EDA_Optimising_NYC_Taxis_J_S_V_Lakshmi_Ashrita_Akella

Objective

In the casestudy, we will be exploring exploratory data analysis (EDA) using a dataset of "Yellow taxi rides in New York City".

Problem Statement – copied from notebook

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

Tasks

1. **Data Loading**
2. **Data Cleaning**
3. **Exploratory Analysis**
4. **Creating Visualizations**
5. **Deriving Insights**

Data provided

1. Pick-up and drop-off dates/times
2. Pick-up and drop-off locations
3. Trip distances
4. Itemized fares
5. Rate types
6. Payment types
7. Driver-reported passenger counts

Data Format

- The data is stored in **Parquet format (.parquet)**.
- The dataset spans from **2009 to 2024**, but this study focuses on **2023 data only**.
- Each month of data is stored in a separate Parquet file (12 files for 2023).
- The data was collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers such as vendors and taxi-hailing apps.

Data Description

The dataset includes various fields describing trip details: (copied from the notebook)

Field Name	**Description**
VendorID	Code indicating the TPEP provider (1 = Creative Mobile Technologies, LLC; 2 = VeriFone Inc.).
tpep_pickup_datetime	Date and time when the meter was engaged.
tpep_dropoff_datetime	Date and time when the meter was disengaged.
passenger_count	Number of passengers in the vehicle (driver-entered).
trip_distance	Distance of the trip in miles.
PULocationID	TLC Taxi Zone where the trip started.
DOLocationID	TLC Taxi Zone where the trip ended.
RateCodeID	Final rate code in effect at the end of the trip (e.g., 1 = Standard rate, 2 = JFK, 3 = Newark, etc.).
store_and_fwd_flag	Indicates if the trip record was stored before being sent (Y = Yes, N = No).
payment_type	Payment method (e.g., 1 = Credit card, 2 = Cash, 3 = No charge, etc.).
fare_amount	Meter-calculated fare amount based on time and distance.
extra	Miscellaneous surcharges (e.g., rush hour charges).
MTA_tax	$0.50 NYC MTA tax automatically applied.
improvement_surcharge	$0.30 surcharge added since 2015.
tip_amount	Tip amount for credit card transactions (cash tips not recorded).
tolls_amount	Total tolls paid during the trip.
total_amount	Final amount charged to passengers (excluding cash tips).
congestion_surcharge	NYS congestion surcharge collected for certain trips.
airport_fee	$1.25 charge for pickups at LaGuardia and JFK airports.

Taxi Zones

Each trip record contains fields (PULocationID, DOLocationID) representing pick-up and drop-off locations using numerical zone identifiers (1-263). These correspond to NYC taxi zones, which can be downloaded and matched to trip records using a table or map.

STEP 1: Data Preparation

First i imported all the necessary libraries as suggested in the notebook:

Recommended versions

numpy version: 1.26.4

pandas version: 2.2.2

matplotlib version: 3.10.0

seaborn version: 0.13.2

1.1 Loading the dataset

Here I loaded the parquet dataset given to my by Upgrad in the assignment by using pandas library to load the relevant file after which the data was sampled in order to take only 5% of the total data with respect to all months.

Then the data was further filtered to only keep 300k values as suggested in the notebook to make the data smaller and easier to handle. this data was then stored in a file to be used further with ease.

The next few steps include data cleaning and preparation for extensive EDA.

STEP 2: CLEANING THE DATA

In this step the data was further cleaned of nan values, outliers, and other unnecessary elements according to the suggestions given in the started notebook. For further information, please go through the notebook.

STEP 3: Explanatory Data Analysis

From here, I have started the analysis of the data using common EDA methods.

3.1.1 Categorizing Variables

Here, I categorized all the variables in to two: numerical and categorical variables.

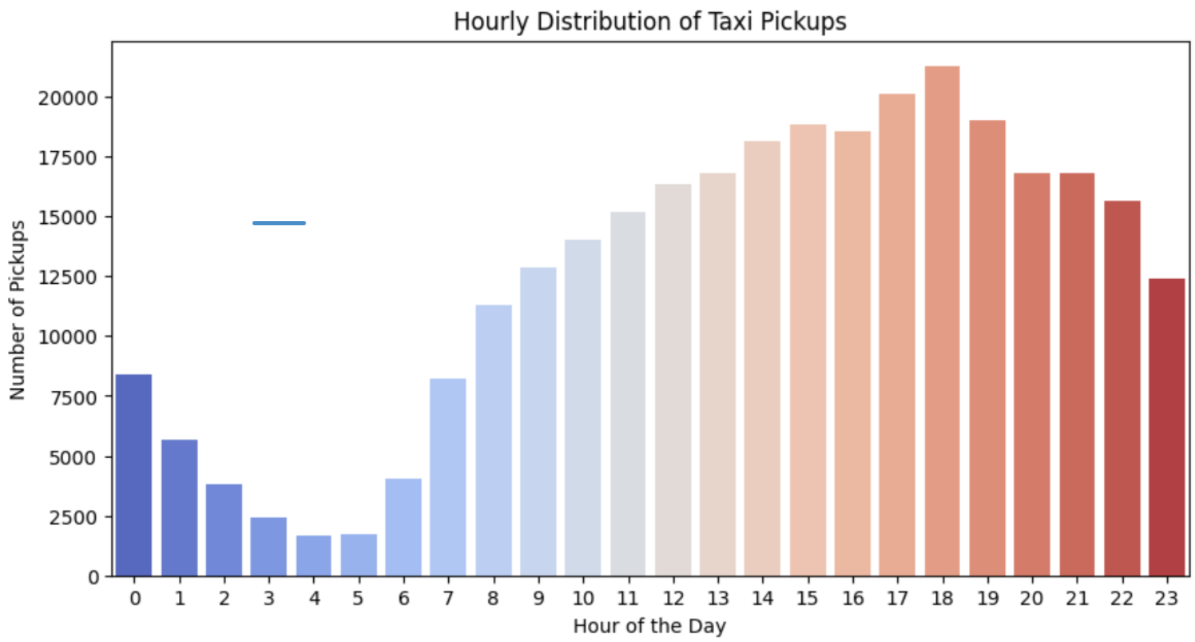
Numerical variables include **counts, distances,** and **monetary values**. IDs such as PULocationID, and DOLocationID are treated as numerical since they represent locations as numbers.

Categorical variables include **identifiers (VendorID)** and **time-based columns (pickup/dropoff times)** because they represent categories rather than continuous values. **payment_type** is also categorical since it refers to different payment methods rather than a measurable quantity.

With that we start our first graphical visualization analysis:

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

3.1.2.1 Find and show the hourly trends in taxi pickups



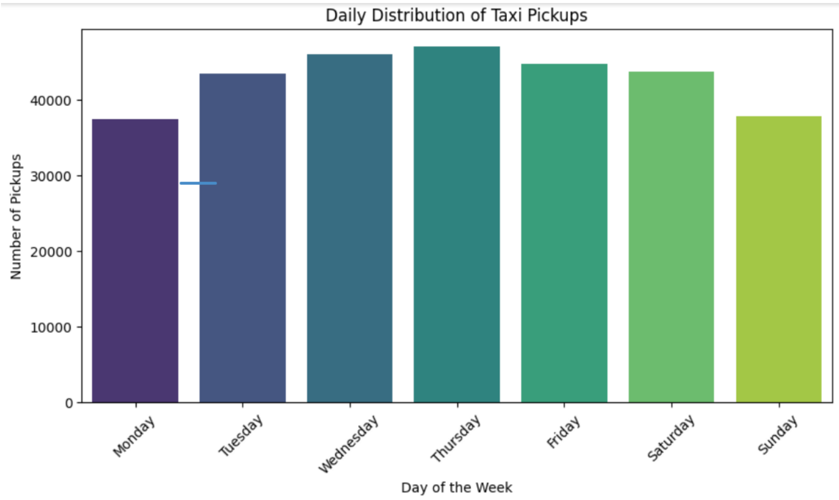
The above graph is the result for the hourly trend of taxi pickups

From this graph we can understand that:

1. The least pickups can be noted between 1AM and 5AM
2. the number of pickups increase significantly after 6AM till 10AM.
3. There is a gradual rise in pickups between 10 AM and 4PM.
4. From 5PM there is a high peak pick up activity that is noted (notably from the evening rush) observing the highest number of pickups at 6PM overall after which

there is a small gradual decrease of the same, though the number of pickups is considerably high.

3.1.2.2 Find and show the daily trends in taxi pickups (days of the week)



From the graph we can tell that,

- Weekdays have more pickups when compared to weekends.
- There is a gradual but not too drastic drop in the number of pick ups during the weekends starting from Friday.
- Highest number of pickups are noted on Thursdays followed closely by Wednesdays.

