Contents

Chapter 1
Introduction

1.1 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.

1.2 Data

Our task is to build a regression model which helps us predict the fare amount for a cab ride depending on various variables. As you can see in the table below, we have the following 6 variables, using which we must correctly predict the fare amount.

- 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

Chapter 2
Methodology

2.1 Pre-Processing

Before we start modelling, we need to look at the data, which basically means exploring and cleaning the data to make them ready for consumption by the model. This is often called as **Exploratory Data Analysis**.

Most analysis (for example regression) requires the data to be normally distributed. We plotted the fare amount (shown in figure below) and it is not normally distributed. The probability density function (shown below in Figure 2.1) helps us to visualize the distribution of data as well as if the variable is normally distributed or not. Next, as a part of cleaning the data, we look at the range of each columns and if each variable is taking acceptable values.

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	15986.000000	15987.000000	15987.000000	15987.000000	15987.000000	15987.000000
mean	15.030453	-72.464446	39.915630	-72.464004	39.898726	2.623131
std	431.213944	10.573270	6.828818	10.569932	6.186185	60.890129
min	-3.000000	-74.438233	-74.006893	-74.429332	-74.006377	0.000000
25%	6.000000	-73.992143	40.734935	-73.991182	40.734651	1.000000
50%	8.500000	-73.981689	40.752603	-73.980167	40.753557	1.000000
75%	12.500000	-73.966819	40.767358	-73.963644	40.768008	2.000000
max	54343.000000	40.766125	401.083332	40.802437	41.366138	5345.000000

Table 2.1 summarizes the training data

Few observations:

1. Fare amount ranges from negative to unusually higher positive number. Additionally, it is skewed and not normally distributed
2. Passenger count: It ranges from 0 to 5334 (with 1 instance of each passenger counts more than 6)
3. Latitude and longitude of train_data vary a lot with the location pointing to NYC, New Jersey, Antarctica, etc.
4. Converting the variables into the required datatype. For example: fare amount should be float/numeric.

2.1.1 Missing value analysis:

We observed that there were 24 missing vales in fare_amount and 55 in passenger_count. The below tables shows the missing percentage for each variable.

	Columns	Missing_percentage
1	passenger_count	0.34
2	fare_amount	0.15
3	pickup_datetime	0.01
4	pickup_longitude	0.00
5	pickup_latitude	0.00
6	dropoff_longitude	0.00
7	dropoff_latitude	0.00

