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

I began the analysis by loading the NYC Yellow Taxi dataset. I used 0.7% sampling size then merged monthly files to create a sample dataframe which is saved as NYC_Taxi_Sampled_2023.parquet with shape (266084, 22)

2. Data Cleaning

2.1. Fixing Columns

2.1.1 Fix the index .

Unnamed index columns were dropped. Date and hour columns were also dropped which were not needed in future analysis. reset_index was used after cleaning

2.1.2 Combine the two airport_fee columns

The dataset include columns with identical names "airport_fee" and "Airport_fee". Both were combined and saved as "airport_fee" in dataframe and "Airport_fee" was dropped.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

```
Proportion of missing values per column:
 VendorID                 0.000000
tpep_pickup_datetime     0.000000
tpep_dropoff_datetime    0.000000
passenger_count          0.033189
trip_distance            0.000000
RatecodeID               0.033189
store_and_fwd_flag       0.033189
PULocationID             0.000000
DOLocationID             0.000000
payment_type             0.000000
fare_amount              0.000000
extra                    0.000004
mta_tax                  0.000041
tip_amount               0.000000
tolls_amount             0.000000
improvement_surcharge    0.000041
total_amount             0.000041
congestion_surcharge     0.033211
airport_fee              0.033196
```

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.

```
Missing values in passenger_count after imputation: 0
```

2.2.3. Handle missing values in RatecodeID

There were 8831 missing RatecodeID values. I imputed them with mode.

```
Missing RatecodeID values: 8831
Missing RatecodeID values after imputation: 0
```

2.2.4. Impute NaN in congestion_surcharge

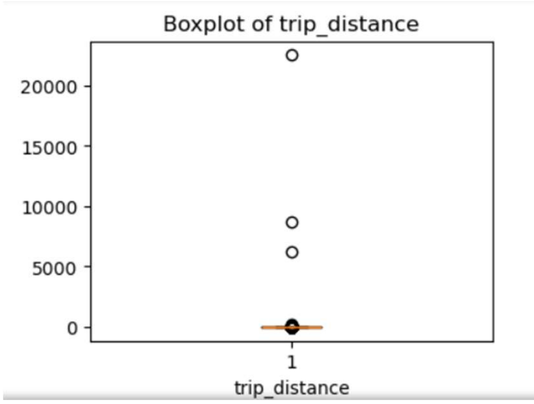
There were 8837 missing congestion_surcharge values. I imputed them with mode.

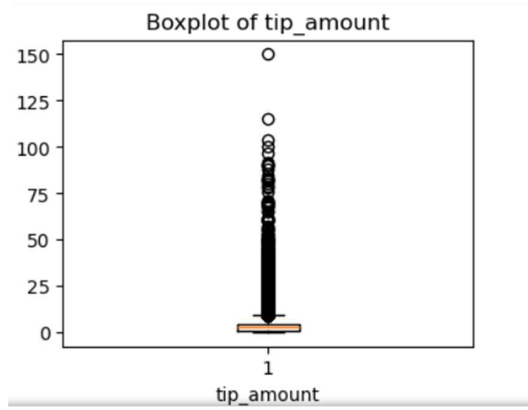
```
Missing congestion_surcharge values: 8837
Missing congestion_surcharge values after imputation: 0
```

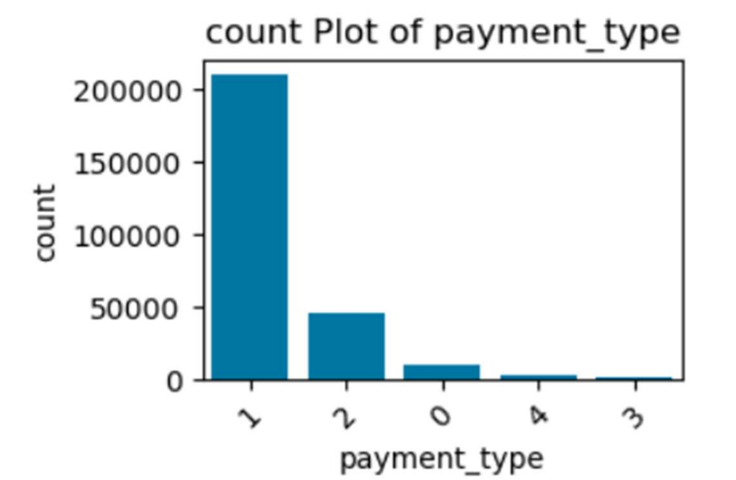
2.3. Handling Outliers and Standardising Values

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

Invalid payment codes were removed, unrealistic distances and fares were filtered out, and scaling was applied where helpful for analysis.







3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

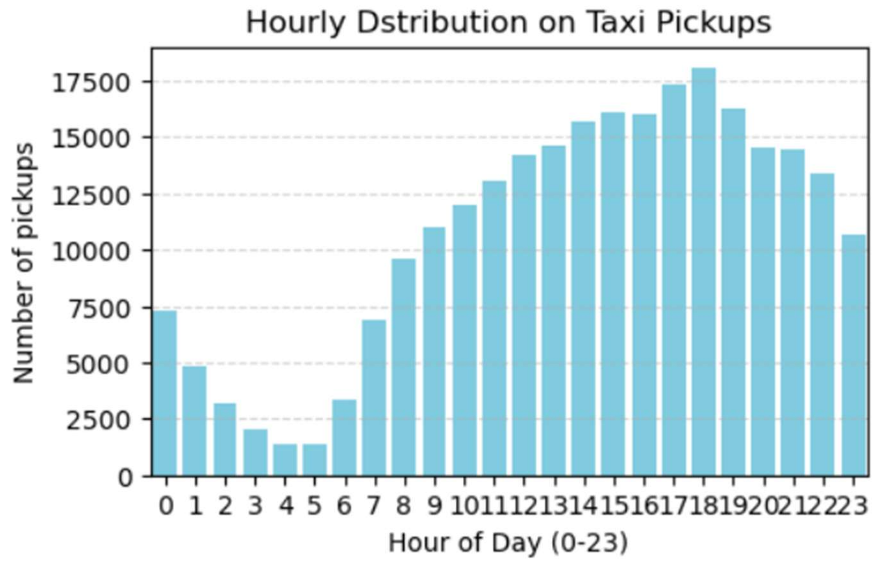
3.1.1. Classify variables into categorical and numerical

