CS215 - Assignment 5

Ujwal L Shankar (23b1050)
Madhav R Babu (23b1060)
Marumamula Venkata Pranay (23b1073)

Contents

Q1.ipynb and Q2.ipynb are identical and contains the codes for both Q1 and Q2.

1 Data Icebreaker

The data contains the sections

- **pickup_community_area** - Community area where the trip started
- **fare** - Fare charged for the trip
- **trip_start_month** - Month the trip started
- **trip_start_hour** - Hour the trip started
- **trip_start_day** - Day the trip started
- **trip_start_timestamp** - Timestamp when the trip began
- **pickup_latitude** - Latitude at the trip start location
- **pickup_longitude** - Longitude at the trip start location
- **dropoff_latitude** - Latitude at the trip end location
- **dropoff_longitude** - Longitude at the trip end location
- **trip_miles** - Distance traveled during the trip in miles
- **pickup_census_tract** - Census tract at the trip start location
- **dropoff_census_tract** - Census tract at the trip end location
- **payment_type** - Payment method used for the trip
- **company** - Company handling the trip
- **trip_seconds** - Duration of the trip in seconds
- **dropoff_community_area** - Community area where the trip ended
- **tips** - Tips given for the trip

1. We classified the data into
    - **Categorical:** pickup_community_area, trip_start_month, trip_start_hour, trip_start_day, payment_type, company, dropoff_community_area
    - **Numerical:** fare, pickup_latitude, pickup_longitude, dropoff_latitude, dropoff_longitude, trip_miles, trip_seconds, tips
    - **Mixed Data Type:** trip_start_timestamp, pickup_census_tract, dropoff_census_tract
2. We converted categorical columns into category data type for faster processing and efficient memory usage. Numerical columns were converted to numeric datatype for performing numerical operations on them. trip_start_timestamp was converted to timestamp data type.
3. Identifing missing values

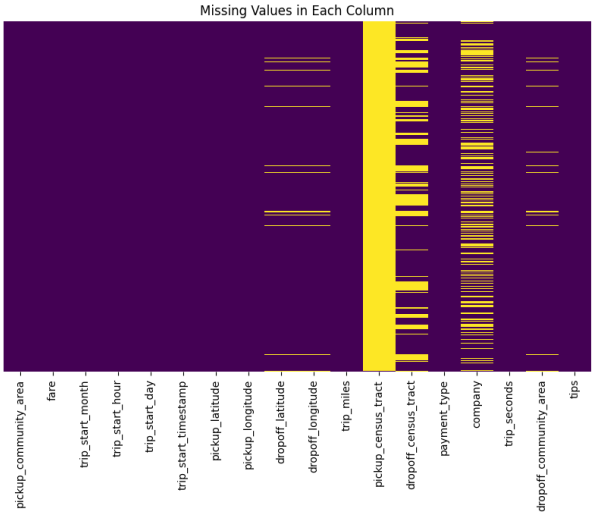

Figure 1: Missing Value Heatmap

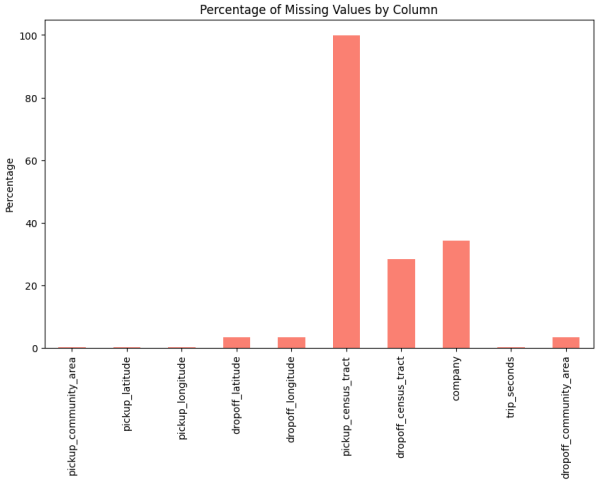

Figure 2: Missing Value Barchart

The column pickup_census_tract has the most number of missing values (150001) followed by company(5140) and dropoff_census_tract (4241).

