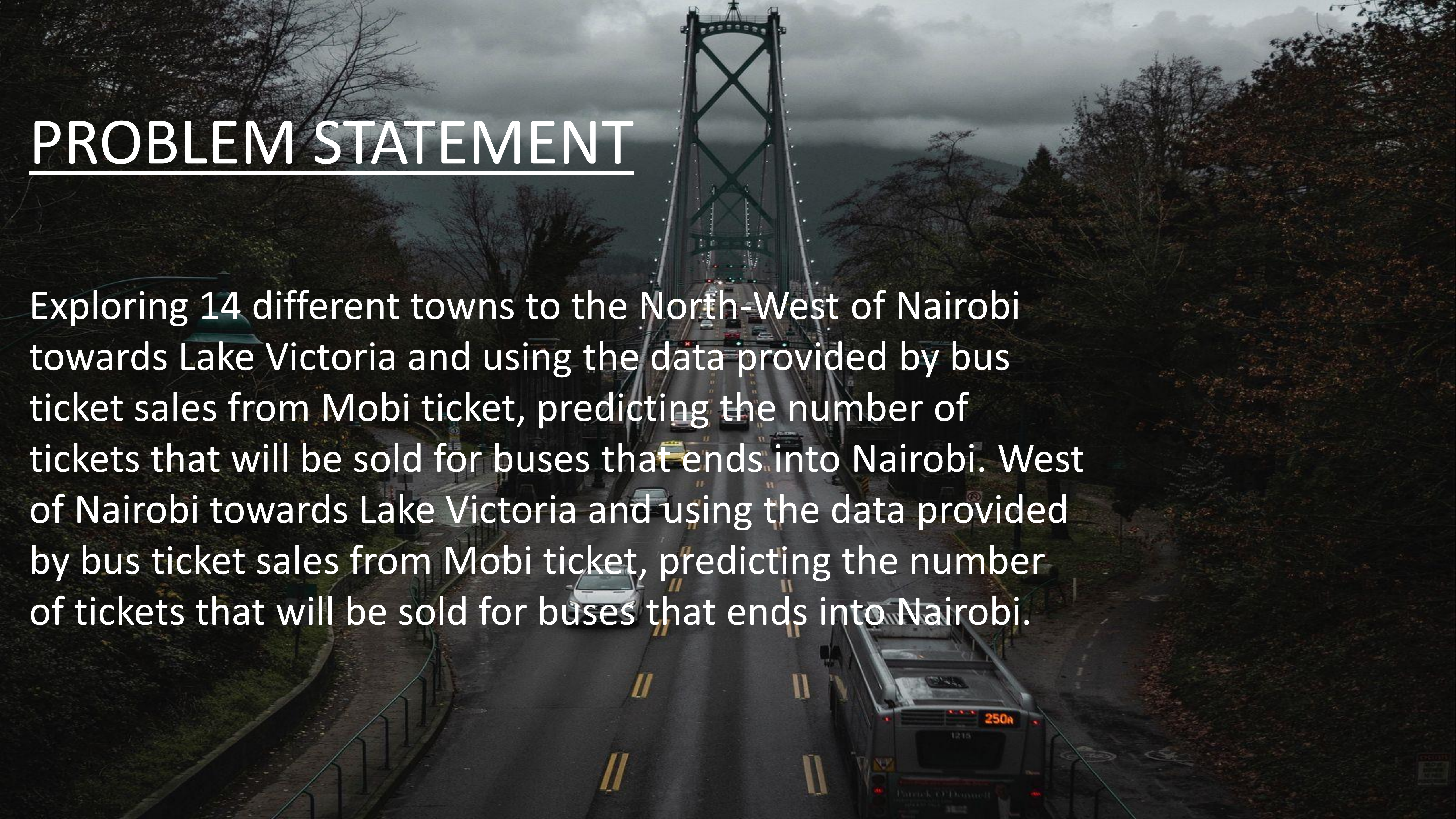
TRANSPORT DEMAND PREDICTION

-BY CHARAN

POINTS TO DISCUSS

DATA SUMMARY
DATA CLEANING
PROBLEM STATEMENT
EDA
FEATURE SELECTION
DATASET PREPARATIONS
APPLYING MODEL
MODEL VALIDATION
CONCLUSION



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.

DATA SUMMARY/CLEANING

SHAPE OF THE DATA
ROW=51645
COLUMNS=10

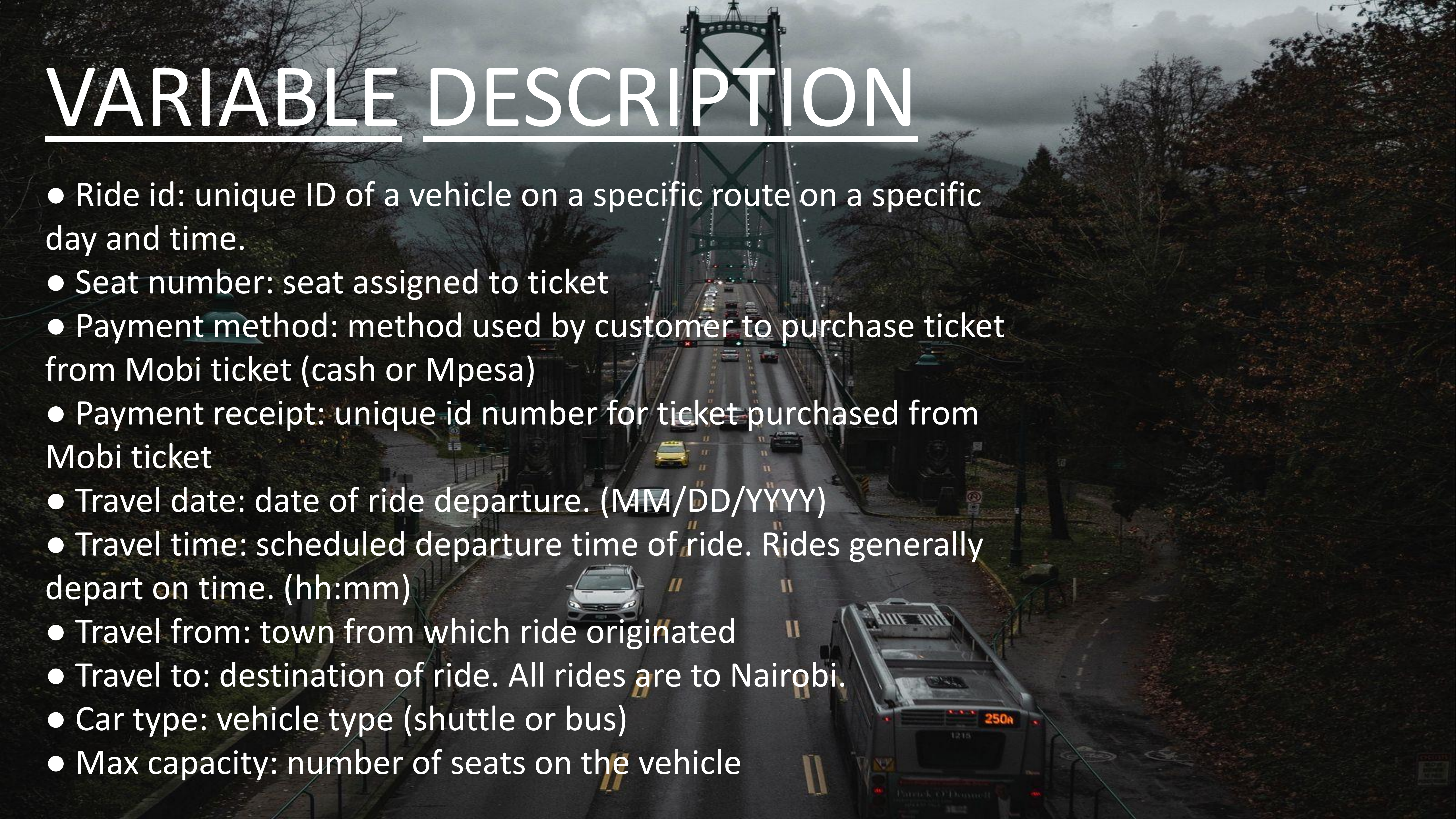
NUMERICAL COLUMNS
‘ride_id’
‘Max_capacity’

