DA 606 Final Project

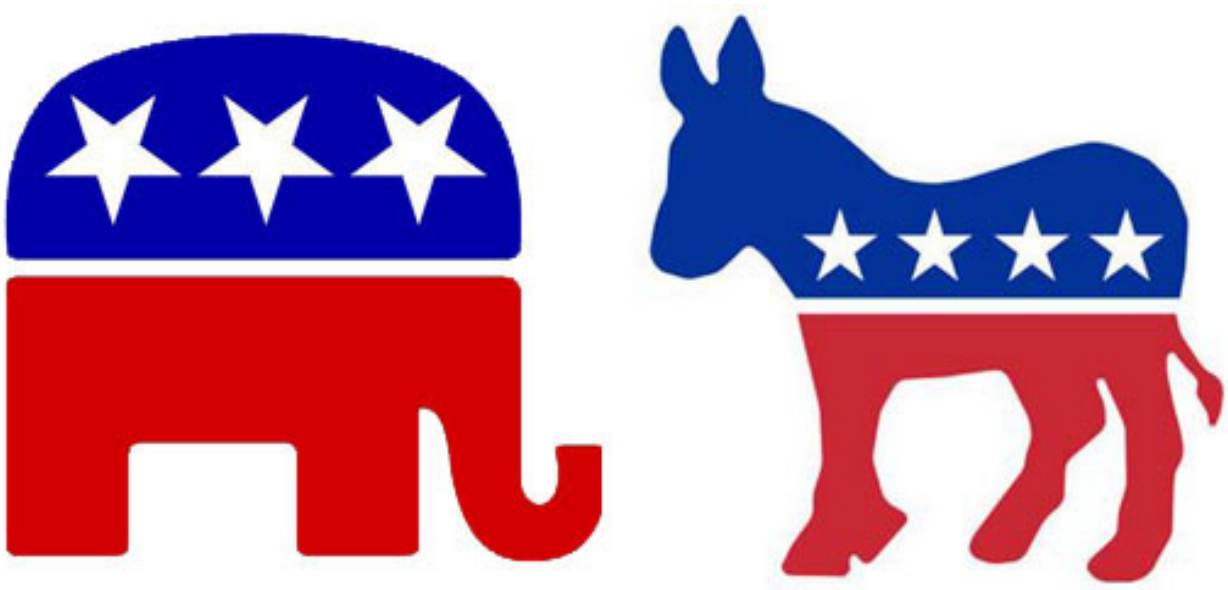
Contents

1. Introduction:

Research question:

What can we learn about the American presidential election process by studying the candidates:

- Number of Contributions
- Total Contributions
- Average Contribution
- Where the candidate falls on the political spectrum
- Who received the largest donations?
- How big are the donation? (1st-3rd Quartiles)
- Are there donation patterns across the left-to-right political spectrum?
- Are there relationsips between donations and other features of the candidate?
- Does testing a subset of the data, due to hardware limits, produce a statistically accurate snapshot of the donations?



Why we should care:

- We should care about the number of, sizes, and totals of contributions to help us see if our government is truly serving its people, or rather serving special interests.

2. Data:

Data collection:

- Election Data was downloaded and loaded from: http://www.fec.gov/disclosurep/PDownload.do
- Political Spectrum information about the candidates was taken from: http://www.huffingtonpost.com/findthebest-/every-2016-candidate-from_b_7562176.html

Cases:

- Each case in this study was an individual campaign contribution.

Variables:

- The two (groups of) variables that will be studied are 1) Contributions, and 2) Candidate Political Spectrum Score:
- Note that after analysis, the most relevant of the contribution statistics will be used (of contribution total, contribution mean, and contribution_count)

Type of study:

This is an observational study, and not an experiment one? (Explain how you've arrived at your conclusion using information on the sampling and/or experimental design.)

Scope of inference:

- Generalizability: The population of interest is the American voting public of 2016. The findings from this analysis need not be generalized as they are the full representation of that data. They may be generalizable to some extent from election year to election year, but as national situations and priorities change, the dynamics of the election model may change too. There is bias in this data in assuming that all contribution sources are the same. There are many attributes that make up an individual, and none of these have been taken into account in this project.
- Causality: I believe the data can be used to establish causal links between the variables of interest. The average donation sizes, in my opinion, seem to be the greatest indicator of the wishes of the Americans with less money, while the largest donations will show the wishes of the Americans with the most money.

3. Exploratory data analysis:

3A) The Candidates, Donations, and the Candidates' Spectrum:

The base donations data is loaded. The finished product relies on reports generated from 2016_donations_ALL.csv (about 225 MB), while during development, smaller samples are used, such as 2016_donations_5000.csv. Below is the code that generated those samples based on the large dataset, as well as the loading of the csv into a data frame:

