

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

```
<class 'pandas.core.frame.DataFrame'>
Index: 3041714 entries, 0 to 3066765
Data columns (total 19 columns):
#   Column                Dtype
---  -
0   VendorID              int64
1   tpep_pickup_datetime  datetime64[us]
2   tpep_dropoff_datetime datetime64[us]
3   passenger_count       float64
4   trip_distance         float64
5   RatecodeID            float64
6   store_and_fwd_flag    object
7   PULocationID          int64
8   DOLocationID          int64
9   payment_type          int64
10  fare_amount           float64
11  extra                 float64
12  mta_tax               float64
13  tip_amount            float64
14  tolls_amount          float64
15  improvement_surcharge float64
16  total_amount          float64
17  congestion_surcharge  float64
18  airport_fee           float64
dtypes: datetime64[us](2), float64(12), int64(4), object(1)
memory usage: 464.1+ MB
```

There are ~3.04 million rows and 19 columns in the data set for one month. When the data for all the months is combined we would have roughly 35 – 40 million rows. Such a large data would not be computationally feasible as there would be high RAM usage, slow groupby and visualizations, notebook could crash or kernel restarts, would be an overkill for trend analysis.

1.1.1. Sample the data and combine the files

We sample by both date and hour, but not just by hour, because sampling by hour would give us the sample data which includes data samples collected from all dates for that specific hour. But our requirement is to have the data samples for each date and each hour in that date. So in this case we have to sample both by hour and by date.

1.1.2. To sample the data, I first created a file list containing the parquet files with each months data. Then for each file, I created two columns in the dataset namely pickup_date and pickup_hour to explicitly record the date and hour for each record, Then for each date I created a dataframe called date_df and for each hour called hour_df. Then I sampled the data from hour_df and added the data to the sampled_data dataframe which we had previously created. After the samples for all the dates for all the hours have been added to the sampled_data, I concatenated this data with empty dataframe df which we had initialized earlier to form the combined dataset.

1.1.3. The dataframe df had 1896400 rows, so in order to reduce the records to somewhere between 250000 – 300000, I again sampled the data with n = 300000 and the final data had 300000 records and stored as a parquet file for efficient reuse in subsequent analysis steps.

Total rows before final sampling: 1896400

Total rows after final sampling: 300000

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

Loaded the parquet file with all the sampled data into a new dataframe df_new. df_new had a RangeIndex: 300000, 0 to 299999 entries. So my index was clean. So I just ensured that the index was reset properly after all the previous concatenations and samplings.

Then I moved onto removing the unnecessary columns pickup_date and pickup_hour which were created only to help with sampling the data using df_new.drop() method. After dropping the 2 columns the dataframe df_new had 20 columns

```
<class 'pandas.core.frame.DataFrame'>
```

RangeIndex: 300000 entries, 0 to 299999

Data columns (total 20 columns):

#	Column	Non-Null Count	Dtype
0	VendorID	300000 non-null	int64
1	tpep_pickup_datetime	300000 non-null	datetime64[us]
2	tpep_dropoff_datetime	300000 non-null	datetime64[us]
3	passenger_count	289687 non-null	float64
4	trip_distance	300000 non-null	float64
5	RatecodeID	289687 non-null	float64
6	store_and_fwd_flag	289687 non-null	object
7	PULocationID	300000 non-null	int64
8	DOLocationID	300000 non-null	int64
9	payment_type	300000 non-null	int64
10	fare_amount	300000 non-null	float64
11	extra	300000 non-null	float64
12	mta_tax	300000 non-null	float64
13	tip_amount	300000 non-null	float64
14	tolls_amount	300000 non-null	float64
15	improvement_surcharge	300000 non-null	float64
16	total_amount	300000 non-null	float64
17	congestion_surcharge	289687 non-null	float64
18	airport_fee	23229 non-null	float64
19	Airport_fee	266458 non-null	float64

dtypes: datetime64[us](2), float64(13), int64(4), object(1)

memory usage: 45.8+ MB

2.1.2. Combine the two airport_fee columns

Because of the naming error in the airport fee column there were two columns with the same type of values, namely airport_fee and Airport_fee. So to set this anomaly right, I first filled the null values in airport_fee column with values in Airport_fee column using df_new.fillna() method and then dropped the Airport_fee column using df_new.drop() method.

After all this had been done df_new had 19 columns with 300000 rows.

After that, I checked for negative values in the fare_amount column and my df_new had no such records where the fare amount was less than 0. Hence, I didn't have to analyse RateCodeID for such columns.

Also when checked for the presence of the negative values in columns, the columns extra, mta_tax, improvement_surcharge, total_amount, congestion_surcharge and airport_fee were found to have negative values.

Negative values in monetary-related columns likely represent refunds, adjustments, or invalid trips. Since such values are not meaningful for fare and revenue analysis, rows containing negative monetary values were removed instead of being corrected.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

"The proportion of missing values was computed for all columns. No missing values were observed in the sampled and cleaned dataset, hence no imputation or removal was required at this stage."

```
# Find the proportion of missing values in each column
```

```
#Calculate proportion of missing values in each column  
missing_proportion = df_new.isnull().mean()
```

```
#Display columns with missing values  
missing_proportion[missing_proportion > 0].sort_values(ascending =  
False)
```

```
Output: - Series([], dtype: float64)
```

2.2.2. Handling missing values in passenger_count

I found no null values in passenger_count column. Hence, there wasn't a need to handle these

2.2.3. Handle missing values in RatecodeID

No missing values in RatecodeID too. Hence, didn't handle these too

2.2.4. Impute NaN in congestion_surcharge

No missing values in congestion_surcharge too since my dataframe didn't have any missing values or NaN in any columns

2.3. Handling Outliers and Standardising Values

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

Outlier analysis was performed on key numerical and categorical columns, including payment type, trip distance, and tip amount, using domain-specific conditions defined in the notebook. These checks were designed to identify logically inconsistent or invalid trip records, such as undefined payment types or unrealistic distance-fare relationships.

Categorical code columns such as passenger_count, RatecodeID, payment_type, and VendorID were standardised to appropriate integer formats to ensure consistency and correct interpretation during analysis. Monetary and distance-related columns were not standardised, as their absolute values are essential for understanding fare distributions, revenue patterns, and trip characteristics during exploratory data analysis.

No additional records satisfied the defined outlier conditions at this stage, indicating that prior data cleaning steps had already addressed most erroneous entries.

3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

3.1.1. Classify variables into categorical and numerical

VendorID - Categorical - Code identifying the vendor

RatecodeID - Categorical - Fare rule categories such as standard, JFK, etc.

PULocationID - Categorical - Zone identifiers, no numeric meaning

DOLocationID - Categorical - Zone identifiers

payment_type - Categorical - Payment method used

tpep_pickup_datetime - Numerical(Temporal) - Time variable used for trend analysis

tpep_dropoff_datetime - Numerical(Temporal)

passenger_count - Numerical(Discrete) - Number of passengers

trip_distance - Numerical(Continuous) - Distance traveled(measured in miles)

pickup_hour - Numerical(Discrete) - Hour of the day

trip_duration - Numerical(Continuous) - Derived from pickup and dropoff times

The following monetary parameters are all Numerical(Continuous):

fare_amount

extra

mta_tax

tip_amount

tolls_amount

improvement_surcharge

congestion_surcharge

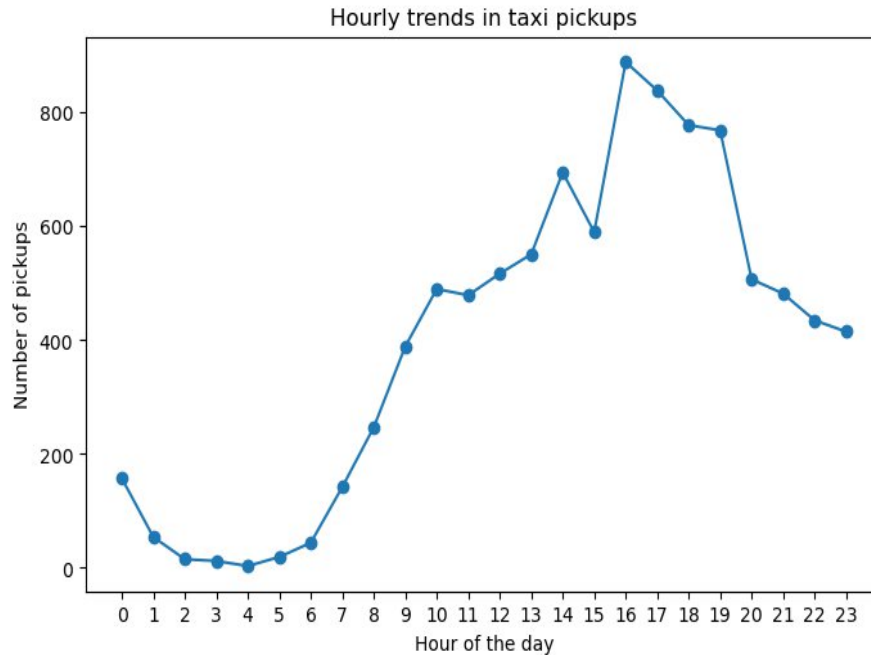
airport_fee

total_amount

Variables were categorised based on whether their values represent measurable quantities or categorical codes. Monetary variables were classified as numerical, as their magnitudes are meaningful for revenue and fare analysis, even when some values are fixed or repeated.

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

Taxi pickup distribution by hours

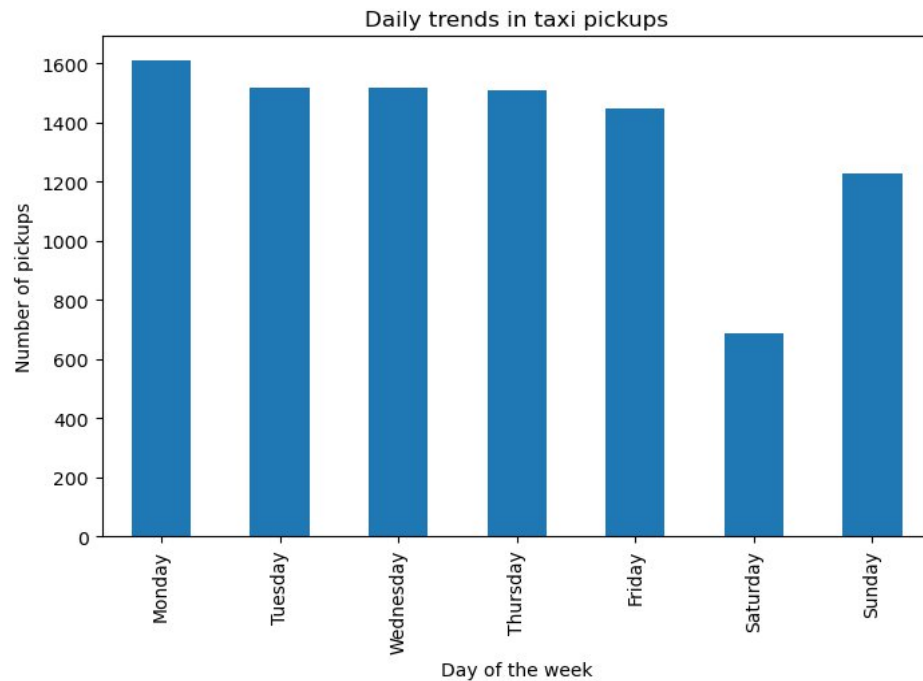


The hourly distribution of taxi pickups shows a clear and structured daily demand pattern. Pickup volumes are lowest during the late-night and early-morning hours (approximately 2 AM to 5 AM), indicating minimal travel activity during this period. Demand begins to rise steadily from early morning, with a noticeable increase after 7 AM, corresponding to the start of daily commuting hours.

Pickup activity continues to grow through the late morning and early afternoon, followed by a pronounced peak in the late afternoon to early evening (around 4 PM to 6 PM). This peak likely reflects evening commute traffic and increased social or leisure-related travel. After this period, taxi demand gradually declines through the night, stabilising at lower levels by late evening.

Overall, the pattern highlights strong alignment between taxi usage and typical urban work and lifestyle schedules, suggesting that driver deployment and fleet availability should be optimised around morning and evening peak hours.

