

Report: Optimizing NYC Taxi Operations

Note: The dataset, shapefile, and **data dictionary (2022)** were provided by the institution. No additional datasets or external data dictionaries were used in this analysis.

Abnormality Data Frame (df_abnormality): A dedicated data frame was created to store the trip records which I flagged as anomalies. This was done to preserve the data instead of discarding it.

1. Data Preparation

The necessary libraries were imported and *gg-plot style* was used for my plots.

1.1. Loading the dataset

I figured that it wouldn't be possible to work with such a large data set, after analyzing a single month's parquet file. Indicating a need for working with a sample of the data.

1.1.1 Sample the data and combine the files

A small percentage of sample was taken from every pickup hour, every pickup day and every pickup month and it was concatenated into a single data frame for further analysis.

Note: I decided to work with **10% random uniform sample of the data** to preserve the trends and to get the right insights from the data.
Estimates of the population are based on the sample.

2. Data Cleaning

Upon inspection I found that the data was unorganized, so I sorted the data, based on the pickup date and time before I could start the Data Cleaning process.

2.1. Fixing Columns

2.1.1. Fix the index

I had to reset the index and the column "store_and_fwd_flag" was identified as it would not add any value to my analysis. Therefore, I dropped this column after analyzing its contents. Additionally, I dropped the "Pickup Day" and "Pickup Hour" columns to avoid overcrowding.

2.1.2. Combine the two `airport_fee` columns

The Two Airport Fee columns were combined into a single column, if the value was present in the first column, it would retain it and if it wasn't present in the first column it would take the value from the second column. A new column was derived. The initial two airport fee columns were dropped.

2.1.3. Fix columns with negative (monetary) values

The column with monetary values was analyzed and the rate code id associated with it was analyzed too. A discovery was made that Standard Rate contributed to the highest negative monetary value.

Six columns were identified with negative monetary value that was a total of **192 rows** had negative value, accounting for **0.0051%** of the data.

Instead of discarding these observations, I assigned it to the data frame I mentioned at the start of my report. To preserve the trips and details of the trip for further investigation.

2.2 Handling Missing Values

2.2.1. Find the proportion of missing values in each column

A loop was used to find the missing values in each column by iterating through the column names. The total number of the column was calculated in the loop and the information provided to us included the Column Name, Number of Missing Values and The Percentage of Missing Values.

Passenger Count, Rate Code ID, Congestion Surcharge and Airport Fee features had 1,30,732 missing values present in the sampled data, accounting for 3.45% of our sampled data.

2.2.2. Handling missing values in `passenger_count`

Upon analysis of the Passenger Count feature, it was discovered that the most frequent occurring value was 1, indicating Solo Passenger Trips accounting for roughly 75% of our sampled data. Approach used was to Impute the Missing Values with the most frequent occurring value.

A separate analysis of value counts revealed 59,469 trips with a passenger count of 0. Based on my domain knowledge Yellow Taxi Trips with zero passengers is not possible since it isn't a delivery service.

These zero-passenger trips were preserved in my anomalies data frame and omitted from my main data frame used for further analysis.

2.2.3. Handle missing values in RatecodeID

To address this issue I loaded the provided shapefile provided using geopandas and converted it to a pandas data frame. Since I wanted to deal with missing values, spatial information such as geometry was omitted while loading the data frame.

I performed two separate merges on Pickup Location ID and Drop-off Location ID. After performing the merge, a consistency check was done to ensure data consistency.

I was able to identify the most frequently occurring Drop Zone for its corresponding Rate Code ID. A dictionary was created where the Drop Location ID served as the key, and the Rate Code ID served as the value. The dictionary was then leveraged to address the missing Rate Code ID issue.

Note : Based on domain knowledge and to ensure consistency, missing values were imputed using the most frequent rate code (mode) per Drop Location ID. Since the Rate Code ID is defined as the final rate code at the end of a trip suggesting a direct link to Drop-Offs.

Rate Code ID 99.0 was present in our data which wasn't present in our data dictionary. The dictionary I created was leveraged to address this issue as well. Similarly, Airport Zones were matched with their respective Rate Code ID.

The number of missing values was reduced quite a bit, the remaining missing values and Rate Code ID 99.0 which wasn't a valid rate code were imputed with the most frequent Rate Code ID that was 1 which represented Standard Rate.

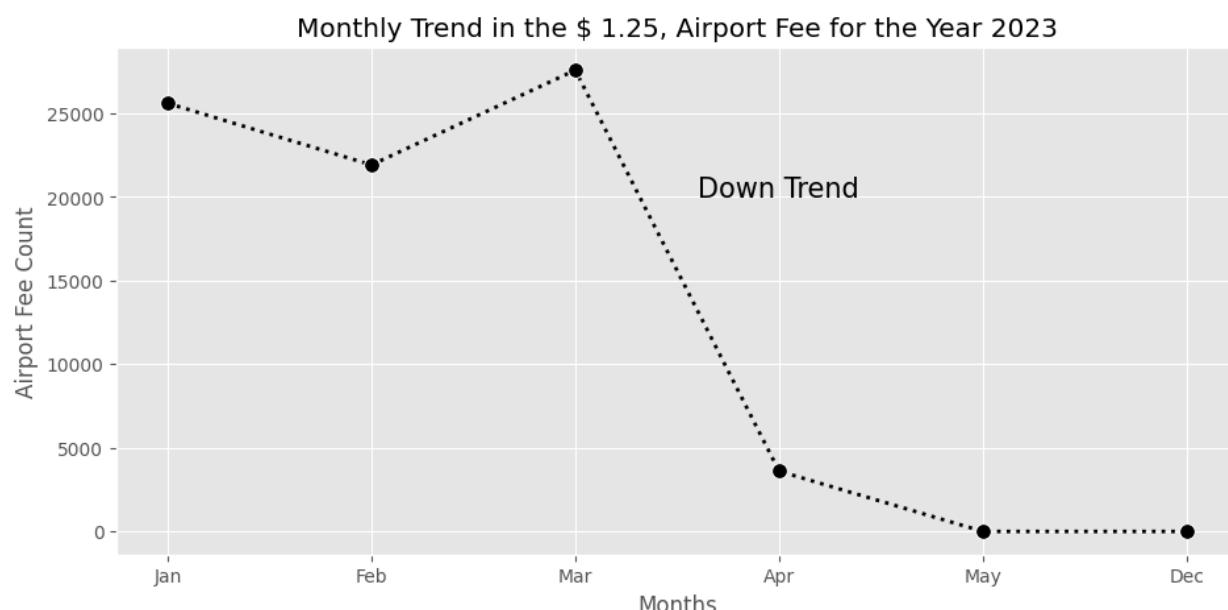
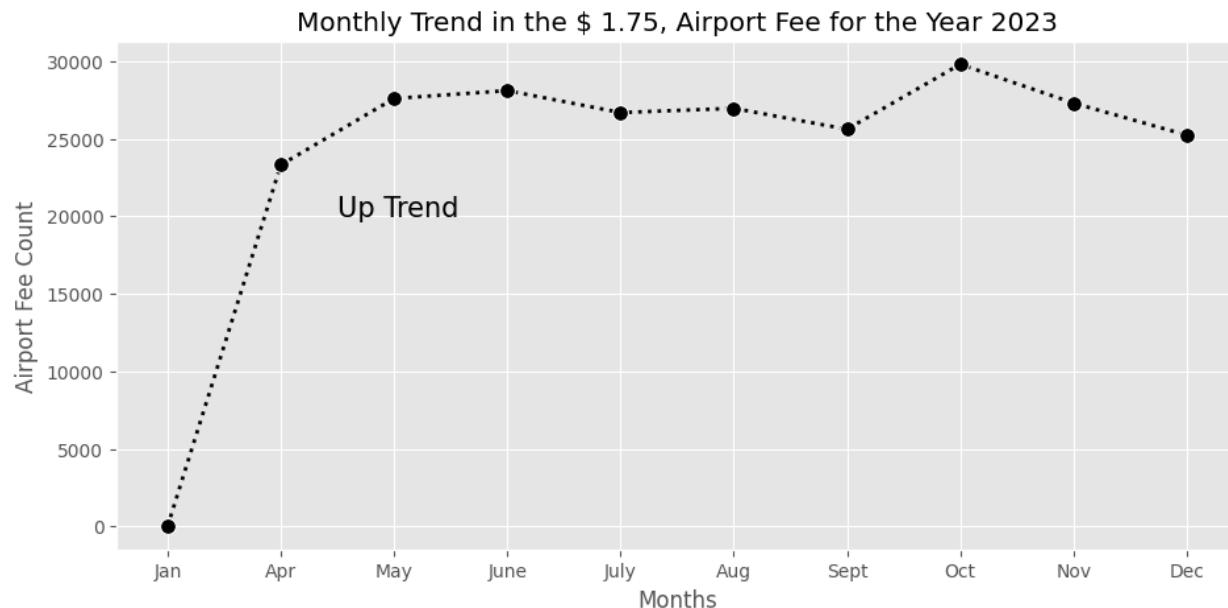
2.2.4. Impute NaN in congestion_surcharge

The most frequently occurring congestion surcharge in New York State was \$2.50 for Yellow Taxis, which accounted for 92.2% of our data. This suggests that \$2.50 is a standard congestion surcharge. Additionally, our analysis indicated that 7.7% of the trips had a congestion surcharge of \$0.0, suggesting that some routes are free from congestion surcharge. Mode Imputation was performed to address the missing values, i.e. \$2.50 was imputed into the missing values.

2.2.5. Handling Missing Values in Other Columns

As per our analysis on the proportion of missing values, we know that Airport Fee feature had missing values too. From our data dictionary, I know that the Airport Fee for taxis is \$1.25 for pick-ups only at LaGuardia and John F. Kennedy Airports.

But our value counts revealed that \$1.75 had a 6.67% occurrence compared to \$1.25 which had a 2.18% occurrence in our sample data. Upon further investigation I found patterns and trends which gave me a strong indication that there was policy change in the month of April which can be observed in our plots.



On the basis of our discovery, imputation was done in a manner to ensure data consistency. Missing values from January to March inclusive, and its corresponding Airport Drop Location ID of LaGuardia and John F. Kennedy Airports were imputed with USD1.25. Missing values from April to December were dealt in the same manner but imputed with USD 1.75.

The remaining missing values were imputed with the most frequently occurring value after cross-verification i.e. USD 0.00, which had a 91.1% occurrence suggesting Pickups that didn't happen at the Airports where the fee was applied.

2.3. Handling Outliers and Standardizing Values

Note: An outlier prediction function was defined which uses the Interquartile Range (IQR) approach a statistical measure to find trips with possible outliers. This technique is frequently used to find odd patterns, trends and possible outliers.

2.3.1 Identifying outliers in Passenger Count

To assess the magnitude of the issue, analysis was carried out which revealed that only 42 trips had an unreasonable number of passengers, i.e. above 6 passengers. These trips were preserved by adding them to my anomalies data frame and were omitted from my primary data frame for further analysis.

2.3.2 Identifying outliers in Vendor ID

A potential outlier / misleading information was identified where the Vendor ID was 6 which wasn't present in our data dictionary. A value count analysis revealed that it accounted for only 0.023% i.e. 884 trips. So, I added it to my anomalies data frame and omitted it from my primary data frame used for analysis.

2.3.3 Identifying outliers in Payment Type

Since our data dictionary made it clear that a value of 5 indicates "Unknown", handling the payment type was easy. To keep things consistent, we changed the misleading payment type value of 0 that we came across to 5. I was able to ensure that 0 was classified as "Unknown" by doing this simple imputation.

2.3.4 Identifying outliers in Airport Fee

As per our earlier analysis, a business-driven approach was taken on the Airport Fee feature, USD 1.25 was imputed for Pickup Zone's LaGuardia and John F. Kennedy Airports for January to March inclusive.

The remaining months were imputed with USD 1.75 where the Pickup Zone's were LaGuardia and John F. Kennedy Airports.

USD 0.00 was imputed for trips where the Pickup didn't happen in LaGuardia and John F. Kennedy Airports.

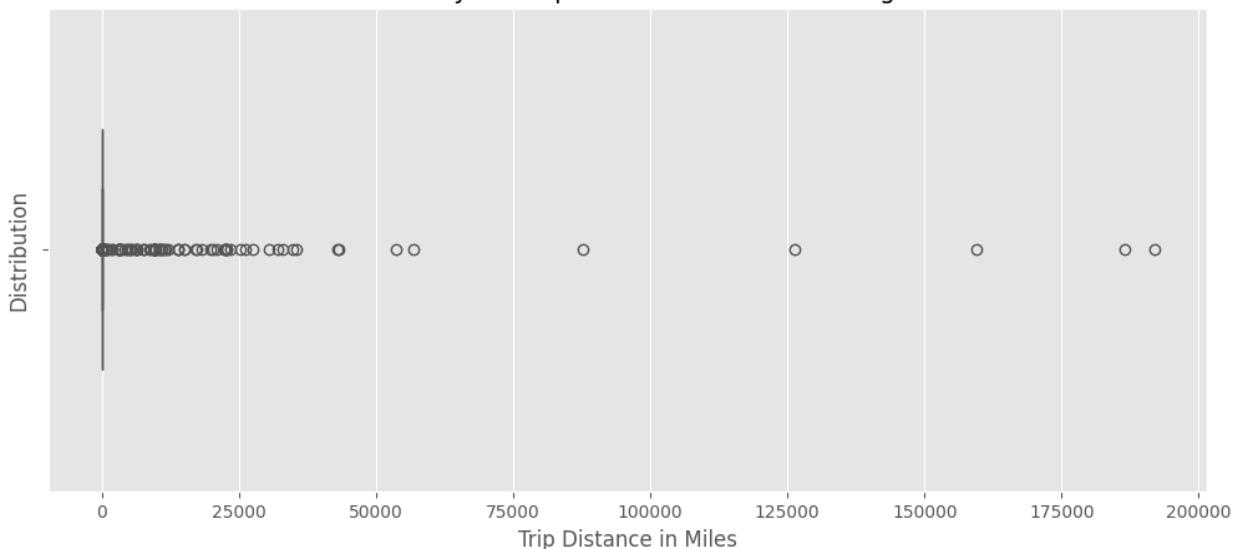
It is important to note that adjusting the airport fees to what we consider the correct amount could create outliers in the total amount column which I addressed at a later stage.

2.3.5 Identifying outliers in Trip Distance

Based on domain knowledge, the first outlier check performed on this feature was if Trip Distance was negative. A statistical analysis revealed that trip distances as high as 1,91,944.96 miles existed, which is also visible in our Box Plot indicating the existence of outliers. In the first box plot, we can see trips with a trip distance higher than 25,000 miles. These values are unrealistic and distort our distribution and analysis.

A cap was applied keeping trips within a range of (0 Miles to 250 Miles both being exclusive). The trips which were Zero Distance Trips or unusual trip distance ≥ 250 Miles was added to our anomalies data frame, keeping only those trips which meet our condition in the primary data frame.

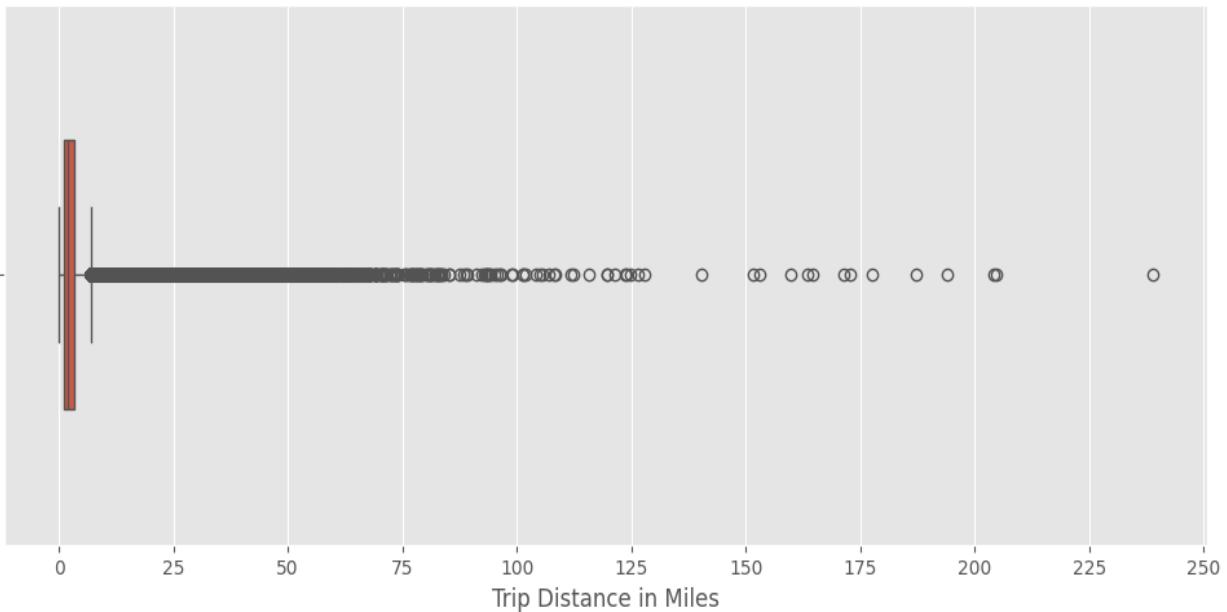
Box Plot to Study the Trip Distance before Handling Outliers



After handling the outliers, the second Box Plot, shows a much cleaner distribution. Majority of our trips lie within the range of Zero to Fifty Miles which is reasonable. By removing extreme outliers, there was a stronger correlation between trip distance and other features than what was observed before.

I decided to keep some rare cases where the Trip Distance was greater than or equal to 150 Miles and had a reasonable Fare Amount. They could be legitimate inter-city trips.

