Massive Data Fundamental: NYC Yellow Taxi Demand Prediction

Ann Lian, Yuhan Ma, Wendy Shi

Introduction

There has been a growing trend of applying intelligent transportation systems and machine learning to enhance the sustainability and profitability of taxi services. For instance, Moreira-Matias et al. (2013) employed time-series forecasting techniques using GPS data to predict short-term taxi demand. Tong et al. (2017) took a different approach by implementing a simple linear regression model enriched with high-dimensional features, including meteorological data (e.g., weather, air quality, humidity) and point-of-interest information (e.g., spatial coordinates and hierarchical categories such as “playground,” “school,” and “hospital”).

Our research aims to develop an easily deployable machine learning model that accurately predicts taxi demand across New York City neighborhoods, supporting both urban mobility planning and private sector decision-making. Our analysis is grounded in the TLC Trip Record Data provided by the New York City Taxi and Limousine Commission, which includes comprehensive records of taxi pickups and drop-offs. For this study, we selected data spanning six months—from August 1, 2024, to January 31, 2025—to capture seasonal variation and emerging demand patterns, specifically in Manhattan.

The primary objective of this research is to build a predictive model that can inform real-time decision-making for transportation network companies (TNCs) such as Uber and Lyft. By identifying zones with elevated demand on specific days, this model could help optimize the allocation of drivers, reduce passenger wait times, and improve operational efficiency. In a broader policy context, our approach can also support city agencies in understanding spatial and temporal mobility trends, which may be used to inform congestion pricing, public transit planning, and equitable service distribution across underserved neighborhoods.

Data Collection & ETL

Overview of Raw Data

The project uses NYC yellow taxi trip record data. The dataset contains detailed records of individual yellow taxi trips in New York City. Each row corresponds to a completed trip and includes key operational, temporal, and spatial attributes such as pickup and dropoff timestamps, passenger count, trip distance, fare amount, and pickup/dropoff locations (by Taxi Zone ID).

For the scope of this project, we filtered the dataset to include only **yellow taxi trips with pickups in Manhattan between August 2024 and January 2025.** This specific focus is motivated by three factors:

- Yellow taxis can pick up and drop off passengers anywhere in NYC without restrictions, unlike green taxis, which are not allowed to pick up street hails in Manhattan south of 96th Street on the East Side, 110th Street on the West Side, and at LaGuardia or JFK airports.
- The time range we choose covers both the seasonal variance (e.g., back-to-school, holidays, winter weather) and weekday/weekends patterns, at the same time controlling our cloud budget at a fair amount.
- The Manhattan area accounts for the majority of taxi activity in NYC. Focusing on this reduces geographic noise.

In addition, our project aims to predict the taxi demand, which is tied to the pickups. This approach aligns with the research of Gangrade et al. (2022), where demand is measured based on the number of pickups in a given area and time interval. Pickup events are a direct indicator of rider-initiated requests, representing where and when taxis are actively being hailed or requested. In contrast, drop-off locations reflect where passengers choose to end their trips, which is influenced by individual trip purposes rather than the underlying demand for taxi services in that area. By focusing on pickup data, we can more accurately capture spatiotemporal patterns of demand.

Data Pipeline: Under Google Ecosystem

To better handle the scale and complexity of our project's data processing needs, we built our pipeline within the Google ecosystem.

Data Storage

First, we created Cloud Storage buckets and uploaded our dataset to **Google Cloud Storage**. After setting the necessary IAM roles—specifically, granting the *Storage Object User* role for upload access and *storage.buckets.list* permission to view buckets—we were able to use the Cloud Console's drag-and-drop interface to upload six monthly data files directly to our bucket from our local files.