CATEGORICAL FEATURE
SEAT_NUMBER
PAYMENT_METHOD
PAYMENT_RECEIPT
TRAVEL_DATE
TRAVEL_TIME
TRAVEL_FROM
TRAVEL_TO
CAR_TYPE

CONSTANT_FEATURE
TRAVEL_TO:’NAIROBI’

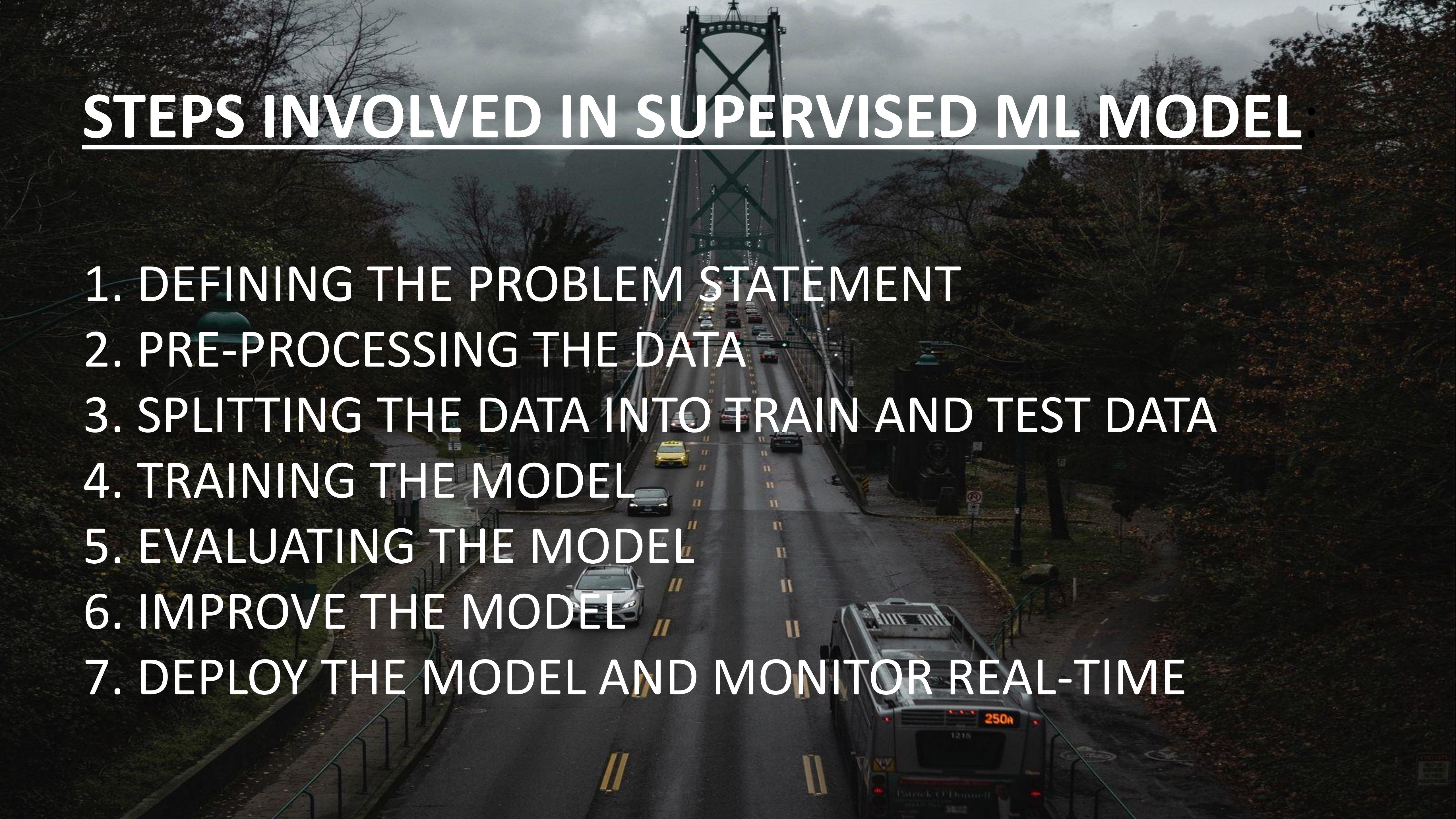
Drop_column
DROPING UNWANTED COLUMNS

MY_DATA
FULLY CLEANED DATASET



VARIABLE 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



STEPS INVOLVED IN SUPERVISED ML MODEL:

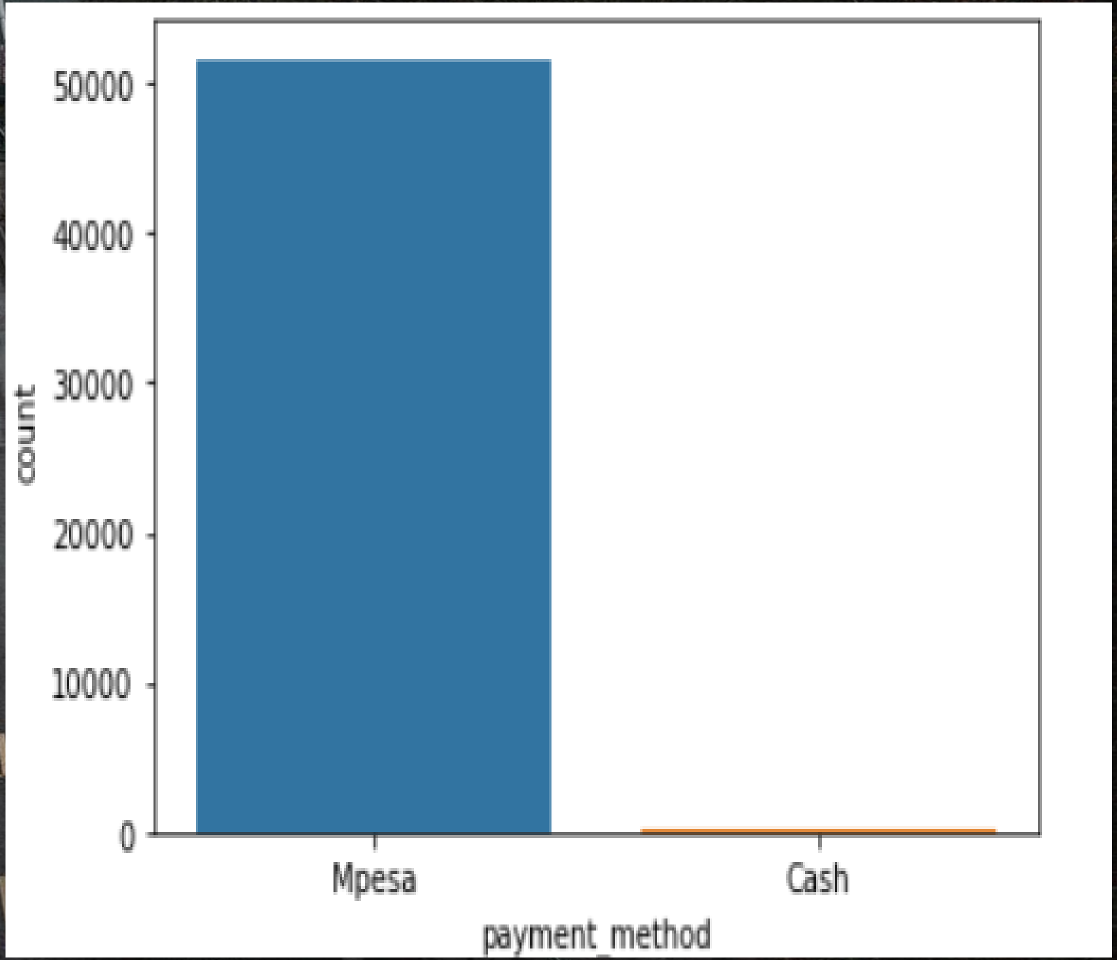
1. DEFINING THE PROBLEM STATEMENT
2. PRE-PROCESSING THE DATA
3. SPLITTING THE DATA INTO TRAIN AND TEST DATA
4. TRAINING THE MODEL
5. EVALUATING THE MODEL
6. IMPROVE THE MODEL
7. DEPLOY THE MODEL AND MONITOR REAL-TIME

EDA(EXPLORATORY DATA ANALYSIS

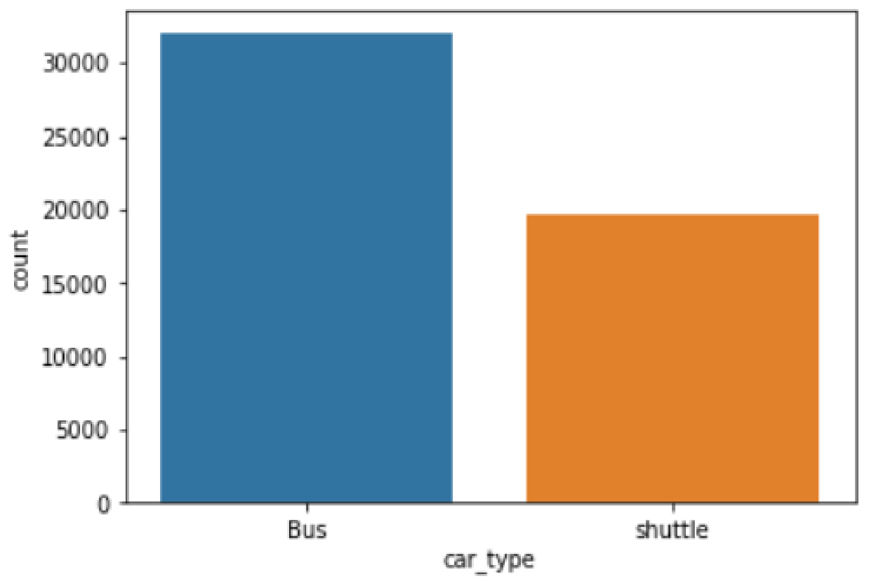
ANALYSING DIFFERENT VARIABLES

PAYMENT METHOD:

There are two types of payment methods people have used To buy the tickets.
Mpesa and cash.
In this bar chart we can see that most of the people have used Mpesa to pay their tickets



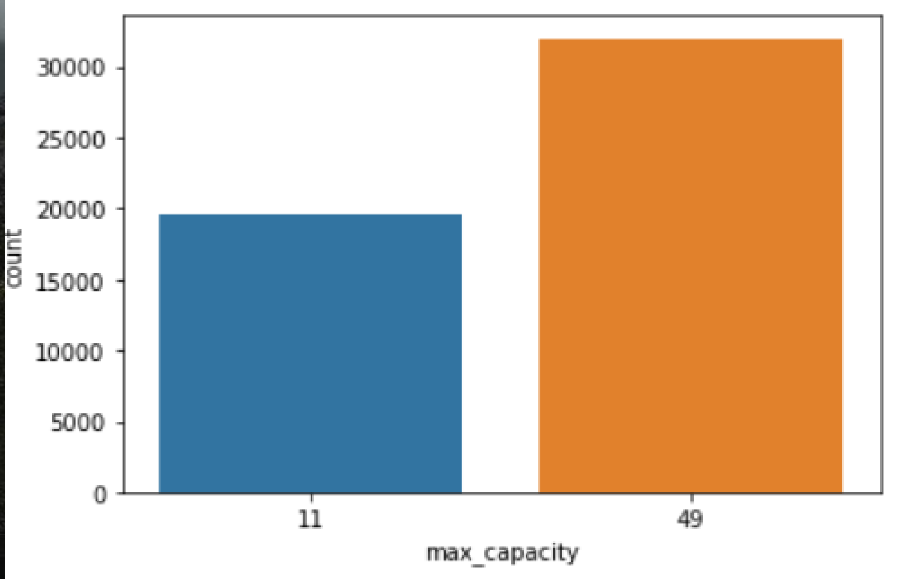
Car type:



There are two different types in car type: (BUS & SHUTTLE).

In the above bar graph we can see that most of them used bus.

