Report: Optimizing NYC Taxi Operations

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

1. Data Preparation

1.1. Loading the dataset

1.1.1. Sample the data and combine the files

Initially, I extracted 500,000 records per month as instructed. After refining the sample further, the final combined DataFrame holds nearly **1,915,511** rows.

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

To fix the index, the DataFrame's index is reset so it starts from zero in a clean sequence, and the old index is removed. After that, unnecessary columns such as the **pickup date** and **pickup hour** are dropped because they are not needed for the analysis.

2.1.2. Combine the two airport_fee columns

The two **airport fee columns** (**airport_fee** and **Airport_fee**) are combined by taking the maximum value between them for each row, ensuring the correct fee is captured even if it appears in only one of the columns. After creating this I combined value, the original duplicate airport fee columns are removed since they are no longer needed.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

	0
VendorID	0.000000
tpep_pickup_datetime	0.000000
tpep_dropoff_datetime	0.000000
passenger_count	3.400659
trip_distance	0.000000
RatecodeID	3.400659
store_and_fwd_flag	3.400659
PULocationID	0.000000
DOLocationID	0.000000
payment_type	0.000000
fare_amount	0.000000
extra	0.000000
mta_tax	0.000000
tip_amount	0.000000
tolls_amount	0.000000
improvement_surcharge	0.000000
total_amount	0.000000
congestion_surcharge	3.400659
airport_fee_combined	3.400659

dtype: float64

2.2.2. Handling missing values in passenger_count

I checked the DataFrame for any rows containing null values, and then filled the missing entries in the **passenger_count** column by replacing them with the most frequently occurring value (the mode), ensuring the column has no gaps.

2.2.3. Handle missing values in RatecodeID

I handled the missing values in the **RatecodeID** column by replacing all null entries with the mode, which is the most common value in that column, ensuring consistency and removing gaps in the data.

This approach is suitable for categorical data like **RatecodeID,** as it preserves the most common pattern in the dataset without introducing bias from rare or extreme values.

2.2.4. Impute NaN in congestion_surcharge

I addressed the missing values in the **congestion surcharge** column by filling all null entries with the **median** value of that column, ensuring the data remains consistent without being skewed by extreme values.

2.3. Handling Outliers and Standardizing Values

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

Payment Type:
Outliers were identified where payment_type had a value of 0, which is not a valid code. These entries were removed from the dataset.

Trip Distance:
Outliers were present in extremely long or suspiciously short trips.

Trips with distance < 0.1 miles but fare > $300 were removed.
Trips with distance > 250 miles were also removed as extreme outliers.
Trips with 0 distance and fare, yet with different pickup and dropoff locations, were treated as invalid and removed.

Tip Amount:
No filtering was applied to tip_amount for zero values since tipping is optional.
However, high-end outliers (very large tips) were implicitly handled through min-max standardization, which scaled values between 0 and 1, minimizing the impact of extreme tips.

3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

3.1.1. Classify variables into categorical and numerical

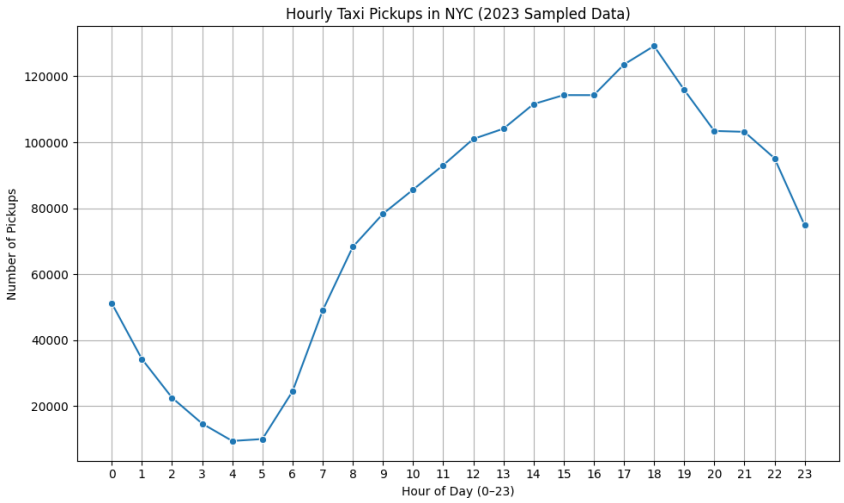
Categorise the varaibles into Numerical or Categorical.

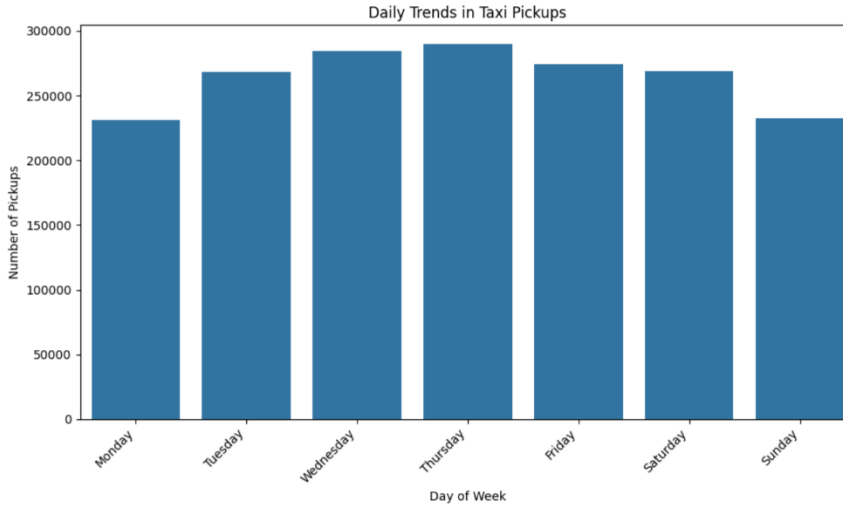
- `VendorID`:
- `tpep_pickup_datetime`:
- `tpep_dropoff_datetime`:
- `passenger_count`:
- `trip_distance`:
- `RatecodeID`:
- `PULocationID`:
- `DOLocationID`:
- `payment_type`:
- `pickup_hour`:
- `trip_duration`:

