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.

1. Data Preparation

1.1. Loading the dataset

1.1.1. Sample the data and combine the files

Following the instructions, I initially sampled **500,000 records** from each monthly Parquet file. I further reduced the sample size to a level where the final combined DataFrame contained approximately **1.89 million rows**.

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

Column names were cleaned by stripping spaces and ensuring consistent formatting.

2.1.2. Combine the two airport_fee columns

The dataset contained two similar columns: **airport_fee** and **Airport_fee**, likely caused by inconsistent column naming across different monthly files.
To resolve this, I created a new column called **airport_fee_combined** by taking the **maximum value across both columns for each row**, ensuring no data was lost. After combining, I **dropped the original columns** to avoid redundancy:

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	3.420903
trip_distance	0.000000
RatecodeID	3.420903
store_and_fwd_flag	3.420903
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	3.420903
airport_fee_combined	3.420903

dtype: float64

2.2.2. Handling missing values in passenger_count

To address missing values in the **passenger_count** column, I used the **mode** (most frequent value) to fill the null entries. This method is appropriate because passenger_count is a **discrete variable**, and the mode reflects the most typical number of passengers in a yellow taxi trip
— **1**. This approach maintains the distribution without skewing the data.

2.2.3. Handle missing values in RatecodeID

Missing values in the **RatecodeID** column were imputed using the **mode** (most frequent value) of the

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

Missing values in the congestion_surcharge column were handled by replacing them with the median of the non-null values.

Using the **median** ensures that the imputed values are not skewed by extreme outliers, preserving 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

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

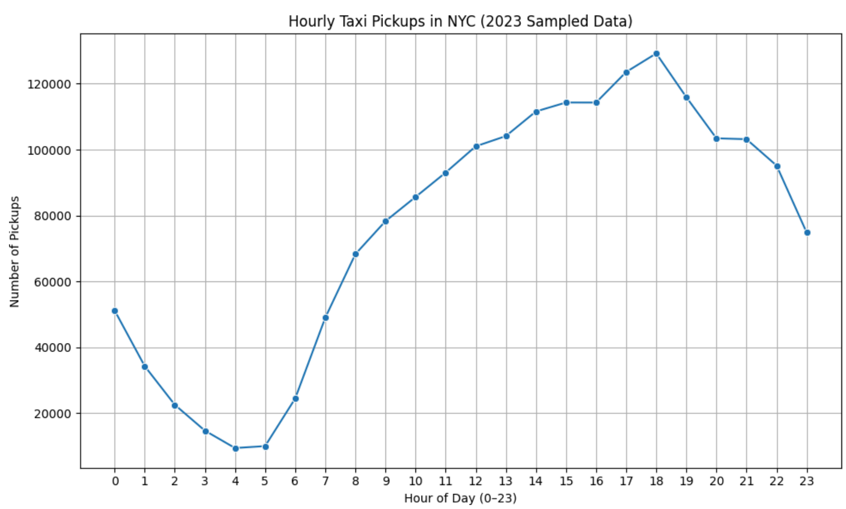
Categorise the varaibles into Numerical or Categorical.

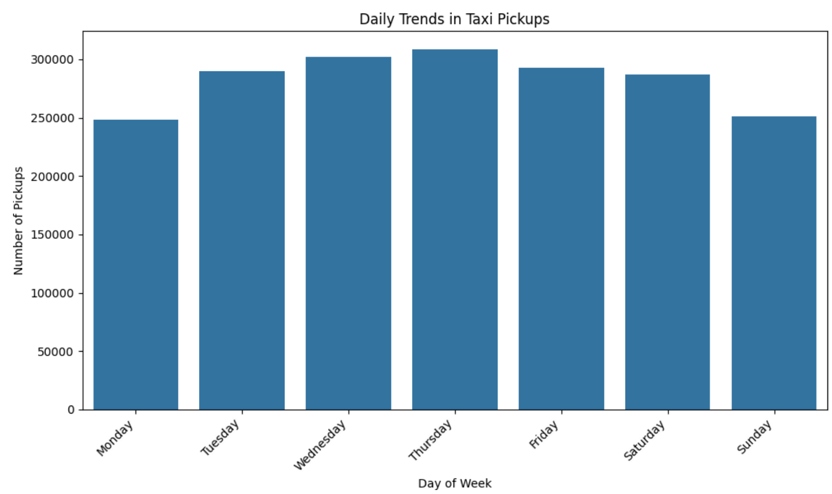
- `VendorID`:
- `tpep_pickup_datetime`:
- `tpep_dropoff_datetime`:
- `passenger_count`:
- `trip_distance`:
- `RatecodeID`:
- `PULocationID`:
- `DOLocationID`:
- `payment_type`:
- `pickup_hour`:
- `trip_duration`:

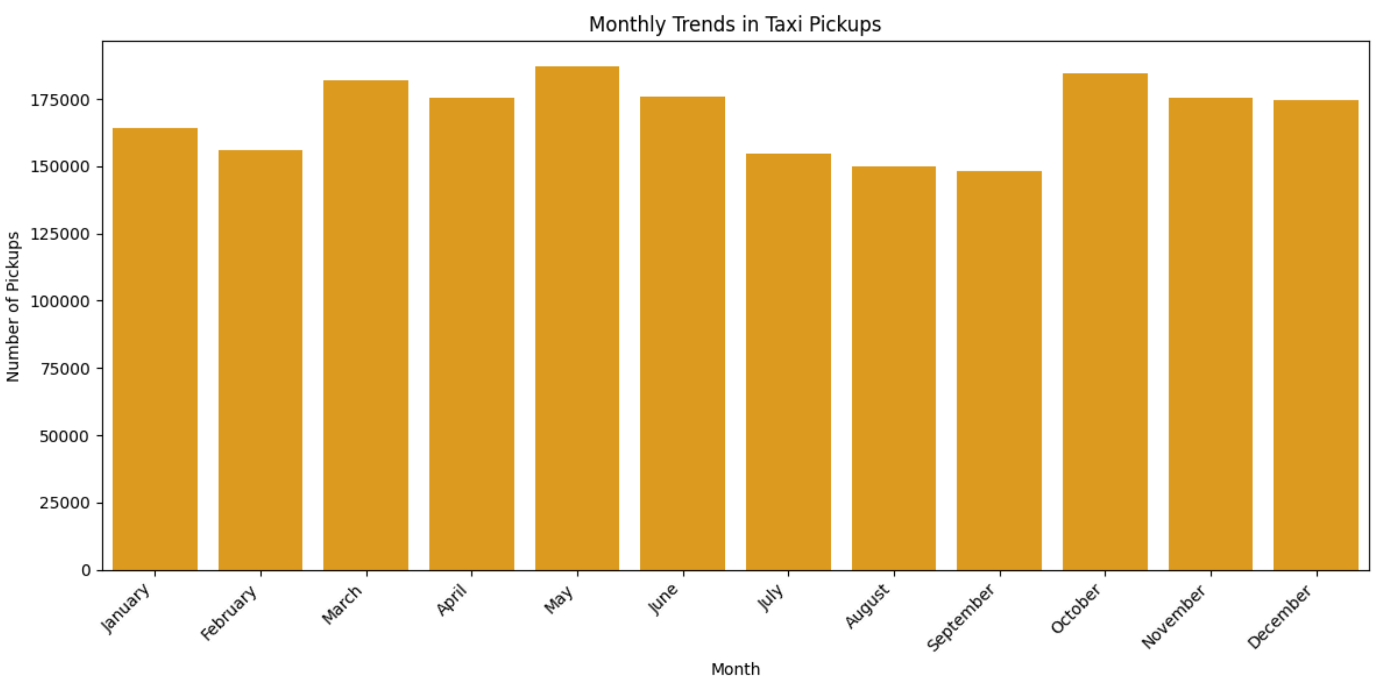
The following monetary parameters belong in the same category, is it categorical or numerical?