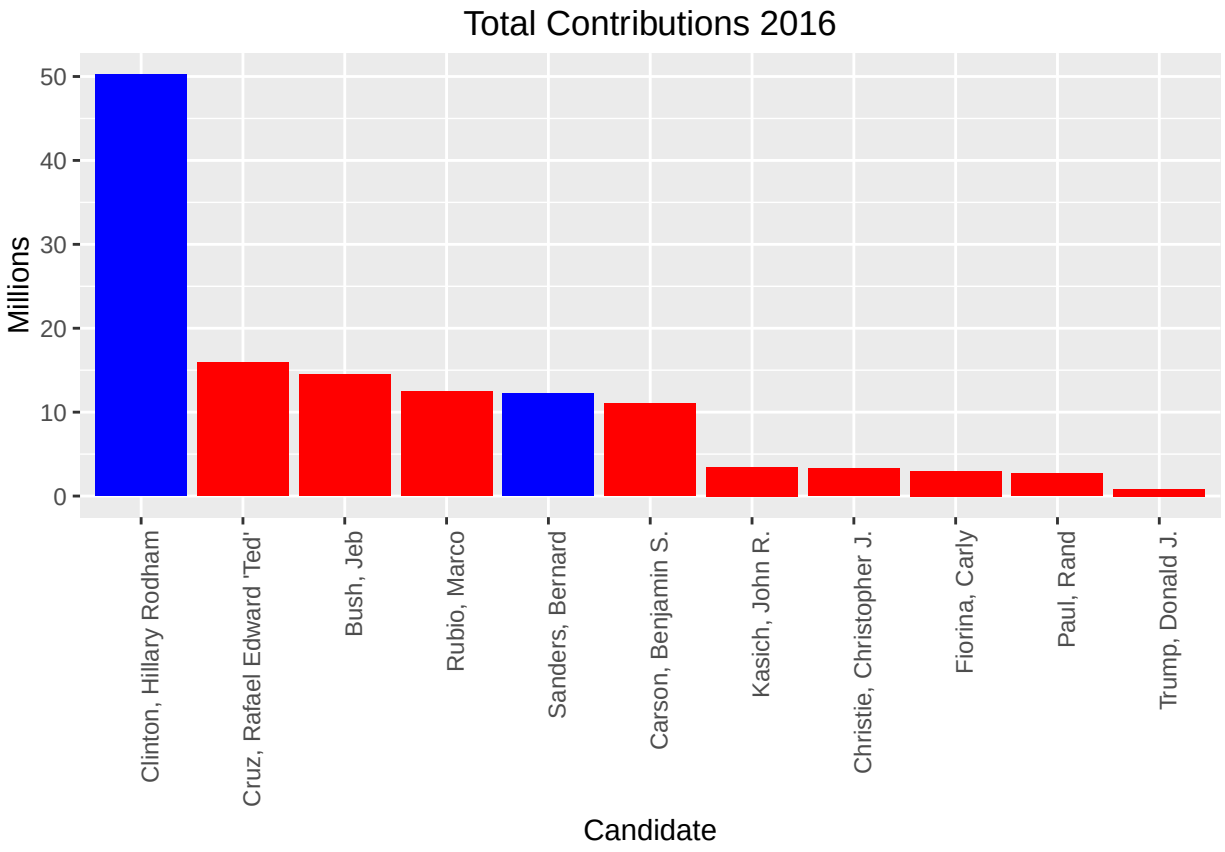
Next, since the presidential candidates were a small set of data, data was manually copied from http://www.huffingtonpost.com/findthebest-/every-2016-candidate-from_b_7562176.html into a data frame.

Below are the relevant features from this set of data:

	cand_nm	contbr_city	contbr_st	contb_receipt_amt	spectrum_score	party
25200	Clinton, Hillary Rodham	NEW YORK	NY	1797625	-6.8	Democrat
447586	Clinton, Hillary Rodham	NEW YORK	NY	1467071	-6.8	Democrat
447767	Kasich, John R.	CINCINNATI	OH	25000	5.0	Republican
30771	Carson, Benjamin S.	ASPEN	CO	18000	5.8	Republican
301848	Kasich, John R.	COLUMBUS	OH	16200	5.0	Republican
420274	Bush, Jeb	PITTSBURG	TX	15000	4.8	Republican

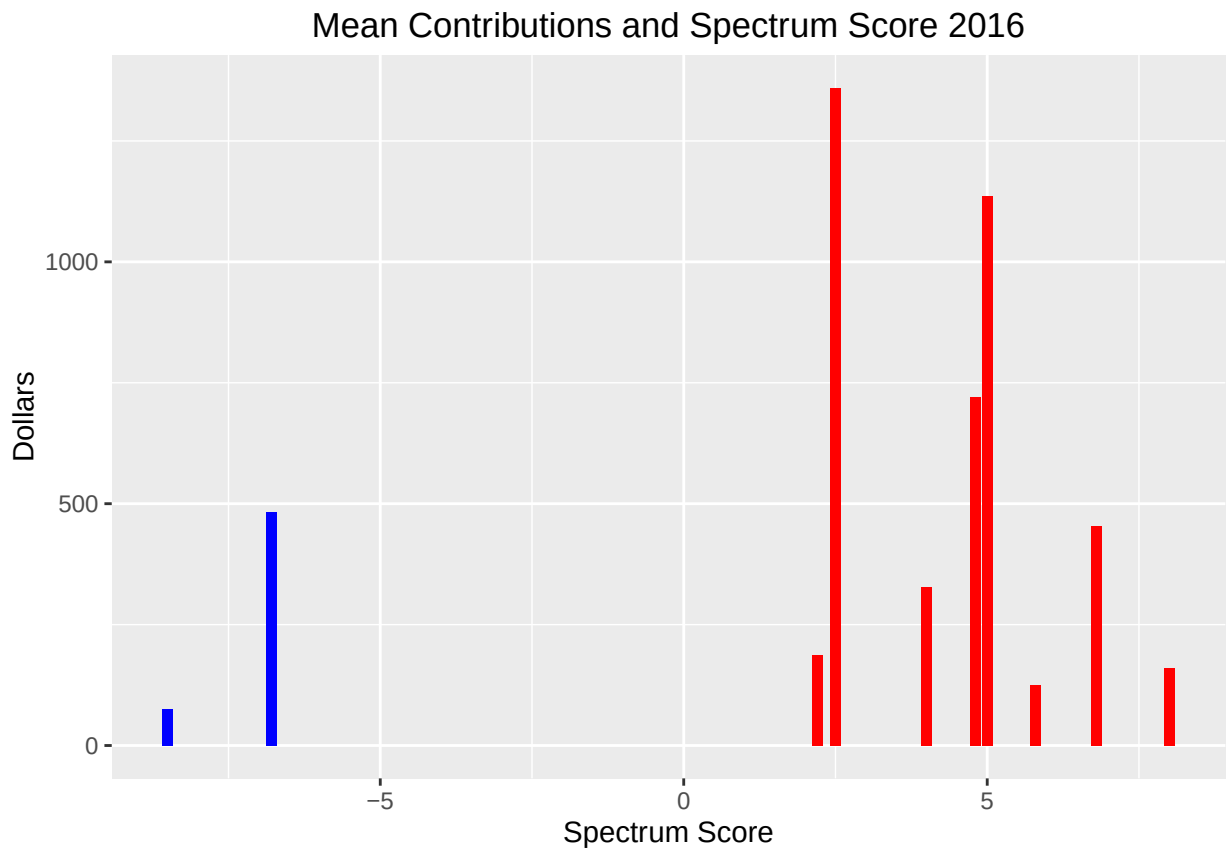
3B) Analyze total contributions per candidate:

	candidate	millions
10	Clinton, Hillary Rodham	50.2837224
4	Cruz, Rafael Edward 'Ted'	15.9249934
1	Bush, Jeb	14.5015898
8	Rubio, Marco	12.4079950
11	Sanders, Bernard	12.1700804
2	Carson, Benjamin S.	11.0681784
6	Kasich, John R.	3.4609306
3	Christie, Christopher J.	3.3025935
5	Fiorina, Carly	2.9850523
7	Paul, Rand	2.7226717
9	Trump, Donald J.	0.8447776



