New York City Yellow Taxi Data

Problem Statement :

As an analyst at an upcoming taxi operation in NYC, you are tasked to use the 2023 taxi trip data to uncover insights that could help optimise 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 :

You need to perform the following steps for successfully completing this assignment:

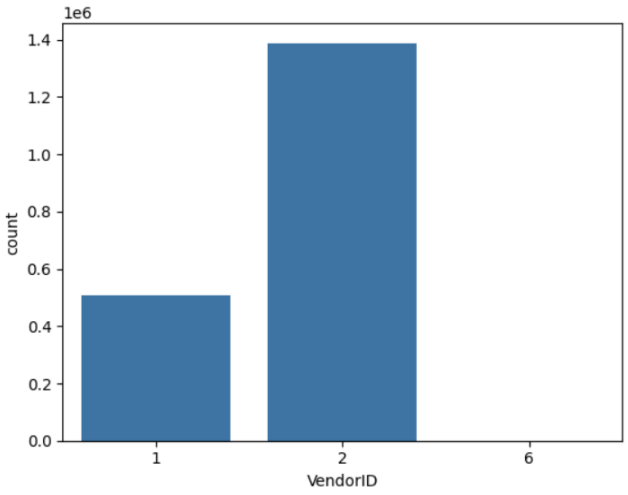
1. Data Loading
2. Data Cleaning
3. Exploratory Analysis: Bivariate and Multivariate
4. Creating Visualisations to Support the Analysis
5. Deriving Insights and Stating Conclusions

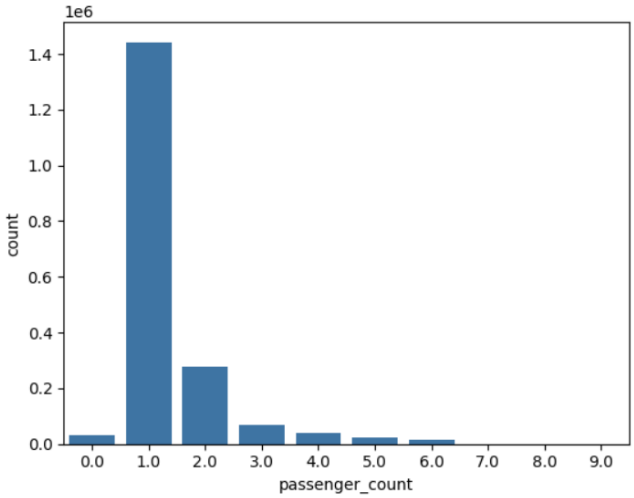
Performance Explanation :

Firstly I have imported the required libraries to the google colab notebook. And then I proceed with understanding versions of libraries. Later I loaded the dataset named ‘2023-1.parquet’ into the colab code cell. I understood that there 12 similar cells like January file. Next I understood the concept of Sampling the data. In sampling the data I took 5 % of data from each and eavery file from a total of 12 months files. I took every hours 5 % data from each day in a month. Similarly done for all 12 months. These files are present in “parquet” file format. So after importing 5 % of sample data I converted it to csv format as I am comfortable in csv format.

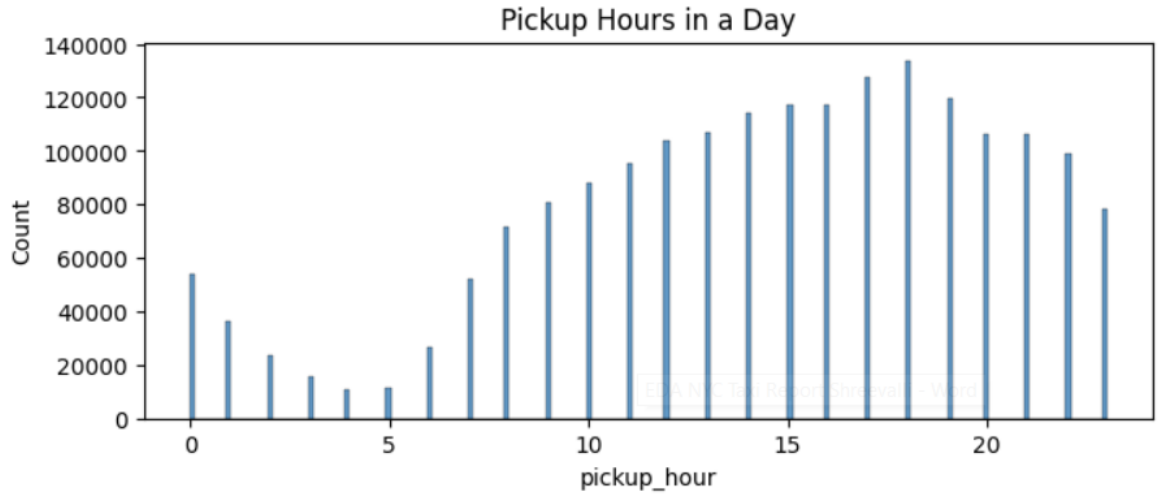
Now I had a glance on data about info, describe, shape, columns for knowing the column names, number of rows and columns, summarization of the dataset. Later I proceed with Data Cleaning . I have observed that there is column named with “Unnamed 0” came because it is storing previous data index or may be due to reading data set 2 times. So I dropped that column as it is unrequired. Also I saw there are 2 column names with “Airport fees” with Caps and smalls in letter difference. So I fixed it. Null values if present in Categorical columns then those are fixed with mode. If null values are present in Numerical columns then those are fixed with median (my preference is median. Even you can choose mean as well depending on your data). Later on proceeded with monetary value analysis.

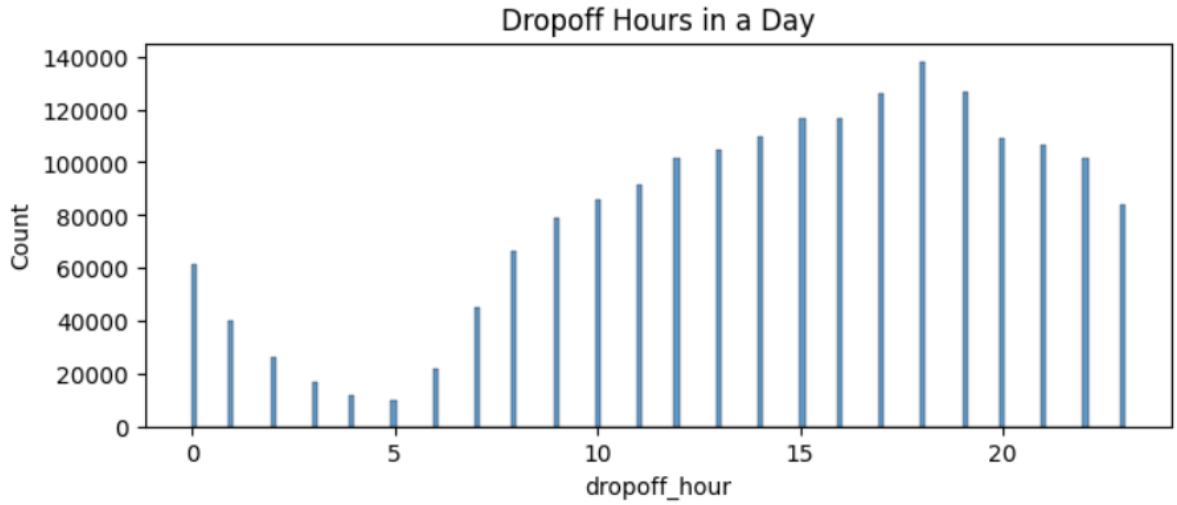
Now I reached Handling Missing values. After checking missing values I found that there are few missing values. These columns include ‘passenger_count’, ’Ratecode_ID’, ‘congestion_surcharge’, ‘store_and_fwd_flag’. Next I checked outliers using boxplot, countplot. Below mentioned are few visualizations of the same :





