Table of Contents

[1. Introduction 2](#_Toc208155953)

[2. Data Overview 2](#_Toc208155954)

[Column summary: 2](#_Toc208155955)

[Observations: 3](#_Toc208155956)

[3. Trip Distance Analysis 3](#_Toc208155957)

[3.1 Box Plot 3](#_Toc208155958)

[3.2 Histogram 4](#_Toc208155959)

[4. Total Fare (total\_amount) Analysis 5](#_Toc208155960)

[4.1 Box Plot 5](#_Toc208155961)

[4.2 Histogram 6](#_Toc208155962)

[5. Tip Amount Analysis 6](#_Toc208155963)

[5.1 Overall Tip Distribution 6](#_Toc208155964)

[5.2 Tip by Vendor 7](#_Toc208155965)

[5.3 Mean Tip by Passenger Count 7](#_Toc208155966)

[6. Passenger Count Distribution 8](#_Toc208155967)

[Observation: 9](#_Toc208155968)

[7. Temporal Analysis 9](#_Toc208155969)

[7.1 Trips by Month 9](#_Toc208155970)

[7.2 Trips by Day of Week 10](#_Toc208155971)

[7.3 Revenue by Day 11](#_Toc208155972)

[7.4 Revenue by Month 12](#_Toc208155973)

[8. Drop-off Locations Analysis 13](#_Toc208155974)

[9. Trip Duration Analysis 14](#_Toc208155975)

[10. Key Insights & Business Implications 14](#_Toc208155976)

[11. Recommendations for Further Analysis 14](#_Toc208155977)

[12. Conclusion 14](#_Toc208155978)

**2017 Yellow Taxi Trip Data Analysis Report Part 2**

# 1. Introduction

The dataset contains information about yellow taxi trips in 2017, including:

* **VendorID** – Taxi service provider
* **Pickup and dropoff timestamps** – When the trip started and ended
* **Passenger count** – Number of passengers
* **Trip distance** – Miles traveled
* **Payment details** – Fare, tip, tolls, and total amount
* **Pickup and dropoff location IDs** – Numeric identifiers for locations

The goal of this analysis is to understand patterns in taxi trips, revenue, tips, distances, passenger distribution, and temporal trends.

# 2. Data Overview

We loaded 22,699 taxi trip records.

## Column summary:

|  |  |  |
| --- | --- | --- |
| Column | Type | Notes |
| VendorID | int | Taxi service provider |
| tpep\_pickup\_datetime | datetime | Trip start time |
| tpep\_dropoff\_datetime | datetime | Trip end time |
| passenger\_count | int | Number of passengers |
| trip\_distance | float | Trip distance in miles |
| RatecodeID | int | Rate type code |
| PULocationID | int | Pickup location ID |
| DOLocationID | int | Dropoff location ID |
| payment\_type | int | Payment method |
| fare\_amount | float | Base fare |
| extra | float | Additional charges |
| mta\_tax | float | MTA tax |
| tip\_amount | float | Tip given |
| tolls\_amount | float | Tolls |
| improvement\_surcharge | float | Extra fee |
| total\_amount | float | Total fare including tip |
| month | str | Month of trip |
| day | str | Weekday of trip |
| trip\_duration | timedelta | Time between pickup and dropoff |

## Observations:

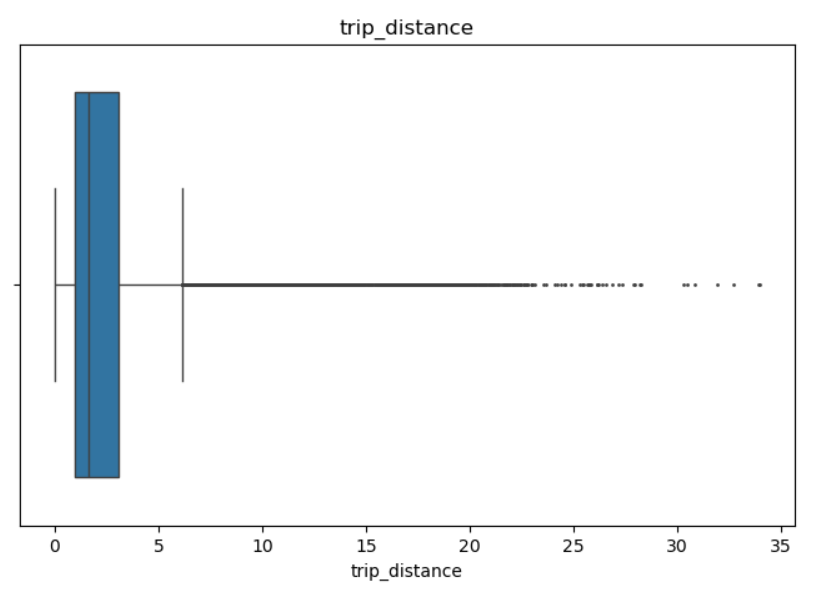
* No missing values in the dataset.
* Data types were adjusted for proper analysis (e.g., datetime conversion for timestamps).

