Part 2 Summary:

The study of potential relationships between congressional elections, industry contributions, and financial shifts continues into statistical testing and further analyses. In the first part, the established goal was to determine whether political contributions by industries to specific candidates had any bearing on election results. A second goal was to see if the election results made any specific impact to the market as a whole. The expectations were that a positive correlation between political contributions and winning an election exists, and that industries would benefit from contributing to the candidate(s) of their choosing.

Initial analysis of the data shows several great disparities within all of the main variables. Among the most noticeable are the total funds raised per candidate (min of $10, max of $21,830,000, mean of $865,900, and median of $566,300) and percentage of votes received per candidate (min of 0, max of 1.00, mean of 0.5289, and median of 0.55). These disparities indicate that the data required reshaping in order to achieve usable results when performing tests. Based on the summary statistics, outliers were identified and removed for further testing.

Histograms provided a high level view of the distribution of numerical and individual variables. Contribution amounts were used in different formats to provide visual conclusions. They show a remarkable difference between winning and losing candidates in total funds raised, contribution sizes by industry, as well as percentage of total funds raised per election. There is also a slight indication that Republicans tend to receive more contributions than their Democratic counterparts. Of the three variables used to test for correlation, the only identifiable correlation is between the percentage of funds candidates receive in their general election and the percentage of votes received. The correlation coefficient is 0.8734, which indicates a very strong correlation.

Apriori rule mining revealed many associations which were in line with initial expectations. While the identified rules strengthened the conviction that the dataset is representative of the real world, they failed to bring much new information to light.  The main takeaway from the frequent itemset mining is that in American congressional politics, challengers face an uphill battle; incumbents have a large fundraising advantage and win an incredibly high proportion of elections in which they participate. Deeper analysis is necessary to determine if the sources of candidates’ funding actually have an impact on election results or if it is only the amount of funding that matters.

Through network analysis of the 33 senators that won their elections in 2014, it is clear that only a few of the industries are responsible for contributing the most money to campaigns.  A caveat to this could be that our bins were too broad to see stark categorical differences between candidates.  One main conclusion to draw would be that industries hedge their bets, and would prefer to have influence over all candidates, regardless of who is elected.  Extension of this would be to check to see how the losers’ top contributors stacked up (though not necessarily with a network).  Since there are 33 winners, the dataset with outliers in other categories was used.  This seemed the most appropriate since the idea is just to compare the industries that link candidates.

The first null hypothesis is that there is no difference between the total contributions received by the incumbent and challenger. However, the student t-test shows that there is a very significant difference skewed towards the incumbent.  The second null hypothesis is that the winner will be based on the total amount of funds raised and the number of supporting industries. The precision of the model is 0.8125881, the recall is 0.8662994, and the F-measure is 0.8385846. From, the ROC curves, it seems that candidate total has a greater effect rather than the number of supporting industries. The third hypothesis is that the outcome of an election for a candidate is based on total contributions, the number of supporting industries, and the incumbency status. The decision tree predicts the results with a high degree of accuracy, precision, and recall. The Lazy Learner (K nearest neighbor) had significantly lower accuracy, but high precision, and recall.  The Naive Bayes model had the same accuracy rate as the Decision Trees, but a high recall rate.

In conclusion of this portion of the study, results of testing and analysis confirmed several expectations and discovered supporting evidence of commonly-held views of American politics. Testing and analysis of the data did not provide any significant evidence of unexpected information. The possibility remains that ground-breaking information can be produced, but it would exist beyond the scope of the data available.