Strategic Taxi Distribution in NYC

Predicting Optimal Deployment with Business and Trip Data

Daksh Agrawal
Student ID: 1340113
Github repo with commit

August 25, 2024

1 Introduction

As Uber and Lyft vie for supremacy in the New York City (NYC) rideshare market, understanding the nuances of hourly demand becomes a pivotal strategy[1]. New York has long been known for its fast-paced lifestyle and High Cost of Living, implying that consumers are highly sensitive towards trip pricing and waiting times. While a multitude of factors governs these, both factors can be optimised by optimising the placement of taxis in the city[1]. In a perfect allocation of taxis to demand, companies can serve demands almost instantly and reduce the cruising miles, which are a factor for the pricing models of both Uber and Lyft[2, 3].

This study uses the TLC dataset for High-Volume For-Hire Vehicles (HV-FHV), as HV-FHV taxis dominated the market in recent years, and have substantial control over their fleets[4]. This study focuses on the timeline of January 2022 to June 2022, a period marked by a sudden decline of Lyft's market share and stock price, allowing us to consider the dimension of competitive landscape[5].

In addition to this, the Legally Operating Businesses dataset provided by the Department of Consumer and Worker Protection (DCWP) available on the NYC OpenData portal was used in this analysis[6]. While tourists usually account for a large chunk of Taxi usage, the local New York citizens are still the primary users of taxi services[7]. Locals typically use the services for commuting to and from work, as well as for nightlife, parties, and day trips, hence concentrating around areas with large number of businesses. Thus, we expect the number of businesses in an area to have a strong impact on the demand in that area.

We also use Open-Meteo's Historical Weather API to fetch basic weather data for New York, recorded at Central Park, as this data is fundamental to effective statistical modelling [8].

The primary audience for this analysis includes Uber and Lyft, given their significant influence over the market and prioritisation of customer service. Further, unlike Yellow and Boro taxis which are fragmented and operate almost independently, Uber and Lyft have some control over driver behaviour. They can execute large-scale optimisation plans through their apps. In addition, these findings may also be used by individual drivers to optimise their plans and policymakers to develop and regulate appropriately.

This study analyses how various factors influence demand, and uses Machine Learning models to predict demand across NYC at any particular time.

2 Preprocessing

Since all of the data is generated by app-based providers Uber and Lyft, it does not rely on data entry and thus, has far fewer errors and discrepancies than datasets for traditional Yellow and Boro taxis.

2.1 Missing Values

Only 2 columns contain missing values, `originating_base_num` and `on_scene_datetime` since they are restricted to accessible vehicles only. Since this information is irrelevant to our analysis, we can safely ignore these columns.

2.2 Outliers

Some outliers that defy set business rules were detected:

1. Non-Positive Trip Miles
2. Non-Positive Trip Duration
3. Negative Tips

Rows containing these issues were dropped and in total comprised of around 0.02% of all rows (21604). Next, there were some more subtle outliers like:

1. Extremely long Trip Durations
2. Pickup / Dropoff Location IDs outside of the defined range

These factors were ignored as they aren't clear violations of rules, and are trivial to our research. Additionally, out-of-city trips assist us with inter-borough analysis.

3 Analysis and Geospatial Visualisation

The primary attribute of interest in this analysis is the hourly demand for rideshare services in different areas of the city. The demand data exhibited a highly skewed distribution, which was log-transformed to approximate a normal distribution, facilitating more accurate modeling. No significant outliers were identified, ensuring the reliability of the transformed data. See Figure 12 (Appendix).

Next, we explore the relationship between demand and various other factors in the TLC dataset, as well as the legally operating business dataset.

3.1 Hour of Day and Day of Week

From Figure 1, there are some clear cyclic patterns in the demand for taxis over the day, with spikes occurring around 8:00 AM-9:00 AM as people travel to work. Next, we see a generally higher demand from the evenings till late at night, as people return from work, and travel for nightlife. Saturdays and Sundays behave wildly differently, with much more late-night demand, possibly due to increased nightlife. Also, the morning spike is missing due to work holidays.

