Data Analytics Project Proposal

2010 US Census - District of Columbia

Sage Foh and Alexia Lutz

The Data

The data we will be using for this project is from the 2010 US Census. The census data is very large, so it is broken up into summary files categorized by state. Further, each state's summary file is actually a collection of multiple files, with each file representing data from a category used in the census such as gender, age, income, etc. For the scope of this project, it is likely that only one state will be used for cleaning and analysis due to the large size of each state summary file. If we look at the census summary for Virginia, for instance, the summary file is almost 3GB in size. For our project, we will be focusing on data collected from the District of Columbia.

Cleaning

The census data for the District of Columbia will need to have each summary file category merged into one complete data file. In order to do this, each summary file will need to be converted from its .sf1 extension to .txt or other workable format. Then, the column headers for the data will need to be determined and imported. Once we have all of our data and appropriate column labels in one file, we'll be able to determine if any further cleaning is necessary.

Analytics

Due to the large amount of various categories in census data, we will be examining the census data questions and responses in order to see if there is any significant correlation between the data categories. If we have time, we can grab data for the District of Columbia using the 1980 and earlier US censuses, and compare responses over time. Another possible goal is to compile together a data table for multiple states and compare the responses given by their residents.

Implementation

We plan on using Python for the majority of our project. Python not only has packages to handle data cleaning and analysis, but is also useful for writing the scripts we will need to convert and merge the summary files. To keep track of project files and progress, we have opted to use Google Drive, but we can also keep a working copy in a Git repository for use by others who may be interested in a cleaned version of the US census data.