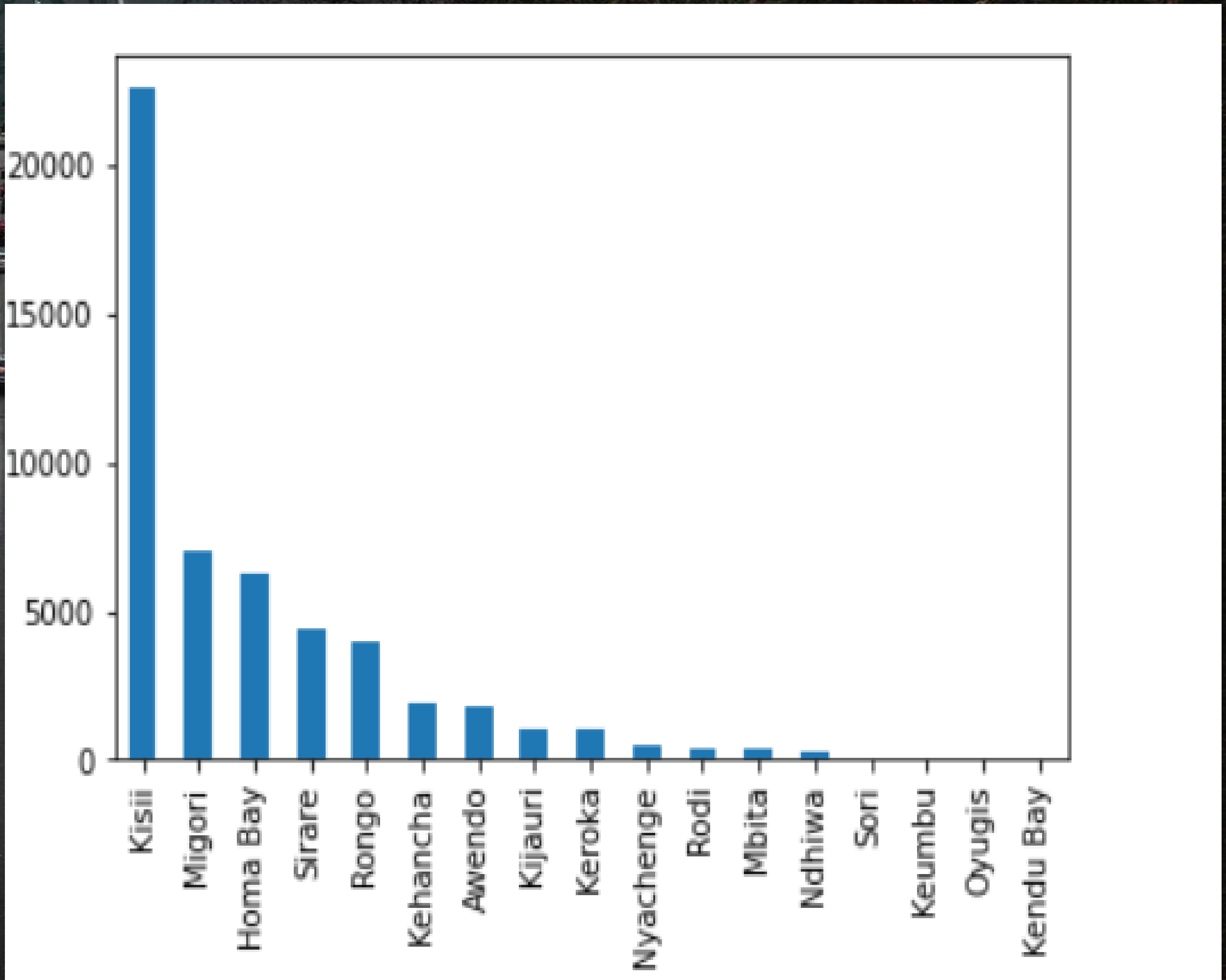
Max capacity:



In the above graph we can see that
Shuttle has maximum capacity =11
And bus has maximum capacity =49

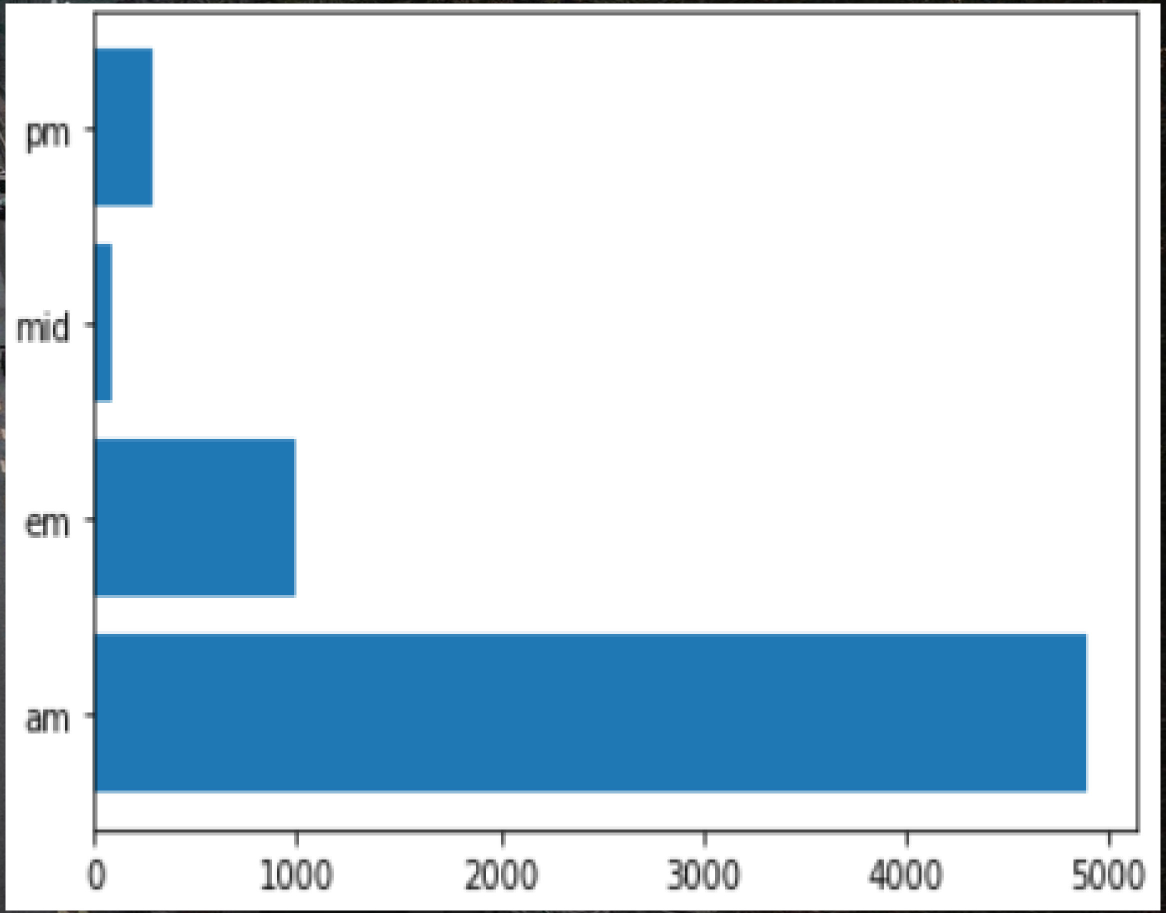
INITIALPOINT OPERATIONS

In this bar plot we can clearly say that most
Of the people travel from kissi.
Less number of people travel from kendu bay .