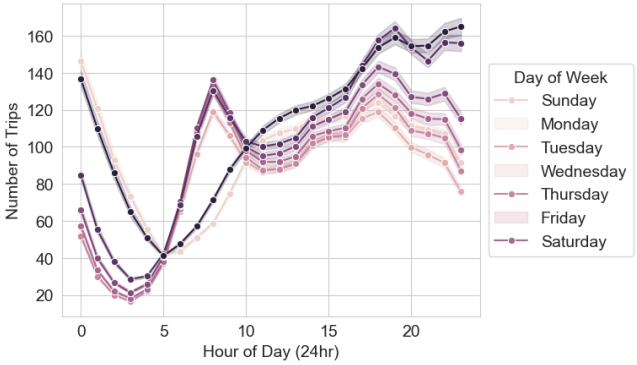

Figure 1: Average Demand by Hour and Weekday

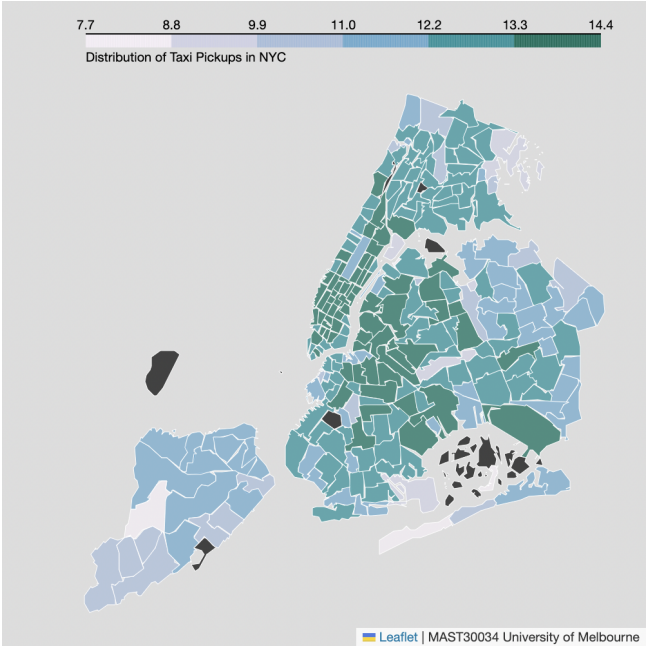

Figure 2: Total Pickups (Log Scale)

3.2 Pickup Zone

The amount of pickups from different areas varies very heavily across the zones in New York, with very high density around popular areas like Central and Southern Manhattan, which are major business and tourism hubs (Figure 2). The demand is also extremely high in airports and major transportation hubs.

3.3 Number of Businesses

While the relationship between demand and the number of businesses in the area is complex at first, it becomes more apparent when we break it down by the borough. From Figure 3, we see a positive trend between the number of businesses in a zone and the demand, but the effect varies heavily by the borough. Yet, we see medium-high Pearson correlation values for all boroughs, indicating it may be a useful feature in our modelling when combined with the borough factor.

