

[SAP] Projektni zadatak - Analiza podataka zdravstvenog pregleda

Statistički najizglednije ime

2025-01-26

Tema projekta

U ovom projektu analiziramo medicinske podatke i pokušavamo uz pomoć statističkih testova odgovoriti na razna pitanja o krvnom tlaku i njegovoj ovisnosti o dobi, indeksu tjelesne mase, životnim navikama i drugim varijablama čije su vrijednosti dostupne u uzorku. Za provedbu analize korišten je programski jezik R.

Inicijalni koraci projekta

Učitavanje podataka

Filtriranje podataka

Početna faza analize uključila je čišćenje podataka korištenjem više kriterija. Eliminirane su sve observacije s neologičnim vrijednostima tlaka (npr. negativne vrijednosti) te slučajevi gdje maksimalni tlak prelazi minimalni. Dodatno, izbačeni su ekstremni BMI-ovi te restrigirani na rasponu 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

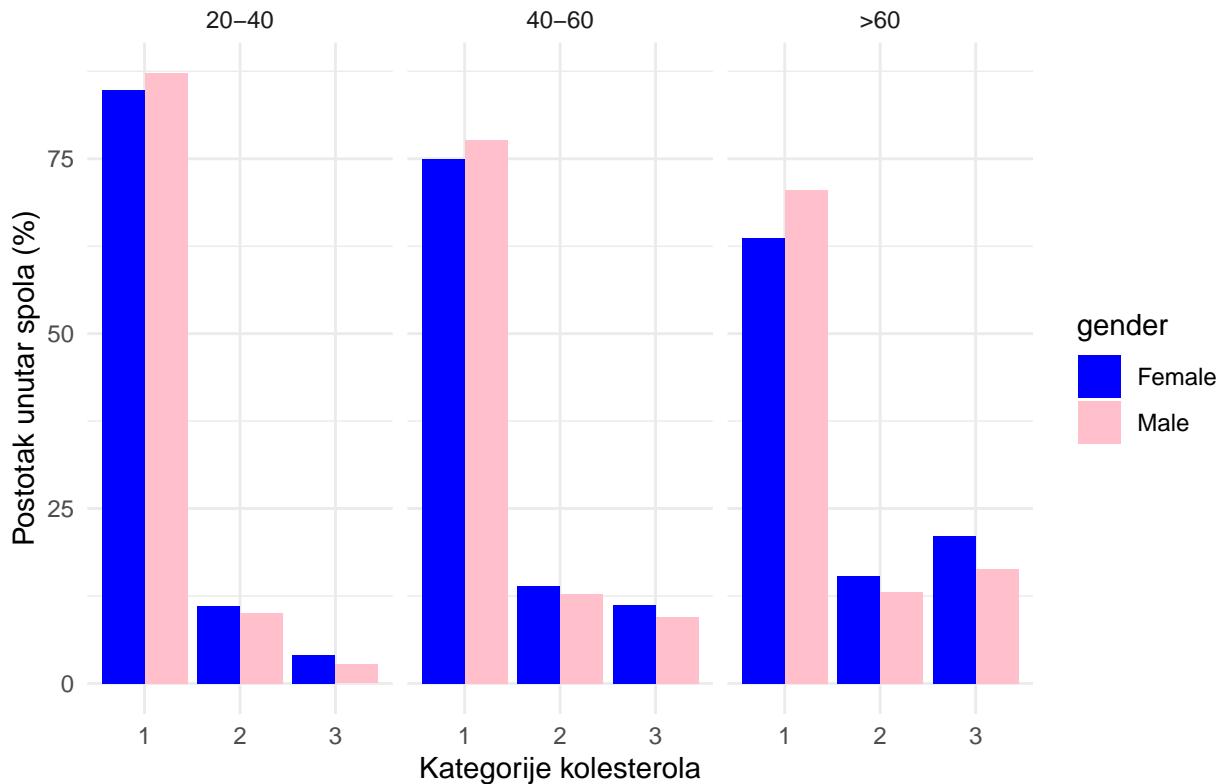
distribution <- filtered_data %>%
  group_by(AgeGroup, gender, cholesterol) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(AgeGroup, gender) %>%
  mutate(percentage = count / sum(count) * 100)

#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 doboj 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 spolove. Vizualno odvajanje kategorija ostvareno je paletom boja: zelenom za zdravu, žutom za rizičnu i ljubičastom za opasnu razinu.

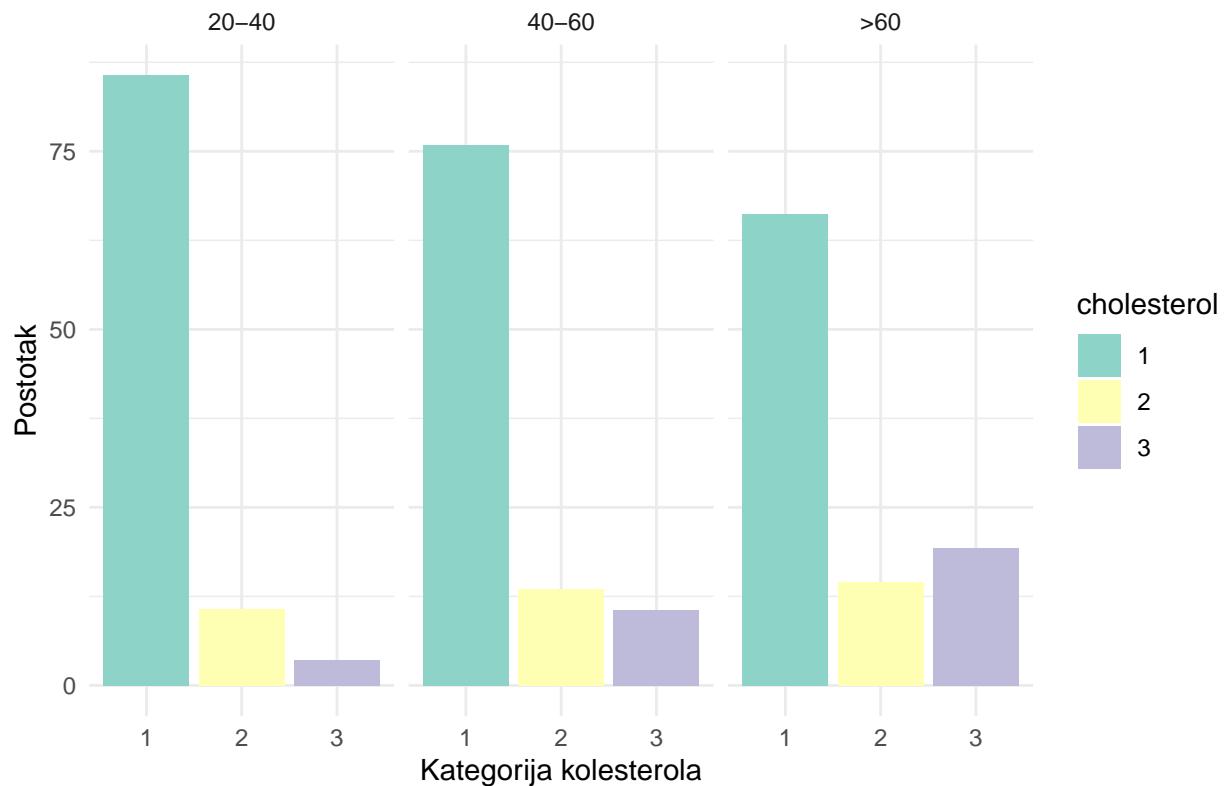
Zdrav (1) - prikazan zelenkastom bojom,
 Rizičan (2) - prikazan žutom bojom,
 Opasan (3) - prikazan ljubičastom bojom.

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:

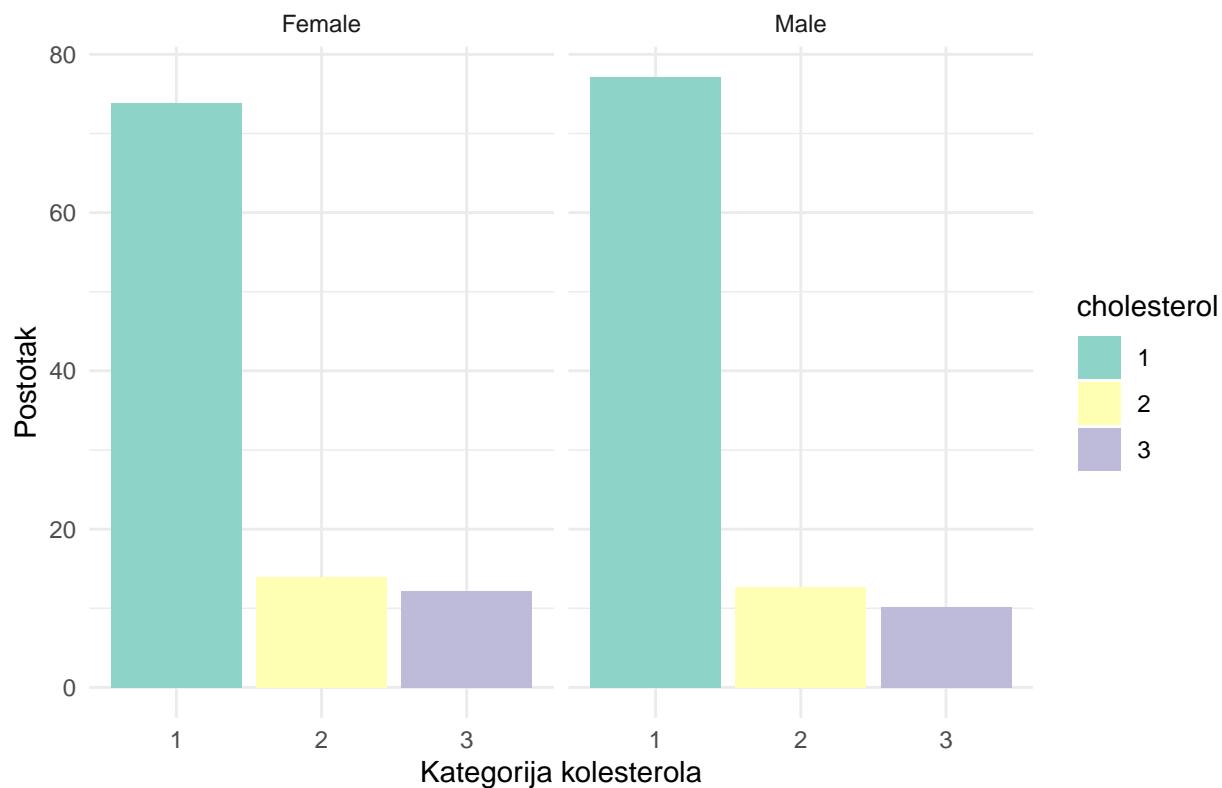
- U svakoj dobroj skupini prevladava zdrava kategorija kolesterola.
- Dobna skupina 20-40 godina ima najzdravije razine kolesterola, s najmanjom zastupljenosću rizične i opasne kategorije.
- Skupina iznad 60 godina i dalje najčešće pripada zdravoj kategoriji, ali ima veću zastupljenost rizičnih i opasnih 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

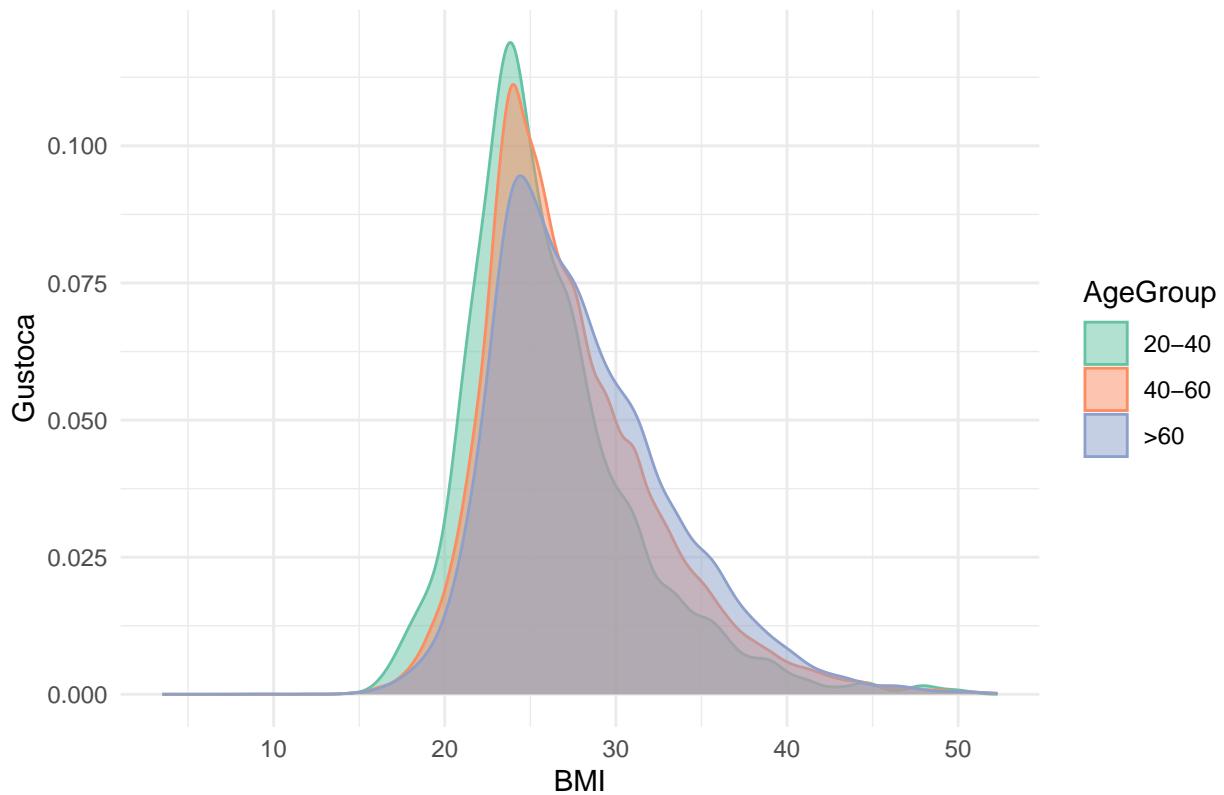
Također nas zanima kako su u uzorku distribuirani indeks tjelesne mase te krvni tlak. Fokusirani grafovi prikazuju češće vrijednosti kako bi se istaknula opća tendencija.

