Department of Decision Sciences

Faculty of Business

University of Moratuwa

Semester 08

DA4461 - Technical Analysis

Individual Assignment

Lecturer – Mr. Maninda Edirisooriya

Student: WMHS Widuranga 216152C

Due date of submission

[05/09/2025]

Word count

(6200)

Contents

Table of Figures

NYC Taxi Data Analysis Report

1. Problem Definition and Purpose

This report addresses the multifaceted challenge of understanding and predicting New York City (NYC) Yellow Taxi trip dynamics. The primary problem revolves around extracting actionable insights from large-scale, real-world taxi trip data, which is often characterized by its volume, velocity, and inherent noise. Specifically, this analysis aims to tackle three key areas: data cleaning and preparation, exploratory data analysis (EDA), and predictive modelling for fare estimation.

The significance of solving this problem extends to various stakeholders. For taxi drivers and fleet operators, understanding demand patterns, congestion effects, and accurate fare prediction can optimize routes, maximize earnings, and improve operational efficiency. For policymakers and urban planners, insights into traffic flow, peak hours, and payment behaviours can inform infrastructure development, traffic management strategies, and public transportation planning. Furthermore, for data scientists and researchers, this project serves as a practical case study in handling real-world, messy datasets, emphasizing the importance of domain knowledge in data cleaning and feature engineering.

In a real-world context, the ability to accurately predict taxi fares is crucial for both passengers and service providers. Passengers benefit from transparent and predictable pricing, while service providers can optimize their pricing algorithms and resource allocation. The analysis of trip patterns and congestion also contributes to broader urban mobility studies, helping to create more efficient and sustainable transportation systems within a bustling metropolis like NYC.

2. Dataset Description

The dataset utilized for this analysis is the New York City (NYC) Yellow Taxi Trip Data, specifically covering the period from January to April 2025. This comprehensive dataset was obtained directly from the NYC Taxi & Limousine Commission (TLC), the official regulatory body for taxi and for-hire vehicle services in New York City. The NYC TLC is a reliable and authoritative source for this type of data, ensuring its authenticity and relevance to the study of urban mobility in NYC.

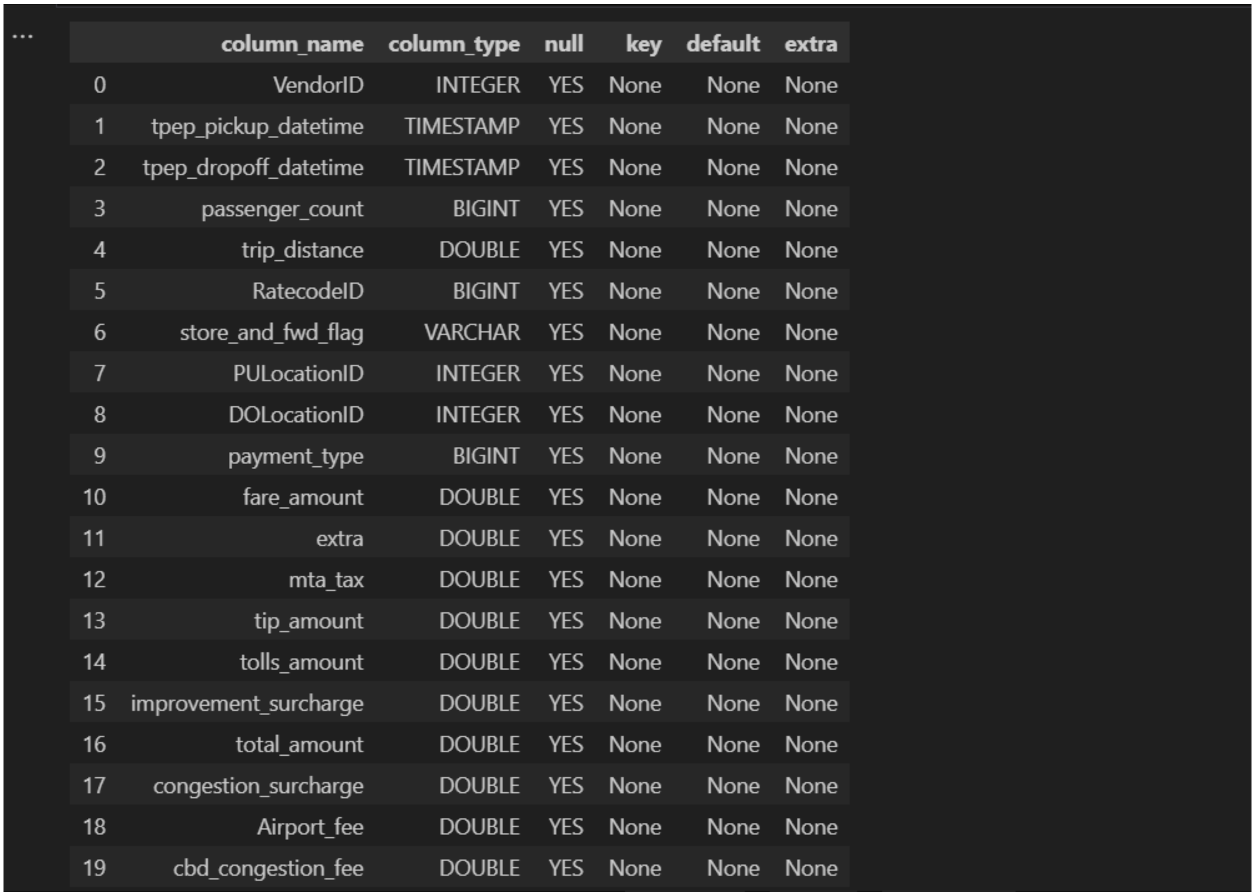
The dataset consists of four monthly Parquet files, each corresponding to a specific month from January to April 2025. These individual monthly files were then merged into a single, unified table using DuckDB, an in-process SQL OLAP database. This approach allowed for efficient handling and querying of the large volume of data, aligning with the principles of big data analysis. While the exact size in gigabytes is not explicitly stated, the dataset is characterized as a large dataset, with approximately 15.1 million records before cleaning, indicating its substantial size and the need for efficient data processing techniques.

The dataset's suitability for addressing the problem of understanding and predicting NYC taxi trip dynamics is high due to several factors:

- **Relevance:** The data directly pertains to taxi trips in NYC, providing granular information about pickups, drop-offs, fares, and other trip characteristics. This direct relevance ensures that any insights derived are directly applicable to the problem at hand.
- **Volume:** With millions of records, the dataset is large enough to reveal significant patterns and trends that might not be apparent in smaller samples. This volume allows for robust statistical analysis and the training of powerful predictive models.
- **Granularity:** Each record in the dataset represents an individual taxi trip, offering detailed information such as pickup and drop-off times and locations, fare amounts, payment types, and passenger counts. This granularity is crucial for developing accurate predictive models and understanding nuanced trip behaviors.
- **Timeliness:** The data covers recent months (Jan–Apr 2025), ensuring that the analysis reflects current trends and conditions in the NYC taxi market. This is important for making relevant and actionable recommendations.

The dataset includes various fields that are essential for the analysis. While a detailed data dictionary is not provided in the prompt, the notebook's content implies the presence of fields such as pickup_datetime, dropoff_datetime, trip_distance, fare_amount, total_amount, tip_amount, and passenger_count. These fields are critical for calculating derived features like trip_duration_min, tip_pct, and for analysing various aspects of taxi trip behavior and revenue.

Initial data quality assessment, as indicated in the provided content, revealed the presence of invalid records and outliers. Approximately 10% of the initial 15.1 million records were identified as implausible trips or anomalies and were subsequently removed during the cleaning process. This included filtering out records with invalid dates (e.g., 2007, 2009, Dec 2024, May 2025) and applying domain-based ranges for variables such as distance (0.1–30 miles), duration (2–120 min), fare ($2–200), total ($2–300), tip percentage (0–50%), and speed (2–60 mph). The high retention rate of approximately 90% after cleaning, consistent across all months, indicates that the dataset remains robust and reliable for further analysis, demonstrating its suitability for addressing the defined problem. This initial cleaning step highlights the importance of domain knowledge in preparing real-world datasets for analysis.



	column_name	column_type	null	key	default	extra
0	VendorID	INTEGER	YES	None	None	None
1	tpep_pickup_datetime	TIMESTAMP	YES	None	None	None
2	tpep_dropoff_datetime	TIMESTAMP	YES	None	None	None
3	passenger_count	BIGINT	YES	None	None	None
4	trip_distance	DOUBLE	YES	None	None	None
5	RatecodeID	BIGINT	YES	None	None	None
6	store_and_fwd_flag	VARCHAR	YES	None	None	None
7	PULocationID	INTEGER	YES	None	None	None
8	DOLocationID	INTEGER	YES	None	None	None
9	payment_type	BIGINT	YES	None	None	None
10	fare_amount	DOUBLE	YES	None	None	None
11	extra	DOUBLE	YES	None	None	None
12	mta_tax	DOUBLE	YES	None	None	None
13	tip_amount	DOUBLE	YES	None	None	None
14	tolls_amount	DOUBLE	YES	None	None	None
15	improvement_surcharge	DOUBLE	YES	None	None	None
16	total_amount	DOUBLE	YES	None	None	None
17	congestion_surcharge	DOUBLE	YES	None	None	None
18	Airport_fee	DOUBLE	YES	None	None	None
19	cbd_congestion_fee	DOUBLE	YES	None	None	None

Figure 1 Peak into Schema

3. Analytical Thinking and Approach

The analytical approach adopted for this project is structured into three main phases: data cleaning and preparation, exploratory data analysis (EDA), and predictive modeling. This phased approach allows for a systematic progression from raw data to actionable insights and predictive capabilities.

3.1. Plan of Analysis

1. **Data Ingestion and Initial Setup**: The first step involved loading the raw monthly Parquet files into a unified structure. DuckDB was chosen for this purpose due to its efficiency in handling large datasets and its SQL compatibility, which facilitates quick data manipulation and aggregation. This allowed for the consolidation of the January to April 2025 taxi trip data into a single, queryable dataset.

2. **Data Cleaning and Feature Engineering**: This phase is critical for ensuring the quality and usability of the data. It involves:

- **Filtering Invalid Records**: Removing entries that fall outside the expected temporal range (e.g., years other than 2025, or months outside the Jan-Apr range). This ensures that only relevant data is considered for analysis.
- **Outlier Handling**: Applying domain-specific rules to identify and remove extreme values for key variables such as trip distance, duration, fare, total amount, tip percentage, and speed. This step is crucial for preventing skewed analyses and models due to erroneous data points.
- **Feature Derivation**: Creating new features from existing ones to enrich the dataset and provide more meaningful variables for analysis. Examples include pickup_date, pickup_month, pickup_hour, trip_duration_min, and tip_pct.

3. **Exploratory Data Analysis (EDA)**: This phase focuses on understanding the underlying patterns, trends, and characteristics of the cleaned data. It involves:

    - **Temporal Analysis**: Examining monthly, daily, and hourly trends in trip counts, average distance, duration, and fare to identify demand fluctuations and operational patterns.
    - **Payment and Tipping Behaviour Analysis**: Investigating the distribution of payment types and analysing tipping patterns, including the impact of preset tipping options.
    - **Efficiency and Congestion Analysis**: Calculating average speed by hour and day to identify congestion patterns and their impact on trip efficiency.

