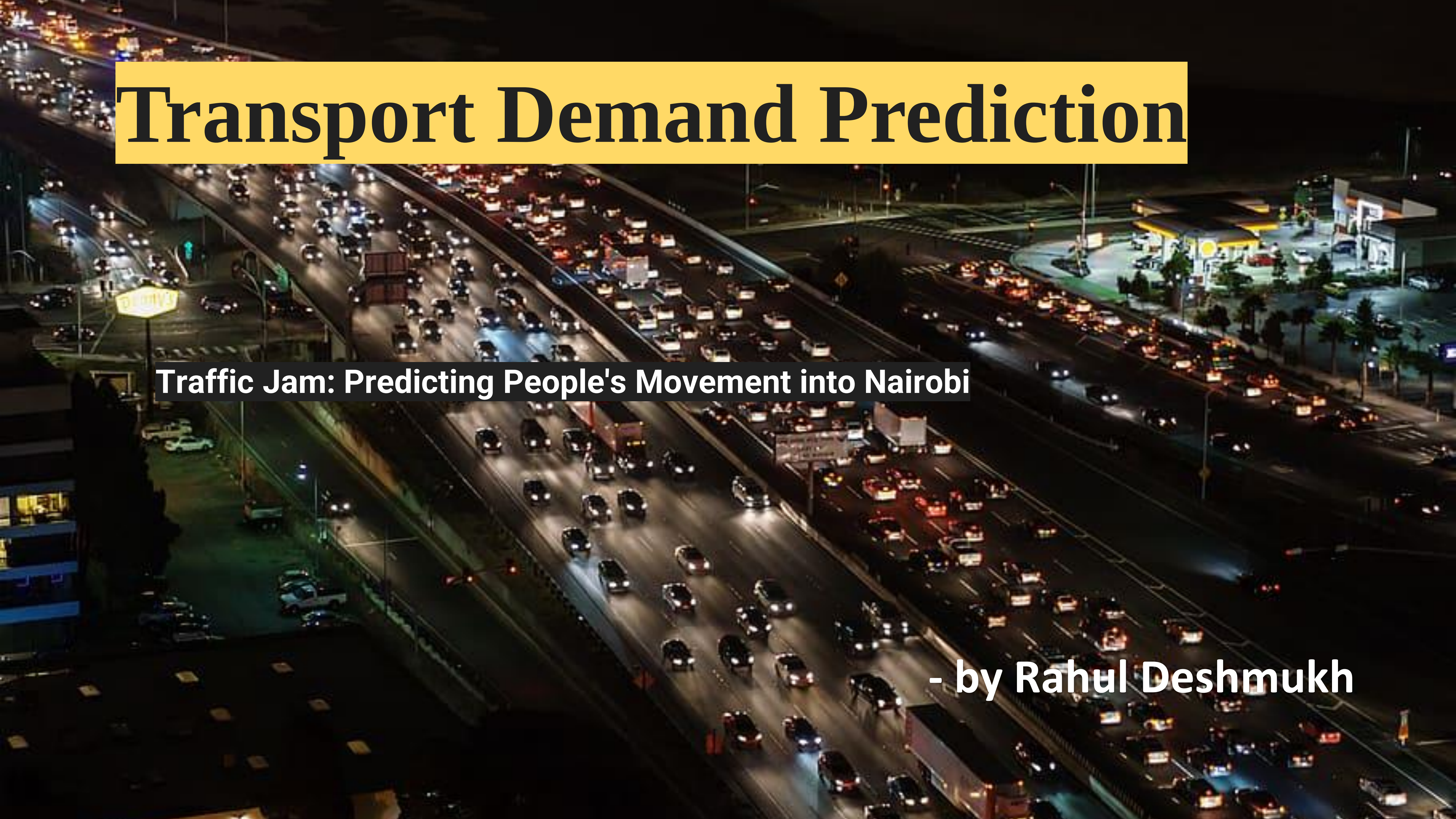


Transport Demand Prediction

Traffic Jam: Predicting People's Movement into Nairobi

- by Rahul Deshmukh

Points To Discussed:

-Data Summary

-Data Cleaning

-Defining problem statement

-EDA

-Feature Selection

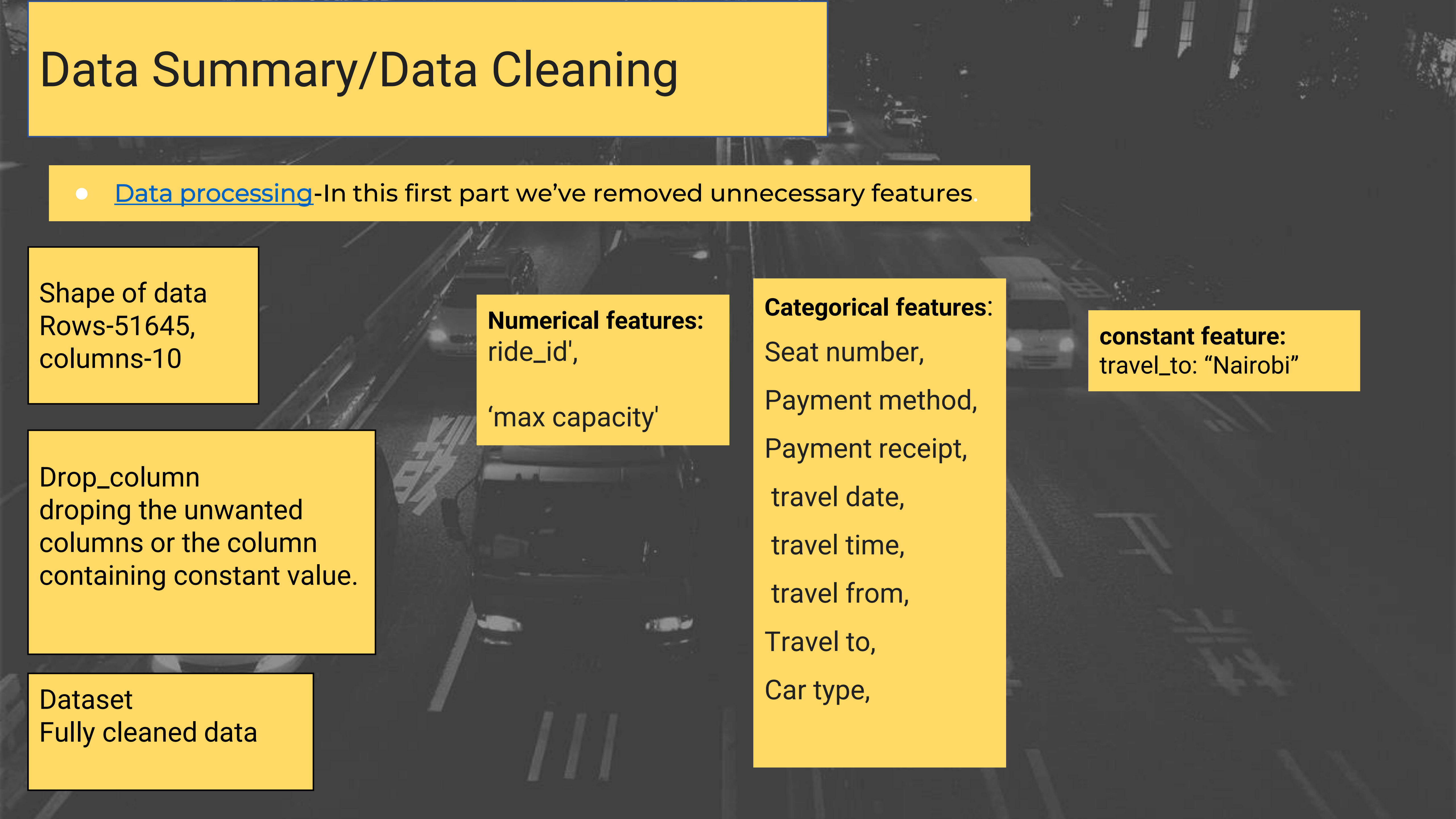
-Preparing dataset for modeling

-Applying Model

-Model validation and selection

-Conclusion

Data Summary/Data Cleaning



- Data processing-In this first part we’ve removed unnecessary features.

Shape of data
Rows-51645,
columns-10

Drop_column
droping the unwanted columns or the column containing constant value.

