MBTI dataset transformation and analysis

Vedran Moškov, Lucija Runjić, Borna Josipović, Lana Bartolović

2024-01-17

Motivacija i opis problema

Istražujemo povezanost između osobnosti, dobivene kroz MBTI test, i fizičkih karakteristika poput držanja tijela, težine i visine. Kroz proučavanje pitanja kao što su "Kako tipovi ličnosti utječu na način držanja?" i "Postoje li razlike u tjelesnoj težini ili visini između različitih tipova ličnosti?" analizom ovih podataka, nastojimo bolje razumjeti veze između mentalnog i fizičkog aspekta individualnosti.

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>
```

```
tail(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    91     92    16     69    130 Female Moderate                 5        0
## 2    92     93    16     58    100 Male   Moderate                 0        0
## 3    93     94    45     62    134 Female Moderate                 0        4
## 4    94     95    43     69    188 Male   Moderate                 2        0
## 5    95     96    28     67    180 Female Low                      0        0
## 6    96     97    43     69    188 Male   Moderate                 4        0
## # 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>
```

```
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")
  )
```

```r
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 te nam ne pomažu u analizi.

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

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

```r
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 centimentri i kilograme.

```r
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.

```r
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
tail(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    16   175.   59   Female Moderate            5      0      5      7 ENFJ
## 2    16   147.   45.4 Male   Moderate            0      0      0      3 ESTP
## 3    45   158.   60.8 Female Moderate            0      4      0      0 ESFJ
## 4    43   175.   85.3 Male   Moderate            2      0      0      0 ENFP
## 5    28   170.   81.6 Female Low                 0      0      0      0 ESFJ
## 6    43   175.   85.3 Male   Moderate            4      0      0      0 ENFP
## # 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)
```

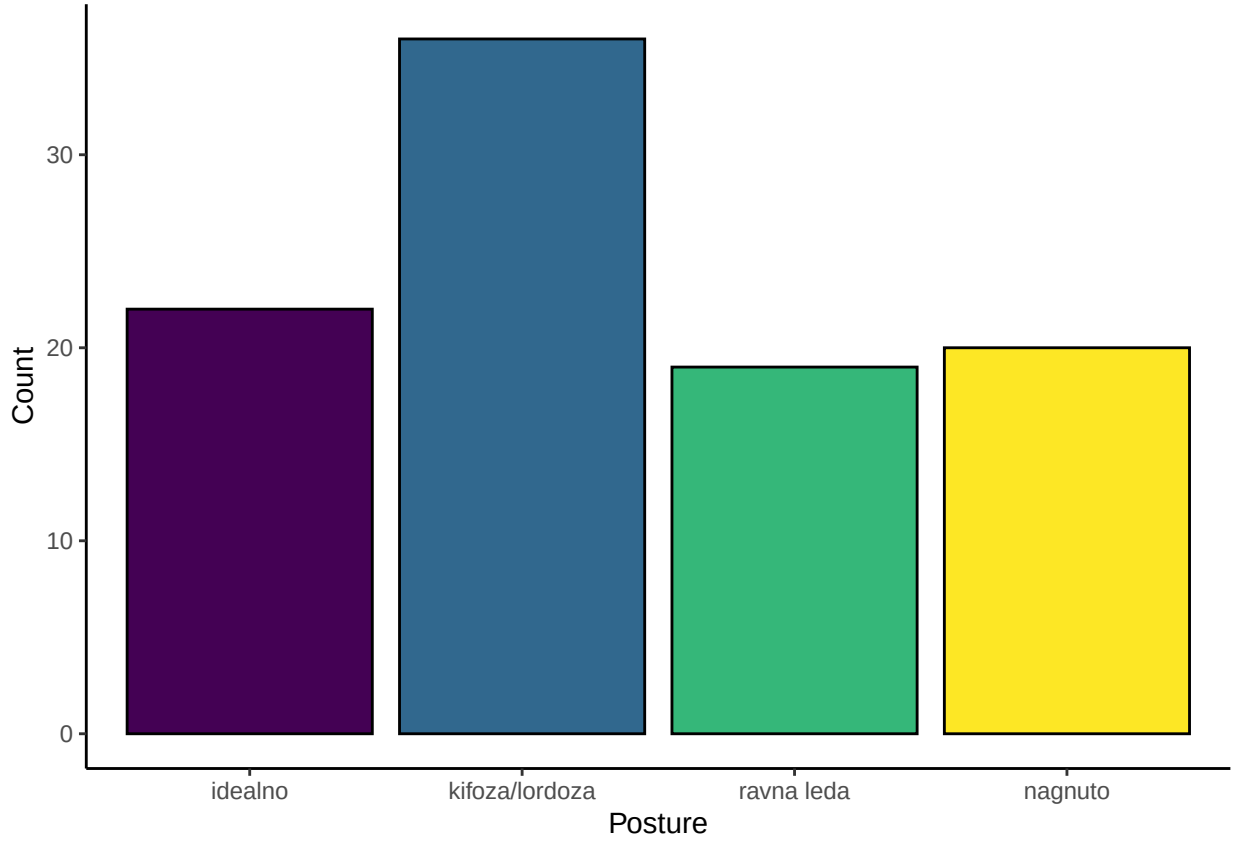
```
## Rows: 97
## Columns: 25
## $ AGE            <dbl> 53, 52, 30, 51, 45, 68, 62, 65, 66, 58, 61, 33, 48, 57,~
## $ HEIGHT         <dbl> 157.5, 175.3, 175.3, 167.6, 160.0, 188.0, 172.7, 154.9,~
## $ WEIGHT         <dbl> 56.7, 71.2, 90.7, 79.4, 90.3, 82.6, 119.3, 64.9, 81.6, ~
## $ SEX            <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

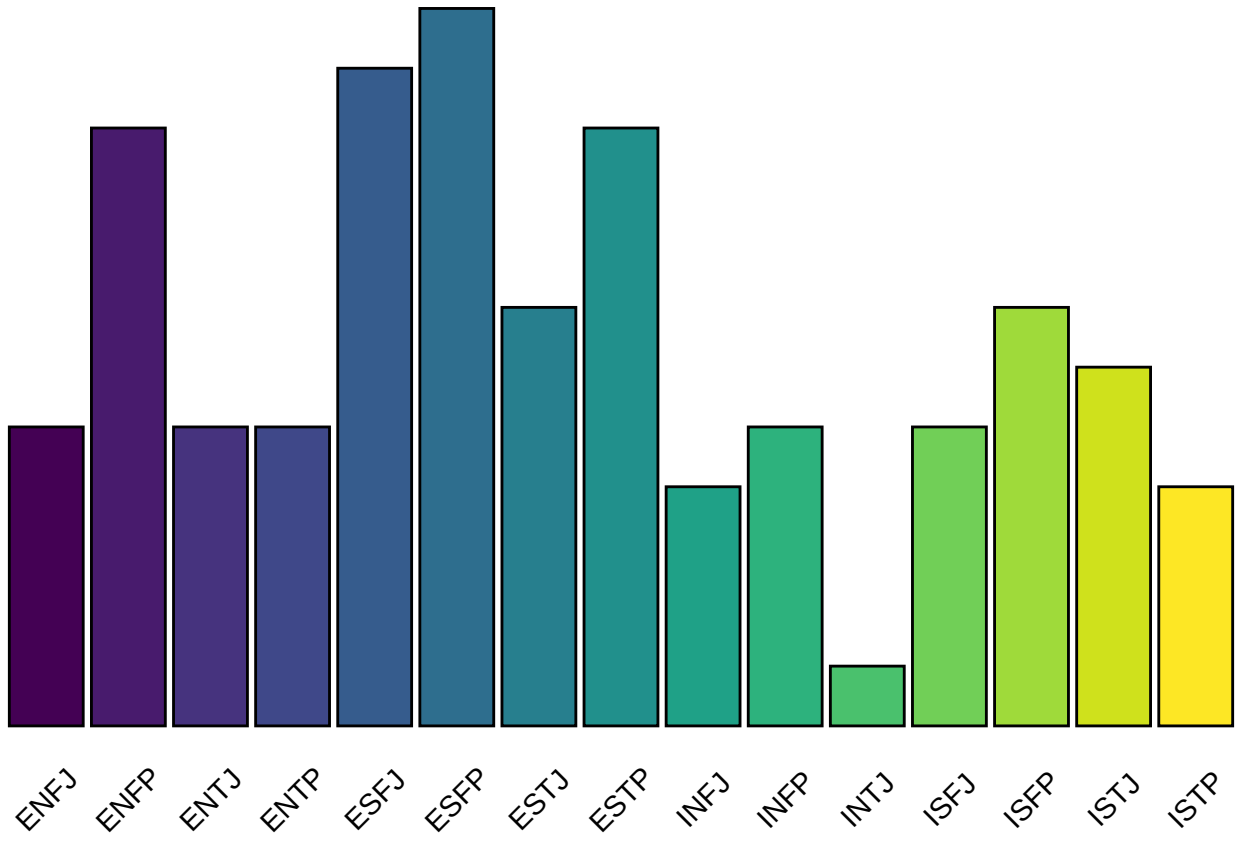
U našem podatkovnom skupu imamo stupce “POSTURE” koji predstavlja kategorije načina držanja i poprima vrijednosti “idealno”, “kifoza/lordoza”, “ravna leđa” i “nagnuto”.

