MAST30034 Project 1

New York Taxi Data Tip Analysis

Can Senyurt
Student ID: 1079752
Github repo with commit

August 25, 2024

1 Introduction

In New York City, where the rhythm of urban life is driven by both people and weather, taxi drivers face the dual challenge of navigating fluctuating weather conditions and maximizing their earnings. This report aims to explore the relationship between these variables by analyzing the interplay between weather data and yellow taxi performance, with a particular focus on optimizing tip earnings for drivers.

Comprehensive data sets that capture both weather parameters and details of trips were used to examine correlations between weather conditions, such as rain, snow, and temperature fluctuations, and taxi metrics, including fare amounts, tips, and duration of the trip.

The findings of this analysis aim to offer actionable recommendations for taxi drivers. By aligning their service strategies with weather-induced demand patterns and operational considerations, drivers can better position themselves to increase their tip earnings. This report not only provides practical tools for individual drivers but also contributes to a deeper understanding of how environmental factors influence urban transportation economics.

1.1 Data

The primary dataset used in this analysis is the **TLC Trip Record Data** provided by the NYC Taxi & Limousine Commission. The dataset includes information such as pickup/drop-off location, trip duration, total amount paid, tolls fees etc. This analysis focuses on the period from December 2023 to May 2024, as this date range was the most recent available at the time of writing this report.

The secondary data set used for the analysis is the weather data provided by Visual Crossing (https://www.visualcross which is a private data collection company that provides large-scale weather data. This dataset includes many features such as hourly temperature, humidity, wind speed and direction, weather conditions, etc.

Dataset	Instances	No. of Features
TLC Yellow Taxi	20169467	19
VC Weather	4391	24

Table 1: Used Datasets

2 Preprocessing

Many pre-processing procedures were applied to both datasets to ensure the quality and the legitimacy of the entries in each dataset. The TLC dataset in particular was not very well kept and required the removal of roughly 35% of the dataset. Resulting to Resulting in approximately 13,110,000 remaining instances. The 2 cleaned datasets (TLC data and weather data) were then combined to be used for visualization and model training.

2.1 TLC Data

- **Records outside of the intended date range** were removed to ensure appropriate weather data could be matched
- **Unreasonably long trips** were removed. This was done by looking at the quantiles and keeping 99.99% of the range, since there were trips that lasted thousands of miles. Only trips that were shorter than 55 miles were kept as this is where the **majority** of the data lied.
- **Removal of short trips** was also necessary as there were trips that were close to/less than 0 miles. An assumption of most people would prefer to walk 0.65 miles rather than taking a taxi was made.
- **A hard limit of 0-7 passengers** were put in place. I allowed for 0 passengers as some people might only give the taxi driver an item for transportation, which could be valid. Any value more than 7 passengers were removed as the NYC only allows up to 6 passengers under normal circumstances with the exception of allowing a child as the 7th passenger.
- **Fare amounts start from $2.5** according to the NYC Taxi Commission so only records of this range were kept.
- **The allowed tip range** was set to $0-$65 as there were many entries with negative and unreasonably high values. The limit of $65 was determined by calculating the 99.99th quantile
- **Only entries with cash payments were removed.** This was done because the dataset guidelines specify that tip amounts are only recorded for credit card payments, which is the main focus of the analysis.
- **Any entry with RatecodeID larger than 6 was removed** to ensure that the data was within the range specified by the TLC.
- **Trips that costed more than $250 were removed** since this is where the 99.99th quantile was found to be.
- **Average speed was calculated** and any entry outside of the range 4.5 mph and 65 mph were removed since the average speed in traffic in New York is around 7 mph (NYC, 2018), and the majority of the data was within the 4.5 - 65 mph range. This was necessary as there were entries that were clearly false with the taxi ride taking a whole day and total distance covered being nearly 0 miles, or a trip lasting only a few seconds yet covering hundreds of miles.
- **Any trip lasting less than 1 minute** was also removed since by assumption, most people would just walk this distance.

2.2 Weather Data

Since this data was provided by a private company, all of the data here was clean and accurate. Only minor alterations were done which included dropping some columns that were not relevant such as which weather station sent the data (which was always the same), and the name, solarradiation, solarenergy, precibprob and preciptype columns. these were dropped because they are not considered significant for daily life.

