**Final Project Report : NYC Taxi Data Analysis**

Submitted by:

* Anuskha Kedar
* Rachit Pandya
* Rishabh Sharma
* Tiantian Cai

**MS in Business Analytics, Northeastern University, Boston**

Course Number: MSIM 6205

Course Name: Data Wrangling

Submitted to: Prof. Carol Lee

Date : 12/04/23

**Executive Summary** : This comprehensive report delves into the analysis of the New York City taxi market, aiming to provide strategic insights for the launch of 'Go,' a new taxi company. The project focuses on understanding transportation dynamics, challenges during peak hours, and positioning recommendations for 'Go.' The analysis incorporates data from diverse sources, including the NYC Taxi and Limousine Commission (TLC) and the NYC Open Data portal, covering variables such as pickup/drop-off locations, fare amounts, and trip distances.

1 . Topic and Questions:

Business Topic: Analyzing the New York City taxi market to facilitate the launch of 'Go.'

Business Questions:

1. What are the transportation dynamics and challenges in NYC, especially during peak hours?

The report delves into the intricate dynamics of transportation in New York City (NYC), particularly focusing on the challenges faced during peak hours. It identifies the highly competitive landscape involving various taxi services such as yellow taxis, green taxis, and for-hire vehicles. The analysis emphasizes the impact of increased demand during peak hours, resulting in longer wait times and elevated prices. External factors like weather conditions, events, and traffic congestion are highlighted as influential contributors to the overall transportation dynamics.

1. How can 'Go' strategically position itself to succeed in this competitive market?

The report provides strategic recommendations for 'Go' to carve a successful niche in the competitive NYC taxi market. It suggests that 'Go' can differentiate itself by prioritizing customer experience, safety, and sustainability. Recommendations include offering personalized services, such as in-car entertainment and refreshments, to enhance the overall customer experience. Safety is underlined through proposed driver training programs and the incorporation of advanced safety technologies. Additionally, the report recommends adopting sustainable practices like employing electric or hybrid vehicles to appeal to environmentally conscious consumers.

Importance: Understanding the NYC taxi market is crucial for guiding operational and strategic decisions for 'Go.'

Target Audience: 'Go' executives, investors, and operational managers.

2. Data Sources and Datasets :

* Data Sources: The analysis utilized data from three main sources: New York City Taxi and Limousine Commission (TLC), NYC Open Data, and Kaggle.
* TLC: Served as the primary source for detailed trip records of yellow taxis, green taxis, and for-hire vehicles, providing essential information crucial for the analysis.
* NYC Open Data and Kaggle: These public sources complemented the analysis by providing additional datasets to enhance the depth and breadth of insights.
* Data Collection Process: The data collection process involved advanced data wrangling and transformation techniques to enhance data relevance. Techniques included uniform date formatting, the introduction of a 'Taxi\_Type' attribute, and strategic DataFrame merging. These measures aimed to ensure consistency, categorize data, and integrate information from diverse sources, enriching the analytical dataset.
* Variables: The datasets included variables such as trip details (pickup/dropoff locations, duration, fare amount, payment type), taxi type, weather conditions, and traffic congestion.

*References:*​

* [New York City Taxi and Limousine Commission (TLC)](https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page)​
* [Yellow Trips Data Dictionary](https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf)​
* [Green Trips Data Dictionary](https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf)​
* [High Volume FHV Trips Data Dictionary](https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_hvfhs.pdf)​
* [Taxi Zone Lookup Table (CSV)](https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page)​
* [Taxi Zone Shapefile](https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page" \t "_blank)​

3. Information Quality :

* Data Accuracy: The accuracy of the analysis depends on the precision of the datasets used. The data cleaning and pre-processing, indicating an awareness of the importance of accurate information.
* Consistency: Working with multiple datasets from different sources raises concerns about consistency. The transformation techniques, such as uniform date formatting and DataFrame merging, likely aimed at ensuring a consistent and cohesive analytical dataset.
* Missing Values: Handling of missing values is a critical aspect of data quality. The data wrangling techniques, and approach helped us to address missing or inconsistent data points.
* Data Relevance: The use of supplementary datasets from NYC Open Data and Kaggle. Ensured the relevance of these datasets to the core analysis for deriving meaningful insights.
* Addressing these concerns likely involved a combination of data cleaning, validation rules, and strategic transformations to enhance the overall quality and reliability of the information used in the analysis.