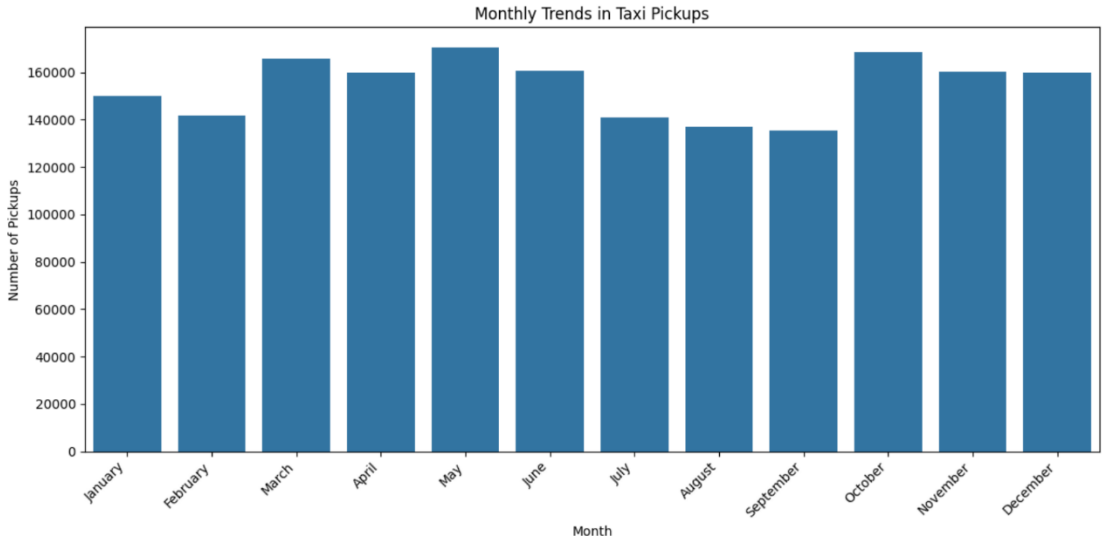
The following monetary parameters belong in the same category, is it categorical or numerical?

- `fare_amount`
- `extra`
- `mta_tax`
- `tip_amount`
- `tolls_amount`
- `improvement_surcharge`
- `total_amount`
- `congestion_surcharge`
- `airport_fee`

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





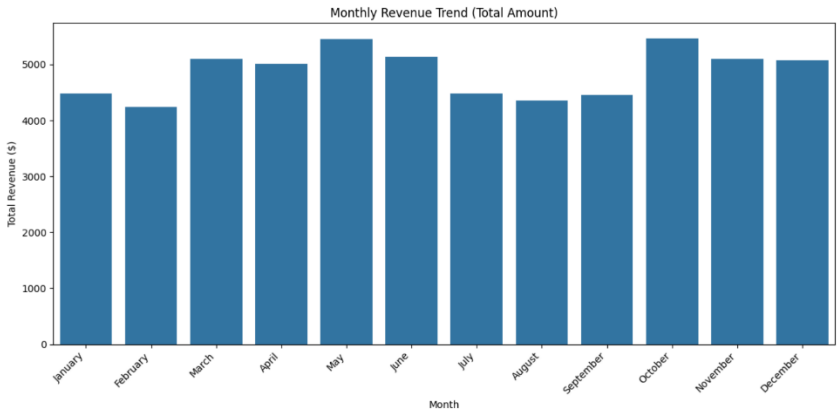


- The analysis shows strong hourly peaks during typical commuting times, especially in the late afternoon and early evening. Mid-week days consistently record the highest number of pickups, indicating increased weekday travel activity.
- Monthly trends highlight seasonal spikes around spring and early autumn. Overall, NYC taxi demand follows predictable daily and seasonal patterns that companies can use for better fleet planning and resource allocation.

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

- Analyzed key columns such as **fare amount**, **tip amount, total amount**, and **trip distance** by checking how many values were zero or negative in each of them. After identifying these issues, I created a filtered version of the dataset that keeps only the rows where all four of these parameters have **non-zero values**, allowing the analysis to focus on valid trips and removing entries that could distort results.
- This filtering helped clean the dataset while keeping real-world behavior like no tipping intact.

3.1.4. Analyze the monthly revenue trends



Monthly Revenue Trends

- Revenue shows strong month-to-month fluctuations, with noticeable peaks in May and October.
- The lowest revenue appears in February, indicating a dip early in the year.
- Despite fluctuations, overall revenue stays within a stable range of $4,300–$5,500.
- The business maintains consistent demand through the year with no extreme drops.

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

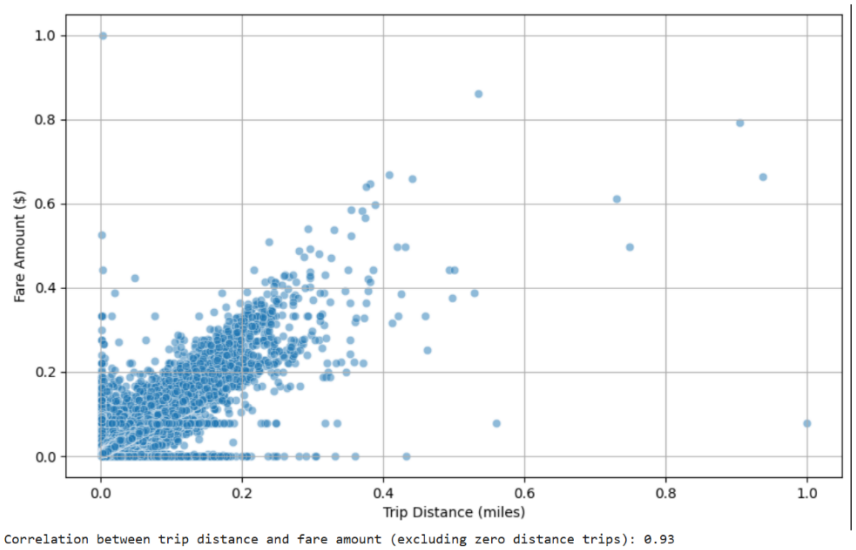
pickup_quarter	total_amount
2022Q4	0.00
2023Q1	23.69
2023Q2	26.73
2023Q3	22.77
2023Q4	26.81

dtype: float64

Quarterly Revenue

- 2023 Q2 and Q4 show the highest quarterly earnings, indicating stronger business activity during these periods.
- 2022 Q4 has a value of *0.00*, likely due to absence or missing data rather than actual revenue drop.
- Revenue grows steadily from 2023 Q1 → Q2, dips slightly in Q3, and increases again in Q4.

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

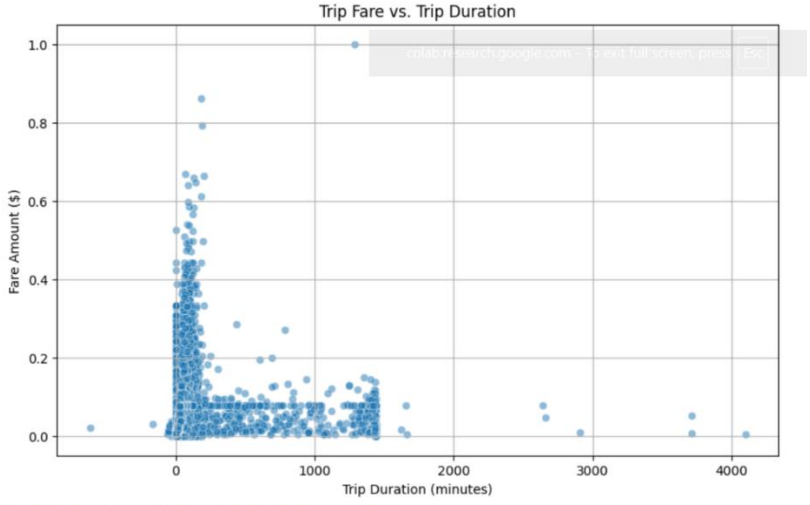



```
Correlation between trip distance and fare amount (excluding zero distance trips): 0.93
```

Trip Distance vs Fare Amount

- There is a very strong positive correlation (0.93) between trip distance and fare amount.
- Longer trips almost always result in higher fares, which validates the pricing model.
- Points are tightly clustered along an upward trend, showing reliable fare calculation based on distance.

