**INTRODUCTION**

**Predict Cab fare Price**

The problem statement designing a system that predicts the fare amount for a cab ride in the city is a classic case of regression. The goal of this project is to predict the fare of a cab ride given information about the pickup and drop off locations, pickup date and time and number of passengers travelling.

In this project we aim to clean the data, visualize different relationship between variables and also figure out new features that are better predictors of taxi fare.

**HYPOTHESIS GENERATION**

A list of hypothesis which generates which in our case will affect the fare of a cab ride.

1. Trip Distance: If the trip distance is more, than fare should be higher.
2. Time of Travel: During peak traffic hours, the cab fare may be higher.
3. Day of Travel: Fare amount may differ on weekdays and weekends.
4. Is the trip to/from airport: These trips generally have a fixed fare.
5. Pickup or Drop off Neighbourhood: Fare may be different based on the kind of neighbourhood.
6. Availability: If a particular location has a lot of cabs available, fares may be lower.

**DATA CLEANING AND EXPLORATION**

In this section we discuss various steps used to clean the data and understand the relationship between variables and use this understanding to create better features.

1. **Distribution of Fare Amount**

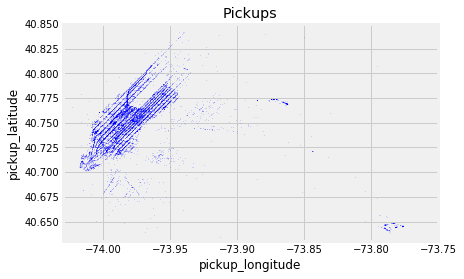
We first looked at the distribution of fare amount and found 3 records where the fare amount was negative. So we excluded the same from the data. Also the fare amount follows a long tail distribution. To understand the data better we take a log transformation after removing the negative values. Also on an analysis we have found that there exists only 9 records where the fare amount is abnormally high and these records are also the only ones greater than 100. So we also exclude them from our data.

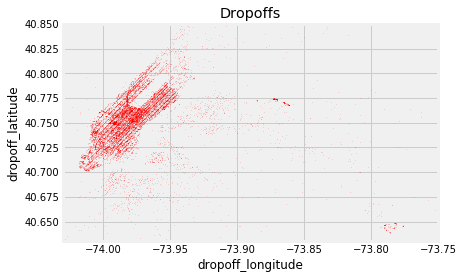
2. **Missing value analysis**

On checking for missing values in the data we found that 25 counts of missing values in Fare amount, 1 count in Pickup Date-time and 55 counts of missing values in Passenger count. As the percentage of missing values is less than 1 % of the data so we have dropped the missing value of Pickup Date-time, imputed 25 cases of Fare amount with Mean of Fare amount and for Passenger count we have replaced it with '0' as these may be cases where the passenger may have booked and cancelled and had to pay for the booking charge.

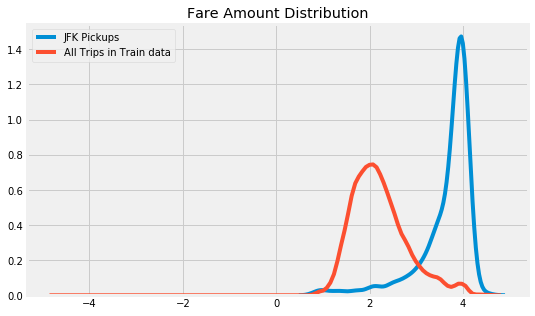
3. **Distribution of Geographical features**

The range of latitudes and longitudes are between 40 to 42 and -72 to -75 respectively. But in the training dataset we observed latitudes and longitudes in the range of 401 and 0 which is not possible. We have found 313 cases where latitude and longitude is 0. Since this data is for cab rides taken in New York (centre latitude and longitude is 40, -74) we remove these rows from our analysis. Such anomalies were not found in "test" data.

