

[SAP] Projektni zadatak - Analiza podataka zdravstvenog pregleda

Statistički najizglednije ime

2025-01-26

```
#Inicijalni koraci projekta ## Učitavanje podataka
```

```
###Filtriranje podataka
```

Početna faza analize uključila je čišćenje podataka korištenjem više kriterija. Eliminirane su sve observacije s neelogičnim vrijednostima tlaka (npr. negativne vrijednosti) te slučajevi gdje maksimalni tlak prelazi minimalni. Dodatno, izbačeni su ekstremni BMI-ovi te restrigirani na raspone unutar NHS inform podataka. (Izvor: <https://www.nhsinform.scot/healthy-living/food-and-nutrition/healthy-eating-and-weight-management/body-mass-index-bmi/>)

```
# Filtriranje besmislenih podataka iz tablice
filtered_data <- healthDATA.modif %>% filter(ap_hi <= 370) %>%
  filter(ap_lo <= 360) %>% filter(ap_hi >= 40) %>% filter(ap_lo >= 0) %>%
  filter(ap_hi >= ap_lo) %>% filter(BMI <= 52.3)

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)

#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      72.3
## 2 Male        77.1
## # A tibble: 2 x 2
##   gender avg_height
##   <fct>     <dbl>
## 1 Female      161.
## 2 Male        170.

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

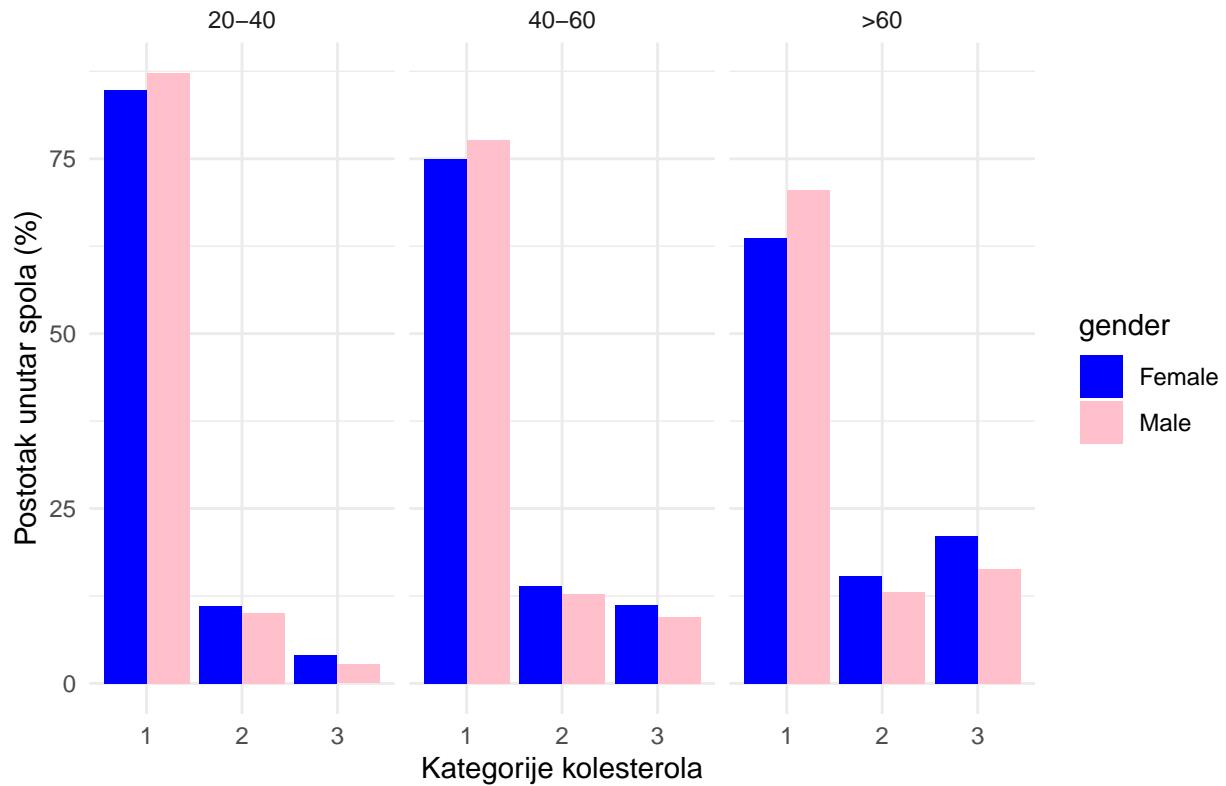
distribution <- filtered_data %>%
  group_by(AgeGroup, gender, cholesterol) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(AgeGroup, gender) %>%
  mutate(percentage = count / sum(count) * 100)

#distribution

ggplot(distribution, aes(x = cholesterol, y = percentage, fill = gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ AgeGroup) +
  labs(title = "Distribucija kolesterola prema spolu i dobnoj skupini",
       x = "Kategorije kolesterola",
       y = "Postotak unutar spola (%)") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "pink"))

```

Distribucija kolesterola prema spolu i dobnoj skupini



Grafički prikazi ilustriraju postotnu zastupljenost tri kategorije kolesterola (1-zdrav, 2-rizičan, 3-opasan) kroz dobne skupine i splove. Vizualno odvajanje kategorija ostvareno je paletom boja: zelenom za zdravu, žutom za rizičnu i ljubičastom za opasnu razinu.

Zdrav (1) - prikazan zelenkastom bojom,

Rizičan (2) - prikazan žutom bojom,

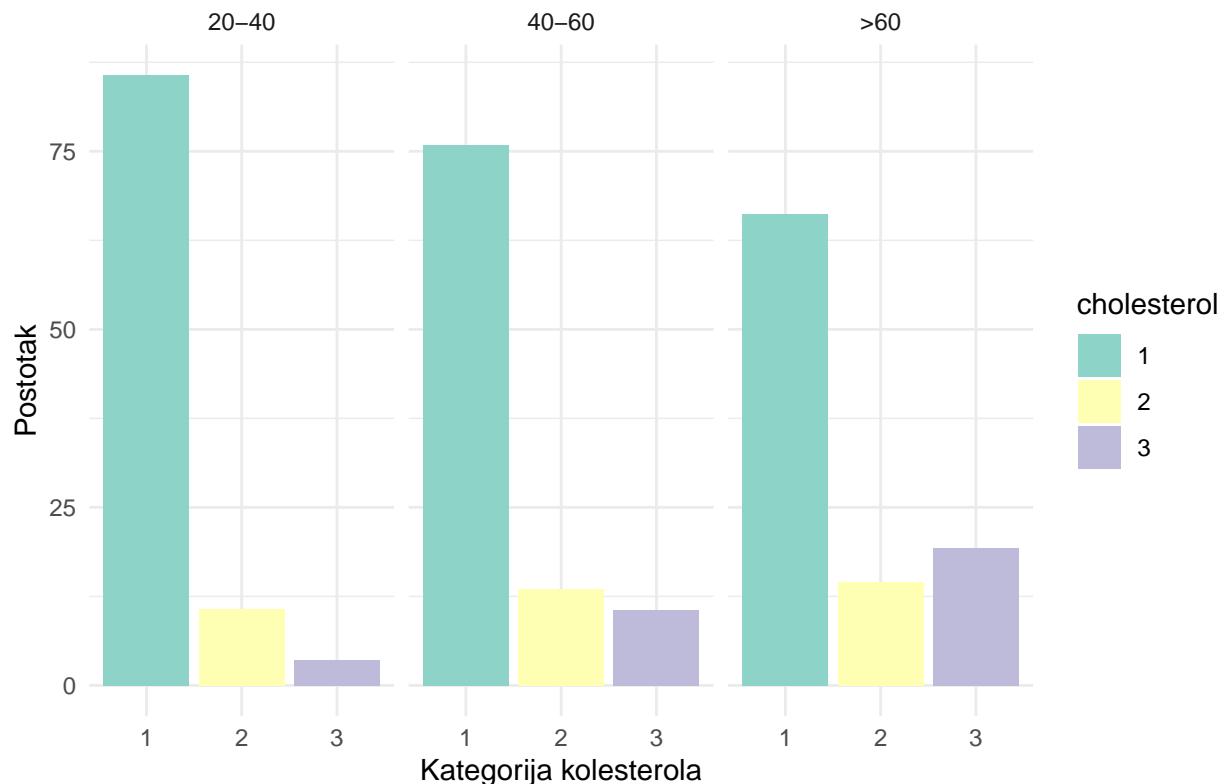
Opasan (3) - prikazan ljubičastom bojom.

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

```
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")
```

Postotak kolesterola unutar dobnih skupina



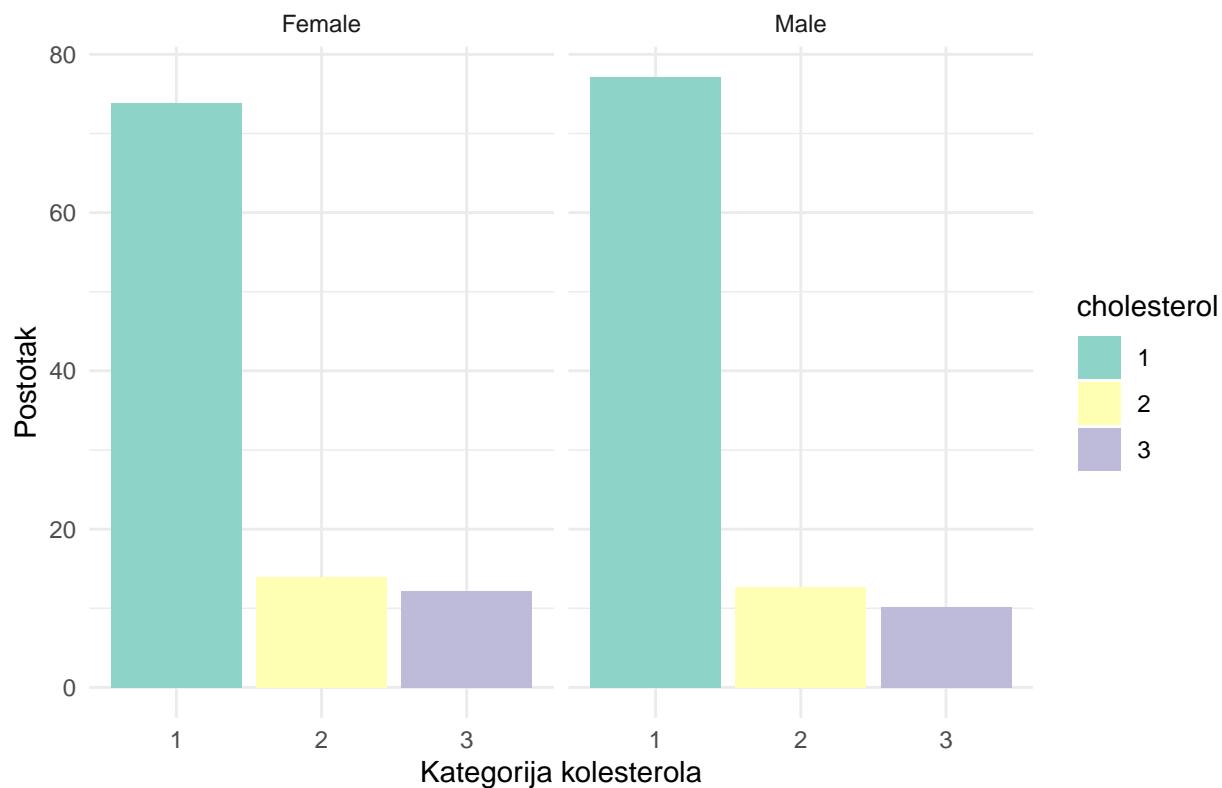
Zaključci sa grafa: - U svakoj dobroj skupini prevladava zdrava kategorija kolesterola. - Dobna skupina 20-40 godina ima najzdravije razine kolesterola, s najmanjom zastupljenosti rizične i opasne kategorije. - Skupina iznad 60 godina i dalje najčešće pripada zdravoj kategoriji, ali ima veću zastupljenost rizičnih i opasnih kategorija. Također, u ovoj skupini opasna kategorija nadmašuje rizičnu.

Sljedeći graf prikazuje distribuciju kolesterola prema spolovima.

```
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")
```

Postotak kolesterola unutar spolova



Zaključci sa grafa: - Zdrava kategorija dominira u svakom spolu. - Razlike između spolova u distribuciji kategorija kolesterola nisu značajne.

##Prikaz dodatnih distribucija

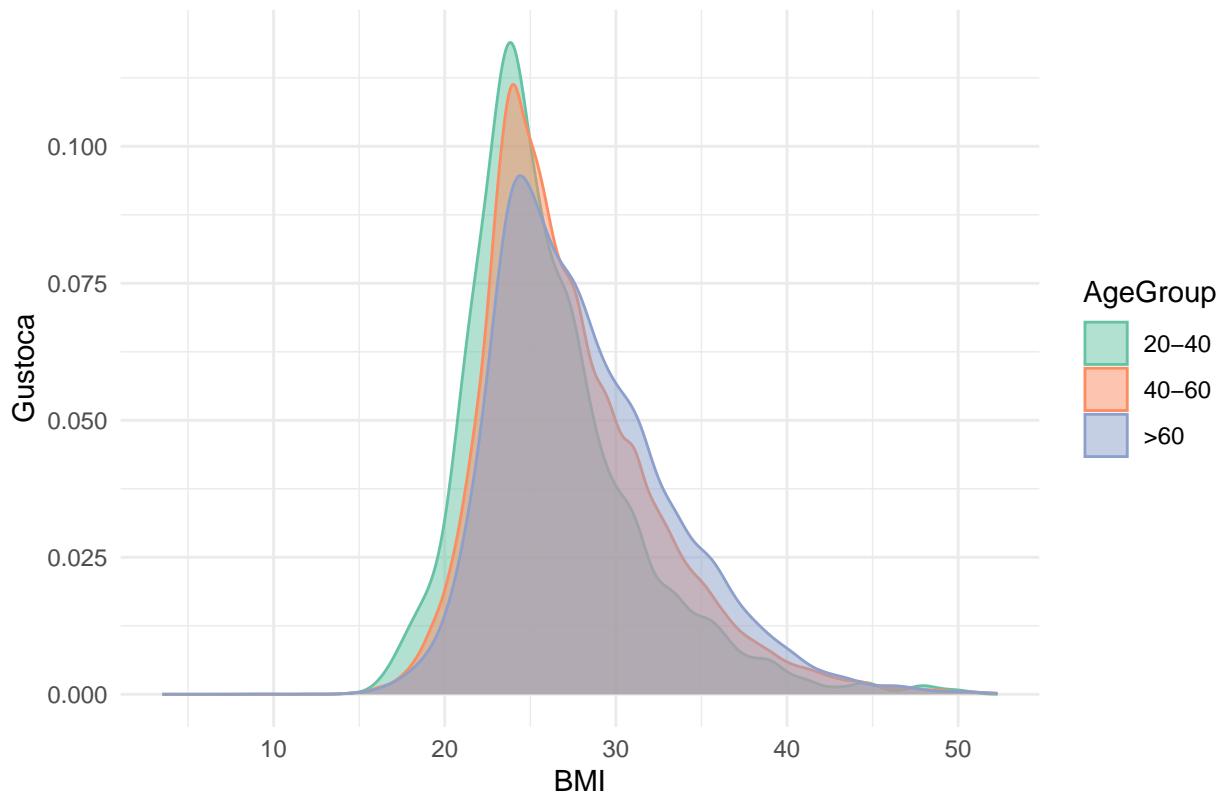
Također nas zanima kako su u uzorku distribuirani indeks tjelesne mase te krvni tlak. Fokusirani grafovi prikazuju češće vrijednosti kako bi se istaknula opća tendencija.

Distribucija indeksa tjelesne mase analizirana je prema dobnim skupinama i spolovima.

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

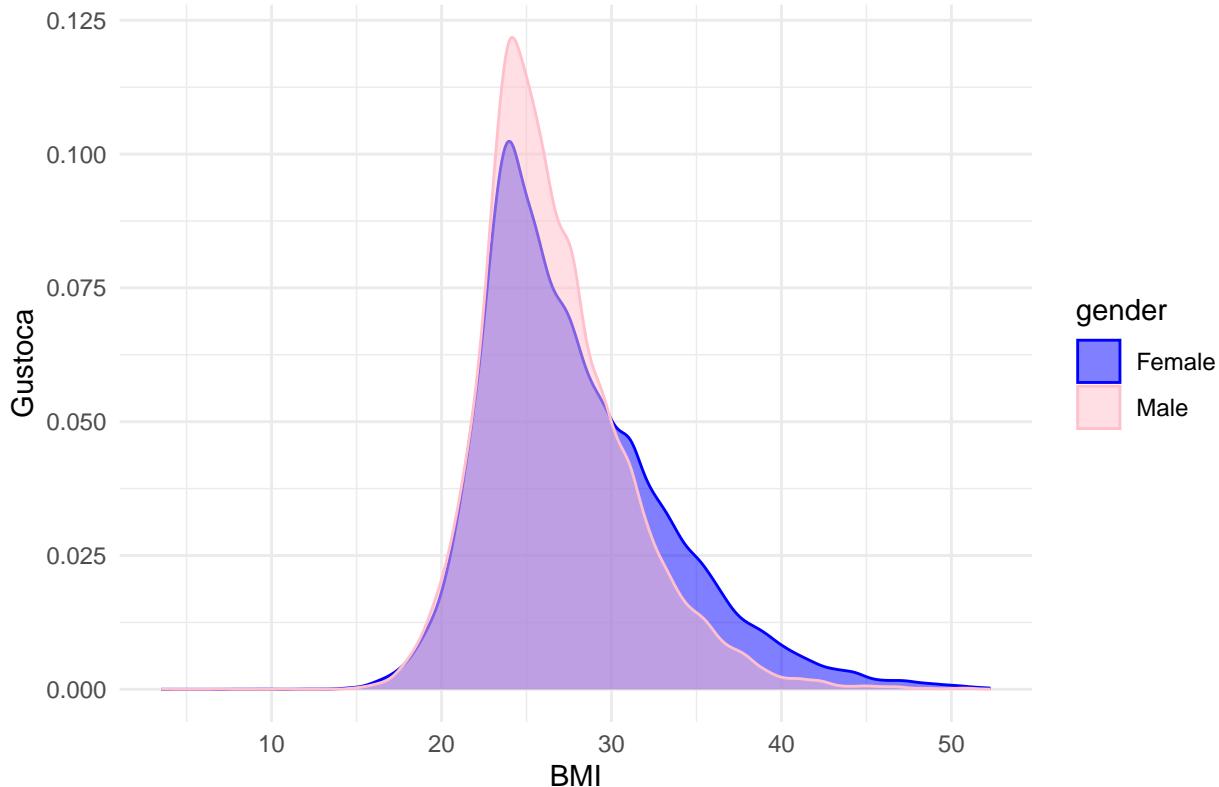
```
ggplot(BMI_filtered_data, aes(x = BMI, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Istaknuta Distribucija BMI-ja prema doboj skupini",
       x = "BMI",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette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

Istaknuta Distribucija BMI-ja prema dobnoj skupini



```
ggplot(BMI_filtered_data, aes(x = BMI, color = gender, fill = gender)) +  
  geom_density(alpha = 0.5) +  
  labs(title = "Istaknuta 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"))
```

Istaknuta distribucija BMI–ja prema spolu



Zaključak - BMI većinom pripada rasponu između 20 i 30, dok su vrijednosti iznad 30 rjeđe i opadaju prema 40.

Distribucija sistoličkog i dijastoličkog krvnog tlaka također je analizirana prema dobnim skupinama i spolovima.

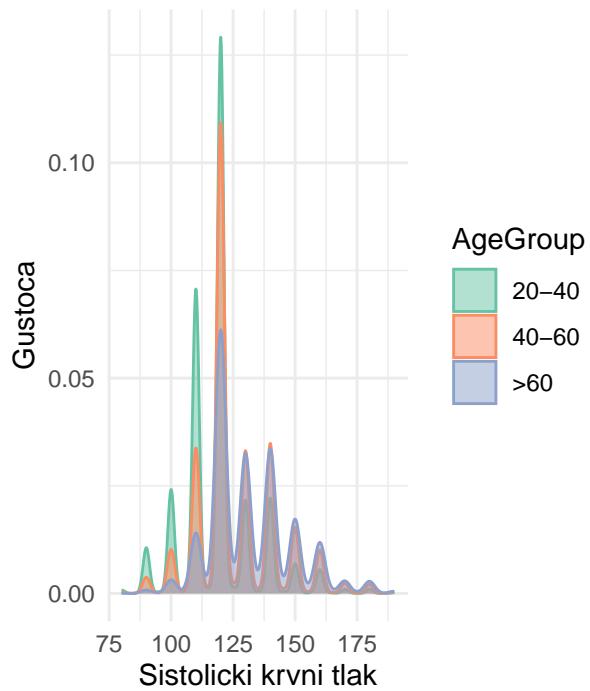
```
ap_filtered_data <- filtered_data %>% filter(ap_hi <= 190) %>% filter(ap_lo <= 130) %>% filter(ap_hi >= 100)

ap_hi_distribution_g1 <- ggplot(ap_filtered_data, aes(x = ap_hi, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Istaknuta distribucija\nsistoličkog krvnog\nntlaka prema dobnoj skupini",
       x = "Sistolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")

