Question 1

A) I)

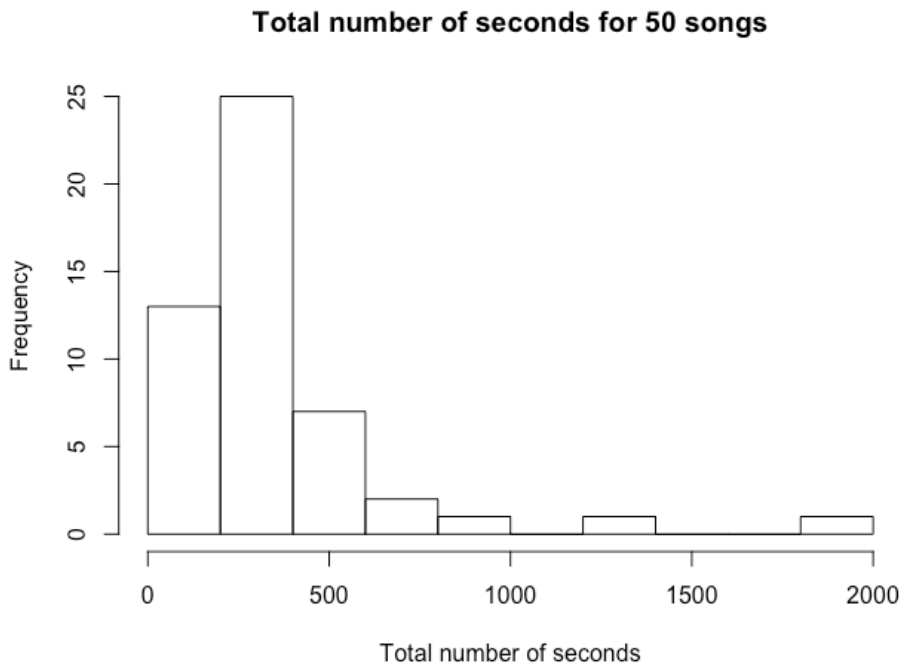

Figure 1: Histogram of total number of seconds for 50 songs that were randomly selected from Spotify for the songs dataset. This histogram shows that it is skewed toward the right tail end indicating that it is a right skewed histogram. It also shows that its measure of center is like to be in between 0 to 500 seconds due to the high level of frequency.

II) In terms of the shape of the histogram, it is right skewed because the distribution of the data is skewed towards the right tail end of the histogram. In terms of the location of the histogram, it is in between 0 to 500 seconds as most that is where most of the data are distributed. In terms of the spread of the histogram, it is actually quite narrow because 90% of the data or more are situated quite close to one another. In terms of outlier, there is a high possibility of that existing due to a small number of data located in the right tail end of the distribution that deviates from the overall data distribution.

B) I)

```
> t.test(Songs$total_secs, mu = 240)

        One Sample t-test

data:  Songs$total_secs
t = 2.62, df = 49, p-value = 0.01168
alternative hypothesis: true mean is not equal to 240
95 percent confidence interval:
 266.5823 441.6177
sample estimates:
mean of x
    354.1
```

Figure 2: This t-test output from R of total duration of 50 songs that were randomly selected from Spotify to test against the hypothesis of the true population mean of 240 seconds for the songs dataset. This output shows that the p-value is in the range of 0.01 - 0.05 indicating a moderate evidence against the null hypothesis.

II) The value of the one-sample t-test test statistic is 2.62.

III) The value of the one-sample t-test p-value is 0.012.

Iv) If null hypothesis were true, then the t-test has a t-distribution with a degree of freedom of 49.

$$t \sim t_{49}$$

V) I reject the null hypothesis because the p-value is smaller than 5% significance level.

C) We are 95% confident that the true population mean for the songs total number of seconds lies between 266.58 and 441.62 seconds.

D) I)

