

Statistička analiza podataka - Projekt Analiza čimbenika rizika za srčane bolesti

Saprofiti

2025-01-24

```
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(magrittr))
suppressPackageStartupMessages(library(tidyr))
suppressPackageStartupMessages(library(readr))
suppressPackageStartupMessages(library(tidyverse))
suppressPackageStartupMessages(library(nortest))
```

Učitavanje podataka

```
data <- read_csv("heart.csv", show_col_types = FALSE)
```

DESKRIPTIVNA STATISTIKA i VIZUALIZACIJA

Prikaz prvih 6 redova

```
print(head(data), width = Inf)
```

```
## # A tibble: 6 x 12
##   Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
##   <dbl> <chr> <chr>          <dbl>      <dbl>      <dbl> <chr>      <dbl>
## 1  40 M   ATA             140        289        0 Normal     172
## 2  49 F   NAP             160        180        0 Normal     156
## 3  37 M   ATA             130        283        0 ST        98
## 4  48 F   ASY             138        214        0 Normal    108
## 5  54 M   NAP             150        195        0 Normal    122
## 6  39 M   NAP             120        339        0 Normal    170
##   ExerciseAngina Oldpeak ST_Slope HeartDisease
##   <chr>          <dbl> <chr>          <dbl>
## 1 N             0 Up             0
## 2 N             1 Flat           1
## 3 N             0 Up             0
## 4 Y            1.5 Flat           1
## 5 N             0 Up             0
## 6 N             0 Up             0
```

Opis podataka

```
dim(data) # broj redaka, broj stupaca (broj primjera, broj varijabli)
```

```
## [1] 918 12
```

```
cat("Broj podataka: ", nrow(data), "\n")
```

```
## Broj podataka: 918
```

```
cat("Broj parametara: ", ncol(data), "\n")
```

```
## Broj parametara: 12
```

```
cat("Imena parametara:", str_wrap(paste(colnames(data), collapse = ", "), width = 70) , "\n")
```

```
## Imena parametara: Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS,  
## RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope, HeartDisease
```

Skup se sastoji od podataka od 918 osoba. Svaki podatak sadrži informacije o 12 razmatраниh parametara (varijabli).

Opis značajki

Age -> dob [godine] Sex -> spol [M: muški, F: ženski] ChestPainType -> vrsta boli u prsima [TA: tipična angina, ATA: atipična angina, NAP: neanginalna bol, ASY: asimptomatska] RestingBP -> krvni tlak u mirovanju [mm Hg] Cholesterol -> razina kolesterola [mm/dl] FastingBS -> krvni šećer [1: ako > 120 mg/dl, 0: inače] RestingECG -> rezultati elektrokardiograma u mirovanju [Normal: normalno, ST: ST-T abnormalnost, LVH: vjerojatna ventrikularna hipertrofija po Estesovom kriteriju] MaxHR -> maksimalni broj otkucaja srca [vrijednost između 60 i 202] ExerciseAngina -> angina tijekom vježbanja [Y: ima, N: nema] Oldpeak -> ST pad tijekom vježbanja ST_Slope -> nagib vrhunca ST segmenta tijekom vježbanja [Up: rastući, Flat: ravni, Down: padajući] HeartDisease -> izlazna klasa [1: srčani bolesnik, 0: normalno]

Numeričke varijable su Age, RestingBP, Cholesterol, MaxHR i Oldpeak. Kategorijske varijable su ChestPainType, RestingBP, FastingBS, RestingECG, ExerciseAngine, ST_Slope te HeartDisease.

Čišćenje podataka

Monotone i konstantne vrijednosti

```
cat("Broj podataka ", nrow(data), "\n")
```

```
## Broj podataka 918
```

```
for (column in colnames(data)){  
  cat("Column: ", column, "ima ", length(unique(data[[column]])), " jedinstvenih vrijednosti\n")  
}
```

```
## Column: Age ima 50 jedinstvenih vrijednosti
## Column: Sex ima 2 jedinstvenih vrijednosti
## Column: ChestPainType ima 4 jedinstvenih vrijednosti
## Column: RestingBP ima 67 jedinstvenih vrijednosti
## Column: Cholesterol ima 222 jedinstvenih vrijednosti
## Column: FastingBS ima 2 jedinstvenih vrijednosti
## Column: RestingECG ima 3 jedinstvenih vrijednosti
## Column: MaxHR ima 119 jedinstvenih vrijednosti
## Column: ExerciseAngina ima 2 jedinstvenih vrijednosti
## Column: Oldpeak ima 53 jedinstvenih vrijednosti
## Column: ST_Slope ima 3 jedinstvenih vrijednosti
## Column: HeartDisease ima 2 jedinstvenih vrijednosti
```

Pošto je broj jedinstvenih vrijednosti za svaku značajku manji od broja redaka, možemo zaključiti da nema monotone vrijednosti (napomena: provjerili smo i ručno kako se ne bi dogodilo da zbog nedostajućih vrijednosti izgleda kao da nema monotonihih vrijednosti). Također, vidimo da nema ni konstantnih vrijednosti jer je broj jedinstvenih vrijednosti za svaku značajku veći od 1.

Nedostajuće vrijednosti

```
data %>% is.na() %>% colSums()
```

```
##           Age           Sex ChestPainType      RestingBP      Cholesterol
##           0             0             0             0             0
##      FastingBS      RestingECG           MaxHR ExerciseAngina           Oldpeak
##           0             0             0             0             0
##      ST_Slope      HeartDisease
##           0             0
```

Nema nedostajućih vrijednosti, međutim, uočili smo veću količinu podataka koji su vrijednost razine kolesterola imali postavljenu na 0, te zato mijenjamo te vrijednosti srednjom vrijednosti preostalih podataka.

```
data$Cholesterol[data$Cholesterol == 0] <- median(data$Cholesterol[data$Cholesterol != 0])
```

Transformiranje kategoričkih značajki u numeričke

Mogli smo koristiti i one-hot encoding za neke značajke, no smatramo da ipak između svih značajki postoji neko uređenje te smo se zato odlučili za label encoding (npr. za značajku ChestPainType može biti TA -> ATA -> NAP -> ASY gdje određeni simptomi postaju manje karakteristični za anginu; ST_Slope također ima uređenje up -> flat -> down). U analizi nismo koristili relativan odnos kategoričkih podataka, tako da odabir enkodiranja nije bio toliko važan.

```
oldCategoricalData <- data
data$Sex <- as.numeric(factor(data$Sex, levels = c("F", "M"), labels = c(0, 1))) - 1
data$ExerciseAngina <- as.numeric(factor(data$ExerciseAngina,
                                         levels = c("N", "Y"), labels = c(0, 1))) - 1
data$ChestPainType <- as.numeric(factor(data$ChestPainType,
                                         levels = c("TA", "ATA", "NAP", "ASY"), labels = c(0, 1, 2, 3)))
data$RestingECG <- as.numeric(factor(data$RestingECG,
                                      levels = c("Normal", "ST", "LVH"), labels = c(0, 1, 2)))
data$ST_Slope <- as.numeric(factor(data$ST_Slope,
                                    levels = c("Up", "Flat", "Down"), labels = c(0, 1, 2)))
```

Deskriptivna analiza numeričkih podataka

Sada kad smo očistili podatke, možemo nastaviti s deskriptivnom statistikom. Prvo ćemo izvući osnovne mjere centralne tendencije, kao što su minimum, maksimum, kvantili i srednja vrijednost, te mjere rasipanja za numeričke podatke te učestalost pojavljivanja po kategorijama za kategoričke.

```
data[-c(2,3,6,7,9,11,12)] %>% summary()
```

```
##           Age           RestingBP           Cholesterol           MaxHR
##  Min.      :28.00   Min.       :  0.0   Min.       : 85.0   Min.       : 60.0
##  1st Qu.:47.00   1st Qu.:120.0   1st Qu.:214.0   1st Qu.:120.0
##  Median :54.00   Median :130.0   Median :237.0   Median :138.0
##  Mean    :53.51   Mean    :132.4   Mean    :243.2   Mean    :136.8
##  3rd Qu.:60.00   3rd Qu.:140.0   3rd Qu.:267.0   3rd Qu.:156.0
##  Max.    :77.00   Max.    :200.0   Max.    :603.0   Max.    :202.0
##      Oldpeak
##  Min.      :-2.6000
##  1st Qu.:  0.0000
##  Median   :  0.6000
##  Mean     :  0.8874
##  3rd Qu.:  1.5000
##  Max.     :  6.2000
```

Iz mjera centralnih tendencija vidimo da su godine i maksimalni broj otkucaja srca malo zakrivljene ulijevo dok su ostale kategorije malo zakrivljene u desno. Također bez računanja možemo primijetiti da zbog jako malog interkvartalnog područja krvni tlak u mirovanju i kolesterol sigurno imaju stršeće vrijednosti.

Izradit ćemo kovarijacijsku matricu za numeričke varijable - vrijednosti prikazuju kovarijance varijable u retku i stupcu, a na dijagonali se nalaze varijance pojedinih varijabli.

```
covMat <- cov(data[-c(2,3,6,7,9,11,12)])
covMat
```

```
##           Age  RestingBP  Cholesterol  MaxHR  Oldpeak
## Age      88.974254  44.427519   22.936698 -91.750920  2.601774
## RestingBP 44.427519 342.773903   83.575197 -52.857808  3.254307
## Cholesterol 22.936698 83.575197 2851.698472 -2.175748  3.104117
## MaxHR     -91.750920 -52.857808 -2.175748 648.228614 -4.363589
## Oldpeak    2.601774  3.254307   3.104117 -4.363589  1.137572
```

Izradit ćemo korelacijsku matricu za numeričke varijable - vrijednosti prikazuju koeficijente korelacije između pojedinih varijabli. To bi nam moglo biti zanimljivo za dobivanje nekog početnog dojma o međusobnim ovisnostima tih varijabli.

```
corrMat <- cor(data[-c(2,3,6,7,9,11,12)])
corrMat
```

```
##           Age  RestingBP  Cholesterol  MaxHR  Oldpeak
## Age      1.00000000  0.25439936  0.045535165 -0.382044675  0.25861154
## RestingBP 0.25439936  1.00000000  0.084532114 -0.112134997  0.16480304
## Cholesterol 0.04553517  0.08453211  1.000000000 -0.001600268  0.05450005
## MaxHR     -0.38204468 -0.11213500 -0.001600268  1.000000000 -0.16069055
## Oldpeak    0.25861154  0.16480304  0.054500050 -0.160690550  1.00000000
```

Pearsonov koeficijent korelacije je realan broj između -1 i 1, gdje veća apsolutna vrijednost prikazuje veću linearnu zavisnost. Iz tablice se čini da varijable nisu toliko korelirane.

Učestalost pojavljivanja vrijednosti za kategoričke značajke

```
for (column in c(2, 3, 6, 7, 9, 11, 12)) {  
  column_name <- colnames(oldCategoricalData)[column]  
  column_values <- oldCategoricalData[[column]]  
  cat("Column:", column_name, "ima vrijednosti:\n")  
  print(table(column_values))  
  cat("\n")  
}
```

```
## Column: Sex ima vrijednosti:  
## column_values  
##   F   M  
## 193 725  
##  
## Column: ChestPainType ima vrijednosti:  
## column_values  
## ASY ATA NAP  TA  
## 496 173 203  46  
##  
## Column: FastingBS ima vrijednosti:  
## column_values  
##    0    1  
## 704 214  
##  
## Column: RestingECG ima vrijednosti:  
## column_values  
##   LVH Normal    ST  
##   188    552   178  
##  
## Column: ExerciseAngina ima vrijednosti:  
## column_values  
##    N    Y  
## 547 371  
##  
## Column: ST_Slope ima vrijednosti:  
## column_values  
## Down Flat   Up  
##   63  460 395  
##  
## Column: HeartDisease ima vrijednosti:  
## column_values  
##    0    1  
## 410 508
```

Analiza numeričkih podataka u raznim kombinacijama

Analiza podataka ovisno o spolu

```
print(data %>% group_by(Sex) %>% summarise(  
  NumOfPatients = n()),
```

```
HeartDiseaseRatio = sum(HeartDisease)/n(),
FastingBSRatio = sum(FastingBS)/n(),
ExerciseAnginaRatio = sum(ExerciseAngina)/n(),
Mean.Age = mean(Age),
Mean.RestingBP = mean(RestingBP),
Mean.Cholesterol = mean(Cholesterol),
Mean.MaxHR = mean(MaxHR),
), width = Inf)
```

```
## # A tibble: 2 x 9
##   Sex NumOfPatients HeartDiseaseRatio FastingBSRatio ExerciseAnginaRatio
##   <dbl>      <int>          <dbl>          <dbl>          <dbl>
## 1     0        193            0.259            0.135            0.223
## 2     1        725            0.632            0.259            0.452
##   Mean.Age Mean.RestingBP Mean.Cholesterol Mean.MaxHR
##   <dbl>      <dbl>          <dbl>          <dbl>
## 1    52.5        132.            255.            146.
## 2    53.8        132.            240.            134.
```

Na prvu uočavamo razlike između srednjih vrijednosti najvećeg broja otkucaja srca (MaxHR) i količine kolesterola (Cholesterol), te, možda najvažnije, razliku u udjelu srčanih bolesnika među muškarcima i ženama. To bi moglo motivirati pitanja o utjecaju spola na vrijednosti tih parametara.