Dataset
Fully cleaned data

Numerical features:
ride_id',

‘max capacity'

Categorical features:

Seat number,

Payment method,

Payment receipt,

travel date,

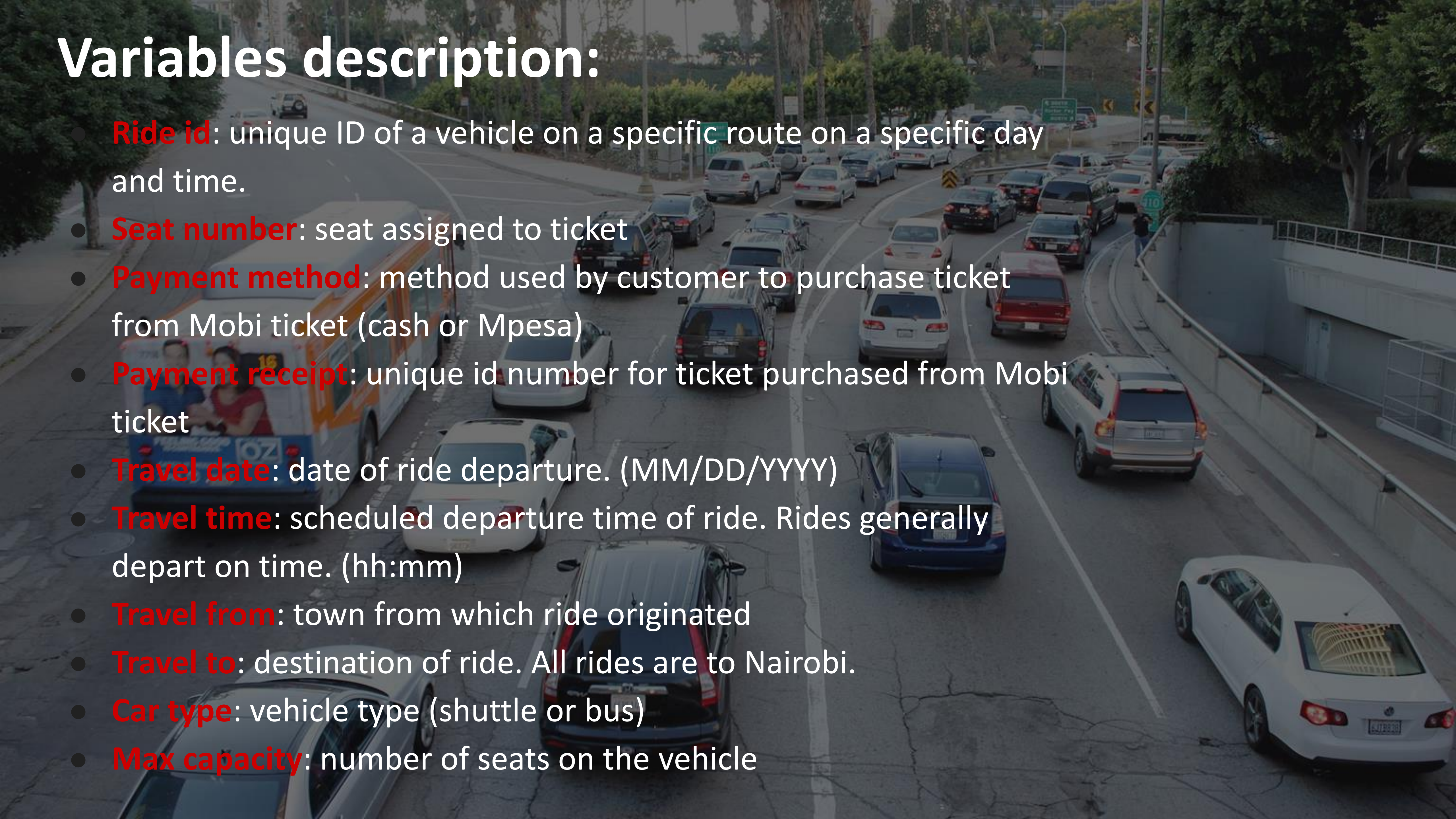
travel time,

travel from,

Travel to,

Car type,

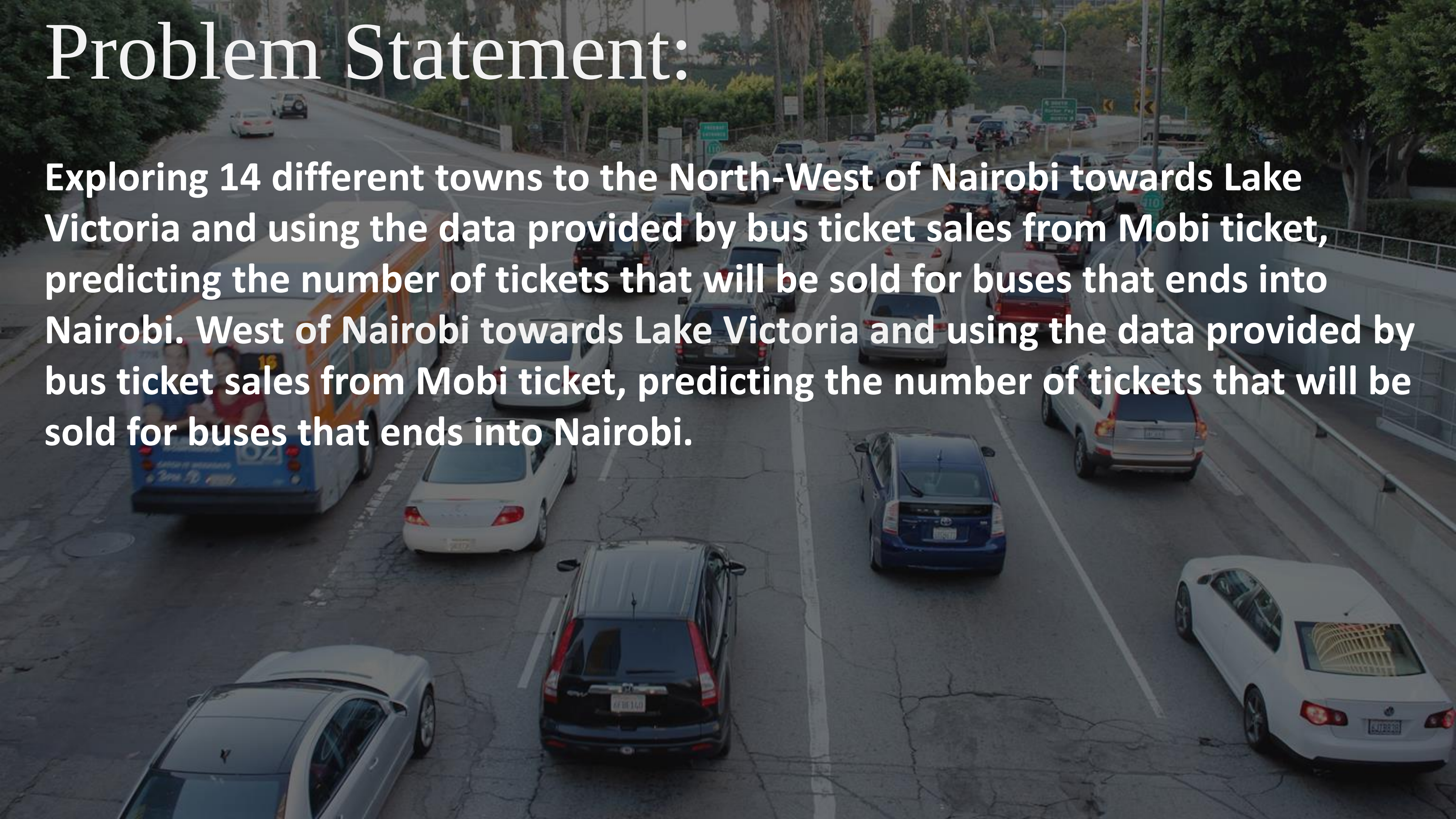
constant feature:
travel_to: “Nairobi”



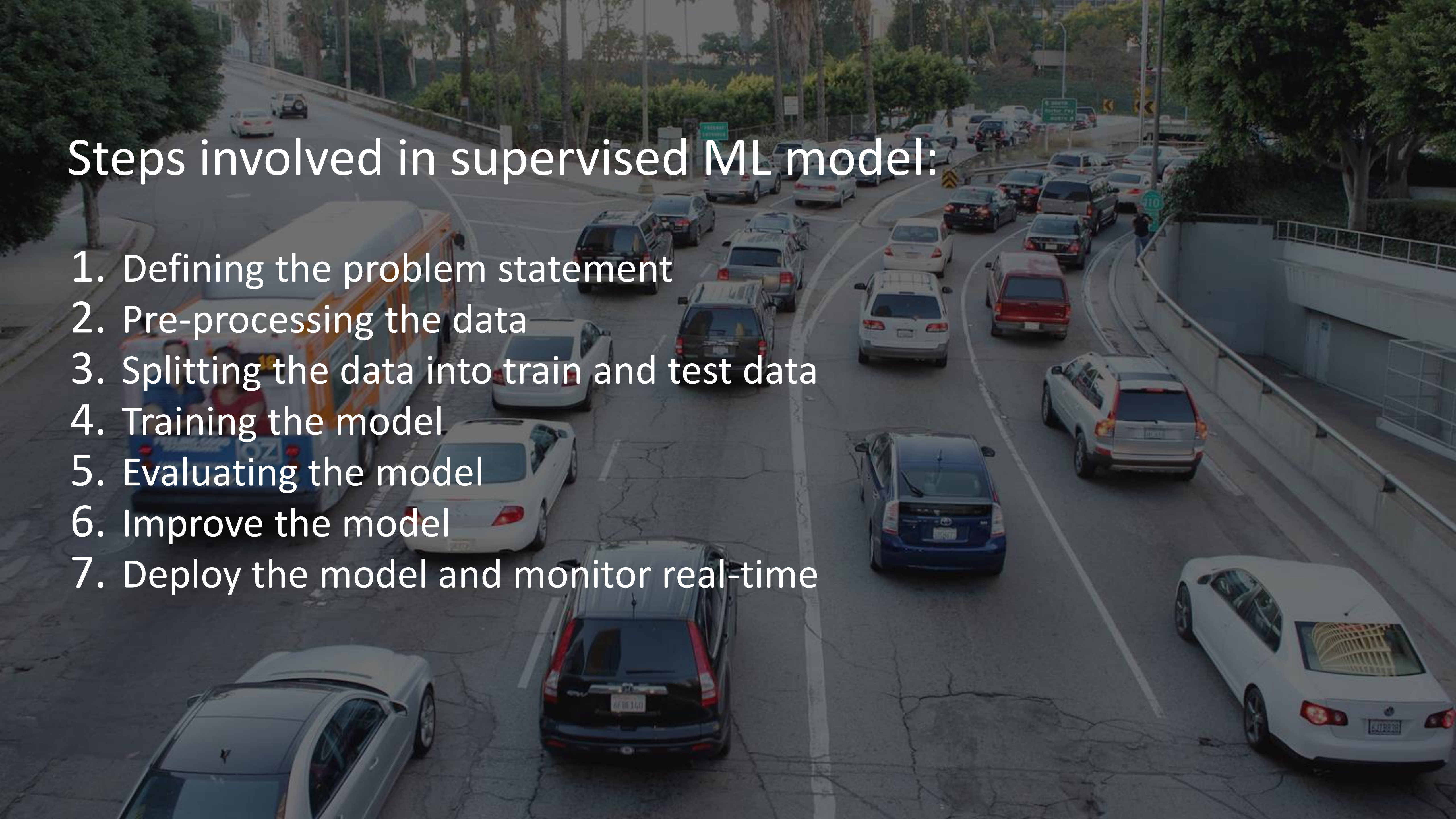
Variables description:

- **Ride id**: unique ID of a vehicle on a specific route on a specific day and time.
- **Seat number**: seat assigned to ticket
- **Payment method**: method used by customer to purchase ticket from Mobi ticket (cash or Mpesa)
- **Payment receipt**: unique id number for ticket purchased from Mobi ticket
- **Travel date**: date of ride departure. (MM/DD/YYYY)
- **Travel time**: scheduled departure time of ride. Rides generally depart on time. (hh:mm)
- **Travel from**: town from which ride originated
- **Travel to**: destination of ride. All rides are to Nairobi.
- **Car type**: vehicle type (shuttle or bus)
- **Max capacity**: number of seats on the vehicle

Problem Statement:



Exploring 14 different towns to the North-West of Nairobi towards Lake Victoria and using the data provided by bus ticket sales from Mobi ticket, predicting the number of tickets that will be sold for buses that ends into Nairobi. West of Nairobi towards Lake Victoria and using the data provided by bus ticket sales from Mobi ticket, predicting the number of tickets that will be sold for buses that ends into Nairobi.



Steps involved in supervised ML model:

1. Defining the problem statement
2. Pre-processing the data
3. Splitting the data into train and test data
4. Training the model
5. Evaluating the model
6. Improve the model
7. Deploy the model and monitor real-time

EDA(EXPLORATORY DATA ANALYSIS):

Analysis of different variables:. Analyzing the payment method

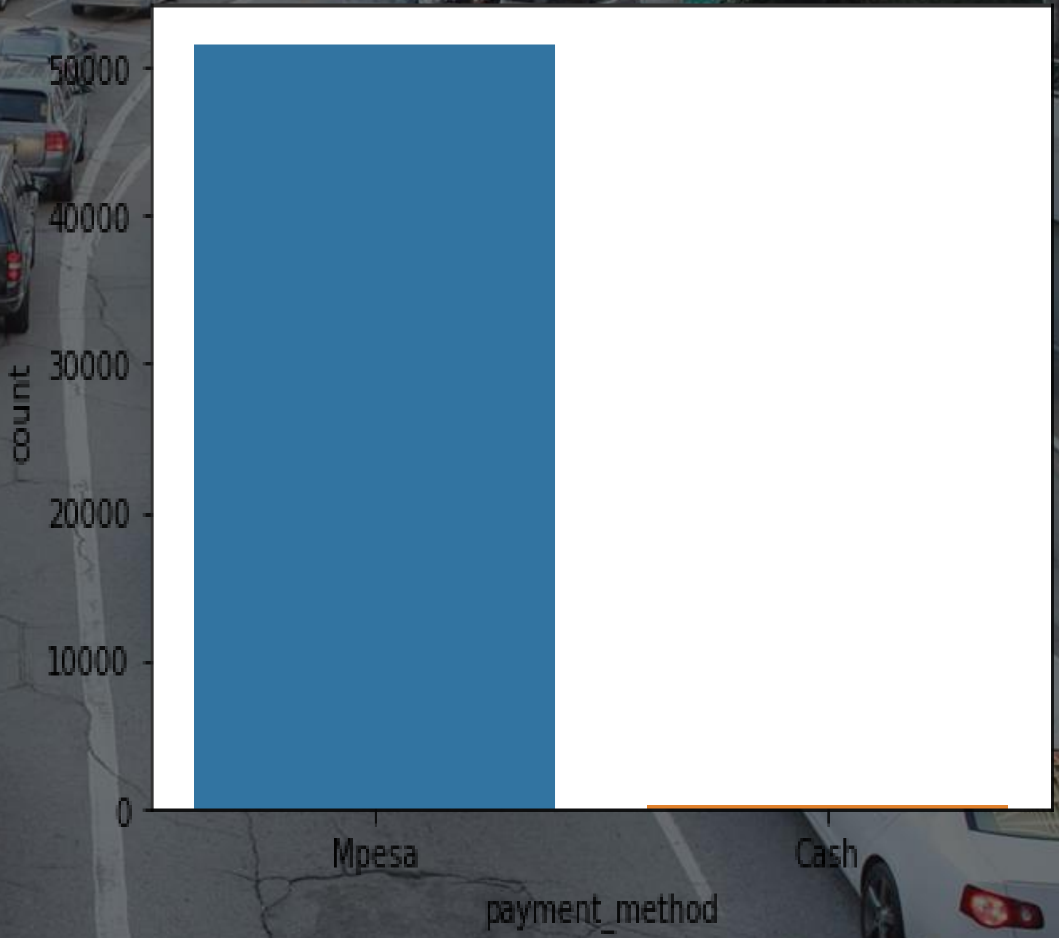
There are two type of payment methods people have used to buy the tickets.

Travelers have used 2 types of payment method.

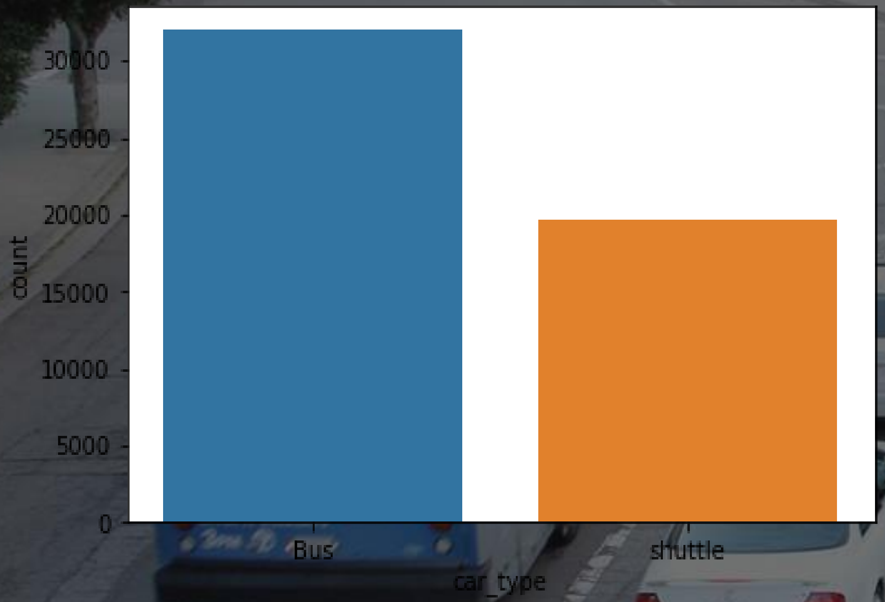
They are Mpesa and cash.

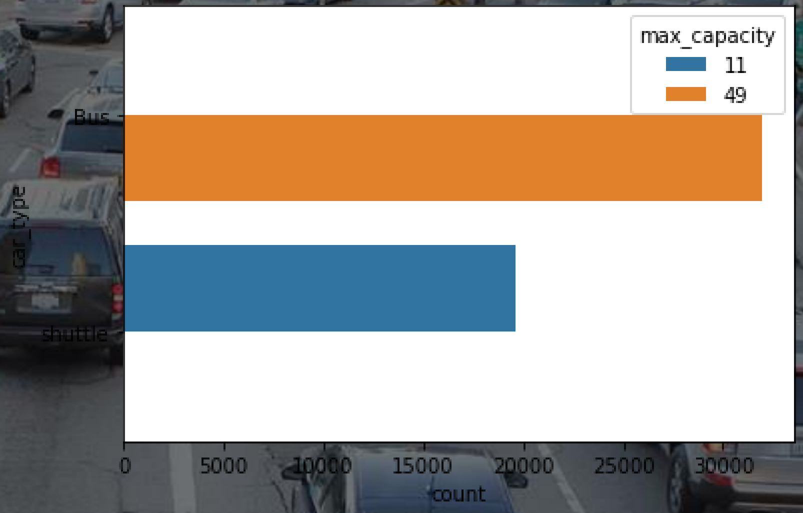
And the most of the people have used Mpesa

to pay for their ticket.



