

Exploratory Data Analysis:

Presidential Campaign Contribution in Ohio

Haleh Dolati

Introduction

I chose the "Financial Contributions to Presidential Campaigns" data for the State of Ohio. I chose Ohio because I went to The Ohio State University and am fairly familiar with the state's political dynamics. Ohio is considered a swing state, which makes it very interesting to study. This data set has the data for 167,259 donations.

Summary of Data Set

Now that I know what kind of data I have in this dataset, I would like to get some general information, like data types.

```
## 'data.frame':   167259 obs. of  18 variables:
## $ cmte_id       : Factor w/ 24 levels "C00458844","C00500587",..
: 15 15 15 15 15 6 12 7 6 7 ...
## $ cand_id       : Factor w/ 24 levels "P00003392","P20002671",..
: 23 23 23 23 23 1 15 12 1 12 ...
## $ cand_nm       : Factor w/ 24 levels "Bush, Jeb","Carson, Benja
min S.",...: 22 22 22 22 22 4 14 19 4 19 ...
## $ contbr_nm     : Factor w/ 44968 levels " JUDGE, ANNE GILDAY",.
.: 35963 35967 35975 33864 33867 17322 8617 16139 36641 25134 ...
## $ contbr_city   : Factor w/ 1360 levels " BATAVIA","45320",...: 2
44 708 248 231 193 274 589 274 596 231 ...
## $ contbr_st     : Factor w/ 1 level "OH": 1 1 1 1 1 1 1 1 1 1 ..
.
## $ contbr_zip    : int  45315 44060 44106 45208 44721 432141210 4
41071232 432022420 450365038 45249 ...
## $ contbr_employer : Factor w/ 13494 levels "", "-","?BECKMAN WEIL S
HEPARDSON, LLC",...: 5791 9973 5791 1283 10564 5791 3169 8464 9973 10547
...
## $ contbr_occupation: Factor w/ 6691 levels "", " CERTIFIED REGISTERE
D NURSE ANESTHETIS",...: 2889 5134 2889 3503 5475 2889 3209 3884 6054 40
2 ...
## $ contb_receipt_amt: num  97.1 53.5 69.4 88.4 -80 ...
## $ contb_receipt_dt : Factor w/ 685 levels "1-Apr-15","1-Apr-16",...:
357 21 199 152 110 69 545 609 452 589 ...
## $ receipt_desc   : Factor w/ 25 levels "", " SEE REATTRIBUTION",..
: 1 1 1 1 1 1 1 1 1 1 ...
## $ memo_cd        : Factor w/ 2 levels "", "X": 2 2 2 2 2 2 1 1 2 1
...
## $ memo_text      : Factor w/ 89 levels "", " SEE REATTRIBUTION",..
```

```

: 1 1 1 1 1 13 1 4 13 4 ...
## $ form_tp      : Factor w/ 3 levels "SA17A","SA18",...: 2 2 2 2
2 2 1 1 2 1 ...
## $ file_num     : int  1146165 1146165 1146165 1146165 1146165 1
091718 1144564 1077404 1091718 1077404 ...
## $ tran_id      : Factor w/ 166816 levels "A0000FD2A304E432AAD5"
,...: 108798 118153 112319 108928 120998 54281 166784 144656 54504 14441
1 ...
## $ election_tp  : Factor w/ 4 levels "", "G2016", "O2016",...: 2 2
2 2 2 4 4 4 4 4 ...

```

Initial Observation from the Summary

This dataset contains individuals' financial contributions to 2016 presidential candidates. It shows whom, how much, when, and where contributed to every single candidate during both primary and general elections.

However, there are some variables in the contribution data that do not give us any insight like the candidates ID number. So I decided to keep only 9 variables:

"cmte_id", "cand_id", "contbr_st", "receipt_desc", "memo_cd", "memo_text", "form_tp", "file_num", and "tran_id"

```

## 'data.frame': 167259 obs. of 9 variables:
## $ cand_nm      : Factor w/ 24 levels "Bush, Jeb","Carson, Benja
min S.",...: 22 22 22 22 22 4 14 19 4 19 ...
## $ contbr_nm     : Factor w/ 44968 levels " JUDGE, ANNE GILDAY",.
.: 35963 35967 35975 33864 33867 17322 8617 16139 36641 25134 ...
## $ contbr_city   : Factor w/ 1360 levels " BATAVIA","45320",...: 2
44 708 248 231 193 274 589 274 596 231 ...
## $ contbr_zip    : int  45315 44060 44106 45208 44721 432141210 4
41071232 432022420 450365038 45249 ...
## $ contbr_employer : Factor w/ 13494 levels "", "-", "?BECKMAN WEIL S
HEPARDSON, LLC",...: 5791 9973 5791 1283 10564 5791 3169 8464 9973 10547
...
## $ contbr_occupation: Factor w/ 6691 levels "", " CERTIFIED REGISTERE
D NURSE ANESTHETIS",...: 2889 5134 2889 3503 5475 2889 3209 3884 6054 40
2 ...
## $ contb_receipt_amt: num  97.1 53.5 69.4 88.4 -80 ...
## $ contb_receipt_dt : Factor w/ 685 levels "1-Apr-15","1-Apr-16",...:
357 21 199 152 110 69 545 609 452 589 ...
## $ election_tp    : Factor w/ 4 levels "", "G2016", "O2016",...: 2 2
2 2 2 4 4 4 4 4 ...

```

I also noticed the zip codes probably need some cleaning. So I'm going to first make sure only 5 digit zip codes are present in the variable. I will replace the unfit zip codes with NA.

```

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##      10   43450   44140   44320   45150   100000         3

```

Since the Min and max zip codes are out of range, I've decided to convert any zip code that does not start with 43, 44, and 45 since all zip codes in Ohio should start with these numbers. The first zip code is 43001 and the last one is 45999.

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	43000	43450	44140	44280	45150	45900	188

Now that I cleaned up the zip code variable, the first questions that came to my mind were: who received more money overall? Which party received more money? How contributions changed general and primary elections? However, I've noticed the party affiliation is missing from this dataset. Therefore, if I want to answer couple of these questions I need to add that to the dataset. First I need a list of candidates:

```
## Candidates:

## [1] "Bush, Jeb" "Carson, Benjamin S."
## [3] "Christie, Christopher J." "Clinton, Hillary Rodham"
## [5] "Cruz, Rafael Edward 'Ted'" "Fiorina, Carly"
## [7] "Graham, Lindsey O." "Huckabee, Mike"
## [9] "Jindal, Bobby" "Johnson, Gary"
## [11] "Kasich, John R." "Lessig, Lawrence"
## [13] "McMullin, Evan" "O'Malley, Martin Joseph"
## [15] "Pataki, George E." "Paul, Rand"
## [17] "Perry, James R. (Rick)" "Rubio, Marco"
## [19] "Sanders, Bernard" "Santorum, Richard J."
## [21] "Stein, Jill" "Trump, Donald J."
## [23] "Walker, Scott" "Webb, James Henry Jr."
```

Based on my initial thought on visualization and analysis, I need to add two other variables. The first one is a new variable for each candidate's party affiliation.

The second one is a categorical one (based on the donation amount) for the level of contribution so I can see if there are any differences between number of small and big donors regarding the candidate that they chose. I chose the following categories:

[0-\$50) [\$50-\$100) [\$100, \$500) [\$500, \$2700) [\$2700,)

My next step is to apply this grouping to all three datasets: election, primary and general.

This data set has information on when each donation was made. In order to make it ready for plotting, I have to convert the contb_receipt_dt (date of donation) into a date type variable so I can track contributions over time. The original format is 11-Jul-16 and I'm going to convert them to 2016-07-11.

And finally I've noticed that the data has the information on contributions for both primary and general elections. I would like to see how donation has changed from primary to general election so the next step for me is to make two separate data frames for them: primary_election and general_election.

Univariate Plots Section

My first question is: How much Ohioan contributed to the 2016 presidential election? But before I answer this question I took a look at the contribution receipt amount data and found out there are some negative contributions (probably refunds). I manually checked couple of those negative records and found out there are positive records from the same contributor for the exact same amount but with a positive sign. Both records for these cancelled transactions need to be excluded. Here is the new summary of election data:

```
## 'data.frame': 164012 obs. of 13 variables:
## $ cand_nm : Factor w/ 24 levels "Bush, Jeb", "Carson, Benjamin S.", ...: 4 4 4 4 4 4 4 4 4 4 ...
## $ contbr_nm : Factor w/ 44968 levels " JUDGE, ANNE GILDAY", .
```