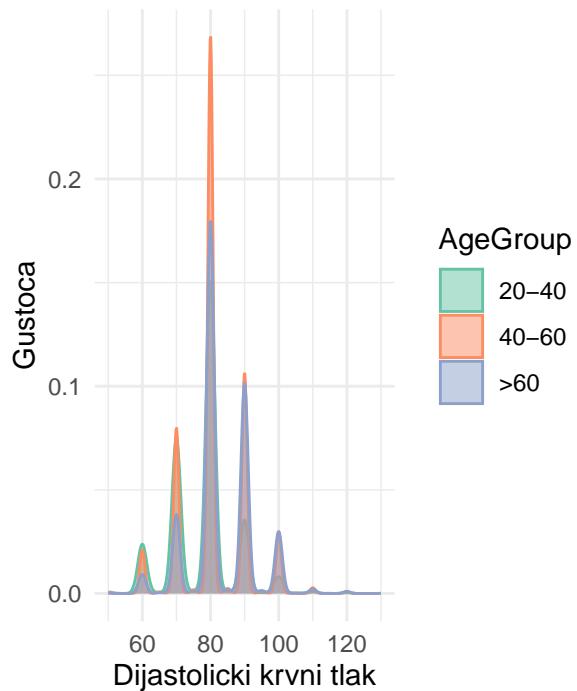
ap_lo_distribution_g1 <- ggplot(ap_filtered_data, aes(x = ap_lo, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Istaknuta distribucija\ndijastoličkog krvnog tlaka\nprema dobnoj skupini",
       x = "Dijastolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")

ap_hi_distribution_g1 + ap_lo_distribution_g1
```

Istaknuta distribucija sistolickog krvnog tlaka prema dobnoj skupini



Istaknuta distribucija dijastolickog krvnog tlaka prema dobnoj skupini

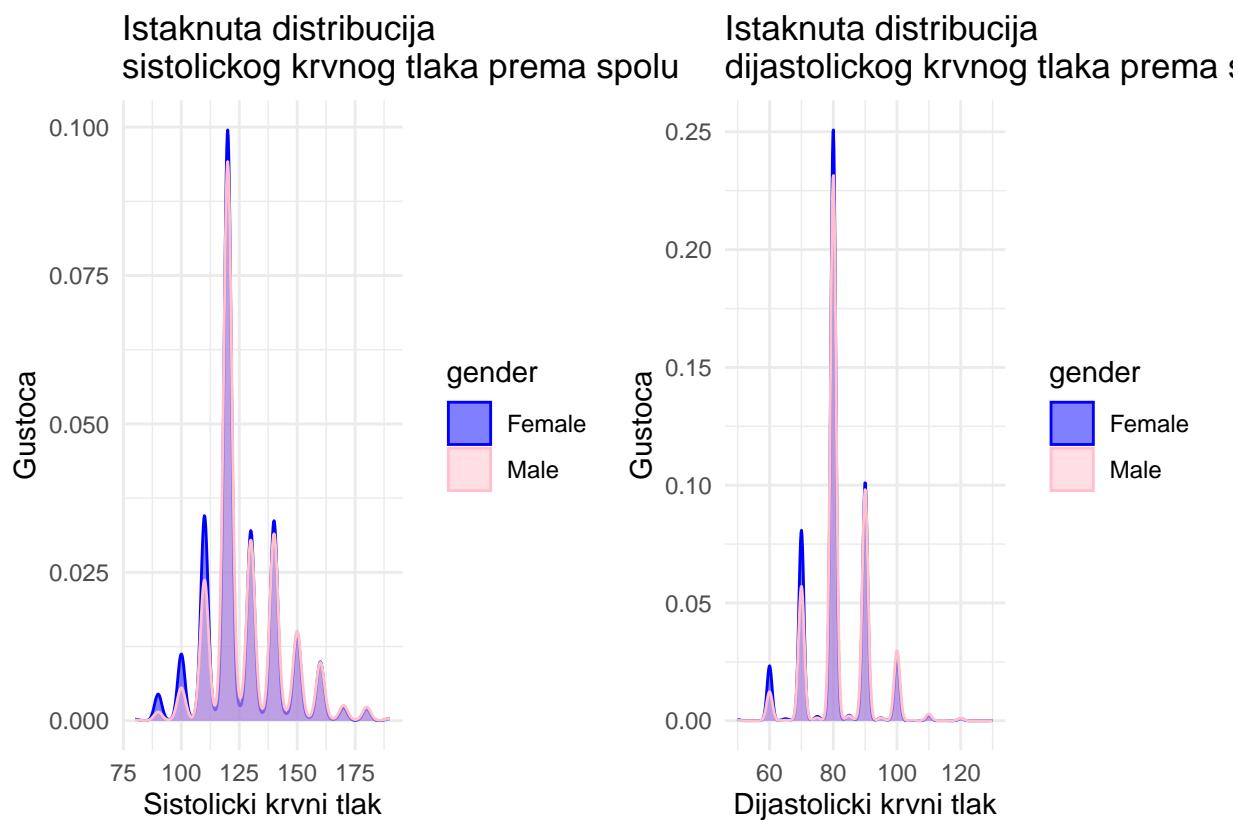


Prema spolu:

```
ap_hi_distribution_g2 <- ggplot(ap_filtered_data, aes(x = ap_hi, color = gender, fill = gender)) +
  geom_density(alpha = 0.5) +
  labs(title = "Istaknuta distribucija\nsistoličkog 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"))

ap_lo_distribution_g2 <- ggplot(ap_filtered_data, aes(x = ap_lo, color = gender, fill = gender)) +
  geom_density(alpha = 0.5) +
  labs(title = "Istaknuta distribucija\ndijastoličkog 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"))

ap_hi_distribution_g2 + ap_lo_distribution_g2
```



Zaključak - Krvni tlak pokazuje višemodalnu distribuciju, pri čemu je najčešća vrijednost 120/80 mmHg.

Analizom distribucije krvnog tlaka primjetili 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? Objasnjenje 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 izmjerениh 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 >= 60) %>% filter(ap_lo >= 40)
```

```

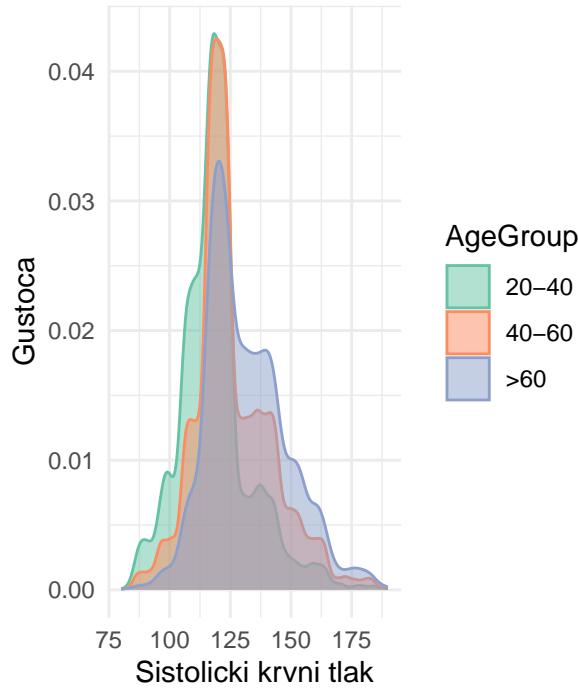
ap_hi_distribution_g3 <- ggplot(ap_filtered_data, aes(x = ap_hi, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribucija sistoličkog\nkrvnog tlaka sa dodanim šumom\nprema dobnoj skupini",
       x = "Sistolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")

ap_lo_distribution_g3 <- ggplot(ap_filtered_data, aes(x = ap_lo, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribucija dijastoličkog\nkrvnog tlaka sa dodanim šumom\nprema dobnoj skupini",
       x = "Dijastolički krvni tlak",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")

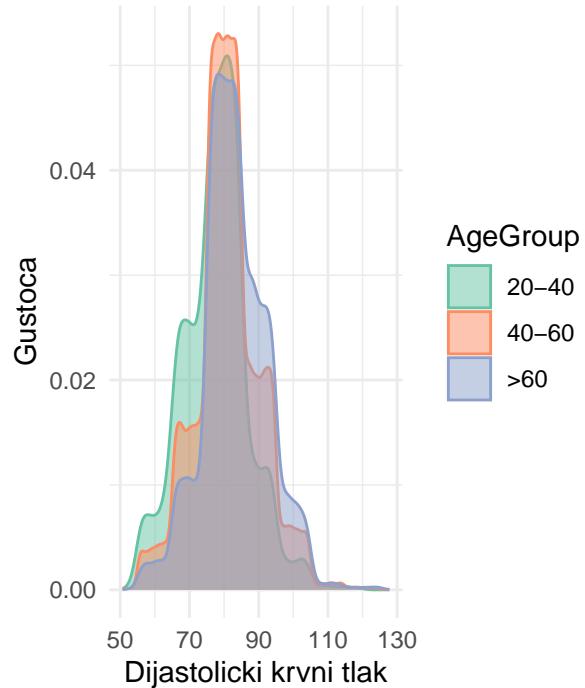
ap_hi_distribution_g3 + ap_lo_distribution_g3

```

Distribucija sistolickog
krvnog tlaka sa dodanim šumom
prema dobnoj skupini



Distribucija dijastoličkog
krvnog tlaka sa dodanim šumom
prema dobnoj skupini



#Zadatak 2 ##Postoji li značajna razlika u prosječnom krvnom tlaku između pušača i nepušača?

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

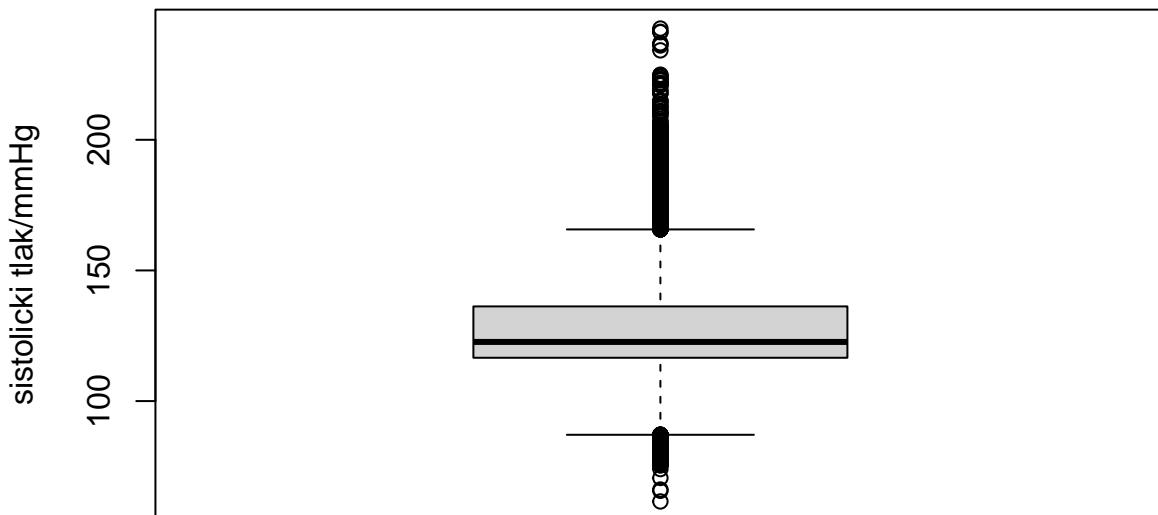
Tlakove pušača i nepušača možemo prvo usporediti grafički pomoću box plotova. U uzorku preostaje oko 6100

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 mjerena 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)

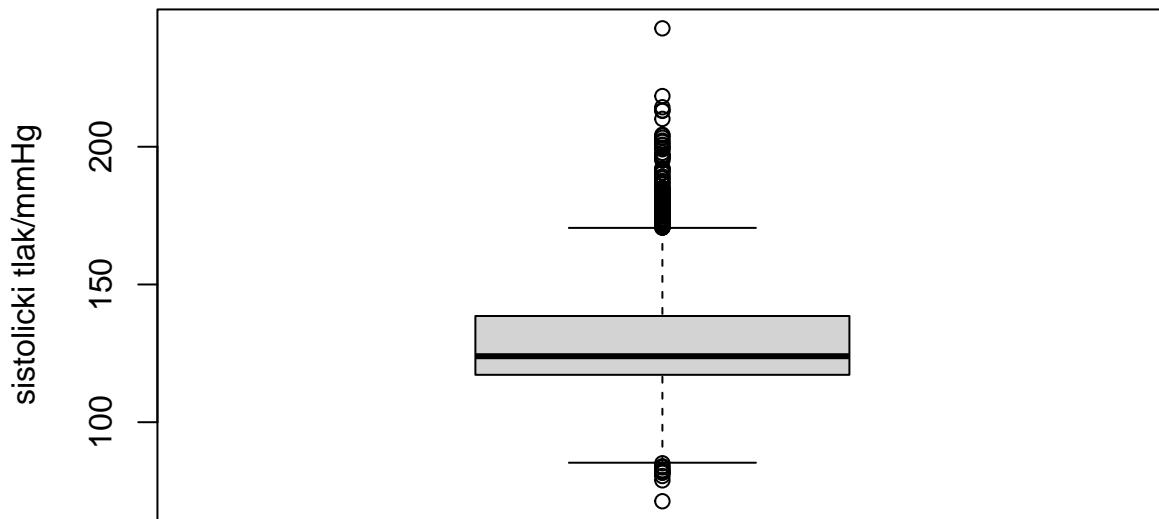
boxplot(nepusaci$ap_hi,
        main='box-plot tlaka nepušača',
        ylab='sistolički tlak/mmHg')
```

box-plot tlaka nepušaca



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

box-plot tlaka pušaca



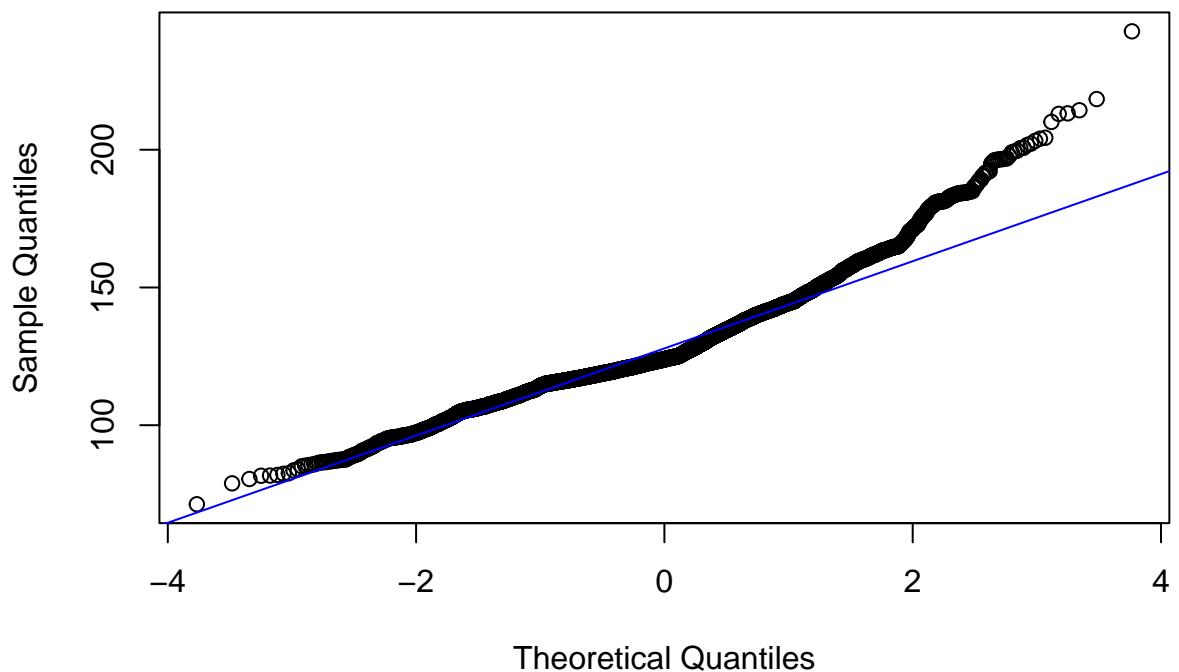
```
## [1] 6034
## [1] 62511
```

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,]

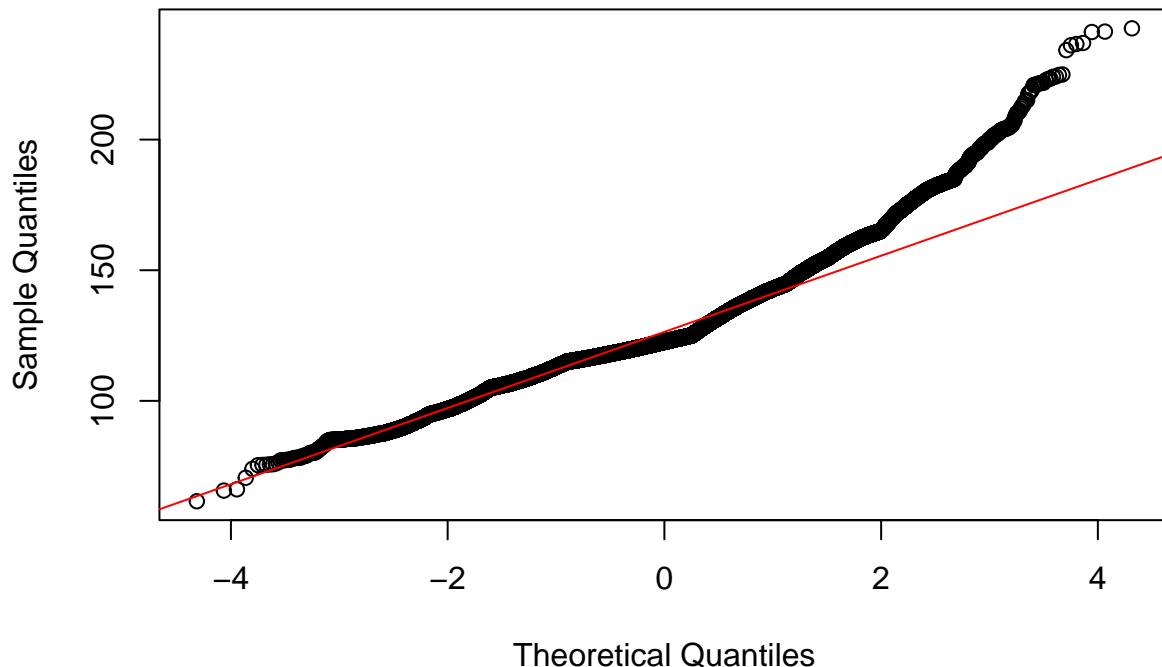
lillie.test(pusaci$ap_hi)
lillie.test(nepusaci$ap_hi)
qqnorm(pusaci$ap_hi, main = "Q-Q plot za tlakove pušača")
qqline(pusaci$ap_hi, col="blue")
```

Q-Q plot za tlakove pušaca



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

Q-Q plot za tlakove nepušaca



```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: pusaci$ap_hi  
## D = 0.11286, p-value < 2.2e-16  
##  
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: nepusaci$ap_hi  
## D = 0.13322, 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 nekim nakon nekoliko sati bez cigareta. U svakom slučaju, ne možemo prepostaviti jednakost varijanci u dalnjim testovima.

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

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

```

## 
## data: nepusaci$ap_hi and pusaci$ap_hi
## F = 0.9053, num df = 62510, denom df = 6033, p-value = 1.19e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.8718424 0.9395030
## sample estimates:
## ratio of variances
## 0.9053046

```

Testiramo hipotezu o jednakosti prosječnih krvnih tlakova u obje grupe koristeći T test za neuparene podatke uz nepoznate i nejednakе varijance na razini značajnosti od 5%. Ne uzimamo pretpostavku da su varijance jednakе 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 = -6.9415, df = 7128.2, p-value = 4.222e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -2.117373 -1.184821
## sample estimates:
## mean of x mean of y
## 126.5013 128.1524

```

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 ##Razlikuje li se prosječni krvni tlak značajno medu skupinama s različitom učestalošću tjelesne aktivnosti?