Analyzing the vehicle type used by the customer and number of seats on the vehicle sell



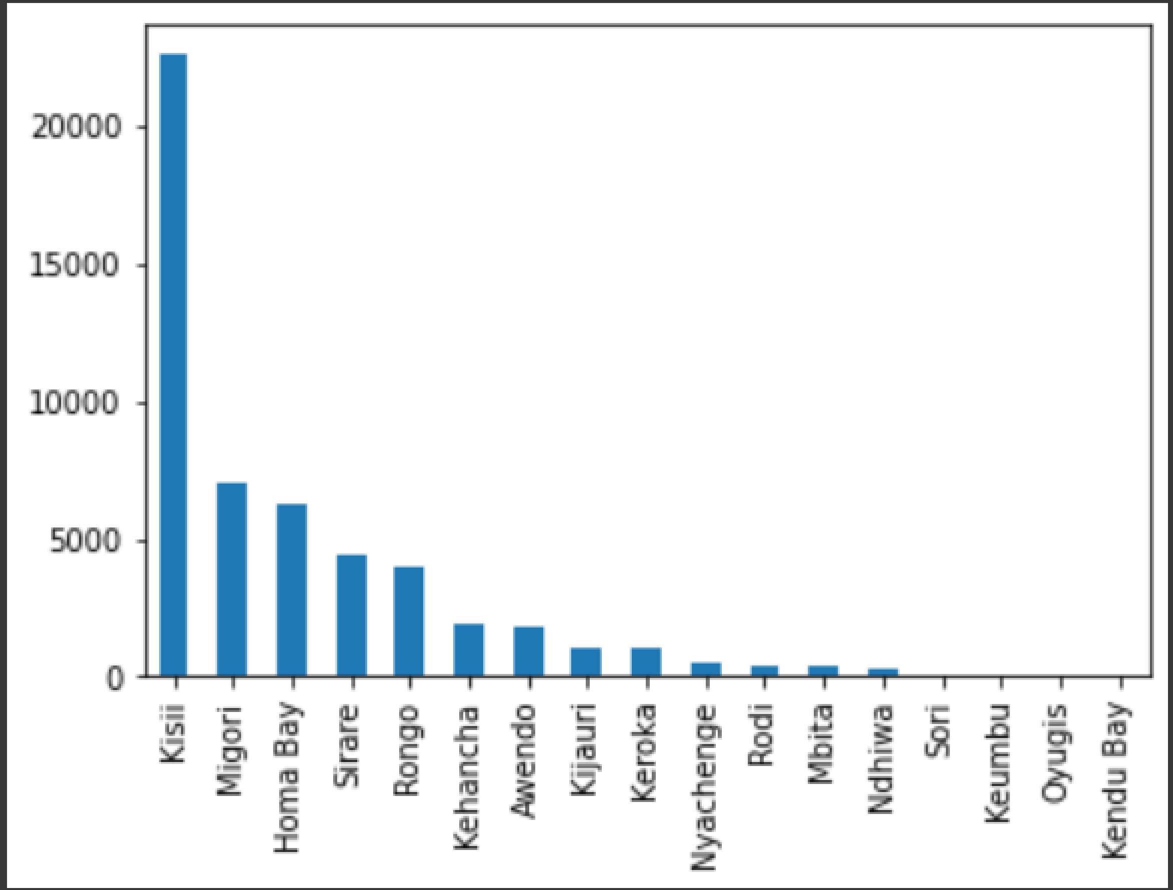


There are 2 different types of car used(shuttle and bus) and most of them are bus.

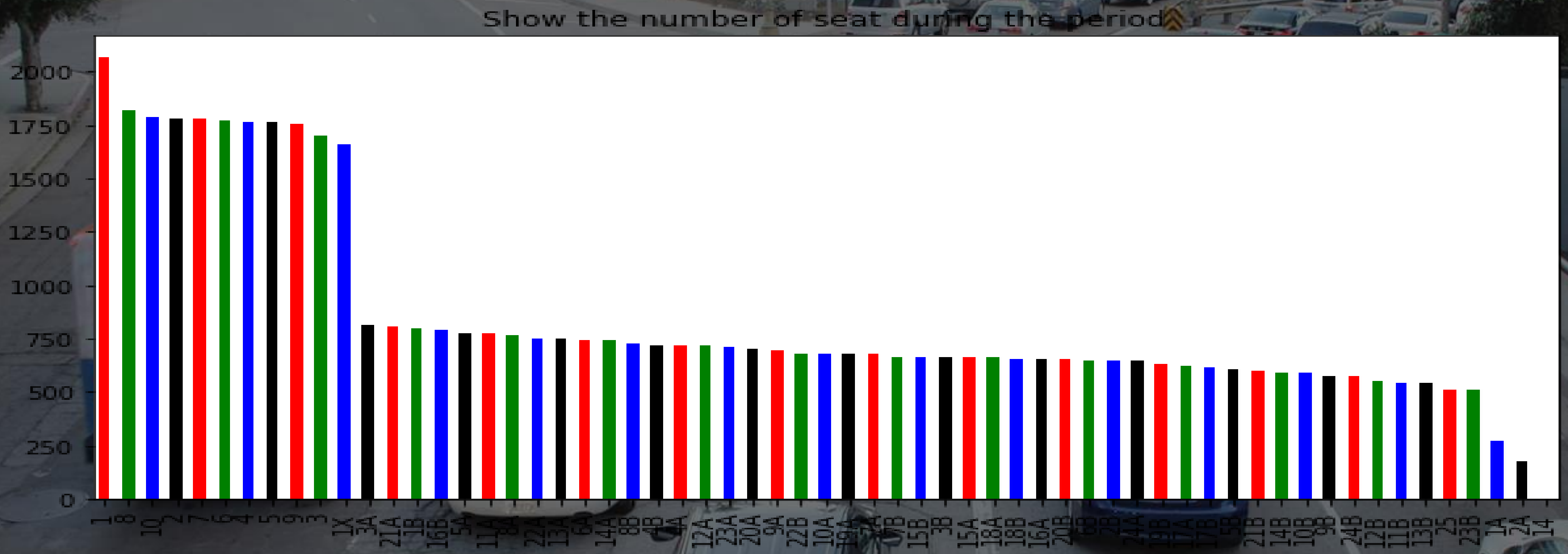
There are two type of cars Bus and shuttle and the maximum capacity of the bus is 49 while shuttle can contain 11 travelers.

Analyzing the travel from town from which ride originated

Most customers travel from Kisii town to Nairobi.



Analyzing the number of seat during the period using bar graph



There are totally 61 unique seats in this dataset. The record of 149 unique days are present in this dataset out of 2 years.

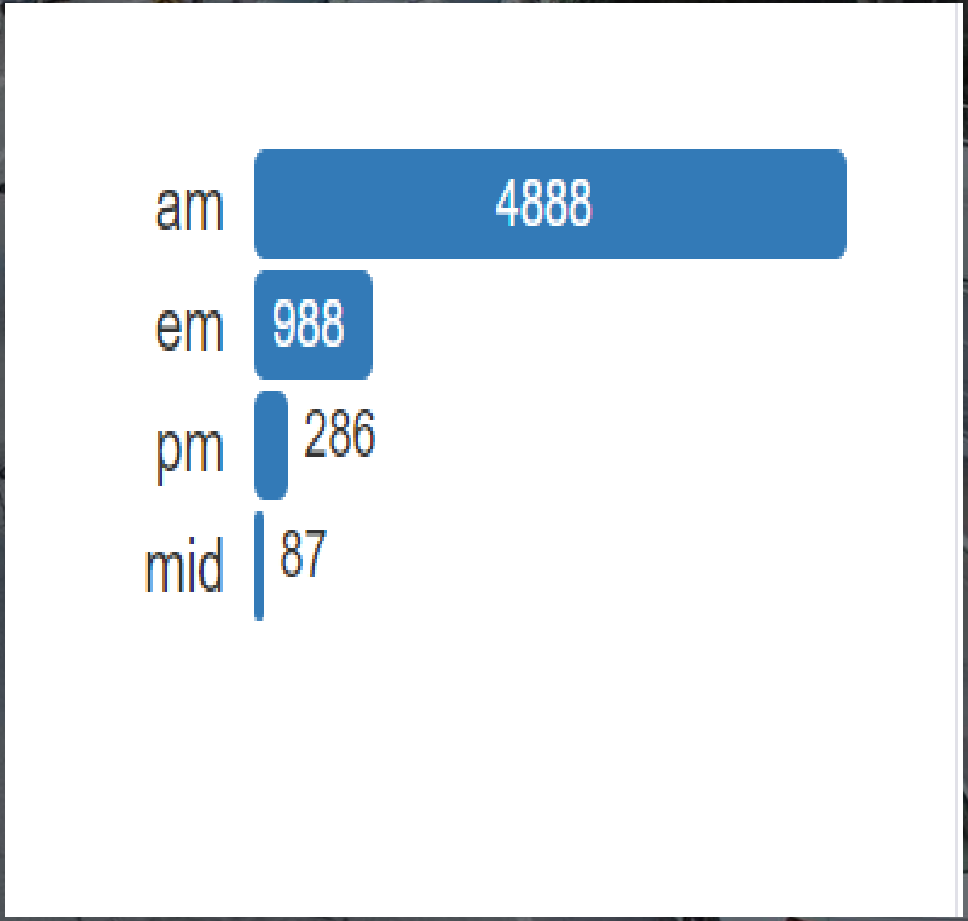
Analyzing the number of persons travelling in which time zone

The most of the people used to travel in

the morning time compared to other times.

