**CAPSTONE PROJECT: MILESTONE REPORT**

**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.

It has 12 different .csv files. Each csv file for a month. The columns (variables) of all the 12 datasets are the same.

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.
* All the 12 datasets were combined together to have a collaborative analysis.
* The numerical variables and the categorical variables were separated to carry out better results.
* After considering all these factors from the data and cleaning up the data, now the data is ready for further analysis.

The below steps were followed to complete the tip prediction:

1. Importing important packages
2. Loading the data
3. Preprocessing and Exploratory Data Analysis (EDA)
4. Splitting taxi trip data into train and test
5. Feature selection and modelling
6. Modelling and plotting the graphs
7. Conclusions
8. Recommendations
9. **IMPORTING IMPORTANT PACKAGES**

**Modular programming** refers to the process of breaking a large, unwieldy programming task into separate, smaller, more manageable subtasks or **modules**. Individual modules can then be cobbled together like building blocks to create a larger application.

There are several advantages to **modularizing** code in a large application like simplicity, maintainability, reusability, scoping.

**Functions**, **modules** and **packages** are all constructs in Python that promote code modularization.

**Module** contents are made available to the caller with the import statement. The import statement takes many different forms, shown below:

import <module\_name>

from <module\_name> import <name(s)>

from <module\_name> import <name> as <alt\_name>

import <module\_name> as <alt\_name>

1. **LOADING THE DATA**

The datasets that we have are all of .csv types. CSV (comma-separated value) files are a common file format for transferring and storing data.

## Steps to Load CSV files to Python Pandas:

1. **Understanding file extensions and file types** – what do the letters CSV actually mean? What’s the difference between a .csv file and a .txt file?
2. **Understanding how data is represented inside CSV files –** if you open a CSV file, what does the data actually look like?
3. **Understanding the Python path and how to reference a file** – what is the absolute and relative path to the file you are loading? What directory are you working in?
4. **CSV data formats and errors** – common errors with the function.
5. **PREPROCESSING AND EXPLORATORY DATA ANALYSIS (EDA)**

The following steps were followed for preprocessing:

1. Viewed the data to understand its structure and fields
2. Listed more information about the given data like the data types, column names, number of rows etc.
3. Took the statistical summary of the data using describe().
4. Handled the missing values in the dataset
5. Combined all the 12 different datasets (.csv files) for a better combined analysis
6. Separated the categorical and the numerical variables
7. Checked the top absolute correlations

The Exploratory Data Analyses are as follows:

1. 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’.

We decided to choose tips over fare as it is a bit more challenging than finding fare. We cannot blindly come to a conclusion on which variable is actually affecting the tips. It clearly depends on the customer whether to provide a tip or no. It also depends on the driver’s behaviour, customer’s mood etc.

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).

However, from the data that we have, we have managed to find the variables that are strongly correlated with tips to help find the solution.

1. Checked the skewness of the target variable



The skewness of target is 7.123919535177937

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.

The **formula** used to find skewness is:

**Skew = 3 \* (Mean – Median) / Standard Deviation**.

This is known as an alternative Pearson Mode **Skewness**.

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. Calculated the summary statistic of the target variable

count 7.564042e+06

mean 1.536023e+00

std 2.636438e+00

min 0.000000e+00

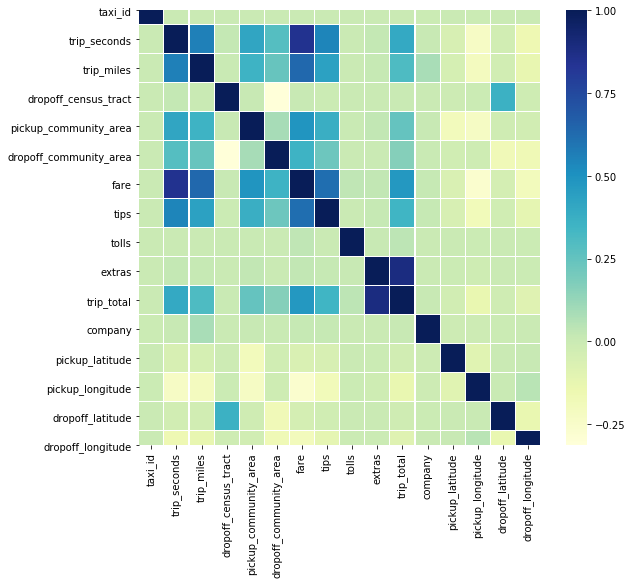
25% 0.000000e+00

50% 0.000000e+00

75% 2.000000e+00

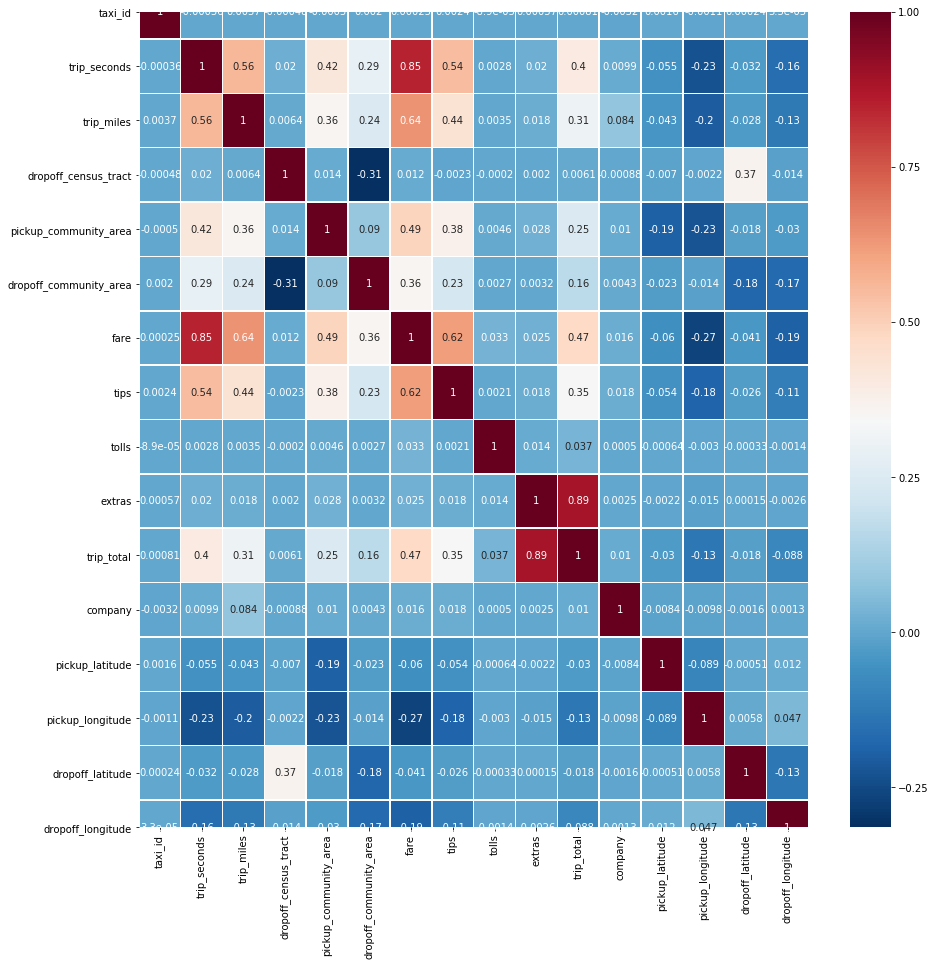
max 3.960000e+02

1. Checked the correlation between the variables



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.

1. Computed the Pearson’s correlation

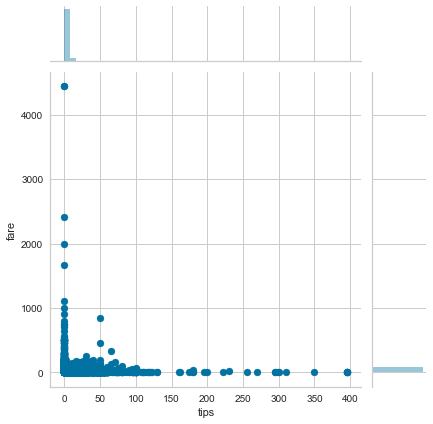


From the above Pearson’s correlation heatmap, we can understand that there are few fields that are correlated with 'tips'. The strongest correlation of tips is with fare. As the fare increases, the tips also tend to increase. Then comes the trip\_seconds and trip\_miles. As the time and distance of the 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.

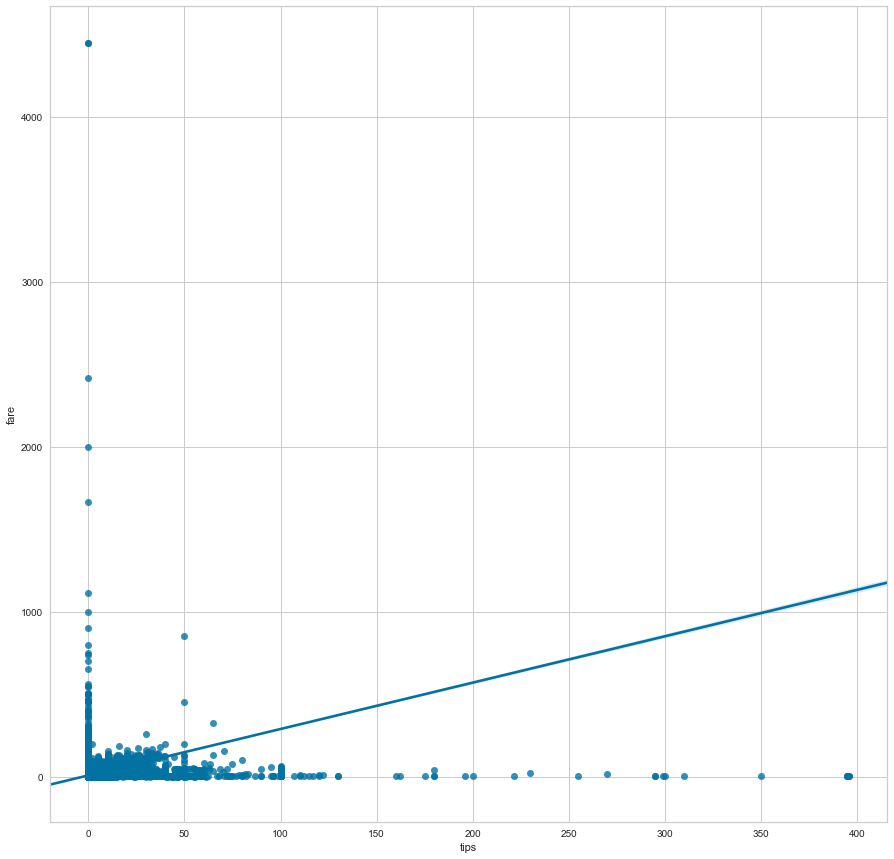
1. Scatter plots and regression plots were checked for the strongly correlated variables with the target variable.

