Assignment 2

Antonio Dehesa

Intent of the application

The purpose of this program is to explore the Iris dataset as the first programming assignment for the Data Mining course, perform exploratory operations, display features of the dataset, and create plots for the dataset.

The plots will be of the sepal length, sepal width, and petal length, which are characteristics in the observations of the dataset.

Dataset to be used, including source

The dataset to be used is the Iris dataset, for which more information can be found here: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/iris.html

Use case

The use case of this application is be to demonstrate basic operations with datasets, as the first steps to take when exploring a dataset for the first time, and show the contents of the iris dataset.

Variables

Sepal.Length: The length of the sepal, which is the outer part of the flower that encloses a developing bud. Sepal.Width: The width of the sepal. Petal.Length: The length of the petal of the flower, which are leaves that surround the reproductive parts of a flower. Petal.Width: The width of the petal of the flower.

Labels

Species: Describes the species of the flower associated to the previous measurements.

Data import

In this application, there is no input needed from the user.

Proposed Libraries

datasets: Used to import the Iris dataset ggplot2: Used to create plots from the dataset

```
knitr::opts_chunk$set(echo = FALSE)

library(datasets)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

Library source

datasets: source -> https://cran.r-project.org/package=dataset ggplot2: source -> https://github.com/tidyverse/ggplot2

Dataset Analysis

We need to import the Iris dataset, for which we will explore the available information.

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

We can see that there are 5 total variables in the dataset: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species. We can also see that the first four variables are of numeric type, while the last variable is of type Factor. Finally, we can see that there are a total of 150 observations in the dataset.

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##          4.3          2.0          1.0          0.1
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##     5.843333     3.057333     3.758000     1.199333
```

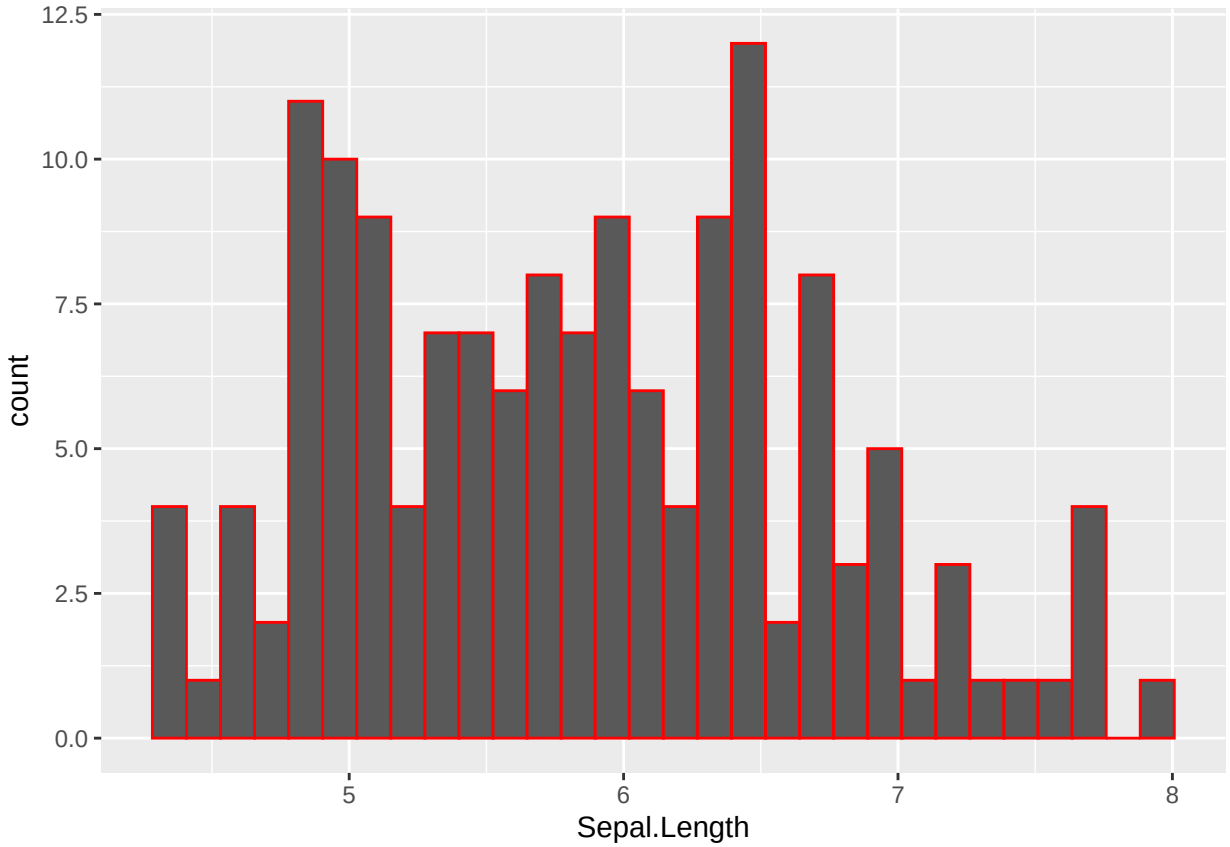
```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##          7.9          4.4          6.9          2.5
```