3.1.2.3 Show the monthly trends in pickups

From this graph we understand that:

- There is high pickup rate during the months of march to June, and October to December. This can indicate that the spring, fall and winter months are seeing high demand in terms of travel.
- July–September Drop:
- Stable Year-End Activity (Nov–Dec):

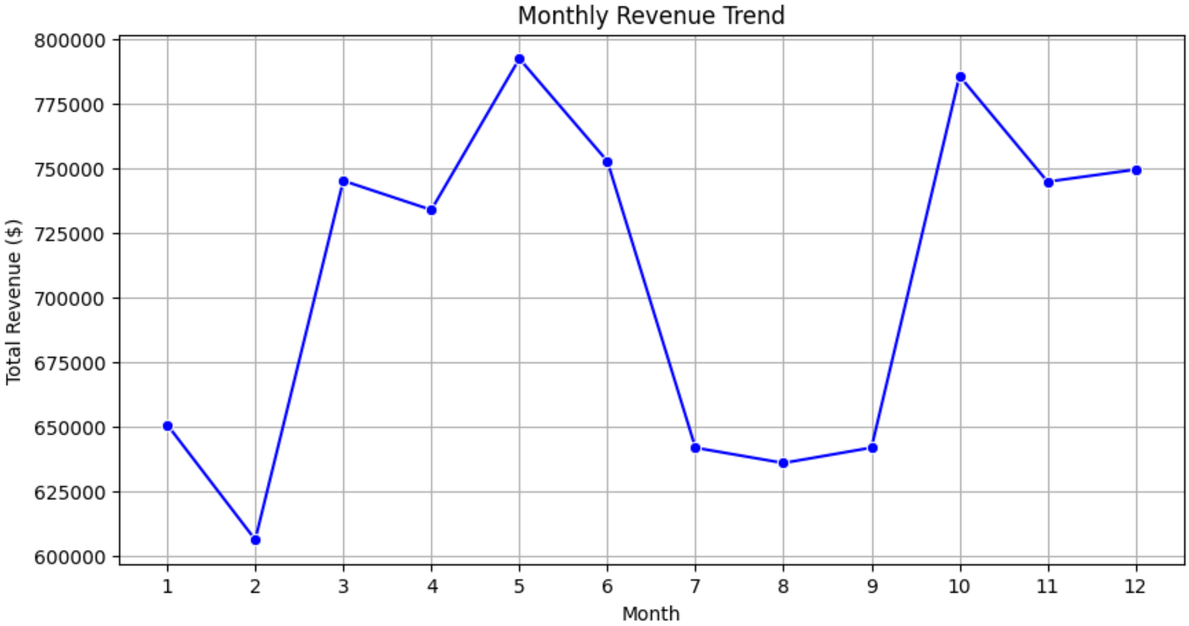
Financial Analysis

For proceeding with this analysis, we need to check if
fare_amount, tip_amount, total_amount, and also trip_distance have zeroes or not. And if they do, we need to selectively remove them.

In this case after finding the number of zeroes in each of these financial parameters, I was able to figure out that fare_amount, trip_distance and total amount can be cleaned of all the zeroes but tip_amount doesn't need any cleaning as tips in general can be 0 at times and depends on the customer.

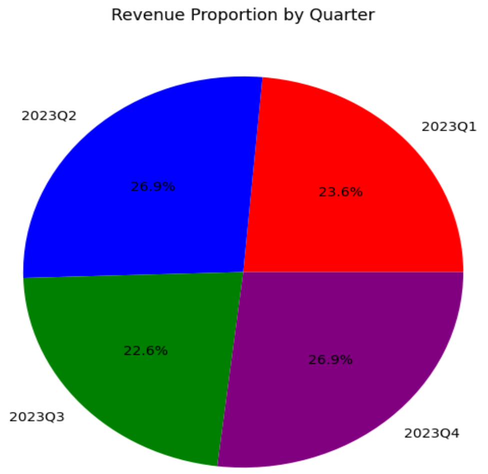
After this starts the analysis:

3.1.4 Analyse the monthly revenue (total_amount) trend



8. There are a lot of fluctuations throughout the graph showing that the total amount trend is a bit unstable.
9. The lowest was observed during the month of February and the highest was observed during the months of May and October
10. While the months of March to June, and October to December are getting total revenues above 725k, the others months seem to be hitting the bottom of grossing only upto 650k.

3.1.5 Show the proportion of each quarter of the year in the revenue

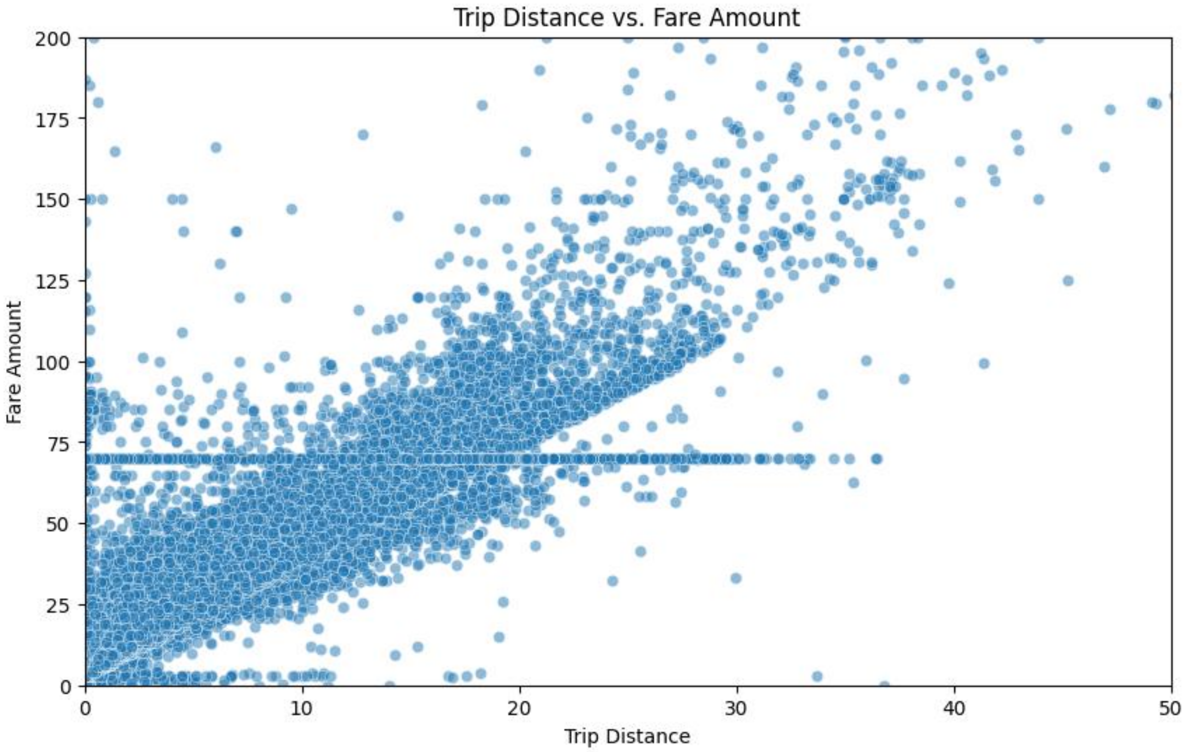


- Quarters 2 and 4 have the highest revenue share of 26.9% of the total revenue.
- Quarter 1 is next in line with a share of 23.6% followed by the lowest which is observed in quarter 3 with a 22.6%.
- This shows that during the spring and fall-winter months, the revenue is at its peak mostly due to the high tourism during spring and the various festivals during the latter whereas during peak winter and summer seasons experience a little dip in the revenue.

3.1.6 Visualise the relationship between trip_distance and fare_amount. Also find the correlation value for these two.

Hint: You can leave out the trips with trip_distance = 0

After leaving out the trips that have distance =0 (which was cleaned before doing this analysis) the below analysis was done.



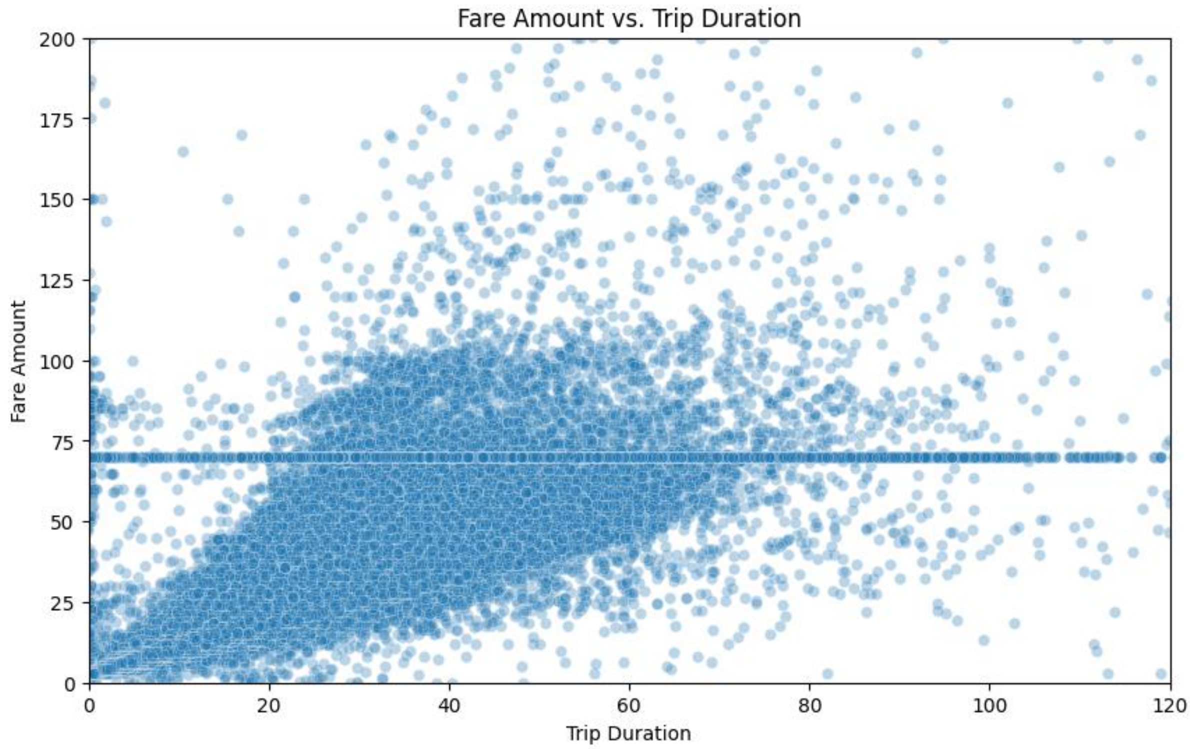
This above scatterplot shows us:

- A strong positive correlation between distance and fare that is as distance increases, fare increases.
- We can also see a lot of clusters of datapoints having less trip distance which means a lot of people travelled less distances and the price ranged between $0 and $75.
- The straight line at $75 showes that there was a fixed amount for some particular zone or place regardless of the distance.

