Final Project

Can Yildiz, Manjima Banerjee

DUE: April 19 @ PM

```
library(tidyverse)
library(ggplot2)
library(gridExtra)
library(latex2exp)
library(boot)
```

```
heart_attack <- read.csv("heart_attack.csv") # Load in the heart attack data
glimpse(heart_attack)
```

```
## Rows: 303
## Columns: 14
## $ age      <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5~
## $ sex      <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1~
## $ cp       <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0~
## $ trestbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1~
## $ chol     <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~
## $ fbs      <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
## $ restecg  <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1~
## $ thalach  <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~
## $ exang    <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
## $ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0~
## $ slope    <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1~
## $ ca       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0~
## $ thal     <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~
## $ target   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
summary(heart_attack)
```

```
##       age             sex               cp           trestbps
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0
##       chol            fbs            restecg          thalach
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##      exang           oldpeak         slope             ca
```

```
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000
##       thal           target
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000
##  Median :2.000   Median :1.0000
##  Mean   :2.314   Mean   :0.5446
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
```

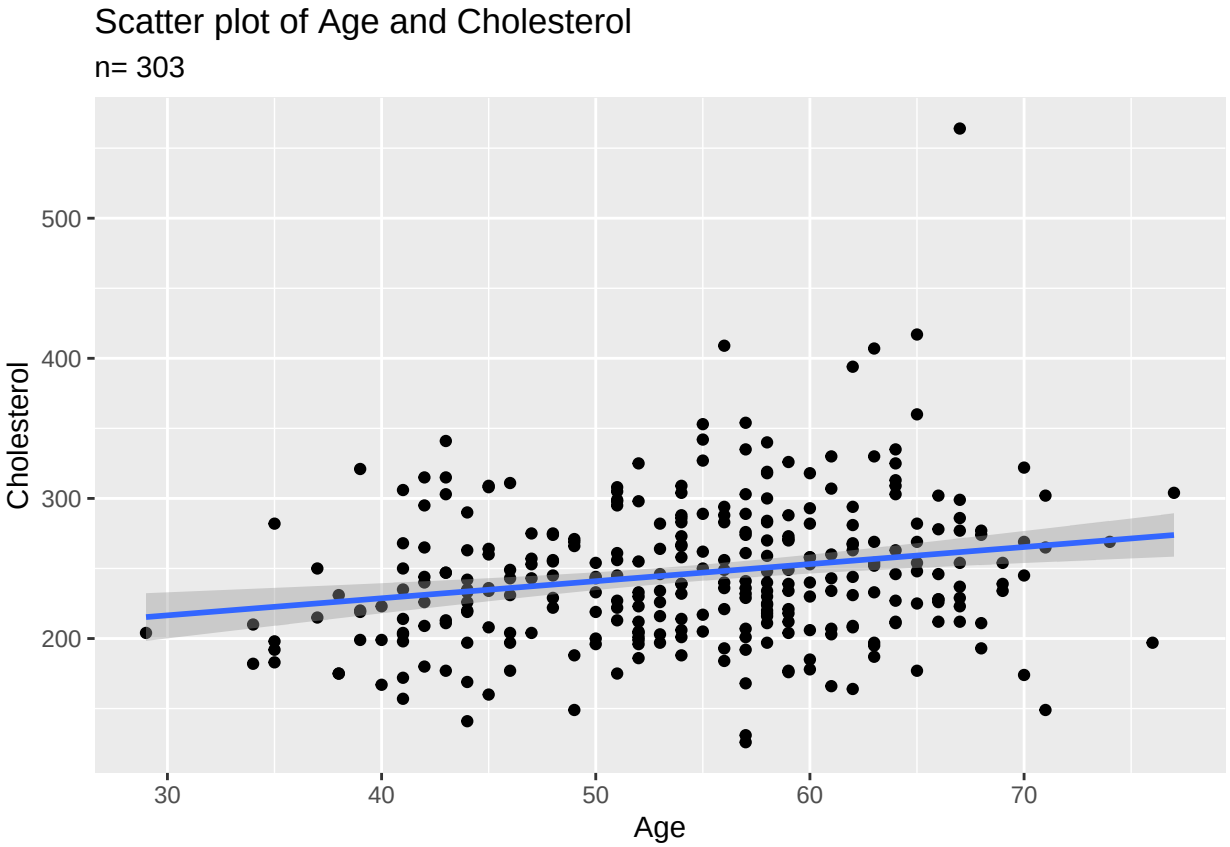
```
# Data cleaning
chol<- heart_attack[complete.cases(heart_attack$chol),]
age <- heart_attack[complete.cases(heart_attack$age),]
```

```
#1. Numerical summary of our Simple regression analysis
lmHeart=lm(age~chol, data= heart_attack)
summary(lmHeart)
```

