Problem Statement

As an analyst at an upcoming taxi operation in NYC, we are tasked to use the 2023 taxi trip data to uncover insights that could help optimize taxi operations. The goal is to analyse patterns in the data that can inform strategic decisions to improve service efficiency, maximise revenue, and enhance passenger experience.

Tasks

We will need to perform the below mentioned tasks to naked an informed data driven strategic decision.

1. Data Loading
2. Data Cleaning
3. Exploratory Analysis: Bivariate and Multivariate
4. Creating Visualisations to Support the Analysis
5. Deriving Insights and Stating Conclusions

Let's perform each of the above mentioned task:

1. Data Loading

Data Format:

The data is stored in parquet files. There are 12 files for each of the months of 2023.

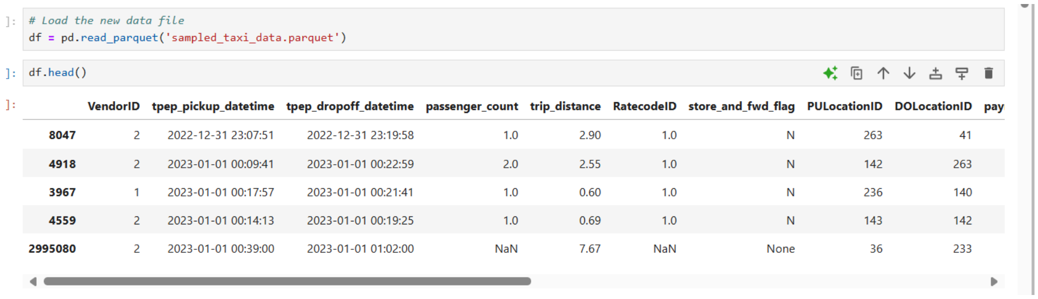
Data Sampling:

Since, for analysis we do not require so many rows of data, we will sample a fraction of data from each of the files.We will take a small percentage of entries for pickup in every hour of a date. So, for all the days in a month, we can iterate through the hours and **select 5% values** randomly from those.

After sampling the data by above logic, we will store the sampled data into a parquet file 'sampled_taxi_data.parquet'

2. Data Cleaning

We will load the sampled data from the parquet file. The loaded data looks like below:



```
]: # Load the new data file
   df = pd.read_parquet('sampled_taxi_data.parquet')
```

```
]: df.head()
```

]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID	pay
8047	2	2022-12-31 23:07:51	2022-12-31 23:19:58	1.0	2.90	1.0	N	263	41	
4918	2	2023-01-01 00:09:41	2023-01-01 00:22:59	2.0	2.55	1.0	N	142	263	
3967	1	2023-01-01 00:17:57	2023-01-01 00:21:41	1.0	0.60	1.0	N	236	140	
4559	2	2023-01-01 00:14:13	2023-01-01 00:19:25	1.0	0.69	1.0	N	143	142	
2995080	2	2023-01-01 00:39:00	2023-01-01 01:02:00	NaN	7.67	NaN	None	36	233	

We perform below data cleaning activities:

Fixing columns:

a. Fix index and drop unnecessary columns like duplicate airport_fee columns
b. Fix columns with negative (monetary) values

Handling missing values:

a. Identify columns with missing values
b. Handling missing values in columns like RatecodeID, passenger_count, congestion_surcharge, etc.

Handling outliers:

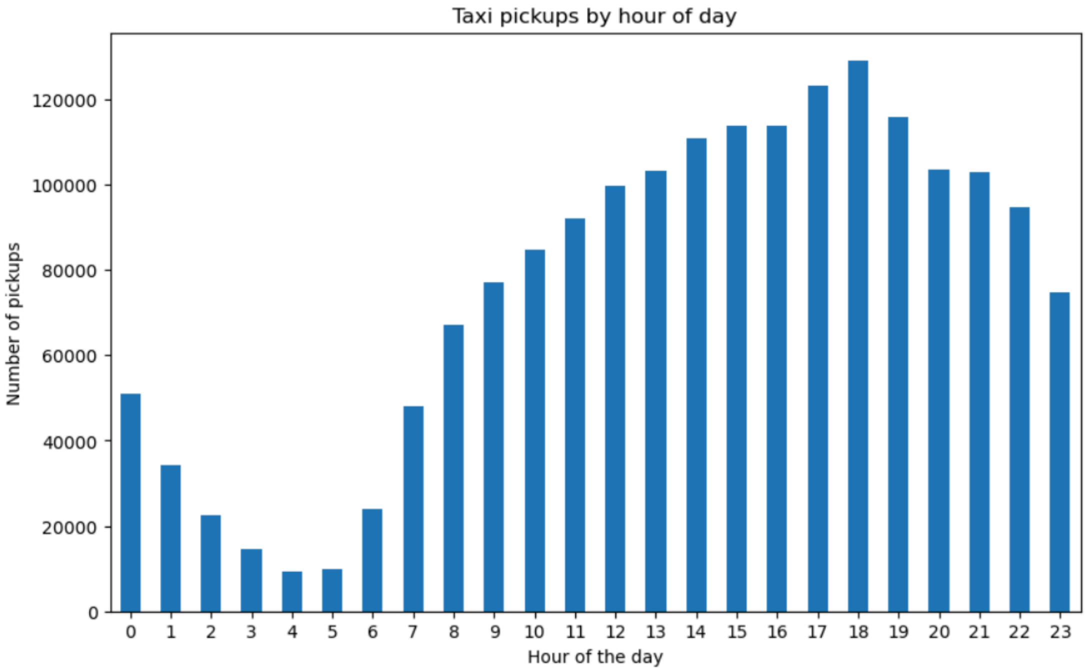
a. Check for potential outliers in data by describing data
b. Analyse outliers in certain columns by boxplot and retrieving IQR.
c. Fix outliers in some of the columns by dropping outliers rows

3. Exploratory Analysis: Bivariate and Multivariate & Visualizations

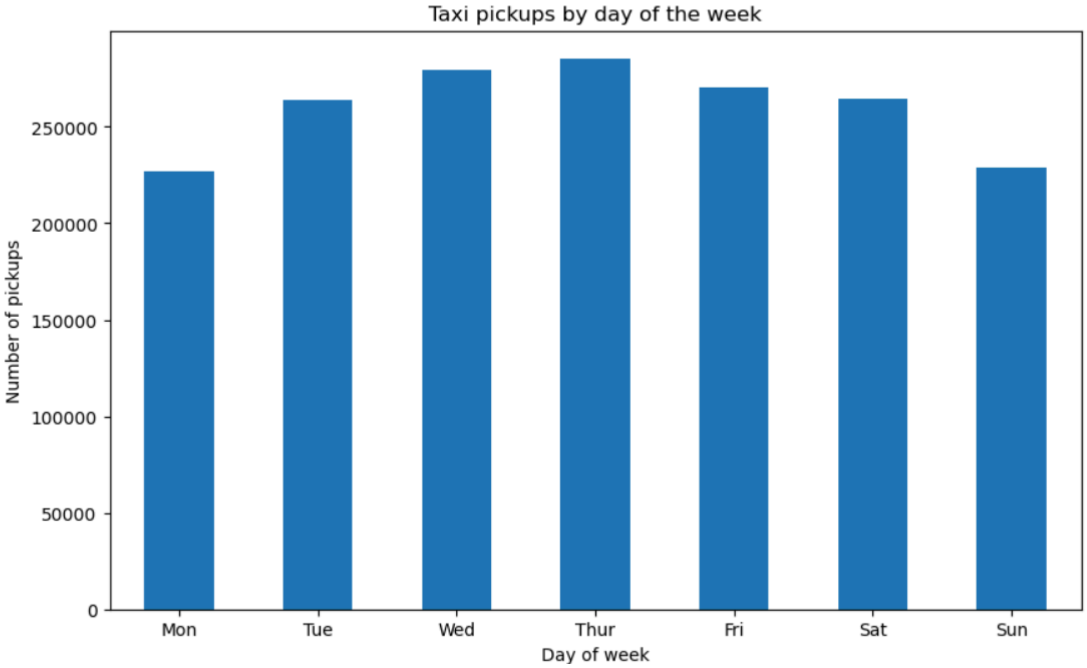
Temporal Analysis:

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

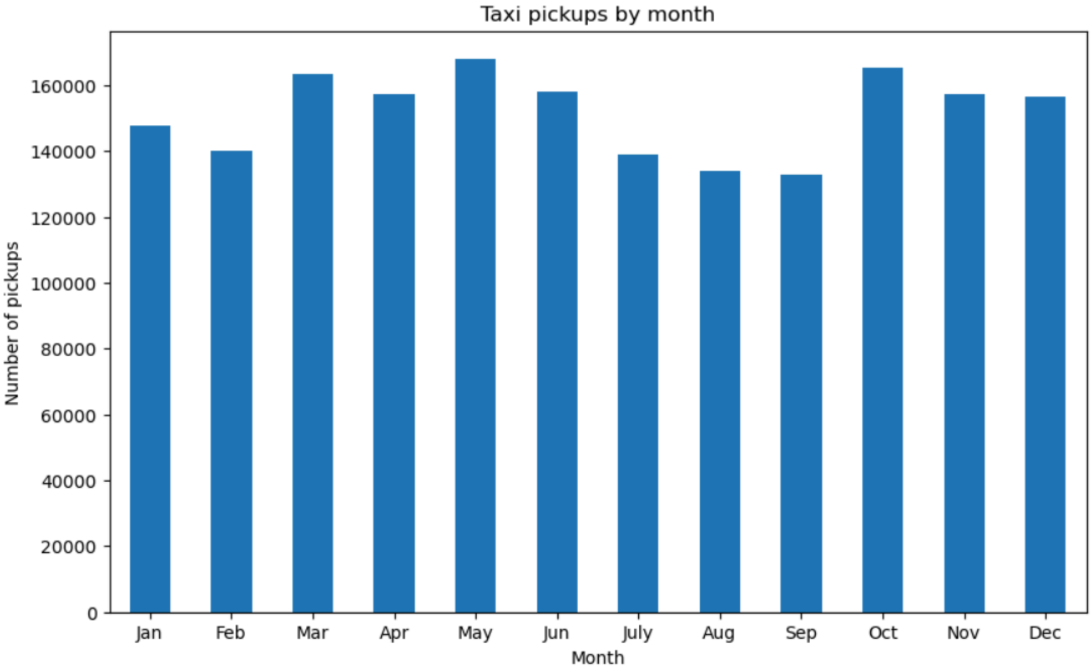
By Hour:



By Week:



By Month:



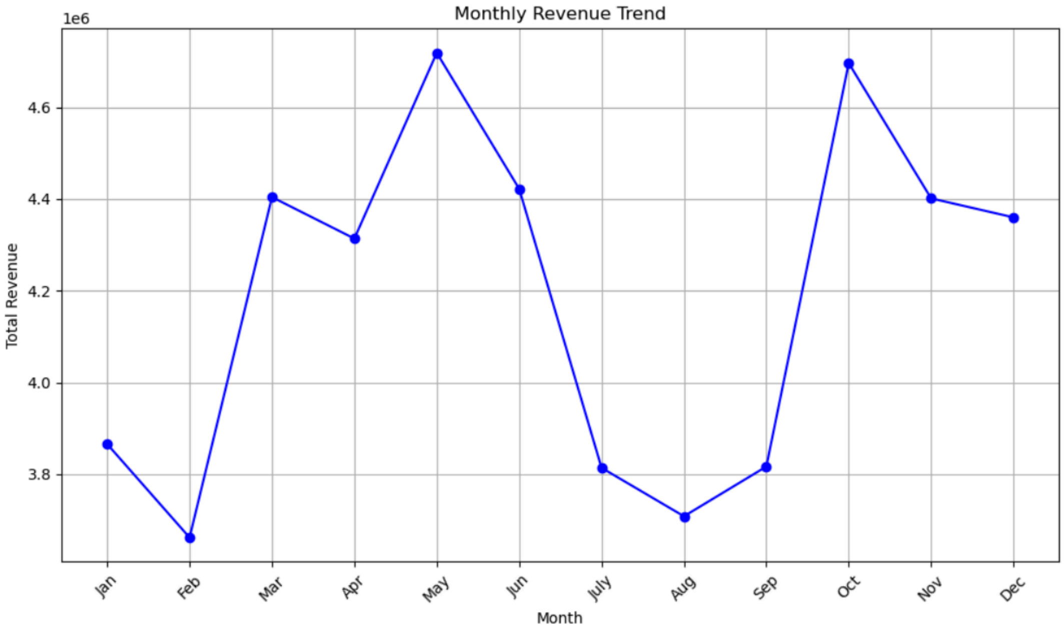
Observations:

- Taxi pickups are more during 5pm to 8pm with 6pm as the peak hour.
- Taxi pickups are more between Tue to Saturday with Thursday being the peak day.

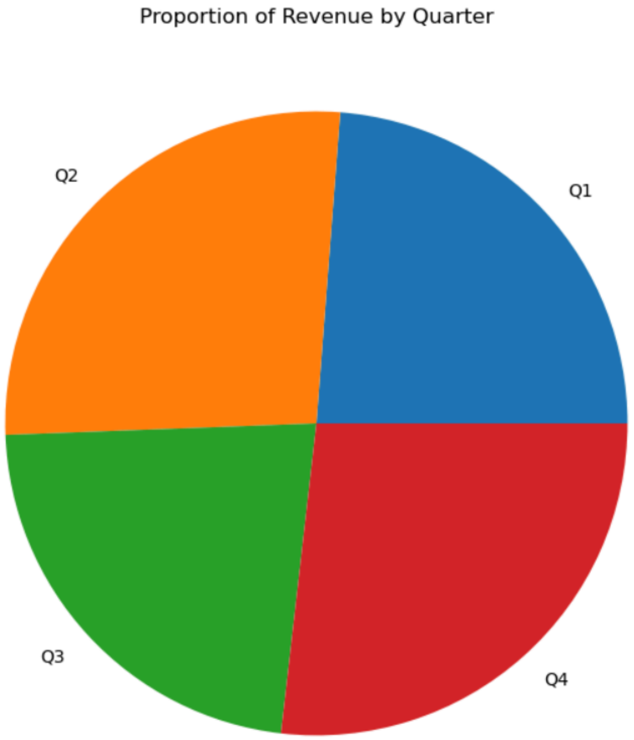
- Taxi pickups are more in the period March - June and Oct to Dec. May month being the peak month and Aug and Sept having lowest pickups.