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

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

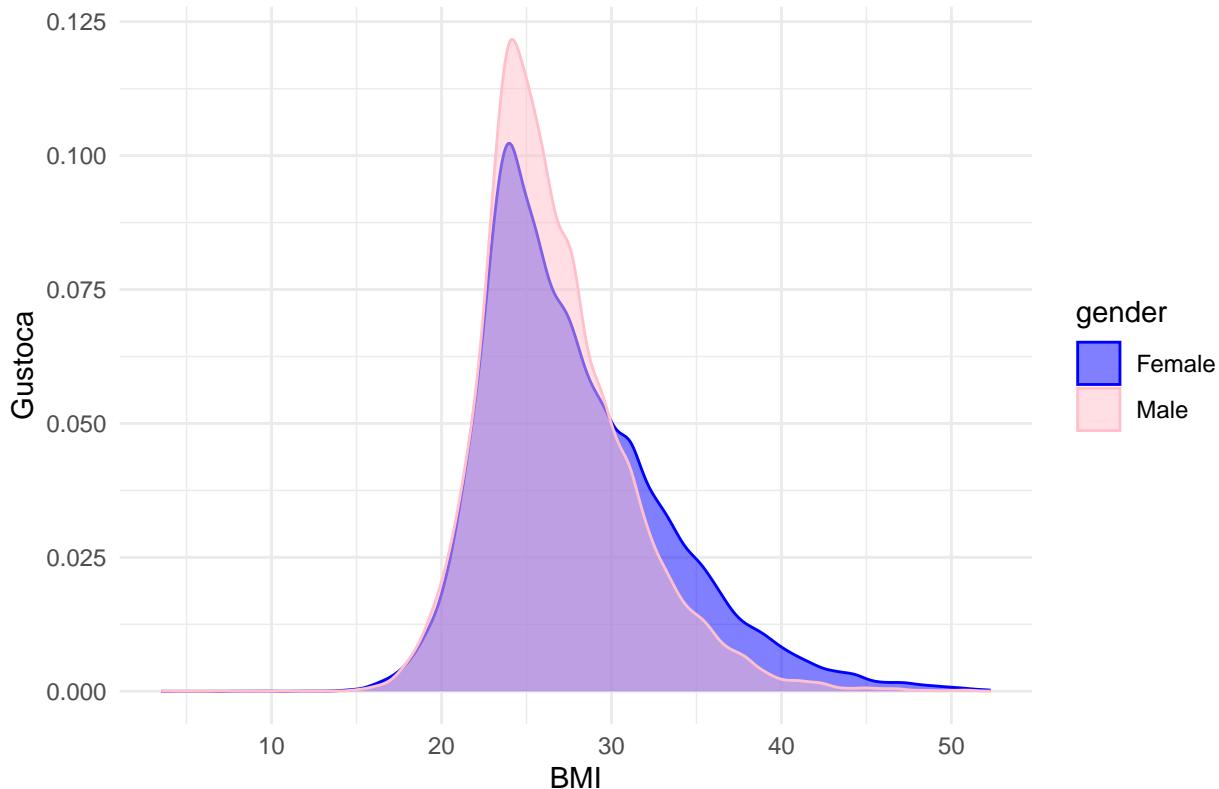
ggplot(BMI_filtered_data, aes(x = BMI, color = AgeGroup, fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = "Istaknuta Distribucija BMI-ja prema doboj skupini",
       x = "BMI",
       y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "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 >= 80) %>% filter(ap_lo >= 50))

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 doboj skupini",
    x = "Sistolički krvni tlak",
    y = "Gustoća") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  scale_color_brewer(palette = "Set2")

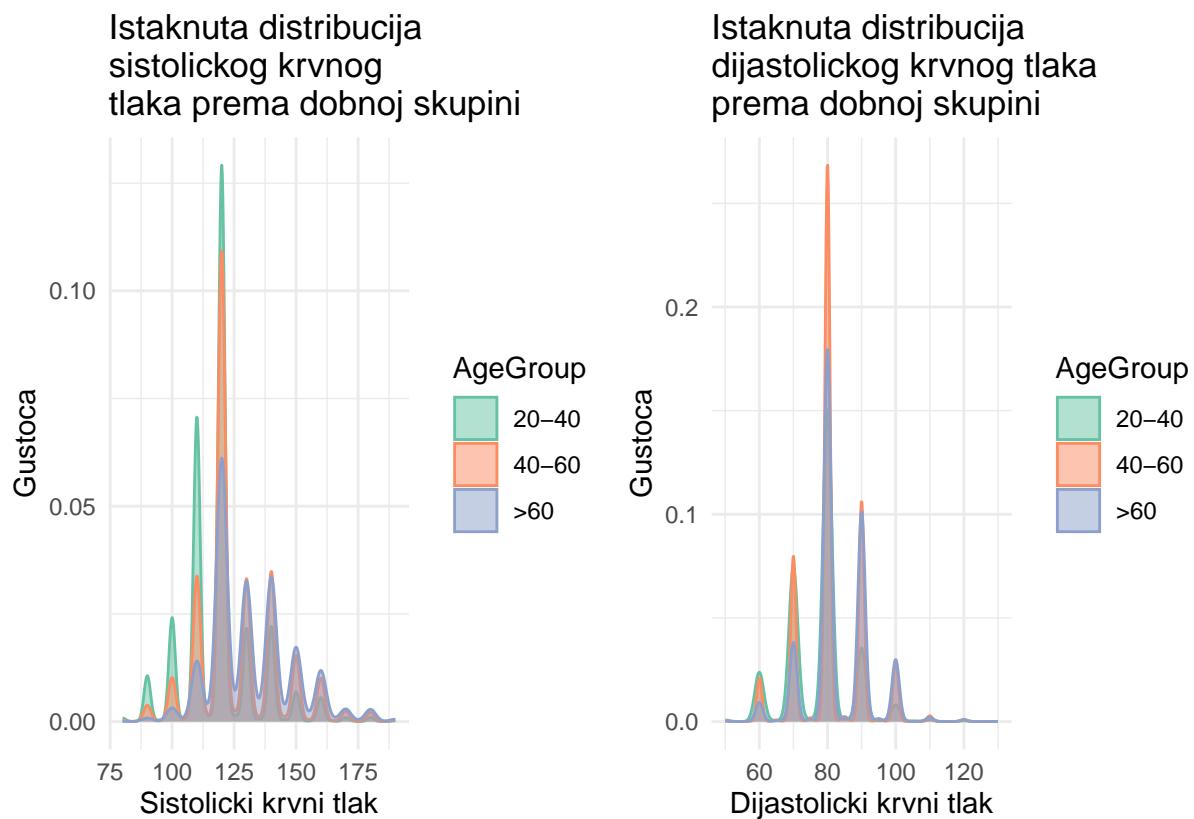
ap_lo_distribution_g1 <- ggplot(ap_filtered_data, aes(x = ap_lo,
  color = AgeGroup,
  fill = AgeGroup)) +
```

```

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

```



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) +

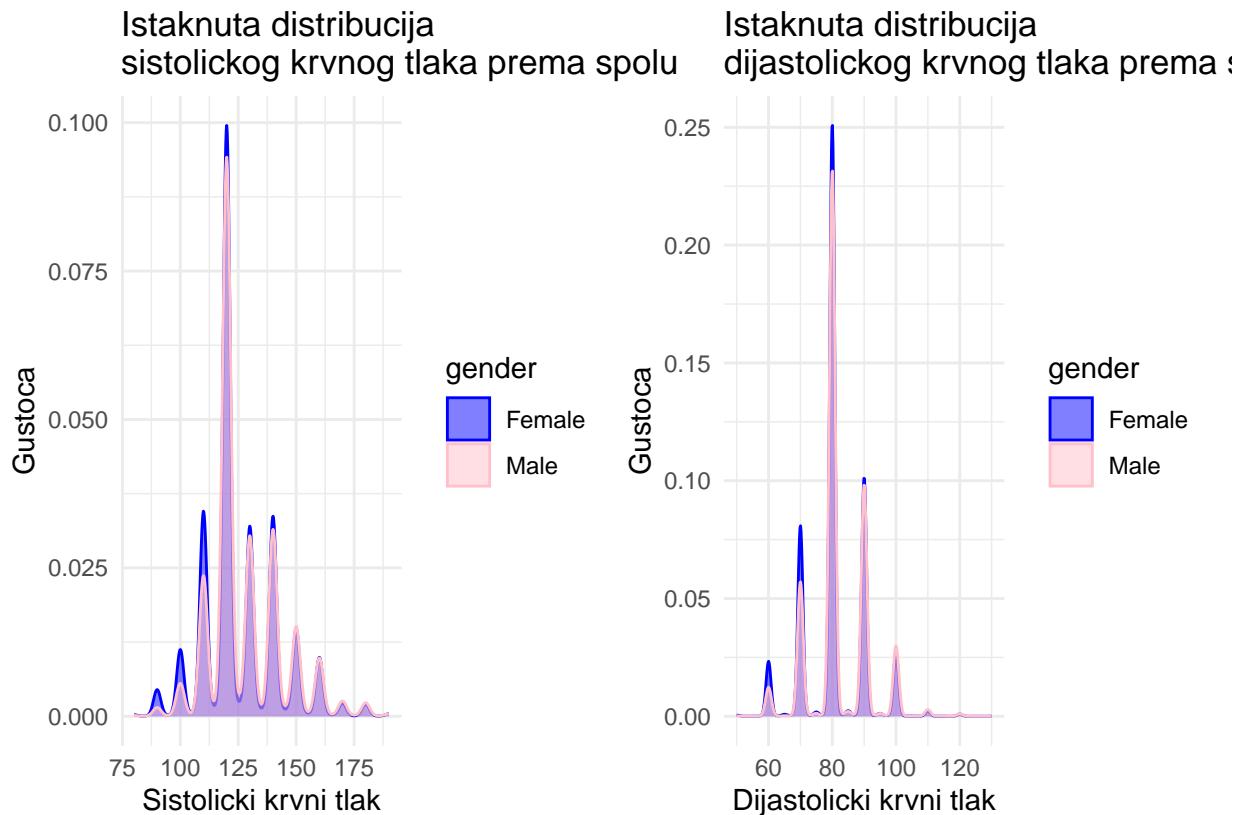
```

```

  labs(title
    = "Istaknuta distribucija\ndijastoličkog krvnog tlaka prema spolu",
    x = "Dijastolički krvni tlak",
    y = "Gustoča") +
  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 pristrandosti, 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 >= 80) %>% filter(ap_lo >= 50))

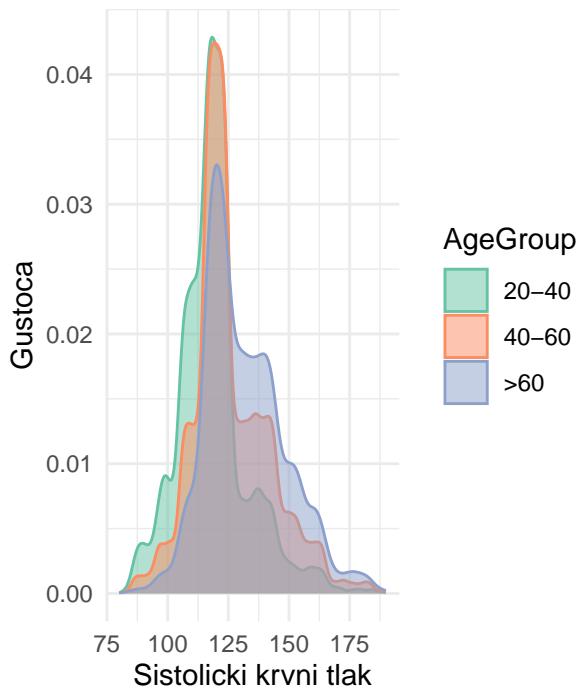
ap_hi_distribution_g3 <- ggplot(ap_filtered_data,
                                 aes(x = ap_hi, color = AgeGroup,
                                     fill = AgeGroup)) +
  geom_density(alpha = 0.5) +
  labs(title = paste("Distribucija sistoličkog\nkrvnog",
                     "tlaka sa dodanim šumom\nprema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

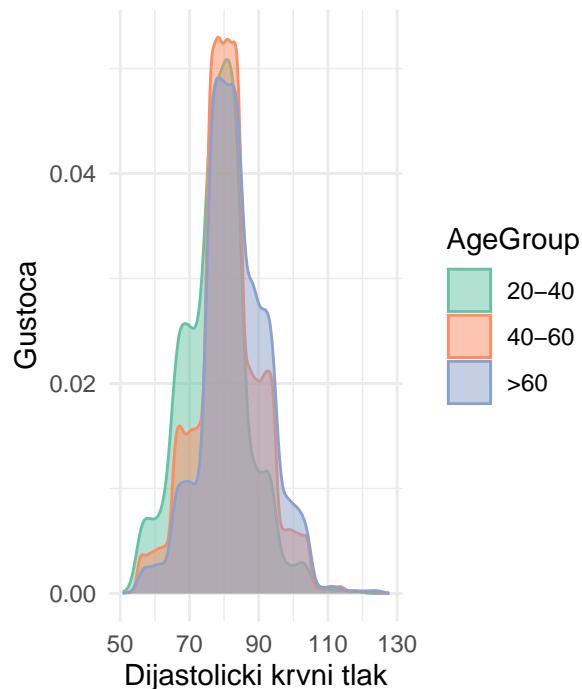
ap_hi_distribution_g3 + ap_lo_distribution_g3