3.1.7 Find and visualise the correlation between:

1. fare_amount and trip duration (pickup time to dropoff time)
2. fare_amount and passenger_count
3. tip_amount and trip_distance

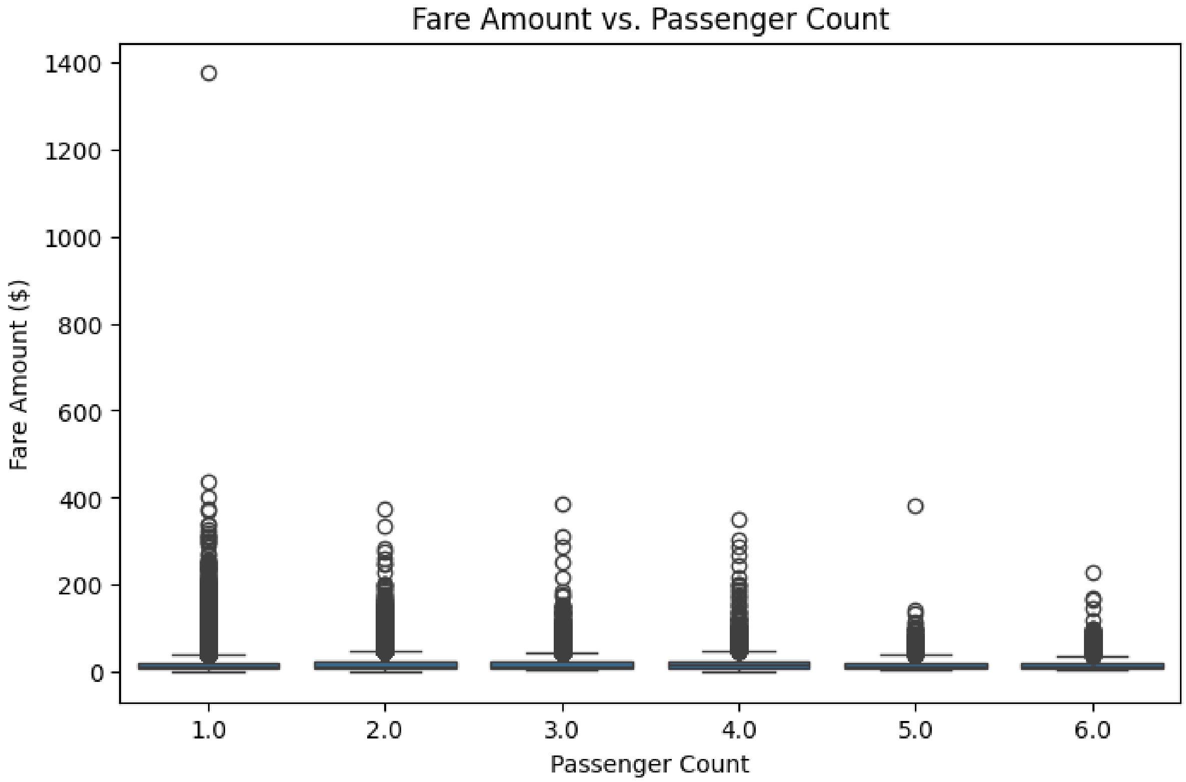
3.1.7.1 fare_amount and trip duration (pickup time to dropoff time)



This graph shows the correlation between trip duration and fare amount through a scatterplot

- there is a very weak correlation between duration of a trip and fare.
- This suggests that duration cannot predict just how much fare amount can be charged.
- Most trips are under fare amount of $75 even if the data is spread across different durations.
- The traight line at 75 again suggests a fixed rate just like the previous graph to certain places irrespective of the time taken.
- There is also very high variability in the fares for those trips above 60 mins which suggests that even though the time taken is long between 1 – 2 hours, the fare is not high in all cases and most cases are recorded under 100$ as well.
- This shows no direct link between the two variables.

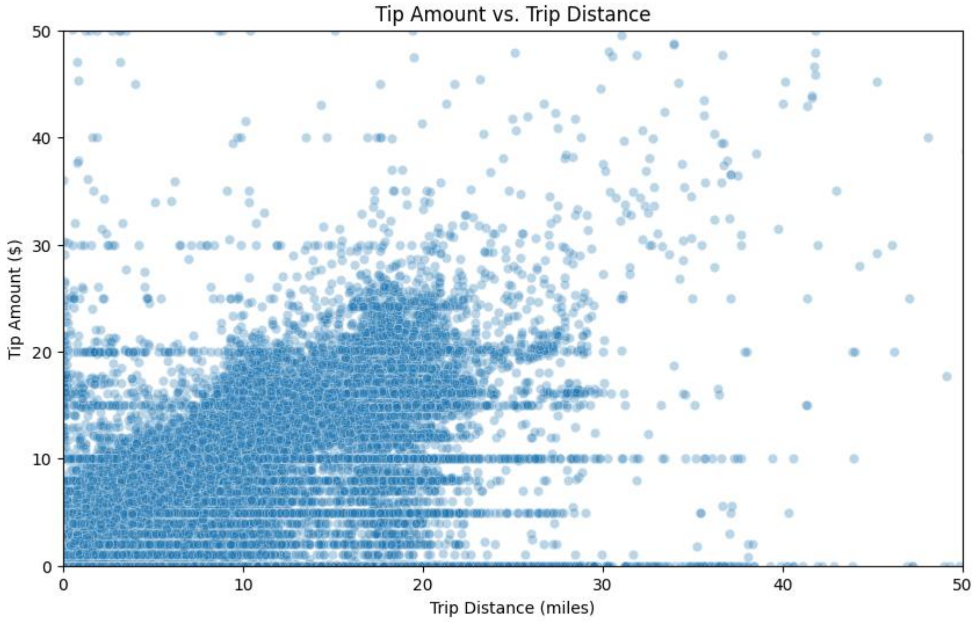
3.1.7.2 Show relationship between fare and number of passengers



Above is a boxplot showcasing the relation between fare and passenger count.

- Here it is made clear that there is almost no correlation between the two as the correlation = 0.041.
- This means passenger count has no impact on fare amount.
- The median fare amount remains static throughout the passenger count.
- The number of trips with only 1 passenger dominate the rest, which makes analysing the rest accurately
- Few extreme fares exist even for 1 passenger trips though they might be outliers or errors ($1400 being way too high to be normal but not ignorable).
- There are many that are taking high fare which might be based on longer trips or tolls but it does not show any correlation between number of passengers and fare.

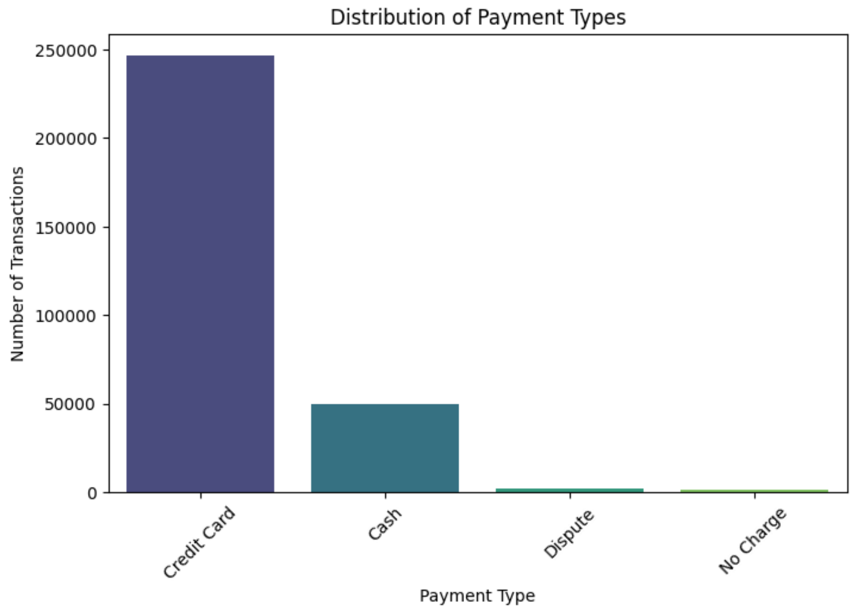
3.1.7.3 Tip Amount vs. Trip Distance



This scatterplot shows the correlation between distance and tip amount.

