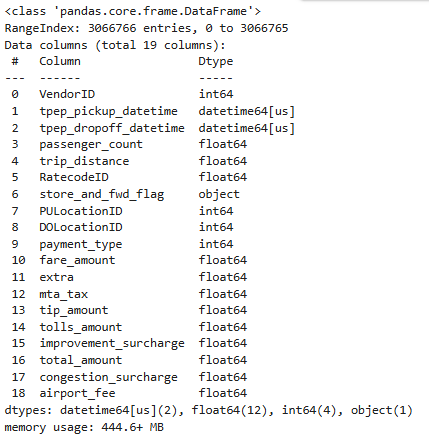
Report: Optimising NYC Taxi Operations

Include your visualisations, analysis, results, insights, and outcomes. Explain your methodology and approach to the tasks. Add your conclusions to the sections.

## Data Preparation

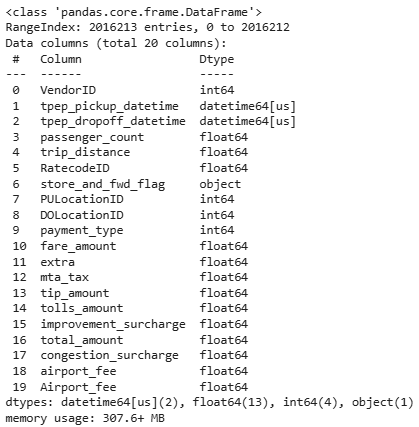
* 1. Loading the dataset
     1. **Sample the data and combine the files**

To initiate data analysis for NYC yellow taxis for the year 2023, data was sampled to study the type and number of columns, data types and data format from the January 2023 file. approximately 30 lacs entries for a single month. To evaluate data set for the entire year, a randomized subset i.e 5% of data from every hour will be studied for data analysis.



**Figure 1: Data information for January 2023 file**

After studying the file information, it was realized that it is a huge data set (~30 lac entries for a single month) which will lead to an increased run time and may lead the system to hang. Hence, 5% of every hour of every day for all 12 months (January to December) is sampled to study the data trend. A dataframe ‘df’ was created for the same. Following are the details of the dataset:



## Data Cleaning

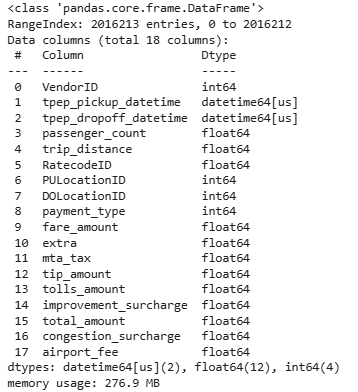
### Fixing Columns

* + 1. **Fix the index**

As a part of data cleaning, the column ‘store\_and\_fwd\_flag’ which denotes whether the trip record was held in vehicle memory was removed as the data availability implies that the data was stored. Further, the column was inferred to have no real usability. Further, the index was reset to avoid out of order indexes.

* + 1. **Combine the two airport\_fee columns**

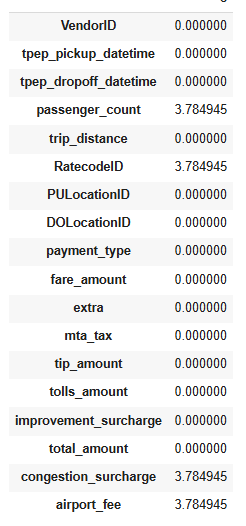
The dataset (df) was found to have two ‘airport\_fee’ columns. Both the columns were combined into a single column) and a new dataset with the name ‘df\_new’ was formed and used for all further evaluations hereon.

**

### Handling Missing Values

* + 1. **Find the proportion of missing values in each column**

The dataset was further studied for its missing values and it was found that out of all the columns, a total of 4 columns had ~3.8 % of missing values.The missing values were handled as below:



* + 1. **Handling missing values in passenger\_count**

The NaN values in passenger\_count column were filled with the median value of the column as a central tendency/ fixed value appraoch was found to be suitable for this column type since, it is an independent factor.

* + 1. **Handle missing values in RatecodeID**

The NaN values in RatecodeID column were filled with the mode value of the column as most used ratecode seemed to be an appropriate substitute.

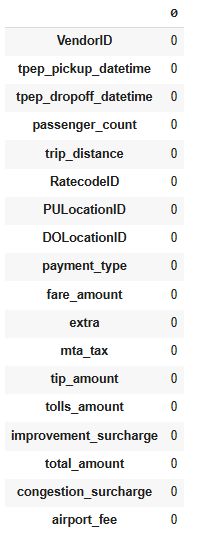
* + 1. **Impute NaN in congestion\_surcharge**

To impute the NaN values in congestion\_surcharge, the dataset was studied for the column’s median and mode. Both median and mode were found to be of the same value (2.5). It denotes that the central value is the most applied value. Hence, the NaN values in congestion\_surcharge column were replaced by the value 2.5.

* + 1. **Impute NaN in airport\_fee**

The missing values in the column airport\_fee were imputed with the value 1.25 for ratecodeID 2 and with the value 0 for other ratecodeIDs as the airport fee is applicable only for JFK and LaGuardia airport pickup.

A final check for null values was done as follows:

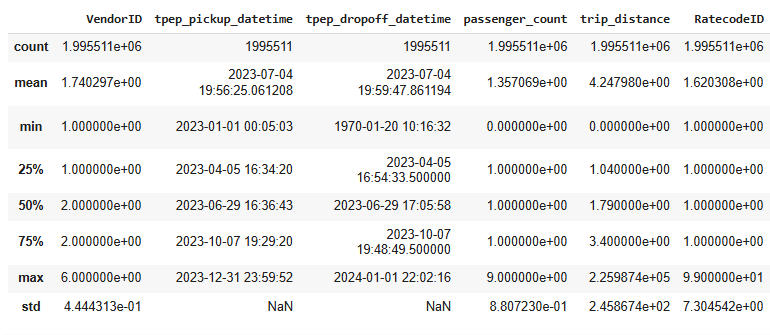


### Handling Outliers and Standardising Values

* + 1. **Check outliers in payment type, trip distance and tip amount columns**

To study the outliers in the dataset, df\_new.describe() function was used. The statistics obtained from the output was studied for difference between the 75th percentile and maximum column value to evaluate outliers. Further the standard deviation was also evaluated in order to study the extent of variation in the data.

