Assignment Report - Exploratory Data Analysis

**NYC Yellow Taxi Trip Data Analysis  
Optimizing Operations Through Data-Driven Insights**

**Name: Ashwin Sai**

Table of Contents

[Exploratory Data Analysis: NYC Yellow Taxi Operations (2023) 1](#_Toc209739263)

[Report Details 1](#_Toc209739264)

[1. Introduction and Objective 4](#_Toc209739265)

[2. Data Understanding 4](#_Toc209739266)

[2.1. Data Source and Description 4](#_Toc209739267)

[2.2. Data Dictionary Overview 5](#_Toc209739268)

[3. Data Preparation and Sampling 6](#_Toc209739269)

[3.1. Initial Data Loading and Assessment 6](#_Toc209739270)

[3.2. Sampling Strategy and Rationale 6](#_Toc209739271)

[3.3. Implementation of Sampling 6](#_Toc209739272)

[3.4. Final Sampled Dataset 7](#_Toc209739273)

[4. Data Cleaning and Preprocessing 7](#_Toc209739274)

[4.1. Initial Data Inspection 7](#_Toc209739275)

[4.2. Structural Corrections 7](#_Toc209739276)

[4.2.1. Index Reset 7](#_Toc209739277)

[4.2.2. Handling Monetary Negatives 7](#_Toc209739278)

[4.3. Missing Value Imputation 8](#_Toc209739279)

[4.3.1. Proportion of Missing Values 8](#_Toc209739280)

[4.3.2. `passenger\_count` Imputation 8](#_Toc209739281)

[4.3.3. `RatecodeID` Imputation 8](#_Toc209739282)

[4.3.4. `congestion\_surcharge` Imputation 8](#_Toc209739283)

[4.3.5. Imputation of Remaining Nulls 9](#_Toc209739284)

[4.4. Outlier Analysis and Treatment 9](#_Toc209739285)

[4.4.1. Descriptive Statistics for Outlier Detection 9](#_Toc209739286)

[4.4.2. Filtering Invalid Trip Records 9](#_Toc209739287)

[4.5. Feature Engineering 10](#_Toc209739288)

[5. Exploratory Data Analysis 10](#_Toc209739289)

[5.1. Variable Classification 10](#_Toc209739290)

[5.2. Temporal Analysis 10](#_Toc209739291)

[5.2.1. Hourly, Daily, and Monthly Pickup Distributions 10](#_Toc209739292)

[5.3. Financial Analysis 12](#_Toc209739293)

[5.3.1. Analysis of Zero-Value Records 12](#_Toc209739294)

[5.3.2. Monthly Revenue Trend 13](#_Toc209739295)

[5.3.3. Quarterly Revenue Distribution 13](#_Toc209739296)

[5.3.4. Correlation Analysis: Distance, Duration, and Fare 14](#_Toc209739297)

[5.3.5. Payment Type Distribution 16](#_Toc209739298)

[5.4. Geospatial Analysis 16](#_Toc209739299)

[5.4.1. Loading and Merging Taxi Zone Data 16](#_Toc209739300)

[5.4.2. Mapping Pickup Density by Zone 16](#_Toc209739301)

[5.5. Detailed Operational Analysis 17](#_Toc209739302)

[5.5.1. Route Speed and Efficiency Analysis 17](#_Toc209739303)

[5.5.2. Peak Hour Trip Volume Analysis 17](#_Toc209739304)

[5.5.3. Weekday vs. Weekend Traffic Patterns 17](#_Toc209739305)

[5.5.4. Analysis of High-Demand Zones 18](#_Toc209739306)

[5.5.5. Pickup-to-Dropoff Ratio Analysis 18](#_Toc209739307)

[5.5.6. Nighttime Operations Analysis 18](#_Toc209739308)

[5.5.7. Fare and Pricing Structure Analysis 18](#_Toc209739309)

[5.5.8. Customer Experience and Tipping Behavior 20](#_Toc209739310)

[5.5.9. Passenger Count Dynamics 20](#_Toc209739311)

[5.5.10. Surcharge Application Analysis 21](#_Toc209739312)

[6. Conclusion: Summary of Analytical Findings 22](#_Toc209739313)

## 1. Introduction and Objective

This report details an exploratory data analysis (EDA) conducted on the 2023 New York City (NYC) Yellow Taxi trip records, provided by the NYC Taxi and Limousine Commission (TLC). The primary objective of this analysis is to use the trip data to uncover significant patterns, trends, and insights that characterize taxi operations within the city.

The scope of this document is strictly analytical and descriptive. It focuses on a systematic examination of the dataset to understand temporal demand cycles, the geographic distribution of taxi activity, the structure of fare composition, and key factors influencing operational metrics. This is achieved through a multi-stage process involving data preparation, cleaning, feature engineering, and visualization. The analysis translates raw trip records into a structured understanding of service efficiency, revenue drivers, and passenger behavior.

Unlike a business strategy document, this report refrains from making prescriptive recommendations. Instead, it serves as a foundational analytical summary, presenting the inherent patterns and relationships discovered within the data. The guiding questions for this exploration are:

* What are the temporal rhythms of taxi demand on hourly, weekly, and monthly scales?
* Which geographic zones are central to taxi activity, and where do imbalances between pickups and dropoffs occur?
* What are the primary drivers of fare amounts, and how do they relate to trip distance and duration?
* What behavioral patterns can be inferred from data on payment types and tipping?

The subsequent sections document the methodology and findings of this exploratory analysis.

## 2. Data Understanding

### 2.1. Data Source and Description

The data for this analysis comprises the NYC Yellow Taxi trip records for the year 2023. The dataset was collected by technology providers authorized by the TLC and is made publicly available. It is organized into twelve separate Parquet files, with each file corresponding to one month of the year.

These records contain a comprehensive set of fields for each trip, including:

* **Temporal Information:** Pickup and drop-off timestamps.
* **Geospatial Information:** Pickup and drop-off locations, identified by a standardized Location ID.
* **Trip Metrics:** Trip distance in miles and passenger counts.
* **Financial Details:** A full breakdown of the fare, including the base amount, tips, tolls, surcharges, and the total amount.
* **Transactional Metadata:** Vendor ID, payment type, and rate code.

The taxi zone Location IDs (ranging from 1 to 263) correspond to specific, defined geographic zones across NYC's boroughs. A separate taxi zone shapefile is used to map these IDs to their physical locations for geospatial analysis.

### 2.2. Data Dictionary Overview

A summary of the key variables from the official TLC data dictionary relevant to this analysis is provided below:

|  |  |
| --- | --- |
| Field Name | Description |
| VendorID | A code indicating the TPEP provider (1=Creative Mobile Tech, 2=VeriFone Inc.). |
| tpep\_pickup\_datetime | The date and time when the meter was engaged. |
| tpep\_dropoff\_datetime | The date and time when the meter was disengaged. |
| passenger\_count | The number of passengers in the vehicle (driver-entered value). |
| trip\_distance | The elapsed trip distance in miles reported by the taximeter. |
| PULocationID | The TLC Taxi Zone ID where the meter was engaged (pickup). |
| DOLocationID | The TLC Taxi Zone ID where the meter was disengaged (dropoff). |
| RatecodeID | The final rate code for the trip (e.g., 1=Standard, 2=JFK, 3=Newark). |
| payment\_type | A numeric code for the payment method (e.g., 1=Credit card, 2=Cash). |
| fare\_amount | The time-and-distance fare calculated by the meter. |
| tip\_amount | Tip amount (automatically populated for credit cards). |
| tolls\_amount | Total amount of all tolls paid during the trip. |
| total\_amount | The total amount charged to passengers, excluding cash tips. |
| congestion\_surcharge | Surcharge collected for trips in designated congestion zones. |
| airport\_fee | Fee for pickups at LaGuardia and JFK Airports. |

This dictionary provides the basis for interpreting the fields during the data cleaning and analysis phases.

## 3. Data Preparation and Sampling

### 3.1. Initial Data Loading and Assessment

The analysis began by loading a single monthly data file (January 2023) to assess its structure and size. The initial file for January 2023 contained over 3 million rows. Aggregating all twelve monthly files would result in a dataset with approximately 36 to 40 million rows. Handling a dataset of this magnitude for exploratory analysis is computationally intensive and can lead to memory overload on standard hardware. Therefore, a sampling strategy was necessary to create a representative, yet manageable, subset of the data for analysis.

### 3.2. Sampling Strategy and Rationale

To ensure the sampled data accurately represents the overall trends present in the full dataset, a stratified sampling approach was selected. The goal was to preserve the temporal distribution of trips across all days and hours of the year. A simple random sample might underrepresent patterns during low-traffic periods (e.g., early mornings) or specific days.

The chosen strategy involves iterating through each day of each month and, for each hour of that day, randomly sampling a fixed fraction of the trips. This ensures that every time-stratum (a specific hour of a specific day) contributes proportionally to the final dataset.

Rationale for Sampling by Date and Hour:

* Sampling by both date and hour, rather than just by hour, is critical for preserving the unique traffic patterns associated with specific days of the week, holidays, or seasons. For example, traffic at 8:00 AM on a Monday is fundamentally different from traffic at 8:00 AM on a Sunday. By stratifying by both date and hour, the sample retains these crucial weekly and seasonal variations, providing a more accurate microcosm of the entire year's activity.
* A 1% sampling fraction (`frac=0.01`) was chosen to balance representativeness with computational feasibility, aiming for a final dataset of approximately 300,000 to 400,000 entries.

### 3.3. Implementation of Sampling

The sampling process was implemented programmatically. The script iterates through each of the twelve monthly Parquet files. For each month, it performs the following steps:

* Loads the monthly data into a pandas DataFrame.
* Engineers two temporary columns, pickup\_date and pickup\_hour, from the tpep\_pickup\_datetime column to facilitate stratification.
* Iterates through each unique date within the month.
* Within each date, it iterates through each of the 24 hours.
* For each hour containing trip data, it randomly samples 1% of the records using the pandas `.sample(frac=0.01, random\_state=42)` method. A `random\_state` is used for reproducibility.
* The hourly samples are concatenated to build a sampled dataset for the current month.
* Finally, the sampled data from all twelve months are concatenated into a single, final DataFrame for analysis.

This method systematically extracts a representative sample from the entire year's data.

### 3.4. Final Sampled Dataset

After executing the sampling procedure across all twelve monthly files, the resulting dataset contained 379,268 records. This size is well within the target range for efficient analysis while being large enough to capture meaningful statistical patterns.

The final sampled DataFrame was saved to a new Parquet file, `nyc\_taxi\_2023\_sampled.parquet`, to allow for direct loading in subsequent analytical steps, bypassing the need to repeat the computationally expensive sampling process.

The shape of the original January file (3,041,714 rows, 19 columns) compared to the final sampled DataFrame for the entire year (379,268 rows, 19 columns after dropping temporary columns) demonstrates the significant reduction in data volume achieved through this process.

## 4. Data Cleaning and Preprocessing

### 4.1. Initial Data Inspection

The sampled dataset was loaded, and its structure and data types were inspected. The inspection revealed 19 columns and 379,268 entries. The data types were generally appropriate, with timestamps correctly parsed as `datetime64` and most numerical fields as `float64` or `int64`. Several columns were identified as having null values (`passenger\_count`, `RatecodeID`, `store\_and\_fwd\_flag`, `congestion\_surcharge`, and `airport\_fee`), requiring handling in the cleaning phase.