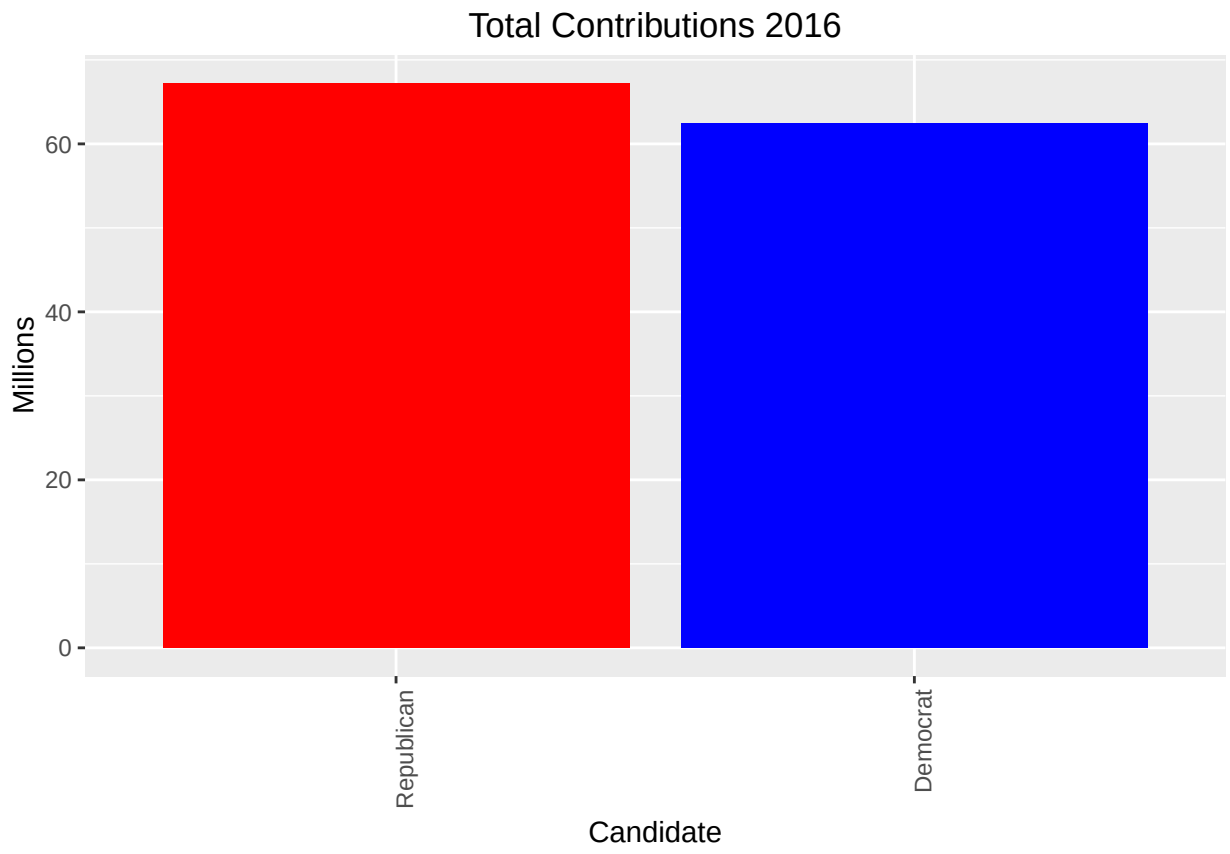
3C) Analyze contribution MEAN by spectrum score:

	spectrum_score	dollars
2	2.5	1359.65151
5	5.0	1135.47590
4	4.8	719.70375
10	-6.8	482.64806
7	6.8	453.70758
3	4.0	328.19642
1	2.2	187.18953
8	8.0	161.28045
6	5.8	126.00816
9	-8.5	74.80948



3D) Contribution Totals by party:

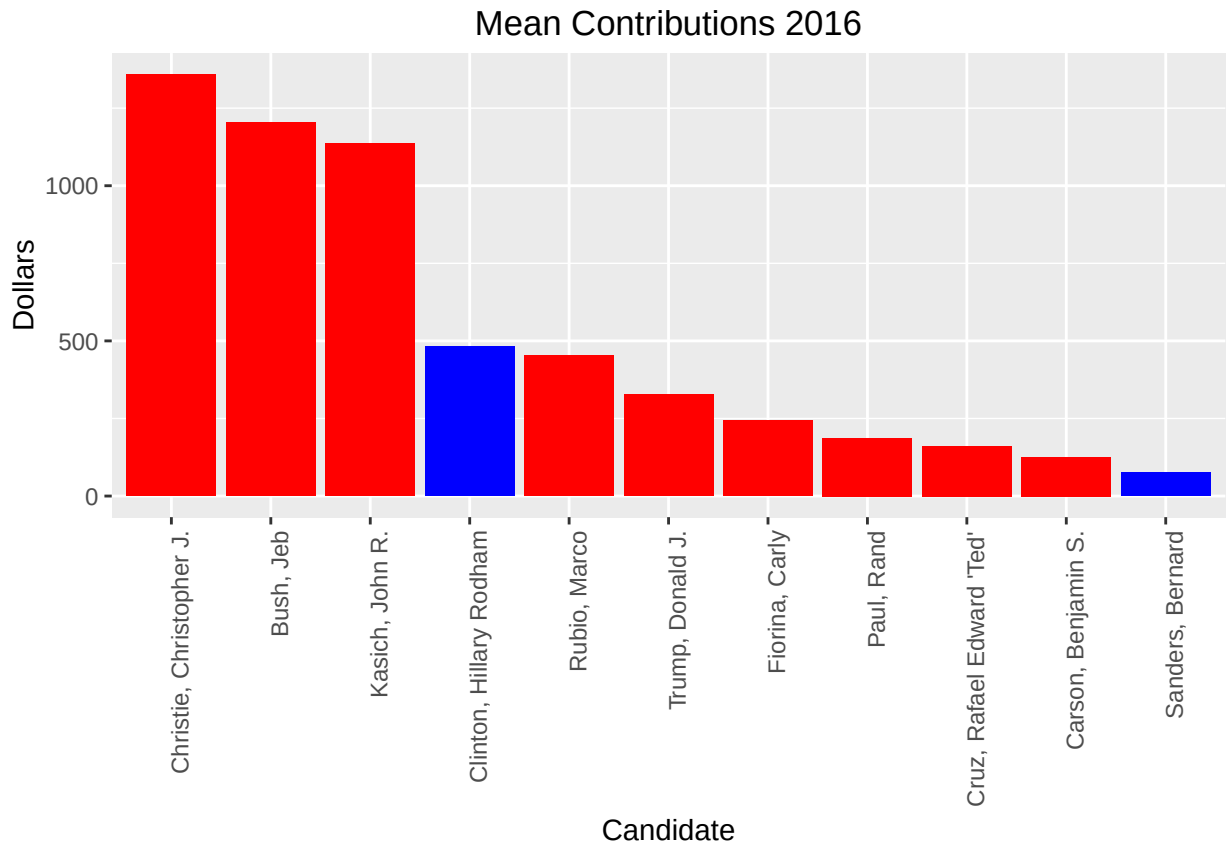
party	millions
Republican	67.21878
Democrat	62.45380



3E) Average Joe (mean contributions by candidate):

Let's take a look at the MEAN contributions per candidate:

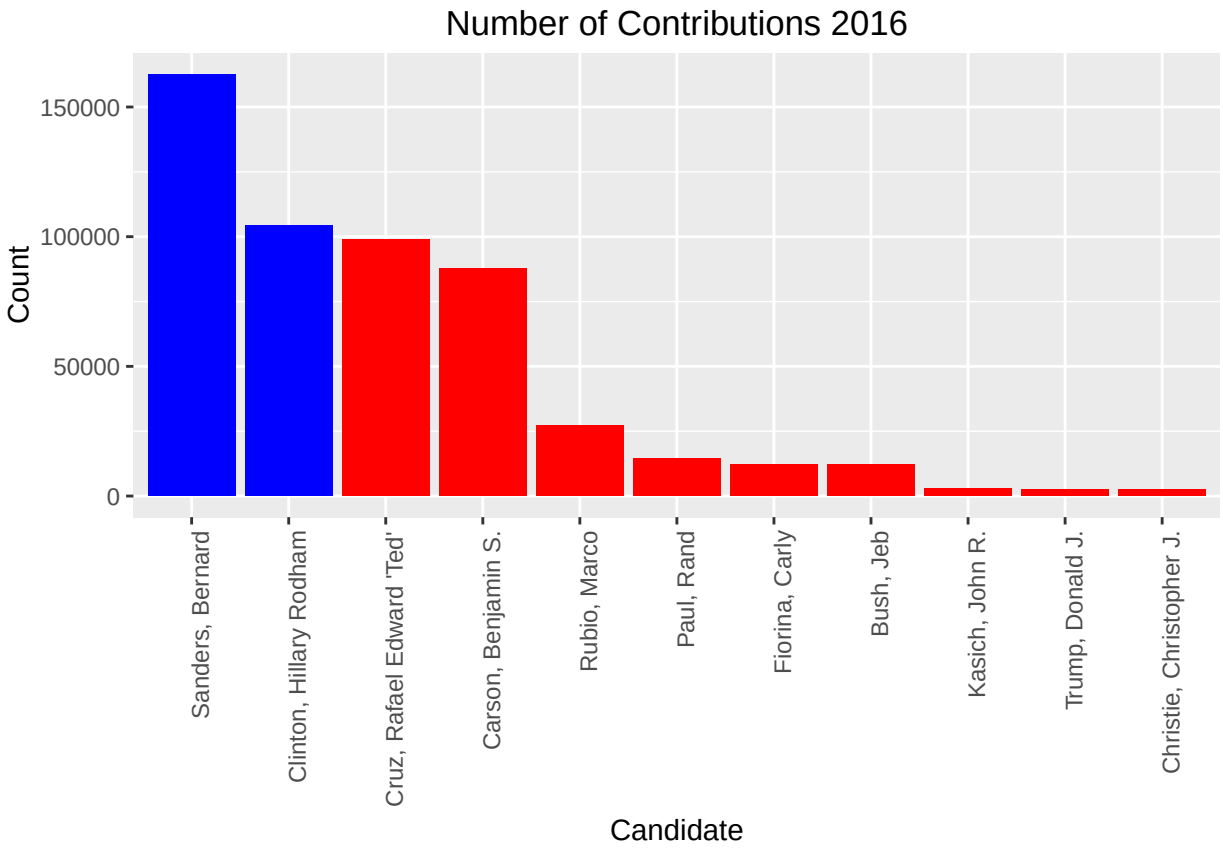
	candidate	dollars
3	Christie, Christopher J.	1359.65151
1	Bush, Jeb	1202.45355
6	Kasich, John R.	1135.47590
10	Clinton, Hillary Rodham	482.64806
8	Rubio, Marco	453.70758
9	Trump, Donald J.	328.19642
5	Fiorina, Carly	243.93661
7	Paul, Rand	187.18953
4	Cruz, Rafael Edward 'Ted'	161.28045
2	Carson, Benjamin S.	126.00816
11	Sanders, Bernard	74.80948