* between target (tips) and 'fare'

Scatter plot:



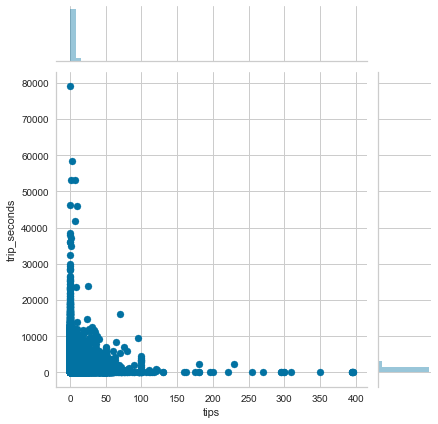
Regression Plot:



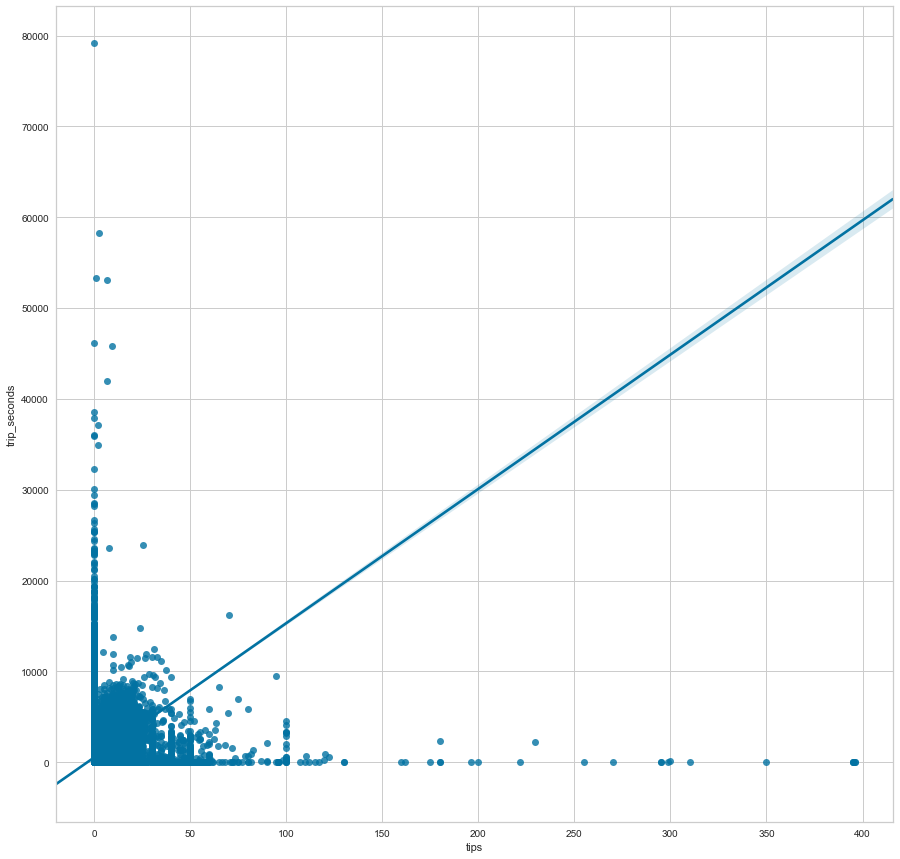
As plotted in the graphs above, we do see that a change in fare definitely has an impact on the tips as well. As the fare increases, the tips value also increases and the vice-versa.

* between target (tips) and 'trip\_seconds'

Scatter Plot:



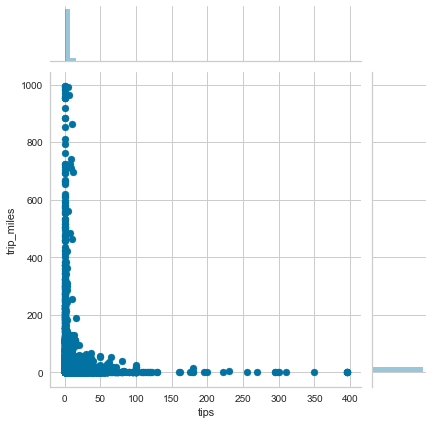
Regression Plot:



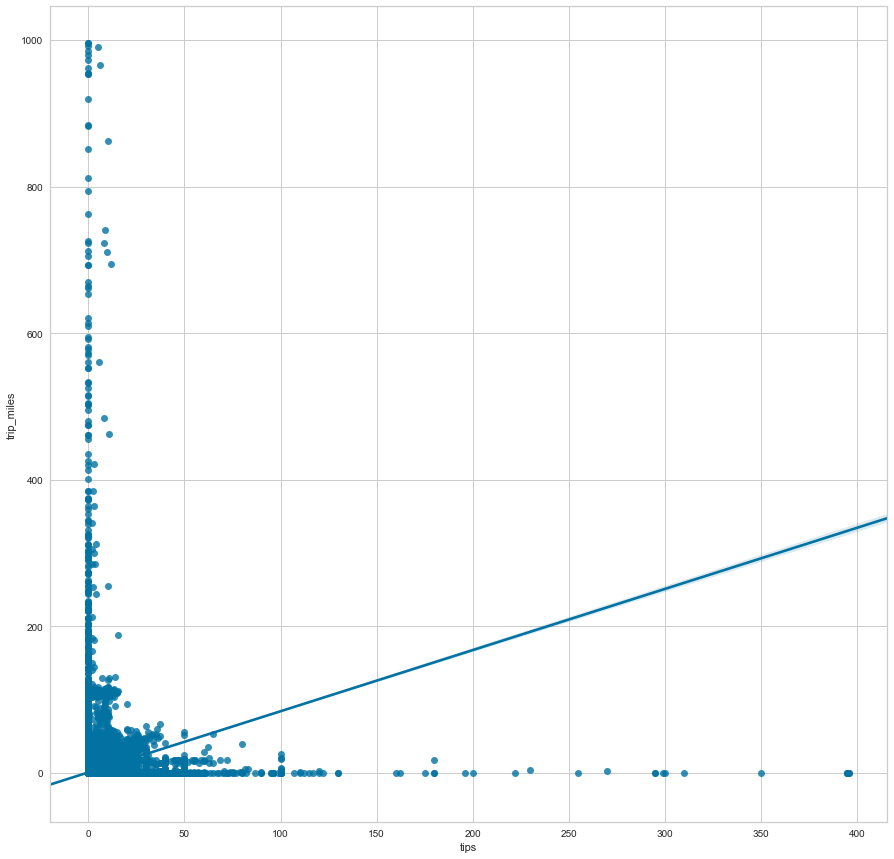
As plotted in the graphs above, we can see that there is a change in tips as and when there is a change in the time of the trip. As the trip\_seconds increases, the tips value also increases and vice-versa.

* target (tips) and 'trip\_miles'

Scatter Plot:



Regression Plot:



As shown in the graphs above, we can see that the tips value has an impact when there is a change in the distance of the trip. As the trip\_miles increases, the tips value also increases and vice-versa.

Listing the linearly correlated columns with Target

tips 1.000000

fare 0.618485

trip\_seconds 0.544883

trip\_miles 0.439544

pickup\_community\_area 0.376597

trip\_total 0.347734

dropoff\_community\_area 0.225430

extras 0.018396

company 0.017947

taxi\_id 0.002392

tolls 0.002108

dropoff\_census\_tract -0.002348

dropoff\_latitude -0.026455

pickup\_latitude -0.054043

dropoff\_longitude -0.112690

pickup\_longitude -0.181162

1. Outlier Detection

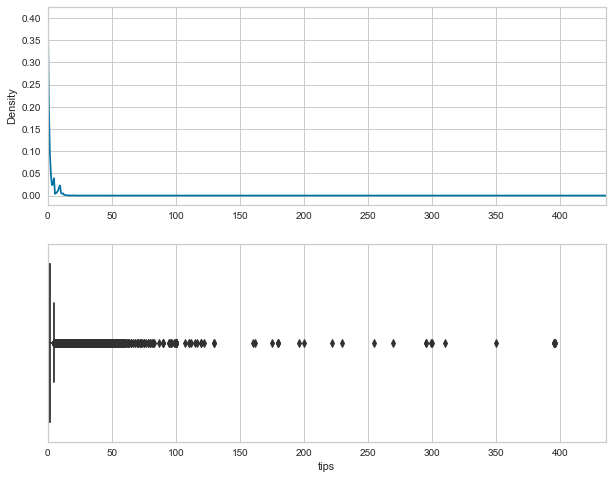
An outlier is an observation that appears to deviate markedly from other observations in the sample. Identification of potential outliers is important for the following reasons:

- An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).

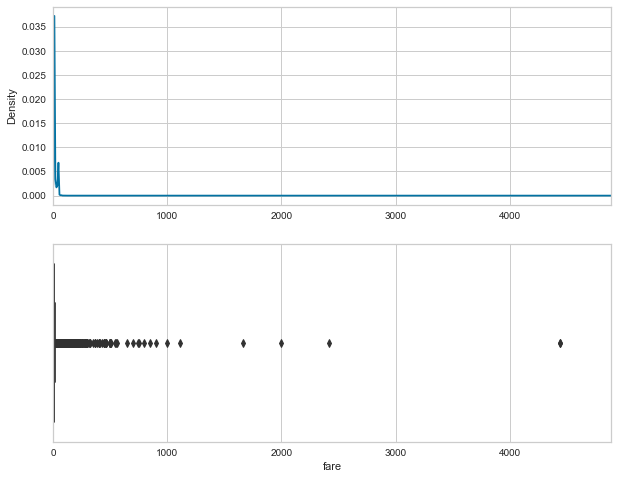
- In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any event, we typically do not want to simply delete the outlying observation. However, if the data contains significant outliers, we may need to consider the use of robust statistical techniques.