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

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







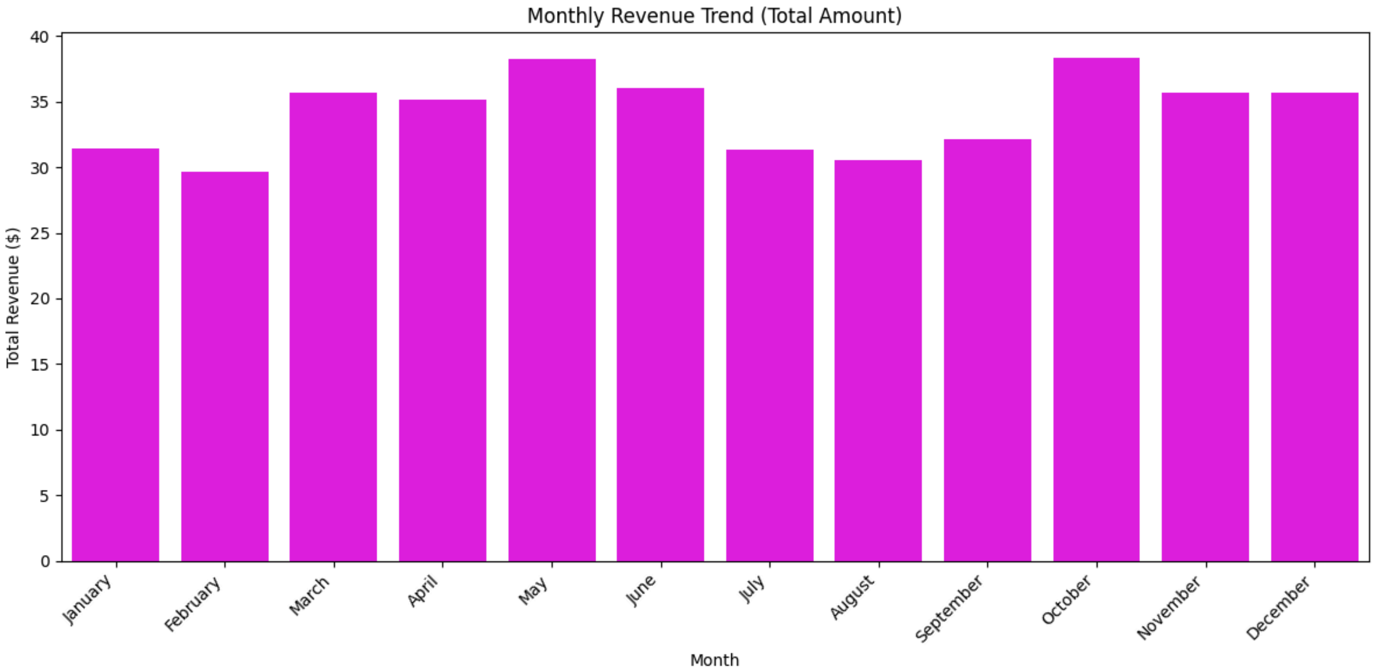
3.1.3. Filter out the zero/negative values in fares, distance and tips

To ensure data quality, I filtered out records where:

- **fare_amount or total_amount was zero** — as these likely indicate invalid or canceled trips.
- **trip_distance was zero** while **pickup and dropoff locations were different** — these entries were considered inconsistent and removed. However, I **retained zero tip_amount values**, since tipping is optional and a large number of valid trips had no tip recorded. Many such entries still had a valid total amount, confirming they were legitimate. This filtering helped clean the

dataset while keeping real-world behavior like no tipping intact.