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

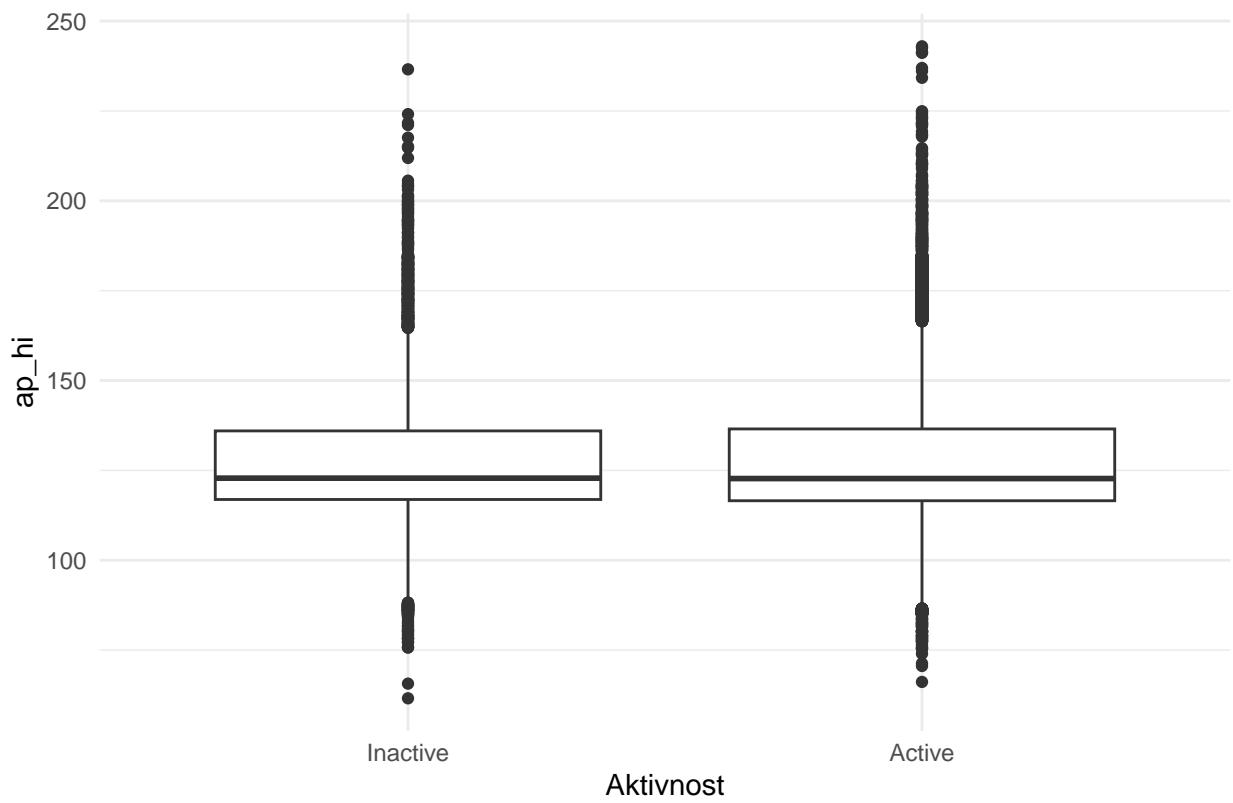
```

# 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")

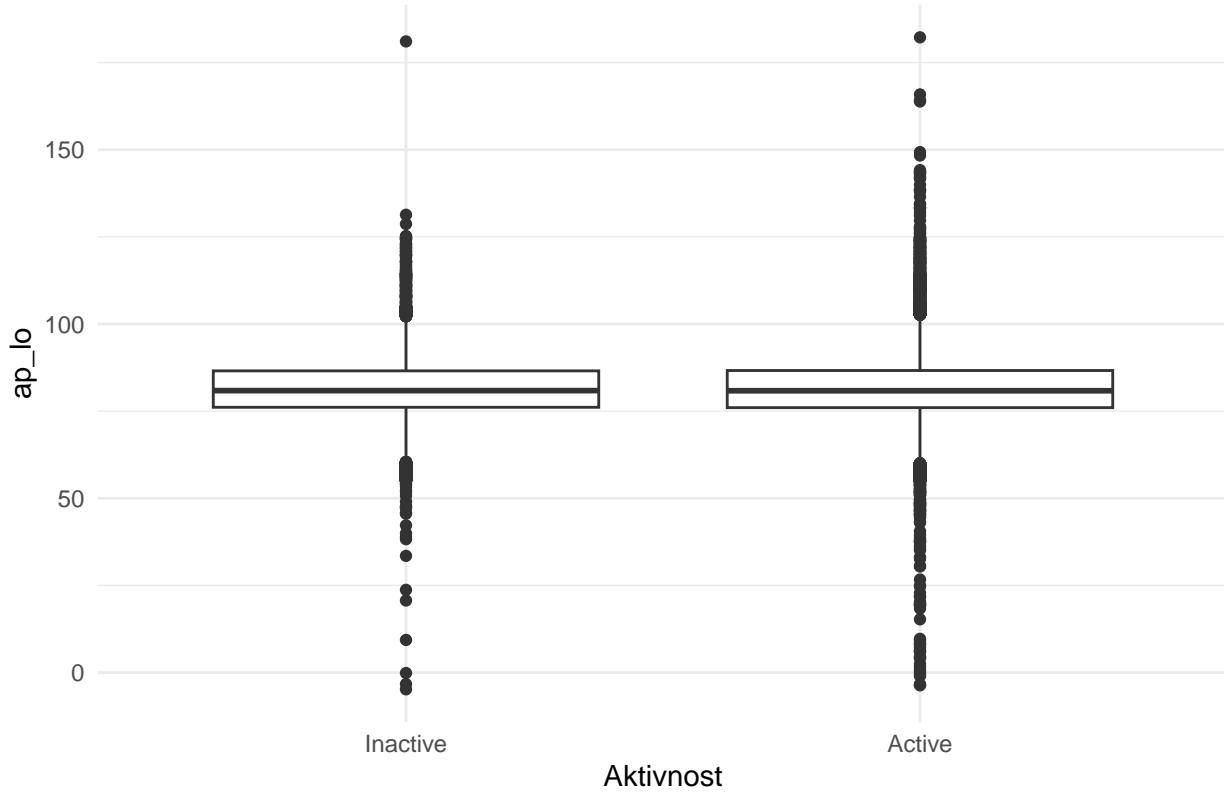
```

Distribucija ap_hi prema aktivnosti



```
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")
```

Distribucija ap_lo prema aktivnosti

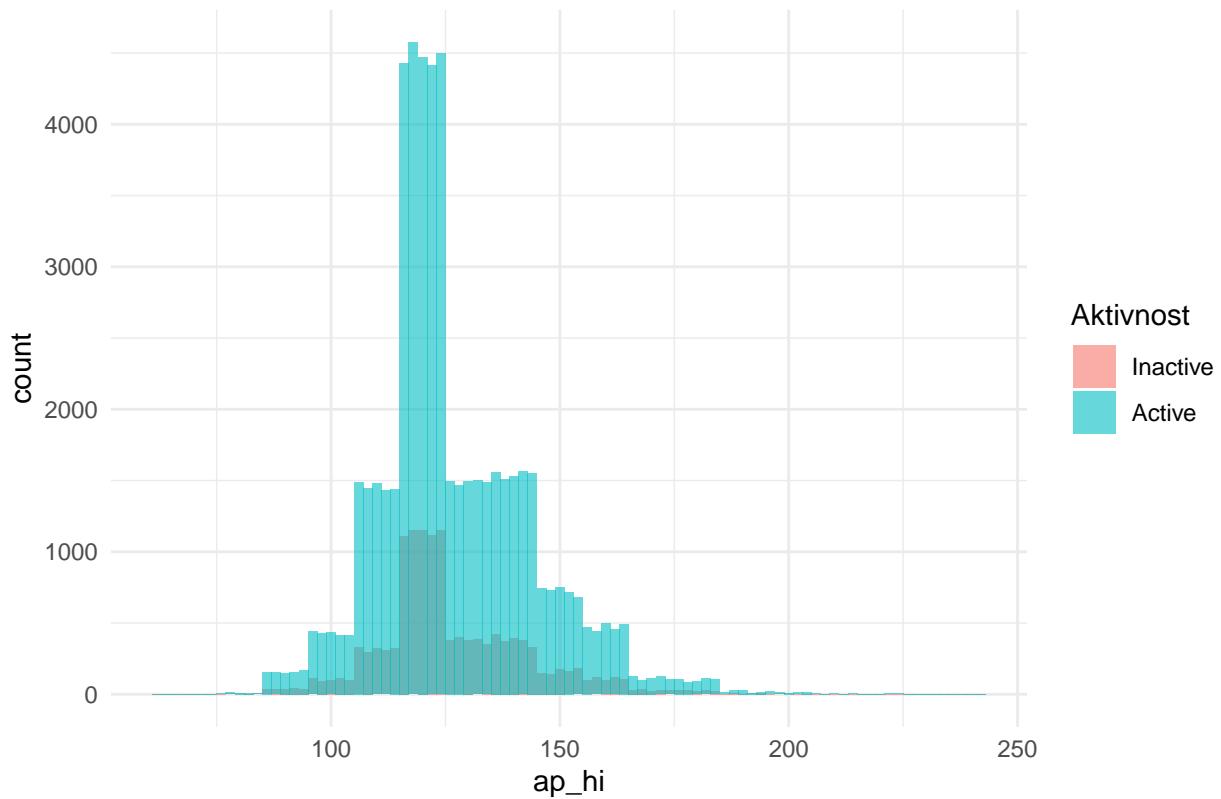


Box plotovi pokazuju slične vrijednosti krvnog tlaka kod aktivnih i neaktivnih osoba. Ipak, daljnje testiranje provedeno je za dublju analizu.

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

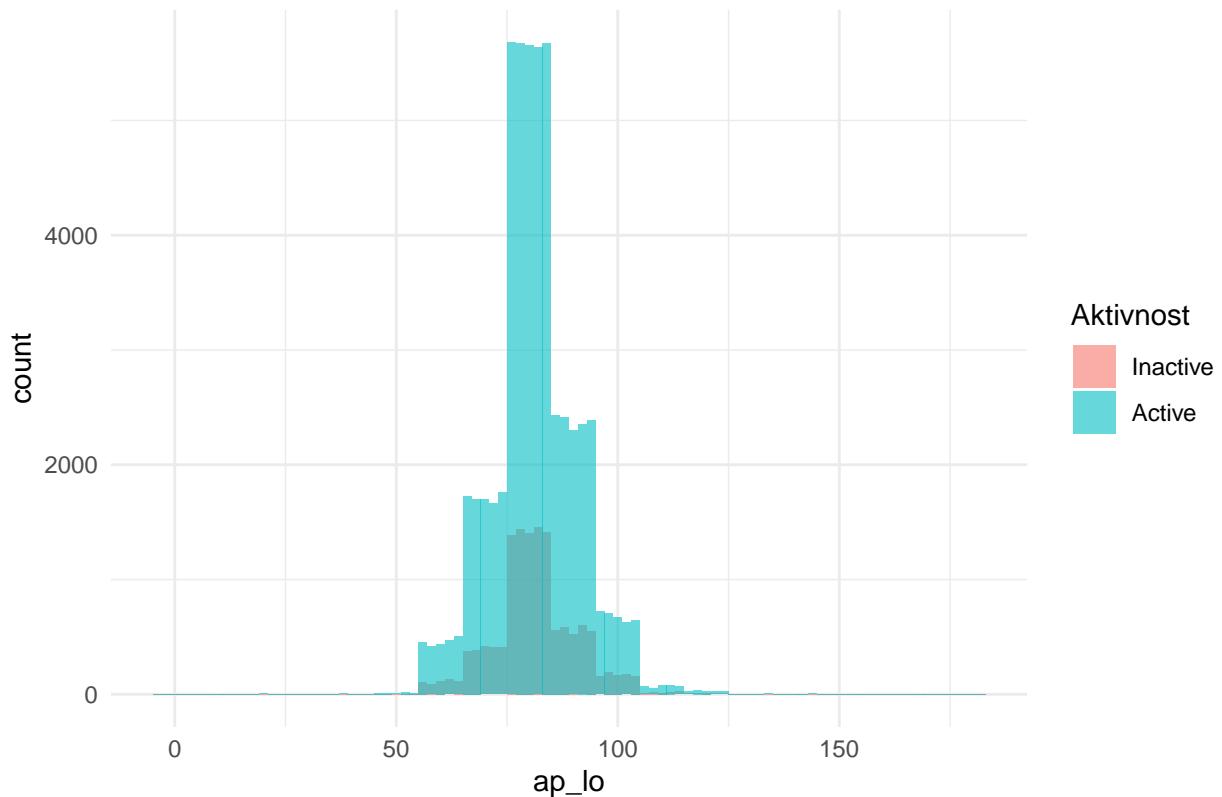
```
# 2) Histogram za ap_hi i ap_lo prema aktivnosti
ggplot(filtered_data, aes(x = ap_hi, fill = factor(active, labels = c("Inactive", "Active")))) +
  geom_histogram(binwidth = 2, alpha = 0.6, position = "identity") +
  theme_minimal() +
  labs(title = "Histogram ap_hi prema aktivnosti", x = "ap_hi",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

Histogram ap_hi prema aktivnosti



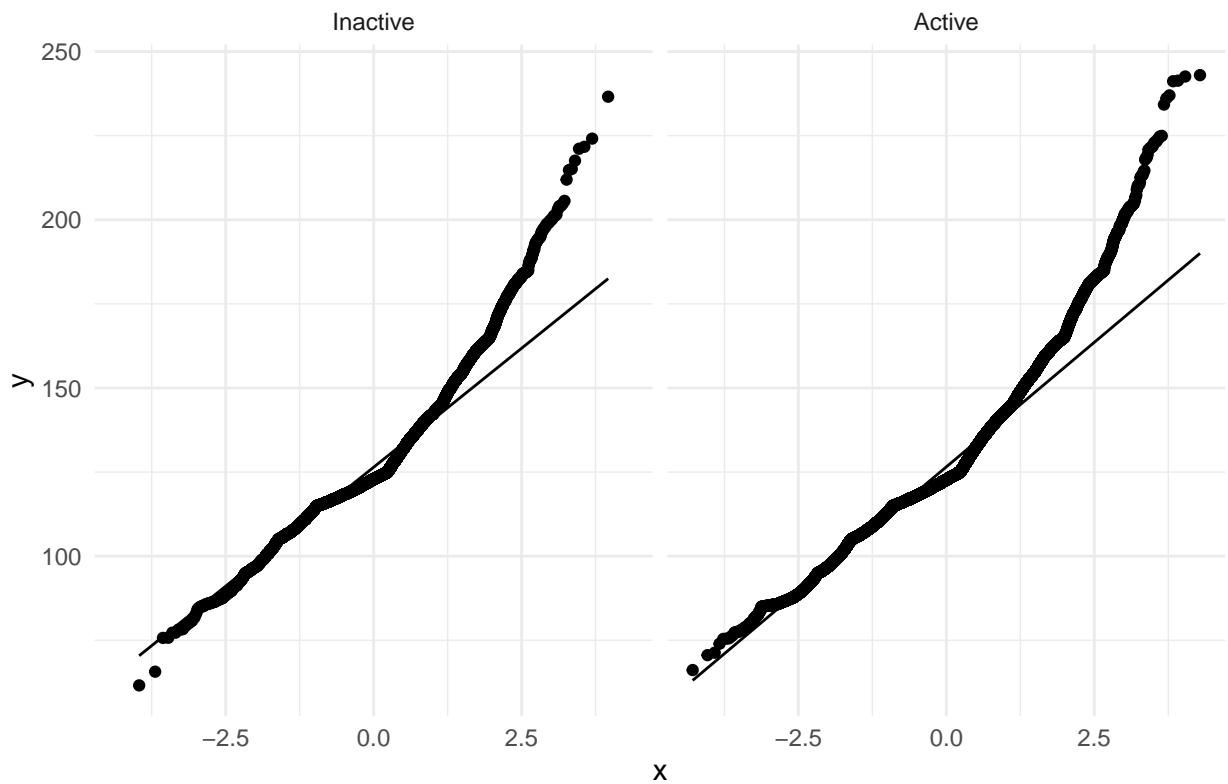
```
ggplot(filtered_data, aes(x = ap_lo, fill = factor(active, labels = c("Inactive", "Active")))) +  
  geom_histogram(binwidth = 2, alpha = 0.6, position = "identity") +  
  theme_minimal() +  
  labs(title = "Histogram ap_lo prema aktivnosti", x = "ap_lo", fill = "Aktivnost")
```

Histogram ap_lo prema aktivnosti



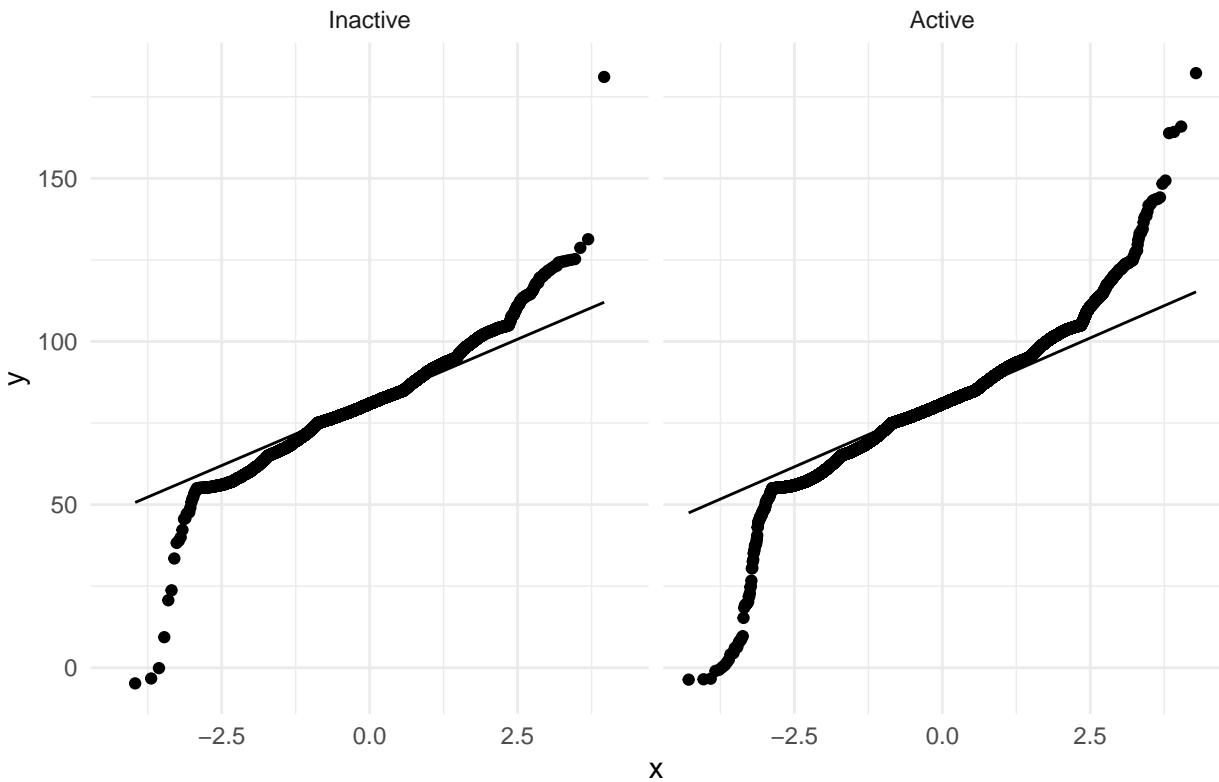
```
# 5) Q-Q Plot za proujedu normalnosti
ggplot(filtered_data, aes(sample = ap_hi)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~ active, labeller = as_labeller(c("0" = "Inactive", "1" = "Active"))) +
  theme_minimal() +
  labs(title = "Q-Q plot za ap_hi prema aktivnosti")
```

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")
```

Q-Q plot za ap_lo prema aktivnosti



Iz plotova vidimo kako distribucija najvjeroatnije 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$ap_hi))

## Warning in ks.test.default(active_data$ap_hi, "pnorm", mean =
## mean(active_data$ap_hi), : ties should not be present for the
## Kolmogorov-Smirnov test

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

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

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$p.value, "\n")
cat("F-test za ap_lo: p-value:", f_test_lo$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 = 1.0124, num df = 55071, denom df = 13472, p-value = 0.3654
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.9857068 1.0396555
## sample estimates:
## ratio of variances
## 1.012432
##
##
## F test to compare two variances
##
## data: active_data$ap_lo and inactive_data$ap_lo
## F = 1.0243, num df = 55071, denom df = 13472, p-value = 0.07829
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.9972808 1.0518631
## sample estimates:
## ratio of variances
## 1.02432
##
##
## F-test za ap_hi: p-value: 0.3654264
## F-test za ap_lo: p-value: 0.07829213

```

Iz rezultata testova vidimo kao i iz plotova kako

```

# 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-test
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)

# 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 = -0.27676, df = 20662, p-value = 0.782
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.3628953 0.2730948
## sample estimates:
## mean of x mean of y
## 126.6379 126.6828
##
##
## ===== t-test za ap_lo =====
##
## Welch Two Sample t-test
##
## data: active_data$ap_lo and inactive_data$ap_lo
## t = -0.71687, df = 20751, p-value = 0.4735
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.2563264 0.1190413
## sample estimates:
## mean of x mean of y
## 81.23678 81.30542
##
##
## ===== 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 = 3.69e+08, p-value = 0.332
## 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 = 369917672, p-value = 0.6016
## alternative hypothesis: true location shift is not equal to 0

# 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()

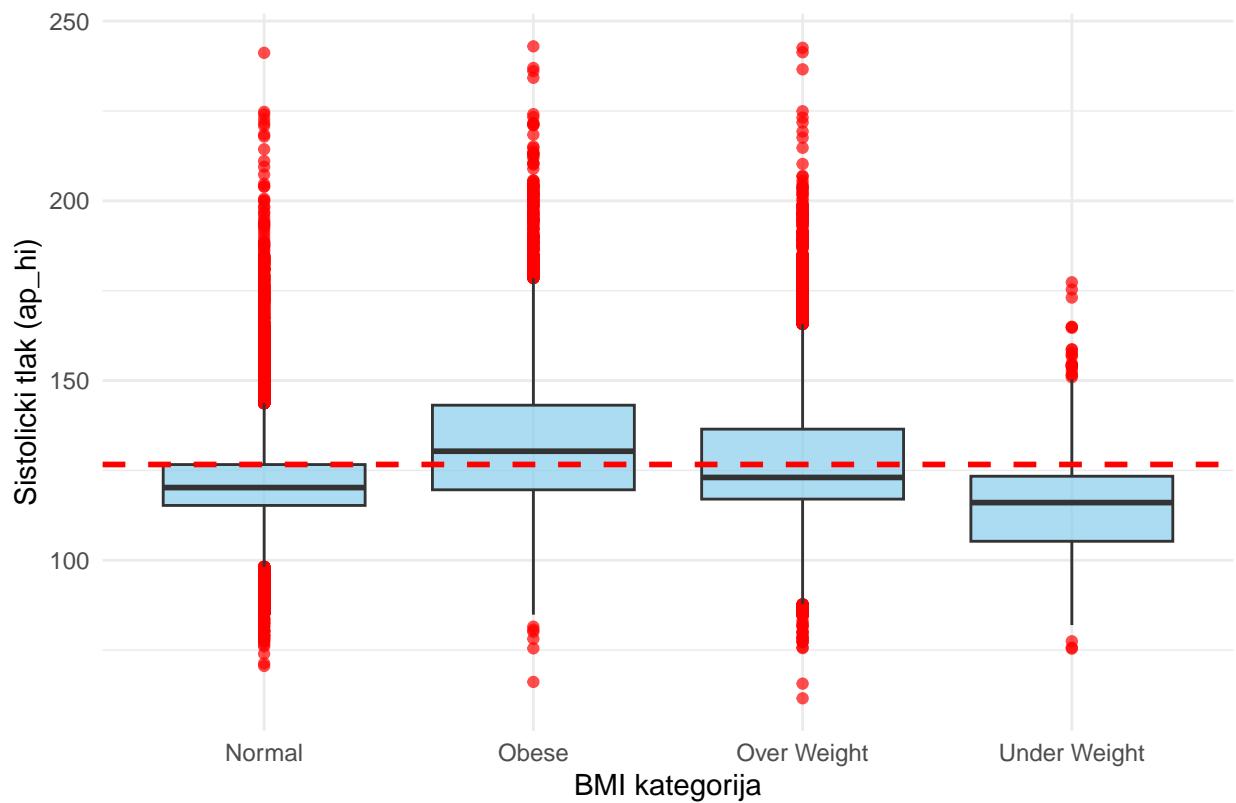
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

