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))
   1. FEC 2004
   2. FEC 2006
   3. FEC 2008
   4. FEC 2010
   5. FEC 2012
   6. 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.

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]](#footnote-1), based off of descriptions of the OpenSecrets industries found on OpenSecrets.org[[2]](#footnote-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 STATS:

Full summary statistics for every variable in the dataset are reported in tables 1 and 2 of the appendix.

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 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 75th/25th 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%].

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **PoldataSPIndustries dataset** | | | | | | | | |
|  | **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 dataset** | | | | | | | | |
|  | **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 |  |
| Votes | 5 | 96600 | 135800 | 206700 | 182600 | 7865000 | NA | 126 |
| Percent | 0 | 0.4038 | 0.5629 | 0.5414 | 0.6676 | 1 | NA | 126 |
| 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 |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |

1. https://eresearch.fidelity.com/eresearch/markets\_sectors/sectors/sectors\_in\_market.jhtml [↑](#footnote-ref-1)
2. https://www.opensecrets.org/industries/slist.php [↑](#footnote-ref-2)