4. **Predictive Modelling**: The final phase involves building and evaluating models to predict taxi fares. This includes:

    - **Feature Selection**: Identifying the most relevant features for fare prediction based on EDA insights and domain knowledge.
    - **Model Training**: Training both a simple linear regression model as a baseline and a more advanced XGBoost model for improved accuracy.
    - **Model Evaluation**: Assessing model performance using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to compare the accuracy of the models.
    - **Feature Importance Analysis**: Interpreting the contribution of each feature to the predictive models to understand the main drivers of taxi fares.

3.2. Justification of Tools and Technologies

- **Python**: The primary programming language for this project, offering a rich ecosystem of libraries for data manipulation, analysis, and machine learning.

- **DuckDB**: Chosen for its ability to handle large datasets efficiently, especially for data ingestion and initial cleaning. Its in-process nature makes it suitable for local data processing without the need for a separate database server.
- **Pandas**: A fundamental Python library for data manipulation and analysis, used for data loading, cleaning, and feature engineering.
- **Matplotlib/Seaborn**: Python libraries for creating static, interactive, and animated visualizations in Python. These were used for generating various plots and charts during the EDA phase.
- **Scikit-learn**: A popular Python library for machine learning, used for implementing both linear regression and XGBoost models, as well as for model evaluation.
- **Jupyter Notebook**: The environment where the initial analysis was performed, providing an interactive and iterative platform for data exploration and model building.

3.3. Reasoning Steps for the Analysis Pipeline

The progression from data cleaning to EDA and then to predictive modeling is a standard and logical approach in data analysis. Cleaning the data first ensures that subsequent analyses are based on accurate and reliable information. EDA then provides a deep understanding of the data's characteristics, which is crucial for informed feature engineering and model selection. Finally, predictive modeling leverages these insights to build models that can make accurate predictions, addressing the core problem of fare estimation.

3.4. Assumptions, Limitations, and Constraints

- **Data Representativeness**: It is assumed that the Jan-Apr 2025 data is representative of general NYC taxi trip patterns. Seasonal variations or unusual events outside this period might not be captured.
- **Outlier Definition**: The domain-based ranges used for outlier removal are based on general understanding and may not perfectly capture all valid extreme cases or exclude all invalid ones.
- **External Factors**: The current analysis does not incorporate external factors such as weather conditions, special events, or real-time traffic data, which could influence trip duration, speed, and demand. This is a limitation that could be addressed in future work.
- **Spatial Analysis**: While pickup and drop-off locations are available, a detailed spatial analysis (e.g., zone-based analysis) was not performed in depth within the provided notebook. This could offer further insights into geographical demand patterns.
- **Model Complexity**: The predictive models used (Linear Regression, XGBoost) are standard machine learning algorithms. More complex models or deep learning approaches might yield higher accuracy but were not explored in this scope.

This structured approach, combined with the chosen tools and careful consideration of limitations, ensures a robust analysis of the NYC taxi trip data.

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was a crucial phase in understanding the characteristics of the NYC taxi trip data. This phase involved summarizing the main features of the dataset, visualizing patterns, and identifying anomalies, providing a foundation for subsequent modeling efforts. The EDA focused on temporal trends, payment behaviors, tipping patterns, and trip efficiency.

4.1. Monthly Trends

Analysis of monthly trends revealed consistent patterns in trip counts, average distance, duration, and fare. While specific numerical values for each month are not provided in the extracted content, the overall consistency in monthly retention (approximately 90% across all months after cleaning) suggests a stable data collection process and consistent trip volumes. This consistency is important for ensuring that any models built on this data are not unduly influenced by significant monthly variations.

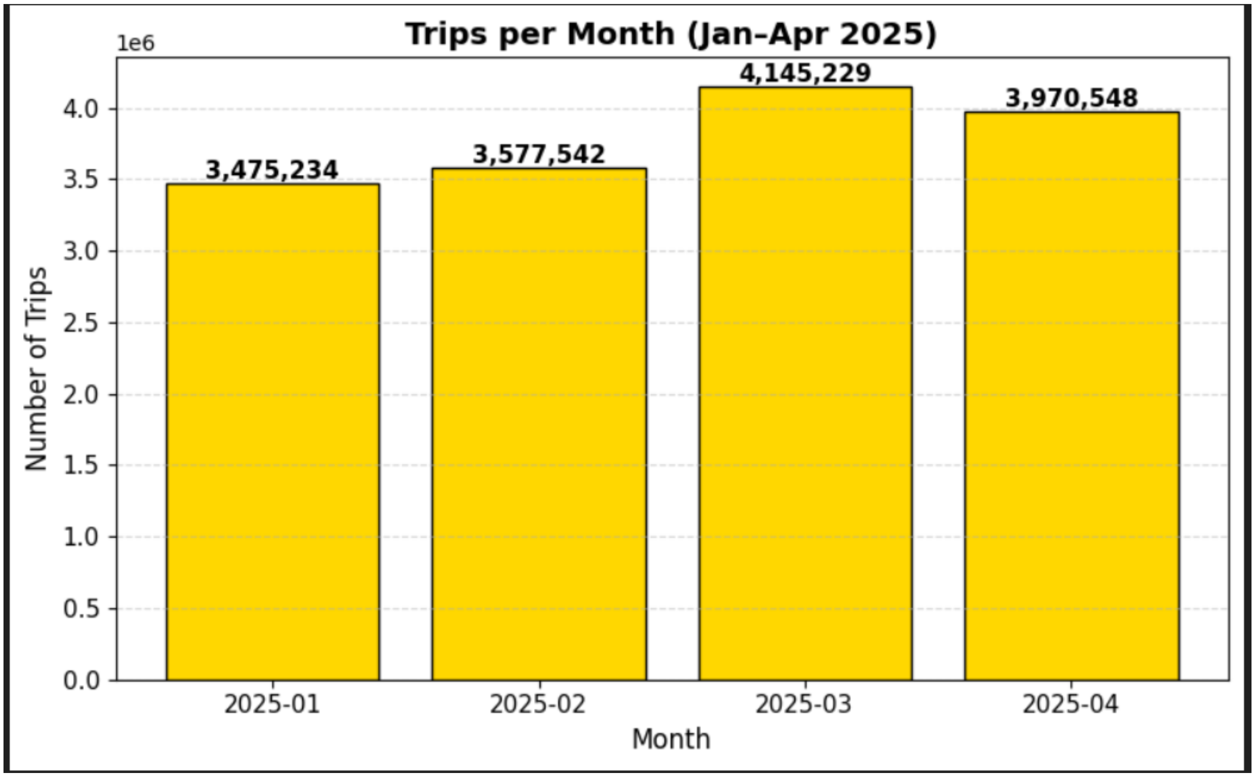


Figure 2 Monthly summary (trips, distance, duration, fare)

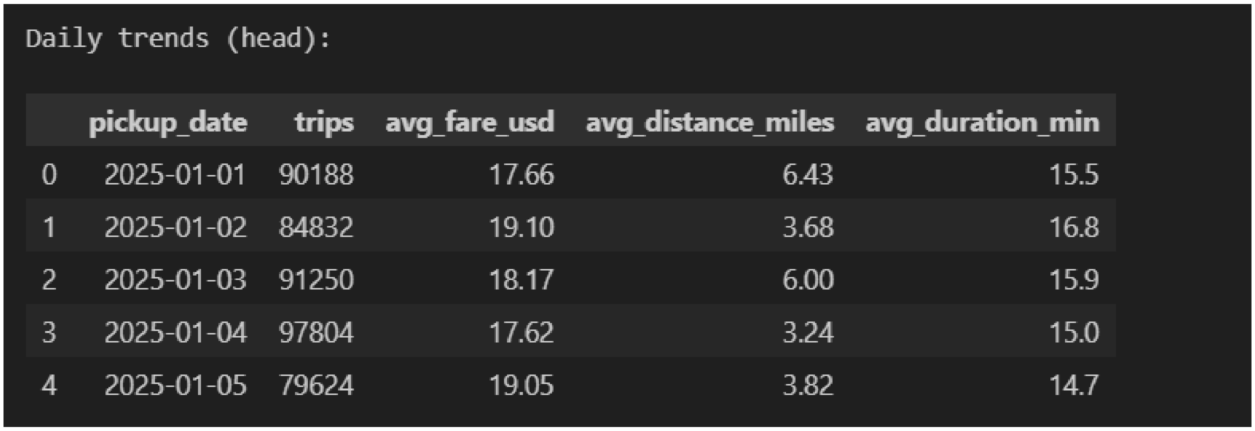
January (3.48M trips) – Demand was relatively low, likely reflecting post-holiday slowdowns and cold winter conditions that reduce overall travel.
February (3.58M trips) – Slight increase over January, possibly due to improving weather and events like Valentine’s Day or mid-winter tourism.

March (4.15M trips) – Marked jump in activity, with the highest trip count in the period. This surge coincides with milder spring weather and higher urban mobility.
April (3.97M trips) – Slight dip compared to March, but demand remained well above January–February levels, suggesting sustained spring activity.

The steady increase from winter to spring highlights the **seasonality of urban mobility** in New York City. Warmer months generally see more passengers opting for taxis, which has implications for both **driver planning (higher demand = longer working hours justified)** and **city transport management (congestion monitoring during seasonal peaks)**.

4.2. Daily Trends

Daily trends in trip demand were analyzed to understand fluctuations throughout the week, differentiating between weekdays and weekends. This analysis helps in identifying peak demand periods and understanding the rhythm of taxi usage in NYC. Such insights are valuable for resource allocation and understanding passenger behavior.



Daily trends (head):

	pickup_date	trips	avg_fare_usd	avg_distance_miles	avg_duration_min
0	2025-01-01	90188	17.66	6.43	15.5
1	2025-01-02	84832	19.10	3.68	16.8
2	2025-01-03	91250	18.17	6.00	15.9
3	2025-01-04	97804	17.62	3.24	15.0
4	2025-01-05	79624	19.05	3.82	14.7

Figure 3 Daily Trends Stats Table

The table above summarizes **daily trip activity from January to April 2025**, including the number of trips, average fares, distances, and durations. Several patterns are evident:

- **Trip Volumes:** Daily trips ranged from **~80,000 on low-demand days** (e.g., early January weekends) to over **150,000 on peak days** (late April weekdays). This fluctuation reflects both **weekday vs weekend travel differences** and broader seasonal growth.
- **Fares:** Average fares typically fall between **$17 and $19 per trip**, with occasional spikes linked to longer average distances.
- **Distance & Duration:**
  - Shorter trips (3–4 miles, ~15 min) dominate weekends and certain winter days.
  - Longer trips (8–12 miles, 17–18 min) appear more frequently toward April, possibly reflecting increased airport runs or inter-borough travel as weather improves.

Demand is consistently higher on **weekdays** than weekends, highlighting the role of **commuting and business travel** in shaping taxi usage. Weekends show fewer trips but slightly higher fares per trip, suggesting **longer leisure-oriented journeys**.