Box Plot to Study the Trip Distance Column After Handling Outliers



2.3.6 Identifying outliers in Pickup and Drop-Off Location ID

On the basis of the shape file provided to us, I was able to identify Pickup and Drop Location IDs which weren't present in the shape file zone's Location ID. 4 Location IDs were flagged as invalid and the reason behind it was mainly each Location ID should represent a Zone or it could lead to missing values while merging. Invalid Location ID Trips were added to our anomalies data frame and omitted from our primary data frame.

The list of invalid locations ids : 57, 105, 264, 265. "Invalid" here strictly means that they cannot be linked to zones and may affect our analysis.

2.3.7 Identifying outliers in Pickup and Drop-Off (Date Time)

Since we are analyzing trends in 2023 NYC Taxi Trips, a rule was made that Pickups that didn't happen in the Year 2023 were flagged as outliers and added to our anomalies data frame and omitted from our working data frame. This rule was made taking into consideration trips that happened on 31st March 2025 and whose drop happened on 1st January 2025.

Unusually, long duration trips were dealt with by keeping only those trips whose trip duration was less than a day in our primary data frame and adding trips with a trip duration more than or equal to a day to our anomalies data frame.

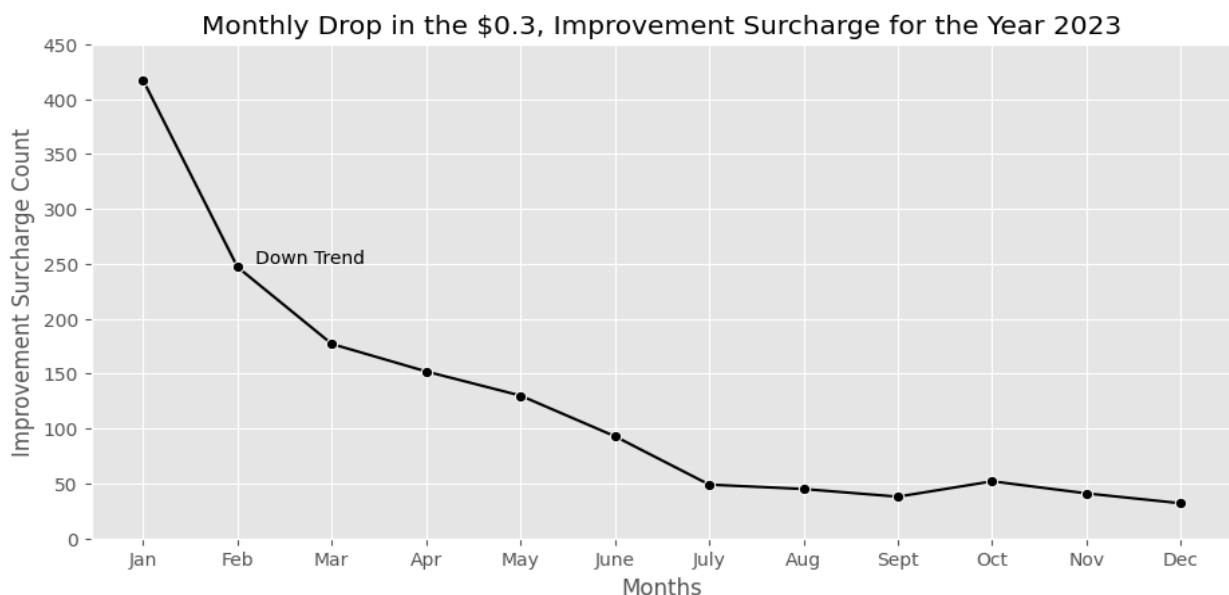
2.3.8 Identifying outliers in Tax on Meter (MTA Tax)

As per our Data Dictionary, a \$0.50 MTA tax is automatically triggered based on the metered rate in use.

In 18,125 Trips MTA Tax wasn't triggered as per the NYC MTA Rules suggesting that the trips could be outside NYC or it could be data quality issues. These trips were added to my anomalies data frame and omitted from my primary data frame which I'll be using for further analysis.

2.3.9 Identifying outliers in Improvement Surcharge

Note : "Our data dictionary indicates that a \$0.30 improvement surcharge is assessed on trips at the flag drop, which began in the year 2015."



A value count analysis revealed that USD 1.00 Improvement Surcharge was levied in 99.95% of the trips which opposes what we have in our data dictionary. USD 0.30 Improvement Surcharge, which was present in our data dictionary, only accounting for 0.04% of the trips.

Upon further investigation, of the \$0.30 improvement surcharge, our plots strongly indicate a policy change in the Year 2023. So, both \$1.00 and \$0.30 were considered as valid amounts for the analysis.

\$0.00 was also present in a few trips (roughly 24 trips in our sampled data), those trips were added to my anomalies data frame and omitted from my primary data frame.

2.3.10 Identifying outliers in Miscellaneous / Extras

As per my business understanding, I established a rule that any value that is not a multiple of 0.25 should be considered an outlier. The number of trips that do not follow our condition was 137 trips which was very small in number. Wherever the condition wasn't met, those trips were added to my anomalies data frame. My primary data frame omitted those trips where the condition wasn't met.

2.3.11 Identifying outliers in Tip Amount

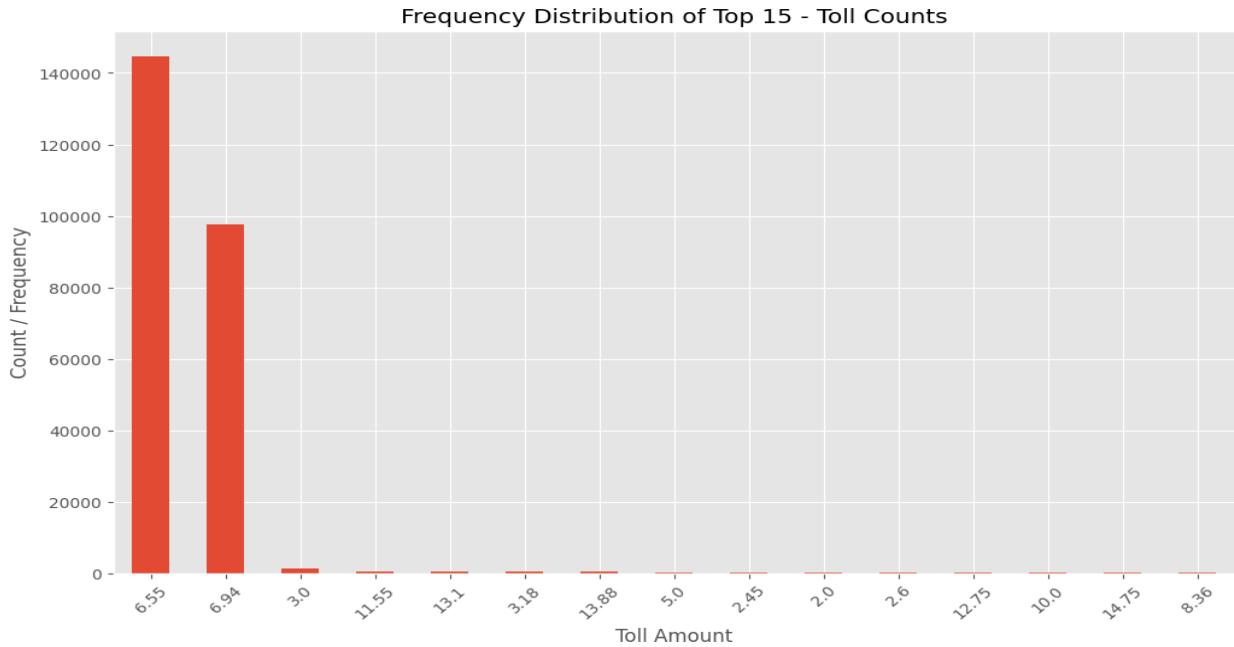
A statistical analysis of the feature reveal tip amounts as high as USD 480.50 and a median of USD 2.88. While it might be possible for a tip amount to exceed the fare amount, it is a very rare occurrence and could be considered misleading. These trips were added to my anomalies data frame and omitted from my primary data frame used for analysis.

To check the effect of our earlier step another statistical and value count analysis revealed unusually high Tip Amounts with a very less frequency of occurrence. The 99th percentile statistical measure was applied which gave me a threshold of USD 16.79

2.3.12 Identifying outliers in Toll Amount

A statistical analysis revealed that 99% of our data lies below the USD 6.94 threshold for the Toll Amount feature. The Bar plot below shows the distribution of various Toll Amounts excluding USD 0.00 since it was present in majority of the trips (3,297,282 Trips as per our sample data).

To handle outliers, a temporary Toll Frequency column was derived from the value count analysis on the Toll Amount feature. Toll frequency less



than or equal to 50, suggests a very low occurrence and could suggest data entry issues. *The above counts are based on 10% sample.*

These trips which didn't meet our conditions, were preserved in our anomalies data frame, and omitted from our primary data frame used for analysis.

2.3.13 Identifying outliers in Fare Amount

A statistical summary revealed presence of fare amount as high as USD 143,163.45 . A quantile analysis revealed that 99% of our data had a Fare Amount less than \$77.60, and 99.9% was less than \$150.00.

Further analysis revealed that the number of trips where Fare Amount is less than USD 2.5 is 437 Trips. The number of trips where Fare Amount is greater than \$150 is 58 Trips. This was a relatively small number.

I capped the feature by keeping only those Fare Amounts which were greater than or equal to USD 2.5 or less than or equal to USD 150. This was done to avoid any misleading information or potential outliers. The trips which didn't meet our expectations were added to my anomalies data frame.

2.3.14 Identifying outliers in Total Amount

I decided to take a strategic approach by dealing with the Total Amount feature at the end since any previous imputations could affect this feature. I created a new total amount column by adding all the monetary values for each trip to get the accurate representation of the Total Amount of each trip which would be useful for Revenue Analysis at a later stage. The old Total Amount column was dropped to avoid confusion or misleading information.

3. Exploratory Data Analysis

Before the EDA Phase could begin, a few initial steps were taken to ensure Data Integrity and Consistency. Data Types for the columns were checked; Column names were standardized where needed. Derived columns were created with their respective names for categorical columns disguised as numeric for visualization and readability purpose. This step was taken to ensure we have a readable and well-organized dataset.

3.2. General EDA: Finding Patterns and Trends

3.2.2. Classify variables into categorical and numerical

Categorical Features : VendorID, RateCodeID-EndofTrip, PULocationID, DOLocationID PaymentType and Pickup Hour

Numerical Features : PassengerCount, TripDistance, Trip Duration, FareAmount, Miscellaneous/Extras, TaxOnMeter, TipAmount, TollAmount, ImprovementSurcharge, CongestionSurcharge, Airport Fee and Total_Amount

Date Time Objects : PickupDateTime and DropoffDateTime.

3.2.3. Analyze the distribution of taxi pickups by hours, days of the week, and months

For the purpose of this analysis three new columns were derived for Pickup Hours, Days of Week and Months to analyze trends and patterns over time, what we call temporal analysis.

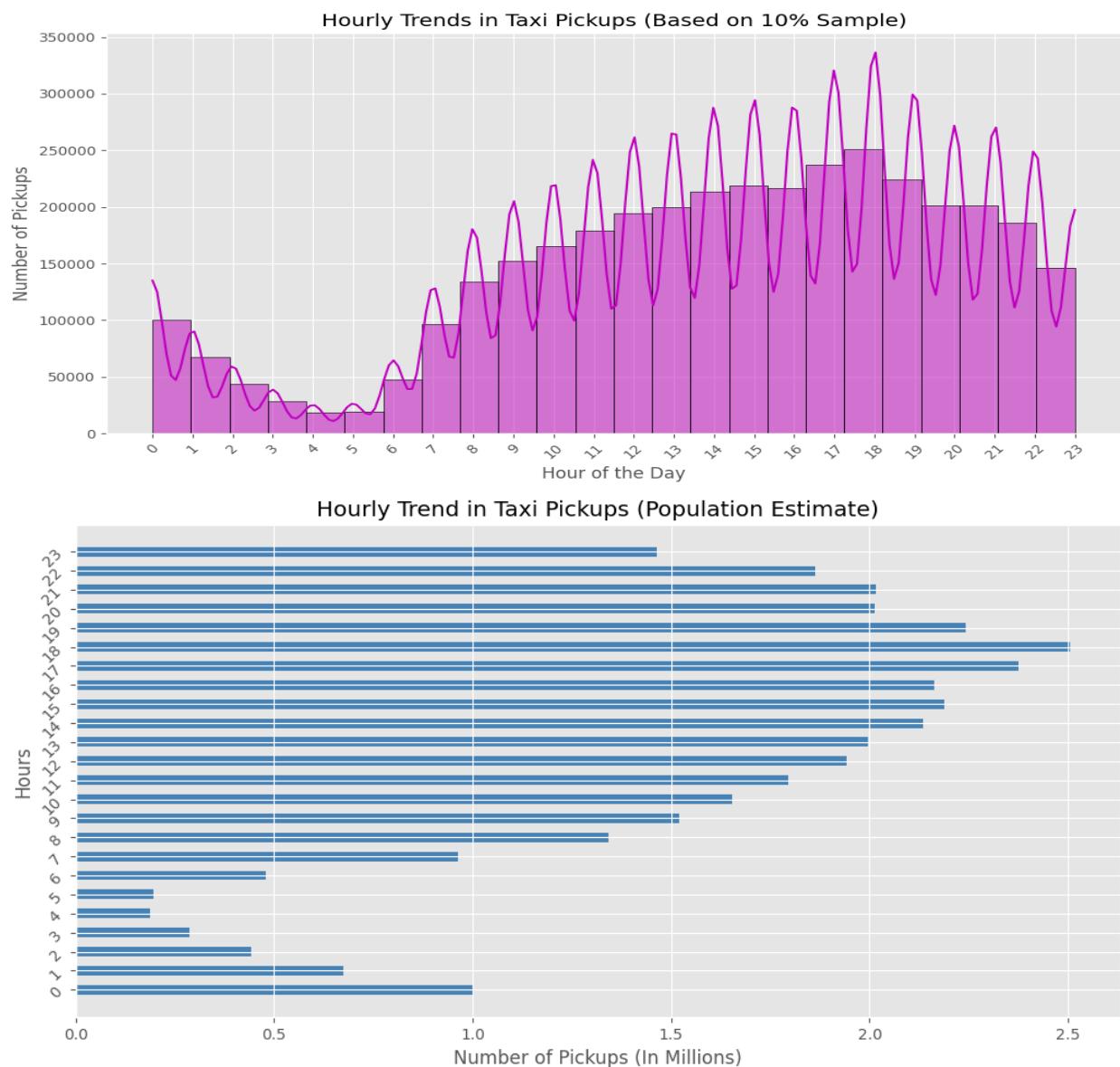
3.2.3.1. Hourly Trend Analysis

Demand starts to decline after midnight i.e. after 12AM. We can see a steady decline up until 5AM.

The Hourly Analysis suggests that the Hourly Demand is more consistent throughout the day and declines at night.

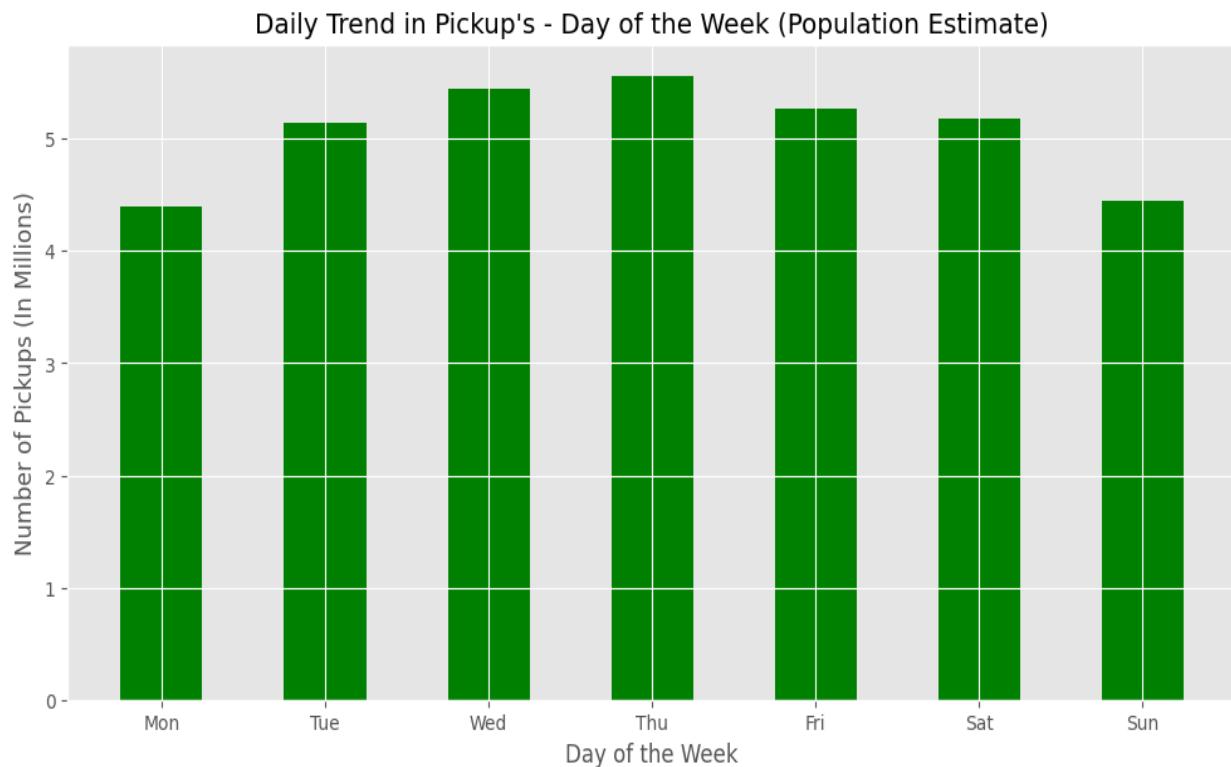
Peak Hours observed are from 5PM to 7PM can be considered as rush hours suggesting a demand for Taxi Cabs post work.
11PM to 5AM can be considered as quiet hours.

