*Create a short document (1-2 pages) in your github describing the data wrangling steps that you undertook to clean your capstone project data set. What kind of cleaning steps did you perform? How did you deal with missing values, if any? Were there outliers, and how did you decide to handle them? This document will eventually become part of your milestone report.*

In order to wrangle and clean the World Bank Gender Statistics data set, I first imported it into a Pandas dataframe. Since the columns in the raw data are years the datapoints were recorded, I melted the dataframe in order to stack the years into rows. Then I dropped NA values from the data frame, to drop any unnecessary rows created for years where no data was listed.

The rows of the raw data contained not only which country the data point was from, but also which indictor variable the point represented. Since I want to look at the various measured variables to see how they are affected by country and year, I decided they need to be listed as columns. So I then unstacked the indicator variable to list each measured indicator as its own column, with the rows representing the country and year the value was collected or measured.

For the time being, I have left any missing values as NaN because listing them as 0 would significantly affect my analysis. Once I have a more clear direction of my analysis, I may have to deal with NaN values in a different way. The same goes for outliers. I used Seaborn to make a boxplot of the entire data frame, but there is just too much data for the boxplot to really be useful.

There are nearly 700 different indicator variables contained in the dataset. This is an overwhelming amount of information to try to tease out something interesting. Once I figure out what my research questions are, it will making filtering out unnecessary data easy. But I also do not want to look over potentially interesting insights. To look at variables that are similar or have similar implications, I may attempt to do some factor analysis to examine correlating variables in order to group them together into a composite score of some kind, which I can then use to run t-tests between the different countries or even the different years.

For now, just to look at something and make sure I know how to run simple correlations, I plotted Contraceptive Prevalence against life expectancy as a scatter plot.