Capstone Project-2

Transport Demand Prediction

Team members

KAMALUDDIN SHAIKH

AMOL RASAM

SHUBHAM JHA

PRETESH AGARWAL

Content

- Problem Statement
- Data Summary
- Dependant Variable
- Ride Origination Towns
- Top 5 place of Origin
- Day wise travel trend
- Hourly travel trend
- Time taken and Distance trend
- Feature Engineering
- ML Models and Metrics
- Feature Importance
- Challenges
- Conclusion

Problem Statement

- Exploring 14 different towns to the northwest of the Nairobi towards Lake Victoria and using the data provided by bus ticket sales from MobiTicket, predicting the number of tickets that would be sold for the buses that ends into Nairobi.
- Build a model that predicts the number of seats that Mobiticket can expect to sell for each ride, i.e. for a specific route on a specific date and time.

Data Summary

Nairobi Transport Data is the dataset of tickets purchased from Mobiticket for the 14 routes from “up country” into Nairobi between 17 October 2017 and 20 April 2018.

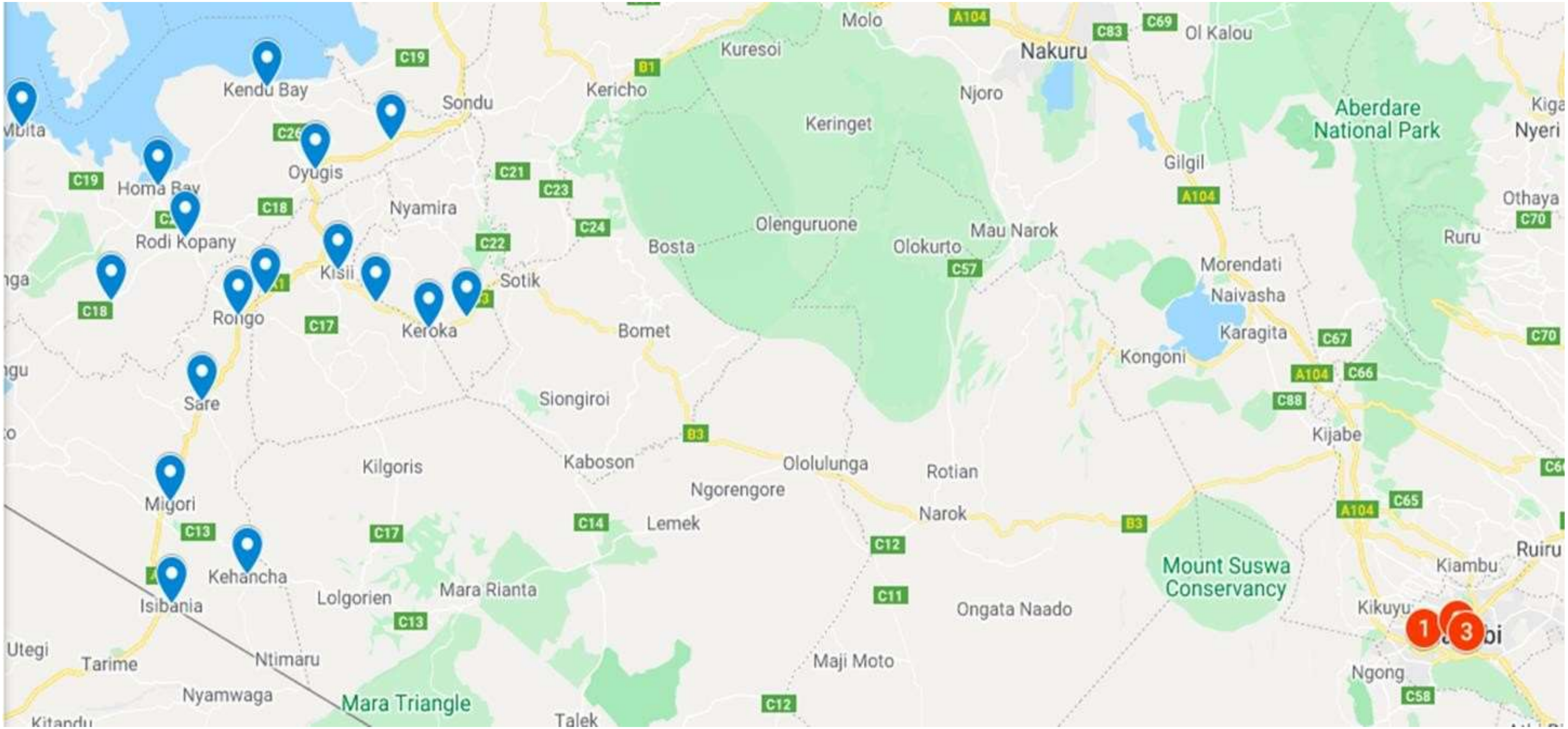
- **ride_id:** unique ID of a vehicle on a specific route on a specific day and time.
- **seat_number:** seat assigned to ticket.
- **payment_method:** method used by customer to purchase ticket from Mobiticket (cash or Mpesa).
- **payment_receipt:** unique id number for ticket purchased from Mobiticket.