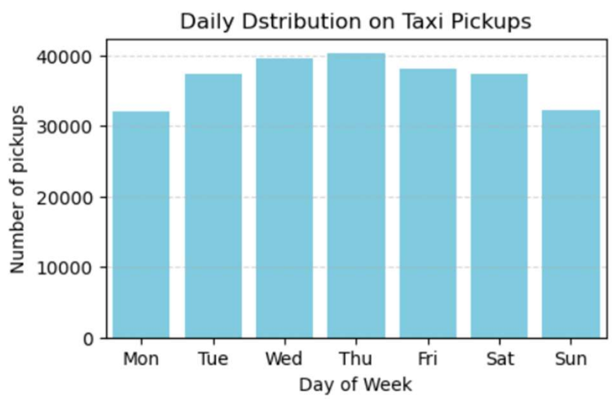
Numerical columns: ['trip_distance', 'trip_duration', 'fare_amount', 'extra', 'mta_tax', 'tip_amount', 'tolls_amount', 'improvement_surcharge', 'total_amount', 'congestion_surcharge', 'airport_fee']

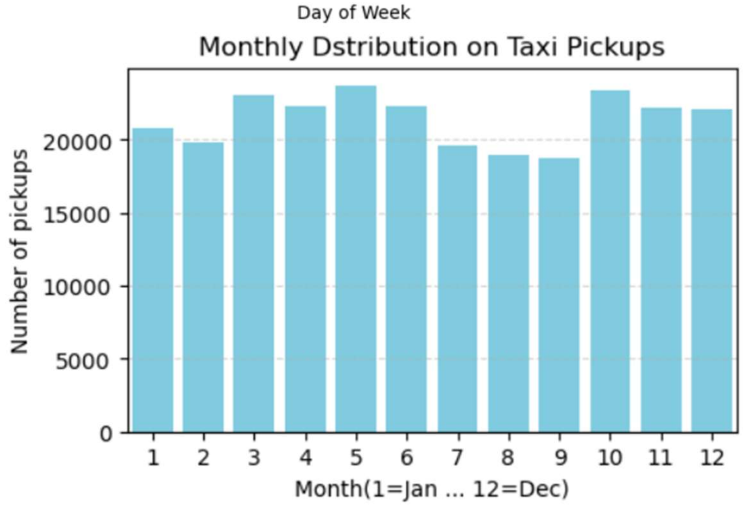
Categorical columns: ['VendorID', 'RatecodeID', 'payment_type', 'passenger_count', 'PULocationID', 'DOLocationID', 'pickup_hour', 'store_and_fwd_flag']

Datetime columns: ['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

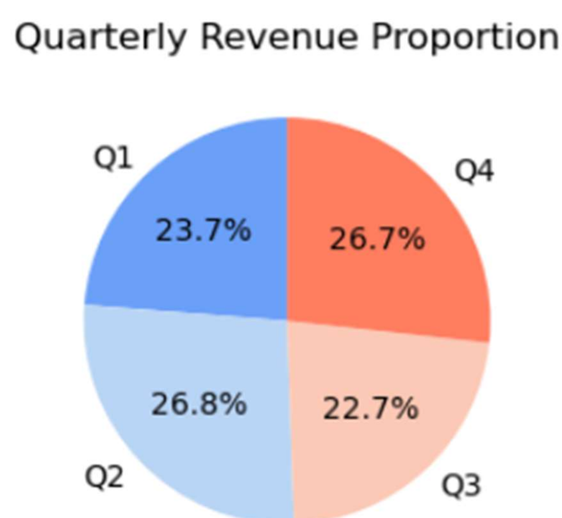
Only tip_amount and trip_distance has zero values.
zero value in tip_amount is valid. but among zero values in trip_distance some are valid and some are invalid entries. I kept all rows which have trip_distance=0 and fare_amount >0.

```
fare_amount: Zero values = 0, negative values = 0
tip_amount: Zero values = 57492, negative values = 0
total_amount: Zero values = 0, negative values = 0
trip_distance: Zero values = 3225, negative values = 0
```

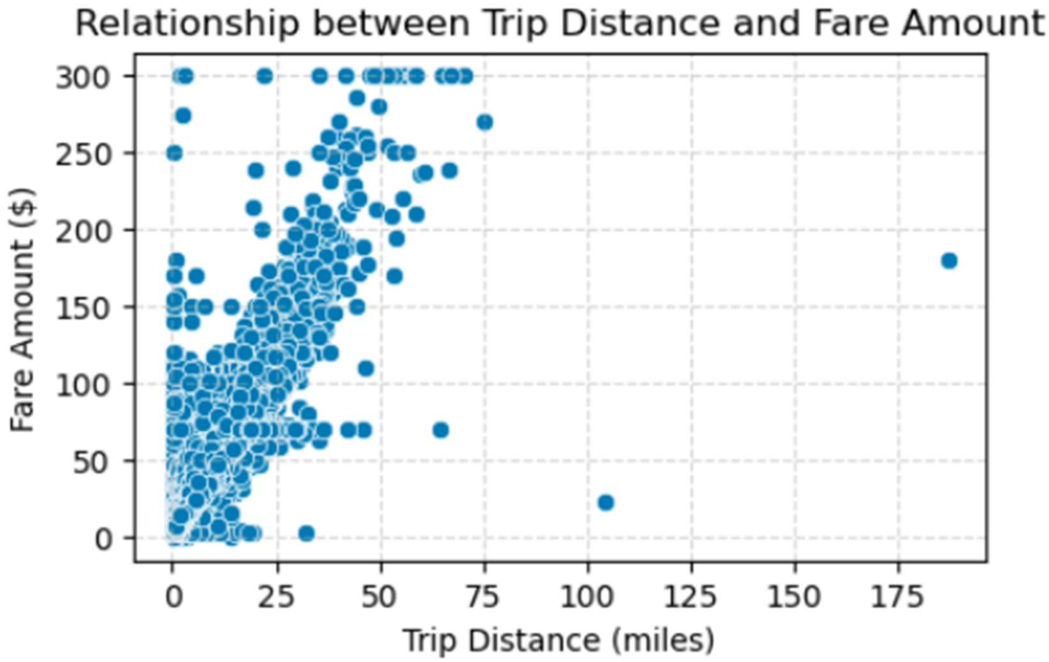
3.1.4. Analyse the monthly revenue trends

	Month	Total_Revenue
0	1	571155.77
1	2	541417.89
2	3	649893.44
3	4	636454.00
4	5	696652.23
5	6	656004.16
6	7	564589.19
7	8	551705.18
8	9	569342.54
9	10	700955.52
10	11	640544.41
11	12	642441.63