Data Preprocessing

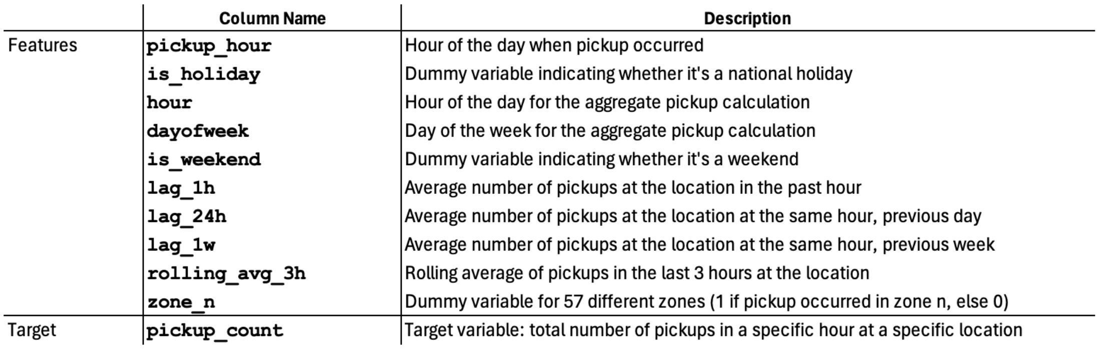
Secondly, we used **BigQuery** to perform data preprocessing for our machine learning pipeline. As Google Cloud's serverless, fully managed data warehouse, BigQuery is specifically designed to handle large-scale structured data with high speed and scalability, and can extract raw data sets from Google Cloud. It allows us to transform, filter, and aggregate data efficiently using familiar SQL syntax—without the need to manage infrastructure, memory, or compute resources manually.

Since our dataset is large, tabular, and based on historical (batch) records rather than real-time streaming, BigQuery offers an ideal solution. It supports advanced SQL features like **window functions** (`LAG`, `ROLLING AVG`, `TIMESTAMP_TRUNC`), which are essential for **time-series feature engineering**. Its distributed architecture allows it to **automatically partition and parallelize queries**, enabling fast scans of massive datasets.

All preprocessing tasks were completed using BigQuery SQL. The SQL script can be found in our GitHub repository.

The processed dataset has the following features and target.

Table 1: Features and Target



	Column Name	Description
Features	`pickup_hour`	Hour of the day when pickup occurred
	`is_holiday`	Dummy variable indicating whether it's a national holiday
	`hour`	Hour of the day for the aggregate pickup calculation
	`dayofweek`	Day of the week for the aggregate pickup calculation
	`is_weekend`	Dummy variable indicating whether it's a weekend
	`lag_1h`	Average number of pickups at the location in the past hour
	`lag_24h`	Average number of pickups at the location at the same hour, previous day
	`lag_1w`	Average number of pickups at the location at the same hour, previous week
	`rolling_avg_3h`	Rolling average of pickups in the last 3 hours at the location
	`zone_n`	Dummy variable for 57 different zones (1 if pickup occurred in zone n, else 0)
Target	`pickup_count`	Target variable: total number of pickups in a specific hour at a specific location

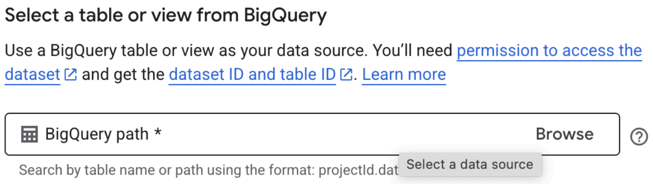
We then save the processed data locally as `Processed.csv` to perform data visualization.

Machine Learning

To build, train, and deploy our machine learning models, we used Vertex AI, Google Cloud's platform for end-to-end ML development. Vertex AI offers a powerful ecosystem for data scientists, allowing us to customize our machine learning or deep learning models without managing infrastructure. It also simplifies model management by automatically scaling compute resources—including CPUs, GPUs, and TPUs—based on the training requirements.

After preprocessing our dataset, we seamlessly imported the cleaned, structured data into Vertex AI to begin model training by selecting the data table from BigQuery (shown in the figure).

Figure 1: Example of importing Data from BigQuery



Select a table or view from BigQuery

Use a BigQuery table or view as your data source. You'll need permission to access the dataset and get the dataset ID and table ID. Learn more

BigQuery path *

Search by table name or path using the format: projectId.dat

Or we can import into our Python scripts:

```
from google.cloud import bigquery
# Load data from BigQuery
client = bigquery.Client()
query = """
     SELECT * FROM `your_project_id.dataset_name.table_name`
"""
```

Vertex AI supports native Python environments and popular libraries such as pandas, scikit-learn, TensorFlow, XGBoost, and NumPy. Using Vertex AI Workbench, we created notebook instances in JupyterLab and ran multiple models directly within the platform. Our complete machine learning scripts and workflows can be found in our GitHub repository.