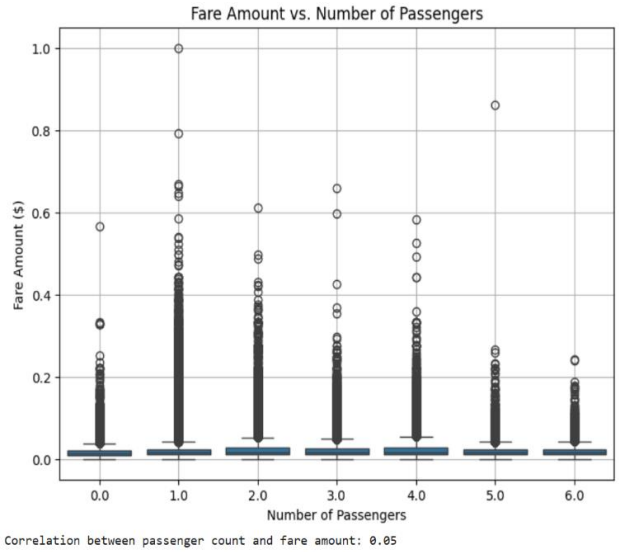
3.1.7. Analyse the relationship between fare/tips and trips/passengers




```
Correlation between trip duration and fare amount: 0.27
```

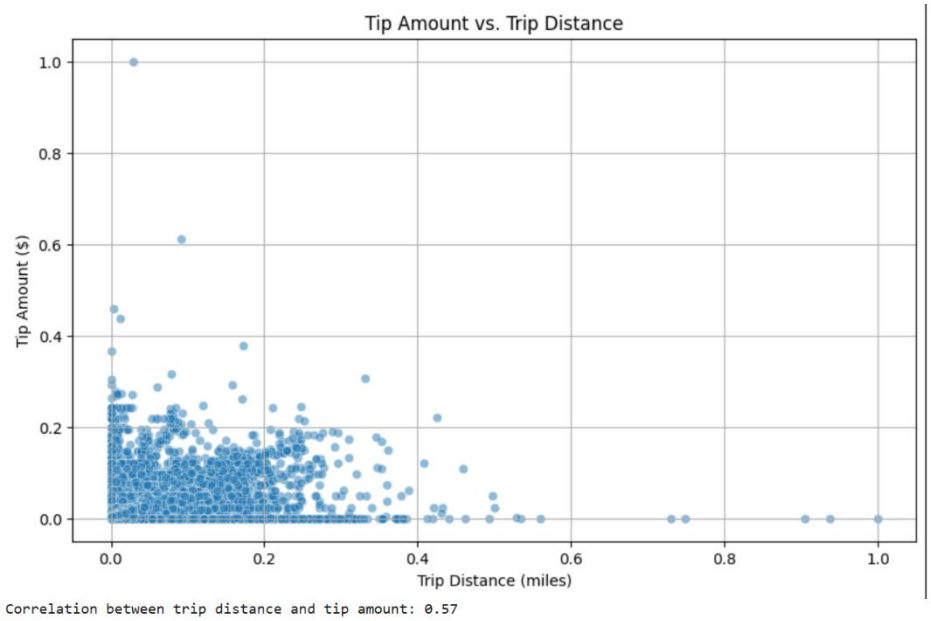
Trip Duration vs Fare Amount

- Correlation between fare and duration is weak (0.27).
- Indicates that fare is not heavily dependent on how long the trip takes.
- Some long-duration trips have unusually low fares, suggesting traffic .



Passenger Count vs Fare Amount

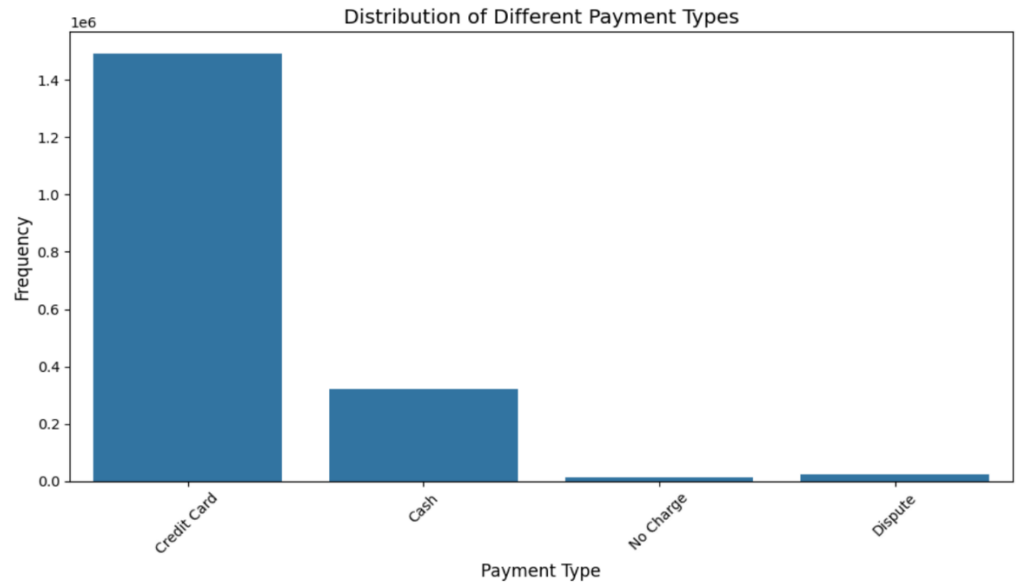
- Correlation is very weak (0.05).
- Fare does not significantly change based on the number of passengers.
- Most trips involve 1–2 passengers, and distribution of fares remains similar across categories.



Trip Distance vs Fare Amount

- There is a very strong positive correlation (0.93) between trip distance and fare amount.
- Longer trips almost always result in higher fares, which validates the pricing model.
- Points are tightly clustered along an upward trend, showing reliable fare calculation based on distance.

