**ETL Project**

Our project focused on reviewing main components in Florida community colleges by choosing a sample of five colleges over all the major territories. In order to retrieve this data, we started doing research for databases that included our needed information; tuition, population, ranking, and graduation rate. There were several education research websites we were able to use that allowed us to either download raw data in excel or filter the information by the specific colleges we wanted, which was also available via excel. We then converted the file into a CSV and started the cleaning process. Working with the data we had, we needed to have the data reflect the specific characteristics we were interested in, such as tuition and graduation rate.

To clean the data and combine all of our data and combine our information into one, we had to use Jupyter Notebook and imported the SQL alchemy and Panda’s module. We removed unnecessary data and cleaned up formats in certain columns as well. When we combined the data that was repeated across our multiple sources of information, we had to reformat/rename certain columns to ease the joining process. This was done when we merged together the data based on college name and ID. One of the very last edits we had to make was removing commas in order to transform the data into integers. By the end of our data cleaning process, we had successfully created a database of our information. At the completion of data cleaning and transformation, we exported everything from quick database diagrams into SQL. We chose to use a SQL database as our datasets composed of information unique to itself. This way we could pull information from separate unique tables to query the exact answer we are looking for.