Analiza ovisno o prisutnosti srčane bolesti

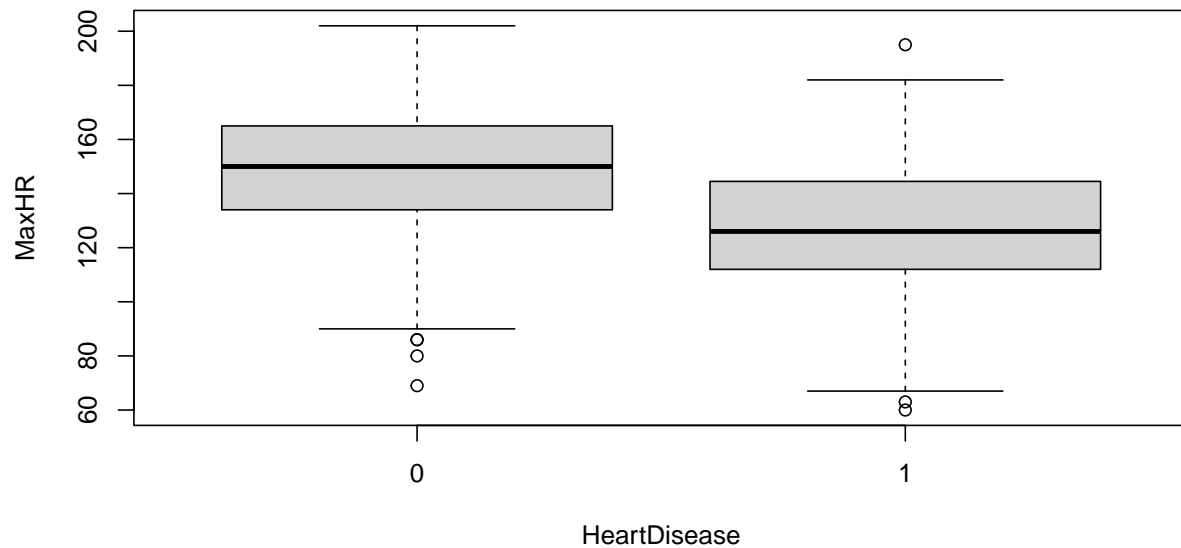
```
print(data %>% group_by(HeartDisease) %>% summarise(
  NumOfPatients = n(),
  FastingBSRatio = sum(FastingBS)/n(),
  ExerciseAnginaRatio = sum(ExerciseAngina)/n(),
  Mean.Age = mean(Age),
  Mean.RestingBP = mean(RestingBP),
  Mean.Cholesterol = mean(Cholesterol),
  Mean.MaxHR = mean(MaxHR),
), width = Inf)
```

```
## # A tibble: 2 x 8
##   HeartDisease NumOfPatients FastingBSRatio ExerciseAnginaRatio Mean.Age
##   <dbl>      <int>          <dbl>          <dbl>          <dbl>
## 1         0        410            0.107            0.134            50.6
## 2         1        508            0.335            0.622            55.9
##   Mean.RestingBP Mean.Cholesterol Mean.MaxHR
##   <dbl>          <dbl>          <dbl>
## 1    130.            239.            148.
## 2    134.            247.            128.
```

Ovdje konkretno vidimo izračunate mjere srednjih vrijednosti parametara za srčane bolesnike i one koji to nisu. Vidimo da se vrijednosti parametara Cholesterol i MaxHR razlikuju po grupama, kao i udio onih s povišenim šećerom i onih s vježbom induciranom anginom. To bi moglo motivirati pitanja može li se na temelju vrijednosti takvih parametara predvidjeti postojanje srčane bolesti.

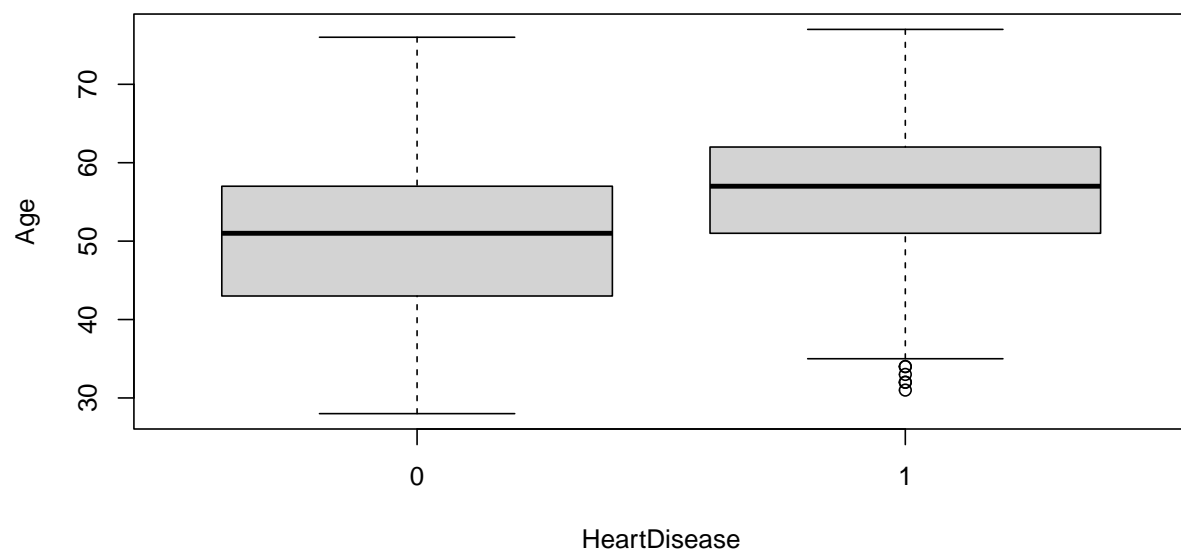
Najbolje je to prvo analizirati boxplot dijagramima, paralelno za srčane bolesnike i zdrave pacijente.

```
boxplot(MaxHR ~ HeartDisease, data = data)
```



Ovaj bi boxplot mogao sugerirati na razliku u vrijednosti MaxHR parametra ovisno o tome je li pacijent srčani bolesnik ili ne - vidimo da za srčane bolesnike vrijedi da su vrijednosti MaxHR-a niže te da medijan bude manji te čak izvan vrijednosti prvog kvartila u skupini srčanih nebolesnika.

```
boxplot(Age ~ HeartDisease, data = data)
```



Ponovno, ovaj boxplot mogao bi sugerirati razliku u dobi pacijenata srčanih bolesnika i nebolesnika - vidimo

da za srčane bolesnike vrijedi da su veće dobi te da medijan bude veći te čak izvan vrijednosti trećeg kvartila u skupini srčanih nebolesnika.

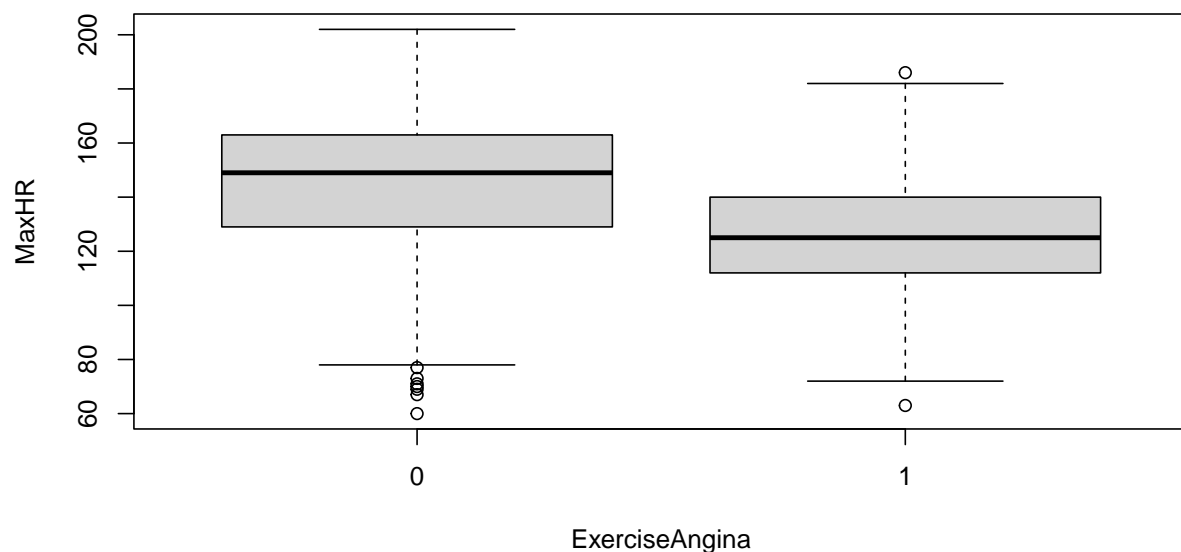
```
unNormalizedData <- data
```

ISTRAŽIVAČKA PITANJA

VAŽNA NAPOMENA: Alfa vrijednost određuje se prije provođenja statističkog testa; mi ćemo u svim testovima uzeti da je $\alpha = 0.05$ odnosno 5%.

1. pitanje: Postoji li razlika u maksimalnom broju otkucaja srca između pacijenata s anginom i onih bez angine?

```
boxplot(MaxHR ~ ExerciseAngina, data=unNormalizedData)
```



Ovaj boxplot sugerira da bi mogla postojati razlika u najvećem broju otkucaja, obzirom da se medijan kategorije s anginom nalazi ispod vrijednosti prvog kvartila onih bez angine.

Kako bismo odgovorili na ovo pitanje trebamo koristiti T-test za dva uzorka. Znamo da T-test ima neke osnovne pretpostavke koje moramo provjeriti, a moramo provjeriti i jesu li varijance podataka jednake ili različite. Prvo ćemo pogledati jesu li podatci uzorkovani iz normalne razdiobe, a nakon toga provesti F-test kako bismo utvrdili jesu li varijance jednake ili različite.

```
withAngine <- data[data$ExerciseAngina == 1,]
withoutAngine <- data[data$ExerciseAngina == 0,]

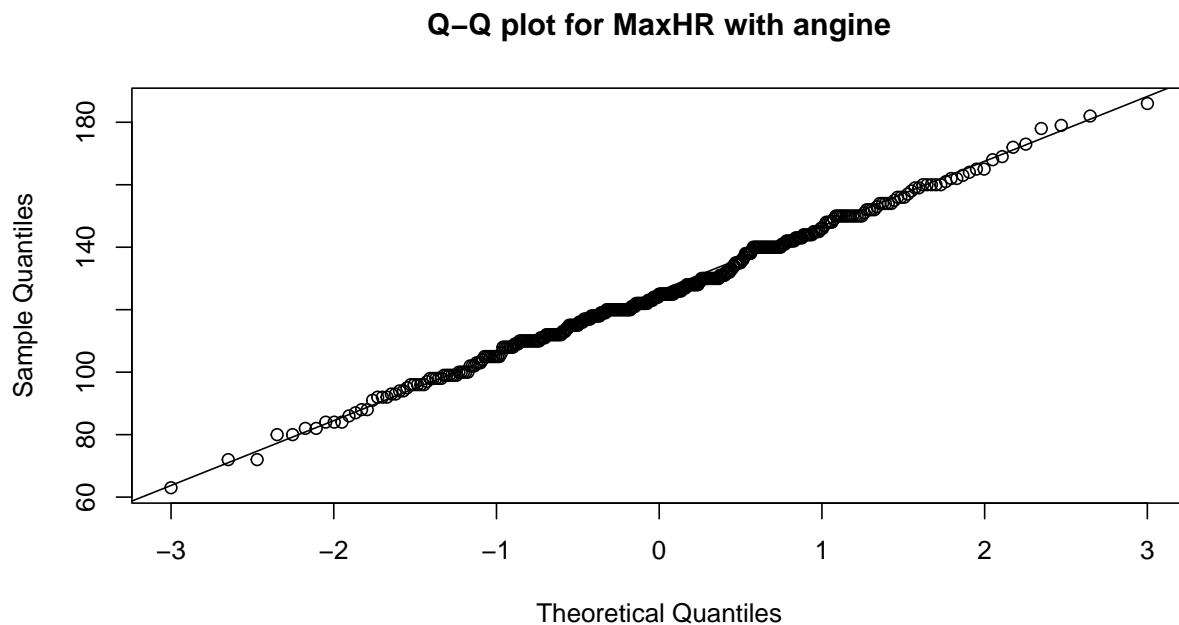
cat("Broj pacijenata s anginom:", nrow(withAngine),
    "\nBroj pacijenata bez angine:", nrow(withoutAngine))
```



```
## Broj pacijenata s anginom: 371  
## Broj pacijenata bez angine: 547
```

