

SUMMER INTERNSHIP PROJECT REPORT

**CASE STUDY ON PREDICTION OF TAXI FARES**

SUBMITTED BY:

POTLURI SAI CHERISH

BOLISETTY SAI GANESH

NALABOTHU VAMSI KRISHNA

MUVVALA AVINASH

PANTHALA KRISHNA SAI

Under the esteemed guidance of

Mr. Mohan P

Senior Data Consultant

SMART BRIDGE

Acknowledgement

We hereby thank **Mr. P Mohan** who has given us the great opportunity to work on this project. He has been a great source of inspiration and her timely support and guidance has helped in the successful completion of the same.

Then we would like to thank **Mr. Manthan Mehta** for his enthusiastic approach and dedication which has indeed been a great source of inspiration and support to us. Also our sincere thanks to **Mr. Daniel Manoj** who guided us throughout the project.

We thank our respective institutions for permitting us to attend this program which is sure to play a significant part in the interviews that we are to face soon. We thank the Jawaharlal Nehru Technological University Hyderabad for providing us with necessary infrastructure to do our work.

**Index**

1. Abstract
   1. Introduction
   2. Objective of Research
2. Problem Statement
3. Data Collection
4. Methodology
   1. Exploratory Data Analysis

5. Findings and Suggestions

6. Conclusion

7. Bibliography and Reference

**Abstract**

The taxi fares are not same always and might vary based upon many factors. This unpredictable nature of the taxi fares is the main output here. In order to study various factors and their effects on the taxi fares many different samples of data were collected. Parameters like payment\_type, surcharge, fare\_amount e.t.c. are found in the samples thus collected. This study reveals that the most of the parameters are exceeding their acceptable limit and hence lay significant impact on the taxi fares.

**Introduction**

A **fare** is the [fee](https://en.wikipedia.org/wiki/Fee) paid by a [passenger](https://en.wikipedia.org/wiki/Passenger) for use of a [public transport](https://en.wikipedia.org/wiki/Public_transport) system: [rail](https://en.wikipedia.org/wiki/Rail_transport), [bus](https://en.wikipedia.org/wiki/Bus), [taxi](https://en.wikipedia.org/wiki/Taxicab), etc. In the case of [air](https://en.wikipedia.org/wiki/Airline) transport, the term *airfare* is often used.

Fare structure is the system set up to determine how much is to be paid by various passengers using a [transit vehicle](https://en.wikipedia.org/wiki/Transit_vehicle) at any given time.

A linked trip is a trip from the origin to the destination on the transit system. Even if a passenger must make several transfers during a journey, the trip is counted as one linked trip on the system.

The fare paid is a contribution to the operational costs of the transport system involved, either partial (as is frequently the case with publicly supported systems) or total. The portion of operating costs covered by fares - the [farebox recovery ratio](https://en.wikipedia.org/wiki/Farebox_recovery_ratio) - typically varies from 30%-60% in North America and Europe, with some rail systems in Asia over 100%.

The rules regarding how and when fares are to be paid and for how long they remain valid are many and varied. Where the fare can be generally be predicted in advance (such as fixed fare systems) fare is usually collected in advance; this is the usual practice of rail and bus systems, who usually require the payment of fares on or before boarding.

Some systems use a hybrid of both, such as a rail system which requires prepayment of the minimum fare in advance, and collecting amounts above the minimum (if the net cost of the trip exceeds the minimum fare) at the end of the trip.

Some systems allow [free transfers](https://en.wikipedia.org/wiki/Free_transfer_(transport)): that is to say that a single payment permits travel within a particular geographical zone or time period. Such an arrangement is helpful for people who need to transfer from one route to another in order to reach their destination. Sometimes transfers are valid in one direction only, requiring a new fare to be paid for the return trip.

A farebox is a device used to collect fares and tickets on streetcars, trains and buses upon entry, replacing the need for a separate conductor. Nearly all major metropolitan transit agencies in the United States and Canada use a farebox to collect or validate fare payment.

Fareboxes did not change again until around 1984, when fares in many larger cities reached $1.00 and the first dollar bill accepting farebox was put into service. In 2006, new fareboxes had the capability of accepting cash, credit, or [smartcard](https://en.wikipedia.org/wiki/Smartcard) transactions, and issuing day passes and transfers for riders.

**Objective of Research**

The given data set is about Prediction Of Taxi Fares. The given data is to analyzed. Various objectives of the given data set are as follows:

a) To study the given data

b) To apply data cleaning methods to remove unknown data from the

given data set.

c) Test the designed model’s working

d) Draw conclusions from the developed model

e) Predict whether the fare total amount will be satisfied by the

customer or not.

**Problem Statement**

The data is related with taxi fares. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access. It is to be predicted whether a customer will satisfied to the fare amount or not.