# 3. Trip Distance Analysis

## 3.1 Box Plot

* **Median distance:** ~1.6 miles (50% of trips ≤1.6 miles)
* **Interquartile range (IQR):** 1–3 miles (middle 50% of trips)
* **Whiskers:** Most trips are ≤6 miles
* **Outliers:** Some trips reach 20–35 miles, showing a few exceptionally long trips

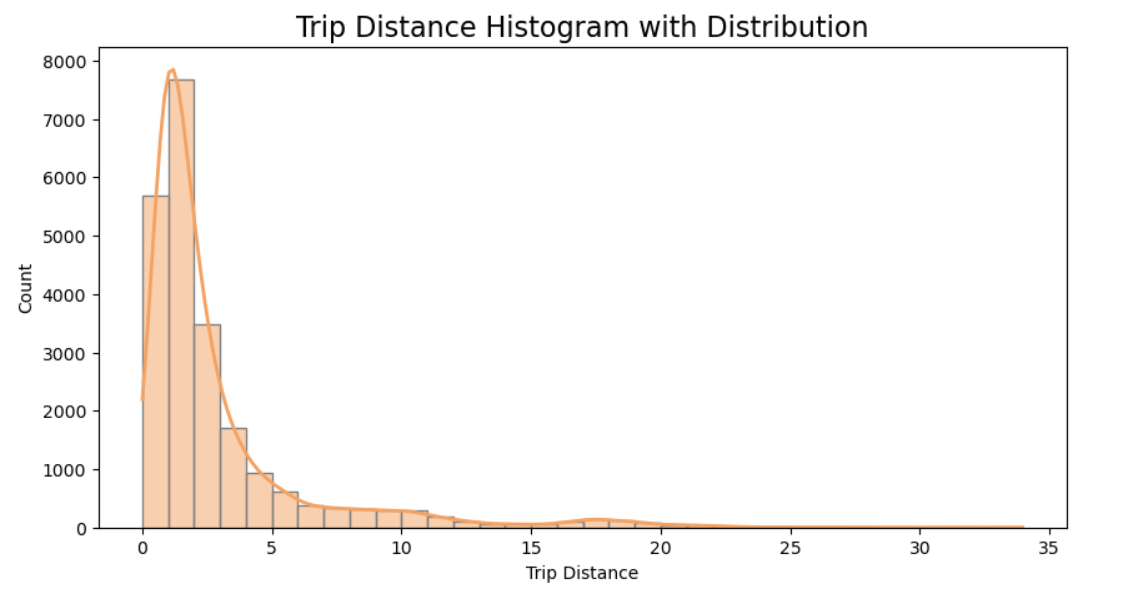
**Interpretation:** Most trips are short, but rare long trips heavily skew the distribution.



## 3.2 Histogram

* The majority of trips are between **0–2 miles**.
* Frequency drops sharply after 3 miles.
* Long trips (>20 miles) are rare but present.