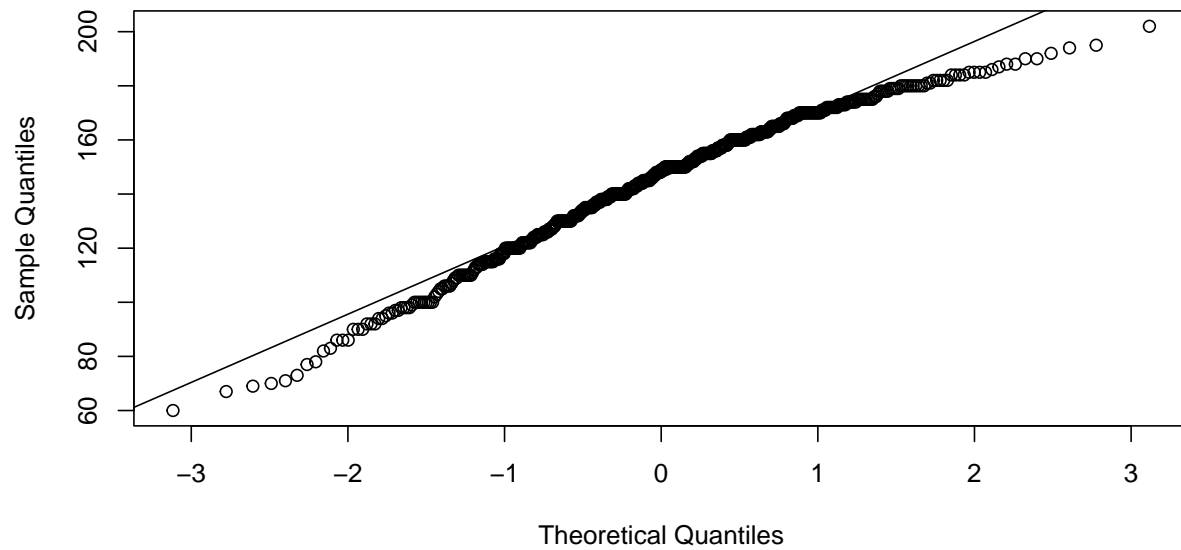
```
maxHRWithAngine <- withAngine$MaxHR  
maxHRWithoutAngine <- withoutAngine$MaxHR
```

```
# crtamo Q-Q plot kako bismo vidjeli odgovaraju li podatci normalnoj raspodjeli  
qqnorm(maxHRWithAngine, main = "Q-Q plot for MaxHR with angine")  
qqline(maxHRWithAngine)
```



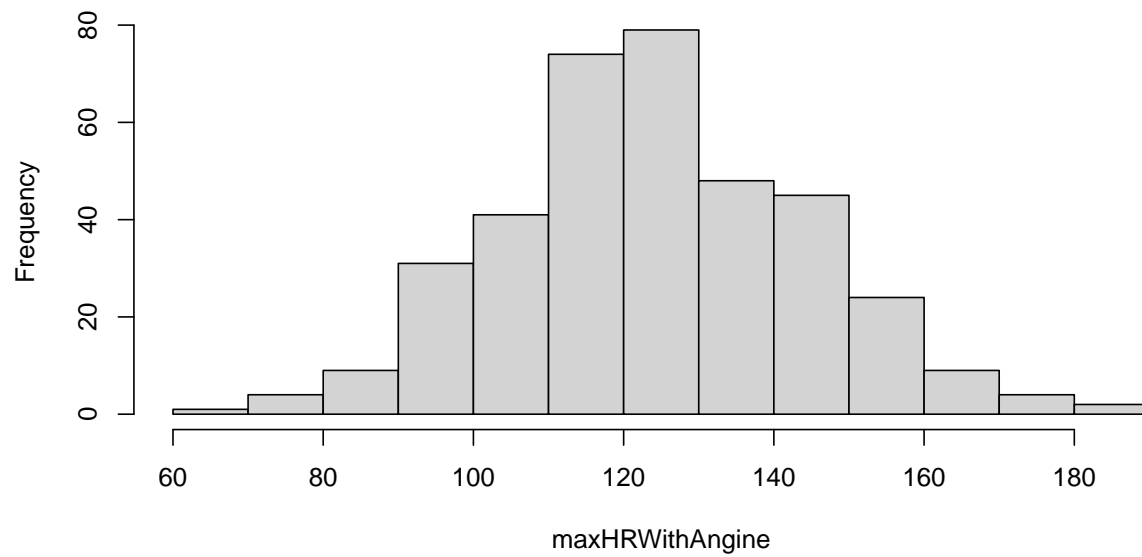
```
qqnorm(maxHRWithoutAngine , main = "Q-Q plot for MaxHR without angine")  
qqline(maxHRWithoutAngine)
```

Q-Q plot for MaxHR without engine

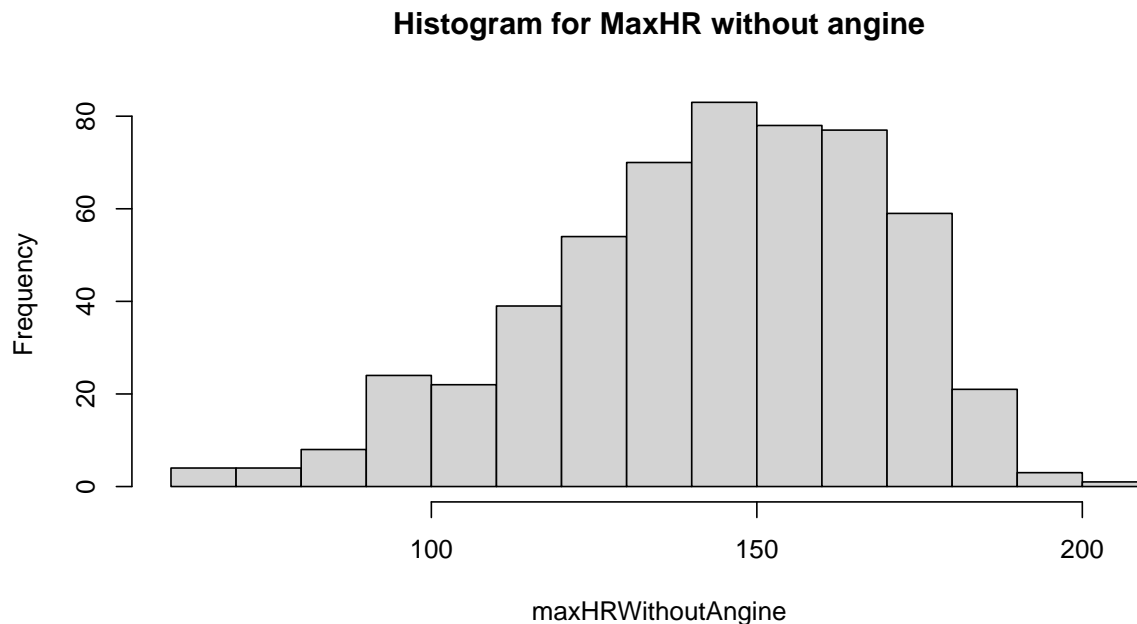


```
# isto radimo s histogramima  
hist(maxHRWithEngine, main = "Histogram for MaxHR with engine")
```

Histogram for MaxHR with engine



```
hist(maxHRWithoutEngine, main = "Histogram for MaxHR without engine")
```



Iz grafova možemo zaključiti da se za osobe s anginom značajka MaxHR ravna po normalnoj razdiobi, dok se za osobe bez angine sredina ravna po normalnoj raspodjeli, no repovi su različiti od normalne raspodjele. Dodatno, histogram za osobe bez angine pokazuje da su podatci malo pomaknuti od sredine 0. Možemo provesti Lillieforsov test kako bi provjerili odgovaraju li podatci normalnoj razdiobi. Pretpostavku nezavisnosti podataka temeljimo na činjenici da su podatci preuzeti iz različitih skupina odnosno od različitih ljudi.

Lilliefors (Kolmogorov-Smirnov) test normalnosti podataka: Provest ćemo Lilliefors test kako bismo provjerili odgovaraju li podatci normalnoj razdiobi. Za provođenje testa mora biti zadovoljen uvjet da su podatci nezavisni.

Hipoteza H0: podatci su uzorkovani iz normalne razdiobe Hipoteza H1: podatci nisu uzorkovani iz normalne razdiobe

```
library(nortest)
lillie.test(maxHRWithoutEngine)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  maxHRWithoutEngine
## D = 0.075696, p-value = 6.45e-08
```

```
lillie.test(maxHRWithEngine)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  maxHRWithEngine
## D = 0.054536, p-value = 0.009992
```

Iz rezultata testa možemo vidjeti da je p vrijednost manja od $\alpha = 5\%$. Test nam ukazuje da postoje statistički značajne razlike između podataka i normalne razdiobe. No kako nam prethodni grafovi pokazuju da su podatci približno normalno raspoređeni i kako je T test dosta robustan na normalnost podataka, možemo pretpostaviti da su podatci uzorkovani iz normalne razdiobe i nastaviti s T-testom.

Nastavljamo s F-testom kako bismo utvrdili jesu li varijance obaju skupova podataka jednake ili različite. Ako imamo dva nezavisna slučajna uzorka (već ranije smo ovo pretpostavili), koji dolaze iz normalnih (provjerili ranije) distribucija s varijancama σ_1^2 i σ_2^2 , tada slučajna varijabla

$$F = \frac{S_{X_1}^2 / \sigma_1^2}{S_{X_2}^2 / \sigma_2^2}$$

ima Fisherovu distribuciju s $(n_1 - 1, n_2 - 1)$ stupnjeva slobode Hipoteza H_0 : varijance su jednake Hipoteza H_1 : varijance su različite

```
var(maxHRWithoutAngine)
```

```
## [1] 655.8826
```

```
var(maxHRWithAngine)
```

```
## [1] 418.2429
```

```
var.test(maxHRWithAngine, maxHRWithoutAngine)
```

```
##
## F test to compare two variances
##
## data: maxHRWithAngine and maxHRWithoutAngine
## F = 0.63768, num df = 370, denom df = 546, p-value = 3.641e-06
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.5298767 0.7700287
## sample estimates:
## ratio of variances
## 0.6376795
```

Vidimo kako je p-vrijednost značajno manja od $\alpha = 5\%$ pa zaključujemo da su varijance različite; radi toga koristimo T-test za dva uzorka s različitim varijancama. Nakon što smo provjerili sve uvjete za T-test, možemo ga provesti.