```

Distribucija sistolickog krvnog tlaka sa dodanim šumom prema dobnoj skupini



Distribucija dijastolickog krvnog tlaka sa dodanim šumom prema dobnoj skupini



Zadatak 2

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

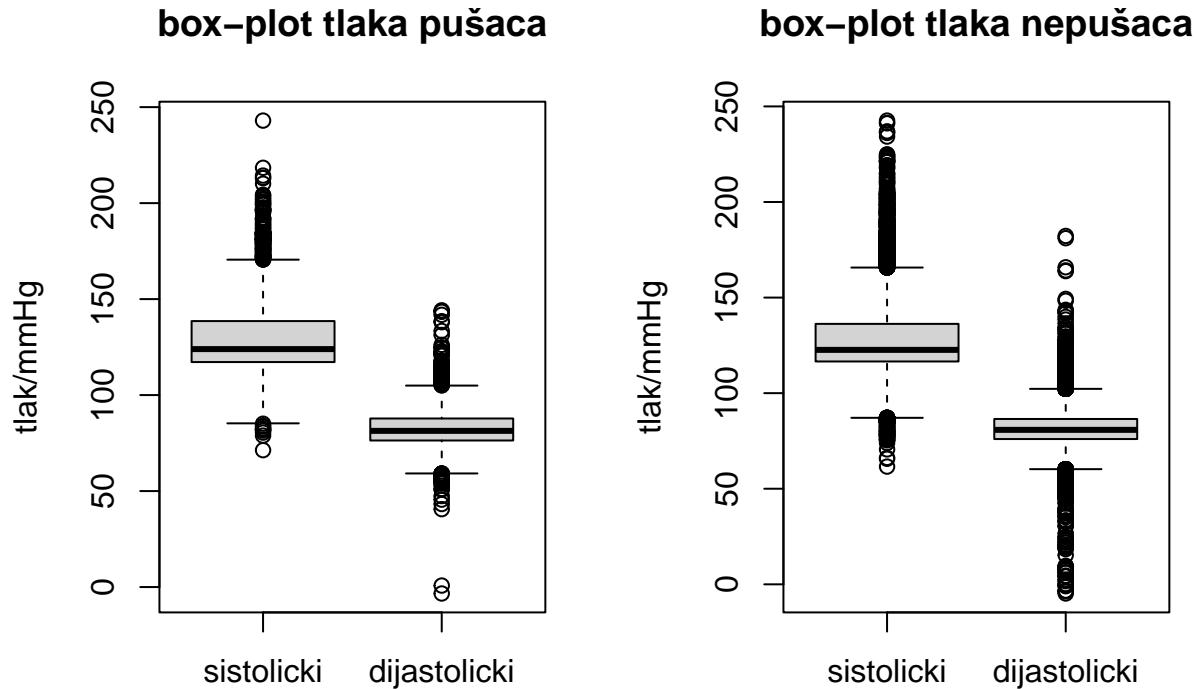
Testiramo postoji li statistički značajna razlika u prosječnim krvnim tlakovima kod pušača i kod nepušača. Pušenje je kategorijalna varijabla (razlikujemo pušače i nepušače, a nema podataka o tome koliko često tko puši).

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 za te dvije grupe. 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)

par(mfrow=c(1,2))
bpd <- data.frame(sistolicki=pusaci$ap_hi, dijastolicki=pusaci$ap_lo)
boxplot(bpd, main='box-plot tlaka pušača',
        ylab='tlak/mmHg')
bpd <- data.frame(sistolicki=nepusaci$ap_hi, dijastolicki=nepusaci$ap_lo)
```

```
boxplot(bpd, main='box-plot tlaka nepušača',
       ylab='tlak/mmHg')
```



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

Provjera preduvjeta testa za jednakost sredina

Najprije provjeravama jesu li podaci iz normalne razdiobe. Kako ne znamo koju konkretnu normalnu razdiobu očekujemo, koristimo Lillieforsovu inačicu Kolmogorov-Smirnovljevog testa uz nivo značajnosti 5%. Posebno testiramo za gornje i za donje tlakove kod pušača i kod nepušača. Također radimo Q-Q plotove za vizualnu usporedbu kvantila razdiobe tlakova s kvantilima normalne razdiobe u sva četiri slučaja.

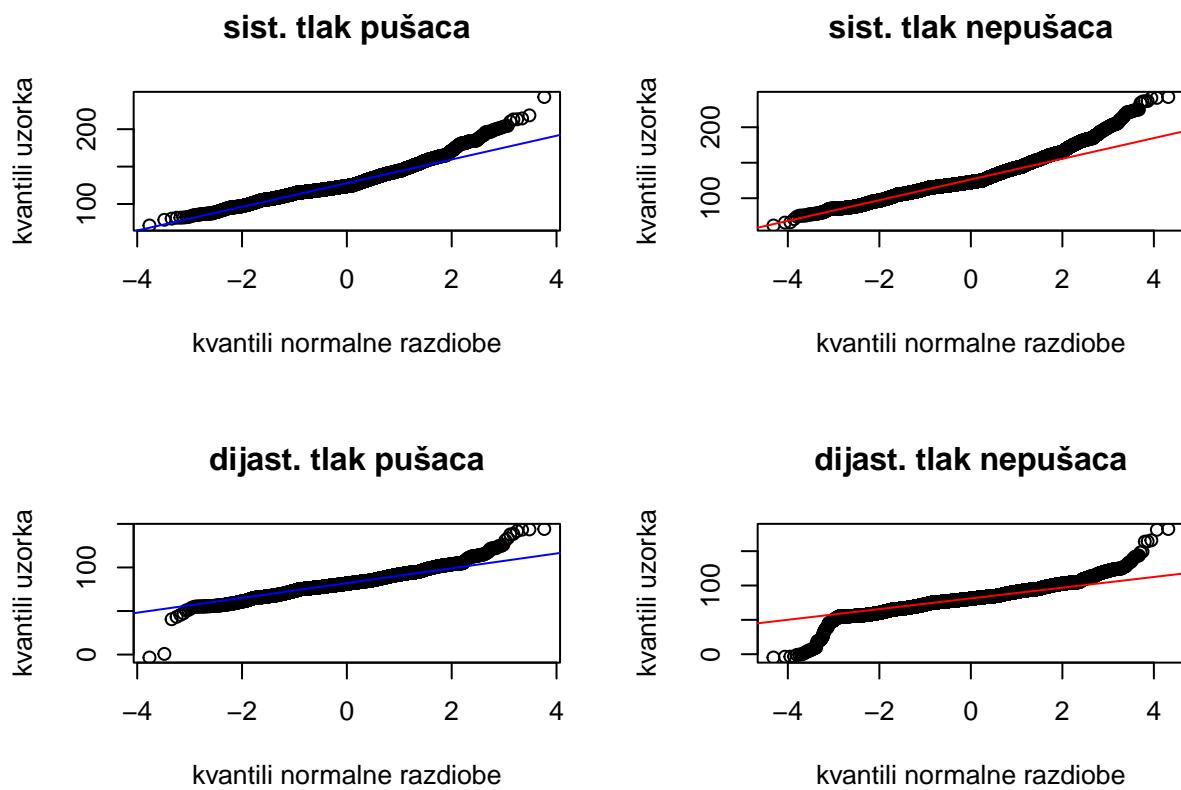
```
pusaci = filtered_data[filtered_data["smoke"]==1,]
nepusaci = filtered_data[filtered_data["smoke"]==0,]

lillie.test(pusaci$ap_hi)
lillie.test(nepusaci$ap_hi)
lillie.test(pusaci$ap_lo)
lillie.test(nepusaci$ap_lo)
par(mfrow = c(2,2))
qqnorm(pusaci$ap_hi, main = "sist. tlak pušača",
       xlab="kvantili normalne razdiobe",
       ylab="kvantili uzorka")
```

```

qqline(pusaci$ap_hi, col="blue")
qnorm(nepusaci$ap_hi, main = "sist. tlak nepušača",
      xlab="kvantili normalne razdiobe",
      ylab="kvantili uzorka")
qqline(nepusaci$ap_hi, col="red")
qnorm(pusaci$ap_lo, main = "dijast. tlak pušača",
      xlab="kvantili normalne razdiobe",
      ylab="kvantili uzorka")
qqline(pusaci$ap_lo, col="blue")
qnorm(nepusaci$ap_lo, main = "dijast. tlak nepušača",
      xlab="kvantili normalne razdiobe",
      ylab="kvantili uzorka")
qqline(nepusaci$ap_lo, col="red")

```



```

##
## 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

```

```

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pusaci$ap_lo
## D = 0.066889, p-value < 2.2e-16
##
##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nepusaci$ap_lo
## D = 0.069755, p-value < 2.2e-16

```

Svi testovi odbacuju nul-hipotezu. Na Q-Q plotovima vidimo da distribucija podataka relativno dobro prati normalnu za vrijednosti oko prosjeka i manje, ali ima vrlo teški gornji rep. To je u skladu s prisutnošću mnogo outliera vidljivih u tom području na box plotu.

Treba provjeriti jesu li varijance grupa jednakih. Provodimo F-test za jednakost varijanci sistoličkih tlakova nepušača i pušača s razinom značajnosti 5%. Dobivamo vrlo malu p-vrijednost, pa odbacujemo nul-hipotezu. Procijenjeni omjer varijanci je oko 0.9 za sistoličke i 0.92 za dijastoličke tlakove, dakle pušači imaju veću varijancu u 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. Neovisno o točnosti te teorije, testovi pokazuju da ne možemo prepostaviti jednakost varijanci u dalnjim testovima.

```

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

```

```

##
## 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
##
##
## F test to compare two variances
##
## data:  nepusaci$ap_lo and pusaci$ap_lo
## F = 0.92314, num df = 62510, denom df = 6033, p-value = 2.177e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8890231 0.9580170
## sample estimates:
## ratio of variances
## 0.9231447

```

Testiramo hipotezu o jednakosti prosječnih krvnih tlakova pušača i nepušača koristeći dvostrani t-test za neuparene podatke uz nepoznate i nejednakne varijance na razini značajnosti od 5%. Test neće biti egzaktan

jer nije zadovljena pretpostavka o normalnosti. Ipak, možemo ga koristiti zbog dovoljne veličine uzorka i robusnosti testa na odstupanje podataka od normalne distribucije. Ne uzimamo pretpostavku da su varijance jednake zbog rezultata F testa.

```
t.test(pusaci$ap_hi, nepusaci$ap_hi, alternative = "two.sided", mu = 0,
       paired = FALSE, var.equal = FALSE, conf.level = 0.95)
t.test(pusaci$ap_lo, nepusaci$ap_lo, alternative = "two.sided", mu = 0,
       paired = FALSE, var.equal = FALSE, conf.level = 0.95)

##
##  Welch Two Sample t-test
##
## data:  pusaci$ap_hi and nepusaci$ap_hi
## 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:
##   1.184821 2.117373
## sample estimates:
## mean of x mean of y
## 128.1524 126.5013
##
##
##  Welch Two Sample t-test
##
## data:  pusaci$ap_lo and nepusaci$ap_lo
## t = 5.8281, df = 7150.6, p-value = 5.849e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.5406252 1.0886254
## sample estimates:
## mean of x mean of y
## 81.99319 81.17856
```

Dobivene P vrijednosti su vrlo male pa možemo odbaciti hipotezu o jednakosti srednjih vrijednosti tlakova na razini značajnosti od 5% i za sistoličke i za dijastoličke. Test pokazuje statistički značajne razlike, no procijenjene razlike sredina su male u odnosu na 30 mmHg koliko otprilike iznosi raspon tlakova koji se smatraju normalnim pa je značaj tih razlika u praksi upitan.

Zadatak 3

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

Za početak, kako bismo dobili dojam o utjecaju tjelesne aktivnosti, crtamo box plotove za sistolički i dijastolički krvni tlak grupirane prema tjelesnoj aktivnosti.

