1 Trip Duration Prediction

Introduction

The NYC Taxi Duration Prediction competition on Kaggle challenges participants to build a model that predicts the total ride duration of taxi trips in New York City. The primary dataset is provided by the NYC Taxi and Limousine Commission and includes information such as pickup time, geo-coordinates, number of passengers, and other variables

2 Exploring Data

2.1 Target Feature

note that we use natural logarithm to have better visualization ex: exp(x)-1 like seconds of (6 in graph)= e(6)-1

we use numpy.expm1

from graph it seems we have some outliers on the right ex: The max trip duration took around 58771 minutes is approximately 40 days so definitely an outlier.

most of trips around 2.2 min to 16.6 min

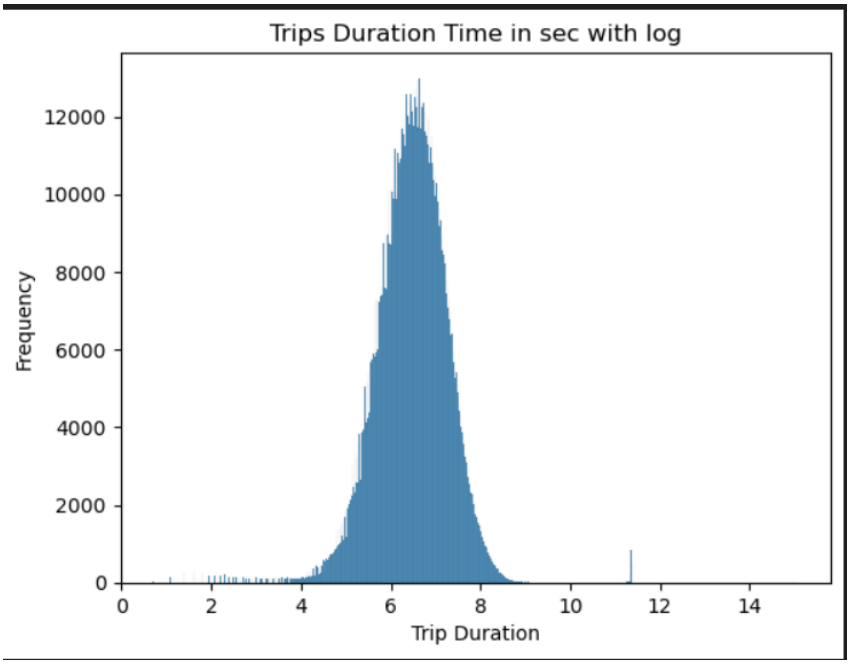

Figure 2-1 trip distribution

2.2 Numerical Features

2.2.1 Discrete

we have discrete numerical features, so they treated like categorical features

My intuition here that till now no clear patterns between trip duration and vendor id bars are relatively the same

