EDA on the Green Taxi trips data set for January 2022

I. Introduction

New York City is one of the busiest cities in the world, with millions of residents and visitors moving through its streets every day. Taxis have long been a popular transportation option for New Yorkers and tourists alike, providing a convenient and efficient way to get around the city. The Green Taxi program, which began in 2013, has added a new type of taxi to the mix, providing a more affordable option for trips outside of Manhattan. This program has since expanded to cover all five boroughs.

In this report, we will conduct an exploratory data analysis (EDA) on the Green Taxi trips data set for January 2022. The purpose of this analysis is to explore the patterns and trends in taxi trips for the new year. Specifically, we will examine the various factors that influence taxi trips, including pick-up and drop-off locations, trip distance, fare amounts, payment types, and more.

By analyzing these factors, we hope to gain insights into the demand for taxi trips in New York City and how these trips vary by time and location. We also hope to identify any unique patterns or trends that may be emerging in the Green Taxi market. Ultimately, this analysis can provide valuable information for taxi drivers, companies, and policymakers, as well as anyone interested in understanding the dynamics of transportation in a bustling city like New York.

II. Data Description

The dataset used in this report is green_tripdata_202201, containing 62,495 records and 20 columns. The dataset consists of taxi trips data in New York City in January 2022. The data contains information about the pickup and dropoff times, locations, fare amounts, and other details about the taxi trips.

The dataset has 6 columns with missing values (Figure 1), including store_and_fwd_flag, RatecodeID, passenger_count, payment_type, trip_type, and congestion_surcharge. The missing values in the store_and_fwd_flag and RatecodeID columns were filled with the most frequent value (mode), while the passenger_count and congestion_surcharge columns were filled with the mean. The missing values in the payment_type and trip_type columns were filled with the most frequent value (mode).

Several new variables were added to the dataset, including the duration of the trip, hour, minute, and day of the week of the pickup and dropoff times, the hour of the day of the pickup time, and the date of the day of the pickup time. A new feature 'Time of day' was also added based on the hour of the day.

The location IDs in the PULocationID and DOLocationID columns were mapped to unique integer values using the dictionary location_dict.

Overall, this dataset provides a comprehensive view of the taxi trips in New York City in January 2022 and allows for the exploration of patterns and trends in taxi trips.

III. Univariate Analysis

This report presents the findings from the univariate analysis conducted on a dataset of taxi rides in New York City. The analysis explores the distribution of various variables such as vendor ID, rate code, location ID, passenger count, trip distance, trip duration, pickup and dropoff times, and day of the week.

Vendor ID (Figure 2): The data shows the count of trips taken by each Vendor for the month of January. Vendor 2 has significantly more trips (53073) than Vendor 1 (9201). This could be due to various factors such as more drivers, better pricing, or better marketing, which might lead to more customers using Vendor 2

Rate Code (Figure 3): The RatecodeID column represents the final rate code in effect at the end of the trip. Here are some explanations for the trends observed in the count of each rate code:

- 1.0: Standard rate - This is the most frequent rate code, with 59582 trips. It represents a regular taxi trip.
- 2.0: JFK - This code is used for trips from JFK airport. It represents only 119 trips, which is much smaller than the standard rate.
- 3.0: Newark - This code is used for trips from Newark airport. It represents only 37 trips, which is much smaller than the standard rate.
- 4.0: Group ride - This code is used for group rides. It represents only 53 trips, which is much smaller than the standard rate.
- 5.0: Negotiated fare - This code is used for trips that have negotiated fares that are different from the standard fare. It represents only 2704 trips, which is relatively small compared to the standard rate.

In summary, the vast majority of trips have the standard rate code, and only a small number of trips have a different rate code. The second most frequent rate code is for negotiated fares, and the other codes represent only a very small number of trips.

Location ID: The dataset contains 62495 unique location IDs for both pickup and dropoff locations. The frequency

of the location ID is highly variable, with some locations appearing in over 200 trips and some appearing only once.

Passenger Count (Figure 4): The majority of trips had only one passenger, with passenger counts ranging from 1 to 6.

