IRIS Dataset:

R Project:

Graphical and Numerical Summaries:

Information on midSetosaSepalLength:

Mean:	4.87
Standard deviation(sd):	0.258414
Quartile 1(Q1):	4.8
Quartile 2(Q2):	4.9
Quartile 3(Q3):	5.1
Quartile 4(Q4):	5.1
IQR:	0.3
Minimum:	4.4
Maximum:	5.1

Information on midVersicolorSepalLength:

Mean:	6.02
Standard deviation(sd):	0.4871687
Quartile 1(Q1):	5.700
Quartile 2(Q2):	5.950
Quartile 3(Q3):	6.275
Quartile 4(Q4):	6.900
IQR:	0.575
Minimum:	5.4
Maximum:	6.9

From the output above, we can determine that:

Both groups are of symmetrical data as the median is approximately equivalent to Q2.

The data centres around 0.3 (IQR) for midSetosaSepalLength and the data centres around 0.575 (IQR) for midVersicolorSepalLength.

The overall spread of data for midSetosaSepalLength is 0.7 (Range) and the overall spread of data for midVersicolorSepalLength is 1.5, showing the variability of the data shown.

Check for Normality of Data:

midVersicolorSepalLength seems to be normally distributed as it only has two outliers.

In comparison, midSetosaSepalLength does not seem to be normally distributed as it has many outliers.

From looking at the histograms and boxplots, we can also see that midVersicolorSepalLength seems to be normally distributed and that midSetosaSepalLength does not seem to be normally distributed.

From the Shapiro-Wilk tests, we can see that the p-value of midSetosaSepalLength is low (0.03778), meaning that there is a low chance that midSetosaSepalLength is normally distributed. On the other hand, we can see the p value of midVersicolorSepalLength is high (0.668), meaning there is a high possibility that it is from normally distributed data.

Confident Intervals and Hypothesis Testing:

One Sample t-test:

Ho: The overall mean midterm score is equal or greater to 3.

Ha: The overall mean midterm score is less than 3.

From the output, we can see that the mean of the sample was 3.763333. The one-sided 99% confidence interval tells us that the mean is less than 4.552071. The p-value of 0.988 tells us that if the mean was 3, the probability of selecting a sample a mean less than or equal to this one would be 98%.

Since the p-value is not less than the significance of 0.01, we cannot reject the null hypothesis that the mean is equal to 3. This means there is no evidence that the sepal length is under 3.

```
> t.test(midsample$petal_length, mu=3, alternative = "less", conf.level = 0.99)

        One Sample t-test

data:  midsample$petal_length
t = 2.3827, df = 29, p-value = 0.988
alternative hypothesis: true mean is less than 3
99 percent confidence interval:
     -Inf 4.552071
sample estimates:
mean of x
 3.763333
```

F test to compare two variances:

Ho: There is no difference between the variances in each group.

Ha: That there is a difference between the variances in each group.

The null or true variance is equal to 1, the p-value is higher than 0.05 as the value is 0.07268. Therefore, we can assume that both sepal lengths have equal variances and we can reject the null hypothesis.

```
> var.test(midVersicolorSepalLength,midSetosaSepalLength)

        F test to compare two variances

data:  midVersicolorSepalLength and midSetosaSepalLength
F = 3.5541, num df = 9, denom df = 9, p-value = 0.07268
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.8827823 14.3086914
sample estimates:
ratio of variances
          3.554077
```

Welch Two Sample t-test:

Ho: That there is no difference between the means in each group.

Ha: That there is a difference between the means in each group.

We can see from the t-test, we can see from the estimated means that there is a difference between the two mean, the estimate mean of midVersicolorSepalLength is 6.02 and the estimate mean of midSetosaSepalLength is 4.87.

```
> t.test(midVersicolorSepalLength,midSetosaSepalLength)

        Welch Two Sample t-test

data:  midVersicolorSepalLength and midSetosaSepalLength
t = 6.5945, df = 13.693, p-value = 1.339e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7751873 1.5248127
sample estimates:
mean of x mean of y
     6.02      4.87
```

Wilcoxon rank sum test with continuity correction:

Ho: When the samples are small, the t-test requires the assumption that data is normally distributed.

Ha: When the samples are small, the t-test requires the assumption that data is not normally distributed.

As we can see from the Wilcoxon rank sum test, we can see that the data is not normally distributed, and the two groups do not have identical population as the p-value is less than .05.

```
> wilcox.test(midVersicolorSepalLength,midSetosaSepalLength)

        Wilcoxon rank sum test with continuity correction

data:  midVersicolorSepalLength and midSetosaSepalLength
W = 100, p-value = 0.0001678
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(midVersicolorSepalLength, midSetosaSepalLength) :
  cannot compute exact p-value with ties
```

Simple Linear Regression:

From the correlation coefficient and graph, we can see that there is a weak downhill (negative) linear relationship. It is of value (-0.1808).

The estimated coefficient of sepal_width is (-0.3890), meaning is a weak, downhill slope. From the summary, we can see that the p-value is greater than 0 (0.3388), meaning that we are confident that the value is quite different from 0.

From the graph, we can see that the regression line does not fit well to the graph as there are many outliers and no points touching the line. We can see the values are scattered away from the line. From the value R squared, we can see that the graph makes up for (0.03271) or 3% of the variance.

Analysis of Data:

We can see that there are two groups of symmetrical data, which have a wide overall spread and centres around the two medians.

We can see that the means are not equal to 3, that there are no differences in variances, that there is a difference in the means and that midSetosaSepalLength is not normally distributed and midVersicolorSepalLength is normally distributed.

From the graph, we can see the data is very scattered and that there is a weak downhill linear relationship, with the value r squared making up 3% of the variance.

The conclusion, beginning that the data is symmetrical, they have a weak relationship and that there is an overall spread of data. Also, midSetosaSepalLength is not normally distributed and midVersicolorSepalLength is normally distributed.