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

As per the given instructions, I first extracted 500,000 records from each monthly Parquet file. Later, I reduced the total sample size so that the combined DataFrame contained about 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 included two columns with nearly identical names — airport_fee and Airport_fee — which likely resulted from inconsistent naming across monthly files. To address this issue, I generated a new column named airport_fee_combined by selecting the maximum value between the two for each record to preserve all information. After the merge, I removed the original columns to eliminate duplication.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

	<unnamed>
VendorID	0.00000
tpep_pickup_datetime	0.00000
tpep_dropoff_datetime	0.00000
passenger_count	3.42090
trip_distance	0.00000
RatecodeID	3.42090
store_and_fwd_flag	3.42090
PULocationID	0.00000
DOLocationID	0.00000
payment_type	0.00000
fare_amount	0.00000
extra	0.00000
mta_tax	0.00000
tip_amount	0.00000
tolls_amount	0.00000
improvement_surcharge	0.00000
total_amount	0.00000
congestion_surcharge	3.42090
airport_fee_combined	3.42090

2.2.2. Handling missing values in passenger_count

To handle the missing values in the passenger_count column, I filled the null entries using the mode, representing the most common value. Since passenger_count is a discrete variable, the mode — typically 1 for yellow taxi trips — effectively preserves the natural distribution of the data without introducing bias.

2.2.3. Handle missing values in RatecodeID

Missing values in the RatecodeID column were filled using the mode, which represents the most frequent category. This method is appropriate for categorical variables such as RatecodeID, as it helps retain the dominant pattern in the dataset while minimizing the influence of rare or outlier values.

2.2.4. Impute NaN in congestion_surcharge

Null values in the congestion_surcharge column were addressed by substituting them with the median from existing non-null entries. The median approach prevents distortion from extreme values, maintaining the original distribution characteristics of the column.

2.3. Handling Outliers and Standardising Values

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

Payment Type: Records with payment_type equal to 0 were identified as invalid since this value does not correspond to a recognized payment code. These entries were removed from the dataset.

Trip Distance: Outliers were detected in cases of unusually long or suspiciously short trips. Trips with a distance below 0.1 miles but a fare exceeding $300 were excluded. Similarly, trips with distances greater than 250 miles were removed as extreme cases. Additionally, records with zero distance and fare but differing pickup and drop-off locations were considered invalid and eliminated.

Tip Amount: Zero values in the tip_amount column were retained, as tipping is optional. Extremely high tip values were managed indirectly through min-max scaling, which normalized the data between 0 and 1, thereby reducing the influence of extreme outliers.

3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

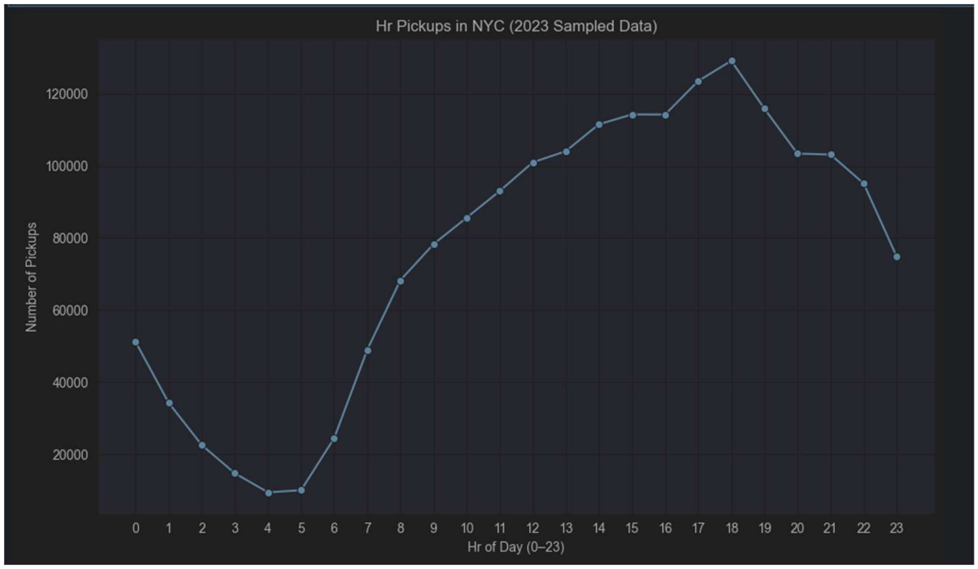
3.1.1. Classify variables into categorical and numerical

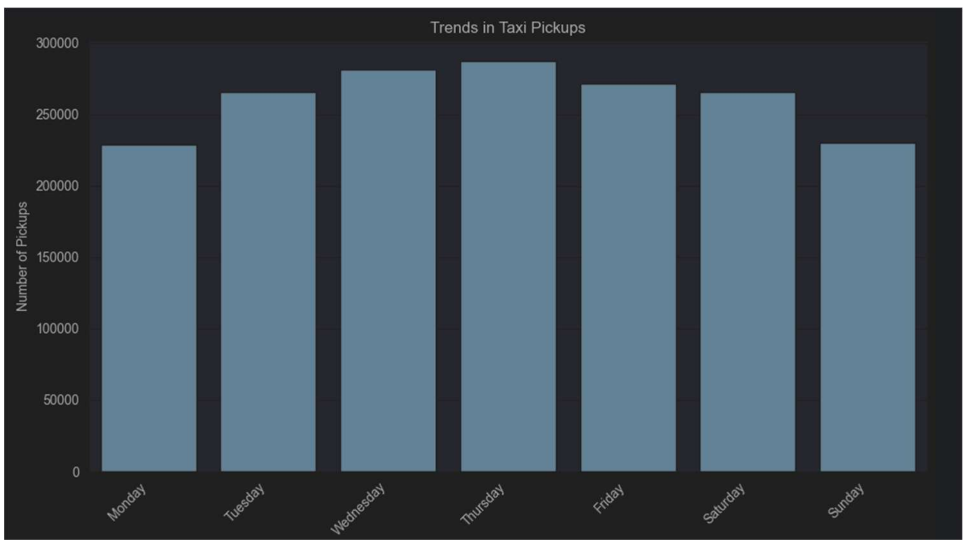
- `VendorID`:
- `tpep_pickup_datetime`:
- `tpep_dropoff_datetime`:
- `passenger_count`:
- `trip_distance`:
- `RatecodeID`:
- `PULocationID`:
- `DOLocationID`:
- `payment_type`:
- `pickup_hour`:
- `trip_duration`:

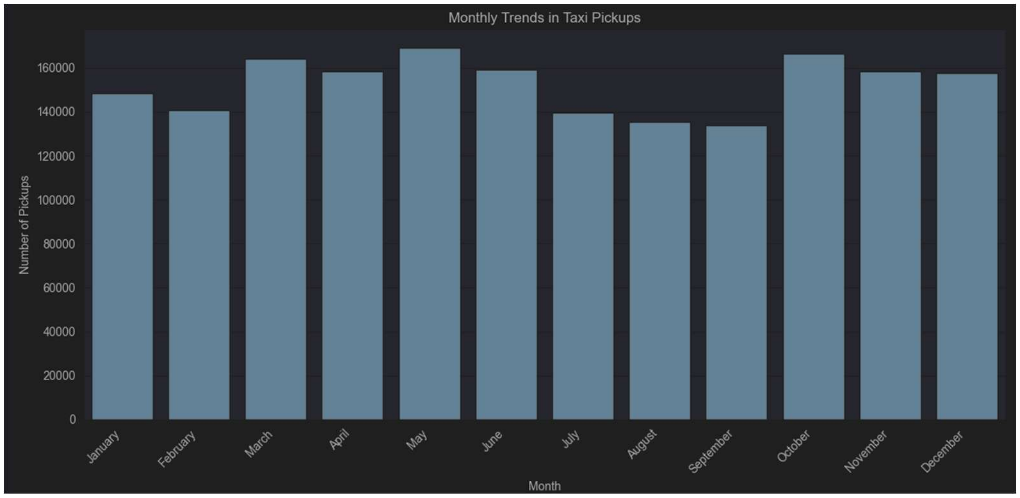
The following monetary parameters belong in the same category, is it categorical or numerical?

- `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


Hr Pickups in NYC (2023 Sampled Data)
Number of Pickups
120000
100000
80000
60000
40000
20000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hr of Day (0–23)


Trends in Taxi Pickups
Number of Pickups
300000
250000
200000
150000
100000
50000
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday



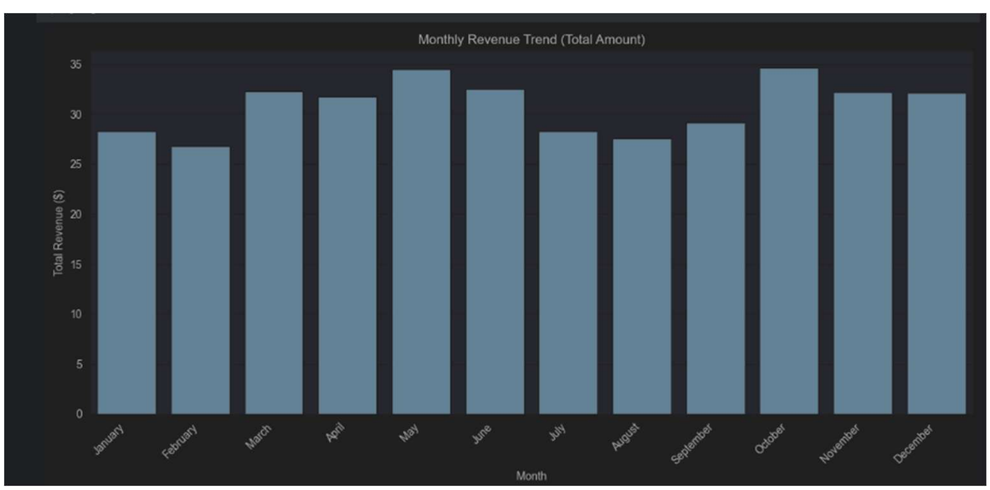
3.1.3. Filter out the zero/negative values in fares, distance and tips

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.

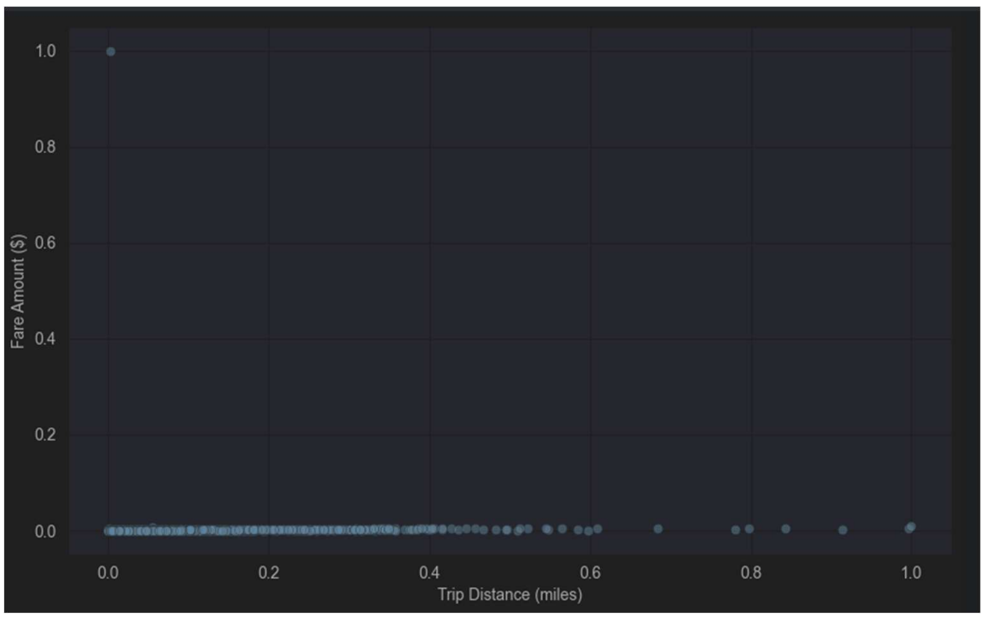
3.1.4. Analyse the monthly revenue trends