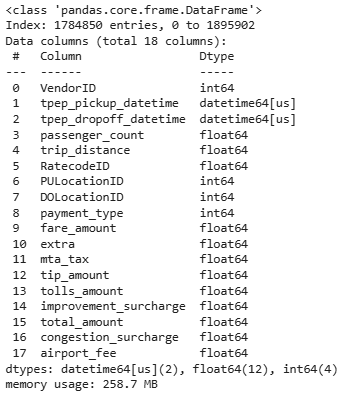
An example of the table is given below with few columns for reference:



From the outliers analysis, it was observed that the columns passenger\_count, **payment type, trip distance and tip amount columns had outliers. These outliers were dealt with as follows:**

1. Rows with passenger\_count>6 were removed
2. Rows with smaller trip\_< 0.1 and fare\_amount > 300 were removed
3. Data set with trip\_distance >250 miles were removed to avoid any extravagantly large distances from the dataset.
4. All the rows with payment\_type 0 were removed as it is not a valid payment type

Following are the details of the dataset after removing outliers and adding a column ‘trip\_distance for future use:



## Exploratory Data Analysis

### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical**

To classify the variables into categorical or numerical, a df\_new.info() function was used. The numerical variables have continuous values, whereas categorical variables have a pre-decided values. Thus following is the list of the categorical and numerical variables:

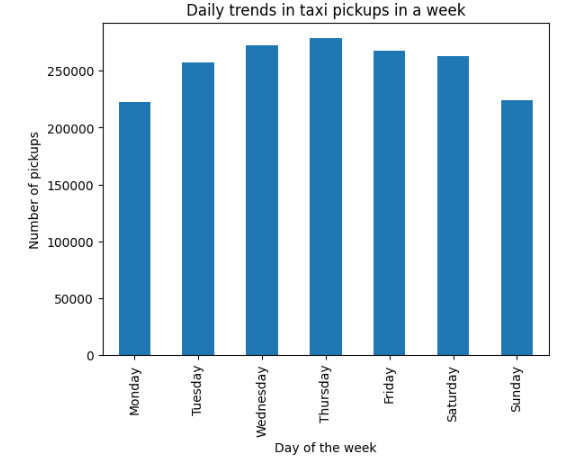
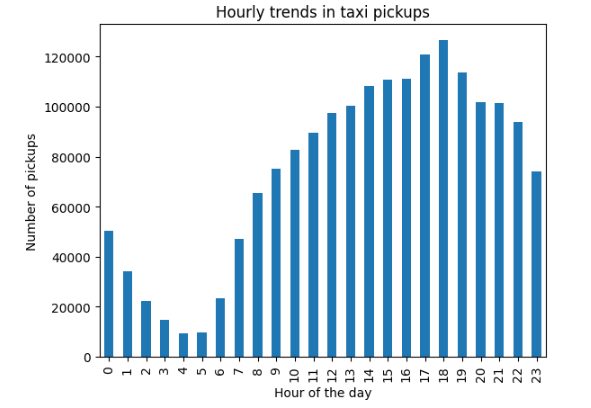
1. Categorical variables:

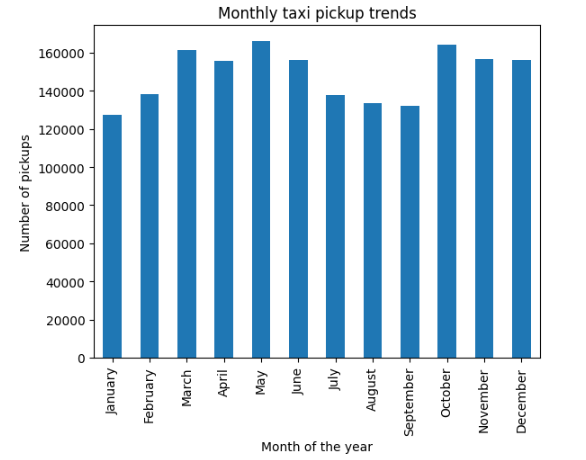
* VendorID:
* tpep\_pickup\_datetime:
* tpep\_dropoff\_datetime:
* passenger\_count:
* trip\_distance:
* RatecodeID:
* PULocationID:
* DOLocationID:
* payment\_type:
* pickup\_hour:
* trip\_duration:

1. Numerical variables:

* fare\_amount
* extra
* mta\_tax
* tip\_amount
* tolls\_amount
* improvement\_surcharge
* total\_amount
* congestion\_surcharge
* airport\_fee
  + 1. **Analyse the distribution of taxi pickups by hours, days of the week, and months**

The distribution of taxi pickups was evaluated by hours, days and months using a barplot as below:



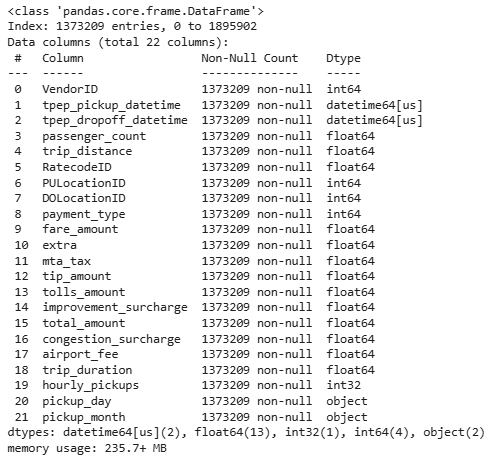


As per the bar plot the maximum pickups take place at 18 hours (6pm) in a day, on Thursdays of the week and in the month of May. Further, in a day, the pickups starts taking flight from 6 am to 6 pm followed by a decline from 7 pm to 5am. In a week, the weekends are observed to have lesser pick ups as compared to the weekdays. Further, in a year, it is observed that July-September has the lowest pickups and october to December has the highest pickups. The change in no. of pickups can be attributed to seasons as June to August is summer, which are tourist heavy and people use taxi services in June followed by a decrease from July to september with again an increase from October-November where people go back from vacation. Travelling still seems to be done till the early winters after which the pickups drastically fall in late winters (January and February).

