Project 2 Write Up

Felix Poley

Dan T. Do

Lucas Vianna

For Project Two, we decided to focus on extracting, transforming, joining, and loading two data sets that were based around COVID-19 vaccination rates and hospital bed availability in the United States. Our COVID-19 vaccination rate data set was provided by ourworldindata.org, and it included roughly 21,000 rows of data that needed to be cleaned and filtered. This dataset showed updated COVID-19 vaccination data from all 50 states, and was downloaded as a .csv file. For our hospital data set, we used a dataset provided by healthdata.gov. The hospital coverage data set had roughly 6,000 rows of data, and was also available for download as a .csv file. Both of these data sets are updated regularly, and contained a robust amount of data. We wanted to choose a topic that not only had large amounts of data, but also had relevance to our lives. Because of the voluminous data sets, we had to cleanse them, erasing incomplete cells or getting rid of data that was not necessary or relevant for this project. Following the cleansing of the data, we joined our two datasets by state. This transformation of our data allowed us to load a cohesive, joined dataset into a relational database. We chose to load our transformed datasets into a relational database for a variety of reasons. Primarily, we believe that our data is best presented in a schema of rows and columns and therefore a relational database is the most appropriate.

Our original datasets were presented in two ways. Firstly, the dataset on hospital coverage in the United States was presented in a set of rows and columns with a variety of column headers, such as state, facility name, and total beds available. The dataset regarding COVID-19 vaccination rates in the United States was presented in the form of a line graph, with options to filter based on state, as well as in the format of a table. Both datasets contained a large amount of data that would not be helpful if we were to have used these datasets to perform an analysis. The hospital coverage in the United States dataset contained columns regarding CMS Certification Numbers, ventilators used, and ICU patients confirmed with Influenza; data that is important, but irrelevant to any potential analysis that would have been conducted. While the hospital coverage in the United States dataset contained irrelevant columns, the COVID-19 vaccination rate in the United States dataset contained a multitude of irrelevant rows. The rows that needed to be cleansed from the COVID-19 vaccination rate in the United States dataset included an array of dates that did not line up with the dates from our hospital data in the United States dataset. For example, the hospital coverage dataset contained just information for the week ending in December 2nd, 2021. Conversely, the COVID-19 vaccination rate in the United States dataset included every single day from January 12th, 2021 and is updated daily. For our two joined datasets to make sense, there needed to be some sort of consistency standards. Had we not cleansed the data of these irrelevant dates, our joined dataset would be difficult to interpret and analyze. After extracting our datasets from the web, cleansing and joining them, they had to be loaded into our relational database. As mentioned earlier, we chose to load our new dataset into a relational database as the best way this data was to be presented, viewed, and analyzed was in a dataset that contained rows and columns. The tabular schema of a relational database offers us a more concise way to present our datasets, directly showing data in comparison in an easy to digest way. The simplicity of the model, coupled with ease of comprehension and data flexibility provided us with a better option for loading our data. Our dataset was loaded into pgAdmin4 as our dataset administrative tool, and displays a joined dataset that consists of the cleansed United States COVID-19 vaccination dataset and the cleansed COVID-19 hospital coverage dataset.

To conclude, for Project Two our group extracted, transformed, and loaded COVID-19 related datasets into a relational database. We decided to focus on COVID-19 related data because of its relevance and spread of data around the United States. Because COVID is still a major presence in the United States and abroad, with considerable amounts of data relating to health, economics, and other topics being posted every day by a variety of sources, we found that it would be a good topic to choose in regards to the transformation part of the project. Because of the seemingly endless ways in which COVID data is displayed we hoped that by choosing related datasets we would be able to focus on the cleansing and joining of our data, taking very large datasets and cleaning them so that they are easily digestible and understandable. Our first dataset was a hospital coverage dataset by state, and our second dataset was COVID-19 vaccination data by state. We cleansed these two datasets by extracting the data then getting rid of unnecessary and irrelevant rows and columns for an easy to comprehend, flexible, and relevant dataset. We cleansed columns from the hospital coverage dataset that served no function, such as reporting source, glove stock supply, and ability to obtain PPE. From the COVID-19 vaccination dataset we cleansed rows, as these contained dates that were not found in the hospital coverage dataset. These two datasets were then joined by states, as opposed to joining by dates. The resulting datasets were then loaded into a PostgreSQL database and then pushed into our Github repository.