Impact of traffic congestion on the tip paid to NYC taxis

Nikhil Kishore, github.com/nk1877, N11258217

Abstract:

This Project tries to understand the trend in the tip amount to quantify customer satisfaction. Various techniques are applied to understand the trend and factors which affect tip. It will try to understand the effect of time and distance on tip percent and tip amount. Liner and polynomial model is used to predict the amount of tip cab drivers will receive depending on factors like duration, distance, day and night time.

Introduction:

"If you tip a taxi driver less than 12 percent of the amount of your fare you will be shortchanging him, in the opinion of Edward Corsi, State Industrial Commissioner" reported in year 1947.Although recently average tip amount paid to cabbies is 15.5%. Tip has become part of culture and good gesture, but could it be used as a measure of goodness? Is tip just part of tradition with no relation to customer satisfaction or is it something more. Paper emphasis on comparing tip against time and distance to understand if customer chooses to pay less if he reaches his destination early or tip is unaffected of quality of service.

he or she his or hers unaffected by the quality

Data:

In order to understand the problem better NYC taxi data for the month of July 2015 is used for the analysis. Analysis required cabs going in one direction between two different destinations. It was imperative to find popular pickup and drop off point for the accurate research and higher number of data points. Therefore, JFK airport is chosen as the pickup point. JFK is chosen as the initial pickup point because of its popularity, next intensity heat map on CartoDP is used to visualize all the destinations where cabs go (Figure A). It could be seen down town Manhattan and some parts of Brooklyn has the highest intensity. Therefore, Manhattan downtown is chosen as the drop off destination for the analysis. Next tip percentages are calculated and added to the data along with time taken for each ride.

Initially data contained approximately 20 million taxi rides which were reduced to 78186 rides after munging and removing all the outliers such as data points where journey distance was more than 40 miles, time duration more than 5 hours and tip more than 100 percent. Since distance between JFK and Manhattan downtown is less than 20 miles and even with the congestion it would take less than 5 hours to reach the drop off point.

all of these choice make sense to simplify the analysis, but they also introduce biases, which should be mentioned

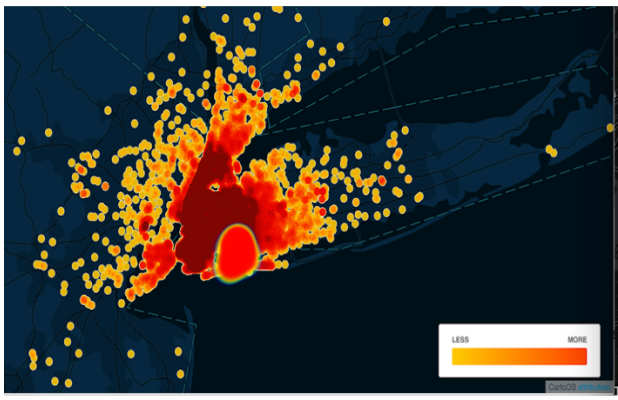

Figure. 1 A

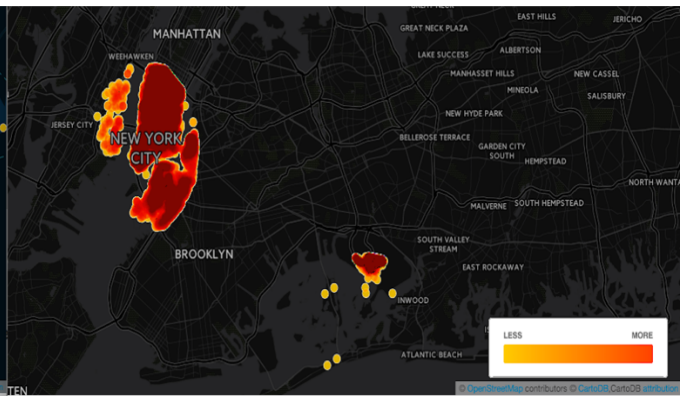

Figure. 1 B

Methodology:

Various tools were used for the analysis of the project. Initially the analysis was started by performing linear fit model between tip percent and trip distance [figure 1]. It was observed that for majority of cases, distance covered to reach the destination was almost similar therefore, observing zero R square value was not shocking. To understand the model further similar model was applied on tip percent versus time. P value observed in ordinary linear square summary displayed some significant dependence of time on tip percent.

