**Chicago taxi trips – tip prediction**

**Introduction:**

This dataset includes taxi trips for 2016, reported to the City of Chicago in its role as a regulatory agency. Taxicabs in Chicago, Illinois, are operated by private companies and licensed by the city. There are about seven thousand licensed cabs operating within the city limits. Licenses are obtained through the purchase or lease of a taxi medallion which is then affixed to the top right hood of the car. To protect privacy but allow for aggregate analyses, the Taxi ID is consistent for any given taxi medallion number but does not show the number, Census Tracts are suppressed in some cases, and times are rounded to the nearest 15 minutes. Due to the data reporting process, not all trips are reported but the City believes that most are. Using this data, an attempt to predict whether a driver will get a tip or not.

**Problem Statement:**

Cab drivers aren't paid on a commission though they pay so many taxes and fees. They either own the cab+medallion (which alone costs 70-80k+car cost) and operate as part of an association, or the lease a cab from a medallion/cab for the week on a given shift from an owner that belongs to a given association. Their pay is all of their fare+tips minus the lease fee, gas, and all of the fees. There are a million fees that they pay out--MPEA/other taxes, credit card processing fees, voucher processing fees, several types of insurance, fines from the city, association fees, etc. Hence tipping a cab driver is always fair enough.

Taxi drivers should be tipped 10 percent to 20 percent of the fare, which you should be able to leave when you pay with a credit card. However, a cash tip is always preferred.

Standard tip for a cab driver is 15%. For under $10.00 fare, usually the cab drivers get the change plus $1.00. For trips (around $35), customers usually go with about $5 - $10, depending on whether the driver drives safely and whether he assists with bags. For $20.00, customers would go with about $3 - $5 tip, assuming the drive was relatively safe. If the behaviour and driving style of the cab driver is not satisfactory, he/she simply gets change to the dollar ($6.45 gets $7).

Hence, trying to predict if a taxi driver gets a tip or not for the rides.

**Dataset:**

The dataset is taken from Kaggle. It includes taxi trips for 2016, reported to the City of Chicago in its role as a regulatory agency. It has more than 10 million observations for each month in the year 2016.

The data includes the following fields:

1. taxi\_id – ID assigned to each taxi

2. trip\_start\_timestamp – date and time when the trip started

3. trip\_end\_timestamp – date and time when the trip ended

4. trip\_seconds – total seconds taken to complete the trip

5. trip\_miles – total miles travelled

6. pickup\_census\_tract – neighbourhood from where the customer was picked up

7. dropoff\_census\_tract – neighbourhood where the customer was dropped off

8. pickup\_community\_area – area of pickup

9. dropoff\_community\_area – area of drop off

10. fare – charge of the taci ride

11. tips – tips given to thr driver

12. tolls – amount paid for tolls

13. extras – extra amount included

14. trip\_total – total trip amount

15. payment\_type – type of payment made

16. company – company to which the taxi belongs

17. pickup\_latitude – latitude from where customer was picked up

18. pickup\_longitude – longitude from where customer was picked up

19. dropoff\_latitude – latitude where customer was dropped off

20. dropoff\_longitude – longitude where customer was dropped off

**Data Limitations:**

The dataset was comprehensive with few missing values. Columns which were not needed for the analysis were removed. The dataset imported to Python was stored in the dataframe where NAN values were checked and removed. In the taxi trip data, few observations had 0 pickup\_latitude, 0 pickup\_longitude, 0 dropoff\_latitude, 0 dropoff\_longitude. Hence, these records were removed as it was of no use having these observations.

**Data Wrangling:**

The dataset was comprehensive with few missing values. It required some cleanup and reformatting. The steps taken are described below. Columns which were not needed for the analysis were removed for example pickup census tract as this column did not have any values included, due to security purposes. The dataset imported to Python was stored in the dataframe where NAN values were checked and removed. These transformations were helpful to conduct preliminary exploration and data visualization. In the taxi trip data, few observations had 0 pickup\_latitude, 0 pickup\_longitude, 0 dropoff\_latitude, 0 dropoff\_longitude. Hence, these records were removed as it was of no use having these observations. After considering all these factors from the data and cleaning up the data, now the data is ready for further analysis.