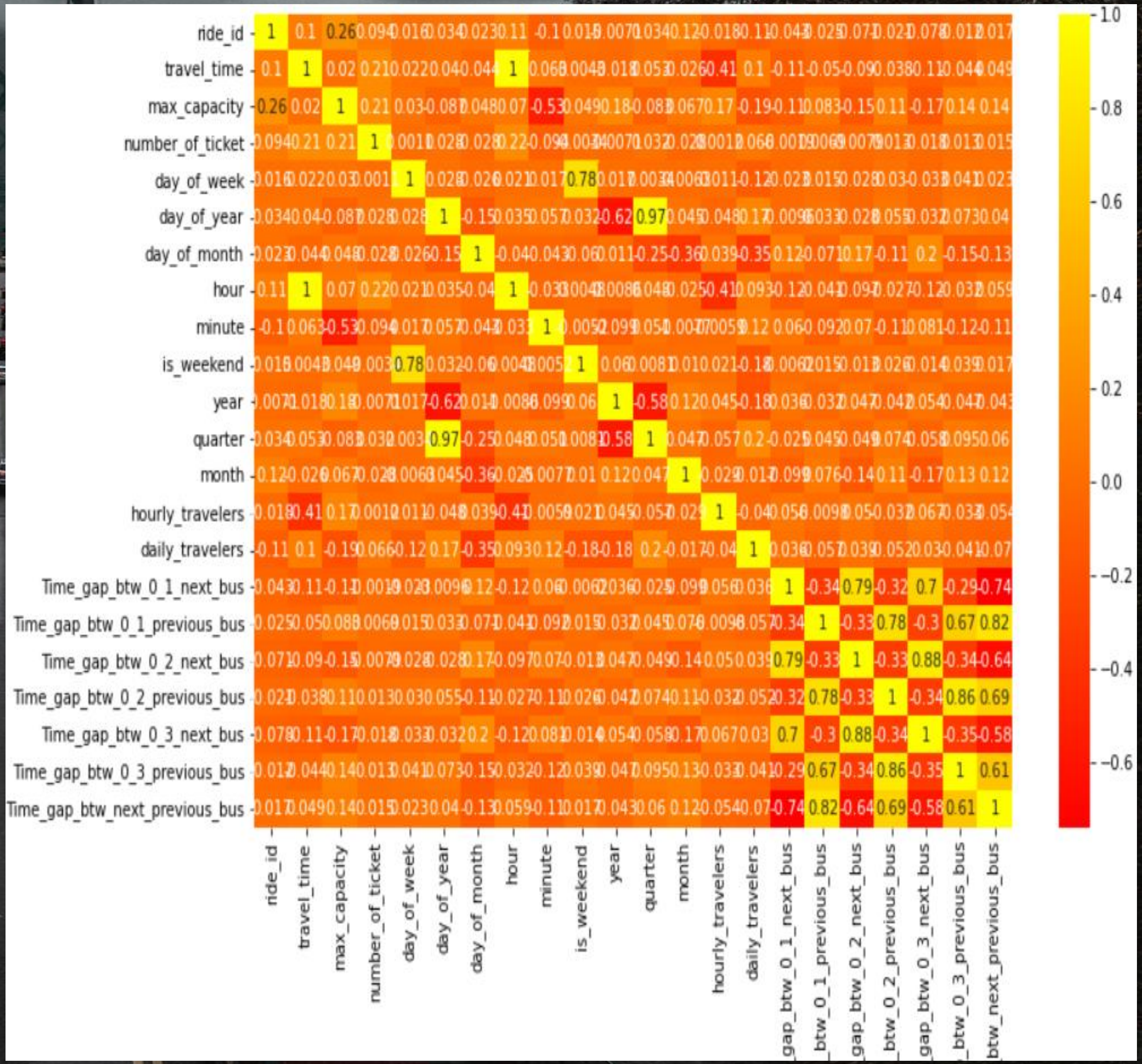
NUMBER OF PEOPLE TRAVELLING IN WHICH TIME ZONE

MOST OF THE PEOPLE USED TO TRAVEL IN THE MORNING TIME, COMPARED TO OTHER TIMES. HERE WE CAN SEE THAT MORE THAN 4500+ PEOPLE TRAVELLED IN THE MORNING.



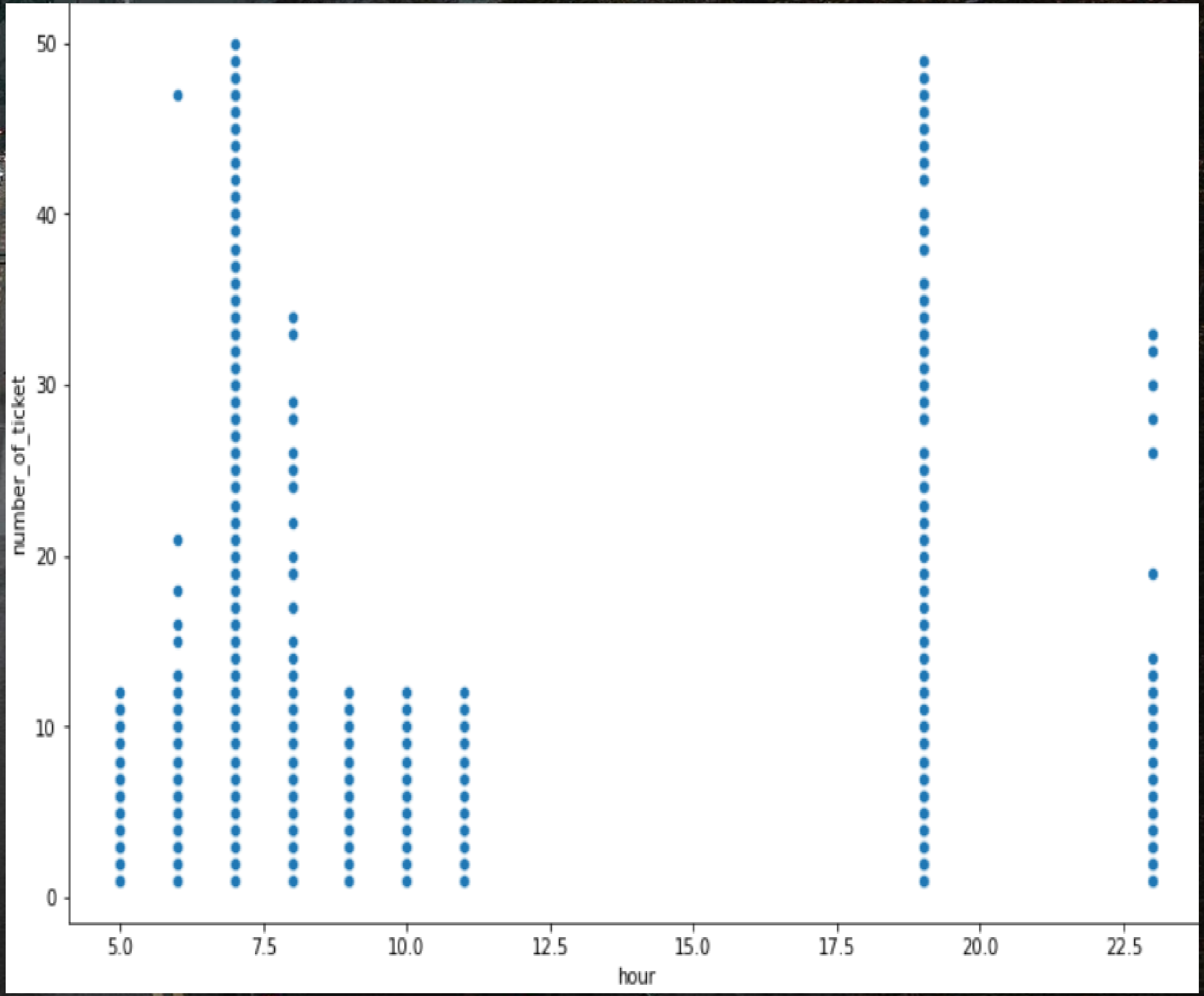
CORRELATION

THIS IS THE 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.



