Report: Optimizing NYC Taxi Operations

Submission by – Neha Shukla

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

In accordance with the provided guidelines, I initially extracted a sample of 500,000 records from each monthly Parquet file. Subsequently, I refined the sample size to ensure that the final combined DataFrame comprised approximately **1.89 million rows**.

2. Data Cleaning

2.1. Fixing Columns

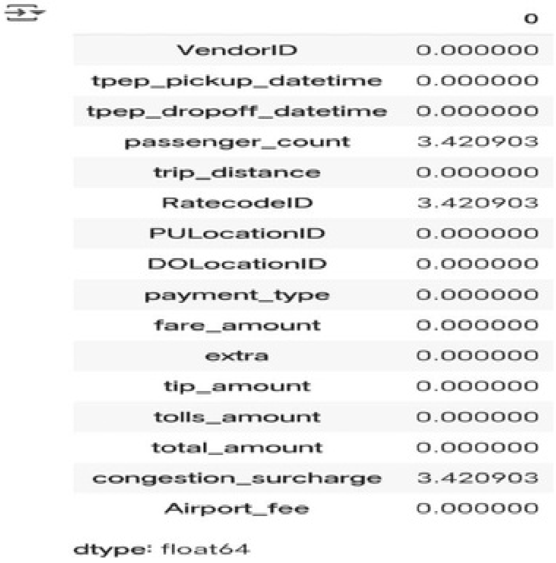
To ensure consistency, column names were standardized by removing extra spaces and applying uniform formatting

2.1.1. Fix the index

Combine the two airport_fee columns.The dataset included two similar columns - airport_fee and Airport_fee — likely resulting from inconsistent column naming across monthly files. To address this, I introduced a new column, Airport_fee, which captures the maximum value between the two columns for each row to prevent data loss. After creating this consolidated column, the original airport_fee and Airport_fee columns were removed to eliminate 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
PULocationID	0.000000
DOLocationID	0.000000
payment_type	0.000000
fare_amount	0.000000
extra	0.000000
tip_amount	0.000000
tolls_amount	0.000000
total_amount	0.000000
congestion_surcharge	3.420903
Airport_fee	0.000000

dtype: float64

2.2.2. Handling missing values in passenger_count

To handle missing values in the passenger_count column, I used the mode (i.e., the most frequently occurring value) to impute the null entries. This method is appropriate because passenger_count is a discrete variable, and the mode — typically **1** — represents the most common number of passengers in a yellow taxi trip. This strategy helps preserve the data distribution without introducing skew.

2.2.3. Handle missing values in RatecodeID

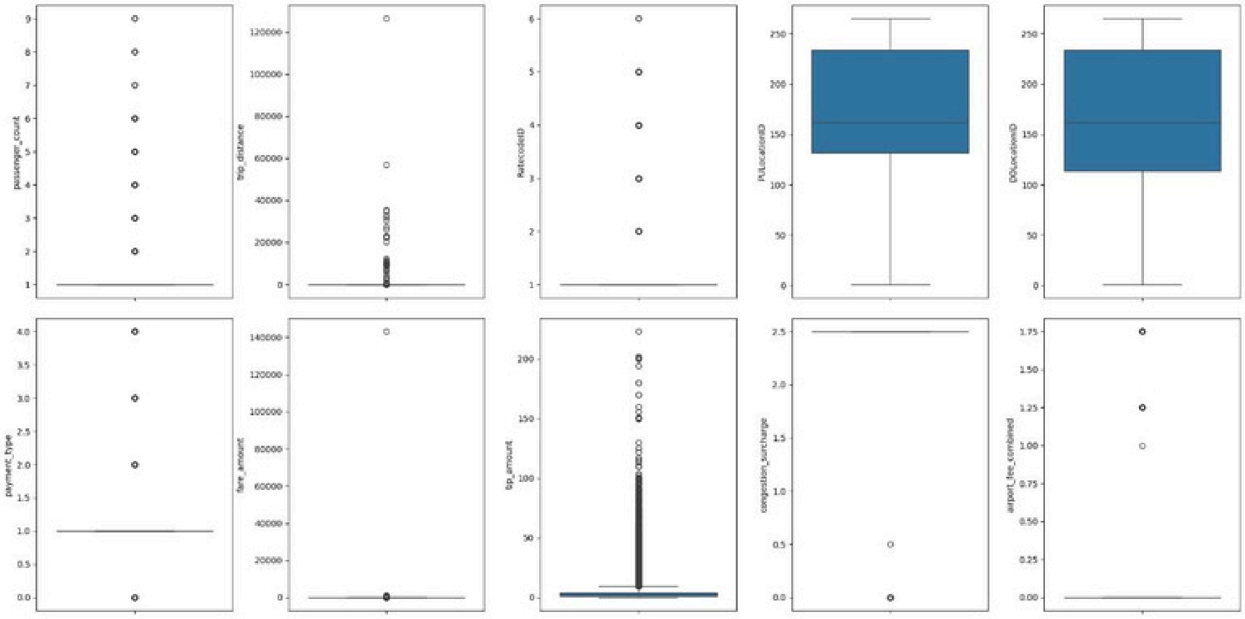
Missing values in the RatecodeID column were also imputed using the mode. Since RatecodeID is a categorical variable, this approach ensures the most frequent category is retained, effectively preserving the dataset's dominant pattern. It also prevents distortion from rare or extreme values, maintaining overall data integrity

2.2.4. Impute NaN in congestion_surcharge

Missing values in the congestion_surcharge column were imputed using the **median** of the non-null values.Using the median helps prevent skewing the data due to extreme outliers, thereby preserving the column’s overall distribution and integrity.

2.3. Handling Outliers and Standardizing Values

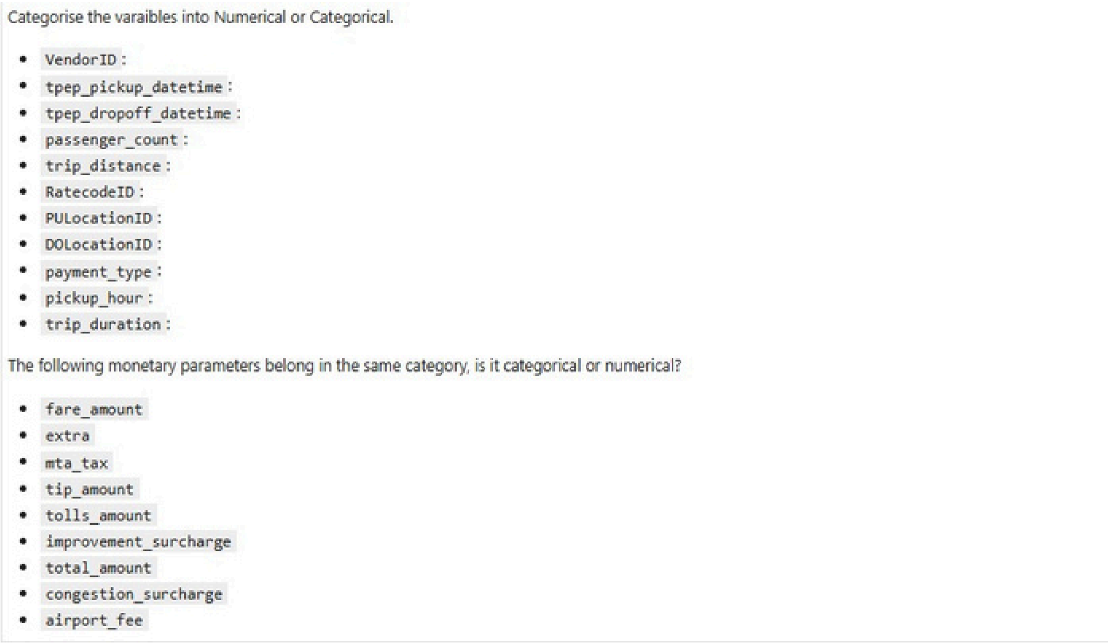
2.3.1. Check outliers in payment type, trip distance and tip amount columns



3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

3.1.1. Classify variables into categorical and numerical



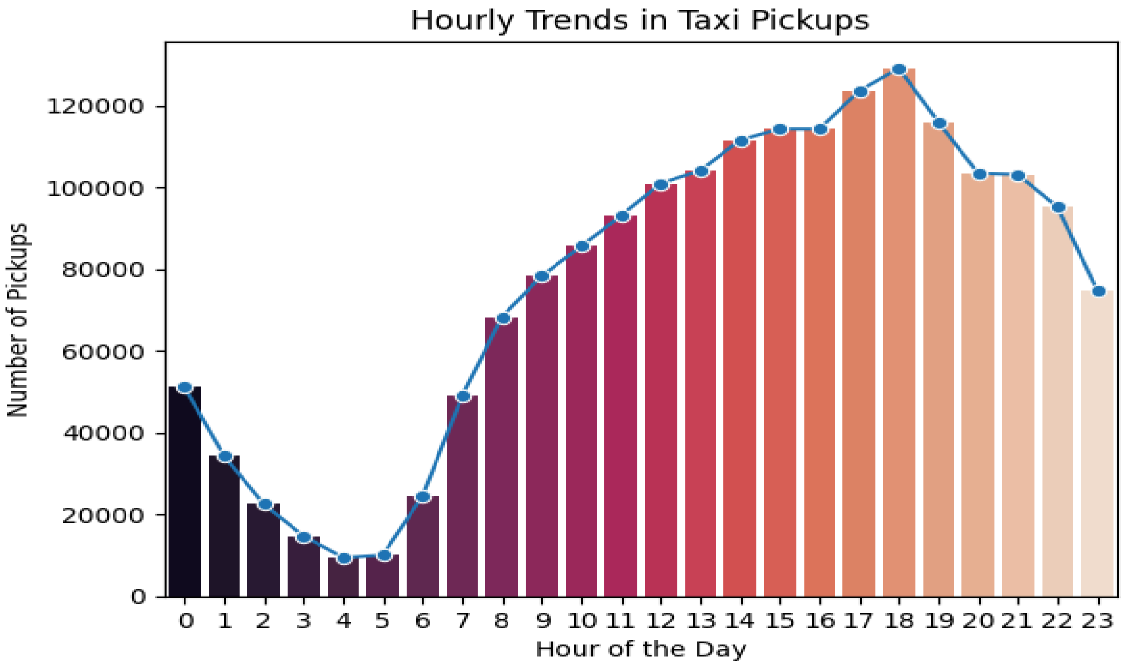
Categorise the varaibles into Numerical or Categorical.

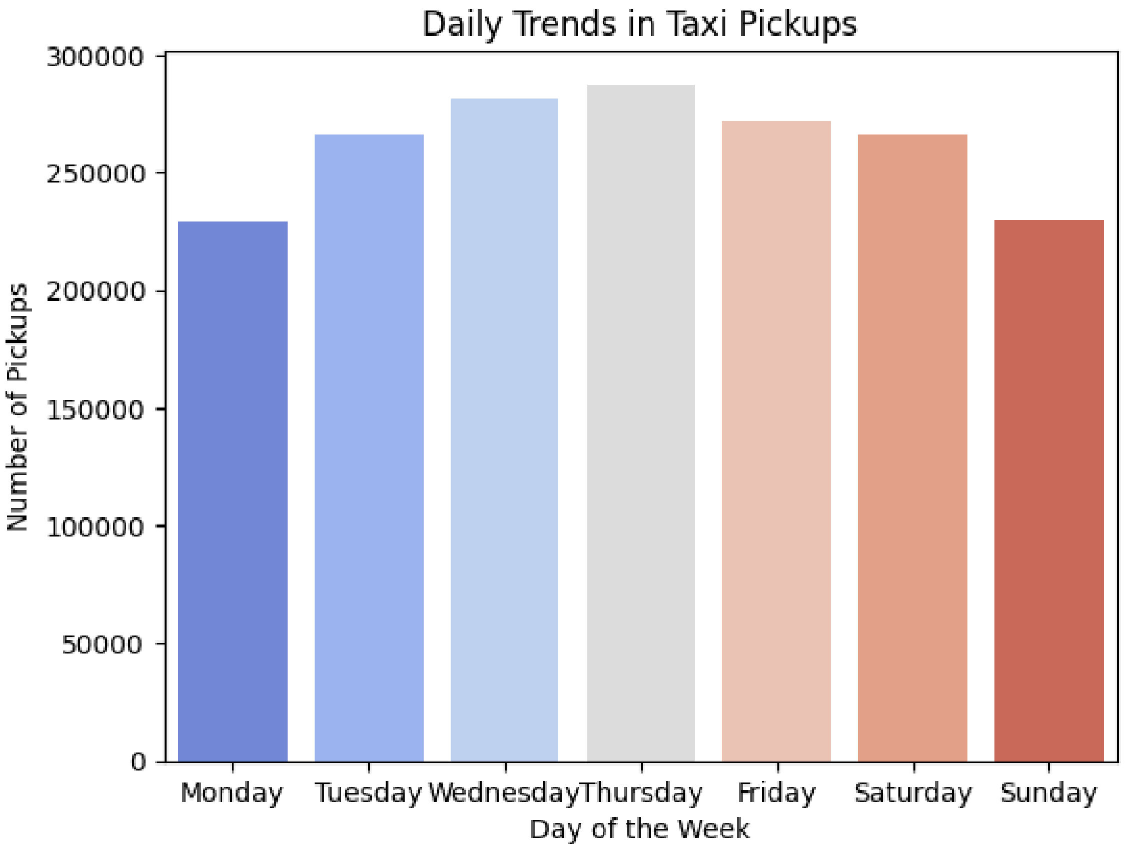
- `VendorID` :
- `tpep_pickup_datetime` :
- `tpep_dropoff_datetime` :
- `passenger_count` :
- `trip_distance` :
- `RatecodeID` :
- `PULocationID` :
- `DOLocationID` :
- `payment_type` :
- `pickup_hour` :
- `trip_duration` :

