UBER DATA ANALYSIS

GROUP 10

BINDU YADDULA

RADHA GUDE

SUPRIYA PATHURI

VAMSI KRISHNA DHIDDI

SCHOOL OF COMPUTING AND ENGINEERING:
UNIVERSITY OF MISSOURI- KANSAS CITY

PRINCIPLES OF DATA SCIENCE

YU LUO

DATE: 12/10/2023

Table of contents

Abstract:

This project, titled "Uber Data Analysis," presents a comprehensive analysis of Uber's ride-sharing data using machine learning techniques to forecast demand. The project explores ride frequency, distance, and purpose, using time-series and categorical analysis. Our findings, illustrated with graphs, offer insights into peak usage times and popular ride types. The study aims to improve ride-sharing efficiency and customer satisfaction by forecasting demand trends. These insights are pivotal for Uber's strategic planning, impacting service delivery and user experience.

1. Introduction:

Project Overview

This project focuses on utilizing machine learning for demand forecasting in ride-sharing, with Uber as the primary case study. Emphasis is placed on the crucial role of forecasting in improving operational efficiency and customer satisfaction.

Background

Demand forecasting is essential in transportation, aiding in fleet management, understanding customer behavior, and strategic planning.

Scope

The study investigates Uber data from 2016, covering rides predominantly in the USA. It aims to analyze ride frequencies, distances, purposes, and user preferences. Beyond data analysis, the study aims to interpret the results in the context of demand forecasting and operational planning for Uber.The project includes visual analysis to find out key findings. This involves generating time-series graphs, categorical analyses, and heatmaps to uncover patterns in ride frequency, purposes, and time-of-day preferences. These visual tools are crucial for a more intuitive understanding of the data.

2. Objectives:

Utilizing Machine Learning for Demand Forecasting: The project focuses around applying machine learning techniques to forecast demand in ridesharing. We plan to find out the behavior and patterns of Uber users throughout a year.

Identifying Key Factors Affecting Demand: Identify the key factors that affect demand in ride-sharing, such as time, location, and purpose of the rides.

Analyzing Trends in Ride-Sharing Demand: The project aims to analyze trends in ride-sharing demand, focusing on various aspects such as ride frequencies, distances, purposes, and user preferences.

Compare two different machine learning models linear regression and random forest classfier and check the performance accordingly.

Research Questions/Hypotheses

The project revolves around:

- Analyzing trends in ride-sharing demand.
- Identifying key factors affecting demand, such as time, location, and purpose.

3. Data Description

Data Sources

The dataset, "My Uber Drives (2016)" from Kaggle, is a credible source of real-world Uber ride data.

Data Characteristics

Dataset Details: https://www.kaggle.com/datasets/zusmani/uberdrives

The data has the ride details in the USA, Sri Lanka, and Pakistan and it spans from January to December 2016.

Attributes: START_DATE, END_DATE: The start and end timestamps of the rides.