3 Visual Analysis

In this section, we explored the relationships between tip amount and various features such as the day, hour, location and weather information.

3.1 Tip Profitability and Map

The first analysis was done by defining tip profitability as:

$$\text{Tip Profitability} = \text{Tip Percent} \times \text{Total Trips from a Location} \div \text{Total Trips}$$

which makes logical sense, since a $5 tip from a trip that costed $10 is better than a $10 tip from a trip that costed $60 since the second trip would have lasted much longer, wasting precious time the driver could have made another quick trip to potentially earn another $5. Furthermore, taxi drivers will want to drive around areas where they are most likely to receive trips from to minimize down-time. The figure below shows the most profitable areas based on this formula.

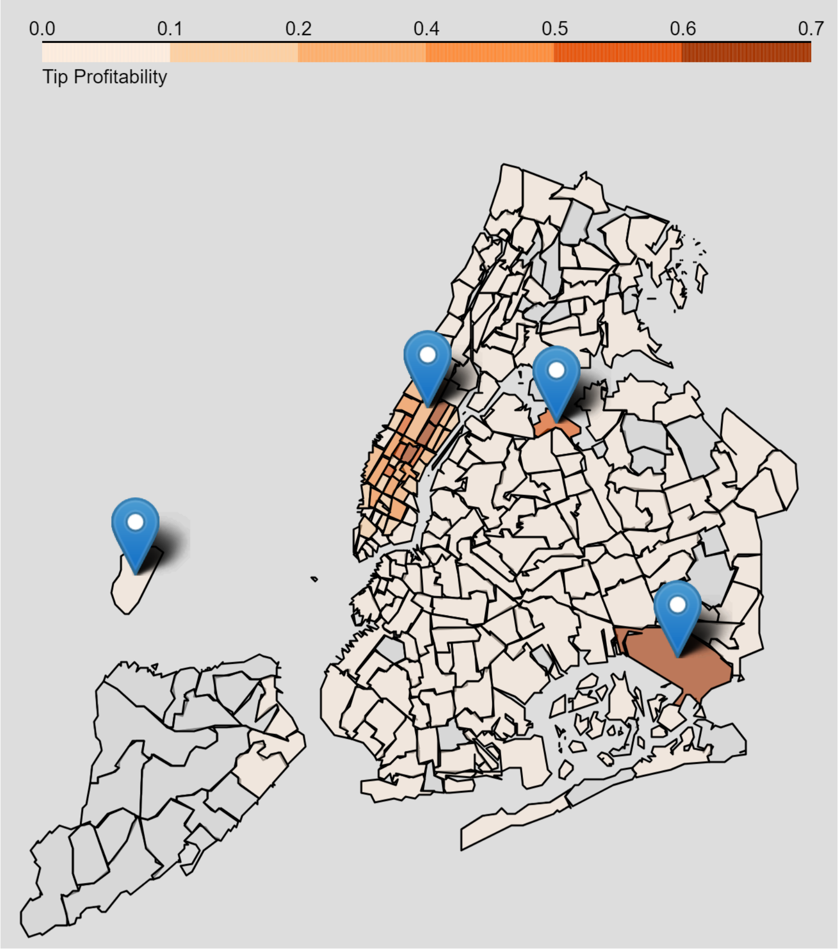

Figure 1: Tip Profitability arond New York

The 2 markers on the right side and the marker on the far left ar the Airports in NYC. According to the data, taxi drivers can expect high tip profit from the JFK and LaGuardia airports (the two on the right) whereas Newark Liberal International Airport is not very profitable. This could be due to less frequent pickups from this region since this airport is located further outside of the city. The other highlighted zone is Central Park which is at the heart of Manhattan, which is the most densely populated part of New York.

3.2 Correlation Map

A correlation map of the numerical values within the combined dataset was created to get an idea of which parameters might be affecting tip amount the most.

