NYC Taxi Tip-Rate Prediction

Shubham Gupta (A53206235)
Rushil Nagda (A53222917)
Sudhanshu Bahety (A53209213)

***Abstract*— Tips given to Taxi Drivers form a substantial part of their income. They also serve as a loose metric of service quality that can be useful to Taxi companies. In this report, we analyze factors that affect tip ratio and train models to predict tips based on features. Our analysis leads us to the conclusion that the most important factors governing tips are geographical factors such as location of the beginning and ending of the trip and trip factors such as trip distance, fare and duration. Multiple baselines were created to judge the effectiveness of our models. We achieved our best results using Gradient Boosted Decision Trees as our prediction model.**

I. INTRODUCTION

New Yorkers take over 200 million Taxi rides per year. Of these, over half are Yellow Taxi rides. Yellow Taxis operate in the Manhattan and the boroughs, but generally make pickups in and around Manhattan. While tips to taxi drivers are usually interpreted as a gesture of gratitude for service, we analyze what other factors can have an impact on tips. For example, do Taxi rides starting or ending in richer parts of the city result in better tips?

We present an analysis of factors affecting tips and try to use this factors to construct models to predict tips.

II. EXPLORATORY ANALYSIS

A. Dataset Description

We used the NYC Yellow Taxi dataset from Kaggle[1]. This dataset consists of Taxi ride information from Jan-June 2016. For our analysis, we have limited ourselves to just January data. This amounts to about 10 million data points. Each data point in the set represents one completed Taxi journey. The dataset consists of a variety of features, sumarized below:

Geographical Features:

- Location (latitude/longitude) of pickup
- Location (latitude/longitude) of dropoff

Temporal Features:

- Date/time of pickup
- Date/time of dropoff

Trip Features

- Trip distance in Miles
- Passenger count

Payment Features

- Fare amount in USD
- Taxes/toll amount in USD
- Tip given in USD
- Payment type (Cash or Credit)

In the following sections we have analyzed these features to get an idea of what features work best when trying to predict the ratio of tip given.

B. Cleaning the Data

The dataset is noisy and has many missing values in all features. We have cleaned the dataset by:

- Removing the data points which have 0 as longitude or latitude for pickup/dropoff
- Removing data points which have zero or negative total bill amount.
- Removing data points which have negative tip
- We have identified the range of latitude and longitude in which over 99% of the datapoints lie. The rest have been ignored as they are outliers, or worse, corrupt.
- We have removed all the transactions which were not electronic, as they all had zero tip. One possible explanation would the that such payments are not recorded properly or perhaps the driver refused to report it to save on taxes.
- We have removed all points where tip was more than 100% of the fare amount as those are outliers and have a significant impact on algorithms which try to optimize MSE.

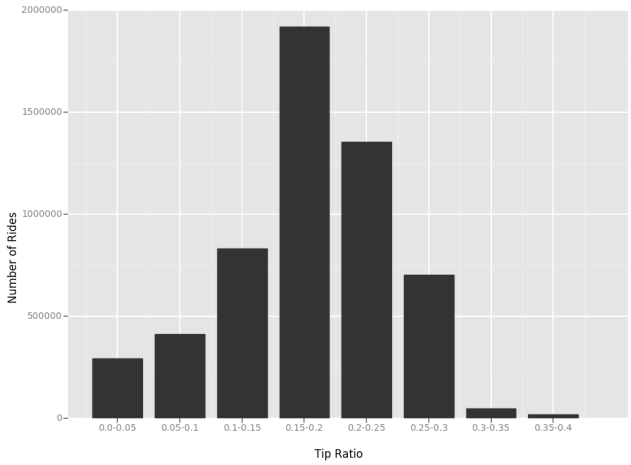
After cleaning the data, we are left with about 7 million data points. The following section presents visualizations and analysis of the cleaned dataset.