Detected outliers using box plots for every variable:

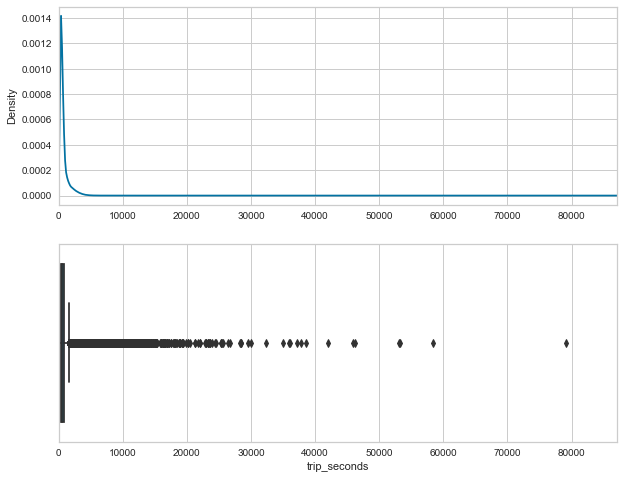
* tips



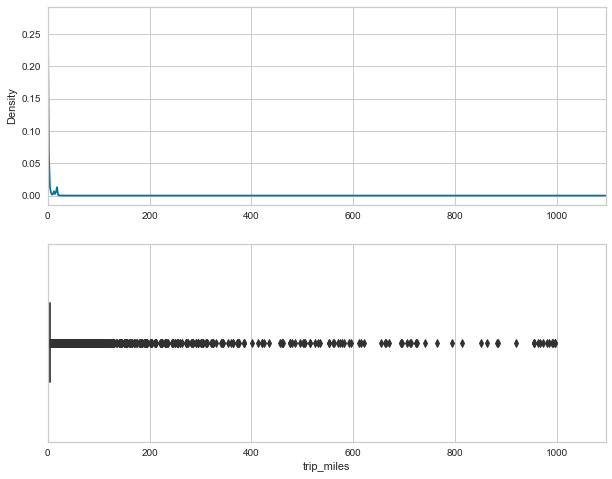
- fare



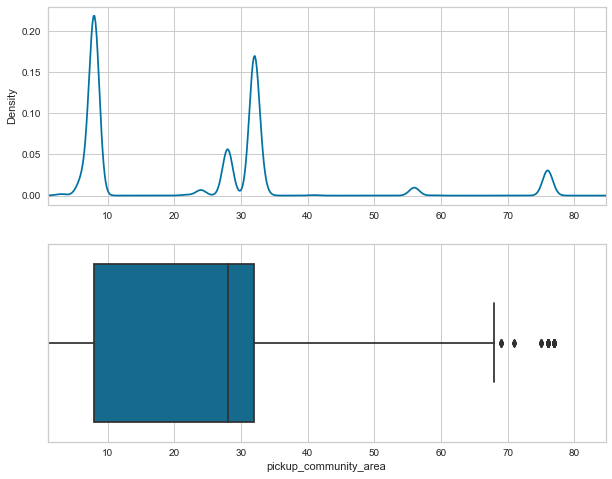
- trip\_seconds



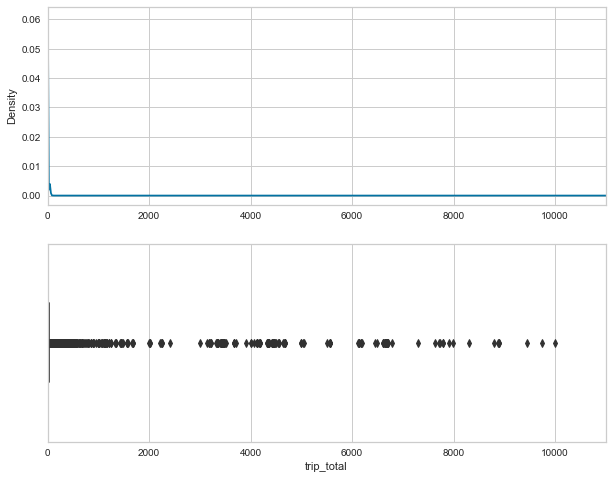
- trip\_miles



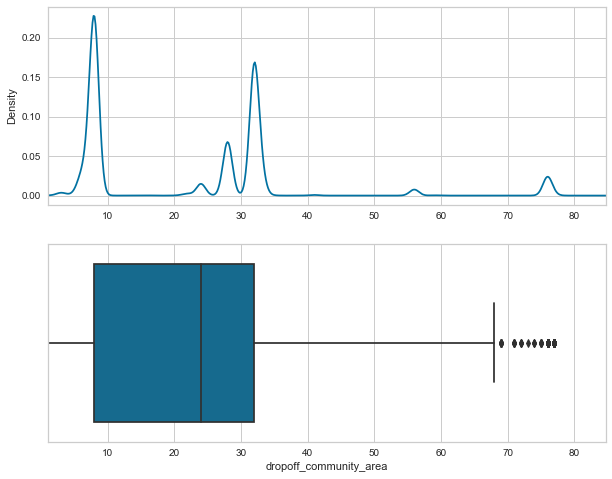
- pickup\_community\_area



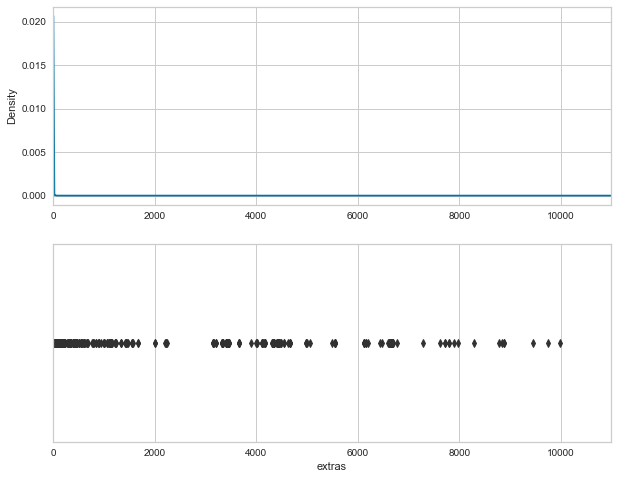
- trip\_total



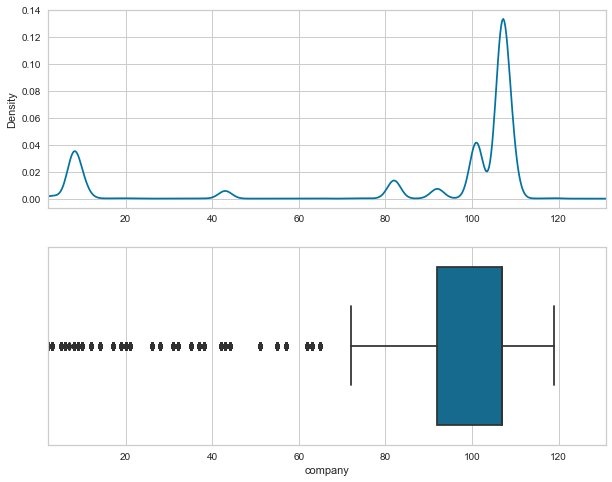
- dropoff\_community\_area



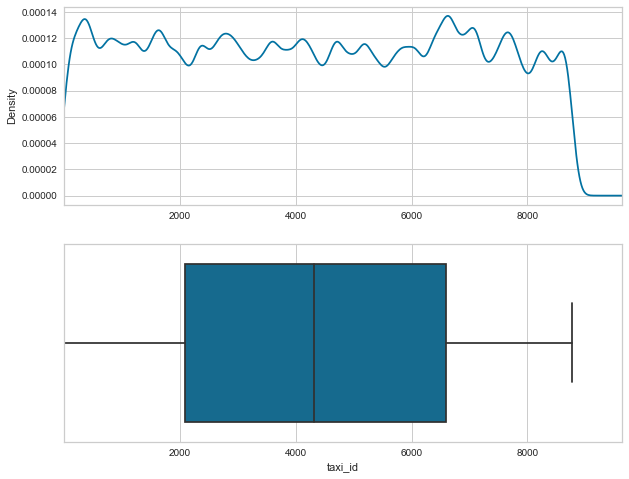
- extras

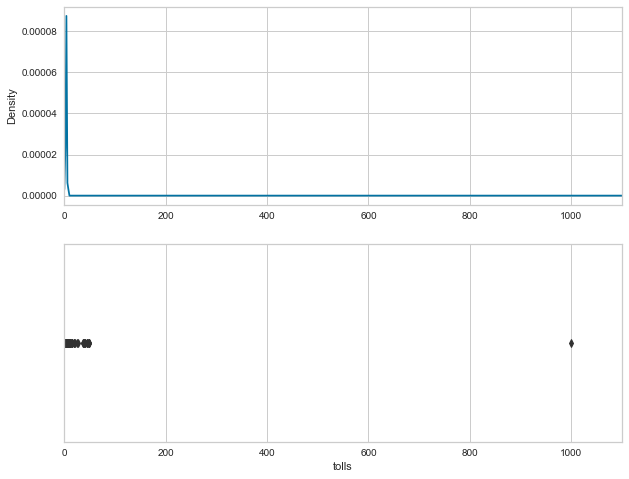


- company

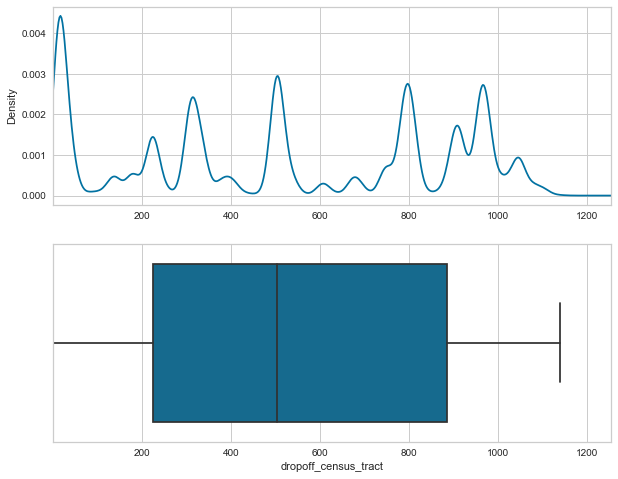


- taxi\_id

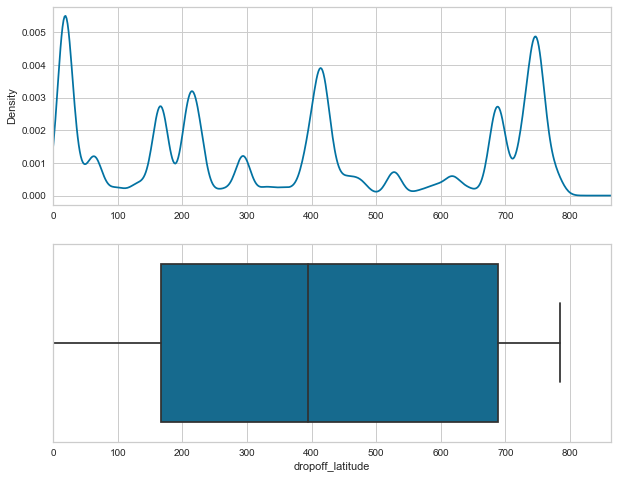


* tolls

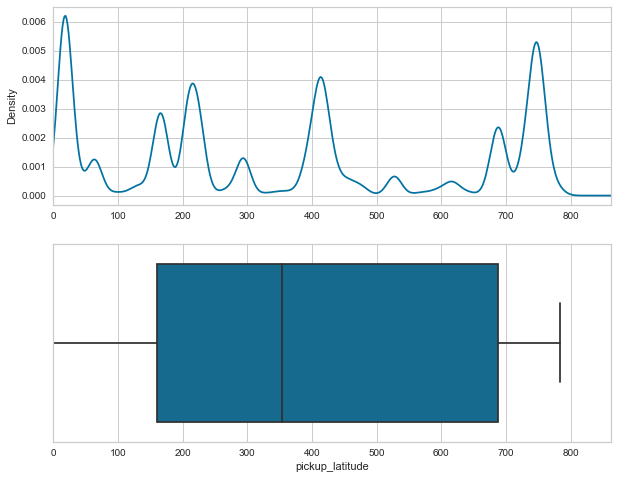
- dropoff\_census\_tract



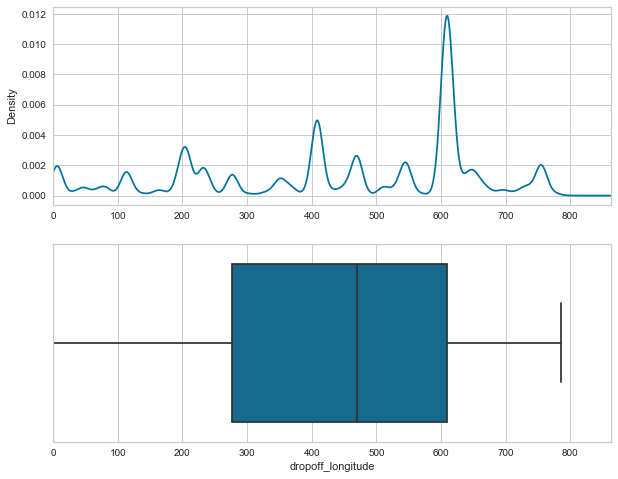
- dropoff\_latitude



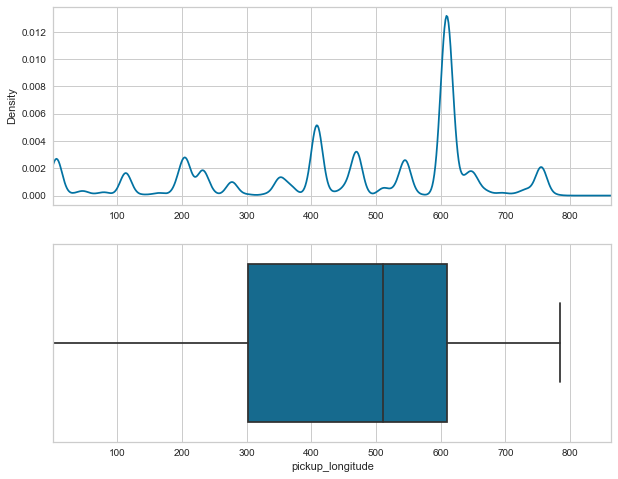
- pickup\_latitude



- dropoff\_longitude



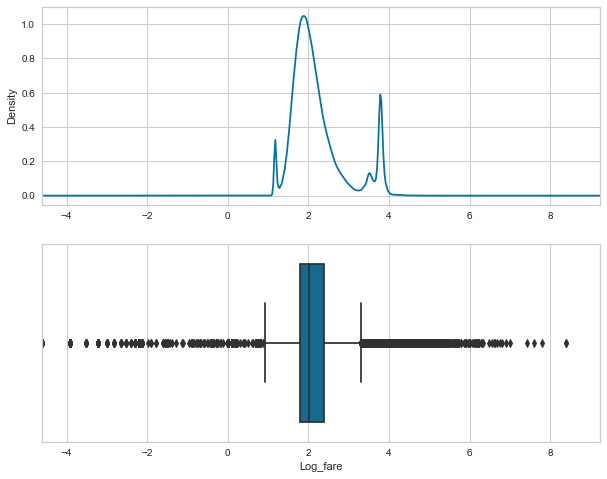
- pickup\_longitude



Checked Outlier Detection Of One Variable - (just to see the process,we worked with one variable and detected ouliers)

Variable selected: fare

We used a log transform to the variable ‘fare’ to transform the data as our target variable is highly skewed.



We loaded the data into Python, removed rows that had missing data. We then used a log transform to transform the data as our target variable is highly skewed.

The **log transformation** can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.

1. **SPLITTING TAXI TRIP DATA INTO TRAIN AND TEST**

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data (80%) is used for training, and a smaller portion of the data (20%) is used for testing.

