**Cab Fare Prediction**

**By: - Kushagra Shroff**

**20-02-2020**

* **INTRODUCTION**

Presently a day's taxi rental administrations are extending with the multiplier rate. The utilization of Data Science can assist the endeavour with providing their clients a superior help than previously.

The objective of this test is to anticipate the toll of a taxi trip given data about the pickup and drop off areas, the pickup date time and number of travellers voyaging.

* **PROBLEM STATEMENT**

You are a cab rental start-up company. You have successfully run the pilot project and now want to launch your cab service across the country. You have collected the historical data from your pilot project and now have a requirement to apply analytics for fare prediction. You need to design a system that predicts the fare amount for a cab ride in the city.

* **DATA**

Number of attributes: 6

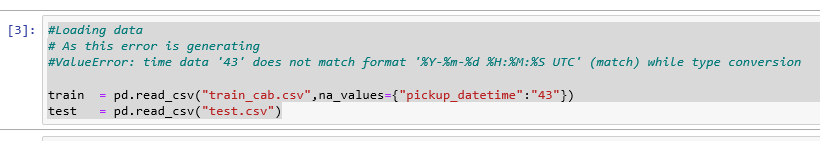
* pickup\_datetime - timestamp value indicating when the cab ride started.
* pickup\_longitude - float for longitude coordinate of where the cab ride started.
* pickup\_latitude - float for latitude coordinate of where the cab ride started.
* dropoff\_longitude - float for longitude coordinate of where the cab ride ended.
* dropoff\_latitude - float for latitude coordinate of where the cab ride ended.
* passenger\_count - an integer indicating the number of passengers in the cab ride.

Missing Values: Yes

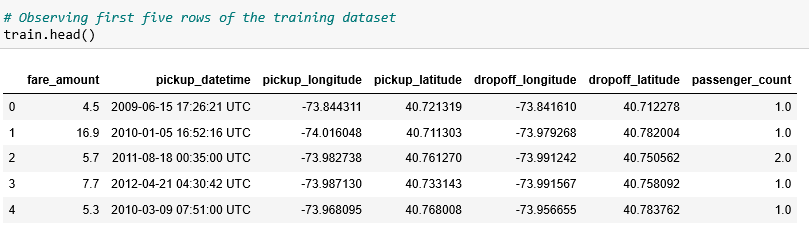
***DESIGN AND IMPLEMENTATION***

1. **Loading the data**

While loading we are also treating an invalid format entry.

