

Capstone Project – 2

Regression

“Transport Demand Prediction”

**Presented By :-
Hasnain Mazhar Rizvi**

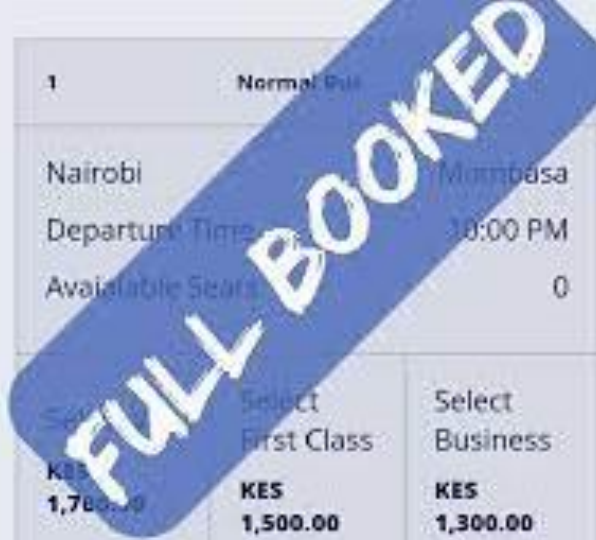
Content

- Problem Statement
- Data Summary
- EDA
- Hypothesis
- Feature Manipulation
- Feature selection
- Data Preprocessing
- ML Implementation
- ML Models and Metrics
- Feature Importance
- Conclusion



Problem Statement

We're going to look at data from Mobiticket, which sells bus tickets. We'll focus on towns northwest of Nairobi towards Lake Victoria. Our goal is to figure out how many tickets will be sold for buses that arrive in Nairobi from those towns.



1	Normal Bus	
Nairobi	Mombasa	
Departure Time	10:00 PM	
Available Seats	0	
Select	Select First Class	Select Business
KES 1,780.00	KES 1,500.00	KES 1,300.00

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
- **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