C. Analysis

Figure 1 is a Histogram of tips across all rides. It can be seen that Tips usually lie between 10-30%. The Mean Value of all tips was about 18.73% and the standard deviation was about 5.8%

Figures 2 and 3 show the average tip ratio based on Pickup and Dropoff Locations respectively. Note that the Dropoff graph is more widespread because Yellow Taxis tend to roam areas such as Manhattan and the airport for pickups while dropoffs could be spread through the city and outer boroughs. In both the plots there is a noticeable dependence of tip ratio on location. This can be attributed to socioeconomic factors. For example, the are north-west of central part is inhabited by mostly students and working class people. These demographics have lower disposable income and hence a lower tendency to tip. Note that to show contrast

Fig. 1: How New Yorkers Tip. Most commutes result in a tip of 15-25%. Almost nobody tips over 40%.



we have eliminated the top 10 percentile and bottom 20 percentiles of tips from these plots.

Fig. 2: Tip Ratio based on Pickup Location. Here we can see that passengers taking rides from Manhattan area, airport tend to tip higher compared to other parts of the city which may be due to high income people living in those parts of the city.

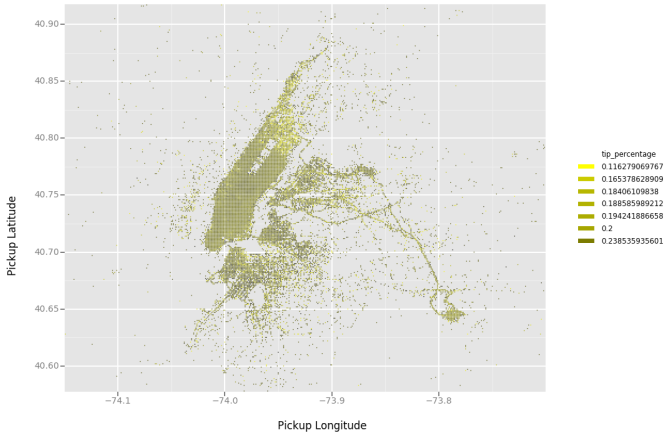


Figure 4 shows the variation in tip ratio based on the time of the day. It can be seen that people tend to tip the least between 3am thorough 5am and tip the most around noon. Figure 5 shows the variation in total payment based on the time of the day. The total Fare (includeing taxes and tolls) tends to spike between 3-6am. This could be because of people taking longer rides when they leave the city after a night out and because of night time charges by cab companies. Either way, this increase in fares could explain the decrease in tipping tendency.

Figure 7 is a plot of the tip percentage and total payment of all our data points. Some interesting patterns can be noticed in this graph. Tips tend to form a Gaussian around about

Fig. 3: Tip Ratio based on Dropoff Location. The pattern over here are similar to that of Fig 2. The major difference that we can see is that dropoff locations are more widespread as compared to pickup locations, which is explained in detail in the Analysis Section.

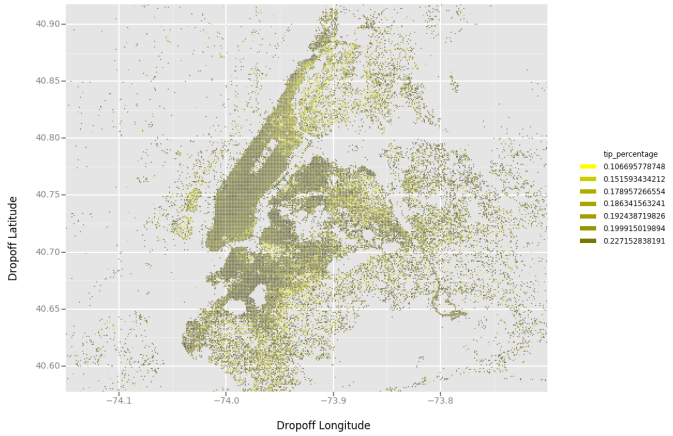
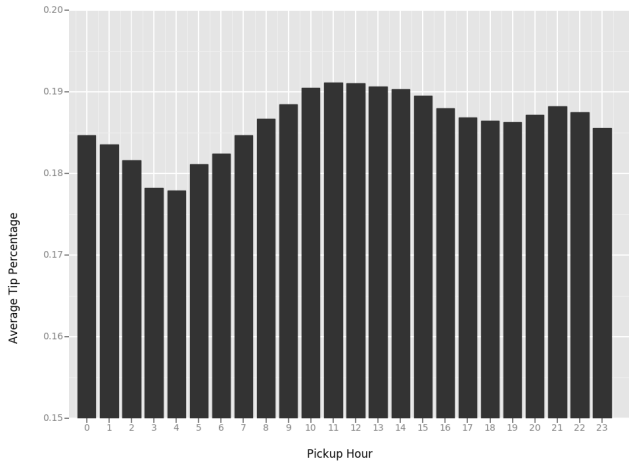