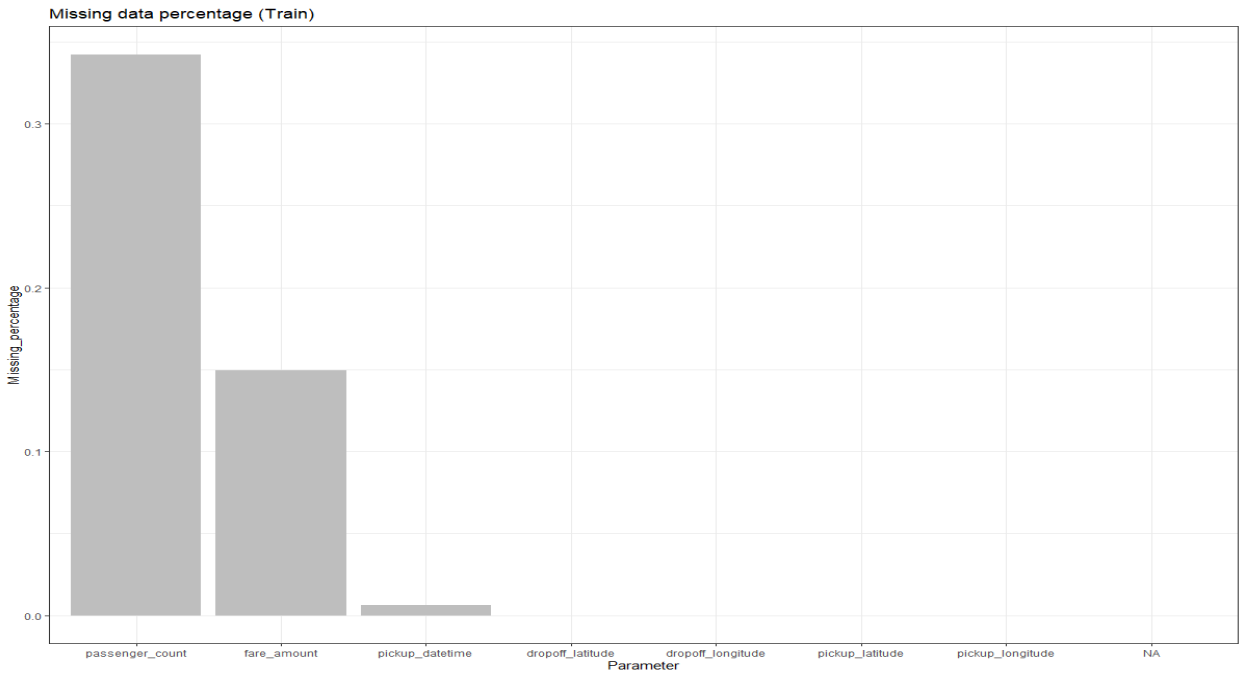

Figure 2.1. Missing value data percentage as a % of total rows

As the missing percentage is small (less than 1 for variable passenger count, fare amount, pickup_datetime), I decided to omit those rows.

2.1.2 EDA and Outlier Analysis

The next step in pre-processing is to analyze the data and remove the outliers or unacceptable values. Outliers/noise result in skewed distribution. In this case, we can clearly observe the skewness in data from probability distribution function. Outliers could be due to incorrect recording/observation or large variability in data. If we don't remove the outliers, it may negatively impact our model especially linear models. We will analyze each variable and will remove or add values to be able to extract the right information.

a. **Passenger Count**: Passenger count varies from 0 to more than thousands. As passenger count in a cab couldn't be this large and additionally for, there is only 1 instance of each passenger count greater than 6, we can remove them considering as outliers. The passenger counts equal to 0 could be due to no passenger showed up (or cancelled the trip) or else it could be the noise. I also cross-checked the test data and it ranges from 1 to 6 so there is no harm in dropping passenger count equal to 0 or more than 6

b. **Coordinates of pickup latitude and longitude**: It seems to have a varied range. It ranges from New York, New Jersey to Antarctica and Equator. The below graph helps us to visualize the coordinates.

   The graph 2.2 and 2.3 show the distribution of latitude and longitude in NYC for train and test data respectively. We have removed the outliers by restricting the range of coordinates to that of test data. We can also observe that there are few points located in water. It doesn't make sense to have cab in water unless it also includes **rental boats.** The test data (which is cleaned data) also has coordinates in water and thus we decided not to remove these points from train data. Next, we calculated distance between pickup and drop locations. It varies from 0 to around 60 miles. The reason I could think of for 0 distance is either the trip was cancelled or the pick up or drop location was incorrectly captured. As the 0 distance is also present in test data, we can conclude that it is not noise but cancellation.

