Data Engineering Challenge

Introduction

We want to thank you for your time to work on this challenge. As a part of the recruitment process, we would like to ask you to complete the following assessment.

Please consider that we encourage you to write clean, maintainable and testable code.

Tech Stack :

- Python (>= 3.7)
- SQL

Feel free to contact us if you need (***erengul.bayram@guidion.net***)

Dimensional Modeling

You will be working on the Airbnb Ratings Dataset. You need to use following datasets :

- https://www.kaggle.com/datasets/samyukthamurali/airbnb-ratings-dataset?select=LA_Listings.csv
- https://www.kaggle.com/datasets/samyukthamurali/airbnb-ratings-dataset?select=airbnb-reviews.csv

1. Make yourself familiar with the dataset.
2. Design a dimension model of the fact and dimension tables according to these datasets for storing data in a data warehouse. Determine relationships between the tables.
3. Write an application to load unstructured data and transform it into a schema that you design. Store transformed data into csv files. (***You can work with sample data, and don't need to upload the whole dataset. Please send also your sample data***)
4. Evaluate data quality of data
5. If it is needed, how would you like to use partitions in your tables and why?
6. Write the SQL queries to answer following questions :
   a. What are the top 5 reviewer_id's who made the most reviews?
   b. In which month were the most reviews done by reviewers?
   c. Define relation between host_response_rate and host_is_superhost if you can find.
7. How would you like to design a data pipeline using this data from beginning to reaching the end user? Which technologies would you like to use?