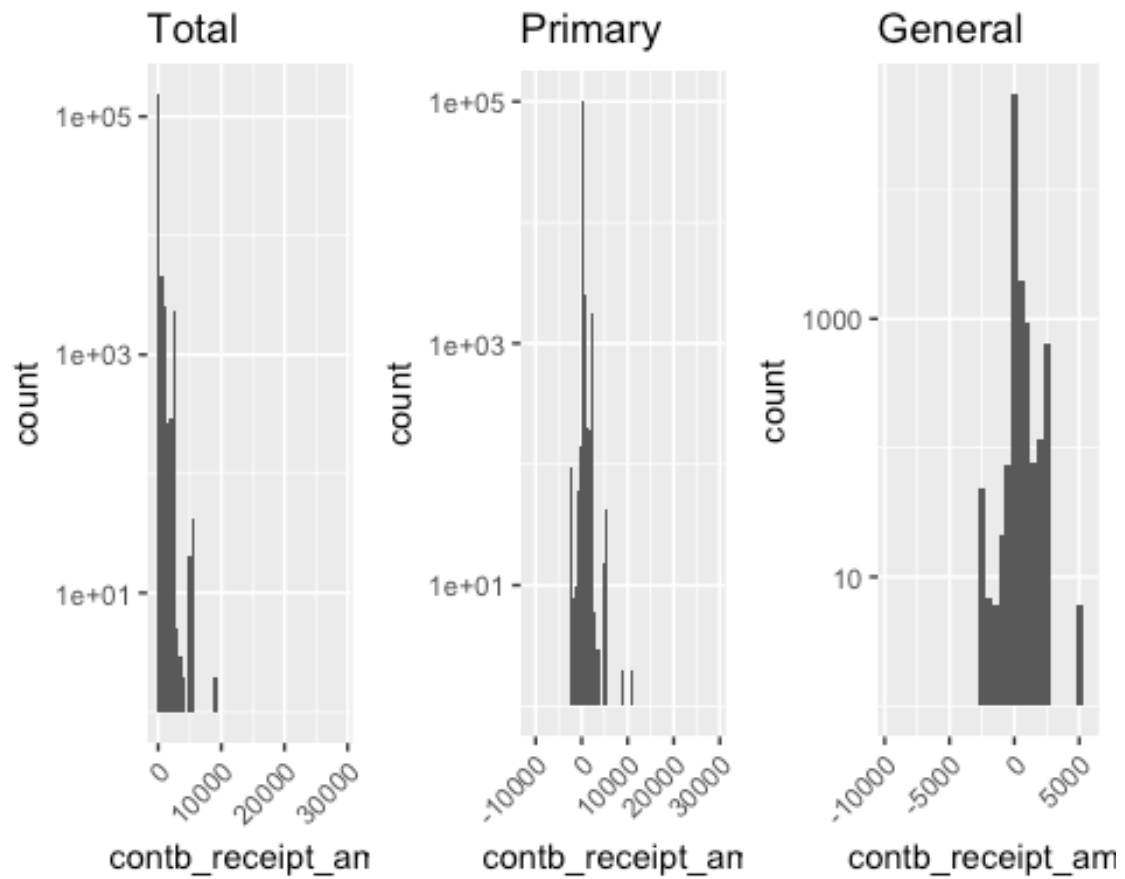
```

.: 738 32418 20895 23631 44908 42431 39245 12849 41980 31054 ...
## $ contbr_city      : Factor w/ 1360 levels " BATAVIA","45320",...: 7
37 231 813 759 231 47 941 231 813 231 ...
## $ contbr_zip       : chr  "44654" "45202" "43054" "44022" ...
## $ contbr_employer  : Factor w/ 13494 levels "", "-", "?BECKMAN WEIL S
HEPARDSON, LLC",...: 5791 10564 3307 9973 10564 8767 1284 9582 8087 8087
...
## $ contbr_occupation: Factor w/ 6691 levels "", " CERTIFIED REGISTERE
D NURSE ANESTHETIS",...: 2889 1152 44 4322 1254 1525 6054 5141 2709 5134
...
## $ contb_receipt_amt: num  0 0.08 0.1 0.15 0.15 0.16 0.17 0.27 0.28
0.55 ...
