Data Merging and Cleaning:

John cleaned, merged FEC data, cleaned and combined NYTimes data with FEC. He also cleaned and merged the OpenSecrets files.

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 averaging the change in the adjusted closing price (variable adjcluse) of the S&P stocks in each sector over each election cycle. 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:

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **PoldataSPIndustries no outliers** | | | | | | | |
|  | **Year** | **Count** | **Party** | **Count** | **Winner** | **Count** |  |
|  | 2004 | 4992 | Dem | 13540 | 0 | 12465 |  |
|  | 2006 | 5093 | Rep | 15873 | 1 | 18024 |  |
|  | 2008 | 4853 | Ind | 1076 |  |  |  |
|  | 2010 | 5477 |  |  | **Incumbent** | **Count** |  |
|  | 2012 | 5136 |  |  | 0 | 14258 |  |
|  | 2014 | 4938 |  |  | 1 | 16231 |  |
|  |  |  |  |  |  |  |  |
| **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 |

Once we had our two base datasets, and had created all of the new variables needed for our analysis, we set about removing outliers and any missing values. Many of the candidates in our dataset had tiny amounts of total contributions, and a few candidates had relatively huge amounts

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)