**ISM 6137 Statistical Data Mining**

**New York Taxi Fare and Tip Analysis**

Group 1: Prasad Acharya, Zakaria El Jbari and Jan Hanousek



Table of Contents

[**Executive Summary 3**](#_Toc55741896)

[**Problem definition and Significance 4**](#_Toc55741897)

[**Prior Literature 4**](#_Toc55741898)

[**Data Source and Preparation 5**](#_Toc55741899)

[**Variable choice 6**](#_Toc55741900)

[**Descriptive Analysis and Data Visualizations 7**](#_Toc55741901)

[**Models 9**](#_Toc55741902)

[**Quality Checks: Assumptions 11**](#_Toc55741903)

[**Results and Recommendations 11**](#_Toc55741904)

[**References 13**](#_Toc55741905)

[**Appendix: R Code 14**](#_Toc55741906)

# [Executive Summary](#_zam6wa5axvq2)

COVID-19 lockdowns have further negatively affected the taxi driver industry in NYC, which has already been struggling with the increased competition from UBER and similar ride sharing services. Our team tries to identify the different driving factors behind tip size and total fare to provide advice to taxi drivers and help them identify the most profitable rides and most profitable conditions to work during.

To do so we use records of yellow taxi trips between June 2019 and July 2020 available at NYC official website by month. This project aims to provide a comprehensive analysis of these trips including 2 different models to understand the marginal effects of important internal and external factors on fares and tips. The emphasis of this project is to explore the historical trends and patterns of taxi trips and extract actionable recommendations for taxi drivers to implement. Due to lack of data on the demographic information of the passengers, this project will aim to partially understand these demographics based on proxy variables created based on location of the trip’s origin and/or destination.

The data analyzed strongly supported that temporal and geographic effects on fares and tips exist on a significant level. The main temporal effect was based on a day cycle where fluctuations in both tips and fares were consistent for different times of the day. While monthly variation was insignificant, slight seasonal spike in fares and tips can be observed during Fall. Very small differences exist between the remaining seasons. Finally, the COVID shutdown period resulted in a decrease in both tips and fares consistent with the hypothesized effect of shutdowns of all non-essential transportation. Geographic effects such as the borough and area type(entertainment, workplace, tourist, residential) were a main interest in this analysis to gain some insight into different customer types based on the origin of their trip. Both these effects revealed strong tipping and fare patterns further supporting that they are accurate proxies for different customer types.

Recommendations provided summarize the findings of all these effects to provide taxi drivers and companies a complete and comprehensive analysis of key external factors that affect their profit. These recommendations will act as a complement to the stakeholders’ domain knowledge to come up with optimal strategies that taxi companies can implement on a large scale to maximize the likelihood of generating the highest revenue possible.

Both models passed the reasonable tests to satisfy their assumptions. The models are expected to generate accurate results due to the large random sample directly taken from the true population of taxi rides in New York City. Some limitations exist in our analysis, mainly the possible slight bias from the data cleaning/transformation and feature engineering process and rationale and the lack of availability of real customer data.

# Problem definition and Significance

Uber has taken over a large portion of the market share for ride services in NYC, significantly hurting the Taxi Business. This effect is so significant that it triggered lawsuits to question Uber’s eligibility to offer the service in not only in the NYC but in several cities around the world.

Many taxi drivers are reporting considerably less in income, and taxi drivers are one of the most common jobs in NYC. Furthermore, the COVID-19 lockdowns in NYC has seriously affected the demand for taxi services, which has further negatively impacted the industry. The following are four business questions that we believe can benefit all taxi drivers if answered through proper analysis:

* What driving factors influence a consumers’ decision to tip higher?
* What strategies can taxi drivers implement to bolster their tips?
* How to select trips that will have the largest total fare?
* What is the effect of COVID-19 lockdowns on the total fare and tip size?

Total fare will be mainly determined by the trip length and distance, with some potential surcharges, but we will use linear regression to try and determine the effect of the variables that the taxi driver can either influence or observe before the trip starts. Since we are also modeling tips, which rely on human behavior, our team understands the theoretical maximum for the accuracy of our model should be low to moderate. The coefficients and parameters of the model we choose, however, may provide us with valuable information and insights to patterns we may not see clearly without a model.

# [Prior Literature](#_s4yjji627r8u)

There have been multiple studies looking at both total fare and tip size for taxi drivers.

Since total fare is calculated from a formula, there is a detailed report[[1]](#footnote-1) if how the metered fares are calculated for the NYC yellow taxis available and the different rate changes are explained. Furthermore, Ce and Quiang (2015) conducted a very thorough study of taxi fares looking at pick-up location, drop-off location and start time of the trip. Based on their analysis of individual taxi id’s they concluded that driver’s income was primarily derived from high income taxi rides rather than frequent trips. Building on their conclusion, we will try to analyze which conditions (that are observable before the trip) are best predictive of higher total fare.

Contrary to that tip size is very dependent on human behavior, which makes the theoretical justification of our model not as robust. However, past research shows several predictors with theoretical explanations behind them. Amir B. et al (2019) shows that tourists and theater goers have higher tips, and this effect persists even when they include trip and weather control variables. The explanation for this effect is that theater goers have higher education and therefore income. Since we do have both pick up and drop off locations, we are able to analyze both tourist and entertainment places, which we then include in our model.

Aydin and Acun (2019) conducted a very thorough study and found that tip is very largely driven by the social pressure or the desire to gain social status. As a result, driving in front, talking with the driver, and wearing nice clothes were all significant and positive predictors of tip size. Since the authors had much more detailed data set, we are unable to use the same predictors, but we can use the location of both pick up and drop off as a proxy for potential social status by of the passenger.