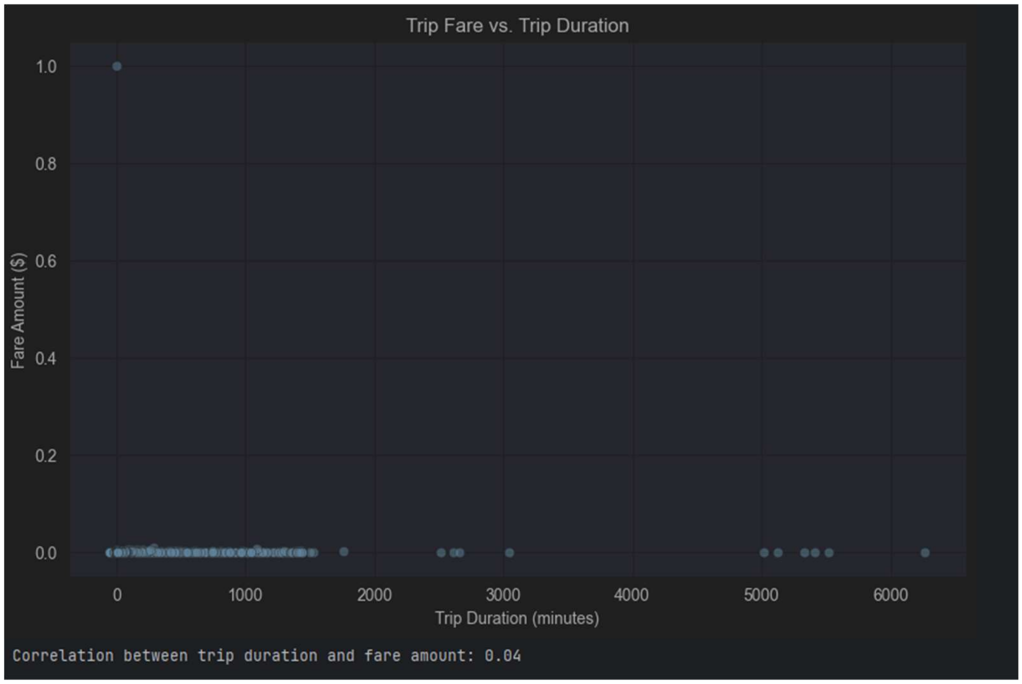
3.1.5. Find the proportion of each quarter's revenue in the yearly revenue

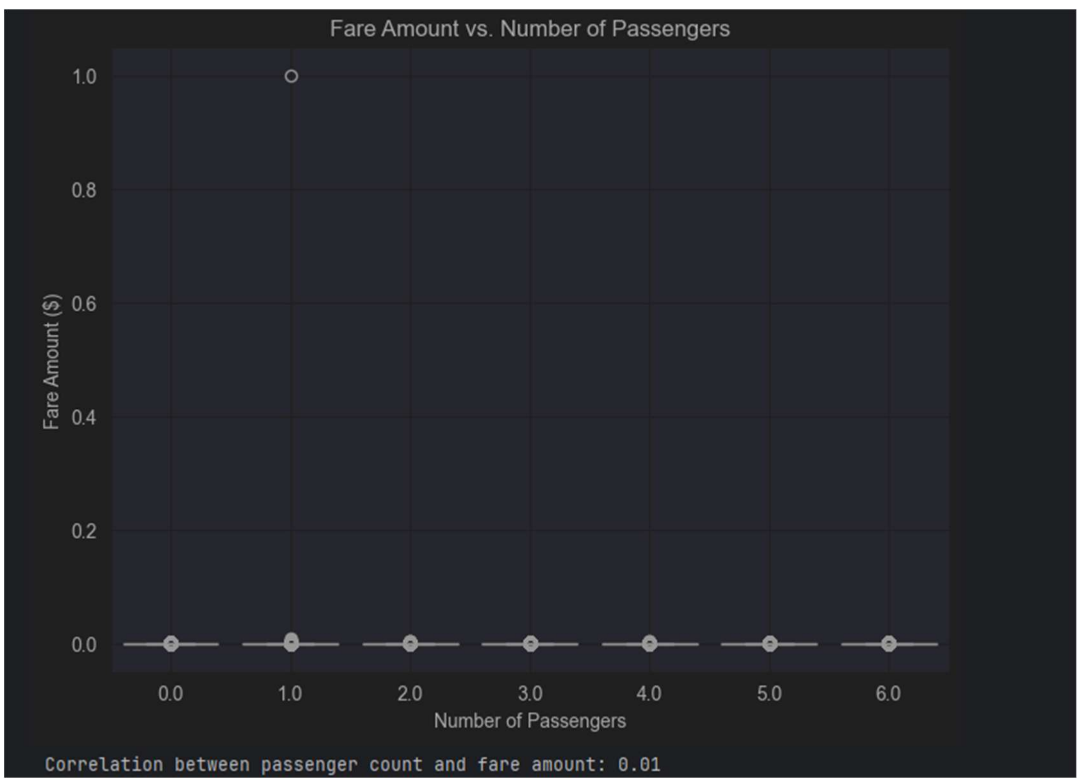
pickup_quarter	total_amount
2022Q4	0.00
2023Q1	23.61
2023Q2	26.68
2023Q3	22.97
2023Q4	26.74

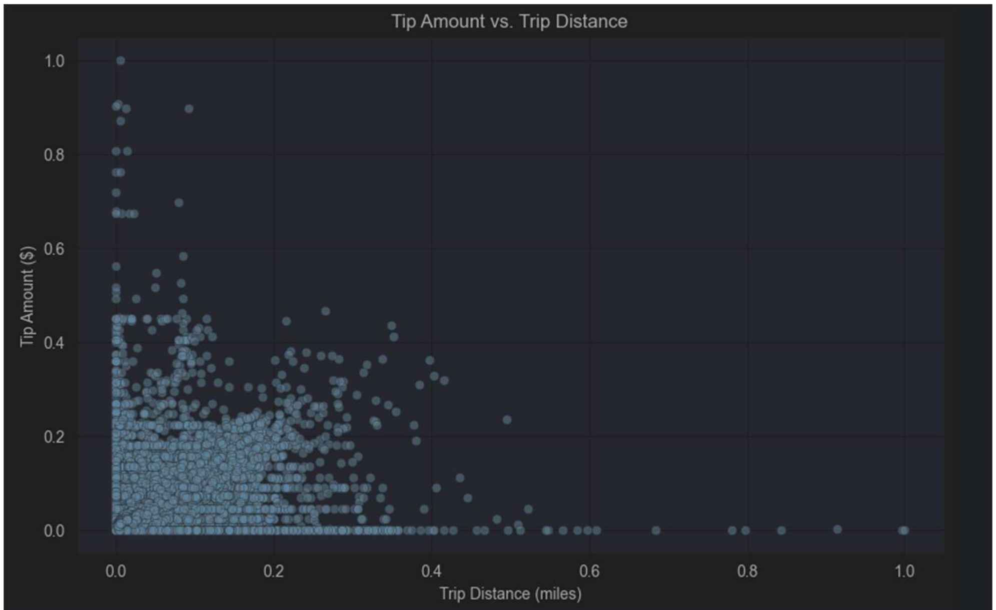
3.1.6. Analyse and visualise the relationship between distance and fare amount



