Lab One

MSDS 598 Spring 2022

Directions

- Submit a .ipynb notebook to Canvas.
- The Notebook should begin with a Markdown cell with your **Name** and the title of the Assignment, **Lab 1**. Failure to do so will result in points lost.
- Use Markdown Cells to **clearly** indicate which code answers which question and to answer short answer questions. Failure to do so will result in points lost.
- The filename for your notebook should be formatted like

  $$FirstName\_LastName\_AssignmentName.ipynb.$$

  Failure to do so will result in points lost.
- This is due on January 30th at Midnight Pacific time. This will not be graded for correctness. Solutions will be posted after the due date.

For many of the following questions you will need to use the penguins dataset. This can be imported either using seaborn by

```
import seaborn as sns
df = sns.load_dataset('penguins')
```

Or it can be imported directly from the seaborn GitHub using

```
df = pd.read_csv(
 'https://raw.githubusercontent.com/mwaskom/seaborn-data/master/penguins.csv')
```

1. What are the variables in this dataset? Which are categorical? Which are continuous? Discrete? (Use pandas commands to answer these questions!)

2. Which columns have missing values? How many?

3. Which species type, on average, weighs more? Does there seem to be a difference between male and female weight?

4. How many species types are there? How many penguins are there for each species type?

5. Create a dataframe of only penguins of species "Adelie".

6. Create a scatter plot where the x values are the first 100 integers in ascending order and the y values are the first 100 integers in descending order.

7. Create *one* figure with four histograms, one for each of the numeric variables in the penguins data

8. Make a scatter plot comparing two of numeric variables in the penguins data. Is there any relationship here?

BONUS (you may need to use Google!)

- Create a new column giving the body mass in pounds.
- Create a bar chart to answer the question "How many penguins are there for each species type?". Now make a horizontal bar chart!
- Create a scatter plot with two of the numeric variables from the penguins dataset. Now incorporate a *third* variable by coloring the scatter plot points according to the `island` variable.