2/21/2023

Exploratory Data Analysis

[Green taxi trips NYC]

Tarun Susanth Sripathi
UNIVERSITY OF BREMEN

Contents

1. Introduction:

The Green Taxi NYC Trip dataset is a publicly available dataset that contains records of trips taken in New York City by green taxis. The dataset provides a wealth of information about taxi trips, including pickup and drop-off times and locations, fare amounts, trip distances, and other relevant information.

It is a valuable resource for data analysis and exploration. It allows analysts to explore patterns and trends in taxi usage, as well as identify areas of high demand for taxi services. The data can also be used to evaluate the impact of policy changes and other interventions aimed at improving taxi services.

The dataset includes data from 2022 January, providing a comprehensive view of taxi usage in New York City over a period of January month.

2. Data:

The 2022_1 NYC Green Taxi trip data set provides information about taxi trips taken in New York City from January 2022. The data set includes over 62495 records of green taxi trips taken in the city.

Dataset: https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-01.parquet

The data set includes the following columns:

- VendorID: The ID of the vendor that operated the taxi. (1=Creative Mobile Technologies, LLC; 2=VeriFone Inc.)
- lpep_pickup_datetime: The date and time the meter was engaged for the trip.
- lpep_dropoff_datetime: The date and time the meter was disengaged for the trip.
- store_and_fwd_flag: Indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server. (Y=store and forward; N=not a store and forward trip)
- RateCodeID: The final rate code in effect at the end of the trip.
- Pickup_longitude: The longitude where the meter was engaged.
- Pickup_latitude: The latitude where the meter was engaged.
- Dropoff_longitude: The longitude where the meter was disengaged.
- Dropoff_latitude: The latitude where the meter was disengaged.
- Passenger_count: The number of passengers in the vehicle.
- Trip_distance: The distance in miles of the trip.
- Fare_amount: The time-and-distance fare calculated by the meter.
- Extra: Miscellaneous extras and surcharges.
- MTA_tax: 0.50 USD MTA tax that is automatically triggered based on metered fare.
- Tip_amount: This field is automatically populated for credit card tips. Cash tips are not included.
- Tolls_amount: Total amount of all tolls paid in trip.
- Ehail_fee: The fee charged by the Taxi and Limousine Commission to drivers who accept a dispatch through the e-hail app.
- Total_amount: The total amount charged to passengers. Does not include cash tips.

The data set is a valuable resource for exploring and analysing taxi usage patterns in NYC. The large size of the data set makes it possible to perform detailed analysis on various aspects of taxi usage, such as pickup and drop-off locations, passenger counts, trip distances, and fare amounts.

3. Data Pre-processing:

Remove columns: Remove columns that are not relevant to the analysis or has a high percentage of missing values.

- Removed the column ‘ehail_fee’ as it consists of nothing.

Filter data: Remove records with extreme values or outliers that may affect the analysis. For example, remove records with trip distance of zero or more than 200 miles.

- Cleaned some outliers from the data like distance of 224481.330 miles in 30 mins, which is practically impossible.

- As there are many outliers, to infer some relations I have filtered the data with following conditions:
  (a) trip_distance between 1 and 50.
  (b) Fare_amount between 2 and 300.
  (c) Trip_duration_min between 1 and 200.
- Removed all the records with tips less than zero which might not be possible.

Handle missing values: Check for missing values and handle them appropriately. For example, remove records with missing values, impute missing values with the median value of the column, or use machine learning algorithms to predict missing values.

- In the process of cleaning, dropped all the null values.
- The following are the percentage of missing values in respective columns:

  (a) ‘store_and_fwd_flag’ has 0.1007 % missing values
  (b) ‘RatecodeID’ has 0.1007 % missing values
  (c) ‘passenger_count’ has 0.1007 % missing values
  (d) ‘payment_type’ has 0.1007 % missing values
  (e) ‘trip_type has’ 0.1007 % missing values
  (f) ‘congestion_surcharge’ has 0.1007 % missing values.

Create new features: Create new features that can provide more insights about the data. For example, calculate the average speed of the trip, or the time of the day when most trips are taken.

- Created a new feature called ‘trip_duration’ which represents the total duration of trip by subtracting ‘lpep_dropoff_datetime’ from ‘lpep_pickup_datetime’.
- As the trip duration is in date time object, its difficult to plot or do any math operations on it. I have converted it the time in minutes and assigned it to new variable ‘trip_duration_min’.

- Extracted the hour of the day, weekday and week day number from ‘lpep_dropoff_datetime’, ‘lpep_pickup_datetime’ and assigned to new variable in data as ‘pickup_day’, ‘dropoff_day’, ‘pickup_day_no’, ‘dropoff_day_no’, ‘pickup_hour’,’ ‘dropoff_hour’.
- Extracted and created another feature time of the day from lpep_dropoff_datetime’, ‘lpep_pickup_datetime’. Which gives which part the day the pickup and drop-off happened. Like Morning, Afternoon, Evening and Late nights.
- Created the feature called avg-speed from the features trip_duration_min and trip distance.

