[SAP] Projektni zadatak - Analiza podataka zdravstvenog pregleda

Statistički najizglednije ime

2025-01-25

#Inicijalni koraci projekta ## Učitavanje podataka

##Filtriranje podataka

Kao prvi korak analize provodi se filtriranje skupa podataka kako bi izbacili besmislene podatke. Izbacuju se podaci gdje su iznosi tlaka nerealni. Također se izbacuju podaci za tlak gdje je izmjereni maksimalni tlak manji od minimalnog. Izbacujemo BMI podatke koji su veći od najvećeg izmjerenog BMI u povijesti.

```
# Filtriranje besmislenih podataka iz tablice
filtered_data <- healthDATA.modif %>% filter(ap_hi <= 370) %>% filter(ap_lo <= 360) %>% filter(ap_hi >=

filtered_data <- filtered_data %>%
  mutate(
    cholesterol = as.factor(cholesterol),
    gender = as.factor(gender),
    AgeGroup = as.factor(AgeGroup)
  ) %>% mutate (AgeGroup = fct_relevel(AgeGroup, "20-40", "40-60", ">60"))

summary(filtered_data)
head(filtered_data)
```

```
##       ...1             id             age        gender        height
##  Min.   :    0   Min.   :    0   Min.   :10798   1:44759   Min.   : 57.0
##  1st Qu.:17497   1st Qu.:25002   1st Qu.:17657   2:23959   1st Qu.:159.0
##  Median :35010   Median :50015   Median :19701             Median :165.0
##  Mean   :35001   Mean   :49975   Mean   :19464             Mean   :164.4
##  3rd Qu.:52485   3rd Qu.:74866   3rd Qu.:21324             3rd Qu.:170.0
##  Max.   :69999   Max.   :99999   Max.   :23713             Max.   :250.0
##      weight           ap_hi           ap_lo        cholesterol      gluc
##  Min.   : 11.00   Min.   : 60.0   Min.   :  0.00   1:51538     Min.   :1.000
##  1st Qu.: 65.00   1st Qu.:120.0   1st Qu.: 80.00   2: 9306     1st Qu.:1.000
##  Median : 72.00   Median :120.0   Median : 80.00   3: 7874     Median :1.000
##  Mean   : 74.11   Mean   :126.7   Mean   : 81.26               Mean   :1.226
##  3rd Qu.: 82.00   3rd Qu.:140.0   3rd Qu.: 90.00               3rd Qu.:1.000
##  Max.   :200.00   Max.   :240.0   Max.   :182.00               Max.   :3.000
##      smoke              alco             active           cardio
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.08792   Mean   :0.05335   Mean   :0.8034   Mean   :0.4947
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##     AgeinYr           BMI             BMICat           AgeGroup
##  Min.   :29.00   Min.   :  3.50   Length:68718       20-40: 3370
```

```
##  1st Qu.:48.00   1st Qu.: 23.90   Class :character   40-60:55715
##  Median :53.00   Median : 26.30   Mode  :character   >60  : 9633
##  Mean   :52.83   Mean   : 27.51
##  3rd Qu.:58.00   3rd Qu.: 30.10
##  Max.   :64.00   Max.   :237.80
## # A tibble: 6 x 18
##    ...1    id   age gender height weight ap_hi ap_lo cholesterol  gluc smoke
##   <dbl> <dbl> <dbl> <fct>   <dbl>  <dbl> <dbl> <dbl> <fct>       <dbl> <dbl>
## 1     0     0 18393 2         168     62   110    80 1               1     0
## 2     1     1 20228 1         156     85   140    90 3               1     0
## 3     2     2 18857 1         165     64   130    70 3               1     0
## 4     3     3 17623 2         169     82   150   100 1               1     0
## 5     4     4 17474 1         156     56   100    60 1               1     0
## 6     5     8 21914 1         151     67   120    80 2               2     0
## # i 7 more variables: alco <dbl>, active <dbl>, cardio <dbl>, AgeinYr <dbl>,
## #   BMI <dbl>, BMICat <chr>, AgeGroup <fct>
```

```
#Analiza skupa podataka
filtered_data <- filtered_data %>%
  mutate(gender = factor(gender,
                         levels = c(1, 2),
                         labels = c("Female", "Male")))

average_weight <- filtered_data %>%
  group_by(gender) %>%
  summarise(avg_weight = mean(weight, na.rm = TRUE))

average_weight

average_height <- filtered_data %>%
  group_by(gender) %>%
  summarise(avg_height = mean(height, na.rm = TRUE))

average_height

str(filtered_data)
head(filtered_data)
```

```
## # A tibble: 2 x 2
##   gender avg_weight
##   <fct>       <dbl>
## 1 Female       72.5
## 2 Male         77.2
## # A tibble: 2 x 2
##   gender avg_height
##   <fct>       <dbl>
## 1 Female       161.
## 2 Male         170.
## tibble [68,718 x 18] (S3: tbl_df/tbl/data.frame)
##  $ ...1       : num [1:68718] 0 1 2 3 4 5 6 7 8 9 ...
##  $ id         : num [1:68718] 0 1 2 3 4 8 9 12 13 14 ...
##  $ age        : num [1:68718] 18393 20228 18857 17623 17474 ...
##  $ gender     : Factor w/ 2 levels "Female","Male": 2 1 1 2 1 1 1 2 1 1 ...
##  $ height     : num [1:68718] 168 156 165 169 156 151 157 178 158 164 ...
##  $ weight     : num [1:68718] 62 85 64 82 56 67 93 95 71 68 ...
```

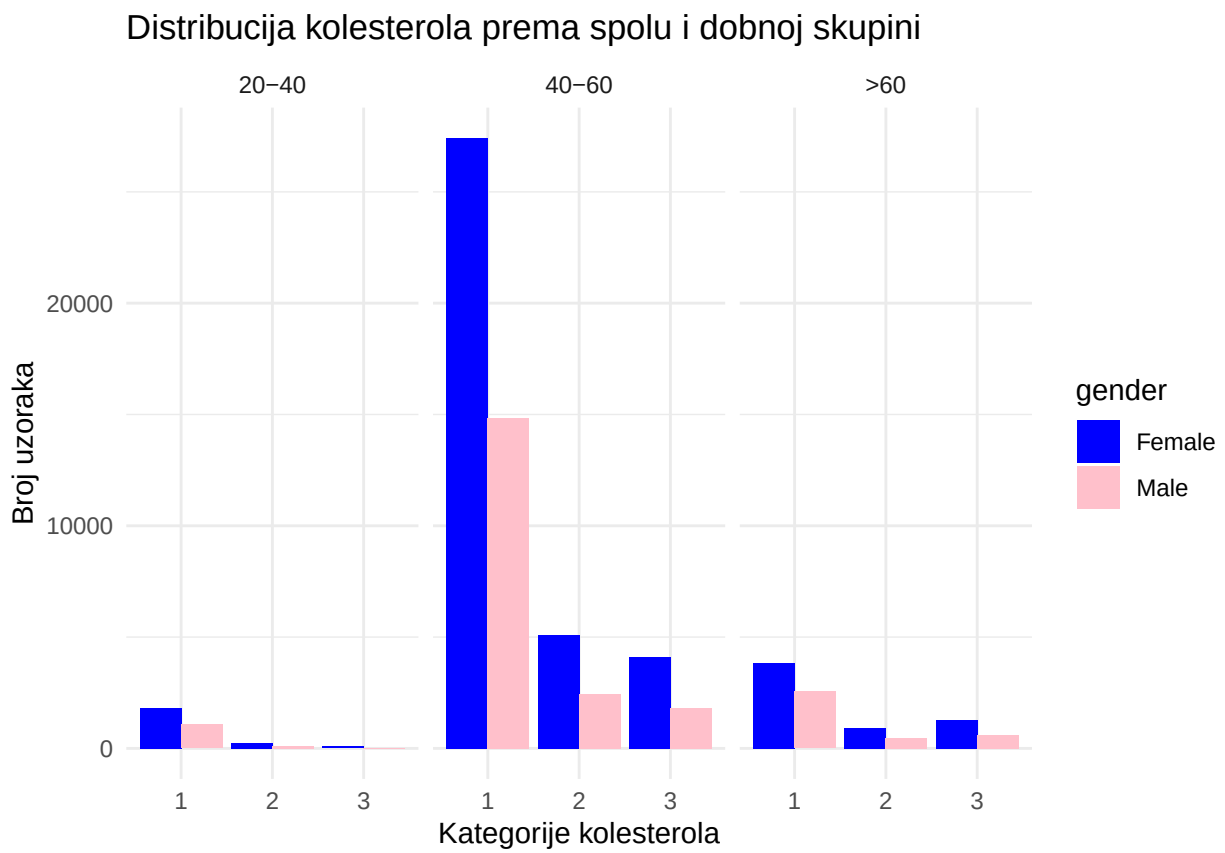
```
##  $ ap_hi      : num [1:68718] 110 140 130 150 100 120 130 130 110 110 ...
##  $ ap_lo      : num [1:68718] 80 90 70 100 60 80 80 90 70 60 ...
##  $ cholesterol: Factor w/ 3 levels "1","2","3": 1 3 3 1 1 2 3 3 1 1 ...
##  $ gluc       : num [1:68718] 1 1 1 1 1 2 1 3 1 1 ...
##  $ smoke      : num [1:68718] 0 0 0 0 0 0 0 0 0 0 ...
##  $ alco       : num [1:68718] 0 0 0 0 0 0 0 0 0 0 ...
##  $ active     : num [1:68718] 1 1 0 1 0 0 1 1 1 0 ...
##  $ cardio     : num [1:68718] 0 1 1 1 0 0 0 1 0 0 ...
##  $ AgeinYr    : num [1:68718] 50 55 51 48 47 60 60 61 48 54 ...
##  $ BMI        : num [1:68718] 22 34.9 23.5 28.7 23 29.4 37.7 30 28.4 25.3 ...
##  $ BMICat     : chr [1:68718] "Normal" "Obese" "Normal" "Over Weight" ...
##  $ AgeGroup   : Factor w/ 3 levels "20-40","40-60",..: 2 2 2 2 2 2 2 3 2 2 ...
## # A tibble: 6 x 18
##    ...1    id   age gender height weight ap_hi ap_lo cholesterol  gluc smoke
##   <dbl> <dbl> <dbl> <fct>   <dbl>  <dbl> <dbl> <dbl> <fct>       <dbl> <dbl>
## 1     0     0 18393 Male      168     62   110    80 1               1     0
## 2     1     1 20228 Female    156     85   140    90 3               1     0
## 3     2     2 18857 Female    165     64   130    70 3               1     0
## 4     3     3 17623 Male      169     82   150   100 1               1     0
## 5     4     4 17474 Female    156     56   100    60 1               1     0
## 6     5     8 21914 Female    151     67   120    80 2               2     0
## # i 7 more variables: alco <dbl>, active <dbl>, cardio <dbl>, AgeinYr <dbl>,
## #   BMI <dbl>, BMICat <chr>, AgeGroup <fct>
```

#Zadatak 1: ##Kakva je distribucija razina kolesterola među različitim dobnim skupinama i spolovima?

```
# Zadatak 1

distribution <- filtered_data %>%
  group_by(gender, cholesterol) %>%
  summarise(count = n(),
            percentage = n() / nrow(filtered_data) * 100)
```

```
## `summarise()` has grouped output by 'gender'. You can override using the
## `.groups` argument.
```

```
distribution

ggplot(filtered_data, aes(x = cholesterol, fill = gender)) +
  geom_bar(position = "dodge") +
  facet_wrap(~ AgeGroup) +
  labs(title = "Distribucija kolesterola prema spolu i dobnoj skupini",
       x = "Kategorije kolesterola",
       y = "Broj uzoraka") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "pink"))
```




```
## # A tibble: 6 x 4
## # Groups:   gender [2]
##   gender cholesterol count percentage
##   <fct>  <fct>       <int>      <dbl>
## 1 Female 1           33059      48.1
## 2 Female 2            6260       9.11
## 3 Female 3            5440       7.92
## 4 Male   1           18479      26.9
## 5 Male   2            3046       4.43
## 6 Male   3            2434       3.54
```

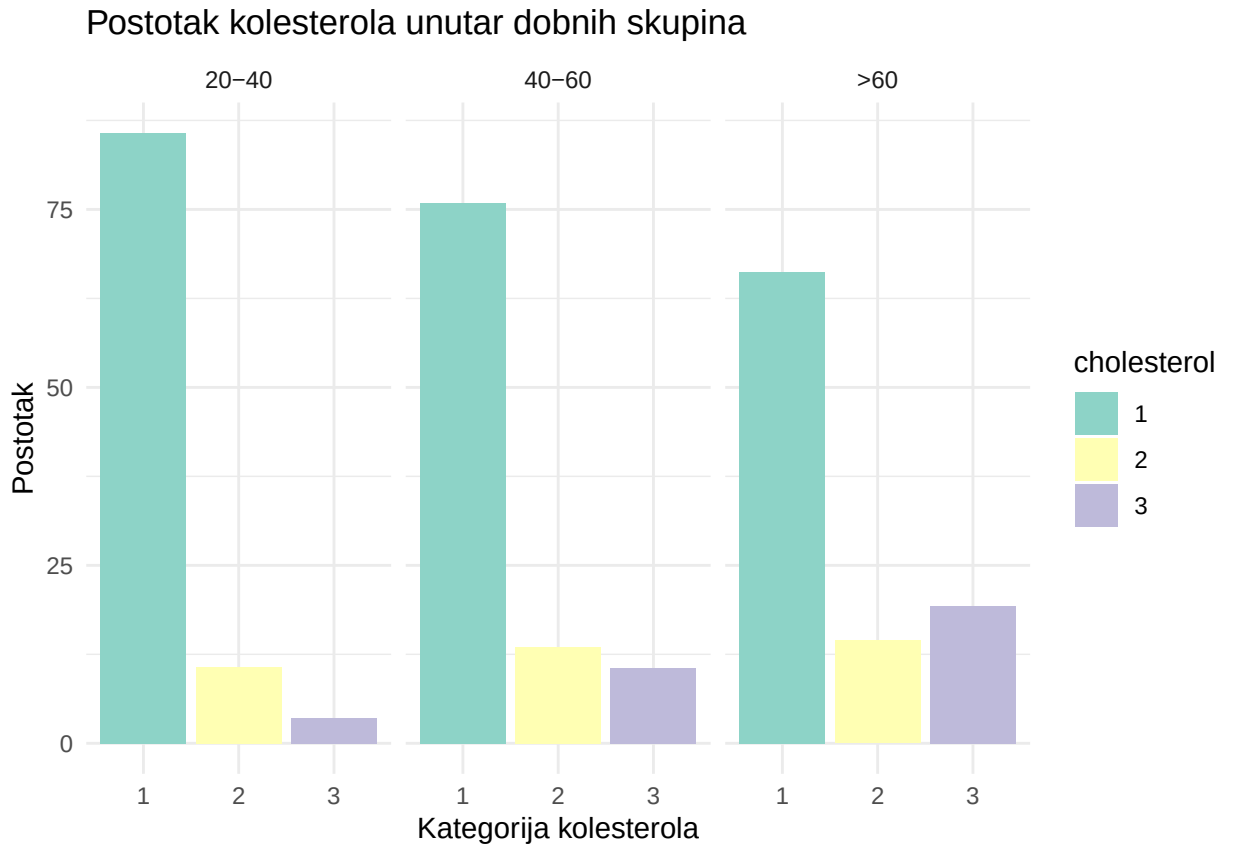
Sljedeći grafovi prikazuju raspodjelu postotka pripadnosti trima kategorijama kolesterola (zdrav, rizičan i opasan). Svaka skupina kolesterola (označena bojama) predstavlja određeni zdravstveni status:

```
Zdrav (1) - prikazan zelenkastom bojom,
Rizičan (2) - prikazan žutom bojom,
Opasan (3) - prikazan ljubičastom bojom.
```

Prvi graf prikazuje raspodjelu unutar tri dobne skupine: 20-40 godina, 40-60 godina i iznad 60 godina.

```r
cholesterol_age <- filtered_data %>%
  group_by(AgeGroup, cholesterol) %>%
  summarise(count = n(), .groups = 'drop') %>%
  group_by(AgeGroup) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(cholesterol_age, aes(x = cholesterol, y = percentage, fill = cholesterol)) +
  geom_bar(stat = "identity", position = "dodge") +
```

```
  facet_wrap(~ AgeGroup) +
  labs(title = "Postotak kolesterola unutar dobnih skupina",
       x = "Kategorija kolesterola",
       y = "Postotak") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3")
```

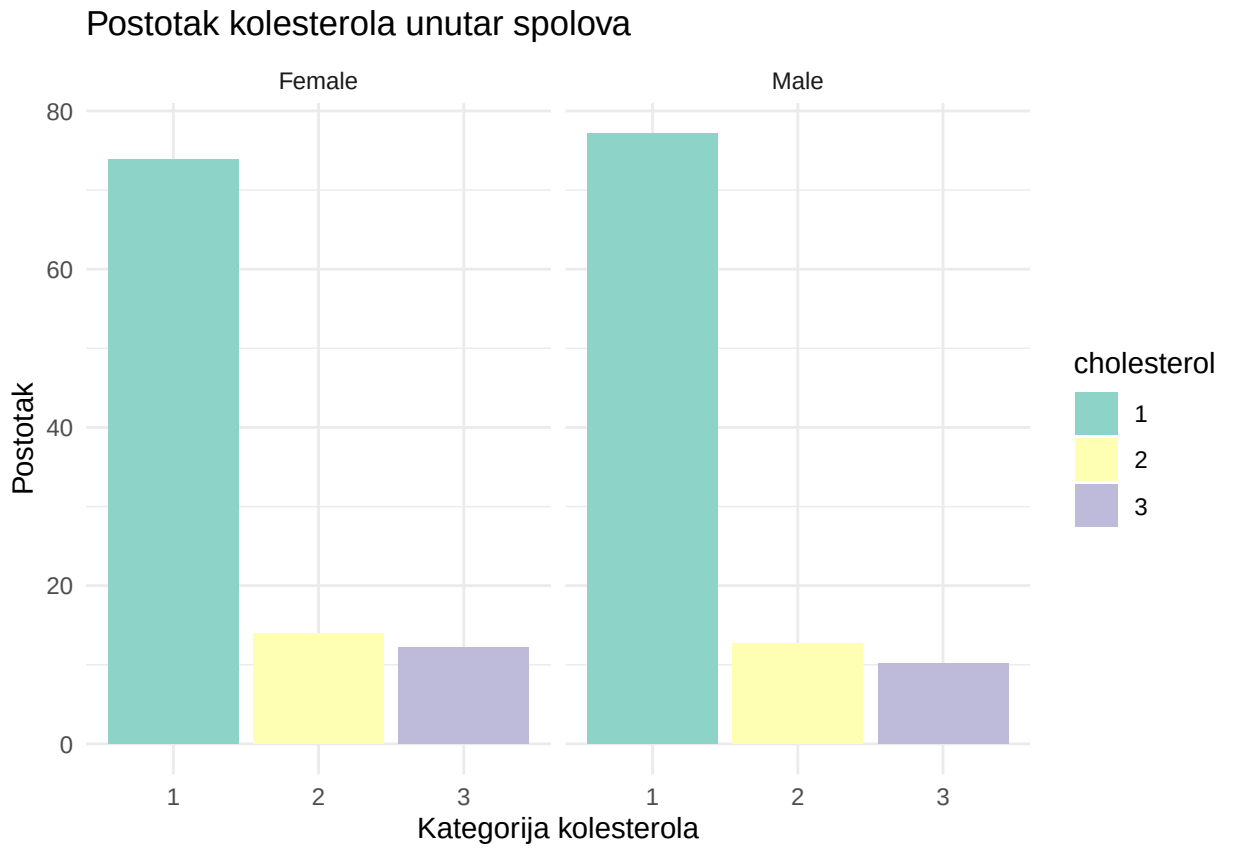


Zaključci sa grafa: - Zdrava kategorija dominira u svakoj dobnoj skupini. - Dobna skupina 20-40 ima najzdravije razine kolesterola, sa najmanjom zastupljenosti rizične ili opasne kategorije. - Dobna skupina iznad 60 godina, iako još uvijek teži zdravoj kategoriji kolesterola, ima češću zastupljenost rizične ili opasne kategorije od ostalih grupa. Također je jedina grupa gdje je opasna kategorija češća od rizične.

Sljedeći graf prikazuje raspodjelu unutar spolova.

```
cholesterol_gender <- filtered_data %>%
  group_by(gender, cholesterol) %>%
  summarise(count = n(), .groups = 'drop') %>%
  group_by(gender) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(cholesterol_gender, aes(x = cholesterol, y = percentage, fill = cholesterol)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ gender) +
  labs(title = "Postotak kolesterola unutar spolova",
       x = "Kategorija kolesterola",
       y = "Postotak") +
  theme_minimal() +
```

```
  scale_fill_brewer(palette = "Set3")
```



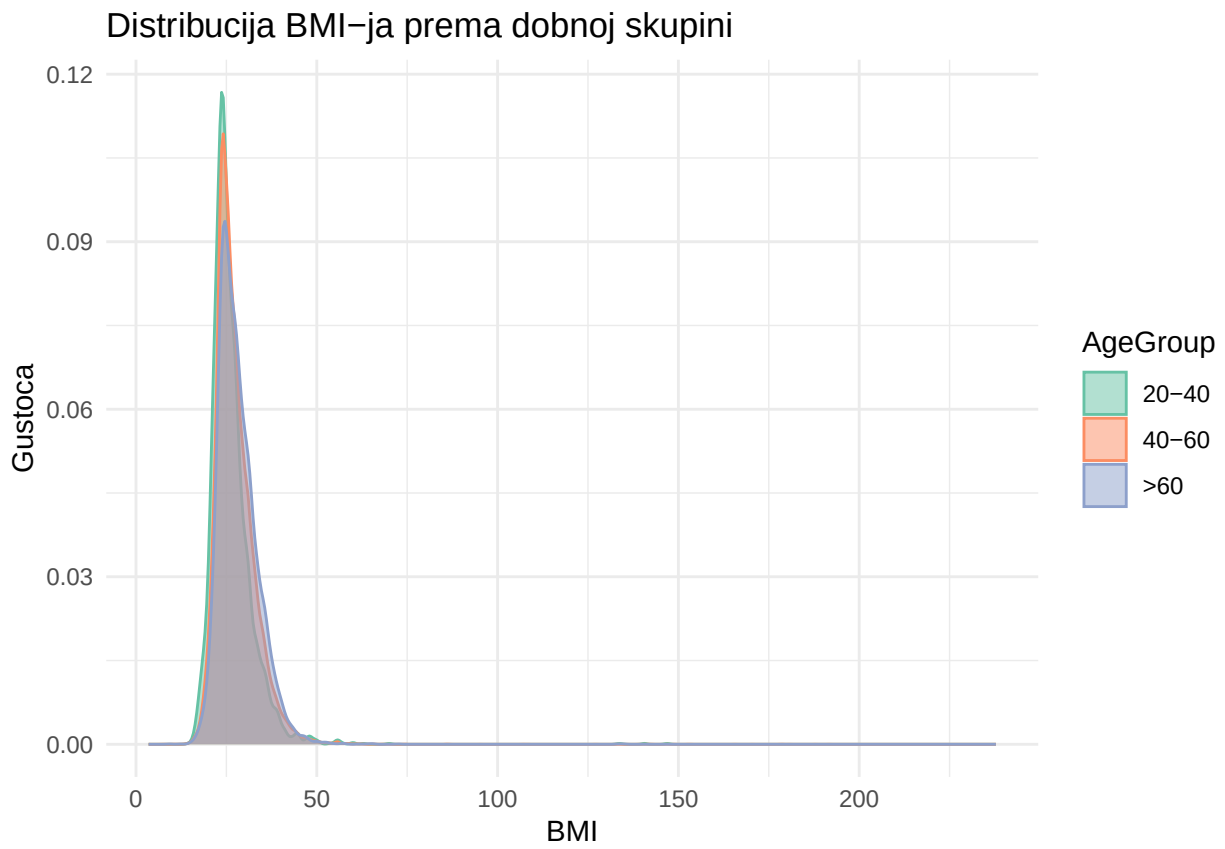
Zaključci sa grafa: - Zdrava kategorija dominira u svakom spolu. - Ne pokazuju se značajne razlike u postotnoj distribuciji kategorija kolesterola.

##Prikaz dodatnih distribucija

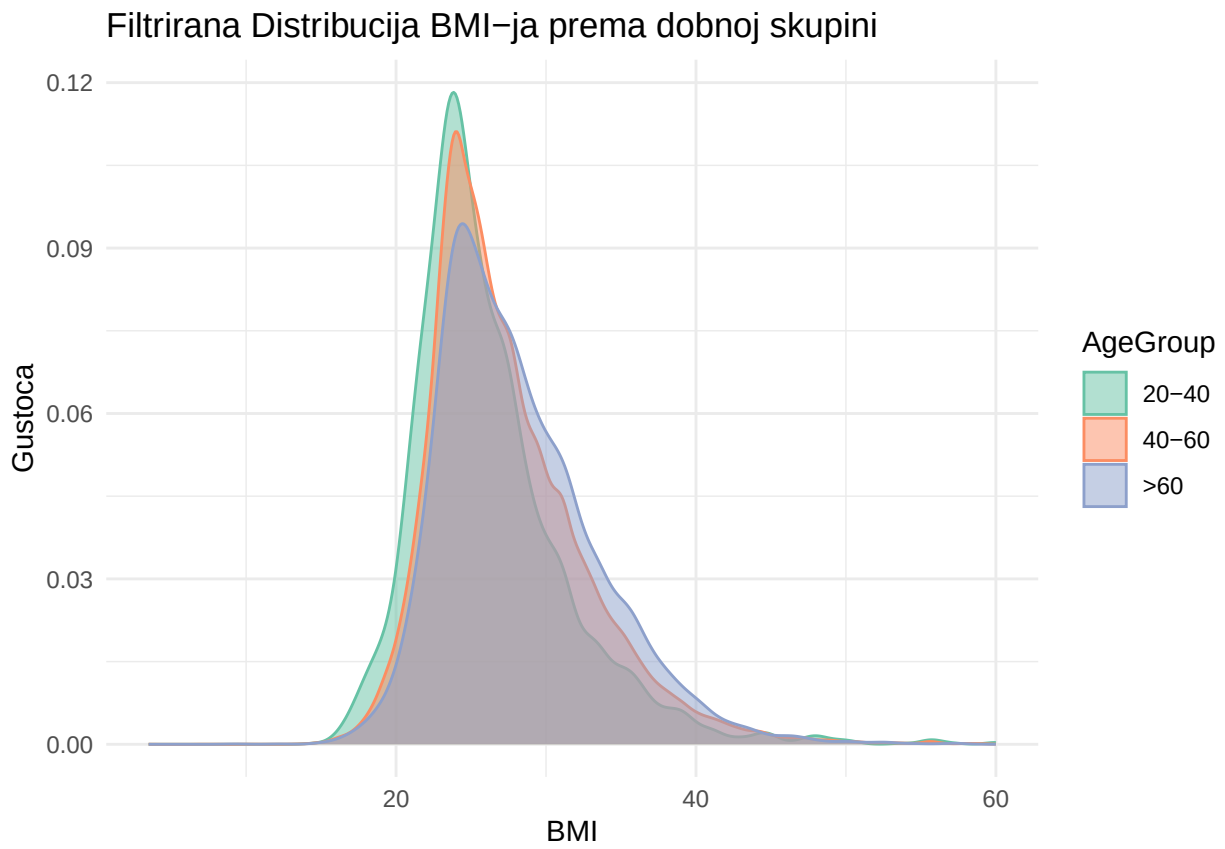
Sljedeći grafovi prikazuju distribucija BMI-ja. Filtrirani grafovi služe kako bi lakše vidjeli tendenciju distribucije, fokusiranjem na češće iznose.

