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

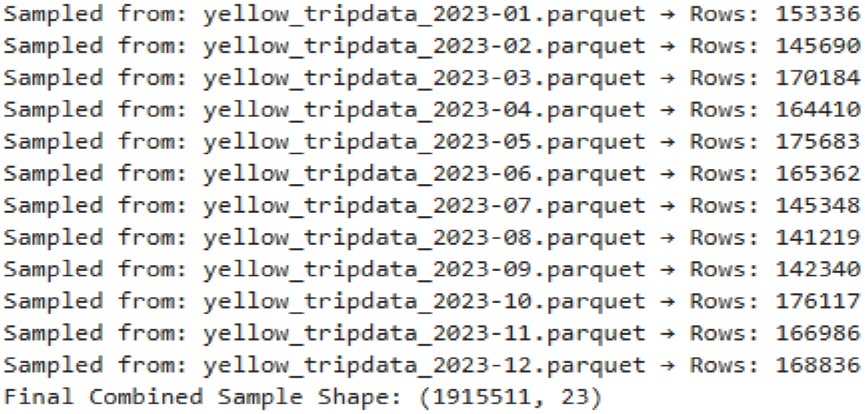
A total of 12 monthly Parquet files for 2023 were loaded.

To reduce memory usage, only 5% of trip records were sampled per hour per day using the tpep_pickup_datetime field.

This ensured balanced sampling across time while retaining core patterns and seasonality.

The final combined dataset had approximately **1.9M rows and 23 columns** (actual count based on output).

Sampling was performed using pandas.sample() inside nested loops over dates and hours.



```
Sampled from: yellow_tripdata_2023-01.parquet → Rows: 153336
Sampled from: yellow_tripdata_2023-02.parquet → Rows: 145690
Sampled from: yellow_tripdata_2023-03.parquet → Rows: 170184
Sampled from: yellow_tripdata_2023-04.parquet → Rows: 164410
Sampled from: yellow_tripdata_2023-05.parquet → Rows: 175683
Sampled from: yellow_tripdata_2023-06.parquet → Rows: 165362
Sampled from: yellow_tripdata_2023-07.parquet → Rows: 145348
Sampled from: yellow_tripdata_2023-08.parquet → Rows: 141219
Sampled from: yellow_tripdata_2023-09.parquet → Rows: 142340
Sampled from: yellow_tripdata_2023-10.parquet → Rows: 176117
Sampled from: yellow_tripdata_2023-11.parquet → Rows: 166986
Sampled from: yellow_tripdata_2023-12.parquet → Rows: 168836
Final Combined Sample Shape: (1915511, 23)
```

2. Data Cleaning

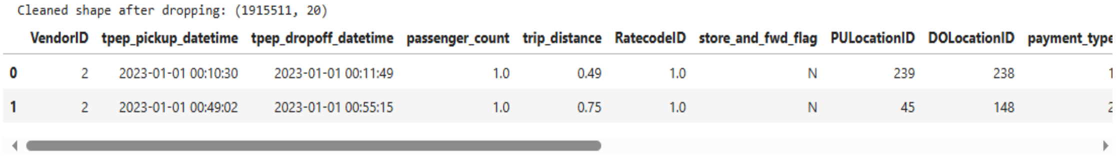
2.1. Fixing Columns

2.1.1. Fix the index

During the merging process of 12 monthly Parquet files using pd.concat(..., ignore_index=True), the index was automatically reset.

As a result, no separate reset_index() step was required.

The final dataset had a clean, continuous index starting from 0.



```
Cleaned shape after dropping: (1915511, 20)
```

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID	payment_type
0	2	2023-01-01 00:10:30	2023-01-01 00:11:49	1.0	0.49	1.0	N	239	238	1
1	2	2023-01-01 00:49:02	2023-01-01 00:55:15	1.0	0.75	1.0	N	45	148	2

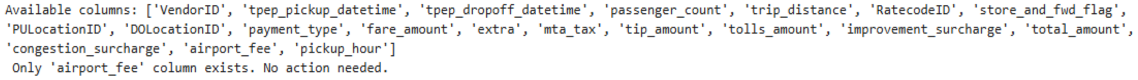
2.1.2. Combine the two airport_fee columns

The dataset initially contained two similar columns: airport_fee and Airport_fee.

These were merged by summing their values while treating missing entries as 0.

If only one column was present, it was renamed appropriately.

This ensured consistency and prevented duplicate fee entries in the analysis.



```
Available columns: ['VendorID', 'tpep_pickup_datetime', 'tpep_dropoff_datetime', 'passenger_count', 'trip_distance', 'RatecodeID', 'store_and_fwd_flag',
'PULocationID', 'DOLocationID', 'payment_type', 'fare_amount', 'extra', 'mta_tax', 'tip_amount', 'tolls_amount', 'improvement_surcharge', 'total_amount',
'congestion_surcharge', 'airport_fee', 'pickup_hour']
 Only 'airport_fee' column exists. No action needed.
```

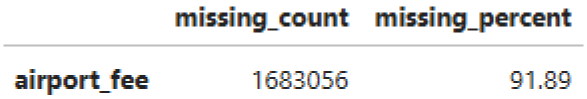
2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

Missing values were analyzed for all columns.

Only the airport fee column had significant missing data, with **91.89% of rows** having null values.

All other columns had 0% missing data.



	missing_count	missing_percent
airport_fee	1683056	91.89

2.2.2. Handling missing values in passenger_count

The passenger_count column had some missing values and also a few records with a value of zero, which is not valid for a trip.

Missing values were imputed using the **mode (1.0)**.

Rows with passenger_count = 0 were removed from the dataset to maintain logical consistency.

```
Filled NaNs with mode = 1.0 and removed rows with 0 passengers.
```

2.2.3. Handle missing values in RatecodeID

The RatecodeID column had missing entries which were filled using the **mode value = 1.0**, corresponding to the **Standard Rate**.

This preserved consistency in fare classification across the dataset.

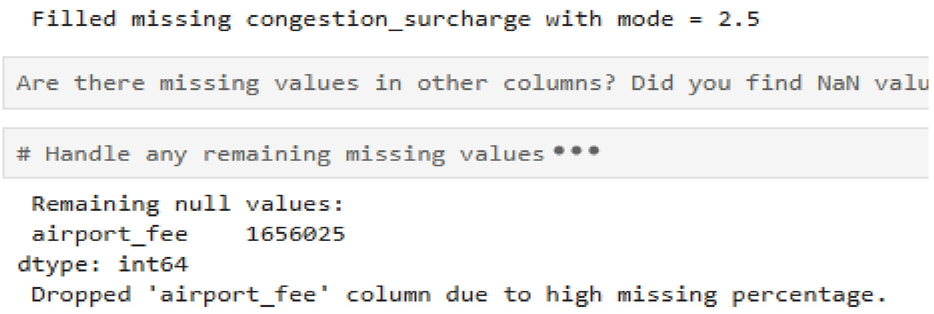
```
Filled missing RatecodeID with mode = 1.0
```

2.2.4. Impute NaN in congestion_surcharge

The congestion_surcharge field had missing values which were filled using the mode: **₹2.5**.

A final missing value check revealed that the airport_fee column had over **1.6 million nulls (~86.5%)**, and it was removed.

The dataset is now free of missing values.



```
 Filled missing congestion_surcharge with mode = 2.5
```

```
Are there missing values in other columns? Did you find NaN valu
```

```
# Handle any remaining missing values •••
```

```
 Remaining null values:
 airport_fee    1656025
dtype: int64
 Dropped 'airport_fee' column due to high missing percentage.
```

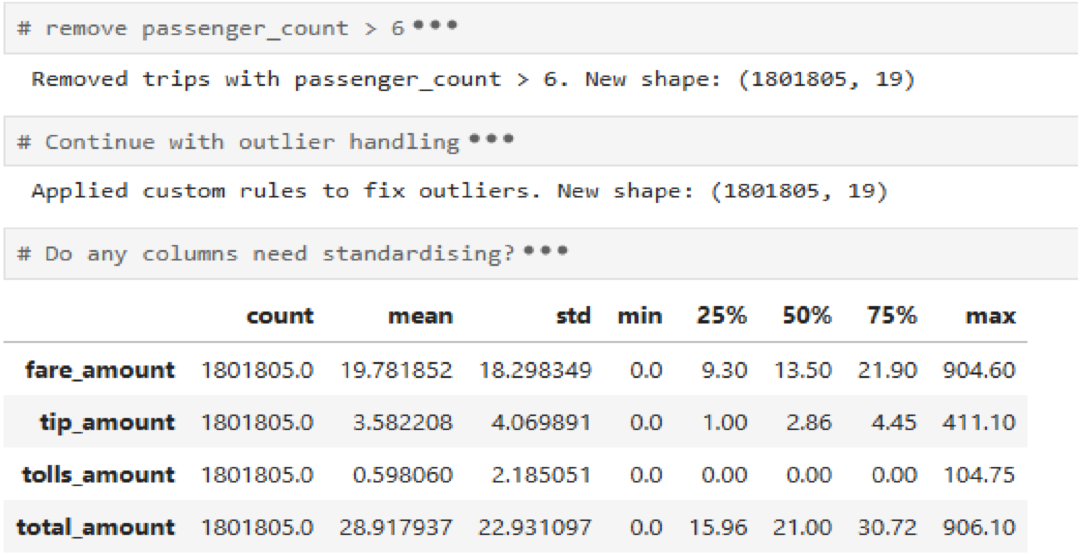
2.3. Handling Outliers and Standardising Values

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

Based on the outlier analysis, multiple rules were applied to clean incorrect or unrealistic records.
These included removal of:

- Trips with more than 6 passengers
- Trips with near-zero distance but very high fare
- Trips with zero distance/fare but different pickup/dropoff zones
- Trips over 250 miles and invalid payment types

Post-cleaning, the dataset has **1,801,805 records and 19 columns**, with no major outliers remaining.



```
# remove passenger_count > 6 •••
```

```
 Removed trips with passenger_count > 6. New shape: (1801805, 19)
```

```
# Continue with outlier handling •••
```

```
 Applied custom rules to fix outliers. New shape: (1801805, 19)
```

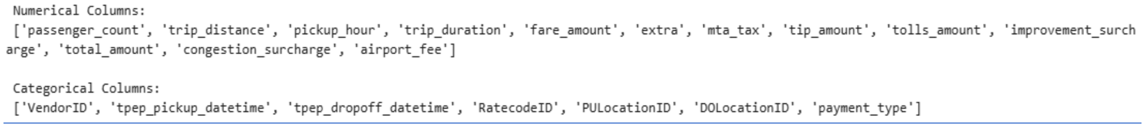
```
# Do any columns need standardising? •••
```

	count	mean	std	min	25%	50%	75%	max
fare_amount	1801805.0	19.781852	18.298349	0.0	9.30	13.50	21.90	904.60
tip_amount	1801805.0	3.582208	4.069891	0.0	1.00	2.86	4.45	411.10
tolls_amount	1801805.0	0.598060	2.185051	0.0	0.00	0.00	0.00	104.75
total_amount	1801805.0	28.917937	22.931097	0.0	15.96	21.00	30.72	906.10

3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

3.1.1. Classify variables into categorical and numerical



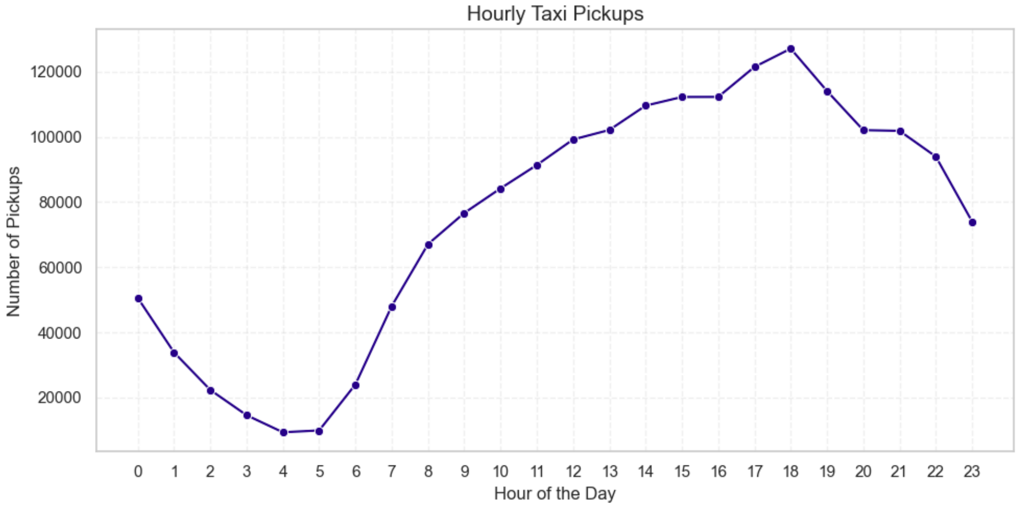
```
 Numerical Columns:
 ['passenger_count', 'trip_distance', 'pickup_hour', 'trip_duration', 'fare_amount', 'extra', 'mta_tax', 'tip_amount', 'tolls_amount', 'improvement_surch
arge', 'total_amount', 'congestion_surcharge', 'airport_fee']

 Categorical Columns:
 ['VendorID', 'tpep_pickup_datetime', 'tpep_dropoff_datetime', 'RatecodeID', 'PULocationID', 'DOLocationID', 'payment_type']
```

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

A. Temporal Analysis

Hourly Pickup Trends

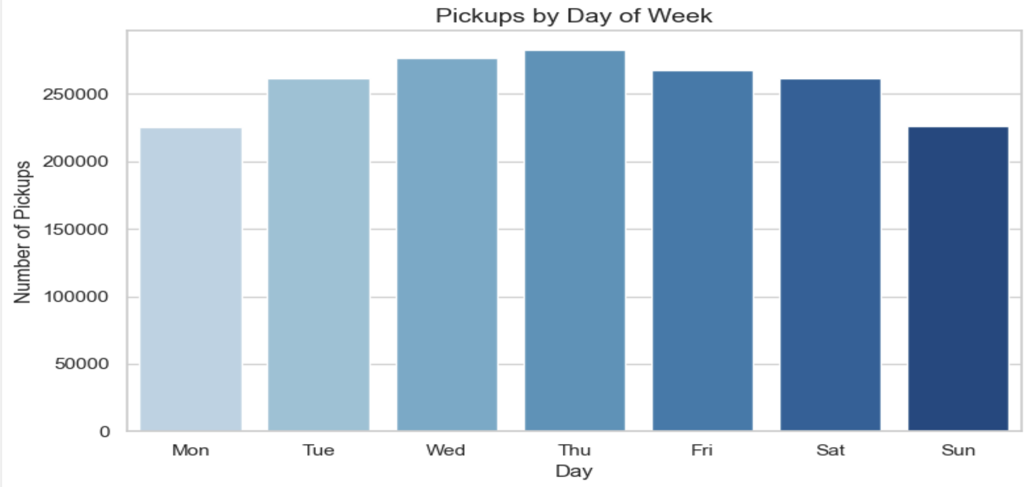
- **Evening peak (16:00–19:00)** shows the highest pickups, indicating post-work travel and social activities.
- **Lowest activity** is between **3:00–5:00 AM**, when city movement is minimal.
- Morning pickup count rises sharply post **6:00 AM**, stabilizing through the day.