* + 1. **Filter out the zero/negative values in fares, distance and tips**

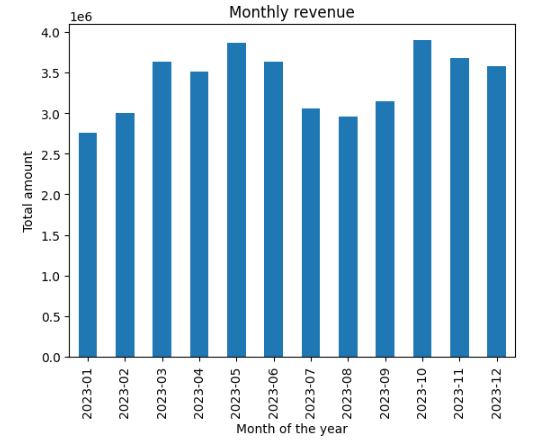
The dataset was then evaluated for zero/negative values for mainly following columns:

fare\_amount', 'tip\_amount', 'total\_amount', 'trip\_distance. As a negative/zero value for this columns is not a viable input, all the zero inputs were removed. A new data set (df\_nonzero) was created and will be used for further data analysis hereon.



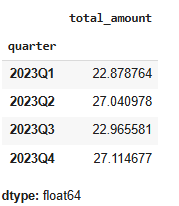
* + 1. **Analyse the monthly revenue trends**

As studied from the monthly revenues, the trend observed is same as that of pickups in a year. it is observed that July-September has the lowest revenue and October to December has the highest revenue. The change in the revenue can be attributed to seasons as in July-September many locals leave the city for vacations. The increase in revenue from October-November is attributed to the fact that people go back to work and kids go back to school. One other reason can be the bad weather when walking on the road is discouraged which leads to increase in taxi rides.



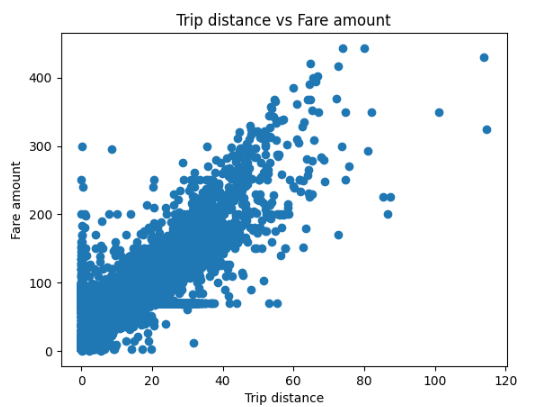
* + 1. **Find the proportion of each quarter’s revenue in the yearly revenue**

The proportion of each quarter’s revenue in the yearly revenue is as follows:



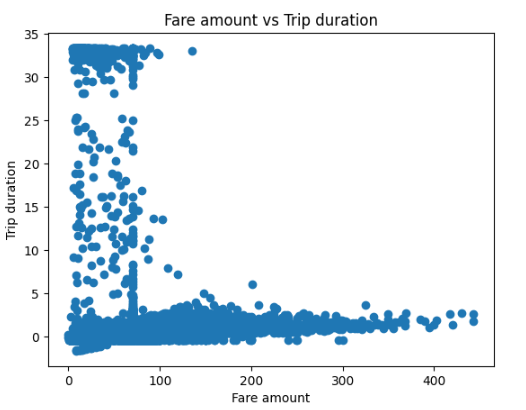
* + 1. **Analyse and visualise the relationship between distance and fare amount**

The relationship between distance and fare amount was evaluated using a scatter plot. From the plot, it is observed that the fare amount increases with the trip distance. The scatter plot for the same can be viewed below:

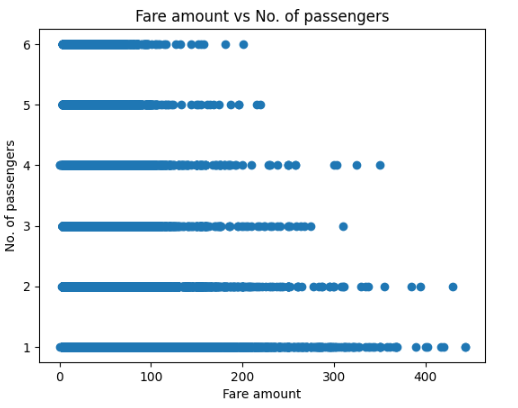
****

* + 1. **Analyse the relationship between fare/trip duration and trips/passengers**

The relationship between trip duration and fare amount was evaluated using a scatter plot. From the plot, it is observed that the fare amount remains at a high value even for short distances which can be attributed to the congestion in the NYC city. Whereas, there are other few instances where the fare amount is low for longer trip durations. This can be a case when the driver forgets to end the ride leading to a longer trip duration.



Further when the relationship between the number of passengers and fare amount was studied, it was observed that with increase in no. of passengers, the fare amount decresed. This means that the long rides are usually taken by less number of passengers since it is difficult to pool for such long diistances. While people maybe pooling up for short distances.