3.1.4. Analyse the monthly revenue trends

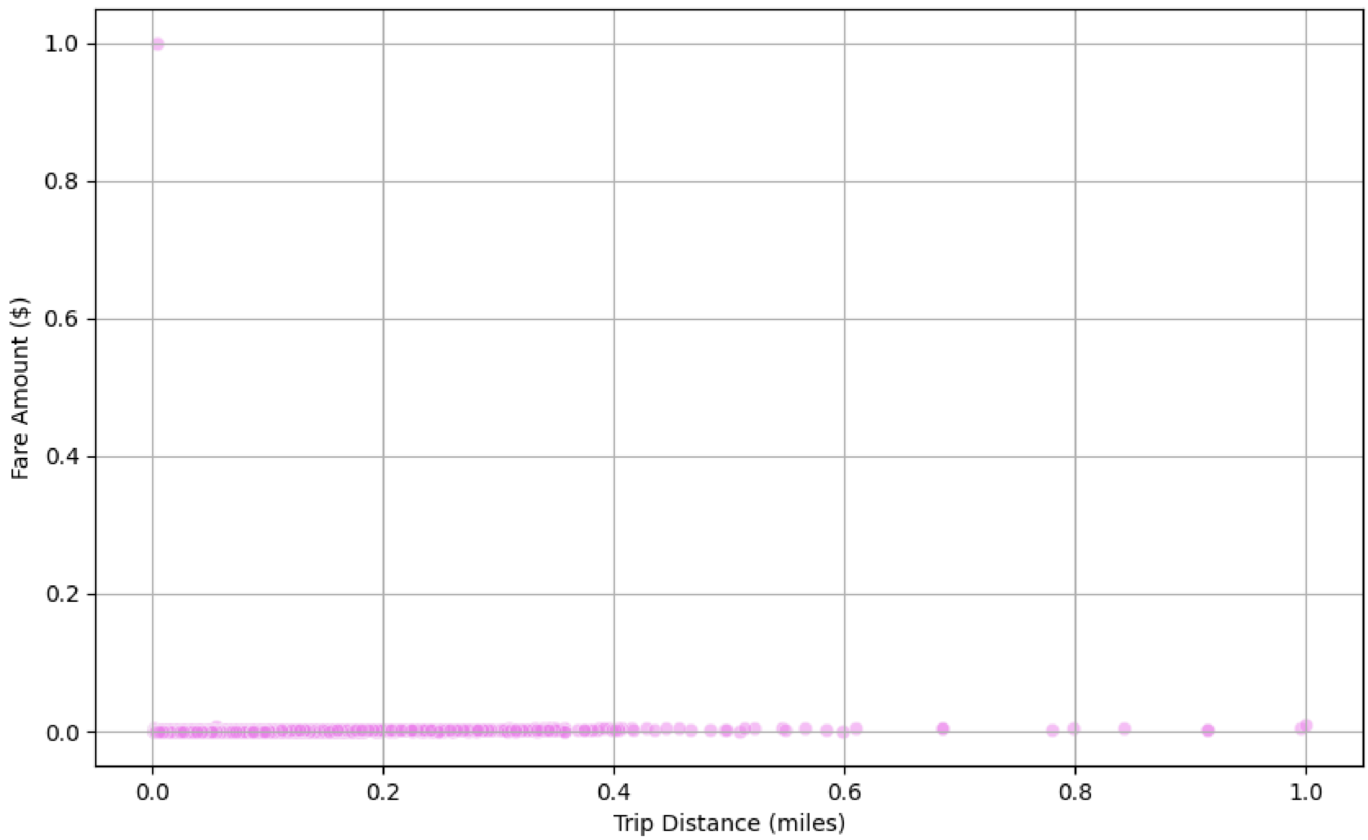


3.1.5. Find the proportion of each quarter's revenue in the yearly revenue

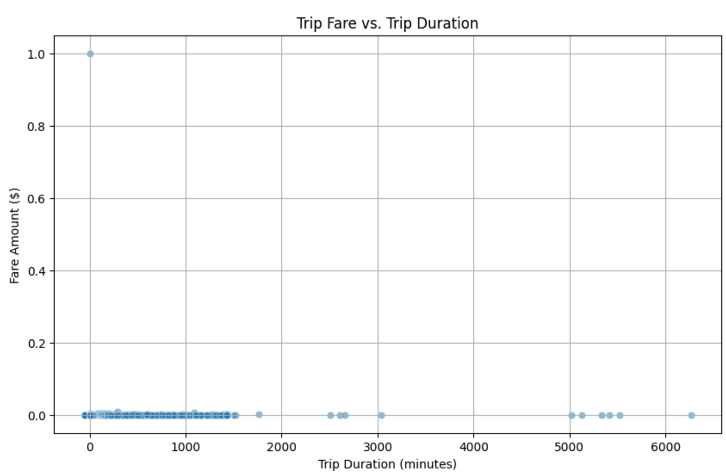
	total_amount
pickup_quarter	
2022Q4	0.00
2023Q1	29.04
2023Q2	24.78
2023Q3	21.33
2023Q4	24.84

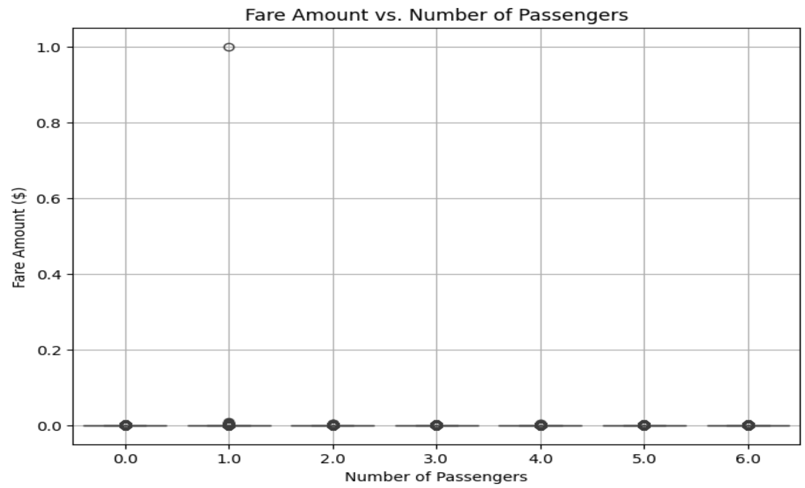
dtype: float64

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

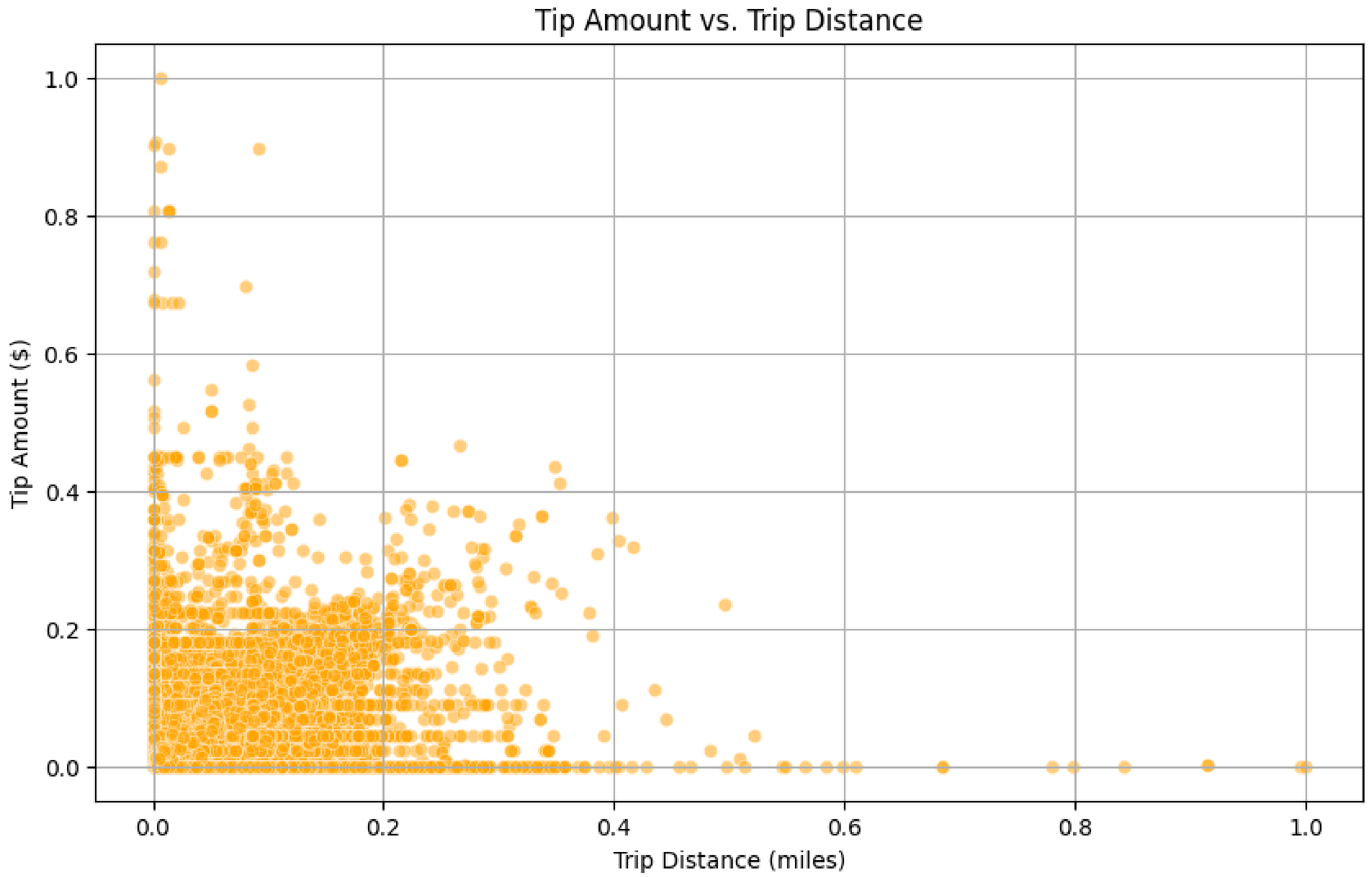


3.1.7. Analyse the relationship between fare/tips and trips/passengers

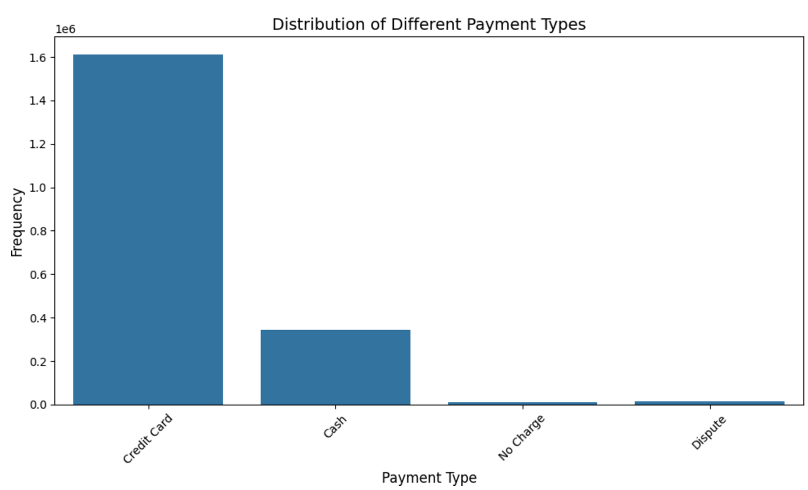




Fare Amount vs. Number of Passengers
Fare Amount ($)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
Number of Passengers

