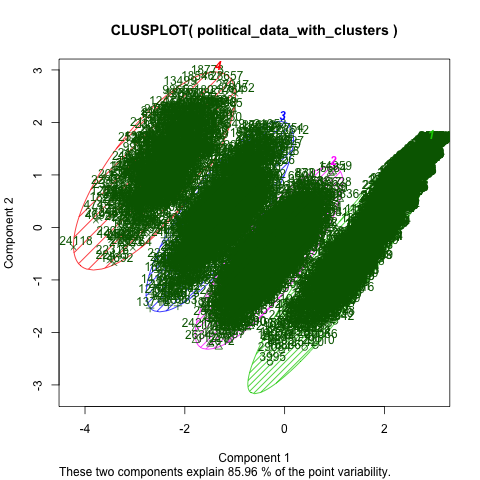
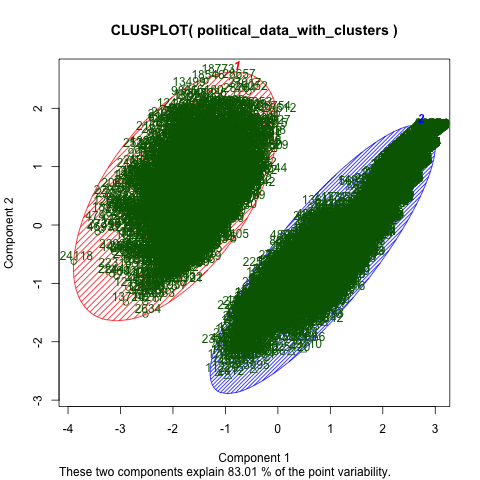
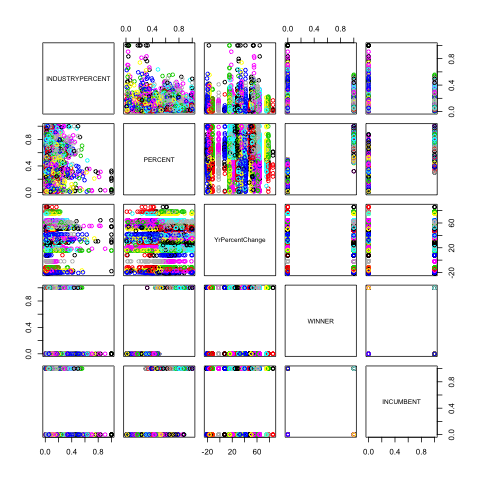
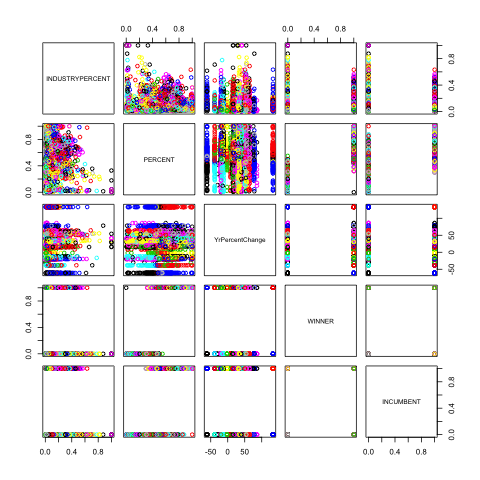


Figure 1 Cluster based on 2014 Congressional Candidates

When attempting the hierarchical clusters, the clusters were based and plotted by year. The “closeness” of candidates is determined by total money contributed by organizations, the number of organizations backed them, and the number of votes that they received. I wanted to also see the differences that the plots would have with or without outliers. From each year, in order to make a readable plot, 50 candidates were sampled from an outlier free dataset and 50 were sampled from a 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 the initial closeness between candidates in 2014 (with outliers removed) is close to zero. This could be because these candidates have close to the same number of votes, amount of funding, or the same number of organizations support them. In the plot including outliers, 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 less 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.



The above K-means clusters were created using the non-outlier datasets that did not include stock data. The cluster was created looking at the total amount of money a candidate received, the number of votes they received, and the number of organizations supporting the the candidate. One cluster was created so that two centers would be created and the other was created with the idea that four centers would be created. 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 these two groups might be incumbent vs challenger or winner vs loser. The plot with 4 clusters was based on the 4 possible pairing of winner vs loser and incumbent vs. 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 2-cluster kmeans, the one with four clusters did not show as much space between division, possibly indicating that looking at four different cases isn’t as telling as looking at only two cases.



DBscans plot were created looking at the dataset including the stock data with outliers and the dataset without outliers. The attributes analyzed were: the industry percentage of contributions to candidate, the percentage of votes a candidate received, the percent change of stock price for an industry and whether the candidate won or is the incumbent. 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 attributes based on an industry’s stock price change during a relevant year. 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 probably indicates no significant relationship between the two attributes appears to exist.