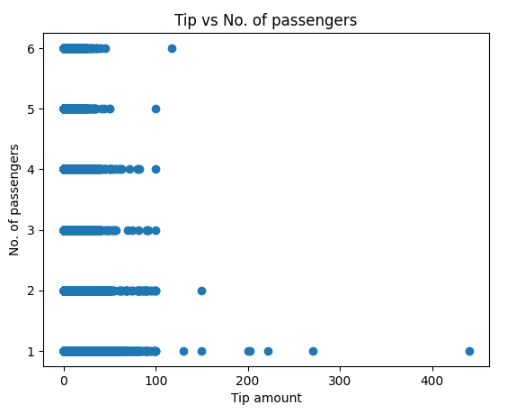


* + 1. **Analyse the relationship between fare/tips and tip/passengers**

The relationship between tip amount and fare amount was evaluated using a scatter plot. From the plot, it is observed that the tip amount surprisingly decrease with increase in fare amount . This trend can be attributed to person paying the amount from his/her pocket and thinking about the actual amount which has to be paid in total.

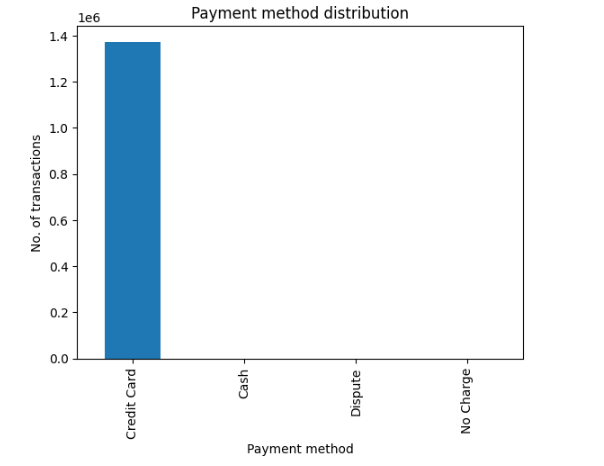
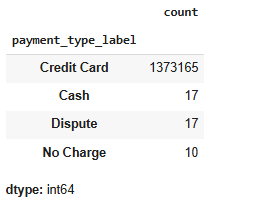
****

The no. of passengers vs tip amount study infers that with increase in no. of passengers, the tip amount decresed. This can be attributed to psychology of the passenger if only a single person is paying for the group, that person does not want to be burdened. Whereas, a single person does not mind paying more tip

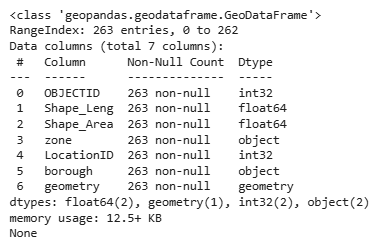
****

* + 1. **Analyse the distribution of different payment types**

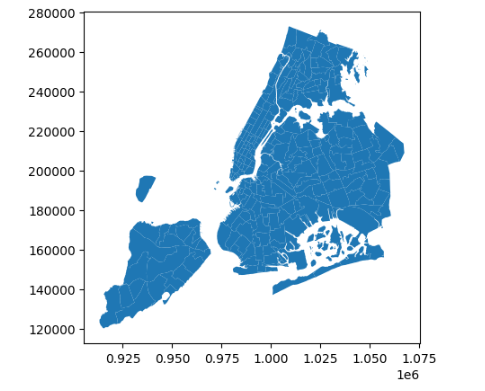
The distribution of different payment methods is evaluated using a bar plot as below. It is observed that the maximum trips were fulfilled using a credit card payment while other methods were used in near to zero instances as compared to credit card.

* + 1. **Load the taxi zones shapefile and display it**

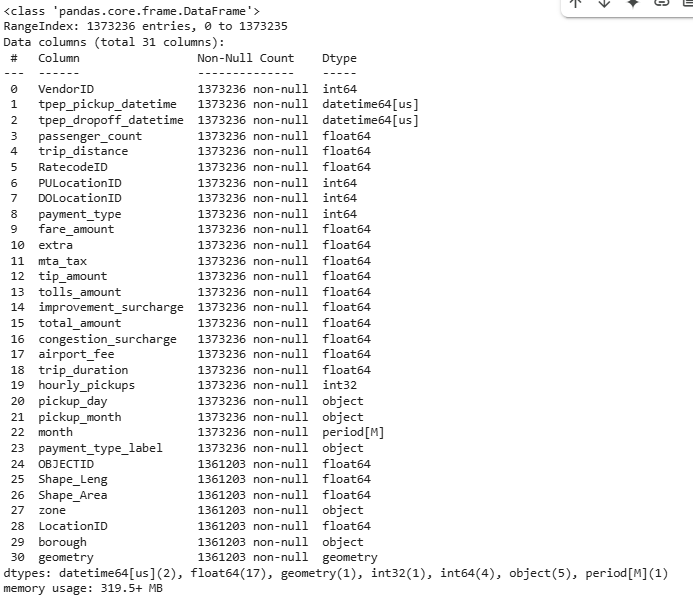


Following are the taxi zones from the shape file:

****

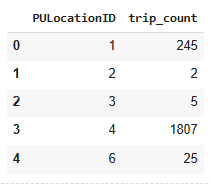
* + 1. **Merge the zone data with trips data**

The zones data was merged with trips data to form a new merged\_df dataset. Following is the information on the said dataset.