Lastly, weather is an important predictor according to Lee and Sohn (2020). They observe that tipping size increases in extreme weather- both temperature and precipitation/snow. This effect can be slightly counterintuitive, since we know that weather affects mood and generally people in better mood tip better. They provide two explanations for this effect. People might be grateful from getting out of the elements into a safety of a taxi and are grateful that they can get easily to their destination. Other explanation might be that people are giving larger tips as a compensation, since driving in heavy rain or snow is more difficult and demanding.

# [Data Source and Preparation](#_ycli8h4krq6k)

We downloaded our data set from <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>. It records all taxi trips that occurred in the city each day, with exact timestamp, location, number of passengers, trip length (miles), fare amount and tip (dollars), surcharges/tolls (dollars), pick-up and drop-off location, and the payment method for yellow cabs.

As we mentioned in the previous section, there have been several predictors identified by the past literature, with weather and passenger personal characteristic being the best of them. Since we do not have any data on the passenger, other than the number of them, we decided to use the pick-up location and drop-off location as a proxy for the type of the traveler and trip.

The location variable provided in the data divides New York city into 273 different zones. There are two types of zones: Yellow zone, which is generally a commercial area, and Borough zone, which is residential area. We manually went through the entire list of different zones and categorized each Yellow zone into one of the following categories: tourist destination, entertainment destination, park and workplace.

There is of course overlap between some of the variables. If the location is a park, we categorize it as such, even though it might be a tourist destination (such as Central Park). We do the same for entertainment destination-if a place could be not only for tourist (SoHo) we only categorize it as an entertainment destination. As a result, the tourist destinations are those places that would only interest tourists (Empire State building, World Trade center, etc.). Lastly, workplace was created because there were both UN headquarters and the Finance district as zones, which did not fit any other description.

For Borough zones, or residential areas, we decided to use the median real estate value to estimate how “rich” the neighborhood is and how potentially rich is the customer. The real estate median prices were manually downloaded from <https://www.neighborhoodscout.com/ny/> . Of course, there was not always a perfect match between the databases-sometimes the neighborhood was very large and there were several median values for the same Borough Zone. In those cases, we would take the median observation amongst them to minimize potential bias.

The weather data for this project was obtained from the Wunderground website that reports hourly weather observations for New York City every day. Due to an absence of open-source API for this information, we decided to scrape the information directly from their website. The website’s source code revealed that the content was not static, therefore we decided to use the Python Selenium package. The data was then integrated into our main data frame by matching the customer pickup date and time to the same date and hour of the day from the weather data frame.

For the data cleaning we firstly restricted our sample to only those trips that paid electronically. This was an important step, since most trips that were paid in cash recorded zero tip, even though there is a high chance that there actually was a tip. As a result, our sample size decreased from 51 million observations to 41 million observations of only trips paid electronically.

We then removed all observations with non sensical records, such as negative trip length or total amount, and trips that had very large fare to trip length ratio. This decreased the number of observations from roughly 41 million to around 39 million. However, only 1.7 million observations were observations after the COVID-19 lockdowns, which was too low to analyze the effect. To combat that we took a random sample of more than 5 million of observations before the lockdowns and all the observations after it. In this final sample we have almost 7 million observations, where 20% of them were after the lockdown started. This left us with large enough sample to be representative of the dataset, but with large enough portion of the data after the lockdown to analyze the effect.

# Variable choice

The variable selection process was designed to emphasize the selection of variables that will both explain the variation in taxi tips and fares, as well as generate useful business insights and recommendations for drivers. The fare model specification is straightforward as the fare amount is mainly driven by the trip time and distance, but other variables such as time of day and pick-up location also add slight variation to the fare amount. All variables that are related to the above explanation for this model were selected for this analysis.

The variable selection process for modeling the tips proved to be more challenging. The main reason for the complexity of this analysis is the fact that the dependent variable value for each trip is the result of a human decision-making process. Considering this fact along with the data at hand, our data preparation and feature engineering process needed variables to try and explain two main factors that mainly drive this decision:

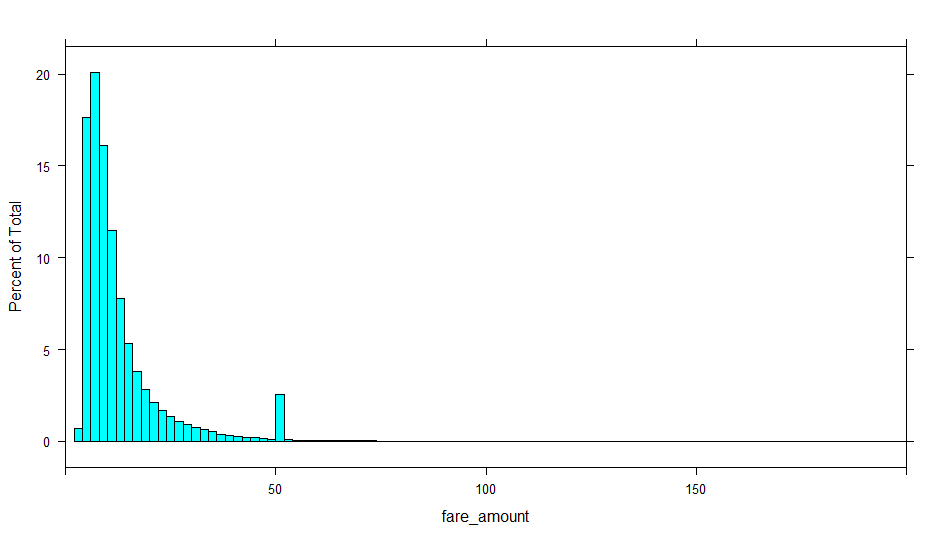
1. The type of passenger(s) with an emphasis on their socioeconomic status
2. Any external factors that may affect the state of mind or mood of the passenger(s)