- This graphs shows moderate positive correlation between the two variables.
- This means that longer trips do get bigger tips but its not perfectly a trend.
- The graph shows that the trips under 10 miles get mostly tip under $10.
- Short distances do get generous tips too but that might just depend upon the customers.
- There is a trend of straight lines at $5, $10 and $20 which these tips were fixed or predetermined.
- Trips above 10 miles are getting high tips and long distances sometimes are left with little to no tips as well. This suggests there is no direct link between the variables.

3.1.8 Analyse the distribution of different payment types (payment_type)



The abovebar graph shows the payment method distribution.

- Credit card payments dominate most.
- This is then followed by cash payments
- There is little dispute regarding either the fare or payment mode and other problems and no charge rides are little to 0

Geographical analysis

3.1.9 data loading

The installation of geopandas was done and the data given is used to visualize and plot the zones accordingly.

3.1.10 Merge the zones data into trip data using the locationID and PULocationID columns.

The necessary columns that is the zones data and original df data were merged to form a new df dataset for usage to do geographical analysis.

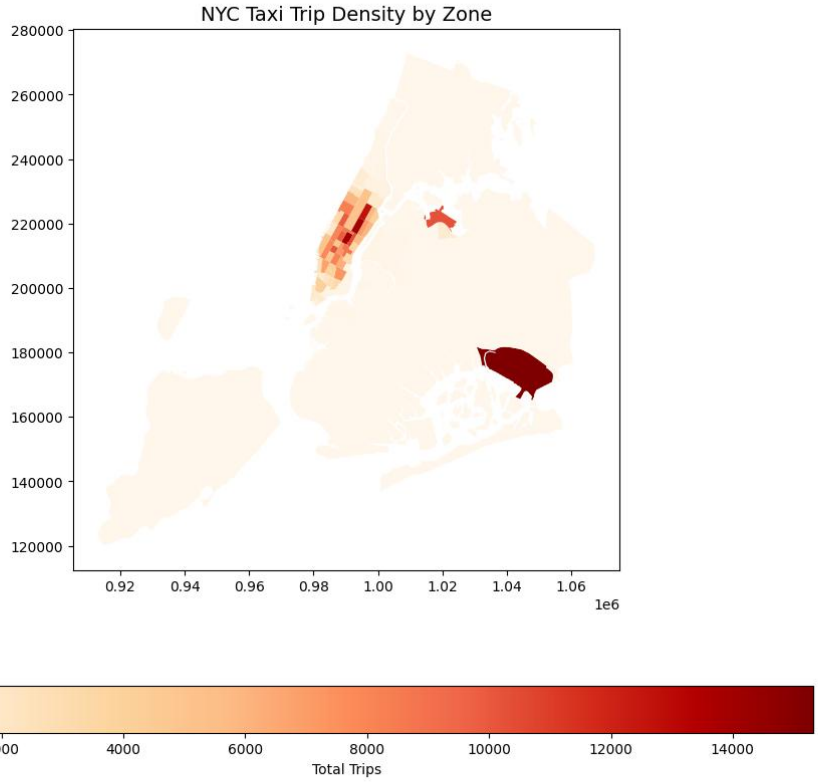
3.1.11

Group data by location IDs to find the total number of trips per location ID

3.1.12

Use the grouped data to add number of trips to the GeoDataFrame.

3.1.13 Plot a color-coded map showing zone-wise trips



The above plot shows that the number of trips that we have taken are concentrated on two specific regions

This is shown in the below table:

	LocationID	zone	borough	trip_count
131	132	JFK Airport	Queens	15327.0
236	237	Upper East Side South	Manhattan	13973.0
160	161	Midtown Center	Manhattan	13961.0
235	236	Upper East Side North	Manhattan	12636.0
161	162	Midtown East	Manhattan	10494.0
137	138	LaGuardia Airport	Queens	10271.0
185	186	Penn Station/Madison Sq West	Manhattan	10169.0
141	142	Lincoln Square East	Manhattan	9895.0
229	230	Times Sq/Theatre District	Manhattan	9767.0
169	170	Murray Hill	Manhattan	8855.0

The table shows the top 10 most travelled to/from.

3.2 Detailed EDA: Insights and Strategies

3.2.1 Identify slow routes by calculating the average time taken by cabs to get from one zone to another at different hours of the day.

Give the formula: Speed on a route *X* for hour *Y* = (*distance of the route X / average trip duration for hour Y*)

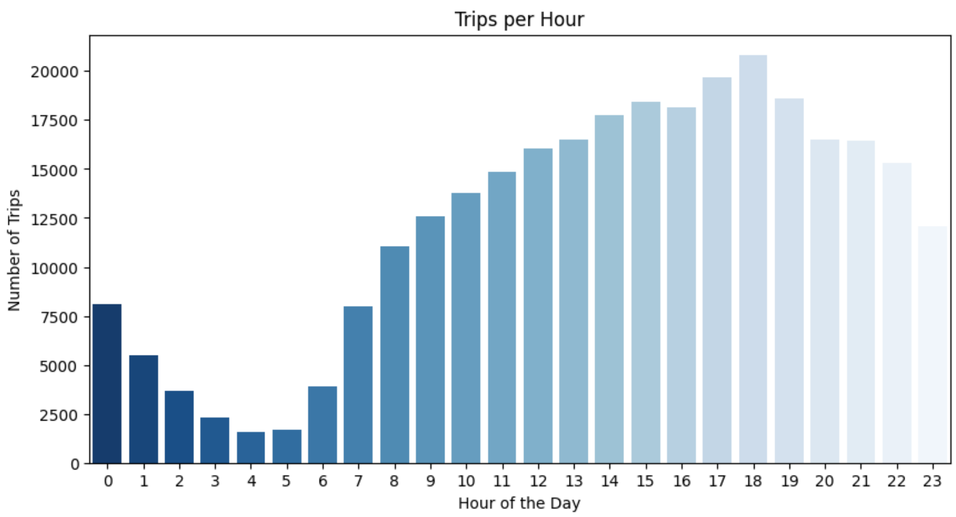
To find the spend we need the trip duration. So I created a new column with the duration of the trip and added it to the df.

Then after I created a column for the average speed of each trip.

After which I grouped the slowest routes based on the speed and displayed them

	hour	PULocationID	DOLocationID	total_distance	total_duration	avg_speed
0	0	263	68	4.64	23.911111	0.194052
1	1	90	33	4.44	23.217778	0.191233
2	2	161	264	2.23	21.744722	0.102554
3	3	246	143	2.48	20.363333	0.121788
4	4	162	162	1.54	23.772778	0.064780

3.2.2 Calculate the number of trips at each hour of the day and visualise them. Find the busiest hour and show the number of trips for that hour.



The graph shows that the busiest time overall is at 6PM with a total of 20810 trips.

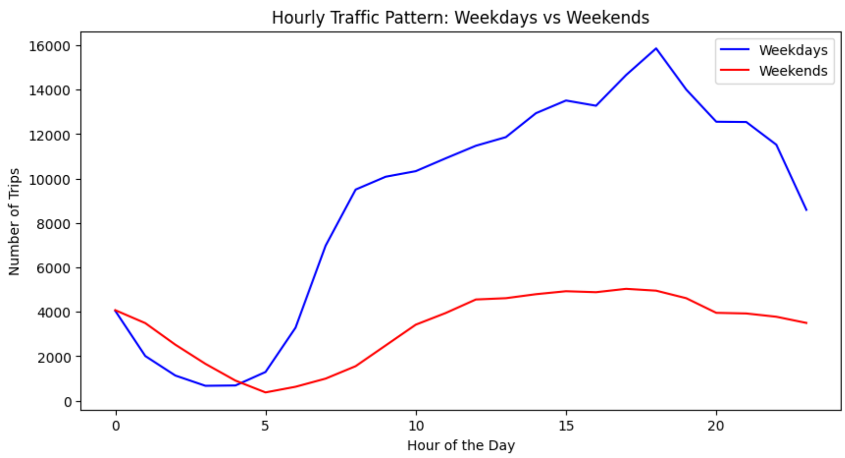
But to find the actual number, you have to scale the number up by the sampling ratio.

