TASK 3

ANALYZING PUBLIC BIKE SHARING RIDERSHIP

INTRODUCTION:

Public bike-sharing systems have become increasingly popular in urban areas, offering an eco-friendly and convenient transportation option. Analyzing public bike-sharing ridership helps in understanding usage patterns, optimizing bike distribution, and improving overall service efficiency. This project aims to analyze and visualize bike-sharing data to derive actionable insights using Python.

OBJECTIVE:

The primary objectives of this analysis are:

- **Understand Ridership Patterns:** Explore data to identify peak usage times, popular routes, and user demographics.
- **Optimize Bike Distribution:** Analyze bike availability and usage to suggest optimal bike distribution strategies.
- **Predict Future Trends:** Utilize predictive algorithms to forecast future ridership trends and demand.

Data:

The dataset used for this analysis includes information on bike trips, such as:

- **Trip Start Time:** When the bike trip started.
- **Trip End Time:** When the bike trip ended.
- **Duration:** Duration of the bike trip.
- **Start Station:** The station where the trip started.

- **End Station:** The station where the trip ended.
- **User Type:** Classification of the user (e.g., subscriber, customer).
- **Weather Conditions:** Weather data during the trip (optional, if available).

TOOLS AND LIBRARIES:

The analysis will be conducted using Python with the following libraries:

- **Pandas:** For data manipulation and analysis.
- **NumPy:** For numerical operations.
- **Matplotlib & Seaborn:** For data visualization.
- **Scikit-learn:** For machine learning algorithms and predictive modeling.

METHODOLOGY:

1. **Data Preprocessing:**
    - **Data Cleaning:** Handle missing values, correct inconsistencies, and remove duplicates.
    - **Feature Engineering:** Create new features from existing data (e.g., trip duration categories, peak hours).
2. **Exploratory Data Analysis (EDA):**
    - **Visualizations:** Generate plots to visualize trip durations, usage patterns, and station popularity.
    - **Descriptive Statistics:** Calculate summary statistics to understand data distribution.
3. **Pattern Analysis:**
    - **Time-Series Analysis:** Identify trends and seasonal patterns in ridership data.

- **Geospatial Analysis:** Map bike-sharing stations and analyze usage by location.

4. **Predictive Modeling:**

   - **Algorithm Selection:** Use regression models (e.g., Linear Regression, Random Forest) to predict future ridership based on historical data.

   - **Model Evaluation:** Assess model performance using metrics like RMSE (Root Mean Squared Error) and R^2 score.

5. **Optimization:**

   - **Bike Distribution:** Analyze usage data to suggest optimal bike distribution strategies and locations for new stations.

ALGORITHMS:

1. **Time-Series Forecasting:**

   - **ARIMA (AutoRegressive Integrated Moving Average):** For forecasting future ridership based on historical trends.

2. **Classification Algorithms:**

   - **K-Means Clustering:** For segmenting users based on their trip patterns and behaviors.

   - **Logistic Regression:** For predicting the probability of a user being a subscriber vs. a customer.

3. **Regression Algorithms:**

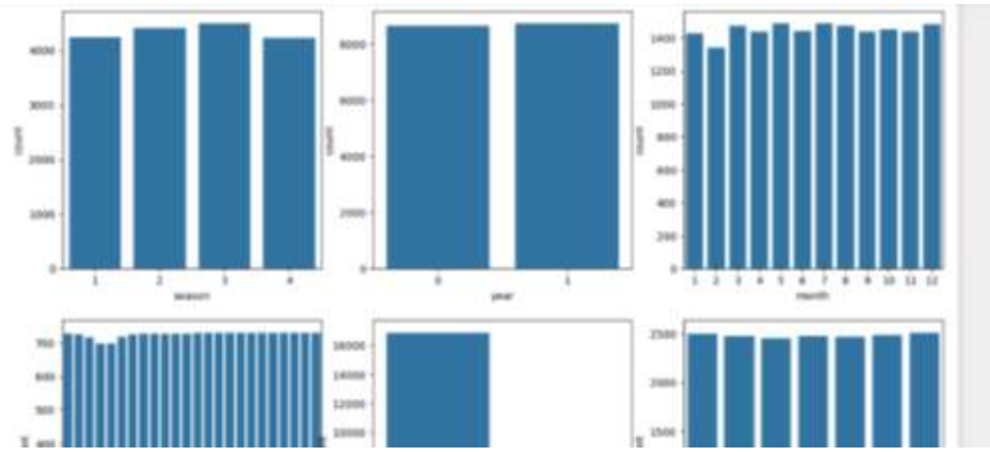
   - **Linear Regression:** For modeling the relationship between ridership and time-related features.

- **Random Forest Regression:** For capturing non-linear relationships and interactions in the data.

DATASET USED:

https://www.kaggle.com/datasets/lakshmi25npathi/bike-sharing-dataset?select=hour.csv

OUTPUT:



```
    data = data.drop(columns = [column],axis=1)
    return data
for col in cols:
    bike_dataset_oh = one_hot_encoder(bike_dataset_oh,col)
```

```
bike_dataset_oh.head()
```

	temp	atemp	humidity	windspeed	casual	registered	count	season_2	season_3	season_4	month_2	month_3	month_4	month_5	month_6	month_7	month_8
0	0.24	0.2879	0.81	0.0	3	13	16	False	False	False	False	False	False	False	False	False	False
1	0.22	0.2727	0.80	0.0	8	32	40	False	False	False	False	False	False	False	False	False	False
2	0.22	0.2727	0.80	0.0	5	27	32	False	False	False	False	False	False	False	False	False	False
3	0.24	0.2879	0.75	0.0	3	10	13	False	False	False	False	False	False	False	False	False	False
4	0.24	0.2879	0.75	0.0	0	1	1	False	False	False	False	False	False	False	False	False	False

```
# Split data into input and label data
X = bike_dataset_oh.drop(columns=['atemp','windspeed','casual','registered','count'],axis=1)
Y = bike_dataset_oh['count']
```

```
# Split data into train and test data
x_train,x_test,y_train,y_test = train_test_split(X,Y,train_size=0.7)
print(X.shape,x_train.shape,x_test.shape)
print(Y.shape,y_train.shape,y_test.shape)
```

```
(17379, 50) (12165, 50) (5214, 50)
```

CONCLUSION:

By leveraging Python and various algorithms, this project aims to deliver a comprehensive analysis of public bike-sharing ridership. The insights derived will be valuable for stakeholders in making data-driven decisions to enhance the efficiency and effectiveness of bike-sharing systems.