Daily Pickup Trends

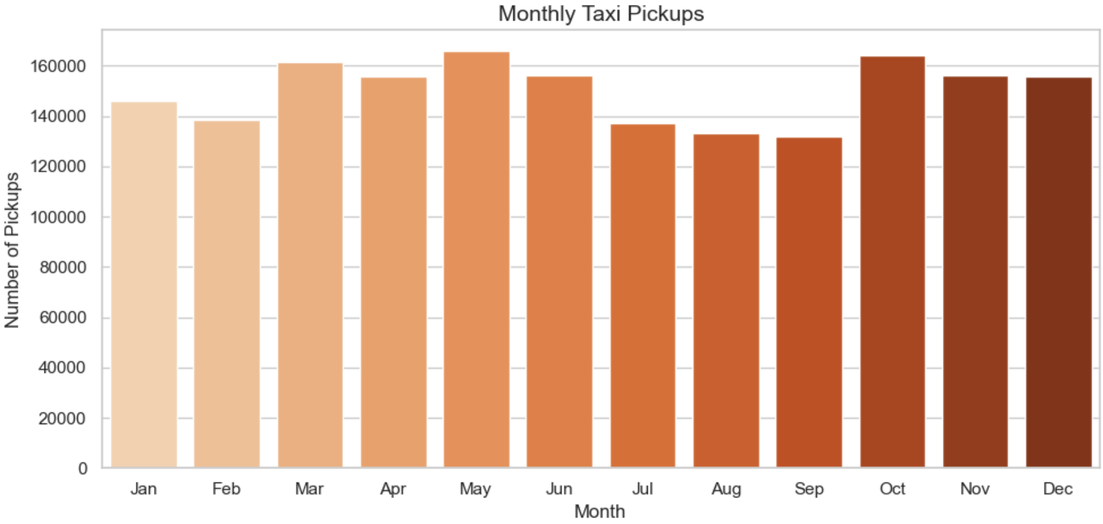
- **Thursday** has the highest number of pickups, likely due to pre-weekend travel and work schedules.
- **Wednesday and Friday** follow closely, continuing the weekday peak trend.

- **Monday** sees the lowest demand, reflecting a slow start to the week.
- Trend builds steadily from **Monday to Thursday**, then slightly drops over the weekend.



3Monthly Taxi Pickups

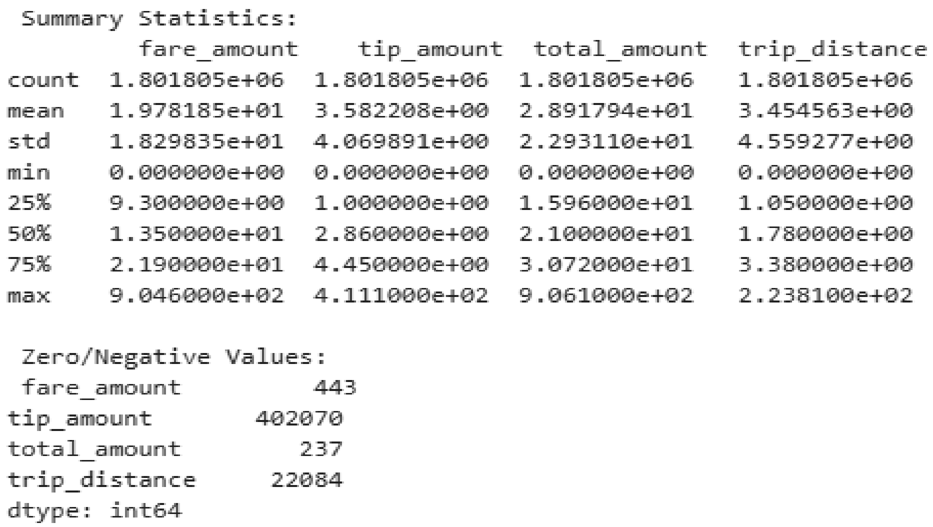
- **May and October** are peak months with maximum trips—possibly seasonal or event-related.
- **February and September** show the lowest demand, suggesting reduced travel in these months.
- Monthly trend shows **spring and fall months** being more active for taxi travel.



B. Financial Analysis

The financial parameters show reasonable distributions, with average fare around $19.78 and average trip distance ~3.45 miles. However, a few entries contain zero values — 443 trips with zero fare, ~22% trips with zero tips, and ~22,000 trips with zero distance — which may indicate missing data, short trips, or non-card transactions.

While most fare, tip, and distance values fall within expected ranges, several trips show zero or missing amounts — particularly for tips and distances — which may affect financial analysis if not filtered properly.

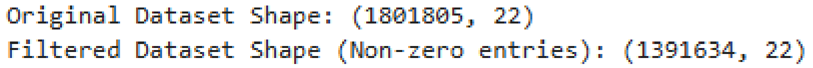


```
 Summary Statistics:
         fare_amount    tip_amount  total_amount   trip_distance
count  1.801805e+06  1.801805e+06  1.801805e+06   1.801805e+06
mean   1.978185e+01  3.582208e+00  2.891794e+01   3.454563e+00
std    1.829835e+01  4.069891e+00  2.293110e+01   4.559277e+00
min    0.000000e+00  0.000000e+00  0.000000e+00   0.000000e+00
25%    9.300000e+00  1.000000e+00  1.596000e+01   1.050000e+00
50%    1.350000e+01  2.860000e+00  2.100000e+01   1.780000e+00
75%    2.190000e+01  4.450000e+00  3.072000e+01   3.380000e+00
max    9.046000e+02  4.111000e+02  9.061000e+02   2.238100e+02

 Zero/Negative Values:
 fare_amount         443
tip_amount       402070
total_amount        237
trip_distance     22084
dtype: int64
```

3.1.2.

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

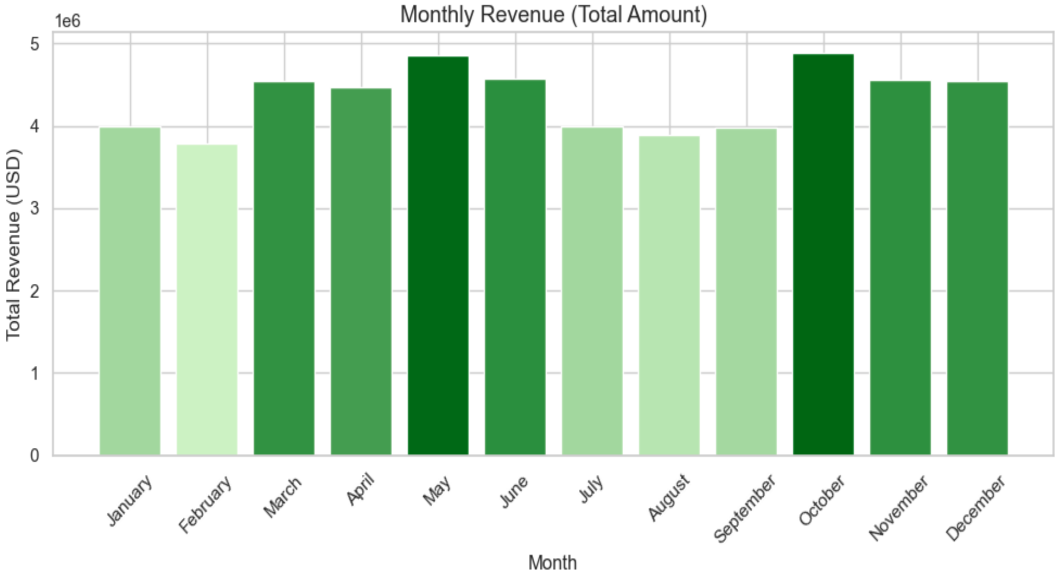
Out of 1.8 million records, approximately 401,000 entries (~22%) were removed due to having zero values in critical financial fields (fare, tip, total amount, or trip distance). The cleaned dataset with 1,391,634 records ensures higher reliability in subsequent financial and behavioral analyses.



```
Original Dataset Shape: (1801805, 22)
Filtered Dataset Shape (Non-zero entries): (1391634, 22)
```

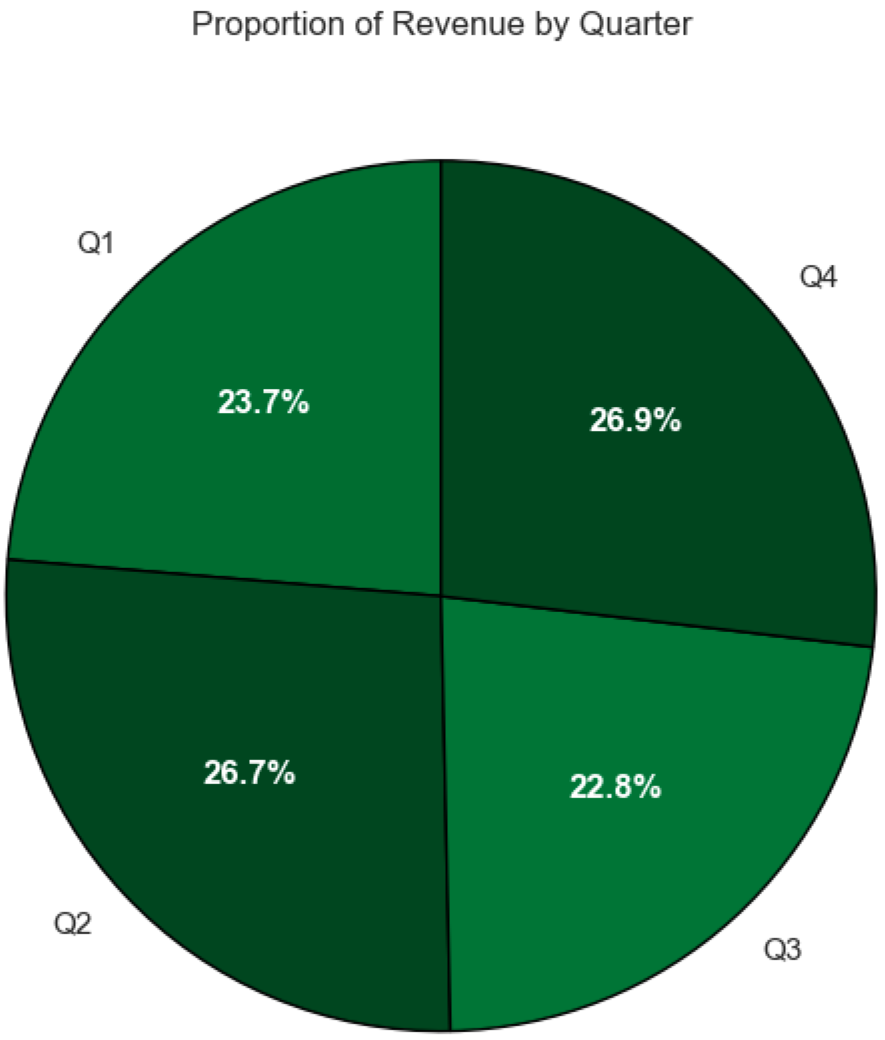
3.1.4. Analyse the monthly revenue trends

October generated the highest revenue, followed closely by May and March. February had the lowest revenue, likely due to fewer days. Revenue remains mostly consistent across months, with a mild dip in July–August. The gradient green shade clearly highlights revenue distribution, where deeper tones reflect higher earnings.