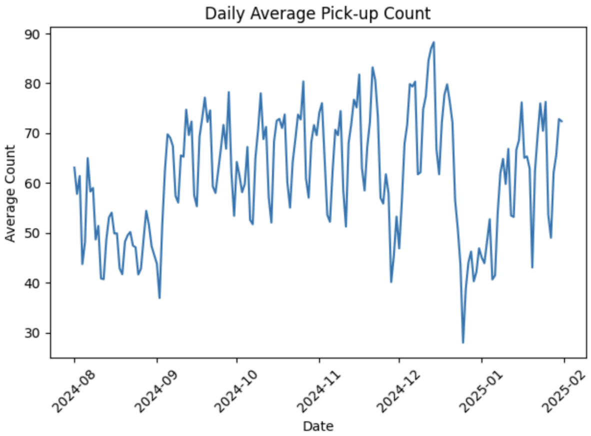
Exploratory Data Analysis

To support model development, we first conducted exploratory data analysis (EDA) to uncover key patterns in taxi demand across Manhattan. This step was essential for understanding the underlying structure of the TLC Trip Record Data and identifying meaningful trends that could inform predictive modeling.

Using six months of hourly pickup data (August 1, 2024 – January 31, 2025), we examined how demand fluctuates across time, space, and day types. Our analysis focused on three core dimensions: temporal variation, including seasonality and hourly trends; spatial concentration of pickups across zones; and demand intensity during peak hours. These insights guided the selection of features such as lag variables, calendar indicators, and zone-level identifiers, ensuring our model captures both routine and irregular demand shifts that affect transportation planning and ride-hailing operations.

Daily pickup volumes reveal clear temporal patterns that reflect broader societal rhythms. Taxi activity increased sharply in early September, aligning with the return of commuters and the start of the academic year. Demand remained elevated through the fall, supporting the notion of sustained weekday mobility needs. A noticeable decline occurred in late December, coinciding with the winter holiday season, followed by a gradual rebound in January. These trends underscore the impact of school calendars, holidays, and seasonal shifts on urban transportation demand.

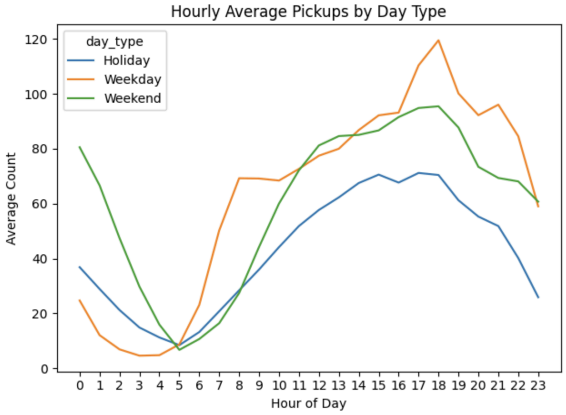
Figure 2: Daily Average Pick-up Count



Hourly trends further illuminate the structure of daily mobility behavior. On weekdays, demand begins rising shortly after 6 AM and peaks between 5–6 PM, consistent with traditional commuting hours. On weekends, activity starts later and peaks in the early evening (6–8 PM), likely reflecting leisure-related travel and nightlife. Holiday demand patterns closely resemble weekends but exhibit lower overall volume, suggesting that both work-related and discretionary travel are reduced during these periods.

