NYC Urban Mobility and Transportation Analysis

Alassanne Sy* Michael Xie† Rida Sohail‡ Ze Hong Wu§

Hunter College - Advanced Visualization Tools

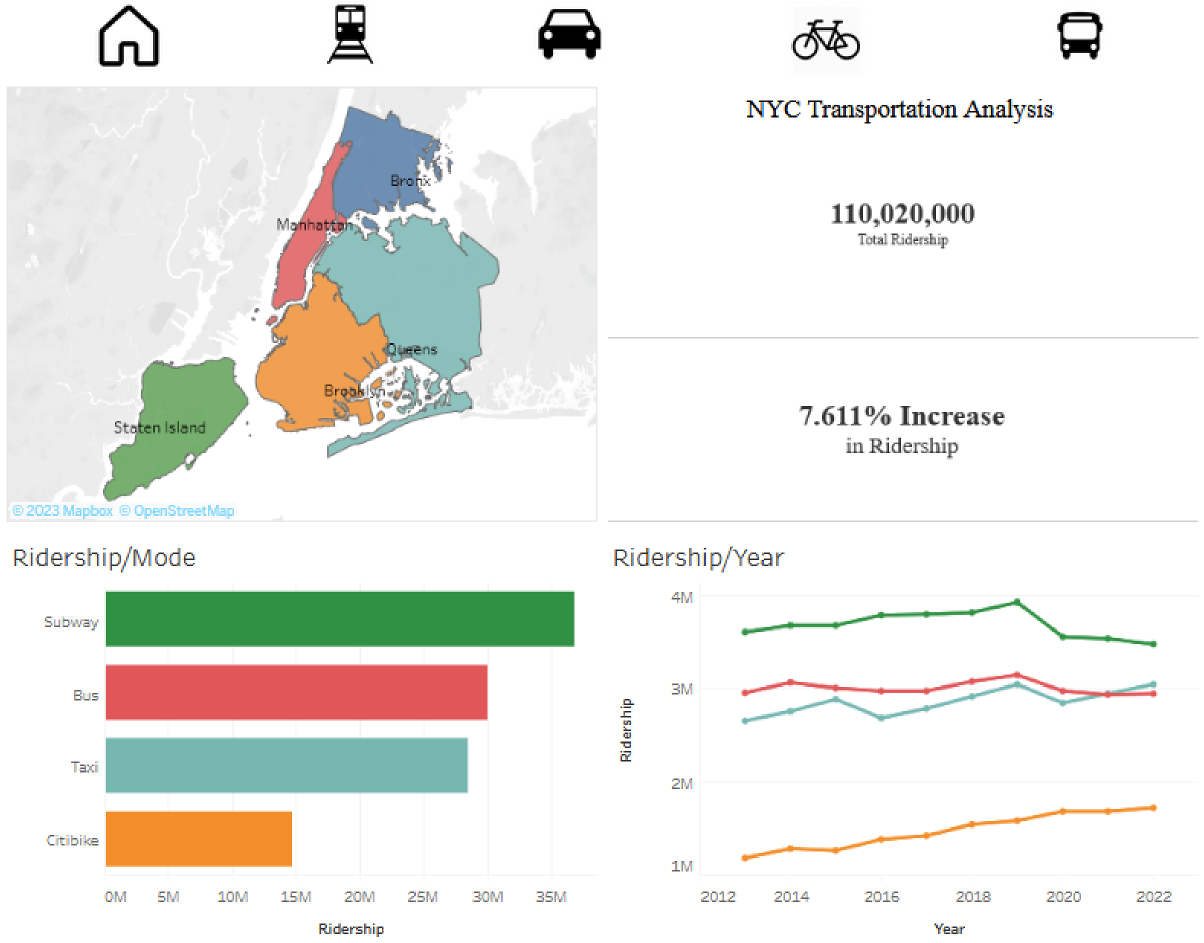

Figure 1: Screenshot of team's dashboard. It was inspired by the foundation living report present on the Tableau's public server.

