Data Science Major Project

Project Topic: - Transport Demand Prediction

PROBLEM STATEMET:

- This challenge asks you to 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.

1. PROJECT OBJECTIVE:

- The objective of this project is to predict the number of seats sold per bus ride based on historical ride data. Accurate predictions help operators:
  - Identify peak demand rides.
  - Improve operational efficiency and revenue management.

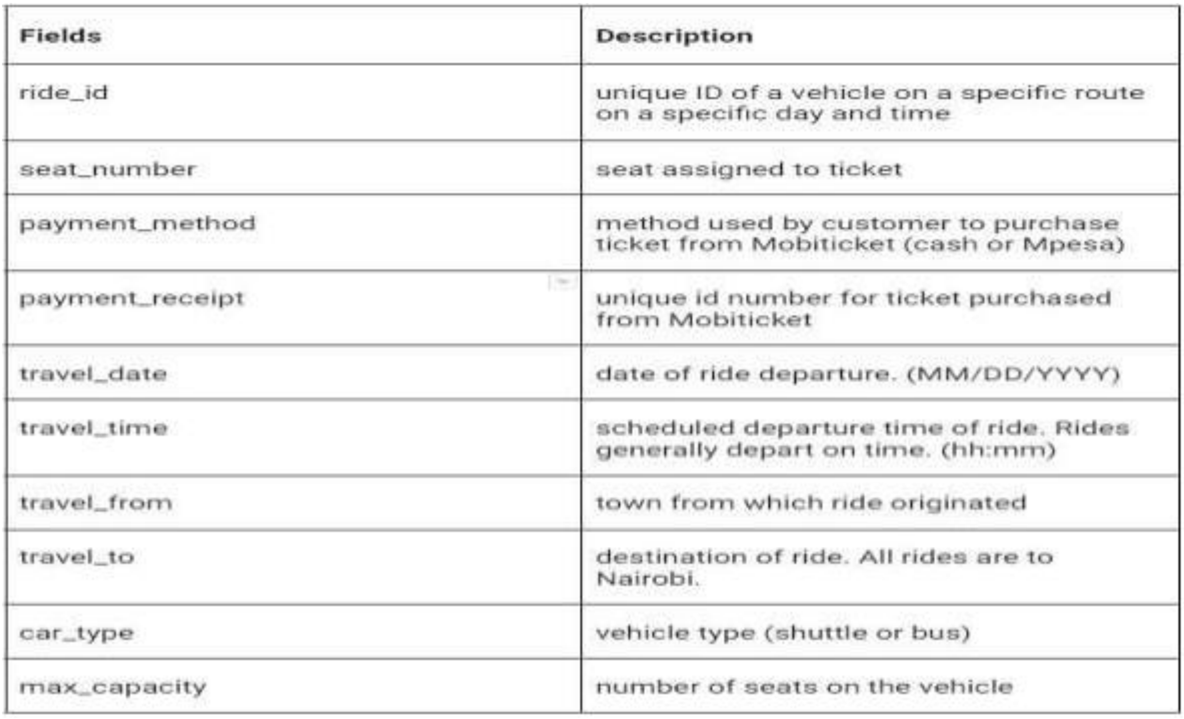
2. TOOLS AND LIBRARIES USED

This project was implemented in **Python** using the following libraries:

Library	**Purpose**
pandas	For data loading, cleaning, aggregation, and manipulation.
numpy	For numerical operations, array manipulation, and mathematical calculations (e.g., IQR computation).
scikit-learn (sklearn)	For machine learning tasks:
	- train_test_split: Split data into training and test sets.
	- RandomForestRegressor: Build and train Random Forest model.
	- mean_absolute_error, mean_squared_error, r2_score: Evaluate model performance.
matplotlib.pyplot	For plotting feature importance and visualizing results.