### 4.2. Structural Corrections

#### 4.2.1. Index Reset

The DataFrame's index was reset to ensure a clean, continuous index from 0 to N-1. This is a standard preprocessing step after concatenation and sampling to prevent issues with index-based operations. No columns were identified as unnecessary for the analysis at this stage.

#### 4.2.2. Handling Monetary Negatives

An investigation into monetary columns revealed the presence of negative values, which are logically invalid for fare components.

|  |  |
| --- | --- |
| Column | Count of Negative Values |
| `fare\_amount` | 0 |
| `extra` | 1 |
| `mta\_tax` | 15 |
| `tip\_amount` | 0 |
| `tolls\_amount` | 0 |
| `improvement\_surcharge` | 15 |
| `total\_amount` | 15 |
| `congestion\_surcharge` | 10 |

These negative values likely represent data entry errors or specific transaction types like refunds that were not properly flagged. For the purpose of this EDA, which focuses on typical trip costs, these values were corrected by taking their absolute value. This approach retains the magnitude of the charge while correcting the invalid sign, assuming the negative was erroneous.

### 4.3. Missing Value Imputation

#### 4.3.1. Proportion of Missing Values

The proportion of missing values in each column was calculated to prioritize imputation efforts.

|  |  |
| --- | --- |
| Column | Percentage Missing |
| `airport\_fee` | 92.18% |
| `congestion\_surcharge` | 3.41% |
| `passenger\_count` | 3.41% |
| `RatecodeID` | 3.41% |
| `store\_and\_fwd\_flag` | 3.41% |
| `(Other columns)` | 0.00% |

The `airport\_fee` column had a very high proportion of missing values. The other columns had an identical missing proportion of approximately 3.41%.

#### 4.3.2. `passenger\_count` Imputation

The `passenger\_count` column contained both null values and entries of zero.

* **Null Values:** The null values were imputed using the median of the column. The median is a robust measure of central tendency for skewed distributions and is less affected by outliers than the mean, making it a suitable choice for this driver-entered field.
* **Zero Values:** A `passenger\_count` of zero is logically invalid for a completed taxi trip. These entries were treated as data errors and were replaced with a value of 1, representing the most common scenario of a single passenger.

#### 4.3.3. `RatecodeID` Imputation

`RatecodeID` null values were imputed with the value `1`, which corresponds to the "Standard rate". This is the most frequent and default rate code, making it the most probable value for the missing entries.

#### 4.3.4. `congestion\_surcharge` Imputation

The `congestion\_surcharge` is applied only to trips within specific zones. Missing values in this column likely indicate that the trip occurred outside these zones and no surcharge was applicable. Therefore, null values were imputed with `0`.

#### 4.3.5. Imputation of Remaining Nulls

The `airport\_fee` column, with over 92% missing values, also represents a situational charge. As with the congestion surcharge, a missing value implies the fee was not applicable. Therefore, all remaining null values across the DataFrame, primarily in `airport\_fee` and `store\_and\_fwd\_flag`, were filled with `0`.

### 4.4. Outlier Analysis and Treatment

#### 4.4.1. Descriptive Statistics for Outlier Detection

A descriptive statistical summary was generated to identify potential outliers.

|  |  |  |  |
| --- | --- | --- | --- |
|  | \*\*trip\_distance\*\* | \*\*fare\_amount\*\* | \*\*total\_amount\*\* |
| \*\*count\*\* | 379268.00 | 379268.00 | 379268.00 |
| \*\*mean\*\* | 3.61 | 20.18 | 29.22 |
| \*\*std\*\* | 49.39 | 233.14 | 233.53 |
| \*\*min\*\* | 0.00 | 0.00 | 0.00 |
| \*\*25%\*\* | 1.05 | 9.30 | 15.96 |
| \*\*50%\*\* | 1.79 | 13.50 | 21.00 |
| \*\*75%\*\* | 3.40 | 21.90 | 30.75 |
| \*\*max\*\* | 22528.82 | 143163.45 | 143167.45 |

The summary revealed several anomalies:

* The maximum `trip\_distance` of over 22,528 miles is physically impossible for a taxi trip and is a clear data error.
* The maximum `fare\_amount` and `total\_amount` are exceptionally high, suggesting data entry errors or non-standard trips not suitable for this analysis.
* The `passenger\_count` had a maximum of 9, which is unlikely for a standard yellow taxi.
* The `payment\_type` contained a value of 0, which is not defined in the data dictionary.

#### 4.4.2. Filtering Invalid Trip Records

Based on the outlier analysis, several filtering criteria were applied to clean the dataset:

* Passenger Count: Trips with a `passenger\_count` greater than 6 were removed, as standard taxis typically do not accommodate more passengers.
* Trip Distance: Trips with a `trip\_distance` greater than 250 miles were filtered out as they are extreme outliers and likely errors.
* Anomalous Fare/Distance: Records where `trip\_distance` was less than 0.1 miles but `fare\_amount` exceeded $300 were removed, as these indicate a likely meter error or a non-standard negotiated fare.
* Payment Type: Trips with `payment\_type` equal to 0 were removed as this is an undefined category.

These cleaning steps resulted in a refined dataset that better represents typical taxi operations.

### 4.5. Feature Engineering

To facilitate deeper analysis, several new features were engineered from the existing columns:

- `trip\_duration`: Calculated as the difference between `tpep\_dropoff\_datetime` and `tpep\_pickup\_datetime`, converted to minutes.

- `pickup\_hour`: Extracted from `tpep\_pickup\_datetime`, this enables hourly trend analysis.

- `weekday`: Extracted from `tpep\_pickup\_datetime`, this enables analysis of weekly patterns.

- `month`: Extracted from `tpep\_pickup\_datetime` for seasonal analysis.

- `quarter`: Derived from `month` to analyze quarterly trends.

- `is\_weekend`: A boolean feature derived from `weekday` to compare weekday and weekend patterns.

- `speed`: Calculated as `trip\_distance` / (`trip\_duration` / 60) to measure average trip speed in miles per hour.

- `fare\_per\_mile`: Calculated as `fare\_amount` / `trip\_distance`.

- `tip\_percentage`: Calculated as (`tip\_amount` / `fare\_amount`) \* 100.

- `distance\_tier`: A categorical variable binning `trip\_distance` into '0-2 miles', '2-5 miles', and '5+ miles' for tiered analysis.

## 5. Exploratory Data Analysis

### 5.1. Variable Classification

The variables in the cleaned dataset were classified into numerical and categorical types to guide the selection of appropriate analytical techniques.

Categorical Variables: `VendorID`, `RatecodeID`, `PULocationID`, `DOLocationID`, `payment\_type`, `pickup\_hour`, `weekday`, `month`, `quarter`, `is\_weekend`, `distance\_tier`.

Numerical Variables: `passenger\_count`, `trip\_distance`, `trip\_duration`, `speed`, and all monetary variables including `fare\_amount`, `tip\_amount`, `total\_amount`, and derived metrics like `fare\_per\_mile`.

### 5.2. Temporal Analysis

#### 5.2.1. Hourly, Daily, and Monthly Pickup Distributions

The distribution of taxi pickups was analyzed across different time scales to identify key demand patterns.

Hourly Trends: The number of taxi pickups exhibits a distinct bimodal pattern throughout the day. A smaller peak occurs during the morning commute around 8:00 AM, followed by a much larger, dominant peak in the evening at 18:00 (6:00 PM). Demand drops significantly after midnight, reaching its lowest point between 4:00 and 5:00 AM.

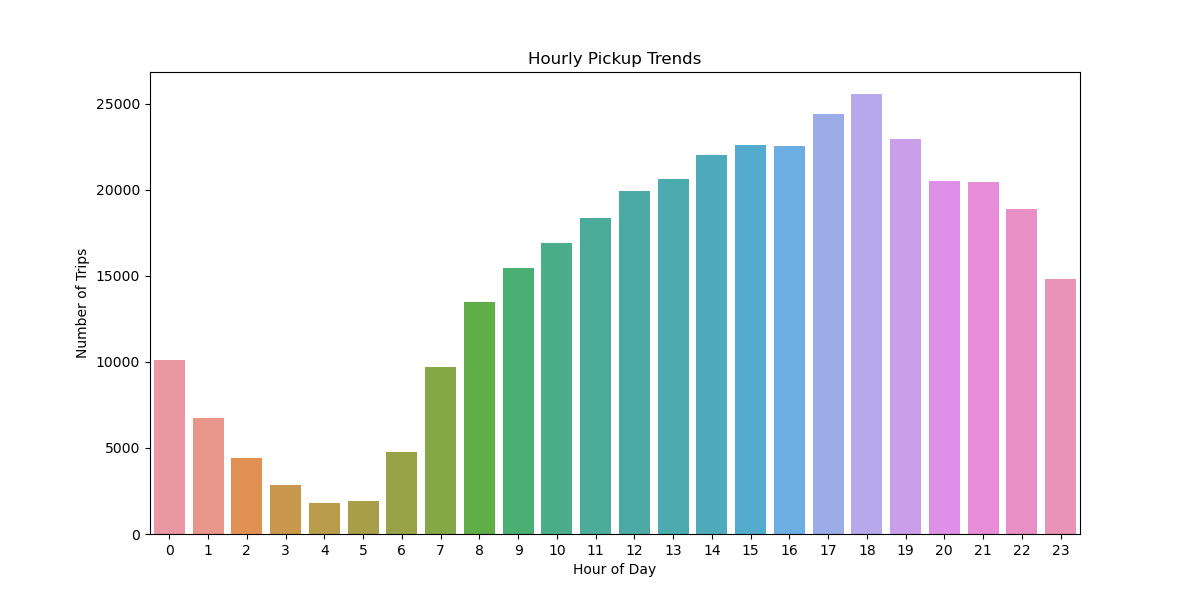


Figure 1

Daily Trends: Analysis of pickups by day of the week shows that demand is highest on Thursdays, closely followed by Fridays and Saturdays. Demand is lowest on Sundays and Mondays.

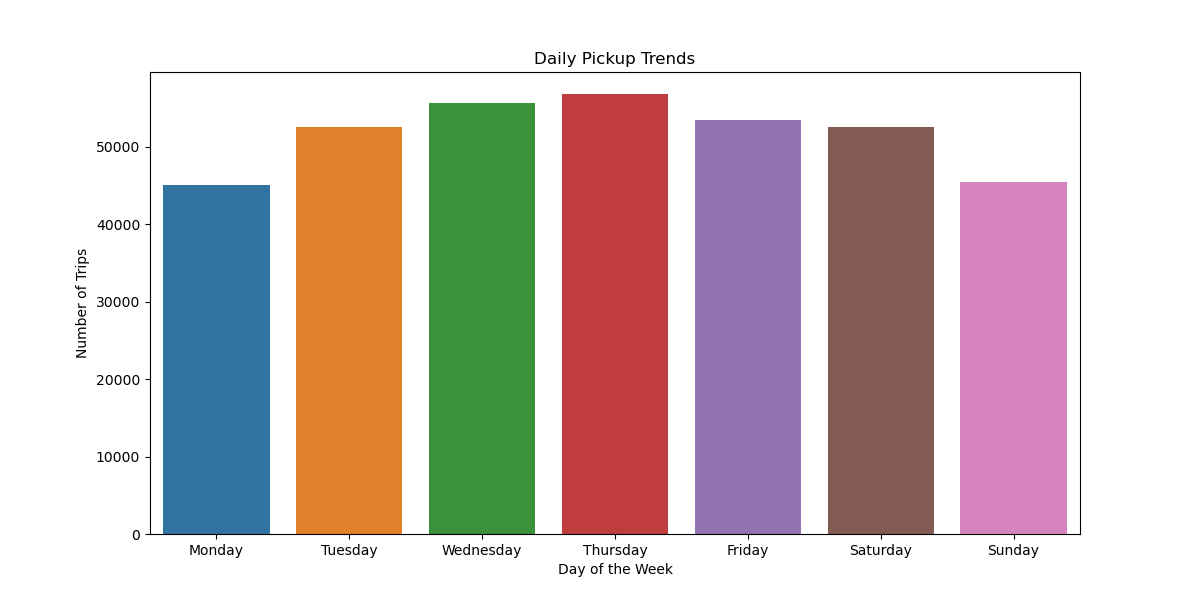


Figure 2

Monthly Trends: On a monthly basis, demand shows seasonal variation. The busiest months are May, March, and October. The summer months of July and August show a slight dip in activity.

