# Project Report and Analysis

*Name: Vipul Vishek and Paramveer Singh* *Subject: Data Science*

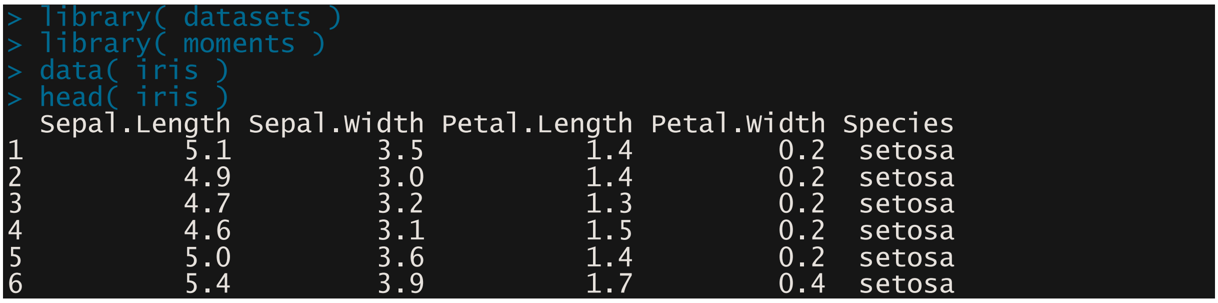
*Prm: 1032201185 and 1032202055* *Submitted to: Mrs. Arti Kharpade*

The code can be accessed *here*

This analysis is based on Iris Dataset. We have used various libraries, namely, datasets, moments, ggplot2, dplyr and have implemented various related functions related to it.

## About the Dataset

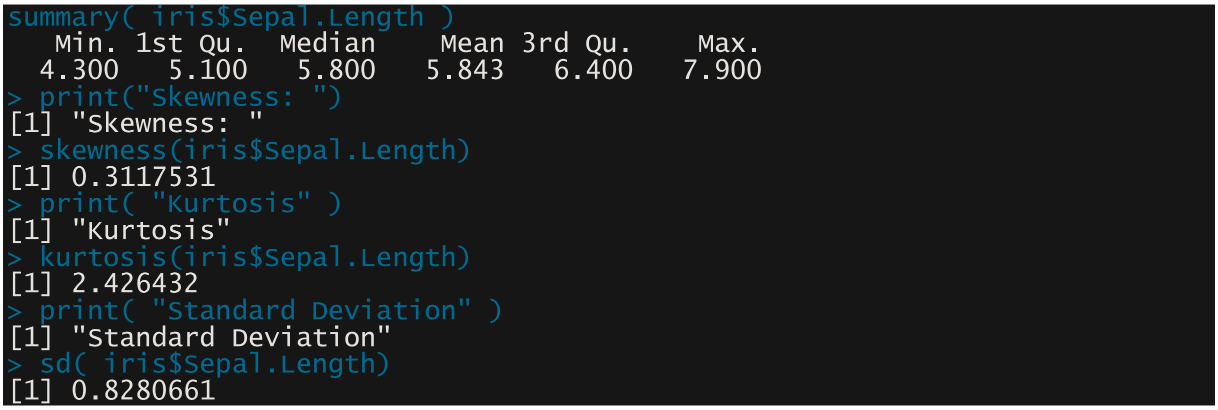
Iris dataset has shape of (150,5) that comprises of 4 Numerical and 1 Categorical column. The column names are Sepal Length , Sepal Width , Petal Length , Petal Width and Species.



