FINAL PROJECT

Introduction and Research Question:

Using our data, we will be analyzing the following questions - **Does the level of Cholesterol increase with age of patients who have had a heart attack? Do male patients have a higher cholesterol level than the female patients?**
The data for our project is taken from all the patients who suffered from a heart attack. It gives us detailed information on the counts for their integral health factors such as Cholesterol, Resting Blood Pressure, Heart Rate, Fasting Blood Sugar, Chest Pain Type and many others. These factors are strongly linked to an individual's pulmonary system and therefore help us check tendency of an individual having a heart attack
The variables that will be key in our analysis are the Cholesterol Level (mg/dL) of the patient, the Gender(Male/Female) and his/her Age (years) as these would best help to draw any conclusions based on our research question.

EDA and Methodology:

In order to solve our first question,'**Does the level of Cholesterol increase with age of patients who have had a heart attack?'**, we perform simple linear regression using two variables - The Cholesterol level(mg/dL) and the respective age of patients(years). The regression line is gently sloped showing a **moderate positive correlation** between Cholesterol and age of a patient who had a heart attack. This means that as the age of an individual increases, the cholesterol level also increases, informing us that elderly people need to be more cautious of their health as they have a higher tendency of having an attack. Moreover, we can also see from our numerical summary that the slope of our regression line is 0.037mg/dL which suggests that with every year in age increasing, the level of cholesterol increases by 0.037mg/dL. We see that the data is extrapolated as when age is 0 years, the cholesterol level is still 45.14mg/dL (y-intercept) which is not possible and therefore is not included in our study. The equation of our linear regression line becomes - $y = 0.037x + 45.145$
Using our boxplot for the age and cholesterol level, we also see that 75% of the patients have a cholesterol level ranging from 220mg/dL to 275mg/dL approximately which suggests that these levels of cholesterol may be alarming and leading to a heart attack if not reduced.

For our second research question,'**Do male patients have a higher cholesterol level than the female patients?'**, we take two population groups for males and females that have mean cholesterol level of males (μ_{males}) and mean cholesterol level of females ($\mu_{females}$) respectively. In order to compare which group has a higher level, we take their respective cholesterol means. Our null hypothesis (H_0: default statement that is assumed for the test) is -

$$H_0 : \mu_{males} = \mu_{females}$$

And our Alternative Hypothesis (H_A: This is the opposite of null hypothesis and is true when we reject null hypothesis) is -

$$H_A : \mu_{males} \neq \mu_{females}$$

Methodology 1:

For our first methodology, we used Statistical inference using hypothesis testing for two population mean groups. For this method, we first found out whether our data satisfies the normality assumption (normal distribution of data) and this is done using a Normal QQ Plot. We see that most of our points lie on or very close to the line with slight deviations on the tails that occur due to basic variability of data. Therefore, we may say that our data is normally distributed.

The next step is to find if both the population groups have equal variances or not. Through this, we will be able to derive evidence for or against our null hypothesis. In order to find information about their variances, we made side by side boxplots for cholesterol levels for males and females, and compared the length of the box (interquartile range) for both the plots. We could see that both the plots vary in length suggesting an unequal assumption to be made.

In order to find our inference, we use Welch's degrees of freedom and T-Distribution that helps us collect our evidence. This is because we have the normality distribution approved and the unequal variances assumption, and when these conditions are satisfied, we use a T- Distribution. Degrees of freedom tells the number of independent values in our data (134.39), our t-test statistic value gives us the value of difference relative to the variation in our sample data (larger t value gives greater evidence against null hypothesis) and our p value which is the the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, also helps us to accept or reject our null hypothesis using evidence.

Methodology 2:

For our second method, we will be using Statistical Inference using Confidence Intervals for two population mean groups of cholesterol which includes sampling from both the groups multiple times and then finding the difference in the means of each of the sample numbers. Here, we will change our assumed hypothesis slightly. The null hypothesis assumed in this approach would be as follows -

$$H_0 : \mu_{males} - \mu_{females} = 0$$

Which shows that the difference between the mean cholesterol levels for both the population is 0, signifying that they are both equal.

