Trip Duration Summary

Table of Contents :

Data Inspection

- **How the data look like :**
  - The Dataset consists of 10 feathers and 1 target.
  - Let's go through each attribute briefly:

- **id** - a unique identifier for each trip
- **vendor_id** - a code indicating the provider associated with the trip record
- **pickup_datetime** - date and time when the meter was engaged
- **dropoff_datetime** - date and time when the meter was disengaged
- **passenger_count** - the number of passengers in the vehicle (driver entered value)
- **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
- **store_and_fwd_flag** - This 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
- **trip_duration** - duration of the trip in seconds
  - Note : trip_duration is our target Variable.

- **Summary Statistics :**
    - We can I describe the dataset as follows:

	count	mean	std	min	25%	50%	75%	max
vendor_id	1229319.0	1.534912	0.498780	1.000000	1.000000	2.000000	2.000000	2.000000e+00
passenger_count	1229319.0	1.664531	1.314509	0.000000	1.000000	1.000000	2.000000	8.000000e+00
pickup_longitude	1229319.0	-73.973446	0.061780	-121.933342	-73.991852	-73.981728	-73.967346	-6.133553e+01
pickup_latitude	1229319.0	40.750928	0.033545	34.359695	40.737370	40.754108	40.768360	5.188108e+01
dropoff_longitude	1229319.0	-73.973395	0.061641	-121.933304	-73.991325	-73.979759	-73.963028	-6.133553e+01
dropoff_latitude	1229319.0	40.751816	0.036341	32.181141	40.735901	40.754532	40.769821	4.392103e+01
trip_duration	1229319.0	959.691748	5263.573404	1.000000	397.000000	662.000000	1075.000000	3.526282e+06

    - We can Notice from the table the following
        - ➢ There are two vendor / taxi companies. **Is there a different speed in each one?.**
        - ➢ Having **Nine passengers Seems challenging**?.
        - ➢ The **max Trip duration** took 3.526282e+06 in sec or 58771 minutes **is approximately 40 days and 1111 minutes so it's outlier** ?

- If we focused to every individual feature, we will found the following :

➢ It looks like the passenger_count is a Categorical Variable so it means the range of people who can't taxi travelled is between [1,8].

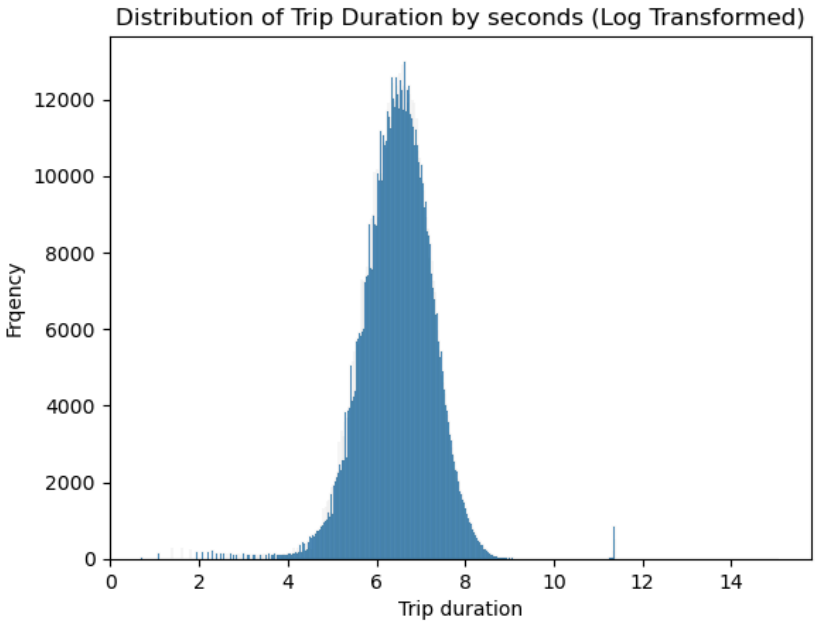
➢ The min number of passengers is 0. It's definitely a noise (Maybe it happens because of an error in the system or the driver forgot to enter the value).

➢ The Time and date format does not help us to get information or knowledge about different time snippets or Months or even days affecting the taxi trip duration , So we need for Temporal Timedate analysis.

