti su grupirane u četiri skupine: "Analysts", "Diplomats", "Explorers" i "Sentinels".

$$H0: \quad \mu_{\text{analysts}} = \mu_{\text{diplomats}} = \mu_{\text{explorers}} = \mu_{\text{sentinels}}$$

$$H1: \quad barem \quad dva \quad \mu_{\text{i}} \quad nisu \quad jednaka.$$

Za hipoteze postavljene na ovaj način, nije preporučljivo koristiti t-test više puta zbog povećanja rizika od greške tipa I.

Stoga smo odabrali ANOVA-u (Analizu varijance) kao odgovarajuću metodu.

ANOVA se primjenjuje uz pretpostavke normalnosti distribucije reziduala, jednakih varijanci između grupa i nezavisnosti podataka unutar grupa.

Prije provedbe ANOVA-e moramo pokazati da su varijance grupa jednake.

Proučimo varijance visina za pojedine grupe osobnosti:

```
vars_height <- aggregate(HEIGHT ~ GROUP, data = dataset, var)
vars_height
```

```
##       GROUP    HEIGHT
## 1  Analysts  82.44418
## 2 Diplomats 110.47998
## 3 Explorers  94.72246
## 4 Sentinels  72.44195
```

Provest ćemo Bartlettov test o jednakosti varijanci koji testira hipoteze:

$H0: \quad \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4.$

$H1: \quad barem \quad dvije \quad \sigma_{\text{i}} \quad nisu \quad jednake.$

```
bartlett.test(HEIGHT ~ GROUP, data = dataset)
```

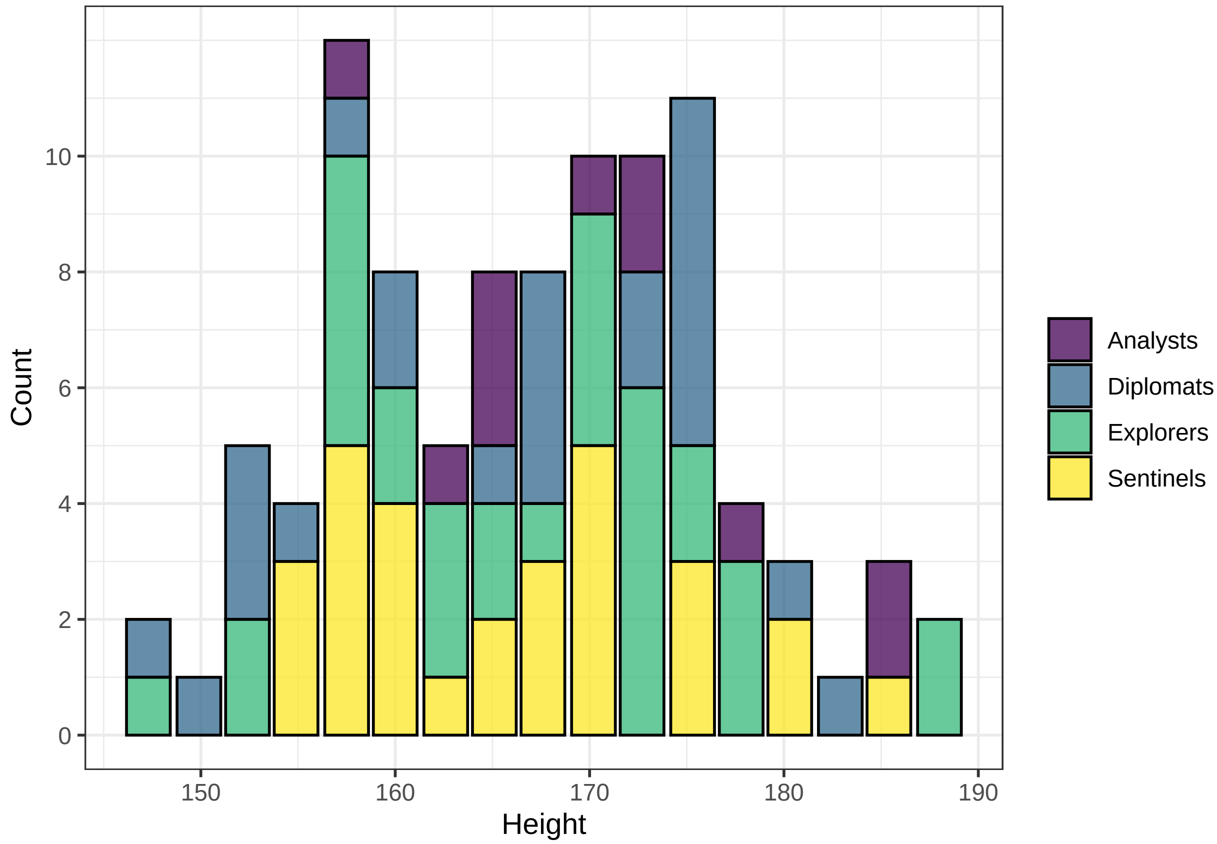
```
##
##  Bartlett test of homogeneity of variances
##
## data:  HEIGHT by GROUP
## Bartlett's K-squared = 1.184, df = 3, p-value = 0.7568
```

Vidimo da je p-vrijednost velika stoga ne odbacujemo nultu hipotezu te zaključujemo da su varijance među grupama jednake.

