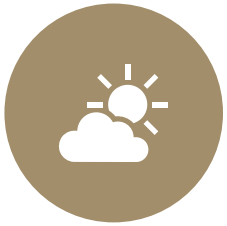
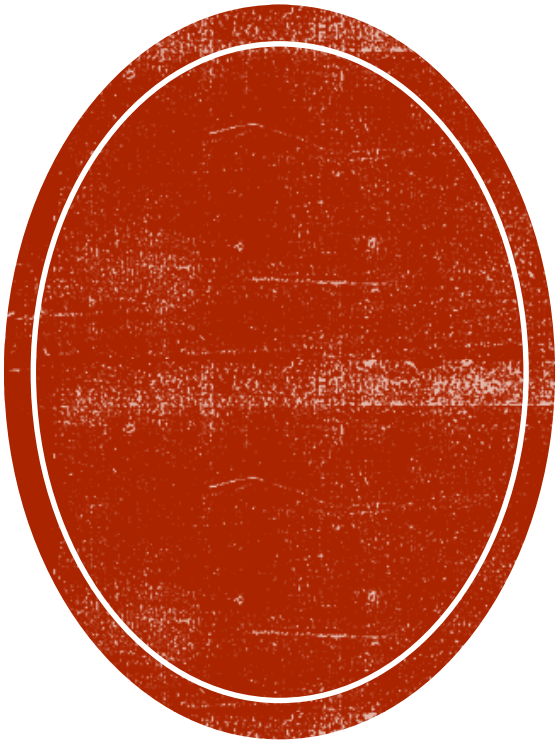
CAB FARE PREDICTION



**PREDICTIONS BASED ON TIME/DISTANCE**

**BY:**

**RAHUL RANJAN**



**Contents**

# 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.

**Number of attributes: ·**

- 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

## 1.2 Data

Data is attached as csv

Snippet of data as shown below : Train Data

| | fare_amount | pickup_datetime | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude | passenger_count |
|---|---|---|---|---|---|---|---|
| 0 | 4.5 | 2009-06-15 17:26:21 UTC | -73.844311 | 40.721319 | -73.841610 | 40.712278 | 1.0 |
| 1 | 16.9 | 2010-01-05 16:52:16 UTC | -74.016048 | 40.711303 | -73.979268 | 40.782004 | 1.0 |
| 2 | 5.7 | 2011-08-18 00:35:00 UTC | -73.982738 | 40.761270 | -73.991242 | 40.750562 | 2.0 |
| 3 | 7.7 | 2012-04-21 04:30:42 UTC | -73.987130 | 40.733143 | -73.991567 | 40.758092 | 1.0 |
| 4 | 5.3 | 2010-03-09 07:51:00 UTC | -73.968095 | 40.768008 | -73.956655 | 40.783762 | 1.0 |

Snippet of data as shown below : Test Data

| | pickup_datetime | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude | passenger_count |
|---|---|---|---|---|---|---|
| 0 | 2015-01-27 13:08:24 UTC | -73.973320 | 40.763805 | -73.981430 | 40.743835 | 1 |
| 1 | 2015-01-27 13:08:24 UTC | -73.986862 | 40.719383 | -73.998886 | 40.739201 | 1 |
| 2 | 2011-10-08 11:53:44 UTC | -73.982524 | 40.751260 | -73.979654 | 40.746139 | 1 |
| 3 | 2012-12-01 21:12:12 UTC | -73.981160 | 40.767807 | -73.990448 | 40.751635 | 1 |
| 4 | 2012-12-01 21:12:12 UTC | -73.966046 | 40.789775 | -73.988565 | 40.744427 | 1 |

## 2.1 Pre Processing

We start with Data Exploratory Analysis and changing the way data looks
We change the behavioral data into categorical columns

## 2.2 Missing Value Analysis

Missing value analysis is done to check is there any missing value present in given dataset. Missing values can be easily treated using various methods like mean, median method, knn method to impute missing value.

