Capstone Project – 2

Transport Demand Prediction

Presented By :-
Aniket Satpute
Kaiwalya Zankar
(AlmaBetter Trainee)

Content-

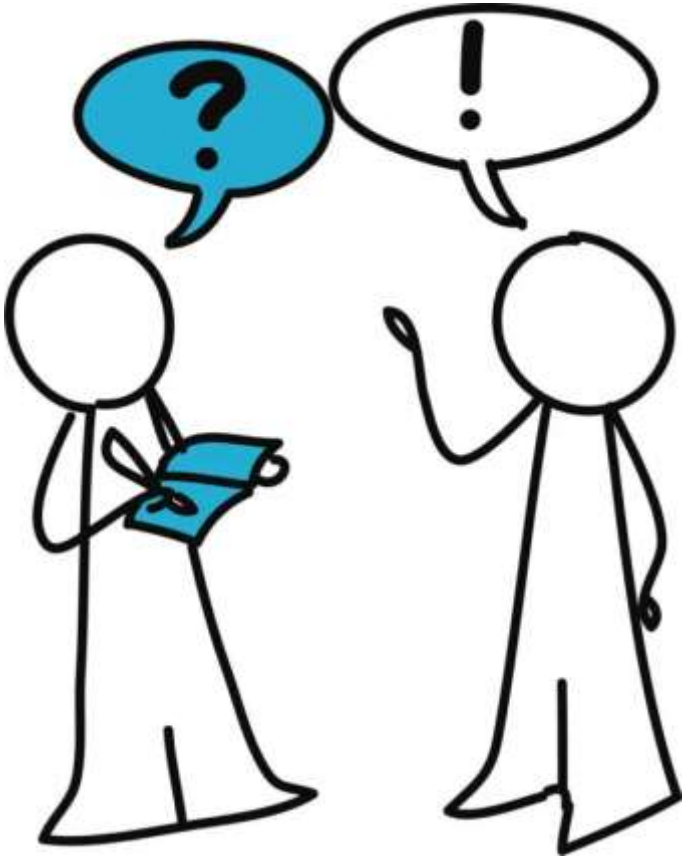
- Introduction
- Scop of project
- Problem statement
- Data summery
- Process flow
- Libraries used
- Data pre procesing
- EDA
- Feature engineering
- ML Model and Evaluation Matrics
- Challenges
- Conclusion
- Q & A

Introduction

- Transport demand forecasting is to predict future transport demand when establishing transport plans within a given budget
- Transport demand is a quantitative input to evaluate supply strategy of transport facilities and land use planning.
- Presented as travel volume based on transport system usage, including transport facilities and transport services.
- The derived demand was created by continuous interaction of transport systems and activity systems

Scope of the Project-

- The scope of project is to create a predictive model using traffic data provided to us and historic bus ticket sales data from Mobiticket to predict the number of tickets that will be sold for buses into Nairobi from cities. The data used to train the model will be historic hourly traffic patterns in Nairobi and historic ticket purchasing data for 14 bus routes into Nairobi from 17 October 2017 and 20 April 2018, and includes the place or origin, the scheduled time of departure, the channel used for the purchase, the type of vehicle, the capacity of the vehicle, and the assigned seat number.



