To Tip or not to Tip? Predicting whether a passenger tips at the end of a taxi ride using a linear regression model.*

Anees Shaikh

12/19/2020

Abstract

The purpose of this analysis is to predict the amount tipped in a taxi ride based on a variety of quantitative factors. Some of these factors include the length of the trip, time of day, number of passengers, and the inclusion of other surcharges. To conduct this analysis, we are using Trip Data collected by the New York City Taxi & Limousine comission(TLC) for the month of May 2019. We employ a linear regression model that is trained on this data and then predicted on a test-set that was generated from the same month.
Keywords: Logistic Regression; Determinants of tipping; Yellow Taxi Data; Taxi Rider behaviour analysis; social norms; human mobility

Introduction

Tipping is an integral part of multiple industries. Any industry with a service aspect to it has some instantiation of tipping. This etiquette of adding more money on top of the bill goes a few centuries back, with a few possible likely origins. Tipping was likely to have began as a practice during the Tudor period(Eeckhout, 2015). In this period, there was an expectation that visitors to private homes would payout a sum of money to the domestic servants at the end of their visit(Eeckhout, 2015). This sum of money would eventually surpass the wages of the servants, making the visitors a more important source of income than the employer(Eeckhout, 2015). The prospect of a large tip would be part of whether a servant would want to work for someone, essentially becoming table stakes(Eeckhout, 2015). Soon after, this practice was temporarily abolished and servants would no longer receive any tips(Eeckhout, 2015). As you can imagine, it did not stay abolished for too long, returning as a practice in restaurants and hotels(Eeckhout, 2015). What's even more interesting is how this practice finds itself most prominent in the United States, despite being initially opposed to it(Mentzer, 2013). The practice would find itself in the US as it was used by the European elite as a display of their wealth(Mentzer, 2013). Soon enough, it became an established cultural norm, similar to form in which we experience it today.

The economic presence of tipping is rather significant. In 2016, cumulative tips were estimated to amount to $36.4 billion in service industries(Lee and Sohn, 2020). While that number is for service industries at large, tips accounted for 18% of annual taxi revenue, which amounts to $445 million(Lee and Sohn,2020).In terms of economic implications, it plays a significant role in the determining of wages. For instance, in Ontario, the minimum wage as of October 2020 is $14.25/hr, and the liquor servers minimum wage is $12.45/hr(Government of Ontario, 2020). The servers wage is even lower than the student minimum wage, which is now $13.40/hr. On their site, the government on Ontario states this wage applies to individuals who serve liquor *and* regularly receive tips or other gratuities. The discrepancy in the wages must in part be explained by the compensation received via tips. It thus becomes quite important to understand the factors that lead to an employee receiving a tip. Understanding these factors is an important aspect of behavioural

*Code available at :https://github.com/aneesshake/tipPrediction economics. In this analysis, we are presented with an opportunity to examine quantitative measures of a taxi ride and their impact on whether a patron is going to tip. This is not the first attempt at such an analysis, Lee and Sohn utilized a cumulative logit model to examine the effects of inclement weather and taxi tipping(Lee and Sohn, 2020). This analysis is another stab in that general direction that draws inspiration from previous research. Anecdotally speaking, examining this effect from a purely quantitative view does not feel complete. I will go into more details about this in the weakness section, but from experience of having known close friends and family that worked jobs where tipping was present, there is more to the equation than just quantitative factors of a ride.

In this paper, we will first begin with an overview of the data, there will be a little discussion about the specifics of the data. In the Model section, I will be providing detail as to what model was selected and why. There will be mention of model diagnostics in addition to decisions made regarding feature selection. In the Results section, I will be presenting the outcomes of this analysis ranging from descriptive statistics that help build context and better understand the problem, to model results and how well I was able to answer the titular question. In the discussion section, I will be discussing the implications of my results and tie it together with prior research. In doing so, I hope to add to the growing body of research that surrounds tipping and the dynamically changing world of transportation. Finally, I will close off the weaknesses of my analysis as well as propose a few next steps.

Data

The data for this analysis is a trip record dataset that was provided by the New York City Taxi and Limousine Commission((NYC Taxi and Limousine Comission, 2019). This is observational data. On their site, they have data going as far back as 2009. This data was made available to the public as a result of a FOIL(Freedom of Information Law) request that was submitted by data researcher, Chris Whong. Now, all this data can be accessed at https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. The data can be accessed in two ways. The first way involves downloading the data from the aforementioned link. The second way is to utilize the NYC Open Data platform. This platform is a free to use tool that allows individuals to download data with a few filters that can be done up front. For instance, on this platform, one could filter down to trips that had a maximum fare of $15. This portal can be accessed at https://opendata.cityofnewyork.us/. Over the years, it has seen growth with an increase in the number of variables recorded. This is a result of changing transportation laws. For instance, an improvement surcharge was levied in 2015 resulting in the creation of a new column that records this. While the TLC provides this data, they do not record it themselves. Instead, it comes from the taxis that usually have a device that tracks fields such as pick-up and drop-off times, the rate code, where the passenger was picked up etc. This device was introduced in 2004 though it did not become mandated until 2008. Taxis are required to have this device provided by TSPs(Technology Service Providers) to track trip data. While the TLC does not directly record it, they do perform routine checks to ensure that it us up the mark. We can thus trust it for the most part. ## Cleaning dataset

It is unclear whether drivers can tamper with these recording devices, and any odd cases will be filtered from the data accordingly. In terms of data quality, a few filters were performed. Since this dataset has been the subject of plenty of previous analyses, there are a few common suggestions made by past researchers. Like Haggag et al.2014, we first started by filtering down to only credit card payments. The reason for this is that non-credit card payments do not have the amount tipped recorded. This is unfortunate as cash tips would likely have changed the amount tipped, owing to people likely just tipping small pocket change. Secondly, I removed a few erroneous records. Records that had 0 recorded for trip distance and trip duration were dropped as well. Finally, records that had the pickup time after the dropoff time were dropped as well. After the application of these filters, the number of records dropped from 7,565,261 records to 5,062,727 records which is roughly a 33% drop in the number of records.

