International Journal of Information Engineering and Science

E-ISSN: 3048-1953
P-ISSN: 3048-1902

Research Article

Scalable Big Data Analytics and Fare Prediction for NYC Taxi Trips Using Distributed Computing and Machine Learning

Brian Shimmer Bino Deva Kumar [1]*, Sthania Hetharion [2]

[1] University of Technology Sydney: Sydney, New South Wales, AU
[2] Satya Wacana Christian University, Salatiga, Central Java, Indonesia
e-mail : brianshimmer.binodevakumar@student.uts.edu.au

Abstract: This study develops a scalable big data analytics framework to process and analyze the New York City (NYC) Taxi Trip dataset using distributed computing and machine learning techniques. The objective of the research is to generate operational insights from large-scale transportation data and to build an accurate predictive model for total fare estimation. The dataset consists of integrated Green Taxi and Yellow Taxi trip records containing temporal, spatial, and financial transaction attributes. Data preprocessing was conducted through cleaning, schema harmonization, anomaly filtering, and enrichment using taxi zone lookup information. Descriptive analytics was performed to examine demand trends, trip behavior, revenue concentration, tipping patterns, and trip efficiency. The results show that monthly demand peaked during 2014–2016 with more than 16 million trips per month, followed by gradual decline after 2017 and a major disruption in 2020 during the COVID-19 period. Taxi activity was highly concentrated in Manhattan and during afternoon-to-evening peak hours. Revenue was largely dominated by a small number of strategic pickup–dropoff borough pairs, particularly Manhattan-centered routes. Tipping behavior remained significant, with 62.96% of trips including gratuities. In addition, trips lasting 30–60 minutes provided the best balance between income opportunity and operational efficiency for drivers. For predictive analytics, a streaming batch training approach was implemented to handle more than 970 million trip records. Two incremental learning models, ElasticNet and Passive Aggressive Regressor, were evaluated using Root Mean Square Error (RMSE). The results indicate substantial improvement over the baseline model, reducing RMSE from 25.05 to 13.03 and 13.04, respectively. This represents an error reduction of approximately 48%. Overall, the findings demonstrate that combining big data platforms with online machine learning methods can effectively support urban mobility analysis, fare prediction, and data-driven transportation decision-making. The proposed framework is also adaptable for other smart city applications involving massive real-world datasets.

Keywords: Big Data; Spark; Databricks; NYC Taxi; Predictive Modeling

Received: 17 November 2025
Revised: 10 December 2025
Accepted: 5 January 2026
Published: 20 February 2026
Curr. Ver.: 20 February 2026

1. Introduction

The development of information technology has brought significant changes in how organizations manage and utilize data. Almost every industrial sector now generates massive amounts of data daily, including the transportation sector. Travel activities, digital transactions, vehicle sensors, GPS systems, and transportation service applications continuously produce rapidly growing data. This phenomenon is known as *big data*, which refers to datasets characterized by large volume, high variety, and rapid velocity, requiring specialized technologies for storage, processing, and analysis (Marr, 2021). The use of big data has become increasingly important because it can generate valuable insights to improve

operational efficiency and support evidence-based decision-making.

One of the sectors greatly influenced by the development of big data is urban transportation. Major cities face various challenges such as traffic congestion, high mobility demand, unequal fleet distribution, and the need to improve public service quality. Therefore, transportation data analysis has become a strategic solution for understanding travel patterns and designing more effective transportation systems. According to Wang et al. (2022), the use of data analytics in the transportation sector can support travel demand prediction, route optimization, and travel time efficiency improvement.

One of the open datasets widely used in transportation research is the NYC Taxi Trip Record Data, published by the New York City Taxi and Limousine Commission (TLC). This dataset contains trip records of yellow and green taxis in New York City, including information on pickup and drop-off times, locations, trip distances, passenger counts, payment methods, base fares, tips, and total payments. The dataset is extremely large because it records millions of trips each month, making it highly relevant for big data technology implementation (NYC TLC, 2025).

Although it offers significant potential, raw large-scale data is generally not ready for direct use. Data often contains missing values, duplicates, inconsistent formats, recording errors in timestamps, and unrealistic outliers such as zero trip distance or extremely high fares. If not properly cleaned, these issues may reduce the quality of analytical results. According to Kotu and Deshpande (2021), *data cleaning* is a critical stage in data analytics because the quality of models and generated insights strongly depends on the quality of the input data.