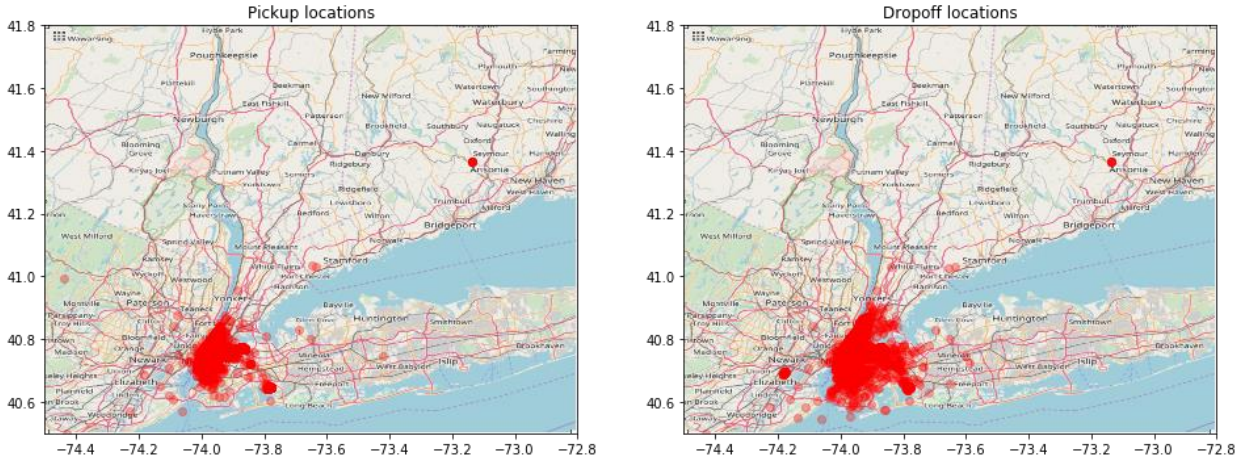

Fig 2.2 The graph represents the coordinates of train data

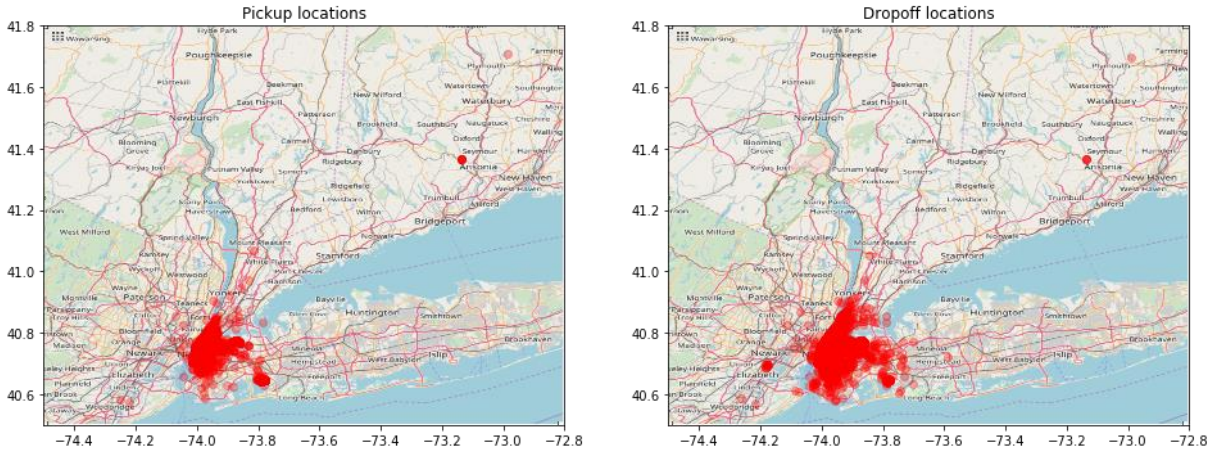

Fig 2.3 The graph represents the coordinates of test data.

I have derived the variable **'pickup_land"** and **"drop _land"** to check if the impact of land vs water is significant. Please note that If pickup_land or drop_land is TRUE, that means the it is on land.
We further calculated distance between pickup and drop location which would be the most important factor in predicting fare amount.

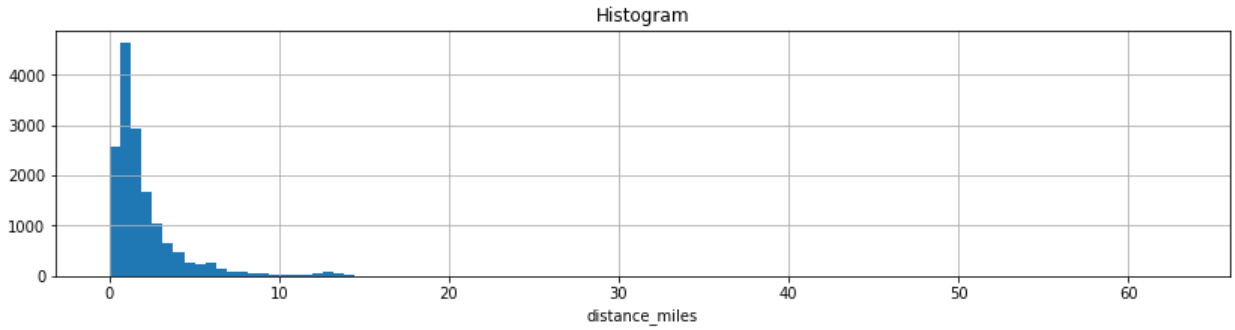

Fig 2.4 shows histogram of distance between pickup and drop locations (in miles)

c. **Fare amount**: Fare amount ranges from negative to more than $53K. Ideally, fare amount can't be negative. Additionally, there would be minimum fare_ amount. Looking at data it seems the minimum amount is $2.5 as the number below it seems the one-off cases (or outliers). I cross checked it online and it seems the minimum fare amount is $2.5. Thus, I have excluded any amount below $2.5. We have also excluded the amount above $400 as it doesn't make sense that cost could go above it. The below plot shows the histogram of fare amount.