Here we can see that 4888 customers

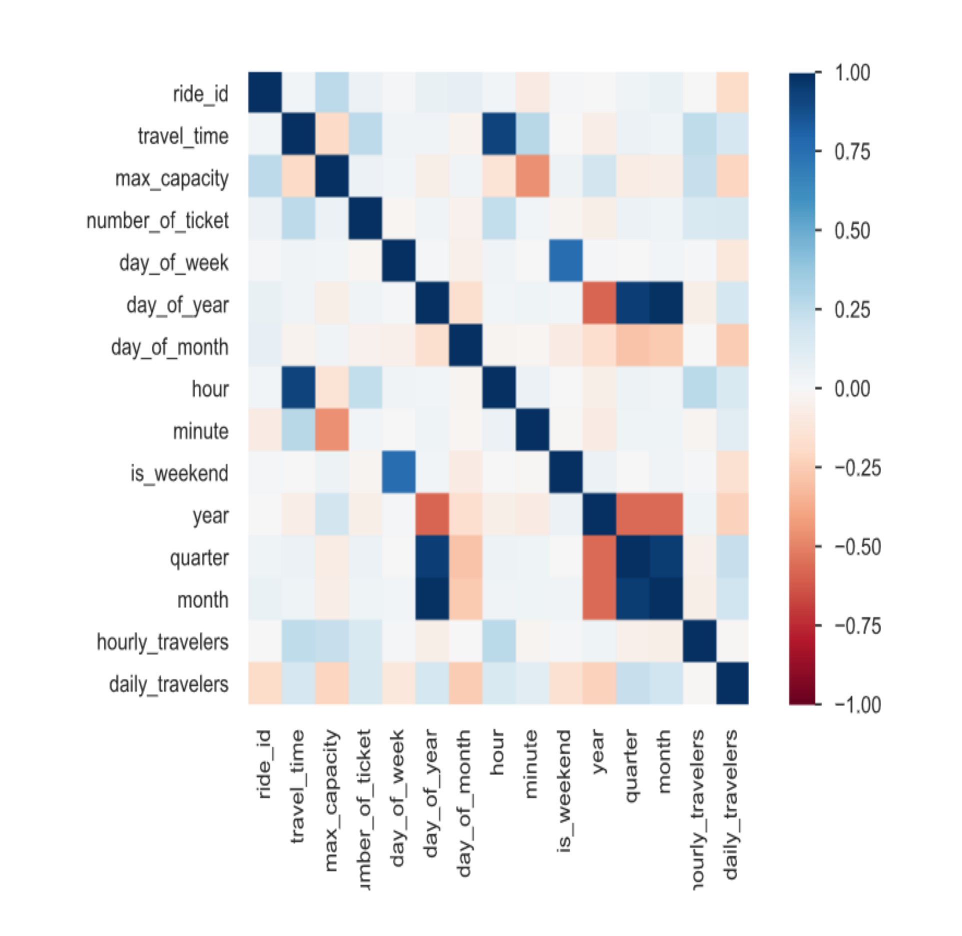
were travelled in the morning.



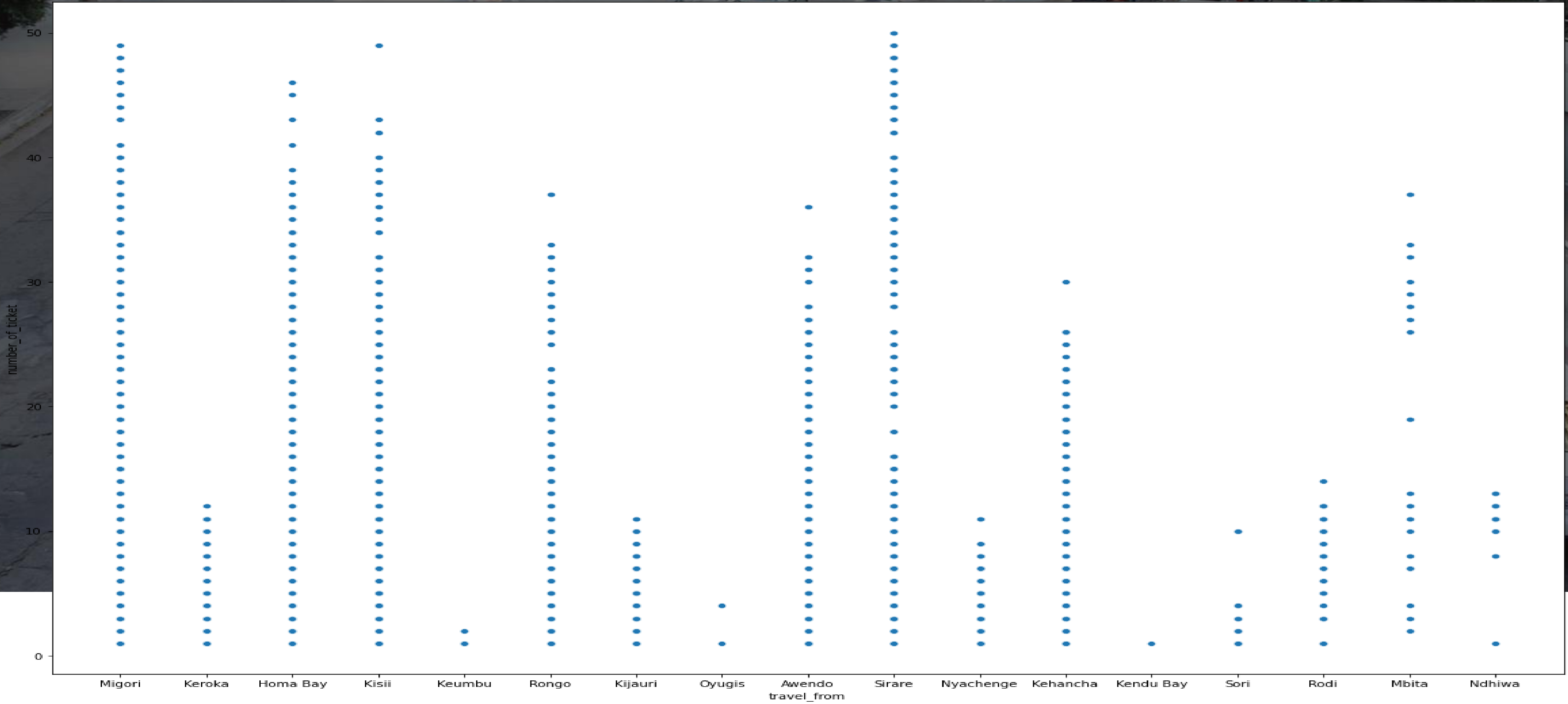
Correlation

Correlation coefficients quantify the association between variables or features of a dataset.

This heatmap of correlation between different independent variables. The matrix shows the coefficients in the squared form colored as per the intensity scale. The map has positive covariance as with the increase in one variable another also increasing.