# 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)

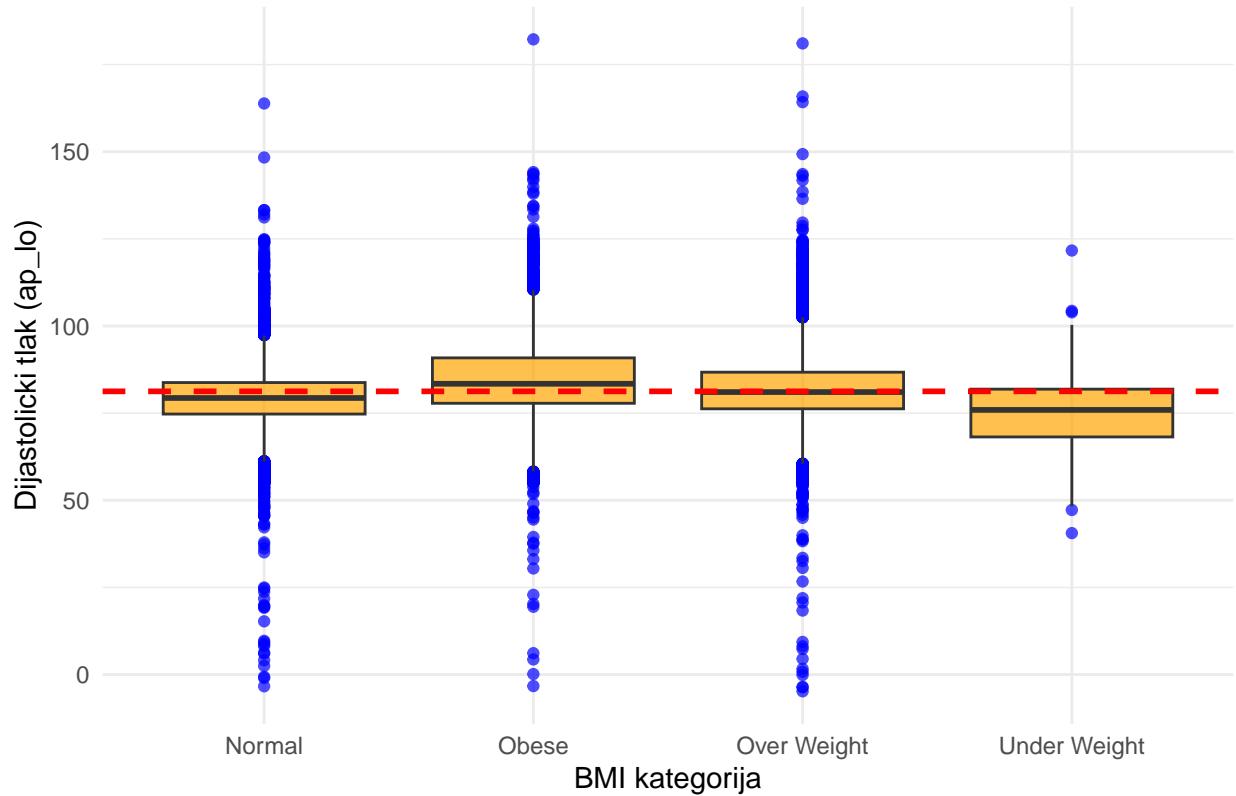
```

Box plot sistolickog tlaka (ap_hi) po BMI kategorijama



```
print(p_lo)
```

Box plot dijastolickog tlaka (ap_lo) po BMI kategorijama



```
# 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") +
    theme_minimal()
}
```

```

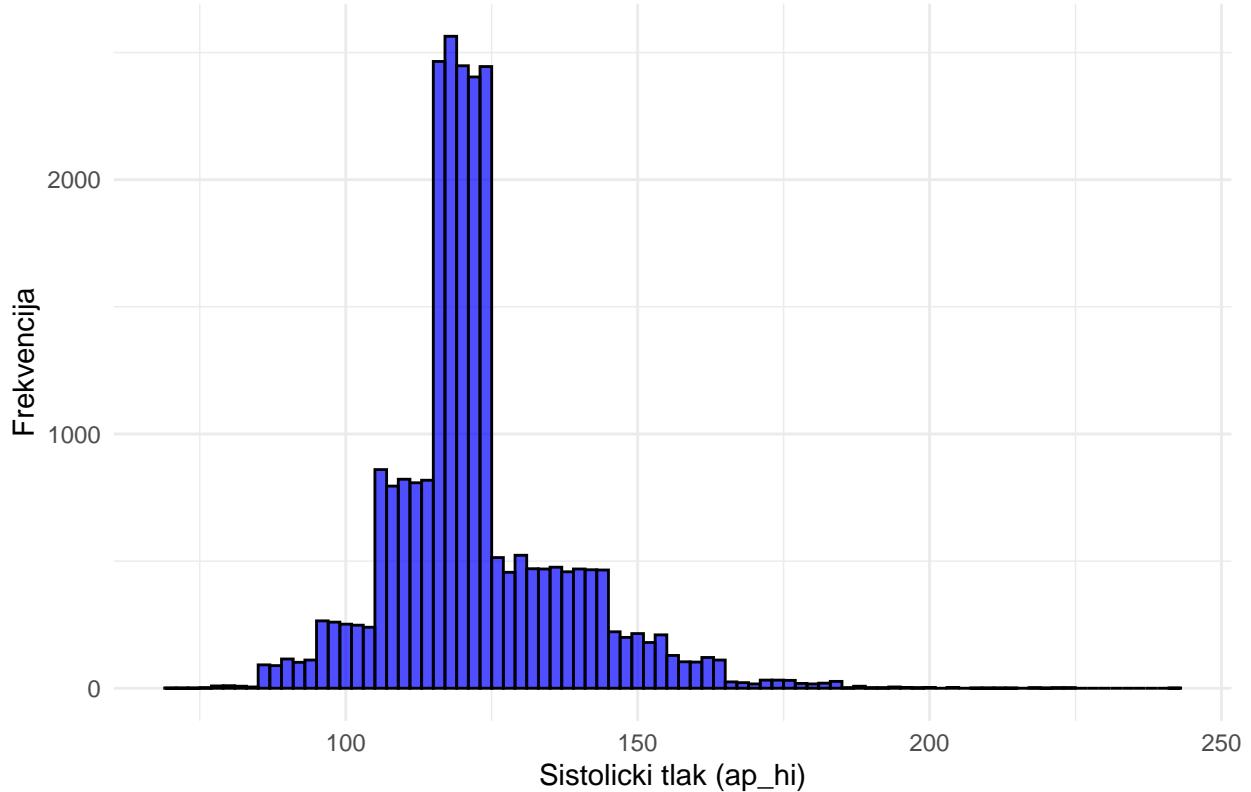
stat_qq_line(color = "red") +
labs(
  title = paste("Q-Q plot - Sistolicki tlak (ap_hi) -", bmi_cat),
  x = "Teorijske kvantile",
  y = "Empirijske kvantile"
) +
theme_minimal()