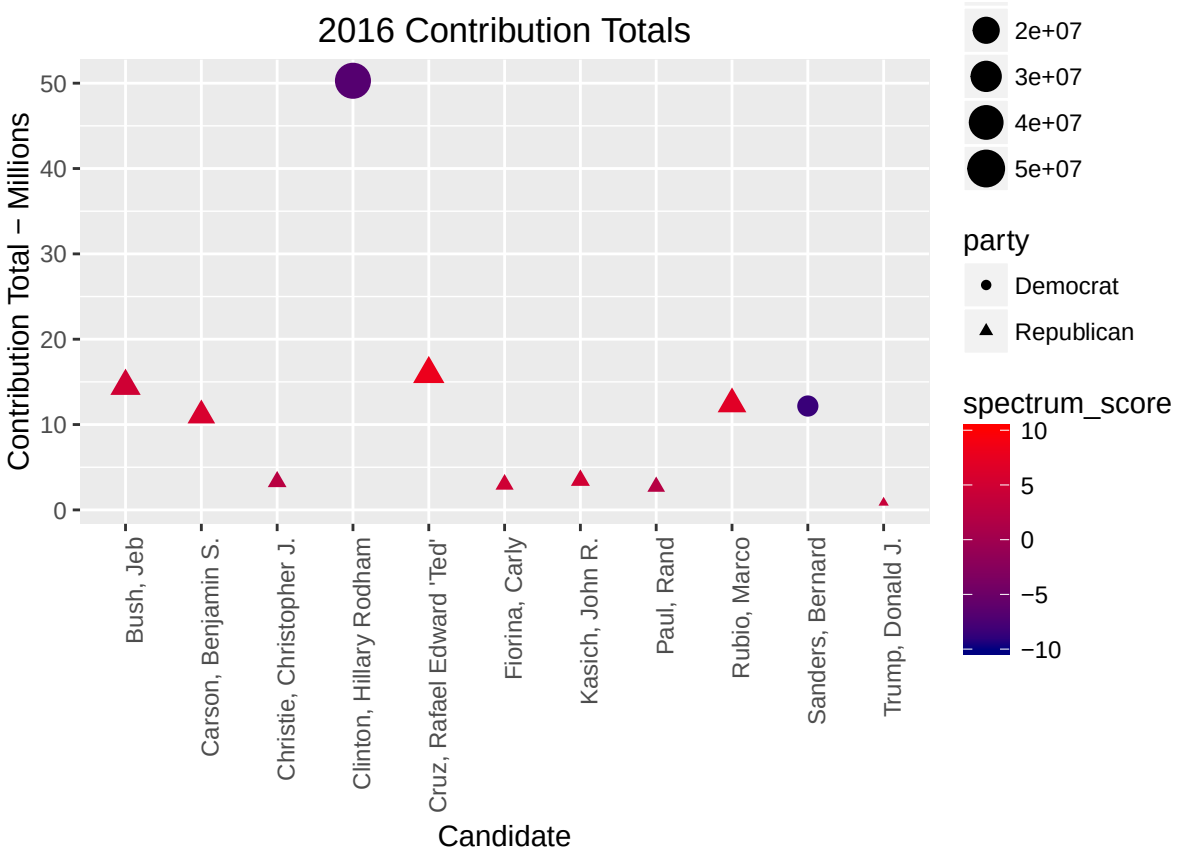
3F) Contribution Counts:

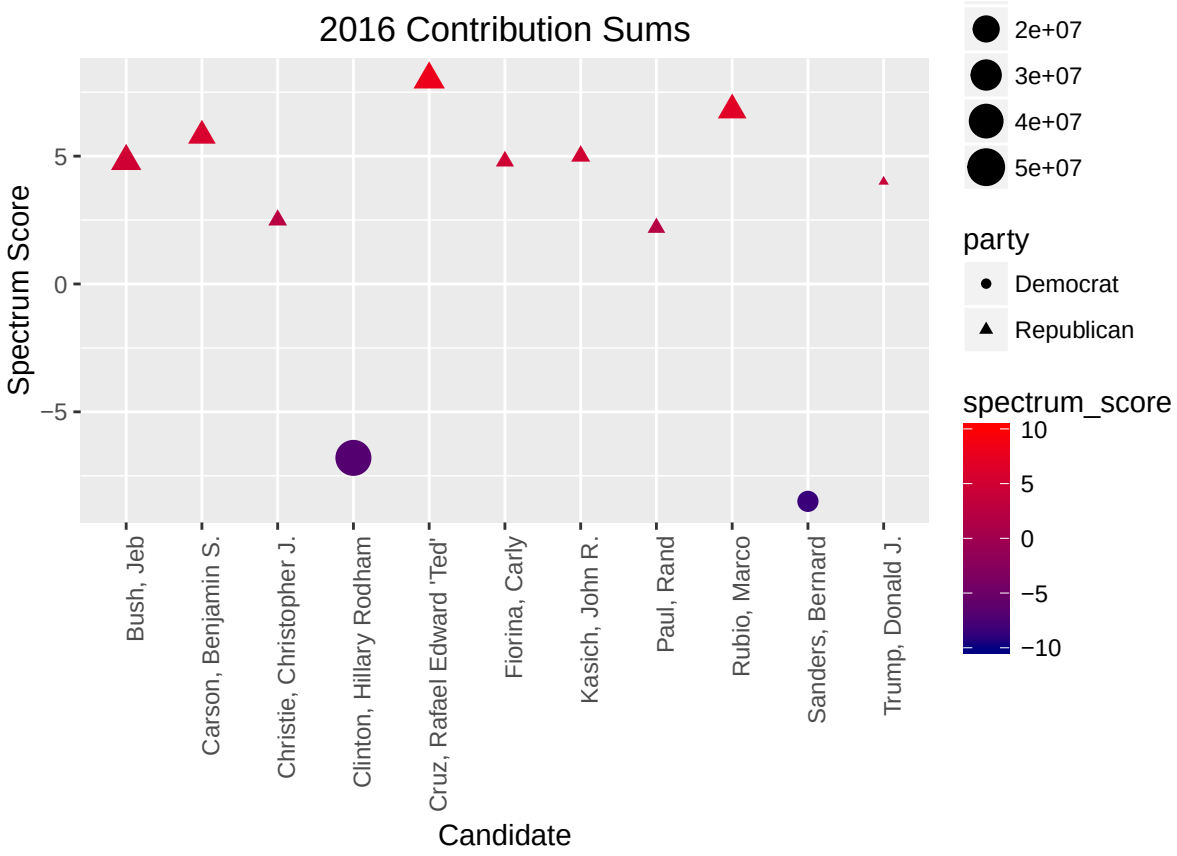
How many UNIQUE contributions per candidate:

	candidate	contr_count
11	Sanders, Bernard	162681
10	Clinton, Hillary Rodham	104183
4	Cruz, Rafael Edward 'Ted'	98741
2	Carson, Benjamin S.	87837
8	Rubio, Marco	27348
7	Paul, Rand	14545
5	Fiorina, Carly	12237
1	Bush, Jeb	12060
6	Kasich, John R.	3048
9	Trump, Donald J.	2574
3	Christie, Christopher J.	2429


Number of Contributions 2016
Count
150000
100000
50000
0
Sanders, Bernard
Clinton, Hillary Rodham
Cruz, Rafael Edward 'Ted'
Carson, Benjamin S.
Rubio, Marco
Paul, Rand
Fiorina, Carly
Bush, Jeb
Kasich, John R.
Trump, Donald J.
Christie, Christopher J.
Candidate

3G) Another vantage point of the data from above in a single plot:





3H) Check the T-Distributions:

```
##
## Call:
## lm(formula = spectrum_score ~ contribution_count + contribution_total +
##     contribution_mean + party, data = summary_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9317 -0.6077  0.2556  0.6071  1.4567
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.393e+01  2.555e+00  -5.451 0.001586 **
## contribution_count  2.457e-05  1.459e-05   1.685 0.143024
## contribution_total  9.772e-08  4.302e-08   2.271 0.063556 .
## contribution_mean  -1.856e-04  1.148e-03  -0.162 0.876883
## partyRepublican     1.747e+01  1.926e+00   9.070 0.000101 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.342 on 6 degrees of freedom
## Multiple R-squared:  0.9622, Adjusted R-squared:  0.9371
## F-statistic: 38.22 on 4 and 6 DF,  p-value: 0.0002093
```

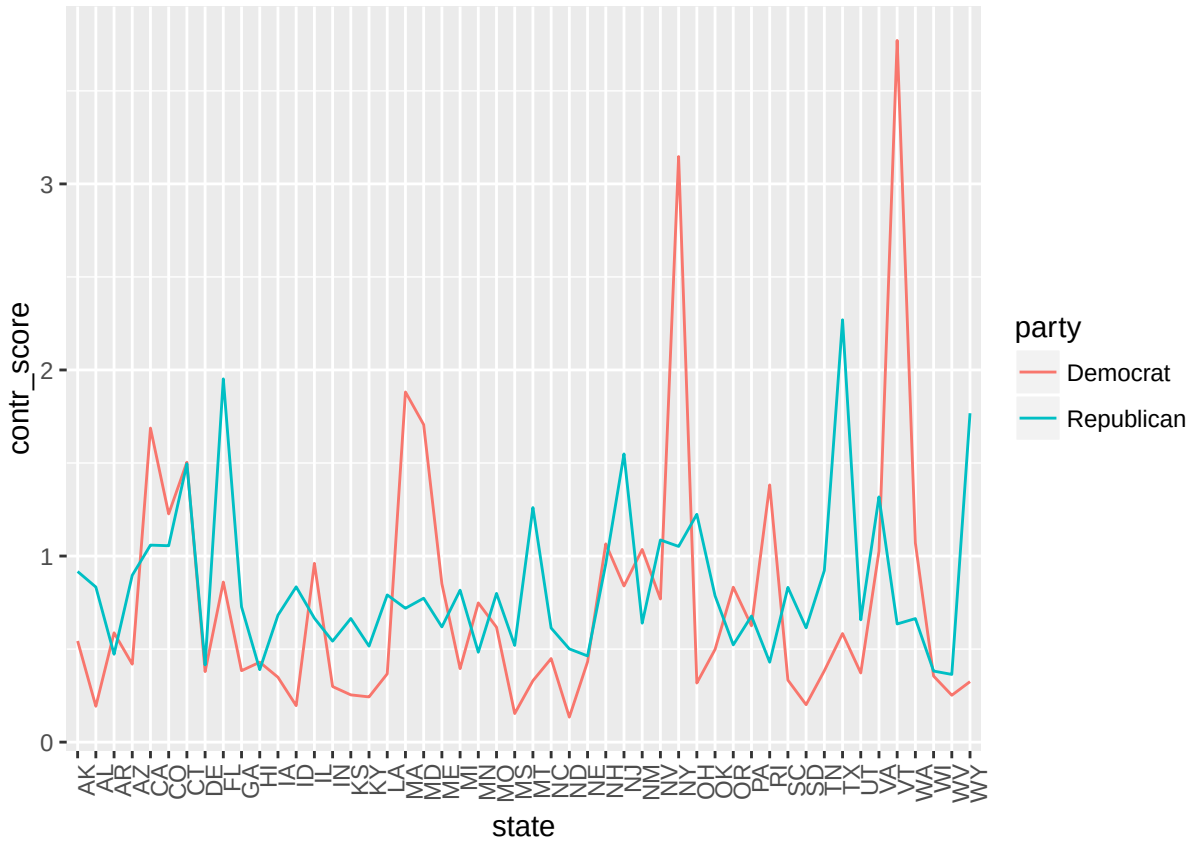
The **partyRepublican** asterisk does not tell us much, as we know that the spectrum score is correlated to the political party.

However, the * in contribution_total shows as unlikely to be by chance.