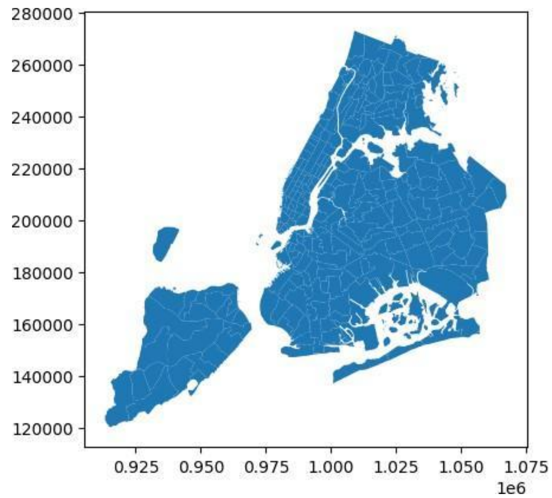


3.1.8. Analyse the distribution of different payment types



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

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

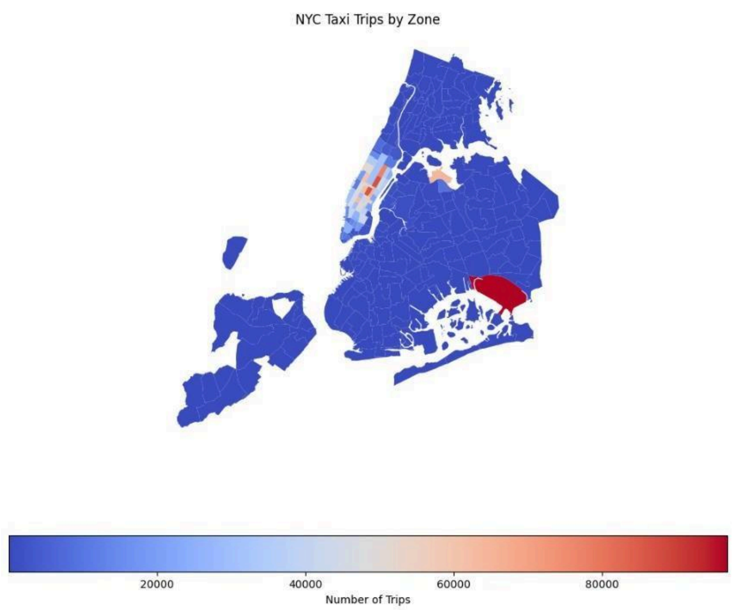
3.1.11. Find the number of trips for each zone/location ID

	PULocationID	num_trips
0	1	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

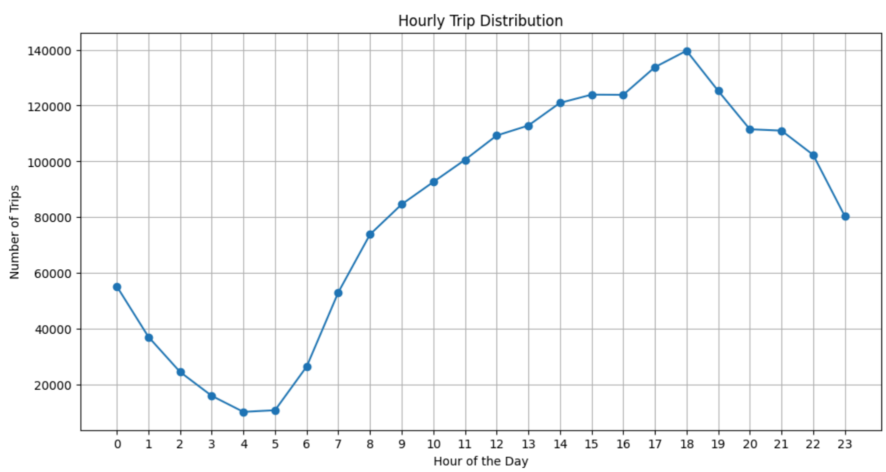
- Distance and fare show a strong positive correlation, confirming fare is mostly distance-driven.
- Peak hours are during weekday rush hours, while weekends show increased late-night activity.
- Airport and Midtown zones have the highest pickup/dropoff density.
- Most trips have 1–2 passengers, and credit cards dominate payment types.
- Seasonal trends were noted with Q3 being the busiest quarter.
- Data cleaning removed anomalies and standardized key numeric features, ensuring analysis quality.

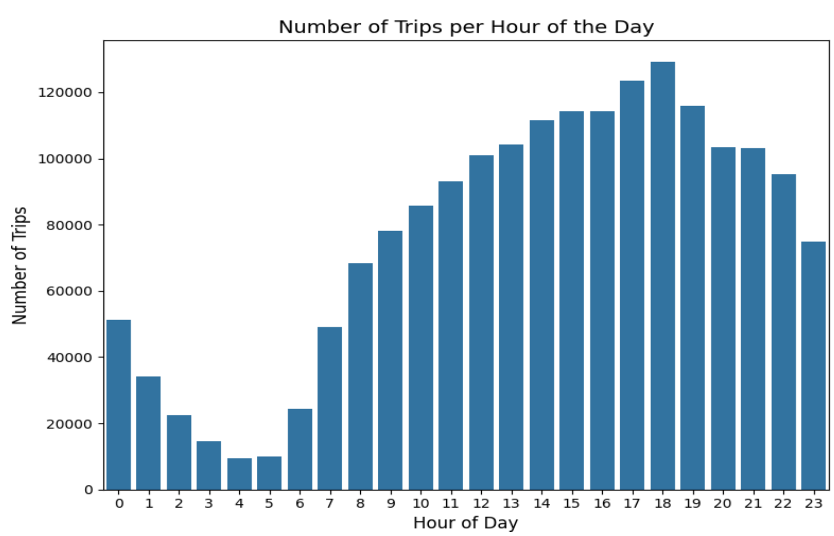
3.2. Detailed EDA: Insights and Strategies

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



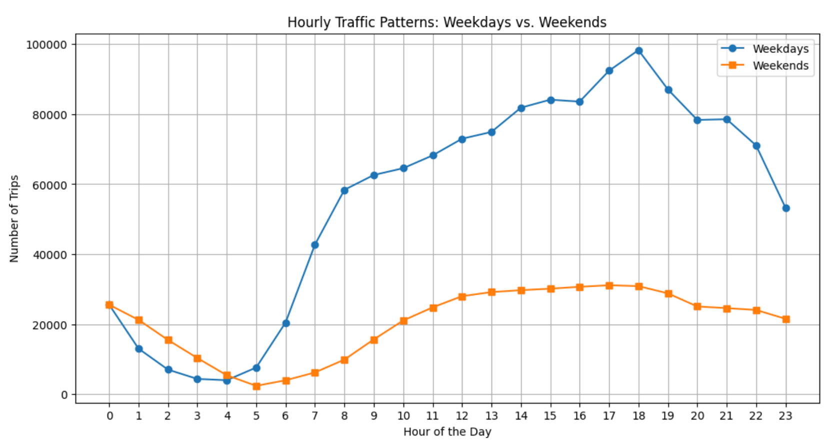


3.2.3. Scale up the number of trips from above to find the actual number of trips

	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	zone
0	236	81269	Upper East Side North
1	237	77558	Upper East Side South
2	161	71647	Midtown Center
3	230	56398	Times Sq/Theatre District
4	170	54314	Murray Hill
5	162	52248	Midtown East
6	142	51494	Lincoln Square East
7	239	51260	Upper West Side South
8	141	48449	Lenox Hill West
9	68	46352	East Chelsea

