**Ally DiStefano, John Menassa, Michelle Thomas, Laasya Nellore, Sai Manoj Gandra**

NYC TAXI Analysis  
mis 587 Final report

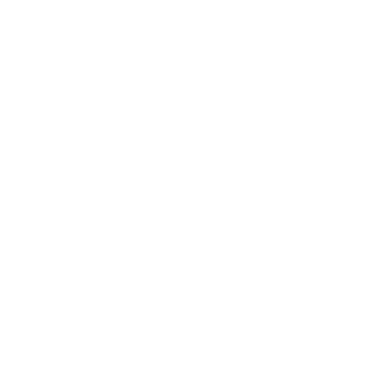


Table of Contents

[Introduction 1](#_Toc986256167)

[Data Warehouse Design 2](#_Toc1740405229)

[Data Preparation 4](#_Toc1496112915)

[Data Exploration 5](#_Toc981081404)

[Data Analysis 7](#_Toc50769397)

[Business Implications 9](#_Toc876206336)

[References 10](#_Toc1463016980)

[Appendix 11](#_Toc25191256)

# Introduction

In the ever-changing urban transportation landscape, taxi services are significantly affected by fluctuating demand based on time, location, and day of the week. Being a New York City taxi driver is a complex business with the rise of rideshares of Uber and Lyft and more people opting to bike. Currently, only 10% of the NYC driver landscape consists of taxis. The main demographic of taxi drivers are middle-aged immigrant men primarily from undeveloped countries. Many drivers work an average of 9.5 hours daily, six days a week.

Another excessive cost of being a NYC taxi driver is medallions. Medallions are limited licenses to taxi drivers that cost up to $1.4M in 2014. A NYC taxi driver needs to own or rent a medallion to work. During the peak of medallion prices, many taxi drivers took out loans to buy or rent a medallion. The price in 2022 was $134K. Many drivers are still paying off their loans and dealing with increased competition. Mastering the art of adapting to these dynamic patterns is critical to boosting profitability and customer satisfaction for taxi operators.

This project delves into the heart of this challenge, focusing on a pivotal question: How can taxi services in NYC better align their operations with varying demand patterns to enhance efficiency and customer experience? To tackle this question, we analyze a comprehensive dataset of NYC taxi trips from January to June 2023. Key attributes under scrutiny include pickup and dropoff times, locations, trip distances, fare amounts, passenger count, types of payment, and various financial parameters. Our goal is to uncover insights that pave the way for more intelligent, more responsive taxi services in the bustling streets of NYC.

Key stakeholders in this initiative include taxi companies, drivers, and regulatory authorities such as the Taxi and Limousine Commission (TLC). Their collective goal is to understand the nuances of taxi service operations, focusing on the factors that drive demand. This analysis is essential for introducing new operational strategies or programs. By examining patterns of pickups and dropoffs, stakeholders can uncover time-based trends, such as peak rush hours or specific peak days of the week affecting demand. This insight is vital for understanding the reasons behind demand fluctuations, which is critical to perfecting resource allocation.

# Data Warehouse Design

To perfect taxi service in NYC, our approach to data warehouse design begins with selecting the business process, which focuses on improving operations and customer satisfaction in response to varying demand patterns. The core of our analytical strategy is to declare the grain at the individual taxi trip level, meaning each record in our fact table corresponds to a single taxi trip, having all relevant details.

The next step in our design involves finding the dimensions crucial for analysis. These include the time dimension, encompassing attributes like AM/PM and Hour to analyze temporal patterns; the date dimension, featuring day, month, and year attributes for more date-specific insights; the pickup and drop-off location dimension, crucial for geographic analysis with attributes like Borough and neighborhood; and the payment type dimension, categorizing the various payment methods used by passengers.