By performing these data cleaning steps, the data set will be ready for exploratory data analysis and will provide more accurate and meaningful insights.

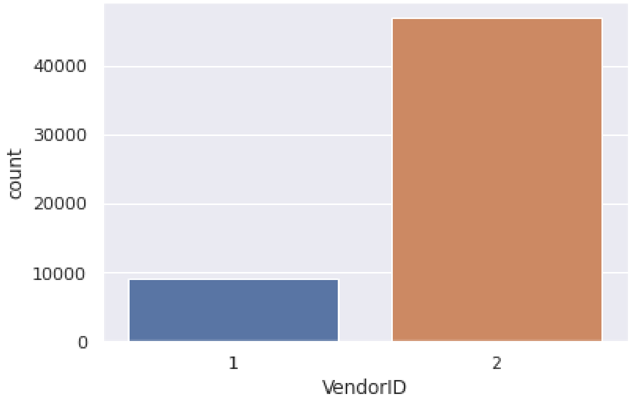
4. Data Exploration:

- There are 2 unique vendor ids.
- There are 9 unique passenger counts
- There are 2 unique values for store_and_fwd_flag, that we also saw in the description of the variables, which are Y and N.
- There are 2 trip types.
- There are 5 payment types.
- There are 5 rate codes.

a. Vendor_ids:

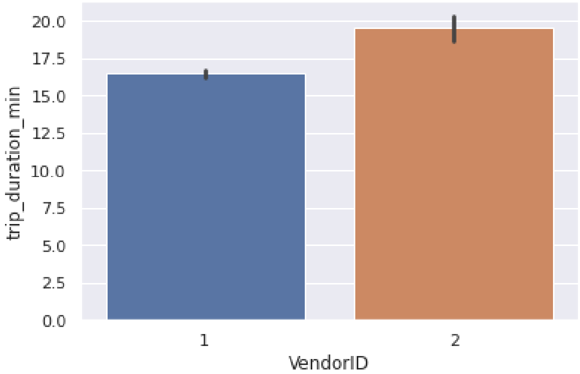
From the data we can infer that there are two vendor's and represented by their id's as 1 and 2.

Let us see which vendor has more trips:

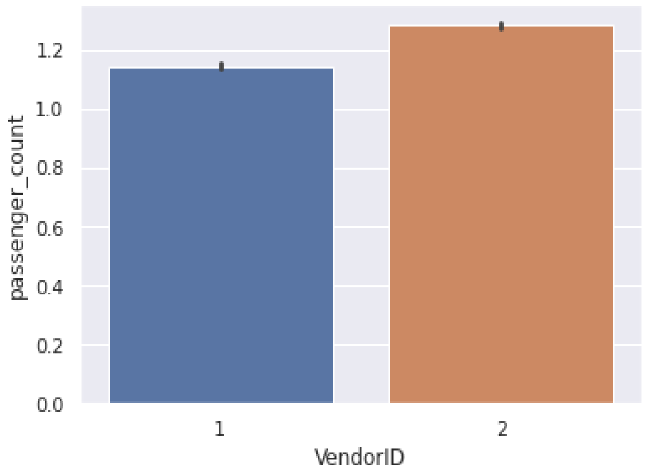


From the above plot we can infer that vendor has more trips.

- ❖ Vendor 2 has 46916 trips and vendor 1 has 9210 trips.
- ❖ May be Vendor 2 has more cars compared to 1. As, vendor 1 has too less trips.



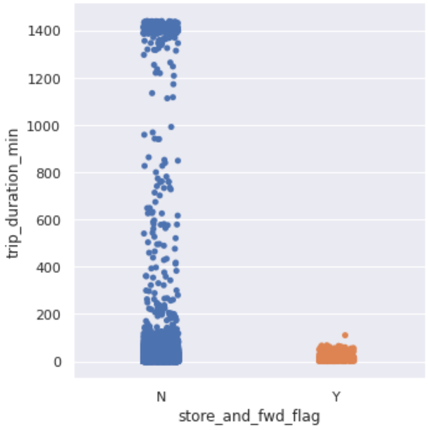
- ❖ Vendor 2 has more trip duration about 20 min of mean and vendor 1 has 16.25 min on an average.



- ❖ Vendor 2 and 1 has similar number of passengers count per trip which is 1.