Financial Analysis:

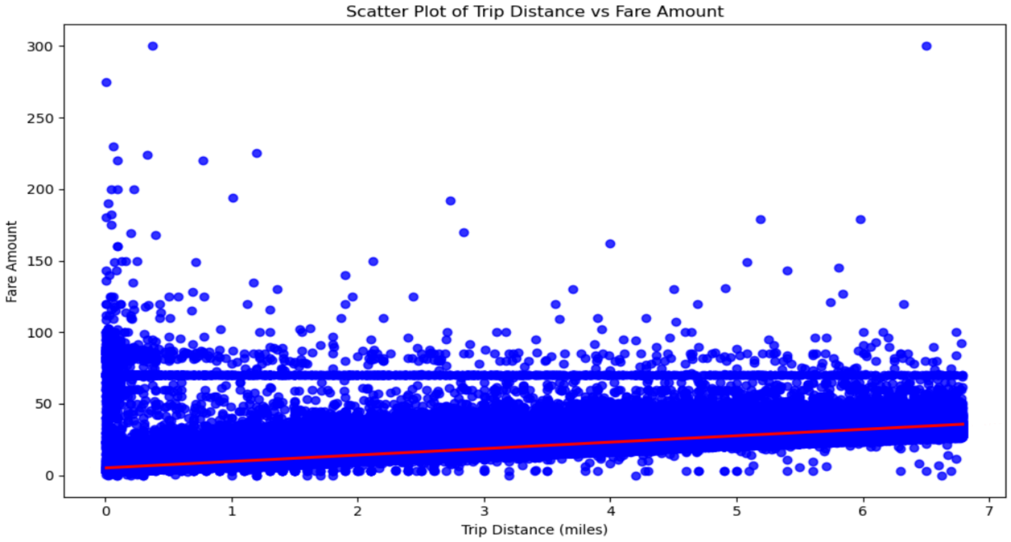
a. Monthly revenue trend:



b. Proportion of revenue by quarter:

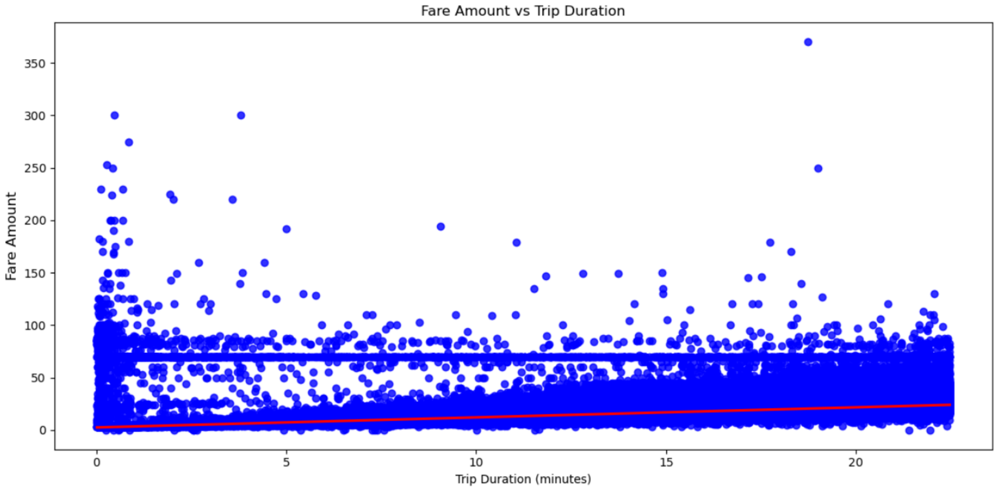


c. **Relation between trip distance and fare amount:**



Correlation coefficient between trip_distance and fare_amount: 0.80

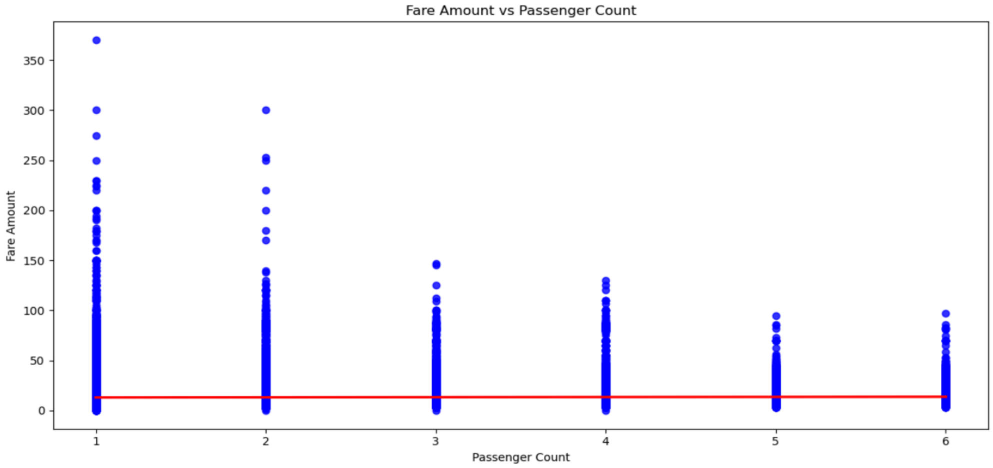
d. **Trip duration and fare amount:**



```
Correlation coefficient between fare_amount and trip_duration: 0.72
```

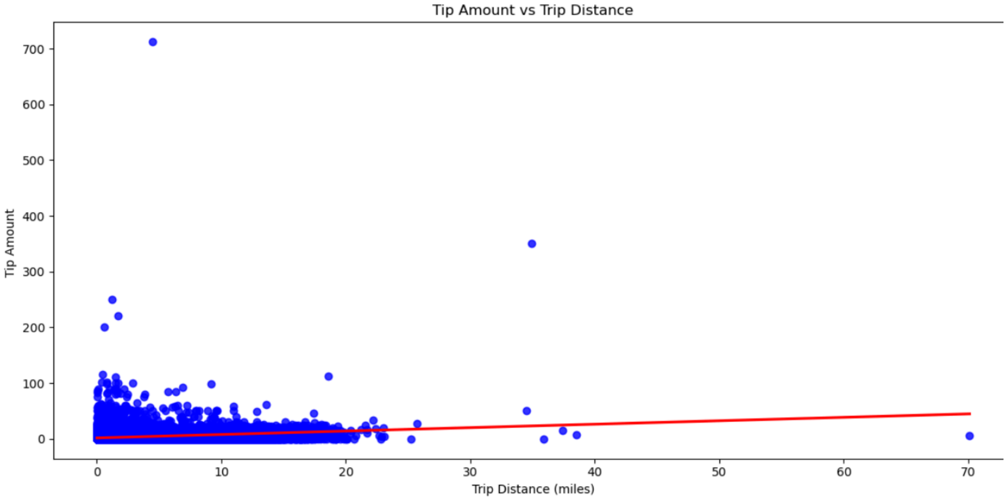
Correlation coefficient between fare_amount and trip_duration: 0.72