Prvi grafovi prikazuju distribuciju BMI-ja po dobnim skupinama.

```
BMI_filtered_data <- filtered_data %>% filter(BMI <= 60)

ggplot(filtered_data, aes(x = BMI, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribucija BMI-ja prema dobnoj skupini",
       x = "BMI",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")
```



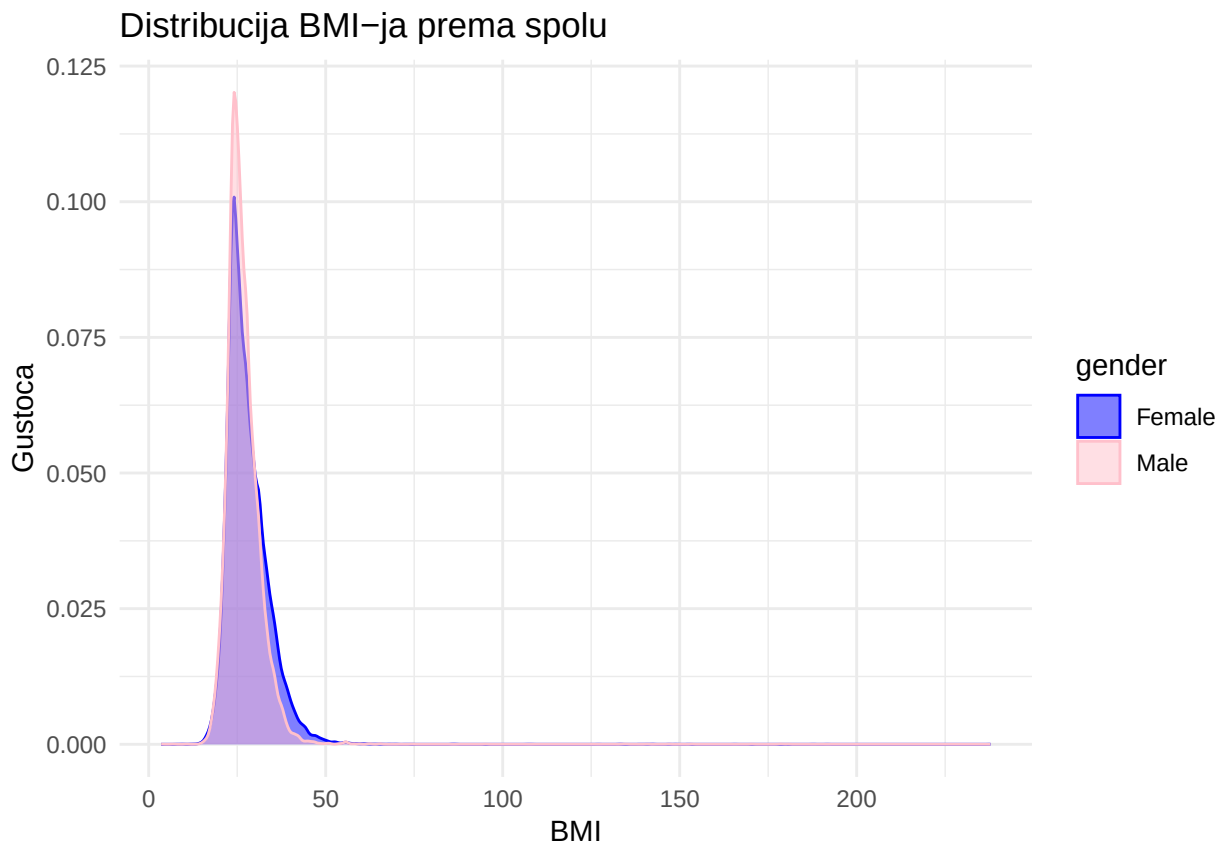

```
ggplot(BMI_filtered_data, aes(x = BMI, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Filtrirana Distribucija BMI-ja prema dobnoj skupini",
       x = "BMI",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")
```



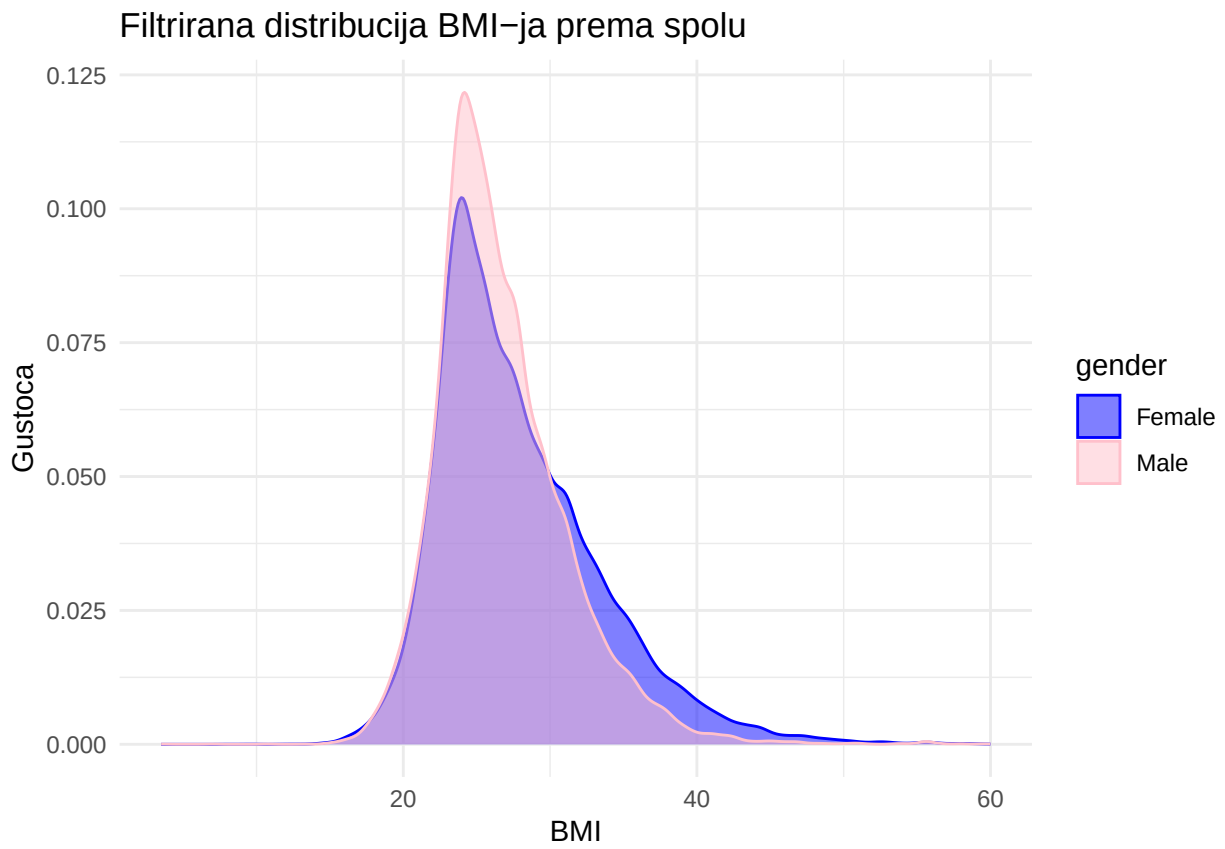
Zaključak - BMI teži iznosima između 20 i 30. Postoji značajna količina podataka koja se nalazi između 30 te opada prema 40. Ostali iznosi nisu bitno reprezentirani.

Idući graf prikazuje distribuciju BMI-ja unutar spolova:

```
ggplot(filtered_data, aes(x = BMI, color = gender, fill = gender)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribucija BMI-ja prema spolu",
       x = "BMI",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "pink")) +
  scale_color_manual(values = c("blue", "pink"))
```




```
ggplot(BMI_filtered_data, aes(x = BMI, color = gender, fill = gender)) +
  geom_density(alpha = 0.5) +
  labs(title = "Filtrirana distribucija BMI-ja prema spolu",
       x = "BMI",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "pink")) +
  scale_color_manual(values = c("blue", "pink"))
```

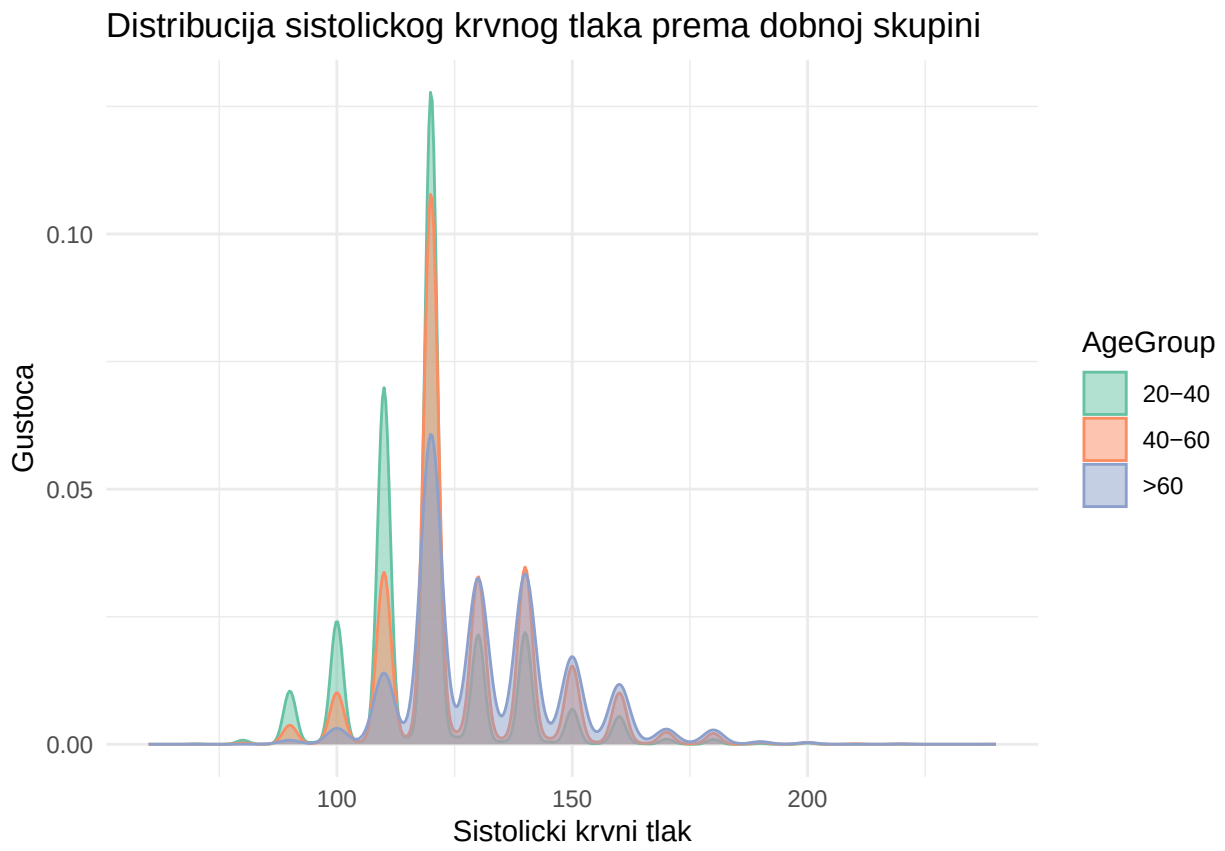


Zaključak - BMI teži iznosima između 20 i 30. Postoji značajna količina podataka koja se nalazi između 30 te opada prema 40. Ostali iznosi nisu bitno reprezentirani.

Sljedeći grafovi prikazuju distribucije iznosa sistoličkog i dijastoličkog krvnog tlaka.. Filtrirani grafovi služe kako bi lakše vidjeli tendenciju distribucije, fokusiranjem na češće iznose.

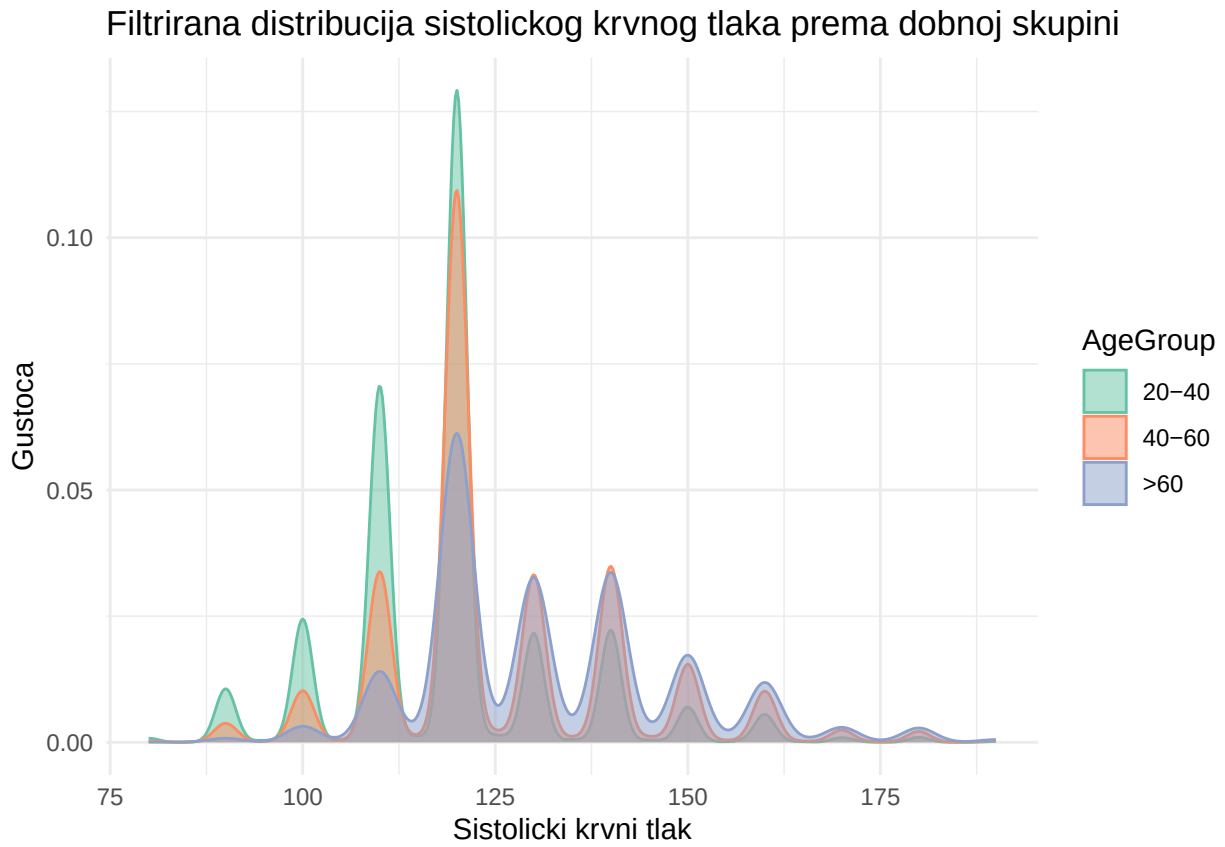
Prvi grafovi prikazuju distribucije krvnog tlaka prema dobnoj skupini:

```
ap_filtered_data <- filtered_data %>% filter(ap_hi <= 190) %>% filter(ap_lo <= 130) %>% filter(ap_hi >=

ggplot(filtered_data, aes(x = ap_hi, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribucija sistoličkog krvnog tlaka prema dobnoj skupini",
       x = "Sistolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")
```

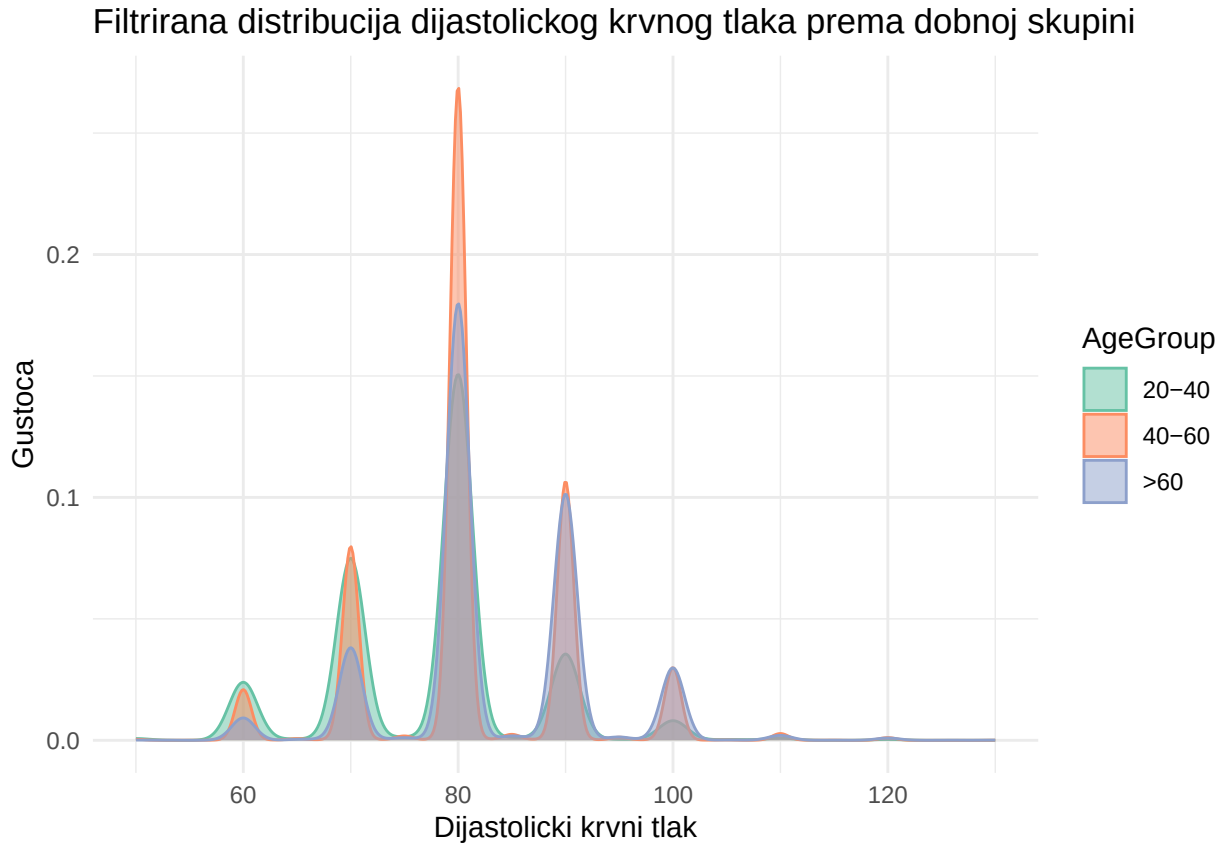



```
ggplot(filtered_data, aes(x = ap_lo, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribucija dijastoličkog krvnog tlaka prema dobnoj skupini",
       x = "Dijastolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")
```


```
ggplot(ap_filtered_data, aes(x = ap_hi, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Filtrirana distribucija sistoličkog krvnog tlaka prema dobnoj skupini",
       x = "Sistolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")
```



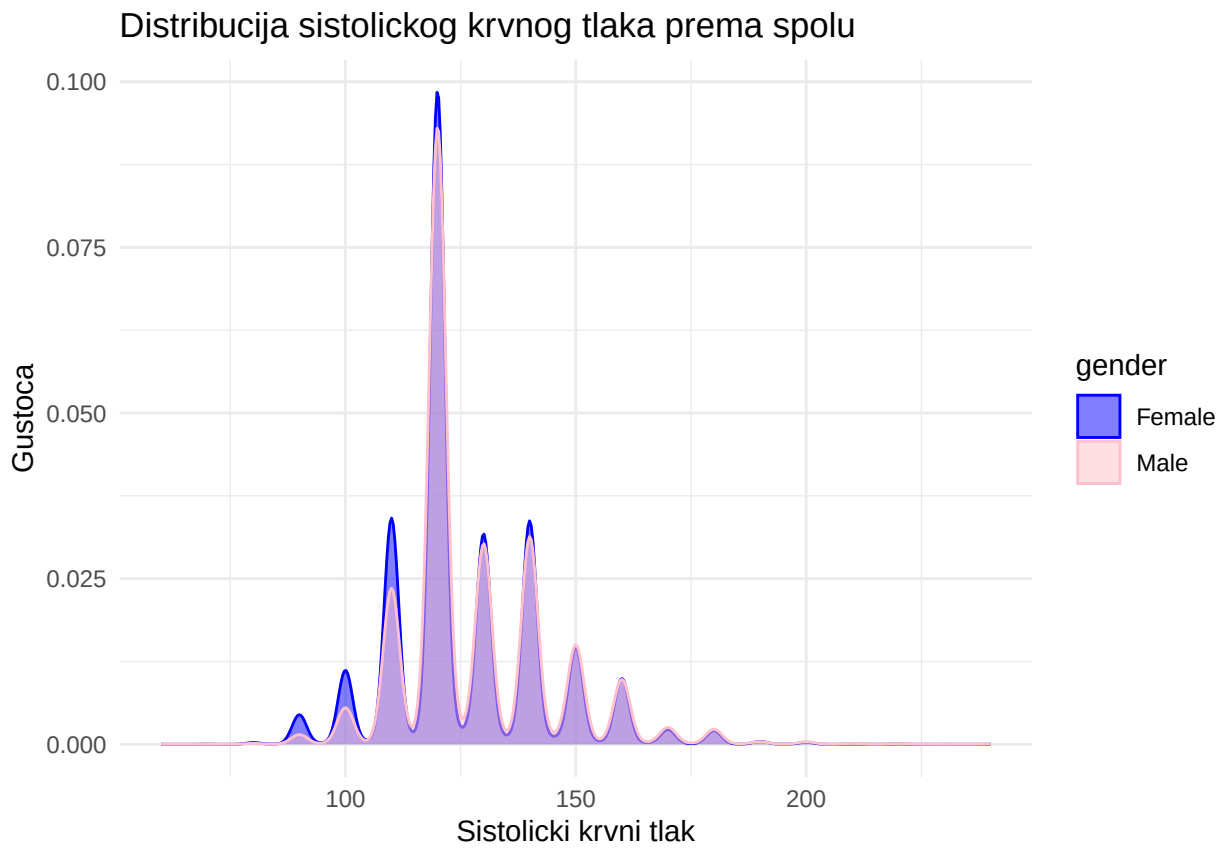

```
ggplot(ap_filtered_data, aes(x = ap_lo, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Filtrirana distribucija dijastoličkog krvnog tlaka prema dobnoj skupini",
       x = "Dijastolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")
```



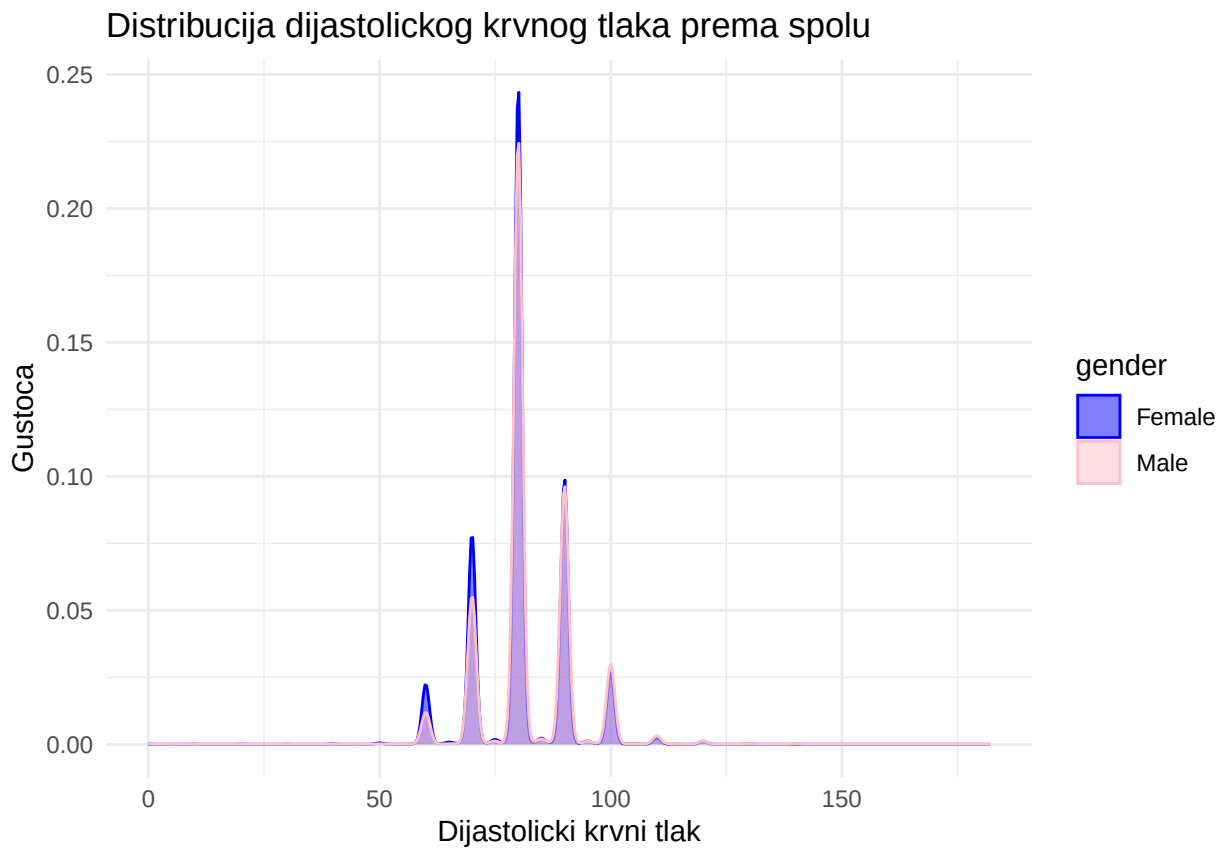
Zaključak - Krvni tlak ima višemodalnu distribuciju. - Najčešći iznos krvnog tlaka je 120/80.

Prema spolu:

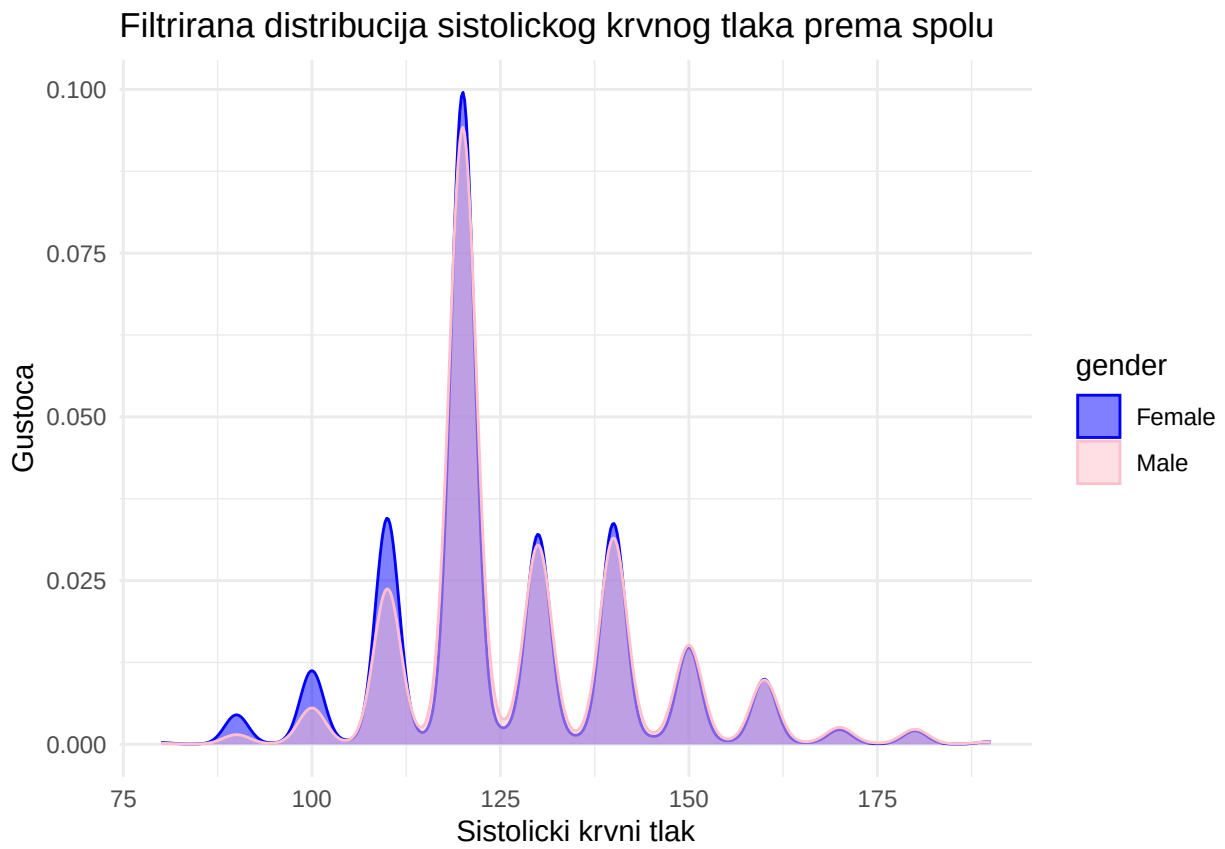
```
ggplot(filtered_data, aes(x = ap_hi, color = gender, fill = gender)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribucija sistoličkog krvnog tlaka prema spolu",
       x = "Sistolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "pink")) +
  scale_color_manual(values = c("blue", "pink"))
```




```
ggplot(filtered_data, aes(x = ap_lo, color = gender, fill = gender)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribucija dijastoličkog krvnog tlaka prema spolu",
       x = "Dijastolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "pink")) +
  scale_color_manual(values = c("blue", "pink"))
```




```
ggplot(ap_filtered_data, aes(x = ap_hi, color = gender, fill = gender)) +
  geom_density(alpha = 0.5) +
  labs(title = "Filtrirana distribucija sistoličkog krvnog tlaka prema spolu",
       x = "Sistolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "pink")) +
  scale_color_manual(values = c("blue", "pink"))
```




```
ggplot(ap_filtered_data, aes(x = ap_lo, color = gender, fill = gender)) +
  geom_density(alpha = 0.5) +
  labs(title = "Filtrirana distribucija dijastoličkog krvnog tlaka prema spolu",
       x = "Dijastolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "pink")) +
  scale_color_manual(values = c("blue", "pink"))
```

Zaključak - Krvni tlak ima višemodalnu distribuciju. - Najčešći iznos krvnog tlaka je 120/80.

