Predicting Hourly Ride-Sharing Demand in New York City: Towards a More Sustainable City

Uma Barnes
Student ID: 1277919
Github repo with commit

August 25, 2024

1 Introduction

Ride-sharing services such as Uber and Lyft have rapidly gained popularity in New York City (NYC), surpassing traditional taxi services - including both green and yellow taxis in total trips taken daily (Schneider, 2022). While these services offer convenience and efficiency, they also pose significant environmental challenges. The cruising miles or time drivers spend waiting for or traveling to pick up passengers, results in substantial and unnecessary carbon emissions, with approximately 40% of their service time spent waiting or en route to a pickup. Studies have shown that ride-sharing services contribute nearly 70% more carbon emissions compared to trips that would have been taken by public transport, cycling, or walking (Anair et al., 2020).

With increasing global temperatures, more frequent extreme weather events, and rising sea levels, the need for effective climate action has never been more critical. Transportation is a major contributor to greenhouse gas emissions, and the continued growth of ride-sharing services exacerbates the challenge of reducing our carbon footprint.

Given these concerns, it is crucial to explore the reasons behind why so many people are opting for ride-sharing services over more sustainable transportation options, such as public transport. This report aims to address this issue by investigating the factors influencing the choice of ride-sharing services and assessing their environmental impact.

We will employ two different machine learning approaches to estimate hourly demand for ride-share trips throughout New York City. By analysing demand patterns, this report seeks to provide valuable insights into the factors driving ride-sharing preferences. These insights are essential for policymakers, environmental advocates, and local transportation authorities, who can use them to develop strategies to mitigate carbon emissions and promote more sustainable transportation alternatives.

1.1 Dataset

This report uses data published by the NYC Taxi and Limousine Commission (NYCTLC), which contains taxi and for-hire vehicle (FHV) trips, including details such as pickup and drop-off times and locations across New York City (NYC) (NYC Taxi and Limousine Commission, 2024) which will be used to calculate hourly rideshare demand. Rideshare vehicles were chosen over green and yellow taxis given their average number of trips massively outweighs that of taxis (Schneider, 2020). We will also utilise another file by the NYCTLC which includes data detailing the shape of all location regions in the city, which will be utilized for visualisation purposes.

To enrich our analysis, we incorporated two additional external datasets. The Metropolitan Transportation Authority (MTA) Service Alerts dataset (Metropolitan Transportation Authority, 2024) includes information on both scheduled and unscheduled disruptions within the public transport network, including subway and bus services . This dataset will be used in an attempt to capture how public transport delays and disruptions impact upon rideshare demand and subsequently what improvements need to be made to these services.

Furthermore, weather data from the US National Centers for Environmental Information (NCEI) (National Centers for Environmental Information, 2024), collected at Central Park, NY, is utilized to investigate how weather conditions such as precipitation, wind, and temperature affect hourly rideshare demand. For instance, it is expected that poor weather might increase people's preference for rideshare services, which offer door-to-door convenience compared to public transport, which involves transiting from the station/bus stop to their destination, also known as last mile transportation (Kåresdotter, et al., 2022)

2 Preprocessing

2.1 Data Wrangling and Feature Selection

2.1.1 TLC Data

- **Time frame adjustment:** Records outside specified time frame 2023/07/01 to 2023/12/12 were removed from our dataset to align with the research requirements
- **Extremely short trip miles:** Entries with total trip distances of 0.5 miles or less were classified as outliers and removed due to their minimal travel distances.
- **Extended Trip Duration:** Trips exceeding 5 hours were deemed outliers and were removed from the dataset as such lengthy trips are unlikely to have occurred entirely within NYC given typical travel times of approximately 2-4 hours across the city and back.
- **Negative numerical records:** Records where driver pay, tips and other additional fee-related columns were negative were completely removed
- **Invalid Pick up Locations:** Trips with pick-up IDs outside of the range specified by the TLC (1-263) were considered to be outliers and were removed from the dataset
- **Extreme trip miles, trip time, driver pay, and other fee-related features:** These were removed by only including the 99.99 the percentile of the data. It was illogical to use the interquartile range method as a far too significant amount of data would be removed.

In total, we removed 1,187,482 records of our original TLC Data set within the specified time frame (1.2%) leaving 96,957,668 entries for our analysis. Whilst multiple of these features were not necessary for our analysis, it was imperative to identify outliers in these features to be able to deem whole records as outliers and subsequently remove them.