```
# Kreiranje boxplota za ap_hi
p1 <- ggplot(filtered_data,
               aes(x = factor(active,
                               labels = c("Inactive",
                                         "Active")), y = ap_hi)) +
```

```

geom_boxplot(fill = "lightblue") +
theme_minimal(base_size = 12) +
labs(title = "Sistolički tlak prema aktivnosti \n",
x = "Aktivnost",
y = "Sistolički tlak") +
theme(
  plot.title = element_text(size = 14, hjust = 0.5),
  axis.title = element_text(size = 10),
  axis.text = element_text(size = 10)
)

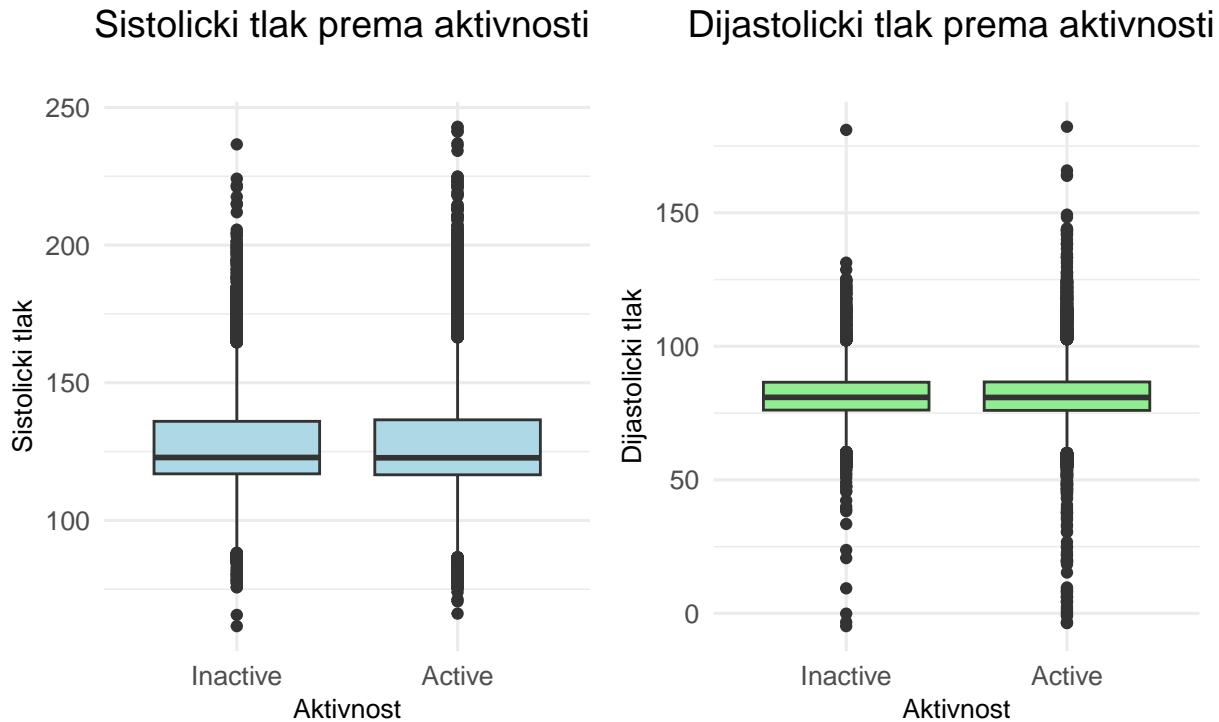
# Kreiranje boxplota za ap_lo
p2 <- ggplot(filtered_data,
  aes(x = factor(active,
    labels = c("Inactive",
    "Active")), y = ap_lo)) +
  geom_boxplot(fill = "lightgreen") +
  theme_minimal(base_size = 12) +
  labs(title = "Dijastolički tlak prema aktivnosti \n",
  x = "Aktivnost",
  y = "Dijastolički tlak") +
  theme(
    plot.title = element_text(size = 14, hjust = 0.5),
    axis.title = element_text(size = 10),
    axis.text = element_text(size = 10)
)

# Grupiranje boxplota horizontalno sa razmakom i višerednim naslovom
combined_boxplots <- p1 + p2 +
  plot_layout(ncol = 2, widths = c(1, 1), guides = "collect") +
  plot_annotation(title = "Boxplotovi krvnog tlaka \nprema aktivnosti") &
  theme(plot.title = element_text(size = 14, hjust = 0.5))

# Prikaz kombiniranih boxplotova
print(combined_boxplots)

```

Boxplotovi krvnog tlaka prema aktivnosti



Iz plotova vidimo da su vrijednosti krvnih tlakova dosta slične kod aktivnih i neaktivnih ispitanika. Ipak, kako se ne bismo zadržali samo na tome, provjerit ćemo jednakost sredina odgovarajućim testovima.

Za određivanje možemo li koristiti t-test, provjeravamo njegove prepostavke. Prvo ćemo ispitati normalnost grupa na temelju histograma i Q-Q plota.

```
# Kreiranje histogramova za ap_hi
hist_hi <- ggplot(filtered_data, aes(x = ap_hi,
                                         fill = factor(active,
                                                       labels = c("Inactive", "Active")))) +
  geom_histogram(binwidth = 2, alpha = 0.6,
                 position = "identity", color = "black") +
  theme_minimal(base_size = 10) +
  labs(title = "Histogram sistoličkog tlaka \nprema aktivnosti",
       x = "Sistolicki tlak",
       fill = "Aktivnost") +
  theme(
    plot.title = element_text(size = 10, hjust = 0.5),
    axis.title = element_text(size = 10),
    axis.text = element_text(size = 10),
    legend.position = "bottom"
  )

# Kreiranje histogramova za ap_lo
hist_lo <- ggplot(filtered_data, aes(x = ap_lo,
                                         fill = factor(active,
                                                       labels = c("Inactive", "Active")))) +
```

```

geom_histogram(binwidth = 2, alpha = 0.6,
               position = "identity", color = "black") +
theme_minimal(base_size = 10) +
labs(title = "Histogram dijastoličkog tlaka \nprema aktivnosti",
     x = "Dijastolički tlak",
     fill = "Aktivnost") +
theme(
  plot.title = element_text(size = 10, hjust = 0.5),
  axis.title = element_text(size = 10),
  axis.text = element_text(size = 10),
  legend.position = "bottom"
)

# Kreiranje Q-Q plotova za ap_hi
qq_hi <- ggplot(filtered_data, aes(sample = ap_hi)) +
  stat_qq(color = "darkblue") +
  stat_qq_line(color = "red") +
  facet_wrap(~ active, labeller =
             as_labeller(c("0" = "Inactive", "1" = "Active"))) +
  theme_minimal(base_size = 10) +
  labs(title = "Q-Q plot sistoličkog tlaka \nprema aktivnosti",
       x = "Teorijske kvantile",
       y = "Empirijske kvantile") +
  theme(
    plot.title = element_text(size = 10, hjust = 0.5),
    strip.text = element_text(size = 10),
    axis.title = element_text(size = 10),
    axis.text = element_text(size = 10)
  )

# Kreiranje Q-Q plotova za ap_lo
qq_lo <- ggplot(filtered_data, aes(sample = ap_lo)) +
  stat_qq(color = "darkgreen") +
  stat_qq_line(color = "red") +
  facet_wrap(~ active, labeller =
             as_labeller(c("0" = "Inactive", "1" = "Active"))) +
  theme_minimal(base_size = 10) +
  labs(title = "Q-Q plot dijastoličkog tlaka \nprema aktivnosti",
       x = "Teorijske kvantile",
       y = "Empirijske kvantile") +
  theme(
    plot.title = element_text(size = 12, hjust = 0.5),
    strip.text = element_text(size = 10),
    axis.title = element_text(size = 10),
    axis.text = element_text(size = 10)
  )

# Grupiranje histogramova u jedan red sa višerednim naslovom
combined_histograms <- hist_hi + hist_lo +
  plot_layout(ncol = 2, widths = c(1, 1)) +
  plot_annotation(title = "Histograme krvnog tlaka \nprema aktivnosti") &
  theme(plot.title = element_text(size = 12, hjust = 0.5))

```

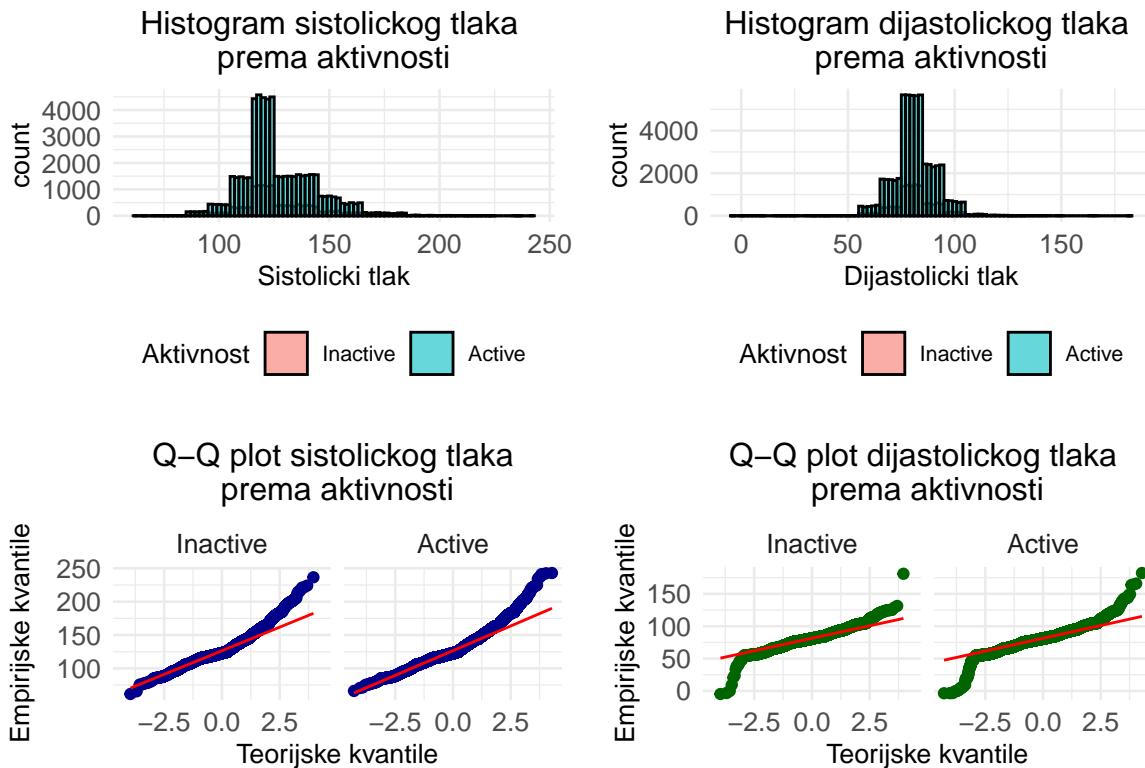
```

# Grupiranje Q-Q plotova u jedan red sa višerednim naslovom
combined_qqplots <- qq_hi + qq_lo +
  plot_layout(ncol = 2, widths = c(1, 1)) +
  plot_annotation(title = "Q-Q Plotovi krvnog tlaka \nprema aktivnosti") &
  theme(plot.title = element_text(size = 12, hjust = 0.5))

# Grupiranje svih histogramova i Q-Q plotova u jedan layout
final_combined <- (combined_histograms / combined_qqplots) +
  plot_layout(ncol = 1, heights = c(1, 1)) &
  theme(
    plot.title = element_text(size = 12, hjust = 0.5),
    plot.margin = margin(10,10,10,10)
  )

# Prikaz kombiniranih grafova
print(final_combined)

```



Iz plotova vidimo da distribucija najvjerojatnije nije normalna, ali za svaki slučaj ćemo za provjeru normalnosti i jednakosti varijanci provesti Kolmogorov-Smirnovljev test i F-test. Koristit ćemo uvijek razinu značajnosti od 5% kao i prije.