With this, we can see the approximate range for the dataset, as we can see the minimum value for each column, the maximum, and the average, except for the Species column, as this is not a numeric column.

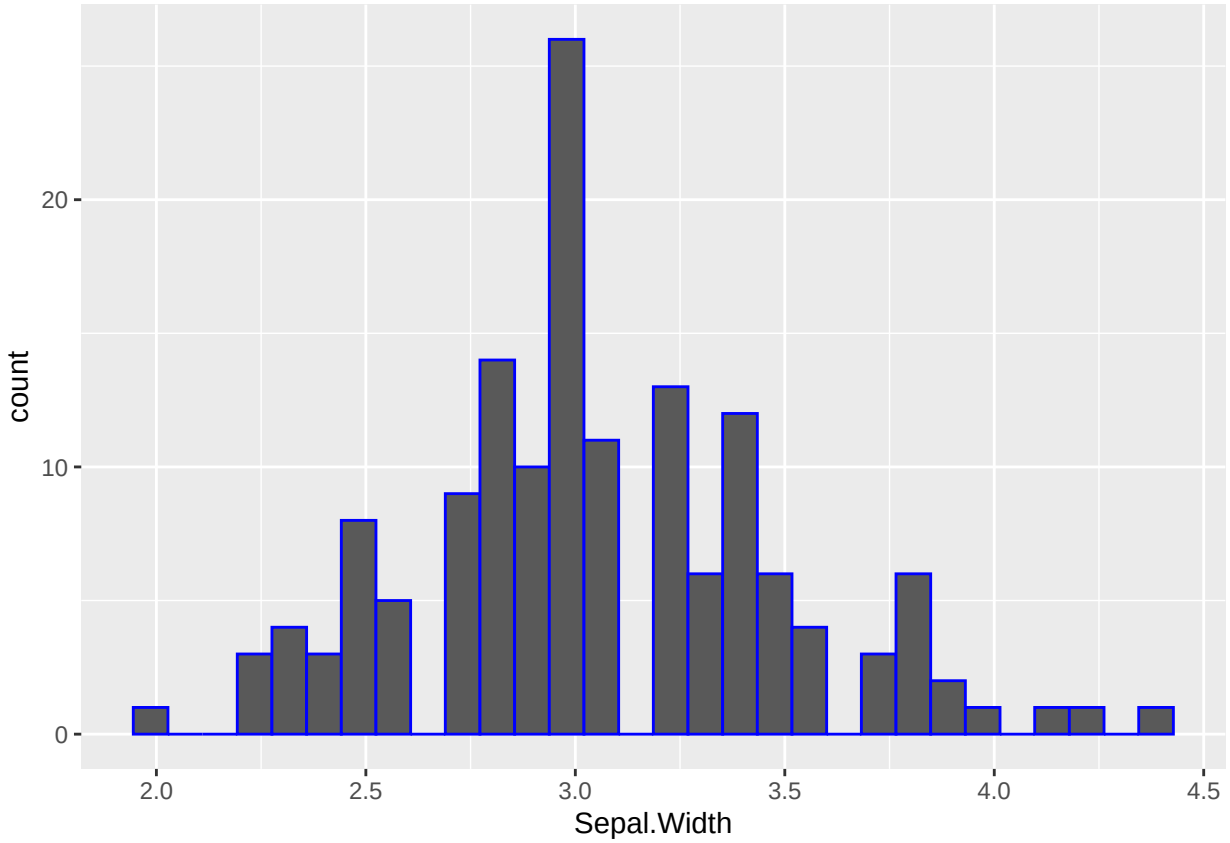