From now onwards, we will be using the train data. We will keep the Test data aside to check the model performance.

1. **FEATURE SELECTION AND MODELLING**

Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested. Having irrelevant features in your data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression.

**LINEAR REGRESSION**

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?

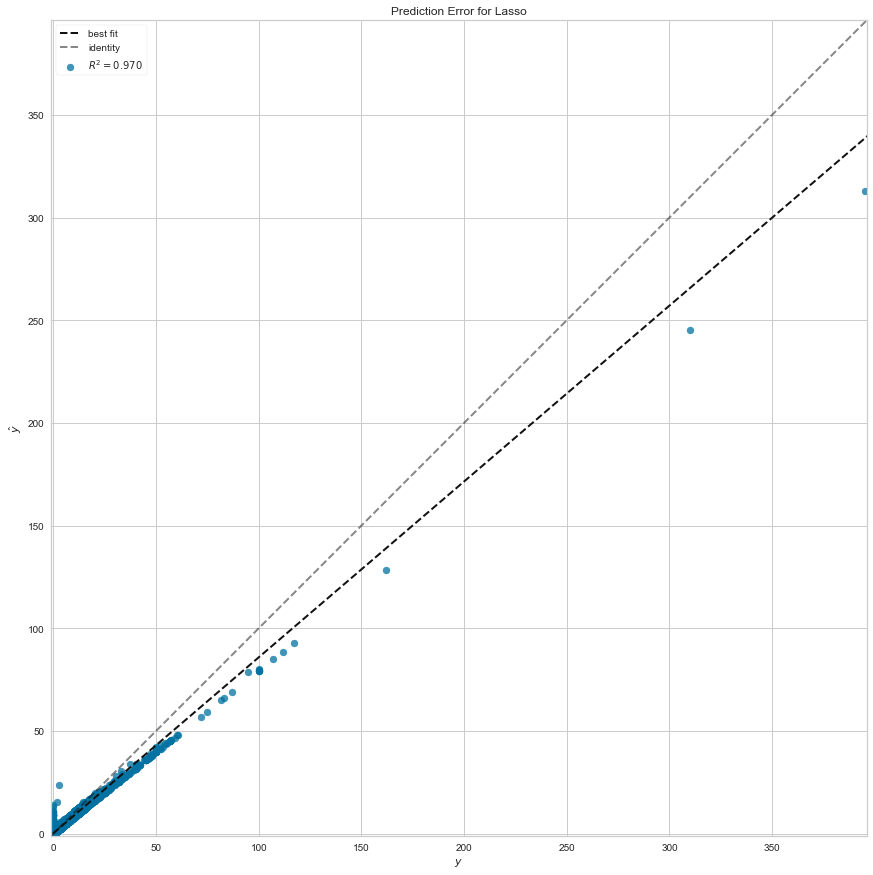
We will be trying a linear regression with multiple combinations of data to find out which are the variables contributing to the success of receiving tips. The combinations are as follows:

1. Linear regression with all the numerical variables
2. Linear regression with all the negatively correlated variables
3. Linear regression with all the variables related to the cost of the trip
4. Linear regression with the distance and duration variables of the trip
5. Linear regression with the variables related to community area
6. Linear regression with the variables related latitudes and longitudes
7. Linear regression with the positively correlated variables
8. **Linear regression with all the numerical variables**

- Actual vs Predicted value table:

|  |  |  |
| --- | --- | --- |
| **Row count** | **Actual** | **Predicted** |
| 0 | 2 | 2.00E+00 |
| 1 | 0 | -6.17E-12 |
| 2 | 9.2 | 9.20E+00 |
| 3 | 0 | -6.20E-12 |
| 4 | 0 | -6.20E-12 |
| ... | ... | ... |
| 1512804 | 1.5 | 1.50E+00 |
| 1512805 | 0 | -6.18E-12 |
| 1512806 | 14.6 | 1.46E+01 |
| 1512807 | 2 | 2.00E+00 |
| 1512808 | 3 | 3.00E+00 |
|  |  |  |
| 1512809 rows × 2 columns | | |

- Prediction-error graph:



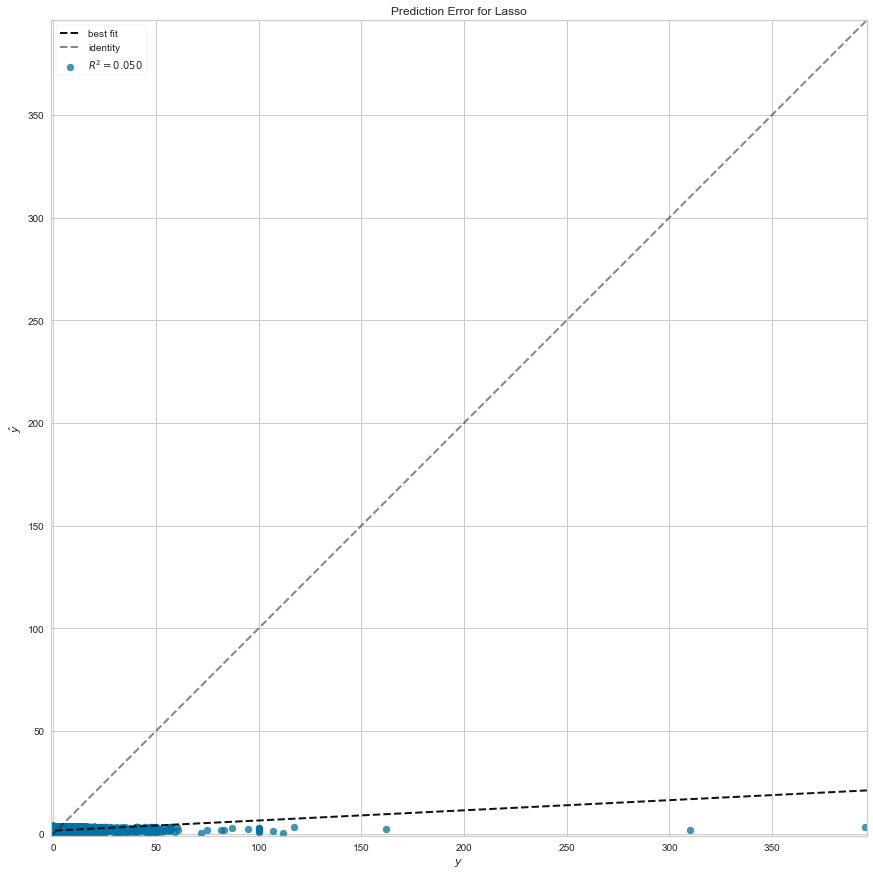
By looking into the predicted values for the linear regression with the numeric variables, we can see that the predictions are very closely successful for the actual values that are greater than 0. However, we do get incorrect (negative) predictions for the actual values that are equal to 0.

1. **Linear regression with all the negatively correlated variables**

* Actual vs Predicted value table:

|  |  |  |
| --- | --- | --- |
| **Row count** | **Actual** | **Predicted** |
| 0 | 2 | 0.787124 |
| 1 | 0 | 0.859128 |
| 2 | 9.2 | 1.578619 |
| 3 | 0 | 1.850171 |
| 4 | 0 | 1.221528 |
| ... | ... | ... |
| 1512804 | 1.5 | 1.282387 |
| 1512805 | 0 | 1.262293 |
| 1512806 | 14.6 | 1.872023 |
| 1512807 | 2 | 0.591227 |
| 1512808 | 3 | 1.200364 |
|  |  |  |
| 1512809 rows × 2 columns | | |

- Prediction-error graph:



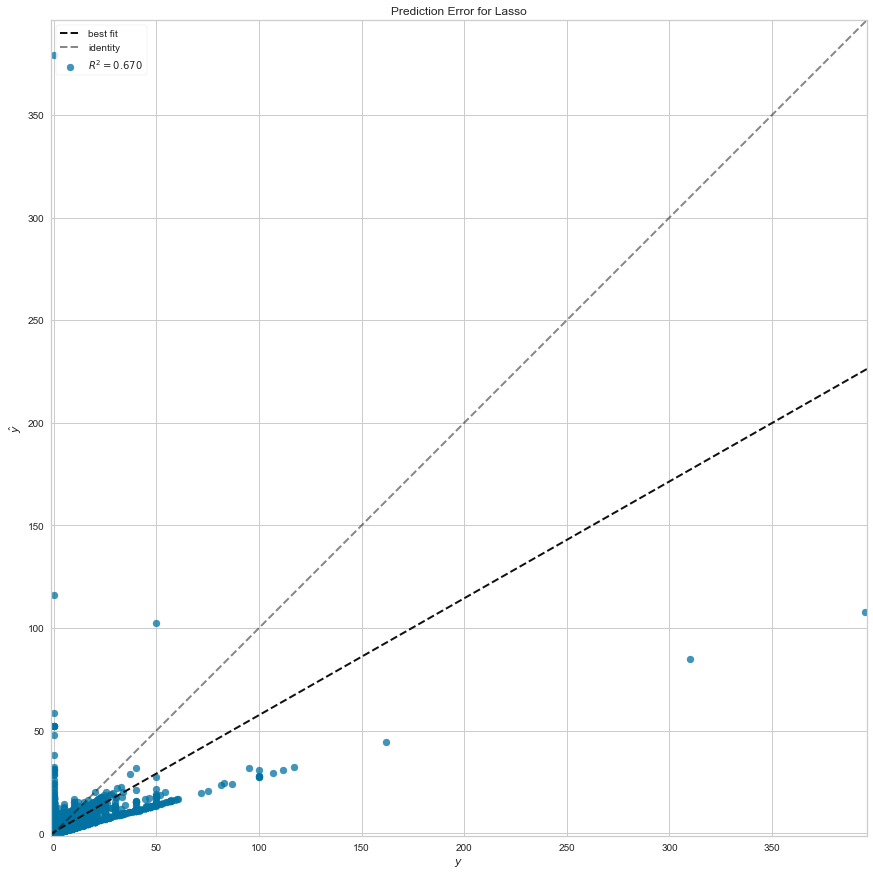
By looking into the predicted values for the linear regression with the negatively correlated variables, we can see that the predictions are very poor. We can hardly find a good prediction in the list. Hence we can come to a conclusion that the negatively correlated variables do not yield a good prediction for our target variable.

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

- Actual vs Predicted value table:

|  |  |  |
| --- | --- | --- |
| **Row count** | **Actual** | **Predicted** |
| 0 | 2 | 1.999991 |
| 1 | 0 | 0.000004 |
| 2 | 9.2 | 9.199989 |
| 3 | 0 | 0.000007 |
| 4 | 0 | 0.000001 |
| ... | ... | ... |
| 1512804 | 1.5 | 1.499998 |
| 1512805 | 0 | 0.000001 |
| 1512806 | 14.6 | 14.599965 |
| 1512807 | 2 | 1.999992 |
| 1512808 | 3 | 2.999989 |
|  |  |  |
| 1512809 rows × 2 columns | | |

* Prediction-error graph:



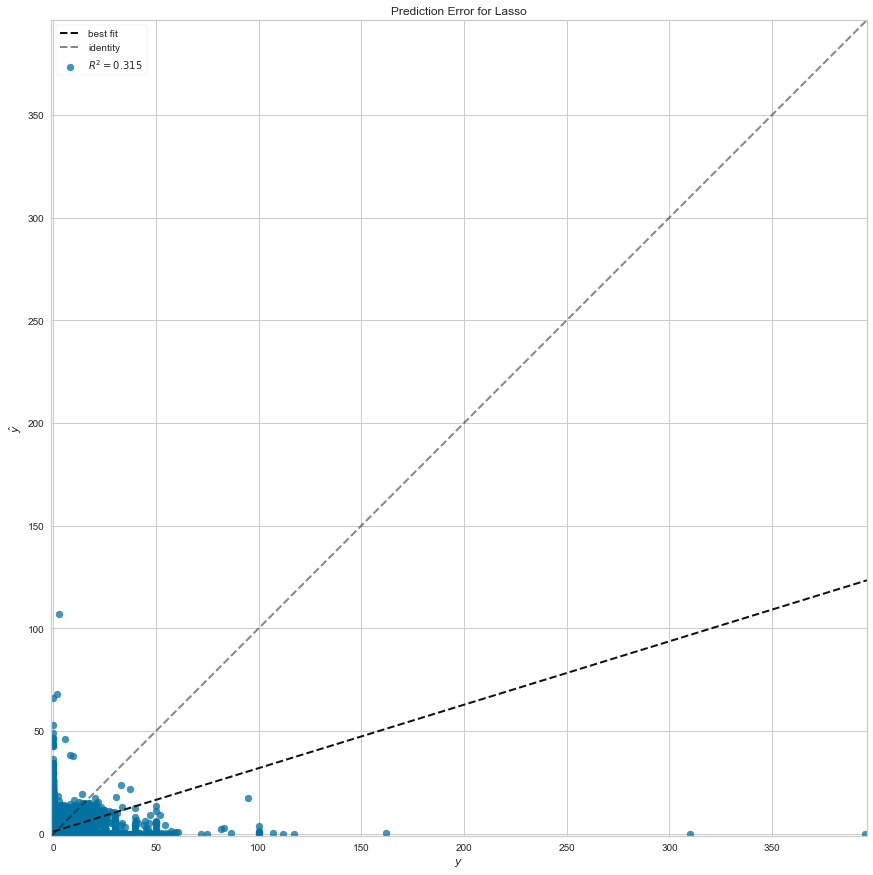
Looking at the predicted values for the linear regression with all the variables related to the cost of the trip, we see that the predictions are closely accurate. We can therefore come to a conclusion that the variables related to the cost of the trip are closely related to the target variable (tips). An increase in the cost variables will definitely have an increase in the target variable and vice versa.