Abstract

New York City is home to a variety of public transportation services, ranging from its extensive subway and bus network to taxis and even city provided bicycles. In this project we analyze the effectiveness and summarize our insights and conclusions in a series of visualizations that provide context on the usefulness of each public transportation method. The link in the teaser image leads to the final version of the visualizations, which provides an overall view of New York City transportation options and can offer insights for relevant readers.

1 Objectives

The project aims to conduct a comprehensive analysis of New York City's transportation system using Tableau to create data visualizations on diverse aspects of urban mobility, such as subway ridership, taxi data, bike-sharing statistics, and bus delays. Each team member will focus on a specific facet, ensuring in-depth exploration. The integration of these visualizations will provide a holistic view of transportation patterns, offering valuable insights for both residents and policymakers. The ultimate objectives include generating actionable recommendations for improvement, creating user-friendly visualizations, serving as an informative resource for policymakers, and facilitating an educational component for transparency. Through an iterative process, the project seeks to remain relevant and contribute to ongoing discussions surrounding the optimization of New York City's transportation infrastructure. Our project aligns with

*e-mail: Alassane.sy48@myhunter.cuny.edu
†e-mail: michael.xie46@myhunter.cuny.edu
‡e-mail: rida.sohail93@myhunter.cuny.edu
§e-mail: zehong.wu@macaulay.cuny.edu

the growing body of research in urban mobility analysis using big data, as demonstrated by studies such as Wang and Ran's survey on Urban Mobility Analysis Using Big Data [1].

All files used and produced as part of this project can be accessed in this GitHub repo.

2 Data Acquisition

The rationale for connecting the datasets lies in the holistic analysis of New York City's transportation system. While each team member focuses on a specific mode of transport, combining these datasets enables the exploration of correlations between different transportation modes. By understanding the pros and cons of each mode, the project aims to draw comprehensive conclusions about the overall efficiency, safety, and viability of various transportation options in the metropolitan context. It is essential to note that correlations discovered in the visualizations do not imply causation, and the project acknowledges the need for careful interpretation of the findings. The interconnected view derived from the integrated datasets contributes to a more informed understanding of urban mobility in New York City.

We used the following data sources:

Dataset Source: Yellow-Taxi-Trip-Data
Rationale: The dataset on Yellow Taxi trips for the year 2020 provides a comprehensive view of taxi operations in New York City. Analyzing this dataset can reveal patterns in daily operations, the number of clients served by taxi drivers, and their workload over a 24-hour period. Insights from this dataset can contribute to understanding the efficiency and demand for taxi services in different time frames.

Dataset Source: Bus-Breakdown-and-Delays
Rationale: The Bus Breakdown and Delays dataset is crucial for assessing the efficiency of the bus transportation system in NYC. Analyzing this dataset can help identify trends in bus delays over time and explore the reasons behind these delays. Insights derived from this dataset can inform recommendations for improving the reliability and effectiveness of the bus transportation network.

Dataset Source: MTA-Subway-Hourly-Ridership
The MTA Subway Hourly Ridership dataset offers a rich source of data to analyze subway usage trends. By exploring this dataset, we can create visualizations that showcase peak hours, popular routes, and other insights related to subway ridership patterns. This information is crucial for understanding the demand and usage patterns within the subway system.

Dataset Source: Citi-Bike-System-Data
The Citi Bike System Data provides a comprehensive dataset for analyzing bike-sharing patterns in NYC. We can explore aspects such as distances traveled, time spent during trips, and CitiBike station locations. Visualizations based on this dataset can offer insights into the popularity, efficiency, and availability of bike-sharing, contributing valuable information for residents and policymakers interested in sustainable transportation options.

3 Methodology

The team's methodology for analyzing New York City's transportation system involved a comprehensive approach, with each member focusing on a specific aspect such as yellow taxi data, bus delays, subway ridership, and bike-sharing patterns. The data cleaning process was meticulously executed across all projects, ensuring that each dataset was well-structured and ready for analysis. Various tools, including Jupyter notebooks and external Python scripts, were employed for data cleaning and feature engineering where necessary. The data cleaning process was meticulously executed using tools such as Jupyter notebooks and external Python scripts, drawing inspiration from established practices outlined in McKinney's ”Python for Data Analysis” [2].

