Report: Optimising NYC Taxi Operations

Include your visualisations, analysis, results, insights, and outcomes. Explain your methodology and approach to the tasks. Add your conclusions to the sections.

Contents

1. Data Preparation 2

2. Data Cleaning 2

2.1. Fixing Columns 2

2.2. Handling Missing Values 2

2.3. Handling Outliers and Standardising Values 3

3. Exploratory Data Analysis 4

3.1. General EDA: Finding Patterns and Trends 4

3.2. Detailed EDA: Insights and Strategies 13

4. Conclusions 27

4.1. Final Insights and Recommendations 27

1. Data Preparation

1.1. Loading the dataset

Sampled the data to ensure manageable processing and **combined files** into a single DataFrame for analysis.

1.1.1. Sample the data and combine the files

I first extracted a sample of **500,000** records from each monthly Parquet file as instructed. Then, I further refined the sample size, ensuring that the final combined DataFrame comprised around **1.89 million rows**.

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

Resolved index inconsistencies, unique trip identifiers. Cleaned column names by removing extra spaces and standardizing formatting for consistency.

2.1.2. Combine the two airport_fee columns

Combined the airport fee columns into a single field to address inconsistencies in naming across monthly files. To preserve all data, I created a new column, airport_fee_combined, by selecting the maximum value from airport_fee and Airport_fee for each row. Once merged, I removed the original columns to eliminate redundancy.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

Identified missing values across fields, addressing gaps in passenger_count, RatecodeID, and congestion_surcharge.

2.2.2. Handling missing values in passenger_count

To handle missing values in the passenger_count column, I filled null entries using the mode (most frequent value). This method preserves the data distribution without introducing bias.

2.2.3. Handle missing values in RatecodeID

To impute missing values in the RatecodeID column, I used the mode (most frequent value). Since RatecodeID is categorical, this approach ensures consistency by preserving the most common pattern in the dataset while avoiding bias from rare or extreme values.

2.2.4. Impute NaN in congestion_surcharge

Filled missing values in the congestion_surcharge column using median imputation. By replacing null entries with the median of non-null values, this approach minimizes the impact of extreme outliers and maintains the integrity of the column's distribution.

2.3. Handling Outliers and Standardising Values

2.3.1. Check outliers in payment type, trip distance and tip amount columns

Identified outliers in trip distance, tip amount, and payment type using percentile-based filtering.

- **Payment Type:** Entries with payment_type equal to 0 (an invalid code) were removed.
- **Trip Distance:**
    - Trips with distances < 0.1 miles but fares exceeding $300 were excluded.
    - Trips longer than 250 miles were considered extreme outliers and removed.
    - Trips showing 0 distance and fare, yet having different pickup and dropoff locations, were treated as invalid and eliminated.
- **Tip Amount:**
    - No filtering was applied to zero values, as tipping is optional.
    - Large tips were handled through min-max standardization, scaling values between 0 and 1 to mitigate the effect of extreme tips.

