Unraveling Tipping Behavior in NYC Taxis

Barbara Noemi Szabo **György Józsa** **Marco Di Francesco**
Niek Pennings **Wesley Joosten**

February 2, 2024

1 Introduction

The city of New York, characterized by its high-density urban environment, serves as the subject of this analytical study focused on human mobility patterns. This research project aims to analyze the data derived from taxi trips, providing insights into traffic trends and tipping behavior. The project is available at the following link: `https://github.com/TheBarbaraIsTaken/BigData-NYCTaxis`.

1.1 About The Datasets

This study utilizes two primary datasets. The first, named "Newyork city Taxi Trip Records Dataset" [1] comprises approximately 15 years of Yellow Taxi trip records in New York City, spanning from January 2009 to September 2023 with a total size of 67 GB. This dataset includes detailed information such as trip distances, pickup, and drop-off times. The second dataset, the "New York City Buildings Database," [2] provides spatial context by categorizing urban structures into 11 types for a total size of 304 MB. This classification is used to examine the relationship between urban environment characteristics and tipping patterns.

1.2 Aims and Scope

The aim of this project is to analyze the factors influencing taxi tipping behavior in New York City. By examining various aspects such as the area of the ride, time of day, distance traveled, and traffic conditions, we seek to understand the patterns and dynamics of tipping behavior in the context of taxi rides.

The scope of the project includes several research questions:

[1] https://www.kaggle.com/datasets/microize/nyc-taxi-dataset
[2] https://www.kaggle.com/datasets/new-york-city/nyc-buildings/

1. **Influence of Area on Tipping:** Investigate how tipping behavior varies based on the area of the ride, as classified by the New York City Buildings Database. This analysis aims to uncover any correlations between the type or location of buildings and the propensity to tip.

2. **Impact of Time, Distance, and Traffic on Tipping:** Explore how tipping behavior during different periods of the day is influenced by factors such as distance traveled, traffic conditions, and duration of the ride. By examining the relationship between tipping amounts and these variables, we aim to discern any patterns or trends that may emerge.

3. **Changes in Ride Distribution Throughout the Day:** Utilize heatmaps to visualize how the distribution of ride destinations changes over the course of a day, particularly during typical office hours. This analysis aims to uncover any temporal variations in ride patterns within the city.

To answer these questions we followed the following pass

1. **Data Preparation:** We conducted data cleaning processes, uploaded datasets, and ensured collaborative permissions were set up for all team members.

2. **Data Processing:** Our team translated the coordinates of buildings, classified trip destinations based on nearby buildings, and repartitioned data for thorough analysis.

3. **Related Works Research:** Extensive research was conducted to contextualize our findings within the existing literature on taxi tipping behavior. This allowed us to gain a broader understanding of the subject and interpret our results effectively.

4. **Visualization:** We created visualizations, including stacked bar charts and line charts, to present our findings in a clear and comprehensive manner. These visualizations offer insights into tipping behavior trends across different variables and facilitate interpretation for stakeholders.

Overall, this project aims to contribute to the understanding of taxi tipping behavior in New York City by analyzing a comprehensive dataset and exploring various influencing factors. The insights gained from this analysis can inform stakeholders in the transportation and service industries, as well as contribute to academic research in the field of consumer behavior and urban transportation.

2 Related Work

In this section, we will have a look at the related works in the area of taxi tipping. Although tipping is a much-studied subject, the behavioural trends are still confusing and hard to predict. From a purely economic standpoint, tipping seems like an irrational action. However, research suggests that tipping behaviour is influenced by a complex mix of social, psychological, and contextual factors, including social norms, customer satisfaction, and perceived service quality [7]. In 2005, the tipping industry in the US had a size of $42 billion [1] and the percentage of transactions that included tips before and after the pandemic has increased significantly, rising from 43.4% to 74.5% [6]. [3] surveyed 12000 Americans and found that 72% noticed an increase in the number of venues where tipping is expected. It is a large industry that has been growing significantly, yet there are still uncertainties about the patterns and the reasons behind the behaviour.