- **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.

Dependant Variable

- In our problem statement we didn’t have Dependant Variable. There are many ways to find our Dependant variable. But we used ride_id which is unique ID of a vehicle on a specific route on a specific day and time and seat_number which is seat assigned to ticket to find our Dependant variable.
- By using Groupby method on ride_id and seat_number we got our Dependant Variable.

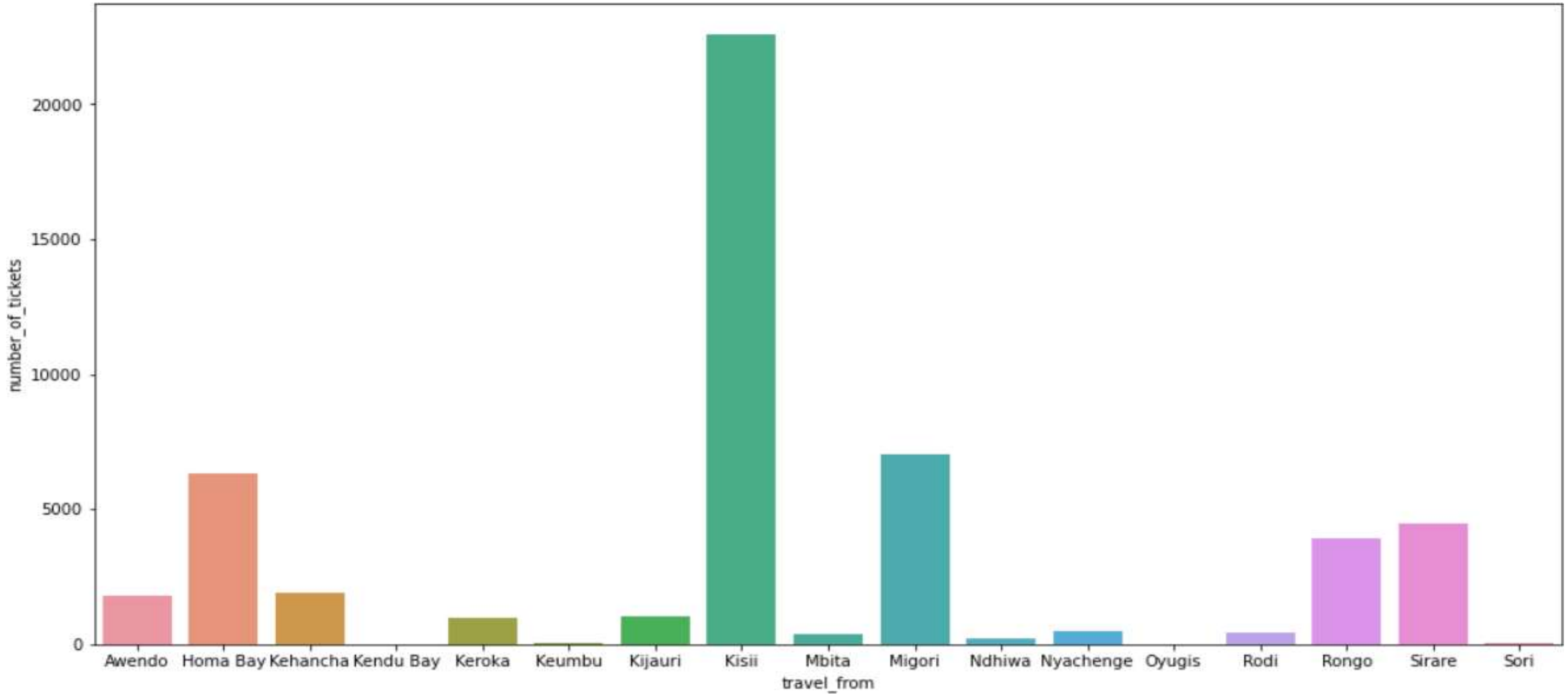
Map


Kendu Bay
Mbita
Sondu
Kericho
Kuresoi
Molo
Nakuru
Ol Kalou
Njoro
Keringet
Aberdare National Park
Nyeri
Gilgil
Oyugis
Homa Bay
Nyamira
Othaya
Rodi Kopany
Olenguruone
