Data Procurement and Management Plan

Mitchell Beckner

1/16/2021

Tool Selection

The primary tool used for this phase of the project will be R and R Studio. Data will be imported into R using the censusapi package and an API key that can be obtained by going to https://api.census.gov/data/key_signup.html. Data wrangling and cleaning will be accomplished primarily using functions from the tidyverse packages. This tool was selected based on my personal experience and comfort level. While the API download process for this data is new to me, I am quite familiar with data cleaning and reshaping in R and find it much easier that attempting the same tasks in Microsoft Excel.

Data Procurement Method

The data set for this project will be created using tables obtained from the United States Census Bureau. R code has been created that will download the list of all the data tables that be obtained. Using that list, the desired data tables will be identified, and the variables included in each selected table will be examined. Specific variables for each table will then be selected or excluded as appropriate and the actual data imported as dataframes. I plan to start with a base table using 10 year census data and add additional data from selected one year tables. I am currently researching how to pull the data by Zip+4.

Data Wrangling and Cleaning

Once the desired data tables have been imported into R, they will be joined using the Zip or Zip+4 columns. Each variable will be inspected to identify missing and/or incorrect values. These values will be replaced, imputed, or deleted as deemed appropriate. Regression or extrapolation methods may be needed to reconsile the one year data with ten year data. As would be expected, this phase of the project will likely be the most time consuming and require the greatest amount of work. My goal will be to assemble the most useful and granular dataset that I can in the time allotted.