As detailed in the data preparation process, we will be using all temporal and geographic variables in this model to try and gain some insight into patterns that form amongst different types of passengers. A compact list and description of these variables can be seen in the table below:

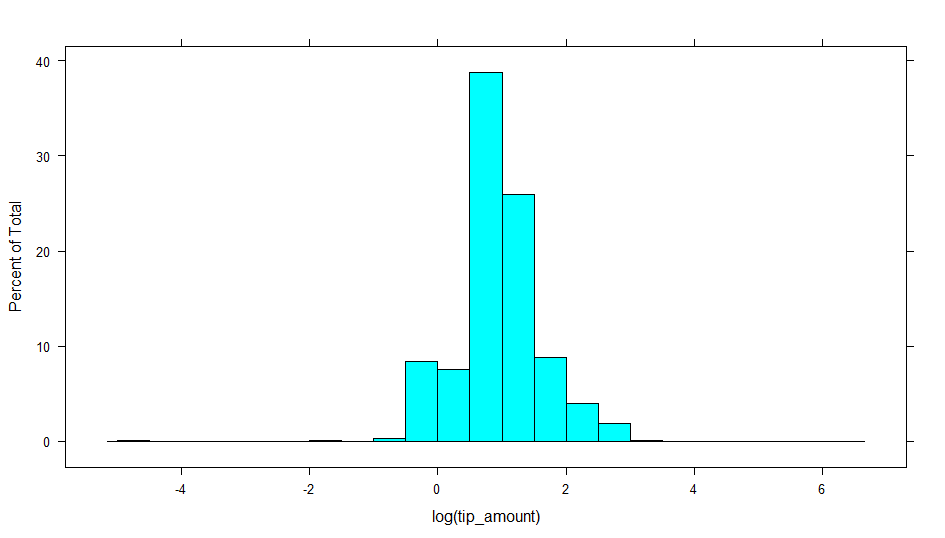
|  |  |
| --- | --- |
| **Variable** | **Description** |
| **Taxi variables** | |
| Trip time | Trip time in minutes |
| Trip distance | Trip distance in miles |
| Passenger count | Total number of passengers |
| **Location variables-both Pick up (PU) and Drop-off (DO)** | |
| Borough | The location borough |
| Tourist | If the location was nonresidential and was categorized as a tourist location. |
| Entertainment | If the location was nonresidential and was categorized as an entertainment spot (shopping centers, theaters, etc.) |
| Workplace | If the location was either the UN Headquarters or the Financial district in Manhattan |
| Park | If the location was nonresidential and was categorized as a park |
| Residential | If the location was categorized as a residential area. |
| Median Real Estate price | If the location was categorized as a residential area this variable specifies the median real estate price in that neighborhood. Zero otherwise. Used as a proxy for the wealth of the passenger. |
| **Weather related** | |
| Extreme temperature | If the temperature was above 86 or below 21 Fahrenheit |
| Good condition | If there is no extreme weather condition (no rain, thunderstorm, hail, snowing, etc.) |
| **COVID-19 related** | |
| COVID lockdown | Equal one when the trip occurred after the lockdown on 8th of March |
| **Time variables** | |
| Year | Year of the trip |
| Month | Month of the trip |
| Hour | Hour of the pick-up |
|  |  |

# Descriptive Analysis and Data Visualizations

The initial descriptive analysis was to understand the distribution of each dependent variable to select the right modeling technique. Here we are showing the histograms for both of the dependent variables.



As can be seen from the histogram above, the fare amount variable, although not completely normal, is normal enough for the robustness of OLS to yield accurate coefficients that explain the marginal effects of each predictor. Another important fact to consider is that the data available represents a large random sample that was obtained from the true population of rides that occurred over the past year.



The tips variable histogram was extremely skewed, and a log transformation stabilized the variable very well. This suggests that we either need to model with a variance stabilizing transformation on the dependent variable(i.e. log(y)) or use Poisson regression.

# Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Dependent variable:* | Fare amount |  |  |  |
|  |  |  |  |  |
| Trip time | 0.003\*\*\* (0.00002) |  | Hour 10 | 1.458\*\*\* (0.010) |
| Trip distance | 2.831\*\*\* (0.0004) |  | Hour 11 | 1.545\*\*\* (0.010) |
| (PU) Brooklyn | -0.427\*\*\* (0.044) |  | Hour 12 | 1.535\*\*\* (0.010) |
| (PU) EWR | 11.625\*\*\* (0.559) |  | Hour 13 | 1.466\*\*\* (0.010) |
| (PU) Manhattan | -0.768\*\*\* (0.042) |  | Hour 14 | 1.547\*\*\* (0.010) |
| (PU) Queens | -1.283\*\*\* (0.042) |  | Hour 15 | 1.668\*\*\* (0.010) |
| (PU) Staten Island | -2.123\*\*\* (0.181) |  | Hour 16 | 1.656\*\*\* (0.010) |
| (PU) Unknown | -0.299\*\*\* (0.045) |  | Hour 17 | 1.583\*\*\* (0.010) |
| (PU) Tourist | -0.578\*\*\* (0.005) |  | Hour 18 | 1.319\*\*\* (0.010) |
| (PU) Entertainment | 0.119\*\*\* (0.004) |  | Hour 19 | 0.910\*\*\* (0.010) |
| COVID lockdown | -1.143\*\*\* (0.003) |  | Hour 20 | 0.604\*\*\* (0.010) |
| Extreme condition | -0.084\*\*\* (0.007) |  | Hour 21 | 0.485\*\*\* (0.010) |
| Good condition | -0.358\*\*\* (0.010) |  | Hour 22 | 0.428\*\*\* (0.010) |
| Hour 1 | -0.200\*\*\* (0.013) |  | Hour 23 | 0.267\*\*\* (0.011) |
| Hour 2 | -0.319\*\*\* (0.015) |  | Constant | 4.750\*\*\* (0.043) |
| Hour 3 | -0.430\*\*\* (0.017) |  |  |  |
| Hour 4 | -0.534\*\*\* (0.020) |  | Observations | 6,957,039 |
| Hour 5 | -0.550\*\*\* (0.018) |  | R2 | 0.904 |
| Hour 6 | -0.227\*\*\* (0.012) |  | Adjusted R2 | 0.904 |
| Hour 7 | 0.491\*\*\* (0.011) |  | Residual Std. Error | 3.525 (df = 6957002) |
| Hour 8 | 1.334\*\*\* (0.010) |  | F Statistic | 1,824,995.000\*\*\* (df = 36; 6957002) |
| Hour 9 | 1.588\*\*\* (0.010) |  |  |  |
|  |  |  | *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |

The output for the best model for the fare dependent variable is shown above. This model was a OLS model of the dollar fare amount as a function of the trip time, trip distance, borough where the passenger was picked up, whether the ride occurred during COVID shutdown, the temperature, weather condition, and finally the time of day when the trip started. This model had a good fit as expected since the fare is calculated using a combination of the outlined factors. The most important takeaways from this model are the marginal effects of the location variables and time of day on tips, since drivers can control these factors and implement strategies to optimize their performance.

The model regarding the tip size was more comprehensive; results are shown on the next page:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dependent variable: | Tip amount rounded |  |  |  |
|  |  |  |  |  |
| Passenger count | 0.005\*\*\* (0.0002) |  | Month 6 | -0.007\*\*\* (0.002) |
| (PU) Brooklyn | 0.457\*\*\* (0.010) |  | Month 7 | 0.0002 (0.001) |
| (PU) EWR | 1.071\*\*\* (0.044) |  | Month 8 | -0.010\*\*\* (0.001) |
| (PU) Manhattan | 0.484\*\*\* (0.010) |  | Month 9 | 0.021\*\*\* (0.001) |
| (PU) Queens | 1.122\*\*\* (0.010) |  | Month 10 | 0.025\*\*\* (0.001) |
| (PU) Staten Island | -0.467\*\*\* (0.038) |  | Month 11 | 0.022\*\*\* (0.001) |
| (PU) Unknown | 0.466\*\*\* (0.010) |  | Month 12 | 0.033\*\*\* (0.001) |
| (PU) Tourist | 0.054\*\*\* (0.001) |  | Hour 1 | -0.015\*\*\* (0.002) |
| (PU) Entertainment | 0.034\*\*\* (0.001) |  | Hour 2 | -0.036\*\*\* (0.002) |
| (PU) Workplace | 0.122\*\*\* (0.001) |  | Hour 3 | -0.035\*\*\* (0.003) |
| (PU) Residential | -0.332\*\*\* (0.002) |  | Hour 4 | -0.004 (0.003) |
| (PU) Median real estate price | 0.00000\*\*\* (0.000) |  | Hour 5 | -0.002 (0.003) |
| (DO) Brooklyn | 0.044\*\*\* (0.003) |  | Hour 6 | -0.058\*\*\* (0.002) |
| (DO) EWR | 0.550\*\*\* (0.004) |  | Hour 7 | -0.051\*\*\* (0.002) |
| (DO) Manhattan | -0.331\*\*\* (0.003) |  | Hour 8 | -0.039\*\*\* (0.002) |
| (DO) Queens | 0.177\*\*\* (0.003) |  | Hour 9 | -0.009\*\*\* (0.002) |
| (DO) Staten Island | 0.318\*\*\* (0.008) |  | Hour 10 | -0.001 (0.002) |
| (DO) Unknown | -0.260\*\*\* (0.004) |  | Hour 11 | 0.008\*\*\* (0.002) |
| (DO) Tourist | 0.144\*\*\* (0.001) |  | Hour 12 | 0.003\* (0.002) |
| (DO) Entertainment | 0.017\*\*\* (0.001) |  | Hour 13 | -0.010\*\*\* (0.002) |
| (DO) Workplace | 0.112\*\*\* (0.001) |  | Hour 14 | 0.010\*\*\* (0.002) |
| (DO) Residential | -0.198\*\*\* (0.001) |  | Hour 15 | 0.011\*\*\* (0.002) |
| (DO) Median real estate price | 0.00000\*\*\* (0.000) |  | Hour 16 | 0.043\*\*\* (0.002) |
| Fare amount | 0.009\*\*\* (0.00001) |  | Hour 17 | 0.036\*\*\* (0.002) |
| Extreme temperature | 0.005\*\*\* (0.001) |  | Hour 18 | 0.023\*\*\* (0.002) |
| Good condition | -0.008\*\*\* (0.002) |  | Hour 19 | 0.013\*\*\* (0.002) |
| COVID lockdown | -0.026\*\*\* (0.002) |  | Hour 20 | 0.014\*\*\* (0.002) |
| Month 2 | 0.004\*\*\* (0.001) |  | Hour 21 | 0.024\*\*\* (0.002) |
| Month 3 | 0.005\*\*\* (0.002) |  | Hour 22 | 0.028\*\*\* (0.002) |
| Month 4 | -0.004\* (0.002) |  | Hour 23 | 0.013\*\*\* (0.002) |
| Month 5 | -0.022\*\*\* (0.002) |  | Constant | 0.766\*\*\* (0.010) |
|  |  |  |  |  |
| Observations | 6,957,039 |  |  |  |
| Note: | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |  |  |  |

The final model for tips was a Quasi-Poisson Regression model, since the Poisson model was overdisperesed (despite the very high number of degrees of freedom). The tip variable was transformed into an integer amount using the ceiling function. This was decided in order to avoid creating artificial zeros by rounding small tips amount to 0 dollars.

