Predicting New York City Taxi Tip Margins

Randy Ngo[1], Hector San Andres Izquierdo[1], Jennifer Park[1], and Angela Lin[1]

[1]Arizona State University, Tempe, AZ 85281, USA

April 29, 2025

Abstract

This study analyzes annual tipping trends in New York City Yellow Taxi rides from 2009 to 2024. Tipping remains a significant income source for taxi drivers, which makes understanding the factors influencing tip amounts crucial. Using a 1% stratified sample of Yellow Taxi Trip Records from 2009, 2014, 2019, and 2024, the data was cleaned, preprocessed, and analyzed through inferential statistics and machine learning models. Exploratory data analysis indicates that the average tip amount along with the percentage has been on the rise. Other findings suggest that the tip percentage surpasses 20% in 2024, up from under 6% in 2009. A one-way ANOVA confirms the differences in tipping behavior over the past decade to be statistically significant.

Modeling to predict tip amount was done using two machine learning models, a Random Forest Regressor and an Artificial Neural Network. The Random Forest Regressor appeared to be the best performing model for all individual years, obtaining an R^2 score of 0.75 in 2024. The feature importance analysis showed that time components, hour of day, day of week, and month, were much less significant than expected. The Artificial Neural Network performed worse consistently with lower accuracy, achieving a peak R^2 value of 0.68.

Through heatmap analyses and aggregating taxi trips in each borough, the geospatial assessment highlighted that Manhattan and Queens had the highest number of trips in addition to the best accuracy in trip tips. In contrast, regions such as Staten Island showed worse model performance, likely due to having less ride records and more variability. During analysis, the data gaps, file corruption, and missing tips for cash payments posed some challenges. In order to maintain rigor and reproducibility for the results, we focused on credit card payments.

1 Introduction and Background

1.1 Introduction

Tipping culture has been widely accepted in many countries around the world, appearing on prompt screens prior to transactions and even on headrest tablets in Uber rides. In the United States, tipping appears as an economic importance that encourages individuals to voluntarily pay tips without any obligation, yet it is often considered very rude not to [1]. Not all cultures expect tips, and some even view it as a disrespectful gesture [2].

Tips are prevalent in various industries, and the transportation sector is no exception. Customer satisfaction is one of highest contributing factors to tip amount received, but other factors can also influence tip behavior.

Taxi cabs have been a key part of the urban transportation industry for decades, providing ride services across the country. In particular, yellow taxis are an iconic part of New York City, the most populous and dense urban region in the United States. NYC employs a substantial amount of America's taxi cabs, which play a major role in meeting the city's travel demands throughout the boroughs [3]. As a result of tipping culture in the United States, taxi drivers' wages rely not only on fares, but heavily on tips, which makes up a substantial amount of their income.

1.2 Background

In the nineteenth century, Americans relied on railways and horses as their primary source of transportation. In the following century, streetcars were popularized and dominated the transportation industry [4]. These shifts in transportation coincided with rapid urbanization and expansion. Currently, more than half of the world's population resides in urban areas [5]. Many urban communities have adopted a basic land-use model, primarily focusing on the spatial distribution of housing instead of a mixed-use development. This has led to more residents in urban areas being car dependent [4]. However, in densely populated cities like New York, owning a car can be impractical due to limited space and heavy traffic congestion. New York City has been able to adapt by investing in alternative transportation such as public transit and taxi services.

Taxis have been a staple of New York City and have rooted themselves as part of the city's culture ever since their introduction in 1907 [6]. More than 400 thousand trips are made by yellow taxis alone per day [7]. Given the high volume of taxi trips, an abundant amount of data is available [5]. This data will be used to conduct our research on tip amount and its influential factors among trips. Some factors are as expected, such as interaction between the driver and passenger [8]. Passenger group size, adverse conditions, and perceived service quality all play a role in tipping as well [9]. Prior research has shown that even weather plays a part in taxi tipping, as days with sunlight were shown to increase tipping percentage by about 0.5 points in NYC [10, 11].

Today, the world is split in terms of tipping, where it is the expectation for consumers to provide a gratuity, and in others, not allowed entirely. In the Unites States, tipping has become more commonplace, with average tip amounts increasing over the years and becoming expected on every transaction. Many establishments have precalculated 15%, 20%, or 25% tips included in the payment process[1]. A study comparing different factors on tipping in restaurants in different countries finds that the United States has the highest prevalence of tipping among the 17 countries analyzed. The models insinuate consumers in the United States improve their self-image and status, feel in control of the interactions with the servers, and derive pleasure and freedom when tipping [2]. Furthermore, a survey conducted on customers found that many individuals prefer tipping as it gives them control over what they pay for provided services, as well as the opportunity to express their gratitude [1]. The commonality of control within an interaction can be observed from the results of the studies. Consumers often find a sense of satisfaction in being able to evaluate and reward a service provider.

Following the COVID-19 pandemic in 2020, many industries experienced major changes in daily interactions and economic behaviors. One of the most noticeable changes was the tipping culture for taxis in the United States. A study conducted in Chicago, measuring the amounts of tip before and after COVID-19, found that people left a significantly higher amount, but the number of people who tipped decreased [12].

1.3 Methodologies in Literature

Machine learning classification and regression has been used in relation to New York City taxi tips, although only using data from one year [13]. One approach by Kim, et al. (2020) [14] uses deep learning and linear regression to forecast demand for yellow taxis and for hire vehicles (FHVs). The study uses New York City's open databases and combines it with weather data to account for external factors that affect demand. Yang and Gonzales (2014) [15] use linear regression models to analyze pickup and drop-off locations to understand the socioeconomic and demographic factors that influence the variation of taxi demand. This study, similar to the previous one mentioned, combines multiple datasets to understand their relation to transit and taxi services. Research by Liao, et al. (2018) [16] also utilizes deep learning neural networks for travel demand prediction. The paper however, is careful in its utilization. Liao, et al. explains that to build an effective model, you need to take into consideration the domain knowledge and model architecture. Similarly to Kim et al., the study builds the model with other deep learning methods, such as Long-Term Short-Term memory to improve its efficiency.

Studies have different approaches to model tipping behavior. Previous research focuses mainly on the behavioral and social conditions that affect tip percentages. However, economic models are also used. Studies focused on economic models suggest that tipping behavior is largely explained by the quality of service [17].

Analysis of tipping behavior also indicates that, although tipping behavior varies depending on the customer, large percentages are rare. The research carried out by Aydin and Acun (2019) [17] found that the distance traveled has a positive correlation with the amount of tip the customer gives. It also collected data on behavioral aspects. They found that seat taken, driver interaction, and driver and clothing presentation impact the amount of tip. Multiple regression analysis with variance analysis modeled the relationship between the different factors and the target variable of tip amount.

Statistical analysis also proved to be viable when looking for tipping trends [9]. Elliott et al. (2017) [18] divides tipping behavior into two groups: tippers and stiffers. Employing mainly descriptive statistics and exploratory analysis, they found that the tipping amount falls mainly into 20%, 25%, and 30% of the total fare amount. These percentages are suggested when paying the fare. They also found that there is a correlation with income amount among the stiffers, or people that leave 0 tip. The study focuses only on yellow taxi cab data and not green taxi cab data, as the latter does not go to Lower Manhattan or south of West 110th Street and East 96th Street, and is used less frequently.

1.4 Project Plan

With such an unstable surrounding environment, it is essential to optimize the income and tip rate for drivers. Our aim is to locate the areas and times of day where drivers make the most money, considering pick-up and drop-off locations as an influential factor.

Since we have access to several years worth of data, we can draw samples from a range of years from the taxi data in order to observe trends over time, such as tip amount, trip durations, and trip volume. Utilizing these variables, as well as location information, we can hypothesize how tipping behavior changes in accordance to each factor, in addition to the introduction of ride-sharing services to NYC and the COVID-19 pandemic.

Supervised machine learning algorithms, Random Forest and an Artificial Neural Network allow us to fully utilize the data we have available. Random Forests have the additional benefit of measuring feature importance, which will be useful for determining the most critical factors in tipping.

2 Methods

2.1 Methods: Data Sources, Cleaning, and Preparation

Our primary dataset, New York City Yellow Taxi Trip Record Data, has numerous features related to each taxi trip, providing insight into factors that could influence tipping.

Dataset Fields Overview:

- `VendorID`: Identifier for the taxi record provider.
- `tpep_pickup_datetime`: Timestamp when the trip started.
- `tpep_dropoff_datetime`: Timestamp when the trip ended.
- `passenger_count`: Number of passengers.
- `trip_distance`: Distance traveled in miles.
- `RatecodeID`: Rate code applied to trip.
- `store_and_fwd_flag`: Indicates whether or not the trip record was stored and forwarded due to connectivity issues (Y/N).
- `PULocationID`: Pick-up location ID referencing NYC taxi zones.
- `DOLocationID`: Drop-off location ID referencing NYC taxi zones.

- `payment_type`: Method of payment for the fare of the trip (range 1-6).
- `fare_amount`: Base fare for the trip.
- `extra`: Additional charges.
- `mta_tax`: Tax imposed by the Metropolitan Transportation Authority.
- `tip_amount`: Gratuity provided by the passenger.
- `tolls_amount`: Toll charges incurred during the trip.
- `improvement_surcharge`: Additional surcharge.
- `congestion_surcharge`: Additional charge for congestion during trip.
- `total_amount`: Total cost of the trip.
- `airport_fee`: Additional fee for trips to the airport.