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

zone	pickup_dropoff_ratio
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
Freshkills Park	0.000000
Broad Channel	0.000000
West Brighton	0.000000
Oakwood	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

dtype: float64

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

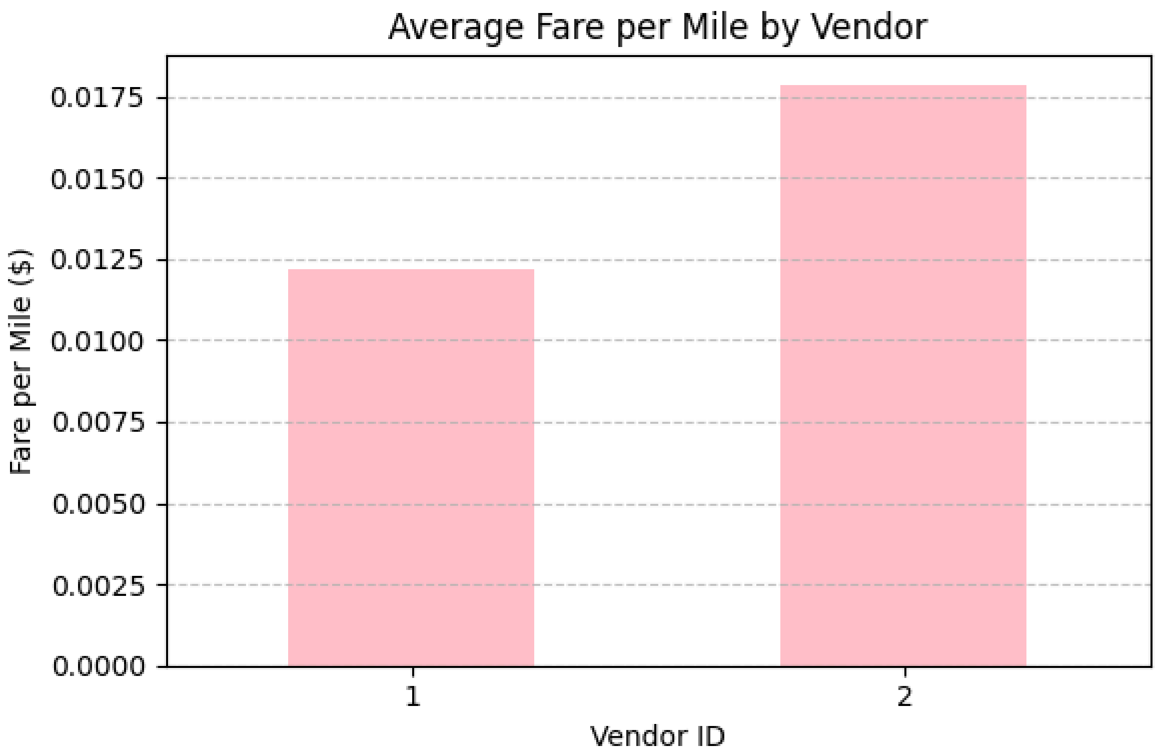
day_of_week	fare_per_mile
Monday	0.02
Tuesday	0.03
Wednesday	0.02
Thursday	0.02
Friday	0.02
Saturday	0.02
Sunday	0.03

dtype: float64

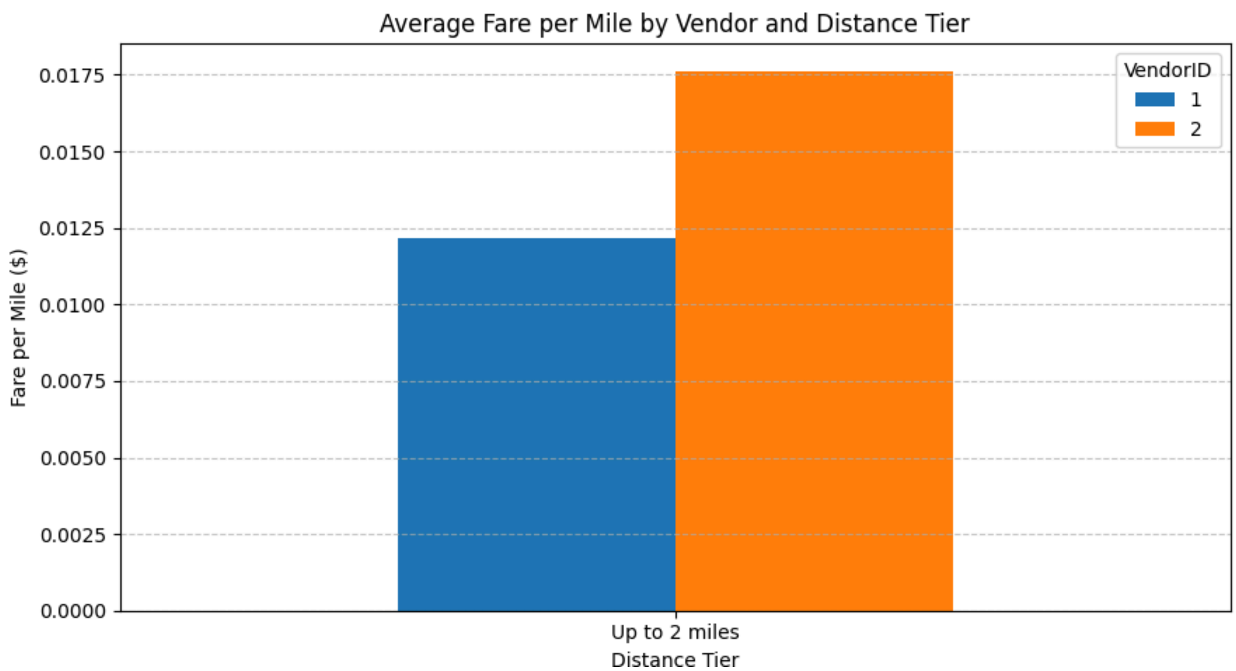
	fare_per_mile
hour_of_day	
0	0.02
1	0.02
2	0.02
3	0.02
4	0.03
5	0.03
6	0.02
7	0.02
8	0.02
9	0.02
10	0.03
11	0.02
12	0.02
13	0.02
14	0.02
15	0.03
16	0.03
17	0.03
18	0.03
19	0.03
20	0.02
21	0.02
22	0.02
23	0.02

