We decide to use the GDP as a barrier to opting former 50% state as our choosing pool. We use the data from the state economic monitor, which include US GDP information. The first step is to read the file and set the Geography column, representing the state name as the index column.

Because the pandemic begins in last year Q2, so we choose the previous year’s Q1 GDP as our GDP barrier. And we use nlargest formula to arrange the state and select the top 50% state as our candidate. Then we use the index to delete the row not necessary. The second step is to read the covid case. We found the information from CDC and use left join to merge it with our previous statement, GDP barrier. Also, we use fillna to fill up all the NA values with zero.

Then we want to choose five states to focus on. So we use nlargest to rank again. This time the ranking factor is the case ratio. Which means we pick up five most severe state among our candidate. Because we need to merge this file later, we also change the index name into “State”. And insert is to visualize better the percentage of confirmed cases.

The five states are Tennessee, Arizona, Wisconsin, New Jersey and Arizona.

Then we start analyzing vaccination data. Because the dataset is a daily accumulation, we use pivot to choose a specific date, in this case, is month-end, and insert the state name instead of the abbreviation to execute merge. We also clean the NA data in the data set. Now we have vaccination data and the five most harsh states.