Problem Statement

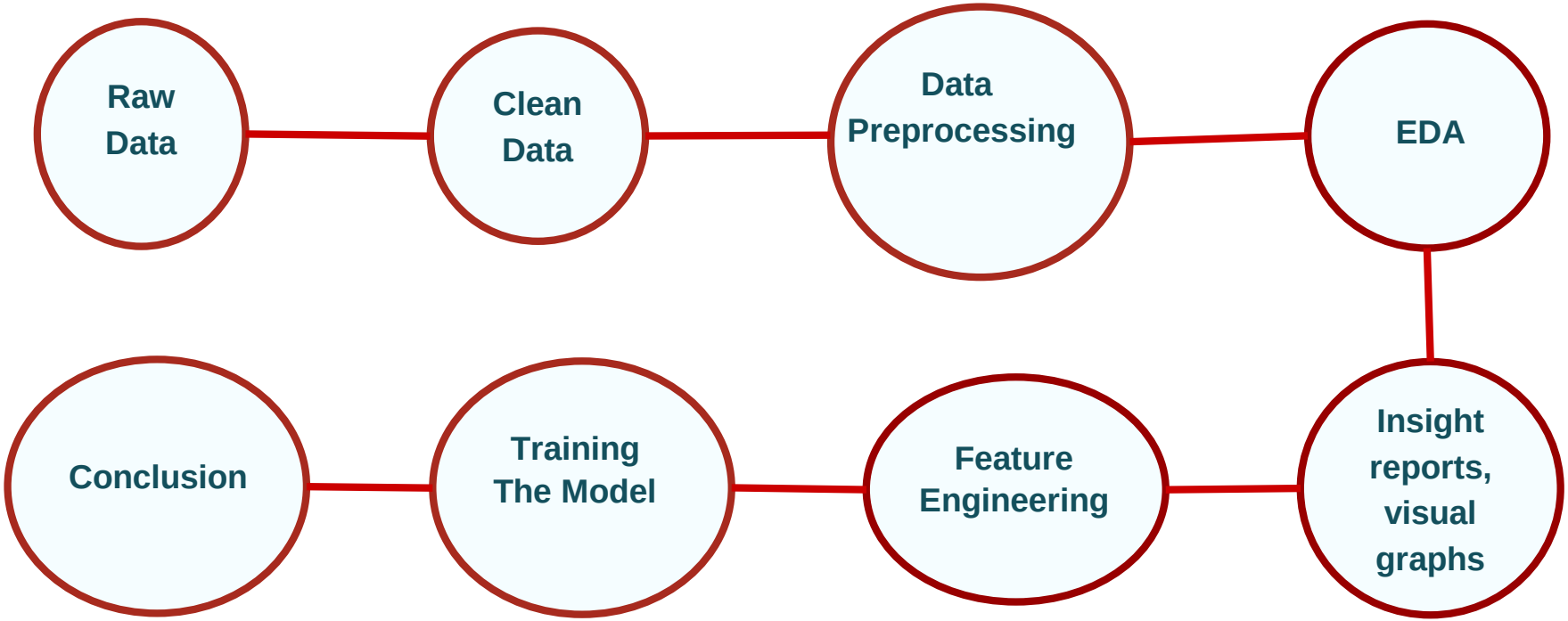
This challenges asks you to build a model that predicts the number of seats that Mobiticket can expect to sell to sell for each ride, i.e. for a specific route on a specific date and time. There are 14 routes in this dataset. All of the routes end in Nairobi and originate in towns to the North-West of Nairobi towards Lake Victoria

Data Summary

This dataset includes the variables from 17 October 2017 to 20 April 2018

- **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 Mobiticket (cash or Mpesa)
- **payment_receipt:** unique id number for ticket purchased from Mobiticket
- **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

Process Flow-



Libraries used-

1. numpy
2. pandas
3. matplotlib.pyplot
4. seaborn
5. datetime
6. Sklearn
7. Math
8. Xgboost
9. warnings

Data Preprocessing

Checking the null values for cleaning the Dataset for further analysis.

ride_id	0
seat_number	0
payment_method	0
payment_receipt	0
travel_date	0
travel_time	0
travel_from	0
travel_to	0
car_type	0
max_capacity	0