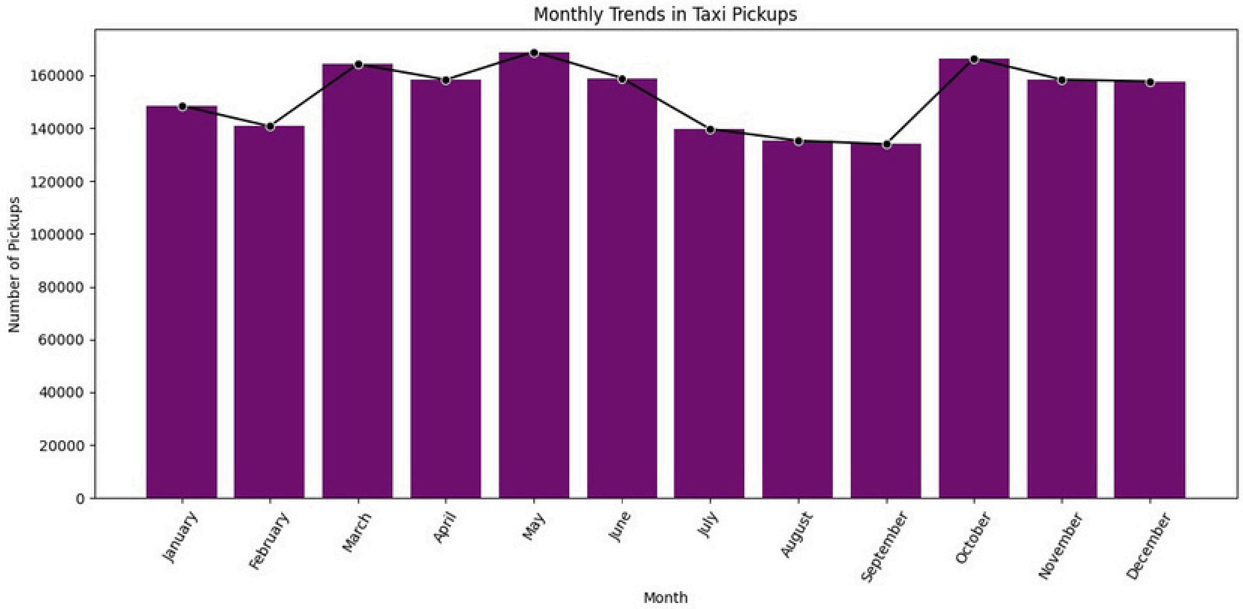
The following monetary parameters belong in the same category, is it categorical or numerical?

- `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





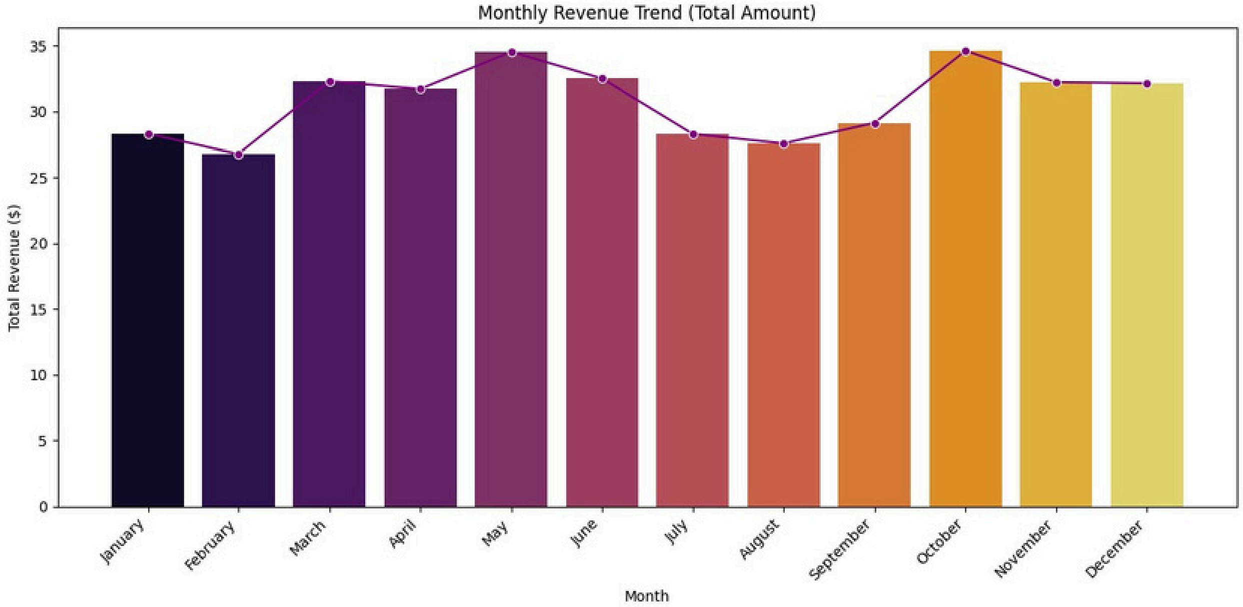


3.1.3. Filter out the zero/negative values in fares, distance and tips

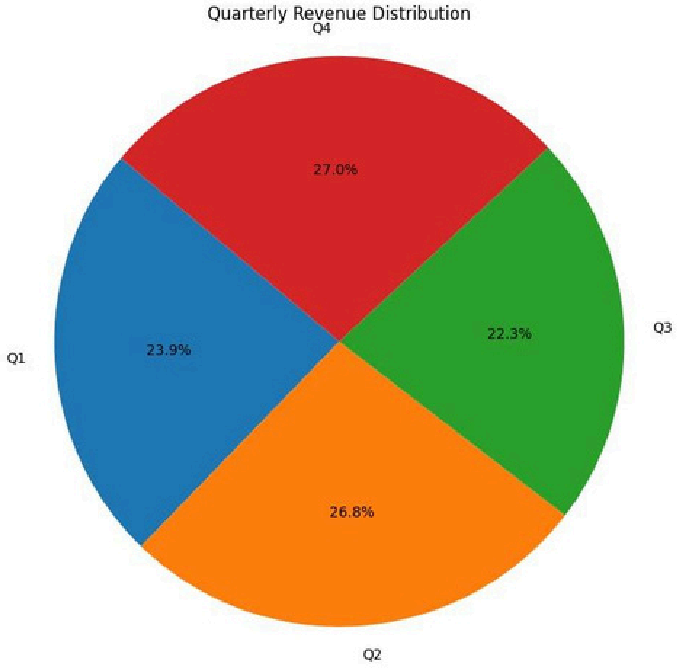
To maintain high data quality, records were filtered out based on the following criteria:

fare_amount or total_amount equal to zero: Such entries were likely due to invalid or canceled trips and were therefore removed.
trip_distance equal to zero while pickup and dropoff locations were different: These entries were flagged as inconsistent and excluded from the dataset.
However, zero tip_amount values were retained, since tipping is optional. Many valid trips did not include a tip but still showed valid total_amount values, confirming their legitimacy.

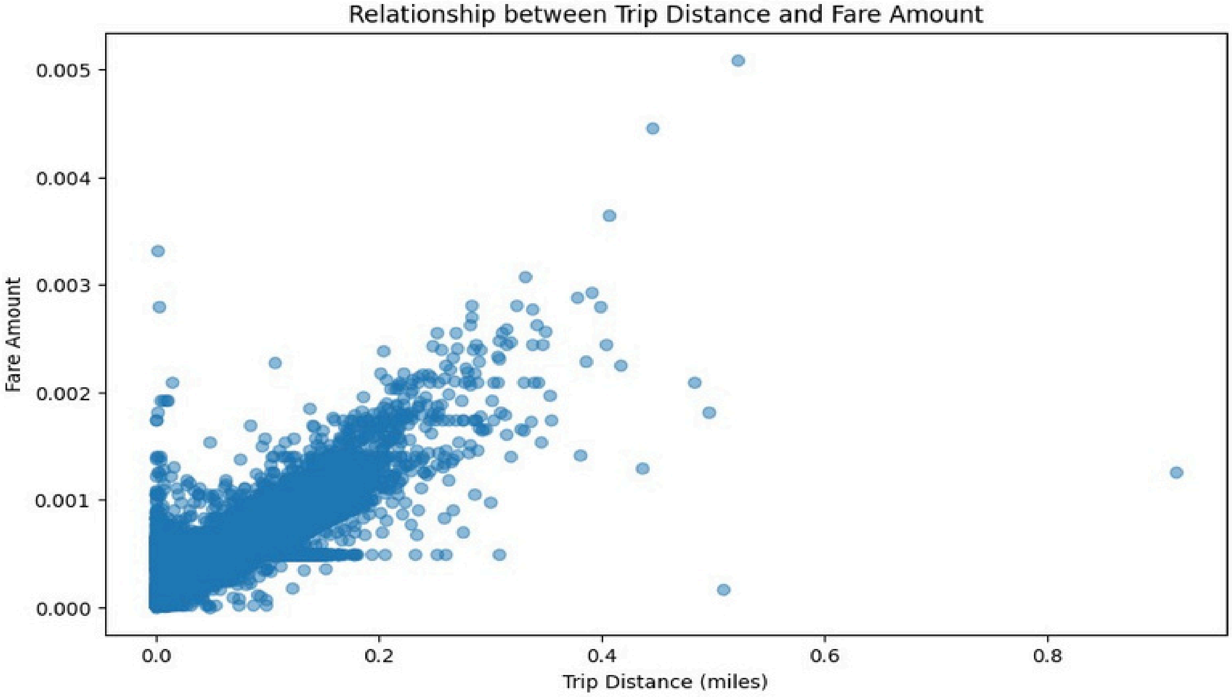
3.1.4. Analyse the monthly revenue trends