Sada smo sigurni da možemo provesti ANOVA-u.

Na sljedećem grafu možemo vidjeti distribucije pojedinih grupa osobnosti s obzirom na njihove visine.

```
ggplot(dataset, aes(x = HEIGHT, fill = GROUP)) +
  geom_bar(color = "black", alpha = 0.75) +
  labs(x = "Height", y = "Count", fill = "") +
  scale_y_continuous(breaks = seq(0, 10, 2)) +
  scale_fill_ordinal() +
  theme_bw()
```



Pogledajmo srednje vrijednosti visina za pojedine grupe osobnosti:

```
means_height <- aggregate(HEIGHT ~ GROUP, data = dataset, mean)
means_height
```

```
##       GROUP   HEIGHT
## 1  Analysts 170.8727
## 2 Diplomats 165.9458
## 3 Explorers 167.3394
## 4 Sentinels 165.8862
```

Sada provedimo test:

```r
model <- aov(HEIGHT ~ GROUP, data = dataset)
summary(model)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## GROUP        3    231   77.09   0.851   0.47
## Residuals   93   8425   90.59
```

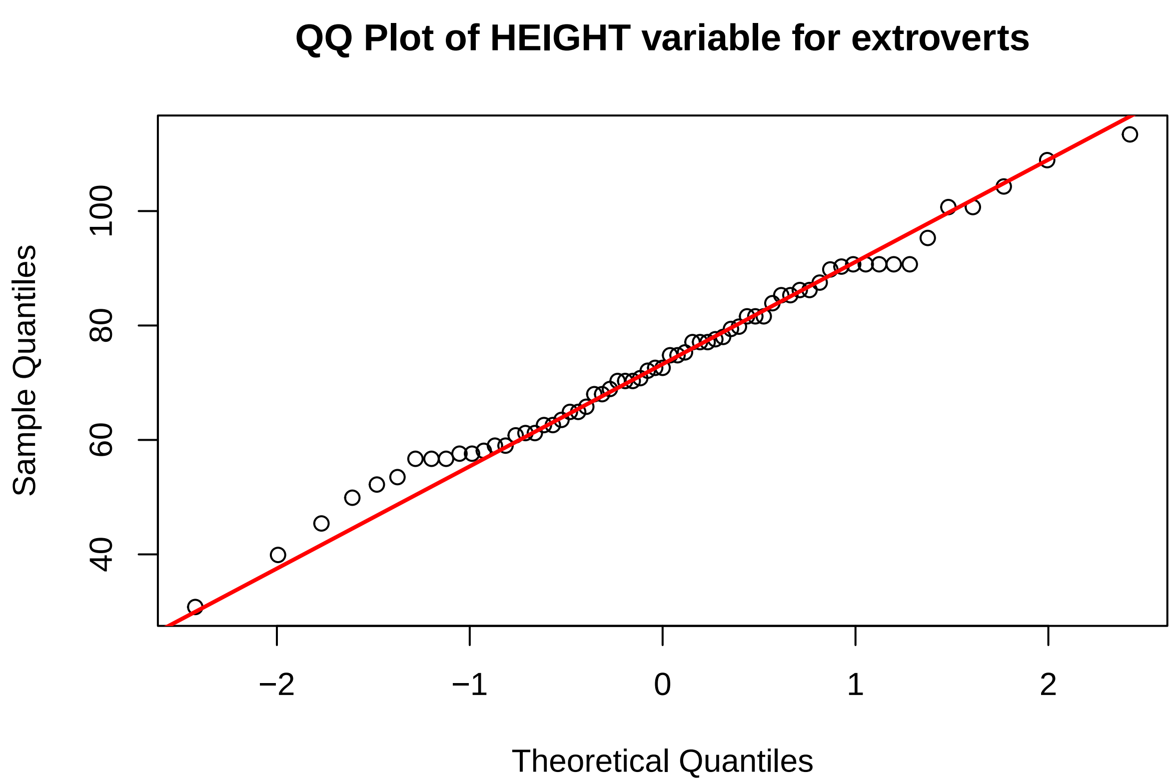
Obzirom na p-vrijednost od 0.47 ne odbacujemo nultu hipotezu. To sugerira da nema dovoljno dokaza za zaključak da postoje značajne razlike u prosječnoj visini između četiri skupine osobnosti.