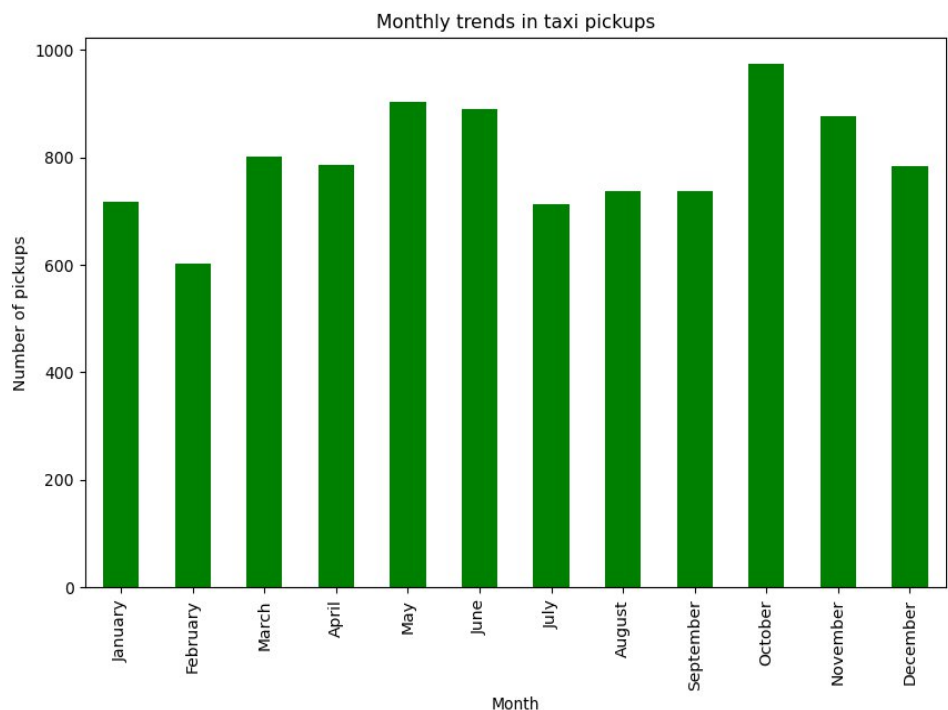
Taxi pickup distribution by days of the week



The daily distribution of taxi pickups shows a clear variation in demand across the week. Weekdays exhibit consistently high pickup volumes, with Monday recording the highest activity, indicating strong reliance on taxis for regular commuting and work-related travel. Demand remains relatively stable from Tuesday through Friday, suggesting predictable weekday travel patterns.

A noticeable decline in pickups is observed on Saturday, reflecting reduced commuting and lower overall travel demand. Pickup volumes increase again on Sunday, likely driven by leisure activities and return travel ahead of the workweek. Overall, the pattern highlights strong weekday demand contrasted with lighter but distinct weekend usage.

Monthly trends in taxi pickups



The monthly distribution of taxi pickups indicates noticeable seasonal variation in demand throughout the year. Pickup volumes are relatively moderate at the beginning of the year, with a dip observed in February. Demand increases during the spring months, peaking around May and June, which may be attributed to improved weather conditions and higher travel activity.

A decline in pickups is observed during the mid-summer months, possibly due to vacations and reduced commuting. Toward the end of the year, taxi demand rises again, with November showing the highest pickup volume, likely driven by holiday travel and increased year-end activity. Overall, the analysis highlights clear seasonal patterns in taxi usage that can inform fleet planning and resource allocation across different times of the year.

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

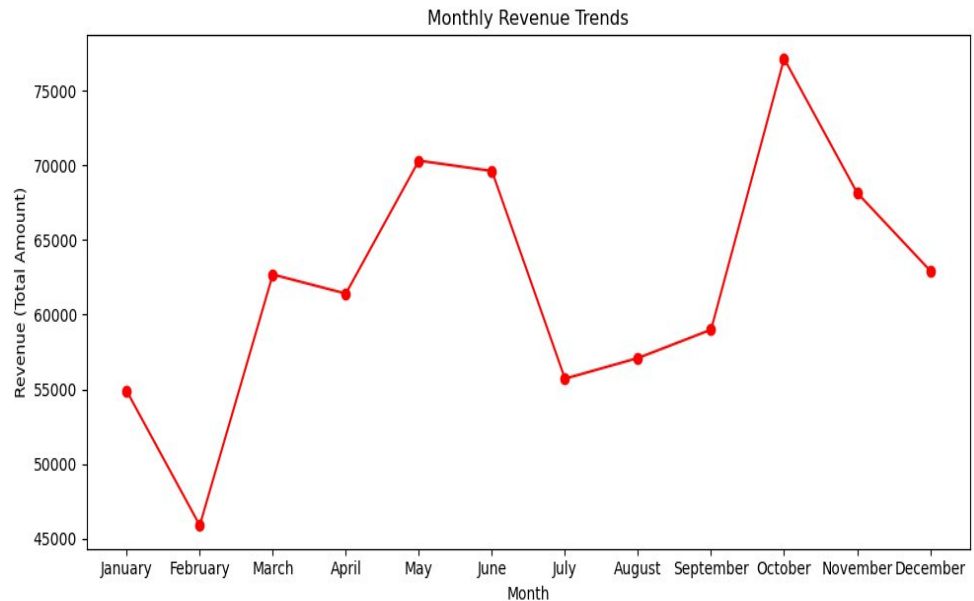
```
(fare_amount      0
tip_amount      1563
total_amount      0
```

```
trip_distance    22
dtype: int64,
fare_amount      0
tip_amount       0
total_amount     0
trip_distance    0
dtype: int64)
```

There weren't any negative value counts, but there were 1563 zero value counts in tip amount and 22 zero value counts in trip distance.

A separate dataframe was created by filtering out zero values in key financial parameters to support cleaner exploratory analysis. Zero values in tip amount were retained, as they represent valid non-tipped trips. Zero trip distances were also preserved, since trips within the same pickup and drop-off zone can legitimately record zero distance. The original dataset was left unchanged to avoid loss of valid information.

3.1.3. Analyse the monthly revenue trends



The monthly revenue trend exhibits clear seasonal variation in taxi earnings throughout the year. Revenue is relatively lower during the early months, with a noticeable dip in February, followed by a steady increase through the spring months, peaking around May and June. A decline is observed during the mid-year period, particularly in July, which may be attributed to reduced commuting and vacation-related travel.

Towards the latter part of the year, revenue rises sharply, reaching its highest level in October, likely driven by increased travel demand, events, and business activity. Although revenue declines slightly in November and December, it remains higher than mid-year levels. Overall, the trend highlights the strong influence of seasonality on taxi revenue, offering opportunities for targeted operational and pricing strategies during peak months.

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

pickup_quarter

2023Q1 0.219487

2023Q2 0.270293

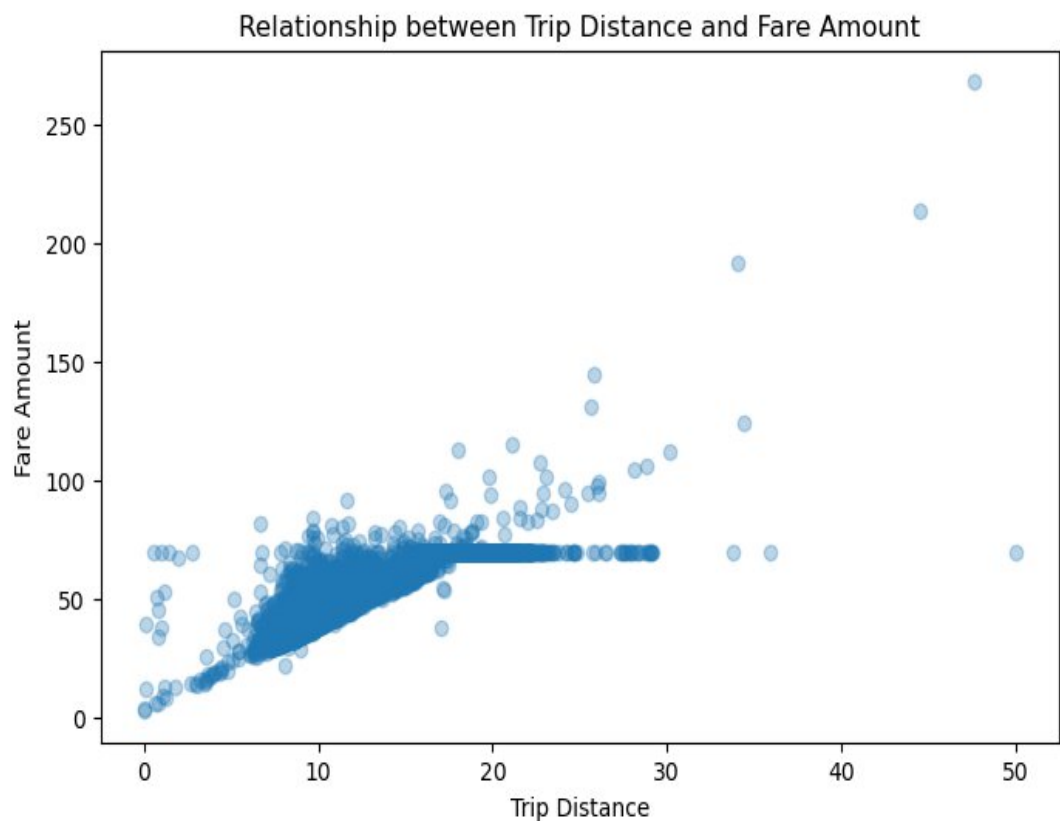
2023Q3 0.230670

2023Q4 0.279550

Freq: Q-DEC, Name: total_amount, dtype: float64

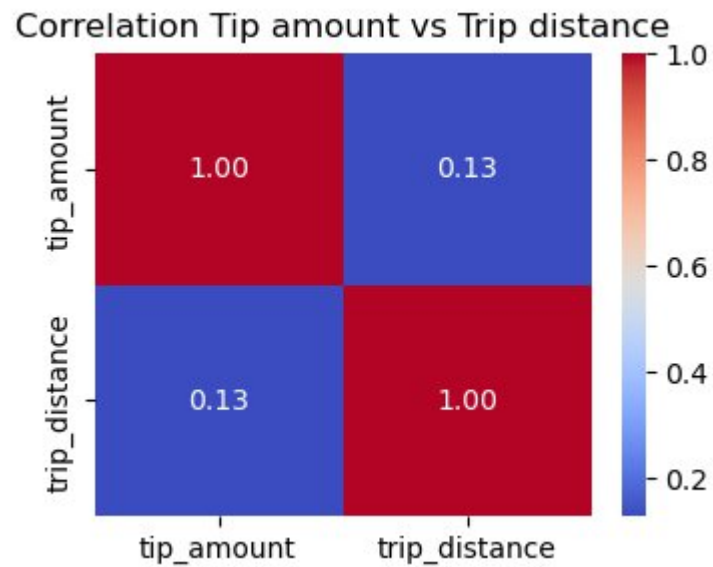
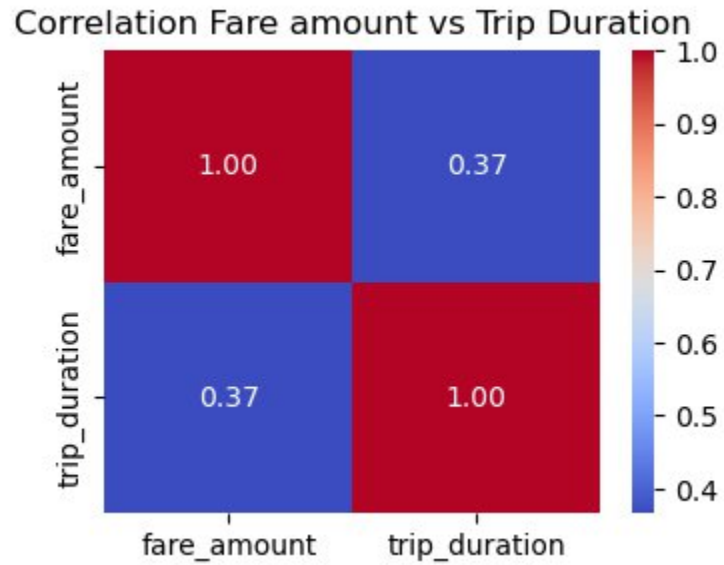
Quarterly revenue analysis shows an uneven distribution of earnings across the year. The first quarter contributes the smallest share of total revenue, while revenue increases through the second quarter. The fourth quarter accounts for the largest proportion of annual revenue, reflecting heightened travel demand toward the end of the year.

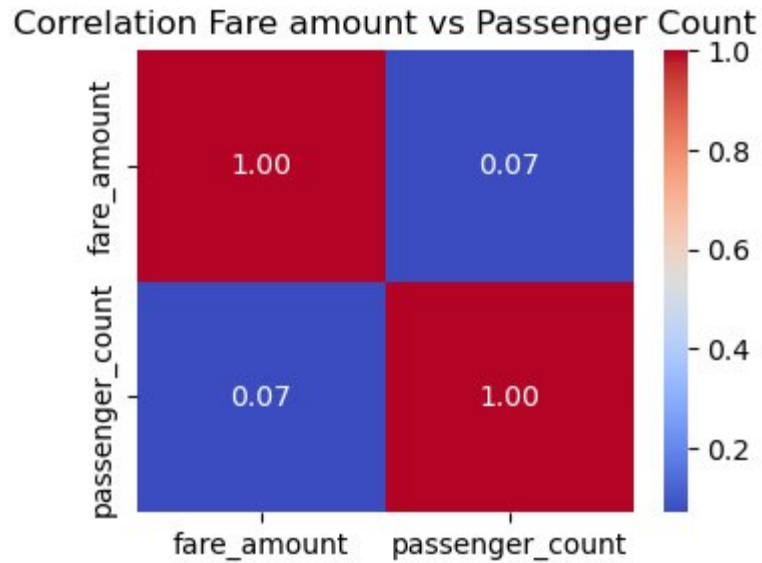
3.1.5. Analyse and visualise the relationship between distance and fare amount



The scatter plot illustrates a strong positive relationship between trip distance and fare amount, indicating that fares generally increase as the distance travelled increases. Most trips are clustered within shorter to medium distances, with fares rising proportionally in this range. A small number of long-distance trips with higher fares are visible, appearing as outliers, but they do not alter the overall positive trend. This relationship confirms that taxi fare calculations are largely driven by distance travelled. Correlation value = 0.893.

