Exploratory Data Analysis (EDA) Report for New York City Yellow Taxi

Parag Parashar

AI/ML Batch - 77

Objective

- Uncover insights to help optimize Taxi Operations

- Analyse patterns in the data to define strategic directions
  - To improve service efficiency
  - To maximise revenue
  - To enhance passenger experience

Dataset Overview

Data Source: 2023 NYC Taxi trip data

- The data was collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers like vendors and taxi hailing apps
- The data is made available as 12 parquet files each representing a month's data
- Data captures
  - Pick-up and drop-ff dates/times
  - Pick-up and drop-off locations
  - Trip distances
  - Itemized fares
  - Rate types
  - Passenger counts
  - Other related data

Data Cleaning & Preprocessing

As it is a very large dataset, and based on the recommendation the report is built of a sampled dataset

Sampling Methodology for Dataset

- Data taken for 5% per hour / per day / per month
- Parquet file created using sampled data from 12 files which is then used in the EDA

Data Cleaning & Preprocessing

- No columns were dropped
- Column mismatch: airport_fee and Airport_fee. Renamed in 1 file to standardize name
- Converted required negative values to positive values
- Dropped records for peculiar case of null values found across columns (same records)
  - Passenger_count, RatecodeID, congestion_surcharge, Airport_fee

Data Cleaning & Preprocessing

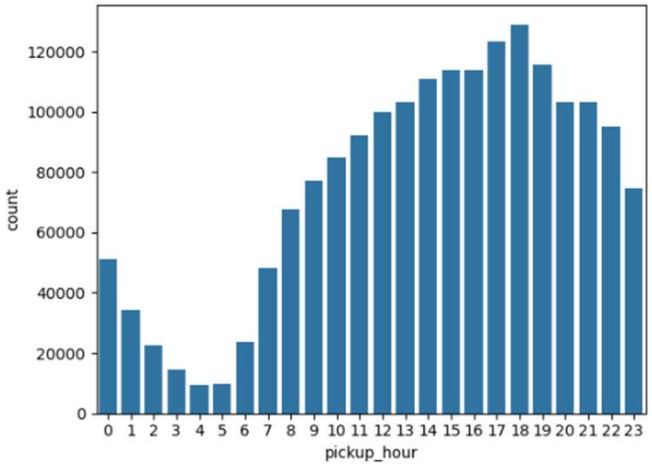
Outlier Detection

- Trip distance = 0, however with different pick-up and drop-off locations
- Fare Amount = 0, however with different pick-up and drop-off locations
- Trip distance is less than 1, however Fare Amount > 300
- Removing trip distance records > 250
- Drop off time earlier than pick up time

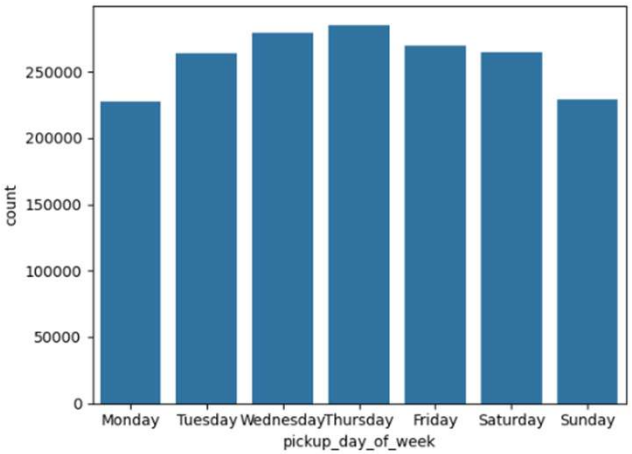
Feature Engineering

- Created columns as and when required for further processing
  - For Dates
  - For Trip_Speed

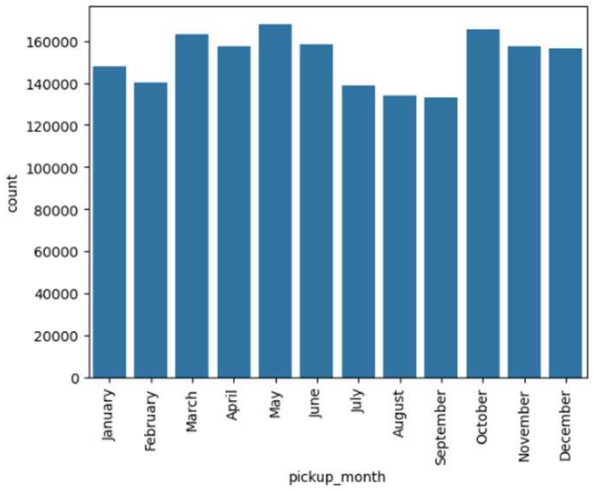
Trip Count Analysis



Trip Count Trend across the day. Busy hours are between 2 pm – 7pm.

