CSE 519 -- Data Science (Fall 2018)

Prof. Steven Skiena

Homework 2: Exploratory Data Analysis in iPython

Due: Tuesday, September 25, 2018 (8:30 AM)

This homework will investigate doing exploratory data analysis in iPython. The goal is to get you fluent in working with the standard tools and techniques of exploratory data analysis, by working with a data set where you have some basic sense of familiarity.

This homework is based New York Taxi Fare Prediction on Kaggle, revolving around predicting the fare of a taxi ride given a pickup and a drop off location. More than just data exploration, you must also join the challenge and submit your model before the deadline, to get a score feedbacked from Kaggle. You are to explore the data and uncover interesting observations about the New York Taxi operations. You will need to submit all your results in a single google form and your code files in three different format (.ipynb, .pdf and .py). Make sure to have your code documented with proper comments and the exact sequence of operations you needed to produce the resulting tables and figures. The submission steps have been discussed below.

Data downloading

First of all, you need to join the challenge and download the data here. The description of the data can also be found at this page.

Python Installation

Instead of installing python and other tools manually, we suggest to install **Anaconda**, which is a Python distribution with package and environment manager. It simplifies a lot of common problems when installing tools for data science. More introduction can be found at here. Installing instruction can be found here. A useful instruction about Anaconda in Youtube can be found here.

If you are an expert of Python and data science, what you need to do is install some packages relevant to data science. Some packages I believe you will definitely use for this homework are as following:

- pandas
- scikit-learn
- numpy
- matplotlib
- seaborn

Tasks (100 pts)

1. Take a look at the training data. There **may be anomalies in the data** that you may need to factor in before you start on the other tasks.
    a. **Clean the data first to handle these issues.**
    b. **Explain what you did to clean the data (in bulleted form).** (10 pt)
2. **Compute the Pearson correlation** between the following: (9 pt)
    a. Euclidean distance of the ride and the taxi fare
    b. time of day and distance traveled
    c. time of day and the taxi fare

    Which has the highest correlation?
3. For each subtask of (2),
    a. **Create a plot visualizing the relation between the variable**s.
    b. **Comment on whether you see non-linear or any other interesting relations**. (9 pt)
4. **Create an exciting plot of your own** using the dataset that you think reveals something very interesting.

    Explain what it is, and anything else you learned. (15 pt)
5. **Generate additional features** like those from (2) from the given data set.

    What additional features can you create? (10 pt)
6. **Set up a simple linear regression model to predict taxi fare**. Use your generated features from the previous task if applicable.
    a. **How well/badly does it work?**
    b. **What are the coefficients for your features?**
    c. **Which variable(s) are the most important one?** (12 pt)
7. Consider **external datasets that may be helpful to expand your feature set.**
    a. **Give bullet points explaining all the datasets** you could identify that would help improve your predictions.
    b. If possible, **try finding such datasets online to incorporate into your training**.
    c. **List any that you were able to use in your analysis**. (10 pt)
8. Now, try to **build a better prediction model that works harder to solve the task.**
    a. Perhaps it will still use **linear regression but with new features**.
    b. Perhaps it will **preprocess features better** (e.g. normalize or scale the input vector, convert non-numerical value into float, or do a special treatment of missing values).
    c. Perhaps it will **use a different machine learning approach** (e.g. nearest neighbors, random forests, etc).
    d. Briefly **explain what you did differently here versus the simple model**.
    e. **Which of your models minimizes the squared error**? (10 pt)
9. **Predict all the taxi fares for instances at file "sample_submission.csv"**.
    a. **Write the result into a csv file and submit it to the website.** You should do this for every model you develop.

b. **Report the rank, score, number of entries, for your highest rank. Include a snapshot of your best score on the leaderboard as confirmation.** (15 pt)

Be honest. This is your first modelling experience, and I am hoping to see you learned something, not just where you are ranked on the leaderboard.

Rules of the Game

This assignment must be done **individually by each student**. It is not a group activity.

1. If you do not have much experience with **Python and the associated tools**, this homework will be a substantial amount of work. Get started on it as early as possible!
2. All of your **written responses will be submitted through a form during submission**. It may make sense to **keep your answers inside your notebook and copy it over into the form when you are ready to submit.**
3. We will discuss topics like linear regression in detail only after the HW is due. Muddle along for now, and we will understand the issues better when we discuss them in the course.
4. To ensure that you are who you are when submitting your models, have your Kaggle profile show your face as well as a Stony Brook affiliation.
5. There are some public discussions and demos relevant to this problem on Kaggle. It is okay for students to read these discussions, but they must write the code and analyze the data by themselves.
6. Our class Piazza account is an excellent place to discuss the assignment. Check it out at piazza.com/stonybrook/fall2018/cse519.