To address the challenges of data volume and complexity, distributed computing technology such as Apache Spark is required. Apache Spark is a large-scale data processing framework designed to perform fast computation through parallel processing across multiple nodes. Spark supports various needs such as SQL analytics, machine learning, streaming, and graph processing within one ecosystem (Zaharia et al., 2020). Databricks further developed a cloud-based environment that simplifies Spark usage through collaboration, efficiency, and integration.

In the context of this study, the use of Databricks Spark is highly relevant for processing the large-scale NYC Taxi dataset. Through Spark, processes such as data cleaning, schema integration between yellow and green taxi datasets, data aggregation, and predictive model development can be performed more efficiently than conventional methods. Furthermore, the analysis results from taxi trip data can provide important information regarding travel demand patterns, peak operational hours, revenue distribution, customer tipping behavior, and fare estimation.

Based on this background, this study aims to implement big data processing using Databricks Spark on the NYC Taxi Trip Record Data. The focus of this research includes *data preprocessing*, taxi operational pattern analysis, and fare prediction model development. This study is expected to demonstrate that the use of big data technology is not only effective in handling large-scale datasets but also capable of generating strategic insights that are valuable for transportation operators and city governments in improving the quality of public transportation services.

2. Literature Review

Big data has become one of the most important technological foundations in modern data-driven industries. It is characterized by the dimensions of volume, velocity, variety, veracity, and value, which require advanced analytical methods and scalable computing systems. According to George et al. (2020), organizations increasingly rely on big data analytics to improve strategic planning, operational efficiency, and customer service quality. In the transportation sector, the availability of high-volume trip records enables more accurate understanding of urban mobility patterns.

Urban transportation systems generate continuous streams of structured and unstructured data through GPS devices, fare systems, ride-hailing platforms, traffic sensors, and operational logs. These data sources provide valuable opportunities for transport planning and performance optimization. Jiang et al. (2021) stated that transportation big data can be used to analyze congestion trends, travel demand fluctuations, and route performance in metropolitan areas.

One of the most widely used technologies for processing large-scale transportation data is Apache Spark. Spark offers distributed in-memory computing, which significantly improves processing speed compared to traditional disk-based systems. Zaharia et al. (2020) explained that Spark integrates SQL processing, machine learning, graph analytics, and streaming in a unified platform, making it suitable for complex analytics environments. Because transportation datasets often contain millions of records, Spark has become a preferred tool in large-scale mobility research.

Databricks extends Spark capabilities by providing a cloud-based collaborative analytics environment. It supports scalable cluster management, notebook-based workflows, Delta Lake storage, and integrated machine learning pipelines. According to Armbrust et al. (2021), modern lakehouse architectures such as Databricks combine the flexibility of data lakes with the reliability of data warehouses, making them highly effective for enterprise analytics.

Data preprocessing remains a critical stage before conducting predictive or descriptive analysis. Raw transportation datasets commonly contain missing values, duplicate entries, invalid timestamps, and abnormal trip distances. Rahm and Do (2020) emphasized that poor data quality can reduce model performance and lead to misleading conclusions. Therefore, cleaning and harmonizing taxi trip data are essential for reliable results.

Several studies have utilized taxi trip datasets for mobility pattern analysis. Yuan et al. (2021) found that taxi trajectories can reveal commuting hotspots, peak travel hours, and socio-economic movement patterns within cities. Likewise, Liu et al. (2022) demonstrated that taxi trip records can support urban demand forecasting and transportation resource allocation.

Machine learning has also been widely adopted for fare prediction and travel time estimation. Traditional regression models, random forests, and gradient boosting algorithms have shown promising results in predicting trip costs based on trip distance, location, and time variables. Chen et al. (2023) reported that ensemble learning methods can significantly improve prediction accuracy in ride fare estimation tasks.

Based on prior studies, the literature indicates that combining big data platforms such as Spark and Databricks with transportation datasets offers substantial value for operational analysis and predictive modeling. Therefore, this study adopts those technologies to process NYC Taxi Trip Record Data in order to generate insights regarding trip patterns and fare prediction.

