Exploratory Data Analysis on NYC TLC's Green Taxi Dataset

Novruz Mammadli – M. Sc. Student of Hochschule Bremen

Abstract — **Exploratory data analysis (EDA) is a powerful tool for understanding any dataset, and in this article, EDA techniques were applied to the green taxi dataset for January 2022 in New York City. The dataset includes detailed trip-level data for green taxis, such as trip distance, pickup and drop off zones, passenger counts, and fare amounts. By analyzing this dataset, insights were gained into the taxi industry for the month of January in NYC. Various EDA techniques were used to identify patterns and trends, such as the busiest times of day and the most popular pickup and drop off locations. The impact of external factors such as weather and holidays on taxi demand also was explored. This analysis provides a detailed understanding of the green taxi industry in NYC for January 2022 and can be used by researchers, policymakers, and industry practitioners to make informed decisions. This article offers a step-by-step guide to conducting EDA on the green taxi dataset, which can serve as a helpful resource for others looking to explore this dataset or similar transportation datasets in the future.**

I. INTRODUCTION

THE taxi industry plays an important role in New York City's transportation ecosystem, and the green taxi service is a vital part of this system. The green taxi service was introduced in NYC in 2013 as a new option for passengers who needed a ride in underserved areas outside of Manhattan. The service was initially designed to provide transportation for people in the outer boroughs, where traditional yellow taxis are less common. Since its introduction, the green taxi service has grown to become a popular option for New Yorkers and visitors alike, with tens of thousands of daily trips in the city [1].

II. INFORMATION OF DATASETS AND DATA CLEANING

A. Information

The shape of dataset is 62494×20, which represents rows and columns respectively. Information about each column was given in provided Jupyter Notebook file. Other datasets, such as Taxi Zones [1] and Weather of NYC in January 2022 [2] also was added to this research to see impacts of pick up and drop off areas and weather conditions to taxi trips.

B. Data Cleaning

The dataset contained some trips with a duration of 0 seconds or 24 hours, which were considered outliers. Durations less than 60 seconds and more than 150 minutes were also removed. And if distance equals to 0 and more than 100 km also removed from dataset. Furthermore, since this dataset contains only January data, three records determined to belong to December were removed.

C. New Columns

New columns were created in the green taxi dataset, such as pickup hour, weekday, and other time periods to better analyze and compare different time periods. Conversion operations were also performed, such as converting miles to kilometers, Fahrenheit to Celsius, and inches to millimeters. In the original dataset, trip duration, average speed (from direct distance) and fare per km information were not available, so these columns were created using the existing data.

III. EXPLORATORY DATA ANALYSIS (EDA)

A. Visualization and Analysis

In this section, exploratory data analysis techniques were applied to the data to gain a better understanding of the dataset. The first analysis focused on the vendors, and as shown in Figure 1 of the notebook, VeriFone Inc. accounted for 87.4% of all taxi trips, which is significantly higher than Creative Mobile Technologies. Figures 2 and 3 in the notebook show the distributions of fare and distance, respectively. Figure 4 in the notebook plots trip duration against trip count, revealing that most trips lasted about 5-10 minutes (~30%). It is worth noting that almost 80% of all trips were between 0-20 minutes.

In **Fig. 1**. (Figures 8, 9, and 10 in the notebook) shows that passengers used green taxis slightly more on Mondays than on other days, with nearly 10,000 trips counted on Mondays. However, the least number of trips were seen on Saturdays, and there were more trips on weekdays than on weekends. The reasons behind these patterns will be discussed in the next chapters. The hourly distribution plot reveals that the peak of taxi trips occurred between 14:00 and 18:00, while between 00:00 and 05:00, the number of trips was less than 1,000. The analysis of monthly trends shows that the number of trips decreased rapidly on the 29th of January and gradually increased on the following days. In Figure 11 in the notebook, a heatmap displays the pickup distribution by weekday and hour. Passengers started to use green taxis at earlier hours on weekdays than on weekends.

The number of passengers who shared trips is also an important factor in taxi trips. Most passengers (over 40,000 trips) traveled alone in green taxis, while the second most common number of passengers was 2, but this only reached 5,000 trips. This suggests that most passengers prefer to travel alone with a taxi. This plot can be seen in Figure 12 in the notebook.

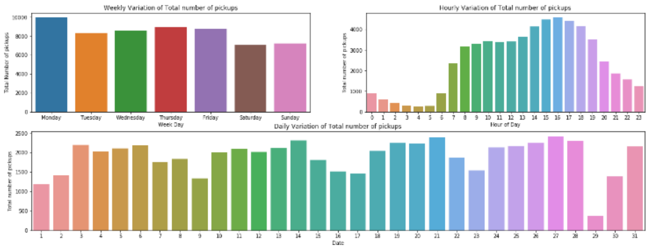

Fig. 1. Plot of weekdays, hours and monthly days vs trip count

Hourly and daily revenues are also important data. If we look at the hourly and daily revenues of green taxis in January, we can see that revenue and the number of trips are linearly correlated.