We can conclude that the hourly trend suggests that demand is not uniform throughout the day as there is sharp dip after 11PM.



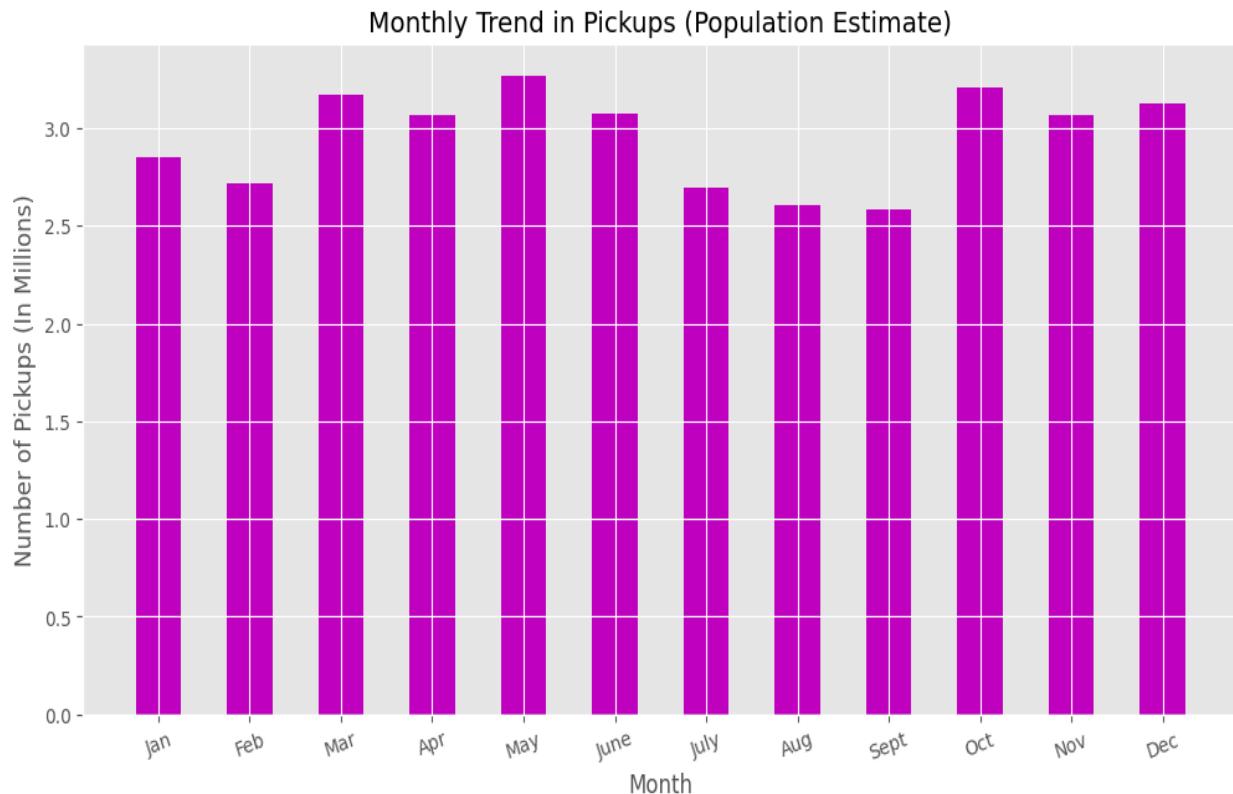
3.2.3.2. Analysis based on Day of the Week

- The most common observation we can make from our plot is that people prefer to rest of Sunday's.
- Monday can be considered a slow start; the demand is almost as slow as on Sunday.
- Thursday can be observed as the Peak Day and Sunday as a Slow Day.
- The Midweek demand is pretty good and strong, with very little fluctuation. Midweek is almost balanced out.
- Saturday being a holiday in the US still has a pretty decent demand.



3.2.3.3. Analysis based on Monthly Trend

- The plots reveal that Monthly Trips, that is Taxi Demand, are higher during the first half-year period. Gradually decreases after June and rises again from October.
- The demand for Yellow Taxi Cab's seems to be the strongest in the month of May. The demand seems to be low during the Summer Months in NYC i.e. July and August.



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

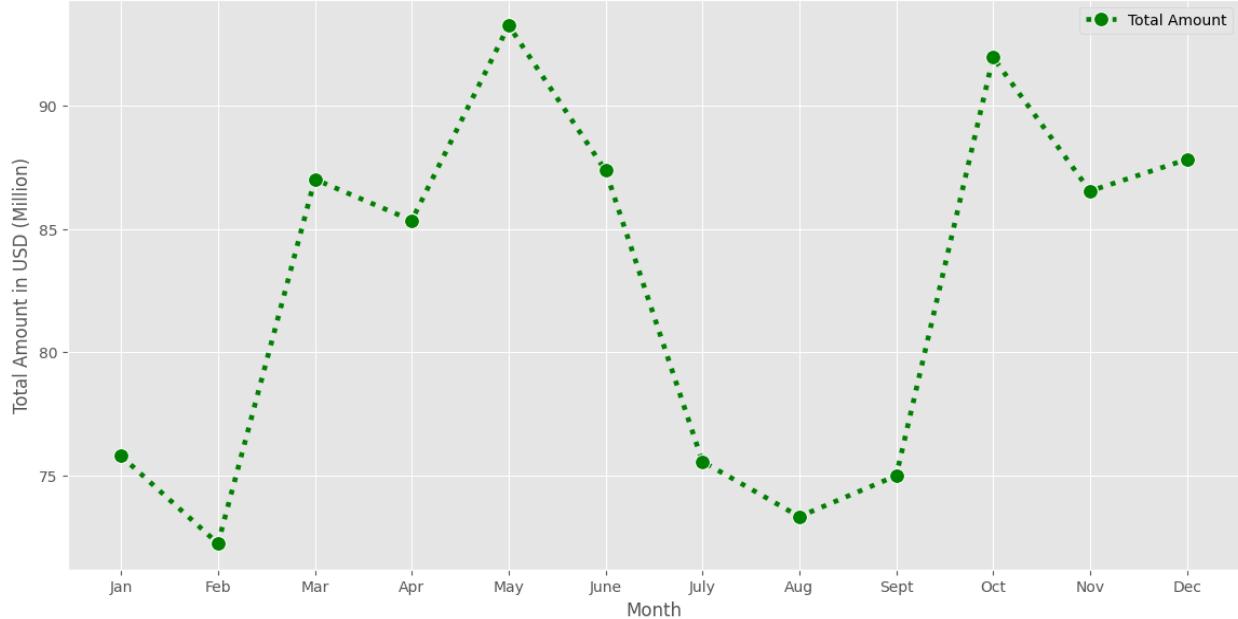
A sanity check was performed to make sure these columns do not contain any zero or negative values.

Negative Amounts, Zero Distance Trips, Fare Amount, Tip Amount and Total Amount were taken care of during my Data Cleaning and Outlier Handling step. A brief description has been provided for the same in the earlier steps.

3.2.5. Analyze the monthly revenue trends

From the plot we can observe that February and August seem to have lower activity due to which the revenue is the lowest and a significant dip is observed. May and October seem to be the most profitable month in terms of revenue. A majority of the months lie above the 85 Million Revenue margin, a few months however lie below 80 Million Revenue.

Estimated Monthly Revenue Trend - Year 2023 - NYC (Population Estimate)

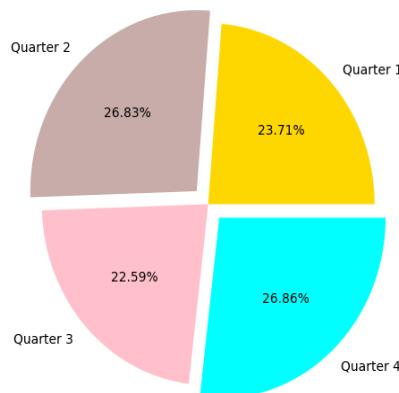


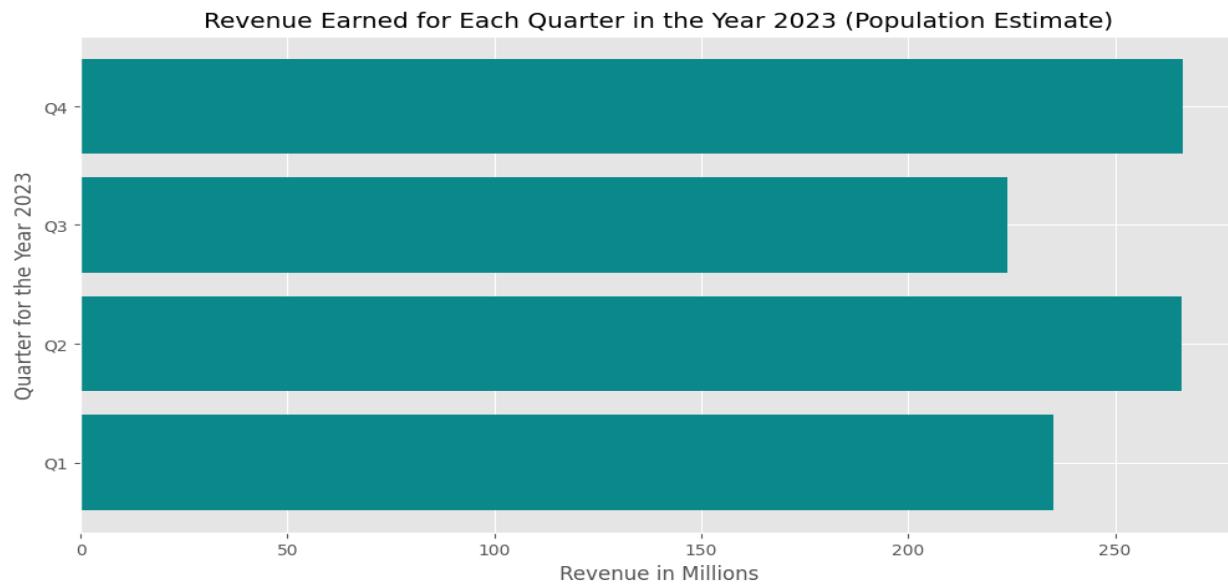
3.2.6. Find the proportion of each quarter's revenue in the yearly revenue

These are the estimated population revenue quarterly based on our sample. Quarter 4 and Quarter 1 are relatively higher.

- Quarter 1 : 23.72% which is \$ 235.04 Million Estimated Population.
- Quarter 2 : 26.83% which is \$ 265.92 Million Estimated Population.
- Quarter 3 : 22.59% which is \$ 223.91 Million Estimated Population.
- Quarter 4 : 26.86% which is \$ 266.24 Million Estimated Population.

Visualizing the Proportion of Each Quarter for the Year 2023 (Population Estimate)

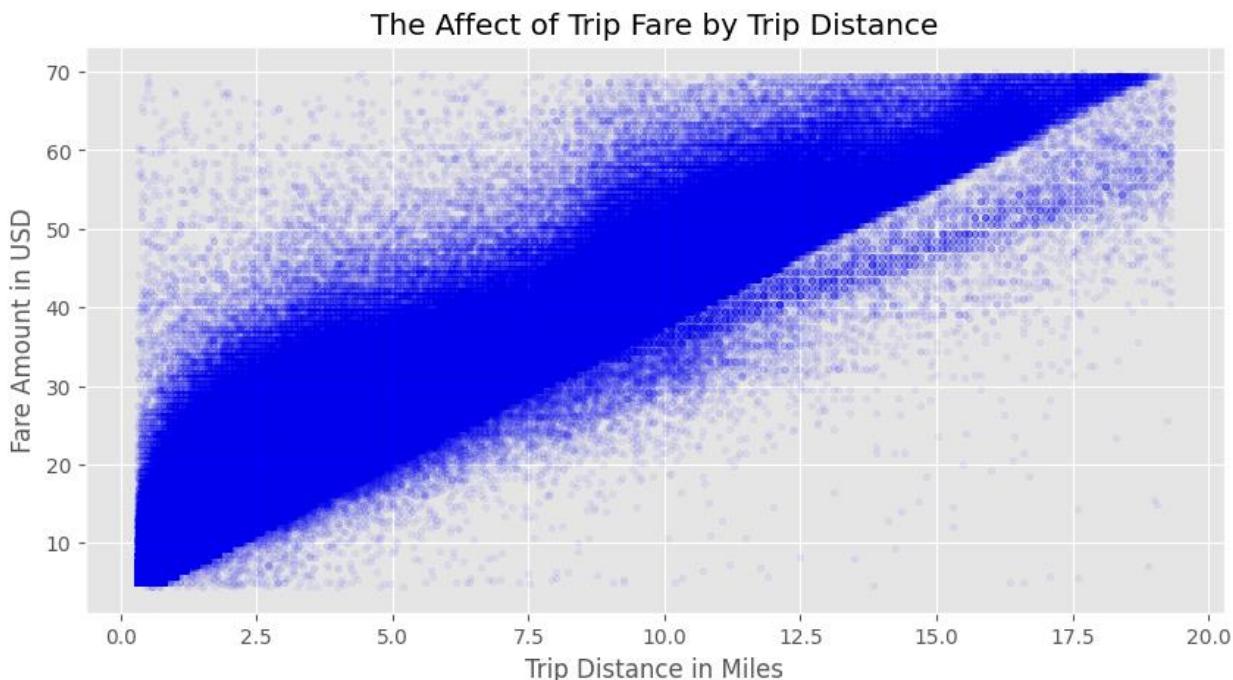




3.2.7. Analyze and visualize the relationship between distance and fare amount

Before I could perform a correlation analysis between the two features, I had to make sure that my data isn't distorted by data entry issues or any potential outliers. A cutoff was decided using a statistical measure 0.01 percentile and 0.99 percentile.

The cutoff wouldn't just help us find the correlation, but it would play an essential role in the visualization process as well. A correlation of 0.95 was



found between the two features, suggesting that there was a strong positive correlation between the features.

In other words, we can say that as the Trip Distance increases the Fare Amount also increases, which is evident from our plot. We can observe a straight line going in the upward direction.

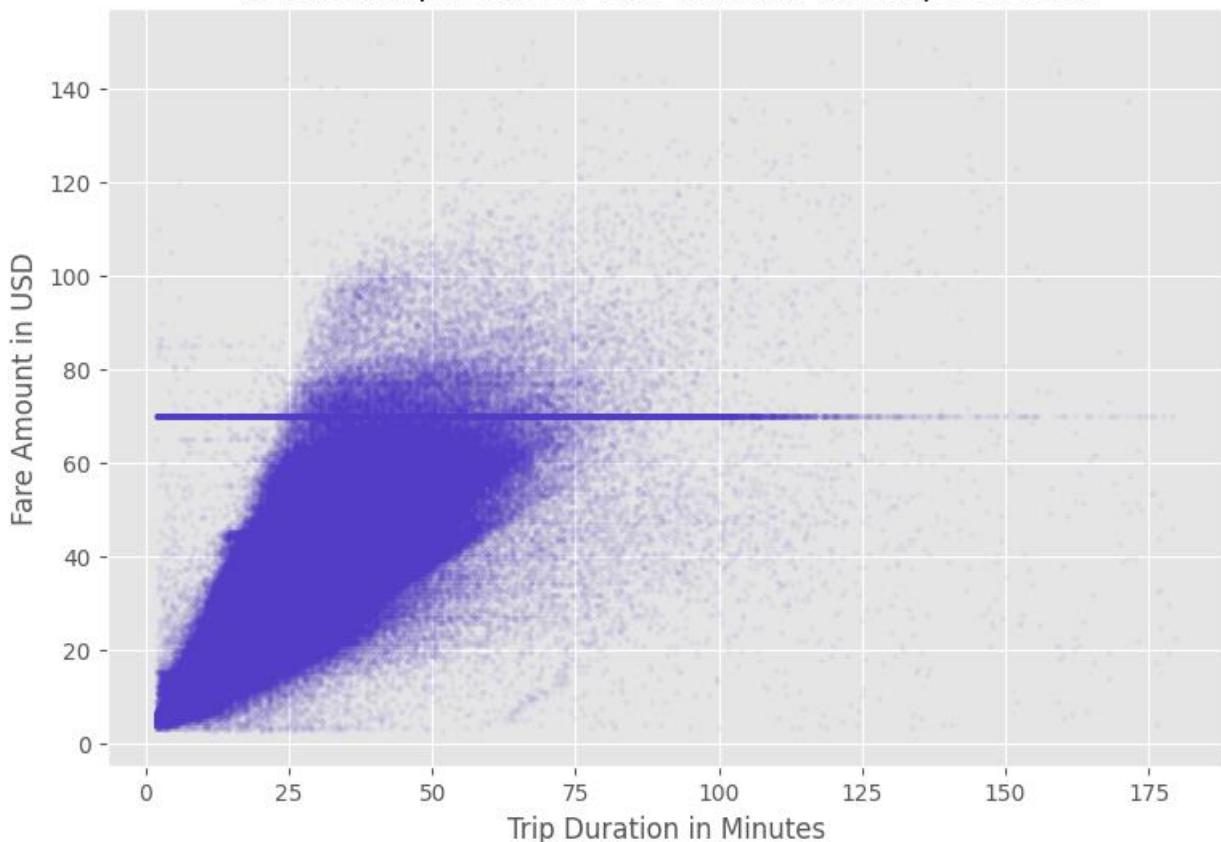
3.2.8. Analyze the relationship between fare/tips and trips/passengers

To find and visualize the correlation between fare amount and trip duration, I had to derive a new metric named Trip Duration. Upon deriving it, I took care of the negative trip duration and added it to my anomalies data frame and omitted it from working data frame.

