Arif Ali (aa1605)
Joshua Kaplan (jk1805)
John Hotchkiss (jsh104)
Timothy Ahn (ta437)
Hongkai Wu (hw271)

Project 2 Write Up

Part 1: Exploratory Analysis

Basic Statistical Analysis and Data Cleaning

Data Merging and Cleaning:

We were left with a number of datasets after Project 1:

1. Election Results (From New York Times (NYT) and Federal Elections Commission (FEC))
    a. FEC 2004
    b. FEC 2006
    c. FEC 2008
    d. FEC 2010
    e. FEC 2012
    f. New York Times 2014
2. Candidate – Industry Connections
3. S&P Financial Data

Each of the above data sets needed to be combined before we could begin to address some of our analytical questions, begin generating descriptive statistics, and test hypotheses. For our group, this was a very large undertaking, with almost every group member doing a piece of it. We broke this into pieces, so first the election results files were stacked, and then the different types of file were merged.

For the election results, the FEC Data came in human-readable spreadsheets from the FEC and the NYT data was scraped. The FEC data needed to be converted to a machine-usable form then stacked, and finally collapsed. The NYT data needed to be reshaped. Once the FEC and NYT data sets were in similar shapes, we had to coerce the variables into the same formats so that when we stacked them to get the full year range we were seeking, the variables would be continuous and appear to have come from the same original file. This was particularly difficult due to the idiosyncrasies of the data, such as the format that each data originator used to store edge cases, such as races where a single candidate ran unopposed.

Note that as part of stacking and aligning the election results data, we also binned party affiliation into 3 bins: Republicans, Democrats, and Independents. In the FEC data in particular, candidates could be listed under multiple parties, subsets of parties, or state-specific parties, to name a few. To align these, we did our best to sort the parties manually based on brief searches of any general party affiliations to the larger Democrat or Republican parties through

Google. Those that didn't have a clear affiliation remained Independent. Candidates that were listed as multiple parties were converted to the bin of the strongest ideological party they were associated with (in the vast majority of cases, some version of Democrat or Republican was part of their association set). When both Democrat and Republican were listed, the one with a greater number of associated votes was chosen (for some FEC data, this was broken out).

Although we originally collected two different sets of data from Open Secrets, we decided that we only needed the information from one of them, therefore we were able to just clean up the one data set with candidate – industry connections.

The next step was to merge the election results (candidate performance during the election) with candidates' funding sources. This was also a considerable effort, beginning with yet more variable cleaning and cajoling to line up between the two files, and ending with a fuzzy merge between the files by name. In order to avoid merge mistakes, this merge was performed using a tiered strategy. For the first merge, the strictest merge rules were applied; among the use of other identifying characteristics of candidates, for the first merge they needed to match by all of the election information (state-year-district) and by full name (first and last). For the second tier, the same criteria were used excluding a first name match. For the third merge, the same criteria as the first merge were used, excluding a last name match. And for the fourth merge, clerical errors in candidate names were corrected so that they would merge. At every step in the merging process, the results were hand-checked to make sure no candidates were merged incorrectly.

The final group of data we merged originated from the daily historical stock information for all tickers that were listed in the Standard & Poor's 500 Index (S&P 500) as of October 14, 2015. This data is from Yahoo! Finance, via Quandl, which is a website that stores and shares financial datasets. The key pieces of information we needed to obtain were the changes in value of each industry on a monthly basis and from one election to the next, beginning on the first day of trading in 2014. This required a multi-step approach. Since the dataset was too large to fit in one CSV file, we had to split them into two separate files. This created duplication of data for 3M (ticker: MMM) and was removed. The two datasets were then merged into a single data frame. The variables kept included date, adjusted close price, ticker, and industry. We also found that there were some dates, such as holidays, where stock information for only a select few stocks were posted but should not have been included. Those dates were easily identified and removed based on frequency of occurrence for all tickers based on date. We then identified the last trading date of each month and removed all other dates. The final step was to add a new variable that gave us monthly changes in adjusted closing stock prices calculated by the quantity of the adjusted close price of each ticker in month i subtracted from the same value in month $i+1$, divided by the adjusted close price of each ticker in month i.

We ended up with two final datasets, to be used in separate analyses. The first dataset, PoldataSPIndustries, consists of, for each candidate/year/industry level observation from every election cycle from 2004-2014, the candidate's political party (party); campaign contribution

amount (amount) and percentage of total contributions (industrypercent) that come from the industry; total campaign contributions (candtotal); incumbent status (incumbent); number of votes received (votes) and percentage of votes received (percent; number of votes divided by total votes cast in the race); election winner status (winner); a variable illustrating how the industry's contribution to the candidate compares to the amounts contributed by other industries (indrank), the total amount of funding all of the candidates in the race received (racetotal), and the percentage of the total race funding that the industry gave to the candidate (racefundperc).

The data we originally scraped from OpenSecrets.org sorted campaign contributions into 95 different industries; in order to compare this data to stock market performance, we sorted these industries into the 10 sectors of the S&P 500[1], based off of descriptions of the OpenSecrets industries found on OpenSecrets.org[2]. Industries which did not fit into an S&P sector were sorted into 3 additional categories; not for profit, not publicly traded, and other. After sorting the OpenSecrets industries into S&P sectors, we collapsed the dataset on S&P sector, adding up the contribution amounts from the OpenSecrets industries contained in each S&P sector.

The second dataset, PoldataSPIndustriesStockData, in addition to all of the data in PoldataSPIndustries, contains data on stock market performance for each of the sectors in the S&P 500, for each election cycle from 2004-2012 (yrpercentchange). We calculated performance for each S&P sector by calculating the cumulative value of all stocks for each sector at the beginning and end of each election cycle (two-year periods) and finding the change in value for each sector. Since not all stocks were listed throughout each cycle, we only included the stocks that appeared at the beginning and end of each term. The 2014 election cycle had to be excluded from any analysis of the stock data, because we didn't think a metric based on the 9 months of data from the 2014 cycle that were available at the beginning of the project would be comparable to the metrics based on 24 months of data in the other election cycles. Since we still wanted to analyze the full political dataset, we decided the best approach would be to keep that dataset, and create a new one to look at the stock data.

Summary Statistics:

We had about the same number of observations for every year in both datasets; as discussed above, the dataset with stock data doesn't have any observations for the 2014 election cycle. There were significantly more Republican candidates than Democrats in both datasets, and both contained a small but not insignificant number of Independent candidates. At first we were surprised that our datasets contained so many more winners than losers, so many more incumbents than challengers, and so many more Republicans than Democrats. However, we realized that this was simply a result of having one observation per industry that donated to each candidate; Republicans tended to have more industries donating to them than

[1] https://eresearch.fidelity.com/eresearch/markets_sectors/sectors/sectors_in_market.jhtml
[2] https://www.opensecrets.org/industries/slist.php

Democrats, and winners and incumbents tended to have more supporting industries than losers and challengers, respectively. We verified this by shrinking the dataset down to unique year/state/race/candidate observations, and observing that the discrepancies vanished.

The PoldataSPIndustries dataset contains 35,082 year/state/race/candidate/industry level observations. In this dataset, the average amount contributed by one industry to one candidate was $122,100, but the standard deviation was nearly $300,000; some candidates received enormous amounts of funding from some industries, while others received very little. The same pattern was evident in the total amount of contributions each candidate received; the average amount was $865,900, but the standard deviation was more than $1.3M. Our analysis of the votes variable revealed that the average candidate received 194,000 votes, and that we had a number of NA observations for the number of votes. These NA's came from uncontested elections; for uncontested elections, the FEC didn't report the number of votes the candidate received. The average amount of funding per race was $1.7M, and the standard deviation was $2.6M, again showing that some races had much greater amounts of funding than others.

The PoldataSPIndustriesStockData dataset contains 20,248 year/state/race/candidate/industry level observations. It has fewer observations because we didn't have stock market data for the 2014 election cycle, and because three of the industries into which we sorted the OpenSecrets industries; not for profit, not publicly traded, and other; are not represented in the stock market. The average contribution amount per industry was $70,380, and the standard deviation was about $180,000. Total contributions per candidate averaged $859,300, with standard deviation of $1.3M. The average number of votes was 206,700, and the analysis revealed that we still had a significant number of NA observations in the votes variable. Total race funds averaged nearly $1.7M, with standard deviation $2.5M. With regard to the new variables from the stock market data, the average adjusted closing value was 2773, and the average year percent change was 28%.

When we took a closer look at the candtotal and votes variables, we discovered that we had a lot of values very close to zero, and just a few values at the high end of the distributions. We decided that for our analysis, we didn't want to look at the marginal candidates who only received a few votes or dollars. We also thought we should exclude some candidates at the top of the distribution, as they likely were special cases, and as such would exhibit different behavior from the middle-of-the-pack candidates we really wanted to look at. We also wanted to exclude any candidates whose ran in uncontested elections, for two reasons, because contributions data from these elections would also be systematically different from the middle-of-the-pack candidates. So, we removed any observations which had values greater than 1 interquartile range above the 75th percentile value for the vote percent or candtotal variables, as well as any observations which had values less than 1 IQR below the 25th percentile values of those variables. We also removed any observations for which we did not have voting results. For the dataset containing stock market data, we removed outliers based on the yrpercentchange variable. We did so because we believed that, in conditions of great economic turmoil, when an industry either rose or dropped a great deal, we would not be able to connect changes in the market value of the industry with the industry's political contributions, since

there were much greater forces at work causing the industry to move on the stock market. So, we removed any observations for which the value of the yrpercentchange variable was greater than/less than 1 interquartile range above/below the $75^{th}/25^{th}$ percentile values. After removing outliers and missing values, we were left with two datasets, containing 29,226 and 15,825 observations, respectively, summary statistics of which are listed in the Appendix.

Finally, it was important for the frequent itemset analysis to bin some of our numerical data into categories, so that we could see look at possible associations between the numerical variables and the other variables in our dataset. So, after outliers and missing values were removed, the total contribution amount and percentage of votes each candidate received in each election cycle was binned into categorical variables (candtotallevel and votepercentlevel, respectively) with four levels (very low, mid-low, mid-high, and high). These levels were calculated by dividing the total range of the variables into four equal segments, and sorting each observation into a segment. For candidate total funding, these bins were [$100, $431,000], ($431,000, $862,000], ($862,000, $1.29M], and ($1.29M, 1.73M], and for vote percent, these bins were [12.3%, 32.4%], (32.4%, 52.6%], (52.6%, 72.7%], and (72.7%, 92.9%].

Summary statistics before outliers were removal are presented in Table 1; summary statistics after outlier removal are presented in Table 2.