Mau Narok
Olokurto
Bosta
Ruru
Kisii
Sotik
Morendati
Naivasha
Rongo
Keroka
Bomet
Karagita
Kongoni
Sare
Siongiroi
Kijabe
Kilgoris
Kaboson
Ololulunga
Rotian
Migori
Ngorengore
Lemek
Narok
Ruiru
Kiambu
Mount Suswa Conservancy
Kehancha
Isibania
Lolgorien
Mara Rianta
Ongata Naado
Kikuyu
Utegi
Tarime
Ntimaru
Maji Moto
Ngong
Nyamwaga
Mara Triangle
Talek
Kitandu
1
3

EDA

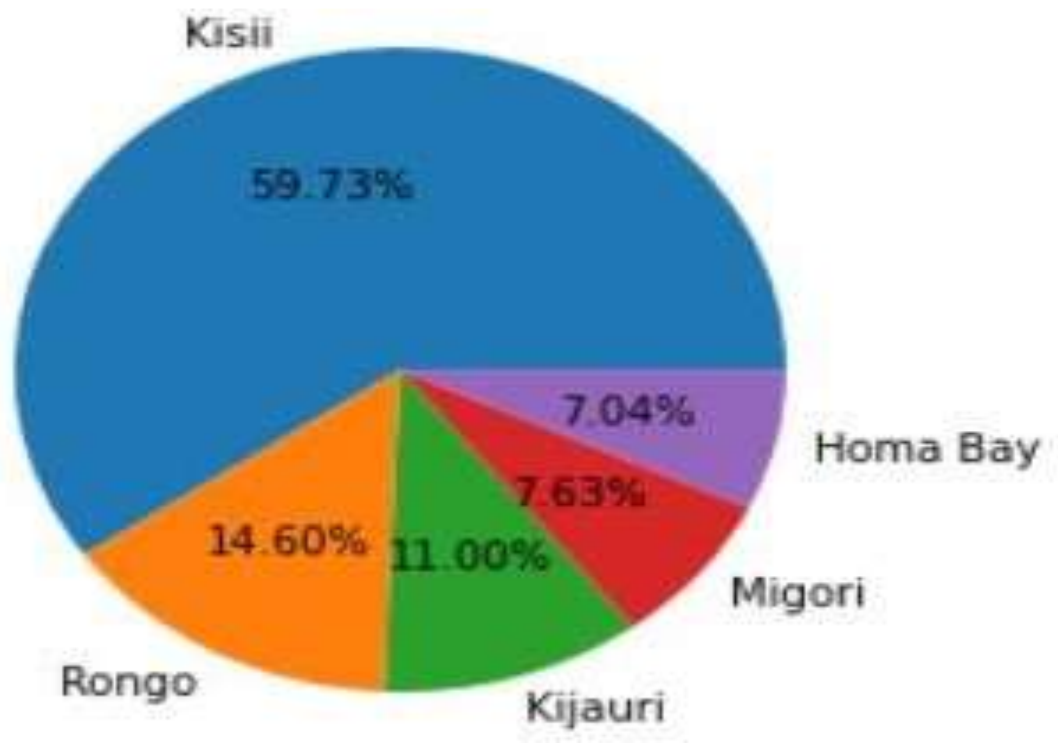
- Ride Origination Towns.
- Top 5 places where most people are coming from.
- Scatter plot of no_of_tickets from origination towns.
- Day wise trend of no of tickets.
- Hourly trend of no of tickets.
- Time taken and Distance relation with no of tickets.