Finally, the heart of our data warehouse is the fact table, where we combine the measurable and quantitative data from each trip. This table includes a broad range of fact measures such as trip distance, fare amount, tip amount, and total amount, along with specific charges like Metropolitan Transportation Authority (MTA) tax, congestion surcharge, and airport fee. It also incorporates essential foreign keys linking to the dimension tables, including TripID, TimeID, PULocationID, DOLocationID, and PaymentTypeID. This comprehensive design enables a multifaceted analysis of taxi services, paving the way for enhanced operational efficiency and customer satisfaction. Below is a detailed analysis of our data warehouse’s dimensions and fact tables.

Dimension Tables

These tables will have the descriptive attributes related to the dimensions we are analyzing.

Time Dimension Table

* TimeID (primary key, a unique identifier for each time record)
* Attributes: (AM/PM, Hour, QuarterHour, Shift, StartOfHour, TimeOfDay)

Date Dimension Table

* DateID (primary key, a unique identifier for each date record)
* Attributes: (Day, DayName, DayNameShort, DayOfWeek, DayOfYear, DayInMonth, Month, MonthName, MonthNameShort, StartOfMonth, EndOfMonth, Year)

Pickup and Dropoff Location Dimension Tables (structure is the same for both)

* PULocationID or DOLocationID (primary key)
* Attributes: (Borough, Service Zone, Zone, Longitude, Latitude, Airport)

Payment Type Dimension Table

* PaymentTypeID (primary key, numeric code showing how the passenger paid for the trip)
* Payment Type Description (description of the payment type)

Top of Form

Bottom of Form

Fact Table

The fact table will have the quantitative data from each trip that we want to measure or analyze.

Foreign keys: TripID, TimeID, PULocationID, DOLocationID, PaymentTypeID

Fact Measures:

* TransactionID – primary key, created during ETL process to uniquely determine each record
* VendorID: A code showing the TPEP provider that supplied the record (Only 2 vendors: 1 = Creative Mobile Technologies, LLC and 2 = VeriFone Inc.)
* RatecodeID: The final rate code in effect at the end of the trip (1 = standard rate, 2 = JFK, 3 = Newark, 4 = Nassau or Westchester, 5 = negotiated fare, 6 = group ride)
* Store and Forward Flag: This flag shown whether the trip record was held in vehicle memory before sending to the vendor (Y = store, N = not a store)
* Tpep Pickup Datetime: The date and time when the meter was engaged
* Tpep Dropoff Datetime: The date and time when the meter was disengaged
* Trip Distance: The elapsed trip distance in miles reported by the taximeter
* Trip Time Duration: Sum of ride durations to understand how long rides are during peak times or specific areas
* Passenger Count: Driver enters value of the number of passengers in the vehicle
* Fare Amount: The time and distance fare calculated by the meter
* Extra: Miscellaneous extras and surcharges of $0.50 during rush hour and $1.00 for overnight charges
* MTA Tax: $0.50 MTA tax that is automatically triggered based on the metered rate in use
* Tip Amount: Automatically populated for credit card tips. Cash tips not included.
* Tolls Amount: Total amount of tolls paid in trip
* Improvement Surcharge: Began in 2015, $0.30 surcharge assessed trips at the flag drop.
* Congestion Surcharge: If surge pricing is applied in congested traffic
* Airport Fee: $1.25 fee for driving to the high traffic area of the LaGuardia and JFK airports
* Total Amount: Sum of fares for all rides in a particular time/location/driver combinations to analyze revenue

The fact table structure for NYC taxi rides has foreign keys that set up connections to corresponding dimensions (Time, Date, Pickup Location, Dropoff Location, and Payment Type). It captures numeric metrics associated with high-demand periods, including durations, averages, and fees. These metrics offer valuable insights into operational efficiency during peak demand times. The organized data model facilitates a comprehensive analysis of historical ride data, enabling data-driven decision-making to enhance driver supply optimization and effectively address peak demand.

The current star schema focuses on a single fact table detailing individual taxi trips ([Figure 1).](#Fig1Star) However, to enhance analysis, additional fact tables could be introduced. For instance, a separate fact table for "Driver Shifts" might track data like shift duration, breaks, earnings, and tips per shift. This expansion allows for a more comprehensive analysis of operations beyond just individual trips.