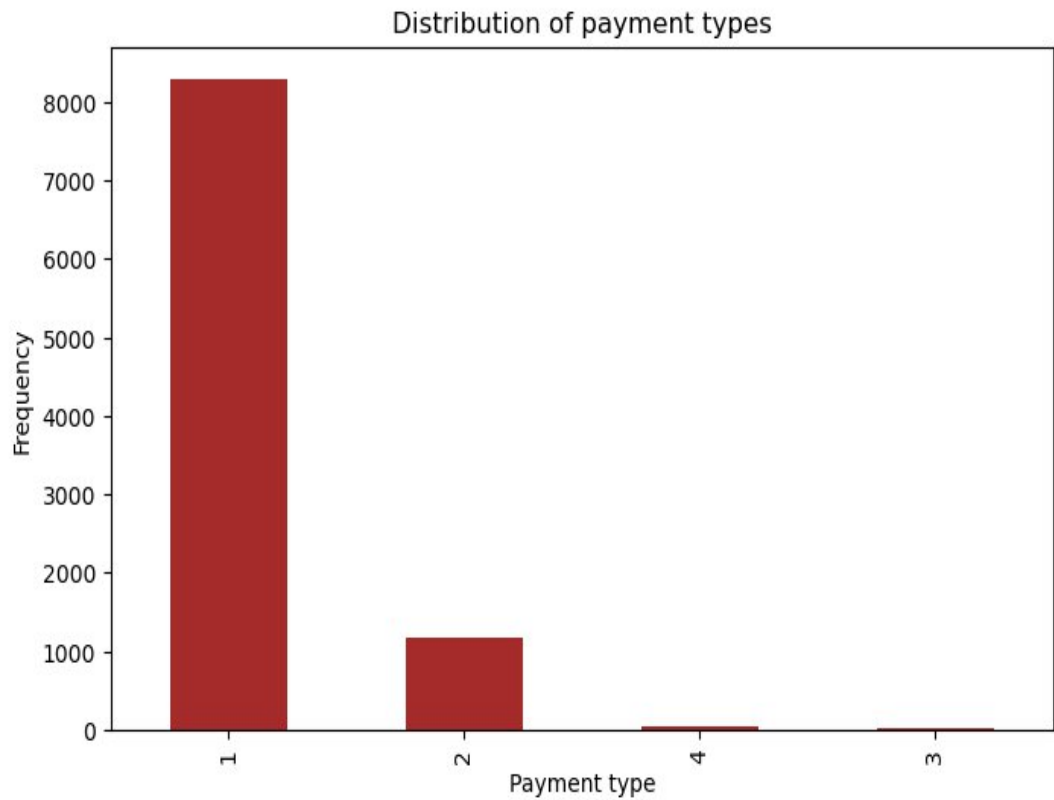
3.1.6. Analyse the relationship between fare/tips and trips/passengers





The correlation analysis reveals a **moderate positive relationship** between fare amount and trip duration (correlation ≈ 0.37), indicating that longer trips in terms of time generally result in higher fares due to time-based fare components. In contrast, the correlation between tip amount and trip distance is **weakly positive** (correlation ≈ 0.13), suggesting that while longer trips may lead to slightly higher tips, tipping behaviour varies significantly among passengers. Lastly, fare amount shows **very little correlation** with passenger count (correlation ≈ 0.07), confirming that taxi fares are primarily influenced by distance and time rather than the number of passengers.

3.1.7. Analyse the distribution of different payment types



payment_type

1 8284

2 1178

4 32

3 24

Name: count, dtype: int64

The distribution of payment types shows a strong preference for **credit card payments (payment_type = 1)**, which account for the vast majority of taxi trips. **Cash payments (payment_type = 2)** form a significantly smaller but still notable proportion of transactions. Other payment types, such as disputes and no-charge trips (payment types 3 and 4), occur very infrequently, indicating that most trips follow standard payment processes.

This distribution highlights the dominance of digital payment methods in NYC taxi operations, which has implications for faster transactions, better tip tracking, and streamlined revenue management.

3.1.8. Load the taxi zones shapefile and display it

The taxi zone shapefile provides geographical boundaries and metadata for each taxi zone in New York City. The LocationID field in the shapefile corresponds directly to the pickup and drop-off location IDs in the trip records, enabling spatial enrichment of trip data. This linkage allows trip activity to be analysed and visualised at the zone and borough levels.

```
<class 'geopandas.geodataframe.GeoDataFrame'>
```

```
RangeIndex: 263 entries, 0 to 262
```

```
Data columns (total 7 columns):
```

```
#   Column      Non-Null Count  Dtype
```

```
---  -----  -
```

```
0   OBJECTID    263 non-null   int32
```

```
1   Shape_Leng  263 non-null   float64
```

```
2   Shape_Area  263 non-null   float64
```

```
3   zone        263 non-null   object
```

```
4   LocationID  263 non-null   int32
```

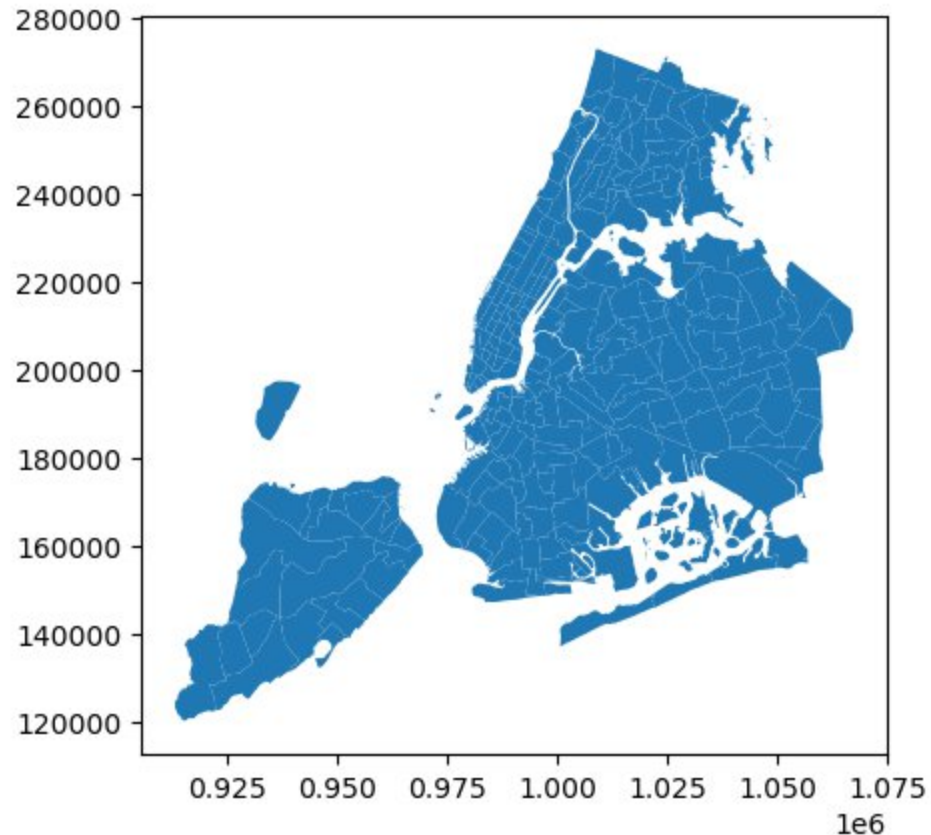
```
5   borough     263 non-null   object
```

```
6   geometry    263 non-null   geometry
```

```
dtypes: float64(2), geometry(1), int32(2), object(2)
```

```
memory usage: 12.5+ KB
```

```
None
```



3.1.9. Merge the zone data with trips data

The trip dataset was merged with the taxi zones shapefile using the pickup location identifier (PULocationID) and the corresponding zone identifier (LocationID). This merge enriches each trip record with pickup zone and borough information, enabling spatial analysis of taxi demand across New York City.

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

The trip data was grouped by pickup location ID to compute the total number of trips originating from each taxi zone. Since each row represents a single trip, the trip count per zone was calculated using row counts rather than summing numerical fields.

Grouping the trip data by pickup location ID reveals a highly uneven spatial distribution of taxi demand. A small number of zones account for a disproportionately large share of trips, with LocationIDs such as **138**, **132**, and **70** recording the highest pickup volumes. In contrast, many locations show very low trip counts, often with only one or two trips, indicating limited taxi activity in those zones within the sampled data.

This concentration of pickups suggests that taxi demand is strongly clustered in specific high-activity areas, highlighting key zones where driver availability and operational focus could be prioritised.

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

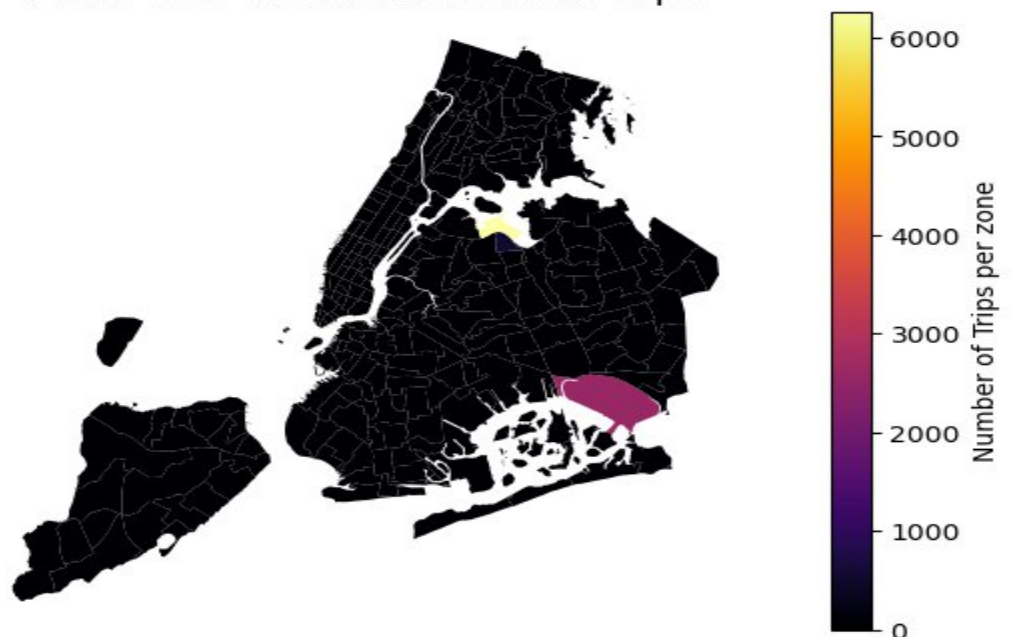
The aggregated trip counts per pickup location were merged back into the taxi zones GeoDataFrame using the LocationID field. This enrichment enables spatial visualisation of taxi demand across zones, including areas with no recorded trips.

Zones with no trips were having NaN values.
For plotting, these were set to 0.

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

The choropleth map highlights the spatial distribution of taxi pickups across New York City. A small number of zones exhibit significantly higher trip volumes, indicating concentrated demand in specific areas, while many peripheral zones show relatively low activity. This spatial clustering of trips suggests that taxi demand is unevenly distributed across the city and is driven by key commercial and transit hubs. Sorting taxi zones by pickup volume reveals that a small number of zones contribute disproportionately to total trips, reinforcing the spatial concentration of taxi demand observed in the choropleth map.

Zone wise distribution of taxi trips



3.1.13. Conclude with results

The exploratory analysis of New York City yellow taxi trip records reveals clear temporal, financial, and spatial patterns that can inform operational and strategic decision-making.

Temporal analysis shows that taxi demand follows a strong daily and weekly rhythm. Pickups are lowest during early morning hours and rise steadily through the day, peaking in the late afternoon and early evening, reflecting work-hour commutes and evening travel. Weekdays consistently record higher trip volumes than weekends, with Saturday showing the lowest demand. Monthly trends indicate seasonal variation, with higher activity during late spring and autumn months, while early-year months exhibit comparatively lower demand.