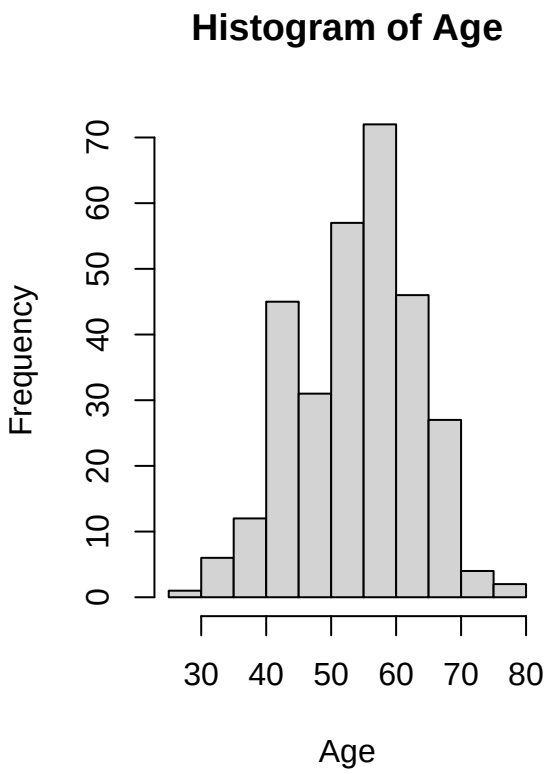
```
##
## Call:
## lm(formula = age ~ chol, data = heart_attack)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.7839  -6.4734   0.4782   6.3221  23.4782
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.145729   2.482848  18.183  < 2e-16 ***
## chol         0.037442   0.009867   3.795 0.000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.887 on 301 degrees of freedom
## Multiple R-squared:  0.04566,    Adjusted R-squared:  0.04249
## F-statistic:  14.4 on 1 and 301 DF,  p-value: 0.0001786
```