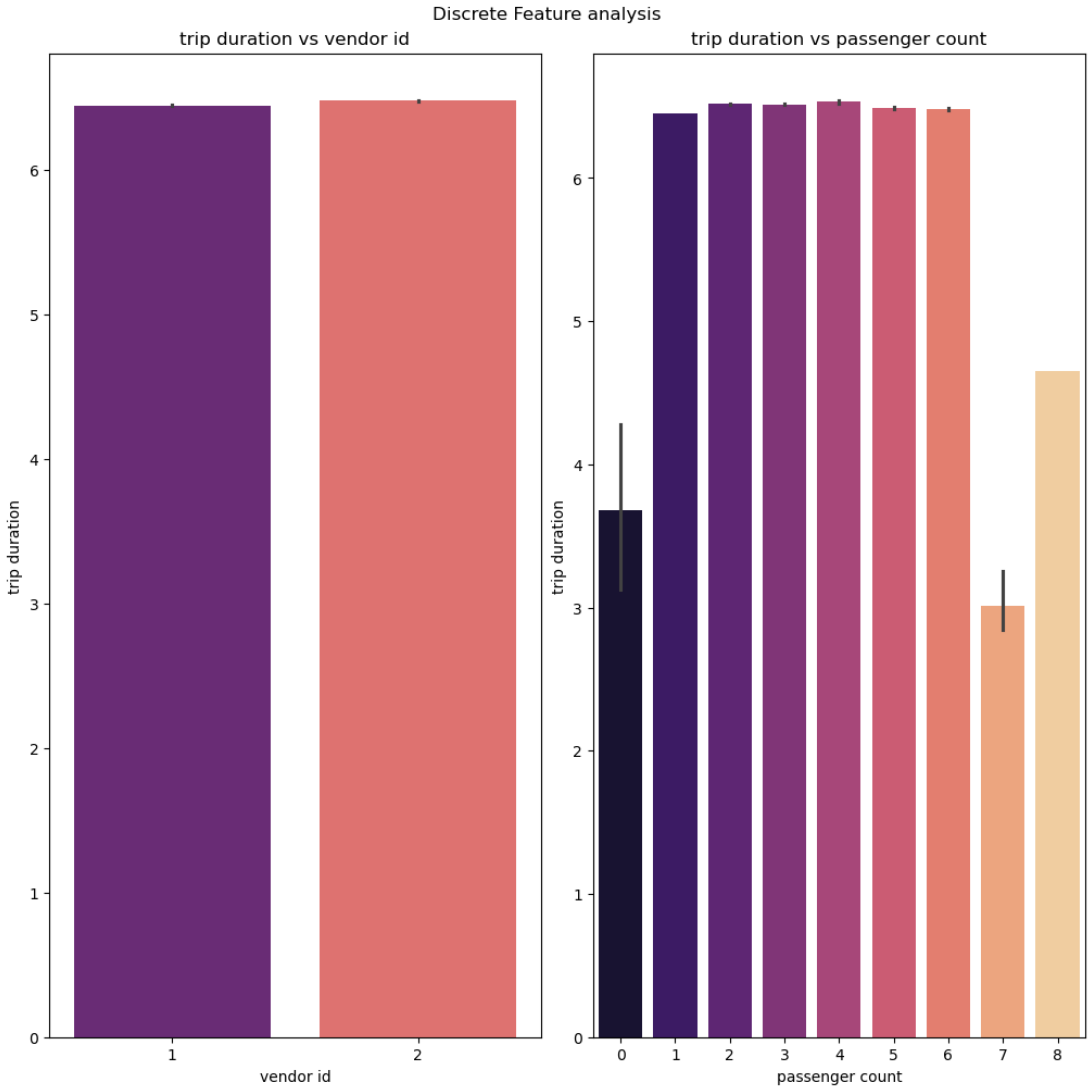

Figure 2-2 Discrete Features Analysis

from 1 to 6 it is constant trip duration from 7 to 8 take less so this might be interesting as may be the standard nums of passenger is from 1 to 6 so it is not having big effect on trip duration but may in 7 or 8 are specific trip so it might have shorter roads etc...

so, bars say for two vendors they have like same vehicles or not make a big difference in trip duration but in first i thought one vendor is faster than other as it more preferred so I might use boxplot as will have more detailed about data noise, outliers

