Data Mining and Machine Learning Lab 2.

In this R lab you will:

Be introduced to Data Exploration in R.

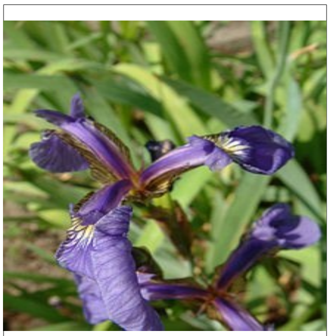
Instructions: Create a file called xxxxxxxx.doc where <xxxxxxxx> is your UCD student number. Write your answers in this file and save it to your own computer so you don't lose your answers. Then upload to the moodle before the end of the lab.

At the top of the file, fill in your details below (delete the 'x' where your information goes):

Name: x
BDIC Student Number: x
UCD Student Number: x

The Iris flower data set

In this lab we will use the *Iris flower data set.* The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936. The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.



Iris setosa

Iris versicolor

Iris virginica

1. **Load the iris data set**

```
> data(iris)
```

This is your **ABT** – 4 descriptive features (Sepal.Length Sepal.Width Petal.Length Petal.Width) and one target feature (Species)

The function dim() returns the dimensions of the data set:

```
> dim(iris)
[1] 150   5
```

Take a look at the attributes of the iris data set:

```
> attributes(iris)
$names
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"     

$row.names
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
 [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
 [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
 [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
 [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
 [91]  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
[109] 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
[127] 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
[145] 145 146 147 148 149 150

$class
[1] "data.frame"
```

Look at the first rows of data using:

```
> 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
```

Look at the last rows of data using:

```
> tail(iris)
```

Or, specify exactly which rows you want to look at e.g. the first three rows:

```
> iris[1:3, ]
```

The first 10 values of Sepal.Length can be extracted using either:

```
> iris[1:10, "Sepal.Length"]
```

or

```
> iris$Sepal.Length[1:10]
```

2. Get summary statistics

Summary statistics are a good place to start when looking at a new dataset. The summary() function provides the summary of each variable in a data.frame.

The syntax and use of the summary() function is: summary(object)

The function returns a table with a column for each of the variables in a data.frame.

For categorical variables (factors, logical, and character variables), the frequency of occurrences is returned. Low frequency levels will be combined in an "other" category if there are too many levels to be displayed.

For numeric variables, the five number summary (median, first quartile, third quartile, min, and max) and the mean are returned.

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
       Species  
 setosa    :50  
 versicolor:50  
 virginica :50  
```

3. The Data Quality Report

We can use the information above to generate the Data Quality Report, but we also need:

- the standard deviation for each feature
- the percentage of instances in the ABT that are missing a value for each feature
- the cardinality of each feature (cardinality measures the number of distinct values present in the ABT for a feature)

Can you get the standard deviation for sepal length?

```
> sd(iris$Sepal.Length)
[1] 0.8280661
```

Question 1:

- What is standard deviation for sepal length? 0.8280661

To get the % missing and the cardinality we can use the Library Hmisc. To install the library tyep:

```
> install.packages("Hmisc")
```

(This will take a few minutes...)

Load the Library:

```
> library("Hmisc")
```

Use the function describe() get get a summery of the iris data set which shows if there is any missing data and the number of distinct values for each feature (i.e. the cardinality):

```
> describe(iris)
iris

 5  Variables      150  Observations
-------------------------------------------------------------------------------
Sepal.Length
    n  missing distinct     Info     Mean      Gmd      .05      .10
  150        0       35    0.998    5.843   0.9462    4.600    4.800
  .25      .50      .75      .90      .95
 5.100    5.800    6.400    6.900    7.255

lowest : 4.3 4.4 4.5 4.6 4.7, highest: 7.3 7.4 7.6 7.7 7.9
-------------------------------------------------------------------------------
Sepal.Width
    n  missing distinct     Info     Mean      Gmd      .05      .10
  150        0       23    0.992    3.057   0.4872    2.345    2.500
  .25      .50      .75      .90      .95
 2.800    3.000    3.300    3.610    3.800

lowest : 2.0 2.2 2.3 2.4 2.5, highest: 3.9 4.0 4.1 4.2 4.4
-------------------------------------------------------------------------------
Petal.Length
    n  missing distinct     Info     Mean      Gmd      .05      .10
  150        0       43    0.998    3.758    1.979     1.30     1.40
  .25      .50      .75      .90      .95
  1.60     4.35     5.10     5.80     6.10

lowest : 1.0 1.1 1.2 1.3 1.4, highest: 6.3 6.4 6.6 6.7 6.9
-------------------------------------------------------------------------------
Petal.Width
    n  missing distinct     Info     Mean      Gmd      .05      .10
  150        0       22     0.99    1.199   0.8676      0.2      0.2
  .25      .50      .75      .90      .95
  0.3      1.3      1.8      2.2      2.3

lowest : 0.1 0.2 0.3 0.4 0.5, highest: 2.1 2.2 2.3 2.4 2.5
-------------------------------------------------------------------------------
Species
    n  missing distinct
  150        0        3