The alternative hypothesis is -

$$H_A : \mu_{males} - \mu_{females} \neq 0$$

Which shows that the difference between the mean cholesterol levels for both the population is not equal to 0, signifying that they are both unequal.

For this, we first find indexes for both males and females and then take 5000 bootstrap samples from both the groups respectively with replacement after creating an empty vector to store these sampled values. For each bootstrapped sample, we find the mean of both groups and store its difference in the empty vector made before.
After we get all our values, we find the probability of 0.975 and 0.025 from our mean differences to create a confidence interval for 95%. The difference in means is called centered means and gives us a range of values for the difference in the means of the cholesterol levels of the two groups.

Findings:

1. From our Hypothesis testing with two mean groups, we see that the mean value of males is 239.2899mg/dL which is smaller than the mean value of the female group of cholesterol that is 261.3021mg/dL. In addition to this, we have a t-test statistic of -3.0244 which is quite large suggesting that we have strong evidence against the null hypothesis. Moreover, we also have a significantly small p value of 0.0029, almost equal to zero. This small p-value gives us very strong evidence to reject our null hypothesis - H_0: $\mu_{males} = \mu_{females}$ .Therefore, due to these findings we may infer that the cholesterol level of males is not equal to the cholesterol level of female patients.

2. Using our second methodology of Statistical inference with confidence intervals, we see that the range found for the difference of our mean cholesterol for the two groups lies between -36.465485mg/dL to -8.7998m/dL. The values found in our project are negative as the population used for (male group) had a smaller value compared to the second group. We are 95% confident that the difference in mean cholesterol levels between males and females is between -36.4654 and -8.79983 units. In this sample, the males have lower mean cholesterol levels than females. Based on this interval, we also conclude that there is statistically significant difference in mean cholesterol levels between males and females group, because the 95% confidence interval does not include the null value, zero. The confidence interval is a range of likely values for the difference in means. Since the interval does not contain zero (difference), we have sufficient evidence to conclude that there is a difference between both the population groups' cholesterol level. This supports our previous methodology for hypothesis testing for both the groups

Conclusions:

After our study for both the research questions, we have evidence for a few inferences from our methodologies. For the first research question, we see that the cholesterol level and age for patients who have had a heart attack are positively correlated, as their age increases, the cholesterol levels also increase. With every year in age increasing, the cholesterol level increases by 0.037mg/dL. This means that elders would have a higher cholesterol level, hence a higher tendency of having a heart attack, compared to individuals younger in age.