Visualization of outputs

First, we can look at the distribution for each of the variables that are going to be plotted

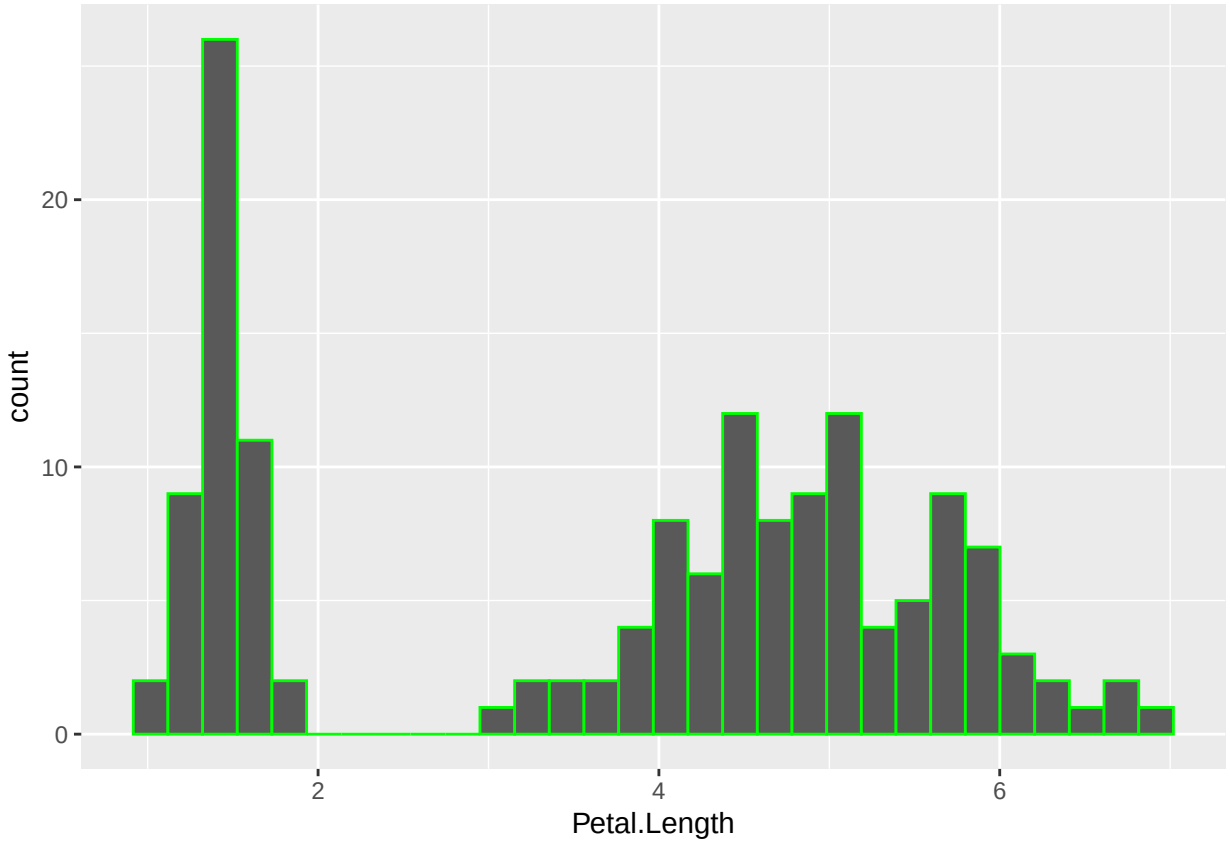
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