**Exploration and Graphs:**

1. Skewness:

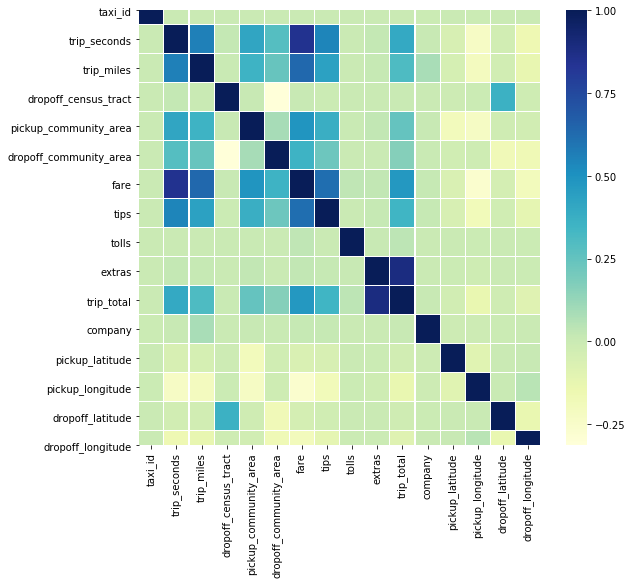


Skewness in statistics represents an imbalance and an asymmetry from the mean of a data distribution. In a normal data distribution with a symmetrical bell curve, the mean and median are the same. In a skewed data distribution, the median and the mean are different values.

If skewness is less than -1 or greater than 1, the distribution is highly skewed. If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

As the skewness is 7.12, we can say that the distribution here is highly skewed.

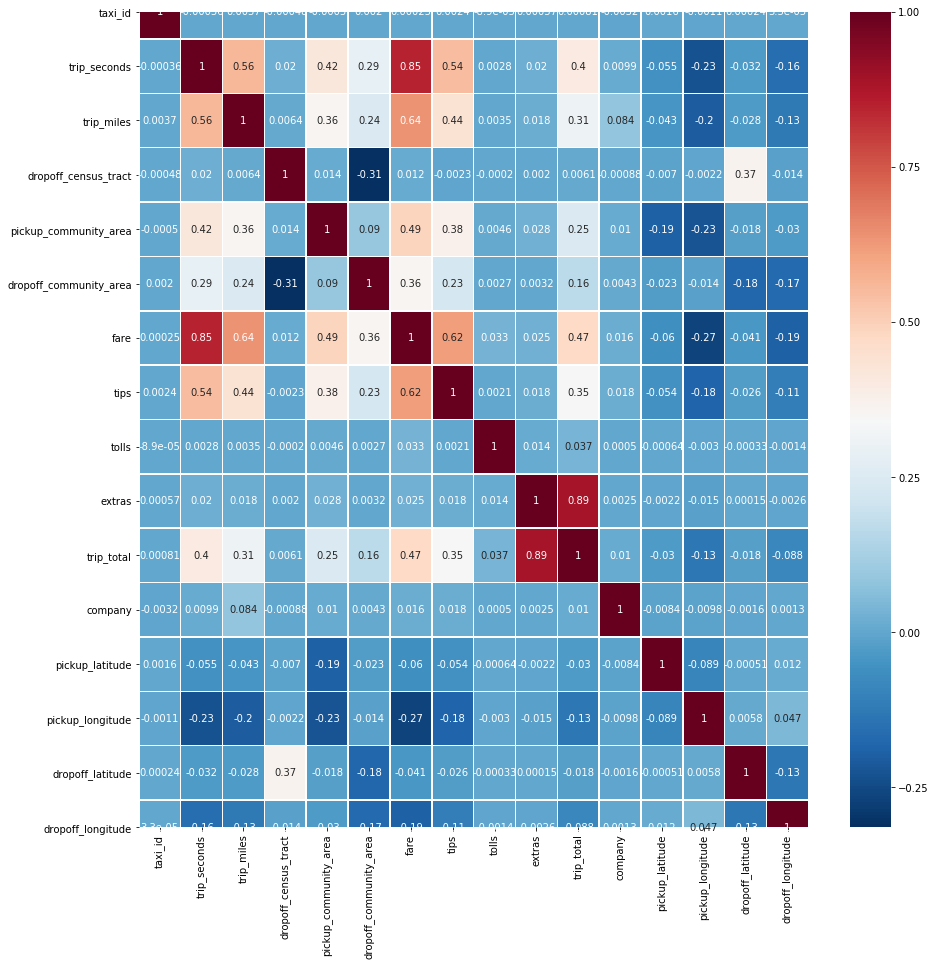
1. Correlation Matrix:



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. A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another.