2.1.2 MTA Service Alerts

- **Irrelevant public transport routes:** Records involving alerts from the Long Island Rail Road (LIRR) and the Metro-North Rail Road (MNR) were removed from the dataset as the LIRR is not within our specified region of NYC and whilst the MNR does connect Manhattan with the Bronx primarily serves commuters regionally traveling to areas such as the Hudson Valley and Connecticut. MTA Bridges and Tunnels (BT) records were omitted as they were not the focus

of our research. Ultimately, we retained records for the Bus and Subway services, as these are the most frequently utilized modes of public transport within NYC.

- **Data Integrity Check:** Alert ID, Update Number and Event ID were also all checked to be non-negative. No records needed to be removed.

2.2 MTA Service Alerts Dataset

- Records which included alerts the Long Island Rail Road (LIRR) and the Metro-North Rail Road (MNR) were removed from the dataset as the LIRR is not within our specified region of NYC and whilst the MNR does connect Manhattan with the Bronx it is used more so to commute people to places like the Hudson Valley, and Connecticut. MTA Bridges and Tunnels (BT) records were also removed as they were not the focus of our research. The main Subway and Bus were included as these services are the most commonly used
- Alert ID, Update Number and Event ID were also all checked to be non-negative
- Columns which recorded affected lines/branches or bus routes , the status, Update Number and were removed

2.2.1 Weather Dataset

- **Feature selection:** The only features retained from the dataset were wind, temperature and dew as these were deemed to be most relevant for predicting rideshare demand.
- **Data Reformatting:** The selected features were initially recorded as comma-separated values within each element. We reformatted these entries to extract individual values for temperature, wind speed, and dew point. Additionally, these values were unscaled by dividing by 10, in line with the provided data dictionary
- **Missing Data Handling:** The data was also checked to see if there were any missing hourly records in the specified range. These records were then imputed using linear interpolation to ensure data completeness.

2.2.2 Feature Engineering and Data Visualisation

- **TLC Data:** For the TLC data hourly and daily demand was extracted from the data and placed into their own columns. Data was also aggregated based on the pickup location and then it's respective borough, so we are able to gain insights into the locational factors at play in rideshare demand. A column identifying a record as occurring on the weekend or weekday was also included as these are likely to significantly impact people's rideshare usage patterns.
- **MTA Service Alert Data:** In the MTA Subway Dataset, we also extracted hourly service alerts for both subway and bus services.

3 Analysis and Geospatial Visualisation

This section will examine the relationships between features within each individual dataset — TLC Data, MTA Service Alerts, and NY Weather Data — as well as the interactions between these datasets.

3.1 Location ID

As illustrated in Figure 1, the distribution of total trips by pick up location ID exhibited significant skewness, which a logarithmic transformation to normalize the data as shown in the accompanying figure.

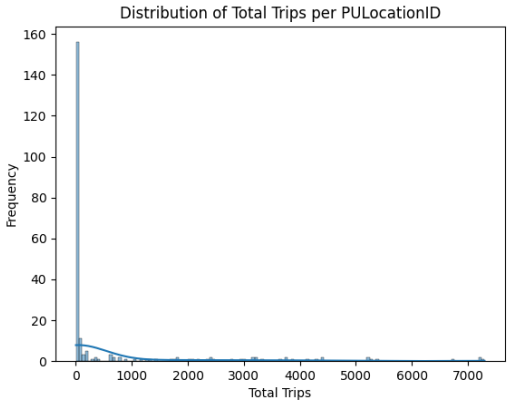

Figure 1: Distribution of Total Trips per Pick-up Location ID

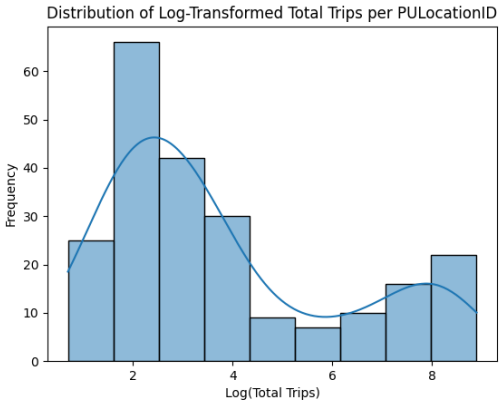

Figure 2: Distribution of Log-Transformed Total Trips per Pick-up Location ID

The chloropleth map of log-transformed pick-ups by location reveals a strong correlation between location and rideshare demand as can be seen in Figure 3. This is supported by Figure 4 which shows that This highlights that pick-ups are predominantly concentrated in Manhattan which is to be expected as this is the central business district and primary economic centre with a high population density. The analysis reveals a notable correlation between rideshare demand and locations within Manhattan. As expected, airports had notably high rideshare pickup numbers, driven by their role in transporting people arriving and departing from the city.