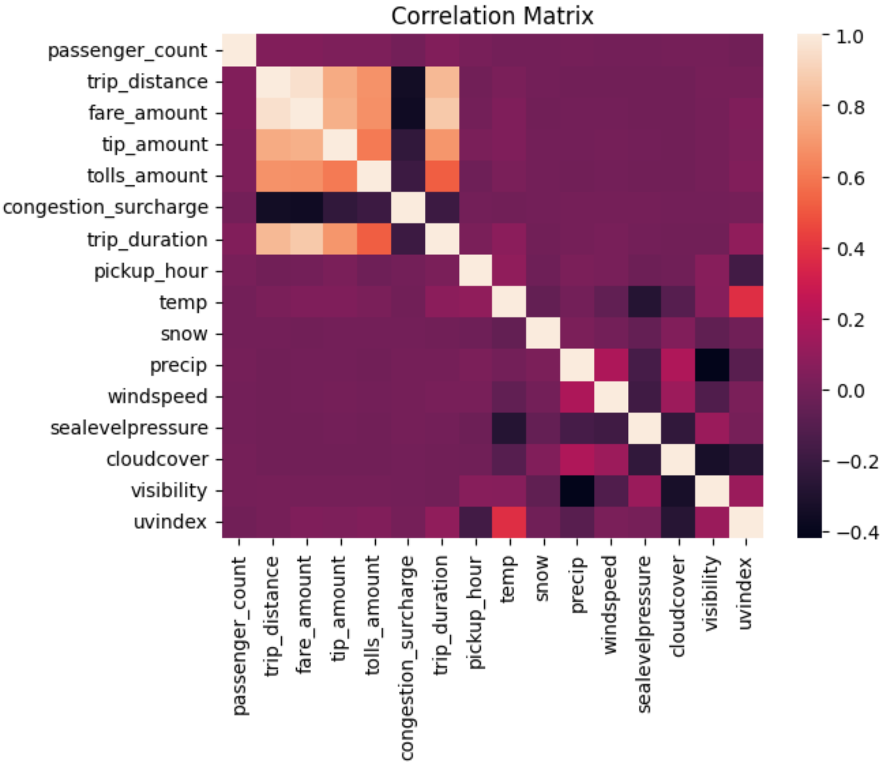

Figure 2: Correlation Matrix of the numerical values

Upon first inspection, there appears to be negligible correlation between tip amount and many features except the expected ones such as trip distance and fare amount. Some things to note however is there appears to be some positive correlation between tolls and tip amount which is interesting to see. Also a negative correlation between congestion and tip amount. This could be due to the passengers getting frustrated while in traffic.

3.3 Heat Maps

Figures 3 and 4 show the average tip amount and the number of trips by hour and day, respectively.