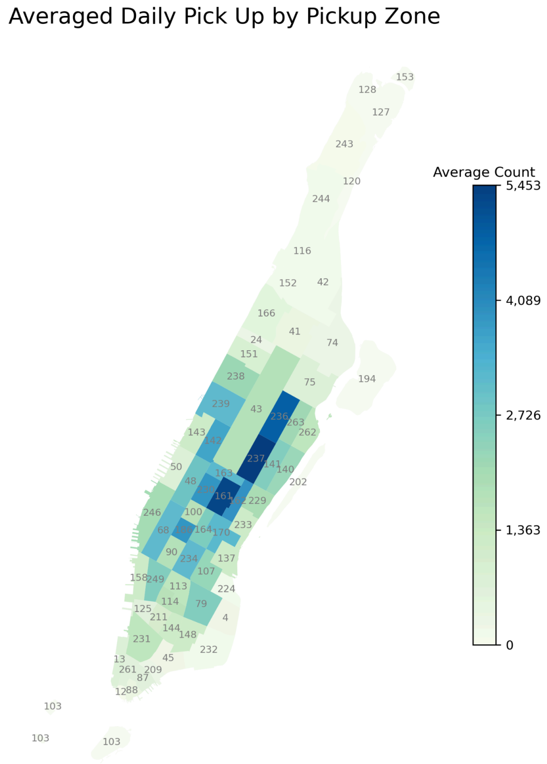
These temporal insights are central to designing a model that not only captures regular commuter flows but also accounts for nonlinear disruptions and seasonal demand shifts—a critical capability for supporting real-time decision-making by TNCs and long-term planning by city agencies.

Figure 3: Hourly Average Pickups by Day Type



Geo-spatial analysis

Figure 4: Geo-spatial Daily Averaged Pickup



As an extension of the exploratory data analysis above, we also perform geo-spatial visualization to help us understand where the hit pickup zones are located.

Taxi demand in Manhattan is highly uneven, with a small number of zones consistently accounting for the majority of activity. The map on the left shows the average daily taxi pickups by zone in Manhattan, with darker blue indicating higher demand. Zones such as 161 (Midtown Center), 237 (Upper East Side South), 236 (Upper East Side North) each average over 5,000 pickups per day, while more than half of all zones see fewer than 1,000. These high-demand areas cluster around Midtown, home to major transit hubs and tourist destinations like Penn Station, Grand Central, and Times Square, which attract large volumes of short, time-sensitive trips. In contrast, lighter zones such as 103, 4, and 127 represent areas with low pickup activity, likely due to low population density or limited accessibility (e.g., parks or islands).

These patterns are useful for both transportation network companies and urban planners. High-demand zones suggest where driver deployment should be prioritized, while low-demand areas might benefit from improved transportation access. From a modeling standpoint, zones with high baseline demand also tend to show greater prediction error, reinforcing the positive relationship between RMSE and average pickups observed in earlier analyses.

While the full-day map showed broader high-demand areas, this time-restricted view highlights the temporal concentration of taxi demand in commercial and transit-heavy zones. Peripheral areas remain lightly shaded, confirming that evening demand is focused in central business districts. This suggests that both model design and driver dispatch strategies should account for time-of-day variation in spatial demand intensity. Understanding these hourly patterns is key for optimizing resource allocation during peak travel periods.

EDA Summary

Our exploratory analysis reveals that taxi demand in Manhattan is shaped by clear temporal cycles and spatial concentration. Demand fluctuates significantly by time of day, day of week, and season, reflecting patterns tied to commuting behavior, leisure activity, and holiday schedules. Spatially, a small number of centrally located zones—particularly around Midtown and major transit hubs—consistently exhibit the highest pickup volumes, while peripheral zones see substantially lower demand.

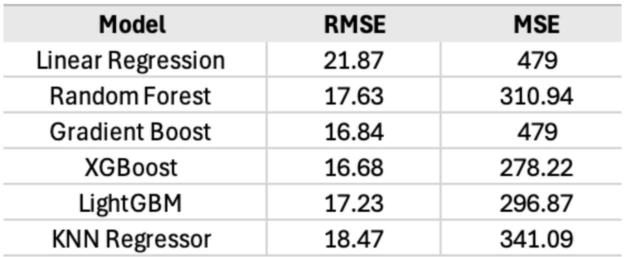
These insights directly informed our feature engineering strategy. By identifying peak periods, high-variance zones, and seasonal shifts, we were able to construct a modeling framework grounded in real-world mobility dynamics. The patterns uncovered through EDA highlight the importance of incorporating time-aware and location-specific features to accurately capture demand variability, laying a strong foundation for the predictive modeling stage of the project.