**Data Collection**

The given data set is related to Taxi Fares. It was taken from the website **kaggle.com**. The website provides various datasets from various domains.

The given data set consists of 8 features and 49,999cases. The features are as follows:

1. vendor id: it describes about id it is categorical data  
2.Payment type:two types CRD or CSH  
3. Fare amount: It describes total distance fare amount   
4. Surcharge: About Taxi Fares. Metered taxi fares are based on a flag-down fare and the distance travelled. Additional surcharges may also apply (e.g. peak period, late night hiring surcharges). If you call for a taxi, a booking fee also applied  
5. Mta\_tax: Most trips in zone are small fares – that means 30% tax on taxi trips while general sales tax is 8.75%. 90% of yellow trips are in the zone, effecting yellow cabs the most and disproportionately from all other private and commercial motorists.  
6. Tip\_amount: Fifteen to 20 percent of your **fare** is a good rule of thumb, but you can **tip** more or less depending on your experience.

7. tolls\_amount: It describes about tollgate amount

8. Total\_amount: the final amount which is sum of all above charges

**Methodology**

The given data set consists of 44,999cases and 8i features out of which one is output and the remaining 8 are input features. As the first step the a few essential packages like **pandas, numpy, matplotlib, seaborn** and **os** are imported into the Jupiter notebook. Now the given data set is imported to the Jupiter note book using pandas. It can be observed that most of the data in the given data set is categorical.

The second step is to check if there is any missing data in the given data set. On applying **is null** function we observe that no visible missing data is present. But on further analyzations it can be observed that eight features have **unknown** values which have to be sorted before applying classification models on the given data. This process is called data cleaning. As the data is categorical in nature, the unknown values are replaced by the mode category of the feature.

Training and testing data and the model with highest efficiency is chosen to draw conclusions regarding the given dataset.

**Exploratory Data Analysis**

Data Analysis

Evaluation

Prediction

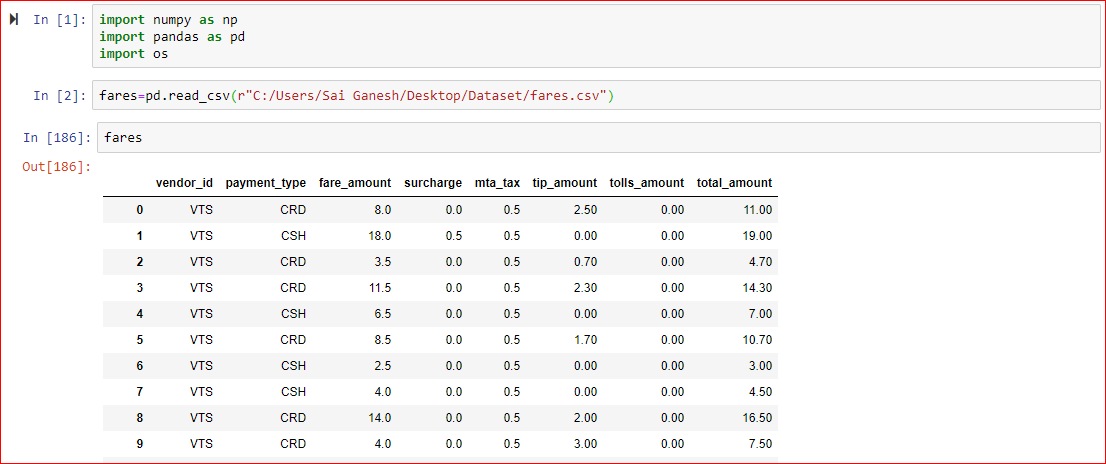
Model Building

Feature Selection

Pre-Processing

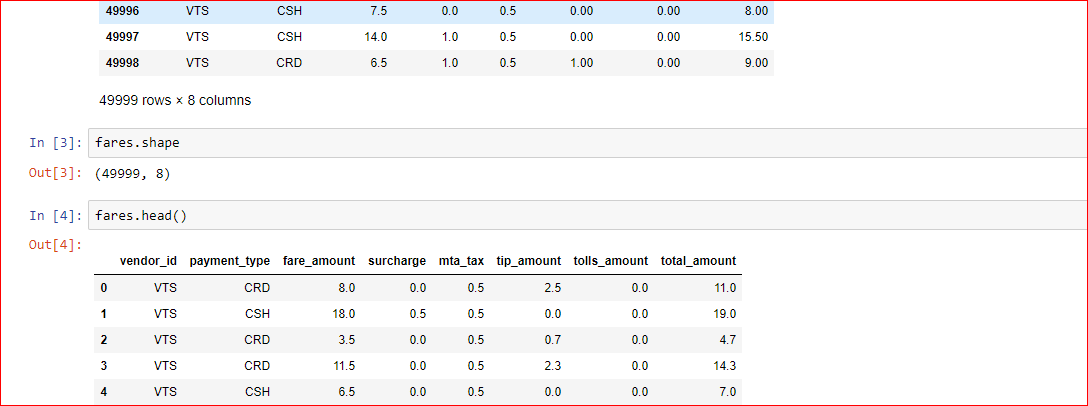
During the process of data cleaning we can observe that the features. Features are vendor\_id,payment\_type,fare\_amount,mta\_tax,tip\_amount,tolls\_amount,total\_amount As the data in these features is numeric in nature, But applying this method directly will draw contradictions. To avoid there a few direct imputations are made to the dataset based on the observations made from the cross tables drawn between the features that contain unknown values.