```
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")
```



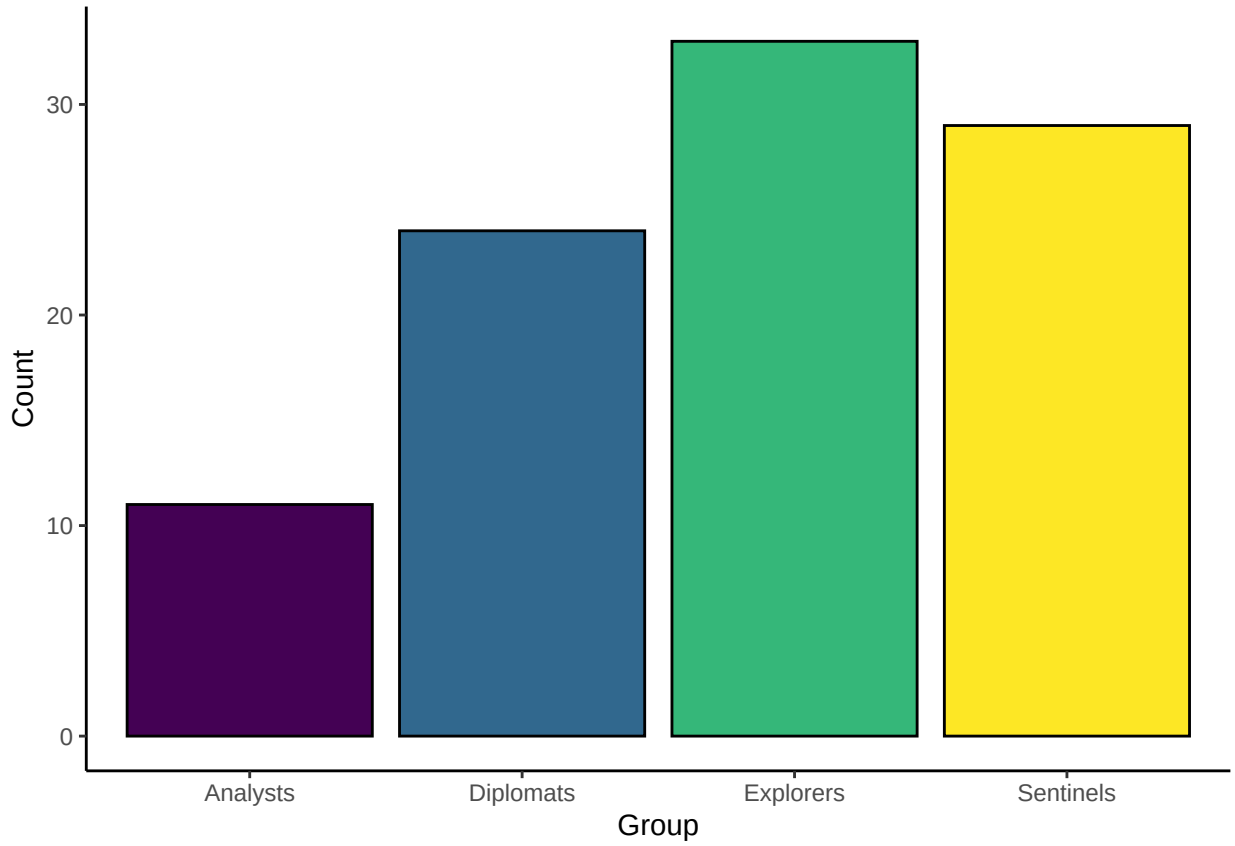
Također imamo stupac “MBTI” koji predstavlja tipove ličnosti i razlikujemo razlikujemo 16 vrsta tipova osobnosti.