There are 18 raw variables in this data-set. These are the variables that are of interest to me:

- Passenger Count- This variable denotes the number of passengers present in the taxi for a specific trip.

I think this variable is likely to have interesting implications on tipping as I suspect that with more people present in the taxi, an individual is more likely to tip. Tipping is a social norm after all, thus there might be an aspect of social conformity observed in the presence of other individuals. It is also noted within the literature of tipping that people tip as a display of status.

- Fare Amount- This variable is the fare that is presented to the customer at the time of pay,
- Trip Distance- This variable records the length traveled in the trip. As such, a longer trip ought to be correlated with a higher fare amount, after all, the fare is charged dependent on the distance traveled.
- Trip Duration- This variable records how long the trip took. I believe this variable can have unpredictable impacts on tipping. For starters, a long trip does not necessarily co-incide with the distance covered. It could be the result of a traffic jam, road work, lack of familiarity with roads etc. Thus, a patron may be in a poorer mood than the beginning of the trip, which is likely to impact their willingness to tip.
- Time of Day- This is a derived variable from the original variable named *tpep_pickup_datetime*. This variable allows us to observe the impact of the time of day and tipping. There is some research to suggest a slight relationship between sunlight and tipping(Devaraj and Patel, 2017).

-Speed- This is a derived variable where I divide the trip distance by trip duration to give us a rough estimate for speed. There are a few erroneous records where speed was 5696km/h. This record was removed.

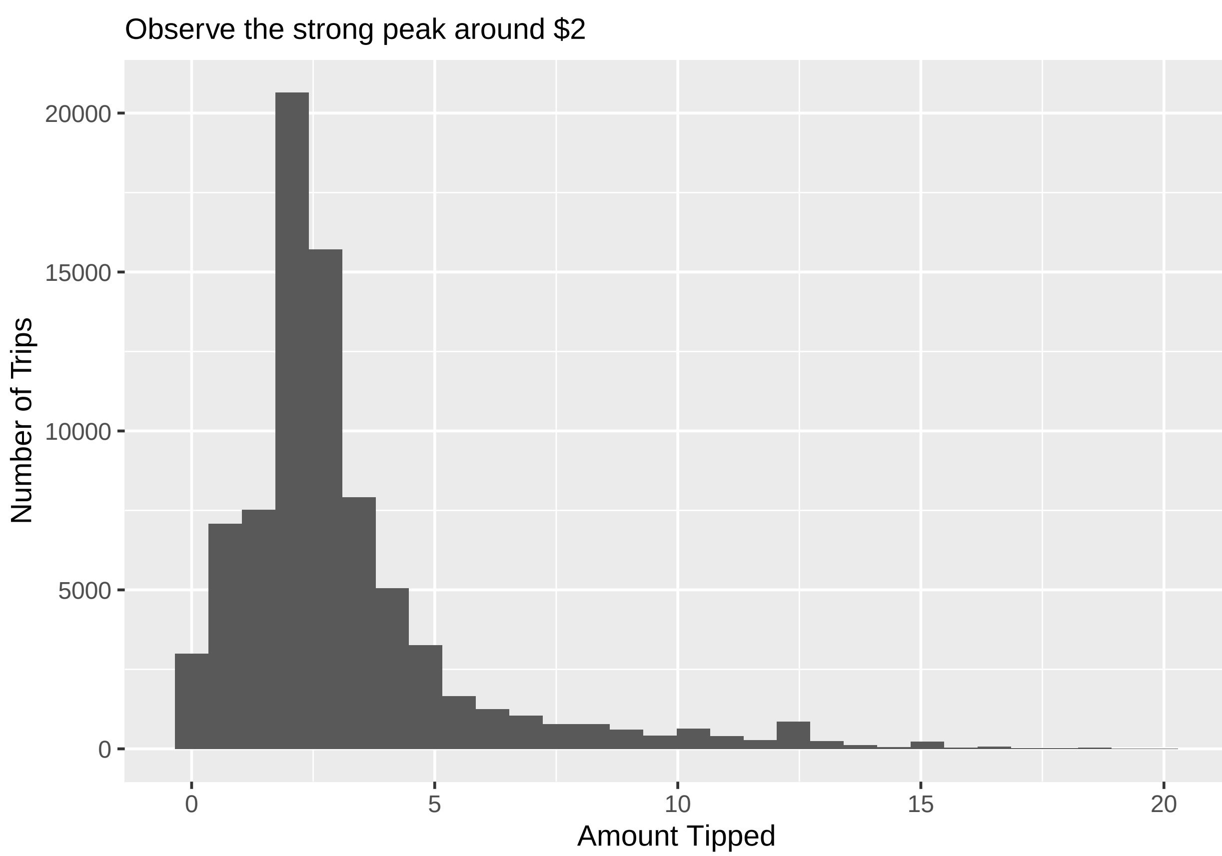


Figure 1: Distribution of tips for Yellow Taxi Cabs in May 2019

Model

To model the relationship between the quantitative factors of a ride and the predicted tip outcome, I am using a linear regression model. A linear regression equation can be represented as