3.2.8.1. Correlation between Fare Amount and Trip Duration

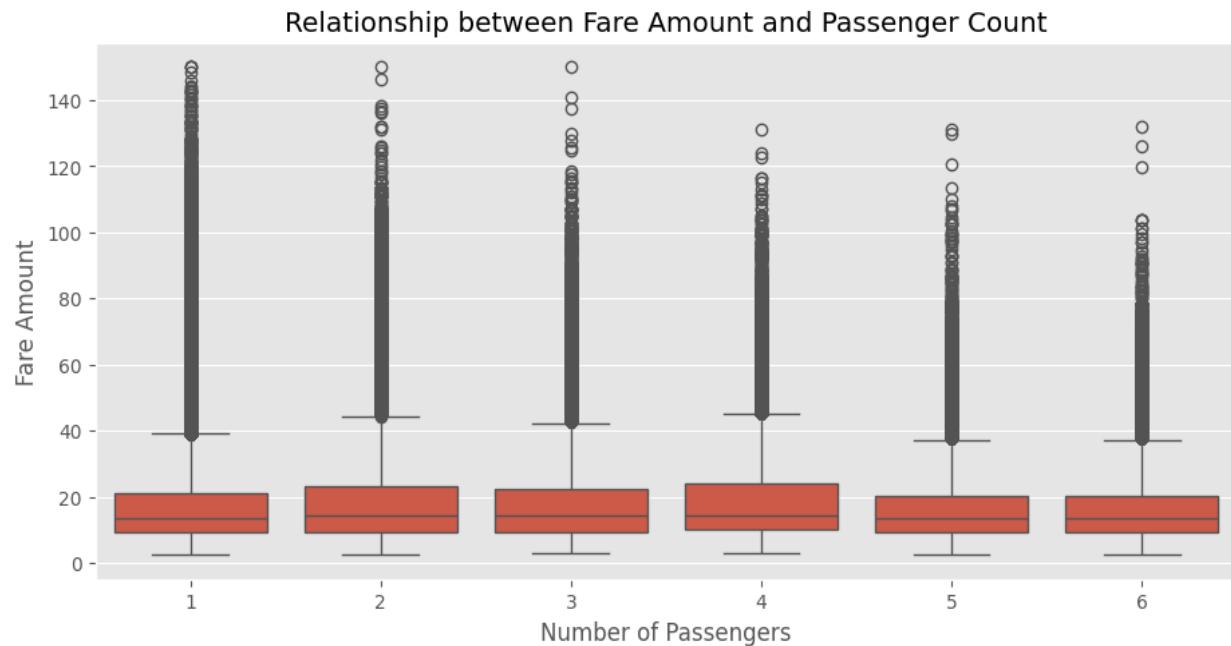
Working with a subset of the data to better understand the correlation between the two features, there was strong positive correlation of 0.88 approx. which suggests that as the trip duration increases the fare amount also increases. The scatter plot shows a strong positive linear relationship. Majority of our trips happen within an hour and cost up to USD 100. The straight line suggests a Flat Rate Fare.

Relationship between Fare Amount and Trip Duration

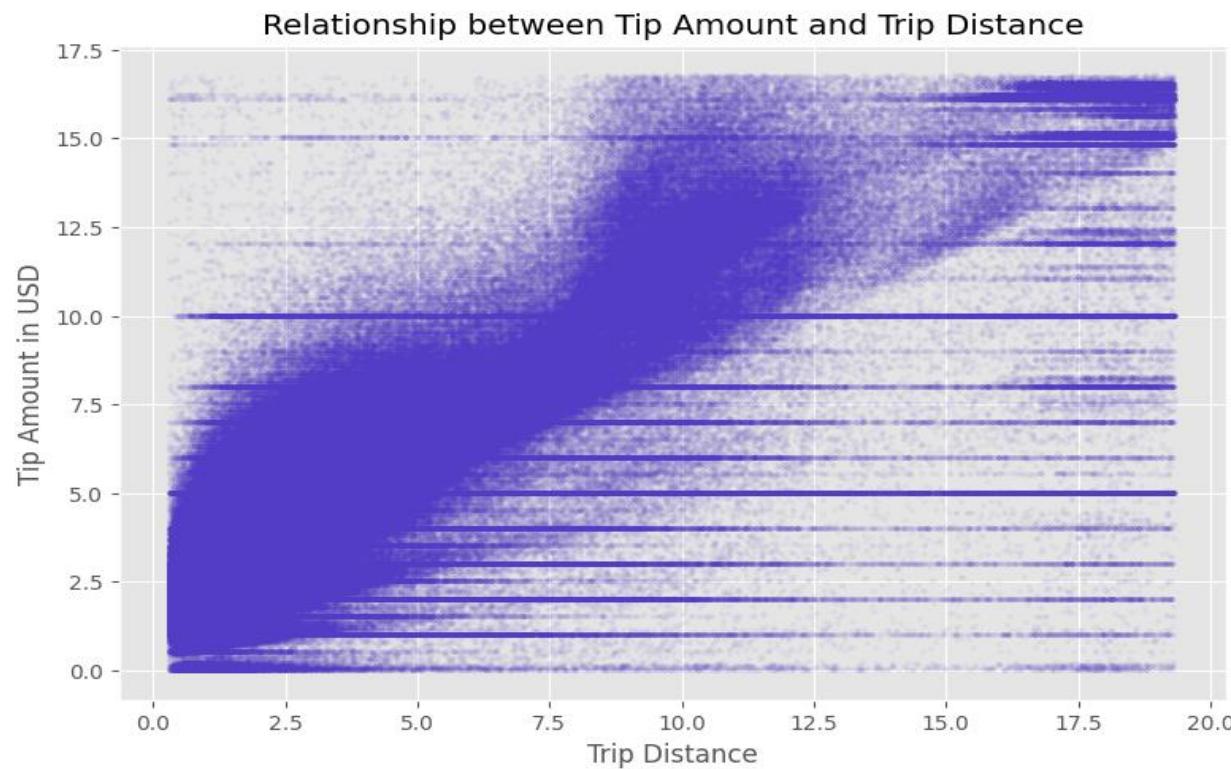


3.2.8.2. Correlation between Fare Amount and Passenger Count

The correlation between the two features were approximate 0.034 which is close to zero. It suggested that there is almost no correlation between the two features.



3.2.8.3. Correlation between Tip Amount and Trip Distance



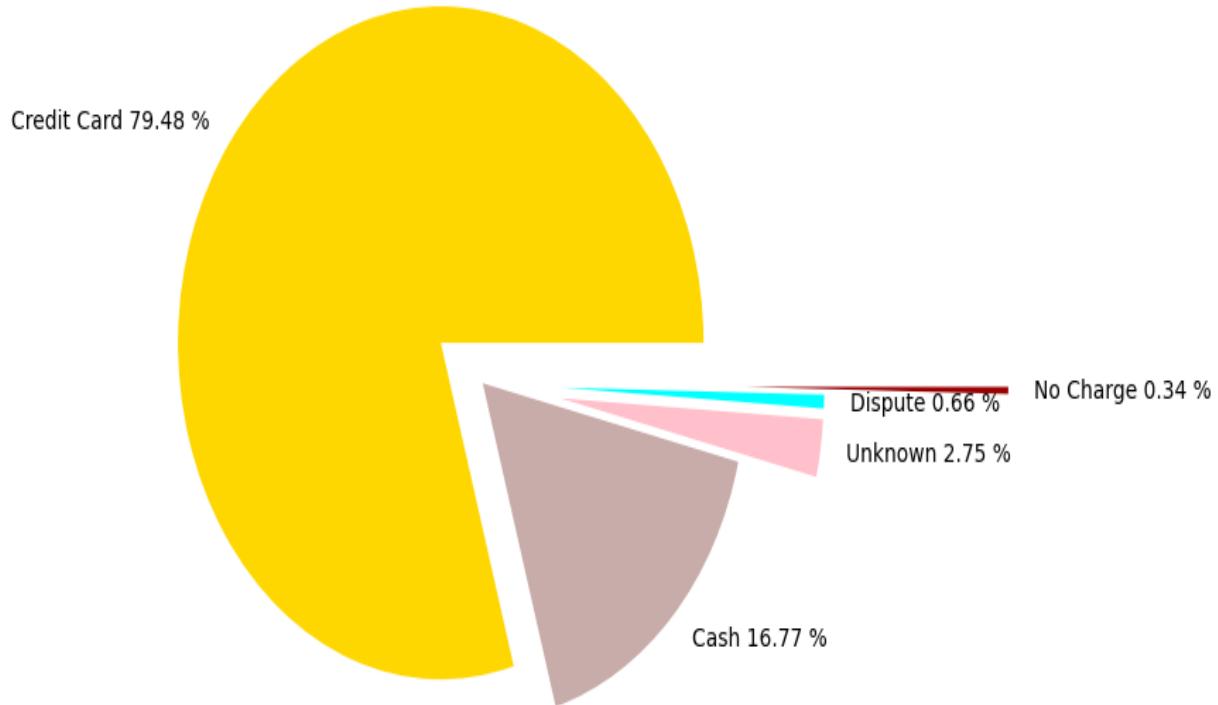
Working with a subset of the data to better understand the correlation between the two features, there was strong positive correlation of 0.82 approx. which suggests that as the trip distance increases, customers tend to tip more. From the scatter plot it is evident that they move in the same upward direction indicating a positive linear relationship. The lines on the scatter plot indicate people tipping the same amount more often.

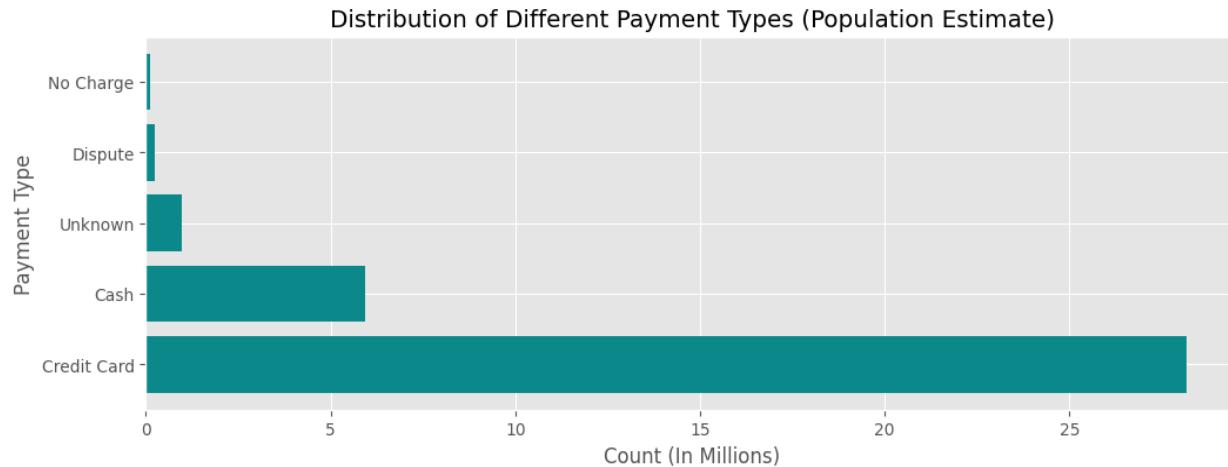
3.2.9. Analyse the distribution of different payment types

We plotted two plots to see the distribution and proportion of the different payment types. Pie Chart and Bar Chart. The Bar Chart clearly shows the distribution of the different payment types, and their count as per our sample data. The Pie Chart shows the proportion of the various payment types in percentages to better understand the distribution of the types of payment.

Credit Card is the most used payment method among all, represent 79.4% approx. Cash is the second highest payment method representing 16.8% approx. People tend to pay with credit card more often and there is hardly any dispute.

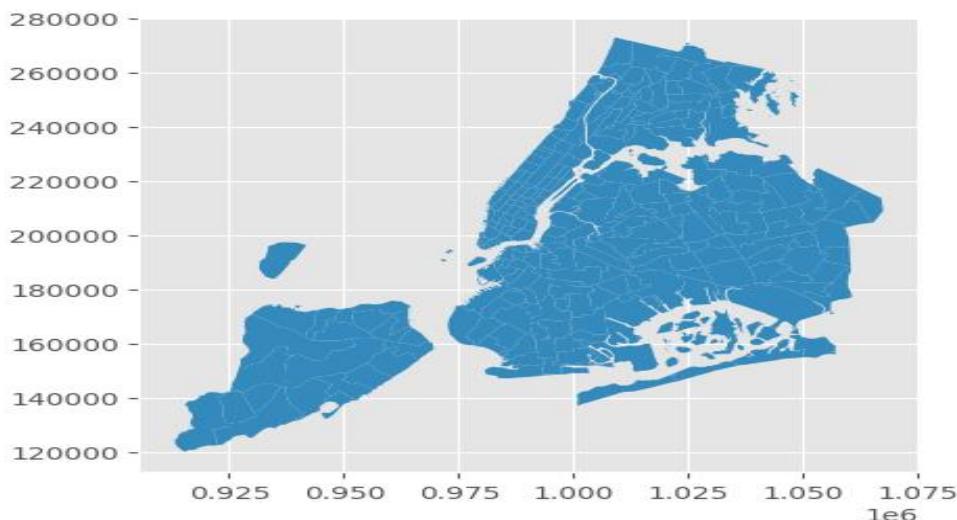
Visualizing the Proportion of Each Quarter for the Year 2023





3.2.10. Load the taxi zones shapefile and display it

geopandas is a python library used for geospatial analysis. The shape file was loaded and displayed via geopandas.



3.2.11. Merge the zone data with trips data

Two merges were required, firstly I merged the Pickup Location ID and the Zone Location ID. Secondly, I merged the Drop Off Location ID and Zone Location ID. This kind of merge would give me insights into the Pickup Zone and Drop Zone, and some more information like borough. This was a crucial step for our further analysis.

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

The merged data was grouped by the Pickup Location ID and a count operation was performed. This would give us the number of trips for each Location ID in the form of a pandas series.

3.2.13. Add the number of trips for each zone to the zones data frame

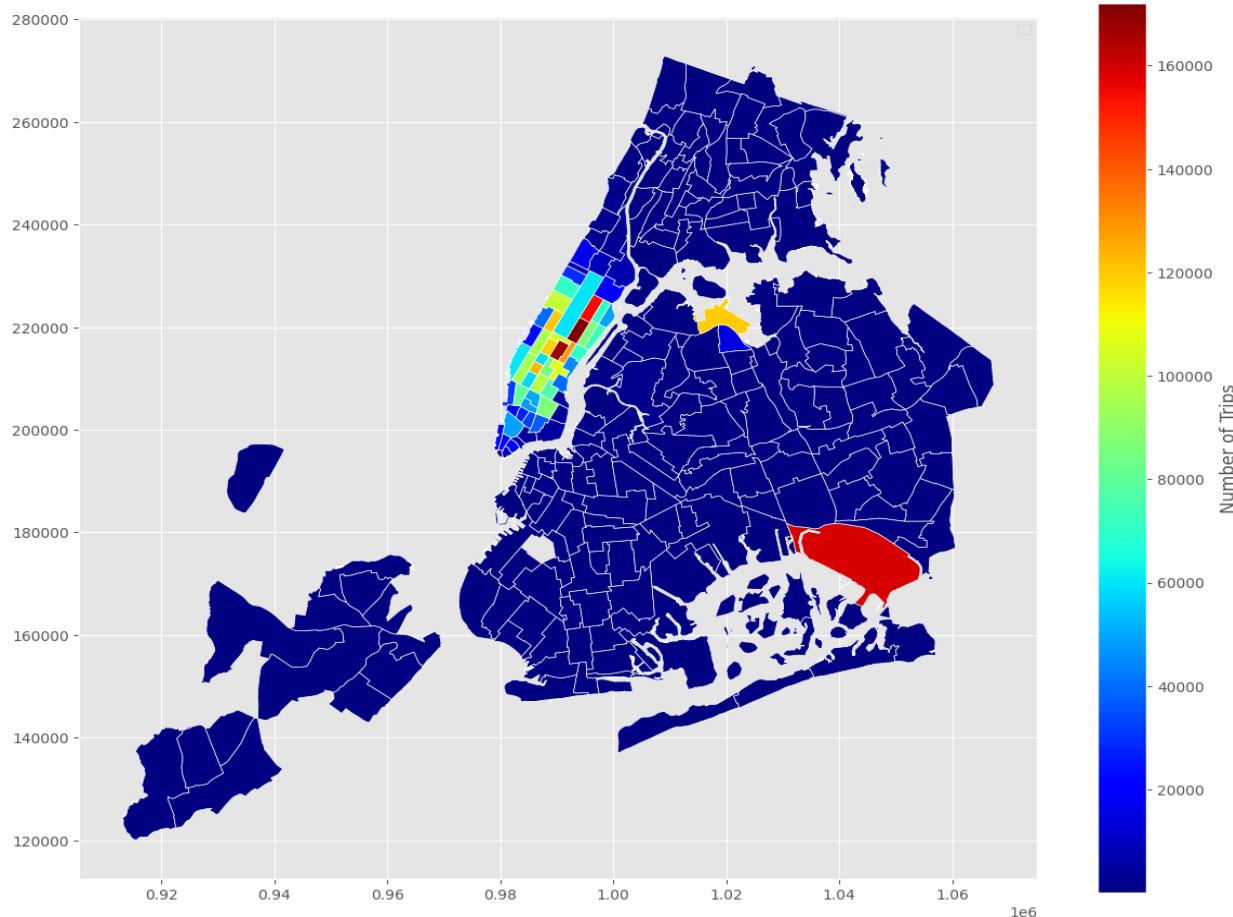
The number of trips which was computed in our previous step. Using the Zone's Location ID feature the series which was created in our previous step was mapped back to the Zones Data Frame.

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

The map is a color map, a visual representation of the map of the New York City. This visualization was performed in order to analyze which areas in New York City tend to have a higher demand and higher volume of Yellow Taxi Trips. The bar on the right displays the number of trips based on the 10% sample data. New York City is divided into five boroughs which can be seen on our color map.