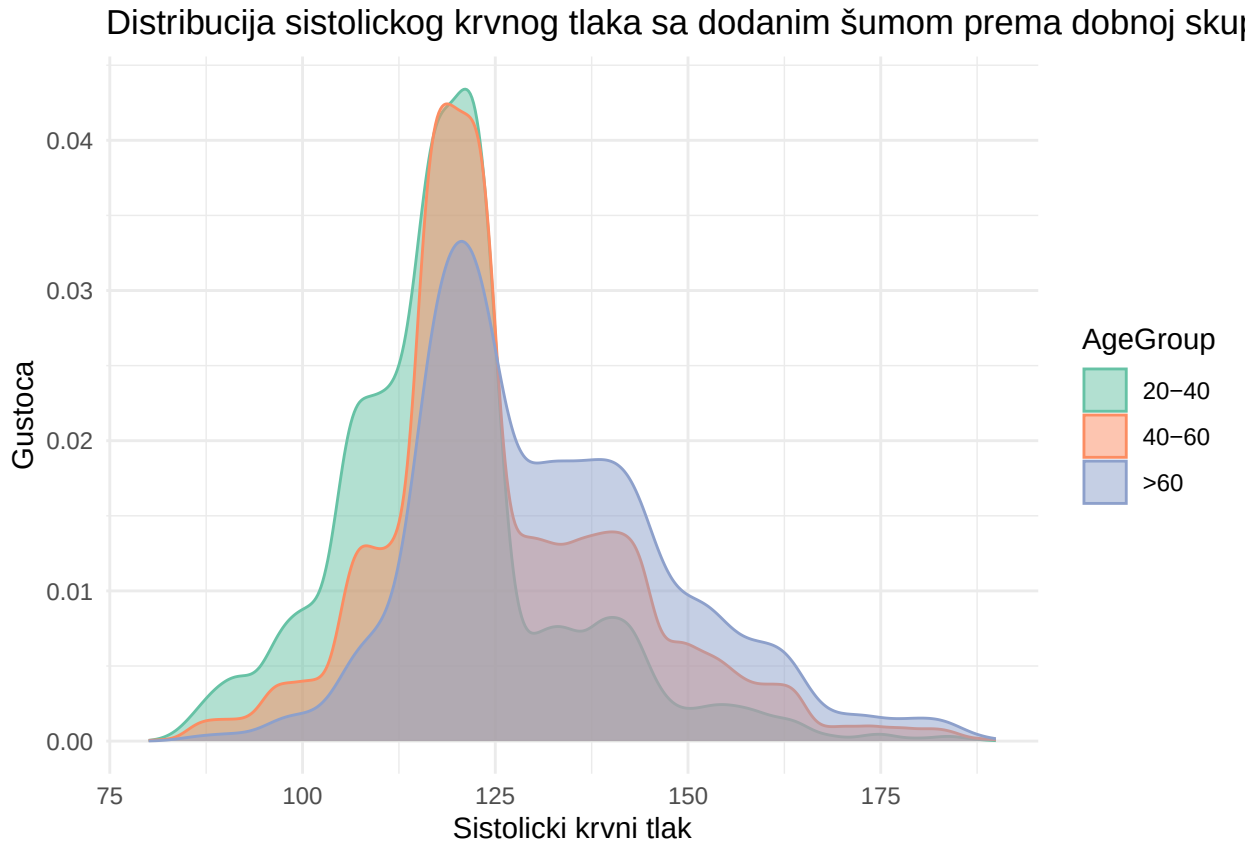
Analizom distribucije krvnog tlaka primijetili smo da se podaci raspodjeljuju višemodalno, što intuitivno nije logično. Na primjer, zašto bi tlak od 80 bio učestaliji od tlaka 85, dok je tlak od 90 također učestaliji od 85? Objašnjenje leži u tendenciji liječnika da zaokružuju vrijednosti krvnog tlaka na brojeve koji završavaju nulom.

Kako je navedeno u dokumentu Svjetske zdravstvene organizacije (WHO) iz travnja 2020., "tehničke pogreške uzrokovane opažačem uključuju sustavne pogreške povezane s ... i suboptimalno bilježenje izmjerenih vrijednosti krvnog tlaka. Primjer je 'preferencija završnog broja', pri čemu opažač zaokružuje izmjerene vrijednosti na preferirani broj, obično nulu." (Izvor: WHO technical specifications for automated non-invasive blood pressure measuring devices with cuff)

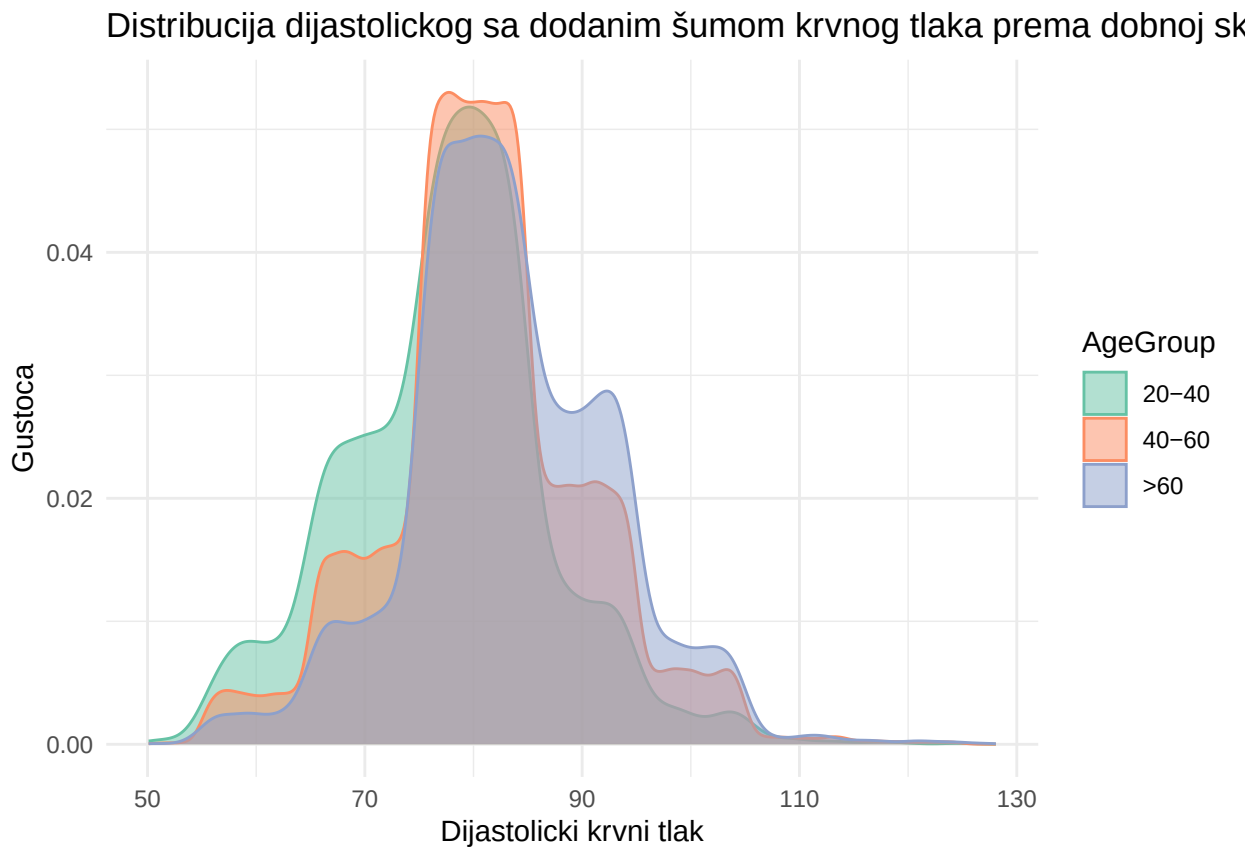
Smatramo da ovaj fenomen značajno utječe na statističke analize. Kako bismo bolje reprezentirali podatke i smanjili utjecaj ove pristranosti, u analizu smo uključili dodavanje uniformnog šuma.

```
set.seed(906)

original_filtered_data <- filtered_data

filtered_data <- filtered_data %>%
  mutate(
    ap_hi = ap_hi + runif(n(), min = -5 , max = 5),
    ap_lo = ap_lo + runif(n(), min = -5 , max = 5)
  )

ap_filtered_data <- filtered_data %>% filter(ap_hi <= 190) %>% filter(ap_lo <= 130) %>% filter(ap_hi >=

ggplot(ap_filtered_data, aes(x = ap_hi, color = AgeGroup, fill = AgeGroup)) +
```

```
  geom_density(alpha = 0.5) +
  labs(title = "Distribucija sistoličkog krvnog tlaka sa dodanim šumom prema dobnoj skupini",
       x = "Sistolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")
```




```
ggplot(ap_filtered_data, aes(x = ap_lo, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribucija dijastoličkog sa dodanim šumom krvnog tlaka prema dobnoj skupini",
       x = "Dijastolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")
```

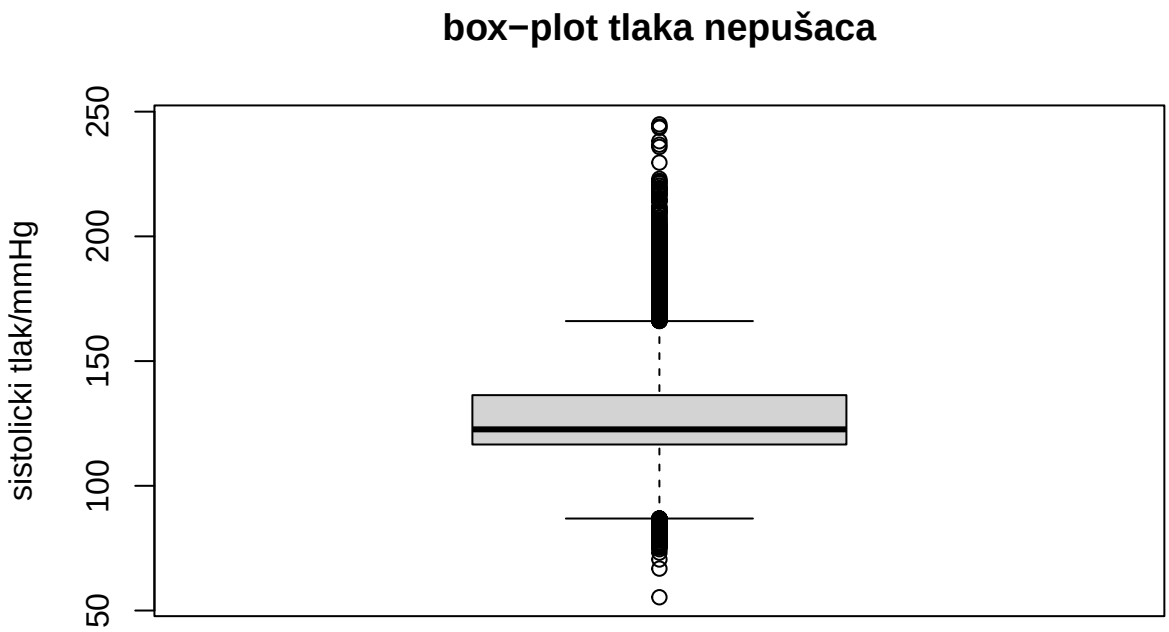


#Zadatak 2

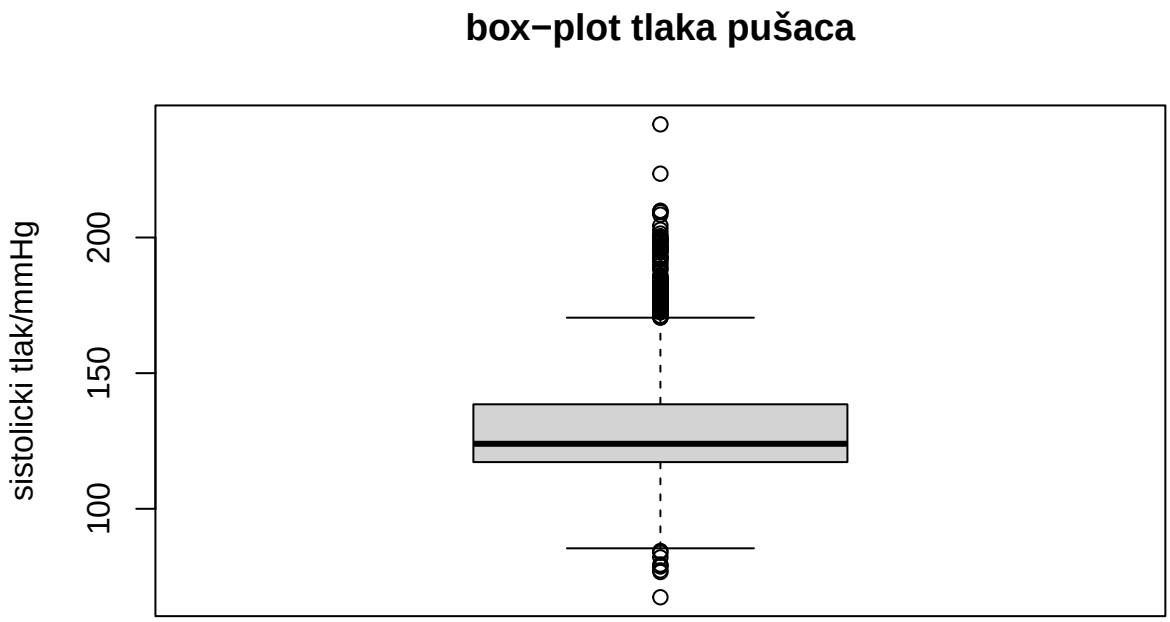
Testiramo postoji li statistički značajna razlika u prosječnim krvnim tlakovima kod pušača i kod nepušača. Pušenje je kategorijska varijabla (razlikujemo pušače i nepušače, a nema podataka o tome koliko često tko puši).

Tlakove pušača i nepušača možemo prvo usporediti grafički pomoću box plotova. U uzorku preostaje oko 6100 pušača i 63000 nepušača nakon eliminacije besmislenih vrijednosti, što znači da imamo dovoljno podataka da kasnije u testiranju možemo koristiti centralni granični teorem. Vidimo da plotovi izgledaju dosta slično. I dalje postoji dosta outliera, pogotovo s visokim tlakom, ali nema ih smisla odbaciti kao pogrešna mjerenja jer je najveći tlak 240, što je sasvim moguća vrijednost.

```
pusaci = filtered_data[filtered_data["smoke"]==1,]
nepusaci = filtered_data[filtered_data["smoke"]==0,]
nrow(pusaci)
nrow(nepusaci)

boxplot(nepusaci$ap_hi,
        main='box-plot tlaka nepušača',
        ylab='sistolički tlak/mmHg')
```



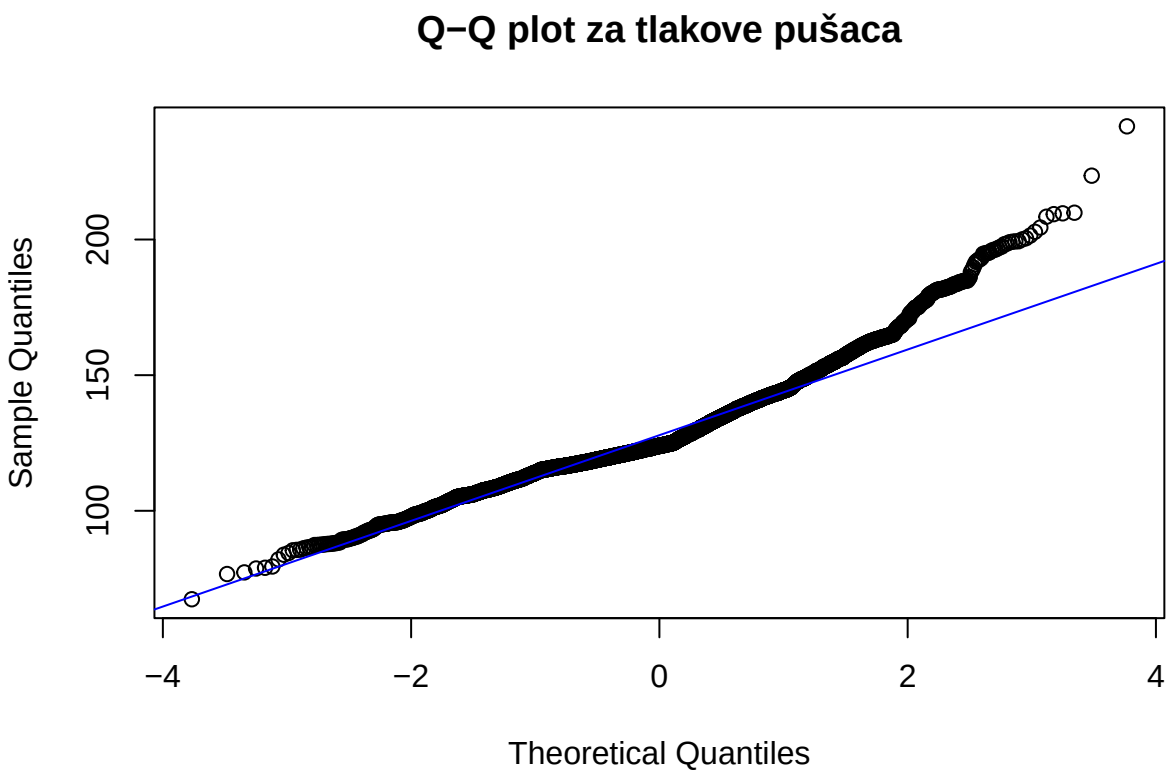

```
boxplot(pusaci$ap_hi,
        main='box-plot tlaka pušača',
        ylab='sistolički tlak/mmHg')
```



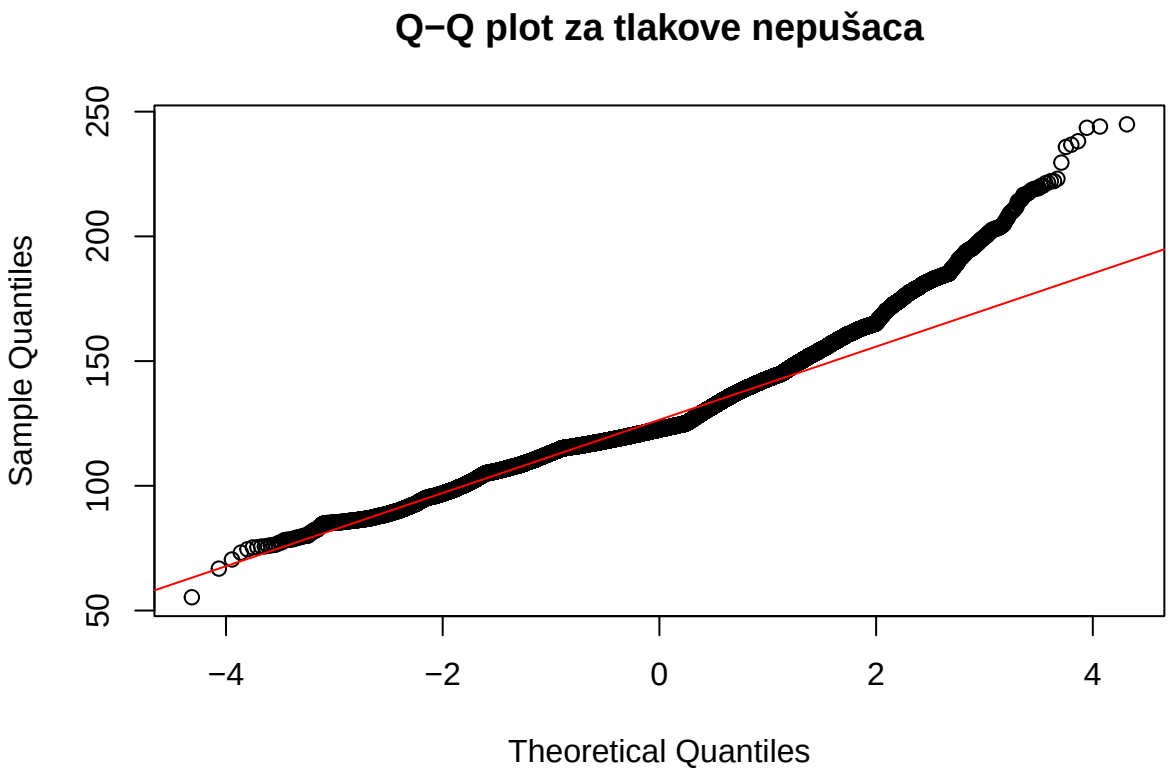

```
## [1] 6042
## [1] 62676
```

Najprije provjeravama jesu li podaci iz normalne razdiobe. Kako ne znamo koju konkretnu normalnu razdiobu očekujemo, koristimo Lillieforsovu inačicu Kolmogorov-Smirnovljevog testa. Posebno testiramo tlakove pušača i tlakove nepušača. Također radimo Q-Q plotove za vizualnu usporedbu kvantila s kvantilima normalne razdiobe.

```
pusaci = filtered_data[filtered_data["smoke"]==1,]
nepusaci = filtered_data[filtered_data["smoke"]==0,]

lillie.test(pusaci$ap_hi)
lillie.test(nepusaci$ap_hi)
qqnorm(pusaci$ap_hi, main = "Q-Q plot za tlakove pušača")
qqline(pusaci$ap_hi, col="blue")
```




```r
qqnorm(nepusaci$ap_hi, main = "Q-Q plot za tlakove nepušača")
qqline(nepusaci$ap_hi, col="red")
```




```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pusaci$ap_hi
## D = 0.11454, p-value < 2.2e-16
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nepusaci$ap_hi
## D = 0.13324, p-value < 2.2e-16
```

Test odbacuje nul-hipotezu (da su podaci slučajan uzorak iz normalne razdiobe) za obje grupe na razini značajnosti od 1%. Na Q-Q plotovima vidimo da distribucija podataka relativno dobro prati normalnu za vrijednosti oko prosjeka i manje, ali ima vrlo teški rep prema većim vrijednostima. To je u skladu s prisutnošću mnogo outliera vidljivih u tom području na box plotu.

Treba provjeriti jesu li varijance grupa jednake. Provodimo F-test za jednakost varijanci sitoličkih tlakova nepušača i pušača. Dobivamo vrlo malu p-vrijednost, pa odbacujemo nul-hipotezu. Procijenjeni omjer varijanci je oko 0.9, dakle pušači imaju veću varijancu u sistoličkom tlaku nego nepušači. Poznato je da pušenje uzrokuje kratkotrajni porast krvnog tlaka. Moguće je tlakovi pušača više variraju jer su nekim pušačima tlakovi izmjereni ubrzo nakon pušenja, a nekima nakon nekoliko sati bez cigareta. U svakom slučaju, ne možemo pretpostaviti jednakost varijanci u daljnjim testovima.

```
var.test(nepusaci$ap_hi, pusaci$ap_hi)
```

```
##
##  F test to compare two variances
```

```
##
## data:  nepusaci$ap_hi and pusaci$ap_hi
## F = 0.90736, num df = 62675, denom df = 6041, p-value = 2.263e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8738485 0.9416141
## sample estimates:
## ratio of variances
##           0.907363
```

Testiramo hipotezu o jednakosti prosječnih krvnih tlakova u obje grupe koristeći T test za neuparene podatke uz nepoznate i nejednake varijance na razini značajnosti od 5%. Ne uzimamo pretpostavku da su varijance jednake zbog rezultata prethodnog testa.

```
t.test(filtered_data[filtered_data["smoke"]==0,]$ap_hi, filtered_data[filtered_data["smoke"]==1,]$ap_hi
```

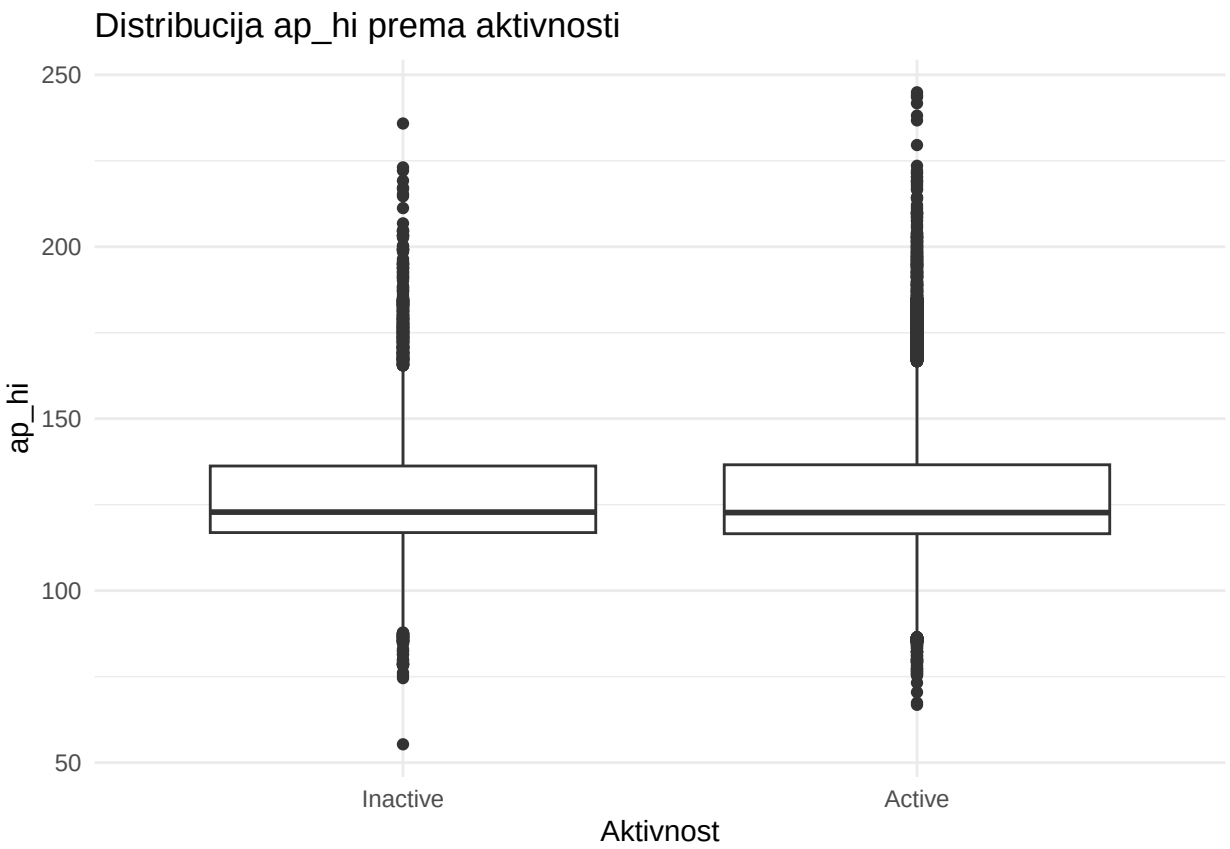
```
##
##  Welch Two Sample t-test
##
## data:  filtered_data[filtered_data["smoke"] == 0, ]$ap_hi and filtered_data[filtered_data["smoke"] ==
## t = -7.0586, df = 7138.8, p-value = 1.84e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.140644 -1.210087
## sample estimates:
## mean of x mean of y
##  126.5168  128.1921
```

Dobivena P vrijednost je vrlo mala pa možemo odbaciti hipotezu o jednakosti srednjih vrijednosti sistoličkih tlakova na razini značajnosti od 5%. Test pokazuje statistički značajnu razliku, no procijenjena razlika sredina je mala u odnosu na 30 mmHg koliko otprilike iznosi širina intervala normalnih tlakova pa je značaj te razlike u praksi upitan.