Scatter plot of number of tickets sold to customers from different cities

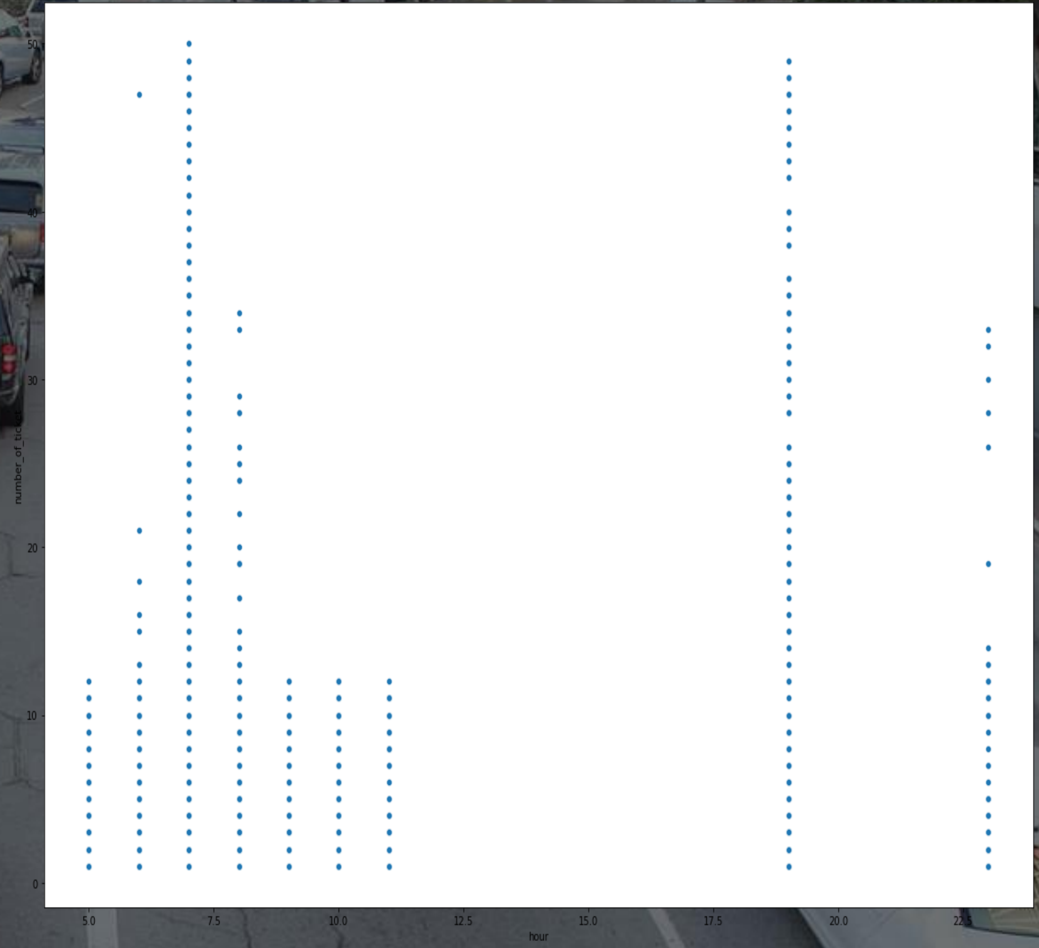


Feature Engineering

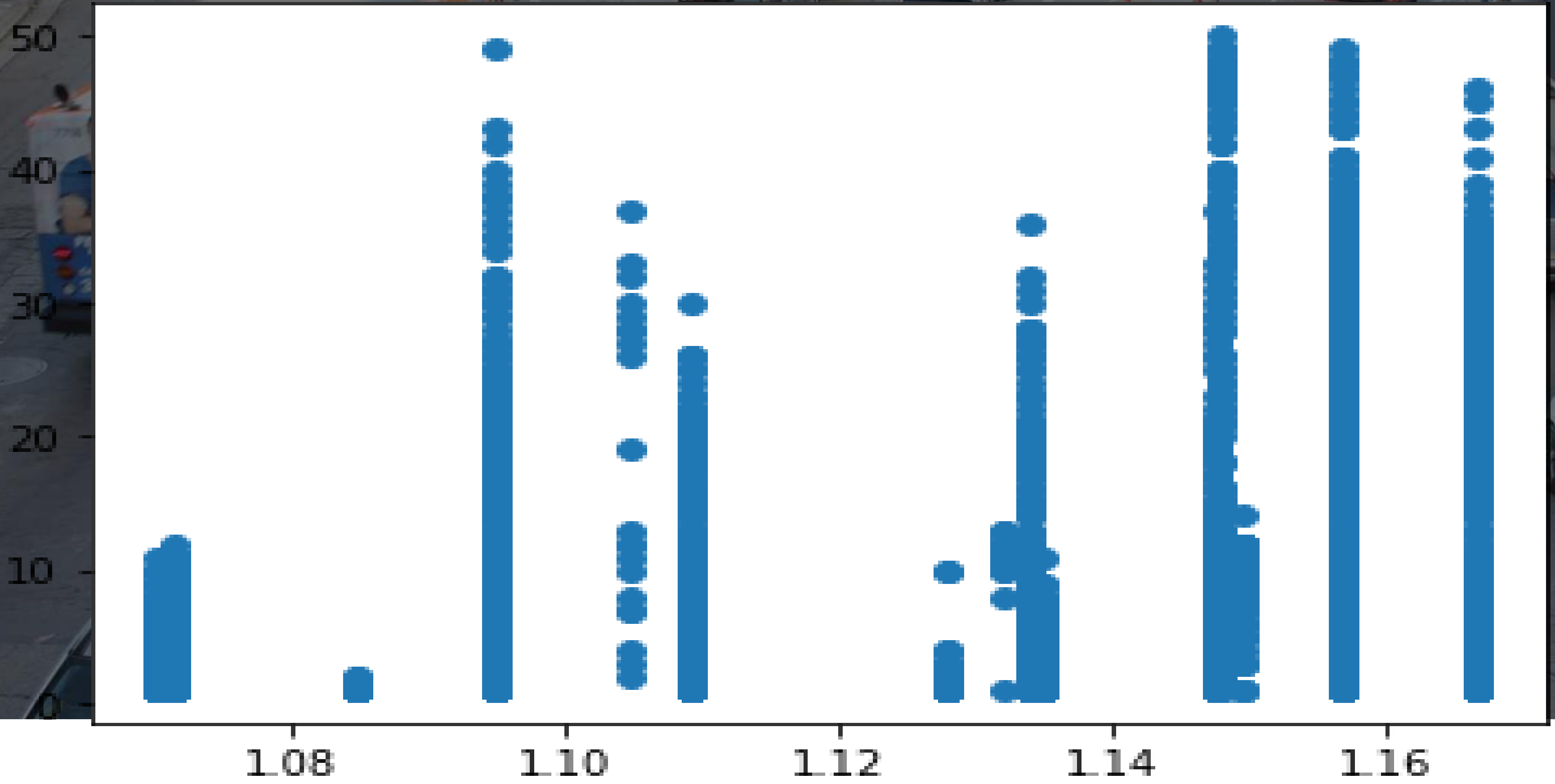
We can see that most of the tickets were sold at 7 AM and 8 PM.

And that seems true because in the morning most of the people go to the work and office.

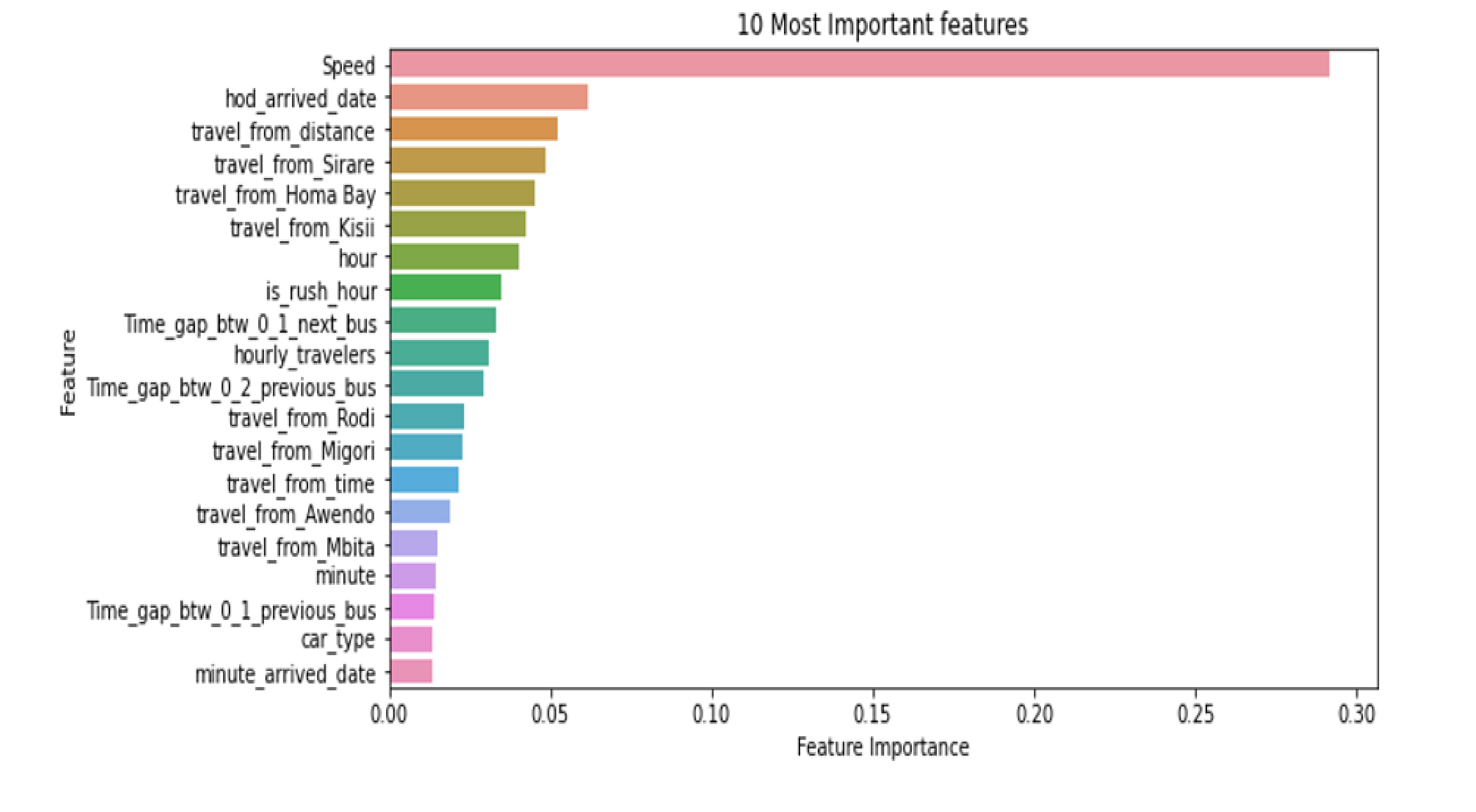
From the above we can say that there is not ride between 12pm to 5.30Pm