Model Evaluation

Model Results

We tested six machine learning models to predict average taxi pickups. XGBoost achieved the best performance, with the lowest RMSE (16.68), followed closely by LightGBM and Random Forest. Tree-based models clearly outperformed linear and distance-based methods, highlighting the importance of capturing nonlinear patterns in the data. Linear Regression had the highest error, while KNN showed moderate performance.

Table 2: Model Performance



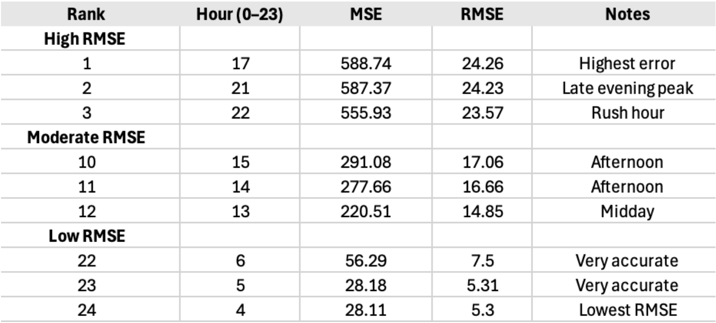
Model	RMSE	MSE
Linear Regression	21.87	479
Random Forest	17.63	310.94
Gradient Boost	16.84	479
XGBoost	16.68	278.22
LightGBM	17.23	296.87
KNN Regressor	18.47	341.09

RMSE Across Hours of the Day

We decided to divide the testing set by hour to evaluate how the model's performance varies throughout the day. As shown in the upper part of the table, the hours with the highest mean squared error (MSE) are 5 PM, 9 PM, and 10 PM, with MSE values approaching 25. However, as noted in the earlier exploratory data analysis (EDA), these hours also exhibit some of the highest hourly average pickup volumes, ranging from approximately 90 to 100 rides per hour.

Therefore, while the root mean squared error (RMSE) is highest during these periods, the magnitude of the prediction errors is relatively small when considered in the context of the overall scale of demand. In other words, despite the elevated RMSE, the predicted values remain reasonably close to the actual values, especially given the higher baseline of taxi activity during these peak hours.

Table 3: MSE and RMSE Across Different Hours



Rank	Hour (0–23)	MSE	RMSE	Notes
High RMSE				
1	17	588.74	24.26	Highest error
2	21	587.37	24.23	Late evening peak
3	22	555.93	23.57	Rush hour
Moderate RMSE				
10	15	291.08	17.06	Afternoon
11	14	277.66	16.66	Afternoon
12	13	220.51	14.85	Midday
Low RMSE				
22	6	56.29	7.5	Very accurate
23	5	28.18	5.31	Very accurate
24	4	28.11	5.3	Lowest RMSE

The middle part of the table shows hours where the RMSE is close to the overall average. These hours exhibit relatively high taxi demand but do not correspond to peak periods. The lower part of the table highlights the hours with the lowest RMSE. These periods also coincide with the lowest taxi demand of the day, resulting in true pickup values that are relatively close to zero across most zones. From the analysis above, we observe a positive relationship between RMSE and average taxi demand throughout the day.

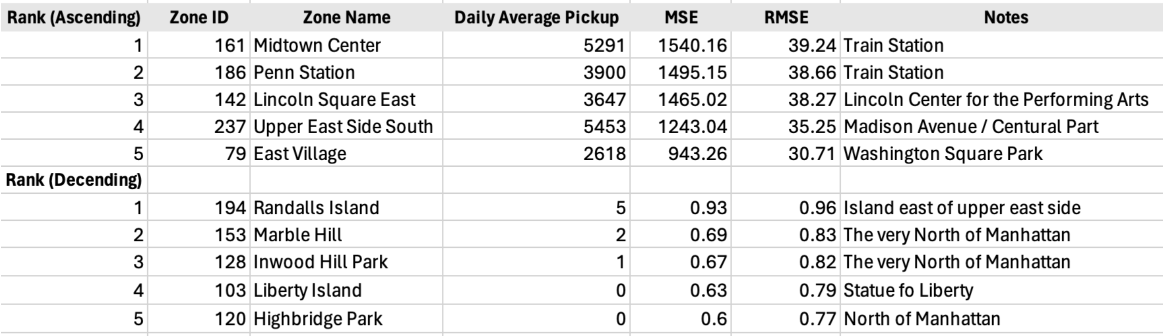
RMSE Across Pick-up Zones

We decided to divide the testing set by hour and analyze how model performance varies across different pickup locations. As shown in the table below, the locations with the highest MSE include Midtown Center, Penn Station, Lincoln Square East, Upper East Side South, and the East Village. These areas are either major transit hubs or popular tourist destinations. As such, taxi demand in these zones is more susceptible to external fluctuations, such as train arrival times or the end of performances and events.