e. **Passenger count and fare amount:**



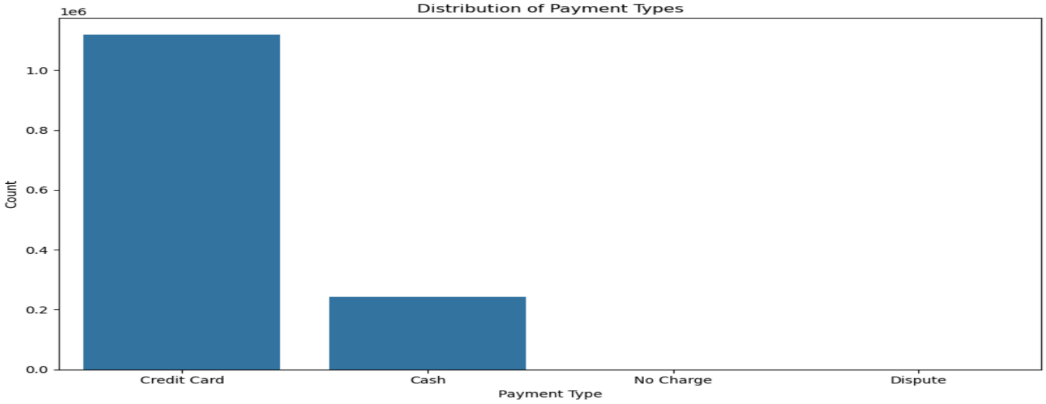
Correlation coefficient between fare_amount and passenger_count: 0.02

f. Trip distance and tip amount



Correlation coefficient between Tip Amount and Trip Distance: 0.43

g. Distribution of payment types



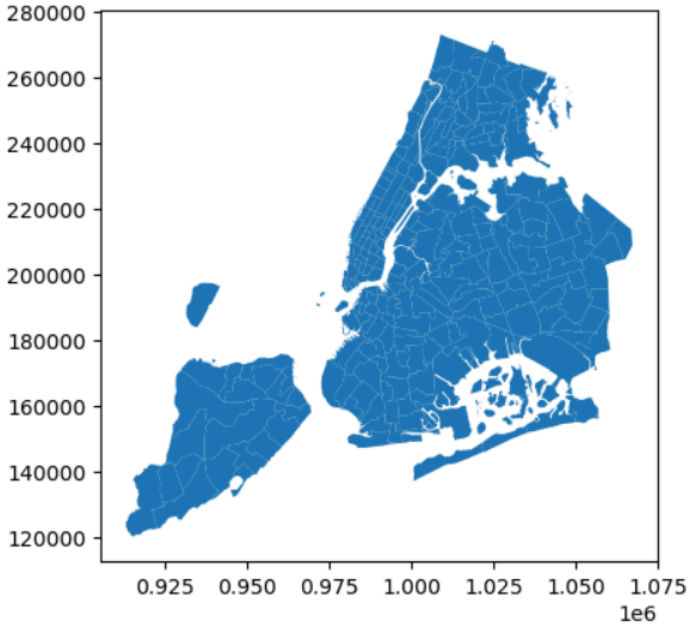
Observations:

- May and October generates highest monthly revenue.Revenue is lowest in Feb and August
- Quarter Q2 and Q4 generate the highest revenue. Q3 has the lowest revenue.
- Trip distance and fare amount are positively correlated. Fare amount increases with the trip distance.
- Trip duration and fare amount are positively correlated. Fare amount increases with the trip duration.
- Fare amount and passenger count is very slightly correlated and the relation is not significant.
- Trip distance and tip amount are moderately positively correlated. Tip amount increases with trip distance.
- Majority of payments are through Credit cards compared to Cash.

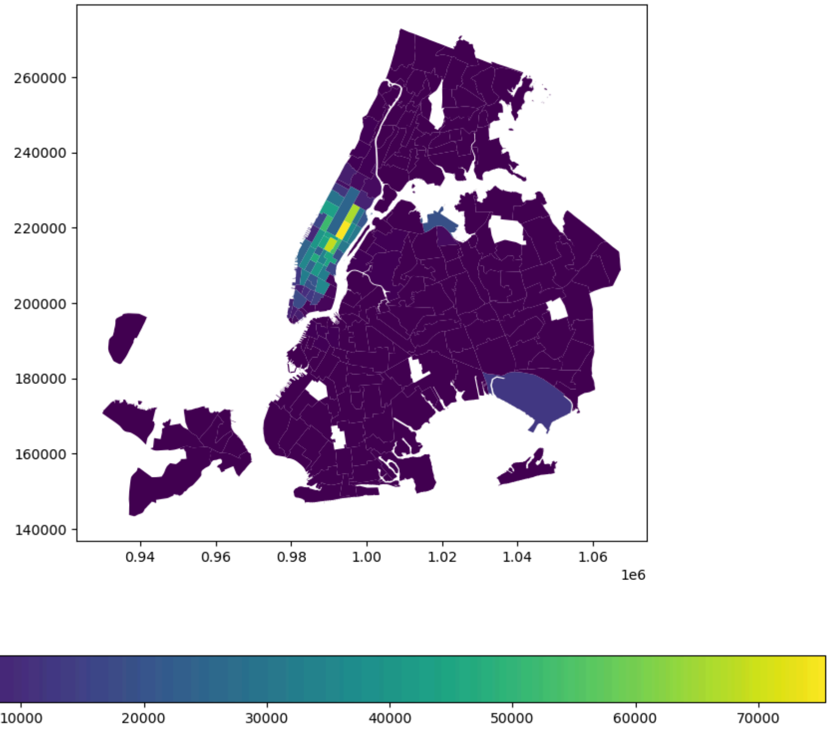
Geographical Analysis:

For this we loaded the taxi_zones.shp file from the taxi_zones folder. We used the GeoPandas library.

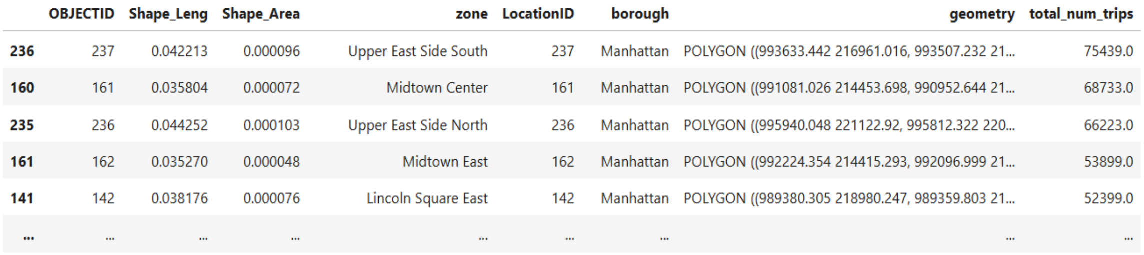
a. **Zone plot after taxi_zones files loading:**



b. Zone wise trips:



c. Zones with highest number of trips



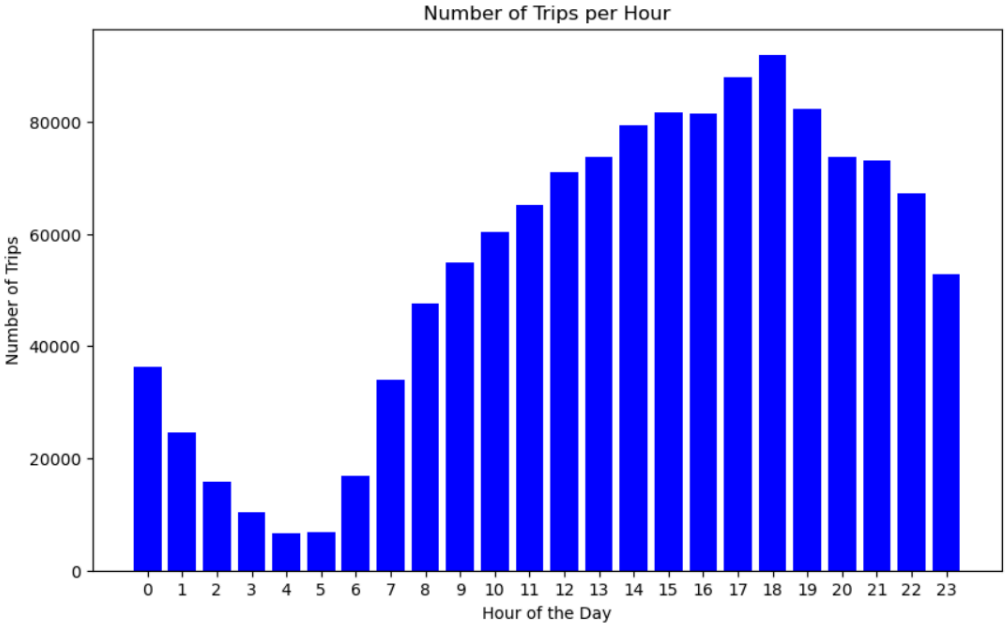
	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry	total_num_trips
236	237	0.042213	0.000096	Upper East Side South	237	Manhattan	POLYGON ((993633.442 216961.016, 993507.232 21...	75439.0
160	161	0.035804	0.000072	Midtown Center	161	Manhattan	POLYGON ((991081.026 214453.698, 990952.644 21...	68733.0
235	236	0.044252	0.000103	Upper East Side North	236	Manhattan	POLYGON ((995940.048 221122.92, 995812.322 220...	66223.0
161	162	0.035270	0.000048	Midtown East	162	Manhattan	POLYGON ((992224.354 214415.293, 992096.999 21...	53899.0
141	142	0.038176	0.000076	Lincoln Square East	142	Manhattan	POLYGON ((989380.305 218980.247, 989359.803 21...	52399.0
...	...	...	...	...	...	...	...	...

<u>Observations</u>:

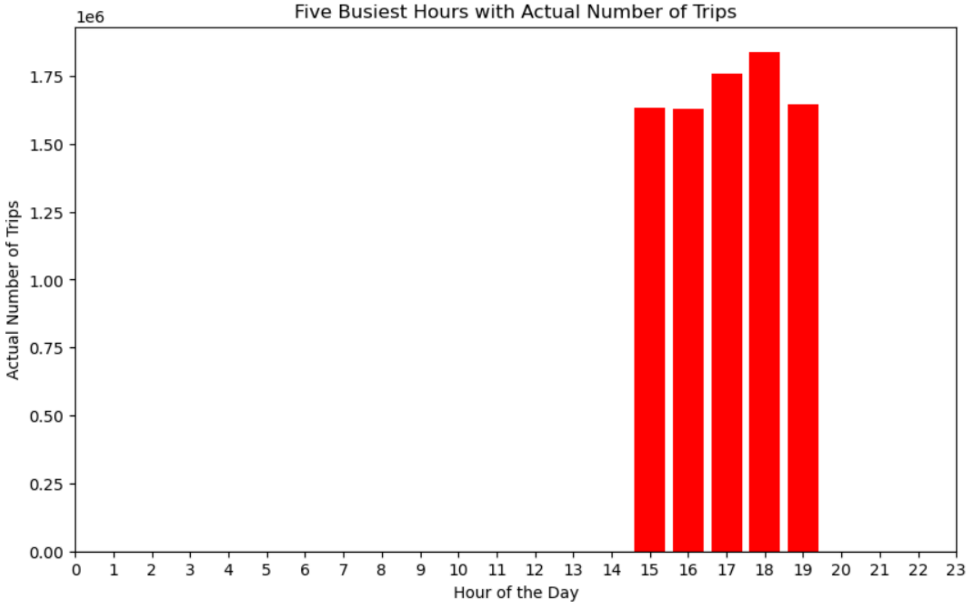
- ○ Zone ‘Upper East Side South’ has highest number of trips followed by ‘Midtown Center’ and ‘Upper East Side North’ and all of them belong to borough ‘Manhattan’

4. Detailed EDA: Insights and Strategies

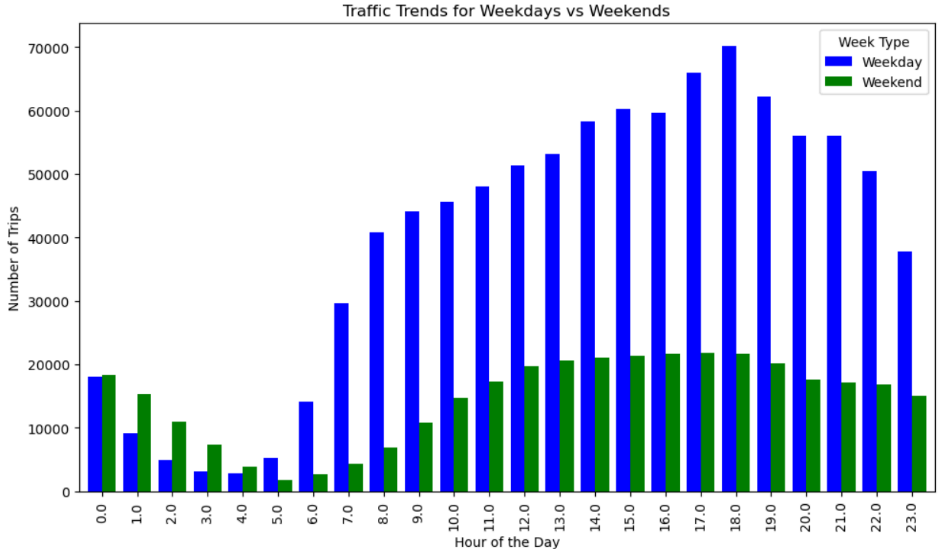
a. Number of trips per hour of day:



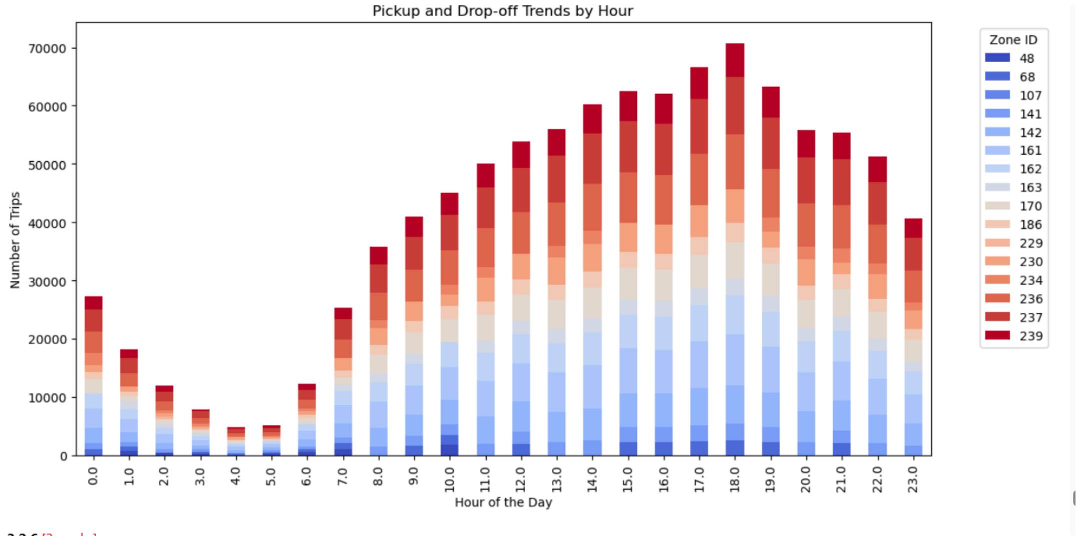
b. Actual number of trips in 5 busiest hour:



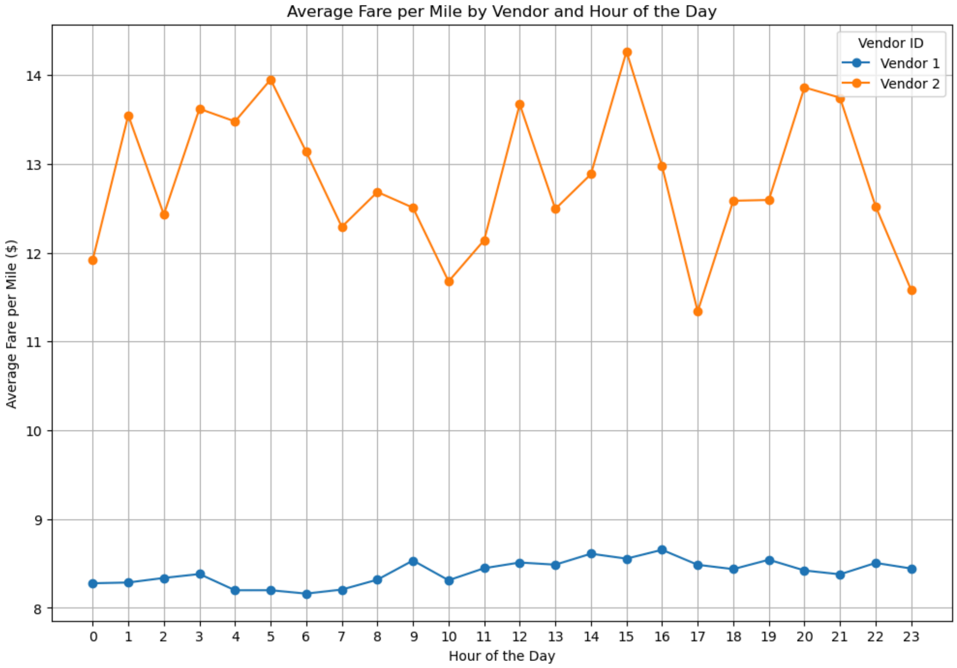
c. Traffic trends for weekday and weekend



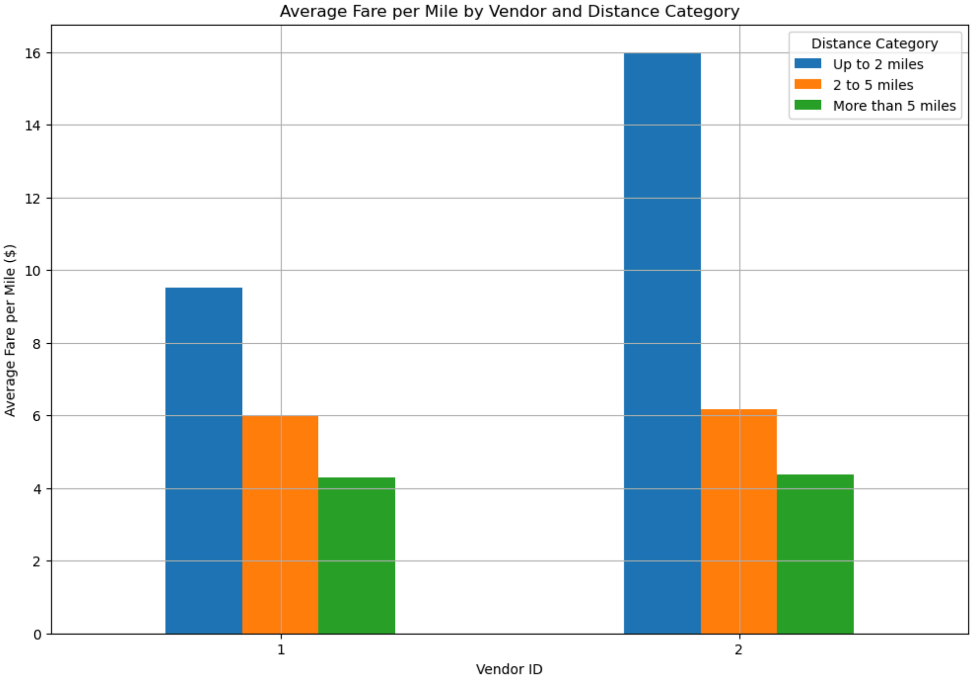
d. Pickup and drop trends by hour:



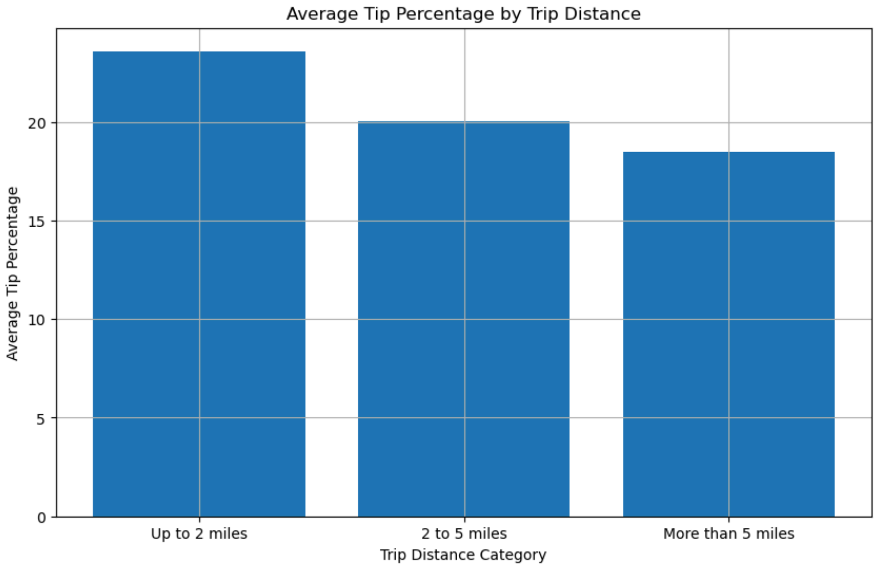
e. Average fare per hour per vendor



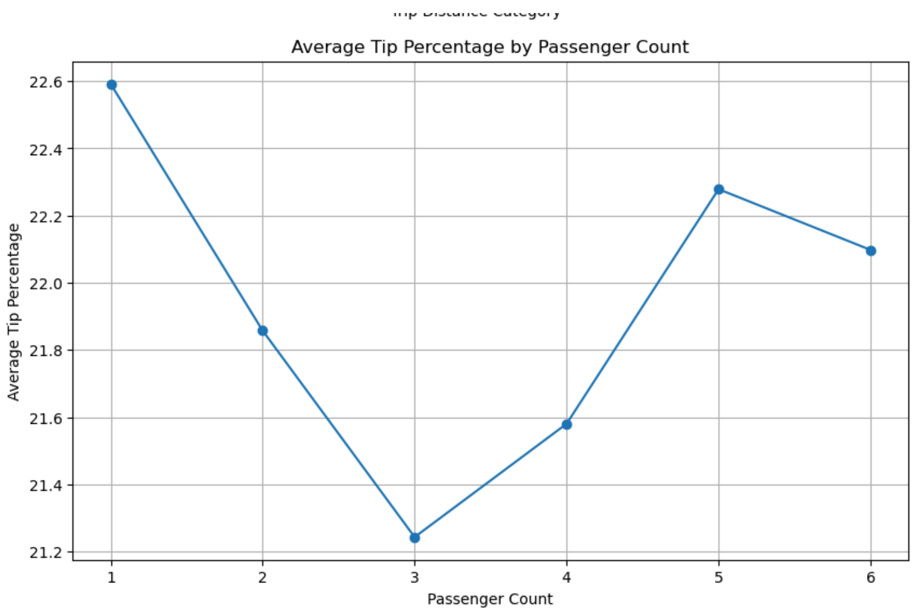
f. Average fare per mile per vendor



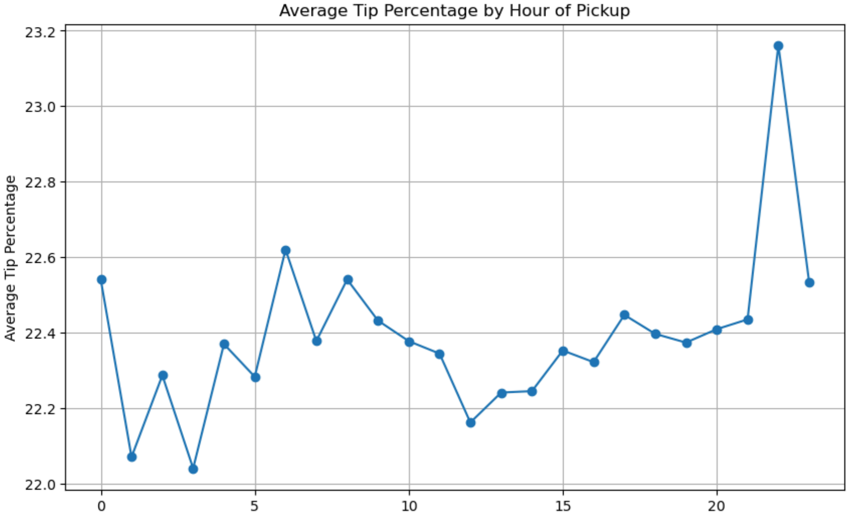
g. Average tip percentage by distance



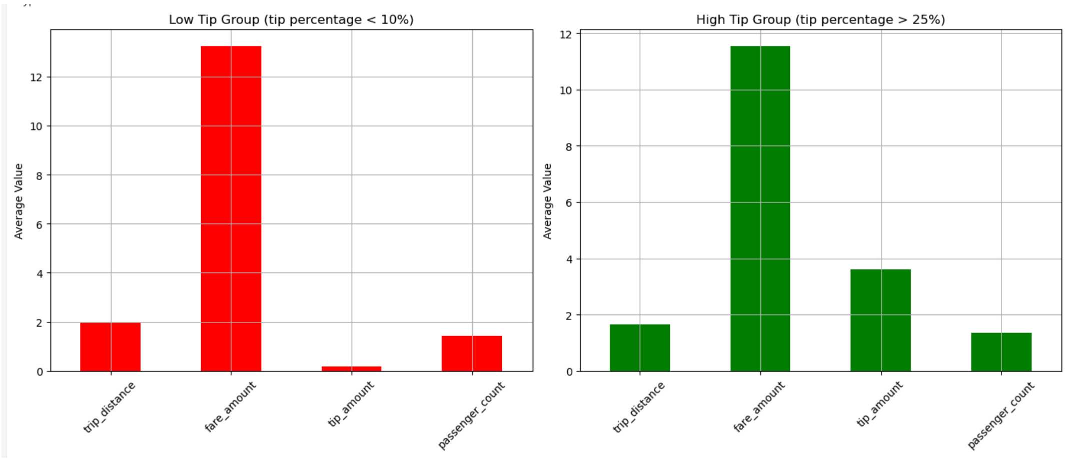
h. Average tip percentage by passenger count



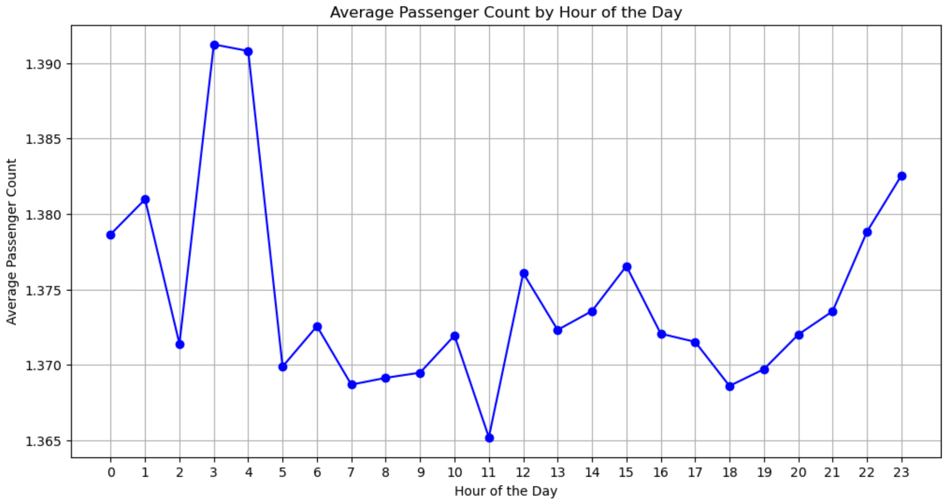
J. Average tip percentage by hour of pickup