```
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")
```



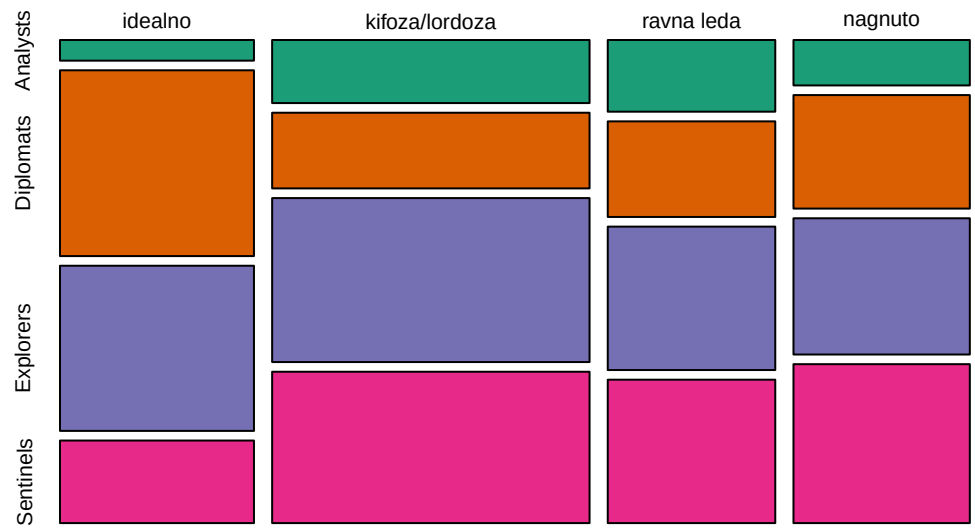
Tih 16 tipova osobnosti radi lakšeg prikaza i analize grupiramo u 4 podskupine u skladu sa Myers-Briggsovim modelom (analiziramo samo prvo slovo tipa osobnosti):

```
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")
```



Vizualiziramo podatke iz stupca "POSTURE" u odnosu na podatke iz stupca "GROUP" kako bismo vidjeli ima li veze između tipa ličnosti i načina držanja.

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



Provest ćemo hi kvadrat 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".

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

```
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

```
check_expected(contingency_table)
```

```
## [1] "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 hi kvadrat testa dobivamo iznimno maleni p-value, iz čega slijedi da odbacujemo nultu hipotezu uz razinu značajnosti od 0.2 i odbacujemo nultu hipotezu te zaključujemo da su ekstrovertnost i način držanja zavisne varijable

Isti postupak ćemo ponoviti i za raspoznavanje/intuiciju, razmišljanje/osjećanje te produđivanje/opažanje.

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)
```

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

Hi kvadrat test nam u ovom slučaju daje p-value od 0.7, iz čega slijedi da ne možemo odbaciti nultu hipotezu te zaključujemo da raspoznavanje/intuicija i način držanja ne ovise jedno o drugome.

H0: POSTURE i TF su nezavisne varijable

H1: POSTURE i TF su zavisne varijable

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

Table 3: Kontingencijska tablica za POSTURE i TF

	F	T
idealno	16	6
kifoza/lordoza	18	18
ravna leđa	13	6
nagnuto	12	8

```r
check_expected(contingency_table)
```

```
## [1] "Table meets the expected values criteria"
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 3.5441, df = 3, p-value = 0.3151
```

Na temelju dobivene vrijednosti p-value 0.32 ne možemo odbaciti nultu hipotezu te zaključujemo da razmišljanje/osjećanje i način držanja ne ovise jedno o drugome.

H0: POSTURE i JP su nezavisne varijable

H1: POSTURE i JP su zavisne varijable

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

Table 4: Kontingencijska tablica za POSTURE i JP

	J	P
idealno	5	17
kifoza/lordoza	18	18
ravna leđa	10	9
nagnuto	11	9

```r
check_expected(contingency_table)
```

```
## [1] "Table meets the expected values criteria"
```

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

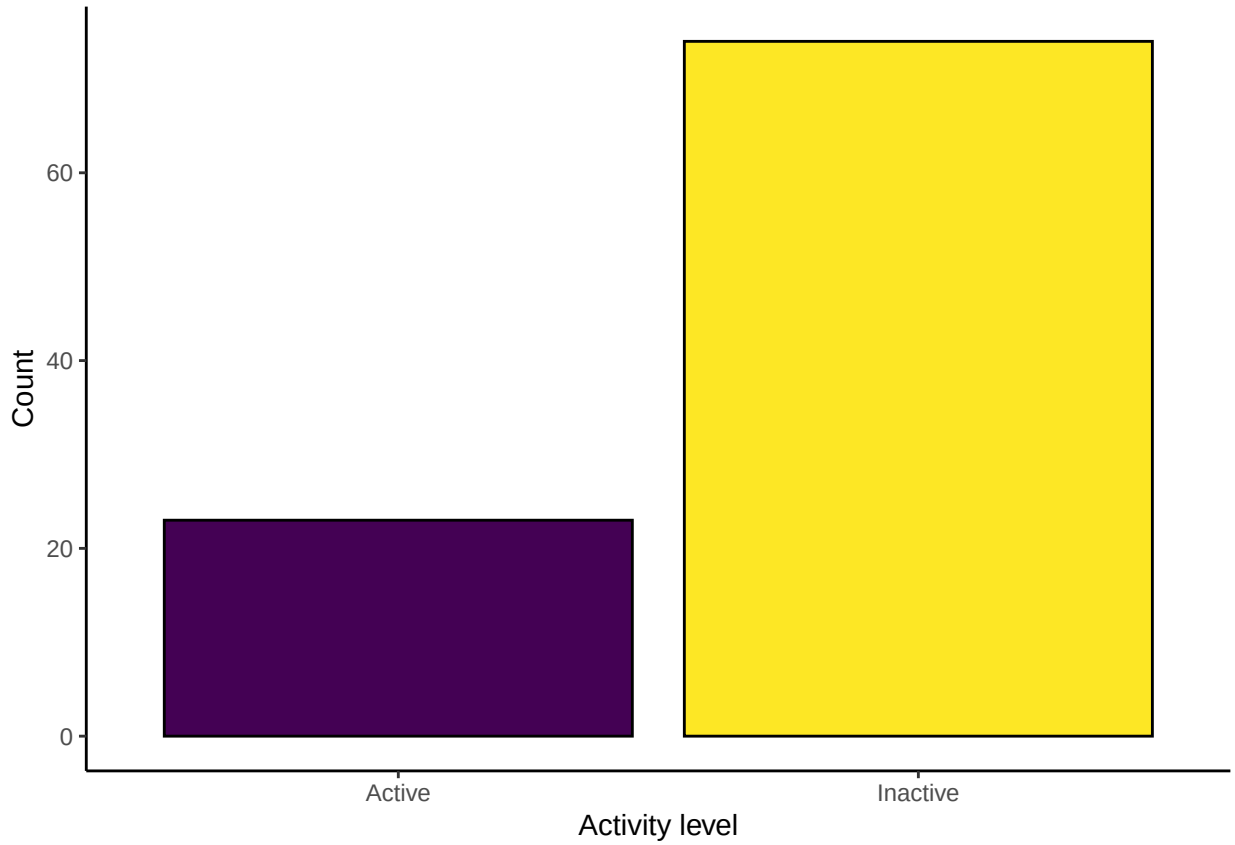
```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 6.0148, df = 3, p-value = 0.1109
```

p-value u ovom slučaju nam iznosi 0.1109 što je manje od razine značajnosti iz čega slijedi da odbacujemo nultu hipotezu te zaključujemo da produživanje/opažanje i način držanja ovise jedno o drugome.