In this taxi service context, a factless fact table can track occurrences or events, such as a table that records the occurrences of specific types of trips (e.g., trips during a special event in the city) without storing any metric like fare or distance.

Implementing of this data warehouse in SQL Server involves creating and managing these tables and ensuring the correct connections through foreign keys. SQL Server's capabilities, such as stored procedures, views, and functions, can be leveraged to query and analyze this data efficiently, providing real-time insights into the taxi service operations. For example, a stored procedure may automatically calculate and update daily revenue figures for each taxi based on the fare amounts and trip distances. By updating revenue figures daily, managers can monitor the financial performance of each cab in real-time. This immediacy allows for decision-making in response to emerging trends or issues, such as identifying underperforming vehicles or drivers. This data can support the implementation of incentive programs, identify training needs, or make operational adjustments to improve efficiency and profitability.

Overall, this data warehouse design provides a robust framework for capturing and analyzing the complexities of taxi services in New York. It enables stakeholders to make informed decisions based on a thorough understanding of various factors affecting taxi demand.

# Data Preparation

Our data preparation journey, a vital part of the ETL (Extract, Transform, Load) process, began with carefully selection of our dataset, followed by an extensive series of data cleaning and preprocessing steps to ensure data integrity and readiness for analysis. The initial phase involved converting each month’s data file from parquet to CSV format using Python, making it more suitable for our analytical tools. We then merged multiple monthly files into a single master file using Python to streamline our dataset and facilitate easier handling in subsequent stages.

Upon closer inspection, post-steps 1-3, we found missing values in the critical pickup and drop-off columns, pivotal for our location-based analysis. Faced with a dataset of considerable size, housing 19 million rows, we strategically chose to drop rows with missing values in these columns. Although this approach slightly reduced the dataset size, its impact on the depth of our information was negligible due to the dataset's vast volume. In handling missing data, real-life scenarios need a flexible approach, ranging from imputation to record removal, contingent on data quantity and criticality. The deliberate removal of incomplete records from our expansive data set ensured the integrity and completeness of our later analysis without compromising a significant loss of information, thereby keeping a robust and reliable dataset.

Moving forward, we delved into the parsing of date and time fields, originally combined in one attribute, an essential step in refining our dataset. This process involved converting the 'tpep\_pickup\_datetime' and 'tpep\_dropoff\_datetime' fields into separate date and time objects. Simultaneously, we extracted pertinent components such as hour, day of the week, and weekend status from the 'tpep\_pickup\_datetime' to create a Time dimension. We did the same thing for the 'tpep\_dropoff\_datetime' so we could utilize both the pickup and drop-off data for analysis.

Once our preprocessing in Python was complete, we imported our Masterfile into PowerBI. Under the transform data function, we meticulously converted each variable to the correct format, a crucial step in data preparation. This step included ensuring that all measures were in numerical formats, enabling them to be accurately averaged and summed across different dimensions. We introduced additional columns into our dataset to enhance our data analysis capabilities further. One notable example is the creation of a new column, a categorical value determined by the values in the airport fee column. If the airport fee exceeds 0, indicating an airport-related trip, the new column reflects this with a corresponding value, like ‘1’. This addition effectively flags all airport trips, facilitating a more focused analysis. With this column in place, we can effortlessly filter for, sort, and conduct targeted analyses on trips associated with airports, thereby streamlining our exploration of this specific segment of the data.

We concluded most of our data preparation process by transitioning to creating dimension tables enriching the dataset for comprehensive analysis. The time dimension table uses the extracted time components: hour, day of the week, weekend status, and holiday status. Similarly, distinct pickup and drop-off locations table formulations use location IDs and associated attributes like neighborhood, longitude, and latitude. Furthermore, we generated a payment type dimension table to map payment type codes to their respective descriptions.

In summary, our meticulous data cleaning process involved date and time parsing, extraction of time components for dimension tables, and judicious handling of missing values. We successfully transformed the dataset into a ‘clean’ state by addressing these facets, primed and ready for the ensuing data exploration and analysis phase. This refined dataset kept the original columns and incorporated newly created time-related columns, effectively mitigating the missing values challenge.