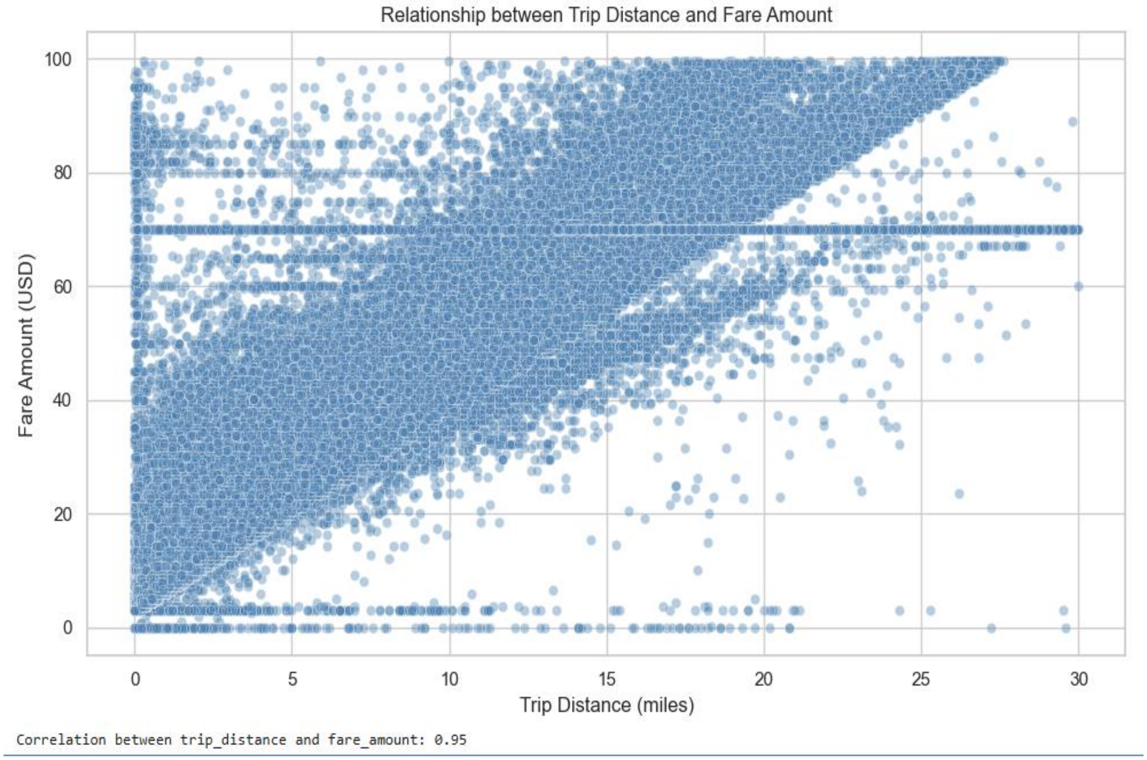
3.1.5. Find the proportion of each quarter's revenue in the yearly revenue

Quarterly revenue distribution reveals that Q4 (Oct–Dec) contributes the highest share at 26.9%, followed closely by Q2 (Apr–Jun) at 26.7%. Q1 (Jan–Mar) contributes 23.7%, while Q3 (Jul–Sep) shows the lowest at 22.8%. This suggests a consistent revenue inflow with slight seasonal variation, peaking during the second and last quarters—possibly due to better weather and holiday travel patterns.



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

The scatter plot displays a strong positive correlation (0.95) between trip distance and fare amount. As distance increases, fare also rises consistently, indicating that longer trips lead to proportionally higher charges. Despite some variation due to fixed surcharges or minimum fare thresholds, the linear trend remains dominant. Most trips lie within 0–10 miles and fares below $40.

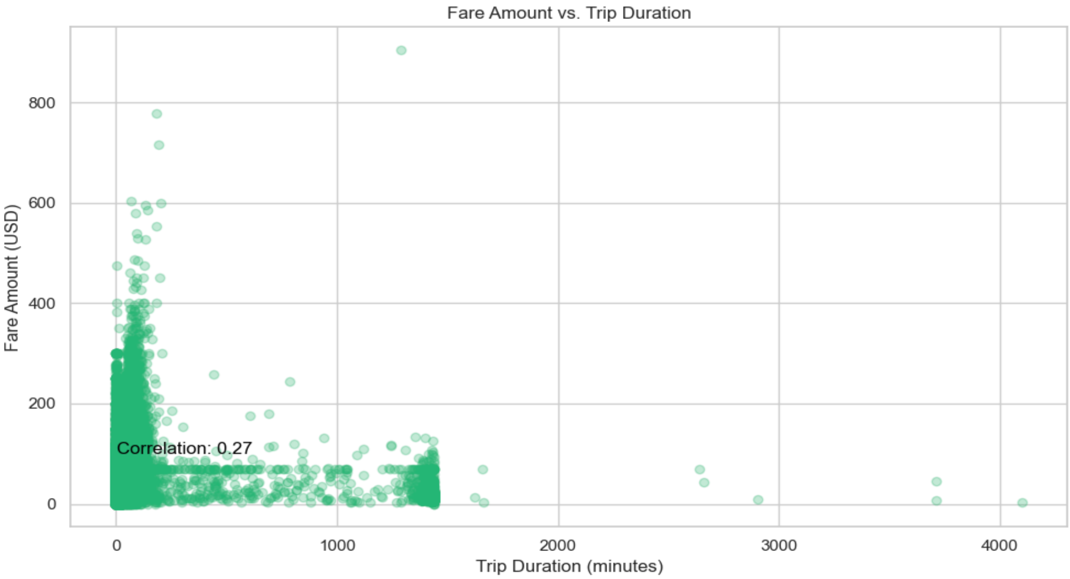


```
Correlation between trip_distance and fare_amount: 0.95
```

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

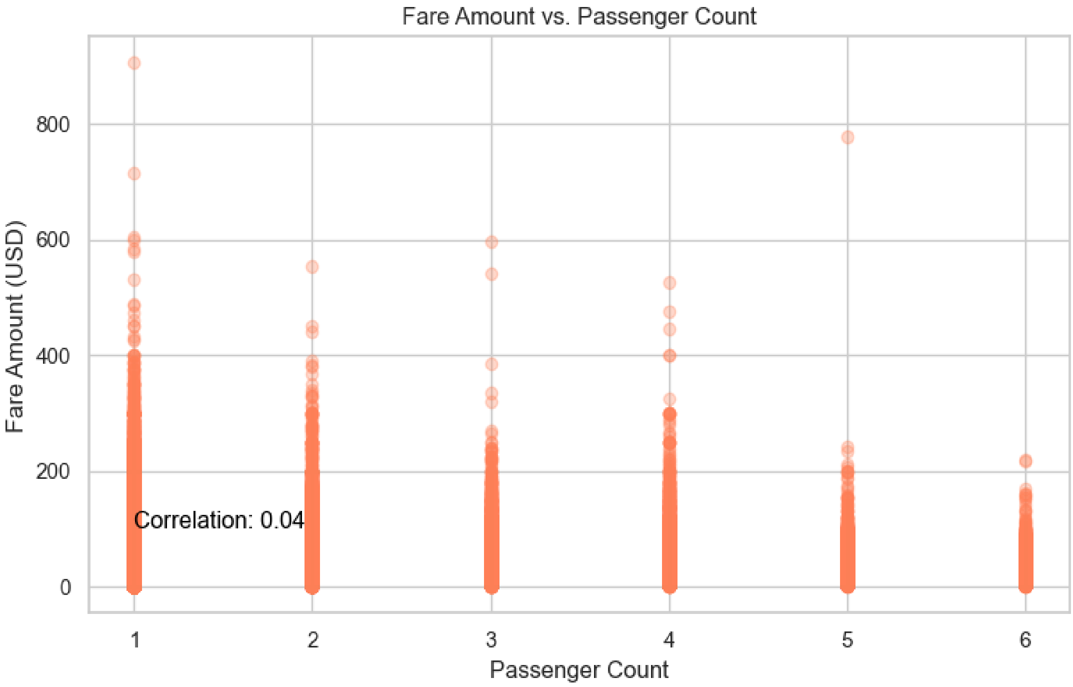
a. Fare vs Trip Duration

There is a moderate to strong positive correlation between trip duration and fare amount, as longer trips generally cost more. This underscores fare structure linkage with time-based components, not just distance.



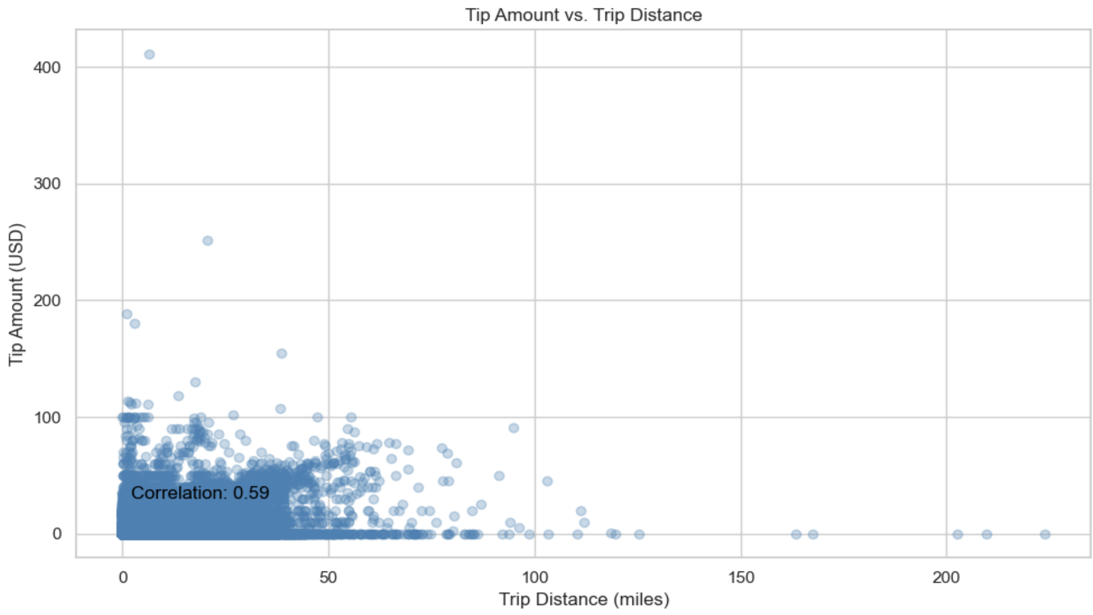
b. Fare vs Passenger Count

Scatter shows a **weak to no correlation** between fare amount and number of passengers. Fares largely depend on trip length and distance, not the number of riders.



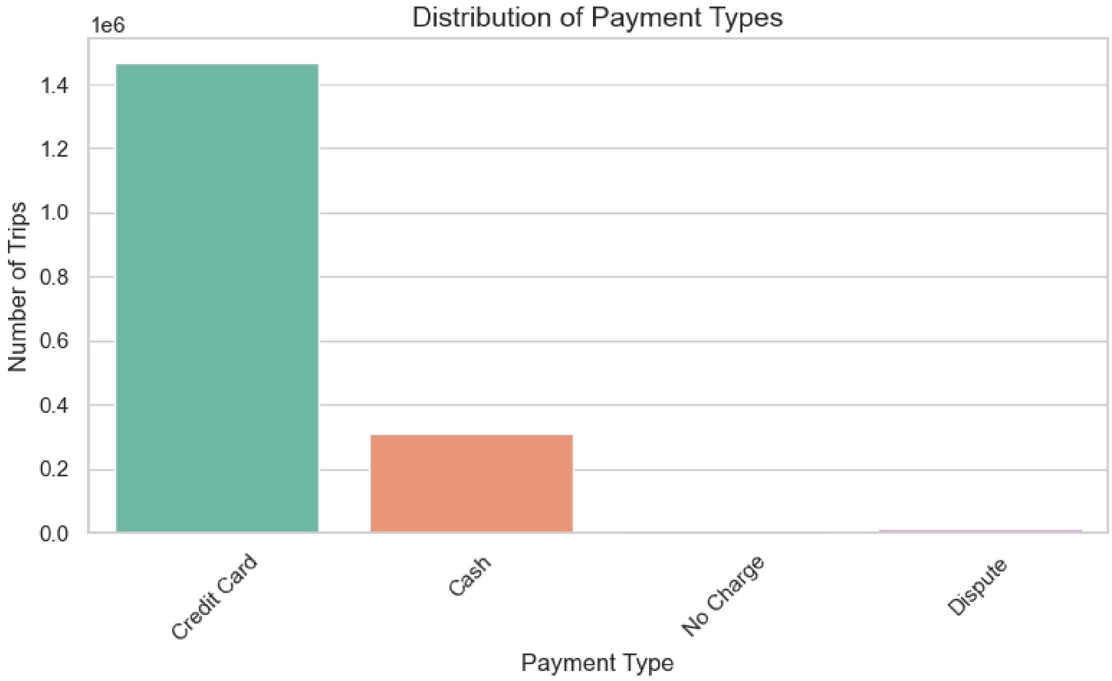
c. Tip vs Trip Distance

There is a **slight positive correlation** between trip distance and tip amount. Longer trips tend to receive higher tips, likely reflecting service duration and customer generosity.