➢ Longitude and latitude feature needed for Geographical data analysis.

Analysis features and Feature Engineering

- **Target Variable:**
  - **First, let's explore the distribution of the trip duration.**



➢ The Target Variable Distribution looks like a right-skewed Gaussian distribution this means that there are more shorter trips than longer trips.

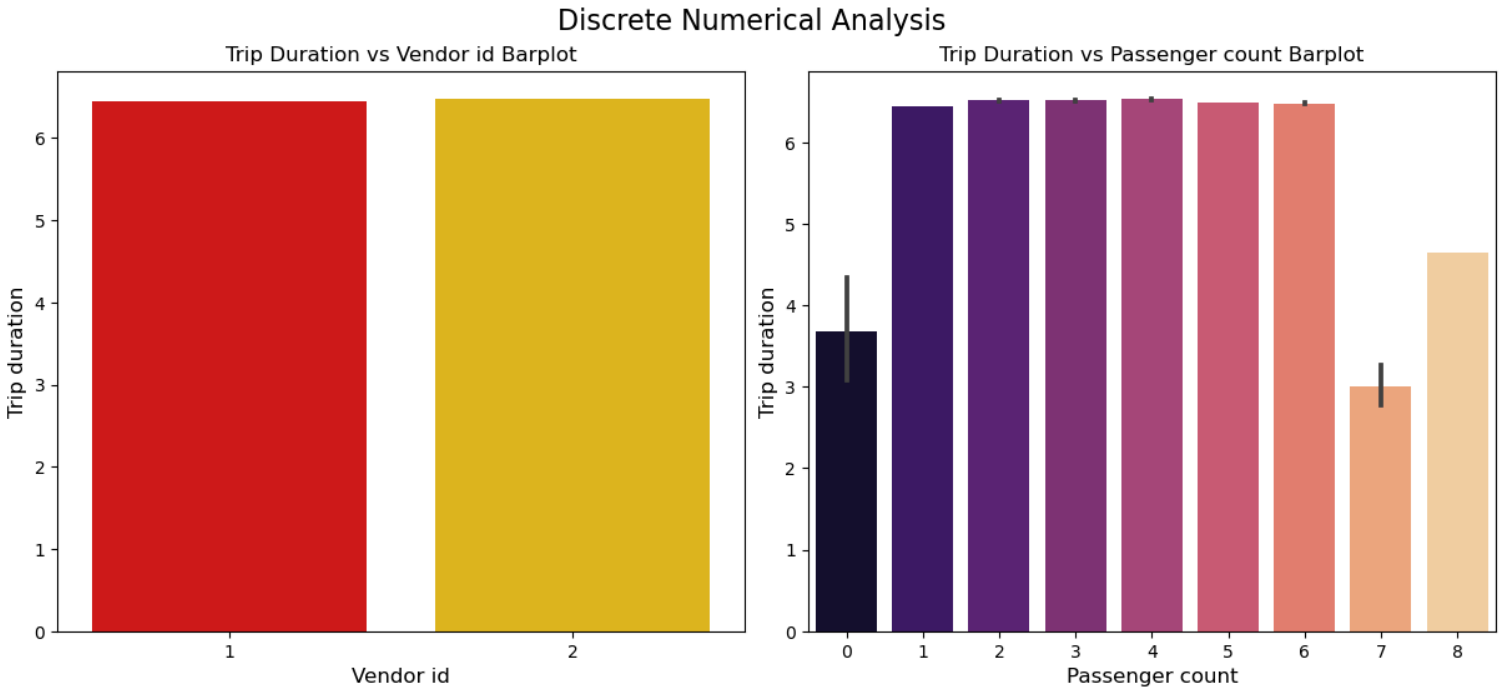
➢ There is a long tail: This means that there are a few very long trips that are outliers compared to the majority of trips.

➢ The peak of the distribution is around 5: This means that most trips are between 150 seconds and 1000 seconds (about 2.5 and 16.7 minutes) long.

★ **Note :**
- we perform the lop1p transform to Visualise better.
- we use np.expm1 to convert to seconds.

➢ The max Trip duration took around 58771 minutes is approximately 40 days and 1111 minutes so definitely outlier.

- **Discrete Numerical Feature:**
  - Vendor id and passenger count is our Discrete Numerical Feature.