Here, we can understand which are the variables that are correlated with each other according to the color gradient.

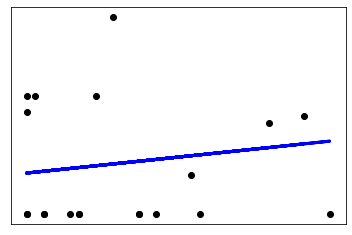
1. Pearson’s Correlation Heatmap



**Pearson’s correlation coefficient** is the test statistics that measures the statistical relationship, or association, between two continuous variables.  It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.  It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

From the above heatmap, we can understand that there are few fields that are correlated with 'tips'. The strongest correlation of tips is with fare. As the fare increases, the tips also tend to increase. Then comes the trip\_seconds and trip\_miles. As the time and distance of the trip increases, the tips given also increases. We can also see that the pickup and dropoff latitudes and longitudes have a negative correlation with tips. This means that as the latitudes and longitude increases, the tips tend to have a decrease

1. Linear Regression Plot



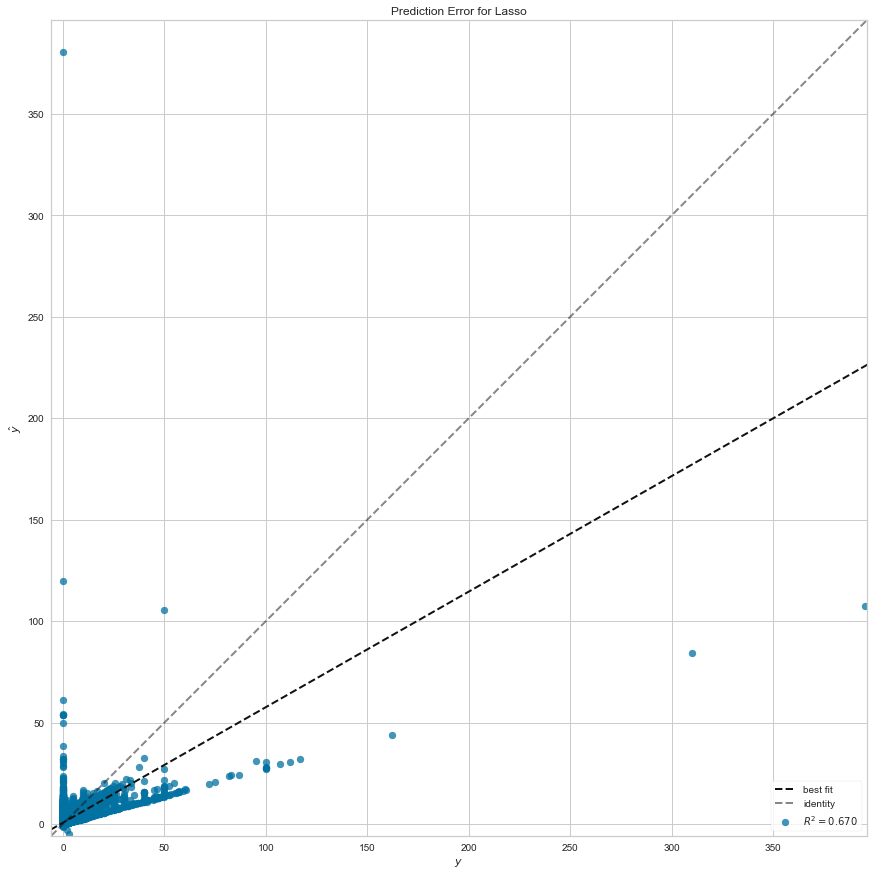
**Linear regression** is a **linear** approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things:

(1) Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

(2) Which variables in particular are significant predictors of the outcome variable?

1. Prediction-error Plot



A **prediction error plot** shows the actual targets from the dataset against the **predicted** values generated by our model. ... Data scientists can diagnose regression models using this **plot** by comparing against the 45 degree line, where the **prediction** exactly matches the model.

Here, the light colored black dotted line indicates the identity and the dark colored black dotted line indicates the best fit of the model.