When it comes to tipping in taxis, there have been various publications [5, 8, 4] on our same dataset, which analyze the Taxi and Limousine Commission (TLC) in New York. In their research, [5] has looked at relations between the average income of the pick-up and drop-off location and the percentage that has been tipped. They have found no real correlation between the average income of the pick-up location, and the amount that was tipped. On both sides of the income spectrum, people tended to pay about 20% as a tip. When it comes to considering the fraction of untipped trips, there was a strong correlation between the average income of the pick-up location and the fraction of tipped trips. Lower-income areas tended to have more untipped trips. Next to that, they found that around 4 AM the highest percentage of untipped rides occurred.

[8] has explored the relationship between tipping behaviour and whether or not the passenger was a tourist. They have tried to estimate whether or not someone is a tourist by identifying the pick-up location as near a hotel or not. This can give a good estimation but does not include every ride. On top of that, a specification has been made to identify trips as theatregoers rather than just tourists. A passenger was identified as a theatregoer if the trip ended within a range of 30 minutes from Broadway and if the time of the trip was around the time a show would've been played. They found that tourists tipped slightly more than locals, with about .2% to .5%. More significantly, they found that theatregoers tipped about .21% to .33% than other rides. Combined, tourists and theatregoers tipped on average .61% to .69% more. This work shows nicely that it is possible to make a classification of rides.

Lastly, in the work of [4], a study is done to find the influence of sunlight on the tipping amount. They compared the tipping percentages with overcast conditions to a transition to clear skies. Careful consideration was given to the transition and the clustering of the data points. They found that tipping increased by .63% when transitioning from overcast conditions to clear skies. It is good to evaluate these findings with other studies that were either inconclusive, or either found a positive or negative correlation. Nonetheless, in this category, [4] were the first to incorporate such a large amount of data.

The downside to these contributions, and something we will also face, is the problem that this dataset, which *is* the first of its kind, unfortunately, does not include much-needed additional demographic information. We can only make estimations about who is a tourist and who is a local. Estimations about income are also averaged and might reveal more about behaviour. Next to that, we know nothing about age, gender, occupation etc. Nonetheless, this dataset allows us to get the first insights into taxi driving on a larger scale without the need for surveys.

3 Methodology

3.1 Data Cleaning

The preprocessing phase of our methodology ensures the integrity and quality of the dataset. In this subsection, we outline the steps undertaken to clean the data, addressing issues such as incorrect information and identifying outliers.

The dataset under consideration spans multiple years, in some instances with slightly different schema and file formats. The files from 2019 were initially provided in CSV format, while those from previous and subsequent years are in Parquet format. To ensure uniformity and to reduce the complexity of our analytical processes, we converted the CSV files from 2019 to Parquet format. During this conversion, we also addressed a column mismatch, ensuring that the schema remained consistent across all files. Notably, files from 2009 and 2010 exhibited missing pickup and drop-off location IDs, among other discrepancies, leading to their exclusion from our analysis. This decision was informed by the lowest percentage of trips involving card-based tips during these years, as depicted in Figure 1. Additionally, we encountered a missing column, "airport_fee," in the 2019 files, necessitating schema adjustments. The subsequent rewriting of all files with a uniform schema not only addressed these issues but also facilitated efficient aggregation, our calculations and enhanced overall data processing speed.