From a financial perspective, total revenue mirrors trip volume trends, increasing during high-demand months and quarters. Quarterly revenue analysis shows that the final quarter of the year contributes the largest share of annual revenue, highlighting the impact of increased travel and commercial activity toward year-end. Fare amount exhibits a strong positive relationship with trip distance and a moderate positive relationship with trip duration, confirming that taxi fares are primarily driven by distance and time. In contrast, passenger count shows negligible correlation with fare amount, indicating that fares are not influenced by the number of passengers. Tip amount displays a weak positive relationship with trip distance, suggesting that longer trips may lead to slightly higher tips, though tipping behaviour varies widely among passengers.

The geographical analysis highlights a highly uneven spatial distribution of taxi demand across the city. A small number of taxi zones account for a disproportionately large share of pickups, with high-activity zones concentrated in central and commercially significant areas. Many peripheral zones exhibit minimal taxi activity, indicating localized demand clusters. The choropleth map reinforces this spatial concentration, clearly identifying zones where taxi operations are most intensive.

Overall, the analysis demonstrates that NYC taxi demand is shaped by predictable temporal cycles, distance- and time-based fare structures, and strong spatial clustering. These insights can help optimise driver deployment, improve revenue forecasting, and enhance service availability in high-demand zones and peak time periods.

3.2. Detailed EDA: Insights and Strategies

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

After filtering route–hour combinations to include only those with sufficient trip volumes, the analysis reveals consistent congestion patterns during morning peak hours (7–9 AM). Several routes originating from the same high-activity pickup zone (notably LocationID 138) repeatedly appear among the slowest routes across consecutive hours. Average speeds on these routes drop significantly during peak periods, with some routes recording speeds as low as 7–14 mph, indicating severe congestion.

The persistence of low speeds across multiple hours and routes suggests systematic bottlenecks rather than isolated delays. These slow routes are likely influenced by high traffic density and commuter demand during peak hours, highlighting specific zones and time windows where operational inefficiencies are most pronounced.

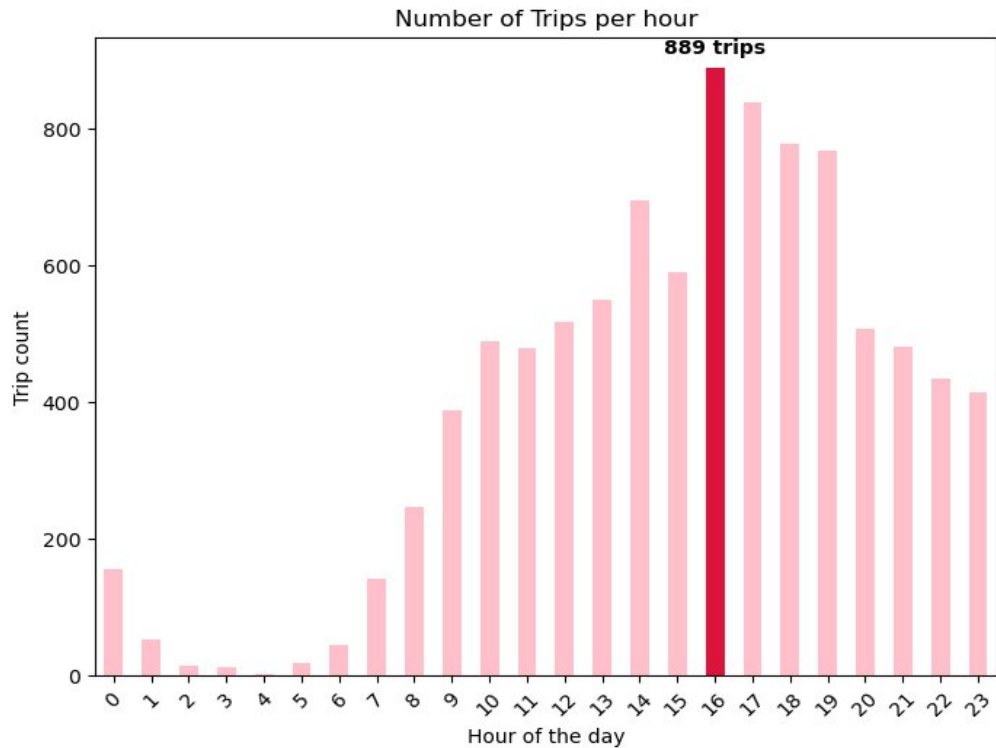
Congestion is route-specific and time-dependent, with peak-hour traffic significantly reducing travel speeds in high-demand urban zones.

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

The hourly analysis of taxi trips reveals a clear demand pattern throughout the day. Trip volumes are lowest during the early morning hours (between 1 AM and 5 AM), indicating minimal travel activity at night. Demand begins to rise sharply from around 7 AM, corresponding to the morning commute, and continues to increase through the afternoon.

The busiest hour of the day occurs around 4 PM–5 PM (hour 16), where the highest number of trips is recorded. This peak reflects evening commute traffic and increased travel demand during late working hours. After this peak, trip counts gradually decline into the late evening.

This pattern highlights critical time windows where taxi availability and fleet allocation should be optimised to efficiently meet passenger demand.

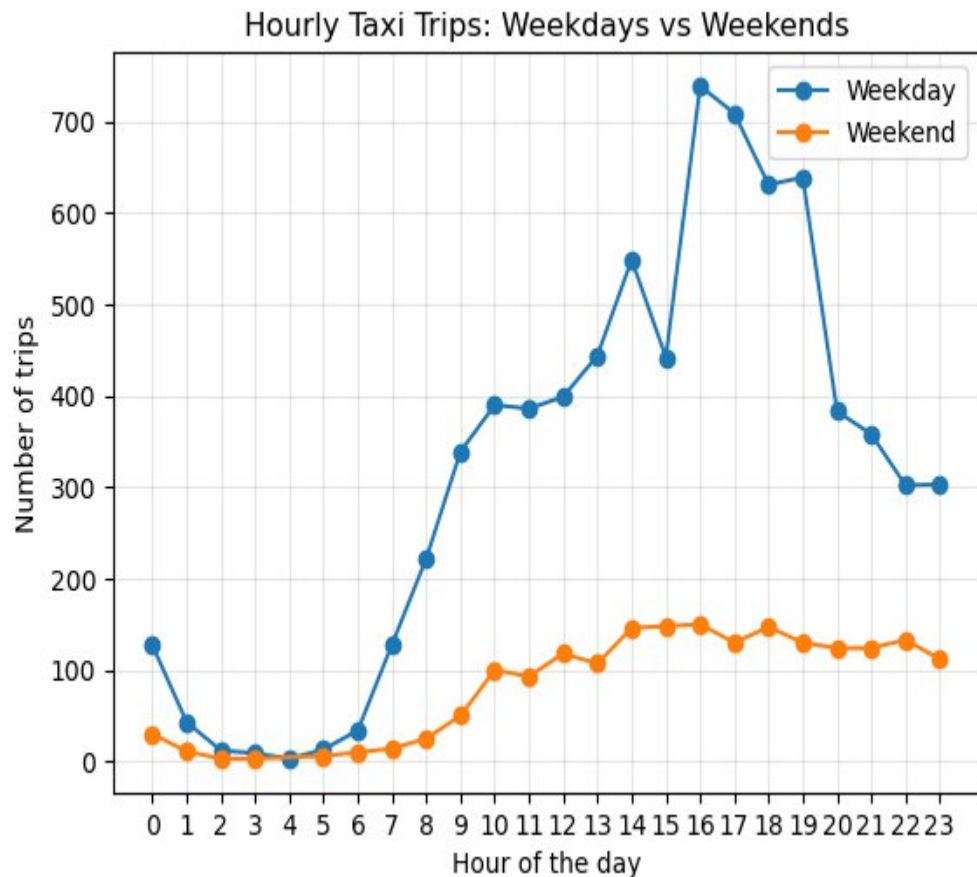


3.2.3. Scale up the number of trips from above to find the actual number of trips

Based on the sampled data, the five busiest hours of the day occur between **2 PM and 7 PM**, with **hour 16 (4–5 PM)** recording the highest number of trips. After scaling the sampled counts using the sampling fraction, the estimated actual number of trips during these peak hours ranges from approximately **13,900 to 17,780 trips per hour**.

The peak hour (4–5 PM) alone accounts for an estimated **17,780 trips**, reinforcing the presence of strong evening commute demand. This concentration of trips during late afternoon and early evening hours highlights critical periods where taxi availability and fleet deployment should be optimized to meet passenger demand efficiently.

3.2.4. Compare hourly traffic on weekdays and weekends



The hourly comparison reveals distinct travel behavior between weekdays and weekends. On **weekdays**, taxi demand shows a pronounced bimodal pattern, with a sharp rise during the **morning commute (7–9 AM)** and a stronger peak in the **evening commute (4–7 PM)**. These peaks reflect work-related travel and regular commuting patterns.

In contrast, **weekend traffic** is more evenly distributed throughout the day, with minimal activity during early morning hours and a gradual increase from late morning onwards. Weekend demand peaks during the **afternoon and early evening**, remaining relatively stable into late evening hours, likely driven by leisure, shopping, and social activities.

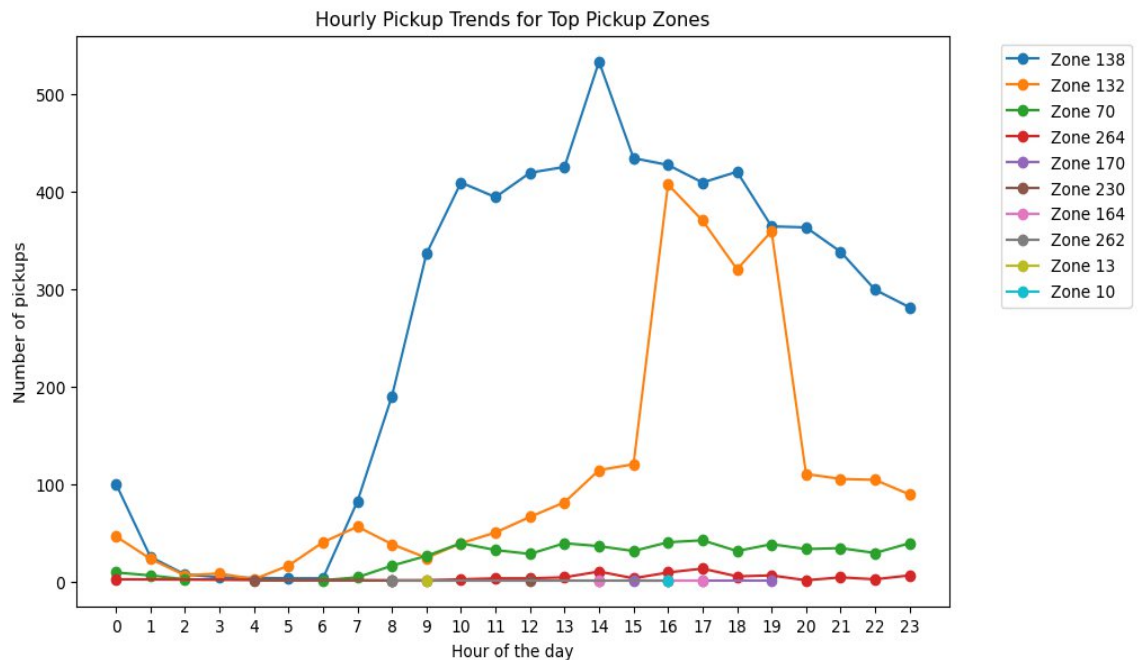
Identifying these busy and quiet hours for weekdays and weekends enables more effective fleet allocation, improved driver scheduling, and reduced idle time during low-demand periods. Such insights can support operational planning and help optimize service availability during peak demand windows.

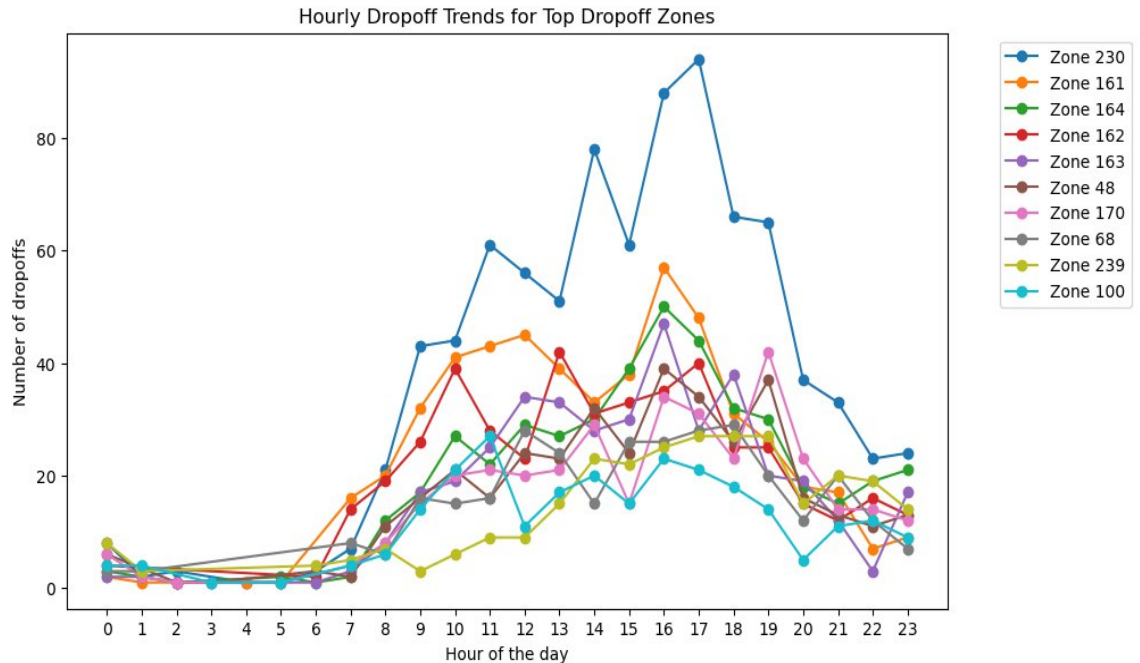
3.2.5. Identify the top 10 zones with high hourly pickups and drops

The analysis of the top 10 pickup zones shows strong time-dependent demand patterns, with several zones exhibiting pronounced peaks during afternoon and evening hours. In particular, Zone 138 consistently records the highest pickup volumes throughout the day, with a sharp increase from morning hours and a peak in the late afternoon, indicating its role as a major demand hub. Other high-pickup zones, such as Zone 132 and Zone 70, also show elevated activity during commute and post-work hours, though with lower overall volumes.