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

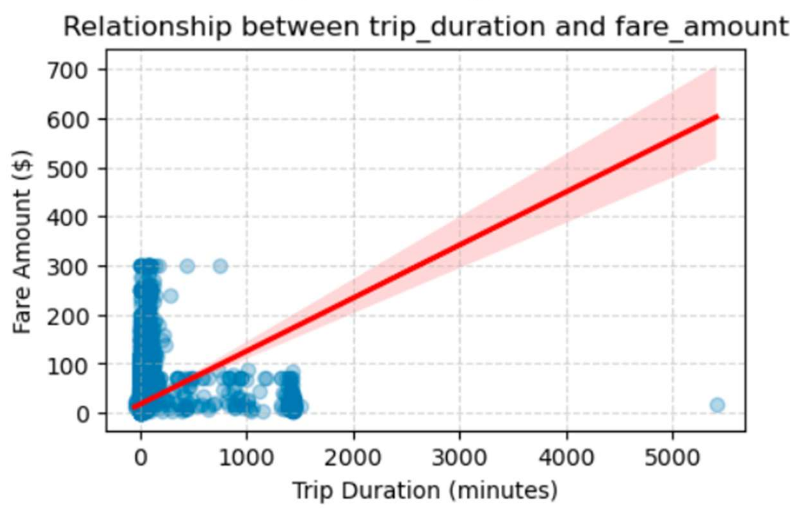


3.1.6. Analyse and visualise the relationship between distance and fare amount

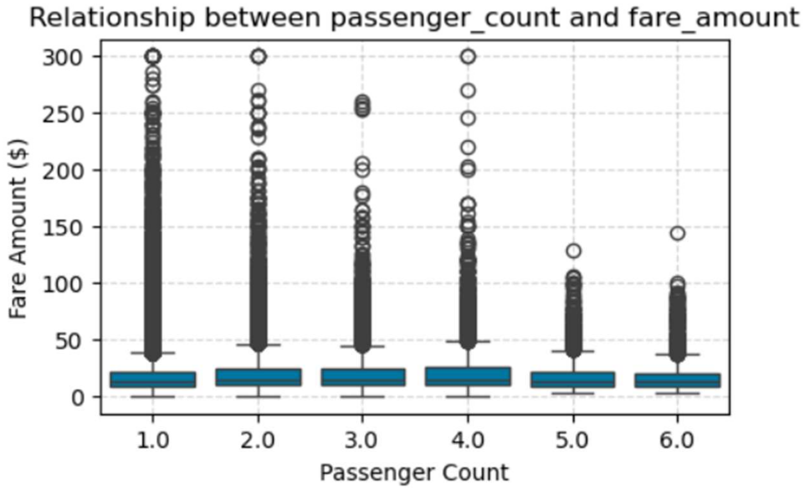


3.1.7. Analyse the relationship between fare/tips and trips/passengers

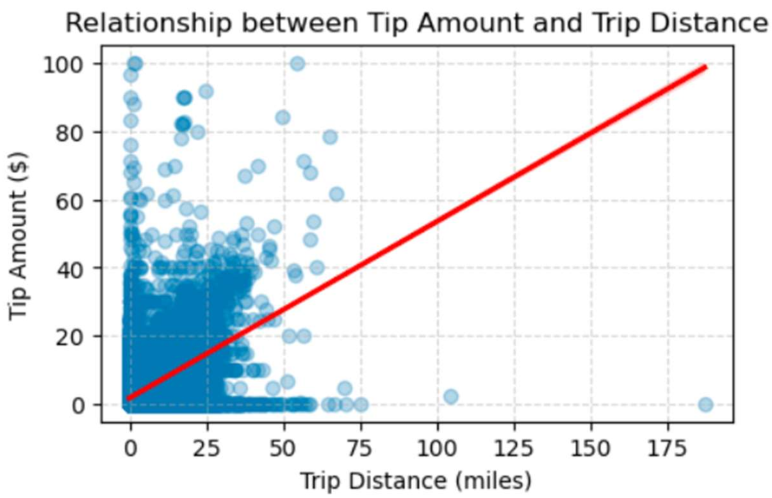
```
Correlation between fare_amount and trip_duration:0.256
```



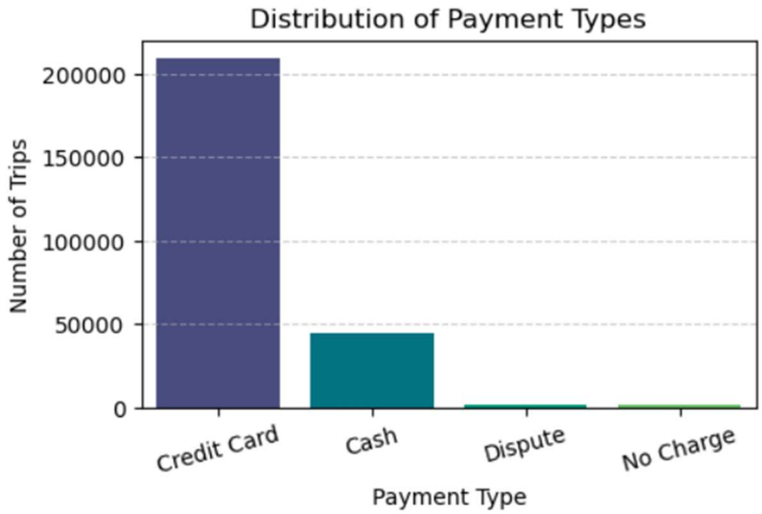
```
Correlation between fare_amount and passenger_count:0.044
```




```
Correlation between tip_amount and trip_distance:0.578
```