Submission

Submit everything through **Google classroom.** As mentioned above, you will need to upload:

1. **The Jupyter notebook all your work is in (.ipynb file)**
2. **Python file (export the notebook as .py)**
3. **PDF (export the notebook as a pdf file)**

For everything else, you will fill out a separate Google Form. These will include the responses to all of the task questions above. You will also need to link your Kaggle profile. It is recommended that you have a look at the Google response form for the questions asked and write all your responses in a local document. Once you feel comfortable with your responses, you may record your final responses in the form and then submit.

Task 1:

1) Explain what you did to clean the data. List each step/method as a separate item.

I performed basic data cleaning which involved:

1. Checking for missing values and removing the rows corresponding to them if any.
2. Removing the rows that have -ve values of Fare amount as Fare cannot be -ve.
3. Truncating longitude and latitude values to fit NY coordinates.
    - As the task is specific to New York City Taxi Fare Prediction, the Longitude and Latitude coordinates should not go far beyond that of New York City.
    - I used the below link to obtain the Boundary Coordinates of New York which are:
      North Latitude: 40.917577
      South Latitude: 40.477399
      East Longitude: -73.700272
      West Longitude: -74.259090
      https://www.mapdevelopers.com/geocode_bounding_box.php
    - I used these coordinates to build a boundary for the allowed Drop_off and Pick_up Longitude, Latitudes.
    - This method of data cleaning did not produce substantial results. On analysing the possible reasons for the bad behaviour of the model, I felt that the narrow possible values for the Longitude and Latitude could be a reason.
    - To verify the same, I checked the boundaries of the Longitudes and Latitudes on the Test Data and found that quite a few samples had coordinates outside the above used boundary.
    - In order to cater to this issue, I used the Test Data to obtain the boundaries on the Latitude and Longitude which came out to be :
      North Latitude: 41.709555
      South Latitude: 40.573143
      East Longitude: -72.986532
      West Longitude: -74.263242
    - Hence, I created the Boundary using the newly obtained coordinated from the Test Dataset.

Task 2:

2.1) Pearson correlation between Euclidean distance and the taxi fare

- 0.8257585563383683

2.2) Pearson correlation between time of day and distance travelled

- -0.030505480979840977

2.3) Pearson correlation between time of day and the taxi fare

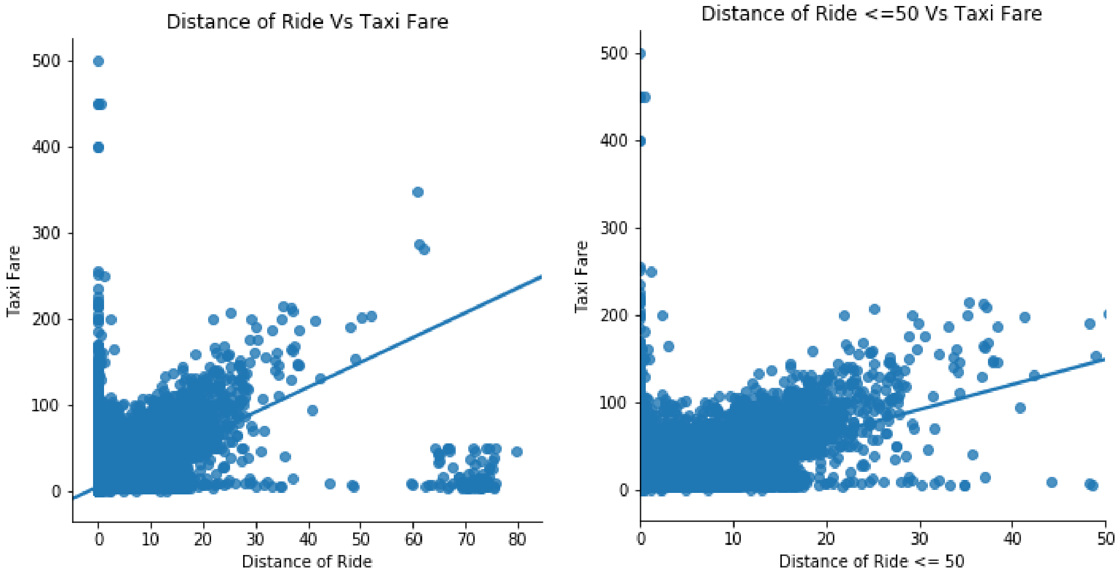
- -0.01927381911008901

2.4) Which has the highest correlation?