4. Methods and Tools :

Methods:

* The data wrangling process involved a combination of established techniques and custom methods tailored to the unique challenges posed by the New York City taxi market analysis.

Tools and Technologies:

* Python: The primary programming language used for data manipulation, analysis, and visualization.
* Pandas and NumPy: Key Python libraries employed for data manipulation and numerical operations.
* Matplotlib and Seaborn: Utilized for creating visualizations that aided in the analysis and presentation of findings.
* Jupyter Notebooks: Provided an interactive and iterative environment for code development and analysis.
* Microsoft PowerBI: Utilized in the creation of dashboards for 2 datasets. Helped analyse the data more comprehensively.
* Looker Studio: Utilized in creating the dashboards for 2 datasets.

Theories and Concepts:

* Data Imputation: Employed to handle missing values and ensure the completeness of the datasets.
* Data Transformation: Techniques such as uniform date formatting and DataFrame merging were applied to enhance data consistency.
* Feature Engineering: Introduced the 'Taxi\_Type' attribute to categorize different types of taxis, enriching the dataset for more nuanced analysis.

Adaptations and Challenges:

* Data Cleaning Challenges: Dealing with outliers, inconsistencies, and missing values required a nuanced approach. Custom data validation rules were developed to identify and address irregularities.
* Integration of Diverse Datasets: Merging data from the TLC, NYC Open Data, and Kaggle posed challenges due to varying formats and structures. Custom scripts were devised to harmonize these datasets seamlessly.

* Updates to Existing Scripts or Techniques: While we utilized standard libraries and techniques, several custom scripts were developed for data validation, and tailored transformations to address the specificities of the NYC taxi market data.

Overall, the methods and tools employed were instrumental in navigating the complexities of the dataset, ensuring data quality, and deriving meaningful insights. The adaptability of our approach allowed us to overcome challenges and present a robust analysis of the NYC taxi market.

1. Data Wrangling Process: The data wrangling process involved several steps to clean, transform, and prepare the raw taxi dataset for further analysis. Below is a detailed description of the executed data wrangling steps:

Data Loading: Loaded raw data from various sources, such as Parquet files, csv and .shp files, into Pandas DataFrames. Explored the structure and content of the datasets to understand column names, data types, and potential issues.

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generated

Cleaning and Standardization: Identified and handled missing values, outliers, and inconsistent data formats. Standardized column names and ensured consistency across datasets for merging.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer program

Description automatically generated

* Created new features by mapping and transforming existing columns, such as deriving payment types, rate code names, and vendor names from corresponding codes. Categorical Mapping: Mapped numerical codes to their respective categorical values for better interpretability. Used mapping conditions for categorical columns like payment types, rate codes, and vendor IDs.
* Utilized conditions to create new columns based on the values of existing columns. For example, created new columns like RateCodeName, PaymentTypeName, and VendorName based on conditions.
* Geospatial Analysis: Leveraged geospatial information (latitude and longitude) for analyzing the distribution of taxi pickups and drop-offs. Visualized the data on maps to identify popular locations and trends.
* Data Validation Rules and Checks: Implemented validation rules to ensure data integrity and accuracy. Checked for consistency in column values, adherence to defined data types, and the absence of unreasonable outliers.

* Memory Management: Managed memory usage, especially for large datasets, to prevent memory errors during processing. Selected a subset of rows based on a predefined maximum memory size to optimize the dataset for further analysis.

Data Export: Exported the cleaned and transformed datasets to CSV files for storage and sharing. Prepared the data for visualization and analysis in tools like Looker.

A screenshot of a computer program

Description automatically generated

Outcome: The final outcome of the data wrangling process is a well-structured, cleaned, and enriched dataset ready for exploratory data analysis and visualization. The dataset contains relevant features, categorical mappings, and geospatial information, providing a solid foundation for addressing business-related questions and insights into the taxi service operations. The validation rules and checks ensure data quality and reliability for downstream analysis.