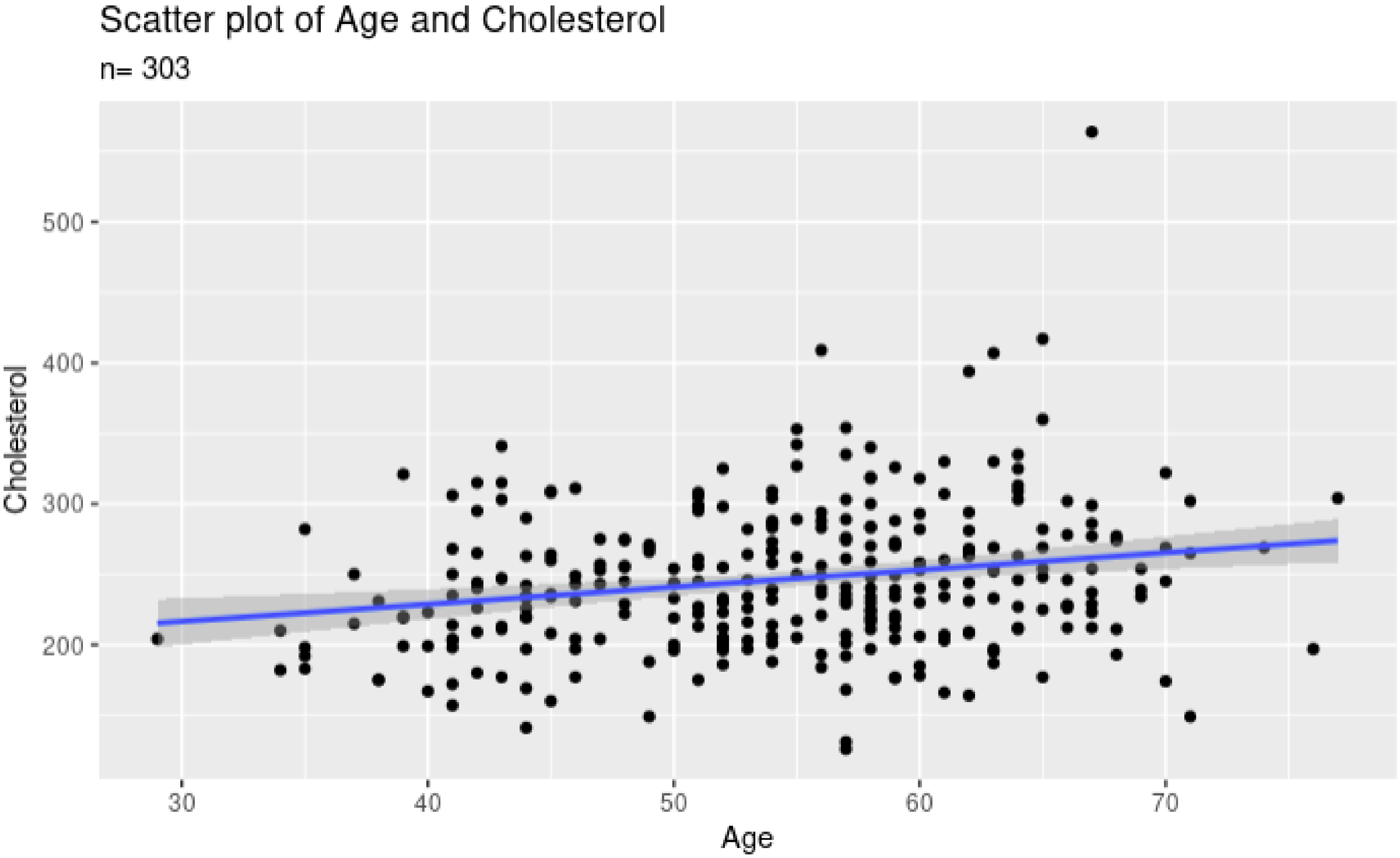
The two methodologies executed also show us that there is a difference in the mean levels of cholesterol for males and females as we see that our t test statistic is quite large and p value, close to zero, signifying that we have strong evidence against the null hypothesis that the means of males and females is equal. An issue that we faced with our second methodology related to bootstrapping confidence intervals was that empirical bootstrapping was used even though we were aware of the distribution of our data to be normal. However, we were able to find accurate results by using the nonparametric approach which is in fact used when we may not wish to guess the distribution of our data. To support our research question further, we also see that in our second methodology, we assumed that the difference between the means of both the population levels is 0, stating that they are both equal. However, after our bootstrapping confidence interval, we saw that as our null value 0 does not lie within our confidence interval, we may reject our null hypothesis and infer that both the population means are unequal. In addition to this, we also see that the value of mean of males - mean of females gives us a negative number, clearly stating that the mean level of cholesterol in males is smaller than the mean level of cholesterol in males. Therefore, we have an answer to our second research question that the male patients do not have a higher cholesterol level than the female patients. Through this research question, it might suggest that females have a higher tendency of having a heart attack than males as their cholesterol levels are higher.