3. Dataset Overview:

The dataset train_revised.csv contains historical bus ride data:



Fields	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 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

4. Data Cleaning

1. **Duplicate Removal**
   Removed duplicate records using ride_id and seat_number.
2. **Standardization**
   Standardized car_type by stripping whitespace and converting to Title Case.
3. **Datetime Conversion**
   Converted travel_date to datetime objects for easy extraction of day/month features.
   Extracted hour from travel_time.
4. **Seat Aggregation**
   Counted number of seats booked per unique ride (grouped by ride_id).
   If count exceeded max_capacity, we capped it at max_capacity to avoid unrealistic overbooking.
5. **Peak Demand Identification**
   Used **Interquartile Range (IQR)**:
   - Q1 = 25th percentile, Q3 = 75th percentile
   - Upper bound = Q3 + 1.5×IQR
     Added binary column is_peak_demand (1 if seats_sold > upper bound).

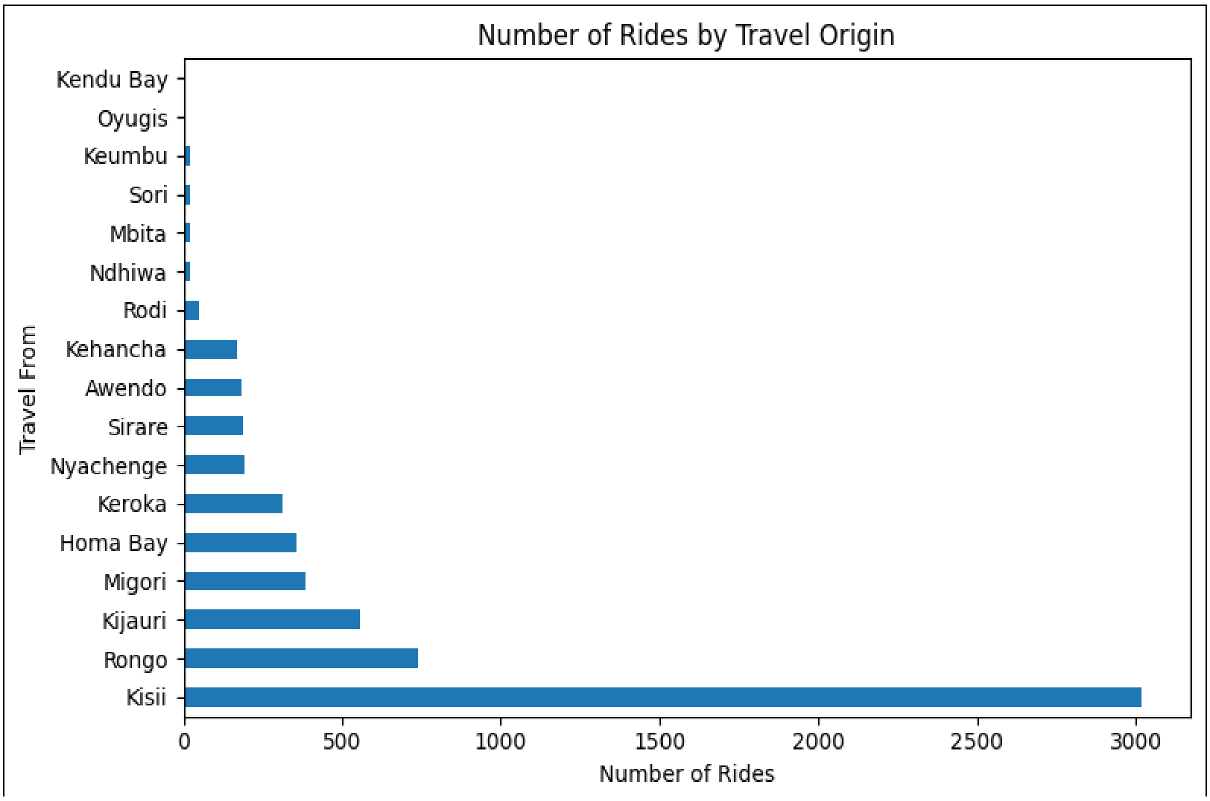
✅ **Result:** After cleaning and aggregation, dataset size reduced to **unique rides**, ready for feature engineering.

5. Feature Engineering

We generated several new features to improve prediction accuracy:

- **Date-based features:**
  - day_of_week (Monday, Tuesday, …)
  - month (1-12)
- **Time-based features:**
  - hour (numeric hour extracted from travel_time)
- **Categorical Encoding:**
  Applied one-hot encoding to travel_from, car_type, and day_of_week to convert categorical values into numeric form.
- **Dropped columns:**
  Removed ride_id, travel_date, travel_time, and travel_to (not directly useful for prediction).