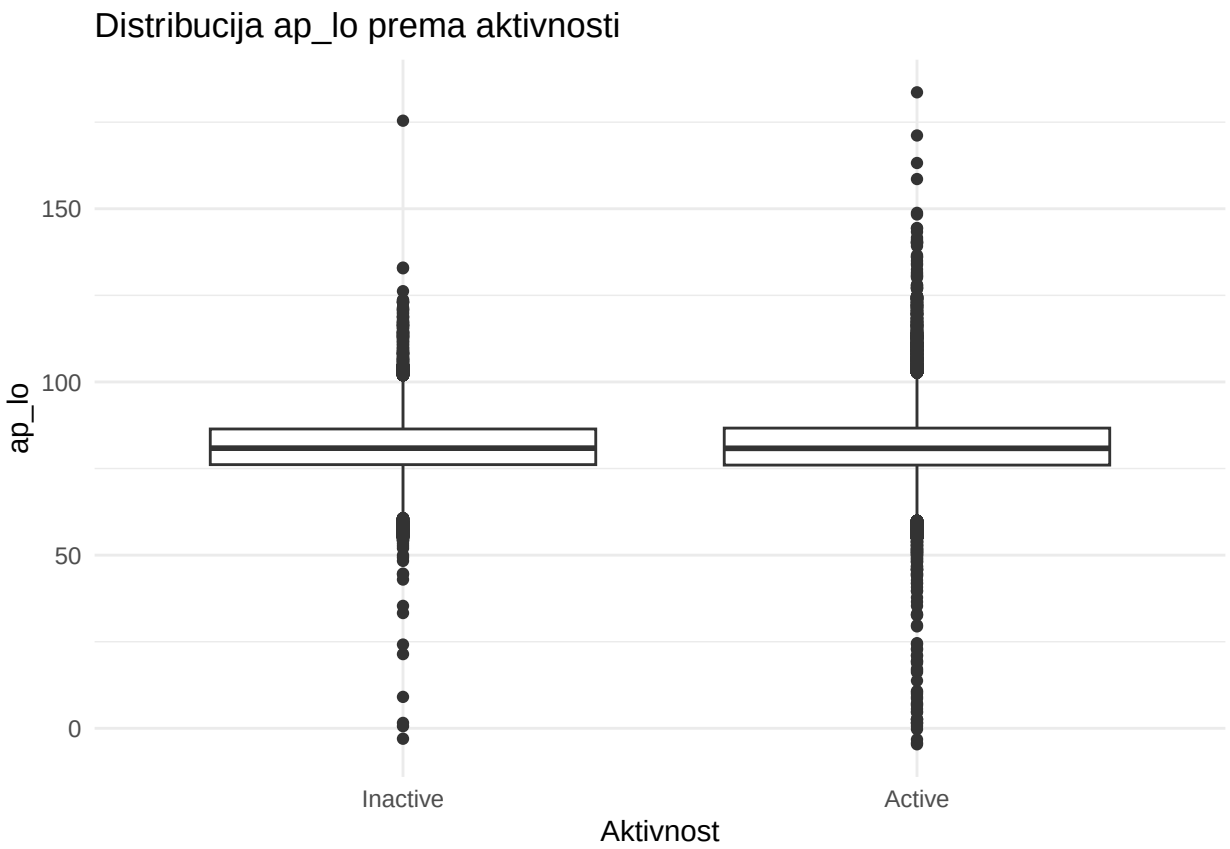
#Zadatak 3

Za početak kako bi dobili dojam o utjecaju tjelesne aktivnosti crtamo box plotove za visoki i niski krvni tlak grupirane po tjelesnoj aktivnosti.

```
# Podjela podataka prema aktivnosti
active_data <- filtered_data %>% filter(active == 1)
inactive_data <- filtered_data %>% filter(active == 0)

# 1) Boxplot za ap_hi i ap_lo prema aktivnosti
ggplot(filtered_data, aes(x = factor(active, labels = c("Inactive", "Active")), y = ap_hi)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Distribucija ap_hi prema aktivnosti", x = "Aktivnost", y = "ap_hi")
```



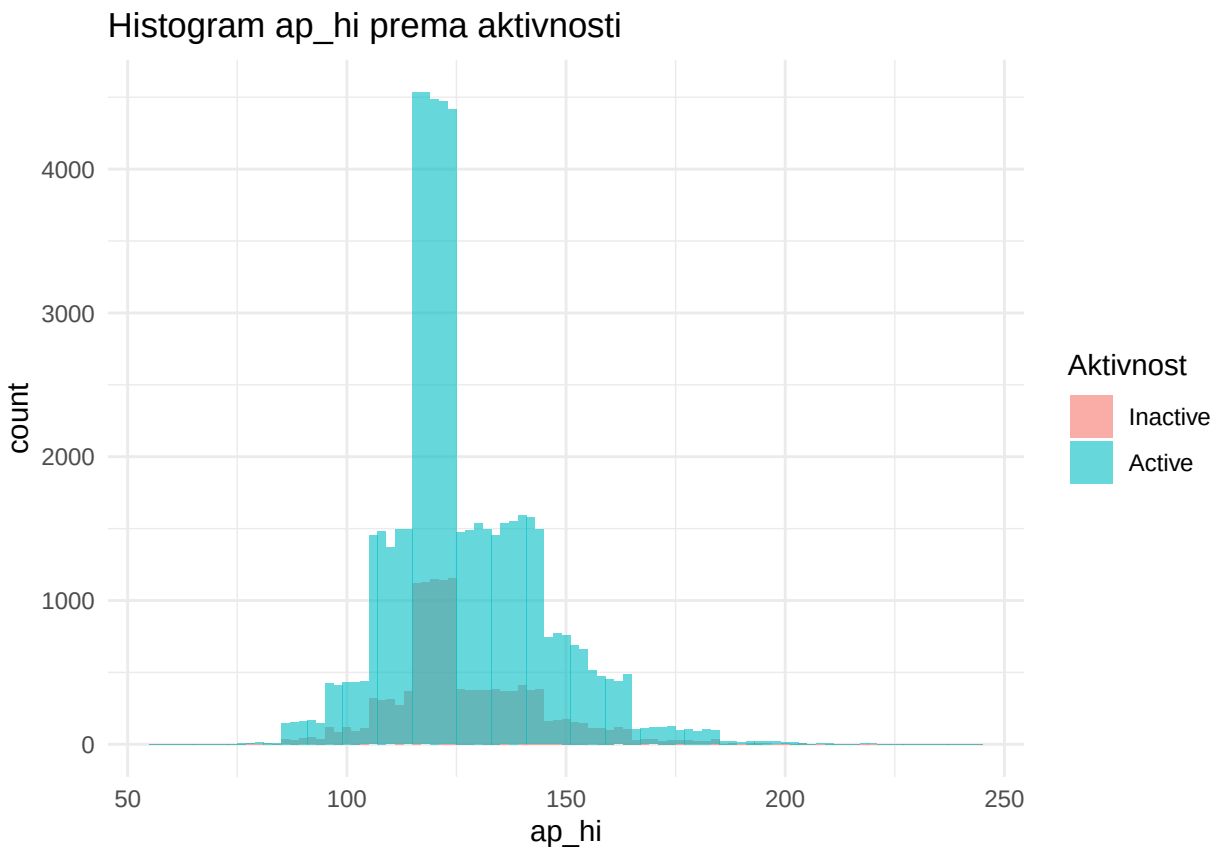

```
ggplot(filtered_data, aes(x = factor(active, labels = c("Inactive", "Active")), y = ap_lo)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Distribucija ap_lo prema aktivnosti", x = "Aktivnost", y = "ap_lo")
```



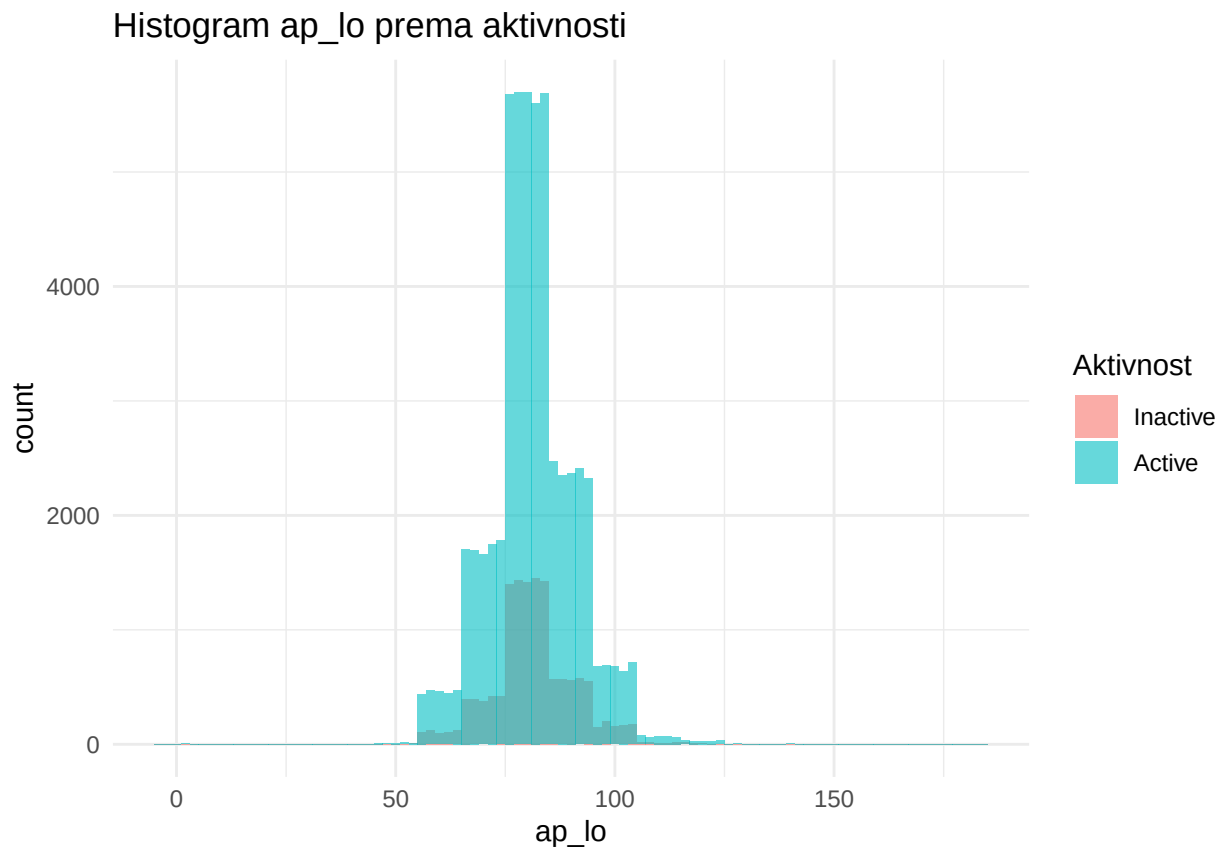
Iz plotova vidimo kako su vrijednosti krvnih tlakova dosla slične kod aktivnih i neaktivnih ljudi. Ipak kako ne bi ostali na tome provjerit ćemo jednakos sredina odgovarajucim testovima.

Za određivanje možemo li koristiti t-test provjeravamo njegove pretpostavke. Prvo ćemo provjeriti normalnost grupa na temelju histograma i qq plota.

```
# 2) Histogram za ap_hi i ap_lo prema aktivnosti
ggplot(filtered_data, aes(x = ap_hi, fill = factor(active, labels = c("Inactive", "Active")))) +
  geom_histogram(binwidth = 2, alpha = 0.6, position = "identity") +
  theme_minimal() +
  labs(title = "Histogram ap_hi prema aktivnosti", x = "ap_hi", fill = "Aktivnost")
```

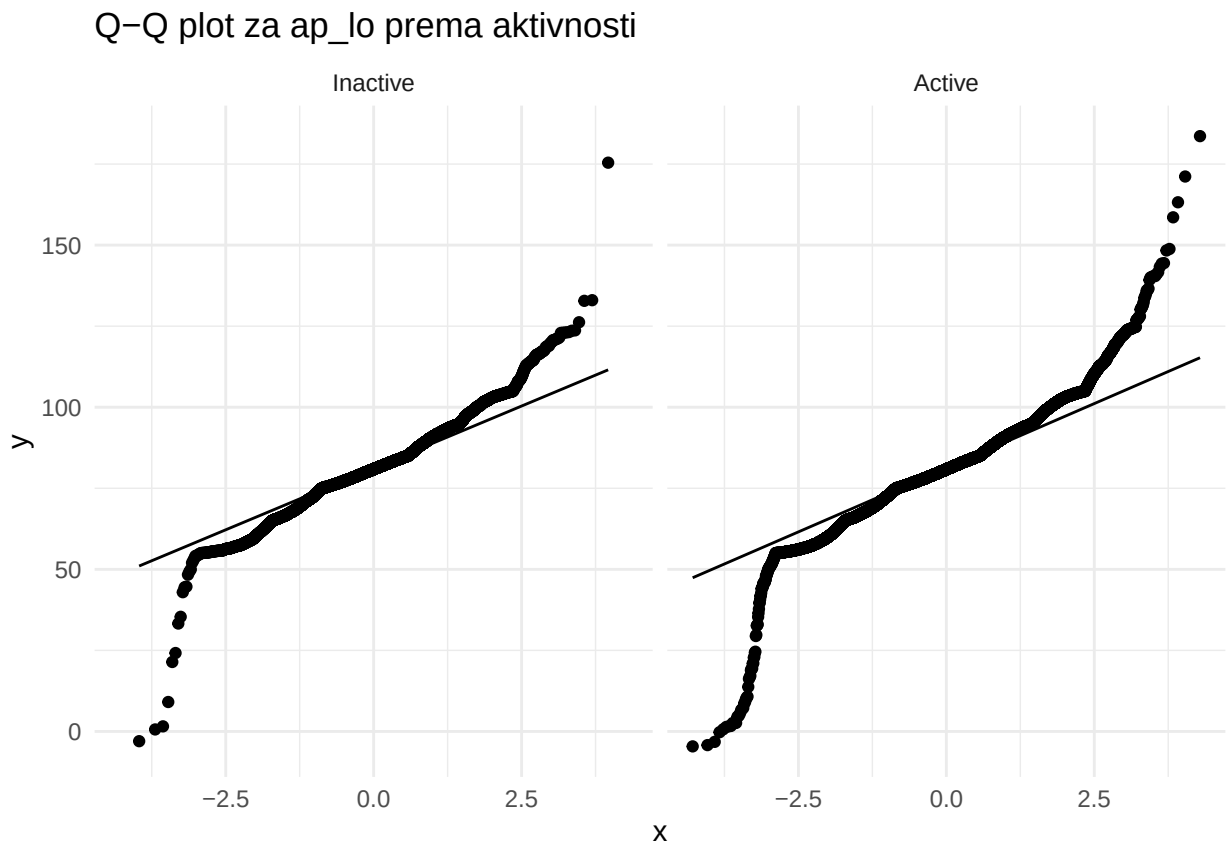



```r
ggplot(filtered_data, aes(x = ap_lo, fill = factor(active, labels = c("Inactive", "Active")))) +
  geom_histogram(binwidth = 2, alpha = 0.6, position = "identity") +
  theme_minimal() +
  labs(title = "Histogram ap_lo prema aktivnosti", x = "ap_lo", fill = "Aktivnost")
```




```r
# 5) Q-Q Plot za provjeru normalnosti
ggplot(filtered_data, aes(sample = ap_hi)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~ active, labeller = as_labeller(c("0" = "Inactive", "1" = "Active"))) +
  theme_minimal() +
  labs(title = "Q-Q plot za ap_hi prema aktivnosti")
```


```
ggplot(filtered_data, aes(sample = ap_lo)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~ active, labeller = as_labeller(c("0" = "Inactive", "1" = "Active"))) +
  theme_minimal() +
  labs(title = "Q-Q plot za ap_lo prema aktivnosti")
```



Iz plotova vidimo kako distribucija najvjerovatnije nije normalna ali za svaki slučaj ćemo za provjeru normalnosti i jednakosti varijanci provesti kolmogorov-smirnovljev test i f-test.

```
# 3) Kolmogorov-Smirnov test za normalnost
cat("\n===== Kolmogorov-Smirnov test za normalnost =====\n")

ks_hi_active <- ks.test(active_data$ap_hi, "pnorm", mean = mean(active_data$ap_hi), sd = sd(active_data
ks_lo_active <- ks.test(active_data$ap_lo, "pnorm", mean = mean(active_data$ap_lo), sd = sd(active_data

ks_hi_inactive <- ks.test(inactive_data$ap_hi, "pnorm", mean = mean(inactive_data$ap_hi), sd = sd(inact
ks_lo_inactive <- ks.test(inactive_data$ap_lo, "pnorm", mean = mean(inactive_data$ap_lo), sd = sd(inact

cat("\nKolmogorov-Smirnov test za ap_hi (Active): p-value:", ks_hi_active$p.value, "\n")
cat("Kolmogorov-Smirnov test za ap_lo (Active): p-value:", ks_lo_active$p.value, "\n")
cat("\nKolmogorov-Smirnov test za ap_hi (Inactive): p-value:", ks_hi_inactive$p.value, "\n")
cat("Kolmogorov-Smirnov test za ap_lo (Inactive): p-value:", ks_lo_inactive$p.value, "\n")

# 4) F-test za homogenost varijance
cat("\n===== F-test za varijance =====\n")

f_test_hi <- var.test(active_data$ap_hi, inactive_data$ap_hi)
f_test_lo <- var.test(active_data$ap_lo, inactive_data$ap_lo)

var.test(active_data$ap_hi, inactive_data$ap_hi)
var.test(active_data$ap_lo, inactive_data$ap_lo)

cat("\nF-test za ap_hi: p-value:", f_test_hi$p.value, "\n")
```

```r
cat("F-test za ap_lo: p-value:", f_test_lo$p.value, "\n")
```

```
##
## ===== Kolmogorov-Smirnov test za normalnost =====
##
## Kolmogorov-Smirnov test za ap_hi (Active): p-value: 0
## Kolmogorov-Smirnov test za ap_lo (Active): p-value: 0
##
## Kolmogorov-Smirnov test za ap_hi (Inactive): p-value: 0
## Kolmogorov-Smirnov test za ap_lo (Inactive): p-value: 0
##
## ===== F-test za varijance =====
##
##  F test to compare two variances
##
## data:  active_data$ap_hi and inactive_data$ap_hi
## F = 1.0061, num df = 55205, denom df = 13511, p-value = 0.6584
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9795407 1.0330750
## sample estimates:
## ratio of variances
##           1.006061
##
##
##  F test to compare two variances
##
## data:  active_data$ap_lo and inactive_data$ap_lo
## F = 1.0336, num df = 55205, denom df = 13511, p-value = 0.01537
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.006352 1.061352
## sample estimates:
## ratio of variances
##           1.033599
##
##
## F-test za ap_hi: p-value: 0.6583691
## F-test za ap_lo: p-value: 0.01536583
```

Iz rezultata testova vidimo kao i iz plotova kako

```r
# Podjela podataka prema aktivnosti
active_data <- filtered_data %>% filter(active == 1)
inactive_data <- filtered_data %>% filter(active == 0)

# 1) Standardni t-test (za normalno distribuirane podatke)
cat("\n===== t-test za ap_hi =====\n")
t_test_hi <- t.test(active_data$ap_hi, inactive_data$ap_hi, var.equal = FALSE)  # Koristi Welchov t-tes
print(t_test_hi)

cat("\n===== t-test za ap_lo =====\n")
t_test_lo <- t.test(active_data$ap_lo, inactive_data$ap_lo, var.equal = FALSE)
print(t_test_lo)
```

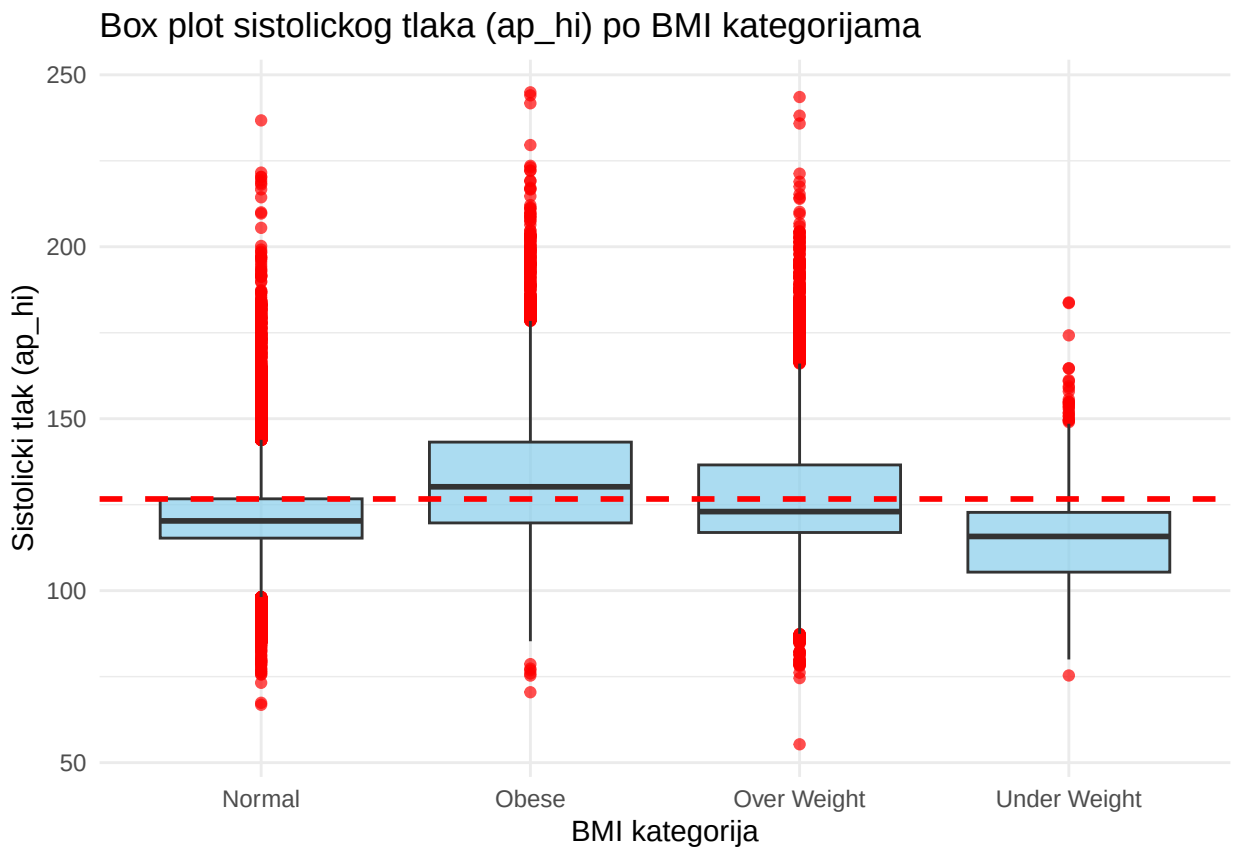
```r
# 2) Neparametrijski Mann-Whitney-Wilcoxon test (za nenormalno distribuirane podatke)
cat("\n===== Mann-Whitney-Wilcoxon test za ap_hi =====\n")
wilcox_hi <- wilcox.test(active_data$ap_hi, inactive_data$ap_hi)
print(wilcox_hi)

cat("\n===== Mann-Whitney-Wilcoxon test za ap_lo =====\n")
wilcox_lo <- wilcox.test(active_data$ap_lo, inactive_data$ap_lo)
print(wilcox_lo)
```

```
##
## ===== t-test za ap_hi =====
##
##  Welch Two Sample t-test
##
## data:  active_data$ap_hi and inactive_data$ap_hi
## t = -0.50551, df = 20677, p-value = 0.6132
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4000184  0.2359906
## sample estimates:
## mean of x mean of y
##  126.6479  126.7300
##
##
## ===== t-test za ap_lo =====
##
##  Welch Two Sample t-test
##
## data:  active_data$ap_lo and inactive_data$ap_lo
## t = -0.64135, df = 20885, p-value = 0.5213
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2484678  0.1259549
## sample estimates:
## mean of x mean of y
##  81.24679  81.30805
##
##
## ===== Mann-Whitney-Wilcoxon test za ap_hi =====
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  active_data$ap_hi and inactive_data$ap_hi
## W = 370247533, p-value = 0.1875
## alternative hypothesis: true location shift is not equal to 0
##
##
## ===== Mann-Whitney-Wilcoxon test za ap_lo =====
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  active_data$ap_lo and inactive_data$ap_lo
## W = 371656426, p-value = 0.5245
## alternative hypothesis: true location shift is not equal to 0
```

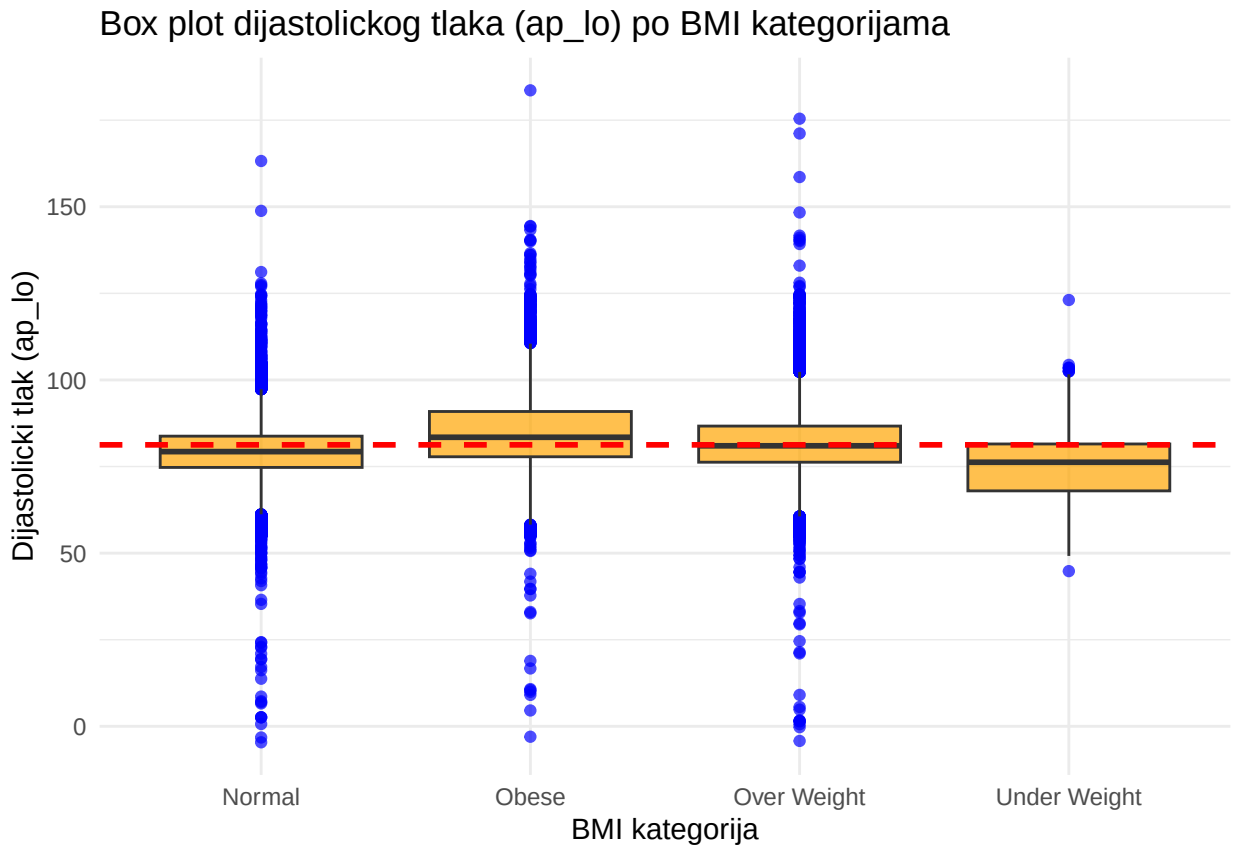
```r
# Boxplot za sistolički tlak (ap_hi) po BMI kategorijama
p_hi <- ggplot(filtered_data, aes(x = BMICat, y = ap_hi)) +
  geom_boxplot(fill = "skyblue", alpha = 0.7, outlier.color = "red") +
  # Crvena isprekidana linija je ukupna sredina (mean) ap_hi
  geom_hline(yintercept = mean(filtered_data$ap_hi),
             color = "red", linetype = "dashed", size = 1) +
  labs(title = "Box plot sistoličkog tlaka (ap_hi) po BMI kategorijama",
       x = "BMI kategorija",
       y = "Sistolički tlak (ap_hi)") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
# Boxplot za dijastolički tlak (ap_lo) po BMI kategorijama
p_lo <- ggplot(filtered_data, aes(x = BMICat, y = ap_lo)) +
  geom_boxplot(fill = "orange", alpha = 0.7, outlier.color = "blue") +
  # Crvena isprekidana linija je ukupna sredina (mean) ap_lo
  geom_hline(yintercept = mean(filtered_data$ap_lo),
             color = "red", linetype = "dashed", size = 1) +
  labs(title = "Box plot dijastoličkog tlaka (ap_lo) po BMI kategorijama",
       x = "BMI kategorija",
       y = "Dijastolički tlak (ap_lo)") +
  theme_minimal()

# Prikaz oba grafa
print(p_hi)
```

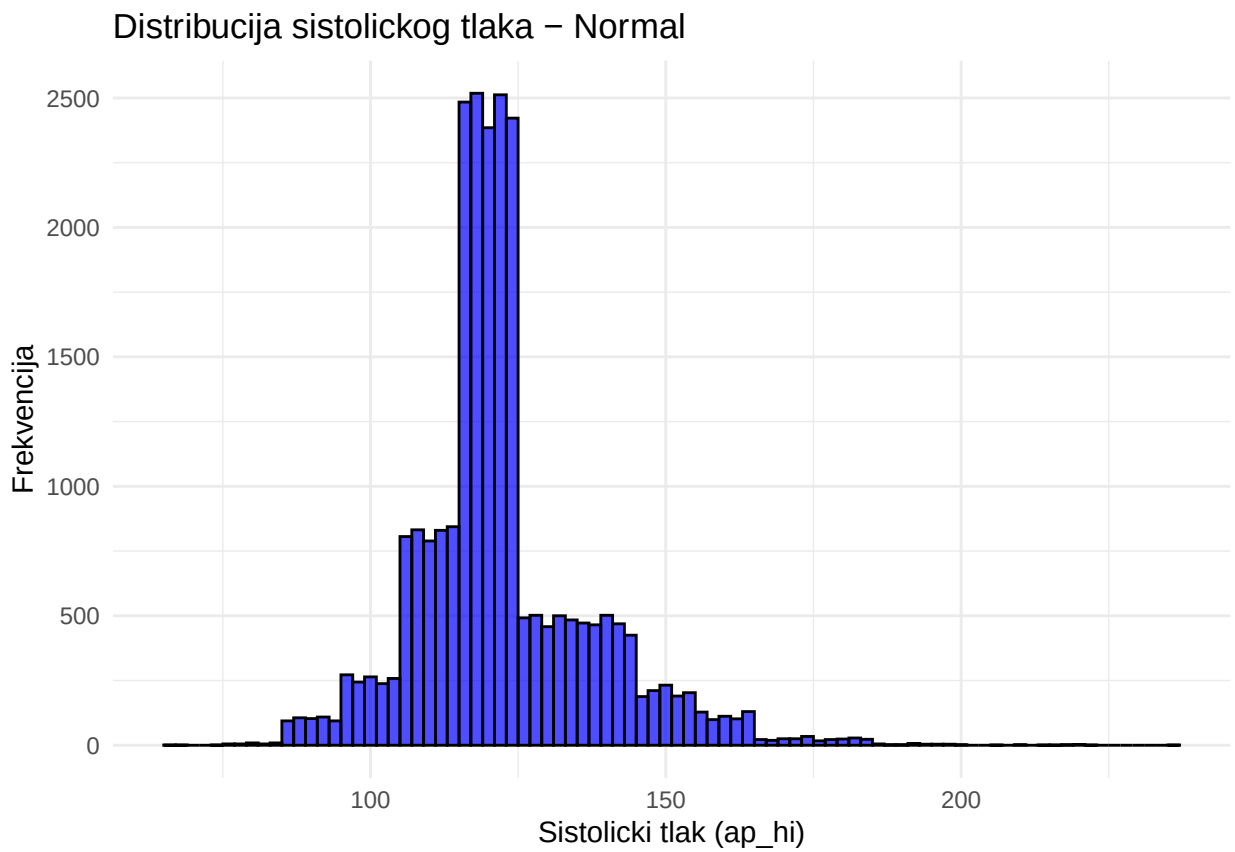

Box plot sistolickog tlaka (ap_hi) po BMI kategorijama
Sistolicki tlak (ap_hi)
250
200
150
100
50
Normal
Obese
Over Weight
Under Weight
BMI kategorija