**Conclusion:** Trip distance distribution is **right-skewed**, dominated by short trips.

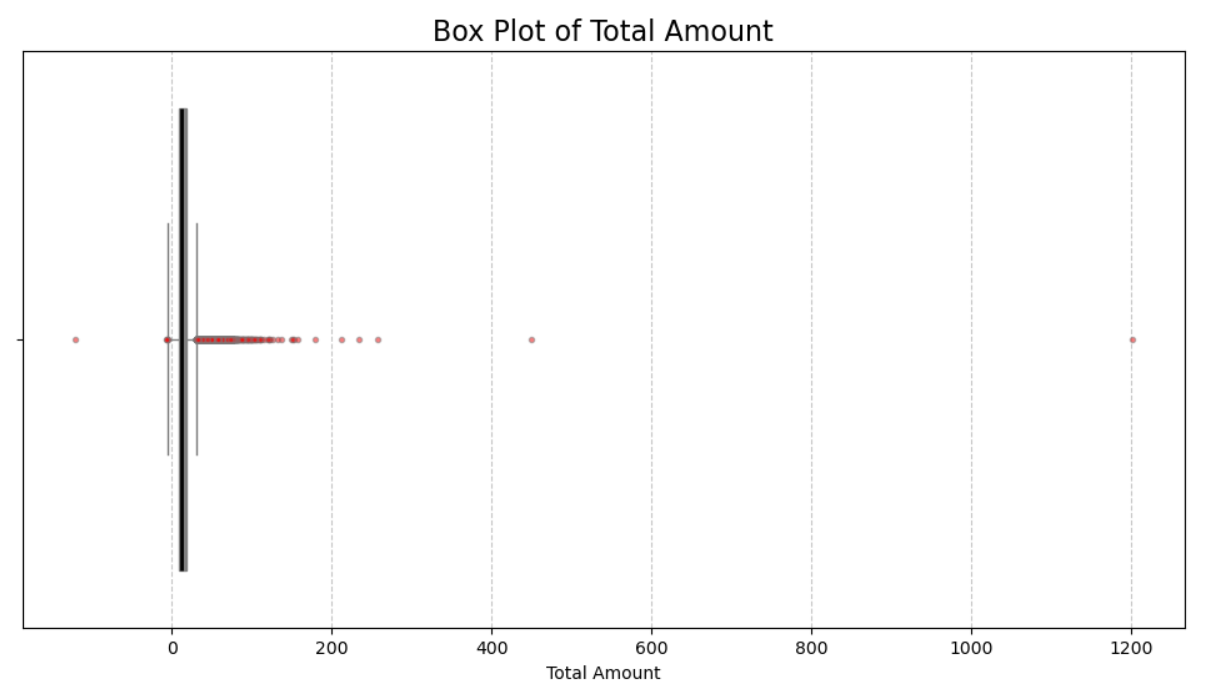


# 4. Total Fare (total\_amount) Analysis

## 4.1 Box Plot

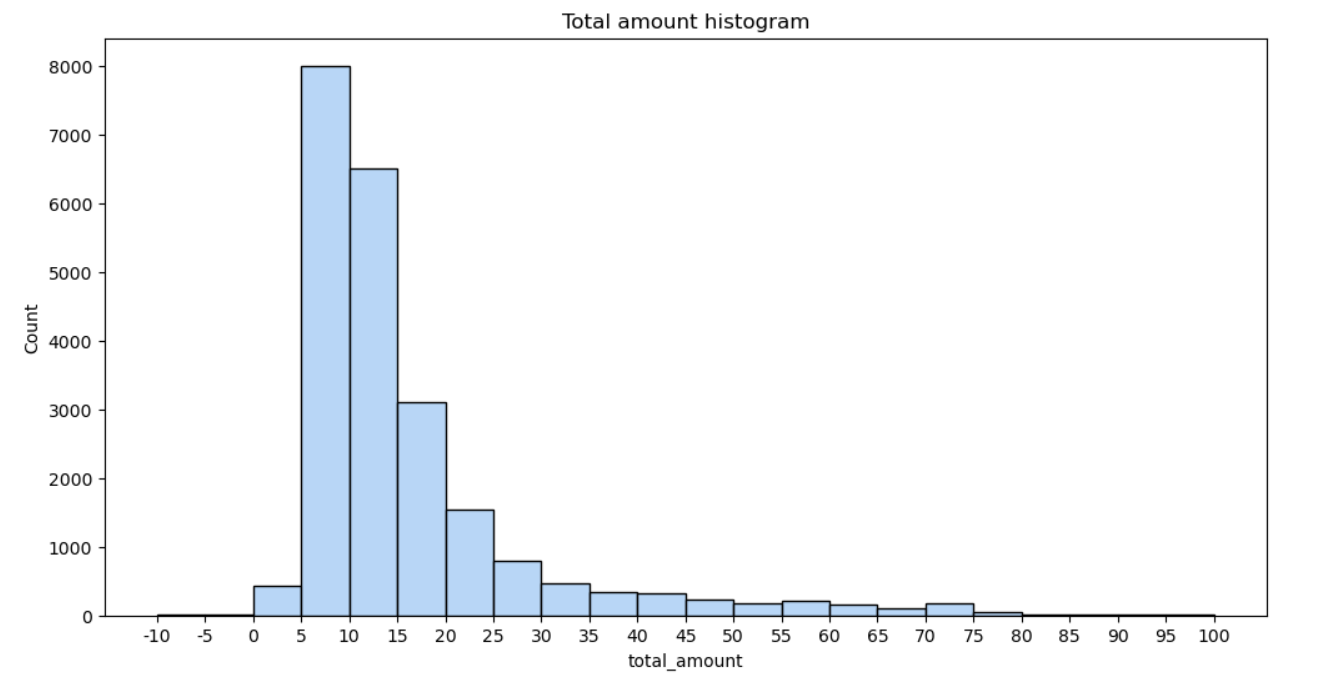
* **Median:** $10
* **IQR:** $7–$15 (middle 50% of trips)
* **Outliers:** Trips with fares exceeding $200, some up to $1,200

