**Project 2- ETL Final Report**

The Miner League

Amy Brunet, PeiChi Hockaday, Trang Bui

**Extract**

For this project, our team wanted to continue our research and analysis on the factors that influence student achievement. With this goal in mind, we began searching for data sets that would allow us to analyze different factors contributing to or hindering student achievement across the United States. In our search, we found three datasets from the years 2013, 2015, and 2017 on poverty rates and median house income in each of the 50 states across varying age groups. We decided to use the three datasets we found for this project in conjunction with one of our data sets from our last project, which encompasses average math and reading scores for students in each of the 50 states over a span of several years. We found the three new datasets (as csv files) from the United States Census Bureau website on the income and poverty rates and the average math and reading scores data (as csv file) was from the National Assessment of Educational Progress website. Location of where we found each of the data sets is linked below:

US Income and Poverty Rates, 2013

<https://www.census.gov/data/datasets/2013/demo/saipe/2013-state-and-county.html>

US Income and Poverty Rates, 2015

<https://www.census.gov/data/datasets/2015/demo/saipe/2015-state-and-county.html>

US Income and Poverty Rates, 2017

<https://www.census.gov/data/datasets/2017/demo/saipe/2017-state-and-county.html>

US Average Math and Reading Scores, 2013-2017

<https://www.nationsreportcard.gov/>

**Transform**

In order to successfully load our data into our database, we first had to do some cleaning and transforming of the data. The first step was to clean and transform the new datasets we procured on the poverty rates and median household income data in the United States. Before loading the csv files into pandas, we first did some cleaning in excel. There were several columns of information that were not necessary for this project, so we removed those columns and the header that was at the top of the file. After the initial cleaning and formatting in excel, we then loaded the three csv files into pandas to create data frames and tables to load into our database. We took each of the data frames for the years 2013, 2015, and 2017 and concatenated them into a single dataframe. Once they were merged together, we re-indexed the dataframe and renamed the columns in order to create consistency with each of our tables.

For the next step, the average math and reading score data was previously cleaned and transformed from our first project, so in order to utilize that data set for this project we loaded the csv file into pandas and manipulated the dataframes with python. The scores data set included data for students in grades 4 and 8, in both math and reading, and for various years in each of the 50 states. From this information, we created three new data frames in pandas in order to give each attribute its own table in the database. We created a year table, which holds the years 2013, 2015, and 2017 and gives them each a year\_id. We then created a subject table, which holds math and reading and gives each subject a subject\_id. Finally, we created a table for the grade level, which includes 4th and 8th grade and gives each a grade\_id.

The final step of transforming the data was to merge each of the aforementioned tables together with the average test score data and with a state table that we created in excel before loading into pandas. The state table we created in excel consists of each of the 50 states, in addition to Washington D.C., and assigns each a state\_id. We merged the state table with the grade, subject, and year table to create a table that displays the average test scores with each of those keys that correspond to it. Finally, we used the same steps to merge the year and state tables with the poverty rates and median household income data.

Through this transformation, we were able to create six different tables to be loaded into our PostgreSQL database.

**Load**

After extracting and transforming our data, we then loaded the data into our PostgreSQL database through pandas and SQLAlchemy. As mentioned above, we loaded six different tables into our database including a state table, a year table, a subject table, a grade table, an average score table with corresponding keys, and an income and poverty rate table also with corresponding state and year keys.

Through this project, we were hoping to learn more about the factors that impact student achievement in the United States. In our first project we looked at funding and discovered that it does not have much of an impact, so we concluded that we would need to do more research on other factors that could be influencing student test scores. By looking at poverty rates and median household income, we hoped to find more evidence either for or against our hypotheses. Although this project does not require any analysis, we plan to investigate this topic further in the future and utilize the database we created to help inform future projects.