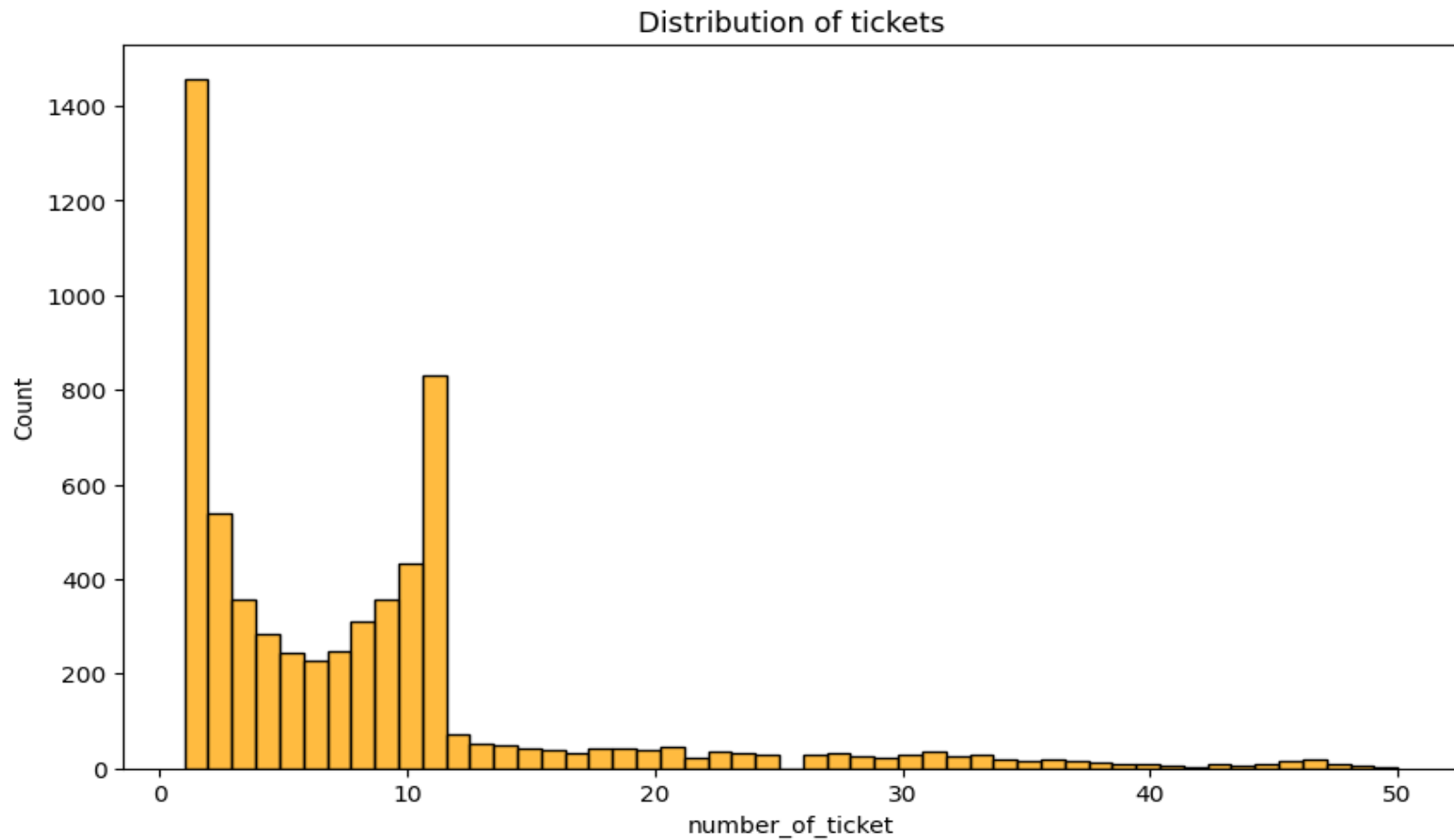
EDA

Know Your Data

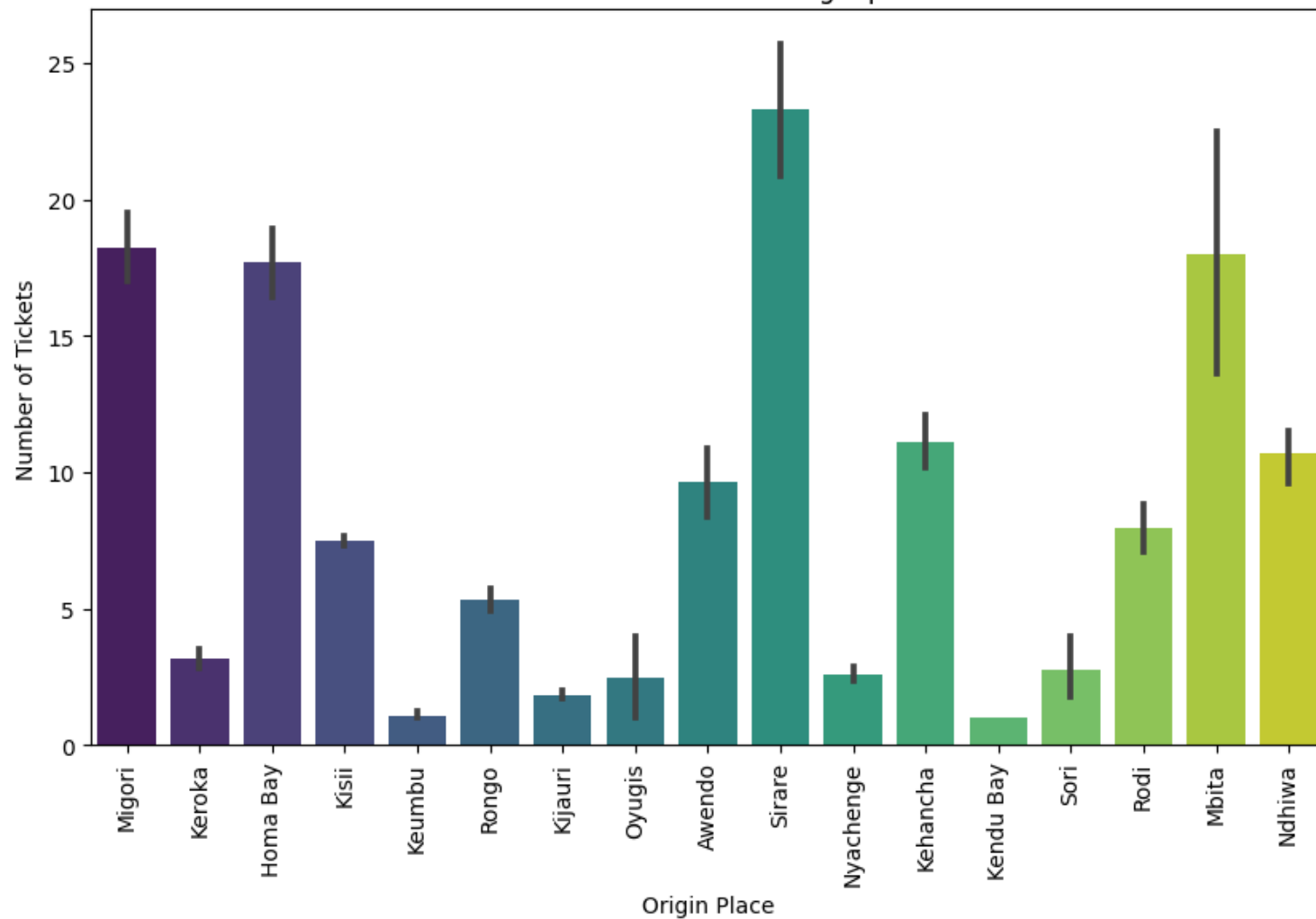
- Importing Basic libraries.
- Loading Dataset
- After doing some basic Python work we understand our dataset and see `head()` and `tail()` And we get to know we have 51645 Rows and 10 Columns.
- Duplicate Values
- Null/Missing Values
- Data Wrangling
- Visualization of data

Note: They didn't give us a target variable so we have to find out it first and then we can visualize the data in a proper manner.