Hipoteza H_0 : nema razlike u maksimalnom broju otkucaja srca između pacijenata s anginom i onih bez angine Hipoteza H_1 : postoji razlika u maksimalnom broju otkucaja srca između pacijenata s anginom i onih bez angine

```
t.test(maxHRWithAngine, maxHRWithoutAngine, var.equal = FALSE)
```

```
##
## Welch Two Sample t-test
##
## data: maxHRWithAngine and maxHRWithoutAngine
## t = -12.594, df = 891.9, p-value < 2.2e-16
```

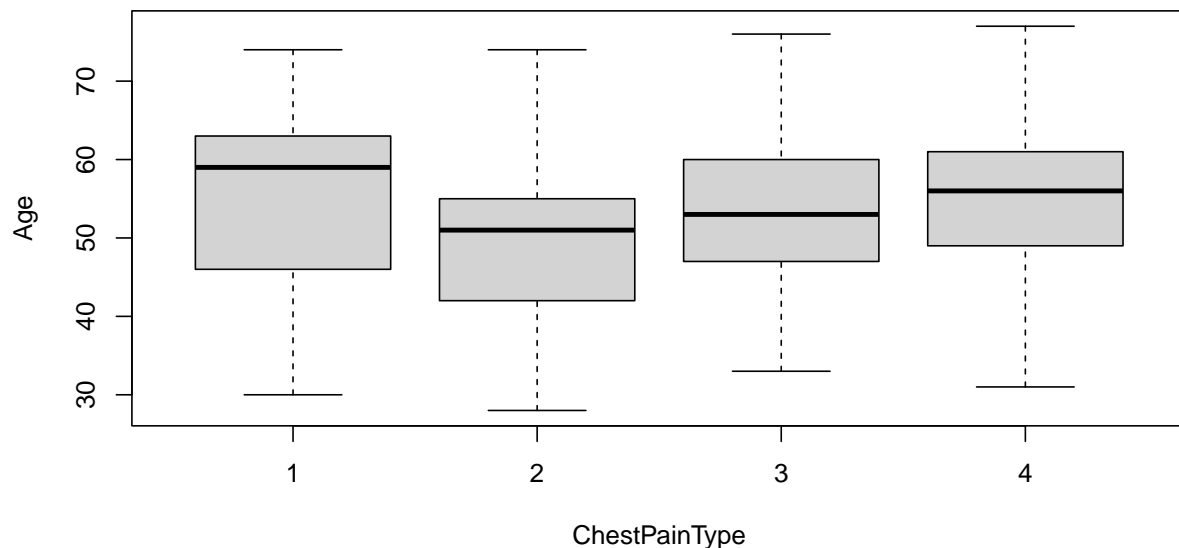
```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -22.20183 -16.21483
## sample estimates:
## mean of x mean of y
## 125.3639 144.5722
```

Iz rezultata vidimo kako je p-vrijednost značajno manja od $\alpha = 5\%$ pa zaključujemo da postoji razlika u maksimalnom broju otkucaja srca između pacijenata s anginom i onih bez angine.

2. pitanje: Postoji li značajna razlika u starosti pacijenata s obzirom na različite vrste boli u prsima?

Kako bismo dobili neku početnu intuiciju, radimo boxplot prikaz.

```
boxplot(Age ~ ChestPainType, data = unNormalizedData)
```



Vidimo da postoje razlike u medijanima pojedinih grupa. Računamo uzoračku sredinu i varijancu pojedinih grupa.

```
print(unNormalizedData %>% group_by(ChestPainType) %>% summarise(
  NumOfPatients = n(),
  Mean.Age = mean(Age),
  Med.Age = median(Age),
  Std.Age = sd(Age),
  Var.Age = var(Age)
), width = Inf)
```

```
## # A tibble: 4 x 6
##   ChestPainType NumOfPatients Mean.Age Med.Age Std.Age Var.Age
```

##	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
## 1	1	46	54.8	59	11.4	131.
## 2	2	173	49.2	51	9.26	85.7
## 3	3	203	53.3	53	9.61	92.3
## 4	4	496	55.0	56	8.76	76.8

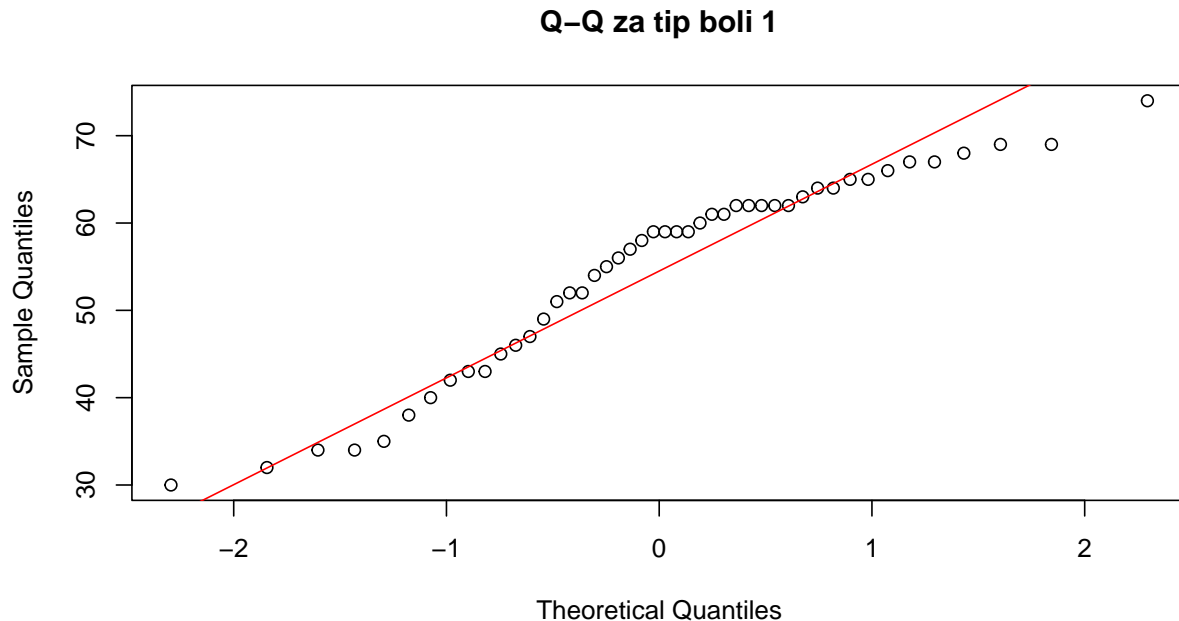
U prvoj je kategoriji 46 pacijenata, u drugoj 173, u trećoj 203 te u zadnjoj 496 pacijenata. Uočavamo i razliku između varijanci po kategorijama.

Test kojim bi ispitali jednakost ovih sredina jest ANOVA (analiza varijanci). Pretpostavke tog testa su nezavisnost, normalnost populacija grupa, kao i jednakost njihovih varijanci. Nezavisnost pretpostavljamo iz činjenice da je svaki podatak od druge osobe.

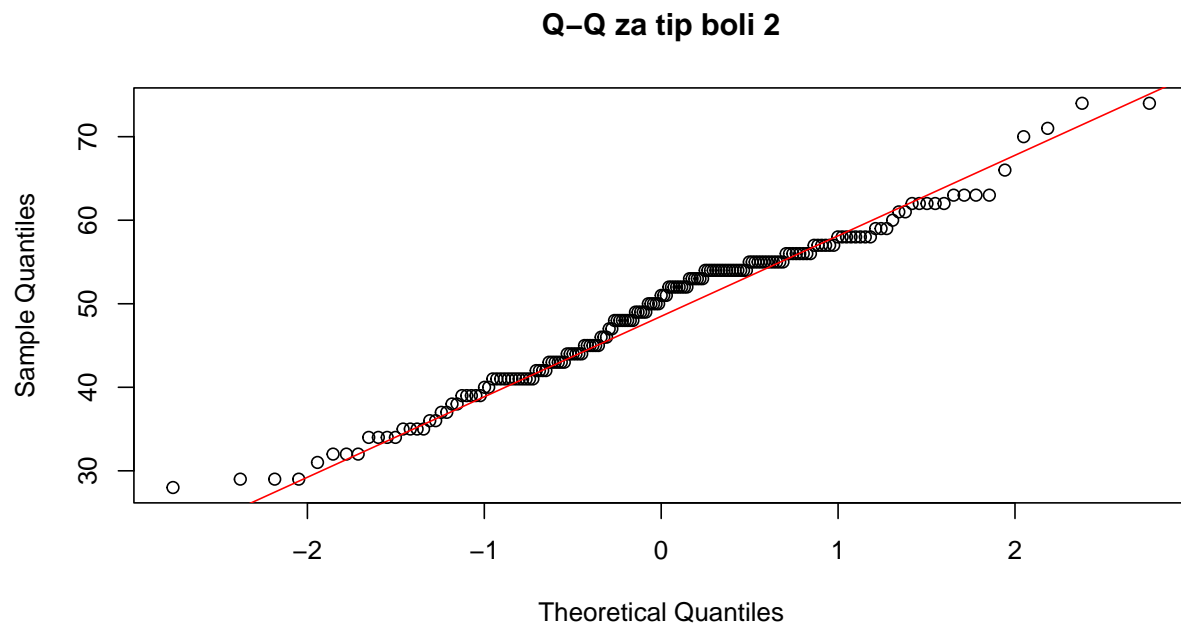
Kako bismo provjerili normalnost podataka iz pojedinih grupa, vizualiziramo ih histogramima i Q-Q plotovima.

```
useddata <- unNormalizedData
agePainType1 <- useddata[useddata$ChestPainType == 1,]$Age
agePainType2 <- useddata[useddata$ChestPainType == 2,]$Age
agePainType3 <- useddata[useddata$ChestPainType == 3,]$Age
agePainType4 <- useddata[useddata$ChestPainType == 4,]$Age

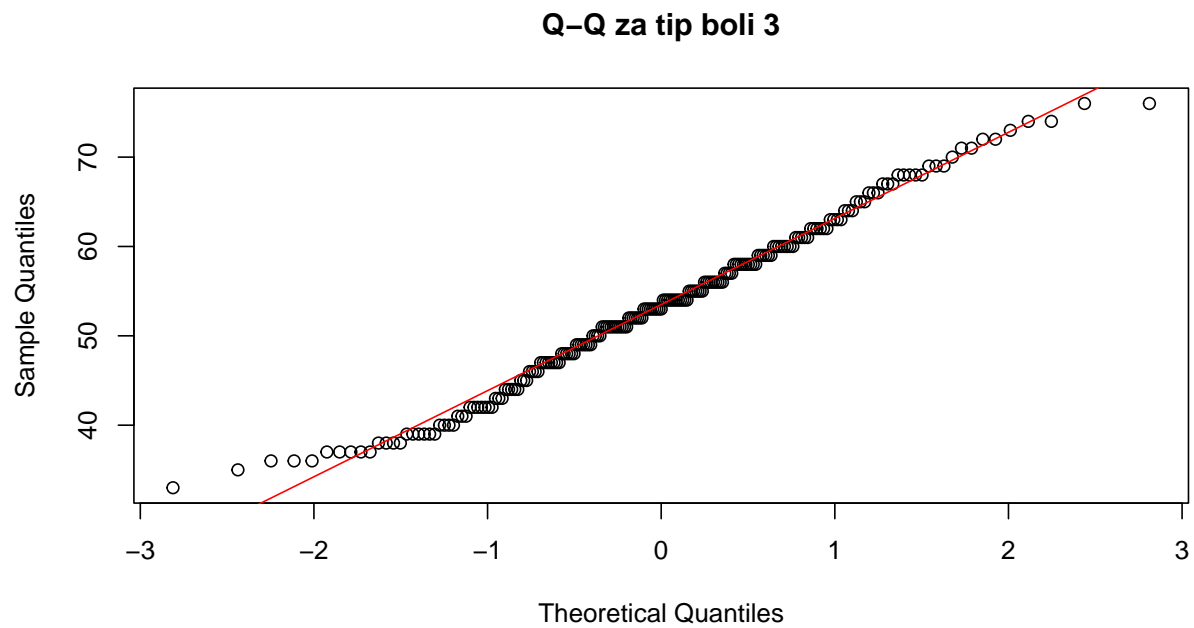
{qqnorm(agePainType1, main = "Q-Q za tip boli 1")
  qqline(agePainType1, col = "red")}
```



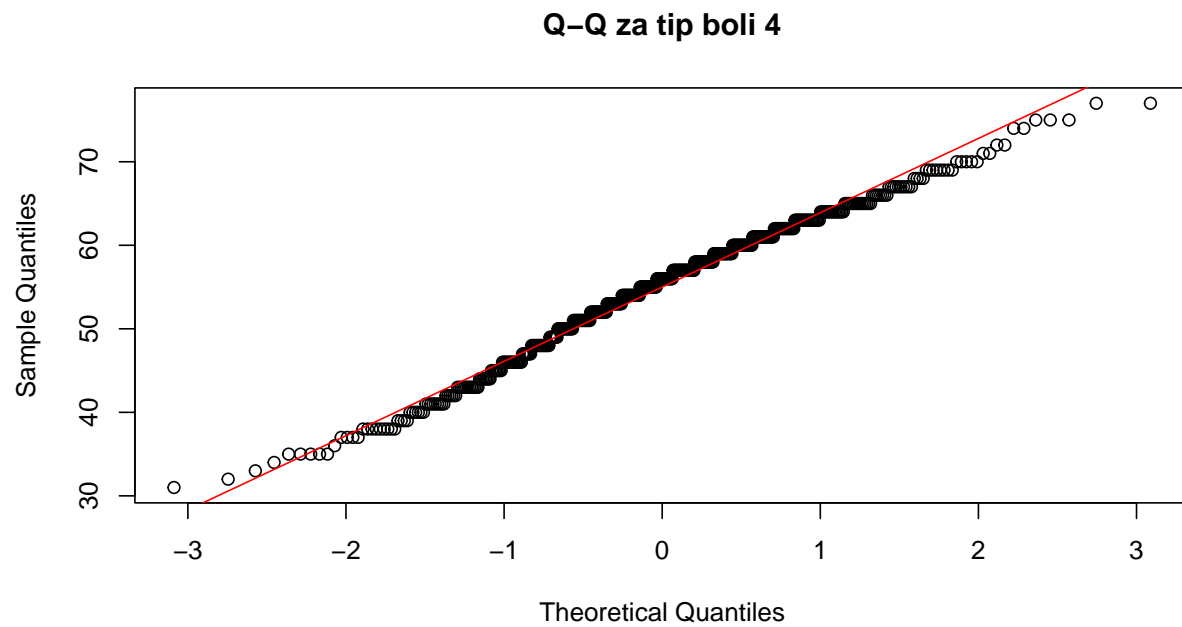
```
{qqnorm(agePainType2, main = "Q-Q za tip boli 2")
  qqline(agePainType2, col = "red")}
```