FEATURE ENGINEERING

WE CAN SEE THAT MOST OF THE TICKETS WERE SOLD BETWEEN 7 AM TO 8 AM. AND THAT SEEMS TRUE BECAUSE IN THE MORNING MOST OF THE PEOPLE GO TO WORK FROM THIS CHART WE CAN SAY THAT THERE IS NO RIDE BETWEEN 12PM TO 5.30PM



TRAINING THE MODEL

FOR MODELING WE TRIED VARIOUS REGRESSION ALGORITHMS LIKE:

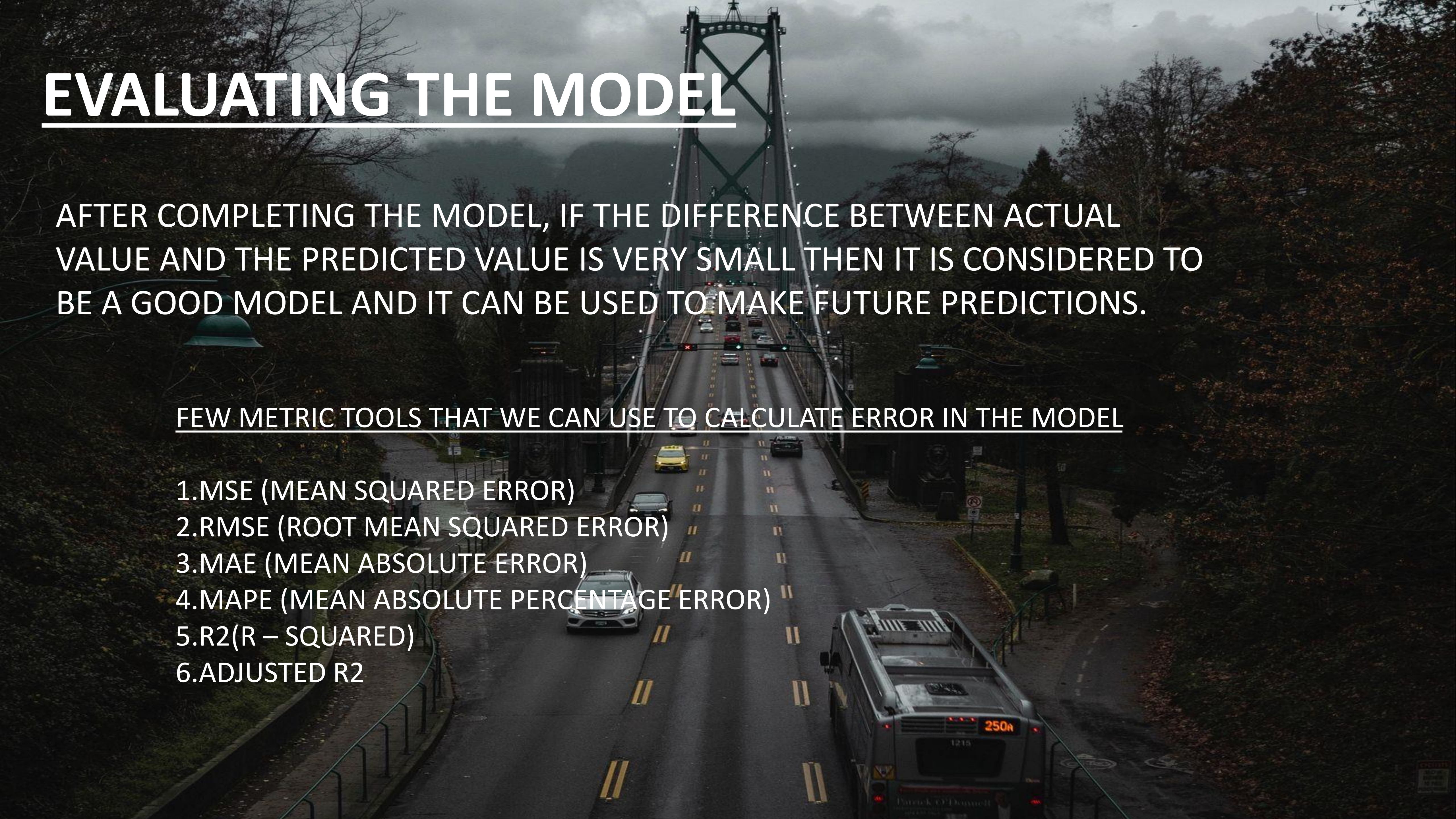
➤ LINEAR REGRESSION

➤REGULARIZED LINEAR REGRESSION (RIDGE AND LASSO)

