**Project 2: Documentation**

First, we collected data on daily covid cases from Kaggle. Then, we collected data on state population from census.gov website. These two datasets are the source datasets for our project. We calculated the total monthly cases as well as daily average cases from the first dataset and added those two variables as two columns to our covid dataset. We then merged this dataset with the state-level population dataset. In the merged dataset we created a new column called ‘infection\_rate’ which is equal to number of monthly cases divided by the population for each state. We saved this merged dataset. This merged dataset is one of our target dataset (titles ‘statepopulation’). Then we used the original covid daily cases dataset and parsed the ‘datetime’ variable in the first dataset into three different columns: year, month, day. We, then, randomly selected four states to monitor the covid cases and infection rates over all the months in 2020. After conducting some data transformations, we saved this dataset as our second target dataset, titled ‘monthlystateinfect’. Those two datasets were then loaded into postgres and we then conducted some queries.