1. Potential Analysis and Results:

* In order to gain insights into the transportation dynamics and challenges in New York City, particularly during peak hours, a comprehensive analysis is undertaken. The exploration of trip characteristics involves a detailed examination of the distribution of trip distances, durations, and fares during peak hours. This in-depth analysis aims to discern the demand patterns for both short and long trips and understand the associated costs, shedding light on the diverse preferences of commuters. Furthermore, a temporal analysis is conducted to unravel the temporal dynamics of trip volumes during peak hours. This involves identifying the peak periods and discerning patterns of demand fluctuation over time. By scrutinizing the temporal aspects of ride demand, the analysis aims to provide a nuanced understanding of the temporal variability in transportation needs, enabling more effective service planning and resource allocation. In addition to temporal insights, a geospatial analysis is employed to visualize the spatial distribution of pickups and drop-offs during peak hours. This geospatial examination unveils the hotspots of ride activity, allowing for the identification of areas characterized by high demand and potential traffic congestion. By mapping these spatial dynamics, the analysis provides valuable information on the geographical intricacies of transportation demand, enabling strategic decision-making related to route optimization and service coverage. The geospatial visualization also aids in pinpointing areas that may require targeted interventions to alleviate congestion or enhance service accessibility. Together, these analytical approaches form a comprehensive strategy to decipher the intricate transportation dynamics in New York City during peak hours, facilitating data-driven decision-making for service providers and contributing to a more efficient and responsive urban transportation system.

* Conducting a robust competitor analysis is crucial for 'Go' to strategically position itself in the highly competitive ride-sharing industry in New York City. This analysis entails a comprehensive exploration of market share and performance metrics among major competitors. By delving into these key indicators, 'Go' can gain insights into the competitive landscape, identifying both strengths and weaknesses. Understanding market share dynamics provides a quantitative measure of 'Go's' standing in relation to its rivals, while an assessment of performance metrics unveils operational efficiency, reliability, and overall service quality. This knowledge forms the basis for strategic decision-making, allowing 'Go' to leverage its strengths and address vulnerabilities to enhance its market standing. Customer satisfaction stands as a pivotal factor in determining the success of ride-sharing services. 'Go' can harness the power of customer reviews and ratings to gauge the user experience and pinpoint areas of excellence and potential improvement. Analyzing sentiment and feedback from users provides valuable qualitative insights, offering a nuanced understanding of customer preferences, pain points, and expectations. Identifying aspects where 'Go' excels allows for strategic reinforcement of those strengths, fostering customer loyalty, while recognizing areas for improvement enables targeted enhancements to elevate overall satisfaction levels. Pricing strategy plays a pivotal role in shaping the market positioning of ride-sharing services. Analyzing the pricing structures of 'Go' and its competitors is essential to assess whether 'Go' is positioned as a cost-effective or premium service. Understanding the price elasticity of demand, comparing fare structures, and evaluating any added value or unique selling propositions associated with pricing allows 'Go' to strategically position itself in the market. This assessment helps 'Go' make informed decisions on pricing adjustments, promotional strategies, and the overall market positioning of its services. In summary, a well-rounded competitor analysis encompassing market share, performance evaluation, customer satisfaction, and pricing strategy empowers 'Go' to navigate the complexities of the ride-sharing industry in NYC. Armed with these insights, 'Go' can fine-tune its operations, enhance customer experiences, and strategically position itself in the market to not only withstand competition but thrive in this dynamic and competitive landscape.

A screenshot of a taxi application

Description automatically generated

GreenTaxi  
The dashboard crafted via Looker offers comprehensive insights into both the green and yellow taxi activities in New York. For the green taxi, the snapshot reveals a total revenue of $1.5 million generated from approximately 72.5 thousand trips, with an average trip distance of 40 miles and a fare amounting to $16.57. Interestingly, the average tip amount is approximately $2. The graph depicting average trip distance indicates that Queens stands out with the highest distance compared to other areas. Manhattan, recording the highest trip count, signifies a bustling location, ideal for business ventures. Delving into fare amount and tips, EWR emerges with the highest figures, specifically showcasing an average tip of $14.55. This denotes a potential lucrative area for taxi operations. The emphasis on congestion charges highlights Manhattan's congestion, with charges averaging around $1.13. The analysis underscores how certain areas like Queens and EWR exhibit significant potential in terms of longer trips and higher fares, while Manhattan, despite its congestion, represents a lucrative market due to its high trip count. This insight could guide strategic decisions for taxi businesses considering new operational areas or optimizing service offerings within these zones.