Veza između fizičke aktivnosti i razine ekstrovertiranosti

Fizičku aktivnost nam predstavlja stupac "ACTIVITY_LEVEL":

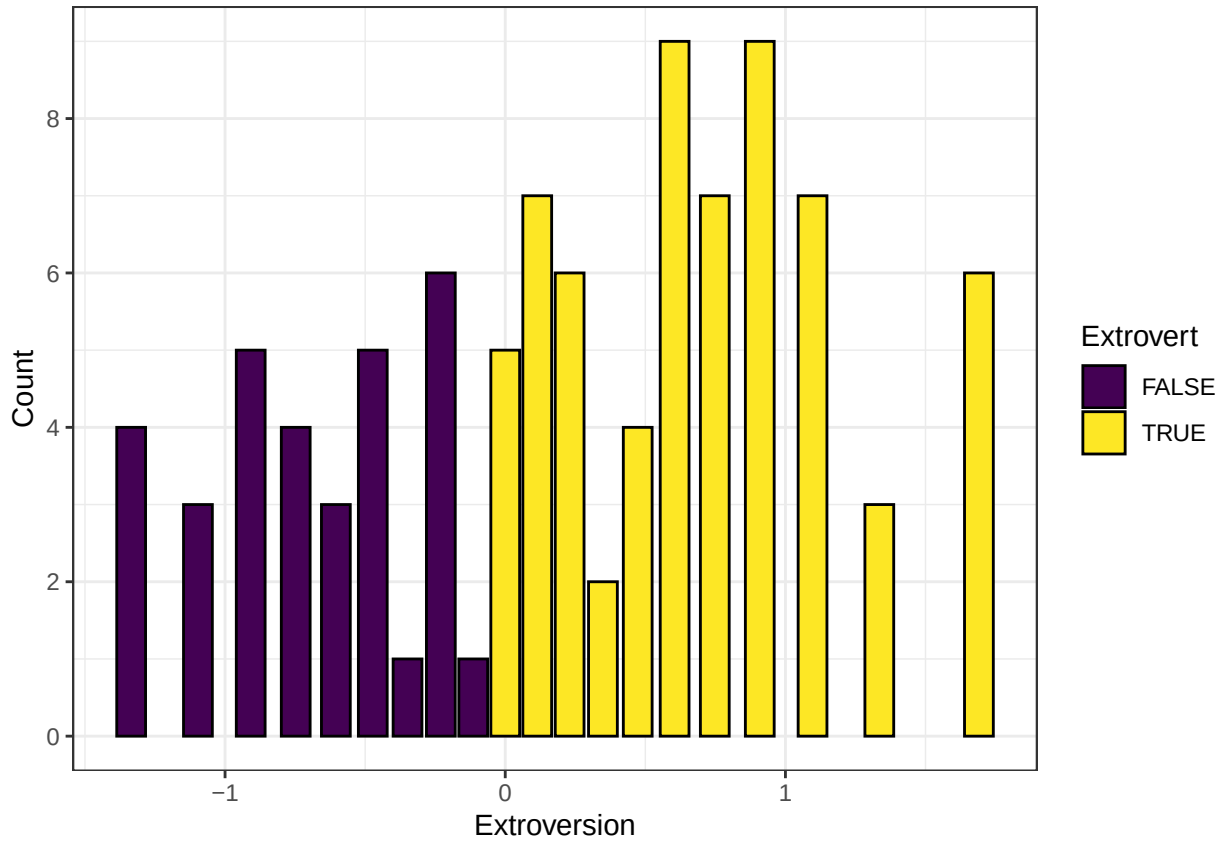
```r
ggplot(dataset, aes(x = IS_ACTIVE, fill = IS_ACTIVE)) +
  geom_bar(color = "black") +
  scale_fill_ordinal() +
  labs(x = "Activity level", fill = "Activity level", y = "Count") +
  theme_classic() +
  theme(legend.position = "none")
```



Prema ovom grafu ljude dijelimo u fizički aktivne i one neaktivne.

Za koeficijent ekstrovertnosti imamo stupac sa 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.

Provest ćemo hi kvadrat test za kategorijske podatke gdje ćemo proučavati stupce "IS_ACTIVE" (stupac koji mjeri razinu aktivnosti osobe) i "IE" (introvert/ekstrovert) podatkovnog okvira.

```
contingency_table <- table(dataset$IS_ACTIVE, dataset$IE)
contingency_table
```

```
##
##              E  I
##   Active    17  6
##   Inactive 48 26
```

```
check_expected(contingency_table)
```

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

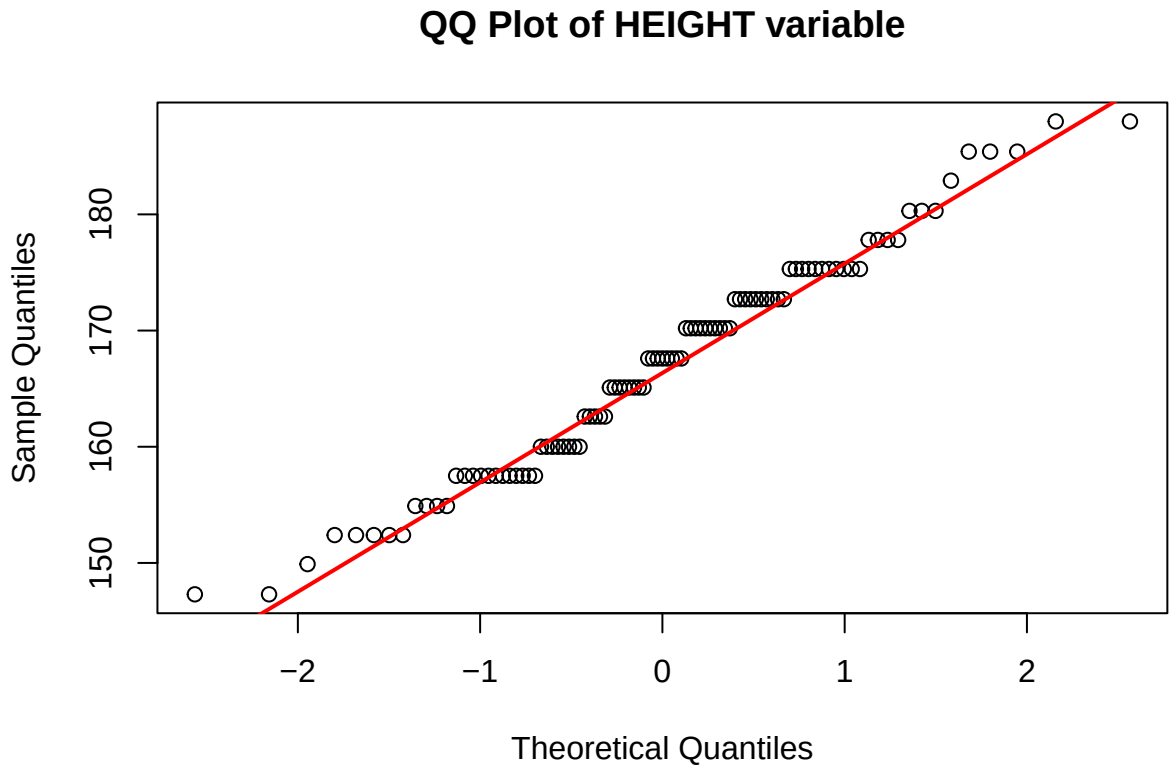
Na temelju rezultata testa zaključujemo da ne postoji veza između fizičke aktivnosti i ekstroverstnosti.