In case data set is vey large and finding missing value is tedious we can also drop the missing value

1. Check the data for any null value.
2. Remove the data which has outliers.
3. Remove the data which are practically not possible for example latitude smaller or grater than -90.
4. Check for any junk data in this training set we have 43 as junk value in pickup_Date column.
5. Convert each data type object into datetime or numeric for our calculatiions.
6. The perfect data type looks like below :

```
fare_amount          float64
pickup_datetime       object
pickup_longitude     float64
pickup_latitude      float64
dropoff_longitude    float64
dropoff_latitude     float64
passenger_count      float64
dtype: object
```

## 2.3 Exploratory Data Analysis:

1. Lets start creating data to be more meaningful data.
2. We will separate the Pickup_datetime column into separate field like year, month, day of the week, etc in both train data and test data.
3. Now as it is also known that in cab service that fare depends upon below things:
   I. Distance travelled
   II. No. of hours/minute the taxi was running.
   III. No. of passengers.

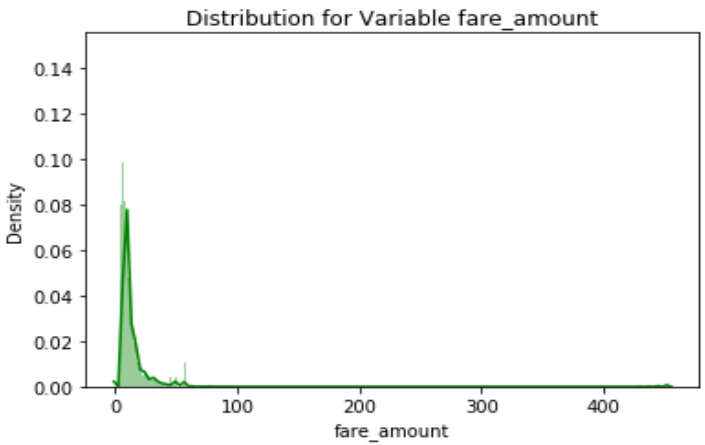
4. We have latitude and longitude info from drop and pickup points so lets calculate distance using haversine formula.
5. Now the new data looks like below with additional columns.

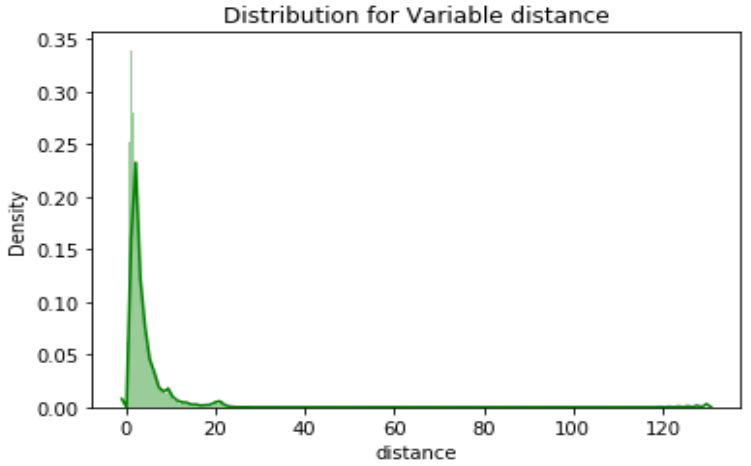
| ount | pickup_datetime | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude | passenger_count | year | Month | Date | Day | Hour | Minute | distance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4.5 | 2009-06-15 17:26:21 | -73.844311 | 40.721319 | -73.841610 | 40.712278 | 1.0 | 2009 | 6 | 15 | 0 | 17 | 26 | 1.030764 |
| 16.9 | 2010-01-05 16:52:16 | -74.016048 | 40.711303 | -73.979268 | 40.782004 | 1.0 | 2010 | 1 | 5 | 1 | 16 | 52 | 8.450134 |
| 5.7 | 2011-08-18 00:35:00 | -73.982738 | 40.761270 | -73.991242 | 40.750562 | 2.0 | 2011 | 8 | 18 | 3 | 0 | 35 | 1.389525 |
| 7.7 | 2012-04-21 04:30:42 | -73.987130 | 40.733143 | -73.991567 | 40.758092 | 1.0 | 2012 | 4 | 21 | 5 | 4 | 30 | 2.799270 |
| 5.3 | 2010-03-09 07:51:00 | -73.968095 | 40.768008 | -73.956655 | 40.783762 | 1.0 | 2010 | 3 | 9 | 1 | 7 | 51 | 1.999157 |