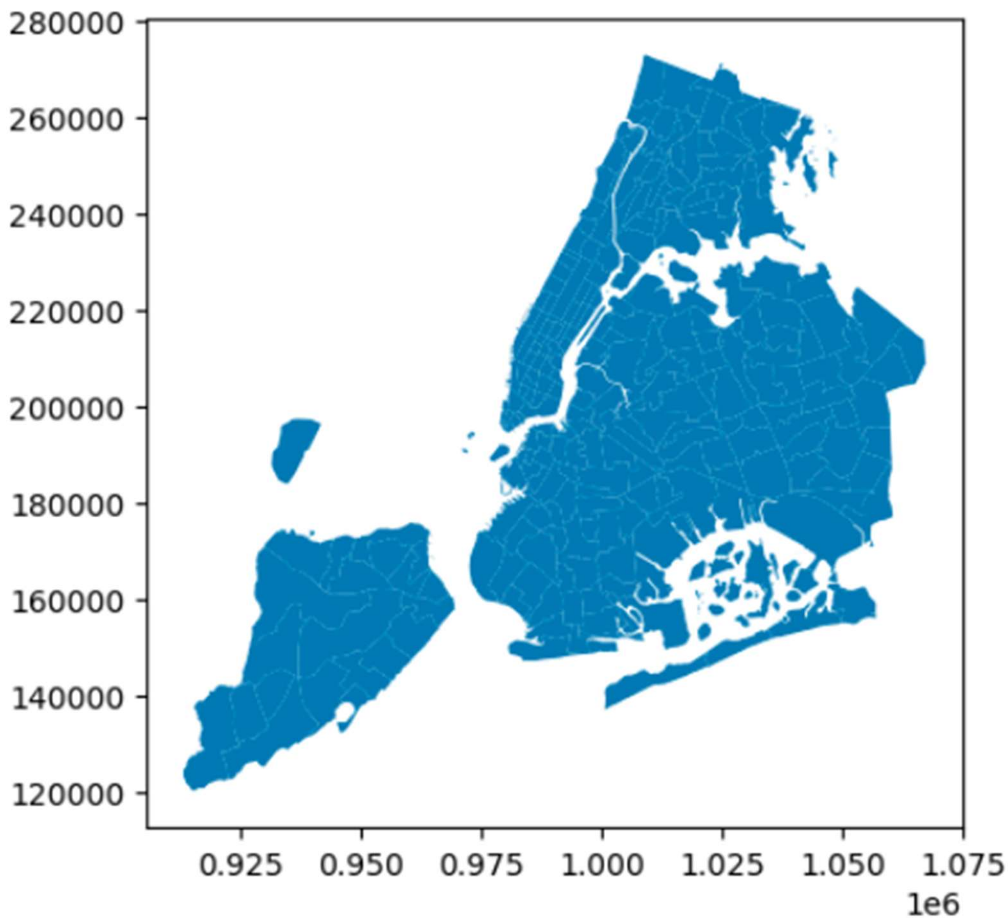


3.1.8. Analyse the distribution of different payment types



3.1.9. Load the taxi zones shapefile and display it

I installed geopandas to load shape file.



3.1.10. Merge the zone data with trips data

Merged the zone data and trip data using locationID and PULocationID

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

Sample of data given below.

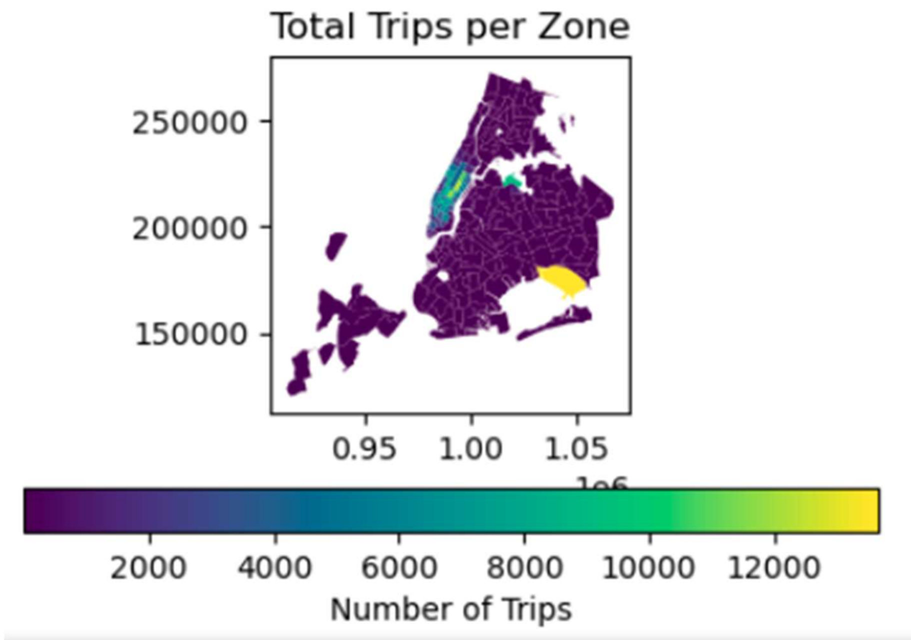
	LocationID	trip_count
0	1.0	38
1	3.0	9
2	4.0	245

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

Merged trip counts back to the zones GeoDataFrame

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry	trip_count
0	1	0.116357	0.000782	Newark Airport	1	EWR	POLYGON ((933100.918 192536.086, 933091.011 19...	38.0
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 103343...	NaN
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 2...	9.0
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...	245.0
4	5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144...	1.0

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

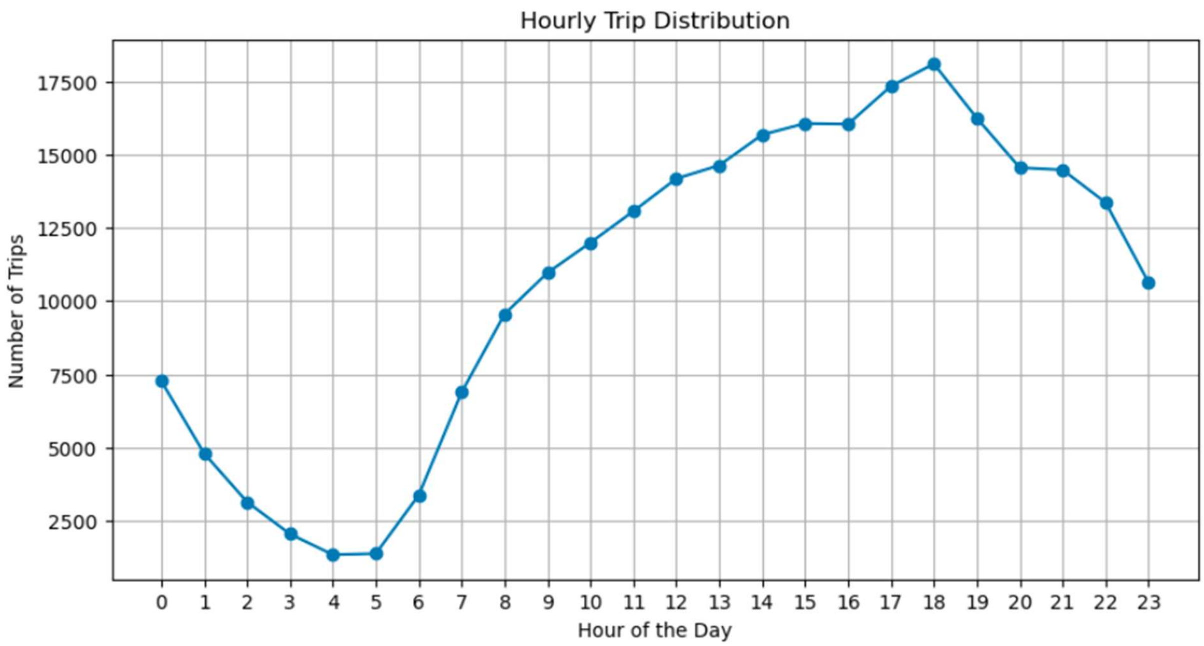
3.2. Detailed EDA: Insights and Strategies