```
> library( datasets )
> library( moments )
> data( iris )
> head( iris )
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

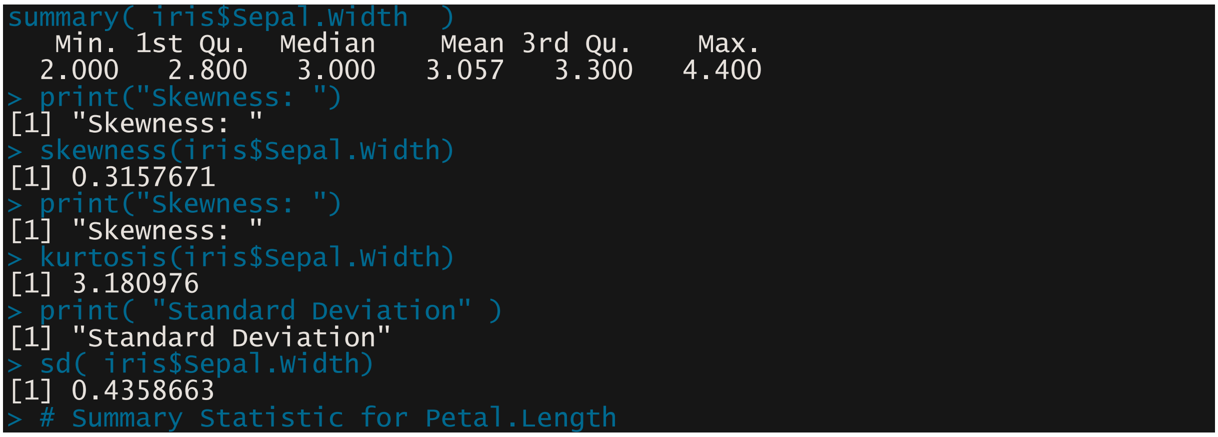
## Univariate Analysis

Sepal Length: The positively skewed feature centres around 5.843 with a spread from 5.100 to 6.400 (Q1 to Q3). It has a moderate variability (IQR), moderate kurtosis (2.426), and a standard deviation of 0.828. Most of the flowers have sepal length in the range of 5 and 7 units.



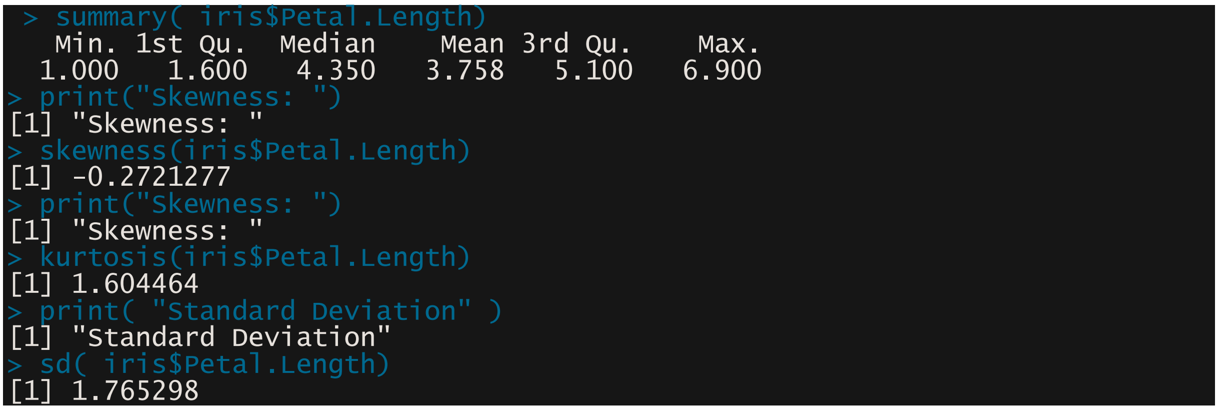
```
summary( iris$Sepal.Length )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.300   5.100   5.800   5.843   6.400   7.900 
> print("Skewness: ")
[1] "Skewness: "
> skewness(iris$Sepal.Length)
[1] 0.3117531
> print( "Kurtosis" )
[1] "Kurtosis"
> kurtosis(iris$Sepal.Length)
[1] 2.426432
> print( "Standard Deviation" )
[1] "Standard Deviation"
> sd( iris$Sepal.Length)
[1] 0.8280661
```

Sepal Width: The feature centres around 3.057 with a range from 2.000 to 4.400 (Q1 to Q3). It exhibits a positively skewed distribution (skewness = 0.316) with moderate kurtosis (3.181). The standard deviation is 0.436, indicating low variability. Hardly a few Species lies in the range of 2.75 and 3.25



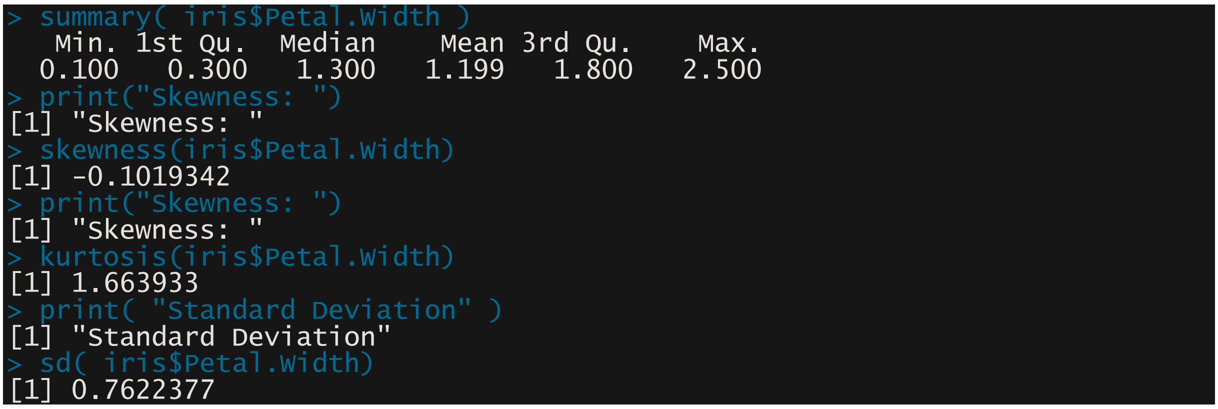
```