3.2.3 Find the actual number of trips in the five busiest hours

hour	count
18	416200
17	393880
19	372380
15	368780
16	363200

dtype: int64

3.2.4 Compare hourly traffic pattern on weekdays. Also compare for weekend.



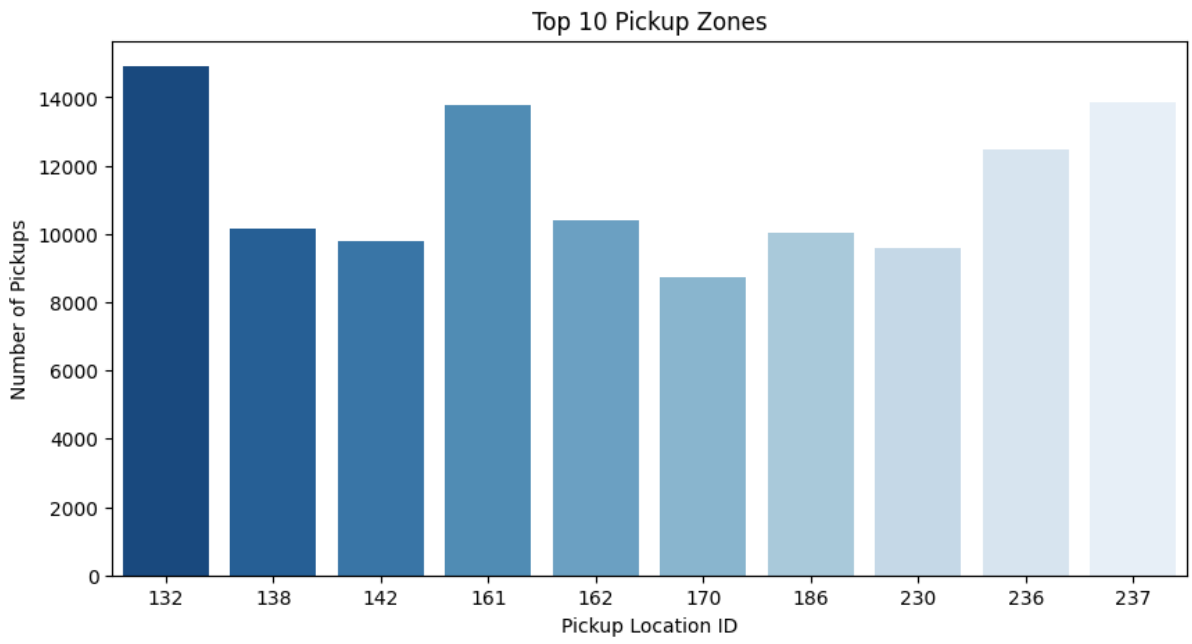
What can you infer from the above patterns? How will finding busy and quiet hours for each day help us?

From the graph we can infer that:

- Weekdays:
  - During the late nights, from midnight to 5 AM very few trips take place.
  - The number of trips pick up after 6AM peaking during 8-9 AM range.
  - Afternoons experience a steady increase.
  - The second peak of the day happens during the 5-7 PM time range.
  - And then there is a gradual decrease in the trips at 9 PM
- Weekends:
  - They are more consistent throughout the day
  - There is no sharp morning peak but there is nor much traffic or trips during the realy morning till 5AM.
  - The busiest hours are recorded between 10 AM and 4 PM
  - Evening activity is also evenly spread and there is no drop at 7PM – 9PM like on weekdays.

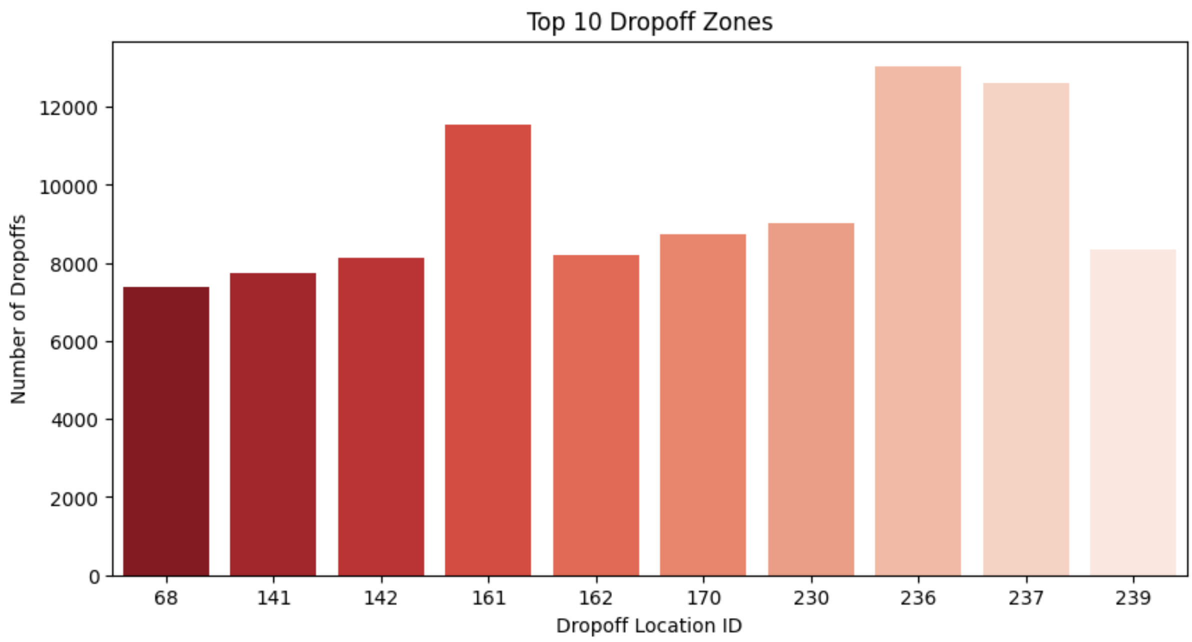
- From these patterns we can infer the busiest hours and the most peak time for travels.
- Also getting this information helps with operations, better management of transportation, etc.
- This also helps in understanding routes and overall trend of the trips.

3.2.5 Identify top 10 zones with high hourly pickups. Do the same for hourly dropoffs. Show pickup and dropoff trends in these zones.



From the pickup zones trend in the bar graph above, we can understand:

- Highest pickups are in zones 132, 161, 236, and 237.
- The rest include 138, 142, 162, 170, 186, and 230



The dropoff zones bar graph shows:

- Highest drops are made at the zones 236, 161, and 237.
- The rest are significantly lesser but are some of the highest drops made overall.

3.2.6 Find the ratio of pickups and dropoffs in each zone. Display the 10 highest (pickup/drop) and 10 lowest (pickup/drop) ratios.

These pick up and drop off ratios indicate how frequent a zone is used for the same.

```
(70     10.213740
 132     4.679008
 138     2.688542
 186     1.531160
 43      1.396557
 114     1.388415
 249     1.375495
 162     1.264645
 142     1.204207
 161     1.195562
 Name: count, dtype: float64,
 201    0.057143
 202    0.054217
 64     0.050000
 160    0.047059
 171    0.039216
 73     0.035714
 53     0.028571
 16     0.020000
 257    0.017699
 1      0.003390
 Name: count, dtype: float64)
```

The given result shows the top 10 pick up and dropoff zones of all and their ratio.

Zones 70, 132 and 138 are used more for pickups than for drops.

The lowest being zones 1, 257, and 16 have lesser pickups than they do drops indicating these places aren't where passengers start new trips that often but end trips.

3.2.7 Identify zones with high pickup and dropoff traffic during night hours (11PM to 5AM)

```
(PULocationID
 79     2572
 132    2305
 249    2000
 48     1692
 148    1625
 114    1405
 230    1260
 186    1073
 164     986
 68      946
 Name: count, dtype: int64,
 DOLocationID
 79     1326
 48     1079
 170    1054
 68      976
 107     907
 141     887
 249     777
 263     755
 148     711
 229     708
 Name: count, dtype: int64)
```

3.2.8 Find the revenue share for nighttime and daytime hours.

Most revenue is generated over the day time which is about 88%

And the revenue split for the night hours is about only 12%

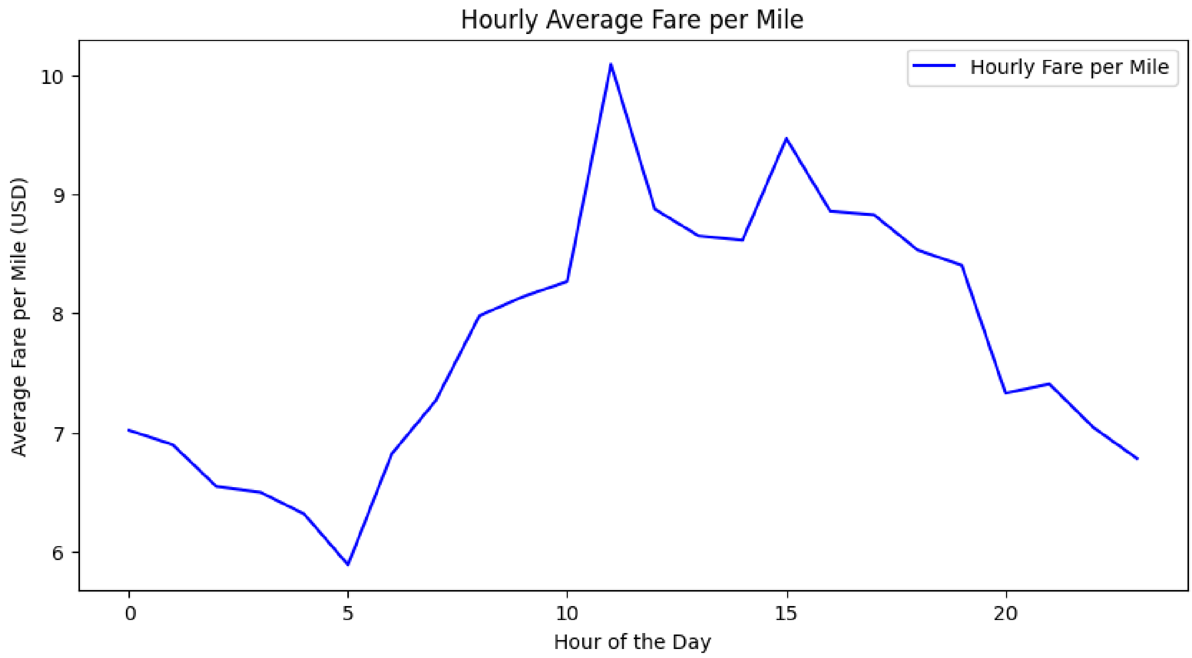
3.2.9 For the different passenger counts, find the average fare per mile per passenger.