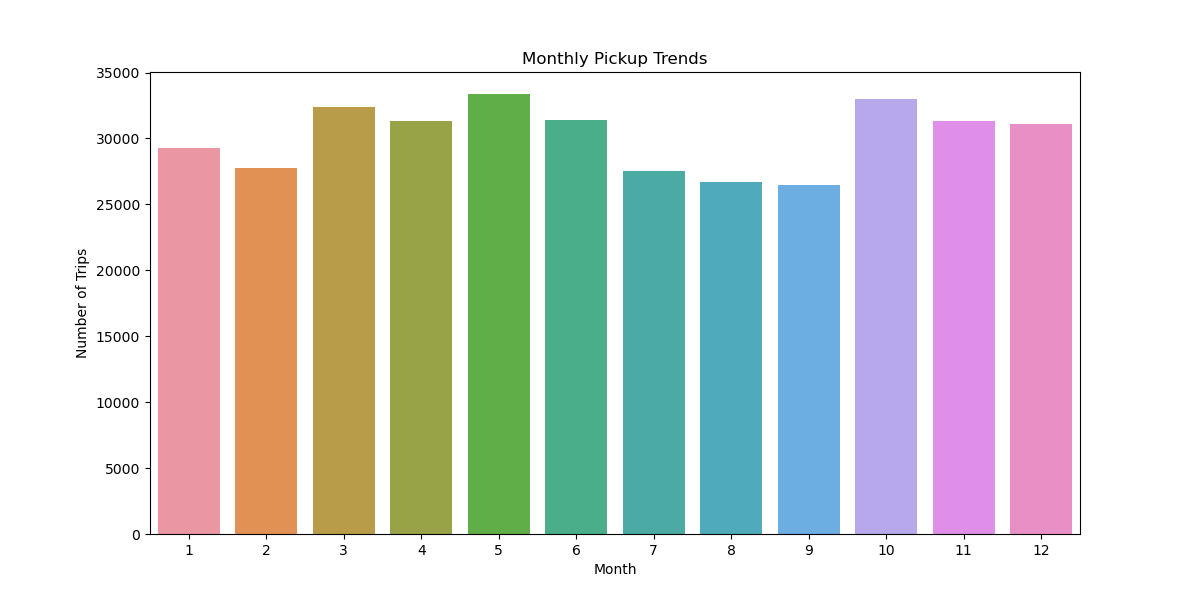


Figure 3

### 5.3. Financial Analysis

#### 5.3.1. Analysis of Zero-Value Records

An investigation of key financial and distance columns revealed a notable number of zero-value entries even after initial cleaning. For subsequent financial and distance-based calculations, these records were filtered out to avoid skewing results. A new DataFrame, `df\_filtered`, was created by removing records where `fare\_amount`, `total\_amount`, or `trip\_distance` was zero, reducing the dataset from 366,312 to 361,720 records for these specific analyses.

#### 5.3.2. Monthly Revenue Trend

The total revenue (`total\_amount`) was aggregated by month to observe seasonal financial performance. The trend shows that revenue is highest in the second and fourth quarters. May and October stand out as particularly high-revenue months, aligning with the peaks observed in monthly trip counts.

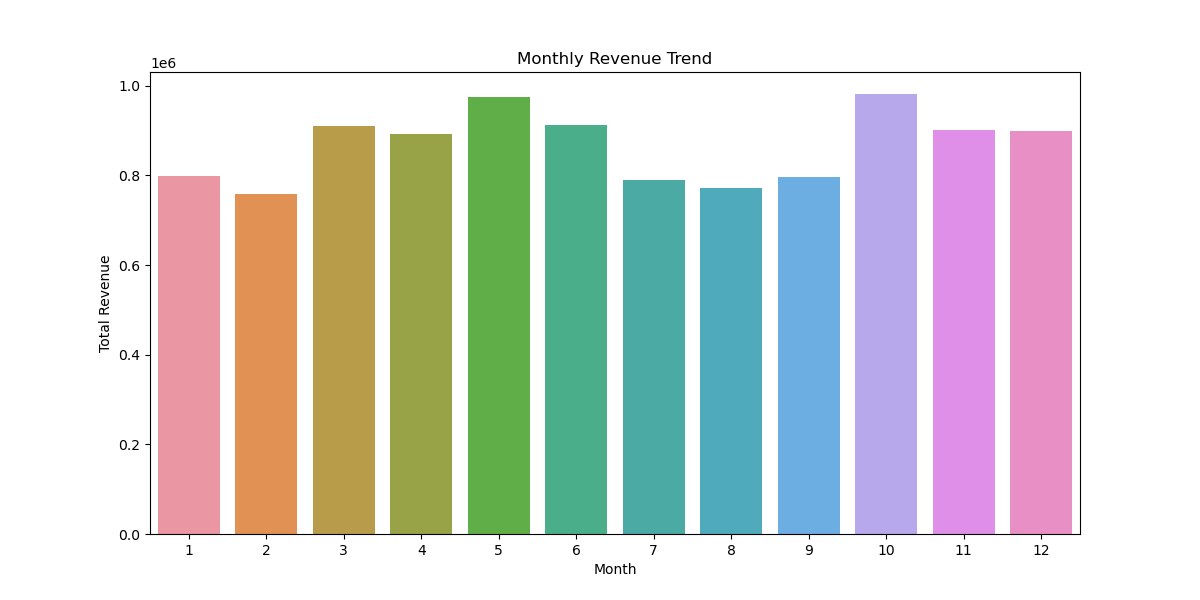


Figure 4

#### 5.3.3. Quarterly Revenue Distribution

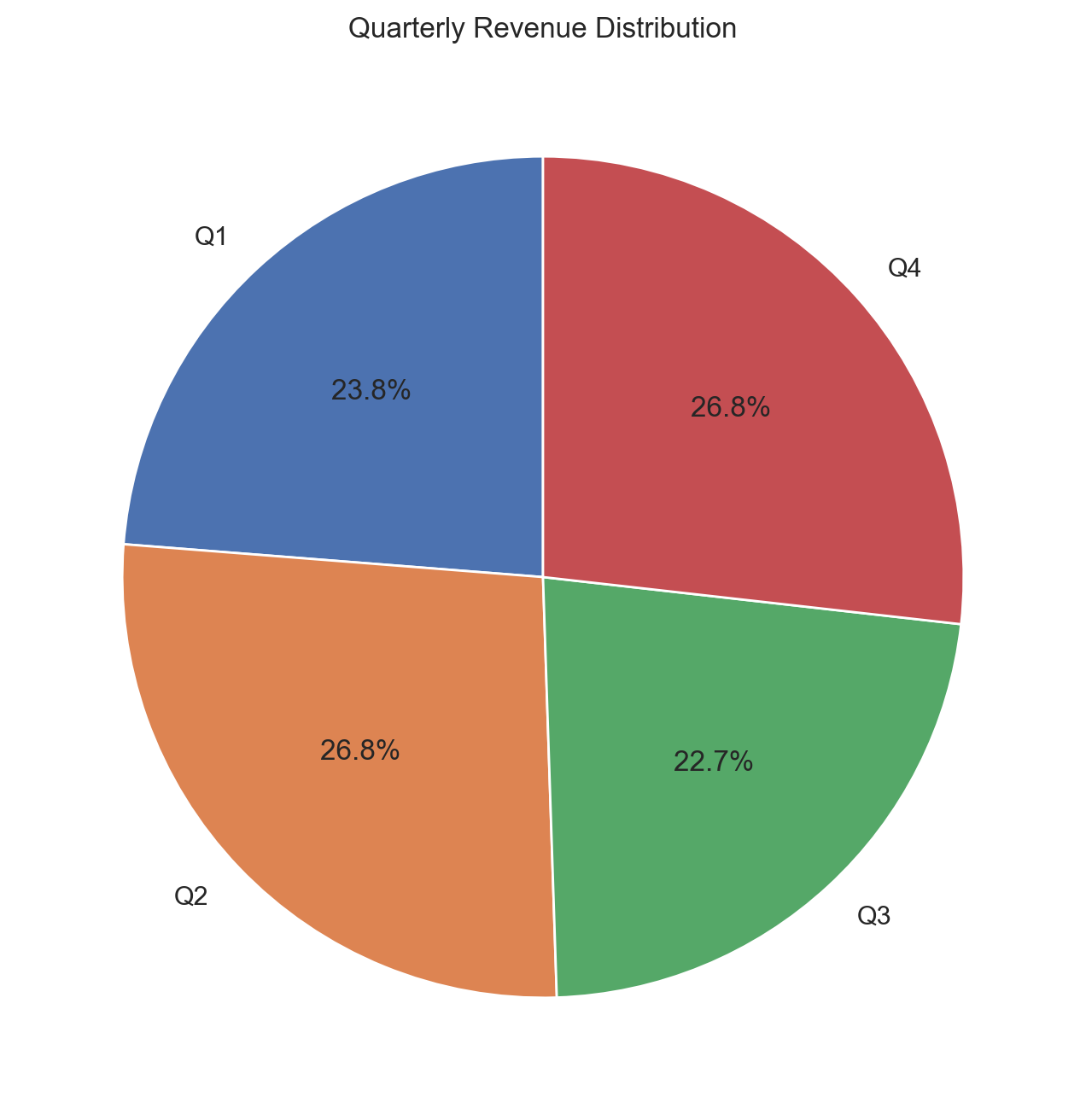
Revenue was grouped by quarter to better understand the yearly financial cycle. The analysis shows that Q2 and Q4 are the most lucrative, each contributing approximately 26.8% of the total annual revenue. Q3 is the least profitable quarter at 22.7%.

Figure 5

#### 5.3.4. Correlation Analysis: Distance, Duration, and Fare

Distance vs. Fare: A scatter plot of trip distance versus fare amount reveals a strong, positive linear relationship. The Pearson correlation coefficient between these two variables is 0.945, indicating that distance is a primary determinant of the final fare.

