Project Report

# Extract

The team decided to gather financial data from Data.gov to process and place into a SQL database. The original file found was a table of Education Funding by School District for 2013. Once we had decided on this table, supplementary data was researched and acquired so future users of the database would be able to perform meaningful analysis. Because of the lack of quality data at the national School District level, the school education funding will be aggregated at the state level.

The datasets used by the team were:

* School education funding by school district:
  + Source: Opendata.gov
  + Format: Downloaded, converted text file to csv
* Unemployment by state
  + Source: Opendata.gov
  + Format: Downloaded as an Excel file
* Education attainment by state
  + Source: Opendata.gov
  + Format: Downloaded as an Excel file
  + Specific Topic: Percentage of adults completing college
* Poverty by state
  + Source: Opendata.gov
  + Format: Downloaded as an Excel file
  + Specific Topics:
    - Total Percentage of people in poverty
    - Percentage of children in poverty

# Transform

Once the data were gathered and reviewed, the next team determined the final table structure in our relational database (SQL). Each of the source files would become their own table. Below are the steps required to transform the source data into the final product that is ready to be loaded into the SQL database. The datasets were transformed in the following ways (as illustrated by the Jupyter Notebook files in the GitHub Repository):

* School education funding by school district
  + Initial file cleanup:
    - Converted original text file into a csv format
    - Reviewed the data and determined the fields that would be useful to keep in the final database
    - Determined the logic for aggregating the data to the state level
      * Count of the number of school districts by state for normalization in the future.
      * Individual records of the state name, abbreviation, and year
      * Sum of all other records per state of the fields that are needed at the state level
  + Detailed cleanup:
    - Determined that the “FIPST” would be the key field for all datasets
    - Created new table with only those fields that are of interest
    - Removed data values that would skew any results. In this case there were multiple rows that contained a “-1” or “-2” instead of a zero. This would impact the summarized values, so they were changed to be zero.
    - Created the count of school districts by state for normalization.
    - Summarized the monetary values so when the state values are shown, the total of all school districts were in the database.
    - Merged the dataframes in Pandas and made sure the column names matched the SQL table field names.
* Unemployment by state
  + Initial file cleanup:
    - Manually cleaned the downloaded Excel file so it would be in a tabular format
      * This involved converting currency to general numbers, taking three rows of columns into one and assigning the appropriate column names
      * It also involved column un-merging, since the conversion to CSV caused the columns to merge
    - Renamed columns for ease of use.
    - Exported to a csv file
  + Detailed cleanup:
    - Used the OS function to join the CSVs onto Pandas. The resulting data showed no errors in rows, columns or the data itself.
* Education attainment by state
  + Initial file cleanup:
    - Manually cleaned the downloaded Excel file so it would be in a tabular format
      * This involved converting currency to general numbers, taking three rows of columns into one and assigning the appropriate column names
      * It also involved column un-merging, since the conversion to CSV caused the columns to merge
    - Renamed columns for ease of use.
    - Exported to a csv file
  + Detailed cleanup:
    - Used the OS function to join the CSVs onto Pandas. The resulting data showed no errors in rows, columns or the data itself.
* Education attainment by state
  + Initial file cleanup:
    - Manually cleaned the downloaded Excel file so it would be in a tabular format
      * This involved converting currency to general numbers, taking three rows of columns into one and assigning the appropriate column names
      * It also involved column un-merging, since the conversion to CSV caused the columns to merge
    - Renamed columns for ease of use.
    - Exported to a csv file
  + Detailed cleanup:
    - Used the OS function to join the CSVs onto Pandas. The resulting data showed no errors in rows, columns or the data itself.

# Load

The last step of the project is to load the data into SQL. The SQL script and Pandas code used for this step are all within the GitHub submission.

Table definitions for School Funding , Unemployment , Poverty , Education with FIPST as a primary key.

Database Schema is as follows:



Data is loaded into the tables and available for query.