Table 1 - Summary Statistics Before Outlier Removal

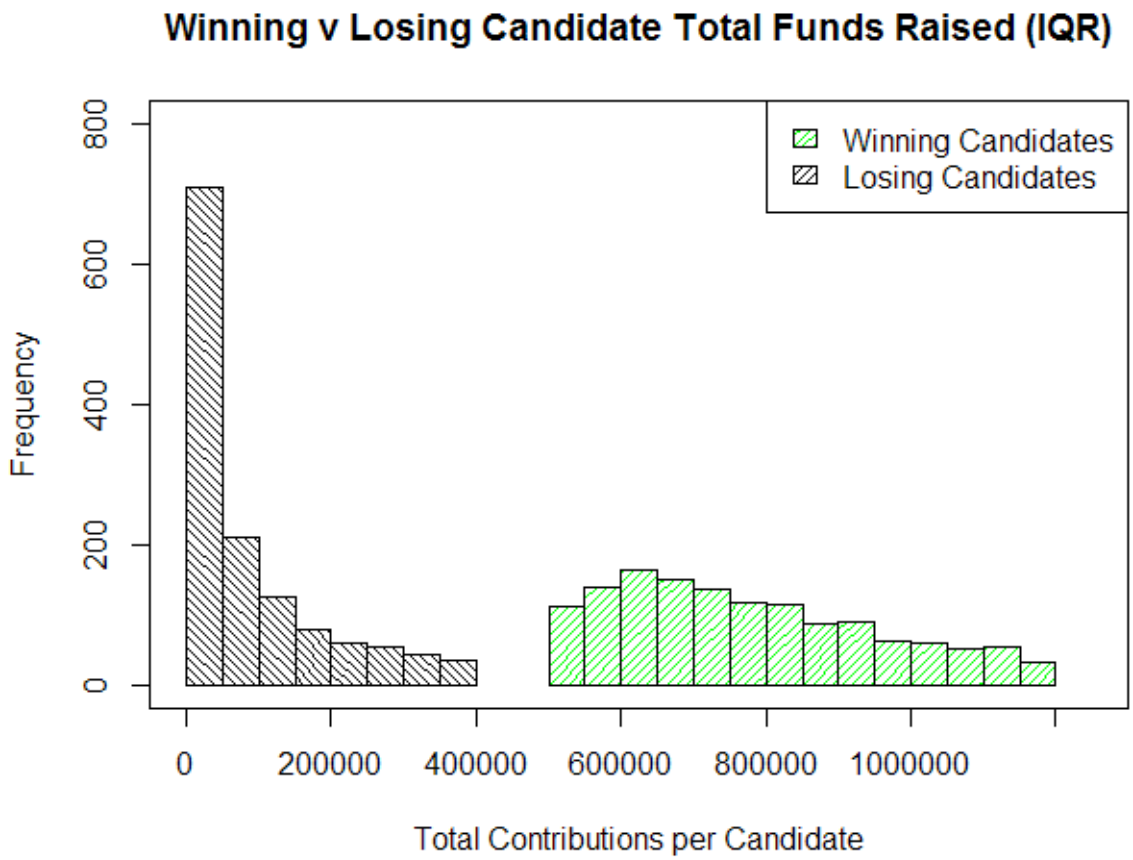
PoldataSPIndustries with outliers								
	Year	**Count**	**Party**	**Count**	**Winner**	**Count**		
	2004	5634	Dem	15539	0	14385		
	2006	5795	Rep	18284	1	20697		
	2008	5580	Ind	1259				
	2010	6184			**Incumbent**	**Count**		
	2012	5885			0	16022		
	2014	6004			1	19060		
Variable	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Std. Dev.**	**NA's**
Amount	10	10000	35950	122100	115600	8829000	296218.9	
Industrypercent	0.001293	0.033	0.08094	0.1538	0.1984	1	0.1839325	
Candtotal	10	198600	566300	865900	961700	21830000	1358704	
Votes	5	88960	125400	194000	174300	7865000	NA	431
Percent	0	0.3915	0.55	0.5289	0.6602	1	NA	431
Totalracefunds	72620	584500	907700	1714000	1743000	32870000	2588232	
Racefundperc	0.00001	0.1867	0.6948	0.5985	0.9842	1	0.3843594	
PoldataSPIndustriesStockData with outliers								
	Year	**Count**	**Party**	**Count**	**Winner**	**Count**		
	2004	3937	Dem	8629	0	7717		
	2006	4021	Rep	11011	1	12531		
	2008	3883	Ind	608	**Incumbent**	**Count**		
	2010	4294			0	8762		
	2012	4113			1	11486		
Variable	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Std. Dev.**	**NA's**
Amount	49	8500	25500	70380	66280	6525000	178624.5	
Industrypercent	0.001293	0.02676	0.05079	0.08256	0.1019	1	0.09603277	
Candtotal	130	246300	570300	859300	949200	21830000	1324489	126
Votes	5	96600	135800	206700	182600	7865000	NA	126
Percent	0	0.4038	0.5629	0.5414	0.6676	1	NA	
Totalracefunds	72620	568700	887100	1651000	1689000	32870000	2513277	
Racefundperc	0.0000266	0.2452	0.7355	0.6194	0.9854	1	0.3750487	
Adjclose	70.09	1487	2225	2773	3571	8184	1901.18	
Yrpercentchange	-0.6101	0.1545	0.3038	0.2818	0.5235	1.332	0.3822045	

Table 2 - Summary Statistics After Outlier Removal

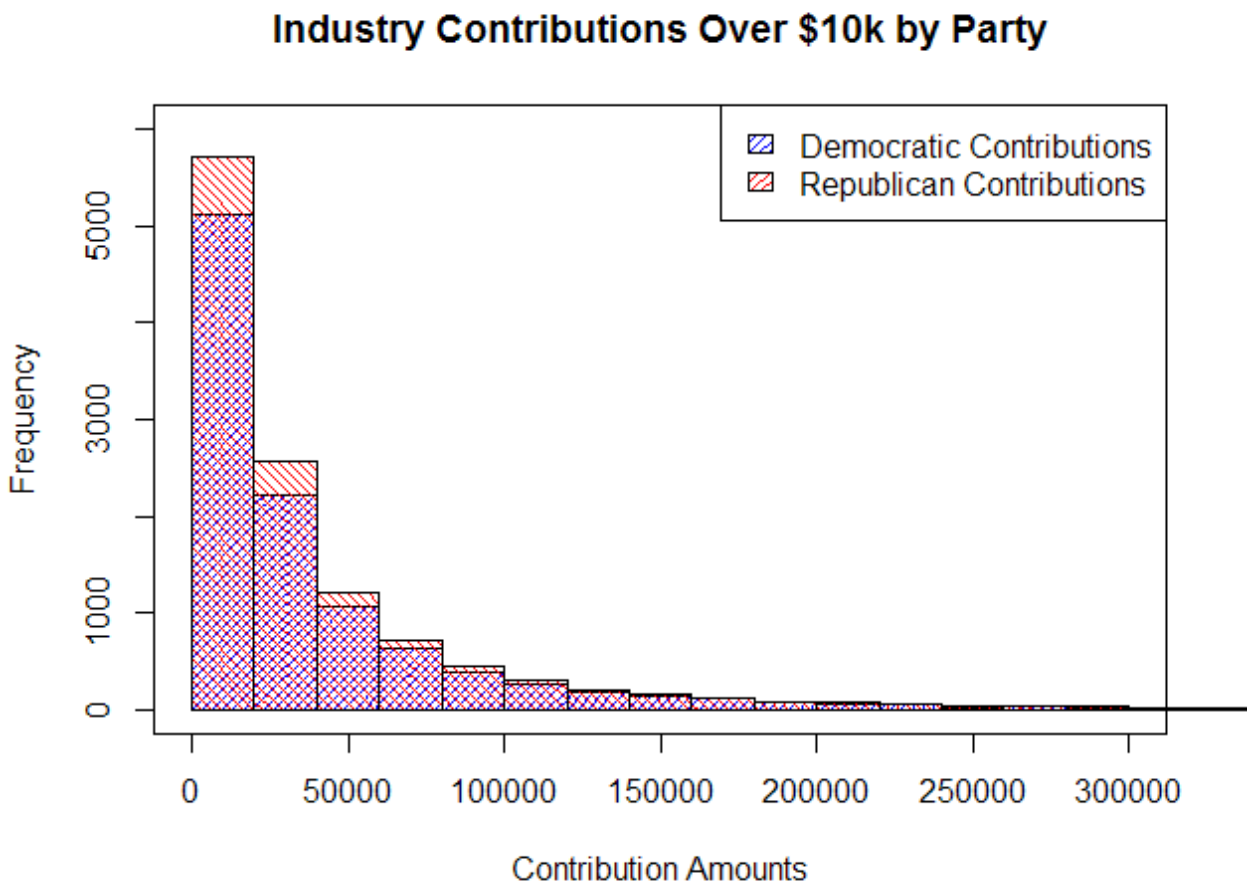
PoldataSPIndustries no outliers							
	Year	**Count**	**Party**	**Count**	**Winner**	**Count**	
	2004	4752	Dem	13239	0	11776	
	2006	4688	Rep	15794	1	17450	
	2008	4705	Ind	193			
	2010	5241			**Incumbent**	**Count**	
	2012	5042			0	13608	
	2014	4798			1	15618	
Variable	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Std. Dev.**
Amount	49	8100	29000	75040	88750	1379000	120266.7
Industrypercent	0.001293	0.03269	0.07939	0.1538	0.1947	1	0.1865573
Candtotal	100	140300	509900	546000	813300	1720000	428264.7
Votes	3713	85110	120100	123400	162400	259500	53989
Percent	0.0008	0.385	0.5551	0.5295	0.6658	1	0.1999814
Totalracefunds	72620	554900	826400	1151000	1378000	22530000	1271807
Racefundperc	0.000012	0.1557	0.696	0.5889	0.9845	1	0.3921799
PoldataSPIndustriesStockData no outliers							
	Year	**Count**	**Party**	**Count**	**Winner**	**Count**	
	2004	3486	Dem	6580	0	5962	
	2006	2201	Rep	8808	1	9863	
	2008	2728	Ind	437	**Incumbent**	**Count**	
	2010	3820			0	6886	
	2012	3590			1	8939	
Variable	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Std. Dev.**
Amount	49	7200	21250	42230	49170	1049000	64267.69
Industrypercent	0.001293	0.02578	0.04742	0.07874	0.09467	1	0.09401986
Candtotal	200	206400	520200	558300	810900	1720000	415769.5
Votes	3713	94380	130800	133000	174000	259500	53770.39
Percent	0.0009	0.4004	0.5717	0.5441	0.6726	1	0.1959033
Totalracefunds	72620	539800	811300	1098000	1311000	22530000	1105324
Racefundperc	0.0000336	0.2192	0.7553	0.6177	0.987	1	0.380596
Adjclose	70.09	1292	2244	2801	3889	8184	1981.615
Yrpercentchange	-0.2081	0.1607	0.3038	0.302	0.5235	0.8584	0.2569981

