G9: Spatio-Temporal Analysis of NYC Taxi Services: Trends and Insights

Member 1 Zhongren Zhao - 20354552, Member 2 Mingen Xu - 20276899,
Member 3 Yueyi Huang - 20322613, Member 4 Sijing Zhang - 20243635

1 Motivation and Problem Statement

New York City's taxi system has generated massive volumes of trip-record data over the past decade, offering a unique opportunity to understand urban mobility patterns at scale. By analyzing these data—from 2009 through 2024—researchers and policymakers can gain deep insights into passenger demand, fare structures, and spatial-temporal trends of the city's iconic Yellow and Green Cabs.

Despite the increasing prevalence of ride-sharing services, such as Uber, this project will focus primarily on leveraging the NYC Taxi dataset to explore core questions around demand fluctuations, pick-up/drop-off hotspots, and fare/policy impacts. Uber will serve only as a minor comparative branch: we will use select Uber metrics where relevant to illustrate how the emergence of app-based ride-hailing services might influence or compare to traditional taxi usage patterns.

By honing in on the foundational taxi data, this research aims to:

- Reveal important demand trends and geographic usage shifts in NYC's taxi market over the years.
- Propose data-driven approaches to enhance service efficiency, ultimately benefiting both passengers and the Taxi Limousine Commission (TLC).
- In addition, with the growing number of online car-hailing services, we will also incorporate Uber's NYC data for comparative analyses alongside traditional taxis to inform better strategies for the development of NYC's taxi system.

2 Research Questions and Methodology

2.1 RQ1: What is the pattern of taxi demand in New York City at different times of day and in different areas?

Motivation: Understanding the temporal and spatial distribution of taxi demand in New York City is essential for improving service efficiency, optimizing fleet allocation, and reducing passenger wait times. Taxi operators and urban planners need data-driven insights to adjust supply in response to demand fluctuations, ensuring better coverage in high-demand areas and reducing congestion in overserved locations.

While previous research has examined taxi demand patterns, most studies either focus on aggregated trends or static spatial distributions, without fully integrating time-dependent variations across different boroughs and peak periods. This research will bridge that gap by conducting a comprehensive spatiotemporal analysis, leveraging statistical modeling and machine learning techniques to uncover significant demand trends. The insights gained can guide both policy-making and operational adjustments.

Proposed Methodology: We use a method that integrates statistical analysis with time series modeling and geospatial techniques to study NYC taxi demand patterns. The methodology consists of the following steps:

Data Collection and Preprocessing

Pull NYC TLC taxi trip records and select important fields like pickup/drop-off locations together with timestamps and trip duration. Perform data cleaning and preprocessing operations to eliminate missing values along with outliers and inconsistencies from the dataset. Translate pickup/drop-off coordinates to taxi zones for organized spatial analysis. Exploratory Analysis and Visualization

Create heatmaps and time series plots in order to clearly show demand variations. Divide demand metrics according to different periods during the day including morning rush, midday, evening rush and late night while distinguishing between weekdays and weekends. Analyze demand variations between boroughs to pinpoint regions with high demand and underserved locations. Statistical and Predictive Modeling

Develop predictive models using ARIMA and LSTM for forecasting taxi demand across multiple regions during different time intervals. Apply spatial clustering methods like K-Means or DBSCAN to find regions with comparable demand patterns. Conduct regression analysis to determine how outside elements like weather conditions and public transportation availability affect the need for taxis. Hypothesis Testing and Key Insights

Hypothesis 1: The taxi demand pattern shows predictable daily and weekly peaks and troughs. Hypothesis 2: The locations with high-demand remain stable over time yet show spatial changes as urban development progresses and external influences emerge. Hypothesis 3: The demand patterns for taxis are heavily affected by external factors which include weather conditions along with economic activity and the accessibility of public transit. This research uses

statistical analysis along with machine learning models and spatial clustering methods to develop actionable insights about NYC taxi demand patterns which will enhance fleet management and urban transportation planning.