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

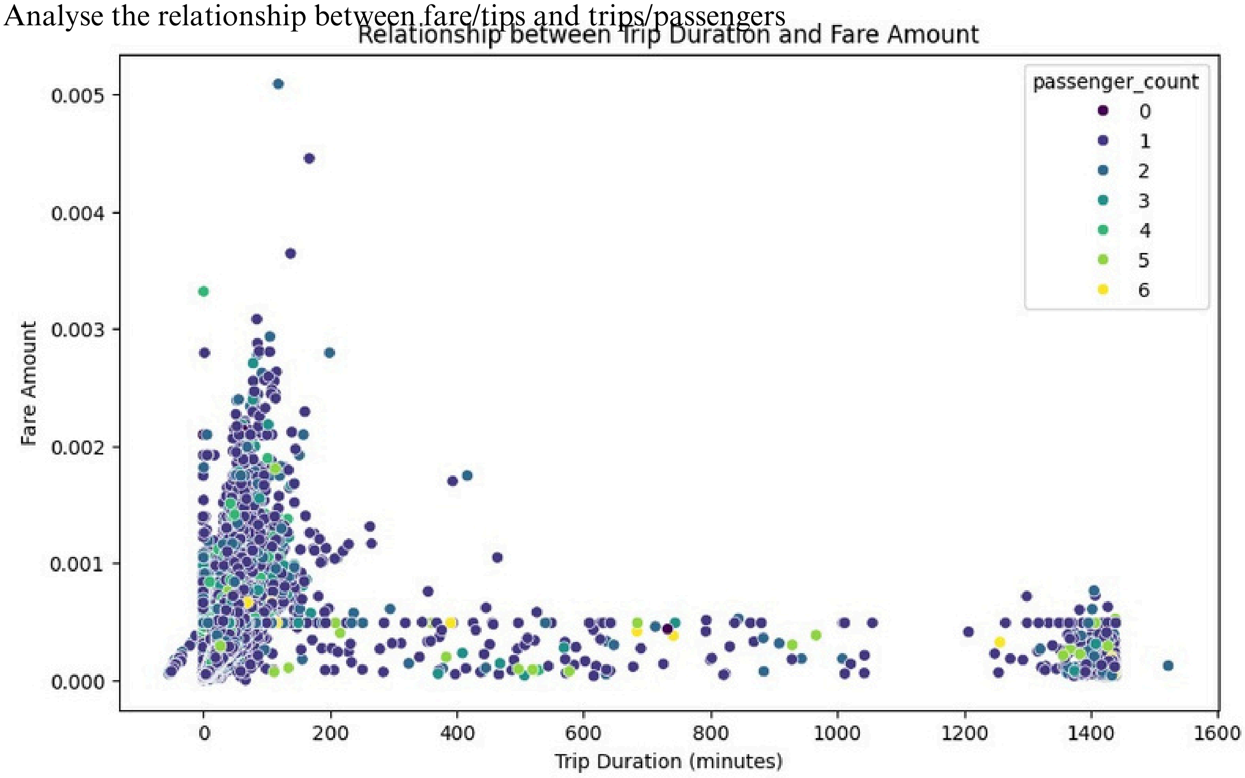


3.1.6. Analyse and visualise the relationship between distance and fare amount

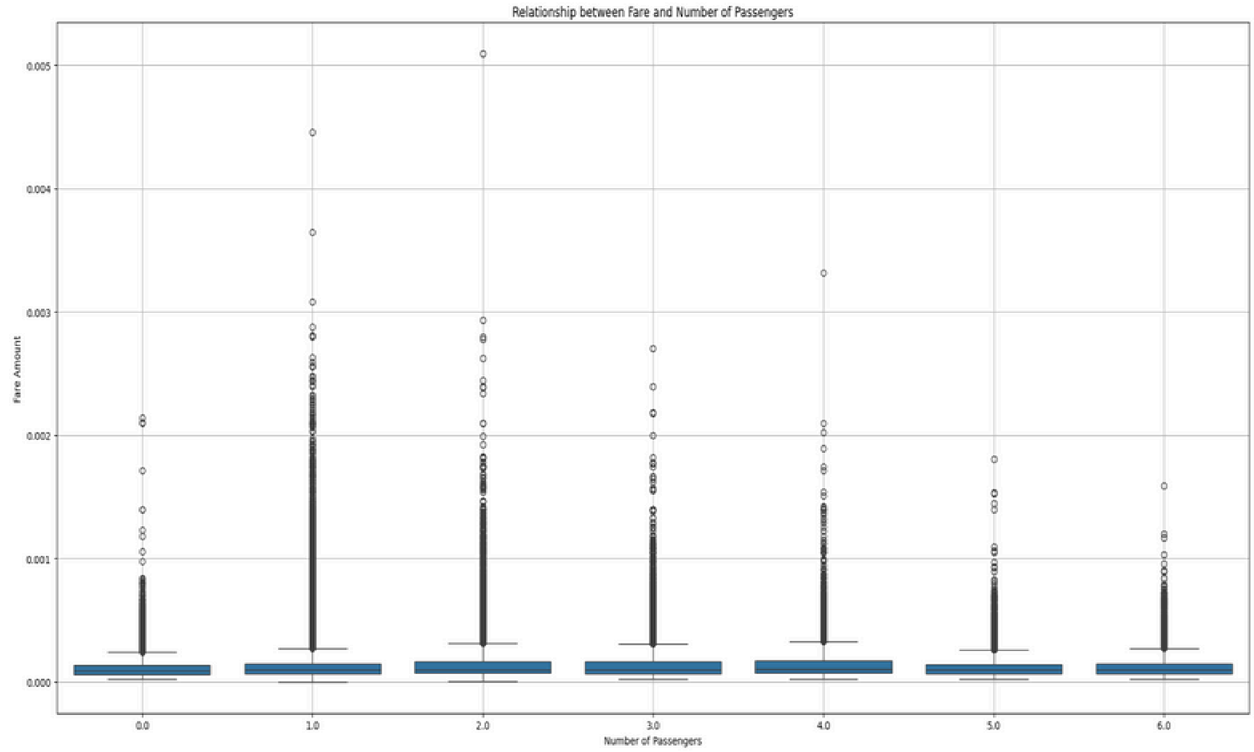


```
Correlation between Trip Distance and Fare Amount: 0.95
```

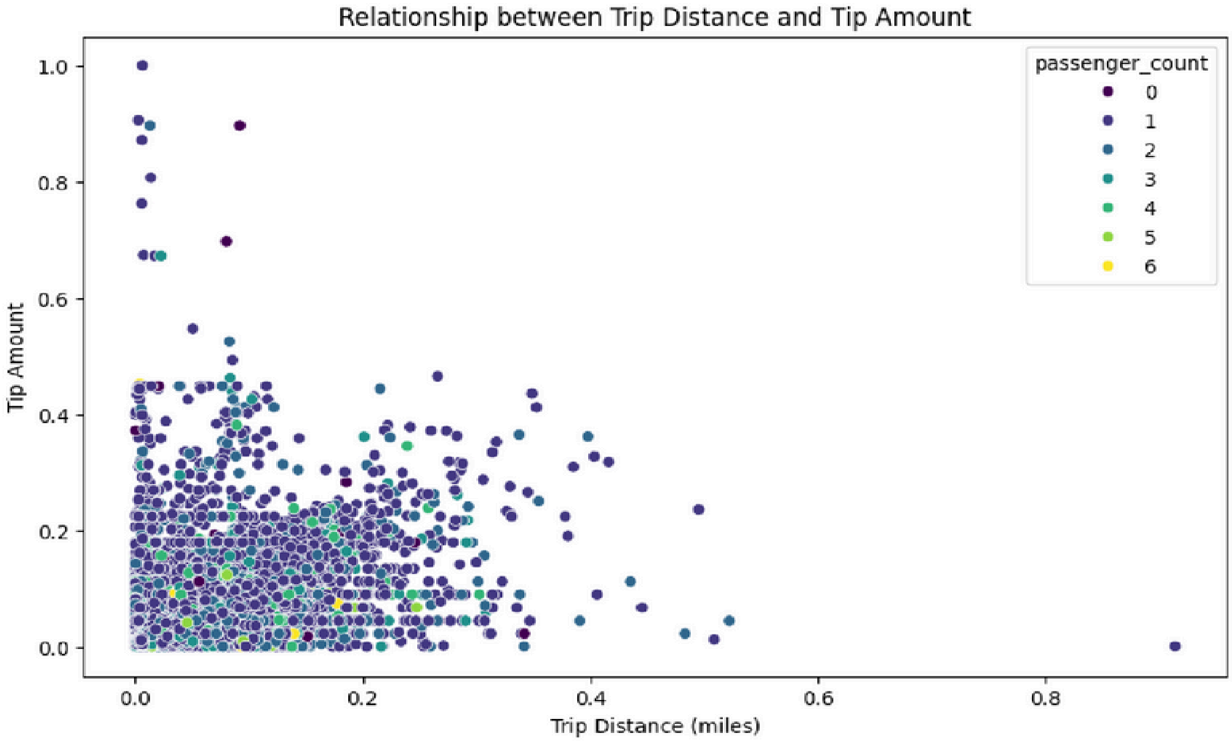
3.1.7. Analyse the relationship between fare/tips and trips/passengers



```
Correlation between Trip Duration and Fare Amount: 0.33
```

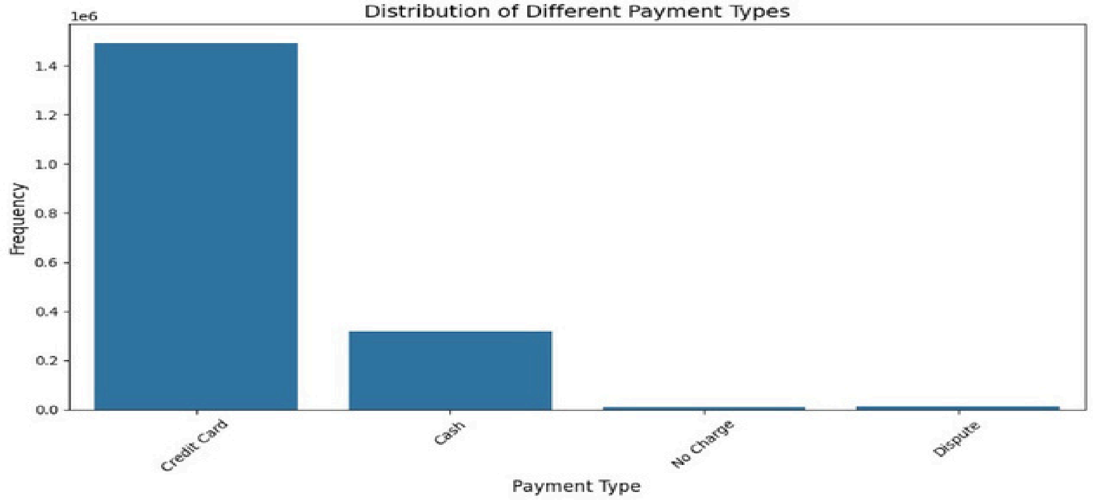


```
Correlation between Passenger Count and Fare Amount: 0.0403
```

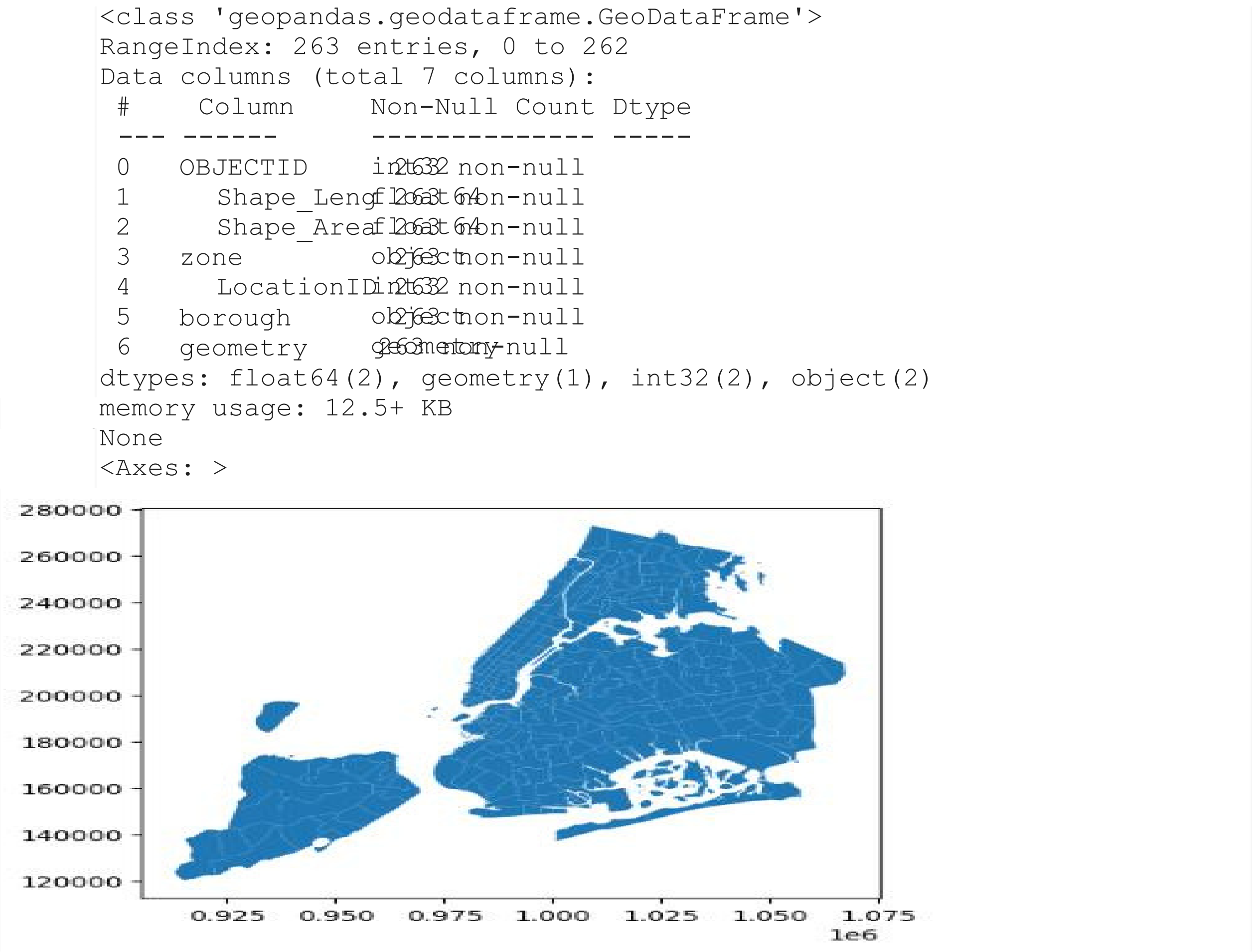


```
Correlation between Trip Distance and Tip Amount: 0.80
```

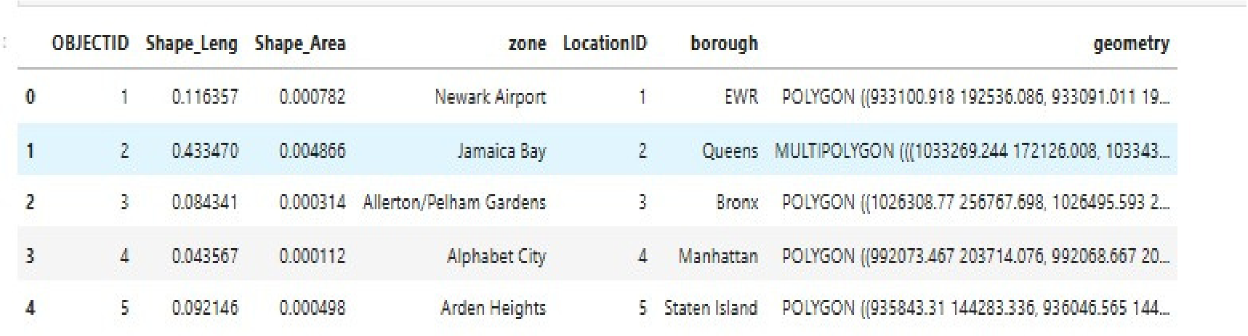
3.1.8. Analyse the distribution of different payment types



3.1.9. Load the taxi zones shapefile and display it



```
<class 'geopandas.geodataframe.GeoDataFrame'>
RangeIndex: 263 entries, 0 to 262
Data columns (total 7 columns):
 #    Column     Non-Null Count Dtype
 --- ------      -------------- -----
 0   OBJECTID    int32 263 non-null
 1     Shape_Leng float64 263 non-null
 2     Shape_Area float64 263 non-null
 3   zone        object 263 non-null
 4     LocationID int32 263 non-null
 5   borough     object 263 non-null
 6   geometry    geometry 263 non-null
dtypes: float64(2), geometry(1), int32(2), object(2)
memory usage: 12.5+ KB
None
<Axes: >
```



	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...