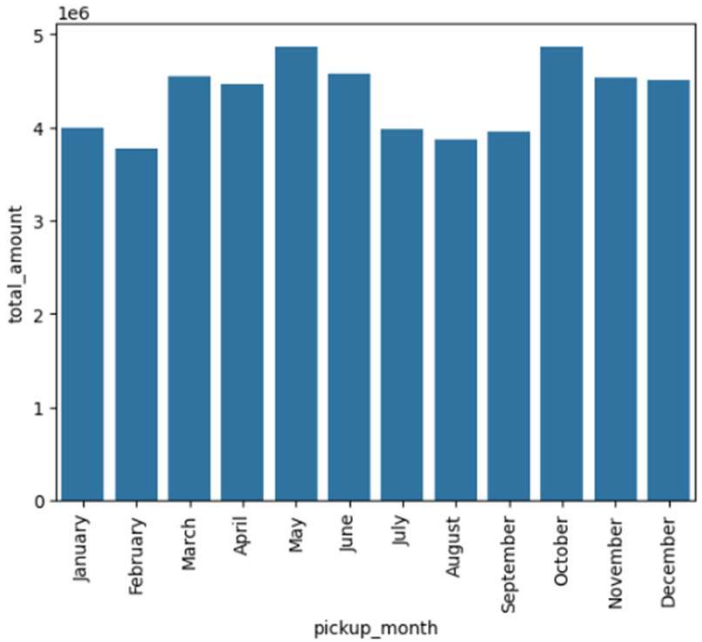


Trip Count Trend across the week. Thursday appears to be the busiest

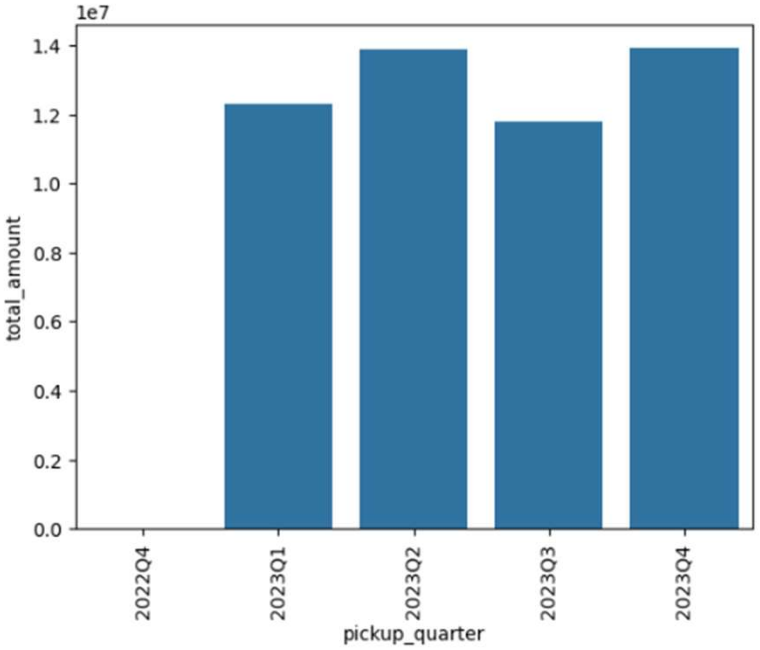


Trip Count Trend across the year. May and October are the busy months.