# Data Exploration

The dataset includes 19,000,000 rows and 20 columns. The next phase involves generating basic statistics for numerical columns and assessing missing values, providing insights into central tendencies, dispersions, and overall data quality. The focus will be on data exploration and understanding, specifically basic statistics for numerical columns.

Data Exploration and Understanding: Basic Statistics for Numerical Columns

TripID:

* Created during ETL process to uniquely identify each trip.

Passenger Count:

* Range: 0 to 6 passengers
* Mean: Approximately 1.56 passengers per trip.

Trip Distance:

* Range: 0 to 105.55 miles
* Mean: Approximately 4.05 miles per trip.
* Note: Some trips cover zero distance, which might need further investigation.

Fare Amount:

* Range: -346 to 454.5 USD
* Mean: Approximately 20.76 USD
* Note: Negative values suggest refunds or data entry errors.

Tip Amount:

* Range: -0.9 to 211.5 USD
* Mean: Approximately 3.39 USD
* Note: Negative tips might indicate data entry errors.

Total Amount:

* Range: -351 to 472.25 USD
* Mean: Approximately 29.17 USD
* Includes negative values, similar to fare amount.

Pickup Hour:

* Range: 0 (midnight) to 23 (11 PM)
* Mean: Most pickups occur around 7:70 AM.
* Missing Values and Data Quality

The dataset does not appear to have missing values in any column, which is a good indicator of data completeness.

Data Granularity and Completeness

The dataset's granularity at the trip level suits our analysis goals. It provides a detailed view of individual taxi operations, essential for understanding overall service patterns.

Data Quality Assessment

Our preliminary data quality assessment revealed anomalies, such as zero-distance trips and negative fare amounts. These irregularities might indicate data entry errors or specific use cases in taxi operations (e.g., canceled rides). These findings will be crucial for the detailed analysis phase, where we will investigate their implications further.

Preliminary Observations

The initial analysis has uncovered several interesting patterns. For example, the average number of passengers per trip is slightly above one, suggesting solo travel is predominant in our dataset. Additionally, the range of trip distances and fares is quite broad, hinting at diverse use cases for taxi services – from short hops to longer journeys.

After running summary statistics on our quantitative variables, we found that the average fare amount is $19.15, and the average trip distance is 4.08 miles. Understanding the relationship between these two variables can be critical for determining pricing strategies. For instance, if the average fare is relatively high compared to the trip distance, it might suggest that customers are willing to pay a premium for the convenience of the service. Conversely, if the fare seems low for the average distance, the service could be positioned as a cost-effective alternative to other modes of transportation.

Continuing with our data exploration, we delved deeper into the distribution of our variables, creating histograms to visualize the spread and central tendency of trip distances and fare amounts. The histograms revealed a right-skewed distribution for both variables, suggesting that while most fares and trips are relatively short and inexpensive, there are many more prolonged and costly outliers. In Python, we took a random sample of 1,000,000 rows and plotted these distributions [(Figure 2](#Fig2TripDist) and [3](#Fig3FareAmtDist)).

To further understand these outliers, we plotted them on a scatterplot against the time and day of the week. This analysis highlighted peak times and days where fares and trip lengths were notably higher, showing rush hours and weekend effects. We also noticed occasional spikes that correspond to dedicated events or holidays, which can drastically alter ride-hailing patterns.

We complemented our quantitative analysis with a qualitative assessment of the service zones. We could pinpoint locations with the highest demand by mapping ride density to specific areas. The visual representation confirmed that the central business district, nightlife hubs, and transportation hubs are hotspots for ride requests, which aligns with our understanding of urban mobility behaviors.

Initial findings suggest that certain areas, such as JFK Airport and the East Village, consistently appear as top taxi pickup locations. This insight could be due to their status as key transport hubs and popular urban areas. These insights form a basis for more in-depth geographical analysis later.