For instance, suppose the average fare per mile for trips with 3 passengers is 3 USD/mile, then the fare per mile per passenger will be 1 USD/mile.

	fare_per_mile_per_passenger
passenger_count	
1.0	8.195292
2.0	4.106408
3.0	2.661634
4.0	2.291200
5.0	1.557353
6.0	1.308618

dtype: float64

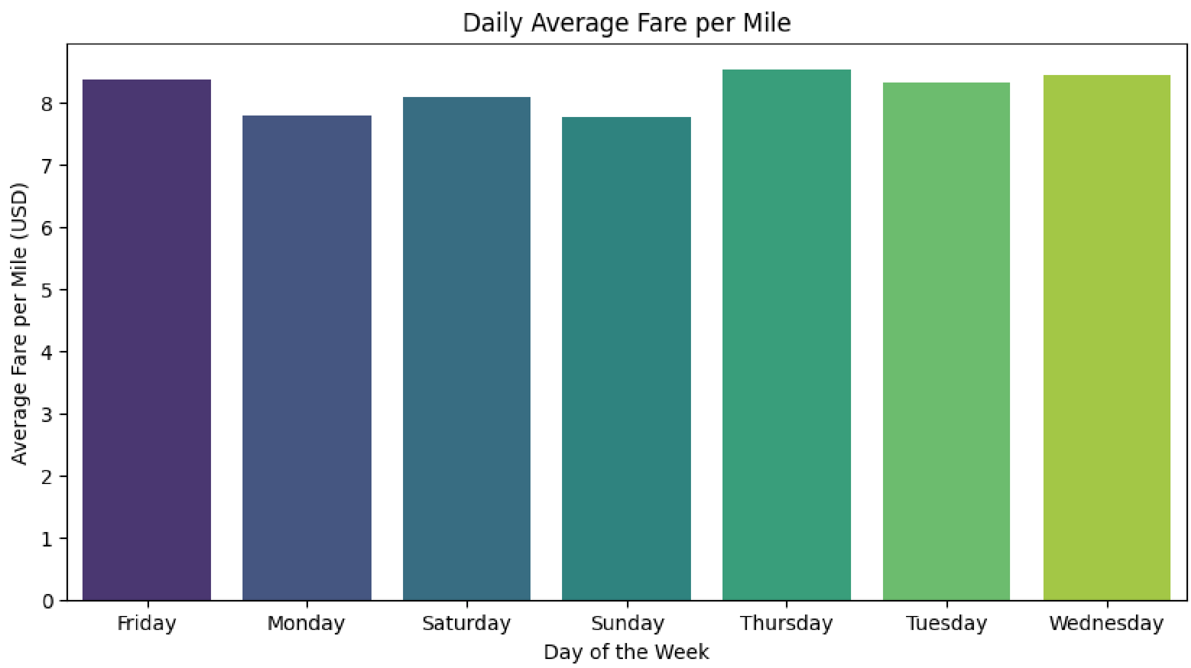
3.2.10 find the average fare per mile by hours of the day and by days of the week



Hourly Trends in Average Fare Per Mile graph

Prices tend to be high during the peak times of the day like at 10 AM, at 3 PM, and then gradually decline.

Early morning and late nights usually record lesser fare amounts.

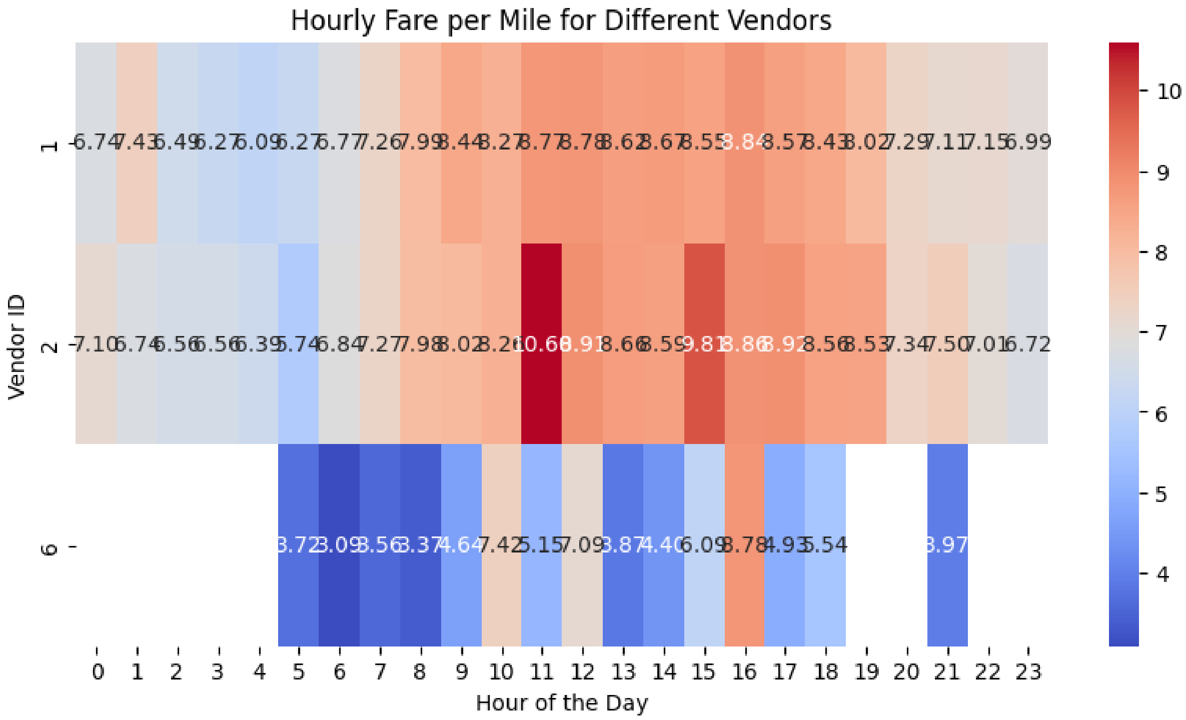


Daily Trends in Average Fare Per Mile graph

Tuesday – thurdays have higher fare per mile whereas the weekends tend to get lower.

Fridays behave a little odder with mixed up fare amounts being the end of the weekdays and start fo the weekend

3.2.11 Analyse the average fare per mile for the different vendors for different hours of the day

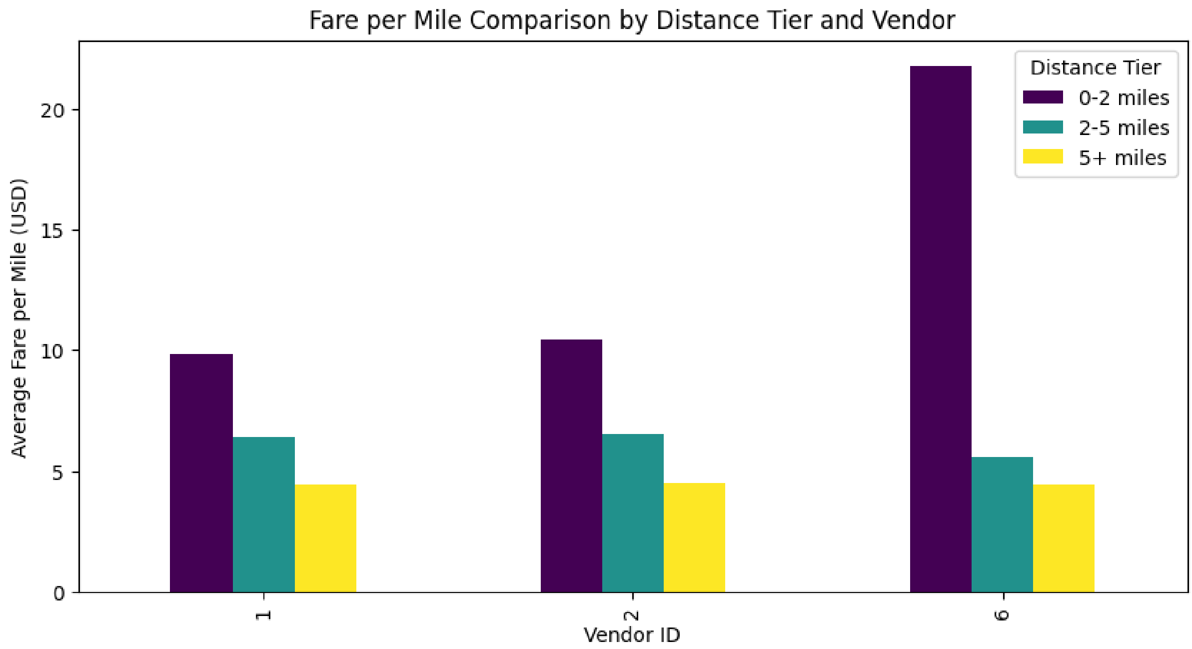


The above heat map provides a comparison between the average fare amount per mile by different vendors at different times of the day.