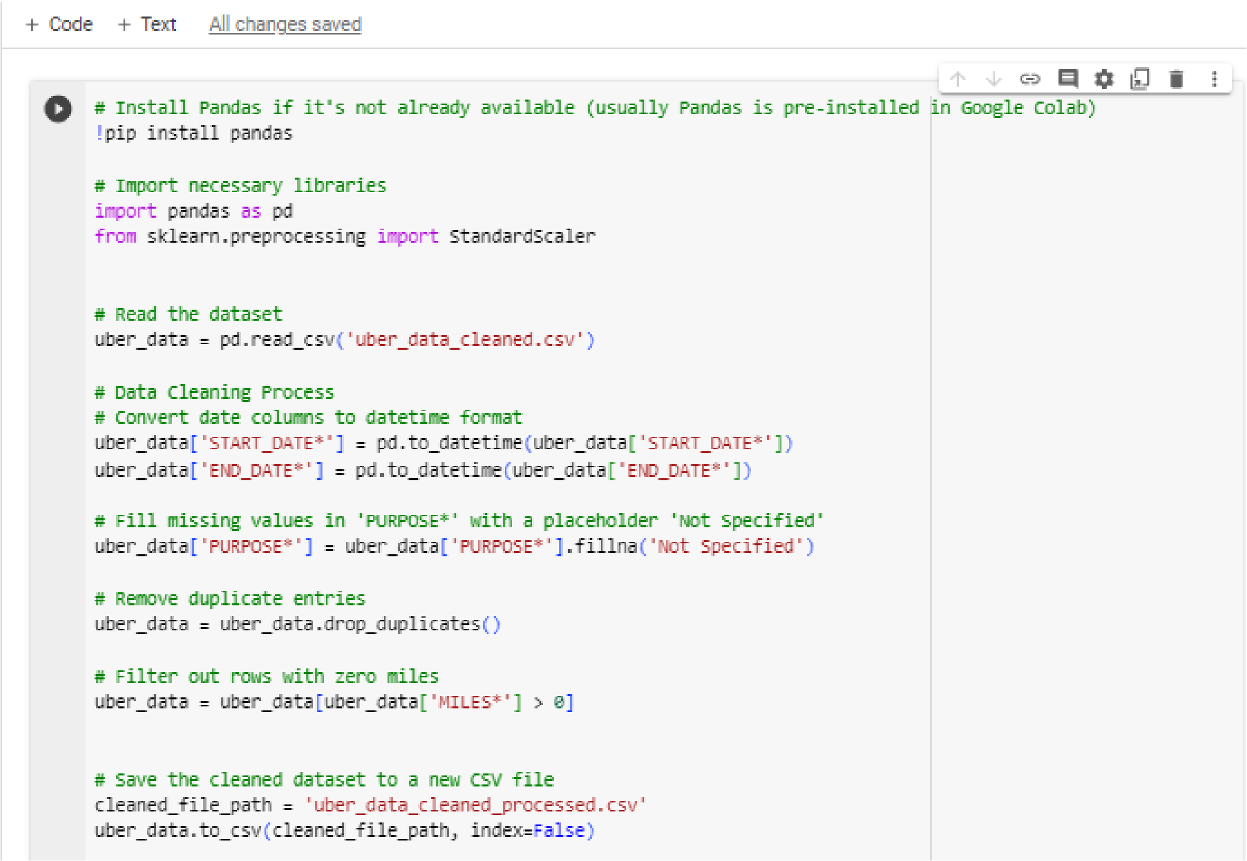
CATEGORY: The category of the ride (Business, Personal).

START and STOP: Start and stop locations of the ride.

MILES: The distance of the ride. PURPOSE*: The purpose of the ride.

Data cleaning method: We have converted the START_DATE and END_DATE into date and time format to get an analysis of time-series. We have also handled the missing values in “PURPOSE” column by using model-based imputations. Data validation has been done and the entries are consistent.

Data Cleaning Process: Transforming Raw Data into Insights



```
# Install Pandas if it's not already available (usually Pandas is pre-installed in Google Colab)
!pip install pandas

# Import necessary libraries
import pandas as pd
from sklearn.preprocessing import StandardScaler


# Read the dataset
uber_data = pd.read_csv('uber_data_cleaned.csv')

# Data Cleaning Process
# Convert date columns to datetime format
uber_data['START_DATE*'] = pd.to_datetime(uber_data['START_DATE*'])
uber_data['END_DATE*'] = pd.to_datetime(uber_data['END_DATE*'])

# Fill missing values in 'PURPOSE*' with a placeholder 'Not Specified'
uber_data['PURPOSE*'] = uber_data['PURPOSE*'].fillna('Not Specified')

# Remove duplicate entries
uber_data = uber_data.drop_duplicates()

# Filter out rows with zero miles
uber_data = uber_data[uber_data['MILES*'] > 0]


# Save the cleaned dataset to a new CSV file
cleaned_file_path = 'uber_data_cleaned_processed.csv'
uber_data.to_csv(cleaned_file_path, index=False)
```

Technologies and Libraries Used:

Python

Pyspark which is a Python API

Pandas and NumPy

Google Colab

Key visual explorations will include:

- Time-series analysis to identify ride frequency patterns over the year.
- Categorical analysis of ride purposes, revealing insights into the primary reasons for Uber usage.
- Analysis of ride distances, exploring their distribution and potential correlations with ride purposes.

4. Methodologies and Insights Generated:

In our study, we use different charts to understand when and why people need rides.

Figure 1 shows us how the number of rides changes each month. This helps us see when rides are most and least needed during the year.