The Taxi and Limousine Commission provides a data dictionary explaining what each variable represents. `Payment_type` ranges from 1 to 6 where 1=credit card, 2=cash, 3=no charge, 4=dispute, 5=unknown, and 6=voided trips. `RatecodeID` ranges from 1-6(1=Standard rate, 2=JFK, 3=Newark, 4=Nassau or Westchester, 5=Negotiated fare, 6=Group ride). `VendorID` is one of two companies: 1=Creative Mobile Technologies LLC or 2=VeriFone Inc.

Each location ID is represented by a number where a comprehensive table of their located borough, taxi zone, and service zone is provided on the website [19]. Being an older set of data, the 2009 Yellow Taxi Data does not have `LocationID`, but rather variables for longitude and latitude coordinates for pick-up and drop-off locations. The airport fee variable is a more recent addition to the information provided, so it is only present in the 2024 dataset.

Conveniently, every column is already encoded, however we shifted over to one-hot encoding from simple label encoding for some features. For example, `LocationID` values range from 1 to 265, so to prevent our models from developing a bias toward larger values, we grouped each location into boroughs and then one-hot encoded the boroughs.

The original dataset has the timestamps for pick-up & drop-off locations in one column each, so we split them into multiple columns as new categorical variables: the time of day (`hour`), the day of the week (`day`), the month (`month`), and year (`year`). We performed feature engineering by creating a new column for the trip duration based on these values.

The variable `total_amount` was dropped from all datasets because it includes the customer's tip and would provide redundant information. `store_and_fwd_flag` was dropped as it does not provide relevant information for predicting tip amount. Because there were a considerable amount of null values, `congestion_surcharge` was removed from the 2009 and 2014 datasets and `RatecodeID` was removed from 2009. Additionally, `airport_fee` was removed from 2024 as it was newly implemented in 2022 and was not likely to provide much insight for trends over time.

During our exploratory data analysis (EDA), we found multiple outliers, like negative tip values or negative trip lengths which we dropped as it would interfere with final analyses and model training. The remaining NaN values were repopulated with the most common value in the respective column as there were few of them compared to the sizes of the datasets.

The dataset is split up into separate parquet files for every month of each year from 2009 to 2024. Observations in each year range from 25 million to over 170 million. Due to the sheer amount of data we have, it is impractical to use all of it for analysis and to train models. Instead, we will take a look at four specific years: 2009, 2014, 2019, and 2024. Increments of five years allow us to observe long-term trends in key time frames, such as before & after the introduction of ride-sharing services in 2011, and the impact of the COVID-19 pandemic in 2020. However, this approach still presents us with more than 400 million observations, so in addition, we will consider only credit card payment types and take 1% subsets from each

of the yearly datasets. Tip amounts for cash payment types are not accounted for, so they are not suitable for this project. The sizes of our new datasets range from 400 thousand to 1.5 million.

We decided upon a 90/10 train/test split for training our models as we have an ample amount of data to work with. We decided to clean the data before splitting, in order to preserve the distribution as much as possible.

2.2 Methods: Inferential Statistics

Inferential statistics are essential in determining trends and patterns within data and in this case, the tip disparity over the past decade and more. In the exploratory data analysis, averages of tip amounts and percentages were calculated for each five year interval. Utilizing Python packages such as `matplotlib`, the values were plotted by variables like hour of the day and month of the year in line graphs, and average tip percentages over the years by day of the week and season in bar graphs. Also, a one-way Analysis of Variance (ANOVA) test was conducted to assess the statistical significance of different means within the tip amounts over the years. To perform this test, the `f_oneway` method from the `scipy.stats` package was used.

2.3 Method: Supervised Machine Learning

We will only be doing supervised learning on our dataset, as the dataset is already properly labeled, so there is no need for any unsupervised learning. Our models of choice are a random forest regressor and an artificial neural network to compare and contrast results.

2.4 Method: Random Forest Regressor

We decided to use a random forest regressor due to its strength against overfitting and its ability to use both categorical and numerical features, handle nonlinear data, and its proficiency when used with very large datasets. Another benefit of random forests is measuring feature importance, which gives us insight into what the most relevant factors are to tipping. Another important characteristic is that it can handle outliers and missing values. We found in our data exploration that, since some of the data is driver-entered, outliers are present in the samples. The hyperparameters were chosen using Randomized Search Cross Validation. Due to the size of our dataset, RandomizedSearchCV was more efficient and less exhausting than grid search cross validation.

2.5 Method: Artificial Neural Network

We used a basic artificial neural network to perform a regression on our data for the sake of comparison. A deep feed-forward neural network (or Multi-Layer Perceptron) provides a way to learn nonlinear relationships in data, potentially resulting in a better fit with the data. Neural networks can be very robust with their hyperparameter tuning as well, with the number of neurons and number of layers acting as levers to use when tuning.

3 Results

3.1 Exploratory Data Analysis

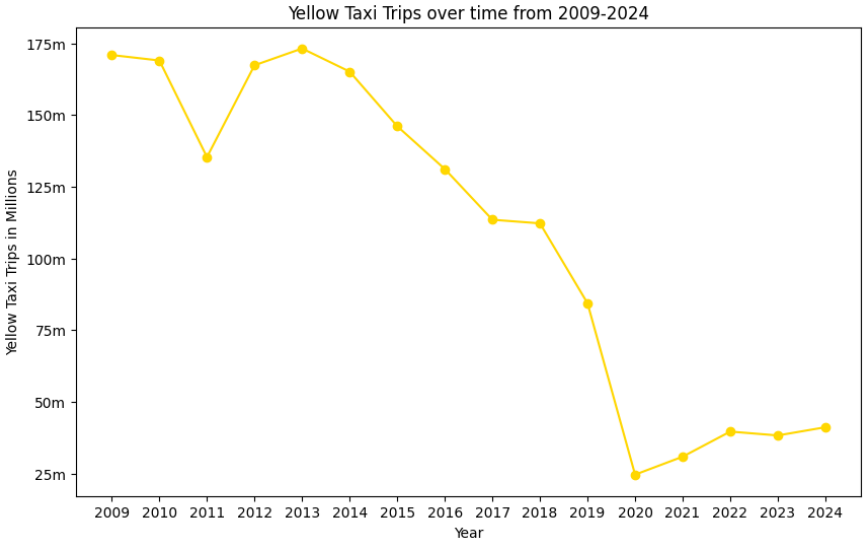

Figure 1: Line graph showing taxi trips made each year from 2009 to 2024

Looking at this trend line, there is a dip in ridership in 2011, which coincides with the launch of Uber in NYC in 2011. Ridership bounced back the year after, however from 2014 onward there was a steady decline until 2020, when the COVID-19 pandemic caused a steep drop in ridership. In the years following the COVID-19 pandemic, ridership has steadily gone up, but remains a fraction of what it was before.

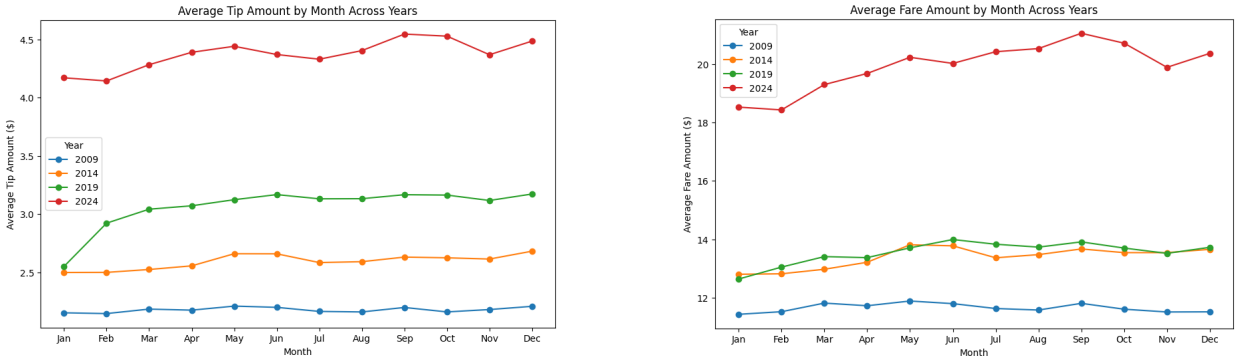

Figure 2: Line graphs showing the average tip amount and average fare amount by month across years

Figure 2 compares the average tip amounts and the average fare amount throughout the years. Although the graphs are not separated by each year, each line is noticeably distinct from one another. In the Average Tip Amount by Month Across Years graph, average tips in 2009 consistently ranged below $1.00. 2014 averaged in the $2.50 range. 2019 averaged about $3.00 with exception to January, which was around $2.50. 2024, the highest, reached averages in the $4.00 to $4.50 range.

Average fare amounts throughout the years also follow a similar pattern. Fares in 2009 are distinctly the lowest, averaging about $11. 2014 and 2019 are close in value with 2019 being slightly higher on average

in the $13 to $14 range. 2024 was about $6.00 more expensive on average than 2019. Additionally, there is a notable $1 to $2 increase in average fare amounts in the beginning of 2024 from January to May. This increase persists as the averages steadily remain at the risen levels, with one dip in November, before increasing again in December.