Data analysis will include the following processes to take place. Initially we will import the python packages from libraries and thus we will be able to access special features.



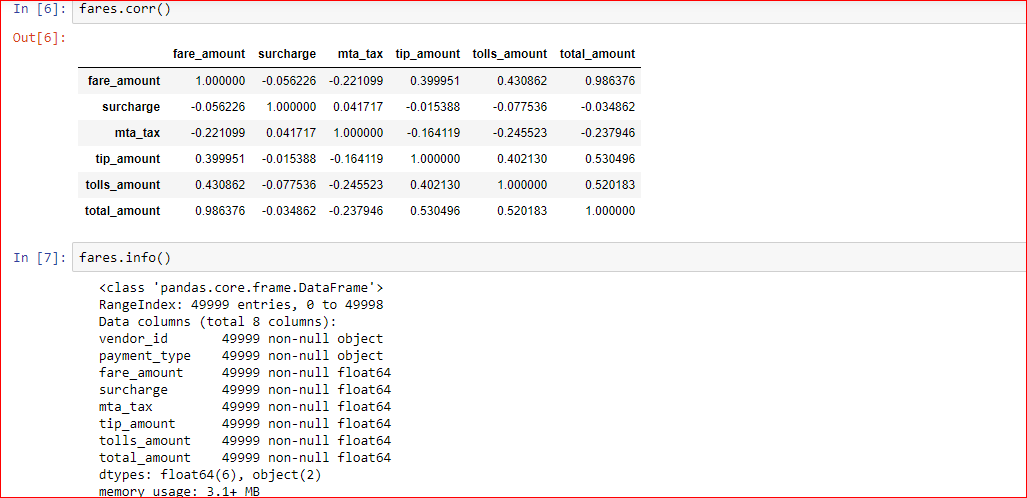
Then we can also import **.csv** files then the data can be displayed as such, we use many functions to display , manipulate and change the data.

One such feature is shape which can be used for displaying the number of rows and columns. Also, we use here the **head()** function which is used to display the top most tuples and the range can be determined by the user themselves.



**Correlation** is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive **correlation** indicates the extent to which those variables increase or decrease in parallel; a negative **correlation** indicates the extent to which one variable increases as the other decreases.

Here by using **.corr()** we can get the correlation values of the attributes, such that we can determine the values which are being negatively effected and are being positively effected

