Akasa Air

IT'S YOUR SKY

Interview Task – Data Engineering & Analytics

Akasa Air
IT'S YOUR SKY

CONTENTS

OBJECTIVE

You are provided with a sample dataset from the aviation industry. Your task is to clean, normalize, and analyze the data to derive relevant insights. The dataset includes various details about flights, such as flight numbers, dates, times, airlines, and delays.

DETAILED REQUIREMENTS

1. DATASET ISSUES AND REQUIREMENTS

a. Inconsistent Date and Time Formats:
   i. DepartureDate and ArrivalDate are in MM/DD/YYYY format.
   ii. DepartureTime and ArrivalTime are in HH:MM AM/PM format.
b. Missing Values:
   i. The DelayMinutes column contains NaN values that need to be handled.
c. Duplicate Flight Entries:
   i. There are duplicate entries for some flights which may need to be addressed.
d. Inconsistent Time Entries:
   i. The ArrivalTime for one entry is later than DepartureTime on the same day which might indicate a data error (e.g., AA1234 on 09/01/2023 with a departure at 08:30 AM and arrival at 10:45 PM).

2. DATA CLEANING:

a. Identify and handle any missing or inconsistent values in the dataset.
b. Ensure all column data types are appropriate (e.g., dates as date types, times as time types).
c. Correct any inconsistencies or errors in times (e.g., arrival time should be later than departure time).

3. DATA NORMALIZATION:

a. Convert DepartureDate and ArrivalDate columns to a standard YYYY-MM-DD format.
b. Convert DepartureTime and ArrivalTime columns to a 24-hour time format (e.g., "08:30" for 8:30 AM).

c. Create a new column for FlightDuration by calculating the difference between DepartureTime and ArrivalTime on the same day.

4. Data Analysis:

a. Analyze the distribution of delays and identify any trends or patterns.
b. Calculate the average delay for each airline.
c. Identify any relationships between flight delays and departure times (e.g., are flights departing later in the day more likely to be delayed?).
d. Determine if there is a significant difference in delays between different airlines.

5. Insights:

a. Provide a summary of the key findings from the data.
b. Analyze the impact of departure times on delays.
c. Compare delay distributions between airlines.
d. Visualize the average delay by airline and the delay distribution using appropriate charts.
e. Offer recommendations based on the analysis, such as suggestions for airlines to improve their punctuality.

6. Documentation:

a. Provide detailed documentation on how to set up and execute the pipeline.

Guidelines

- You are free to choose any programming language, framework, or technology stack for the implementation. (Python, SQL would be good to have)
- It is recommended to use a database to store data.
- Focus on code quality, maintainability, and scalability of the pipeline.
- You may use third-party libraries or frameworks, if necessary, but clearly mention any dependencies.

Deliverables

- Cleaned and Normalized Dataset: Submit the transformed dataset in a CSV format.
- Analysis Report: A detailed report including:
  - Summary of data cleaning and normalization steps.
  - Insights derived from the data analysis.
  - Visualizations (e.g., bar charts, histograms) illustrating the key findings.
  - Recommendations based on the analysis.
- Code: Provide the code used for data cleaning, normalization, and analysis. Ensure it is well-documented and reproducible.

Note: We are looking for your understanding of the problem, your approach to solving it, and your ability to design and implement a basic version of the data pipeline, **with all requirements met**.

Please complete the task within the given time frame and submit the deliverables for evaluation. All the Best!

Sample Data set

File Name: aviation_data.csv

FlightNumber,DepartureDate,DepartureTime,ArrivalDate,ArrivalTime,Airline,DelayMinutes

AA1234,09/01/2023,08:30 AM,09/01/2023,10:45 AM,American Airlines,15

DL5678,09/01/2023,01:15 PM,09/01/2023,03:30 PM,Delta,5

UA9101,09/01/2023,05:00 PM,09/01/2023,07:15 PM,United Airlines,25

AA1234,09/01/2023,08:30 AM,09/01/2023,10:45 PM,American Airlines,30

DL5678,09/02/2023,02:00 PM,09/02/2023,04:10 PM,Delta,NaN

UA9101,09/02/2023,05:00 PM,09/02/2023,07:15 PM,United Airlines,20

AA1234,09/02/2023,08:30 PM,09/03/2023,10:45 AM,American Airlines,60

DL5678,09/03/2023,01:00 PM,09/03/2023,03:30 PM,Delta,10

UA9101,09/03/2023,03:00 PM,09/03/2023,05:20 PM,United Airlines,NaN

AA1234,09/03/2023,08:30 AM,09/03/2023,10:00 AM,American Airlines,15

DL5678,09/04/2023,12:30 PM,09/04/2023,02:40 PM,Delta,25

UA9101,09/04/2023,07:00 PM,09/04/2023,09:15 PM,United Airlines,45

Thank You!