- Euclidean distance and the taxi fare - highest
- time of day and distance travelled
- time of day and the taxi fare

Task 3:

3.1.1) Visual plot depicting the relationship between the distance of the ride and the taxi fare



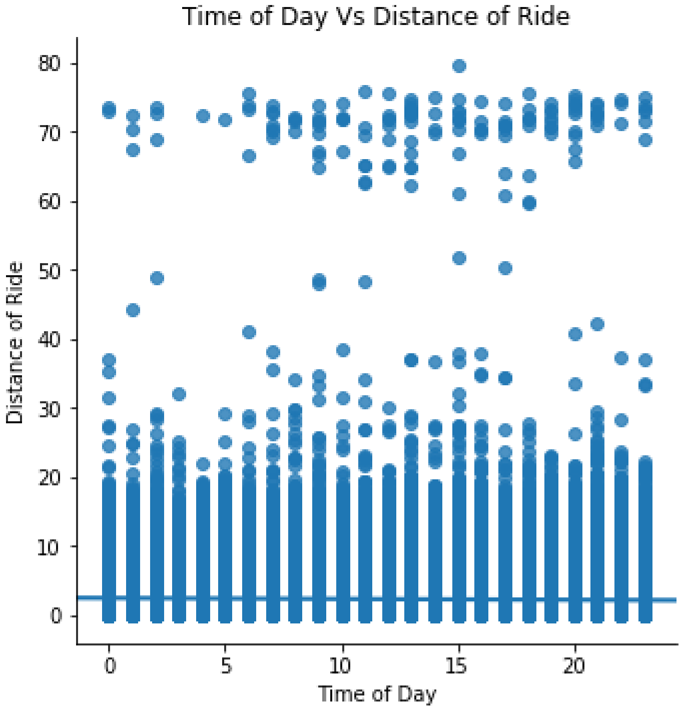
-

3.1.2) Comment on whether you see non-linear or any other interesting relations (based on 3.1.1)

- In an ideal scenario, we can say that the Distance Travelled and Taxi Fare should be linearly related. But, it is very obvious that there are a lot of external factors that affect the relation between the two and hence we cannot say that they are completely linearly related.
- Although, as we can see in the attached plot there does exist a nearly linear relationship between the two for a lot of data samples. This is more clear in the Right hand side plot which caps the distance travelled to 50.
- One interesting aspect I think could be a depiction of the above plot is that the line describing the data is quite below the y=x line, i.e the fare for higher distances does not increase in proportion to how it does for the distances clustered around the centre.
- This is evident In the left hand side plot where we can see a cluster of data samples at very high distances (60-80) but with quite less fare. This (as previously discussed in class) could be due to the Trips to the Airport which are quite long, but have fixed prices.

- Also, as we know Pearson Correlation depicts the strength of linearity between different parameters. A Pearson Correlation of 0.8257585563383683 does indicate quite a strong linear correlation between the Distance Travelled and Taxi Fare.

3.2.1) Visual plot depicting the relationship between the time of day and distance travelled

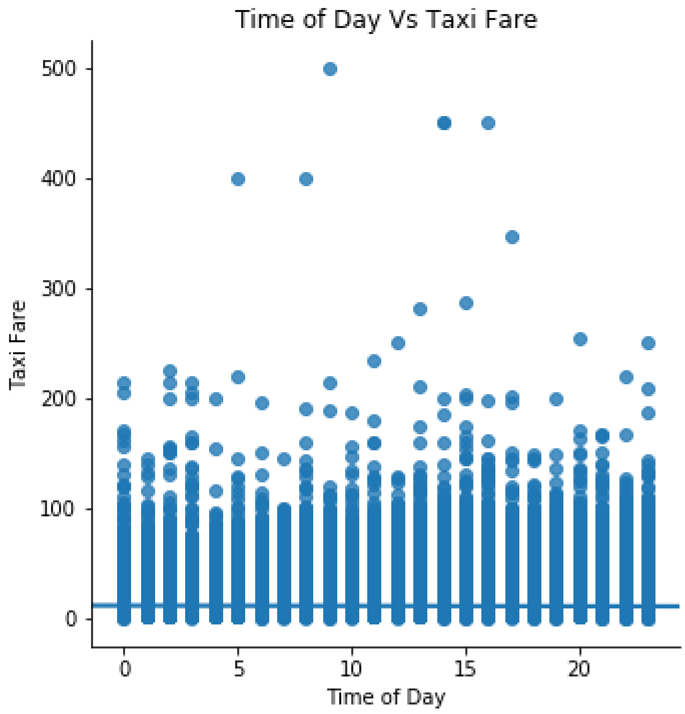


-

3.2.2) Comment on whether you see non-linear or any other interesting relations (based on 3.2.1)

- Based on the attached plot, I see that Time of the Day and Distance travelled are not quite linearly related.
- The Pearson Correlation between the two is very small (-0.030505480979840977) which means that they aren't related much.
- Although, one interesting thing to observe is that there is a range of distances that are mostly never taken despite any part of the day, like the distances between 40-60 are very less taken. And in my assumption, I think that the distances below that could indicate the daily commuters from home to office or vice versa, and the above half could indicate the ones who travel to the airports which tend be quite far away.
- The average distance travelled may be high at the centre (i.e through the day than early or late night) but apart from this, there seems to be no definite linear relationship between time and the distance travelled.

