5.3 Mid Term Project

I chose a data set from Kaggle, called the Human Freedom Index, <https://www.kaggle.com/gsutters/the-human-freedom-index>. The purpose of the index is to present a picture of the overall freedom in the world. A larger purpose is to explore what is meant by freedom and to better understand its relationship to other social and economic phenomena. The dataset is available as a CSV file.

To begin, I examined the file in Excel to get a feel for its structure. Next, I imported the data into Python as a list of dictionaries, making each row into a dictionary. The original headers are brief and do not fully explain what is captured in each field. Using the website’s explanation for each column, I created a table with the original header name and the full name. I saved this as a CSV file and also imported that into Python. I created a new file using the cleaned headers in place of the original ones. This is the file that I continued to use throughout the project.

Viewing the data as lines of dictionaries is not very user friendly. I used the format function, and printed the data in a question and format display. I showed this for just the first row of data, but this could be used to display any row of data that needs to be examined.

Next, I identified any missing values in the data. When an answer was not present, the field was left empty. There were no NA answers. The missing values will need to be taken into account when statistical methods, such as mean and standard deviations, are to be performed on the data, as they could skew the results. I also examined the type of data in each field, such as digit, boolean, text, etc. which displayed the number of empty records for each category.

To find any duplicate rows, I created a unique variable for each row made up of the Year and ISO code categories. Since each country has a unique ISO code, and the data is by year, there should be one record for each country per year. Using this variable, I created a set of all of these unique codes and tested its length. Since its length matched the number of records in the dataset, I deduced that there were no duplicate records.

Finally, I looked into Fuzzy Matching. Most of the categories were numerical, but I looked at the Region category. I did a ratio compare to find results for any rows that were close to Europe. This helped me find any sub regions of Europe that were included. This matching would be very helpful if users had been inconsistent in keying in the various regions and sub regions in the world. After examining the data and replacing the headers, the file was saved as a CSV file.