When comparing the two graphs side-by-side, the difference between average tips is larger than the difference between the average fare amount in 2014 and 2019. This could be a result of changes within tipping culture in the United States as it is more expected of consumers to tip and to tip more for services provided to them [20]. Additionally, the lines in both graphs remain distinctly separate and follow the same order, average amounts increasing as time has progressed. This may be correlated with the significant inflation that has occurred throughout the last decade, making many things more expensive than before, including services such as taxi rides.

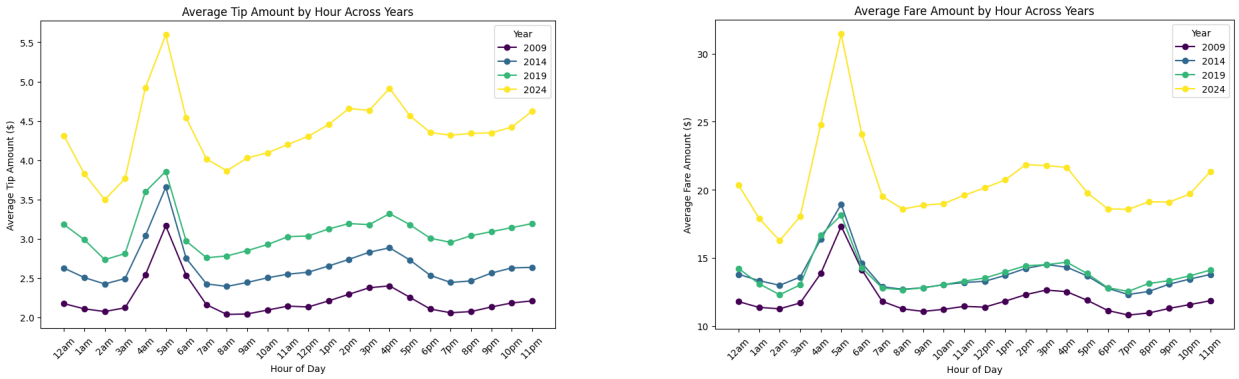

Figure 3: Line graphs showing the average tip amount by hour and average fare amount by hour across years

Figure 3 depicts average tip and fare amounts through each hour of the day. Both graphs, once again, have distinctly separated averages between the years with the exception of fare amount averages in 2014 and 2019 which intersect throughout the day. The average tips and fares spike significantly at 5 am and slightly increase towards 4 pm. This could be indicative of more frequent or further trips at these hours, or better paying customers in general. A potential cause could rise from airport rides as early morning flights are common for airline travel hours.

Between the two graphs, average tips have been increasing at a more even and steady rate over the years with averages increasing roughly a dollar every five years. In comparison, average fare amounts in 2014 and 2019 are incredibly close, overlapping each other. There is a significant five dollar increase in average fare amount between 2019 and 2024 compared to the two dollar increase between 2009 and 2014. There is a steady incline of trip costs during the the afternoon hours, but it is not a significant improvement in tips during this hour, though the most recent 2024 set shows more variation.

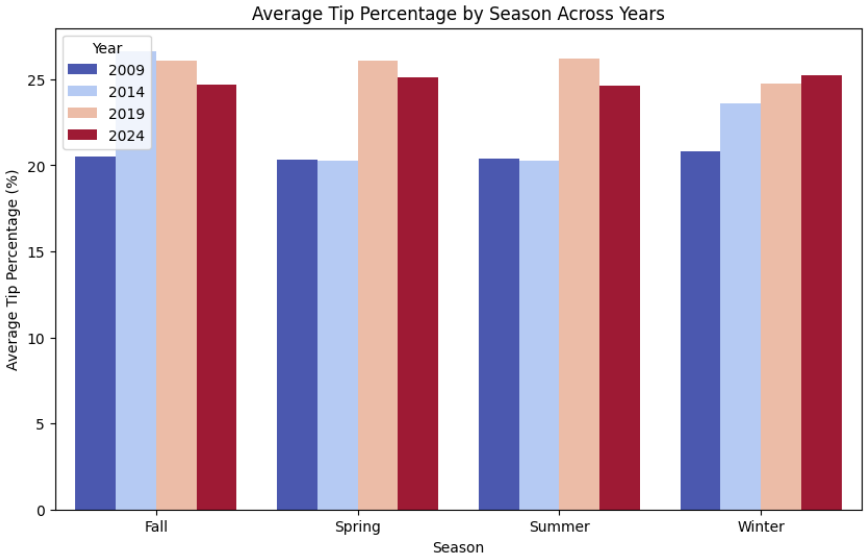

Figure 4: Average tip percentage by season

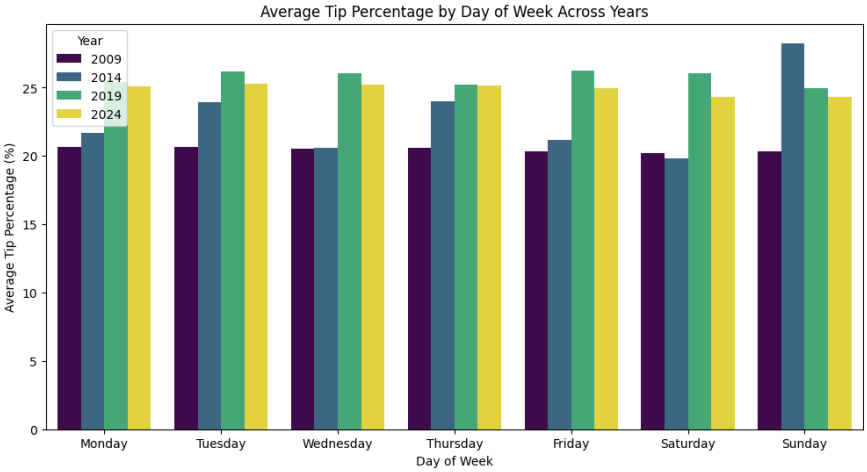

Figure 5: Average tip percentage by day of the week

Figures 4 & 5 explore the relationships between average tip percentage by season and days of week across the years. One of the most noticeable trends is the average percentages from 2014 to 2019 have widely increased. One exception is in the 2014 Sunday averages which could be due to outliers that are observed in the following heatmaps, as there are no other exceptions such as this in these graphs. This suggests gradual improvement in tipping culture which can be influenced by a variety of factors. This does not always remain consistent average tip percentages in 2024 are often lower than 2019.

Additionally, the data suggests that tip percentages can fluctuate depending on the day of the week, with certain days having higher tips than others. Overall, weekdays have slightly higher tip percentages than the weekends.

The data also reveals seasonal impact on tipping behavior. Tip percentages tend to peak during certain times of the year, particularly in the winter months, which coincides with the holiday season, when people tend to be more gratuitous, but is not always the case as the 2019 averages are higher in the other seasons.

The general percentage of tips vary greatly during the decade, as there is a sharp increase in tip percentages from 2009 and 2014 to 2019. As fare prices have steadily increased over the years, tipping percentages have also risen. The average tip percentage in 2009 hovered around 20% which is similar in 2014, averaging 25% in 2019, and hovering in the 20%-25% in 2024. This suggests that customers are tipping at higher rates, allowing the dollar amount received by drivers to increase.

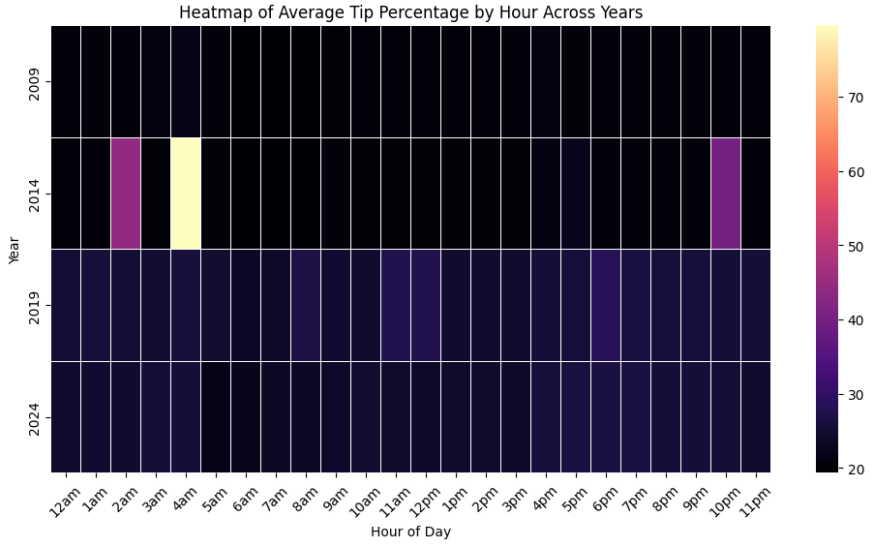

Figure 6: Average tip percentage by hour

Figure 6 shows fairly consistent trends all throughout 2009 around 20%, however in 2014 there are notable peaks at 2 am, 4 am, and 10 pm. This could indicate outliers during these hours with 2 am and 10 pm averaging in the 50% range, and 4 am in the 80% range. Otherwise, similarly to 2009, average tips remain in the 20% range. In 2019 and 2024, we can see a shift in tipping behavior as averages begin to vary and increase to the 30% to 40% range.

