EE 590 Basics of Data Analysis and Machine Learning
Assignment #3: Flight Delay Analysis

Due Date: Tuesday, October 24, 2023

Objective:
Explore the factors affecting flight delays using Exploratory Data Analysis (EDA) techniques and Python programming.

Background:
Flight delays can be a major concern for airlines and passengers. By analyzing the factors causing delays, airlines can optimize their operations, and passengers can make informed travel choices. In this project, you will dive into a dataset containing flight delay information to uncover patterns and insights.

Dataset Details:

FLIGHT_NUMBER:	Unique Identifier for each flight
MONTH:	The month of the flight
DAY:	The day of the month of the flight
DAY_OF_WEEK:	The day of the week of the flight
Airline:	Airline operating the flight
ORIGIN_AIRPORT:	Departure airport
DESTINATION_AIRPORT:	Arrival airport
SCHEDULED_DEPARTURE:	Scheduled departure time
DEPARTURE_TIME:	Actual departure time
SCHEDULED_ARRIVAL:	Scheduled arrival time
ARRIVAL_TIME:	Actual arrival time
ARRIVAL_DELAY:	Delay in minutes

Project Tasks:

1. **Data Pre-processing:**
    - Load the flight delay dataset.
    - Handle missing values.
    - Provide a statistical summary of the dataset (mean, median, mode, standard deviation, etc.).
    - Understand the shape and structure of the dataset (number of rows/columns, data types).

2. **Univariate Analysis:**
    - Plot the distribution for each feature using histograms or density plots.
    - If there are categorical variables, showcase the frequency of each category using bar plots.
    - Describe any patterns or anomalies you observe from the univariate analysis.

3. **Initial Data Exploration:**
    - Check the first few rows to understand the dataset.
    - Get summary statistics for each column.
    - Identify the percentage of flights delayed by more than 15 minutes.

4. **Visual Analysis:**
    - Visualize the distribution of flight delays.
    - Compare average delays across different airlines.
    - Investigate the impact of the day of the week on delays.
    - Analyze if there's any pattern of delays based on the scheduled departure hour.
    - Understand the correlation between the various numeric variables.
    - Check if there's any seasonality in delays.

5. **Deep Dive Analysis:**
    - Identify the origin-destination pairs with the highest average delays.
    - Analyze the 10 most frequent flight routes and their average delays.
    - Visualize the relationship between the origin and destination airports using a heatmap.

6. **Insights & Recommendations:**
    - Document the insights derived from the visualizations and deep-dive analysis.
    - Provide recommendations for airlines to reduce delays.
    - Advise passengers on the best times to fly or airlines to pick to avoid delays based on your findings.

7. **Bonus Task (for advanced students):**
    - Use a simple linear regression model to predict flight delays based on the features in the dataset. Evaluate the performance of your model.
    - Apply feature engineering to improve the model's performance.

Submission Details:

- A Jupyter Notebook or Python script containing all the code for the above tasks.
- A report (PDF/Word) summarizing your findings, including relevant visualizations.
- If attempting the bonus task, include your model performance metrics in the report.

Evaluation Criteria:

1. Completeness and accuracy of the analysis.
2. Quality and relevance of visualizations.
3. Depth and clarity of insights drawn from the dataset.
4. Quality of the final report and recommendations provided.