Ride Origin Towns



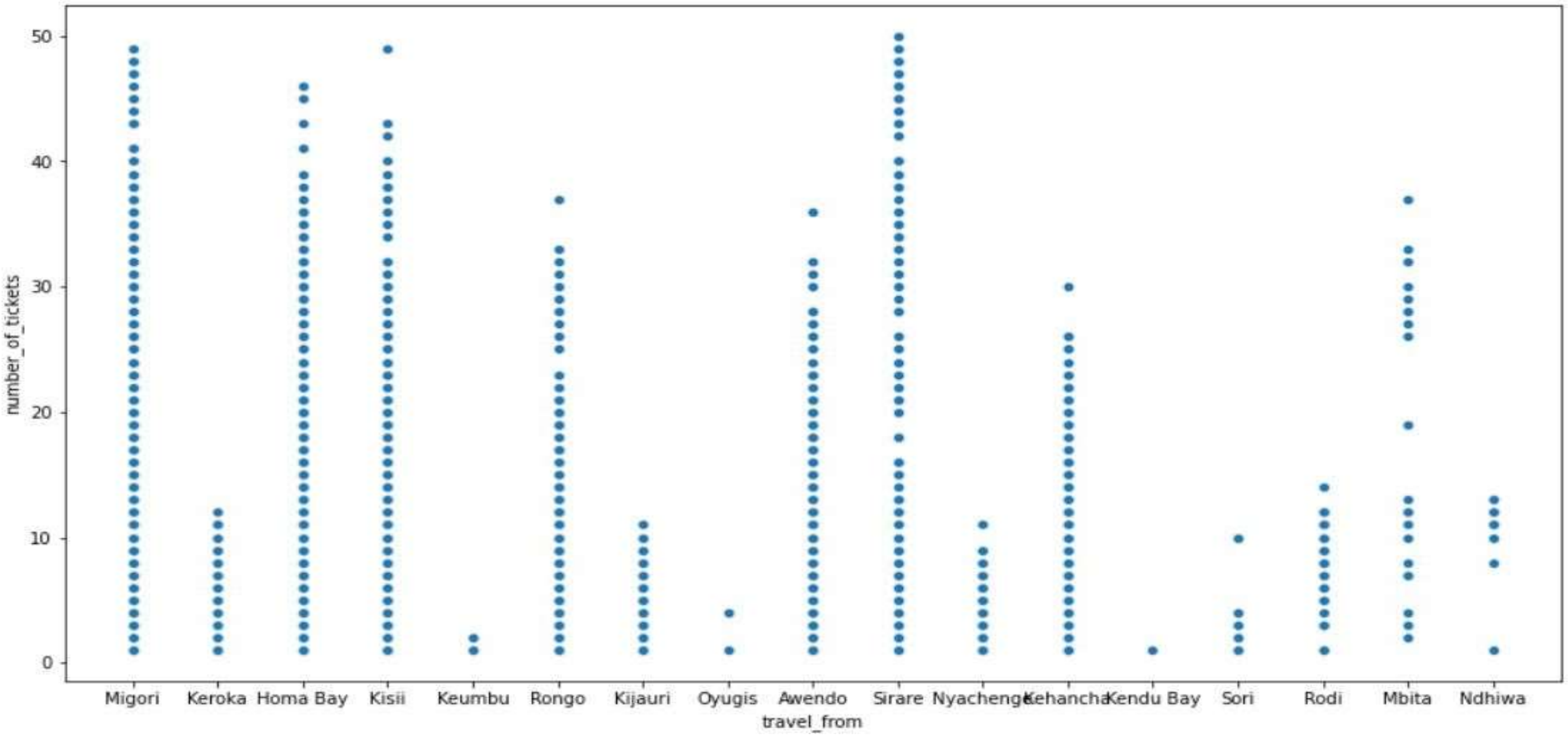
Kisii is the top place from where the most number of rides originate.

Top 5 places of Origin.



