A study on Geospatial and Temporal Data Analysis on New York City Taxi Trip Data

1st Fatin Ishraq
Dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology

2nd Parvez Ahammed
Dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology

3rd Tabassum Tara Lamia
Dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology

4th Tahmid Sattar
Dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology

***Abstract*—With the increasing availability of urban data, new opportunities for data-driven research are emerging, facilitating evidence-based decision-making and strategy development. This study focuses on a substantial dataset of taxi rides, offering insights into various aspects of urban life, such as economic activity, human behavior, and mobility trends. The analysis involves pre-processing, feature engineering, and exploratory analysis of taxi data, which includes geographic information and several ride-specific variables, to understand transportation dynamics in New York City. Using Apache Spark, the study examines taxi drivers' wait times for their next passenger based on location. The dataset, sourced from the NYC Taxi and Limousine Commission, utilizes geospatial and temporal data to extract insights. The findings indicate that drivers can find passengers more quickly in Manhattan compared to the Bronx, providing valuable information for enhancing taxi service efficiency and understanding mobility patterns in New York City.**

***Index Terms*—Taxi Services, Geospatial Analysis, Spatial-Temporal Patterns, Machine Learning Algorithms**

I. Introduction

New York City is widely known for its yellow taxis, and hailing one is just as much a part of the experience of visiting the city as eating a hot dog from a street vendor or riding the elevator to the top of the Empire State Building. Residents of New York City have all kinds of tips based on their anecdotal experiences about the best times and places to catch a cab, especially during rush hour and when it's raining. But there is one time of day when everyone will recommend that you simply take the subway instead: during the shift change that happens between 4 and 5PM every day. During this time, yellow taxis have to return to their dispatch centers (often in Queens) so that one driver can quit for the day and the next one can start,and drivers who are late to return have to pay fines.To understand taxi economics, one key statistic is utilization—the percentage of time a cab is on the road and carrying passengers. A major factor affecting utilization is the passenger's destination. For example, a taxi that drops off passengers near Union Square during midday is likely to find its next fare within minutes. In contrast, a taxi dropping someone off at 2AM on Staten Island may need to drive back to Manhattan before picking up another passenger. We aim to quantify these effects and determine the average time it takes for a cab to find its next fare based on the borough where it dropped off its passengers—Manhattan, Brooklyn, Queens, the Bronx, Staten Island, or elsewhere (such as Newark International Airport).

For this analysis, we're only going to consider the fare data from January 2013, which will be about 2.5 GB of data after we uncompress it.One strategy that experienced data scientists deploy when working with a new data set is to add a try-catch block to their parsing code so that any invalid records can be written out to the logs without causing the entire job to fail.

II. Related Works

Venkatesh Subramaniam et al. utilized geospatial and temporal data which is made to good use in spark and the insights are derived.The New. York City taxi business is one of the interesting fields for data analysis. The analysis is done on spark and it concentrate on the duration of wait time for the drivers after a successful ride based on location. The data used is from NYC Taxi and Limousine department. There are two types of taxi cabs that run in New York City (Splechtna et al., 2016). One is New York City Yellow Taxi cab that we are analyzing in this project. The other type is New York City Green Taxi cab which is also popular but not as popular as yellow taxi cab.That is the reason we analyze only New York City yellow taxi cab in this analysis.The workflow of the Geospatial and Temporal Data Analysis on the New York City Taxi Trip Data also follows similar method to that of research papers but the efficiency of filtering out invalid data and having a seamless flow process makes the difference[1].

Nivan Ferreira et al. discusses the analysis of taxi trip data in New York City, highlighting its value as a sensor of urban life. It explores patterns, anomalies, and useful information derived from the data. So this introduces us to a new visual query model to simplify complex spatio-temporal queries, addressing limitations in existing tools. The proposed model supports various query types and enables interactive exploration of large datasets. The system, TaxiVis, is designed to facilitate visual exploration of taxi trip data, catering to the needs of domain experts in economics and

traffic engineering. It emphasizes the importance of unified data selection and visual analysis to enhance exploratory and confirmatory analyses. The model allows users to compose queries visually, refine them iteratively and explore query results effectively.[2]

Omid Ghaffarpasand et al. analyzes the limitations of current methods in understanding vehicle emissions and energy consumption patterns. It introduces a new approach using telematics data to analyze urban mobility and road dynamics. By mapping vehicle speed variations, the study aims to estimate fuel consumption and air pollutant emissions. The methodology involves segment-based analysis and calculating vehicle-specific power. The study focuses on the West Midlands region in the UK, utilizing data from GPS-connected vehicles. Key aspects include the spatial and temporal scope of the study, segment-based estimation of vehicle-specific power, and geospatial estimation of segment slopes. The goal is to provide a detailed understanding of road dynamics for better transport planning and decision-making.[3]

Sašo Džeroski et al. highlighted various approaches to monitor forest attributes using remote sensing data for efficiency and accuracy. Combining Landsat and LIDAR data improves forest attribute estimation. Hudak et al. (2002) demonstrated LIDAR's accuracy in characterizing forest structure when integrated with Landsat imagery, and Lefsky et al. (2002) discussed LIDAR's effectiveness in detailing vertical forest structure. Wulder et al. (2012) emphasized the cost-effectiveness and scalability of satellite data for large-scale forest monitoring, addressing traditional field method limitations. Saarela et al. (2018) showed that predictive models based on remote sensing data reliably map forest attributes, supporting sustainable management. These findings underscore the importance of integrating multi-source remote sensing data for comprehensive forest monitoring.[4]

