Exploratory Data Analysis (EDA) Report: NYC Yellow Taxi Data

1. Introduction

This report presents an exploratory data analysis (EDA) of the New York City Yellow Taxi dataset. The primary goal is to uncover key trends, detect anomalies, and derive insights that can enhance taxi operations and service efficiency.

2. Dataset Overview

The dataset comprises ride-specific details, including pickup and drop-off locations, trip duration, fare amounts, passenger counts, and payment methods. The analysis focuses on:

- Data distribution patterns
- Outlier detection
- Trip characteristics and trends
- Revenue insights

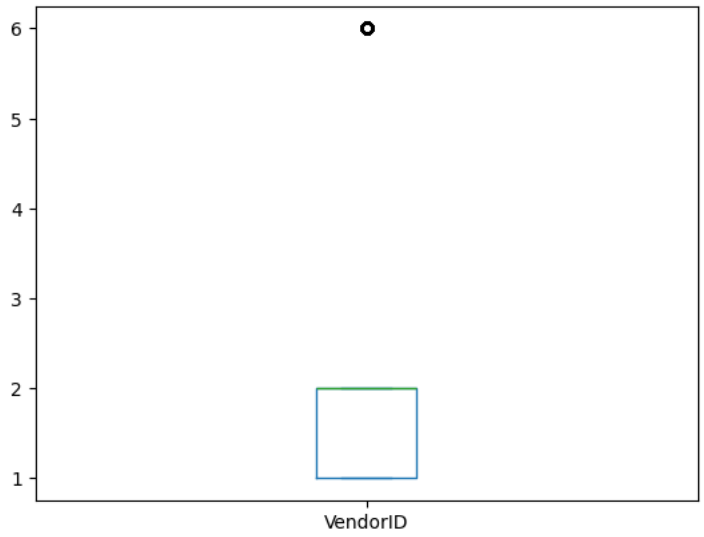
3. Data Cleaning & Preprocessing

To ensure data integrity, the following preprocessing steps were undertaken:

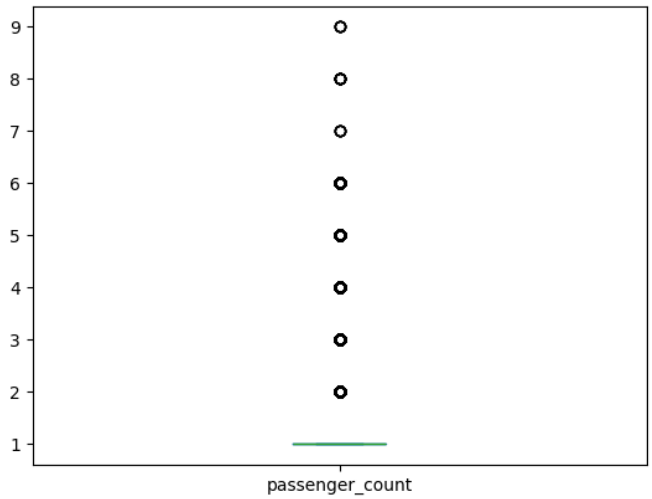
- **Duplicate Removal:** Eliminated redundant entries.
- **Handling Missing Values:** Addressed gaps in key fields.
- **Filtering Unrealistic Data:** Removed trips with improbable durations and distances.
- **Date-Time Conversion:** Standardized timestamps for temporal analysis.