Now as we know where where outliers are present , I concentrated to remove those outliers by using “Capping & Flooring”. Next continued with further Analysis on Dataset. Next created separate columns for Pickup Time and Pickup Date. Similarly done for Dropoff also. Below chart represents Pickup hours in a day.

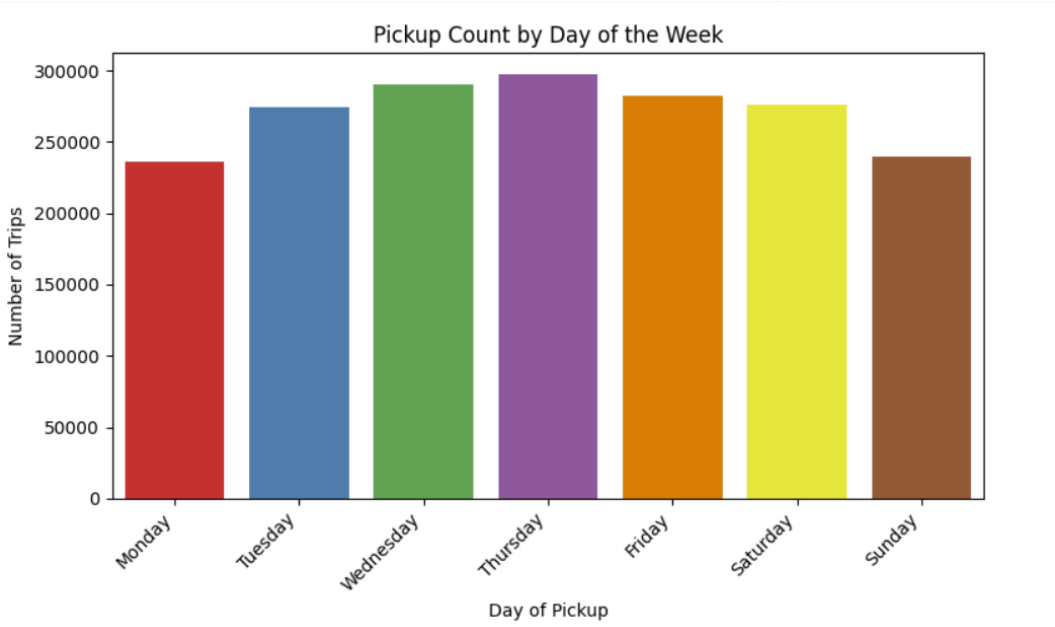




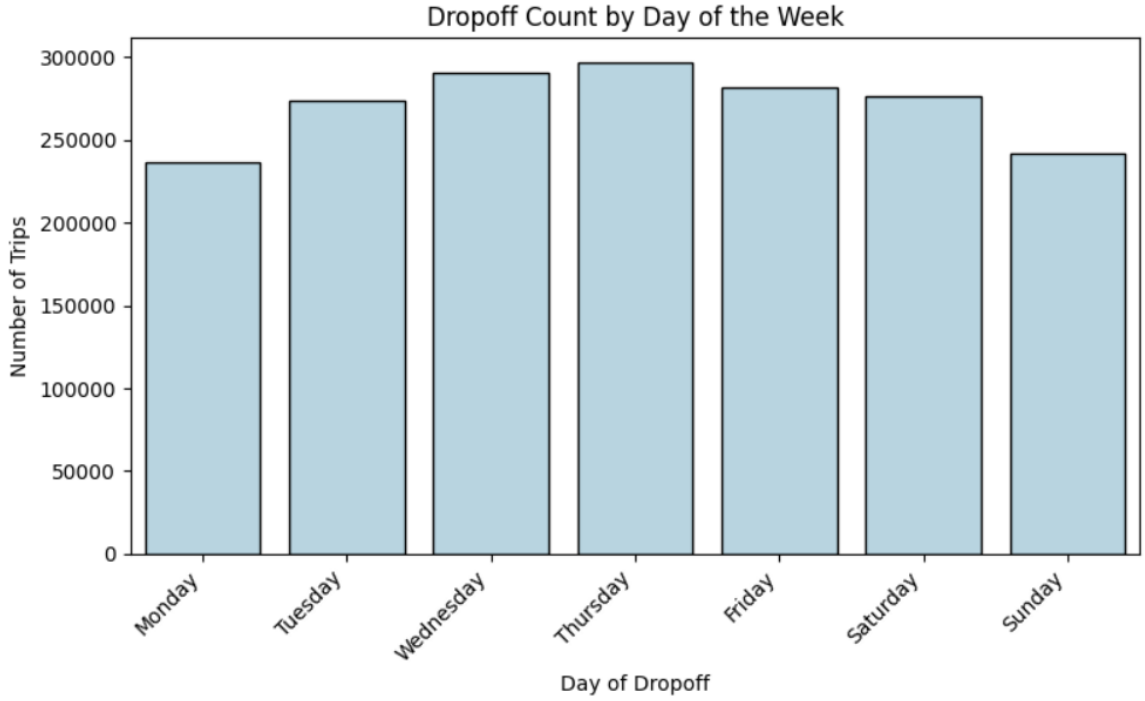
Got an insight that Pickup hours in a day and Dropoff hours in a day are same almost. Next worked on monthly trends of bookings. Later on further analysis is performed on zero data point values and negative values. Analysis goes wrong if negative values are present So filtered data with less than or equal to zeros . So now correct required data is present. Later on groupby analysis is performed on columns. Performed Analysis on Monthly revenue generation, Day waise bookings, speed of taxis, busiest pickup hour, busiest dropoff hour, top pickup zones, top dropoff zones and etc. Next worked with geopandas to analyse geographic locations. Imported taxi_zones.shp file to analyse and visualize data

related to geographic locations. As per numerical or categorical column data graphical visualizations are done.