4. We performed the following imputation:

   - Mean Imputation for numeric columns
   - Mode imputation for Categorical columns and Mixed Columns.

2 Rush Hour Rush

2.1 Trip Frequency by Time of Day, Month, and Day

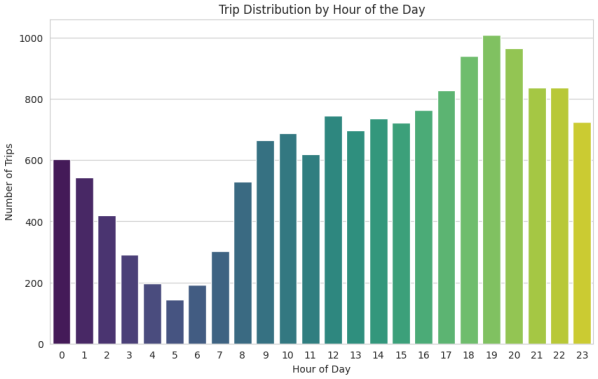

Figure 3: Trip Frequency by Hours of the Day

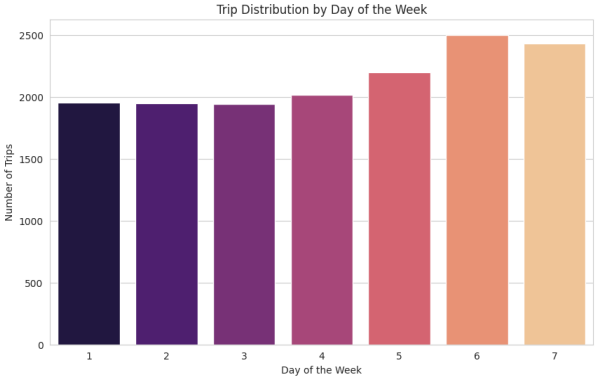

Figure 4: Trip Frequency by Day of the Week

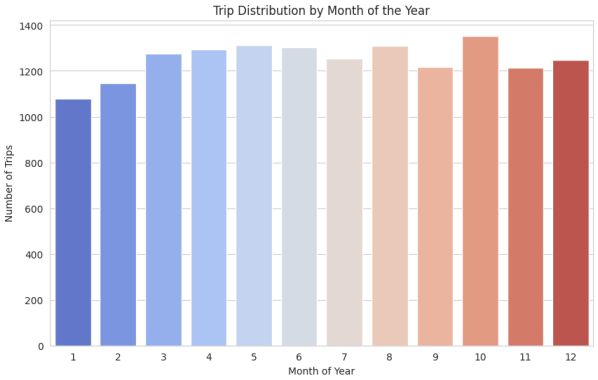

Figure 5: Trip Frequency by Month of the Year

a)

b) By looking at the plots we can infer that

- Trip Frequency vs Hour of the Day: Frequency of the trips peaks at 7 PM and there there exists a general increase in trips from 5 PM to 10 PM. There is a decreasing trend from 12 AM till 5 AM, and then starts increasing till it peaks at 7 PM.

- Trip Frequency vs Day of the Month: Trip frequency increases during weekend with peaking during saturdays.
- Trip frequency is low during January and February and it peaks at October. Freuqency is generally high during March, April, May, June, August and October.

These plots can be used to manage the demand by: hiring more workers around the peak months or adding temporary staff during weekends or increasing the number of workers in a particular shift when the frequency is high, etc.

2.2 Peak Hours and Distance Analysis

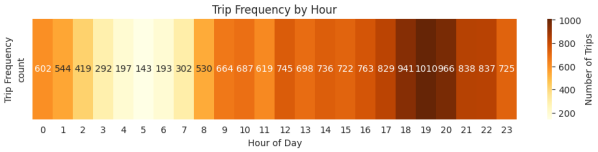

Figure 6: Heatmap of trip frequency by the hour

a) From this it is evident that the peak hours are 6 PM, 7 PM, and 8 PM.

b) Histogram for the the distribution of trip distances during peak hours without removing outliers indicate that a lot of data with 0 distance is present.