# 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()

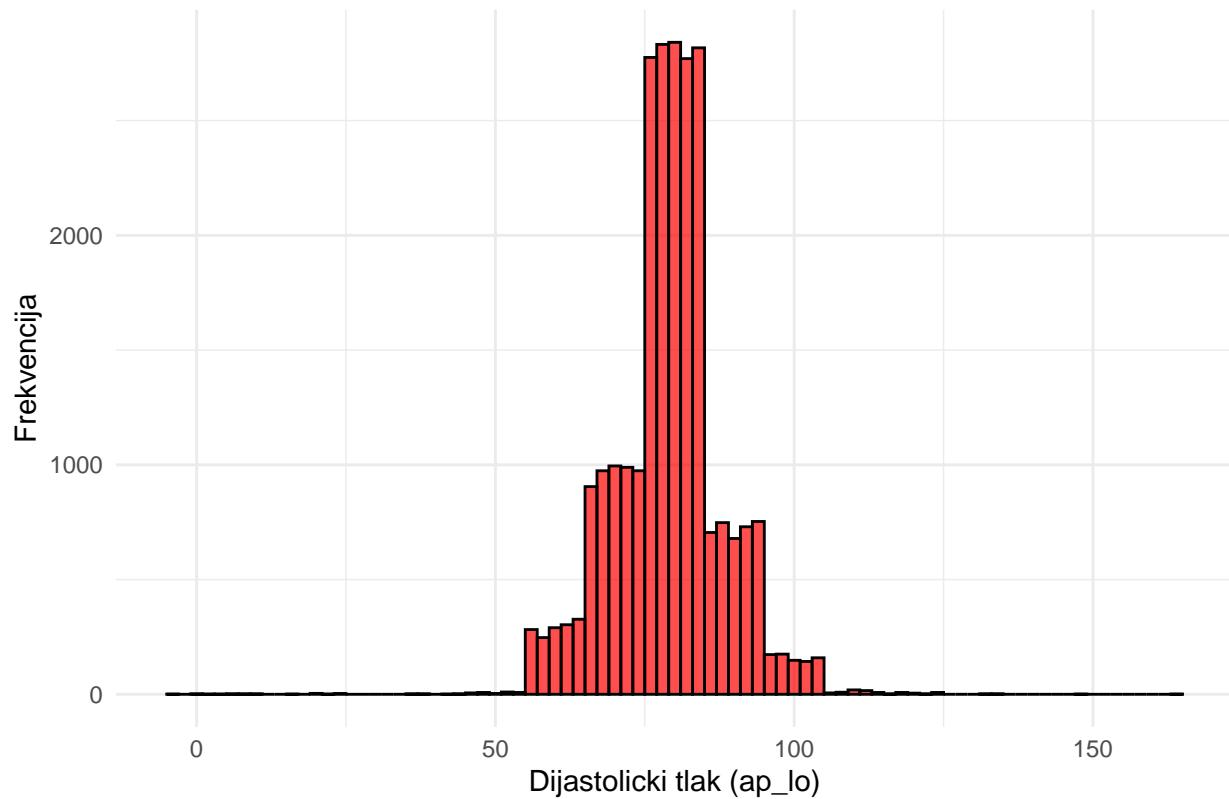
# Prikaz grafova
print(p1)
print(p2)
print(p3)
print(p4)
}

```

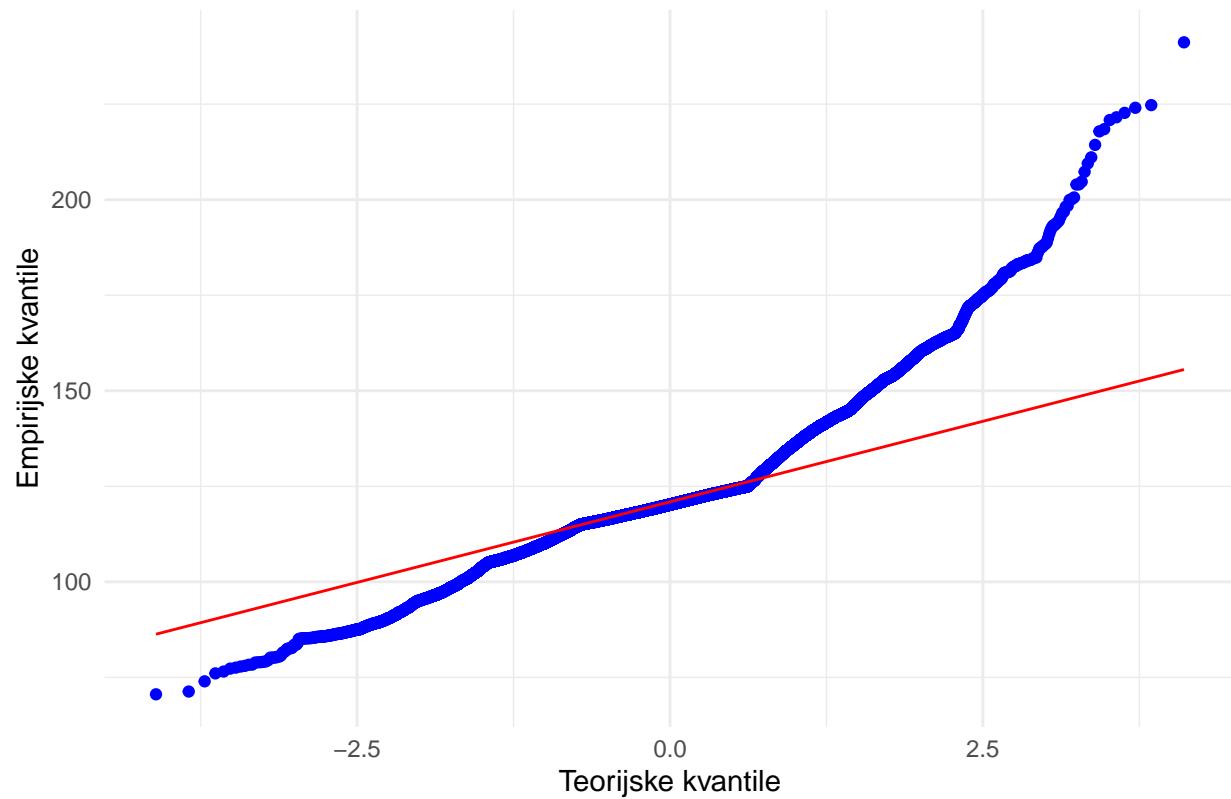
Distribucija sistolickog tlaka – Normal



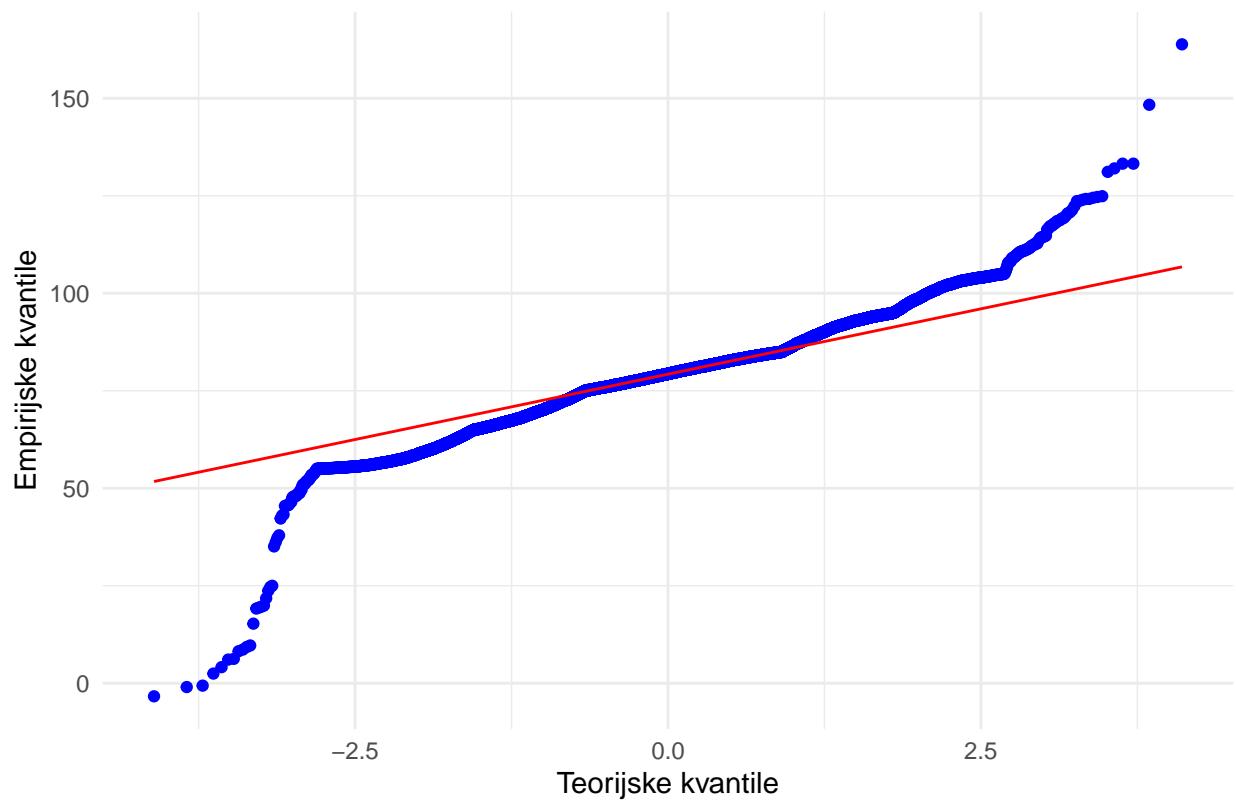
Distribucija dijastolickog tlaka – Normal



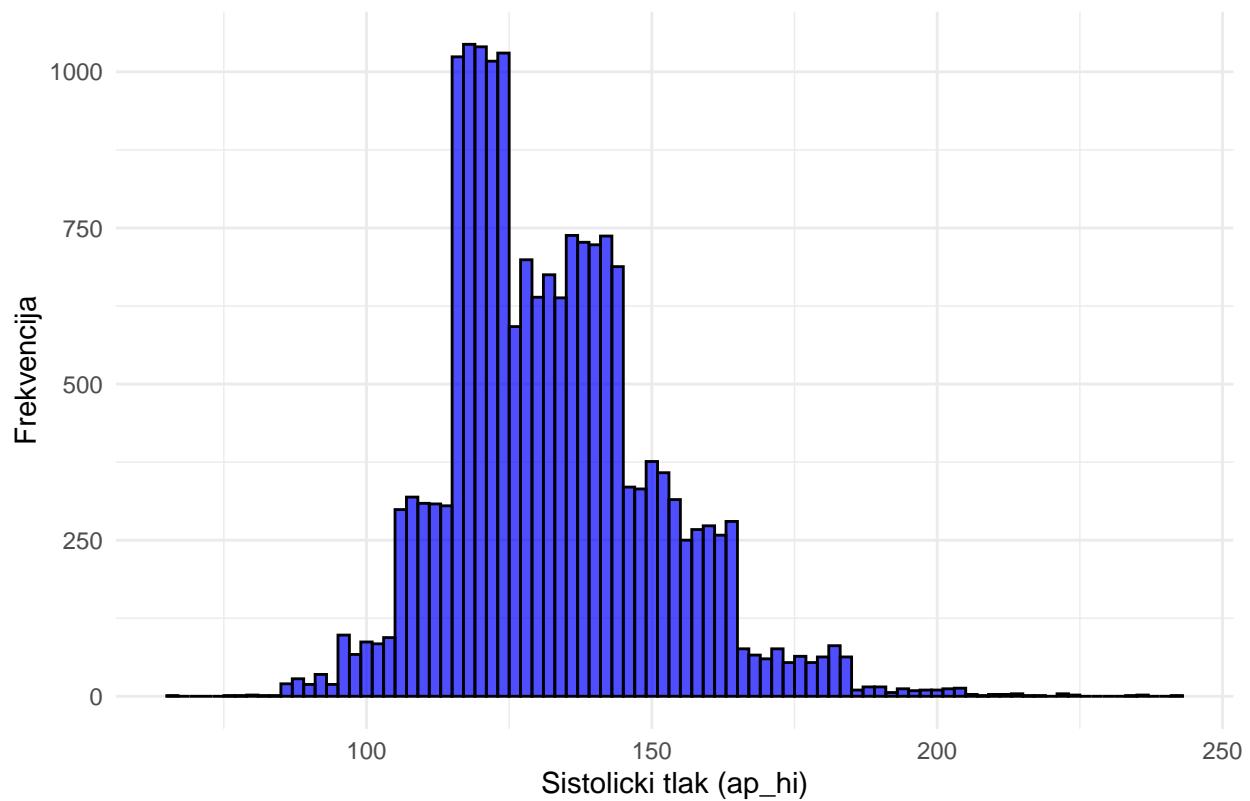
Q–Q plot – Sistolicki tlak (ap_hi) – Normal



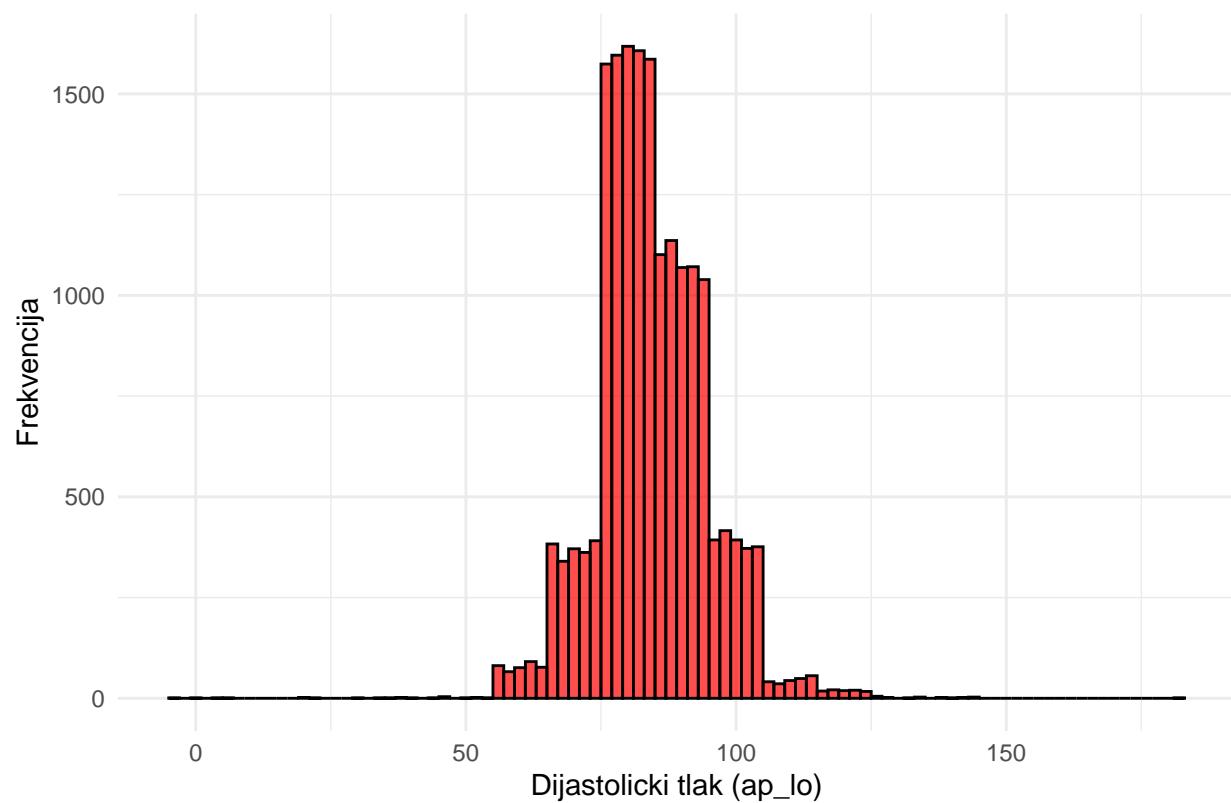
Q–Q plot – Dijastolicki tlak (ap_lo) – Normal



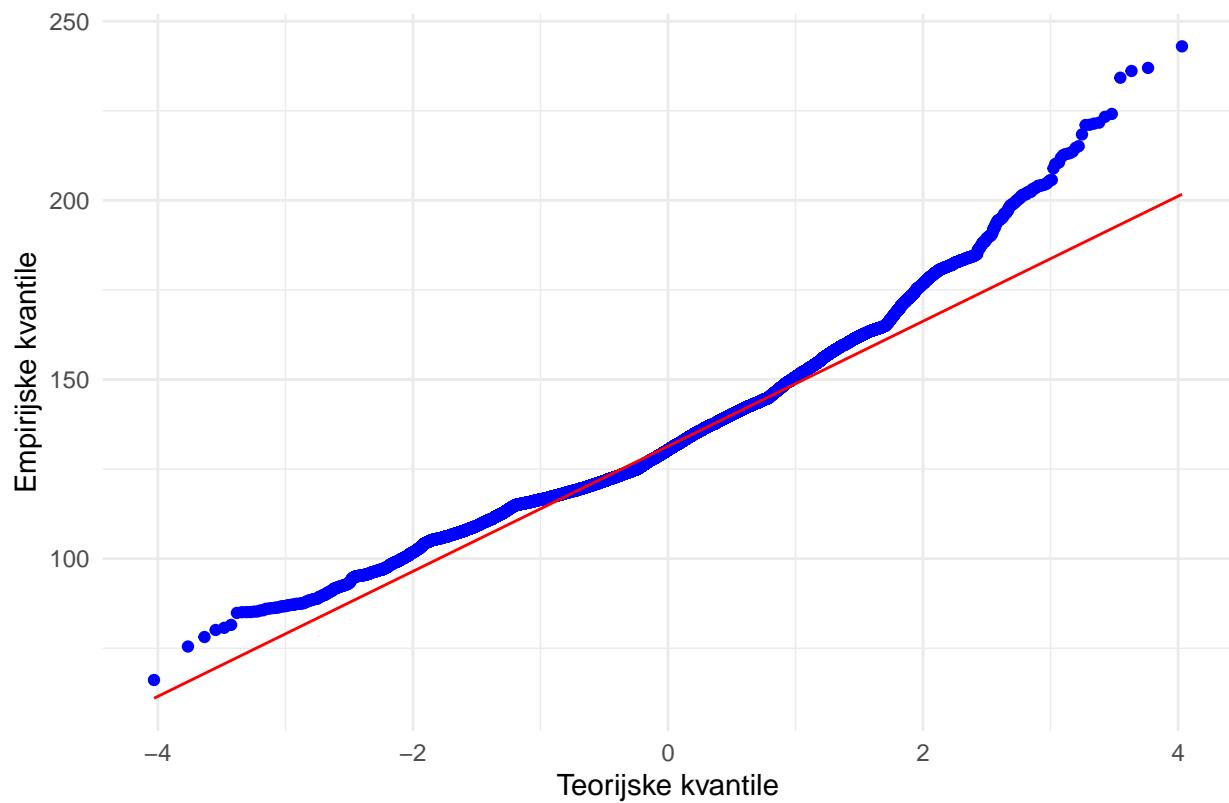
Distribucija sistolickog tlaka – Obese



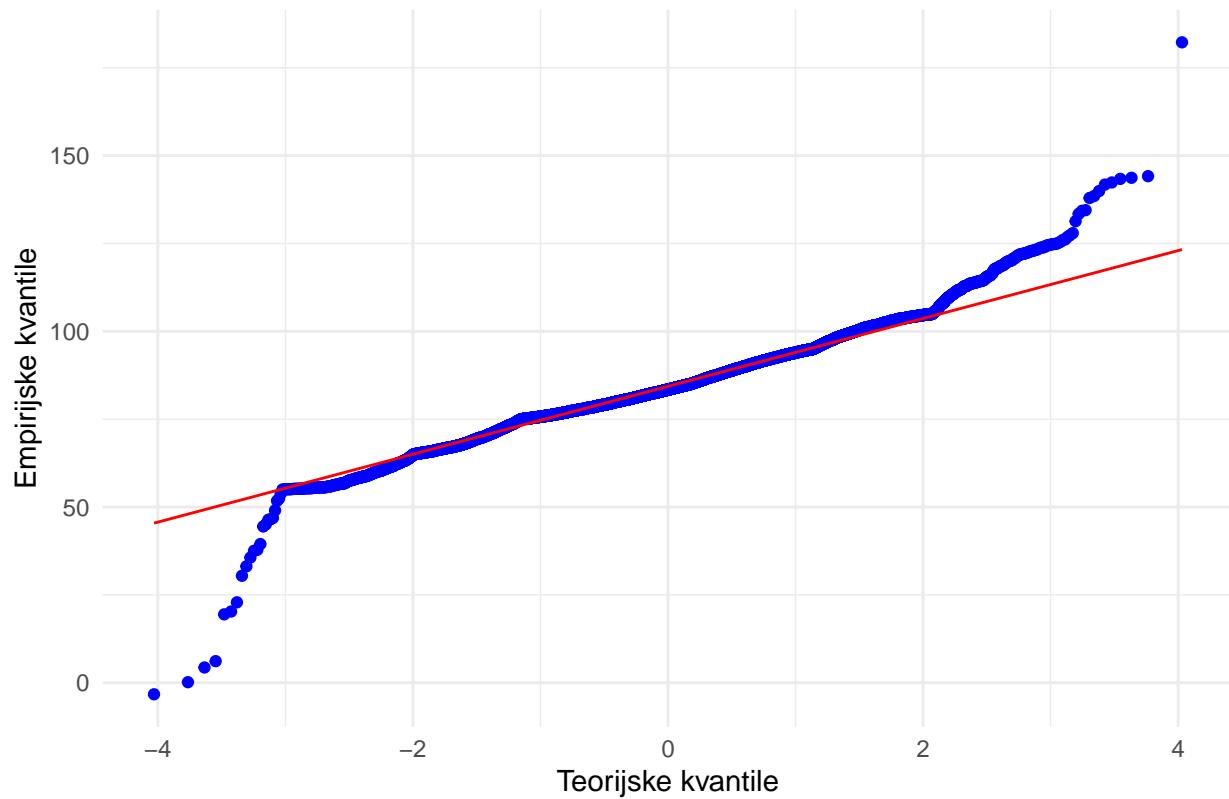
Distribucija dijastolickog tlaka – Obese



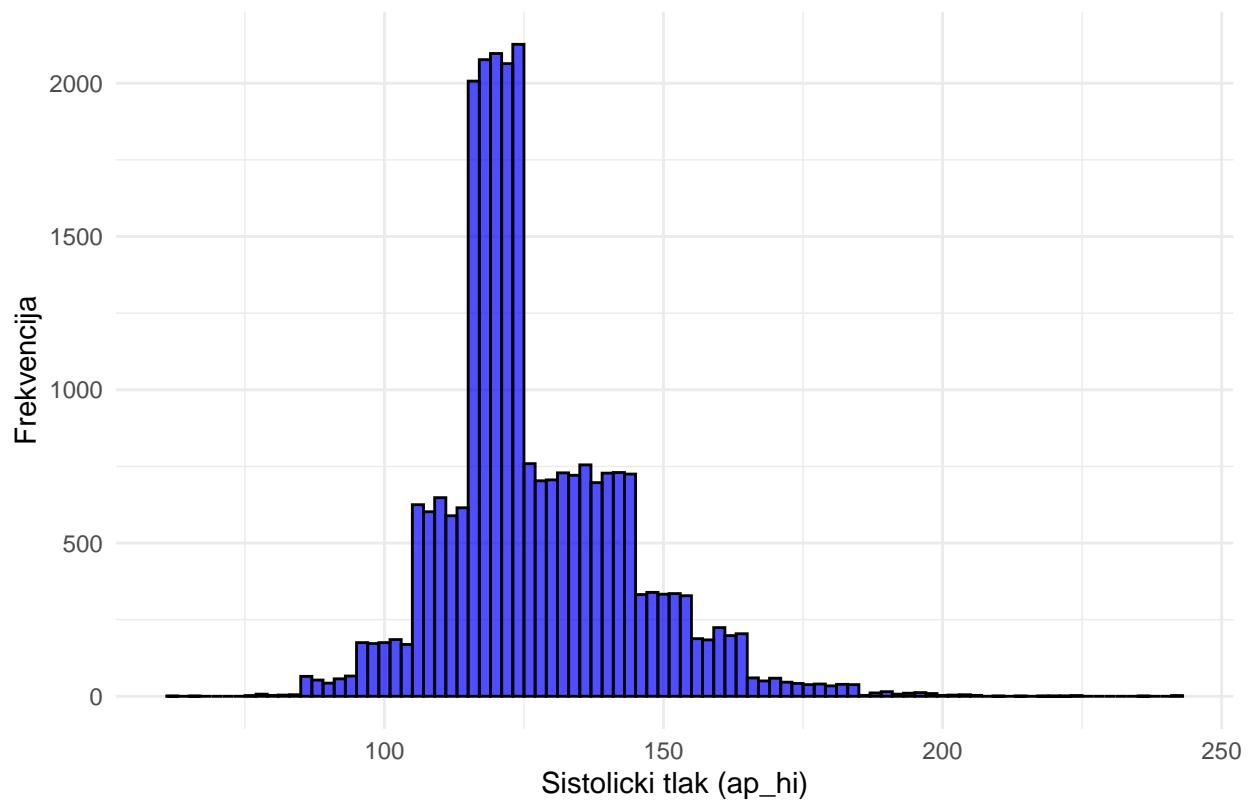
Q–Q plot – Sistolicki tlak (ap_hi) – Obese



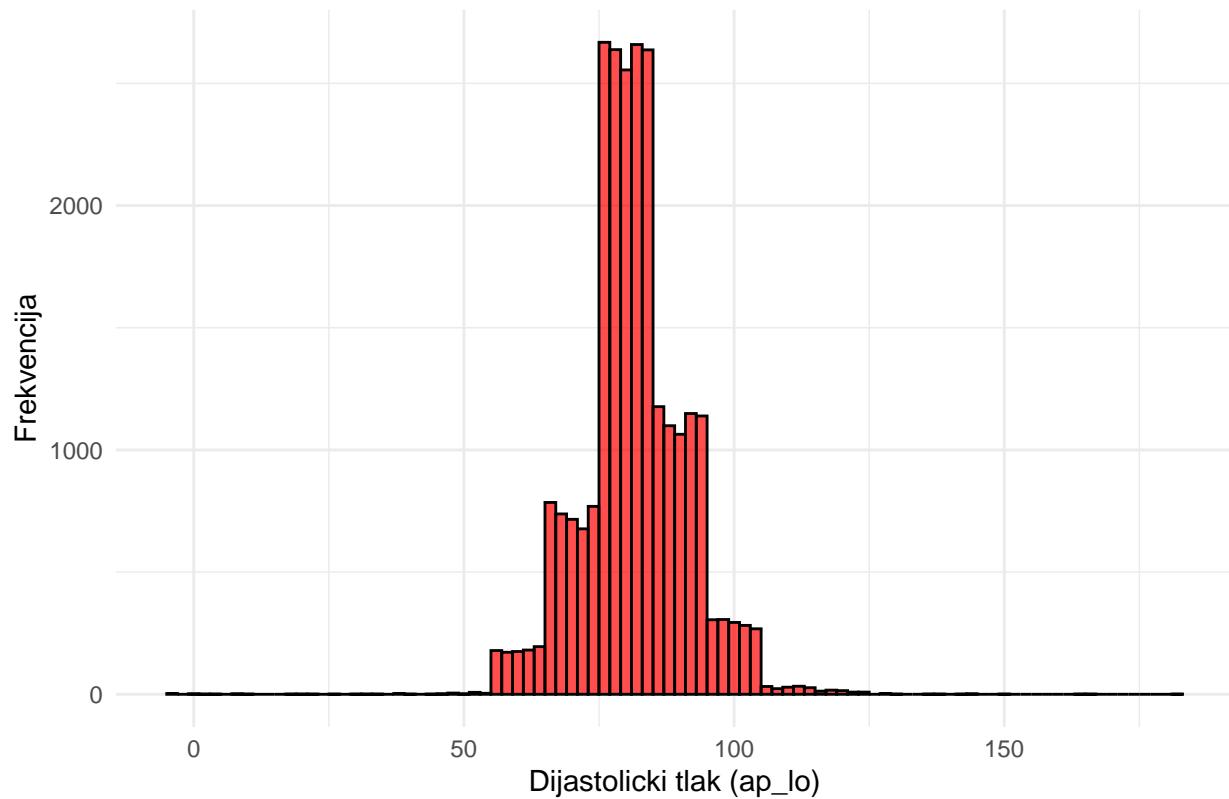
Q–Q plot – Dijastolicki tlak (ap_lo) – Obese



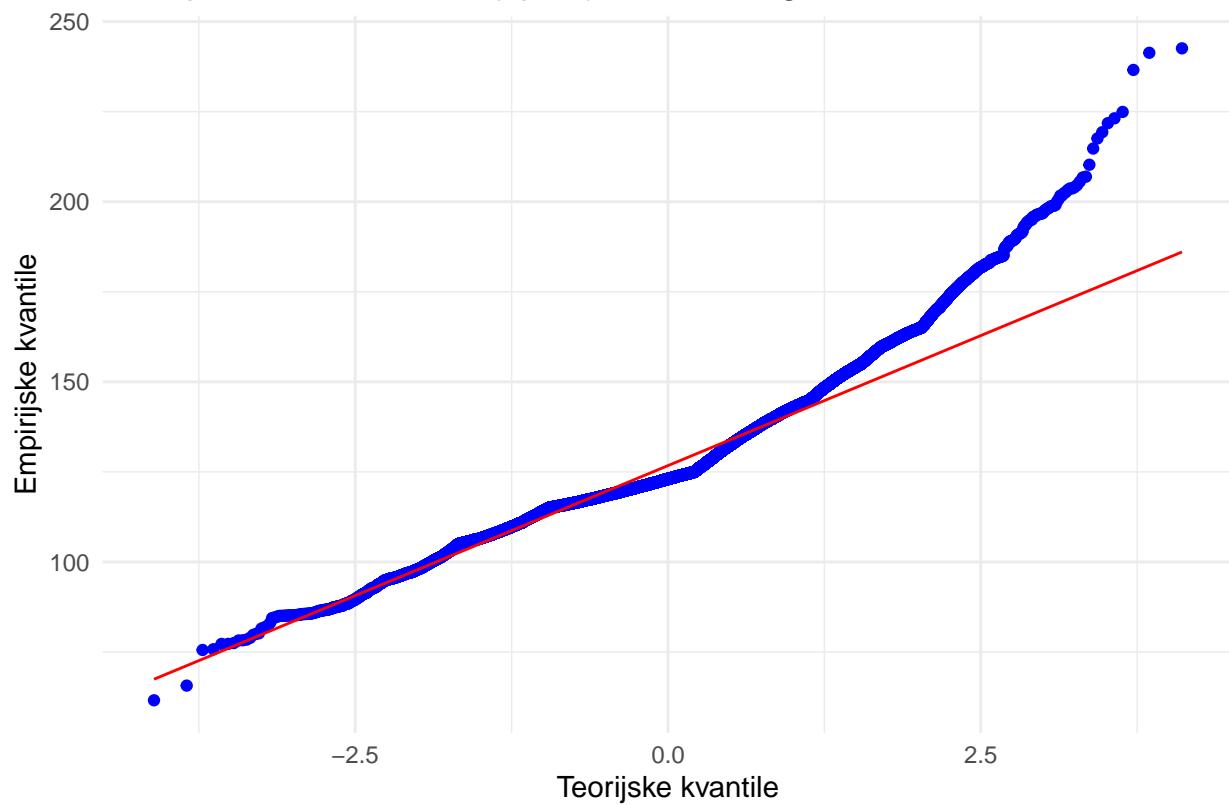
Distribucija sistolickog tlaka – Over Weight



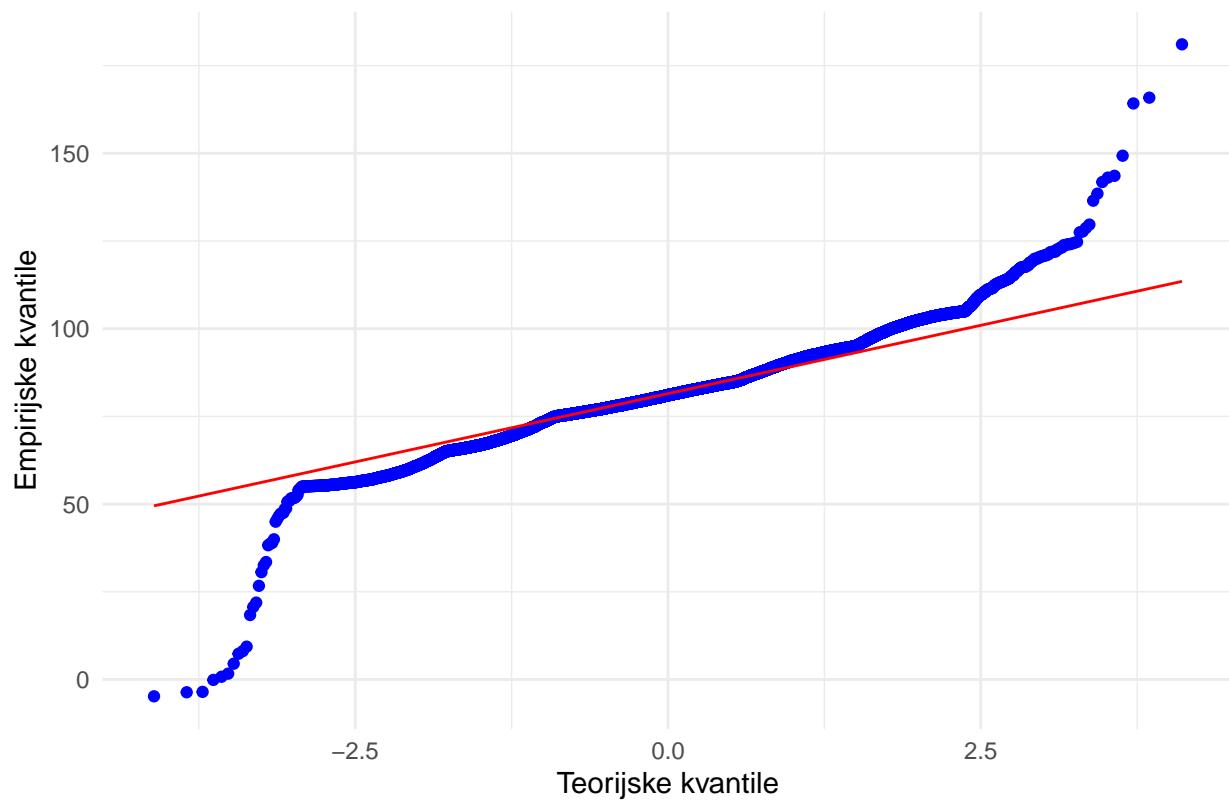
Distribucija dijastolickog tlaka – Over Weight



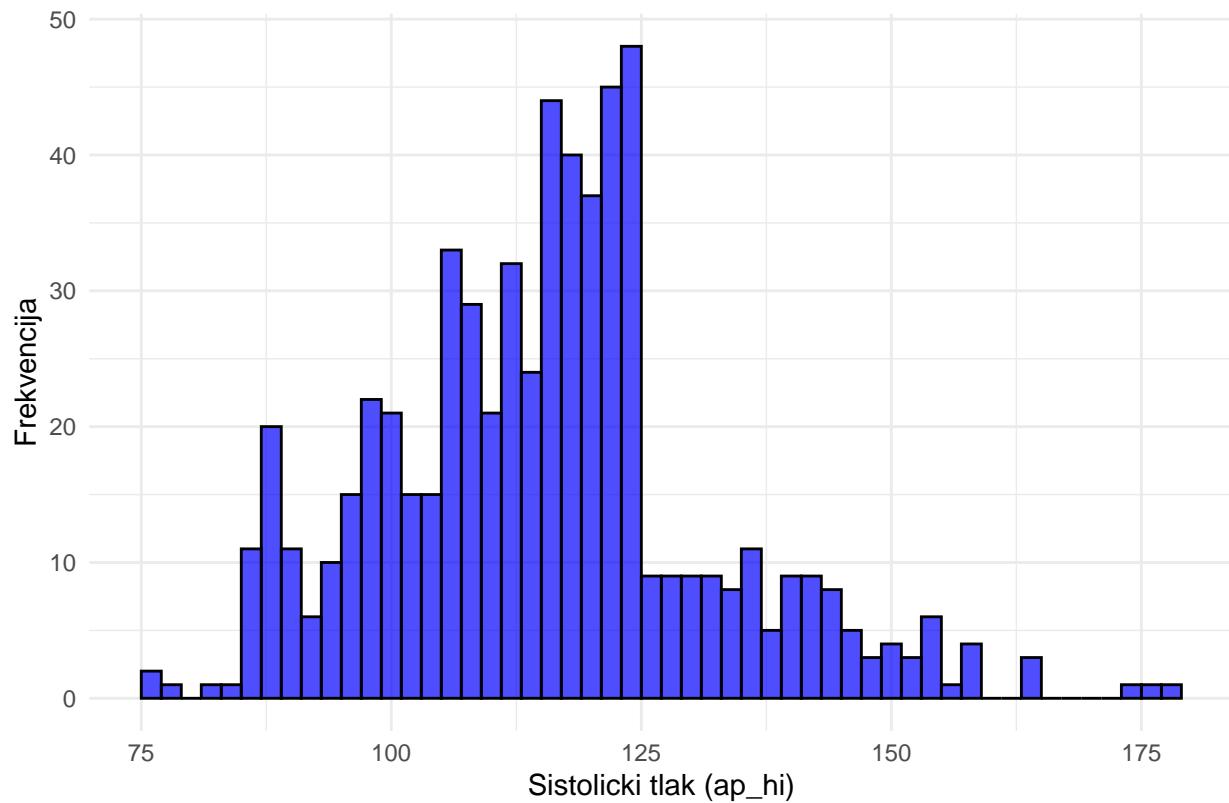
Q–Q plot – Sistolicki tlak (ap_hi) – Over Weight



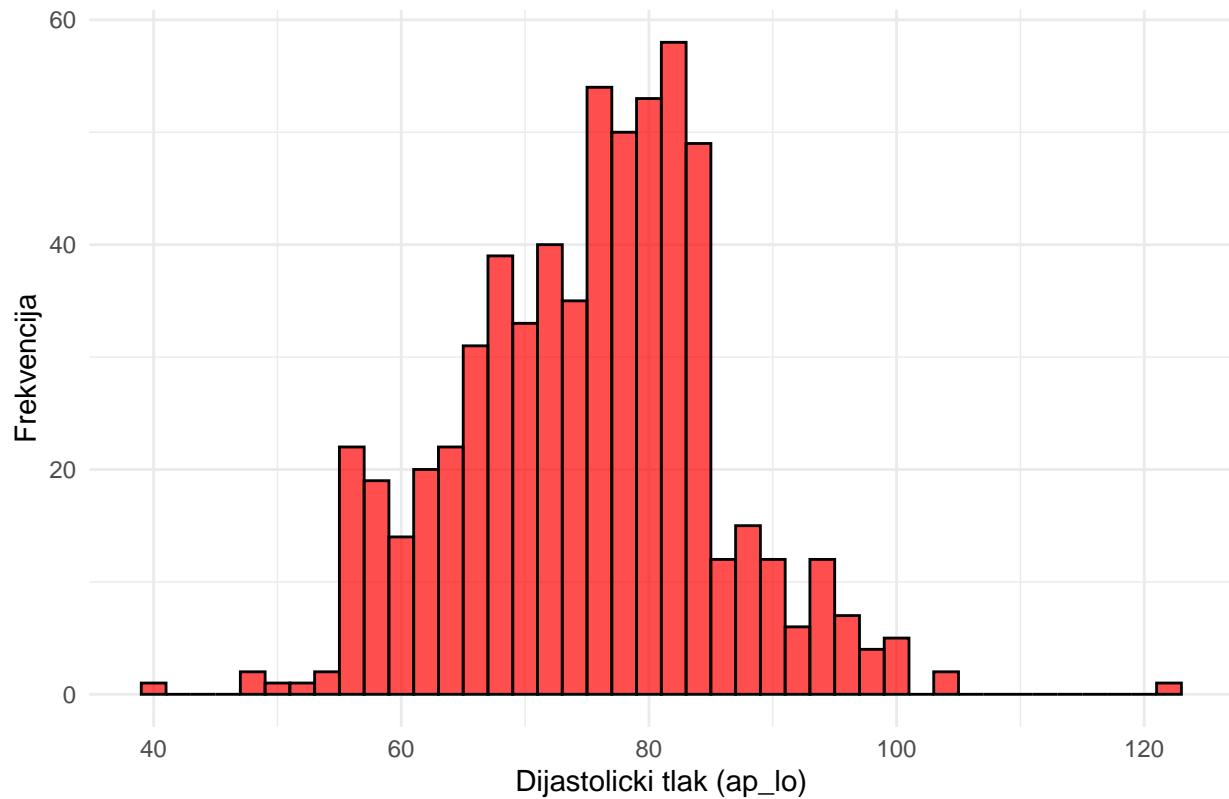
Q–Q plot – Dijastolicki tlak (ap_lo) – Over Weight



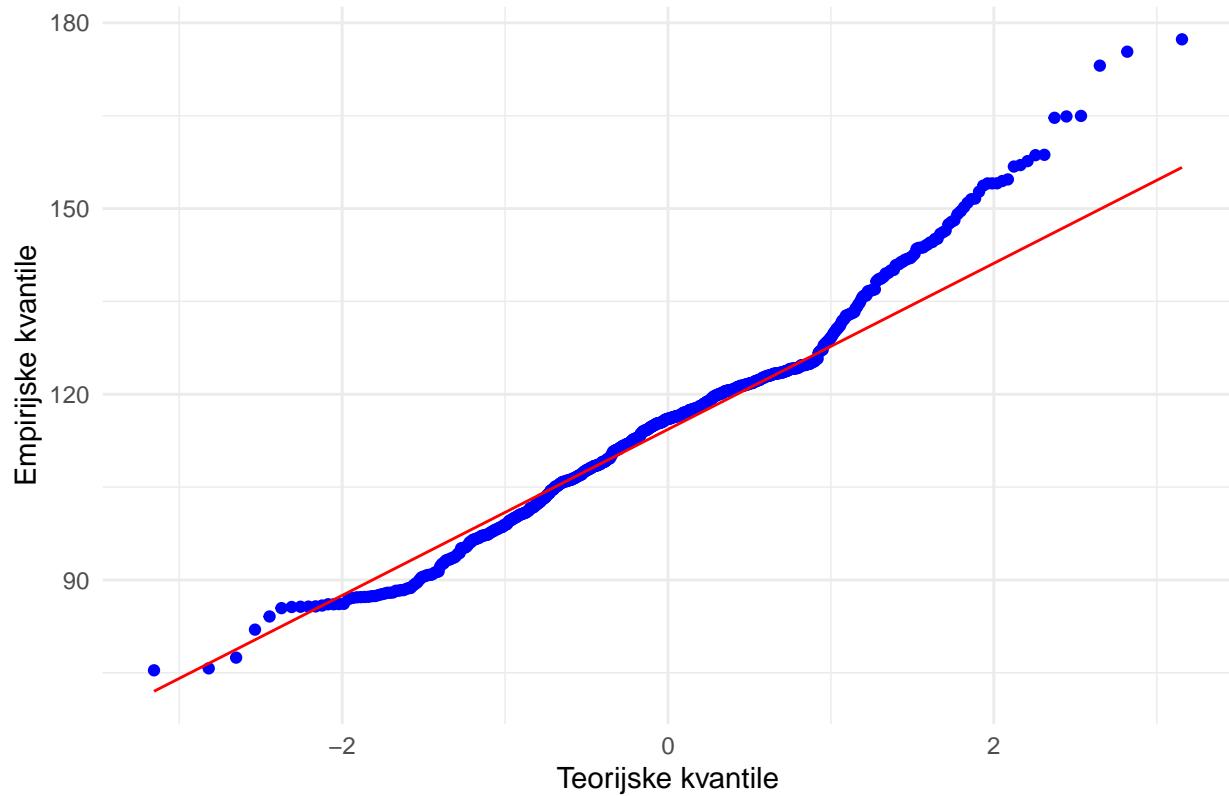
Distribucija sistolickog tlaka – Under Weight



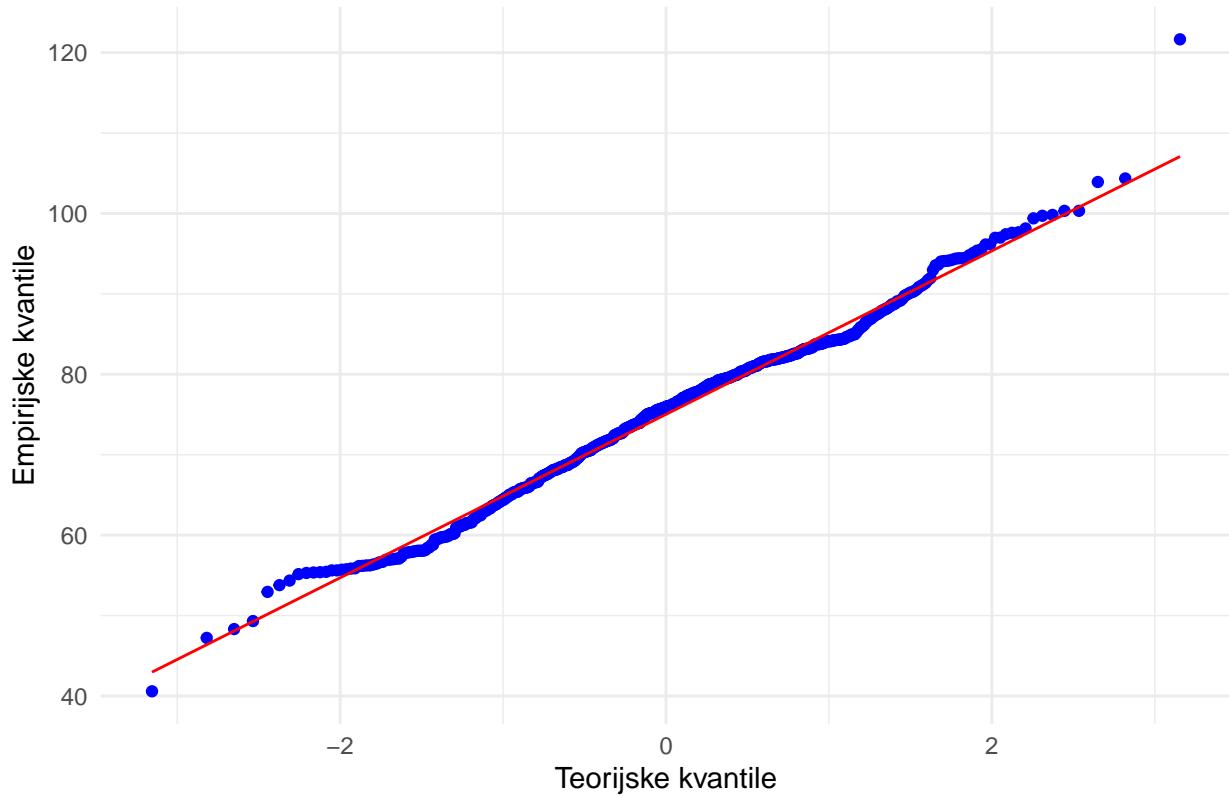
Distribucija dijastolickog tlaka – Under Weight



Q–Q plot – Sistolicki tlak (ap_hi) – Under Weight



Q-Q plot – Dijastolicki tlak (ap_lo) – Under Weight