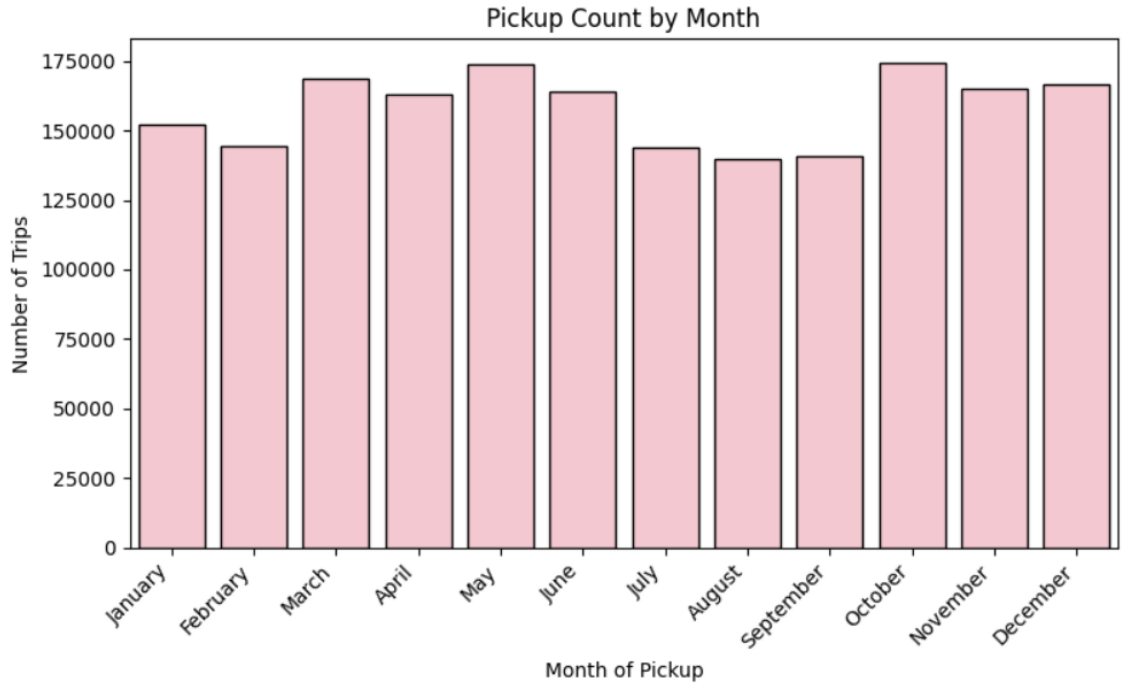
Below is the Analysis of “Pickup Count by Day of the Week” :

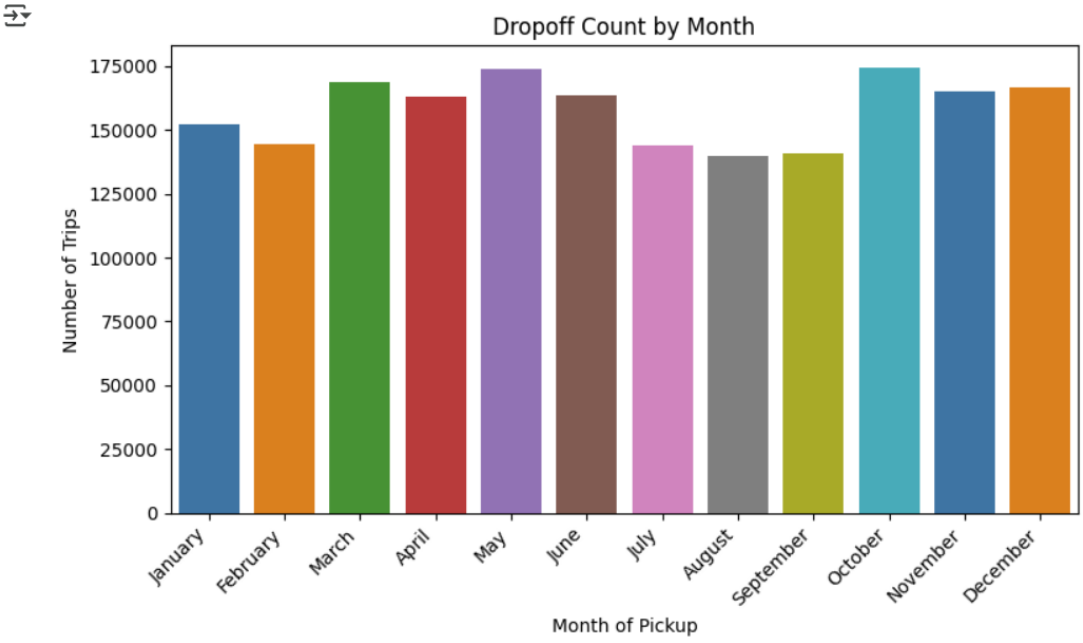


Below is the Analysis of “Dropoff Count by Day of the Week” :

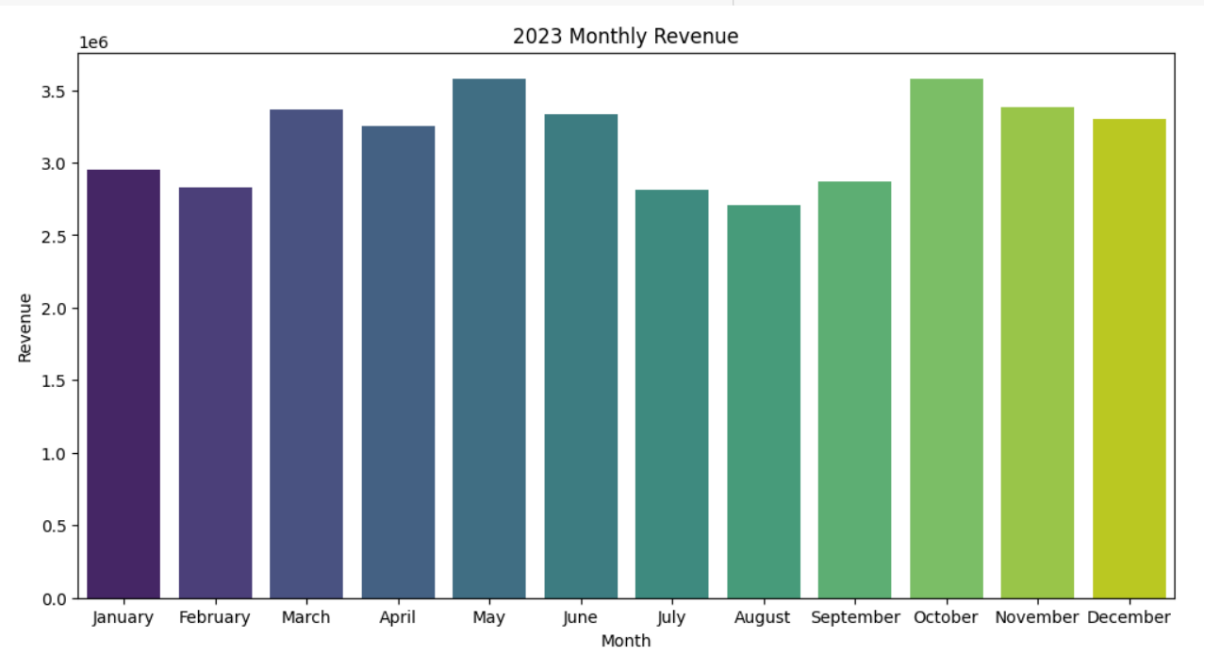


Below images says the Pickup count by Month and Dropoff Count by Month :