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

	PULocationID	DOLocationID	pickup_hour	avg_speed_mph
15904	113	113	13	0.025129
44025	226	145	18	0.026569
57262	260	129	17	0.040746
60053	264	237	15	0.043036
42540	209	232	13	0.043579
16448	113	235	22	0.048105
42140	193	193	13	0.052326
5047	50	43	8	0.059525
37590	164	100	21	0.067827
21922	134	265	15	0.073831

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

```
Busiest hour: 18
Number of trips during busiest hour: 18093
```

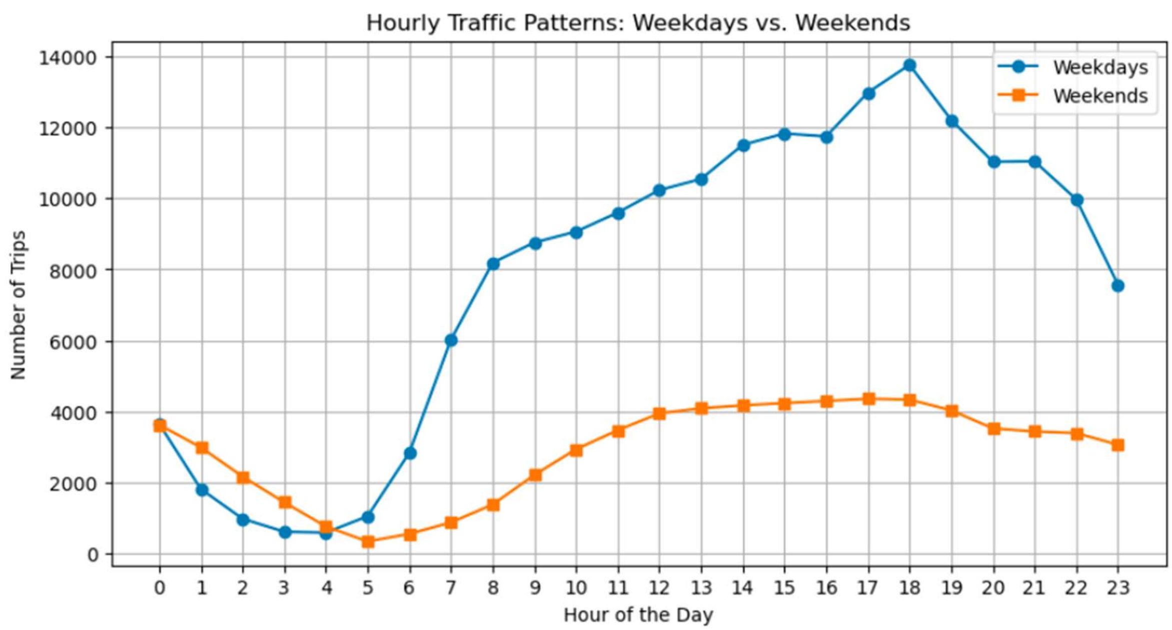


3.2.3. Scale up the number of trips from above to find the actual number of trips

Scaled up the dataframe. Below is the list of five busiest hours with the trip counts.

18	18093
17	17333
19	16243
15	16061
16	16037

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
117	132	13639	JFK Airport
213	237	12135	Upper East Side South
145	161	12057	Midtown Center
212	236	10812	Upper East Side North
146	162	9309	Midtown East
123	138	9001	LaGuardia Airport
166	186	8762	Penn Station/Madison Sq West
206	230	8603	Times Sq/Theatre District
127	142	8394	Lincoln Square East
154	170	7587	Murray Hill

```
Top 10 dropoff zones
```

	LocationID	Dropoff_Trips	zone
226	236	11487	Upper East Side North
227	237	10770	Upper East Side South
154	161	10042	Midtown Center
220	230	7817	Times Sq/Theatre District
163	170	7654	Murray Hill
155	162	7326	Midtown East
135	142	7288	Lincoln Square East
229	239	7185	Upper West Side South
134	141	6665	Lenox Hill West
67	68	6579	East Chelsea

```
Bottom 10 Pickup/Dropoff Ratios:
```

	zone	Pickup_Trips	Dropoff_Trips	pickup_dropoff_ratio
119	Highbridge Park	0.0	11	0.0
181	Pelham Bay Park	0.0	2	0.0
155	Mariners Harbor	0.0	1	0.0
202	Saint George/New Brighton	0.0	4	0.0
200	Rossville/Woodrow	0.0	3	0.0
25	Breezy Point/Fort Tilden/Riis Beach	0.0	4	0.0
111	Green-Wood Cemetery	0.0	3	0.0
217	Stapleton	0.0	3	0.0
64	Crotona Park	0.0	1	0.0
63	Country Club	0.0	9	0.0

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

```
Top 10 Pickup/Dropoff Ratios:
```

	zone	Pickup_Trips	Dropoff_Trips	pickup_dropoff_ratio
75	East Elmhurst	1187.0	145	8.186207
131	JFK Airport	13639.0	2890	4.719377
137	LaGuardia Airport	9001.0	3101	2.902612
183	Penn Station/Madison Sq West	8762.0	5797	1.511471
41	Central Park	4399.0	3158	1.392970
245	West Village	5734.0	4202	1.364588
114	Greenwich Village South	3352.0	2489	1.346726
161	Midtown East	9309.0	7326	1.270680
160	Midtown Center	12057.0	10042	1.200657
104	Garment District	4202.0	3532	1.189694

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

```
Top 10 Pickup zones during night hours (11pm to 5am):
pickup_zone
East Village                    2161
JFK Airport                     1915
West Village                    1766
Lower East Side                 1372
Clinton East                    1366
Greenwich Village South         1187
Times Sq/Theatre District       1154
LaGuardia Airport                896
Penn Station/Madison Sq West     883
Midtown South                    834
```

```
Top 10 Dropoff zones during night hours (11pm to 5am):
dropoff_zone
East Village       1165
Clinton East        960
Murray Hill         858
Gramercy            817
East Chelsea        810
Lenox Hill West     734
Yorkville West      715
West Village        675
Flatiron            632
Lower East Side     624
```

3.2.8. Find the revenue share for nighttime and daytime hours

```
Nighttime Revenue Share: 12.25%
Daytime Revenue Share: 87.75%
```

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

Below is the list of average fare per mile for different passenger count.

```
passenger_count
1.0    16.536024
2.0     9.267888
3.0     6.186545
4.0     6.816272
5.0     2.609565
6.0     2.104039
```