The paper "Taxi Pricing Analysis under Government Price Regulation: A Case Study of Shenzhen Taxi Market" by Lixian Lin, Yuling Zhang, and Liang Ge provides significant insights into the regulation of taxi prices under government oversight. This study develops a practical model to determine the average price per taxi ride, aiming to maximize consumer surplus while ensuring the sustainability of taxi operators. The authors propose a four-step method for estimating initial charges, initial distances, and distance-based charges for different distance segments, addressing common issues such as short-distance ride refusals by taxi drivers.[5]

The authors build upon a foundation of previous research on taxi service economics under various regulatory frameworks, such as the works of Douglas (1972), De Vany (1975), Arnott (1996), Cairns and Liston-Heyes (1996), and Yang et al. (2002). These studies have explored factors like customer waiting times and the complex interplay between customers and operators, but the models they propose are often too intricate for practical implementation by government bodies[5].

By focusing on the Shenzhen taxi market, this study illustrates the applicability of their proposed pricing model in a real-world context. The model is designed to be simpler and more feasible for government use, recommending variable distance-based charges to ensure fairness and efficiency. This method contrasts with traditional fixed-distance charges and aims to prevent issues such as drivers favoring long-distance over short-distance rides[5].

In summary, this paper contributes to the body of literature on regulated taxi pricing by presenting a more practical and adaptable approach, grounded in a comprehensive case study of Shenzhen. This makes it a valuable reference for policymakers and researchers interested in urban transportation economics and regulation.

III. Dataset

The dataset used in this analysis comprises New York City taxi trip data from January 2013, amounting to approximately 2.5 GB before decompression. Although this study focuses specifically on data from January 2013, the dataset is part of a larger collection that spans the entire year. We chose to analyze only the January data due to the large size of the full dataset, which makes it more manageable. Analyzing a single month's data allows the code to run faster, enabling quicker model building, testing, and debugging. This one-month dataset is sufficient to initiate our analysis and assess the model's performance.

Another dataset used in this study is the *yellow_tripdata_2020-01.parquet*, part of the New York City Taxi and Limousine Commission (TLC) trip record data. This dataset provides a comprehensive record of taxi trips taken in New York City during January 2020. The data is stored in the Parquet format, which is efficient for both storage and query performance, making it suitable for large-scale data processing.

The key features of this dataset include:

- **Vendor ID**: An identifier for the taxi service provider (e.g., two-letter code indicating which company dispatched the taxi).
- **Pickup Datetime**: The date and time when a passenger was picked up.
- **Dropoff Datetime**: The date and time when a passenger was dropped off.
- **Passenger Count**: The number of passengers in the taxi.
- **Trip Distance**: The distance of the trip in miles.
- **Rate Code ID**: An identifier for the rate type (e.g., standard rate, JFK Airport rate, Newark Airport rate, etc.).
- **Store and Forward Flag**: A flag indicating whether the trip record was stored in the vehicle's memory before being sent to the vendor.
- **Pickup Location**: Latitude and longitude coordinates of the pickup point.
- **Dropoff Location**: Latitude and longitude coordinates of the dropoff point.

- **Payment Type**: The payment method used by the passenger (e.g., credit card, cash, no charge, etc.).
- **Fare Amount**: The fare for the trip.
- **Extra**: Any extra charges incurred (e.g., night surcharge, peak hour surcharge).
- **MTA Tax**: A tax collected by the Metropolitan Transportation Authority.
- **Tip Amount**: The tip amount given by the passenger.
- **Tolls Amount**: Any tolls paid during the trip.
- **Improvement Surcharge**: A surcharge applied to the trip to fund infrastructure improvements.
- **Total Amount**: The total cost of the trip, including all charges and taxes.

This dataset is instrumental for various analyses related to urban transportation. Researchers can explore trip durations, distances, fare structures, and passenger behaviors. For instance, the dataset allows for the examination of peak travel times, popular routes, and the economic aspects of taxi operations. This information can be used to develop predictive models for traffic congestion, optimize route planning, and improve dynamic pricing models. Moreover, the data provides valuable insights for urban planners and policymakers aiming to enhance transportation infrastructure and services in New York City.

IV. Conclusion

References

[1] E. Nagaesvara, M. Aadhithya, and A. Malini, "An Analysis of Machine Learning Techniques for Forest Cover Type Classification," presented at the 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), DOI: 10.1109IC3I59117.2023, 2023.

[2] N. Ferreira, J. Poco, H. T. Vo, J. Freire and C. T. Silva, "Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips," in IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pp. 2149-2158, Dec. 2013

[3] Omid Ghaffarpasand, Francis D. Pope, Telematics data for geospatial and temporal mapping of urban mobility: New insights into travel characteristics and vehicle specific power, Journal of Transport Geography, Volume 115, 2024, 103815, ISSN 0966-6923, https://doi.org/10.1016/j.jtrangeo.2024.103815. (https://www.sciencedirect.com/science/article/pii/S0966692324000243)

[4] UCI Machine Learning Repository, "Forest Cover Type Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset. [Accessed: May 25, 2024].

[5] L. Lin, Y. Zhang, and L. Ge, "Taxi Pricing Analysis under Government Price Regulation: A Case Study of Shenzhen Taxi Market," in *2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*, Changchun, China, Dec. 2011, pp. 1534-1539. doi: 10.1109/TMEE.2011.6199500.