```
# 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: 17948
##
## >> 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: 3.471774e-05
##
## >> Kolmogorov-Smirnov test - ap_lo <<
##   p-value: 0.09497015
##
## =====
##   Test homoskedasticnosti (BartlettTest)
## =====
##
## Bartlett test: ap_hi ~ BMICat
##
##   Bartlett test of homogeneity of variances
##
## data: ap_hi by BMICat
## Bartlett's K-squared = 889.45, 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 = 262.98, df = 3, p-value < 2.2e-16

# Kruskal-Wallis za sistolicki 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 = 4373.8, 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 = 3071.9, df = 3, p-value < 2.2e-16

#Zadatak 4 ##Kakav je odnos izmedu BMI-a i krvnog tlaka te možemo li predvidjeti krvni tlak na temelju
BMI-a, dobi i učestalosti tjelesne aktivnosti?
```

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?"

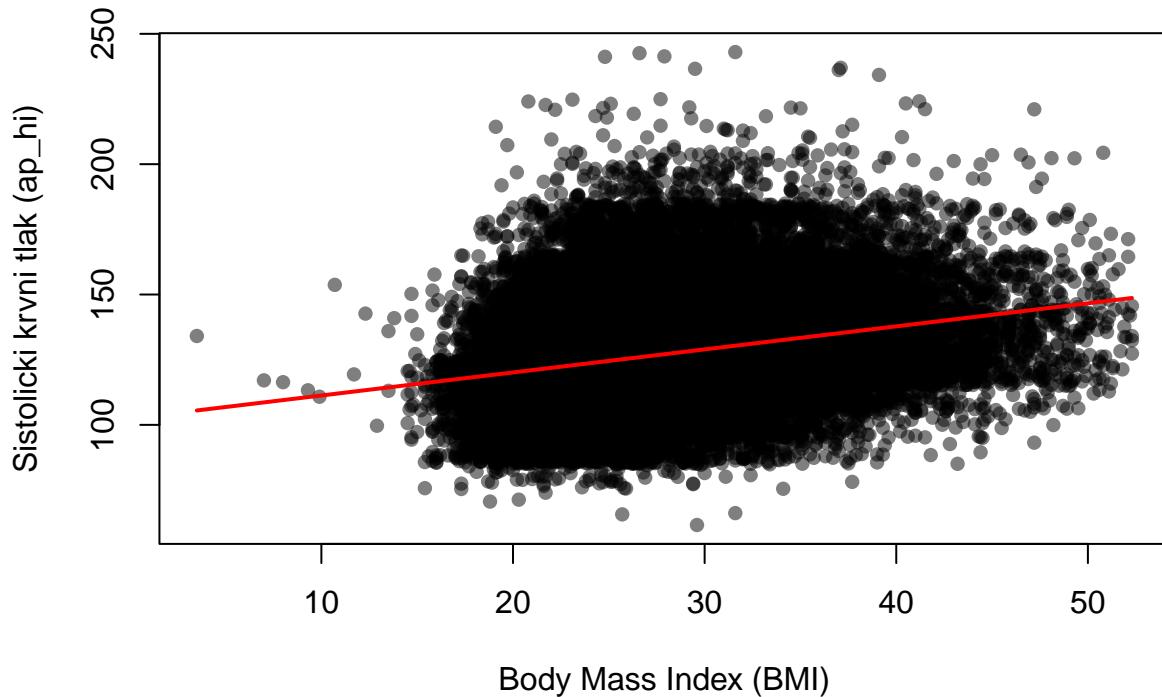
Sada ćemo metodom najmanjih kvadrata pokušati uspostaviti vezu između BMI-a i krvnog tlaka.

```
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)
```

Odnos sistolickog krvnog tlaka i BMI-a



```
summary(fit.ap_hi)

##
## Call:
## lm(formula = ap_hi ~ poly(BMI, 1), data = filtered_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -67.004 -9.693 -2.844  8.594 116.824 
##
## Coefficients:
```

```

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.266e+02  6.239e-02 2030.04   <2e-16 ***
## poly(BMI, 1) 1.179e+03  1.633e+01    72.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.33 on 68543 degrees of freedom
## Multiple R-squared:  0.07064, Adjusted R-squared:  0.07062
## F-statistic:  5210 on 1 and 68543 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. Takoder, 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 = 0.10152, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: standardized_residuals
## D = 0.10152, p-value < 2.2e-16

```

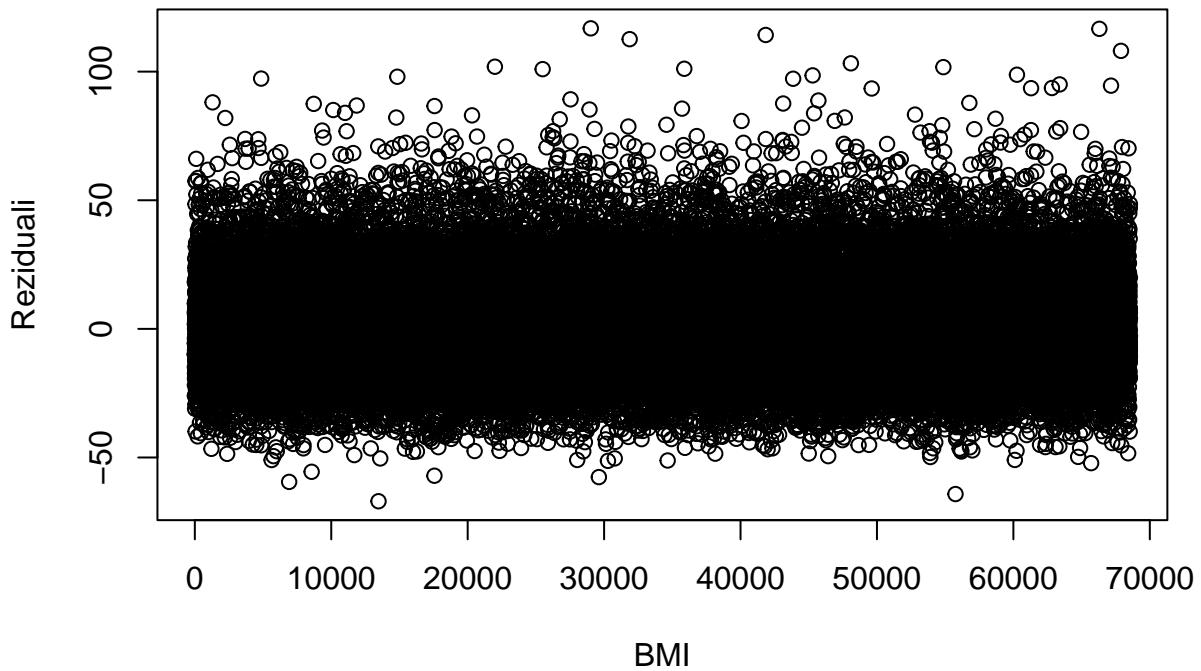
Reziduali padaju na testu normalnosti (i KS i Lilliefors). Možemo također vidjeti kako se reziduali ponašaju grafički:

```

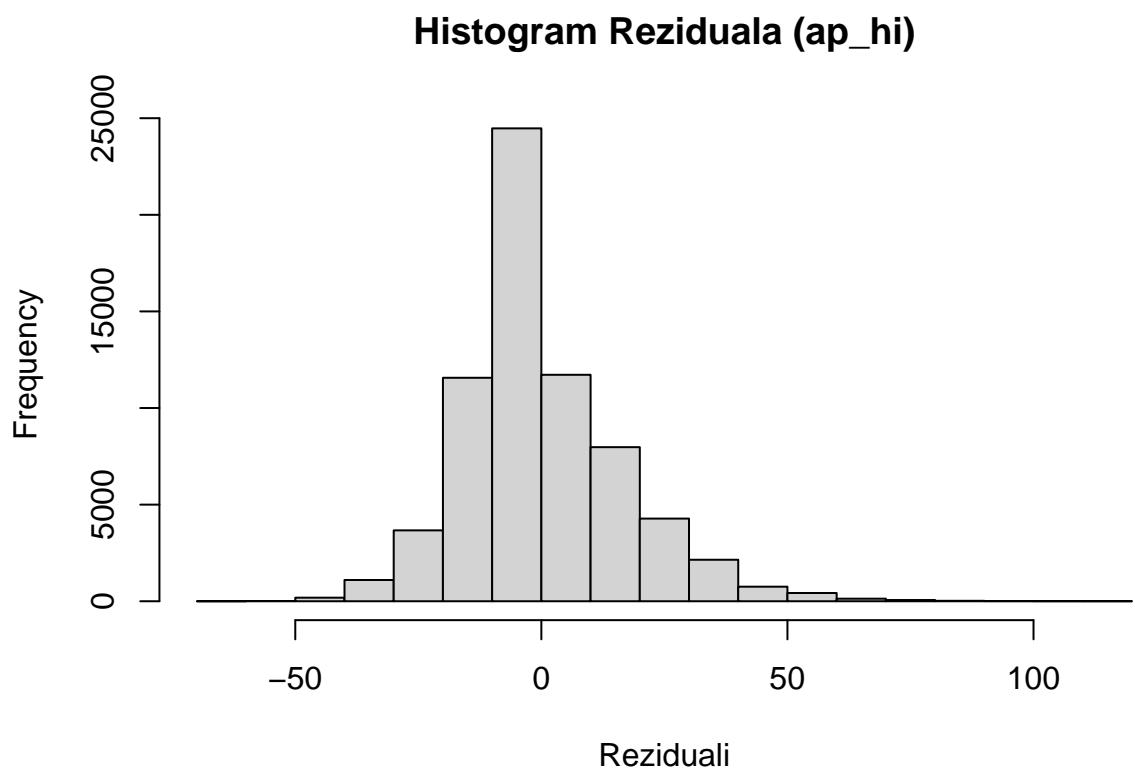
plot(fit.ap_hi$residuals,
      main = "Graf reziduala (ap_hi)",
      ylab = "Reziduali", xlab = "BMI")

```

Graf reziduala (ap_hi)

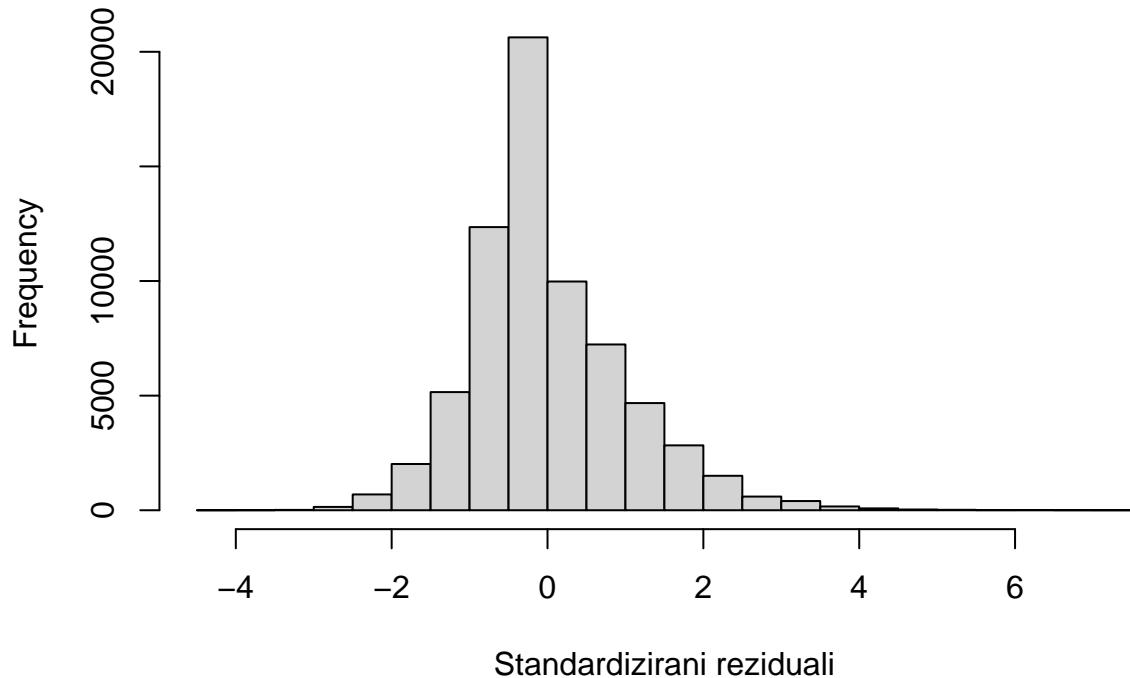