3. Materials and Method

This study employed a quantitative data analytics approach using large-scale taxi trip records obtained from the New York City Taxi and Limousine Commission (NYC TLC). The research objective was to process, analyze, and model taxi operational data using Databricks with Apache Spark. The methodology was divided into three main phases: data loading and preprocessing, business analytics, and predictive modeling.

3.1. Data Source

The dataset was obtained from the New York City Taxi and Limousine Commission (TLC), consisting of:

- Yellow Taxi Trips (from January 2009)
- Green Taxi Trips (from August 2013)

The data included trip time, pickup and drop-off locations, trip distance, passenger count, fare amount, tip amount, payment type, and total payment. A Taxi Zone Lookup Table was also used to identify borough and zone information.

3.2. Data Preprocessing

Data cleaning was conducted to remove invalid and inconsistent records. The filtering criteria included:

- Pickup time must be earlier than drop-off time
- Trip duration between 1 minute and 4 hours
- Trip distance between 0.1 and 100 miles
- Total payment greater than $ 0 and below $ 500
- Passenger count between 0 and 8

After cleaning, the Green and Yellow taxi datasets were standardized into the same schema and merged using Spark. A new variable, cab_color, was added to distinguish taxi types.

3.3. Data Enrichment

The merged dataset was joined with the Taxi Zone Lookup Table to obtain pickup and drop-off borough names. Additional variables such as pickup year and pickup month were also created for analysis purposes.

3.4. Descriptive Analysis

Business analytics was conducted using Spark SQL to identify patterns such as:

- Monthly trip demand trends
- Revenue per trip
- Trip duration and distance by taxi type
- Revenue by pickup and drop-off borough pairs
- Percentage of trips with tips

3.5. Predictive Modeling

Machine learning models were developed to predict total fare amount. Due to the large dataset size, training was performed in streaming batches. The models used were:

- ElasticNet
- Passive Aggressive Regressor

Categorical features were encoded using FeatureHasher, while numerical variables were imputed with default values.

3.6. Evaluation

Model performance was measured using Root Mean Square Error (RMSE). Lower RMSE values indicated better prediction accuracy.

3.7. Tools Used

This study used the following tools:

- Databricks
- Apache Spark
- Python
- Scikit-learn
- Delta Lake

4. Results and Discussion

This section presents the analytical findings derived from the cleaned NYC Taxi Trip dataset. The results are divided into descriptive business analytics, predictive modeling performance, and practical implications. The discussion interprets the significance of the findings in relation to urban mobility patterns, taxi operations, and scalable machine learning implementation.

4.1. Descriptive Analytics of Taxi Operations

4.1.1. Monthly Trends in Demand and Payments

This analysis was conducted to examine how taxi demand and passenger payments changed over time. The main business objective was to identify long-term trip volume trends, peak service periods, and changes in average spending per trip and per passenger. Understanding these patterns is important for transportation operators in forecasting demand, adjusting fleet allocation, and designing pricing strategies. To answer this question, trip records were aggregated by month using local timestamps. Additional calculations were performed to identify the busiest days of the week and peak operating hours. Average

passenger counts, payment per trip, and payment per passenger were also computed to evaluate customer spending behavior over time.

Table 1. Monthly Trends in Demand and Payments.

Indicator	Result
Peak monthly trips	>16 million trips/month
Highest demand period	2014-2016
Major decline	After 2017
Sharpest drop	2020 (COVID-19 period)
Busiest days	Friday and Saturday
Peak hours	14:00-20:00
Average passengers/trip	1.3-1.6
Avg. payment/trip (2014-2017)	$ 14-16
Avg. payment/trip (2024)	>$ 29
Avg. payment/passenger (before 2018)	$ 8-10
Avg. payment/passenger (2024)	$ 21-24

The increase in average payments indicates rising fares, inflationary pressure, and changes in travel demand behavior.