Usporedba srednjih težina za ekstroverte i introverte

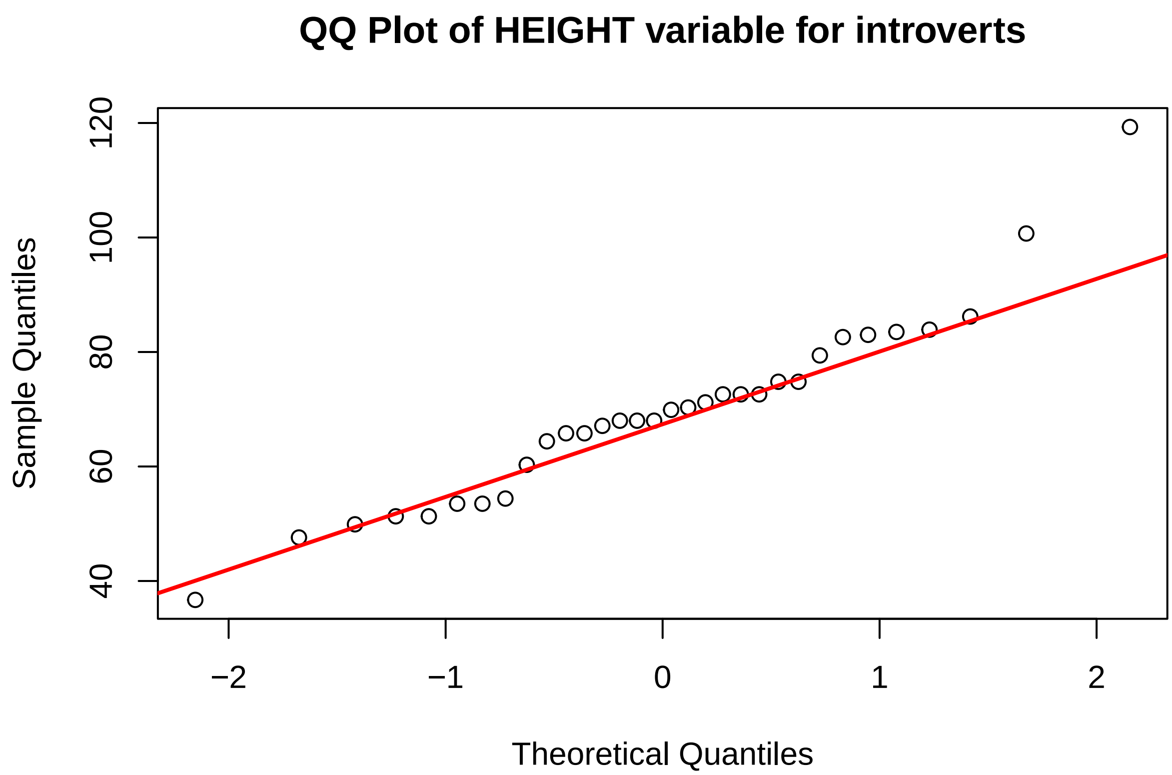
Obzirom da u ovom koraku varijablu WEIGHT grupiramo u dvije grupe, težine ekstroverata i introverata, moramo provjeriti je li varijabla WEIGHT normalno distribuirana za svaku grupu.

```r
weight_extroverts <- subset(dataset, IE == "E")$WEIGHT
weight_introverts <- subset(dataset, IE == "I")$WEIGHT
```

```r
qqnorm(weight_extroverts, main = "QQ Plot of HEIGHT variable for extroverts")
qqline(weight_extroverts, col = "red", lwd = 2)
```



```r
qqnorm(weight_introverts, main = "QQ Plot of HEIGHT variable for introverts")
qqline(weight_introverts, col = "red", lwd = 2)
```



Sada provodimo Shapiro-Wilkov test normalnosti varijable HEIGHT za svaku grupu:

```r
shapiro.test(weight_extroverts)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight_extroverts
## W = 0.99249, p-value = 0.9642
```

Uz dobiveni p-value od 0.96 i razinu značajnosti od 0.05 ne odbacujemo nultu hipotezu te zaključujemo da je varijabla HEIGHT normalno distribuirana za grupu ekstroverata.

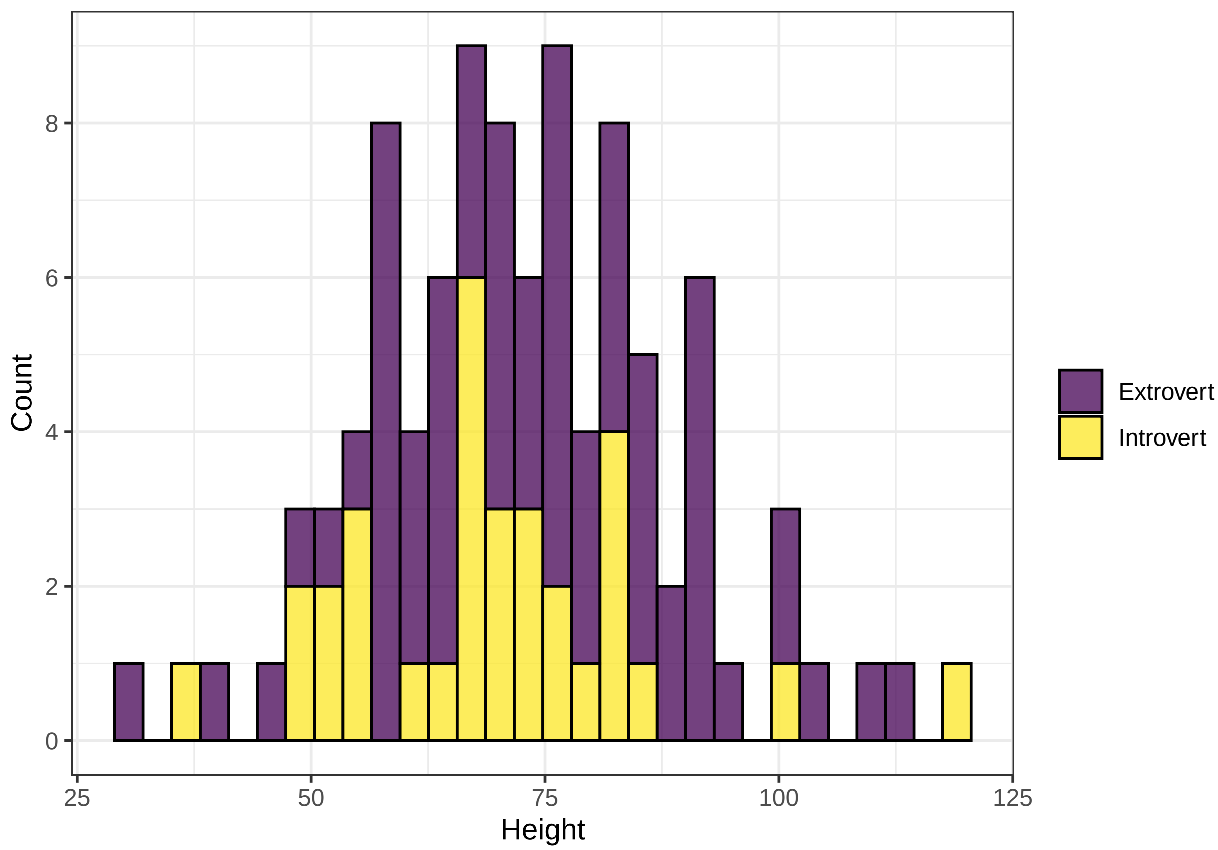
```r
shapiro.test(weight_introverts)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight_introverts
## W = 0.94796, p-value = 0.1261
```