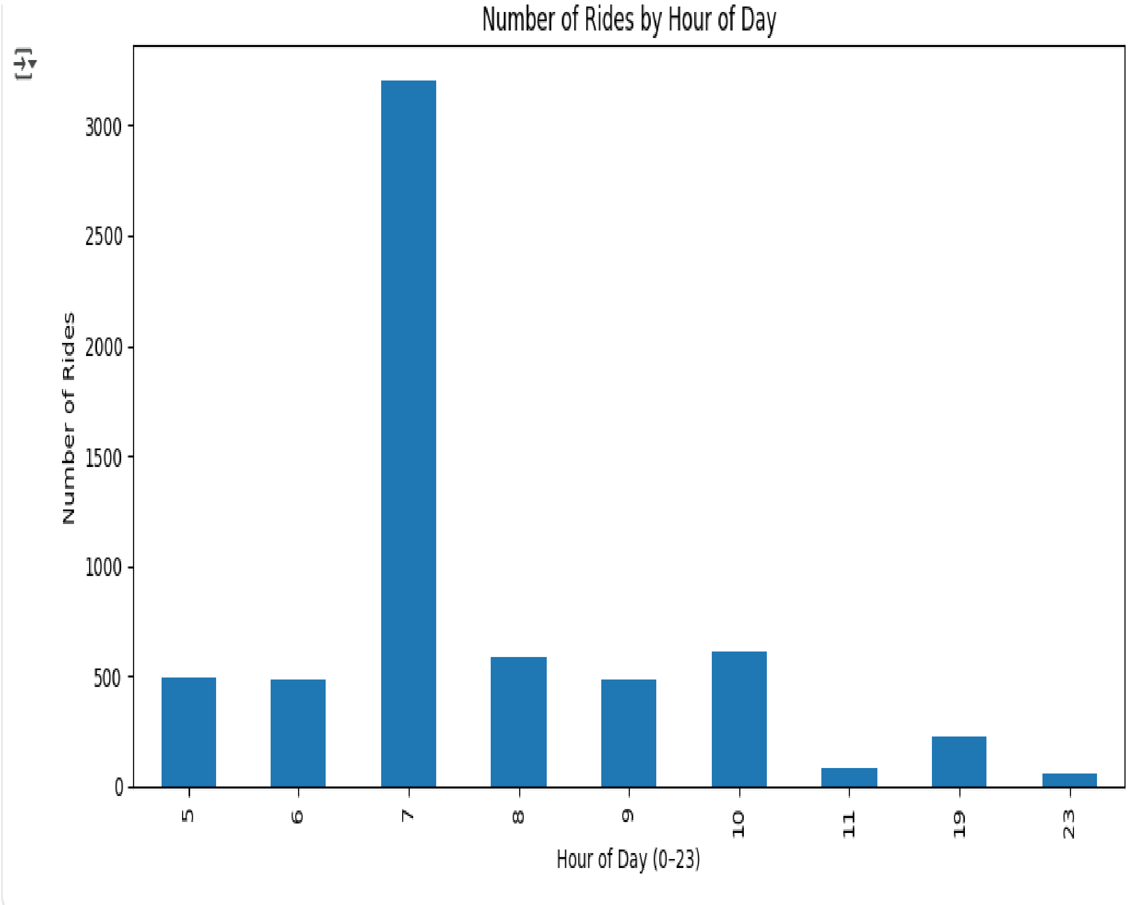
6. Data Visualization:



Key Observations:

- **Kisii** has the **highest number of rides**, significantly more than any other location (almost 3000 rides).

- **Rongo** and **Kijauri** follow, but their numbers are much lower than Kisii's.
- Other moderately contributing locations include **Migori**, **Homa Bay**, and **Keroka**.
- Locations like **Keumbu**, **Sori**, **Mbita**, **Ndhiwa**, and **Oygis** have **very few rides**.