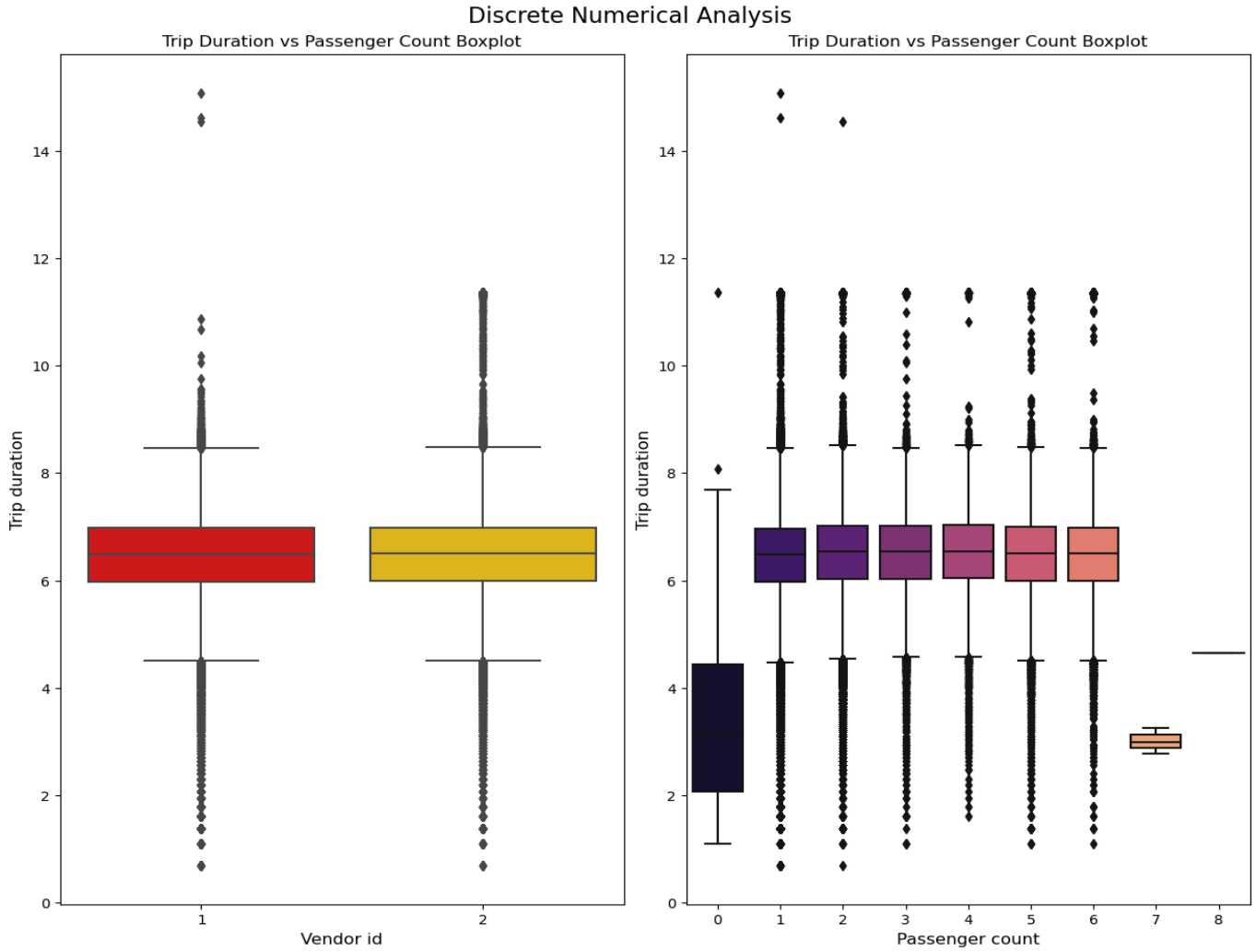
➢ Trip duration and Vendor ID: It's difficult to discern any clear patterns or trends between trip duration and vendor ID from the bar chart. The bars appear to be spread out relatively evenly across the x-axis, suggesting no significant difference in trip duration among vendors.

