<https://www.kaggle.com/shireennagdive>

Task1: Explain what you did to clean the data. List each step/method as a separate item.

I have performed the following steps to clean my data:

1. Removed all rows having observations with missing entries (NaN).
2. Removed all rows having 0 as a column value since any of the column values in the data set cannot be zero.
3. To optimize storage in memory, changed datatypes of columns in the dataframe to following:

{'fare\_amount': 'float32',

'pickup\_longitude' : 'float32',

'pickup\_latitude': ‘float32',

'dropoff\_longitude': 'float32',

'dropoff\_latitude': 'float32',

'passenger\_count': 'uint8'

}

1. Removed fare amounts which are less than $2.50 as it is the minimum fare for a taxi in NYC.
2. Removed fare amounts which are greater than $100.00 as maximum value of fare within NYC cannot exceed $100.00
3. Removed pickup\_latitude, pickup\_longitude, dropoff\_longitude, dropoff\_latitude which are outside New York City. I took the center of NYC area and then added 100 miles in four directions as the boundary.
4. Removed trip distance(haversine) travelled less than 0.5 miles.
5. Removed trip distance travelled greater than 40 miles as it will be outside New York City.
6. Removed passenger counts greater than 6 as 6 is the maximum number of people who can travel.

Task2: Distance and fare: 0.90282

Time of the day and distance travelled: -0.03429

Time of the day and Taxi fare: -0.0200

Task3: Plot:

1. Distance of the ride and taxi fare

According to my analysis, distance of the ride and taxi fare is linear in relationship. The fare increases as the distance increases which is also obvious as fare directly depends on distance travelled.

1. Time of the day and distance travelled

There is not much co-relation here. There is one interesting observation that more distance is covered in night hours which is obvious due to less traffic. Also, 4 to 5 p.m. is the traditional hour for cabs to change shifts, I think these drivers might prefer to pick up passengers who want to travel long distances. Therefore, I can expect a higher distance covered in the shift hours.

1. Time of the day and taxi fare: There is not much correlation. Since I expect a higher distance to be covered in the shift hours, and distance is directly proportional to fare, the fare is maximum during the traditional hour for cabs to change shifts i.e 4-5 p.m.

Task4

Apart from fares being correlated to only time of the day, I observed that fares are correlated to day of the week as well. The distance travelled is maximum on Sunday. On weekends, especially Sunday, most of the people have holidays at their work. Hence, they might travel more on weekends with their families in NYC. Also, college students have holidays on Sundays, so I expect them to contribute to the number of taxi rides on Sundays.

Task5: I have created the following additional features:

1. Latitude Difference
2. Longitude Difference
3. Trip Distance
4. Day of the week
5. Pickup Hour
6. Pickup Date

Task 6:

Intercept -422.2146

Trip Distance: 5.8863

Pickup Longitude: 4.3611

Pickup Latitude: 7.391

Dropoff Longitude: -12.0981

Dropoff Latitude: -10.9935

Passenger Count: 0.0386

Absolute Latitude Difference: -160.592

Absolute Longitude Difference: -49.3634

Pickup Hour: 0.0144

The variable Absolute Latitude Difference is the most important one. It means that if all other parameters are constant, then the difference between the pickup and drop off latitude decreases the fare by a value of 160.592. Also, trip distance is a very important variable. If other parameters are constant except trip distance, the model predicts the fare to increase by a value of 5.88 which is close to the actual charge per mile of a taxi. Similarly, pickup latitude, longitude and drop off latitude, longitude have an effect on the fare. Number of passengers in the taxi does not much affect the fare according to the prediction model.

Task 6.2:

Linear Regression gave a root mean squared error (RMSE) of 3.8035 while Random Forest gives a RMSE of 3.5573. Since the model was very well trained, it performs quite well. Since Random Forest has regularization in-built unlike Linear Regression, it performs better than Linear Regression.

Task7:

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

* Weather Data: Severe weather conditions like snow fall, heavy rains and precipitation can have a huge impact on the trip duration. Also, longer routes might have to be taken due to bad road conditions. Thus, longer trip durations and routes directly affect the fare.
* Accidents: Accidents can have a huge impact on the traffic and hence will increase the trip duration. Also, taxi driver might have to take different routes in case the road is blocked affect the trip fare. Thus, longer trip durations and routes directly affect the fare.
* NYCHolidays: On holidays such as Independence Day, Memorial Day etc. there might be a dip in taxi counts.
* Traffic: Traffic conditions on a particular route directly affects the fare of the trip as more trip time means an increased fare.

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

NYC\_holidays

Task8:

I used the feature holiday in my model from the NYC Holidays dataset. I found this dataset on <https://www.kaggle.com/anirudh796/nycholidays>

It contains holidays for the year 2016,2017 and 2018

I added a new feature in my model named ‘holiday\_or\_not’ and set it equal to 1 or 0 based on the corresponding date in my training set. 1 means that date is a holiday based on the date present in the holidays dataset and 0 means it is not a holiday.

I tried two machine learning models:

Linear Regression

Random Forest

Random Forest Model performed the best out of the two

Linear Regression gave a root mean squared error (RMSE) of 3.8035 while Random Forest gives a RMSE of 3.5573. Random Forest is able to discover more complex dependencies at the cost of more time for fitting. Decision Trees in Random Forest are handling data and relationships better than the regression model. But since taxi fare has a linear dependency on the parameter Trip Distance, I get almost similar results from both algorithms with a little difference in RMSE

sf3.803dsd344343

3.8 3.8035

T