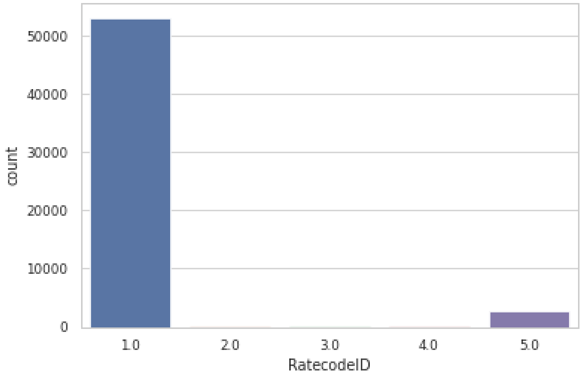
b. store_and_fwd_flag:

- ❖ store_and_fwd_flag has 99% N and 1% Y. which means only 1% of data is stored and forwarded in the vehicle memory as it does not have connection with server.



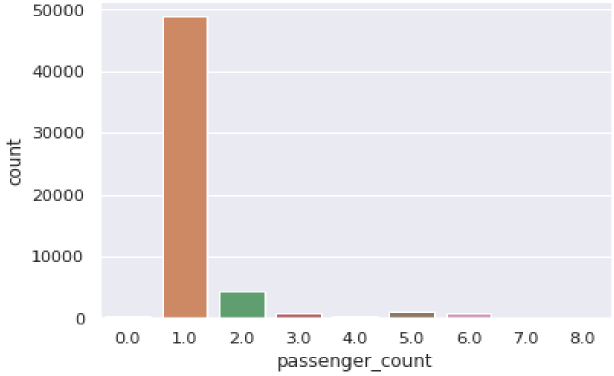
- ❖ Trip Duration is generally longer for trips whose flag wasn't stored.

c. RatecodeID :



The majority of people have travelled with standard price and very less percentage people travel while negotiating.

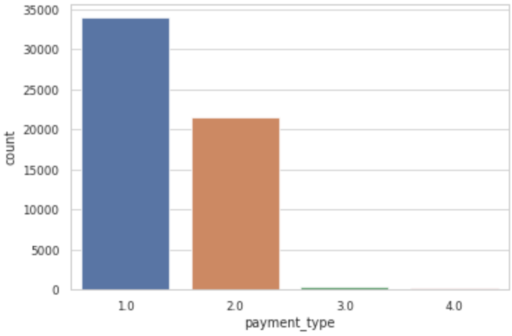
d. Passenger count:



- There are more single passengers. Maybe they are traveling back and forth from work.
- So, The vendors can implement more small cars for better service in more areas and less carbon foot print.
- There are even 0, 7 and 8 as well. So, the zero people may be the error of driver or noted deliberately to complete a greater number of rides.
- So, deleted those values which are expected as outlier and some fake data.
- Here are the values of trips with passenger counts.

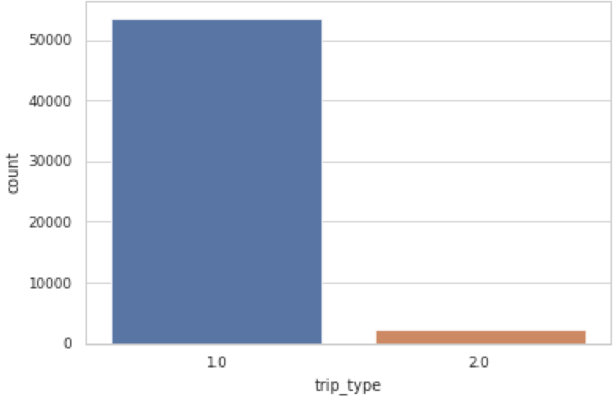
1	48845
2	4240
3	726
4	207
5	1084
6	805

e. Payment_type:



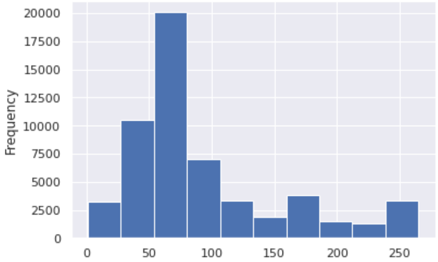
The majority people have preferred to pay with credit card and followed by cash.

f. Trip type:



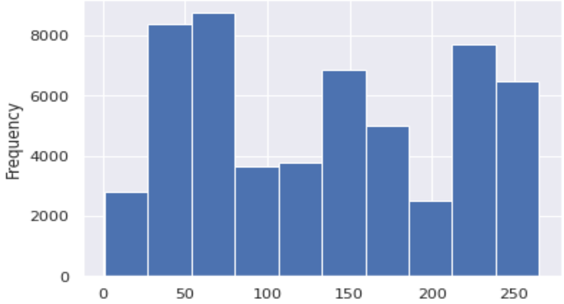
The majority are Street hails than actual dispatch.

g. PULocationID:



The major pick-up zones are 50 t0 75. May be we can arrange a taxi stations here.