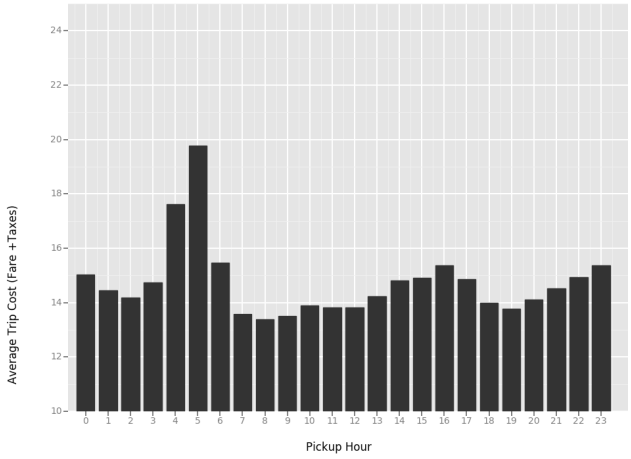
Fig. 4: Tip Ratio based on Hour of Day. The variations in the tip ratio at different hours of day can be explained by looking at fare amount in Fig5. Increase in fare amount results in decrease in tip ratio.



17% for lower amounts. The horizontal lines in the graph correspond to the preset tip percentages (20, 25 and 30) that a passenger sees when they swipe their credit card. The asymptotic curves correspond to common tip values (such as $10 or $20).

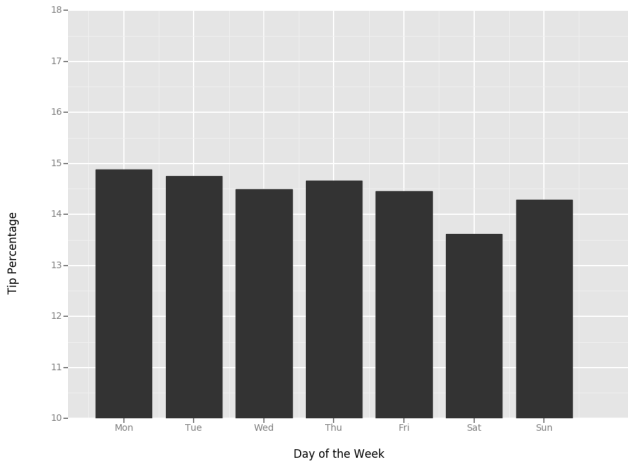
Figure 8 shows the relationship between trip distance and tip ratio. There is a marked decreasing tendency in tip ratio as the trip distance rises. This can be attributed to the fact that the fare amount rises when an increase in distance and people tend to tip less for larger fares. Another explanation is the fact that when people tip common values such as $10 or $20, this shows as a larger tip percentage for shorter trips. Figure 9 shows the influence on the tip percentage of the

Fig. 5: Total Payment based on Hour of Day. The fare amount usually hikes during 3 - 6 in the night, which results in drop of tip ratio (See Fig 4). Similar patterns are seen for different hours of day.



average speed of the trip. Contrary to what one might expect, faster trips do not necessarily result in better tips.

Fig. 6: Total Payment based on Hour of Day. The fare amount usually hikes during 3 - 6 in the night, which results in drop of tip ratio (See Fig 4). Similar patterns are seen for different hours of day.



III. Predictive Task

There are multiple types of predictions that are possible on this dataset. One such task could be to predict the dropoff location based on pickup location and other factors. Another task could be to predict temporal demand. But the one we find most interesting is to predict the tip (or rather ratio of tip to total fare) given all other features.

Fig. 7: Tip Percentage vs Total Paymnent (Not including tip). We can see three fine lines at 20%, 25%, 30%. We can also see different asymptotic curves in the graph. Each curve corresponds to a fixed tip amount.