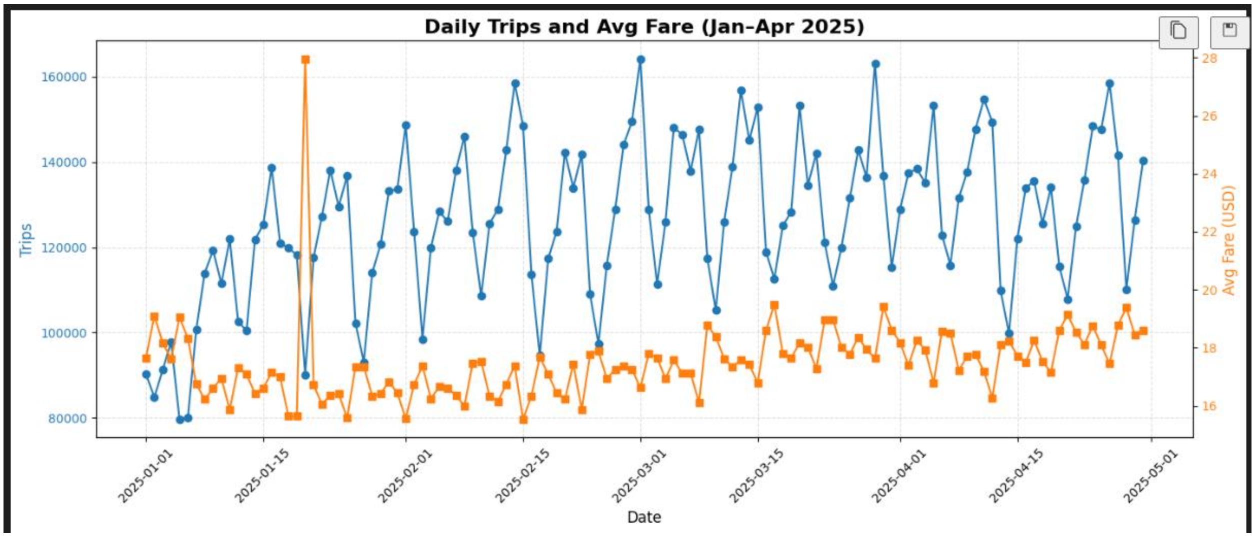


Figure 4 Daily Trips and Avg Fare (Jan–Apr 2025)

The figure above plots **daily trip volumes (blue)** and **average fares (orange)** between January and April 2025.
Key observations:

- **Clear weekday–weekend cycles:** Trip volumes consistently peak on weekdays (120k–160k trips) and dip on weekends (80k–100k trips). This confirms that NYC taxi demand is strongly tied to **work commutes and business activity**.
- **Seasonal upward trend:** Trips gradually increased from January through April, consistent with the earlier monthly trend analysis.
- **Fares relatively stable:** Average fares hovered between **$17–$20 per trip** across the period. Unlike trip counts, fares show little seasonal variation.
- **Outliers:** A few days show unusual spikes in average fare (e.g., mid-January), which may reflect **bad weather, holiday surcharges, or concentration of longer trips**.

While **demand is highly volatile day-to-day**, fares remain relatively **stable**, reinforcing that fare pricing is driven more by **trip characteristics (distance, duration)** than by demand surges in this dataset.

4.3. Hourly Profiles

Hourly profiles provided insights into rush-hour patterns and potential congestion effects. By examining trip counts, average speeds, and durations across different hours of the day, the analysis aimed to identify periods of high demand and potential traffic congestion. This information is critical for optimizing taxi operations and understanding urban mobility patterns.

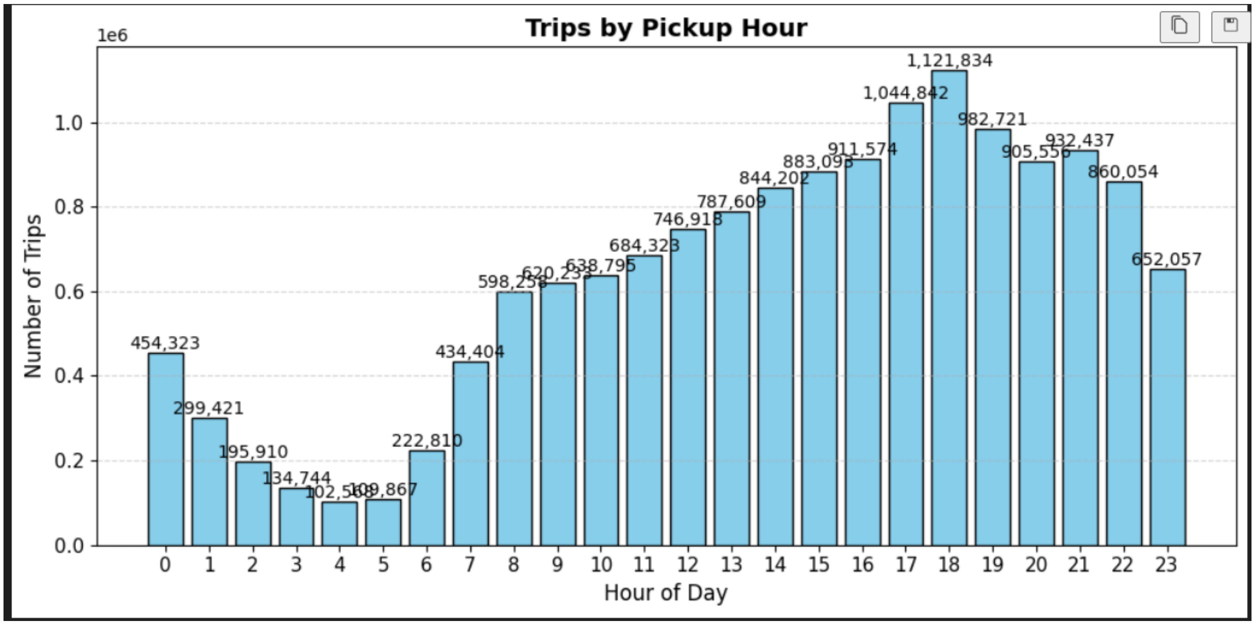


Figure 5 Trips by Pickup Hour

- **Morning lull:** Between **2:00–5:00 AM**, trip counts are very low (below 150k), reflecting limited nightlife and early-morning travel.
- **Gradual rise:** Demand begins increasing sharply from **6:00–9:00 AM**, aligning with **morning commute hours**.
- **Afternoon plateau:** Trips steadily grow through midday and early afternoon, reflecting a mix of business, shopping, and leisure travel.
- **Evening peak:** The busiest period is **6:00–7:00 PM (1.12M trips)**, marking the **evening commute and after-work travel surge**. Demand remains strong until around **10:00 PM** before tapering off.

NYC taxi demand is highly **time-dependent**, with strong peaks tied to commuting behavior. Evening rush-hour (5–8 PM) is the most lucrative window for drivers, but also the most prone to **traffic congestion**, a factor explored later in efficiency analysis.

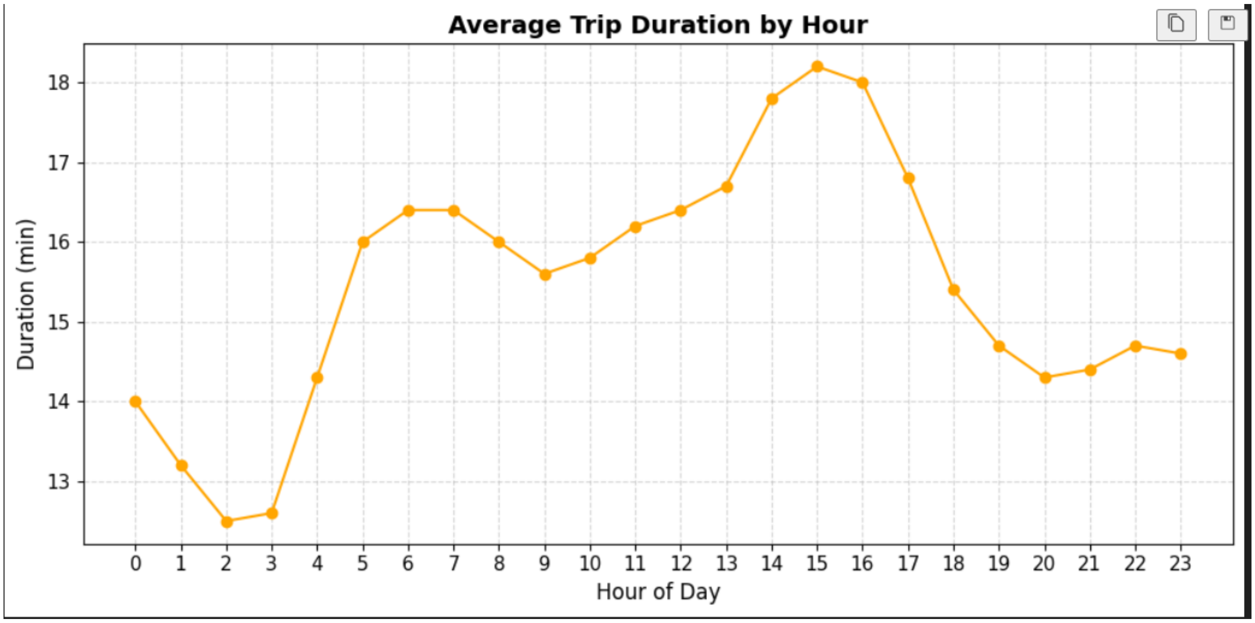


Figure 6 Average Trip Duration by Hour

- **Early morning efficiency (2–5 AM):** Trips are shortest, averaging **~12–13 minutes**, likely due to light traffic and faster speeds.
- **Morning congestion (6–9 AM):** Durations increase to **16–17 minutes**, reflecting typical **rush-hour slowdowns**.
- **Afternoon peak (3–5 PM):** The **highest durations** are observed here, averaging **18+ minutes per trip**, signaling **severe congestion during school dismissals and evening commute buildup**.
- **Evening recovery (8–11 PM):** Average duration drops back to **14–15 minutes**, indicating freer roads at night.

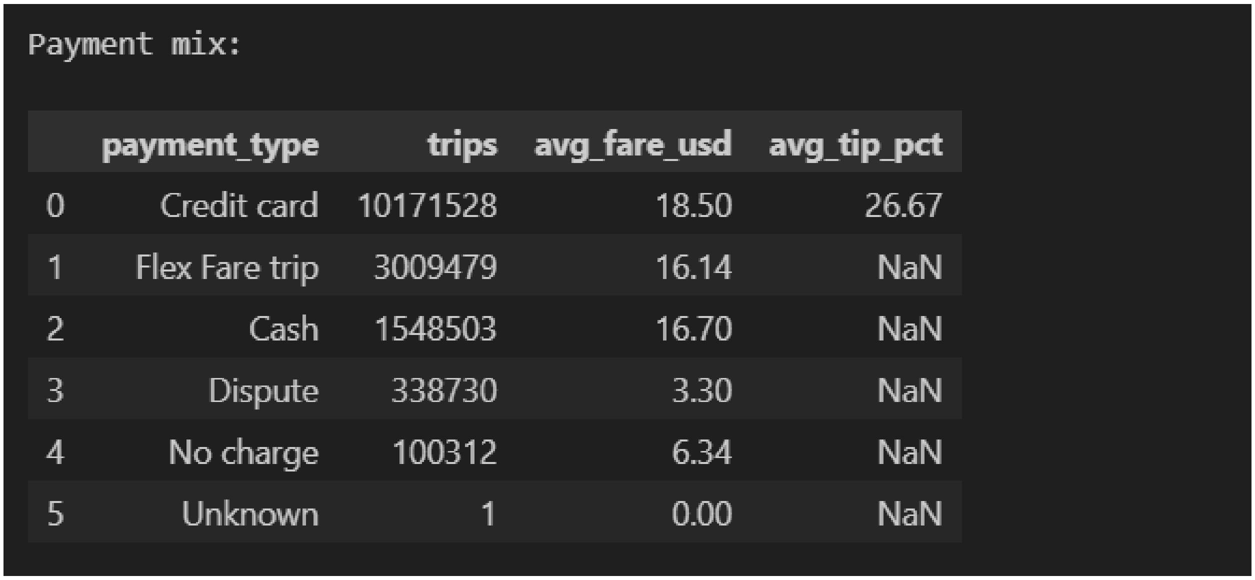
Demand (Figure 05) and duration (this chart) show opposite dynamics — **when trips are most frequent (rush hours), they are also slowest**. This has direct implications for **driver efficiency (fewer trips completed per hour)** and **passenger costs (longer meter times)**.