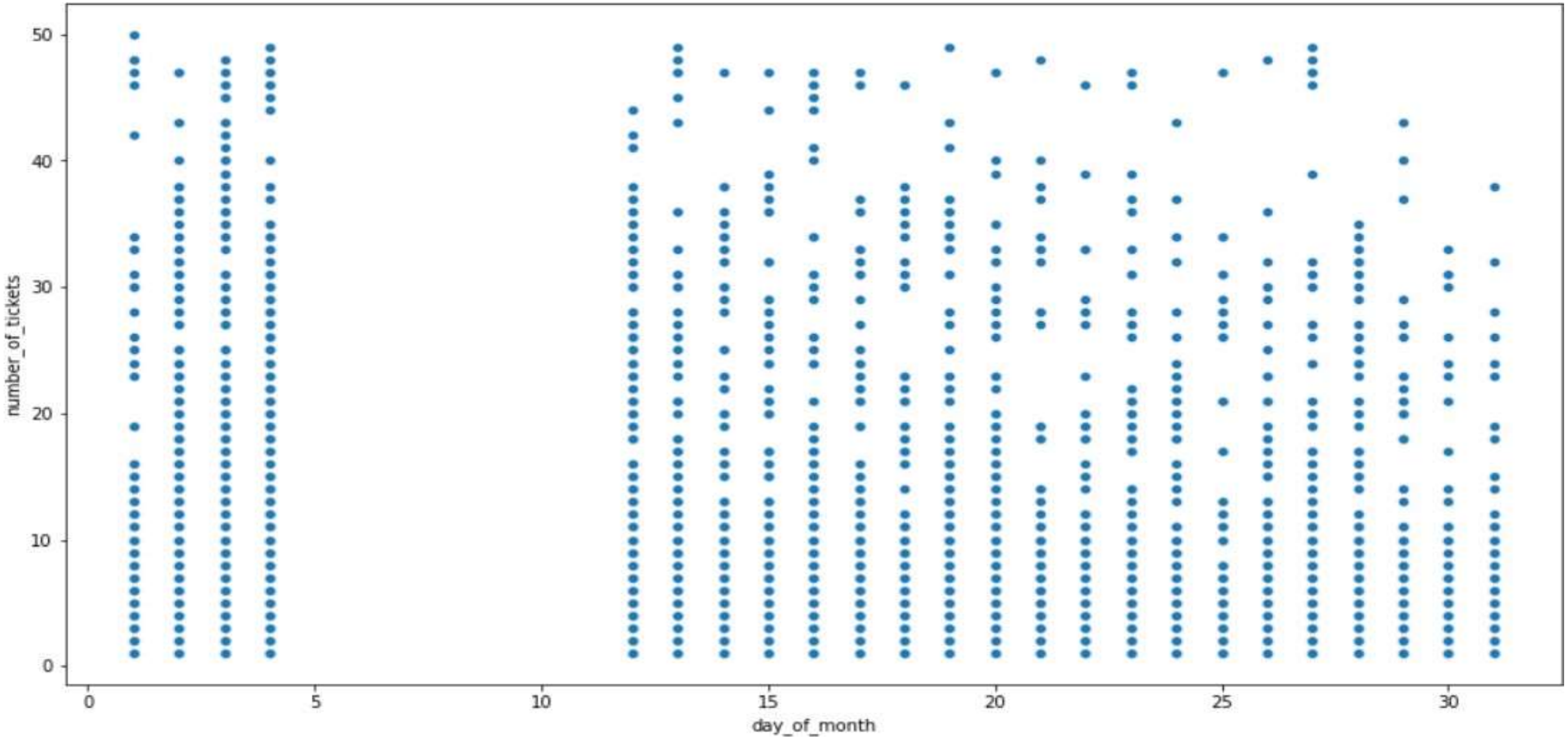
Most people are travelling from Kisii 59.73%, followed by Rango- 14.60%, kijauri-11%, Migori-7.63%, Homa Bay-7.04% etc.