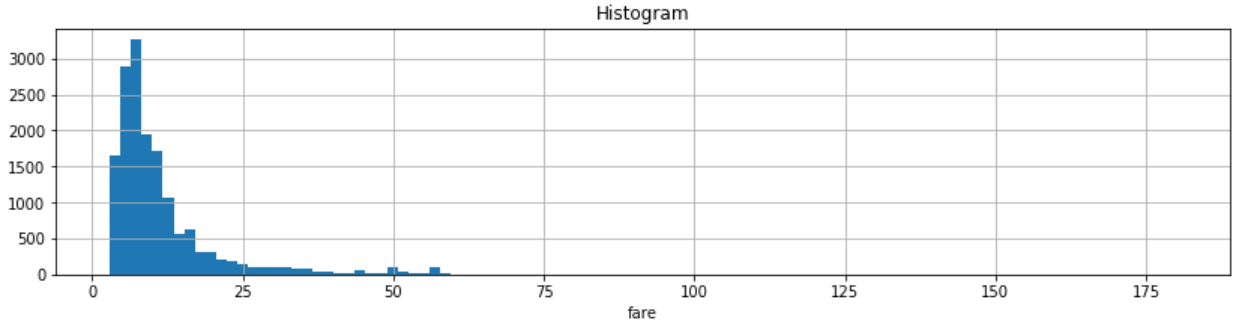

Fig 2.5 Histogram of fare amount

We can see spike near $50. On further analyzing it, we came to know that it is because the fare from jfk airport is almost flat at 50. The below graph shows the same.

d. **Datetime**: We looked at date time variable and extracted the information (like date, year, hour, day of week, session, etc.) Observing the histogram for each, we made the following observations:
   a. The demand is almost same across all months
   b. The demand is more during night resulting in higher cost per mile. Similarly, the demand varies across months and years. Day of week doesn't impact the fare amount significantly.

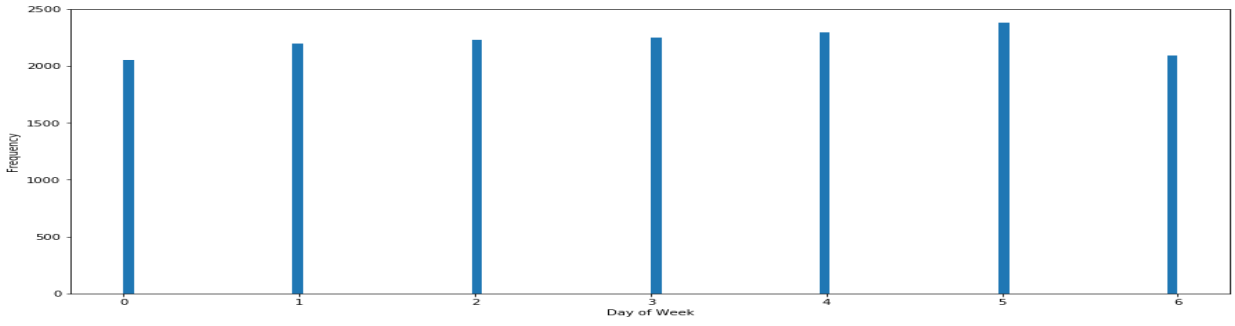

Fig 2.6 Histogram of cab demand on different weeks of day (Sunday as 0)