Next to investigate further multivariate regression is preformed with time and distance as repressors for change in tip percent [figure 2]. There was a positive change in predicted tip percent versus actual tip percent. R square for the analysis was 0.12 percent with coefficient for time -0.0695 and coefficient for distance 0.0634. Therefore, multivariate analysis suggests that increase in tip is affected by decrease in time and increase in distance.

there is definitely something wrong with this plot: the dependent and independent variables are sswitched for duration vs tip i think

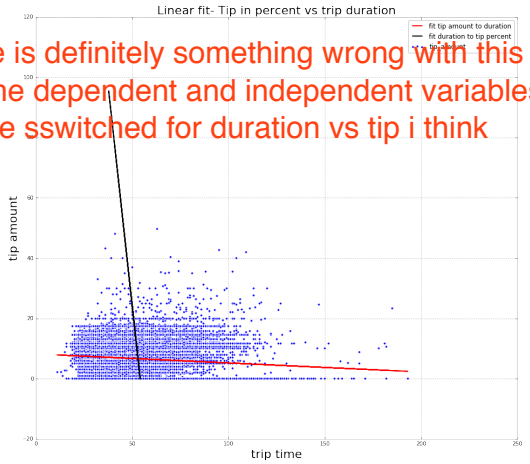

Figure. 2

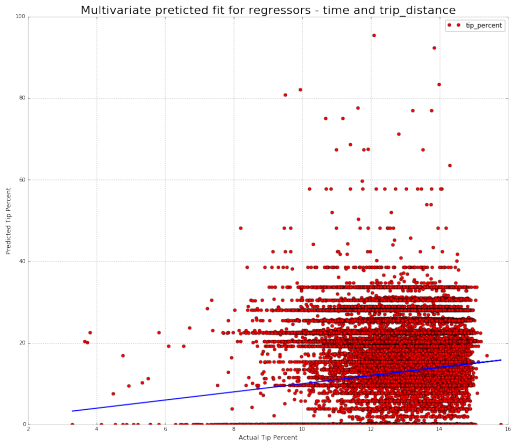

Figure. 3

Next tip pattern of day versus night is observed to find any visible difference in tip. R square for multivariate regression on tip percent versus time and distance during the day was low with value of 0.04 but it was still significant with p value 0 for time and 0.001 for distance. Therefore, it could be suggested that time and distance are influencing factors for tip during the day.

Similarly, for night analysis it was observed that R square is slightly higher than day with value of 0.06. P value for time is significant with value 0 and beta coefficient -0.0653, although p value for distance is not significant during night time with value higher that 5 percent [Appendix A, Table 1].

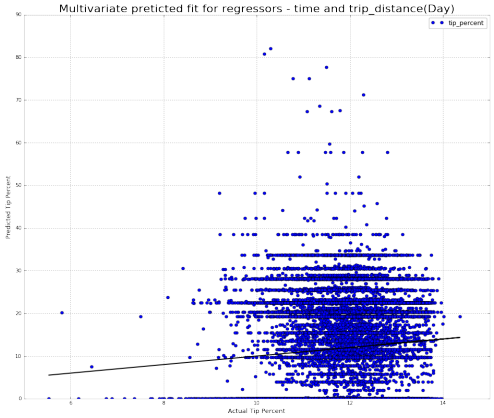

Figure. 4

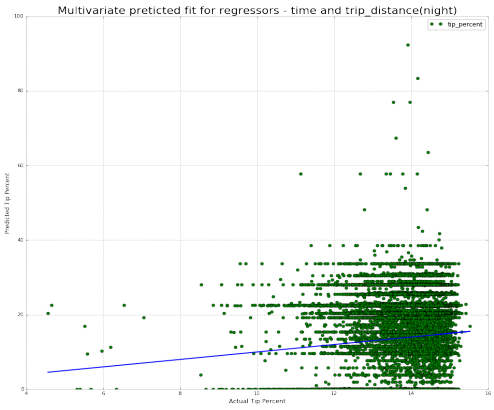

Figure 5

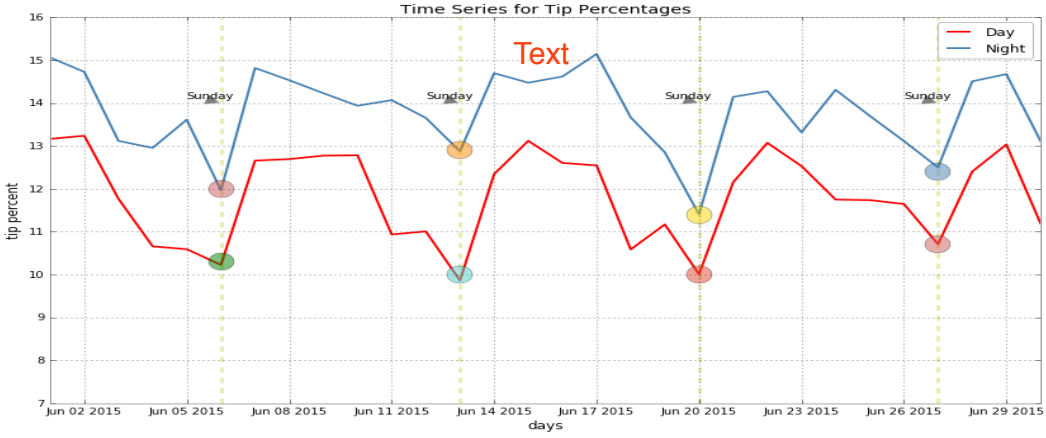

Figure 6

LABELS TOO SMALL EVERYWHERE!!