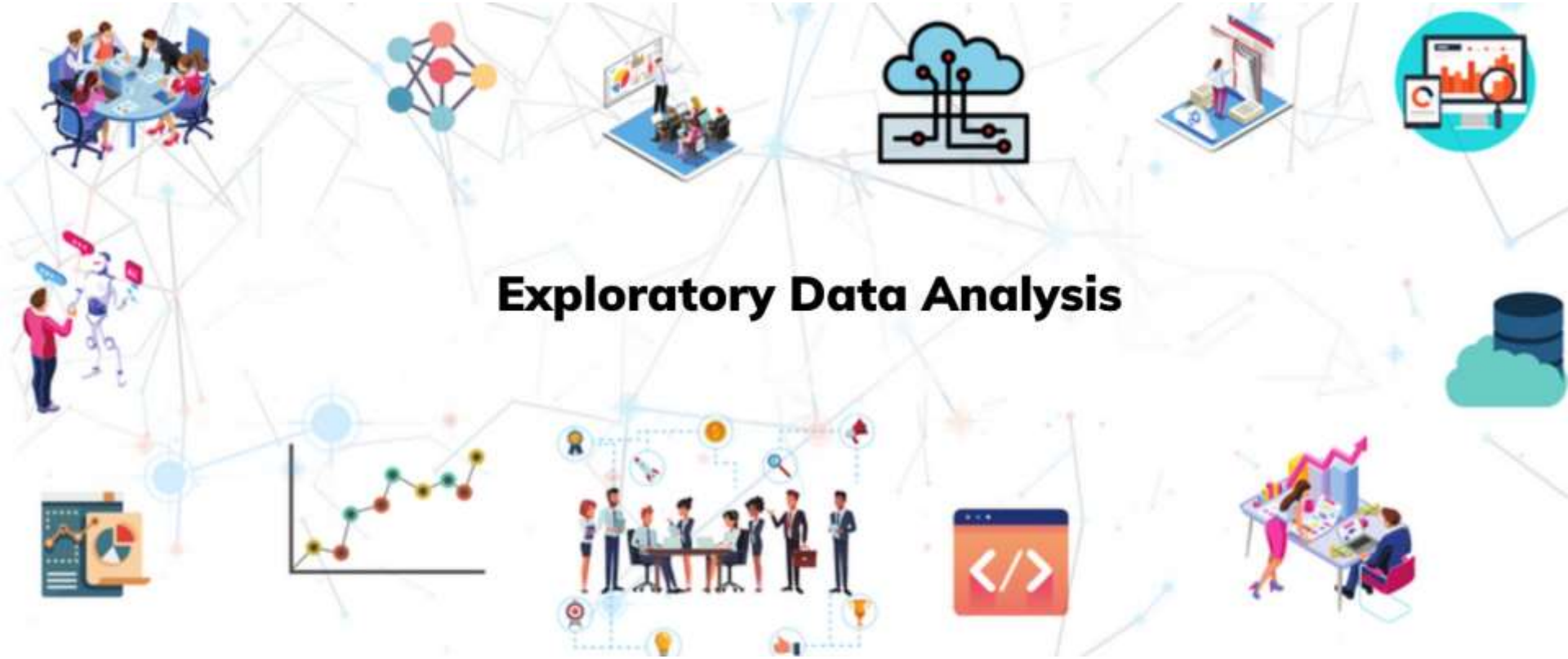
dtype: int64

We do not see any null values in the dataset.

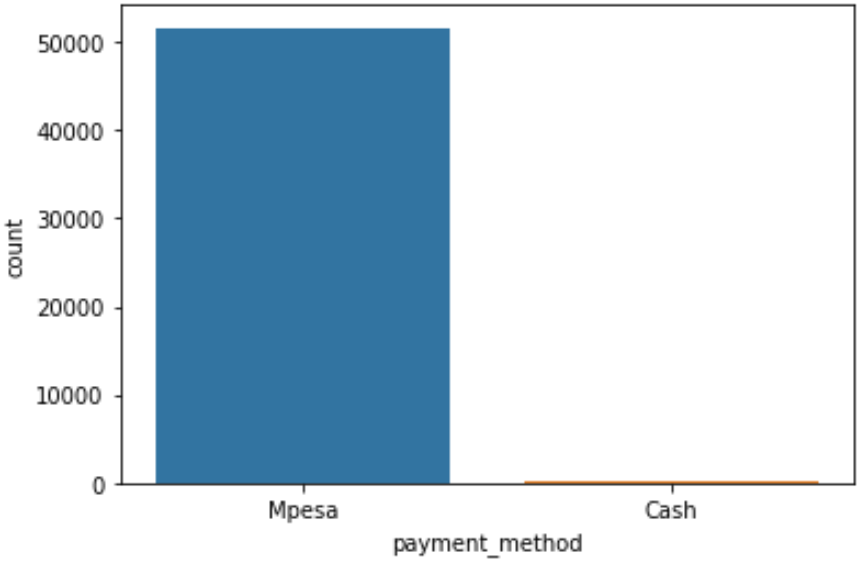
Checking the unique values for Analyzing the dataset for Further analysis.