- The Dark Blue color represents low number of trips. (*Low Demand Zones*)
- The Light Blue, Light Green, Yellow color represents moderate number of trips. (*Moderate Demand Zones*)
- The Orange and Red color represents the highest number of trips. (*Higher Demand Zones*)

Visualizing Zones by Number of Trips (Sample)



3.2.15. Conclude with results

Taxi Demand is seen to follow a clear trend or a pattern we can say based on our analysis. Our staffing model should be focused on deploying drivers during peak hours, peak days and peak months. Reducing the staff when the demand is low will save us on operational costs. Better Revenue Management.

Targeted marketing promotions can be done for the low demand months or seasons. We can focus on recruiting new staff for the higher demand months, during the low demand months. Increase the incentives of drivers during the high demand days, months, peak hours like evening time.

Our Fare structure is directly tied to trip duration and trip distance. We could keep a reasonable fare amount which would lead to customer satisfaction and profitability.

Helps in Resource Management on a Larger Scale based on the Monthly Trend Analysis. February and August can be predicted as months with lower activity managing supply during these months is crucial. Increasing the number of driver's for months when the demand is higher.

There is a strong positive relationship between the monthly trend in pickups and the monthly revenue generated. Quarter 2 and Quarter 4 are the highest contributors in terms of revenue share. Quarter 1 and Quarter 3 are significantly lower. Quarterly planning could be done to address this issue.

Common tipping behavior trends were observed, could be because of credit card being used the most people tend to swipe the card and tip the driver as well.

3.3. Detailed EDA: Insights and Strategies

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

I analyzed variations by time of day and location to identify bottlenecks or inefficiencies in routes. Pickup Zones, Drop Zones, Number of Trips, Average Mile Per hour and the Pickup Hour was observed.

Analysis revealed that the feature derived "Miles Per Hour" had unreasonably high miles per hour. So, I worked with a subset of the data, where the miles were below 100 Miles and Number of trips was above 24.

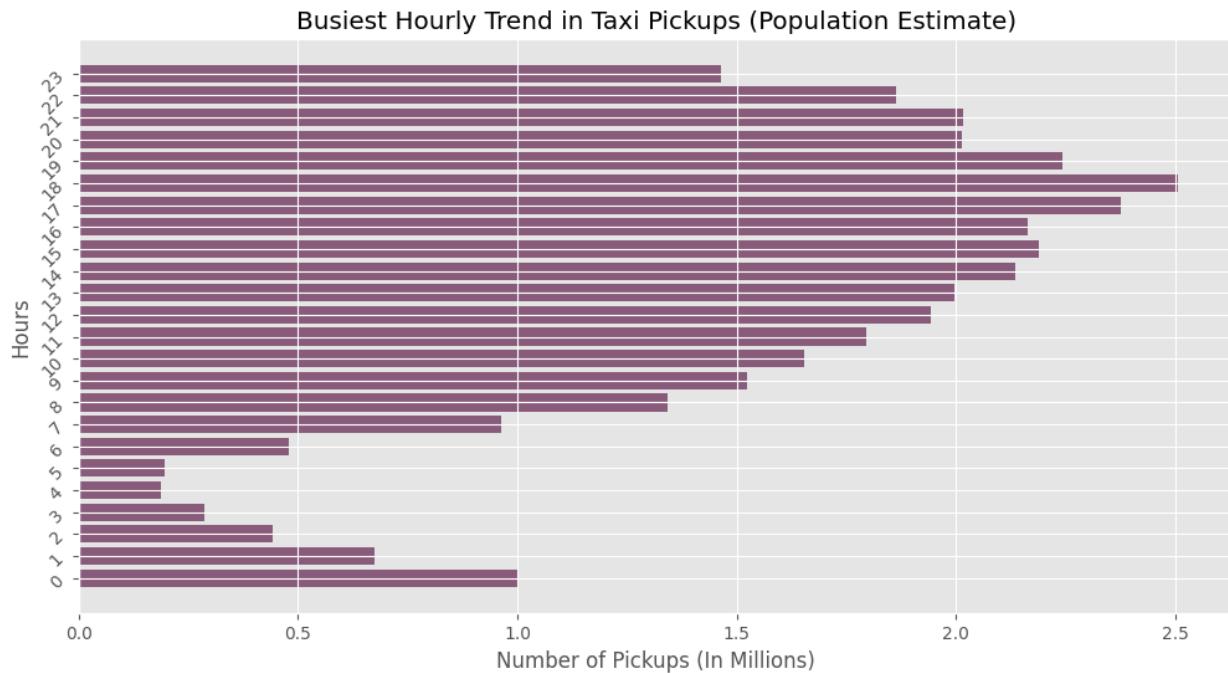
Frequently occurring names were Garment District, Penn Station/Madison Sq West, Times Sq/Theatre District, Midtown Center.

This analysis confirmed that a significant amount time is spent by our taxi drivers in heavy congestion areas, which directly impacts our operations. The analysis provides a view of the operational inefficiencies which allow us to make adjustments to our dispatching strategy and routing strategy and better time management for drivers.

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

Our analysis which was based on 10% of the sample data showed the busiest hour as 6PM. The Number of Trips for the busiest hour as per the population is 25,06,840 Trips.

The visualization is based on the 10% sample of the data. Based on this analysis bonuses can be given to driver's who put in the extra mile and work during the peak hours to support the demand.



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

Since the sample we collected was a random uniform sample, the relationship between sample size and population size can be estimated.

This is a Report on the Top Five Busiest Hour's of the day :

- **Rank 1 :** For 6 PM, The Estimated Population is : 2.51 Million Trips.
- **Rank 2 :** For 5 PM, The Estimated Population is : 2.38 Million Trips.

- **Rank 3** : For 7 PM, The Estimated Population is : 2.24 Million Trips.
- **Rank 4** : For 3 PM, The Estimated Population is : 2.19 Million Trips.
- **Rank 5** : For 4 PM, The Estimated Population is : 2.16 Million Trips.

3.3.5. Compare hourly traffic on weekdays and weekends

The analysis was based on the sample of my data i.e. 10% but will hold true for the population as well.

A total of eight plots were created to provide us with the right insights and trends we are looking for into the hourly traffic pattern on weekdays and weekends as a whole as well as individual days. This was an in-depth analysis. We were able to identify Busy Hours and Quiet Hours for weekdays and weekends.

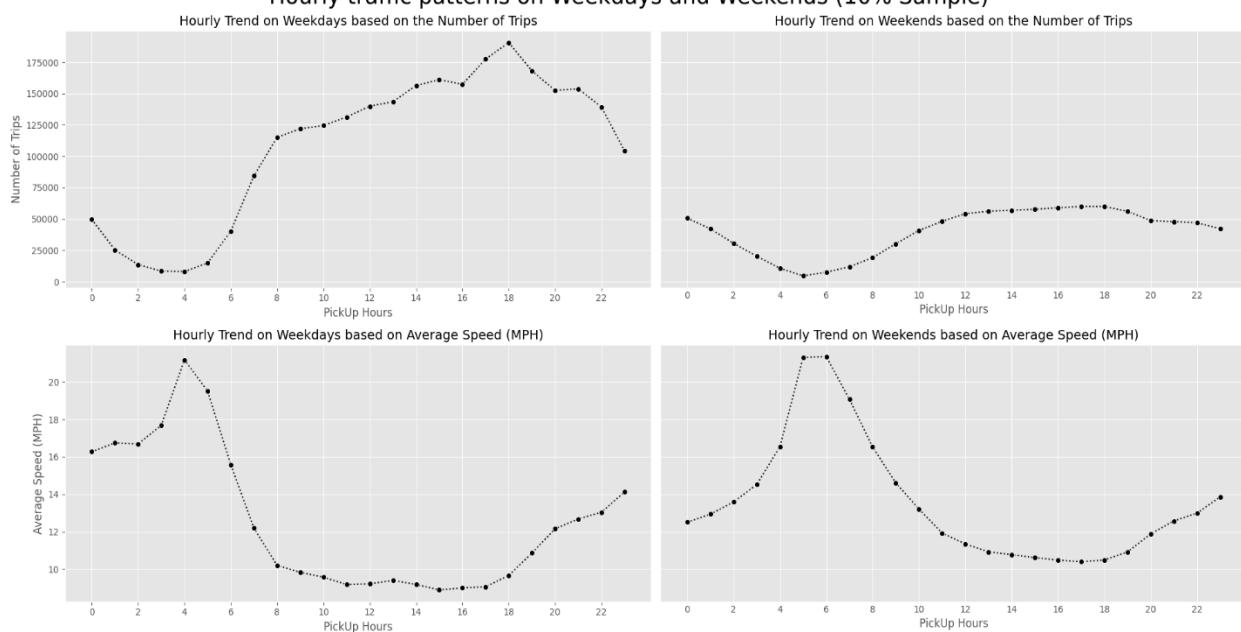
From the plots we can tell that the Average Speed and the Number of trips is correlated. As the number of trips increases the Average Speed decreases and vice versa.

The number of trips is more on Weekdays than on Weekends.

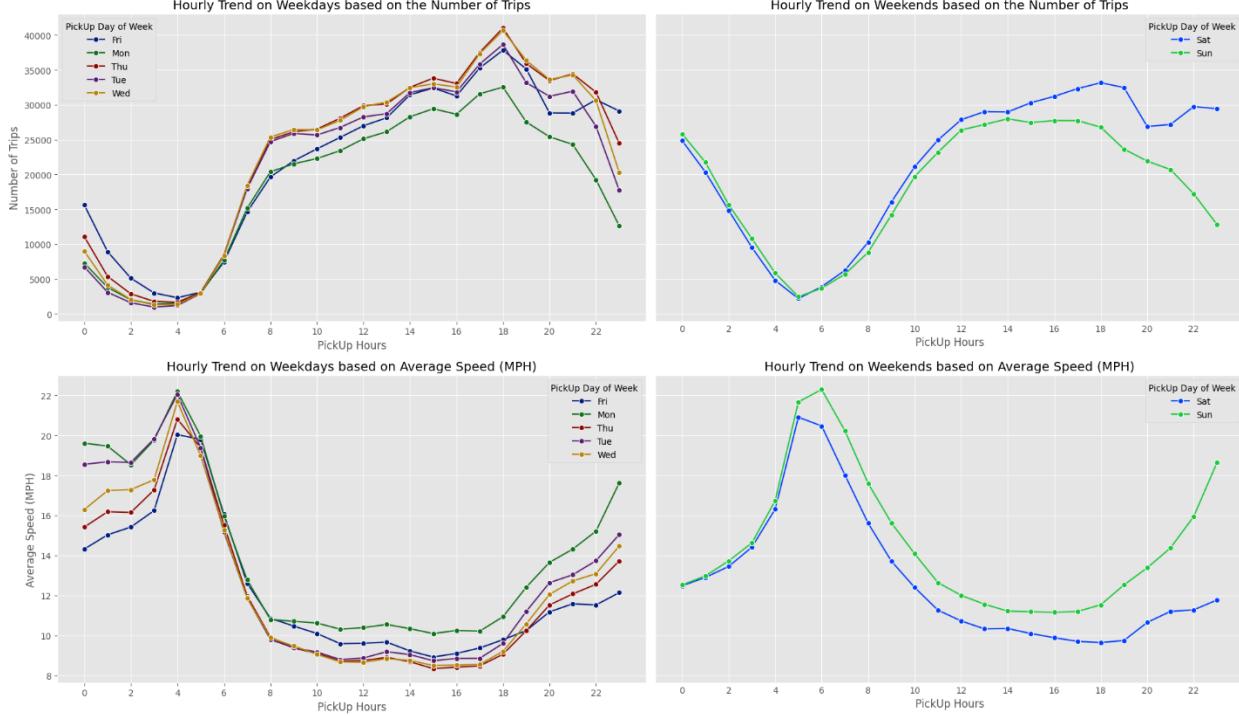
- Busy Hours (Weekdays) : 8AM to 6PM. Spike starts at 6AM.
- Busy Hours (Weekends) : 8AM to 6PM. Spike starts at 8AM.
- Quiet Hours (Weekdays) : 12 AM to 6AM. High Average Speed.
- Quiet Hours (Weekends) : 12 AM to 8AM. High Average Speed.

This kind of analysis is useful in predicting busy hours and quiet hours, it will help us in resource allocation. Driver's can be given breaks when demand is low and can work during surge hours to meet the demand.

Hourly traffic patterns on Weekdays and Weekends (10% Sample)



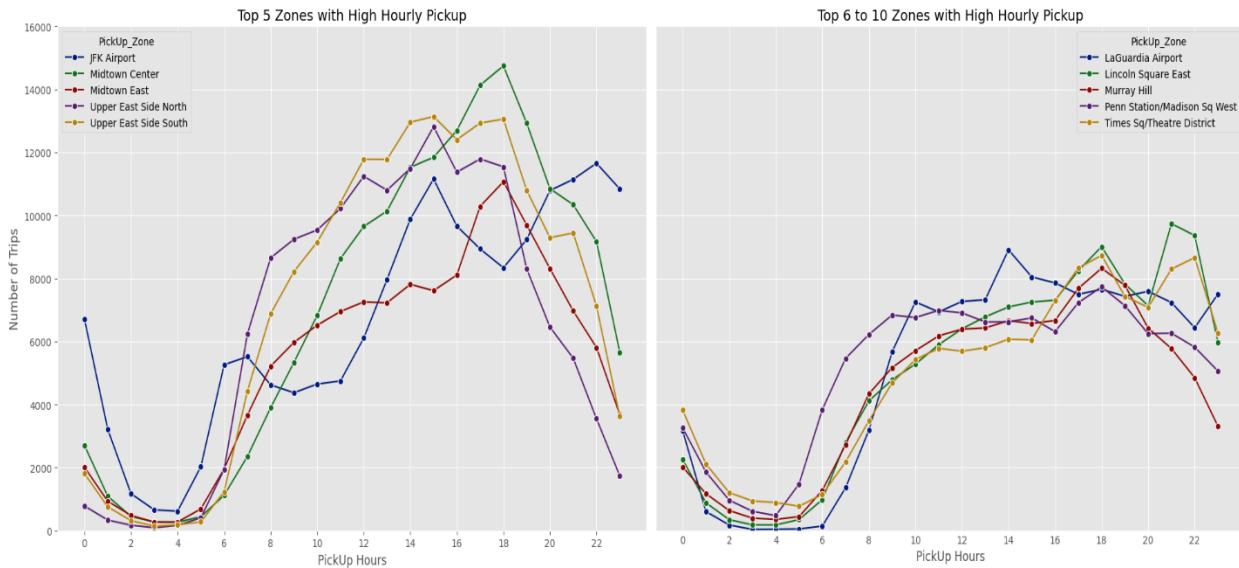
Hourly traffic patterns on Weekdays and Weekends for Each Day of the Week (10% Sample)



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

A total of four plots was created to provide us with the right insights and trends we are looking for into the high hourly pickups and drops. This kind of analysis reveals that the high demand zone has a unique hourly trend which must be taken into account. The demand is usually from 6AM to 6PM but the demand need varies across zones.

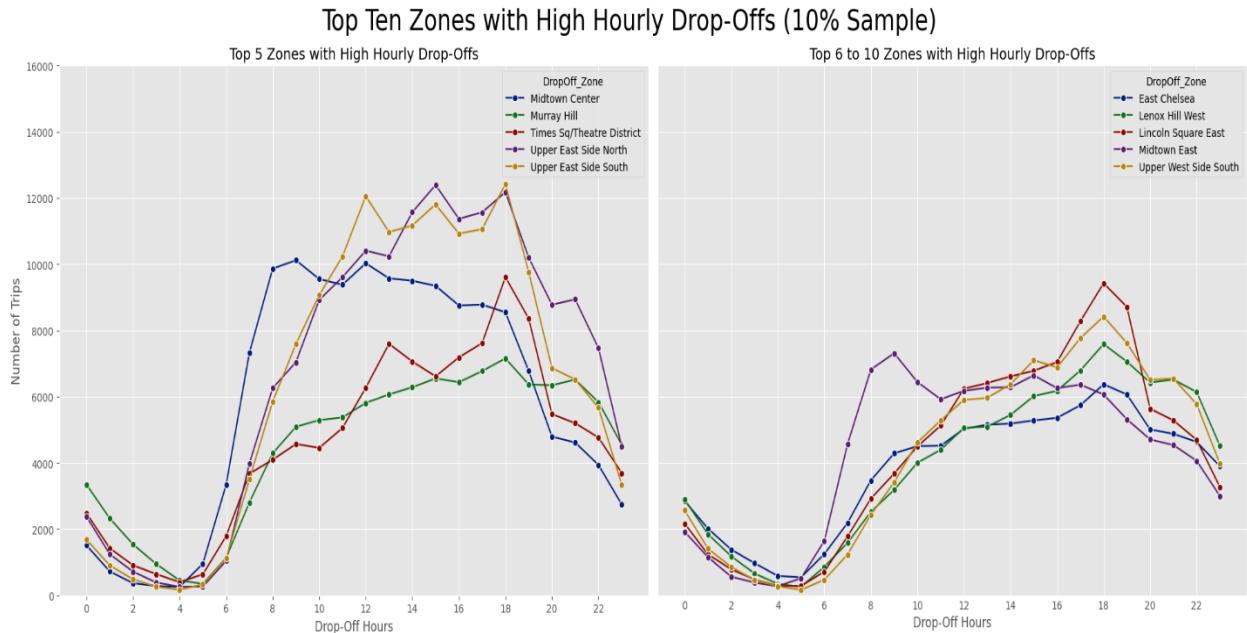
Top Ten Zones with High Hourly Pickup (10% Sample)



Airport Zones are in the top ten pickup high demand zones, JFK Airport and LaGuardia Airport. They could be considered as a major source of revenue.

Presence of Midtown Center, Midtown East, Murray Hill, Upper East Side North, Upper East Side South appear in both Pickup and Drop Top 10 High Demand Zones. Indicating the need for strategic dispatchment of taxis in these areas.

Penn Station/Madison Sq West, Times Sq/Theatre District, Midtown Center appear in the Top 10 Pickup Zones however, we must take into account that these are slower routes while planning a strategy.