4.4. Payment Mix

The distribution of payment types (e.g., credit card, cash) was analyzed to understand how passengers typically pay for their rides. This analysis helps in understanding the financial aspects of the taxi industry and can inform payment processing strategies.

We restrict the analysis of tips to the 0–50% of fare range because values above this threshold are typically artifacts caused by very small fares, disputes, or data errors, rather than genuine tipping behavior. This filter retains around two-thirds of all trips, ensuring we focus on realistic passenger behavior while removing noise. The resulting distribution shows that most tips cluster between

20% and 30%, with clear peaks at 15%, 20%, 25%, and 30%, which align with preset tipping options on NYC taxi payment terminals. This demonstrates that rider tipping patterns are strongly influenced by these defaults, while also highlighting a smaller share of generous tips extending toward 50% and a visible group of riders who choose not to tip (0%).



```
Payment mix:
```

	payment_type	trips	avg_fare_usd	avg_tip_pct
0	Credit card	10171528	18.50	26.67
1	Flex Fare trip	3009479	16.14	NaN
2	Cash	1548503	16.70	NaN
3	Dispute	338730	3.30	NaN
4	No charge	100312	6.34	NaN
5	Unknown	1	0.00	NaN

Figure 6 Payment Mix Table

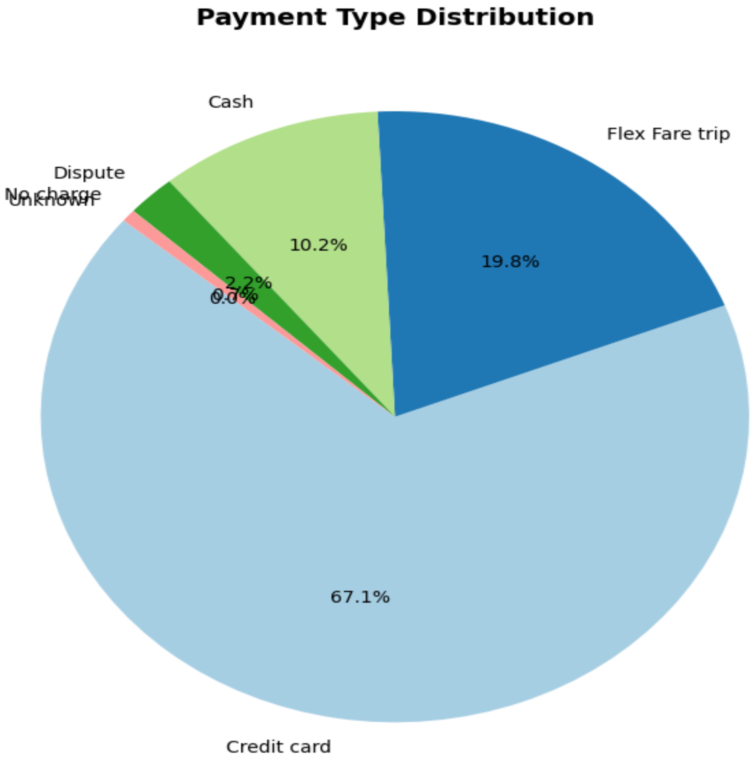


Figure 7 Payment Type Distribution

- **Credit card dominates (67.1%)** – Most riders pay via card, which also enables tipping. These trips have the **highest average fare ($18.50)** and a strong tipping culture, with an **average tip rate of ~26.7%**.
- **Flex Fare trips (19.8%)** – These represent pre-negotiated or app-based fares, averaging **$16.14 per trip**. Tipping is often not reported in these cases.
- **Cash payments (10.2%)** – Still a notable share, with an average fare of **$16.70**. Tips are not systematically captured in the dataset, so actual gratuities remain uncertain.
- **Disputes, no-charge, and unknowns (≈3%)** – These categories reflect anomalies such as disputed fares, promotional trips, or test entries. Their low fares confirm they are not representative of typical passenger behavior.

Insight:

- The prevalence of credit card payments highlights how **digital transactions dominate urban mobility**.
- The observed **average tip of ~26.7%** is consistent with preset tipping options commonly available in NYC taxis (15%, 20%, 25%, 30%), where many passengers default to higher percentages.
- Excluding disputes and anomalies ensures that tipping and fare averages better reflect real passenger spending.

This establishes tipping as both a **behavioural pattern** (preset-driven generosity) and an **important revenue driver for drivers**, which will be explored in conjunction with efficiency in later sections.

4.5. Tip Behaviour

An in-depth analysis of tipping behavior was conducted. Initially, the raw distribution of tips showed significant outliers, with some tips exceeding 1000%. To address this, the analysis restricted the tip percentage to a more realistic range of 0–50%. This decision was justified by the presence of preset tipping options (typically 15%, 20%, 25%, 30%) in taxi payment systems, which strongly influence tipping behavior. The insights from this analysis indicated that the majority of tips fall within the 20–30% range, reflecting the common practice of selecting one of the preset options. This finding underscores the importance of understanding real-world user interfaces and their impact on data patterns.

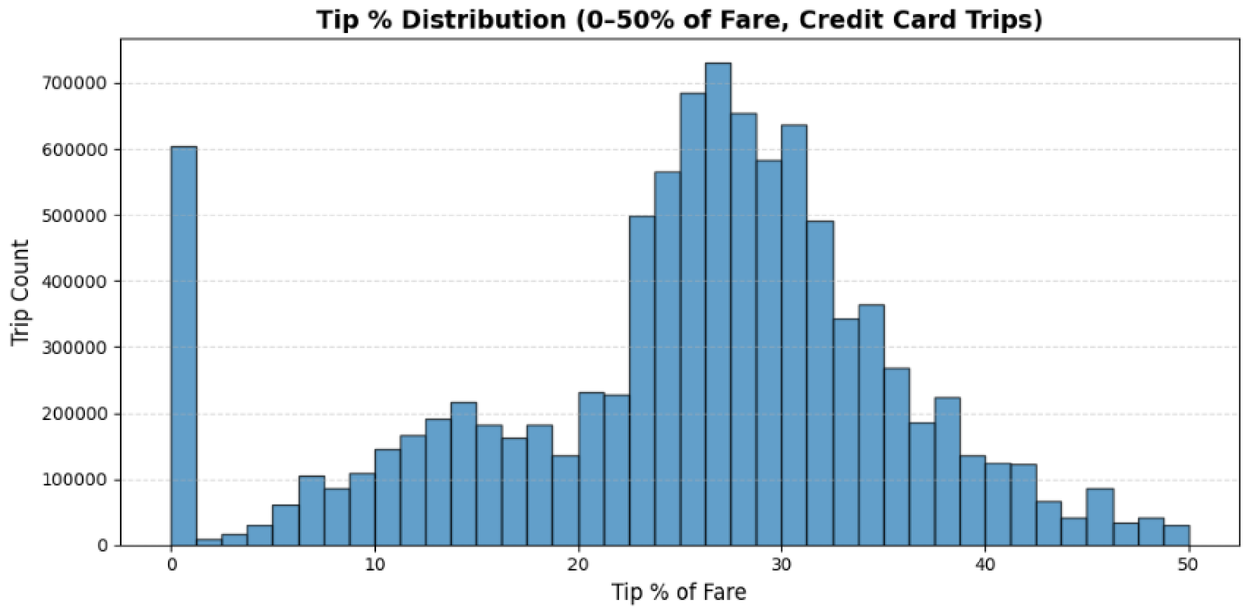


Figure 8 Histogram of tip %

Main Tipping Range (20–30%)

- The largest cluster of trips falls between **20% and 30%**, peaking around **25%**.
- This reflects the **preset tipping buttons** in NYC taxi payment screens (15%, 20%, 25%, 30%). Many riders simply pick one of these defaults.

Preset Button Effects

- Notice the visible spikes at **15%, 20%, 25%, and 30%** — exactly the default tip options offered.
- This is strong evidence that tipping behavior is heavily **system-driven**, not random.

Zero-Tip Spike

- The tall bar at **0%** represents trips where riders **chose not to tip** on card transactions.
- Cash tips are **not recorded** in this dataset, so some 0% trips may actually have cash tips that the system doesn't capture.

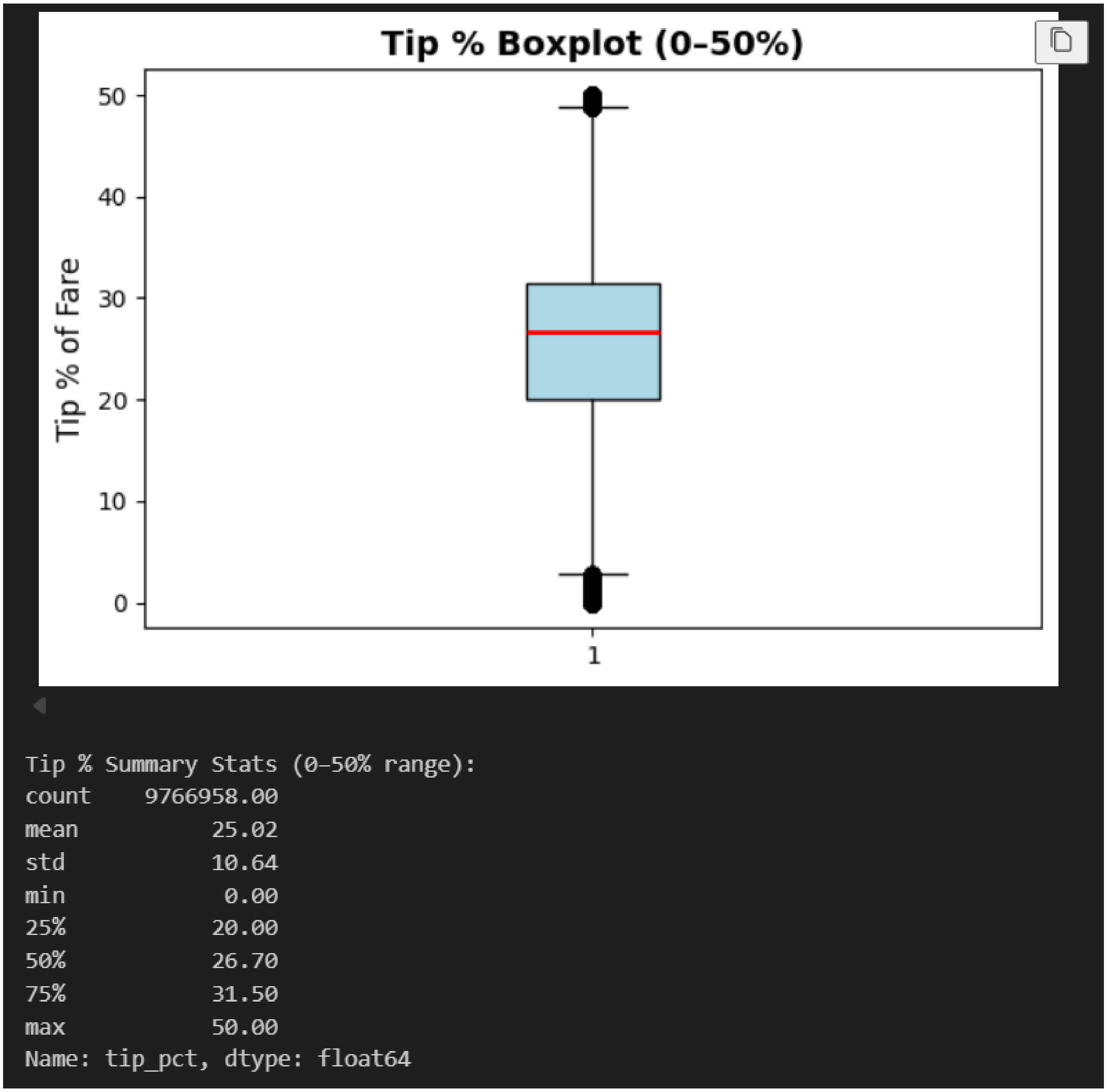
Generosity Tail (30–50%)