Dobiveni p-value za WEIGHT za introverte iznosi 0.1261 što je veće od 0.05, stoga ne odbacujemo nultu hipotezu te zaključujemo da je varijabla HEIGHT normalno distribuirana za grupu introverata, što znači da možemo provesti t-test.

Na sljedećem grafu možemo vidjeti distribuciju introverata/ekstroverata s obzirom na njihove težine.

```
ggplot(dataset, aes(x = WEIGHT, fill = IE)) +
  geom_histogram(color = "black", bins = 30, alpha = 0.75) +
  labs(x = "Height", y = "Count", fill = "") +
  scale_y_continuous(breaks = seq(0, 10, 2)) +
  scale_fill_ordinal(labels = c("Extrovert", "Introvert")) +
  theme_bw()
```



Sada ćemo usporediti srednje vrijednosti težina ekstroverata i introverata koristeći stupce “IE” i “WEIGHT” koristeći t-test.

Prije provedbe t-testa moramo provesti f-test da provjerimo jednakost varijanci budući da su nepoznate.

Pogledajmo prije varijance težina introverata:

```
vars_weight <- aggregate(WEIGHT ~ IE, data = dataset, var)
vars_weight
```

```
##   IE   WEIGHT
## 1  E 270.7246
## 2  I 263.5132
```

F-test ćemo provesti s hipotezama:

$H0: \quad \sigma_{\text{introvert}} = \sigma_{\text{extrovert}}$

$H1: \quad \sigma_{\text{introvert}} \neq \sigma_{\text{extrovert}}$

```r
f_test <- var.test(HEIGHT ~ IE, data = dataset)
f_test
```

```
##
##  F test to compare two variances
##
## data:  HEIGHT by IE
## F = 0.87582, num df = 64, denom df = 31, p-value = 0.6418
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4577772 1.5681367
## sample estimates:
## ratio of variances
##          0.8758191
```

F-test daje p-vrijednost 0.6418 te zaključujemo da su varijance jednake.

Sada možemo provesti t-test uz jednake, ali nepoznate varijance.

Pogledajmo srednje vrijednosti težina ekstroverata i introverata.

```r
means_weight <- aggregate(WEIGHT ~ IE, data = dataset, mean)
means_weight
```

```
##   IE   WEIGHT
## 1  E 73.71692
## 2  I 69.46875
```

Test provodimo s hipotezama:

$H0: \quad \mu_{\text{introvert}} = \mu_{\text{extrovert}}$

$H1: \quad \mu_{\text{introvert}} \neq \mu_{\text{extrovert}}$

```r
t_test <- t.test(HEIGHT ~ IE, data = dataset, var.equal = TRUE)
t_test
```

```
##
##  Two Sample t-test
##
## data:  HEIGHT by IE
## t = -0.032885, df = 95, p-value = 0.9738
## alternative hypothesis: true difference in means between group E and group I is not equal to 0
## 95 percent confidence interval:
##  -4.160115  4.024538
## sample estimates:
## mean in group E mean in group I
##        166.9385        167.0062
```

Zbog velike p-vrijednosti ne odbacujemo nultu hipotezu te zaključujemo da su srednje vrijednosti visina za introverte i ekstroverte jednake.

Tip ličnosti na temelju pojedinih karakteristika

Predviđanje paramatera ličnosti na temelju ostalih karakteristika

Sada ćemo pokušati predvidjeti karakteristike ličnosti na temelju pojedinih obilježja.

Ne možemo izračunati točnu vezu s MBTI tipom osobnosti jer je to kategorijska varijabla, stoga ćemo pokušavati predvidjeti koeficijente za pojedino svojstvo (npr. ekstrovertnost, intuitivnost, itd.) koristeći linearnu regresiju i uspoređivanjem dobivenih koeficijenata za isto svojstvo, u našem primjeru za J (prosuđivanje) i P (opažanje).