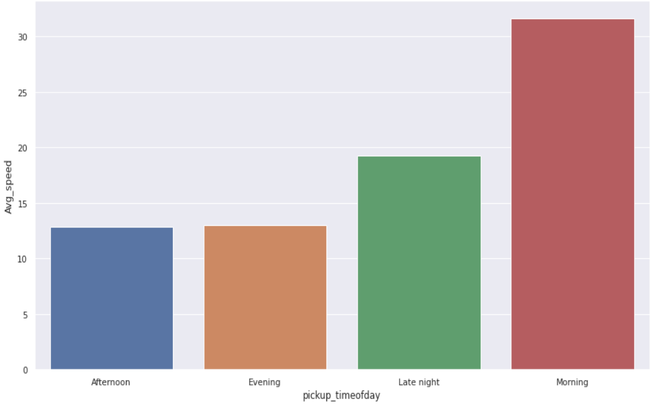
h. DOLocationID:



- The major drop off zones are 25 to 75 and followed by 200 to 250.
- More taxis should be place in the zones between 50 and 250 by the vendors.
- Even local government should run more services in these zones.

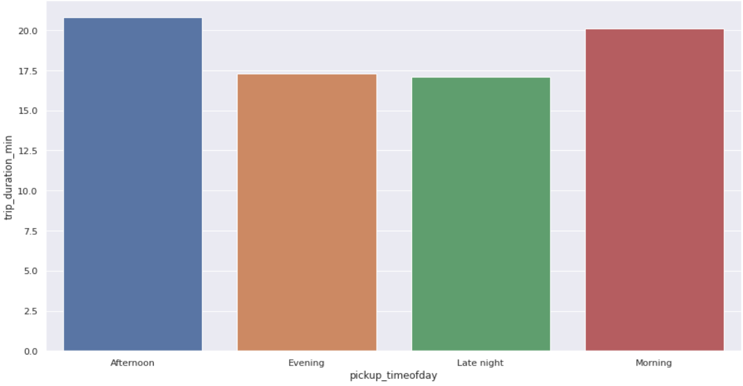
i. Average Speed:

- The mean average speed is 14.55 miles per hour.

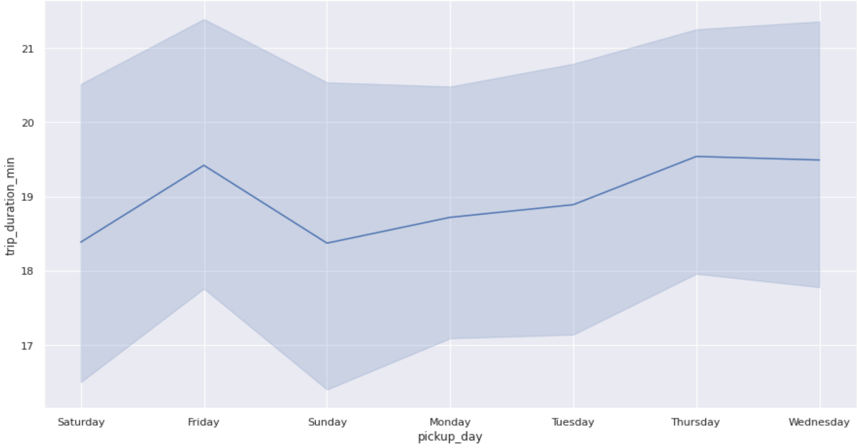


- The average speed is more in mornings.

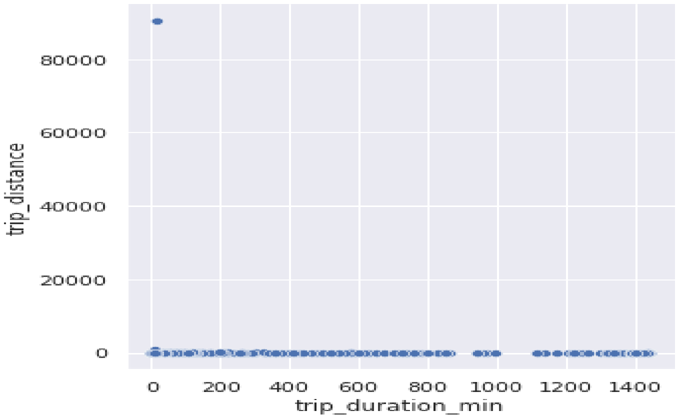
j. Trip_duration_min:



- Most of the long duration trips are in day time like morning and afternoons combined. As, the roads are busy while people are commuting to work.