$$Y = \beta_0 + \beta_1 X_i$$

Previous researchers have used a wide variety of models based on how they framed the dependent variable. Lee and Sohn(2020) used a cumulative logit model to identify the relationship between weather and taxi tipping. Their dependent variable was modeled as an ordinal-dependent variable where the tip amount was broken down into three classes- more than 20% of the base fare, default tip option, or less than 20% of the base fare. The reason they chose a cumulative logit model is primarily because of the dependent variable, and the fact that the percentages don't follow a specific distribution. Tjahjadi uses a linear regression model to predict the tip amount(Tjahjadi, 2019), much like I am. Devaraj and Patel(2017) use a linear regression model again to model the impact of sunlight on taxicab tipping, though their response variable is the ratio of tip to the fare amount. As we can see, there are a wealth of options at hand and ultimately model selection boils down to a few factors: how the dependent variable is framed- is it ordinal, categorical, continous; how the predictor variables are identified. For my purpose, I used a stepwise regression to find the best combination of predictor variables. A stepwise regression is a process by which predictor variables are iteratively added/removed to find the combination that reduces prediction error. It chooses the best option that minimizes the AIC-Akaike Information criterion.

Table 1: Checking model collinearity

	GVIF	Df	GVIF^(1/(2*Df))
VendorID	1.00	1	1.00
speed	1.36	1	1.17
fare_amount	1.35	1	1.16
RatecodeID	1.11	1	1.05
time_of_day	1.11	3	1.02

Though a stepwise regression gives the best predictor variables, it does not take into account the issue of multicollinearity. Multicollinearity is when variables are quite linearly related. From our host of possible predictor variables, there is a high degree of multicollinearity. Variables such as trip distance and trip duration are highly linearly related. This seems to negate the assumption that I had earlier stated in the Data section; that it was possible to have a long trip duration without necessarily covering long distances due to possible traffic, roadwork etc. The multicollinearity was checked using the *vif()* function from the *car* package(Fox and Weisberg, 2019). All variables that had a VIF(Variance of Inflation) greater than 4 were rejected.

Table 2: Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	0.32	0.03	12.37	0.00
VendorID	0.10	0.01	8.84	0.00
speed	0.01	0.00	9.41	0.00
fare_amount	0.18	0.00	310.79	0.00
RatecodeID	0.12	0.01	8.23	0.00
time_of_day4pm-8pm	0.04	0.02	2.36	0.02
time_of_day5am-12pm	-0.09	0.02	-5.78	0.00
time_of_day8pm-5am	-0.01	0.02	-0.68	0.49

The model can thus be represented as:

$TipAmount = 0.32{+}0.10{*}VendorID{+}0.01{*}Speed{+}0.18{*}FareAmount{+}0.12{*}RatecodeID{+}0.04{*}TimeOfDay4pmTO8pm{-}$

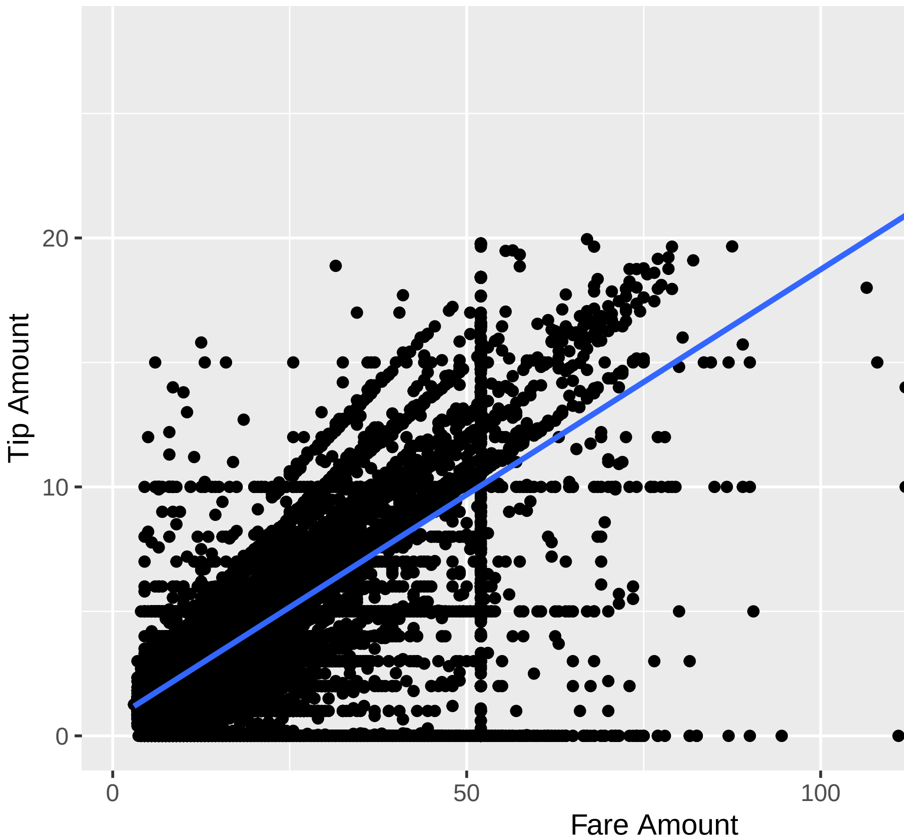
The term that represents whether a ride was between the time of 8pm-5am is removed as it does not reach statistical significance as it has a p-value higher than our alpha of 0.05 .This model can be interpreted by plugging in values for the variables, and then multiplying them by the coefficients that are in front of each variable. We are assessing model accuracy based on the Residual Standard Error(RSE), R-squared, and adjusted R-squared. Our model has a Residual Standard Error of 1.52. This means that the observed tip amount deviates from the true regression line by approximately 1.5 units on average. We can calculate the percentage error as the