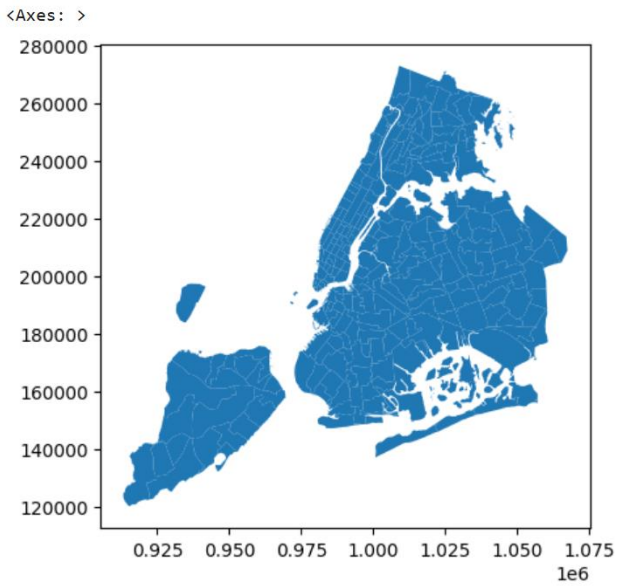
3.1.8. Analyse the distribution of different payment types




```
payment_type
1    80.66
2    17.33
3     0.66
4     1.35
Name: count, dtype: float64
```

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

Merge was performed : zones data into trip data using the `**locationID**` and `**PULocationID**` columns.

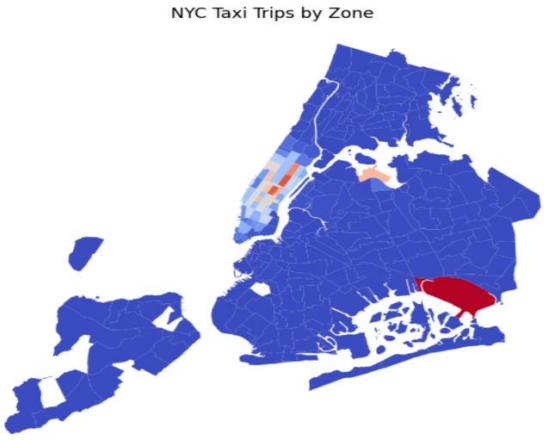
3.1.11. Find the number of trips for each zone/location ID

	PULocationID	num_trips
0	1	246
1	2	2
2	3	47
3	4	1861
4	5	19

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	246.0
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	47.0
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	19.0

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



3.1.14. Conclude with results

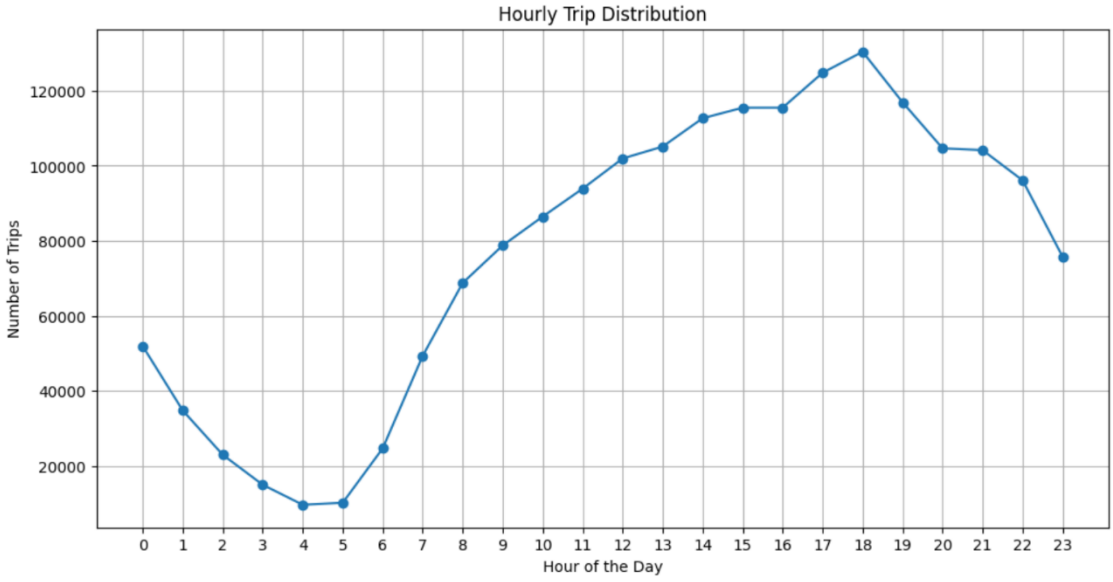
- There is a clear positive relationship between trip distance and fare, indicating that fares generally increase with longer distances.
- **Weekday** travel peaks during **morning and evening** rush hours, whereas **weekends** tend to see more **late-night trips**.
- The highest trip concentrations occur around **Airport areas** and **Midtown**, which act as major pickup and drop-off hubs.
- Most rides involve **1–2 passengers**, and **credit card** payments remain the most commonly used method.
- Seasonal patterns show that the **third quarter** experiences the highest trip volumes.
- After cleaning the dataset, anomalies were removed and important numerical columns were standardized, improving the overall reliability of the analysis.

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
111555	237	193	9	0.000084
1079	7	149	12	0.000096
86692	182	250	19	0.000103
91728	209	209	14	0.000120
41637	114	193	21	0.000138
76318	162	138	22	0.000176
121710	260	129	17	0.000182
3083	13	211	0	0.000256
94251	216	216	7	0.000264
1506	10	145	11	0.000267

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



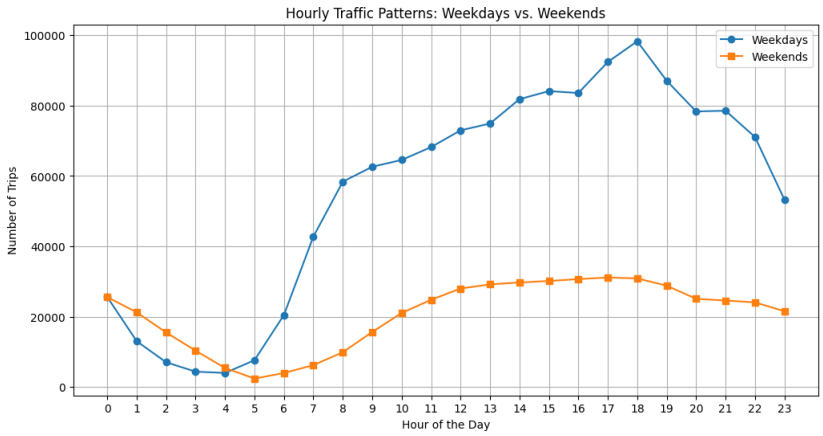

```
Busiest hour: 18
Number of trips during busiest hour: 130389
```

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

pickup_hour	count
18	130389
17	124796
19	116917
16	115489
15	115483

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	98789	JFK Airport
1	237	87946	Upper East Side South
2	161	86405	Midtown Center
3	236	77635	Upper East Side North
4	162	66102	Midtown East
5	138	64858	LaGuardia Airport
6	186	64201	Penn Station/Madison Sq West
7	230	61880	Times Sq/Theatre District
8	142	61186	Lincoln Square East
9	170	54490	Murray Hill

Top 10 Dropoff Zones:

	LocationID	Dropoff_Trips	zone
0	236	82124	Upper East Side North
1	237	78416	Upper East Side South
2	161	72751	Midtown Center
3	230	57314	Times Sq/Theatre District
4	170	54633	Murray Hill
5	162	52088	Midtown East
6	239	51928	Upper West Side South
7	142	51818	Lincoln Square East
8	141	48926	Lenox Hill West
9	68	46336	East Chelsea

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

zone	pickup_dropoff_ratio
Freshkills Park	0.000000
Green-Wood Cemetery	0.000000
Oakwood	0.000000
Grymes Hill/Clifton	0.000000
Rossville/Woodrow	0.000000
Heartland Village/Todt Hill	0.023256
Saint George/New Brighton	0.025000
West Brighton	0.041667
Highbridge Park	0.044444
Newark Airport	0.044711

dtype: float64

zone	pickup_dropoff_ratio
East Elmhurst	7.602473
JFK Airport	4.462418
LaGuardia Airport	2.904523
Penn Station/Madison Sq West	1.590196
Greenwich Village South	1.380461
Central Park	1.358355
West Village	1.338495
Midtown East	1.269045
Garment District	1.190779
Midtown Center	1.187681

dtype: float64

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

pickup_zone	PULocationID
East Village	15524
JFK Airport	13707
West Village	12573
Clinton East	9877
Lower East Side	9714
Greenwich Village South	8804
Times Sq/Theatre District	7912
Penn Station/Madison Sq West	6362
Midtown South	6110
LaGuardia Airport	6094

dtype: int64

dropoff_zone	DOLocationID
East Village	8373
Clinton East	7004
Murray Hill	6070
Gramercy	5631
East Chelsea	5523
Lenox Hill West	5194
Yorkville West	4991
West Village	4974
Times Sq/Theatre District	4453
Flatiron	4392

dtype: int64

3.2.8. Find the revenue share for nighttime and daytime hours

```
Nighttime Revenue Share: 12.09%
Daytime Revenue Share: 87.91%
```

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	4.054886
2.0	2.363664
3.0	1.534000
4.0	1.477359
5.0	0.661306
6.0	0.593923

dtype: float64

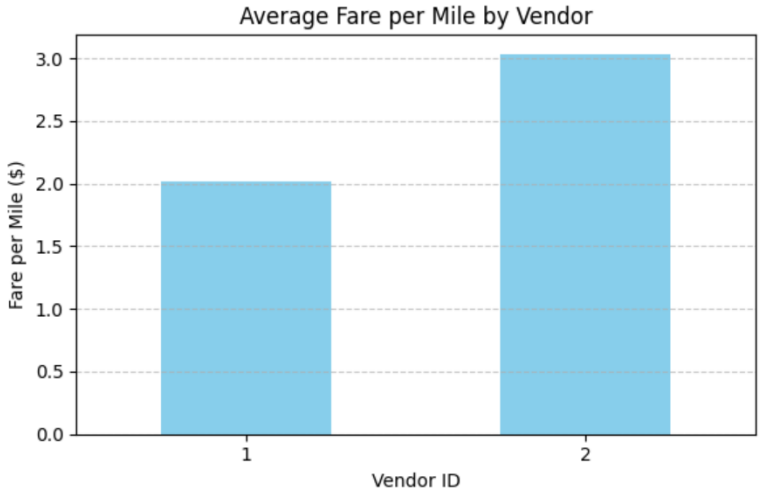
3.2.10. Find the average fare per mile by hours of the day and by days of the week

day_of_week	fare_per_mile
Monday	4.15
Tuesday	4.15
Wednesday	4.22
Thursday	4.41
Friday	4.30
Saturday	4.11
Sunday	4.28

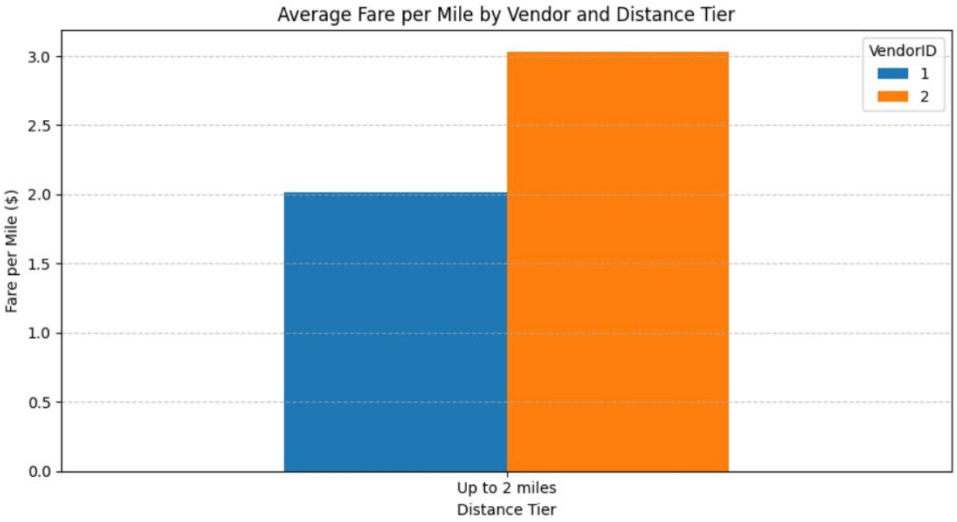
dtype: float64

hour_of_day	fare_per_mile
0	4.06
1	3.83
2	3.96
3	4.18
4	5.12
5	4.85
6	3.53
7	3.68
8	3.62
9	4.03
10	4.04
11	4.03
12	4.22
13	4.20
14	4.40
15	4.62
16	5.11
17	4.79
18	4.46
19	4.63
20	3.84

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



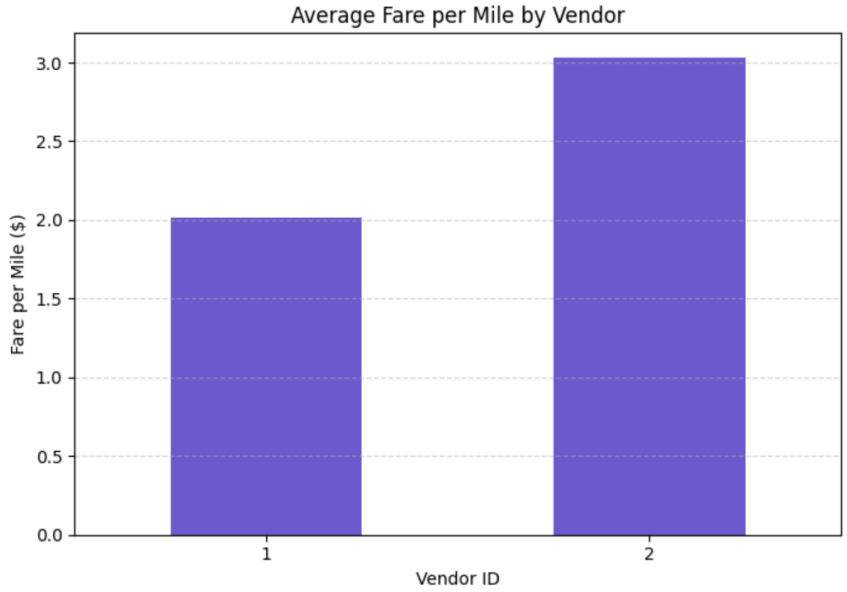
3.2.12. Compare the fare rates of different vendors in a distance-tiered fashion