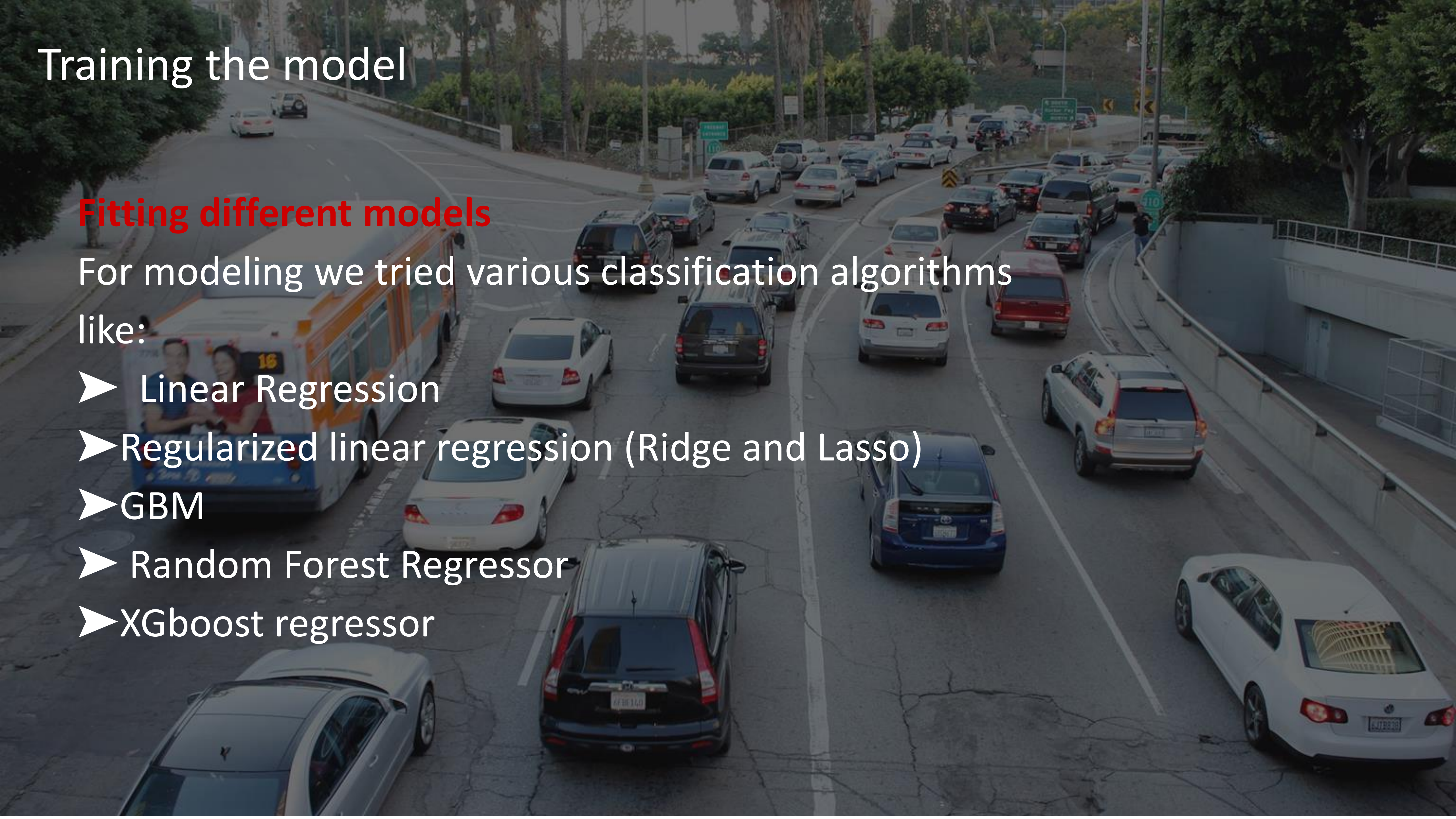


The following figure shows the scatter plot of number of tickets sold to customers verses the speed(the speed here is the time travelled by customers from different distances)



Futuristic Features





Training the model

Fitting different models

For modeling we tried various classification algorithms like:

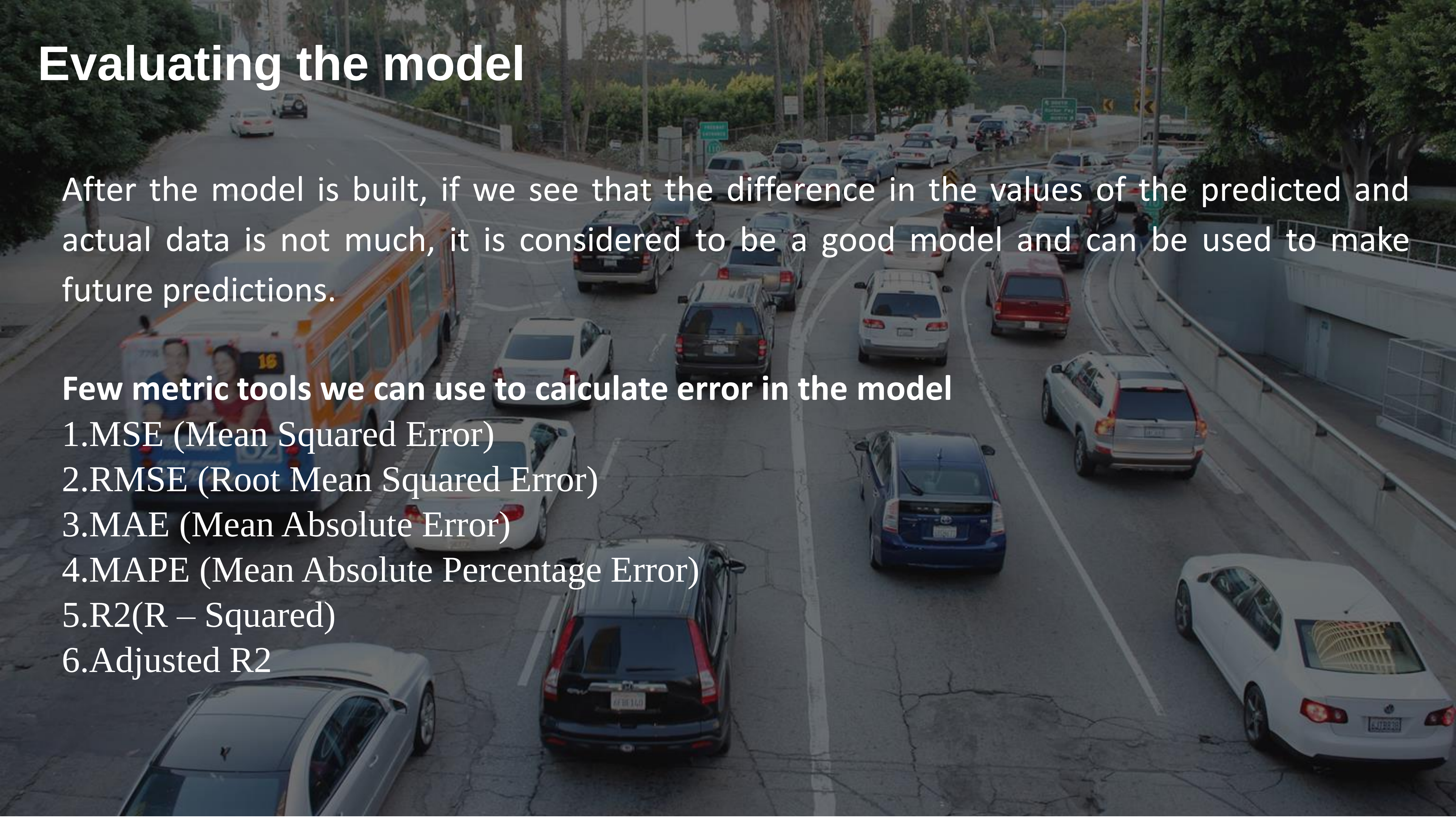
➤ Linear Regression

➤Regularized linear regression (Ridge and Lasso)

➤GBM

➤ Random Forest Regressor

➤XGboost regressor



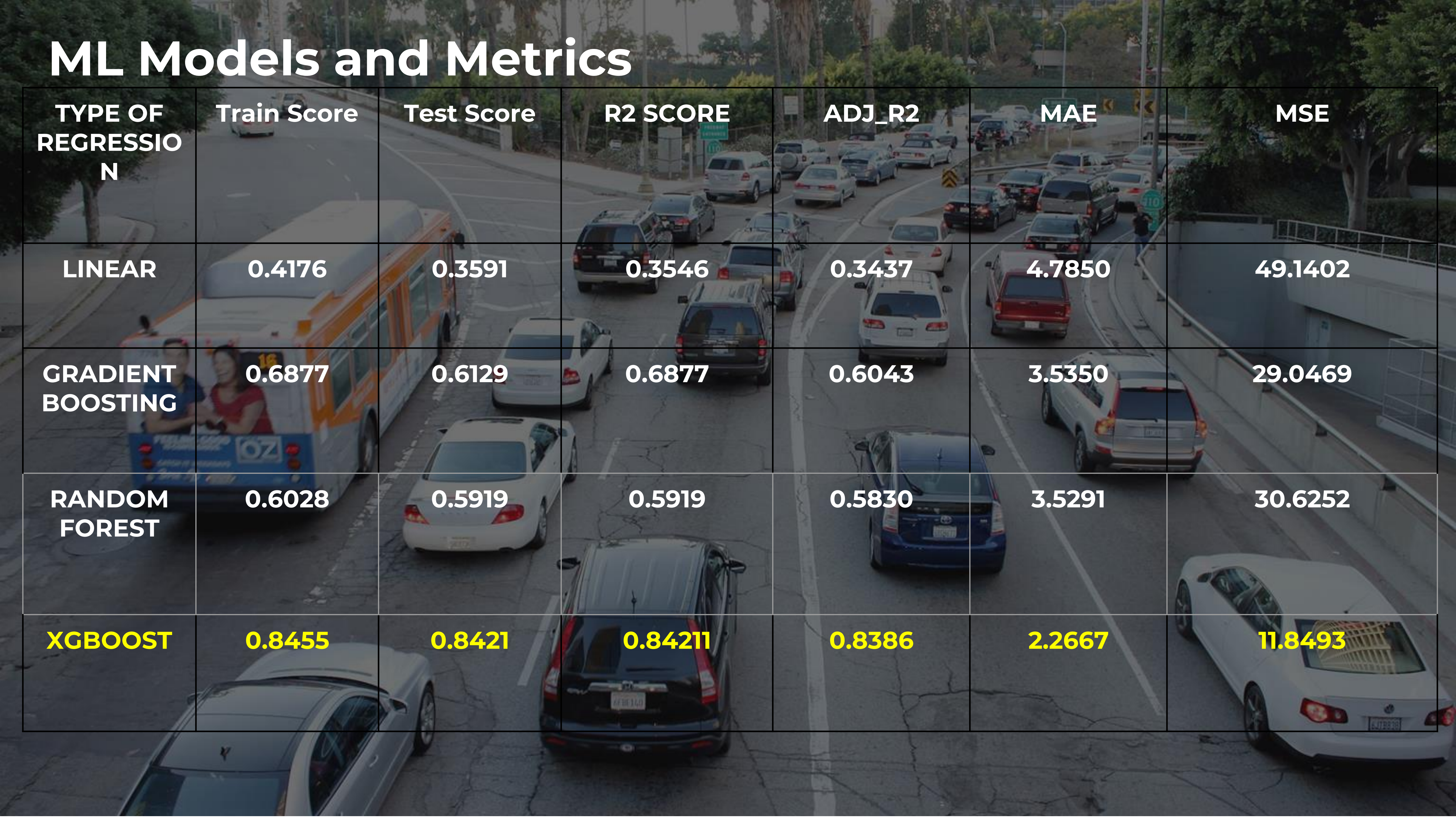
Evaluating the model

After the model is built, if we see that the difference in the values of the predicted and actual data is not much, it is considered to be a good model and can be used to make future predictions.