3.3.1) Visual plot depicting the relationship between the time of day and the taxi fare



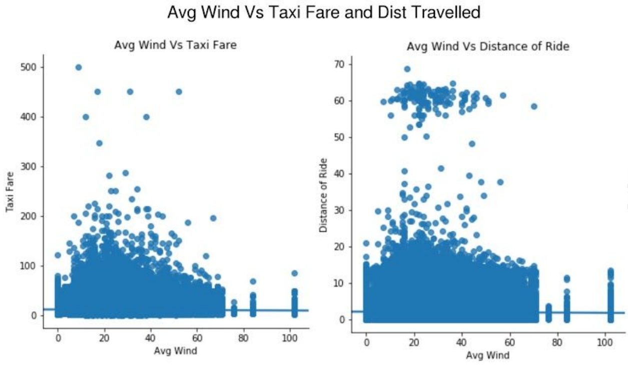
-

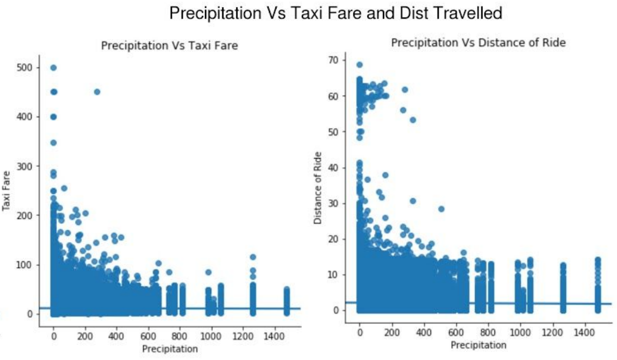
3.3.2) Comment on whether you see non-linear or any other interesting relations (based on 3.3.1)

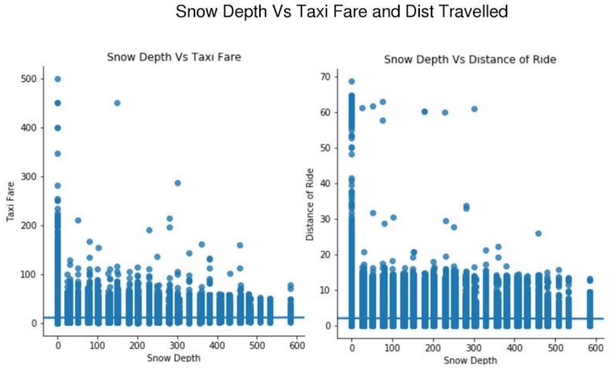
- In similar line with the above explanation, I see that there does not exist that strong a linear relationship between Time of the Day and Taxi Fare either.
- These are also share a very small Pearson Correlation () indicating that they are not much related as well.
- The interesting aspect about this plot are the data points that lie at high Taxi Fares in the middle of the day, and some at the end of the day. With an assumption that most travel to the airport happens either at early or late hours and with the fact that the fare for these trips is fixed, these anomalous data points would not be of the trips to the airport.
- Instead I feel that the reason for the high fare during midday and later could be due to the increased traffic during that time that could have lead to less available cabs with surged prices, thus leading to some really high data points.
- Apart from that, these two are not as strongly related to each other.

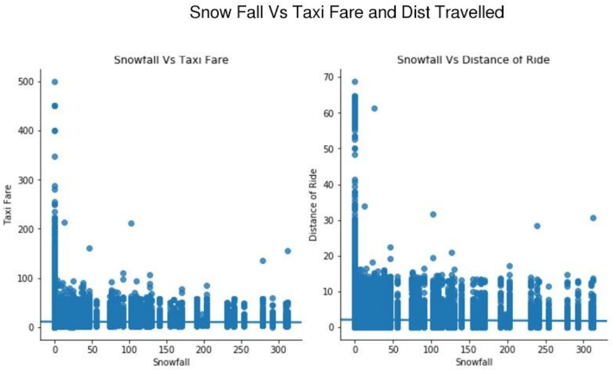
Task 4:

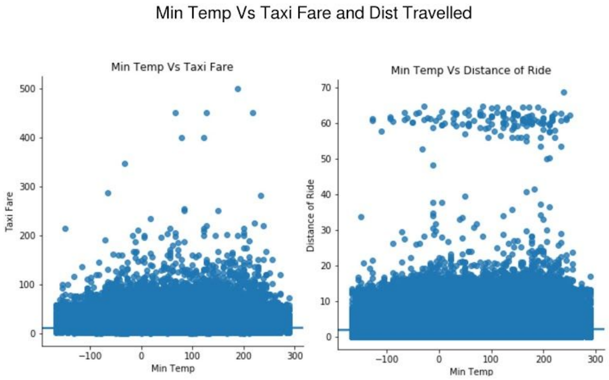
4.1) An exciting plot of your own using the data set that you think reveals something very interesting

