Project 2 Group:3

For the extraction process we needed to first determine what files we wanted to look at for the subject of our project. We determined the subject to be Covid-19 cases from the start of the pandemic to current. Once this was determined we set out to find various databases that kept track of the cases of covid both World Wide and throughout America. The datasets we found were through Kaggle and Data World websites and all three datasets were CSV files, which made it easier for us to work with. We had to read the CSV files into our Pandas dataframe and from there we were able to determine which columns which would best reflect the information we wanted to share, as well as find out what type of data the information in those CSV files fell under to helps transform the information into a more manageable dataframe to gather information.

For the transform portion of the project we first had to check to see if there were any null variables within the datasets we were using. Since our datasets were significantly large, we decided to condense them to give a more digestible amount of information to present. While merging the datasets together we managed to merge all the csv files on ‘Date’ so that there were no problems moving the Datasets to PgAdmin4.

For the load portion of the project we decided to use PgAdmin4. This is a relational database we have been using for the vast majority of our time working with databases so it seemed to be the most fit for our needs. We found it would be helpful as any interest in connecting two different tables together would rely on having a primary key to match the tables. The tables we had at the very end were just versions of our cleaned up dataframes from Pandas. We did not do any merging of tables as we felt the data told stories that were too different and would just further confuse and dilute our main goal of simplifying the ETL process.