summary( iris$Sepal.Width  )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   2.800   3.000   3.057   3.300   4.400 
> print("Skewness: ")
[1] "Skewness: "
> skewness(iris$Sepal.Width)
[1] 0.3157671
> print("Skewness: ")
[1] "Skewness: "
> kurtosis(iris$Sepal.Width)
[1] 3.180976
> print( "Standard Deviation" )
[1] "Standard Deviation"
> sd( iris$Sepal.Width)
[1] 0.4358663
> # Summary Statistic for Petal.Length
```

Petal Length: The feature has a central tendency around 3.758, with a range from 1.000 to 6.900 (Q1 to Q3). It exhibits a negatively skewed distribution (skewness = -0.272) and moderate kurtosis (1.604). The standard deviation is 1.765, indicating moderate variability.



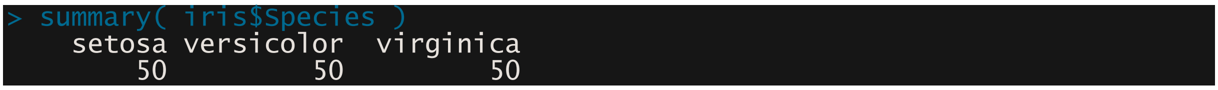
```
 > summary( iris$Petal.Length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.600   4.350   3.758   5.100   6.900 
> print("Skewness: ")
[1] "Skewness: "
> skewness(iris$Petal.Length)
[1] -0.2721277
> print("Skewness: ")
[1] "Skewness: "
> kurtosis(iris$Petal.Length)
[1] 1.604464
> print( "Standard Deviation" )
[1] "Standard Deviation"
> sd( iris$Petal.Length)
[1] 1.765298
```

Petal Width : The feature is centered around 1.199 with a range from 0.100 to 2.500 (Q1 to Q3). It shows a slightly negatively skewed distribution (skewness = -0.102) and moderate kurtosis (1.664). The standard deviation is 0.762, indicating moderate variability.