$$(RSE)/average(TipAmounts)$$

. In our case, that value is 48.6%. We have an R-Squared and adjusted R-Squared Value of 0.63. A number closer to 1 indicates that this model can explain a large proportion of the variability in a given dataset. Since our number is reasonably close to 1, we can conclude that it accounts reasonably well for the variation in the data.

Results

The entirety of this analysis was done using the R language(R Core Team, 2020). To aid in the graphing of the images, we utilized the ggplot() function that is found in the ggplot2 package(Wickham, 2016) that is present



in the tidyverse library(Wickham et al, 2019).
There are a few interesting results from the above figure. Firstly, there are quite a few rides where the fare amount exceeds $50 but the tip is still 0. We also see a mostly linear relationship though there is a quite a lot of variance.

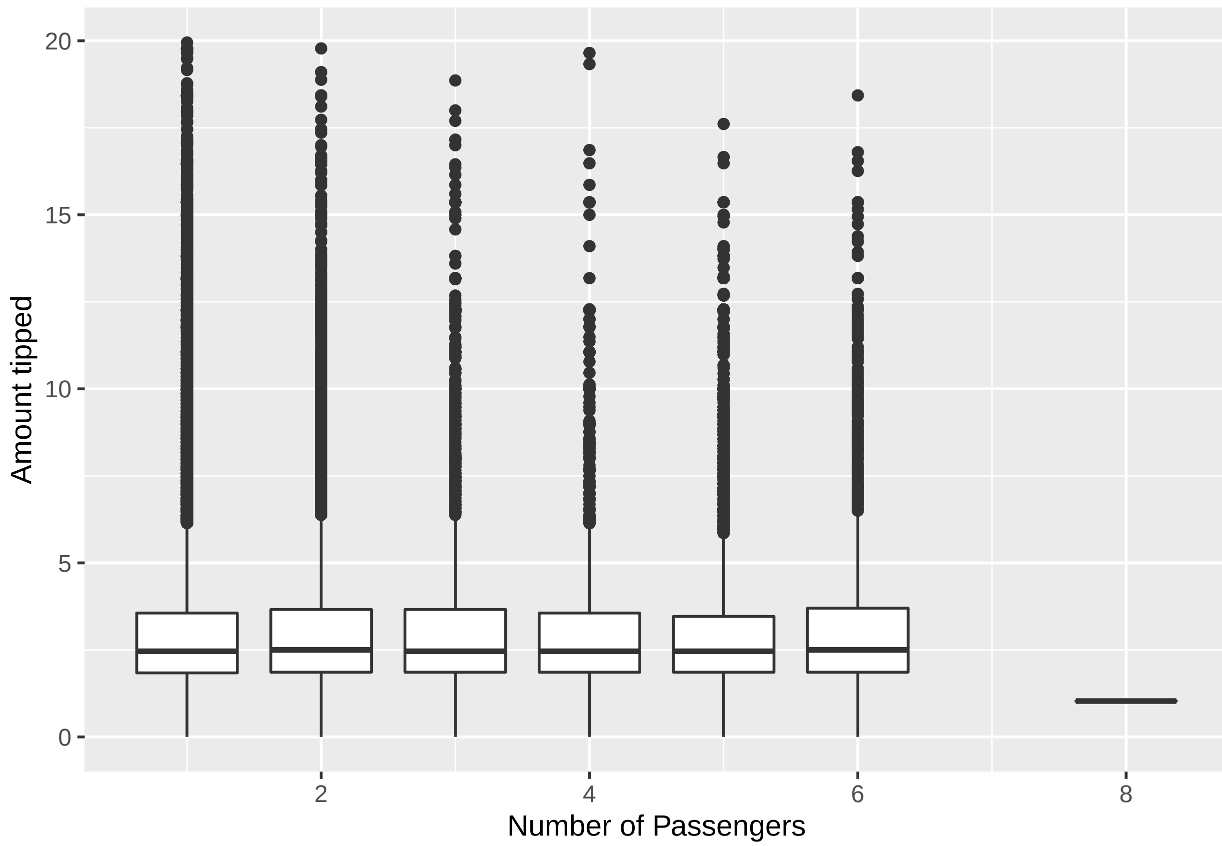
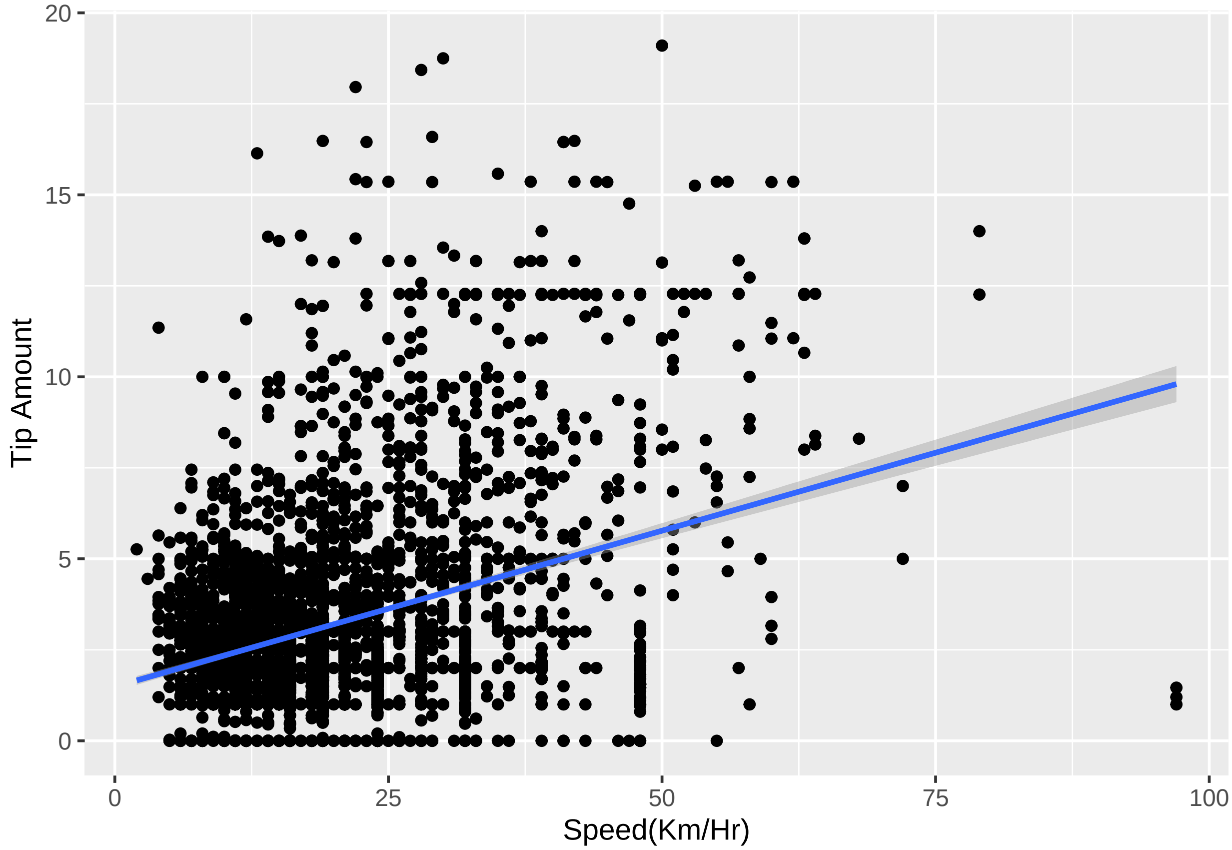