3.3.7. Find the ratio of pickups and drop-offs in each zone

Based on our analysis we were able to identify the top ten zones with the highest and lowest ratio of pickups and drop-offs. The zones which have a higher ratio are high demand zones and the ones with a lower ratio are lower demand zones. A higher ratio also indicates that a zone has more pickups than drop-offs while a low ratio indicates the opposite.

This kind of analysis is useful for zone-wise targeting and transportation planning and dispatch.

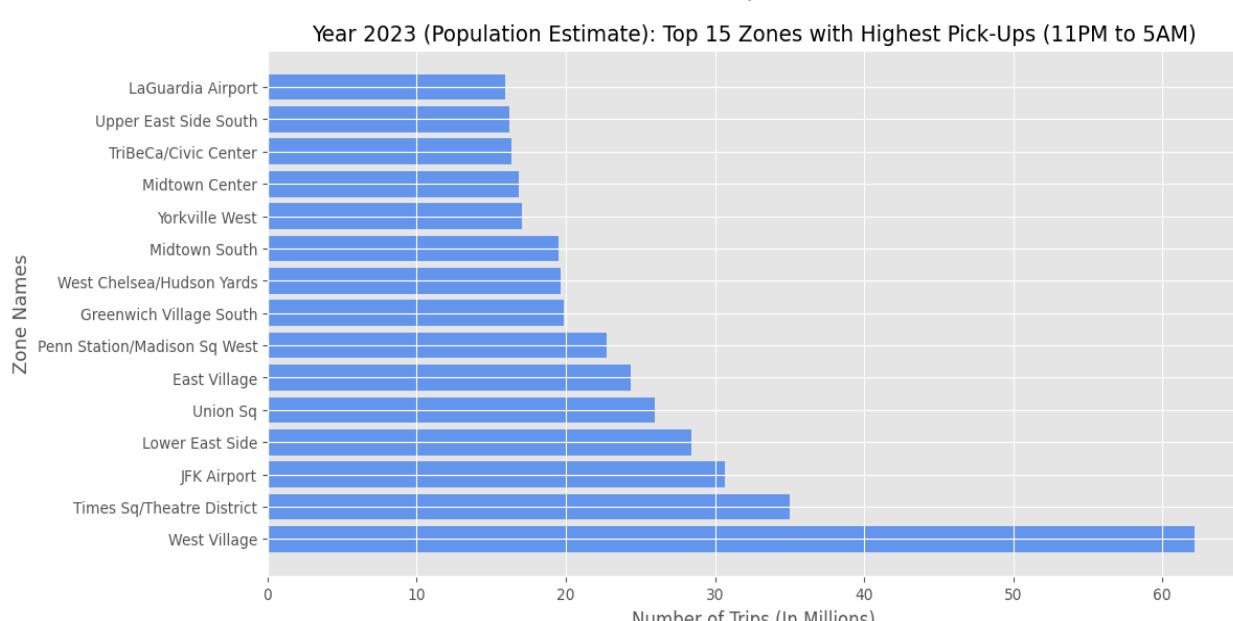
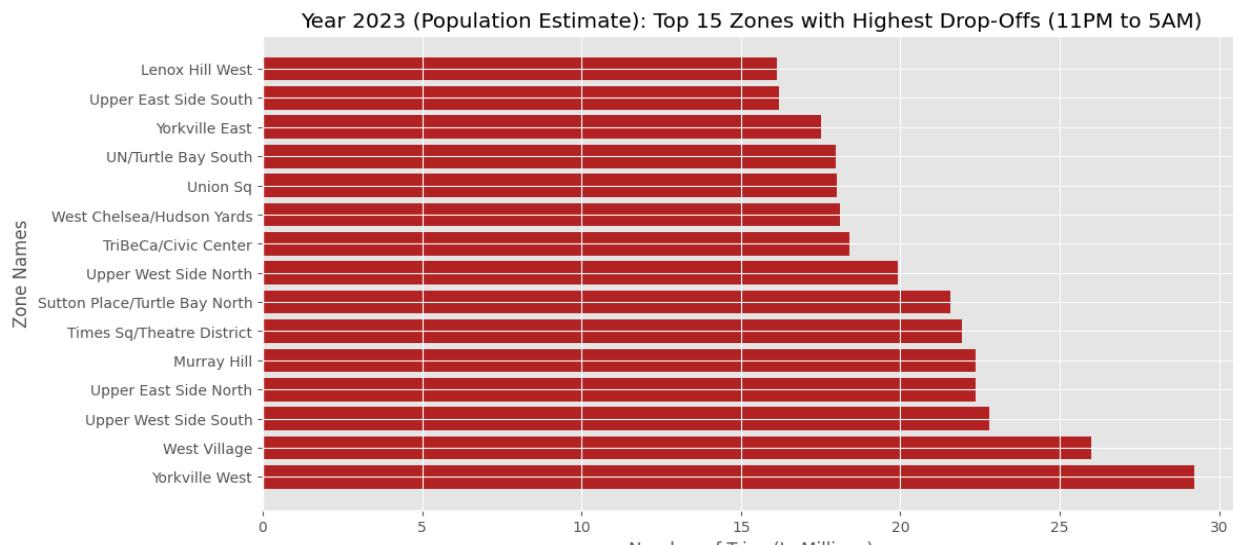
Top 10 Lowest Pickup / Drop-off Ratio : Heartland Village/Todt Hill, Breezy Point/Fort Tilden/Riis Beach, West Brighton, Broad Channel, Forest Park/HIGHLAND PARK, Windsor Terrace, Douglaston, Whitestone, Bay Terrace/Fort Totten, Riverdale/North Riverdale/Fieldston.

Top 10 Highest Pickup / Drop-off Ratio : East Elmhurst, JFK Airport, LaGuardia Airport, Penn Station/Madison Sq West, Greenwich Village South, Central Park, West Village, Midtown East, Garment District, Midtown Center.

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

For the purpose of this analysis, I have considered 11PM to 5AM as Night Hours.

From our analysis and plots these Zones have high Pickup and Drop-Off traffic during Night Hours (11PM to 5AM) : West Village, Times Sq/Theatre District, Union Sq, West Chelsea/Hudson Yards, Yorkville West, TriBeCa/Civic Center and Upper East Side South.



West Village has the highest number of pickups during night hours and second highest number of drop-offs during night-time hours, indicating that the area's nightlife is pretty good. Yorkville West has the highest number of drop-offs but a lower number of pickups compared.

This kind of analysis is useful to strategically place driver's in zone's with a strong nightlife. Surge Prices / Night-time charges could be applied at these zones. Based on our analysis we should target zones like West Village.

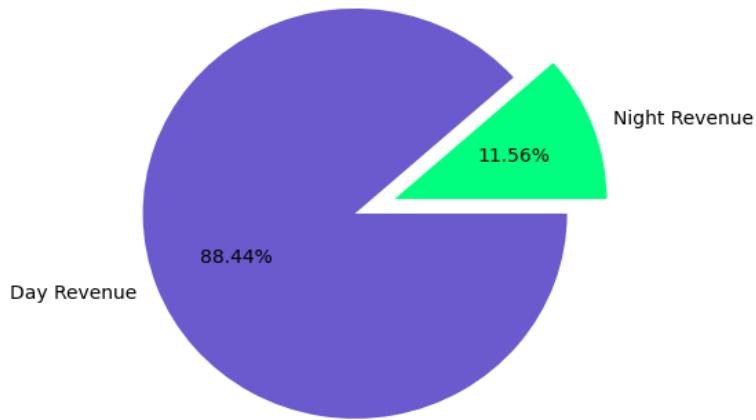
3.3.9. Find the revenue share for nighttime and daytime hours

The Revenue Amount Generated for Night-Time Hours (11PM to 5AM) : 11.56% as per our sample data, which will hold true for the population.

The Revenue Amount Generated for Day-Time Hours (5AM to 11PM) : 88.44% as per our sample data, which will hold true for the population.

We can say that 88.44% of the revenue is generated during the day and 11.56% of the revenue is generated at night.

Revenue Share for Nighttime and Daytime Hours

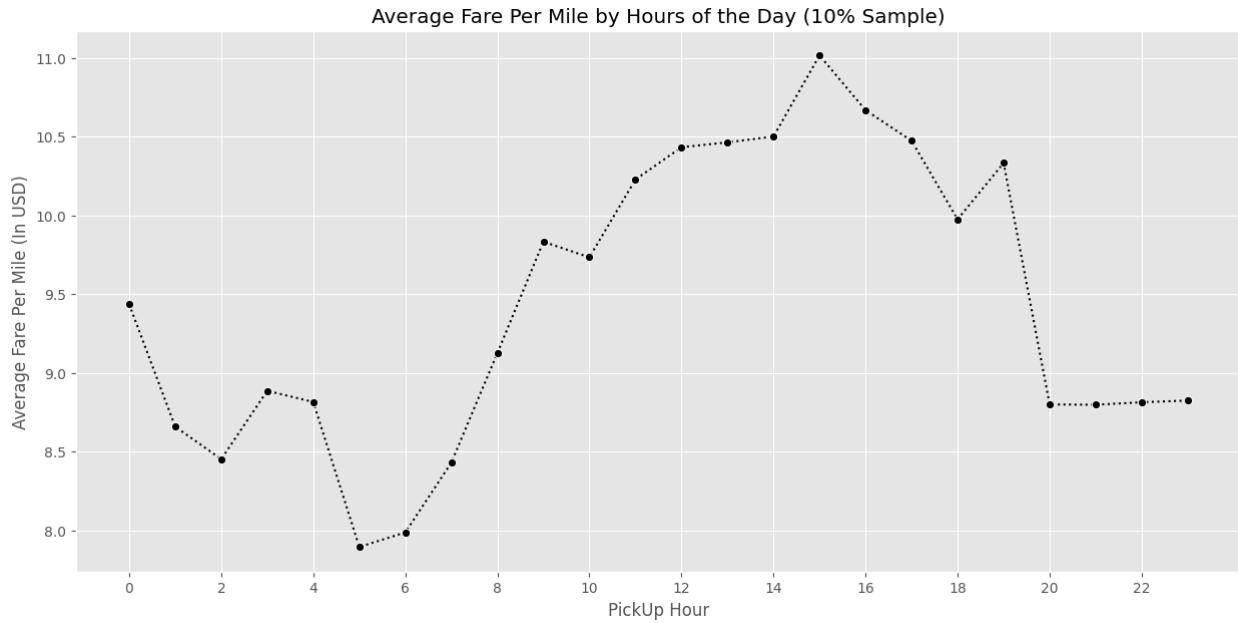


3.3.10. For the different passenger counts, find the average fare per mile per passenger

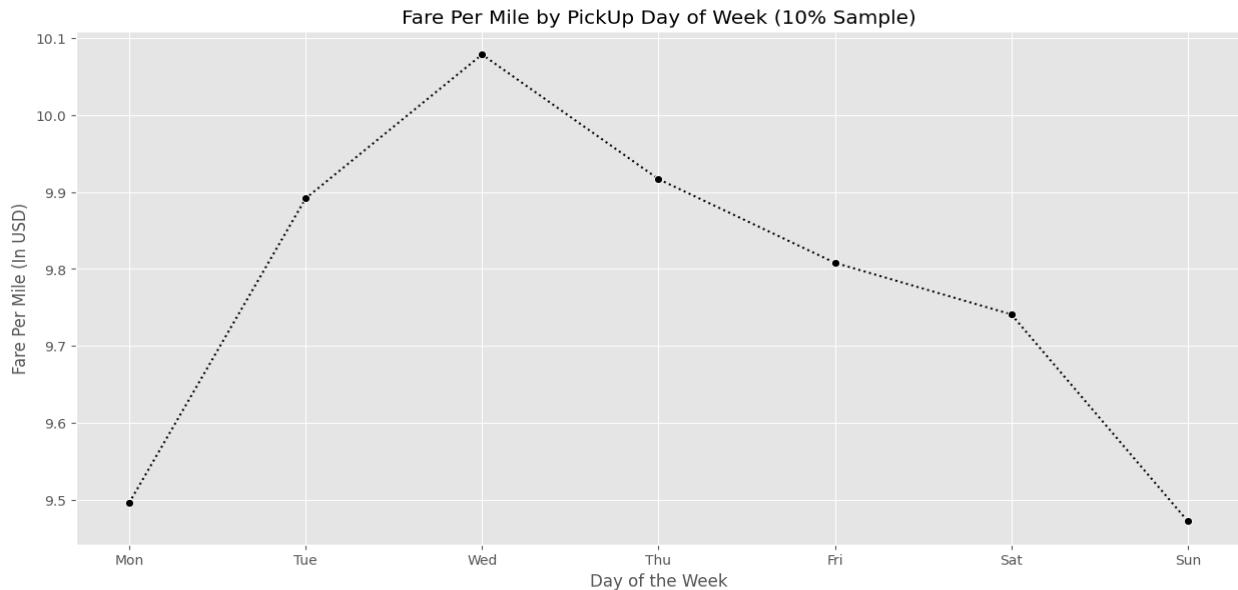
This was an important part of our analysis to determine the Fare Per Mile (per passenger). *Our Analysis Revealed : As the Passenger Count increased the Fare Per Mile Per Passenger decreased.* This pattern indicates that shared rides are more cost-effective for passengers, while it does not reduce the revenue for the Vendor since the Fare Amount remains the same irrespective. It is a Win-Win situation for both the Vendor and the Customers.

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

We analyzed the relationship between Average Fare Per Mile based on the Hour of the Day and Day of the Week. Two plots were created to study the relationship.



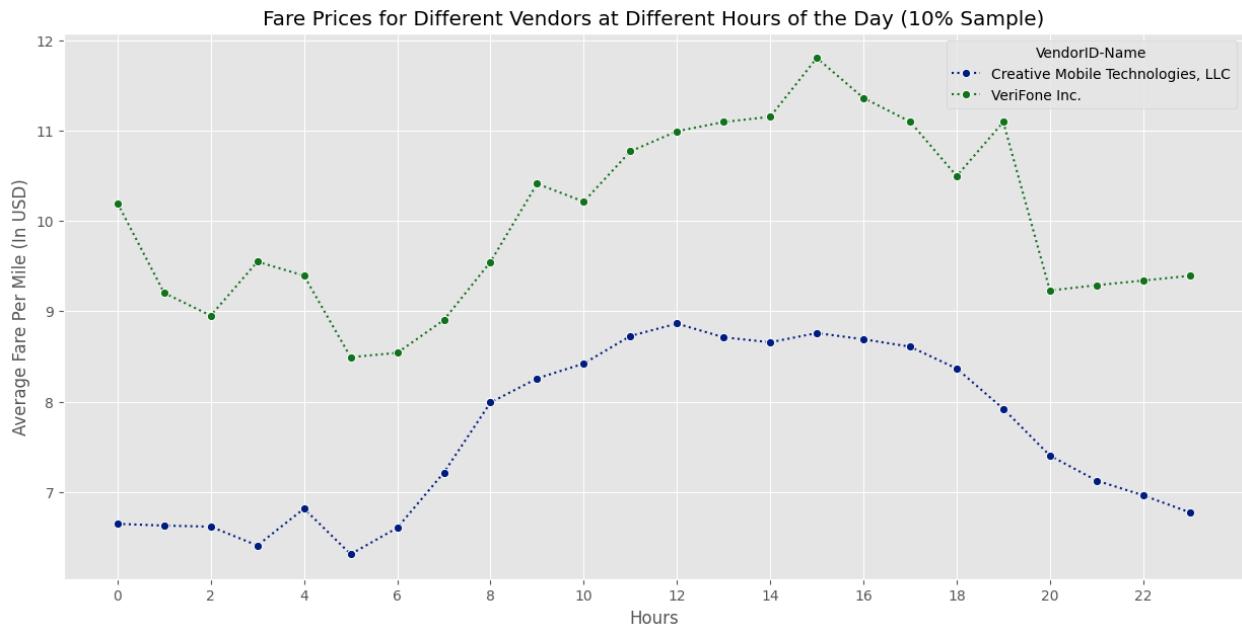
The lowest Average fare per mile is at 5AM (Early Morning) and the highest is at 3PM (Afternoon). Average fare per mile increases after 6AM and declines after 4PM. Between 8PM to 12AM the average fare stabilizes with very little variation. Companies can leverage this information and adjust fare prices based on the hour of the day. It can be used to identify new opportunities which would lead to maximizing revenue. Fare can be increased at night.



Sunday and Monday has the lowest Average fare per mile, which makes sense based on our earlier analysis these are day's where the demand is low. When the demand is high the prices peak, as we can observe the mid-week prices. Companies can plan more aggressively to meet revenue target's mid-week and plan for lower revenue on days with low demand. This kind of analysis is useful for informing businesses how they can strategically plan and utilize their resources.

3.3.12. Analyze the average fare per mile for the different vendors

We compared the average fare per mile for different vendors. Creative Mobile Technologie's LLC and VeriFone Inc. This analysis gave us insights about different vendors having different pricing.