****

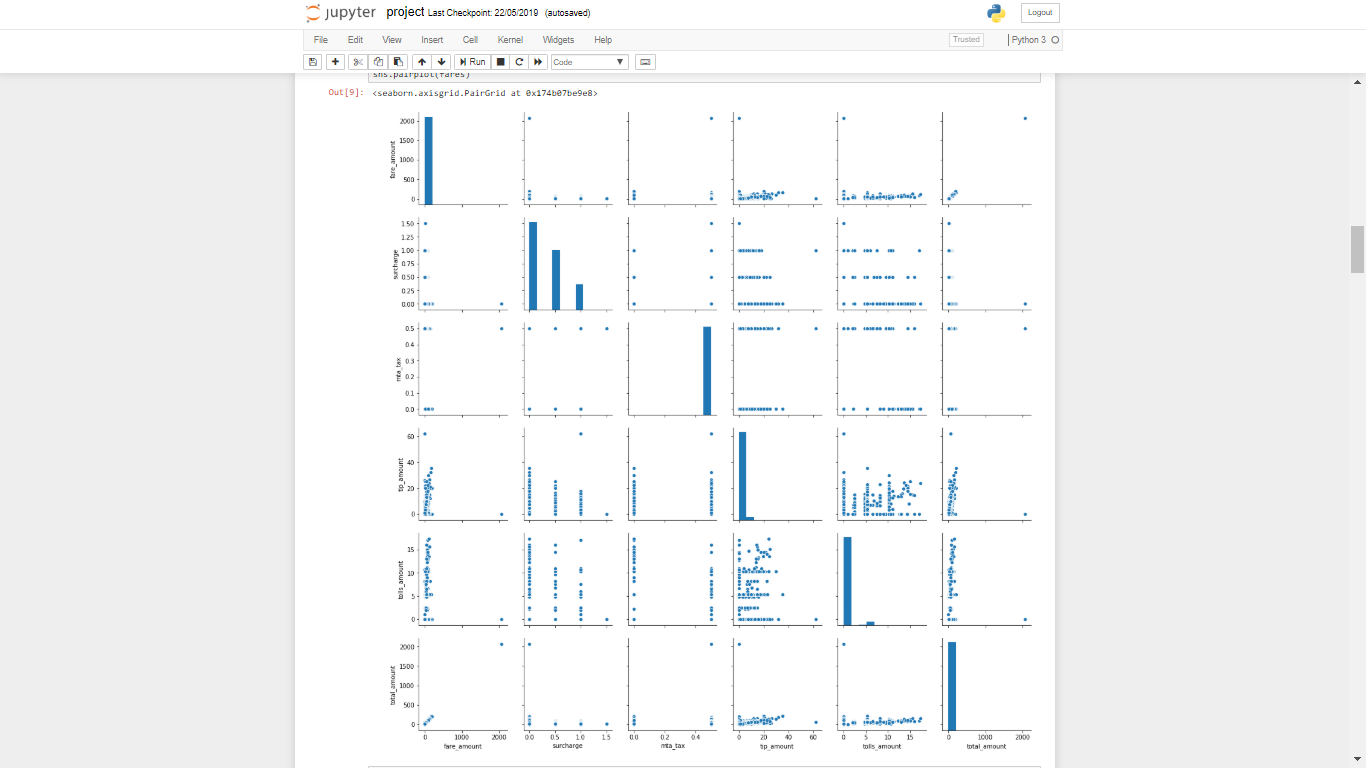
**.info()** is used to get the information about how many columns are there in total, datatype of the objects and whether they are null or not etc.

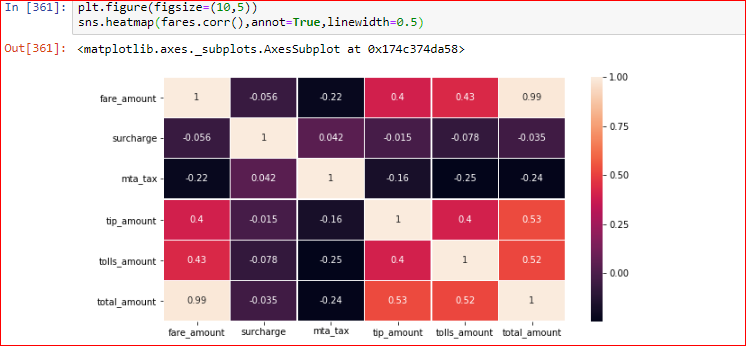
The function **lambda** is used to find out whether there are any null values or not the **axis=0** is for column wise searching and **axis =1** is for row wise searching.

****

Function **.describe()** is used to describe the give dataset in terms of quartiles and it also gives values of mean, count, std, min and max values as well.

We will use the library **seaborn()** which can be used to display the graphical relations between each and every attributes and we can get the graphs in many variants they can be scatter plots or bar graphs etc.

****In previous we are using scatter plots to represent the data model heat map is used to represent the data in blocks positive is in whitish colour and negative represents in black colour it also represents all the features in the given data.

****

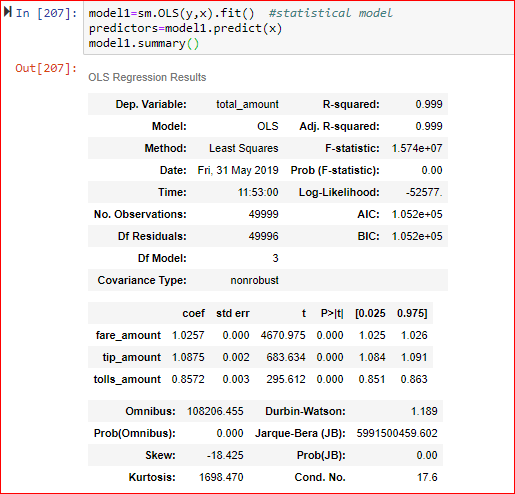
The function **.summary()** is used to get the summary of the data being used and here we have few key values. Them being.

**Dep.variable** – Which is the variable which is depended on inputs, in simple terms the dependent variable is the output variable. R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regressionmodel. ... It may also be known as the coefficient of determination.. **Adj.R-squared** is used to address the population.**P** is the probability and its value is always compared to **α**(0.05) :

If **P<α** we reject the null hypothesis and accept the alternate hypothesis.

If **P>α** we accept the null hypothesis and reject the alternate hypothesis.