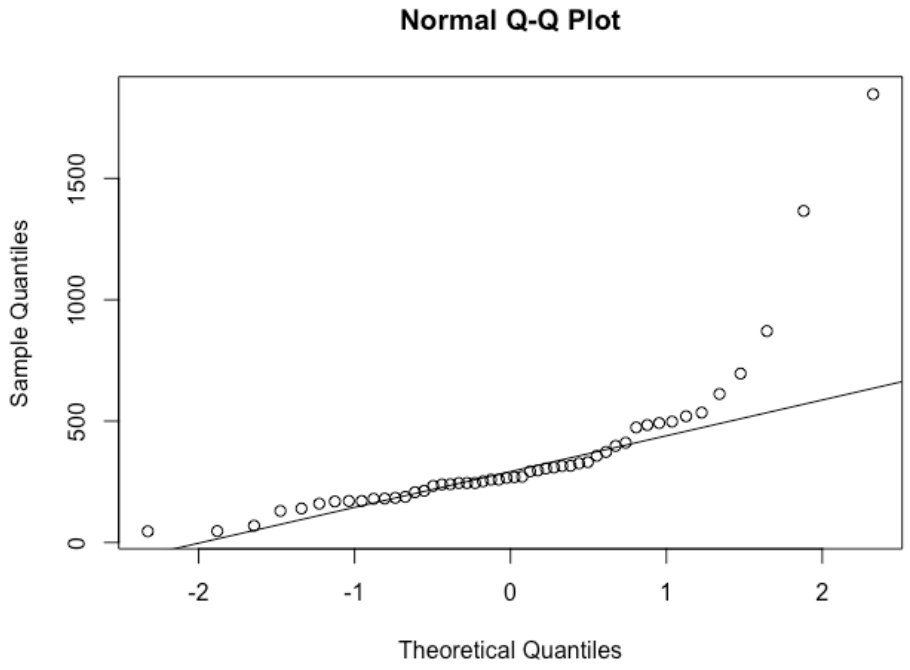

Figure 3: Normal Quantile-Quantile Plot total number of seconds for 50 songs that were randomly selected from Spotify for the songs dataset. It shows that some of the data are not along the line rather they are trailing upwards from the line.

II) Some of the data are skewed upwards away from the straight line which made the assumption of normality for the data not resounding. Therefore, the normality is not met for this dataset. But based on the central limit theorem, since most of the data falls on diagonal line, we can say that the data is approximately normal.

Question 2

A)

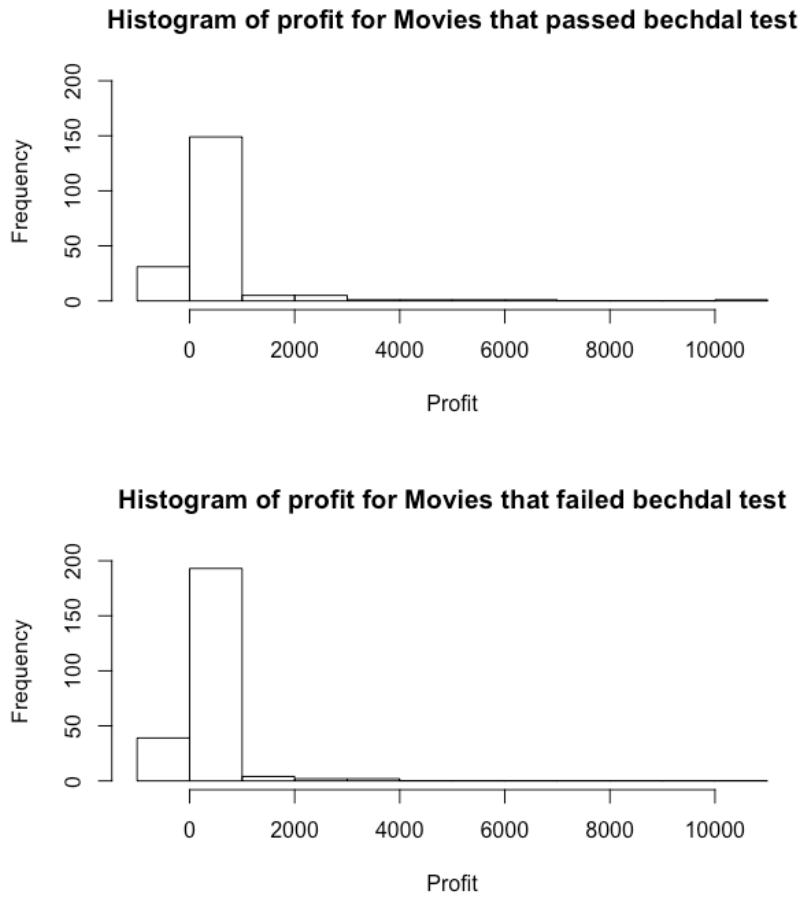

Figure 1: Two Histograms of profit for movies that passed or failed the bechdal test for the movies dataset. The first histogram shows that its highest frequency of movies that is between 0 to 1000 profit range is 150 movies and it also shows that its data are distributed beyond 10000-profit mark. The second histogram shows that its highest frequency of movies that is between the same range is greater than 150 movies and its data are distributed up to the 4000-profit mark.