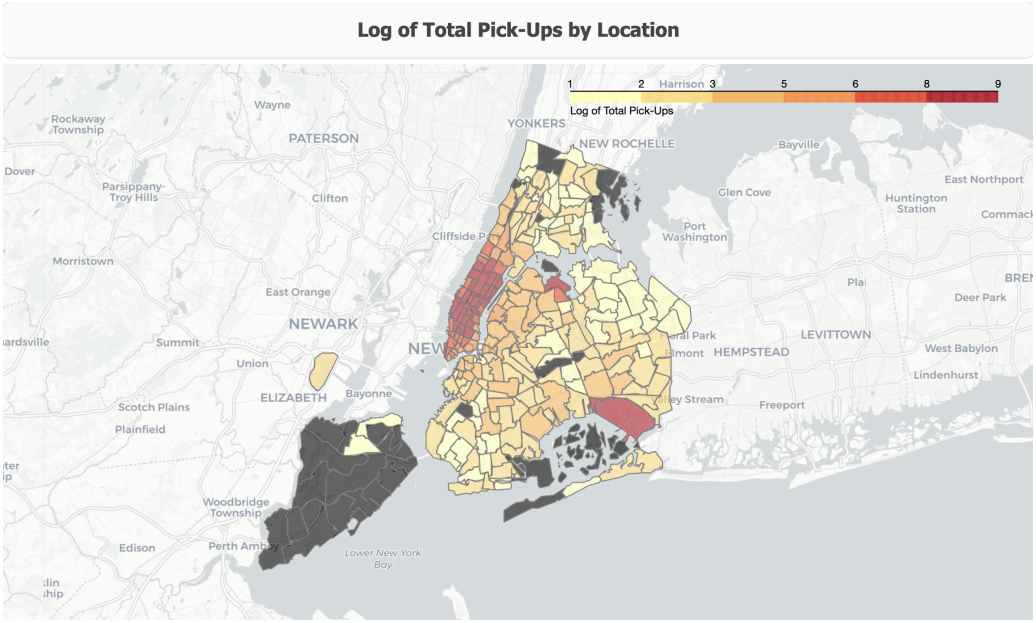

Figure 3: Chloropleth map of log of total pick-ups by location

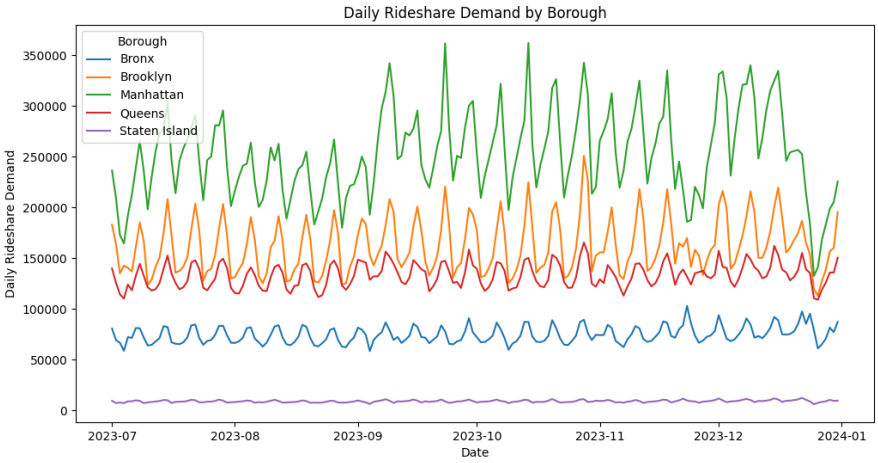

Figure 4: Graph showing daily rideshare demand by borough

3.2 Day of Week

There was also significant correlation found between the day of the week and the the total number of rideshare trips, with rideshare starting at its lowest point on Monday and gradually increasing throughout the week, with the highest demand typically occurring on Saturday, Friday and then Sunday. Additionally, Pearson's correlation was calculated for the hour of the day and rideshare demand (0.24) which suggested a potential correlation between demand and hour of the day.