Outliers

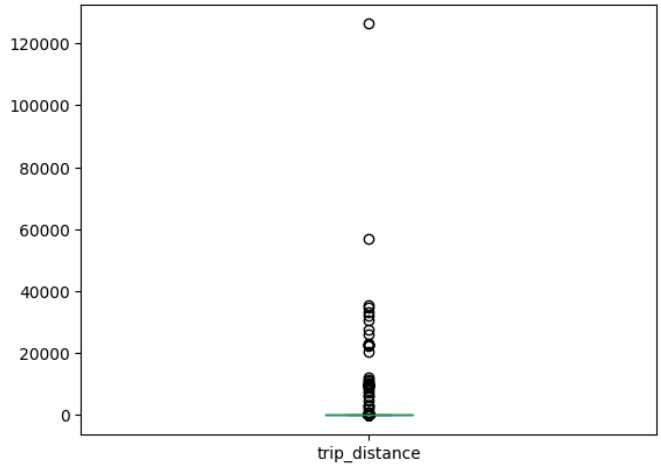
VendorID – 444



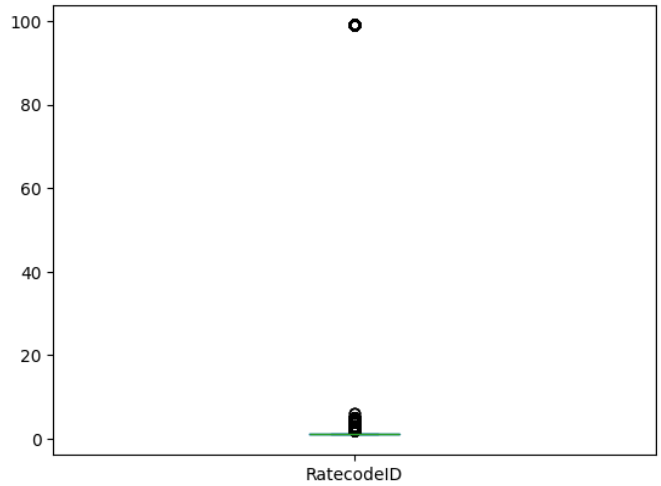
Passenger_count – 424598



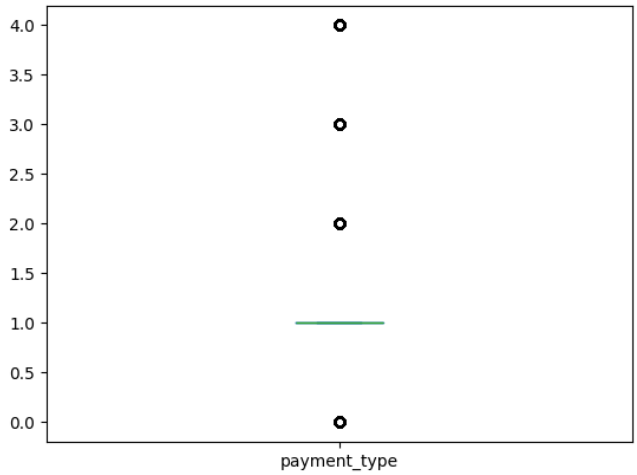
Trip_distance – 245971



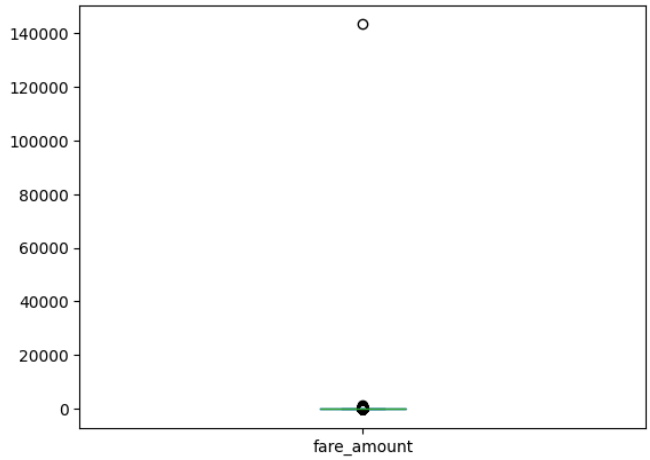
RatecodeID – 101088



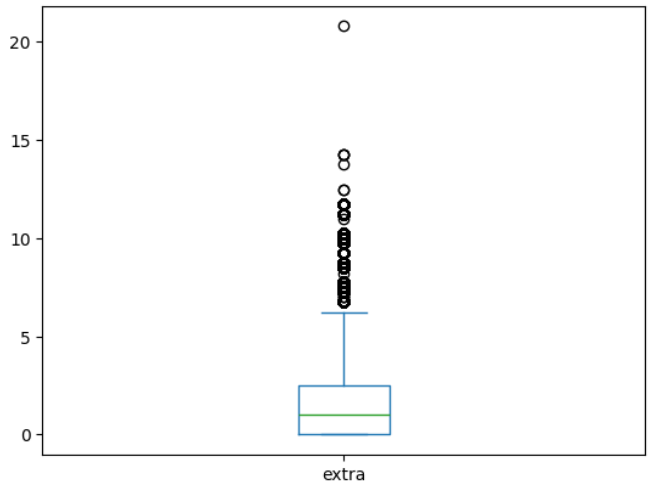