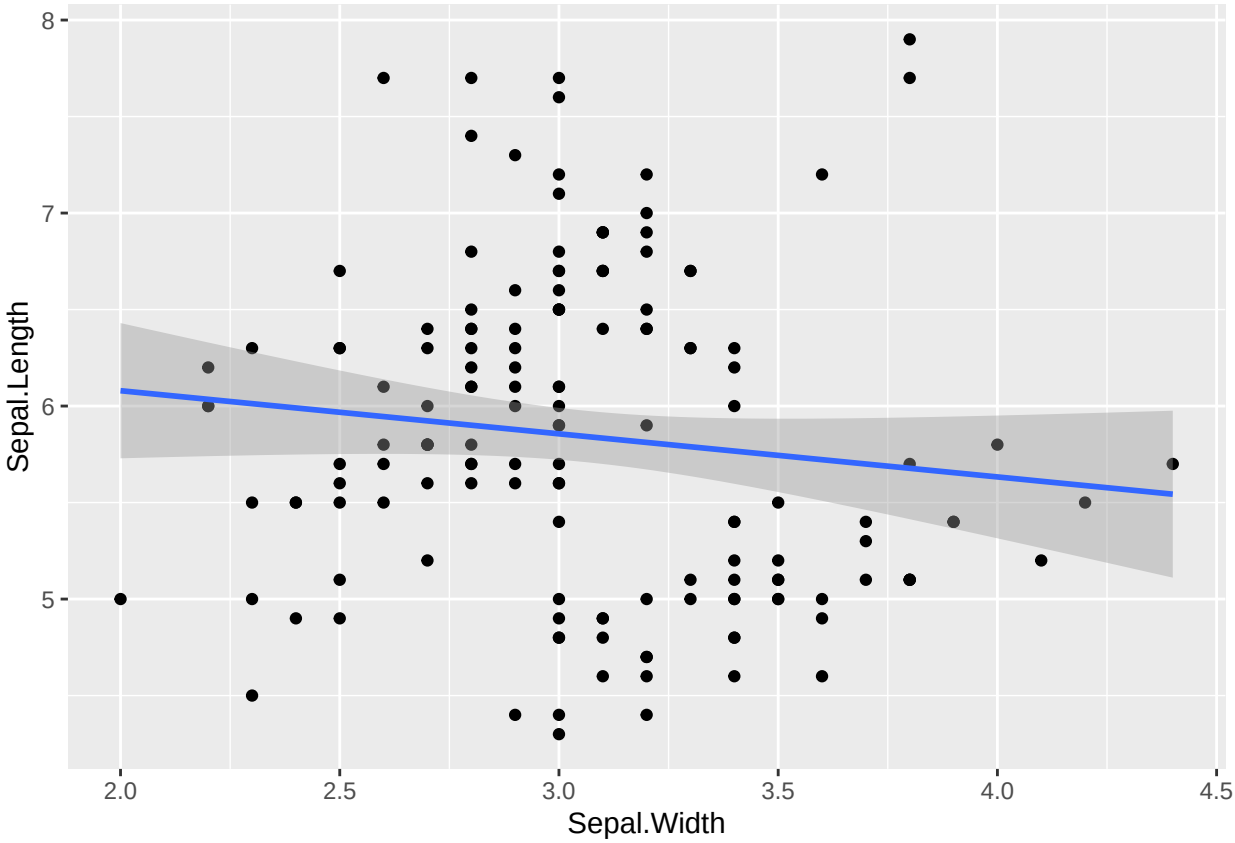

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can see that the Sepal Width is normally distributed, while sepal length and petal length are not. Sepal length seems to be completely randomly distributed, while petal length is split in two groups, with a large gap between them.