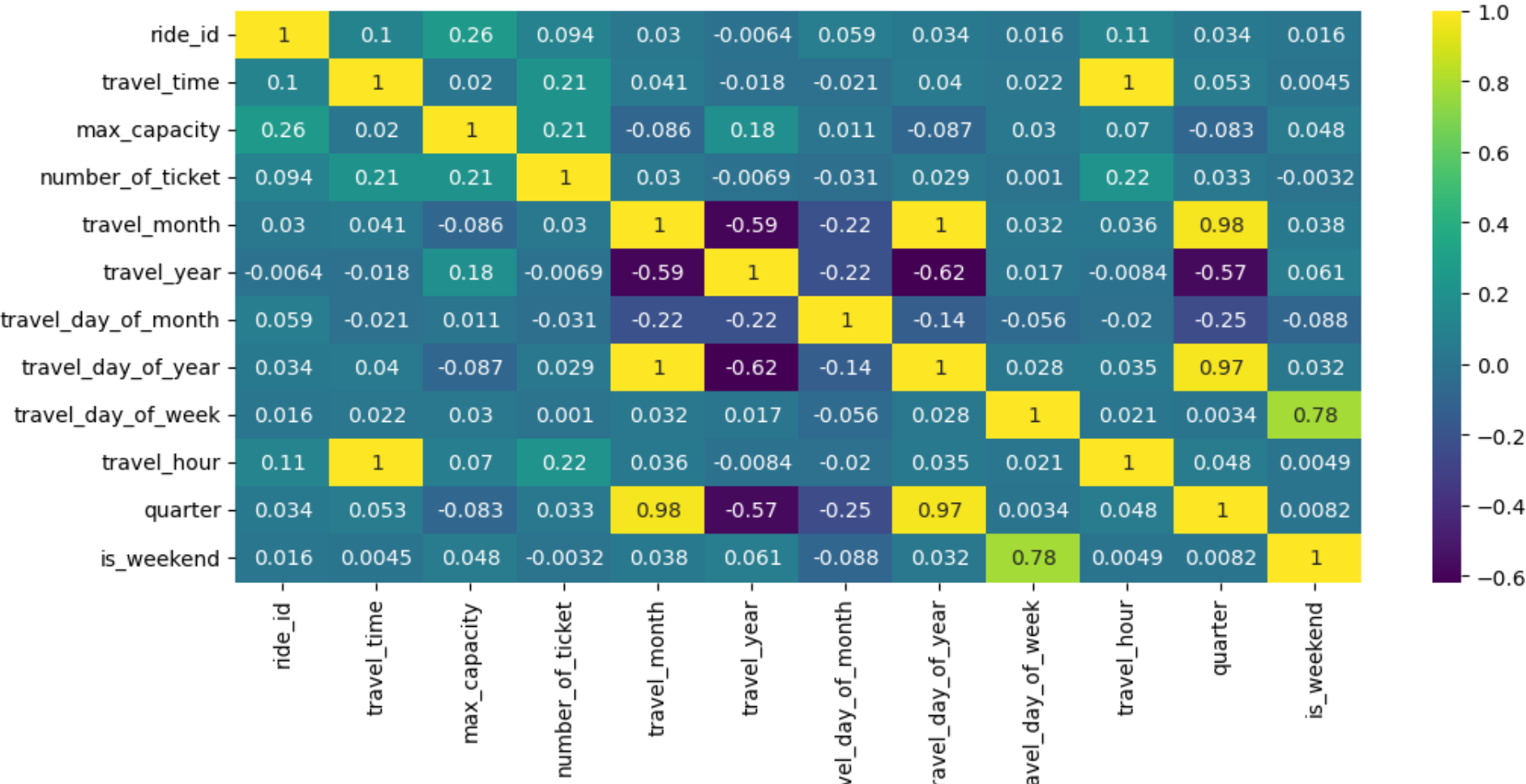
Visualization



Total tickets from each origin place



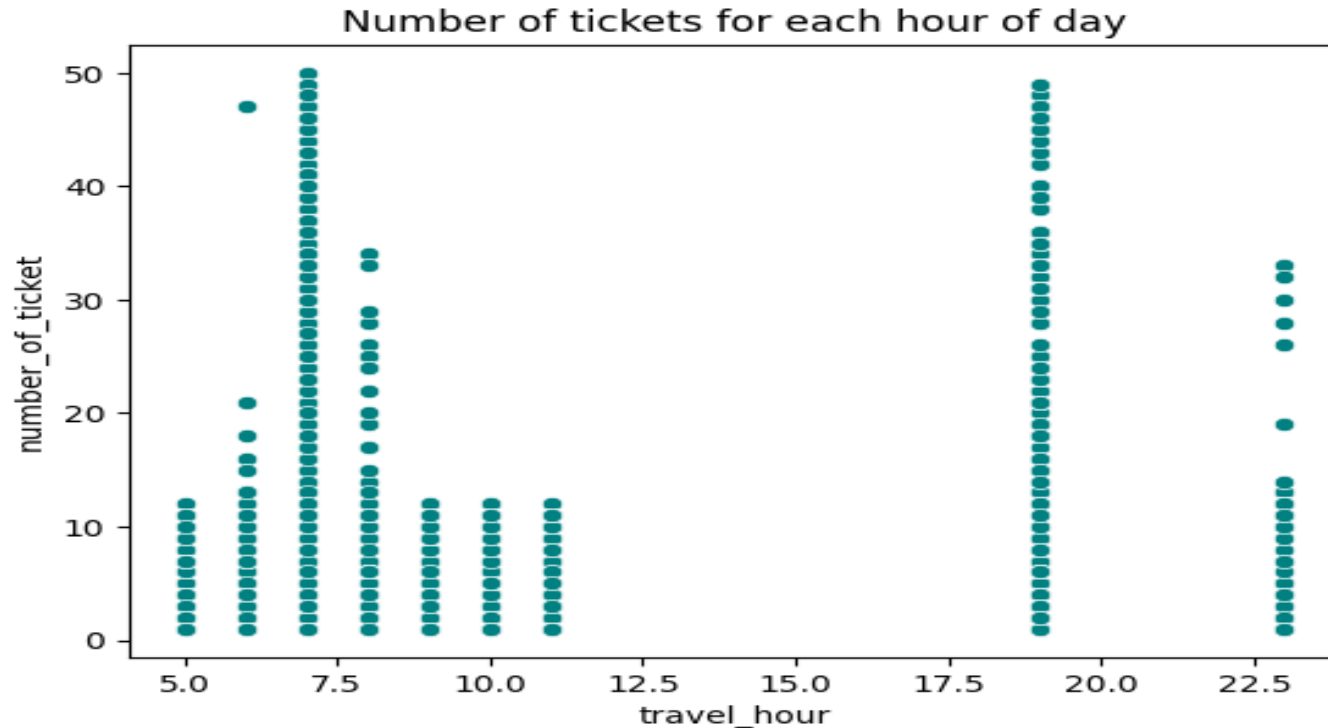