```

active_data <- filtered_data %>% filter(active == 1)
inactive_data <- filtered_data %>% filter(active == 0)

# 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 one-sample
## 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: 1.259129e-227
## =====
## Kolmogorov-Smirnov test za ap_hi (Inactive): p-value: 3.781441e-206
## Kolmogorov-Smirnov test za ap_lo (Inactive): p-value: 1.793305e-64
## =====
## ===== 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 vidimo da možemo odbaciti nultu hipotezu o normalnosti distribucija, dok iz F-testa vidimo da ne odbacujemo nultu hipotezu o jednakosti varijanci dviju distribucija. Unatoč nenormalnosti distribucija, provest ćemo t-test za usporedbu sredina budući da je robustan na nenormalnost uzorka. Ipak, provest ćemo i neparametarsku alternativu t-testu, Mann-Whitney-Wilcoxonov test. Za Man_Whitney-Wilcoxonov test prepostavljamo da je krvni tlak distribuiran jednako (do na translaciju) za aktivne i neaktivne ljude.

```

# Podjela podataka prema aktivnosti

# 1) Standardni t-test
cat("\n==== t-test za ap_hi =====\n")
t_test_hi <- t.test(active_data$ap_hi, inactive_data$ap_hi, var.equal = TRUE)
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 = TRUE)
print(t_test_lo)

# 2) Neparametrijski Mann-Whitney-Wilcoxon test
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 =====
##
## Two Sample t-test
##
## data: active_data$ap_hi and inactive_data$ap_hi
## t = -0.27572, df = 68543, p-value = 0.7828
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.3640767 0.2742763
## sample estimates:
## mean of x mean of y
## 126.6379 126.6828
##
##
## ===== t-test za ap_lo =====
##
## Two Sample t-test
##
## data: active_data$ap_lo and inactive_data$ap_lo
## t = -0.71166, df = 68543, p-value = 0.4767
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.2576922 0.1204071
## 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

```

Nakon provedbe testova vidimo da ne možemo odbaciti nultu hipotezu prema kojoj su srednje vrijednosti tlakova aktivne i neaktivne grupe jednake. Stoga, ne možemo tvrditi da se prosječni krvni tlak značajno razlikuje među skupinama s različitom učestalošću tjelesne aktivnosti.

Kao dodatno pitanje, zanima nas i razlikuje li se značajno prosječni krvni tlak među skupinama s različitim BMI kategorijama. Za početak, kako bismo dobili dojam, crtamo box plotove koji nam daju početni uvid u podatke.

```

# Kreiranje boxplota za ap_hi po BMICat
p_hi <- ggplot(filtered_data, aes(x = BMICat, y = ap_hi)) +
  geom_boxplot(fill = "skyblue", alpha = 0.7, outlier.color = "red") +
  geom_hline(yintercept = mean(filtered_data$ap_hi),
             color = "red", linetype = "dashed", size = 1) +
  labs(title = "Sistolički tlak po BMI kategorijama \n",
       x = "BMI kategorija",
       y = "Sistolički tlak") +
  scale_x_discrete(labels = c("Normal"      = "Normal",
                             "Obese"        = "Obese",
                             "Over Weight" = "Over\nWeight",
                             "Under Weight" = "Under\nWeight")) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(size = 14, hjust = 0.5),
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10)
  )

## 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.

# Kreiranje boxplota za ap_lo po BMICat
p_lo <- ggplot(filtered_data, aes(x = BMICat, y = ap_lo)) +
  geom_boxplot(fill = "orange", alpha = 0.7, outlier.color = "blue") +
  geom_hline(yintercept = mean(filtered_data$ap_lo),
             color = "red", linetype = "dashed", size = 1) +
  labs(title = "Dijastolički tlak po BMI kategorijama \n",
       x = "BMI kategorija",
       y = "Dijastolički tlak") +
  scale_x_discrete(labels = c("Normal"      = "Normal",
                             "Obese"        = "Obese",
                             "Over Weight" = "Over\nWeight",
                             "Under Weight" = "Under\nWeight")) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(size = 14, hjust = 0.5),
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10)
  )

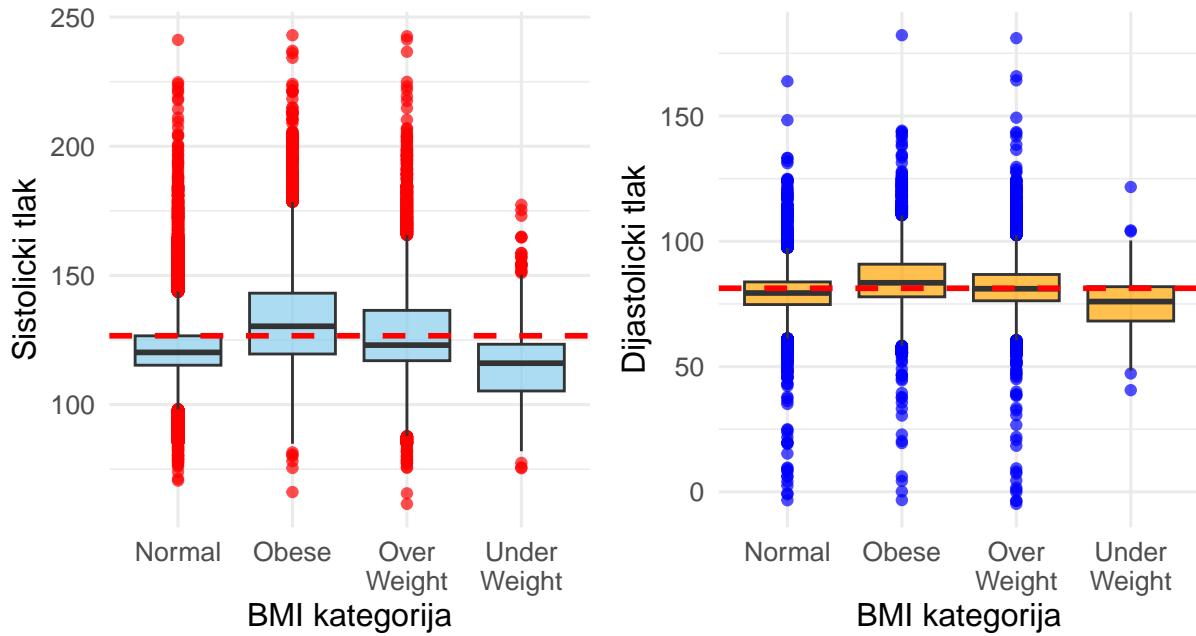
combined_boxplots_bmi <- p_hi + p_lo +
  plot_layout(ncol = 2, widths = c(1, 1), guides = "collect") +
  plot_annotation(title = "Boxplotovi krvnog tlaka po BMI kategorijama \n") &
  theme(plot.title = element_text(size = 14, hjust = 0.5))

print(combined_boxplots_bmi)

```

Boxplotovi krvnog tlaka po BMI kategorijama

Sistolicki tlak po BMI kategorijama Dijastolicki tlak po BMI kategorijam



Prema box plotovima vidimo da je za očekivati kako se prosječni tlakovi značajno razlikuju među različitim skupinama, ali to moramo potvrditi odgovarajućim statističkim testovima. U ovom slučaju testiramo jednakost sredina četiri različite skupine, za što je potrebno koristiti ANOVA-u. Ipak, prije toga moramo provjeriti pretpostavke ANOVE. Prvo ćemo nacrtati histograme i Q-Q plotove kako bismo stekli dojam o normalnosti skupina.

```
# 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)

  hist_hi <- ggplot(data_subset, aes(x = ap_hi)) +
    geom_histogram(binwidth = 2, fill = "blue", color = "black", alpha = 0.7) +
    labs(title = "Distribucija sistoličkog tlaka",
         x = "Sistolicki tlak", y = "Frekvencija") +
    theme_minimal(base_size = 12)

  hist_lo <- ggplot(data_subset, aes(x = ap_lo)) +
    geom_histogram(binwidth = 2, fill = "red", color = "black", alpha = 0.7) +
    labs(title = "Distribucija dijastoličkog tlaka",
         x = "Dijastolički tlak", y = "Frekvencija") +
    theme_minimal(base_size = 12)

  qq_hi <- ggplot(data_subset, aes(sample = ap_hi)) +
    stat_qq(color = "darkblue") +
```

```

stat_qq_line(color = "red") +
  labs(title = "Q-Q plot - Sistolički tlak",
       x = "Teorijske kvantile",
       y = "Empirijske kvantile") +
  theme_minimal(base_size = 12)

qq_lo <- ggplot(data_subset, aes(sample = ap_lo)) +
  stat_qq(color = "darkgreen") +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q plot - Dijastolički tlak",
       x = "Teorijske kvantile",
       y = "Empirijske kvantile") +
  theme_minimal(base_size = 12)

hist_combined <- hist_hi + hist_lo +
  plot_layout(ncol = 2) +
  plot_annotation(title = paste("Histogramme krvnog tlaka - BMI Kategorija:",
                                bmi_cat)) &
  theme(plot.title = element_text(hjust = 0.5))

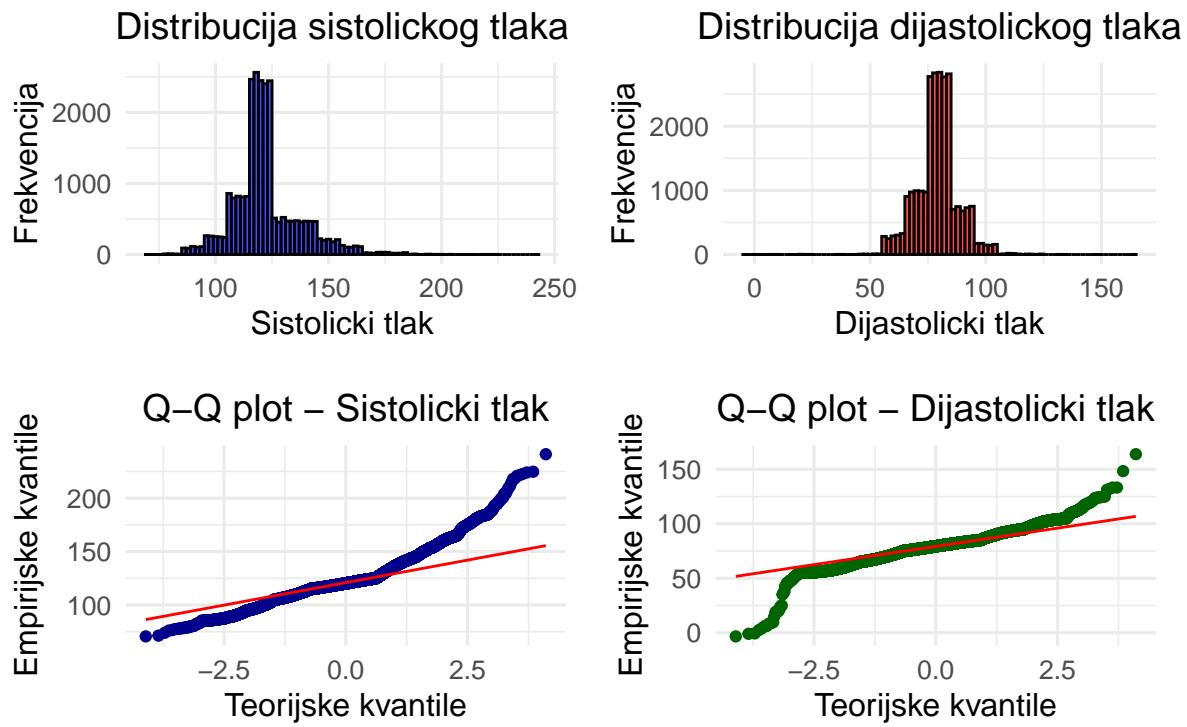
qq_combined <- qq_hi + qq_lo +
  plot_layout(ncol = 2) +
  plot_annotation(title = paste("Q-Q Plotovi krvnog tlaka - BMI Kategorija:",
                                bmi_cat)) &
  theme(plot.title = element_text(hjust = 0.5))

combined_plots <- (hist_combined / qq_combined) +
  plot_layout(ncol = 1) +
  plot_annotation(title = paste("Analiza krvnog tlaka - BMI Kategorija:",
                                bmi_cat)) &
  theme(plot.title = element_text(hjust = 0.5),
        plot.margin = margin(10, 10, 10, 10))

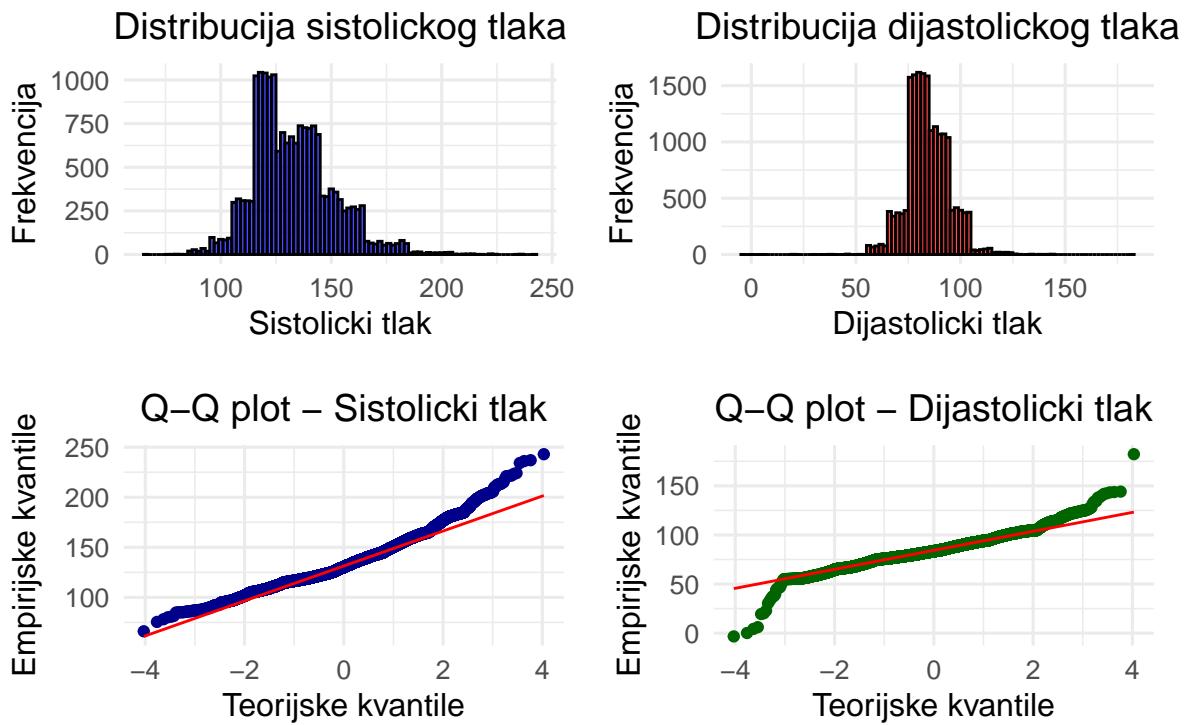
print(combined_plots)
cat("\\\\newpage")
}