**Observation:** While most fares are low, a small number of expensive trips heavily influence total fare distribution.



## 4.2 Histogram

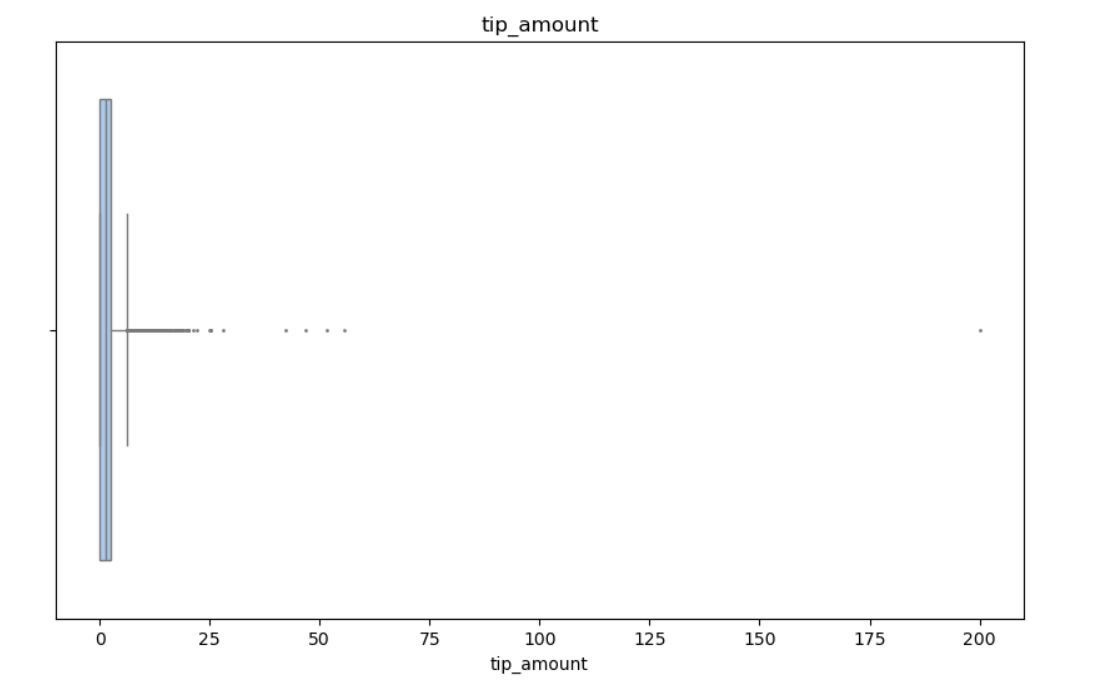
* Most trips fall between **$5–$10**.
* Very few trips have fares >$50, confirming the right-skew.



# 5. Tip Amount Analysis

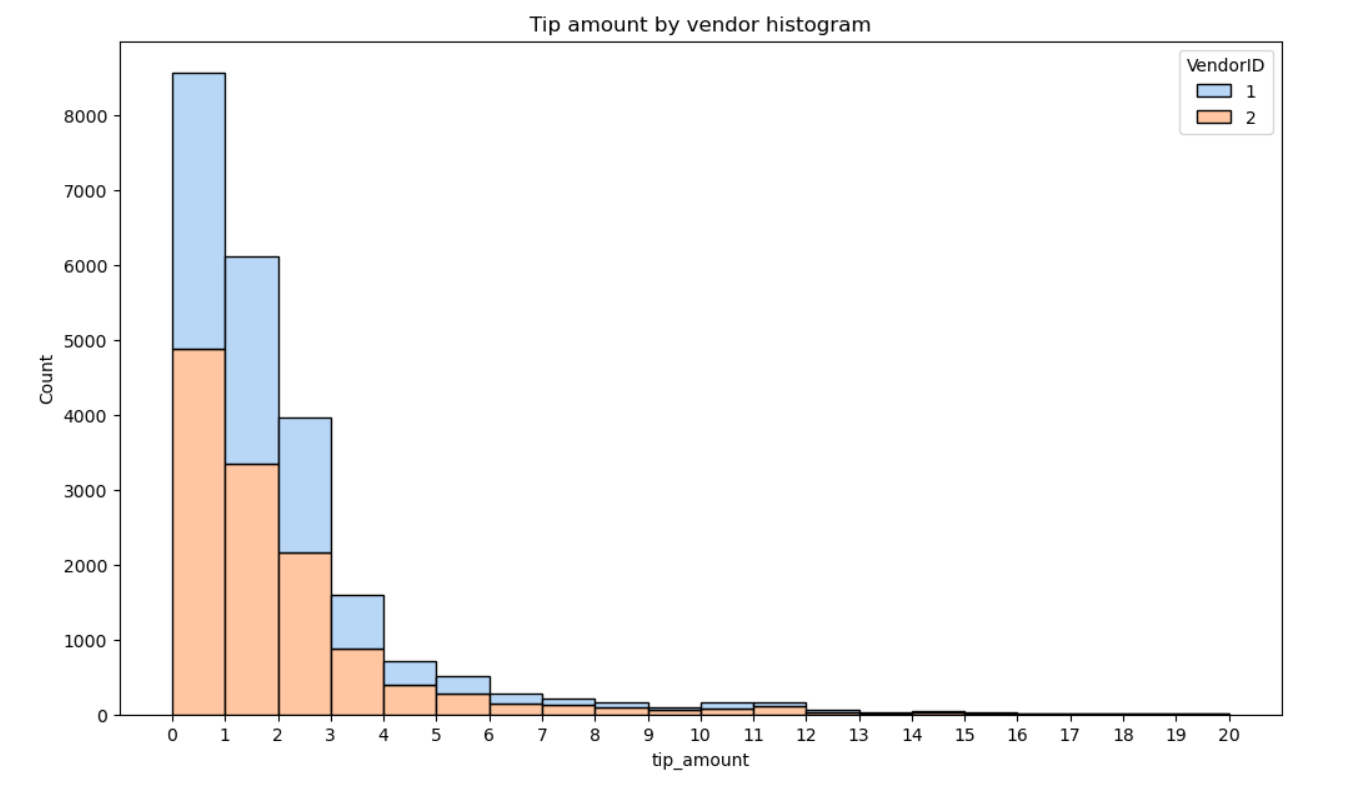
## 5.1 Overall Tip Distribution

* **Median tip:** ~$1
* **IQR:** $0–$2.5
* **Outliers:** Tips over $200 are rare but exist
* Many trips have **$0 tips**, possibly due to cash payments or no tipping.