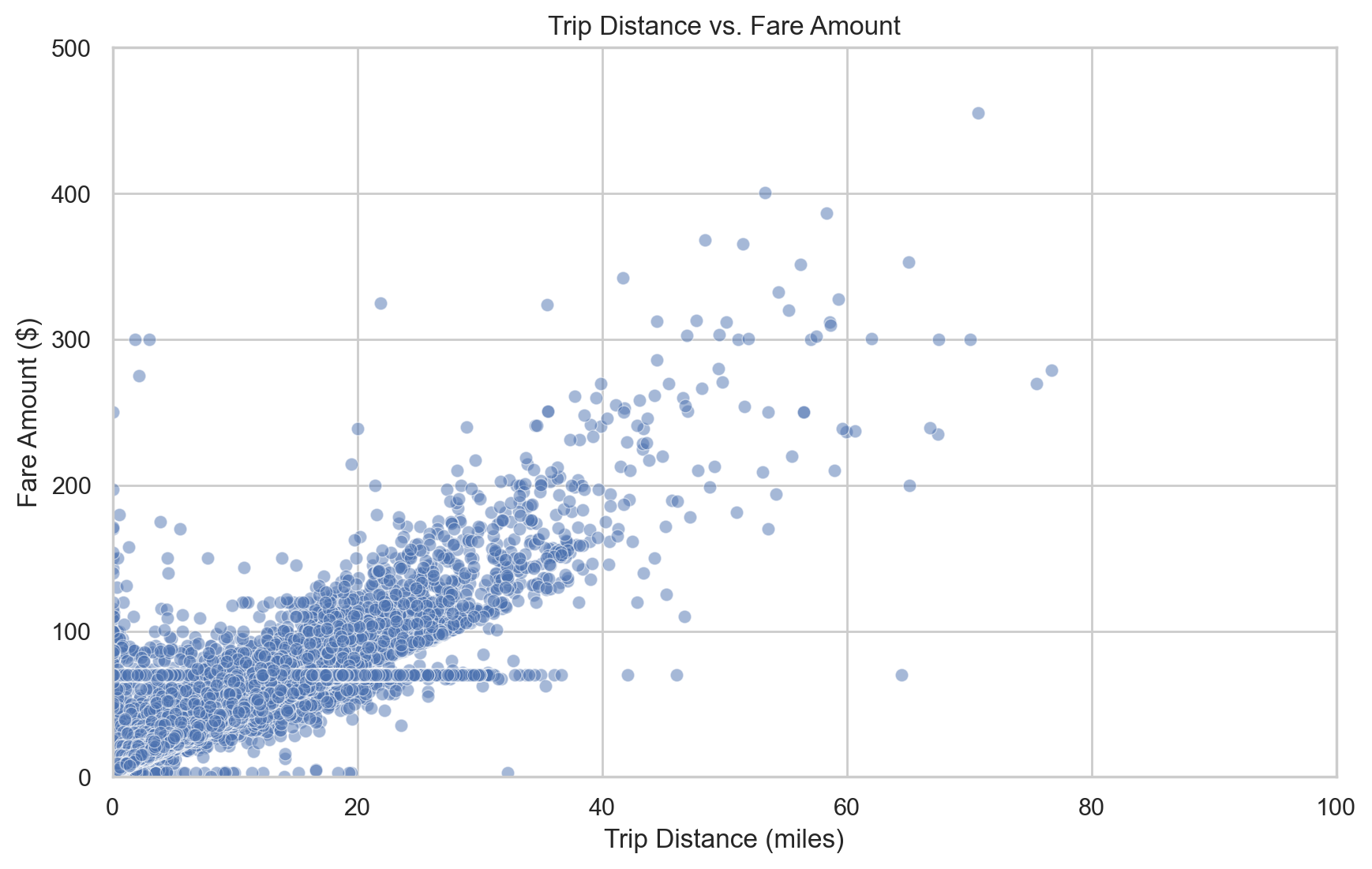


Figure 6

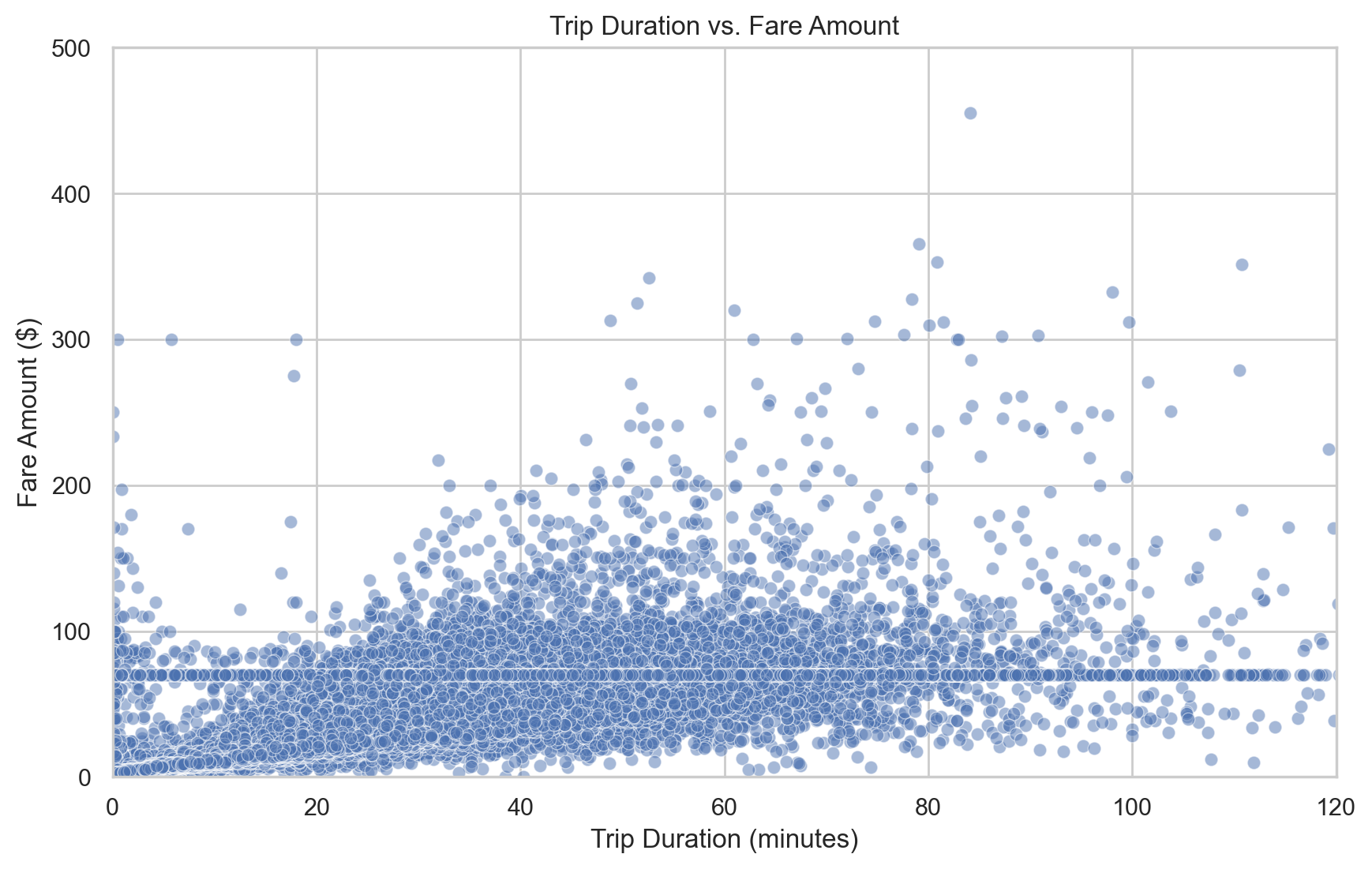
Duration vs. Fare: A scatter plot comparing trip duration and fare amount shows a positive correlation with a Pearson coefficient of 0.278. This suggests a moderate positive relationship, indicating that while longer trips generally cost more, other factors like traffic create variability.

Figure 7

Fare by Passenger Count: The data shows that fares tend to be slightly higher for trips with more passengers, though the difference is not substantial.

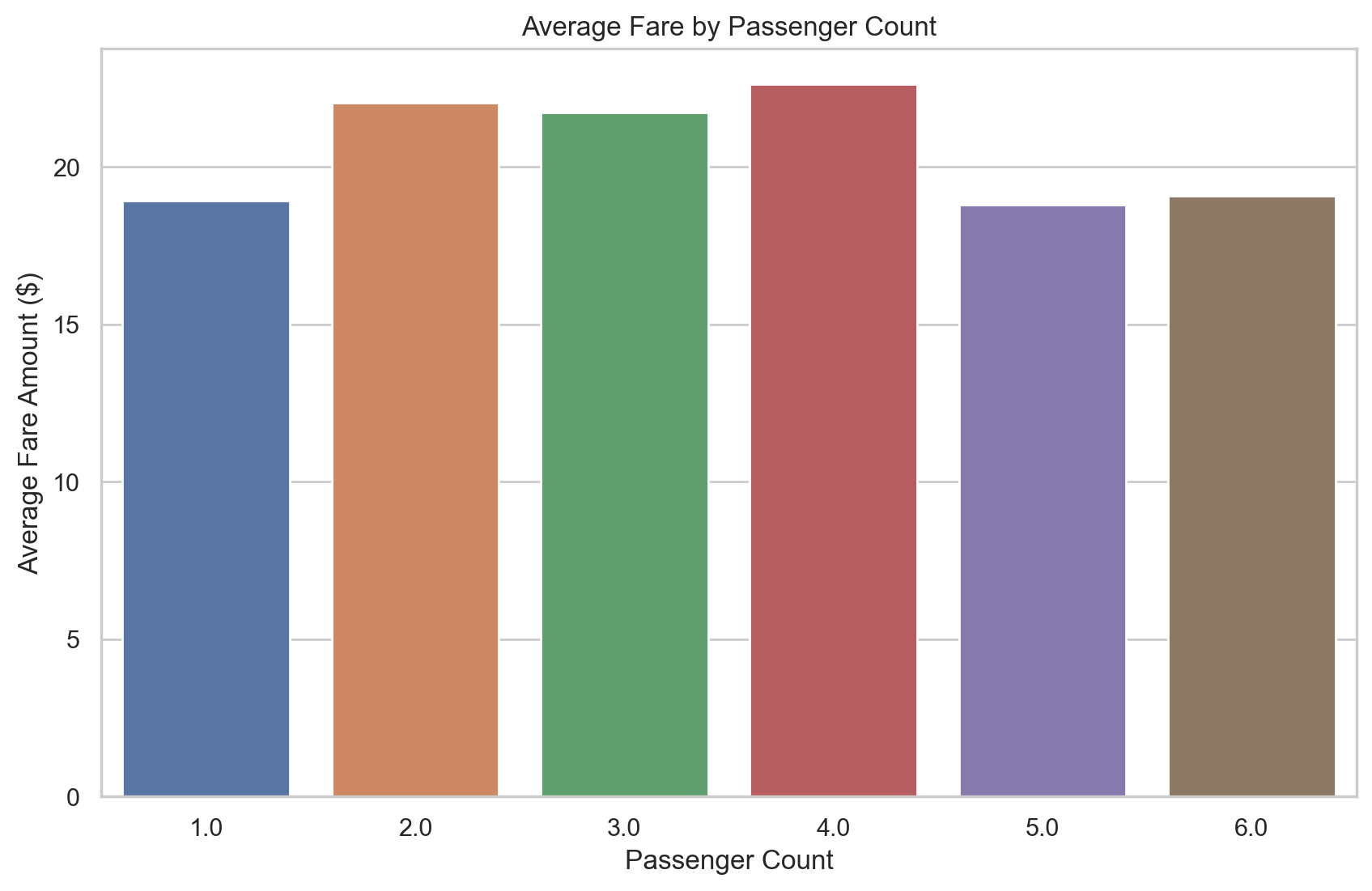


Figure 8

Tip vs. Distance: The scatter plot shows a positive correlation between tip amount and trip distance, with longer trips generally receiving higher tips in absolute terms.