```

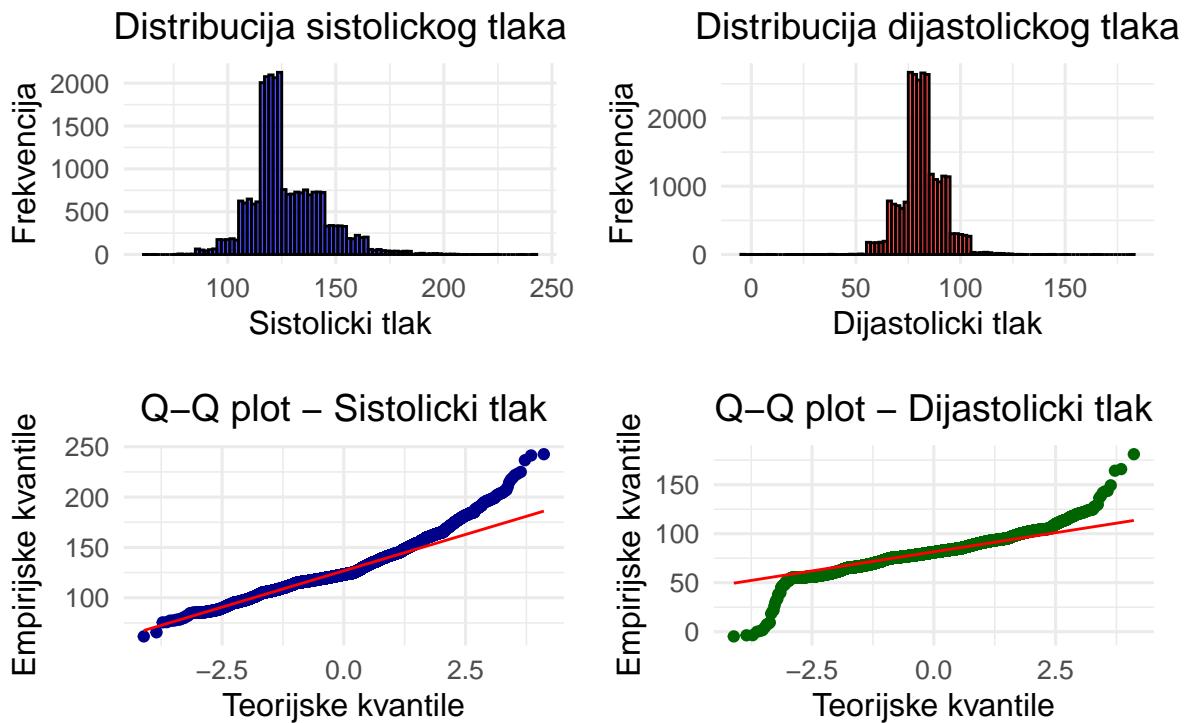
Analiza krvnog tlaka – BMI Kategorija: Normal



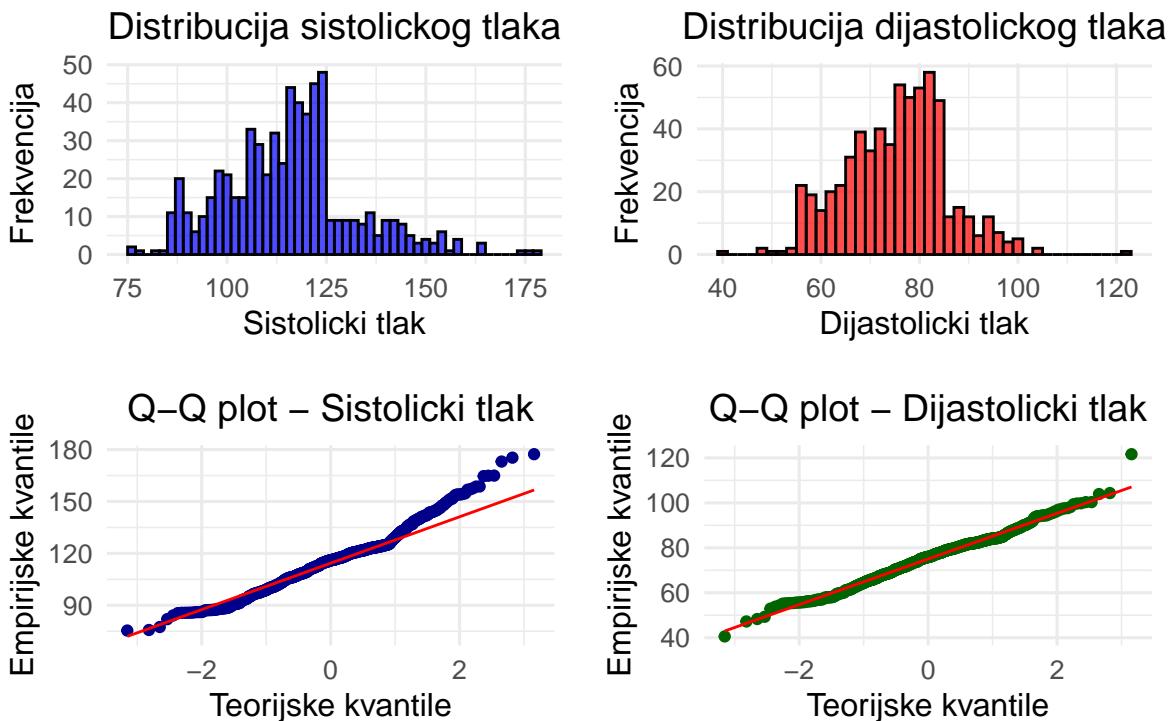
Analiza krvnog tlaka – BMI Kategorija: Obese



Analiza krvnog tlaka – BMI Kategorija: Over Weight



Analiza krvnog tlaka – BMI Kategorija: Under Weight



```
## \newpage\newpage\newpage\newpage
```

Nakon crtanja histograma i Q–Q plotova dobivamo dojam da skupine ne ispunjavaju uvjet normalnosti, ali svejedno ćemo provesti Kolmogorov-Smirnovljev test kako bismo tu tvrdnju dodatno provjerili. Također, provest ćemo i Bartlettov test za provjeru homoskedastičnosti, koja je još bitnija prepostavka ANOVE.

```
# 1) Test normalnosti po grupama (Kolmogorov-Smirnov)

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 homoskedastičnosti (BartlettTest)\\n")
cat("=====\\n")

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

# Bartlettov test za dijastolički tlak
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: 3.36491e-138
##
## =====
## BMI Category: Obese
## Broj zapisa u ovoj kategoriji: 17948
##
## >> Kolmogorov-Smirnov test - ap_hi <<
##     p-value: 1.120204e-72
##
## >> Kolmogorov-Smirnov test - ap_lo <<
##     p-value: 2.221413e-56
##
## =====
## 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: 5.618071e-113
##
## =====
## 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 homoskedastičnosti (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

```

Iščitavanjem p-vrijednosti iz testova možemo odbaciti pretpostavke o normalnosti i homoskedastičnosti skupina. Stoga moramo odustati od provedbe ANOVE te se okrećemo njezinoj neparametarskoj alternativi, Kruskal-Wallisovu testu.

```

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

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

## ----- Kruskal-Wallis Test za sistolički tlak -----
##
## 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 dijastolički tlak -----  

##  

## Kruskal-Wallis rank sum test  

##  

## data: ap_hi by BMICat  

## Kruskal-Wallis chi-squared = 3071.9, df = 3, p-value < 2.2e-16

```

Nakon provedbe Kruskal-Wallisova testa vidimo da niske p-vrijednosti sugeriraju odbacivanje početne hipoteze o jednakosti sredina te prihvaćamo alternativu da se stvarne sredine razlikuju među različitim BMI skupinama.

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.

```

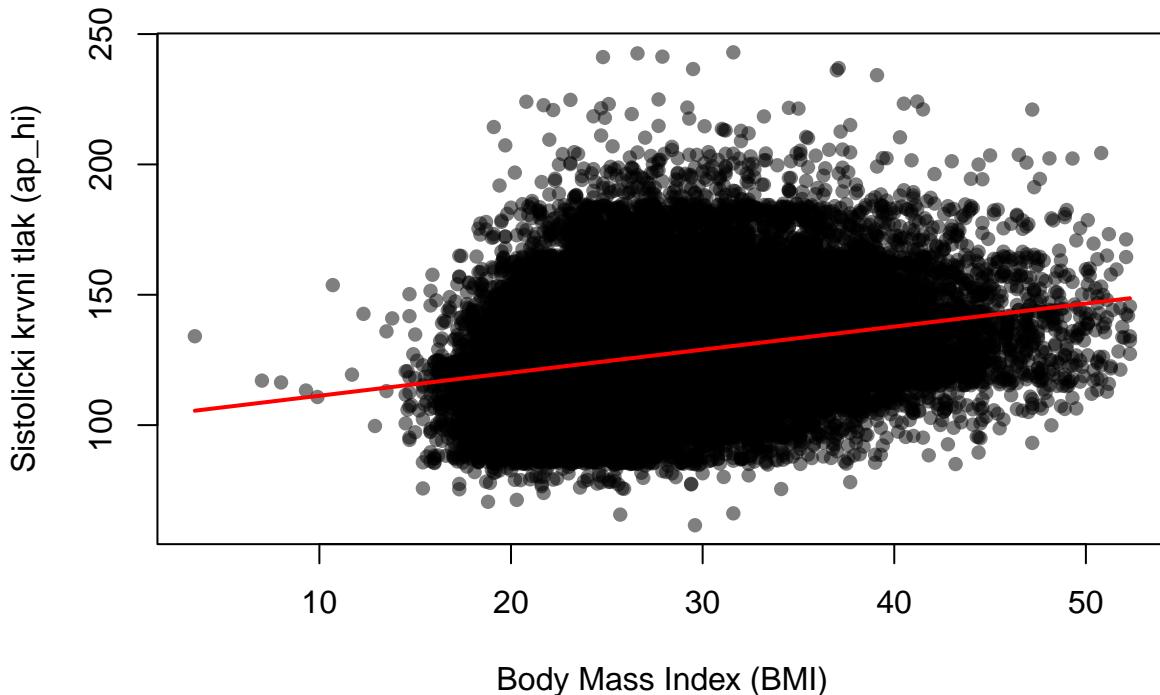
fit.ap_hi <- lm(ap_hi ~ poly(BMI, 1) , data = filtered_data)

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. Također, R^2 vrijednost je 0.07064 (R^2_{adj} je 0.07062) 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:

```
plot1 <- ggplot(data.frame(x = filtered_data$BMI,
                             y = fit.ap_hi$residuals), aes(x = x, y = y)) +
  geom_point() +
  labs(title = "Graf reziduala (ap_hi)", x = "BMI", y = "Reziduali") +
  theme_minimal()

plot2 <- ggplot(data.frame(x = fit.ap_hi$residuals), aes(x = x)) +
  geom_histogram(bins = 20, fill = "blue", color = "black") +
  labs(title = "Histogram Reziduala (ap_hi)",
       x = "Reziduali", y = "Frekvencija") +
  theme_minimal()

plot3 <- ggplot(data.frame(x = rstandard(fit.ap_hi)), aes(x = x)) +
  geom_histogram(bins = 20, fill = "green", color = "black") +
  labs(title = "Standardizirani reziduali (ap_hi)",
       x = "Standardizirani reziduali", y = "Frekvencija") +
  theme_minimal()