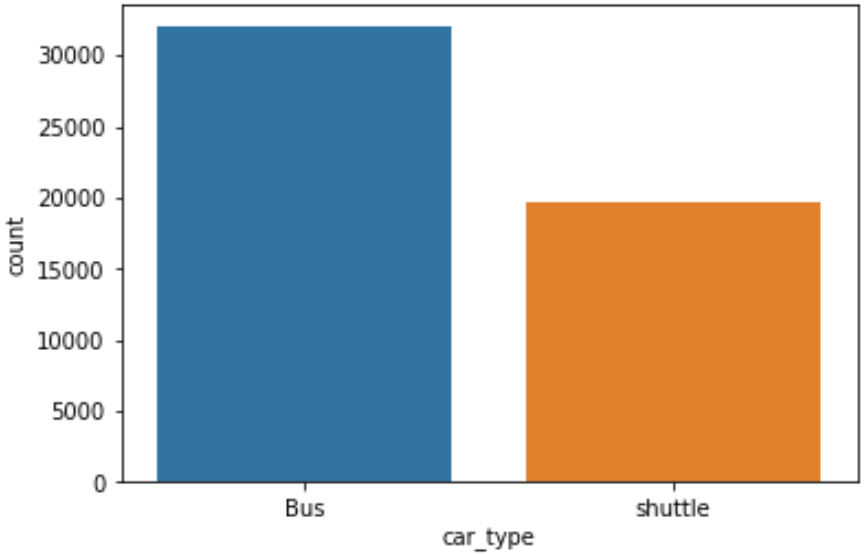
ride_id	6249
seat_number	61
payment_method	2
payment_receipt	51645
travel_date	149
travel_time	78
travel_from	17
travel_to	1
car_type	2
max_capacity	2