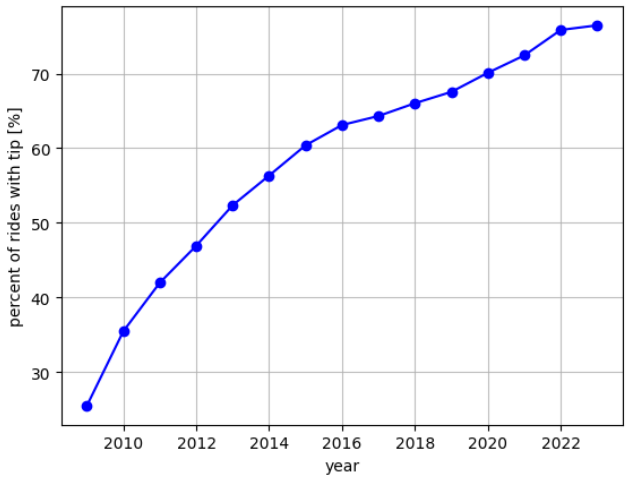


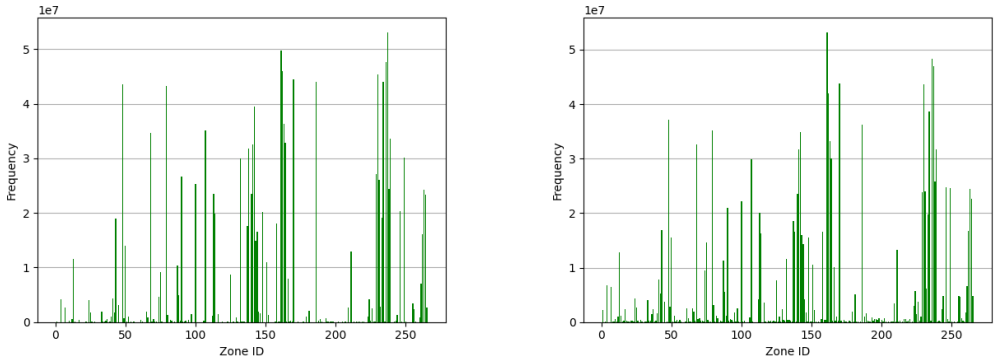
Figure 1: The percent of rides where tip was given for each year from 2009 until 2023

3.1.1 Handling Invalid Records

During data cleaning, records containing invalid pickup or drop-off timestamps were identified and removed from the dataset. This step ensures the temporal coherence of the data and eliminates any anomalies caused by inaccurate timestamp entries.

Instances with negative values for trip distance, travel time, or tip amounts were considered erroneous and were therefore excluded from the dataset. Cleaning such discrepancies is essential for maintaining the reliability of the information and preventing skewed analysis results.

Pick-Up and Drop-Off Locations: In our analysis, the zones function as critical variables for assessing spatial dynamics in the New York City taxi dataset. The data in these columns are error-free, with location IDs ranging from 1 (Newark Airport) to 265 ("Unknown"). The distribution of these zone IDs is illustrated in Figure 2, which provides a quantitative analysis of the frequency of various pick-up and drop-off zones in the dataset.



(a) Pick up (b) Drop off

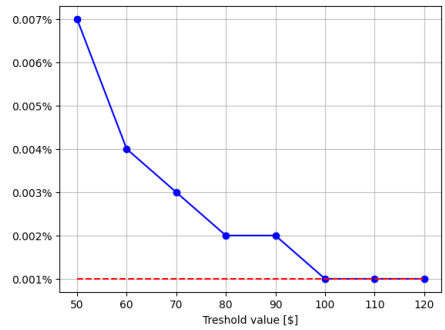
Figure 2: Distribution of zone IDs

3.1.2 Outlier Analysis

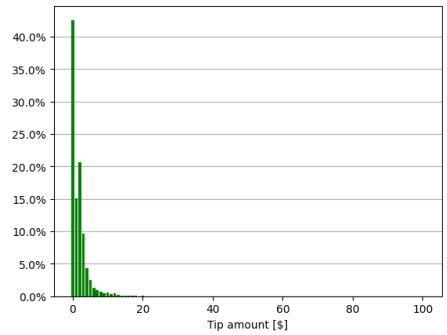
To augment the reliability of our analytical approach, a comprehensive evaluation of the dataset is performed to detect and isolate outliers. Outliers, defined as data points that significantly deviate from the norm, can compromise the integrity of statistical metrics and reduce the precision of the results. Utilizing statistical techniques and data visualization tools, we identify and exclude data entries that exhibit extreme values in critical parameters, including trip distance, travel duration, and tip amounts. This process ensures the elimination of data points that may potentially skew the dataset, thereby preserving the validity of our statistical analysis.