➢ When the number of passenger groups from [1 to 6] take constant trip duration and the number of passenger groups from [7 to 8] take less trip duration, Possible explanations:

- Vehicle Capacity : It's possible that the vehicles used by both vendors have a maximum capacity of 6 passengers. When there are 6 or fewer passengers, the vehicles are operating at their maximum capacity, and the trip duration remains constant because the vehicles are fully utilized.

- Vehicle Type : It's also possible that the two vehicle vendors have different types of vehicles in their fleet. One vendor might have larger vehicles capable of accommodating more passengers, while the other vendor might have smaller vehicles. The larger vehicles can comfortably accommodate 7 to 8 passengers, resulting in shorter trip durations.
- Trip purpose: Trips with 7-8 passengers might be more likely to be for specific purposes (e.g., airport shuttle, group sightseeing) that involve shorter distances or routes optimised for efficiency.

➢ If **vendor of the taxi not effect with trip duration so idea of vehicle type and vehicle capacity not correct** so we need to use a boxplot to detect If we just dealing with some random noise or passengers groups from [7 to 8] just travel less than another groups.

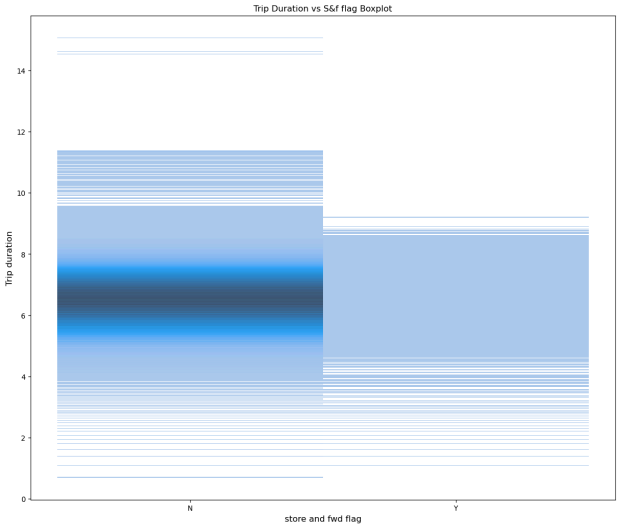


- **Trip duration distribution:** The distribution of trip durations appears to be positively skewed for both groups, meaning there are more shorter trips than longer trips. This is especially evident for the group with lower passenger counts (left boxplot).
- **Median trip duration:** The median trip duration, represented by the horizontal line within the box, is longer for the group with higher passenger counts (right boxplot). This suggests that trips with more passengers tend to take longer overall.
- **Outliers:** There are a few outliers for both groups, represented by the circles beyond the whiskers. These are individual trips that were much longer than the majority of trips in their respective groups and this **Support conclusion passengers groups from [7 to 8] just travel less than another groups because Trip purpose.**

- ## Categorical variable:

➢Most taxi trips are sent to the vendor in real-time ("N").

➢The **most trips sent to the vendor in real-time ("N") likey take more Trip duration.**