3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

3.1.1. Classify variables into categorical and numerical

Categorical Variables

VendorID

RatecodeID

PULocationID

DOLocationID

payment_type

Numerical Variables

passenger_count

trip_distance

pickup_hour

trip_duration

fare_amount

extra

mta_tax

tip_amount

tolls_amount

improvement_surcharge

total_amount

congestion_surcharge

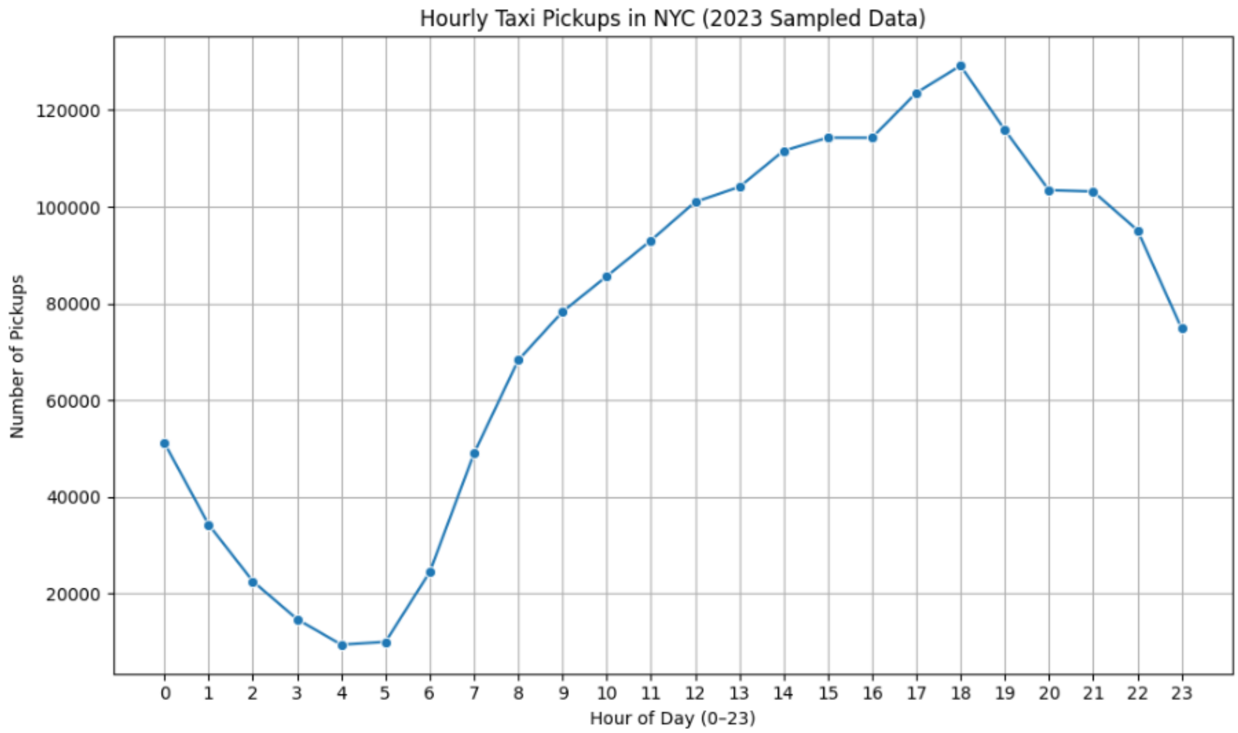
airport_fee

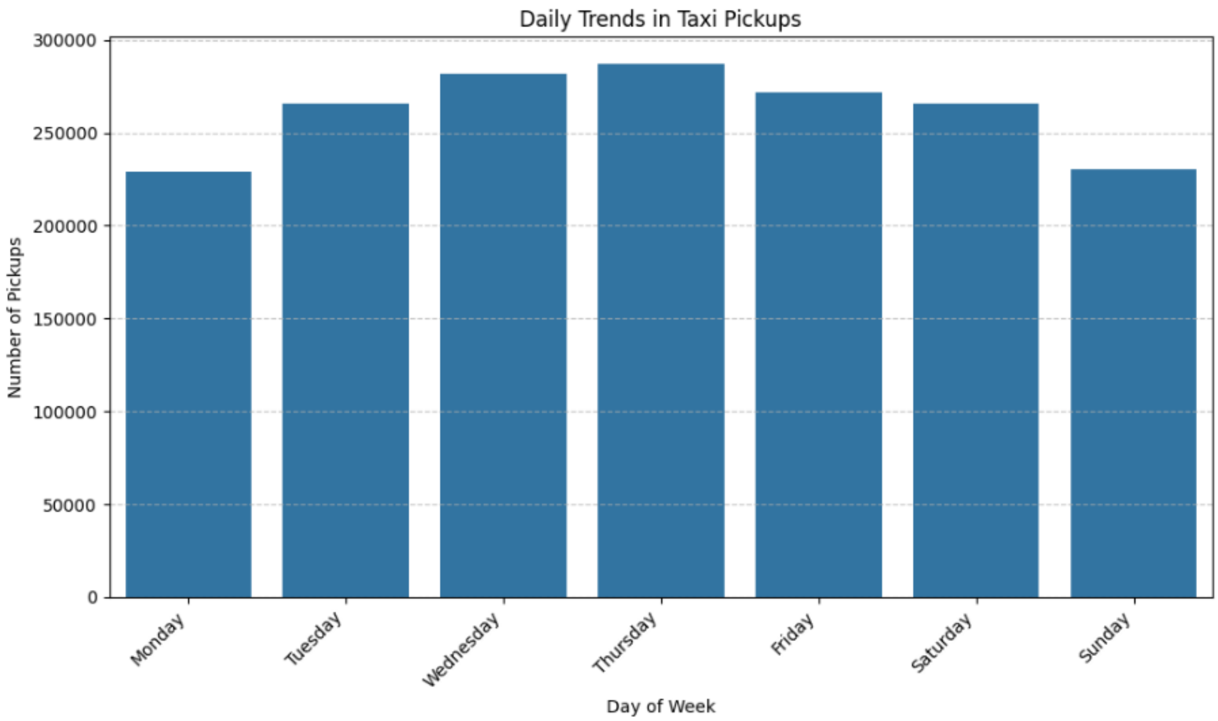
Timestamp Variables

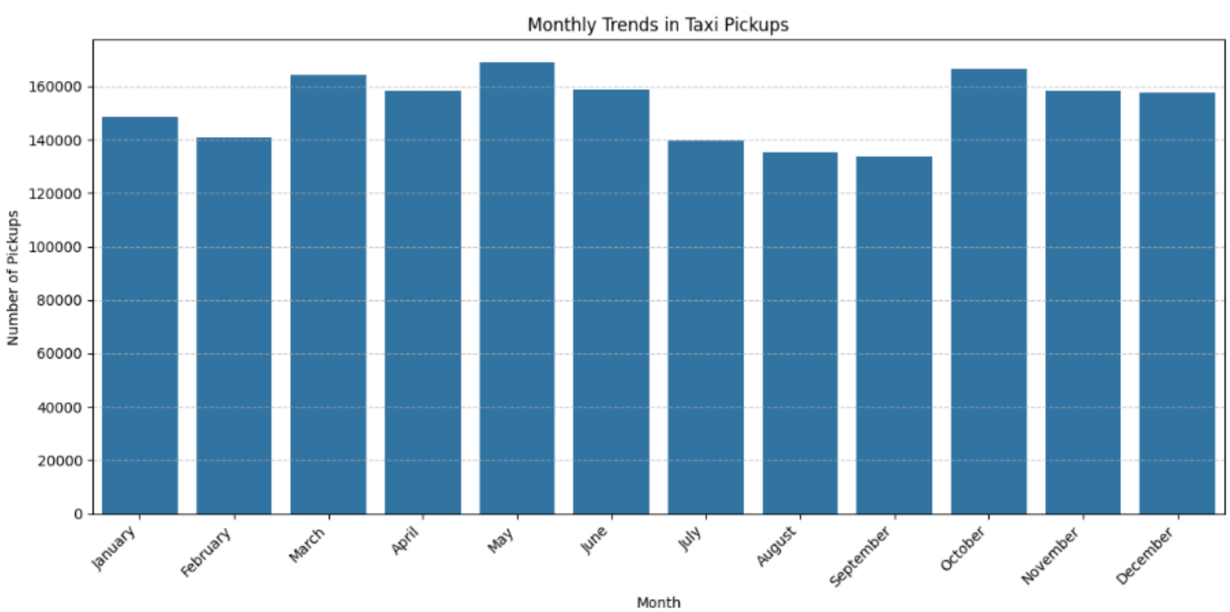
tpep_pickup_datetime

tpep_dropoff_datetime

3.1.2. Analyse the distribution of taxi pickups by hours, days of the week, and months







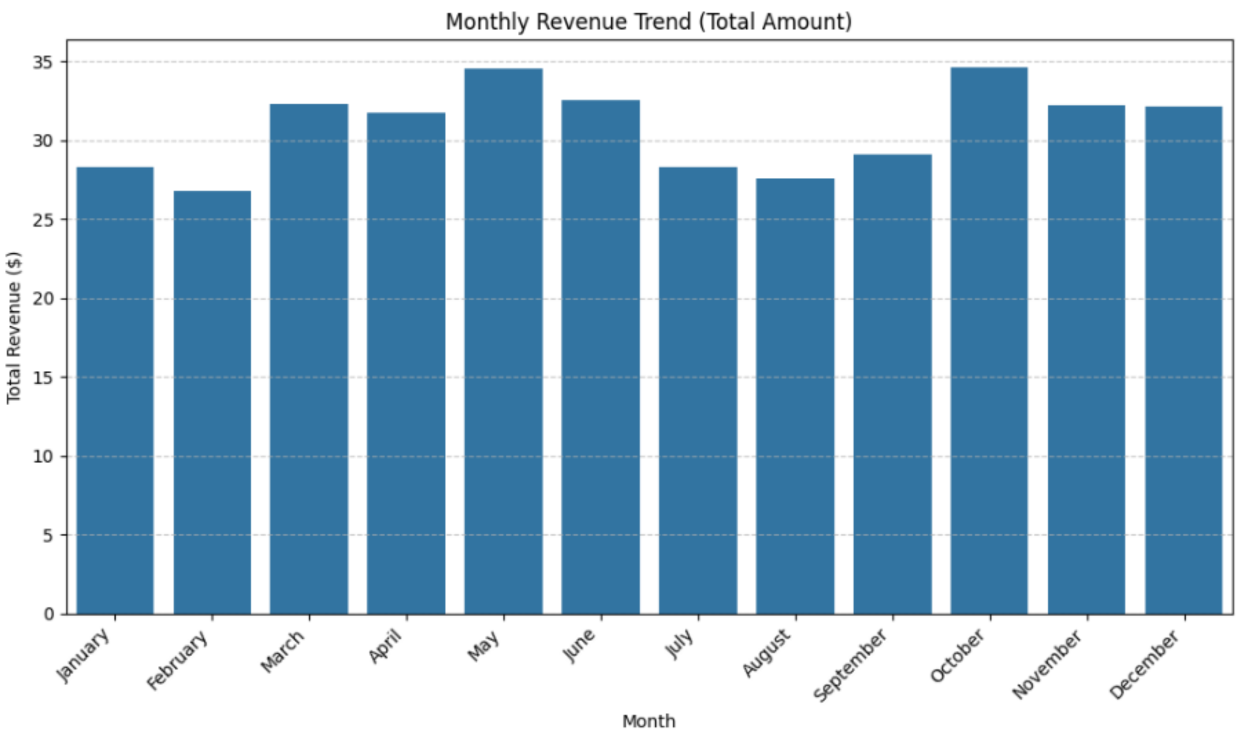
3.1.3. Filter out the zero/negative values in fares, distance and tips

To maintain data integrity, I removed records where:

- fare_amount or total_amount was zero, as these likely represented invalid or canceled trips.
- trip_distance was zero despite different pickup and dropoff locations, indicating inconsistencies.

However, I kept entries with zero tip_amount, since tipping is optional and many valid trips had no recorded tip. These entries still contained a valid total_amount, confirming their legitimacy. This filtering ensured a cleaner dataset while preserving real-world behaviors such as no tipping.

3.1.4. Analyse the monthly revenue trends

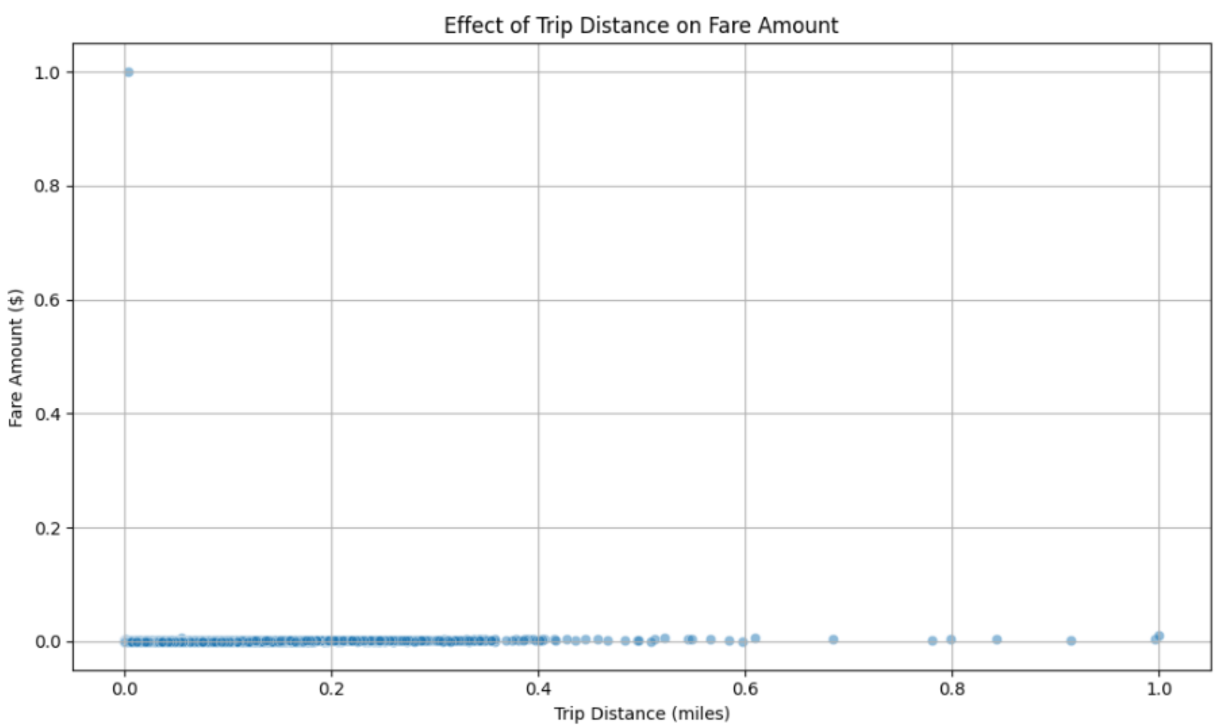


3.1.5. Find the proportion of each quarter's revenue in the yearly revenue

	total_amount
pickup_quarter	
2022Q4	0.00
2023Q1	23.61
2023Q2	26.68
2023Q3	22.97
2023Q4	26.74