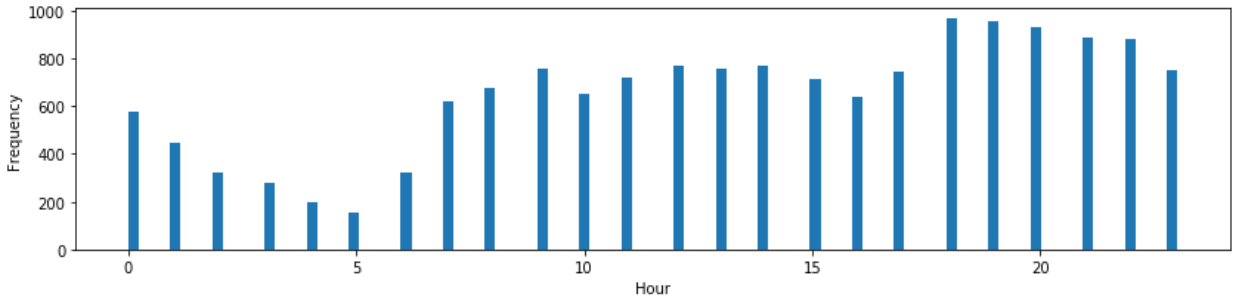

Fig 2.7 Cab demand at different hours of day

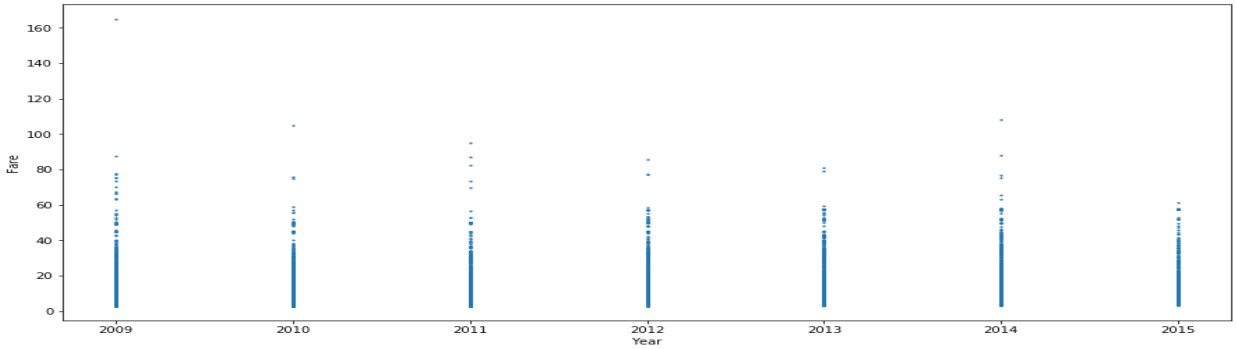

Fig 2.8 Scatter plot showing relatioinship betweenfare amount vs years

Now, we will see the relationship between fare amount and various variables.

i) **Fare amount vs Distance (in miles):**
The below scatter plot shows the distribution between distance and fare amount. We can see that fare amount increases with increase in distance, but the costs are smaller for large distance (around 60 miles) We can also notice that this long distance and small costs combination is for pickup location around **Seymour**

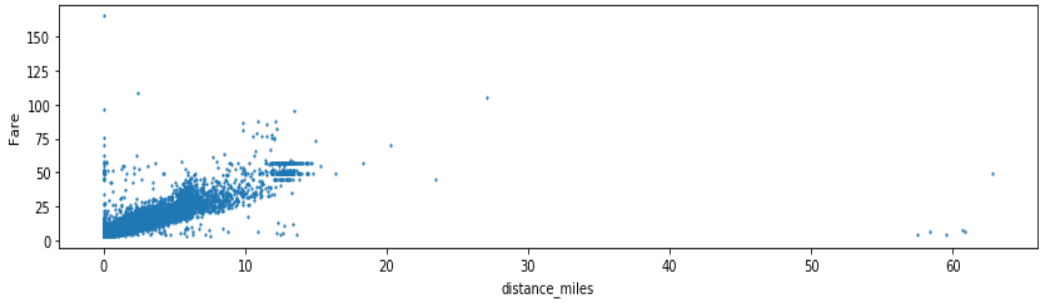

Fig 2.

- The distance ranges from 0 to ~60 miles. Distance could be 0 either in case of cancellation or noise. As 0 distance is also present in test data, we assumed that it couldn't be noise. We observe the fare amount varies significantly for 0 miles distance. As the cancellation charges

should be fixed, we took mean of all the costs with 0 distance which is around 10 miles. We cross checked online an dit seems Uber charges $10 as cancellation charges.