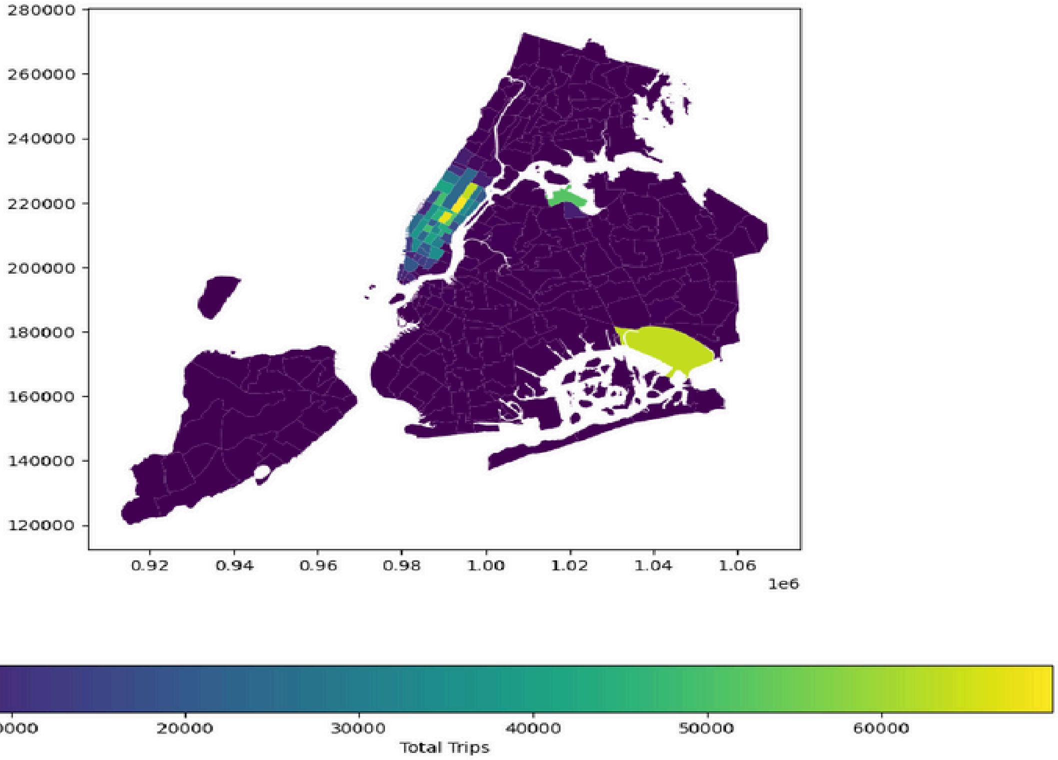
3.1.10. Merge the zone data with trips data

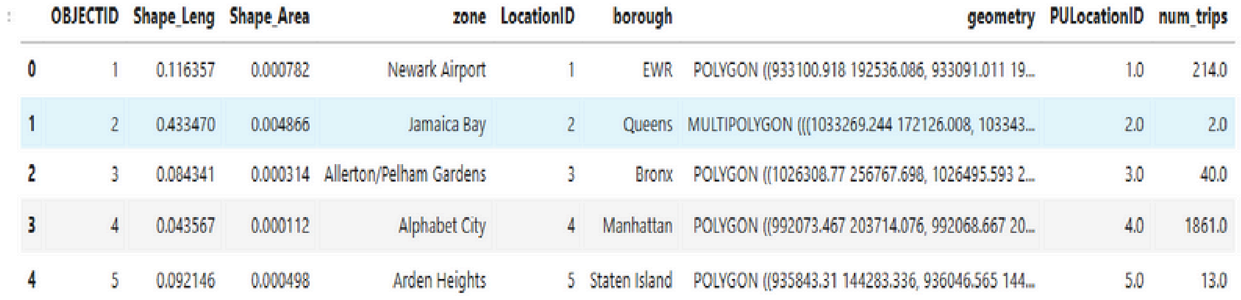
The zones dataset was merged into the trip dataset using the locationID from the zones data and the PULocationID from the trip data as the key columns.

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

PULocationID	***total_Trips***
1	35
2	2
4	1403
6	1
7	253

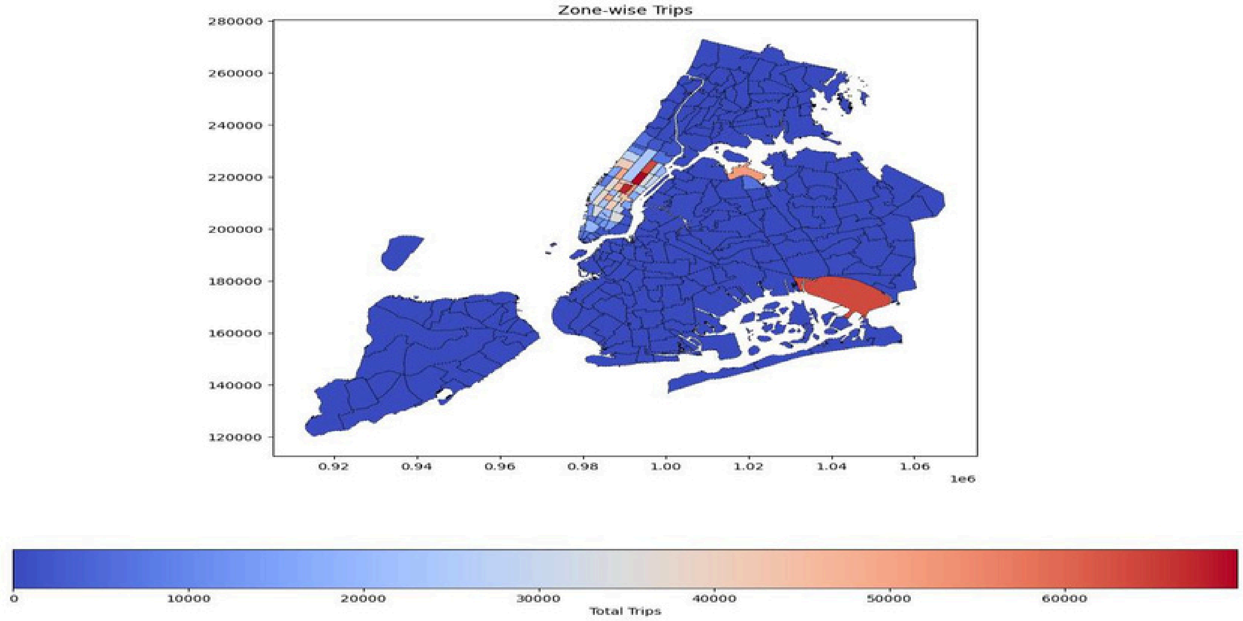
3.1.12. Add the number of trips for each zone to the zones dataframe





	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

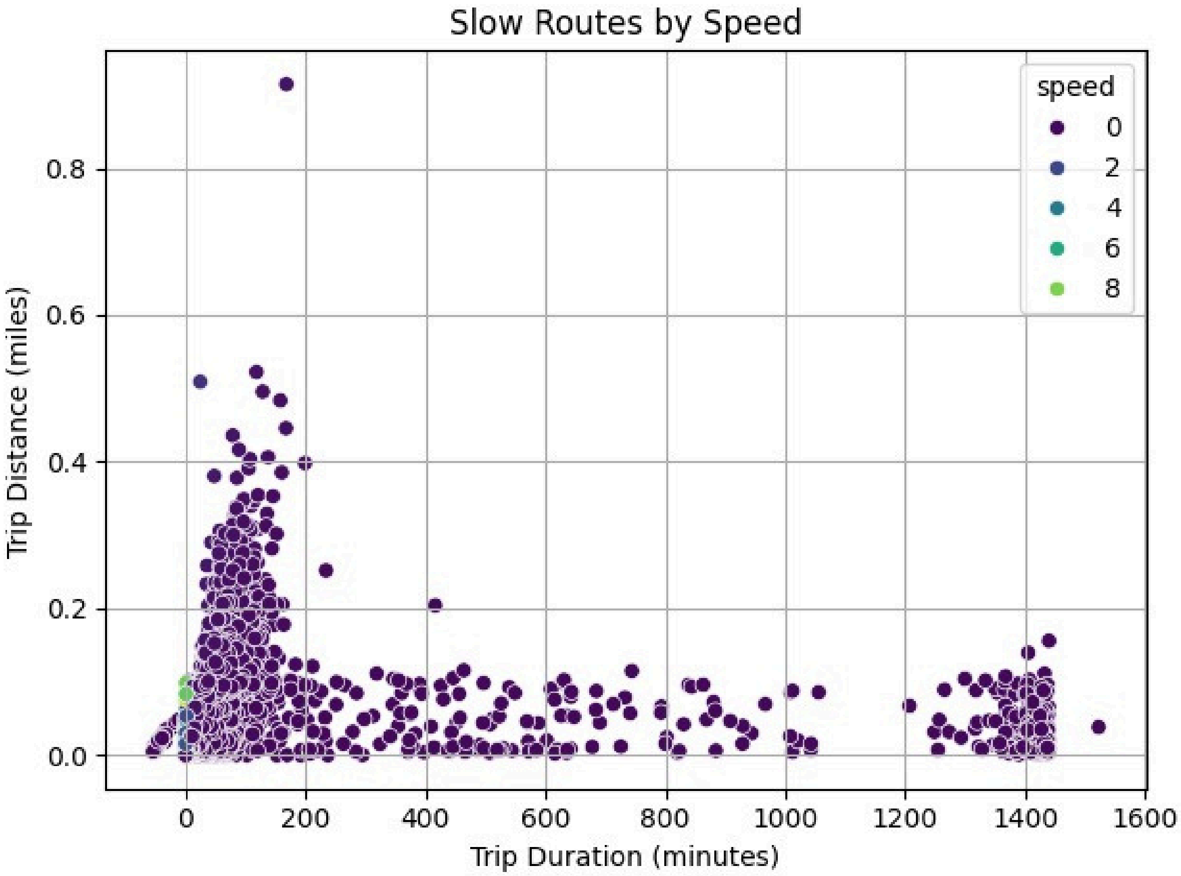
Conclude with results A strong positive correlation was observed between distance and fare, indicating that fares are predominantly distance-driven.

- Weekday peak hours align with rush hour traffic, while weekends exhibit a rise in late-night activity.
- Airport and Midtown zones show the highest concentration of pickups and drop-offs.
- The majority of trips involve 1–2 passengers, with credit cards being the most common payment method.
- Seasonal patterns emerged, with Q3 (July–September) identified as the busiest quarter.
- Rigorous data cleaning was conducted to remove anomalies and standardize numeric features, thereby enhancing the accuracy and reliability of the analysis.

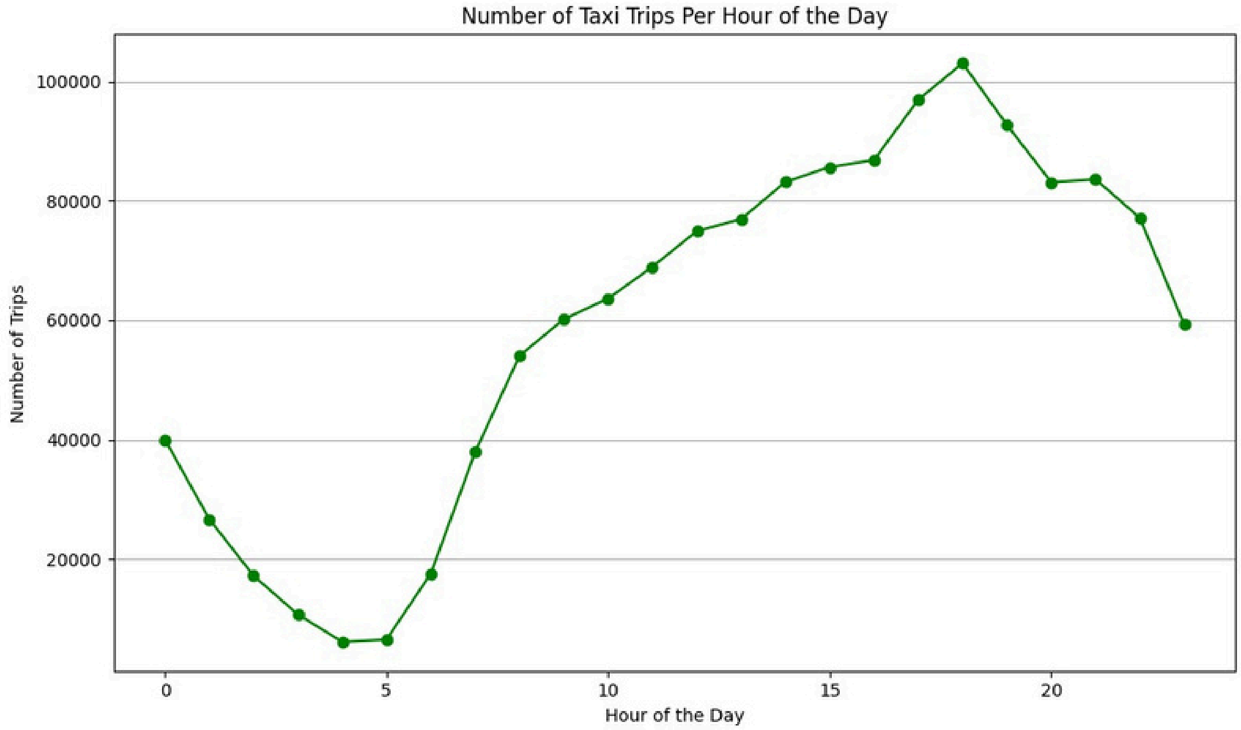
3.2. Detailed EDA: Insights and Strategies

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

PULocationID	DOLocationID	tpep_pickup_datetime	trip_duration_derived	trip_distance	speed
1	1	2023-02-06 16:26:31	0.116667	0.000293	0.150626
1	1	2023-02-14 13:13:04	0.116667	0.000244	0.125521
1	1	2023-03-06 12:55:36	0.316667	0.000244	0.046245
1	1	2023-03-09 19:02:51	0.083333	0.000195	0.140584
1	1	2023-03-24 11:41:59	0.116667	0.000244	0.125521