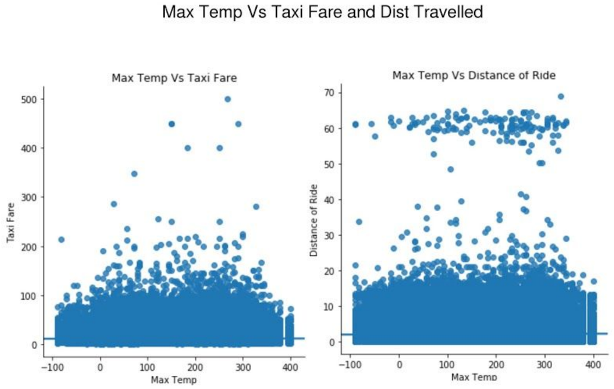












-

4.2) Explain what it is, and anything else you learned. (based on 4.1)

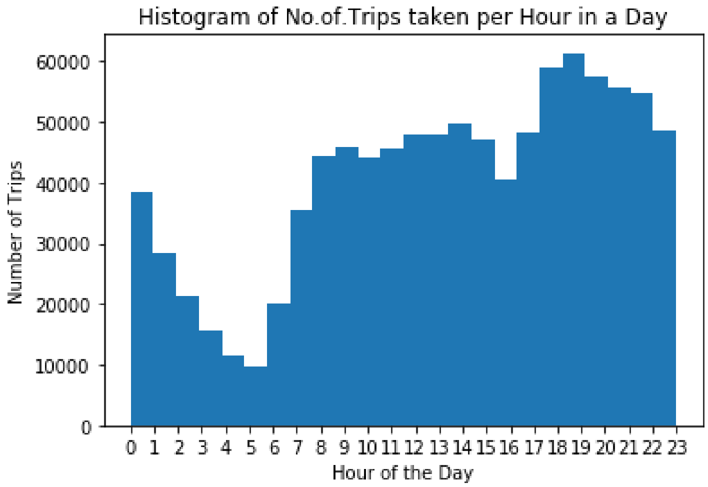
- The above plots indicate the New parameters that I have used from an External New York Weather Dataset (explained in detail in below questions).
- The parameters are ['avg_wind', 'max_temp', 'min_temp', 'precipitation', 'snow_depth', 'snowfall']. I plotted all these variables against Distance Travelled and also the Taxi Fare to

see how they were related to the two and also to check whether these features would help improve my existing model or not and I found some interesting observations.

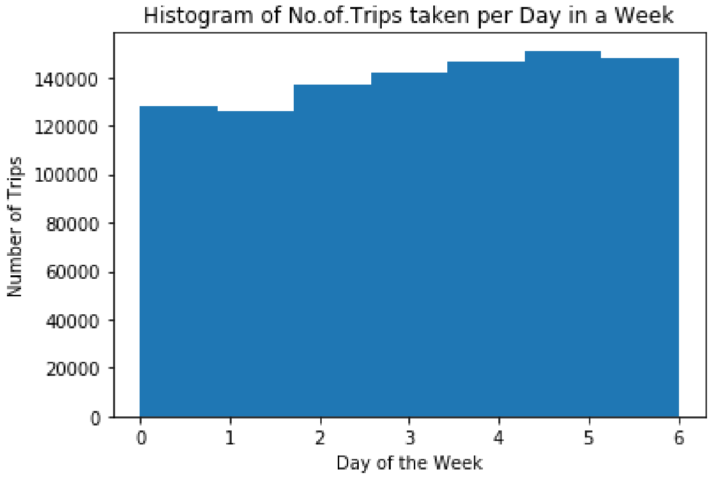
  - Among all the above plots, one common observation I can make is that there is quite a clear relation between the Distance travelled and Taxi Fare, because for the points at the top in the left plots, there are corresponding samples in the right plot as well indicating that Trip Distance and Taxi Fare are indeed correlated.
  - Now comparing the features one by one:
  - Avg Wind - We can clearly see that for higher winds the distance travelled is lesser, thus in turn leading the taxi fare to be lesser.
  - Precipitation - Similar to Avg Wind, the distance travelled and the fair are low when Precipitation is high.
  - Snow Depth and Snow Fall - Here surprisingly I initially expected a decrease in both the rides and the fairs as I thought the availability of cabs would become lesser. But I see that there is quite a constant curve which could mean that during the snow, people continue to use the cabs all the more due to the harsh weathers.
  - Min Temp and Max Temp - This plot was very interesting indicating perfectly that the distance travelled and taxi fares are high when the temperatures are relatively bearable and they reduced when they go under a certain temperature.