Trip Distance (Figure 5): The distribution of trip distance is highly skewed, with the mean distance being 78.03 miles and a standard deviation of 2914.49 miles.

Trip Duration (Figure 6): The distribution of trip duration is also highly skewed, with a mean duration of 1143.10 seconds and a standard deviation of 4700.99 seconds.

Pickup and Dropoff Times (Figure 7, 8): The most common pickup and dropoff hours are between 4:00 PM to 6:00 PM (16:00 to 18:00), followed by 6:00 PM to 8:00 PM (18:00 to 20:00). The frequency drops during late night and early morning hours, with the lowest number of pickups and dropoffs between 4:00 AM to 5:00 AM. This trend is expected since most people use taxis during rush hours to commute to and from work, while there is less demand during late night and early morning hours.

Day of the Week (Figure 9, 10): Both pickup and dropoff day of week counts show similar trends.

The highest number of trips occur on weekdays, with the most trips on Mondays and the fewest on Fridays. This suggests that the majority of the rides are used for commuting or work-related purposes.

The number of trips decreases on weekends, with Sundays having the fewest trips. This trend is expected since weekends are usually for leisure and relaxation.

Time of Day (Figure 11): The highest frequency of trips was in the ”Evening” category, which suggests that a large number of trips occur during the evening hours. The next most frequent category was ”Morning”, followed by ”Afternoon”. The ”Late night” category had the fewest number of trips, suggesting that fewer people take trips during the late night hours.

Overall, the univariate analysis revealed some interesting patterns in the dataset.

IV. Bivariate Analysis

Bivariate analysis was conducted to identify any trends and patterns in the January trip data. The relationships between various variables were examined to gain insights into how they might be related to one another. The variables analyzed included trip distance, fare amount, passenger count, pickup day of the week, and pick up hour of the day.

A comparison of trip distance and fare amount (Figure 12) revealed that short trips tended to have higher fare amounts than longer trips, which may be due to a fixed starting fee. For longer trips, the fare amount was more consistent and typically ranged between near to 0 miles.

Passenger count was found to be related to trip distance (Figure 13), with the majority of trips being taken by single passengers or those traveling with just one other person. This could be due to a variety of factors, including personal errands, business trips, or commuting to work.

A comparison of the pickup day of the week and mean trip duration (Figure 14), (Figure 15) revealed longer trip durations on weekdays, particularly on Thursdays, and shorter trip durations on weekends. This could be due to differences in commuter traffic and leisurely or recreational travel patterns.

V. Conclusion

In conclusion, through our exploratory data analysis, we have been able to identify trends and patterns in taxi trips for the month of January in New York City. The analysis of the data has provided us with insight into the relationships between various variables in the dataset.

We observed that for very short trips, the fare amount is significantly higher than for other distances, which could be attributed to some fixed starting fee. Longer trips, however, seem to have more consistent fare amounts. Additionally, we found that trips with single passengers or 1.2 passengers had the highest frequency of occurrence.

Furthermore, we found that trip durations were longer on weekdays and shorter on weekends, which could be attributed to the difference in commuter traffic. Finally, we noted that there were two peaks in the distribution of pickup hours, with the highest number of taxi rides being taken during the morning and evening rush hours, reflecting the work and commuting patterns of people in New York City.

These trends and patterns can help us understand the behavior of taxi riders in New York City, which can be useful for taxi companies and city planners. With this information, they can better optimize their services to cater to the needs of the riders and improve the efficiency of transportation systems in the city.

Overall, EDA is an important step in the data analysis process, as it provides us with a preliminary understanding of the dataset, and helps us identify potential relationships and trends that can be further explored in subsequent analysis.