In the data analysis phase, team members delved into the specifics of their respective datasets, identifying correlations and trends that would later inform the creation of insightful visualizations. Notably, the team faced challenges, such as handling time-based data and merging disparate datasets, which required thoughtful solutions to ensure the accuracy and relevance of the visualizations.

3.1 Responsibilities:

The Yellow Taxi Trip, Bus Breakdowns and Delays, MTA Subway Ridership, and Citi Bike System sub-dashboards were created by Alassanne Sy, Michael Xie, Rida Sohail, and Ze Hong Wu, respectively.

3.2 Data Cleaning:

We imported the 2020 NYC Yellow Taxi database from NYC Open Data into R for data cleaning and size reduction. The filter function was utilized to specifically choose the relevant data for further analysis.

We imported the Buses dataset and converted the .csv file into a DataFrame using a Jupyter notebook. The data cleaning process was executed in Python, and we meticulously addressed each data point by going through the dataset cell by cell. One key challenge involved managing time-based data, particularly the ”How Long Delayed” and ”Occurred On” fields. To ensure uniformity, ”How Long Delayed” entries were transformed into minute units, converting all values to integers and omitting textual descriptions. The ”Occurred On” data was formatted into a datetime structure and further dissected into day, month, and year columns, facilitating potential future analyses. Geographical data underwent reformatting to align with Tableau's expectations, such as transforming ”Standardized Boro” to ”City, State” format (e.g., ”Manhattan, New York”).

We imported the NYC MTA subway ridership data from the state's website for our visualization. The data cleaning process involved removing duplicate values and filling null values using pandas DataFrames in a Jupyter notebook. An additional step was taken to manually re-group subway lines that were erroneously grouped.

We imported the CitiBikes dataset from the NYC Citi Bikes website. This dataset required relatively little data cleaning, as it already contained well-formatted data with little to no problematic entries. The dataset underwent a feature engineering step to create additional columns, describing the day of the month for each bike ride, whether the ride occurred during rush hour and/or a weekend, and the duration of the bike rides in minutes. Feature engineering was performed through an external Python script.

3.3 Data Analysis:

During the analysis of the Yellow Taxi data, we systematically went through each column, focusing on vendor, time, location, payment, tip, tax, and distance. This process aimed to identify significant correlations and understand the dataset's key features, guiding the purpose of the visualizations.

During the analysis of the Buse Breakdown and Delays data, we systematically examined each column, identifying significant correlations between date, location, delay duration, and reasons for delays. This approach aimed to enhance the comprehensibility of the subsequent visualizations.

During the analysis of the MTA Subway Ridership data, we created two tables containing the data used for the visualization. One table contained geospatial data, and the other contained ridership information along with subway line groups. To plot the subway tracks on the map of New York, we joined the geospatial data source with the NYC MTA ridership data using the route ID parameter. For the bar chart visualization, ridership data was used, and parameters such as group and route-id were blended with the other table to ensure consistency between visualizations.

During the analysis of the Citi Bikes data, we focused on a combination of the spatial and temporal availability of Citi Bike services. This information can inform readers and travelers on whether they should consider using a Citi Bike for transit or seeking other options depending on the situation they are in.

3.4 Data Representation:

Our approach to representing the Yellow Taxi data draws insights from Tufte's seminal work, "The Visual Display of Quantitative Information" [3], emphasizing clarity and effectiveness in visual communication. We created three visualizations/dashboards to represent the analyzed yellow taxi data. The first visualization showcased a histogram displaying the average fare amount based on trip distance. This visualization also provided insights into each vendor's expenses, including tips, taxes, and more. The second visualization, in the form of a circle, presented detailed information about each vendor's trip, encompassing trip distance, fare amount, tip, extra charges, taxes, etc. The last visualization focused on the evolution of fare amounts based on pick-up and drop-off times.

