Urban Explorer: Visualizing NYC Taxi Data

Jinqin Xiong, Anthony Chen, Dong Tang, Songling Li

1 Introduction

The taxi system in New York City is an integral part of urban life, providing convenient transportation services for residents and visitors alike. This project aims to leverage a vast dataset of taxi trips provided by the New York City Taxi & Limousine Commission (TLC), in conjunction with OpenStreetMap and Java D3 technology, to create a dynamic data visualization website. Through this website, users can intuitively explore and analyze passenger travel habits, traffic flow, and the quality of taxi services. Dataset Overview: The dataset used in this project covers trip records of yellow taxis in New York City, including the following information:

- Date/time of passenger pick-up and drop-off: Providing specific times for trip start and end, allowing analysis of peak periods and passenger travel habits throughout the day.
- Pick-up and drop-off locations: Geographical location information, including latitude and longitude, aiding in identifying hotspots for passenger pick-up and drop-off.
- Trip distance: Straight-line distance from pick-up to drop-off location, useful for analyzing trip lengths and traffic flow within the city.
- Fare details: Including trip fare, tips, etc., assisting in evaluating the economic characteristics of taxi services.
- Rate type: Identifying whether trips are billed according to standard rates or other special rates.
- Payment type: Records of passenger payment methods, such as cash or credit card.
- Passenger count reported by the driver: Providing information on the number of passengers riding in the taxi.

The dataset is stored in Parquet format, with approximately 1.5 billion rows of data (50 GB) as of 2018, covering trip records from 2009 to 2018. These data, collected and provided by the NYC TLC, reflect detailed trip information for yellow taxis within New York City. By dynamically visualizing this vast and detailed dataset, this project aims to provide an intuitive, interactive platform to help users gain insights into the usage of taxis in New York City, including passenger travel patterns, distribution of travel times, and traffic hotspots by geographic location. Additionally, the project will explore the quality of taxi services, providing data support for urban planning, traffic management, and public policy.

2 One-sentence description

This project will develop a website where users can explore taxi trip data in New York City. The website will offer various interactive data visualization tools, including but not limited to timeline views, geographic heatmaps, and dynamic flow charts, to showcase taxi usage and traffic flow during different time periods.

3 Project Type

This project is for an interactive website containing multiple dynamic data visualization elements. The website will be divided into several sections, each focusing on showcasing different aspects of the data, such as hotspots for passenger pick-up and drop-off, peak periods throughout the day, and more.

4 Audience

The target audience for this project includes policymakers, transportation planners, researchers, and the general public. Providing intuitive data visualizations for these users will help them better understand traffic flow patterns in the city and the usage of taxi services, enabling them to make more informed decisions and strategies.

5 Approach

5.1 Details

Data Preparation: Process and clean the New York City taxi trip data, identifying key fields for visualization. Technology Stack Selection: Use Java as the backend language, D3.js to build dynamic data visualizations, and OpenStreetMap to display geographic information. Design and Development: Design the user interface and user experience, develop data visualization components, and website functionality. Testing and Deployment: Perform comprehensive testing to ensure website usability and data accuracy, then deploy the website to servers.

5.2 Evidence for Success

The key to the success of the project lies in providing a user-friendly interface that showcases taxi data in a dynamic and interactive manner, helping users easily understand complex datasets.

6 Best-case Impact Statement

If the project succeeds, it will become an important tool for understanding and analyzing traffic patterns and taxi service usage in New York City, positively impacting urban planning and the formulation of transportation policies.

7 Major Milestones

Data Preprocessing and Cleaning:

1. Process and clean the New York City taxi trip data.

2. Identify and extract key fields necessary for visualization.

Basic Framework and User Interface Design:

1. Develop the basic framework of the website.

2. Design the user interface to ensure usability and intuitive navigation.

Development and Integration of Data Visualization Components:

1. Develop the data visualization components using D3.js.

2. Integrate the visualization components into the website framework.

Testing and Deployment:

1. Conduct thorough testing to ensure the website's functionality and data accuracy.

2. Deploy the website to servers for public access.

8 Obstacles

8.1 Major obstacles

Large-scale and Complexity of Data: Processing and analyzing approximately 1.5 billion rows of data requires robust data processing capabilities and optimized data storage solutions. Integration of Geographic Data: Accurately integrating taxi data with OpenStreetMap geographic information requires precise geocoding and data matching techniques. Performance Optimization: To ensure smoothness and responsiveness of dynamic data visualization, performance optimization is needed for frontend display and data queries.

8.2 Minor obstacles

User Experience Design: Designing an intuitive and user-friendly interface, especially when dealing with complex data visualizations and interactions.

9 Resources Needed

Compute Resources: Powerful servers for data processing and storage, as well as high-performance web servers for hosting the dynamic data visualization website. Development Tools: Java development environment, D3.js library, Geographic Information System (GIS) tools, MapLibre GL JS, and software for data processing and analysis. Geographic Data Source: Obtain access permissions to the OpenStreetMap API for integrating geographic information and displaying maps. Dataset: Taxi trip dataset provided by the New York City Taxi & Limousine Commission.