1. **Linear regression with the distance and duration variables of the trip**

- Actual vs Predicted value table:

|  |  |  |
| --- | --- | --- |
| **Row count** | **Actual** | **Predicted** |
| 0 | 2 | 0.512796 |
| 1 | 0 | 1.79846 |
| 2 | 9.2 | 5.413272 |
| 3 | 0 | 2.455773 |
| 4 | 0 | 0.801127 |
| ... | ... | ... |
| 1512804 | 1.5 | 1.75427 |
| 1512805 | 0 | 0.993348 |
| 1512806 | 14.6 | 6.201426 |
| 1512807 | 2 | 0.897238 |
| 1512808 | 3 | 1.382866 |
|  |  |  |
| 1512809 rows × 2 columns | | |

* Prediction-error graph:



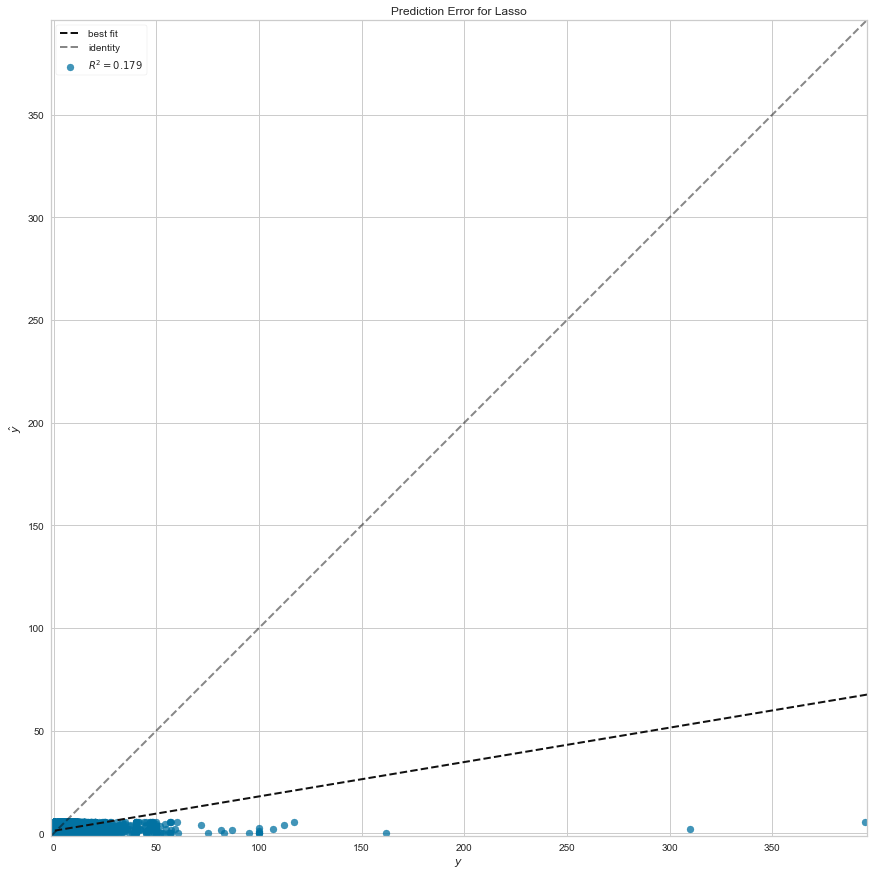
Seeing the predicted values for the linear regression with the variables related to the distance and duration of the trip, we can see that the predictions are moderate. We can hardly find a good prediction or a bad prediction in the list. Therefore, we can say that the duration and distance variables when taken separately do not play a vital role in predicting the target.

1. **Linear regression with the variables related to community area**

- Actual vs Predicted value table:

|  |  |  |
| --- | --- | --- |
| **Row count** | **Actual** | **Predicted** |
| 0 | 2 | 1.561498 |
| 1 | 0 | 1.030958 |
| 2 | 9.2 | 3.642602 |
| 3 | 0 | 1.530894 |
| 4 | 0 | 0.296451 |
| ... | ... | ... |
| 1512804 | 1.5 | 0.786122 |
| 1512805 | 0 | 1.030958 |
| 1512806 | 14.6 | 2.377555 |
| 1512807 | 2 | 1.030958 |
| 1512808 | 3 | 0.90854 |
|  |  |  |
| 1512809 rows × 2 columns | | |

* Prediction-error graph:



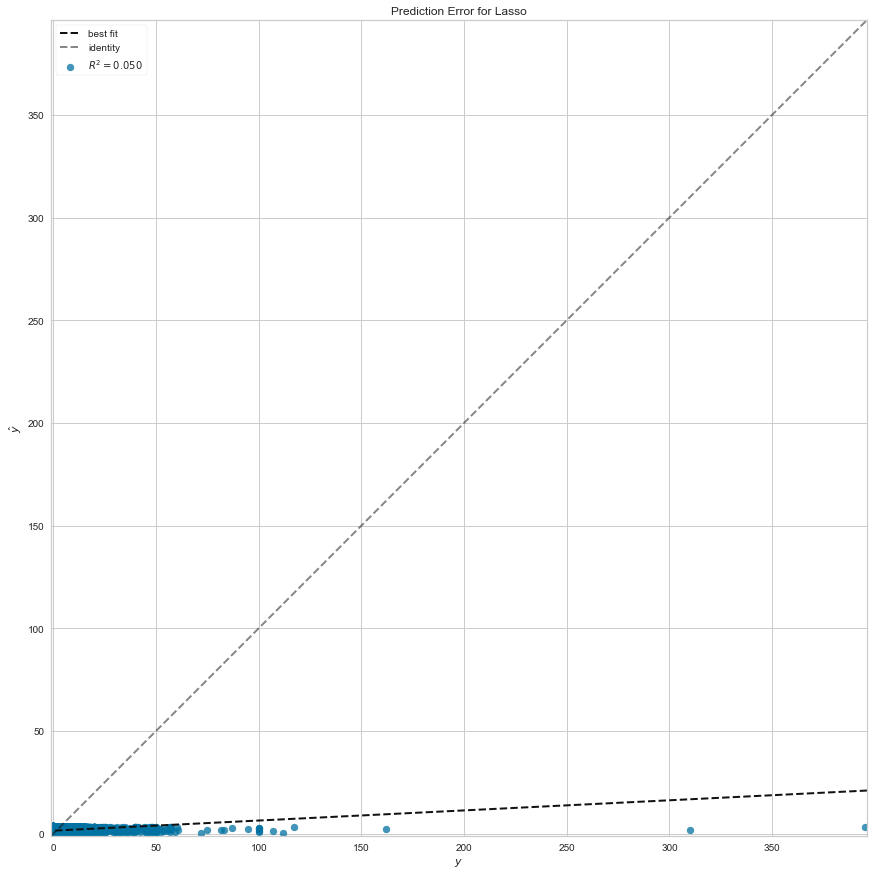
By looking into the predicted values for the linear regression with the variables related to community area, we see that the predictions are not too good. We can hardly find a perfect prediction in the list. Hence we can come to a conclusion that the variables related to community area do not make much change in the prediction for our target variable.

1. **Linear regression with the variables related latitudes and longitudes**

- Actual vs Predicted value table:

|  |  |  |
| --- | --- | --- |
| **Row count** | **Actual** | **Predicted** |
| 0 | 2 | 0.825822 |
| 1 | 0 | 0.886632 |
| 2 | 9.2 | 1.58686 |
| 3 | 0 | 1.816661 |
| 4 | 0 | 1.181276 |
| ... | ... | ... |
| 1512804 | 1.5 | 1.281314 |
| 1512805 | 0 | 1.289858 |
| 1512806 | 14.6 | 1.880837 |
| 1512807 | 2 | 0.607813 |
| 1512808 | 3 | 1.192335 |
|  |  |  |
| 1512809 rows × 2 columns | | |

* Prediction-error graph:

****

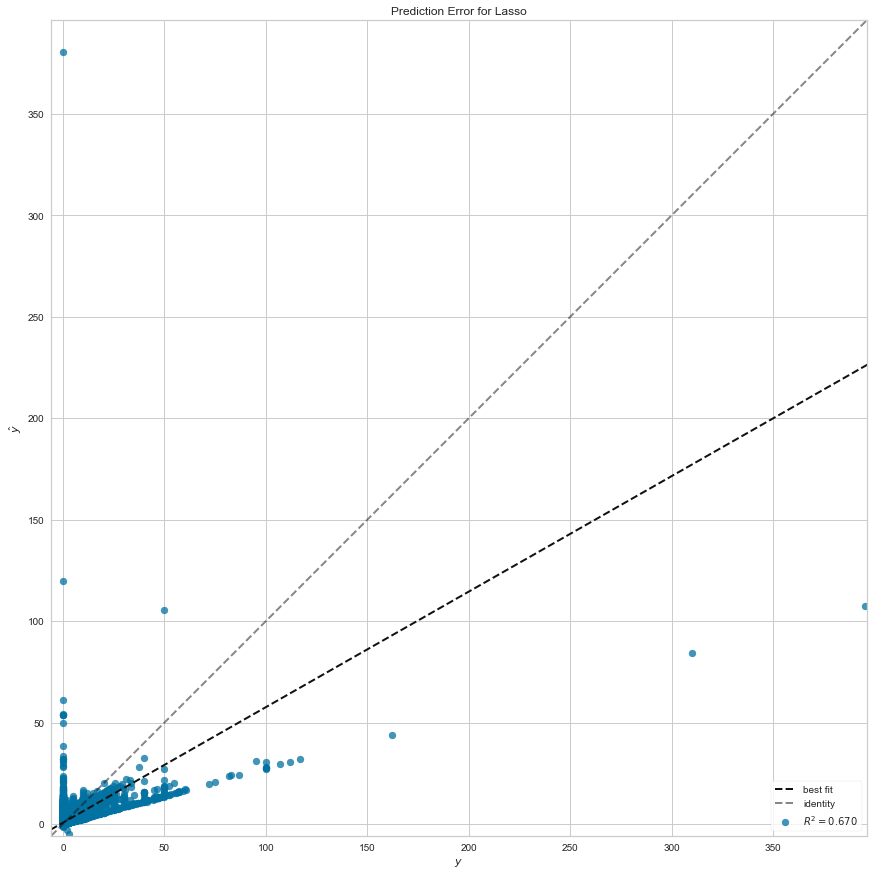
Variables related to latitudes and longitudes have a negative correlation with the target variable. Checking the predicted values for the linear regression with the variables related to latitudes and longitudes, we can see that the predictions are not good. We can hardly find a good prediction in the list. Hence we can conclude that the variables related to community area are not directly proportional to the prediction for our target variable.