Histograms and Correlations

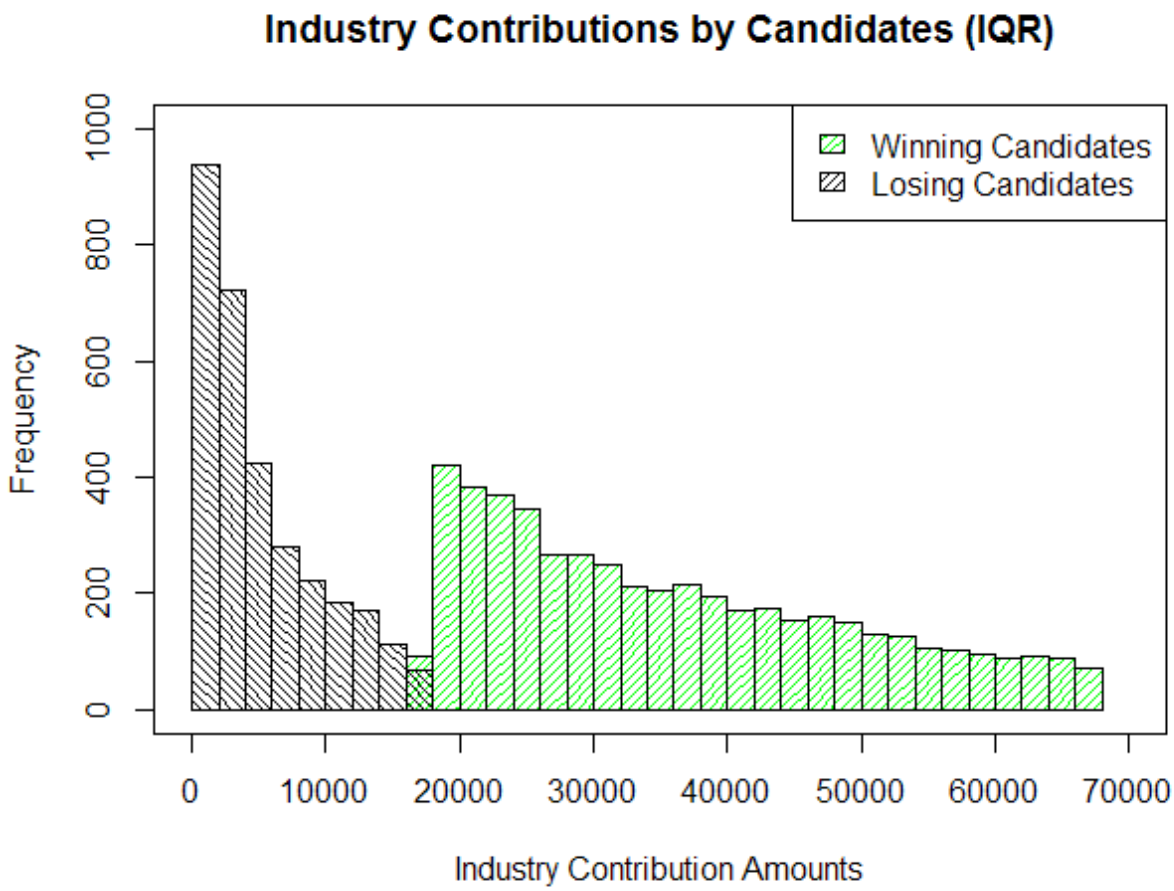
Histograms



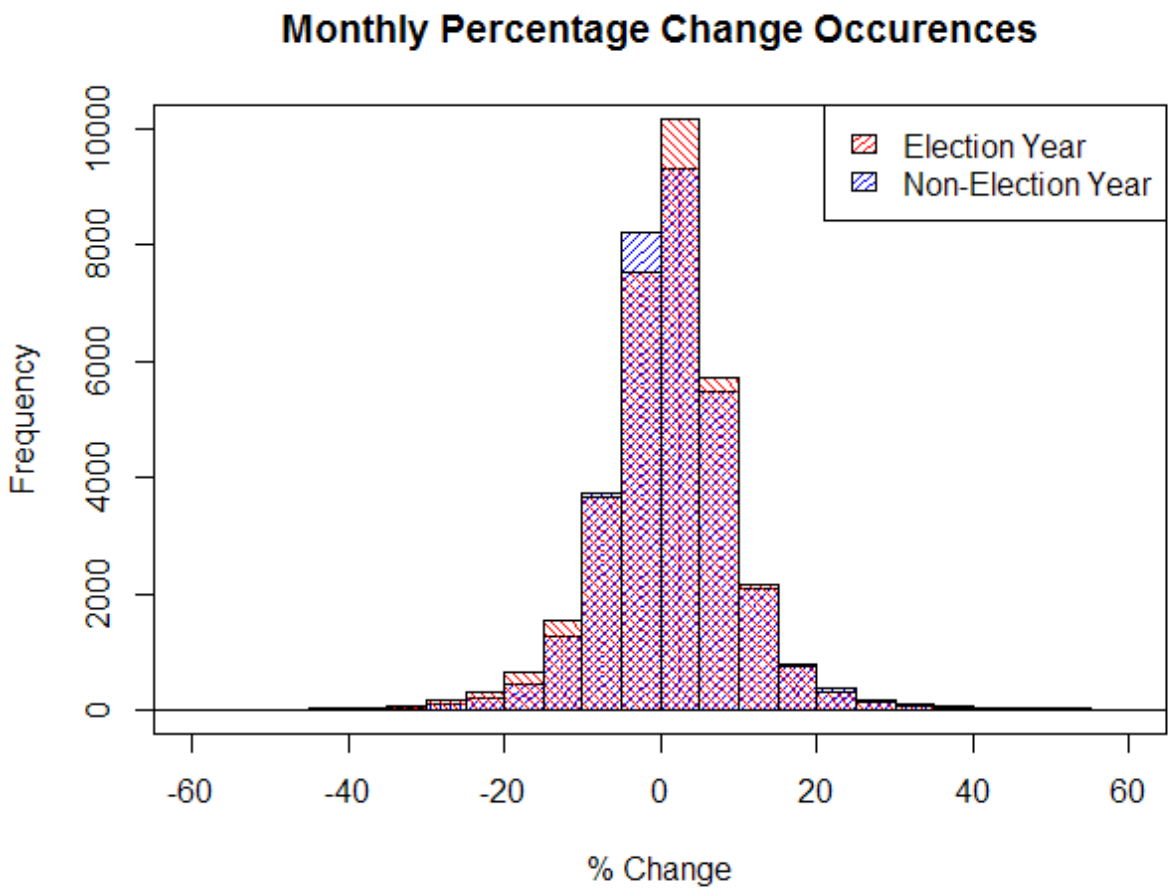
The above histogram provides the total amount of funds raised by each candidate for each Congressional race. Each race is distinguished by year, state, and district. There is a clear dichotomy of funds raised by winning and losing candidates. Not only does this show that the winning candidate always raises more funds than the losing candidate of a particular congressional race, but from the data used (interquartile range), every single winning candidate raised more funds than every single losing candidate of every single race. The data used was the interquartile range of total contributions, which provides a better visual representation without losing the utility of the histogram.



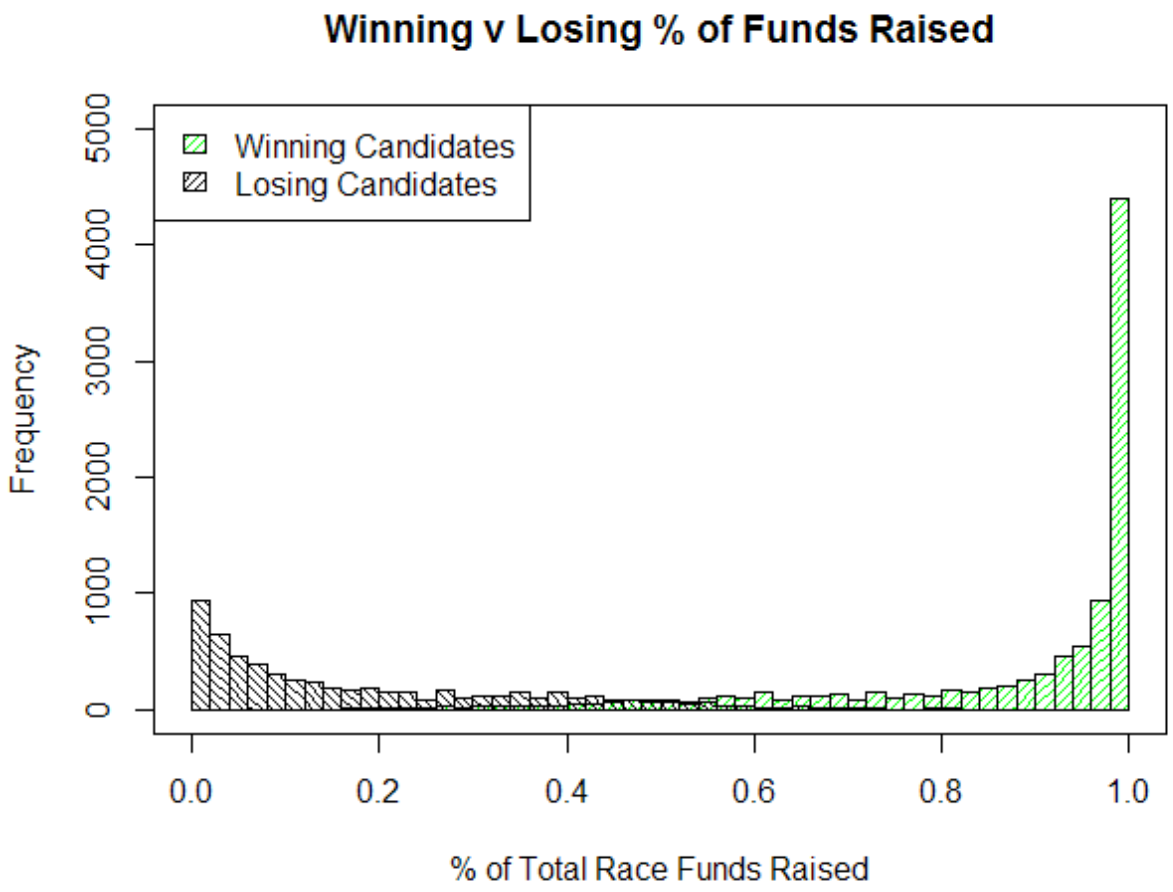
In an effort to show the difference in moneys contributed to the Democratic and Republican parties, the histogram provides visual evidence that Republicans have received the majority of all contributions. The individual contribution amounts are the collective sums of all funds raised from a particular industry to a specific candidate. At nearly every visible level, Republicans received more funds than their Democratic counterparts. Amounts below $10,000 and above $300,000 are considered outliers for the purposes of this histogram and are not represented to provide a better visual representation while maintaining the integrity of the analysis.



When breaking down each candidate's total funds raised by industry, the visual evidence appears to show either one of two things, or possibly both: candidates who eventually win their congressional race are better fund raisers, in general, or the more funds an industry contributes to a particular candidate, the higher the likelihood that their chosen candidate will win. In the latter case, this would lead us to believe that there is an underlying purpose of supporting one candidate over another which would be of some benefit to the particular industry itself. The data used was the interquartile range of total contributions, which provides a better visual representation without losing the utility of the histogram.



For each month from the beginning of 2004 to September 2015, the percentage change of all adjusted closing prices for stocks listed on the S&P 500 index were calculated and recorded. All percentages occurring in odd-numbered years were classified as “non-election years” while even-numbered years were classified as “election years.” The histogram shows a relatively normal distribution with the balance slightly in favor of positive values, and slightly better performance in election years compared to non-election years.