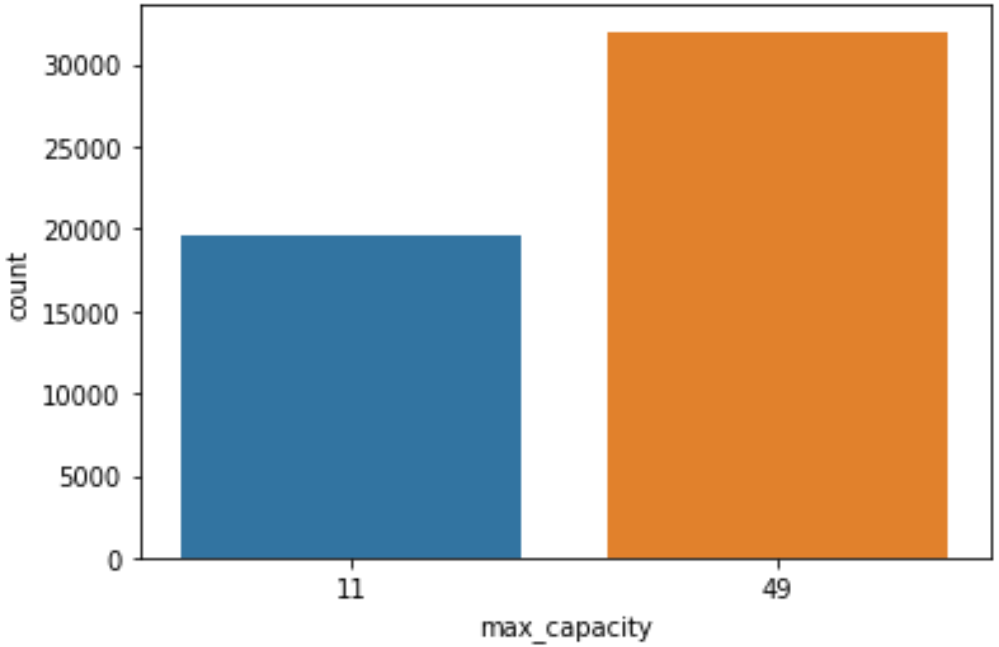
dtype: int64


Exploratory Data Analysis

Values count for payment mode, car type, maximum capacity

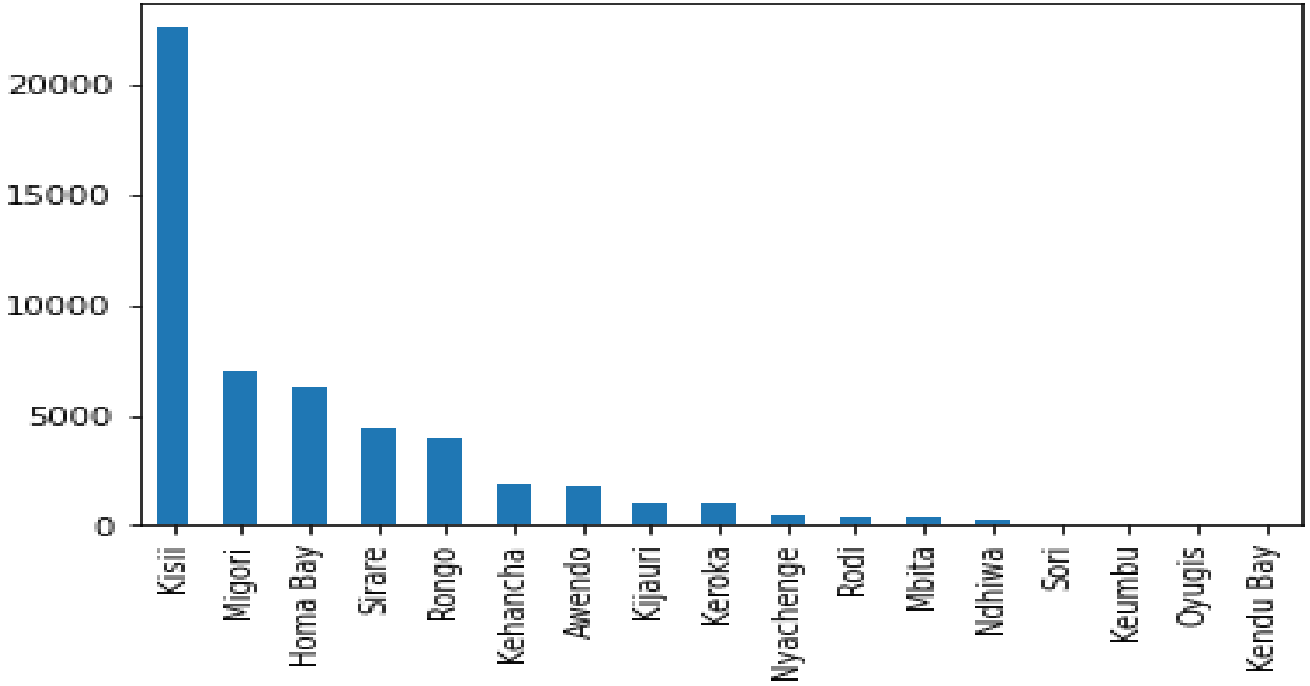






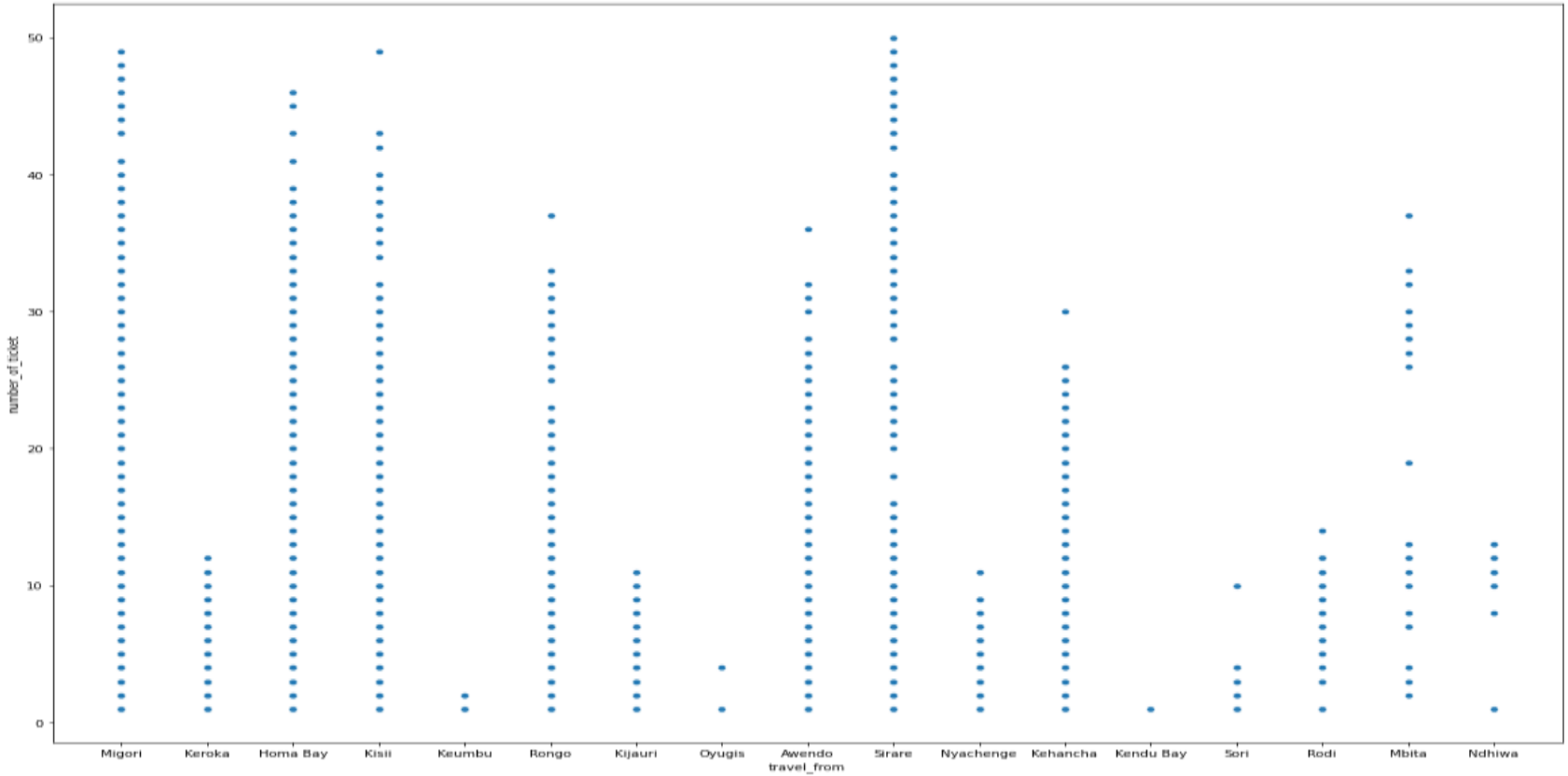
There are two types of payment methods people have used to buy the tickets
There are two types of vehicles bus and shuttle and the maximum capacity of the bus is 49 while shuttle can contain 11 travelers