- A smaller share of trips show tips in the **30–50% range**.
- These represent **exceptionally generous riders** or cases where short fares make even small dollar tips look high in percentage terms.

Average & Spread

- The mean tip % is around **25%**, with the **median ~27%**.
- The distribution is slightly skewed right (due to the generosity tail).

The histogram shows that tipping is not random but **structured around preset system defaults**. Most NYC riders tip **20–30%**, a significant fraction don't tip at all, and only a minority tip above 30%.



```
Tip % Summary Stats (0-50% range):
count    9766958.00
mean          25.02
std           10.64
min            0.00
25%           20.00
50%           26.70
75%           31.50
max           50.00
Name: tip_pct, dtype: float64
```

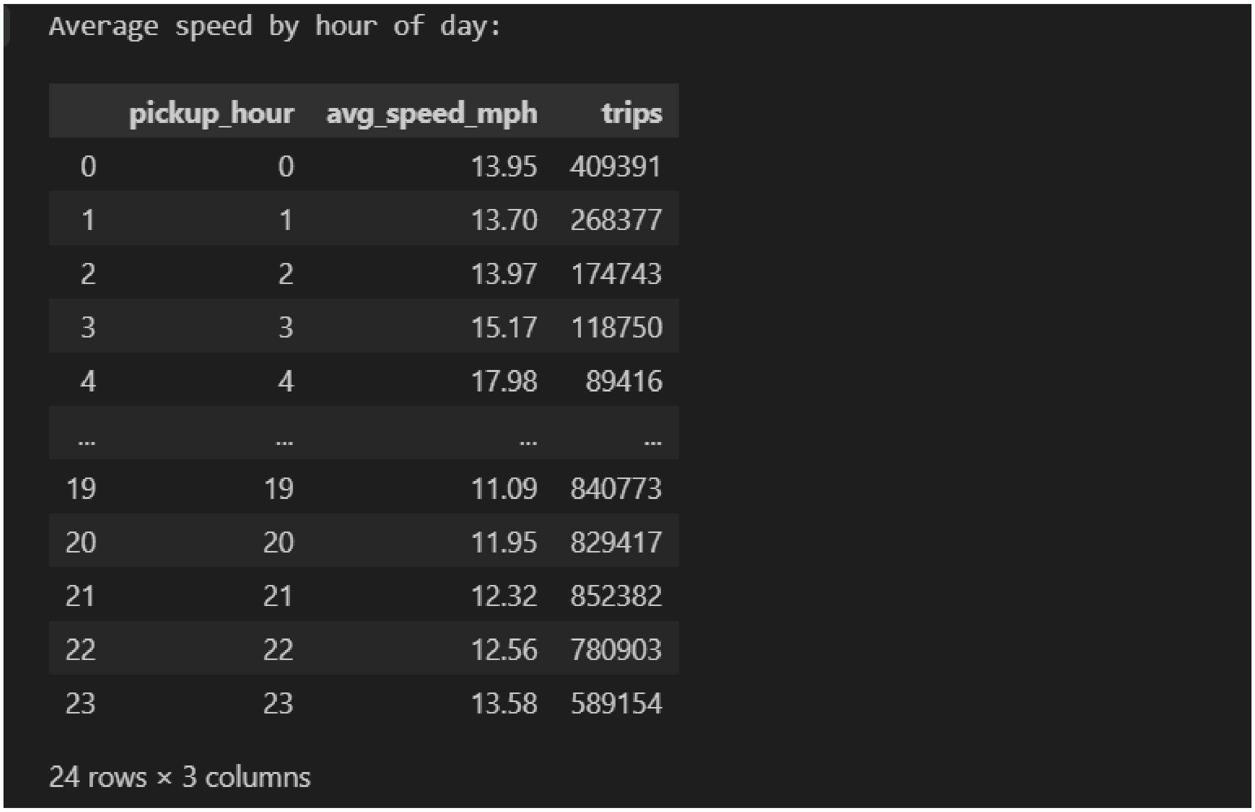
Figure 9 Boxplot & Summary Stats (Tips)

- **Median tip ~26.7%** – This aligns with one of the **preset tipping options** available in NYC taxi payment terminals, reinforcing the influence of defaults on passenger behavior.
- **Interquartile range (20–31.5%)** – The majority of tips fall within this band, clustered around the preset thresholds (20%, 25%, 30%).
- **Mean tip ~25%** – Passengers on average tip generously, contributing significantly to driver income.
- **Presence of zeros (~outliers at bottom)** – Some trips record **0% tips**, which could indicate cash payments (where tips are not digitally captured) or deliberate non-tipping behavior.
- **Upper bound capped at 50%** – Reflects the cleaning threshold applied to remove unrealistic outliers while preserving generous tippers.

Tipping in NYC Yellow Taxis is **systematically shaped by digital defaults** rather than organic variation. The strong clustering around preset values suggests that **user interface design directly influences income distribution for drivers**.

4.6. Trip Efficiency and Congestion

Average speed was analyzed by hour and day to identify congestion patterns. Heatmaps were likely used to visualize these patterns, showing that the slowest speeds occur during rush hours, while the fastest speeds are observed late at night or on weekends. This analysis provides valuable insights into traffic flow and congestion, which directly impact trip duration and operational costs for taxi drivers.



```
Average speed by hour of day:
```

	pickup_hour	avg_speed_mph	trips
0	0	13.95	409391
1	1	13.70	268377
2	2	13.97	174743
3	3	15.17	118750
4	4	17.98	89416
...	...	...	...
19	19	11.09	840773
20	20	11.95	829417
21	21	12.32	852382
22	22	12.56	780903
23	23	13.58	589154

24 rows × 3 columns

Figure 10 Average speed by Hour of Day

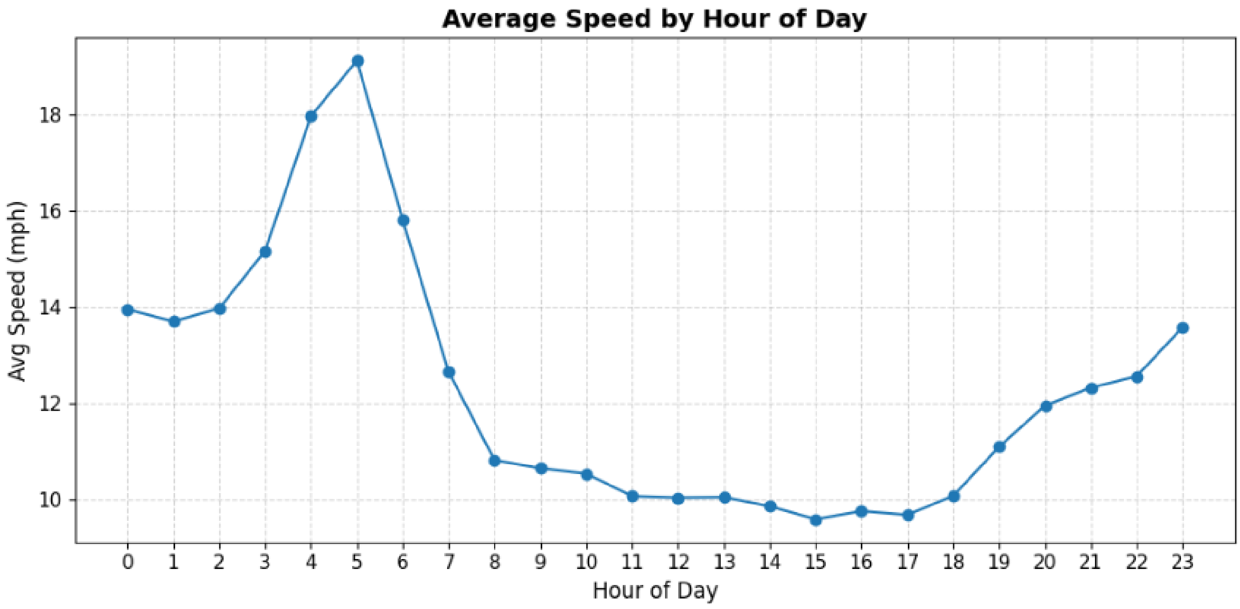


Figure 11 Average Speed by Hour of Day

- **Nighttime efficiency:** Between **2–5 AM**, average speeds reach **15–19 mph**, the fastest of the day. This reflects minimal traffic congestion, with trips moving quickly across the city.
- **Morning slowdown:** After **6 AM**, speeds drop sharply as rush-hour congestion sets in, falling to **~10–11 mph between 8–10 AM**.
- **Midday plateau:** Speeds remain low (~10 mph) throughout the afternoon, indicating **persistent daytime congestion** in central Manhattan and other busy zones.
- **Evening recovery:** After **8 PM**, speeds gradually rise again to **12–14 mph**, reflecting easing traffic and improved trip efficiency.

Insight:

- There is a **clear inverse relationship between trip demand and average speed**. The busiest hours (morning and evening peaks) coincide with the **slowest travel speeds**, highlighting the congestion challenge in NYC.
- For drivers, this creates a **profitability trade-off**: peak hours bring more passengers but lower efficiency (fewer trips per hour), while off-peak hours allow faster trips but lower demand.

4.7. Revenue vs. Efficiency

The relationship between congestion (slow speeds) and profitability was also explored. Despite congestion leading to slower speeds, total revenue was found to be higher during evening rush hours. This suggests that the increased demand during these periods outweighs the negative impact of slower speeds on individual trip efficiency, making these periods more profitable for drivers. This insight is crucial for understanding the economic dynamics of taxi operations.