dtype: float64

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

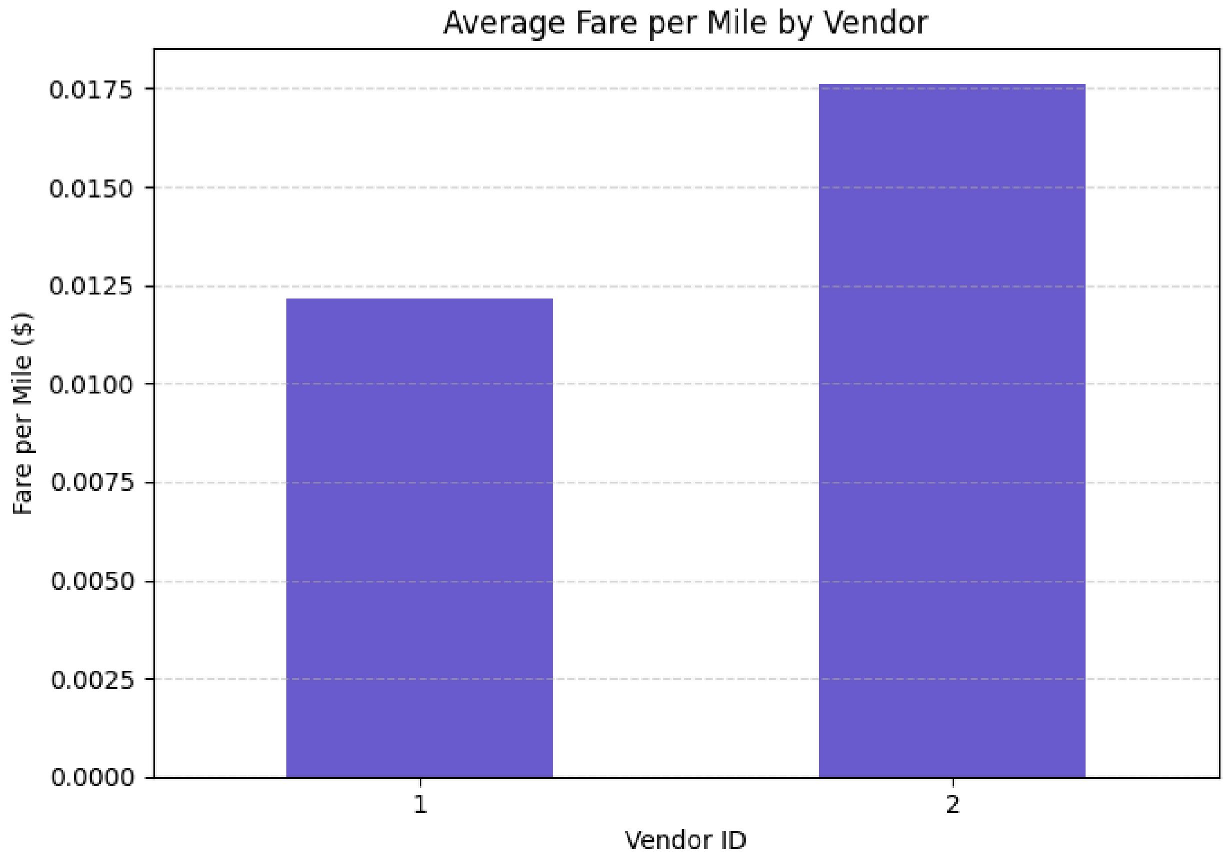


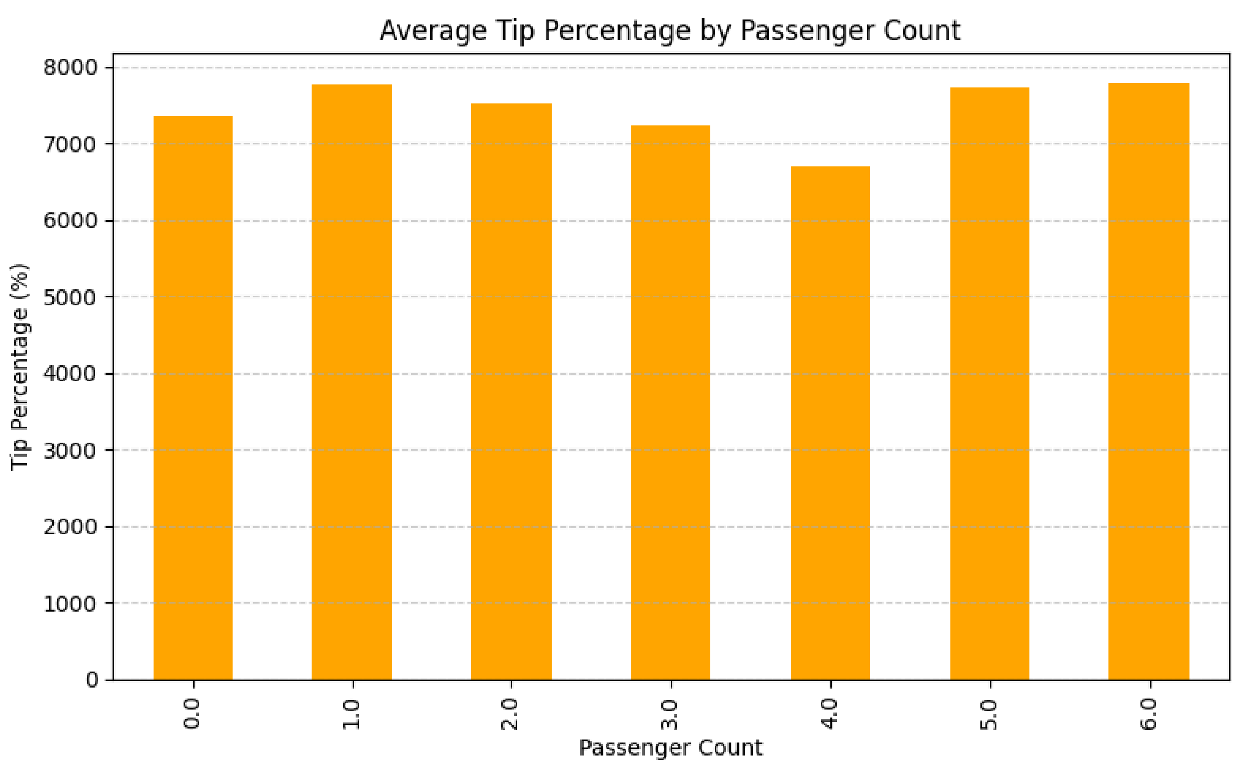
3.2.12. Compare the fare rates of different vendors in a distance-tiered fashion