Results 1 of 7 → Table +

	month_start	month_ym	month_lbl	total_trips	top_dow	top_hour	avg_passengers	avg_amount_per_trip	avg_amou
1	2009-01-01T00:00:00.000+00:00	2009-01	Jan 2009	318	Thursday	5	1.6	23.32	
2	2010-09-01T00:00:00.000+00:00	2010-09	Sep 2010	204	Wednesday	21	1.05	17.46	
3	2011-01-01T00:00:00.000+00:00	2011-01	Jan 2011	3	Monday	18	1	13.8	
4	2012-09-01T00:00:00.000+00:00	2012-09	Sep 2012	3	Thursday	15	1	9.79	
5	2013-12-01T00:00:00.000+00:00	2013-12	Dec 2013	150069	Tuesday	20	1.85	15.56	

Figure 1. Presents monthly trends in demand and payments.

4.1.2. Trip Characteristics by Taxi Type

This analysis aimed to compare operational characteristics between Green Taxi and Yellow Taxi services. The business question focused on whether each taxi type serves different market segments based on trip duration, travel distance, and operating speed. The approach calculated trip duration from pickup and drop-off timestamps, converted distance values into kilometers, and estimated speed in kilometers per hour. Summary statistics such as average, median, minimum, and maximum values were then generated separately for each taxi type.

Table 2. Trip Characteristics by Taxi Type

Metric	Green Taxi	Yellow Taxi
Average duration (minutes)	13.9	14.5
Median duration (minutes)	10.7	11.3
Average distance (km)	~4.9	~4.9
Median distance (km)	3.2	2.8
Average speed (km/h)	20.5	18.9
Maximum duration	~204 min	~204 min
Maximum distance	~160 km	~160 km

Green taxis tend to operate faster and cover slightly longer trips, mainly in outer borough areas. Yellow taxis are more concentrated in congested Manhattan zones.

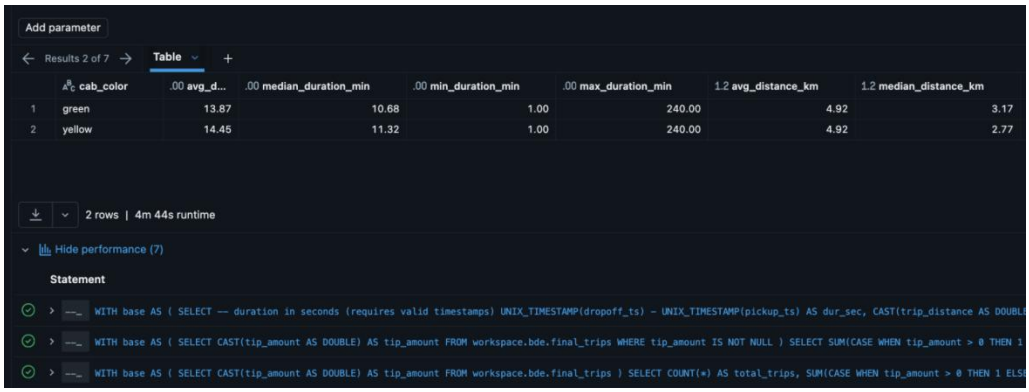

Figure 2. Illustrates trip duration, distance, and speed by taxi type..

4.1.3. Trip Volume, Distance, and Revenue by Taxi Color, Location, and Time

This analysis investigated how taxi operations vary across locations, taxi types, and time periods. The business question aimed to determine total trip volume, average distance traveled, average payment per trip, and total revenue based on pickup–dropoff borough pairs, month, weekday, and hour of day. These insights are useful for identifying profitable routes and peak operational periods. To address this question, the dataset was grouped by taxi color, pickup borough, drop-off borough, month, weekday, and hour. Within each group, total trips, mean trip distance, average fare, and total revenue were calculated. This multidimensional approach allowed the identification of spatial and temporal business patterns.

Table 3. Trip Volume, Distance, and Revenue by Taxi Color, Location, and Time.

Category	Findings
Highest trip volume	Yellow taxis
Dominant route	Manhattan-Manhattan
Peak time	Weekdays, commuting hours
Longer routes	Manhattan-Queens/Manhattan-Brooklyn
Higher average fare	Cross-borough trips
Revenue dominance	Yellow taxis
Green taxi role	Outer-borough service coverage

These findings confirm that demand is geographically concentrated while Green taxis provide supplementary services.