1. **Linear regression with the positively correlated variables**

- Actual vs Predicted value table:

|  |  |  |
| --- | --- | --- |
| **Row count** | **Actual** | **Predicted** |
| 0 | 2 | 1.997836 |
| 1 | 0 | 0.001962 |
| 2 | 9.2 | 9.190472 |
| 3 | 0 | 0.003075 |
| 4 | 0 | 0.001323 |
| ... | ... | ... |
| 1512804 | 1.5 | 1.499923 |
| 1512805 | 0 | 0.001518 |
| 1512806 | 14.6 | 14.590097 |
| 1512807 | 2 | 1.998851 |
| 1512808 | 3 | 2.997788 |
|  |  |  |
| 1512809 rows × 2 columns | | |

* Prediction-error graph:



Looking at the predicted values for the linear regression with all the positively correlated variables, we see that the predictions are very closely accurate. We can hence come to a conclusion that the positively correlated variables are directly proportional to the target variable. They are very closely related to the target variable (tips). An increase in the cost variables will definitely have an increase in the target variable and in the same way, a decrease in these variables will also show a decrease in the target variable.

1. **MODELLING AND PLOTTING THE GRAPHS**

**R-squared:** R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

**Key Limitations of R-squared**

R-squared cannot determine whether the coefficient estimates and predictions are biased.

R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!

**Are Low R-squared Values Inherently Bad?**

No! There are two major reasons why it can be just fine to have low R-squared values.

In some fields, it is entirely expected that the R-squared values will be low. For example, any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%. Humans are simply harder to predict than, say, physical processes (here, giving tips).

Furthermore, if your R-squared value is low but you have statistically significant predictors, you can still draw important conclusions about how changes in the predictor values are associated with changes in the response value. Regardless of the R-squared, the significant coefficients still represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant. Obviously, this type of information can be extremely valuable.

The linear regression plots are made for all the following combinations of variables:

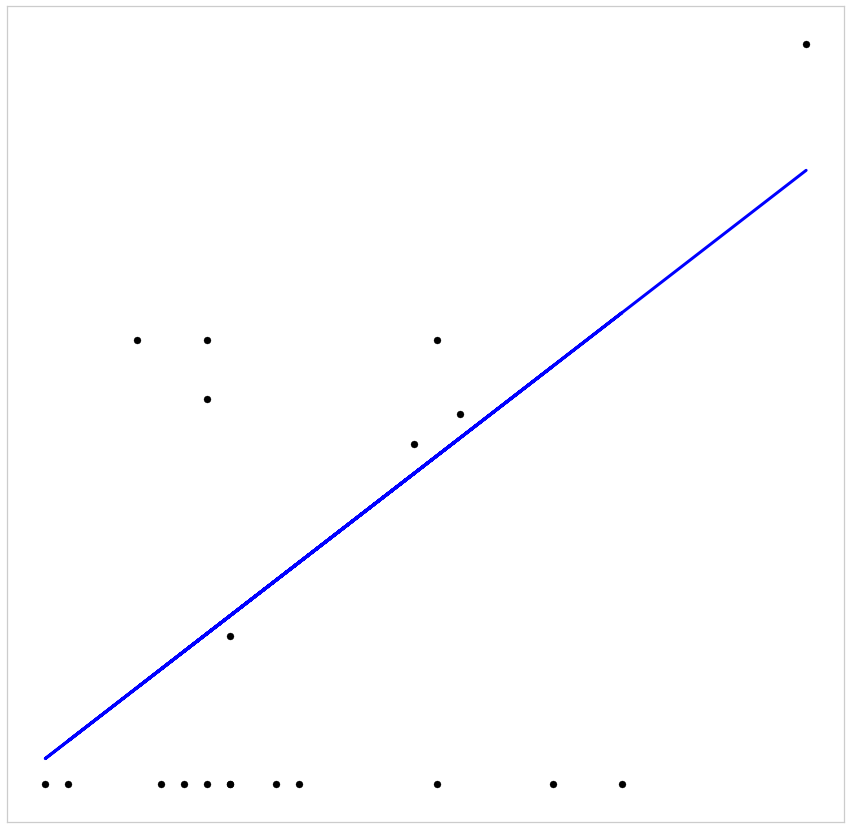
1. All the numerical variables
2. All the negatively correlated variables
3. Variables related to the cost of the trip
4. Distance and duration variables
5. Variables related to community area
6. Variables related to latitudes and longitudes
7. All positively correlated variables
8. **All the numerical variables**

Coefficients: [[0.00200759]]

Mean squared error: 2.28

Coefficient of determination(R-squared): 0.01

Linear Regression plot for the all the numerical variables

****

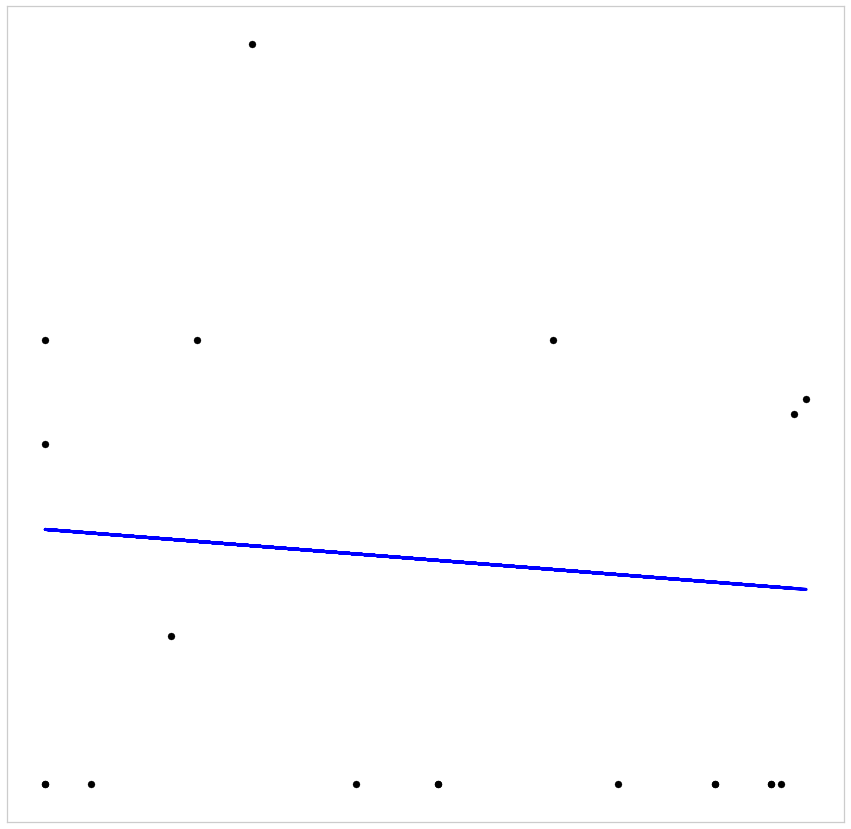
1. **All the negatively correlated variables**

Coefficients: [[-0.00053288]]

Mean squared error: 2.40

Coefficient of determination(R-squared): -0.04

Linear Regression plot for all the negatively correlated variables

****

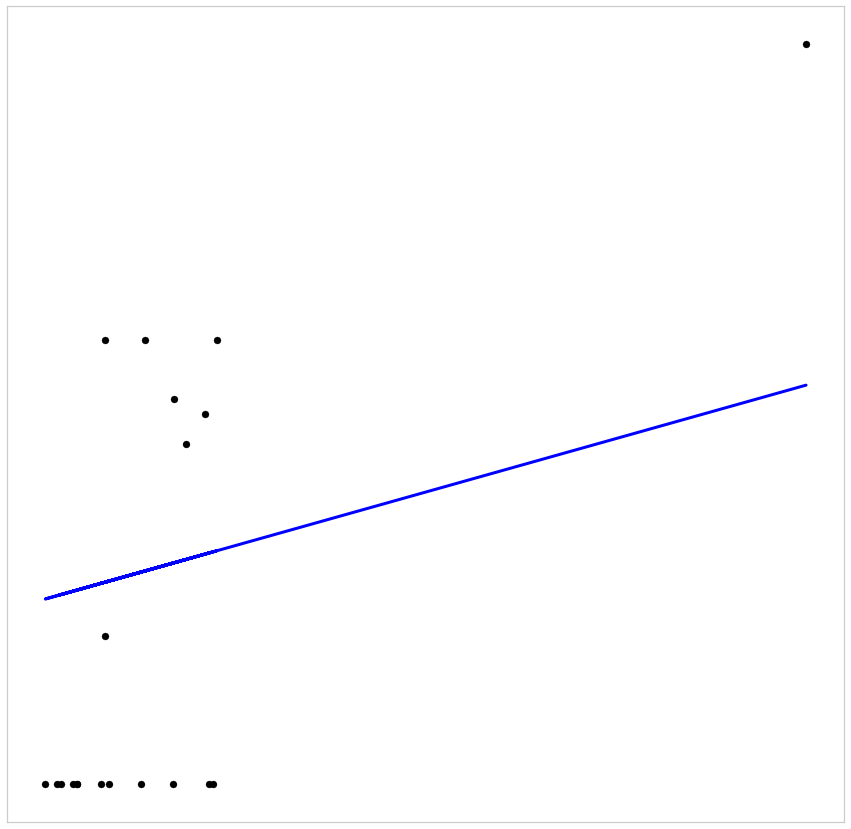
1. **Variables related to the cost of the trip**

Coefficients: [[0.03025912]]