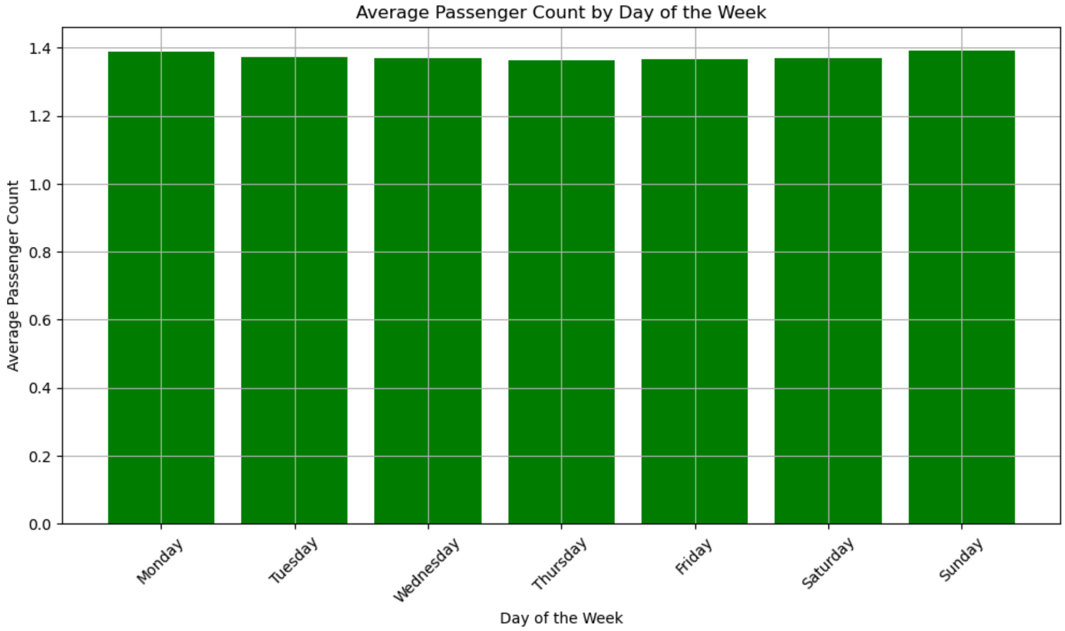
K. Low and High tip group percentage



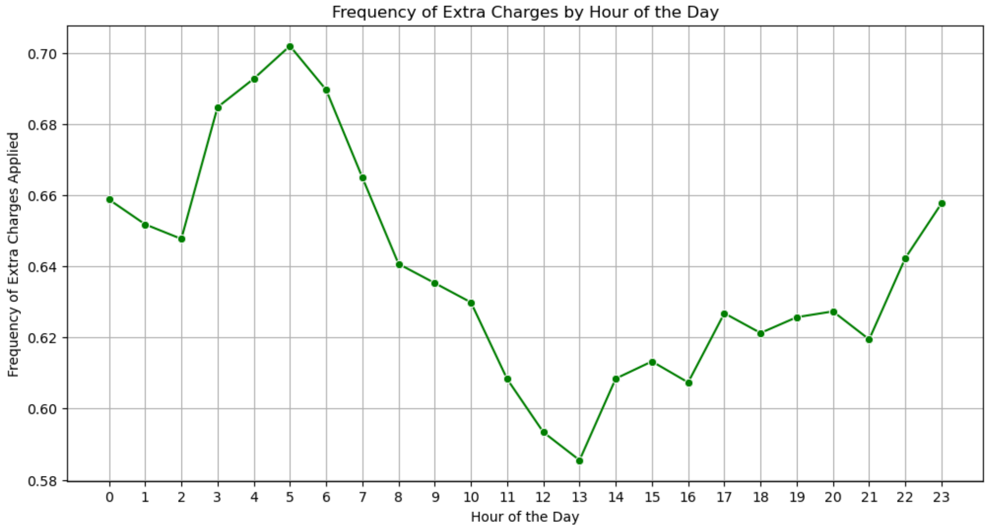
L. Average passenger count by hour of the day



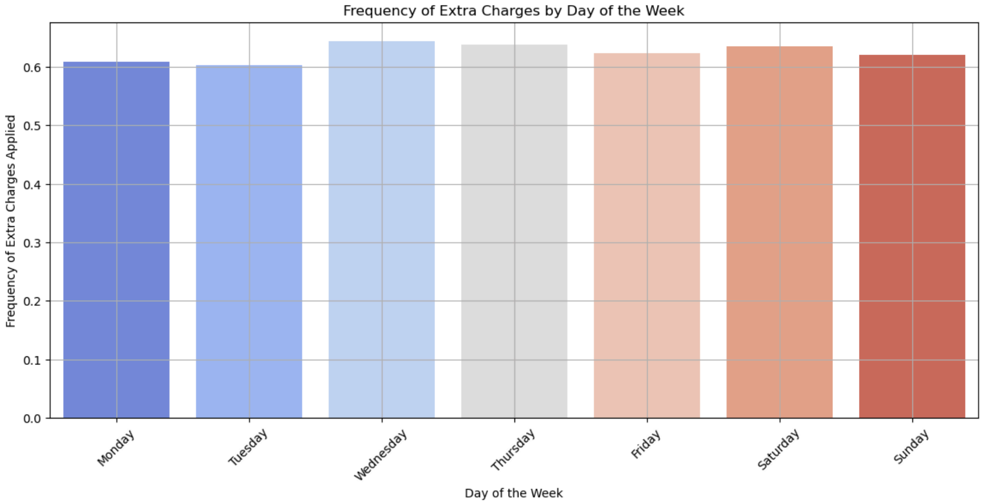
M. Average passenger count by day of week



N. Frequency of extra charges by hour of the day



O. Frequency of extra charges by day of week



<u>Observations</u>:

- The busiest hour is 6pm with 91927 trips.
- Busiest hours are 3pm to 7pm with highest at 6pm.
- Early morning hours have lesser trips
- Weekday has higher demands than weekends
- For weekday, the rush hours are 3pm to 7pm
- FOr weekends, demand is more from 12pm to 8pm
- Zone ID 239, 237, 236, 234 has highest number of trips during peak hours (i.e 3pm to 8pm)
- For night hours (11pm to 5am), zone id 237, 161,236,162,142 are top 5 zones with higher pickups
- Revenue share of daytime is 88.58% while nightime revenue share is 11.42%

- With increase in passenger count, fare per mile per passenger increases. So, single passenger trip is highest with 11
- Day of week does not have any significant trend for average fare per mile
- Hour of day does not have any significant trend for average fare per mile
- Fare per mile for Vendor 2 is higher than for Vendor 1
- Tip percentage is higher for trip distance upto 2 miles and decreases gradually post that.
- Tip percentage increase post 4pm and is highest after 8pm
- Night time has more extra charges applied to the fare. It is more from 11pm to 6am with highest during 5am.

5. Final Insights and Recommendations

Detailed Observations:

- Taxi pickups are more during 5pm to 8pm with 6pm as the peak hour.
- Taxi pickups are more between Tue to Saturday with Thursday being the peak day.
- Taxi pickups are more in the period March - June and Oct to Dec. May month being the peak month and Aug and Sept having lowest pickups.
- May and October generates highest monthly revenue.Revenue is lowest in Feb and August
- Quarter Q2 and Q4 generate the highest revenue. Q3 has the lowest revenue.
- Trip distance and fare amount are positively correlated. Fare amount increases with the trip distance.
- Trip duration and fare amount are positively correlated. Fare amount increases with the trip duration.
- Fare amount and passenger count is very slightly correlated and the relation is not significant.
- Trip distance and tip amount are moderately positively correlated. Tip amount increases with trip distance.
- Majority of payments are through Credit cards compared to Cash.
- Zone ‘Upper East Side South’ has highest number of trips followed by ‘Midtown Center’ and ‘Upper East Side North’ and all of them belong to borough ‘Manhattan’
- The busiest hour is 6pm with 91927 trips.
- Busiest hours are 3pm to 7pm with highest at 6pm.
- Early morning hours have lesser trips
- Weekday has higher demands than weekends
- For weekday, the rush hours are 3pm to 7pm
- FOr weekends, demand is more from 12pm to 8pm
- Zone ID 239, 237, 236, 234 has highest number of trips during peak hours (i.e 3pm to 8pm)
- For night hours (11pm to 5am), zone id 237, 161,236,162,142 are top 5 zones with higher pickups
- Revenue share of daytime is 88.58% while nighttime revenue share is 11.42%
- With increase in passenger count, fare per mile per passenger increases. So, single passenger trip is highest with 11

- Day of week does not have any significant trend for average fare per mile
- Hour of day does not have any significant trend for average fare per mile
- Fare per mile for Vendor 2 is higher than for Vendor 1
- Tip percentage is higher for trip distance upto 2 miles and decreases gradually post that.
- Tip percentage increase post 4pm and is highest after 8pm
- Night time has more extra charges applied to the fare. It is more from 11pm to 6am with highest during 5am.

Recommendations:

- It should be ensured that there is availability during peak hours and weekdays.
- Promotional offers need to be rolled out during weekends, off hours to increase demand and revenue.
- From a revenue perspective, it has been observed that May and October generate the highest monthly revenue. Revenue is lowest in Feb and August Quarter Q2 and Q4 generate the highest revenue. Q3 has the lowest revenue. Revenue share of daytime is 88.58% while nighttime revenue share is 11.42% . Hence,for the months and quarters with lowest revenue, there should be more focus to increase revenue.