Technical Test on Data Quality

This document contains two use cases, with the associated data sets. The candidate is free to choose one of them.

As part of this technical assessment, we invite you to work on a set of simulated data relating to flight tracking.

The goal is to allow us to assess your knowledge of Python and the Pandas library for manipulating and analysing complex data from multiple data sources and moderate volumes.

Please do not hesitate to contact us if you need further details during implementation.

We would appreciate any initiative to offer additional analyses or relevant visualizations that would highlight interesting aspects of the data provided.

In order to allow us to best assess your skills, **we ask you to send us your Python project the day before the interview**.

We invite you to provide us with structured and commented code (imagine that another colleague had to make changes).

During the interview, you will have the opportunity to present your solution for about ten minutes, followed by a Q&A session of about five minutes.

Please read the following pages for use-cases descriptions.

USECASE N°1

You have several files provided:

- A master file (*flights.csv*) containing the tracking data of flights operated over a day, with detailed information on each flight (identifier, departure and arrival airports, scheduled and actual schedules, type of aircraft, status).

- A file (*scope.csv*) listing the specific flight identifiers on which further analysis is expected.

- A file (*weather.csv*) of the weather conditions recorded at a certain time for certain airports.

- An aircraft maintenance file (*maintenance.csv*), detailing the latest operations carried out on different types of aircraft.

For each question below, your main task is to develop a Python project using the Pandas library, to analyze this data. Please do save your code (in a Jupyter Notebook, for example) to be able to share it during the discussion of your work.

Specifically, we expect you to be able to:

- Load all of these CSV files, making sure you **understand the types of data they contain**
- Filter flight tracking data to **keep only those with identifiers in the scope.csv file**
- For these targeted flights, **calculate the departure and arrival delays** (difference between the actual time and the scheduled time). You will then need to **identify the flights that have experienced a significant delay** (e.g. more than 30 minutes) **and extract the list**.
- Integrate weather data with flight information. The goal is to be able **to associate each flight with its weather conditions at departure**, in order to explore a possible correlation between weather conditions (strong wind, low visibility, rain, fog) and observed delays.
- Attach aircraft maintenance data to the corresponding flights. This will make it possible **to analyse whether any flights have taken place shortly after major maintenance on the aircraft**, and to identify these cases.

USECASE N°2

You have several files provided:

- flights.gz: consisting of flights at the start of December
- capacity.txt: consisting of capacity in flights at the start of December
- revenues.bz2: consisting of flight revenues at the start of December

For each question below, your main task is to develop a Python project using the Pandas library, to analyze this data. Please do save your code (in a Jupyter Notebook, for example) to be able to share it during the discussion of your work.

Questions:

1. Read the capacity.txt file. Are there any issues reading this file? If so, how did you solve them?
2. What are the different columns and types of columns in the capacity.txt? If you have a column that is a different type than expected, why is that the case and can you fix the data/original file so that each column is formatted how you expect? (hint: check the values in the columns for outliers)
3. Are there any dates that seem out of place compared to the rest of the file?
4. Read in the revenue file (hint, check the extension & consider which separator to use). What is the total revenue over the first week of December?
5. Read in the flights file (hint: check both the extension and consider file encoding). The headers are *flight_number, carrier, origin, desintation, plane, departure_date*. What are the different types of planes?
6. All files can be merged based on flight number and departure date. What is the capacity (seats) of a Boeing 777-300? And what is the average number of passengers (pax) the first week of December between Amsterdam (AMS) and New York (JFK) going both directions?

7. AF and KL are the carrier symbols of Air France-KLM. Assume company only services those flights. What is the revenue of AF flights on the 5th of December? What is the revenue for all KL flights? What is the revenue of all AF flights? Is there anything odd about the summation of the carrier revenues? (Hint: about your full investigation up till now)
8. Create a new (compressed) file consisting of all the AF and KL flights you have, with their respective seats, pax, revenue and BLF. BLF (Book-load-factor) = pax/seats. How did you handle missing data?
9. Assuming the flights, capacity and revenues files are tables in a SQL database like PostgreSQL. The column/data formatting issues have been solved before creating these tables. Create an SQL query to get the mean and standard deviation of all Flights with a capacity (seats) above 290.