Root-mean-square error (**RMSE**) (or sometimes root-mean-square*d* error) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an [estimator](https://en.wikipedia.org/wiki/Estimator) and the values observed. The RMSD represents the square root of the second [sample moment](https://en.wikipedia.org/wiki/Sample_moment) of the differences between predicted values and observed values or the [quadratic mean](https://en.wikipedia.org/wiki/Quadratic_mean) of these difference

****

**Correlation**: It is the strong relation between two variables Correlation always lies between 0 to 1 weak correlation lies between 0 to 0.5 strong correlation lies between 0.5 to 1.**Accuracy**: if rmse value is low and R- square value is high then Accuracy is high. **OLS** means ordinary least squares it is used to estimate the un known parameters in the model

**Linear Regression**

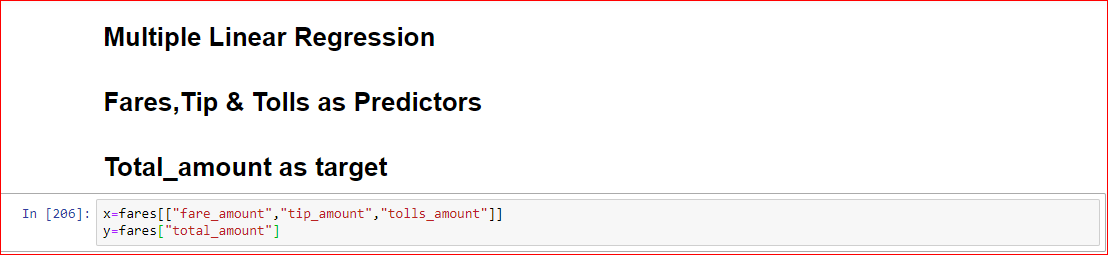
It is one of the regression model in linear regression the predicted out put value is in numerical.Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical.

There are 3 major uses for multiple linear regression analysis. First, it might be used to identify the strength of the effect that the independent variables have on a dependent variable.

Second, it can be used to forecast effects or impacts of changes. That is, multiple linear regression analysis helps us to understand how much will the dependent variable change when we change the independent variables. For instance, a multiple linear regression can tell you how much GPA is expected to increase (or decrease) for every one point increase (or decrease) in IQ.

Third, multiple linear regression analysis predicts trends and future values. The multiple linear regression analysis can be used to get point estimates.

When selecting the model for the multiple linear regression analysis, another important consideration is the model fit. Adding independent variables to a multiple linear regression model will always increase the amount of explained variance in the dependent variable (typically expressed as R²). Therefore, adding too many independent variables without any theoretical justification may result in an over-fit model.

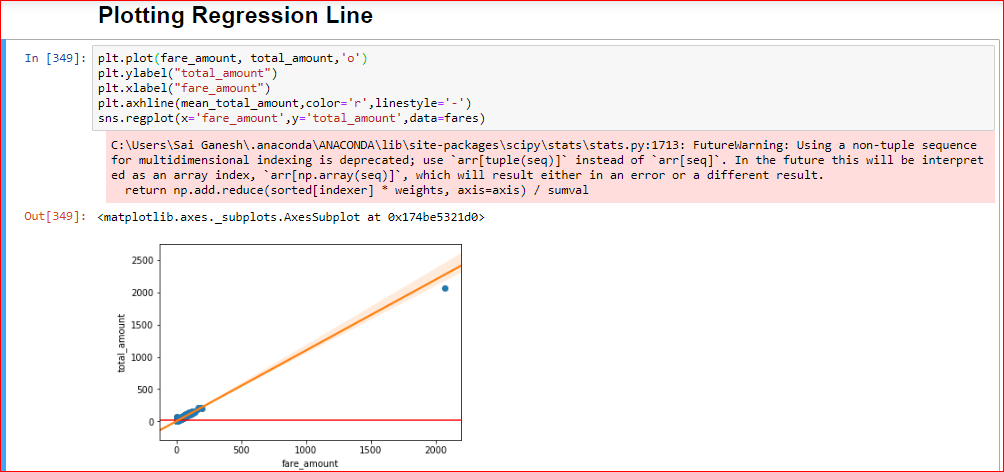
****

****

Here we will calculate the mean. Then we will plot the graph using the **matplotlib** library and the graph we were to plot is going to be scatter plot.

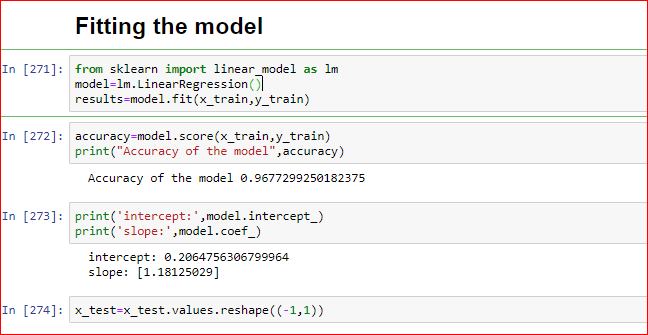
****

Then we will plot the regression line. Here the regression line will be joining the points which are most similar .