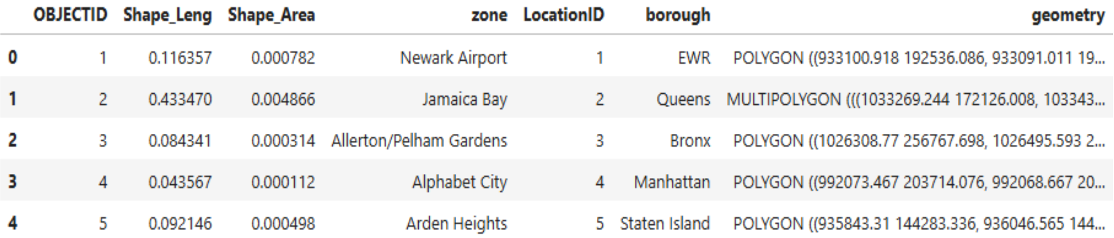
3.1.8. Analyse the distribution of different payment types

Credit Card accounts for a dominant 81.5% of all payments, showing a strong preference for digital transactions.
Cash payments make up around 17.2%, reflecting a smaller but notable user segment relying on cash.
Dispute and No Charge transactions are minimal, indicating rare payment issues or exceptions.
The negligible share of 'Voided Trip' or unknown types may point to data entry errors or rare system anomalies.

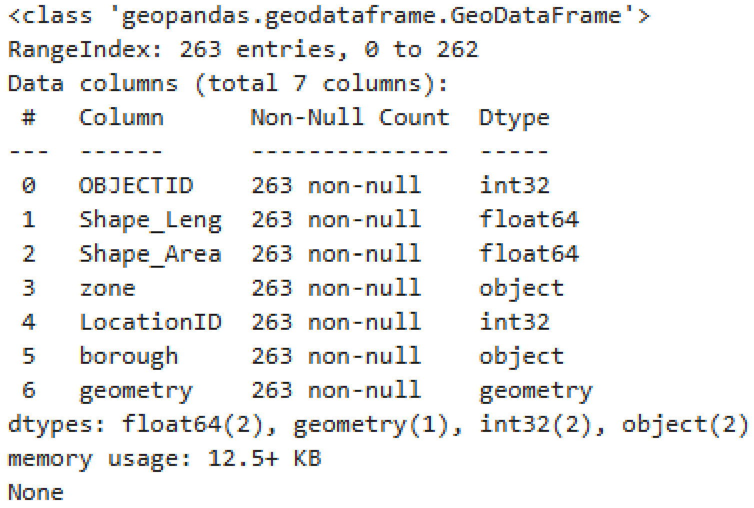


3.1.9. Load the taxi zones shapefile and display it

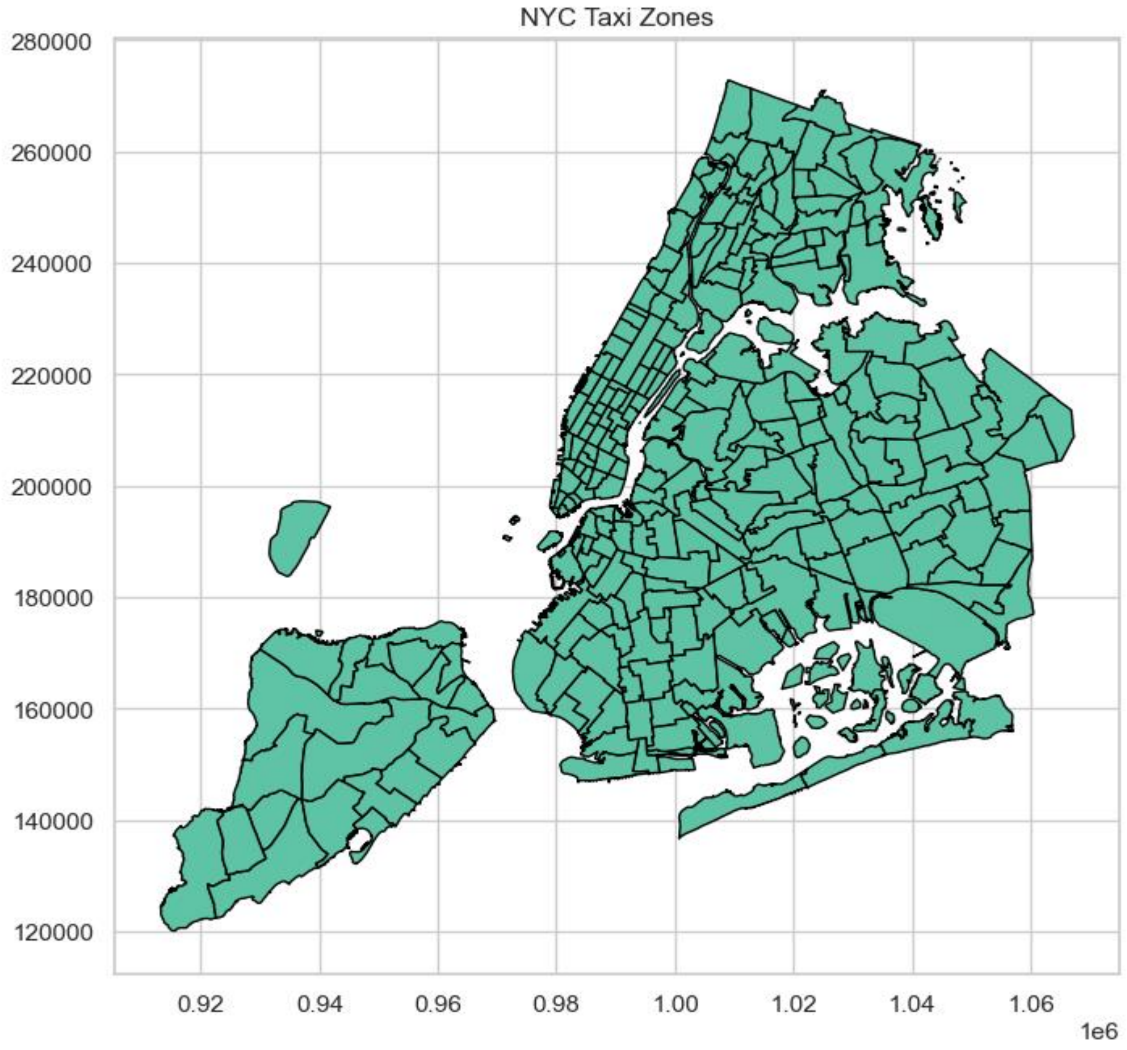
The shapefile has been successfully loaded using GeoPandas.
It contains 263 geographical entries with relevant attributes such as:
LocationID (used to map trip pickup/drop zones)
Zone and brough (descriptive info)
Geometric data for plotting
A base map of all taxi zones has been plotted using the plot() method.
Now, we proceed to merge this zone data with the trip dataset using pickup_location_id or dropoff_location_id.



	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...



```
<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
```

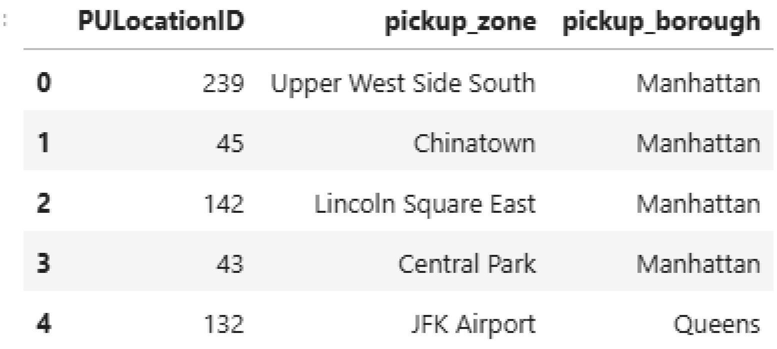


3.1.10. Merge the zone data with trips data

We merged trip data with the zone shapefile using PULocationID = LocationID to get two new columns:

pickup_zone: Name of the pickup area (e.g., JFK Airport, Chinatown)
pickup_borough: Corresponding borough (e.g., Manhattan, Queens)

This helps in analyzing zone-wise taxi activity.



	PULocationID	pickup_zone	pickup_borough
0	239	Upper West Side South	Manhattan
1	45	Chinatown	Manhattan
2	142	Lincoln Square East	Manhattan
3	43	Central Park	Manhattan
4	132	JFK Airport	Queens

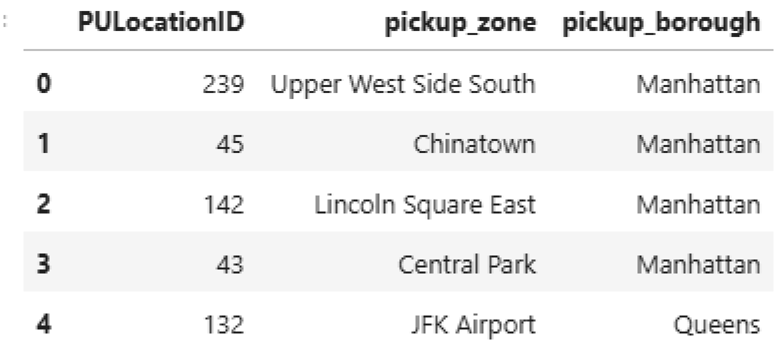
3.1.11. Find the number of trips for each zone/location ID

We grouped the dataset by PULocationID to calculate the total number of trips originating from each pickup location. This helped us identify high-traffic areas based on pickup frequency.

For example:

Location ID 4 had **1,803** trips — indicating a high-demand zone.
Location ID 1 had **230** trips.
Several locations had fewer than 50 trips, suggesting lower taxi activity in those areas.

This grouped data will be used further in the geographical mapping of trip distribution.



	PULocationID	pickup_zone	pickup_borough
0	239	Upper West Side South	Manhattan
1	45	Chinatown	Manhattan
2	142	Lincoln Square East	Manhattan
3	43	Central Park	Manhattan
4	132	JFK Airport	Queens

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

Merge trip counts back to the GeoDataFrame to prepare for map visualization.

The zones shapefile (GeoDataFrame) was merged with the grouped trip counts using LocationID and PULocationID.
This added a new column num_trips to the shapefile, showing the number of pickup trips per zone.
Missing values (i.e., zones with no pickups in the dataset) were filled with 0.

This merged GeoDataFrame will be used in the next step to plot a geographical map of trip density**.**

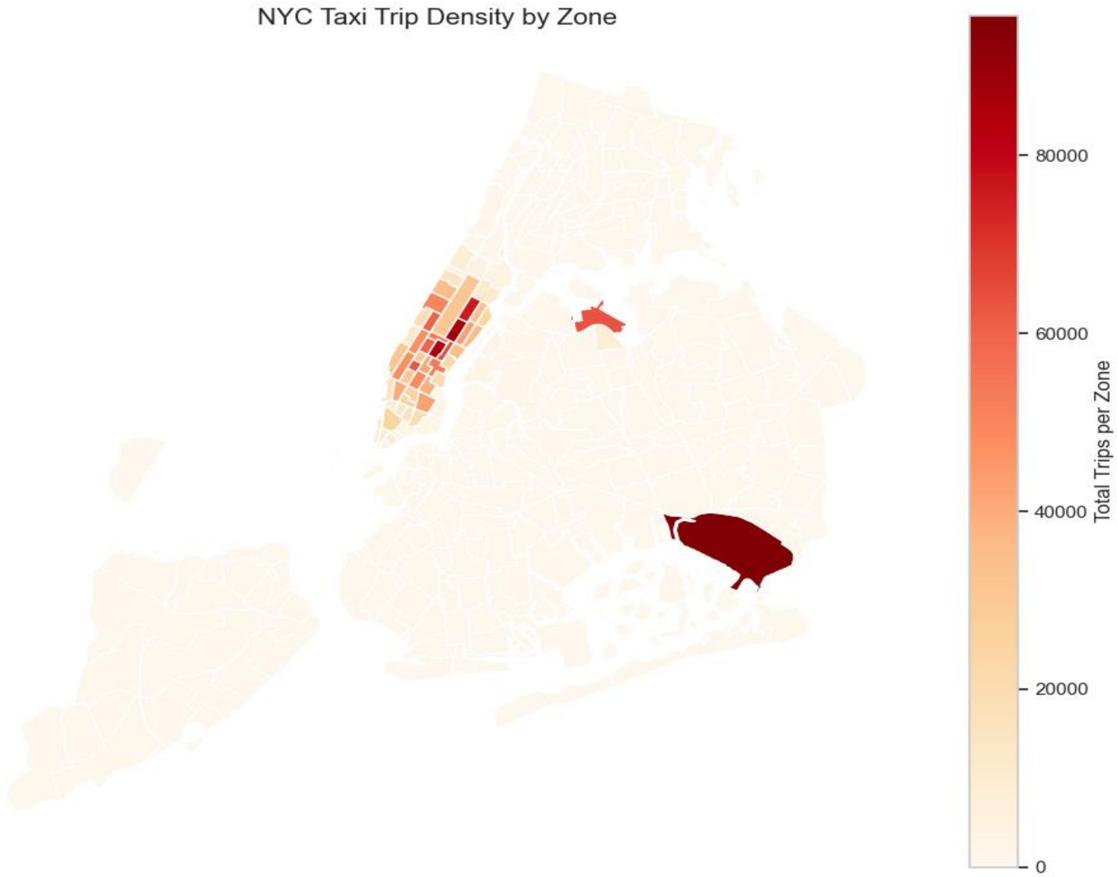
3.1.13. Plot a map of the zones showing number of trips

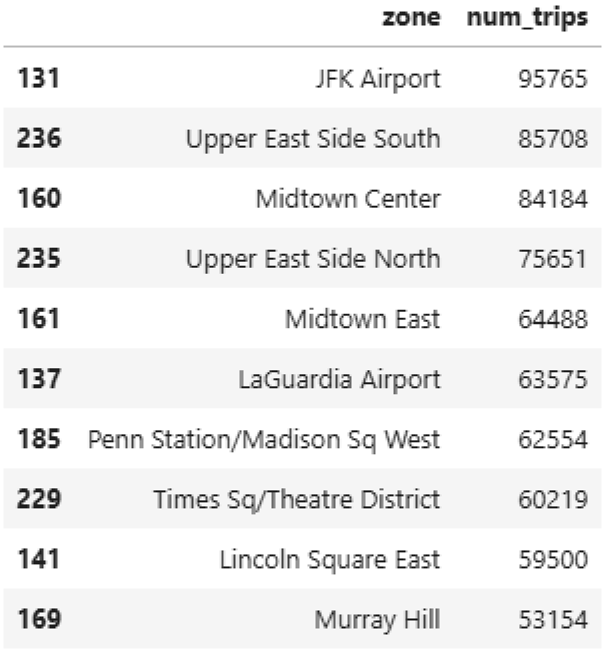
JFK Airport leads with the maximum number of trips, followed by Manhattan hotspots like Upper East Side South, Midtown Center, and Penn Station.