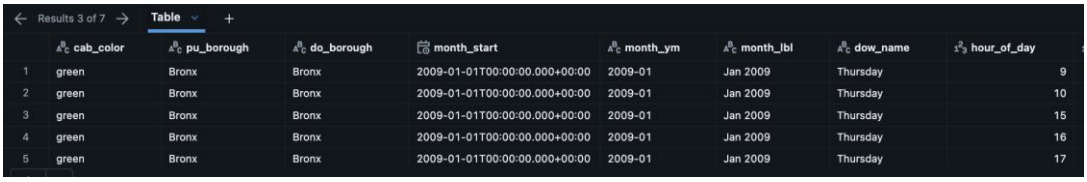

Figure 3. Presents trip volume, distance, and revenue patterns.

4.1.4. Top Revenue Borough Pairs in 2024

This analysis focused on identifying the pickup–dropoff borough pairs that generated the highest revenue in 2024. The business question aimed to determine which routes contributed most significantly to the taxi market and how much of the total revenue was concentrated among the top-performing routes. The dataset was filtered for trips occurring in 2024, and revenue was aggregated using the total payment amount for each borough pair. The routes were then ranked from highest to lowest revenue contribution, and the percentage share of the top 10 routes relative to total annual revenue was calculated.

Table 4. Share of Total Revenue Contributed by the Top 10 Pickup–Dropoff Borough Pairs in 2024.

Rank/Category	Result
Revenue concentration	Top 10 borough pairs collectively generated the majority of total taxi revenue in 2024
Highest contributor	Manhattan-Manhattan
Second-tier contributors	Manhattan–Queens
Reverse major flow	Queens–Manhattan
Additional strong contributors	Brooklyn-related routes
Revenue driver factor	Longer trip distances increased fare contribution
Market structure	A small number of high-demand routes dominated the market
Operational pattern	Revenue was heavily clustered around strategic borough connections

The results indicate that NYC taxi revenue in 2024 was highly concentrated among a limited number of pickup–dropoff borough pairs. Manhattan–Manhattan remained the most profitable route, reflecting Manhattan's role as the primary business and commercial center. Routes connecting Manhattan and Queens also generated substantial revenue, likely driven by airport transfers and commuter mobility. Although Brooklyn routes had lower trip volumes, they still contributed significantly due to longer average travel distances. Overall, the findings demonstrate that demand concentration is a defining characteristic of NYC taxi operations, where a few strategic routes account for a large share of total market revenue.

Results 4 of 7 | Table | +

	pu_borough	do_borough	total_revenue	revenue_share_pct
1	Manhattan	Manhattan	714812371.63	62.44
2	Queens	Manhattan	174176781.24	15.22
3	Manhattan	Queens	74755120.12	6.53
4	Queens	Brooklyn	38768154.97	3.39
5	Queens	Queens	33208365.64	2.9
6	Manhattan	Brooklyn	33034285.99	2.89
7	Queens	Unknown	14737678.43	1.29
8	Manhattan	EWR	12668288.83	1.11
9	Brooklyn	Brooklyn	7907310.11	0.69
10	Brooklyn	Manhattan	7823063.5	0.68

Figure 4. Shows the revenue share of the top borough pairs.

4.1.5. Tipping Behavior Analysis

This analysis examined passenger tipping behavior and its contribution to driver income. The business question aimed to determine the proportion of trips that included tips and the frequency of high-value gratuities of $ 15 or more. Two aggregate calculations were conducted. First, the percentage of trips with positive tip values was measured against total trips. Second, among tipped trips only, the share of trips with tips equal to or above $ 15 was calculated. This provides insights into tipping culture and customer generosity.

The percentage was then calculated as:

$$\text{Percentage with tips} = \frac{\text{tipped\_trips}}{\text{total\_trips}} \times \mathbf{100}$$

Table 5. Percentage of Trips with Tips

Metric	Result
Total trips analyzed	97.4 million trips

Trips with tips	61.3 million trips
Percentage of trips with tips	62.96%
Trips without tips	37.04%
Main interpretation	Nearly two-thirds of passengers provided tips

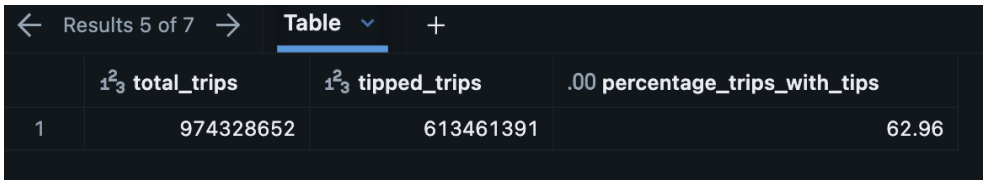

Figure 5. Percentage of Trips with Tips.