Time series for day and night is used to compare changes in tip percent [Figure 6]. Similar time series is plotted to see changes in tip amount. Both shows similar trend with drastic low on tip percent and tip amount on Sundays [Figure 6].

that is very very interesting. it would be interesting to see if it holds for other-origin destination pairs

Conclusions:

Project revealed some interesting facts about tipping pattern in taxis of New York City. Although analysis covered only two different points in New York City, lot of facts were uncovered. Mainly following three conclusions could be deducted from the analysis

1. During the linear and multivariate analysis, it was witnessed that time affects the tip percent therefore it could be suggested that delay in dropping the customer at its destination affects the tip percentage negatively. However, it is hard to hypnotize the reason for delay. Multivariate regression shows 1.2% of R square, which is low but impact of time and distance combined is still significant. Time again is showing negative coefficient but distance is showing positive coefficient. Therefore, it could be suggested that for a unit change in time tip percent would decrease by 0.06 and for unit change in distance tip would increase by 0.0634 units.
2. After segregating the data between day and night it was witnessed that time series trend of tip percentages behaves similarly for both day and night [Figure 6]. At every significant drop of tip percentage during the day there was drop on tip percentage at night too. Result was similar with major spikes as well. Therefore, prediction of customer's behavior about paying tip during the night could be predicted by their tip paying behavior at day and vise-versa.
3. The Tip percentage amount was lowest on Sundays for all the week throughout the month in both cases i.e. days and night [figure 6]. Perhaps lower tip amount on Sunday's exist because people travelling on holidays are not on higher budget limit.

In terms of quantifying happiness of customer's taxi trip via their tip is very hard. Although with average of 12.84 percent of tip and median of 15.90 it could be assumed mostly New Yorkers prefer paying only the customary tip where possible. Tip payment changes with predictable frequency for example we can easily say next Sunday will be a bad day in terms of tips for cab drivers when compared to weekdays.

Future work:

Project sets root for lot of work which could be done to its extension. It would be fascinating to cover the following points for the future.

- Cover wider range of data to find trend and similarity for overall tipping behavior as it could be translated into many applications like improving customer service, understanding the factors that influences human tipping behavior.
- Fit time series of tip amount with equation y=|cosx| to test for improved R square. Visually time series gives the y=|cosx| trend therefore it would be interesting to see the result. Also if plot fits perfectly then prediction of tip amount for particular day and time will be easier.

cos may be not the best model cause the time behavior is not smooth

- Develop tipping application to analyze tipping pattern of an area to help tourists and travelers understand the range of tipping amount in an area which would translate to the quality of service.

Link to I-python notebook

https://github.com/nk1877/PUI2015_nkishore/blob/master/Extra%20Credit/PUI%20Extra%20Credit.ipynb

Appendix A

Table 1

Data set	*Test Type*	*Versus*	*R square*	*Beta coefficient*	*P value*	*Confidence interval*
Complete Data	Linear	Tip Percent Versus Time	0.008	-0.0296	0.000	-0.032 -0.027
Complete Data	Polynomial	Tip Percent Versus Time	0.012	-0.0437	0.000	-0.053 -0.034
Complete Data	Multivariate	Tip percent versus distance and time	0.012	Time= -0.0694 Distance= 0.0634	Time= 0.000 Distance= 0.000	Time= -0.074 to -0.065 Distance= 0.034 to 0.093
Day Time Data	Multivariate	Tip percent versus distance and time	0.004	Time= -0.0502 Distance= 0.0672	Time= 0.000 Distance= 0.001	Time= -0.058 to -0.043 Distance= 0.027 to 0.108
Night Day Nime	Multivariate	Tip percent versus distance and time	0.006	Time= -0.0653 Distance= 0.0226	Time= 0.000 Distance= 0.366	Time= -0.075 to -0.056 Distance= -0.026 to 0.071

Bibliography

- "Consumer Tipping: A Cross-Country Study", Michael Lynn, George M. Zinkhan and Judy Harris,*Journal of Consumer Research*,Vol. 20, No. 3 (Dec., 1993), pp. 478-488

- "How Much Do You Tip Cabbies?", EMILY S. RUEB, New York Times(January,2013), available at http://cityroom.blogs.nytimes.com/2013/01/02/how-much-do-you-tip-cabbies/

- "The impact of tipping recommendations on tip levels" David B. Strohmetz, available at http://www.sciencedirect.com/science/article/pii/S0010880401810260