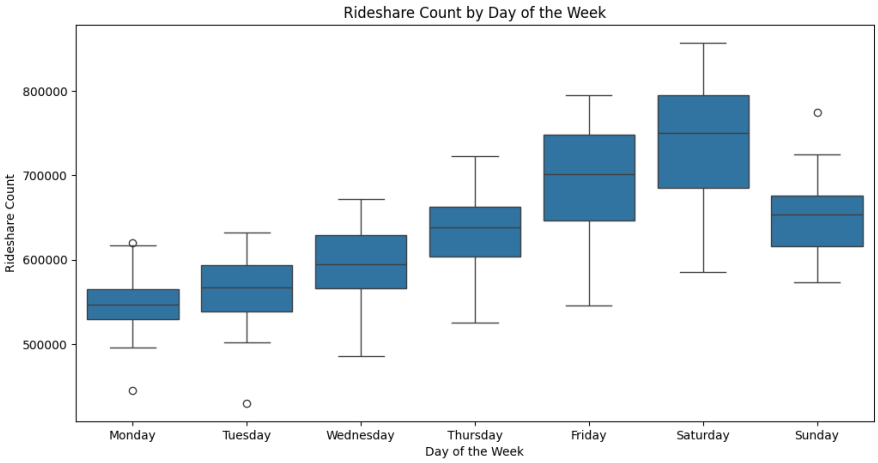

Figure 5: Boxplot comparing total daily rideshare rides for each day of the week

3.3 Weather

Due to the strong correlation between the temperature and dew point (with Pearson's Correlation Coefficient 0.89), temperature was retained and dew was removed. This choice not only reduces the dimensionality of the dataset, making it more efficient to process, but also helps avoid multicollinearity.

4 Modelling

Given the large volume of the dataset and the significant impact of location on demand, both Linear Regression (LR) and Random Forest Regression (RFR) were performed individually for each borough.

4.1 Linear Regression

Linear Regression is a fundamental algorithm used to model the relationship between a target variable and one or more features. The goal is to find the best-fit line (in the case of one feature) or hyperplane (in the case of multiple features) that predicts the target variable. We assumed no interactions between the predictor variables which resulted in the final model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \epsilon \tag{1}$$

where β_0 is the intercept, β_1 are the effects of the hour of the day, β_2 are the effects of the location, β_3 are the effects of the wind, β_4 are the effects for the temperature, β_7 are the effects of the amount of public transport disruptions and ϵ is the error.

4.2 Random Forest Regression

The supervised-learning ensemble method, Random Forest Regression was chosen for its ability to handle non-linear relationships and interactions between features more effectively. Given the potential complexity in predicting rideshare demand based on the factors considered, RFR is well-suited for capturing these intricate relationships.

It was thought that when predicting rideshare demand based on the aforementioned factors, a slightly more complex relationship might be underlying which RFR is suitable for. Additionally, it can assess feature importance, which helps us to understand which factors (e.g. weather conditions, hour of the day) are most influential in predicting demand. The ensemble nature of RFR is also more robust to overfitting and would be a good fit for the diverse charactistics of the data.

5 Results and Discussion

The performance of both the linear regression model and the Random Forest Regression model were assessed across different boroughs using the Root Mean Square Error (RMSE) which can be seen in Table 1. RMSE is a commonly used metric for assessing the accuracy of a regression model, where lower values indicate better model performance.

	RMSE	
Borough	**LR**	**RFR**
Manhattan	124.77	**86.73**
Queens	111.54	**58.15**
Bronx	44.37	**28.98**
Staten Island	13.23	**10.08**
Brooklyn	102.34	**62.47**

Table 1: Performance of models on the test set

In summary, the Random Forest Regression model demonstrated superior accuracy compared to the Linear Regression model across all boroughs. This suggests that the data exhibits complex

relationships that a simple linear model cannot effectively capture. The Random Forest model's ability to handle intricate interactions among features and its lower RMSE underscore its suitability for forecasting taxi demand in New York City, particularly in areas with varied and dynamic demand patterns.

