EXPLORATORY DATA ANALYSIS OF NEW YORK CITY TAXI TRIP DATASET

Md. Mahmudul Hasan

Digital Media & Informatics.
University of Bremen

1. Introduction

This is an extensive exploratory data analysis using Python and data visualization tools like matplotlib and seaborn for the New York City Taxi Trip (2022).

Predicting the length of taxi rides in New York City using parameters like route locations or the pickup date and time is the aim of this playground activity. We use pandas to import the dataset in order to begin the exploratory data analysis.

2. Dataset description and analysis

In this New York City Taxi Trip dataset there are 20 columns which means there are 20 types of variables to do the exploratory data analysis. Among the variable we have datetime data. To use this data, we have to convert in by the code in the Fig 1. Then we add three new columns for day, month and hour which is also extracted from the datetime variable.

```
df['lpep_pickup_datetime']= pd.to_datetime(df['lpep_pickup_datetime'])
df['lpep_dropoff_datetime']= pd.to_datetime(df['lpep_dropoff_datetime'])
df['pickup_day']=df['lpep_pickup_datetime'].dt.day_name()
df['dropoff_day']=df['lpep_dropoff_datetime'].dt.day_name()
```

Fig 1. Converting the data

3. Passenger Count in every Trips

The bar graph (Fig 2) shows the distribution of passenger counts in taxis. The x-axis represents the number of passengers in each taxi, ranging from 0 to 8, while the y-axis represents the count of specific the number.

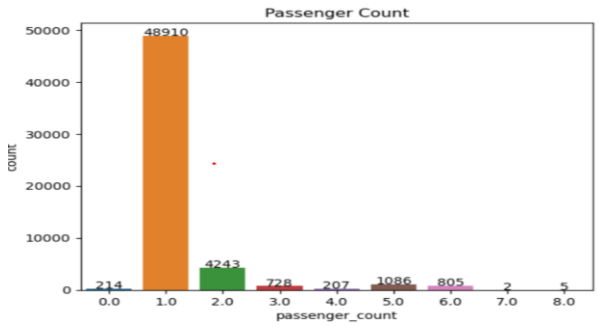

Fig 2. Number of passengers Count

A significant majority of taxis, precisely 48,910, have only one passenger. Taxis with two passengers are also common but significantly less frequent at a count 4,243. The frequency diminishes drastically for higher passenger counts.

4. Trip count in every weekday

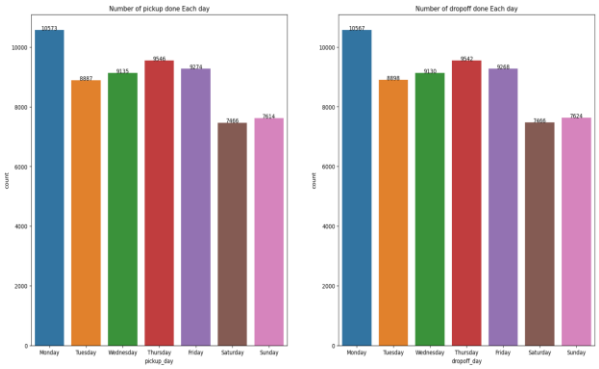

Fig 3. Number of trips in weekdays

Based on this graph, we can conclude that both pickups and drop-offs peak on Monday, with 10,567 instances respectively, before experiencing a gradual decline throughout the week.

5. Trips count per-day.

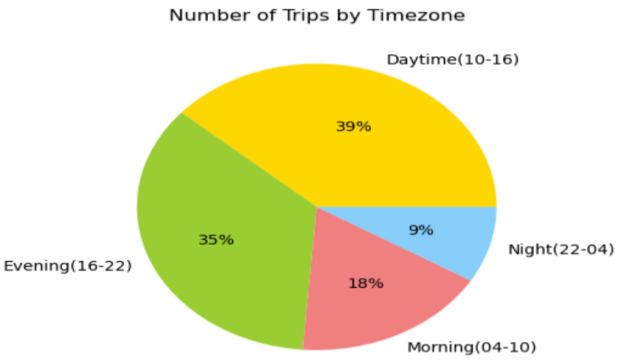

Fig 4. Number of trips per hour

The pie chart titled “Number of Trips by Time zone” visually represents the distribution of trips taken during different times of the day. The largest segment, colored in yellow and labeled as “Daytime,” accounts for 39% of the total trips, indicating a preference for travel during these hours. The green “Evening” section follows closely at 35%, while the “Morning” and “Night” time zones are represented by smaller segments at 18% and 9%, respectively.

6. Trips distance.

Here we can see that there er many trips under 0 km.