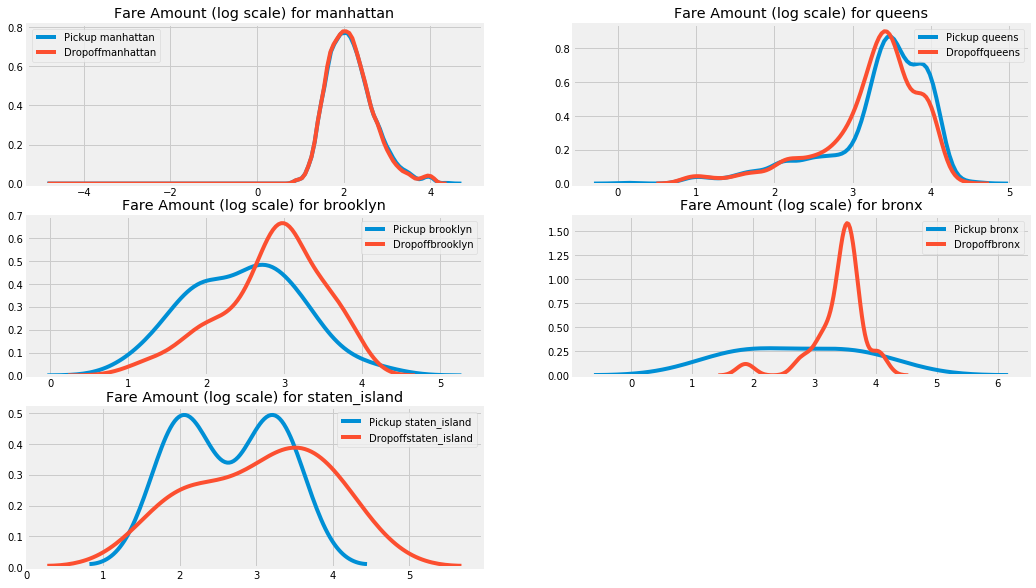




We can see that there is a high density of pickups near JFK and La Guardia airports. We then looked at the average fare amount for pickups and drop offs to JFK, compared to all trips in the train data and observed that fare was higher for airport trips. Based on this observation we created features to check whether a pickup or drop was to or from any of the 3 airports in NYC.

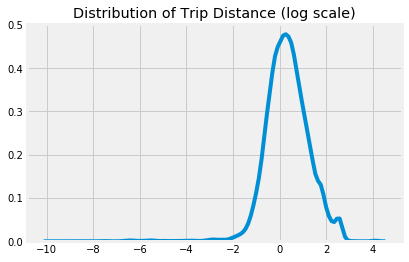


The next step was to check whether our hypothesis of fare from certain neighbourhoods are higher than the rest, based on the 5 boroughs New york city is divided - Manhattan, Queens, Brooklyn, Staten Island and Bronx. Each pickup and dropoff was grouped into these 5 neighbourhoods. Our hypothesis was right except for Manhattan and Queens , for every other neighbourhood there was a difference in the pickup and drop off fare distribution. Also, Queens has a higher mean pickup compared to other neighbourhoods.

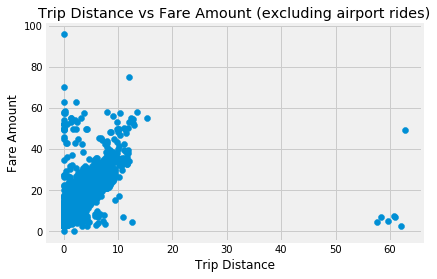
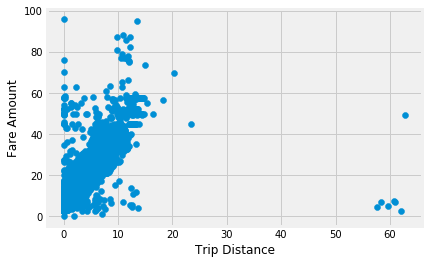


4. **Distribution of Trip distance**

Using the pickup and drop off coordinates we calculate the trip distance in miles . Trip distance just like fare amount follows long tail distribution , we take a log transformation to make it close to normal distribution



One of our hypothesis was that the fare amount should increase with fare distance. A scatter plot between trip distance and fare amount showed that though there is a linear relationship, the fare per mile (slope) was higher, and there were only a few trips whose distance was greater than 50 miles, but fare was very low. To check if this was the case because of airport trips, we removed airport trips and plotted the distribution. We then observed that fare per mile was still lower but a small point was noted where the fare was considerably higher.

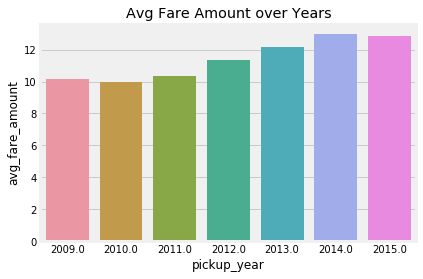


The next step was to see if there was a particular region where the trip distance greater than 50 miles was observed. This showed that these trips are from lower Manhattan.

