Homework Two

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, **Homework 2**. 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 February 28th at 6:30 PM Pacific time. Solutions will be posted one week after the due date.

For the following questions you will need to use the `taxis` dataset which can be obtained via

```
df_taxis = sns.load_dataset('taxis')
```

1. Suppose we want to use the variables `distance` and `passengers` to predict `total` which represents the total fare.

(a) Use linear regression to create a model prediction `total` from `distance` and `passengers`. Report the R^2.

(b) Incorporate another variable `taxi color` into the model and then report the R^2.

2. As in the fourth lecture, create a length of ride variable. Use the new feature you've engineered to further improve the model and report the R^2.

3. Create a validation set using ten percent of the data, and use the complement to create a training set. (Make sure there is no overlap in observations between the validation and training set).

4. Use this validation and the R^2 metric to decide whether adding length of ride to the variables from Q1 improves the model.

5. Engineer one new feature from the data and repeat Q4. (Hint: you can use powers, differences, ratios, products of other features!)

6. Why shouldn't you use `fare` or `tip` to predict `total`?