Geospatial and Temporal Analysis of NYC Taxi Trips Using PySpark

Submitted by

Hasin Md. Daiyan, 20200204055
Shahariar Hossain Remon, 20200204097
Ekram Chowdhury, 20200204103

Submitted to

Ms. Nusrat Jahan
Ms. Nawrin Tabassum

Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh

January 15, 2025

Introduction

Urban transportation systems generate massive amounts of data daily, providing opportunities to analyze and optimize mobility patterns. In this project, we aim to investigate the geospatial and temporal dynamics of New York City taxi operations using the NYC Taxi Trip dataset. Leveraging PySpark, a distributed data processing framework, we plan to handle the scale and complexity of the dataset effectively. Our methodology involves preprocessing the dataset to clean and extract relevant features, followed by geospatial and temporal analyses. Geospatial analysis will focus on mapping trip patterns and identifying hotspots, while temporal analysis will examine trends in demand over time. Additionally, we aim to integrate these dimensions to uncover spatiotemporal patterns and provide actionable insights. Through this study, we seek to demonstrate how large-scale data analytics can help optimize urban transportation systems, improve resource allocation, and inform decision-making for city planners. The findings of this project can also guide further applications in transportation research and policy development.

Related Works

The rapid growth of urbanization and advancements in transportation systems have led to an increased emphasis on analyzing large-scale transportation datasets. Among these, the New York City (NYC) taxi trip dataset is widely used for studying urban mobility patterns, transportation planning, and geospatial-temporal trends. PySpark, a distributed computing framework, is extensively applied in handling the scale and complexity of such datasets due to its efficiency and scalability.

0.1 Geospatial Analysis in Urban Transportation

Geospatial analysis helps study spatial distributions, identify hotspots, and optimize routes. Zhao et al. in Geospatial Analysis of Urban Mobility Patterns Using Big Data [1] highlighted its importance in understanding urban mobility. Techniques like DBSCAN and k-means clustering are used to detect dense activity zones (Jiang et al., Clustering High-Demand Taxi Zones: A Case Study on New York City [2]). Apache Sedona extends PySpark for spatial operations, as shown by Shekhar et al. in Spatial Big Data Challenges and Opportunities [3].

0.2 Temporal Analysis of Transportation Data

Temporal analysis examines variations in transportation demand over time, including daily and seasonal trends. Gößling et al. in Temporal Trends in Urban Taxi Use [4] identified peak and off-peak periods crucial for resource allocation. PySpark simplifies time-series analysis, as demonstrated by Wang et al. in Anomaly Detection in Urban Mobility Patterns Using PySpark [5], where events like weather disruptions were analyzed.

0.3 Integrating Geospatial and Temporal Analysis

Integrating geospatial and temporal dimensions enables spatiotemporal hotspot identification and demand forecasting. Zhang et al. in Spatiotemporal Analysis of Urban

Transportation Systems [6] used Random Forest and Gradient Boosting for weather-related demand prediction. PySpark, with libraries like MLlib, provides robust tools for spatiotemporal data processing, as discussed by Chen et al. in Machine Learning Approaches for Spatiotemporal Predictions in Big Data [7].

0.4 Graph-Based and Visual Analytics Approaches

Huang et al. in TrajGraph: A Graph-Based Visual Analytics Approach [8] introduced a framework using graph theory to identify transportation hubs.
Zhang et al. in VisDPT: Visual Exploration of Differentially Private Trajectories [9] employed differential privacy mechanisms for secure trajectory analysis.

Summary of Related Works

Paper Title	**Approach**	**Model**	**Result**
Zhao et al. (2020) - "Geospatial Analysis of Urban Mobility Patterns Using Big Data"	Geospatial clustering using DBSCAN and spatial density analysis	Clustering algorithms (e.g., DBSCAN)	High-demand areas for taxis were identified near commercial hubs, transportation hubs, and tourist hotspots.
Jiang et al. (2018) - "Clustering High-Demand Taxi Zones: A Case Study on New York City"	Comparative analysis of clustering techniques for high-demand zones	DBSCAN and k-means clustering	Identified high-demand clusters including airports and nightlife areas. Recommended dynamic pricing and fleet rebalancing strategies.
Shekhar et al. (2019) - "Spatial Big Data Challenges and Opportunities"	Review of spatial big data challenges and solutions using Apache Sedona	Spatial indexing and partitioning with Sedona integration	Improved performance of spatial queries and detected demand-supply mismatches in ride-hailing services.
Gößling et al. (2019) - "Temporal Trends in Urban Taxi Use"	Temporal trend analysis focusing on peak and off-peak variations	Time-series analysis	Highlighted peak-hour demand during work commutes and seasonal variations. Suggested dynamic fare systems.