Ordinarily, there would be a concern over excess zeros, which can be addressed through hurdle or zero-inflated models. In our case, however, we do not believe that the non-tippers are predicted using a different **model** than we are presenting (or rather there are no other available predictors). Furthermore, the portion of zero tips in our sample is very low (around 3%), which leads us to conclude that excess zeros are not an issue for this data.

Since the focus of this model is to understand any noticeable tipping patterns for similar trips, we are focusing on the variables that taxi drivers can control, such as pick up location and time of day. The model takeaways and recommendations will be discussed in the following section.

# Quality Checks: [Assumptions](#_aqifice7ntvm)

The **fare** OLS model assumptions were all satisfied by the best model presented-Linearity, Homoscedasticity, Normality, Independence and Multicolineraity. The test results were more difficult to interpret, because of the statistical paradox (and our very large sample size), but since fare is calculated as a linear function of time and distance traveled, OLS seems to be theoretically and practically the best choice.

The **tip** model was Quasi-Poisson, since normal Poisson regression was overdispersed. It also met the assumption of Independence and Multicolinearity. Additonally, as we mentioned previously, we do not believe there to be excess zeroes-there is very small number of observations with zero tips and we do not believe there is a different data generating process for not tipping other than we have controled for the tip size model.

# Results and Recommendations

The results and recommendations for taxi drivers to bolster their **tips and fares** are outlined below and categorized by variable types. Our findings show that the location variables did a good job at capturing significant tipping and fare patterns. The temporal variables showed that while short term variation in tips and fares caused clear strong patterns, long term cyclicality was not present in the data with the exception of a decline in both fares and tips post COVID shurdowns. This is important observation, since the trip frequency, which is not modelled, also decreased drastically under the COVID 19 lockdowns Finally, the weather variables’ marginal effects were not significant in increasing tips, contradicting previous research.

**1.Location Analysis**

Borough Pick-up Results

|  |  |
| --- | --- |
| Tips: | Our base level for this analysis was the Bronx Borough. In comparison to the base level, trips originating in Queens, EWR, Manhattan, and Brooklyn were 112%, 107.1%, 48.4%, and 45.9% higher on average, respectively. Trips in Staten Island were 46.7% lower than our base level on average. |
| Fare: | Compared to the Bronx borough, fares for trips starting in Newark (EWR) airport were $11.74 higher on average. |

Recommendations:

Considering that the goal is to maximize fares and tips, taxi drivers should place themselves in areas where customers are more likely to tip generously while resulting in a large fare. Our analysis suggests that taxi drivers should aim to position themselves in the EWR (Newark airport) borough given the fact that the tips are only 5% lower on average than the Queens borough, but fares are usually over $12 higher! It is important to note that the fare amount is not as important as the tips since large fares can be accredited to longer trips that consume more time. This means that the while the driver will make more on a large fare compared to a smaller one, this does not consider the fact that the driver could accomplish multiple short trips in the same time as a long one. With this in mind, location marginal effects on tips should be emphasized over fares when making a decision about the optimal target location. The remaining boroughs are listed in the order of importance to boosting trips: Queens, Manhattan, Brooklyn, Bronx, and Staten Island. We strongly recommend that taxi drivers avoid Staten Island and Bronx as a starting point as tips tend to be lower than other boroughs by somewhere between 160% and 45% depending on the borough.

**2.Area Sub-Type Results:**

Because the boroughs are huge sections that include different types of areas, we created additional categorical variables that yielded very interesting results and insights to explain taxi tip patterns. Most notably, trips that started or ended in a residential location had severely lower tips on average (33.2% and 19.8% respectively). For the non-residential locations, trips that ended in either workplace or tourist locations had higher tips. This is supporting previous research that showed that tourists tend to out-tip the locals. However, unlike workplace trips, fares for tourist trips are slightly lower. Lastly, we see that trips originating in entertainment areas have higher tips, which is again supported by research, since tipping is largely influenced by good mood. The fare was not significantly impacted by area subtype(less than $1).

Pattern Observations and Effect rationale:

Locals are most likely to start their trips from residential areas and the negative results may be attributed to locals not feeling the need to tip taxi drivers.

The workplace variable is centered around locations such as the financial district and the UN building where a lot of these professionals make a very comfortable living and rely on taxis for transportation due to their convenience compared to subways or other alternate transportation methods. We can see that the effects are very consistent for drop off and pick up possibly showing that these professionals tip as a percentage consistently regardless of their mood.

The tourist variable also showed that both a tourist are origin and destination yield higher tips on average, the tourist destination proved to have double the impact on tips. This could be explained by the fact that people are usually in a favorable mood and intend to spend money on their touristic activities when they are headed to such places, hence the higher tip. Since they spend some money during their day, they will likely still tip higher due to their mood but not as high because they probably already spent money on their other activities.