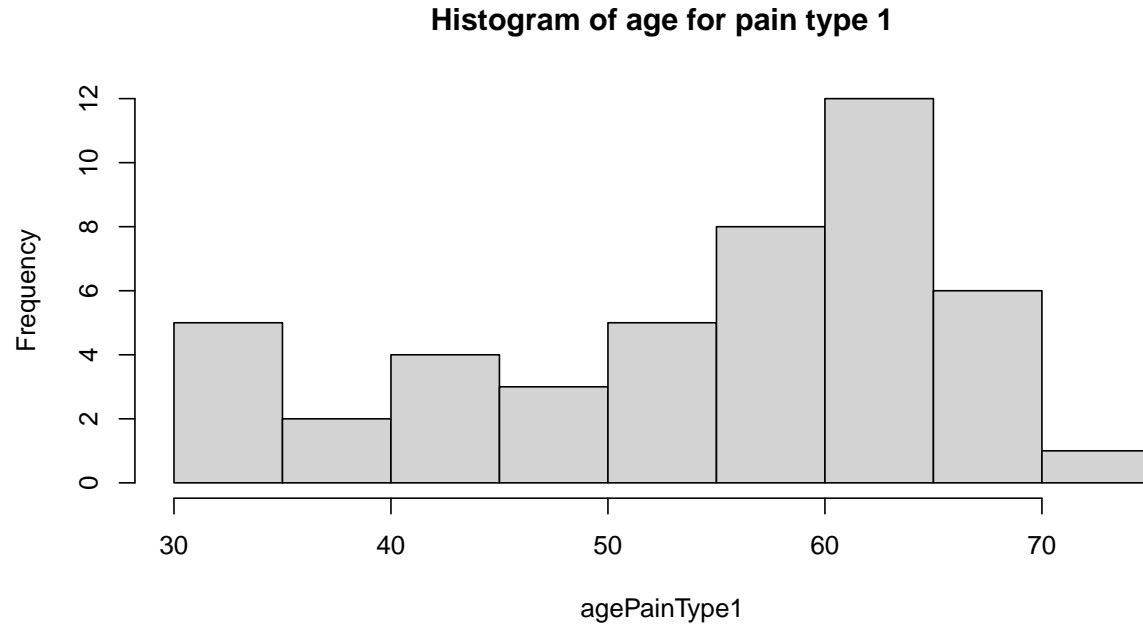
```
{qqnorm(agePainType3, main = "Q-Q za tip boli 3")
 qqline(agePainType3, col = "red")}
```



```
{qqnorm(agePainType4, main = "Q-Q za tip boli 4")
 qqline(agePainType4, col = "red")}
```

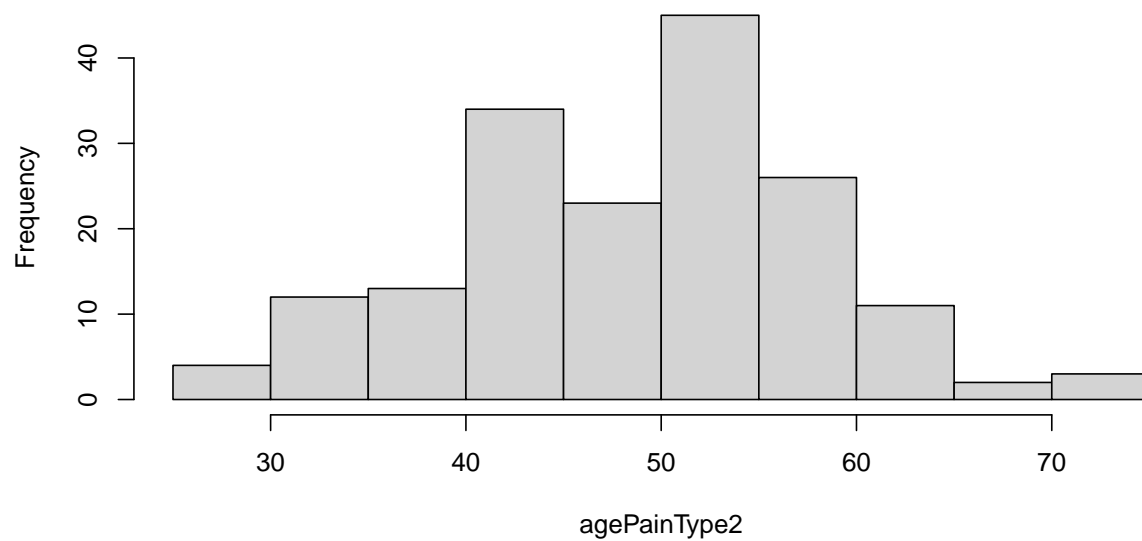


```
hist(agePainType1, main = "Histogram of age for pain type 1")
```



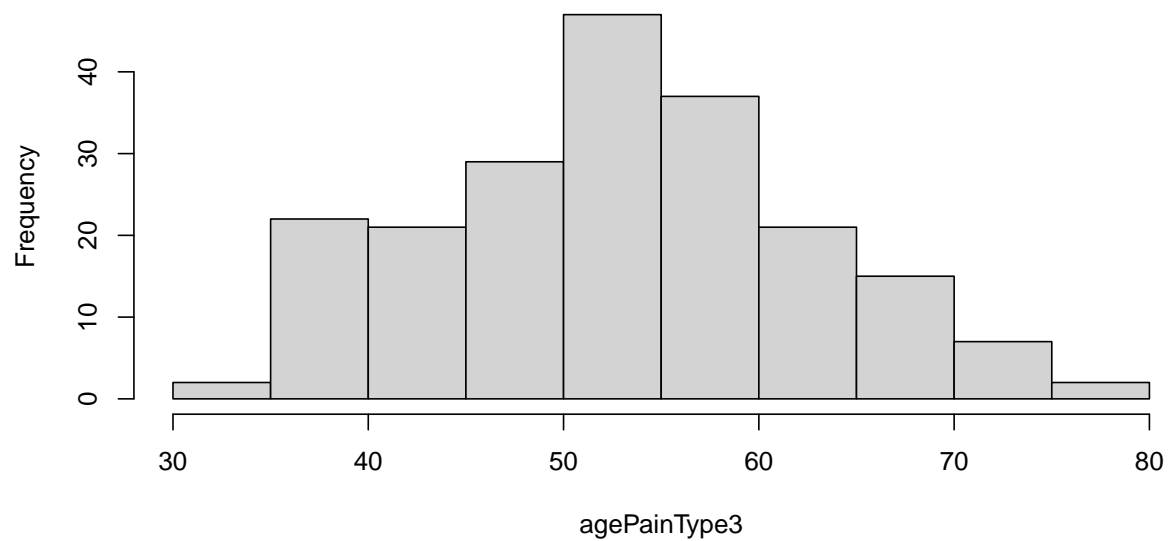
```
hist(agePainType2, main = "Histogram of age for pain type 2")
```


Histogram of age for pain type 2

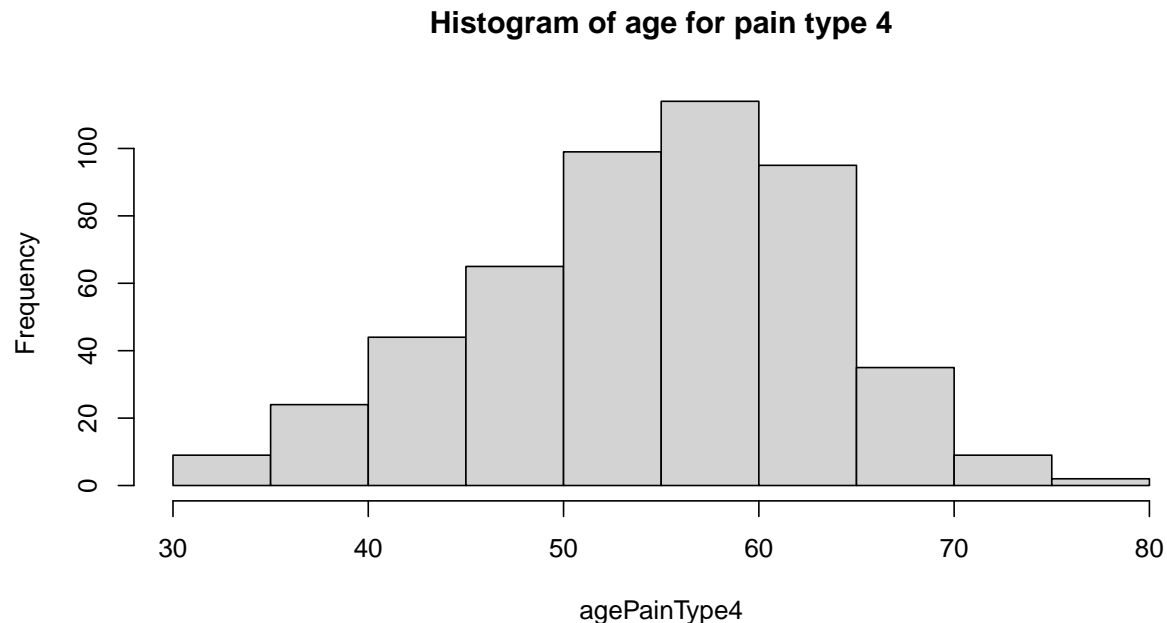


```
hist(agePainType3, main = "Histogram of age for pain type 3")
```

Histogram of age for pain type 3



```
hist(agePainType4, main = "Histogram of age for pain type 4")
```



Najslabiju pripadnost normalnoj distribuciji pokazuju grafovi grupe 1. Pokušali smo razne transformacije dobi (npr. x^n), međutim nismo uspjeli istovremeno postići bolje rezultate na svim grupama. Budući da su ostali grafovi zadovoljavajući, pretpostavljamo normalnost i nastavimo s analizom ovakvih podataka.

Kako bismo ispitali jednakost varijanci po grupama, koristimo Bartlettov test. Nulta hipoteza tog testa jest da su varijance po svim grupama jednake, dok je alternativna hipoteza da za barem jedan par grupa vrijedi da to nije tako.

```
bartlett.test(Age ~ ChestPainType, unNormalizedData)
```

```
##
## Bartlett test of homogeneity of variances
##
## data: Age by ChestPainType
## Bartlett's K-squared = 8.0402, df = 3, p-value = 0.04519
```

Na razini značajnosti od 5% bismo odbacili nultu hipotezu, odnosno podaci sugeriraju da varijance po grupama nisu iste. Međutim, ANOVA test je relativno robustan na nejednakost varijanci, pa nastavljamo.

Konačno, vršimo ANOVA test.

```
anova <- aov(Age ~ factor(ChestPainType), data=unNormalizedData)
summary(anova)
```

```
##               Df Sum Sq Mean Sq F value    Pr(>F)
## factor(ChestPainType)  3   4280   1426.8    16.87 1.14e-10 ***
## Residuals          914   77309     84.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vidimo da na bismo na razini značajnosti odbacili nultu hipotezu, odnosno podaci sugeriraju da postoji razlika u starosti ovisno o vrsti boli u prsima.

Korištenje parametarskog ANOVA test gdje su pretpostavke normalnosti i jednakosti varijanci prekršene može dovesti do povećanja pogreške prvog red, odnosno, vjerojatnije je da ćemo dobiti da postoje značajne razlike u sredinama, što je moguće da se dogodilo i ovdje. Najveći je problem nesrazmjer u veličinama pojedinih grupa te bi najbolje bilo prikupiti još podataka o dobi grupe 1.

Puno nas toga odvraća od parametarskog ANOVA testa, pa se odlučujemo i za provedbu njegove neparametarske alternative koja ne zahtjeva te pretpostavke - Kruskal-Wallisova testa. Uvjet za primjenjivost Kruskal-Wallisovog testa jest da je veličina svakog uzorka (grupe) barem 5, što je kod nas slučaj. Nulta hipoteza je da su sve sredine po grupama jednake, a alternativna da za barem jedan par grupa to ne vrijedi.

