Assignment: Linear and Logistic Regression

Objective:
This assignment aims to solidify your understanding of linear and logistic regression, two essential techniques used by data analysts to understand relationships and make predictions based on data. You will work with real-world datasets and apply Python to perform regression analysis, evaluate models, and interpret results in the context of business or data-driven decisions.

Part 1: Linear Regression

Context:
You are working as a data analyst for a transportation agency in X. Your task is to analyze factors affecting taxi fares and build a predictive linear regression model to estimate trip fares.

Dataset:
A CSV file (`trip.csv`) containing the following columns:

- **Trip Start Timestamp:** When the trip started, rounded to the nearest 15 minutes.
- **Trip End Timestamp:** When the trip ended, rounded to the nearest 15 minutes. ● **Trip Seconds:** Time of the trip in seconds.
- **Trip Miles:** Distance of the trip in miles.
- **Fare:** The fare for the trip.
- **Tips:** The tip for the trip. Cash tips generally will not be recorded.
- **Tolls:** The tolls for the trip.
- **Extras:** Extra charges for the trip.
- **Trip Total:** Total cost of the trip, the total of the previous columns.
- **Payment Type:** Type of payment for the trip.

Tasks:

1. **Data Preprocessing**
   a. Load the dataset and inspect its structure.
   b. Check for and handle missing values, if any.
   c. Perform exploratory data analysis (EDA) to understand relationships between variables and then perform feature engineering.
2. **Model Building and Evaluation**
   a. Split the dataset into training and test sets (e.g., 80% training, 20% testing).
   b. Build a linear regression model to predict fares.
   c. Evaluate the model using metrics like:
      1. Mean Absolute Error (MAE)

2. Mean Squared Error (MSE)
3. R^2 score
4. Adjusted R^2 score

3. **Insights and Recommendations**
    1. **Key Findings**
        - Summarize the key insights of taxi fares based on your analysis.
        - Identify trends or anomalies in the data, such as peak hours for high fares or anything associated with higher/lower fares.
        - Identify the most significant features influencing fares.
    2. **Recommendations**
        - Provide actionable recommendations for stakeholders, such as optimizing pricing strategies or planning for peak demand.

Part 2: Logistic Regression

Context:
You are a data analyst working for Airline X. Your task is to predict customer satisfaction based on various demographic and behavioral attributes. The goal is to identify patterns in customer satisfaction levels and provide actionable insights that the airline can use to improve its services and enhance customer experiences.

Dataset:
A CSV file (`airlines.csv`) containing the following columns:

- **Satisfaction**: Indicates the satisfaction level of the customer.
- **Age**: Age of the customer.
- **Customer Type**: Type of customer: 'Loyal Customer' or 'Disloyal Customer'.
- **Type of Travel**: Purpose of travel: 'Business travel' or 'Personal Travel'.
- **Class**: Class of travel: 'Business', 'Eco', or 'Eco Plus'.
- **Flight Distance**: Distance of the flight in kilometers.
- **Seat Comfort**: Rating of seat comfort provided during the flight (1-5).
- **Departure/Arrival Time Convenient**: Rating of the convenience of departure/arrival time (1-5).
- **Food and Drink**: Rating of food and drink quality provided during the flight (1-5).
- **Gate Location**: Rating of gate location convenience (1-5).
- **Inflight Wifi Service**: Rating of inflight wifi service satisfaction (1-5).
- **Inflight Entertainment**: Rating of inflight entertainment satisfaction (1-5).
- **Online Support**: Rating of online customer support satisfaction (1-5).
- **Ease of Online Booking**: Rating of ease of online booking satisfaction (1-5).
- **On-board Service**: Rating of on-board service satisfaction (1-5).

- **Leg Room Service**: Rating of leg room service satisfaction (1-5).
- **Baggage Handling**: Rating of baggage handling satisfaction (1-5).
- **Check-in Service**: Rating of check-in service satisfaction (1-5).
- **Cleanliness**: Rating of cleanliness satisfaction (1-5).
- **Online Boarding**: Rating of online boarding satisfaction (1-5).
- **Departure Delay in Minutes**: Total departure delay in minutes.
- **Arrival Delay in Minutes**: Total arrival delay in minutes.

Tasks:

1. **Data Preprocessing**
   a. Load the dataset and inspect its structure.
   b. Check for and handle missing values, if any.
   c. Perform EDA to understand the distribution of features and their relationship with the target variable, and then then perform feature engineering.

2. **Model Building and Evaluation**
   a. Split the dataset into training and test sets.
   b. Build a logistic regression model.
   c. Evaluate the model using metrics such as:
      - Accuracy score
      - Confusion matrix
      - Precision, Recall, and F1-score (for each class)
      - ROC-AUC score and ROC Curve
4. **Insights and Recommendations**
   1. **Key Findings**
      - Summarize the key insights of airline customer satisfaction based on your analysis.
      - Summarize the factors most strongly associated with customer satisfaction based on model coefficients.
   2. **Recommendations**
      - Suggest improvements to specific areas such as check-in processes, in-flight service, or punctuality.

Submission Requirements

Submit a zip file that should include two Jupyter notebooks (for part 1 and part 2) with your code, visualizations, and outputs for both parts. Ensure your code is well-commented and readable by using markdown in your notebooks.