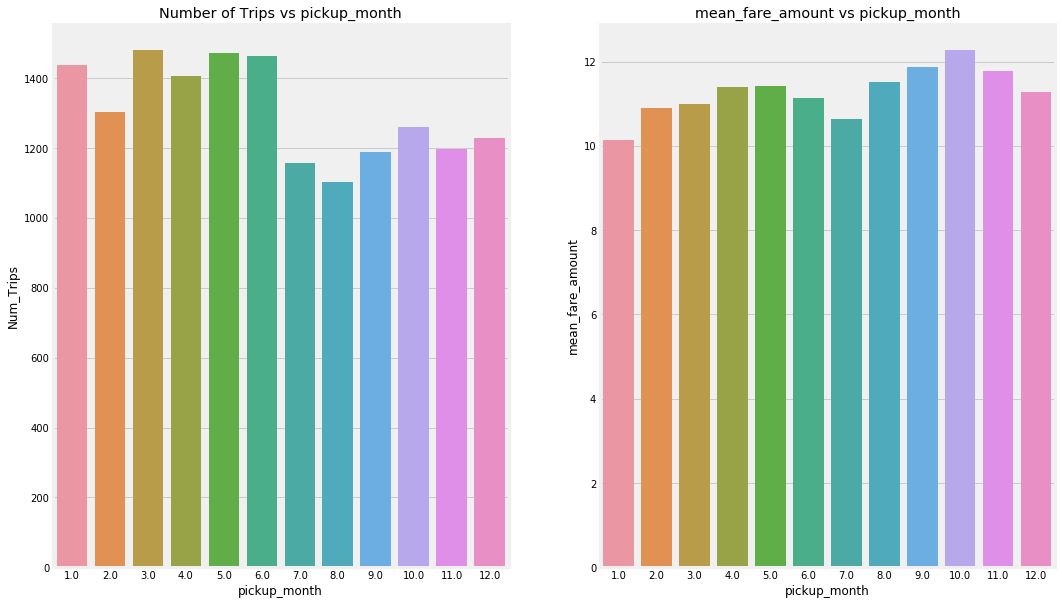
5. **Distribution of Pickup date time.**

The first step to analyse how the fares have changed over time, is to create features like hour, day of the week, day, month, year from pickup datetime.

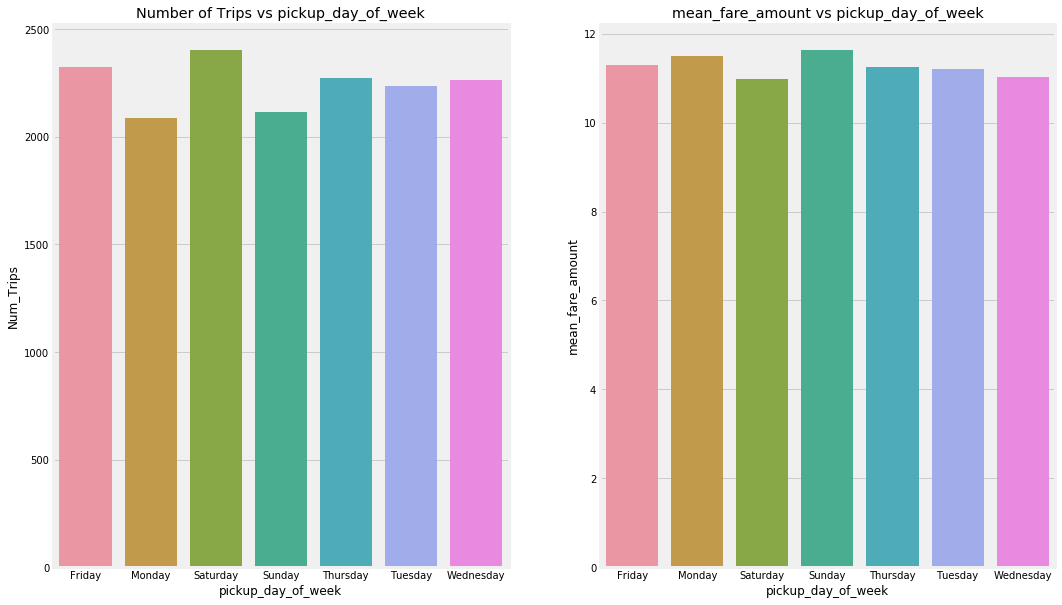
As expected over the years average fare has increased.