```
kruskal.test(Age ~ ChestPainType, data=unNormalizedData)

##
##  Kruskal-Wallis rank sum test
##
## data:  Age by ChestPainType
## Kruskal-Wallis chi-squared = 48.599, df = 3, p-value = 1.588e-10
```

Vidimo da je p-vrijednost jako malena, što znači da odbacujemo nultu hipotezu, odnosno, podaci sugeriraju da postoji značajna razlika u starosti ovisno o tipu boli u prsima.

3. pitanje: Možemo li predvidjeti prisutnost srčane bolesti na temelju maksimalnog broja otkucaja srca i starosti?

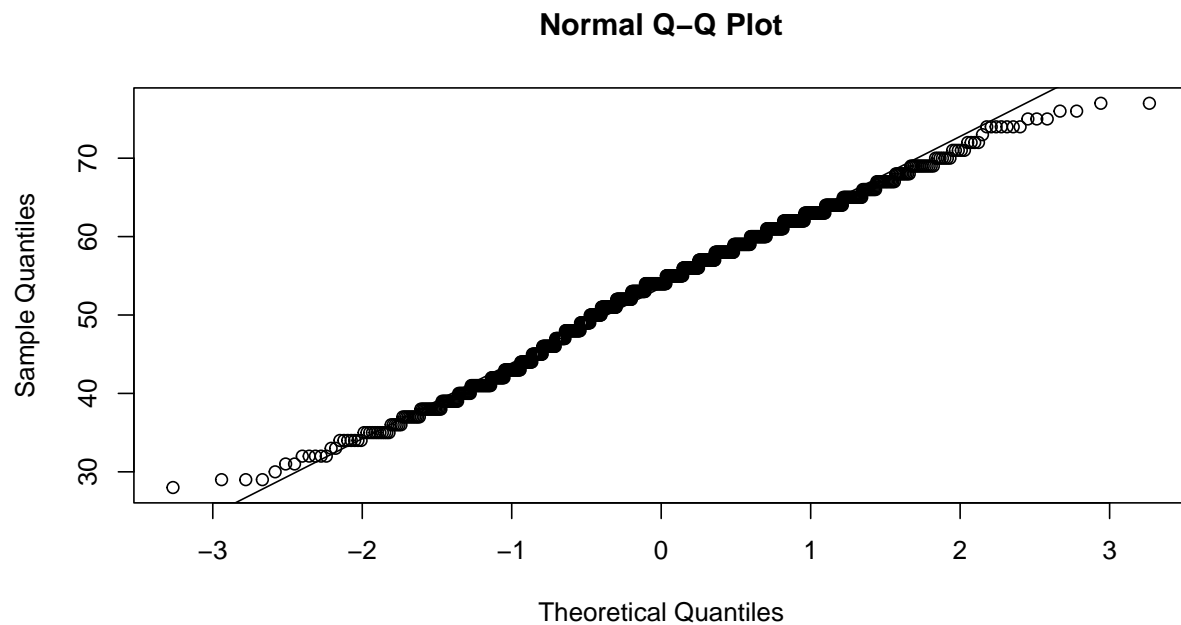
Za pronalazak odgovora na treće pitanje koristimo logističku regresiju. Linearna regresija nije dobar izbor jer je problem klasifikacijski. Linearna regresija loše klasificira jer funkcija pogreške kažnjava i točno klasificirane primjere koji su daleko od granice odluke te zbog toga stršeće vrijednosti mogu značajno utjecati na pomak granice odluke (čak i kada su točno klasificirane). Također, linearna regresija nam vraća vrijednosti izvan intervala [0, 1].

Za logističku regresiju trebamo provjeriti nekoliko pretpostavki: 1. Ne smije biti multikolinearnosti u podacima 2. Podatci moraju biti nezavisni (ulazi su nezavisne varijable dok je izlaz y zavisna varijabla) 3. Nema nedostajućih vrijednosti i klase imaju balansirani broj primjera (u našem slučaju oba uvjeta su ispunjena, provjera je napravljena u fazi čišćenja podataka)

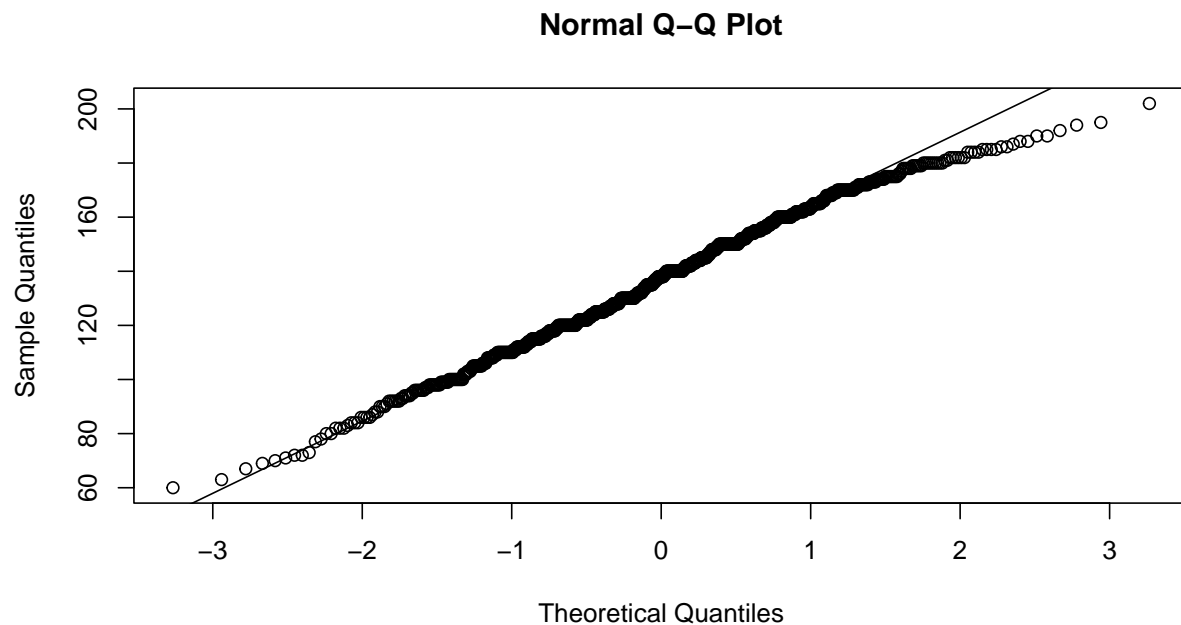
Kako bismo mogli rezultate logističke regresije interpretirati kao vjerojatnost moraju biti zadovoljene i navedene pretpostavke: (1) primjeri iz obiju klasa normalno distribuirani oko srednje, prototipne vrijednosti (tj. izglednost je Gaussova gustoća vjerojatnosti) (2) da postoji linearna zavisnost između izvora šuma koja je u obje klase identična (tj. kovarijacijska matrica je dijeljena), onda izlaz logističke regresije doista odgovara aposteriornoj vjerojatnosti oznake y za primjer x.

Ako ove pretpostavke ne vrijede, onda nemamo teorijski model uz koje bi izlaz logističke regresije odgovarao aposteriornoj vjerojatnosti. Međutim, u praksi se time previše ne zamaramo, tj. izlaz logističke regresije tumačimo kao vjerojatnost neovisno o tome koliko podatci doista odgovaraju navedenim pretpostavkama. Tako da navedene dvije pretpostavke nećemo detaljno provjeravati one služe kao teorijska podloga. Radi potpunosti ćemo nacrtati qq plot za provjeru normalnosti.

```
#normalnost
qqnorm(oldCategoricalData$Age)
qqline(oldCategoricalData$Age)
```



```
qqnorm(oldCategoricalData$MaxHR)
qqline(oldCategoricalData$MaxHR)
```



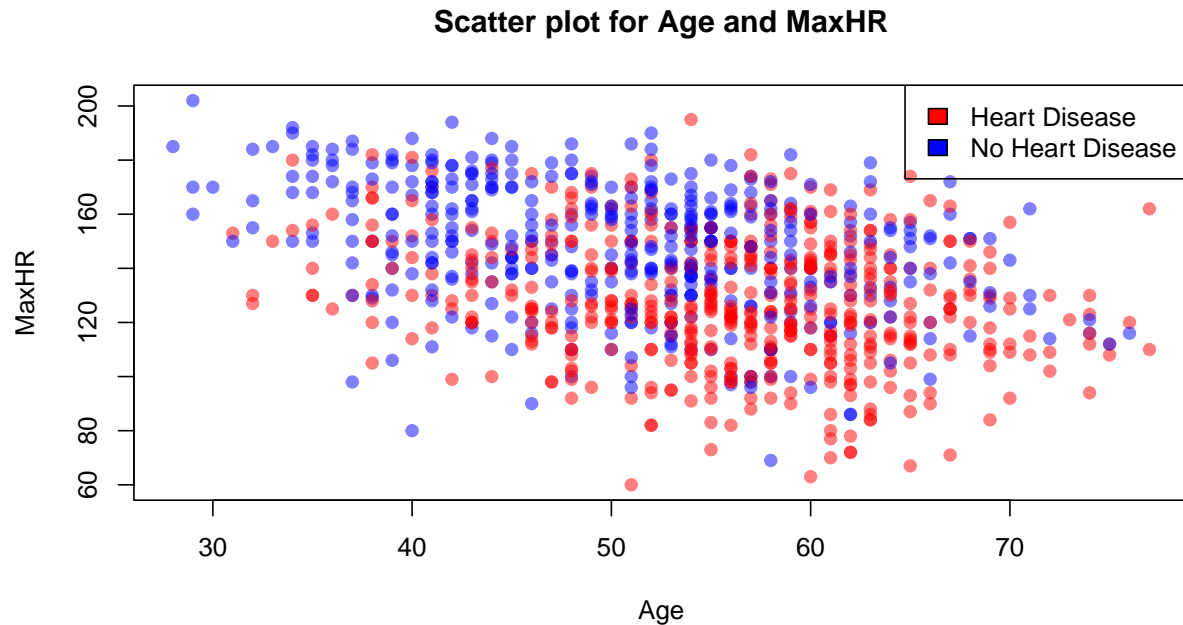
Iz qq plota možemo vidjeti kako podatci prate normalnu razdiobu s malim odstupanjima u repovima.

Prije provedbe logističke regresije poželjno je vizualizirati podatke kako bismo dobili bolji uvid u njih. U našem slučaju možemo koristiti graf raspršenja jer imamo samo dvije varijable.

```

colors <- ifelse(oldCategoricalData$HeartDisease == 1, rgb(1,0,0,0.5), rgb(0,0,1,0.5))
plot(oldCategoricalData$Age, oldCategoricalData$MaxHR, col = colors,
     main = "Scatter plot for Age and MaxHR", xlab = "Age", ylab = "MaxHR",
     pch = 19)
legend("topright", legend = c("Heart Disease", "No Heart Disease"), fill = c("red", "blue"))

```



Iz grafa je vidljivo kako su srčane bolesti rjeđe kod mlađih ljudi koji imaju veći maksimalan broj otkucaja srca, dok stariji ljudi s manjim maksimalnim brojem otkucaja srca češće imaju srčane bolesti.

Provjera pretpostavki logističke regresije

Funkcijom `cor` tražimo korelaciju između naših dviju varijabli.

```

library(ggcorrplot)

predictor <- data.frame(data$Age, data$MaxHR)
cor(predictor)

```

```

##           data.Age data.MaxHR
## data.Age      1.0000000 -0.3820447
## data.MaxHR    -0.3820447  1.0000000

```

```

#ggcorrplot(cor(predictor), hc.order = TRUE, type = "lower", lab = TRUE)

```

Dobili smo korelaciju od -0.38 što je relativno slaba korelacija, pa možemo nastaviti s provođenjem logističke regresije.