Scatter plot



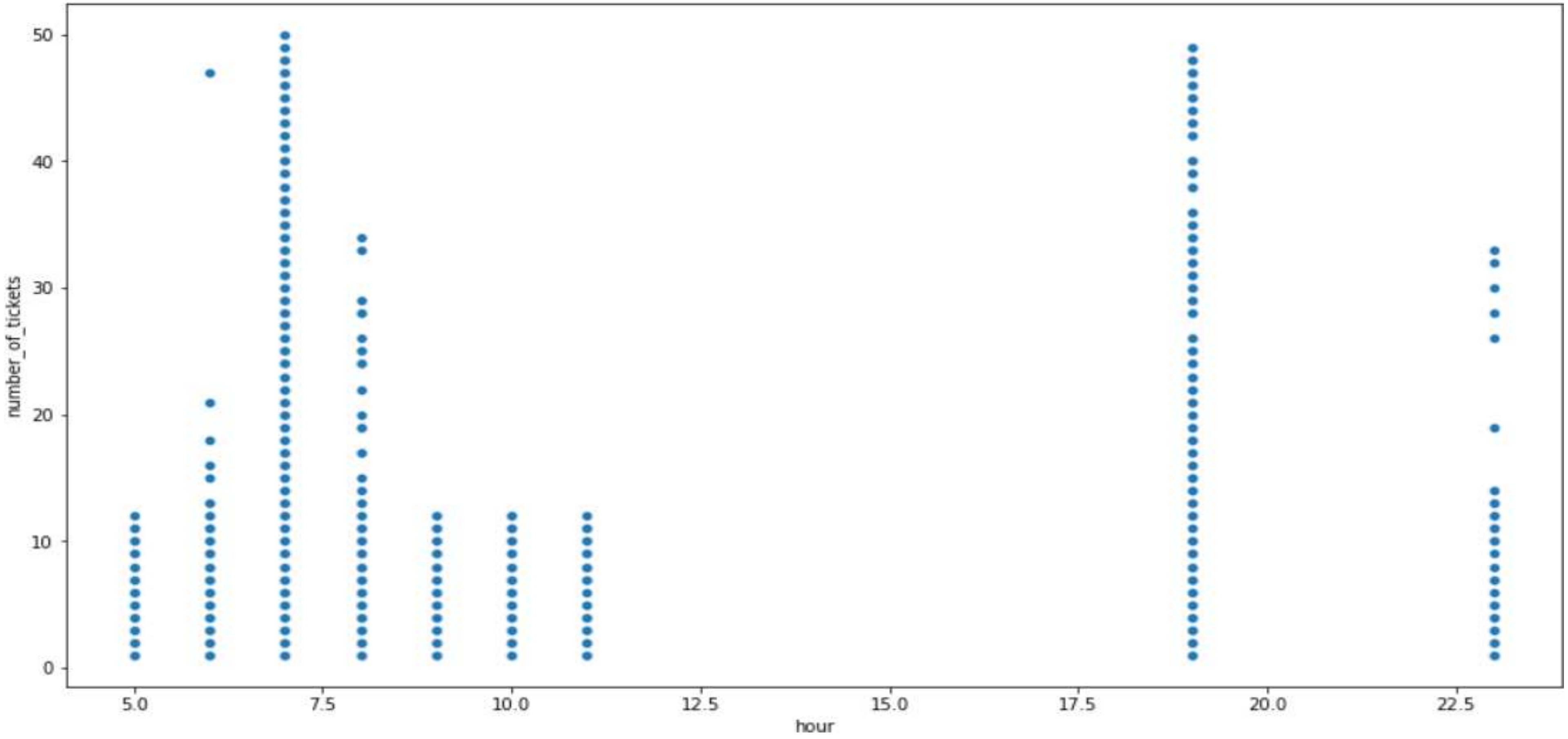
Scatter plot of travel_from by number of tickets.

Day wise Travel Trend



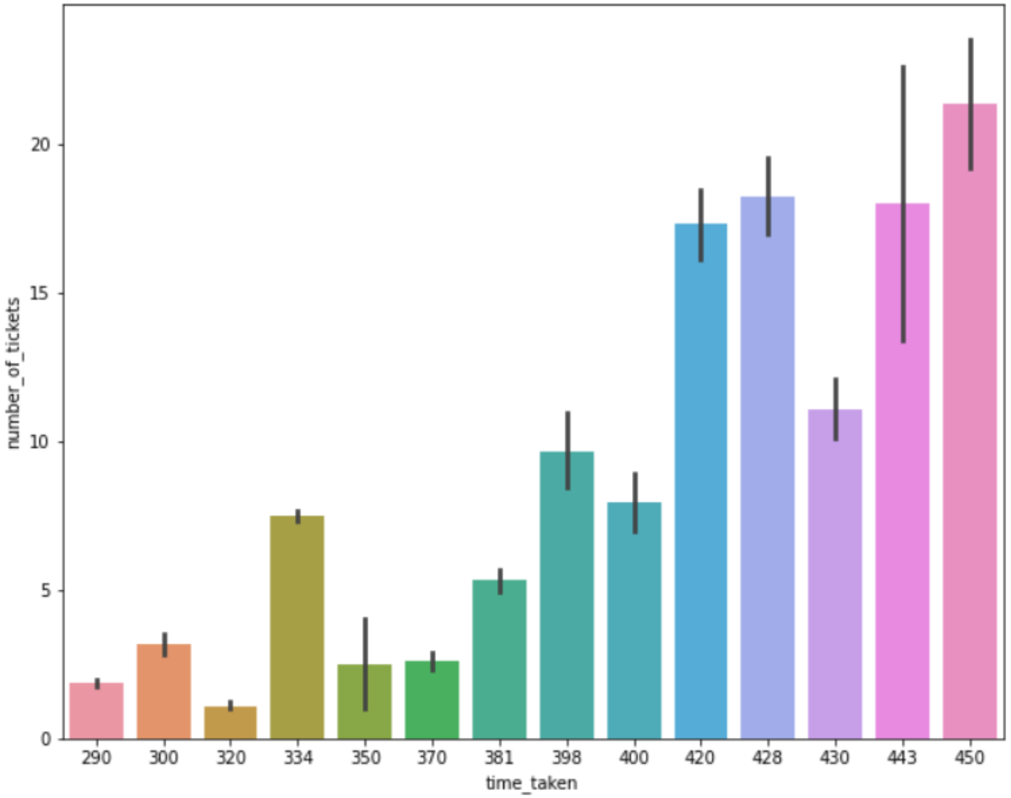
There are no rides between 5th to 10th of every month, but this might be because of missing data or some public Holiday.