Osoba pripada kategoriji J ili P ovisno o tome koji je koeficijent (predviđan linearnom regresijom) veći.

Najprije je potrebno testirati više modela linearne regresije odabirom različitih regresora kako bi pronašli najbolji model.

Za pojedine modele moramo izračunati R^2 vrijednost, koja nam govori koliko dobro model objašnjava varijabilnost podataka. Model s najvećom R^2 vrijednosti je najbolji model.

Nakon što odaberemo najbolji model, izračunat ćemo koeficijente za J i P, te na temelju njih odrediti tu kategoriju.

Prvi model za J i P karakteristike:

```r
model <- lm(formula = J ~ HEIGHT + SEX + AGE, data = dataset)
model_summary <- summary(model)
r_squared1 <- model_summary$r.squared

model <- lm(formula = P ~ HEIGHT + SEX + AGE, data = dataset)
model_summary <- summary(model)
r_squared2 <- model_summary$r.squared

cat("R^2 za predikciju slova J prvim modelom:", r_squared1 %>% round(4),
    "\nR^2 za predikciju slova P prvim modelom:", r_squared2 %>% round(4), "\n")
```

```
## R^2 za predikciju slova J prvim modelom: 0.1589
## R^2 za predikciju slova P prvim modelom: 0.1337
```

Drugi model za J i P karakteristike:

```r
model <- lm(formula = J ~ HEIGHT + WEIGHT + SEX + AGE + ACTIVITY_LEVEL, data = dataset)
model_summary <- summary(model)
r_squared1 <- model_summary$r.squared
```

```
model <- lm(formula = P ~ HEIGHT + WEIGHT + SEX + AGE + ACTIVITY_LEVEL, data = dataset)
model_summary <- summary(model)
r_squared2 <- model_summary$r.squared

cat("R^2 za predikciju slova J drugim modelom:", r_squared1 %>% round(4),
    "\nR^2 za predikciju slova P drugim modelom:", r_squared2 %>% round(4), "\n")
```

```
## R^2 za predikciju slova J drugim modelom: 0.1765
## R^2 za predikciju slova P drugim modelom: 0.1514
```

Treći model za J i P karakteristike:

```
model <- lm(formula = J ~ HEIGHT + WEIGHT + SEX + AGE + ACTIVITY_LEVEL +
                          PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4,
            data = dataset)
model_summary <- summary(model)
r_squared1 <- model_summary$r.squared

model <- lm(formula = P ~ HEIGHT + WEIGHT + SEX + AGE + ACTIVITY_LEVEL +
                          PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4, data = dataset)
model_summary <- summary(model)
r_squared2 <- model_summary$r.squared

cat("R^2 za predikciju slova J trećim modelom:", r_squared1 %>% round(4),
    "\nR^2 za predikciju slova P trecim modelom:", r_squared2 %>% round(4), "\n")
```

```
## R^2 za predikciju slova J trećim modelom: 0.2055
## R^2 za predikciju slova P trecim modelom: 0.1817
```

Vidimo kako je najprikladniji ispao treći model jer ima najveće R^2 vrijednosti u iznosima od 0.2055 za predikciju karakteristike J i 0.1817 za predikciju karakteristike P.

Testirajmo sada točnost modela

```
model <- lm(formula = J ~ HEIGHT + WEIGHT + SEX + AGE + ACTIVITY_LEVEL +
                          PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4,
            data = dataset)
model_summary <- summary(model)
probs1 <- predict(model, newdata = dataset, type = "response")

model <- lm(formula = P ~ HEIGHT + WEIGHT + SEX + AGE + ACTIVITY_LEVEL +
                          PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4,
            data = dataset)
model_summary <- summary(model)
probs2 <- predict(model, newdata = dataset, type = "response")

slova <- ifelse(probs1>probs2, "J", "P")
vrijednosti <- ifelse(slova == substr(dataset$MBTI, 4, 4), 1, 0)
uk <- sum(vrijednosti)/length(vrijednosti)

cat(paste0("Točnost: ", 100*uk %>% round(4), "%"))
```

```
## Točnost: 62.89%
```

Kod procjene pripada li osoba kategoriji J ili P, dobili smo točnost od 62.89%, što je zadovoljavajuće. Ovo je dobar rezultat s obzirom na to da je naš model bio jednostavan odnosno nije sadržavao puno regresora. Još jedan problem na koji smo naišli je bila činjenica da podaci nisu bili ujednačeni, to jest određenih osobnosti nije bilo željene količine.

