Process

GitHub: https://github.com/jawhite1612/database_project

States Census Data

All state census data was taken from the US Census, specifically from the website: https://www.census.gov/mycd/. All of the csv files that come from this site are separated by state and include a plethora of information on each of the different congressional districts of every state. We used a python script in order to combine information from these csvs into different relations. The script only extracts the attributes that we found most interesting (all of which are modeled by our diagram in Phase B). There was a lot of data transformation that had to be done; since the csvs were per state, and each state had each of its districts as column attributes whereas districts are tuples in our schemas, we had to do more than simply delete columns. Instead of describing the whole transformation process, I have linked our GitHub repository that contains the script called *stateDataParser.py* that creates the state-related tables. The data used to populate those tables can be found in the director *states/* found on the same repository.

Election Data

Similar to the states data, election data was parsed and relation files were made using a python script. However, the data that we received from this source required much less processing. It basically required removing a few columns that would not add to our project. The election data was taken from an MIT database (https://electionlab.mit.edu/data). We decided to focus on presidential, senate, and US house races from 2016-2018. These years were chosen since our census data is up to date as of 2018, and we wanted to include a presidential election as well. We then combined all of the election data into two relations (Election and Candidate) using a python script. The python script is included in the GitHub as *electionDataParser.py*, and uses data from the *elections/* directory.