IS 452 Foundations Information Processing Fall 2017

Final Project Proposal

Nanzhu Liu

UIN: 655394878

**Project Type**

Programming project

**Project Summary**

The project is to estimate demographic data related to children under age 5. The data is updated annually, and the project is done every year. I used to combine them manually and by using Access. It took me a lot of time. The structure is similar every year, so the process is repetitive. One problem I met was some of the data are not reasonable since they are from two different sources. To make them closer to reality, some adjustments need to be done. To achieve this and save time, I want to try to program the data processing in Python. There are mainly three parts. Part one is to read all data sets into Python. Part two is to combine those data sets by their common attributes. Part three is to balanced data to make them look reasonable.

I want to use the knowledge of Python I learned from this class to create a python file which can accomplish data combination and preparation for me.

**Dataset**

2015 American Community Survey (ACS) 5-year Estimates from <https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t>

2015 Bridged-Race Postcensal Population Estimates at National Vital Statistics System (NVSS) from <https://www.cdc.gov/nchs/nvss/bridged_race/data_documentation.htm>

ACS contains varieties of data sets and has large data sets, but the numbers are not as accurate as NVSS. NVSS contains less kinds of data set and just has total number of children under 5. I need the number of children in every age. In this project, I will collect population, population by race, 100% federal poverty level, 200% federal poverty level, 400% federal poverty level and language data from ACS and I will also collect population data by counties from NVSS because this is more accurate. All the data sets are about children under age 5 in Illinois divided by different regions. The regions include counties, townships, elementary school district, unit school district and municipalities. The total records will be about 4000 with 41 columns.

**Deliverables & Project Process**

1. Collect and prepare population data by counties from NVSS. Create codes to read sas data file, keep the needed columns and rename their headers– Done by midpoint check in
2. Collect and prepare population, population by race, 100% federal poverty level, 200% federal poverty level, 400% federal poverty level and language data by counties, townships, elementary school districts, unit school districts and municipalities from ACS. Create codes to read csv data files, keep or delete columns, rename headers, aggregate several columns –Done by midpoint check in
3. Merge all the data files based on their common attribute. Write codes to merge data.
4. Get the rate of each column from ACS data file. Create a function to calculate the rate at one time
5. Balanced the data based on the data from NVSS and the rate. Create a function to balance the data.
6. Output the data into a new file.

**Challenges**

1. Read data files. In the class, I know how to read txt file, but I don’t know how to read sas or csv format file.
2. I need to use keep and delete functions for columns. I’m not sure if I can use replace functions. If this doesn’t work, I have to find out something else.
3. I know how to use operators to do math problem for one numeric variable or one matrix in Python. However, I don’t know how to do it for multiple matrixes. I will try to use for loop.
4. I need to use merge function. I don’t know if Python provides the function or not. If not, I have to write one by my own.

To solve those challenges, I will look for solutions by searching online, reading books, and reading papers.