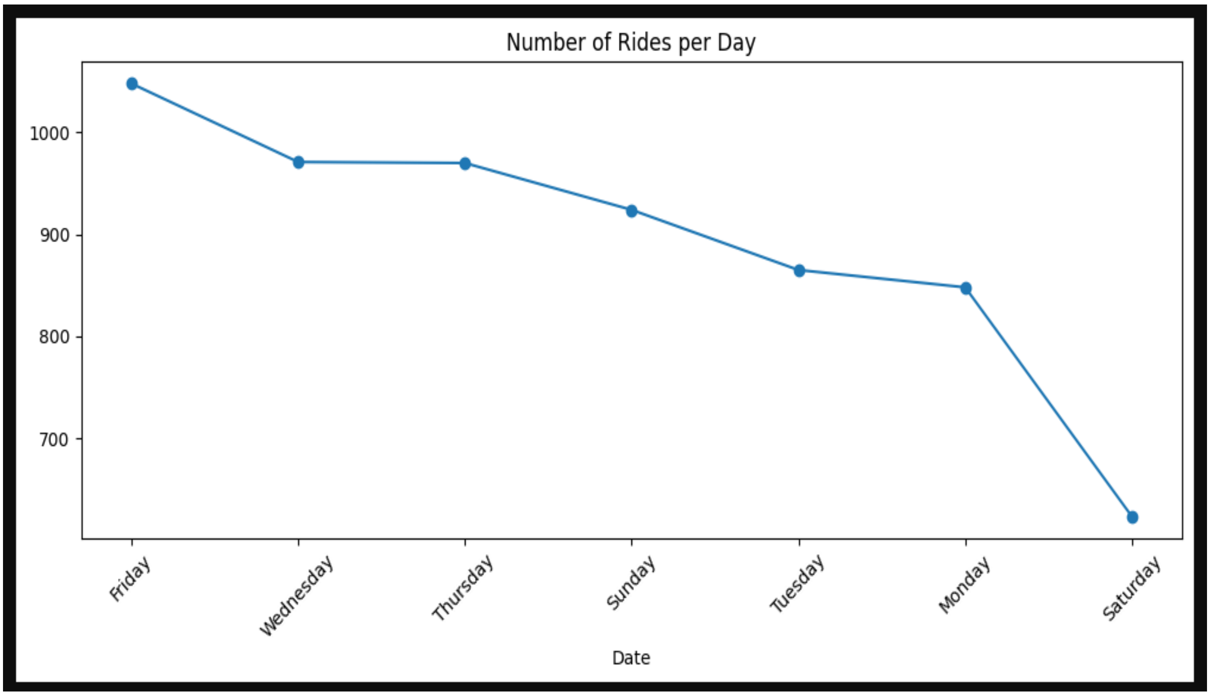


🔎 Key Observations

1. **Peak Hour is 7 AM** –
   The bar for **7:00** is much higher than all other hours, meaning a majority of rides are concentrated early in the morning.
   This indicates **commuting or school/work rush hour demand**.

2. **Moderate Demand in 5 AM, 6 AM, 8 AM, 10 AM** –
   There are smaller peaks around 5–6 AM and 8–10 AM, showing early-morning and late-morning trips.

3. **Very Low Demand in Evening & Late Night** –
   19:00 (7 PM) has a small number of rides, and 23:00 (11 PM) & 11:00 have the lowest.



Key Observations:

- **Friday** has the **highest number of rides** (just over 1050).
- Ride numbers **gradually decline** from Friday through to **Monday**.
- **Saturday** has the **lowest ride count**, with **fewer than 650 rides**.
- **Wednesday** and **Thursday** have nearly the same number of rides (just under 1000).
- There is a **notable drop** from **Monday to Saturday**, suggesting weekends may see reduced demand.

Insight Summary:

- **Weekdays (especially mid-to-late)** are the busiest.
- **Saturday** has the **least ride activity**, indicating a potential opportunity (or drop in demand) for services on weekends.

7. Train-Test Split

We split the data into **training set (80%)** and **test set (20%)** using `train_test_split`.

This ensures the model is trained on historical data and evaluated on unseen data.