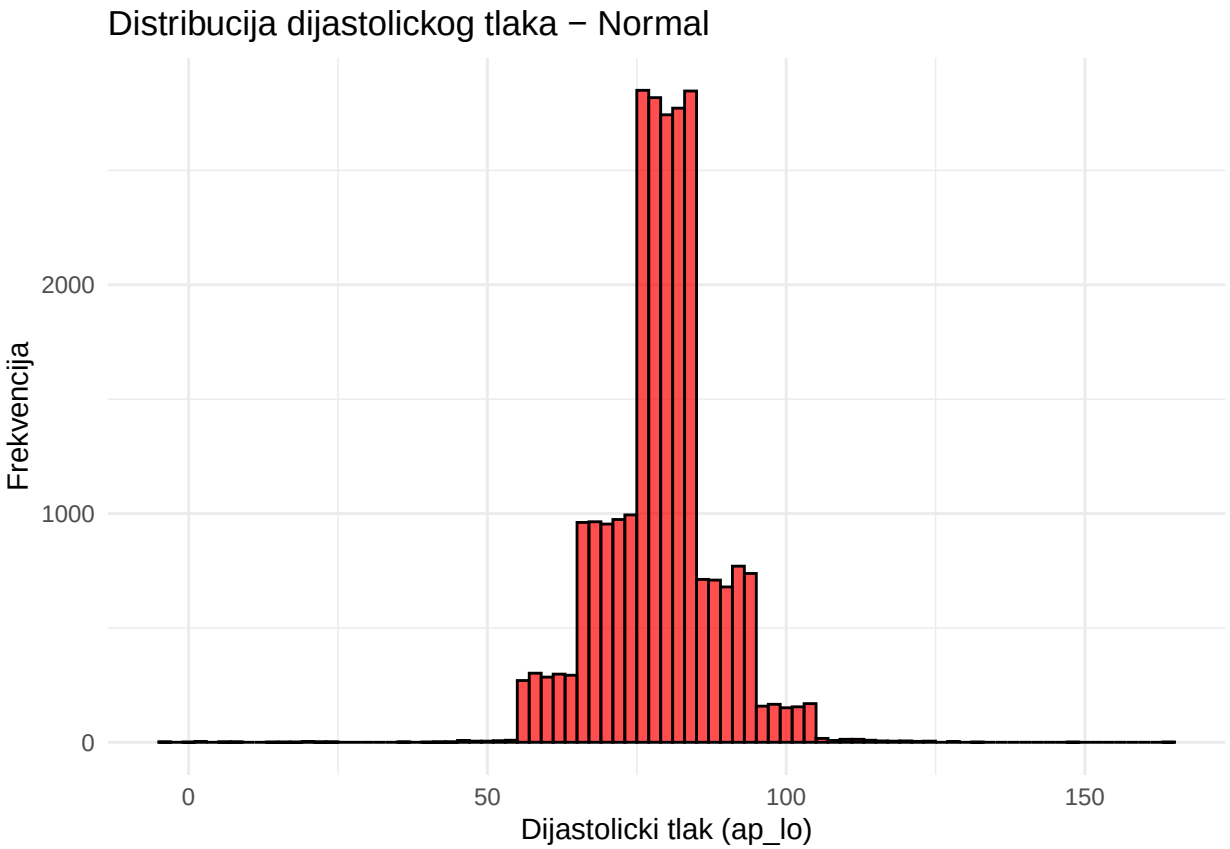
```
print(p_lo)
```

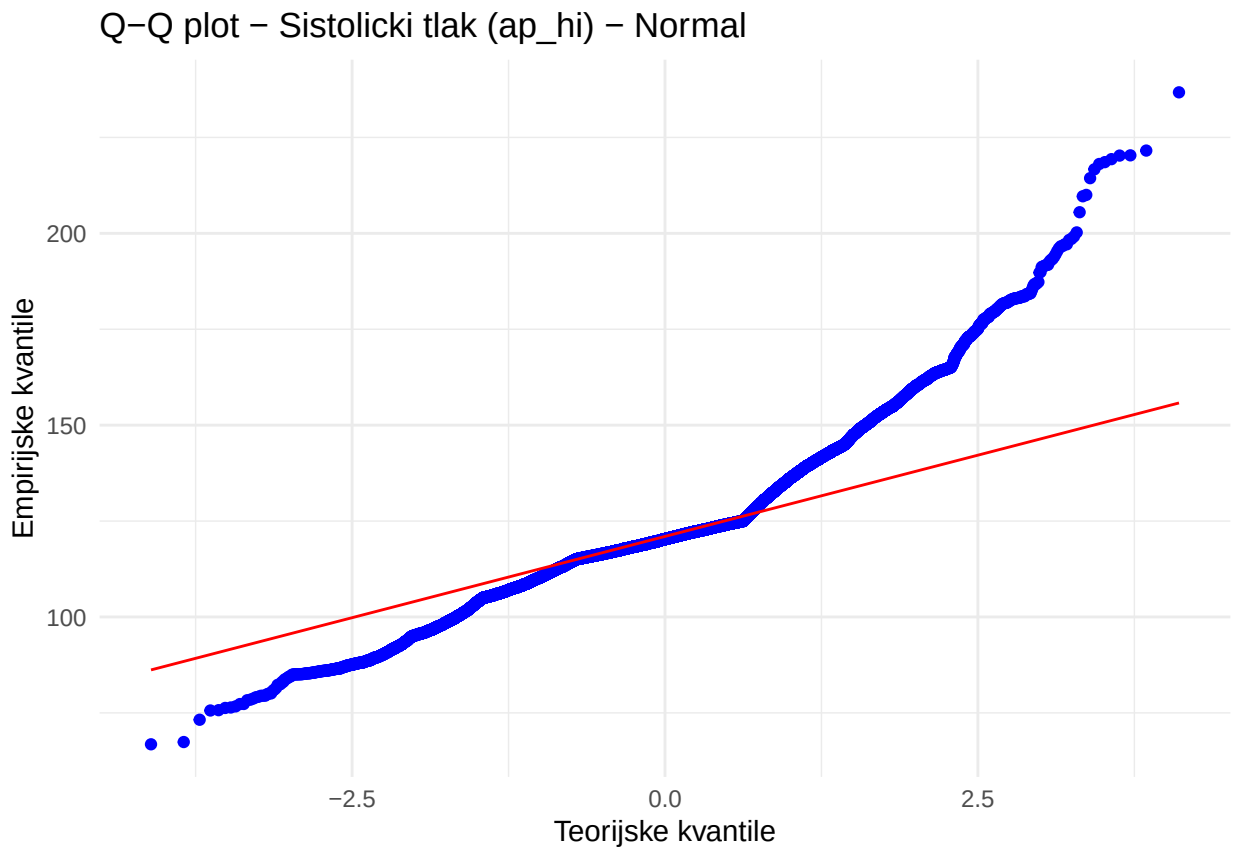


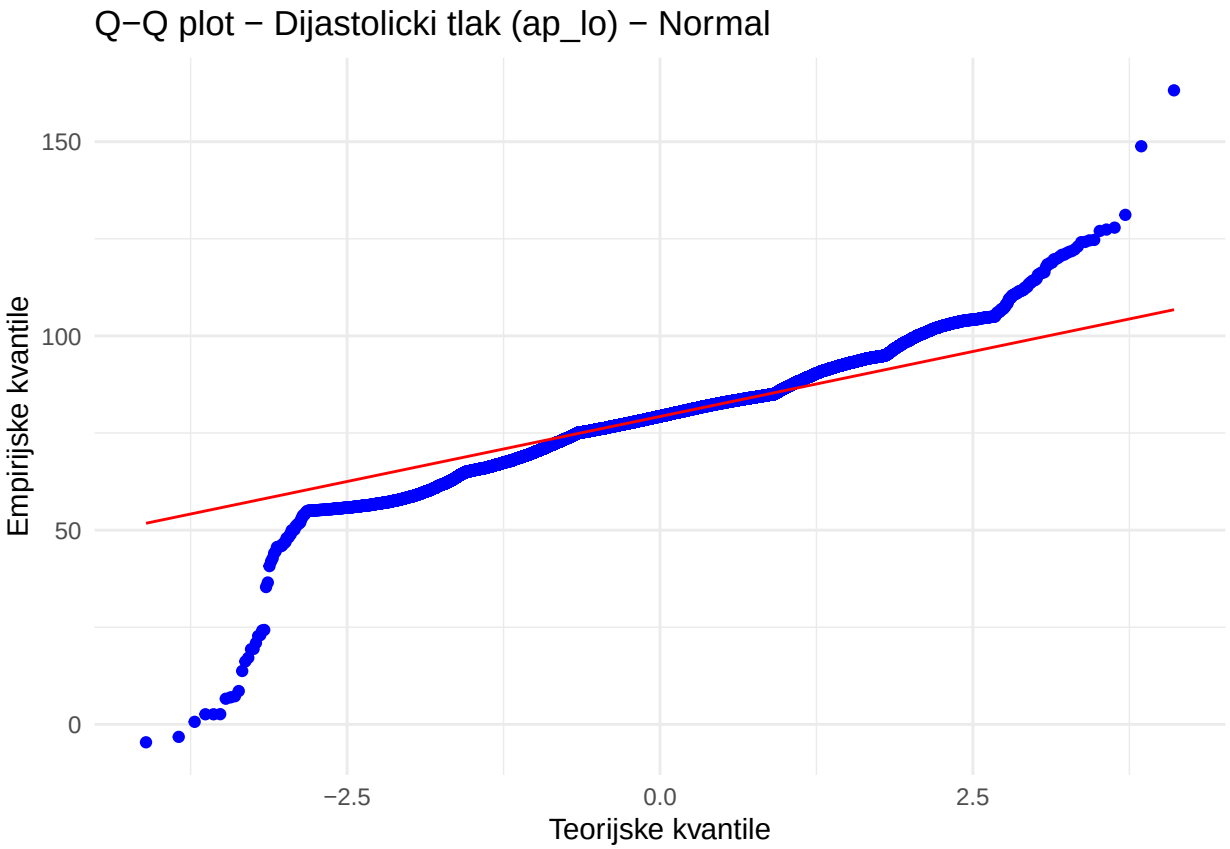

```r
# Lista jedinstvenih BMI kategorija
bmi_categories <- unique(filtered_data$BMICat)

# Kreiranje histograma i QQ plotova za svaku BMI kategoriju
for (bmi_cat in bmi_categories) {

  # Filtriranje podataka za trenutnu BMI kategoriju
  data_subset <- filtered_data %>% filter(BMICat == bmi_cat)

  # Histogram za sistolički tlak (ap_hi)
  p1 <- ggplot(data_subset, aes(x = ap_hi)) +
    geom_histogram(binwidth = 2, fill = "blue", color = "black", alpha = 0.7) +
    labs(title = paste("Distribucija sistoličkog tlaka -", bmi_cat),
         x = "Sistolicki tlak (ap_hi)", y = "Frekvencija") +
    theme_minimal()

  # Histogram za dijastolički tlak (ap_lo)
  p2 <- ggplot(data_subset, aes(x = ap_lo)) +
    geom_histogram(binwidth = 2, fill = "red", color = "black", alpha = 0.7) +
    labs(title = paste("Distribucija dijastoličkog tlaka -", bmi_cat),
         x = "Dijastolički tlak (ap_lo)", y = "Frekvencija") +
    theme_minimal()

  # QQ plot za sistolički tlak (ap_hi)
  p3 <- ggplot(data_subset, aes(sample = ap_hi)) +
    stat_qq(color = "blue") +
```

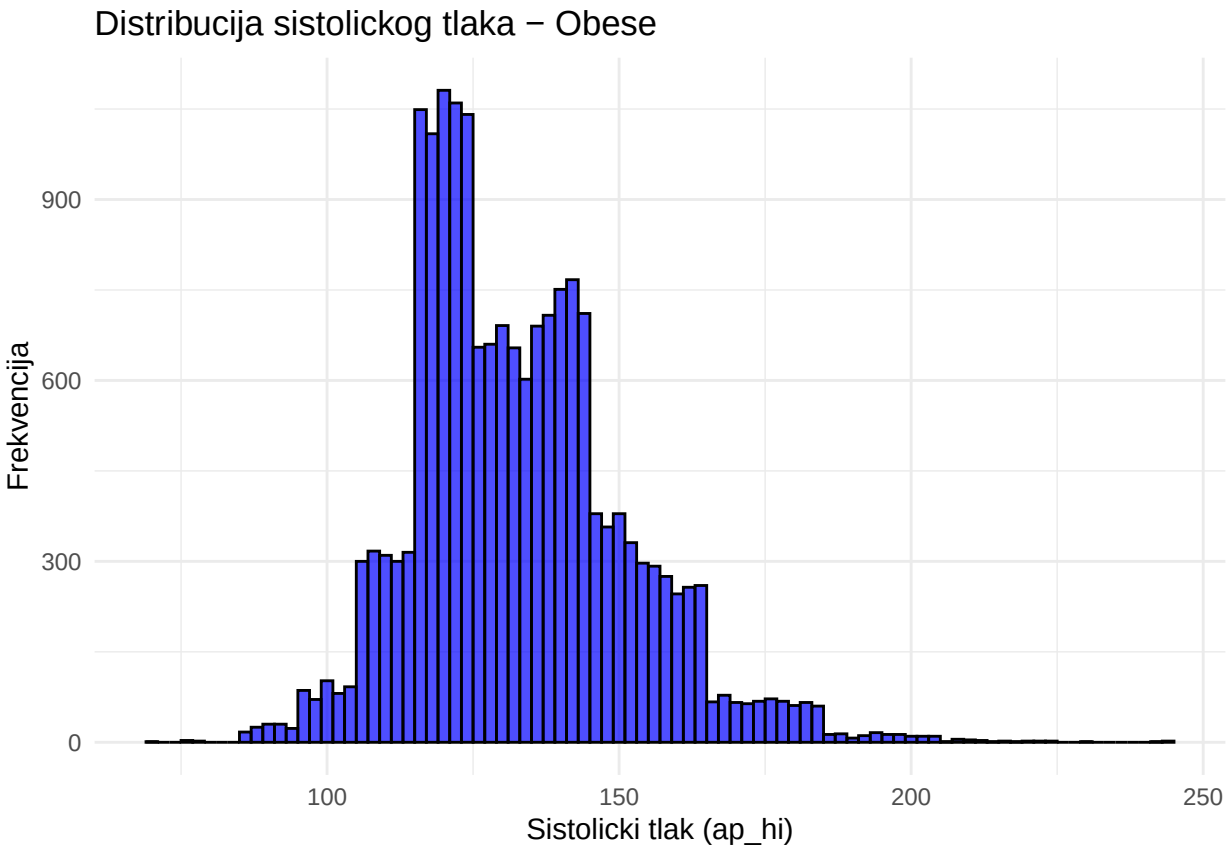
```
    stat_qq_line(color = "red") +
    labs(
      title = paste("Q-Q plot - Sistolicki tlak (ap_hi) -", bmi_cat),
      x = "Teorijske kvantile",
      y = "Empirijske kvantile"
    ) +
    theme_minimal()

  # QQ plot za dijastolički tlak (ap_lo)
  p4 <- ggplot(data_subset, aes(sample = ap_lo)) +
    stat_qq(color = "blue") +
    stat_qq_line(color = "red") +
    labs(
      title = paste("Q-Q plot - Dijastolički tlak (ap_lo) -", bmi_cat),
      x = "Teorijske kvantile",
      y = "Empirijske kvantile"
    ) +
    theme_minimal()

  # Prikaz grafova
  print(p1)
  print(p2)
  print(p3)
  print(p4)
}
```

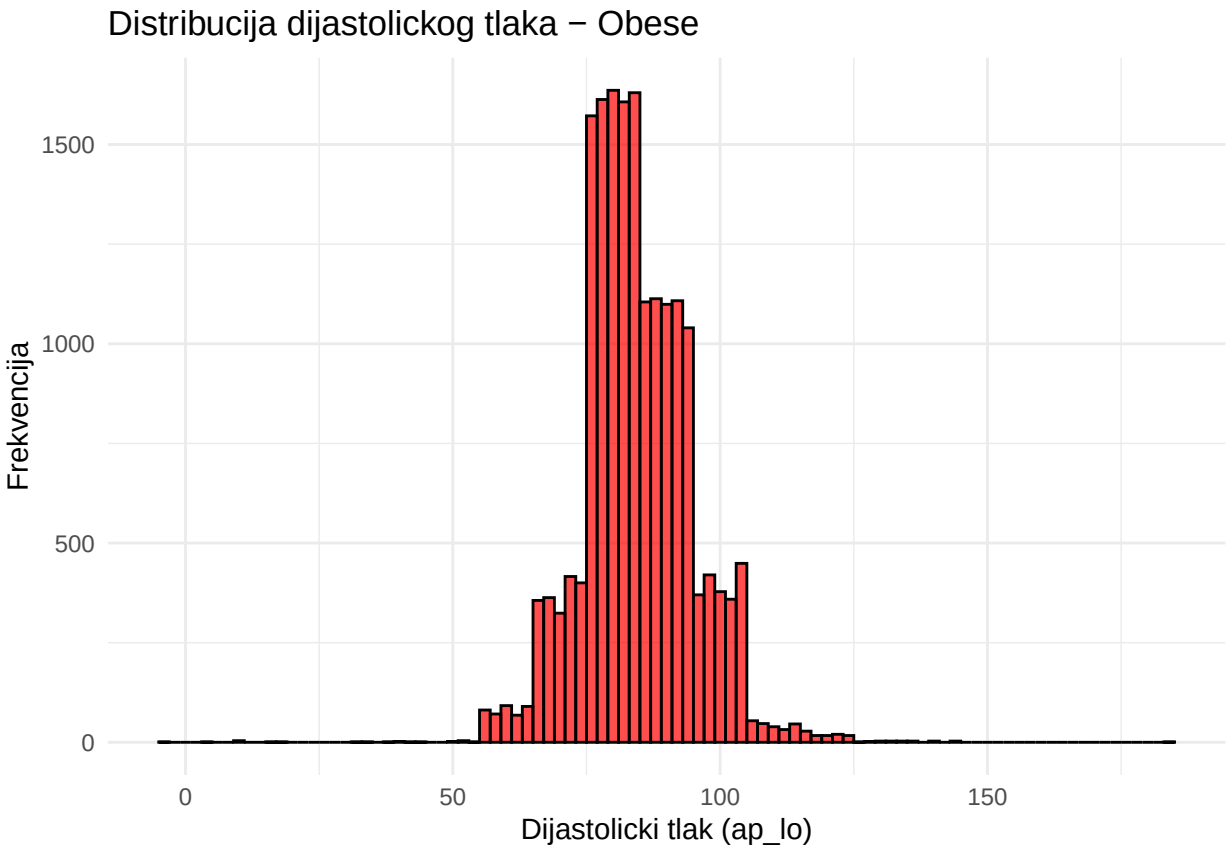


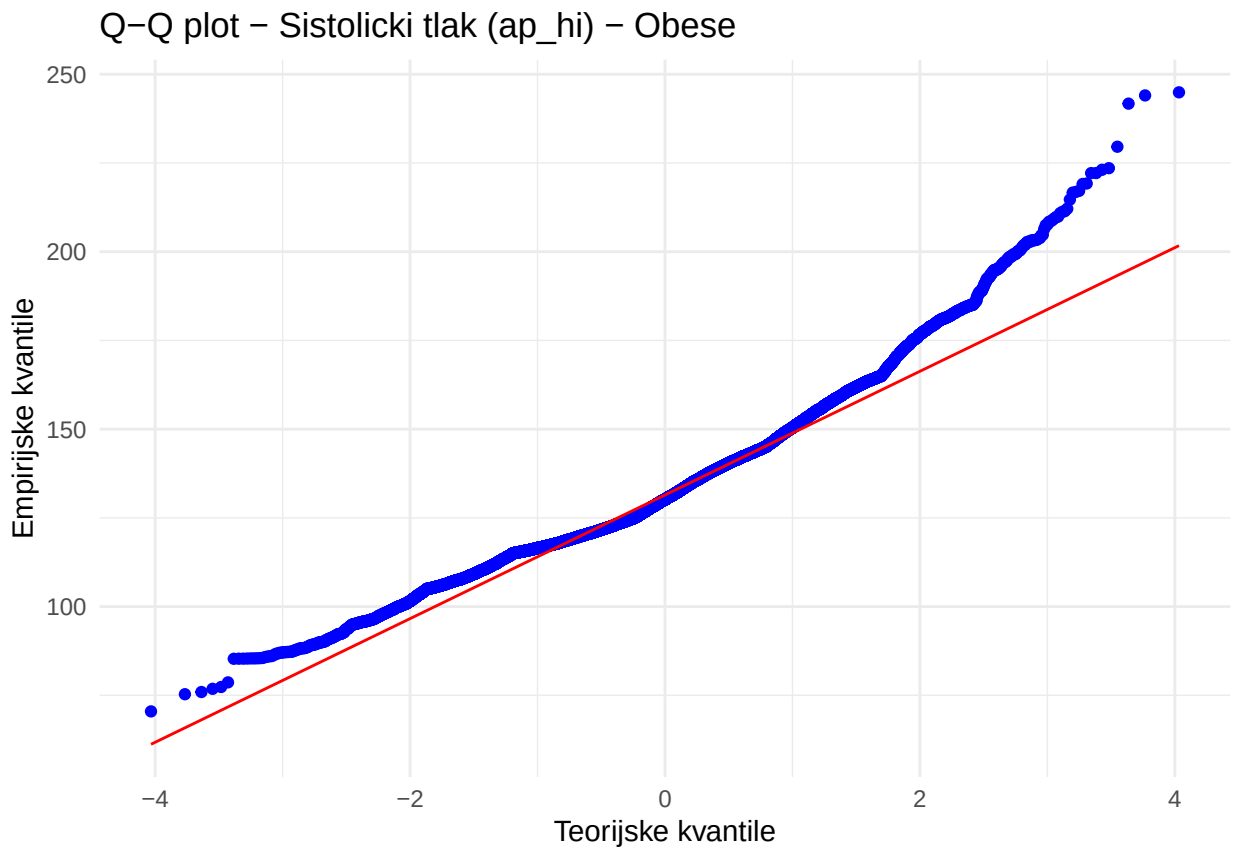

Distribucija dijastolickog tlaka − Normal
Frekvencija
2000
1000
0
0
50
100
150
Dijastolicki tlak (ap_lo)

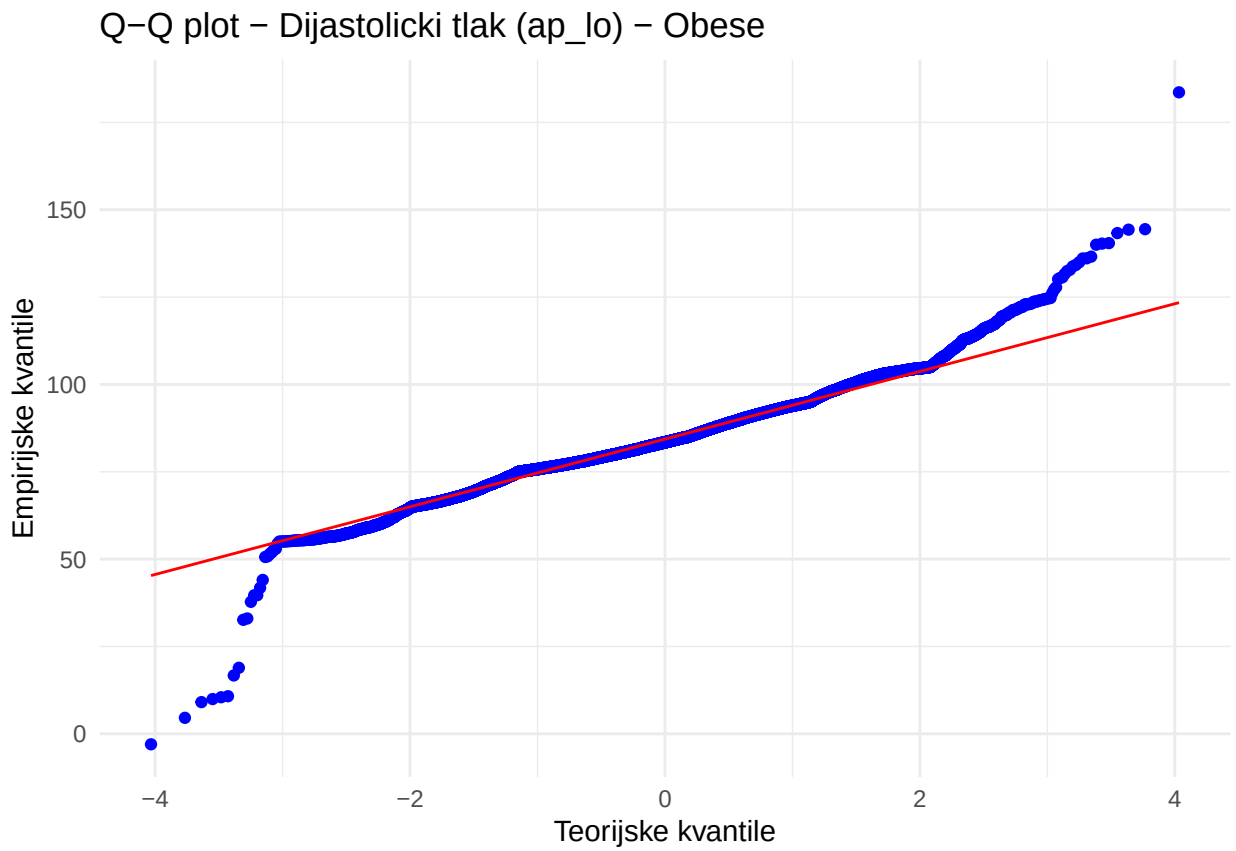

Q−Q plot − Sistolicki tlak (ap_hi) − Normal
Empirijske kvantile
200
150
100
−2.5
0.0
2.5
Teorijske kvantile

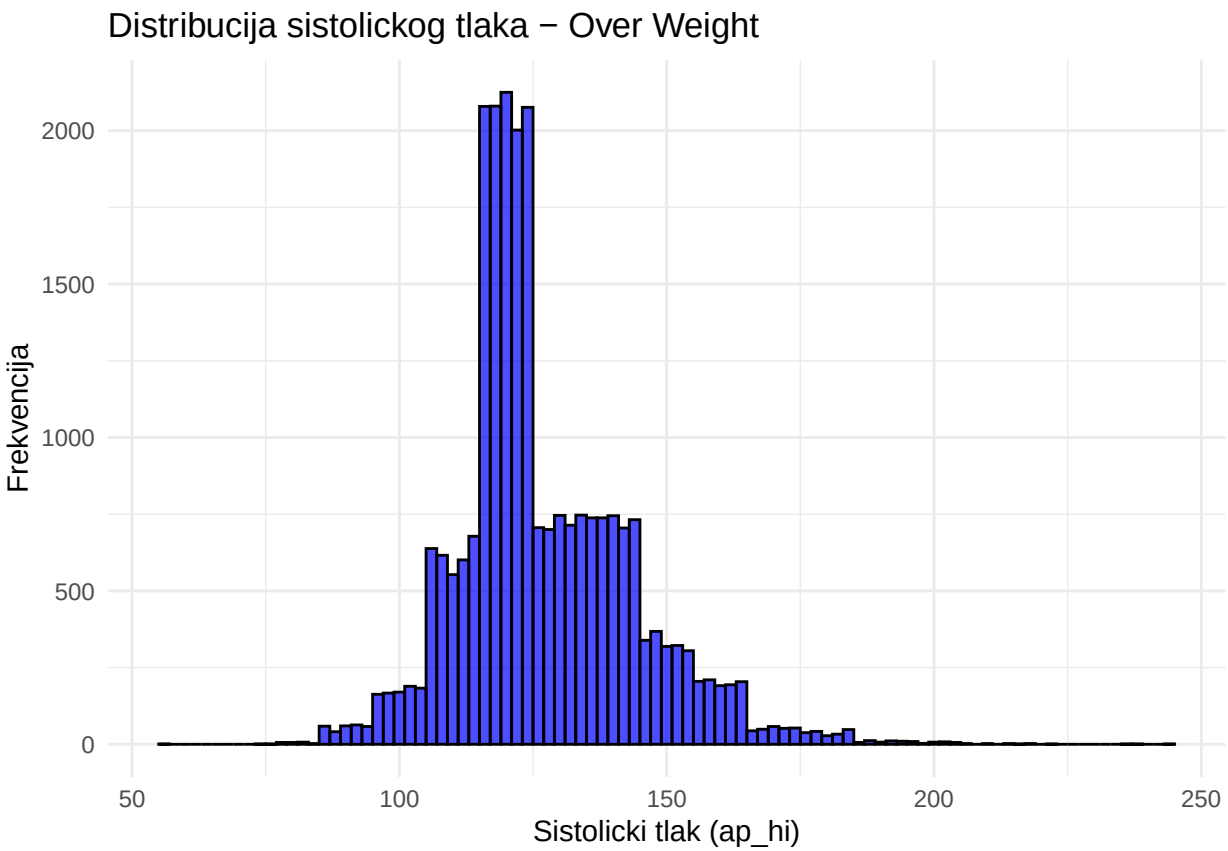

Q−Q plot − Dijastolicki tlak (ap_lo) − Normal
Empirijske kvantile
150
100
50
0
−2.5
0.0
2.5
Teorijske kvantile