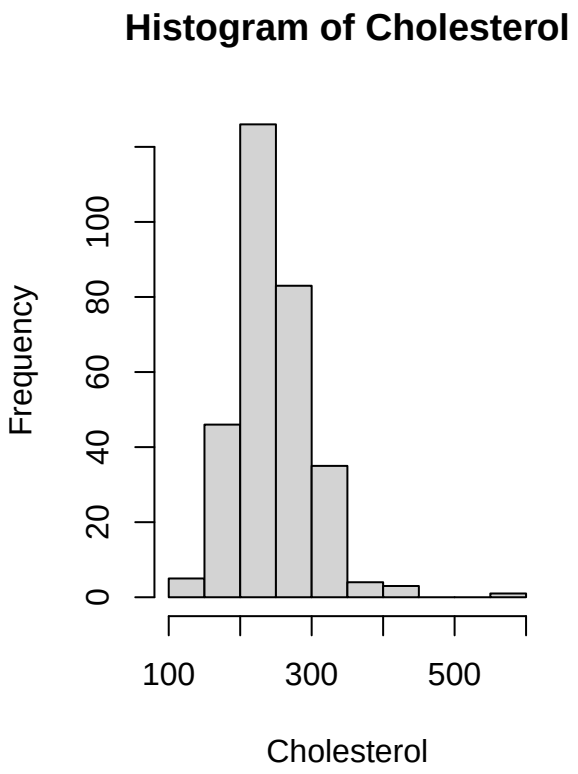
```
# 2.Simple Linear Regression Analysis for Age and Cholesterol Levels.
ggplot(heart_attack) +
geom_point(aes(x= age, y= chol) ,
na.rm=T)+
labs(x='Age',y='Cholesterol',
title='Scatter plot of Age and Cholesterol', subtitle= "n= 303") +
geom_smooth(method=lm, aes(x=age, y= chol))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

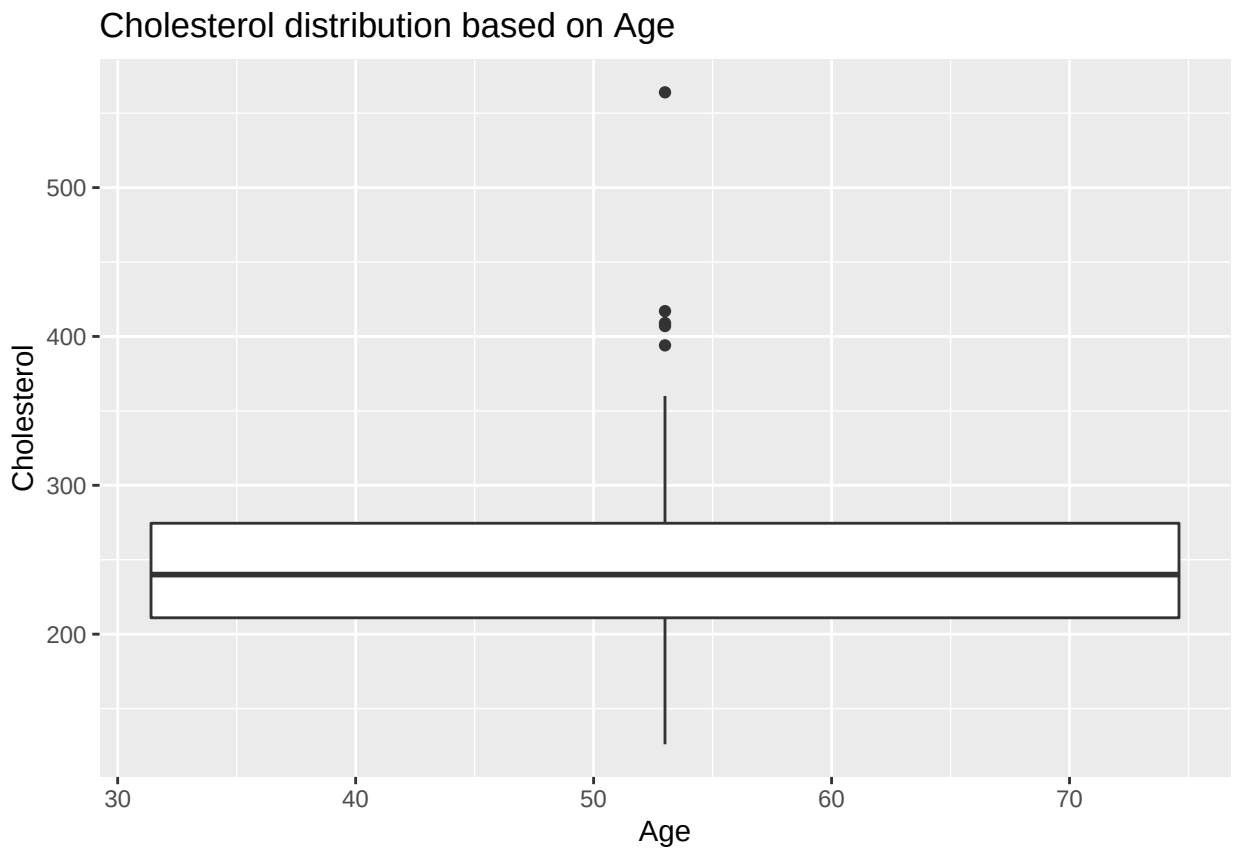