Visualization



Hypothesis 1:

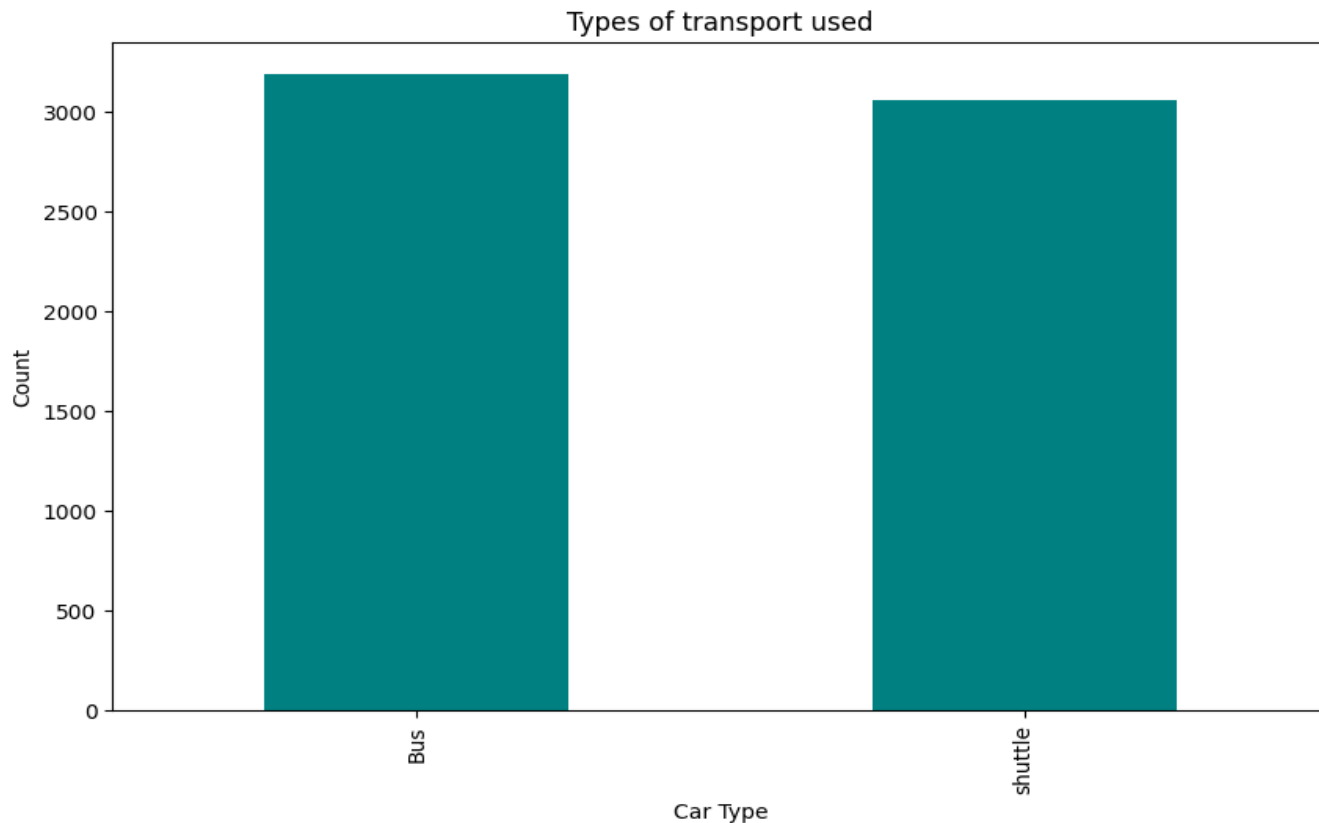
We Found Three Hypotheses after lots of charts so let's discuss them one by one.