- Apart from the above graph, I have also plotted the below which are available in the Jupyter Notebook.

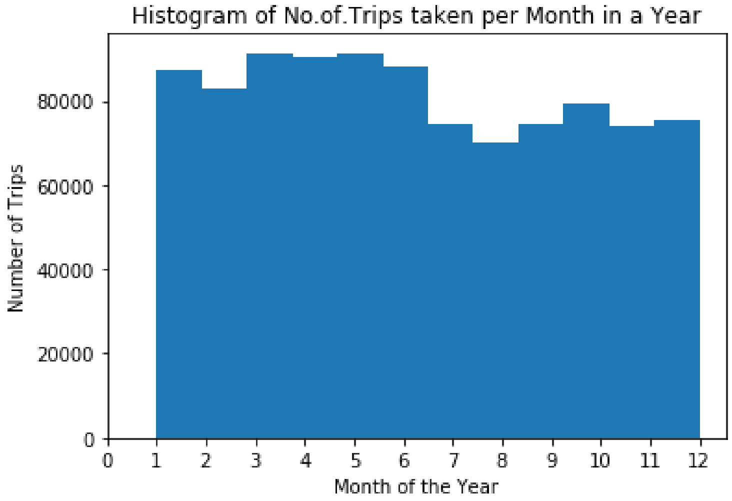
  - Histogram between No.of.Trips and Pick up Hour



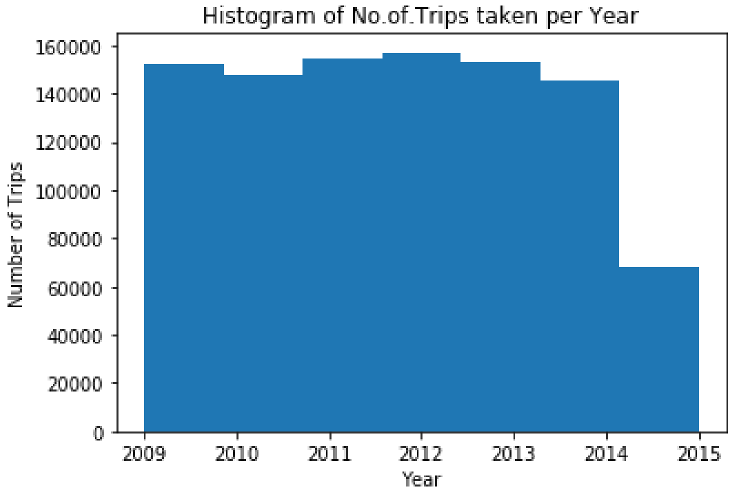
  - Histogram between No.of.Trips and Pick up week



- ○ Histogram between No.of.Trips and Pick up Month



- ○ Histogram between Avg Fare and Pick up Year



Task 5:

5) What additional features can you create? List each feature as a separate item.

- From the pickup_datetime (1-1-2009 to 6-30-2015) timestamp in the dataset, I have extracted the following
    - Pick up Year - int64 - Ranges from 2009 to 2015
    - Pick up Month - int64 - Ranges from 1 - 12 (indicating the month of the year)
    - Pick up Day - int64 - Ranges from 1 - 31 (indicating the day of the month)
    - Pick up Day of Week - object - Ranges from ‘Monday’ to ‘Sunday’ (indicating the day of the Week)
        - This has also been changed to an integer to indicate from 0-6 for each day.
    - Pick up Hour - int64 - Ranges from 0 to 23 (indicating the hour of the day)
- Apart from the above I also extracted the Euclidean and Haversian Distance based on the given Pickup, Drop Off Longitudes and Latitudes.
- These were the extra features that I extracted from the already available dataset that was given to work with. I have discussed about the features extracted from external datasets in Que 7.

Task 6:

6) What are the coefficients for your model? What variable(s) are the most important one?

- The coefficients for my model including the features that have been extracted using the external set are as follows for the following features.