Ponovimo testiranje za još jednu karaktristiku, npr. razmišljanje i osjećanje

Prvi model za T i F karakteristike:

```
model <- lm(formula = F ~ HEIGHT + SEX + AGE, data = dataset)
model_summary <- summary(model)
r_squared1 <- model_summary$r.squared

model <- lm(formula = T ~ HEIGHT + SEX + AGE, data = dataset)
model_summary <- summary(model)
r_squared2 <- model_summary$r.squared

cat("R^2 za predikciju slova F prvim modelom:", r_squared1 %>% round(4),
    "\nR^2 za predikciju slova T prvim modelom:", r_squared2 %>% round(4), "\n")
```

```
## R^2 za predikciju slova F prvim modelom: 0.1647
## R^2 za predikciju slova T prvim modelom: 0.1719
```

Drugi model za T i F karakteristike:

```
model <- lm(formula = F ~ HEIGHT + WEIGHT + SEX + log(AGE) + ACTIVITY_LEVEL +
                          PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4,
            data = dataset)
model_summary <- summary(model)
r_squared1 <- model_summary$r.squared

model <- lm(formula = T ~ HEIGHT + WEIGHT + SEX + log(AGE) + ACTIVITY_LEVEL +
                          PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4,
            data = dataset)
model_summary <- summary(model)
r_squared2 <- model_summary$r.squared

cat("R^2 za predikciju slova F drugim modelom:", r_squared1 %>% round(4),
    "\nR^2 za predikciju slova T drugim modelom:", r_squared2 %>% round(4), "\n")
```

```
## R^2 za predikciju slova F drugim modelom: 0.2049
## R^2 za predikciju slova T drugim modelom: 0.2095
```

Treći model za T i F karakteristike:

```
model <- lm(formula = F ~ HEIGHT + WEIGHT + SEX + AGE + ACTIVITY_LEVEL +
                          PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4,
            data = dataset)
model_summary <- summary(model)
r_squared1 <- model_summary$r.squared
```

```
model <- lm(formula = T ~ HEIGHT + WEIGHT + SEX + AGE + ACTIVITY_LEVEL +
                          PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4,
            data = dataset)
model_summary <- summary(model)
r_squared2 <- model_summary$r.squared

cat("R^2 za predikciju slova F trecim modelom:", r_squared1 %>% round(4),
    "\nR^2 za predikciju slova T trecim modelom:", r_squared2 %>% round(4), "\n")
```

```
## R^2 za predikciju slova F trecim modelom: 0.2034
## R^2 za predikciju slova T trecim modelom: 0.2078
```

Drugi i treći modeli su znatno bolji od prvog, što je i očekivano s obzirom na to da su sadržavali više regresora, no treći model se ipak pokazao malo bolji od drugog.

Testirajmo sada točnost modela

```
model <- lm(formula = F ~ HEIGHT + WEIGHT + SEX + log(AGE) + ACTIVITY_LEVEL +
                          PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4,
            data = dataset)
model_summary <- summary(model)
probs1 <- predict(model, newdata = dataset, type = "response")

model <- lm(formula = T ~ HEIGHT + WEIGHT + SEX + log(AGE) + ACTIVITY_LEVEL +
                          PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4,
            data = dataset)

model_summary <- summary(model)
probs2 <- predict(model, newdata = dataset, type = "response")

slova <- ifelse(probs1>probs2, "F", "T")
vrijednosti <- ifelse(slova == substr(dataset$MBTI, 3, 3), 1, 0)
uk <- sum(vrijednosti)/length(vrijednosti)

cat(paste0("Točnost: ", 100*uk %>% round(4), "%"))
```

```
## Točnost: 70.1%
```

Točnost danog modela iznosi 70.1%, što je znatno bolje od slučajnog odabira te smatramo to dovoljno dobrom predikcijom.

Probajmo sad koristiti i koeficijente drugih karakteristika kao regesore u modelu

Četvrti model za T i F karakteristike:

```
model <- lm(formula = J ~ HEIGHT + SEX + PAIN_3 + PAIN_4 + S + F + T + E + I,
            data = dataset)
model_summary <- summary(model)
r_squared1 <- model_summary$r.squared

model <- lm(formula = P ~ HEIGHT + SEX + PAIN_3 + PAIN_4 + S + F + T + E + I,
            data = dataset)
model_summary <- summary(model)
```

```r
r_squared2 <- model_summary$r.squared

cat("R^2 za predikciju slova J četvrtim modelom:", r_squared1 %>% round(4),
    "\nR^2 za predikciju slova P četvrtim modelom:", r_squared2 %>% round(4), "\n")
```

```
## R^2 za predikciju slova J četvrtim modelom: 0.359
## R^2 za predikciju slova P četvrtim modelom: 0.3172
```

Vidimo kako smo dobili veći R^2 izbacivši neke neznačajne regresore, a ubacivši koeficijente za neke druge karakteristike. To nam govori da unatoč tome što bi karakteristike trebale biti neovisne jedna od druge, na temelju nekih je moguće bolje predvidjeti drugu.
Testirajmo model

```r
model <- lm(formula = J ~ HEIGHT + SEX + PAIN_3 + PAIN_4 + S + F + T + E + I,
            data = dataset)
model_summary <- summary(model)
probs1 <- predict(model, newdata = dataset, type = "response")

model <- lm(formula = P ~ HEIGHT + SEX + PAIN_3 + PAIN_4 + S + F + T + E + I,
            data = dataset)

model_summary <- summary(model)
probs2 <- predict(model, newdata = dataset, type = "response")

slova <- ifelse(probs1>probs2, "J", "P")
vrijednosti <- ifelse(slova == substr(dataset$MBTI, 4, 4), 1, 0)
uk <- sum(vrijednosti)/length(vrijednosti)

cat(paste0("Točnost: ", 100*uk %>% round(4), "%"))
```

```
## Točnost: 60.82%
```

Vidimo kako smo koristeći druge koeficijente kao regresore dobili točnost skoro dobru kao i kod prijašnjeg modela

Predviđanje MBTI skupine na temelju pojedinih karakteristika

Sada ćemo pomoću logističke regresije pokušati predvidjeti skupinu MBTI tipa na temelju ostalih karakteristika

Prisjetimo se, skupine MBTI tipova su "Analysts", "Diplomats", "Explorers" i "Sentinels".