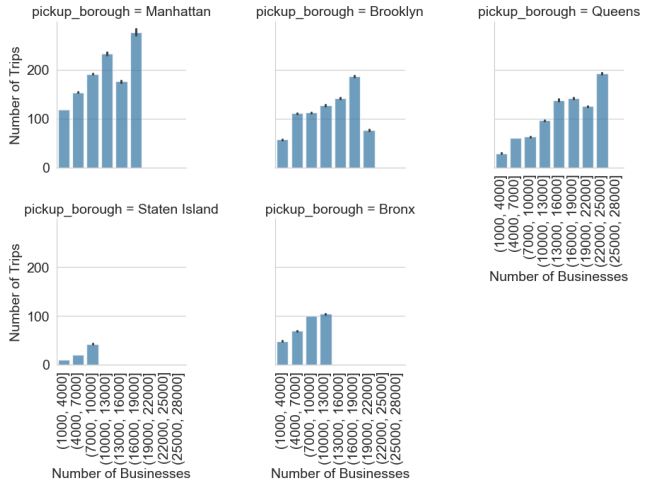

Figure 3: Number of Businesses vs Number of Trips

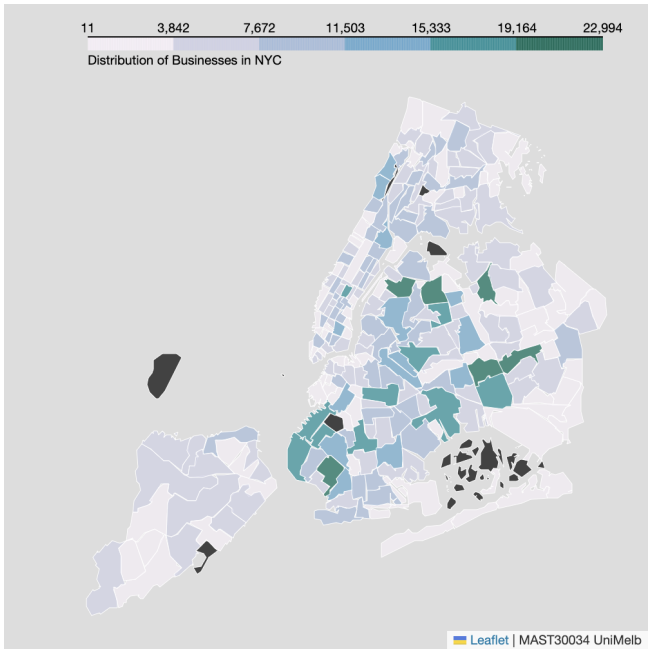

Figure 4: Distribution of Businesses

Borough	Pearson Correlation
Bronx	0.5784
Brooklyn	0.3681
Manhattan	0.4096
Queens	0.3016
Staten Island	0.4881

Table 1: Pearson Correlation between Number of Businesses in Area and Demand

3.4 Inter-Borough Trips

From Figure 11 (Appendix) we observe that most trips in NYC are within borough trips, with rare cross-borough trips. Most of these cross-borough trips start or end in Manhattan, hinting at the general importance of Manhattan in NYC life. From previous research, NYC residents tend to live in the same borough as they work in and prefer public transport for long cross-borough travels[7]. Further analysis may be done on the cross-borough trips and may be used by the Department of Transportation to optimise the public transport network.

3.5 Surges

Surges in demand are a critical opportunity for taxi providers to make use of unusually high demands. Usually, they may happen due to special events like concerts, parties or sports. Both companies can charge additional surge charges on these trips, and drivers concentrating around these in advance can heavily improve customer experience and even help organisers and planners ensure a smooth outflow[2, 3]. In our analysis, we define a surge as an event when the demand is more than double the median demand in that area. From Figure 5, we see that surges are most common late at night from 1:00

AM to 4:00 AM, which from Figure 1, is usually a low-demand time. Drivers may tend to take breaks during those times and if they are correctly identified, we can encourage them to design their schedule around the surge for maximum profitability. Next, we analyse the surges by location using Figure 6 and observe that surges are much more common in some areas than others, especially around parks and event venues, further confirming our theory about special events driving surges.