Moreover, although these locations exhibit high RMSE values, they also have some of the highest daily average pickup volumes in New York City. This suggests that even small relative errors can result in larger absolute deviations in areas with high baseline demand.

We also examined the locations with the lowest RMSE. These include Randall’s Island, Marble Hill, Inwood Hill Park, Liberty Island, and Highbridge Park. These zones are either relatively secluded or primarily accessible by boat, which contributes to their consistently low taxi demand. These zones are either relatively secluded or primarily accessible by boat, which contributes to their consistently low taxi demand.

Table 4: MSE and RMSE Across Different Locations



Rank (Ascending)	Zone ID	Zone Name	Daily Average Pickup	MSE	RMSE	Notes
1	161	Midtown Center	5291	1540.16	39.24	Train Station
2	186	Penn Station	3900	1495.15	38.66	Train Station
3	142	Lincoln Square East	3647	1465.02	38.27	Lincoln Center for the Performing Arts
4	237	Upper East Side South	5453	1243.04	35.25	Madison Avenue / Centural Part
5	79	East Village	2618	943.26	30.71	Washington Square Park
Rank (Decending)						
1	194	Randalls Island	5	0.93	0.96	Island east of upper east side
2	153	Marble Hill	2	0.69	0.83	The very North of Manhattan
3	128	Inwood Hill Park	1	0.67	0.82	The very North of Manhattan
4	103	Liberty Island	0	0.63	0.79	Statue fo Liberty
5	120	Highbridge Park	0	0.6	0.77	North of Manhattan

Extensions

This project developed a machine learning model to predict taxi demand across New York City neighborhoods using temporal and spatial features derived from TLC Trip Record data. By incorporating variables such as day of the week, hour of the day, pickup location, and multiple lagged demand indicators (3-hour, 24-hour, and 1-week), the model captured overall trends with reasonable accuracy. However, as revealed through our error analysis, high-demand zones like Penn Station and Midtown Center remain more challenging to predict due to their volatility and sensitivity to external factors.

Looking forward, there are several key areas for improvement. First, integrating external data sources, such as train arrival schedules for major transit hubs, could help account for irregular spikes in demand tied to intermodal transfers. Similarly, incorporating weather data (e.g., temperature, precipitation, snow) may enhance the model's ability to adjust for seasonal and short-term variability, especially during extreme conditions when taxi usage patterns deviate from the norm.

Additional feature engineering could include event-related data (e.g., concerts, sports games, parades) or real-time traffic congestion indicators, both of which affect supply and demand dynamics. Exploring non-linear models, ensemble methods, or even spatio-temporal deep learning architectures may also improve prediction performance in complex, high-variance areas.

Ultimately, building a more context-aware, adaptive model will better serve both private-sector actors like Uber and Lyft and public agencies aiming to optimize urban mobility and reduce congestion through data-informed decision-making.

Reference

Gangrade, A., Pratyush, P., & Hajela, G. (2022). Taxi-demand forecasting using dynamic spatiotemporal analysis. *ETRI Journal*, *44*(4), 624–640. https://doi.org/10.4218/etrij.2021-0123

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., & Damas, L. (2013). Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, *14*(3), 1393-1402.

New York City Taxi & Limousine Commission. (n.d.). *TLC trip record data*. NYC.gov. Retrieved May 8, 2025, from https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Tong, Y., Chen, Y., Zhou, Z., Chen, L., Wang, J., Yang, Q., ... & Lv, W. (2017). The simpler the better: a unified approach to predicting original taxi demands based on large-scale online platforms. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1653-1662).