```
# Histogram of Age and Cholesterol Levels.
par(mfrow=c(1,2))
hgA <- hist(x=heart_attack$age,,xlab= "Age", main= "Histogram of Age")
hgC<-hist(x=heart_attack$chol, xlab= "Cholesterol",
main= "Histogram of Cholesterol" )
```

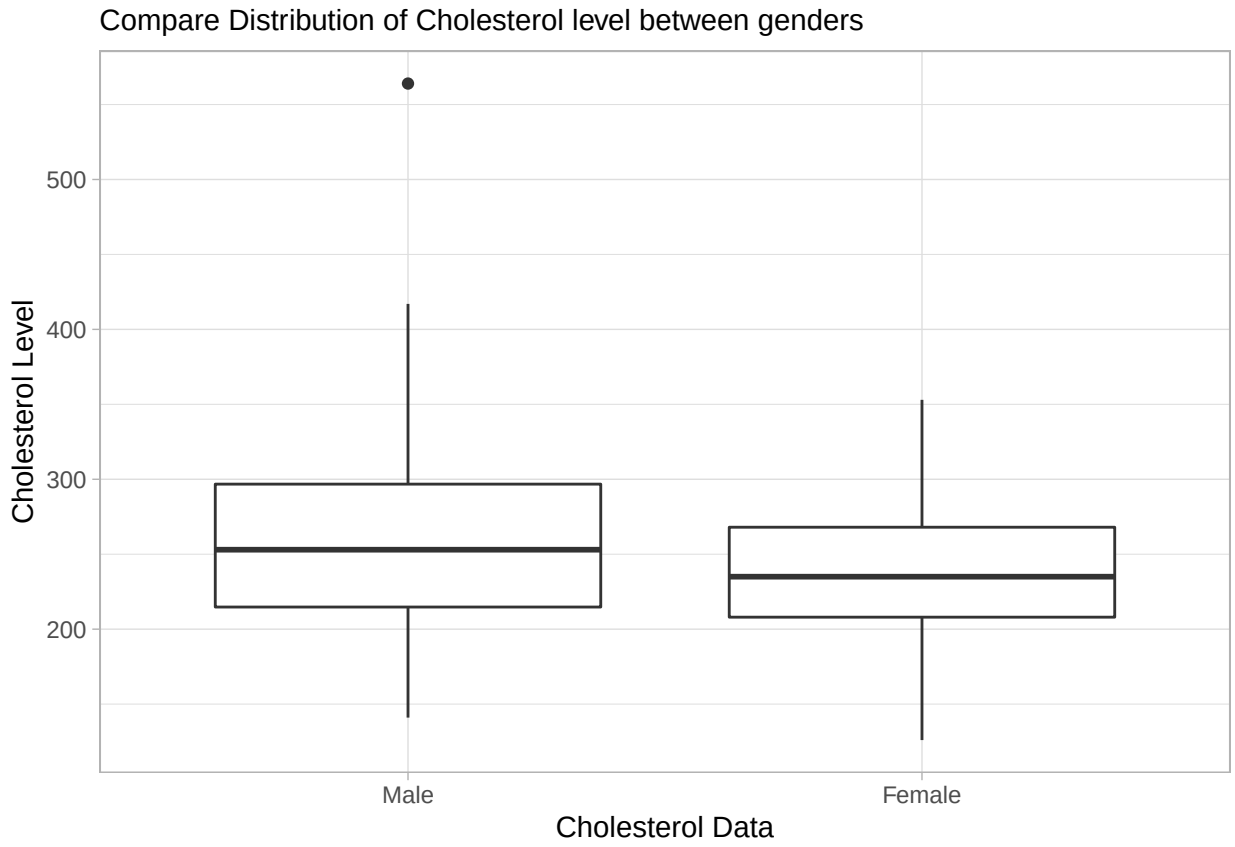




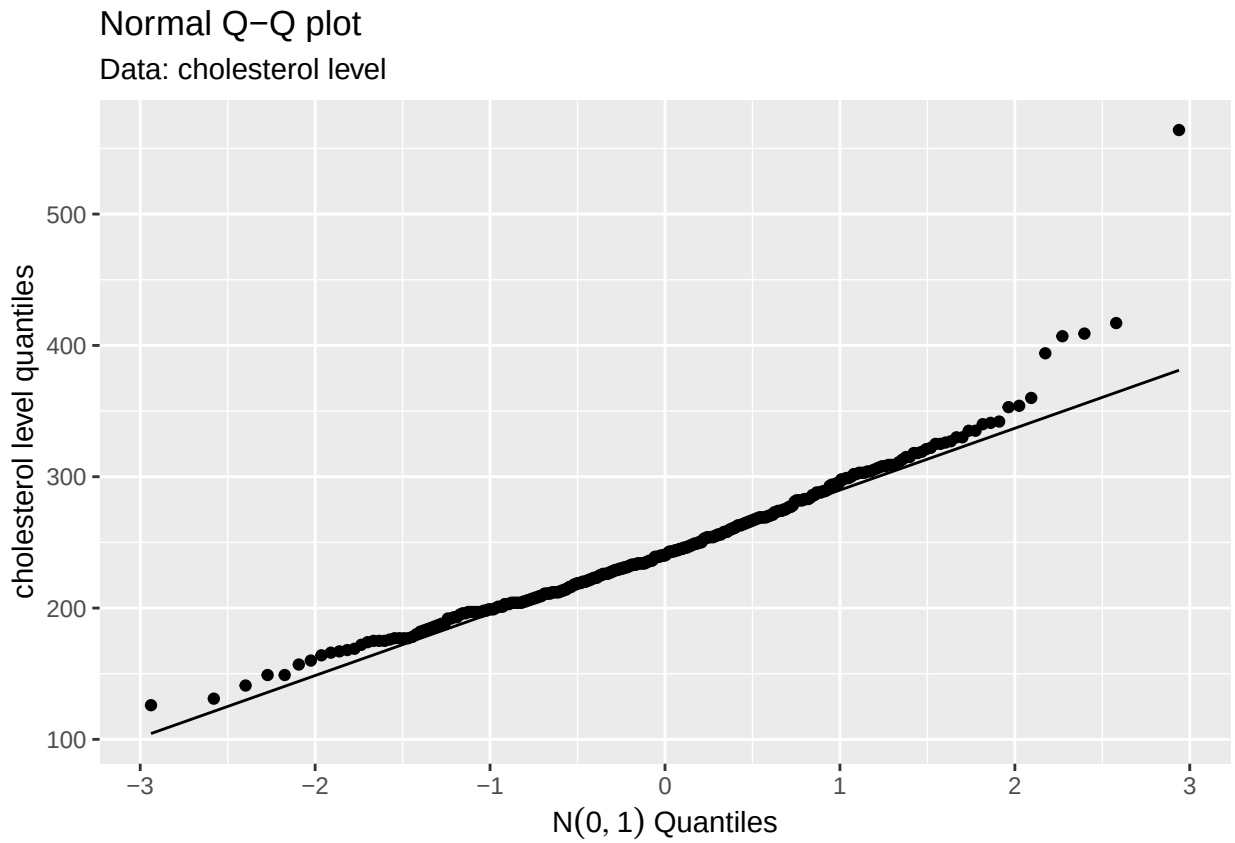