['avg_wind', 'dropoff_latitude', 'dropoff_longitude', 'hav_distance', 'max_temp', 'min_temp', 'passenger_count', 'pickup_day', 'pickup_hour', 'pickup_latitude', 'pickup_longitude', 'pickup_month', 'pickup_year', 'precipitation', 'snow_depth', 'snowfall']
where ['avg_wind', 'max_temp', 'min_temp', 'precipitation', 'snow_depth', 'snowfall'] have been extracted from the external New York Weather dataset that I have considered. (explained in more detail in Que 7)

Coefficients: obtained are as follows:

```
[[ 1.60929997e-04 -1.36371540e+01  1.07266298e+01  3.16495014e+00
  -1.10849266e-03  4.61756423e-04  3.70720804e-02  1.90155047e-03
   9.94131511e-03 -2.05127099e+01  1.40138653e+01  7.91423303e-02
   5.36076031e-01  2.81981289e-04  1.78000353e-05 -9.99671939e-04]]
```

This is the order of importance of the feature based on its coefficient. As this is linear regression , higher coefficient would mean that that particular feature is of more importance.
'pickup_longitude' : 1.40138653e+01
'dropoff_longitude' : 1.07266298e+01
'hav_distance' : 3.16495014e+00
‘pickup_year' : 5.36076031e-01

'pickup_month' : 7.91423303e-02
'passenger_count' : 3.70720804e-02
'pickup_hour' : 9.94131511e-03
'pickup_day' : 1.90155047e-03
'min_temp' : 4.61756423e-04
'precipitation' : 2.81981289e-04
'avg_wind' : 1.60929997e-04
'snow_depth' : 1.78000353e-05
'snowfall' : -9.99671939e-04
'max_temp' : -1.10849266e-03
'dropoff_latitude' : -1.36371540e+01
'Pickup_latitude' : -2.05127099e+01

The coefficients for my model including the features that have been extracted using the external set are as follows for the following features.

['avg_wind', 'dropoff_latitude', 'dropoff_longitude', 'hav_distance', 'max_temp', 'min_temp', 'passenger_count', 'pickup_day', 'pickup_hour', 'pickup_latitude', 'pickup_longitude', 'pickup_month', 'pickup_year', 'precipitation', 'snow_depth', 'snowfall']

> where ['avg_wind', 'max_temp', 'min_temp', 'precipitation', 'snow_depth', 'snowfall'] have been extracted from the external New York Weather dataset that I have considered. (explained in more detail in Que 7)

Coefficients: obtained are as follows:

```
[[ 1.60929997e-04 -1.36371540e+01  1.07266298e+01  3.16495014e+00
  -1.10849266e-03  4.61756423e-04  3.70720804e-02  1.90155047e-03
   9.94131511e-03 -2.05127099e+01  1.40138653e+01  7.91423303e-02
   5.36076031e-01  2.81981289e-04  1.78000353e-05 -9.99671939e-04]]
```

And, this is the order of importance of the feature based on its coefficient. As this is linear regression, higher coefficient would mean that that particular feature is of more importance.

'pickup_longitude' : 1.40138653e+01
'dropoff_longitude' : 1.07266298e+01
'hav_distance' : 3.16495014e+00
‘pickup_year' : 5.36076031e-01
'pickup_month' : 7.91423303e-02
'passenger_count' : 3.70720804e-02
'pickup_hour' : 9.94131511e-03
'pickup_day' : 1.90155047e-03
'min_temp' : 4.61756423e-04
'precipitation' : 2.81981289e-04
'avg_wind' : 1.60929997e-04

'snow_depth' : 1.78000353e-05
'snowfall' : -9.99671939e-04
'max_temp' : -1.10849266e-03
'dropoff_latitude' : -1.36371540e+01
'Pickup_latitude' : -2.05127099e+01

How well or bad did the model do? Make sure to discuss your error metrics.

- I have used 3 Error Metrics to evaluate my models.
    - Root Mean Square Error (RMSE)
    - Mean Squared Error (MSE)
    - R2 Score
    - And the goal is to try to maximize R2 score (1 being the best) while minimizing RMSE and MSE along with ensuring that the model does not overfit.
- My Linear regression model using the above features achieved an RMSE = 5.255454, MSE = 27.62 and R2 score = 0.70.
- For a basic linear model (that could be considered as the baseline model) I believe that it did quite okay providing an RMSE = 5.60765 on the Test Data. But, this is definitely not a good model and has to be improved upon.
- Also, I think for data as sparse and vast as that we have, we may not be able to model it linearly as it involves a lot of real world parameters and in general might contain lots of outliers along with many other underlying dependencies among the features. Hence, linear regression did not produce that good a result for the problem.
- Improvisations upon the baseline model have improved the RMSE score for Test Data drastically and this has been discussed further in Que 8.

Task 7:

7.1) Give bullet points explaining all the data sets you could identify that would help improve your predictions

- The NYC TLC Trip Data from http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
    - This website contains data about Yellow and Green Cabs since 2009 monthwise for every day until now.
    - And they contain very detailed information about the Total Charges for the customer like Fare_amount, Extra, MTA_Tax, Improvement_Surcharge, Tip_Amount and Tolls_Amount.
        - http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf
        - http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf
    - They also contain information about the pickup date and time when the meter was engaged, and this could have been used to merge this data to our existing data based on the timestamp of the pickup.

