CMPS 263

Data Wrangling Procedure

Brady Goldman, James Byron, Jon Selberg

March 16, 2017

1 Data Retrieval

- **System Requirements**
  - Java SE Runtime Environment 8
  - Internet Explorer (Not Edge) or Mozilla Firefox
- **Data Retrieval from ACS using DataFerrett**
  - Before running DataFerrett, be sure to disable your popup blocker for `https://dataferrett.census.gov` and allow Java to run in your browser.
  - Go to `https://dataferrett.census.gov/LaunchBetaDFA.html` and enter your ucsc email in the login popup. Then click the top tab that says **Step1**.
  - From the left column, dropdown into the American Community Survey dataset, then dropdown into the Public Use Microdata Sample. View variables from the 2015 dataset.
  - Only select the topic that says **Population**, then search the variables.
  - Double click variables with names DEAR, DEYE, DPHY, OCCP, SCHL, and WAGP. Select all values for each varable with the exception of "Not in universe - missing" for DPHY.
  - Proceed to **Step2** tab. Click button that says **Make A Table**.
  - Drag and drop OCCP variable to left-most column. Then drag DEAR and WAGP in that order to the adjacent columns and Click the green button that says **GO Get Data**. When the data has replaced the question mark placeholders, save the table as a .csv file. Clear spreadsheet and repeat for DEYE and DPHY.

- Drag and drop the OCCP variable in the left-most column again, but this time drag only SCHL to the column adjacent. Generate data just like above and save to a file titled "education_attained_by_occupation.csv".
- Place each csv file in a folder labeled "data".

2 Data Post-Processing

- **System Requirements**
  - Python 2.7.x with numpy and pandas installed
  - Excel or Libre office
- **Narrowing Education Categories**
  - Open file "education_attained_by_occupation.csv" in Libre office or Excel. This should have counts for each highest level of education attained by occupation. Group these into the following five categories: Less Than High School; High School; Some College; Bachelor Degree; and Master, PhD, or Professional Degree.
  - Sum counts from Columns D-S to obtain total by occupation for "Less than High School", Columns T-U for "High School", V-X for "Some College", Y for "Bachelors", and Z-AB for "Master, PhD, or Professional Degree". Place these in the next five columns after the last.
- **Mapping expected highest level of education attained by occupation**
  - To convert the counts for highest level of education attained to a label for each occupation, we compute the following expected value for each:

$$\frac{\sum_{i=0}^{4} i * \text{Count}_i}{\text{Total Count}}$$

    Where i is the index corresponding to each category of highest degree obtained. The result rounded to the nearest whole number will designate the corresponding category as the label.
  - Download and move expected_education_mapper.py just outside of the directory labeled "data". Run script to generate csv file with categories mapped to each occupation.