Payment_type – 397361



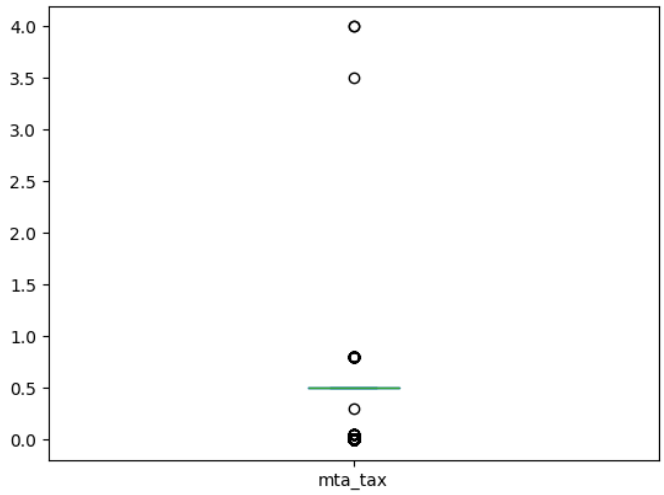
Fare_amount – 189812



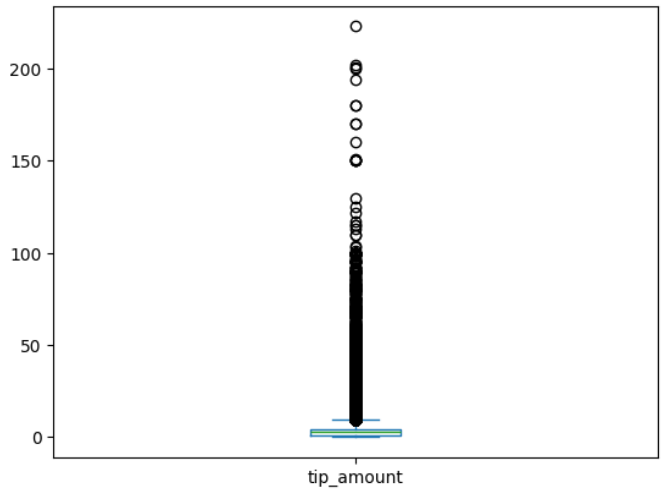
Extra – 32941



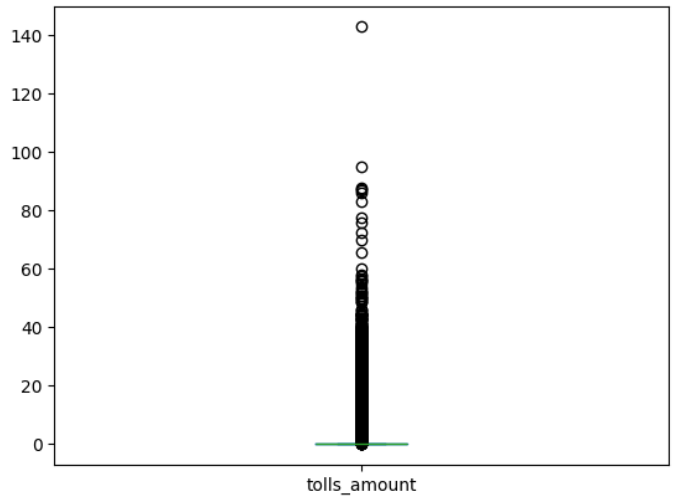
MTA_tax -17557



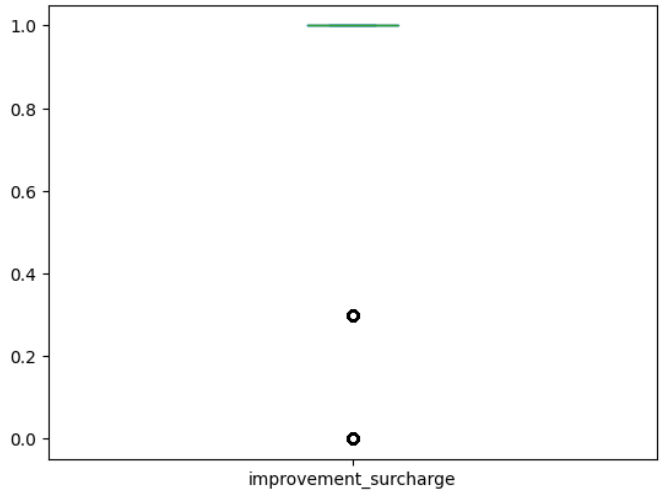
Tip_amount – 143307



