We extracted 3 separate datasets from different locations. The Population Dataset was from US Census csv file. We dropped the first row and last 7 on the import as those were notes and not data. The US Education Datasets Unification Project data was a csv file from Kaggle. We dropped the first row on import as it had unnecessary notes. The GDP data was scraped from Wikipedia using Pandas. It was a HTML table, so it was ingested as a data frame.

Some of the datasets contained years and states that were not included in others. The Census dataset was for the years of 2010 to 2018. The GDP dataset were from years 2011 to 2018. Finally, the education data had data from 1995 until 2016. We decided that we should only load the data for the years between 2011 and 2016.

We also limited the states to the fifty states and the District of Columbia The School had US Territories in it, but only for a few years, so it was excluded. We limited it to states and the District of Columbia. The web scrape GDP data included a row for the United States. Because this data can be calculated from all of the other states, we considered that it was redundant and not necessary to be uploaded.

We decided it best to make a primary key that was year\_state. As the education dataset had this key already, we ensured that the other two datasets had keys that matched that format. That way they could match in a way that they could be merged in pandas. The first part of this process was to normalize the state names. We dropped the dot from the state in the census data, replaced any spaces with underscores for both the census and GDP datasets, and concatenated the year to the front of the states for the census and GDP datasets.

The datasets did not have the same table shapes. The GDP and census data had years in the columns and the states in the row. To merge these, we create a function to transform the data frames so that they were one column of data that had one index that was the combined primary key of year\_STATE. We made a function for this as this transform was required twice. The web scrape included a row for the United States. Because this data can be calculated from all of the other states, we considered that it was redundant and not necessary to be uploaded.

After all of these steps were completed, we merged all three of the datasets using the primary key using pandas. The final outcome of this merge was one table.

We transformed the data to be loaded as a single table. We chose a single table because all of the data is related and can be used to calculate other data features. For example, since we have the population of each state, and the GDP per capita, we can calculate the total GDP per state from this data for future analysis. We chose mySQL so that we have the potential to house the data in a remote location, if necessary. Additionally, doing this allows direct queries to sort, filter, or aggregate this data. It also facilitates the integration with visualization utilities.