Christopher Anderson
Professor Birdsong
January 19, 2020

DSC 540 Data Preparation

Mid-Term Project Summary

Introduction

This brief paper will summarize the steps and challenges I ran into during this mid-term project, and it will outline particular decisions I had to make while transforming and cleaning the data.

Steps and Challenges

There are a handful of steps that I went through with this project, and this is the order in which I experienced them:

1. Select a dataset
2. Modify the dataset
3. Standardize the data
4. Look for outliers and bad data
5. Check for duplicates
6. Perform fuzzy matching

One of the more difficult tasks I faced was finding a dataset that was applicable to our learning in the course. It was easy to find datasets; however, to get a dataset that was a good match for what the requirements were for this mid-term project took me a bit. I eventually located something that fit the bill — and simultaneously piqued my election-year curiosity — when I found The General Social Survey dataset provided by NORC at the University of Chicago. A brief explainer of the GSS:

> *"Since 1972, the General Social Survey (GSS) has provided politicians, policymakers, and scholars with a clear and unbiased perspective on what Americans think and feel about such issues as national spending priorities, crime and punishment, intergroup relations, and confidence in institutions."*

After downloading a dataset for the 2018 year, I quickly realized my first challenge: I would need to convert the SPSS data (which is provided in a SAV file format) into something that I could use. This was a one-to-one match for what we did in this course, so I felt very comfortable doing so using the same SPSS to CSV format conversion steps as outlined in Chapter 7 of *Data Wrangling With Python*.

Furthermore, similar to what the author outlined in that chapter, I also had to transform the heading data from the GSS dataset into something more human-readable. Since the downloaded dataset uses mnemonics (a few examples include "ABANY" and "ABCARE") for the data's column headers, I had to find out what these strange words meant, and learned that there is a GSS Codebook Index, which I could only find in PDF format. So, just as in previous week's work in this course, I had to work through importing content from this PDF file into something I could use to cross-reference the header data from the dataset so that I could provide more meaning header information for my project.

Decisions to Make

A few of the decisions I made along the way in the journey of this project included limiting the number of columns/headers and rows in the original downloaded dataset into something much smaller, while also still fitting into the requirements of the project. The default 2018 GSS dataset had over 1,000 columns/headers and I felt that was simply way too much to work with for this project, so I trimmed this down to the low 20s. I also limited the total number of rows to something just a bit over 1,000. In a real world situation, I would not have done this, of course.