- The shape of the distribution of profit for the movies that passed the bechdal test is right-skewed because its data trails to the right tail end of the dataset. The shape of the distribution of profit for the movies that passed the bechdal test is right-skewed as well due to the same reason as the first distribution.
- Both distributions are not normally distributed because they are not symmetric.

B) I) $H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

Where μ_1 is the mean of profit for movies that fail the bechdal test and μ_2 is the mean of profit for movies that pass the bechdal test.

II)

```
> t.test(profit~bechdal, data = Movies)

        Welch Two Sample t-test

data:  profit by bechdal
t = -2.6915, df = 248.22, p-value = 0.007596
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -386.23760  -59.82417
sample estimates:
mean in group FAIL mean in group PASS
          241.3306           464.3615
```

Figure 2: The two-sample t-test output from R of profit for movies that passed or failed the bechdal test for the movies dataset to test if there is any difference in mean profit between the movies that passed the bechdal test and failed the bechdal test. This output shows that the true difference in means is not equal to 0 indicating that there is a strong evidence against the null hypothesis.

II) The observed value of the test-statistic is –2.692.

III) If the null hypothesis is true, then the test-statistic will have a t-distribution with a degree of freedom 248.22.

Iv) The p-value is 0.0076.

V) Since the p-value is smaller than the 5% significance level, I will reject the null hypothesis and conclude that there is a very strong evidence that there is a statistically significant difference in the mean profit between the movies that failed or passed the bechdal test.

C) We are 95% confident that the true mean profit of the movies that passed the bechdal test is in between 59.824 to 386.238 dollars more than that of movies that failed the bechdal test.

D)

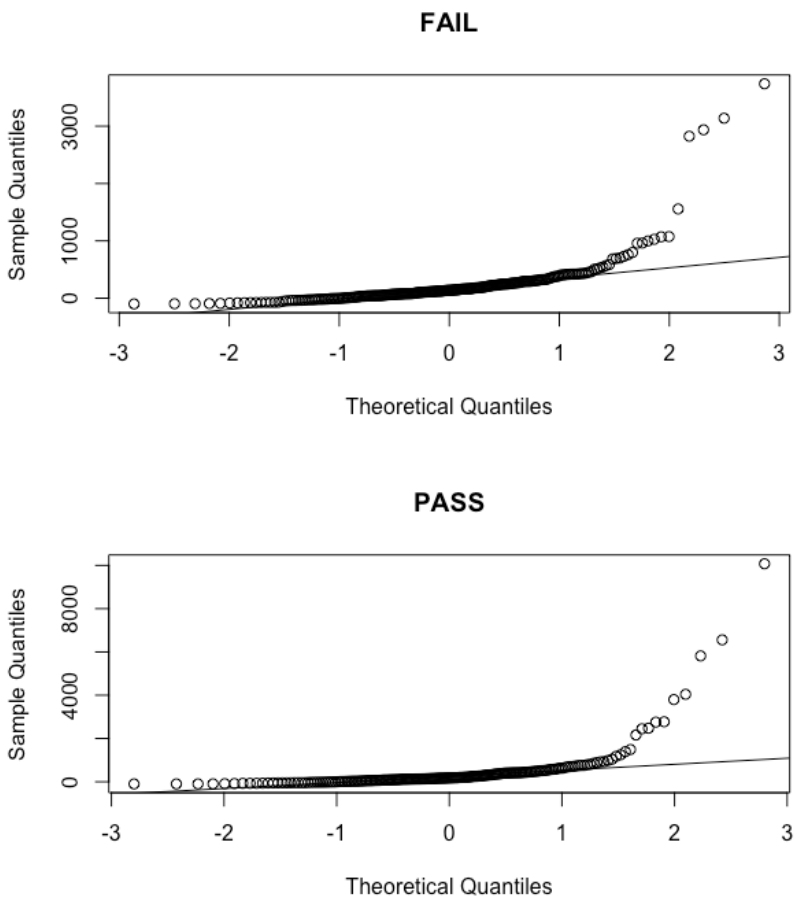

Figure 3: Two Normal Quantile-Quantile Plots of mean profit for the movies that failed the bechdal test and the movies that passed the bechdal test for the movies dataset. Both plots show that some of the data aren't falling on the diagonal line in both plots instead they diverge upwards into the right tail end.

Based on the QQ-plot, we do not see the observations on either group to fall along the diagonal line. Therefore the assumption of normality is not reasonable for both groups. However, there are more than 30 observations in each group so we can invoke the central limit theorem and say that we expect the sample means of both groups to be normally distributed.