1. There are No travel activities taking place during the afternoon.



Hypothesis 2:

2. The number of buses(N_b) used in traveling is the same as the number of shuttles(N_s) used.

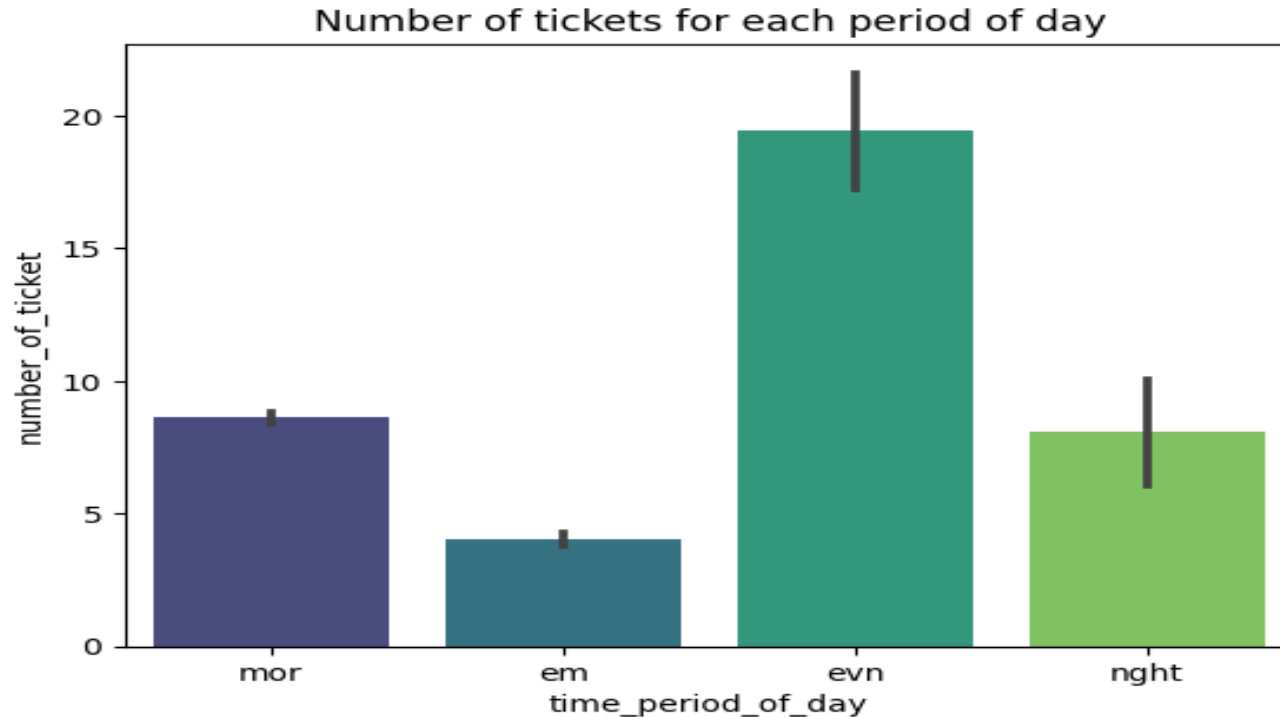


Hypothesis 3:

3. Most tickets are sold in the morning.

let's: H_0 = Most tickets are sold in the morning.

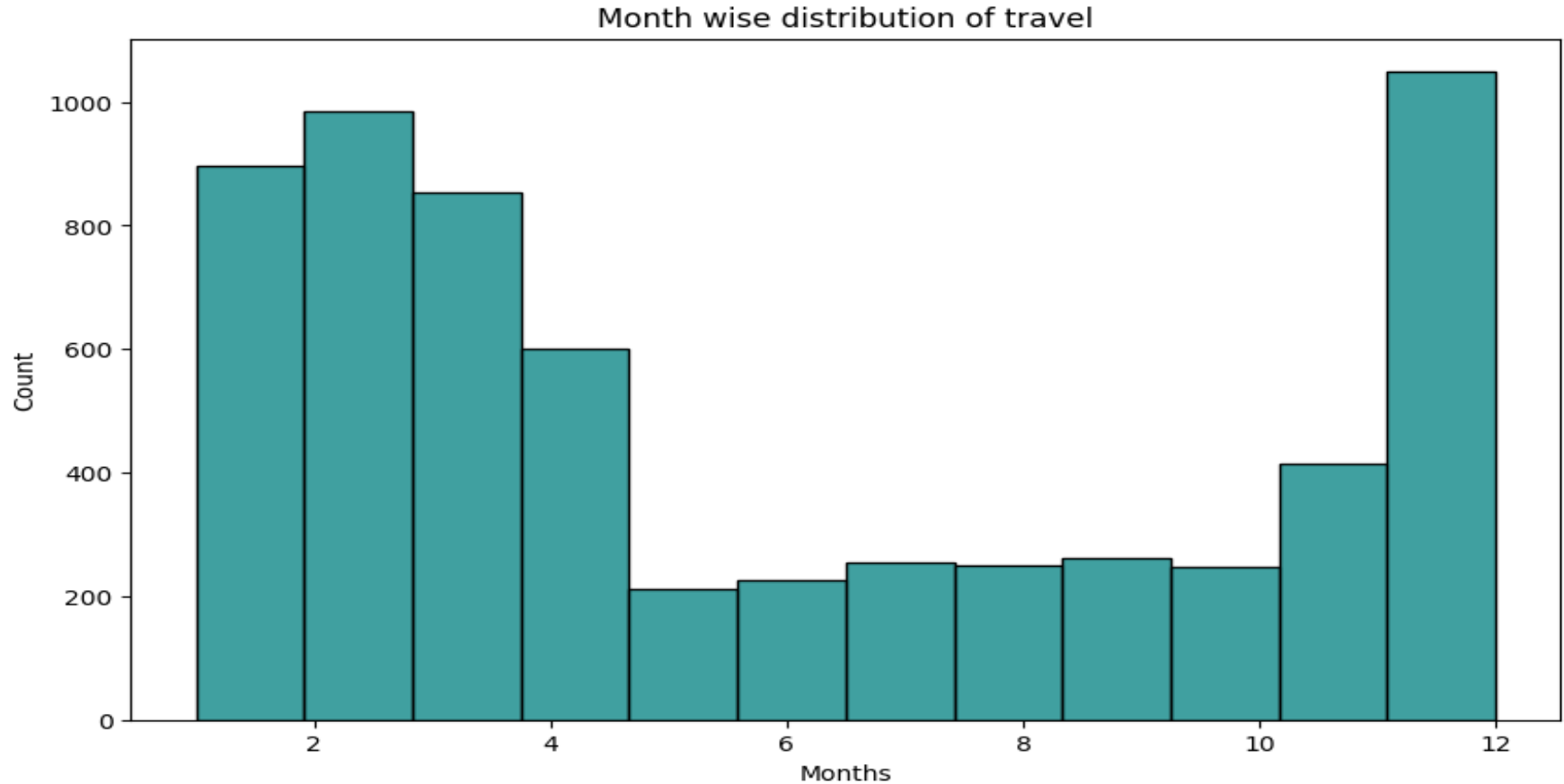
H_1 = Most number of tickets are NOT sold in the morning