8.Model Selection: Random Forest Regressor

Hyperparameters

Parameter	Value
n_estimators	**200**
min_samples_split	**10**
min_samples_leaf	**4**
max_depth	**None**
bootstrap	**True**

Evaluation

Metric	Score
Mean Absolute Error (MAE)	3.3125
Root Mean Squared Error (RMSE)	20.9472
R² (Coefficient of Determination)	0.7396

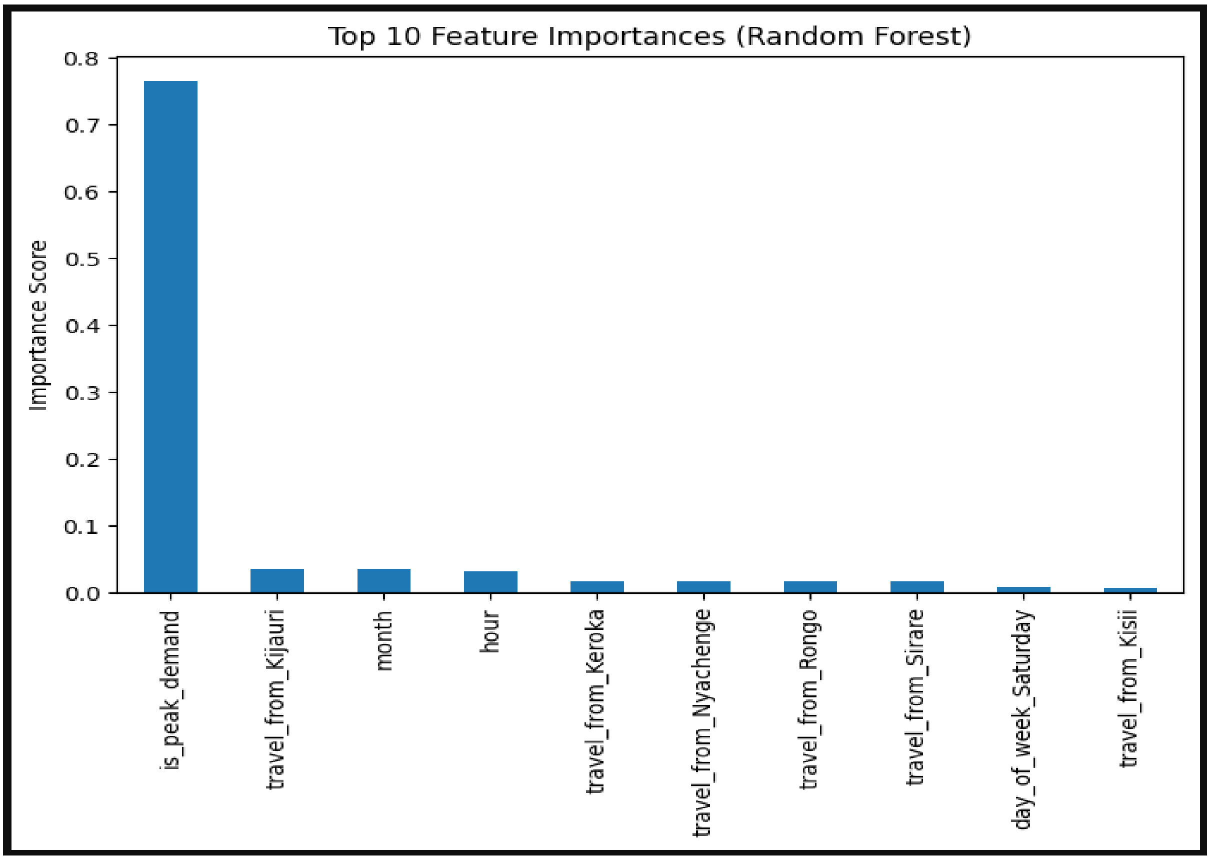
Why Random Forest ? :

We use **Random Forest** because:

- It gives **better predictive performance** than simple models.
- It can model **complex patterns** in data.
- It provides **interpretability** through feature importances.
- It is **robust, scalable, and reliable** for production use.

9. Feature Importance (Random Forest)

The model identifies which features most affect predictions



- A bar chart of top 10 features was plotted for clear understanding.
  This visualization helps stakeholders see which variables matter most for seat prediction.

12. Conclusion

- **Random Forest Regression** is the best model for predicting seat demand.
- Achieved **73.96% R²**, indicating strong predictive power.
- Top factors:Peak_demand, hour of travel, and month.
- This model can help bus operators optimize routes and schedule more efficiently.