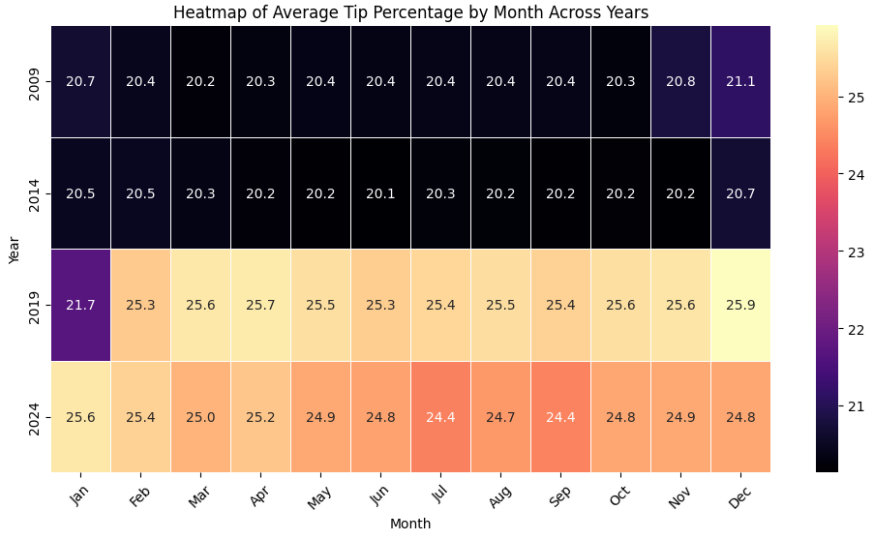

Figure 7: Average tip percentage by month

Figure 7 provides visual representation of how tipping behavior has changed over the years, with emphasis on how tip percentages fluctuate across different times of the year. In 2009 and 2014, tip percentages were consistently low, staying below 21% throughout the year. By 2019, there was a visible increase, with tipping percentages ranging between 25% and 26%. The biggest shift occurring between January 2019 to February 2019, where the percentages jumped from 21% into the 25% range and stabilized in 2024.

In 2014, 2019, and 2024, October, November, and December displayed the highest tip percentages of the year. In particular, December tips spiked significantly more in comparison to the rest of the year, with 2014 reaching 20.7%, 2019 at 25.9%, and 2024 peaking at 24.8%. The tipping percentages tend to stabilize from April to August and do not show dramatic increases or decreases.

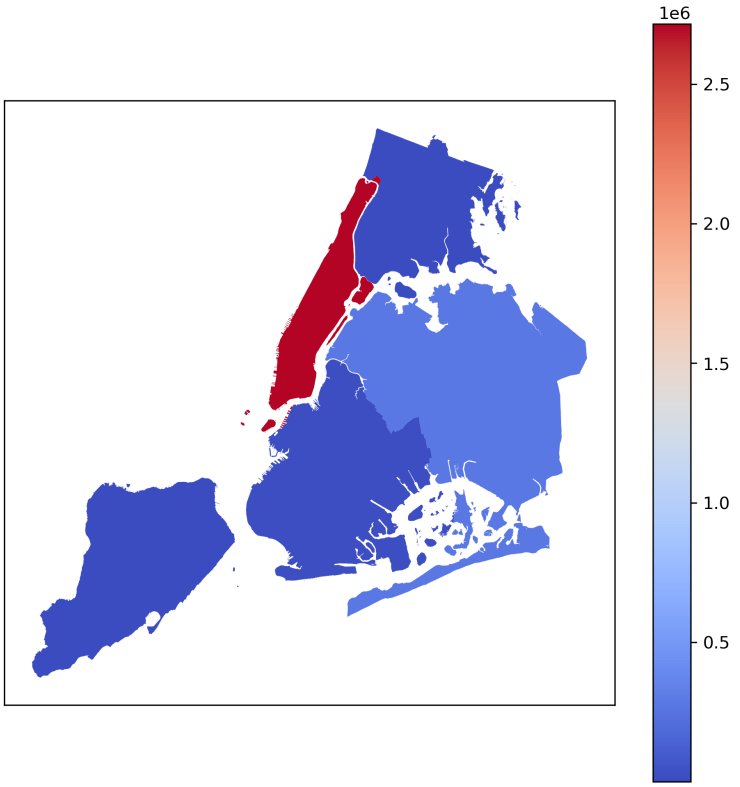

Figure 8: Total drop-offs heatmap by borough. Descending from top: Bronx, Manhattan, Queens, Brooklyn, Staten Island

Figure 8 is a heatmap showing which boroughs have the most number of drop-offs. Manhattan is the most popular borough by a wide margin, then followed by Queens, Brooklyn, the Bronx, and Staten Island. This coincides with research that found many taxi rides are concentrated in the more populated areas of the state [20]. Although Manhattan is one of the smallest boroughs, it is one of the more densely-populated boroughs. With such a dense population in a constricted space, it is easier to find riders because the streets are fuller, creating plentiful opportunities. In comparison, larger and less densely-populated areas are less likely to find drivers a rider as quickly, and destinations not as likely to be close.

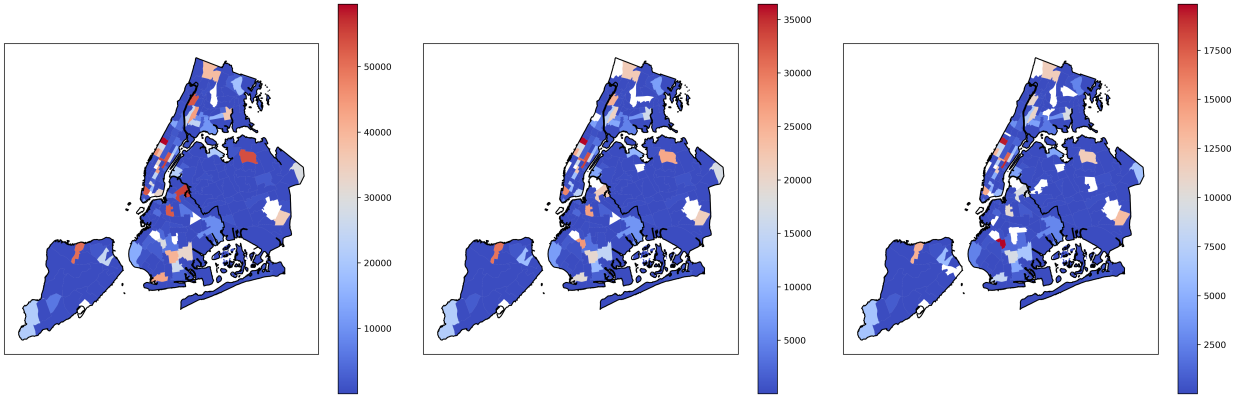

Figure 9: Heatmaps of pick-up zone frequencies in 2014, 2019, and 2024, respectively. 2009 was dropped due to the dataset storing location values differently

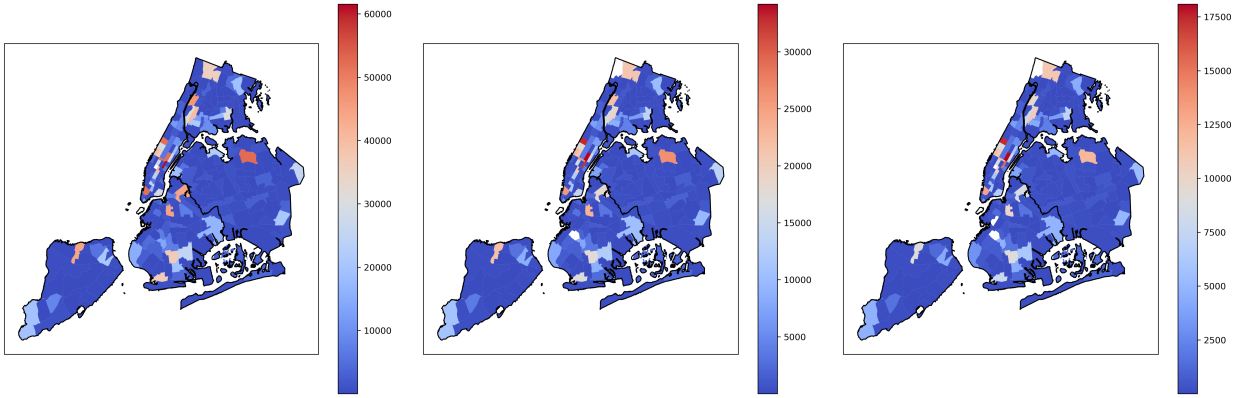

Figure 10: Heatmaps of drop-off zone frequencies in 2014, 2019, and 2024, respectively

The heatmaps for the pick-up and drop-off locations are very similar with small variations. The biggest trend we can observe is that taxi trips dropped by approximately half each time. For the most part, the heatmaps are largely the same when accounting for sample size, although there are a few zones in Manhattan that saw greater frequency as time went on, while there were zones in Queens and Brooklyn that saw decreased frequency during the same time period. The zones with the highest number of pick-ups are in Manhattan and Brooklyn. Brooklyn has the highest population and Manhattan has the highest GDP. In both pick-up and drop-off, Manhattan has the highest variation. In the other boroughs, the locations are concentrated in one or two zones. Zone 133 has a notable decrease in frequency of more than 5000.