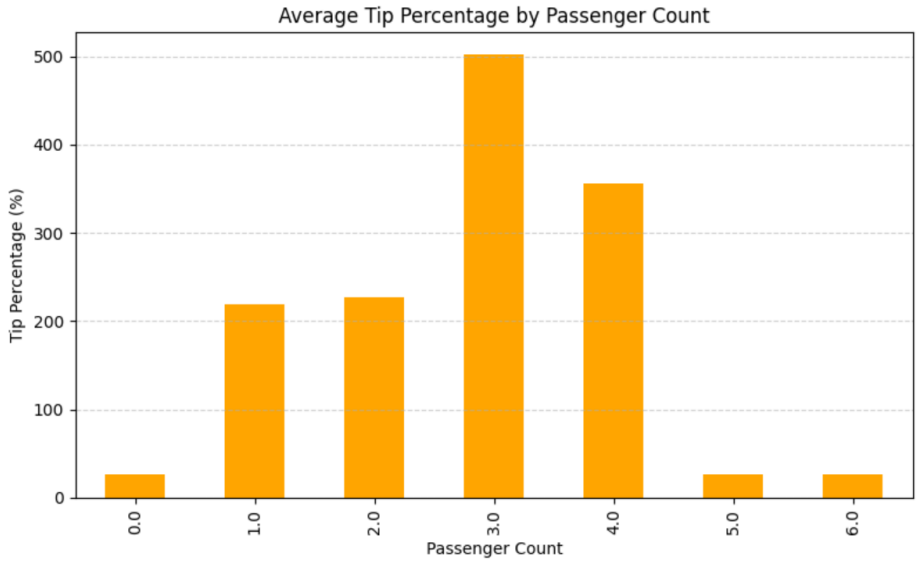


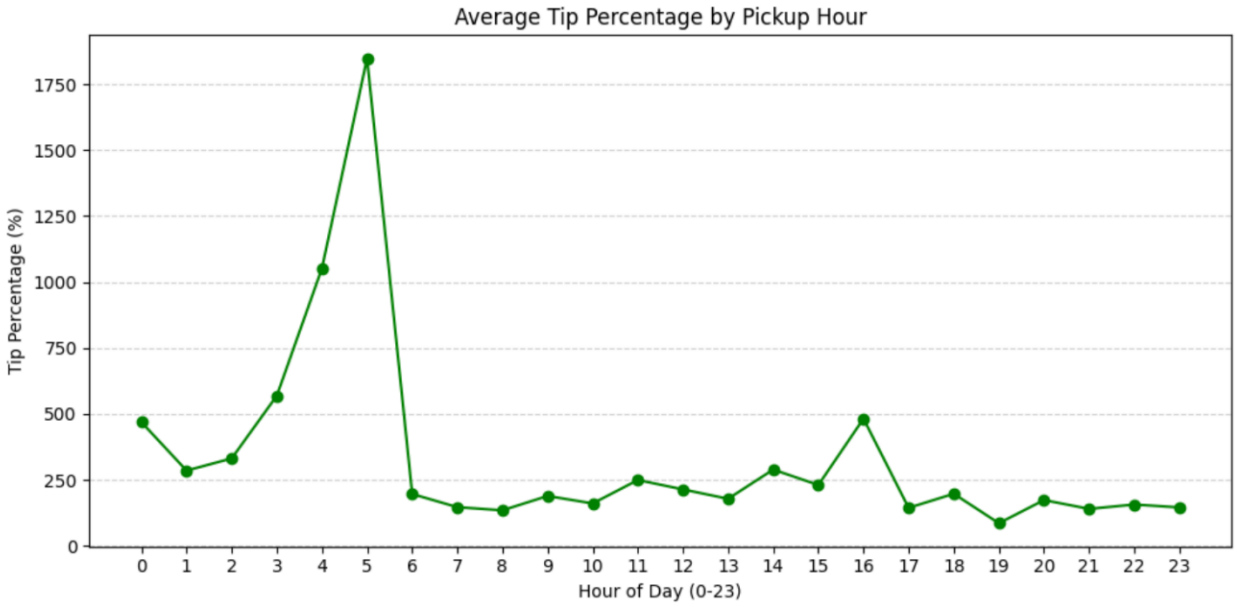
3.2.13. Analyse the tip percentages

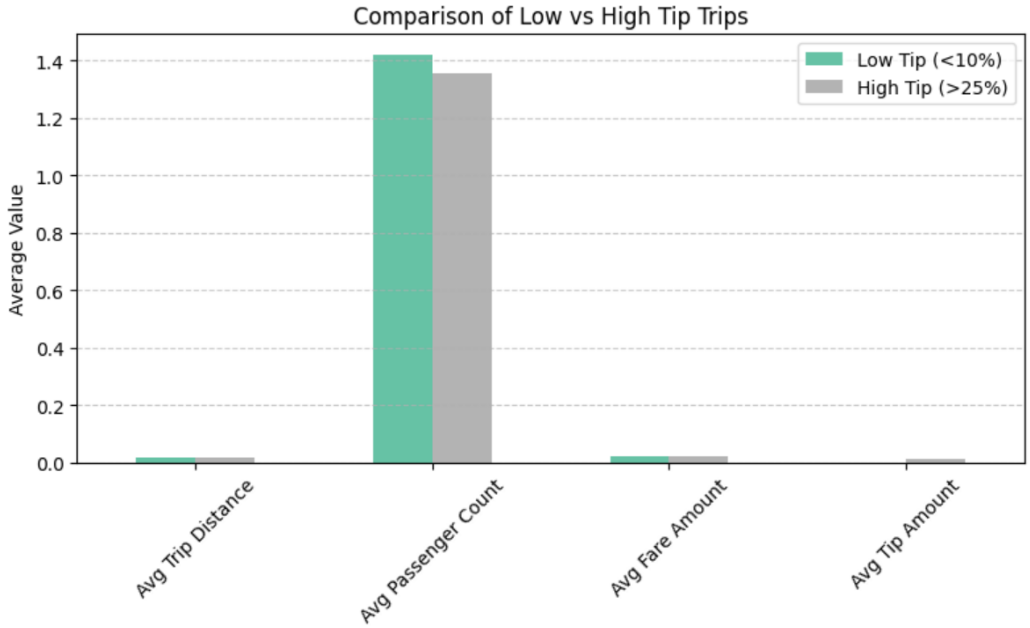
```
Average Tip Percentage by Distance:
distance_category
Up to 2 miles        227.065949
2 to 5 miles                NaN
More than 5 miles           NaN
Name: tip_percentage, dtype: float64

Average Tip Percentage by Passenger Count:
passenger_category
1 passenger       219.658282
2-3 passengers    281.552950
4+ passengers     189.119227
Name: tip_percentage, dtype: float64

Average Tip Percentage by Time of Pickup:
pickup_times_category
Midnight to 6 AM    547.055798
6 AM to Noon        182.936718
Noon to 6 PM        256.631003
6 PM to Midnight    150.591821
Name: tip_percentage, dtype: float64
```

```
Most Common Low Tip Scenarios:
distance_category  passenger_category  pickup_times_category
Up to 2 miles      1 passenger         Noon to 6 PM             120101
                                       6 PM to Midnight          90770
                                       6 AM to Noon              75362
                   2-3 passengers      Noon to 6 PM              36849
                                       6 PM to Midnight          30133
                   1 passenger         Midnight to 6 AM          27433
                   2-3 passengers      6 AM to Noon              16212
                   4+ passengers       Noon to 6 PM               8880
                   2-3 passengers      Midnight to 6 AM           7254
                   4+ passengers       6 PM to Midnight           7217
```