Few metric tools we can use to calculate error in the model

1.MSE (Mean Squared Error)
2.RMSE (Root Mean Squared Error)
3.MAE (Mean Absolute Error)
4.MAPE (Mean Absolute Percentage Error)
5.R2(R – Squared)
6.Adjusted R2

ML Models and Metrics

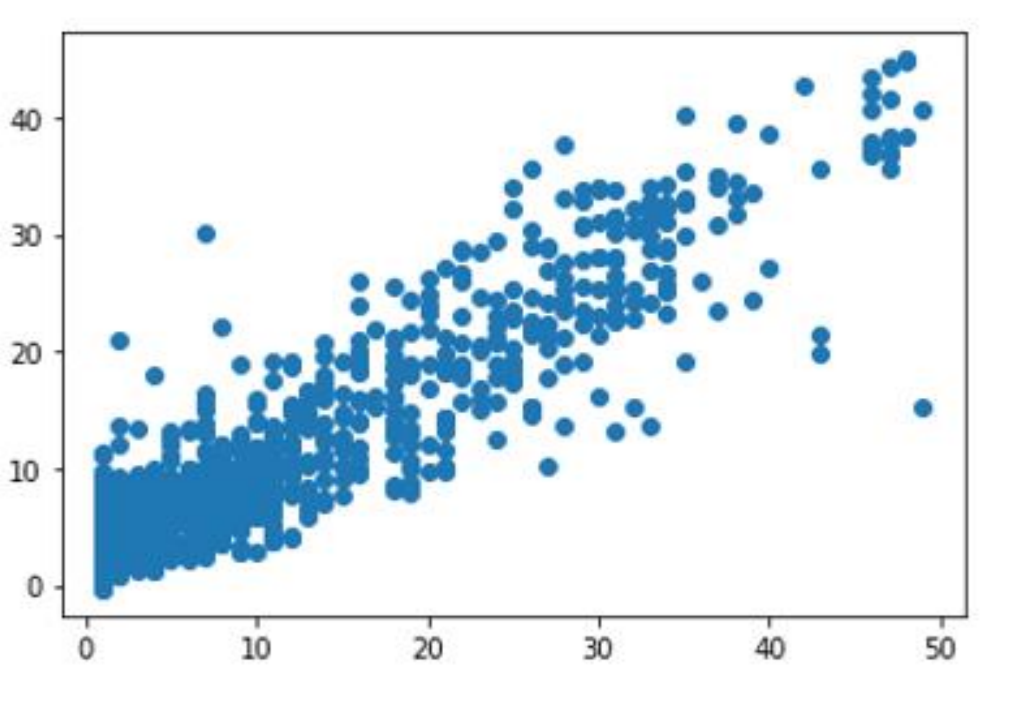


TYPE OF REGRESSION	Train Score	Test Score	R2 SCORE	ADJ_R2	MAE	MSE
LINEAR	0.4176	0.3591	0.3546	0.3437	4.7850	49.1402
GRADIENT BOOSTING	0.6877	0.6129	0.6877	0.6043	3.5350	29.0469
RANDOM FOREST	0.6028	0.5919	0.5919	0.5830	3.5291	30.6252
XGBOOST	0.8455	0.8421	0.84211	0.8386	2.2667	11.8493

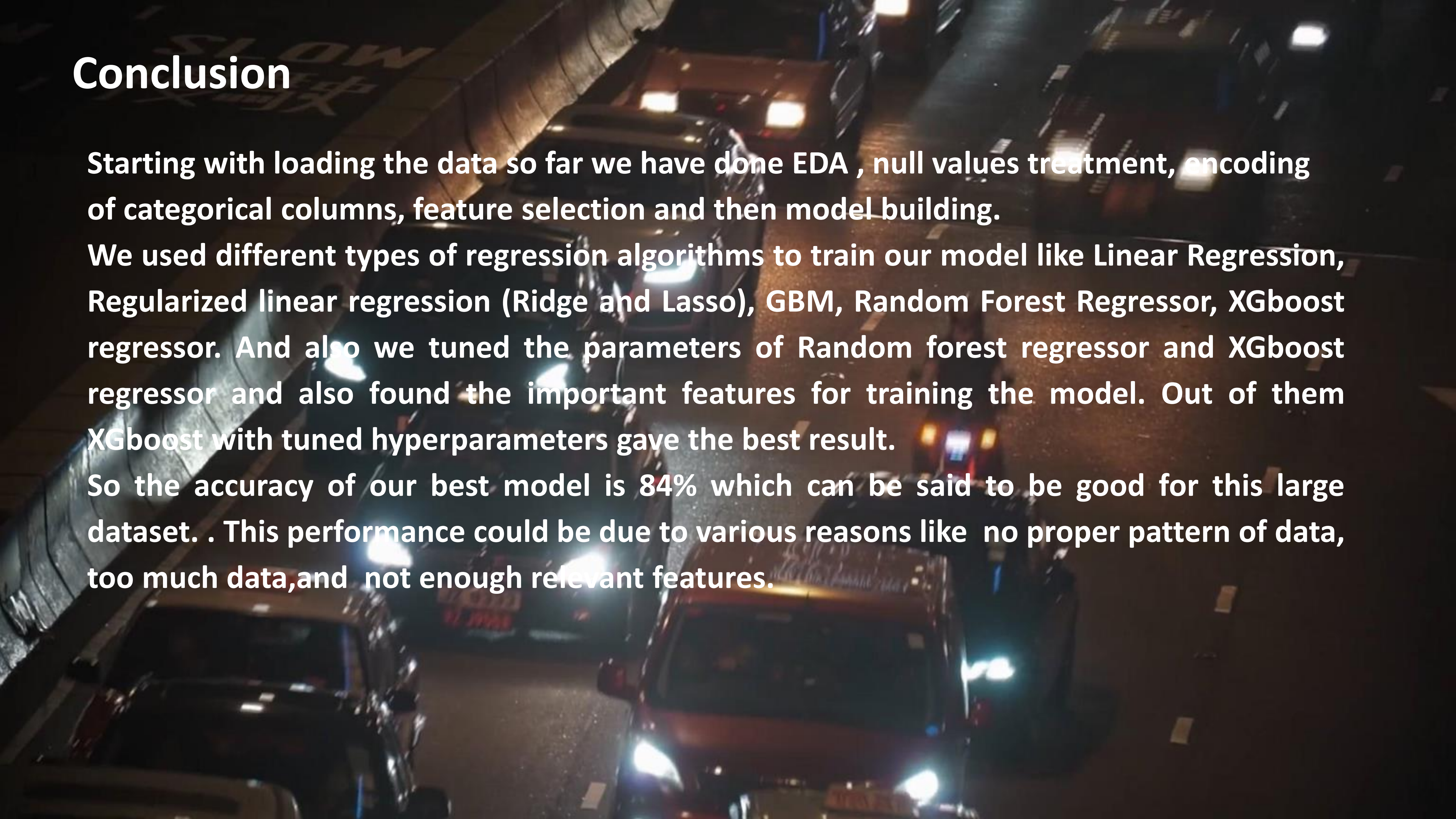
Scatter plot of test and predicted values

XGboost regressor:

We used different types of regression algorithms to train our model like Linear Regression, Regularized linear regression (Ridge and Lasso), GBM, Random Forest Regressor, XGboost regressor. And also we tuned the parameters of Random forest regressor and XGboost regressor and also found the important features for training the model . Out of them XGboost with tuned hyperparameters gave the best result.



XGBoost expects to have the base learners which are uniformly bad at the remainder so that when all the predictions are combined, bad predictions cancels out and better one sums up to form final good predictions.



Conclusion

Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then model building.

We used different types of regression algorithms to train our model like Linear Regression, Regularized linear regression (Ridge and Lasso), GBM, Random Forest Regressor, XGboost regressor. And also we tuned the parameters of Random forest regressor and XGboost regressor and also found the important features for training the model. Out of them XGboost with tuned hyperparameters gave the best result.

So the accuracy of our best model is 84% which can be said to be good for this large dataset. . This performance could be due to various reasons like no proper pattern of data, too much data,and not enough relevant features.

Thank you