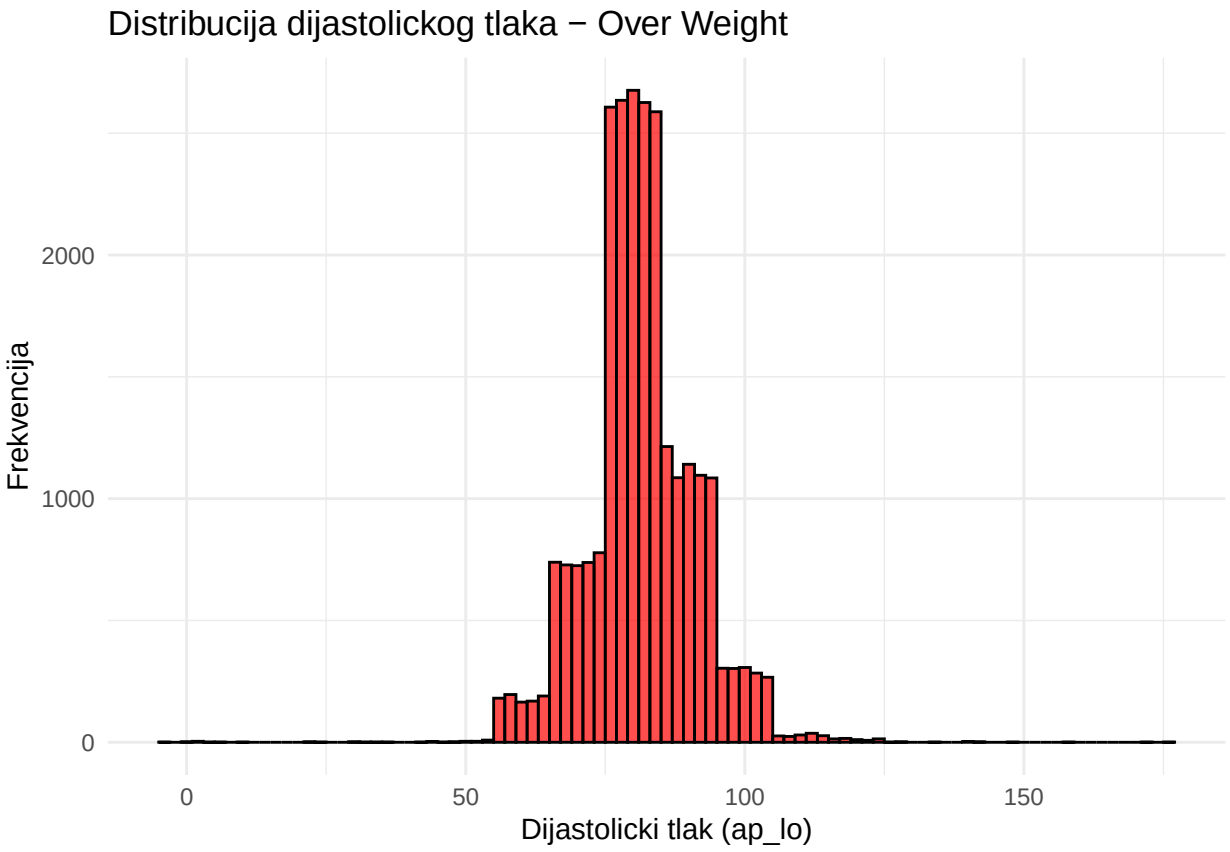

Distribucija sistolickog tlaka − Obese
Frekvencija
900
600
300
0
100
150
200
250
Sistolicki tlak (ap_hi)

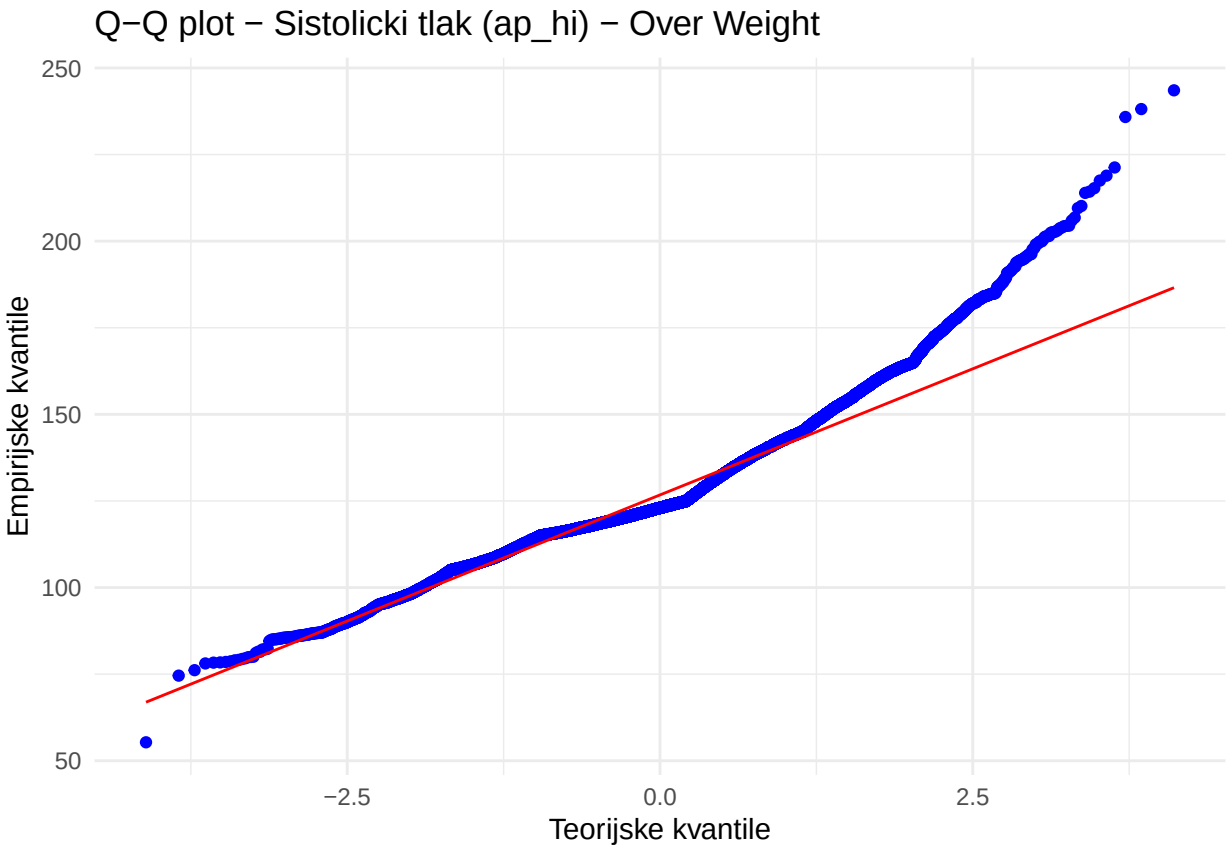

Distribucija dijastolickog tlaka − Obese
Frekvencija
1500
1000
500
0
0
50
100
150
Dijastolicki tlak (ap_lo)

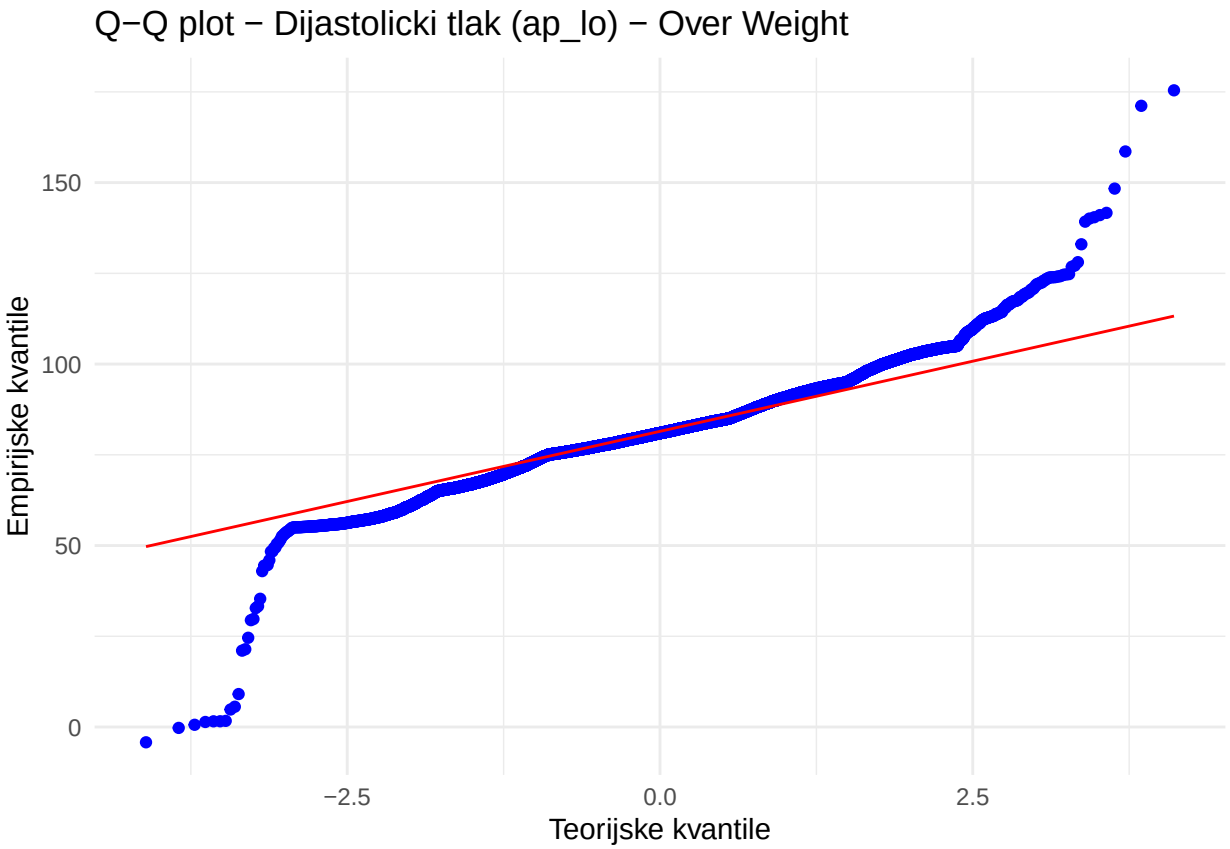

Q−Q plot − Sistolicki tlak (ap_hi) − Obese
Empirijske kvantile
Teorijske kvantile
250
200
150
100
−4
−2
0
2
4

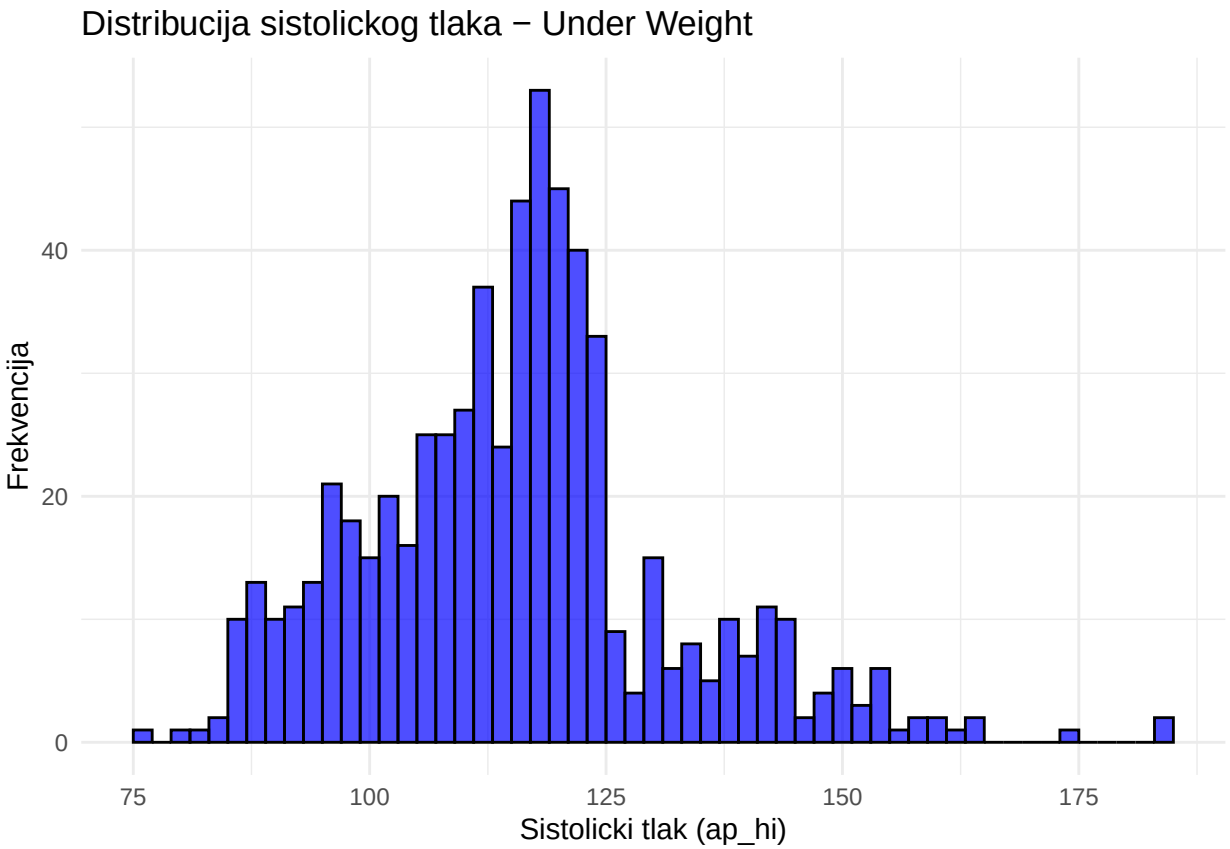

Q−Q plot − Dijastolicki tlak (ap_lo) − Obese
Empirijske kvantile
Teorijske kvantile
150
100
50
0
−4
−2
0
2
4

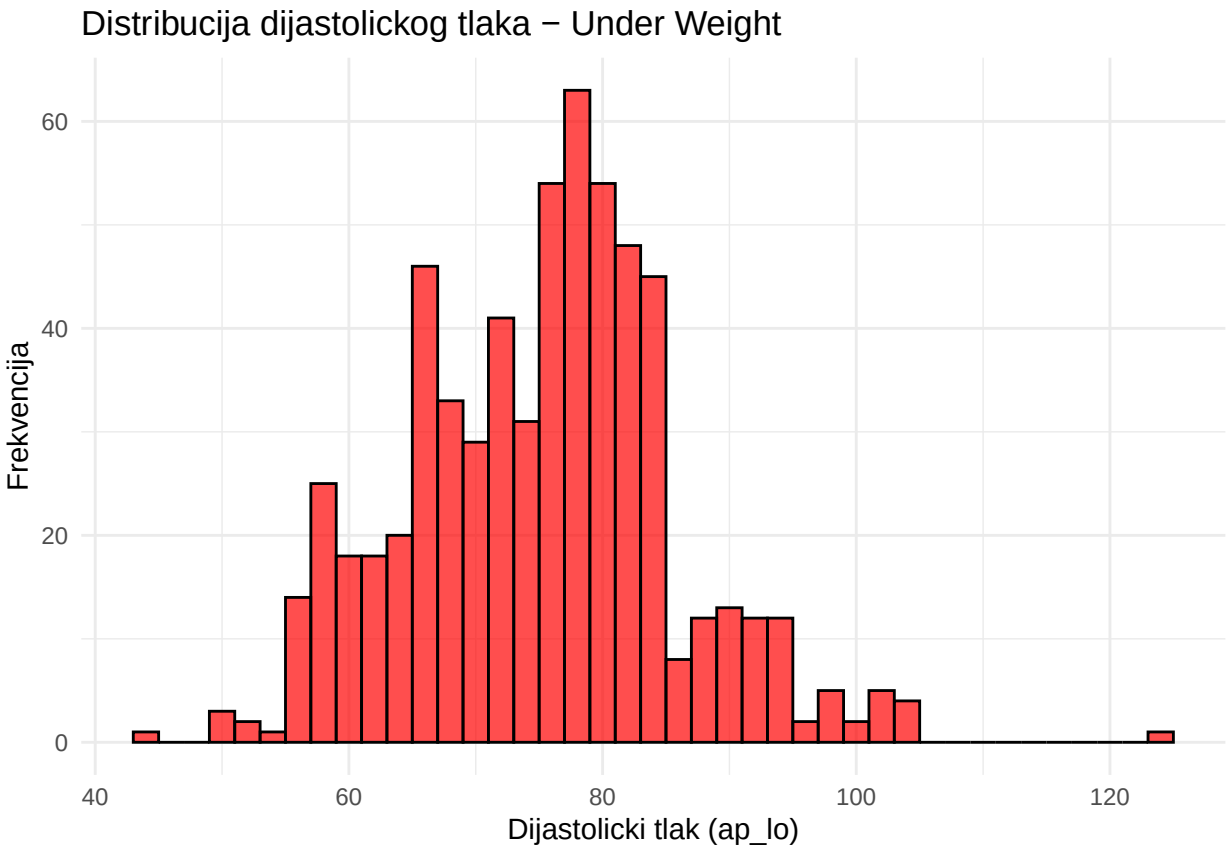

Distribucija sistolickog tlaka − Over Weight
Frekvencija
2000
1500
1000
500
0
50
100
150
200
250
Sistolicki tlak (ap_hi)

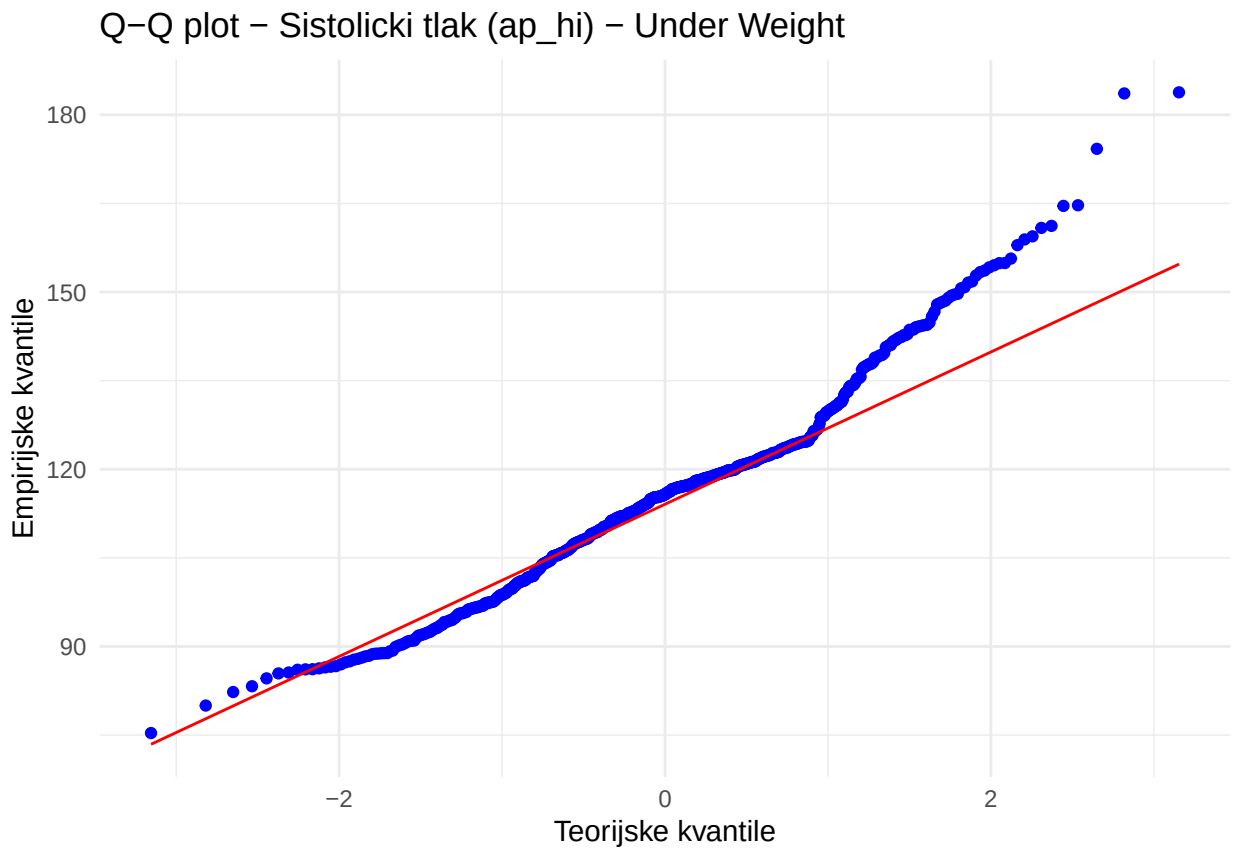

Distribucija dijastolickog tlaka − Over Weight
Frekvencija
2000
1000
0
0
50
100
150
Dijastolicki tlak (ap_lo)

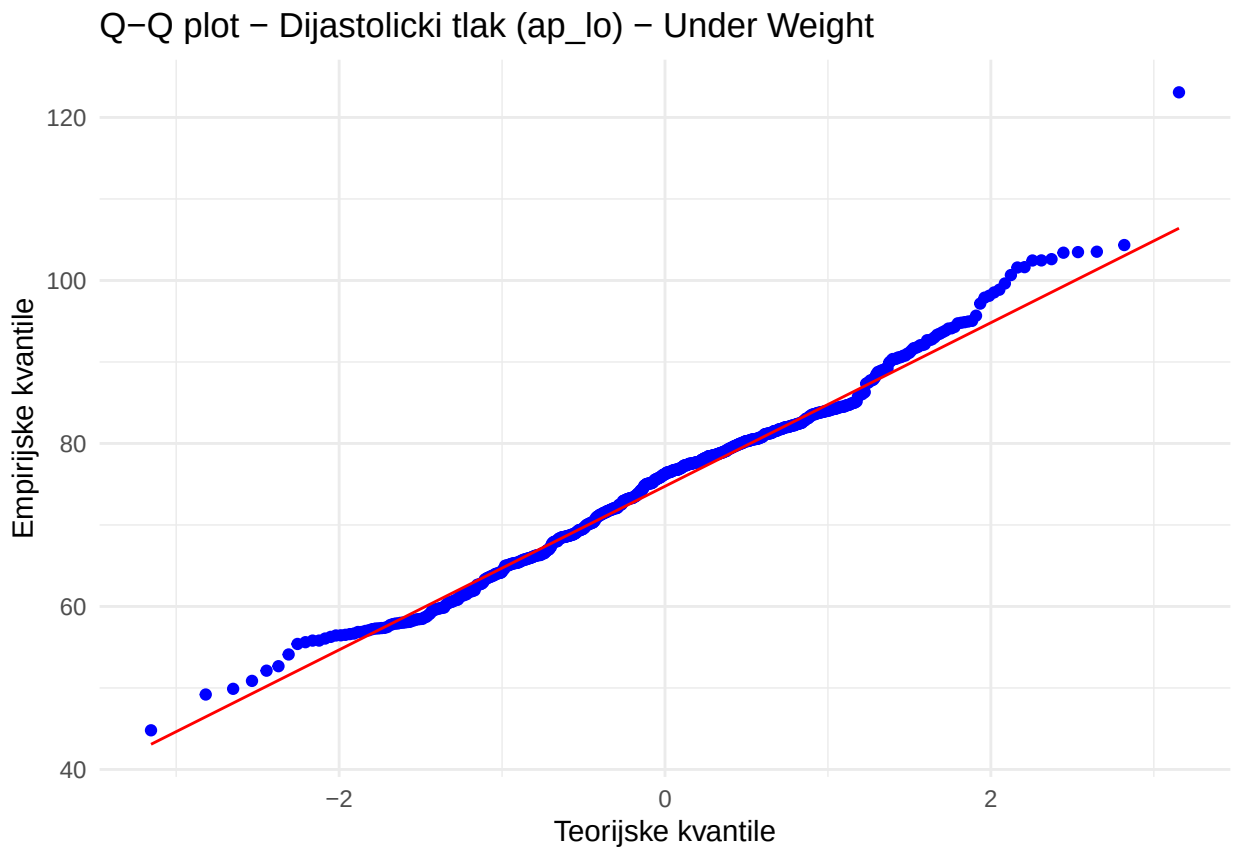

Q−Q plot − Sistolicki tlak (ap_hi) − Over Weight
250
200
150
100
50
Empirijske kvantile
−2.5
0.0
2.5
Teorijske kvantile


Q−Q plot − Dijastolicki tlak (ap_lo) − Over Weight
Empirijske kvantile
150
100
50
0
−2.5
0.0
2.5
Teorijske kvantile


Distribucija sistolickog tlaka − Under Weight
Frekvencija
0
20
40
75
100
125
150
175
Sistolicki tlak (ap_hi)


Distribucija dijastolickog tlaka − Under Weight
Frekvencija
0
20
40
60
40
60
80
100
120
Dijastolicki tlak (ap_lo)


Q−Q plot − Sistolicki tlak (ap_hi) − Under Weight
Empirijske kvantile
180
150
120
90
−2
0
2
Teorijske kvantile




```
# 1) Test normalnosti po grupama (Kolmogorov-Smirnov)
bmi_categories <- unique(filtered_data$BMICat)

for (bmi_cat in bmi_categories) {

  data_subset <- filtered_data %>%
    filter(BMICat == bmi_cat)

  cat("=====================================================\n")
  cat("BMI Category:", bmi_cat, "\n")
  cat("Broj zapisa u ovoj kategoriji:", nrow(data_subset), "\n")

  # Kolmogorov-Smirnov test za ap_hi
  ks_hi <- ks.test(
    data_subset$ap_hi,
    "pnorm",
    mean = mean(data_subset$ap_hi),
    sd   = sd(data_subset$ap_hi)
  )
  cat("\n>> Kolmogorov-Smirnov test - ap_hi <<\n")
  cat("  p-value:", ks_hi$p.value, "\n")

  # Kolmogorov-Smirnov test za ap_lo
  ks_lo <- ks.test(
    data_subset$ap_lo,
    "pnorm",
```

```
    mean = mean(data_subset$ap_lo),
    sd   = sd(data_subset$ap_lo)
  )
  cat("\n>> Kolmogorov-Smirnov test - ap_lo <<\n")
  cat("  p-value:", ks_lo$p.value, "\n\n")
}

cat("=================================================\n")
cat("     Test homoskedasticnosti (BartlettTest)\n")
cat("=================================================\n")

# Bartlettov test za ap_hi
bartlett_hi <- bartlett.test(ap_hi ~ BMICat, data = filtered_data)
cat("\nBartlett test: ap_hi ~ BMICat\n")
print(bartlett_hi)

# Bartlettov test za ap_lo
bartlett_lo <- bartlett.test(ap_lo ~ BMICat, data = filtered_data)
cat("\nBartlett test: ap_lo ~ BMICat\n")
print(bartlett_lo)
```

```
## ================================================
## BMI Category: Normal
## Broj zapisa u ovoj kategoriji: 24885
##
## >> Kolmogorov-Smirnov test - ap_hi <<
##   p-value: 0
##
## >> Kolmogorov-Smirnov test - ap_lo <<
##   p-value: 0
##
## ================================================
## BMI Category: Obese
## Broj zapisa u ovoj kategoriji: 18121
##
## >> Kolmogorov-Smirnov test - ap_hi <<
##   p-value: 0
##
## >> Kolmogorov-Smirnov test - ap_lo <<
##   p-value: 0
##
## ================================================
## BMI Category: Over Weight
## Broj zapisa u ovoj kategoriji: 25090
##
## >> Kolmogorov-Smirnov test - ap_hi <<
##   p-value: 0
##
## >> Kolmogorov-Smirnov test - ap_lo <<
##   p-value: 0
##
## ================================================
## BMI Category: Under Weight
## Broj zapisa u ovoj kategoriji: 622
```

```
##
## >> Kolmogorov-Smirnov test - ap_hi <<
##   p-value: 7.520888e-06
##
## >> Kolmogorov-Smirnov test - ap_lo <<
##   p-value: 0.06618708
##
## ================================================
##      Test homoskedasticnosti (BartlettTest)
## ================================================
##
## Bartlett test: ap_hi ~ BMICat
##
##  Bartlett test of homogeneity of variances
##
## data:  ap_hi by BMICat
## Bartlett's K-squared = 890.75, df = 3, p-value < 2.2e-16
##
##
## Bartlett test: ap_lo ~ BMICat
##
##  Bartlett test of homogeneity of variances
##
## data:  ap_lo by BMICat
## Bartlett's K-squared = 251.8, df = 3, p-value < 2.2e-16
```

```r
# Kruskal-Wallis za sistolički tlak (ap_hi)
kruskal_hi <- kruskal.test(ap_hi ~ BMICat, data = filtered_data)
cat("----- Kruskal-Wallis Test za ap_hi -----\n")
print(kruskal_hi)

# Kruskal-Wallis za dijastolički tlak (ap_lo)
kruskal_lo <- kruskal.test(ap_lo ~ BMICat, data = filtered_data)
cat("\n----- Kruskal-Wallis Test za ap_lo -----\n")
print(kruskal_lo)
```

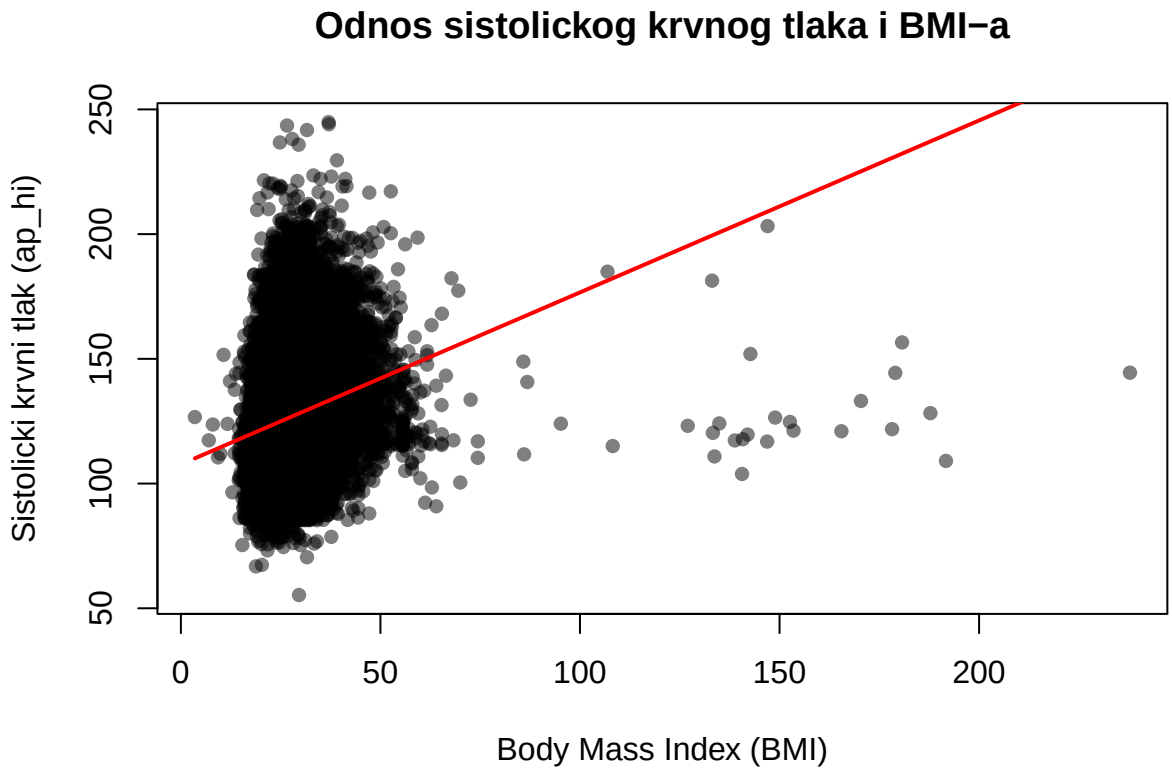
```
## ----- Kruskal-Wallis Test za ap_hi -----
##
##  Kruskal-Wallis rank sum test
##
## data:  ap_hi by BMICat
## Kruskal-Wallis chi-squared = 4363.9, df = 3, p-value < 2.2e-16
##
##
## ----- Kruskal-Wallis Test za ap_lo -----
##
##  Kruskal-Wallis rank sum test
##
## data:  ap_lo by BMICat
## Kruskal-Wallis chi-squared = 3133.8, df = 3, p-value < 2.2e-16
```

#Zadatak 4

Želimo odgovoriti na sljedeće pitanje: “Kakav je odnos izmedu BMI-a i krvnog tlaka te možemo li predvidjeti krvni tlak na temelju BMI-a, dobi i učestalosti tjelesne aktivnosti?”