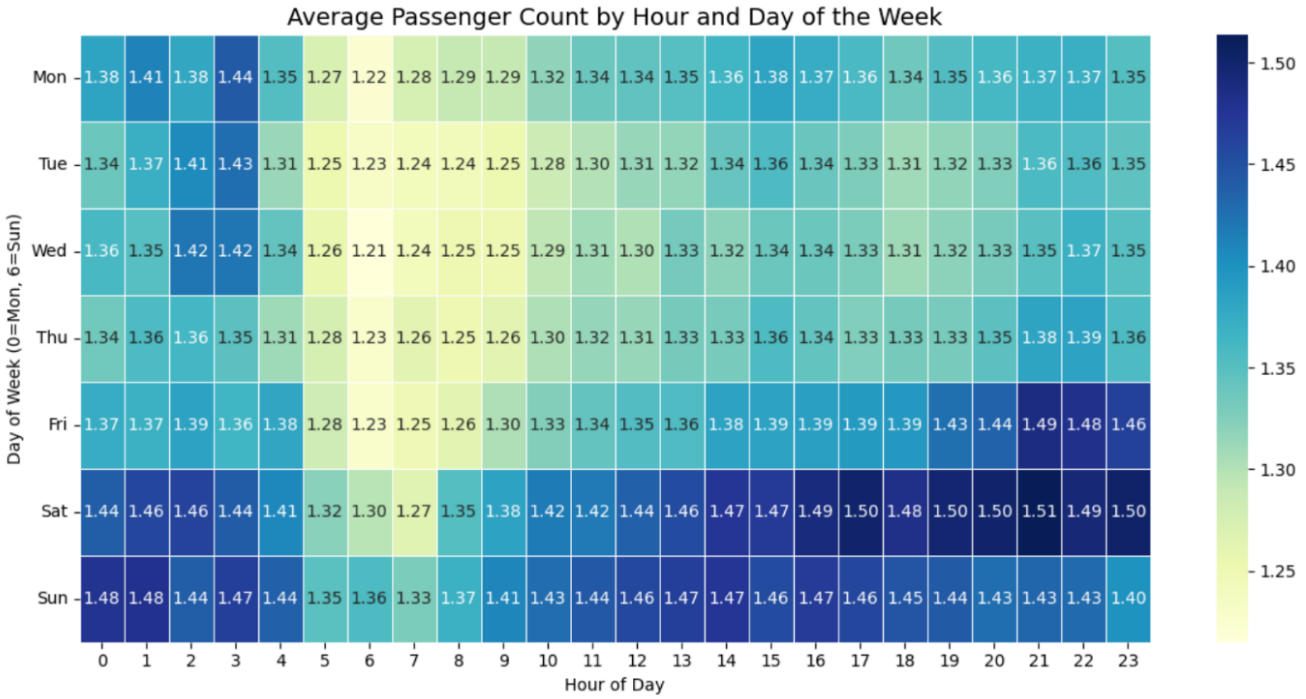

Average Fare per Mile by Vendor
Fare per Mile ($)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
Vendor ID


Average Tip Percentage by Passenger Count
Tip Percentage (%)
500
400
300
200
100
0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
Passenger Count

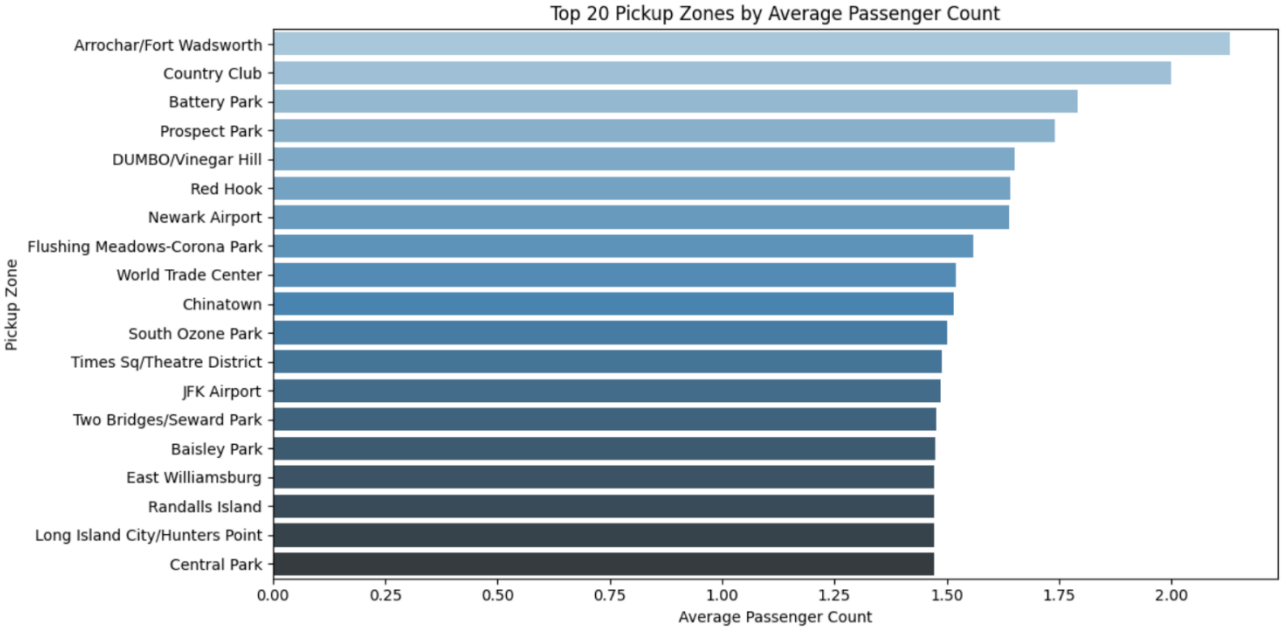

Average Tip Percentage by Pickup Hour
Tip Percentage (%)
0
250
500
750
1000
1250
1500
1750
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day (0-23)