Wang et al. (2021) - "Anomaly Detection in Urban Mobility Patterns Using PySpark"	Anomaly detection for urban mobility using PySpark and temporal aggregation	Time-series anomaly detection	Detected anomalies during events like snowstorms and public gatherings. Proposed resource allocation strategies.
Zhang et al. (2022) - "Spatiotemporal Analysis of Urban Transportation Systems: Insights from NYC Taxi Data"	Spatiotemporal correlation analysis using integrated weather and trip data	Random Forest and Gradient Boosting for predictive modeling	Revealed weather's impact on trip durations and identified spatiotemporal hotspots. Developed predictive models for trip durations and demand forecasting.
Chen et al. (2023) - "Machine Learning Approaches for Spatiotemporal Predictions in Big Data"	Application of machine learning models for demand forecasting	Random Forest, Gradient Boosting, and PySpark's MLlib integration	Achieved high accuracy in demand prediction. Improved anomaly detection with integrated weather and temporal patterns.
Huang et al. (2020) - "TrajGraph: A Graph-Based Visual Analytics Approach to Studying Urban Network Centralities Using Taxi Trajectory Data"	Graph-based modeling of trajectories and centrality analysis	Graph theory metrics (e.g., degree and betweenness centrality)	Identified critical transportation hubs, optimizing traffic flow and urban planning.
Zhang et al. (2018) - "A Demonstration of VisDPT: Visual Exploration of Differentially Private Trajectories"	Differential privacy mechanisms for trajectory analysis	Privacy-preserving visual analytics	Balanced privacy and functionality while maintaining data utility for trajectory analysis.

Dataset

For this analysis, we plan to use the publicly available New York City Taxi Trip dataset, which is accessible through the NYC Open Data platform. The dataset includes detailed information about taxi trips in NYC, such as pickup and dropoff locations, timestamps, trip distances, and fare amounts. We intend to focus on data from January 2013, approximately 2.5 GB in size after decompression, as a representative sample for our study. Each row in the dataset corresponds to a single taxi ride and contains the following attributes:

- **Trip ID**: A unique identifier for each ride.

- **Pickup and Dropoff Times**: Timestamps indicating the start and end times of the trip.
- **Pickup and Dropoff Locations**: Latitude and longitude coordinates of the trip's origin and destination.
- **Fare Information**: Details of the fare, including the total amount, tips, and surcharges.
- **Driver and Vehicle Information**: Anonymized identifiers for drivers and medals.

To prepare the dataset, we plan to perform cleaning and preprocessing tasks, such as handling missing values, filtering out invalid records (e.g., zero or out-of-bound coordinates), and converting timestamps to a suitable format for temporal analysis.

Proposed Methodology

This study aims to analyze the geospatial and temporal patterns within the NYC Taxi Trip dataset. The methodology is structured into several key stages, as outlined below.

Model Diagram

The following steps describe the workflow for our proposed methodology:

1. **Data Collection and Preprocessing:**
   - We will download the dataset from the NYC Open Data platform.
   - Next, we will extract relevant features, such as timestamps and geospatial coordinates.
   - Finally, we will handle missing or invalid records and filter trips with zero or out-of-bound coordinates.
2. **Geospatial Analysis:**
   - We will use the Esri Geometry API to perform spatial operations, such as mapping pickup and dropoff locations to NYC boroughs.
   - GeoJSON files containing NYC borough boundaries will be integrated into the analysis pipeline.
   - Trips will be classified based on their geospatial characteristics (e.g., within a borough, inter-borough, or outside NYC).
3. **Temporal Analysis:**
   - We will conduct time-series analysis to identify peak and off-peak hours, seasonal trends, and anomalies in trip patterns.
   - PySpark SQL and UDFs will be used to compute durations and intervals between trips.