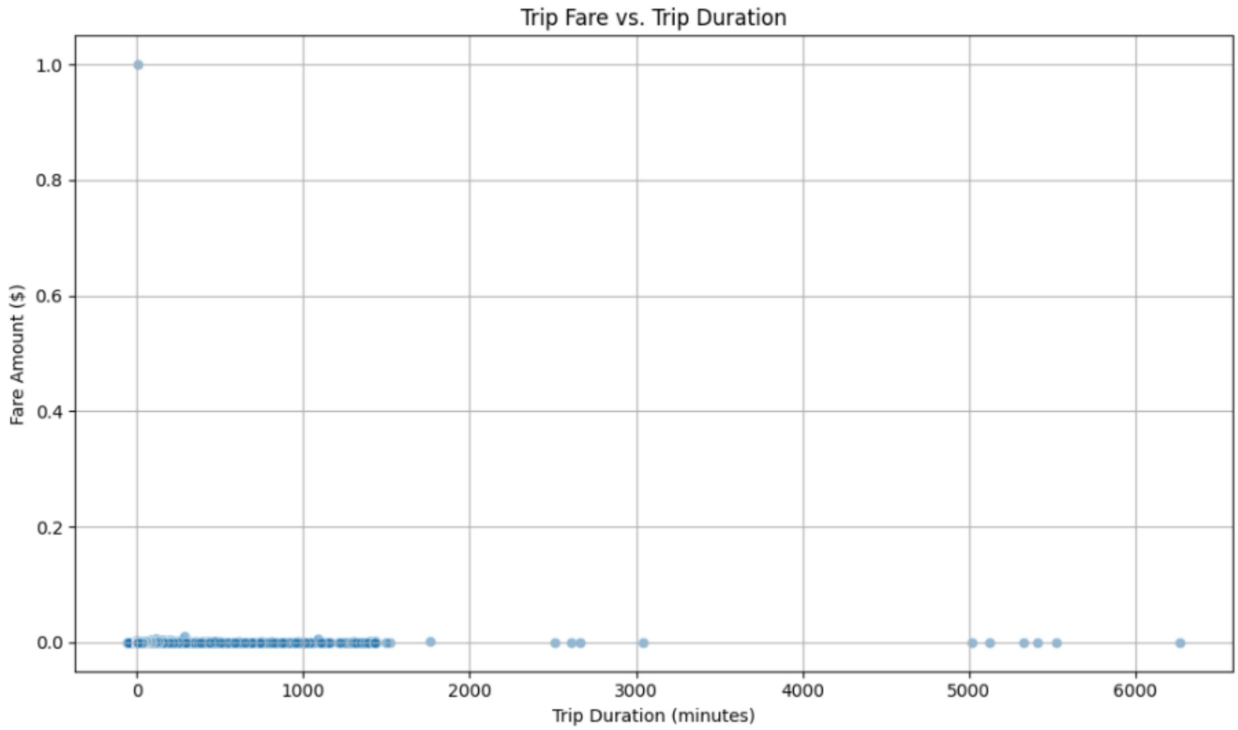
dtype: float64

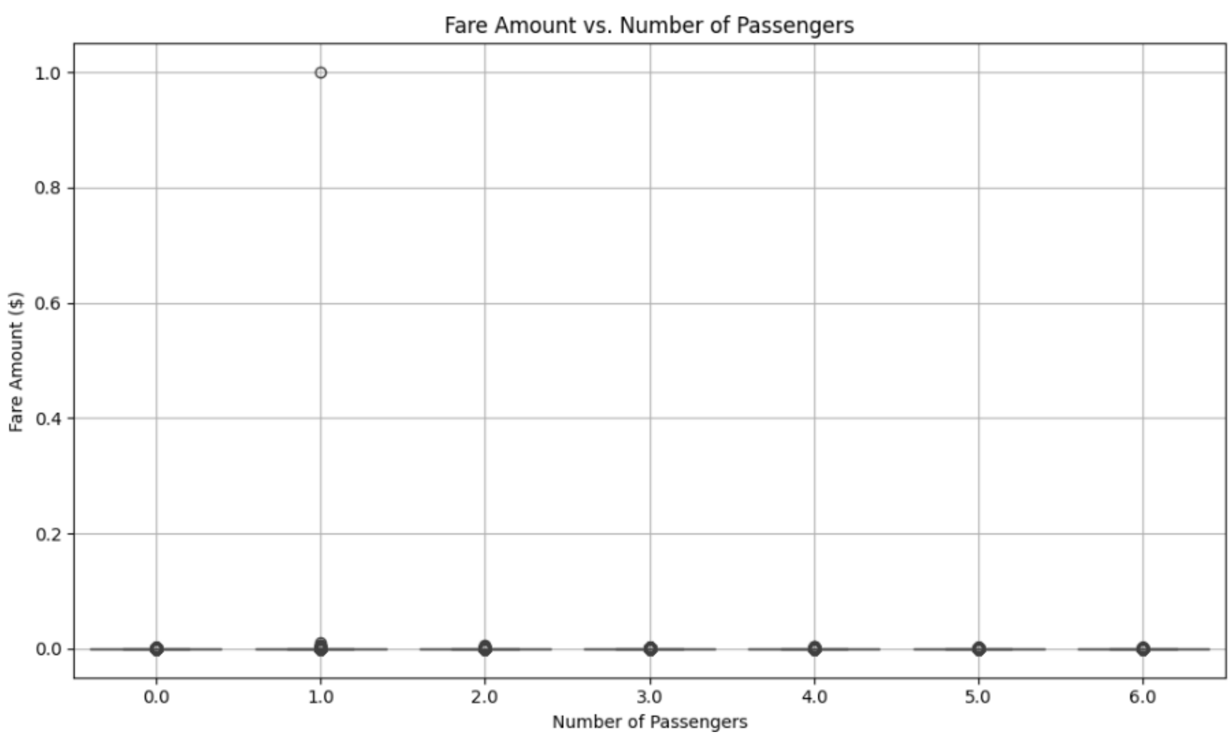
3.1.6. Analyse and visualise the relationship between distance and fare amount

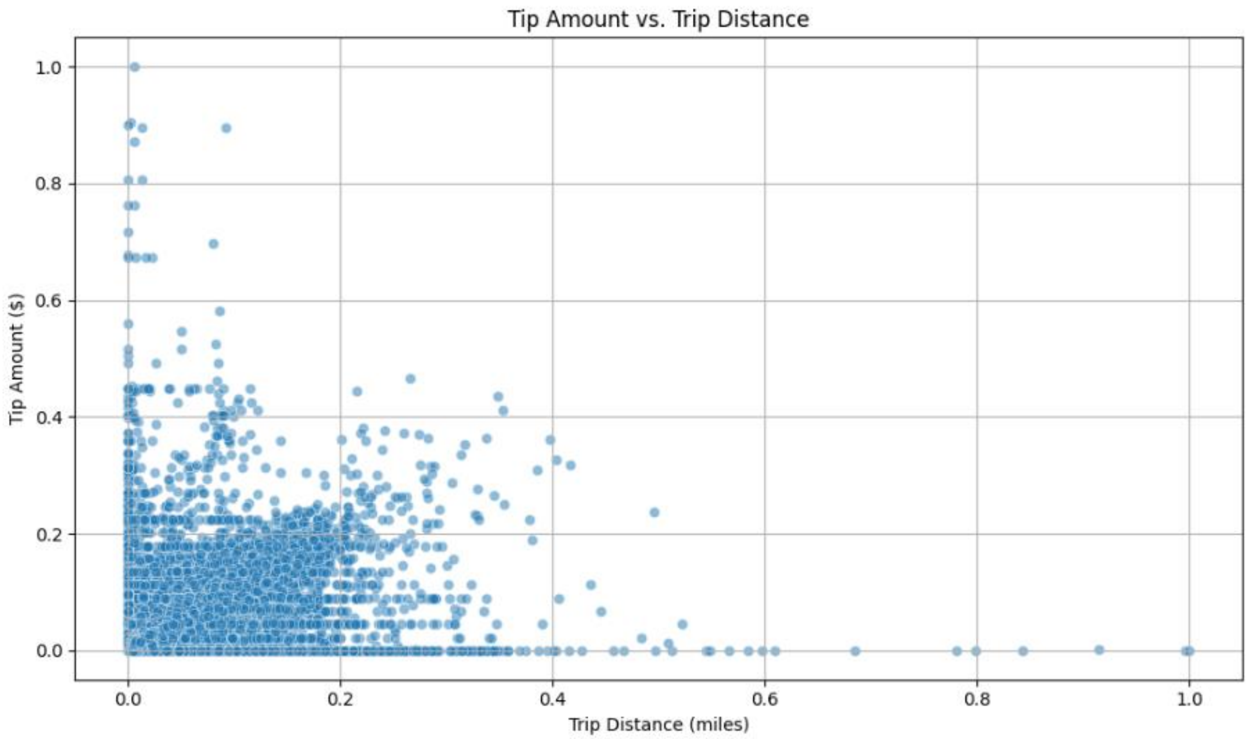



```
Correlation between trip distance and fare amount (excluding zero-distance trips): 0.16
```

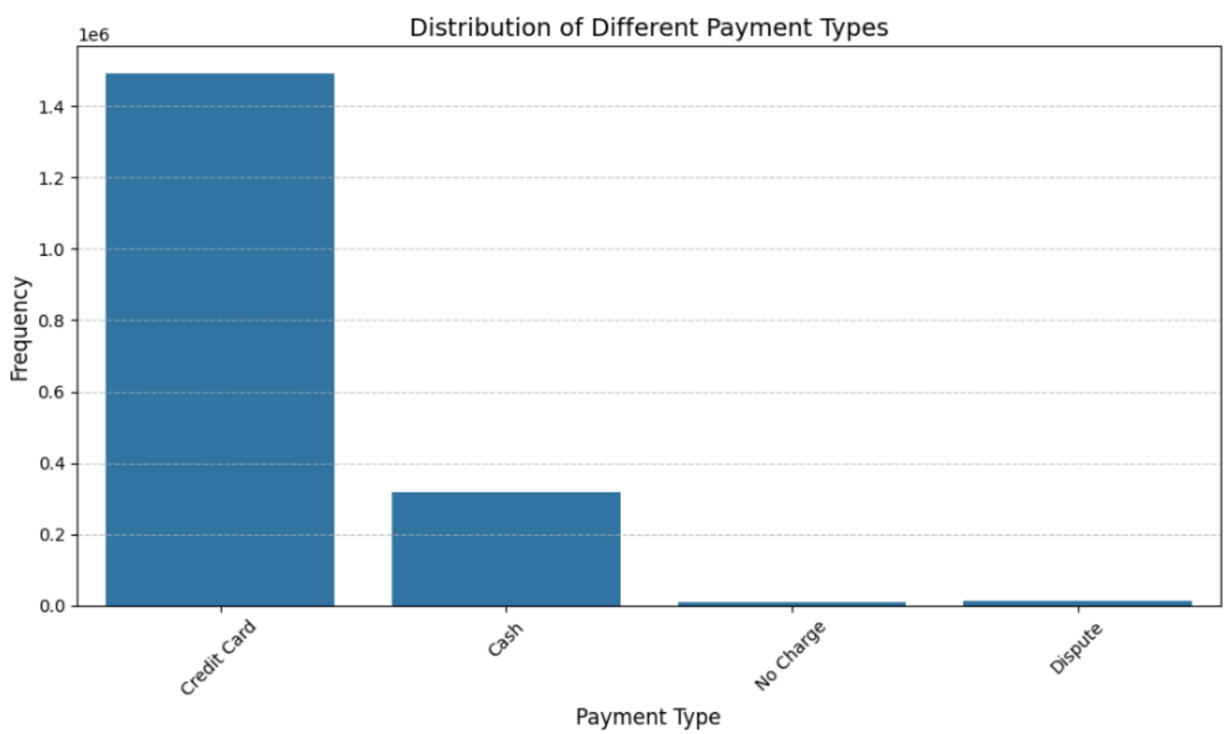
3.1.7. Analyse the relationship between fare/tips and trips/passengers