## 5.2 Tip by Vendor

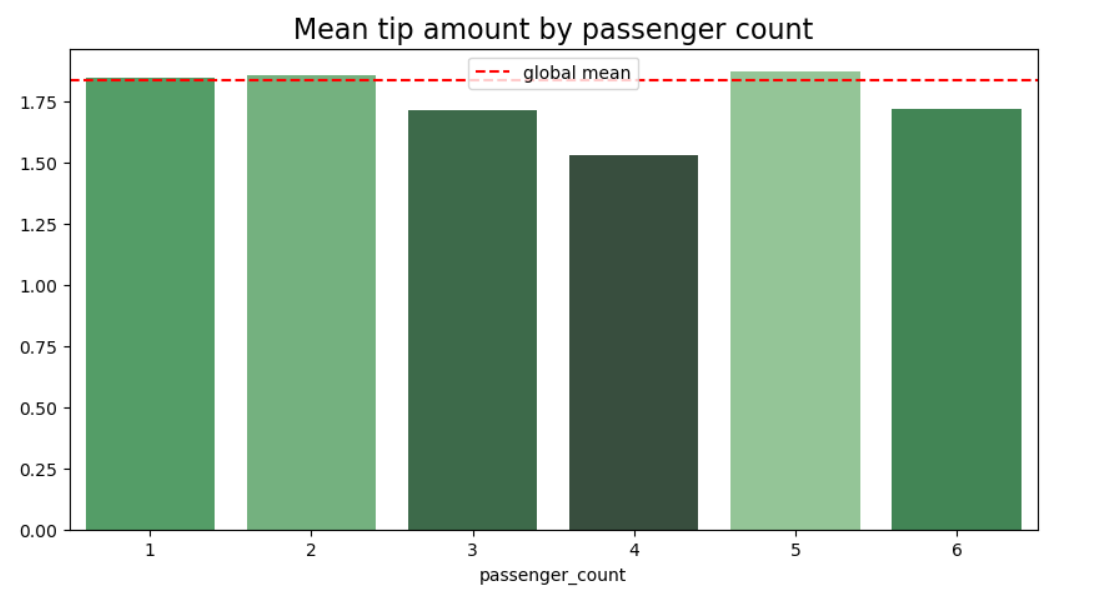
* Most common tips: $0–$5
* For tips >$10, **VendorID 2** dominates slightly, especially for $10–$12 tips.



## 5.3 Mean Tip by Passenger Count

|  |  |
| --- | --- |
| Passenger Count | Average Tip |
| 0 | $2.14 |
| 1 | $1.85 |
| 2 | $1.86 |
| 3 | $1.72 |
| 4 | $1.53 |
| 5 | $1.87 |
| 6 | $1.72 |

**Observation:**

* Trips with **zero passengers** show the highest average tip, likely due to canceled trips, data anomalies, or driver-only trips.
* Tips do not increase linearly with passenger count; mid-sized groups (3–4) tip less, whereas 5-passenger trips tip slightly more.
* 

# 6. Passenger Count Distribution

|  |  |
| --- | --- |
| Passenger Count | Number of Trips |
| 1 | 16,117 |
| 2 | 3,305 |
| 3 | 953 |
| 4 | 455 |
| 5 | 1,143 |
| 6 | 693 |
| 0 | 33 |

## Observation:

* Most trips involve **single passengers**.
* Trips with 0 passengers are rare.
* Trips with 5 passengers are slightly higher than trips with 3–4 passengers.

