ETL Project Report

Team: DataManiacs

Members: Kelly Cash, Sophie Qiao, Betsy James, Joe Kim, Grace Sun

3/20/2021

# Extraction

Four datasets were extracted from Kaggle, USDA, and Worldometer and were used to fulfill the analysis objectives. We exported the data to CSV files from three of the four websites: U.S. state vaccinations, population, and poverty. The web scraping method was applied to extract Covid-19 data from the website.

# Transformation

The raw data exported and scraped from the websites contain some extra data that is not useful for our analysis. Therefore, after using Python Pandas to transform the data into data frames, we started the data cleaning process, including: remove the unnecessary rows and columns, drop “NaN”s, and convert data types.

The USA COVID-19 vaccinations dataset from Kaggle contained daily COVID-19 vaccination data from the U.S. by location from 1/12/21 to 3/18/21. We chose to focus on the 50 states and Washington, D.C. as locations. Upon examining the dataset, we noticed that it also contained additional territories or categories that did not fit into our analysis, such as “Bureau of Prisons” or “Virgin Islands.” These additional locations were searched using loc, and those rows were deleted. Any rows with “NaN”s were dropped as well as columns that did not fit our project goals. Finally, the data were grouped by location, and data for each column was summed. The data frame with each State listed once could be easily joined with our other datasets.

The poverty dataset contains 2019 U.S. poverty data. We dropped the columns that do not fit our analytical objectives, such as poverty numbers of 17 and under population. The final dataset contains the total poverty number, Lower\_Bound\_90CI poverty, and Upper\_Bound\_90CI poverty numbers by states. To match with other datasets, we have to drop the row contains U.S. national data.

Similarly, we dropped the columns that contain population data from 2010 to 2018 and only kept the most recent U.S. population data for 2019.

The covid-19 data is the only dataset we have to extract by using Web Scraping and pandas. To be consistent with the process we have done to other datasets. We dropped the Covid data for additional territories. The final columns included in the analysis are State, total cases, total deaths, active cases, total cases per million population, death per million population, and total recovered population.

# Load:

Postgres, the relational database management system, is what we used to create our ETL Project database, schemas and store the final dataset product. We created an ETL Project database and wrote four sets of queries to create the four schemas based on our decisions on columns and data types. After cleaning the four datasets in Python Pandas, we used SQL Alchemy to load our data to the ETL project database built on Postgres. In addition, we ran some tests in Pandas to ensure the data we load to Postgres database are accurate.