Trip Fare vs. Trip Duration
Fare Amount ($)
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
6000
Trip Duration (minutes)


Fare Amount vs. Number of Passengers
Fare Amount ($)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
Number of Passengers

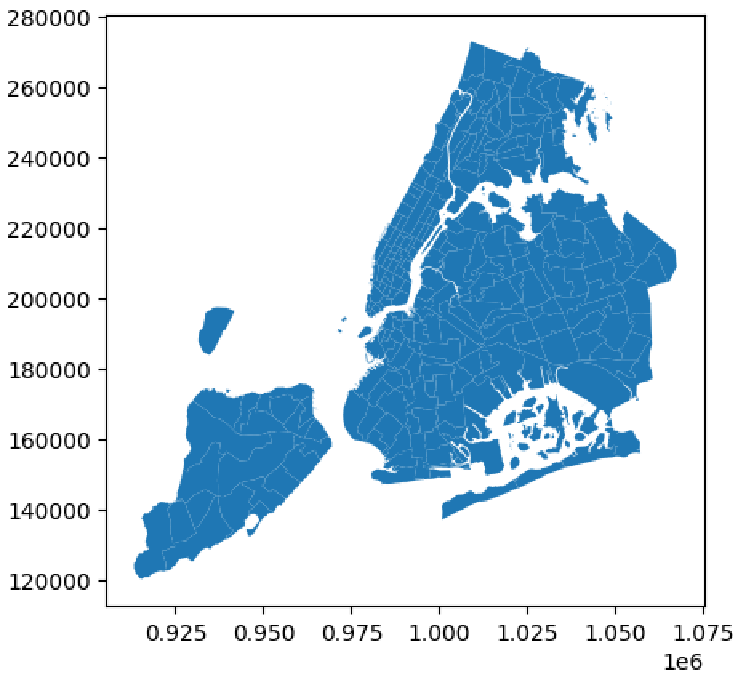


3.1.8. Analyse the distribution of different payment types



3.1.9. Load the taxi zones shapefile and display it

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry
0	1	0.116357	0.000782	Newark Airport	1	EWR	POLYGON ((933100.918 192536.086, 933091.011 19...
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 103343...
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 2...
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...
4	5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144...



3.1.10. Merge the zone data with trips data

Merged zones data with trip data using the locationID and PULocationID columns.

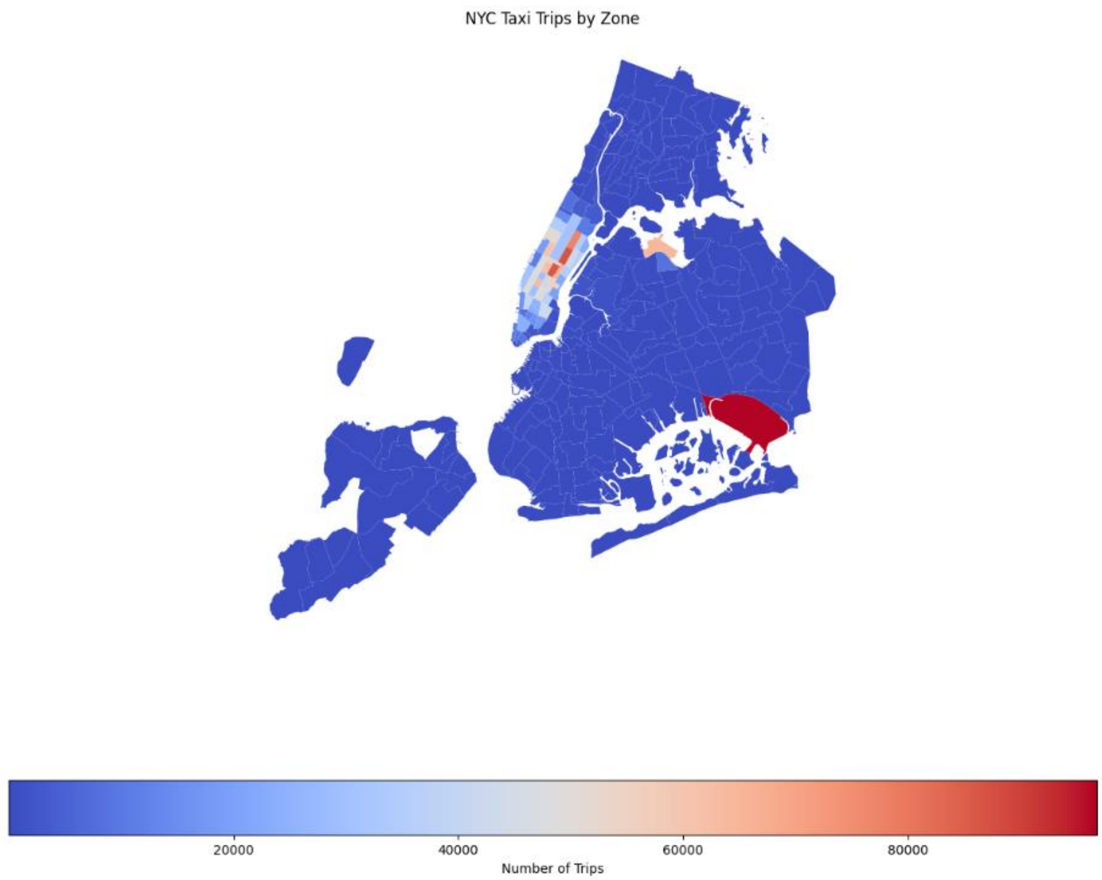
3.1.11. Find the number of trips for each zone/location ID

	PULocationID	num_trips
0	1	214
1	2	2
2	3	40
3	4	1861
4	5	13

3.1.12. Add the number of trips for each zone to the zones dataframe

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry	PULocationID	num_trips
0	1	0.116357	0.000782	Newark Airport	1	EWR	POLYGON ((933100.918 192536.086, 933091.011 19...	1.0	214.0
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 103343...	2.0	2.0
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 2...	3.0	40.0
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...	4.0	1861.0
4	5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144...	5.0	13.0

3.1.13. Plot a map of the zones showing number of trips



3.1.14. Conclude with results

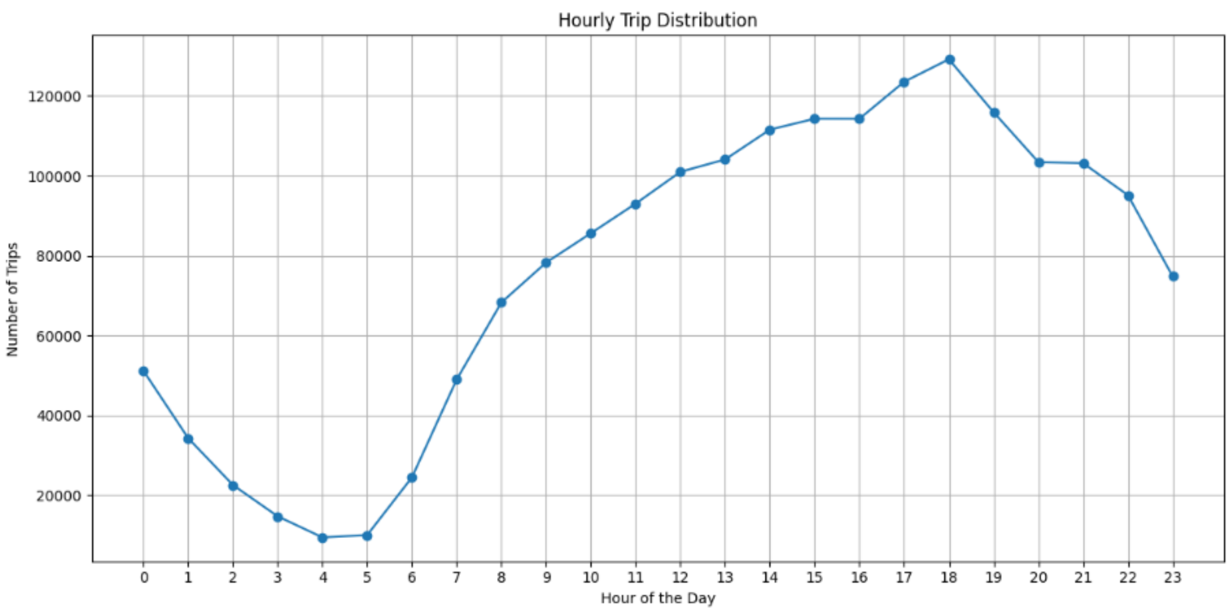
- Fare primarily depends on distance, showing a strong correlation. Weekday rush hours are peak times, while weekends see more late-night trips.
- Airport and Midtown zones have the highest trip density.
- Most rides carry 1–2 passengers, with credit cards as the preferred payment.
- Q3 is the busiest.