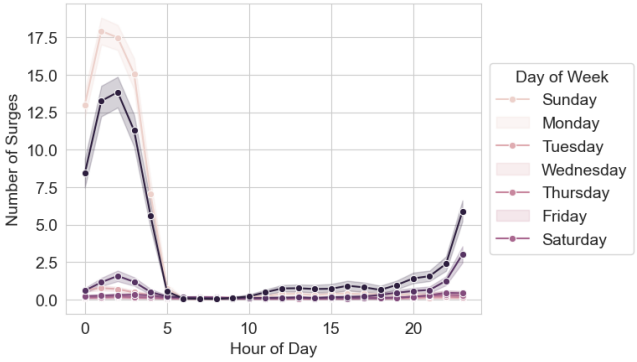

Figure 5: Average Number of Surges by Hour and Day

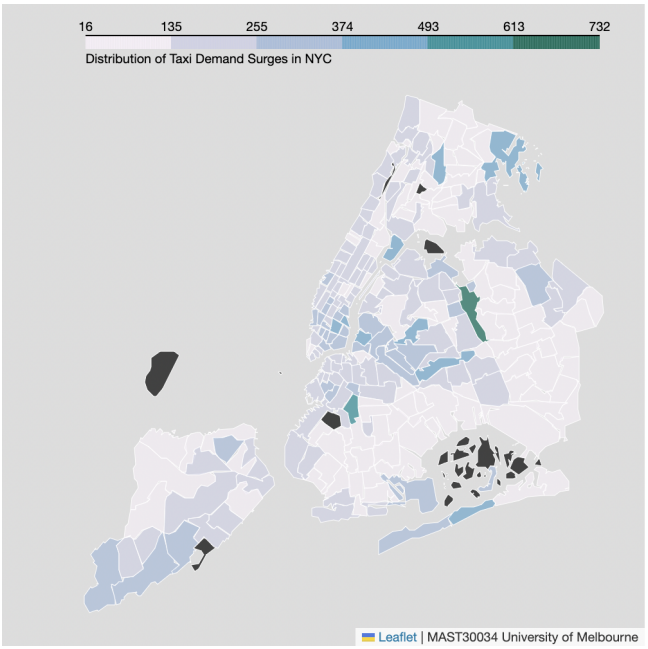

Figure 6: Locations with Most Frequent Surges

4 Modelling

4.1 Chosen Factors

To ensure the model is practically useful, we include readily available factors to model the demand in a particular location.

- Hour of Day and Day of Week (One-Hot Encoded)
- Number of Businesses in the respective zone
- Weather (Temperature, Relative Humidity, Rain, Snowfall, Wind Speed).
- Is Airport or not
- Borough (One-Hot Encoded)

4.2 Assumptions

This model makes some inherent assumptions, that should be considered while interpreting the output of these models, and in practical usage.

- We treat the hour of the day as a continuous variable, despite it being inherently cyclic in nature, to avoid feature space explosion and thus, over-fitting.

- We assume the size and distribution of businesses remain stable over time, and thus influence demand in predictable ways.
- Lastly, since zone-wise weather data is not available, we assume that the data collected at Central Park applies across New York and there aren't major variations within the city. Further, we assume that future weather predictions are reasonably accurate for the practical usage of the model.

4.3 Chosen Models

Three Machine Learning models, namely Linear Regression (LR), Decision Tree Regressor (DTR) and XGBoost (XGB) were implemented to model the rideshare demand.

Firstly, Linear Regression (with L2 regularisation) was chosen due to its simplicity and interpretability. Since we standardise the data, we can look at the respective regression coefficients as an indicator of the relative importance of each feature.

Next, we move to some more complex models to gain a higher accuracy. App-based ride share businesses are known to not care about model interpretability over performance, and the loss of simplicity would be acceptable in most cases.

Since we have some important categorical predictors, we can use a Decision Tree Regressor to effectively model their impacts separately.

XGBoost, a widely recognized model for exceptional performance on tabular data, was included due to its superior performance in complex datasets[9]. It is an enhanced version of the Gradient Boosting Algorithm and includes L1 and L2 regularisation[9].

5 Discussion

In this study, we chose Mean Absolute Error (MAE) as our performance metric. While Root Mean Squared Error (RMSE) is the metric used by training algorithms, we use MAE for practical inferences as it enables easier interpretation and comparison.

