# **CTA1 – Option 1**

Austin Brown

Colorado State University

MIS500: Foundations of Data Analytics

Dr. Jamia Mills

9/26/2021

### Task 1 – Display The First Six Rows Of The Iris Dataset

#### **Input**

**Figure 1**

R Script, Task 1

Text

Description automatically generated with low confidence

#### **Output**

**Figure 2**

R Output, Task 2

Table

Description automatically generated

#### **Description**

Within RStudio I’ve assigned the dataset, limited to the first 6 rows via the function head(), to the variable *data*. I then return *data* as well as the Sys.Date() function to have the terminal display the desired results from the dataset as well as the system date.

### Task 2 – Display Summary Statistics For All Attributes Of The Iris Dataset

#### **Input**

**Figure 3**

R Script, Task 2



#### **Output**

**Figure 4**

R Output, Task 2

A picture containing text

Description automatically generated

#### **Description**

Unlike the previous task, I do not assign any portion of the Iris dataset to another variable – I run the summary() function against the entire dataset in order to get summary statistics for every attribute. The summary statistics are as follows:

1. Minimum
2. 1st Quartile
3. Median
4. Mean
5. 3rd Quartile
6. Maximum

### Task 3 – Select Any Two Attributes of the Iris Dataset, Display Selected Graphs and Stats

#### **Input**

**Figure 5**

R Script, Task 3

Text

Description automatically generated

#### **Output**

**Figure 6**

**Figure 7**

R Output (Graphical), Task 3

R Output, Task 3

Chart, box and whisker chart

Description automatically generatedText

Description automatically generated

#### **Description**

For this phase we were tasked with displaying the Mean, Median, and Range of two attributes within the Iris dataset as well as displaying a histogram and box plot of each. As the Mean, Median, and Range are displayed through the use of the summary() function I employ it here while specifically targeting my chosen attributes. As you can see from the labels, my chosen attributes for this analysis are *Sepal Length* and *Petal Length*.

I also display the requested visualizations with preferential parameters such as axis labels and increment controls. I show multiple visualizations here through partitioning the graphical output area via the par() function.

### Task 4 – Display The Correlation & Plot Of The Two Attributes Used In Task 3

#### **Input**

**Figure 8**

R Script, Task 4

Text

Description automatically generated with medium confidence

#### **Output**

**Figure 9**

**Figure 10**

R Output (Graphical) Task 4

R Output, Task 4

Chart, scatter chart

Description automatically generatedGraphical user interface, text, application

Description automatically generated with medium confidence

#### **Description**

I use the cor() function to analyze the correlation between my previously selected Attributes. With a return of .8718 as a correlation value, we can confirm that the two attributes in question are strongly correlated with one another in a positive manner.

I also display a scatter plot, plot(x,y), of the data here that visibly shows the positive relationship between the two attributes.

### Task 5 – Explain The Work And Interpret The Results

This CTA assignment had us perform some basis summary statistics on one of R’s pre-built data sets. Our work here provided us a general view of the data and two attributes of our choosing at a descriptive level. These actions act as stepping stones to further predictive and inferential analysis.

Our first step was to establish a return of data, limited to the first 6 rows, as well as return some descriptive summary statistics on the set as a whole. Using the summary() function against the Iris dataset allowed us to easily get a holistic view of how the data set and it’s attributes were constructed. Using the summary() return of the Minimum, 1st Quartile, Mean, Median, 3rd Quartile, and Maximum allows for a quick mental visualization of how the data exists. At this point, we have the statistical values for each attribute that show use measures of central tendency as well as 3 separate classifications of object types, though we will not be breaking those types out in this assignment.

From this point forward we hone in on two attributes of our choosing, my choice being Sepal Length and Petal Length. After running the same summary statistics and getting the measures of central tendency against the two attributes as well as creating two sets of visualizations via a Box & Whisker and Histogram, we can the structure of our two attributes. Based on the output, Sepal Length can be described as having a somewhat focused but relatively normal distribution with the highest frequency of results being between 5 and 6. Petal Length, on the other hand, is Bi-Modal in nature, having two points of frequency highs between 1 and 2 then 4 and 5. This occurrence makes it difficult to interpet that particular attribute through averages.

Our final step was to take our first action in comparing these two attributes through finding their correlation with one another. This is easily done in R through the cor() function and resulted in a value of .8718 between the two values. This value represents what is called pearson’s correlation coefficient, which is used in linear and multiple regression to determine variable relation strength and direction. As this value was positive and nearing the maximum correlation of 1 (correlation falls between -1 and 1) it can be interpreted as being a strong positive correlation between these two attributes. Note that this does not mean that one causes the other, just that there is a strong trend between the two.

Conclusions I have been able to draw throughout the assignment are as follows: the two chosen attributes, Sepal Length and Petal Length, give us an example of a normal and bi-modal distribution; and these two attributes have strong correlation with one another that should be investigated. I can also say that it may be necessary, depending on the methods used, to normalize that bi-modal data of petal length as it will interfere with analyses dependent on mean and error calculations. I would be interested to continue on and conduct some hypotheses tests and predictive models from this data set.

# References

N/A