## $ contb_receipt_dt : Date, format: "2015-08-24" "2016-03-09" ...
## $ election_tp      : Factor w/ 4 levels "", "G2016", "O2016",...: 4 2
2 2 2 2 2 2 2 2 ...
## $ party            : chr  "Democrat" "Democrat" "Democrat" "Democra
t" ...
## $ donation_level   : Factor w/ 5 levels "[0-$50)","[$50-$100)",...:
1 1 1 1 1 1 1 1 1 1 ...
## $ abs_amount       : num  0 0.08 0.1 0.15 0.15 0.16 0.17 0.27 0.28
0.55 ...
## $ sum              : num  1 0.03 0.18 0.25 0.3 0.31 0.33 0.44 0.55
0.83 ...

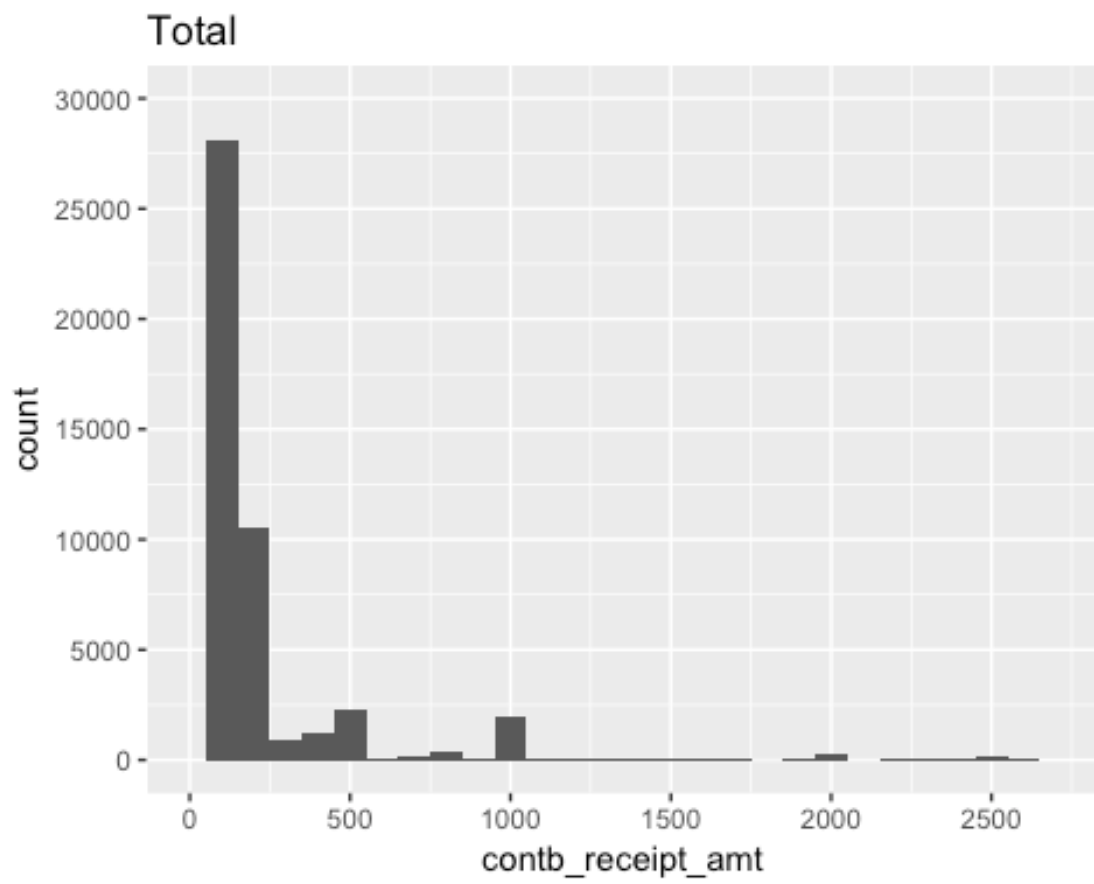
## Total dollars donated during 2016 presidential election: 20128635
## Total dollars donated 2016 presidential primary election: 13366731
## Total dollars donated 2016 presidential general election: 6470701
## Total count of donations: 164012
## Total number of donors: 44399