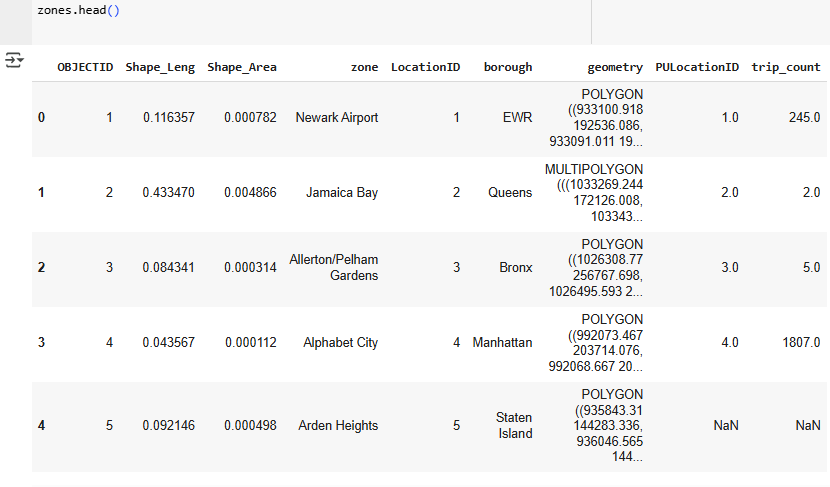
* + 1. **Find the number of trips for each zone/location ID**

Following are the number of trips for each location:



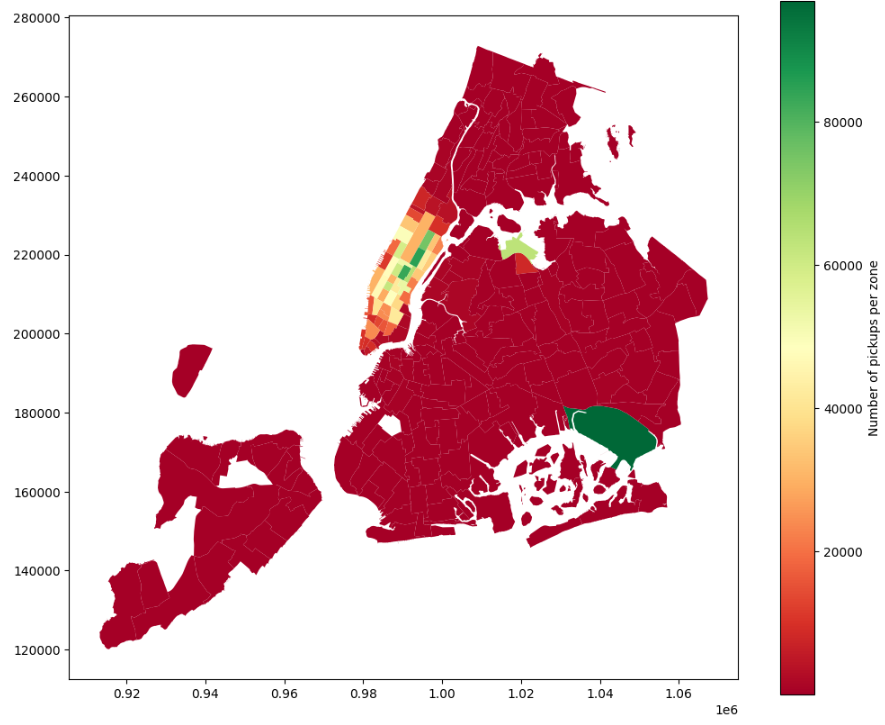
* + 1. **Add the number of trips for each zone to the zones dataframe**

Following is the output after adding the number of trips to the zones dataframe:



* + 1. **Plot a map of the zones showing number of trips**

Below is the map of the zones

****

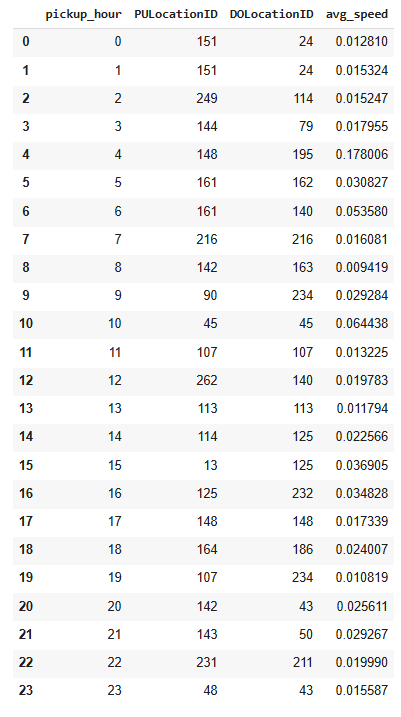
* + 1. **Conclude with results**

As per the plot in section 2.1.12, green colour denotes highest number of trips, yellow denotes average number of trips and red as lowest number of trips. If the zones are actually overlapped with NYC map, JFK and Laguardia airport shows the maximum trips.

### Detailed EDA: Insights and Strategies

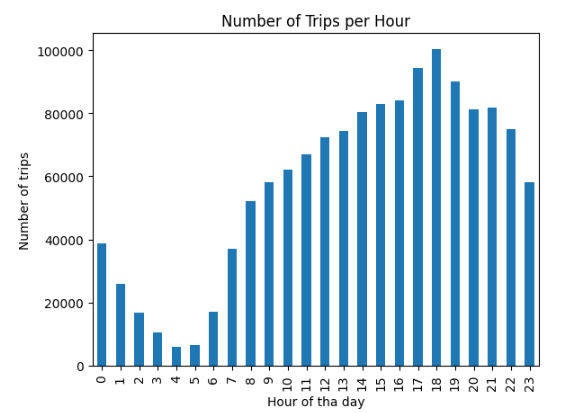
* + 1. **Identify slow routes by comparing average speeds on different routes**

The dataset was evaluated to check the slow routes by comparing average speeds. It was observed that the slowest routes are:



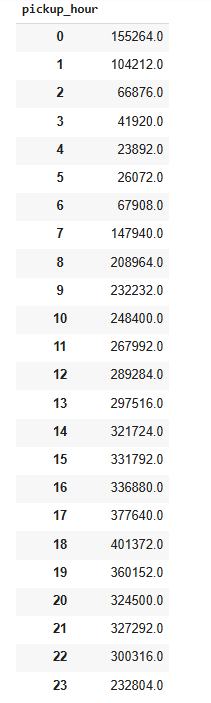
* + 1. **Calculate the hourly number of trips and identify the busy hours**

The hourly number of trips were evaluated and the busiest hours were from 5 pm to 7 pm with 6pm trips as the highest.