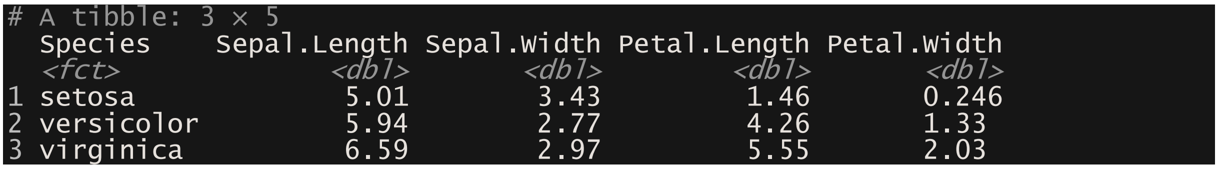
```
> summary( iris$Petal.Width )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.100   0.300   1.300   1.199   1.800   2.500 
> print("Skewness: ")
[1] "Skewness: "
> skewness(iris$Petal.Width)
[1] -0.1019342
> print("Skewness: ")
[1] "Skewness: "
> kurtosis(iris$Petal.Width)
[1] 1.663933
> print( "Standard Deviation" )
[1] "Standard Deviation"
> sd( iris$Petal.Width)
[1] 0.7622377
```

Species : There are in total 50 samples each of Setosa, Versicolor, Virginica.

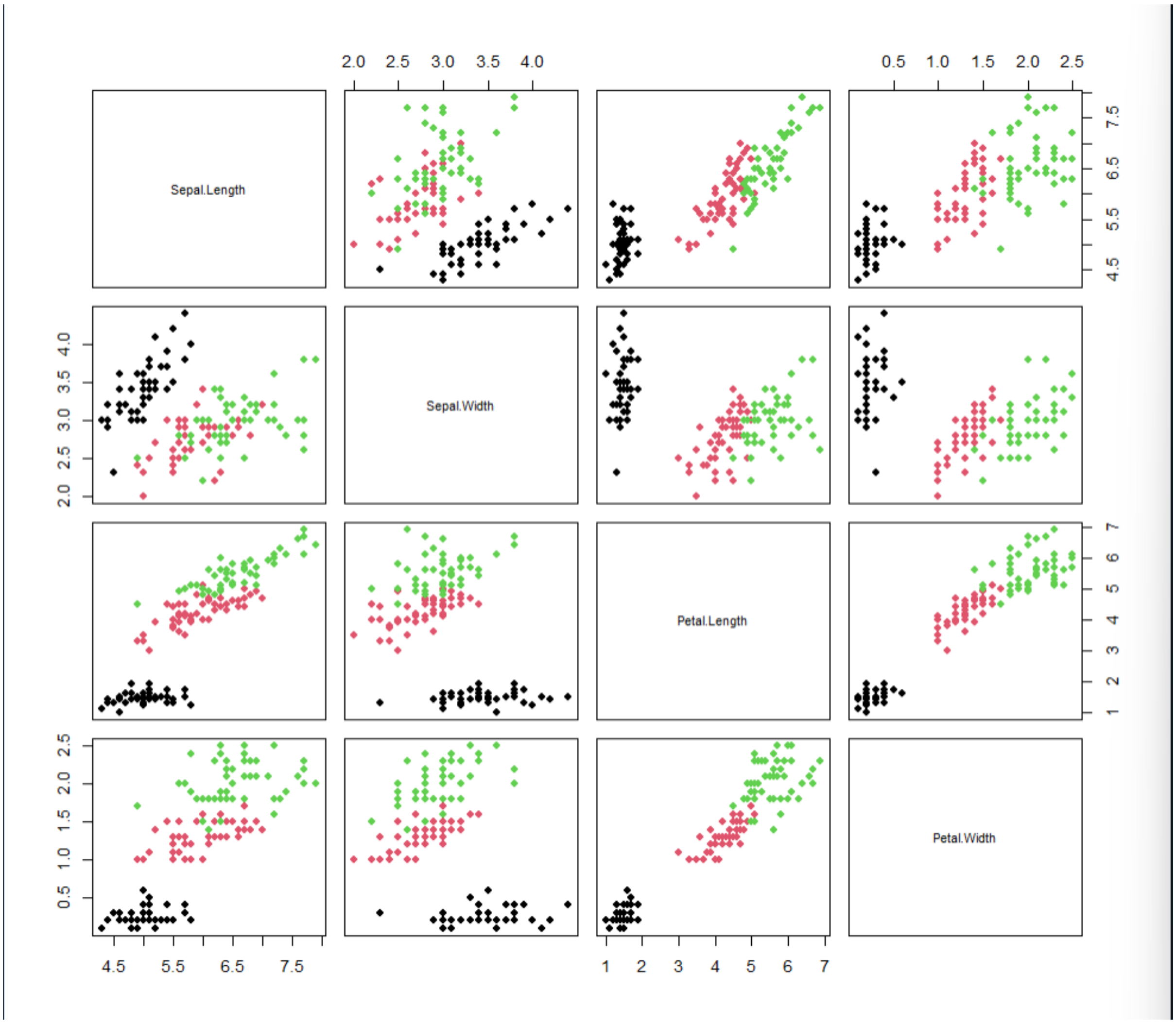


```
> summary( iris$Species )
    setosa versicolor  virginica 
        50         50         50 
```

## Bivariate Analysis



```
# A tibble: 3 × 5
  Species    Sepal.Length Sepal.Width Petal.Length Petal.Width
  <fct>             <dbl>       <dbl>        <dbl>       <dbl>
1 setosa             5.01        3.43         1.46       0.246
2 versicolor         5.94        2.77         4.26       1.33 
3 virginica          6.59        2.97         5.55       2.03 
```

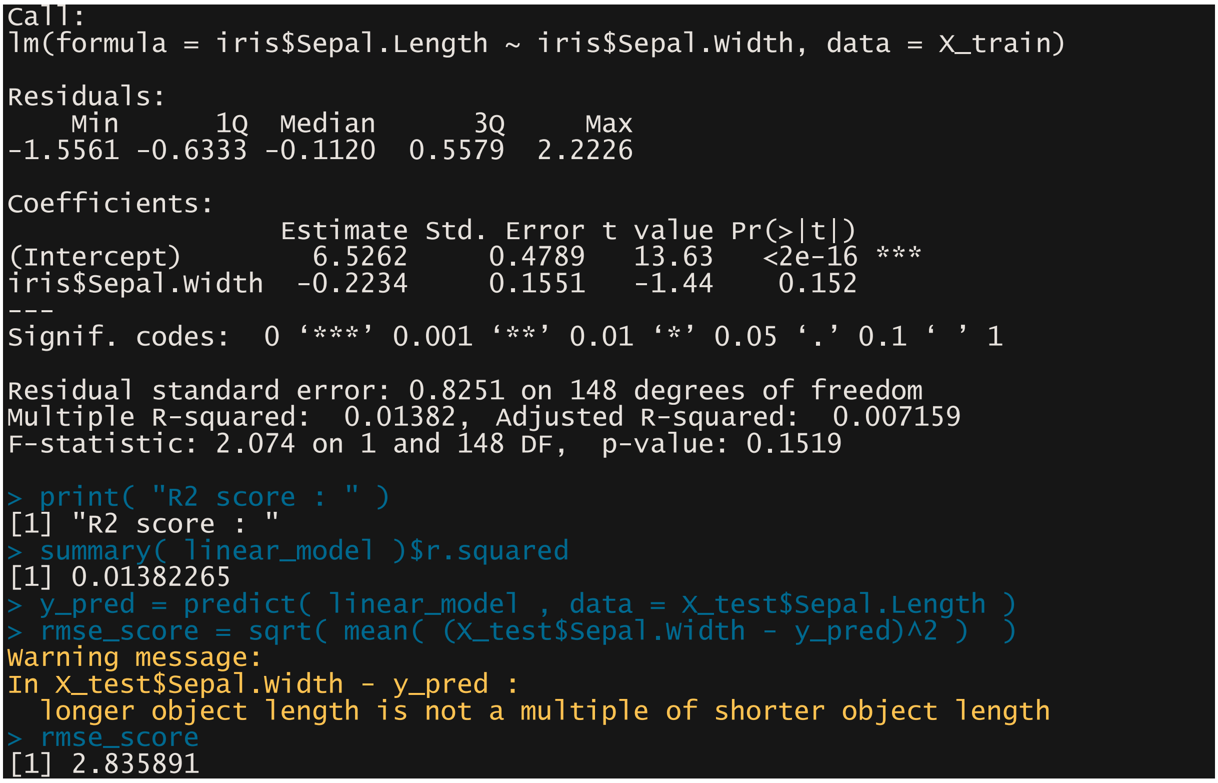


This pair plot clearly shows how is one feature correlated to the other. For instance, Petal Length is highly correlated to Petal width as one could see that the scatter plot is grows linearly, that too with less sparsity. Also, one could notice that there exists hardly any correlation between Setosa species' Petal Length and Sepal Width.