Value            setosa versicolor  virginica
Frequency            50         50         50
Proportion        0.333      0.333      0.333
-------------------------------------------------------------------------------
```

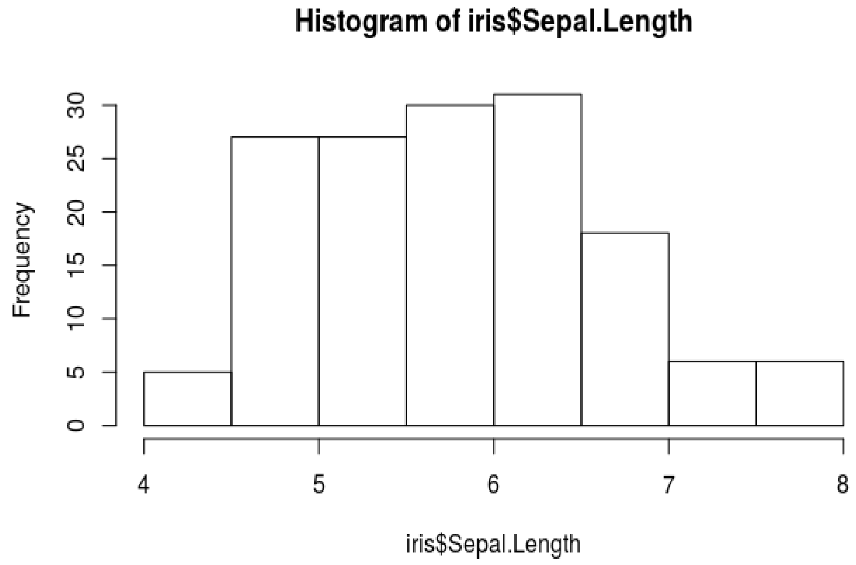
Question 2:

- What is the cardinality i.e. the number of “`distinct`” values for:
  - Sepal.Length 35
  - Sepal.Width 23
  - Petal.Length 43
  - Petal.Width 22
  - Species 3

4. Histograms

Next plot histograms for your 4 descriptive features. For Sepal length you can do this:

```
> hist(iris$Sepal.Length)
```



The histogram will appear in RStudio. Click on the “export” button and save the histogram and copy it into your .doc file which you will upload to moodle at the end of the lab. Repeat this for the other 3 descriptive features.

5. Bar Plot

Create a bar plot for the target feature (Species):

Species is “Categorical” not numerical i.e. see what you get when you type:

```
> iris$Species
```

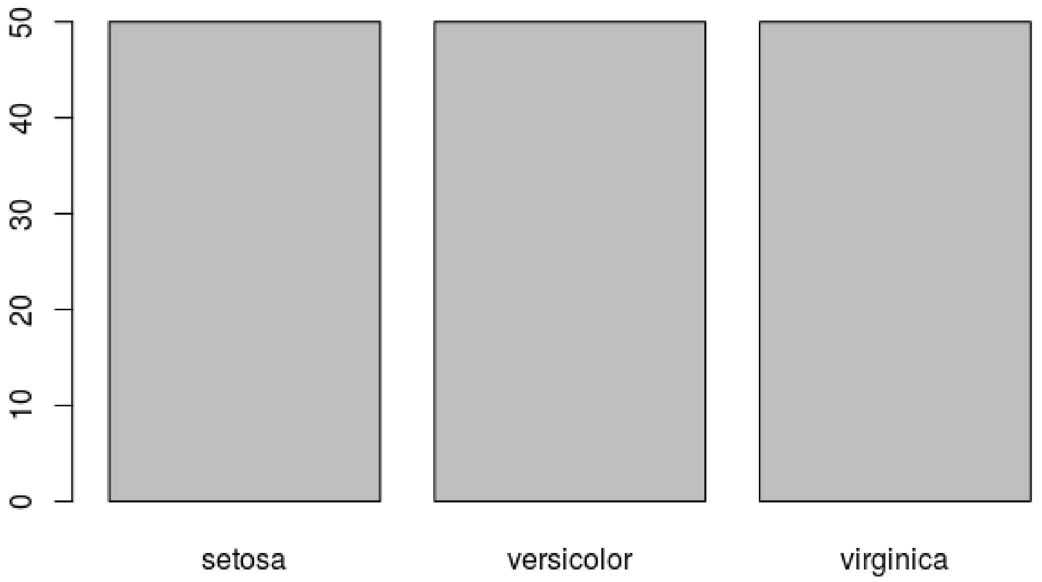
So, before we can plot the data we must create a table:

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

We can then create the bar plot:

```
> barplot(table(iris$Species))
```

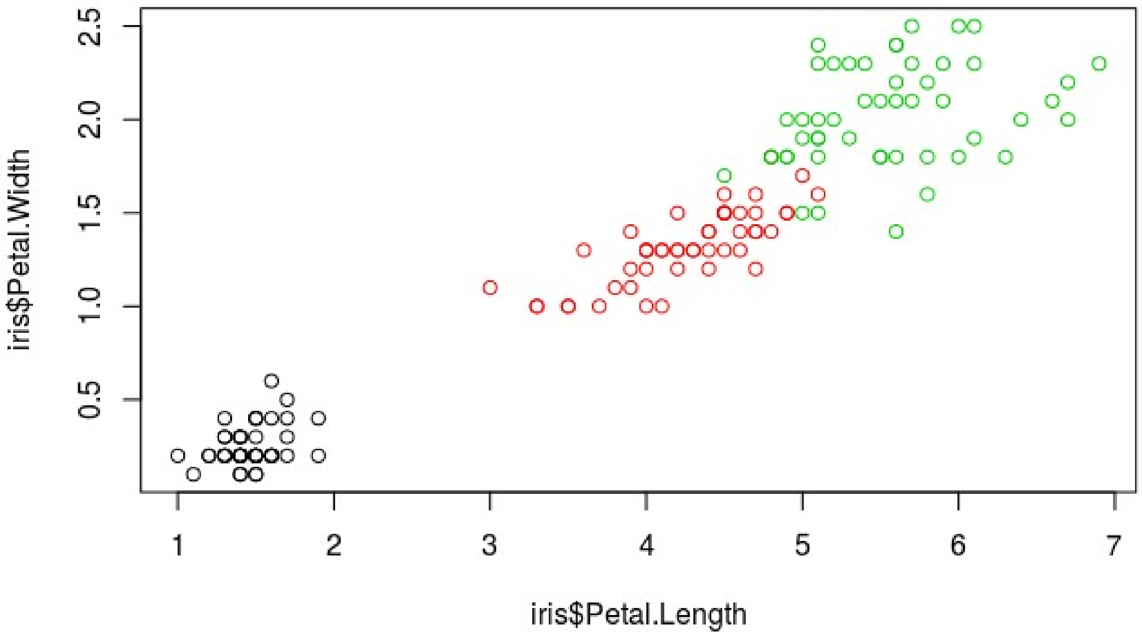
Create the bar plot and export it, save it and copy it into your file for upload.



6. Scatter plots

We can use scatter plots to show the relationship between two numeric variables as follows:

```
> plot(x=iris$Petal.Length, y=iris$Petal.Width, col=iris$Species)
```



Export the plot and save it into your file for upload.

Repeat for Sepal.Length and Sepal.Width

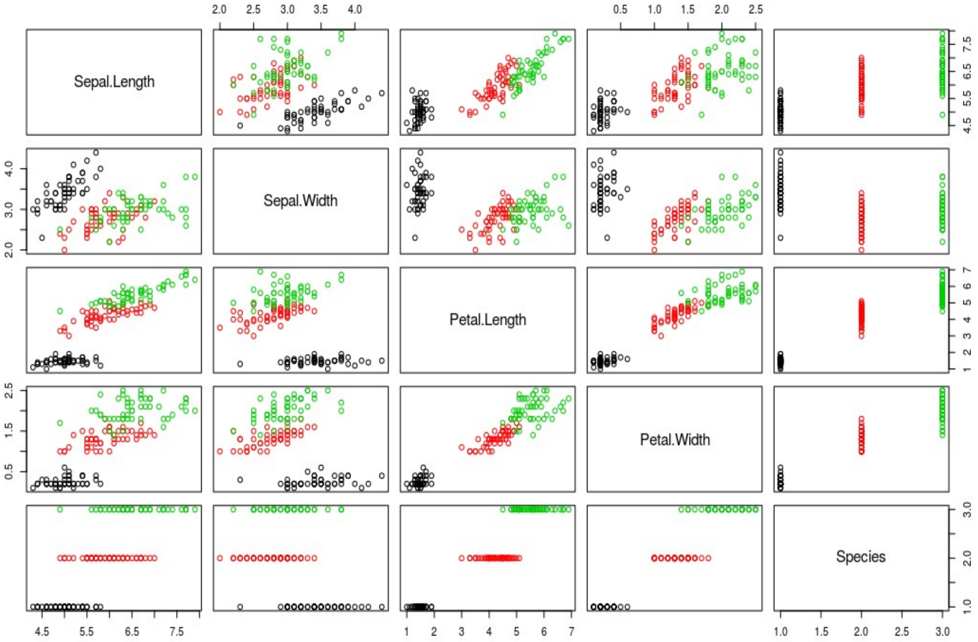
<u>Question 3</u>: What do you notice about the two plots (i.e. does it look like there is a correlation between either pair of features)?

- The plots show that there is a strong correlation between Petal length and width (i.e. as petal length increases so does petal width) but there is not a correlation between Sepal length and width.

7. Scatter plot matrix

A scatter plot matrix shows the relationship between several numeric variables and can be plotted as follows:

```
> pairs(iris, col=iris$Species)
```



8. Correlation

Another useful method to explore the relationships within a dataset is to examine the correlation between the variables.

The syntax and use of the cor() function

cor(object)

Returns a matrix of the correlations.

The object passed to cor needs to be a two dimensional object with a type of numeric.

The iris data set is a two dimensional object, though not all the variables are numeric. So we will exclude the non-numeric variable (Species) from our call to cor().

```
> cor(iris[, 1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

9. Correlation matrix visualization

Install the package 'corrplot' and include the library:

```
> install.packages('corrplot')

> library('corrplot')
```

Create the correlation matrix:

```
> correlationMatrix <- cor(iris[, 1:4])
```

Create the plot:

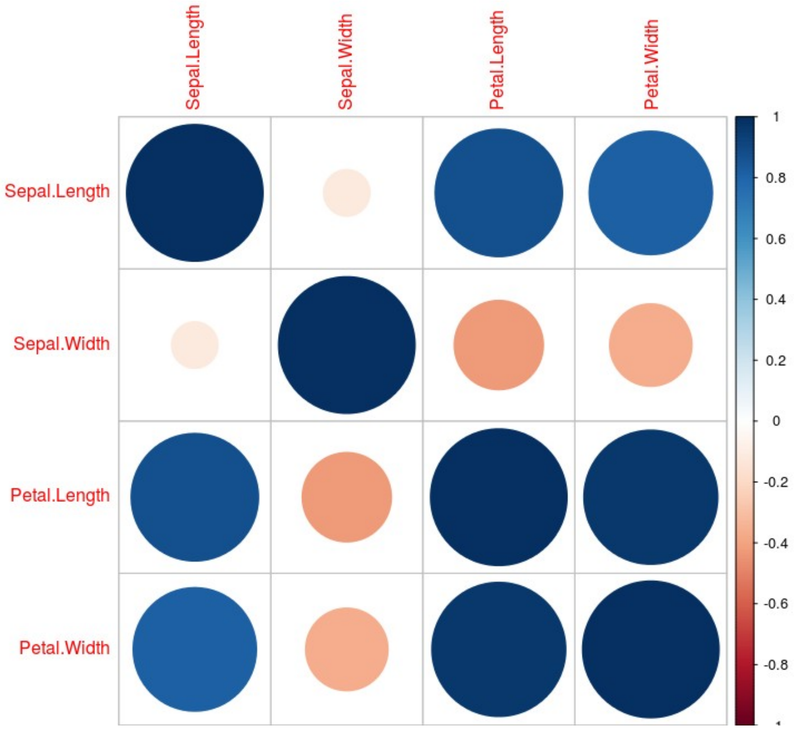
```
> corrplot(correlationMatrix, method = "circle")
```

Export the plot and save it into your file for upload.

<u>Question 4</u>: What does the colour of the circle signify? What does the size of the circle mean? (hint `?corrplot`)

- The sizes of the circles are scaled according to the absolute value of the strength of the correlation to draw attention to those pairs of features with the strongest relationships.

- Blue represents positive correlations and red represents negative or inverse correlations.


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width