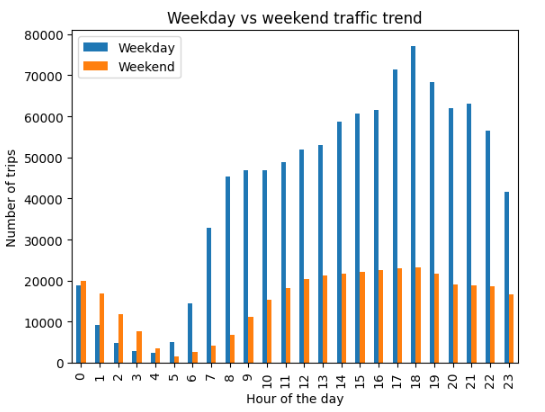
* + 1. **Scale up the number of trips from above to find the actual number of trips**

While performing a scale up from 25% of the data set, the actual number of trips per hour is as follows:



* + 1. **Compare hourly traffic on weekdays and weekends**

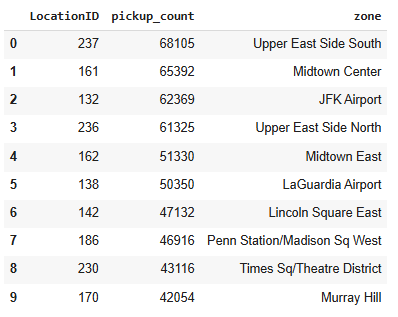
Following is the comparison of hourly traffic on weekdays (blue) and weekends(orange) performed using a barplot:

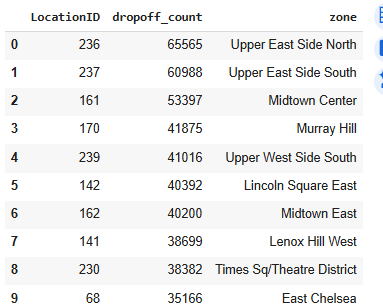


From the bar plot it is observed that, weekends have a slighlty higher number of trips during night time (12 am to 4am) due to the night life in NYC with an all time high day trips on weekdays.

* + 1. **Identify the top 10 zones with high hourly pickups and drops**

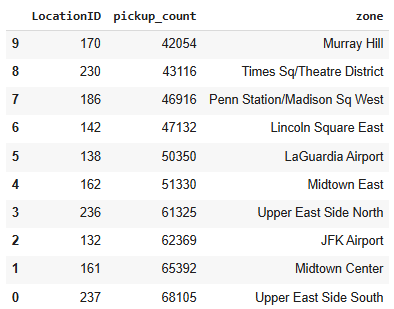
The top 10 zones with high hourly pickups and drops are as follows:

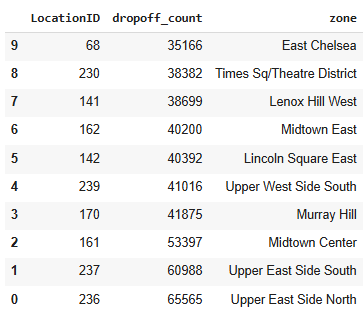




* + 1. **Identify the top 10 zones with high hourly pickups and drops**

The bottom 10 zones with high hourly pickups and drops are as follows:



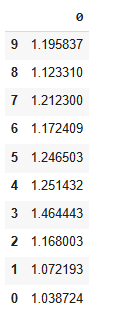


* + 1. **Find the ratio of pickups and dropoffs in each zone**

The ratio of pickups and dropoffs for each of the top 10 zones mentioned in section 2.2.5 are:

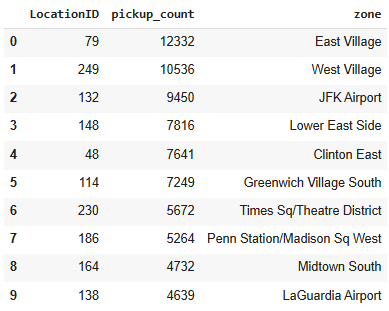


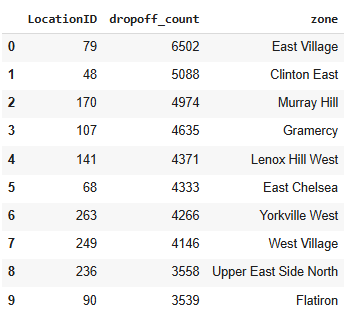
The ratio of pickups and dropoffs for each of the bottom 10 zones mentioned in section 2.2.5 are:



* + 1. **Identify the top zones with high traffic during night hours**

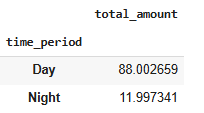
The top zones with high traffic during night hours are as follows**:**

****



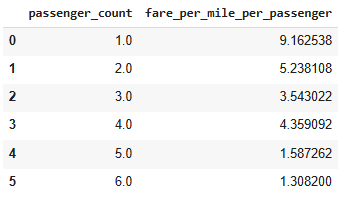
* + 1. **Find the revenue share for nighttime and daytime hours**

The revenue share for nighttime and daytime hours are as follows:



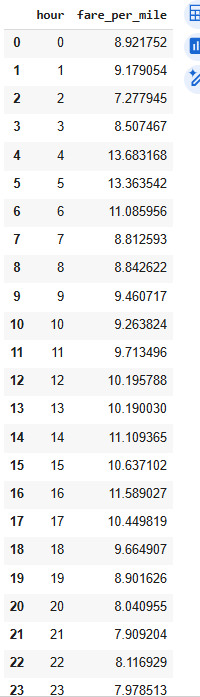
* + 1. **For the different passenger counts, find the average fare per mile per passenger**

For the different passenger counts, the average fare per mile per passenger is as follows:

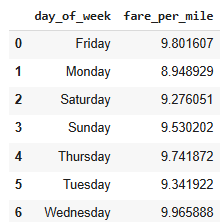


* + 1. **Find the average fare per mile by hours of the day and by days of the week**

The average fare per mile by hours of the day is as follows:

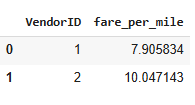


The average fare per mile by by days of the week is as follows:



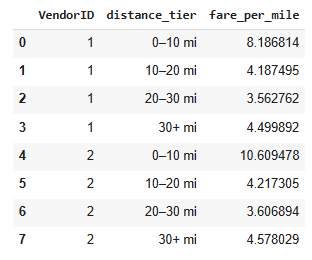
* + 1. **Analyse the average fare per mile for the different vendors**

The average fare per mile for the different vendors is as follows:



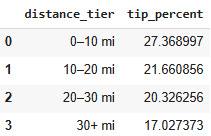
* + 1. **Compare the fare rates of different vendors in a distance-tiered fashion**

The fare rates of different vendors in a distance-tiered fashion are:

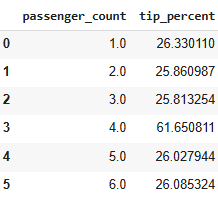


* + 1. **Analyse the tip percentages**

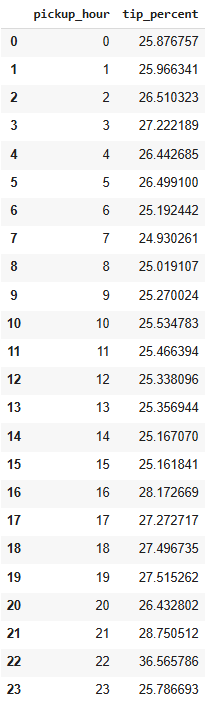
The tip percentage was analyzed based on the distance tier and following is the result when evaluated :



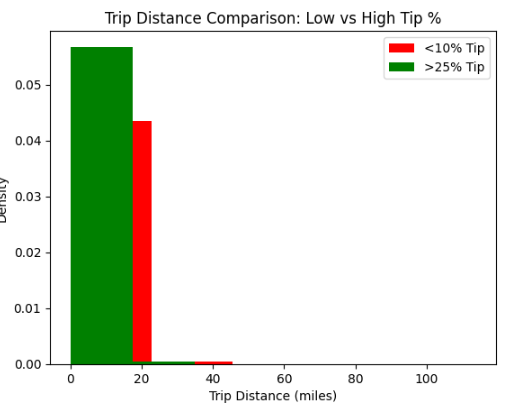
The tip percentage was analyzed based on the passenger count and following is the result when evaluated :



The tip percentage was analyzed based on hourly basis and following is the result when evaluated :



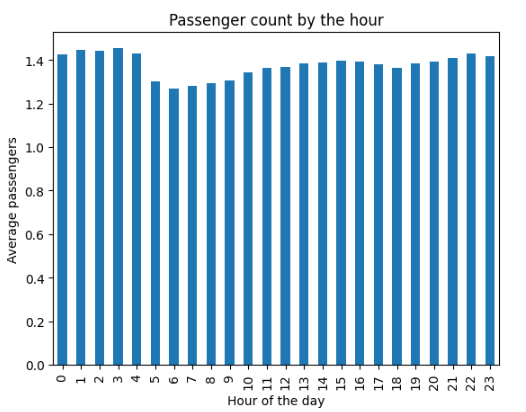
The trips were compared with tip percentage < 10% to trips with tip percentage > 25%



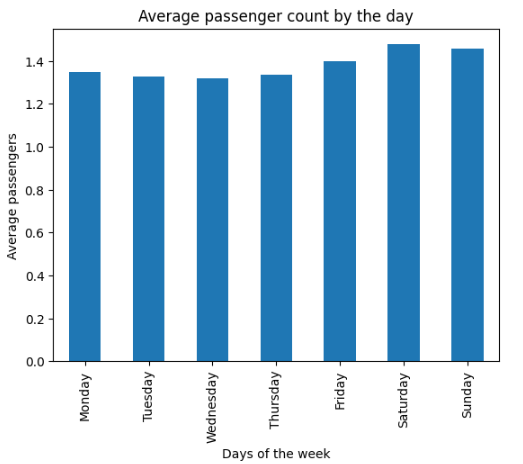
It was observed that people have tipped >25% for short distance trips (~<18 miles and 21-35 miles) and about 10% tip for distances between ~18-21 miles and >35 miles. Overall, the tipping amount is more for short distances as compared to long distance rides.

* + 1. **Analyse the trends in passenger count**

The trend in passenger count was analyzed using a bar plot. From the plot, it was observed that the highest passenger count is in the night to early morning period (9 pm to 4 am) attributing to the passenger thinking about its safety and wanting to travel in group during non-busy hours.

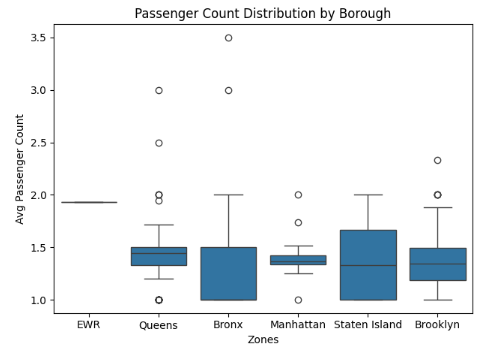


When passenger count was evaluated over the week, it was observed that the passenger count is more on the weekends which shows that people travel in group over the weekends assuming they are going to have a good time and relax.



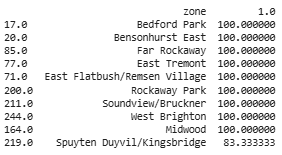
* + 1. **Analyse the variation of passenger counts across zones**

The variation of passenger counts across zones is evaluated using box plot. As per the evaluation, the central tendency is same for all zones except for EWR (upper central tendency) and Bronx (lower central tendency) region. The passenger counts are evenly distributed in the Queens, Staten island and Brooklyn zones. Overall, it can be said that the no. of pasengers travelling per ride in all the zones are 1-2 except for EWR zone which has a passenger count constantly at 2.

