R Project 1

```
ClevelandHD <- read.csv("C:/Users/saket/Downloads/ClevelandHD.csv")
HungaryHD <- read.csv("C:/Users/saket/Downloads/HungaryHD.csv")
```

1. How many of the 303 patients were diagnosed with heart disease?

```
sum(ClevelandHD$presenceHD==1)
```

```
[1] 139
```

2. Make a histogram and a qqplot of the data in the "maxheartrate" column.

```
qqnorm(ClevelandHD$maxheartrate)
qqline(ClevelandHD$maxheartrate)
```

3. For this column, report the sample mean, median, sample standard deviation, and 95% confidence interval for the population mean.

```
mean(ClevelandHD$maxheartrate)
```

```
median(ClevelandHD$maxheartrate)
```

```
sd(ClevelandHD$maxheartrate)
```

```
t.test(ClevelandHD$maxheartrate)
```

```
    One Sample t-test

data:  ClevelandHD$maxheartrate
t = 113.84, df = 302, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 147.0212 152.1933
sample estimates:
mean of x
 149.6073
```

4. Make a side-by-side boxplot of the maximum heart rate for those patients diagnosed with heart disease vs. those patients not diagnosed with heart disease.

5. Do a t-test to test whether the maximum heart rate is significantly different for those patients diagnosed with heart disease vs. those patients not diagnosed with heart disease. Report both the p-value and the 95% confidence interval for the difference of the population means. Clearly state your conclusion.

```
heart.data.withDisease <- ClevelandHD[ClevelandHD$presenceHD==1,]
heart.data.withoutDisease <- ClevelandHD[ClevelandHD$presenceHD==0,]
t.test(heart.data.withDisease$maxheartrate, heart.data.withoutDisease$maxheartrate)
```

```
    Welch Two Sample t-test

data:  heart.data.withDisease$maxheartrate and heart.data.withoutDisease$maxheartrate
t = -7.8579, df = 272.27, p-value = 9.106e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23.90912 -14.32900
sample estimates:
mean of x mean of y
  139.259   158.378
```

6. Do a Wilcoxon-Mann-Whitney test to test whether the maximum heart rate is significantly different for those patients diagnosed with heart disease vs. those patients not diagnosed with heart disease. Report the p-value and clearly state your conclusion. Do the conclusions of the t-test and the Wilcoxon-Mann-Whitney test agree? Discuss.

```
wilcox.test(heart.data.withDisease$maxheartrate, heart.data.withoutDisease$maxheartrate)
```

```
    Wilcoxon rank sum test with continuity correction

data:  heart.data.withDisease$maxheartrate and heart.data.withoutDisease$maxheartrate
W = 5806.5, p-value = 1.861e-13
alternative hypothesis: true location shift is not equal to 0
```

The conlcusions do not change between t-test and Wilcoxon test. The QQplot shows the maxheartrate might not be normally distributed and as such t-test might not be the most appropriate test here. Wilcoxon-test is always valid (no assumptions on the distribution). Though in this case, applying a t-test also results in a pvalue of the same order as the wilcoxon test, ideally you would choose only one test to test your hypothesis. It might be more approrpriate to choose wiclxon test.

7. Make a histogram and a qqplot of the data in the "chol" column.

```
hist(ClevelandHD$chol)
```

8. For this column, report the sample mean, median, sample standard deviation, and 95% confidence interval for the population mean.

```
mean(ClevelandHD$chol)
```

```
median(ClevelandHD$chol)
```

```
sd(ClevelandHD$chol)
```

```
t.test(ClevelandHD$chol)
```

9. Make a side-by-side boxplot of the cholesterol values for those patients diagnosed with heart disease vs. those patients not diagnosed with heart disease.

```
boxplot(chol ~ presenceHD, data=ClevelandHD)
```

10. Do a t-test to test whether cholesterol is significantly different for those patients diagnosed with heart disease vs. those patients not diagnosed with heart disease. Report both the p-value and the 95% confidence interval for the difference of the population means. Clearly state your conclusion.

Check for normality:

```
qqnorm(ClevelandHD$chol)
qqline(ClevelandHD$chol)
```

```
t.test(heart.data.withDisease$chol, heart.data.withoutDisease$chol)
```

```
    Welch Two Sample t-test

data:  heart.data.withDisease$chol and heart.data.withoutDisease$chol
t = 1.4924, df = 298.64, p-value = 0.1366
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.815018 20.484170
sample estimates:
mean of x mean of y
 251.4748  242.6402
```

11. Do a Wilcoxon-Mann-Whitney test to test whether cholesterol is significantly different for those patients diagnosed with heart disease vs. those patients not diagnosed with heart disease. Report the p-value and clearly state your conclusion. Do the conclusions of the t-test and the Wilcoxon-Mann-Whitney test agree? Discuss.

```
wilcox.test(heart.data.withDisease$chol, heart.data.withoutDisease$chol)
```

```
    Wilcoxon rank sum test with continuity correction

data:  heart.data.withDisease$chol and heart.data.withoutDisease$chol
W = 12998, p-value = 0.03536
alternative hypothesis: true location shift is not equal to 0
```

In this case, the conclusion between t-test and wicoxon test are different (reject null if using wilcoxon but not if using a t-test.). Thought the wiclxon tedt pvalue is only border line. The data does appear to be normally distributed and as such a t-test might be more "powerful" in this setting. Ideally, you would have choosen just one test (t-test in this case having establsiehd the normality of chol column)

12. Propose your own question to answer with this dataset (for example, is the maximum heart rate significantly different for men and women?). Clearly state the question you are trying to answer, the test you performed, the results of the test, and your conclusion. If there are any relevant plots include them as well.

13. In the "Content/R stuff" section of Blackboard I have placed another file called "HungaryHD.csv". Download this dataset onto your computer, and import it to R. This is a similar dataset to the Cleveland dataset, except the patients are all from Hungary. Repeat the test you proposed in question #12 on this dataset. Do you reach the same conclusion? Discuss.