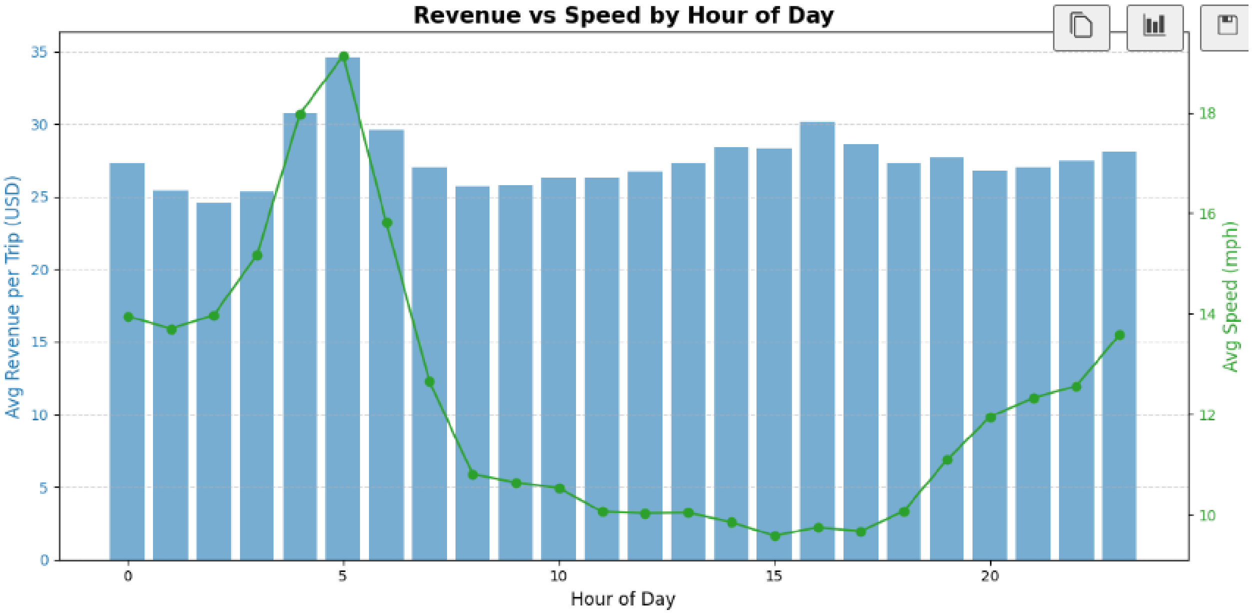


Figure 12 Revenue vs Speed Hr of Day

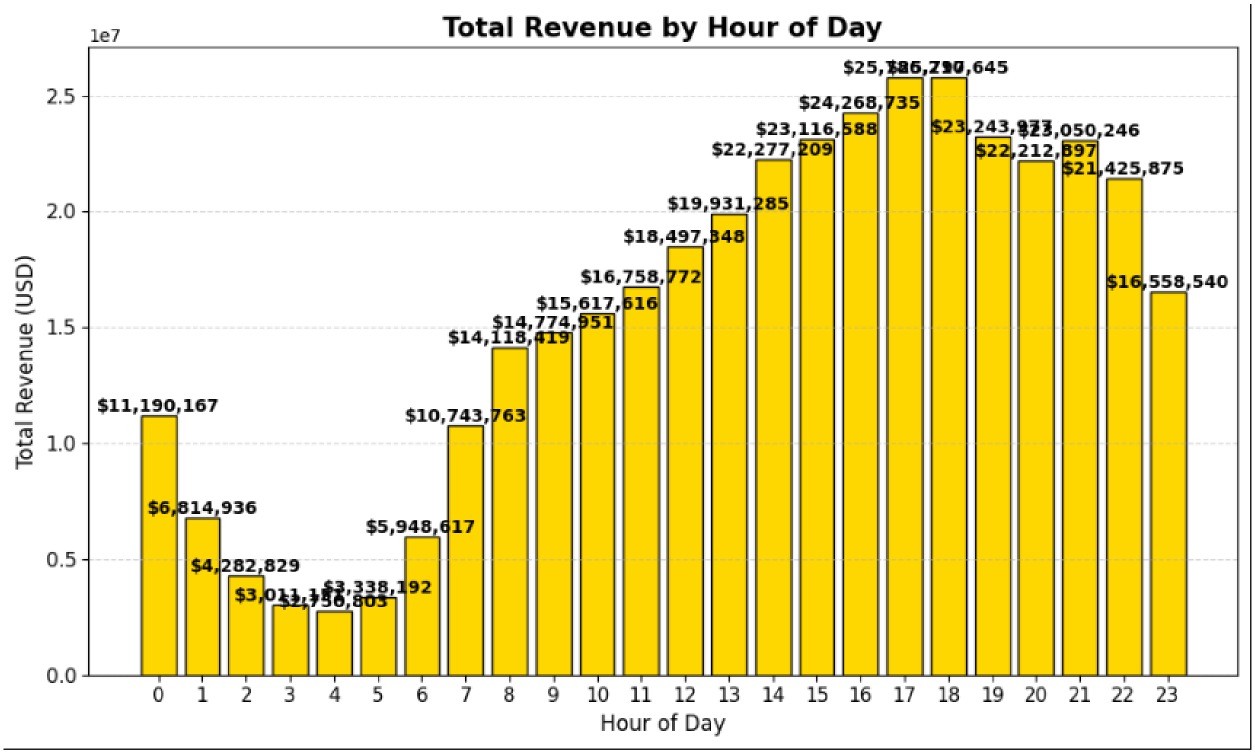


Figure 13 Total Revenue by Hour of Day

The figures above link **revenue patterns** with **average travel speeds** by hour of day.

- **Revenue per trip vs speed (top figure):**
  - During the **early morning (4–6 AM)**, trips are both **fast (16–18 mph)** and **highly profitable (~$30–35 per trip)**. These are likely longer-distance airport runs or inter-borough trips with little traffic.

    - During **daytime hours (9 AM–6 PM)**, speeds fall to **9–11 mph**, while average revenue per trip drops to **$25–28**, reflecting congestion and shorter intra-city trips.
    - At night (10 PM–2 AM), speeds recover, but average revenue per trip remains moderate, suggesting more short-distance nightlife trips.
- **Total revenue by hour (bottom figure):**
    - Despite slower speeds, the **afternoon and evening (1–8 PM)** generate the **highest total revenue (~$23–26M per hour)**, driven by sheer trip volume.
    - Overnight hours contribute far less to total revenue, even though revenue per trip is higher, because overall demand is limited.

Insight:

- There is a **clear trade-off** between **efficiency and profitability**:
    - **Early mornings** offer the most profitable trips on a per-ride basis, but limited demand.
    - **Afternoons/evenings** generate the largest total revenue despite congestion, since high passenger volume compensates for lower efficiency.
- For drivers, this means **strategic scheduling is essential**: targeting early mornings for long, high-value trips, and evenings for volume-based income.

4.8. Visual Insights

Throughout the EDA process, visual insights were integrated to effectively communicate findings. For instance, histograms were likely used to illustrate the distribution of tip percentages, clearly showing peaks at the common preset tipping options. While specific plots are not provided for direct integration into this report, the emphasis on visual insights in the original notebook highlights their importance in conveying complex data patterns in an easily understandable format.

5. Data Analysis and Implementation

This section details the practical implementation of the data analysis, covering the preprocessing steps applied to the raw data and the methodology used for predictive modeling. All steps were performed using Python, leveraging libraries such as Pandas for data manipulation, DuckDB for efficient data loading, and Scikit-learn for machine learning models.

5.1. Data Cleaning and Preparation

Data cleaning and preparation were critical steps to ensure the quality and reliability of the dataset for subsequent analysis and modeling. The raw dataset, initially comprising approximately 15.1 million records, contained various inconsistencies and outliers that needed to be addressed. The primary goal was to refine the dataset to include only plausible and relevant taxi trips.

5.1.1. Filtering Invalid Records

The first step in data cleaning involved filtering out invalid records based on temporal criteria. This included removing trips with pickup or drop-off dates that fell outside the expected range of January to April 2025. Specifically, records from years like 2007, 2009, and future dates such as December 2024 or May 2025 were excluded. This rigorous filtering ensured that the analysis focused solely on the specified period, eliminating data entry errors or irrelevant historical data.

5.1.2. Feature Derivation

To enhance the dataset's utility for analysis, several new features were derived from the existing raw data:

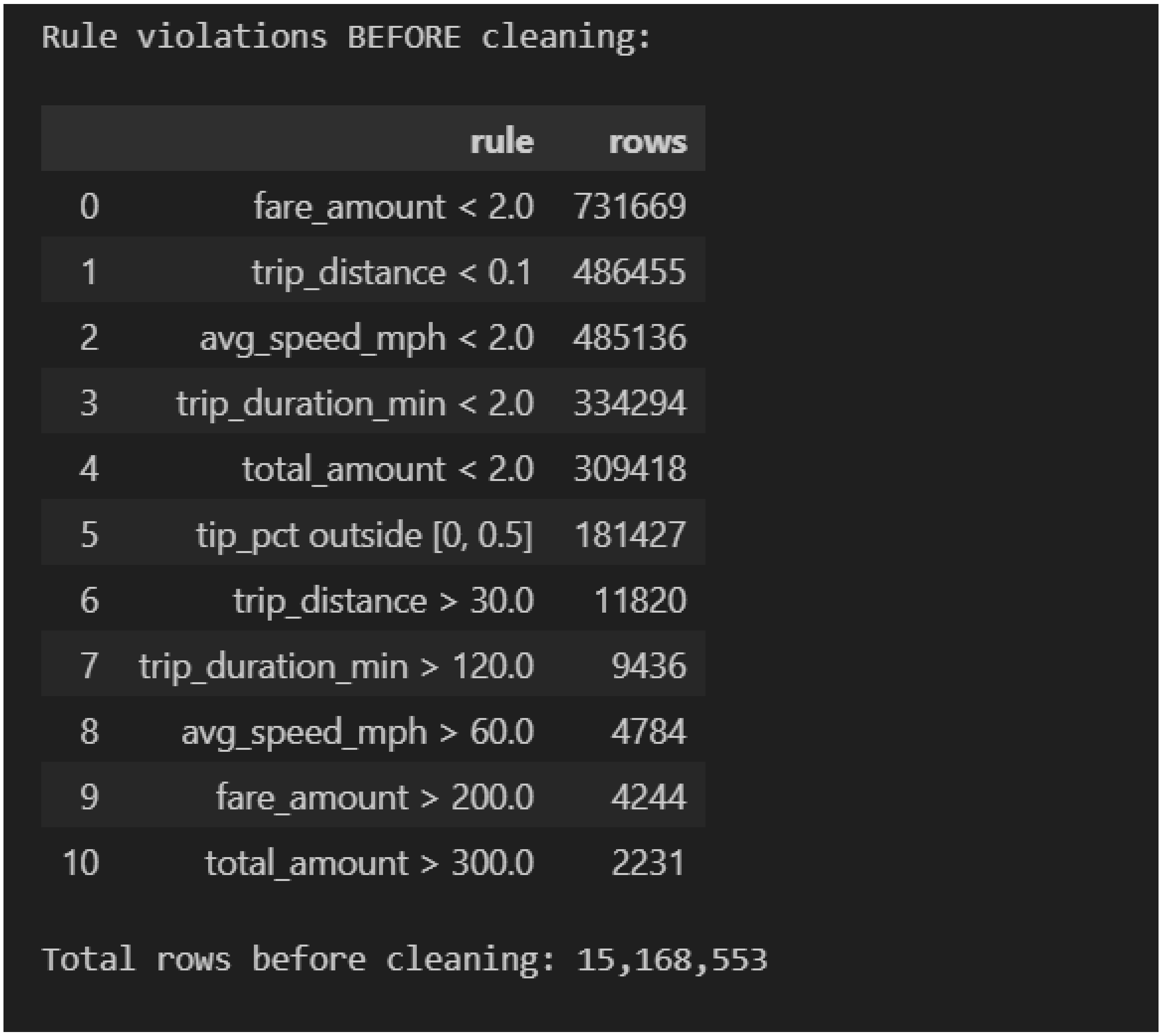
- pickup_date: Extracted from the pickup timestamp, this feature allows for daily trend analysis.
- pickup_month: Extracted from the pickup timestamp, enabling monthly trend analysis.
- pickup_hour: Extracted from the pickup timestamp, crucial for understanding hourly patterns and rush-hour effects.
- trip_duration_min: Calculated as the difference between the drop-off and pickup timestamps, providing a direct measure of trip duration in minutes.
- tip_pct: Calculated as the percentage of the tip amount relative to the total fare, specifically for credit card transactions. This metric is vital for analysing tipping behavior.

