Megan Jacobs

8/24/2021

CGT 270 – 007

Lab 01 – Acquiring Data

The data set that I chose to pick was Global Burden of Disease done by the Institute of Health Metrics and Evaluations (IHME). This data set has a total of 7 columns. There are two columns that are in string formats. These two columns are for the specific country and the code related to it. Two of the columns are in integer/decimal format which tracks the number of deaths and the death rate per 100,000. While one is in a date format, specifically the year of that particular row. Finally, the last two columns are in categorical columns of sex of the person and what age group it falls under. For the Global Burden of Disease, I will give it an 18 out of 20 rating. This is because it is a bit out of date and because it gives age ranges instead of giving exact ages but the rest of it checks out. The dependent variable is country and the independent variable for this dataset is mortality rate. This dataset will be the one that I will continue to use this semester!

I have chosen two other datasets to compliment it which include, death in the United States from 2005 to 2015 and Cancer rates by US State from 2013. I will first start with Death in the United States from 2005 to 2015. I give it a rating of 16 out of 20 due to the fact that it is not current, it has been said that it did not account for exact age of the participants and introduced some errors within the data for the Death in the United States. This data set has many and I mean many columns. A majority of these columns are in an integer format while some other columns are string because it has certain codes that connect the data to a certain cause of death. Finally, there are some columns that are categorical like whether or not the person is a male or female or is married, widowed, divorced or single. The dependent variable of this dataset is the year, and the independent variable is the death rate.

I gave a 16 out of 20 for the Cancer rates by US State based on the Evaluation Rubric. This is because it is a bit outdated going back to 2013 and that it was an individual that took data from many different sources and did not check if fallacies or errors were introduced. The data sources This data set has 3 columns, one being a categorical format since it is the abbreviation for each state. The next one is also a string and a bit messy sine it uses 2 numbers and a word in it to establish the range. Finally, the last column is an integer for the rate of death due to cancer in each state. The dependent variable of this dataset is the state that is chosen, and the independent variable is cancer rate for that state.

The target audience for these datasets can include those in the medical field ranging from doctors and nurses to people who work of the CDC trying to see which countries could be more affected by COVID-19 due to the burden of other diseases within the country. All the information is telling me that it is all a bit outdated due to the fact the two of the datasets are from 2015 and the cancer rate by state does not specify a date range that it is taken from.