ETL project

This project was focused on extracting, transforming, and loading data. In this case, coronavirus data was chosen, first, because the information is constantly been updated and also due to the importance of the situation at this moment. Having said this, the objective of the project was to successfully perform this stages, that is, extract data from different sources, transform this data for further manipulation and analysis, and load it to a database ready to pull it from here. Below is a description of each of the phases of the process.

Extract

To obtain data for the chosen topic an API was used. The request method was used and an API key was required as well. The response from the request method was a JSON() and it contained data from 214 countries with information about the number of deaths, cases, people recovered, new deaths, new cases, critical cases, active cases, cases per 1 million population, total tests and tests per 1 million population. Having this data, other methods were used to leave data ready for its analysis, which will be explained in the transform stage.

Transform

After we requested the data, we decided to gather all the countries’ info into groups by continent, G7, and G20. For this, we performed two methods.

1. Got the list of the country members by continent from another table hosted on the internet, then matched that list with the countries retrieved by the request.
2. For the smallest groups (G7, G20), we looked directly into the indexes of those countries in the request results.By the time we built the groups, we did some “country\_name” changes so both names in the list and the request matched. Eg. From “United Kingdom” to “UK”.Once all the groups were done, we had to transform all the numeric variables into integer values, since all the data was string datatype. For this, we created a function “replace(region\_list)” to do the task.

Load

The reason why Mongo was chose are not different from what has been discussed during the lessons. NoSQL allow you to handle large volumes of data either, structured or unstructured and also, because of its more frequent implementation in different companies. In this case, two methods were proposed to load the data to Mongo after the transformation was made:

1. The first one consisted of creating a single dictionary that contains each of the regions and groups. Then, the connection to mongo was established and the database was created. It should be mentioned, that inside of this database, only a single collection was created, and by using the update() method from pymongo, the dictionary with all the regions and groups was loaded. Also, it must be said that with this method the way to extract the data from the Mongo database is by using the find() method from pymongo, and then iterating on this element and using libraries like pandas.
2. The second method proposed was to create a collection for each region and group inside the database. Then, by using the method insert\_one() from pymongo, an iteration was made on each list to load the data. Unlike from the first method, the way to query documents, in this case, is by using directly pymongo query and projection operators