3.2. Detailed EDA: Insights and Strategies

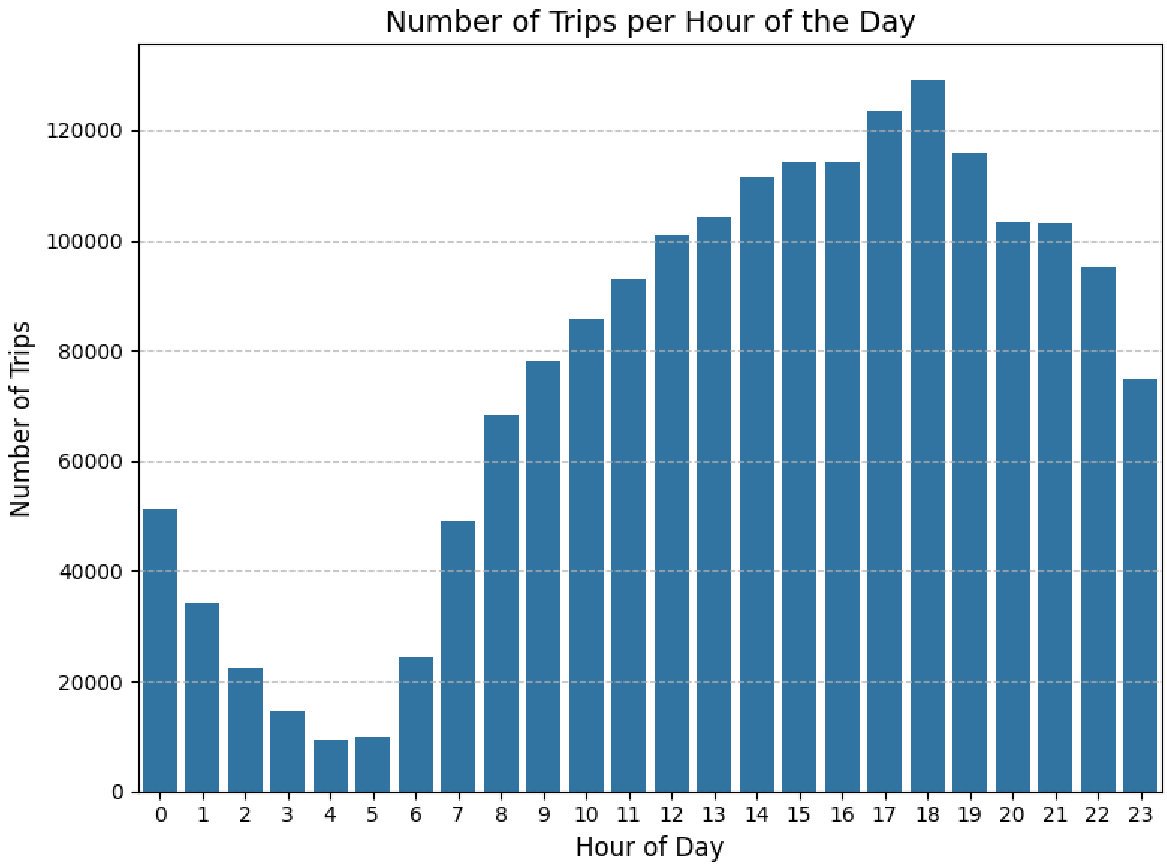
3.2.1. Identify slow routes by comparing average speeds on different routes

	PULocationID	DOLocationID	pickup_hour	avg_speed_mph
102294	232	65	13	0.000026
114929	243	264	17	0.000038
61252	142	142	5	0.000116
120428	258	258	1	0.000128
33393	100	7	8	0.000193
6451	40	65	21	0.000229
39490	113	235	22	0.000235
89226	194	194	16	0.000239
95261	226	145	18	0.000253
9705	45	45	10	0.000290

3.2.2. Calculate the hourly number of trips and identify the busy hours




```
Busiest hour: 18
Number of trips during busiest hour: 129190
```

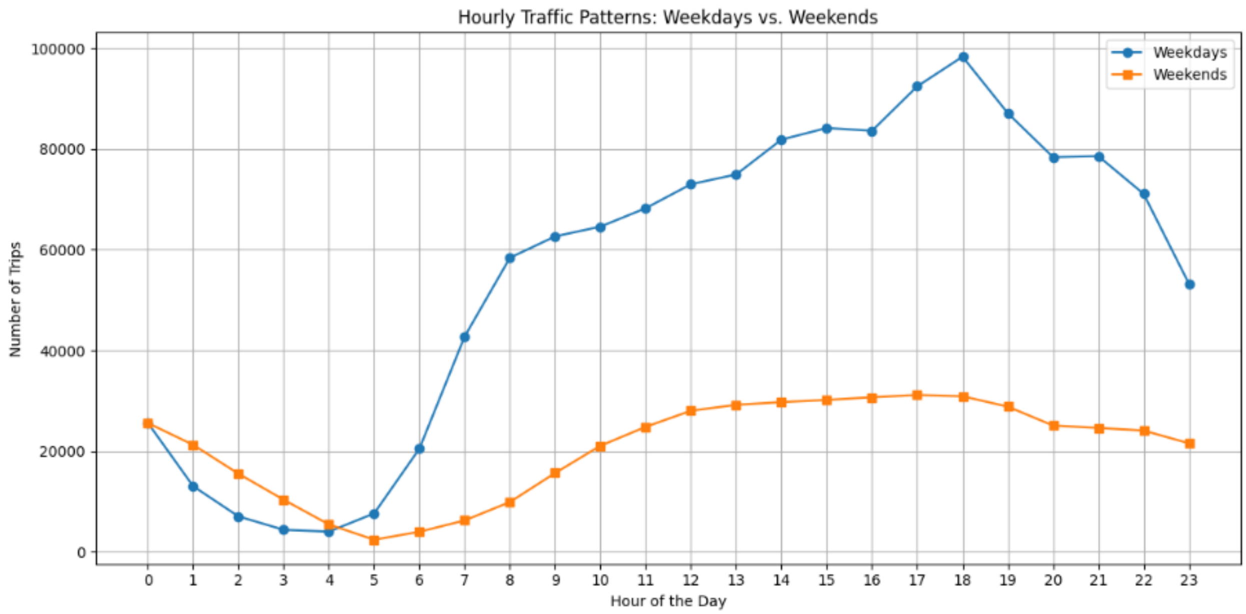


3.2.3. Scale up the number of trips from above to find the actual number of trips

	count
pickup_hour	
18	129190
17	123563
19	115920
15	114301
16	114289

dtype: int64

3.2.4. Compare hourly traffic on weekdays and weekends



3.2.5. Identify the top 10 zones with high hourly pickups and drops

```
Top 10 Pickup Zones:
```

	LocationID	Pickup_Trips	zone
0	132	96827	JFK Airport
1	237	86905	Upper East Side South
2	161	85948	Midtown Center
3	236	77517	Upper East Side North
4	162	65634	Midtown East
5	138	64177	LaGuardia Airport
6	186	63471	Penn Station/Madison Sq West
7	230	61315	Times Sq/Theatre District
8	142	60887	Lincoln Square East
9	170	54493	Murray Hill

```
Top 10 Dropoff Zones:
```

	LocationID	Dropoff_Trips
0	236	81269
1	237	77558
2	161	71647
3	230	56398
4	170	54314
5	162	52248
6	142	51494
7	239	51260
8	141	48449
9	68	46352

3.2.6. Find the ratio of pickups and dropoffs in each zone

	pickup_dropoff_ratio
zone	
East Elmhurst	8.320717
JFK Airport	4.617626
LaGuardia Airport	2.884489
Penn Station/Madison Sq West	1.582187
Central Park	1.374760
Greenwich Village South	1.374743
West Village	1.326222
Midtown East	1.256201
Midtown Center	1.199604
Garment District	1.191880

dtype: float64

zone	pickup_dropoff_ratio
West Brighton	0.000000
Broad Channel	0.000000
Oakwood	0.000000
Freshkills Park	0.000000
Breezy Point/Fort Tilden/Riis Beach	0.025641
Stapleton	0.029412
Windsor Terrace	0.038259
Newark Airport	0.040233
Grymes Hill/Clifton	0.043478
Ridgewood	0.052525

dtype: float64

3.2.7. Identify the top zones with high traffic during night hours

pickup_zone	PULocationID
East Village	15339
JFK Airport	13399
West Village	12352
Clinton East	9797
Lower East Side	9535
Greenwich Village South	8720
Times Sq/Theatre District	7776
Penn Station/Madison Sq West	6233
Midtown South	5962
LaGuardia Airport	5947

dtype: int64

dropoff_zone	DOLocationID
East Village	8239
Clinton East	6641
Murray Hill	6085
Gramercy	5627
East Chelsea	5551
Lenox Hill West	5122
West Village	4896
Yorkville West	4878
Lower East Side	4321
Times Sq/Theatre District	4297

dtype: int64

3.2.8. Find the revenue share for nighttime and daytime hours

Nighttime Revenue Share: 12.06%

Daytime Revenue Share: 87.94%

3.2.9. For the different passenger counts, find the average fare per mile per passenger

passenger_count	fare_per_mile_per_passenger
1.0	0.024175
2.0	0.013309
3.0	0.008308
4.0	0.008498
5.0	0.003936
6.0	0.003173

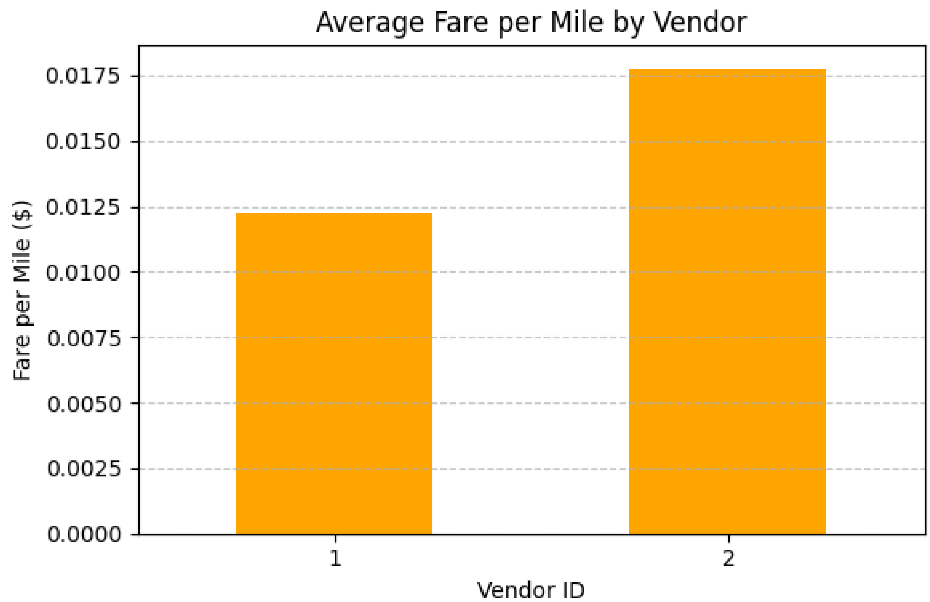
3.2.10. Find the average fare per mile by hours of the day and by days of the week

	fare_per_mile
day_of_week	
Monday	0.02
Tuesday	0.03
Wednesday	0.02
Thursday	0.02
Friday	0.02
Saturday	0.02
Sunday	0.03