Sada ćemo metodom najmanjih kvadrata pokušati uspostaviti vezu između BMI-a i krvnog tlaka.

```
fit.ap_hi <- lm(ap_hi ~ poly(BMI,1) , data = filtered_data)

plot(filtered_data$BMI, filtered_data$ap_hi,
     main = "Odnos sistoličkog krvnog tlaka i BMI-a",
     xlab = "Body Mass Index (BMI)",
     ylab = "Sistolički krvni tlak (ap_hi)",
     pch = 16, col = rgb(0, 0, 0, 0.5))


sorted_index <- order(filtered_data$BMI)
lines(filtered_data$BMI[sorted_index],
      fit.ap_hi$fitted.values[sorted_index],
      col = "red", lwd = 2)
```




```
summary(fit.ap_hi)
```

```
##
## Call:
## lm(formula = ap_hi ~ poly(BMI, 1), data = filtered_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -130.765   -9.598   -3.076    8.972  117.504
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.267e+02  6.275e-02 2018.42   <2e-16 ***
## poly(BMI, 1) 1.054e+03  1.645e+01   64.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.45 on 68716 degrees of freedom
## Multiple R-squared:  0.05637,    Adjusted R-squared:  0.05635
## F-statistic:  4105 on 1 and 68716 DF,  p-value: < 2.2e-16
```

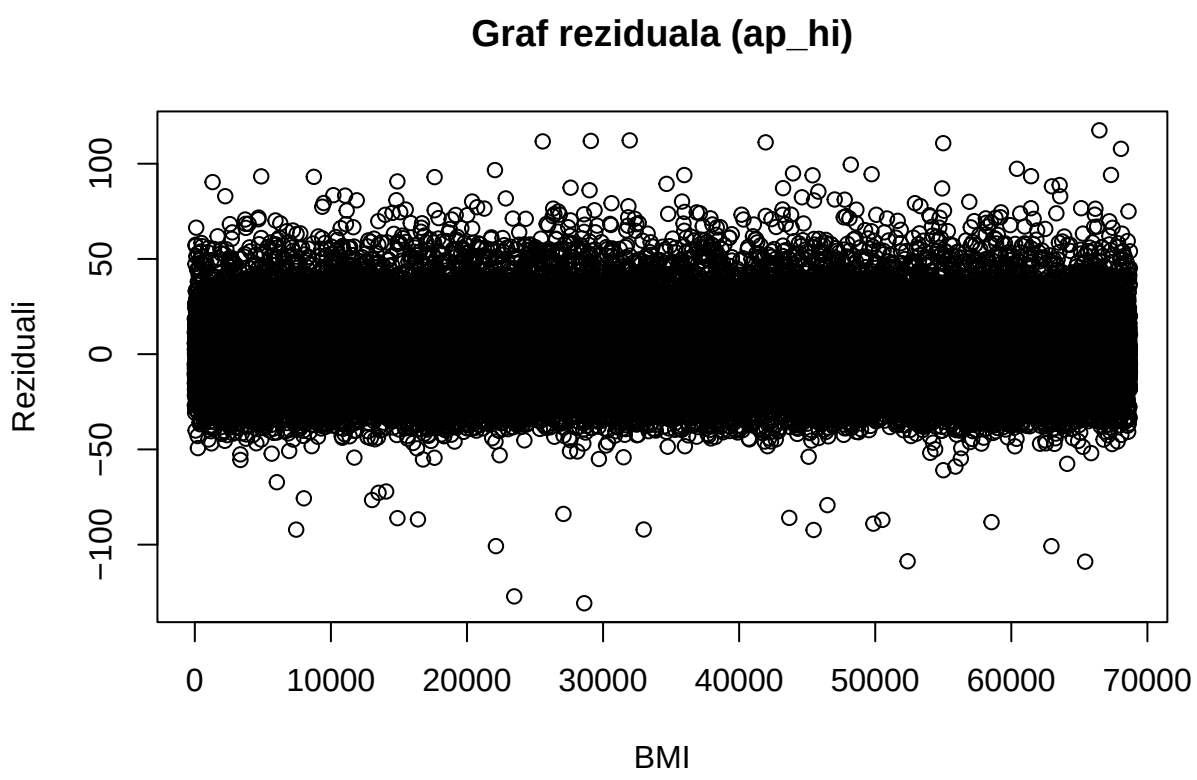
Iznad možemo vidjeti graf raspršenja između sistoličkog tlaka i BMI-a kao i pravac linearne regresije koji smo izračunali iz podataka. Pokušavali smo linearnu regresiju s polinomima viših stupnjeva, ali su svi stupnjevi bili veoma slični pravcima i nisu poboljšavali vrijednost R^2. Zbog toga smo dali prednost najjednostavnijem modelu, a to je naravno pravac. Vidimo blagi pozitivan trend, ali se iz p vrijednosti vidi da je značajnost regresora skoro pa zanemariva. Također, R^2 vrijednost je 0.0564 (R^2_{adj} je 0.05638) što ukazuje na loš fit modela, no mi ćemo svakako sada nastaviti s analizom reziduala.

```
standardized_residuals <- rstandard(fit.ap_hi)
ks_test <- ks.test(standardized_residuals, "pnorm")
print(ks_test)

require(nortest)
lilliefors_test <- lillie.test(standardized_residuals)
print(lilliefors_test)
```

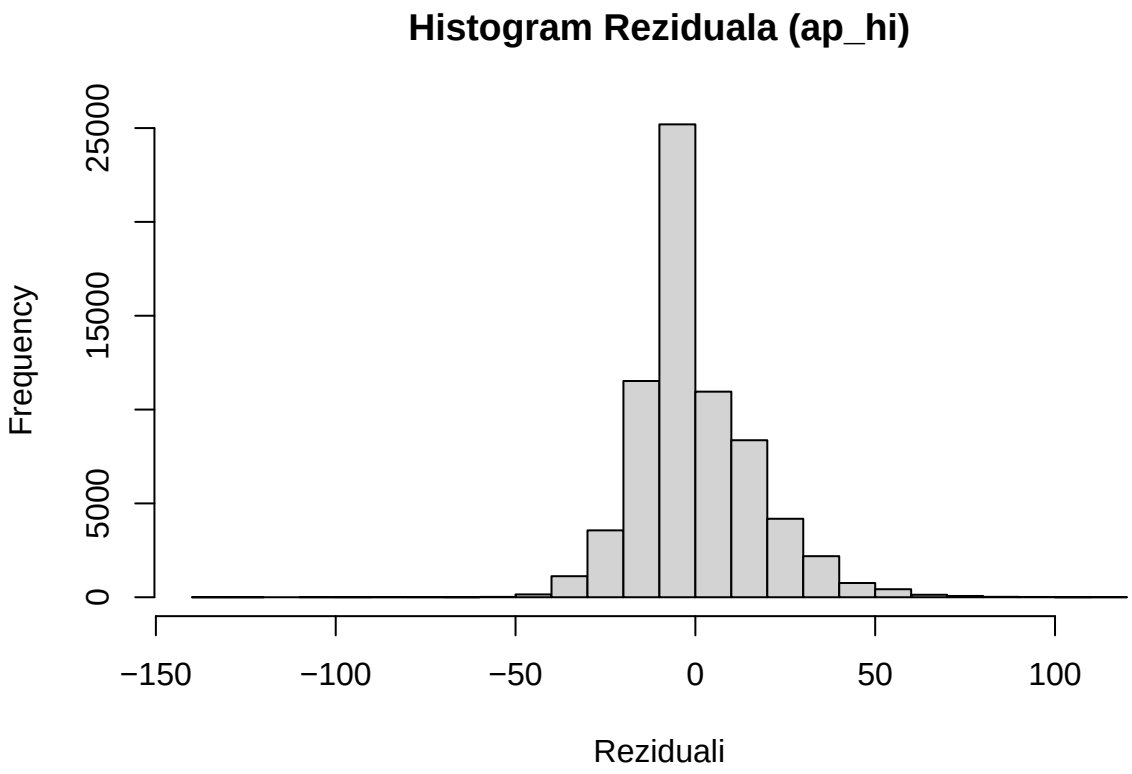
```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  standardized_residuals
## D = 0.10822, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  standardized_residuals
## D = 0.10821, p-value < 2.2e-16
```

Reziduali padaju na testu normalnosti (i KS i Lilliefors). Možemo također vidjeti kako se reziduali ponašaju grafički:

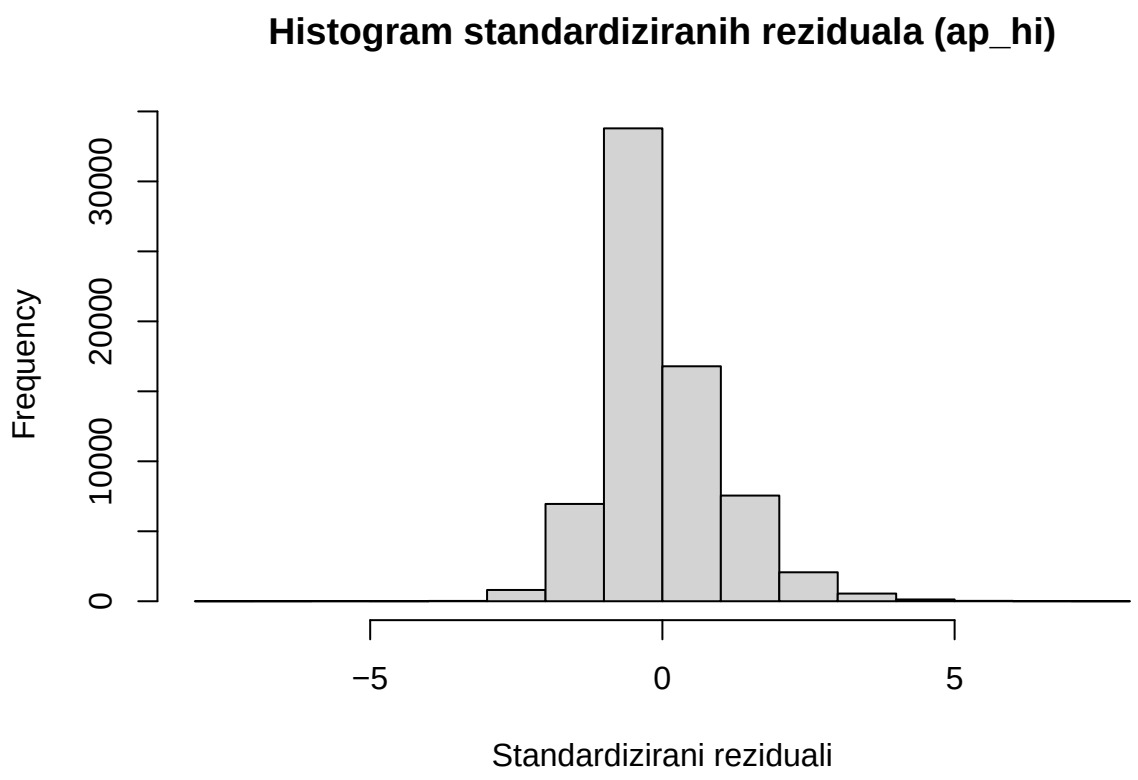
```
plot(fit.ap_hi$residuals,
     main = "Graf reziduala (ap_hi)",
     ylab = "Reziduali", xlab = "BMI")
```



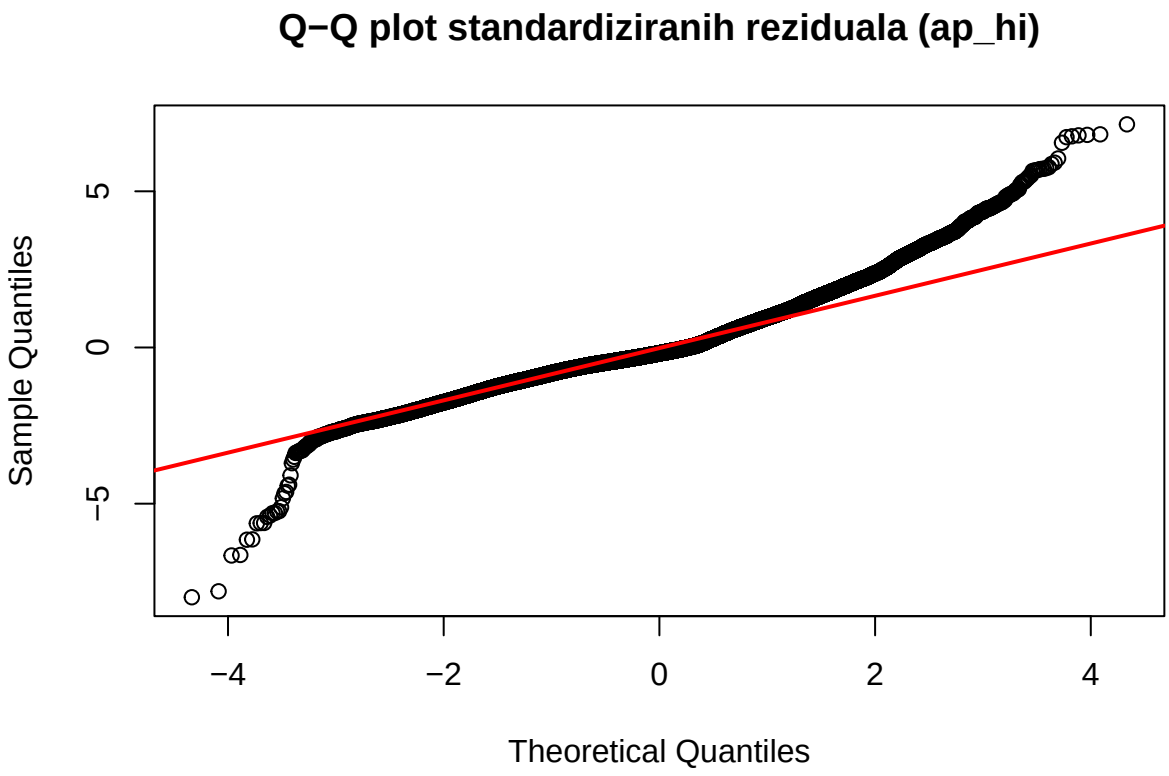

```
hist(fit.ap_hi$residuals,
     breaks = 20,
     main = "Histogram Reziduala (ap_hi)",
     xlab = "Reziduali")
```



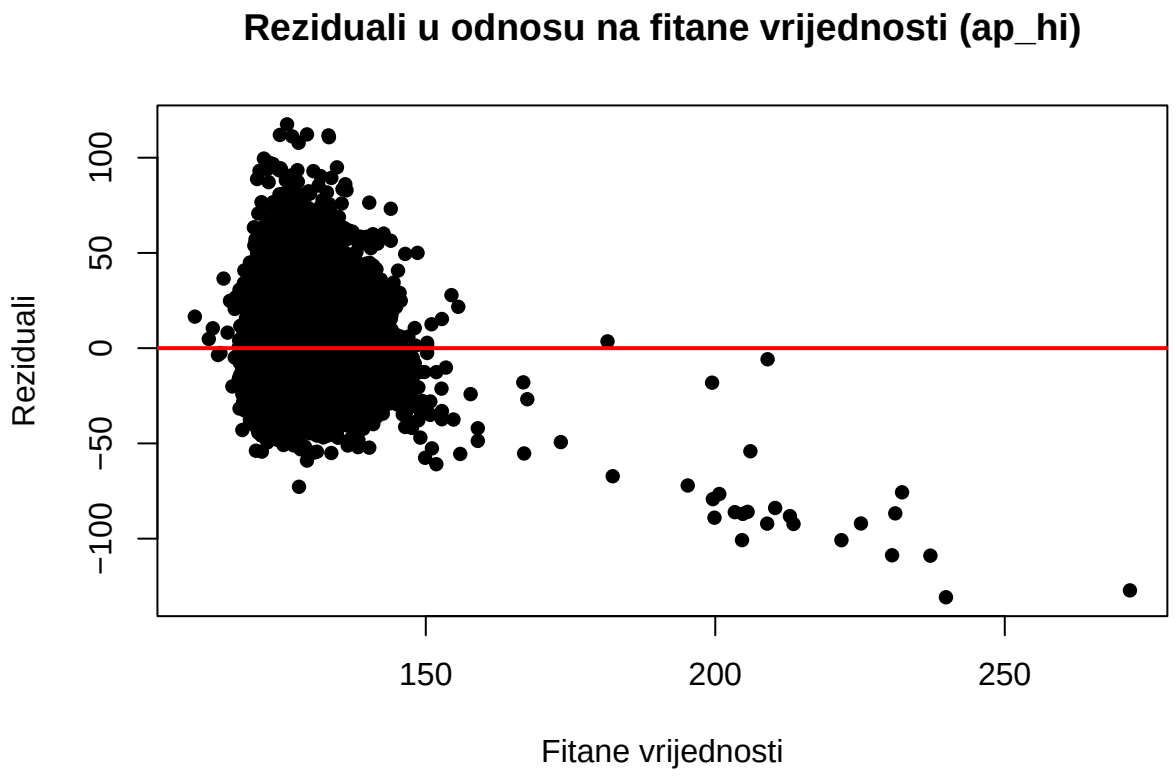

```
hist(rstandard(fit.ap_hi),
     breaks = 20,
     main = "Histogram standardiziranih reziduala (ap_hi)",
     xlab = "Standardizirani reziduali")
```




```
qqnorm(rstandard(fit.ap_hi),
       main = "Q-Q plot standardiziranih reziduala (ap_hi)")
qqline(rstandard(fit.ap_hi), col = "red", lwd = 2)
```



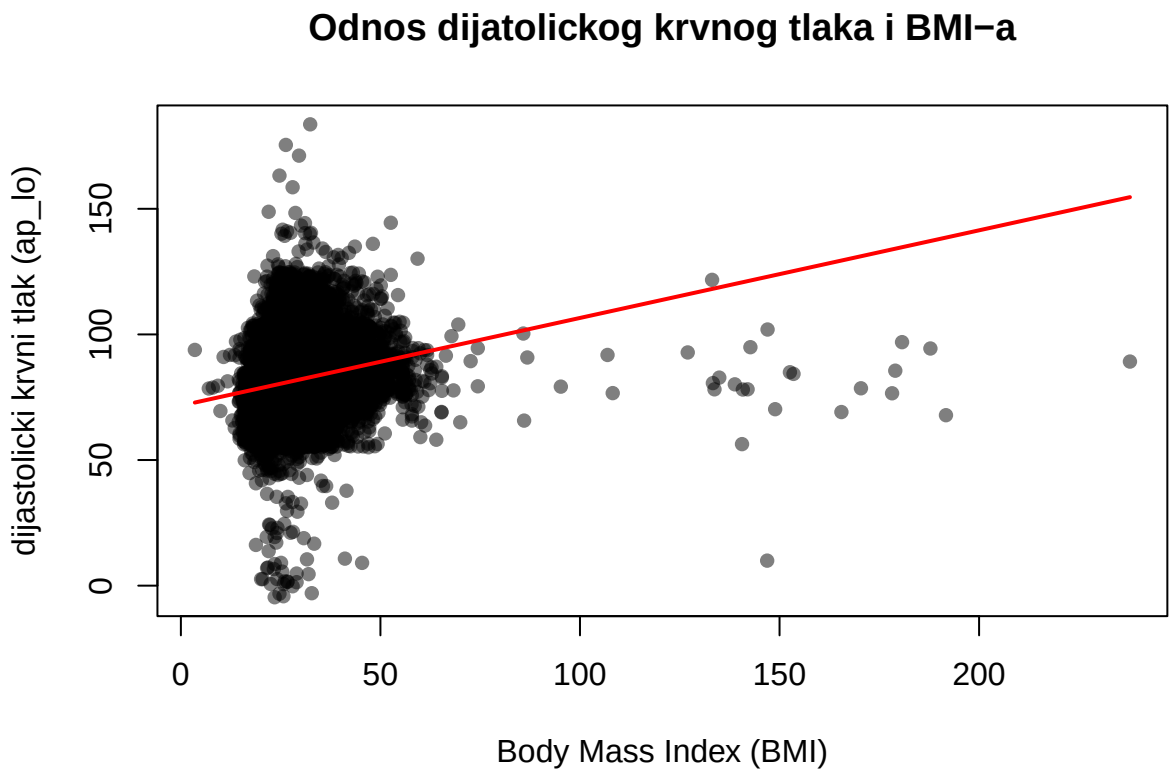

```r
plot(fit.ap_hi$fitted.values, fit.ap_hi$residuals,
     main = "Reziduali u odnosu na fitane vrijednosti (ap_hi)",
     xlab = "Fitane vrijednosti", ylab = "Reziduali", pch = 16)
abline(h = 0, col = "red", lwd = 2)
```



Q-Q plot nam govori da ova razdioba ima lakše repove od normalne, ali ovo svakako nije normalna distribucija. Sada možemo zaključiti da je nemoguće predvidjeti sistolički krvni tlak iz BMI-a (iz ovih podataka).

Za dijastolički krvni tlak ponavljamo isti postupak.

```
fit.ap_lo <- lm(ap_lo ~ poly(BMI, 1) , data = filtered_data)
plot(filtered_data$BMI, filtered_data$ap_lo,
     main = "Odnos dijatoličkog krvnog tlaka i BMI-a",
     xlab = "Body Mass Index (BMI)",
     ylab = "dijastolički krvni tlak (ap_lo)",
     pch = 16, col = rgb(0, 0, 0, 0.5))


sorted_index <- order(filtered_data$BMI)
lines(filtered_data$BMI[sorted_index],
      fit.ap_lo$fitted.values[sorted_index],
      col = "red", lwd = 2)
```




```
summary(fit.ap_lo)
```

```
##
## Call:
## lm(formula = ap_lo ~ poly(BMI, 1), data = filtered_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -112.962   -5.197   -0.270    5.091  100.650
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   81.25884    0.03755 2164.02   <2e-16 ***
## poly(BMI, 1) 533.42601    9.84338   54.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.843 on 68716 degrees of freedom
## Multiple R-squared:  0.04099,    Adjusted R-squared:  0.04097
## F-statistic:  2937 on 1 and 68716 DF,  p-value: < 2.2e-16
```

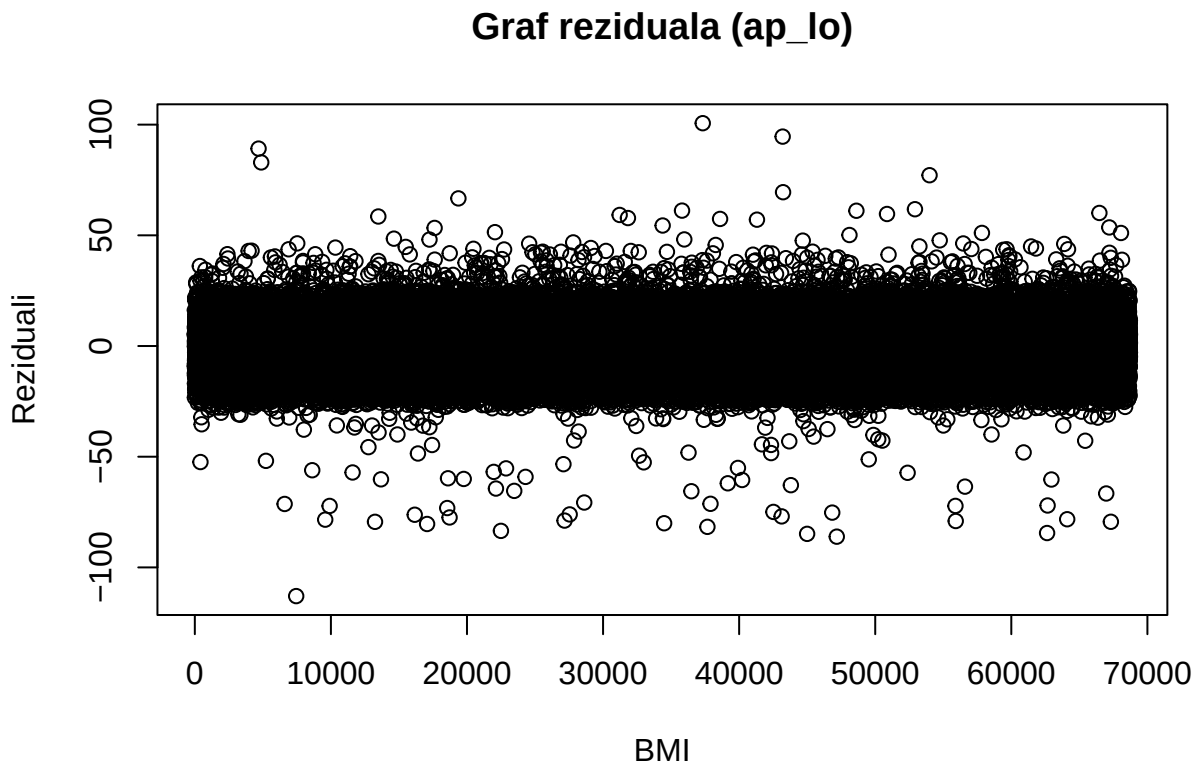
Zadržat ćemo model pravca iz istog razloga kao i za sistolički tlak. Vidi se blagi pozitivan trend, ali vidimo (iz p vrijednosti) da regresor ima jako malenu značajnost. Također, R^2 vrijednost je sada 0.04008 (R^2_{adj} je 0.04007) što opet ukazuje na loš fit modela, no mi ćemo svakako opet nastaviti s analizom reziduala. Analiza reziduala

```r
standardized_residuals <- rstandard(fit.ap_hi)

ks_test <- ks.test(standardized_residuals, "pnorm")
print(ks_test)

lilliefors_test <- lillie.test(standardized_residuals)
print(lilliefors_test)
```

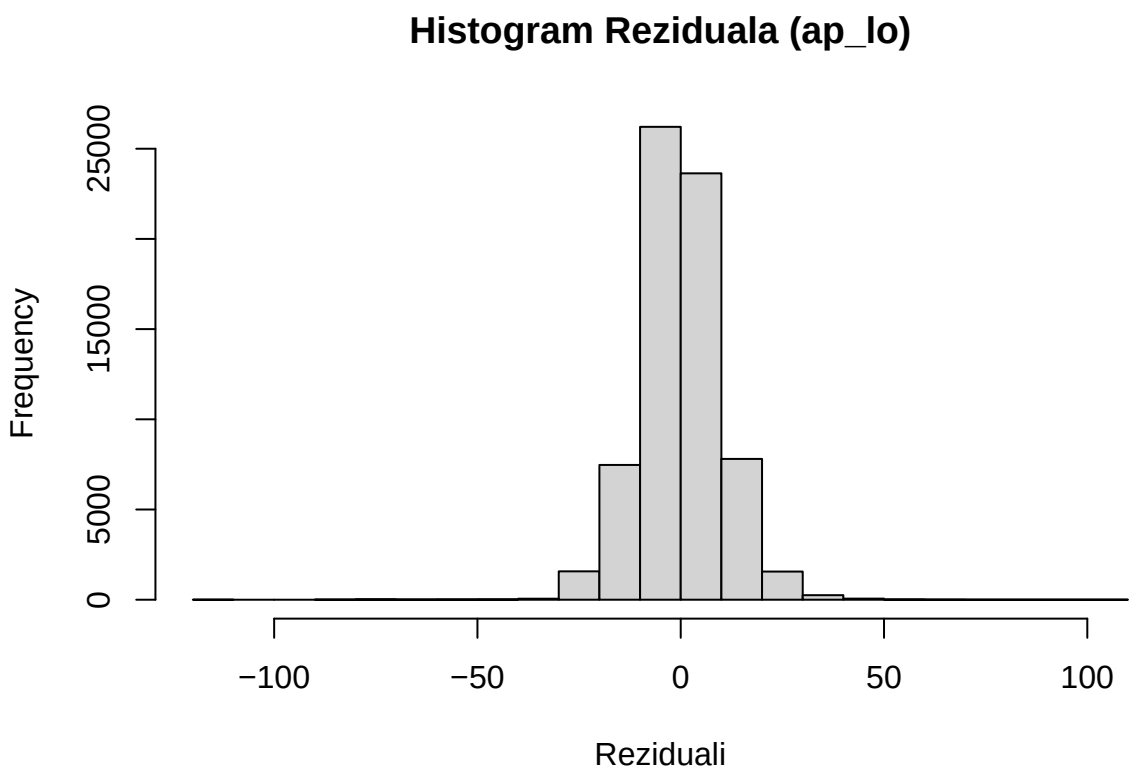
```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  standardized_residuals
## D = 0.10822, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  standardized_residuals
## D = 0.10821, p-value < 2.2e-16
```