****

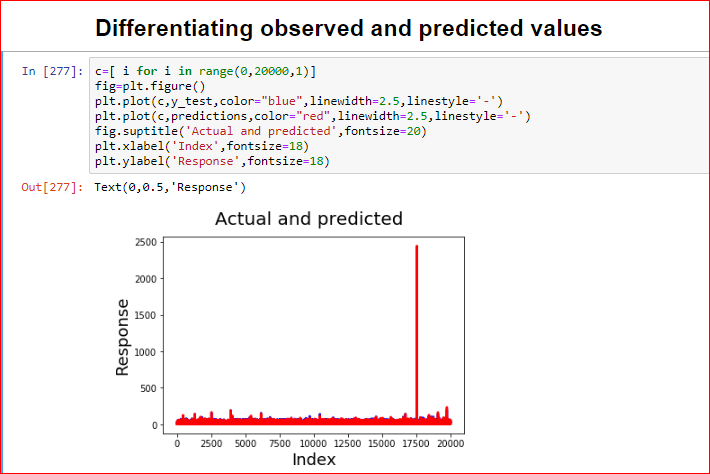
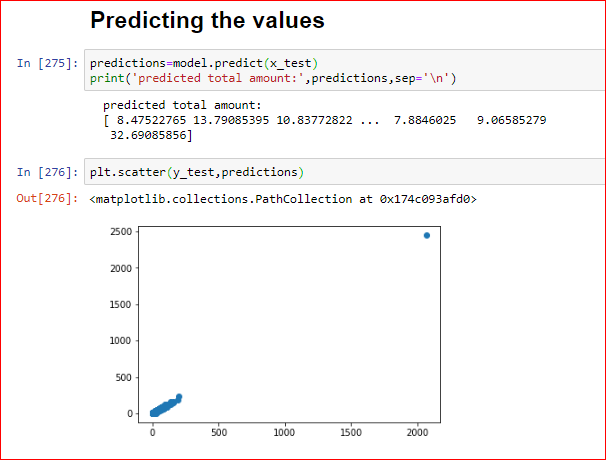
Seaborn is a Python data visualization library based on [matplotlib](https://matplotlib.org/). It provides a high-level interface for drawing attractive and informative statistical graphics.****

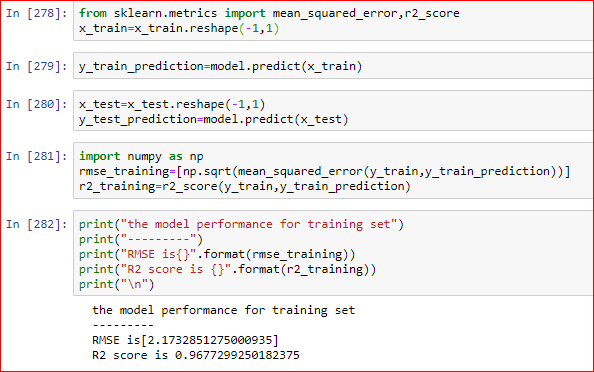
Next, we will try and split the data into training data and test data. Here we will use the Sci-Kit Learn python library function. Here we will mention the **test\_size** which will be the testing value which shall always be less than that of **training\_size** for better accuracy.The **random\_state** is used to take the values on a random.

****

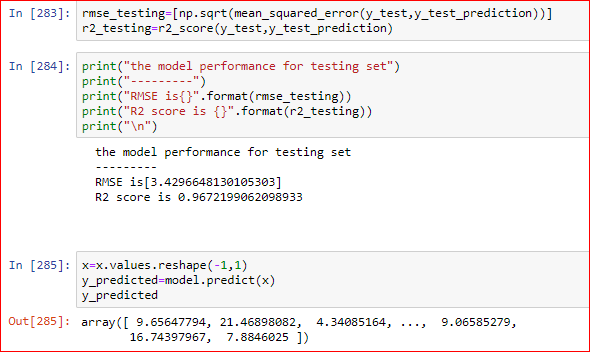
Then we will be fitting the values, fitting is used to fit the data values which are mostly given by the user.

We will next predict the values using **predict** function which is used to predict the test values. Then we draw a scatter plot.