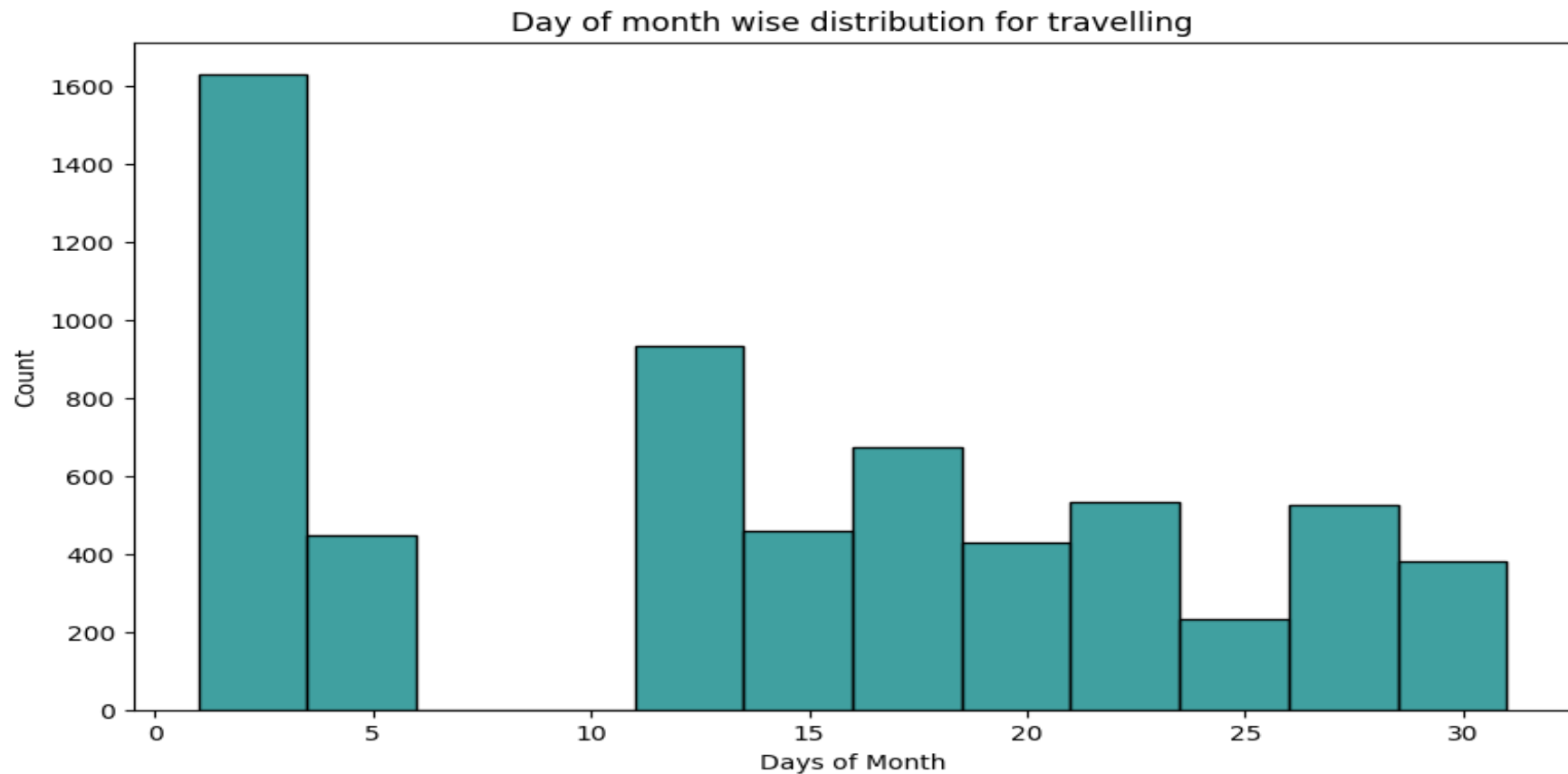
Feature Manipulation and selection

1. **By Extracting variables from variables we found these things :**
2.
 1. **Some months have a higher frequency of travel.**
 2. **Some days in the year have a very high frequency of travelers, while others have a really low frequency. This is because of more traveling being done in some months than others.**
 3. **Some days in a month have a higher frequency of travel than others.**
 4. **Some periods in the day have a high frequency of travel.**

Feature Selections