Towns from which these routes originate



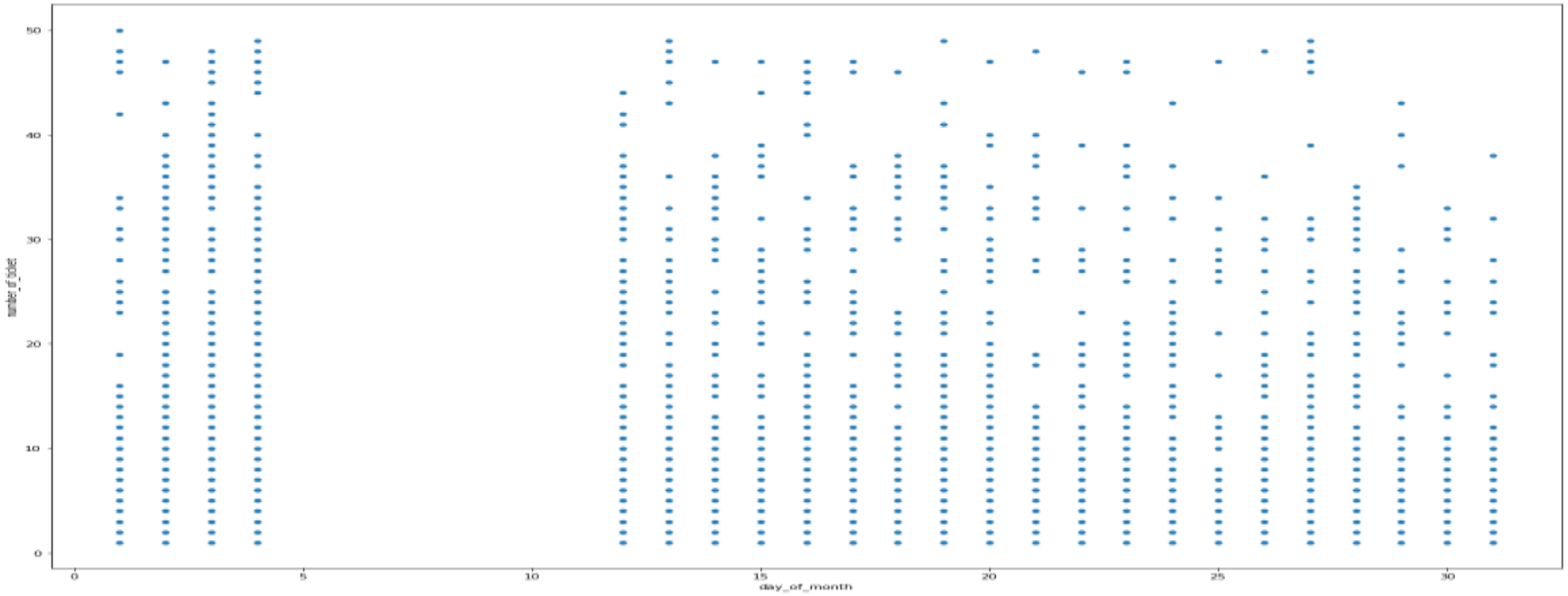
Kisii is the top place from where the more number of rides originate.