Each row represents a vendor, anc each column, a specific time of day, and the colours represent the fare amount per mile.

- The fares are higher during the peak times for most fo the vendors the peak hours being 9AM to 6 PM
- Late nights are lesser in general
- Vender 2 shows a spike in fare at 12PM and 4PM
- Vender 6 has significantly lower prices and fares when compared to vendors 2 and 1 the lowest recorded between 6AM and 10 AM
- Vendor 1's prices remain high throughout the day with a slight dip in the morning before gradually increasing again.

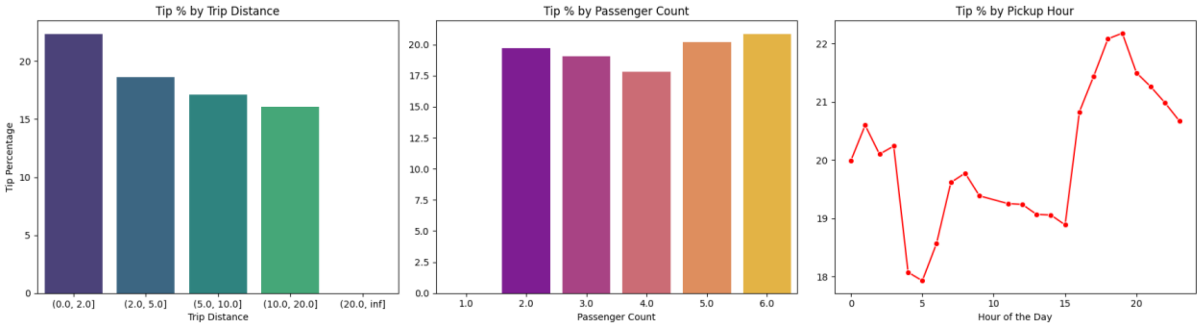
3.2.12 Compare the fare rates of the different vendors in a tiered fashion. Analyse the average fare per mile for distances upto 2 miles. Analyse the fare per mile for distances from 2 to 5 miles. And then for distances more than 5 miles.



The bar chart above compares the average fare per mile to different vendors differentiated by tip distances.

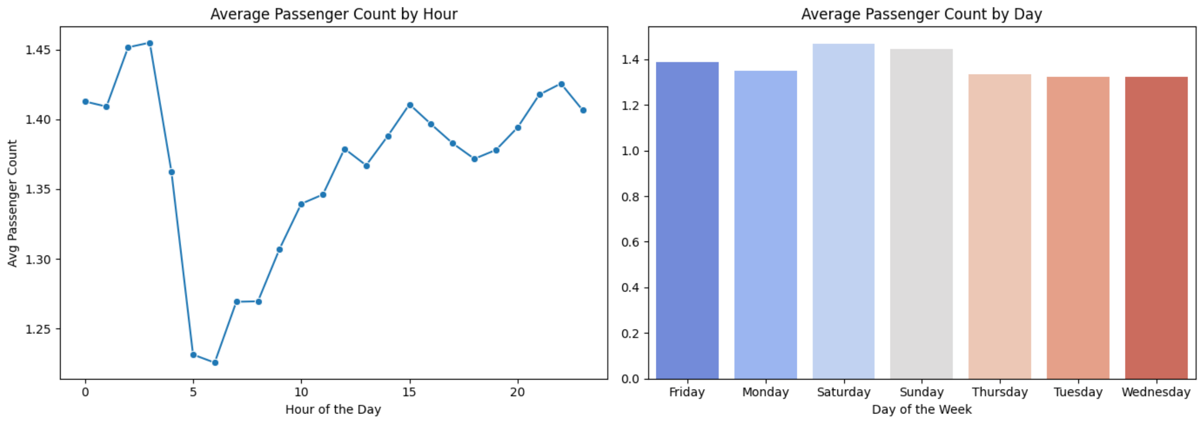
- Vendor 6 charges the highest for short trips whereas 1 and 2 have similar price ranges
- Fare price drops across all vendors when it comes to medium trips as the distance increases with vendor 6 lowering the fare per mile more than the other 2.
- Long trips fare remains same across all three vendors.

3.2.13 Analyse average tip percentages based on trip distances, passenger counts and time of pickup. What factors lead to low tip percentages?



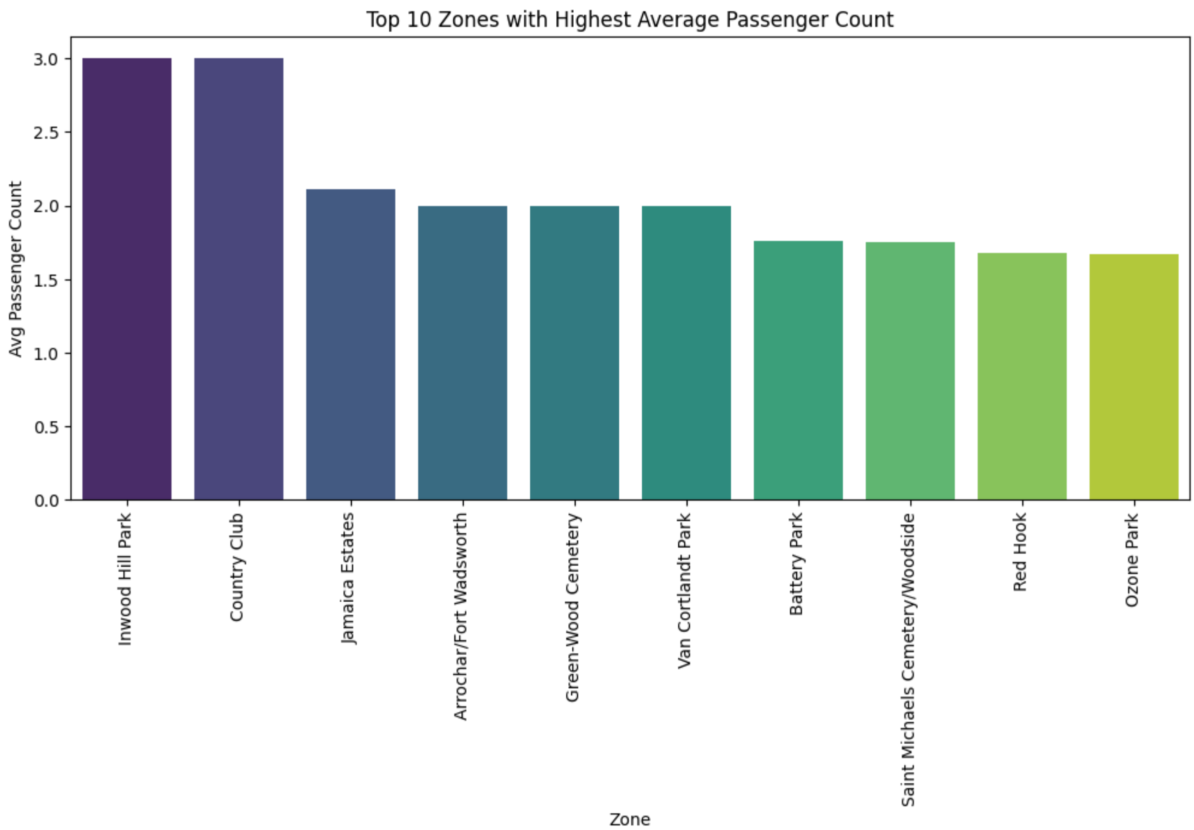
- Trip distance
  - Shorter trips receive higher tips
  - Longer the trips lower the tips get
  - this is mostly because at the trip prolongs, it feels more transactional but shorter the trip is, the more personal interaction makes it comfortable.
- Passenger count
  - Solo riders and small groups of 2 to 3 people tip upto 20%
  - Larger groups tip a little better upto 21 percent
  - This is mostly because large groups tend to feel the pressure of tipping more fairly.
- Pickup time
  - Late night and early mornings show lesser tips
  - Evenings are when tips are highest upto 22%
  - This is possibly due to the type of passengers moving in the evenings are probably more generous as they might be tourists or going somewhere.
  - But mornings are mostly dull and passengers are tired and moody due to which it would affect the tip

3.2.14 Analyse the variation of passenger count across hours and days of the week.



- Passenger count across hour of the day
    - Passenger count is higher during midnights with a sharp decline during the early mornings
    - Count increases gradually during the day from 8AM and peaks after 4PM again then later gets stabilized at night
- Passenger count by the day of the week
    - Highest count is seen on the weeneds
    - Weekdays are mostly consistent with the count and remains stable.
    - There is a slight dip in the midweek from Tuesday to Thursday