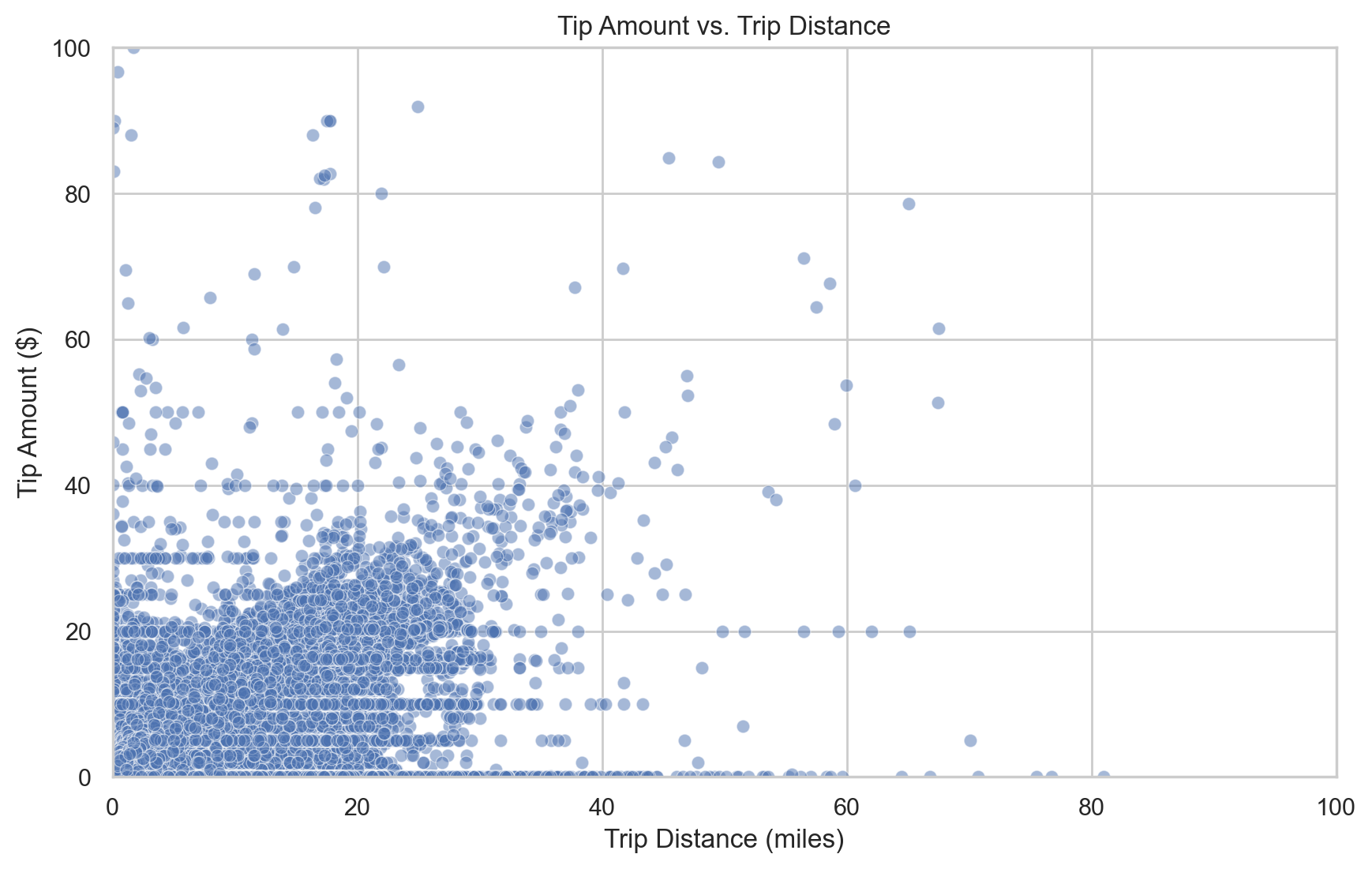


Figure 9

#### 5.3.5. Payment Type Distribution

The analysis of payment types shows a clear passenger preference for credit card (Type 1) over cash (Type 2). "No charge" (Type 3) and "Dispute" (Type 4) trips represent a very small fraction of total transactions.

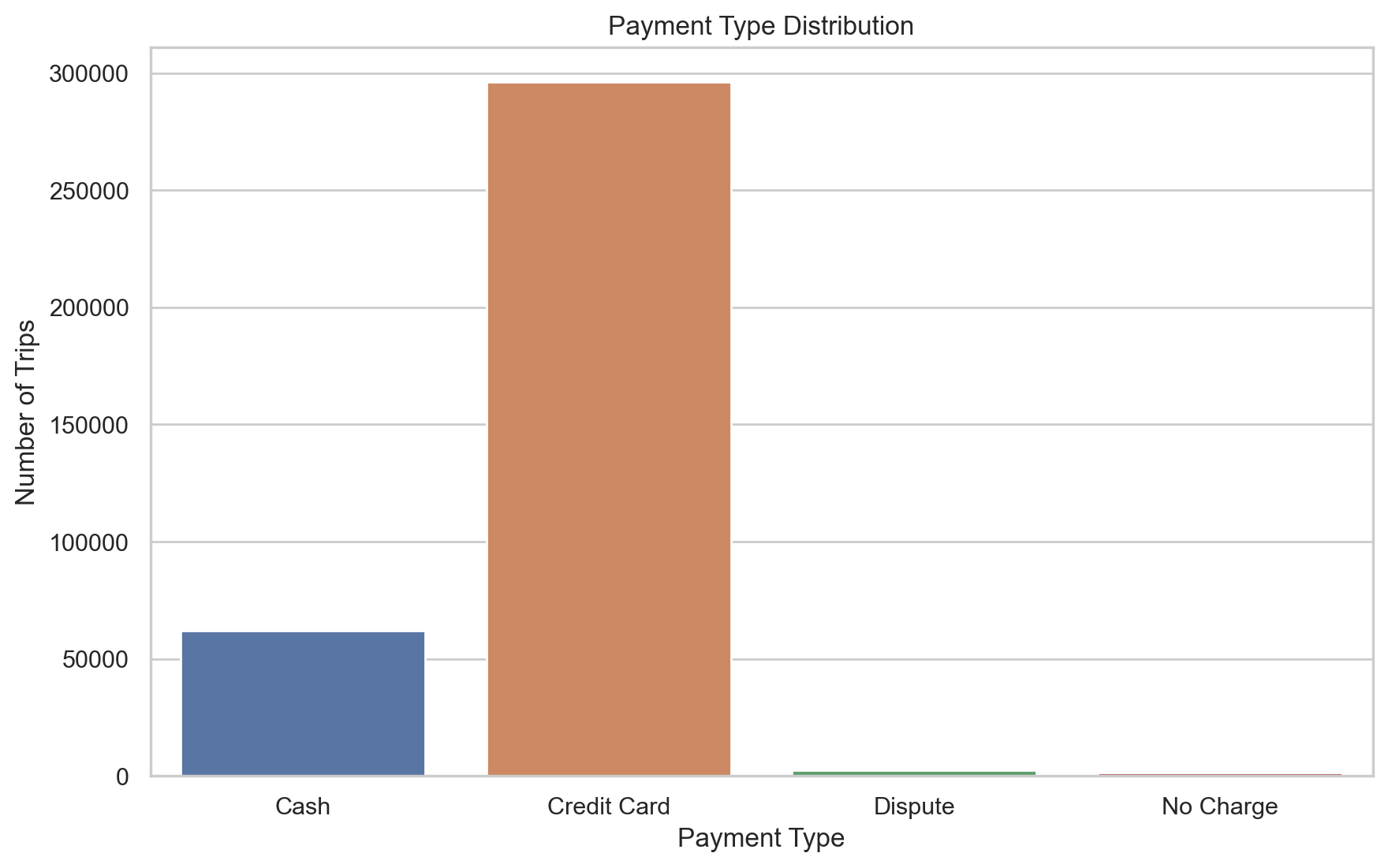


Figure 10

### 5.4. Geospatial Analysis

#### 5.4.1. Loading and Merging Taxi Zone Data

To analyze the geographic distribution of trips, the `taxi\_zones.shp` shapefile was loaded using the GeoPandas library. This file contains the geometric polygons for each of the 263 official TLC taxi zones. The trip data DataFrame was then merged with this GeoDataFrame on the `PULocationID`, enriching the trip data with geographic information for each pickup.

#### 5.4.2. Mapping Pickup Density by Zone

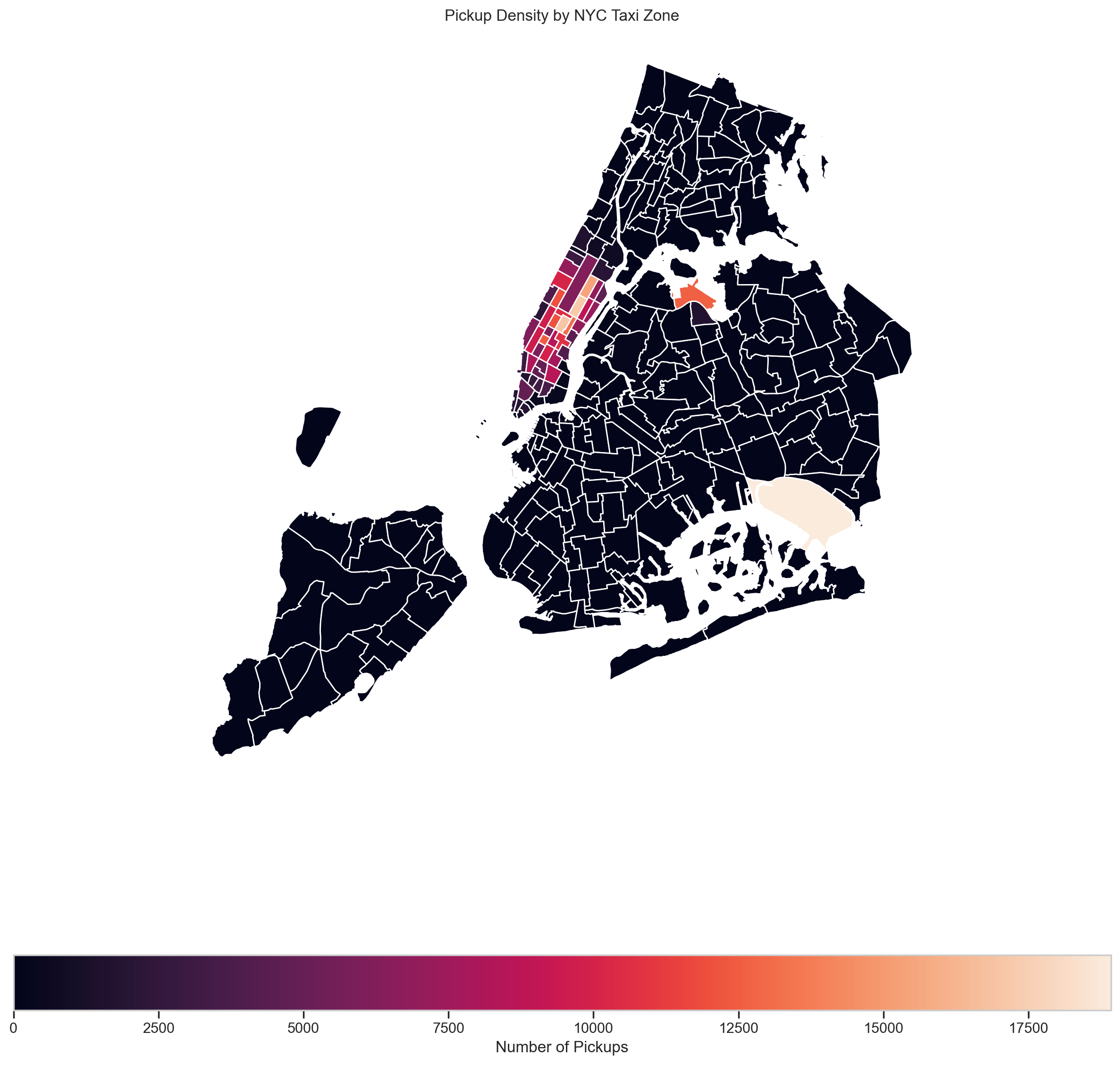
A choropleth map was generated to visualize the concentration of taxi pickups across NYC. The map clearly illustrates that taxi activity is heavily concentrated in Manhattan. Other boroughs show significantly lower pickup density.