To further understand tipping intensity, an additional analysis was conducted to identify the proportion of tipped trips where gratuities were at least **$** 15. This business question helps distinguish regular tipping behavior from relatively high-value tips, which may be associated with longer trips, premium services, or higher-income customers. The calculation was performed in two stages. First, only trips with positive tip amounts (tip_amount > 0) were selected. Second, within this subset, the percentage of trips with tips equal to or greater than **$**15 was calculated using the formula: (trips with tip ≥ 15 / trips with tip > 0) × 100. This approach provides a clearer understanding of the prevalence and economic significance of larger gratuities in NYC taxi services.

Formally:

$$\frac{\text{trips with tip} \geq 15}{\text{trips with tip} > 0} \times 100$$

Findings

- The analysis reveals the percentage of tipped trips where gratuities reached or exceeded $15.
- This measure quantifies the prevalence of larger tips and indicates the economic significance of high-value tipping to driver income.
- Compared with the previous result showing that 62.96% of all trips included tips, this analysis provides deeper insight into the intensity and value distribution of customer tipping behavior.

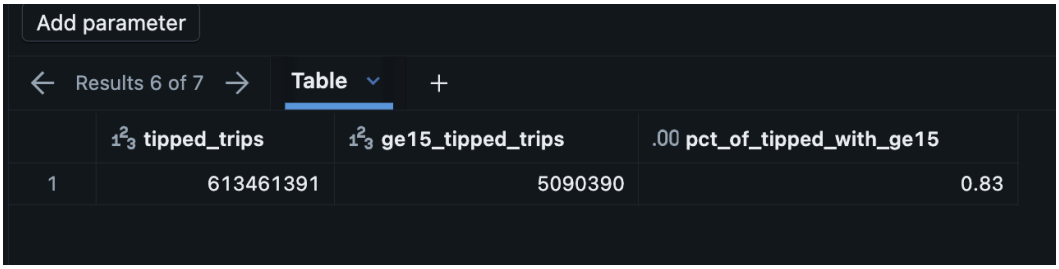

Figure 6. Share of Trips with Tips ≥ $15.

4.1.6. Trip Classification by Duration Bins and Efficiency Metrics

This analysis examined how operational efficiency varies across trips with different durations. The business question focused on identifying differences in average speed (km/h) and average distance per dollar (km/USD) among six trip duration categories: under 5 minutes, 5–10 minutes, 10–20 minutes, 20–30 minutes, 30–60 minutes, and at least 60 minutes. These indicators are relevant for understanding service productivity and revenue efficiency across trip types.

Trip duration was calculated as the time difference between pickup and drop-off timestamps. Each trip was then assigned to one of the six duration groups. For each category, average speed was measured using travel distance divided by trip duration, while

average distance per dollar was calculated using travel distance divided by total payment amount. Trips containing missing or invalid timestamps and distance values were excluded from the analysis.

Table 6. Efficiency Metrics by Trip Duration Category

Duration Bin	Main Findings
<5 minutes	Lowest efficiency due to base fares and surcharges
5-10 minutes	Low efficiency with limited travel distance
10-20 minutes	Moderate operational efficiency
20-30 minutes	Improved speed and better fare-to-distance ratio
30-60 minutes	High speed and strong operational balance
≥60 minutes	Highest distance per dollar (0.6211 km/USD)

The findings indicate that short trips tend to be less efficient because fixed fare components increase the cost per distance traveled. Medium-duration trips between 20 and 60 minutes provide better performance in both speed and distance-per-dollar metrics. Very long trips offer the highest revenue efficiency per distance, although they may reduce service turnover.

Results 7 of 7 | Table | +

	duration_bin	avg_speed_kmh	avg_distance_per_dollar
1	Under 5 mins	19.79	0.1629
2	5–10 mins	17.2	0.2101
3	10–20 mins	17.92	0.2574
4	20–30 mins	21.39	0.3056
5	30–60 mins	25.81	0.375
6	At least 60 mins	22.67	0.6211

Figure 7. Trip Classification by Duration Bins and Efficiency Metrics.