```
#3. Boxplot for Cholesterol levels and Age of Patients
ggplot(heart_attack)+ geom_boxplot(aes(x=age, y=chol,group=1)) +
  labs(x= "Age", y= "Cholesterol",
title= "Cholesterol distribution based on Age")
```




```
#4.Side by Side Boxplots for Male and Female Cholesterol levels
# Create vectors containing the observed data for gender
index.sex <- heart_attack$sex == "1"
obs.male <- heart_attack$chol[index.sex]
obs.female <- heart_attack$chol[!index.sex]


dat <- tibble(genders=c(heart_attack$chol[index.sex],
heart_attack$chol[!index.sex]),type = c(rep('male', 207), rep('female', 96)))

ggplot(dat, aes(x= type, y= genders)) +
geom_boxplot() + labs(x ="Cholesterol Data",
y = "Cholesterol Level",
subtitle = "Compare Distribution of Cholesterol level between genders") +
scale_x_discrete(labels = c ( "Male","Female"))+ theme_light()
```




```
#5. Normal QQ Plot for Distribution of Cholesterol Levels

x_bar <- mean(heart_attack$chol)
heart_attack %>%
ggplot(aes(sample = chol))+
geom_qq()+
geom_qq_line()+
labs(x = TeX(r'($N(0,1)$ Quantiles)'),
y = "cholesterol level quantiles",
title = "Normal Q-Q plot",
subtitle = "Data: cholesterol level")
```




```
t.test(obs.female,obs.male) #Two-Sample T-test with Unequal Variance
```

```
##
##  Welch Two Sample t-test
##
## data:  obs.female and obs.male
## t = 3.0244, df = 134.39, p-value = 0.002985
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.617474 36.406982
## sample estimates:
## mean of x mean of y
##  261.3021  239.2899
```

```
#Bootstrap confidence interval for the difference in means
# Create vectors containing the observed data for gender
index.sex <- heart_attack$sex == "1"
obs.male <- heart_attack$chol[index.sex]
obs.female <- heart_attack$chol[!index.sex]

set.seed(238)
B <- 5000
boot.mean.diff <- c()

for(b in 1:B){
# Bootstrap sample for each gender
boot.male <- sample(obs.male, replace = TRUE)
```

```r
boot.female <- sample(obs.female, replace = TRUE)
# Compute difference in bootstrap means
boot.mean.diff[b] <- mean(boot.male) - mean(boot.female)
}
```

```r
#Calculating the  95% confidence intervals
ci.mean.diff <- quantile(boot.mean.diff, probs = c(0.025, 0.975))
ci.mean.diff
```

```
##       2.5%      97.5%
## -36.465485  -8.799838
```