```

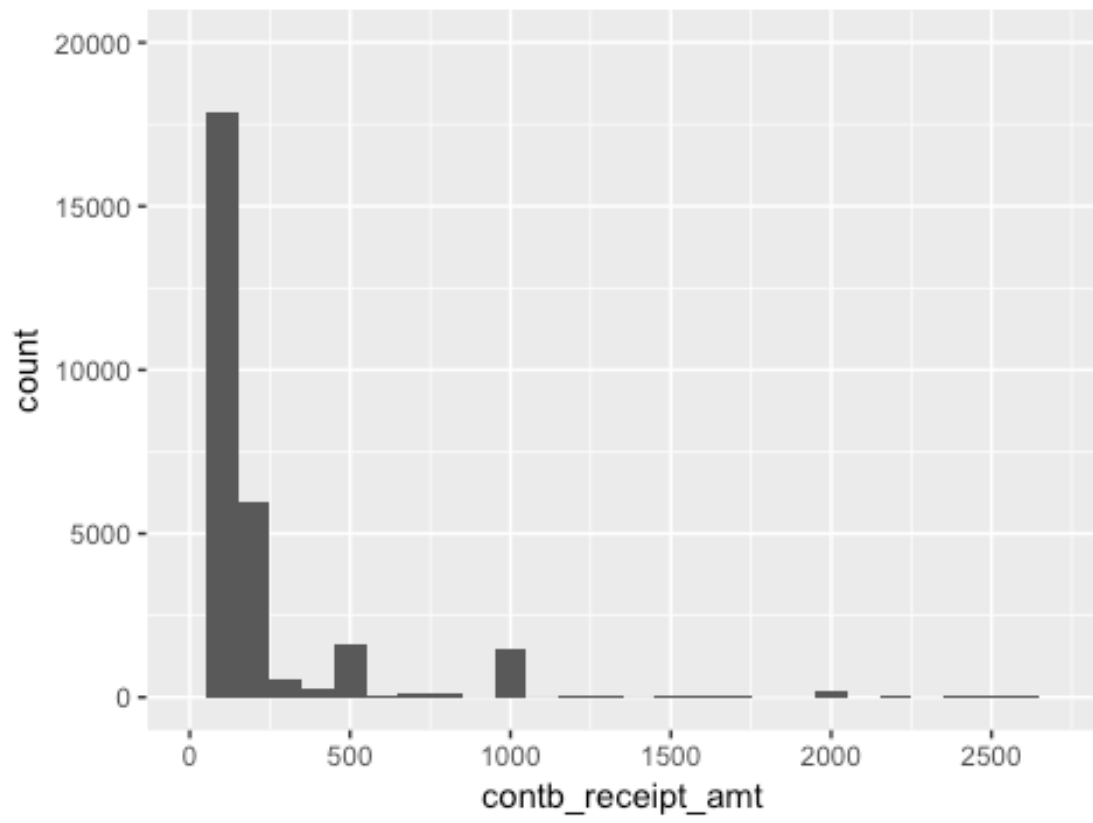
Well, majority of the contributions happened during the primaries. I also noticed the sum of contributions during primary and general election is lower than total contributions. I checked the data and it is because there were other types of denotable election type in the state of Ohio during the presidential race. Let's take a look at the distribution of these donations based on the amount of them.