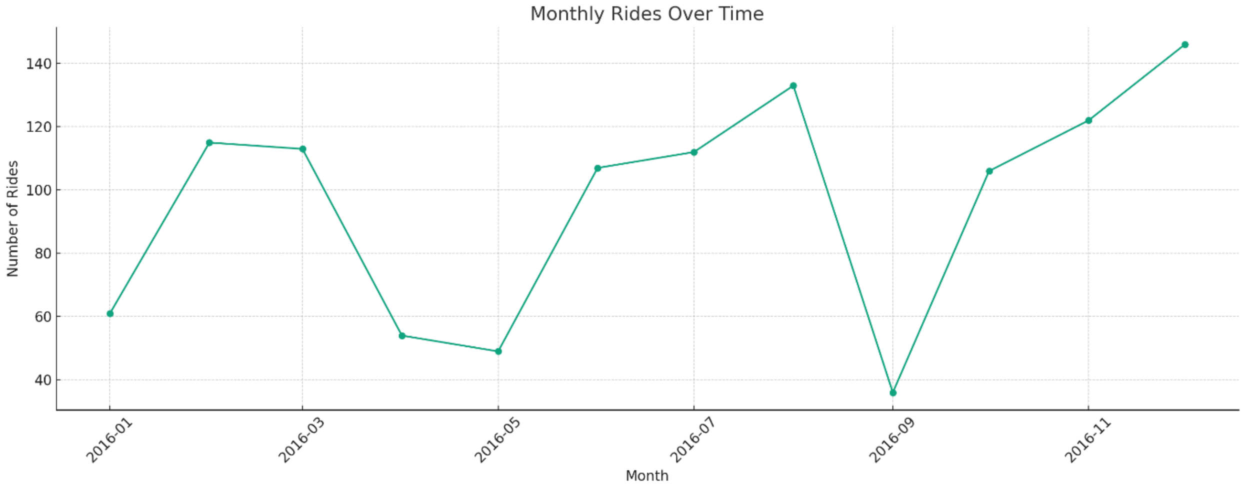


Figure 2 is a chart that tells us how far most people travel. It's useful for figuring out how long the average trip is.

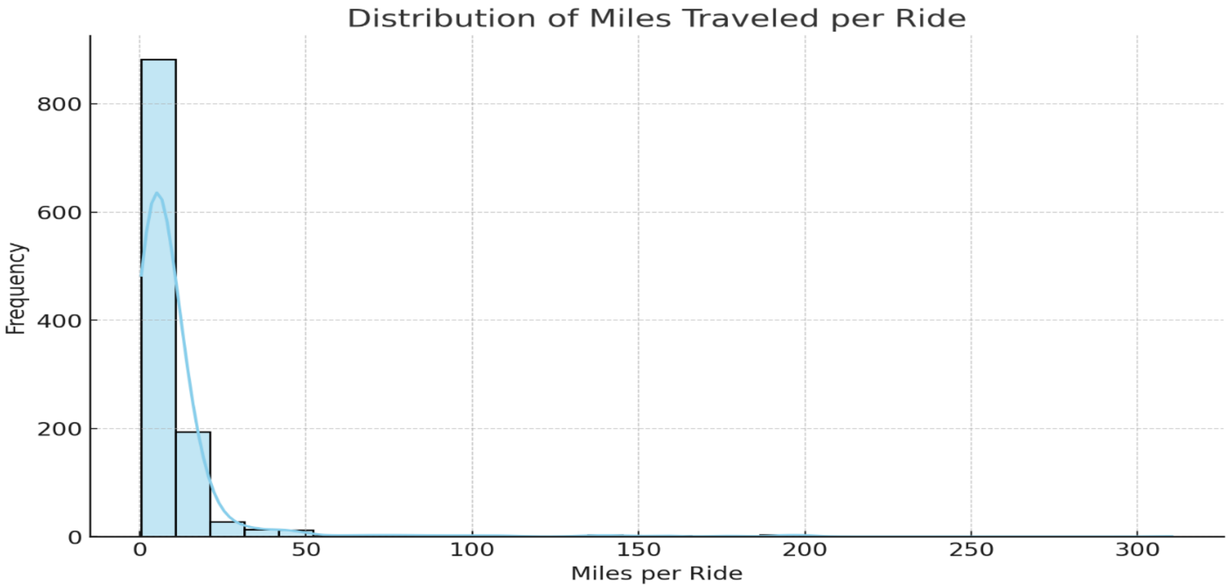


Figure 3 is another chart that gives more detail about how far people travel, like showing the longest and shortest trips. It helps us spot unusual trips.