Tip Amount: Upon analyzing the tip amount data within the dataset, we identified outliers with tip amounts that disproportionately affect the statistical representation. Detailed analysis indicates that the maximum tip recorded is over 100 million dollars, an amount considered non-representative in the taxi service context. As Figure 3a shows, only 0.001% of the data entries record tips above $100. We categorize these data points as outliers and remove them from the dataset. Additionally, we eliminate entries with negative tip amounts to maintain

data accuracy. The refined data, as shown in Figure 3b, offers a more realistic distribution of tip amounts, mitigating the impact of these extreme outliers on the overall analysis.



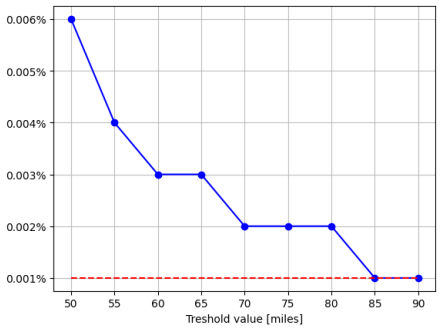
(a) The percentage of records that are greater than a threshold value with blue. The red line shows the boundary of 0.001%



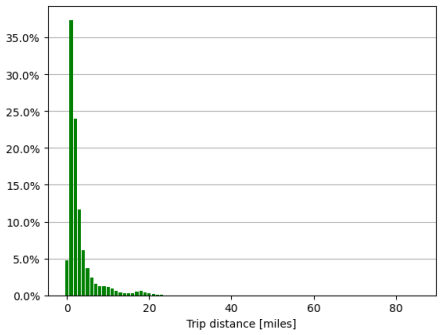
(b) The distribution of tip amount after filtering the outliers and negative values.

Figure 3: Tip Amount

Trip distance: In conducting the analysis of trip distances using the established methodology, we observed complexities arising from the heterogeneity in the distances covered by taxi journeys. Taxis, by nature, can traverse beyond the borders of New York City, complicating the identification of outliers. Our examination reveals that exceptionally long trips are relatively infrequent. Figure 4a visually encapsulates this insight, illustrating that a mere 0.001% of trips encompass distances exceeding 85 miles (approximately 136.79 km). Figure 4b displays the distribution of trip distances limited to a more reasonable range, between 0 and 85 miles.



(a) The percentage of records that are greater than a threshold value (blue line). The red line shows the boundary of 0.001%



(b) The distribution of trip distances in the range of 0 and 85 miles.

Figure 4: Trip Distance

3.2 Zoning

Since we want to investigate the relationship between the type of destination and tipping, we need a classification for the destination of each trip. Since we only know the taxi zone where a trip ended, we have to classify each taxi zone. The buildings dataset [2] includes different types of information regarding the zoning (i.e. the purpose) of buildings. To determine a label for each taxi zone, we first group all of the buildings per taxi zone and then use the zoning information from the dataset to determine a classification. We use two different types of information.

Figure 5: Zoning of taxi zones

Firstly, the dataset includes information on the zoning *per building*. Only using this information is not representative of the taxi zone as a whole, since buildings can have vastly different sizes and are of different importance for the zone. However, this information is well suited to determine the location of parks, as it includes buildings that are parts of parks.

Secondly, the dataset includes information on the floor area per building designated for certain usages (e.g. retail, factory, or residential). We sum the area per usage type of all the buildings within a zone and label the zone by taking the largest value. This gives us, unfortunately, only a distinction between residential and commercial areas, since the commercial type in the dataset can be used for different usages (e.g. manufacturing, offices) even though these also have separate categories. There does not seem to be enough information to easily distinguish different types of commercial zones.

After these two steps, there are still some zones left without a label, because the buildings dataset does not include a single building in these zones. This brings us to the next step of labeling the taxi zones. Since we assume that airport trips have a different influence on

tipping as compared to other commercial zones, we label all the airports by hand, there are three major airports in New York City, each of which is its own taxi zone, which was partly yet unlabeled. This leaves only Rikers Island without a label, which is fully covered by a jail, which we also labeled by hand.