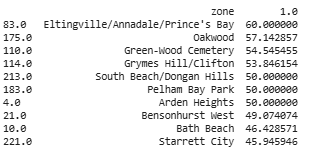


* + 1. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.**

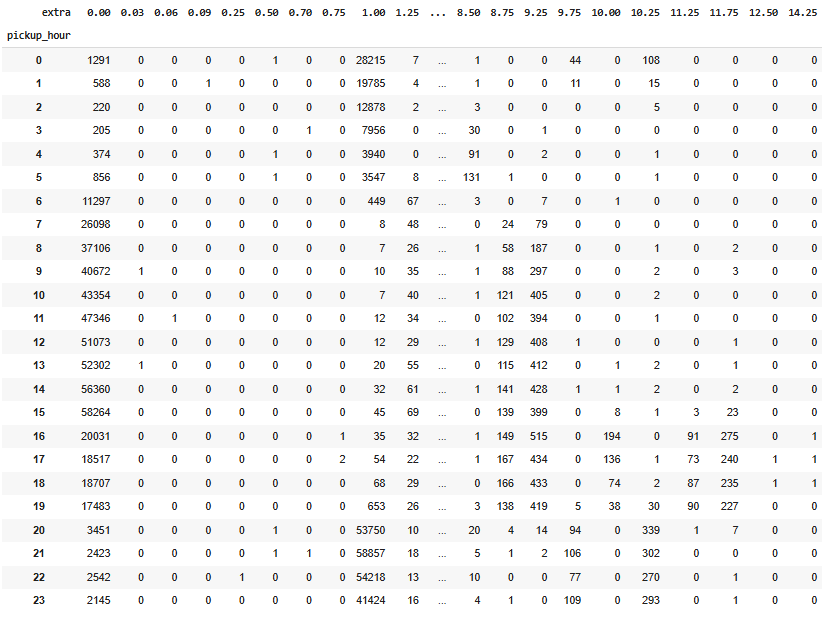
Here are the top 10 pickup zones when extra charges are applied more frequently. As per the evaluation an extra charge of 100$ is levied more frequently at the pick up zones.

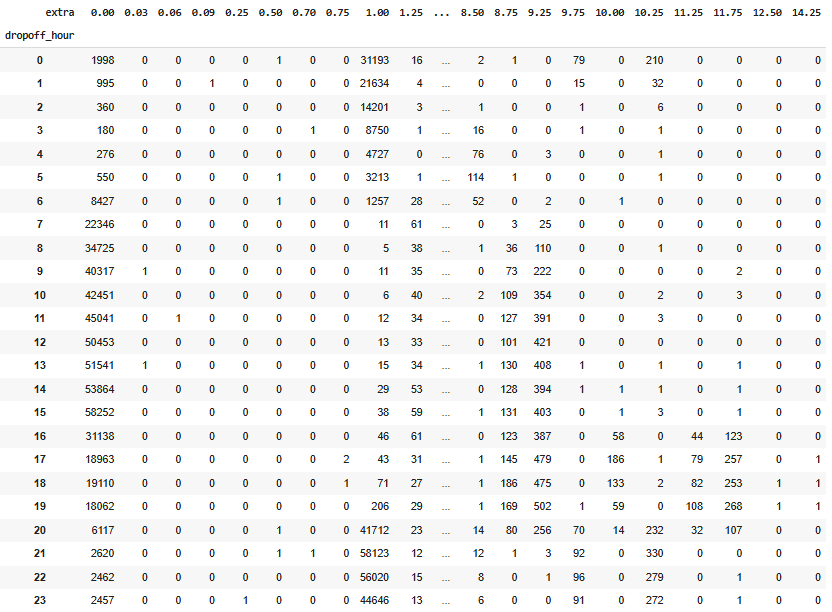


Here are the top 10 dropoff zones when extra charges are applied more frequently. As per the evaluation an extra charge of 50-60$ is levied more frequently at the drop off zones.



Here are the times and amount of extra charges applied during the day. It is observed that the most frequently applied amount is 1$ throughout the day. An increased extra charge (8-9$) is applied during office hours i.e 7 am to 7pm. This is true for both pick up and drop off locations.



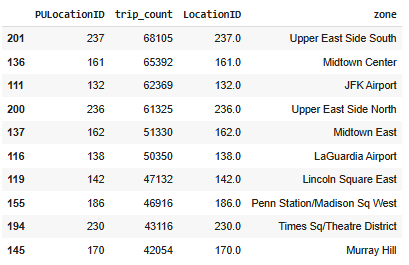


## Conclusions

### Final Insights and Recommendations

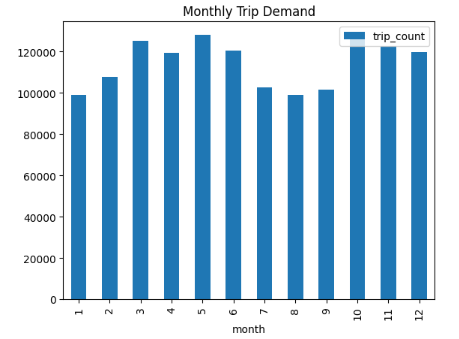
* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**

Reposition idle taxis toward high-pickup zones. Here are the top 10 zones:



* + 1. **Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analysing trip trends across time, days and months.**

Monthly trends show peaks in tourist-heavy months, increase the cab availability in the area for those months. Free cabs from remote areas or new cabs can be diverted to NYC region.



* + 1. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**

High demand (and tipping) during evening rush and weekend nights are observed. A dynamic pricing can be implemented with slight surcharges during peak hours (e.g., 5–8 PM, Fri/Sat nights) as seen before.