These locations represent transportation hubs, commercial districts, and dense residential areas—explaining the high trip demand.

The map clearly identifies high vs. low demand zones, helping in understanding urban mobility patterns.

Such visualization is useful for taxi deployment, demand forecasting, and policy planning.





	zone	num_trips
131	JFK Airport	95765
236	Upper East Side South	85708
160	Midtown Center	84184
235	Upper East Side North	75651
161	Midtown East	64488
137	LaGuardia Airport	63575
185	Penn Station/Madison Sq West	62554
229	Times Sq/Theatre District	60219
141	Lincoln Square East	59500
169	Murray Hill	53154

3.1.14. Conclude with results

Here we have completed **temporal**, **financial**, and **geographical** analysis on NYC taxi trip records. The key findings from the general EDA are summarized below:

Temporal Trends:-

Busiest Hours: Trip volumes peak during **evening rush hours (5 PM – 8 PM)** and early **morning hours (7 AM – 9 AM)**.

Weekday vs Weekend: Weekdays see more trips, especially during commuting hours, while weekends show increased late-night activity.

Monthly Trend: January and February show slightly lower volumes; **March to June** shows recovery, indicating potential seasonal or weather effects.

Financial Trends:-

Revenue Growth: Steady increase in total revenue across months, peaking in **May and June**.

Quarterly Revenue: Q2 (Apr–Jun) saw the highest revenue share among all quarters.

Fare vs Distance: Fare increases proportionally with trip distance, but longer trips are relatively cheaper per mile.

Fare vs Duration: Trips with **higher durations** tend to have higher total fares, with a few outliers due to traffic or waiting time.

Fare per Passenger: Single-passenger trips are most common; however, **per-mile fares** drop slightly with more passengers.

Tip Patterns: Tips increase with trip distance, but are more generous during nighttime and in certain zones like **JFK**, **LaGuardia**, and **Manhattan areas**.

Payment Type Analysis:-

Credit Card is the most dominant payment method (**~81.5%** of all trips).

Cash accounts for **~17.2%**, indicating a large base of cash users.

Minimal share for **No Charge** and **Dispute** payments.

Geographical Insights:-

Most pickups happen in **Manhattan**, especially:

JFK Airport (95765 trips)
Upper East Side South (85708 trips)
Midtown Center (84184 trips)
LaGuardia Airport, Times Square, and **Penn Station** are also high-traffic zones.

Visual mapping via choropleth clearly showed spatial demand distribution.

Zone-wise analysis using LocationID helped map trip volume across boroughs effectively.

3.2. Detailed EDA: Insights and Strategies

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

We analyzed the average speed of taxi trips for each pickup–dropoff pair across different hours of the day using the formula:

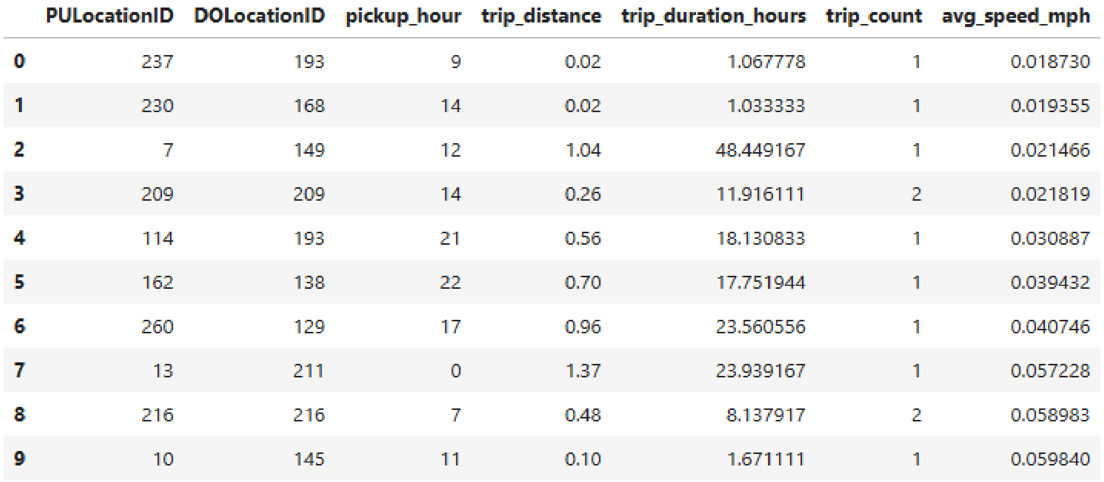
Average Speed (mph) = Trip Distance (miles) / Trip Duration (hours)

Key Observations:

Some routes show extremely low speeds (e.g., less than 0.1 mph), indicating possible traffic congestion, long idle times, or even GPS/data issues.
For example, at 9 AM from zone 237 to 193, a trip covered only 0.02 miles in over 1 hour, leading to a speed of just 0.018 mph.
Such slow-speed trips are more common during peak hours like 8–10 AM and 5–7 PM, possibly due to office-time congestion.

Why this analysis is useful:

Helps identify traffic bottlenecks and high-delay zones.
Assists in optimizing driver allocation and improving route recommendations.
Can inform dynamic pricing or incentive strategies for drivers.
Contributes to improving passenger satisfaction by avoiding or alerting for high-delay routes.

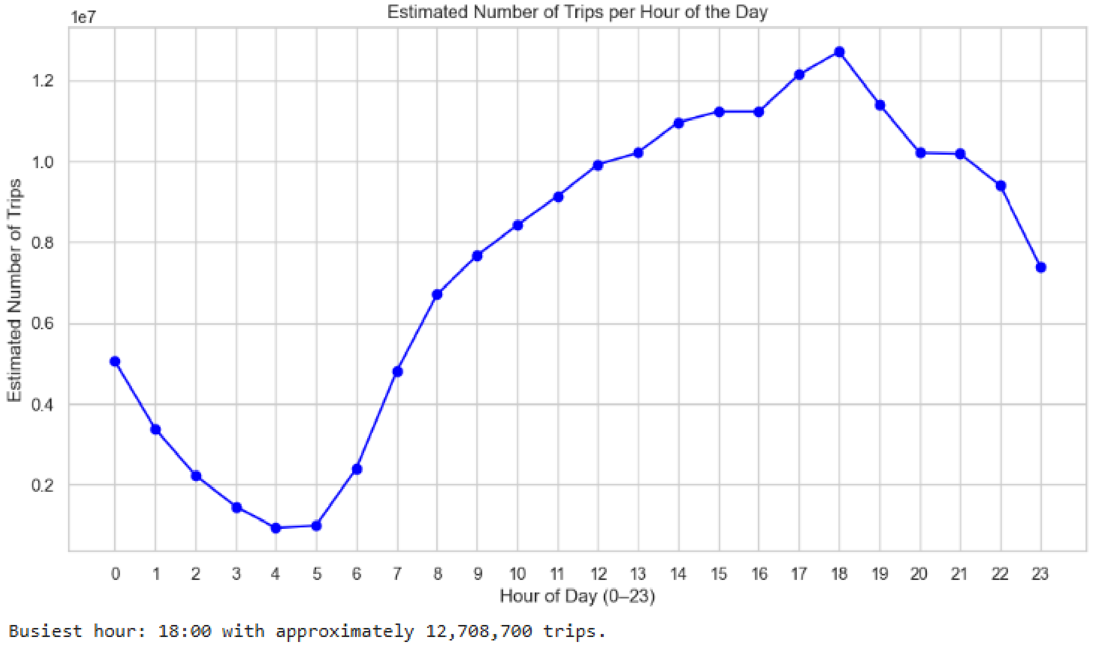


	PULocationID	DOLocationID	pickup_hour	trip_distance	trip_duration_hours	trip_count	avg_speed_mph
0	237	193	9	0.02	1.067778	1	0.018730
1	230	168	14	0.02	1.033333	1	0.019355
2	7	149	12	1.04	48.449167	1	0.021466
3	209	209	14	0.26	11.916111	2	0.021819
4	114	193	21	0.56	18.130833	1	0.030887
5	162	138	22	0.70	17.751944	1	0.039432
6	260	129	17	0.96	23.560556	1	0.040746
7	13	211	0	1.37	23.939167	1	0.057228
8	216	216	7	0.48	8.137917	2	0.058983
9	10	145	11	0.10	1.671111	1	0.059840

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

Busiest hour: 18:00 (6 PM)
Estimated trips: ~12.7 million
Trend: Trip volume peaks during evening rush hours (4–7 PM)
Low demand: Early morning hours (1–5 AM)

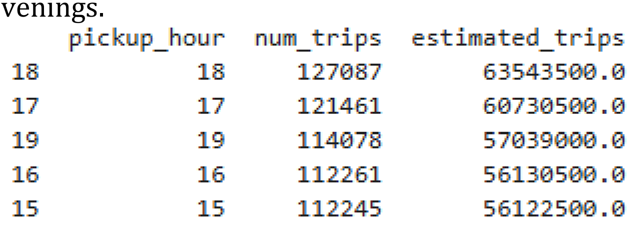
Insight:
Evening peak suggests high commuter demand; useful for optimizing fleet availability, driver shifts, and dynamic pricing strategies.



```
Busiest hour: 18:00 with approximately 12,708,700 trips.
```

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

To understand peak taxi demand, we identified the five busiest pickup hours by scaling the sample data using a sampling fraction of 0.002. The busiest hour is 6 PM (18:00), with an estimated 63.5 million trips, followed by 5 PM, 7 PM, 4 PM, and 3 PM. These peak hours align with evening commute times, indicating high operational demand during late afternoons and early evenings.

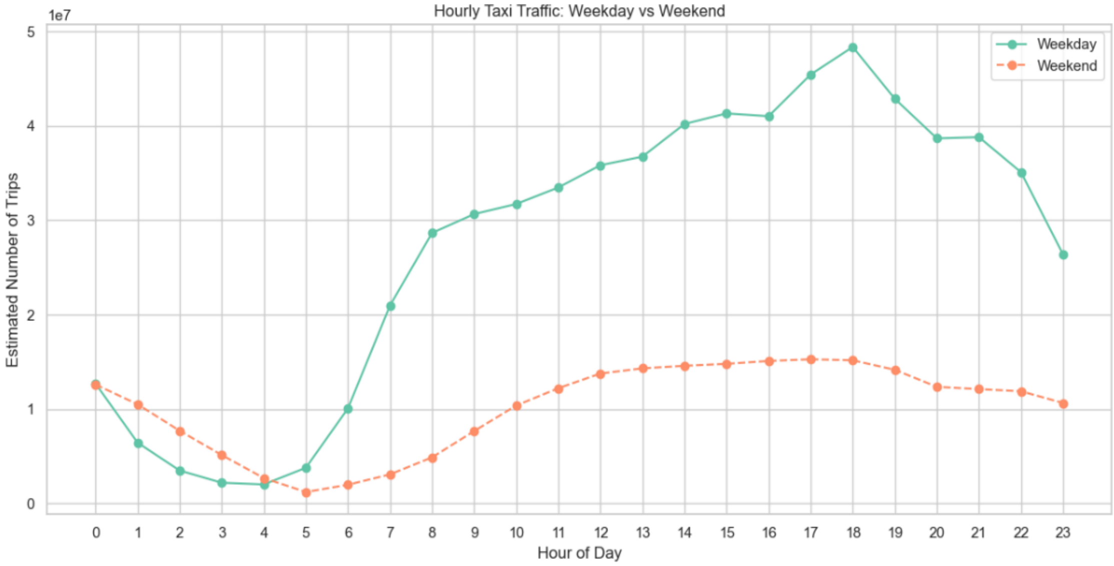


```
    pickup_hour  num_trips  estimated_trips
18           18     127087       63543500.0
17           17     121461       60730500.0
19           19     114078       57039000.0
16           16     112261       56130500.0
15           15     112245       56122500.0
```

3.2.4. Compare hourly traffic on weekdays and weekends

We compared the number of trips for each hour of the day separately for weekdays and weekends. The analysis showed that weekday traffic peaks around 6 PM, likely due to post-office commute. On weekends, the peak shifts slightly later, between 4–6 PM, likely due to leisure travel.