Recommendations:

* Taxi drivers should combine knowledge about the boroughs and specific type of area to choose the area with maximum likelihood of higher tips and fares. As far as type of areas, taxis should primarily target workplace areas, followed by touristic areas.
* Taxis should try to avoid residential areas as much as possible

**3.Time effects**

1. Hours of the day

Tips are the highest at 4pm and lowest at 6am. Time patterns show a very smooth and consistent patterns. The time window where tips are the most favorable is 4pm-11pm. Possible explanations include that those hours constitute the end of workdays for most people, which can again affect the mood. Another important factor to consider is that these are rush hour peak times so customers might be extra grateful that they were able to secure a ride during these times and they are showing their gratitude through higher tips. Further evidence supporting this rationale is that the least favorable times for tips are the morning. This could be explained simply by the fact that most people are generally either tired, stressed and/or in bad moods since they are headed to work.

Recommendations:

* Strategically choose mid-day through midnight shifts for driving to maximize tips with an emphasis on the 4pm-11pm window.
* Avoid early morning shifts if possible

1. Monthly Patterns/Seasonality

The month variable was introduced to control for any cyclical/seasonal variation in tips. The results yielded very interesting results as far as seasonal as opposed to monthly variation. We can see that Fall is the most favorable season for tips, followed by Spring, Winter, and Summer, respectively. It is important to note that while the tips during the months of the fall season were between 1.4% and 3.3% higher than rest of the months, the differences between the rest of the months and seasons were not very significant(between 0.0004% and 1%).

1. COVID-19 lockdowns

We see that tips for taxi rides that occurred during New York’s COVID shutdown were 2.6% lower on average than trips that happened prior to the shutdowns**.** Shutdown had a negative impact on tips due to the economic depression that it caused for most people in the country and especially in NYC due to stricter COVID safety legislation. Our analysis shows that the lockdowns not only affected the frequency of trips, but also the type of trips and tip size, which is generally unobserved by policy makers.

**4**.**Weather**

The effect of the uncomfortable weather variables was positive as hypothesized. Poor conditions such as thunderstorms, heavy snow, or rain, seemed to increase tips by 1% compared to favorable weather conditions such as fair or overcast and trips started during extremely hot and cold temperatures resulted in tips that were 0.5% higher on average than trips in moderate temperatures. However, we see that the marginal effect was not very significant showing that while these external factors positively affect some customers’ decision to tip, other customers rely on a standard tipping strategy that is not impacted by any external factors. Since driving in the extreme conditions or temperatures can be more difficult, it might be optimal to avoid working during these conditions, despite the positive estimated effect.

# References

“Trip Fare Estimation Study from Taxi Routing Behaviors and Localizing Traces” Liu, Ce, Qu, Qiang   
*2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015*

“Do Tourists Tip More Than Local Consumers? Evidence from Taxi Rides in New York City” Amir B. Ferreira Neto, Adam Nowak, Amanda Ross *International Regional Science Review* | *Vol 42, Issue 3-4, 2019*

“An investigation of tipping behavior as a major component in service economy: The case of taxi tipping” Asli Elif Aydin, Yüksel Acun *Journal of Behavioral and Experimental Economics Volume 78, February 2019, Pages 114-120*

# Appendix: R Code

rm(list= ls())

library(data.table)

library(AER)

library(car)

library(lattice)

library(splitstackshape)

library(base)

library(corrplot)

library(tidyverse)

library(MASS)

library(stargazer)

#install.packages("pscl")

library(pscl)

library(Hmisc)

setwd("D:/USF Statistical data mining/Stat\_data\_mining/Taxi")

fread("yellow\_tripdata\_2020-06.csv")->taxi1

fread("yellow\_tripdata\_2020-05.csv")->taxi2

fread("yellow\_tripdata\_2020-04.csv")->taxi3

fread("yellow\_tripdata\_2020-03.csv")->taxi4

fread("yellow\_tripdata\_2020-02.csv")->taxi5

fread("yellow\_tripdata\_2020-01.csv")->taxi6

fread("yellow\_tripdata\_2019-12.csv")->taxi7

fread("yellow\_tripdata\_2019-11.csv")->taxi8

fread("yellow\_tripdata\_2019-10.csv")->taxi9

fread("yellow\_tripdata\_2019-09.csv")->taxi10

fread("yellow\_tripdata\_2019-08.csv")->taxi11

fread("yellow\_tripdata\_2019-07.csv")->taxi12

#the format is the same, as are the colnames->can easily just rbind the whole data

rbind(taxi1,taxi2)->taxi

rbind(taxi,taxi3)->taxi

rbind(taxi,taxi4)->taxi

rbind(taxi,taxi5)->taxi

rbind(taxi,taxi6)->taxi

rbind(taxi,taxi7)->taxi

rbind(taxi,taxi8)->taxi

rbind(taxi,taxi9)->taxi

rbind(taxi,taxi10)->taxi

rbind(taxi,taxi11)->taxi

rbind(taxi,taxi12)->taxi

rm(taxi1,taxi2,taxi3,taxi4,taxi5,taxi6,taxi7,taxi8,taxi9,taxi10,taxi11,taxi12)

taxi %>% filter(payment\_type!=2)->taxi#losing close to 10 million observation,

#which is around 20%

#the 4 other categorires apart from cash are from not completed trips!