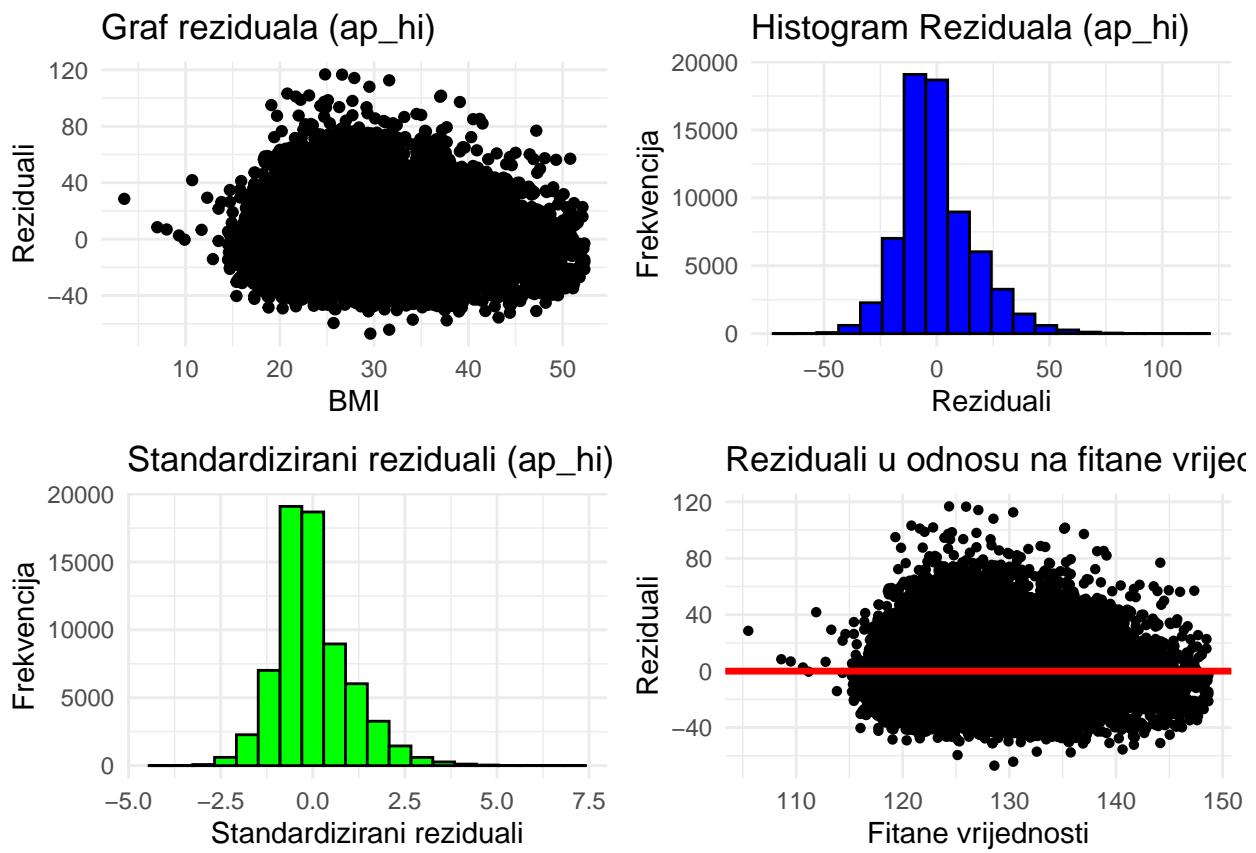
plot4 <- ggplot(data.frame(sample = rstandard(fit.ap_hi)),
                 aes(sample = sample)) +
  stat_qq() +
  stat_qq_line(color = "red", lwd = 1.2) +
  labs(title = "Q-Q plot standardiziranih reziduala (ap_hi)",
       x = "Teorijske kvantile", y = "Standardizirani reziduali") +
  theme_minimal()
```

```

plot5 <- ggplot(data.frame(x = fit.ap_hi$fitted.values,
                            y = fit.ap_hi$residuals), aes(x = x, y = y)) +
  geom_point(pch = 16) +
  geom_hline(yintercept = 0, color = "red", linetype = "solid", lwd = 1.2) +
  labs(title = "Reziduali u odnosu na fitane vrijednosti (ap_hi)",
       x = "Fitane vrijednosti", y = "Reziduali") +
  theme_minimal()

grid.arrange(plot1, plot2, plot3, plot5,
             heights = c(1,1), widths = c(1,1))

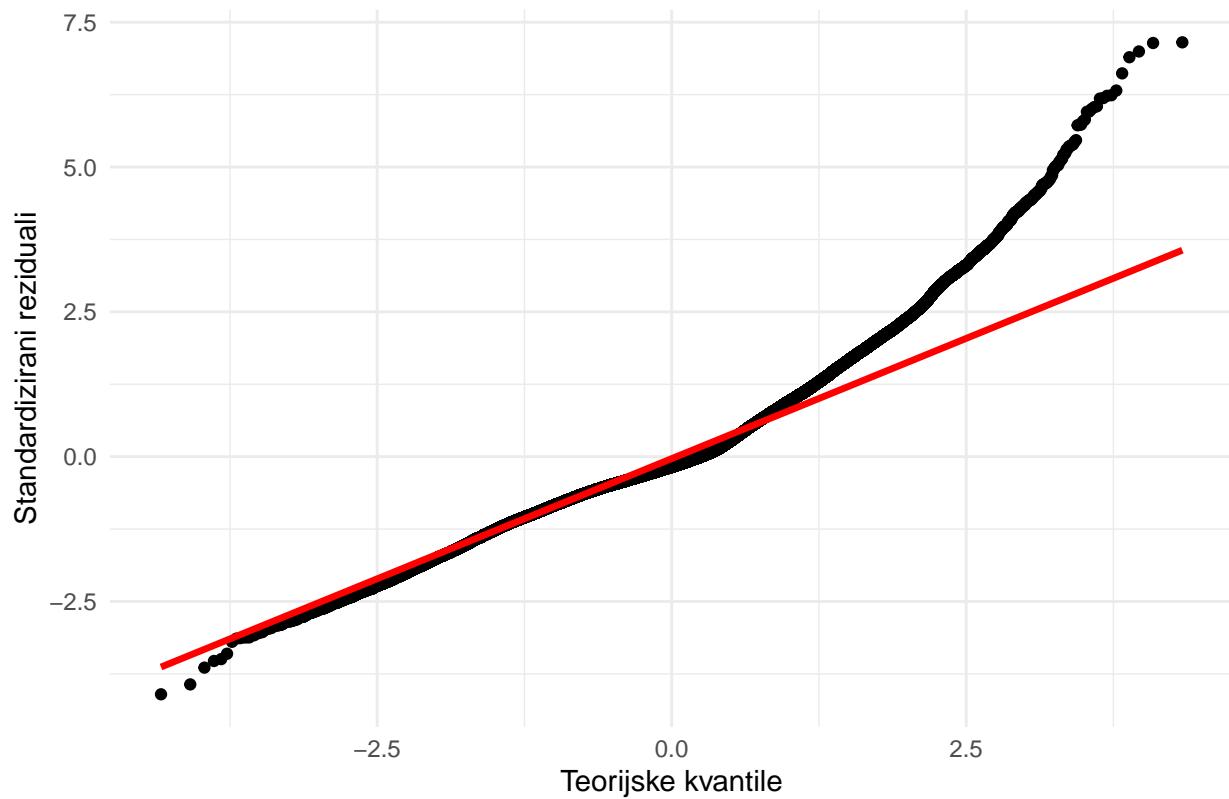
```



```
# (plot1 + plot2) / (plot3 + plot5)
```

```
plot4
```

Q–Q plot standardiziranih reziduala (ap_hi)



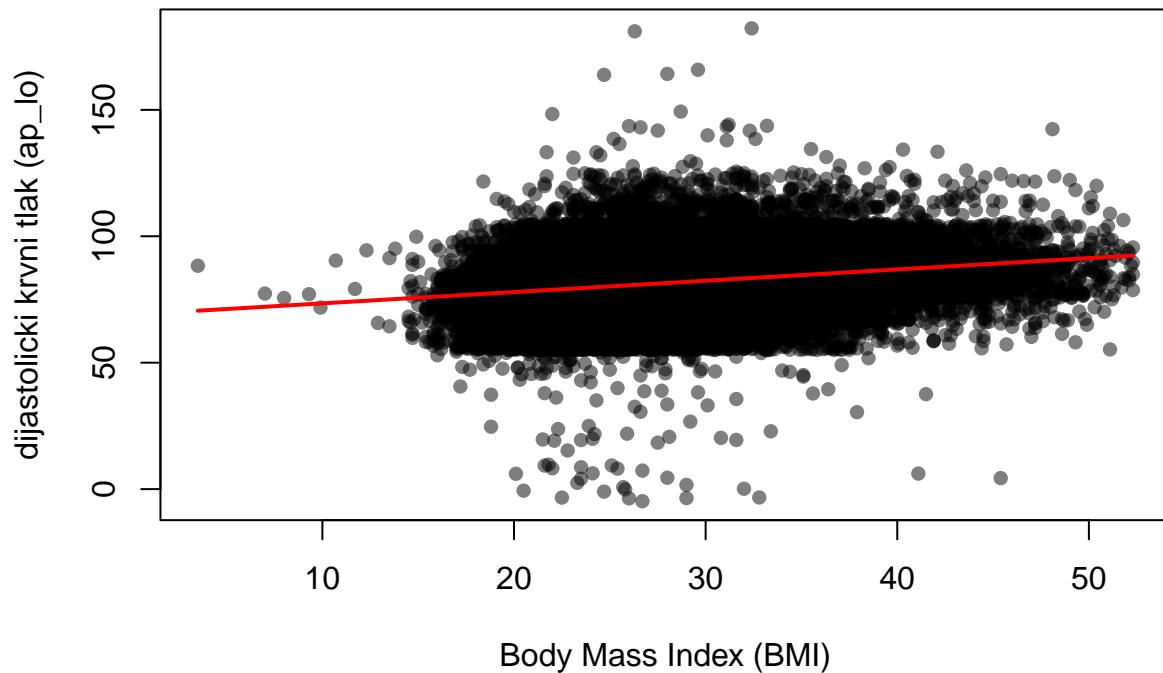
Q–Q plot nam govori da ova razdioba ima teže repove od normalne, ali ovo svakako nije normalna distribucija. Sada možemo zaključiti da je nemoguće predvidjeti sistolički krvni tlak iz BMI-a (iz ovih podataka).

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 0.05161 (R^2_{adj} je 0.0516) 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:

```

plot6 <- ggplot(data.frame(x = filtered_data$BMI, y = fit.ap_lo$residuals),
                 aes(x = x, y = y)) +
  geom_point() +
  labs(title = "Graf reziduala (ap_lo)", x = "BMI", y = "Reziduali") +
  theme_minimal()

plot7 <- ggplot(data.frame(x = fit.ap_lo$residuals), aes(x = x)) +
  geom_histogram(bins = 20, fill = "blue", color = "black") +
  labs(title = "Histogram Reziduala (ap_lo)", x = "Reziduali",
       y = "Frekvencija") +
  theme_minimal()

plot8 <- ggplot(data.frame(x = rstandard(fit.ap_lo)), aes(x = x)) +
  geom_histogram(bins = 20, fill = "green", color = "black") +
  labs(title = "Standardizirani reziduali (ap_lo)",
       x = "Standardizirani reziduali", y = "Frekvencija") +
  theme_minimal()

plot9 <- ggplot(data.frame(sample = rstandard(fit.ap_lo)),
                 aes(sample = sample)) +
  stat_qq() +
  stat_qq_line(color = "red", lwd = 1.2) +
  labs(title = "Q-Q plot standardiziranih reziduala (ap_lo)",
       x = "Teorijske kvantile", y = "Standardizirani reziduali") +
  theme_minimal()

plot10 <- ggplot(data.frame(x = fit.ap_lo$fitted.values,
                            y = fit.ap_lo$residuals), aes(x = x, y = y)) +
  geom_point(pch = 16) +