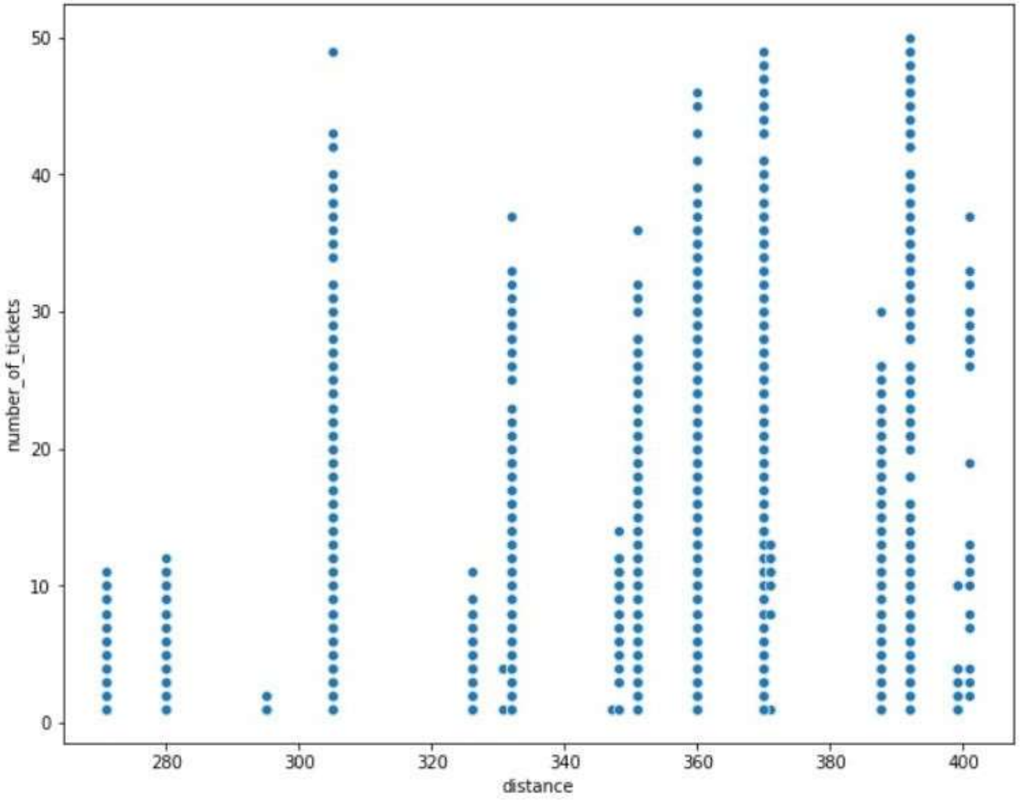
Hourly Travel Trend



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.

Time taken and Distance Trend





- Time taken to reach Destination have positive relation with number of tickets. distance also have some what positive relation with no of tickets.
- No of ticket sold is increasing with increase in time taken and distance.