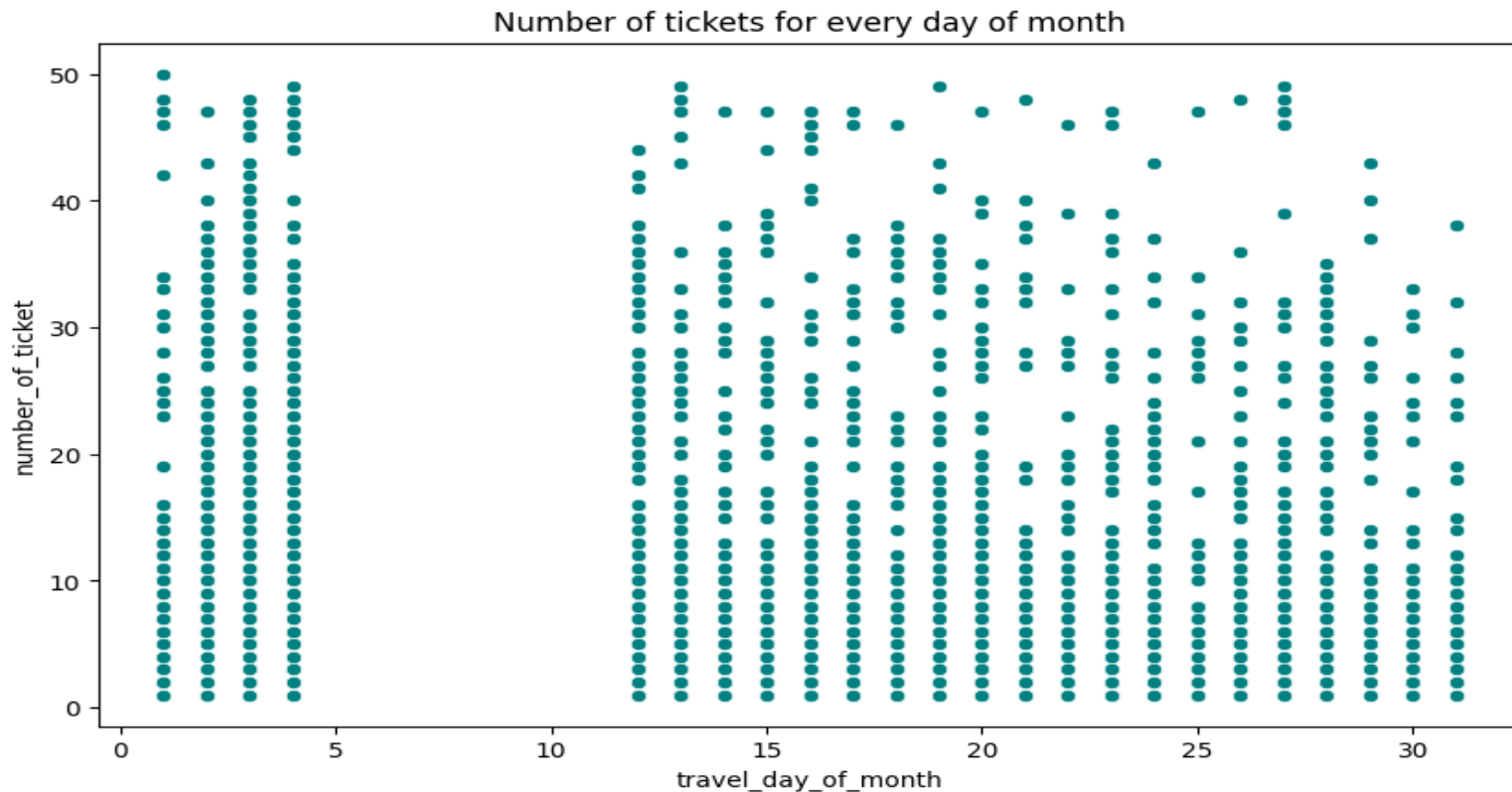


Feature Selections



Feature Selections

- Similar to the above Graph, it's a seaborn scatter plot for another view of visualization.



Data Preprocessing :

To enhance the performance of the model, additional features have been generated. These new features aim to provide more relevant information and contribute to improved predictions.

- Some columns `travel_month`, `travel_day_of_year`, `travel_day_of_month` and `time_period_of_day` have skewed data, to handle this, new weight wise columns have been created by taking log transformation.