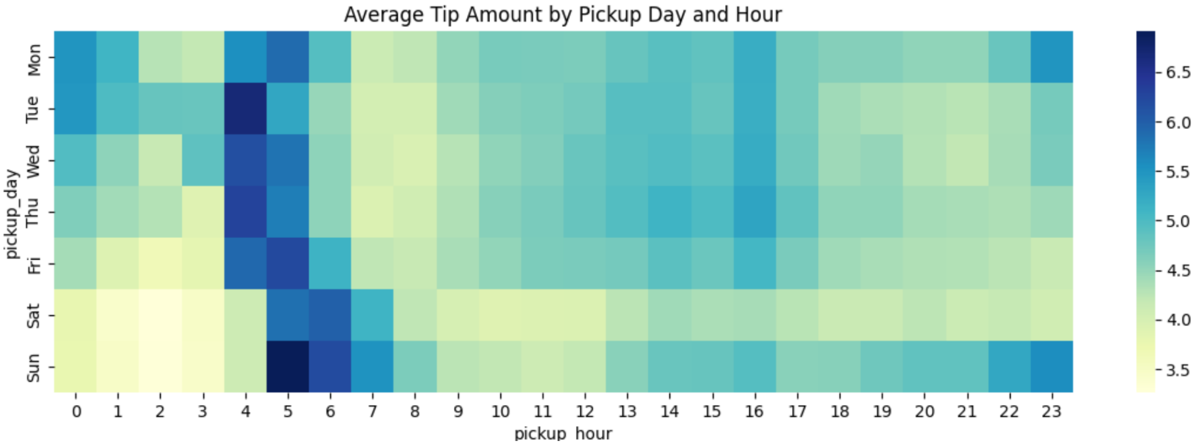

Figure 3: Avg Tip amount by day and time

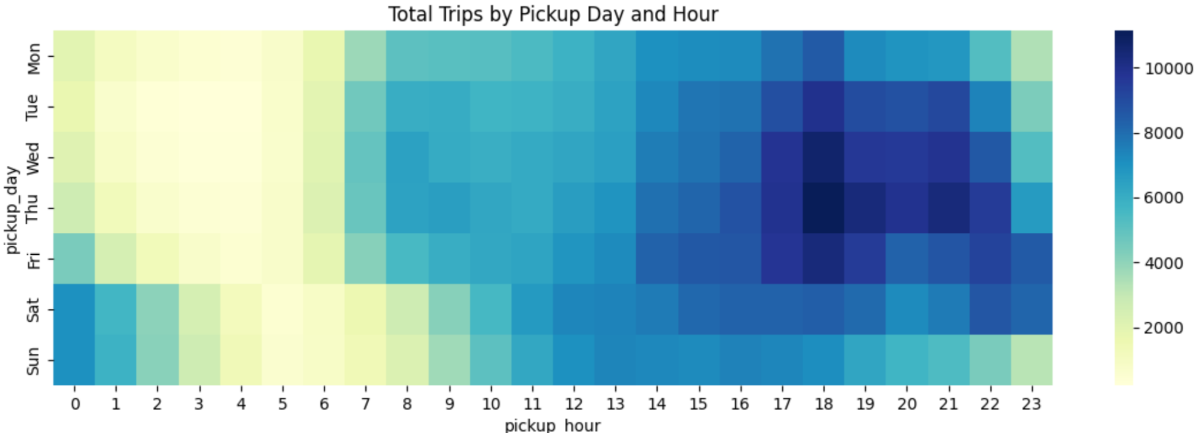

Figure 4: Trip counts by day and time

The first thing that stands out from these two plots is they appear to be almost exactly opposites of each other, which is an inconvenience for the taxi drivers since in the very early mornings (4AM-5AM) they can expect higher tip amounts, whereas around 5PM-7PM they can expect more amounts of trips, but never both simultaneously.

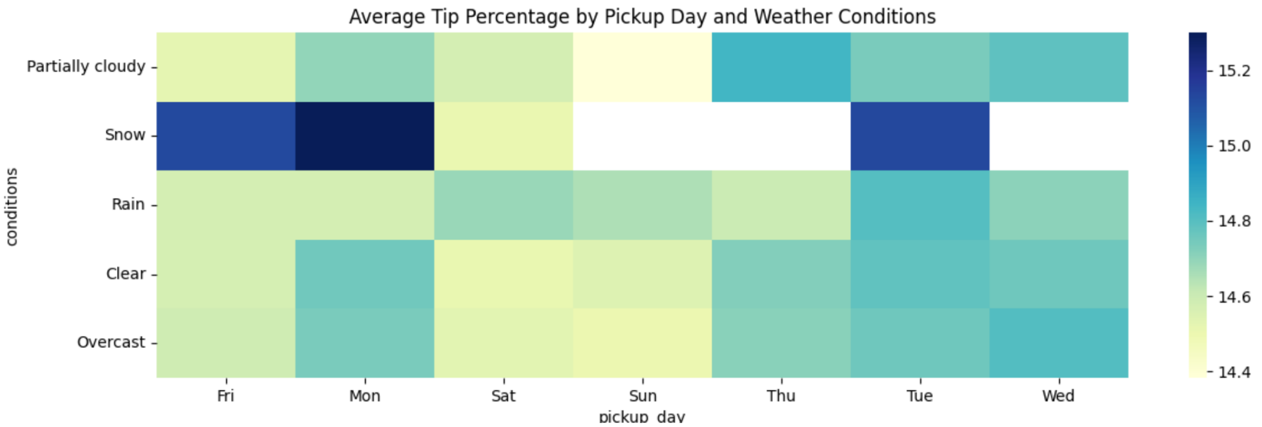

Figure 5: Tip Percentage based on day and weather conditions

Weather Condition	Avg. Tip %
Clear	14.66
Overcast	14.66
Partially Cloudy	14.65
Rain	14.67
Snow	15.03

Table 2: Average Tip % Based on Weather Condition

As it can be seen from both Figure 5 and Table 2, passengers are more likely to tip higher on snowy conditions compared to every other weather forecast. However, this could be due to a lower amount of data from snowy days as they are less common compared to other conditions (for example, there weren't any snowy Sundays, Thursdays or Wednesdays in this random sample of 7% of the entire data).