Revenue Analysis

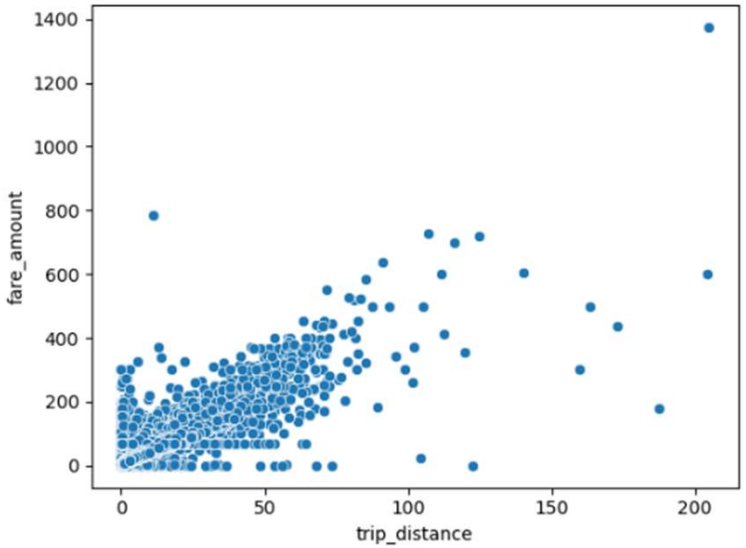


Taxi Revenue across the year by month. Summer holiday months show lower revenue. The number of visitors are generally higher in these months, and these months are not getting monetized.

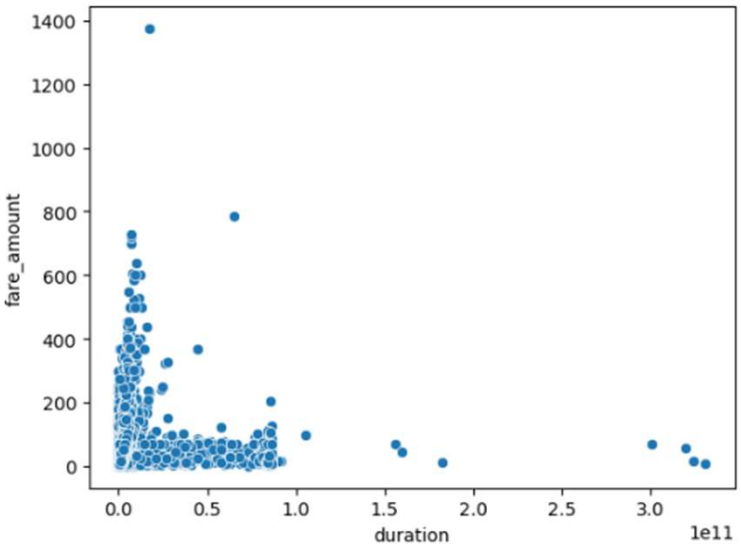


Taxi revenue by the quarter.