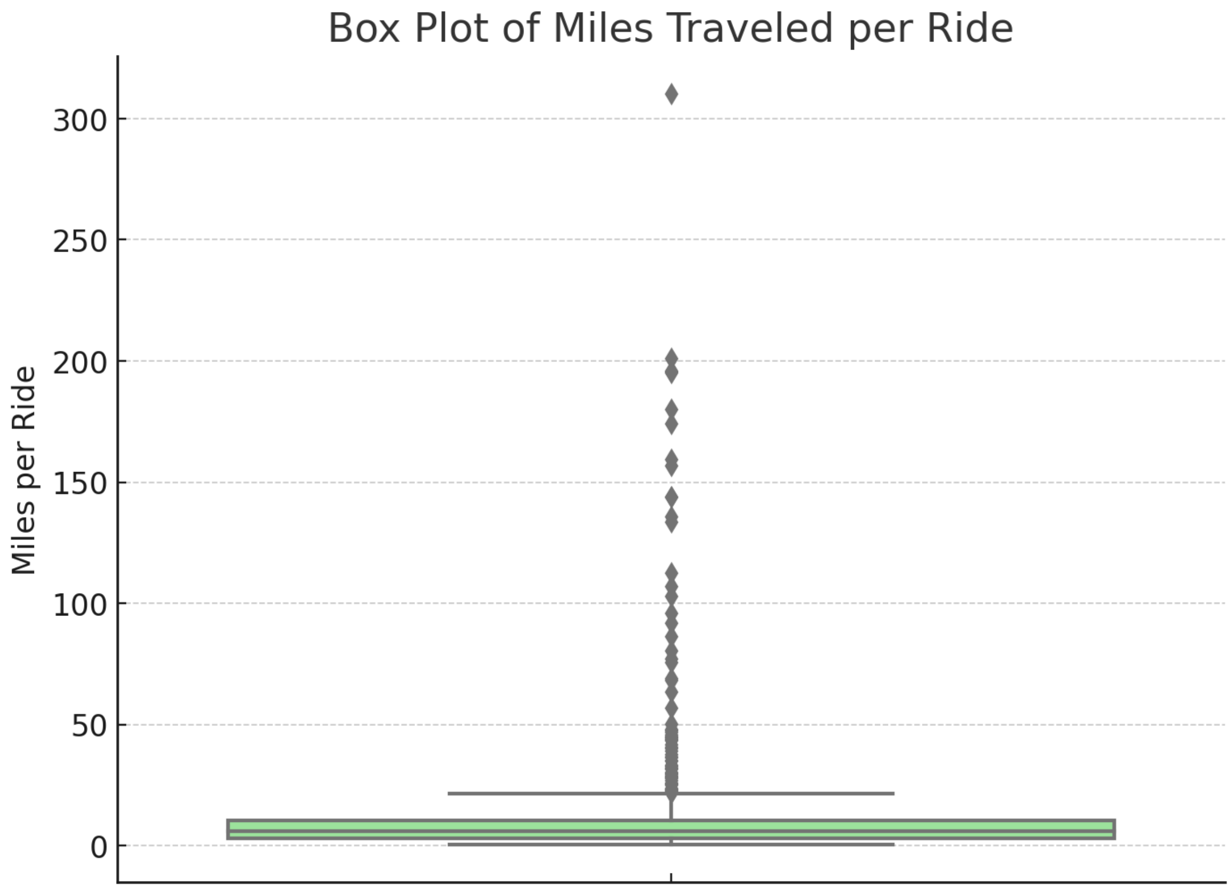


Figure 4 is a map of colors that shows when during the week and what time of day people take the most rides. This is helpful for knowing when we need more drivers.