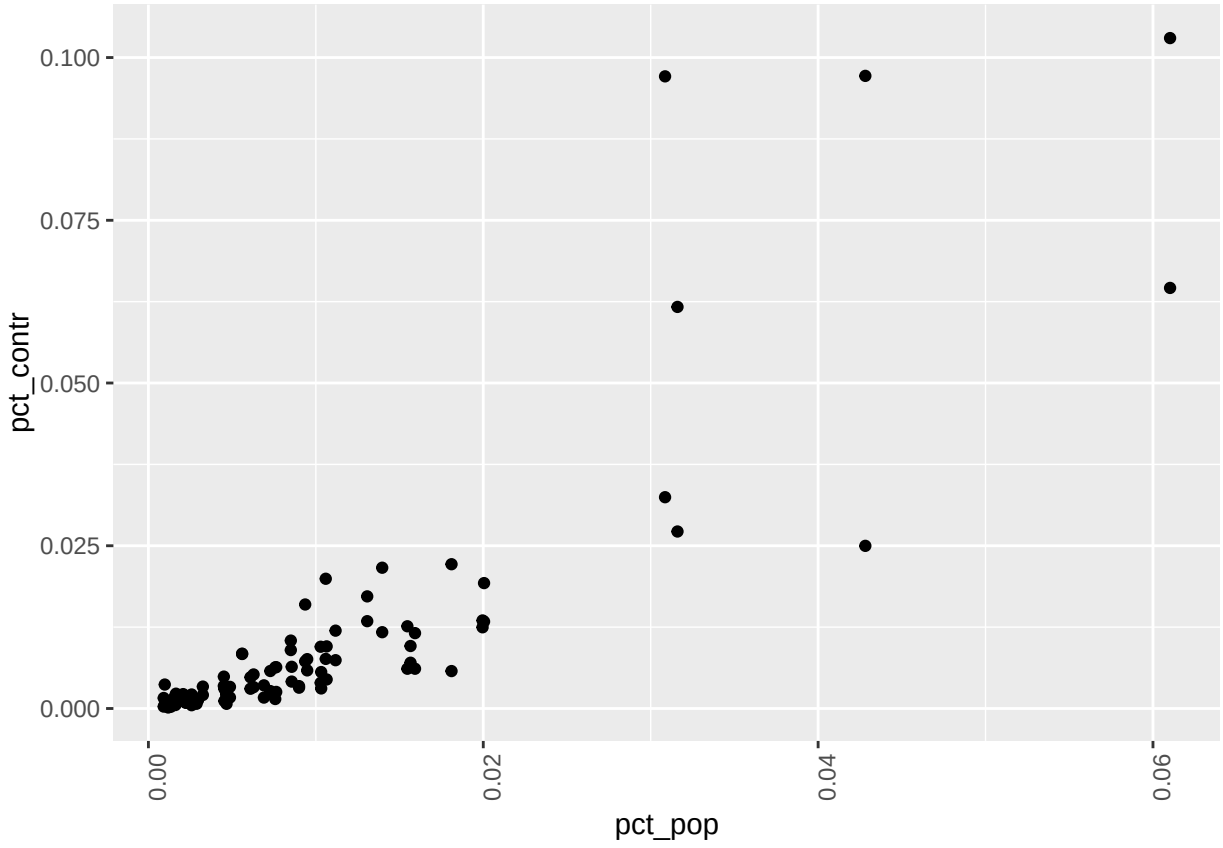
3I) Population % vs. Donation %:

http://www.census.gov/popest/data/national/totals/2015/files/NST-EST2015-alldata.csv

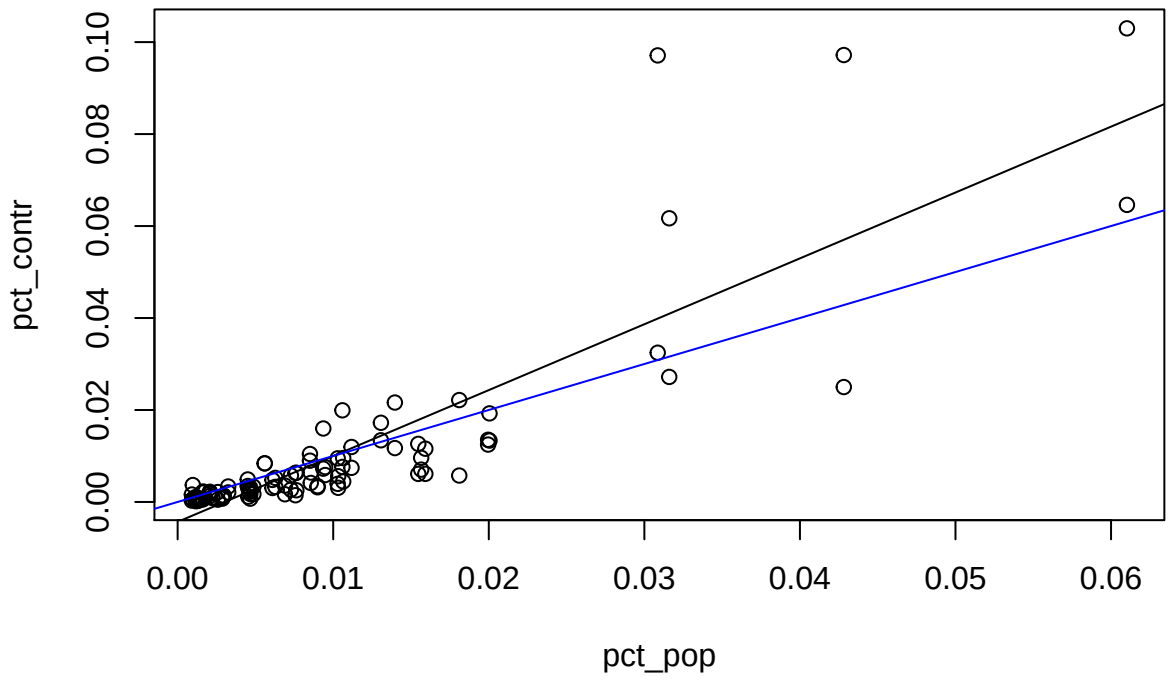
state	POPULATION	party	contrb_total	pct_pop	pct_contr	contr_score
AL	4858979	Democrat	184206.59	0.0075745	0.0014585	0.1925557
AL	4858979	Republican	797399.46	0.0075745	0.0063136	0.8335414
AK	738432	Democrat	78998.78	0.0011511	0.0006255	0.5433831
AK	738432	Republican	133299.84	0.0011511	0.0010554	0.9168861
AZ	6828065	Democrat	563318.98	0.0106440	0.0044602	0.4190376
AZ	6828065	Republican	1205195.26	0.0106440	0.0095425	0.8965119
AR	2978204	Democrat	344434.18	0.0046426	0.0027272	0.5874191
AR	2978204	Republican	276989.25	0.0046426	0.0021931	0.4723944
CA	39144818	Democrat	13007544.31	0.0610214	0.1029911	1.6877857
CA	39144818	Republican	8159785.45	0.0610214	0.0646075	1.0587678




```
##
## Call:
## lm(formula = contr_score ~ us_pop + party, data = state_donations_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6210 -0.3717 -0.1692  0.1429  3.0157
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.75530    0.08189   9.224 5.85e-15 ***
## us_pop                NA         NA      NA       NA
## partyRepublican  0.08129    0.11580   0.702    0.484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.579 on 98 degrees of freedom
## Multiple R-squared:  0.005004,   Adjusted R-squared:  -0.00515
## F-statistic: 0.4928 on 1 and 98 DF,  p-value: 0.4843
```




```
##
## Call:
## lm(formula = pct_contr ~ pct_pop, data = state_donations_df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.032024 -0.003900  0.000393  0.002800  0.057231
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.004324   0.001319  -3.279  0.00144 **
## pct_pop      1.432375   0.088027  16.272  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00982 on 98 degrees of freedom
## Multiple R-squared:  0.7299, Adjusted R-squared:  0.7271
## F-statistic: 264.8 on 1 and 98 DF,  p-value: < 2.2e-16
```