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

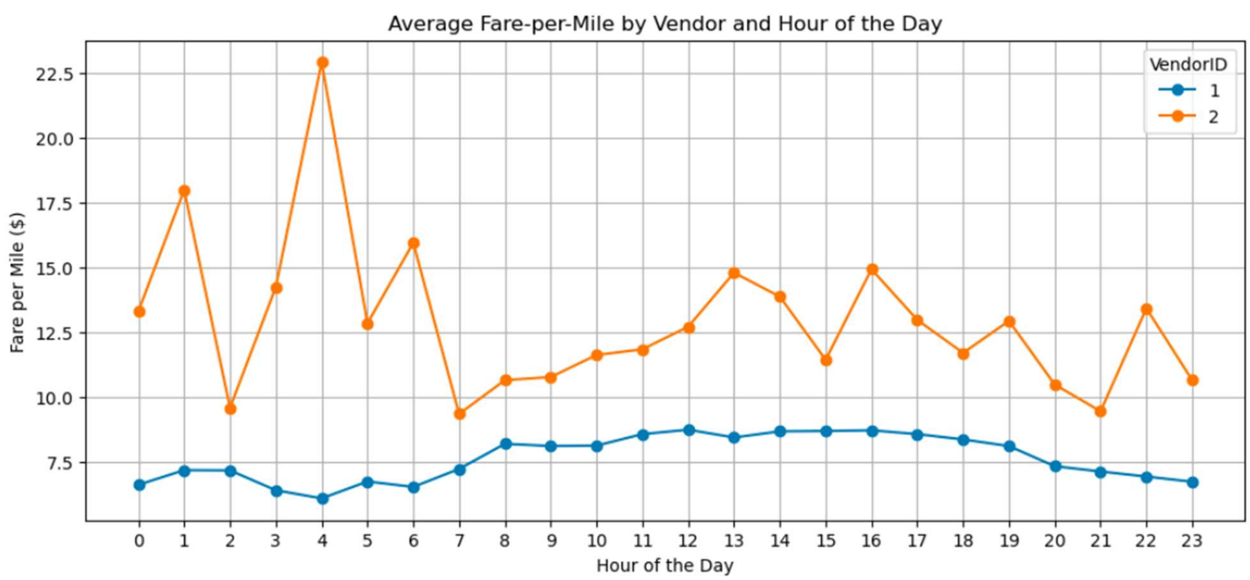
<u>average fare per mile by hours of the day</u>

```
hour_of_day
0     17.03
1     21.40
2     14.49
3     17.92
4     28.43
5     16.53
6     18.86
7     13.22
8     15.00
9     15.16
10    15.70
11    16.34
12    17.18
13    18.40
14    17.87
15    15.85
16    20.97
17    18.80
18    17.82
19    18.86
20    15.09
21    14.22
22    17.28
23    14.64
```

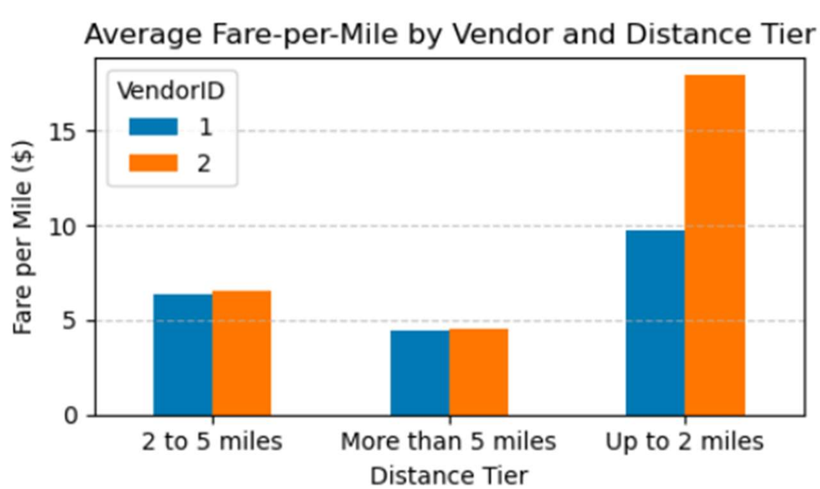
<u>fare per mile by days of the week</u>

```
day_of_week
Monday       15.66
Tuesday      17.19
Wednesday    17.91
Thursday     19.23
Friday       15.59
Saturday     16.52
Sunday       16.85
```

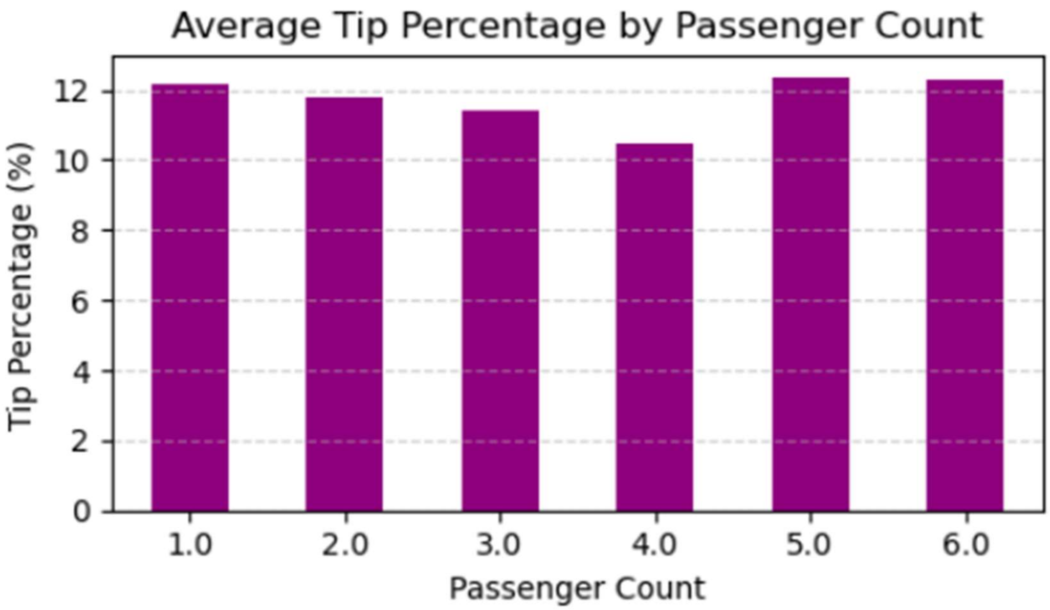
3.2.11. Analyse the average fare per mile for the different vendors