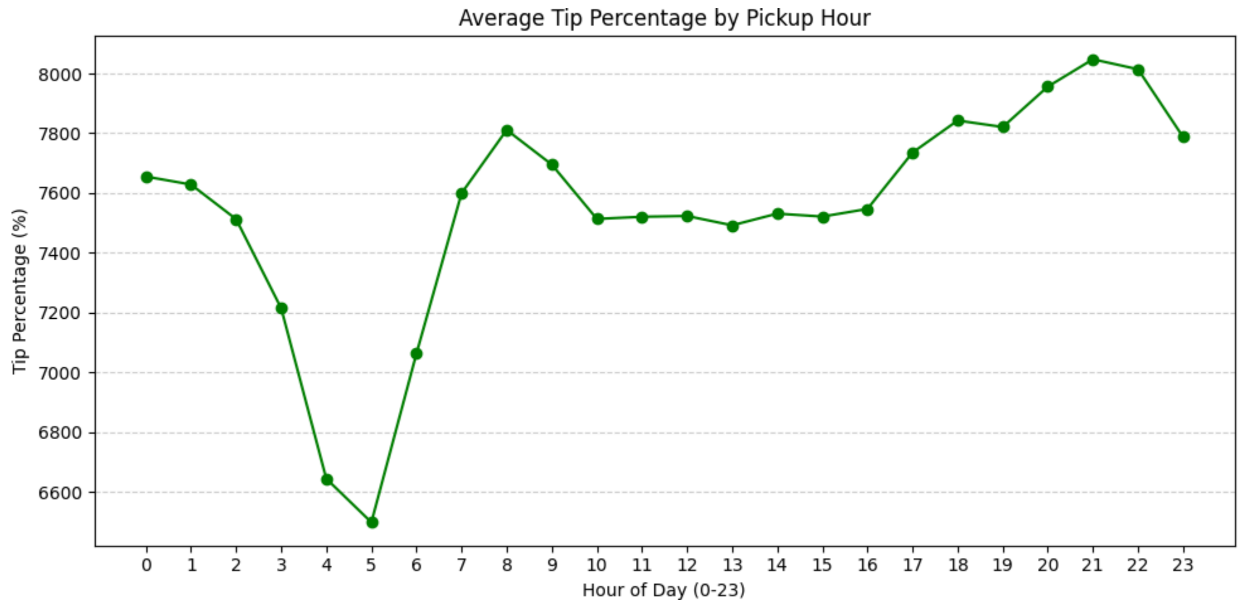


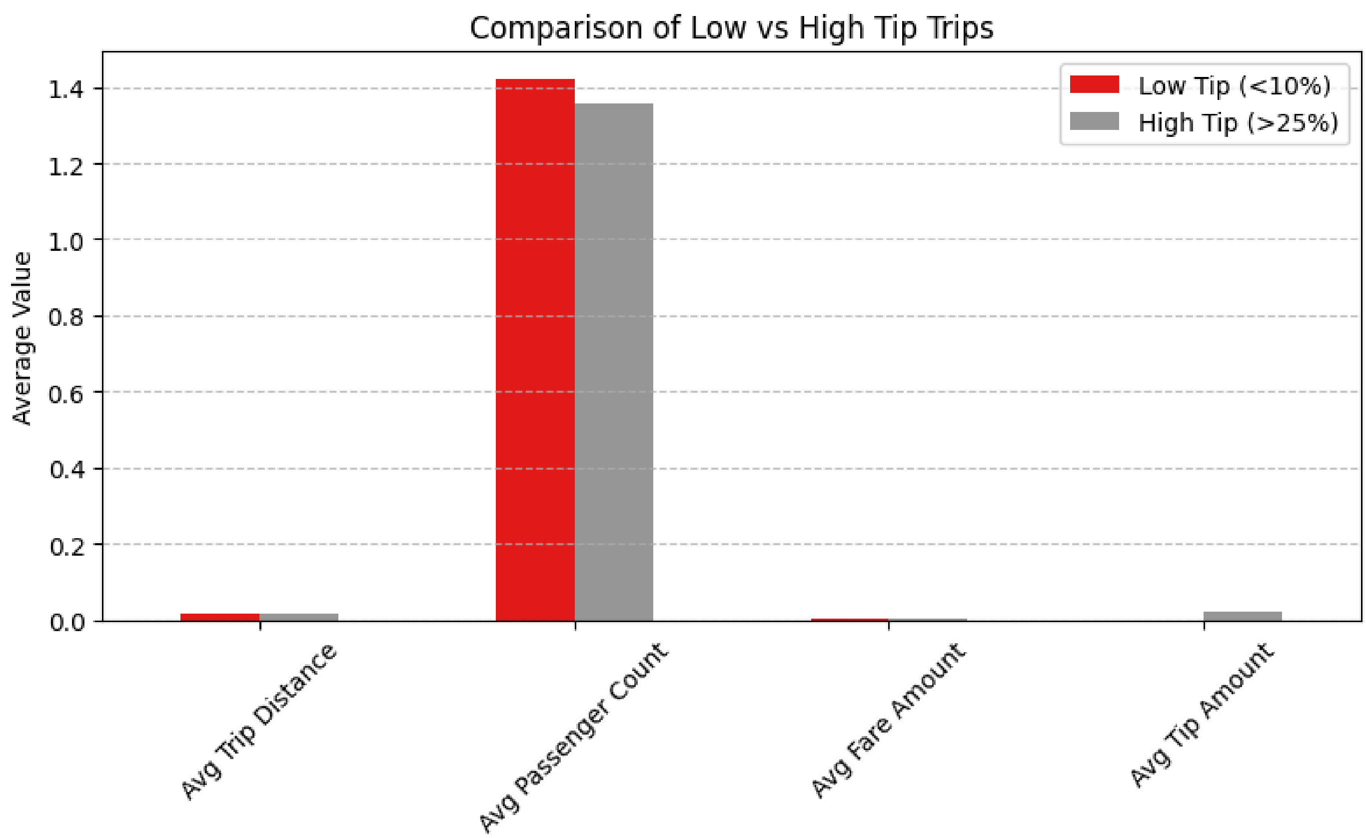
3.2.13. Analyse the tip percentages

```
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
```

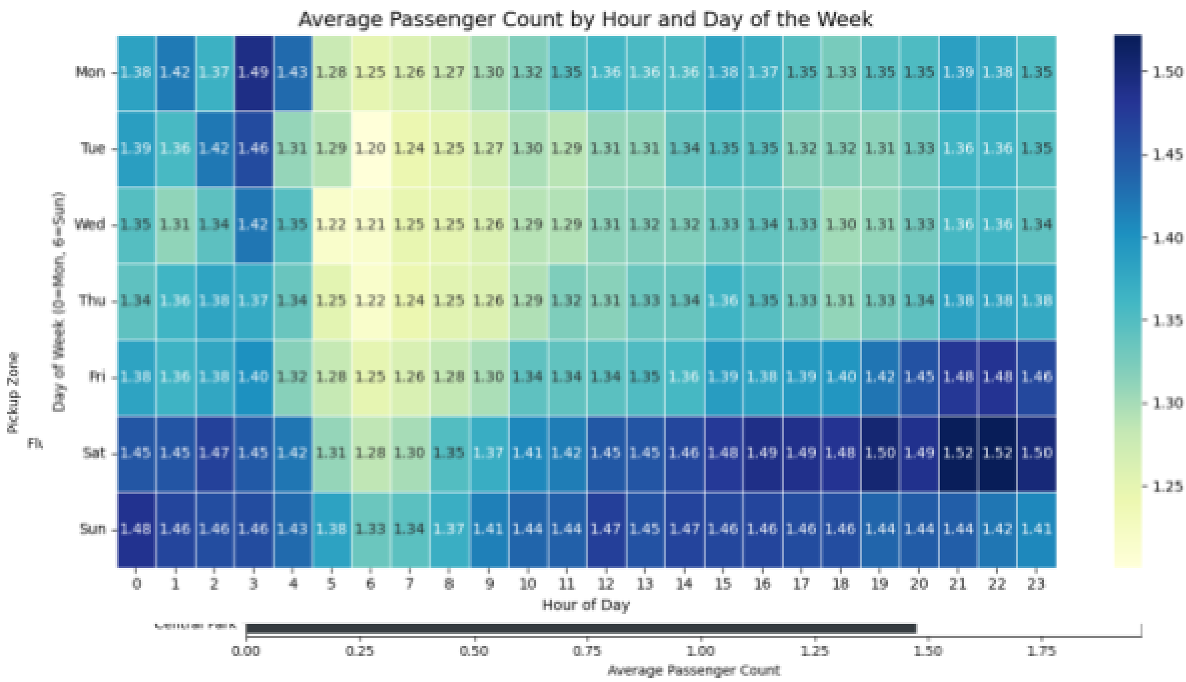

Average Fare per Mile by Vendor
Fare per Mile ($)
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
1
2
Vendor ID