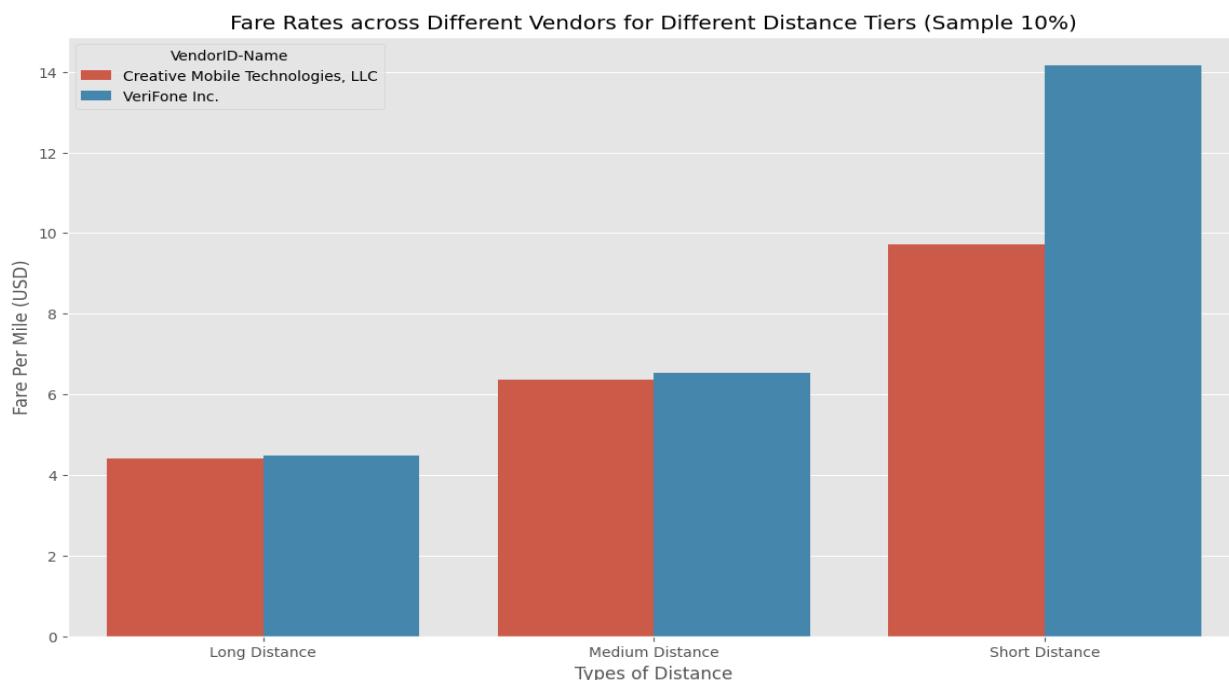


Creative Mobile Technologie's LLC has a higher average fare per mile. Verifone's average fare per mile peak is around 12PM whereas Creative Mobile Technologie's peak is around 3PM. It could indicate that these two vendors are serving different zones or maybe different groups of people. It could be that Verifone is usually for longer-distance trips so the average fare per mile is low and Creative usually takes short-distance trips.

This kind of analysis enables businesses to make informed decision. If a new vendor was to step in, he now has valuable information about the pricing model of different vendors at different time of the day. It helps us analyze competitor's behavior. This kind of analysis helps businesses make a better pricing model.

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

The analysis revealed that the fare amount and distance tier is inversely correlated which wasn't visible in our earlier analysis. Both vendors charge a higher fare for short-distance trips and the lowest fare for long-distance trips. This is a strong indication that short-distance trips are likely the most profitable due to an initial fixed fare. VeriFone Inc. consistently overcharges customers compared to Creative Mobile Technologies, LLC.



The plot suggests that the fare is heavily influenced by base fare. So, we could consider a pricing model with a higher base fare.

For passengers travelling a short distance, Creative Mobile Technologies, LLC, is more cost effective. The choice of Vendor doesn't matter much for medium and long distance.

From a business point of view, we have an option to adapt VeriFone Inc. pricing model. An alternative way would be to use a more competitive rate to attract customers who care about their spending habit.

3.3.14. Analyse the tip percentages

I analyzed the relationship of tip percentage with trip distance; passenger counts and time of pickup.

Short Distance trips tend to have a higher tip percentage compared to Medium and Long-Distance Trips. Trips with solo passengers and six

passenger tips more generously. Trips with three and four passenger's tip the least. The tips are observed to be the highest in the evening. Tipping percentage suggest that people tip less in the Morning and Afternoon time compared to Evening and Night-time.

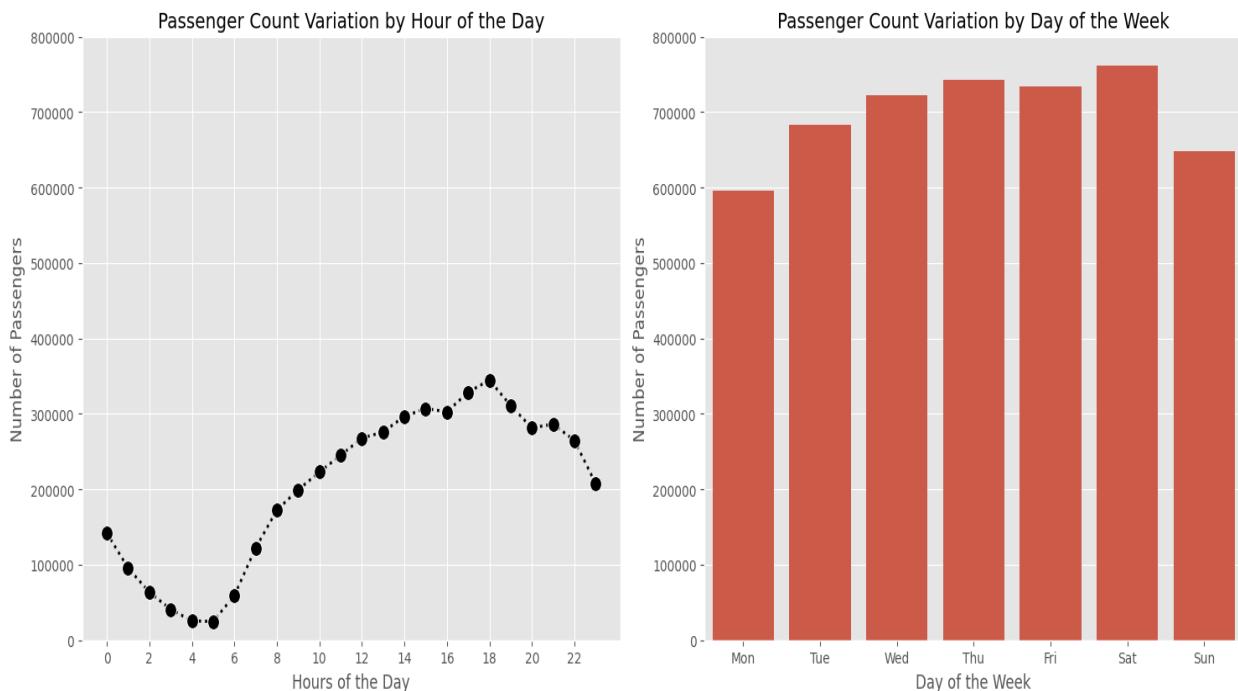
Tip amount feature has a stronger relationship with trip distance and time of the day compared to passenger count.

Further analysis suggested that short distance trips may have a higher tip percentage but also have an average low tip percent compared to medium and long-distance trips. Time of the day plays an important role in the range of tipping behavior. Further analysis confirmed our belief that solo passengers are more generous and their tipping range is more.

3.3.15. Analyze the trends in passenger count

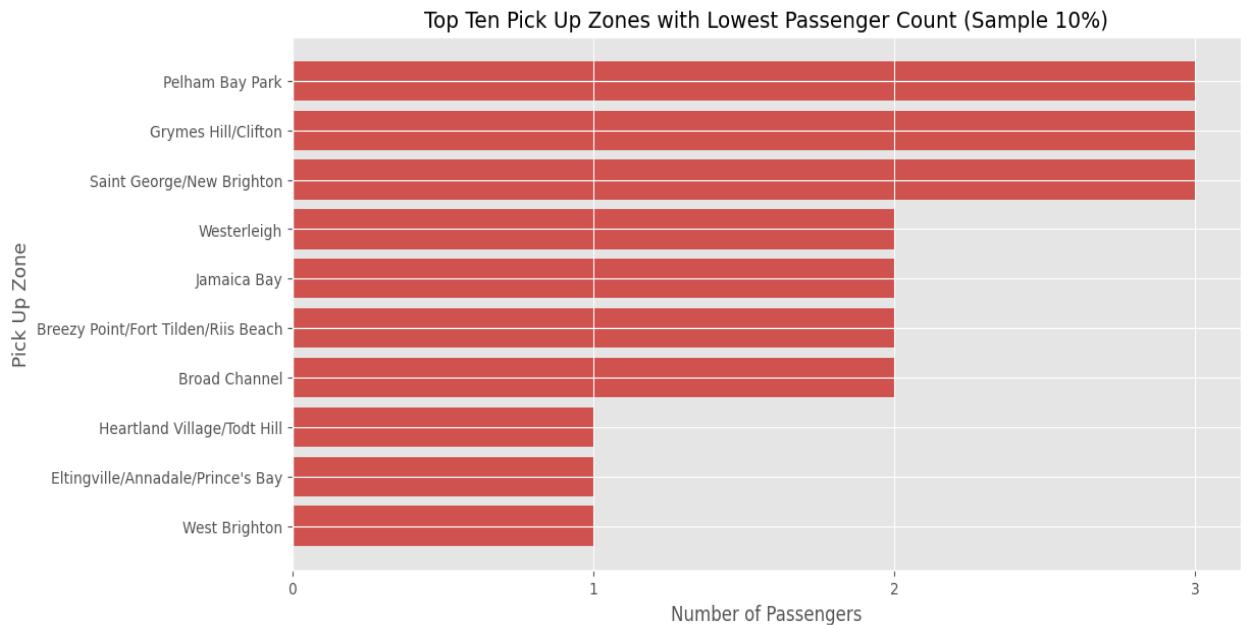
From the plot we can observe that the passenger count is lowest in the early morning, highest in the evening and rises steadily throughout the day. Passenger count shows a clear variation across the week. This kind of analysis also suggests differences between weekdays and weekends. This kind of analysis is useful if the vendor is providing group ride services to customers. Public transports can be deployed during peak rush to avoid overcrowding. This kind of analysis is beneficial in improving the overall customer experience and reducing operational inefficiencies.

Variation of Passenger Count across Hours and Days of the week (Sample 10%)

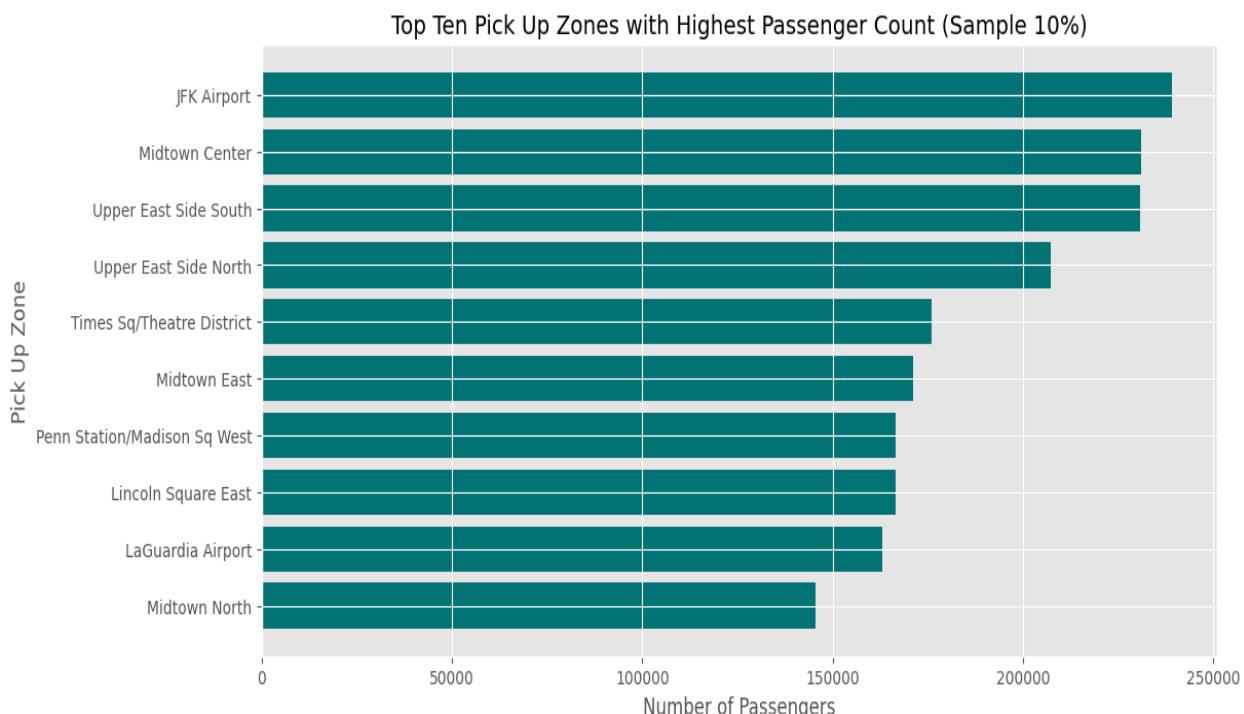


3.3.16. Analyse the variation of passenger counts across zones

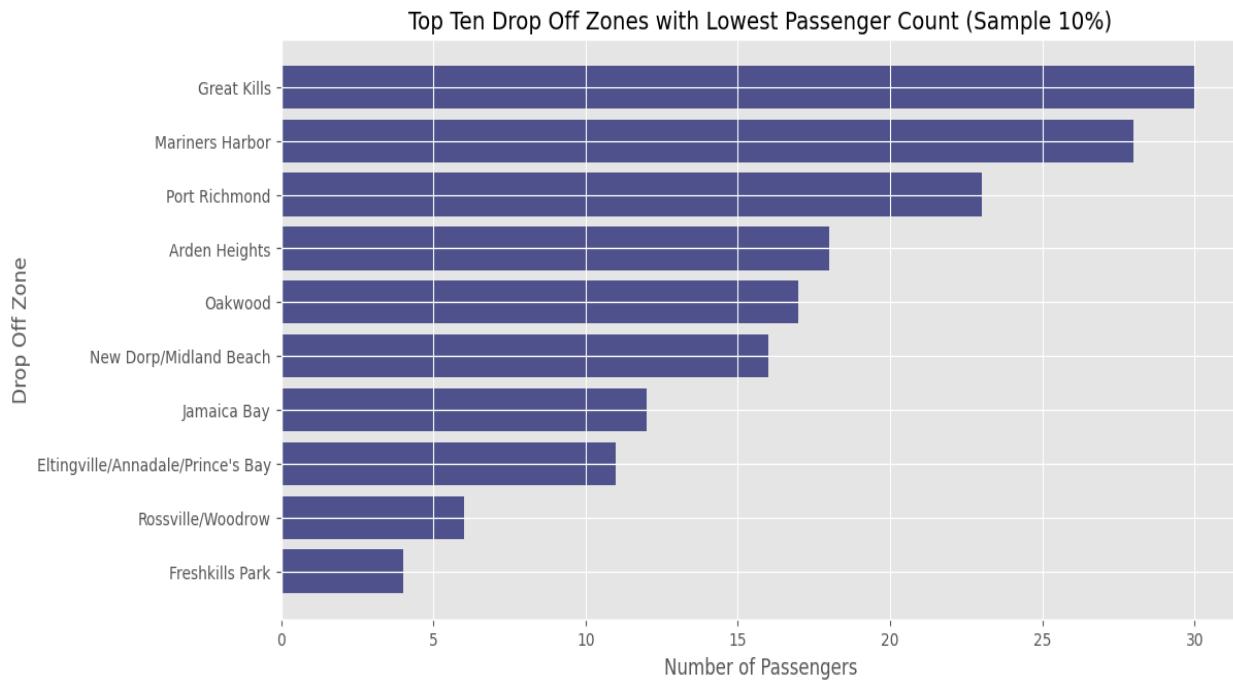
This kind of analysis helps us to spot the imbalance of passenger counts across zones. Four plots were made for the purpose of this analysis.



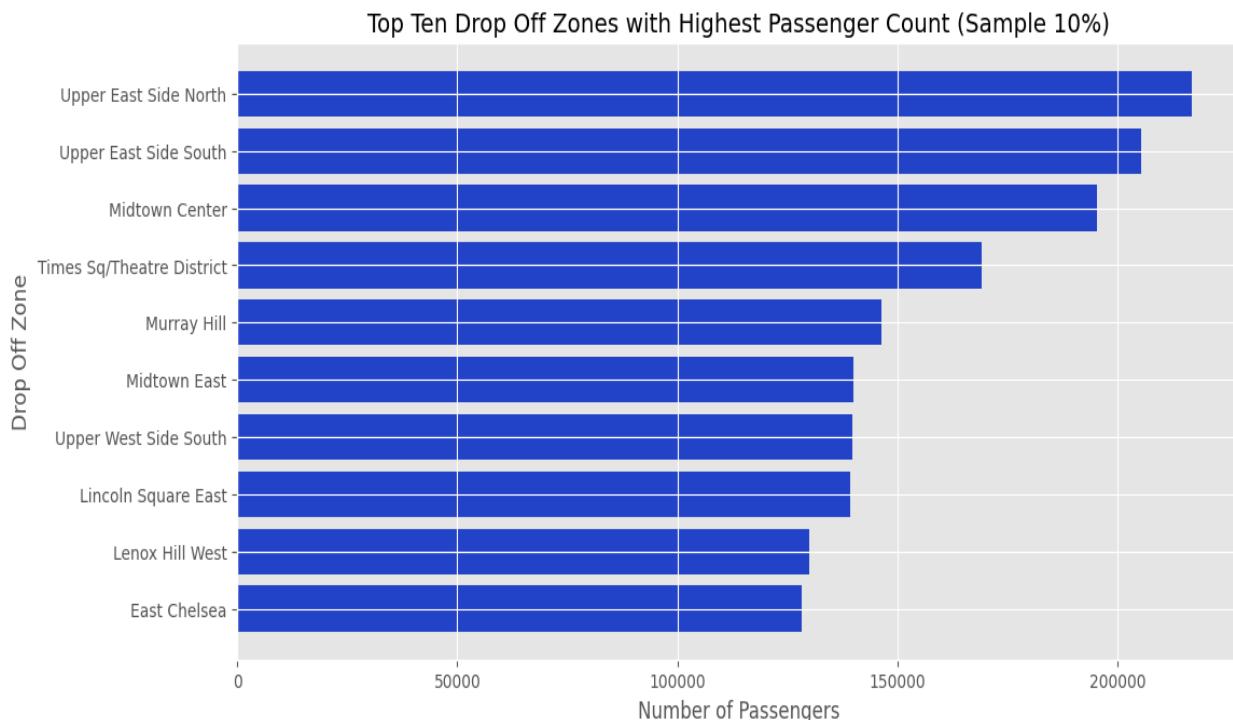
We can see a significant imbalance in demand between the lowest and highest passenger count pickup and the same for drop-offs zones as well. This is a key factor in deciding which zone's we should target and which zone's to not target.