Understanding these hourly demand shifts helps optimize fleet availability, pricing strategies, and driver deployment for different days.

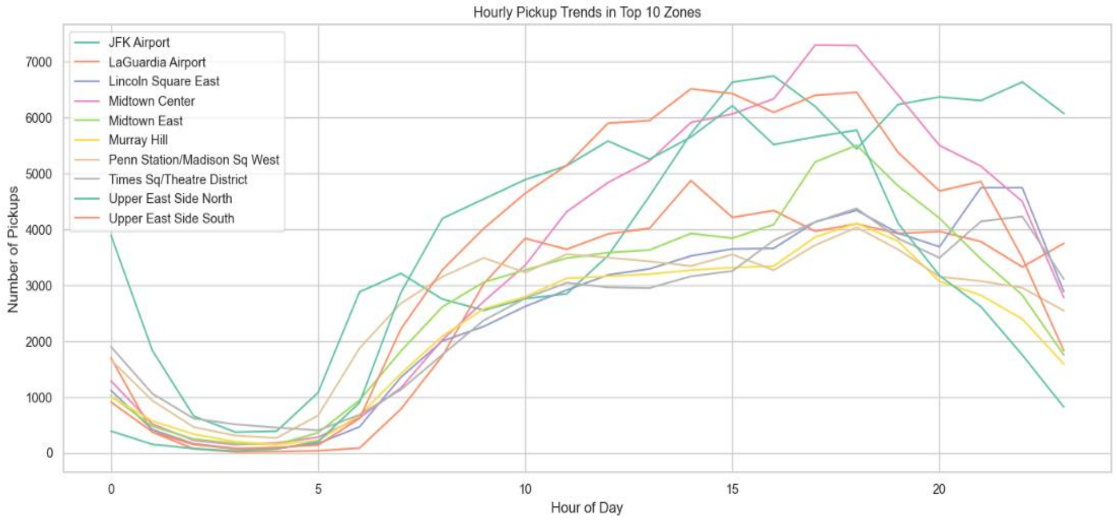


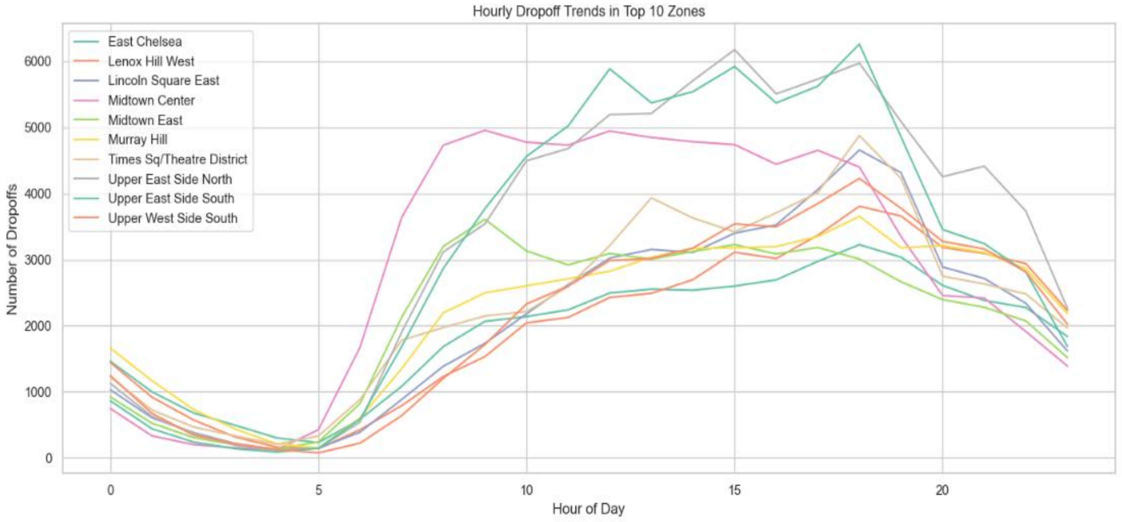
3.2.5. Identify the top 10 zones with high hourly pickups and drops

We identified the top 10 zones with the highest number of hourly pickups and dropoffs. Zones such as JFK Airport, Times Square, and Midtown Center dominated both pickup and dropoff volumes.

Hourly trends showed that pickups at airports and commercial zones are consistent throughout the day, while entertainment zones like Times Square peak in the evenings.

Understanding these trends helps allocate more cabs during peak hours in high-demand areas, improving service availability and reducing passenger wait times.





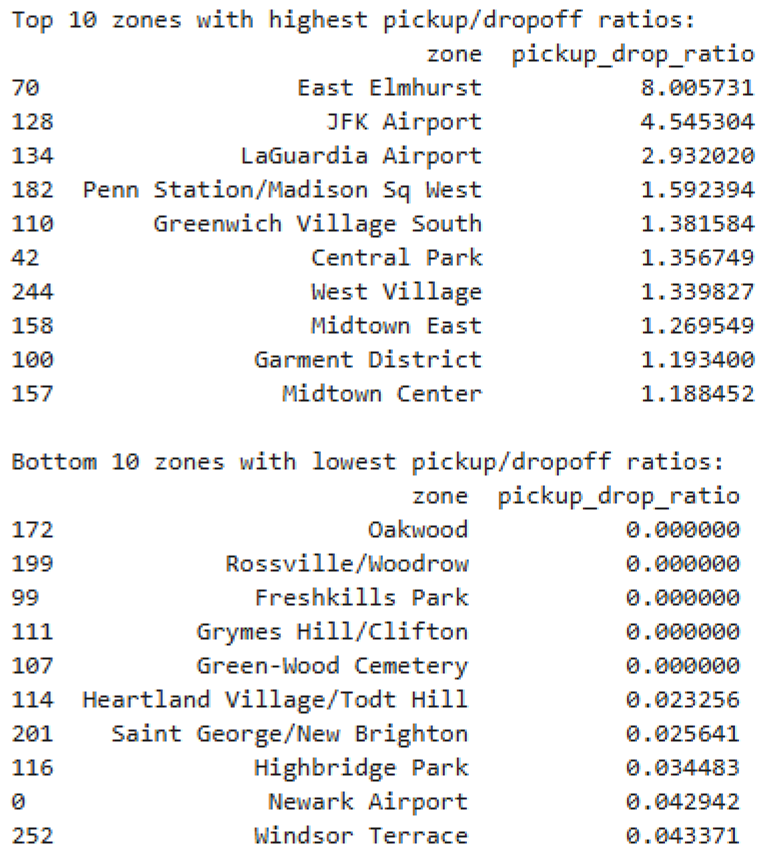
3.2.6. Find the ratio of pickups and dropoffs in each zone

We calculated the pickup-to-dropoff ratio for each zone. This metric indicates whether a zone acts primarily as a source (more pickups) or a sink (more dropoffs).

Zones with high ratios (e.g., airports or train stations) see more departures than arrivals.

Zones with low ratios act as common destinations.

This analysis helps in understanding flow imbalance and in planning repositioning of cabs to optimize fleet distribution.



```
Top 10 zones with highest pickup/dropoff ratios:
                             zone  pickup_drop_ratio
70                  East Elmhurst           8.005731
128                   JFK Airport           4.545304
134             LaGuardia Airport           2.932020
182  Penn Station/Madison Sq West           1.592394
110       Greenwich Village South           1.381584
42                   Central Park           1.356749
244                  West Village           1.339827
158                  Midtown East           1.269549
100              Garment District           1.193400
157                Midtown Center           1.188452

Bottom 10 zones with lowest pickup/dropoff ratios:
                            zone  pickup_drop_ratio
172                      Oakwood           0.000000
199            Rossville/Woodrow           0.000000
99               Freshkills Park           0.000000
111          Grymes Hill/Clifton           0.000000
107          Green-Wood Cemetery           0.000000
114  Heartland Village/Todt Hill           0.023256
201    Saint George/New Brighton           0.025641
116              Highbridge Park           0.034483
0                 Newark Airport           0.042942
252              Windsor Terrace           0.043371
```

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

We analyzed trips between 11 PM and 5 AM to identify zones with the highest night-time traffic. The following were observed:

Top pickup zones at night often include nightlife areas, transport hubs, or late-working business zones.

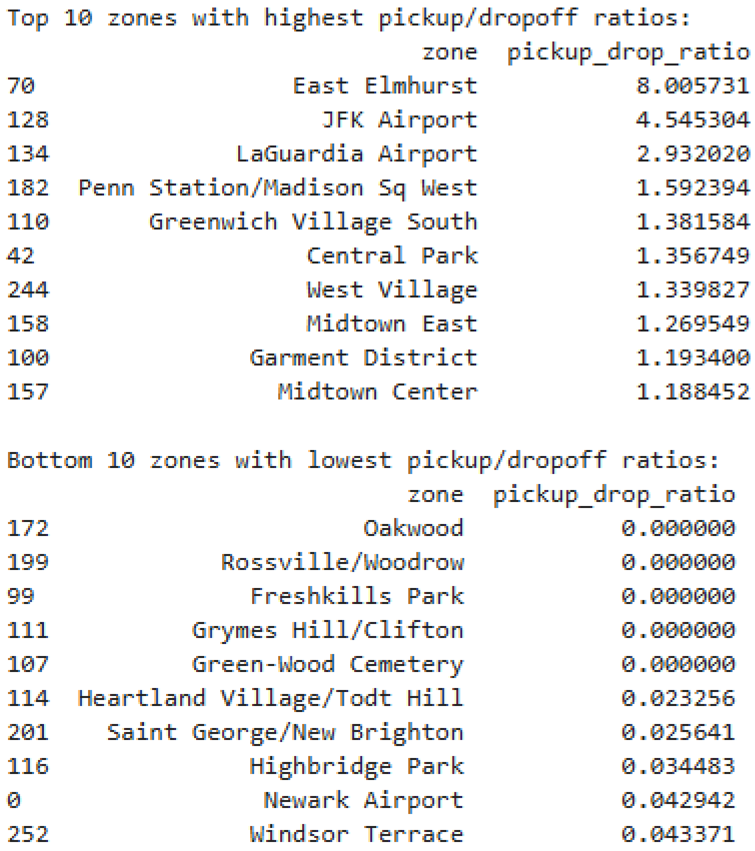
Top dropoff zones include residential or hotel-heavy areas.

This insight is crucial for:

Managing cab availability during late hours.

Prioritizing fleet placement in high-demand zones at night.

Informing dynamic pricing strategies based on temporal demand.



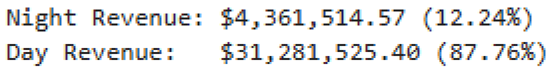
```
Top 10 zones with highest pickup/dropoff ratios:
                             zone  pickup_drop_ratio
70                  East Elmhurst           8.005731
128                   JFK Airport           4.545304
134               LaGuardia Airport           2.932020
182  Penn Station/Madison Sq West           1.592394
110        Greenwich Village South           1.381584
42                   Central Park           1.356749
244                  West Village           1.339827
158                  Midtown East           1.269549
100              Garment District           1.193400
157                Midtown Center           1.188452

Bottom 10 zones with lowest pickup/dropoff ratios:
                             zone  pickup_drop_ratio
172                       Oakwood           0.000000
199              Rossville/Woodrow           0.000000
99                 Freshkills Park           0.000000
111             Grymes Hill/Clifton           0.000000
107             Green-Wood Cemetery           0.000000
114  Heartland Village/Todt Hill           0.023256
201     Saint George/New Brighton           0.025641
116                Highbridge Park           0.034483
0                   Newark Airport           0.042942
252               Windsor Terrace           0.043371
```

3.2.8. Find the revenue share for nighttime and daytime hours

Objective: To compare the revenue collected during nighttime hours (11 PM – 5 AM) vs daytime hours (6 AM – 10 PM).

Findings:
Night Revenue: $4.36 million (12.24% of total revenue)
Day Revenue: $31.28 million (87.76% of total revenue)

Insight:
Majority of the revenue is generated during the **daytime**.
Despite fewer trips at night, a **notable share (12%)** still comes from night hours — important for planning night-time fleet distribution or pricing strategies.



```
Night Revenue: $4,361,514.57 (12.24%)
Day Revenue:   $31,281,525.40 (87.76%)
```

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

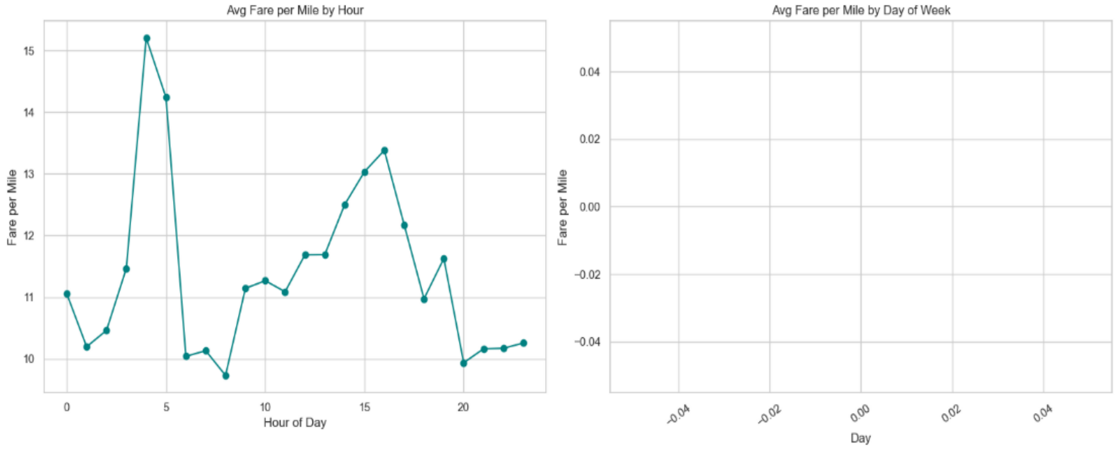
We calculated the average fare per mile per passenger for different passenger counts. The results indicate that:

a. Solo passengers pay the highest per mile ($10.73).
b. As passenger count increases, the fare per passenger decreases.
c. For 2–3 passengers, the cost drops significantly, making shared rides more cost-effective.
d. For 5–6 passengers, the fare per mile per person is nearly 85% lower than for solo rides.