Figure 11

The top pickup zones, sorted by trip count, confirm this observation, with JFK Airport, Upper East Side South, and Midtown Center being the top three.

### 5.5. Detailed Operational Analysis

#### 5.5.1. Route Speed and Efficiency Analysis

The average speed for trips was calculated to identify routes prone to congestion. The routes with the lowest positive speeds often involve travel through central Manhattan during rush hour, indicating significant operational inefficiencies due to traffic.

#### 5.5.2. Peak Hour Trip Volume Analysis

A detailed hourly analysis of trip volume confirmed that the single busiest hour is 18:00 (6:00 PM), with 25,837 trips in the sampled dataset. Scaling this by the inverse of the 1% sampling fraction projects to approximately 2,583,700 trips during this hour across the year. The top five busiest hours were identified and scaled similarly.

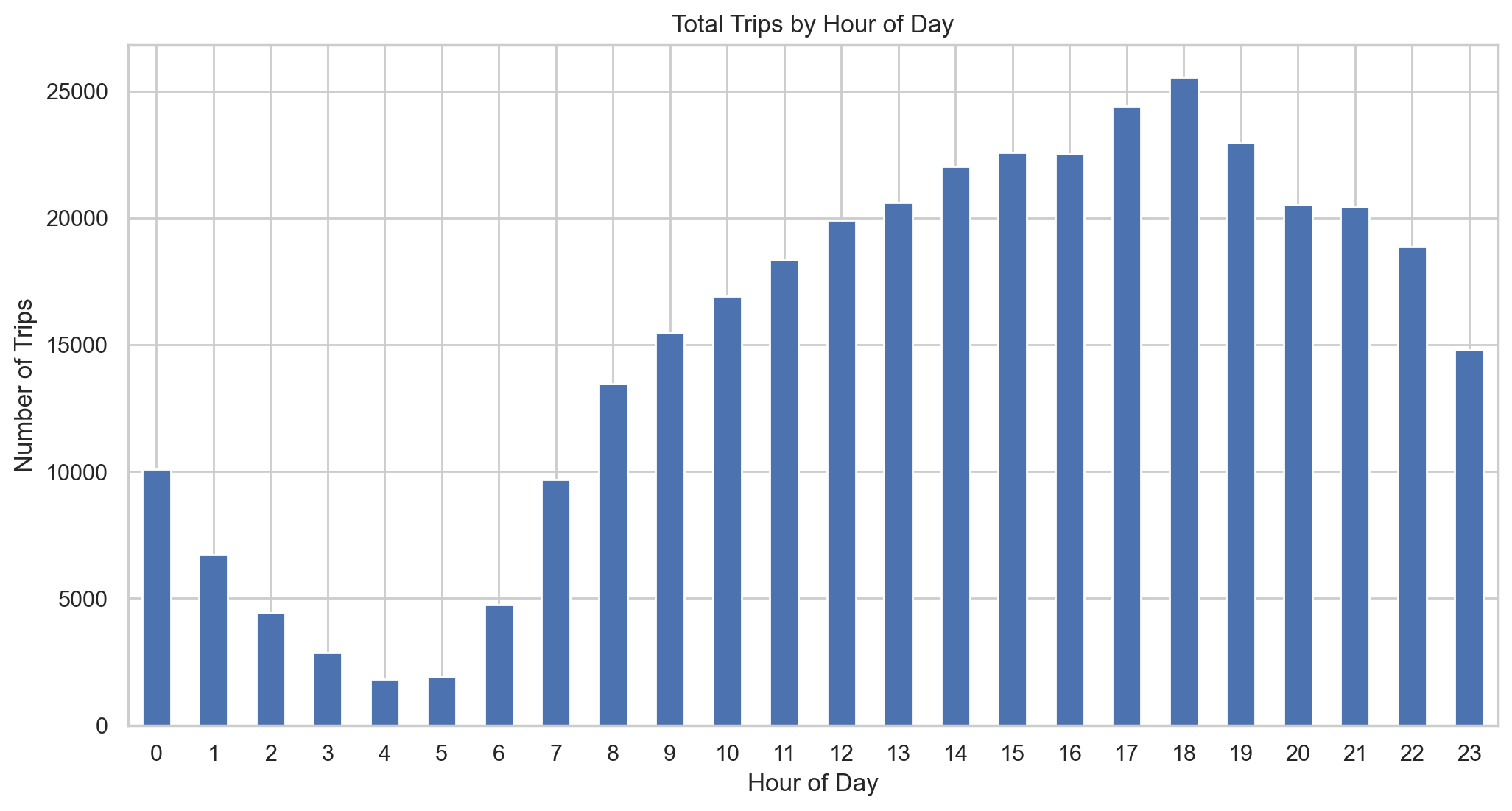


Figure 12

#### 5.5.3. Weekday vs. Weekend Traffic Patterns

A comparative analysis of hourly pickup patterns for weekdays versus weekends revealed distinct behaviors. Weekdays exhibit the classic bimodal commuting pattern. Weekends show demand building more gradually, peaking in the late afternoon/evening and staying higher later into the night, reflecting leisure activities.

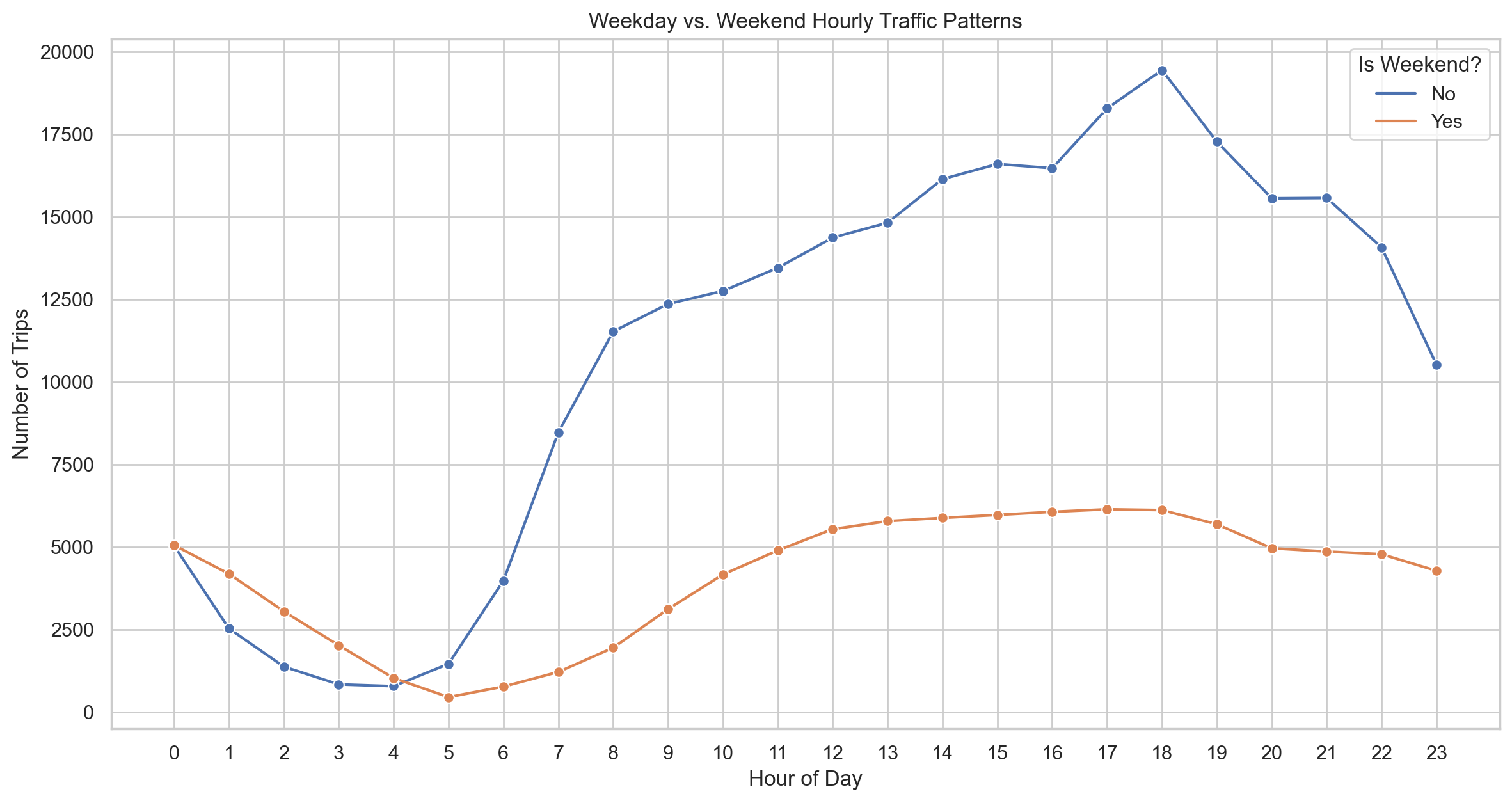


Figure 13

#### 5.5.4. Analysis of High-Demand Zones

The analysis identified the top 10 pickup and dropoff zones by trip volume. Top pickup zones include JFK Airport, Upper East Side South, and Midtown Center. Top dropoff zones largely overlap but also include other central Manhattan neighborhoods.

#### 5.5.5. Pickup-to-Dropoff Ratio Analysis

The ratio of pickups to dropoffs was calculated for each zone to identify geographic imbalances.

- High Ratios (More Pickups): Zones like JFK Airport (4.67) and LaGuardia Airport (2.91) are major trip originators.  
- Low Ratios (More Dropoffs): Several residential zones in outer boroughs are primarily destinations.