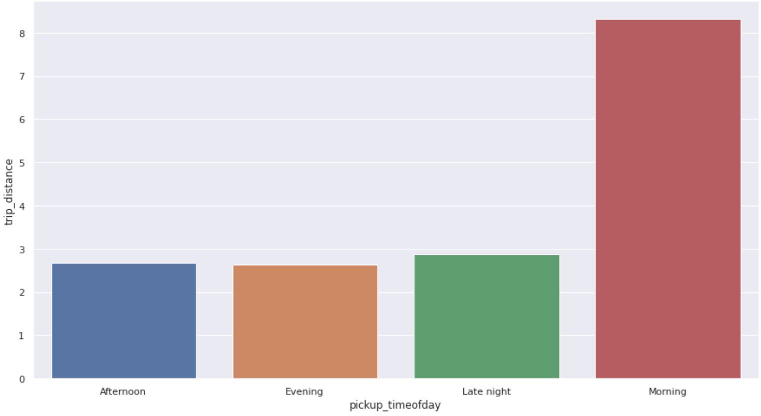


- Trip duration longest on Fridays closely followed by Thursday and Wednesday.

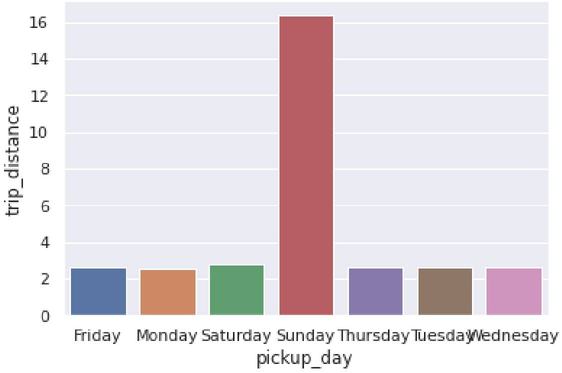


- From the plot above, though distance is recorded ass 0 but trip duration is more. There are 3274 records.
- One reason can be that drop off coordinates were not recorded.
- Another reason one can think is that for short trip durations, maybe the passenger changed their mind and cancelled the ride after some time.

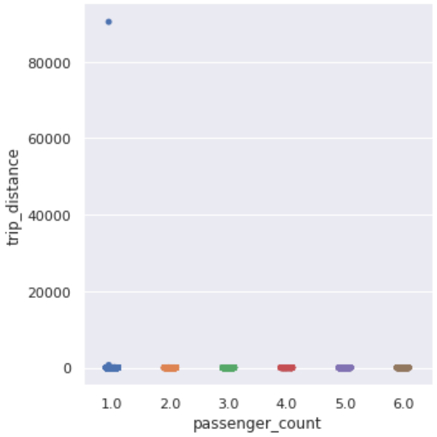
k. Trip_distance:



- Most distance are covered in Mornings.

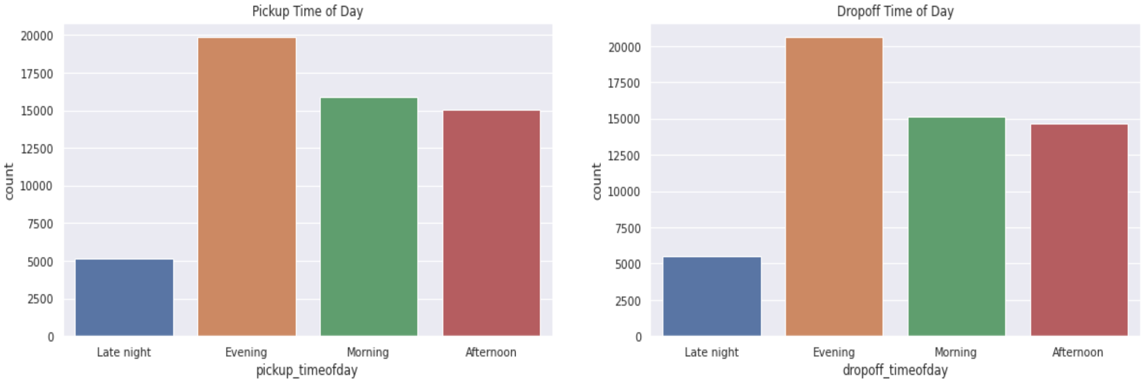


- The long-distance trips are on Sundays compared to any other week, Due to weekend.