The final result can be seen in figure 5.

4 Results

Heat map tip per area: Our preliminary spatial analysis, focusing on tipping activity by zone and excluding nonspecific areas, is visualized in figure 6. It is evident that Newark Liberty Airport is in the zone with the highest tipping, followed by Staten Island. When these findings are compared with zoning classifications (see figure 5), a trend emerges revealing that residential zones generally exhibit lower tipping rates compared to commercial areas.

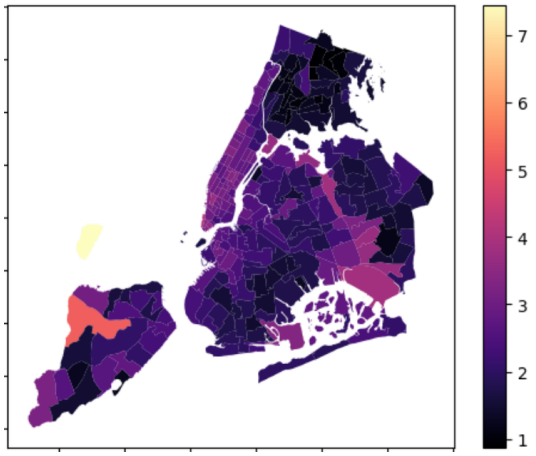
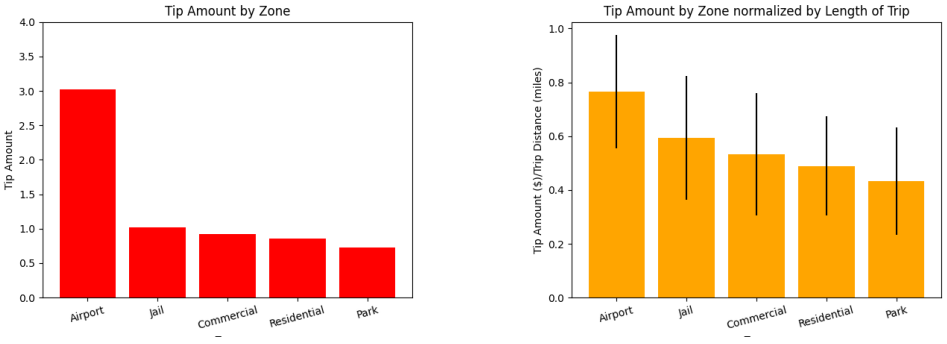


Figure 6: Heat map highlighting tip amount per zone.

Average tipping per zone: The disparity in tipping amounts across zones is examined in two distinct ways, as illustrated in Figures 7a and 7b. The former displays raw tipping data, while the latter provides a perspective normalized by trip length. The normalization process highlights variations in tipping behavior relative to the distance of the journey. This was done given the intuition that passengers are inclined to tip more on lengthier trips, reflecting a tipping culture that scales with total cost rather than a fixed amount each time. In both cases, the airport zone exhibits a higher tipping average, while the park has the lowest tipping.



(a) Not normalized

(b) Normalized by distance

Figure 7: Tip Distribution Across Zones

Tipping over an hour of the day: Our main goal for this paper was to understand how tipping might be influenced by factors such as the current traffic or the time of day. To achieve this, we defined *traffic flow* as the average speed of the taxi during its ride. Since traffic is dynamic, we have analyzed the change in traffic over the time of the day. Similarly, the average tip amount is also subject to change during the day, so we have plotted both to see if there is any correlation between the two. As shown in plot 8, there seems to be a connection between how fast the taxi moves through the city and the tip given to the driver.