In this section, the effects of snow and rain on taxi trips will be briefly discussed. **Fig. 2** shows the height of snow and rain on the ground in January 2022. As mentioned before, on the 29th of January, taxi trips decreased rapidly and there was nearly 19cm of snow on the ground. This led people to stay at home, and it also made it difficult to drive taxis. On other days when the number of taxi trips decreased, it was observed that it was rainy or snowy. However, on the 7th of January, when it was both snowy and rainy, taxi trips looked normal. The reason for this could be the holiday of Orthodox Christmas, as everyone wants to celebrate it with their family and may use taxis to reach their desired destination.

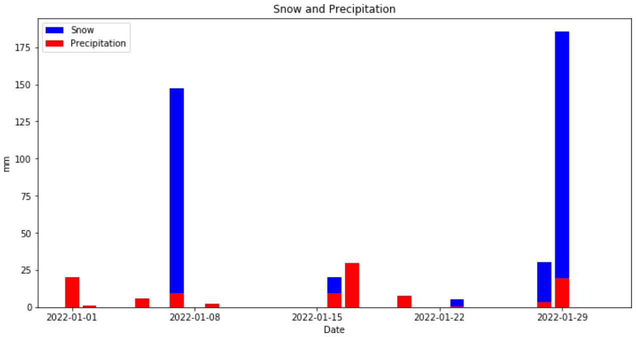

Fig. 2. Plot of Snow and Rain

TABLE I
Top 5 Pick-up and Drop off Zones.

Top 5 Pick-up Zones	Top 5 Drop off Zones
East Harlem North	East Harlem North
East Harlem South	East Harlem South
Central Harlem	Central Harlem North
Morningside Heights	Central Harlem
Central Harlem North	Upper East Side North

Table I shows Top 5 pick-up and drop off zones in New York City, which Top 10 zones presented in Figure 16. and 17. East Harlem North, East Harlem South and Central Harlem zones are the most frequent zones in this both pick-up and drop off zones which shows that green taxis worked in these areas too much.

Figure 18. presents the trip type which 1 refers to street hail, 2 refers to dispatch, more than 90% of trip types were street hail type.

Payment type is also another important data here, there were 6 different payment types [1]:

- 1: Credit Card
- 2: Cash
- 3: No charge
- 4: Dispute
- 5: Unknown
- 6: Voided trip

In Figure 19. it is observed that most of the payments paid by Credit Card and then paid with Cash. Payment 6 is not observed, and rest of the payment methods were too less.

Figure 20 shows a plot of trip duration vs. trip distance, which reveals that most of the trips were under 40 km and 80 minutes. Figures 21 and 22 display the hourly average speed and the week daily average speed, respectively. These graphs show that the peak of average speed was observed during periods with fewer taxi trips (fewer trips between 0:00 and 05:00 for the hourly graph and on Saturdays and Sundays for the daily graph).

B. Correlation

To plot correlation absolute values of correlation used, instead of getting correlation between -1 and 1, here aimed to get values between 0 and 1. Which -1 correlated variables also correlated in a spatial domain, but they are vectors with opposite directions.

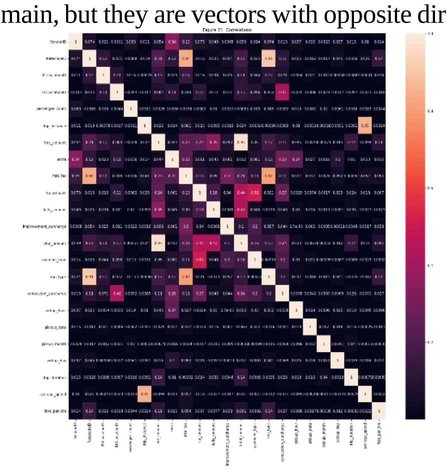

Fig. 3. Absolute correlation of dataset

This correlation shows trip distance and average speed are correlated. And other correlated values can be seen from the correlation plot.

IV. CONCLUSION

Weekends there are less demands to green taxis than other weekdays. Weather conditions and holidays are also affecting the trip numbers. Revenue is dependent of the number of trips. The top pick-up and drop-off locations remain nearly consistent, which could help green taxi drivers plan their routes to maximize their earnings. Tip amount only correlated with payment method which shows that the customers who pay with cash tend to give higher tips compared to those who pay with credit card.

V. REFERENCES

[1] "TLC Trip Record Data," TLC Trip Record Data - TLC. [Online]. Available: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. [Accessed: 16-Feb-2023].

[2] National Centers for Environmental Information (NCEI), "Climate Data Online," Climate Data Online (CDO) - The National Climatic Data Center's (NCDC) Climate Data Online (CDO) provides free access to NCDC's archive of historical weather and climate data in addition to station history information. | National Climatic Data Center (NCDC). [Online]. Available: https://www.ncei.noaa.gov/cdo-web/. [Accessed: 16-Feb-2023].

[3] "Exploratory Data Analysis on NYC taxi trip duration dataset." [Online]. Available: https://www.analyticsvidhya.com/blog/2021/01/exploratory-data-analysis-on-nyc-taxi-trip-duration-dataset/. [Accessed: 16-Feb-2023].