#we cannot use that data, since it is different from the completed trips with competely

#different dynamic..the total sum of observations of the 4 categories that I am excluding

#is less than half a million-we still have enough observations

#taxi %>% filter(payment\_type==6) %>% nrow()

taxi %>% filter(payment\_type==1)->taxi

#adding my data

read.csv("taxi+\_zone\_lookup.csv")->zone

nafill(zone$tourist,fill=0)->zone$tourist

nafill(zone$entertainment,fill=0)->zone$entertainment

nafill(zone$park,fill=0)->zone$park

nafill(zone$workplace,fill=0)->zone$workplace

nafill(zone$other,fill=0)->zone$other

summary(zone)

gsub(zone$median\_realstate,pattern="\\$",replacement="")->zone$median\_realstate

gsub(zone$median\_realstate,pattern="\\,",replacement="")->zone$median\_realstate

zone$median\_realstate %>% as.numeric()->zone$median\_realstate

ifelse(zone$service\_zone=="Boro Zone",1,0)->zone$residential

zone %>% select(-c(Zone,service\_zone,other))->zone

#first pick up

colnames(zone)=c("PULocationID","borough\_pu","median\_rlst\_pu","tourist\_pu","entert\_pu",

"park\_pu","workplace\_pu","residential\_pu")

left\_join(taxi,zone,by="PULocationID")->taxi

#then drop off

colnames(zone)=c("DOLocationID","borough\_do","median\_rlst\_do","tourist\_do","entert\_do",

"park\_do","workplace\_do","residential\_do")

left\_join(taxi,zone,by="DOLocationID")->taxi

fwrite(taxi,file="taxi\_data.csv")

#fread("taxi\_data.csv")->taxi

########data cleaning

taxi <- subset(taxi, trip\_distance>0 & fare\_amount>0 & extra>=0 & mta\_tax>=0 & tip\_amount>=0 & tolls\_amount>=0 & improvement\_surcharge>=0 & total\_amount>0 & congestion\_surcharge>=0 & passenger\_count>0)

#lose almost 2 million obs

summary(taxi)

taxi$rate\_fare=taxi$fare\_amount/taxi$trip\_distance

taxi %>% filter(!((rate\_fare>100)&(total\_amount>100)))->taxi

#losing about 2,5 thousand obs

summary(taxi)

#fwrite(taxi,file="taxi\_data.csv")

#fread("taxi\_data.csv")->taxi

#adding weather

readxl::read\_excel("WeatherData.xlsx")->weather

weather$...1=NULL

#formating

format(weather$date\_time,"%H")->weather$hour

as.Date(weather$date,format=c("%Y-%m-%d"))->weather$date

#deleting duplicates

weather %>% select(date,hour) %>% duplicated()->weather$dups

weather %>% filter(dups==FALSE)->weather

weather$dups=NULL

#creating hour and date for taxi

str\_sub(taxi$tpep\_pickup\_datetime, -8, -7)->taxi$hour

as.Date(taxi$tpep\_pickup\_datetime)->taxi$date

left\_join(taxi,weather,by=c("date","hour"))->taxi

#removing unnecessary variables

taxi$`Dew Point`=NULL

taxi$Wind=NULL

taxi$`Wind Gust`=NULL

taxi$date\_time=NULL

taxi$time\_24=NULL

taxi$Time=NULL

taxi$date=NULL

taxi$hour=NULL

#some data transformation so that we can use them in the regression

str\_replace(taxi$Temperature,pattern = "F",replacement = "")->taxi$Temperature

str\_trim(taxi$Temperature)->taxi$Temperature

taxi$Temperature %>% as.numeric()->taxi$Temperature

taxi %>% drop\_na(Temperature)->taxi

str\_replace(taxi$Humidity,pattern = "%",replacement = "")->taxi$Humidity

str\_trim(taxi$Humidity)->taxi$Humidity

#humiodity was in %

taxi$Humidity %>% as.numeric()->taxi$Humidity

str\_replace(taxi$`Wind Speed`,pattern = "mph",replacement = "")->taxi$`Wind Speed`

str\_trim(taxi$`Wind Speed`)->taxi$`Wind Speed`

taxi$`Wind Speed` %>% as.numeric()->taxi$`Wind Speed`

str\_replace(taxi$Pressure,pattern = "in",replacement = "")->taxi$Pressure

str\_trim(taxi$Pressure)->taxi$Pressure

taxi$Pressure %>% as.numeric()->taxi$Pressure

str\_replace(taxi$Precip.,pattern = "in",replacement = "")->taxi$Precip.

str\_trim(taxi$Precip.)->taxi$Precip.

taxi$Precip. %>% as.numeric()->taxi$Precip.

summary(taxi)

df=taxi

rm(taxi)

#need to create a subsample to use

# Make a covid variable based on shutdown date

df$covid= ifelse(sample$tpep\_dropoff\_date>'2019-03-08',1,0)

df %>% filter(covid==1)->df\_covid

df %>% filter(covid==0)->df\_non

subsample = sample(1:nrow(df\_non), size=round(0.15\*nrow(df\_non)), replace=FALSE)

df\_non <- df\_non[subsample,]

rm(df)

rbind(df\_non,df\_covid)->df

rm(df\_covid,df\_non,subsample)

# Further filtering

# get rid of any fares under 2.50 because that's the nyc minimum

df %>% filter(fare\_amount>=2.5)->df

# Engineered features

# Get rid of 27 categories into a dummy for condition

# bad conditions include any of the categories below

df$good\_condition= ifelse(df$condition== "Snow"|

df$condition=="Rain / Windy"|

df$condition=="Heavy Rain"|

df$condition=="Rain" |

df$condition=="Heavy T-Storm"|

df$condition=="Thunder in the Vicinity"|

df$condition=="Thunder"|

df$condition=="Light Rain with Thunder"|

df$condition=="Thunder / Windy" |

df$condition=="T-Storm",0,1)

# Temperature should only make a difference when causes an inconvenience/discomfort

df$extreme\_temp= ifelse(df$temperature>86| df$temperature<21, 1,0)

# round tip to integer for hurdle model using ceiling to avoid rounding small tips to 0

df$tip\_amount\_int= ceiling(df$tip\_amount)