travel_from v/s number_of_ticket



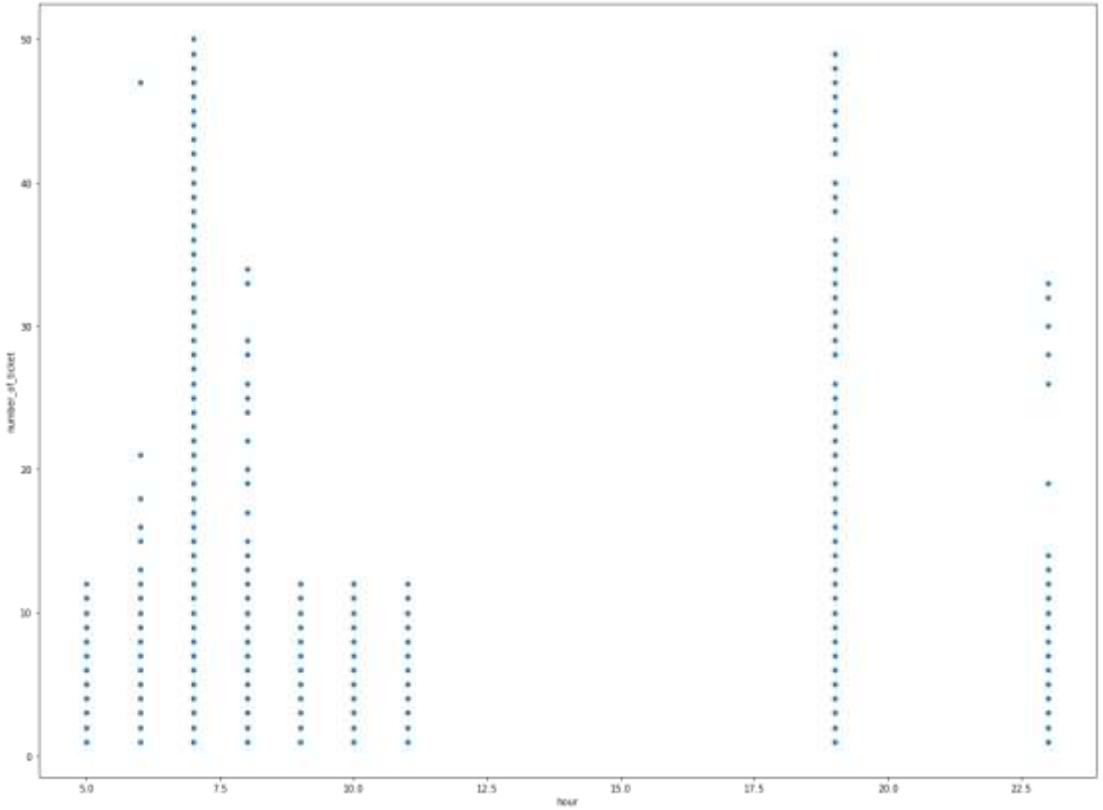
Scatter plot of travel_from by number of tickets

Day_of_month v/s number_of_ticket



We can see that there is the gap between 5-11 in the day of the month.we can assume that there is the holiday of public transport between these days. We can also say that the number of tickets in all the days of month are same.