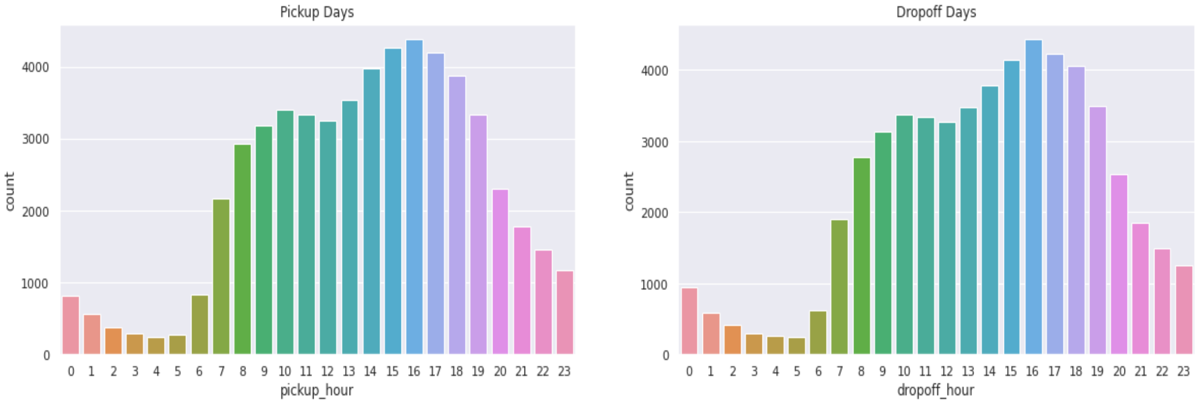


- Some longer Distance are covered by 1 passenger trips.
- We can see there are 2893 trips with 0 km distance.
- The reasons for 0 km distance can be:
    - The drop off location could not be tracked.
    - The driver deliberately took this ride to complete a target ride number.
    - The passengers cancelled the trip.
    - We will analyse these trips further in bivariate analysis.

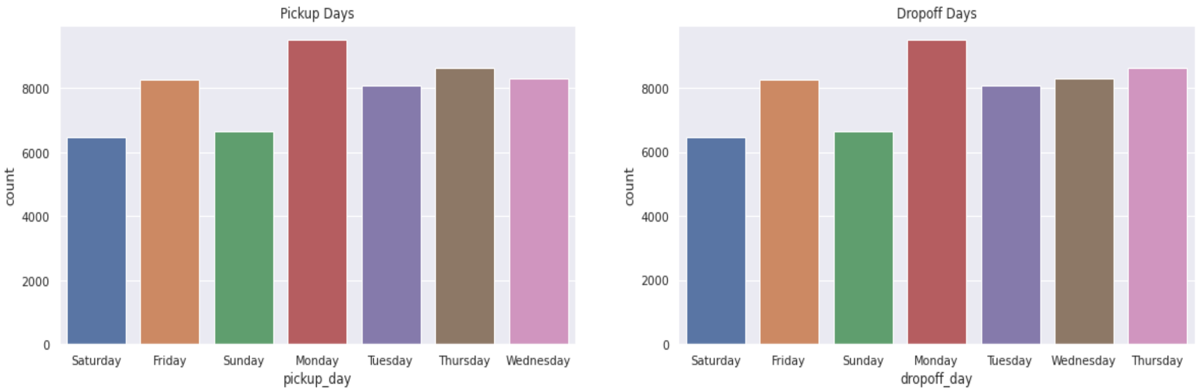
I. Pickup and drop-off time of the day:



- As we can infer from above evenings are busiest.

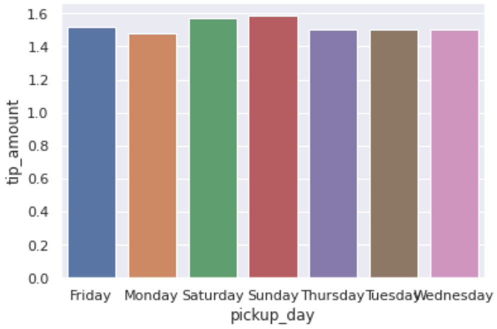


- We see busiest hours are from 8 am to 7 pm and that makes sense as this is the where people travel to and back from the work.

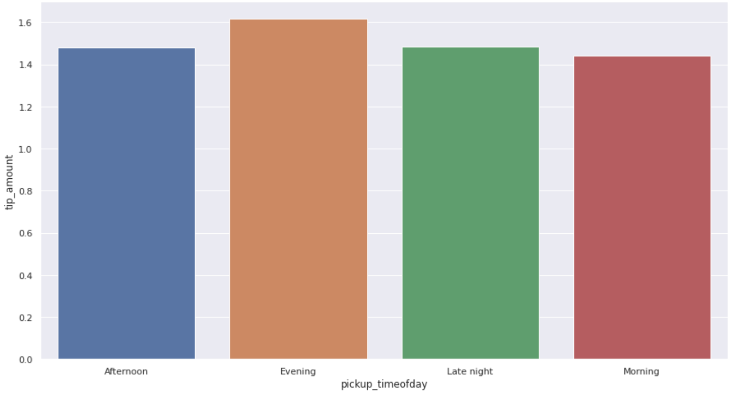


- ❖ We see Monday to Friday are the busiest days than the weekends as the reason of more people commuting to their work more on taxis.

m. Tip_amount:

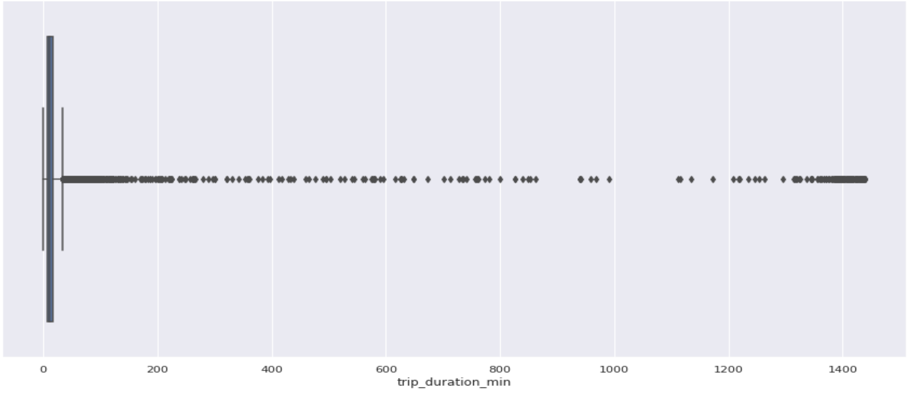


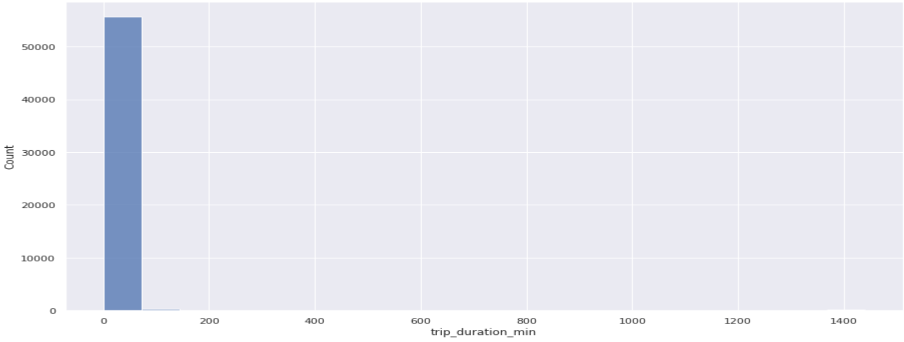
- ❖ The contribution to the tips amount is on weekends.



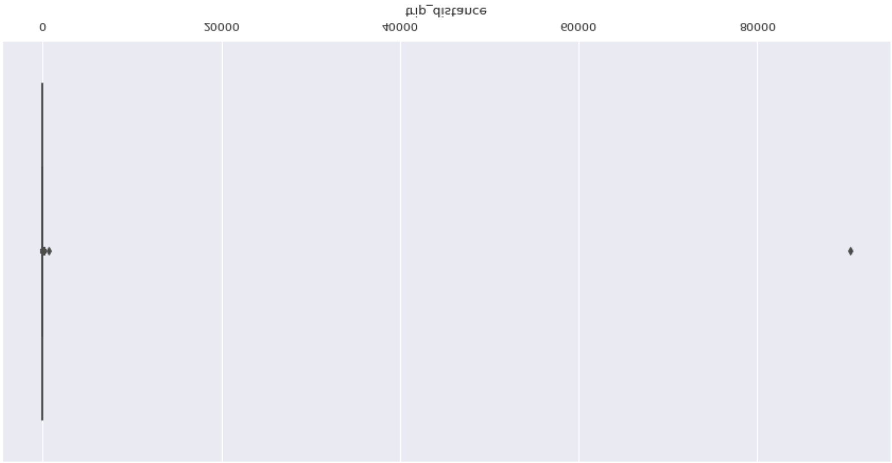
- ❖ The contributions to tips is more on Evenings, followed by rest which more or like similar.