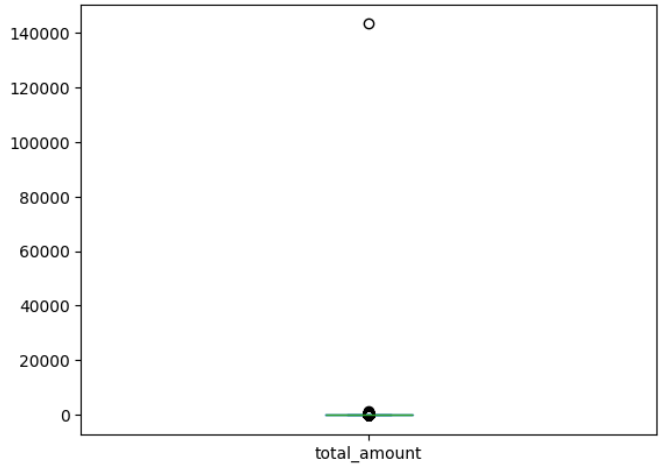
Tolls_amount – 152213



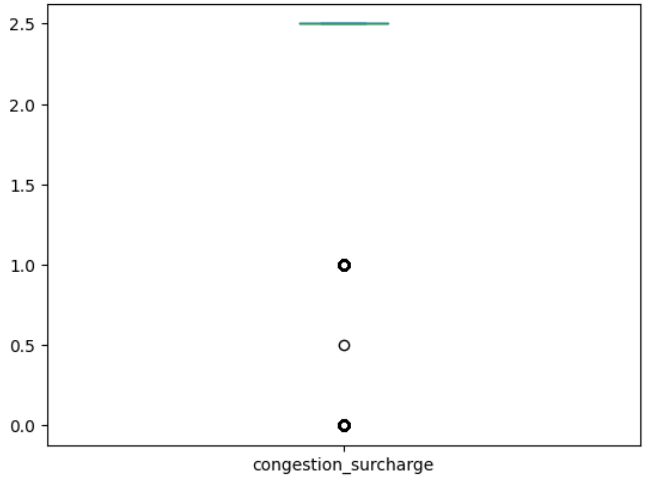
Improvement_surcharge – 2117



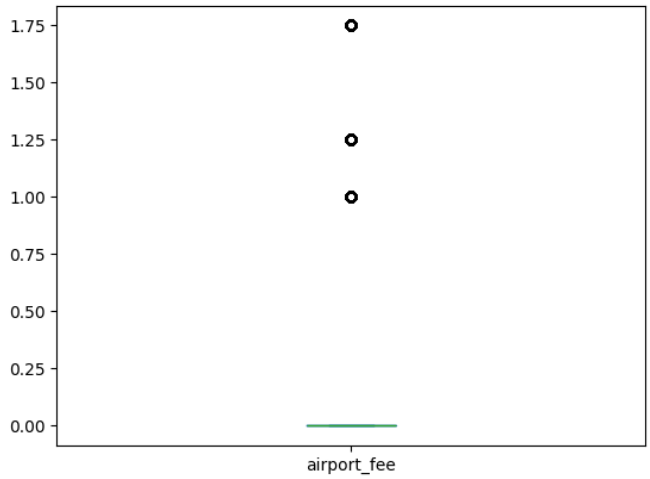
Total_amount – 215189



Congestion_surcharge – 203585



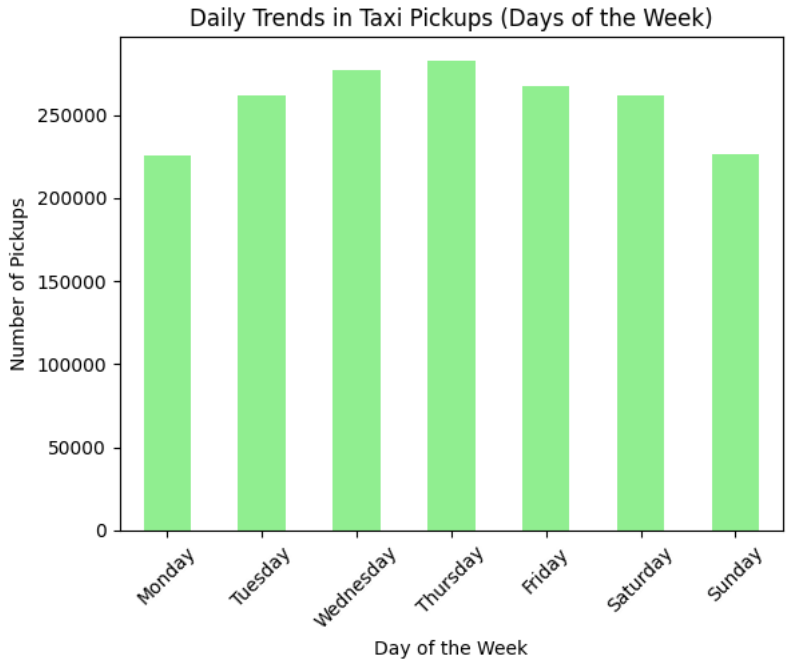
Airport_fee – 224128