Hours vs number_of_tickets



We can see that the 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

Feature Engineering

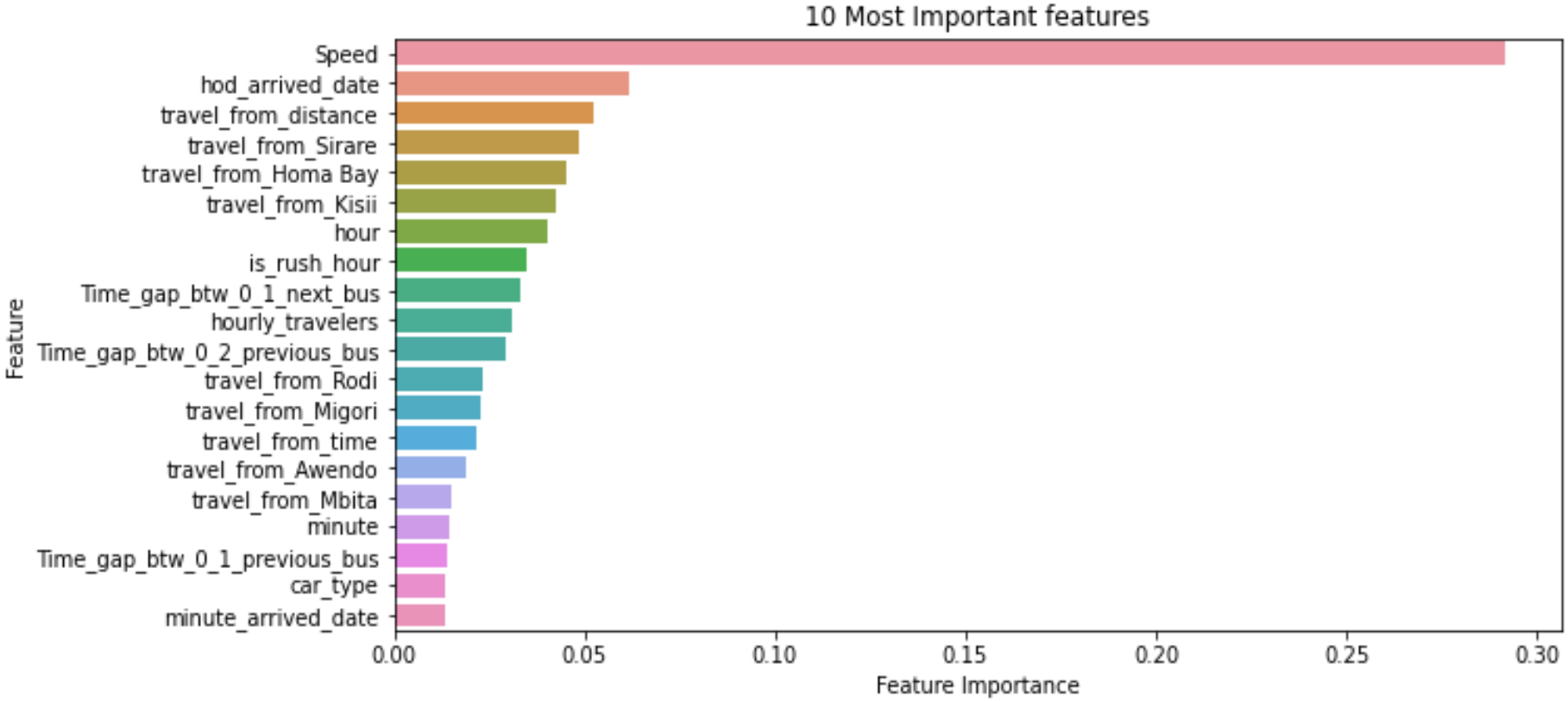
I have added some more features like as follows:

- day_of_week
- Day_of_year
- Day_of_month
- year_woy
- Hour
- minute
- is_weekend
- year
- quarter
- month

ML Models and evaluation Metrics

TYPE OF REGRESSION	Train Score	Test Score	R2 SCORE	ADJ_R2	MAE	MSE
LINEAR	0.41531	0.35462	0.35467	0.34765	4.74747	48.435119
LINEAR-LASSO	0.41406	0.35476	0.35476	0.347804	4.74177	48.42415
LINEAR-RIDGE	0.4302051	0.4838009	0.35535	0.34129	4.74177	48.42415
GRADIENT BOOSTING	0.67633	0.608508	0.608508	0.60467	3.54003	29.39045
RANDOM FOREST	0.62769	0.62338	0.623387	0.615172	3.375124	5.31642
XGBOOST	0.845594	0.842112	0.842112	0.838668	2.2667203	11.8493008

Feature Importance



Challenges

- Feature engineering – to get the more required features that will ease the further analysis
- What should be the dependent variables
- To filter the given data
- Feature to be selected to get the required output

Conclusion

We used different type of regression algorithms to train our model like, linear regression, regularized linear regression (Ridge and lasso), GBM, Random Forest Regressor, XG Boost regressor, and also we tuned the parameters of Random forest regressor and XG Boost regressor and also found the important features for training the model. Out of them XG Boost with tuned hyperparameters gave the best result.

Q & A