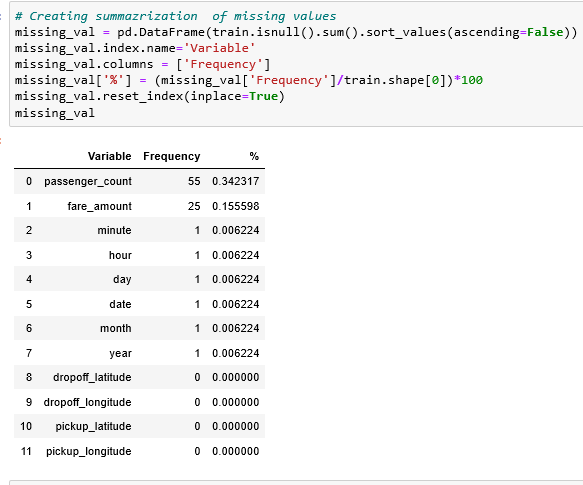


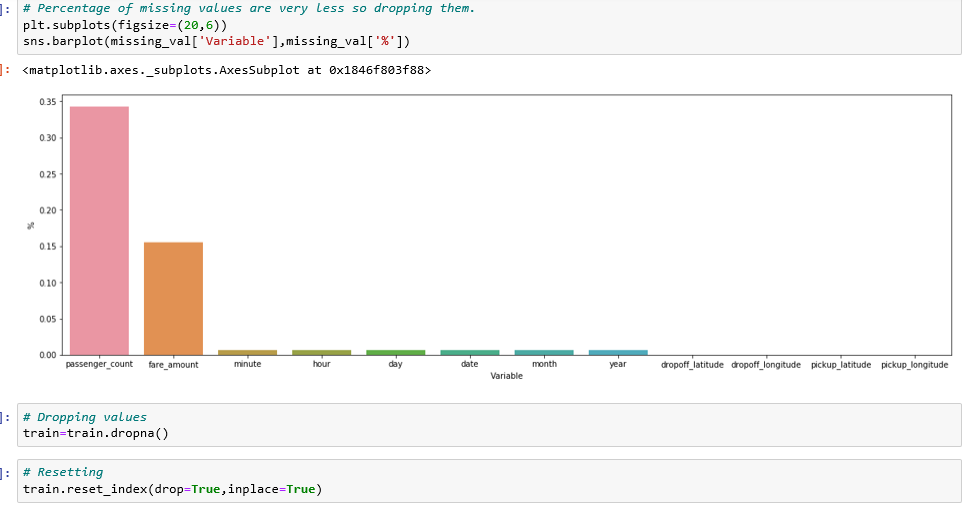
In our case the dataset is already provided, which has 16067 rows and 7 columns.



1. **Data Pre-Processing**
2. ***Missing Value Analysis***

Missing information can happen in view of nonresponse, no data is accommodated at least one things or for an entire unit. Once in a while missing qualities are brought about by the analyst. For instance, when the information assortment is done inappropriately or mix-ups are made in information passage, it is called Human Error.





1. ***Outlier Analysis***

An Outlier is an item that veers off essentially from the remainder of the articles. They can be brought about by estimation or execution mistake. The investigation of exception information is alluded to as anomaly examination or anomaly mining.

A basic method to approach to distinguish exception is to utilize boxplot.

On the off chance that there is a lot of information in anomaly, at that point erasing them isn't a productive way, since we will lose parcel of information.

# 1.> Negative fares as well as ultra-higher fares (fare>434) which are practically impossible will be removed as they are outliers.

# 2.> Null values are already removed from the fare\_amount.

# 3.> Fare can't be zero so also removing that entry.

# 4.> Latitude valid range for norther and southern hemisphere is -90 to 90 degrees.

# 5.> Longitude valid range for northern and southern hemisphere is -180 to 180 degrees.

# 6.> Most of the latitude and Longitude values are identical to NYC coordinates that is 40 N (+)

And 74 W (-) approximately.

# 7.>This data is about taxis running in NYC.

# 8.> Some coordinates have zero as there value which is not practical as we know the location.

# 9.>Also by getting the range of latitudes we come to know about various outliers in data.

# 10.>Some picking and dropping coordinates are same which accounts for useless data will remove them after calculating trip distance.



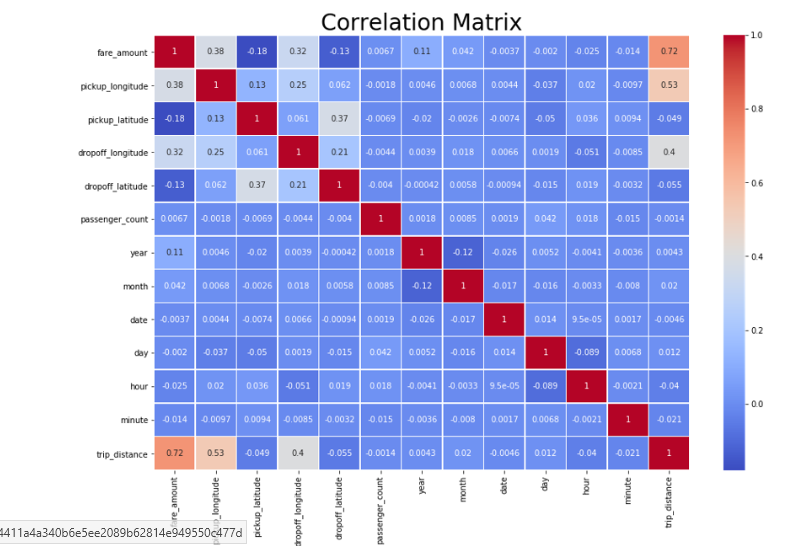
1. ***Feature Selection***

For modelling, we have to select the columns r independent variables that will highly contribute to the dependent variable.