Figure 2: Boxplot of Amount Tipped by the number of passengers present

Figure(3) shows the average amount tipped and spread of the tip by the number of passengers present. There is practically no difference when the number of passengers is increased.



We see an increase in tip as speed increases. This is likely due to a higher speed potentially signalling shorter trips and/or reduced congestion.

From the figure above, we see no significant difference in the spread of tips based on the time of day.

Table 3: Correlation Matrix between predicted and actuals

	X1	X2
X1	1.00	0.79
X2	0.79	1.00

Table 3 indicates that there is a 79% correlation accuracy. This means that actuals and predicted values have similar directional movement.

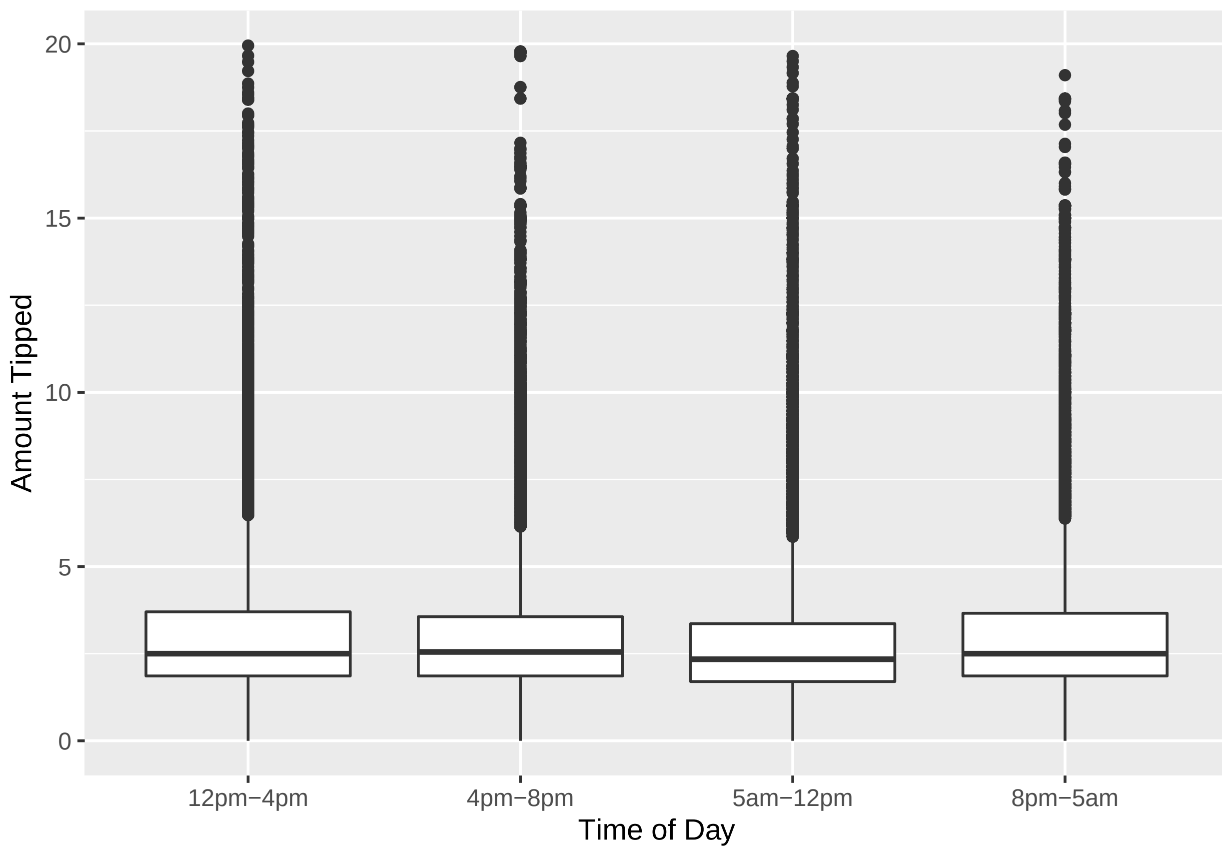
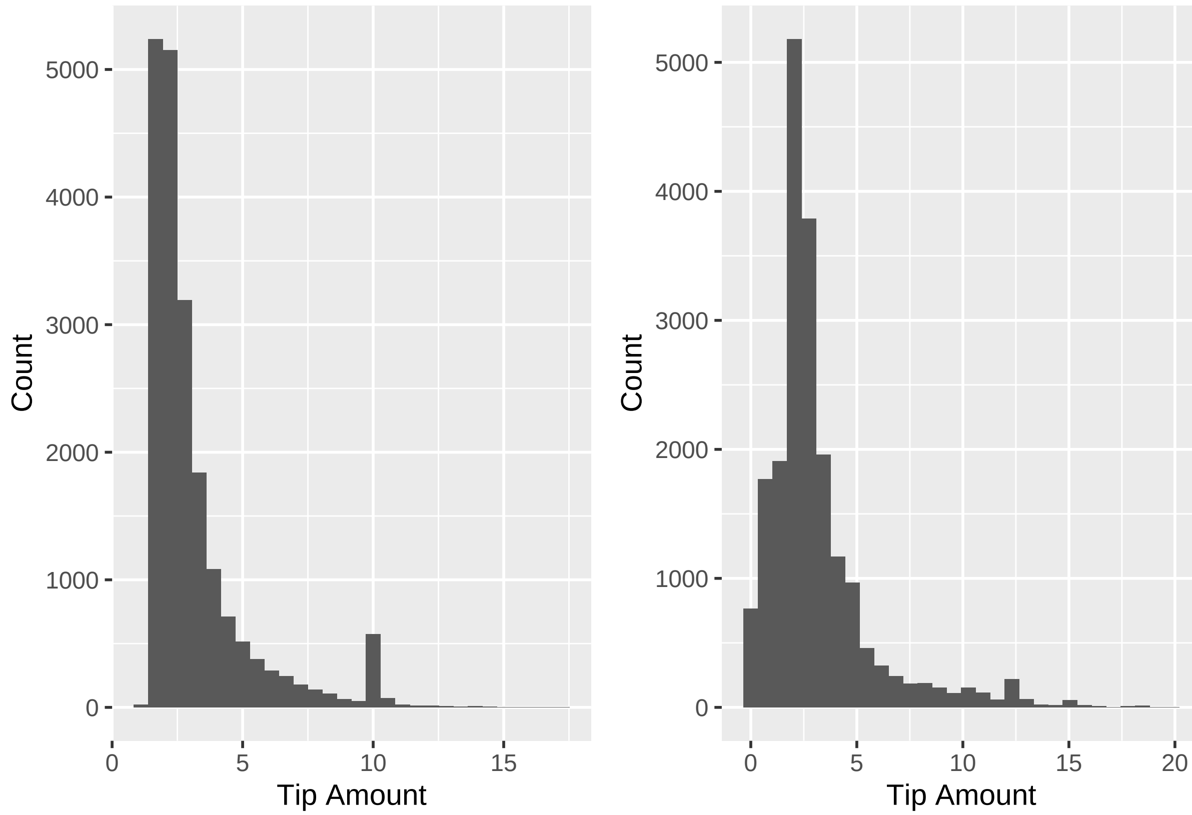


Figure 3: Boxplot investigating spread of tip amount by time of day



We observe a shift in distribution of the tips when predicted, namely they move left changing the norm of the data.