From Table 2, we see that while all models perform reasonably well, XGBoost significantly outperforms Linear Regression and Decision Trees in all boroughs, indicating a high complexity in the problem. With an error of 36 rides in bustling areas like Manhattan with thousands of rides every hour, and even less for other boroughs, we see that our model is largely successful in modelling the complexity of taxi demands in New York, and is suitable for practical applications.

From Figure 7, we observe that Linear Regression and Decision Tree underfit on the pickup hour of day feature, which might be highly affecting their accuracy. Meanwhile, XGBoost is highly successful in modelling hourly fluctuations, with some bias towards underestimating the demand, which may be due to the generally increased demands in June, as per Figure 13 (Appendix). If this is the case, models trained on larger datasets spanning multiple years may be able to perform even better.

Figure 7: Errors on Test Data by Hour

Borough	MAE LR	MAE DTR	MAE XGBoost
Bronx	26.87	23.80	**11.55**
Brooklyn	67.65	55.99	**20.40**
Manhattan	79.79	68.44	**35.99**
Queens	38.98	35.09	**13.39**
Staten Island	8.94	9.31	**3.81**

Table 2: Performance of Models on Test Data

Lastly, we analyse the errors by location using Figures 8, 9 and 10, and we see that our models, especially XGBoost are highly reliable in most zones across NYC, but face trouble in some specific zones and airports. Notably, from Figure 2, the demand is much higher in these areas, and so even if the numeric error is high, the relative error is still very low.