APPENDIX

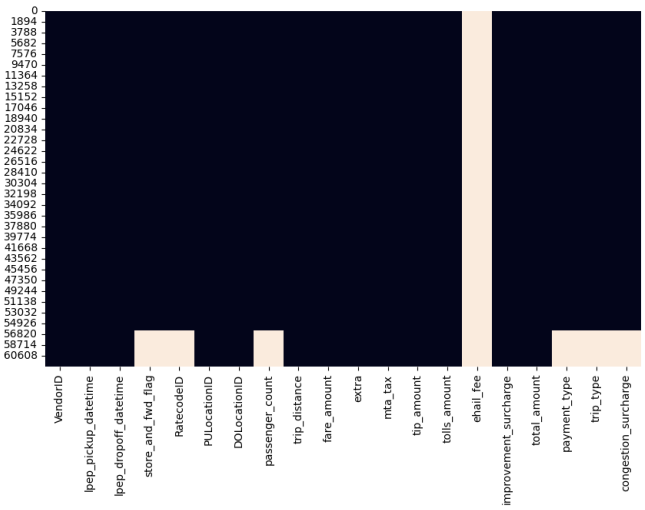

Fig. 1. Heatmap to visualize the pattern of missing values

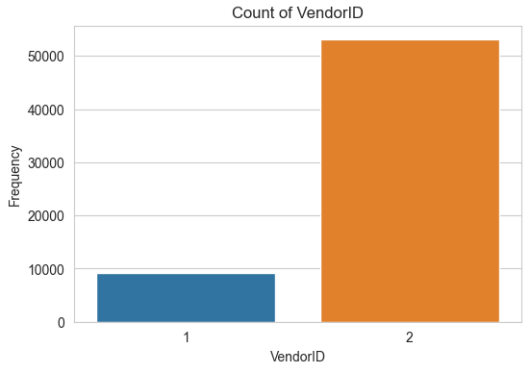

Fig. 2. Count of VendorID

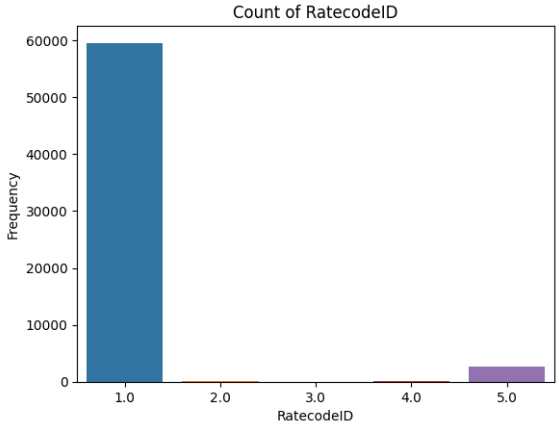

Fig. 3. Count of RatecodeID

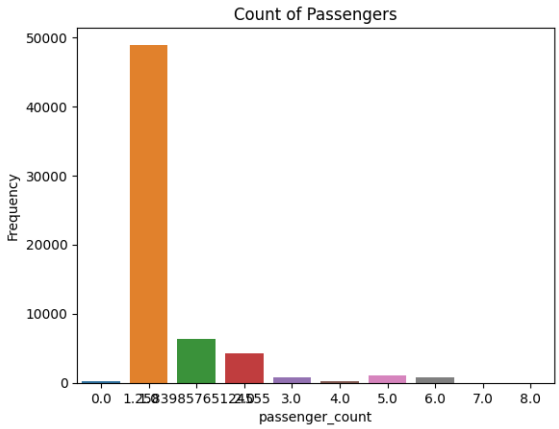

Fig. 4. Count of Passengers

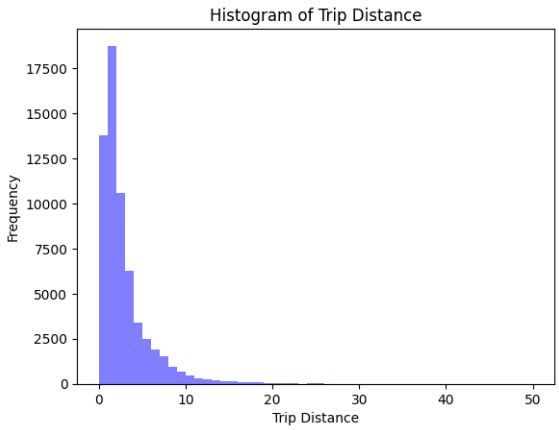

Fig. 5. Histogram of Trip Distance