5.1.3. Outlier Handling

Outlier handling was performed using domain-based ranges to remove implausible trip records. This approach leverages real-world knowledge of taxi operations to define reasonable boundaries for key variables. The following ranges were applied:

- **Distance**: Trips with distances less than 0.1 miles or greater than 30 miles were considered outliers. This filters out extremely short or long trips that might be erroneous or represent non-standard travel.
- **Duration**: Trips shorter than 2 minutes or longer than 120 minutes were excluded. This helps in removing very short, likely erroneous trips, and excessively long trips that might indicate unusual circumstances or data errors.
- **Fare**: Fares outside the range of $2 to $200 were removed. This range accounts for minimum fares and excludes extremely high fares that could be data entry errors or special, non-standard trips.
- **Total Amount**: Similar to fare, total amounts outside the $2 to $300 range were excluded.
- **Tip Percentage**: Tips were restricted to a range of 0% to 50%. This is particularly important given the observation of extremely high tip percentages (e.g., >1000%) in the raw data, which are likely due to data errors or unusual circumstances. The 0-50% range aligns with typical tipping practices.

- **Speed**: Calculated speed (distance/duration) was constrained to between 2 mph and 60 mph. This filters out trips with unrealistically low speeds (e.g., stationary vehicles) or excessively high speeds (e.g., data errors).



```
Rule violations BEFORE cleaning:
```

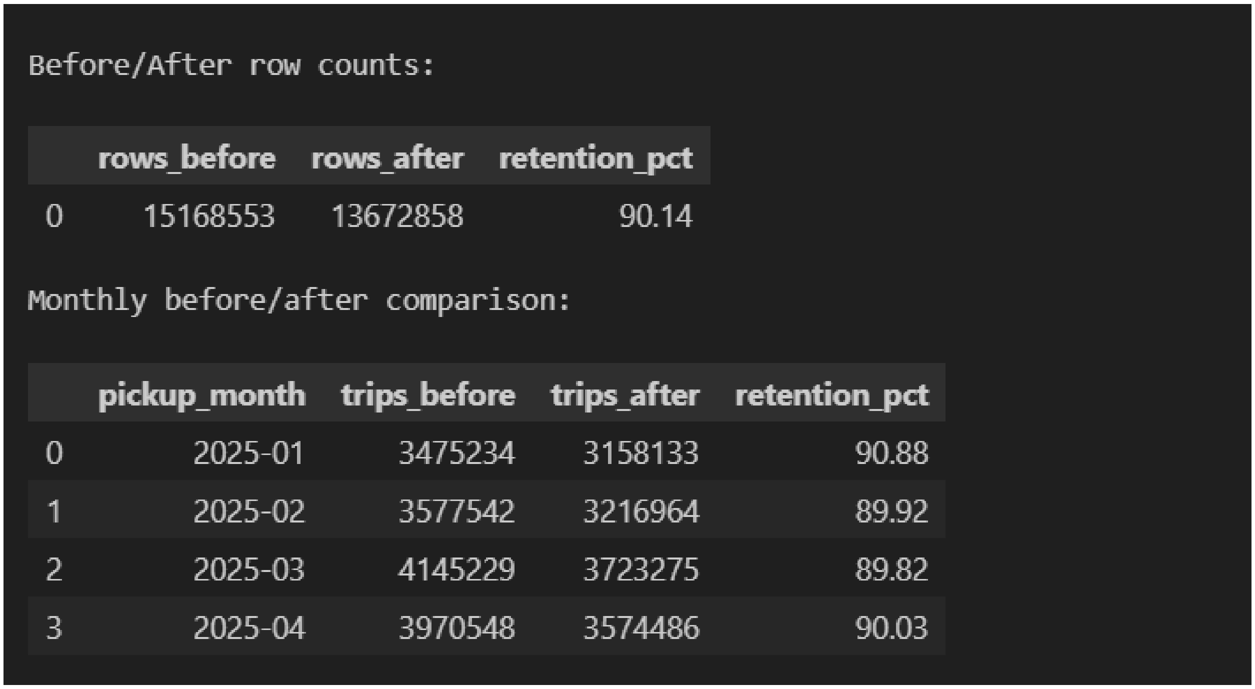
	rule	rows
0	fare_amount < 2.0	731669
1	trip_distance < 0.1	486455
2	avg_speed_mph < 2.0	485136
3	trip_duration_min < 2.0	334294
4	total_amount < 2.0	309418
5	tip_pct outside [0, 0.5]	181427
6	trip_distance > 30.0	11820
7	trip_duration_min > 120.0	9436
8	avg_speed_mph > 60.0	4784
9	fare_amount > 200.0	4244
10	total_amount > 300.0	2231

```
Total rows before cleaning: 15,168,553
```

Figure 14 Rule Violations before cleaning the dataset

5.1.4. Data Quality Assessment

After applying these cleaning steps, a data quality assessment was performed by comparing the row counts before and after cleaning. The dataset was reduced from approximately 15.1 million records to 13.6 million records, representing a retention rate of approximately 90%. This high retention rate, coupled with the consistency of retention across all months, indicates that the cleaning process effectively removed anomalies without significantly reducing the dataset's size or representativeness. The removal of approximately 10% of the data was justified as these records represented implausible trips or anomalies that would have skewed the analysis.



Before/After row counts:

	rows_before	rows_after	retention_pct
0	15168553	13672858	90.14

Monthly before/after comparison:

	pickup_month	trips_before	trips_after	retention_pct
0	2025-01	3475234	3158133	90.88
1	2025-02	3577542	3216964	89.92
2	2025-03	4145229	3723275	89.82
3	2025-04	3970548	3574486	90.03

Figure 15 Comparison of Before vs After Data frame

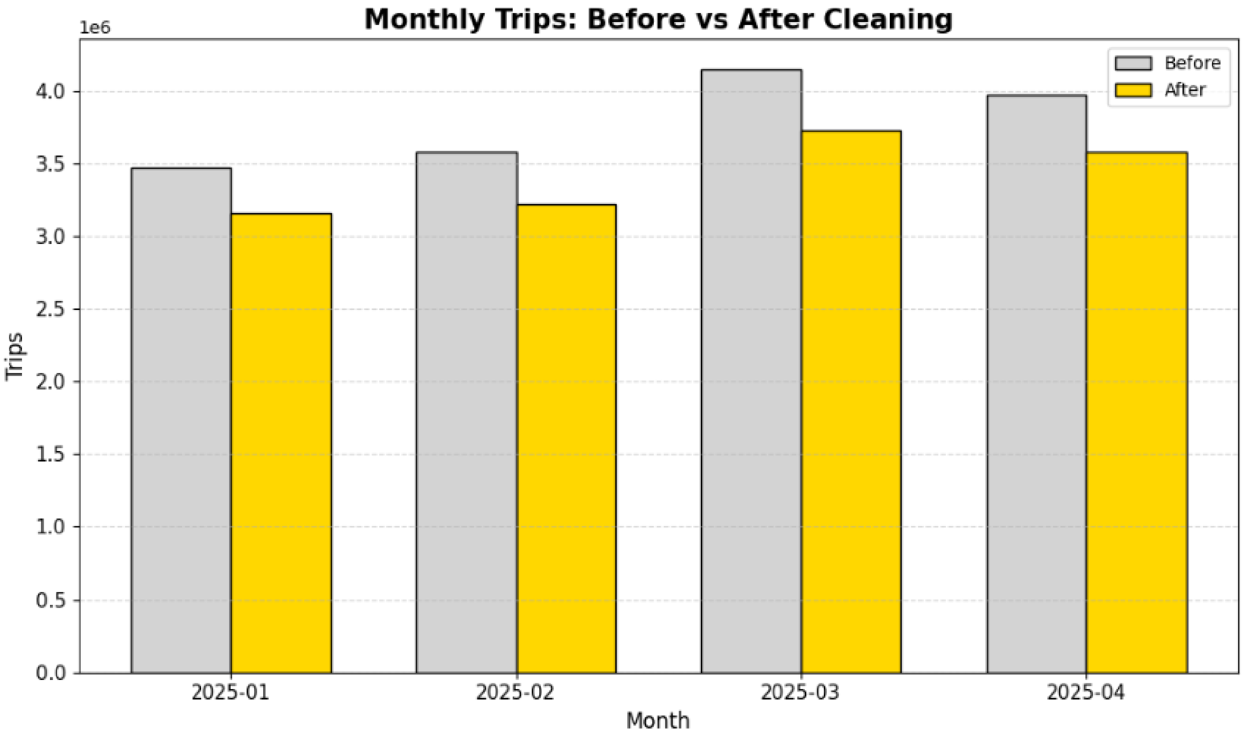


Figure 16 Monthly Trips Before vs After

5.2. Predictive Modelling

The objective of predictive modeling was to accurately estimate taxi fares based on various trip features. Two types of models were implemented and compared: Linear Regression as a baseline, and XGBoost for improved accuracy.