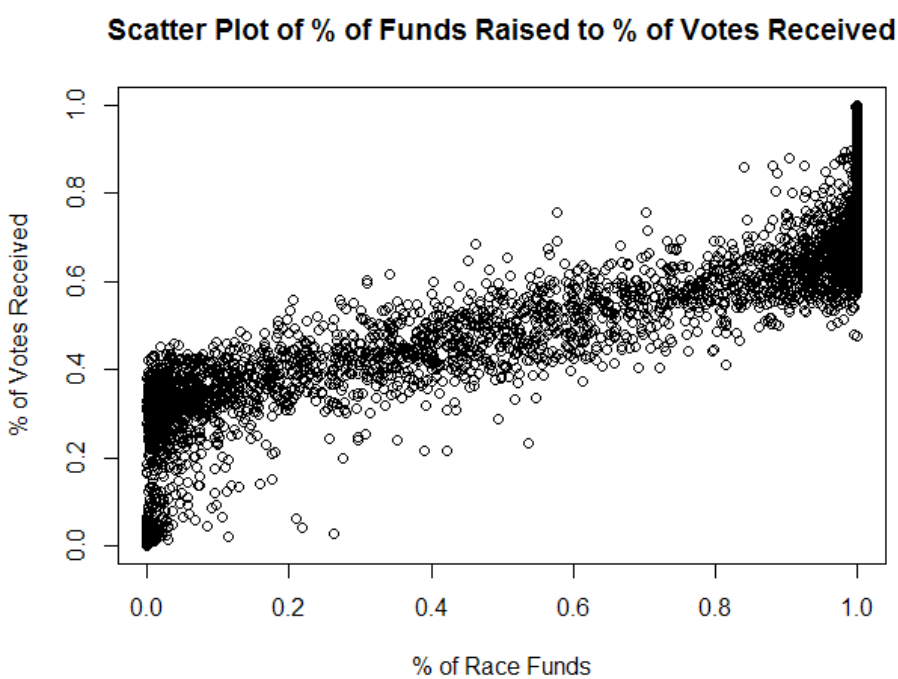
This histogram illustrates the widely-accepted belief that money wins elections. For each election race (distinguished by year, state, and district), each candidate was assigned a percentage of the funds they raised relative to the entire amount of contributions to all candidates involved in the race. It is evident that the higher percentage of funds a candidate raises compared to their competition, the more likely a victory will result. Interestingly, there are some candidates who were able to win their election with less than 20% of total funds raised in a race, and conversely, candidates with more than 80% of all funds raised who lost the election.

Correlations:

	%VotesReceived	%ofRaceFunds	IndValue%Change
%VotesReceived	1.00000000	0.87342601	-0.02004721
%ofRaceFunds	0.87342601	1.00000000	-0.01925322
IndValue%Change	-0.02004721	-0.01925322	1.00000000

The three variables used to test for correlation were the percentage of votes received by a candidate in an election, the percentage of funds an individual candidate raised from the total amount raised by all candidates, and the percent change of the value of an industry from the beginning of a congressional term to the end of the term. The percentage of votes received by a candidate is the most important factor in determining an election winner. The percentage of funds raised relative to the entire race gives us a measure of support from industries. The industry value percentage change will allow us to observe potential effects the contributions made to candidates may have.

%VotesReceived / %ofRaceFunds



As seen in the histogram "Winning v Losing % of Funds Raised," there is a very high correlation between the amount of funds raised by a candidate and the amount of votes they receive. A correlation coefficient of over 0.873 indicates a very strong linear relationship

between the two variables. In most cases shown in the scatter plot, raising more than 50% of a particular race's funds resulted in receiving more than 50 % of the total votes.

%VotesReceived / IndValue%Change

With a correlation coefficient of -0.020, there is virtually no correlation between the two variables. This result is unremarkable, as there is no intuitive relationship between the percentage of votes a candidate receives relative to the percent change in value of an industry that supports the candidate financially. There may be some type of relationship between the change in value of an industry with the winning or losing candidates that have particular stances on specific issues, but that is beyond the scope of our data.

%ofRaceFunds / IndValue%Change

The lowest apparent correlation is between the percentages of funds collected by a candidate and the change in value of an industry between votes. The calculated correlation coefficient is -0.019. Again, there would be no reason to believe that the percentage of funds a candidate collects would have any type of correlation with the change in value of an industry that contributes to the candidate without subsetting the data into more detailed variables.

Cluster Analysis:

Hierarchical Clustering

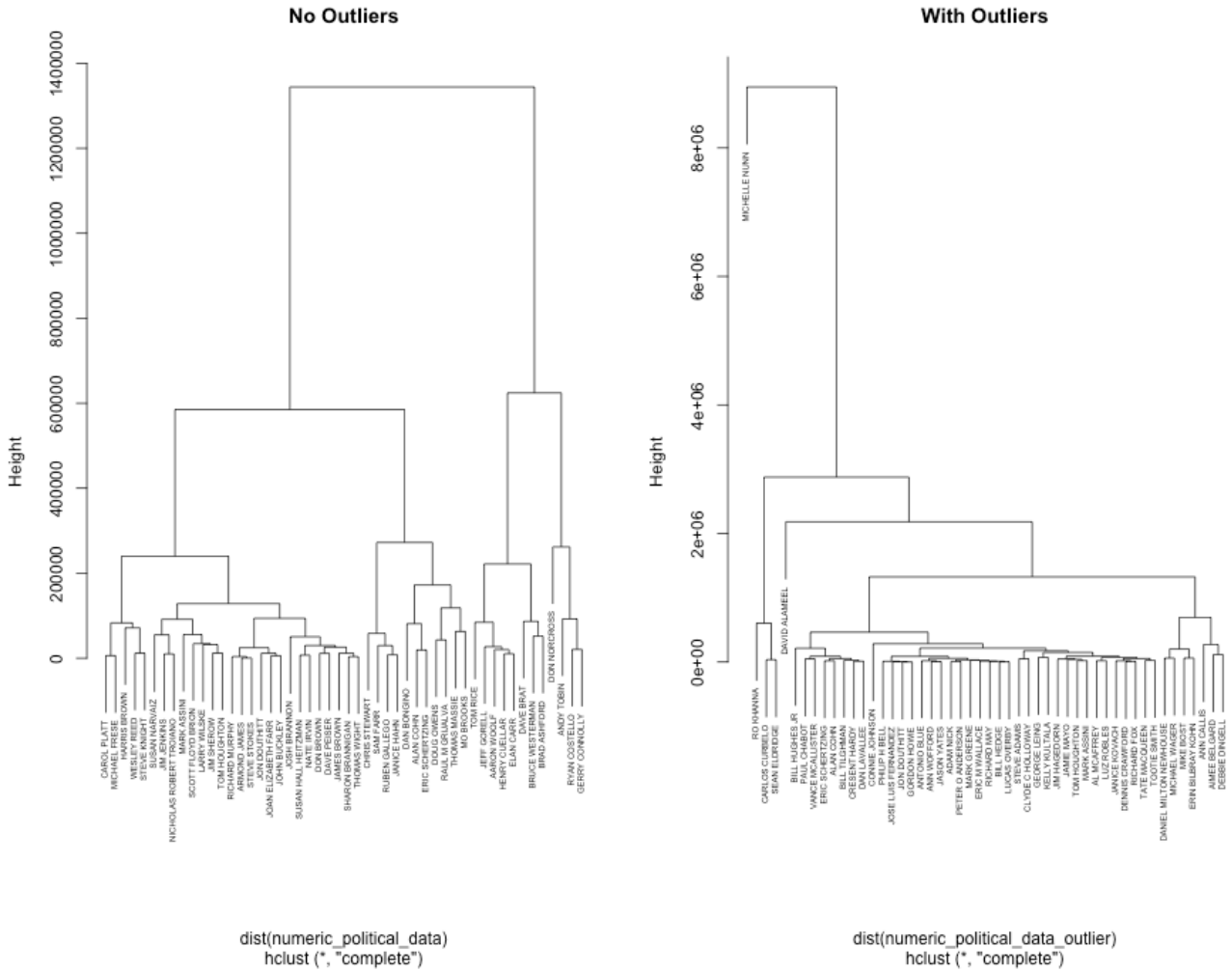
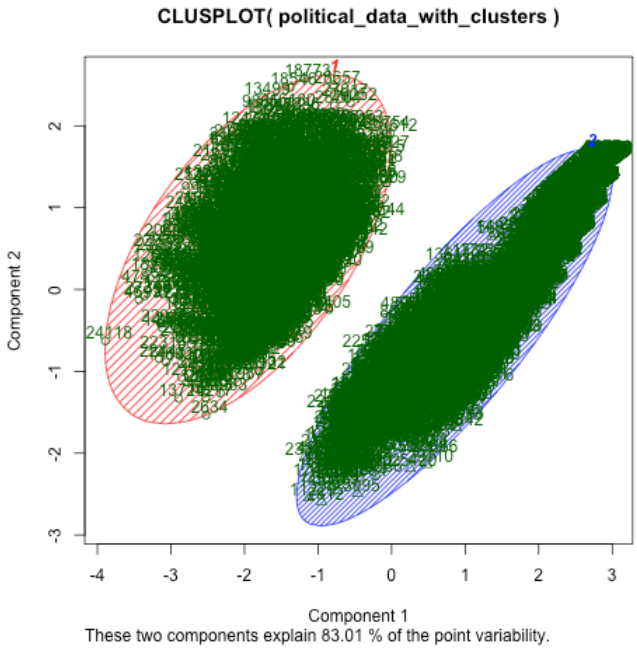


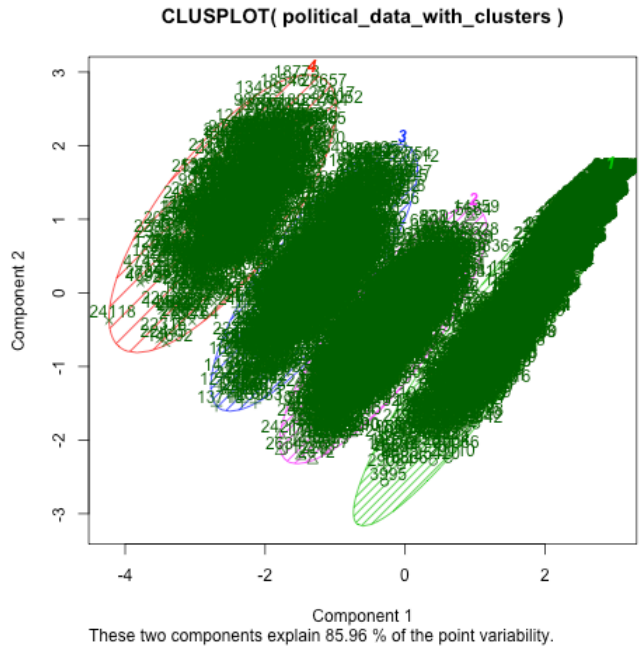
Figure 1 Cluster based on 2014 Congressional Candidates

For the purpose of making hierarchical clusters which weren't too cluttered, the clusters were plotted by year. The "closeness" of candidates is determined by total money contributed, the number of industries that backed them, and the number of votes they received. We also wanted to see the differences that including outliers in the plots would make. From each year, in order to make a readable plot, 50 candidates were sampled from the full dataset and 50 were sampled from the dataset with outliers removed. This was done because a Senatorial candidate tends to be determined from a significantly greater pool of votes than a House candidate. It's interesting to see that when outliers are excluded from the analysis, candidates in 2014 are extremely close (closeness is nearly zero). This could be because these candidates received a similar number of votes, amount of funding, or the same number of organizations

supported them. When outliers are included, the structure holds in similar ways, with the exception of some stand-alone candidates like Michelle Nunn and David Alameel. Both of these were democratic candidates in deep red states, so less funding or fewer votes could have resulted in both candidates being placed further away from the clusters. Please note, that for the years 2004, 2006, 2008, 2010, and 2012, there are corresponding plots in the submission folder.

