ETL Project Summary
Bercin Cenik and Dann Taylor

Idea: We wanted to compare the number of students in a state to the number of people incarcerated.

Process: Using Google, we found three different websites providing starting data:
- thesentencingproject.org - provided incarceration rates at the state level for those in prison and in jail
- US Education Datasets: Unification Project from Kaggle showing the trend of enrollment numbers, expenditures and test scores for each state
- The US Census website, providing resident populations per state for 2010 through 2018

Extract: A file was created from each of the above sources.

The Census and Kaggle provided fairly clean CSV files while data from the sentencingproject.org required manual extraction.

After preparing the datasets, each was read into a Jupyter Notebook using pandas. The data structure was examined ensuring each file could relate to each other in a database.

We then created a database using MYSQL allowing us to connect files. As each file had a state column, we connected the files on this variable/field. The data frames were loaded into the MySQL database and the queries were tested. Each query results was successful, showing for the example the enrollment numbers, and the prison population and the population per state.

Ideally these query and data could then be used to compare the number of students against the number of people incarcerated per 100,000 people. The school enrollment data did include budget information, however the sentencingproject.org dataset did not. If this information was available, we'd be able to compare how much a state spends per student vs per incarceration. We are able to compare the number of people per 100,000 in each environment which is interesting.