Graphs and Numerical Summaries:

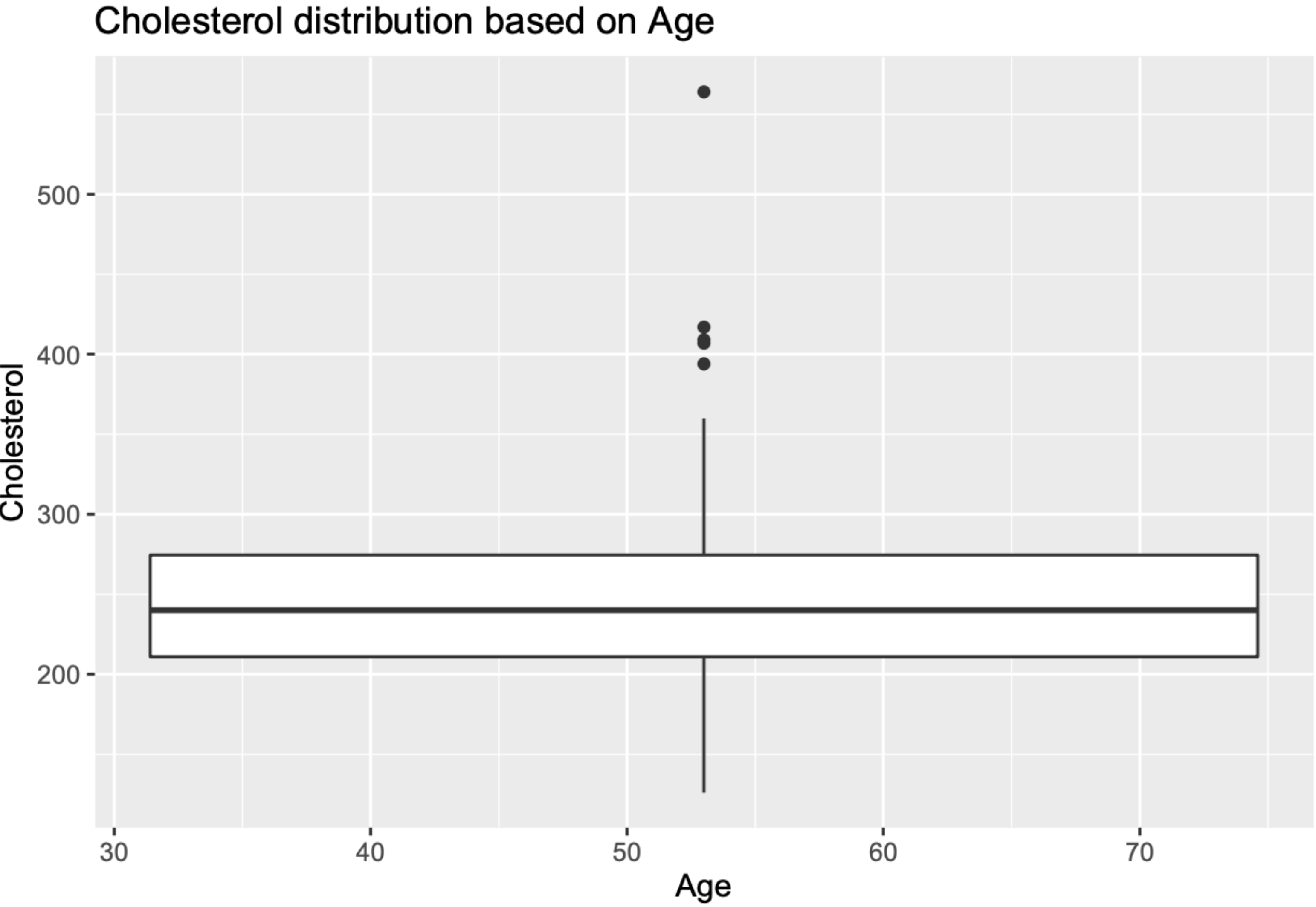
1. Numerical summary of our Simple regression analysis

```
##
## 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
```