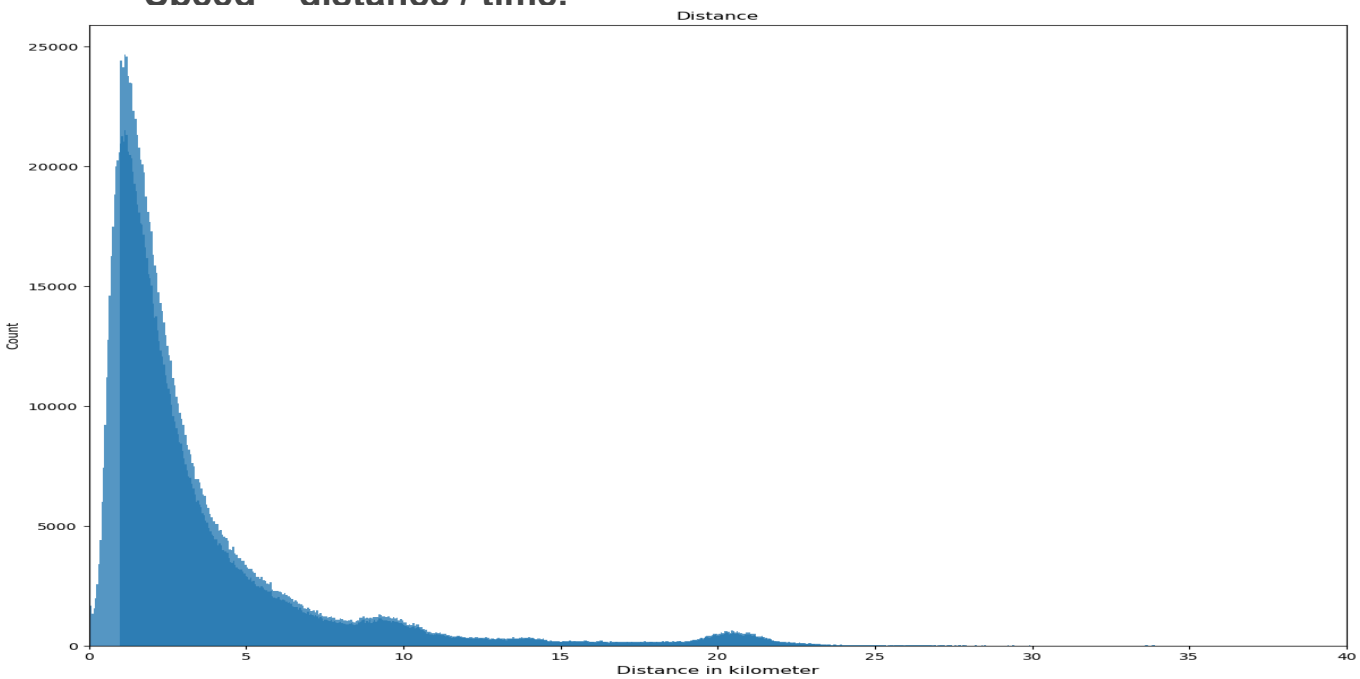
- ## Geographical Data:

➢Now let's analyse Latitude and longitude as geographical coordinates.

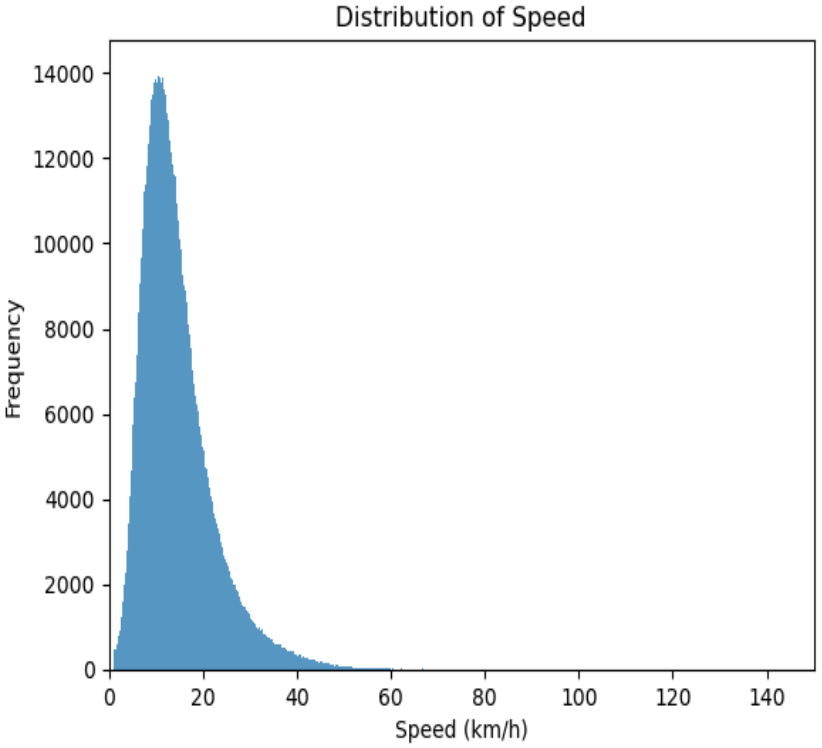
➢We can use **haversine distance** to get more information from geographical coordinates

➢ Looks like most of the trip goes from less then 1 kilometre to 25 kilometres.