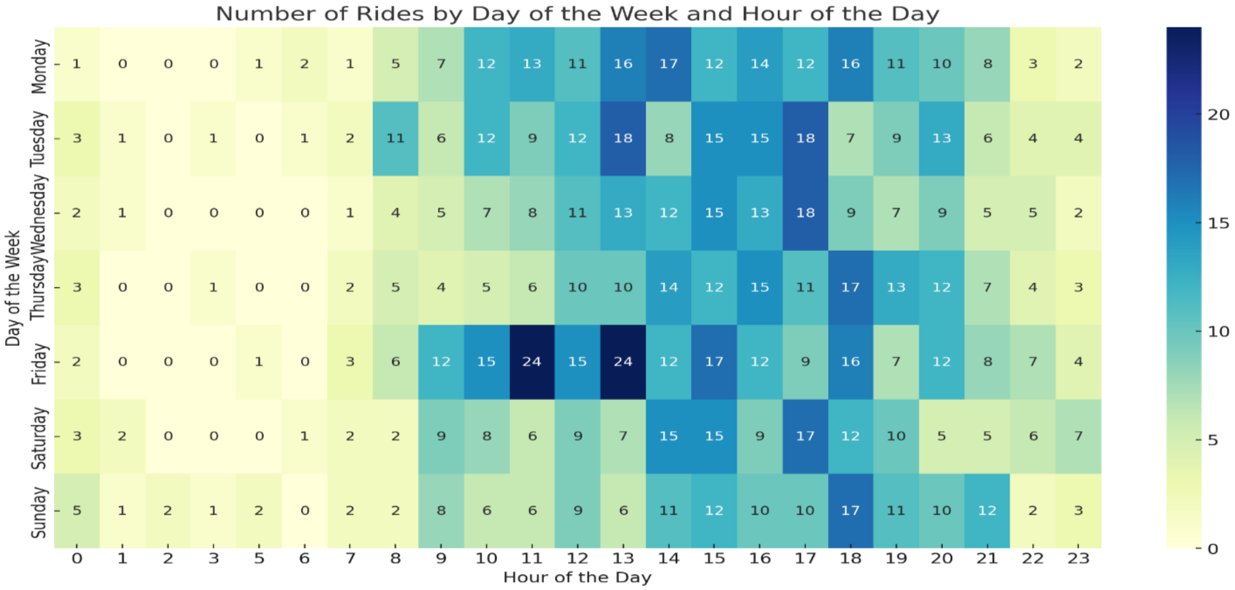
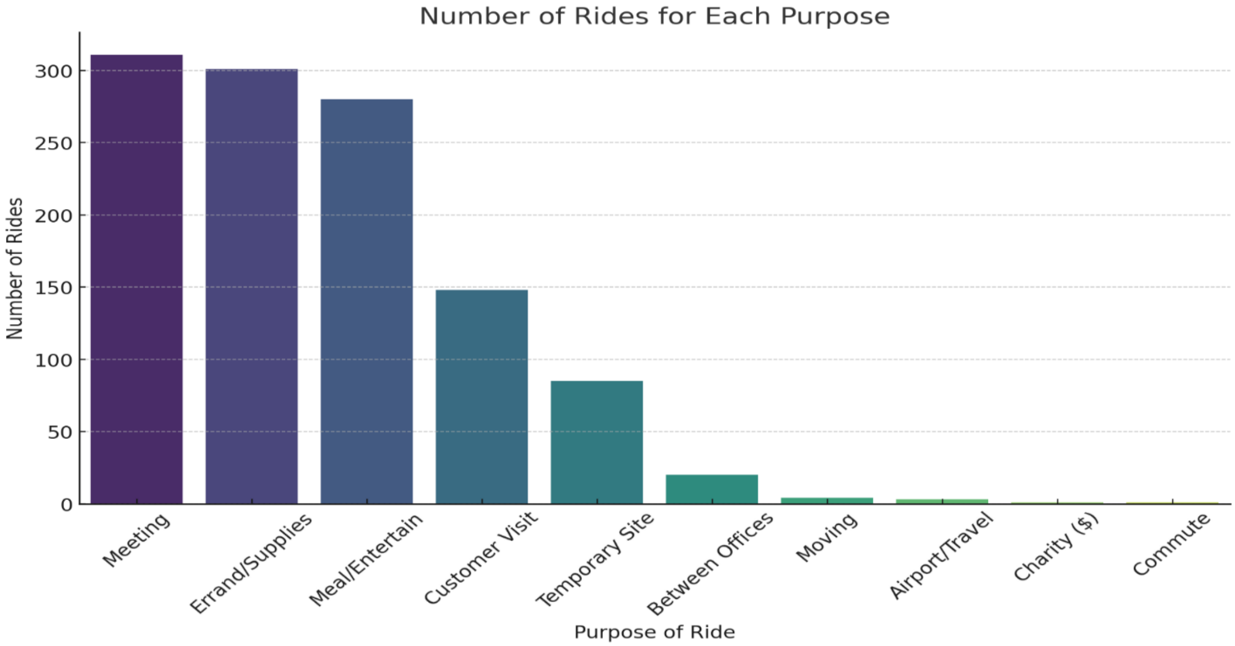


Figure 5 shows why people are taking rides, for work or shopping. This helps us understand what kind of trips are most common. All these charts together help us get a good picture of when and why people need rides, which is important for planning and improving our ride service.



5. Modeling Details:

Linear Regression: Serves as a baseline model. It is straightforward and provides an understanding of linear relationships in the data.