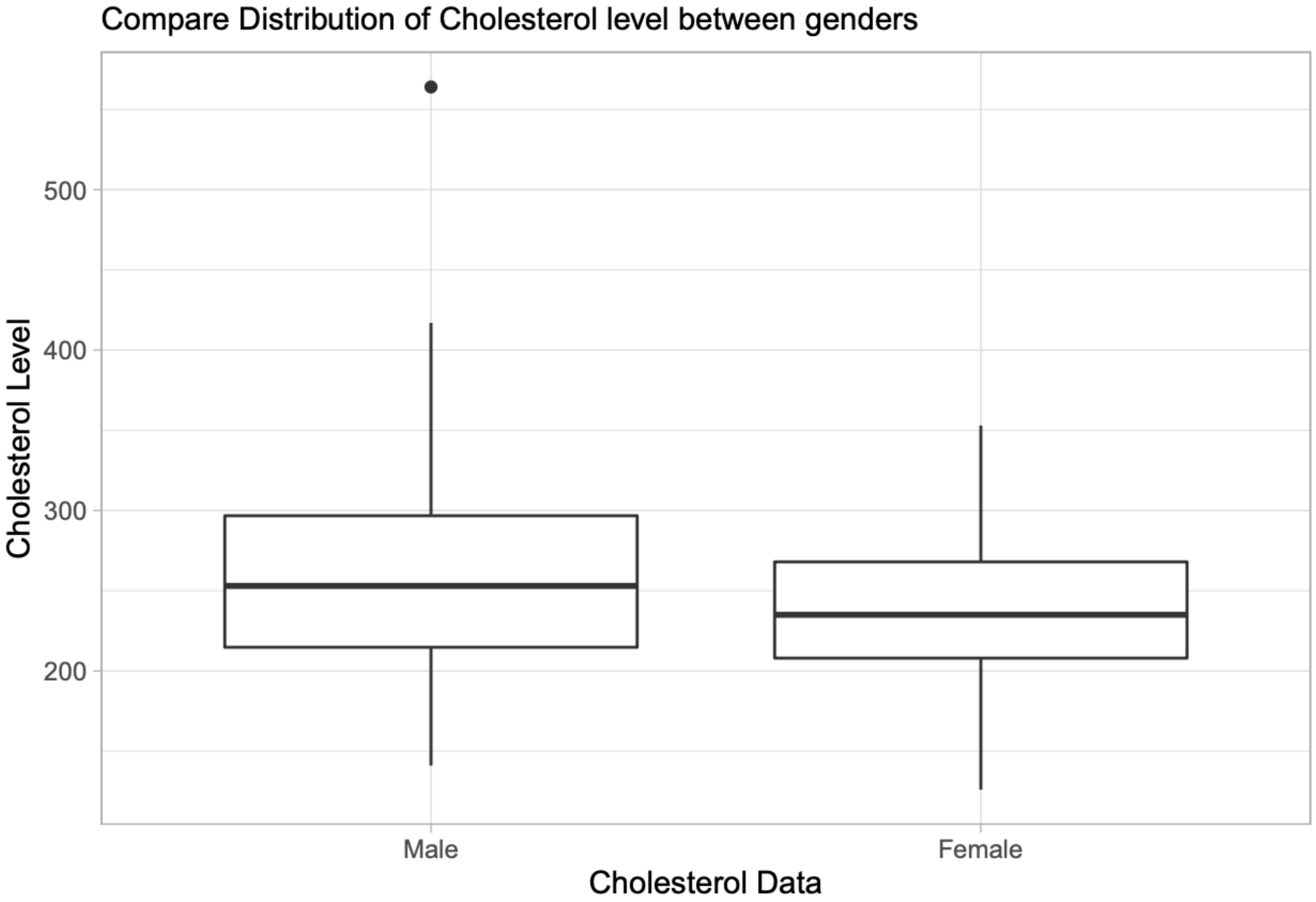
2. Linear Regression Analysis for Age and Cholesterol levels