Finally, we conducted a correlation analysis to find relationships between the variables. We found an expected moderate positive correlation between trip distance and fare amount ([Figure 4).](#Fig4DistVsFare) However, we also uncovered less intuitive correlations, such as a negative correlation between ride density and average trip speed during peak hours, likely due to increased traffic congestion.

Future Directions for Analysis

Based on our exploratory findings, further analysis will focus on understanding the reasons behind the wide range of trip distances and fares and the implications of frequent pickups in certain areas. These insights will be pivotal for developing strategies to enhance taxi service efficiency and customer satisfaction.

Through this comprehensive data exploration, we have laid the groundwork for more advanced analyses, such as predictive modeling and optimization algorithms, to enhance service efficiency and customer satisfaction.

# Data Analysis

Upon completing the data exploration phase, we transitioned to analyzing our findings. Our data warehouse analysis reveals fascinating patterns in NYC’s taxi usage. A prominent trend is the surge in ride volume during the evening hours between 5 pm-10 pm and early mornings between 1am-4am ([Figure 5).](#Fig5TotRidePickupAvgDist) This peak aligns with the city's bustling nightlife and commuting patterns, suggesting that New Yorkers heavily rely on taxis for transportation during these times.

Furthermore, average trip distance mirrors this pattern, peaking during evenings and early mornings. This trend likely reflects longer commutes and journeys to social events, reinforcing the association between taxi usage and specific daily activities. Interestingly, the average trip distance steadily decreases as the day progresses, suggesting shorter journeys like errands or lunch breaks.

The data also reveals a distinct geographical distribution of taxi demand. Manhattan stands out as the hub of activity, with most rides originating and terminating within its boundaries. This finding is unsurprising, considering Manhattan's status as the most populous and densely populated borough, teeming with daily activity. Brooklyn and Queens, with their significant populations and major landmarks, also witness a high volume of taxi trips. Conversely, demand remains lower in the Bronx and Staten Island, likely due to their lower population densities and limited major attractions.

Weekend and holiday periods also see a spike in taxi demand, reflecting the increased travel activity associated with these times. This observation further emphasizes taxis' crucial role in NYC’s transportation ecosystem, catering to the city’s dynamic needs across different times and situations.

This information is crucial for taxi dispatching and staffing as it suggests the most and least busy times of the day, enabling more efficient resource allocation. Subsequent analysis focused on identifying key taxi service points ([Figure 6](#Fig6TopPickup) and [Figure 7](#Fig7TopDropoff)). The prominence of JFK Airport as a primary pickup point aligns with the expected high taxi demand at airports. This observation is further corroborated by the ride density map [(Figure 8).](#Fig8Map) Additionally, Midtown Center and Upper East Side South emerge as popular pickup locations due to their blend of residential, commercial, and entertainment offerings These areas are also noted as top drop-off points, highlighting their status as significant hubs for taxi activity in pickups and dropoffs.

This data provides valuable insights for dispatchers and fleet managers, enabling them to make informed decisions about stationing taxis during different times to maximize service efficiency. By focusing on areas with high demand, such as JFK Airport, Midtown Center, and Upper East Side South, taxi availability can be optimized, minimizing passenger wait times and improving overall service delivery.

These identified zones could be prime candidates for surge pricing during peak demand periods for taxi services utilizing dynamic pricing models. By implementing higher fares in areas like JFK Airport, Midtown Center, and Upper East Side South during peak hours, service providers can incentivize drivers to operate in these locations and ensure adequate availability for passengers. Furthermore, dynamic pricing can help regulate demand and manage resource allocation effectively, contributing to a more efficient and balanced taxi ecosystem.

Understanding the spatial distribution of taxi demand through data visualization tools like the ride density map allows for improved decision-making in crucial areas like dispatching, resource allocation, and pricing. By capitalizing on these insights, taxi services can effectively enhance operational efficiency, cater to passenger needs, and ultimately contribute to a more sustainable and user-friendly transportation system.