3.2 Descriptive Statistics

year	count	mean	std	min	0.25	0.50	0.75
2009	1552914	5.361647	15.893624	0	0	0.000000	6.060606
2014	1654123	12.296494	430.205705	0	0	11.764706	21.212121
2019	843984	18.901476	351.256764	0	0	22.000000	28.250000
2024	404146	19.506719	173.533294	0	0	23.238866	28.602151

The descriptive statistics show results similar to figure 2. The year 2009 had the lowest tip amount. The tip is still 0 until the second quartile. 2014 saw the greatest variance in the amount of tip. Between 2019 and 2024 there is a small increase in the mean and quartiles but a lower variation in the data.

A correlation analysis was also conducted to test the assumption that the tip increases with fare amount, i.e. the two have a positive linear relationship. The correlation coefficient is 0.29, showing a small correlation. A similar analysis was conducted between the trip distance and tip amount. The coefficient is 0.47 showing a higher but still negligible correlation.

3.3 Inferential Statistics

From all of these charts, a common trend was observed: a general increase of tip amounts over the last 15 years. Within the line graphs comparing tip amounts, 2009 averages hovered in the 50 cent range, 2014 in the $1.50 range, 2019 at $2-$2.50, and 2024 $3-$3.50. In addition to the years having minimal variation, the lines for each year were distinctly separate with no intersection between the monthly average tips, increasing at each passing interval. Based on these results, there is reason to believe that time, year specifically in this case, would be an appropriate and effective measure in predicting the tip margin of a given ride.

For the ANOVA test, H0, the null hypothesis, assumes that all of the tip amount means are equal, whereas HA, the alternative hypothesis, implies that at least one of the means is different. For this test, an alpha = 0.05 is assumed.

F-statistic	P-Value	Significant
188228.92	0.0	Yes

Table 1: Results of ANOVA test F-statistics comparing tip amount in each year

In process concluded F-statistic: 188228.92, and p-Value: 0.0. Since the p-value, 0, is less than the alpha level, 0.05, the null hypothesis is rejected and there is statistical significance to assume the alternative hypothesis, at least one of the means for tip amount during each of the given years is different than the others. Although, it may be questionable for a p-value to be 0, similar trends can be observed in the EDA as well.

3.4 Correlation Analysis of Taxi Tip Margins

To identify the significance of each variable with respect to `tip_amount`, inferential statistics will be performed. We decided to create separate correlation matrices for our selected years to capture newly implemented variables throughout the decade. Encoded categorical variables were excluded in this analysis as it could potentially skew the results. Categorical variables `VendorID`, `PULocationID`, `DOLocationID`, and `RatecodeID` were dropped.

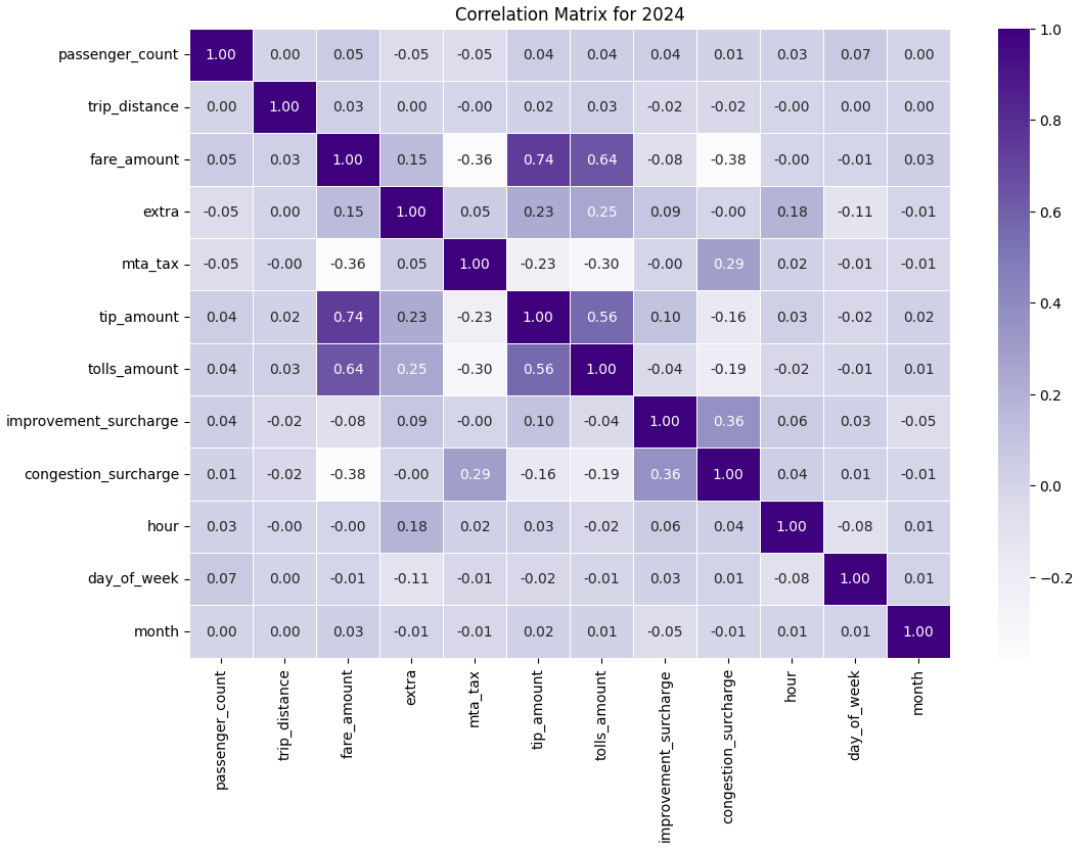

Figure 11: Correlation Matrix for 2024

In 2024, `tip_amount` was highly correlated to the `fare_amount` at 0.74 and `tolls_amount` at 0.56. The `tolls_amount` and `fare_amount` are correlated at 0.64. The remaining correlations are, overall, weak. `Tip_amount` and `trip_distance` are weakly correlated with `extra`, both sitting at 0.02. `Improvement_surcharge` and `congestion_surcharge` have a 0.36 correlation, `congestion_surcharge` and `mta_tax` have a 0.29. `Extra` and variables `hour` and `fare_amount` have correlations of 0.18 and 0.15, respectively. Other values have less, if not zero correlation between each other in the 2024 dataset.

Figure 12: Correlation Matrix for 2019

Within the 2019 dataset, `tip_amount` is highly correlated with `fare_amount` at 0.72 and `trip_distance` at 0.68. There is a moderate connection between `tip_amount` and `tolls_amount` at 0.55 and a weak connection with `passenger_count` at 0.01.

The highest correlation values, overall, are associated with `fare_amount`. There is a moderate correlation between it and three variables. With `tolls_amount` there is a 0.55 correlation, `tip_amount` as stated previously, and `trip_distance` has 0.68. Other correlations are moderately weak or lower: `extra` versus `congestion_surcharge`, `passenger_count`, and `hour` are 0.2, -0.17, and 0.13, respectively; `mta_tax` and `tolls_amount`, `fare_amount`, `congestion_surcharge`, `improvement_surcharge`, and `trip_distance` are -0.30, -0.32, 0.20, 0.08, and -0.15, respectively. `Congestion_surcharge` and `trip_distance` have a -0.18 correlation whereas the remaining values are lesser or equal to -0.10 or 0.10.