This analysis helps with better resource and time management. It could be that the lowest pickup and drop-offs zone's are likely to be areas with a lesser population or underdeveloped.

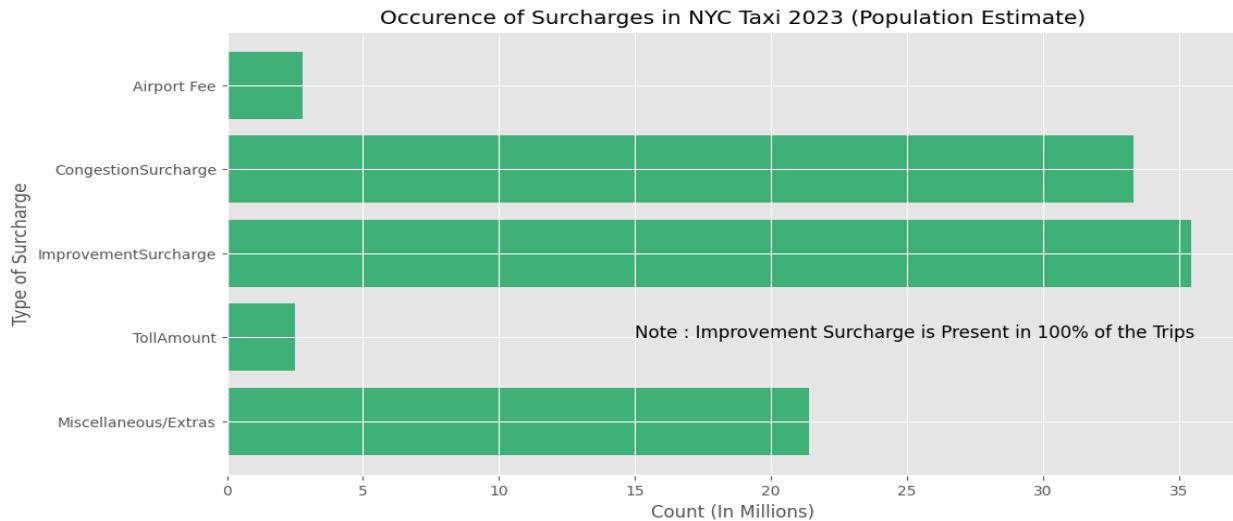


To meet demand and ensure high profitability, vendor's must target zone's with higher pickup and drop-offs since it is a clear indication for high demand.

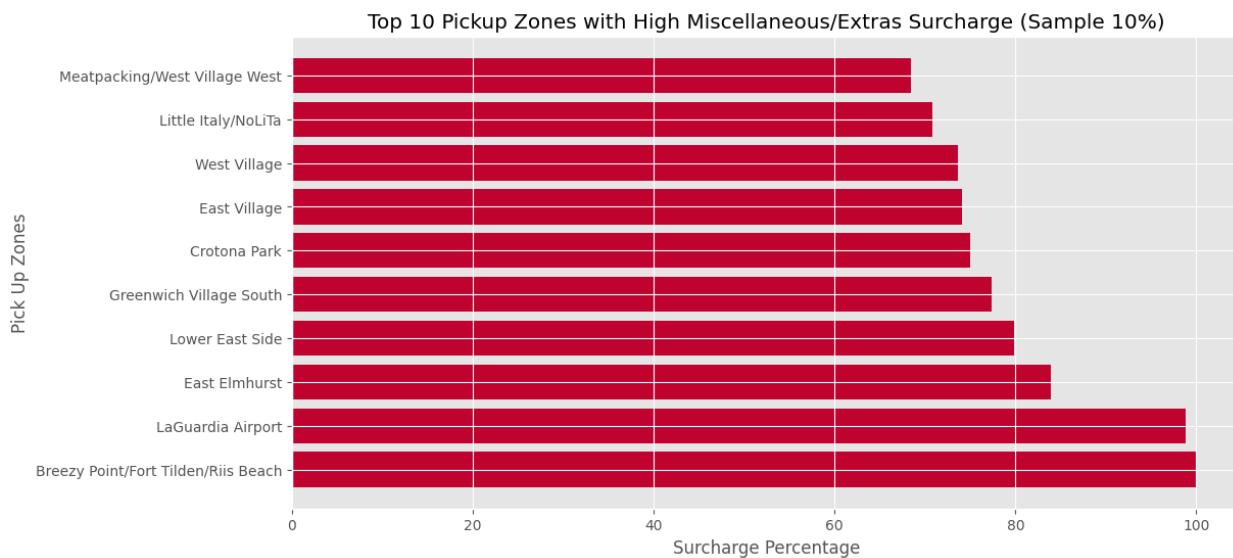


3.3.17. Analyze the pickup/drop-off zones or times when extra charges are applied more frequently.

Our analysis revealed that Improvement Surcharge was levied in all the trips as shown in the plot. Further analysis revealed that Airport Fee was levied only in 2 zones, JFK and LaGuardia Airport. Airport Fee and Improvement Surcharge was removed from further analysis since they do not provide any extra value.

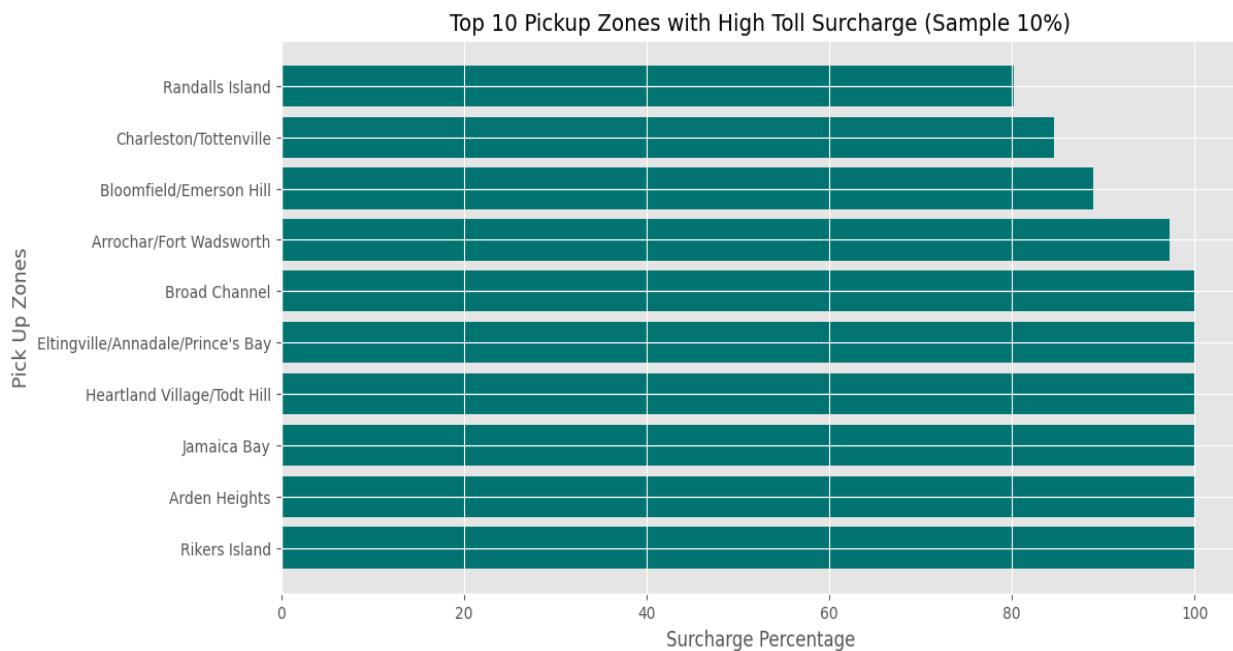


Based on the congestion surcharge feature, analysis revealed that the top Congestion Zones are From these Boroughs : Manhattan, Queens and Staten Island. Congestion surcharge is levied accordingly.

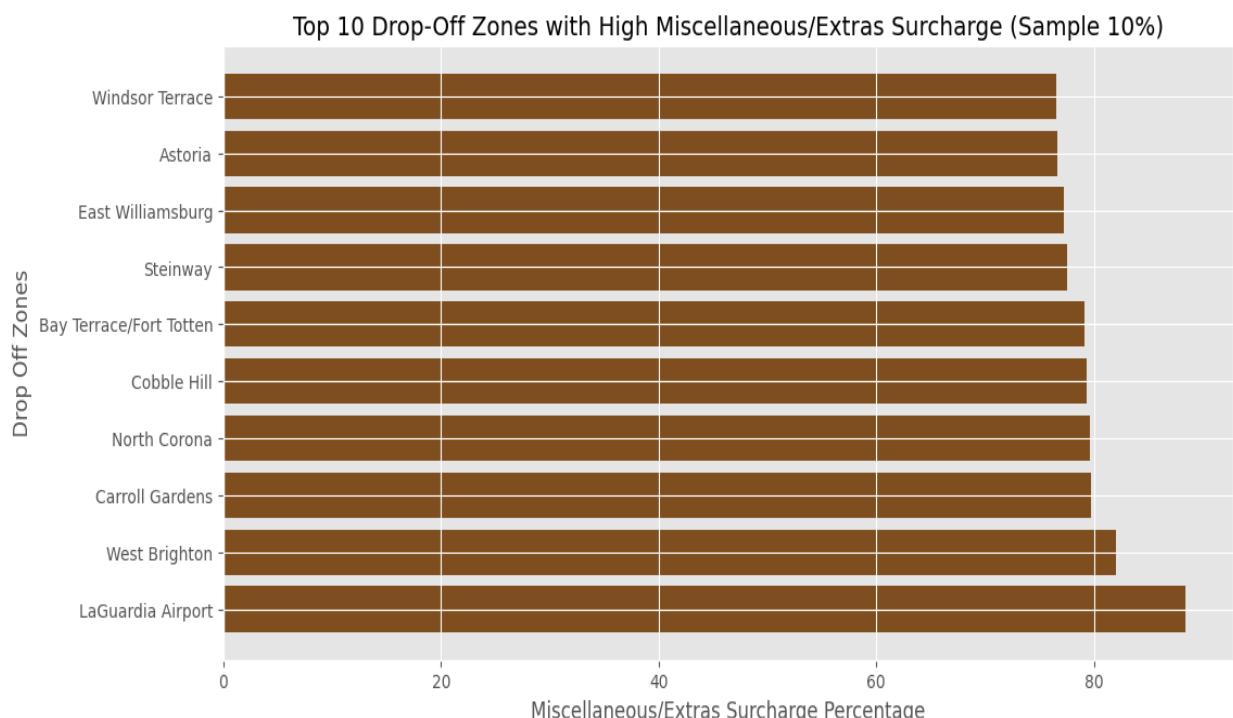


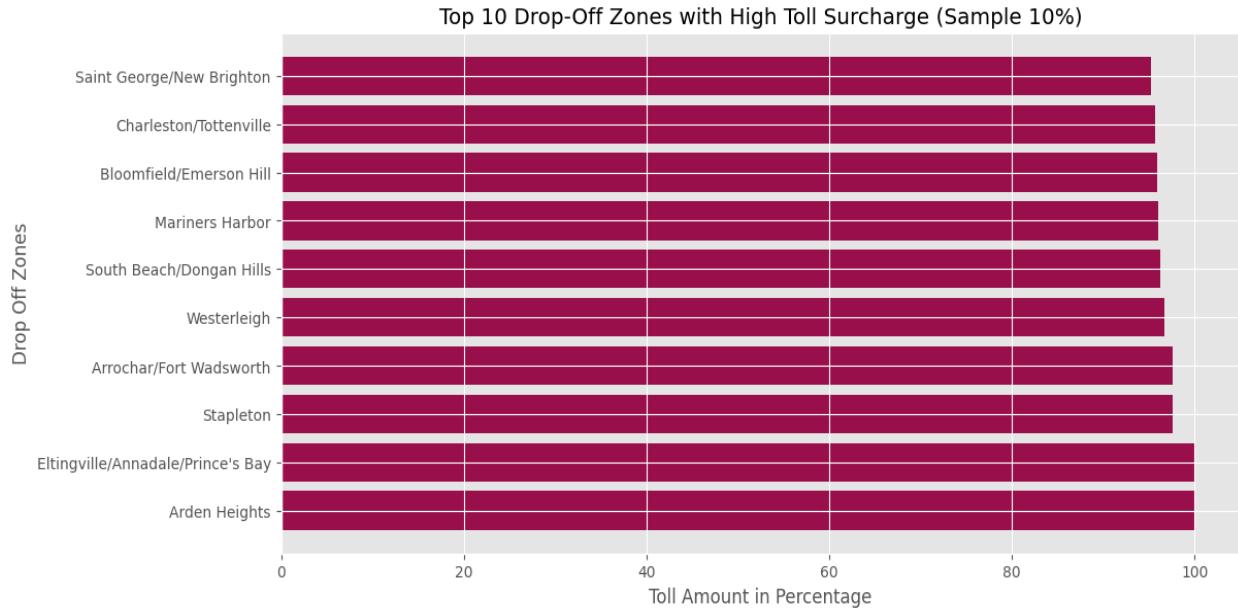
Further analysis was carried out and focused on the key drivers driving Miscellaneous / Extra Surcharge and Toll Amount Surcharge. The analysis reveals that these features are predictable and not random. This provides

an opportunity for improve the fare estimate, a more accurate fare estimate would surely lead to happier customers.



Some Pickup and Drop Off zones are observed to have a 100% Toll amount surcharge and can be seen in our visuals. Understanding these patterns is essential for businesses to introduce a dynamic pricing model taking these factors into consideration.





4. Conclusions

4.2. Final Insights and Recommendations

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

- *Forecasting of the demand pattern* could lead to better dispatching and use of resources.
- *Incentives* to drivers complete trips in Higher Demand and Slower Routes. Drivers who work during peak hours or week-offs should be given an incentive. *Bonuses and Incentives* for drivers who meet the demand at zones with a higher demand.
- *Avoid over supplying drivers* to routes with lower demand and higher traffic.
- *Zone-wise targeting and transportation planning and dispatch* based on our pickups and drop-offs ratio analysis. A higher ratio indicates that a zone has more pickups than drop-offs while a low ratio indicates the opposite.
- *Nightlife*: Based on our analysis, dispatching driver's to zones where the nightlife is higher compared to other zones.
- *Long-term resource management* strategies can be implemented on the basis of the Monthly trends.
- *Tracking Idle Time of Drivers* : Analyzing this metric if more information was provided could help us draw more insights into operational inefficiencies.

- *Vendors can make deals with companies to provide transport.* Since, Evening Commute time is likely to be associated with office staff.
- *Seasonal Dispatch / Monthly Dispatch* to meet the months with higher demand, we can ensure that drivers are available.

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

- Our staffing / dispatching model should be focused on *deploying drivers during peak hours, peak days and peak months.*
- *Airport Zones* were identified as a major source of revenue, and it is a high in demand zone for pick-ups. Airport Fees are levied as well.
- Presence of Midtown Center, Midtown East, Murray Hill, Upper East Side North, Upper East Side South appear in both Pickup and Drop Top 10 High Demand Zones. Indicating the need for *strategic dispatchment of taxis in these areas.*
- *Strategic Planning based on Plots* : The high demand zone has a unique hourly trend which must be taken into account. The demand is usually from 6AM to 6PM but the demand need varies across zones.
- *Slower Routes but High Demand* : Penn Station/Madison Sq West, Times Sq/Theatre District, Midtown Center appear in the Top 10 Pickup Zones however, we must take into account that these are slower routes while planning a strategy.
- *Night-Time Analysis* : This kind of analysis is useful to strategically place driver's in zone's with strong nightlife. Surge Prices / Nighttime charges could be applied at these zones.
- *Driver Recruiting and Training and Vehicle Maintenance* can be done during months like February and August when the demand is low.
- *Utilizing Pickup and Drop-off Ratio Analysis for Strategic Positioning of Cabs* : The zones which have a higher ratio are high demand zones and the ones with a lower ratio are lower demand zones. A higher ratio also indicates that a zone has more pickups than drop-offs while a low ratio indicates the opposite.

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

- *Dynamic Pricing Model* can be introduced on the basis of frequent occurring tolls and miscellaneous surcharge.
- *Short Distance Trips* attract a base fare, focusing on these trips can drive revenue significantly. Our reports suggest that Short distance trips are more profitable than medium and long-distance trips.
- From a business point of view, we have an option to *adapt VeriFone Inc. pricing model*. An alternative way would be to *use a more*

competitive rate to attract customers who care about their spending habit.

- *Focus on Quarter 2 and Quarter 4*, these are the months that generate the maximum revenue. Planning should be done accordingly.
- We could use a competitive marketing strategy by *promoting group rides* without any extra fees or strings attached. *Or we could implement a small surcharge on group rides* to increase the revenue from our trip without impacting the customers.
- *Reducing the staff* when the demand is low will save us on operational costs which in short leads to better revenue management.
- *Surge Prices / Night-time charges* could be applied to zones with a higher nightlife.
- *Surge Pricing* could be implemented on Slow Routes having Higher Demand. To meet the demand.
- *Based on our hourly analysis* : companies can leverage this information and adjust fare prices based on the hour of the day. It can be used to identify new opportunities which would lead to maximizing revenue.
- *Targeted marketing* promotions can be done for the low demand months or seasons with the intention of maximizing revenue.
- We could make an informed decision regarding the *Congestion Pricing Strategy* as per the congestion prone zones with higher demand.