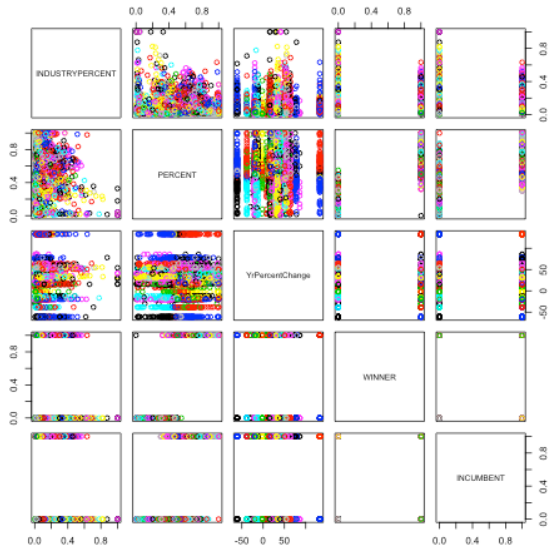
K-means Clustering

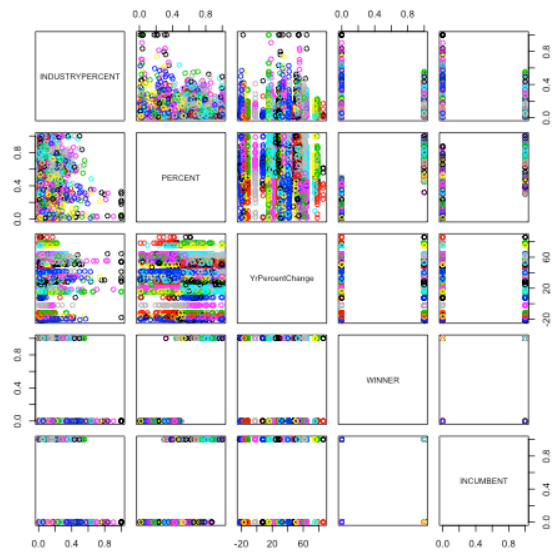




The above K-means clusters were created using the datasets that did not include stock data, after outliers were removed. The variables taken into account in creating the clusters were the total amount of money a candidate received, the number of votes they received, and the number of industries that supported them. One cluster plot was created so that there would be two clusters, and the other was created to show four clusters. The 2-cluster plot is interesting because it shows that there are stark differences in terms of class between two types of candidates. Based on the hypothesis tests, it is possible that these two groups might be incumbent vs. challenger or winner vs. loser. The plot with 4 clusters was based on the 4 possible pairings of winner/loser and incumbent/challenger. The assumption that could be made is that clusters 4 and 1 (see labels on plot) could be incumbents who won and challengers that lost and and 2 and 3 could be the remaining two cases. Unlike the k-means plot with 2 clusters, the four cluster plot did not show as much space between the clusters, possibly indicating that looking at four different cases isn't as telling as looking at only two cases.

DBSCAN Clustering





DBSCAN plots were created using the dataset which included stock data, with outliers and without outliers. The attributes analyzed were the percentage of contributions to each candidate by industry, the percentage of votes a candidate received, the percent change in stock price for an industry, and the winner and incumbent indicators. All plots comparing winners and incumbents create very polar structures along the borders. This is probably due to the fact that both winners and incumbents are binary (either 1 or 0). It's interesting to see the plots between voting "percent" and "yrpercentchange," an attribute based on an industry's stock price change during an election cycle. In both DBSCAN plots, there appears to be a linear-like structure occurring. Industry percent vs voting percent seems to be an incoherent scatter plot, which indicates that no significant relationship between the two attributes exists.

Association Rules / Frequent Itemset Mining Analysis

We ran association rule mining on the political subset of our data, consisting of political contributions to candidates by industry, as well as election results. Apriori rule mining was run to find rules with a minimum confidence level of .2, under three different support levels (.4, .2, .05). Eclat rule mining also was utilized with the same support level but, as it yielded nearly identical results to the Apriori algorithm, the analysis focuses on the rules generated by the Apriori algorithm. A selection of rules deemed interesting is found at the end of this section, and files containing all of the rules generated by the Apriori algorithm are available in the submission folder.

A variety of characteristics occurred frequently with the incumbent variable, a binary variable which represents whether or not a candidate was an incumbent. Incumbent and election winner occurred together in 50.9% of the observations in the dataset; incumbent candidates won 95.3% of reelection opportunities, and 85.3% of all election winners were

incumbents (rules 1, 2). This is in line with the historical average proportion of incumbents who win reelection[3], providing evidence that our dataset is representative of the real world. Another way in which the rule mining analysis mirrored popular opinions about American politics is in the relationship it revealed between a candidate's fundraising and election success. 72.4% of candidates who raised a very low amount of funding lost their election bids, and 74.9% of election losers raised a very low amount of funding, while 77.9% of those who raised a high amount of funding won their election bids (rules 3, 4, 5).

Challenger (non-incumbent) and election loser also frequently occurred together, in 37.8% of the observations. 81.1% of challengers lost, and 93.8% of losing candidates were challengers (rules 6, 7). The incumbent variable also had a strong relationship with candidate funding levels; in general, incumbents received more funding than challengers. Challenger was frequently associated with very low levels of campaign contributions, occurring together in 31.4% of observations; 67.5% of challenger candidates received a very low amount of funding, and 75.5% of all candidates who received a very low amount of funding were challengers (rules 8,9). On the other end of the funding distribution, 76.5% of candidates who received a high amount of funding were incumbents, and 70.1% of candidates who received a mid-high amount of funding were incumbents (rules 10, 11).

Apriori association rule mining revealed some interesting frequent itemsets featuring the political party variable. Over the time period examined (2004-2014), belonging to the Republican party was frequently associated with being elected. 62.6% of Republican candidates won their elections, and 56.6% of election winners were Republicans (rules 14,15). Interestingly, 57% of Democratic candidates also won election over the time period (rule 16). There weren't enough independent candidates in the dataset to show up in the association rules even at support level .05, but one takeaway from the Democratic and Republican results is that Independent candidates generally do not do very well in American congressional politics; out of the 193 Independent candidates in the dataset, only 23 were elected.

By ranking every industry that contributed to each candidate by contribution amount, we hoped to determine if the breakdown of a candidate's contributions affects the candidate's performance in elections. However, even with a minimum support level of .05, there were only two combinations of the industry rank variable, an industry, and another variable that occurred frequently enough to get picked up by the association rule generation. The "industry" not for profit occurred with an indrank rank of 1 (meaning nonprofits were the candidate's primary source of funding) and a very low level of funding in 5.4% of the observations; 49.3% of candidates whose primary industry was not for profit received a very low level of funding (rule 18). Considering that overall, 41.6% of candidates received a very low level of funding, this is not a very significant result; candidates whose main contributor was the not for profit industry were slightly more likely to receive a very low level of funding than the average candidate. Not for profit also occurred with an indrank of 1 and the election loser indicator in 5.8% of cases; 52.8% of candidates whose primary source of funding was nonprofits lost their elections (rule

[3] https://www.opensecrets.org/bigpicture/reelect.php

19). This is a fairly significant result; only 40.3% of candidates in the dataset lost their elections, so candidates whose largest contributor was nonprofits were much more likely than average to lose.

Apriori rule mining revealed many associations which were in line with our expectations; while the rules we found strengthened our conviction that our 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, life is hard for challengers; 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 candidate's funding actually have an impact on election results, or if it is only the amount of funding that matters.

Table 3 - Interesting Frequent Itemsets

	Rule	Support	Confidence	Description
1	{INCUMBENT=1} => {WINNER=1}	0.509512078	0.953451146	95.3% of incumbents won reelection
2	{WINNER=1} => {INCUMBENT=1}	0.509512078	0.853352436	85.3% of winners were incumbents
3	{CANDTOTALLEVEL=Very Low} => {WINNER=0}	0.301615	0.724322104	72.4% of candidates who raised very low amounts lost
4	{WINNER=0} => {CANDTOTALLEVEL=Very Low}	0.301615	0.748556386	74.9% of losing candidates raised very low amounts of money
5	{CANDTOTALLEVEL=High} => {WINNER=1}	0.055601177	0.779002876	77.9% of candidates who raised a high amount won
6	{INCUMBENT=0} => {WINNER=0}	0.378053788	0.811948854	81.1% of challengers lost
7	{WINNER=0} => {INCUMBENT=0}	0.378053788	0.938264266	93.8% of losers were not incumbents
8	{CANDTOTALLEVEL=Very Low} => {INCUMBENT=0}	0.314309177	0.754806902	75.5% of candidates who raised a very low amount were challengers
9	{INCUMBENT=0} => {CANDTOTALLEVEL=Very Low}	0.314309177	0.675044092	67.5% of challengers raised a very low amount of funding
10	{CANDTOTALLEVEL=High} => {INCUMBENT=1}	0.05460891	0.765100671	76.5% of candidates who raised a high amount of funding were incumbents
11	{CANDTOTALLEVEL=Mid-High} => {INCUMBENT=1}	0.111510299	0.70116179	70.1% of candidates who raised a mid-high amount were incumbents
12	{VOTEPERCENTLEVEL=Very Low} => {INCUMBENT=0}	0.114863478	0.995551601	99.6% of candidates who received a very low vote percent were not incumbents
13	{VOTEPERCENTLEVEL=High} => {INCUMBENT=1}	0.111510299	0.952923977	95.3% of candidates who received a high vote percent were incumbents
14	{PARTY=R} => {WINNER=1}	0.338328885	0.626060529	62.6% of Republican candidates won
15	{WINNER=1} => {PARTY=R}	0.338328885	0.566647564	56.6% of election winners were Republicans
16	{PARTY=D} => {WINNER=1}	0.257955245	0.569453886	57% of Democratic candidates won
17	{WINNER=1} => {PARTY=D}	0.257955245	0.432034384	43.2% of election winners were Democrats
18	{PRIMARY.INDUSTRY=Not for profit,indrank=1} => {CANDTOTALLEVEL=Very Low}	0.054061452	0.493133583	49.3% of candidates whose primary industry was not for profit raised a very low amount of funding
19	{PRIMARY.INDUSTRY=Not for profit,indrank=1} => {WINNER=0}	0.057893656	0.528089888	52.8% of candidates whose primary industry was not for profit lost

Network Analysis:

It is clear how the most straight forward implementation of a network to our data, or maybe the most obvious, might be bi-modal. Such a network would have nodes belonging to both candidates and industries. In particular, the visualization of such a network might make clusters of candidates much clearer. However, it is also clear that multimodal networks, even if the network under consideration is just a bimodal network, quickly become difficult to interpret and manage.[4] With this in mind, and in accordance with the assignment, we decided to consider the network of candidates and industries, but converted to the unimodal domain.