****

****

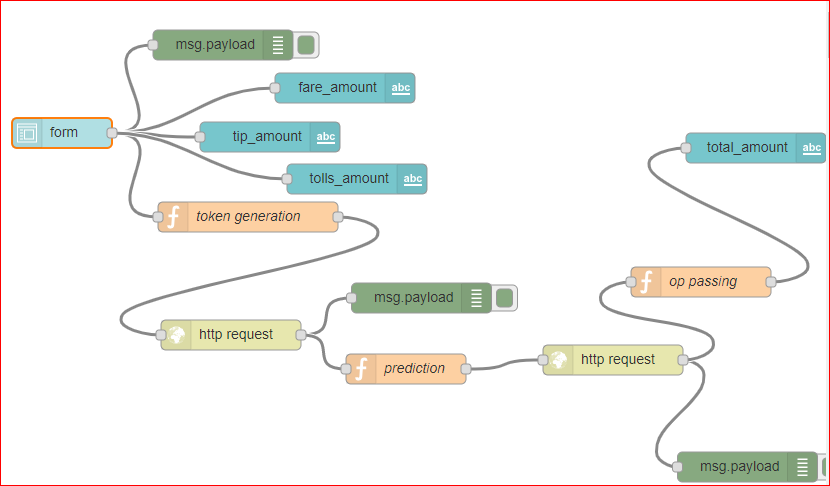
Here we find the RMSE and R2 square and for an accurate model the R2 square should be high and the RMSE should be low. Here we will calculate the RMSE and R2 values for both training and test set.