Note that in the above chart, the mean abline is blue, while the actual is black. It seems that the states that represent larger percentages of the overall population actually donate a higher **Percentage** than their state counterparts.

4. Inference:

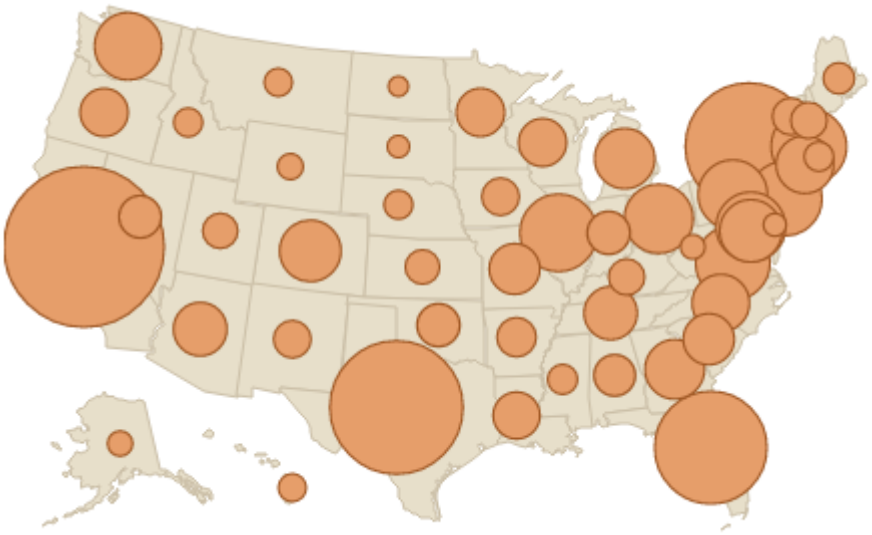
It seems that the difference in the numbers of candidates running in each party makes it difficult to do any nominee-to-nominee comparisons. With that said, let's look to some of the other possible models and see if they have any more strong correlations:

```
##
## Call:
## lm(formula = contribution_count ~ spectrum_score + party, data = summary_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39198 -20131 -15833  25062  48063
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       222969      52583   4.240  0.00284 **
## spectrum_score     11704       6146   1.904  0.09333 .
## partyRepublican  -251080      81276  -3.089  0.01491 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 33300 on 8 degrees of freedom
## Multiple R-squared:  0.7115, Adjusted R-squared:  0.6393
## F-statistic: 9.863 on 2 and 8 DF,  p-value: 0.006932
```

```
##
## Call:
## lm(formula = contribution_total ~ spectrum_score + party, data = summary_df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -16180480  -3937072  -1565424   4097702  16180480
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      57113967   14432552   3.957  0.00419 **
## spectrum_score    3383930    1686861   2.006  0.07976 .
## partyRepublican -66151272   22307920  -2.965  0.01800 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9140000 on 8 degrees of freedom
## Multiple R-squared:  0.6534, Adjusted R-squared:  0.5667
## F-statistic:  7.54 on 2 and 8 DF,  p-value: 0.01444
```

```
##
## Call:
## lm(formula = contribution_mean ~ spectrum_score + party, data = summary_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -587.1 -326.6 -186.8  416.6  619.2
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -283.40     780.86  -0.363    0.726
## spectrum_score    -73.48      91.27  -0.805    0.444
## partyRepublican  1219.37    1206.95   1.010    0.342
##
## Residual standard error: 494.5 on 8 degrees of freedom
## Multiple R-squared:  0.1347, Adjusted R-squared:  -0.08158
## F-statistic: 0.6228 on 2 and 8 DF,  p-value: 0.5605
```



5. Conclusion

My Findings:

- There were other interesting data sets available, such as WHO the donations came from, when in the campaign lifecycle they came, etc.
- I normally pictured these data sets with very "distinct" columns, ie - with truly distinct meanings. This data set showed me that you can create fields with different meanings, all based on the same data (ie - dollars per contribution, total contributions, total number of contributions - all the same data, but different underlying meanings.)
- Based on this project, some questions that I would be curious about in the future are:
- Which types of groups are donating the large amounts?
- Which companies are truly just 'bribing' their candidates by contributing to all the candidates.
- System memory limits really were an issue. I knew that this was a problem in industry, but was not really expecting to get so many out-of-memory errors running Rmd scripts.