These imbalances create deadheading challenges for drivers.

#### 5.5.6. Nighttime Operations Analysis

An analysis focused on night hours (11 PM to 5 AM) was conducted. Night operations account for 11.97% of total revenue. The geographic focus shifts significantly to nightlife districts like the East Village and West Village, as well as airports.

#### 5.5.7. Fare and Pricing Structure Analysis

Fare per mile was analyzed across different segments.

Hourly and Daily Fare Per Mile: Average fare per mile varies by time of day, peaking during late-night hours. It is lowest during the morning rush hour due to high congestion. Daily analysis shows that fares per mile are highest on Sundays.

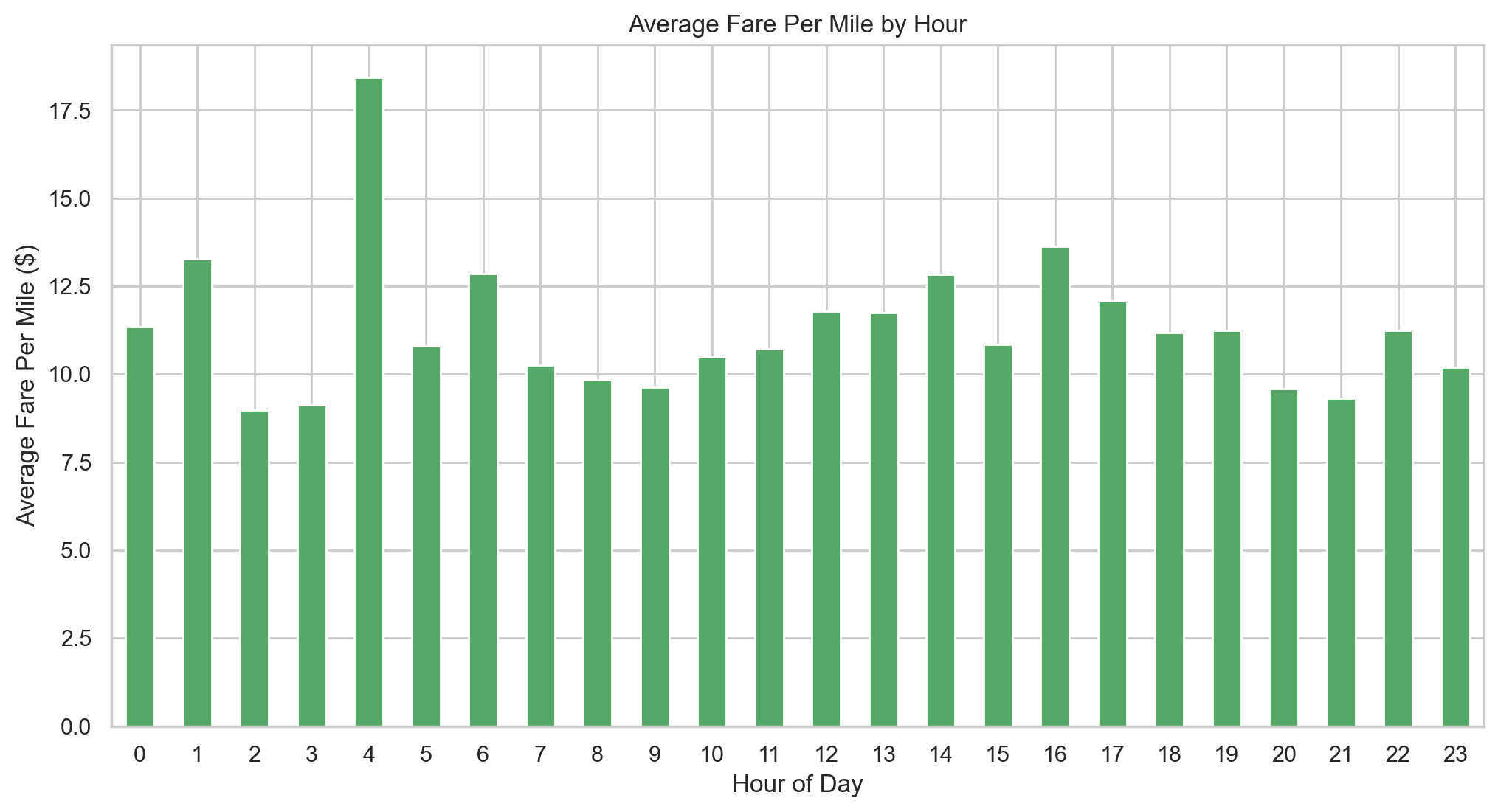


Figure 14

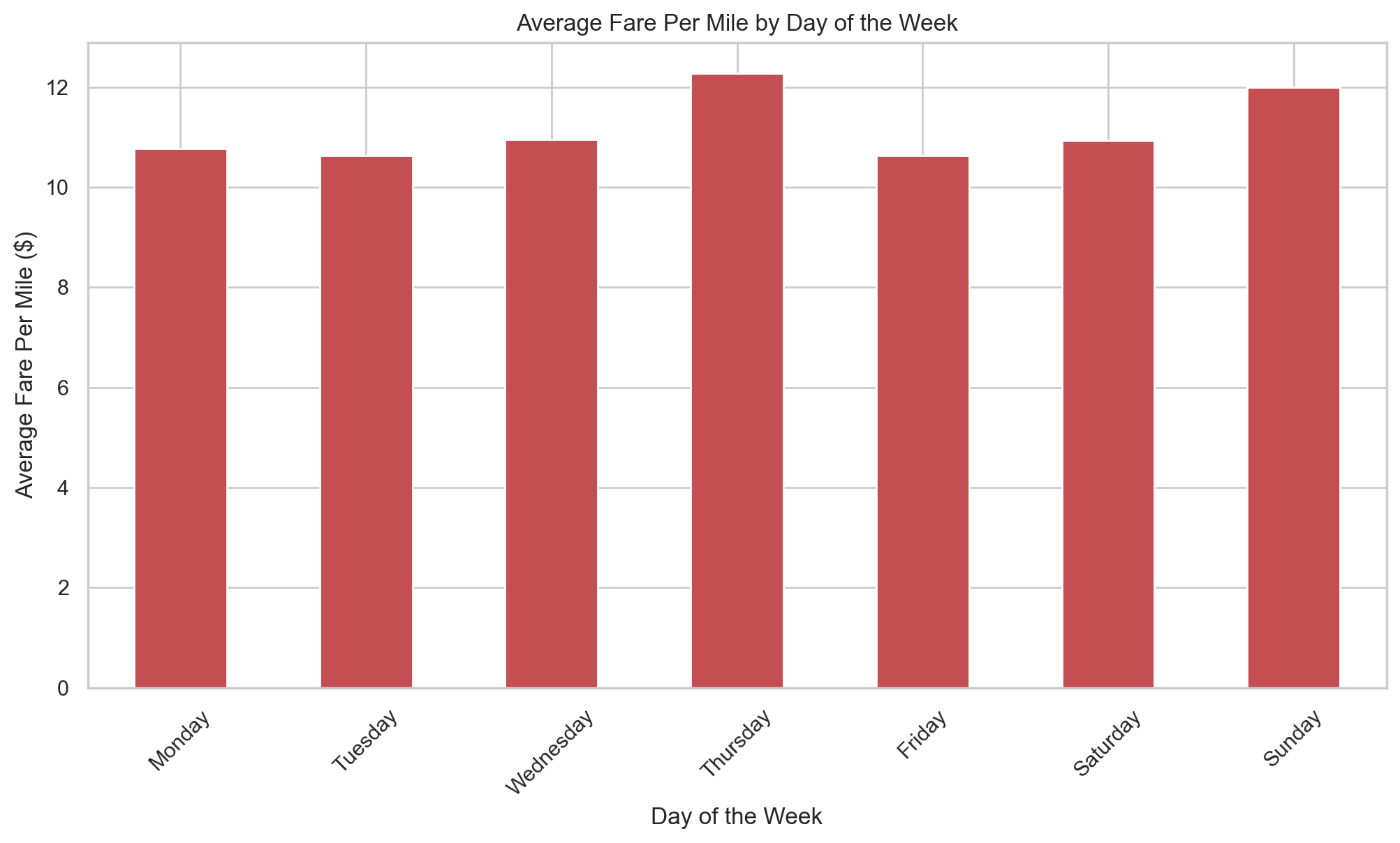


Figure 15

Vendor Fare Rates: The two primary vendors (Vendor 1: CMT, Vendor 2: VeriFone) exhibit different pricing patterns. Vendor 2 generally has a higher fare per mile, especially for short-distance trips (0-2 miles), suggesting different meter or surcharge rules.

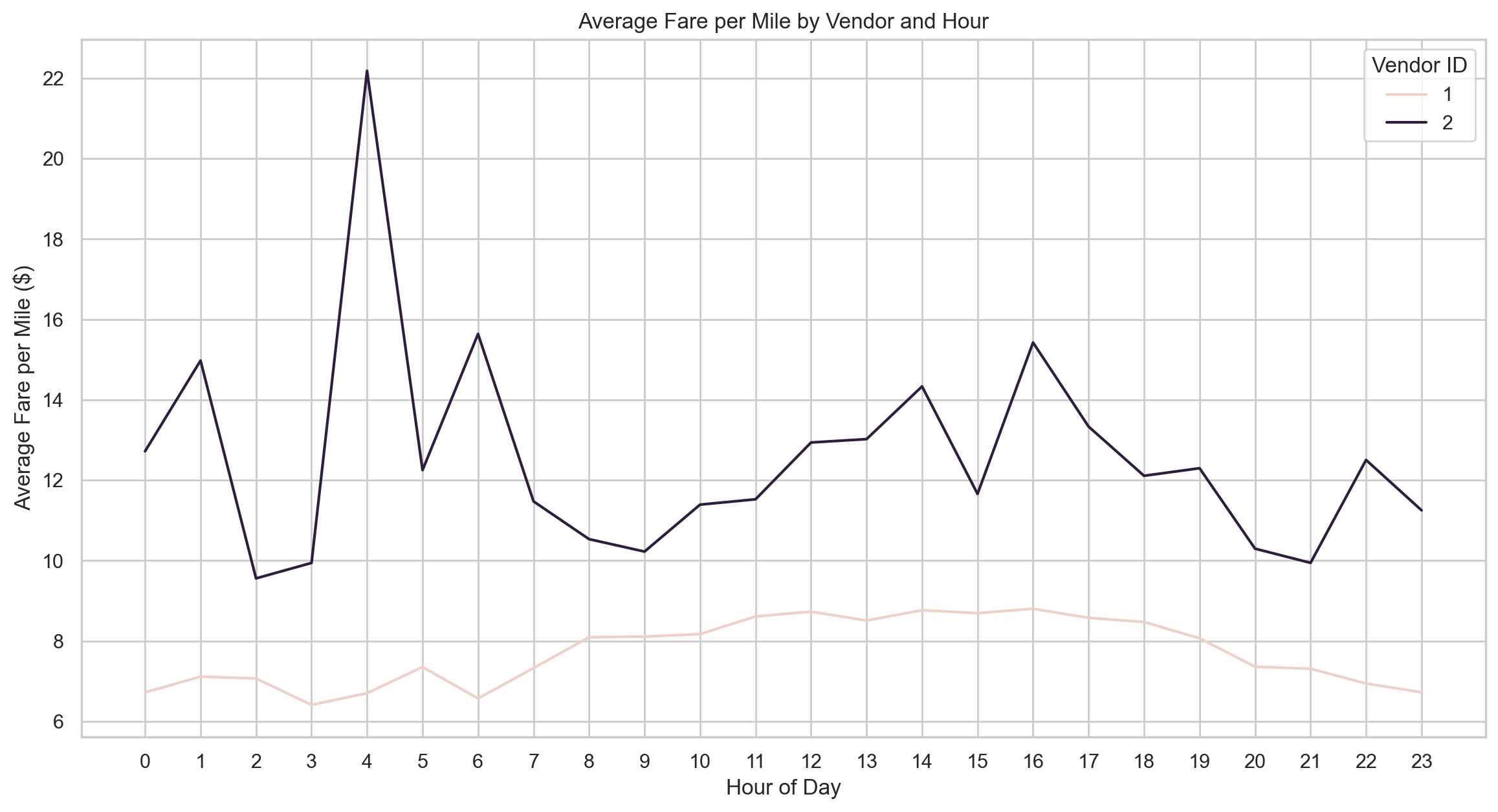


Figure 16

#### 5.5.8. Customer Experience and Tipping Behavior

Tip percentage analysis revealed that tip percentage tends to decrease as trip distance increases. There is no strong pattern between passenger count and tip percentage.