- We also observed that for distance around 60 miles (pick up around Seymour), the cost per mile is significantly lower. This seems outlier. In this case, either distance is outlier or the fare amount. We cross checked it in test data for the combination of distance and pickup coordinates and distance is not outlier. Thus, we treated the fare amount as an outlier and imputed them.
- We also observed that there was spike around $50 and it corresponds to the airport. Thus, we included extra variables, which are distance to/from airports. The below graph shows how the costs are flat for pickup from jfk airport

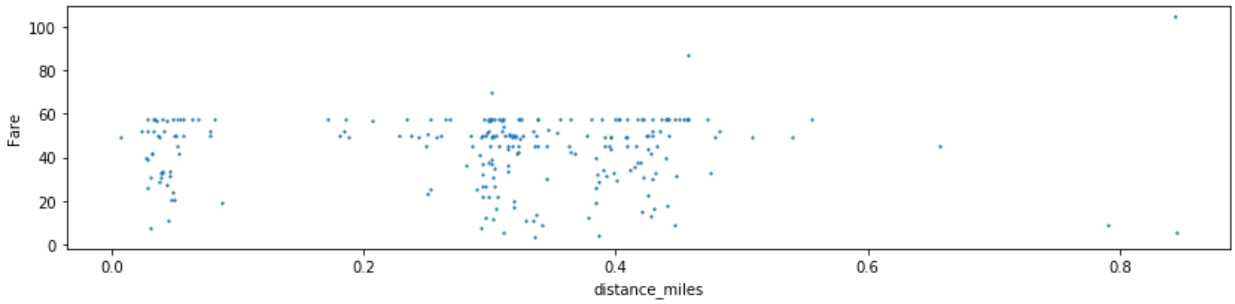

Fig 2.10 Distance from jfk airport vs Fare amount

2.1.3 Feature Selection

Before we perform modelling, it's important to assess the significance of each independent variable in predicting the target variable. There is a possibility that many variables in our analysis are not important to the problem of prediction. Feature selection is basically selecting relevant features for the model construction. We need to perform feature selection as some variables may carry same information or may be information not relevant in predicting the target variable. This in turn increases the overheads. Thus, feature selection helps us to deal with relevant variables and avoid the problem of multi collinearity. In this case we have used **Correlation Analysis** for continuous variable and **ANOVA** (Analysis of variance) for categorical variable.

Correlation

From correlation analysis, we didn't find any strong correlation between independent variables.