Manhattan's comparatively poorer performance may be attributed to more intricate patterns of rideshare demand within the borough which were not accounted for in our model. As the most densely populated borough in New York City, Manhattan experiences a wide range of economic activities and a higher level of variability in demand. This complexity, driven by diverse economic activities and a highly dense population, results in more challenging and less predictable demand patterns, which can be difficult for models such as LR and RFR to accurately predict. However, considering the number of trips taken per hour can reach well above 10,000 especially in Manhattan, an error of around 83 trips per hour is sufficiently accurate.

6 Recommendations

Based on our modelling, we can conclude that RFR is a better suited approach for modelling hourly rideshare demand for rideshare services in New York, likely due to the complex non-linear nature of the relationships between the various factors and demand.

Manhattan and Brooklyn's evidently complex and high demand patterns, likely driven by dense populations and diverse activities, highlight the need for increased public transport frequency and capacity, particularly during peak hours. Enhancements could include extending subway hours, adding more trains and buses, and improving transit connections.

As mentioned previously rideshare vehicles often spend a significant amount of time cruising for passengers, which contributes to traffic congestion and higher emissions. To reduce cruising miles, real-time data could be used to guide drivers to areas with higher demand as predicted by our model, minimizing the need for unnecessary driving and consequently carbon emissions.

The correlation between service disruptions (e.g., delays and stops) and ride-share demand suggests that unreliable public transport pushes people toward ride-share options. The government should invest in infrastructure upgrades to minimize disruptions and delays.

Incorporating real-time data analytics to adjust public transport schedules based on conditions like weather and traffic could better match supply with demand. For example, during poor weather, increasing bus services or providing real-time updates on services can help encourage people to still use public transport.

Areas with lower public transport coverage may see higher ride-share demand due to the convenience of door-to-door service. Improve accessibility by enhancing last-mile connectivity (Kåresdotter et al., 2022) could involve expanding bike-sharing programs, introducing shuttle services in areas poorly served by public transport, and ensuring that all neighbourhoods have easy access to transit hubs.

Ultimately, ongoing assessment and monitoring of these strategies and of the modelling of rideshare demand will ensure they remain effective and responsive to commuter needs, promoting more sustainable transportation choices and reducing reliance on rideshare services.

7 Conclusion

This analysis revealed key factors influencing ride-share demand across New York City's boroughs, highlighting the impact of factors such as weather, public transport service disruptions, and location.

The findings suggest that improving public transport reliability, particularly in high-demand areas like Manhattan and Brooklyn, could effectively reduce the reliance on ride-sharing. Additionally, by optimizing rideshare operations to minimize cruising miles and guiding drivers to areas with higher demand, the city can decrease congestion and emissions.

References

[1] Anair, D., Martin, J., Pinto de Moura, M. C., & Goldman, J. (2020). *Ride-hailing's climate risks: Steering a growing industry toward a clean transportation future*. Union of Concerned Scientists. https://www.ucsusa.org/resources/ride-hailing-climate-risks

[2] Coordinated Public Transit-Human Services Transportation Plan for NYMTC Region • Final Chapter 3. Overview of New York City. (n.d.). https://www.nymtc.org/portals/0/pdf/CPT-HSP/NYMTC%20coord%20plan%20NYC%20CH03.pdf

[3] Kåresdotter, E., Page, J., Mörtberg, U., Näsström, H., & Kalantari, Z. (2022). *First Mile/Last Mile Problems in Smart and Sustainable Cities: A Case Study in Stockholm County*. *Journal of Urban Technology*, 29(2), 115–137. https://doi.org/10.1080/10630732.2022.2033949

[4] Metropolitan Transportation Authority. (2024, May). *MTA service alerts: Beginning April 2020*. https://data.ny.gov/Transportation/MTA-Service-Alerts-Beginning-April-2020/7kct-peq7/about_data

[5] National Centers for Environmental Information. (n.d.). *Global hourly data*. National Oceanic and Atmospheric Administration. https://www.ncei.noaa.gov/access/search/data-search/global-hourly?bbox=40.801,-73.983,40.767,-73.949&startDate=2021-10-01T00:00:00&endDate=2022-04-30T23:59:59

[6] New York City Taxi and Limousine Commission. (2024). *TLC trip record data*. New York City Government. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[7] Schneider, T. W. (2022). *Taxi and ridehailing app usage in New York City*. Todd W. Schneider. https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/