Exploratory Data Analysis: A case study on exploratory data analysis of the famous New York Taxi Trip dataset

Author: Sarah Islam Momo

I. Abstract

This case study presents an exploratory data analysis of the famous New York Taxi Trip dataset. The dataset comprises detailed information on taxi trips in New York City, including pickup and drop-off locations, trip duration, fare amount, and other variables. My analysis is conducted using various techniques of data visualization and statistical analysis and hypotheses to extract insights from the data. The study involves data pre-processing and cleaning to ensure data accuracy, followed by exploratory data analysis techniques such as data visualization and statistical analysis to uncover patterns and trends.The findings of this study reveal interesting patterns and trends in taxi usage, such as the busiest pick-up and drop-off locations, the most common trip duration etc.

II. Introduction

EDA is a fundamental step in data analysis that allows us to gain an understanding of the data and its underlying structure, providing insights that can guide further analysis.[1] New York Taxi Trip dataset, contains detailed records of numerous numbers of taxi trips made in New York City over several years. This dataset has been widely used for various data analysis projects due to its size, complexity, and potential insights. In this case study, I will perform an exploratory data analysis (EDA) on the New York Taxi Trip dataset to uncover patterns and insights from the data.

III. Observations from initial Data Analysis

From the initial data analysis, I observed that the New York Taxi Trip dataset contains 62495 rows and 20 columns which represents the records of taxi trips made in New York City in the year 2022. The dataset includes various trip attributes, such as pickup and drop-off times and locations, trip distances, fare amounts, passenger counts etc. I have also observed that the dataset contains some missing values and outliers, which may need to be addressed during the data cleaning process. My initial observations are listed below:

- The column ehail_fee contains null values.
- Although there are auto-generated IDs for each trip, there is no specific attribute to specify each trip individually. For better understanding, I have added one with the Trip_ID prefix as a new attribute.
- There are no invalid entries in the dataset, such as symbols like −, ?, ⋆ etc.
- The column pickup_month only contains January and December
- The column dropoff_month only contains January, February and December. I assume this can be possible if a trip start at the last day of January and end in February.
- There is no column which contains information about the trip duration and vehicle speed. But there are information like pickup and dropoff time, trip distance by which these new features can be created.
- There are columns such as store_and_fwd_flag, RatecodeID, passenger_count, payment_type, trip_type, congestion_surcharge attributes which contain 6295 missing values.
- 214 entries with 0 passenger count.

IV. Result: Univariate Analysis

At first ehail_fee which contained null value is dropped to remove outlier from the dataset. Entries containing 0 passenger count or 7 or 9 is removed considering them as outliers. I assume there were trips that initially were accepted by the drivers but later cancelled by the passenger. By looking at the columns which contain trip information such as store_and_fwd_flag, RatecodeID, passenger_count, payment_type, trip_type, congestion_surcharge carrying missing values this statement can be justified. Also, I noticed that these columns were replaced by NAN values. Initially, I tried to handle these missing values by creating new columns with NAN.[2] However, looking carefully at the passenger count information, this is clearly noticed that there are trips which include 0, 7 and 9 passengers which is irrelevant. After dropping these rows, the missing values were also optimized. Most of the trips took 0 - 10 minutes to complete i.e. approx 600 secs. Some of the trips were 30 minutes long. Though there were not significant amount of difference between the two Vendors, vendor 1 has higher trip count then the other one. There are 3274 trip records with 0 miles distance. This might be because of the reason that, customers changed mind and cancelled the journey just after accepting it or there might be other reasons. Most of the rides are completed between 1-10 miles with some of the rides with distances between 10-20miles. According to these data I considered 10 as a difference point between long and short trip. After that I found among the trips that 3566 Trips were Short and 19 Trips were Long. Most of the passengers travelled solo. The number of trips for both pick up and drop off were higher on Mondays and lower on Sundays. This might be because of the fact that Offices start in Mondays and Sunday is vacation. Also, Early morning pickups seems consistently low. Pickups seems to be consistent across the week at 15 Hours. Late night pickups were low too. The hours between 09 to 12 hrs seems busy.

V. Result: Bivariate Analysis

I made several hypotheses about having relations between the trip attributes and according to that analysed the dataset. Though Vendor 1 had more trips comparing to Vendor 2, Vendor 2 has trips carrying more passengers. Trip duration in mid day were larger than other part of the day. Also, among the week days trips which were recorded in Thursday and Friday were lengthy. Trip distance in early morning specifically 07 hrs are long. On Sundays the trip distance are longer than any other days of the week. This might me because that, people tends to travel to longer distances to visit friends and family in weekends to celebrate off day. Observing trip duration and store and forward flag data, this is clearly visible that most of the trips records were held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server. Observing the vehicle speed it is noticed that, late night trips are faster comparing to other time of the day, assuming the reason being in late nights there are less number of vehicles on the road.