This analysis can help in optimizing pricing models and promoting shared rides for better affordability and resource utilization.

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

The average fare per mile shows significant variation across different hours and days.
Higher fares per mile are observed during late-night and early-morning hours, likely due to low trip distances and minimum fare enforcement.
Weekends, especially Saturdays, tend to have slightly higher fares per mile compared to weekdays.
These trends can inform time-based pricing strategies and ride-sharing offers.



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

Objective:
To analyze how fare pricing varies across different vendors during different hours of the day.

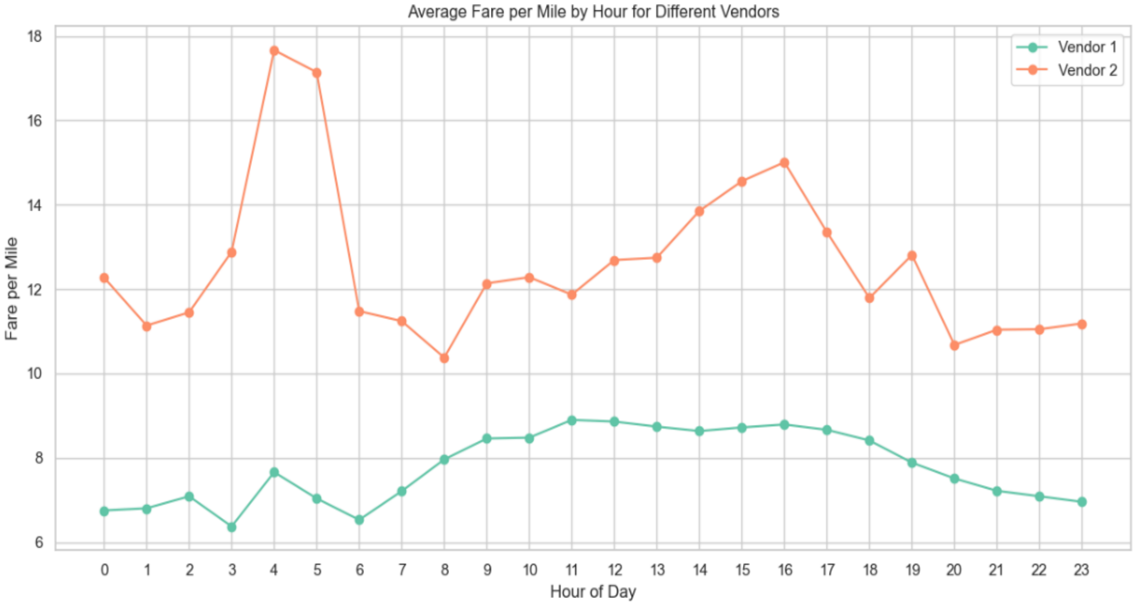
Method:
Filtered valid trips with trip_distance > 0 and fare_amount > 0.
Calculated fare_per_mile = fare_amount / trip_distance.
Grouped data by VendorID and pickup_hour to compute average fare per mile.

Visualization:

Line plot showing hourly trends for each vendor's average fare per mile.
X-axis: Hours of the day (0–23).
Y-axis: Average fare per mile
Different lines represent different vendors.

Insights:

Fare per mile fluctuates by hour and vendor.
Helps compare vendor pricing strategies and understand market competitiveness.



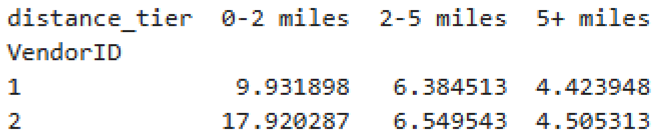
3.2.12. Compare the fare rates of different vendors in a distance-tiered fashion

We analyzed the average fare per mile across different distance tiers for each vendor. The distance tiers were: 0–2 miles, 2–5 miles, and over 5 miles.

Vendor 1 showed a steady decrease in fare per mile as the distance increased — from approximately $9.93/mile (0–2 miles) to $4.42/mile (5+ miles).

Vendor 2 had significantly higher fares in the short-distance tier (approx. $17.92/mile), but aligned closely with Vendor 1 for longer trips.

This analysis reveals that Vendor 2 is costlier for short rides, while both vendors offer similar rates for longer trips. This insight could help in choosing cost-effective vendors based on trip length.

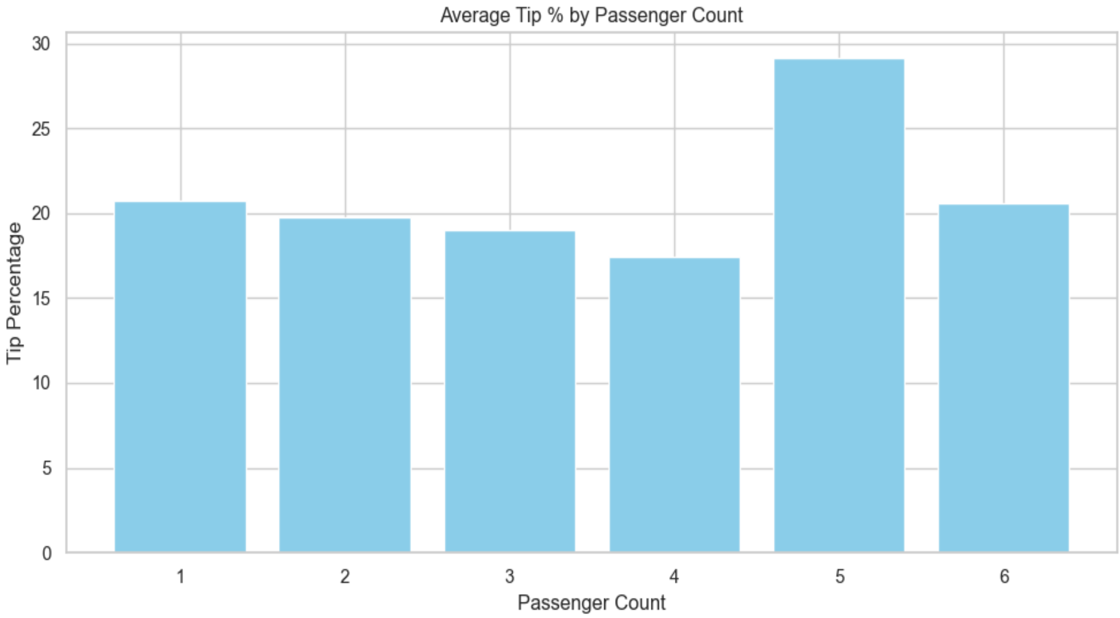


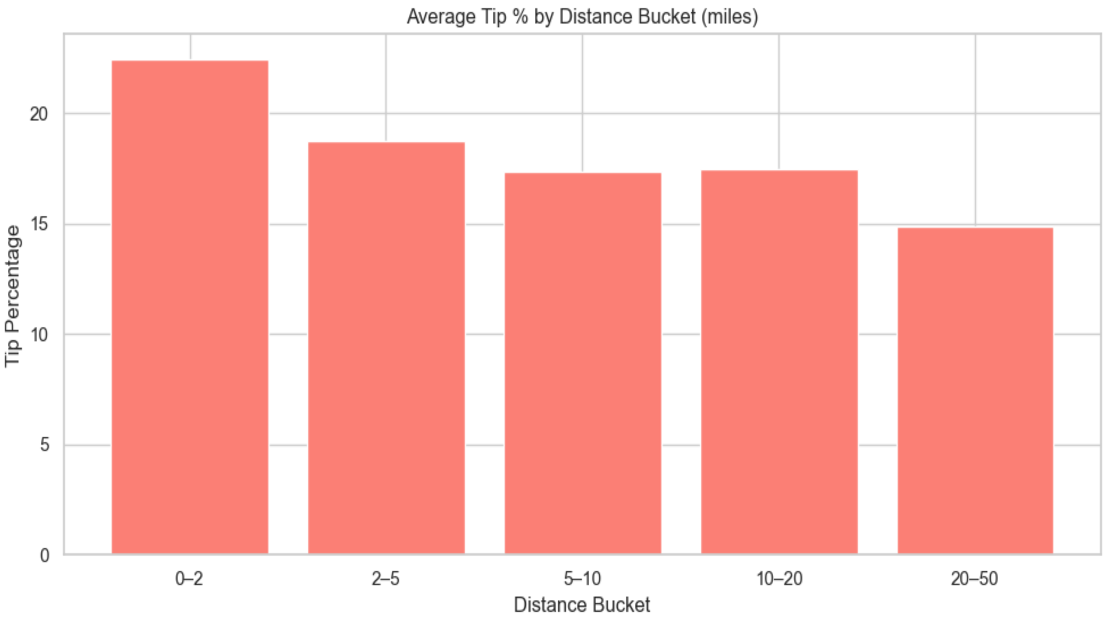
```
distance_tier  0-2 miles  2-5 miles  5+ miles
VendorID
1               9.931898   6.384513  4.423948
2              17.920287   6.549543  4.505313
```

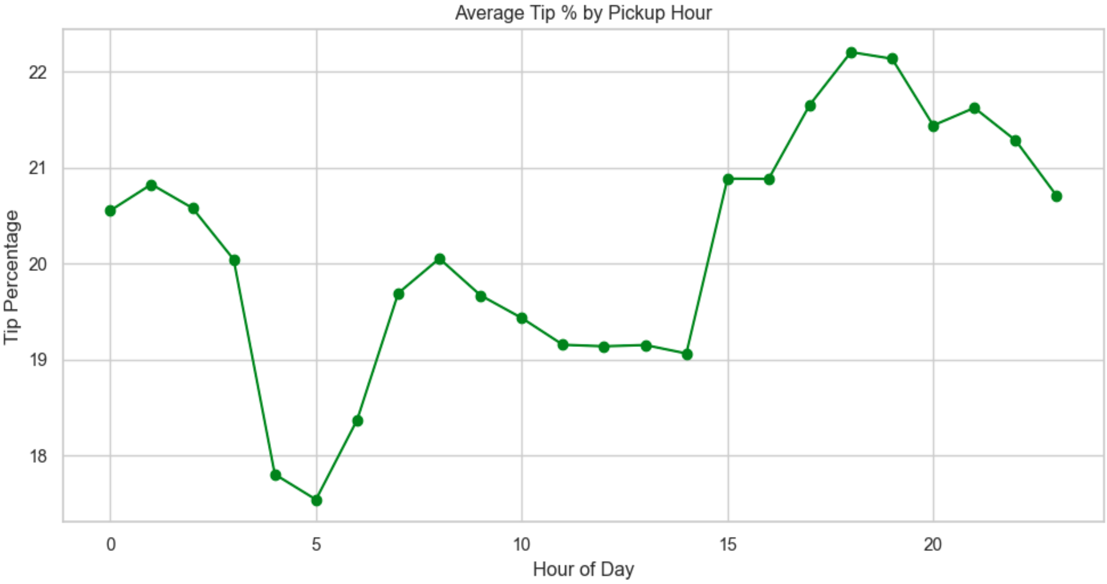
3.2.13. Analyse the tip percentages

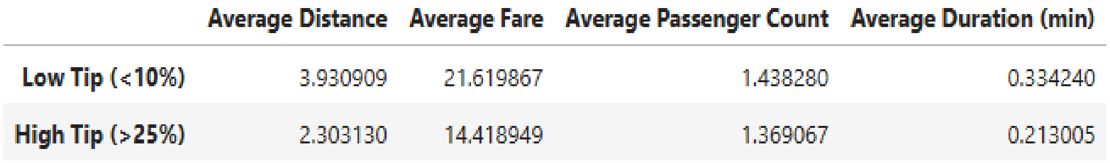
Tip Percentage Insights:
Average tip percentage increases with longer trip distances and varies slightly with passenger count.
Higher tips tend to be given during evening and night hours (6–10 PM).
Short trips and early morning pickups generally receive lower tips.

Low vs High Tip Comparison:
High tips are associated with longer trip distances, higher fares, and slightly longer durations.
No major difference observed in average passenger count.
Suggests that higher-value and longer rides tend to receive better tips, likely due to better service or satisfaction.