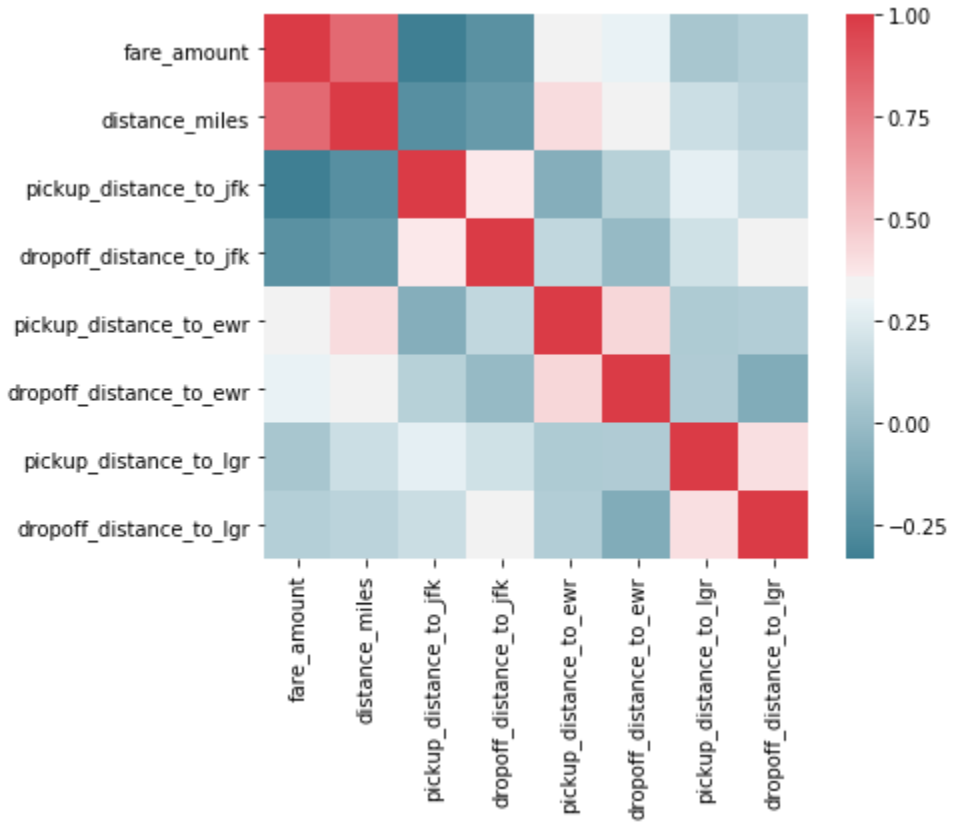

Figure 2.11 Correlation analysis of all continuous variables

ANOVA:

We have rejected variables day of week, pickup_land, drop_land, date and passenger count and consider month, year, hour and sessions important variables as per output of ANOVA in the below table

	sum_sq	df	F	PR(>F)
Year	1.698414e+04	1.0	191.269758	3.034502e-43
Residual	1.375906e+06	15495.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Month	2.365271e+03	1.0	26.356859	2.872450e-07
Residual	1.390525e+06	15495.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Date	6.107818e+01	1.0	0.679485	0.409777
Residual	1.392829e+06	15495.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Day_of_Week	1.434349e+00	1.0	0.015956	0.899482
Residual	1.392889e+06	15495.0	NaN	NaN

	sum_sq	df	F	PR(>F)
Hour	1.311439e+03	1.0	14.602656	0.000133
Residual	1.391579e+06	15495.0	NaN	NaN

	sum_sq	df	F	PR(>F)
pickup_land	5.146342e+02	1.0	5.727088	0.016717
Residual	1.392376e+06	15495.0	NaN	NaN

	sum_sq	df	F	PR(>F)
drop_land	3.210314e+02	1.0	3.57209	0.058777
Residual	1.392569e+06	15495.0	NaN	NaN

	sum_sq	df	F	PR(>F)
session	2.881672e+03	3.0	10.706346	5.017529e-07
Residual	1.390008e+06	15493.0	NaN	NaN

	sum_sq	df	F	PR(>F)
passenger_count	2.722363e+01	1.0	0.302851	0.582108
Residual	1.392863e+06	15495.0	NaN	NaN

2.2.4 Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization as range of values of raw data varies widely, the objective functions may not work properly without normalization. Thus, is important to go for feature scaling. There are two ways of feature scaling: Standardization or normalization. The choice between two depends on distribution of data if our data is normally distributed, we go for standardization, else we scale it from 0 to 1. In our project, our data is not uniformly distributed we will use **Normalization** as Feature Scaling Method.

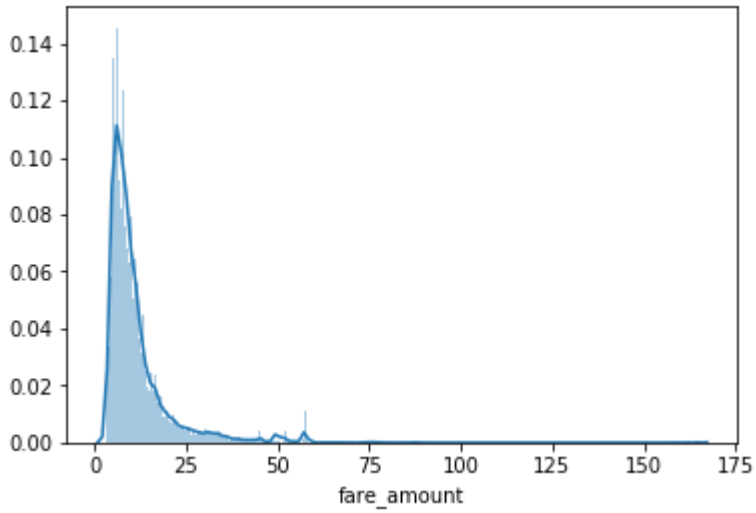

Fig 2.12 the impact of normalization shown as an example for the target variable fare amount

But as our variables are of same units, scaling won't add much value.