Model logističke regresije na temelju danih varijabli izbacuje vektor vrijednosti veličine c što je u našem slučaju 4, budući da imamo 4 različita MBTI tipa.

Svaka od tih vrijednosti je pomoću funkcije `softmax` preslikana na interval [0, 1] gdje suma svih vjerojatnosti za pojedine klase iznosi 1.

Taj vektor se zatim "ubacuje" u funkciju `argmax` koja vektor enkodira tzv. "one-hot" encodingom, koji ima vrijednost 1 na indeksu najveće vrijednosti, a 0 na ostalim indeksima.

Taj vektor se konačno uspoređuje s vektorom stvarnih vrijednosti i računa se točnost modela.

U našem prvom modelu ćemo dodati jedan skriveni sloj s 3 neurona, a u drugom modelu 2 skrivena sloja sa 4 i 2 neurona. Parametar `linear.output` postaviti na `TRUE` kako bi se klasična sigmoidalna funkcija u skrivenom sloju zamijenila linearnom ReLU funkcijom koja u praksi pokazuje bolje rezultate.

Definirajmo sada funkciju koja će nam pomoći u evaluaciji modela.

```
evaluate_model <- function(model, dataset) {
  GROUP <- dataset$GROUP
  one_hot_encoding <- c("Analysts", "Diplomats", "Explorers", "Sentinels")
  predictions <- predict(model, dataset, type = "response")
  predicted_classes <- max.col(predictions)

  accuracy <- sum(one_hot_encoding[predicted_classes] == GROUP) / length(GROUP)
  cat("Točnost modela: ", 100*accuracy %>% round(4), "%")
}
```

Provedimo konačno logističku regresiju na našem skupu podataka.

Prvi model će kao regresore koristiti samo visinu, težinu, dob i spol ispitanika.

```
set.seed(131213121)

encoded_SEX <- ifelse(dataset$SEX == "M", 1, 0)
dataset$encoded_SEX <- encoded_SEX

model <- neuralnet(formula = GROUP ~ HEIGHT + WEIGHT + AGE + encoded_SEX,
                   data = dataset,
                   hidden = c(3),
                   linear.output = T,
                   stepmax = 1000,
                   learningrate = 0.01
                   )

evaluate_model(model, dataset)
```

```
## Točnost modela:  35.05 %
```

Točnost ovog modela iznosi 35.05%, što je bolje od slučajnog odabira, ali i dalje nedovoljno dobro.

Drugi model će kao regresore koristiti PAIN_1, PAIN_2, PAIN_3 i PAIN_4, koji predstavljaju bol u vratu, bol u gornjem dijelu leđa, bol u srednjem dijelu leđa i bol u donjem dijelu leđa, respektivno.

```r
set.seed(1312131)

model <- neuralnet(formula = GROUP ~ PAIN_1 + PAIN_2 + PAIN_3 + PAIN_4,
                   data = dataset,
                   hidden = c(4, 2),
                   linear.output = T,
                   stepmax = 100000,
                   learningrate = 0.01
                   )

evaluate_model(model, dataset)
```

