Data Engineering Challenge

Proposed by Dell Technologies

Solved by Lucas Webber Molin

December 29, 2021 - January 2, 2022

Exercise 1 - Loading Data

First, I accessed the websites provided in the challenge documentation, and analyzed the JSON and CSV files to understand what I had in hand. Then I created a PostgreSQL database locally on my computer, and a table for each of the two files. I sent the CSV file to the countries_of_the_world_csv database table using the DBeaver application, where I also wrote the queries used in this challenge. For the JSON file, I created a Python project through the PyCharm IDE.

The script for creating the tables is attached under the name create_tables.sql in the SQL_Script folder. The data load is in the covid19_insert.sql and countries_of_the_world_csv_insert.sql files also in the SQL_Script folder.

Exercise 2 - Create a Pipeline

In the main.py script, attached to the Load_Script folder, I wrote a code to search the JSON file from the website and insert it into the covid19 database table. On the website, it says that the file is updated weekly, so to add only new data I implemented a condition to only insert records of dates longer than the longest date already stored in the database. To avoid duplicate records, I created a unique constraint. To automate the extraction, a task could be created in the Windows task scheduler to run weekly.

I noticed that between the two data sources there were the same countries with the name spelled differently. I used the query that is in the full_outer_join_countries.sql file attached to the SQL_Script folder to identify which countries they were. I performed the correction manually, looking up the countries on Google, using the covid data source name as the main one.

Exercise 3 - Create a View

The script for creating the view is in the view_ex3_create.sql file attached to the SQL_Script folder. And in the same folder is the view data sample in the view_ex3_insert.sql file.

Exercise 4 - Queries

The queries written in SQL for this exercise are in the queries_ex4.sql file attached to the SQL_Script folder, along with the performance analysis.

Exercise 5 - Data from other sources

As an additional data source, I chose tourism data for each country to identify correlations between the number of visitors and the number of COVID-19 cases. Seeing that small countries like Aruba, for example, had many cases for a few inhabitants, I wondered if there was a correlation.

Source: https://onestep4ward.com/most-visited-countries-in-the-world/
The data were extracted manually.

The data load is in the tourism_countries_insert.sql file, along with the tourism_countries_create.sql table creation file, in the SQL_Script folder.

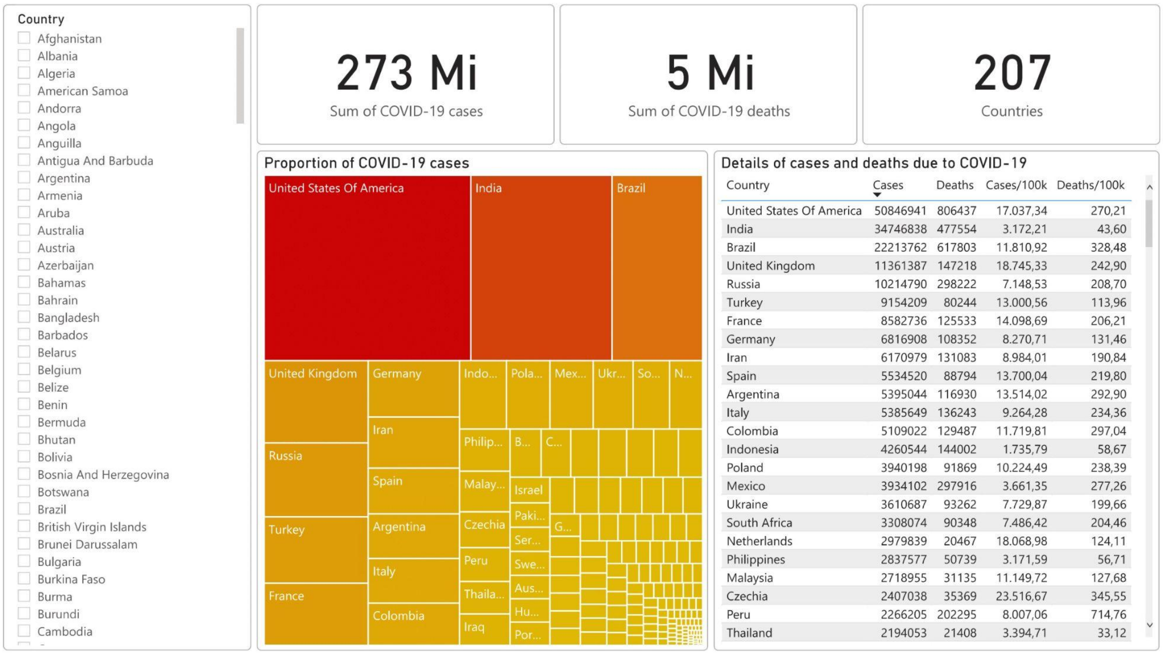
Pearson's correlation calculation was done in Python, according to the correlation.py file attached to the Data_Analysis folder, using the query from the view_cases_visitors.sql file attached to the SQL_Script folder. The coefficient result was 0.60, indicating a moderate correlation.

