DA5020 – Assignment *2*

This and all subsequent assignments must be done using R Markdown Notebooks. Be sure to properly "chunk" your code and add comments to each chunk. Make it clear which question you are answering, and each chunk of code should accomplish a small well-defined task.

Using the built-in dataset **msleep**, answer the following questions and demonstrate the use of dplyr verbs and pipes. Use the help function to learn more about the dataset e.g. **?msleep**

You may want to complete the two data camps Intermediate R and Data Manipulation with dplyr before attempting this assignment and will earn bonus points by completing them.

Question 1 — *(+5 Bonus Points, All instructions must be followed)* (Both questions must be completed to obtain the bonus points)

1. Complete the data camp "Intermediate R and complete a seperate datacamp titled "Data Manipulation with dplyr" (two data camp courses and both must be completed to earn the bonus, no partial credit). A link to Data Camp will be shared within the course, so that you can access the resource for free using your Northeastern email. Please submit the statement of accomplishment from both data camps to earn credit (must also have your name on it).

2. Load the **msleep** dataset and inspect the dimensions and properties of the data e.g. what type of observations are recorded in the dataset, are any missing, etc. Summarize your findings/understanding about the dataset based on what you have observed.

Question 2 — (10 points)

Calculate the proportion of each category of **vore** as a percentage and visualize the results using a bar chart. Comment on the results.

Question 3 — (10 points)

Filter the data to extract data for just omnivores. Hint: vore == 'herbi' and vore == 'carni'. Calculate the mean **sleep_total** for that group.

Question 4 — (10 points)

Create a scatterplot showing the relationship between **bodywt** and **brainwt**. Comment on any correlation that is visually apparent. Tip: if you rescale the x and y axis using a logarithmic scale, it may help you to interpret the visualization better e.g. you can append the following to ggplot: scale_x_log10() and scale_y_log10().

Question 5 — (5 points)

Calculate the Pearson coefficient of correlation in R, to evaluate the strength of the correlation between **bodywt** and **brainwt**. Did the results support your original assumptions from question 4?

Question 6 — (15 points)

Determine which mammals are outliers in terms of **sleep_total**. Outliers, for the sake of this question, are defined as values that are more than 1.5 standard deviations from the mean. Display the **name** and **sleep_total** of the mammals which are outliers.

Submission Details

** Note: Do not merely print the output for each question but always explain the results from your code. Points will be deducted for any missing explanation.

- **Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd)**.

- Name your .R script using the following format: DA5020.A1.FirstName.LastName.R, where FirstName.LastName is your first and last name. If you submit an Rmd file, you should use DA5020.A1.FirstName.LastName.Rmd

- Ensure that you number each question correctly.

- We must be able to run your code. You will not receive credit for any code that does not run. Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd). Name your .Rmd file, DA5020.A2.FirstName.LastName.Rmd and your PDF/HTML DA5020.A2.FirstName.LastName.{pdf,html}, where FirstName.LastName is your first and last name.

- The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it.) Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.

- Not submitting a knitted PDF or HTML will result in reduction of 30 points.

- Not submitting the .Rmd file (or both) will result in a score of 0.

- Describe answers for full marks