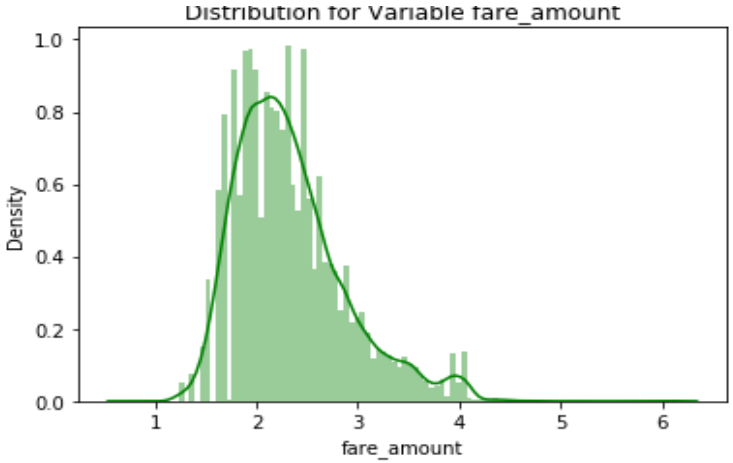
## 2.4 Feature Scaling

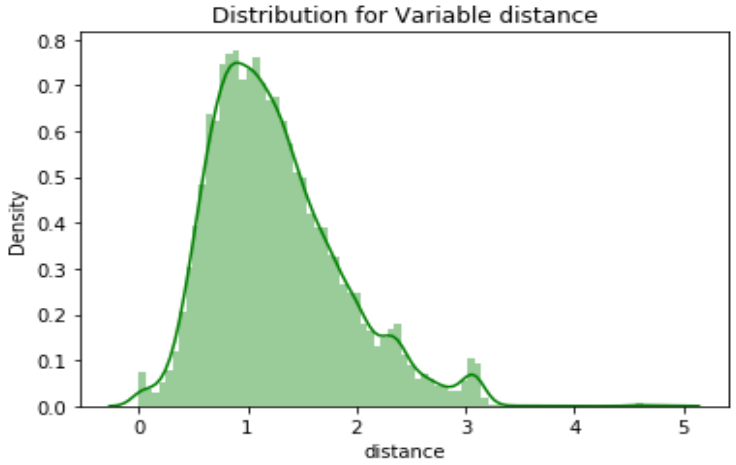
We do feature scaling to see if there is any skewness in data.
The data should be in good symmetry . If it is not then we make it symmetrical by doing normalization.

We observe the data symmetry in fare_amount and distance and found it is ok.