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

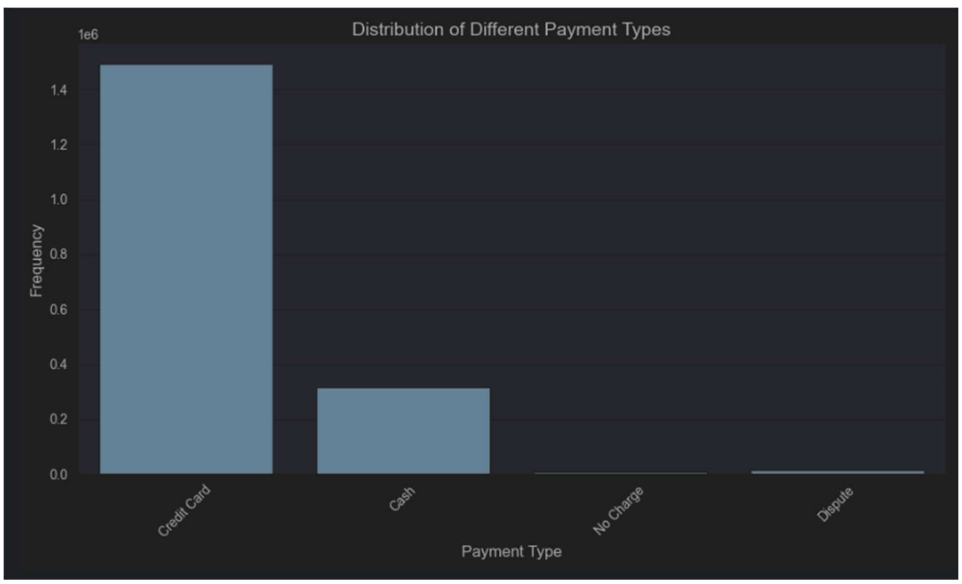






Correlation between trip distance and tip amount: 0.57

3.1.8. Analyse the distribution of different payment types

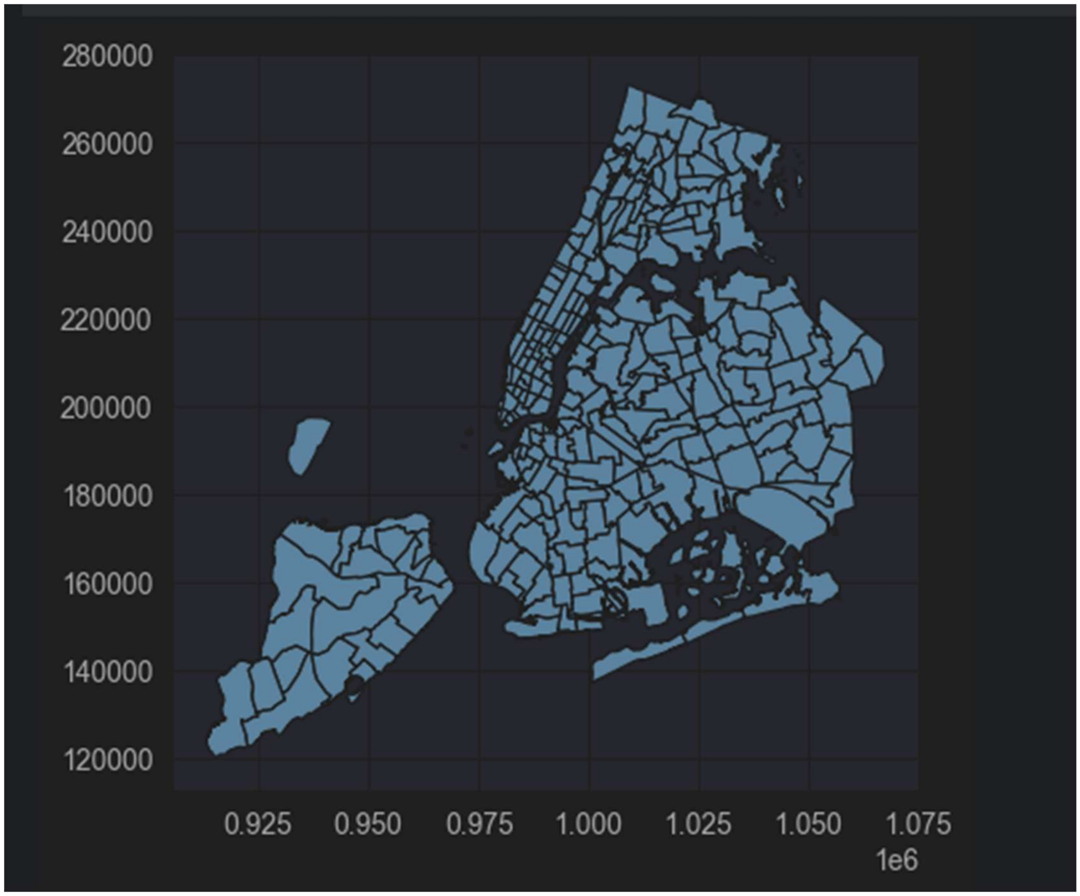


3.1.9. Load the taxi zones shapefile and display it

5 rows 5 rows × 7 cols

	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 192572.175, 933088.585…
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 1033439.643 170883.946, 10…
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 256638.616, 1026567.2…
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 203711.502, 992061.716…
4	5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144173.418, 936387.922 …

Now, if you look at the DataFrame created, you will see columns like: `OBJECTID`,`Shape_Leng`, `Shape_Area`, `zone`, `LocationID`, `borough`, `geometry`.



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

5 rows ⌄ 5 rows × 2 cols

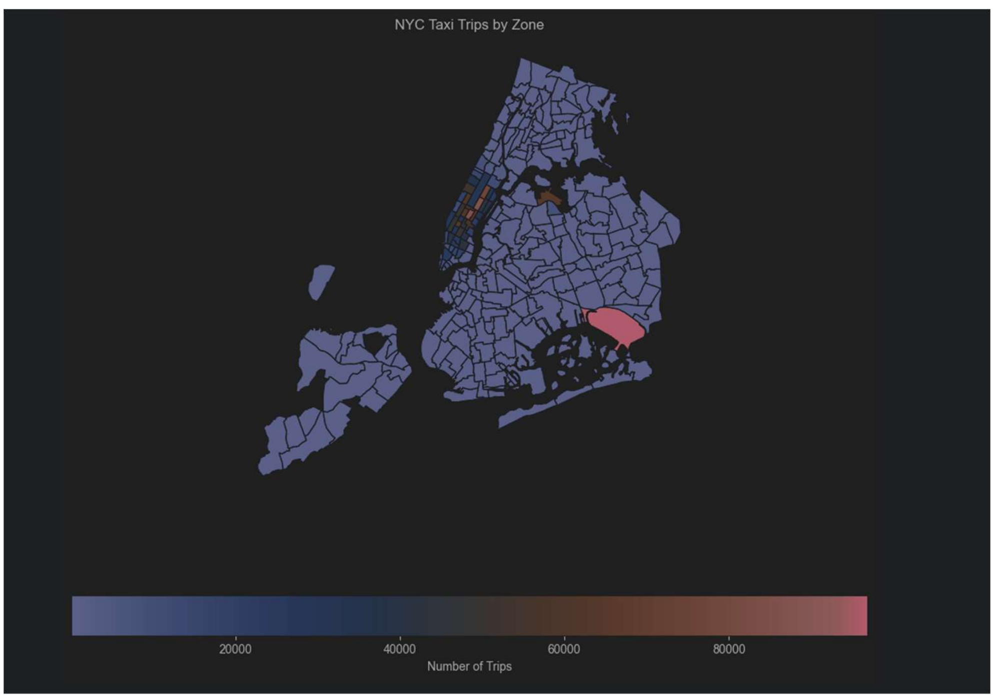
	PULocationID	num_trips
0	1	214
1	2	2
2	3	40
3	4	1861
4	5	13

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

5 rows ⌄ 5 rows × 9 cols

	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 192572.175, 933088.585…
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 1033439.643 170883.946, 10…
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 256638.616, 1026567.2…
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 203711.502, 992061.716…
4	5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144173.418, 936387.922 …

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.