```
hist(fit.ap_hi$residuals,
      breaks = 20,
      main = "Histogram Reziduala (ap_hi)",
      xlab = "Reziduali")
```



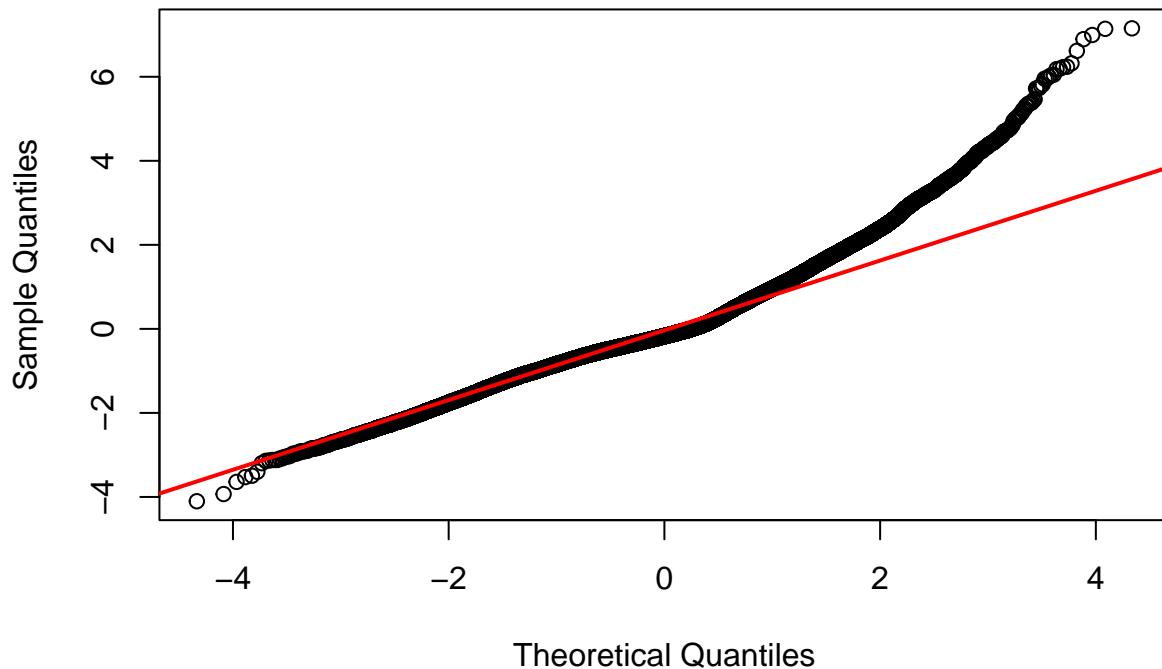
```
hist(rstandard(fit.ap_hi),
      breaks = 20,
      main = "Histogram standardiziranih reziduala (ap_hi)",
      xlab = "Standardizirani reziduali")
```

Histogram standardiziranih reziduala (ap_hi)



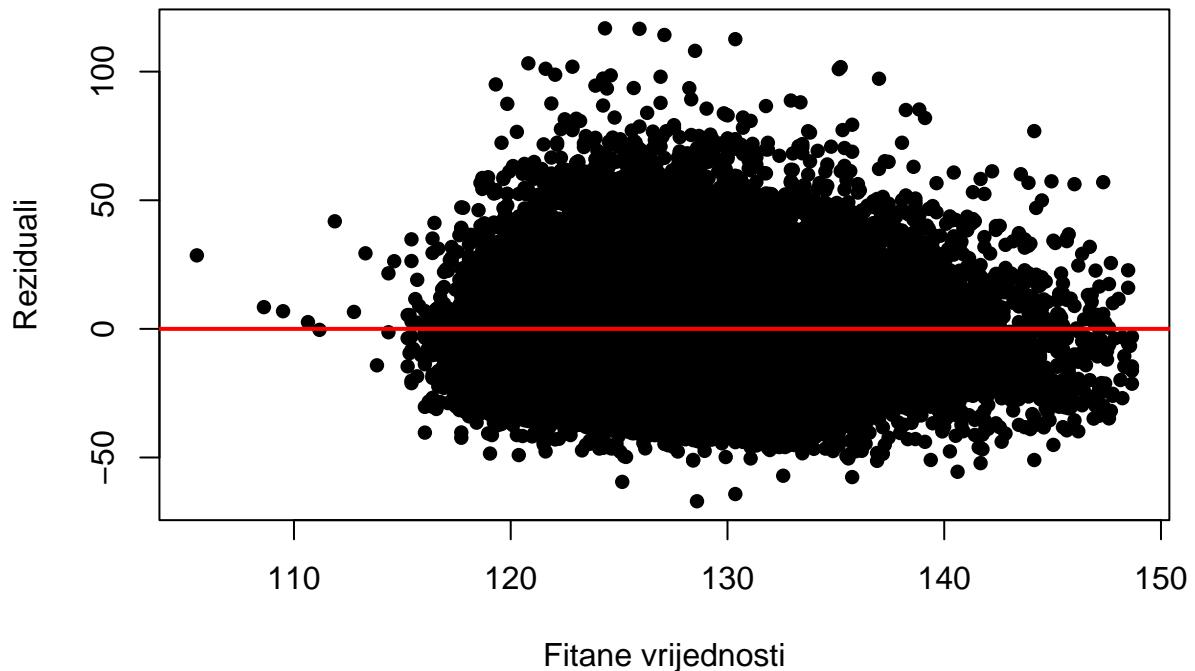
```
qqnorm(rstandard(fit.ap_hi),  
       main = "Q-Q plot standardiziranih reziduala (ap_hi)")  
qqline(rstandard(fit.ap_hi), col = "red", lwd = 2)
```

Q-Q plot standardiziranih reziduala (ap_hi)



```
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)
```

Reziduali u odnosu na fitane vrijednosti (ap_hi)



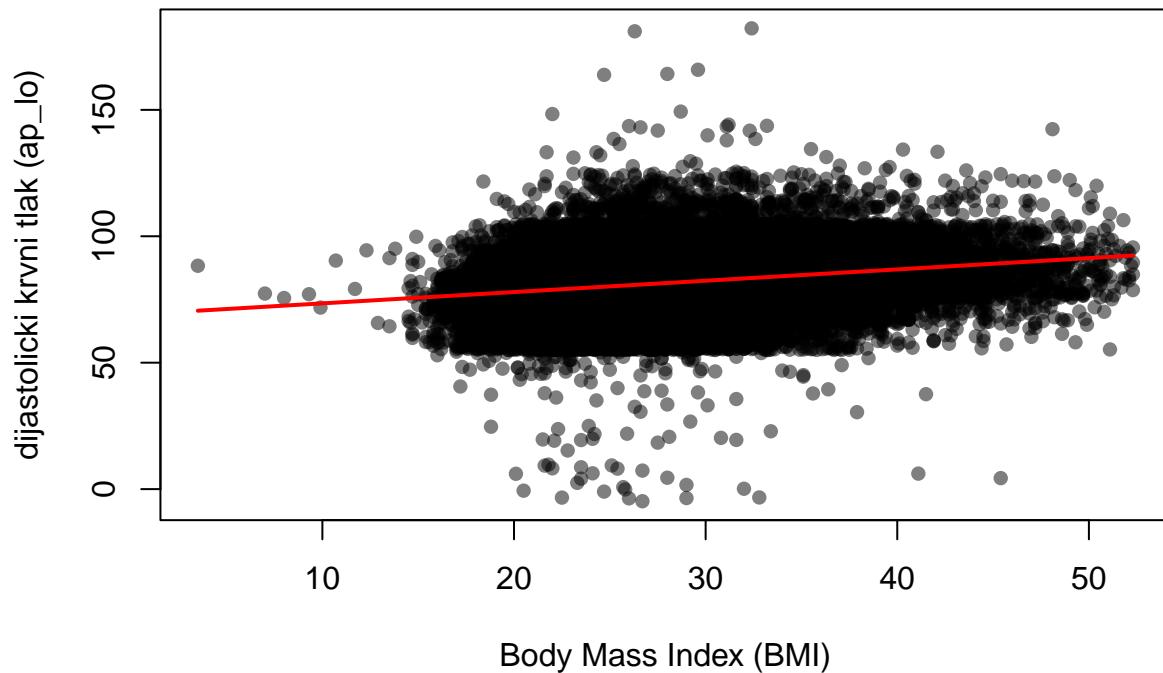
Q-Q plot nam govori da ova razdioba ima lakše repove od normalne, ali ovo svakako nije normalna distribucija. Sada možemo zaključiti da je nemoguće predvidjeti sistolički krvni tlak iz BMI-a (iz ovih podataka).

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)
```

Odnos dijatolickog krvnog tlaka i BMI-a



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

```
##  
## Call:  
## lm(formula = ap_lo ~ poly(BMI, 1), data = filtered_data)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -86.961  -5.286  -0.190   5.144 100.310  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  81.25027   0.03733 2176.65 <2e-16 ***  
## poly(BMI, 1) 596.88753   9.77292   61.08 <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 9.773 on 68543 degrees of freedom  
## Multiple R-squared:  0.05161,    Adjusted R-squared:  0.0516  
## F-statistic:  3730 on 1 and 68543 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

```

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 = 0.10152, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  standardized_residuals
## D = 0.10152, p-value < 2.2e-16

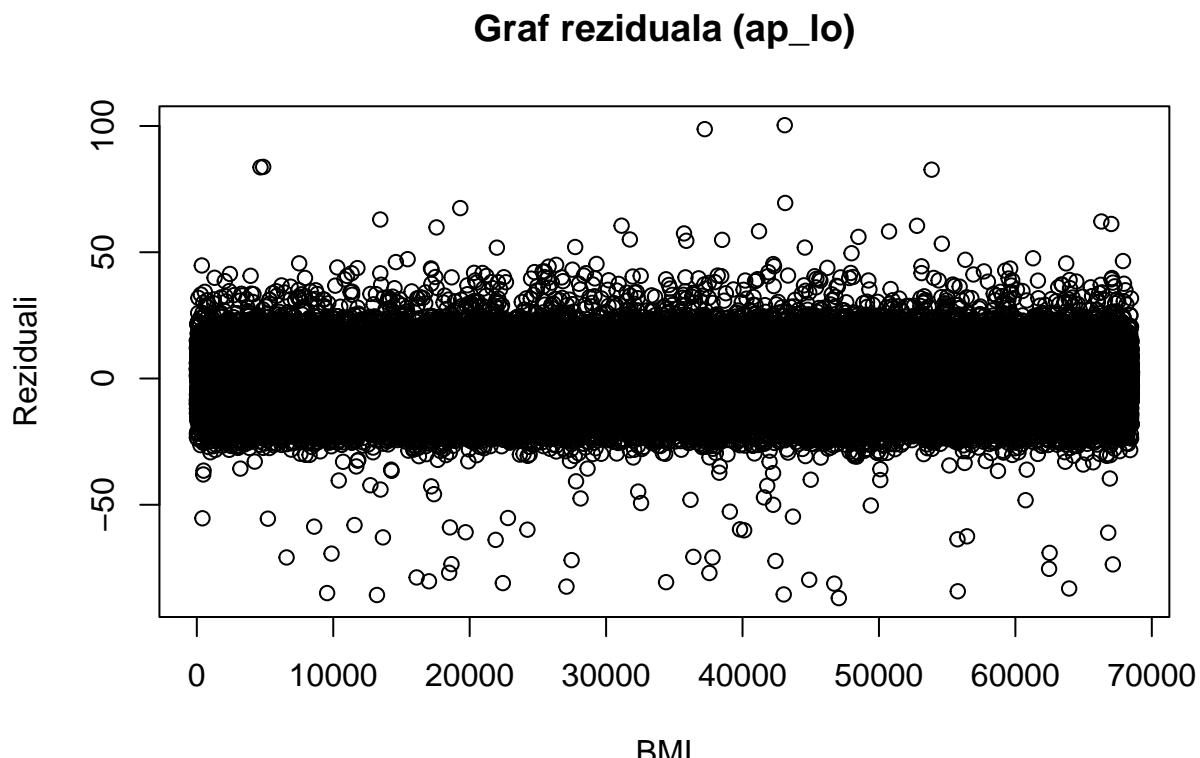
```

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

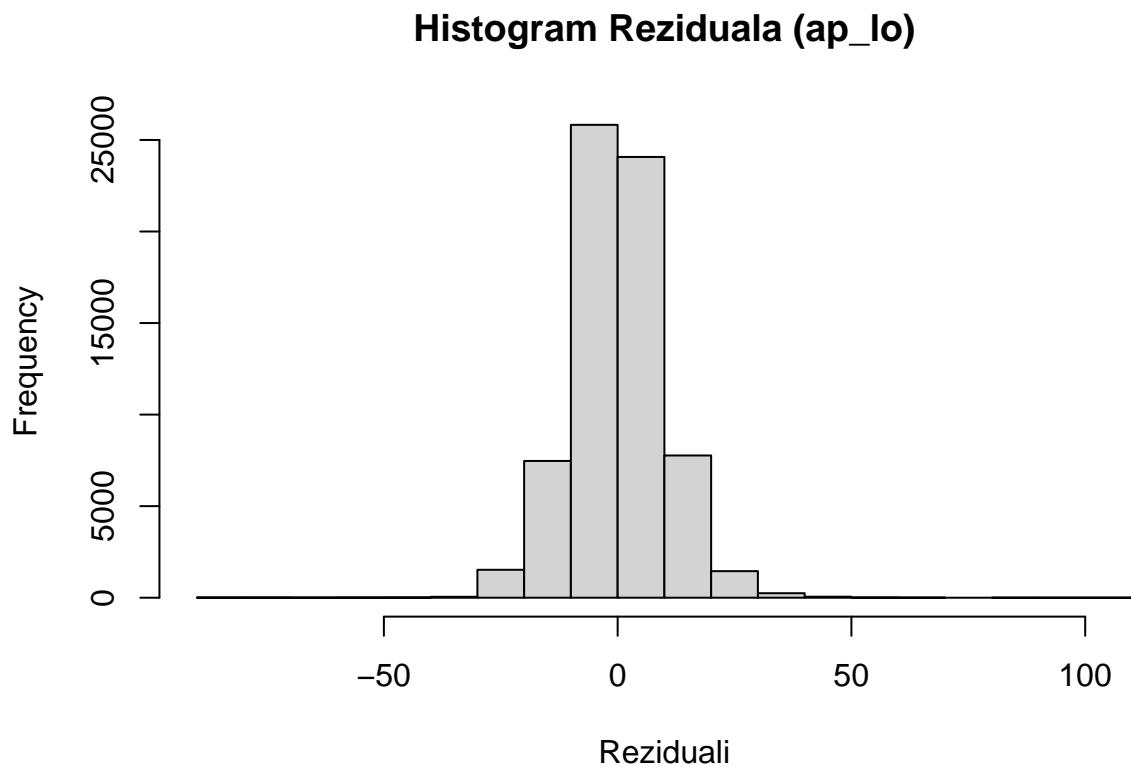
```

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

```

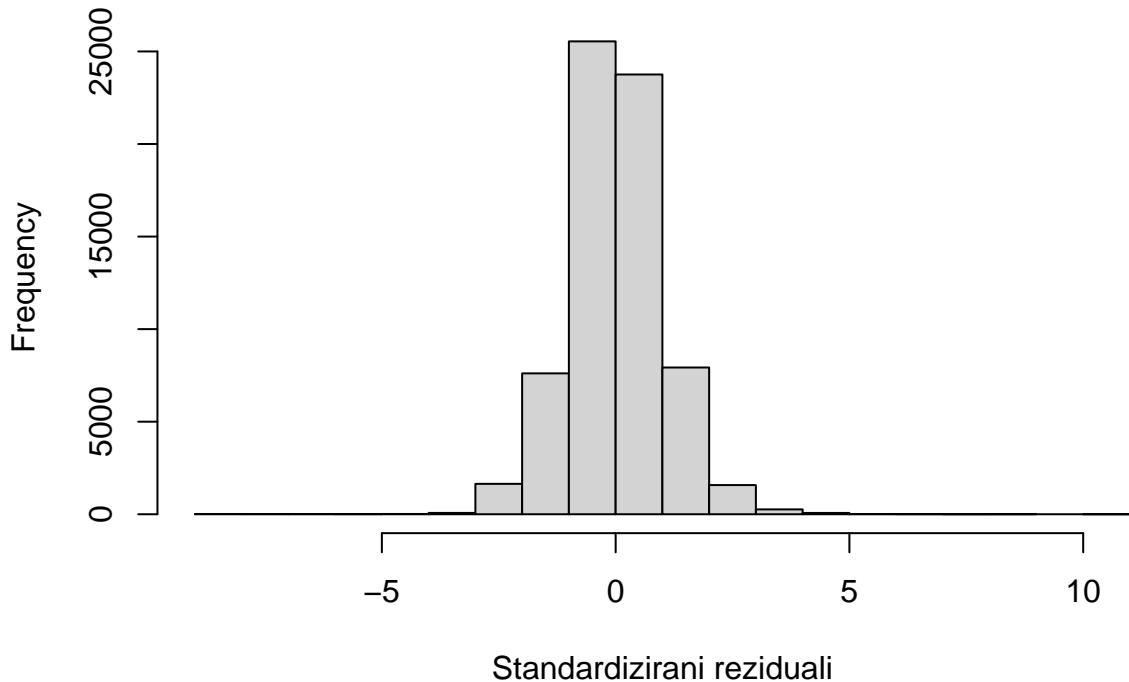