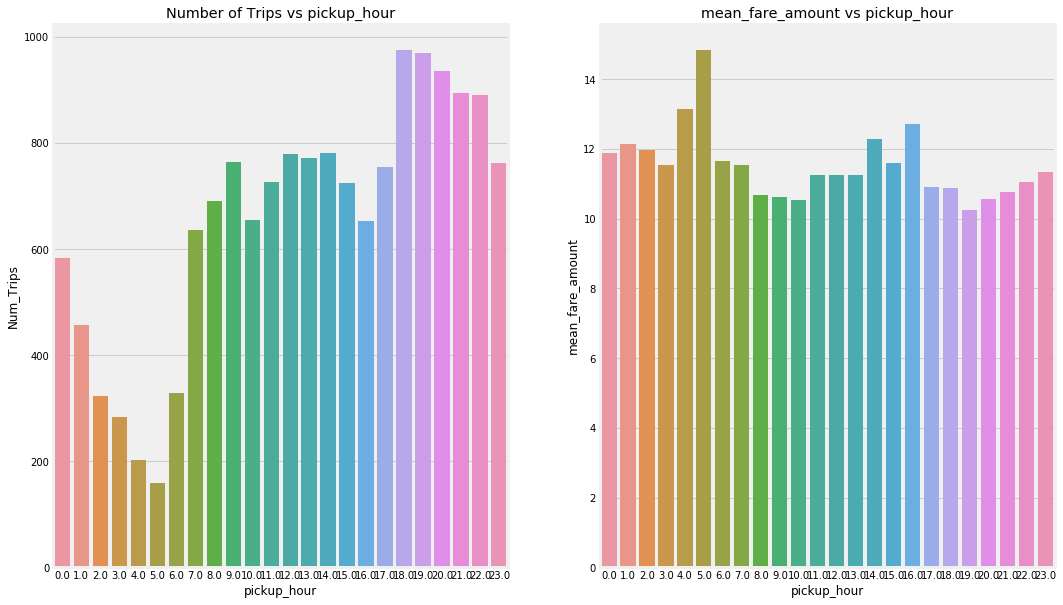
Over months though there has been fewer pickups from July to December, the average fare is almost constant across months.



We observed that though the number of pickups are higher on Saturday, the average fare is lower. On Sunday and Monday though the number of trips are lower, average fare amount is higher.



The average fare amount at 5 am is the highest while the number of trips at 5 am is the least. This is because at 5 am most of the trips are to the airport. The number of trips are highest in 18 & 19 hrs.



**MODEL BUILDING**

1. **Linear Regression:** A good first choice of model for establishing a baseline on a regression task is a simple linear regression. Based on the strength of the relationships between some of the features and the target, we can expect a linear model to do exceptionally well.

The linear regression model trained on different parameters achieved a validation Root Mean Square error of 5.4162.

2. **Decision Tree:** Next we choose Decision tree algorithm to build a model and predict. Decision tree has helped us achieve a Root Mean Square error of 5.5875.

3. **Random Forest:** The Random Forest is a more flexible model than the linear regression which means it has reduced bias. It can fit in the data better. The random forest has a low variance meaning it can generalize to new data. Random Forest outperforms both the above model with a Root Mean Square error of 4.1647.

RMSE (Root Mean Square error ) was used to evaluate the model s as RMSE method is more accurate. By squaring the errors before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is more desirable when large errors are particularly undesirable.

Mean absolute deviation (MAD) and Mean absolute percentage error (MAPE) is the same. The difference between them is that MAPE measures the deviation from actual data in terms of percentage. The similarity between them is that they both measure the absolute error. So both the negative and positive errors cancel out each other.

**NOTE: The analysis and the model reported is best viewed in the Python code block as the code is clear and Python gives us an advantage of using minimalistic functions as most of the functions are predefined in Python and its syntax rules enable developers to build applications with consice and readable code base.**

**On the other hand the code in R is not precise and does not show the proper analysis. As I am a newbie to R platform and this is m first code block in R, I was not able to replicate the exact analysis.**