➤GBM

➤ RANDOM FOREST REGRESSOR

➤XGBOOST REGRESSOR

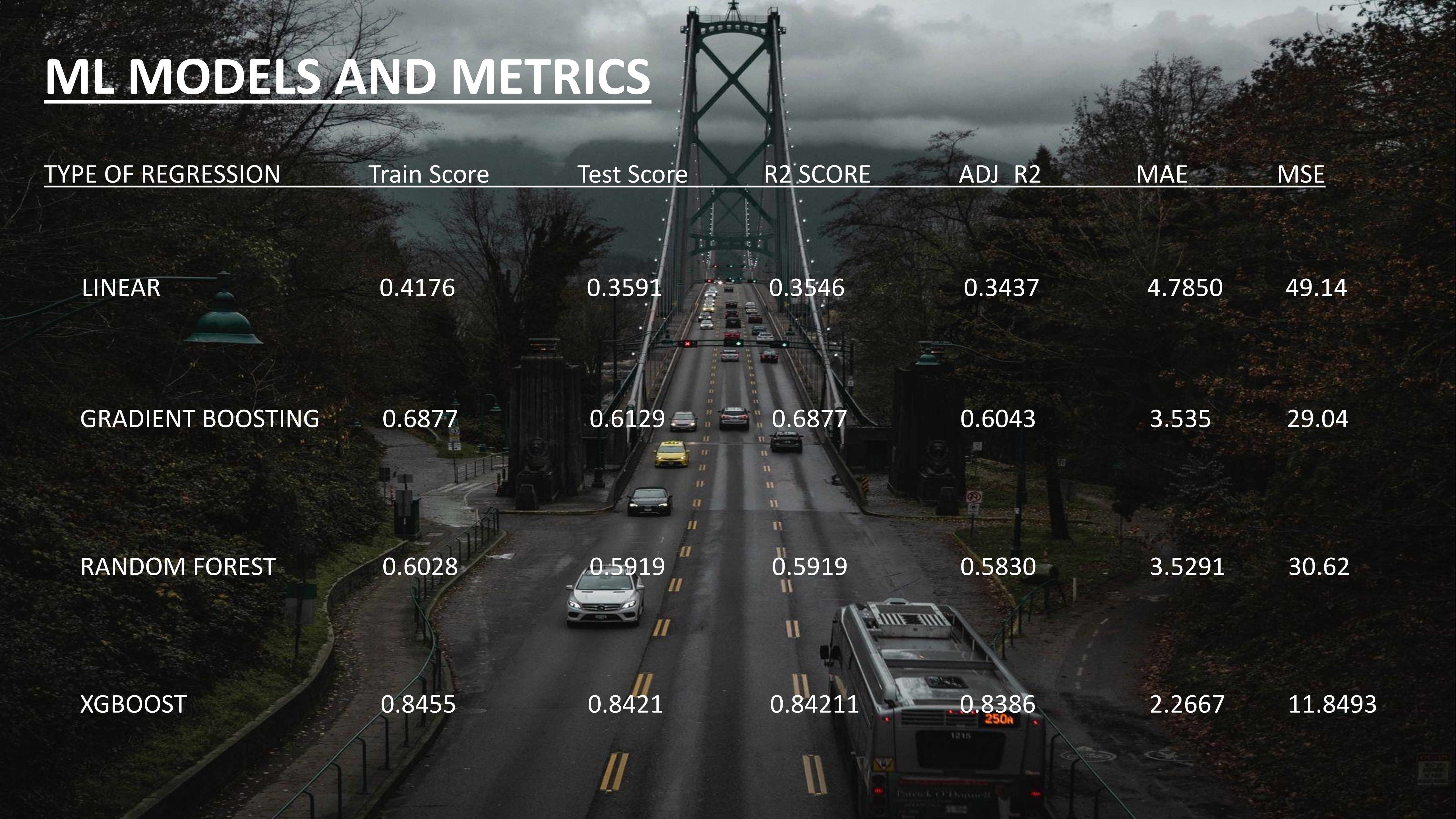


EVALUATING THE MODEL

AFTER COMPLETING THE MODEL, IF THE DIFFERENCE BETWEEN ACTUAL VALUE AND THE PREDICTED VALUE IS VERY SMALL THEN IT IS CONSIDERED TO BE A GOOD MODEL AND IT CAN BE USED TO MAKE FUTURE PREDICTIONS.