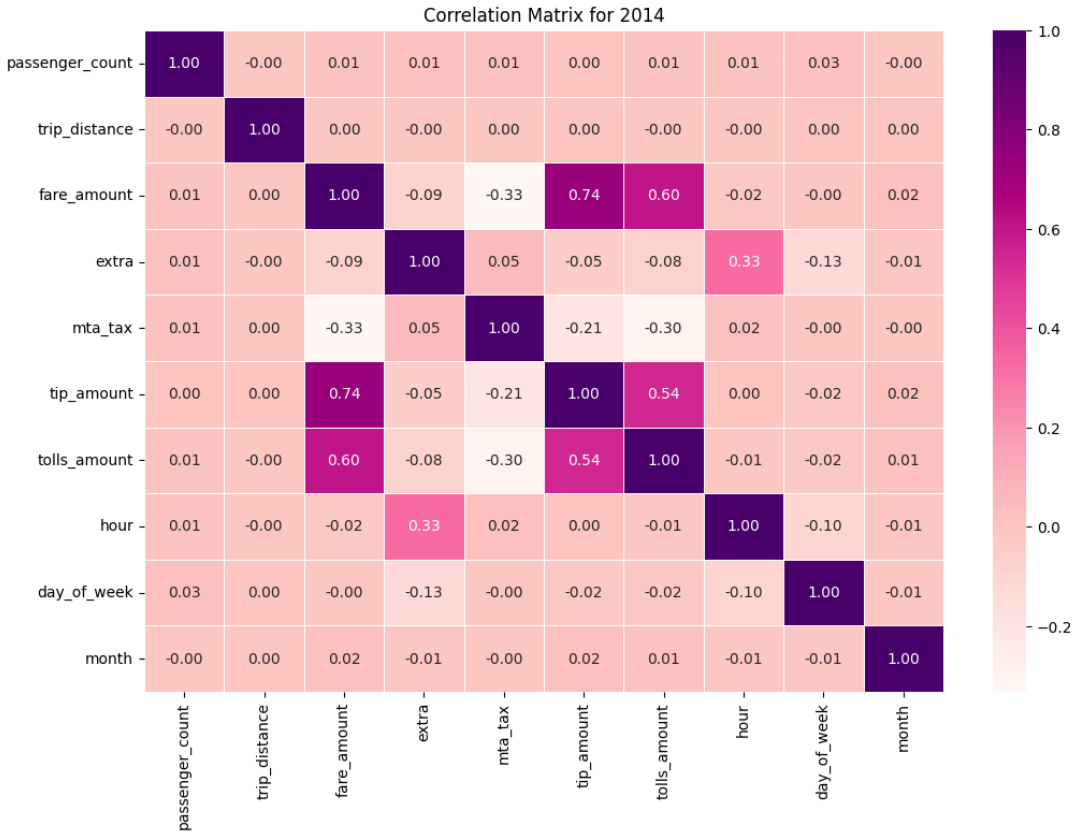

Figure 13: Correlation Matrix for 2014

In the correlation matrix for 2014 variables, `tip_amount` is highly correlated with `fare_amount` at 0.74; moderate with `tolls_amount` at 0.60; and weak with `mta_tax` at -0.21.

Two other correlations are moderate: `fare_amount` and `tolls_amount` at 0.60; and `extra` and `hour` at 0.33. Other pairings are weakly correlated or lower, of them, the highest are `fare_amount` and `mta_tax` at -0.33 and `tolls_amount` to `mta_tax` at -0.30.

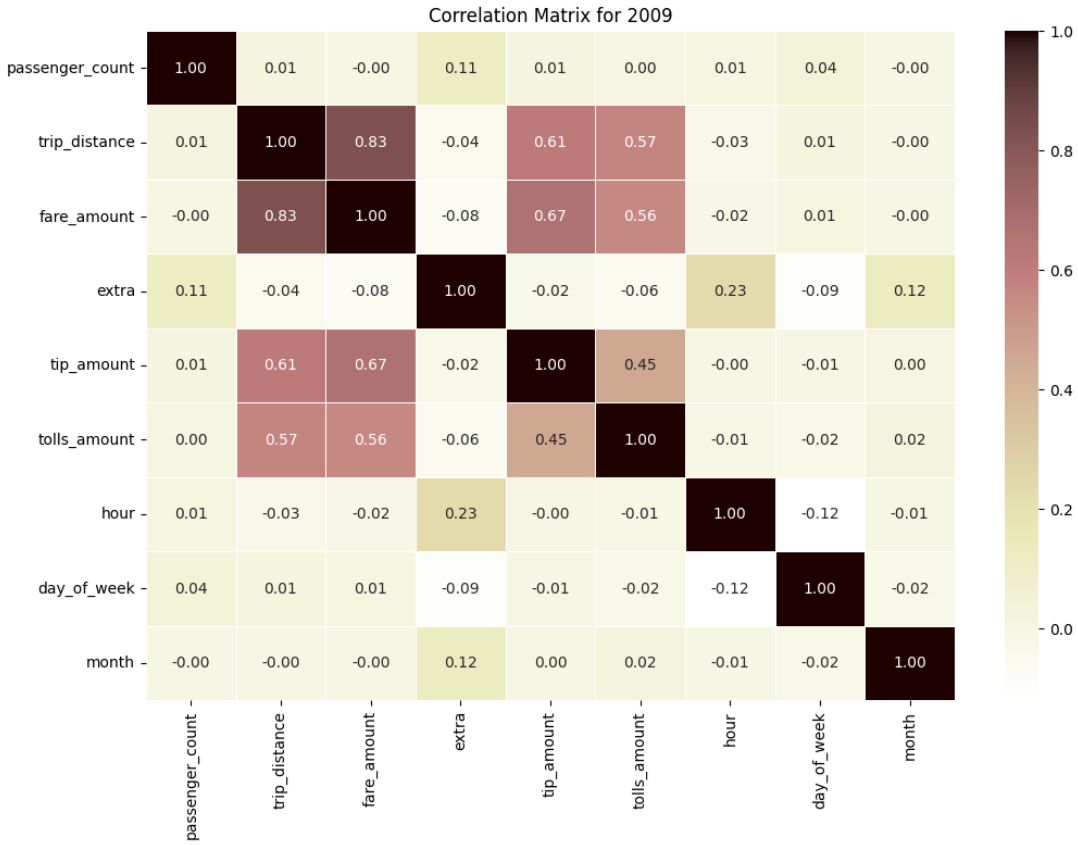

Figure 14: Correlation Matrix for 2009

Among the correlations to `tip_amount` in 2009, there are moderately weak connections to `passenger_count` and `extra` at 0.01 and -0.02 respectively.

There is a high correlation between `trip_distance` and `fare_amount`, the highest among any of the correlations in 2009, at 0.83. The correlations between `tip_amount` and `fare_amount` sits at 0.67, `tip_amount` and `trip_distance` at 0.61, and `tip_amount` and `tolls_amount` follows at 0.45 The remaining correlations are extremely low or zero.

Overall, there were some moderate correlations to the independent variable `tip_amount` in all of the datasets being observed. Often, those correlation values were the highest among the values in the entire year's matrix.

Correlations between dependent variables varied in the moderate or lower range, with the exception of two, one from the 2009 dataset and the other from the 2024 dataset; between `trip_distance` and `fare_amount` at 0.83 in 2009 and 0.88 in 2024. The two variables share the highest correlation between all other variables and datasets being observed. As both are dependent variables, there is potential for multicollinearity if both variables are included in a model. This runs the risk of adding redundant information to predictive models, thereby creating excess noise and throwing off efficacy of the methods used. Otherwise, correlations between dependent variables are not significantly large and are not at high risk for causing this issue. Our analyses helped eliminate variables of insignificance to our final analysis, maximizing tip margins for taxi drivers in the New York region.

3.5 Random Forest Regressor

Initially, we planned on doing Grid Search Cross Validation, however the computational cost was too expensive, so we settled on using Randomized Search Cross Validation. We tested max features, number of estimators, minimum sample leafs and minimum sample split. Our best model had 20 estimators, a minimum sample split of 4, a minimum sample leaf of 3, and a square root of total features. The hyperparameters help reduce overfitting, balance computational cost and accuracy, and control the architecture of the trees. We did a 90-10 train/test split on the combined dataset. Then, we examined how well the model performed for predicting each year and extracted the feature importance in predicting tip amount.

year	R^2
2009	0.7124
2014	0.7162
2019	0.7180
2024	0.7529
All	0.5932

Table 2: Random Forest performance for each year and combined

The model performed best for the 2024 year. Overall, it was consistent in 2009, 2014, 2024 with a small improvement each year. After we had the model, we wanted to look at feature importance to determine what affects the tip outcome in our model.

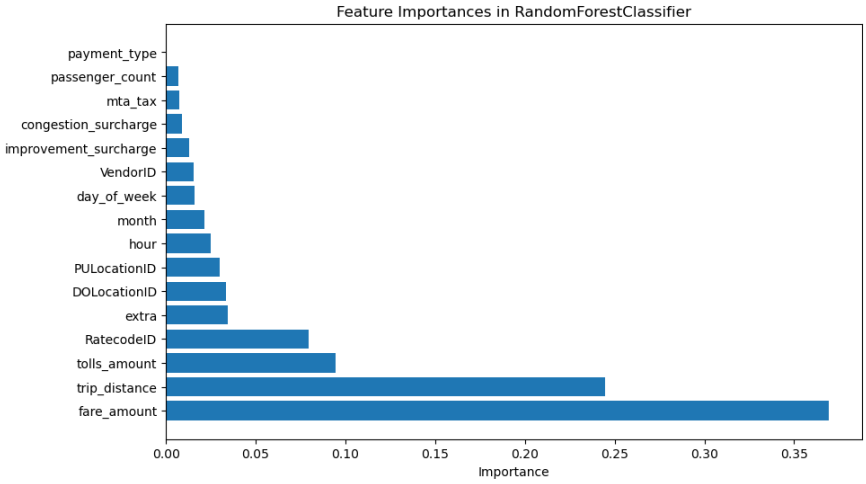

Figure 15: Feature importance

As anticipated, fare amount is the biggest indicator of tip amount, followed by trip distance, which is accurate to our prior assumptions. Contrary to our beliefs however, the timing of the trips play a relatively minor role in prediction. Then, we grouped each zone into the boroughs to look at the model performance throughout New York City. We used pick-up location and the test data for our model.