2.2 RQ2:Is there a significant relationship between trip duration and departure time across different boroughs?

Motivation: The traffic conditions in New York City show notable differences between boroughs and time periods. These variations stem from factors including population concentration, commuting behaviors, and roadway design. Such differences directly affect travel times, creating difficulties for taxi services in predicting accurate trip durations. Analyzing how trip duration correlates with departure time in specific boroughs is essential for enhancing driver allocation tactics, refining navigation strategies, and boosting passenger satisfaction.

Additionally, precise travel time predictions can decrease passenger disputes, facilitate equitable pricing calculations, and assist drivers in optimizing their work schedules. Earlier research has concentrated mainly on citywide traffic trends without detailed investigation of inter-borough discrepancies. This study seeks to fill that research void by systematically assessing how departure timing affects trip duration within New York City's boroughs, yielding practical recommendations for taxi businesses and city infrastructure planners.

Proposed Methodology: To analyze the relationship between trip duration and departure time across boroughs, the study will follow these steps:

Data Preparation

Extract relevant data fields from the NYC Taxi dataset, including pickup/drop-off locations, trip duration, and timestamps. Clean the dataset by removing outliers, such as extremely short or long trips, and filter for trips with valid timestamps and geographic data. Map pickup locations to boroughs (Manhattan, Brooklyn, Queens, Bronx, Staten Island) using taxi zone identifiers. Grouping and Segmentation

Divide the dataset into time periods (e.g., morning peak, midday, evening peak, late night) to capture temporal variations. Group trips by boroughs and time periods, creating subsets of data for more granular analysis. Exploratory Data Analysis (EDA)

Visualize the distribution of trip durations across boroughs and time periods using boxplots, histograms, and heatmaps. Investigate trends such as the average trip duration during different periods in each borough and identify outliers or anomalies. Statistical and Predictive Analysis

Employ a multiple linear regression model to quantify the impact of departure time and borough on trip durations. The model will include interaction terms to capture combined effects of borough and time period. Hypothesis: Trip durations are significantly longer during peak hours, especially in boroughs with high traffic congestion, such as Manhattan. Use ANOVA tests to determine if the differences between boroughs and periods are statistically significant. Machine Learning Approaches (Optional)

Apply random forest regression or XGBoost models to predict trip durations, incorporating additional features such as weather conditions and day-of-week indicators to enhance prediction accuracy.

Validation and Results Interpretation Model performance evaluation and results interpretation

Using R square value, average absolute error (MAE), and average square root error (RMSE) and other indicators evaluation model prediction accuracy

Interpret the impact of the average period and the schedule of the administrative division through the regression coefficient analysis (for example: Manhattan early peak peak period coefficient = 0.78, P ¡0.001)

Develop heatmaps and interactive dashboards to illustrate the connections between trip length, time of day, and the various boroughs. Offer advice to taxi companies regarding time-sensitive dispatch strategies, and to urban planners on reducing traffic bottlenecks in areas with significant impact. Hypotheses to be Examined

Hypothesis 1: Trip durations are noticeably longer during the morning and evening rush hours, especially in densely populated boroughs like Manhattan.

Hypothesis 2: Boroughs with sparser populations, such as Staten, exhibit less fluctuation in trip durations throughout the day.

Hypothesis 3: Outside influences, such as weather patterns and the accessibility of public transportation, may intensify the correlation between trip durations and departure times.

By integrating statistical analysis and machine learning techniques, this research aims to provide a thorough understanding of how trip lengths change depending on the time and location, yielding practical information for refining taxi services and the administration of urban transit.

2.3 RQ3: How has the widespread adoption of Uber in New York City influenced traditional taxi ridership and revenue patterns across different boroughs?

Motivation: Ever since Uber showed up in New York City, things haven't been the same for getting around. Taxis, which used to be the main way to grab a ride, are now up against these app-based services that are just so much easier for a lot of people. Though Uber, can order a car right from your phone, the prices change depending on demand, and they seem to be everywhere. Younger people, especially those who are always on their phones, seem to love it, but older folks or people who don't use smartphones as much might still rely on regular taxis.