2.2 Modeling

Once we are done with the preprocessing, the next step is to build the models using the processed data. As our target variable is continuous, we are using regression models to predict the target variable. Following are the models which we have built –

2.2.1 Linear Regression

Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous data. It can only be applied if the data follows some assumptions. The most important assumption is to check the multicollinearity. While analyzing the data using VIF, we found that data doesn't have multicollinearity problem.
The RMSE and R-square values are shown below

Linear Regression	R	PYTHON
RMSE Train	4.65	4.950
RMSE Test	5.03	5.031
R^2 Test	0.7267	0.687

2.2.2 Decision Tree

A decision tree is a supervised learning approach that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. The advantage of this model is that it can be easily understood by the business users. The RMSE value and R^2 value for our project in R and Python are –

Decision Tree	R	PYTHON
RMSE Train	4.364	5.004
RMSE Test	4.70	4.743
R^2 Test	0.760	0.722

2.2.3 Random Forest

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees randomly to have more accuracy in dataset. The randomness in building the number of trees and choosing the variable leads to the name random forest. The RMSE value and R^2 value for this project in R and Python is as follows:

Random Forest	R	PYTHON
RMSE Train	1.633	1.499
RMSE Test	4.160	3.74
R^2 Test	0.8138	0.8266

2.2.4 Gradient boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Gradient boosting	R	PYTHON
RMSE Train	3.239	3.192
RMSE Test	3.980	3.717
R^2	0.829	0.829

Which metrics and why?
We have chosen RMSE as it doesn't cancel out positive and negative errors and additonally penalize larger error and thus is the more useful metrics here. It is one of the most useful metrics for time series analysis.

Cross validation score:

Using the validation set approach where we split the data into train and test, the above tables show RMSE values for train and test. We also tried 5-fold approach for GBM and the score averages out to 0.8272.

Chapter 3
Conclusion

The next step in the process is to evaluate the model and decide which model is suitable for our data.

3.1 Model Evaluation

In the previous chapter we observed the **Root Mean Square Error** (RMSE) and **R-Squared** Value for different models. **Root Mean Square Error** (RMSE) is the standard deviation of the residuals (prediction **errors**). Residuals are measure of spread of data from regression line data points and RMSE is the SD of residuals. In other words, it tells you how concentrated the data is around the line of best fit. Whereas **R-squared** is a relative measure of fit, **RMSE** is an absolute measure of fit. Lower values of **RMSE** and higher value of **R-Squared Value** indicate better fit.

3.2 Model Selection

From the observation of all **RMSE Value** and **R-Squared** Value we have concluded that Random forest **model** has minimum value of RMSE, and Gradient Boosting has the second least. R-square is approximately same for both, but we chose RMSE train and test and it differs significantly for Random Forest implying the probability of overfitting. But it's almost same in case of Gradient Boosting. *Thus, we chose the Gradient Boosting model*

3.3 Predicting fare amount for test data:

Using gradient boosting model , we predicted the fare amount for test data and the below figure depicts the relationship between fare amount and distance in miles.

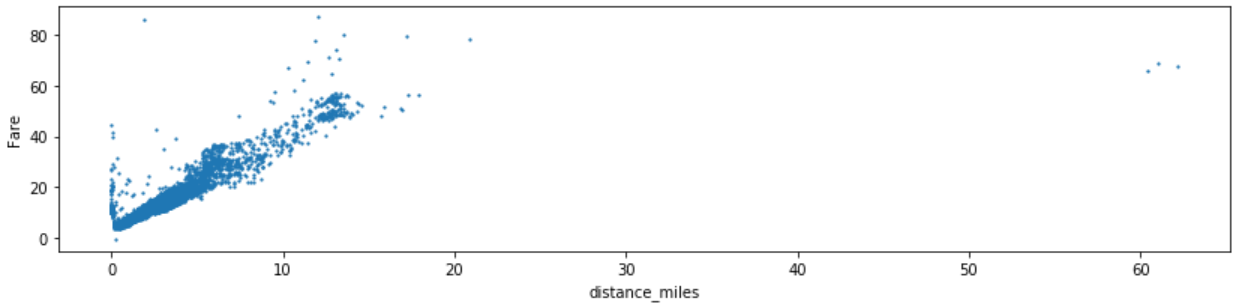

Fig 3.1 Figure depicting the fare amount for test data vs fare amount