Now we can plot Sepal length vs Sepal width:

```
## `geom_smooth()` using formula = 'y ~ x'
```

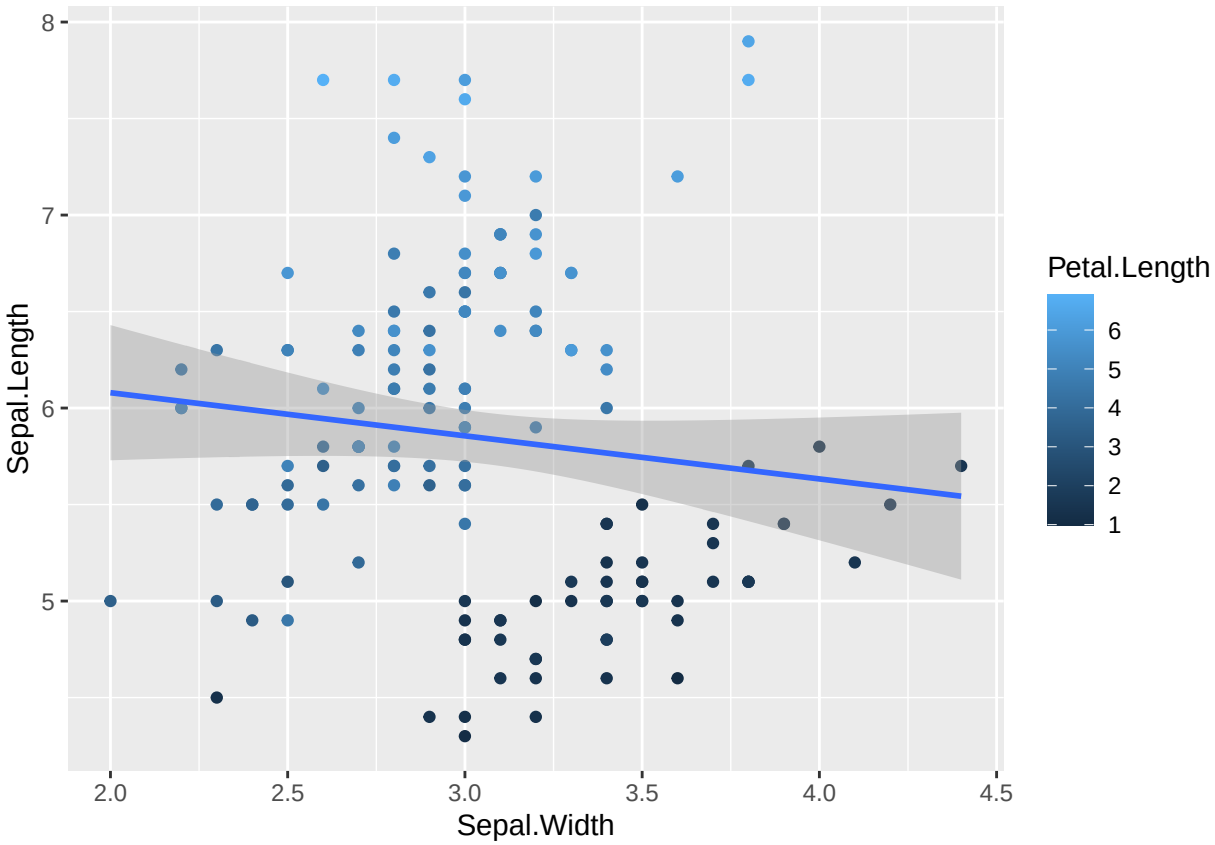


By creating a plot of sepal length vs sepal width, we cannot see any obvious or clear relationship. By using a linear model to fit a line for the points, we can see that there really is no direct relationship between the sepal length and the sepal width.

We can also add the petal length as a third factor. In this case, we can include it as a color gradient to the plot.

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



Analysis of results

We can see a relationship, in which most of the points on the bottom-right part of the plot tend to be darker, which with the color gradient code, we know it means that most of these points have a petal length within 1 and 3, while most of the points on the upper-left part of the plot tend to be lighter, which means they are between 4 and 6.

In other words, the higher the sepal width, the lower the sepal length, the petal length would decrease. The lower the sepal width, the higher the sepal length, the petal length would increase.