If there is no dependent variable, we have to find the collinearity between all the variables and if there is high collinearity between two variables, we can remove any one variable and keep the other.

# Dropping latitude and longitudes as they are no longer needed as we have calculated distance.

# After extracting date time features dropping that column too.



1. ***Feature Scaling***

Feature Scaling is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

* Rescaling (min-max normalization)
* Mean Normalization
* Standardization (Z-score Normalization)
* Scaling to Unit etc.

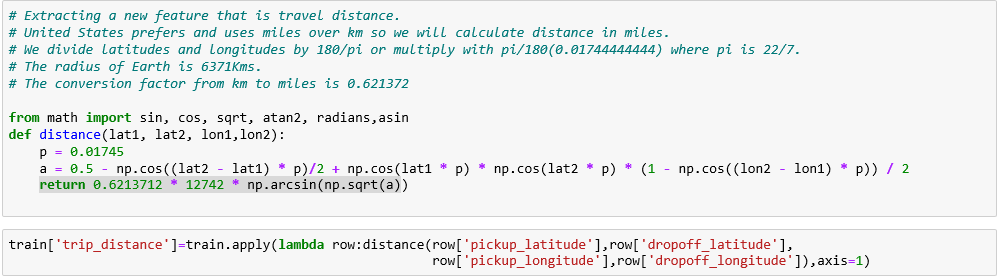
Standardization has to be applied when the data is normally/uniformly distributed that is graph is having a bell shaped curve.

In our case, we are taking the log of variables and again there inverse for meaningful imputations.



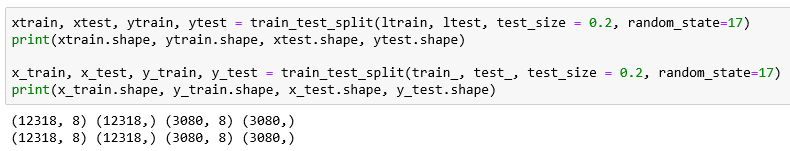
1. ***Feature Extraction***

* Extracting a new feature that is travel distance.
* United States prefers and uses miles over km so we will calculate distance in miles.
* We divide latitudes and longitudes by 180/pi or multiply with pi/180(0.01744444444) where pi is 22/7.
* The radius of Earth is 6371Kms.
* The conversion factor from km to miles is 0.621372

******

1. **Model Creation**

Some of the time the information or the business goals loan themselves to a particular calculation or model. Different occasions the best methodology isn't so obvious. As you investigate the information, run the same number of calculations as you can; think about their yields. Base your decision of the last model on the general outcomes. Here and there you're in an ideal situation running a troupe of models all the while on the information and picking a last model by looking at their yields.



1. **Important Conclusions**

#1.> Solos are taking more rides than other passenger counts which is practical in nature. The cab company can produce attractive offers for solos to capture the market more as well as focus on their other segments of passengers.

#2.> Most of the rides are booked in the evening time duration from 6pm to 8pm which is also the rush hour.

#3.> Most of the revenue is generated from passengers travelling solos or in pair thus the company can introduce attractive offers for travellers in pair like sharing of fare when the route is same to increase revenue.

#4.> There's no boom in the cab market during weekends but somewhat same every day.

**Results** (referred from python code file name pypro1)

1. **Linear Regression**

* Mean Absolute Percentage Error: 18.04347362748669
* Which means the model is 81.95652637251331 percentage correct.

1. **Random Forest**

* Mean Absolute Percentage Error: 20.286595192115932.
* Which means the model is 79.71340480788407 percentage correct.

1. **Decision Tree**

* Mean Absolute Percentage Error: 31.222825443032427.
* Which means the model is 68.77717455696757 percentage correct.

1. **KNN**

* Mean Absolute Percentage Error: 24.93735357401775.
* Which means the model is 75.06264642598225 percentage correct.

1. **Gradient Boosting Regressor**

* Mean Absolute Percentage Error: 17.620068362751855,
* Which means the model is 82.37993163724815 percentage correct.