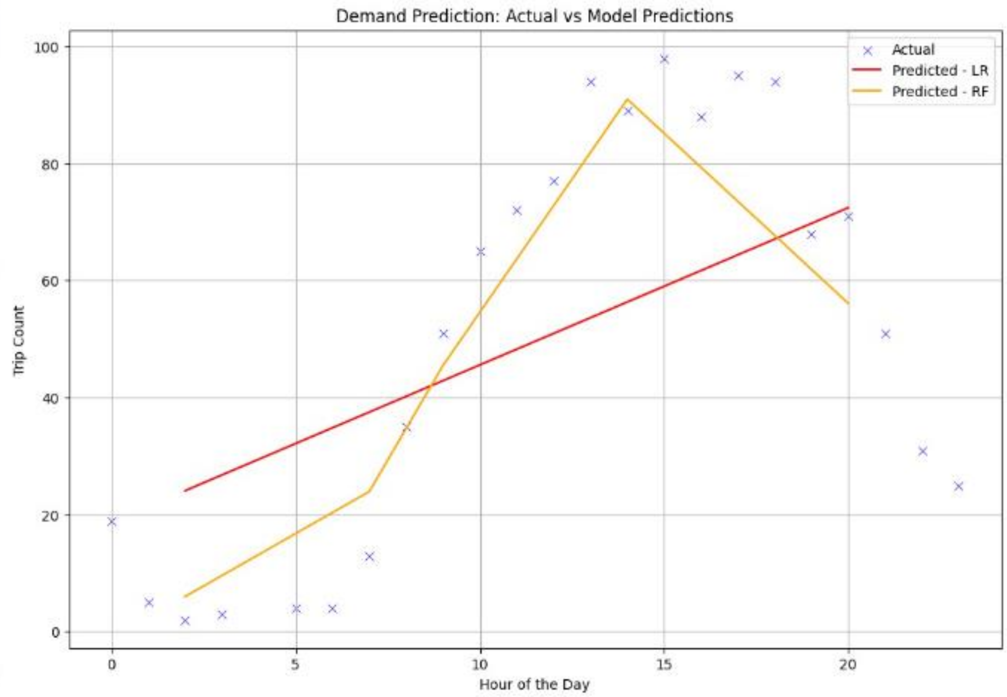
Expectation: We expect this model to capture basic trends in demand across different hours. However, its performance might be limited in handling complex patterns or non-linear relationships.

Random Forest: An advanced model compared to Linear Regression. It can handle non-linearities and complex interactions between features.

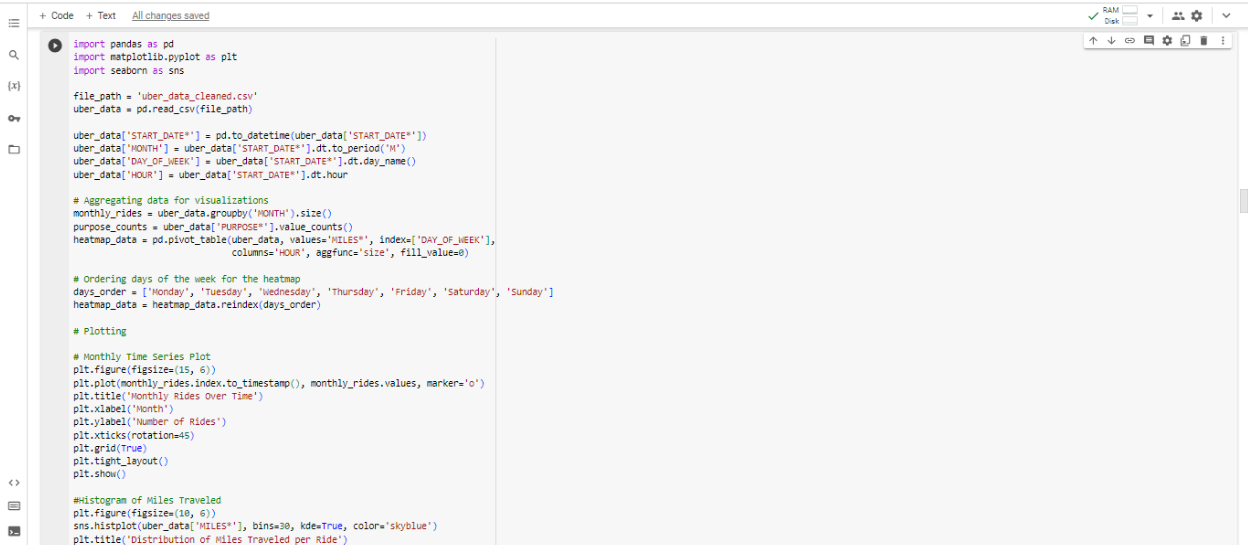
Expectation: This model is anticipated to provide a more accurate forecast of demand, especially in capturing variations that the linear regression may miss. It should perform better in terms of error metrics like RMSE (Root Mean Squared Error)