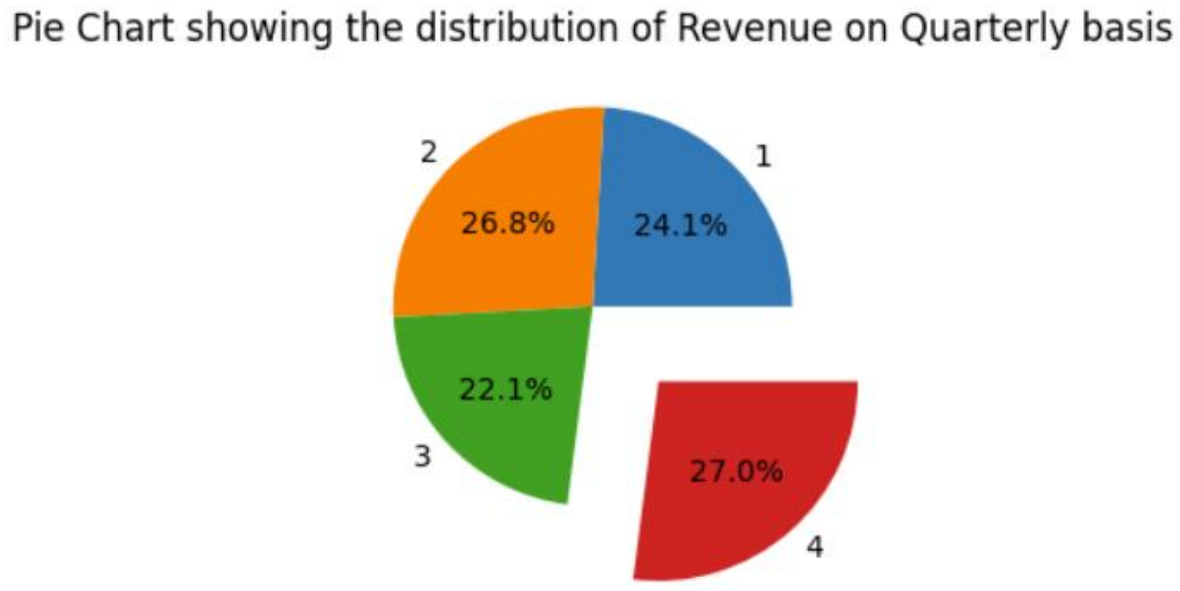




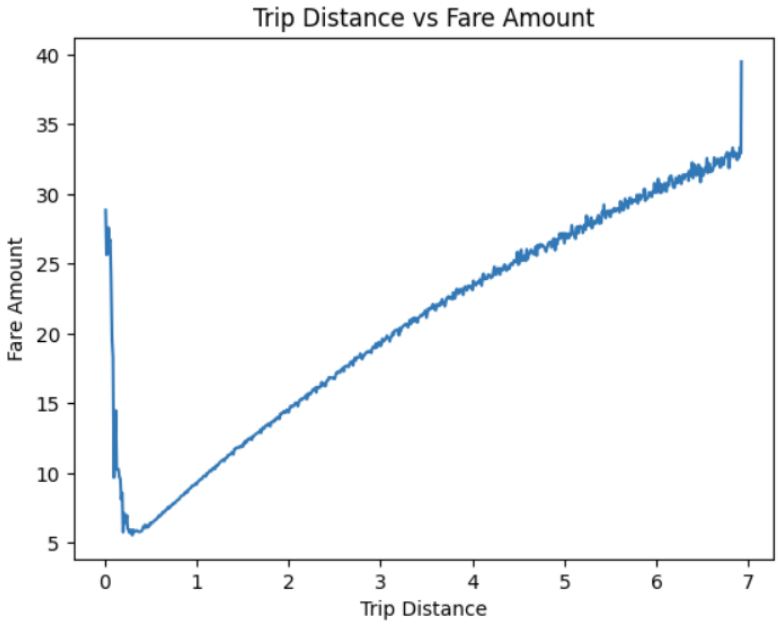
Below image potrays about Monthly Revenue generation.



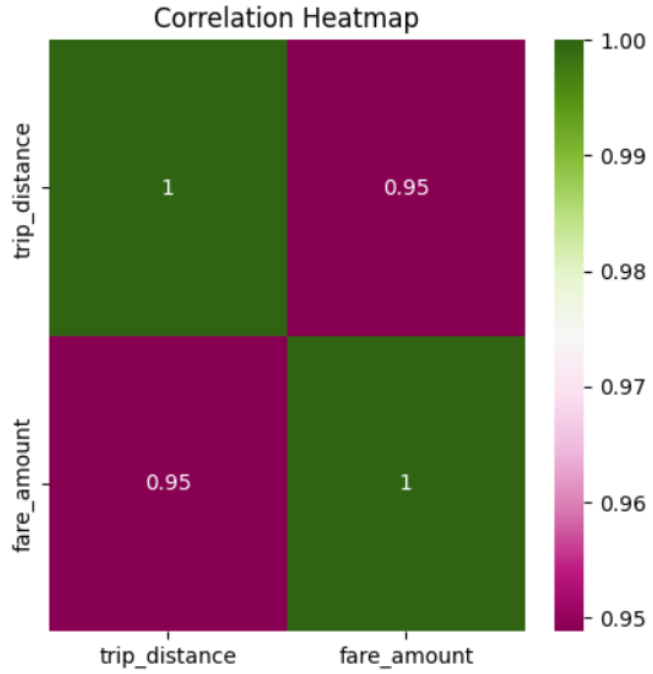
Below image potrays Pie Chart showing the distribution of Revenue on Quarterly Basis :