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

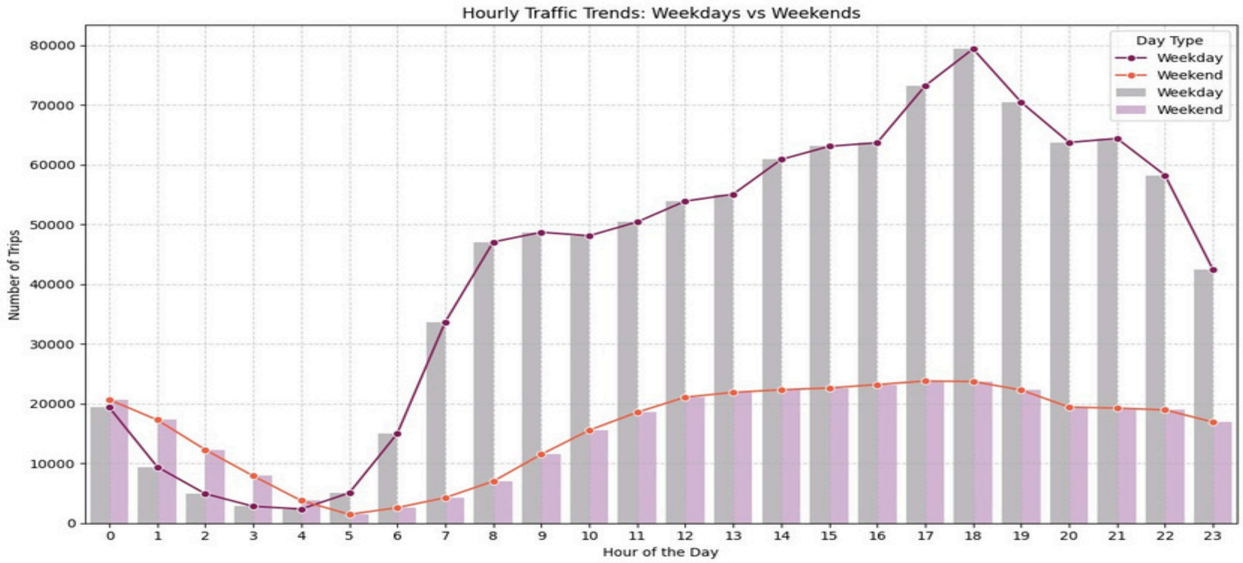


```
The five busiest hours:
pickup_hour
18 17 19 103059
15        96953
Name: count, dtype: int64
       92730
       86841
       85666
```

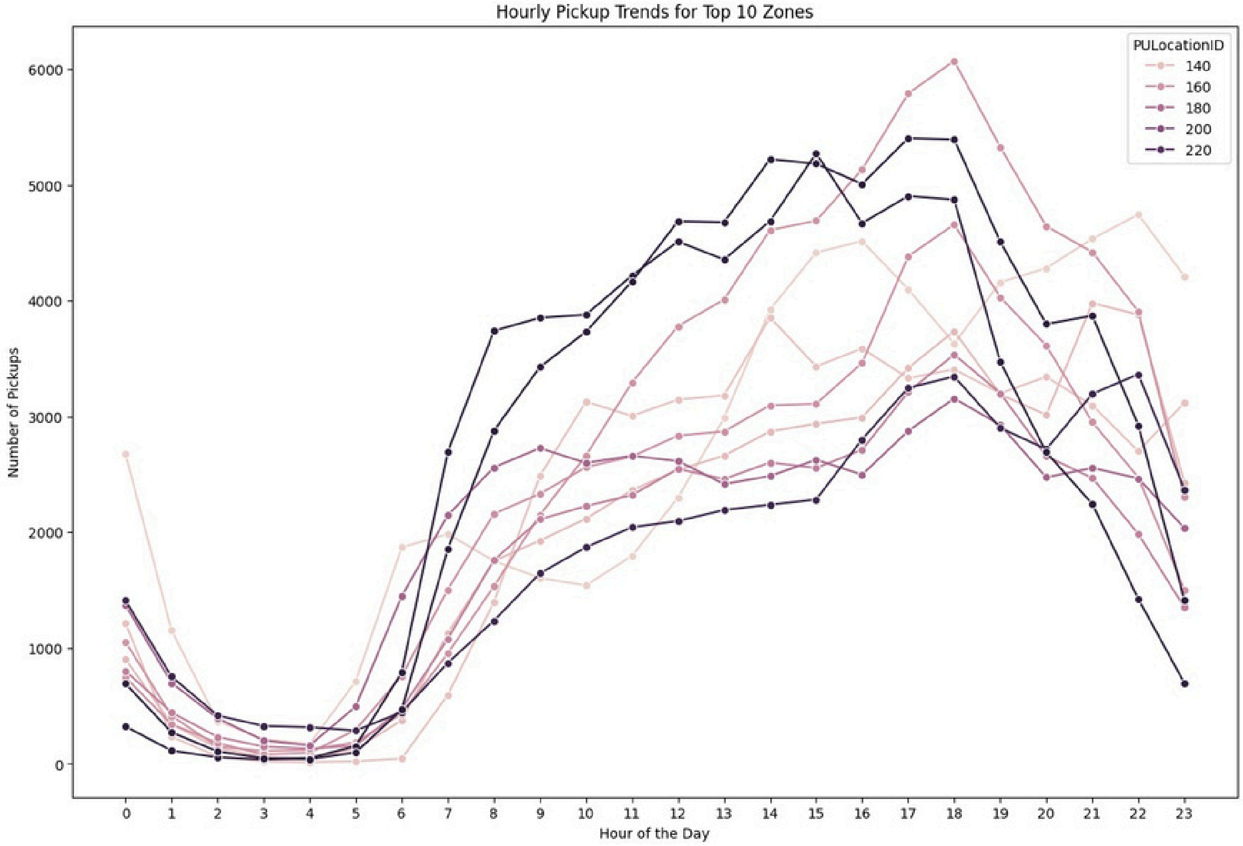
3.2.3. Scale up the number of trips from above to find the actual number of trips

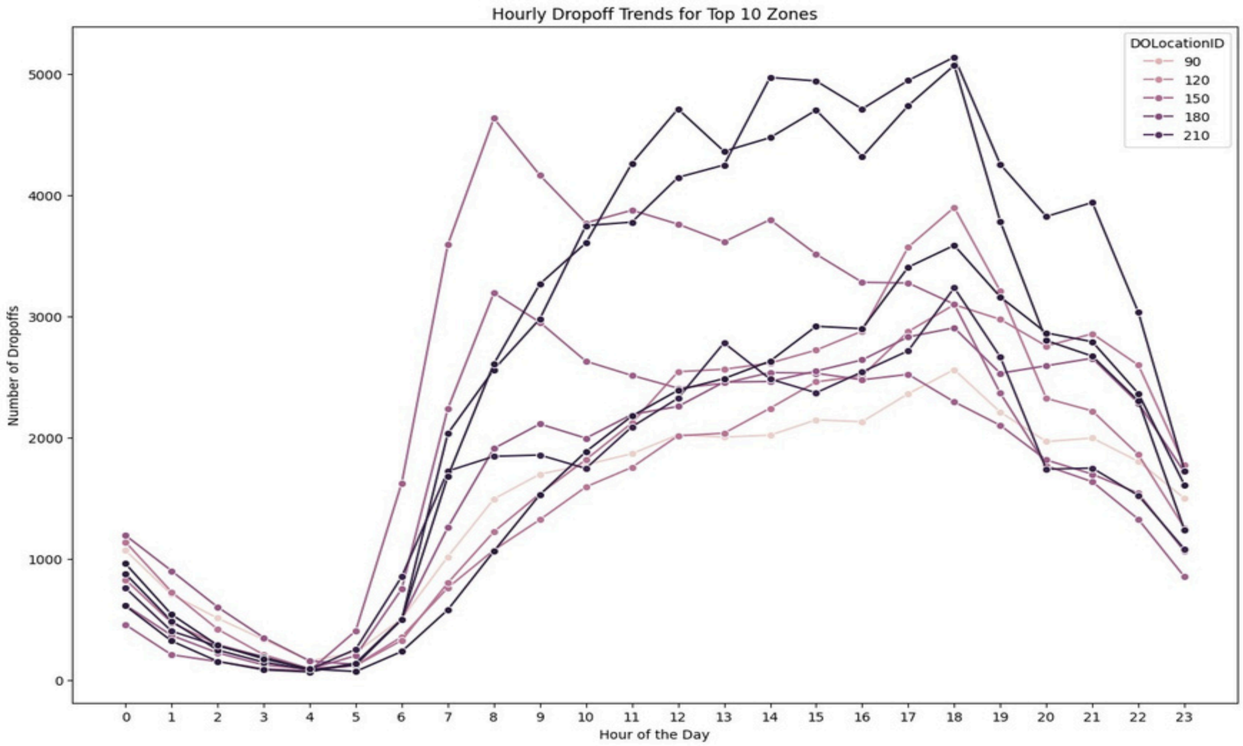
```
Actual number of trips in the five busiest hours (scaled from sample):
pickup_hour
18     2061180
17     1939060
19     1854600
16     1736820
15     1713320
Name: count, dtype: int64
```