A screenshot of a taxi application

Description automatically generatedA screenshot of a computer

Description automatically generated

YellowTaxi  
The Looker dashboard provides a detailed overview of the yellow taxi operations in New York City. Yellow taxis account for approximately 48.4 million trips, with an average trip distance of 4.2 miles, a fare of $15.93, and an average tip of $2.96. State Island emerges as the area with the highest average trip distance, evident in the bar chart representation. However, Manhattan dominates the total trip count, covering nearly 50% of all yellow taxi trips. Similar to the green taxi, EWR exhibits the highest fare and tip amounts. Notably, when comparing congestion charges, Manhattan's charges for yellow taxis are almost double that of green taxis, signaling heightened congestion. This disparity underscores the significance of congestion charges in Manhattan for yellow taxis, indicating a higher operational cost. The insights gleaned could aid taxi businesses in strategic decision-making, particularly in optimizing service provision and considering the impact of congestion charges in different areas.

A screenshot of a graph

Description automatically generated

1. Additional Data and Analysis: if more time were available, collecting the following additional data could enhance the analysis:

* Weather Data: Incorporating weather data (e.g., temperature, precipitation) could reveal correlations between weather conditions and taxi usage patterns.
* Economic Indicators: Including economic indicators (e.g., unemployment rates, GDP) may help understand the impact of economic conditions on taxi ridership.
* Event Data: Gathering data on major events (e.g., sports events, concerts) could uncover spikes in taxi demand during such occasions.
* Traffic Data: Integrating real-time traffic data would enable analyzing how traffic conditions affect trip durations and taxi availability.
* Special Promotions and Discounts: Obtaining data on special promotions or discounts offered by taxi services could provide insights into their impact on ridership.

Potential New Analysis: With additional data, the following analyses could be conducted:

* Weather Impact on Ridership: Explore the correlation between weather conditions and taxi ridership to understand how factors like rain or extreme temperatures affect travel behavior.
* Economic Factors Analysis: Analyze the relationship between economic indicators and taxi usage to identify trends related to economic conditions.
* Event-Driven Demand Analysis: Examine how major events impact taxi demand, identifying patterns related to event types, timings, and locations.
* Traffic Congestion Analysis: Investigate how traffic congestion influences taxi ride durations and customer preferences.
* Promotion Effectiveness Assessment: Evaluate the effectiveness of promotional activities or discounts in boosting taxi ridership.

1. Contributions:
   1. Anuskha Kedar: Researcher responsible for summarizing external materials and providing insights, building 2 visualization dashboards with the help of Microsoft PowerBI.
   2. Rachit Pandya: Data analyst focusing on building the visualization dashboard in Looker.
   3. Rishabh Sharma: Lead data wrangler responsible for data pre-processing and extraction.
   4. Tintian C: Data specialist responsible for managing data and integration with tools.

1. References:

* [New York City Taxi and Limousine Commission (TLC)](https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page" \t "_blank)​
* [Yellow Trips Data Dictionary](https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf" \t "_blank)​
* [Green Trips Data Dictionary](https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf" \t "_blank)​
* [High Volume FHV Trips Data Dictionary](https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_hvfhs.pdf)​
* [Taxi Zone Lookup Table (CSV)](https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page" \t "_blank)​
* [Taxi Zone Shapefile](https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page" \t "_blank)​

Bibliography (External Material)

* CAMERER , C., BABCOCK , L., LOEWENSTEIN , G., & THALER, R. (1997). LABOR SUPPLY OF NEW YORK CITY CABDRIVERS: ONE DAY AT A TIME. The Quarterly Journal of Economics, 407-441.
* Hochmair, H. H. (2016). Spatiotemporal Pattern Analysis of Taxi Trips in New York City. Journal of the Transportation Research Board(2542), 45-56.
* Morrison, C. N., Mehranbod, C., Kwizera, M., Rundle, A. G., Keyes, K. M., & Humphreys, D. K. (2022). Ridesharing and Motor Vehicle Crashes: A Spatial Ecological Case-Crossover Study of Trip-Level Data. HHS Public Access, 118-123.
* Ulak, M., Yazici , A., & Aljarrah, M. (2020). Value of convenience for taxi trips in New York City. ELSEVIER, 85-100.
* Williams, Z., & College, M. (2018). Rideshare Companies Uber and Lyft Strongly Affect Taxi and Rental Car Industries. Journal of Management and Innovation, 1-21