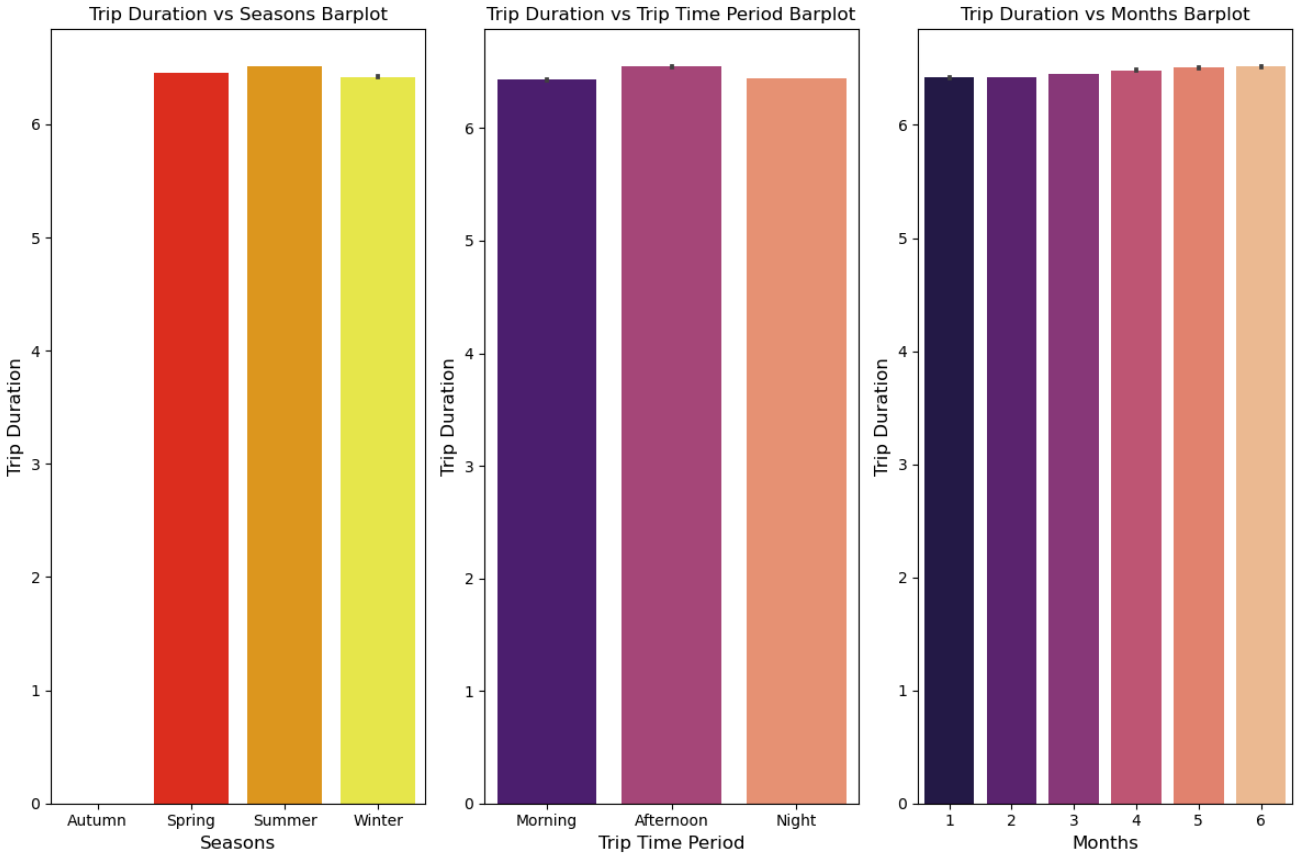
➢ We **can use trip duration as a time feature and calculate Speed = distance / time.**