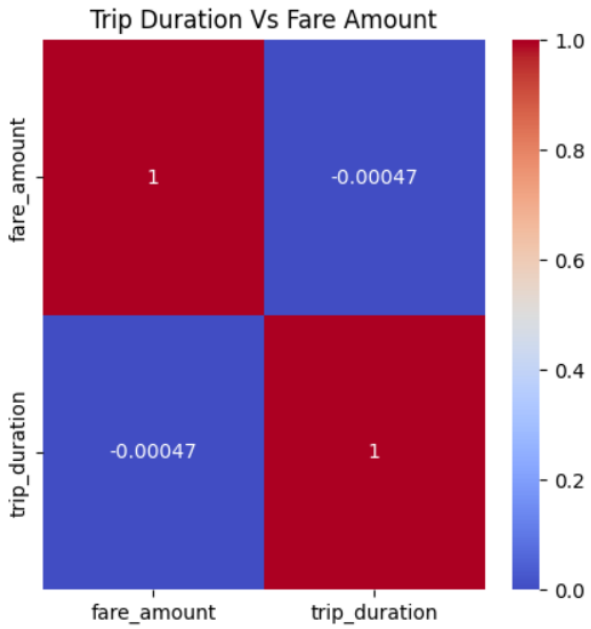
Below visual shows the graph between Trip Distance and Fare Amount :



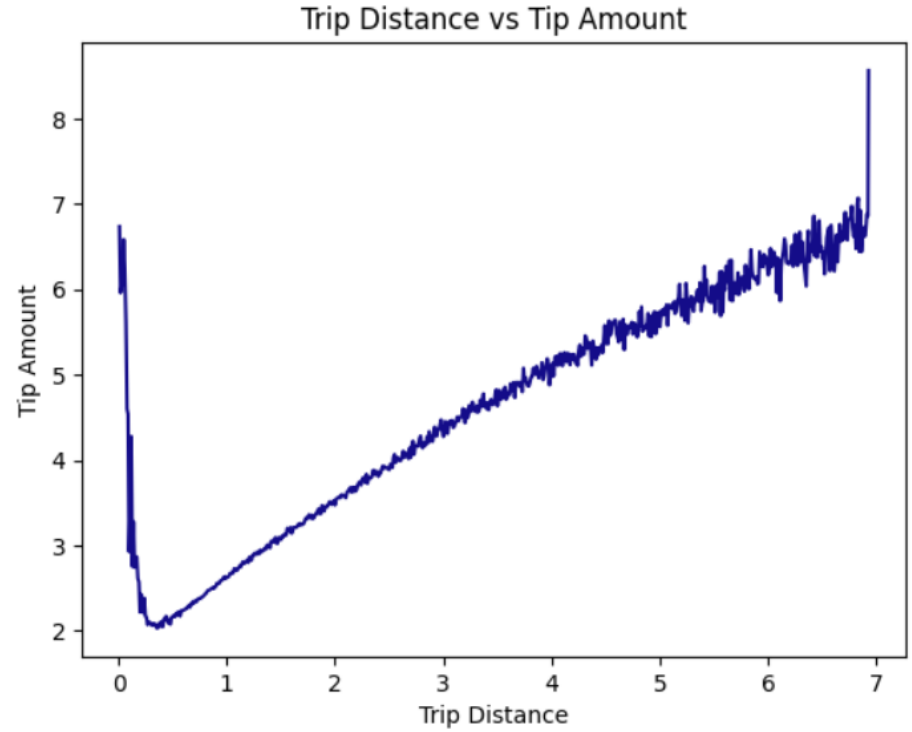
Visualize the Correlation between trip_distance and fare_amount using heatmap



Below visual represents Trip Duration and Fare Amount :



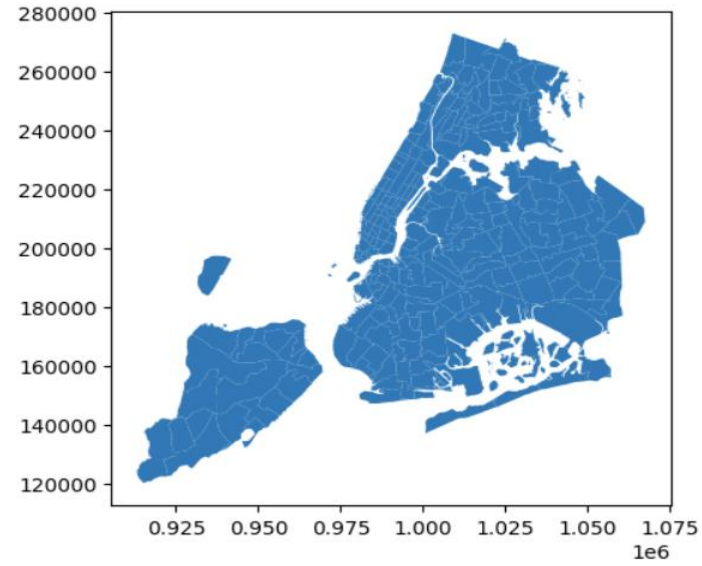
Trip Distance Vs Tip Amount :



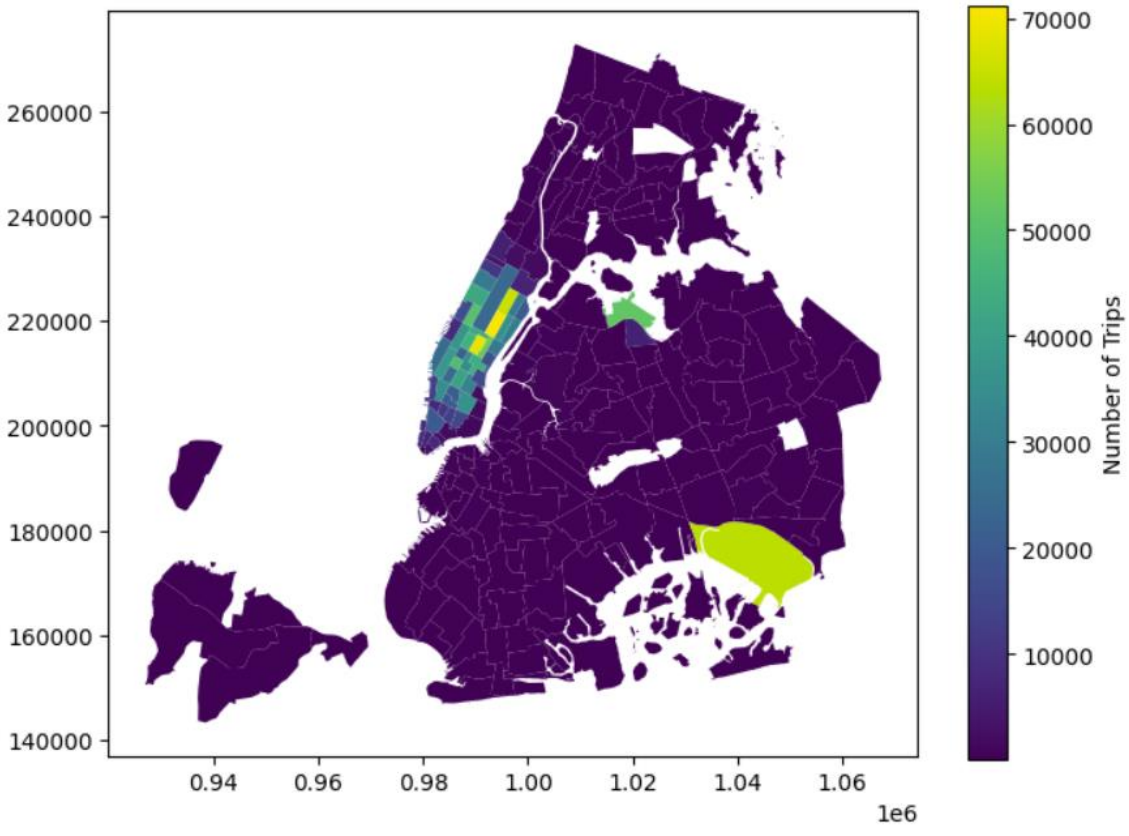
Trip_Duration Vs Fare Amount :