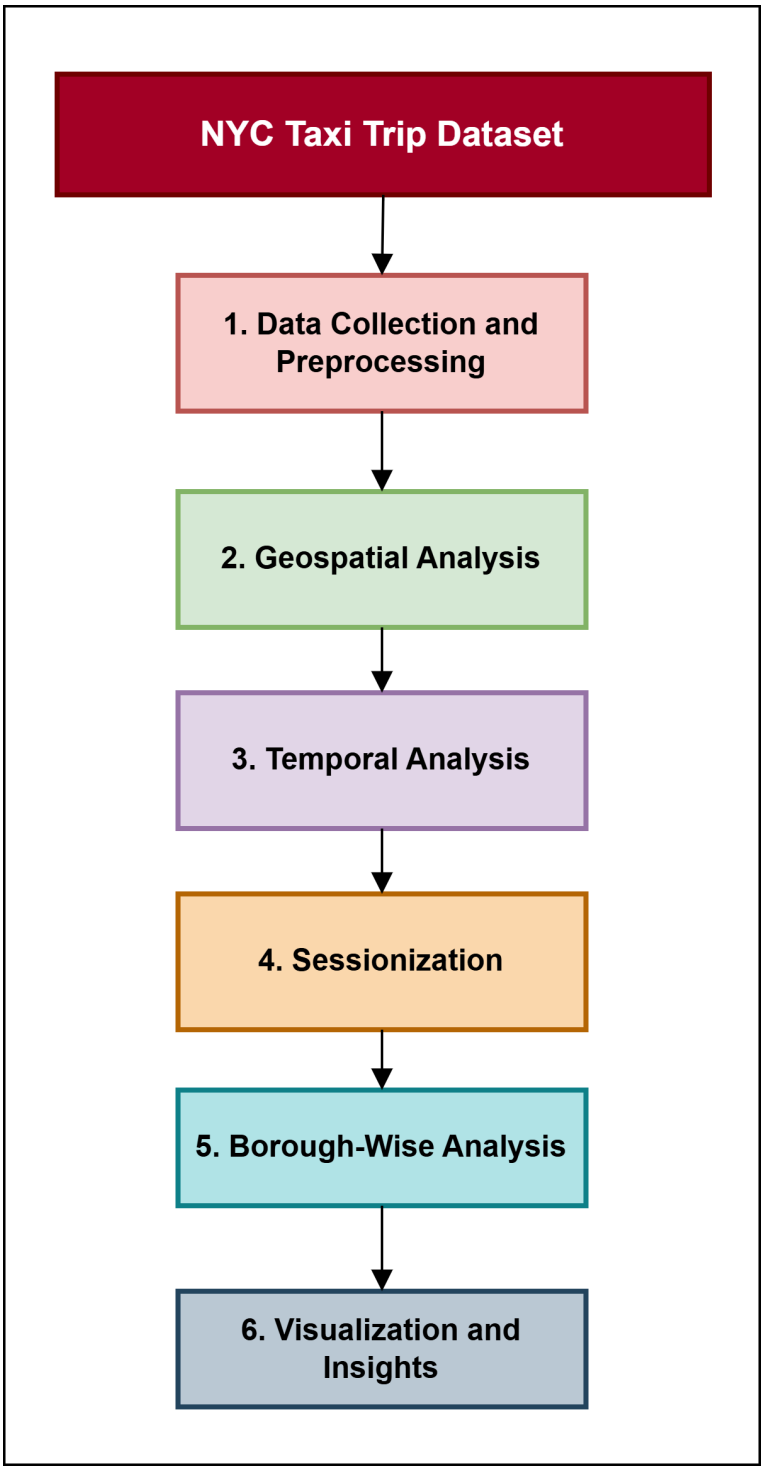

Figure 1: Proposed Methodology.

4. **Sessionization:**
    - Trips will be grouped by driver and sorted chronologically to analyze session patterns.
    - Inter-trip durations will be computed to study the time taken for drivers to find the next fare.
5. **Borough-Wise Analysis:**
    - We will analyze trip distributions and durations across different boroughs.
    - Factors affecting taxi availability and downtime, such as traffic and destination popularity, will be investigated.
6. **Visualization and Insights:**
    - Geospatial visualizations, including heatmaps and trip density maps, will be generated.
    - Temporal visualizations, such as hourly and daily trip trends, will be created to derive actionable insights.

In the Figure 1, we have outlined the key steps involved in our analysis, including data collection, preprocessing, geospatial and temporal analysis, sessionization, and visualization to derive actionable insights.

Through this methodology, we aim to uncover meaningful insights into NYC's taxi operations and propose actionable recommendations for improving urban transportation.

Conclusion

Our project aims to analyze the NYC Taxi Trip dataset to uncover geospatial and temporal patterns in urban mobility. By leveraging PySpark for scalable data processing, we plan to identify high-demand zones, peak usage times, and driver activity trends. The findings will offer actionable insights for optimizing resource allocation, improving traffic management, and supporting urban planning. Through this study, we highlight the importance of big data analytics in addressing real-world transportation challenges and enhancing city infrastructure.

References

1. Zhao, X., Wang, S., & Ye, Z. (2020). "Geospatial Analysis of Urban Mobility Patterns Using Big Data." Journal of Urban Studies.
2. Jiang, Y., He, Y., & Li, Z. (2018). "Clustering High-Demand Taxi Zones: A Case Study on New York City." Transportation Research Record.
3. Shekhar, S., Xiong, H., & Zhou, C. (2019). "Spatial Big Data Challenges and Opportunities." Springer.

4. Gößling, S., Cohen, S., & Hares, A. (2019). "Temporal Trends in Urban Taxi Use." Urban Transport Journal.

5. Wang, J., Lin, H., & Zhang, T. (2021). "Anomaly Detection in Urban Mobility Patterns Using PySpark." IEEE Transactions on Big Data.

6. Zhang, P., Chen, Y., & Wu, M. (2022). "Spatiotemporal Analysis of Urban Transportation Systems: Insights from NYC Taxi Data." Elsevier.

7. Chen, L., Yang, D., & Sun, W. (2023). "Machine Learning Approaches for Spatiotemporal Predictions in Big Data." Journal of Applied Artificial Intelligence.

8. Huang, X., Zhao, Y., Yang, J., Zhang, C., Ma, C., & Ye, X. (2020). "TrajGraph: A Graph-Based Visual Analytics Approach to Studying Urban Network Centralities Using Taxi Trajectory Data."

9. Zhang, J., Wang, F., & Li, M. (2018). "A Demonstration of VisDPT: Visual Exploration of Differentially Private Trajectories."