For the Bus Breakdowns and Delays data we created four visualizations to represent the analyzed data. The first visualization displayed average bus delay times by year, featuring an interactive dropdown for annual comparisons. This allowed users to explore average delay durations across different times of the year. The subsequent trio of graphs delved into the reasons behind the delays, juxtaposing duration against causes, with the first graph incorporating location data. The third visualization was a geographic map color-coded to represent delay durations in the tri-state area, providing a spatial understanding of the delays. The final visualization, a heatmap, highlighted peak traffic times, pinpointing the 6-7 AM window as the most congested.

The visualization for the Subway Ridership data was inspired by a visualization present on Tableau's public server[1]. We created two visualizations, a map and a bar chart. The map displays NYC's subway tracks in their respective colors, grouped together based on the tracks they operate on. The bar chart showcases route-ids of subway lines, grouped based on their respective groups, with the aggregate count of passengers. We used two line graphs, one for the starting points and the other for the endpoints of the bars, combined and synchronized to create a horizontal bar chart effect.

We also created two pie charts to represent the distribution of people using a particular payment method for riding subways and the distribution of people taking the subway from a specific borough. These visualizations were incorporated by integrating TabPy with Tableau and utilizing Python scripts.

For the Citi Bikes data we produced two map plots displaying the location of Citi Bike and generic bike racks and a pair of bar and line charts displaying the approximate level of Citi Bike usage at various hours of the day and days of the month. The map plots displayed each known Citi Bike and generic bike rack as a point on a street map of New York City. The bar chart displays the total number of Citi Bike trips undertaken per hour of day, color coded by whether the trip occurred on a weekend or not, with noticeable spikes in the morning and evening rush hours. The line chart displays the total number of bike trips undertaken per day of the month, color coded by whether the trips occurred during rush hour or not, with an easily observable trend of rush hour bike trip spikes during weekdays.

3.5 Data Interaction:

We incorporated data interaction elements into the Yellow Taxi visualizations to enhance user engagement and understanding. The visualizations allowed users to explore average fare amounts, vendor expenses, trip details, and fare evolution based on specific criteria, providing a dynamic and interactive experience.

Throughout the Bus Breakdowns and Delays visualizations, we maintained a focus on data interactivity and clarity. Various filters and graphic types were employed to effectively differentiate and highlight the complexities of bus breakdowns in NYC.

For the Subway Ridership visualizations, hovering over a specific subway line on a barchart highlights the corresponding subway track on the NYC map. This provides users with a more interactive way to visualize the NYC subway system.

3.6 Summary:

In summary, the team's methodology encompassed thorough data cleaning, in-depth data analysis, thoughtful data representation through a variety of visualizations, and the integration of data interaction to foster a dynamic exploration of transportation patterns in New York City.

4 Conclusion

In conclusion, our team has undertaken a thorough exploration of various facets of New York City's transportation system, shedding light on the intricate dynamics of yellow taxi operations, bus delays, subway ridership, and bike-sharing patterns. The data-driven insights gleaned from our visualizations offer a glimpse into the complexities of urban mobility in one of the world's busiest cities.

By refining our visualizations and weaving them into a compelling narrative, we aspire to provide residents and policymakers with a nuanced understanding of the urban mobility landscape. Ultimately, our goal is to empower individuals to make informed decisions about their transportation choices while providing valuable insights that contribute to the ongoing discourse on optimizing the transportation infrastructure in New York City.

Acknowledgments

The authors wish to express their gratitude to Professor Wole Oyekoya for his invaluable guidance and support throughout the duration of the course. His expertise and encouragement significantly contributed to the success of this project.

References

[1] J. Wang and B. Ran. Urban mobility analysis using big data: A survey. *IEEE Access*, 6(4):17394–17409, 2018.

[2] Wes McKinney. *Python for data analysis*. " O'Reilly Media, Inc.", 2022.

[3] Edward R Tufte. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 2001.

[1] https://public.tableau.com/app/profile/skybjohnson