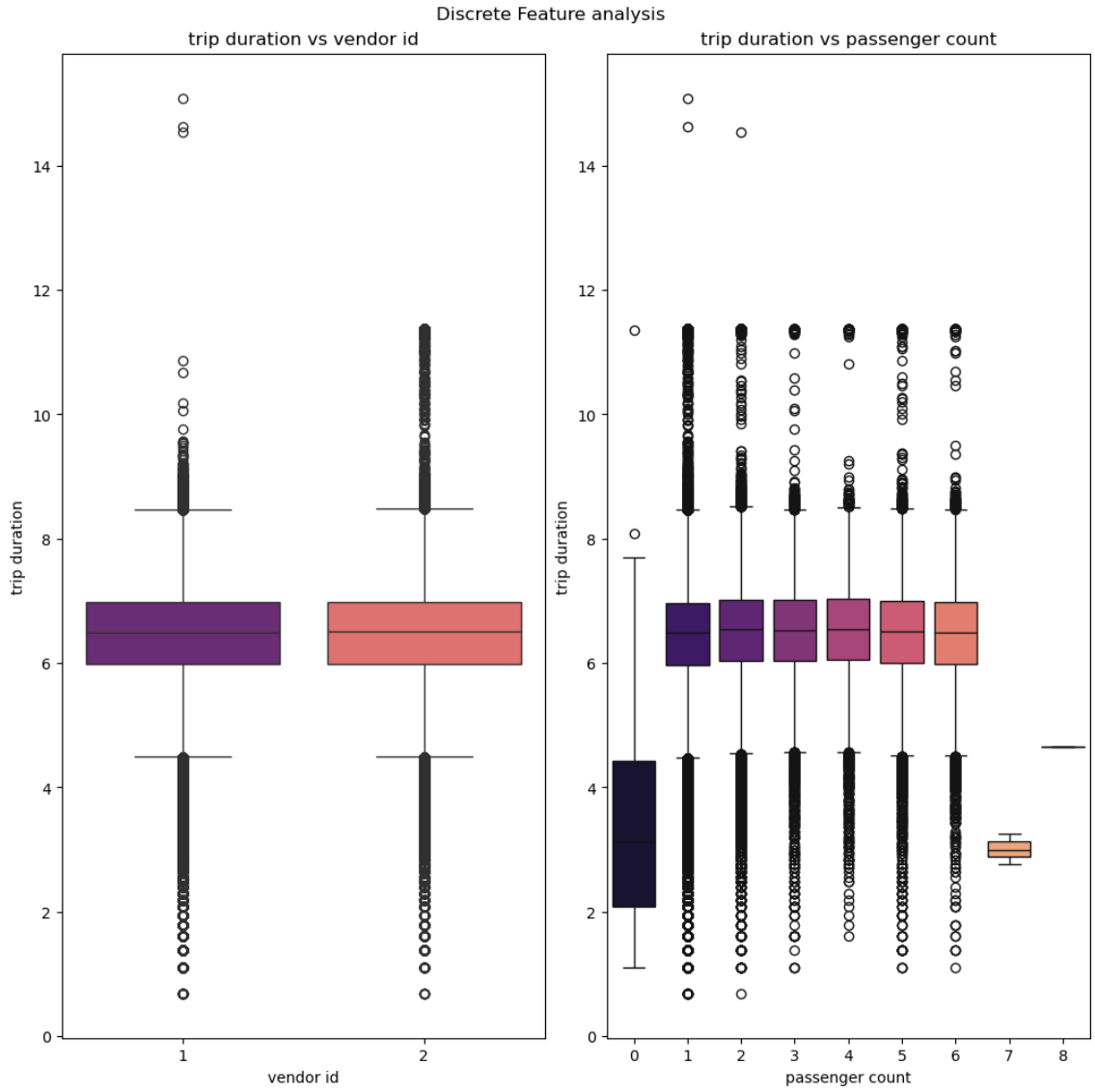

Figure 2-3

from this boxplot we have intuition that there are short trips more than long trips