In contrast, the top 10 dropoff zones display more distributed hourly patterns, with dropoffs peaking slightly later in the day compared to pickups. Zones such as Zone 230 and Zone 161 show higher dropoff activity during afternoon and early evening hours, suggesting these areas act as key destinations for work or leisure trips. Several zones appear in both the pickup and dropoff top-10 lists, highlighting their importance as major mobility hubs within the city.

Overall, the comparison reveals that pickup activity tends to peak earlier than dropoff activity, reflecting commuting flows from residential or mixed-use zones to commercial and activity-dense areas. Understanding these zone-specific hourly trends can help improve driver positioning, reduce empty cruising time, and enhance service availability in high-demand zones during peak hours.





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

The pickup-to-dropoff ratio was calculated for each taxi zone to understand directional imbalances in trip demand. Zones with the highest ratios (e.g., Zones 70, 138, and 132) exhibit significantly more pickups than dropoffs, indicating that these areas primarily act as trip origin zones, such as residential neighbourhoods or major entry points. In contrast, zones with the lowest ratios (e.g., Zones 162, 163, and 230) experience substantially more dropoffs than pickups, suggesting that they function mainly as destination zones, likely associated with commercial or employment centres.

This imbalance highlights opportunities for improving operational efficiency by repositioning drivers from dropoff-heavy zones to nearby pickup-heavy zones, thereby reducing idle time and increasing driver utilisation.

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

Night-hour analysis (11 PM to 5 AM) reveals a distinct set of high-activity zones compared to daytime patterns. Zones such as 138, 132, and 70 emerge as the most active pickup locations during night hours, indicating areas with sustained late-night travel demand. In contrast, dropoff activity during the same period is concentrated in zones such as 230, 164, and 262, suggesting these areas function as key night-time destinations.

The partial overlap between high pickup and dropoff zones highlights the presence of certain zones with continuous activity across the day, while other zones show strong directional demand during night hours.

Identifying these night-specific demand patterns can help optimize overnight driver allocation, reduce idle cruising, and improve service availability during late-night operations.

3.2.8. Find the revenue share for nighttime and daytime hours

Night revenue: 47456.61

Total revenue: 744880.6699999999

Day revenue: 697424.0599999999

Night Revenue Share: 6.37%

Day Revenue Share: 93.63%

The revenue analysis reveals a significant imbalance between daytime and nighttime taxi operations. Nighttime hours (11 PM to 5 AM) contribute approximately 6.37% of the total revenue, while daytime hours account for the remaining 93.63%. This indicates that although taxis continue to operate overnight, most of the revenue is generated during daytime and evening hours.

The relatively low revenue share during nighttime reflects reduced trip volumes despite potentially longer trip durations or higher per-trip fares. This insight highlights that nighttime operations are demand-constrained and may require targeted pricing or incentive strategies to improve profitability and driver participation during late-night hours.

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

The average fare per mile per passenger decreases steadily as passenger count increases, indicating effective cost sharing among passengers. Single-passenger trips incur the highest per-passenger cost, while multi-passenger trips distribute fares more efficiently.

passenger_count

1 6.457052

2 3.071608

3 1.996512

4 1.539231

5 1.315774

6 1.082721

dtype: float64

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

Hourly Patterns

The average fare per mile shows a clear time-of-day variation. Early morning hours (around 3 AM–6 AM) have the lowest fare per mile, likely due to lighter traffic and smoother travel conditions. From 8 AM to 12 PM, the fare per mile increases sharply, peaking around 9–10 AM, which aligns with morning commute congestion.

During the afternoon and early evening, fares per mile remain relatively high, before gradually declining at night. This indicates that congestion and travel delays during peak hours significantly increase fare efficiency per mile.

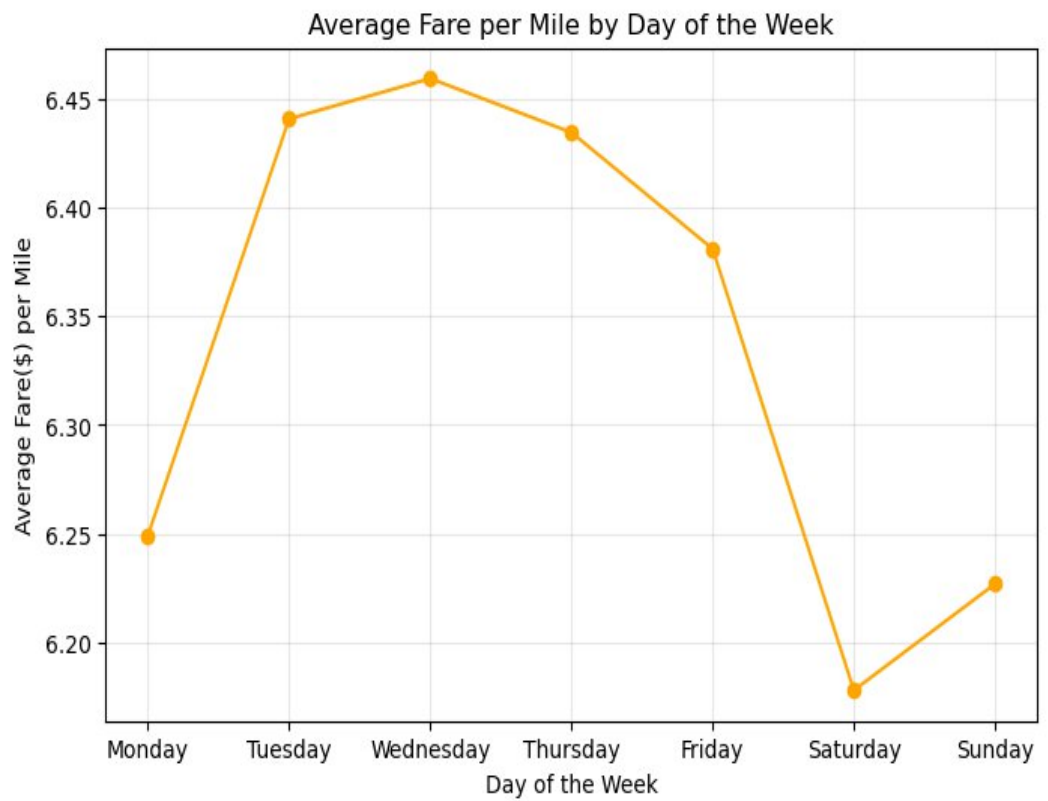
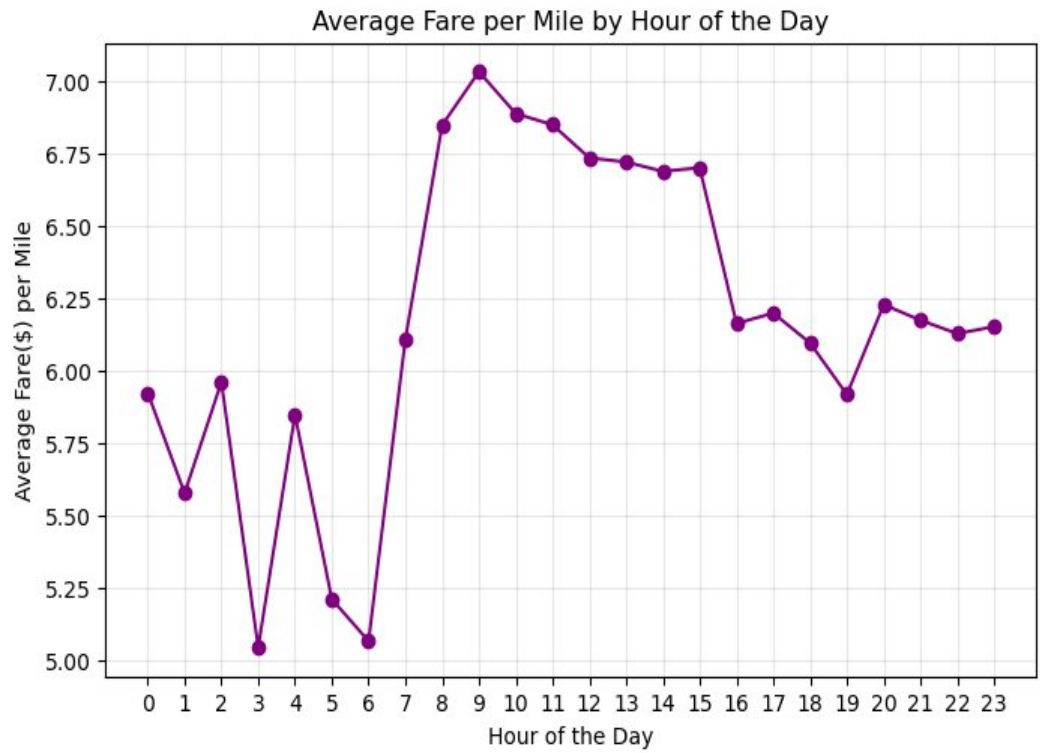
Day-of-Week Patterns

Across days of the week, mid-weekdays (Tuesday to Thursday) exhibit the highest average fare per mile, with Wednesday showing the maximum. This suggests consistent commuter demand and congestion during working days.

In contrast, weekends (Saturday and Sunday) show slightly lower fare per mile values, reflecting smoother traffic flow and more leisure-oriented trips with fewer delays.

Key Takeaway for Pricing Strategy

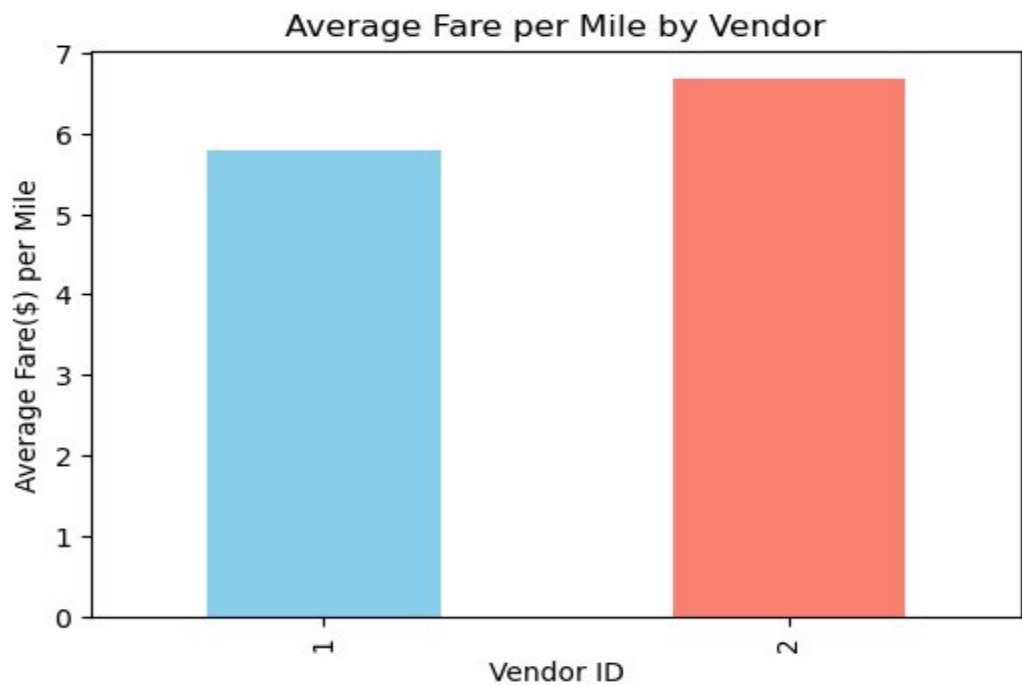
These patterns suggest that time-based pricing adjustments can be effective. Peak commuting hours and mid-weekdays naturally generate higher fare efficiency, while early mornings and weekends may require discounts or demand-stimulating strategies. Dynamic pricing aligned with congestion and demand can help optimize both revenue and service availability.