Razlika u visini/težini obzirom na tip ličnosti

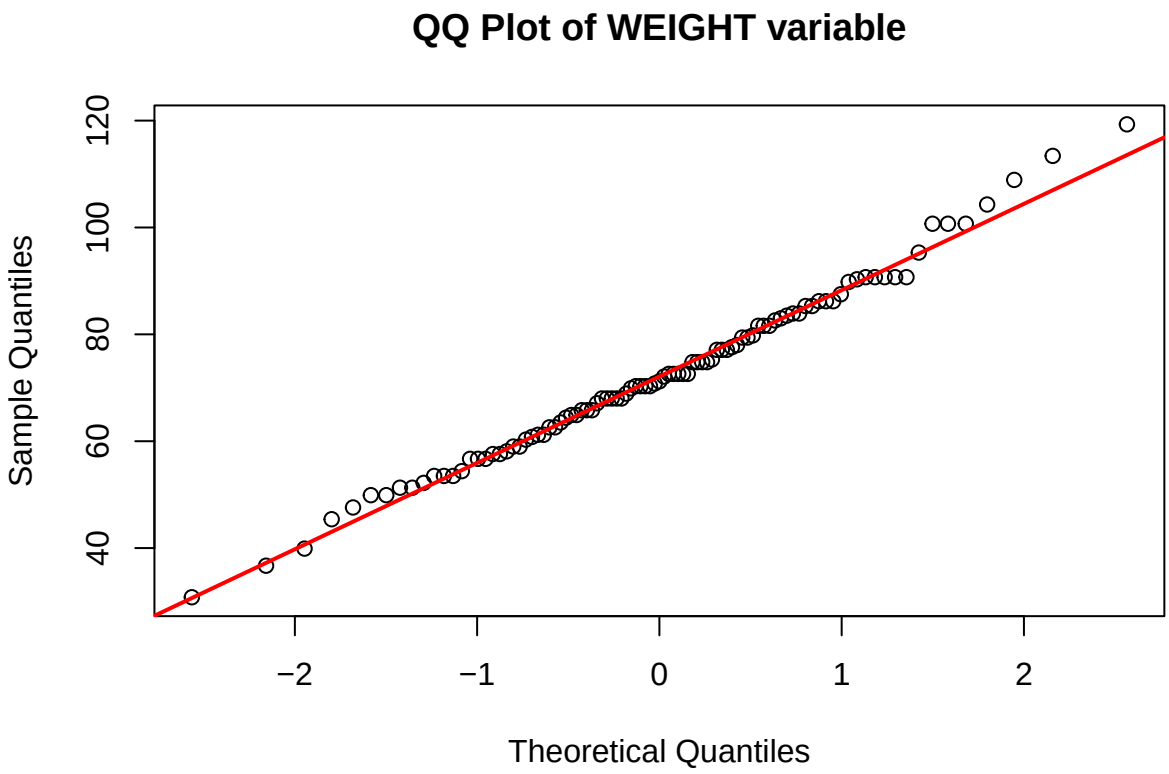
U našem podatkovnom skupu imamo stupce "WEIGHT", "HEIGHT", "GROUP" (stupac u kojem je 16 tipova osobnosti raspoređeno u 4 grupe) te stupac "IE" (u kojem je sadržan podatak o tome je li osoba introvert ili ekstrovert) koje ćemo koristiti u analiziranju visine i težine obzirom na tip ličnosti.

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)
```




```
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.