3.2.4. Compare hourly traffic on weekdays and weekends

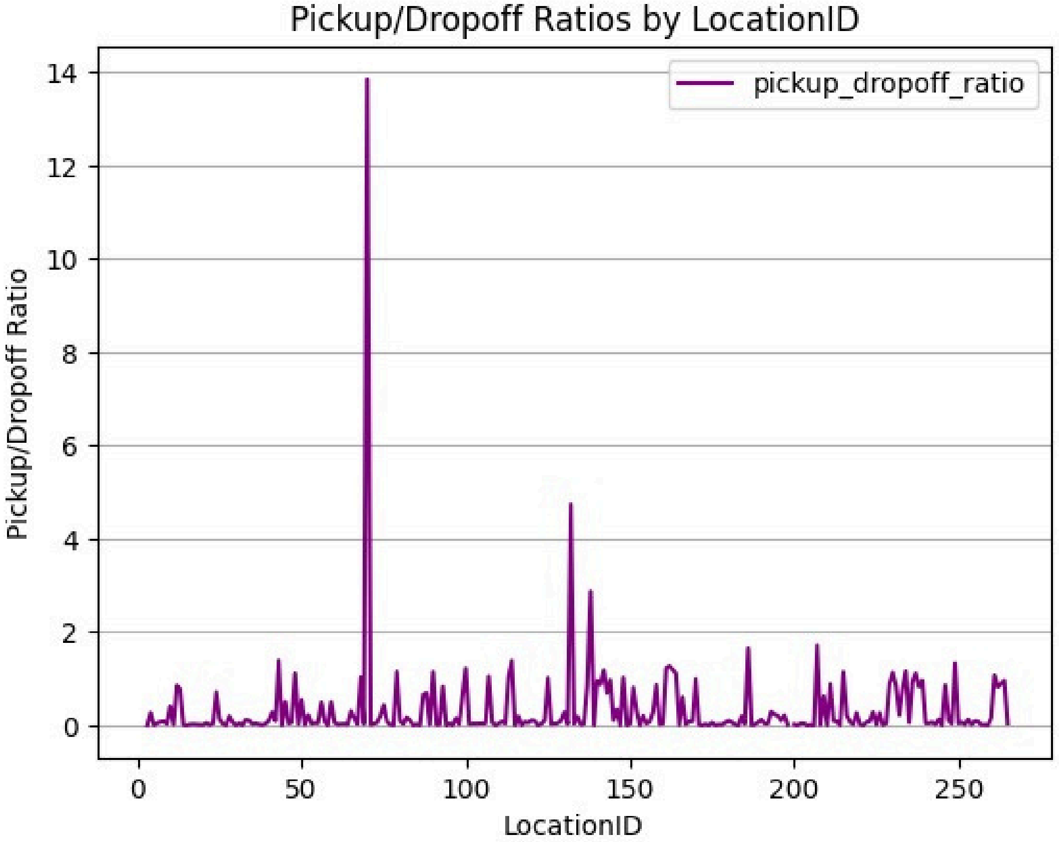


3.2.5. Identify the top 10 zones with high hourly pickups and drops

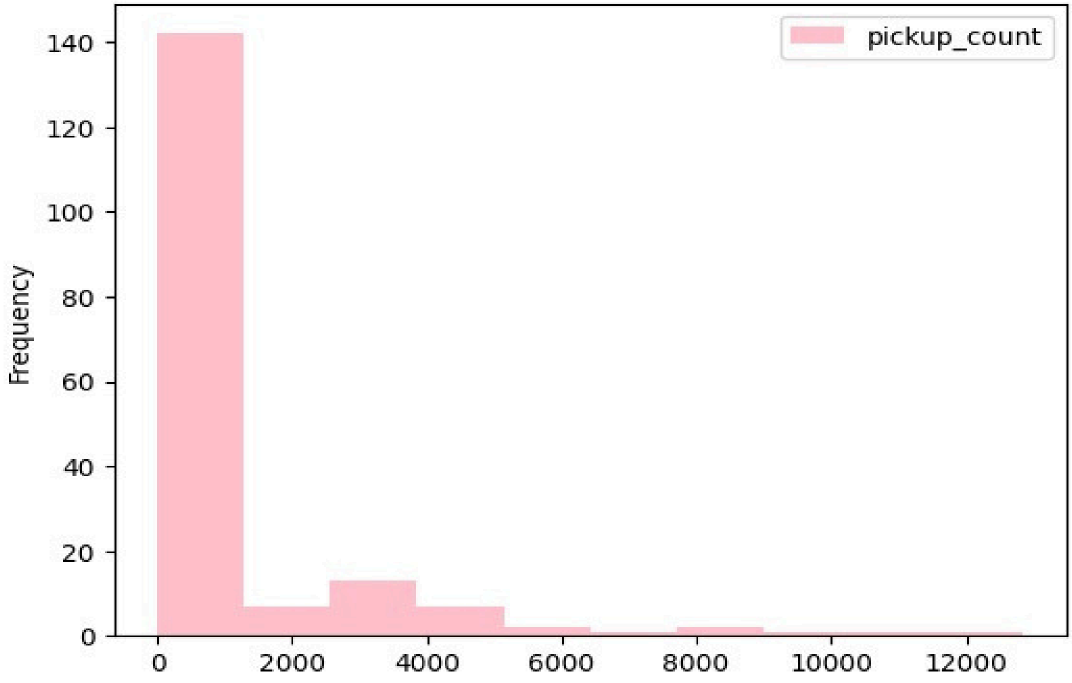


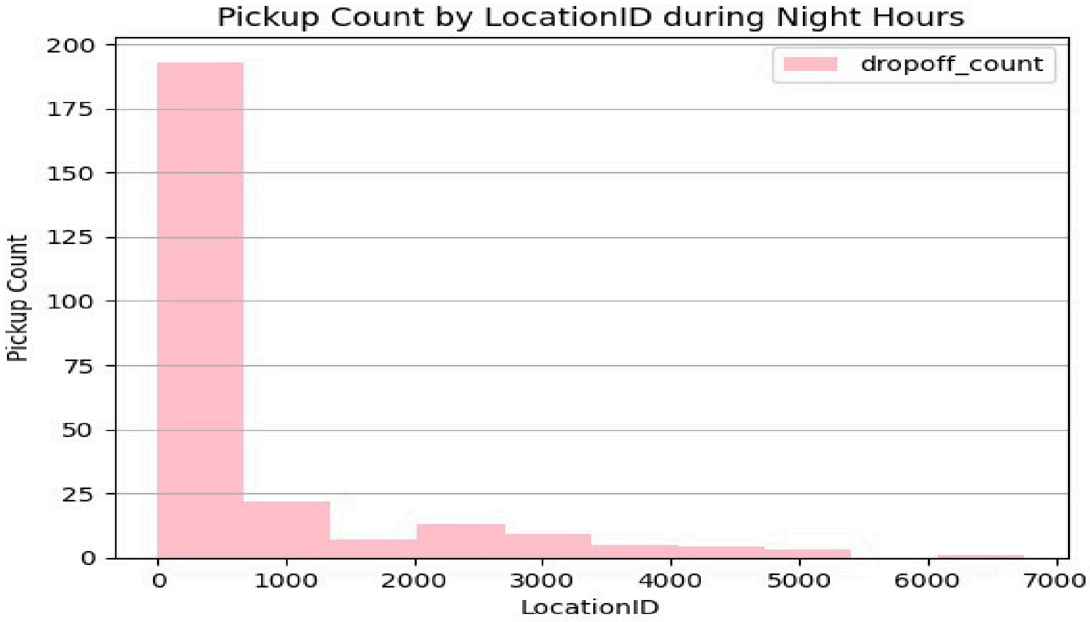


3.2.6. Find the ratio of pickups and dropoffs in each zone

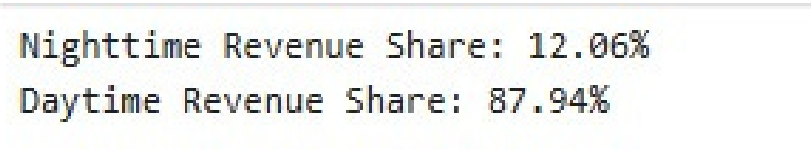


3.2.7. Identify the top zones with high traffic during night hours



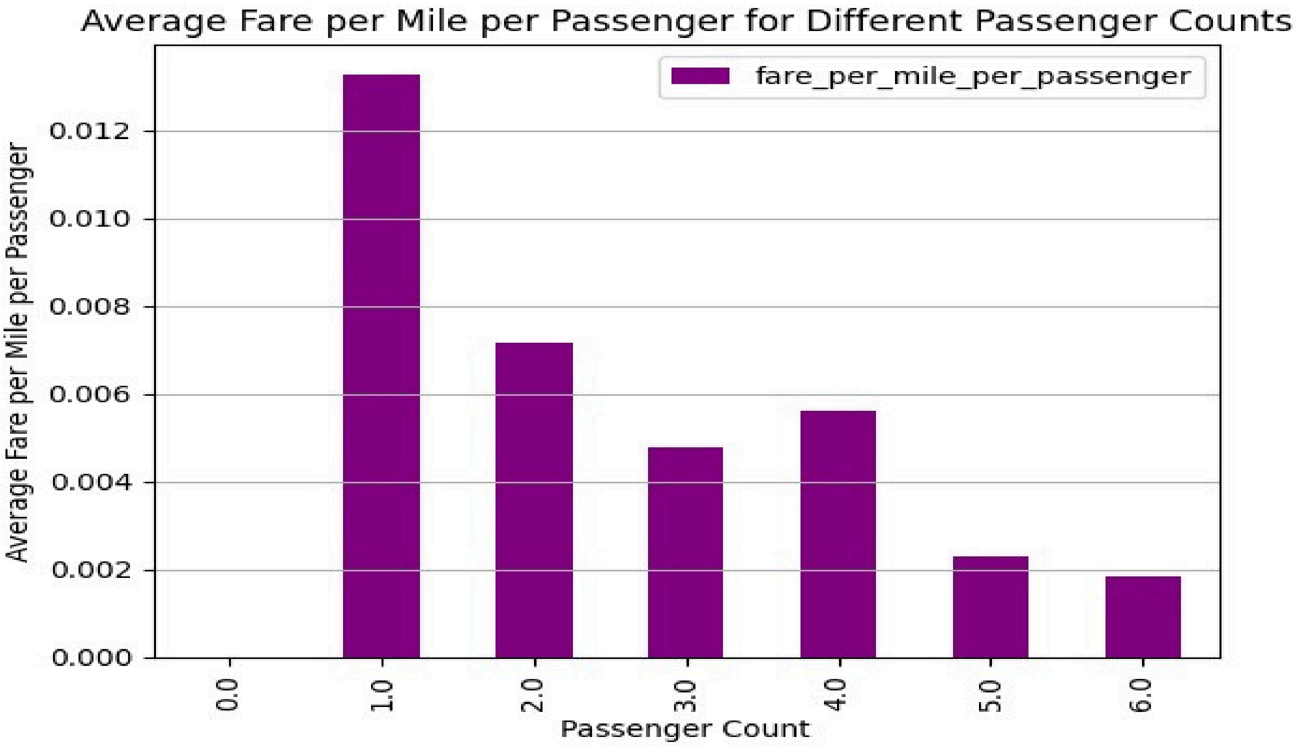


3.2.8. Find the revenue share for nighttime and daytime hours

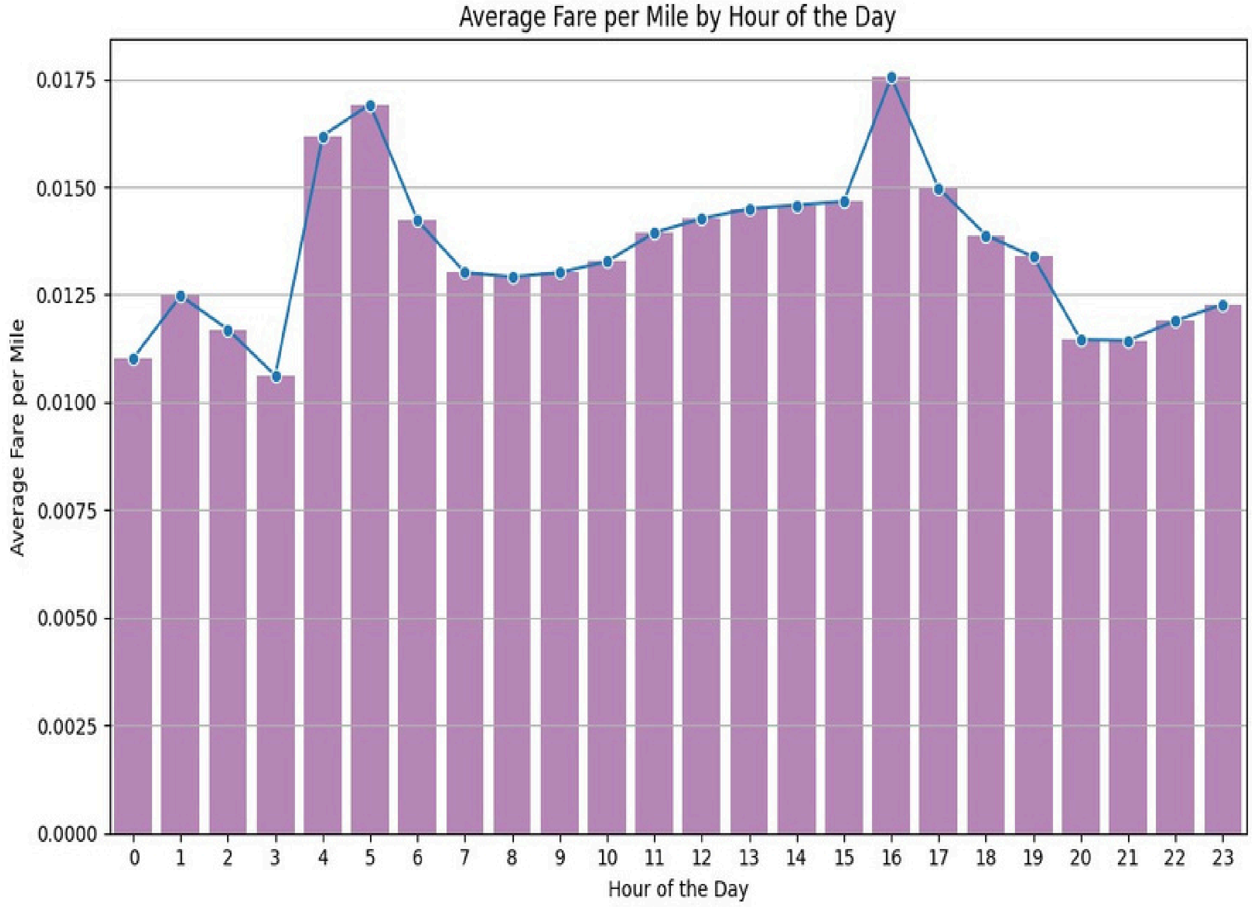


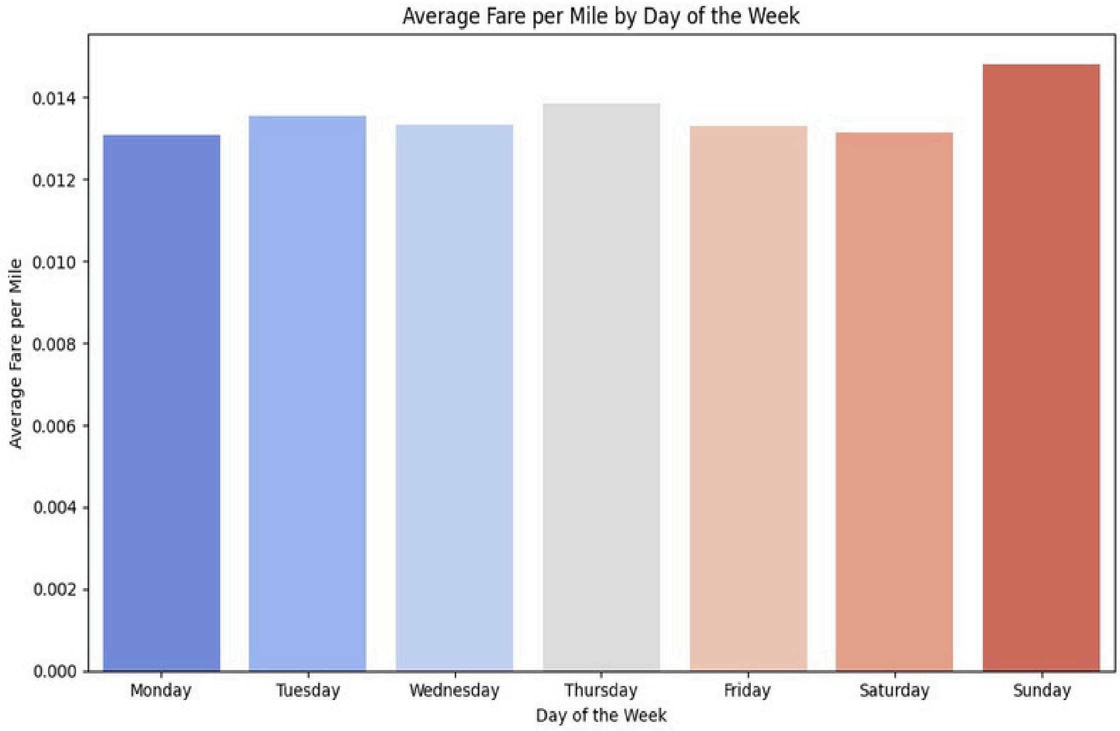
```
Nighttime Revenue Share: 12.06%
Daytime Revenue Share: 87.94%
```