Trend Analysis

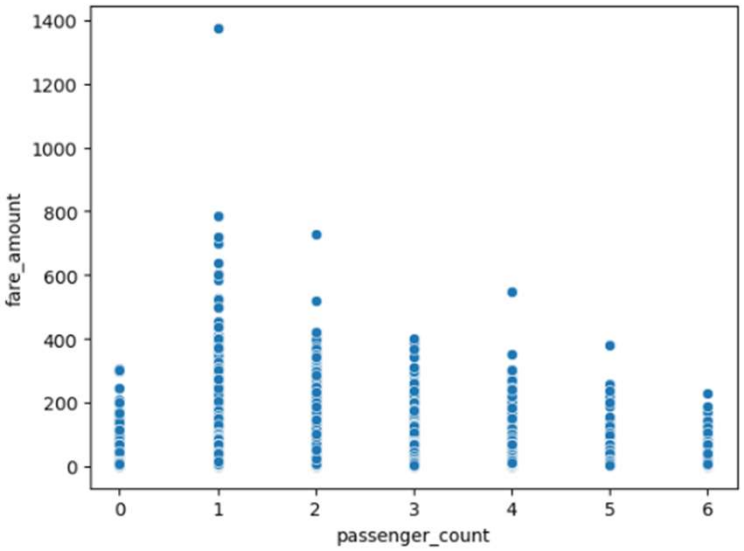


Fare amount increases relatively to the trip distance

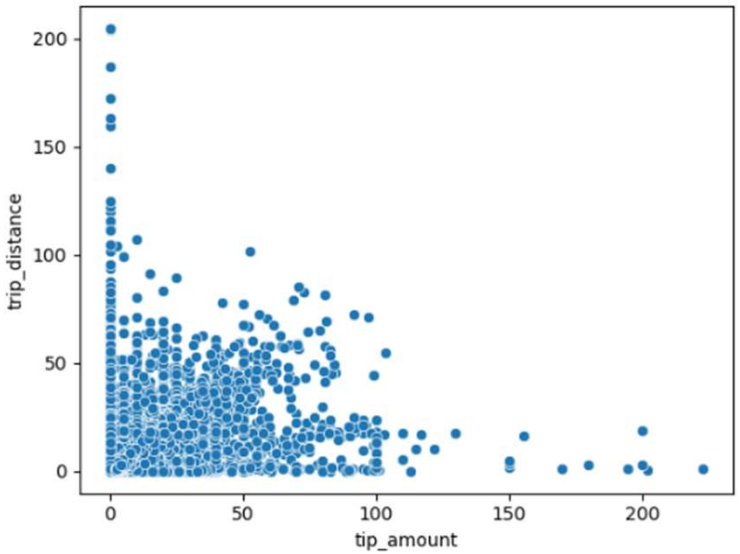


A relationship cannot be established between the trip duration and the fare amount

Trend Analysis



1 passenger trips appear to be good as rides as they end up in higher fares



Tip amount and the trip distance do not have a well-defined relationship. Shorter rides are also getting good tips

Key Insights

- Trip counts are highest in March, May and October
- Trip counts were highest in the evening between 2 pm – 7pm
- The revenue collection supports the trip count number
- The nighttime revenue stands at 11% of the overall revenue and this aligns with the proportion of night trips to day trips

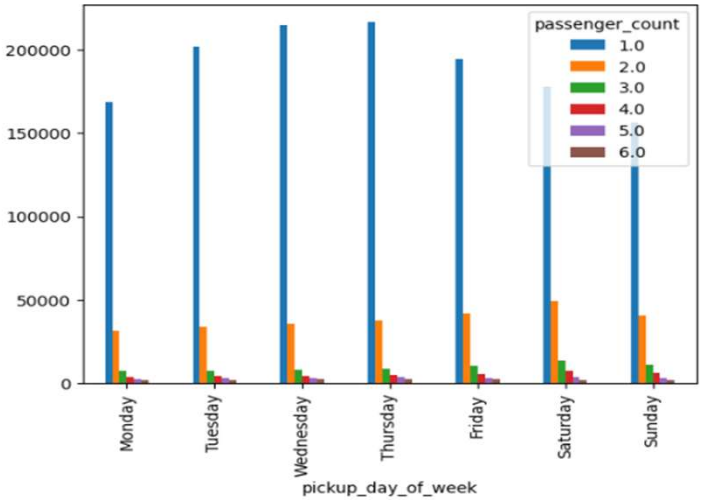
- The average speed for a trip were the slowest during 2 pm – 7pm, indicating the trips were during high congestion period

- High number of pick-ups at the JFK Airport
- High number of drop-offs at the Upper side of NY
- Most trip durations are in the range of 12 mins – 30 mins

- Weekend trips show an uptick in the 2 passenger count rides indicating family trips

- The summer months of Jul, Aug, Sep show a drop in revenue when it is expected there will be a higher visitor turnout

Additional Visualizations



Trips tiered by passenger count

Conclusion

The nighttime revenue stands at 11% of the overall revenue and this aligns with the proportion of night trips to day trips

- Changing the number of taxis availability of day vs night need not change

Next Steps / Recommendations

Using the pick-up/drop-off ratio, it can be concluded that

- For a high pick-up/drop-off ratio: Taxis are sitting idle till the next pick-up or have to travel to the next location to get a pick-up
- For a low pick-up/drop-off ratio: There is a shortage of taxis to serve that location and a lost business opportunity

The conclusions were based on a sample data set of 5% per hour per day per month. To validate the conclusions, I would run the notebook with 10% or 15% of the data or even a different 5% sample data.

An analysis of the routes served by respective vendors can be done to check if either of the vendors can improve their revenues either by serving the routes not served or increasing the number of taxis to serve the high-demand routes.