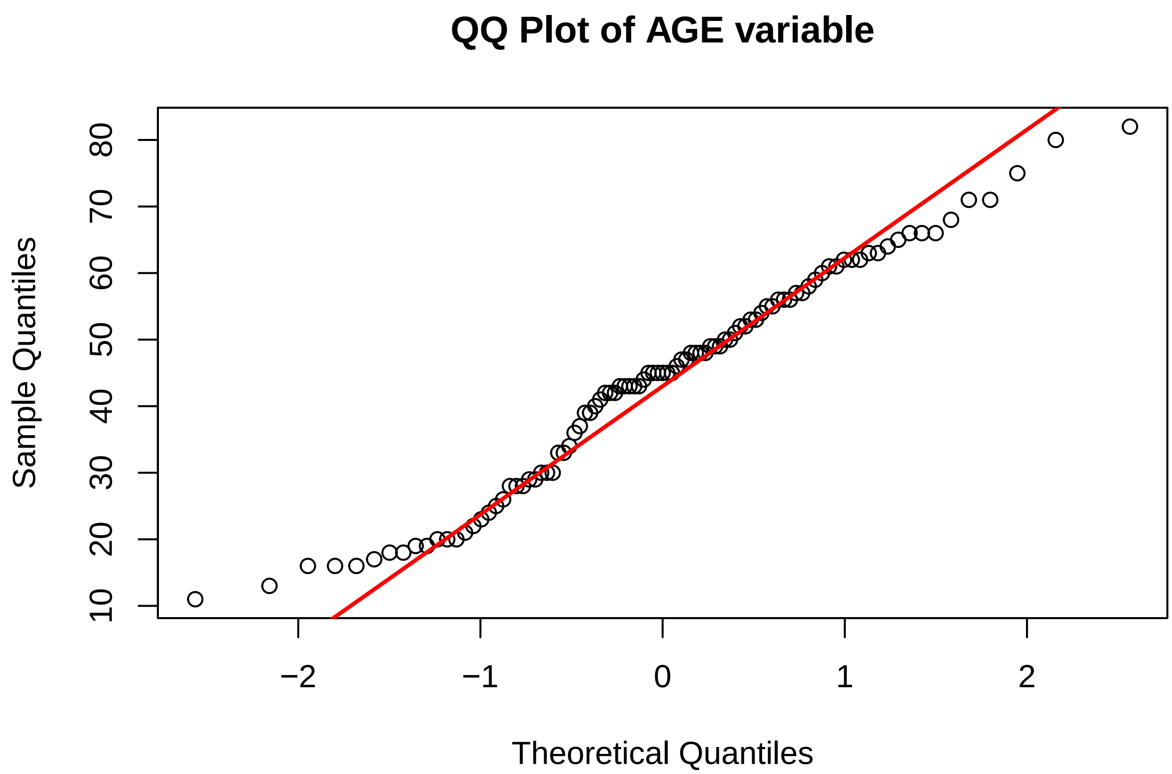
```
## Točnost modela:  60.82 %
```

Iznenađujuće, ovaj model ima točnost od 60.82%, što je znatno bolje od prethodnog modela.

Je li udio ekstrovertnih ljudi isti kod ljudi iznad i ispod 45 godina

Prije nego krenemo na testiranje, moramo provjeriti je li naša pretpostavka o normalnoj distribuciji podataka točna.

```
qqnorm(dataset$AGE, main = "QQ Plot of AGE variable")
qqline(dataset$AGE, col = "red", lwd = 2)
```



Iz grafa se čini da je pretpostavka o normalnoj distribuciji točna, no provjeravamo tu hipotezu i Shapiro-Wilkovim testom o normalnosti.

```
shapiro.test(dataset$AGE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset$AGE
## W = 0.97566, p-value = 0.06816
```

Uz p-value od 0.068 uz razinu značajnosti od 0.05 ne možemo odbaciti nultu hipotezu stoga zaključujemo da varijabla AGE ima normalnu distribuciju.

Također, poznavajući centralni granični teorem, znamo da će se uz dovoljno veliki uzorak, distribucija srednjih vrijednosti približiti normalnoj distribuciji, stoga ćemo koristiti z-test za testiranje.

Kako bismo testirali je li postotak ekstroverata isti kod mladjih i starijih ljudi, koristit ćemo Z-test o dvije proporcije, uz $\alpha = 0.05$, a hipoteze postavljamo ovako

$H0: \ postotak \ ekstroverata \ je \ isti \ neovisno \ o \ godinama \ (p_1 - p_2 = 0)$

$H1: \ postotak \ ekstroverata \ se \ razlikuje \ ovisno \ o \ godinama \ (p_1 - p_2 \neq 0)$

```r
subset_result <- subset(dataset, AGE < 46 & substr(MBTI, 1, 1) == "E")
Eyoung <- nrow(subset_result)
total_group1 <- sum(dataset$AGE <= 45)

subset_result <- subset(dataset, AGE > 45 & substr(MBTI, 1, 1) == "E")
Eold <- nrow(subset_result)
total_group2 <- sum(dataset$AGE > 45)

z_test_result <- prop.test(c(Eyoung, Eold),
                           c(total_group1, total_group2),
                           alternative = "two.sided")

z_test_result
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(Eyoung, Eold) out of c(total_group1, total_group2)
## X-squared = 0.32824, df = 1, p-value = 0.5667
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1325616  0.2834567
## sample estimates:
##    prop 1    prop 2
## 0.7058824 0.6304348
```

Provedeni z-test nam daje p-value 0.5667 koji je znatno veći od razine značajnosti 0.05 iz čega zaključujemo kako na temelju ovih podataka nemamo razloga sumnjati u H0 stoga ju niti ne odbacujemo.

Zaključak

Pokazali smo kako postoji povezanost između razine ekstrovertnosti i načina držanja. Ljudi koji se drže uspravno u prosjeku su ekstrovertniji od ljudi koji se drže pogrbljeno. S druge strane, utvrdili smo kako ne postoji povezanost intuicije/raspoznavanja i načina držanja. Omjer ekstrovertnih i introvertnih ljudi s obzirom na njihovu razinu aktivnosti je približno jednak, što znači da aktivnost ne utječe na ekstrovertnost. Koristeći ANOVA-u zaključili smo kako su prosječne težine također jednake za svaku od 4 MBTI skupine, a koristeći t-test kako su prosječne visine introverata i ekstroverata jednake, odnosno da nema značajne veze između težine i MBTI skupine ni visine i toga je li osoba ekstrovert. Pokazali smo i kako je postotak ekstroverata isti kod mlađih i starijih ljudi. Dali smo par primjera modela koje smo koristili za predikcije pojedinih karakteristika osobnosti na temelju drugih osobina koristeći linearnu regresiju. Na kraju smo napravili i dvije neuronske mreže koje su predviđale MBTI skupinu ispitanika na temelju njihovih fizičkih karakteristika.