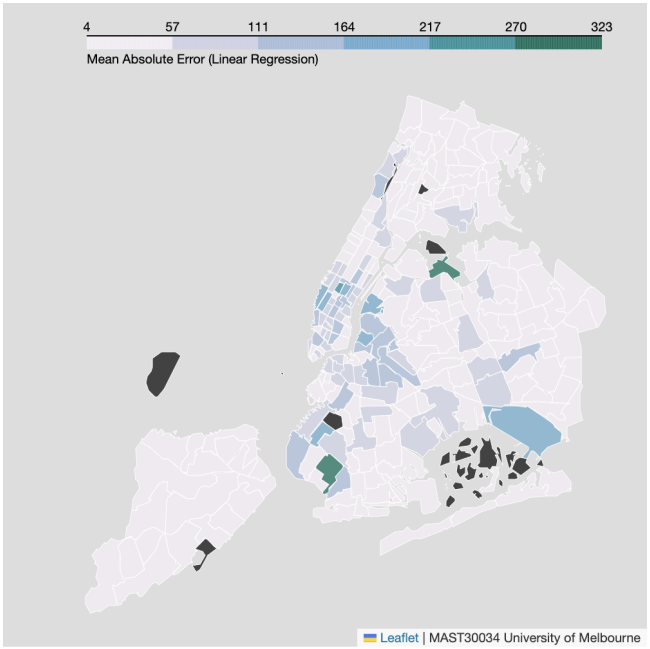

Figure 8: Linear Regression Errors

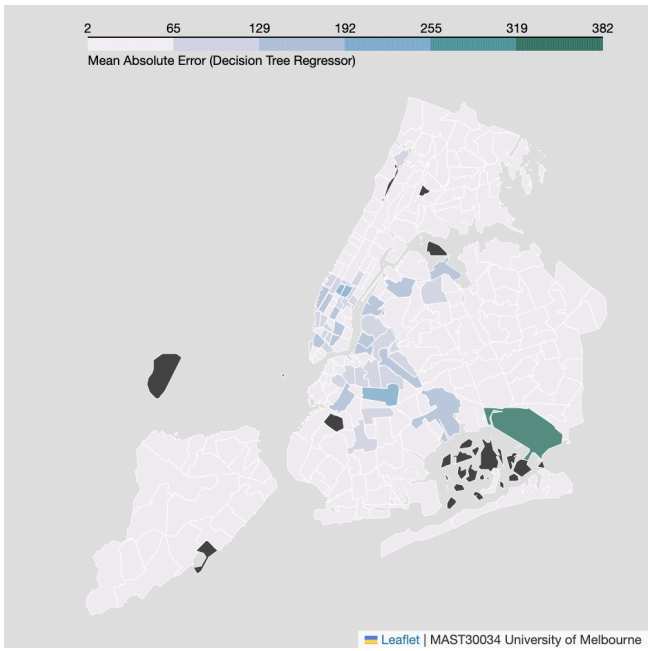

Figure 9: Decision Tree Regressor Errors

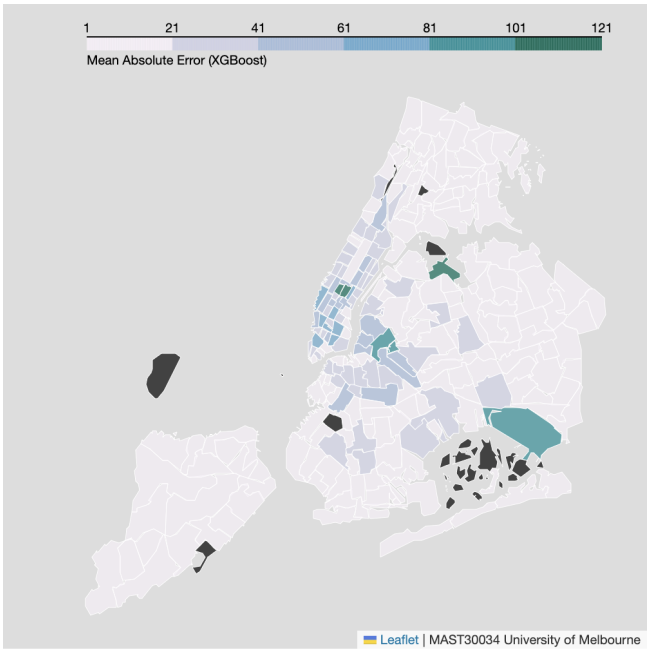

Figure 10: XGBoost Errors

6 Recommendations

As analysed above, the XGBoost model is very effective in modelling demand in most areas of New York. This can be used in real-time by ride-share companies to reorganise their fleet to meet the varying demand across the city. Their app-based nature allows them to send out notifications to divert drivers to high-demand locations and enhance user satisfaction and driver earnings. While this might increase cruising miles, the assurance that movements are made in the right direction helps negate their impact. Next, this can allow strategic management of driver rest times as companies can estimate the demand and allow excess drivers to rest. This can also help in electric car adoption as this makes strategic charge time scheduling possible. Lastly, these companies can invest in keeping a close eye on events in areas with high surge probabilities to predict them effectively and get the fleet prepared around those areas.

7 Conclusion

From our analysis, we see that the XGBoost model is highly effective at predicting the demand for most areas in NYC, especially with an overall mean of 95.63 and a very high standard deviation of 105.67 for the target variable. The external dataset of legally operating businesses helped us effectively model the demand by zone, which would have otherwise been a challenging task, as the amount of activity in different zones varies a lot across the city. In addition, the weather data helped us with the finer details of prediction.

For future improvements, companies may want to create separate models for airport demands as they might rely on airport flight schedules much more than the factors in current research. Next, the companies may develop overhead models that use these demand predictions to choose the best actions for individual drivers, optimising placement and rest times.

References

[1] City of Melbourne. *Microclimate Sensor Readings*. https://data.melbourne.vic.gov.au/Environment/Microclimate-Sensor-Readings/u4vh-84j8. Accessed: 2022-08-01.

[2] Jessica Phillips. "How Uber's dynamic pricing model works". In: *Uber Blog* (). URL: https://www.uber.com/en-GB/blog/uber-dynamic-pricing/ (visited on 08/25/2024).

[3] Lyft. *Ride pricing and charges*. https://help.lyft.com/hc/ru/all/articles/115012925707-Ride-pricing-and-charges. Accessed: 2024-08-25.

[4] NYC Taxi & Limousine Commission. *TLC Trip Record Data*. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2024-08-25.