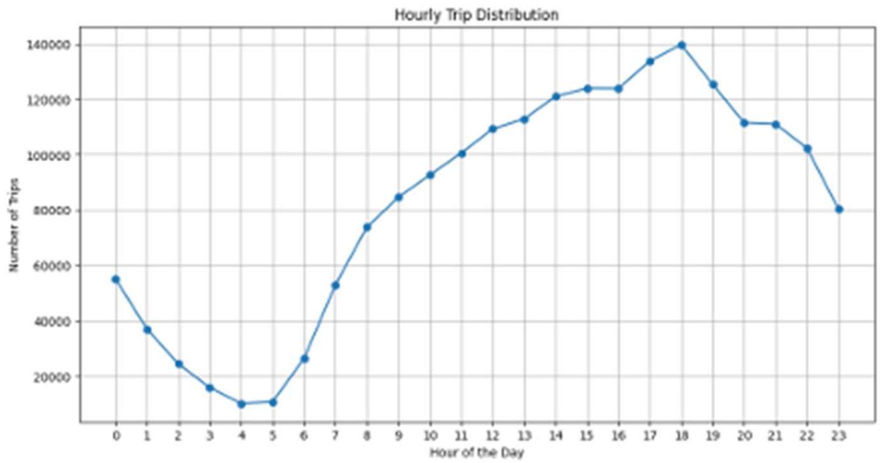
• Data cleaning removed anomalies and standardized key numeric features, ensuring analysis quality.

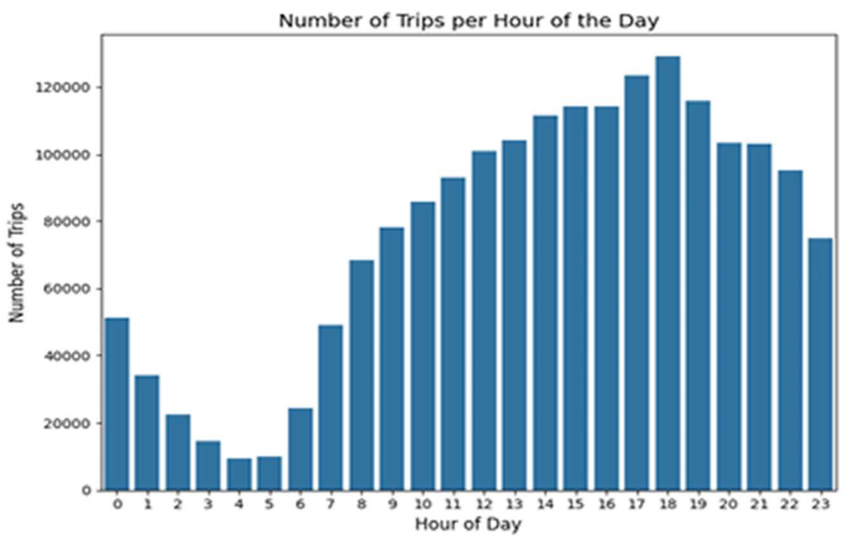
3.2. Detailed EDA: Insights and Strategies

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



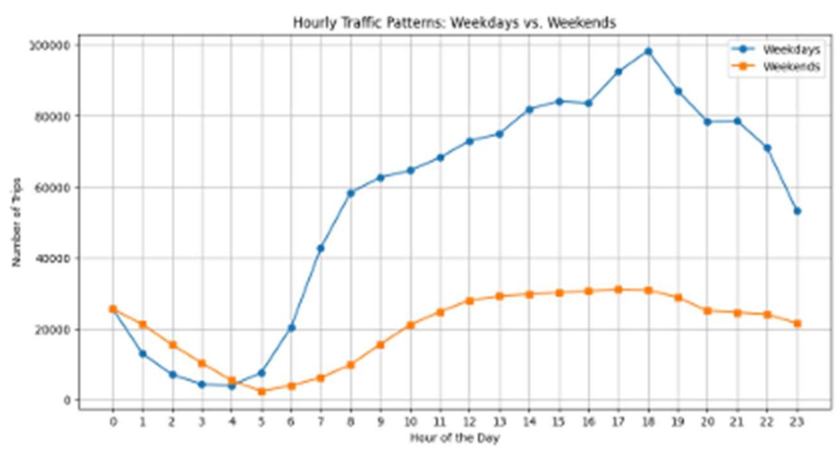


3.2.3. Scale up the number of trips from above to find the actual number of trips

pickup_hour	count
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

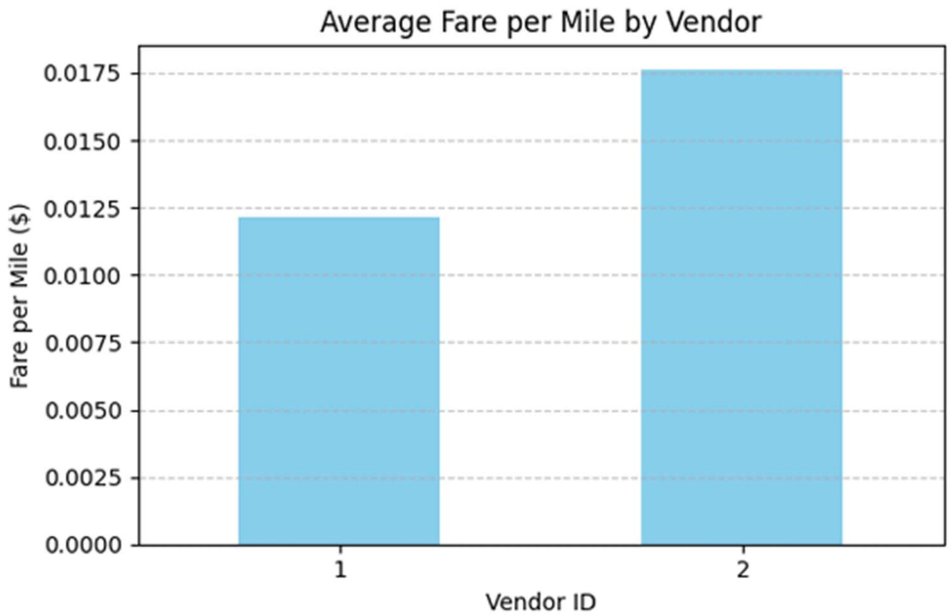
day_of_week	fare_per_mile
Monday	0.02
Tuesday	0.03
Wednesday	0.02
Thursday	0.02
Friday	0.02
Saturday	0.02
Sunday	0.03