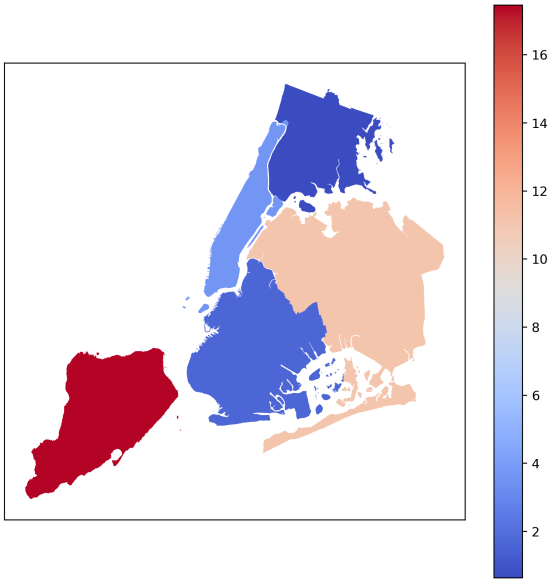

Figure 16: Predicted tip amount by model for each borough in 2024

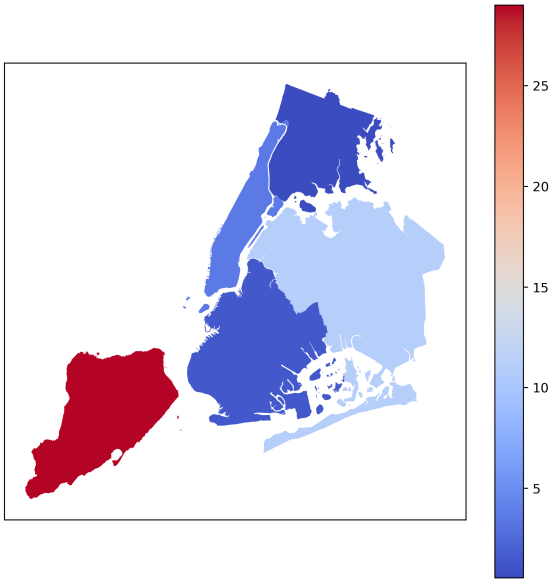

Figure 17: Actual tip amount by model for each borough in 2024

The model performed best in Manhattan and Queens. This aligns with our observations from the EDA when counting the number of trips in those areas. Staten Island had the biggest drop in accuracy. Based on these results, our model deployment could be used best for high-transit areas on a neighborhood or borough level analysis.

3.6 Artificial Neural Network

We tried several variations and modifications on the neural network, such as adding more layers, adjusting the number of neurons in each layer, adding regularization, changing the batch size, and changing the number of epochs, but ultimately the results were largely the same and the model never reached beyond an R^2 value of 0.68.

year	R^2
2009	0.6727
2014	0.6696
2019	0.6697
2024	0.6619
All	0.6429

Table 3: Neural Network performance for each year and combined

We settled on a training time of 20 epochs with 3 layers of 64 neurons, as those settings for the hyper-parameters resulted in the best and most consistent average results across the years.

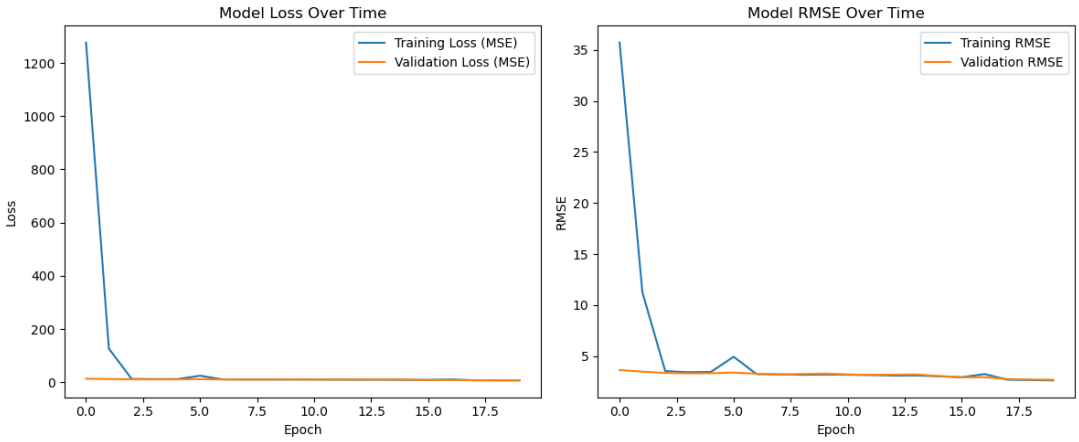

Figure 18: Training graphs for 2009

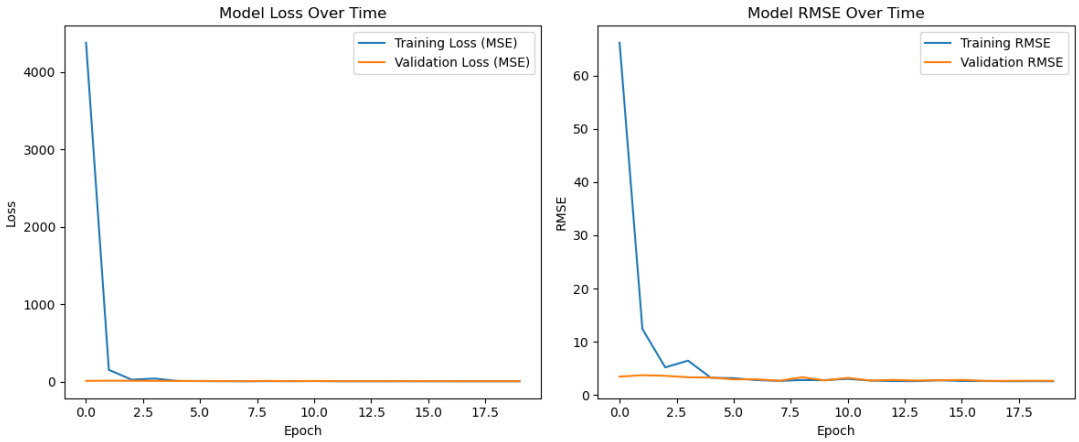

Figure 19: Training graphs for 2014

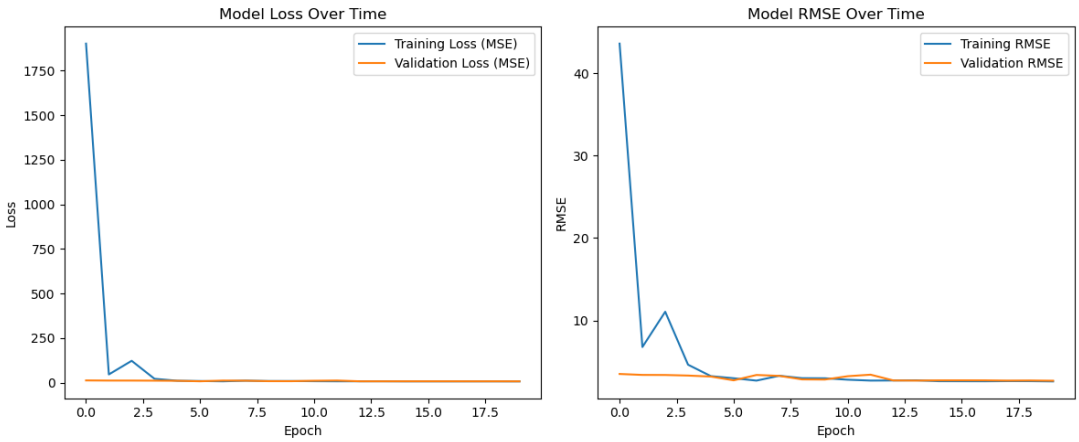

Figure 20: Training graphs for 2019

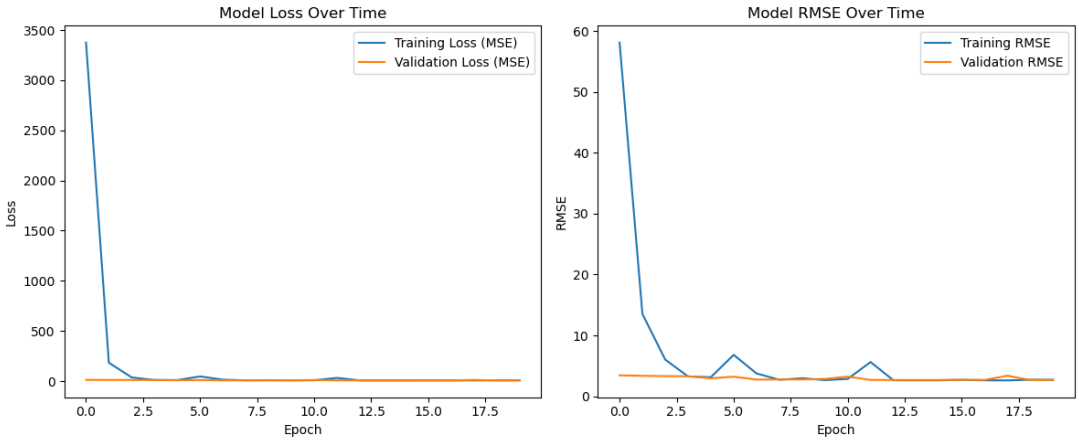

Figure 21: Training graphs for 2024

The training graphs in figures 8-11 show the loss and RMSE over time. Each of the graphs levels off after about 5 epochs, so we felt a training time of 20 epochs was appropriate to avoid overfitting. Interestingly, incorporating regularization elements into the model resulted in worse overall performance.