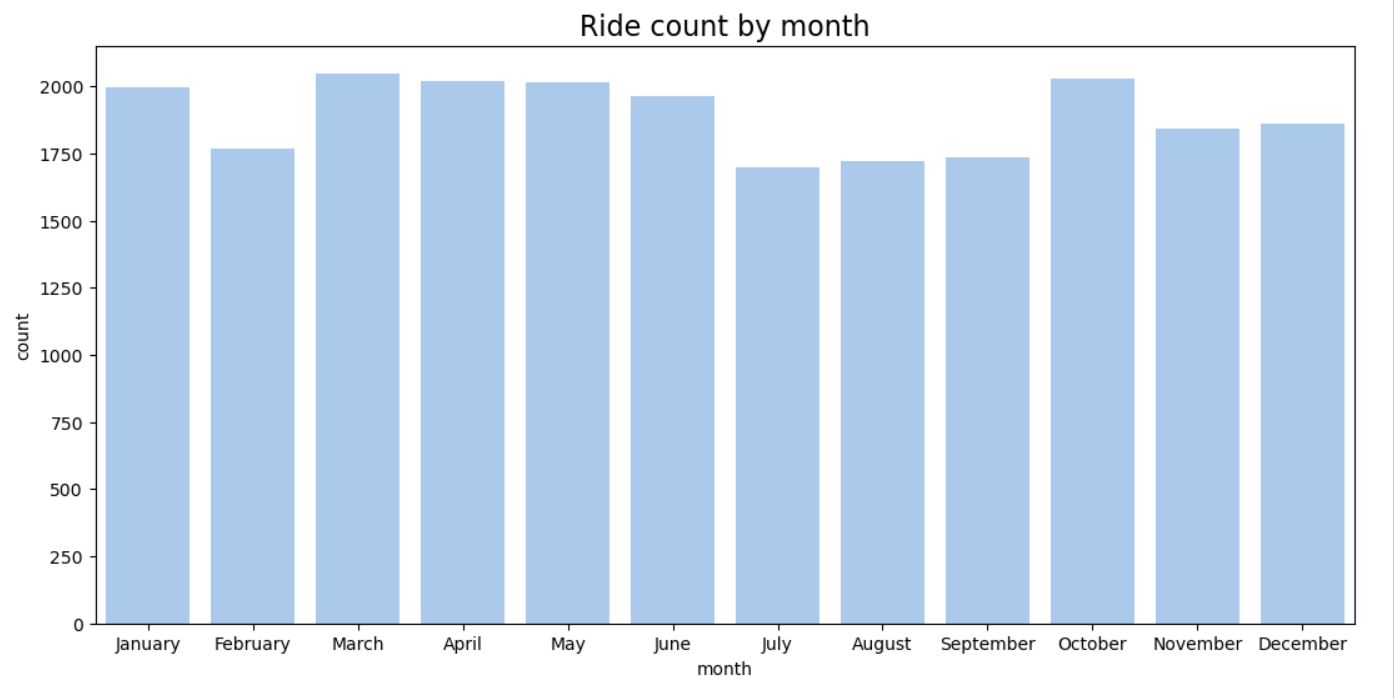
# 7. Temporal Analysis

## 7.1 Trips by Month

|  |  |
| --- | --- |
| Month | Trip Count |
| January | 1,997 |
| February | 1,769 |
| March | 2,049 |
| April | 2,019 |
| May | 2,013 |
| June | 1,964 |
| July | 1,697 |
| August | 1,724 |
| September | 1,734 |
| October | 2,027 |
| November | 1,843 |
| December | 1,863 |

**Observation:**

* March and October have the highest number of trips.
* July has the lowest, possibly due to seasonal factors.

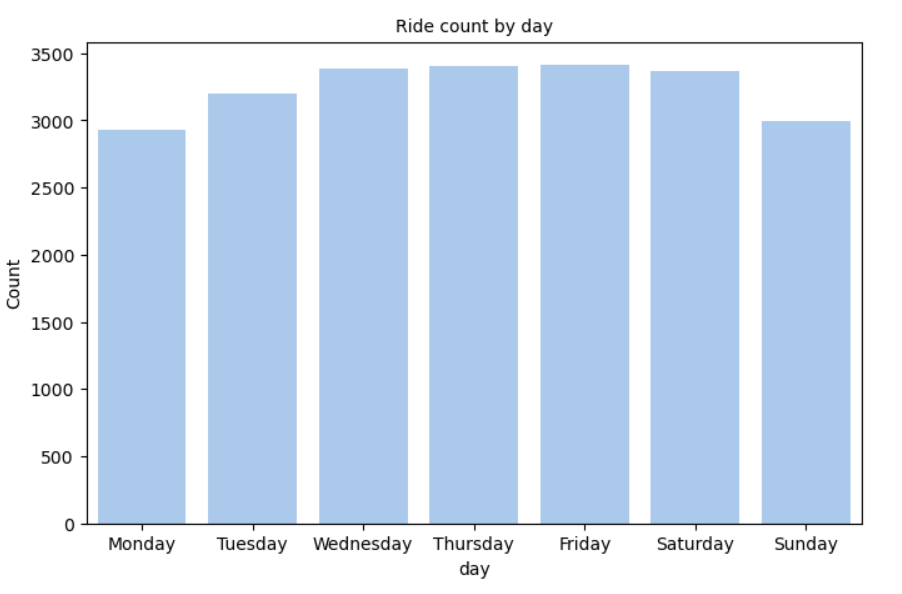


## 7.2 Trips by Day of Week

|  |  |
| --- | --- |
| Day | Trip Count |
| Friday | 3,413 |
| Thursday | 3,402 |
| Wednesday | 3,390 |
| Saturday | 3,367 |
| Tuesday | 3,198 |
| Monday | 2,931 |
| Sunday | 2,998 |

**Observation:**

* Mid-week and Friday have slightly higher trip counts.