dtype: float64

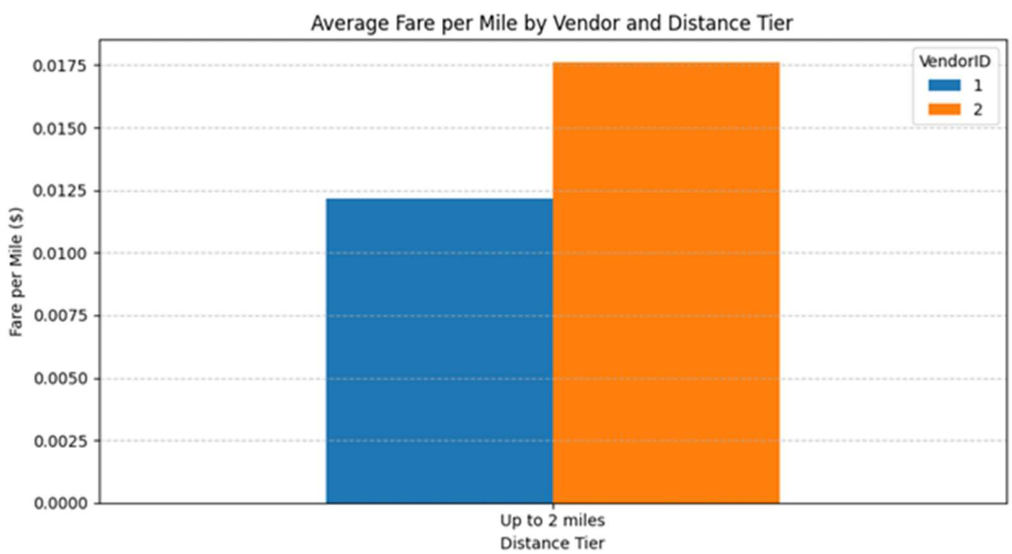
hour_of_day	fare_per_mile
0	0.02
1	0.02
2	0.02
3	0.02
4	0.03
5	0.03
6	0.02
7	0.02
8	0.02
9	0.02
10	0.03
11	0.02
12	0.02
13	0.02
14	0.02
15	0.03
16	0.03
17	0.03
18	0.03
19	0.03
20	0.02
21	0.02
22	0.02
23	0.02

dtype: float64

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

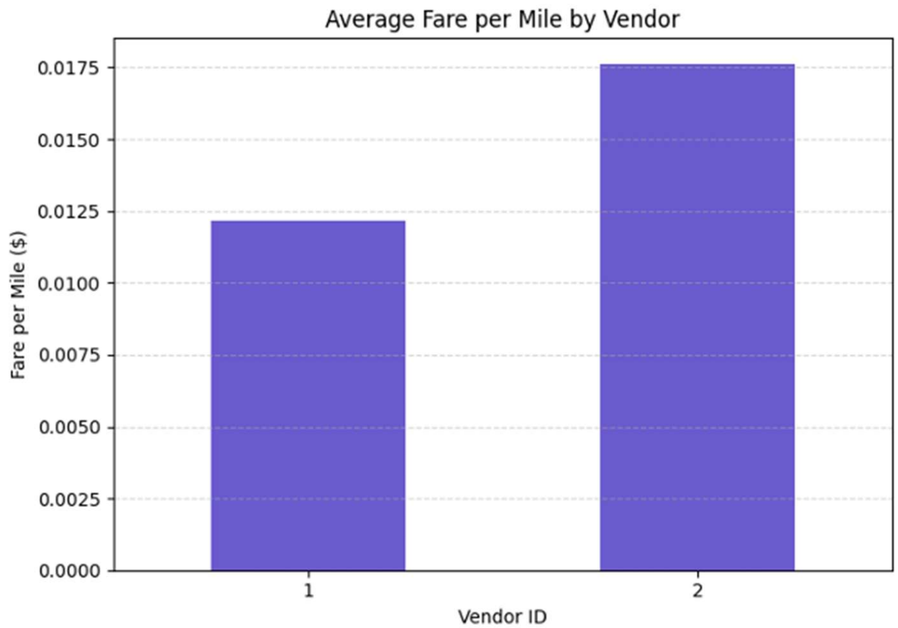


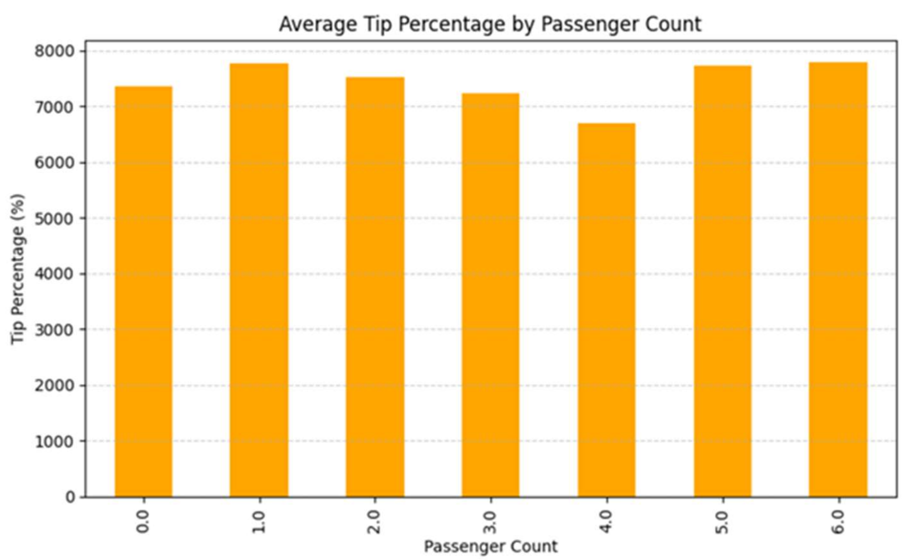
3.2.12. Compare the fare rates of different vendors in a distance-tiered fashion