4 Implications and Discussion

4.1 Popular times & places for taxi trips

As we found in our EDA, Manhattan is the most popular borough for taxi trips to no surprise. Similarly, peak hours for taxi trips and tipping were also as expected, peaking at 5 a.m. and 5 p.m., which are before and after common working hours respectively. However, our Random Forest model ranked `hour`, `day of week`, `month`, and `year` all on the lower end of feature importance, leading us to believe that these factors are not as relevant to tipping as initially projected.

4.2 Tip amounts over time

Our EDA revealed that over time, tip amounts in addition to tip percentages have increased in the observed time frame of 2009 to 2024. This is supported by the results of our models, where the models trained on the whole dataset performed worse than the ones trained on individual years, due to the variation in tipping trends for each year.

4.3 Machine learning tip predictions

Both models we employed were moderately successful. Between the two, the Random Forest Regressor had consistently higher R^2 values and proved to be the more effective model for this problem. Due to the efficacy of our model, we conclude that machine learning is indeed a viable solution to predicting tip amounts.

4.4 Challenges

One of the main challenges to consider in our analysis is the integrity of the data. We found many outliers, negative values, and null values. The most inconsistencies were present in the payment type column, as cash payments entries did not record tip amount accurately. Additionally, there were limitations in handling the amount of data and determining how to represent results and questions through it. The New York City Taxi and Limousine Commission provides data on trips taken in yellow taxis, accounting for every single ride engaged in a registered medallion taxi.

As a result, the datasets are extremely large in size, making analysis unfeasible using all of the data. To combat this issue, the datasets for this project are taken from five-year intervals since the beginning of the posted data, 2009, to the most recently completed year, 2024. The goal is to still gather as comprehensive of an analysis as possible, while reducing the amount of observations being manipulated.

Despite reducing the datasets down to the years 2009, 2014, 2019, and 2024, there remains to be a substantial amount of data. In 2009 alone, the largest annual dataset, contained over 170 million rows, and in 2024, the smallest dataset, recorded over 20 million observed taxi rides in the entire year. In order to mitigate this issue, the datasets in this project consist of 1% subsets of the original dataset sizes within the five-year intervals. To ensure there is a proper representation of the original information, the subsets were randomly selected the full year's data. Since the sample size of the datasets is still significantly large, the belief is that there is enough information for analyses to be statically significant in due to the size.

Overall consistency in the datasets varied as variables were added, dropped, changed, etc. For example, airport fee became a new addition to the collected yellow taxi data in 2022, so it was only present in one of the datasets used in this project. Since there is lack of information to work with, the variable was dropped from the working dataset. A similar case appeared with the improvement surcharges in the earlier years, many of the observations having mainly, if not all, null values.

To deal with null and unreasonable values, a combination of actions were taken. Given there is no reliable way to determine the cause behind unrealistic values and their actual values, the rows containing such values were dropped entirely to prevent them from skewing the machine learning models in negative and unproductive ways. The majority of the dropped rows contained negative values for certain variables and did not include outliers in the data. For null values, the observations were repopulated with the most common value in the respective variable.

To maintain consistency in the data, all payment methods other than credit card transactions were dropped. Since the dataset contained information on cash tips that could not be accounted for and the majority of the cash payments had tips of zero, it was decided that using the most reliable and traceable payment type, credit card payments, would be the most beneficial for any future models.

When testing different machine learning methods, there were issues with performance time and computational complexity. Even with the reduced 1% subsets of the original data, the cost of running the models was substantial. An example of this occurred when running the grid search method, none of the devices were able to allocate enough memory to produce any sort of results. Thus, the chosen alternative was the randomized search method due to its efficiency, making hyperparameter tuning more convenient and attainable.

4.5 Next Steps

More statistical approaches to the data would be a beneficial for improving this project. More random sampling or different methods of sampling could ensure a more unbiased working dataset. To get more accurate results of the tips over time, further feature engineering would help optimize the predictions and reduce overfitting in the machine learning models. To truly encompass all of the data, one would need to utilize a device with strong processing capacity to perform models more efficiently, or combine data from all of the years.

Another step to get a more comprehensive approach of the factors that influence tipping amount could be to incorporate other datasets. Weather data can be used to see how snowy and rainy days impact customer tipping behavior.

5 Conclusion

From the results of our analysis and modeling of the NYC Yellow Taxi data, there are clear trends highlighting tipping behavior to maximize driver income.

Over the 15-year period from 2009 to 2024, tipping behavior has shifted significantly, with a gradual increase in average tip amounts and percentages. In accordance with expectations, fare amount was the most highly associated among all the collected variables. In addition, trip distance had significant feature

importance, highlighting how longer rides tend to result in higher tips. Time-based variables such as the hour of day and day of week were less influential than initially expected. However, visual patterns in the data displayed some temporal trend, particularly around peak commuting hours and holiday seasons.

Geographically, we found that Manhattan and Queens not only had the highest number of trips, but was also the best performing model. These areas are also more consistent in tipping behavior compared to other boroughs like Staten Island, which had more variability and fewer trips. This highlights the population density and travel demand across New york City.

To maximize tips, New York City yellow taxi drivers should prioritize longer trips and higher-priced fares, as these factors are more highly associated with greater tip margins. Additionally, targeting popular hours, specifically common work times, can help boost earnings. As the economy continues to shift, it is important for drivers to take advantage of every opportunity to increase profit.

For New York City yellow taxi drivers to maximize tips, higher-priced fares and trips of further distance are more likely to result in higher tips margins, as well as popular hours, specifically common work times, had peaks in average tips. As the economy continues to change, it is important for drivers to take advantage of every opportunity to increase profit.

References

[1] Ofer H. Azar. The economics of tipping. *Journal of Economic Perspectives*, 34(2):215–36, May 2020.

[2] Graham Ferguson, Carol M. Megehee, and Arch G. Woodside. Culture, religiosity, and economic configural models explaining tipping-behavior prevalence across nations. *Tourism Management*, 62:218–233, 2017.

[3] Fangru Wang and Catherine L Ross. New potential for multimodal connection: Exploring the relationship between taxi and transit in new york city (nyc). *Transportation*, 46(3):1051–1072, 2019.

[4] Gerald S. Goldstein and Leon N. Moses. Transportation controls and the spatial structure of urban areas. *The American Economic Review*, 65(2):289–294, 1975.

[5] Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Cláudio T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, 2013.

[6] NYC Taxi and Livery Commission. THE EARLY YEARS: 1907 - 1935.

[7] Mehmet Baran Ulak, Anil Yazici, and Mohammad Aljarrah. Value of convenience for taxi trips in new york city. *Transportation Research Part A: Policy and Practice*, 142:85–100, 2020.

[8] BSW Amber Raymond and Kenneth Cramer. Taxi tipping in new york city (2014-2017): Reciprocity in hailed vs. dispatched cab fares. *OpRpRp*, page 61, 2020.

[9] Ye Hu and Rex Yuxing Du. Passenger group size and tipping: An empirical study of 50 million nyc yellow taxi rides. *Available at SSRN 4705908*, 2024.

[10] Won Kyung Lee and So Young Sohn. A large-scale data-based investigation on the relationship between bad weather and taxi tipping. *Journal of environmental psychology*, 70:101458, 2020.

[11] Srikant Devaraj and Pankaj C Patel. Taxicab tipping and sunlight. *Plos one*, 12(6):e0179193, 2017.

[12] Sarah Conlisk. Tipping in crises: Evidence from chicago taxi passengers during covid-19. *Journal of Economic Psychology*, 89:102475, 2022.

[13] Hejingyu Huang. Prediction of new york taxi tip behavior based on machine learning classification and regression methods. In *Proceedings of the 2023 2nd International Conference on Public Service, Economic Management and Sustainable Development (PESD 2023)*, pages 686–698. Atlantis Press, 2024.

[14] Taehooie Kim, Shivam Sharda, Xuesong Zhou, and Ram M Pendyala. A stepwise interpretable machine learning framework using linear regression (lr) and long short-term memory (lstm): City-wide demand-side prediction of yellow taxi and for-hire vehicle (fhv) service. *Transportation Research Part C: Emerging Technologies*, 120:102786, 2020.

[15] Ci Yang and Eric J Gonzales. Modeling taxi trip demand by time of day in new york city. *Transportation Research Record*, 2429(1):110–120, 2014.

[16] Siyu Liao, Liutong Zhou, Xuan Di, Bo Yuan, and Jinjun Xiong. Large-scale short-term urban taxi demand forecasting using deep learning. In *2018 23rd Asia and south pacific design automation conference (ASP-DAC)*, pages 428–433. IEEE, 2018.

[17] Asli Elif Aydin and Yüksel Acun. An investigation of tipping behavior as a major component in service economy: The case of taxi tipping. *Journal of behavioral and experimental economics*, 78:114–120, 2019.

[18] David Elliott, Marcello Tomasini, Marcos Oliveira, and Ronaldo Menezes. Tippers and stiffers: An analysis of tipping behavior in taxi trips. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–8. IEEE, 2017.

[19] Taxicab and Livery Passenger Enhancement Programs. TLC Trip Record Data.

[20] Diego Correa, Kun Xie, and Kaan Ozbay. Exploring the taxi and uber demands in new york city: An empirical analysis and spatial modeling 2, 2017.

Appendix A

Project Code: https://github.com/rngo1214/CapstoneProject