with higher passenger than 6 tends to be a special short trip

we have some outliers that takes longer than their group

2.3 Categorical Features

-Most trips are sent in real time and tend to have more duration

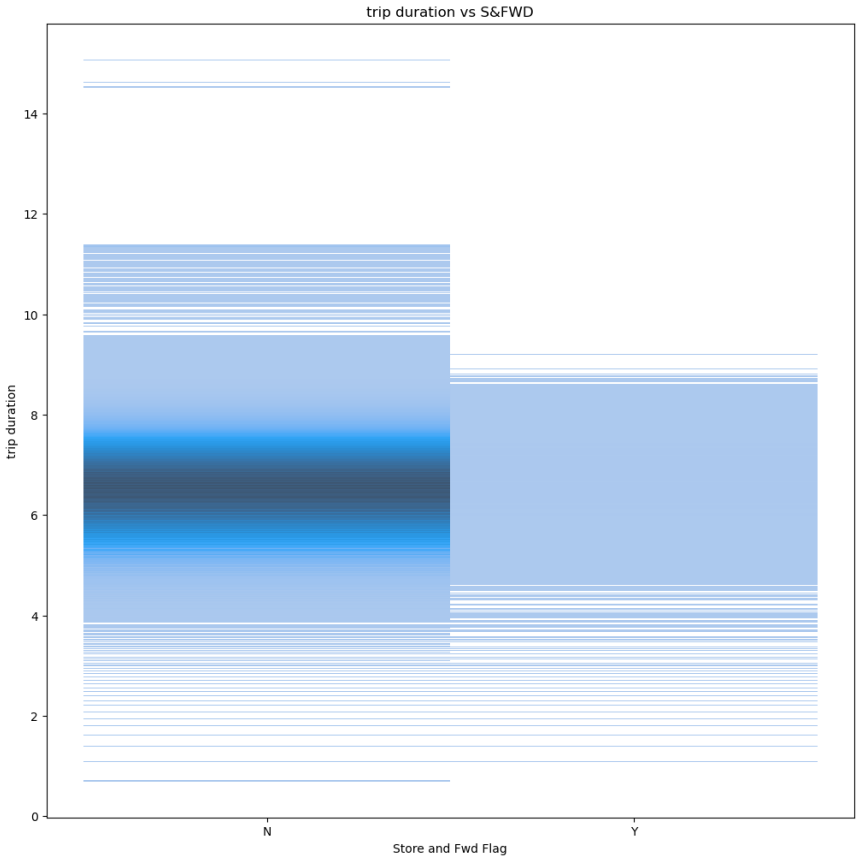

Figure 2-4

2.4 Geographical Data

Looks like Most of trips are 1km to 25 km

Most of trips are [1-40]km/h

