Jacob Miller

Professor Sonstein

DSCI.623.01

18 October 2019

Midterm – Part 3 – Post-Processing

There was very minimal post-processing necessary to produce the final executive summary report. All the data processing happened in the *Query\_Script* program, the only exception being the visualizations. I very much wanted to produce the entire package in a Bash script since I had no prior knowledge of the language prior to this report and found it very enjoyable to learn. However, I had to move to Python/Matplotlib for the plotting to ensure the report could be completed on time, as that language is far more familiar to me.

The Matplotlib script, aptly named “Midterm\_Plots”, is very straightforward. When run from within the *Combined\_Files* directory, the four processed data files (*Combined\_Education, Combined\_Income, Combined\_Mortality, Combined\_Terrorism*) are loaded into Pandas dataframes using *pd.read\_csv*. As these data files are tab delimited, it’s necessary to set the *delimiter* parameter to “\t” to override the default comma delimiter. To align all the dataframes as closely as possible, information from 2017 is dropped from the Terrorism file since the Mortality/Fertility/Income files only run through 2016.

When plotting directly from the dataframes, it is necessary to transpose them so that columns represent categorical variables (countries) while rows represent consecutive years. Several other plot parameters are set (x- and y-axis ranges, for example) to make the plots appear visually consistent, and then are simply plotted individually.