PySpark Assignment in Databricks: Analyzing and Modeling Flight Delays Data

Objective: This assignment will guide students through an end-to-end data analysis project using PySpark in Databricks. Students will load, clean, and explore a dataset, perform SQL queries and data analysis using Spark SQL, and build a predictive model to classify or regress based on flight delays.

Dataset: Use the **Flight Delays and Cancellations** dataset from Kaggle or a similar open dataset. This dataset contains information on flight delays, cancellations, and causes from different airlines, which makes it suitable for exploration, cleaning, and analysis using PySpark.

Step 1: Data Loading and Initial Exploration

1. **Load Data**: Import the Flight Delays dataset into Databricks using PySpark DataFrames.
    - Upload the dataset file (e.g., flights.csv) to the Databricks environment.
    - Load the data as a Spark DataFrame.
2. **Initial Exploration**: Perform a preliminary analysis to understand the data.
    - Display the schema of the DataFrame to understand the column types.
    - Show a sample of the data (e.g., the first 5-10 rows).
    - Count the number of rows and columns.

Questions:

- How many rows and columns does the dataset contain?
- What are the data types of each column?

Step 2: Data Cleaning and Transformation

1. **Handle Missing Values**: Identify columns with missing values and decide how to handle them.
    - Drop rows or columns with excessive missing values.
    - Impute missing values if applicable (e.g., fill with mean or median for numerical columns)
2. **Feature Engineering**: Create new features that may be useful for analysis.
    - Convert the FLIGHT_DATE column to extract features like DAY_OF_WEEK or MONTH.
    - Create binary columns indicating if a flight was delayed (ARRIVAL_DELAY > 0).
3. **Data Type Conversion**: Convert columns to appropriate data types if necessary.

Questions:

- Which columns had missing values, and how did you handle them?
- What new features did you create, and why?

Step 3: Exploratory Data Analysis (EDA) using Spark SQL

1. **Register DataFrame as a Temporary Table**: Register the DataFrame as a temporary table to use SQL queries.

2. **SQL Queries**: Perform the following analyses using Spark SQL:
    - Calculate the average delay time for each airline.
    - Identify the top 5 airports with the most delayed departures.
    - Determine the most common reason for flight cancellations.
3. **Visualization**: Use Databricks visualizations or matplotlib to visualize key findings.
    - Plot the average delay by airline.
    - Visualize delay patterns over days of the week or months.

    Questions:
    - Which airlines had the highest average delays?
    - What patterns did you observe in delays by day of the week?

Step 4: Conclusion and Documentation

1. **Summary**: Write a brief summary of your findings.

    - Summarize key insights from the EDA and SQL queries.
2. **Documentation**: Provide clear documentation of each step and any code you used.

Submission Requirements

- Submit the Databricks notebook file containing all code and outputs.
- Answer the questions in each step in markdown cells within the notebook.
- Include any visualizations generated during the analysis.