5. Results:

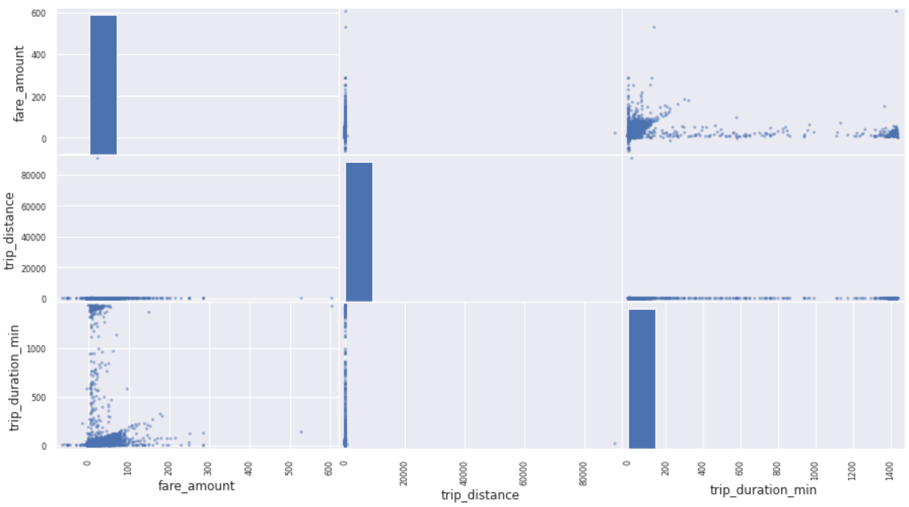




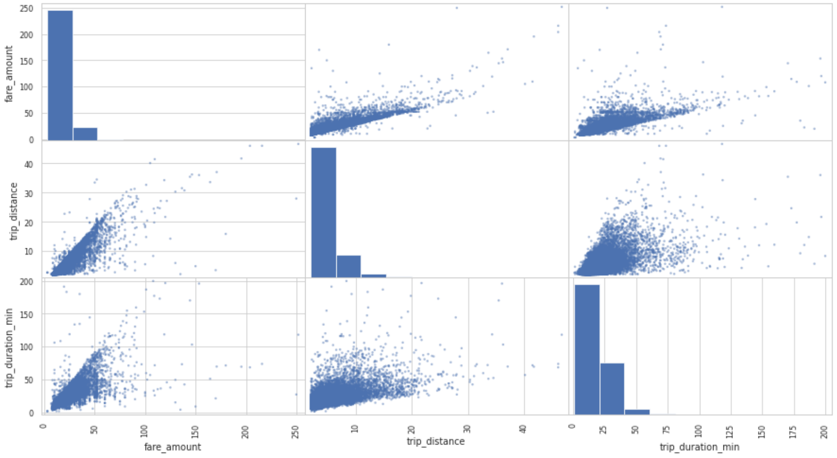
- ❖ As, we can see from the above there are many outliers around zero.



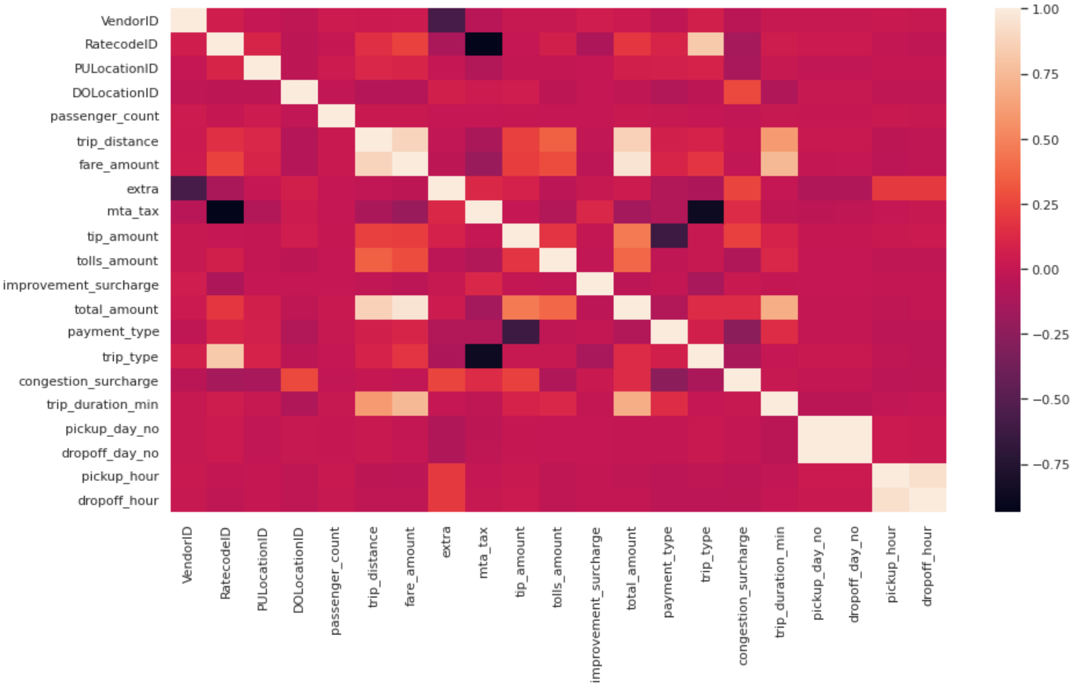
- ❖ Here as well there are many outliers on the either sides.



- ❖ The Scatter matrix comparing and finding out relation between fare_amount, trip_distance and trip_duration_min.
- ❖ It can be seen from the chart, there are significant outliers for all three columns. Lets remove outliers and plot again.



- The outlier's removal helped us to corelate. We can Clearly see that there is strong correlation between fare_amount, trip_distance, trip_duration_min.



- From the heat map we can see more corealtions like
  - RateCodeID vs trip_type
  - Trip_distance vs total_amount and more.