Location wise booking overview :

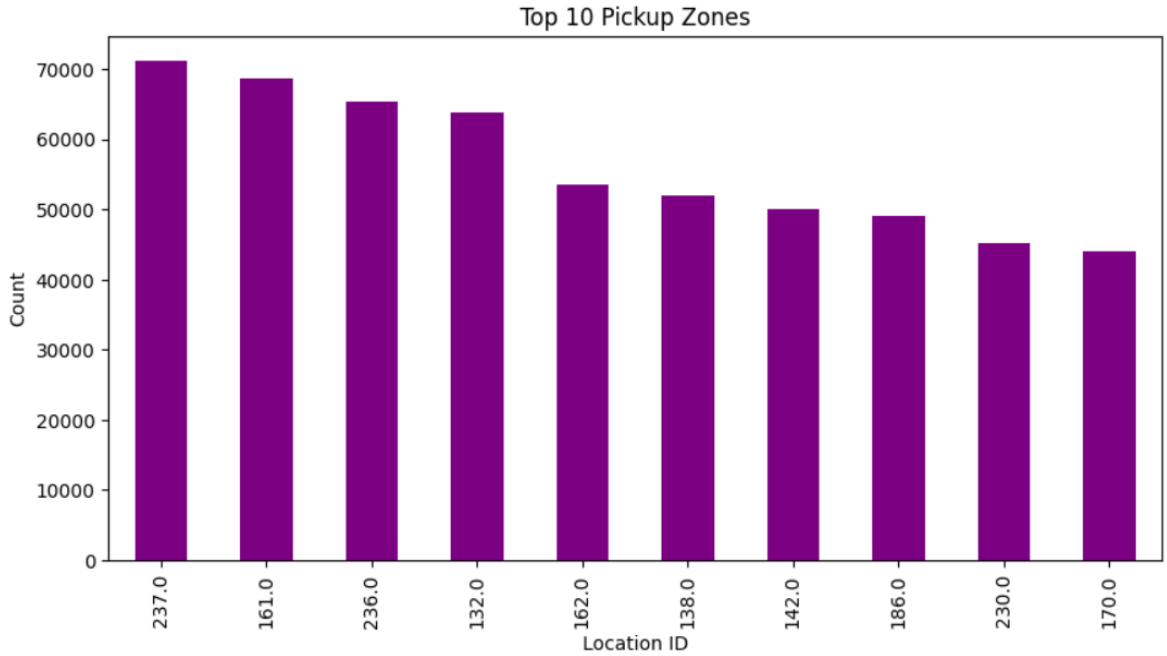
<Axes: >

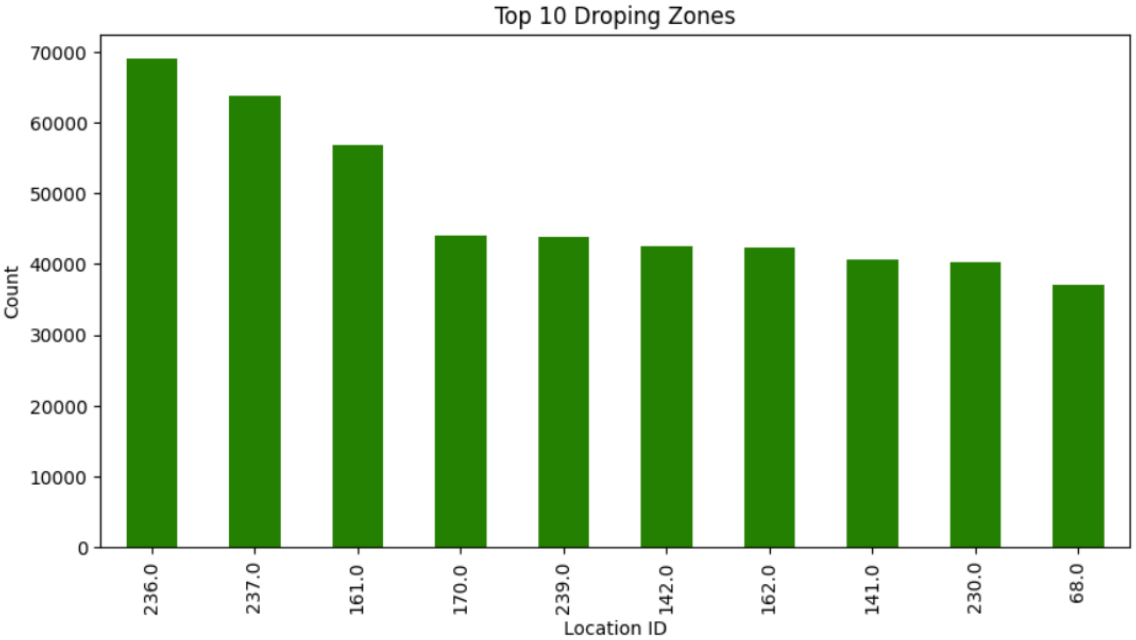


Below visualization gives booking done location wise clearly :