With all these findings, we crafted a dashboard in PowerBI to display these key metrics, providing a comprehensive overview of taxi service data [(Figure 9).](#Fig9Dashboard) The dashboard shows the average fare amount, $19.15, and an average trip distance of 4.08 miles. Most payments, at 81%, are made with credit cards. Rides peak in the early evening, with a significant morning surge in trip distance. The highest ride counts occur on Thursdays, while Mondays see the fewest. Manhattan consistently registers the highest demand among boroughs, with minor month-to-month variations in ride frequency.

After data exploration, the analysis revealed that taxi demand peaks in the early evening, likely due to work commutes and social activities, with a significant drop in the early morning hours. Demand gradually rises as the morning progresses. Additionally, trip distance is consistent but spikes in the early morning, suggesting airport commutes. Key pickup and drop-off points include JFK Airport, Midtown Center, and Upper East Side South, indicating these as high-traffic areas for taxis. This data is instrumental for optimizing taxi fleet distribution and could inform dynamic pricing strategies during high-demand periods.

Ultimately, the data-driven approach employed through the Power BI dashboard and data exploration provides a comprehensive understanding of taxi service dynamics. By leveraging these insights, taxi services can optimize their operations, enhance service delivery, and cater to the evolving needs of passengers, contributing to a more efficient and user-friendly transportation system.

# Business Implications

The exploratory analysis of the taxi ride data reveals several key insights into demand patterns. Most prominently, Manhattan has higher monthly taxi demand than other boroughs, suggesting a prime opportunity for the concentration of taxi fleets. Manhattan zones like East Village and Midtown experience increased demand and are central to entertainment and commercial districts. A weekday usage cycle appears with Thursday evening peaks potentially tied to post-workday demand and much lower Sunday rides reflecting weekend declines. Notably, demand significantly drops from 2 am to 5 am during the early morning hours.

The average trip distance of around 4 miles supplies a stable operational metric for factors like fueling and maintenance. Payment data shows that over 80% of rides are paid for by credit card, compared to only 17% with cash. Spatially, central hotspots Midtown Center and the airports generate huge volumes showing locations to distribute more taxis around potentially. The hourly, daily, and neighborhood-level demand fluctuations suggest taxi companies can significantly boost efficiency by dynamically adjusting operations, asset distribution, and staffing based on analytical findings.

The insights from exploring granular taxi trip patterns supply a foundation for data-driven strategies to improve customer service, responsiveness, and operational efficiency. Key business implications include the dynamic redistribution of taxi fleets to Manhattan zones with the highest demand to reduce wait times. Zone-based fleet optimization can use knowledge of neighborhood-level demand variances to achieve alignment of vehicle supply and driver staffing to when and where rides are most needed. Taxi companies can focus on driver incentives and rides during weekday rush hour or Thursday evenings when volumes spike. Improving operational efficiency includes fuel consumption planning, preventative maintenance operations, and parts inventory management using predictable mileage. The relative stability of an average trip distance of around 5 miles supplies useful indicators for improving operational efficiency and streamlining credit card payment systems to align with 81% transaction preference.

Specific applications could involve:

* Perfecting zone-based distribution.
* Implementing demand-based dynamic pricing surges.
* Developing predictive fleet maintenance models.
* Simplifying payment channels.
* Strategically distributing marketing by demand profiles.
* Leveraging data to forecast taxi volumes needed for major city events.

Data-driven automation and analytics applications present immense potential based on these demand insights. This includes predictive fleet maintenance applications forecasting part servicing needs based on miles traveled data. Dynamic pricing applications could institute temporary surge increases during high demand using automatic triggers - balancing supply and demand optimally. Similarly, marketing optimization algorithms can automatically shift budget allocations to smooth weekend leisure variability as analytical models prescribe.

With a comprehensive understanding of the neighborhood, hourly, and weekday variability in taxi demand, taxi drivers can boost reliability, profitability, and customer experience through analytical, insights-based resource planning, automated decision tools, and data-optimized execution. NYC can keep the iconic yellow taxis for decades to come.