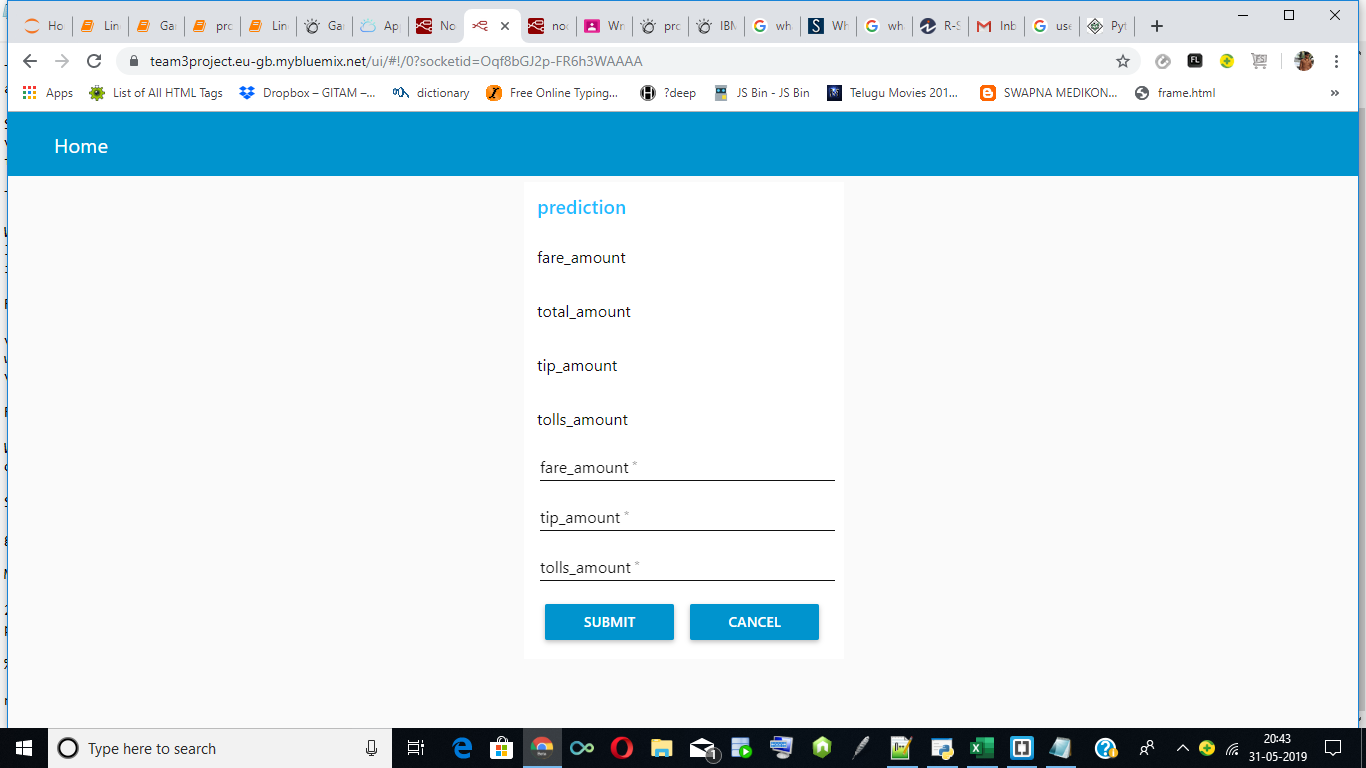
****

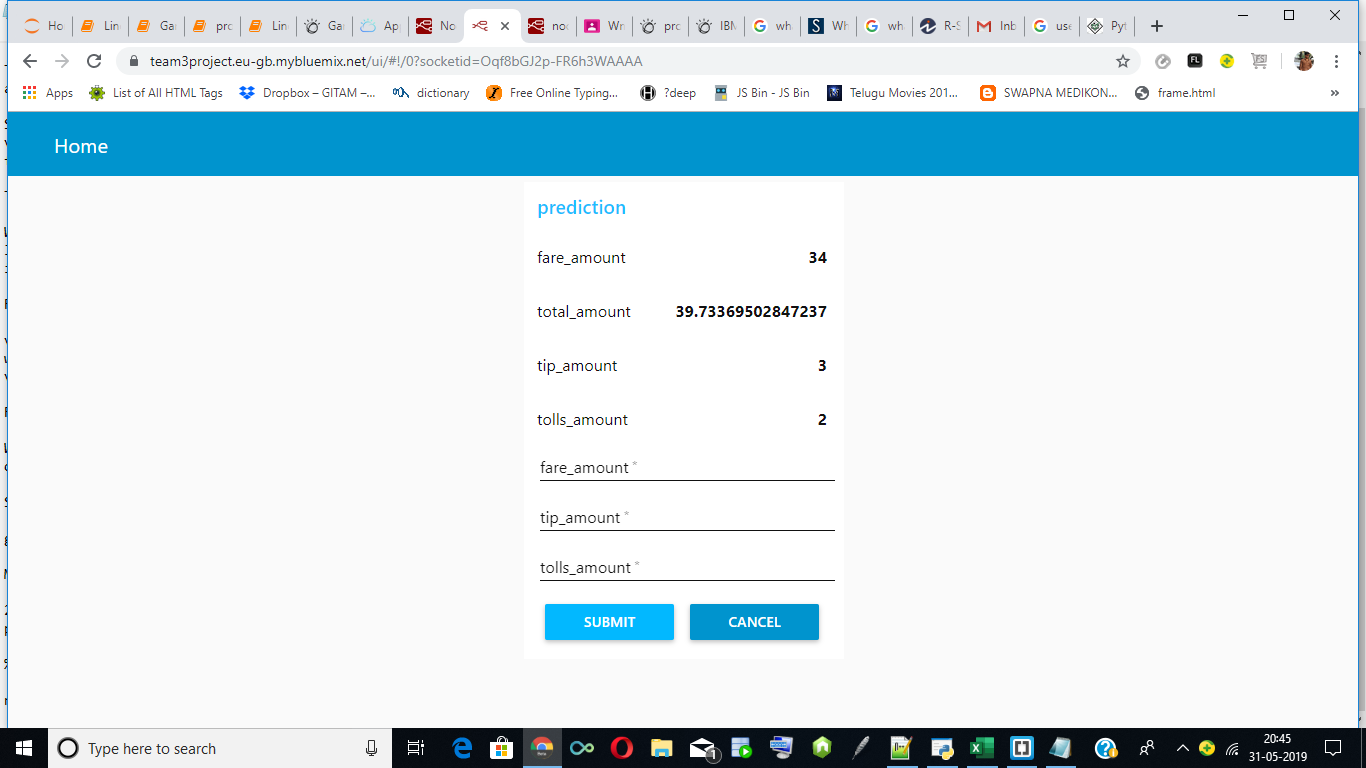
**NODE RED**

Node-RED is a programming tool for wiring together hardware devices, APIs and online services. Primarily, it is a visual tool designed for the Internet of Things, but it can also be used for other applications to very quickly assemble flows of various services.

NODE RED FLOW







**Findings and Suggestions**

Through Exploratory Data Analysis, we use linear regression models each were built for data split in 60-40 and partition.

60-40 split

 The Accuracy for single linear regression is 96.2%. It had a R square value for training is **0.967** and RMSE value for training is 2.17.

 The Accuracy for single linear regression is 96.2%. It had a R square value for testing is 0.16and RMSE value for testing is 3.429.

 The Accuracy for multiple linear regression is 99.9

The slope for single linear regression model is 1.1812 and intercept is 0.2064.The slope for multiple linear regression model when we took input variables fare amount toll\_amount,tip\_amount and intercept is [0.9971,1.007,0.9798] and intercept is 0.8494.

**Conclusion**

Of all the linear regression models built using different data splits, the linear regression model built using 60-40 data split with linear regression model showed comparatively more accuracy, Rsquare, RMSE score.

From the above parameters it can be concluded that

Linear and multiple linear regression model with 60-40 data split is a better model to analyze the given data set.

In this project, we have proposed methods for prediction for prediction of taxi fares using machine learning techniques. The system was implemented by comparing all other models and their performance was evaluated by using fare and taxes we determined the total charge this model used to reduce burden on taxi drivers.

**Bibliography and Reference**

1. www.kaggle.com

2. www.quora.com

3. www.wikkipedia.com