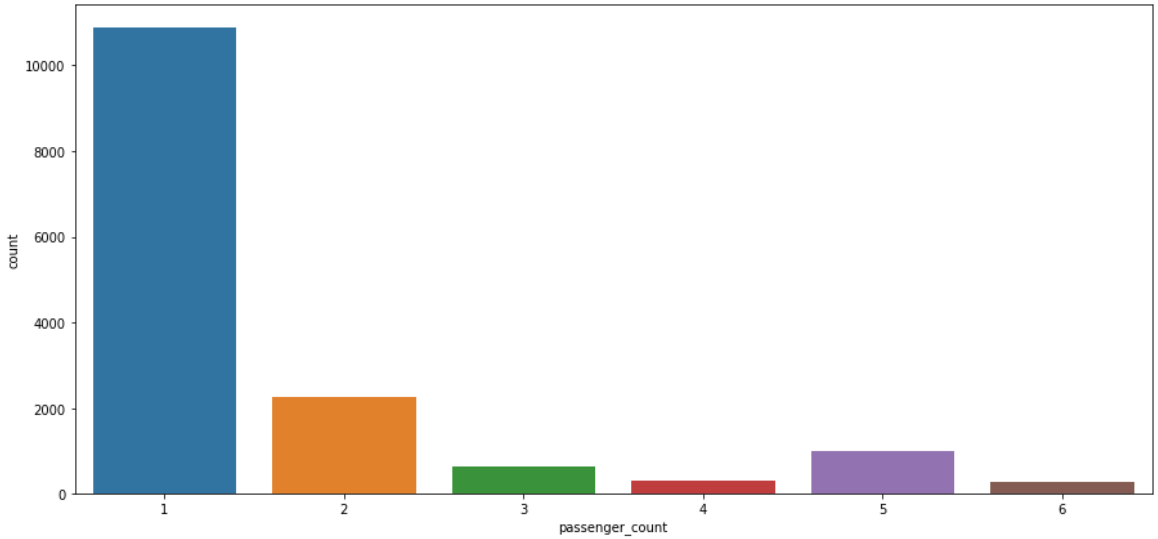



## 2.5 Feature Selection/feature visualization

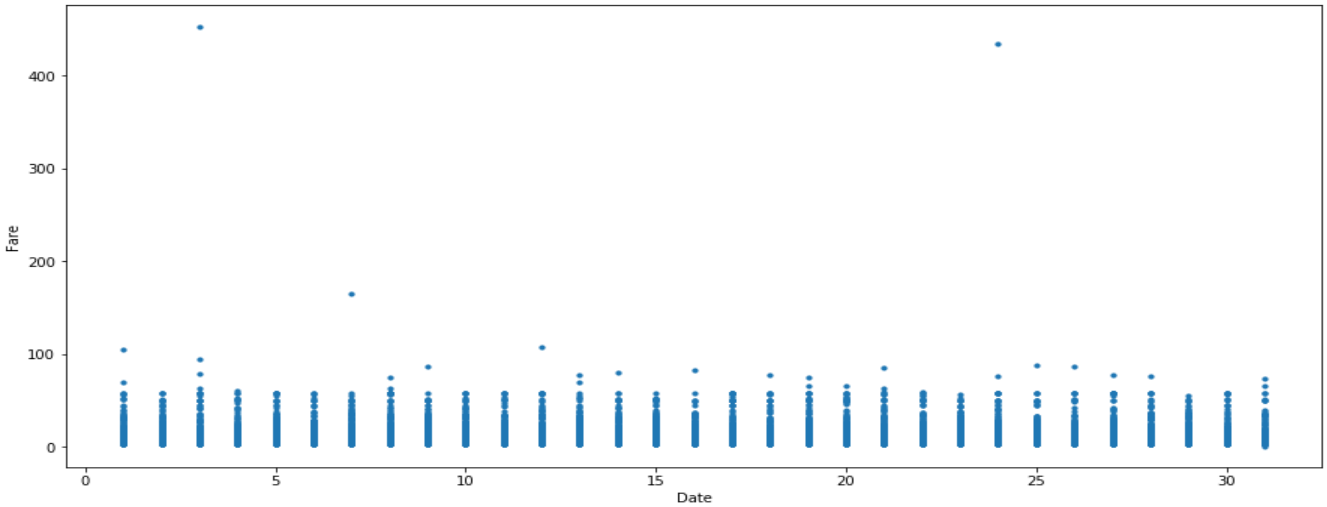
Visualizing the data in various ways to find any relationship between the features.
Passenger Data --What people prefers travelling alone or with group?