    - The no.of data points available in each single month of a particular year is enormous (e.g 14092415 data point for January 2009). Hence, this plethora of data on the division of the taxi price could definitely help improve our prediction model.
- Weather Dataset from https://www.ncdc.noaa.gov/cdo-web/datasets
    - This website contains the Daily Summaries of the Weather in required cities for the required period of time. And they have the data available from 1869-01-01 until so far.
    - They provide large set of features like Snowfall, Snow Depth, Min Temperature, Max Temperature, Precipitation, Air Wind etc to name a few.
        - https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/GHCND_documentation.pdf
    - Along with the above details regarding the weather, they provide the year/date/month on which that particular weather was observed. Hence, this can be used to merge this data with our existing data.
    - Weather plays a huge role in the movement of taxis. The availability, prices, and demand everything can change based on the weather. Hence, incorporating details about the weather into our dataset would definitely boost our model.

7.2) List any external data sets that you were able to use in your analysis. (in bulleted form)

- I have used the Weather Dataset from January 1st 2009 to November 11th 2015 (as our dataset spans from January 1st 2009 to June 30th 2015) in my analysis.

Task 8:

8.1) Did you use any new features from an external dataset? If so, source the location and name them.

- Yes, I have used the following new features from the external Weather Dataset.
    - ‘Date’
    - 'Precipitation'
    - 'Snow Depth'
    - 'Snowfall'
    - 'Maximum Temperature'
    - 'Minimum Temperature'
    - 'Average Wind'

I have referenced the following link for information regarding the Weather Dataset
https://sdaulton.github.io/TaxiPrediction/
And have taken the above data from
https://raw.githubusercontent.com/sdaulton/TaxiPrediction/master/data/nyc-weather-data.csv
(that was provided in the above github.io link)

8.2) Did you preprocess the features further? If so, what did you do?

- Yes, I have again performed basic Pre-processing similar to before.
    - In this dataset -9999 indicate missing values, hence replaced such values with 0
    - Parsed the 'Date' feature to extract the following in order to be able to merge this data to our existing data.
        - 'pickup_year'
        - 'pickup_month'
        - 'pickup_day'
    - The provided data was initially parsed to have the above fields. Hence, using these 3 fields, the initial data has been merged with this external Weather Data.

8.3) Did you try different machine learning models? List all that you tried and which performed the best.

- Yes, apart from the Linear Regressor I tried various other Machine Learning Models like:
    - Decision Trees
    - Random Forest Regressor
    - XGB Regressor (without any Hyperparameter Tuning) and
    - XGB with Hyperparameters
- Among all the above, XGB with Hyperparameters worked the best providing a score of 3.17502 on the Test data. (which is a very high improvement on the Linear Model that was done initially which got a score of 5.60765)
- Also, I tried splitting the data in different proportions to obtain the Train and Validation set and a split of 2/3rd Train and 1/3rd Validation Set along with XGB with Hyperparameters worked the best.
- Instead of manually splitting the data every time, I also performed K-Fold Cross Validation for the different modules. But I did not see significant betterment in the model through this.

8.4) Did you note any improvements from any of the changes you made above? Elaborate your thoughts on any improvement or lack thereof.

- I noted significant change in the performance of my models as I went through from a simple Linear Model to a complex XGB Booster.
- The RMSE score for the different models on the Test Data was as follows:
    - Linear Regression - RMSE score = 5.60765
    - Decision Trees - RMSE score = 5.88059 (was worse than Linear Model)
    - Random Forest Regressor - RMSE score = 3.64299 (drastic improvement from here on)
    - XGB Regressor without parameters - RMSE score = 3.55638
    - XGB Regressor with Parameters - RMSE score after tuning the parameters few times = 3.17502
- My understanding from this practice mainly was that real world datasets (considering most dataset are actually real world) are extremely complicated and visualizing this kind of data in order to understand the data better requires efforts. In other words, a simple

linear model is not always sufficient to best model a data BUT should definitely be the way to start as it gives us a better understanding into the data spread and feature dependencies.

- I also understood the concept of 'Overfitting' a lot better as I went through multiple training attempts where my Training RMSE was quite impressive but it did not work well on the Test Data Set. Learning Regularization techniques to solve these issues is one thing that I am looking forward to.
- Also, I ran the test on the entire dataset using XGB (without parameters) and it performed way worse than when I consider 1000000 sample (for all my above experiments). I believe this could be due to the fact that there could have been more noise/outliers in the whole set, or that too much data let to extreme overfitting as Train data was 55M while test data was only 9414.

Task 9:

9.1) Report your best achieved rank

9.2) Report the score you received for your best rank

9.3) Report the total number of entries you made during the course of this challenge

9.4) Include a snapshot of your best score on the leader board as confirmation

- ADD FILE