Comparison of Low vs High Tip Trips
Low Tip (<10%)
High Tip (>25%)
Average Value
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Avg Trip Distance
Avg Passenger Count
Avg Fare Amount
Avg Tip Amount

3.2.14. Analyse the trends in passenger count

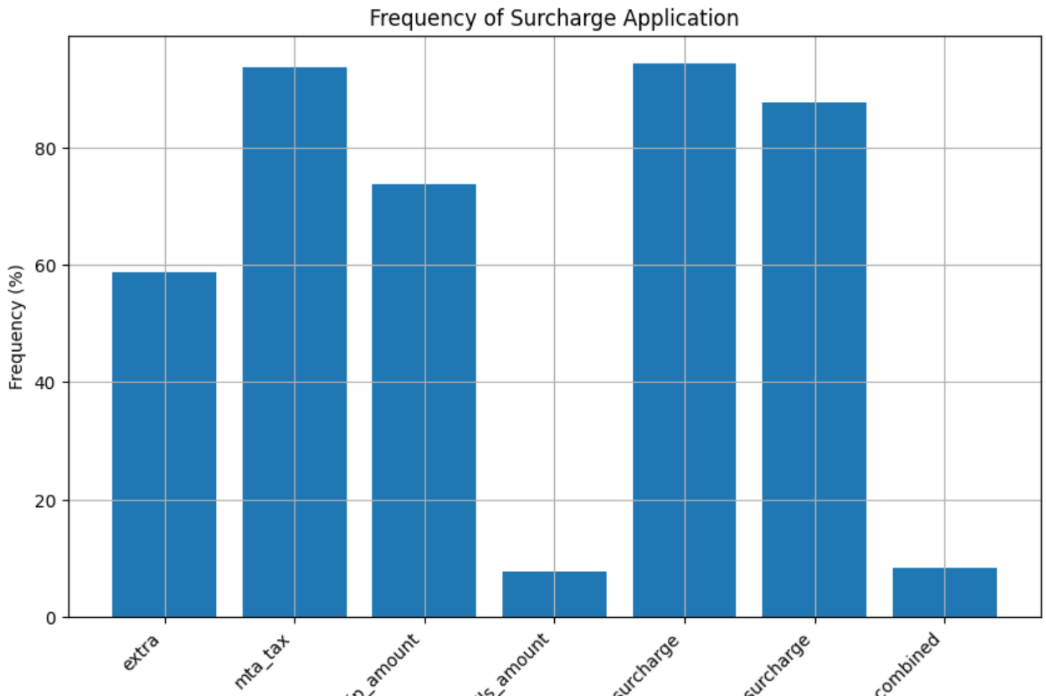


3.2.15. Analyse the variation of passenger counts across zones



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

```
Frequency of Surcharge Application (%):
extra                    58.774969
mta_tax                  93.795128
tip_amount               73.795870
tolls_amount              7.632324
improvement_surcharge    94.398936
congestion_surcharge     87.673420
airport_fee_combined      8.296167
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 Insights:

- Time Trends: Demand peaks during rush hours, weekends, and certain months.Nightlife zones show strong late-night activity.
- Financial Patterns: Fares rise with distance and duration. Shared rides offer savings. Tip percentages vary based on ride quality and trip characteristics.
- Geographical Trends: Airports, business hubs, and popular destinations attract the most trips. Some areas show pickup–dropoff imbalances. Nightlife districts become hotspots after dark.

- Vendor/Surcharges: Fare structures differ across vendors, with frequent surcharges and tiered pricing based on distance.

Recommendations:

- Demand: Focus resources on high-demand zones and peak timings; strengthen late-night availability; promote shared/group rides.
- Supply: Allocate more taxis to busy areas, use dynamic pricing, encourage repositioning, and offer driver incentives during low-traffic hours.
- Customer Experience: Improve driver training, expand payment options, and actively promote ride-sharing.
- Continuous Improvement: Use data to monitor patterns, gather customer feedback, and collaborate with city officials to improve overall mobility.

Conclusion Story

By analyzing travel patterns across time, zones, and trip characteristics, NYC taxi services can better understand rider needs. Aligning taxi supply with high-demand periods and locations helps reduce wait times and boosts efficiency. Optimizing pricing through insights on distance, duration, and tipping improves both revenue and customer satisfaction. Enhanced service quality and flexible payment options strengthen the overall rider experience. Ongoing data monitoring and feedback ensure that strategies stay relevant and effective. Together, these data-driven improvements create a more reliable, efficient, and customer-focused taxi system for the city.

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.

Strategically Positioning of Cabs:

- **Hourly Demand Alignment**: Deploy cabs according to changing demand patterns across the day—peak commuting hours, late-night travel spikes, and low-traffic intervals—while considering monthly and seasonal variations.
- **Weekly Movement Patterns**: Focus on office zones and commercial corridors during weekdays, and shift taxis toward leisure areas, residential neighborhoods, and event locations on weekends or holidays.
- **Geographical Priority Zones**: Increase cab presence in consistently busy regions like airports, major transit hubs, and nightlife districts, while addressing zones with uneven pickup and drop-off volumes.
- **Analytics-Based Optimization**: Utilize demand forecasting tools, real-time data feeds, and predictive modeling to reposition taxis dynamically and quickly adapt to sudden shifts in rider activity.
- **Tech-Enabled Coordination**: Use GPS heatmaps, tracking systems, and dashboard insights, supported by regular communication with drivers and cooperation with city authorities to maintain efficient fleet distribution.

By implementing these strategies, taxi companies and drivers can optimize cab positioning to meet customer demand, minimize wait times, and enhance efficiency in NYC.

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

Data-Driven Pricing Strategy Enhancements

- **Real-Time Demand Pricing:** Implement a dynamic fare structure that adjusts automatically based on rider demand, driver availability, weather conditions, and traffic congestion. Increase prices during heavy rush periods and provide lower rates during slow hours to maintain rider interest.
- **Distance & Zone-Based Tiers**: Introduce flexible pricing tiers where short-distance trips remain affordable, while longer journeys follow a structured, distance-based tariff. Add zone-specific pricing for airports, business districts, and nightlife zones where demand patterns differ significantly.
- **Ride-Sharing Incentives:** Promote shared rides through group discounts, loyalty offers, and reduced per-person fares to increase occupancy levels, lower operational costs, and attract budget-conscious passengers.
- **Optimized Surcharges:** Review surcharge patterns using historical and real-time insights; apply peak surcharges only when justified by demand spikes. Ensure transparency so passengers clearly understand when and why extra charges are added.
- **Market & Competitor Analysis:** Track pricing strategies of rival vendors and adjust fares to stay competitive. Emphasize service quality, reliability, and unique features to justify slightly higher fares where value is delivered.
- **Continuous Data Evaluation:** Use analytics dashboards, A/B pricing

experiments, and customer feedback to refine fare models regularly. Update pricing rules dynamically to maximize revenue without compromising customer satisfaction**.**