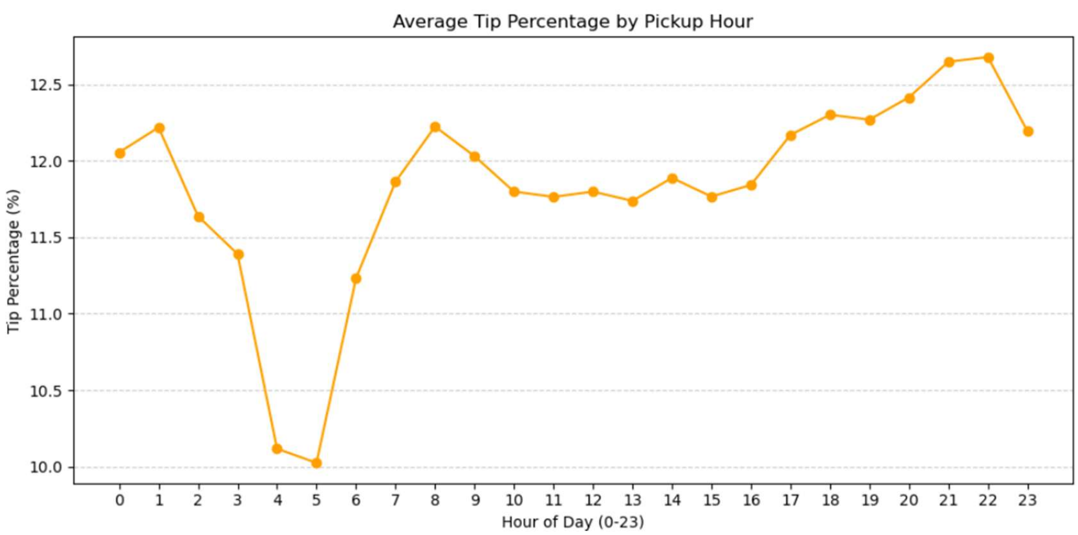


3.2.12. Compare the fare rates of different vendors in a distance-tiered fashion

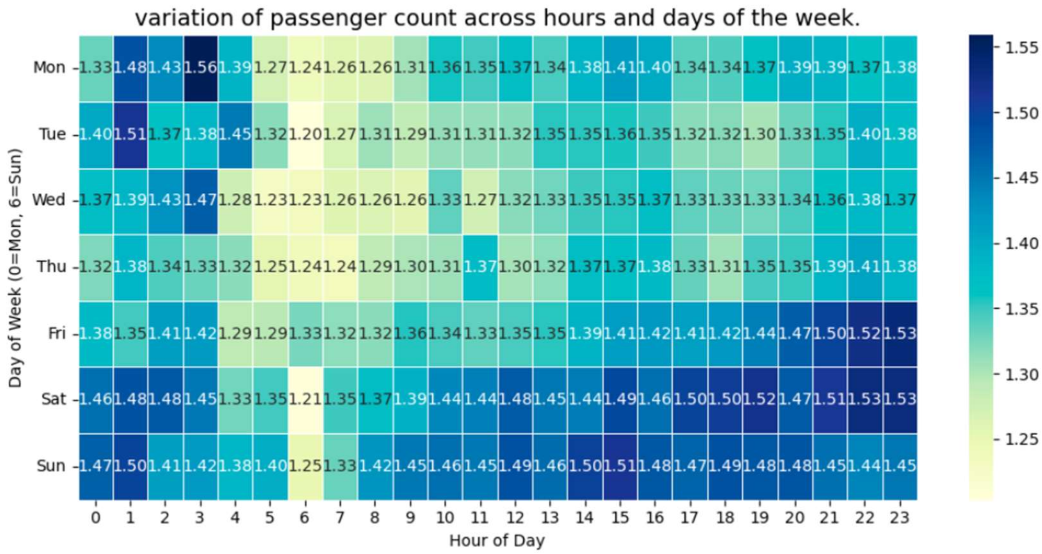


3.2.13. Analyse the tip percentages

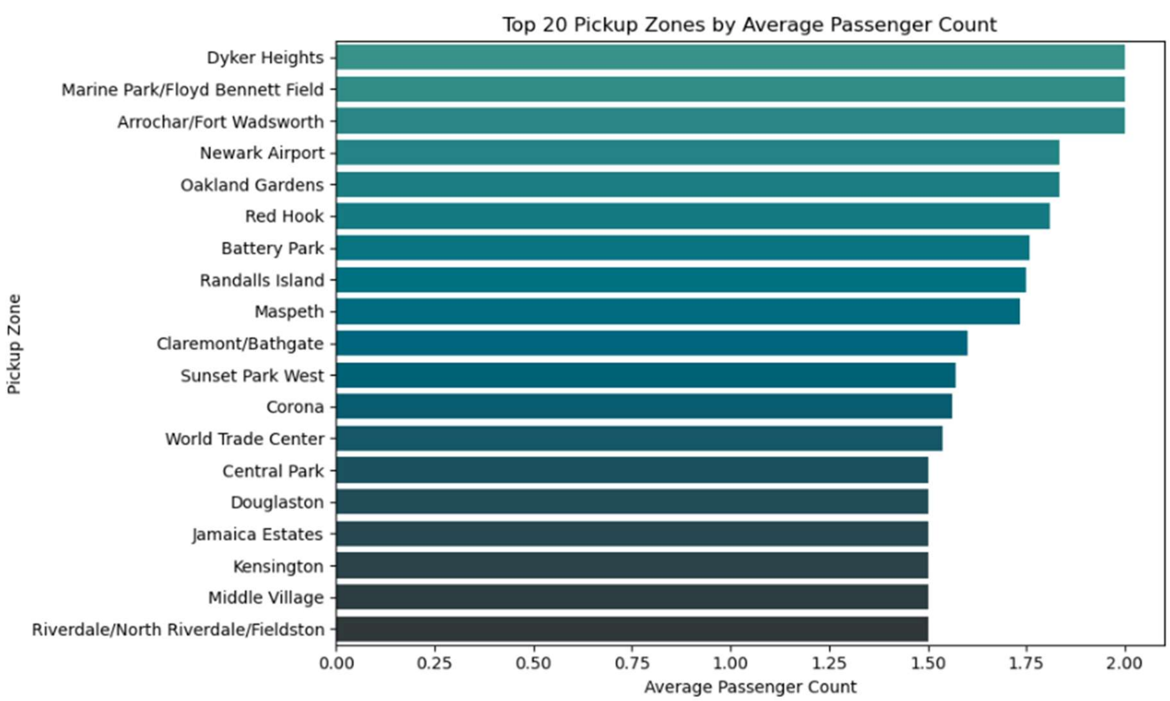




3.2.14. Analyse the trends in passenger count

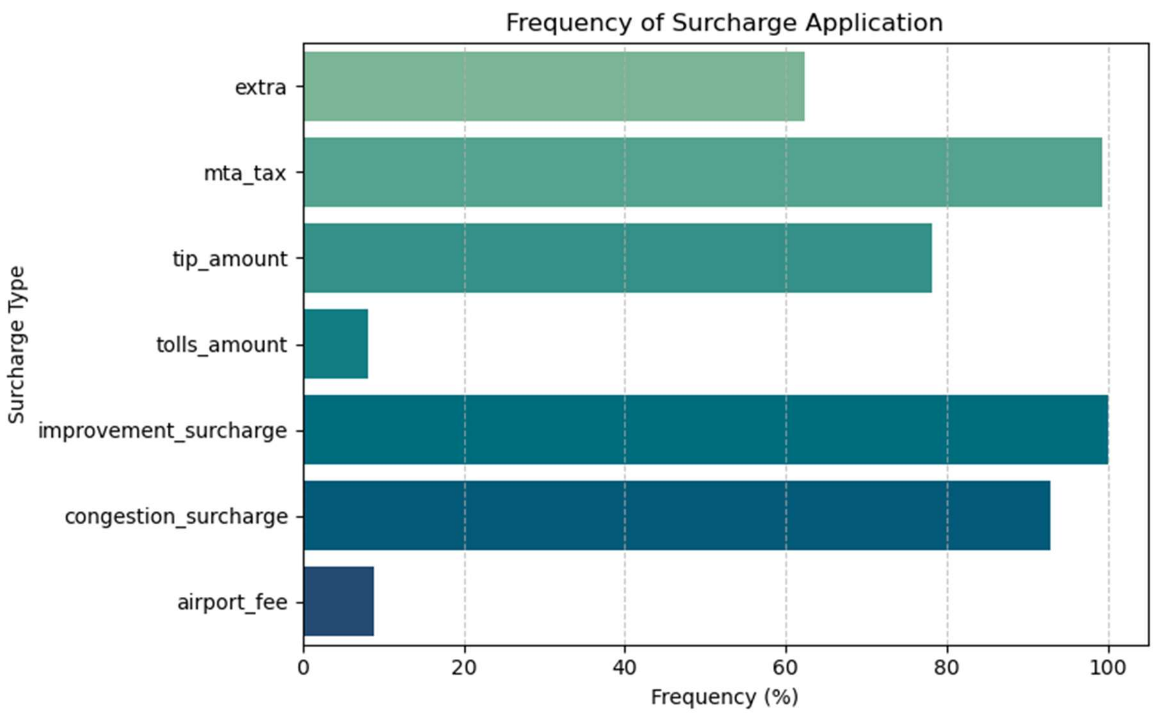


3.2.15. Analyse the variation of passenger counts across zones



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

```
Frequency of Surcharge Application (%):
extra                     62.295857
mta_tax                   99.388013
tip_amount                78.147263
tolls_amount               8.085788
improvement_surcharge     99.998819
congestion_surcharge      92.912530
airport_fee                8.808437
```



4. Conclusions

4.1. Final Insights and Recommendations

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

Recommendations for Strategy and Optimization

1. Demand-aware vehicle allocation

- The hourly plots and heatmap show clear morning and evening peaks e.g., 8:00 appears in the analysis. Allocate more vehicles to identified high-demand zones during those peak windows.

- Maintain a smaller, steady presence in entertainment/nighttime areas shown in the heatmaps to preserve availability overnight.

2. Zone-level dispatching

- The zone/heatmap outputs list specific zones and airport corridors (examples appear in above analysis: Midtown, JFK, LaGuardia, Chelsea, Upper East/Upper West, Bronx, Queens, Staten Island). Treat the high-density zones as priority dispatch areas and increase vehicle concentration there.

- For under-served zones visible in the zone analysis, consider targeted measures (operational reallocation or routing adjustments) to improve coverage.

3. Traffic-aware routing and path optimization

- optimize travel paths using shortest-path routing while factoring in congestion. Use traffic-aware routing to avoid slow corridors identified in route-speed plots and reduce trip times.

4. Reduce idle time and improve driver utilization

- Prioritize assigning the nearest available driver to a forecasted pickup to reduce deadhead travel and idle minutes—this is supported by the “driver utilization” and “idle” mentions in the analysis above.

5. Dynamic / predictive dispatching

- Implement a demand-prediction component (using the hourly and monthly patterns already plotted) to plan vehicle distribution before peak windows and major seasonal spikes.

6. Fare and congestion considerations

- Use the fare and surcharge information that appears in the plots to inform time-based operational decisions (for example, prioritizing airport and congestion-affected trips when surcharges apply).

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.

1. Morning (around 8:00 AM):

- The hourly pickup plots show a noticeable increase in demand starting around 8:00 AM, consistent with commuters traveling from residential areas toward central business zones.