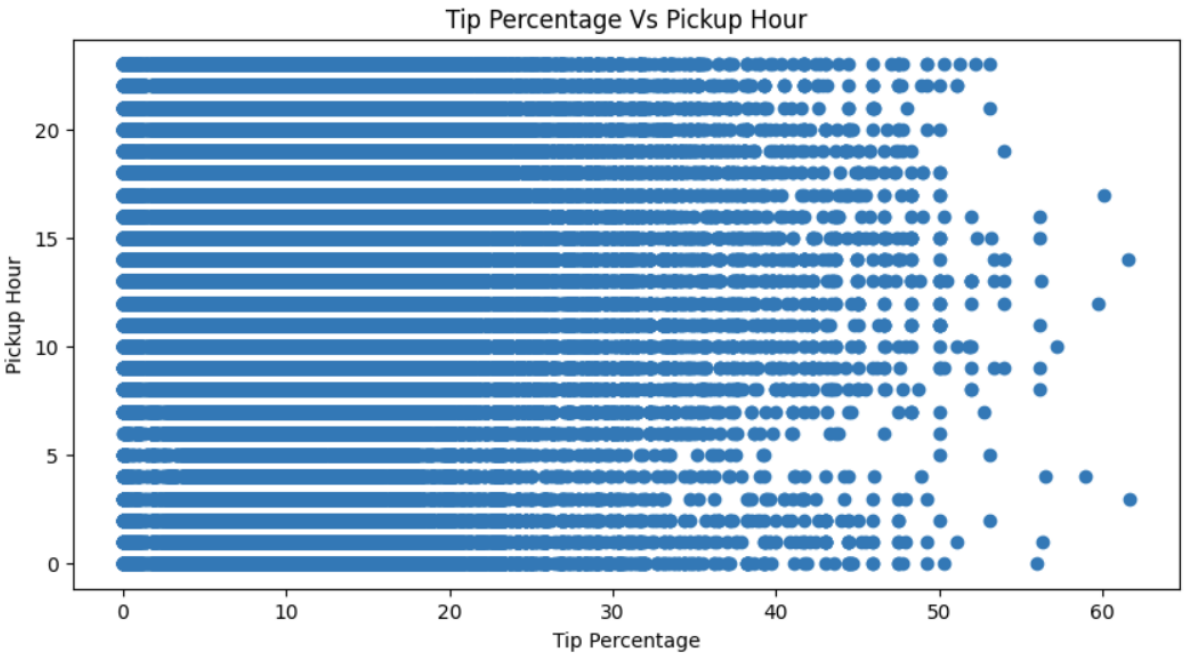


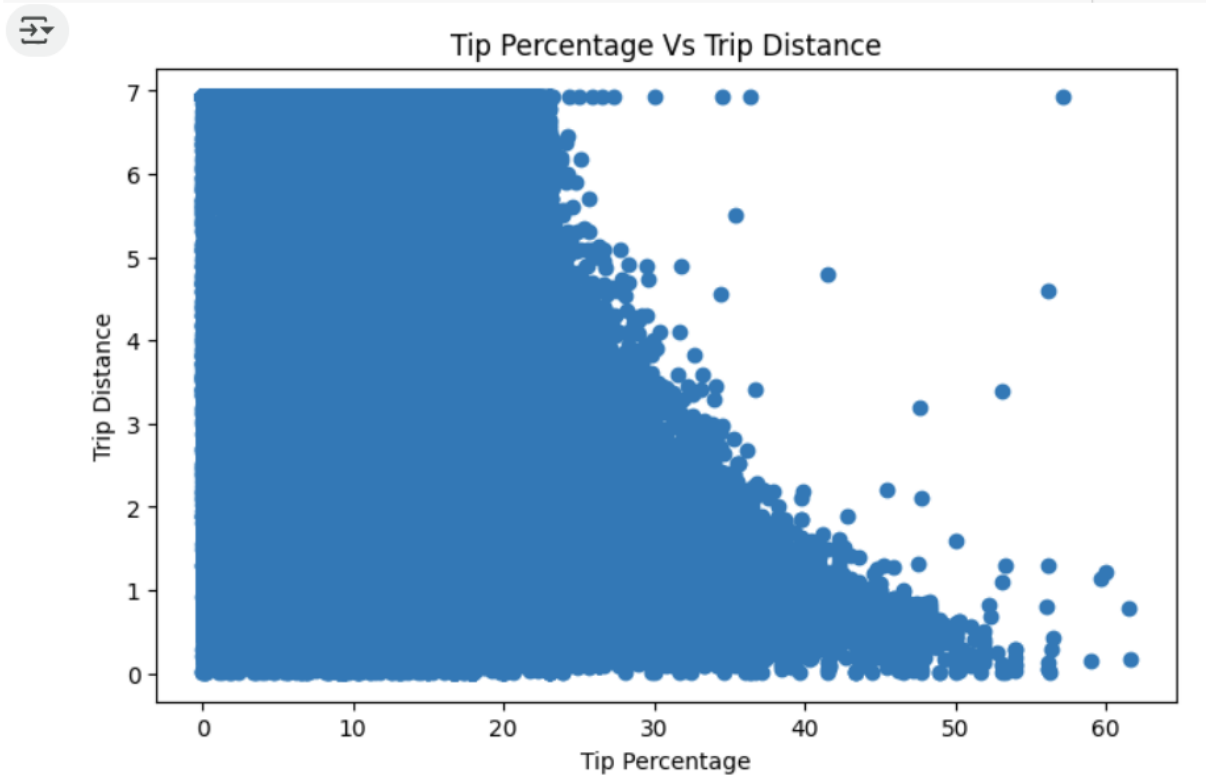
Below visualization shows top 10 Pickup and dropping zones :





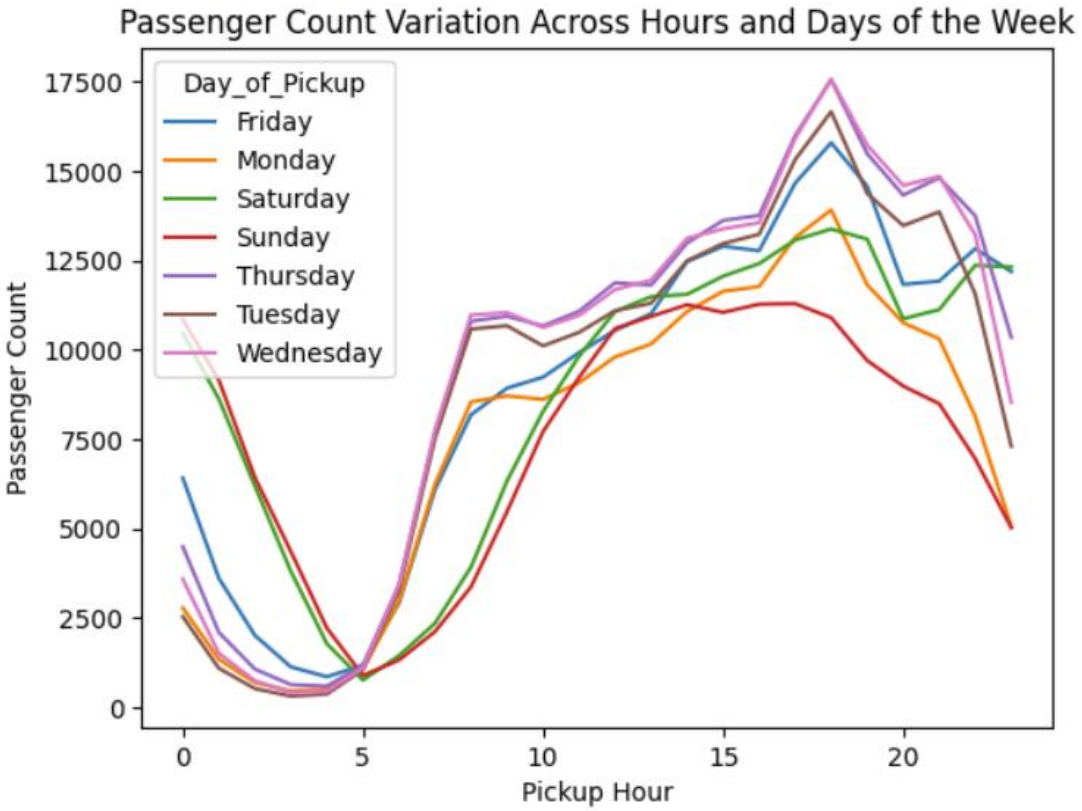
Below visualization shows Tip Percentage varying with othere factors :