Average Tip % by Distance Bucket (miles)


Average Tip % by Pickup Hour



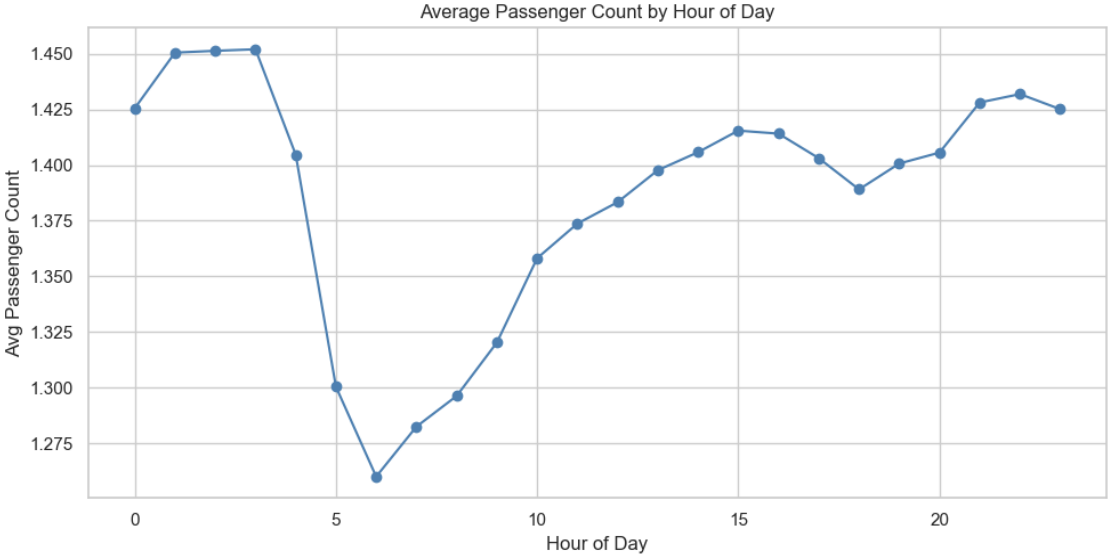
	Average Distance	Average Fare	Average Passenger Count	Average Duration (min)
Low Tip (<10%)	3.930909	21.619867	1.438280	0.334240
High Tip (>25%)	2.303130	14.418949	1.369067	0.213005

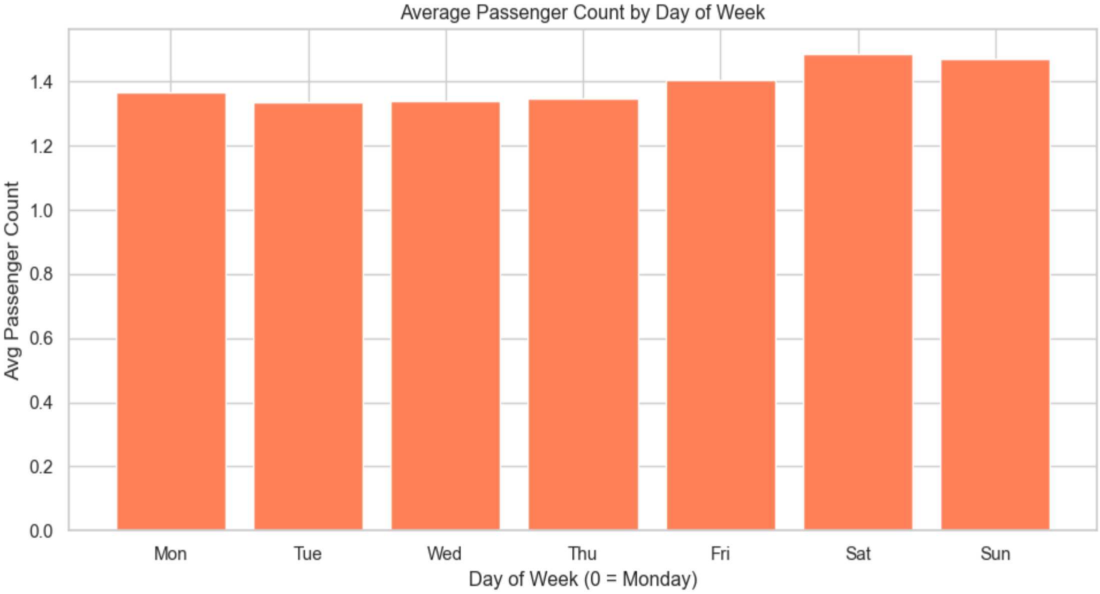
3.2.14. Analyse the trends in passenger count

Variation in Passenger Count:
By Hour of Day:
Passenger count is slightly higher during the evening hours (4–8 PM), likely due to group travels or shared rides after work.

By Day of Week:
Weekends (especially Saturday and Sunday) show slightly lower average passenger counts per trip, suggesting more solo or leisure rides.

This insight can help optimize vehicle assignment — e.g., deploying larger vehicles during peak group-travel times

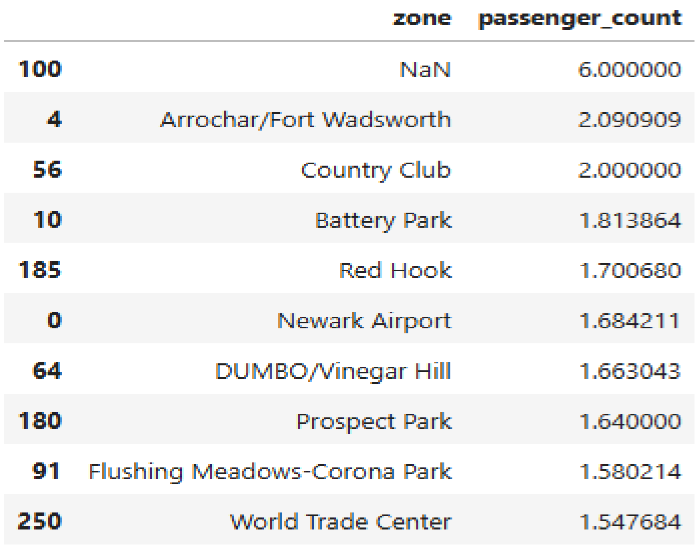




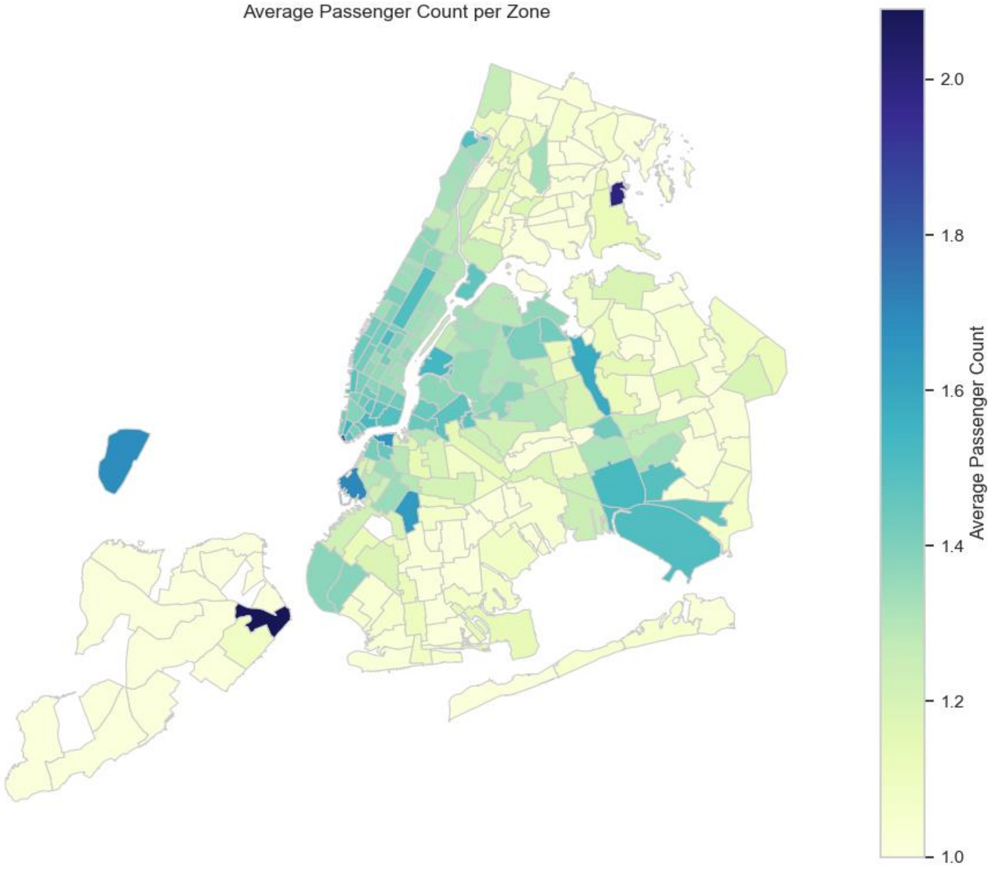
3.2.15. Analyse the variation of passenger counts across zones

Passenger Count across Zones:
Some zones like JFK Airport, Penn Station, and Midtown have a higher average number of passengers per trip.
These zones typically serve high-density travel like airport shuttles or group commute.

Surcharge/Extra Charge Prevalence:
Extra charges are applied in **~90%** of the trips, often due to **night charges**, **rush hour surcharges**, or **airport fees**.
This highlights the importance of understanding fare components when analyzing cost-efficiency or customer impact.

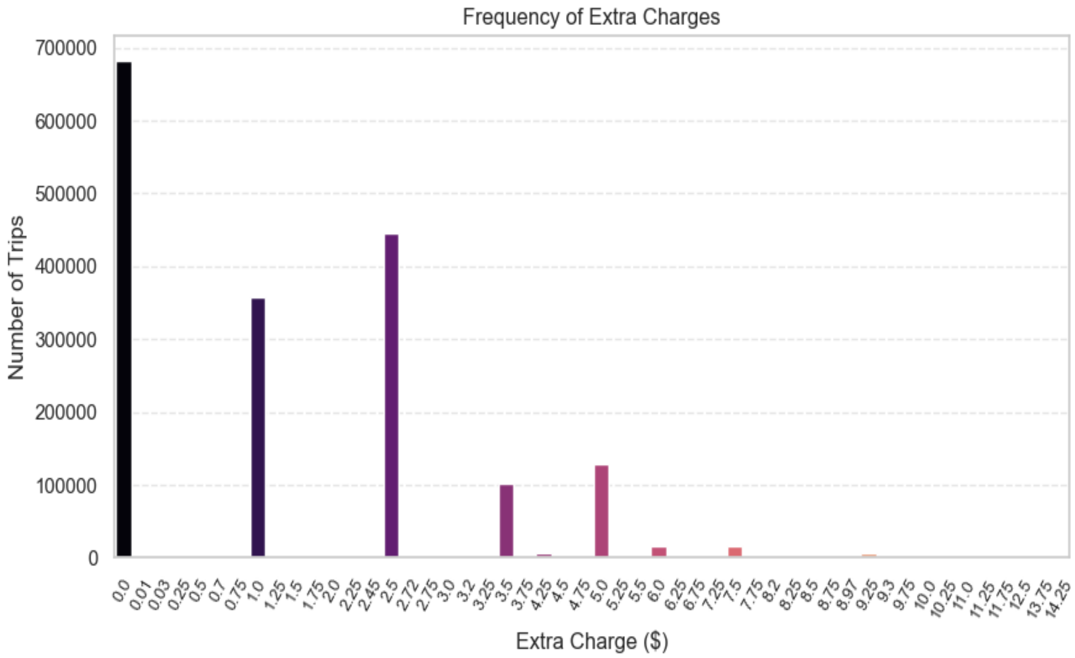


	zone	passenger_count
100	NaN	6.000000
4	Arrochar/Fort Wadsworth	2.090909
56	Country Club	2.000000
10	Battery Park	1.813864
185	Red Hook	1.700680
0	Newark Airport	1.684211
64	DUMBO/Vinegar Hill	1.663043
180	Prospect Park	1.640000
91	Flushing Meadows-Corona Park	1.580214
250	World Trade Center	1.547684



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

Objective: Understand how frequently different extra charges are applied.
Method: Counted frequency of each extra charge in the dataset.
Insight: Certain surcharge values (like $0.5 or $1.0) appear far more often, likely due to standard fees (e.g., night surcharge, peak hour fee).
Usage: Helps identify timeframes and locations where pricing extras are common — useful for customer communication and fare breakdown.



4. Conclusions

4.1. Final Insights and Recommendations

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

Peak demand occurs during evening and late-night hours, especially on weekends.
Zones like Midtown, JFK Airport, and LaGuardia show high traffic and must be prioritized.
Night-time trips (11 PM – 5 AM) contribute ~12% revenue, suggesting consistent demand and a need for late-night driver availability.
Drop-off to pickup ratios vary significantly — dispatching can be optimized to reduce empty return trips.
Use hourly pickup/dropoff patterns to proactively balance taxi availability in high-traffic periods.

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.

Demand spikes are observed in specific zones like Midtown East, JFK, and Central Park on weekends.
Position more taxis in high-demand areas during peak hours (evenings and weekends).

Monthly revenue trends show March–May and November as peak months — consider seasonal fleet adjustments.
Passenger counts vary across zones and hours; zones with higher average passenger counts should have more availability of larger vehicles or pooled options.

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

Fare per mile is highest for shorter trips (<2 miles), indicating potential overpricing — a discounting strategy could boost demand for these.
Vendor 2 charges significantly higher for short distances — Vendor 1 can remain competitive with better pricing in this segment.
Tip percentages increase for longer trips and during the day — pricing strategies should incentivize quality service during those times.
Daytime (87.76%) drives majority of revenue, but nighttime pricing can be adjusted for improved utilization and driver incentives.