Homework 3 (50 points)

Data Processing, Viz and Supervised Learning

Due: Sunday, June 23, 11:59 pm

Instructions

Please carefully read and follow the instructions below to complete your homework assignment.

- The expected form of submission is an IPython notebook.
- Each Part (1,2,3 ...) should start with a Level 1 heading that says Part X.
- Copy each question to your notebook, and make the text bold. Answer the question in the cells below it
- Except for code concerning creating a visualization, for everything else,
  - Write each line of code in a new cell.
  - Show either:
    - Full output, e.g., Number of null values
    - Partial output, e.g., a sample of the dataframe you created.
  - Use your judgment on when to provide full or partial output.
- Unless explicitly mentioned, I expect you to find the answer using programming. Do not write the answer by looking at something. Find it using code. Your answer should be the output of your code.
- Some questions may require text responses. Type the answer in a markdown cell.
- All answers should use pandas or sklearn functions when possible. **Do NOT rebuild the wheel.** Use pandas or sklearn functions when available; failure to do so may result in no credit or penalty. If you cannot find the function, ask in Piazza or reach out to Bennett.

  Example: To count the NaN values in a column, do not iterate over the entire column and count the number of NaN. There is a one-line solution in pandas. Use that.
- While the last homework had incentivized good formatting in notebooks by assigning separate credit for it, in this homework, there is no such credit. But poorly formatted notebooks will receive a penalty up to 4 credits.
- When in doubt, reach out for assistance. We are here to guide you as you build your skills.

Grade Distribution

Submission	**Total**	**% Contribution to Final Grade**	**Impact of 1 point of Submission in Final Grade**
HW1	14	10.5	0.75
HW2	50	19.5	0.39
HW3	50	20	0.4
HW4	50	20	0.4
Final Project	-	20	-
Weekly Checkin	-	5	-
Participation Activities	-	5	-
Total	-	100	-
Bonus Questions	5	5	1

Tabela 1: Grade Distribution

1 Project Setup [0 points]

- Download the dataset from the website. Place the file in the same folder that you have your Ipython notebook. So your code will be pd.read_csv('filename.csv'). ie no folders in paths.
- Read the data into a dataframe named data_df.
- Show 5 random rows of the dataframe.

2 Dataset Description [2 points]

1. Print a concise summary of a DataFrame. Output expected below.

[1 point]

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1156 entries, 0 to 1155
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   START_DATE  1156 non-null   object
 1   END_DATE    1155 non-null   object
 2   CATEGORY    1155 non-null   object
 3   START       1155 non-null   object
 4   STOP        1155 non-null   object
 5   MILES       1156 non-null   float64
 6   PURPOSE     653 non-null    object
dtypes: float64(1), object(6)
memory usage: 63.3+ KB
```

2. Generate descriptive statistics. Output expected below. [1 point]

	MILES
count	1156.000000
mean	21.115398
std	359.299007
min	0.500000
25%	2.900000
50%	6.000000
75%	10.400000
max	12204.700000

3 Data Processing [20 points]

3.1 Start Date [3 points]

1. Check if there are any NaN values in the Start Date column. [1 point]
2. Extract the time part of the start date string, and save it to a new column named start time. [1 point]
3. Convert the column Start date to an appropriate datatype. [1 point]

3.2 End Date [4 points]

1. Check if there are any NaN values in the End Date Column. [0 point]
2. How many NaN values are present in the column? [1 point]
3. Display the row or rows with the NaN values for the End Date Column. [1 point]
4. Remove that row from the data_df dataframe. [1 point]
5. Extract the time part of the end date string, and save it to a new column named end time. [0.25 point]
6. Convert the column End Date to an appropriate datatype. [0.25 point]
7. Confirm the changes to datatype by using the function you used in Q2.1 [0.5 point]

3.3 Category [2 points]

1. Check if there are any null values in the column Category [0 points]
2. Print the labels in the column Category. [1 point]
3. Print the number of data points for each label in the column Category. [1 point]

3.4 Start [6 points]

1. Check if there are any null values in the column Start [0 points]
2. What are the different start locations for the rides in the dataset? [1 point]
3. Create a new dataframe with the name of the start location and a number of rides with that start location. [1 point]
4. Which is the most popular ride start location? [1 point]
5. Amoung all the start locations you listed, you will notice that one location is not the name of an actual location. What label was used instead of an actual location here? [1 point]
6. How many rows have that label? [1 point]
7. Display 10 sample rows with that label as the start location. [1 point]

3.5 Stop [1 point]

1. Check if there are any null values in the column Stop [0 points]
2. What are the different end locations for the rides in the dataset? [0.2 point]
3. Create a new dataframe with the name of the end location and the number of rides with that end location. [0.2 points]
4. Which is the most popular ride end location? [0.2 points]
5. Among all the end locations you listed, you will notice that one location is not the name of an actual location. What label was used instead of an actual location here? [0.2 points]
6. How many rows have that label? [0.2 points]

3.6 Miles [1 point]

1. Are there any null values in the column miles? [0 points]
2. What are the average and median miles for rides? [1 point]

3.7 Purpose [3 points]

1. What are the different purposes recorded for the rides? [1 point]
2. What is the most popular **known** purpose? [1 point]
3. What percentage of total rides were for the purpose above? [1 point]

4 Data Exploration [6 points]

Play around with the data, ie look into the data. Give three insights about rides that you found. *Example: The most popular start location for the rides are ..., The busiest time ie time with the most rides are..., The most popular start and end locations for business rides are...*

5 Data Visualisation [8 points]

Use Matplotlib for all the coding questions in this section. Give appropriate title, x and y-axis labels, and legend where applicable.

1. Show a donut chart displaying the purpose of the trips, **where the purpose is known.** [1 point]
2. Show a histogram with the distribution of miles driven. Use 10 equal-width bins. [3 points]
3. Show any other visualization of your choice to visualise something from the dataset. [4 points]

6 Linear Regression [10 points]

6.1 Data Preparation [5 points]

1. The features we will use are the start location, category, and purpose. Perform label encoding on the three features and store them in a variable **X** [2 points]
2. Store the miles driven in **Y** [1 point]
3. Split the dataset (X, Y) you prepared above into Train (80%), Validation (10%), and Test (10%) splits. Use the sklearn function, you may have to use the function twice. [2 points]

6.2 Model Training [4 points]

Train the following models using the data you have. Save the test performance (metric: MSE) of each of the models on the validation dataset. You can choose whatever parameters you think fit, or just go with the default. It is up to your discretion.

- Linear Regression
- Lasso
- Ridge

6.3 Pick the best model [1 point]

Pick the top-performing model from the 3 models above. Report its performance on the test data.

7 Final notes [4 points]

Text Response

1. Why are datasets split into train, validation, and test? [1 point]
2. What is overfitting and underfitting? [1 point]
3. How does Lasso differ from Ridge ? [2 points]