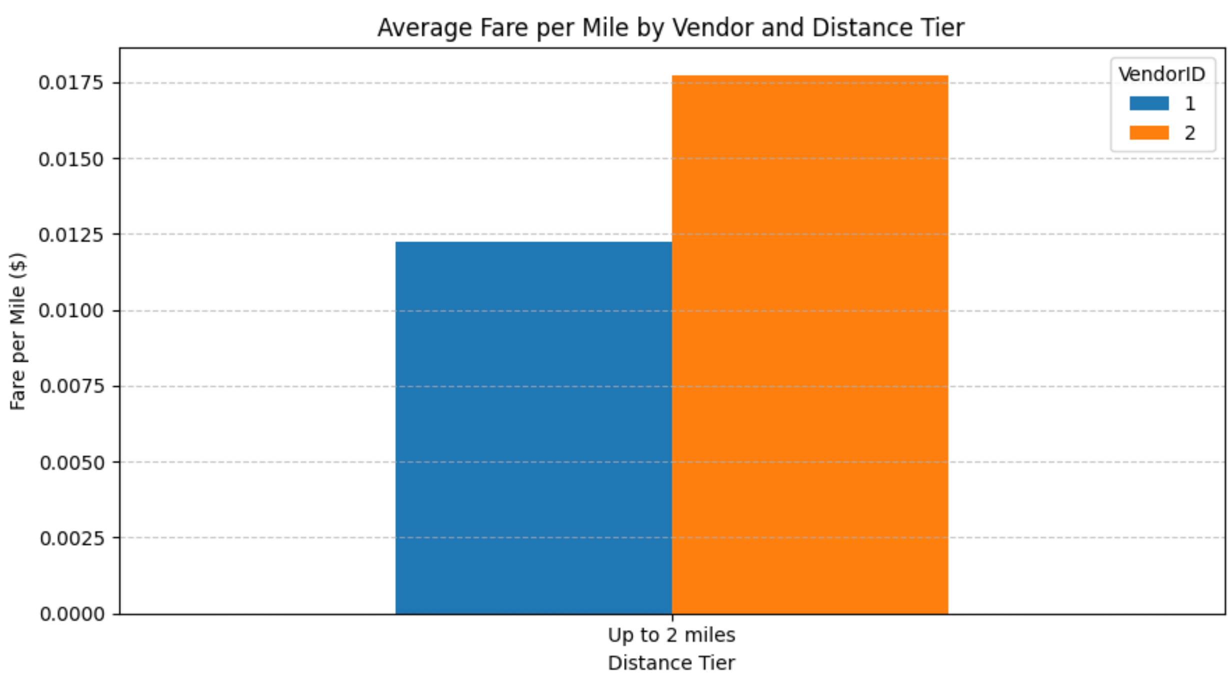
dtype: float64

	fare_per_mile
hour_of_day	
0	0.02
1	0.02
2	0.02
3	0.02
4	0.03
5	0.03
6	0.02
7	0.02
8	0.02
9	0.02
10	0.03
11	0.02
12	0.02
13	0.02
14	0.02
15	0.03
16	0.03
17	0.03
18	0.03
19	0.03
20	0.02
21	0.02
22	0.02
23	0.02

3.2.11. Analyse the average fare per mile for the different vendors

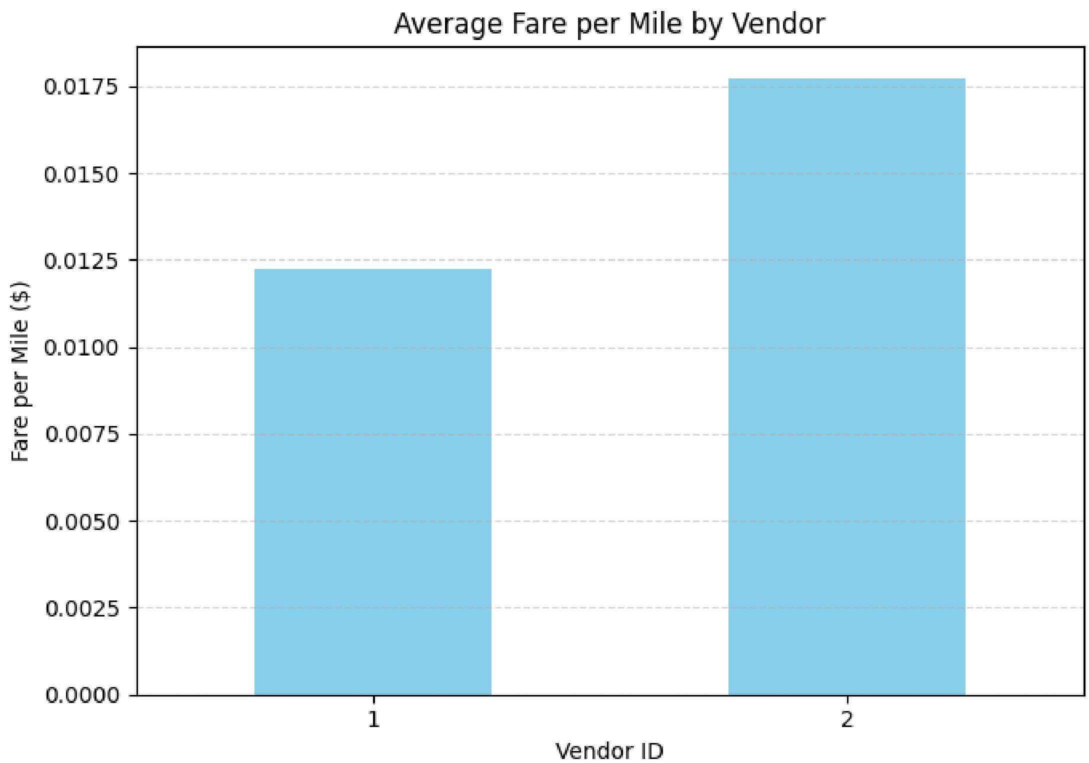


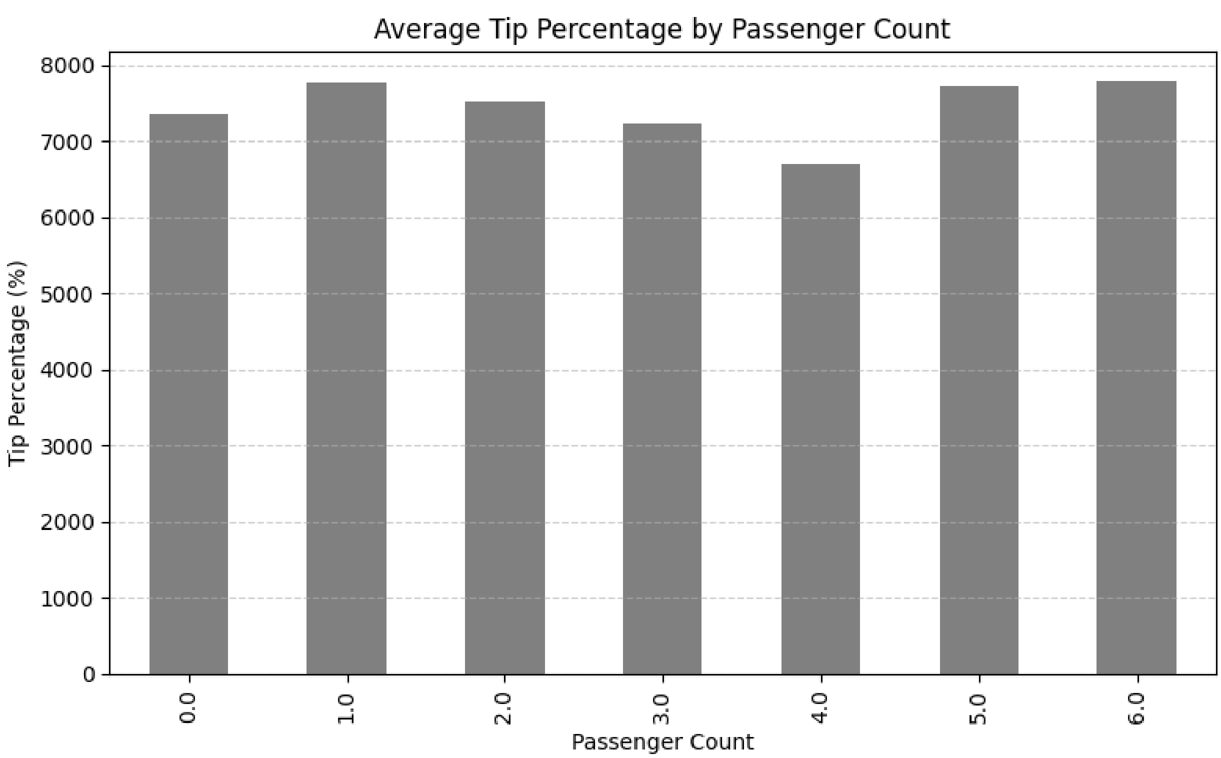
3.2.12. Compare the fare rates of different vendors in a distance-tiered fashion