- Recommended action: Position a higher share of cabs in Midtown, the Financial District, and nearby commercial areas during this window.

2. Evening Rush Hours (after 5 PM):

- The same hourly analysis shows another strong increase during evening hours as passengers return home or travel to leisure zones.
- Recommended action: Shift part of the fleet toward Chelsea, Upper East Side, and Upper West Side, where the pickup heatmaps in your plots show recurring evening activity.

3. Night and Midnight Hours:

- According to hourly and night-hour plots, trip volumes remain steady through late-night and early-morning hours, especially around JFK and LaGuardia airport corridors and entertainment regions.
- Recommended action: Keep a smaller but steady number of cabs positioned near airport terminals and nighttime activity zones to serve passengers arriving from late flights or nightlife areas.

Weekday vs Weekend:

- The weekday/weekend comparison plots indicate that weekend nights have higher trip counts in zones with restaurants and nightlife (e.g., Chelsea, Financial District).
- Recommended action: Increase late-evening coverage in these areas on Fridays and Saturdays while slightly reducing weekday midnight deployments to optimize utilization.

Outer Borough Balancing:

- Heatmaps show fewer pickups in Bronx, Queens, and Staten Island compared to central Manhattan.
- Recommended action: Use short-term predictive logic to temporarily reallocate cabs from low-demand Midtown blocks to these outer boroughs when utilization is low.

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

Time-of-Day Fare Differentiation:

- Since both the 8:00 AM morning and evening hours exhibit clear demand peaks, a small peak-hour multiplier can be applied during those windows to balance supply and demand without overpricing off-peak riders.

Congestion-Linked Adjustments:

- Recommendation: Integrate surcharge dynamically, applying it during hours and routes where the route-speed and fare analyses indicate frequent slow travel (for example, evening trips between Midtown and the airports). This ensures fair compensation for time lost in traffic and aligns price with trip duration.

Distance-Tiered Fare Refinement:

- The fare vs. distance plots show the fare increasing roughly linearly with distance.

- Recommendation: Review short-distance (under 2 miles) and long-distance (>10 miles) trip fare efficiency. Slightly adjust the base fare or minimum charge for very short trips to cover idle time, while maintaining per-mile rates for long airport rides.

Vendor-Level Pricing Balance:

- Differences in fare and tip distributions can guide competitive pricing—keeping base fares aligned across vendors while allowing small variations in surcharge handling or promotional discounts to attract riders.