CSC343 Project Phase 2

Terry Tu, Owen Zhang

Friday, November 19, 2021

Design Decisions

At the beginning, we find the main key for every table to make them connect, then we change each table's name to let them represent their role in our research question and change their variable name to make them keep the same name for one variable and easy to understand. For example, we make the countryCode as the main key and change them as the same name for every table to connect them and we make the name in the covidImpact as total_cases_per_million, total_death_per_million, total_cases, total_deaths to let them easy for readers understand.

Feedbacks from TA in phase 1

1. Since TA make the comment that our division into relations avoids redundancy and unnecessary nulls and our division into relations is well justified throughout. Then we keep the design to divided these csv in to small tables depend on their relationships like country dictionary table, region dictionary table and sub-region dictionary table. Also further more we divided the covidImpact table into more pieces to make them more easier to clarify their role in our research question.
2. TA said in our phase 1 the only problem is that our research question could be drill down to deeper levels of detail and more diversified. Therefore we decide to add more muti-angle analysis the impact connection between the Covid-19 and countries with different characterized. We separate the covidImpact into three tables that for total death and total cases table, the health care level table and income level table. Which could make our research question in project more interesting and more comprehensively.

Justification of Design

In our project, We combined economic knowledge and geography knowledge to make our design more professional. We explore the unemployment rate from different angles, including the distinct economic entity in different continents, countries in various income levelsthe heath care facility level, and the severity that the country inflects during the pandemic.

We make the country code the key through datasets since the national country code is the same, which could help us connect each dataset accuracy by country. In the dataset of the Covid-19 Data, we keep the data of total cases for people suffering from and death by the Covid-19 and the data of the death and cases per million people, which can help us make the data more comprehensively.

We also separated the Country Information Data Set into three sub-data relations: country, Region, and Sub-Region. Since most of our question only requires the country code and country name. It is a good idea to separate the region and sub-region information from the Country table. When needed to combine the country and the country's continent, we can use the regionCode and subRegionCode to integrate with the Region and SubRegion table.

We only keep the income level column and country code in the income countries data set since the original data set containing a suit separate income group connected to our project questions.

For the unemployment data set, we will be using the country code and the country's unemployment rate for the years 2019 and 2020. We chose these two particular years since the year 2019 represents the unemployment rate of prepandemic, and 2020 represents the unemployment rate during the pandemic. Comparing these two years will give us the information we need to analyze the impact of COVID-19 on a country's unemployment rate.

Data Cleaning Process

1. We firstly downloaded all the CSV files from the links in our Phase 1 includes continents.csv, covid_data.csv, income_data.csv, unemployment_rate.csv

2. We make the data cleaning and rename in load_data.sql to make them connect with the same key countryCode. Also we removed all unnecessary null value so that they won't make the program exist error while running.

3. We removed the row for Antarctica in the continents.csv file, since this row is very special and it violates our key constraint. Also, this is only one row and it is not useful to our project.

4. In the schema.ddl we use "COPY" to copy these columns in the csv files we need in to our tables name import_...._data. These data tables are just temporary before our cleaning process.

5. Then, for the cleaning process we separated the Country Information Data Set into three sub-data relations: country, Region, and Sub-Region. We only keep the countries' unemployment rate of the year 2019 and 2020 and drop off data in other years. Remove the Region and Special Notes columns. Only keep the country code and the income group the country belongs to. For our research question we only keep the countries' unemployment rate of the year 2019 and 2020 and drop off data in other years. Also we rename the columns in this step in order to these columns have the same name in our schema.

6. We also deleted a few irrelevant columns from our data sets such as 'special note' and 'Indicator Code' since these columns are not useful to our data base.