10 Related Publications

1. ”Spatiotemporal Pattern Analysis of Taxi Trips in New York City” investigates the spatial and temporal patterns of taxi trips in NYC using a dataset of 29 million trip records. The study examines factors influencing trip generation and attraction, including the role of airports and variations in travel speed throughout the day. Through negative binomial regression modeling and spatial filtering, the paper predicts taxi trip numbers based on infrastructure, socioeconomic, and land use variables. It identifies districts with increased taxi use and inadequate public transit service, offering insights for decision-making on investments in transit infrastructure to enhance mode share.

2. ”Spatial-Temporal Analysis of Urban Mobility using Taxi Dataset,” authored by Pratyush Kumar and Varun Singh, was published in November 2023. This study delves into the spatial-temporal analysis of urban mobility using taxi dataset. It utilizes aggregated Uber trip data from 2016 to 2019 for New Delhi, India. The authors employ Python-based techniques, including big data analytics, machine learning, and probabilistic programming, to predict travel time by utilizing the Uber Movement Dataset of New Delhi across various origins and destinations. Time series forecasting is conducted using ARIMA, Holt-Winters, Facebook Prophet, and a global model, highlighting the variance between actual and predicted travel time. Spatial analysis is also performed for different wards to discern fluctuations in trip volume. The findings of this study are valuable for urban planning and gaining a deeper understanding of human mobility patterns in New Delhi.

3. ”Taxi Techblog 2: Leaflet, D3, and other Frontend Fun” By Chris Whong - This article offers an excellent starting point for transforming GeoJSON into SVG layers closely integrated with Leaflet, by providing examples with D3 + Leaflet. Leaflet is an open-source JavaScript library for easily adding interactive maps to web pages. Mapbox is a source of hosted tiles that allows you to easily style your maps.

4. “Creating an NYC taxi data visualization with KeyLines” - This publication discusses the visualization of NYC taxi data using the KeyLines toolkit, emphasizing the visualization of connected data on maps and the analysis of patterns like pick-up and drop-off locations, total fare, and trip distances. The approach showcases how graph visualization techniques can offer insightful analysis of complex datasets (Cambridge Intelligence).

5. “NYC Taxi Visualization by Daniel Beckwith & Aditya Nivarthi” - This project provides a detailed exploration of taxi trip times, costs, and the importance of taxis across different NYC zones. It presents data in an interactive format, allowing users to filter data by month and taxi zone, offering insights into the operation and usage of taxis within the city (Sizmw).

6. “HEAVY.AI: NYC Taxi Ride Data Visualization” - Offers an example of using HEAVY.AI for visualizing every taxi ride in NYC over a 7-year period. It integrates taxi trip data with the building footprint of every store within proximity of pickup and dropoff locations, enhancing the understanding of taxi ride patterns in relation to commercial points of interest (HEAVY.AI).

7. “A Day in the Life: NYC Taxis” - This visualization portrays the daily activities of a single NYC yellow taxi in 2013, including where it operated, the revenue it generated, and its activity over 24 hours. It utilizes data provided by the Taxi and Limousine Commission and presents an engaging narrative on the life of a taxi through visual data storytelling (Chris Whong).

8. NYC taxi data analysis by transdim - Offers a comprehensive analysis of taxi flow in Manhattan, aggregating hourly taxi flow data between pickup and dropoff locations. The publication uses data science techniques to group and visualize the complex dataset, highlighting the dynamics of taxi movement within specific urban areas (transdim).

11 Define Success

The success criteria for this project entail the creation of a dynamic, interactive data visualization website that effectively showcases New York City taxi data, enabling users to gain a deep understanding of urban traffic patterns and taxi service usage. Additionally, a successful project should feature excellent user experience design, ensuring that all functionalities are easily accessible and user-friendly. Lastly, the project’s success also hinges on its impact among the target audience, including policymakers, transportation planners, and the general public, as reflected in positive feedback and engagement.

12 References

1. Hochmair, Hartwig. (2016). Spatiotemporal Pattern Analysis of Taxi Trips in New York City. Transportation Research Record: Journal of the Transportation Research Board, 2542, 45-56. DOI: 10.3141/2542-06.

2. Kumar, Pratyush & Singh, Varun. (2023). Spatial-Temporal Analysis of Urban Mobility using Taxi Dataset. DOI: 10.21203/rs.3.rs-3630229/v1.

3. https://chriswhong.com/open-data/taxi-techblog-2-leaflet-d3-and-other-frontend-fun/

4. https://cambridge-intelligence.com/visualizing-nyc-taxi-cab-data/

5. https://sizmw.github.io/nyc-taxi-vis/

6. https://www.heavy.ai/demos/taxis

7. https://chriswhong.github.io/nyctaxi/

8. https://transdim.github.io/dataset/NYC-taxi/

9. https://medium.com/@muhammadaris10/nyc-taxi-trip-data-analysis-45ecfdcb6f91

10. https://c2smart.engineering.nyu.edu/wp-content/uploads/69A3551747124-Impact-of-Ride-sharing-in-New-York-City.pdf

11. https://en.wikipedia.org/wiki/OpenStreetMap

12. https://wiki.openstreetmap.org/wiki/Using$_{O}penStreetMap$