Take Home Exercise (Data Scientist)

We expect this exercise to take approximately 3-4 hours. Please read the entire document before starting. Provided with the exercise are two datasets: ***site_ridership.csv*** *and* ***pickups.csv****.*

May Mobility has multiple operational sites around the world. One of the operational metrics we are interested in is our ridership numbers. In this exercise, we provide some ridership data we collect manually at one of our sites. The site is a circulator route meaning that it operates like a bus service going around in a loop during service hours. Riders hop on and off the vehicle just like a bus service and that information is recorded manually by site staff using Google Forms.

Additional information:

- Each vehicle is only allowed one household at a time in the vehicle. There is no ride sharing at this site.
- Each vehicle can hold up to a max of 4 riders in a household.
- Service hours are 7am - 7pm.

The dataset file *site_ridership.csv* consists of the following columns (all columns are manually entered in the Google Form except for timestamp):

- **timestamp** (timestamp when Google Form is submitted)
- **pickup** (number of riders picked up)
- **dropoff** (number of riders dropped off)
- **stop** (stop name)
- **vehicle** (vehicle id)
- **time** (approximate time of pickup)
- **date** (date of pickup)
- **name** (initials of person filling out google form)

The dataset file *pickups.csv* consists of the same columns as *site_ridership.csv* with the addition of:

- **row_id** (row unique identifier)

Please return an output file with the answers to the following questions. Also please provide the code with comments in a script or notebook that was used to answer the questions. Please provide any necessary commentary about what you are doing and why.

Note: State any assumptions you make about the data in your solution that was not already given in the question.

Question 1: EDA

Perform an EDA and provide a summary of the data (specifically *site_ridership.csv*). Target the most important features of the dataset first. Does the dataset match your expectations?

Question 2: Data Insights

Are there any data insights that you can provide that would be of interest to stakeholders of the site/company? Feel free to add a plot for clarity if you think it is needed.

Question 3: Modeling I

Suppose the site team wants to predict what the ridership will be in the future for planning purposes. Return a prediction about what the ridership will be on these dates Nov. 15-21. In the commentary, tell us how well you think your model will do.

Submit a comma delimited file called *riders.csv* with the following 2 columns:

```
date,riders
Nov 15,<your prediction>
Nov 16,<your prediction>
…
Nov 21,<your prediction>
```

We will use *riders.csv* on a private test dataset to see how well your model does.

Question 4: Modeling II

Suppose for some unknown reason, during the month of November, the site team only entered pickup data. We would like to fill out this dataset by predicting where the associated dropoff locations are. Predict the dropoff stop based on the input file *pickups.csv*. In the commentary, tell us how well you think your model will do.

Submit a comma delimited file called *dropoff.csv* with the following 2 columns:

```
row_id,dropoff_stop
1,<your prediction>
2,<your prediction>
…
363,<your prediction>
```

We will use *dropoff.csv* on a private test dataset to see how well your model does.

Appendix

The table contains descriptions of the 9 route stops for some context.

Stop	Description	Longitude	Latitude
Bus	Bus stop on a major transit line	-86.16168	39.77285
Dentist	School of Dentistry	-86.17895	39.77467
Doctor	Pediatrician’s office	-86.17496	39.77926
Admin	Administrative building	-86.17433	39.77459
Hospital	Campus hospital	-86.17557	39.77567
Lime	Bus stop on campus	-86.18376	39.77473
Parking	Campus parking lot	-86.18121	39.77882
School	School of Art and Design	-86.17148	39.77148
University	University lecture hall	-86.17575	39.77271