[5] Jessica Bursztynsky. "Lyft shares drop 30% on disappointing guidance". In: *CNBC* (May 4, 2022). URL: https://www.cnbc.com/2022/05/04/lyft-shares-plunge-29percent-on-disappointing-guidance.html (visited on 08/25/2024).

[6] NYC OpenData. *Legally Operating Businesses*. https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh/about_data. Accessed: 2024-08-25.

[7] NYC Planning. *The Ins and Outs of NYC Commuting*. https://www.nyc.gov/assets/planning/download/pdf/planning-level/housing-economy/nyc-ins-and-out-of-commuting.pdf. Accessed: 2024-08-25.

[8] Patrick Zippenfenig. *Open-Meteo.com Weather API*. 2023. DOI: 10.5281/zenodo.7970649. URL: https://open-meteo.com/.

[9] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

Appendix

This section includes additional plots that weren't covered in the main analysis of the report but might be of further interest.

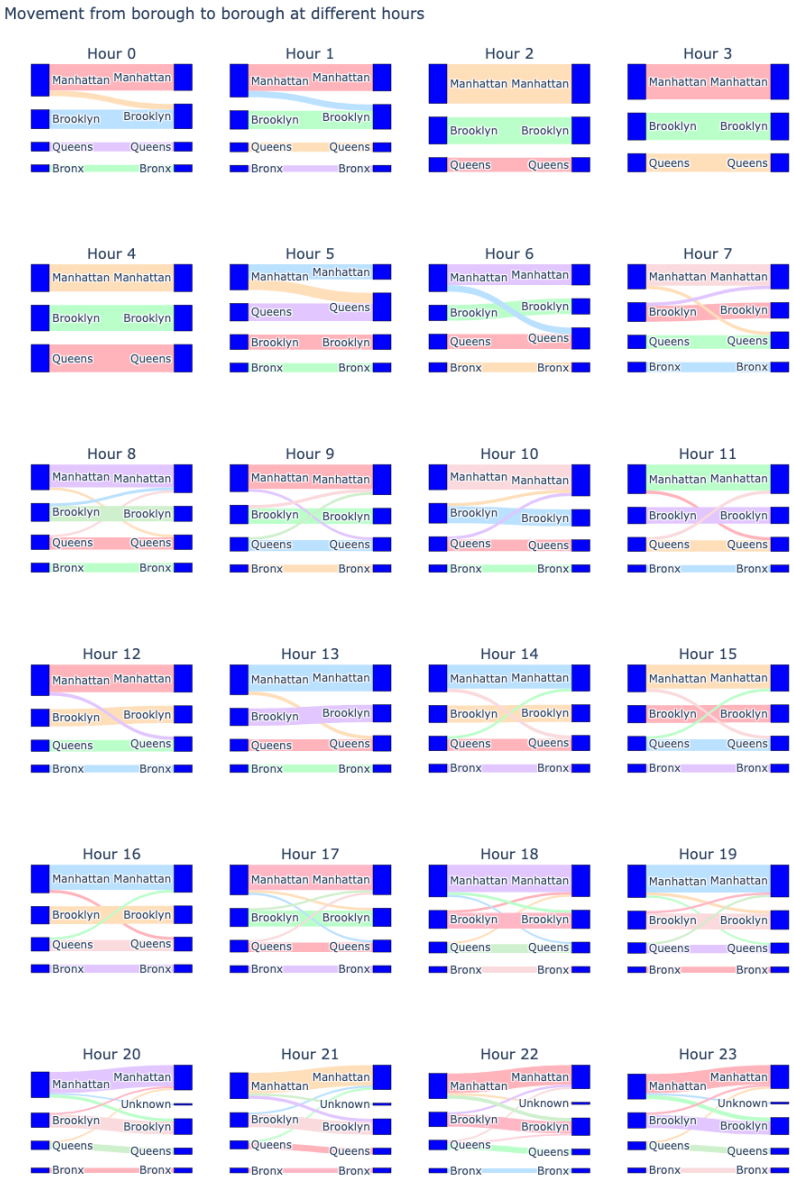

Figure 11: Cross borough movements in NYC, across the day

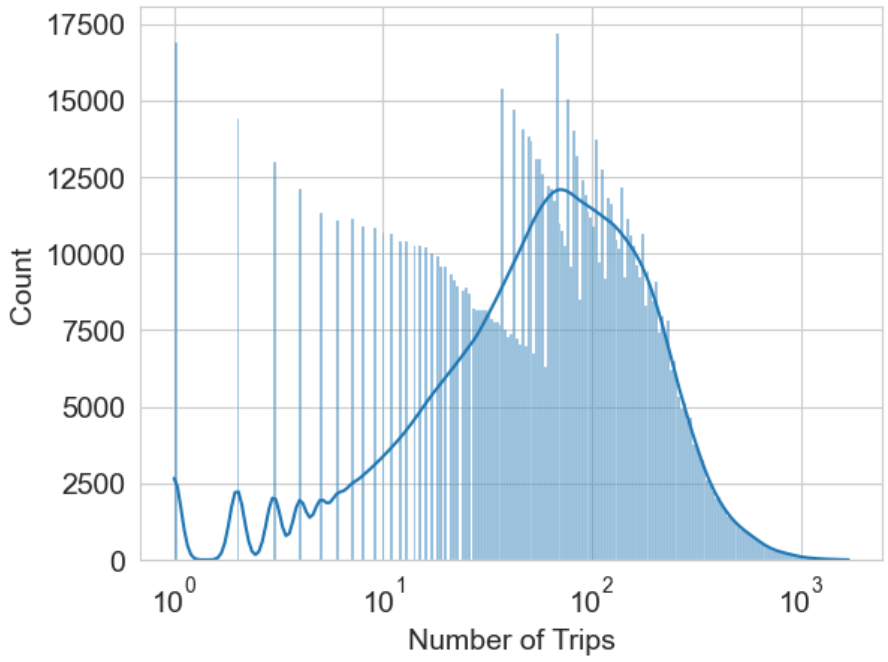

Figure 12: Distribution of Label - Number of Trips in Zone per Hour

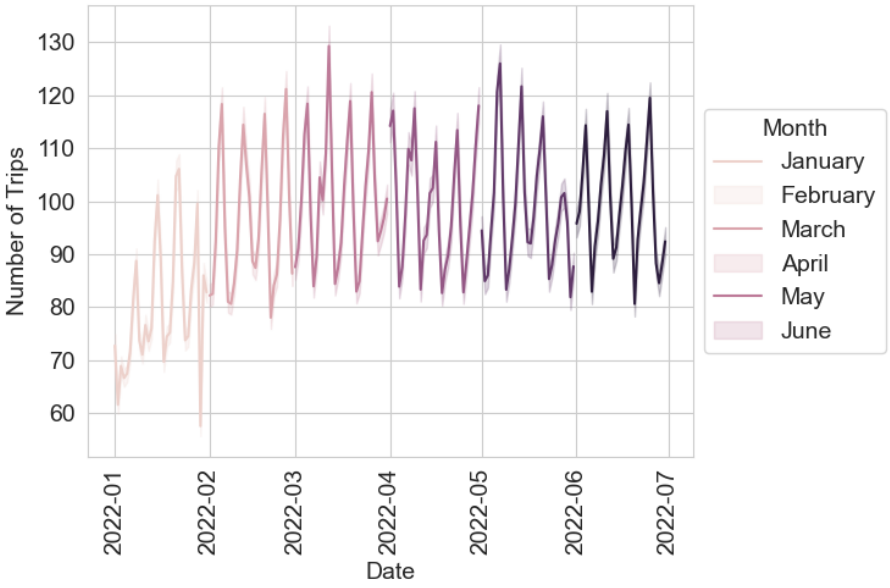

Figure 13: Demand Timeline