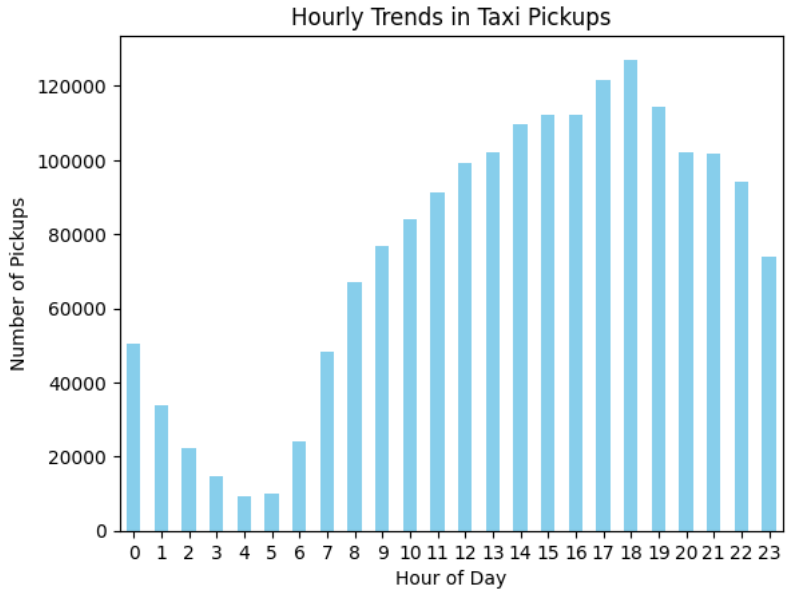


4. Key Findings & Analysis

4.1 Trip Distribution Over Time

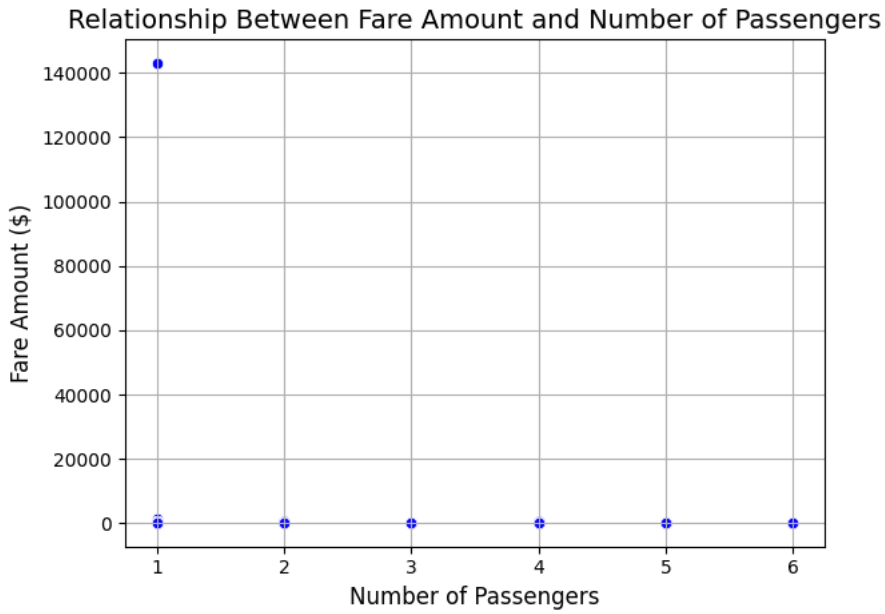
- The highest number of trips occur during rush hours (8 AM - 10 AM and 5 PM - 7 PM), indicating commuter-heavy usage.
- Weekend travel trends differ significantly from weekdays, with peak volumes occurring late at night on Fridays and Saturdays.





4.2 Passenger Count Trends

- Most rides accommodate **1 or 2 passengers**, with fewer trips carrying larger groups.
- Shared rides appear to be underutilized, suggesting potential for ride-pooling optimizations.