In order to make a suitably small matrix, we used a subset of our data to build it. The subset was determined by taking all of the winning candidates for senate seats in 2014, and their two largest sponsors. What we were hoping to see was something interesting about the primary industries that supported those candidates. Perhaps the Republicans and Democrats would be in two clusters, or perhaps different industries focused on different candidates by region of the United States. What we ended up finding was that it looks as though industries hedge their bets. Using the top two industries, as we did here, we connect all of the winning senators a remarkable amount. There is a strong possibility that this is because of the binning strategy applied to the industries. However, an analysis of the results follow below.

The resulting dataset after the subsetting above results contains 33 senators. These senators each have two rows with industries that could connect them directly with other candidates. Keep in mind the following summary statistics for the senators' dataset:

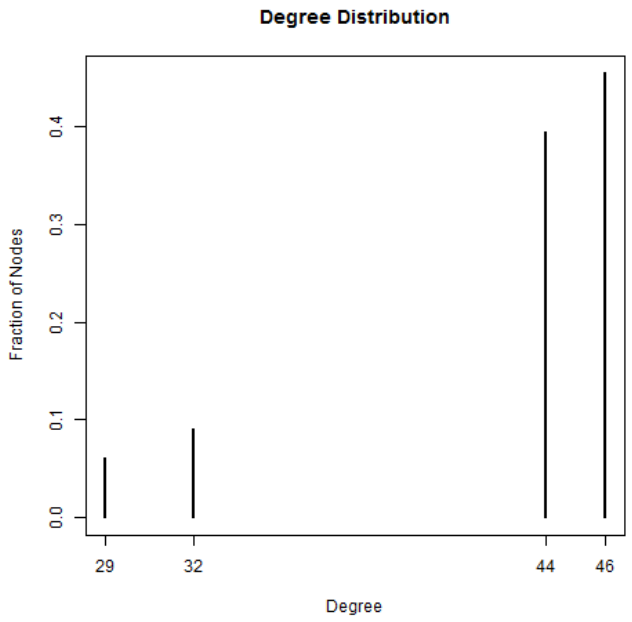
Industry	Consumer Staples	Energy	Financials	Not for profit	Not publicly traded
Count	1	1	18	30	16

After converting our network using a strategy by Solomon Messing found in his blog, we applied the example from Blackboard and implemented our network using the R igraph package.[5] Since the original dataset had 33 senators, we end up with a network with 33 nodes. Amazingly, with just the top two industries per candidate, the candidates can connect to one another with a mean degree of 42.9 (we also tried this statistic after removing multiple edges, finding a mean degree of 31.6). Of course, the only way to get more than 32 degrees is to connect to a candidate through both of the top two industries. Only 5 of the 33 candidate failed to connect to every other candidate. Please find a degree distribution for the network with multiple edges, below.

---

[4] Source: Scott Weingard, "Networks Demystified 9: Bimodal Networks", http://www.scottbot.net/HIAL/?p=41158

[5] Source: blog, https://solomonmessing.wordpress.com/2012/09/30/working-with-bipartiteaffiliation-network-data-in-r/

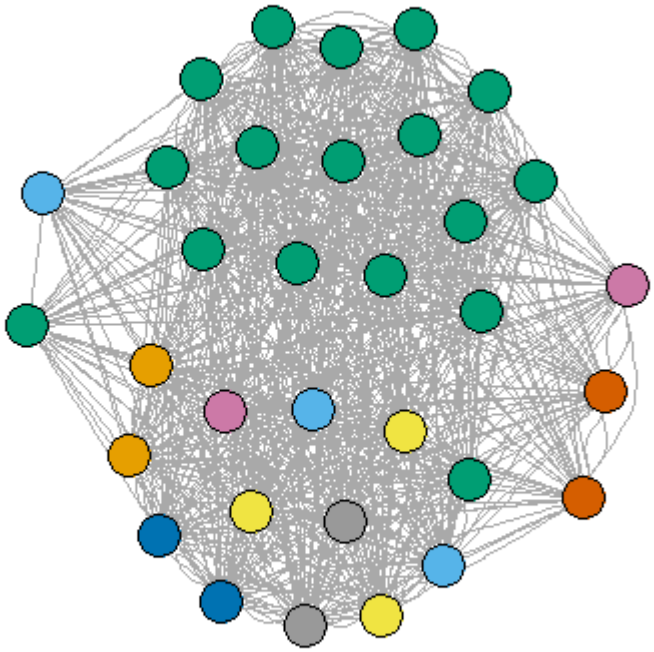


Unsurprisingly, this was also borne out in the betweenness statistic, so it didn't provide any additional value. The 5 that didn't connect to every other candidate had betweenness scores of 0 since the other candidates don't need to pass through any nodes to get to everyone. Every other candidate had a betweenness score of 0.2142857 due to the 5 being able to cross through them 1 time to get to any that they can't get to otherwise.

The mean clustering coefficient was 0.56. For the 5 less connected candidates, in two cases there was a coefficient of 1, and for the other 3 it was close to 1 at 0.88. All of the other were around 0.5. This shows that 2 of the 5 less connected candidates were the only winning candidates for one of their sponsoring industries (of those where that industry was one of the top two sponsors). That would be the only way to get a coefficient equal to 1. The other three, however, share whatever industry it is that they have in common, bringing down the coefficient.

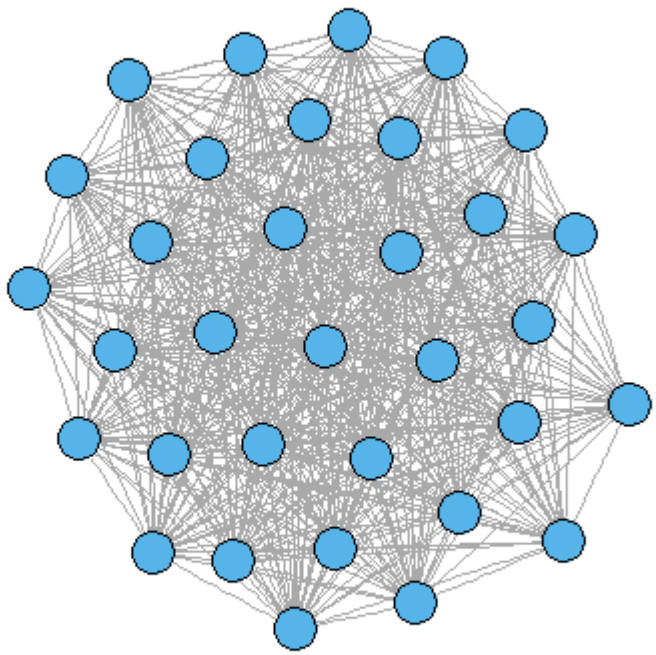
Given the above statistics, we already have a pretty clear idea of the remaining statistics. The graph density is greater than 1 due to the multiple edges at 1.34. The graph diameter is equivalent to 2, which indicates that even the 5 candidates that stand out have a first connection the same network. The number of connected components is equivalent to 1 (If just one industry is picked for each Senator – essentially turning the network into clustering – there are four components). Finally the largest k-core was 39. These are all indicators of just how many of the senators share both top supporters.

Please find the edge-betweenness network, below.



As you can see, and as we guessed based on the network statistics, the network is very intertwined. We did take a look at and tried to analyze the green cluster in particular, and this appears to be associated primarily with candidates that had both “Not for Profit” and “Financials” as their supporting industries. It is not clear why this cluster was able to snag the two dots not in its prime area. It is not clear why the bottom cluster was not also colored the same, as it must represent the combination of “Not for Profit” and “Not Publicly Traded”. The five that are more separated from the rest are clearly our 5 disparate points, with the two loners on the left holding “Consumer Staples” and “Energy” as one of their two top supporters, respectively, and the three on the right sharing “Financials” and “Not Publicly “Traded” in common.

Please find the modularity network, below. For this network, multiple edges were removed since that was a requirement of the greedy algorithm to run on this network.



Not surprisingly, the greedy algorithm determined that all of the nodes were part of the same community. Once multiple edges are removed it is very hard to see any significant pattern in the visualization of the network.

Part 2: Predictive Analysis

Hypothesis Testing:

Student's T-test

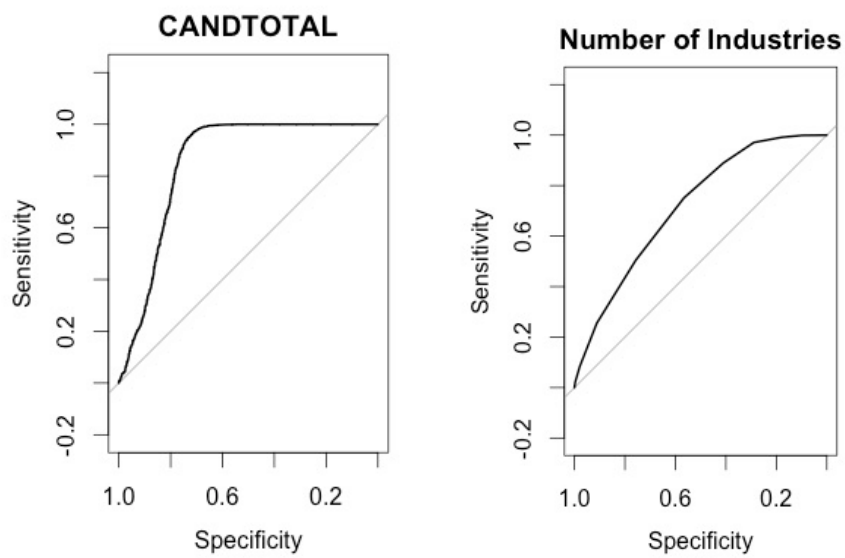
For the Parametric Statistical Tests, the dataset, "PoldataSPIndustriesStockData no outliers", was used. The student's t-test was run on one of the three hypotheses developed, while the logistic regression model was run on a possible linear relationship. In hindsight, it would have been interesting to run cross-validation on the model, but, unfortunately, time was a factor in forgoing it. ROC curves were provided on both of the attributes tested by the logistic regression model.

The first null hypothesis tested was: there is no difference between the total contributions to an incumbent and the total contributions to a challenger. We ran the student's

t-test (not pairwise) to test this null hypothesis. We got a p-value of <2.2e-16 and, given that this a social science analysis, the threshold for statistical significance should be .05. The p-value crosses this threshold and is well within the rejection region, so it is statistically significant. Thus, we reject the null hypothesis in favor of the alternative; there is a difference in the amount of contributions incumbents and challengers receive. From looking at the means of the two categories of contributions, it is clear that the incumbent receives a higher amount of contributions than the challenger. This is also verified by looking at the association rules.