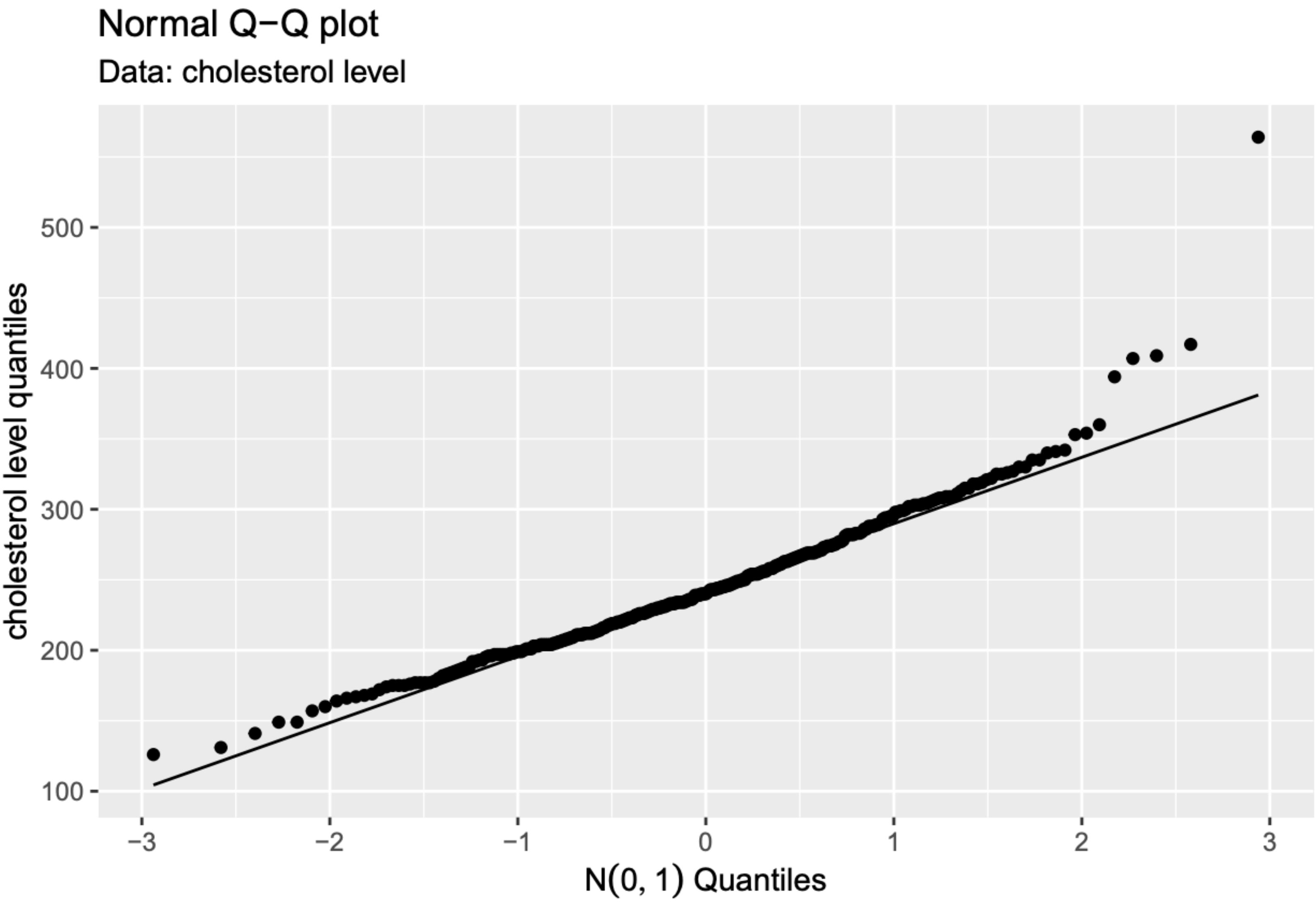
3. **Boxplot for Cholesterol levels and Age of Patient**



4. **Side by Side Boxplots for Male and Female Cholesterol levels**



5. Normal QQ Plot for Distribution of Cholesterol Levels



Appendix -
Data source link - https://www.kaggle.com/datasets/pritsheta/heart-attack
Prit Sheta, 2021-09-26. Heart Attack Dataset (Version 1)

REFERENCES:

Estimating the Difference in Two Population Means. Lumen. (n.d.). Retrieved April 16, 2022, from https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/estimating-the-difference-in-two-population-means/

1.3.5.3. Two-sample t-test for Equal means. (n.d.). Retrieved April 17, 2022, from https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm

2.1 - What is Simple Linear Regression? PennState Eberly College of Science. (n.d.). Retrieved April 19, 2022, from https://online.stat.psu.edu/stat462/node/91/