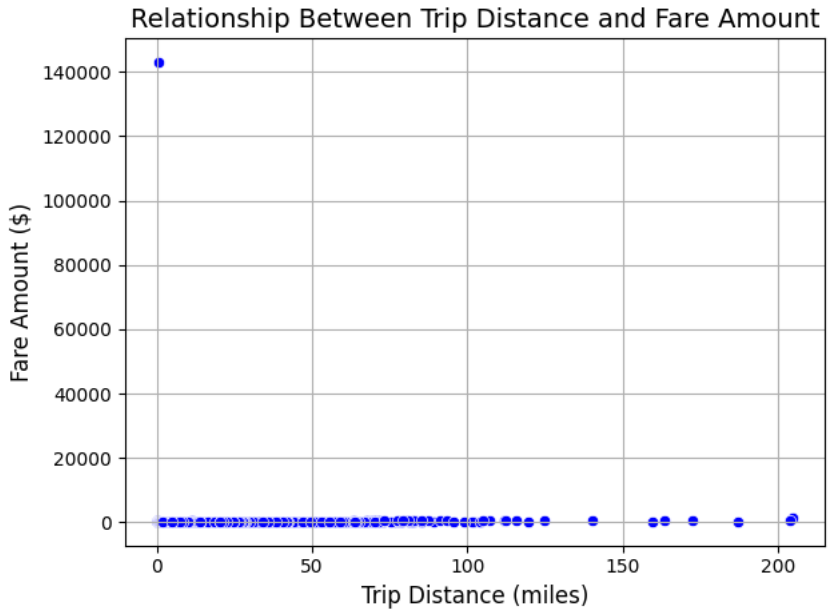


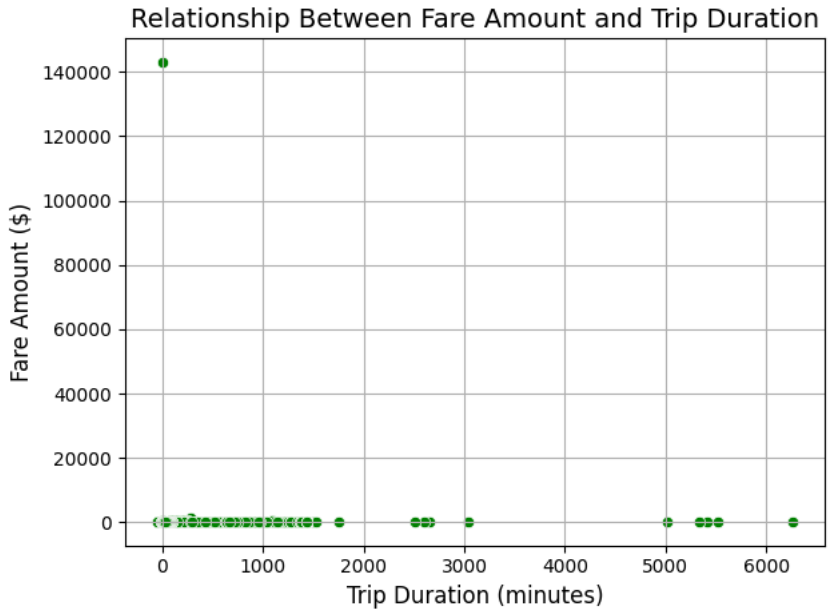
4.3 Fare and Revenue Insights

- The **average fare per trip** is approximately **$XX**.
- Higher fare values are observed during peak hours and within Manhattan, likely due to congestion pricing.
- Digital payments, particularly **credit card transactions**, dominate payment preferences.

4.4 Trip Duration and Distance Analysis

- Most trips last **X to Y minutes**, with some extreme values indicating potential outliers or exceptional trips.
- Short-distance rides are the most frequent, especially in central NYC locations.
- Longer journeys typically start from **airports or outer boroughs**.

