Introduction

The data that we have collected is from an American community survey collected over 5 years of data. Originally our hope was to have data collected over time, but due to the way the data was collected over 5 years and not instances in time that would allow us to do a time series analysis. Instead what we have done is looked at certain different demographics and gone through a cleaning process to have the data organized for others to download. The different demographics that we looked at were frequency by age, sex, food stamps, healthcare, employment, poverty status, and schooling for each state. In addition we parsed through and examined median income based on sex and employment status as part time or full by for each state.

There are a few interfaces online that display similar data but only allow you to see the data in their pre-determined graphics. What we do is allow the user to see a sample of images of what you can do with the data and allow the user to download the data so that they can do similar analytics and graphics on the data itself. Especially given the vast amount of data cleaning it takes to get the data in a tidy format, allowing users to download the data is an extremely important feature.

Cleaning

The data cleaning that was done used the acs package in R in order to first get he data offline. There are different ways to go through and search for the data, this can be done by looking into data at a state, region, city demographic in order to grab different frequencies. We would then use a function called acs.search, what this did is allowed us to see all the different tables for a certain variable of interest. For example we would do an acs.search on employment, and every table that broke down frequencies by state for different employment categories would be given to us.

At this point we would have a list of tables to look into for a variable such as employment and we would have to go through each of those tables to find ones that were of interest. A lot of tables would give things such as margin of errors for estimates and other summaries that we would not care to display, so those had to be filtered out when finding the correct tables.

After a table of interest was identified the cleaning of that table was then done. The format of the columns needed to be cleaned by removing certain characters that were not unique to that individual column. We then would remove certain rows that we did not need for example totals, if we were given gender as male and female having a column for total would be seen as unnecessary and would be removed from the dataset. Furthermore, keeping this column would create issues later on in our cleaning when string splitting to create columns in our final step.

After this was done the columns were still given in a non tidy format for example a column header would be “Sex.by.Age.by.Employment.Status.for.the.Population.16.Years.and.over..Female..75.years.and.over..Not.in.labor.force”. We would have first gone through and removed “Sex.by.Age.by.Employment.Status.for.the.Population.16.Years.and.over..” At that point we may need to go through and fix ages so that they are in a more appropriate bucketing (some ages would be only one year whereas others would be ranges of 5 years). Once this was done we would make each of the row names, which were the states into a column and melt by state.

Once the dataset was melted we still needed to go through and strsplit by our variable column to create different columns to have our data in a final format. In our example above “Female..75.years.and.over..Not.in.labor.force” would be split on “//.//.” Which are two dots and creating a new variable for gender, age, and employment status. At this point our dataset is finally in a tidy format and can be used to create visualizations, or do any type of data analytics on itself. To create visualizations we would use other tools such as plyr which allowed us to summarize the data in different ways or create new columns such as state totals. Creating a new column such as state totals is very important given the dynamic of our analysis since certain states populations differ so vast, having a percentage as opposed to frequency is much more informative.

This process although seemingly could be used on every dataset through a function, could not be done in such a way given the different format datasets were given to us. For example removing different columns such that string splitting could be done in an efficient manner had to be examined for each individual dataset to ensure that we were correctly identifying variables of interest. Furthermore, there are certain datasets that have tables A-H on the same dataset. What these different columns A-H stand for is a different race by demographic instead of including a race column in one individual table. Because of this, tables that examine individual races had to be done separately and joining together by race after the initial cleaning process was finalized. For tables A-H a function was created to parse through all tables within a particular demographic, but once we cleaned a different set of A-H tables we had to create an entirely new function for those tables.

The dataset in the final tidy format is what was released to the users who use our shiny application. We found that the acs package is very helpful in terms of getting the data offline into a nice format to be cleaned, but there was still much cleaning to be done. Because of this we found it to be very important to allow users to go through and download the data in a cleaned format on their own.

Shiny

Given the interactive nature of what we are looking to do Shiny was the best option to display our data to users. Our shiny application has two conditional panels

Future work