The figures of Diagrams containing these results analysis can be found in the annex section of this document.

VI. Conclusion

In conclusion, the exploratory data analysis of the New York Taxi Trip dataset has provided valuable insights into the patterns and trends of taxi trips in New York City. Through the analysis of various variables such as passenger count, trip distance, trip duration, pickup and drop-off locations, and time of day, I have gained a better understanding of the factors that influence taxi trips in the city. I observed from the provided dataset that the majority of taxi trips in New York City are short trips, with a trip distance of less than 10 miles and a trip duration of less than 10 minutes. I also observed that there is a high demand for taxi services during Mondays. Overall, the insights gained from the exploratory data analysis of the New York Taxi Trip dataset can be useful for various stakeholders in the transportation industry, such as taxi companies, city planners, and policymakers, to make data-driven decisions and improve the efficiency and effectiveness of the transportation system. Further analysis and modeling on this dataset can also provide more detailed insights into the taxi industry in New York City. [2]

VII. References

[1] https://arrow.apache.org/ [2] https://www.analyticsvidhya.com/ [3] https://www.nyc.gov/

VIII. Appendix

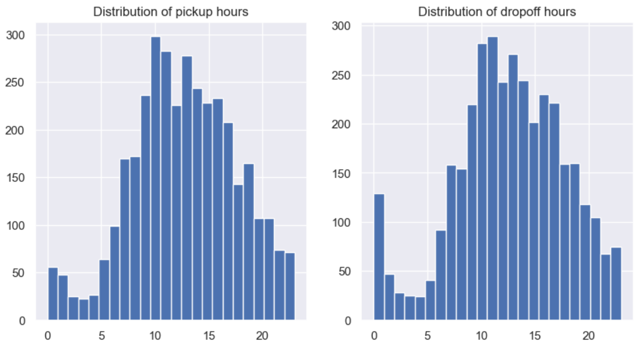

Figure 1. Distribution of trips per pickup hourse

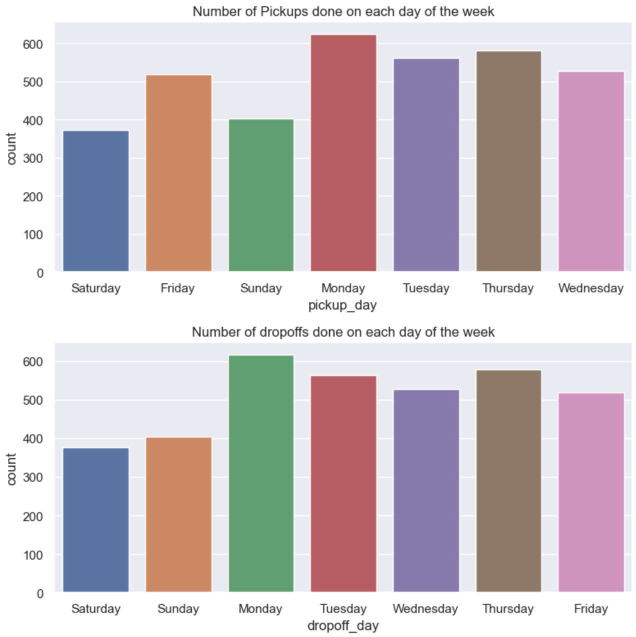

Figure 2. Number of pickups each day of the week

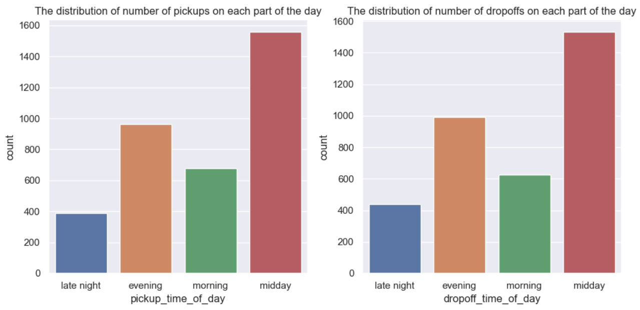

Figure 3. Number of pickups in several times of the day

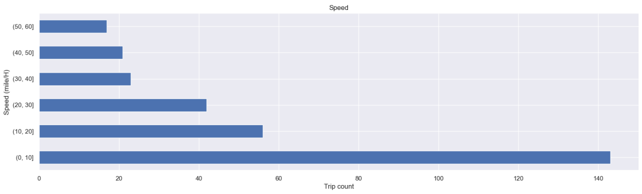

Figure 4. Vehicle Speed

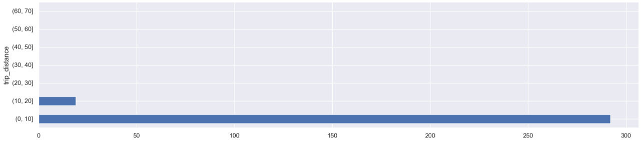