```
hist(fit.ap_lo$residuals,
     breaks = 20,
     main = "Histogram Reziduala (ap_lo)",
     xlab = "Reziduali")
```



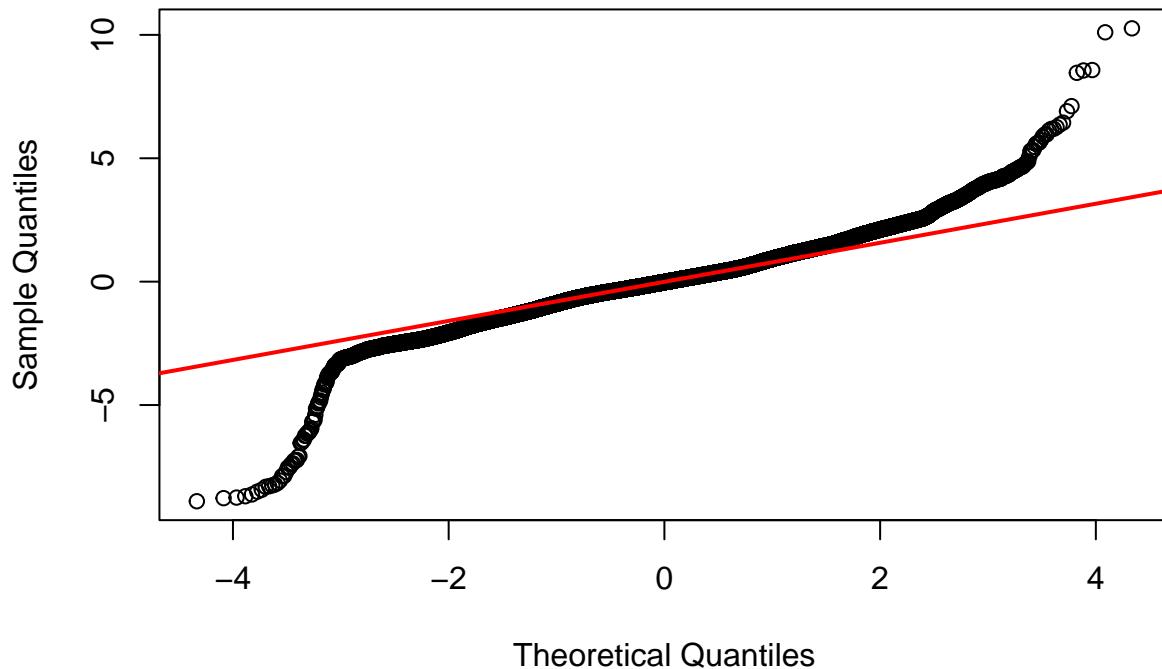
```
hist(rstandard(fit.ap_lo),
      breaks = 20,
      main = "Histogram standardiziranih reziduala (ap_lo)",
      xlab = "Standardizirani reziduali")
```

Histogram standardiziranih reziduala (ap_lo)



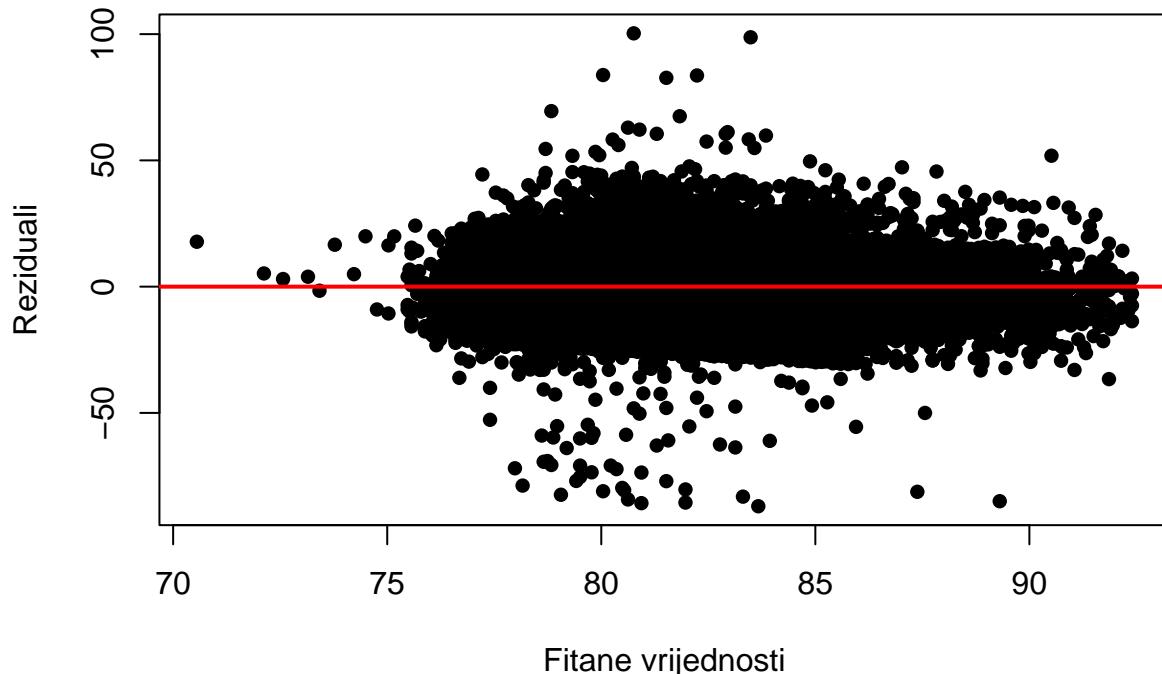
```
qqnorm(rstandard(fit.ap_lo),  
       main = "Q-Q plot standardiziranih reziduala (ap_lo)")  
qqline(rstandard(fit.ap_lo), col = "red", lwd = 2)
```

Q-Q plot standardiziranih reziduala (ap_lo)



```
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)
```

Reziduali u odnosu na fitane vrijednosti (ap_lo)



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 regresiji 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 kategorijalska 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 -0.01480864 -0.01013567
## [2,] -0.01480864  1.00000000  0.10526297
## [3,] -0.01013567  0.10526297  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)
```

```

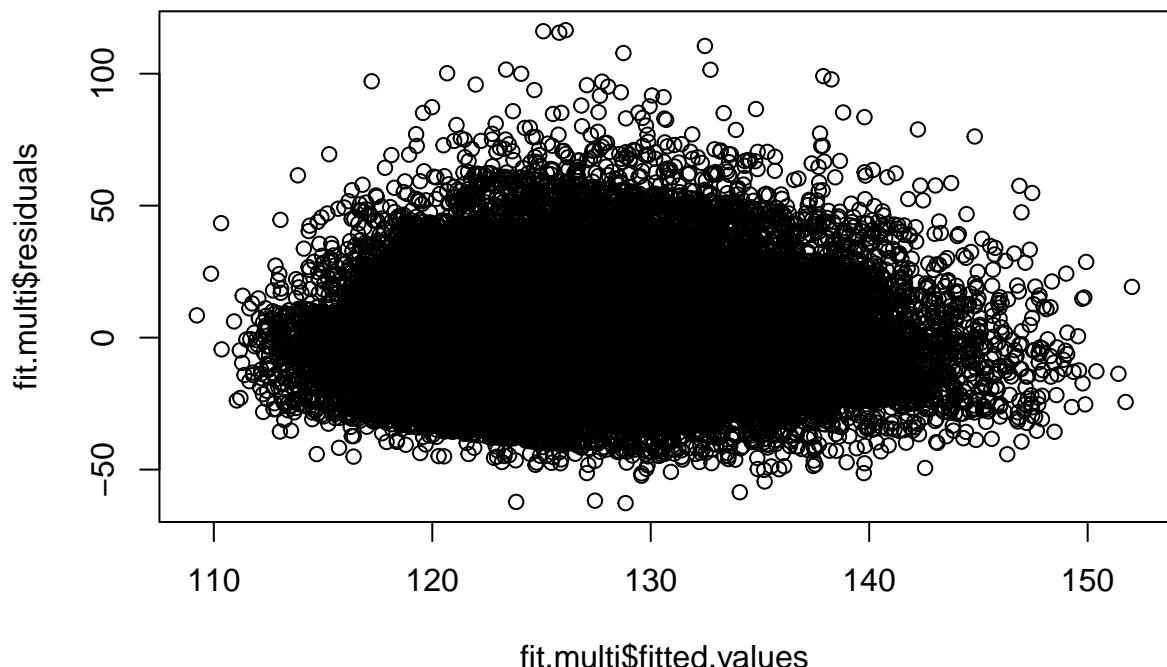
## 
## Residuals:
##   Min     1Q Median     3Q    Max 
## -62.701 -9.811 -2.666  8.121 116.472 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 80.031600  0.573624 139.519 <2e-16 ***
## BMI         0.820780  0.012101  67.830 <2e-16 ***
## active      0.189236  0.154243   1.227    0.22    
## AgeinYr     0.453819  0.009105  49.841 <2e-16 ***  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 16.05 on 68541 degrees of freedom
## Multiple R-squared:  0.1031, Adjusted R-squared:  0.1031 
## F-statistic:  2628 on 3 and 68541 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)
```



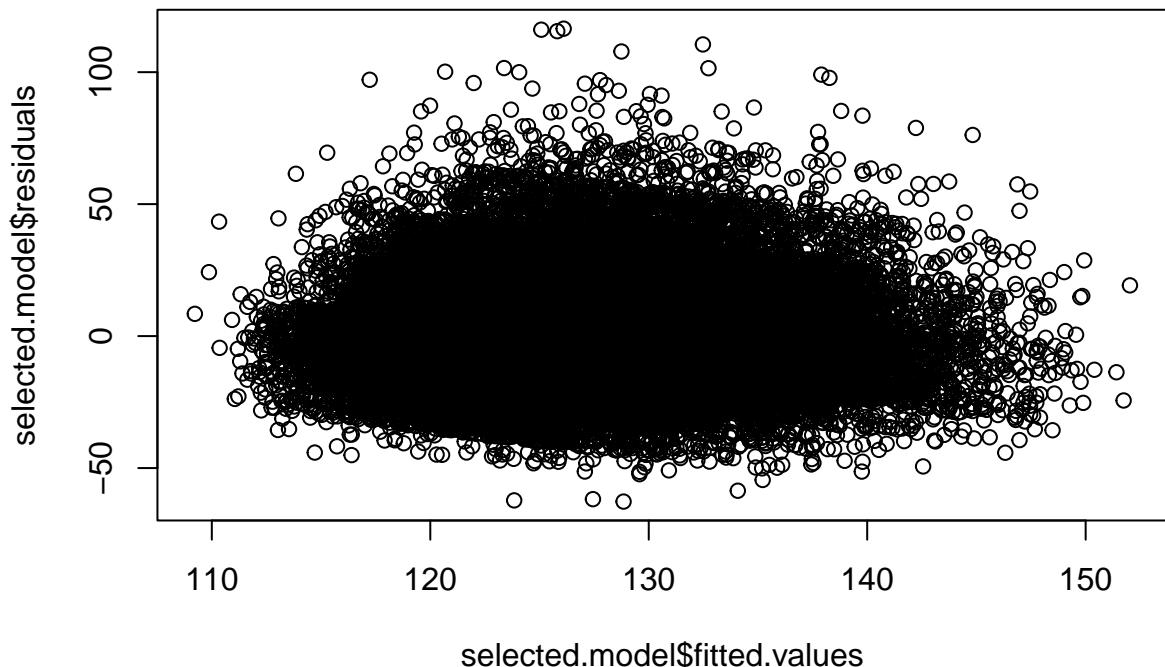
```
#KS test na normalnost
ks.test(rstandard(fit.ap_hi), 'pnorm')
```

```

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

##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: rstandard(fit.ap_hi)
## D = 0.10152, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fit.ap_hi)
## D = 0.10152, p-value < 2.2e-16
selected.model = fit.multi
plot(selected.model$fitted.values,selected.model$residuals)

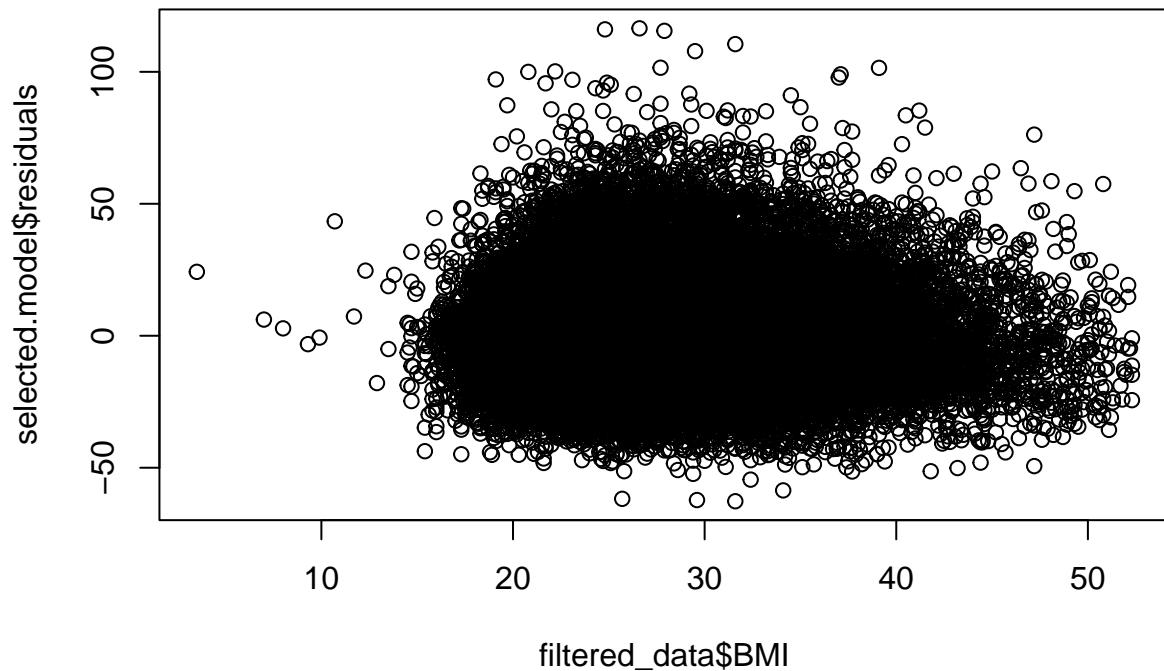
```



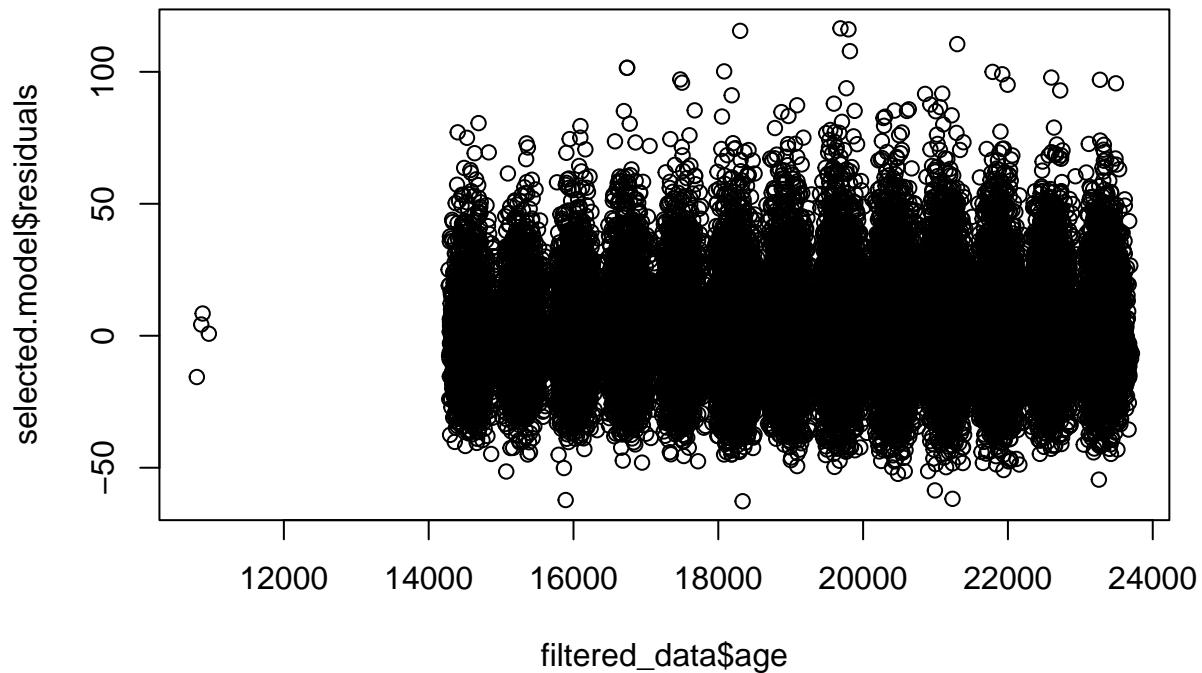
```

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

```



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

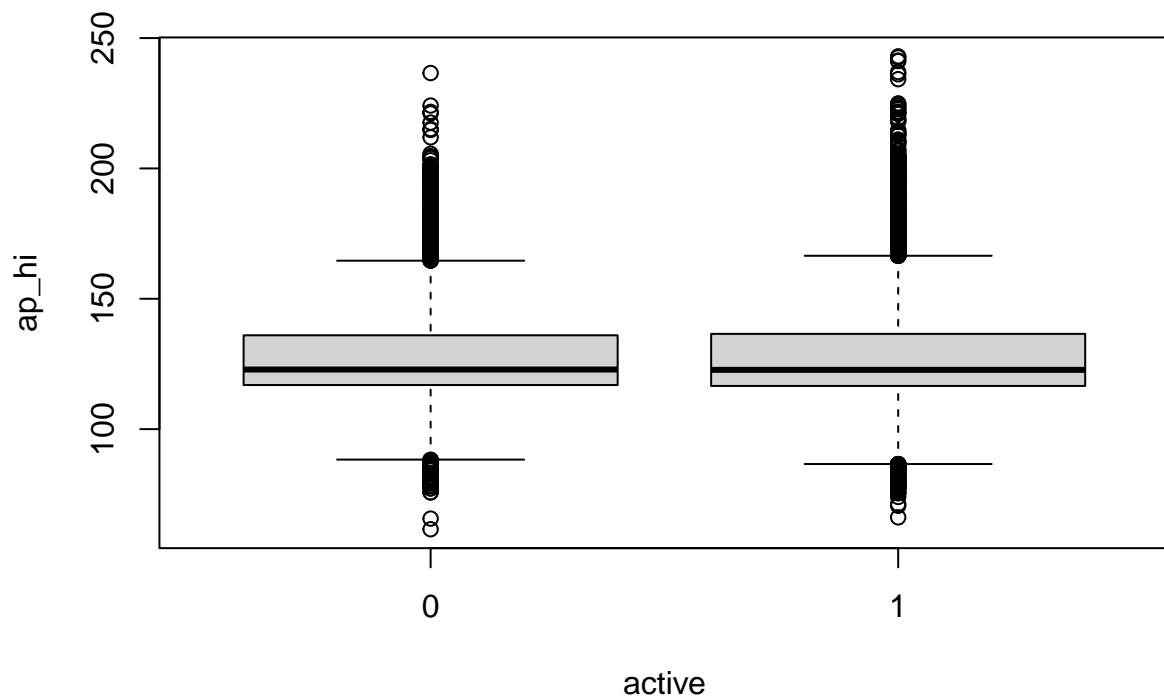


```

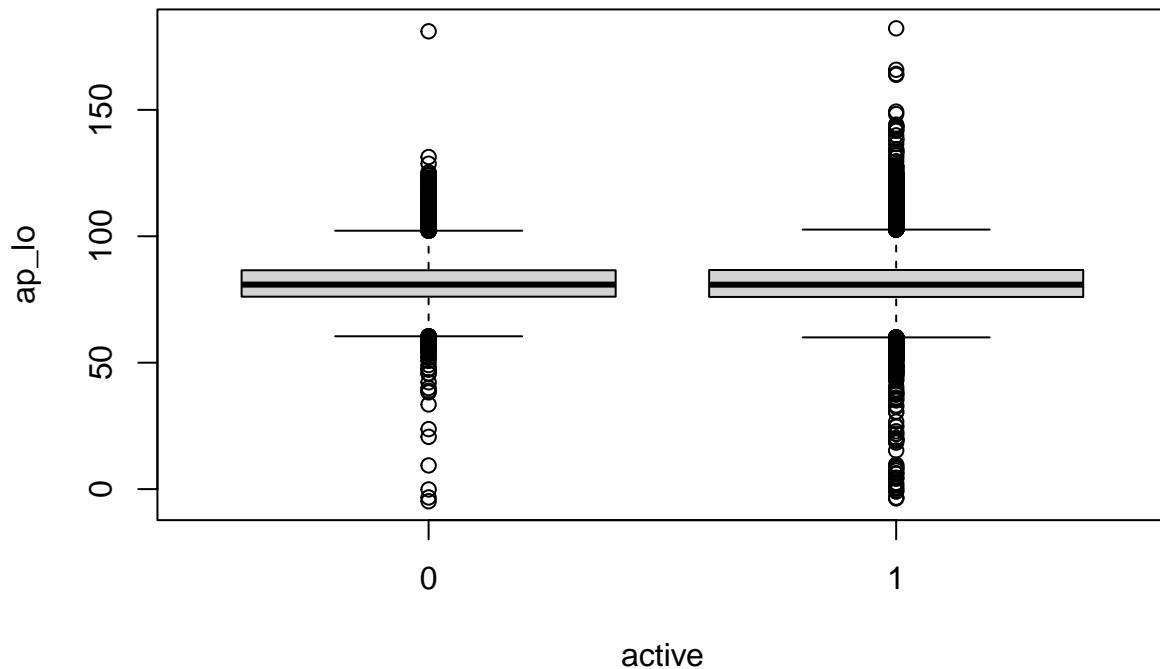
ks.test(rstandard(fit.multi), 'pnorm')
require(nortest)
lillie.test(rstandard(fit.multi))

##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: rstandard(fit.multi)
## D = 0.089646, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fit.multi)
## D = 0.089646, 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] 126.6828
## [1] 126.6379
## [1] 81.30542
## [1] 81.23678

```

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

#Zaključak projekta

Na temelju provedene analize možemo zaključiti da skup podataka o zdravstvenim informacijama nudi važne uvide, ali i pokazuje određene nedostatke koji utječu na kvalitetu interpretacije. Uočena je visoka prisutnost zdravih kategorija kolesterola u svim skupinama, dok starije dobne skupine pokazuju povećan rizik za opasne razine kolesterola. Slično tome, distribucija BMI-ja i krvnog tlaka ukazuje na određene tendencije, ali analize sugeriraju da predikcija krvnog tlaka na temelju BMI-a i dodatnih varijabli ima ograničenu pouzdanost.

Pristranosti, poput zaokruživanja izmjerjenih vrijednosti krvnog tlaka, dodatno otežavaju interpretaciju i ukazuju na potrebu za unapređenjem metoda prikupljanja podataka. Također, razlike u krvnom tlaku između pušača i nepušača, kao i između aktivnih i neaktivnih osoba, nisu bile praktično značajne.

Unatoč ograničenjima, analiza je ukazala na smjerove za daljnje istraživanje, osobito u pogledu povezanosti zdravstvenih parametara i načina prikupljanja podataka. Ovi uvidi mogu poslužiti kao osnova za buduća istraživanja i razvoj pouzdanijih prediktivnih modela.