Feature Engineering

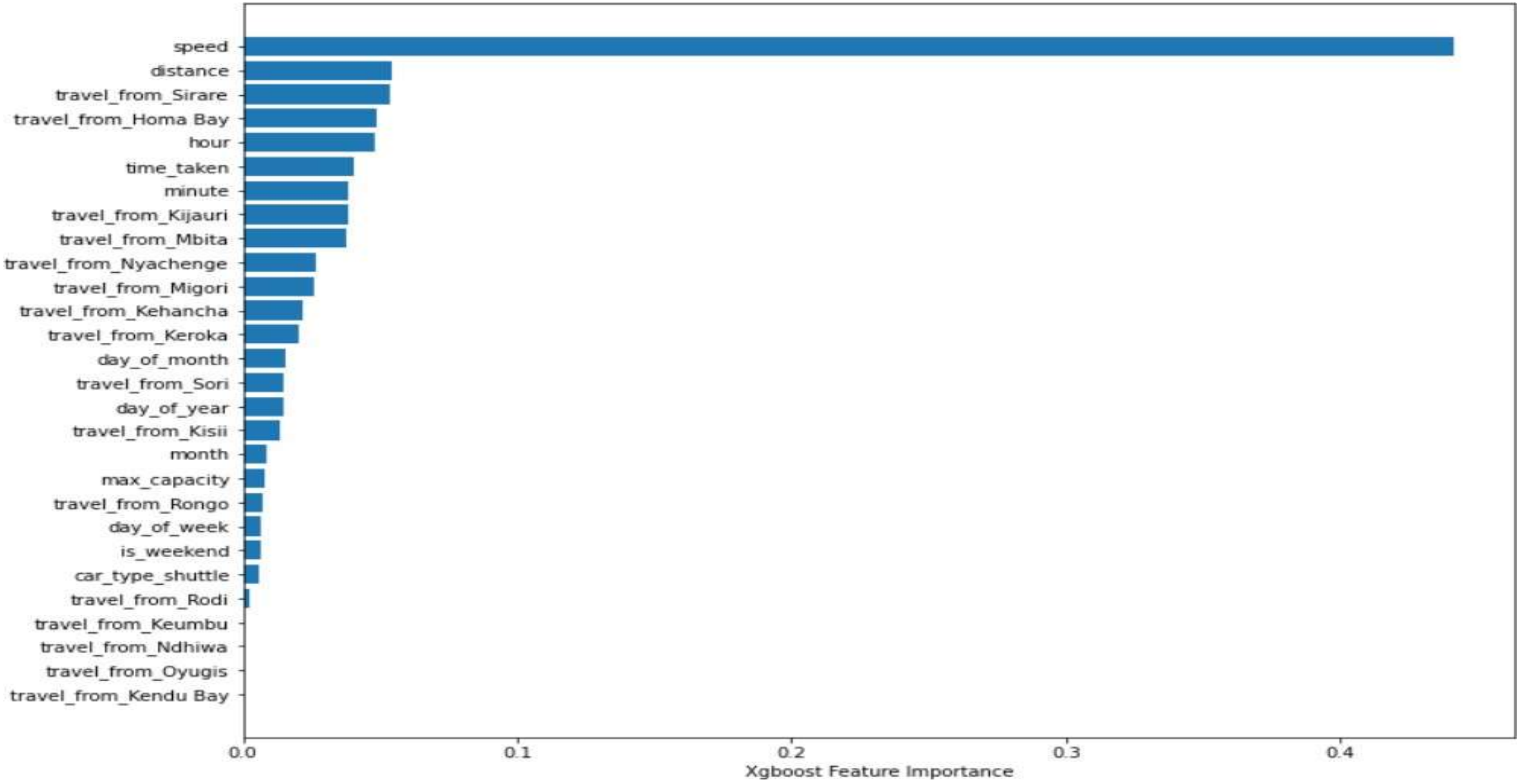
Using domain knowledge to extract features from raw data, by which the performance of the model can be improved.

- Number_of_tickets
- Speed
- Distance
- Time_taken
- Hour
- Minute
- Month
- Day_of_week
- Day_of_month
- Is_weekend

ML Models and Metrics

Model Name	Train Accuracy	Test Accuracy	r2_score	Adjusted r2_score	MSE	RMSE	MAE	MAPE
Linear Regression	0.366001	0.404588	0.404588	0.396211	46.063907	6.787040	4.643707	151.217350
Lasso Regression	0.339320	0.344299	0.344299	0.335073	49.506884	7.036113	5.006458	181.637236
Ridge Regression	0.374090	0.384397	0.384397	0.376073	46.479378	6.817579	4.739805	164.651974
Decision Tree	0.587940	0.555827	0.555827	0.549334	33.536021	5.791029	3.974281	138.058099
Random Forest	0.649787	0.665061	0.665061	0.659980	25.288646	5.028782	3.381311	115.978124
Xgboost	0.792211	0.806540	0.806540	0.803605	14.606686	3.821869	2.579984	84.516617

Feature Importance



Challenges

- Finding the Dependant Variable
- Feature engineering
- Model Training and performance improvement
- Tuning Hyperparameter.

Conclusion

- As we have implemented six different models to predict the number of seats that Mobiticket can expect to sell for each ride. Linear Regression, Regularized linear regression (Ridge and Lasso), Decision Tree, Random Forest Regressor and Xgboost Regressor. Xgboost regression model performed the best among them.
- Our Model will help Mobiticket and Bus operators to anticipate the number of tickets they can expect to sell for each ride.

Thank You