Discussion

Our results indicate a slew of interesting results. For starters, we observe that a few variables that were assumed to be important in the prediction of a tip value, turned out to be either statistically insignificant or subject to multicollinearity. In his analysis titled “A Qualitative Analysis of the Influences of Tipping”, Patrick Tjahjadi found nearly identical variables to predict the tip amount(Tjahjadi, 2019). However, his model had two values that were different. For starters, his model included trip duration as being statistically significant. As far as I can see from his research, there is no check for multicollinearity though he does use the same stepwise regression to find the model that minimizes the AIC. It is possible that given that he is working with a significantly larger dataset, 7.5 million versus my sampled dataset of 99k, the significance of variables may only be noticeable at a larger dataset. Secondly, his data is also older, he uses the entirety of 2015 yellow taxi data, whereas I use just the month of May from 2019. I am in essence working with a very comparable niche of his data.That said, there are still quite a few similarities. Both our analyses found that the number of passengers had no bearing on the amount tipped. This is an interesting finding as it seems to be at odds with some of the literature of tipping. A conceptual framework of Tipping is given by Devaraj and Patel.(2017) where they state that some of the non-economic motivations for tipping include “abiding by social norms, increasing social esteem and social status, and enhancing server welfare, among others”. They go on to say that “Tipping is largely a norm-driven behaviour”(Devaraj and Patel, 2017). One would reasonably assume that when a patron is subject to increased visual scrutiny of how much they tip, they would tip more. Figure 4 provides evidence to the contrary. There is no meaningful difference between groups as defined by the number of passengers present. Since tipping is a bit of an American cultural norm, at least by virtue of its extensive prevalance, do tourists have trouble adapting to this norm? A take on this is provided by Neto et al(2017). In their work, they use the proximity of pickup zones to hotels to determine if a passenger is a tourist or a New Yorker. They conclude that tourists tip more than locals, leaving 0.51% more than a local(Neto et al, 2017). This provides some colour on the faceted nature of the social norms of tipping. The other connection we saw was between time of day and tipping. It is an interesting result as it seems to confirm prior conclusions by Tomasini et al(2017). They noticed that there were no statistically significant differences in tipping behaviour across different times and days(Tomasini et al, 2017). However, a theme that I've noticed is the use of variables in this dataset as a proxy for other things-pickup zone for tourist classification, percentage of tip to fare to determine which top option they pressed on the machine. Another such approximation is time of day for sunlight. Devaraj and Patel(2017) do explicity use sunlight levels, however it does intuitively follows that there is more sunlight during the day. They found that increased exposure to sunlight drove tipping up by 0.5 to 0.7 percentage points(Devaraj and Patel, 2017). Thus, even though my model indicates a negative impact on tipping with early morning rides, there is still some reason to believe that for its inadvertent impact, it may drive tipping up. While time of the day was observed as a feature, and previous research has suggested that day of week did not have an impact(Tomasini et al, 2017), I reckon that time of the year might be a feature. We know that there is evidence to suggest an effect of weather on tipping(Lee and Sohn, 2020) and that there is evidence of self validation that an individual receives when tipping(Devaraj and Patel, 2017). Thus, it becomes clear that in a city like New York, weather and general ambience of the city varies throughout the year. Take for instance, the setting of an average December. December typically has colder weather, shorter days, reduced exposure to sunlight. From the literature, we could conclude that December would see a decrease in tips, as colder weather is likely to have a negative effect on tipping(Devaraj and Patel, 2017, Lee and Sohn, 2020). This is mostly a result of the impact of weather on mood. However, December is also the holiday season for New York. Anecdotally, there is a buzz in New York, Christmas lights are visible in most places, and generally speaking there is a festive spirit. I contend that the generous gift giving nature of the season might result in higher tips. This is not just conjecture, there is a growing body of evidence that highlights the power of suggestion. In a 2013 study conducted by Jacob et al, one group of patrons of a restaurant were shown altruistic quotes and one was not. The study concluded that the group that was exposed to the quotes showed a significant increase in tipping behaviour, in both male and female patrons(Jacob et al, 2013). With that in mind, it becomes evident that the gift-centric marketing of the time is likely to drive an increase in tips. An interesting analysis would be to evaluate the impact of these factors that seemingly work against each other. For this analysis, I made the decision to only focus on one month of data. This decision was made on the following basis. Each file is quite large, even after applying a few filters. The problem is that each month of data is roughly 650MB large, making this close to a 7.2GB file. Even with a PC that was equipped with 24GB RAM, I was unable to combine the resultant dataframes of each month. A potential solution is to sample a limited amount of data from each month and then combine all of them. However, my RAM was still limiting this. While it is hard to say what exactly is "big data", I think this is getting pretty close. With this wealth of data that might grow in complexity or in number of records, I don't think a solution that involves sampling is necessarily the best way to move forward. Instead, I think a more technology conscious solution is to use a cloud technology. The suite of cloud computing softwares offered by major cloud providers such as Amazon, Google, Databricks are a suitable solution. Using a big data processing framework such as SparkR that allows you to utilize the benefits of distributed cloud computing would greatly smooth over this roadblock. As a reference point, check out: https://aws.amazon.com/blogs/big-data/build-machine-learning-powered-business-intelligence-analyses-using-amazon-quicksight/. As mentioned in the introduction, a purely qualitative approach feels incomplete. For instance, the history of tipping indicates that there was a disparity in tipping received based on gender(Eeckhout, 2015). This disparity is not only found in gender, it is also found in race. Research from Ayres and colleagues has also alluded to another instantiation of prejudices in tipping. Using a regression, they concluded from limited data and some caveats that a driver "might expect to earn 56.5% lower tip from an African-American passenger than from a white passenger"(Ayres et al, 2005). They go on to mention that this has this effect of drivers not stopping to pick up minority passengers due to perceived reduced income. Secondly, they also note that African-American drivers received tips that were a third less than a comparable white driver(Ayres et al, 2005). As we can see, tipping is not simply a matter of ride length or the presence of a passenger. It is a complex setting with so many variables at play. Future enhancements of this kind of research need the inclusion of qualitative dimensions. An analysis that at the very least includes the ethnicity of the driver and/or gender would, in my opinion, greatly enrich this field of study.