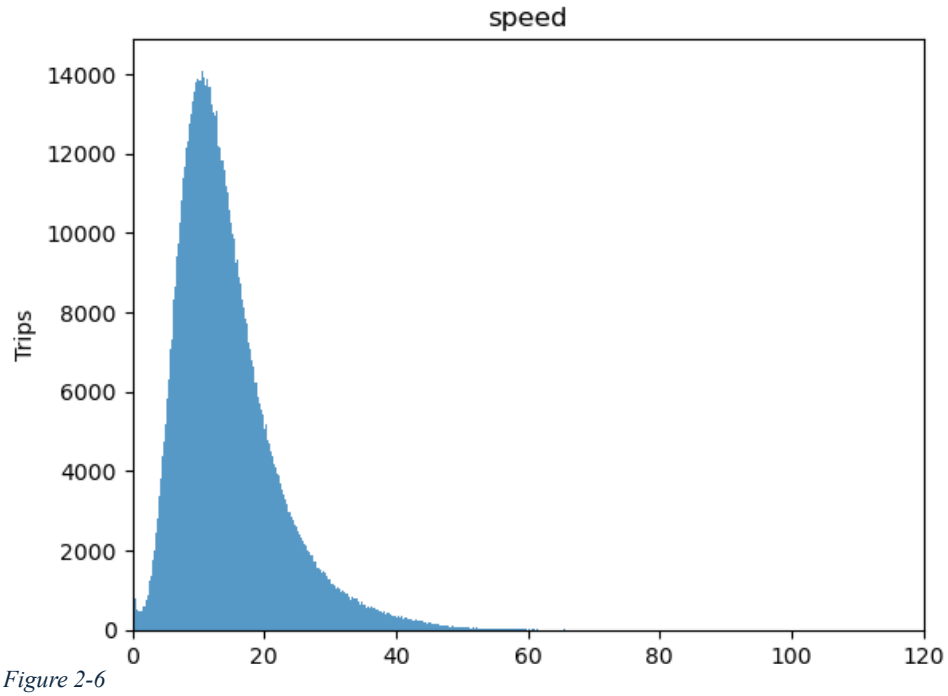

Figure 2-6

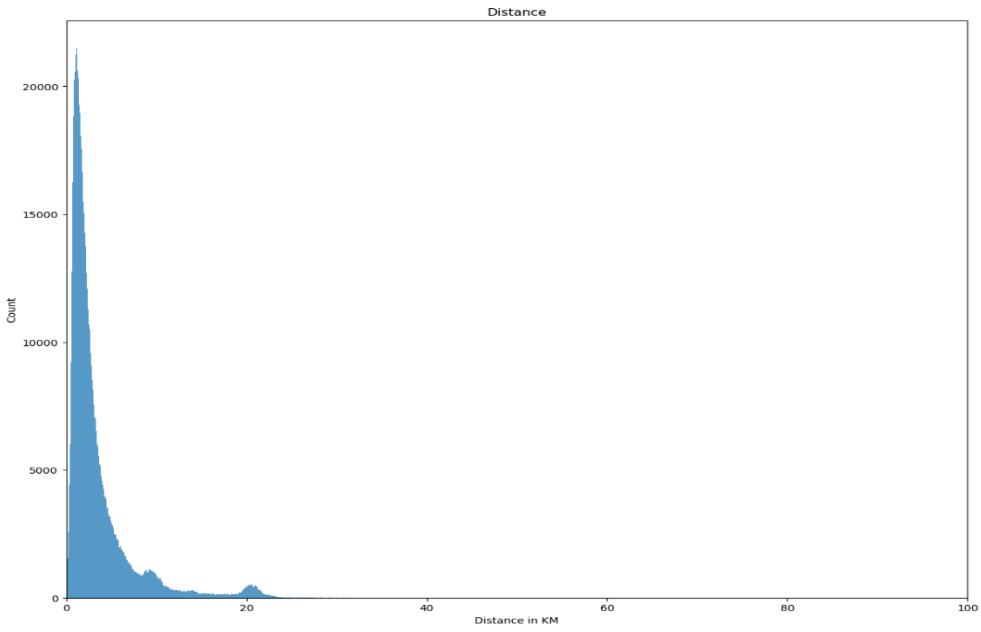

Figure 2-5

This distribution’s right skewed So we can use a transformation

Specially log transformations

which can improve the performance of our Linear model

2.5 Temporal/Time-Date Analysis

I am trying to find a relationship between time and day of weak with trip durations