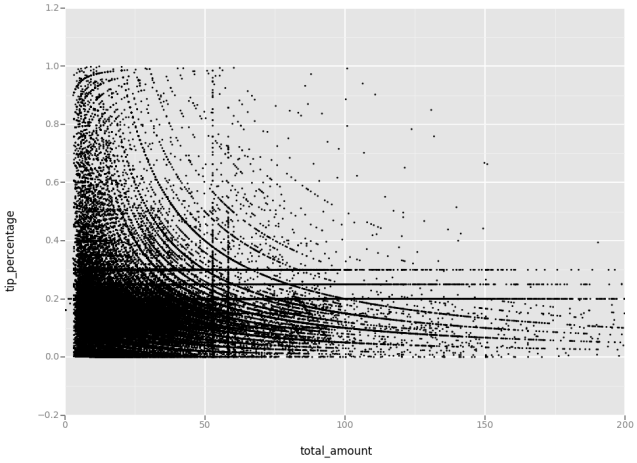
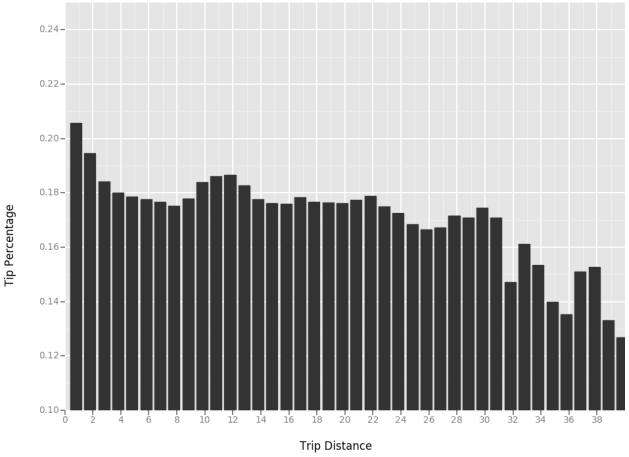


Fig. 8: Tip Percentage based on trip distance. We can see that as the trip distance increases tip ratio decreases which could be accounted due to the fact of increase in fare amount.



A. Evaluation

We predict tip ratio, which is defined as:

$$\text{tip ratio} = \frac{\text{tip amount}}{\text{total fare}} \tag{1}$$

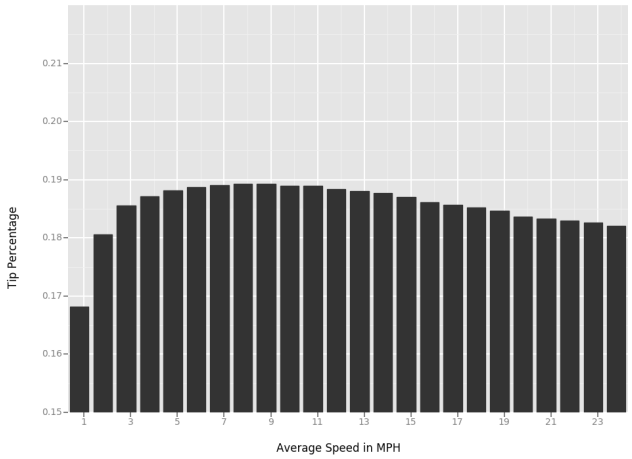
where total fare includes fare amount, taxes and tolls.

We evaluate models based on their *RMSE* which is defined as:

$$RMSE = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{2}$$

where $\hat{y}_i$ is the predicted tip ratio by our model and y_i is the actual tip ratio.

Fig. 9: Tip Percentage based on speed of the trip. The tip ratio increases with increase in speed until 11 MPH and then decreases. Shorter tips are usually within the city and from Fig3 we can see that tips near Manhattan are more. Moreover, longer trips outside the city accounts for high average speed and due to more fare amount, tip ratio are usually lower.



IV. Existing Literature

Analysts at Bloomberg [2] have used this dataset to identify tipping patterns. They have reached to similar conclusions as us even though they used the dataset which was older compared to ours. They note that riders like to tip the presets (20%, 25% etc) and a lot of riders like to tip simple 0%. They also note that people tip more in the evening and less during the night, which is similar to what we have observed.

Jos M. Camacho [3] has used the same data from a different year to try and predict the tip percentage of a rider. Although, he has done this prediction in a different way, by converting this to a classification task by manually dividing tips into buckets of 5% each.

Github user @sdaulton [4] has analyzed this data to see the most common locations for pickups and dropoff. He shows that central Manhattan is one of the location for pickups. He also shows that pickup and dropoff location are the most important features, as we have also argued.