Logistic Regression

The second null hypothesis that was tested involved a logistic regression model. The idea behind the model is to predict who the winner will be based on the total amount of money raised and the number of supporting industries. From the confusion matrix, we know the following: the precision of the model is 0.8125881, the recall is 0.8662994, and its F-measure is 0.8385846. Below are the ROC plots of both variables. The accuracy against the training data is 0.8184791; however, the number seems low because the prediction should have better matched against the actual results. This may mean that the data might not be best tested using a model that assumes normally distributed data.



From the ROC curves, it seems like the logistic model predicts with a surprising degree of accuracy. This could signify that the model may be over fitting. This is probably because the model is predicting the values based upon itself. As might be expected, the total amount of money raised seems to have a bigger impact on accuracy than the number of industries.

Confusion Matrix:

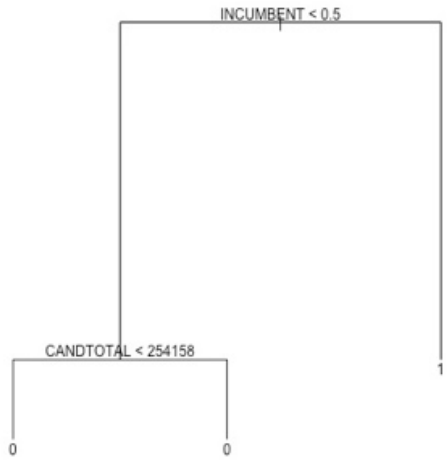
	lose	win
Lose	1273	399
win	267	1730

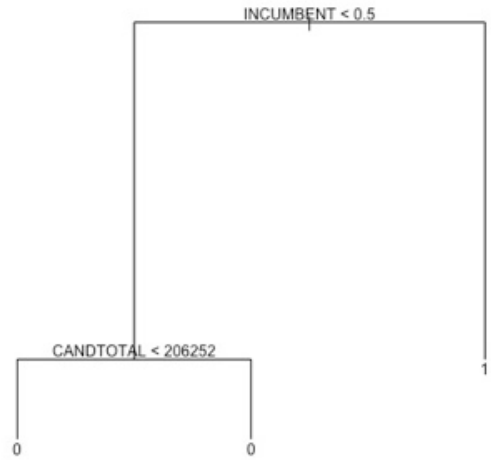
Decision Tree

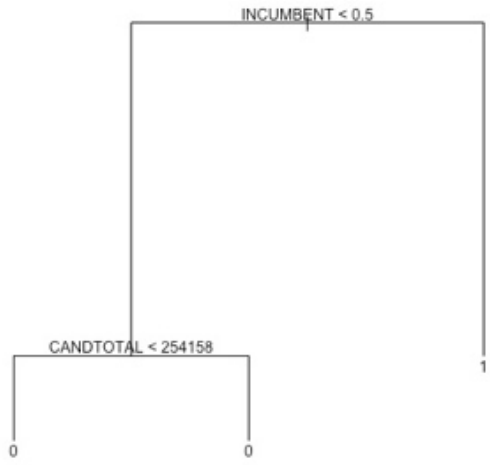
For the data driven predictive models, we looked into the following third hypothesis: the outcome of an election for a candidate, is determined by total contributions, the number of supporting industries, and the candidate's incumbency status. For each of the three models, cross-validation was run 5 times. The training set comprised 80% of the dataset, selected at random, and the test dataset comprised the remaining 20%.

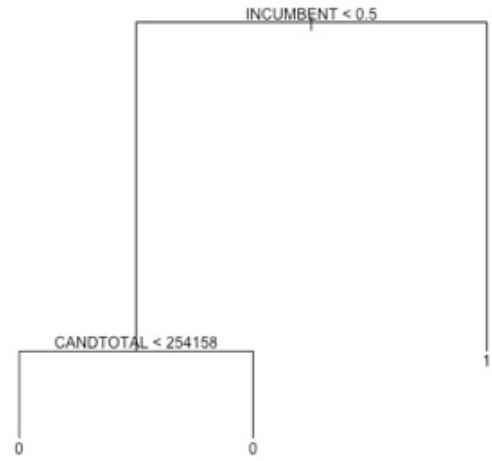
Cross-validation of the decision tree model generated the following trees and parameters (please note that for all confusion matrices, 0 denotes losing and 1 denotes winning).

Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
0 / 1 0: 314 / 70 1: 16 / 334	0 / 1 0: 316 / 58 1: 14 / 346	0 / 1 0: 313 / 60 1: 20 / 341	0 / 1 0: 321 / 59 1: 17 / 337	0 / 1 0: 332 / 52 1: 21 / 329
accuracy: 0.882834	accuracy: 0.901907	accuracy: 0.891008	accuracy: 0.896458	accuracy: 0.900545
precision: 0.826733	precision: 0.856436	precision: 0.850374	precision: 0.85101	precision: 0.86352
Recall: 0.954286	Recall: 0.96111111	Recall: 0.944598	Recall: 0.9519774	Recall: 0.94
F-measure: 0.88594	F-measure: 0.90576	F-measure: 0.89501	F-measure: 0.89867	measure: 0.90013

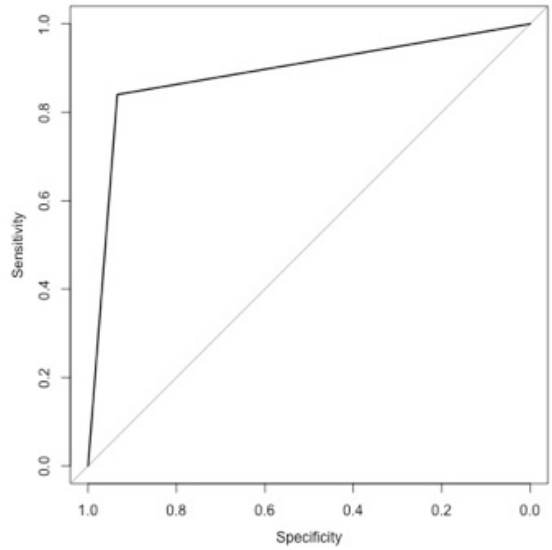


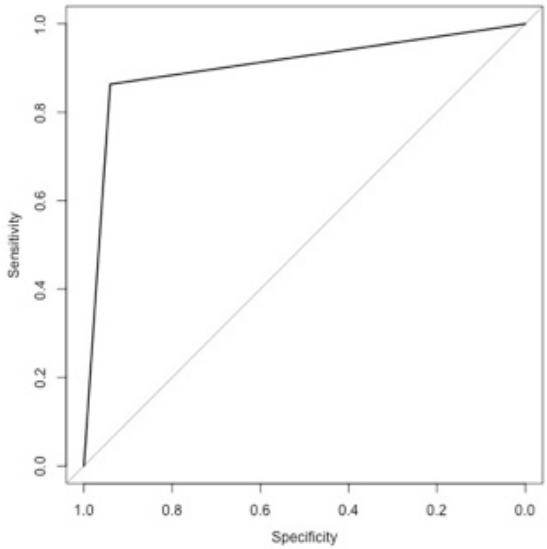


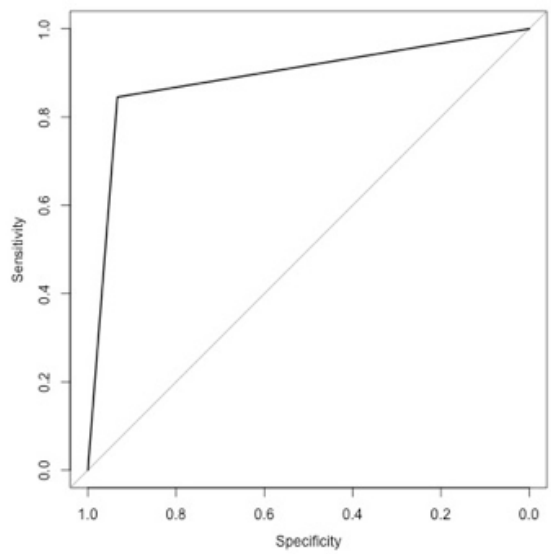


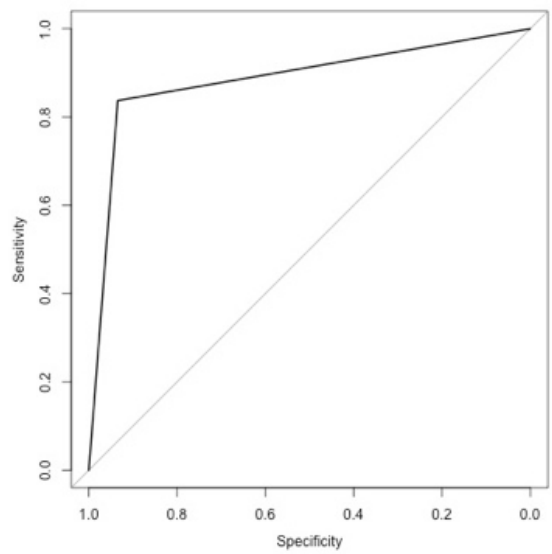


Due to an issue in the for loop for the cross-validation, we weren't able to output the fourth of the five decision trees. However, from structure of the four trees shown above, it is obvious that the candidate total variable does not significantly impact whether a candidate wins or not, given their status as incumbent or challenger.







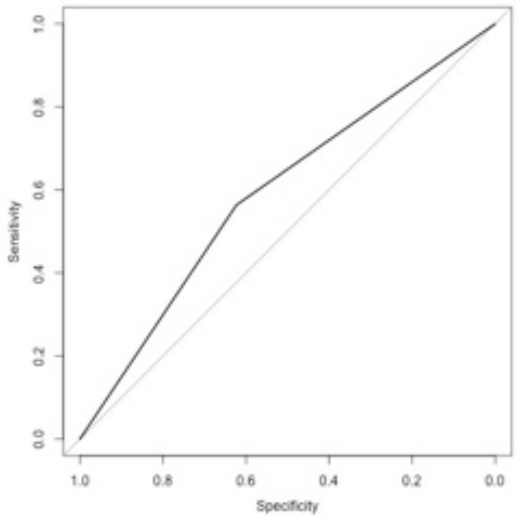


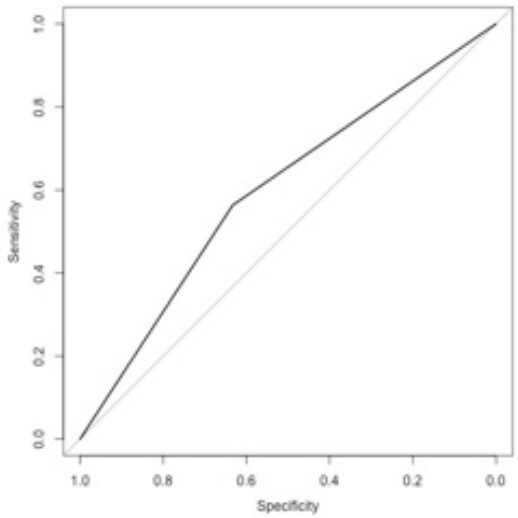
From the ROC curves, it seems like the model does a good job of predicting which candidate will win. Something to note is that the decision tree seems to associate winning the election with being an incumbent. The accuracy rates, tree structures, and confusion matrices all also seem to confirm this idea. This model confirms a deeply held belief about congressional politics; incumbents win. However, the model may not be the most useful because its precision is low relative to the other parameters.