Results:

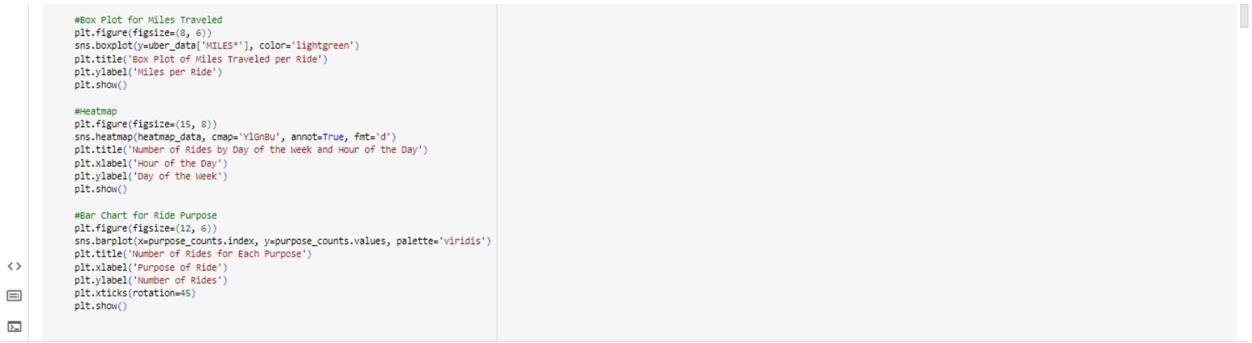
We have compared the performance of both models based on metrics such as RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error). The Linear Regression model provided a basic understanding and a baseline performance level. The Random Forest model showed superior performance due to its complexity and ability to handle diverse data patterns. Visualizations have been created to compare the actual demand data against the predictions made by each mode.