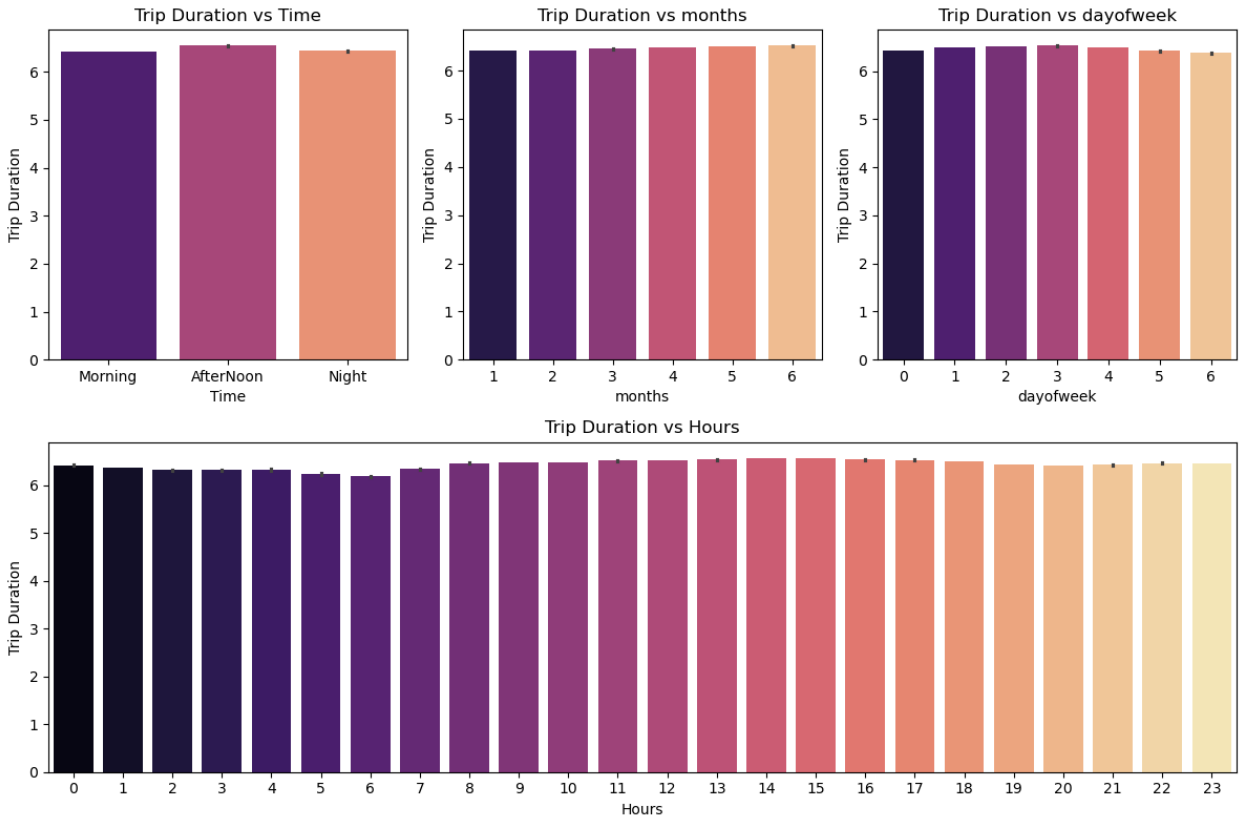

Figure 2-7

- looks like most trip durations are afternoon and this okay as it is normal in this time "Crowd"

- months of 4,5,6 have longer duration may be to the start of summer, traveling in these months

- duration takes longer on rush hours and this might be of the fact that is middle of day most people are out doing different activities

- short duration during evening and morning due to fact people adore just making purpose trips

2.6 Correlation Analysis

- There is positive relation trip duration with pickup longitude, drop-off longitude and passenger count.

- There negative relation trip duration with pickup latitude and drop-off latitude.