H0: Varijabla je normalno distribuirana H1: Varijabla nije normalno distribuirana

```
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 0.2, što znači da odbacujemo nultu hipotezu i zaključujemo kako varijabla HEIGHT 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, što znači da uz razinu značajnosti 0.2 ne možemo odbaciti nultu hipotezu i zaključujemo da je varijabla WEIGHT uistinu normalno distribuirana.

Iako varijabla HEIGHT nije normalna osim što njen histogram donekle nalikuje na normalnu distribuciju, pozivamo se na robusnost ANOVA testa i T-testa te ga ipak provodimo na podacima uz pretpostavku da su podaci nezavisni stoga ćemo nad rezultatima tih testova i donositi zaključke

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

H0: mi1 = mi2 = mi3 = mi4

H1: barem dva mi-i nisu jednaka.

Za hipoteze postavljene na ovaj način, nije preporučljivo koristiti t-test više puta zbog povećanja rizika od greške tipa I. Stoga, odabrali smo ANOVA (Analiza 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: varijance su jednake. H1: barem dvije varijance nisu jednake.

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

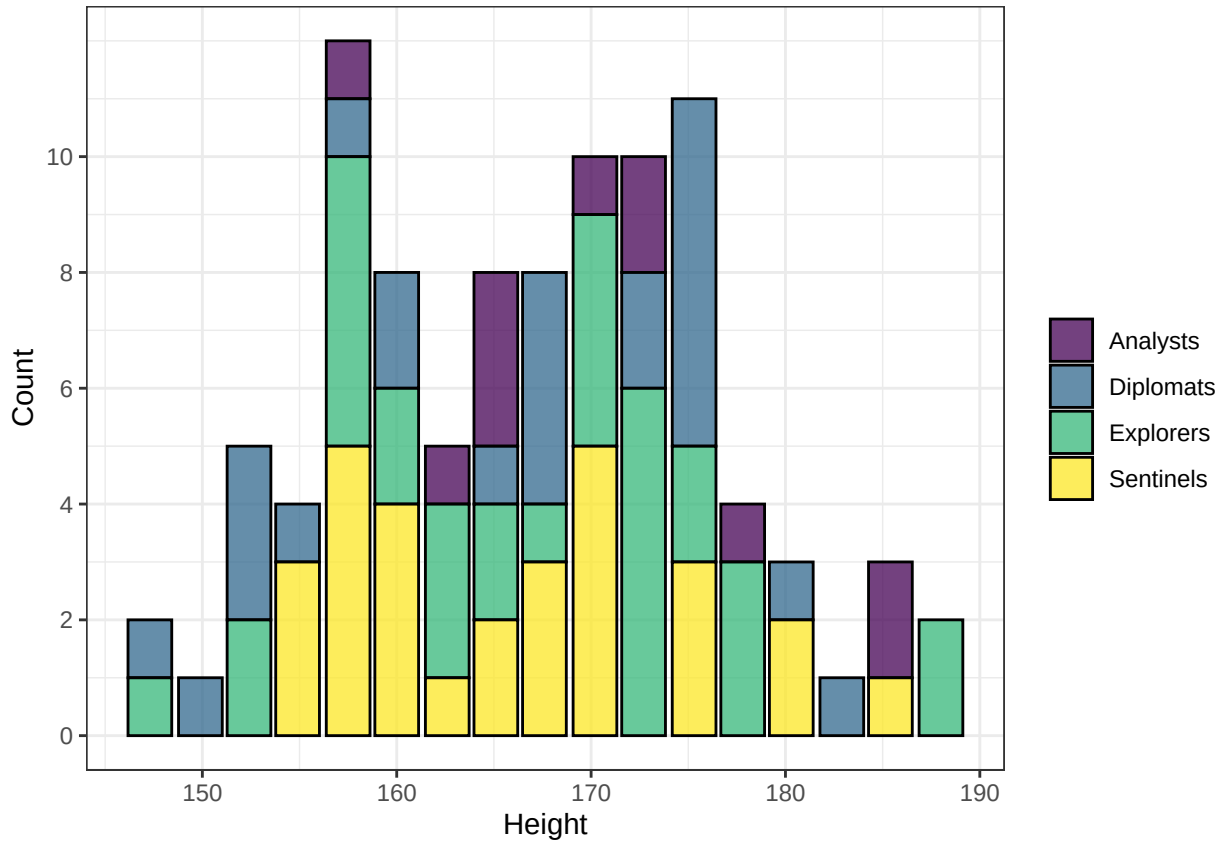
```
##
##  Bartlett test of homogeneity of variances
##
## data:  HEIGHT by GROUP
## Bartlett's K-squared = 1.184, df = 3, p-value = 0.7568
```

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

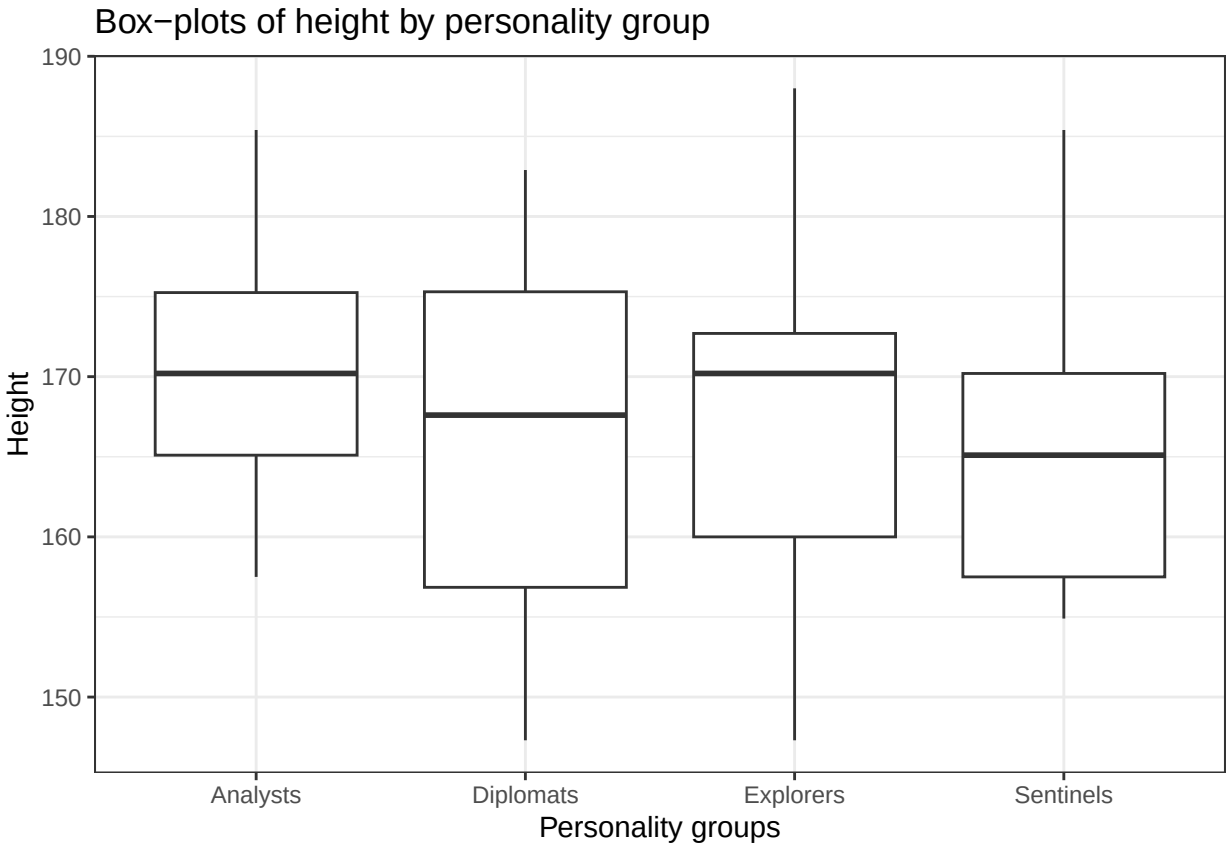
Sada smo sigurni da možemo provesti ANOVA-u.

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()
```




```
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.

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
```

Sad provedimo test:

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

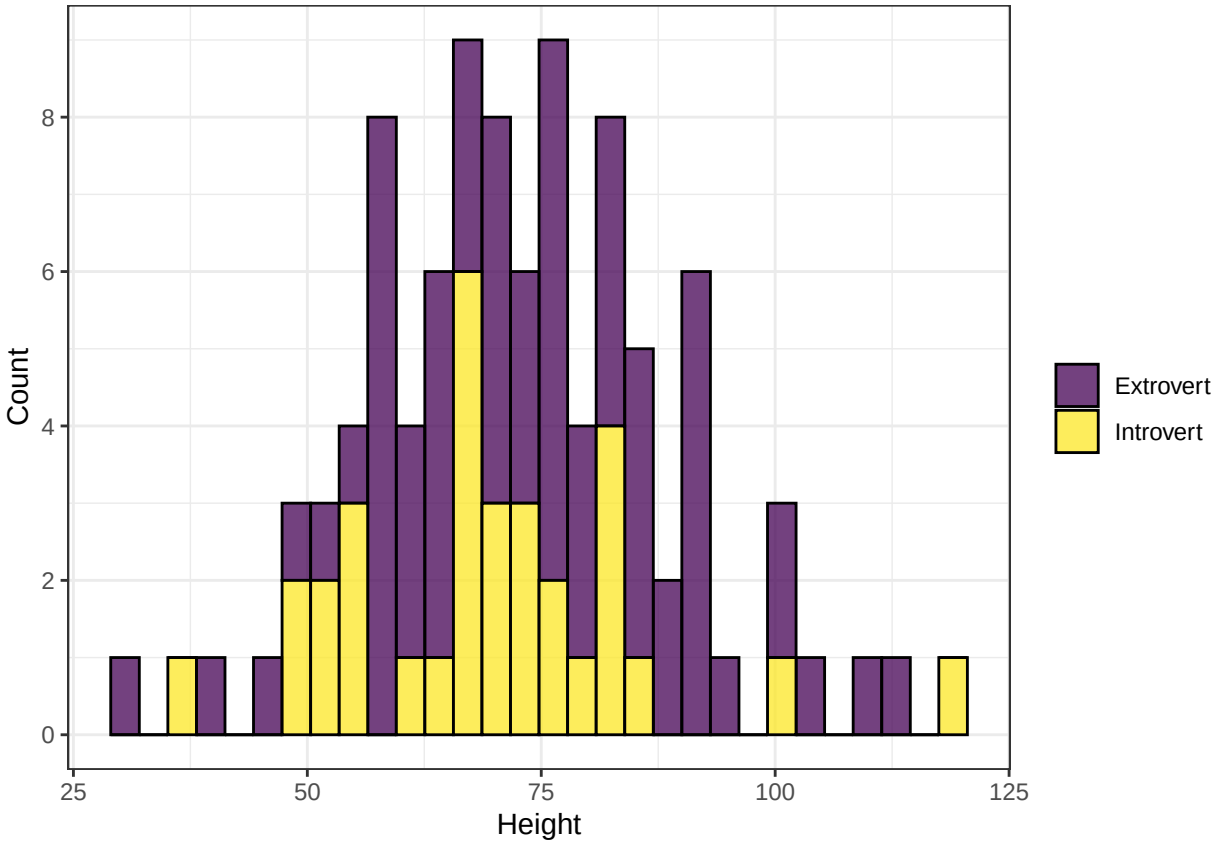
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## GROUP        3    231   77.09   0.851   0.47
## Residuals   93   8425   90.59
```

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

Na sljedećem grafu možemo vidjeti distribuciju introverata/ekstroverata 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 profesti 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 sa hipotezama:

H0: varijance su jednake

H1: varijance nisu jednake

```
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.

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

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

Test provodimo sa hipotezama:

H0: srednje vrijednosti težine jednake su za ekstroverte i introverte

H1: srednje vrijednosti težine nisu jednake za ekstroverte i introverte

```
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

U ovom zadatku pokušavamo predvidjeti tip ličnosti na temelju pojedinih karakteristika. Nismo mogli izračunati točnu vezu s MBTI tipom osobnosti jer je to diskretna varijabla, pa smo umjesto nje pokušavali 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 spada u kategoriju J ili P ovisno o tome koji je koeficijent (predviđan linearnom regresijom) veći. Najprije je bilo potrebno testirati više modela linearne regresije, tj. odabrati koje regresore koristiti u modelu. Za pojedine modele smo izračunali R^2 vrijednost, koja nam govori koliko dobro model objašnjava varijabilnost podataka. Model s najvećom R^2 vrijednosti je najbolji model. Nakon što smo odabrali najbolji model, izračunali smo koeficijente za J i P, te smo na temelju njih odredili tu kategoriju.

```
print("Prvi model za J i P karakteristike")
```

```
## [1] "Prvi model za J i P karakteristike"
```

```
model <- lm(formula = dataset$J ~ dataset$HEIGHT +
              dataset$SEX +
              dataset$AGE
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova J prvim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova J prvim modelom 0.158856531111037"
```

```
model <- lm(formula = dataset$P ~ dataset$HEIGHT +
              dataset$SEX +
              dataset$AGE
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova P prvim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova P prvim modelom 0.133730520215013"
```

```
print("Drugi model za J i P karakteristike")
```

```
## [1] "Drugi model za J i P karakteristike"
```

```
model <- lm(formula = dataset$J ~ dataset$HEIGHT +
    dataset$WEIGHT +
    dataset$SEX +
    dataset$AGE +
    dataset$ACTIVITY_LEVEL
  )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova J drugim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova J drugim modelom 0.176452683031228"
```

```
model <- lm(formula = dataset$P ~ dataset$HEIGHT +
              dataset$WEIGHT +
              dataset$SEX +
              dataset$AGE +
              dataset$ACTIVITY_LEVEL
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova P drugim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova P drugim modelom 0.151419748023853"
```

```
print("Treci model za J i P karakteristike")
```

```
## [1] "Treci model za J i P karakteristike"
```

```
model <- lm(formula = dataset$J ~ dataset$HEIGHT +
              dataset$WEIGHT +
              dataset$SEX +
              dataset$AGE +
              dataset$ACTIVITY_LEVEL +
              dataset$PAIN_1 +
              dataset$PAIN_2 +
              dataset$PAIN_3 +
              dataset$PAIN_4
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova J trecim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova J trecim modelom 0.205510806944524"
```

```
model <- lm(formula = dataset$P ~ dataset$HEIGHT +
              dataset$WEIGHT +
              dataset$SEX +
              dataset$AGE +
              dataset$ACTIVITY_LEVEL +
              dataset$PAIN_1 +
              dataset$PAIN_2 +
              dataset$PAIN_3 +
              dataset$PAIN_4
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova P trecim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova P trecim modelom 0.181734475013676"
```

Vidimo kako je najprikladniji ispao treći model jer ima najveću R kvadrat vrijednost. Testirajmo točnost modela

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

model <- lm(formula = dataset$P ~ dataset$HEIGHT +
              dataset$WEIGHT +
              dataset$SEX +
              dataset$AGE +
              dataset$ACTIVITY_LEVEL +
              dataset$PAIN_1 +
              dataset$PAIN_2 +
              dataset$PAIN_3 +
              dataset$PAIN_4
            )
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)

print(paste("Točnost", uk))
```

```
## [1] "Točnost 0.628865979381443"
```

Kod procjene spada li osoba u kategoriju J ili P, dobili smo točnost od 0.6289, što je zadovoljavajuće. Ovo je dobar rezultat s obzirom na to da je naš model bio jednostavan, tj. nije sadržavao puno regresora, a i podaci su bili dosta neujednačeni, tj. određenih osobnosti nije bilo željene količine.