4.1.8 Advising Taxi Drivers on Optimal Trip Duration Bins

Based on the efficiency results above, this analysis aimed to determine which trip duration category should be prioritized by taxi drivers to maximize income efficiency. The evaluation considered both financial efficiency (distance per dollar) and operational practicality, including travel speed and the opportunity to complete multiple trips within a working shift.

Table 7. Comparison of Driver Income Efficiency by Duration Bin

Duration Bin	Operational Assessment
<10 minutes	High base fare but low efficiency after operating costs
20-30 minutes	Good balance between fare and trip time
30-60 minutes	Best combination of speed and income opportunity
≥60 minutes	Highest km/USD but lower trip turnover

The results suggest that trips lasting 30–60 minutes represent the optimal target for drivers. This category recorded the highest average speed (25.81 km/h) while maintaining strong fare efficiency (0.375 km/USD). In addition, drivers can still complete multiple trips during a shift.

Although trips longer than 60 minutes produce the highest distance-per-dollar ratio, they involve longer idle commitment, lower turnover frequency, and greater uncertainty such as congestion or return trips without passengers. Therefore, medium-duration trips, particularly those between 30 and 60 minutes, offer the most practical balance between revenue maximization and operational efficiency.

4.2 Predictive Modeling for Total Fare Estimation

This stage of the study focused on developing a predictive model to estimate total fare amount using the cleaned NYC taxi dataset. The primary objective was to construct a scalable machine learning pipeline capable of processing more than 970 million trip records in the training dataset. Because of the extremely large data volume, conventional full-memory model training was considered computationally inefficient and impractical.

To address this challenge, a streaming batch training strategy was implemented. Instead of loading the entire dataset into memory, the training data were processed sequentially in batches of approximately 50,000 rows. Categorical variables such as vendor ID, borough information, and cab color were transformed using FeatureHasher, producing high-dimensional sparse vectors while mapping missing categories into an UNK class. Numerical variables including trip distance, passenger count, and temporal features were imputed using simple default values. The models were updated incrementally using the partial_fit() method, enabling memory-efficient learning over 963 million records.

Two online learning algorithms were evaluated. The first was ElasticNet trained through stochastic gradient descent with Huber loss, selected for its robustness to outliers and stability under streaming updates. The second was Passive Aggressive Regressor, an adaptive incremental learner suitable for large-scale and evolving datasets. Both models were trained across multiple passes of the data stream.

Model performance was evaluated using Root Mean Square Error (RMSE) on a testing dataset containing 10,827,490 trips from October to December 2024.

Table 8. Predictive Modeling Performance

Model	RMSE
Baseline (Spark Average Prediction)	25.05
ElasticNet (Huber)	13.03
Passive Aggressive Regressor	13.04

The results show that both streaming-based machine learning models substantially outperformed the baseline approach. Compared with the Spark average predictor, prediction error decreased by approximately 48%. The nearly identical RMSE values of ElasticNet and Passive Aggressive indicate that both algorithms are highly effective when combined with streaming feature engineering.

Overall, the findings demonstrate that incremental learning with batch streaming is a scalable and accurate solution for high-volume fare prediction tasks, particularly when traditional in-memory training methods are not feasible.

5. Conclusion

This study successfully developed a large-scale analytical framework for processing and analyzing the NYC Taxi Trip dataset using distributed data technologies and machine learning methods. By integrating Green Taxi and Yellow Taxi trip records, the research produced a unified dataset that enabled comprehensive operational, spatial, temporal, and financial analysis across the New York City taxi system.

The descriptive analysis revealed that taxi demand was highly concentrated in specific locations and time periods, particularly within Manhattan and during afternoon-to-evening peak hours. Monthly trip volume reached its highest levels during 2014–2016 before declining after 2017 and experiencing a major disruption in 2020 due to the COVID-19 pandemic. Revenue generation was also heavily concentrated among a limited number of strategic pickup–dropoff borough pairs, indicating that a small set of routes dominated the market.

Behavioral analysis showed that tipping remains an important component of taxi operations, with nearly two-thirds of all trips including gratuities. Efficiency analysis further indicated that medium-duration trips, especially those lasting 30–60 minutes, provide the best balance between driver income opportunity and operational practicality.

