Are there any data problems in the dataset?

* Both the cardiovascular dataset and the poverty rate dataset are untidy.
* The datasets have differing levels of detail. One goes down to the county level, and the other to the state level. One has entries every 5 years; the other has entries annually.

The goal is to combine the two datasets in order to create a dataset that will allow us to answer the question, How does the Poverty Rate affect the cardiovascular mortality rate for each state?

The following are the steps I took to clean the capstone project data set.

Process for cleaning the poverty dataset.

1. Because of the lack of granularity in the cardiovascular dataset, I determined that I would pull the following years from the poverty dataset (1980, 1985, 1990, 1995, 2000, 2005, 2010)
2. Added a “Years” column to the corresponding dataset and filled it with the corresponding year.
3. There were no missing data entries in the resulting dataset.
4. Using the outlier definition of Q1 – 1.5\*IQR and Q3 + 1.5\*IQR, I sliced the poverty data by state and found that there were 39 instances a state has had a value that was considered an outlier when compared to that particular state’s other values.

Process for cleaning the cardiovascular dataset.

1. I filtered out the county level data and kept the state level data.
2. Tidied up the data by melting the data and replacing the variable values with their correct years.
3. Got rid of the confidence intervals that were included in the same column as the mortality rate.
4. Converted all the columns except for the state column into floats.
5. Using the same outlier condition as above, there are no outliers in the mortality rates.

Process for joining the data.

1. Replaced all the ‘District of Columbia’ values to ‘D.C’ in the cardio dataset because the cardio dataset had a different notation for Washington DC than the poverty dataset.
2. Changed the column value in the cardio dataset from Location to State.
3. Got rid of the FIPS column in the cardio dataset.
4. Merged the two datasets based on the State and the Year.
5. I also converted the mortality rate unit of measure from per 100,000 to 1,000 because the poverty rate unit of measure was per 1,000 people.

How did you handle the outliers?

* I identified 39 values that were considered outliers in the poverty dataset. However, I figured that the important outliers would lie in the poverty rate column. There were 15 poverty rate outliers.
* However, these outliers would not occur in a vacuum. If the poverty rate were an outlier, then we would expect there to be a significant change in the state population or the number of people in poverty.
  + We find that the number of significant outliers decrease from 15 to 8.
  + After further investigation, these outliers just seem to be from normal variation.
* I have determined that there is no need to drop the outliers. They do not seem to be incorrect entries.