**Exercise 1 - Getting Started with R and Exploring Iris Data**

Top of Form

|  |
| --- |
|  |

**Part 1** – Write an R script with the following commands in the same order as listed below.

The script contents

* Line 1 - a comment #This is my first R script
* Line 2 - Sys.time() command
* A screenshot of a computer

  Description automatically generated
* Line 3 - Sys.info() command
* Text

  Description automatically generated
* Line 4 - R.version command
* Text

  Description automatically generated
* Line 5 – leave blank
* Line 6 – the command to display the first 5 rows in the iris data
* Text

  Description automatically generated
* Line 7 – the command to display the maximum value of the Petal.width variable
* Text

  Description automatically generated
* Line 8 – the summary command to display the statistics only for Petal.width variable
* Text

  Description automatically generated
* Line 9 – the command to display the help page for plot command
* A screenshot of a computer

  Description automatically generated with medium confidenceLine 10 – the hist command to build the histogram for the Petal.width variable
* A screenshot of a computer

  Description automatically generated with medium confidence

Save the file as your first initial Last name\_exercise1 (for example, Yelena Bytenskaya saves the file as YBytenskaya\_exercise1)

**Part 2 Part 1 script output and iris data exploration**

1. **Explain what information could the R commands reveal about the iris data? To answer this question, you may need to review the tutorial in Part 0**
   * + Head(iris,5) just shows the first 5 rows which is a nice sneak peak into the iris dataset. This comes in handy to get an idea of the format of the data set before cleaning.
     + Max(iris$Petal.Length) goes through the column Petal.Length and finds the max value
     + Summary(iris.$Petal.Width) gives a summary of the data you can find in the column Petal.Width. This is especially helpful when exploring the data and seeing if there are any preliminary trends you can become aware of
     + Hist(iris$Petal.Width) This is similar to using the summary command because it helps give you an idea of the pattern of data in the column but helps visualize to find a trend faster than numbers might.

**Part 3 Data Visualization**

1. **Explain what the hist command on line 10 in Part 1 reveals about Petal.width variable in the iris data**
   * + The histogram in line 10 helps visualize where the Petal.Width data is concentrated around what widths in this case. It shows that the data is normal but also has a gathering around the shorter lengths too. The visuals help the human brain to see patterns quickly as it is easier to comprehend the spacing in images instead of nunbers like we would see with the summary command.
2. **Why is it important for the data-mining tool to have visualization capabilities?**
   * + Similar to the answer above, visualization in data mining is important to help identify any trends and patterns in the data set before cleaning and utilizing the data set to help answer any questions. This is helpful because it is a quick tool for the human mind to identify any possible trends or any outliers in the data before beginning the data cleaning. We could see that in our previous course when we were using Cognos Analytics and can see it in real world data sets that have not yet been cleaned.

**Part 4: Is an open source data mining tool better than a commercial one? Why or why not? To answer this question, consider the key differences between open source and close source tools.**

For usability, commercial software is generally better than open source. This is because even though it is expensive, it is most likely well supported so any issues a user may run into, can generally be resolved whereas open source does not have the support that a new user may need to the data mining tool. Also, since commercial data mining tools would want to keep their users, they would also try to keep in mind the ease of understanding for the user and therefore the user interface is likely to be easier to navigate and understand. As a user gets more advanced, the preference will likely lean more toward open source data mining tools for the flexibility and cost. After all, as users get more advanced, so will their use cases and commercial tools cannot always have the specific solution that a more advanced user is looking for and that is when open source tools can help.

Bottom of Form