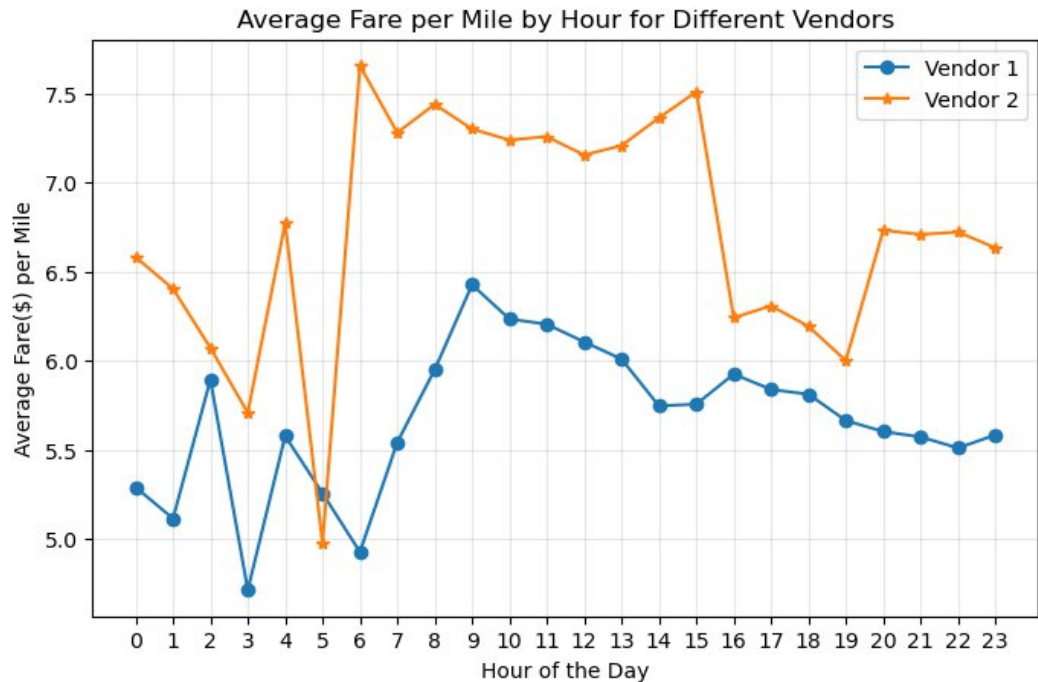


3.2.11. Analyse the average fare per mile for the different vendors

The vendor-level analysis shows that Vendor 2 consistently records a higher average fare per mile than Vendor 1, both overall and across most hours of the day. Hourly trends indicate that the difference between vendors is most pronounced during daytime and peak travel hours, suggesting vendor-specific pricing strategies or route characteristics that lead to higher revenue efficiency for Vendor 2.

During late-night and early-morning hours, fare per mile values for both vendors decline and converge slightly, indicating more uniform pricing under lower demand conditions. Overall, Vendor 1 exhibits relatively stable fare efficiency across the day, while Vendor 2 demonstrates higher variability and stronger monetization during peak periods.





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

This analysis compared the average fare per mile charged by different vendors across three distance tiers: short trips (≤ 2 miles), medium trips (2–5 miles), and long trips (> 5 miles). To ensure meaningful and stable comparisons, trips with trip distances less than 0.5 miles were excluded from the analysis.

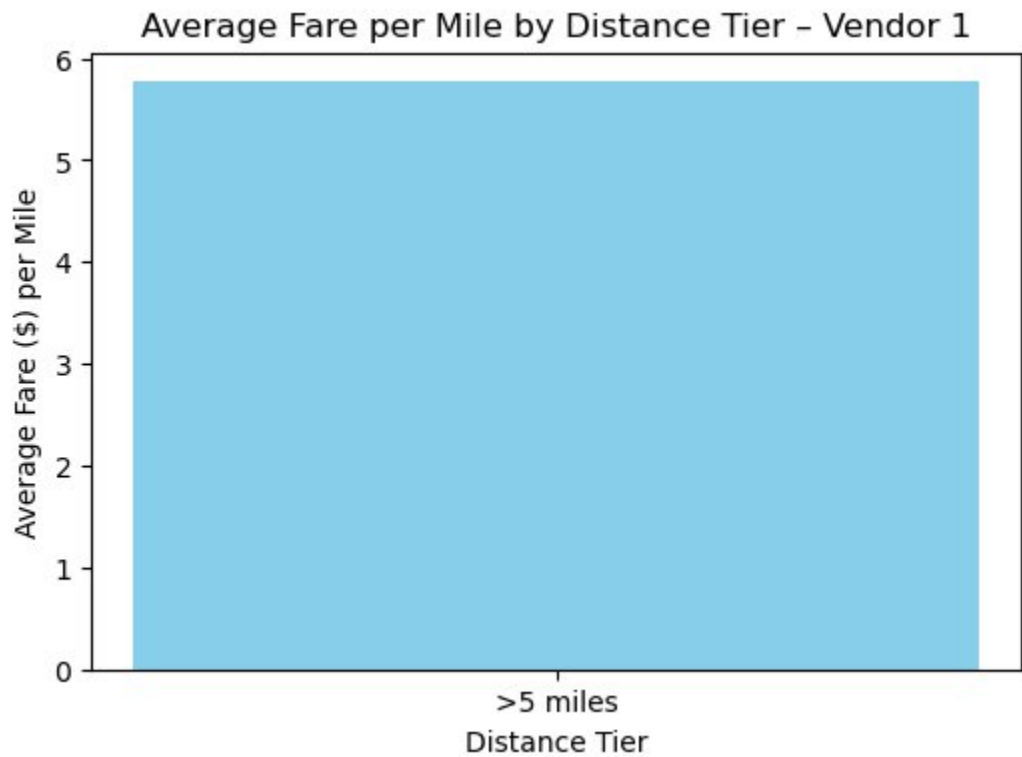
Very short trips are heavily influenced by minimum fare policies, flag-drop charges, and fixed surcharges, which significantly inflate fare-per-mile values due to near-zero travel distances. Including such trips leads to distorted and economically unrealistic results, particularly in the shortest distance tier. Applying a minimum distance threshold of 0.5 miles helps mitigate this distortion and improves the reliability of fare-per-mile estimates.

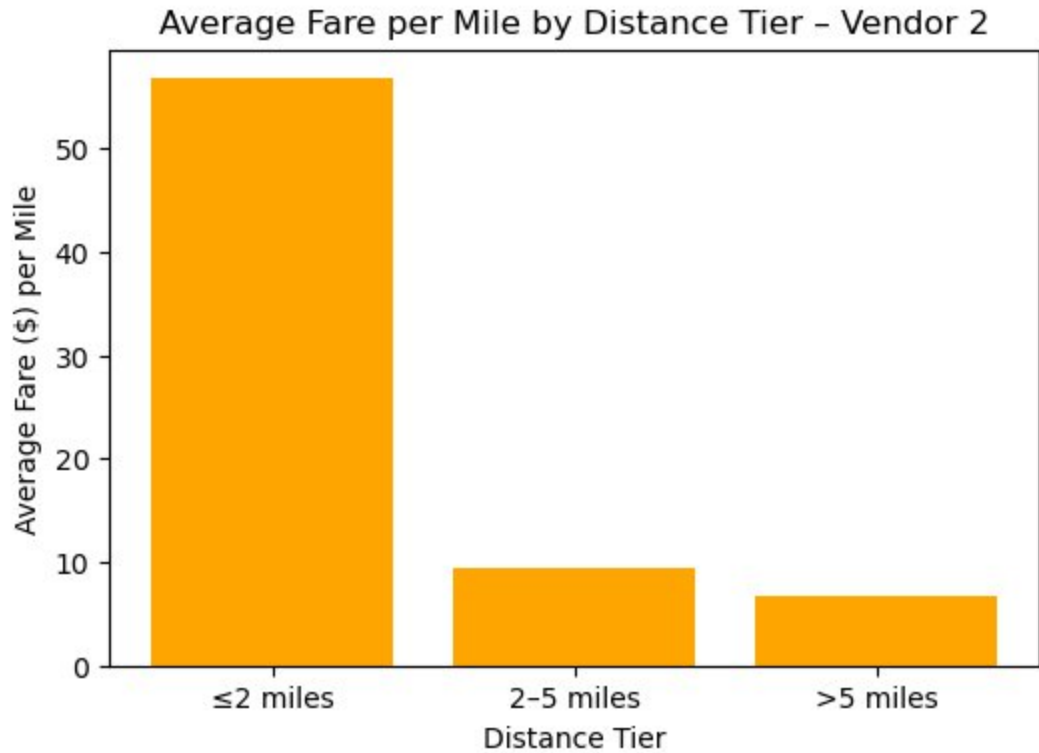
Even after filtering, the ≤ 2 mile tier continued to show inflated fare-per-mile values, indicating that fare-per-mile is not a stable metric for very short trips. As trip distance increases, fare-per-mile values stabilise and become more suitable for vendor comparison. The > 5 mile tier provides the most reliable insights into vendor pricing behaviour.

In the long-distance tier, Vendor 2 consistently records a higher average fare per mile than Vendor 1, suggesting stronger monetisation on longer routes. This pattern aligns with earlier vendor-level analyses and

indicates potential differences in pricing strategies or route characteristics across vendors.

Overall, the tier-based analysis highlights that fare-per-mile comparisons should be interpreted with caution for short trips, while long-distance trips offer the most meaningful basis for evaluating vendor pricing efficiency.