Max Woolf [5] has used ggplot to see the total revenue based on location of pickups. This blog inspired us to use ggplot as out plotting tool in python as it looks far cleaner than matplotlib.

Yunrou Gong et. al. [6] has given a good visualization of taxi demand by using 3 different datasets - NYC Yellow taxi, Green taxi and Uber dataset. The usual trend shown is that yellow taxis are much more in demand than the other competitors. Their exploratory analysis revealed that demand greatly depended on zipcode, hour of day, day of week, seasonal changes(mainly during rainy season). For modelling their data, they used Linear Regression, Ridge Regression, XGBoost, Random Forest and Multiple Regression. XGBoost outperformed all other ML Models to predict the taxi demand.

Jay Gopalakrishnan [7] worked on dataset consisting of trip details of 1.237 billion rides. Their analysis revealed that pickups are concentrated around a few locations like Manhattan and airports. They also analyzed that peak times and found that 6 10PM and 8 10 AM were rush hours, and the demand was low 4 5PM. A possible explanation they came around was that around 5PM is the shift change occurs where cab drivers are heading back to the garage. They also note that there has been a significant drop in taxi usage from 2012 to 2016, mostly due to rise of Uber, lyft, etc.

Ramkumar Chandrasekaran, Microsoft [8] used MicrosoftML to predict New York Taxi Tips as a binary classification of predicting whether or not a tip was paid for a trip. If the tip amount was > 0, then it was labelled as 1 else 0. He didn't do any feature engineering but used various different models like Logistic Regression, Fast Linear model (SDCA), Fast Tree, Fast Forest, Neural Network. In order to compare the model, he plotted ROC and calculated area under ROC. In his case, Fast Forest performed best as compared to others.

Aiko Liu [9] investigates the tipping behavior of passengers as a measure of their satisfaction on two different datasets - Yellow Taxi and Green Taxi. He observes that tipping rate changes with hours and weekdays which is non-trivial. He notes that people are more likely to tip higher during rush hours, as expected. One of his finding that coincides with ours is that tipping is negatively correlated with the average speed of the ride. This is counter intuitive as one would expect that a faster ride is more satisfactory and one would tip higher.

V. Models Used

In order to better understand our models performance, we created a few simple baselines against which we evaluated our models. We have created three baseline predictors, as described below.

A. Baselines

1) Mean: The first baseline is to just predict the average tip ratio of training set, for every data point in test set. RMSE of 0.05807 was obtained with this simple baseline

2) Mean based on pickup location: From our training set, we compute the average tip ratio for each pickup location. For each point in test set, we predict the tip ratio as the average tip ratio of the pickup location. RMSE of 0.05793 was obtained for this baseline, which is an improvement on the previous baseline.

3) Mean based on dropoff location: Here we do apply a very similar approach as before, only changing pickup location by dropoff location. Here we obtain RMSE of 0.05787 which is very similar to that of previous baseline.

Now, we would like the rest of our models to outperform all three of our baselines.

B. Feature Engineering

A few of the features that we have used as is:

- Distance traveled during the ride
- Total fare of the trip
- Latitude/Longitude of pickup location
- Latitude/Longitude of dropoff location

Features that have been engineered:

- Day of the week
- One-hot encoding of pickup hour
- One-hot encoding of pickup location
- One-hot encoding of dropoff location

Before we did one hot encoding of location (latitude longitude pairs), we rounded the latitude and longitude to nearest two decimal place to decrease granularity. This results in the formation of a grid of about 4000 boxes in New York. We encode each pickup and dropoff location based on which box that location lies in.

Features that were tried but removed as including them increased the propensity of our models to overfit:

- Number of passengers
- One-hot encoding of number of passengers
- Date of the month

C. Scalability issues

As the number of training points are over 5 million, we cannot use one hot encoding of locations in linear regression and random forests. But xgboost, which is library for gradient boosted trees uses sparse representation of dataset, hence we have tried GBDT with one hot encoding.

D. Linear Regression

This is first approach we tried It is reasonable to assume that the tip ratio is linearly dependent of some of its features, like trip distance.

Linear regression performed better than we would expected with RMSE of 0.05674 which is a clear improvement over baselines.

E. Random Forest Regression

RFR uses an ensemble of decision tree for regression. They have proved very effective in regression problems previously and are known to overfit less than many models if correctly used. We have set the parameters by grid search. We notice that with good regularization and other parameters like max_leaf_nodes, it outperforms linear regression.