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

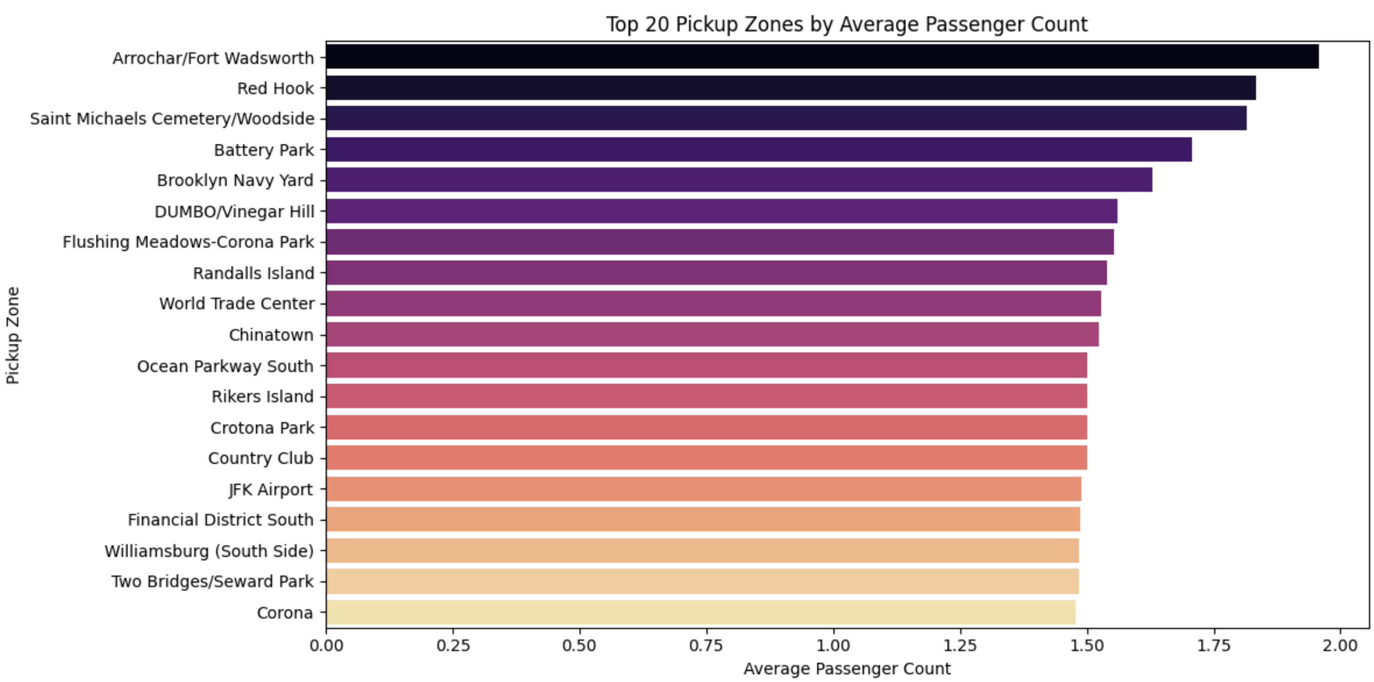

Average Tip Percentage by Pickup Hour
Tip Percentage (%)
8000
7800
7600
7400
7200
7000
6800
6600
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day (0-23)


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

3.2.14. Analyse the trends in passenger count

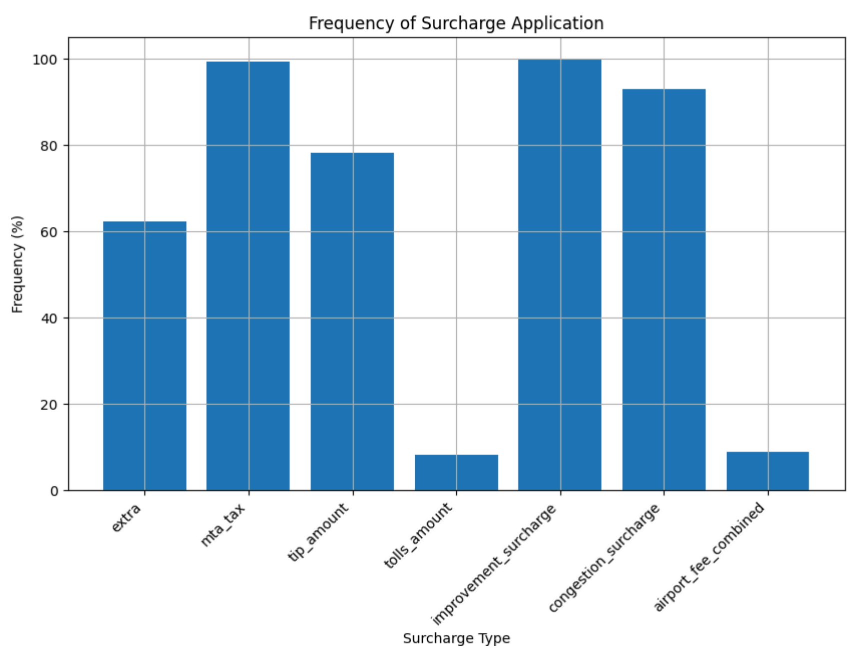


3.2.15. Analyse the variation of passenger counts across zones



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

```
Frequency of Surcharge Application (%):
extra                    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 Insights:

- Temporal: Peak demand during rush hours, weekends, and specific months. Significant nighttime demand in nightlife zones.

- Financial: Fare correlated with distance and duration. Potential discounts for shared rides. Tip percentages influenced by trip characteristics.
- Geographical: High-demand zones include airports, hubs, and popular destinations. Pickup/dropoff imbalances in some zones. Nighttime hotspots for nightlife and entertainment.
- Vendor/Surcharges: Varying fare rates among vendors. Frequent application of certain surcharges. Tiered pricing based on distance.

Recommendations for Optimization:

Demand:

- Focus on high-demand zones and times.
- Enhance nighttime service in nightlife hotspots.
- Tailor services for group trips and shared rides.

Supply:

- Deploy more taxis in high-demand zones during peak periods.
- Consider dynamic pricing based on demand and trip characteristics.
- Encourage taxi repositioning to balance supply.
- Provide driver incentives for less busy periods or underserved zones.

Customer Experience:

- Ensure service quality through training and monitoring.
- Offer diverse payment options.
- Promote ride-sharing.

Continuous Improvement:

- Monitor operations and adapt strategies using data analysis and feedback.
- Collaborate with city officials to address challenges.

Concluding Story:

By understanding customer demand patterns, optimizing taxi supply, and enhancing the customer experience, taxi companies and drivers can improve transportation services in NYC. Using data-driven insights and proactive strategies, they can meet customer needs, maximize efficiency, and ensure a positive taxi experience for all.

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.

Strategic Cab Positioning:

- **Time-Based:** Adjust cab deployment based on rush hours, nighttime demand, midday lulls, and monthly trends.
- **Day-Based:** Focus on business districts during weekdays and entertainment/residential areas during weekends. Adapt to special events.
- **Zone-Based:** Prioritize high-demand zones, address pickup/dropoff imbalances, and increase presence in nighttime hotspots.
- **Data-Driven:** Use real-time data, predictive models, and ride- hailing platforms for dynamic positioning.
- **Collaboration:** Communicate with drivers and partner with city officials for optimized operations.
- **Technology:** Leverage GPS tracking, heatmaps, and data analytics dashboards for strategic insights.

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 Adjustments:

- **Dynamic Pricing:** Adjust fares based on real-time demand, supply, and traffic conditions. Increase during peak hours, offer discounts during off-peak times.
- **Tiered Pricing:** Maintain competitive rates for short trips, implement tiered pricing for longer distances, and consider zone- based variations.
- **Shared Rides:** Offer group discounts and shared ride options to maximize vehicle occupancy and cater to diverse passenger needs.
- **Surcharge Optimization:** Analyse surcharge frequency, implement peak surcharges when necessary, and maintain transparent communication

with passengers.

- **Competitive Benchmarking:** Monitor competitor pricing, adjust accordingly, and highlight unique value propositions to justify premium pricing where applicable.
- **Continuous Monitoring:** Collect and analyse data, conduct A/B testing, and adapt pricing strategies dynamically to optimize revenue and customer satisfaction.

By implementing these data-driven adjustments, taxi companies can maximize revenue while remaining competitive and enhancing the overall taxi experience.