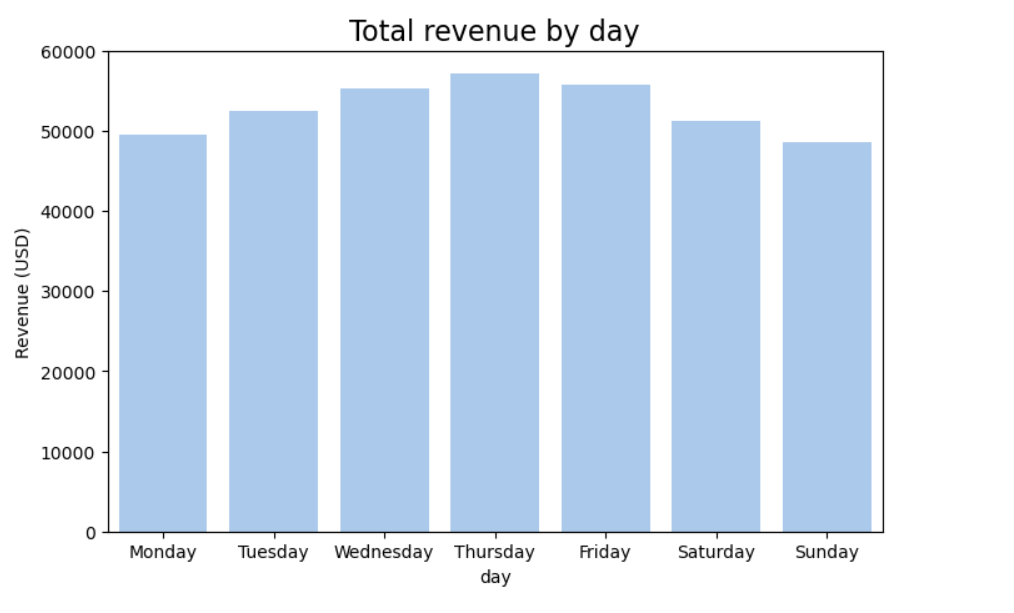


## 7.3 Revenue by Day

|  |  |
| --- | --- |
| Day | Total Revenue |
| Thursday | $57,182 |
| Friday | $55,819 |
| Wednesday | $55,310 |
| Tuesday | $52,527 |
| Saturday | $51,195 |
| Monday | $49,574 |
| Sunday | $48,624 |

**Observation:**

* Revenue roughly follows trip counts, with Thursday and Friday generating the most revenue.

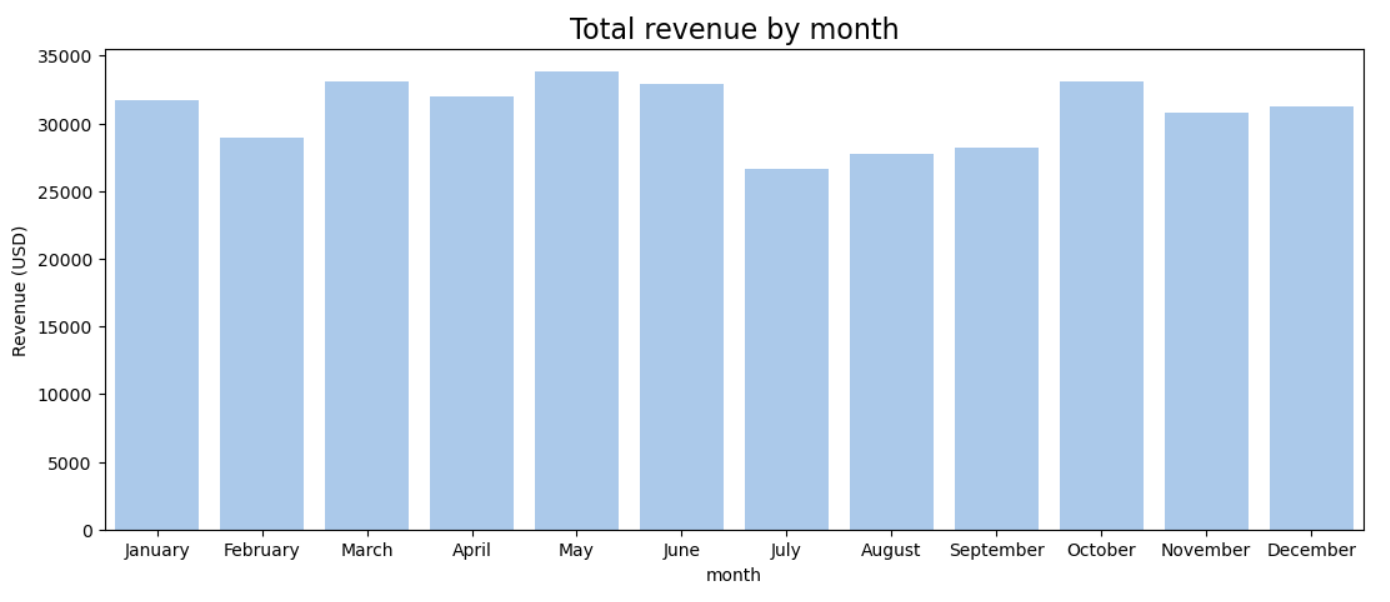


## 7.4 Revenue by Month

|  |  |
| --- | --- |
| Month | Total Revenue ($) |
| January | 31,735.25 |
| February | 28,937.89 |
| March | 33,085.89 |
| April | 32,012.54 |
| May | 33,828.58 |
| June | 32,920.52 |
| July | 26,617.64 |
| August | 27,759.56 |
| September | 28,206.38 |
| October | 33,065.83 |
| November | 30,800.44 |
| December | 31,261.57 |

