Data Cleaning

In order to properly score the data, two factors needed to be considered. One was potential outliers in the quantitative attributes within our data. In other cases, it was simply if the Data itself was complete and formatted correctly.

In order to determine if there were any outliers in the attributes, a confidence interval of 99% was set up. This involved looking for any quantitative values that were greater or less than three times the standard deviation than the mean of the entire attribute. The percentage of values outside this confidence interval were then recorded as a percentage for each attribute. All the outlier percentages were then averages over the total number of attributes in each data set to give us the outlier component.

The second component was slightly more involved and required two subcomponents. The first subcomponent required seeing if in a given attribute, the total number of missing values. Generally, in the quantitative sections, missing values were already handled with zero values; however, in some cases they were blank as well. Thus the total number of blank values in the attributes were added up and then added to the second subcomponent. This subcomponent was designed specifically for the string attributes. The string attributes are a big component on the method by which the data sets will eventually merged; thus basic formatting techniques were required. The basic formatting that was agreed upon was analyzing if the character strings followed proper case. The method involved splitting up the strings and counting the number of words in said string. That value was compared to the number of upper case letters in the string. For all strings in an attribute that did not match it were summed of to make up the second subcomponent of the second component.

The two scores were weighted equally and subtracted from 1 to represent a score that was reflective of a percentage. After running through all the data sets, some basic assumptions were validated. In the case where data was pulled from the API, the percentage was significantly higher than non API obtained data. The worse scoring data set was scrapped from the New York Times. Due to the way it was scrapped, the attributes remained strings, in order to compensate for the string attributes, some required some slight editing of clearly numeric attributes. There were two major surprises to the assumptions made. One was the FEC data that was obtained in Excel Format, these scored significantly better than the New York Times Data Set, yet were still no where close to the scores of the API obtained Data sets. Another major assumption differing score was data scraped from OpenSecrets (no obtained from the API). This scored significantly better compared to the New York Times and the FEC Data. This was probably due to the method by which the Data was scrapped.

Possible Drawbacks to the method is understanding why certain attributes had missing data. It could be in place of a zero, meaning there was no actual information gained. Another issue involved proper case. The issue found was that there are some words that are common, such as ‘for’, which generally don’t need to follow proper case. These instances negatively impacted the second component; however, it generally occurred on attributes that are probably not going to be used for anymore than footnotes. The one major drawn back of proper casing occurs with names of candidates. The name Bill O’Reilly would be flagging as unclean data because the name is two words, yet there are three upper case letters. This should have some drawback, but the subset of names that follow this case are in the low minority and can easily be fixed when the merging occurs.