It's really important for the city and the people in charge of transportation to figure out how Uber's presence is impacting taxis. They need to make sure things are fair for everyone, that people can still get around, and that resources are being used wisely. And, by looking at how Uber affects taxi use and how much money taxis are making in different parts of the city, we can learn about whether everyone has equal access to transportation and help planners fix any gaps in service, especially in areas that are already struggling. This research should help us understand the challenges taxis are facing and give some ideas on how to make things fairer for everyone.

Proposed Methodology: Proposed Methodology To analyze Uber's impact on traditional taxi ridership and revenue,

this study will use NYC Taxi and Limousine Commission (TLC) trip records alongside publicly available Uber datasets, focusing on matching time frames for consistency. The methodology involves the following steps:

Data Preparation

Traditional Taxi Data: Collect Yellow and Green Cab trip records, including pickup/drop-off locations, timestamps, fares, and trip durations. Uber Data: Use publicly available Uber trip records, including trip volumes, timestamps, and spatial distribution of rides. Data Cleaning: Address missing values, remove anomalies, and align both datasets to the same time intervals and geographic boundaries (e.g., taxi zones or boroughs). Temporal and Spatial Segmentation

Segment trip data by boroughs (e.g., Manhattan, Brooklyn, Queens) to study regional disparities. Divide data into time periods (e.g., peak hours, off-peak hours, weekdays, weekends) to capture temporal trends in ride-sharing and taxi usage. Exploratory Data Analysis (EDA)

Generate heatmaps to visualize shifts in demand and revenue across boroughs. Compare ride volumes, average fares, and trip durations between traditional taxis and Uber over time. Statistical and Predictive Modeling

Regression Analysis: Examine the relationship between Uber's market penetration and traditional taxi ridership. Analyze the effect of external variables on both services. Time-Series Analysis: Identify long-term trends in taxi revenue and ridership as Uber's adoption increases. Detect seasonal or cyclical patterns in taxi usage that align with Uber's expansion. Hypothesis Testing

Hypothesis 1: Uber's adoption leads to a significant decrease in traditional taxi ridership, particularly in high-density areas like Manhattan.

Hypothesis 2: Revenue losses for traditional taxis are most pronounced during peak hours, when Uber's dynamic pricing is competitive.

Hypothesis 3: Boroughs with fewer ride-hailing users exhibit minimal impact on traditional taxi demand. Visualization and Policy Recommendations

Use interactive dashboards and geographic visualizations to highlight Uber's influence on traditional taxis across boroughs and periods. Provide policy suggestions, such as fare adjustments, fleet size regulations, or hybrid taxi-Uber partnerships, to address disparities and improve transportation equity. Expected Outcome By comparing Uber and traditional taxi data, this study will identify critical shifts in ridership patterns and revenue impacts across boroughs. It will also offer actionable insights for balancing competition and accessibility in NYC's evolving transportation landscape. These findings aim to inform policy-making, guide service optimization, and promote sustainable urban mobility solutions.

3 Dataset

We will primarily use the NYC Taxi Limousine Commission (TLC) trip records from 2009 through 2024, which include ride information for Yellow and Green Cabs. Each record contains pickup/drop-off timestamps, geographic coordinates, fare amounts, and other trip-level variables (tip amounts, and payment types). These data will allow us to identify temporal patterns and spatial distributions (high-demand neighborhoods).

I also plan to use the fhvhv dataset (obtained from Kaggle), which is some online ride-hailing service records, which can be useful data when we need to compare with NYC taxis.

4 Group Member Contributions

In this project, everyone in the team worked hard. Zhongren completed the part of MOTIVATION AND PROBLEM STATEMENT, and looked for additional data sets to help the data analysis in the future project, and complete most of the research question's methodology. Yueyi proposed some novel research questions and described the motivation for the research in detail. After group discussion, she decided to adopt them as our research question 2. At the same time, she proposed some feasible research methods on this issue.

References