Mean squared error: 1.91

Coefficient of determination (R-squared): 0.17

Linear Regression plot for the variables related to the cost of the trip

****

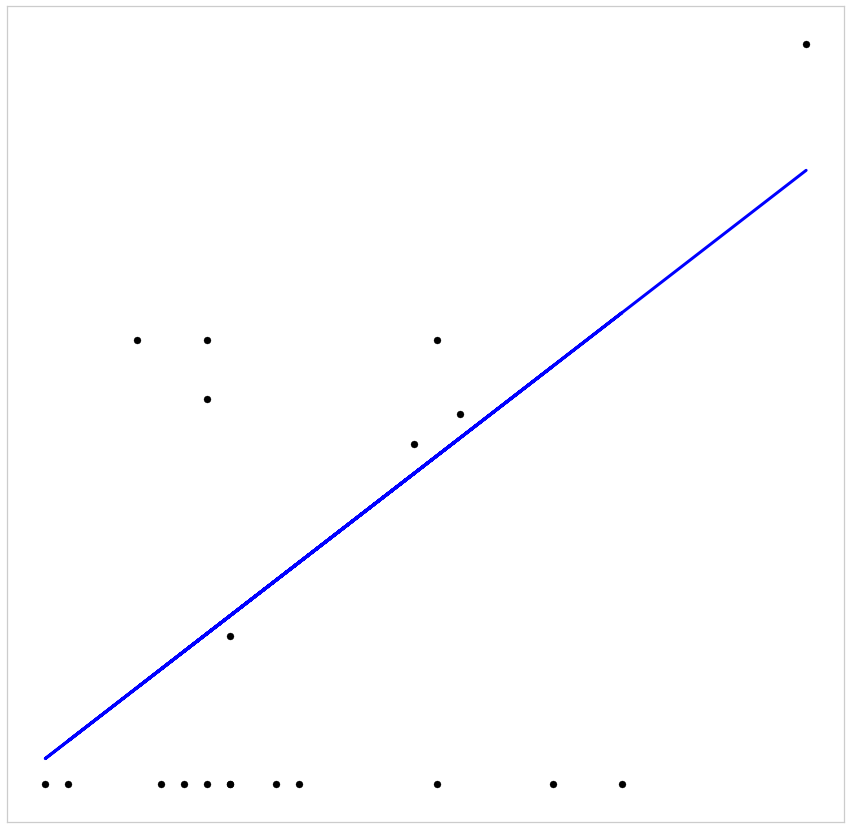
1. **Distance and duration variables**

Coefficients: [[0.00200759]]

Mean squared error: 2.28

Coefficient of determination (R-squared): 0.01

Linear Regression plot for the distance and duration variables

****

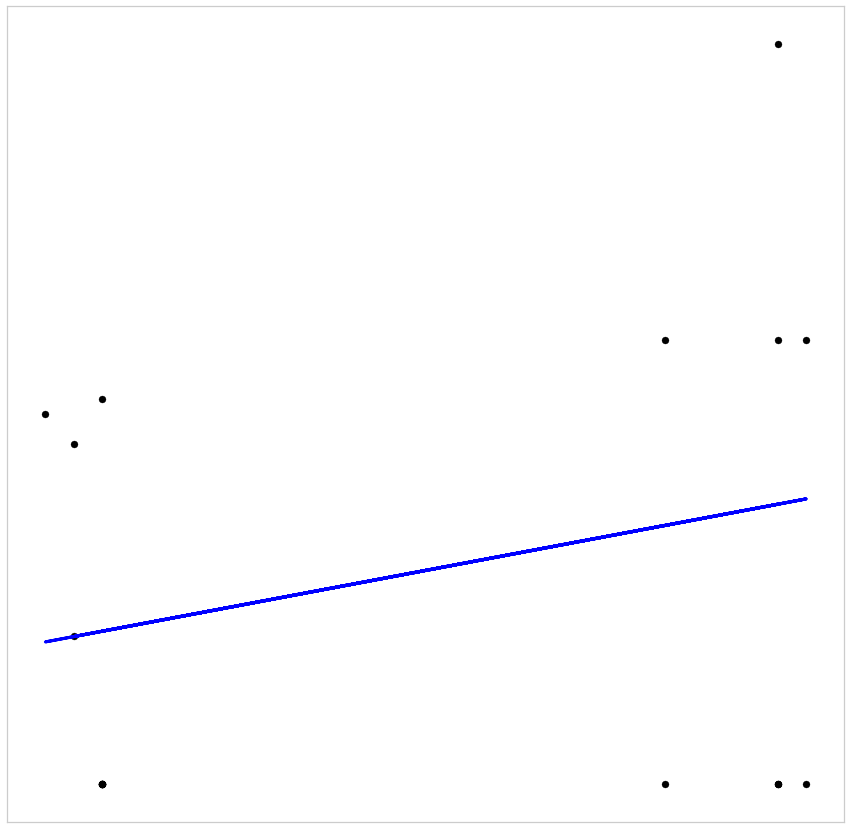
1. **Variables related to community area**

Coefficients: [[0.03578704]]

Mean squared error: 2.25

Coefficient of determination (R-squared): 0.02

Linear Regression plot for the variables related to community area

****

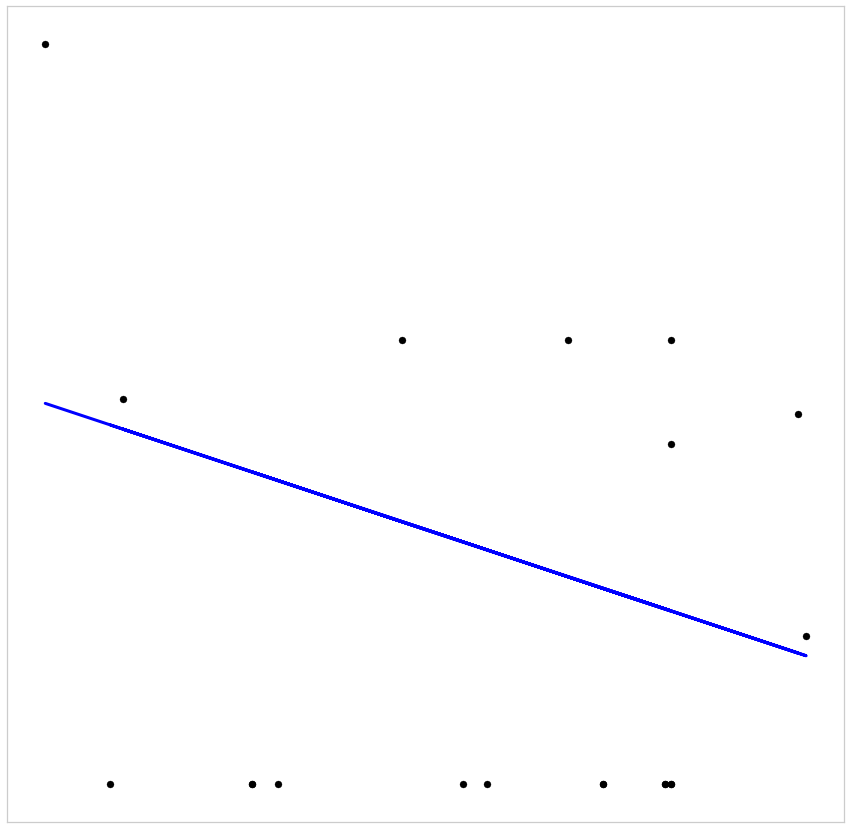
1. **Variables related to latitudes and longitudes**

Coefficients: [[-0.0023203]]

Mean squared error: 2.52

Coefficient of determination (R-squared): -0.09

Linear Regression plot for the variables related to latitudes and longitudes

****

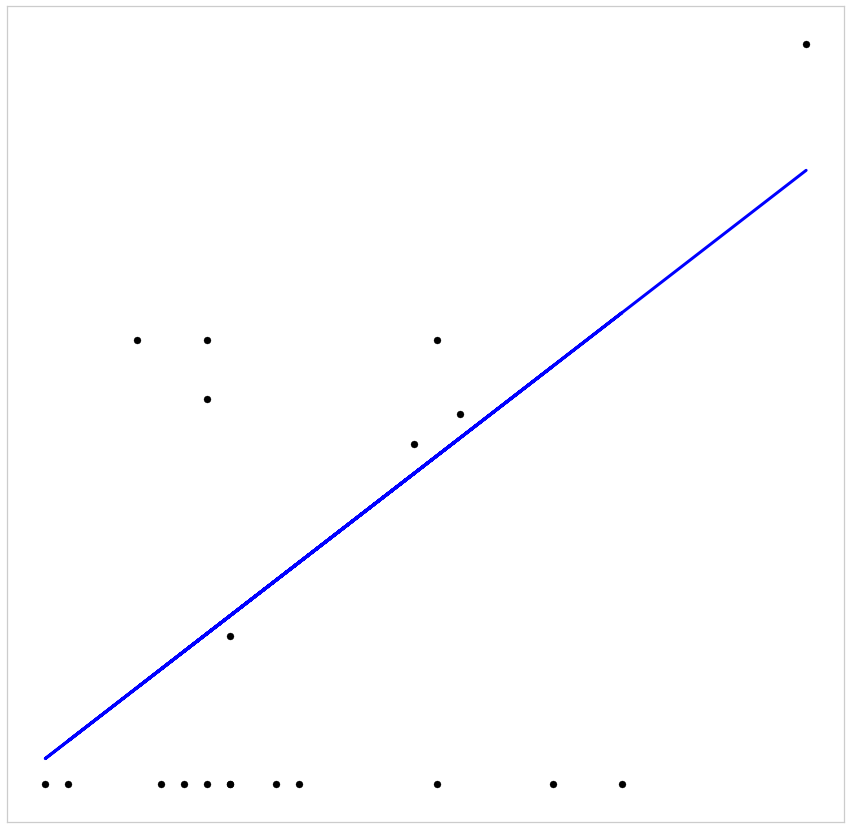
1. **All positively correlated variables**

Coefficients: [[0.00200759]]

Mean squared error: 2.28

Coefficient of determination (R-squared): 0.01

Linear Regression plot for the positively correlated variables



Even though the calculated R-squared values are low, keeping in mind the limitations of R-squared, we can still conclude that the changes in the predictor values are associated with changes in the response value.

1. **CONCLUSIONS**

Based on these analysis, the possibilities of predicting the tips for a taxi ride depend mainly on the below mentioned factors:

1. Tips can be predicted using the below combinations:

-fare, extras and trip\_total

-taxi\_id, trip\_seconds, trip\_miles, pickup\_community\_area, dropoff\_community\_area, fare, extras, trip\_total and company

1. While we have been able to improve the results, this may not be enough for the predictions as the customer reviews is not included. This is definitely a major factor in terms of considering whether to give a tip as it is a part of human behavior.
2. **RECOMMENDATIONS**
   1. Including the customer reviews for further analysis will be helpful to get more precise predictions as it is directly related to providing tips.
   2. To improve the accuracy in the prediction we can try using different algorithms as well