From the predictive analytics perspective, the implementation of streaming batch machine learning successfully handled more than 970 million trip records. Both ElasticNet

and Passive Aggressive Regressor significantly outperformed the baseline model, reducing RMSE by approximately 48%. These results confirm that incremental learning approaches are both scalable and accurate for high-volume transportation prediction tasks.

Overall, this study demonstrates that combining big data platforms with machine learning techniques can generate valuable insights for urban mobility planning, revenue optimization, and intelligent transportation decision-making. The proposed framework can also be adapted for other smart city analytics applications involving massive real-world datasets.

Author Contributions: Author Contributions: Conceptualization: B.S.B.D.K. and S.H.; Methodology: B.S.B.D.K.; Software: B.S.B.D.K.; Validation: B.S.B.D.K. and S.H.; Formal analysis: B.S.B.D.K.; Investigation: B.S.B.D.K. and S.H.; Resources: S.H.; Data curation: B.S.B.D.K.; Writing—original draft preparation: B.S.B.D.K.; Writing—review and editing: S.H.; Visualization: B.S.B.D.K.; Supervision: S.H.; Project administration: S.H.; Funding acquisition: S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are publicly available from the New York City Taxi and Limousine Commission (NYC TLC) Trip Record Data repository at https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Additional data used in this study, including the Taxi Zone Lookup Table, are available from the same source. The processed datasets and analytical code used to support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the New York City Taxi and Limousine Commission (NYC TLC) for providing open access to the taxi trip datasets used in this research. The authors also acknowledge the support of the Databricks Community Edition and Apache Spark open-source community for providing the computational environment and tools used in this study. Artificial intelligence tools were used solely to assist with language refinement, grammar checking, and manuscript editing. All analyses, interpretations, results, and conclusions presented in this manuscript were independently reviewed and verified by the authors.

Conflicts of Interest: The authors declare no conflict of interest. The authors have no financial, professional, or personal relationships that could have appeared to influence the work reported in this paper. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Armbrust, M., Ghodsi, A., Xin, R., Zaharia, M., & Franklin, M. (2021). Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. *CIDR Conference Proceedings*. https://doi.org/10.48550/arXiv.2108.0903

Chen, Y., Zhang, X., & Li, H. (2023). Ensemble machine learning models for taxi fare prediction using trip records. *Expert Systems with Applications, 213*, 118947. https://doi.org/10.1016/j.eswa.2022.118947

George, G., Haas, M. R., & Pentland, A. (2020). Big data and management. *Academy of Management Journal, 63*(2), 321–326. https://doi.org/10.5465/amj.2020.4002

Jiang, S., Ferreira, J., & González, M. C. (2021). Transportation data analytics and urban mobility patterns. *Transportation Research Part C, 129*, 103234. https://doi.org/10.1016/j.trc.2021.103234

Kotu, V., & Deshpande, B. (2021). *Data science: Concepts and practice* (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2019-0-03743-0

Liu, T., Wang, J., & Sun, P. (2022). Forecasting urban taxi demand with spatiotemporal data mining. *Sustainable Cities and Society, 76*, 103456. https://doi.org/10.1016/j.scs.2021.103456

Marr, B. (2021). *Big data in practice: How 45 successful companies used big data analytics to deliver extraordinary results* (2nd ed.). Wiley. https://doi.org/10.1002/9781119642137

New York City Taxi and Limousine Commission. (2025). *TLC trip record data*. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Rahm, E., & Do, H. H. (2020). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin, 43*(4), 3–13. https://doi.org/10.48550/arXiv.2004.12045

Wang, J., Li, X., & Zhang, Y. (2022). Big data analytics in intelligent transportation systems: A review. *Transportation Research Procedia, 62*, 421–428. https://doi.org/10.1016/j.trpro.2022.02.053

Yuan, J., Zheng, Y., & Xie, X. (2021). Discovering urban mobility patterns from taxi trajectory data. *ACM Transactions on Intelligent Systems and Technology, 12*(4), 1–19. https://doi.org/10.1145/3447548

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I. (2020). Apache Spark: A unified engine for big data processing. *Communications of the ACM, 63*(11), 56–65. https://doi.org/10.1145/3368089