It certainly is not an easy ask as there are privacy issues and consent issues. For starters, a driver would have to be comfortable disclosing some personally identifiable information, as would the passenger. This current data-set is anonymized which may explain why it is so openly accessible. Adding these dimensions runs the risk of making it harder to access. Another possible exploration is using simulated data. The TLC gives an aggregated breakdown of the drivers' nationality, spoken languages etc. Thus, we could these distrubutions as proxies and perhaps integrate them as a feature. In many studies, the response variable is modelled as a percentage of the base fare. The reason for this is meaningful as it is a corroboration of the tip amount suggedted on the payment machine- 20%, 25%, 30%. Haggag and Paci(2014) found that when the default suggestion is increased, taxi drivers see a decrease in the amount received. A neat experiment would be running a variation of the suggested tips. For instance, run an A/B/C test where group A has no suggested tips, group b has the current treatment, and group c has a different suggested tip. It would be interesting to see the differences in these groups and the effect of this treatment on mean tipping between the groups. The TLC could use this data to come up with default suggestions that maximize tip outcome. There is some evidence to suggest that when passengers choose a custom tip option when presented with the default options, they generally tip less(Tomasini et al, 2017). Still, a more cohesive experiment might yield interesting results. Another variable that I think would greatly influence the predictive power of a model is the inclusion of a ride rating. This record would show what the patron thought of the quality of the ride and driver. Theories behind motivation of tipping often include some discussion of it being a reward for a good service and that it influences service quality(Becker et al, 2012). A driver rating then gives us a much more substantial predictor of ride happiness than something like a congestion surcharge.

The beauty of this dataset is what it enables in tangential research questions. I believe that this industry can provide a great many opportunities for experimenting with projects to increase tipping. One study found that when a restaurant provided free candies to its customers, they saw a 2.78% increase in the average tip(Becker et al, 2012). Thinking about it intuitively, it makes sense. There is a great deal of literature on the impact of mood on tipping behaviour. Anecdotally, think about how at the end of every dental cleaning, they give you a free little toothbrush kit. Even though, I think its safe to assume that most people have toothbrushes. If you have children, you may be familiar with how the pediatrician or family doctor gives their child patients a chocolate to reinforce a good experience of the place. The TLC could run many campaigns to increase tipping outcomes. One such campaign could be the inclusion of a city map for newcomers/tourists. In conclusion, tipping is an interesting social construct that can be quite polarizing. It might be the source of an awkward social encounter where a patron does not feel the need to tip but is publicly shamed for abstaining. It is a substantial source of income for plenty, driving something as large as wage policy. The benefit of a tip prediction model is that may provide a fair value that recognizes the needs of the consumer and the provider.

References

Ayres, Ian & Vars, Fredrick & Zakariya, Nasser. (2003). To Insure Prejudice: Racial Disparities in Taxicab Tipping. Yale Law Journal. 114. 10.2139/ssrn.401201.

B. Ferreira Neto, Amir & Nowak, Adam & Ross, Amanda. (2018). Do Tourists Tip More Than Local Consumers? Evidence from Taxi Rides in New York City. International Regional Science Review. 42. 016001761879843. 10.1177/0160017618798431.

Ben Hamner and Michael Frasco (2018). Metrics: Evaluation Metrics for Machine Learning. R package version 0.1.4. https://CRAN.R-project.org/package=Metrics

Devaraj, S., & Patel, P. C. (2017). Taxicab tipping and sunlight. PloS one, 12(6), e0179193. https://doi.org/10.1371/journal.pone.0179193

Eeckhout, P. (2015). Waiters, Waitresses, and their Tips in Western Europe before World War I. International Review of Social History, 60(3), 349-378. doi:10.1017/S0020859015000504

Elliott, David & Tomasini, Marcello & Oliveira, Marcos & Menezes, Ronaldo. (2017). Tippers and Stiffers: an Analysis of Tipping Behavior in Taxi Trips. 10.1109/UIC-ATC.2017.8397523.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Haggag, K., & Paci, G. (2014). Default Tips.

Jacob, Céline & Guéguen, Nicolas & Ardiccioni, Renzo & Sénémeaud, Cécile. (2013). Exposure to Altruism Quotes and Tipping Behavior in a Restaurant. International Journal of Hospitality Management. 32. 299-301. 10.1016/j.ijhm.2012.03.003.

Jarek Tuszynski (2020). caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc. R package version 1.18.0. https://CRAN.R-project.org/package=caTools

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Lee, Won & Sohn, So. (2020). A large-scale data-based investigation on the relationship between bad weather and taxi tipping. Journal of Environmental Psychology. 70. 101458. 10.1016/j.jenvp.2020.101458.

Mentzer, M. S. (2013). The payment of gratuities by customers in the United States: an historical analysis. International Journal of Management, 30(3), 108+.

NYC Taxi and Limousine Comission. (2019). TLC Yellow Taxi Trip Record Data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Thomas Lin Pedersen (2020). patchwork: The Composer of Plots. R package version 1.0.1. https://CRAN.R-project.org/package=patchwork

Tjahjadi, Patrick. (2019). A Qualitative Analysis of the Influences of Tipping.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

Yihui Xie (2019) TinyTeX: A lightweight, cross-platform, and easy-to-maintain LaTeX distribution based on TeX Live. TUGboat 40 (1): 30–32. http://tug.org/TUGboat/Contents/contents40-1.html