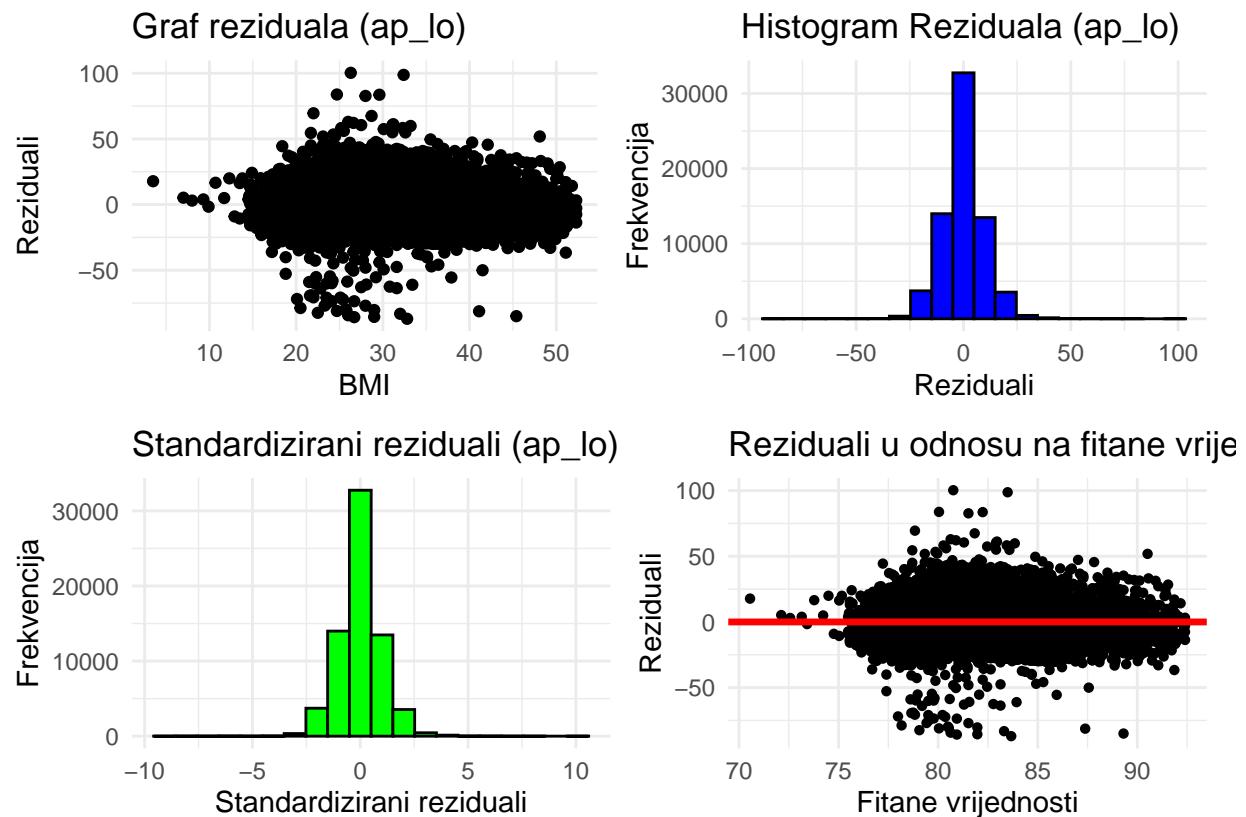
```

```

geom_hline(yintercept = 0, color = "red", linetype = "solid", lwd = 1.2) +
labs(title = "Reziduali u odnosu na fitane vrijednosti (ap_lo)",
x = "Fitane vrijednosti", y = "Reziduali") +
theme_minimal()

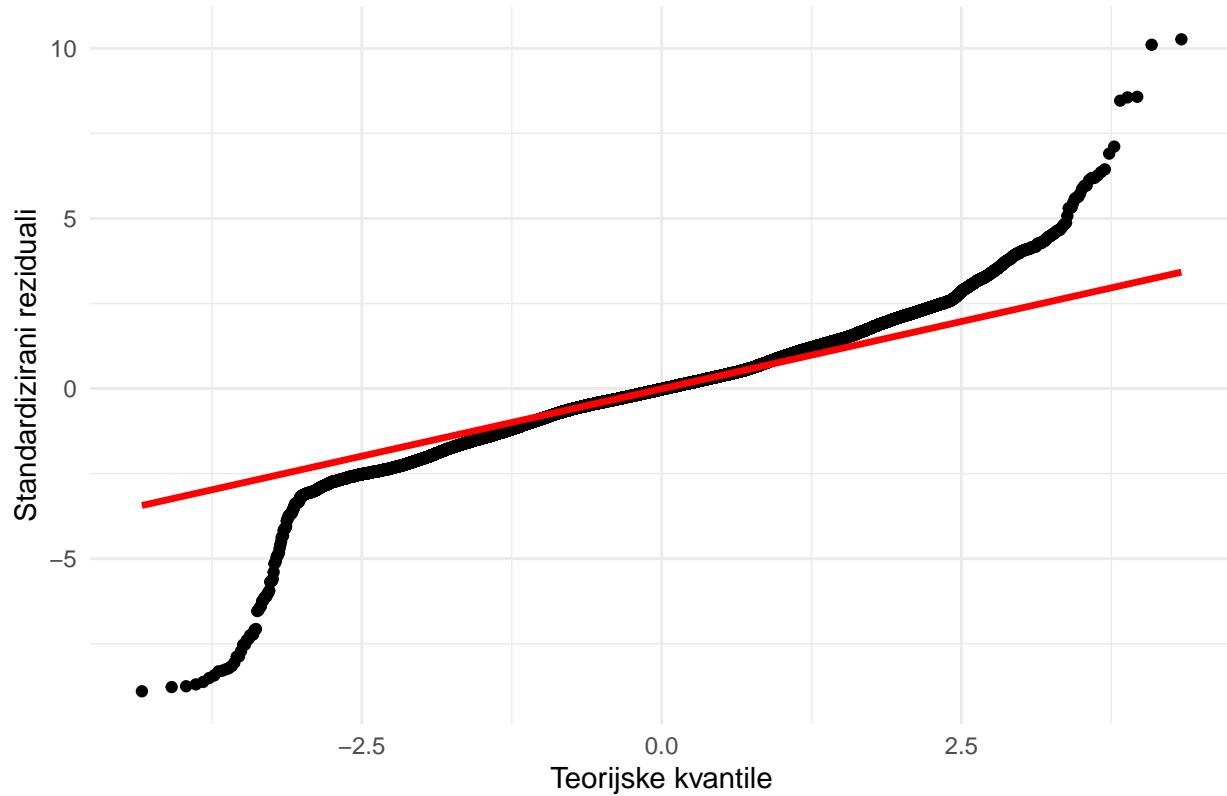
(plot6 + plot7) / (plot8 + plot10)

```



plot9

Q-Q plot standardiziranih reziduala (ap_lo)



Grafički možemo reći da reziduali imaju teže repove. 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.

Višestruka regresija za sistolički i dijastolički krvni tlak

```
#ako maknete regresore koji su manje značajni R^2 pada
fit.multi_hi <- lm(ap_hi ~ BMI + active + AgeinYr, filtered_data)
#fit.multi = lm(ap_hi ~ AgeinYr + active, filtered_data)
```

```

summary(fit.multi_hi)

#ako maknete regresore koji su manje značajni R^2 pada
fit.multi_lo <- lm(ap_lo ~ BMI + active + AgeinYr, filtered_data)
#fit.multi = lm(ap_lo ~ AgeinYr + active, filtered_data)
summary(fit.multi_lo)

## 
## 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
## 
## Call:
## lm(formula = ap_lo ~ BMI + active + AgeinYr, data = filtered_data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -86.632  -5.402  -0.210   5.224  98.615 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 59.86135  0.34652 172.748   <2e-16 ***  
## BMI         0.42170  0.00731  57.689   <2e-16 ***  
## active      0.04342  0.09318   0.466     0.641   
## AgeinYr     0.18551  0.00550  33.726   <2e-16 ***  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9.693 on 68541 degrees of freedom
## Multiple R-squared:  0.06709, Adjusted R-squared:  0.06705 
## F-statistic:  1643 on 3 and 68541 DF,  p-value: < 2.2e-16

```

Vidimo da je jedini značajan regresor mjeru tjelesne aktivnosti, ali smo odlučili zadržati ostale regresore zbog njihove interakcije u višestrukoj regresiji. Naime, za ovu kombinaciju regresora dobili smo najbolju vrijednost od R_{adj} . R_{adj} kod dijastoličkog tlaka je manji nego kod sistoličkog tlaka i smatramo kako bi regresori za sistolički i dijastolički tlak trebali biti isti jer je riječ o usko vezanim fizikalnim veličinama.

Nastavimo s analizom reziduala. Prvo testiramo normalnost:

```
#KS test na normalnost
print("Testovi za sistolički tlak")

ks.test(rstandard(fit.ap_hi), 'pnorm')

lillie.test(rstandard(fit.ap_hi))

print("Testovi za dijastolički tlak")

ks.test(rstandard(fit.ap_lo), 'pnorm')

lillie.test(rstandard(fit.ap_lo))

## [1] "Testovi za sistolički tlak"
##
##  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
##
## [1] "Testovi za dijastolički tlak"
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: rstandard(fit.ap_lo)
## D = 0.049727, p-value < 2.2e-16
## alternative hypothesis: two-sided
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fit.ap_lo)
## D = 0.049728, p-value < 2.2e-16
```

Iz KS i Lillieforsovog testa dobivamo da reziduali nisu normalno distribuirani. Pogledajmo kako se oni ponašaju grafički.

```
selected.model = fit.multi_hi
selected.model_ap_lo = fit.multi_lo

plot1 <- ggplot(data.frame(x = selected.model$fitted.values,
                           y = selected.model$residuals), aes(x, y)) +
  geom_point() +
  labs(title = "ap_hi residuals", x = "fitted.values", y = "residuals")
```

```

plot2 <- ggplot(data.frame(x = filtered_data$BMI,
                           y = selected.model$residuals), aes(x, y)) +
  geom_point() +
  labs(title = "ap_hi residuals", x = "BMI", y = "residuals")

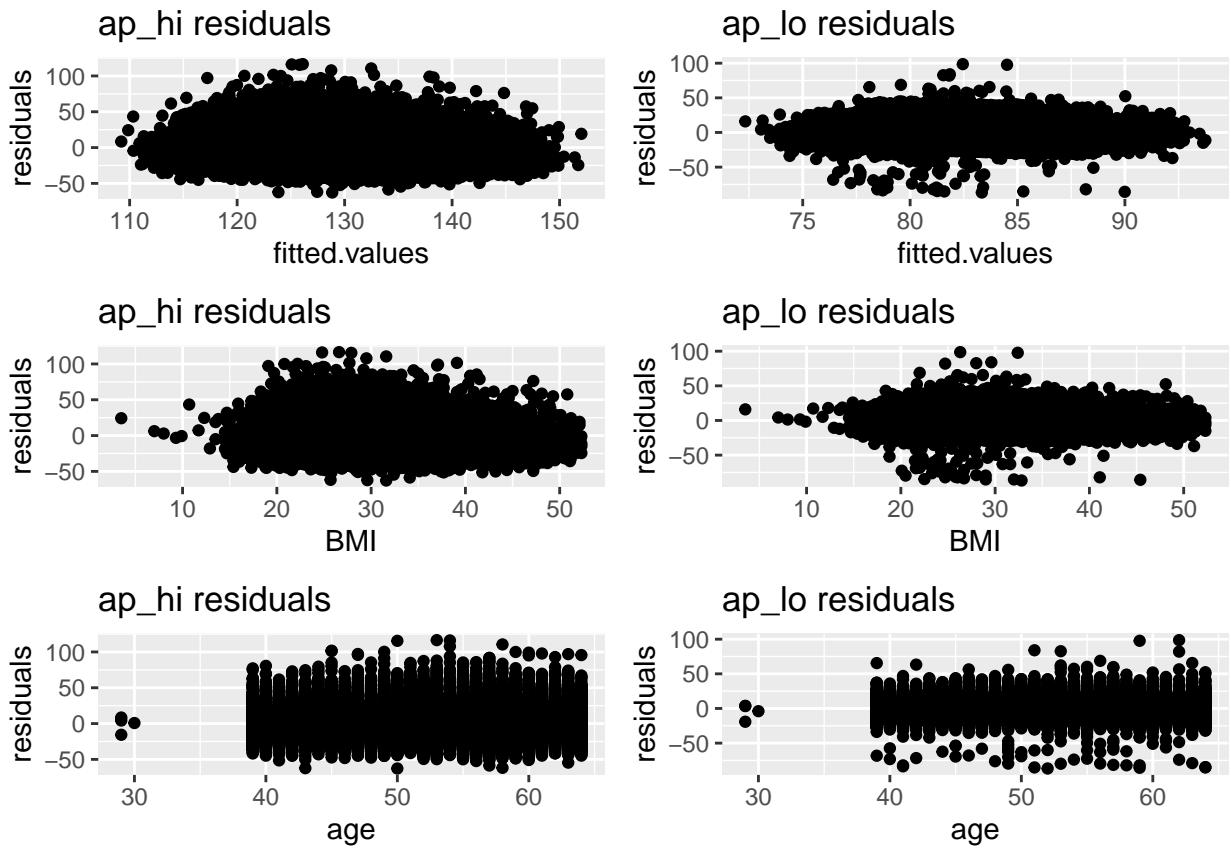
plot3 <- ggplot(data.frame(x = filtered_data$AgeinYr,
                           y = selected.model$residuals), aes(x, y)) +
  geom_point() +
  labs(title = "ap_hi residuals", x = "age", y = "residuals")

plot4 <- ggplot(data.frame(x = selected.model_ap_lo$fitted.values,
                           y = selected.model_ap_lo$residuals), aes(x, y)) +
  geom_point() +
  labs(title = "ap_lo residuals", x = "fitted.values", y = "residuals")

plot5 <- ggplot(data.frame(x = filtered_data$BMI,
                           y = selected.model_ap_lo$residuals), aes(x, y)) +
  geom_point() +
  labs(title = "ap_lo residuals", x = "BMI", y = "residuals")

plot6 <- ggplot(data.frame(x = filtered_data$AgeinYr,
                           y = selected.model_ap_lo$residuals), aes(x, y)) +
  geom_point() +
  labs(title = "ap_lo residuals", x = "age", y = "residuals")
grid.arrange(plot1, plot4, plot2, plot5, plot3, plot6,
             heights = c(1,1,1), widths = c(1,1))

```



Zaključak je da reziduali nisu distribuirani normalno.

Zaključak Iz konteksta medicine smatramo da je dobivena R^2_{adj} premalena kako bismo mogli predviđati sistolički i dijastolički 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 izmjerениh 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.