RMSE of 0.05654 was obtained which is again an improvement.

F. Gradient Boosted Decision Trees

XGBoost was used for this task as it has been known to win many kaggle competitions with similar prediction tasks. Multiple combination of features were used here with multiple variations in parameters, and the best RMSE obtained was 0.05232. This beats both RFR and LR with a significant margin.

VI. Results

We divided our dataset into Training and Test sets in an 80:20 ratio. This resulted in about 5.5 million training points and 1.5 million test points. Performance of our models and the baseline models on the test set are presented in Table I.

TABLE I: Summary of Results

Model	RMSE
Baseline: Mean	0.05807
Baseline: Mean based on pickup location	0.05793
Baseline: Mean based on dropoff location	0.05787
Linear Regression	0.05674
Random Forest Regression	0.05654
Gradient Boosted Decision Trees	0.05232

Fig. 10: Importance of different Features calculated by our GBDT Model

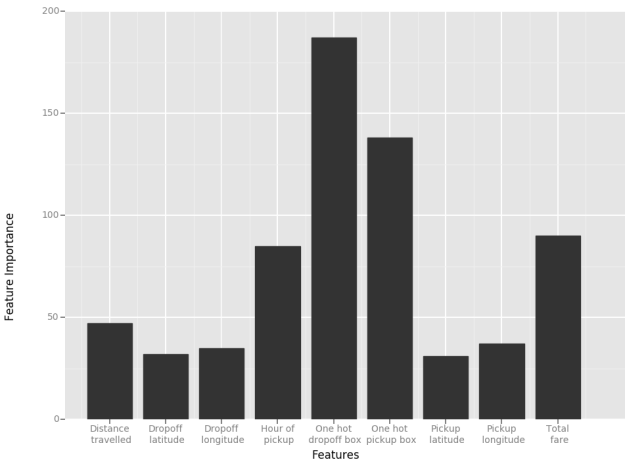


Figure 10 shows the importance of the more important features calculated by the GBDT Model. Location buckets are particularly important, more so than latitude and longitude represented as real values.

We consider it to be a considerable win as any increase in the tip received by the driver goes a long way as this is their primary source of income.

As noted above, GBDT outperforms other models as it gives special attention to data points which are hard to predict for and their tip ratio has been poorly predicted in the past. This kind of mechanism is lacking in other techniques and hence they can't adapt with time giving more attention to points which perform poorly.

We can further decrease our RMSE if we are presented with information of the rider and driver for each trip. Tipping habits of riders and quality of service of drivers matter towards tipping at least as much as pickup locations, etc.

A. Conclusion

We conclude that viewing tips as solely a measure of service quality can be misleading. A large number of factors such as location, trip distance, trip duration and fare have a

significant effect on tip. There are also some temporal effects on tip, such as the time of day and day of week.

REFERENCES

[1] NYC Yellow Taxi Trips, January-June 2016
https://www.kaggle.com/nyctaxi/yellow-taxis

[2] Here's How Much You Should Be Tipping Your Cab Driver
https://www.bloomberg.com/news/articles/2014-07-31/heres-how-much-you-should-be-tipping-your-cab-driver

[3] nyc-taxi-tip-predictor
https://github.com/josemazo/nyc-taxi-tip-predictor

[4] http://sdaulton.github.io/TaxiPrediction/

[5] How to Visualize New York City Using Taxi Location Data and ggplot2
http://minimaxir.com/2015/11/nyc-ggplot2-howto/

[6] Predict New York City Taxi Demand
http://blog.nycdatascience.com/student-works/predict-new-york-city-taxi-demand/

[7] Analyzing 1.2 Billion NYC Taxi Rides
https://medium.com/@gopalaj61/analyzing-1-2-billion-nyc-taxi-rides-83ea8012827e#.2yv0bch80

[8] Predicting NYC Taxi Tips using MicrosoftML
https://blogs.msdn.microsoft.com/microsoftrservertigerteam/2017/01/17/predicting-nyc-taxi-tips-using-microsoftml/

[9] NYC Taxi Riders' Tipping Behavior Analysis
http://www.datasciencecentral.com/profiles/blogs/nyc-taxi-riders-tipping-behavior-analysis