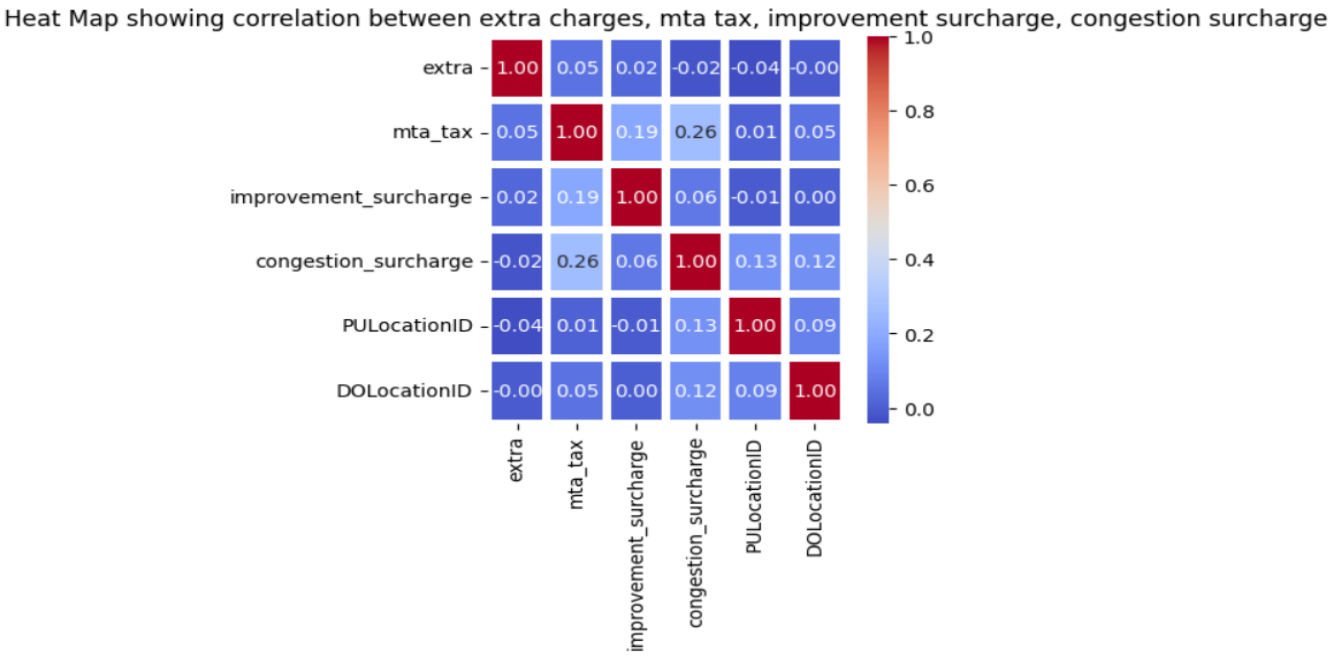


Passenger Count Variation Across Hours and Daya of the Week :

```
<Figure size 1200x1200 with 0 Axes>
```



Below is the mentioned Heatmap across various factors related to tax :



Conclusion :

- Optimizing Routing :

Engage best Tour guider or driver of that particulat place / city / town. Tour Guider or driver who knows that place well will actually suggest good ideas of routing vehicles in correct, traffic less direction. This inturn helps the taxi company in reducing the usage of fuel to taxis. Also Taxi company gets good reputation because of dropping customers in desired location on time. Also the drivers get good ratting because of this.

- Dispatching Vehicles based on Demand patterns :

If Operational demand patterns are observed very clearly the, there is no confusion for which vehicle to go in which direction. Taxis which good milage must be selected to go to pickup customers who are in city outskirts or any far away region. As per observation in horizontal bar graph most of the top 4 Location id's are from 237.0, 161.0, 236.0, 132.0 . Above mentioned 4 location ID's must be given first preference to send more number of taxis over ther. Later on other Location id's comes into preference.

- Operational Inefficiencies :

Operational inefficiencies must not be done by NYC Taxi company. It includes drivers unknowingly going in wrong direction which inturn consumes fuel for the vehicle. This is a huge loss to the company. All the rides which were booked are notting synched with company application which means comany people can't know which vehicle is going in which direction. This must be taken care by officials of the company. Otherwise company may go into Loss.

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

- After visualiaing how passenger count varies across hours and days, we can conclude that Sunday has the least number of bookings which number of taxis required is less in comparision to other days in any month. When compared most of the customers book rides on Thursdays

and Wednesdays. Next highest bookings happed on Tuesday and next comes friday and so on.

- As per the pattern observerd in days of any month as per that number of vehicles must be present accordingly. As per Quarterly patterns it is observed that Q4 is mostly used which means last 3 months in year 2023 is mostly used by customers. Q3 months have lowest ride bookings. As per the Quarterly analysis also NYC Taxi company can hire number of drivers.
- As per Monthly revenue generated also we can analyse. Revenue is generated to ride booking in that specific months. According to monthly revenue generated October and May month have most of the money generated. Next comes March month with Second highest revenue generation. As per per Montly revenue generation also number of drivers hired can be impacted. Also Monthly Pickups and Monthly Drops off can impact number of vehicles required and number of drivers required.
- As per Location ID i.e Pickup Location ID and Drop Location ID also Vehicles must be available at that particular spot on time.

Other Suggestions :

Most of bookings are done by VeriFone Inc Vendor. Other Vendor named Creative Mobile Technologies also took rides but less in comparision to VeriFone Inc Vendor. So number of vehicles must ne increased for Verifone Inc Company in future. So that employees of VeriFone Inc can book more.

Pricing Strategy must be changes because there are numerous taxes on rides. If tax rate increases then number of customers using this application for ride booking decreases. So make sure to finalise 1 or 2 tax rates per ride for any customer. If these changes are implemented then only Revenue generation will increase soon in future.