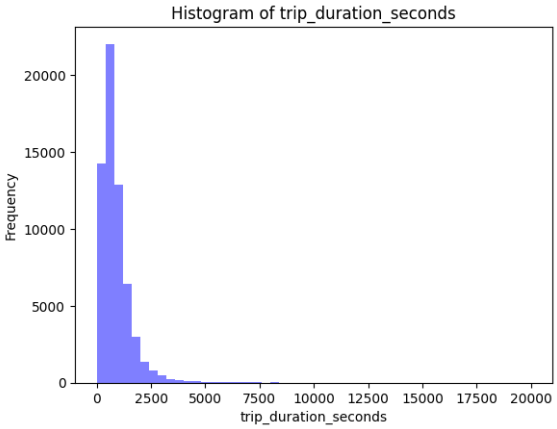

Fig. 6. Histogram of trip duration in seconds

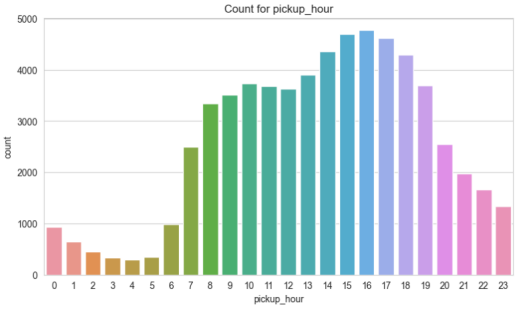

Fig. 7. Count for pickup hour

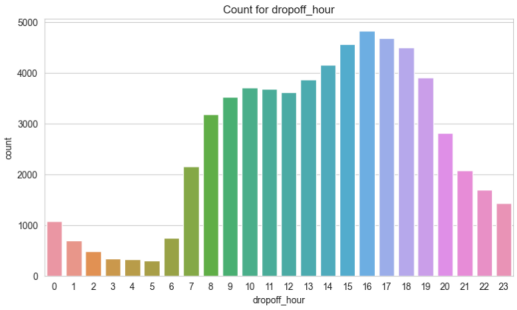

Fig. 8. Count for dropoff hour

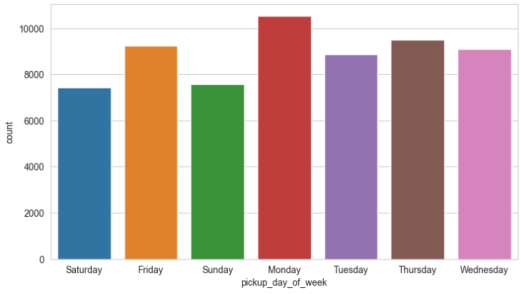

Fig. 9. Trip Count for pickup day of week

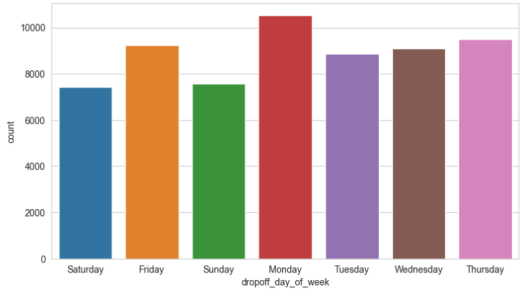

Fig. 10. Trip Count for dropoff day of week