Exercise 6 - Report

To present the data, I chose to build a dashboard. The tool I use and master in my current job is protected by a confidentiality agreement, so I chose Power BI for this challenge. But I'm a beginner at Power BI. The project file is attached in the Data_Analysis folder of this zip under the name dell2021.pbix

I built a simple dashboard with general data for total cases and deaths and for every 100,000 inhabitants. The dashboard contains a filter that allows you to select multiple countries; totalizers that show the sum of the number of cases, deaths and the number of filtered countries; a treemap showing the proportion of cases among the filtered countries; and a table with details for each country, which can be sorted by the chosen column (country, number of cases, number of deaths, case rate and death rate for every 100,000 inhabitants.

See the dashboard in the image below or in the attached dell2021.pbix file in the Data_Analysis folder.



Country
☐ Afghanistan
☐ Albania
☐ Algeria
☐ American Samoa
☐ Andorra
☐ Angola
☐ Anguilla
☐ Antigua And Barbuda
☐ Argentina
☐ Armenia
☐ Aruba
☐ Australia
☐ Austria
☐ Azerbaijan
☐ Bahamas
☐ Bahrain
☐ Bangladesh
☐ Barbados
☐ Belarus
☐ Belgium
☐ Belize
☐ Benin
☐ Bermuda
☐ Bhutan
☐ Bolivia
☐ Bosnia And Herzegovina
☐ Botswana
☐ Brazil
☐ British Virgin Islands
☐ Brunei Darussalam
☐ Bulgaria
☐ Burkina Faso
☐ Burma
☐ Burundi
☐ Cambodia

273 Mi
Sum of COVID-19 cases

5 Mi
Sum of COVID-19 deaths

207
Countries

Proportion of COVID-19 cases

Details of cases and deaths due to COVID-19

Country	Cases	Deaths	Cases/100k	Deaths/100k
United States Of America	50846941	806437	17.037,34	270,21
India	34746838	477554	3.172,21	43,60
Brazil	22213762	617803	11.810,92	328,48
United Kingdom	11361387	147218	18.745,33	242,90
Russia	10214790	298222	7.148,53	208,70
Turkey	9154209	80244	13.000,56	113,96
France	8582736	125533	14.098,69	206,21
Germany	6816908	108352	8.270,71	131,46
Iran	6170979	131083	8.984,01	190,84
Spain	5534520	88794	13.700,04	219,80
Argentina	5395044	116930	13.514,02	292,90
Italy	5385649	136243	9.264,28	234,36
Colombia	5109022	129487	11.719,81	297,04
Indonesia	4260544	144002	1.735,79	58,67
Poland	3940198	91869	10.224,49	238,39
Mexico	3934102	297916	3.661,35	277,26
Ukraine	3610687	93262	7.729,87	199,66
South Africa	3308074	90348	7.486,42	204,46
Netherlands	2979839	20467	18.068,98	124,11
Philippines	2837577	50739	3.171,59	56,71
Malaysia	2718955	31135	11.149,72	127,68
Czechia	2407038	35369	23.516,67	345,55
Peru	2266205	202295	8.007,06	714,76
Thailand	2194053	21408	3.394,71	33,12

A hypothesis that I wanted to test is the correlation between the number of cell phones and the number of cases, because it is common to distribute fake news about COVID-19 through social networks. The result of Pearson's correlation calculation was 0.55, which indicates a weak correlation. I imagine that more specific data would be needed to identify whether there is a correlation between the amount of fake news spread and cases of COVID-19. The calculation was done in Python, according to the correlation.py file attached to the Data_Analysis folder, using the query from the view_cases_phones.sql file attached to the SQL_Script folder.

As requested, the conclusion of the data analysis is in the conclusion.txt text file in the Text_File_Conclusions folder.