```
logisticModel <- glm(HeartDisease ~ MaxHR + Age, data = data, family = binomial)
summary(logisticModel)
```

```
##
## Call:
## glm(formula = HeartDisease ~ MaxHR + Age, family = binomial,
##      data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.666240   0.738178   3.612 0.000304 ***
## MaxHR       -0.032387   0.003420  -9.469 < 2e-16 ***
## Age          0.037807   0.008494   4.451 8.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1262.1  on 917  degrees of freedom
## Residual deviance: 1082.6  on 915  degrees of freedom
## AIC: 1088.6
##
## Number of Fisher Scoring iterations: 3
```

Nakon što smo proveli logističku regresiju možemo analizirati dobivene rezultate.

Vidimo kako je koeficijent MaxHR jednak -0.83 što sugerira da se s povećanjem maksimalnog broja otkucaja srca smanjuje vjerojatnost imanja srčane bolesti. Koeficijent uz Age je 0.3566 što sugerira da starije osobe s većom vjerojatnošću imaju srčanu bolest. Dobiveni rezultati slažu se s vizualizacijom podataka koju smo napravili ranije.

P-vrijednosti su male: $2 * 10^{-16}$ za MaxHR i $8.55 * 10^{-6}$ za Age, iz čega možemo zaključiti kako oba koeficijenta imaju statistički značajan utjecaj.

Dodati ćemo matricu konfuzije da pomogne u interpretaciji točnosti modela.

```
# matrica konfuzije
confusionMatrix <- table(data$HeartDisease, predict(logisticModel, type = "response") > 0.5)
rownames(confusionMatrix) <- c("FALSE", "TRUE")
confusionMatrix
```

```
##
##          FALSE TRUE
## FALSE    246  164
## TRUE     122  386
```

Iz nje računamo:

```
accuracy = sum(diag(confusionMatrix)) / sum(confusionMatrix)
precision = confusionMatrix[2,2] / sum(confusionMatrix[,2])
recall = confusionMatrix[2,2] / sum(confusionMatrix[2,])
specificity = confusionMatrix[1,1] / sum(confusionMatrix[,1])

cat("Točnost: ", accuracy, "\n")
```

```
## Točnost: 0.6884532
```

```
cat("Preciznost: ", precision, "\n")
```

```
## Preciznost: 0.7018182
```

```
cat("Odziv: ", recall, "\n")
```

```
## Odziv: 0.7598425
```

```
cat("Specifičnost: ", specificity, "\n")
```

```
## Specifičnost: 0.6684783
```

Međutim, te mjere ovise o realnom broju iz intervala $<0, 1>$ koji smo postavili za graničnu vrijednost izlaza iz logističke regresije za dvije kategorije (u našem slučaju je to 0.5). Možemo promatrati i vrijednost pseudo-Rsquared, koja nam daje informaciju koliko je naš procijenjeni model blizu null modelu (tj. modelu koji koristi samo slobodni član) i ne ovisi o odabranome pragu.

```
Rsq = 1 - logisticModel$deviance/logisticModel$null.deviance  
Rsq
```

```
## [1] 0.1422405
```

Vidimo da je model s ovim regresorima malo prikladniji našim podacima od null modela.

Predikcija

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
```

```
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
## cov, smooth, var
```

```
# Predikcije
```

```
predictions <- predict(logisticModel, type = "response")
```

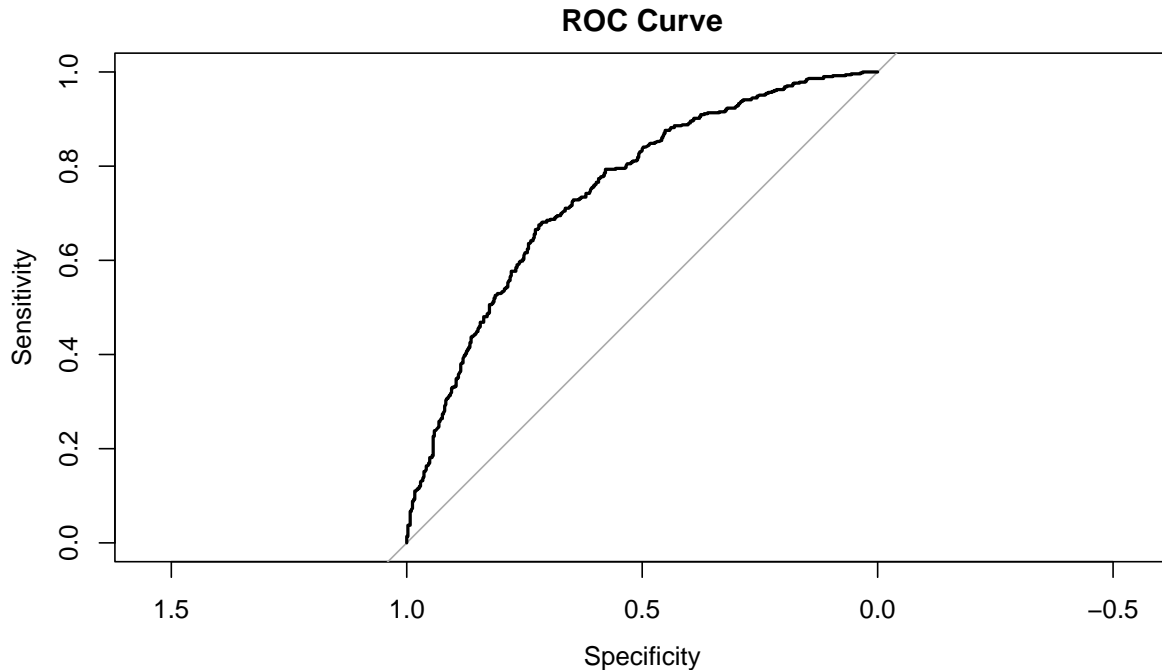
```
# ROC krivulja
```

```
roc_curve <- roc(data$HeartDisease, predictions)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(roc_curve, main = "ROC Curve")
```



```
# AUC vrijednost  
auc(roc_curve)
```

```
## Area under the curve: 0.7467
```

Tzv. “area under the curve” predstavlja površinu ispod ROC krivulje. Znamo kako je ROC krivulja jednostavno pravac pod kutem od 45 stupnjeva ako je predikcija potpuno slučajna. Što je veća površina ispod grafa, bolja je sposobnost predikcije. Idealno je AUC jednak 1.

Za naš slučaj AUC je 0.7467, iz čega zaključujemo kako naš model ima dobru sposobnost predikcije. Poboljšanje rezultata eventualno bismo mogli postići tako da bismo u predikciju dodali još relevantnih značajki.

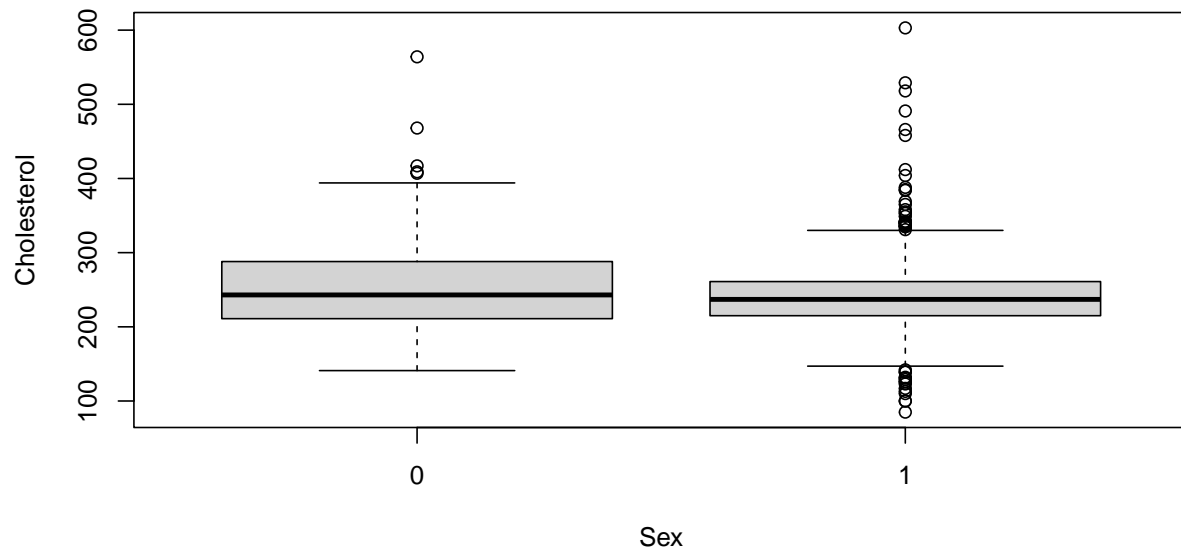
4. pitanje: Postoji li razlika u razini kolesterola između muškaraca i žena?

```
male <- data[data$Sex == 1,]  
female <- data[data$Sex == 0,]  
  
cat("Broj žena: ", nrow(female),  
    "\nBroj muškaraca: ", nrow(male))
```

```
## Broj žena: 193  
## Broj muškaraca: 725
```



```
boxplot(Cholesterol ~ Sex, data = data)
```

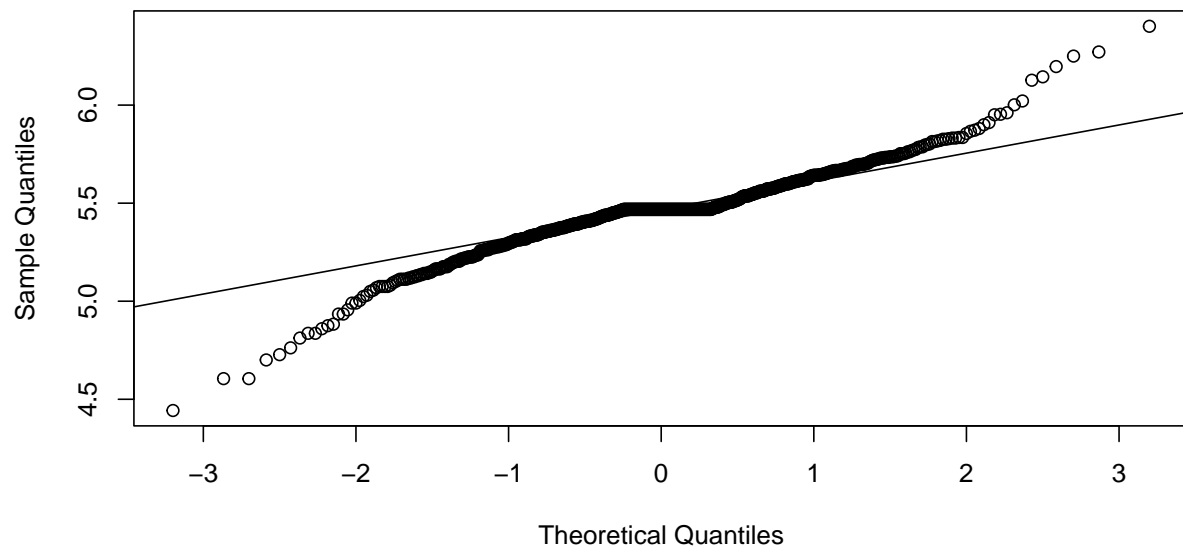


Ovaj box plot prikazuje da nema velike razlike u medijanima razine kolesterola po spolu, međutim, primjećujemo veliki broj stršećih vrijednosti među muškarcima.

Kao i u prvom istraživačkom pitanju koristimo T-test za dva uzorka. Opet ćemo prvo provjeriti normalnost podataka i nakon toga varijanci pomoću F-testa. Opet pretpostavljamo nezavisnost podataka jer su podatci uzeti iz različitih skupina.