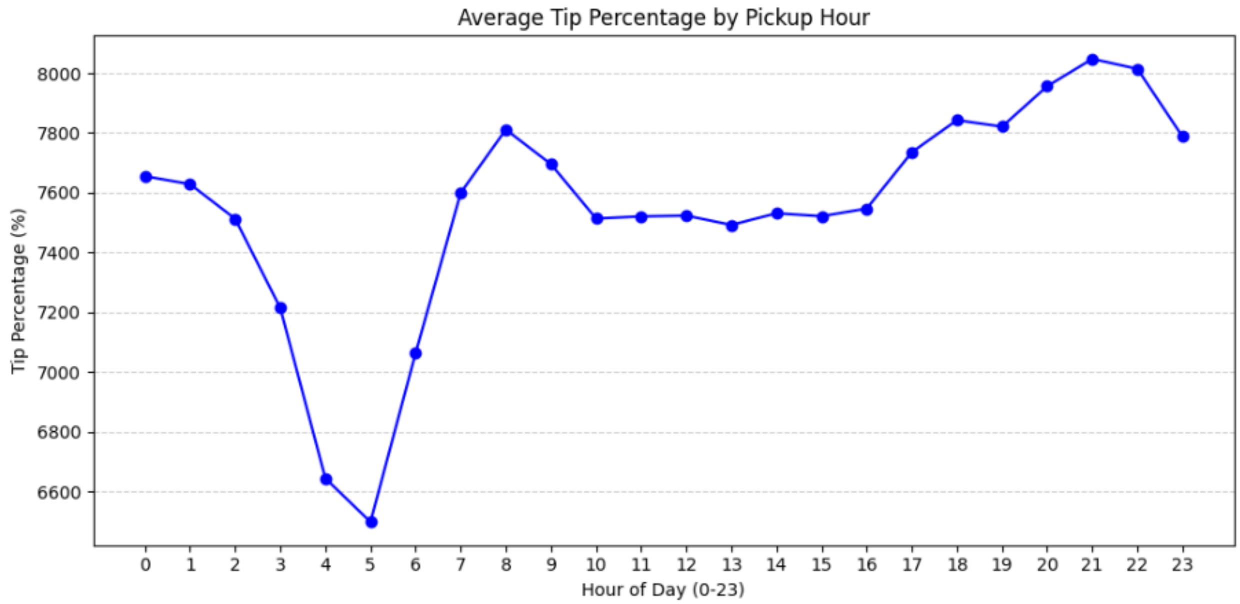


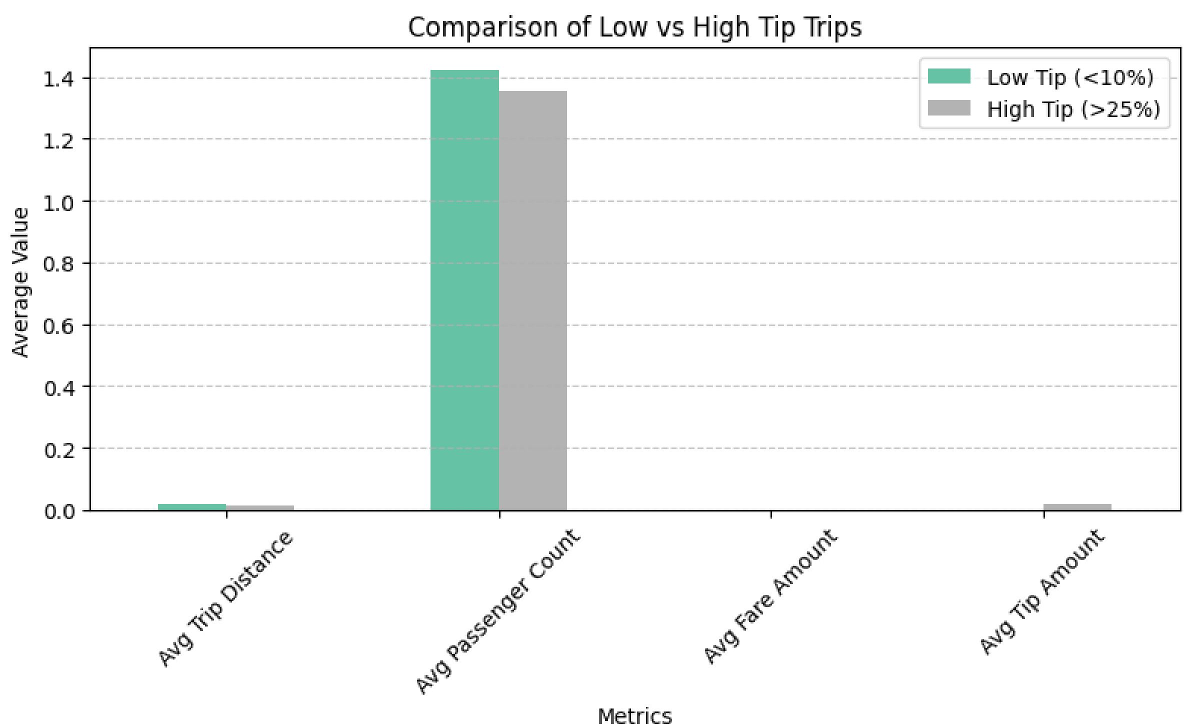
3.2.13. Analyse the tip percentages

```
Average Tip Percentage by Distance:
distance_category
Up to 2 miles        7676.350688
2 to 5 miles                 NaN
More than 5 miles            NaN
Name: tip_percentage, dtype: float64

Average Tip Percentage by Passenger Count:
passenger_category
1 passenger       7762.079995
2-3 passengers    7462.690167
4+ passengers     7236.778000
Name: tip_percentage, dtype: float64

Average Tip Percentage by Time of Pickup:
time_category
Midnight to 6 AM    7434.382746
6 AM to Noon        7585.160093
Noon to 6 PM        7562.828478
6 PM to Midnight    7911.194588
Name: tip_percentage, dtype: float64

Most Common Low Tip Scenarios:
distance_category  passenger_category  time_category
Up to 2 miles      1 passenger         Noon to 6 PM        110058
                                       6 PM to Midnight     80830
                                       6 AM to Noon         70189
                   2-3 passengers      Noon to 6 PM         34091
                                       6 PM to Midnight     27288
                   1 passenger         Midnight to 6 AM     23999
                   2-3 passengers      6 AM to Noon         15073
                   4+ passengers       Noon to 6 PM          8455
                                       6 PM to Midnight      6563
                   2-3 passengers      Midnight to 6 AM      6311
dtype: int64
```

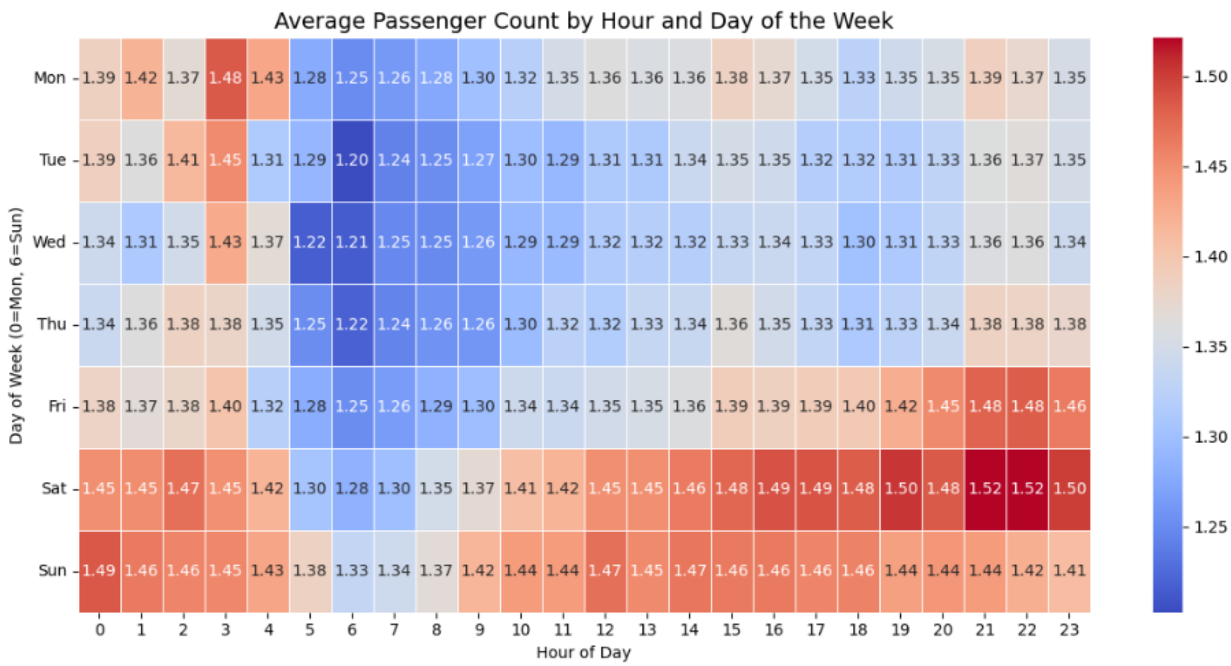

Average Fare per Mile by Vendor
Fare per Mile ($)
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
1
2
Vendor ID


Average Tip Percentage by Passenger Count
Tip Percentage (%)
0
1000
2000
3000
4000
5000
6000
7000
8000
0.0
1.0
2.0
3.0
4.0
5.0
6.0
Passenger Count