**Observation:**

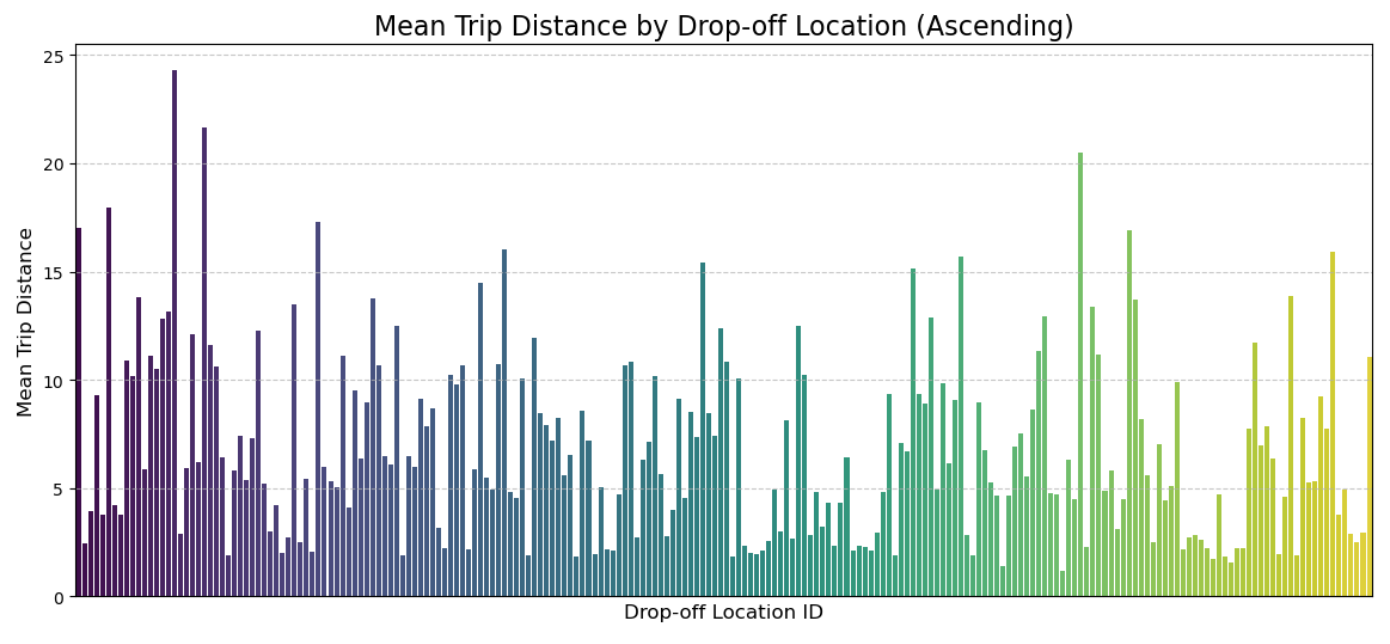
* Revenue patterns are fairly consistent with trip counts, but higher fares may affect totals.



# 8. Drop-off Locations Analysis

* **Unique Drop-off IDs:** 216
* Mean trip distance by drop-off location ranges from **1.2 miles** to **24.3 miles**.
* Distribution of trips: A few high-traffic locations dominate, likely near tourist attractions or transport hubs.
* Not all location IDs are consecutive; 49 IDs are missing, which can be handled for plotting purposes.

**Conclusion:** Drop-off locations are **geographically diverse**, but most traffic is concentrated at a few locations.



# 9. Trip Duration Analysis

* Calculated as **tpep\_dropoff\_datetime − tpep\_pickup\_datetime**.
* Observed durations match expectations: short trips have short durations, long-distance trips have longer durations.
* Can be used in future models to predict trip efficiency and fare estimation.

# 10. Key Insights & Business Implications

1. **Short Trips Dominate:** Most taxi trips are under 5 miles, with occasional long-distance trips.
2. **Fare & Tip Skew:** Total fare and tip amounts are heavily right-skewed due to outliers, which should be considered in predictive models.
3. **Passenger Trends:** Single passengers dominate; zero-passenger trips are anomalies.
4. **Temporal Patterns:** Revenue and trip volume peak mid-week and in certain months.
5. **Drop-off Location Concentration:** A few locations get most of the traffic; this may guide driver allocation or pricing strategy.
6. **Trip Duration:** Derived durations can help analyze efficiency and predict travel time.

# 11. Recommendations for Further Analysis

* Investigate **trips with zero passengers or zero distance** to understand anomalies.
* Use **trip duration, distance, and fare** for predictive modeling (fare estimation or demand prediction).
* Analyze **tips and payment types** to optimize customer satisfaction strategies.
* Explore **geospatial analysis** if coordinates are available to improve driver allocation.

# 12. Conclusion

Through comprehensive EDA:

* We identified key trends, outliers, and temporal patterns.
* Visualizations revealed actionable insights for revenue optimization, operational efficiency, and passenger behavior.
* The dataset is clean, and additional features (trip duration, month, day) were derived to enable further analysis.

This analysis provides a solid foundation for predictive modeling, strategic planning, and operational decision-making.