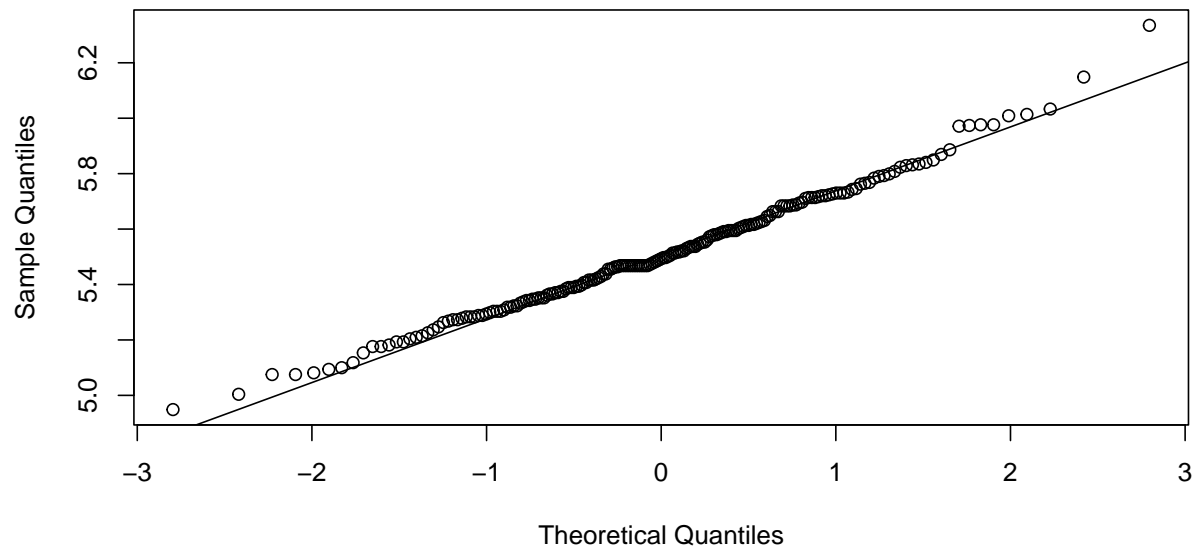
```
# crtamo Q-Q plot da vidimo odgovaraju li podatci normalnoj raspodjeli  
cholesterolMale <- log(male$Cholesterol)  
cholesterolFemale <- log(female$Cholesterol)  
  
qqnorm(cholesterolMale, main = "Q-Q plot for males")  
qqline(cholesterolMale)
```

Q-Q plot for males

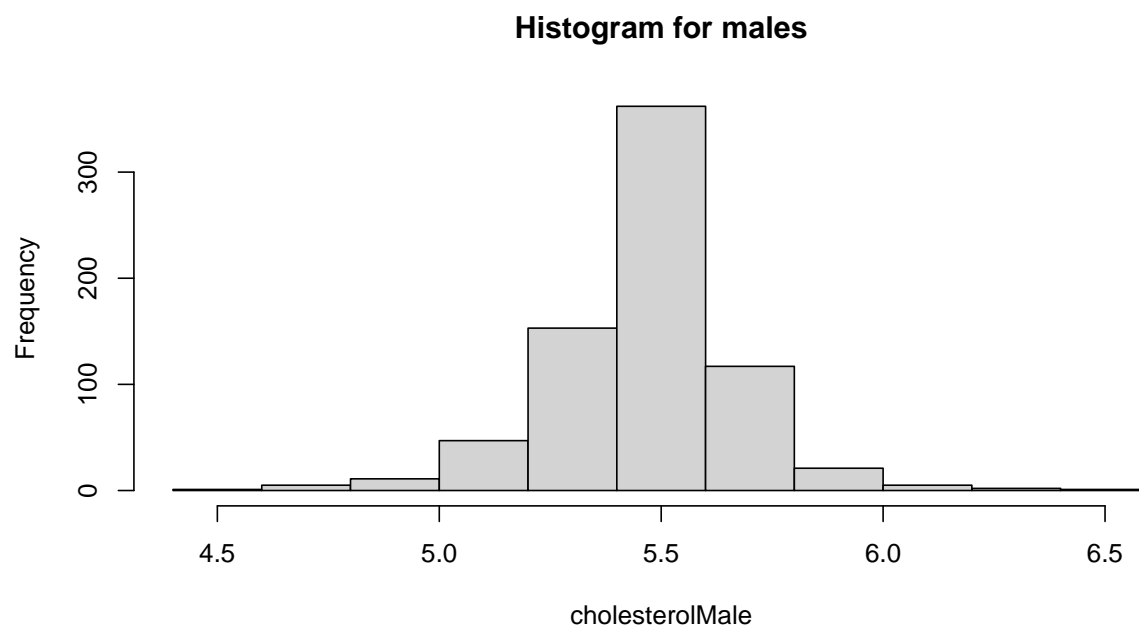


```
qqnorm(cholesterolFemale , main = "Q-Q plot for females")
qqline(cholesterolFemale)
```

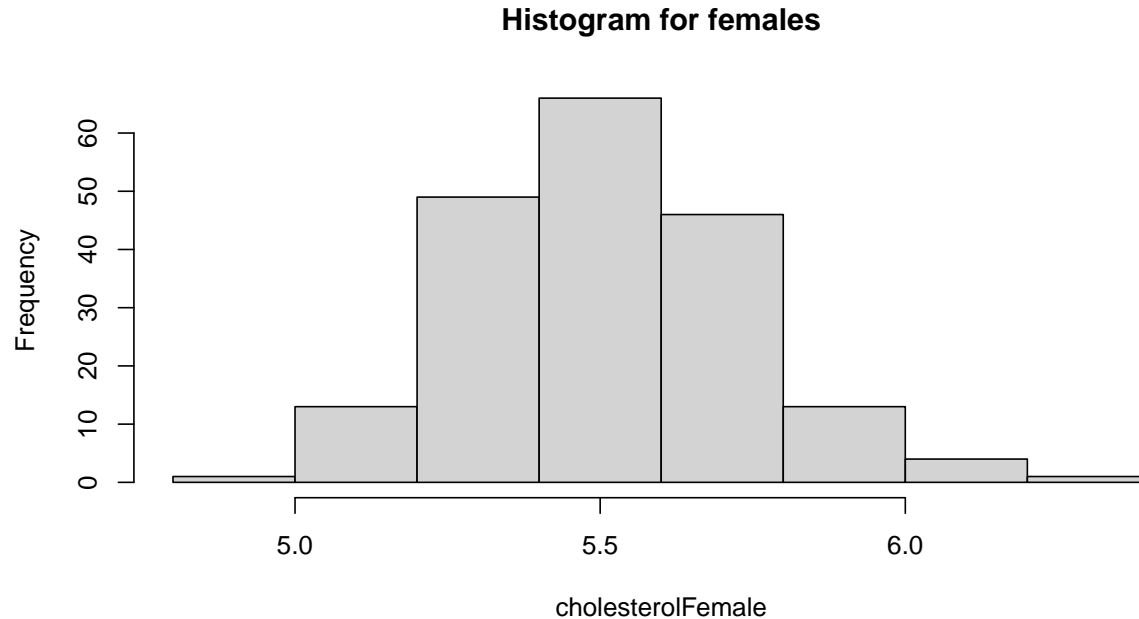
Q-Q plot for females



```
# histogrami
hist(cholesterolMale, main = "Histogram for males")
```



```
hist(cholesterolFemale, main = "Histogram for females")
```



Koristeći logaritamsku funkciju za transformaciju dobivamo bolje poklapanje s normalnom. Vidimo kako se podatci (Q-Q plot za muškarce) ne ravnaju savršeno po normalnoj razdiobi na repovima, ali iz histograma je sličnost dovoljno velika da bi mogli nastaviti s provođenjem testa (mala odstupanja od normalne razdiobe neće utjecati na rezultate zato što je T-test robusan na normalnost podataka) i dovoljno dobro funkcionira za “zvonolike” distribucije.

Nakon što smo utvrdili nezavisnost i normalnost podataka, provodimo F-test kako bismo utvrdili jesu li

varijance jednake ili različite. Sav tekst i pretpostavke iz prvog zadatka vezane za F-test možemo prenijeti i na ovaj zadatak. Sada ćemo radi jednostavnosti samo napisati hipoteze (nezavisnost i normalnost smo provjerili i pretpostavili)

Hipoteza H0: varijance su jednake (omjer varijanci je 1) Hipoteza H1: varijance su različite

```
var(cholesterolMale)
```

```
## [1] 0.04269296
```

```
var(cholesterolFemale)
```

```
## [1] 0.05115543
```

```
var.test(cholesterolMale, cholesterolFemale)
```

```
##
## F test to compare two variances
##
## data:  cholesterolMale and cholesterolFemale
## F = 0.83457, num df = 724, denom df = 192, p-value = 0.1044
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6611057 1.0379219
## sample estimates:
## ratio of variances
##      0.8345733
```

Opet vidimo kako je p-vrijednost veća od $\alpha = 5\%$ pa ne možemo odbaciti hipotezu da su varijance jednake. Zato koristimo T-test s jednakim i nepoznatim varijancama.

Hipoteze za T-test: H0: nema razlike u maksimalnom broju otkucaja srca između muškaraca i žena H1: postoji razlika u maksimalnom broju otkucaja srca između muškaraca i žena

```
t.test(cholesterolMale, cholesterolFemale, var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data:  cholesterolMale and cholesterolFemale
## t = -3.1513, df = 916, p-value = 0.001678
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08734506 -0.02030357
## sample estimates:
## mean of x mean of y
##  5.460091  5.513915
```

Iz rezultata vidimo kako je p-vrijednost manja od $\alpha = 5\%$ pa zaključujemo da postoji razlika u razinama kolesterola između muškaraca i žena. Korištenje parametarskog T-testa na distribucije koje odstupaju od normalne povećavaju vjerojatnost pogreške prvog reda. Također, iz boxplota smo uočili mnogo outliera. Iz tih razloga se odlučujemo provesti i neparametarsku inačicu T-testa, Wilcoxonov test predznačnih rangova

koji testira jednakost medijana, što je centralna mjera koja bolja podnosi stršće vrijednosti od aritmetičke sredine. Pretpostavke su da je distribucija simetrična i kontinuirana. Iz grafova možemo vidjeti da distribucije zadovoljavaju pretpostavke Wilcoxonovog testa.

Hipoteza H0: medijani razina kolesterola među spolovima su isti Hipoteza H1: medijani razina kolesterola među spolovima nisu isti

```
wilcox.test(cholesterolMale, cholesterolFemale, paired = FALSE, alternative = "two.sided")

##
## Wilcoxon rank sum test with continuity correction
##
## data: cholesterolMale and cholesterolFemale
## W = 61539, p-value = 0.009808
## alternative hypothesis: true location shift is not equal to 0
```

Vidimo da je p-vrijednost manja od $\alpha = 5\%$ te zaključujemo da podaci sugeriraju da postoji razlika između vrijednosti medijana razine kolesterola po spolovima.

5. pitanje: Postoji li razlika u udjelima pojavnosti srčanih bolesti među spolovima?

Kako bismo provjerili postoji li razlika u udjelu srčanih bolesnika među muškarcima i ženama, radimo test homogenosti.

H0: spol i pojava srčanih bolesti su nezavisni H1: postoji zavisnost spola i pojave srčanih bolesti

```
data_copy = data.frame(data);
tracemem(data)==tracemem(data_copy);
```

```
## [1] FALSE
```

```
untracemem(data_copy);
```

Računamo opažene i očekivane frekvencije te na njima provodimo Chi-kvadrat test. To je neparametarski test, čija je pretpostavka da su očekivane frekvencije svakog para razreda barem 5.

```
tbl_heart_disease = table(data_copy$Sex,
                           data_copy$HeartDisease)
added_margins_tbl_heart_disease = addmargins(tbl_heart_disease)

cat("Opažene frekvencije:")
```

```
## Opažene frekvencije:
```

```
print(added_margins_tbl_heart_disease)
```

```
##
##      0    1 Sum
## 0   143  50 193
## 1   267 458 725
## Sum 410 508 918
```

```

for (col_names in colnames(added_margins_tbl_heart_disease)){
  for (row_names in rownames(added_margins_tbl_heart_disease)){
    if (!(row_names == 'Sum' | col_names == 'Sum')){
      cat('Očekivane frekvencije za razred ', col_names, '-', row_names, ': ',
          (added_margins_tbl_heart_disease[row_names, 'Sum']
           * added_margins_tbl_heart_disease['Sum', col_names])
          / added_margins_tbl_heart_disease['Sum', 'Sum'], '\n')
    }
  }
}

```

```

## Očekivane frekvencije za razred 0 - 0 : 86.19826
## Očekivane frekvencije za razred 0 - 1 : 323.8017
## Očekivane frekvencije za razred 1 - 0 : 106.8017
## Očekivane frekvencije za razred 1 - 1 : 401.1983

```

Pretpostavke su zadovoljene pa nastavljamo s testom.

Hipoteza H0: očekivane vrijednosti su jednake opaženima ($o_{ij} = e_{ij}$ za i, j iz $\{0, 1\}^2$) Hipoteza H1: za neki par očekivanih i opaženih vrijednosti vrijedi da nisu jednaki

```
chisq.test(tbl_heart_disease, correct=F)
```

```

##
## Pearson's Chi-squared test
##
## data:  tbl_heart_disease
## X-squared = 85.646, df = 1, p-value < 2.2e-16

```

Vidimo da je p-vrijednost manja od $\alpha = 5\%$ što znači da odbacujemo hipotezu H0 u korist H1 koja tvrdi da za neki par očekivanih i opaženih vrijednosti vrijedi da nisu jednake, odnosno da postoji zavisnost između pojave srčanih bolesti i spola.