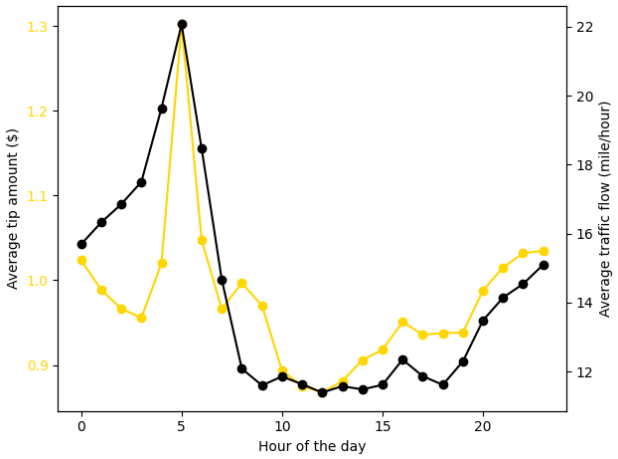


Figure 8: Average tip amount (yellow line) over traffic flow (black line).

Tipping over traffic: To fully understand the connection between traffic flow and tipping, we directly compared the two, the result of which can be seen on plot 9.

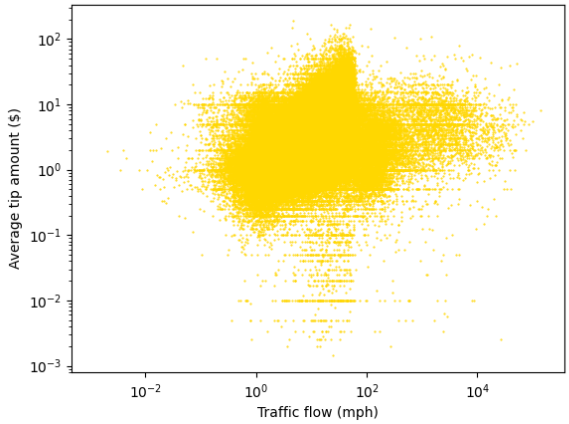


Figure 9: Average tip amount (y-axis, log scale) over traffic (x-axis).

Taxi activity over hour of the day: After the striking opposition of plots 8 and 9, we wanted to see how much the traffic might influence tipping. A good approximation of the

business of a time of day is the number of taxis actively used. This makes sense, since taxis and personal cars tend to be active at the same time. Another added benefit of analysing how many taxis are on the road at each given time is that we can estimate how accurate the data is for each time period. At a time where few taxis are on the road, the average tip has a higher probability of being atypical.

As shown in figure 10, the number of taxis actively used varies wildly during the day. Particularly, at around 3-5 AM, only about 40% of the average number of taxis are active.

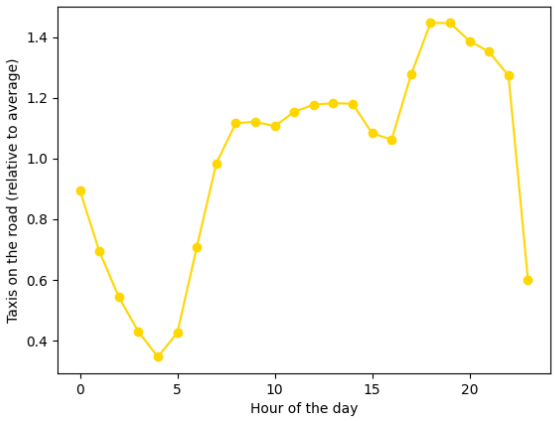


Figure 10: Number of taxis on the road at each hour of the day, compared to the average.

5 Discussion

5.1 Zoning

An older version of the dataset included the coordinates of the pickup and drop-off location for each ride, this would have allowed us to look at a few buildings close to the location to determine the type of destination. Unfortunately, this information was no longer available, leaving us with only the pickup and drop-off *taxi zone* for each trip. We can, hence, only make a rough guess on the type of destination for each trip based on the prevalence of types of destinations within the complete taxi zone. This has to be done by aggregating the information of all the buildings within a taxi zone in some way, but how to do this in a way that is representative of reality is not immediately clear. Using the sum of floor area used for different purposes seems like a fair way to label a taxi zone since floor area is a good measure of the amount of space used for a certain purpose in a taxi zone. However, the amount of space used might not correspond to the actual destination type of the taxi trips within the taxi zone. Ideally, we would compare the destination type of known taxi trips to the labeling of taxi zones we determined to see whether our labeling is representative and potentially adjust our labeling process, but unfortunately, such information is not available to us.