# References

Baldwin, S., et al. (July 2023). *New York City taxis fight for survival against Uber and Lyft*. CNBC. https://www.cnbc.com/2023/07/15/new-york-city-taxis-fight-for-survival-against-uber-and-lyft.html

Litterst, R. (September 2022). *The* *ups and downs of NYC’s taxi medallion.* The Hustle. https://thehustle.co/09062022-nyc-taxi-medallion

Mujawar, I., et al. (August 2021). *Sleep behavior of New York City taxi drivers compared to the general US population*. National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9249005

# Appendix

[*Figure 1 – Star Schema*](#Bookmark1) *(click figure title to return to relevant paragraph)*

A screenshot of a computer screen

Description automatically generated

*[Figure 2 –Trip Distance Distribution](#Bookmark2)*

*A graph of a trip distance

Description automatically generated*

*[Figure 3 –Fare Amount Distribution](#Bookmark3)*

*A graph of a number of blue bars

Description automatically generated*

*[Figure 4 – Scatterplot of Trip Distance vs Fare Amount](#Bookmark4)*

*A blue dotted line with white text

Description automatically generated*

*[Figure 5 – Total Rides by Pickup Times and Average Distance](#Bookmark5)* A graph of a line

Description automatically generated with medium confidence

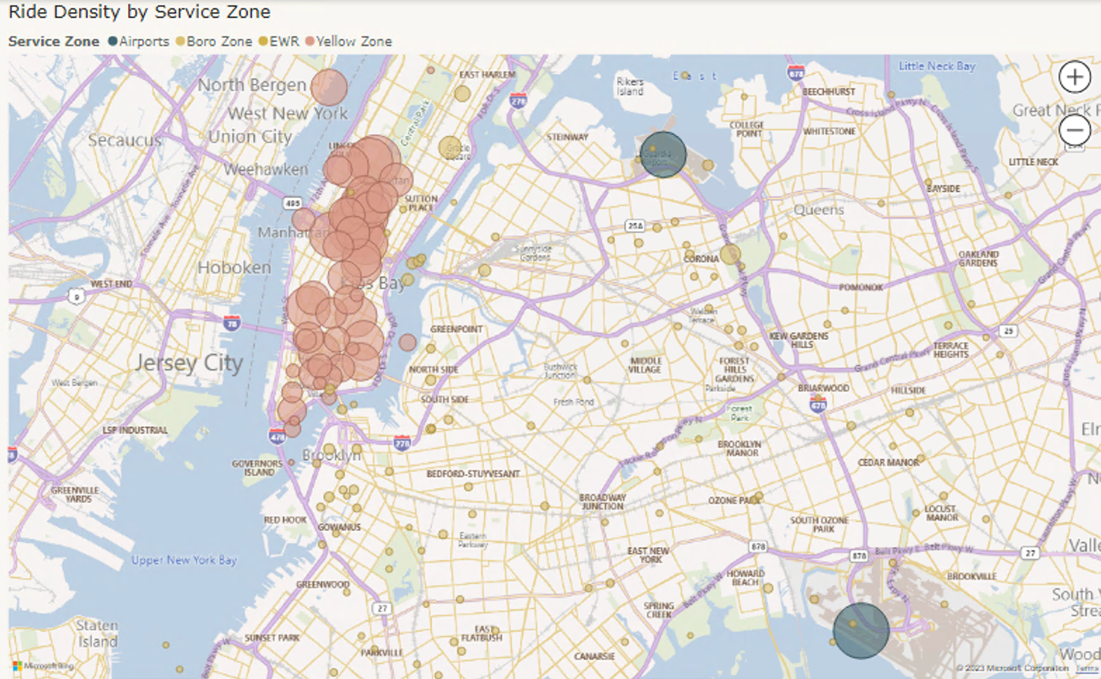
*[Figure 6 – Top 10 Pickup Locations](#Bookmark6)* A graph of a bar

Description automatically generated with medium confidence

*[Figure 7 – Top 10 Dropoff Locations](#Bookmark7)*

A graph of a graph

Description automatically generated with medium confidence

*[Figure 8 – Ride Density by Service Zone](#Bookmark8)*

*[Figure 9 – PowerBI Dashboard](#Bookmark9)*

A screenshot of a graph

Description automatically generated