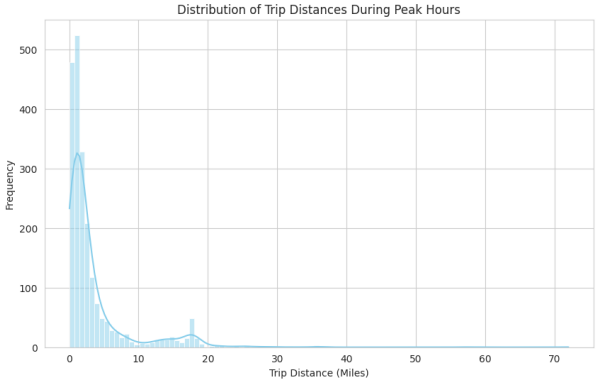

Figure 7: Trip Distance during Peak Hours with outliers

After removing the outliers using IQR rule (find the first and third quartile and removing data points that doesn't fall within Q1-1.5*(Q3-Q1) and Q3+1.5*(Q3-Q1)) we obtained the following plot.

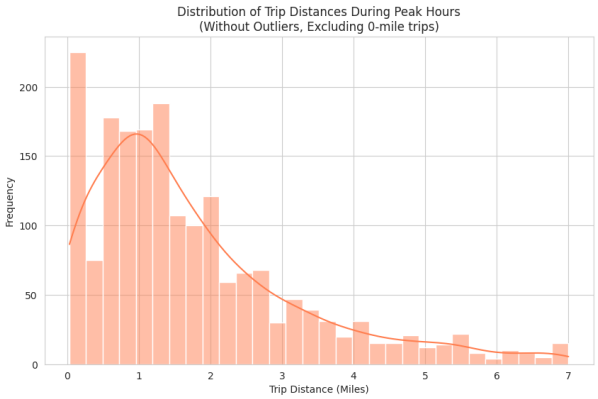

Figure 8: Trip distances during peak hours

c) The distance trends indicate that the during peak hours most trips are short distance trips this indicates a scenario where congestion and longer wait times significantly affect taxi service efficiency. As demand for taxis increases in congested areas, the time taxis spend in traffic reduces the overall efficiency of the service, and passengers experience longer waits.

2.3 Trip Duration and Distance Distribution

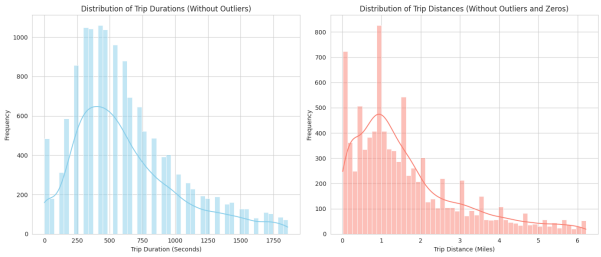

Figure 9: Histogram for trip duration and trip distance after removing outliers

a) From the figure we can see that both trip distance and trip duration peaks around a value and decreases to zero slowly after that.

b) Scatter plot of trip duration vs trip distance after removing outliers.

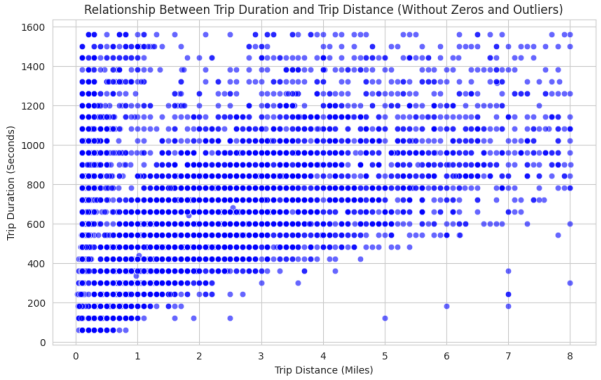

Figure 10: Trip Duration vs Trip Dsitance

c) There is a high occurrence of zero in the trip distance this is a noticeable outlier that affects the plot. On analyzing the trip duration vs trip distance plot we can see that there is a general increasing trend since the duration of the trip tend to increase as distance increase, however there exist shorter trips with longer duration implying the existence of traffic congestion, hence analyzing these routes can help in reducing the trip duration.