FEW METRIC TOOLS THAT 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

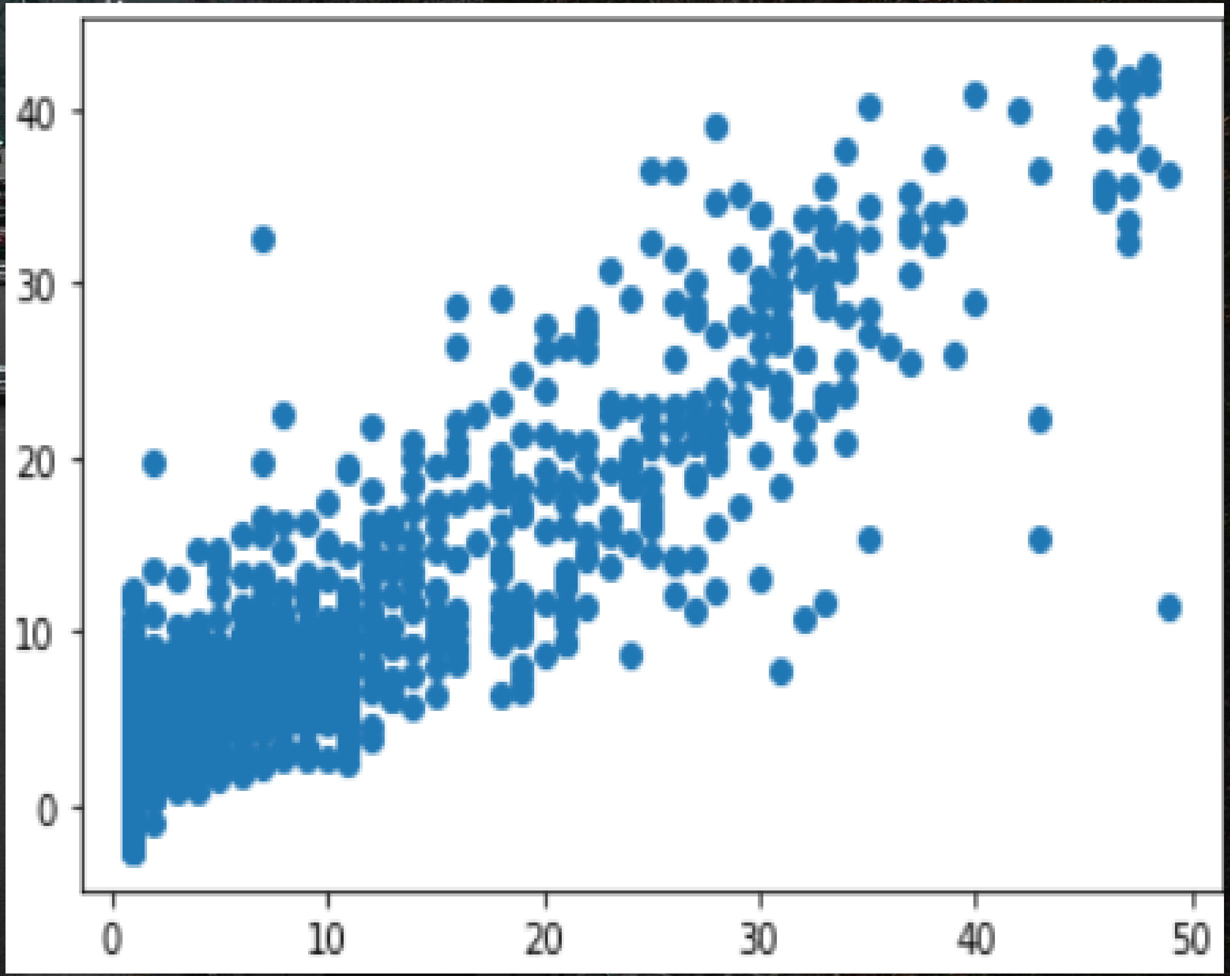


ML MODELS AND METRICS

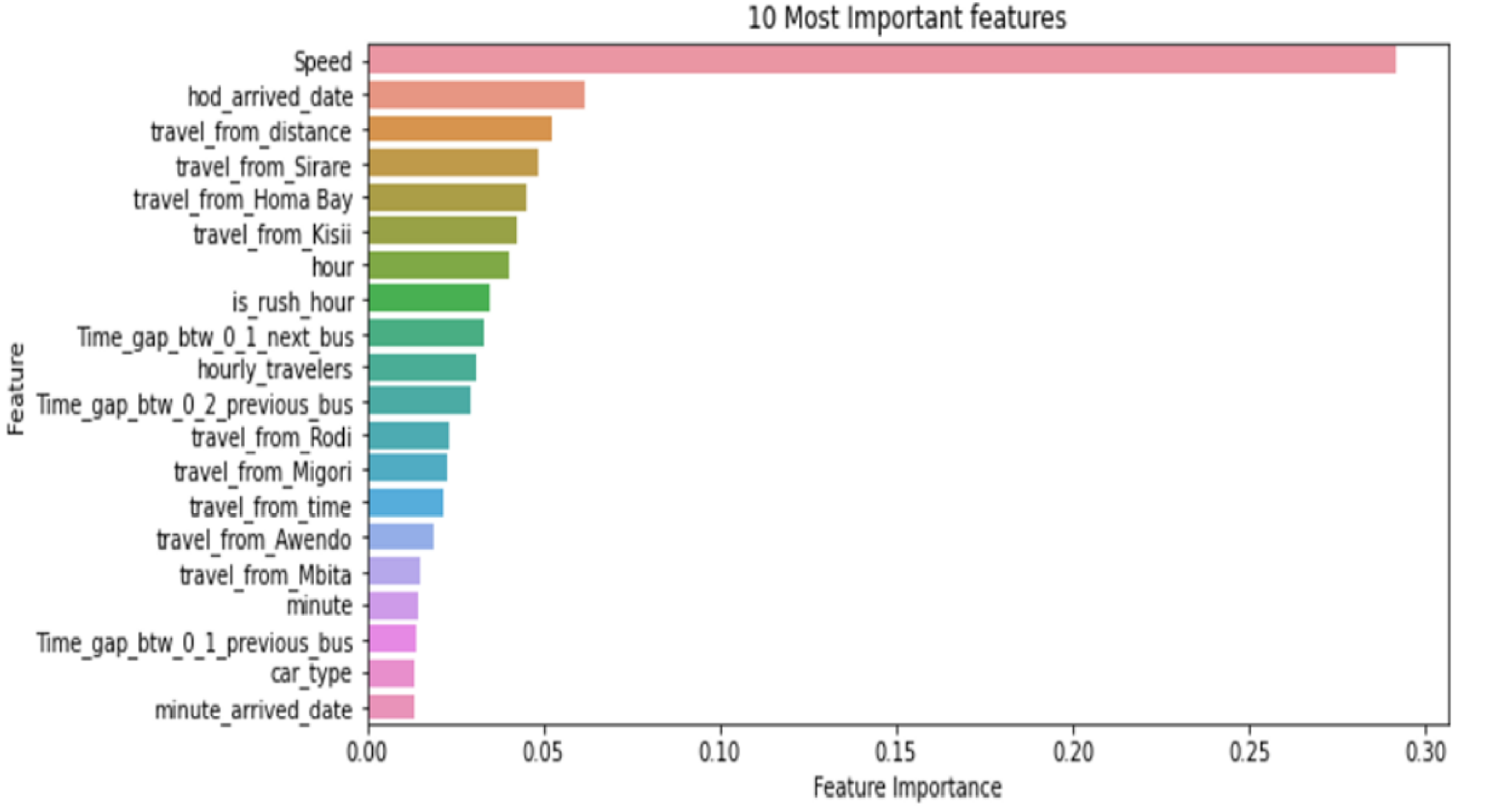
TYPE OF REGRESSION	Train Score	Test Score	R2 SCORE	ADJ R2	MAE	MSE
LINEAR	0.4176	0.3591	0.3546	0.3437	4.7850	49.14
GRADIENT BOOSTING	0.6877	0.6129	0.6877	0.6043	3.535	29.04
RANDOM FOREST	0.6028	0.5919	0.5919	0.5830	3.5291	30.62
XGBOOST	0.8455	0.8421	0.84211	0.8386	2.2667	11.8493

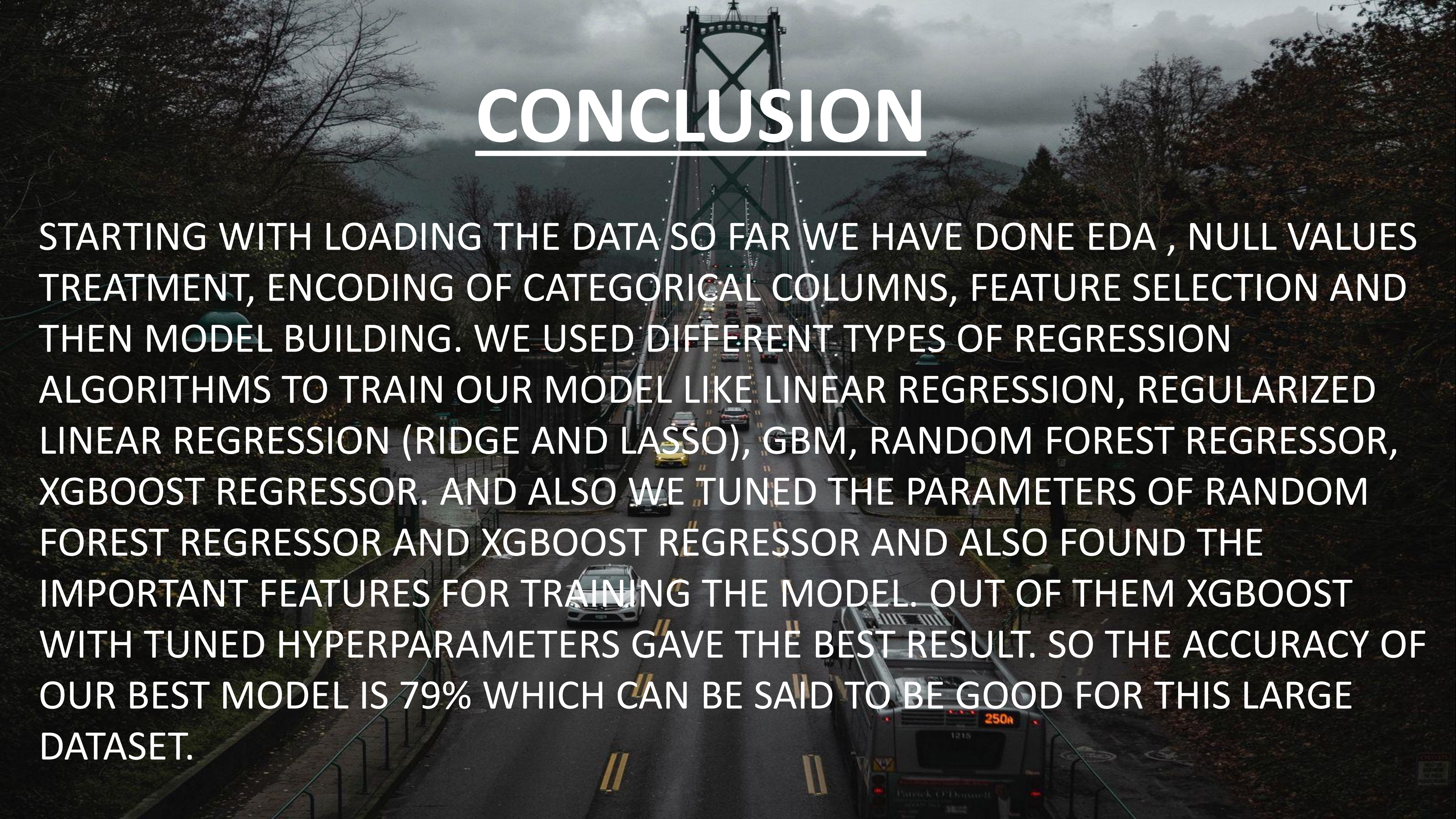
SCATTER PLOT OF TEST AND PREDICTED VALUES

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.



IMPORTANT FEATURES





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 79% WHICH CAN BE SAID TO BE GOOD FOR THIS LARGE DATASET.

THANK
YOU