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

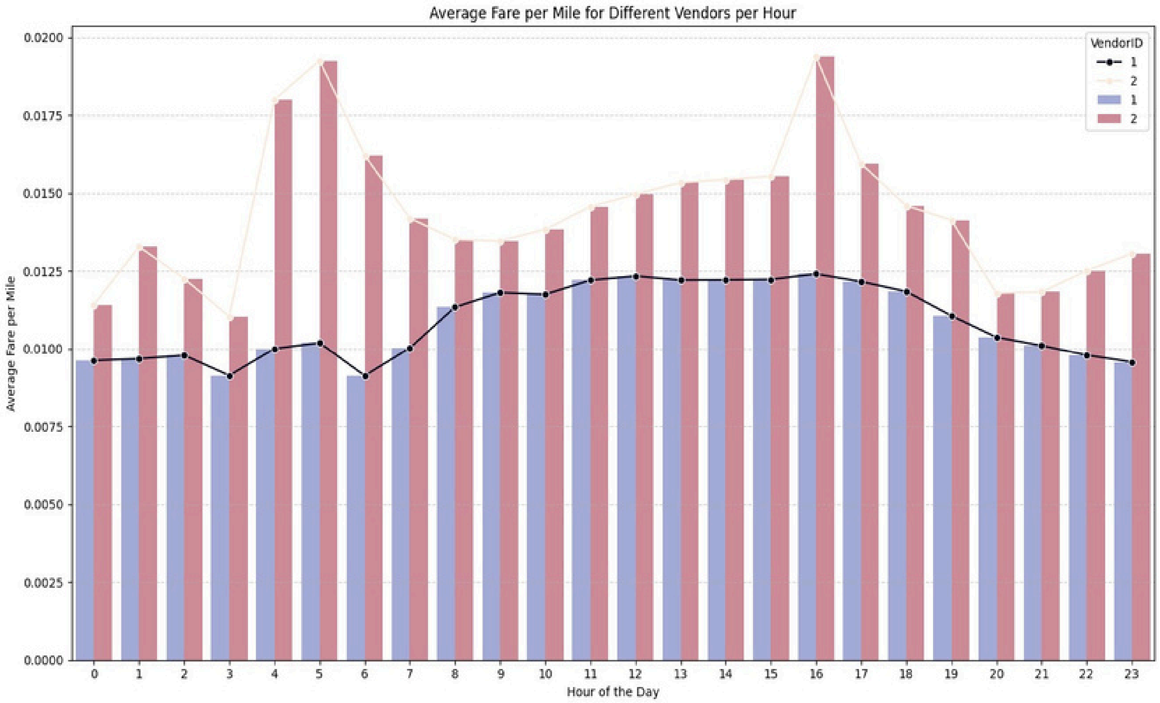


3.2.10. Find the average fare per mile by hours of the day and by days of the week

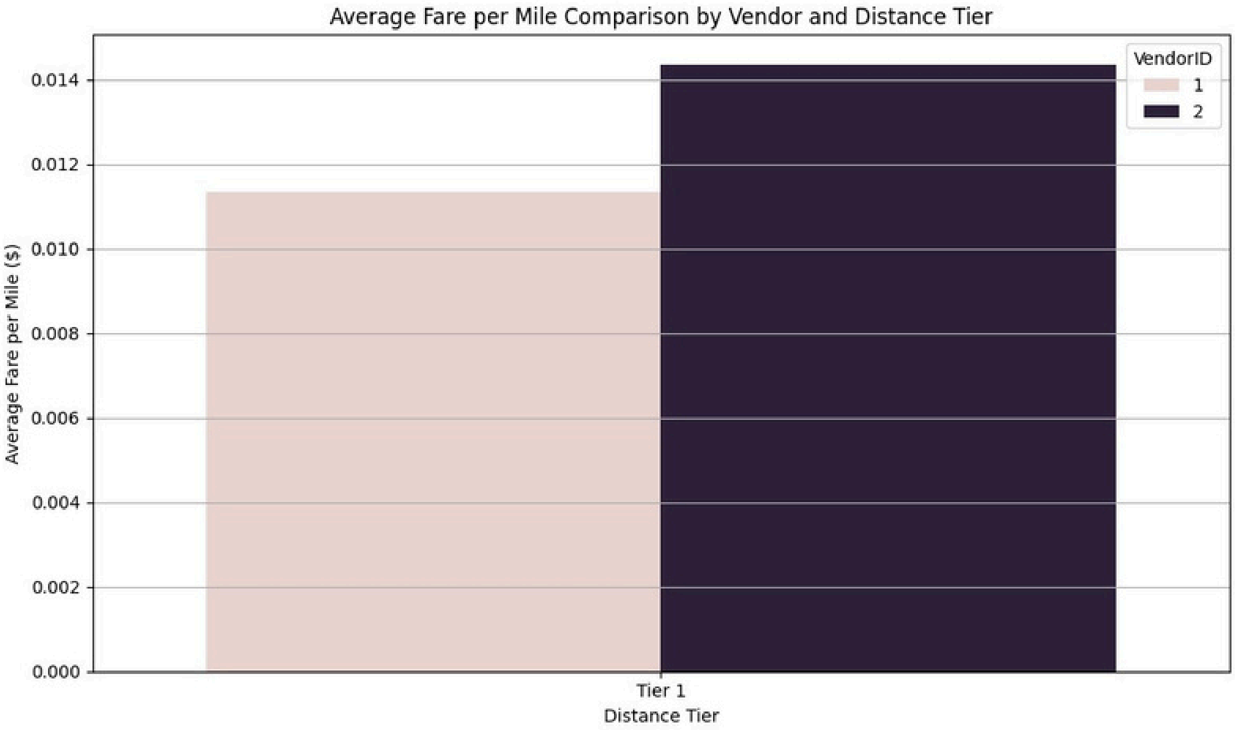




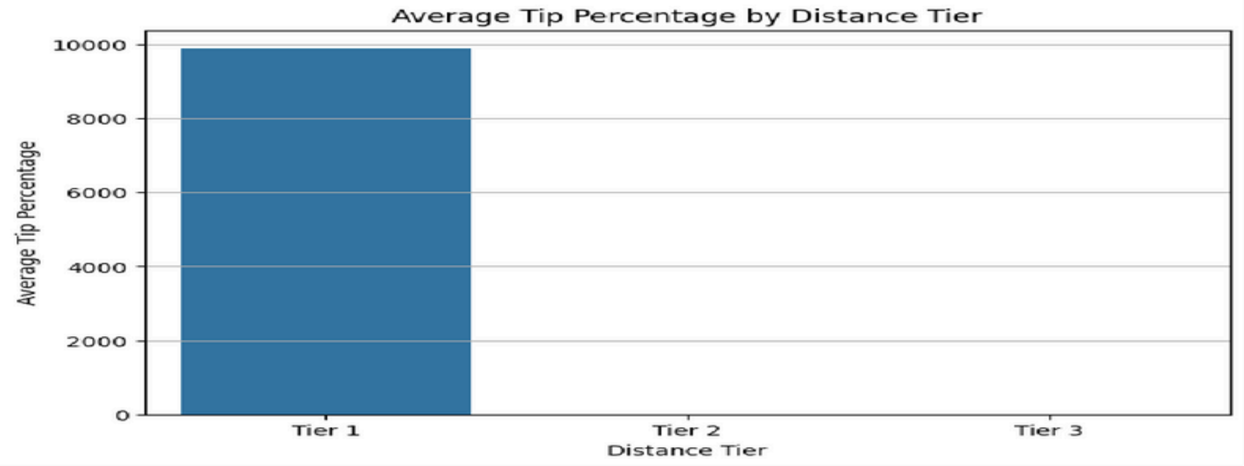
3.2.11. Analyse the average fare per mile for the different vendor

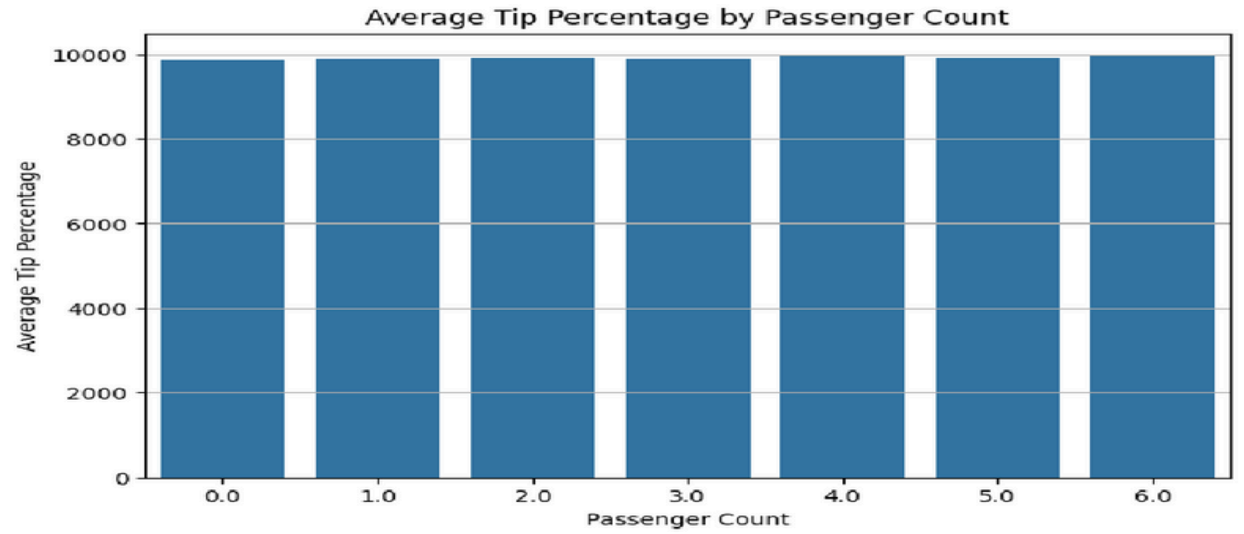


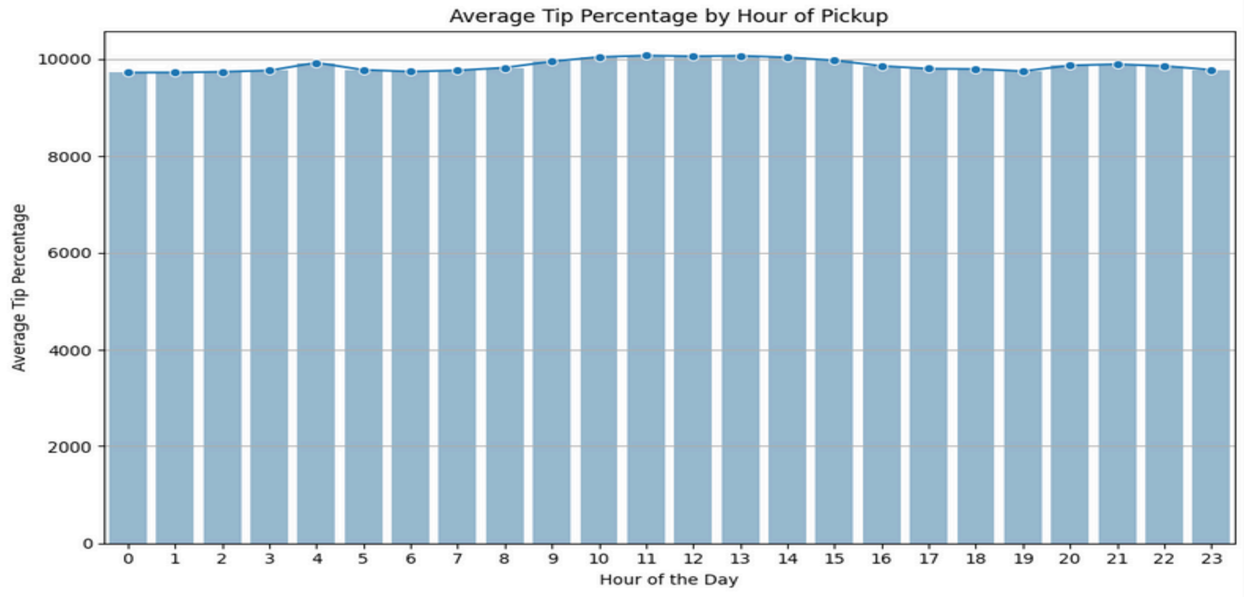
3.2.12. Compare the fare rates of different vendors in a distance-tiered fashion