Možemo ponoviti testiranje za još jednu karaktristiku, npr. razmišljanje i osjećanje

```
print("Prvi model za F i T karakteristike")
```

```
## [1] "Prvi model za F i T karakteristike"
```

```
model <- lm(formula = dataset$F ~ dataset$HEIGHT +
              dataset$SEX +
              dataset$AGE
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova F prvim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova F prvim modelom 0.16473518225415"
```

```r
model <- lm(formula = dataset$T ~ dataset$HEIGHT +
              dataset$SEX +
              dataset$AGE
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova T prvim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova T prvim modelom 0.171902851748343"
```

```r
print("Drugi model za F i T karakteristike")
```

```
## [1] "Drugi model za F i T karakteristike"
```

```r
model <- lm(formula = dataset$F ~ dataset$HEIGHT +
              dataset$WEIGHT +
              dataset$SEX +
              log(dataset$AGE) +
              dataset$ACTIVITY_LEVEL +
              dataset$PAIN_1 +
              dataset$PAIN_2 +
              dataset$PAIN_3 +
              dataset$PAIN_4
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova F drugim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova F drugim modelom 0.204905451765094"
```

```r
model <- lm(formula = dataset$T ~ dataset$HEIGHT +
              dataset$WEIGHT +
              dataset$SEX +
              log(dataset$AGE) +
              dataset$ACTIVITY_LEVEL +
              dataset$PAIN_1 +
              dataset$PAIN_2 +
              dataset$PAIN_3 +
              dataset$PAIN_4
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova T drugim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova T drugim modelom 0.20948149065612"
```

```r
print("Treci model za F i T karakteristike")
```

```
## [1] "Treci model za F i T karakteristike"
```

```
model <- lm(formula = dataset$F ~ dataset$HEIGHT +
              dataset$WEIGHT +
              dataset$SEX +
              dataset$AGE +
              dataset$ACTIVITY_LEVEL +
              dataset$PAIN_1 +
              dataset$PAIN_2 +
              dataset$PAIN_3 +
              dataset$PAIN_4
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova F trecim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova F trecim modelom 0.203402187853853"
```

```
model <- lm(formula = dataset$T ~ dataset$HEIGHT +
              dataset$WEIGHT +
              dataset$SEX +
              dataset$AGE +
              dataset$ACTIVITY_LEVEL +
              dataset$PAIN_1 +
              dataset$PAIN_2 +
              dataset$PAIN_3 +
              dataset$PAIN_4
            )
model_summary <- summary(model)
r_squared <- model_summary$r.squared
print(paste("R kvadrat za predikciju slova T trecim modelom", r_squared))
```

```
## [1] "R kvadrat za predikciju slova T trecim modelom 0.207816499112624"
```

Drugi model se pokazuje kao najbolji, testiramo

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

model <- lm(formula = dataset$T ~ dataset$HEIGHT +
              dataset$WEIGHT +
              dataset$SEX +
              log(dataset$AGE) +
              dataset$ACTIVITY_LEVEL +
```

```
              dataset$PAIN_1 +
              dataset$PAIN_2 +
              dataset$PAIN_3 +
              dataset$PAIN_4)

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)

print(paste("Točnost", uk))
```

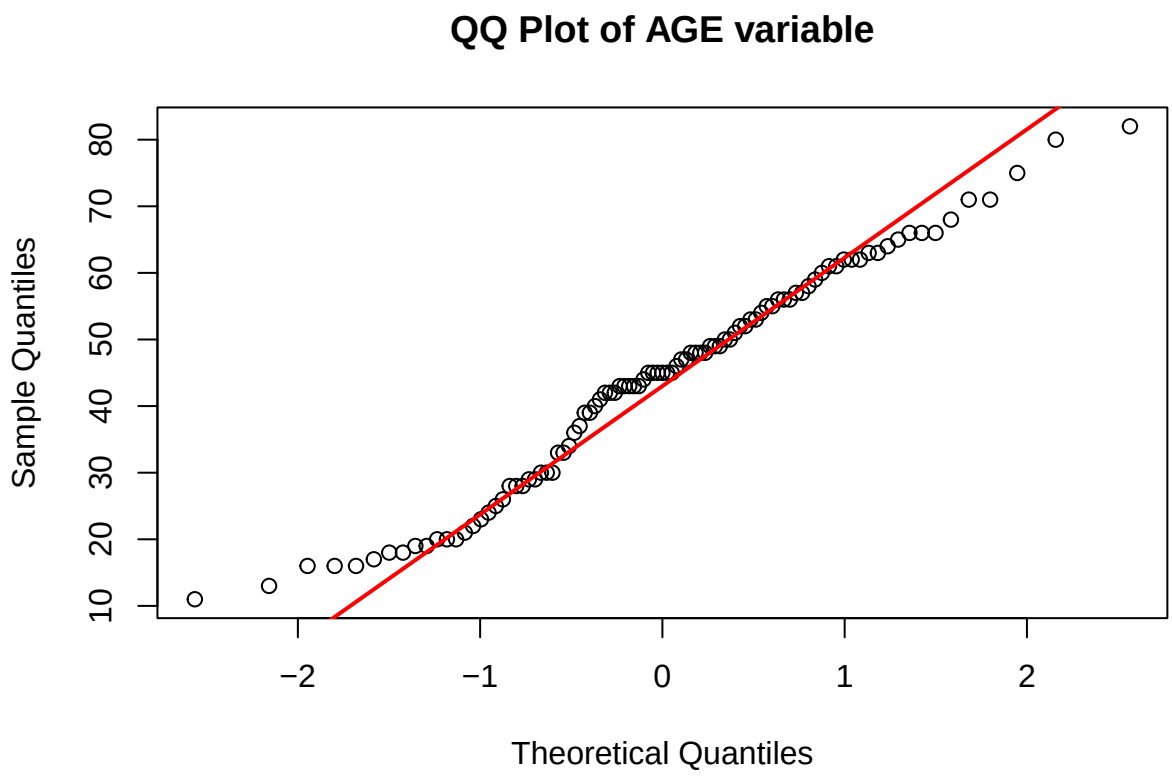
```
## [1] "Točnost 0.701030927835051"
```

Dobivena je točnost 0.7010

Je li postotak 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 0.2 odbacujemo nultu hipotezu i zaključujemo da varijabla AGE nije normalno distribuirana.

Iako varijabla AGE nije normalno distribuirana, poznato nam je da su z-test i t-test robusni na normalnost podataka, pa ćemo ih koristiti za testiranje. 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 za dvije proporcije, uz alfa=0.05, a hipoteze postavljamo ovako H0 : postotak ekstroverata je isti neovisno o godinama (p1 - p2 = 0) H1 : postotak ekstroverata se razlikuje ovisno o godinama (p1 - p2 != 0)

```
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)

# Perform the z-test for proportions
z_test_result <- prop.test(c(Eyoung, Eold),
                           c(total_group1, total_group2),
                           alternative = "two.sided")

# Print the result
print(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
```

Iz polja p-value koji je veći od 0.2 zaključujemo kako na temelju ovih podataka nemamo razloga sumnjati u H0 te ju s toga ne odbacujemo. ***

Zaključak

(—prazan prostor za zaključak—)