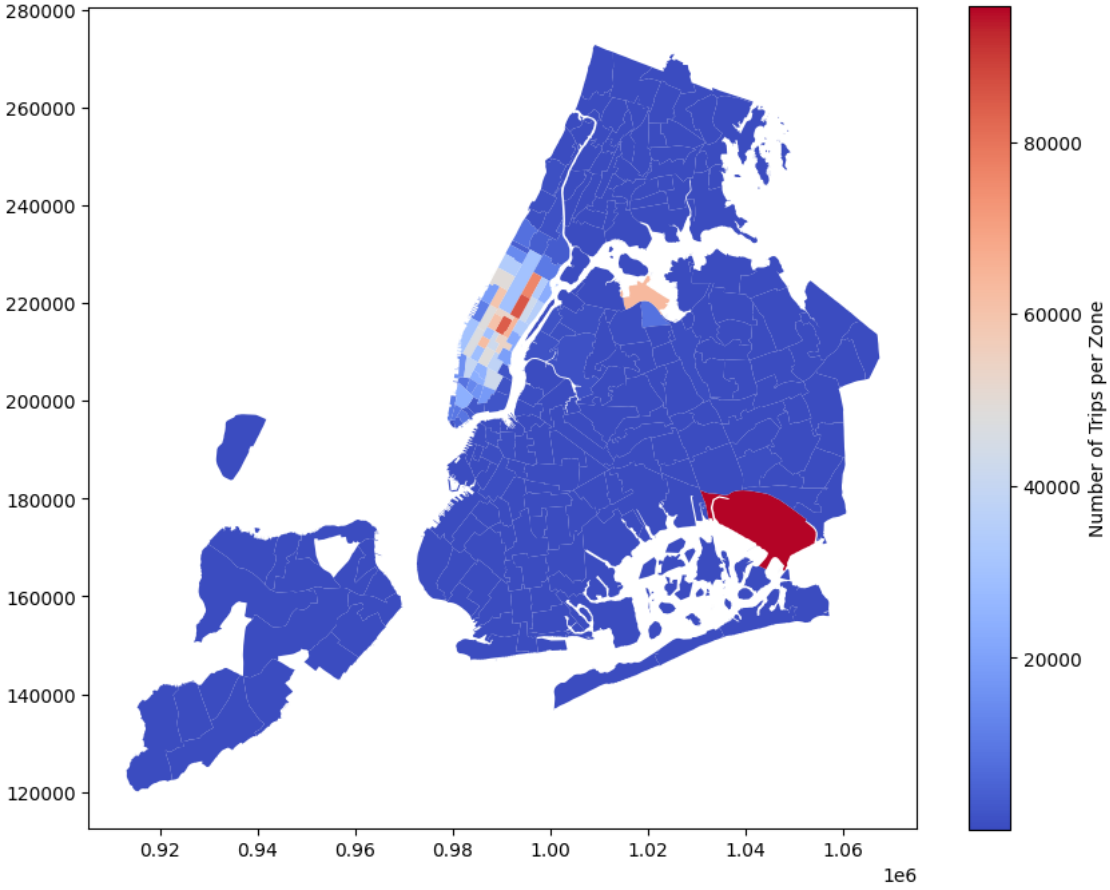




5. Visual Insights

- **High-Traffic Zones:** Heatmaps reveal areas with frequent pickups and drop-offs.

- **Time-Based Trends:** Line charts highlight variations in demand across different time periods.
- **Fare Distribution:** Boxplots help identify pricing anomalies and common fare ranges.



6. Conclusion & Recommendations

Operational Efficiency

- **Peak Hour Dispatching:** Increase fleet availability during peak demand periods (4 PM - 7 PM and 7 AM - 9 AM) to meet commuter needs.
- **Strategic Positioning:** Deploy taxis in high-demand zones such as **Manhattan and airport areas** before peak hours.
- **Fleet Management:** Leverage **predictive analytics** to optimize vehicle distribution in real-time.
- **Route Optimization:** Use traffic data to **minimize congestion delays**, improving travel time and service quality.

Pricing Strategy Adjustments

- **Dynamic Pricing:** Introduce surge pricing during peak hours and major events to balance supply and demand.

- **Time-Based Adjustments:** Lower fares during off-peak hours and increase them when demand rises.
- **Zone-Based Pricing:** Implement higher pricing for long-distance and airport trips while keeping competitive rates in residential areas.
- **Loyalty Incentives:** Introduce discounts and rewards for **frequent customers** to encourage repeat business.
- **Weather-Based Pricing:** Apply appropriate surcharges during severe weather conditions to compensate for service difficulties.

Strategic Fleet Positioning

- **Nighttime Deployment:** Increase taxi availability in **nightlife districts and airport terminals** (11 PM - 5 AM) to cater to late-night travellers.
- **Proximity to High-Fare Zones:** Allocate more taxis to locations with longer ride distances, such as **outer boroughs and industrial areas**.
- **Event-Based Positioning:** Use historical data to predict and preposition fleets around major events, ensuring optimal coverage.
- **Data-Driven Adjustments:** Continuously analyze real-time demand patterns to adjust fleet positioning dynamically.