One could also use some classification algorithm, such as Logistic Regression ( in case of Linearly Separable Data ) or in this case SVM (Support Vector Machine (Soft Margin) ) to train a model to classify the Species based on only the Sepal Length, Sepal Width , Petal Length and Petal Width.

# Regression Model

We have implemented Simple Linear Regression Algorithm on the dataset, where one could input Sepal Length and get Sepal Width. For that purpose, we have trained our model on 75% (112,2) data and tested on only 25% data (38,2). The model after training was not so satisfactory as our rmse and r2 score was quite low



```
Call:
lm(formula = iris$Sepal.Length ~ iris$Sepal.Width, data = X_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5561 -0.6333 -0.1120  0.5579  2.2226 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        6.5262     0.4789   13.63   <2e-16 ***
iris$Sepal.Width  -0.2234     0.1551   -1.44    0.152    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8251 on 148 degrees of freedom
Multiple R-squared:  0.01382,	Adjusted R-squared:  0.007159 
F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519

> print( "R2 score : " )
[1] "R2 score : "
> summary( linear_model )$r.squared
[1] 0.01382265
> y_pred = predict( linear_model , data = X_test$Sepal.Length )
> rmse_score = sqrt( mean( (X_test$Sepal.Width - y_pred)^2 )  )
Warning message:
In X_test$Sepal.Width - y_pred :
  longer object length is not a multiple of shorter object length
> rmse_score
[1] 2.835891
```

This also gives us info about the F-statistic and P-value ( whose value if greater than 0.5 signifies that the distribution is close to Normal Distribution). The intercept and Slope value helps us find the regression line.

**Eqn. of the Line = (-0.22333611)*Test + 6.5262226**

# Conclusion:

By performing univariate analysis on the Iris dataset, we can individually explore each feature's distribution and characteristics. For instance, we can observe the ranges, central tendencies, and spreads of sepal length, sepal width, petal length, and petal width. Bivariate analysis allows us to uncover potential relationships between pairs of features, enabling insights into how changes in one variable relate to changes in another. Simple regression, applied to predict one variable based on another, can reveal linear relationships and provide a straightforward model for prediction. For instance, predicting petal length based on sepal length could be examined through simple regression. These analyses collectively offer a comprehensive understanding of the dataset's patterns, relationships, and predictive capabilities, contributing to informed decision-making in various analytical and modeling contexts.