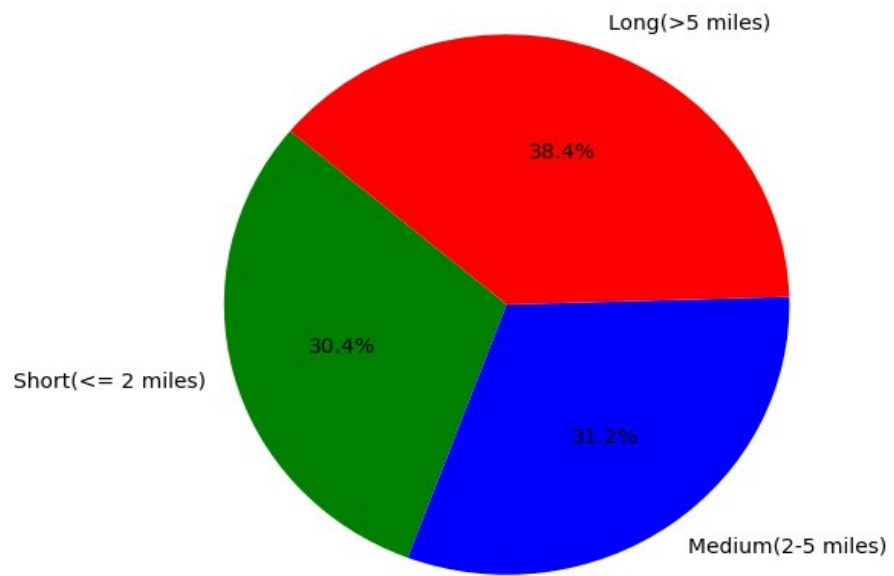




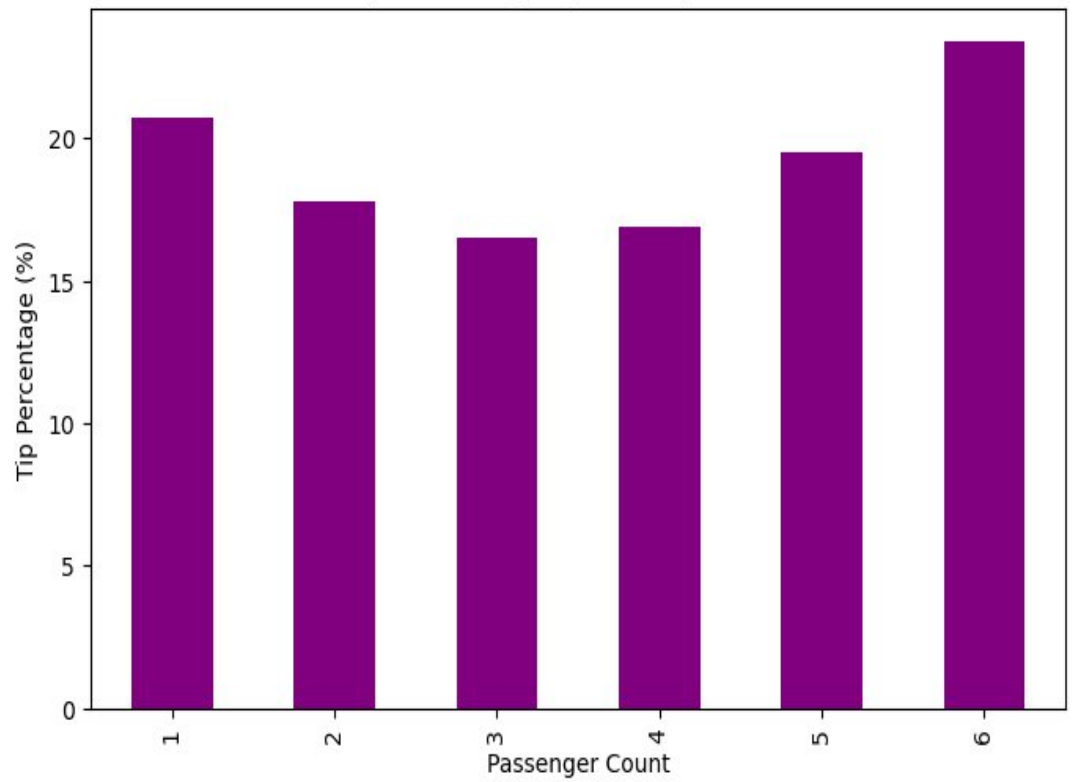
3.2.13. Analyse the tip percentages

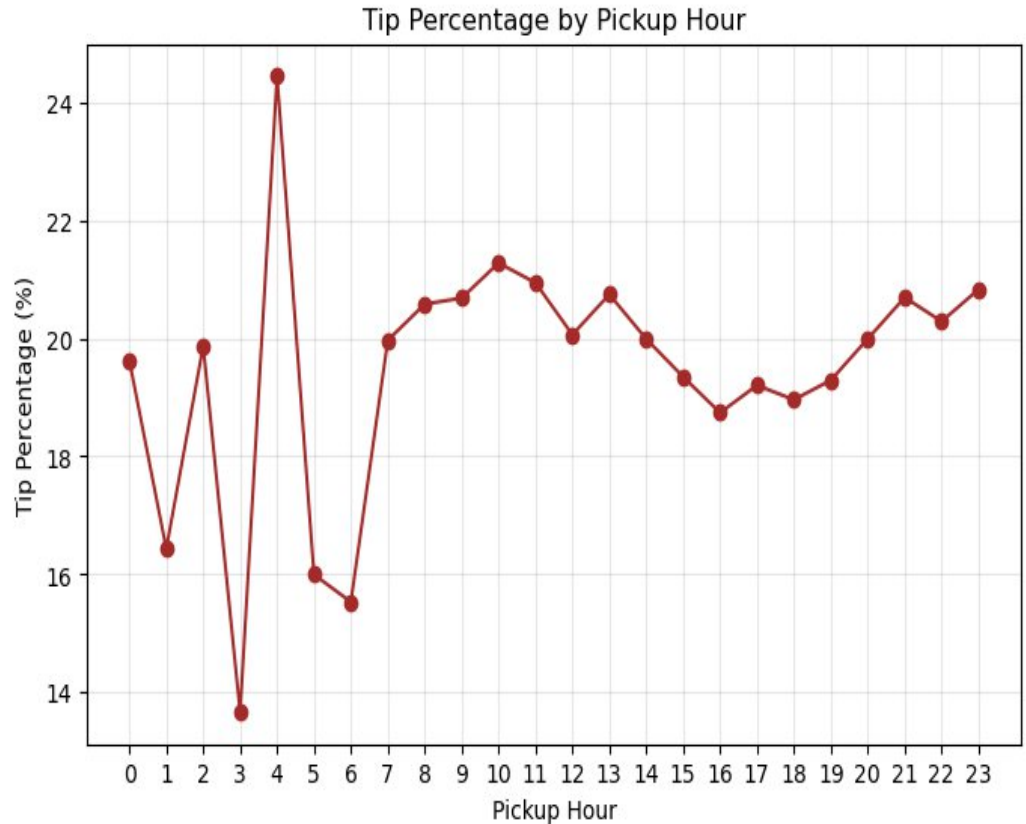
Tip percentages vary meaningfully with trip distance, passenger count, and pickup time. Longer trips attract higher tip percentages, likely due to increased interaction and perceived service value. Solo riders and larger passenger groups tend to tip more generously, while mid-sized groups show lower tipping behavior, possibly due to shared responsibility. Temporally, tipping is lowest during late-night hours (2–4 AM), likely driven by passenger fatigue or urgency, and higher during early morning and evening periods. Overall, low tip percentages are associated with short trips, late-night travel, and mid-sized passenger groups.

Tip Percentage by Trip Distance



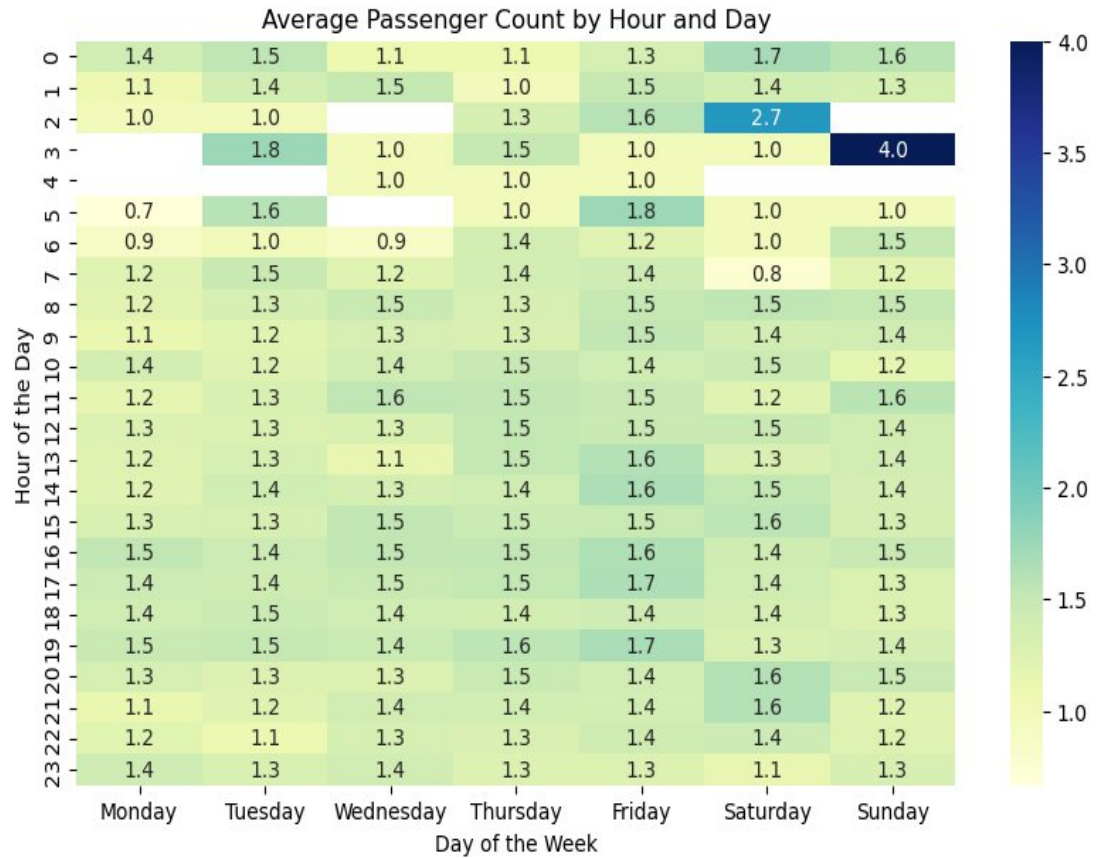
Tip Percentage by Passenger Count





3.2.14. Analyse the trends in passenger count

The variation of passenger count across hours and days was analysed using the mean passenger count per trip. The mean was chosen instead of the sum to capture typical group size rather than overall traffic volume. The analysis shows that taxi trips are predominantly solo across all times, with marginally higher passenger counts during evening hours and weekends, indicating increased group or leisure travel. Weekday mornings show the lowest average passenger counts, reflecting commuter-driven solo trips.



3.2.15. Analyse the variation of passenger counts across zones

To understand how passenger demand varies geographically, the average passenger count was analysed across different pickup zones. Instead of using total passenger counts, the mean passenger count per trip was used to avoid bias from zones with very high trip volumes and to capture the typical group size travelling from each location.

The analysis shows that most zones have an average passenger count close to 1, indicating that single-passenger trips dominate taxi usage across New York City. This is consistent with the general nature of urban taxi travel, where individual commuters and solo travellers form the majority of users.

However, a small subset of zones exhibits higher average passenger counts, suggesting different travel behaviour. These zones are likely associated with:

Airports and transit hubs,

Tourist-heavy areas,

Residential zones where group travel (families or shared rides) is more common.

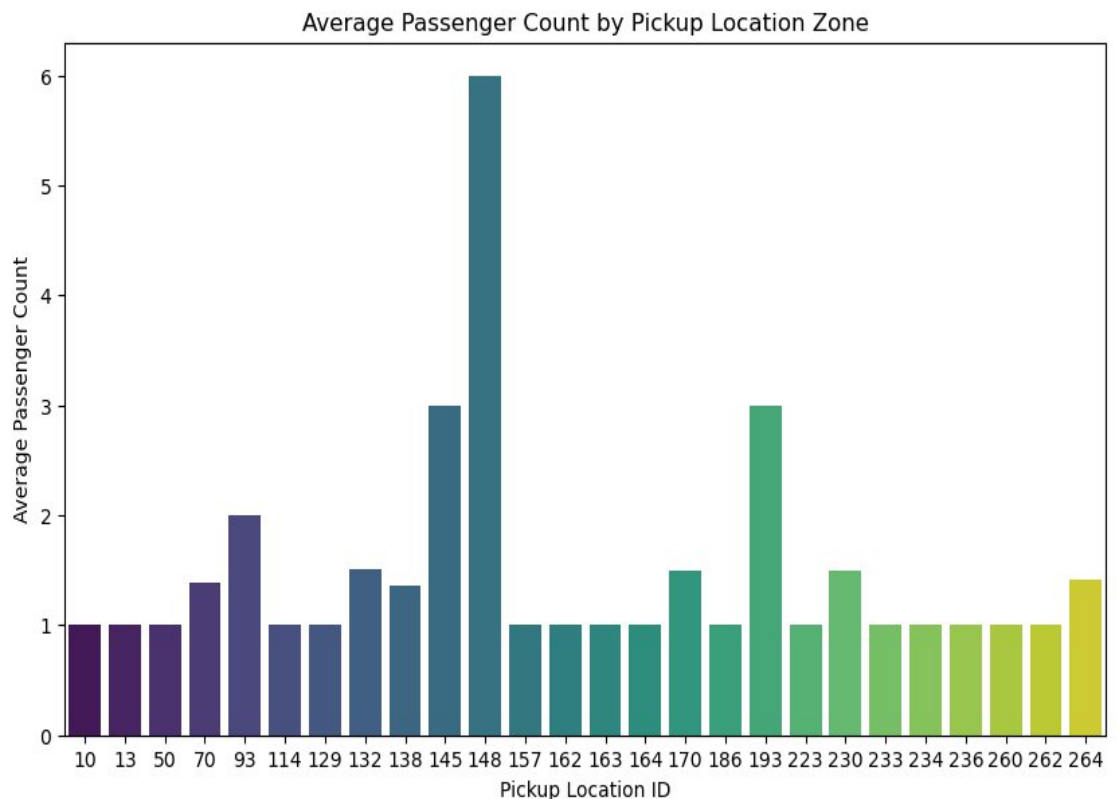
By merging the passenger count statistics with the taxi zone GeoDataFrame, zones with no sampled trips were retained and assigned zero values. This ensured spatial completeness and prevented biased exclusion of low-activity zones, especially important for geographical comparisons.

Overall, this analysis highlights that:

Passenger occupancy is relatively uniform across most zones,

Certain zones consistently show higher group travel,

These insights can be useful for fleet allocation, ride-sharing strategies, and vehicle type planning (e.g., deploying larger vehicles in zones with higher average occupancy).