4 Modelling

4.1 Lasso Linear Regression

Since the aim of this model is to predict tip amount by using both numerical and categorical features, one possible model is to use a Linear Regression model which is capable of handling both types of data. The reason for using a Lasso model is to ensure that the model does not overfit the high amount of data and try to keep a simple model that would generalize well into real world scenarios.

4.2 Random Forest Regression

As Random Forest Regression model was chosen for the second model as it also could handle both numerical and categorical values. Although losing out on interpretability, the RFR model was expected to perform better than the Lasso model since it is a more powerful ML Algorithm.

4.3 Model Results & Discussion

After training and tuning both models over a random split of 80/20 train/test split and performing 3 fold cross validation, both models performed quite similarly with R2 scores of 0.62 for both models and RMSE values of 2.42 and 2.45 for the Lasso and RFR models respectively. Although these values are not ideal, they are acceptable as it is quite challenging to predict how much a person will tip even if we knew the person personally.

What is more important however, is the weighting of each feature and how they affect the model.

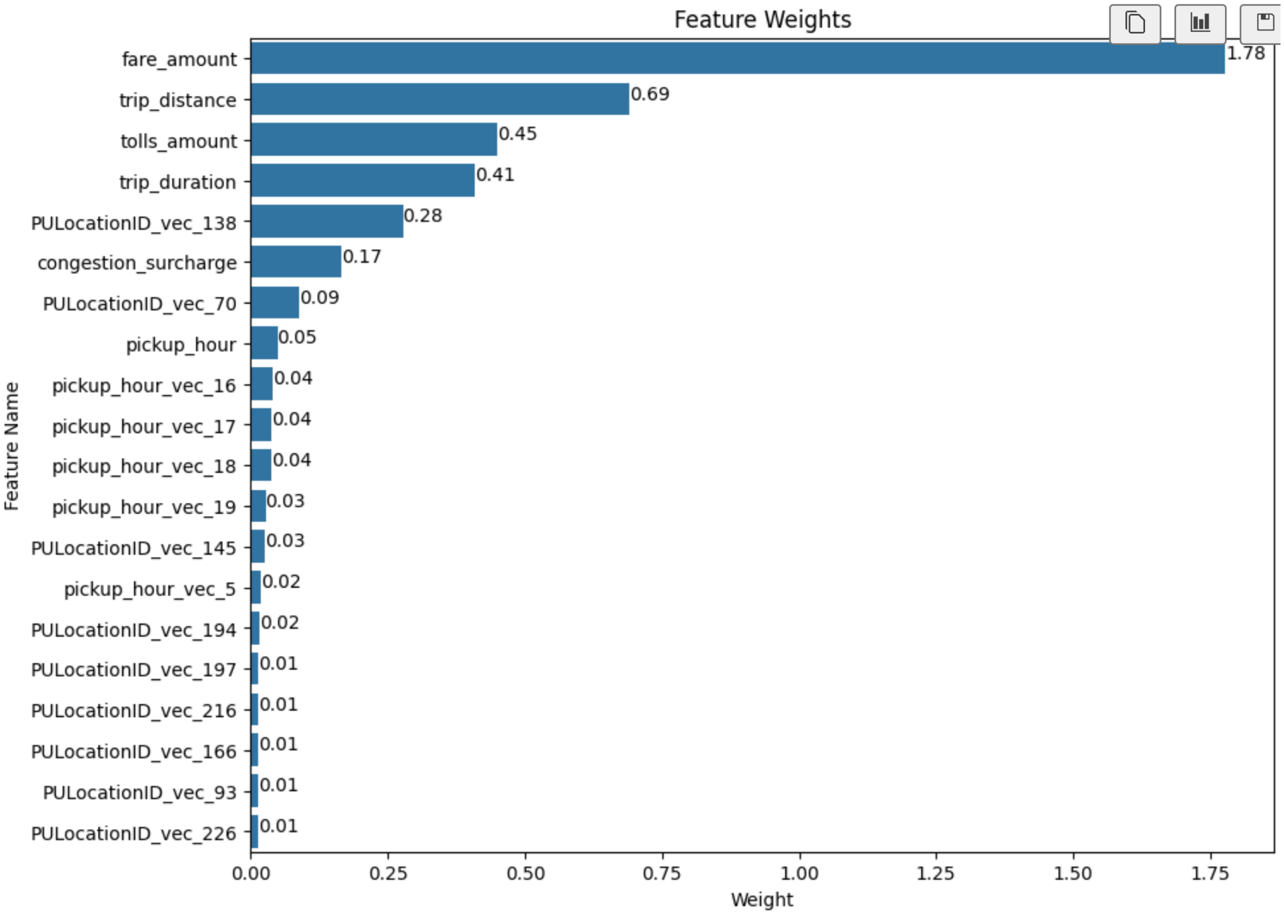

Figure 6: Feature weighting of the Lasso Regression Model

As previously observed, fare amount, trip distance, and tolls are the most significant features in predicting tip amounts. However, this model presents a notable divergence: congestion appears to have a positive relationship with tip amounts, contrary to the findings from the earlier heatmap.

Other key insights include the identification of the top five pick-up locations for passengers: LaGuardia Airport (Queens), East Elmhurst (Queens), Bayside (Queens), Bedford (Brooklyn), and Bedford Park (Bronx), as determined using the lookup table provided in the data folder. Additionally, the optimal times for passenger pick-ups are between 4 PM and 7 PM.

While the model was not highly accurate in predicting tip amounts, it successfully identified general patterns that align with our visualizations. However, there are notable discrepancies: the model underemphasized JFK Airport and Manhattan compared to Figure 1, which was unexpected. Furthermore, the average tip amounts depicted in Figure 3 contrast with the model's findings, as Figure 3 indicates poor tipping during the 4 PM to 7 PM window, whereas the model assigns high importance to these hours. Figure 4 shows that these hours are among the busiest, suggesting they are highly profitable for taxi drivers. This timeframe could explain why congestion is a significant factor in predicting tip amounts, likely due to increased traffic as the workday ends.

