**Chicago taxi trips – tip prediction**

**(Statistical Inference)**

After cleaning the data, the following steps were performed as a part of Inferential Statistics:

1. Identified the target variable
2. Checked for the skewness of the data
3. Checked the summary statistics of the target variable
4. Checked the correlation of variables with the target variable
5. Tried a Pearson’s correlation heatmap to find out the exact percentage of correlation
6. Scatter plots and linear regression plots were done with the strongly correlated variables.
7. Outlier detection was done
8. Linear Regression was carried out with different combinations of variables and the target variable.

**Identified the target variable**

As our goal is to predict the tips of each taxi ride, it was very simple to identify the target variable. Here, our target variable is ‘tips’.

**Checked for the skewness of the data**

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.

We calculated the skewness of the target variable using the function skew().

The skewness of target obtained is 7.123919535177937. As the skewness is 7.12, we can say that the distribution here is highly skewed.



**Checked the summary statistics of the target variable**

Summary statistics of Target Variable is as follows :-

count 7.564042e+06

mean 1.536023e+00

std 2.636438e+00

min 0.000000e+00

25% 0.000000e+00

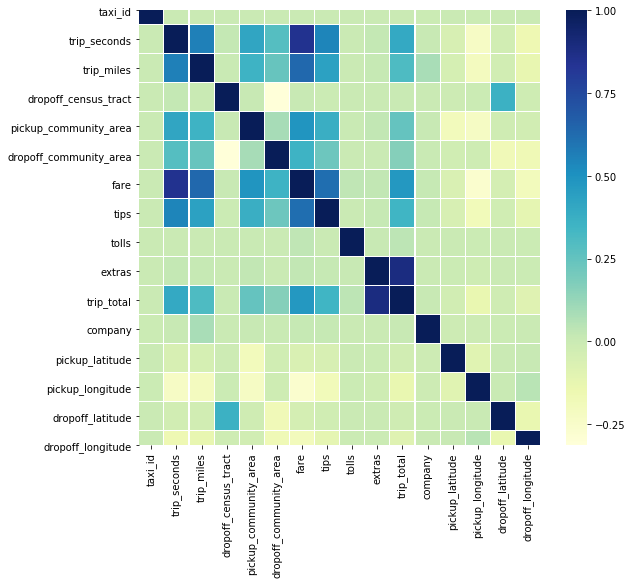
50% 0.000000e+00

75% 2.000000e+00

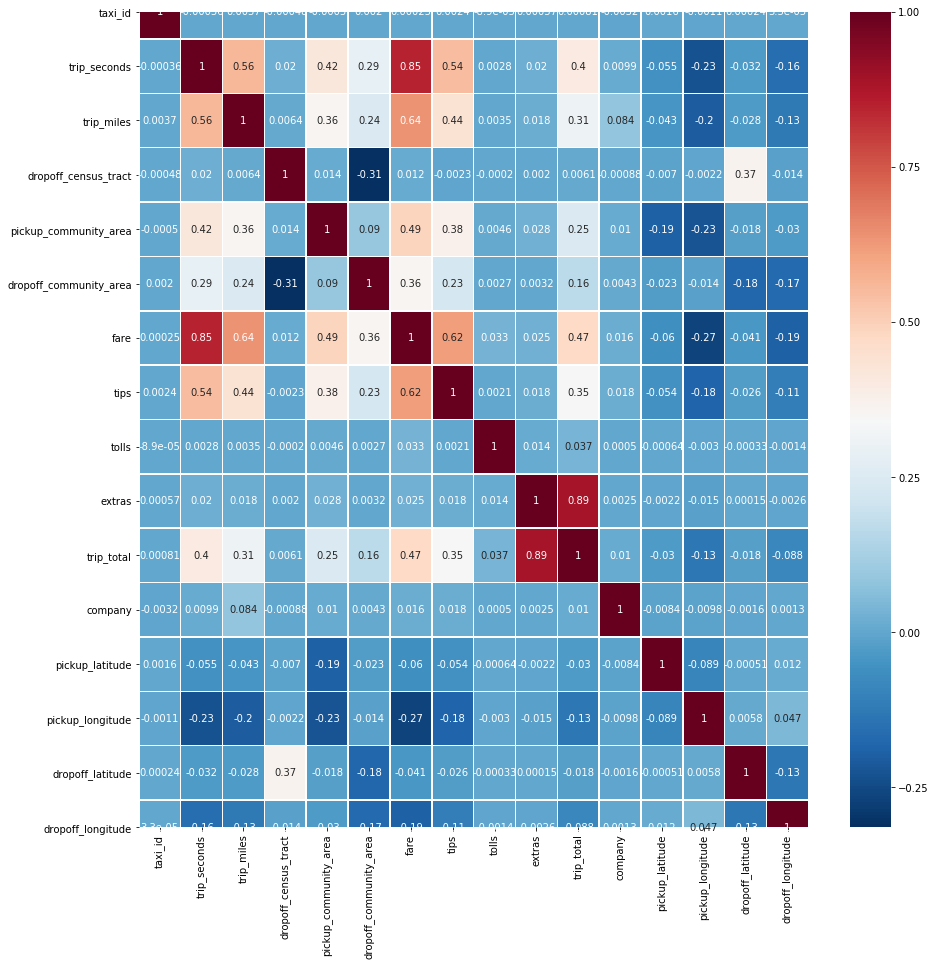
max 3.960000e+02

**Checked the correlation of variables with the target variable**

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.



**Tried a Pearson’s correlation heatmap to find out the exact percentage of correlation**



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.

**Scatter plots and linear regression plots were done with the srongly correlated variables.**

The strongly correlate variables are: fare, trip\_seconds and trip\_miles.

Hence scatter plots and regression plots were done for tese 3 variables.

Findings are as follows:

1. For the variable ‘fare’, we did see that a change in fare definitely has an impact on the tips as well. As the fare increases, the tips value also increases and vice-versa.
2. For the variable ‘trip\_seconds’, we saw that there is a change in tips as the time of the trip changes. As the trip\_seconds increases, the tips value also increases and vice-versa.
3. For the variable ‘trip\_miles’, we see that the tips value changes along with the change in the distance of the trip. As the trip\_miles increases, the tips value also increases and vice-versa.

**Outlier detection was done**

We loaded the data into Python, removed rows that had missing data. We then used a log transform to transform the data.

We also tried 1 variable separately to check the findings. We checked the for the variable ‘fare’ and were able to see the outliers present in column 'fare'.

**Linear Regression was carried out with different combinations of variables and the target variable.**

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?

We tried the linear regression with different combinations of the variables.

However, only 2 of the combinations yielded a very good prediction. The 2 combinations are:

i. Linear regression with all the variables related to the cost of the trip

ii. Linear regression with the positively correlated variables