1. `Time_gap_btw_0_1_next_bus`
2. `Time_gap_btw_0_1_previous_bus`
3. `Time_gap_btw_0_2_next_bus`
4. `Time_gap_btw_0_2_previous_bus`
5. `Time_gap_btw_0_3_next_bus`
6. `Time_gap_btw_0_3_previous_bus`
7. `Time_gap_btw_next_previous_bus`

8. ...

ML Implementation

- **Five models are implemented to predict the transport demand to Nairobi. These include: Linear Models:**

- 1. Linear Regression**
- 2. Lasso Regression (L1)**
- 3. Ridge Regression (L2)**

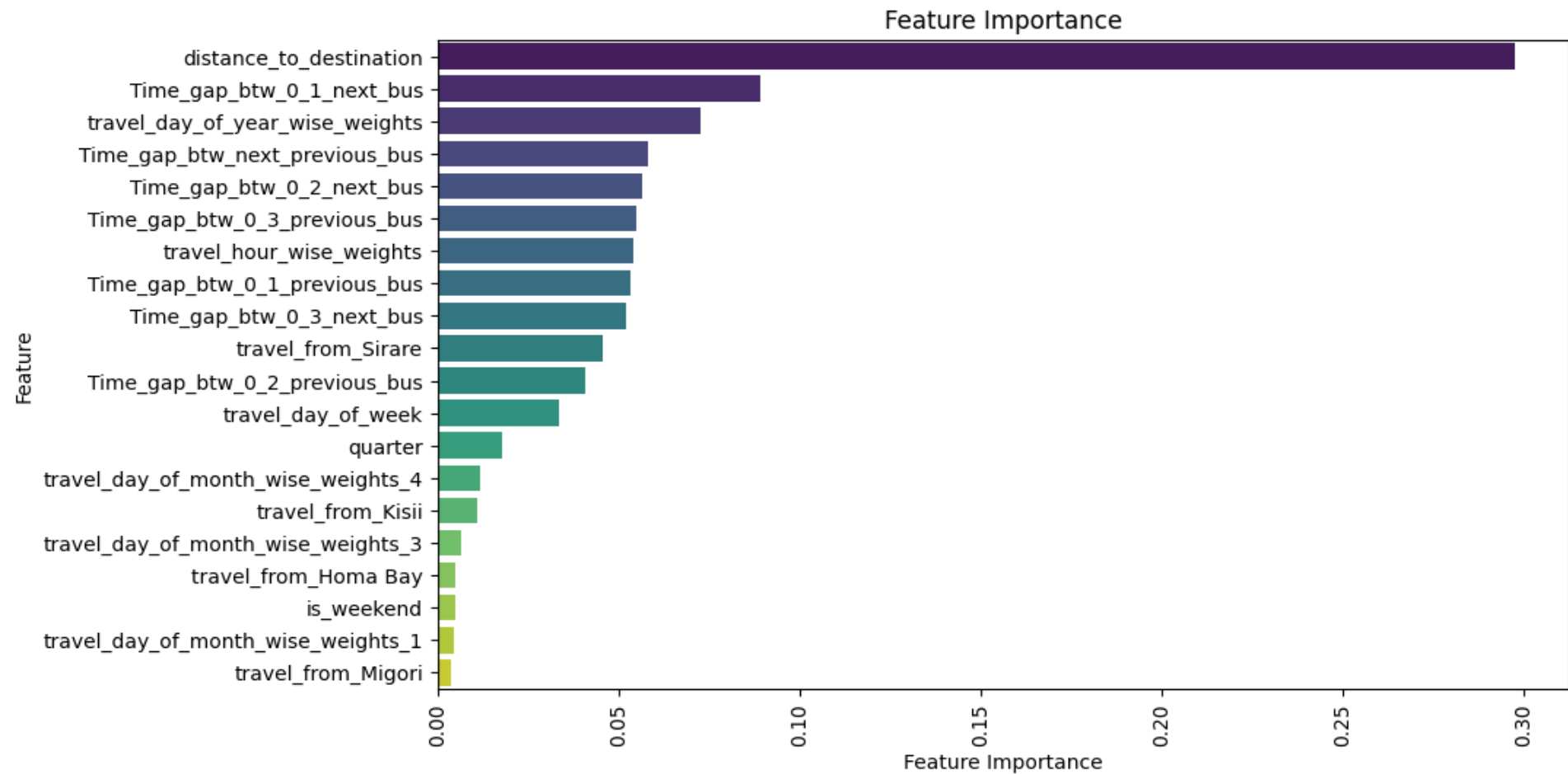
Ensemble Models:

- 1. XGBoost**
- 2. Random Forest**

ML Models and Metrics

Models	R2 - Score	Adjusted R2 - Score
Linear Regression	0.35	0.33
Lasso (L1)	0.248	0.225
Ridge(L2)	0.356	0.337
XGBoot	0.882	0.887
Random Forest	0.944	0.946

Feature Importance



Conclusion :

- In our project, we employed diverse regression techniques to forecast transportation demand from different locations to Nairobi.
- Utilizing the available data, we engineered both the target variable and several other features crucial for enhancing our model's predictive performance.
- This project resulted in a hyperparameter-tuned Random Forest-based predictive model, achieving an R2 score of 94.4% for training data and 94.6% for test data.

Thank You AlmaBetter!