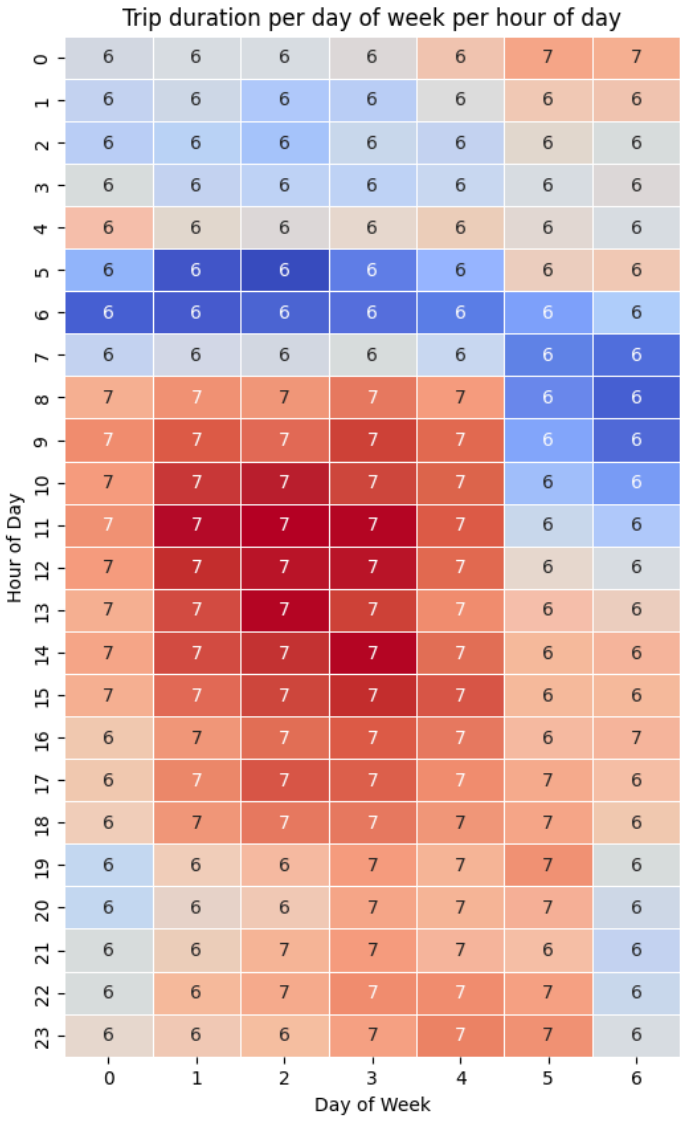
➢ Most of the trip goes at a speed in the range 1-40 Km/h.



- **Temporal/Time-date Analysis :**
  - Using **data/time we can get new information like Months/ day/ Morning or afternoon or night / season** for each trip and now we can **try to find how it affects trip duration.**



- ➢ The **longer trip durations during summer might be attributed to vacations and holidays, which lead to increased traffic on the roads.**
- ➢ **Longer trip durations in the afternoon can be explained by higher levels of crowding during** that time of day.
- ➢ **April, May, and July experience longer trip durations compared to other months.**



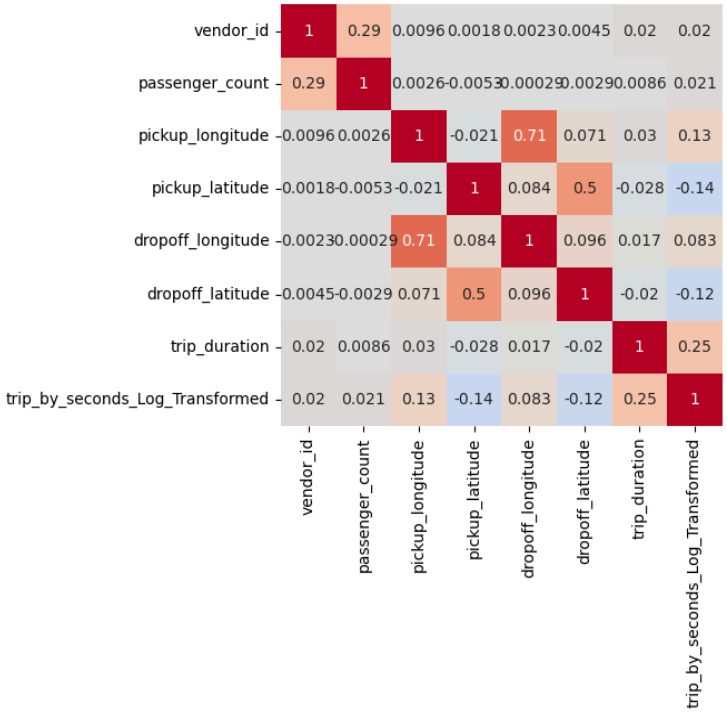
➢ **Trip durations are generally longer on weekends than on weekdays (Saturday and Sunday).** This is likely due to the fact that people are more likely to take longer trips for leisure purposes on weekends.