Reziduali padaju na testu normalnosti (i KS i Lilliefors). Možemo također vidjeti kako se reziduali ponašaju grafički:

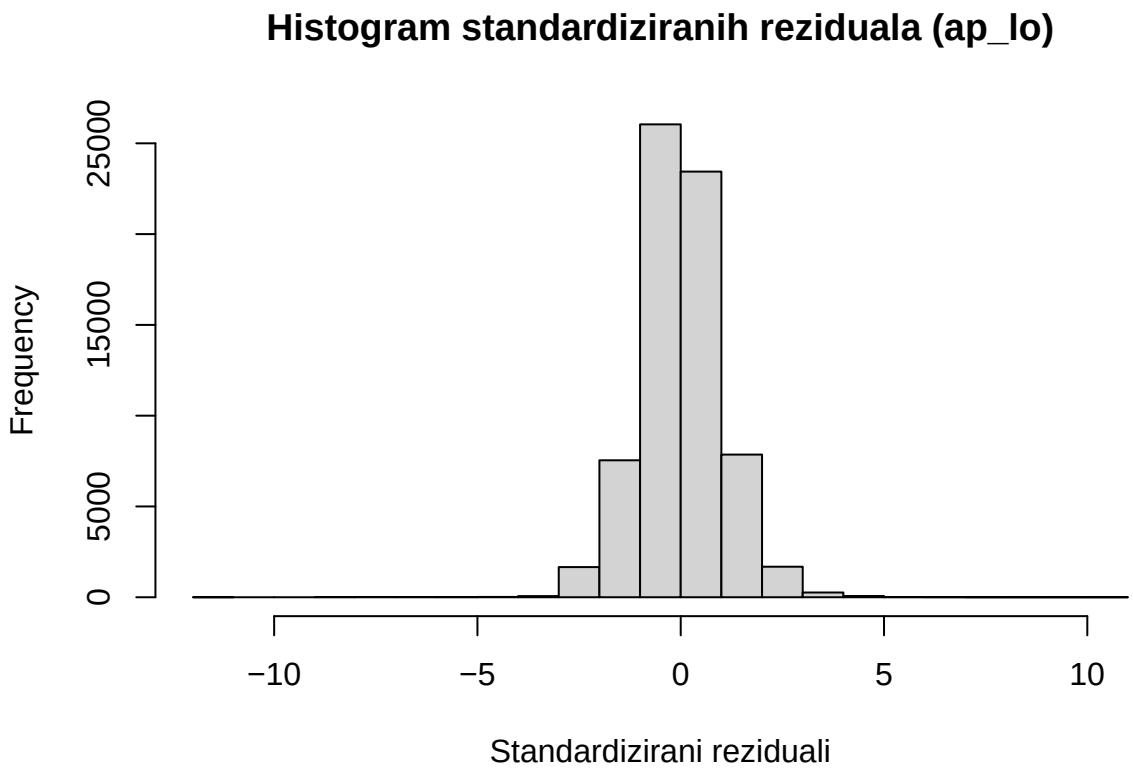
```r
plot(fit.ap_lo$residuals,
     main = "Graf reziduala (ap_lo)",
     ylab = "Reziduali", xlab = "BMI")
```



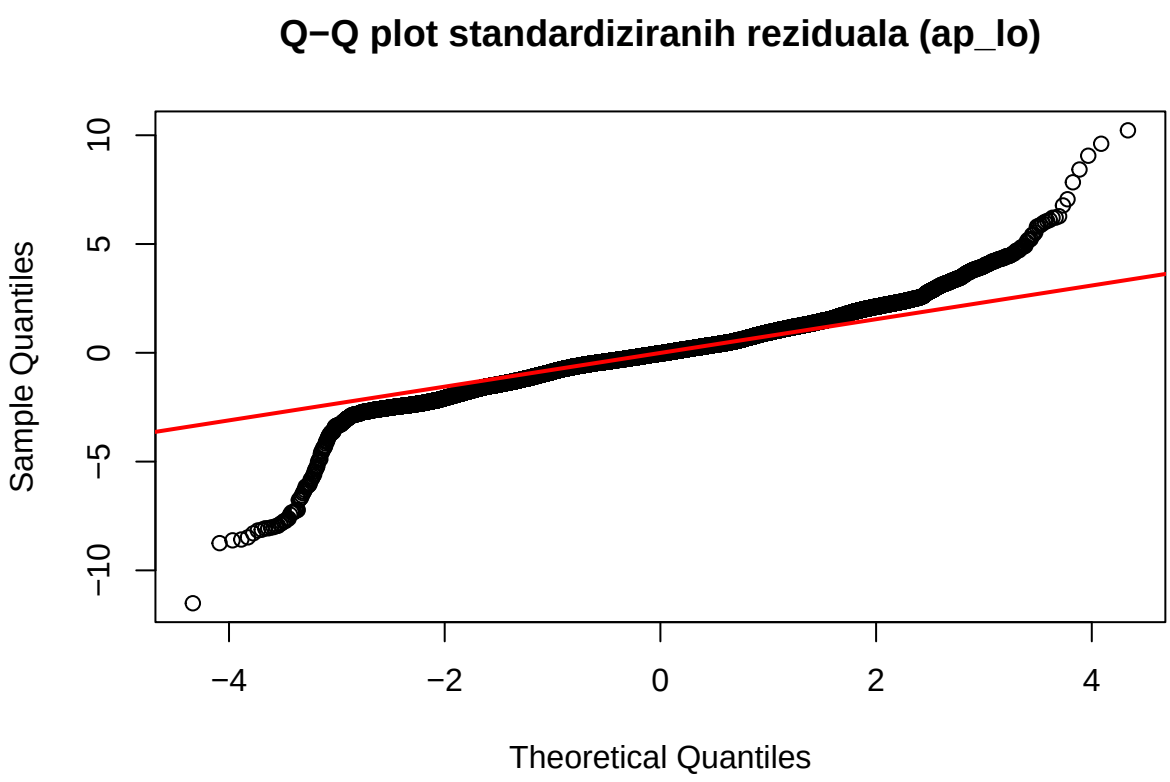
```r
hist(fit.ap_lo$residuals,
     breaks = 20,
     main = "Histogram Reziduala (ap_lo)",
     xlab = "Reziduali")
```



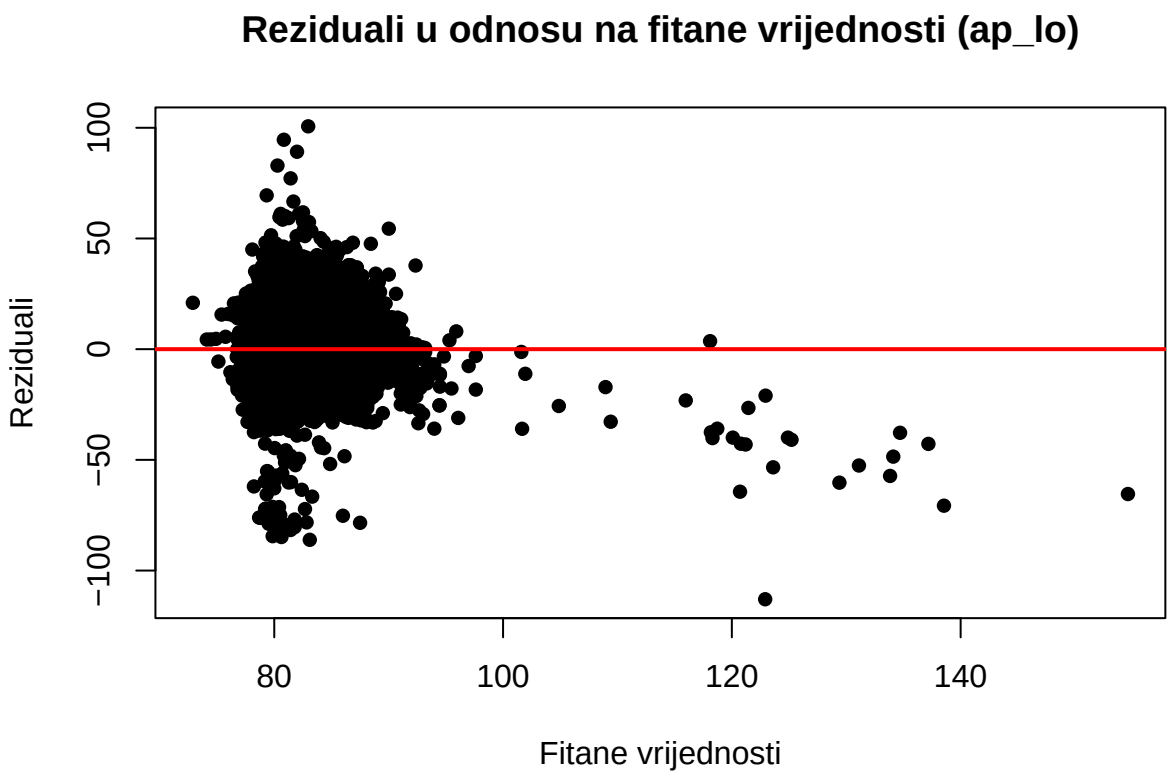

```r
hist(rstandard(fit.ap_lo),
     breaks = 20,
     main = "Histogram standardiziranih reziduala (ap_lo)",
     xlab = "Standardizirani reziduali")
```




```
qqnorm(rstandard(fit.ap_lo),
       main = "Q-Q plot standardiziranih reziduala (ap_lo)")
qqline(rstandard(fit.ap_lo), col = "red", lwd = 2)
```




```r
plot(fit.ap_lo$fitted.values, fit.ap_lo$residuals,
     main = "Reziduali u odnosu na fitane vrijednosti (ap_lo)",
     xlab = "Fitane vrijednosti", ylab = "Reziduali", pch = 16)
abline(h = 0, col = "red", lwd = 2)
```



Grafički možemo reći da reziduali imaju teže repove, ali se ne ponašaju baš pravino. Nemoguće je (na temelju ovih podataka) predvidjeti dijastolički krvni tlak iz BMI-a.

Obratimo sada pažnju na drugi dio problema: "Možemo li predvidjeti krvni tlak na temelju BMI-a, dobi i učestalosti tjelesne aktivnosti?"

Kada radimo na višestrukoj regrešiji želimo da nam regresori budu međusobno "dovoljno" nezavisni, inače ne možemo interpretirati rezultate. Stoga računamo kovarijancu za sve parove od BMI, starosti i tjelesne aktivnosti. NAPOMENA: S obzirom da je tjelesna aktivnost binarna kategorijska varijabla nije loša ideja staviti ju u model višestruke regresije.

```
cor(cbind(filtered_data$active,filtered_data$BMI,filtered_data$AgeinYr))
```

```
##             [,1]        [,2]        [,3]
## [1,]  1.00000000 -0.01566573 -0.01019637
## [2,] -0.01566573  1.00000000  0.08960381
## [3,] -0.01019637  0.08960381  1.00000000
```

Iz kovarijanci možemo zaključiti da su varijable "dovoljno" nezavisne. Veću zavisnost vidimo između BMI i starosti, što ima smisla jer kako starimo naša visina se toliko ne mijenja koliko naša masa, pa je normalno da će BMI ovisiti o starosti, no svakako možemo pretpostaviti nezavisnost i zbog toga što je najstarija osoba u uzorku ima 64 godina, što nije dovoljno staro da krene značajno odumiranje mišićnog tkiva.

```
fit.multi <- lm(ap_hi ~ BMI + active + AgeinYr, filtered_data) #ako maknete regresore koji su manje zna
#fit.multi = lm(ap_hi ~ AgeinYr + active, filtered_data)
summary(fit.multi)
```

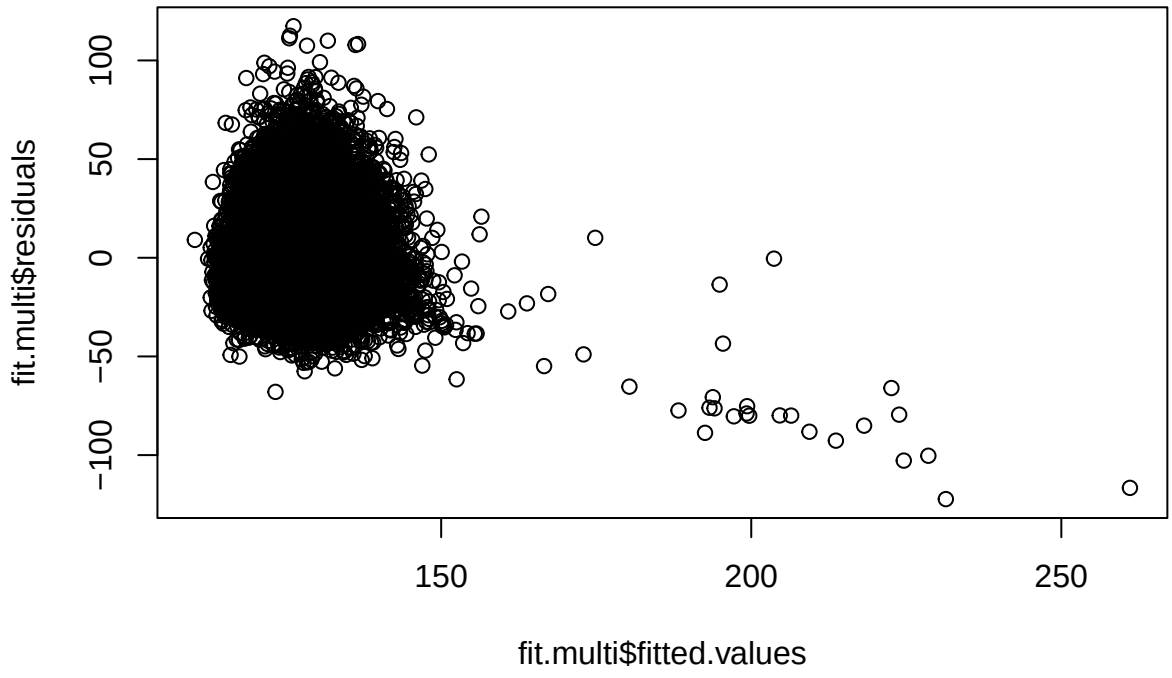
```
##
## Call:
## lm(formula = ap_hi ~ BMI + active + AgeinYr, data = filtered_data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -122.296   -9.869   -2.807    8.270  117.350
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 84.199255   0.560580 150.200   <2e-16 ***
## BMI          0.640825   0.010604  60.434   <2e-16 ***
## active       0.146562   0.154988   0.946    0.344
## AgeinYr      0.467887   0.009135  51.217   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.15 on 68714 degrees of freedom
## Multiple R-squared:  0.09107,    Adjusted R-squared:  0.09103
## F-statistic:  2295 on 3 and 68714 DF,  p-value: < 2.2e-16
```

Vidimo da je jedini značajan regresor mjera tjelesne aktivnosti, no s trenutnim odabirom regresora dobivamo najbolju R^2 vrijednost tako da smo ih odlučili zadržati.

Nastavimo s analizom reziduala. Prvo testiramo normalnost:

```
plot(fit.multi$fitted.values,fit.multi$residuals)
```

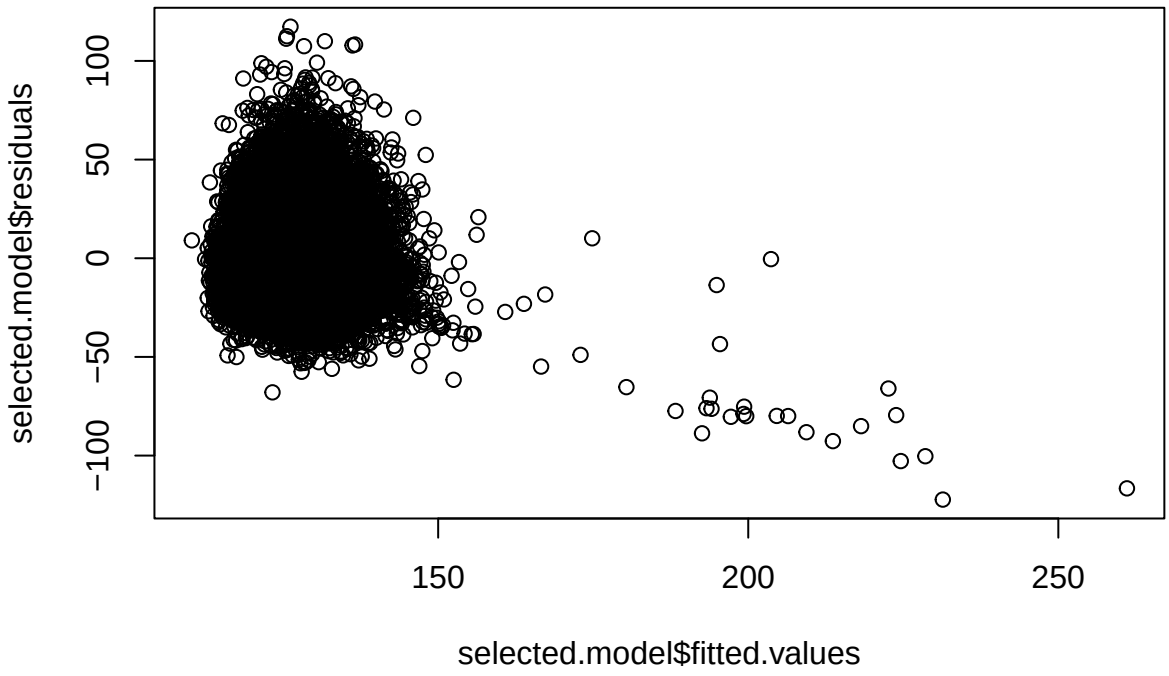



```
#KS test na normalnost
ks.test(rstandard(fit.ap_hi),'pnorm')
```

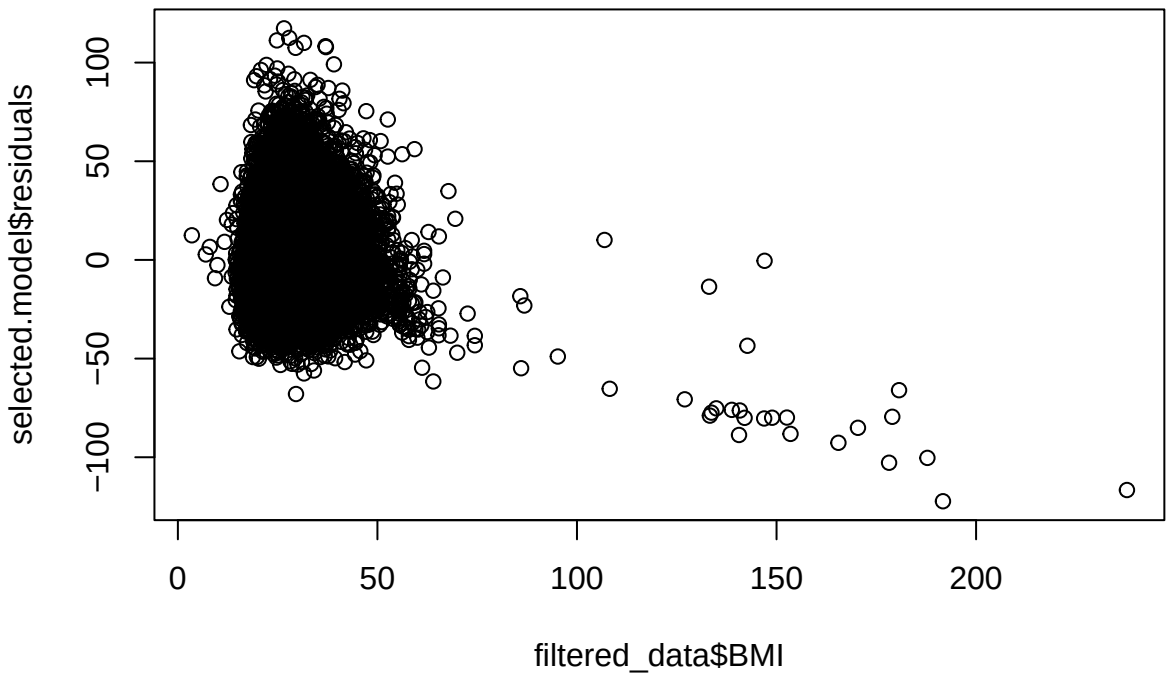
```
require(nortest)
lillie.test(rstandard(fit.ap_hi))
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.ap_hi)
## D = 0.10822, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.ap_hi)
## D = 0.10821, p-value < 2.2e-16
```

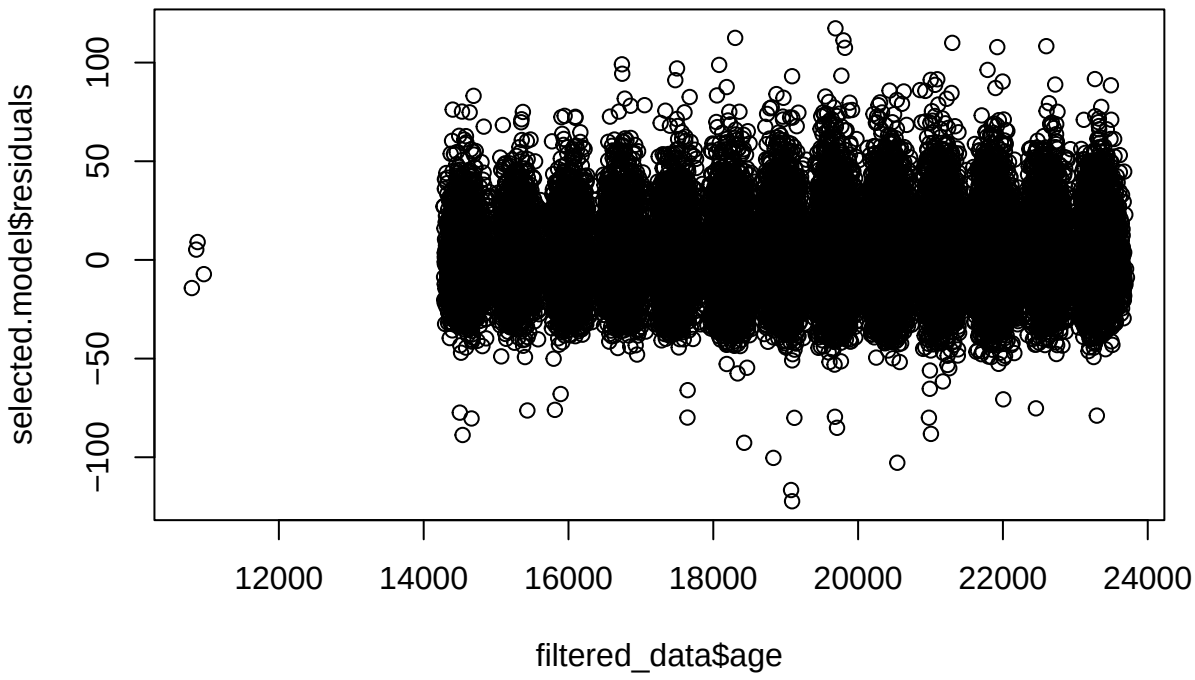
```
selected.model = fit.multi
plot(selected.model$fitted.values,selected.model$residuals)
```




```
plot(filtered_data$BMI, selected.model$residuals)
```



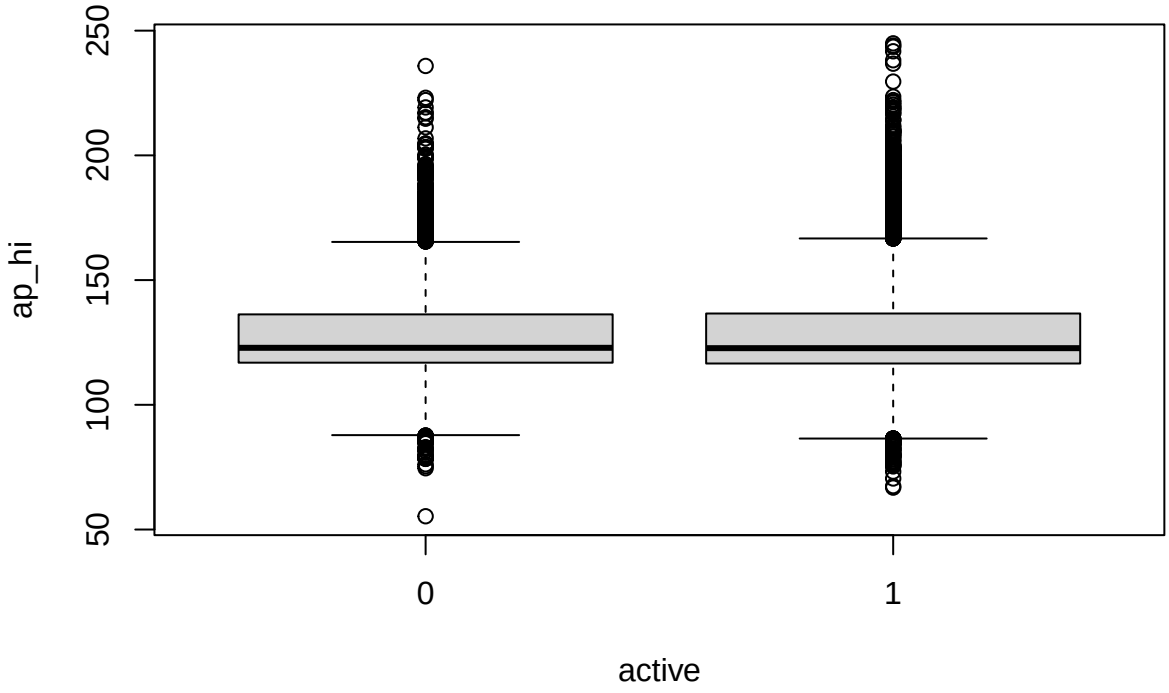

```
plot(filtered_data$age, selected.model$residuals)
```



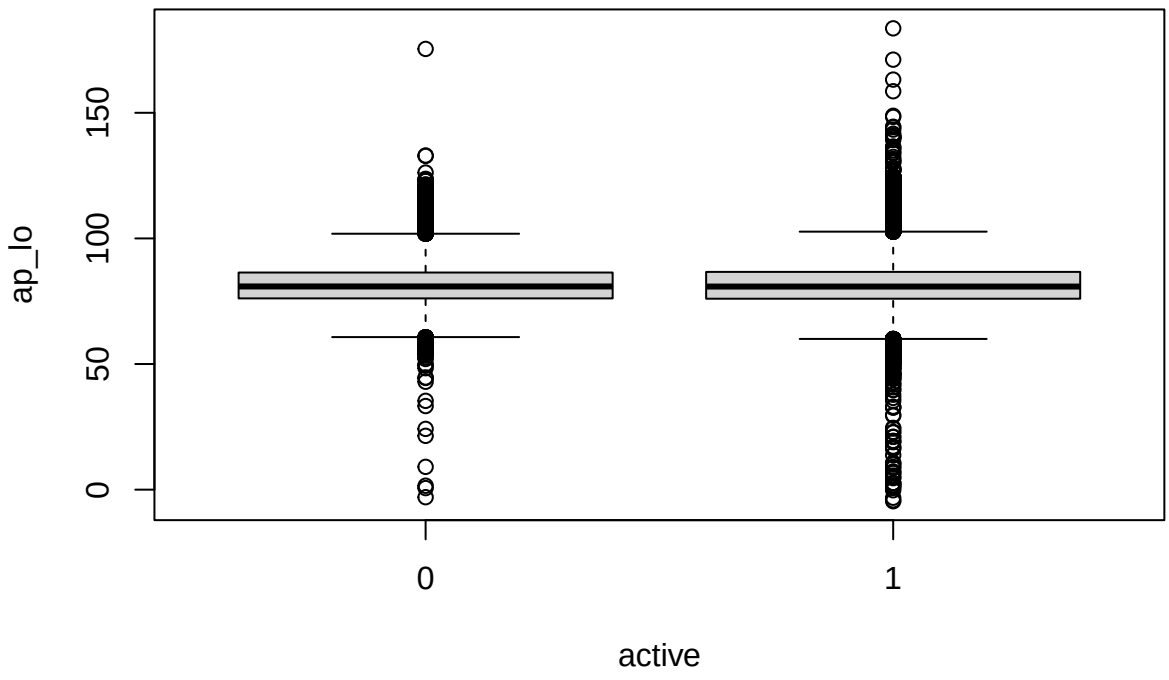

```
ks.test(rstandard(fit.multi),'pnorm')
require(nortest)
lillie.test(rstandard(fit.multi))
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.multi)
## D = 0.091074, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.multi)
## D = 0.091072, p-value < 2.2e-16
```

```
boxplot(ap_hi~active,data=filtered_data)
```




```
boxplot(ap_lo~active,data=filtered_data)
```




```
notA <- subset(filtered_data, active == 0)
A <- subset(filtered_data, active == 1)
mean(notA$ap_hi)
mean(A$ap_hi)
mean(notA$ap_lo)
mean(A$ap_lo)
```

```
## [1] 126.73
## [1] 126.6479
## [1] 81.30805
## [1] 81.24679
```

Iz grafova gore se ne čini kao da tjelesna aktivnost uopće utječe na krvni tlak.