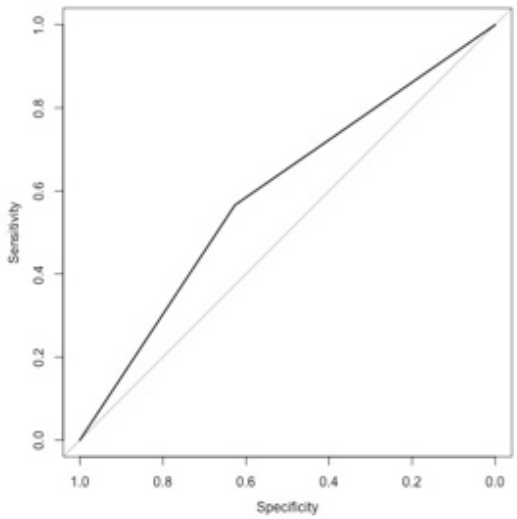
Lazy Learner

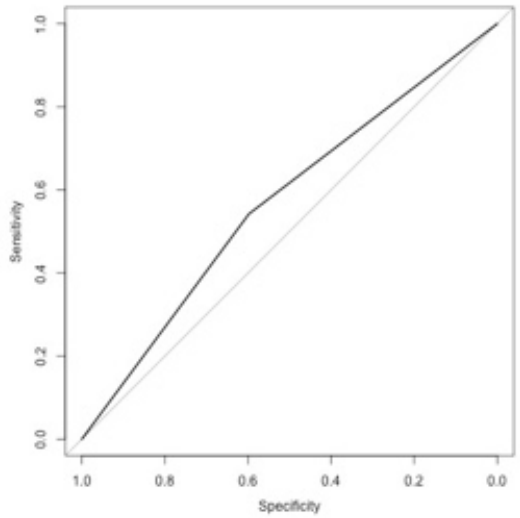
For the KNN (lazy learner) algorithm, the k value was benchmarked for values between 1 to 20; the benchmark was based on finding the k value that resulted in the most accuracy. This could have been improved if, for each k, the model was tested approximately 1000 times, but given the size of the dataset, one time was enough. In the end, the K value that was used was 17.

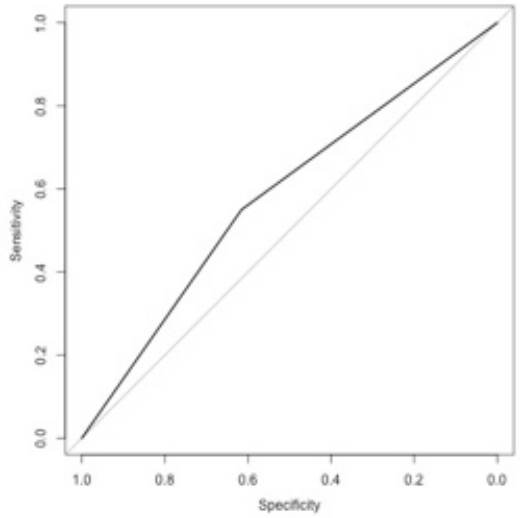
Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
0 1 0 256 16 1 86 376	0 1 0 244 18 1 80 392	0 1 0 247 14 1 96 377	0 1 0 236 13 1 89 396	0 1 0 229 20 1 111 374
accuracy: 0.576294	accuracy: 0.611716	accuracy: 0.572207	accuracy: 0.588556	accuracy: 0.5722
precision: 0.959184	precision: 0.956098	precision: 0.96419	precision: 0.96822	precision: 0.94924
Recall: 0.81385281	Recall: 0.830508	Recall: 0.797040	Recall: 0.816495	Recall: 0.771134
measure: 0.880562	F-measure: 0.88889	F-measure: 0.87269	F-measure: 0.88591	F-measure: 0.85097


Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0


Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0


Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0


Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0

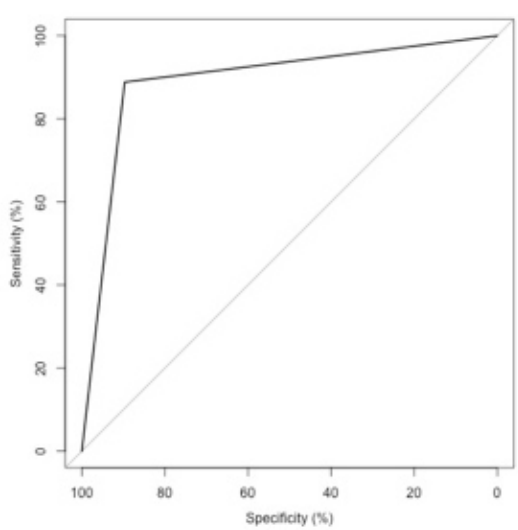

Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0

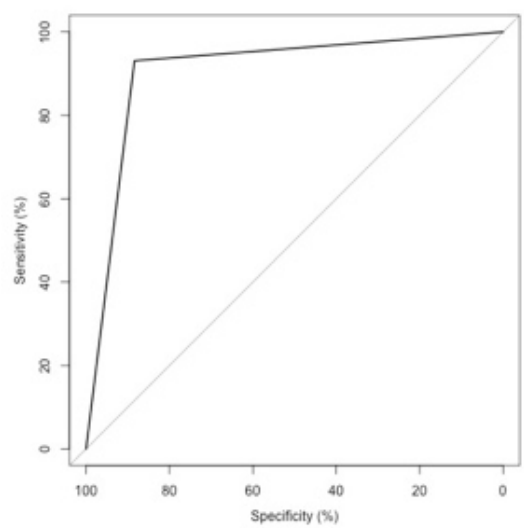
Based on the accuracy, this model is much less accurate than the decision tree model. On the other hand, the precision is very good, meaning that the amount of true positives is very high. The downsize is that there is a significant amount of false positives that are detected. This model may not be the best for overall detection. The issue may be a result of the heavy influence of the incumbent variable.

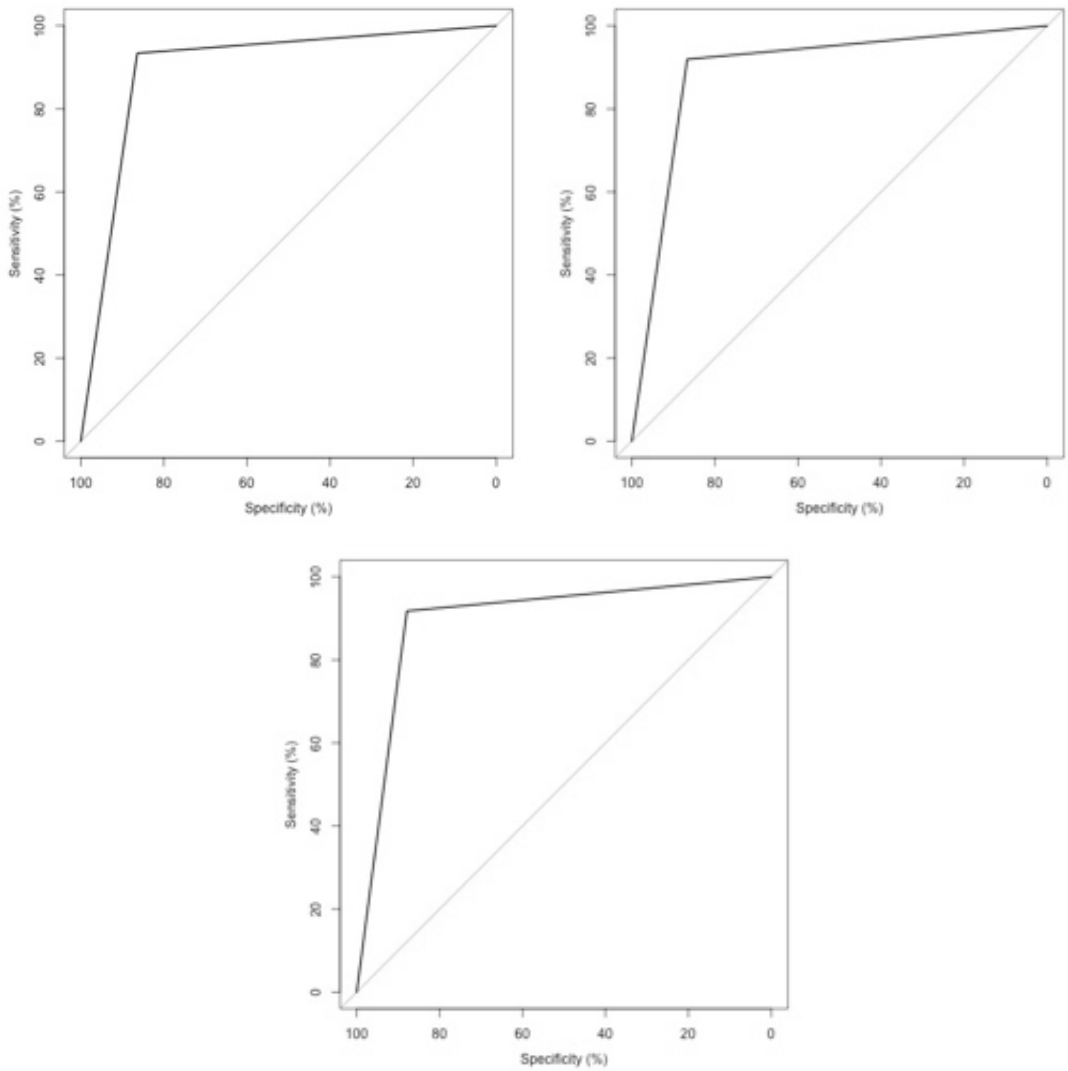
Naïve Bayes

For the Naïve Bayes algorithm, the following results were observed in terms of confusion matrices, accuracy of model, precision, recall, and the F-measure:

Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
0 / 1 0: 333 / 54 1: 19 / 328	0 / 1 0: 324 / 60 1: 18 / 332	0 / 1 0: 327 / 58 1: 17 / 332	0 / 1 0: 321 / 62 1: 15 / 336	0 / 1 0: 328 / 49 1: 18 / 339
accuracy: 0.900545	accuracy: 0.89373	accuracy: 0.897820	accuracy: 0.895095	accuracy: 0.908719
precision: 0.858639	precision: 0.84694	precision: 0.85128	precision: 0.84422	precision: 0.87371
Recall: 0.94524496	Recall: 0.9485714	Recall: 0.9512893	Recall: 0.95726496	Recall: 0.9495798
F-measure: 0.89986	F-measure: 0.89487	F-measure: 0.8985	F-measure: 0.8972	F-measure: 0.91007







From the results of the confusion matrices, the accuracy, precision, recall, and F-measures, are equal to or better than the other two models. The ROC curves cover a bigger area than than either KNN or Decision trees. The false negative numbers are significantly smaller as well, which is also evidenced by the values for the precision and recall.

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 had any specific impact on 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. A logical 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; contributions are skewed in favor of incumbents. 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 candidate's 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.