➢ Trip durations **are generally shorter during the morning and evening rush hours.** This is likely due to the fact that people are more likely to be making shorter trips for commuting purposes during these times.

➢ Trip durations are **generally longer in the middle of the day**. This is likely due to the fact that people are more **likely to be making longer trips for shopping, errands, or other activities during these times.**

Correlation Analysis

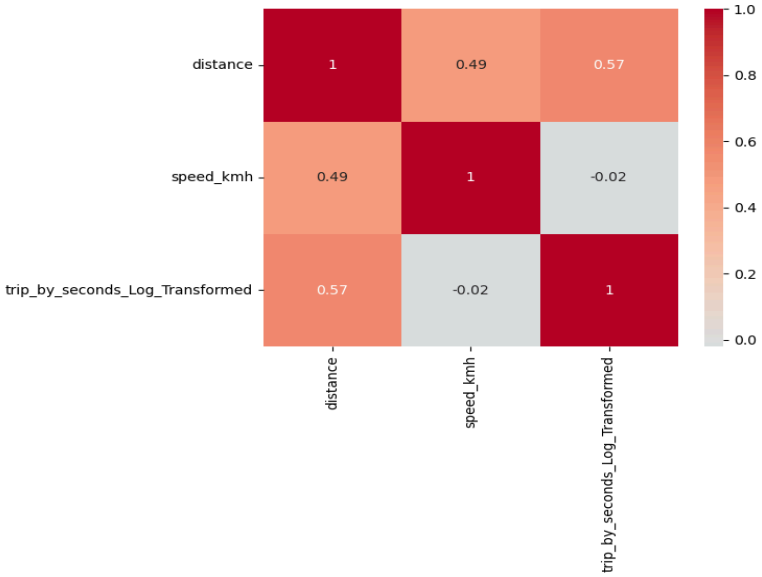
- First lets see correlation table for Original data:

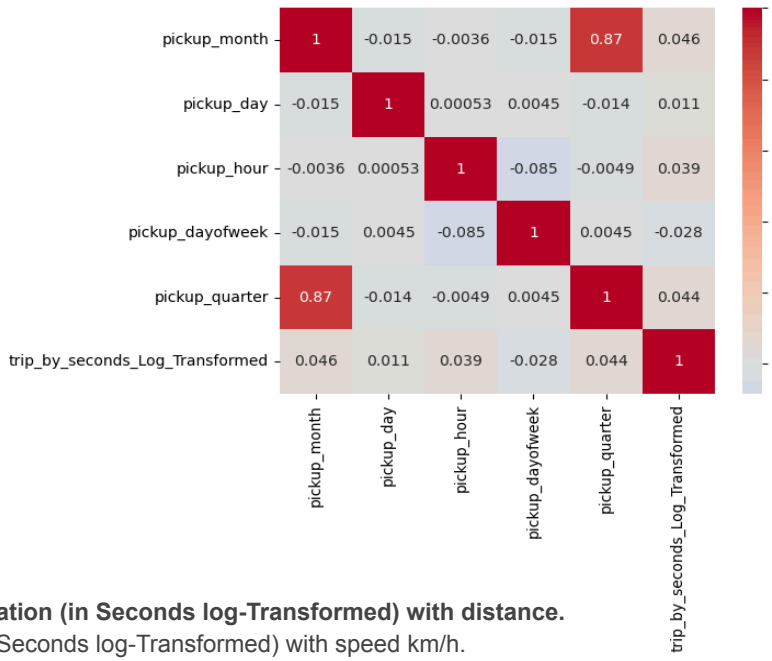


➢ Their positive relation trip duration (in Seconds log-Transformed) with pickup longitude ,dropoff longitude and passenger count.

➢ There negative relation trip duration (in Seconds log-Transformed) with pickup latitude and dropoff latitude.

- Second, let's see the correction of our new feature:





➢ Their **strong positive relation trip duration (in Seconds log-Transformed) with distance.**
➢ Their negative relation trip duration (in Seconds log-Transformed) with speed km/h.
➢ **Conclusion we have a pretty Good Relationship between the distance features and trip duration so we can use it in modelling.**

Modelling