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

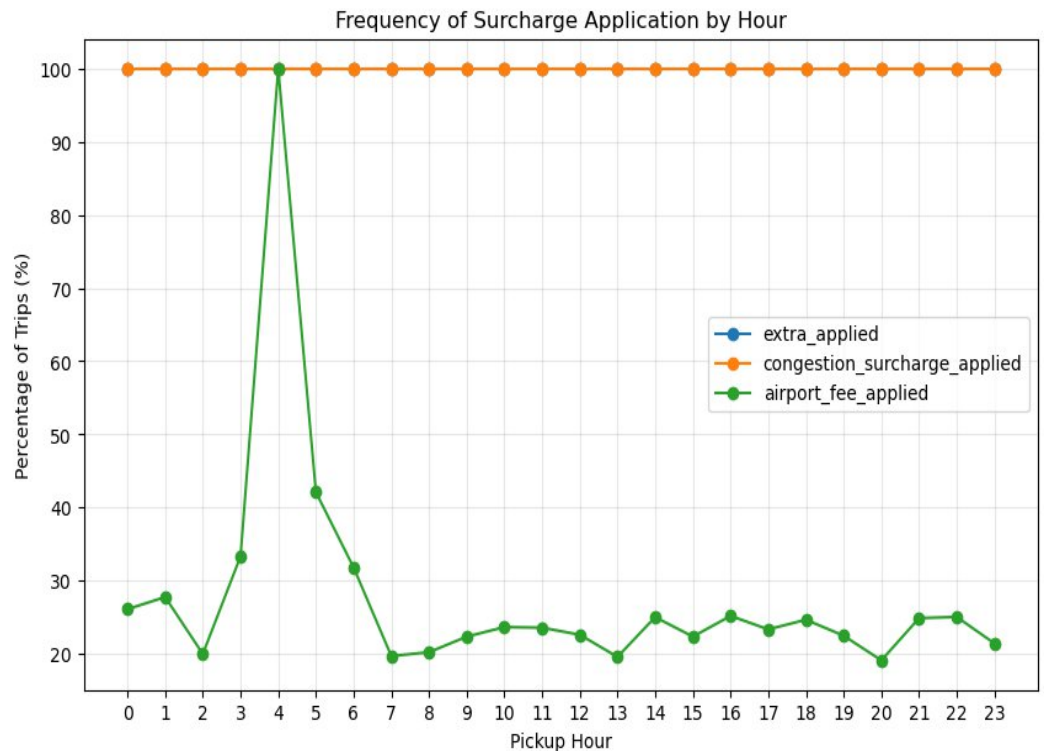
We analyzed the frequency of surcharge application across pickup hours and pickup zones to identify when and where extra charges are applied more frequently.

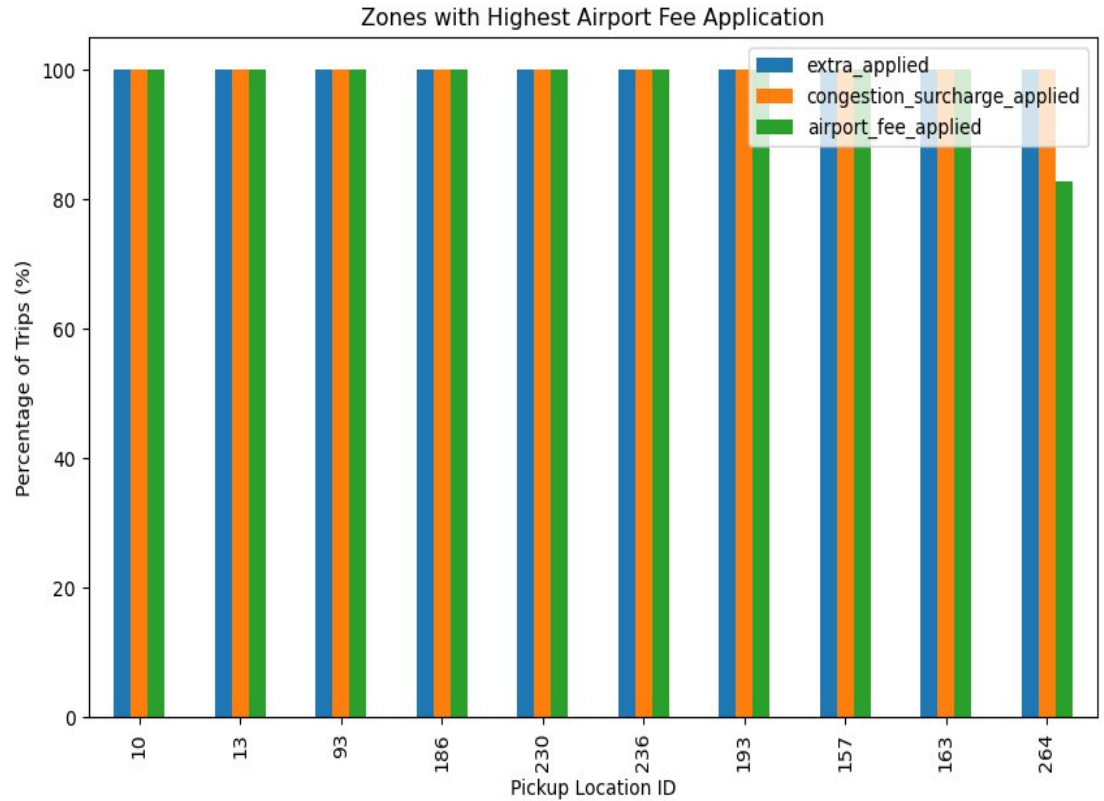
The analysis shows that both the *extra surcharge* and *congestion surcharge* are applied to nearly 100% of trips across all hours and zones. This indicates that these surcharges are mandatory components of the fare structure in the sampled dataset and therefore do not vary meaningfully by time or location.

In contrast, the *airport fee* exhibits strong temporal and spatial variation. Airport fees are applied to approximately 20–30% of trips during most hours of the day, with a significant spike during early morning hours (around 3–5 AM), where airport-related trips dominate. This reflects increased airport travel during off-peak public transport hours.

Zone-level analysis further confirms that airport fees are concentrated in specific pickup zones corresponding to airports and nearby access areas. Several zones show nearly 100% airport fee application, reinforcing the location-specific nature of this surcharge.

Overall, the findings indicate that while congestion-related surcharges are uniformly applied, airport fees are driven primarily by pickup location and time of day, making them a key variable for pricing strategy and operational planning.





4. Conclusions

4.1. Final Insights and Recommendations

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

Based on the temporal, geographical, and operational analyses conducted, several actionable recommendations can be made to optimize routing and dispatching strategies, improve fleet utilization, and reduce inefficiencies.

1. Time-based Dynamic Dispatching

The analysis shows distinct peak demand during morning (8 - 10 AM) and evening (4 - 7 PM) hours, especially on weekdays.

Dispatch systems should proactively reposition vehicles toward high-demand zones before these peak windows begin, rather than reacting after congestion builds up.

During late-night and early-morning hours (11 PM - 5 AM), demand is lower but concentrated in specific zones; targeted dispatching in these hours can reduce idle times.

Recommendation:

Implement time-aware dispatch algorithms that adjust vehicle availability and routing priorities based on hour-of-day demand patterns.

2. Zone-based Supply Rebalancing

Certain pickup zones consistently exhibit high pickup-to-dropoff ratios, indicating that vehicles frequently leave these zones and are not returning at the same rate.

Conversely, zones with high dropoffs but low pickups tend to accumulate idle vehicles.

Recommendation:

Use pickup/dropoff imbalance metrics to trigger automated vehicle rebalancing, ensuring that high-pickup zones maintain sufficient supply without over-crowding low-demand areas.

3. Congestion-Aware Routing

The slow-route analysis revealed that specific zone pairs during peak hours exhibit significantly lower average speeds, indicating congestion bottlenecks.

Persistently routing vehicles through these slow corridors increases trip duration and reduces fleet efficiency.

Recommendation:

Integrate hour-specific speed profiles into routing logic to avoid consistently slow routes during peak congestion periods and dynamically reroute vehicles through faster alternatives.

4. Weekend vs Weekday Dispatch Differentiation

Weekday traffic patterns show sharper peaks tied to work commutes, while weekends display more evenly distributed demand across daytime hours.

Applying the same dispatch strategy across all days leads to inefficiencies.

Recommendation:

Adopt separate dispatch models for weekdays and weekends, with weekday models prioritizing rush-hour clustering and weekend models focusing on steady, zone-based availability.

5. Airport and Surcharge-Heavy Zone Optimization

Zones associated with airport fees and higher surcharge frequency show predictable demand and revenue patterns.

Vehicles serving these zones often experience longer trip durations but higher per-trip value.

Recommendation:

Designate special dispatch pools for airport and surcharge-heavy zones to ensure experienced drivers and optimized routing are assigned, minimizing turnaround time while maximizing revenue efficiency.

Summary

By combining time-aware dispatching, zone-based rebalancing, congestion-sensitive routing, and day-specific strategies, fleet operations can be significantly optimized. These recommendations directly address the observed inefficiencies in trip duration, zone imbalance, and peak-hour congestion, enabling better customer service and higher operational efficiency.

4.1.1. Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analysing trip trends across time, days and months.

The spatial and temporal analysis of trip data reveals clear patterns in how demand varies across zones, hours of the day, days of the week, and months. These insights can be leveraged to strategically position cabs in advance, thereby reducing passenger wait times and improving fleet utilization.

1. Peak-Hour Zone Prioritization

Zones such as high-traffic commercial and transit areas consistently show elevated pickup volumes during weekday morning (8 - 10 AM) and evening (4 - 7 PM) peaks.

These zones often experience rapid depletion of available cabs due to high pickup-to-dropoff ratios.

Strategy:

Pre-position additional cabs in these high-demand zones 30 - 60 minutes before peak hours, ensuring adequate availability during rush periods and minimizing missed demand.

2. Night-Time and Late-Hour Concentration

Nighttime analysis (11 PM - 5 AM) indicates that demand becomes highly concentrated in a smaller subset of zones, including nightlife areas, transport hubs, and airport-connected locations.

Despite lower overall demand, trip value and tip percentages tend to be relatively higher during these hours.

Strategy:

Reduce fleet dispersion at night and concentrate cabs in identified night-active zones, allowing better coverage with fewer vehicles and reducing idle cruising.

3. Weekday vs Weekend Positioning

Weekday trips are driven largely by work commutes and follow predictable spatial patterns.

Weekend demand shifts toward leisure, residential, and entertainment zones, with peaks occurring later in the day.

Strategy:

Adopt day-specific positioning strategies:

Weekdays: focus on business districts, transit zones, and residential-to-commercial corridors.

Weekends: increase coverage near leisure hotspots, shopping areas, and nightlife zones.

4. Seasonal and Monthly Adjustments

Monthly trends indicate variations in demand intensity, likely influenced by weather conditions, holidays, and travel seasons.

Certain zones experience disproportionate increases during specific months.

Strategy:

Use historical monthly demand data to adjust baseline fleet allocation, increasing coverage in consistently high-growth zones during peak seasons while scaling back during low-demand months.

5. Balancing Zone Imbalances

Zones with high dropoff volumes but low pickups tend to accumulate idle cabs, while high-pickup zones experience shortages.

Strategy:

Implement dynamic rebalancing rules that periodically move idle cabs from surplus zones to deficit zones, ensuring a more even spatial distribution of supply.

Summary

Strategic cab positioning should be driven by time-aware, zone-specific, and day-sensitive insights. By proactively aligning vehicle availability with observed demand patterns across zones and timeframes, operators can significantly enhance service reliability, reduce inefficiencies, and maximize revenue opportunities.

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

The pricing analysis highlights meaningful variations in fare efficiency across time, distance, vendor, passenger count, and trip context. These insights can be used to design a dynamic and competitive pricing strategy that maximizes revenue without adversely affecting customer satisfaction.

1. Time-Based Dynamic Pricing

Average fare per mile increases during peak daytime hours, particularly between 8 - 11 AM and 4 - 7 PM, reflecting higher demand and congestion.

Nighttime hours generate a smaller share of revenue but show relatively stable fare efficiency.

Pricing Adjustment:

Introduce mild peak-hour price multipliers during consistently high-demand windows while maintaining base fares during off-peak periods. This balances revenue optimization with affordability.

2. Distance-Tiered Pricing Optimization

Short-distance trips (≤ 2 miles) show disproportionately high fare-per-mile values due to fixed base charges.

Medium and long trips exhibit more stable and predictable fare-per-mile behavior.

Pricing Adjustment:

Maintain competitive base fares for short trips to remain attractive, while optimizing per-mile rates for medium and long-distance trips, where price sensitivity is lower and revenue potential is higher.

3. Vendor-Specific Competitive Positioning

Vendor-level analysis reveals differences in average fare per mile and pricing behavior across hours.

One vendor consistently charges higher fares per mile, particularly during peak periods.

Pricing Adjustment:

Adopt competitive benchmarking against higher-priced vendors by offering slightly lower per-mile rates during peak hours, capturing price-sensitive demand without triggering price wars.

4. Passenger Count-Aware Pricing

Fare per mile per passenger decreases as passenger count increases, indicating cost-sharing effects.

Larger groups generate higher total fares despite lower per-person costs.

Pricing Adjustment:

Encourage group travel by maintaining favorable per-passenger pricing for multi-passenger trips, increasing total fare value while enhancing perceived fairness.

5. Zone and Context-Sensitive Surcharges

Airport-related zones and early-morning hours show higher surcharge frequency and trip value.

These trips are less price-sensitive due to urgency and limited alternatives.

Pricing Adjustment:

Sustain or slightly enhance context-specific surcharges (e.g., airport pickups, late-night hours) while clearly communicating fare components to maintain customer trust.

Summary

A balanced pricing strategy should combine time-based adjustments, distance-tier optimization, competitive vendor benchmarking, and context-aware surcharges. By aligning prices with observed demand elasticity and trip characteristics, operators can maximize revenue while preserving competitiveness and customer satisfaction.