#### 5.5.9. Passenger Count Dynamics

The average number of passengers per trip varies by time and day. Counts are highest during evening/late-night hours and on weekends, reflecting group leisure travel.

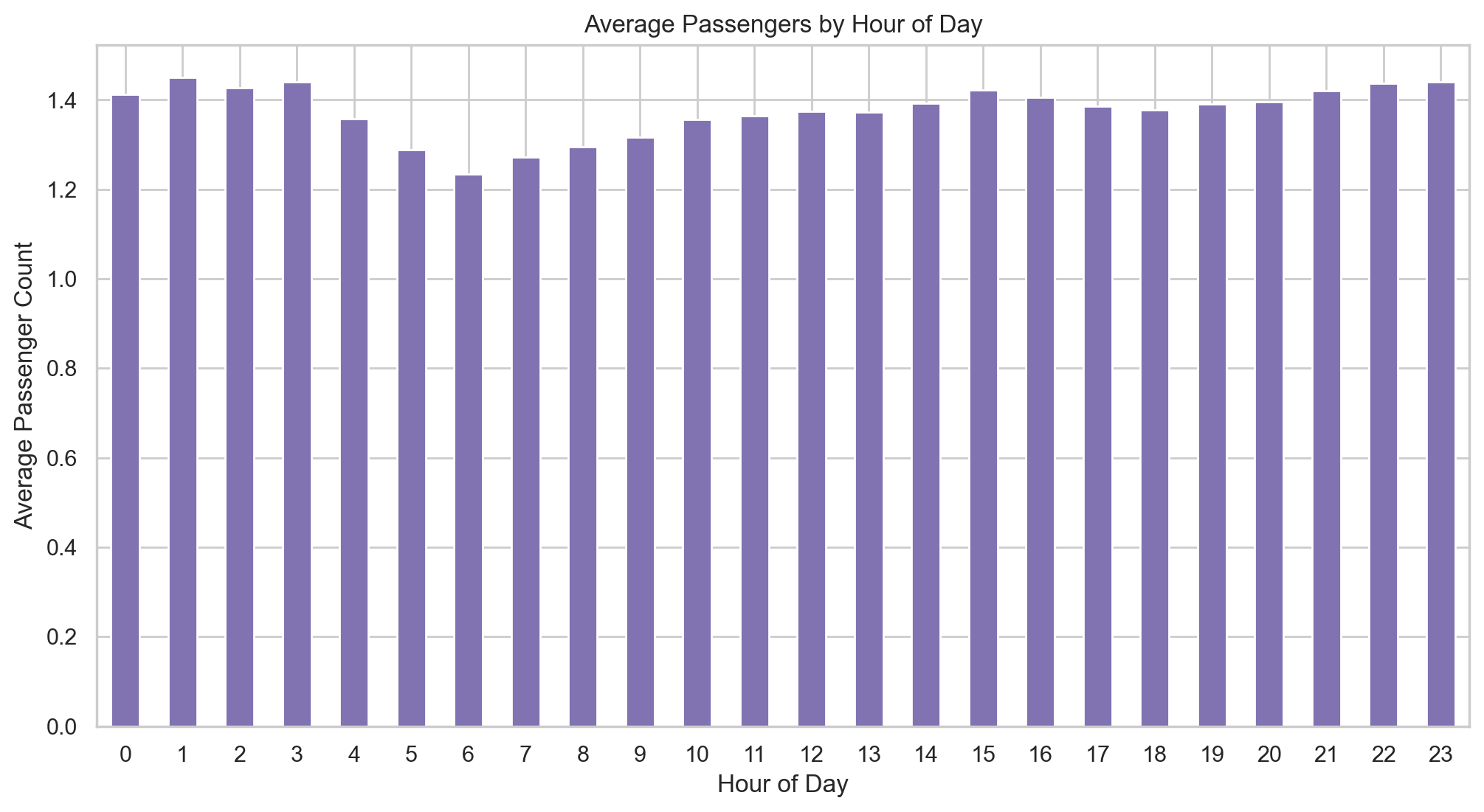


Figure 17

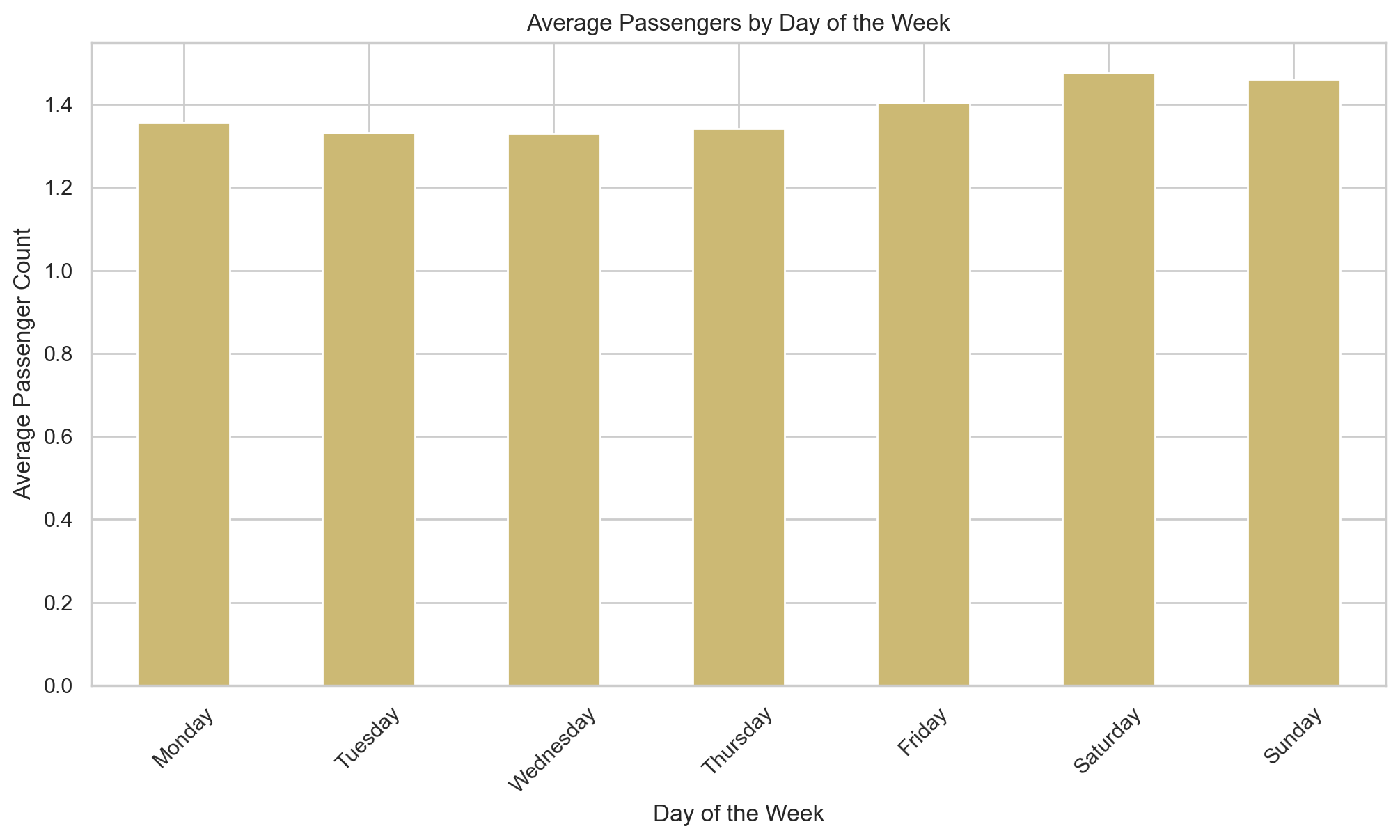


Figure 18

#### 5.5.10. Surcharge Application Analysis

The frequency of various surcharges was analyzed:

- Improvement Surcharge & MTA Tax: Applied to nearly 100% of trips.  
- Congestion Surcharge: Applied to 92.4% of trips, indicating most trips travel through the Manhattan congestion zone.  
- Extra Charges: Applied to 61.9% of trips, likely corresponding to rush hour and overnight surcharges.  
- Tolls: Applied to only 8.0% of trips.

This confirms that regulatory surcharges are a standard component of almost every fare.

## 6. Conclusion: Summary of Analytical Findings

This exploratory data analysis of the 2023 NYC Yellow Taxi dataset has yielded several key findings that characterize the city's urban mobility patterns and taxi operational environment.

The primary analytical findings are as follows:

1. Predictable Temporal Rhythms: Taxi demand follows clear and consistent hourly, daily, and monthly cycles. The most significant demand peak occurs on weekdays at 18:00 (6:00 PM). Weekends exhibit a different pattern, with demand building later and peaking in the evening, aligned with leisure activities. Seasonally, demand and revenue are highest in the second and fourth quarters.

2. Geographic Concentration and Imbalance: Taxi activity is overwhelmingly concentrated in Manhattan. This concentration creates significant operational imbalances, with certain zones acting as net generators of pickups (e.g., airports, business districts) and others as net receivers (e.g., residential areas). This necessitates driver deadheading, reducing overall fleet efficiency.

3. Fare Structure and Congestion Impact: While trip distance is the strongest predictor of the final fare (correlation 0.945), trip duration also plays a significant role. The weaker correlation with duration (0.278) suggests the substantial impact of traffic congestion, where time-based charges accumulate even over short distances.

4. Payment and Tipping Behavior: There is a clear passenger preference for credit card payments over cash. This choice has a direct impact on recorded revenue, as credit card transactions are associated with significantly higher and more consistently recorded tips.

In summary, this EDA provides a data-grounded narrative of the NYC taxi ecosystem. It quantifies the predictable nature of demand, highlights key areas of operational inefficiency, deconstructs the primary drivers of fare revenue, and offers insights into passenger behavior. These findings serve as a comprehensive analytical foundation for understanding the complex dynamics of urban transportation.