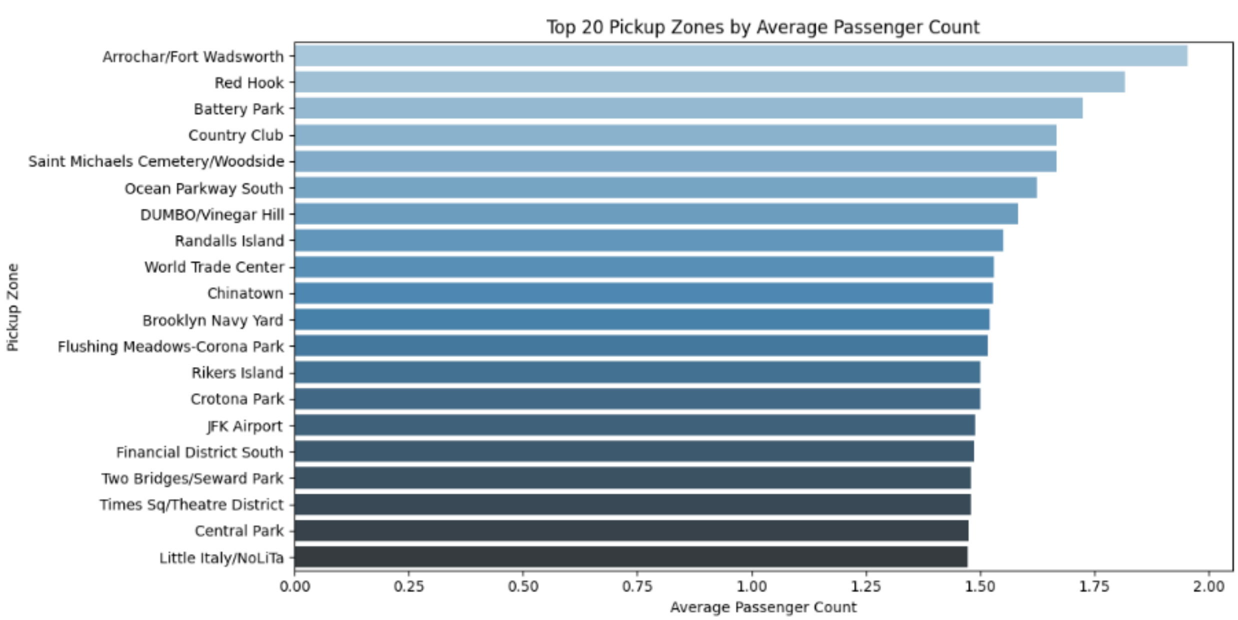




3.2.14. Analyse the trends in passenger count

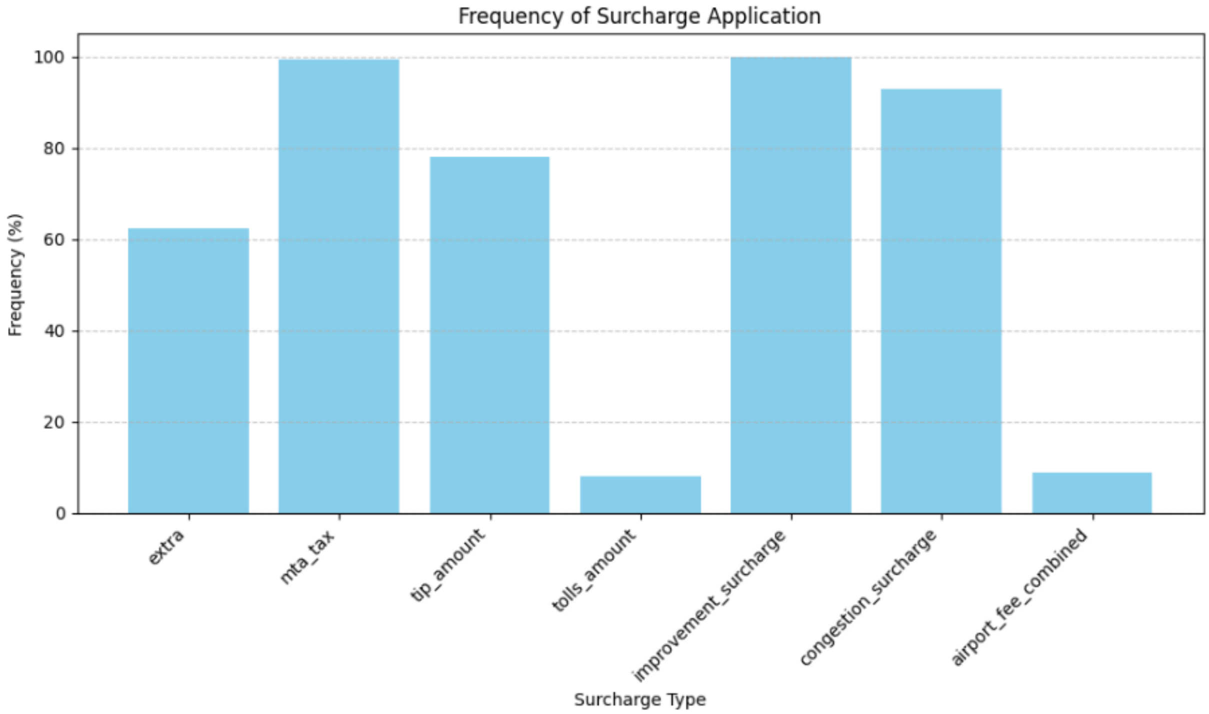


3.2.15. Analyse the variation of passenger counts across zones



3.2.16. Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.

```
Frequency of Surcharge Application (%):
extra                    62.312583
mta_tax                  99.357465
tip_amount               78.127946
tolls_amount              8.095659
improvement_surcharge    99.990323
congestion_surcharge     92.915310
airport_fee_combined      8.782154
dtype: float64
```



4. Conclusions

4.1. Final Insights and Recommendations

4.1.1. Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.

Key Findings from the Analysis

Based on the analyses of trip patterns, fare structures, tip behavior, passenger variations, and zone-based demand, several insights emerge:

- Peak Demand Hours & Days: The busiest hours for trips tend to be late evening (around 6 PM - 11 PM) and early morning rush hours. Weekends see a surge in nightlife-related pickups, while weekdays show steady demand during commute hours.
- High-Traffic Zones: Top pickup and drop-off zones are concentrated around commercial hubs, airports, and entertainment districts. Nighttime demand shifts towards downtown and nightlife districts, while daytime demand leans toward office areas and transit hubs.
- Fare & Tip Patterns: Short-distance trips tend to have higher tip percentages, while longer trips may have lower tipping rates. Higher fares per mile are charged for short-distance rides, likely due to base fare influence.
- Passenger Trends: Rush hour trips generally have fewer passengers per vehicle, while weekend rides involve more group travelers. Airports and tourist zones see a higher average passenger count per trip compared to local city zones.
- Surcharge Application: Extra charges such as congestion fees, airport surcharges, and tolls are applied frequently.

Strategic Recommendations for Demand Optimization

- Demand Forecasting by Time & Location: Focus fleet allocation based on hourly demand trends. Adjust dispatching strategies for commercial zones during weekdays and entertainment zones on weekends for maximum ride efficiency.
- Zone-Based Supply Adjustments: Ensure higher vehicle availability around train stations, airports, and tourist spots, especially during their busiest hours. Redirect supply towards business hubs during commute times and downtown nightlife areas for late-evening rides.
- Fare & Incentive Optimization: Dynamic pricing can be used during peak hours and high-demand zones. Provide targeted promotions for low-demand periods or zones with fewer trips to balance supply and demand.
- Routing & Dispatch Efficiency: Identify routes with slow traffic speeds and optimize dispatching accordingly to avoid delays.
- Customer Satisfaction & Tip Optimization: Encourage passenger-friendly service standards in areas with historically lower tip percentages. Identify low tip scenarios and improve service models where needed.

By implementing these insights, ride service providers can enhance trip efficiency, improve customer satisfaction, and optimize earnings.

4.1.2. Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analysing trip trends across time, days and months.

key positioning strategies:

- Peak-Hour Allocation: During morning rush hours (6 AM - 9 AM), cabs should be stationed at residential areas, transit hubs, and office districts. Evening peak (5 PM - 8 PM) requires more availability near business zones, shopping areas, and entertainment hubs.
- Zone-Based Deployment: Position weekend fleets near nightlife spots, tourist attractions, and airport terminals for late-night surges. On weekdays, focus on commercial hubs and business districts to maximize regular commuter rides.
- Dynamic Rebalancing: If demand suddenly surges in one area due to events, weather changes, or public transport delays, cabs should dynamically shift to accommodate these short-term spikes. Data-driven heatmaps and real-time tracking can help balance vehicle supply efficiently.
- Traffic & Routing Optimization: Cabs should be strategically placed near main roads, avoiding heavy congestion points while ensuring quick access to busy pickup locations. Using predictive analytics, drivers can be guided toward high-demand areas before the surge happens.
- Fare & Demand-Based Prioritization: During low-demand hours, incentivizing drivers through priority zones (hotspots with frequent riders) ensures consistent availability without over-supply. Using zone-specific surge pricing, fleet operators can keep vehicles available where demand is highest, ensuring profitability.

By implementing these strategies, cab fleets can efficiently meet customer demand, reduce idle time, and optimize driver earnings.

4.1.3. Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.

To maximize revenue while staying competitive, the pricing strategy should be data-driven, adjusting dynamically based on demand patterns, trip characteristics, and external factors. Here are some key adjustments:

Dynamic Pricing Model

- Implement real-time surge pricing based on demand fluctuations. Higher rates during peak hours and lower fares in off-peak times ensure steady revenue without overpricing.
- Factor in event-based demand spikes, adjusting pricing near stadiums, concerts, and transit hubs before crowds arrive.

Distance-Based Fare Optimization

- Short-distance rides should have competitive base fares but slightly higher per-mile rates to balance profitability.
- Long-distance fares could include discounted per-mile rates to encourage riders to opt for extended trips.

Time-of-Day Pricing Adjustments

- Nighttime and early-morning rides should include premium pricing for safety and limited availability.
- Midday pricing should be optimized to encourage more riders during lower-demand hours.

Zone-Based Pricing

- High-demand pickup zones (airports, business districts) should have slightly increased base fares due to traffic congestion and longer wait times.
- Low-demand zones should use competitive pricing and discounts to increase ridership.

Subscription & Loyalty Discounts

- Offer subscription-based ride discounts for frequent commuters.
- Reward loyal riders with fare discounts after a set number of trips.

Competitive Benchmarking

- Regularly analyze competitor rates and adjust pricing models to stay within a reasonable range.

By implementing these strategic pricing adjustments, vendors can maximize earnings, attract more customers, and ensure balanced supply-demand economics.