Data Engineering Challenge

Desired Output: A Document, with all the steps necessary to execute the proposed solution and conclusions derived from data analysis.

- Load Script(s)
- SQL Script(s)
- Text file(s) with the conclusions.

Exercise 1: Loading Data

Programmatically (Using Python/R or other programming language) retrieve JSON data from europe.eu (link below) and load the data to a PostgreSQL Database (install it locally).

Data source 1: Covid Data: https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19 (use the Json file)

Load the CSV table Countries of the word.

Data source 2: Countries data: https://www.kaggle.com/fernandol/countries-of-the-world/data?select=countries+of+the+world.csv

Alternative: Load the data manually (using some importing tool).

Exercise 2: Create a Pipeline

Create a Data Pipeline that extracts the last version of the data (Covid-19 data, Data Source 1) and adds to the PostgreSQL database only the new records. (Note the Data source 1, Covid-19 dataset, changes one time per day)

Exercise 3: Create a View

Create a view with the data of the table "Countries of the word" with the latest number of cases, "Cumulative_number_for_14_days_of_COVID-19_cases_per_100000" and date when the Information was extracted.

Exercise 4: Queries

1- What is the country with the highest number of Covid-19 cases per 100 000 Habitants at 31/07/2020?
2- What is the top 10 countries with the lowest number of Covid-19 cases per 100 000 Habitants at 31/07/2020?
3- What is the top 10 countries with the highest number of cases among the top 20 richest countries (by GDP per capita)?
4- List all the regions with the number of cases per million of inhabitants and display information on population density, for 31/07/2020.
5- Query the data to find duplicated records.
6- Analyze the performance of all the queries and describes what you see. Get improvements suggestions.

Exercise 5: Enrich the information with any other piece of data (available on the web) and justify the choice.

Exercise 6: Produce a report showing the most useful discoveries you have made. Graphical representation is welcome.

Optionally find some patterns in the Covid-19 Dataset using the enriched data. You can use Python/R/Excel/etc**.**

Countries of the World

Acknowledgements

Source: All these data sets are made up of data from the US government. Generally they are free to use if you use the data in the US. If you are outside of the US, you may need to contact the US Govt to ask.
Data from the World Factbook is public domain. The website says "The World Factbook is in the public domain and may be used freely by anyone at anytime without seeking permission."
https://www.cia.gov/library/publications/the-world-factbook/docs/faqs.html
https://www.kaggle.com/fernandol/countries-of-the-world/data?select=countries+of+the+world.csv