- no huge relation

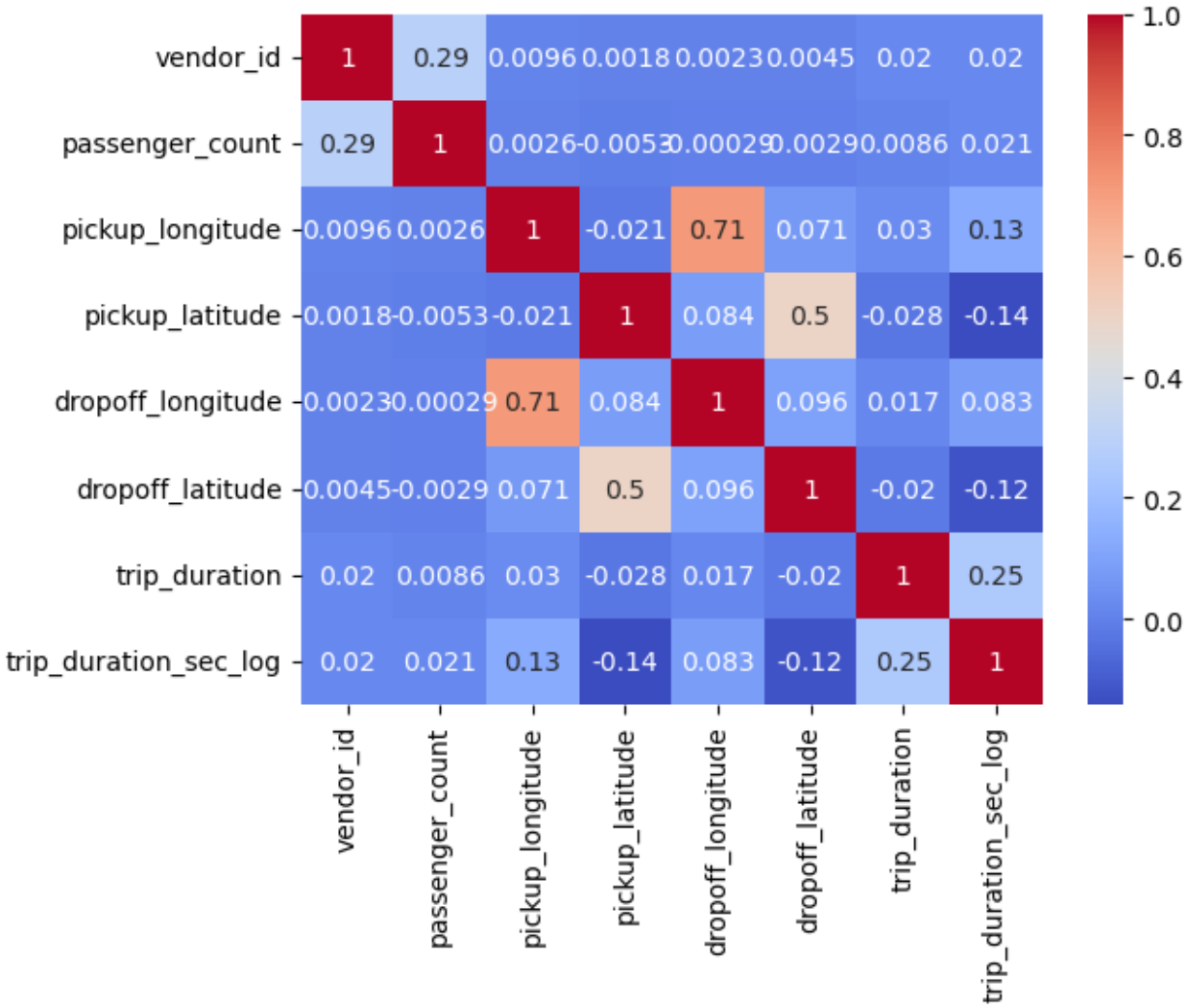

Figure 2-8

- There is a strong positive relation between trip duration and distance.

- There is negative relation between trip duration and speed kmh.

- I think we Can make use of distance feat in modeling

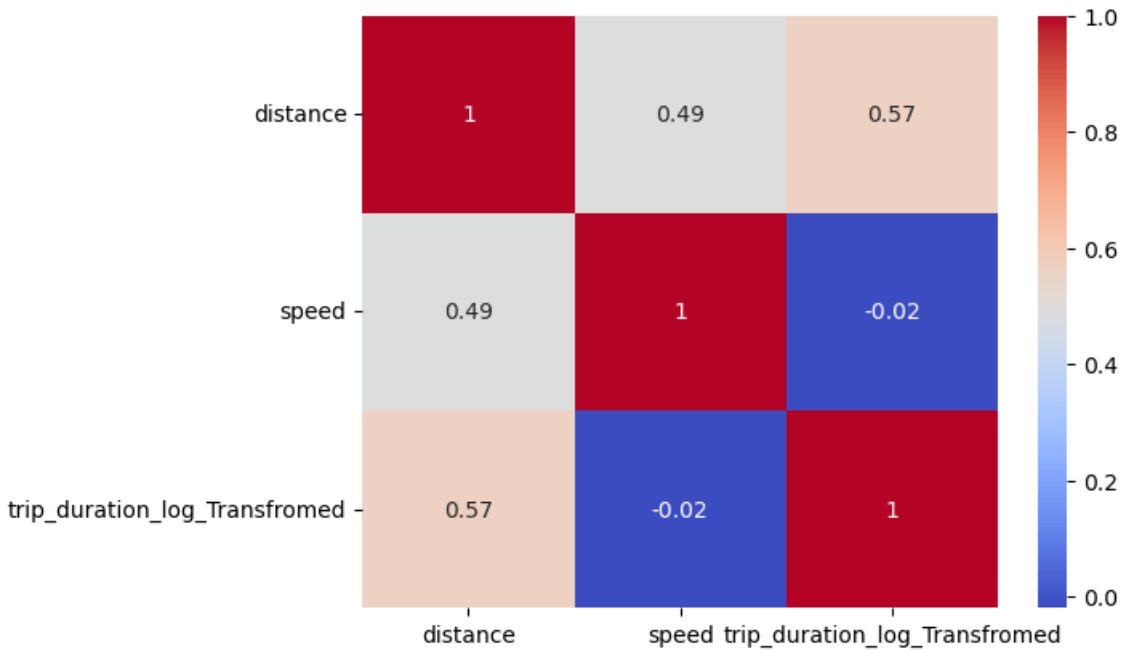

Figure 2-9

3 Modeling

We perform One hot encoding for the categorical feature and for the numerical feature We scale the data by standard scalar then do Polynomial Features (degree=6) finally We use log transform for the data. As previously discussed, we perform log transformation because the distance data is right skewed

3.1 Results

Table 1 Performance Metrics for model

Metric	Train	Validation
R2 Score	0.7255	0.675

3.2 Future Work

Having a type of version control for the data or model scores is very beneficial for error analysis and verifying assumptions.

In this project, we observed the following insights:

• Feature selection consistently improves model performance.

• Outlier removal using z-score improve model performance.