5.2.1. Problem Definition and Feature Selection

The problem was defined as predicting the fare_amount using a set of relevant trip features. Based on domain knowledge and insights from EDA, the following features were selected for modeling:

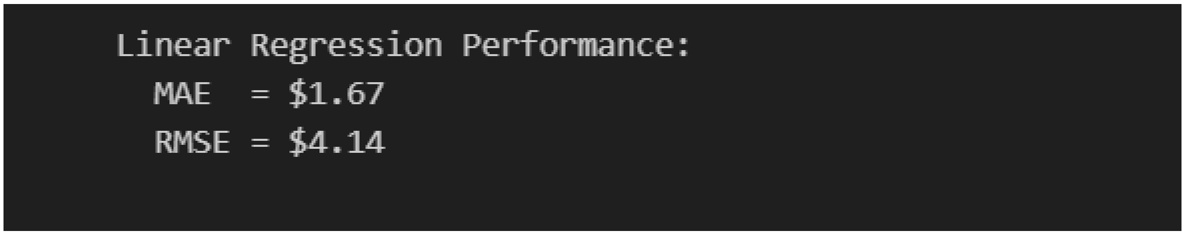
- trip_distance
- trip_duration_min
- pickup_hour
- pickup_month
- passenger_count

5.2.2. Linear Regression Model

A Linear Regression model was trained as a simple, interpretable baseline. This model assumes a linear relationship between the features and the target variable (fare amount). The results obtained were:

- **Mean Absolute Error (MAE)**: Approximately $1.67
- **Root Mean Squared Error (RMSE)**: Approximately $4.14

The coefficients of the Linear Regression model confirmed that trip_distance and trip_duration_min were the primary drivers of fare amount, which aligns with the expected fare structure of taxi services.



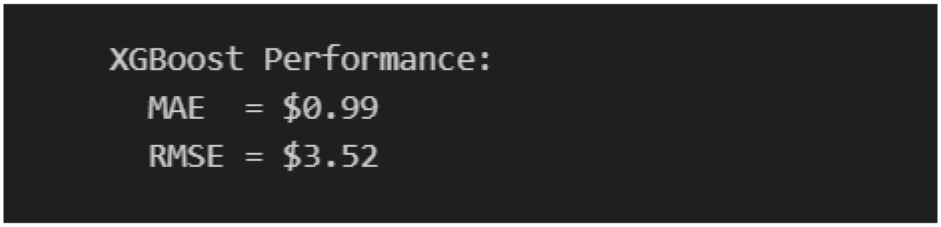
```
Linear Regression Performance:
  MAE  = $1.67
  RMSE = $4.14
```

Figure 17 Linear Regression Performance

5.2.3. XGBoost Model

An XGBoost (eXtreme Gradient Boosting) model was then implemented to achieve higher predictive accuracy. XGBoost is an ensemble learning method that uses a gradient boosting framework, known for its efficiency and performance in various machine learning tasks. The results for the XGBoost model were:

- **Mean Absolute Error (MAE)**: Approximately $0.99
- **Root Mean Squared Error (RMSE)**: Approximately $3.52



```
XGBoost Performance:
  MAE  = $0.99
  RMSE = $3.52
```

Figure 18 XGBoost Performance

This represents a significant improvement of approximately 40% in MAE compared to the Linear Regression model, demonstrating the superior predictive power of XGBoost for this problem.

5.2.4. Feature Importance Analysis

An analysis of feature importance for the XGBoost model provided insights into the relative contribution of each feature to fare prediction:

- trip_distance: Identified as the strongest driver of fare, which is intuitively correct as fares are primarily distance-based.
- trip_duration_min: The secondary most important feature, indicating that trip duration also significantly influences the fare, especially in congested conditions.
- pickup_hour: Showed a minor effect on fare, suggesting that time of day has some, but not a dominant, influence.
- passenger_count and pickup_month: These features were found to have negligible impact on fare prediction, implying they are not significant determinants of fare in this dataset.

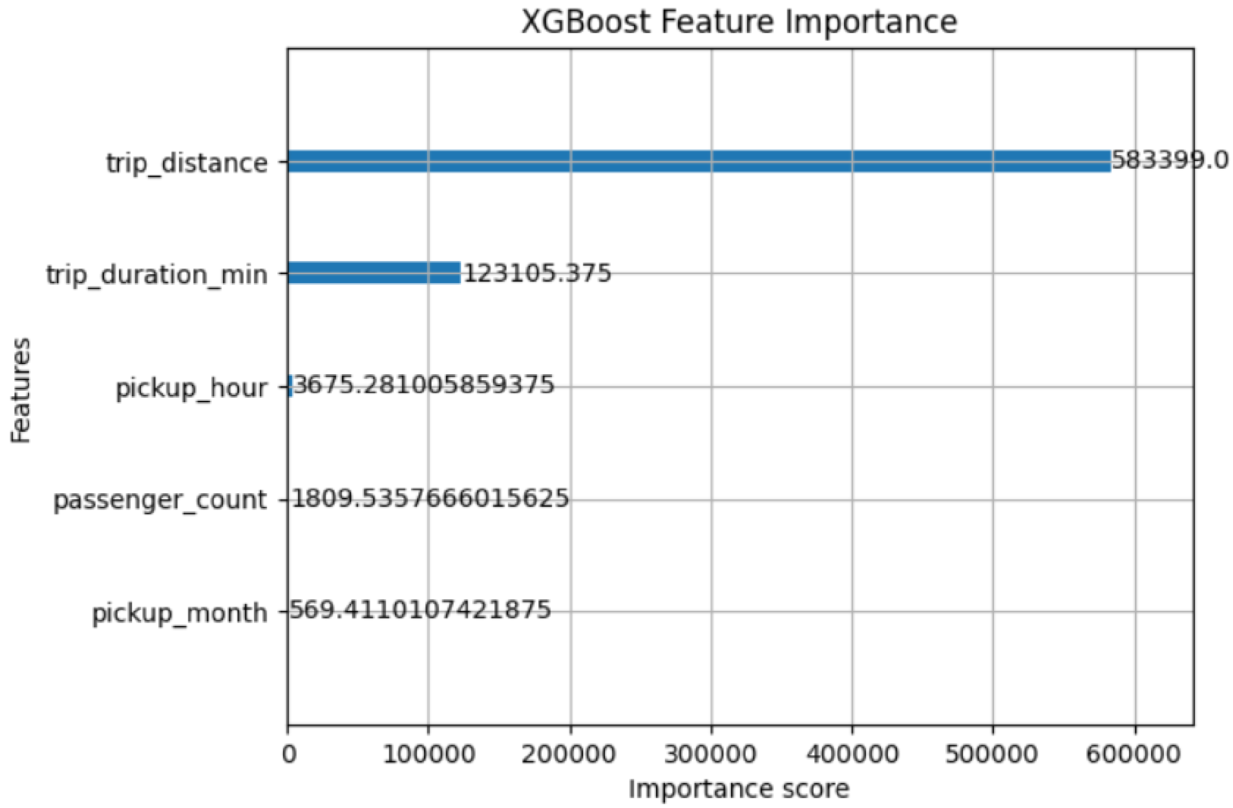


Figure 19 XGBoost Feature Importance

5.2.5. Model Comparison

Comparing the two models, Linear Regression served as an interpretable baseline, providing clear insights into the linear relationships between features and fare. However, the XGBoost model demonstrated significantly higher accuracy, making it more suitable for real-world fare prediction where precision is paramount. The 40% improvement in MAE highlights the effectiveness of more advanced machine learning techniques for complex predictive tasks.

6. Results and Interpretation

The results of this comprehensive analysis provide significant insights into NYC taxi trip dynamics, covering data quality, temporal patterns, payment behaviors, and the effectiveness of predictive models for fare estimation.

6.1. Data Quality and Cleaning Impact

One of the most crucial results was the successful cleaning and preparation of the raw NYC taxi data. Out of an initial 15.1 million records, approximately 13.6 million records were retained after rigorous filtering and outlier handling. This 90% retention rate, consistent across all months (January to April 2025), demonstrates the robustness of the cleaning methodology. The removal of approximately 10% of the data was justified by the identification of implausible trips and anomalies (e.g., trips with invalid dates, extreme distances, durations, or fares). This cleaning process ensures that all subsequent analyses and model training are based on high-quality, reliable data, which is fundamental for drawing accurate conclusions in real-world big data scenarios.

6.3. Predictive Modelling Performance

The predictive modeling phase demonstrated the feasibility of accurately estimating taxi fares using trip features, with a clear advantage for more advanced machine learning techniques.

6.3.1. Model Comparison

- **Linear Regression**: Served as an interpretable baseline, yielding a Mean Absolute Error (MAE) of approximately $1.67 and a Root Mean Squared Error (RMSE) of approximately $4.14. This model confirmed that trip_distance and trip_duration_min are the primary linear drivers of fare.
- **XGBoost**: Significantly outperformed the linear model, achieving an MAE of approximately $0.99 and an RMSE of approximately $3.52. This represents an approximate 40% improvement in MAE, highlighting the superior predictive power of XGBoost in capturing complex, non-linear relationships within the data.

6.3.2. Feature Importance

Feature importance analysis from the XGBoost model provided clear insights into the main determinants of taxi fares:

- **Distance (Strongest Driver)**: trip_distance was unequivocally identified as the most influential feature, which aligns perfectly with the standard fare calculation mechanisms in the taxi industry.
- **Duration (Secondary Driver)**: trip_duration_min was the second most important feature, underscoring that time spent in traffic or during the trip also significantly contributes to the final fare, especially in a city prone to congestion.
- **Minor/Negligible Factors**: pickup_hour had a minor effect, suggesting some hourly fare variations (e.g., surge pricing or time-of-day adjustments), while passenger_count and pickup_month were found to have negligible impact on fare prediction. This indicates that these factors, while potentially relevant for other analyses (like demand forecasting), do not significantly drive the fare amount itself.

6.4. Value and Impact of Results

The results of this analysis have several practical implications:

- **For Drivers**: The insights into demand patterns, profitable hours, and the main drivers of fare can help drivers optimize their working hours and routes to maximize earnings.

- **For Policymakers/Urban Planners**: Understanding congestion patterns and their impact on revenue can inform decisions regarding traffic management, infrastructure investments, and public transportation integration.
- **For Passengers**: The high accuracy of the XGBoost model suggests that reliable fare estimation tools can be developed, leading to greater transparency and trust in taxi services.
- **For Data Science**: The project demonstrates the effectiveness of a structured data analysis pipeline, from meticulous data cleaning to advanced predictive modeling, emphasizing the critical role of domain knowledge in achieving meaningful results from real-world datasets.

In summary, the analysis successfully transformed raw, complex taxi data into actionable intelligence, providing a clear understanding of the factors influencing NYC taxi trips and enabling accurate fare prediction. The findings confirm the importance of data quality, the power of EDA in uncovering hidden patterns, and the superior performance of advanced machine learning models for real-world predictive tasks.

References

Donovan, B., & Work, D. B. (2017). *New York City taxi trip data cleaning and analysis* (Technical Report). University of Illinois at Urbana-Champaign.

New York City Taxi and Limousine Commission. (2025). *Data dictionary: Yellow taxi trip records (March 2025)*. NYC TLC. https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

New York City Taxi and Limousine Commission. (n.d.). *Taxi fare information*. NYC TLC. Retrieved September 4, 2025, from https://www.nyc.gov/site/tlc/passengers/taxi-fare.page

Schaller, B. (2010). *New York City traffic congestion: The impact of taxis and for-hire vehicles*. Schaller Consulting.

Yadav, R., & Saha, S. (2019). *An analysis of NYC taxi trip data*. Proceedings of the International Conference on Data Science and Engineering (pp. 178–183). IEEE. https://doi.org/10.1109/ICDSE.2019.00038

Castillo, M., & Petrie, R. (2010). Discrimination in the marketplace: Evidence from the New York City taxi industry. *American Economic Review, 100*(5), 2361–2388. https://doi.org/10.1257/aer.100.5.2361