3.2.15 Analyse the variation of passenger counts across zones

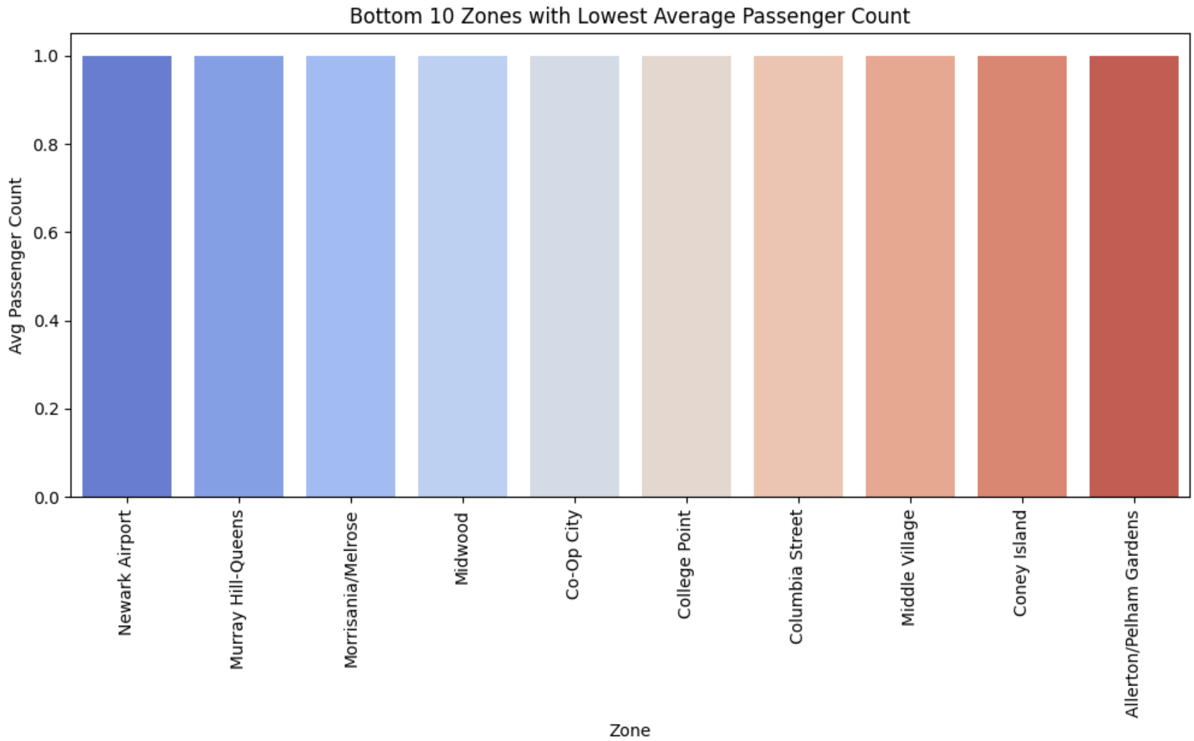


The highest average count of passengers with 3 per trip is recorded at inwood hill park and country club

Jamaica estates, arrorchar/fort Wadsworth and Van Cortlandt park also show high passenger count with 2 to 2.1

These places are most likely where shared or group rides take place. Including family places, a lot of these places are where families tend to attend together which increases passenger count.

Graph 2: bottom 10 zones with lowest average passenger count:

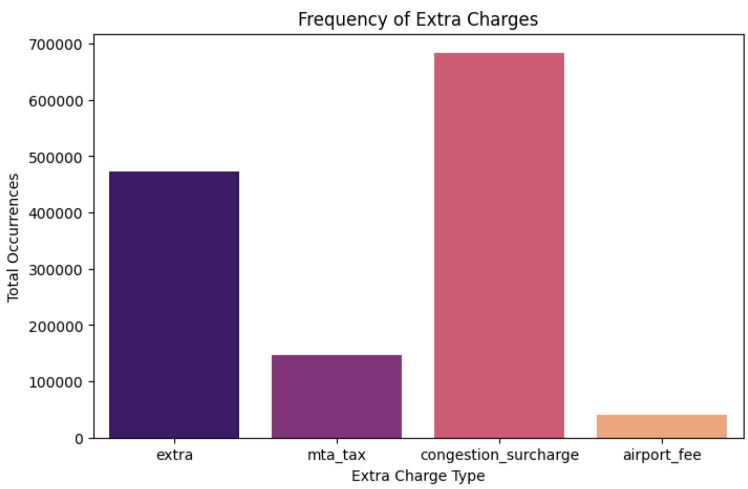


The average count in all these zones is close to 1 indecating that most rides here are single/ solo rides.

Newark airport, college point, and co-op city being in this list suggest that these places get mostly individual travellers.

This might also be where public transport is less efficient and so people depend on cabs more even though they are travelling solo.

3.2.15.1 Find out how often surcharges/extra charges are applied to understand their prevalence



Airport fee is pretty less occurring when compared to the other surcharges.

The highest and most occurring surcharge is congestion surchargewhich is nearly occurred abiut 700k times

Which is then followed by extra charges mostly for long distances etc.

MTA tax is applied less often but is still high when compared to airport fee.

	extra
zone	
LaGuardia Airport	10033
Midtown Center	8841
Upper East Side South	7951
JFK Airport	7417
Upper East Side North	6547
Midtown East	6381
Times Sq/Theatre District	6297
Lincoln Square East	6108
Penn Station/Madison Sq West	5780
Midtown North	5402

When looking at different zones, laguardia airport has the most number of extra charges applied

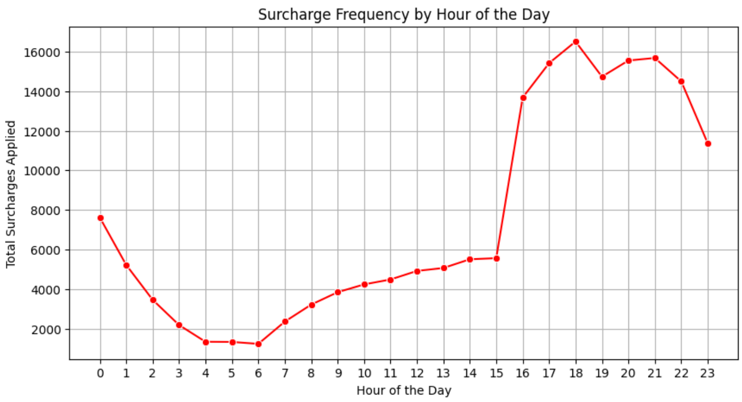
Midtown center, upper east side and JFK airport also have high occurances of extra changes overall.

The rest seem to be tourist heavy zones like the times square, Lincoln square and penn station.

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

```
Top 10 Pickup Zones with Extra Charges:
 zone
LaGuardia Airport               10033
Midtown Center                   8841
Upper East Side South            7951
JFK Airport                      7417
Upper East Side North            6547
Midtown East                     6381
Times Sq/Theatre District        6297
Lincoln Square East              6108
Penn Station/Madison Sq West     5780
Midtown North                    5402
Name: extra, dtype: int64

Top 10 Dropoff Zones with Extra Charges:
 zone_dropoff
Upper East Side North        7462
Upper East Side South        6966
Midtown Center               5850
Murray Hill                  5285
Times Sq/Theatre District    5249
Upper West Side South        5106
Lincoln Square East          4905
Lenox Hill West              4874
Clinton East                 4671
Midtown East                 4455
Name: extra, dtype: int64
```



Analysis of Extra Charges by Zone and Time of Day

- Pick up zones with most extra charges are airports, and tourist heavy places like JFK, LaGuardia, Times Square, Penn station etc.
- Dropoff zones with the most extra charges seen to be upper east side north and south, midtown and times square.
- Extra charges by the hour of the day:
    - Surcharges tend to peak during 4 to 10 PM with a sharp rise at 4PM.
    - A second peak is seen at midnight 12AM
    - The lowest is recorded between 3 and 6 AM

STEP 4: CONNCLUSION

4.1 Final Insights and Recommendations

Conclude your analyses here. Include all the outcomes you found based on the analysis.

Based on the insights, frame a concluding story explaining suitable parameters such as location, time of the day, day of the week etc. to be kept in mind while devising a strategy to meet customer demand and optimise supply.

Peak Demand Hours are as follows:

- 8 AM - 10 AM, 5 PM - 8 PM (Weekdays)
- 10 PM - 2 AM (Weekends)

High Revenue Zones:

- Midtown Manhattan
- Financial District

- Upper East side
- Times Square
- Airports

Pricing Optimization:

- Nighttime fares provide a lot to revenue

Operational Efficiency:

- positioning should match high demand areas

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

- As mentioned above, using the peak hours and the calculated sow routes can help with efficiency.
- Check for traffic congestion and average speeds which will tell exactly how to optimize routing to a particular place.
- Knowing that longer trips reduce tips, it would be best to make sure to take the shortest route to increase the chances of getting a higher tip.
- Also, reducing empty return trips by rerouting cabs to places that have a high drop and pick up rate will help or by showing the nearest high pick up zone will also increase changes of getting higher revenue.

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.

- Opt for the high revenue zones mentioned above by stationing cabs at those locations more often than not.
- Most of the night time revenue is obtained from entertainment hubs and tourist heavy areas.

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

- Night time fare optimization will help in increasing the overall revenue by increasing the fares after 11PM thus creating a balance of supply and demand.
- Increasing the prices slightly at peak hours during weekdays will help in increasing the revenue for the day without hurting the frequency of customers.
- Offering off peak hour discounts, where normally lesser rides are available, will increase the customer base by attracting people toward the slashed prices.
- Offering premium services to business travellers at higher prices will increase demand, revenue and customer trust.
- Having fixed prices to airports will also help with increasing transparency with the customer base and attract new people toward the cabs.