# Get hour from tpep\_pickup to control for hour differences

df$hour= substr(df$tpep\_pickup\_datetime, 12, 13)

# Make sure there are 24 unique hours for each one of the day

df$hour= as.factor(df$hour)

# Make pick up and drop off datetime instead of char

df$tpep\_dropoff\_datetime= as.POSIXct(df$tpep\_dropoff\_datetime)

df$tpep\_pickup\_datetime= as.POSIXct(df$tpep\_pickup\_datetime)

# Calculate trip time

df$trip\_time= difftime(df$tpep\_dropoff\_datetime,df$tpep\_pickup\_datetime, units = 'mins')

#trip could not realistically last less than 3 minutes

df %>% filter(trip\_time>=3)->df

# Correlations for important columns that are numeric

num\_cols= c('trip\_time','passenger\_count','trip\_distance','fare\_amount','tip\_amount','total\_amount','tourist\_pu','tourist\_do','entert\_pu','entert\_do','park\_pu', 'park\_do','workplace\_do','residential\_do','residential\_pu','workplace\_pu', 'rate\_fare','extreme\_temp','precip','wind speed','humidity')

num\_sample= df[,num\_cols]

# correlations for float or int columns including dummies

correlations=cor( num\_sample %>%

select\_if(~is.numeric(.)))

#visualize correlations

corrplot(correlations, method="circle")

#Store correlations for y variable vs independent

fare\_cor=correlations["fare\_amount",]

tip\_cot=correlations["tip\_amount",]

# Make improvement surcharge and congestion categorical

df$improvement\_surcharge= as.factor(df$improvement\_surcharge)

df$congestion\_surcharge= as.factor(df$congestion\_surcharge)

# Fix data types

df$month= as.factor(df$month)

#adding tip as a fraction of total fare(including the tip)

df$tip\_amount/df$total\_amount->df$tip\_fraction

#adding the variable which is =1 if there was a tip and 0 otherwise

ifelse(df$tip\_amount>0,1,0)->df$yes\_tip

#Build Fare model

#Summary stats

describe(df$fare\_amount)

#Distribution of fares

histogram(~fare\_amount, data = df, xlim= c(0,50))

histogram(~fare\_amount, data = df , xlim= c(0,200), breaks=500)

histogram(~log(fare\_amount), data = df)

# Look at correlations

fare\_cor

#(IF you leave default parameters, the graph is squished to one side)

# Look into any trends of fare prices over time

bwplot(~fare\_amount | month, data = df, xlim= c(0,50))

bwplot(~fare\_amount| year, data = df, xlim= c(0,50))

# See how covid shutdown affected fares

bwplot(~fare\_amount | covid, data = df, xlim= c(0,50))

rm(num\_sample)

# ols mode

ols\_fare= lm(fare\_amount~trip\_time+trip\_distance+borough\_pu+tourist\_pu+entert\_pu+covid+extreme\_temp+good\_condition+hour, data= df)

ols\_fare\_log= lm(log(fare\_amount)~trip\_time+trip\_distance+borough\_pu+tourist\_pu+entert\_pu+covid+extreme\_temp+good\_condition+hour, data= df)

summary(ols\_fare)

stargazer(ols\_fare,ols\_fare\_log,out="stargazer\_taxi\_fare\_ols.html",single.row = T)

rm(ols\_fare,ols\_fare\_log)

#Assumptions

#Multicollinearity

vif(ols\_fare)

#Independence

dwtest(ols\_fare)

#Build Tips models

# Look at tips distribution

histogram(~tip\_amount, data = df)

histogram(~log(tip\_amount), data= df,breaks=30)

#Look at correlations

#tip\_cor

# Look into any trends of tips over time

histogram(~tip\_amount | month, data = df)

histogram(~tip\_amount| year, data = df)

# See how covid affected tips

histogram(~tip\_amount | covid, data = df)

histogram(~log(tip\_amount) | covid, data = df)

#boxplot visualizations

bwplot(~tip\_amount | month, data= df, xlim = c(0,50))

#bwplot(~log(tip\_amount) | month, data= df, xlim = c(0,50))

bwplot(~tip\_amount| year, data= df, xlim = c(0,50))

#bwplot(~log(tip\_amount)| year, data= df, xlim = c(0,50))

df %>%

glm(tip\_amount\_int~ passenger\_count +borough\_pu +tourist\_pu +entert\_pu+workplace\_pu +

residential\_pu+ median\_rlst\_pu+ borough\_do +tourist\_do+ entert\_do +

workplace\_do +residential\_do+ median\_rlst\_do +fare\_amount+month+ hour+

extreme\_temp+good\_condition+covid, family=poisson (link=log), data=.)->poisson1

summary(poisson1)

stargazer(poisson1,out="stargazer\_taxi\_tip\_Poisson\_only.html",single.row = T)

df %>%

glm(tip\_amount\_int~ passenger\_count +borough\_pu +tourist\_pu +entert\_pu+workplace\_pu +

residential\_pu+ median\_rlst\_pu+ borough\_do +tourist\_do+ entert\_do +

workplace\_do +residential\_do+ median\_rlst\_do +fare\_amount+month+ hour+

extreme\_temp+good\_condition+covid, family=quasipoisson (link=log), data=.)->quasipoisson1

summary(quasipoisson1)

stargazer(quasipoisson1,out="stargazer\_taxi\_tip\_QuasiPoisson\_only.html",single.row = T)

vif(quasipoisson1)

dwtest(quasipoisson1)

fwrite(df,"taxi\_subsample.csv")

1. Found at <https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page> [↑](#footnote-ref-1)