2.4 Seasonal Trends in Taxi Usage and Distance

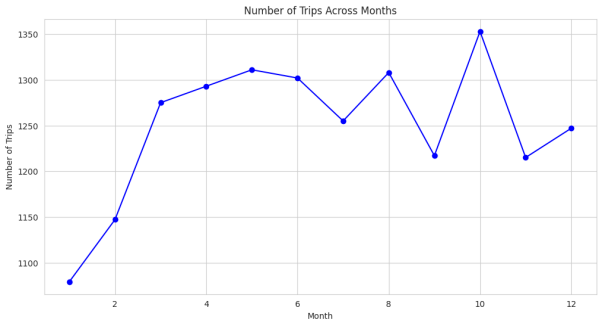

Figure 11: Number of trips per Month

a) The plot shows monthly trip counts with an upward trend from January to June, stability around mid-year, and a peak in October. There's a sharp drop in November, with a slight recovery in December. This pattern indicates seasonal demand fluctuations, possibly due to holidays in October.

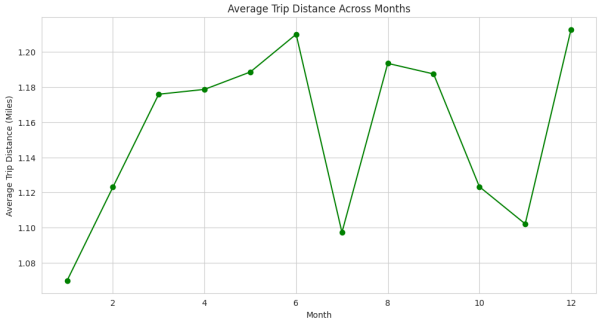

Figure 12: Average trip distance per Month

The plot reveals a peak in average trip distance in June and December, with a notable dip in July and October.

b) To prepare for high-demand periods:

- Increase drivers in October (trip spike) and adjust shifts in June/December (longer trips).
- Vehicle Maintenance: Allocate budget for extra maintenance and fuel in peak months.
- Inventory: Stock up on spare parts and fuel to reduce downtime and control costs.

2.5 Distance vs. Time of Day

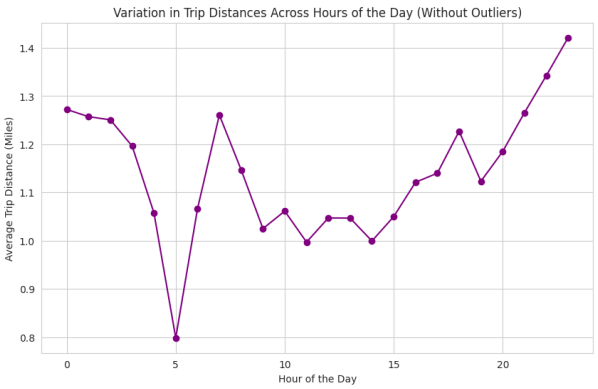

Figure 13: Average distance per Hour after removing outliers

a)

b)

- Early Morning Dip: From 3 AM to 5 AM, there is a sharp drop in trip distance, reaching the lowest point around 5 AM.
- Gradual Increase Throughout the Day: After 5 AM, average trip distances start to increase gradually, peaking again around 8 AM. This may correspond to morning commute times when people travel longer distances to work or school.
- Afternoon Stability: Between 10 AM and 3 PM, average trip distances remain relatively stable and lower compared to peak hours. This period likely represents midday trips, which are shorter on average, possibly for running errands or shorter travel needs.
- Evening Surge: From around 5 PM onward, there is a noticeable rise in trip distances, reaching the highest point by 11 PM. This surge in distance could reflect people going back from the office.
- Late-Night Peak: The highest average trip distance occurs late at night around 11 PM, possibly due to fewer short, local trips and more long-distance journeys.

3 Fare Frenzy

Q3.ipynb contains the answers in its markdown section

4 Data Mayhem

Q4.ipynb contains the answers in its markdown section