5.2 Area Analysis

The analytical outcomes deduced from the 'tip per area' heat map and the 'average tipping per zone' bar chart indicate a pronounced disparity in tipping across different zones. The airport's position as the locus of the highest tipping aligns with the hypothesis that a wealthier demographic predominates among taxi patrons in that area. Additionally, the comparative analysis of residential and commercial zones reveals a distinct variance in tipping tendencies, with residential zones consistently demonstrating a lower propensity towards tipping. This pattern hints at underlying socio-economic dynamics influencing tipping behaviors within these urban localities.

5.3 Tipping Analysis

As shown in plot 8, there seems to be a connection between how fast the taxi moves through the city and the tip given to the driver. During the hours of the day when the taxis can move faster, people tend to tip more. This is not necessarily a causal link though. Our initial assumption was that there would be a direct correlation between these values, since a smoother, faster taxi ride might make the passenger more satisfied, therefore giving a higher tip. Our results however show a different story. It seems from the plot that there is no meaningful correlation between traffic flow and tipping. This might be explained by traffic flow having only a minor effect on tips, while other unrelatable factors such as the location of the ride or the personality of the driver exert a greater effect.

Figure 10 offers some insight as to why tipping is so extreme at the time of 3-5 AM, but it does not fully explain the effect. One possibility is that at such a time very few taxi drivers are working to begin with, and as such demand is a lot higher compared to the supply. Another possible explanation is that people who use taxis at such hours are probably traveling to atypical areas, such as to an airport, and therefore exhibit atypical tipping practices.

6 Conclusion

In this study, we conducted an examination of tipping behavior in New York City taxis, using the New York City Taxi Trip Dataset and the New York City Buildings Database. Our analysis involved statistical techniques to uncover insights into the determinants of tipping patterns in an urban taxi context. Our main findings indicate significant variances in tipping behaviors across different urban zones. A higher tendency to tip in commercial zones, especially at major hubs like airports, was observed, showing the socio-economic factors influencing tipping. The study also revealed a complex interplay between tipping behavior and variables such as trip distance, time of day, and traffic conditions. Notably, our findings suggest a relationship between traffic flow and tipping, diverging from conventional assumptions.

The study was constrained by the lack of comprehensive demographic data in the dataset, limiting the depth of our behavioral analysis. Despite this, our approach, incorporating spatial data, adds a new view to consumer behavior in the context of urban transportation that is not present in the literature and may be useful for future works.

References

[1] AZAR, O. H. Strategic Behavior and Social Norms in Tipped Service Industries. *The B.E. Journal of Economic Analysis & Policy 8*, 1 (3 2008).

[2] CITY OF NEW YORK. New York City - Buildings Database, 10 2016.

[3] DESILVER, D., AND LIPPERT, J. Tipping Culture in America: Public Sees a Changed Landscape. Tech. rep., Pew Research Center, 11 2023.

[4] DEVARAJ, S., AND PATEL, P. C. Taxicab tipping and sunlight. *PLOS ONE 12*, 6 (6 2017), e0179193.

[5] ELLIOTT, D., TOMASINI, M., OLIVEIRA, M., AND MENEZES, R. 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)* (8 2017), IEEE, pp. 1–8.

[6] LORSCH, E. Tipping in the United States has gotten out of control, experts say., 5 2023.

[7] LYNN, M. Service gratuities and tipping: A motivational framework. *Journal of Economic Psychology 46* (2 2015), 74–88.

[8] NETO, A. B. F., NOWAK, A., AND ROSS, A. Do Tourists Tip More Than Local Consumers? Evidence from Taxi Rides in New York City. *International Regional Science Review 42*, 3-4 (5 2019), 281–306.