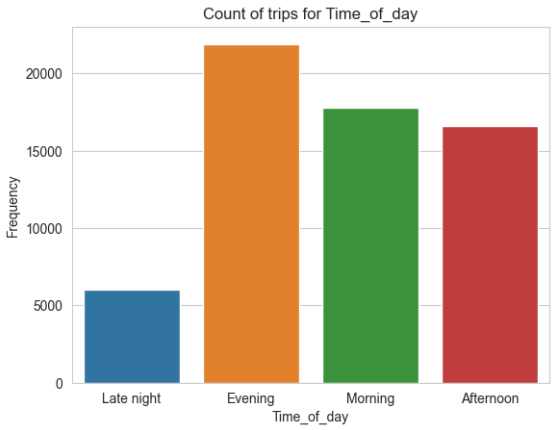

Fig. 11. Count of trips for Time of day

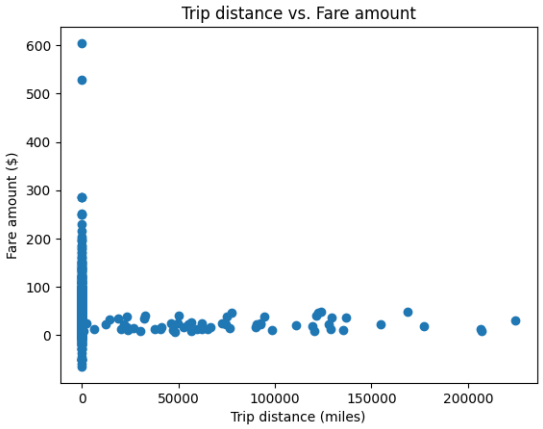

Fig. 12. Trip distance vs. Fare amount

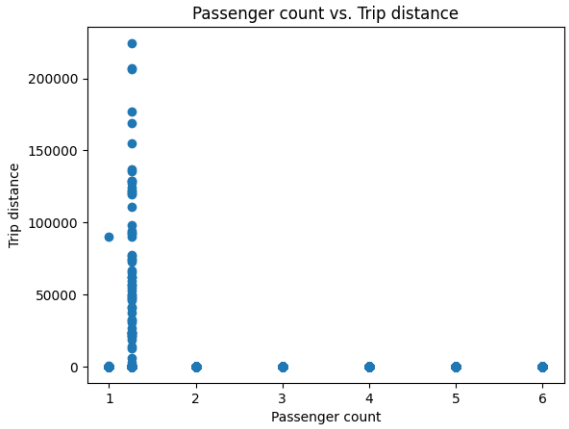

Fig. 13. Passenger count vs. Trip distance

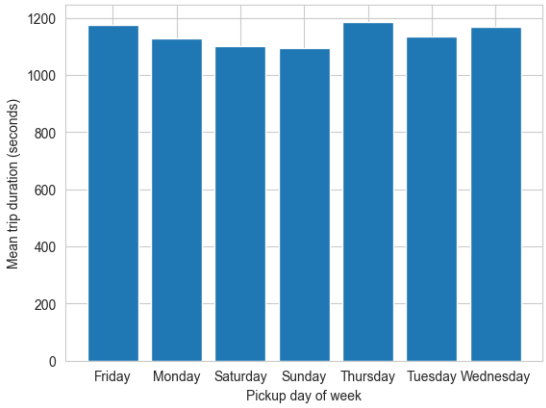

Fig. 14. Mean trip duration (seconds)

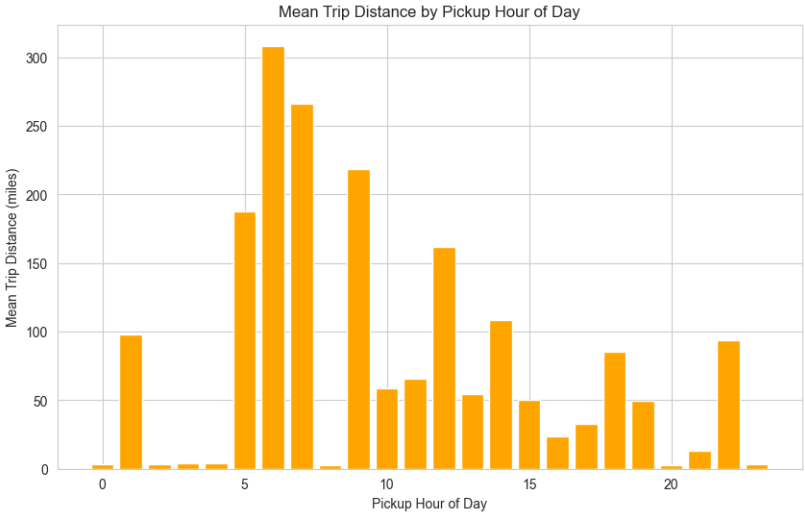

Fig. 15. Mean Trip Distance by Pickup Hour of Day