Figure 5. Trip distance

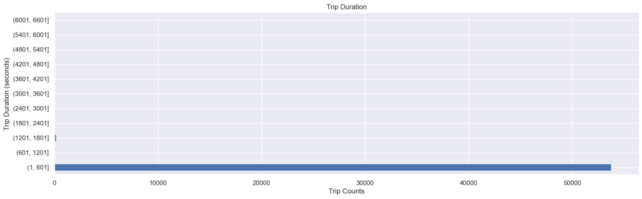

Figure 6. Trip duration

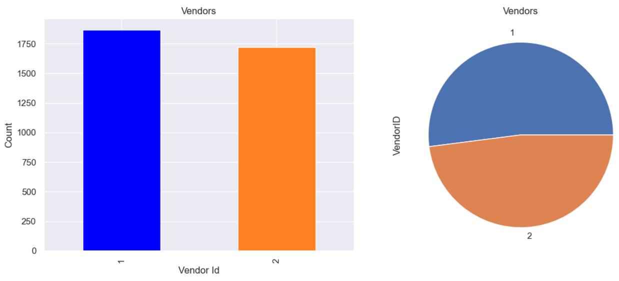

Figure 7. Vendor ID

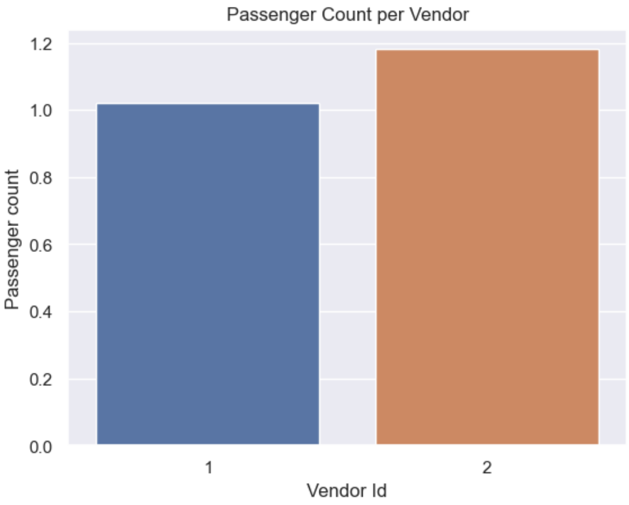

Figure 8. Passenger count per vendor

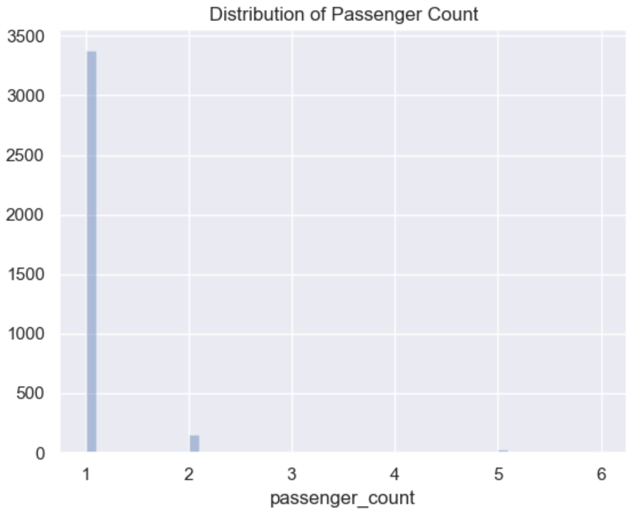

Figure 9. Passenger count

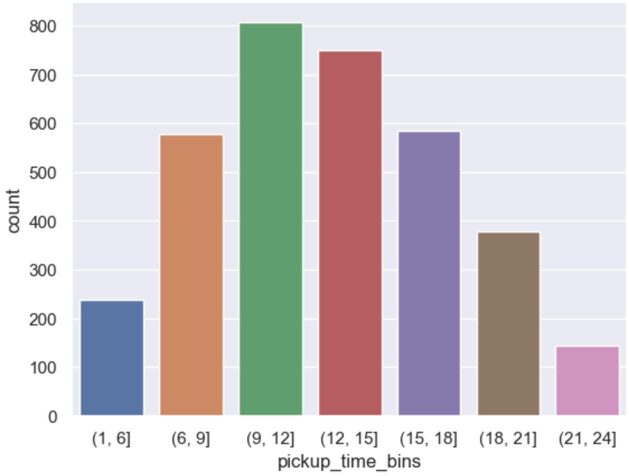

Figure 10. Number of pickups in different time bins

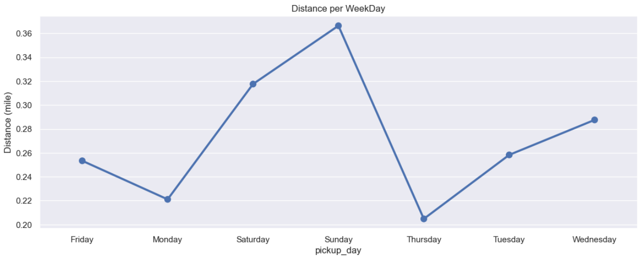

Figure 11. Relation between distance and pickup day

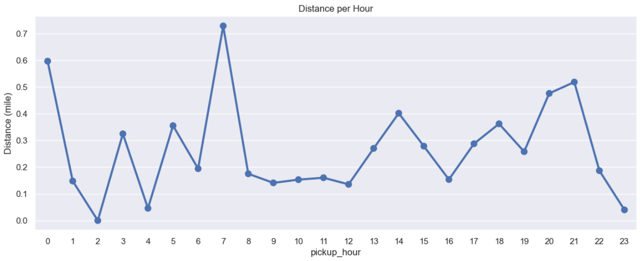

Figure 12. Relation between distance and pickup hour

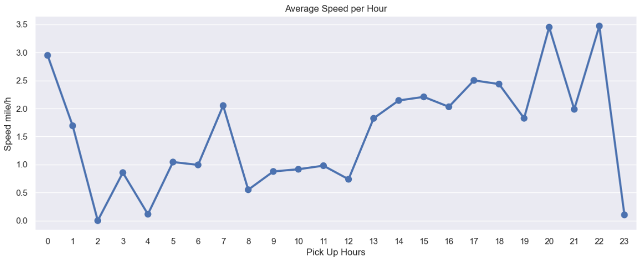

Figure 13. Relation between speed and pickup hour

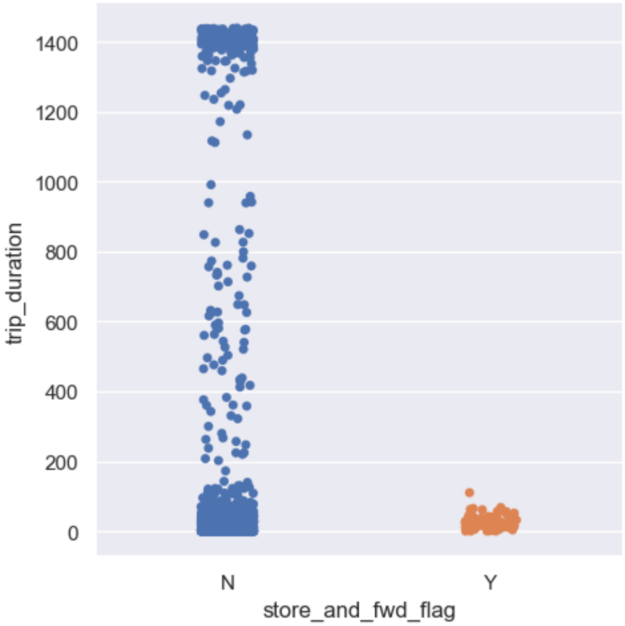

Figure 14. Relation between store and forward flag with trip duration

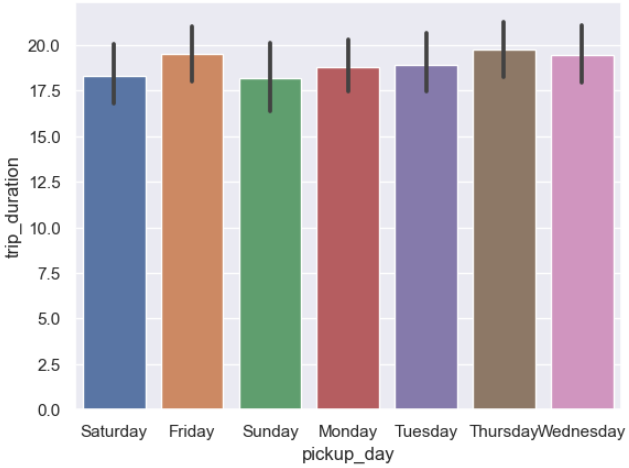

Figure 15. Relation between trip duration and pickup day