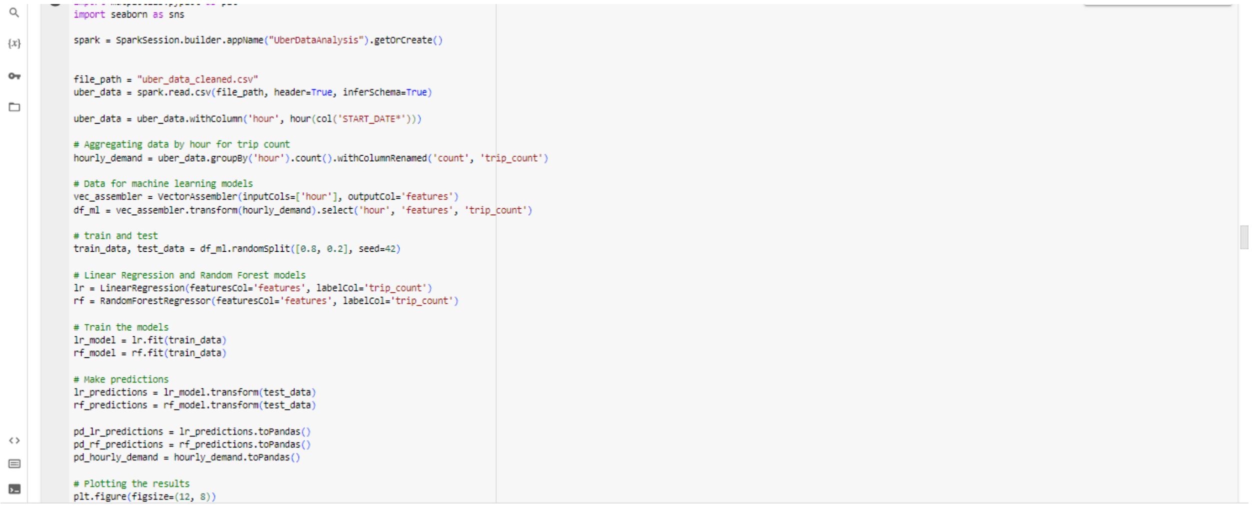
6. **Code Overview**:



```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file_path = 'uber_data_cleaned.csv'
uber_data = pd.read_csv(file_path)

uber_data['START_DATE*'] = pd.to_datetime(uber_data['START_DATE*'])
uber_data['MONTH'] = uber_data['START_DATE*'].dt.to_period('M')
uber_data['DAY_OF_WEEK'] = uber_data['START_DATE*'].dt.day_name()
uber_data['HOUR'] = uber_data['START_DATE*'].dt.hour

# Aggregating data for visualizations
monthly_rides = uber_data.groupby('MONTH').size()
purpose_counts = uber_data['PURPOSE*'].value_counts()
heatmap_data = pd.pivot_table(uber_data, values='MILES*', index=['DAY_OF_WEEK'],
                              columns='HOUR', aggfunc='size', fill_value=0)

# Ordering days of the week for the heatmap
days_order = ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']
heatmap_data = heatmap_data.reindex(days_order)

# Plotting

# Monthly Time Series Plot
plt.figure(figsize=(15, 6))
plt.plot(monthly_rides.index.to_timestamp(), monthly_rides.values, marker='o')
plt.title('Monthly Rides Over Time')
plt.xlabel('Month')
plt.ylabel('Number of Rides')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()

#Histogram of Miles Traveled
plt.figure(figsize=(10, 6))
sns.histplot(uber_data['MILES*'], bins=30, kde=True, color='skyblue')
plt.title('Distribution of Miles Traveled per Ride')
```



```
#Box Plot for Miles Traveled
plt.figure(figsize=(8, 6))
sns.boxplot(y=uber_data['MILES*'], color='lightgreen')
plt.title('Box Plot of Miles Traveled per Ride')
plt.ylabel('Miles per Ride')
plt.show()

#Heatmap
plt.figure(figsize=(15, 8))
sns.heatmap(heatmap_data, cmap='YlGnBu', annot=True, fmt='d')
plt.title('Number of Rides by Day of the Week and Hour of the Day')
plt.xlabel('Hour of the Day')
plt.ylabel('Day of the Week')
plt.show()

#Bar Chart for Ride Purpose
plt.figure(figsize=(12, 6))
sns.barplot(x=purpose_counts.index, y=purpose_counts.values, palette='viridis')
plt.title('Number of Rides for Each Purpose')
plt.xlabel('Purpose of Ride')
plt.ylabel('Number of Rides')
plt.xticks(rotation=45)
plt.show()
```



```
import seaborn as sns

spark = SparkSession.builder.appName("UberDataAnalysis").getOrCreate()


file_path = "uber_data_cleaned.csv"
uber_data = spark.read.csv(file_path, header=True, inferSchema=True)

uber_data = uber_data.withColumn('hour', hour(col('START_DATE*')))

# Aggregating data by hour for trip count
hourly_demand = uber_data.groupBy('hour').count().withColumnRenamed('count', 'trip_count')

# Data for machine learning models
vec_assembler = VectorAssembler(inputCols=['hour'], outputCol='features')
df_ml = vec_assembler.transform(hourly_demand).select('hour', 'features', 'trip_count')

# train and test
train_data, test_data = df_ml.randomSplit([0.8, 0.2], seed=42)

# Linear Regression and Random Forest models
lr = LinearRegression(featuresCol='features', labelCol='trip_count')
rf = RandomForestRegressor(featuresCol='features', labelCol='trip_count')

# Train the models
lr_model = lr.fit(train_data)
rf_model = rf.fit(train_data)

# Make predictions
lr_predictions = lr_model.transform(test_data)
rf_predictions = rf_model.transform(test_data)

pd_lr_predictions = lr_predictions.toPandas()
pd_rf_predictions = rf_predictions.toPandas()
pd_hourly_demand = hourly_demand.toPandas()

# Plotting the results
plt.figure(figsize=(12, 8))
```

7. Conclusion:

The analysis revealed significant patterns in user behavior, such as preferred travel times, frequent destinations, or common trip distances. These insights can be useful for Uber in tailoring their services to better meet customer needs.

The findings can inform strategic business decisions for Uber. This includes optimizing the allocation of drivers during peak hours, adjusting pricing based on trip distances and durations. By understanding the trends and patterns in the data, uber can make informed decisions to enhance efficiency and customer satisfaction.

The project highlights how valuable data analysis is for making informed decisions in any business. It opens opportunities for using data in new ways, like predicting future trends or understanding customer needs better. It shows that using data wisely can lead to smarter business moves and better customer experiences.