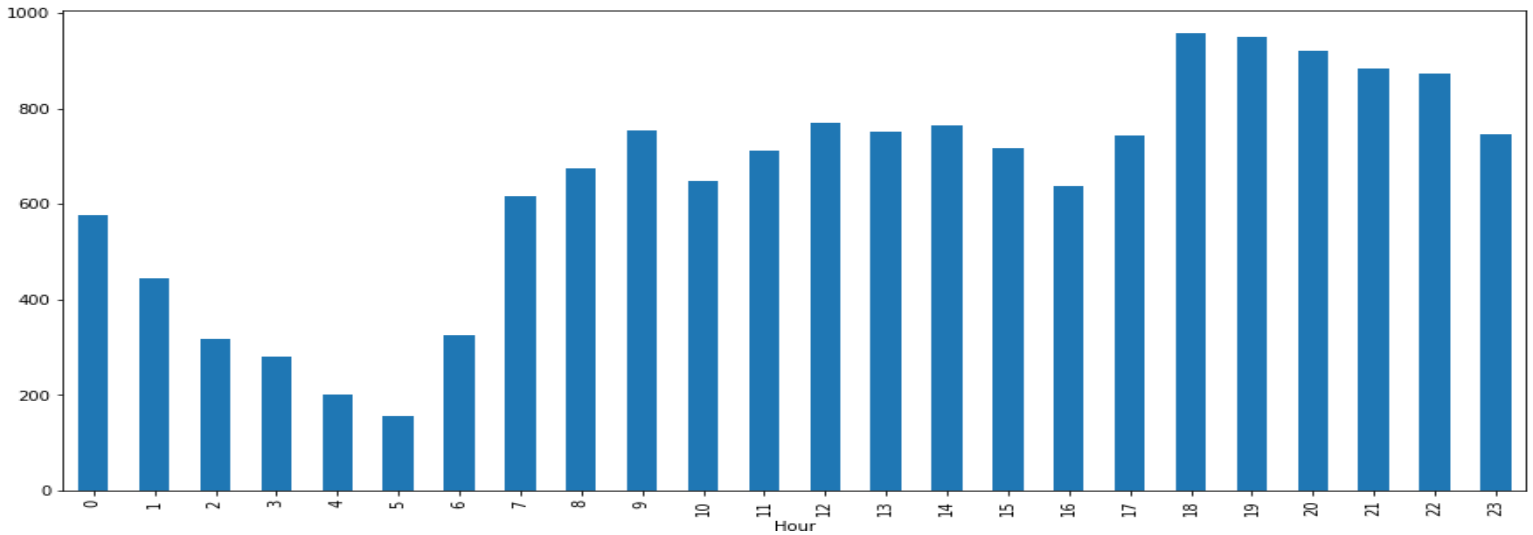

We observe :
People prefers to travel alone

## 2.5 Feature Selection/feature visualization

Relation between date and fare—Uniform reltion

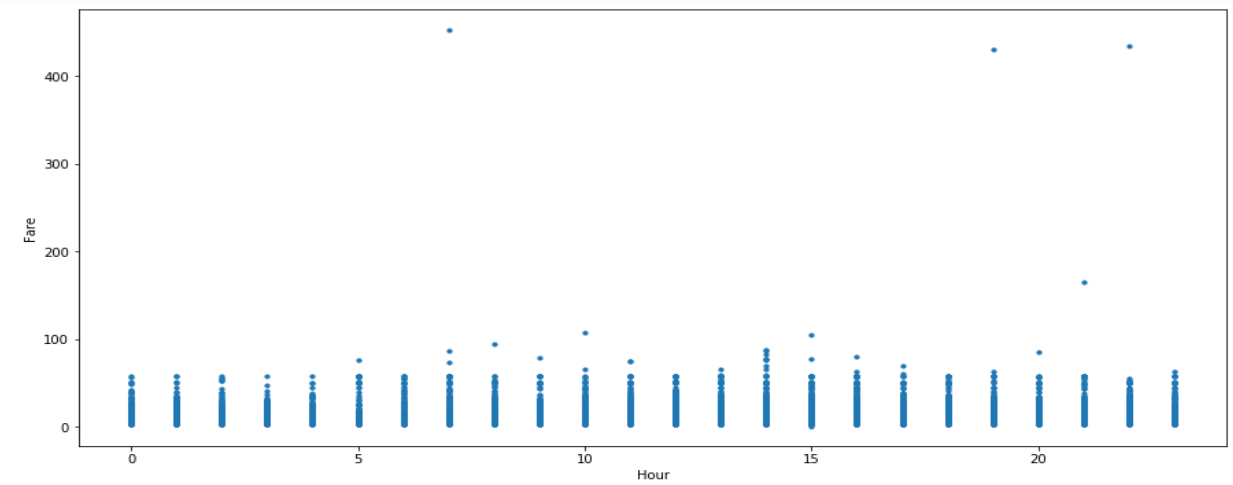



Relation between hour and cabs – From 7 am to 23 pm more cabs

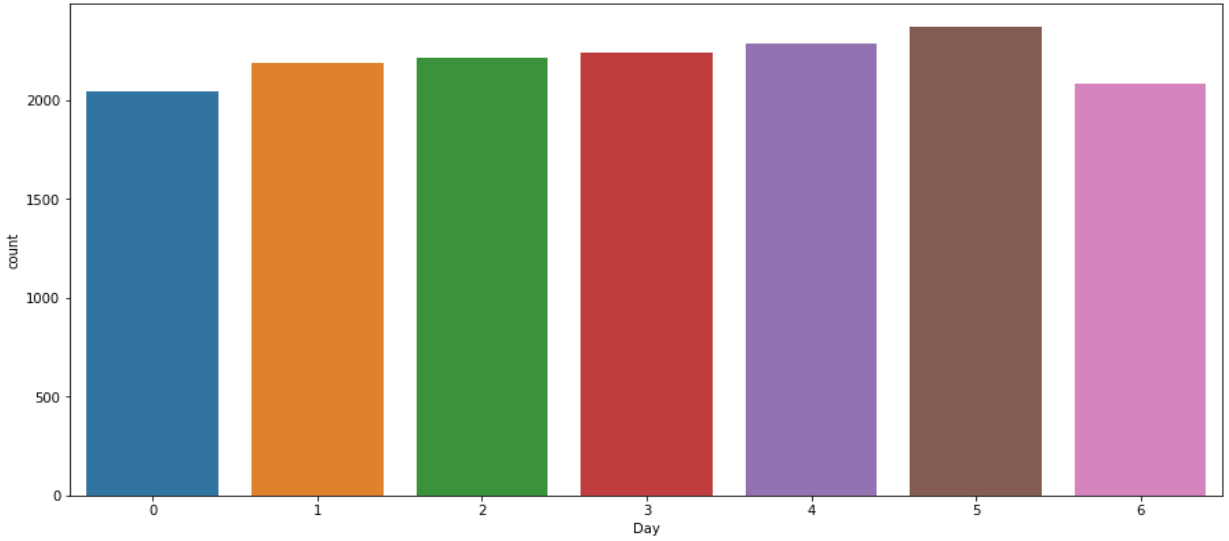


## 2.5 Feature Selection/feature visualization

Relationship between Time and Fare—More during 7 am to 23 pm

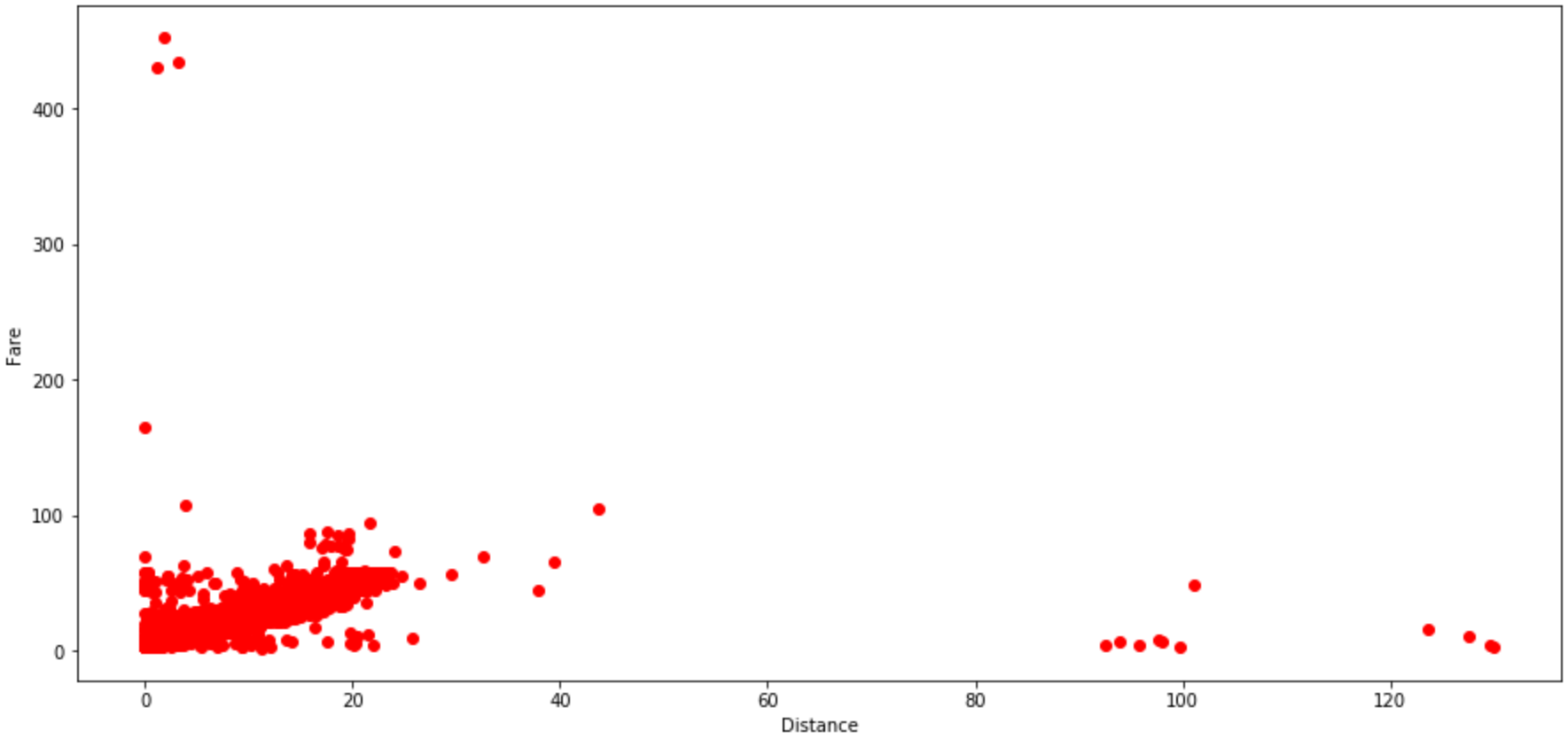



Relation between day and cabs– Not much



## 2.5 Feature Selection/feature visualization

Relationship between Distance and Fare –More the distance more is the fare



**Things confirmed after analysis:**

1. Passengers like travelling alone
2. More distance means more fares.
3. Cabs booked between 7 am to 23 Pm will have high fares due to rush hours..

Lets starts predictions.

So the problem statement is of Regression type so we can use:

1. Linear Regression
2. Decision Tree
3. Random Forest etc.

We apply Linear Regression, Decision tree and Random Forest :
After applying in all the 3 case we find the MAPE, RMSE.

MAPE- Mean Absolute Percentage Error– How much % error our model has.
Lower the MAPE better is the accuracy score and a good model for our case

RMSE – Root Mean Sqaured Error – How much much % error our model has .
Lower the RMSE better is the model .

Below are the calculated values , based upon this model is selected.

We see Random Forest as clear winner. So we will Random forest for predictions

| Model | MAPE in % |
|---|---|
| Linear Regression | 36.78 |
| Decision Tree | 32.31 |
| Random Forest | 9 |

| Model | RMSE in % |
|---|---|
| Linear Regression | 9.71 |
| Decision Tree | 8.45 |
| Random Forest | 3.17 |

| Model | Rsquared in % |
|---|---|
| Linear Regression | 60 |
| Decision Tree | 72 |
| Random Forest | 92 |

# THE END