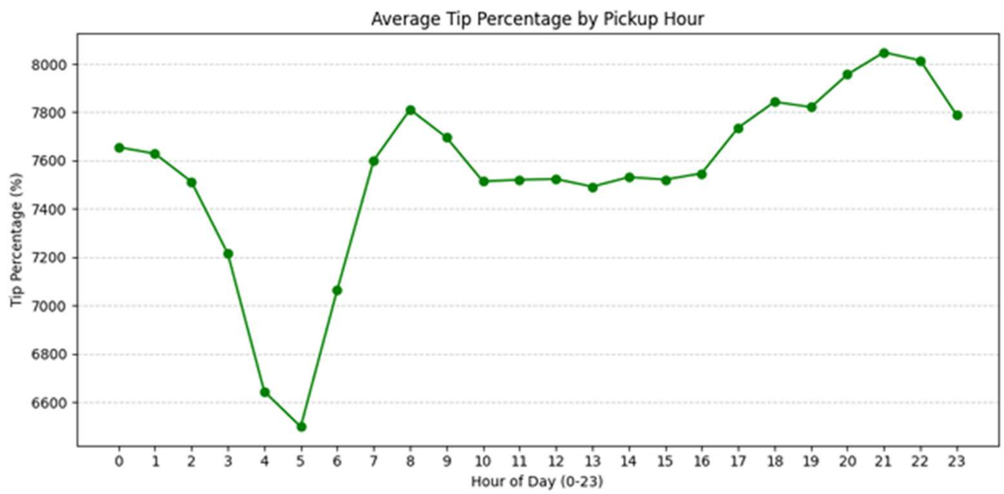


3.2.13. Analyse the tip percentages

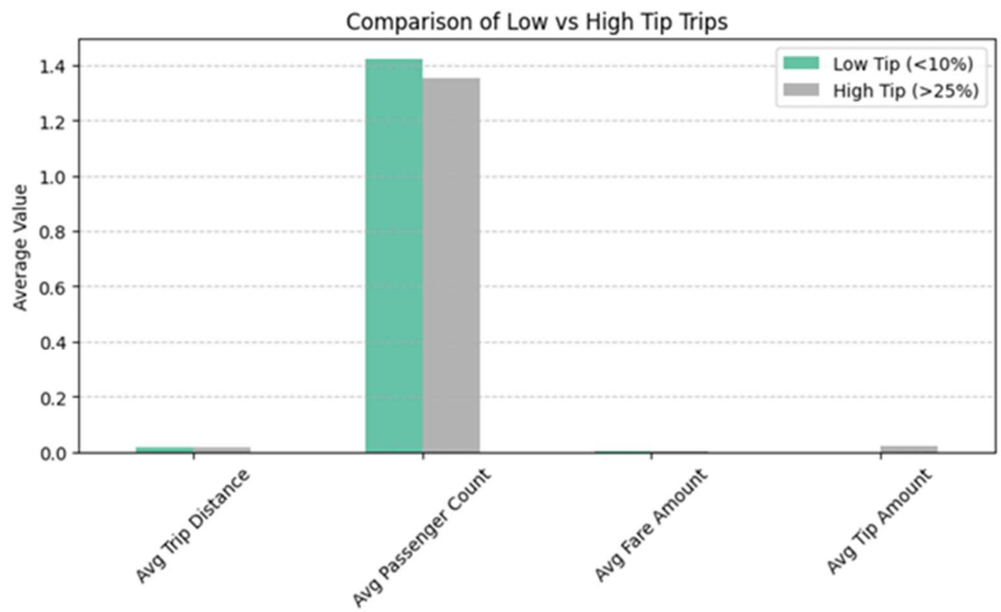
```
Average Tip Percentage by Distance:
distance_category
Up to 2 miles        7676.350688
2 to 5 miles                 NaN
More than 5 miles            NaN
Name: tip_percentage, dtype: float64

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


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

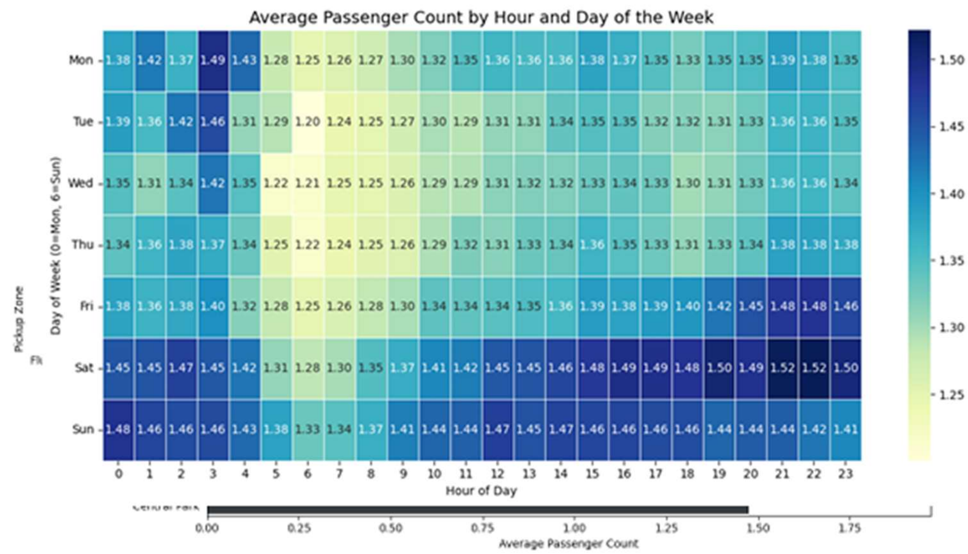


Average Tip Percentage by Pickup Hour
Tip Percentage (%)
Hour of Day (0-23)

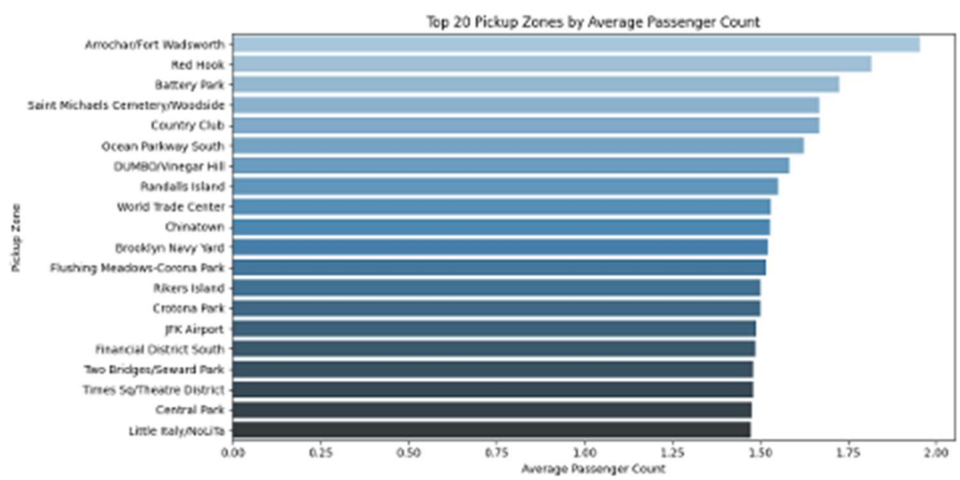


Comparison of Low vs High Tip Trips
Low Tip (<10%)
High Tip (>25%)
Average Value
Avg Trip Distance
Avg Passenger Count
Avg Fare Amount
Avg Tip Amount

3.2.14. Analyse the trends in passenger count

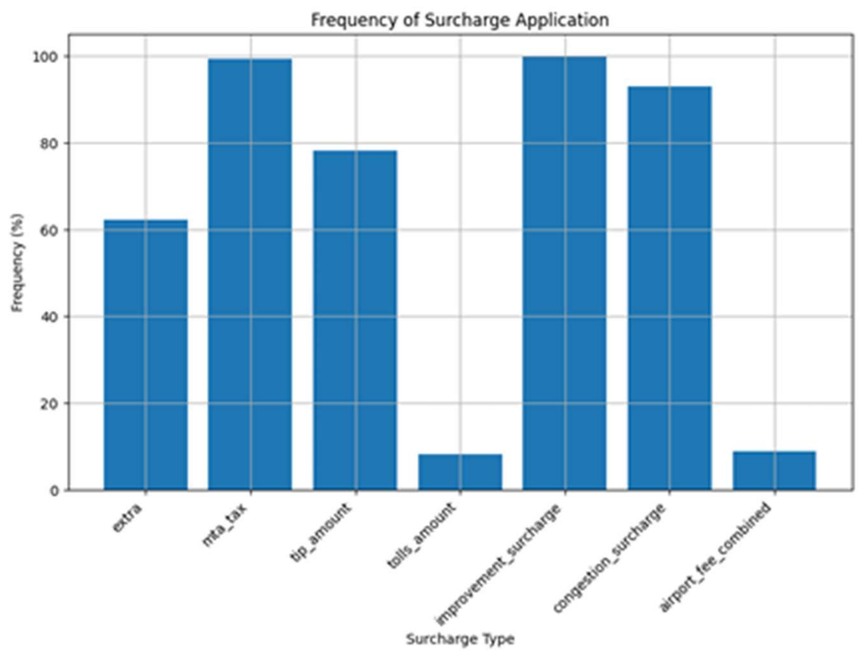


3.2.15. Analyse the variation of passenger counts across zones



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 Trends: Peak taxi demand occurs during morning and evening rush hours, on weekends, and varies by month. Nighttime demand is notably high in entertainment and nightlife districts.

-Financial Patterns: Fares generally increase with trip distance and duration. Shared rides often benefit from discounted rates. Tips tend to correlate with specific trip features.

-Geographical Insights: Airports, transportation hubs, and tourist hotspots experience the highest demand. Some areas show an imbalance between pickups and drop-offs. Nighttime activity is concentrated around popular nightlife spots.

-Vendor and Surcharges: Fare structures differ across taxi vendors, with particular surcharges frequently applied. Pricing commonly follows a tiered model based on travel distance.

Recommendations for Optimization:

Managing Demand:

-Concentrate resources in densely traveled zones and during peak times.

-Improve nighttime coverage in areas with thriving nightlife.

-Customize offerings for groups and promote carpool options.

Adjusting Supply: -Increase taxi availability in busy zones during high-demand periods. -Explore dynamic pricing models responsive to demand fluctuations. -Encourage repositioning tactics to balance taxi distribution. -Incentivize drivers to service less busy areas or times.

Enhancing Customer Experience: -Maintain high service standards with driver training and oversight. -Expand payment methods to increase convenience. -Actively promote ride-sharing to improve utilization.

Continuous Improvement: -Use ongoing data analysis and rider feedback to refine strategies. -Partner with city agencies to address regulatory and logistical 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.

Key Strategies:

```
Time-Oriented Deployment:
-Increase cab availability during peak rush hours and late-night periods, especially in nightlife zones.
-Reduce presence during midday slow periods to maximize efficiency.
-Monitor and adapt to monthly travel patterns to anticipate demand shifts.
```

Day-Specific Focus:

```
-Target business districts on weekdays where commuter traffic is highest.
-Focus on residential and entertainment districts on weekends.
-Adjust strategies dynamically for large events or festivals that draw crowds.
```

Zone-Targeted Positioning:

```
-Prioritize areas with consistent high demand such as airports, transit hubs, and tourist spots.
-Address imbalances between pickups and drop-offs by repositioning taxis accordingly.
-Boost presence in known nightlife hotspots during evening hours.
```

Data-Driven Operations:

```
-Utilize real-time data, heatmaps, and demand forecasting models to inform positioning.
-Leverage ride-hailing platform analytics for evidence-based decision-making.
-Use GPS data for dynamic vehicle repositioning aligned with current demand patterns.
```

Collaborative Efforts:

```
-Maintain clear communication channels with drivers for real-time updates.
-Partner with city officials to facilitate smooth operations during high-demand periods or events.
```

Technological Leverage:

```
-Implement GPS tracking, heatmaps, and data dashboards for strategic insights.
-Use predictive analytics to proactively position taxis ahead of demand surges.
```

Expected Outcomes: By systematically applying these strategies, taxi operators can improve service availability, reduce passenger wait times, and optimize resource utilization. Such targeted, flexible positioning enhances overall operational efficiency and customer satisfaction 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:** Modify prices in response to real-time factors such as demand spikes, taxi availability, and traffic conditions. Raise fares during busy periods and provide discounts during slower times to balance demand.
- **Tiered Pricing:** Keep competitive pricing for short-distance rides, implement graduated fare levels for longer trips, and consider location-based pricing differences to reflect zone-specific costs..
- **Shared Rides:** Encourage shared rides by offering reduced rates for groups or passengers willing to share, improving vehicle utilization and accommodating varied rider preferences.
- **Surcharge Optimization:** Evaluate how often additional fees are applied, impose surcharges strategically during high-demand periods, and ensure passengers are clearly informed about these charges.
- **Competitive Benchmarking:** Continuously monitor competitor fare policies, adapt pricing to stay competitive, and communicate unique service features to justify any higher prices.
- **Continuous Monitoring:** Regularly gather and analyze pricing data, perform tests like A/B experiments, and adjust pricing methods dynamically to maximize both revenue and customer satisfaction.

By adopting these flexible, data-informed pricing practices, taxi companies can enhance profitability while maintaining a fair and attractive cost structure for riders.