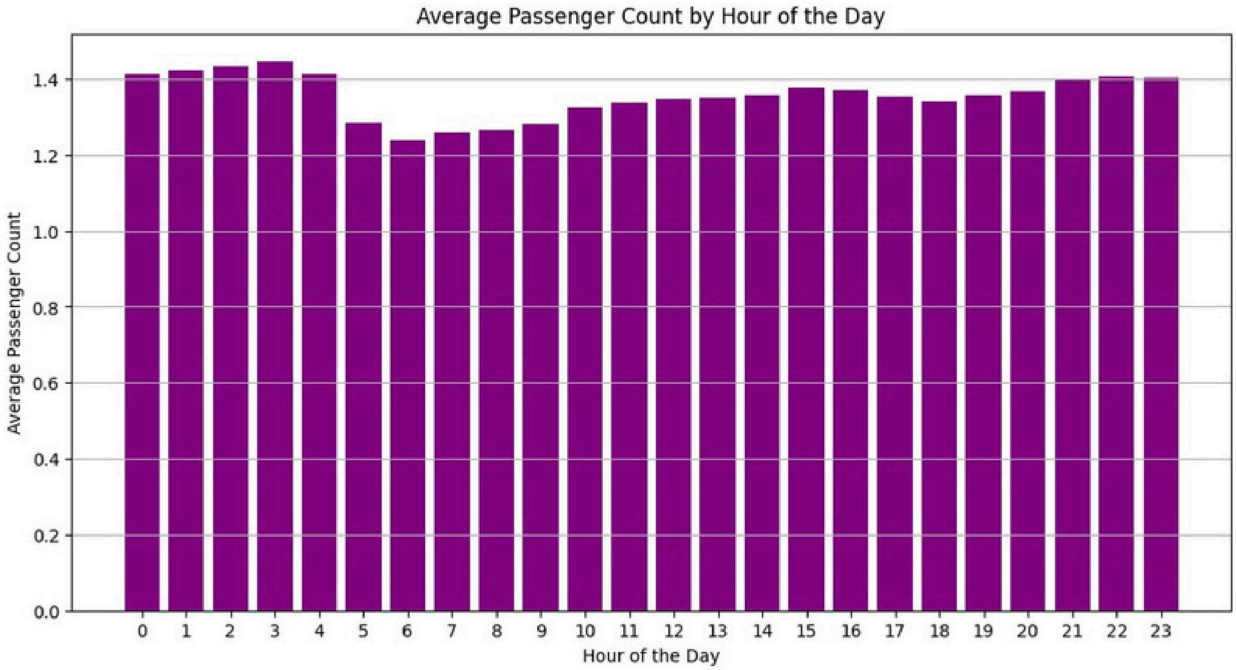
3.2.13. Analyse the tip percentages



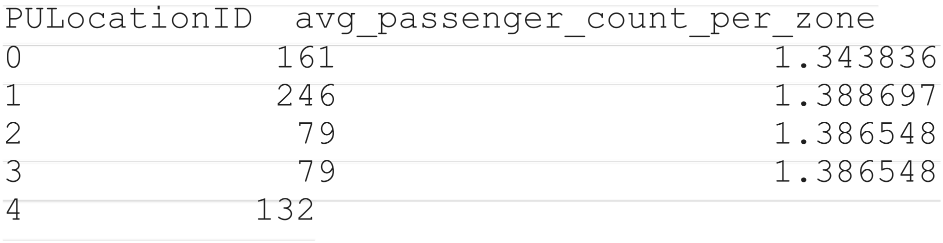




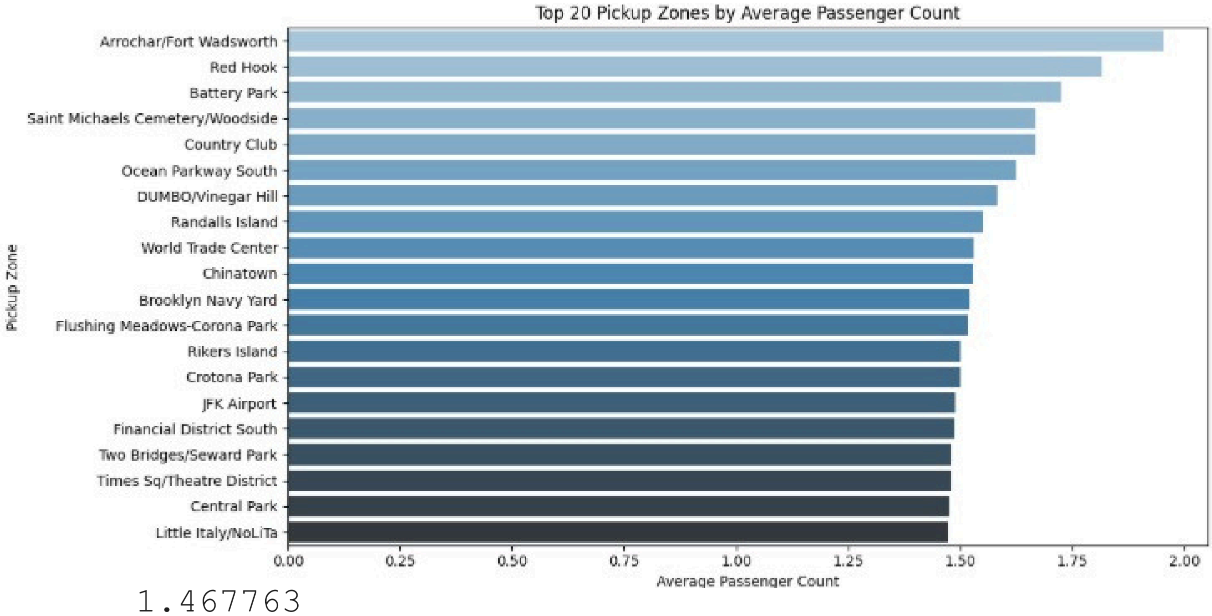
3.2.14. Analyse the trends in passenger count



3.2.15. Analyse the variation of passenger counts across zones

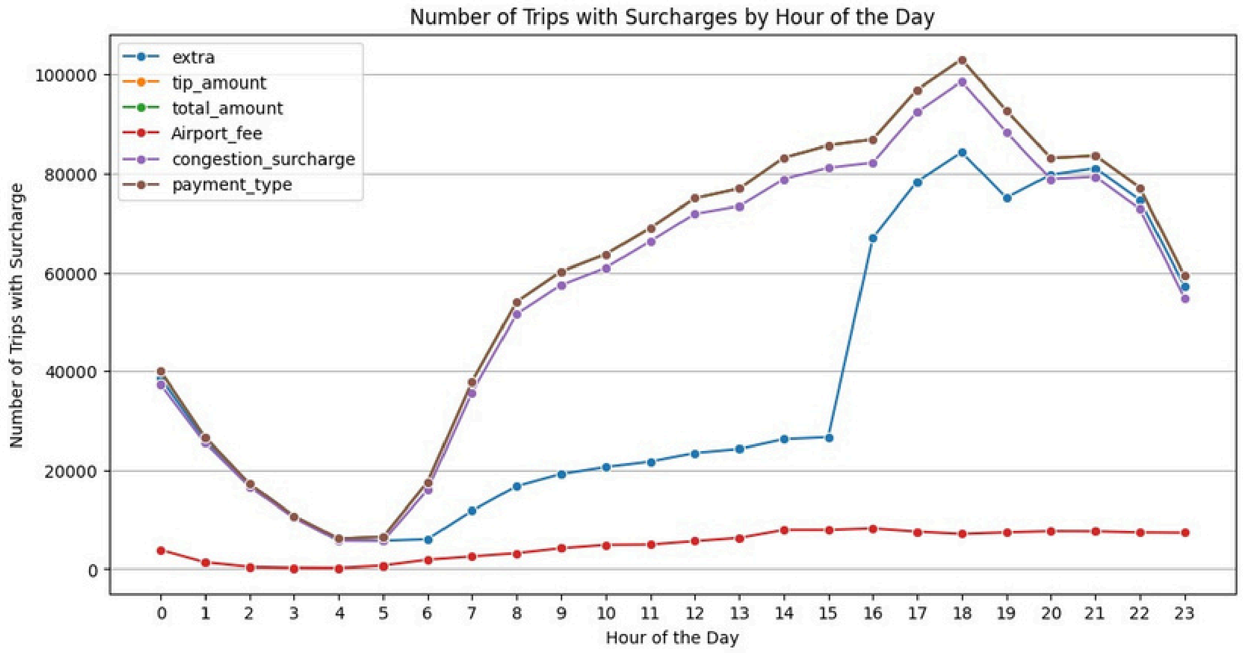


```
PULocationID  avg_passenger_count_per_zone
0          161                      1.343836
1          246                      1.388697
2           79                      1.386548
3           79                      1.386548
4         132
```



```
1.467763
```

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



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 to Optimize Routing and Dispatching: Increase Cab Availability During

□ Peak Daytime Hours (Based on Section 3.2.2) Deploy
additional cabs between 6:00 AM and 10:00 PM to address high demand during peak daytime
periods. This adjustment will help reduce passenger wait times and improve service efficiency.

□ Implement Surge Pricing in High-Demand Zones During Peak Hours Introduce dynamic pricing
in areas experiencing high daytime demand to better balance supply and demand. This
incentivizes drivers to move toward high-demand zones and increases profitability during

□ busy
hours.

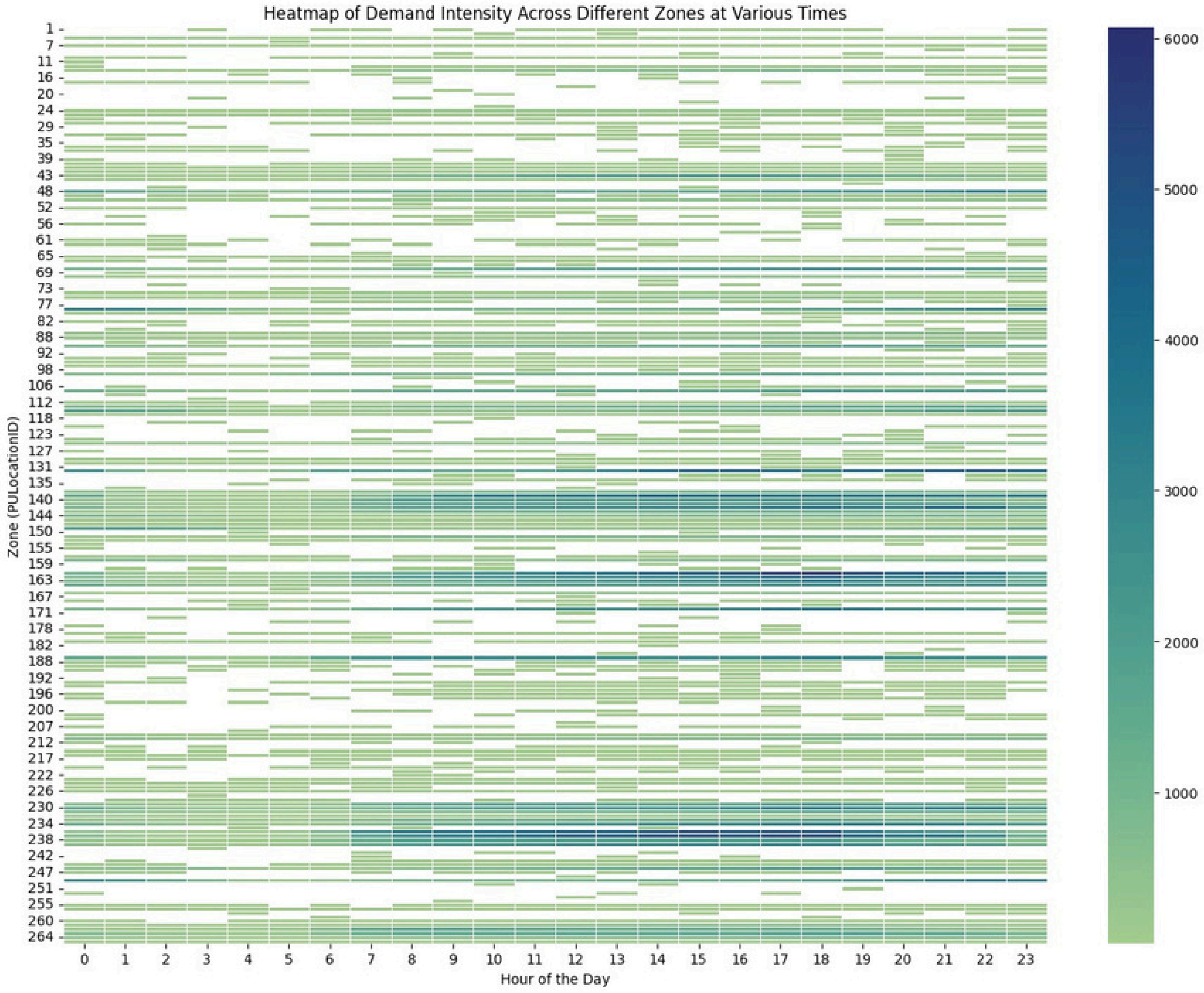
Adjust Fare Rates Based on Time of Day and Day of the Week (Sections 3.2.4 & 3.2.10)
Analyze average fare-per-mile trends to implement time-based and day-based pricing strategies.
For example, apply higher rates during weekend evenings and weekday rush hours, while
offering discounts during low-demand periods.

□ Expand Nighttime Coverage in High-Demand Zones (Section 3.2.7) Increase the number of active cabs between 11:00 PM and 5:00 AM in zones with consistent late-night demand. This improves service coverage during less active hours and caters to nightlife, airport, and shift-worker travel needs.

□ Implement Intelligent Repositioning Algorithms Introduce routing algorithms that automatically

reposition idle or underutilized cabs to areas with anticipated demand surges. This data-driven
dispatching approach enhances operational efficiency and maximizes cab utilization.

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.



Dentify High-Demand Zones by Time of Day ▪ Use the heatmap of trip counts across pickup zones and hours to pinpoint zones with consistently high demand (e.g., during morning and evening rush hours). ▪ Example: Zones showing strong activity between 7–10 AM and 4–8 PM should have increased cab presence during those intervals.

□Weekday vs. Weekend Demand Patterns(Derived from Section 3.2.4) ▪ Weekday mornings and evenings often correspond to work-related commutes—focus on business districts and transportation hubs. ▪ Weekend demand may shift toward leisure zones (e.g., shopping, entertainment areas)—redeploy fleet to match this spatial pattern.

Match Cab Types with Trip Distances

▪ Position shorter-trip focused vehicles (e.g., sedans) in zones with high short-distance demand. ▪ Use larger or premium cabs in zones with longer average trip distances to optimize cost-efficiency and customer service.

Rebalancing Through Predictive Dispatch

▪ Use real-time data and historical patterns to reposition idle cabs to areas with expected demand surges. ▪ For example, after morning peaks in residential areas, shift cabs toward commercial districts for afternoon coverage. Continuous Monitoring and Feedback

▪ Update zone-based positioning strategies regularly based on ongoing trip trend data. ▪ Integrate feedback loops to fine-tune allocation by time, day, and seasonal demand fluctuations. Visual Aid Reference:

The heatmap titled "Heatmap of Demand Intensity Across Different Zones at Various Times" effectively reveals temporal and spatial trip density, guiding evidence-based cab deployment strategies.

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

The pricing strategy to maximize revenue while maintaining competitive rates with other vendors:

⌧ Monthly revenue is very low in July, August, September company can offer competative price as compared to other vendor during these month which can incrase pickup during that time and also revenue will increase

⌧Correlation between Trip Duration and Fare Amount is 0.32 which is very low. Company can impose waiting charge for the ride which will increse the corrreation between these two variables.

⌧Fare amount depended on count of pessenger can also increase the revenue for the company.

⌧Consider using machine learning models to predict demand elasticity for various distances. This would allow more precise price adjustments.