5 Conclusion & Recommendations

This report conducted an initial exploration of the TLC dataset, integrating it with weather data to predict tip amounts earned by yellow taxi drivers in New York City and uncover potential passenger behavior patterns. The analysis, which included visualizations and two trained models, revealed some discernible tipping patterns. Specifically, tipping behavior appears to be influenced primarily by the time of day and pick-up locations. Optimal tipping times are between 4 PM and 7 PM, aligning with the end of the workday, while top pick-up locations include LaGuardia Airport and Manhattan. These areas are high-traffic zones with significant traveler activity, which likely contributes to the observed patterns.

Weather data did not significantly impact the Lasso model's predictions. The high tipping percentage observed in Figure 5 might be attributed to the limited data available for snowy days, given their relative rarity. Although the models provided useful insights, they might not capture all the nuances of tipping behavior. A more sophisticated model, such as a Neural Network, could potentially better identify underlying patterns. However, predicting tipping amounts with high accuracy remains challenging due to numerous influencing factors, such as individual passenger preferences and the context of their day.

Recommendations for taxi drivers include working around Manhattan, Queens, and LaGuardia Airport, especially during peak hours to transport people getting out of work and kids getting out of school. They can expect higher tips on longer trips but this is trivial as the percentage of the tip could still be the same. A slightly more useful metric to check could be to look at the tip profitability as calculated in the Visualizations section which also agrees with the Manhattan and LaGuardia Airport recommendations, but also includes JFK Airport.

6 References

New York City Department of Transportation. (2018). New York City mobility report 2018. https://www.nyc.gov/ht

New York City Taxi and Limousine Commission. (n.d.). TLC trip record data. New York City Government. Retrieved August 25, 2024, from https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Visual Crossing. (n.d.). New York City, USA weather history. Visual Crossing. Retrieved August 25, 2024, from https://www.visualcrossing.com/weather-history/New+York+City%2CUSA

Wikipedia contributors. (n.d.). Demographics of New York City. Wikipedia, The Free Encyclopedia. Retrieved August 25, 2024, from https://en.wikipedia.org/wiki/Demographics_of_New_York_City