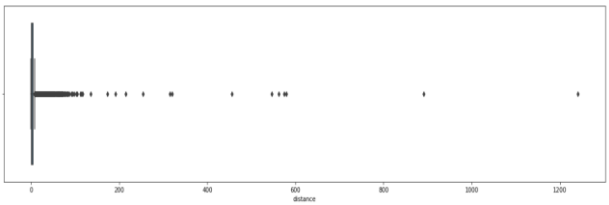

Fig 5. Trip Distance

7. Relationship between the fare amount and tip amount

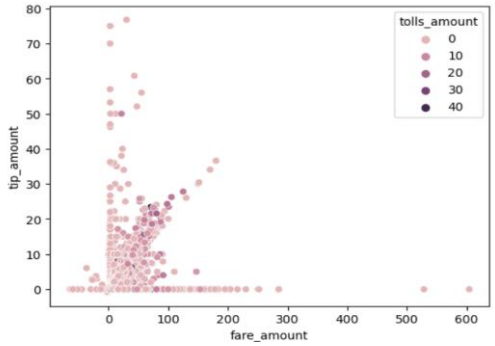

Fig 6. Fare amount and tip amount

The graph is a scatter plot that illustrates the relationship between the fare amount and tip amount, with a color-coded representation to indicate varying toll amounts. The data points are scattered, showing a concentration of lower fare and tip amounts, with fewer instances of higher values. The color intensity increases with the toll amount, providing insights into how tolls impact the correlation between fares and tips.

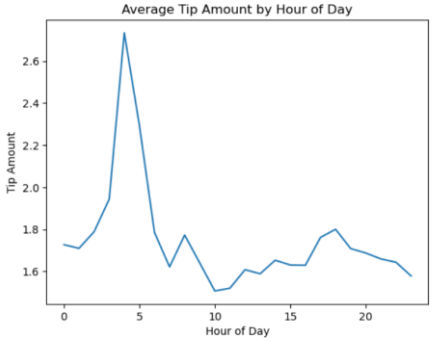

Fig 7. Average Tip per hour.

The graph titled “Average Tip Amount by Hour of Day” shows a distinct pattern in tipping behavior over a 24-hour period. It shows a significant spike in the average tip amount around the 5th hour of the day, reaching nearly 2.6 units before sharply declining. For the rest of the day, tips fluctuate mildly but remain relatively low, hovering around 1.6 to 1.8 units.

8. Best hour to move a location in a Day.

In this graph we can find out the number of trips made to Location ID 74-75 at different hours of the day because this area has a maximum number of trips.

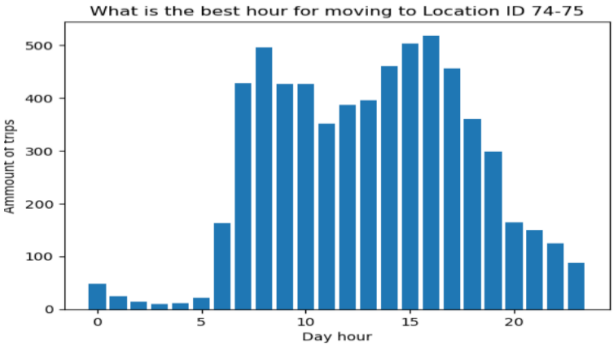

Fig 8. Higher chance of getting another trip from the same location

A significant increase in trips is observed starting from the 10th hour, peaking at around the 15th hour with over 500 trips. After reaching its peak, there’s a gradual decline in trip frequency as it moves towards later hours. The graph shows a strong correlation between the lpep_pickup_datetime and DOLocationID variables. Specifically, it indicates the hours during which taxis should drop off passengers at location IDs 74 and 75 to increase their chances of getting another trip from that same location.

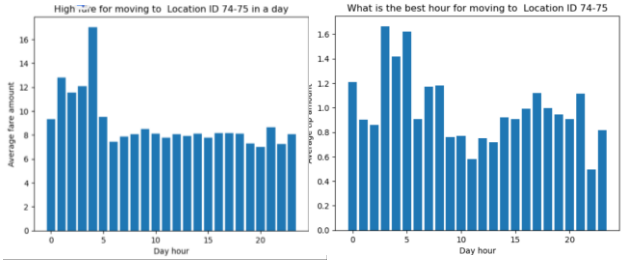

Fig 9. High tip and fare in Location Id 74 and 75.

The right side of the graph indicates that taxis cover long-distance trips in the fifth hour for the same location, resulting in higher fare amounts. On the left side of the graph, the third and fifth hours show high tips amount due to the high fare or long-distance trips.

9. Conclusion

In conclusion, the data presented in the graphs can be used to optimize the transportation of goods and people to Location ID 74-75, schedule trips during the peak hours for maximum efficiency, and plan logistical operations.