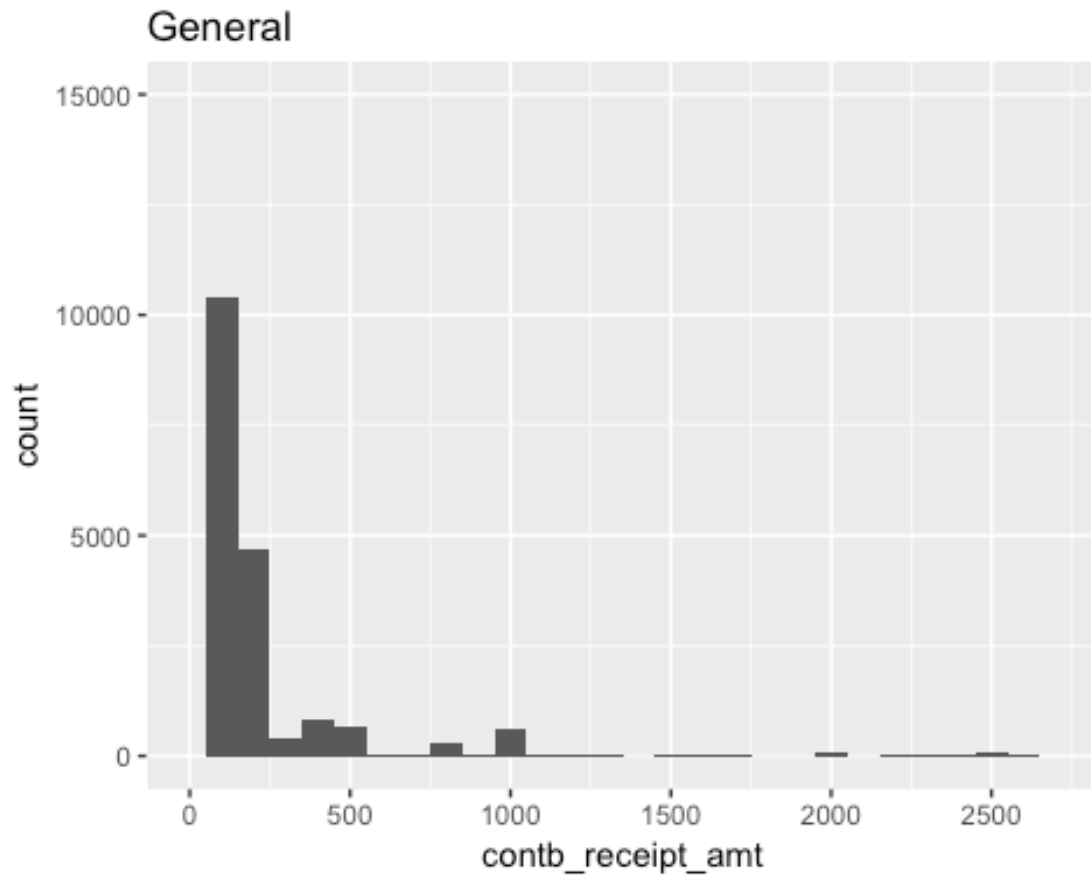


It looks like majority of the donations were small donations. However, maximum legal contribution by individuals is \$2700. I'm going to take a look at those donations that were made by individuals, not the corporations. Also I'm not interested in refunds, so the negative ones are going to be excluded as well.

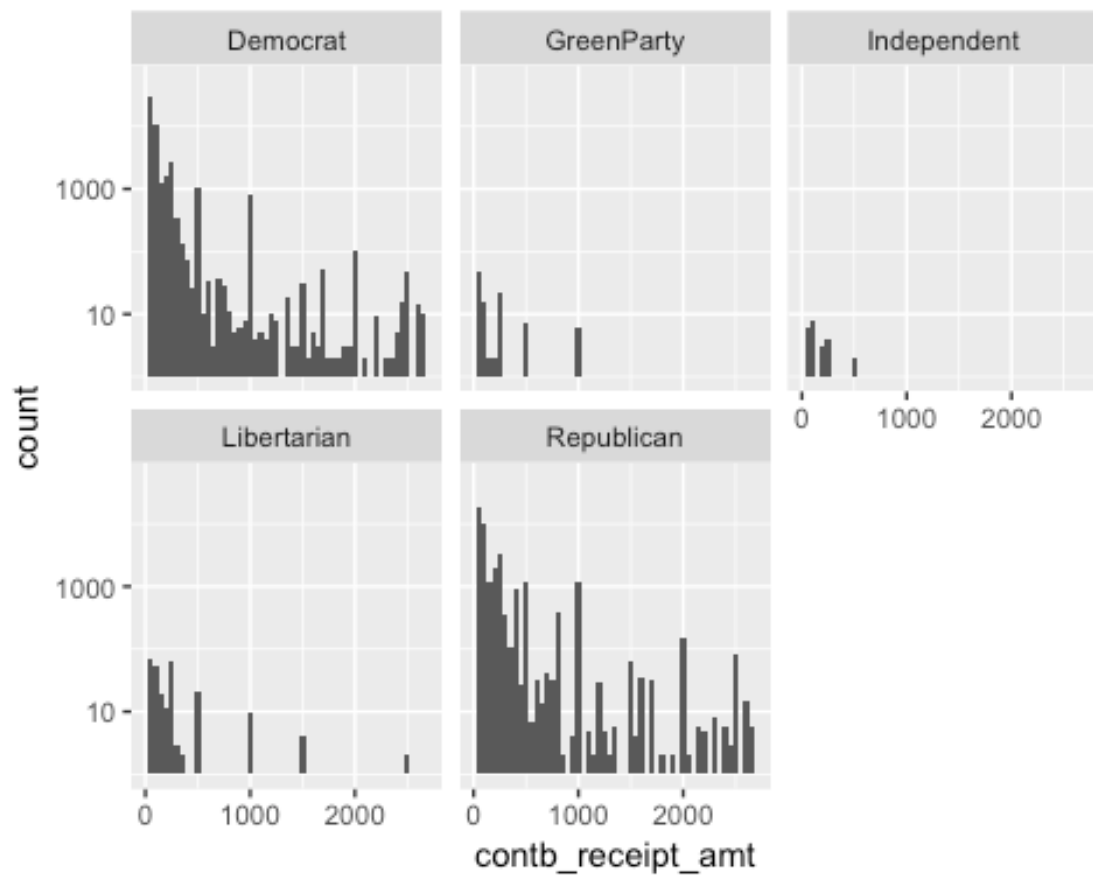


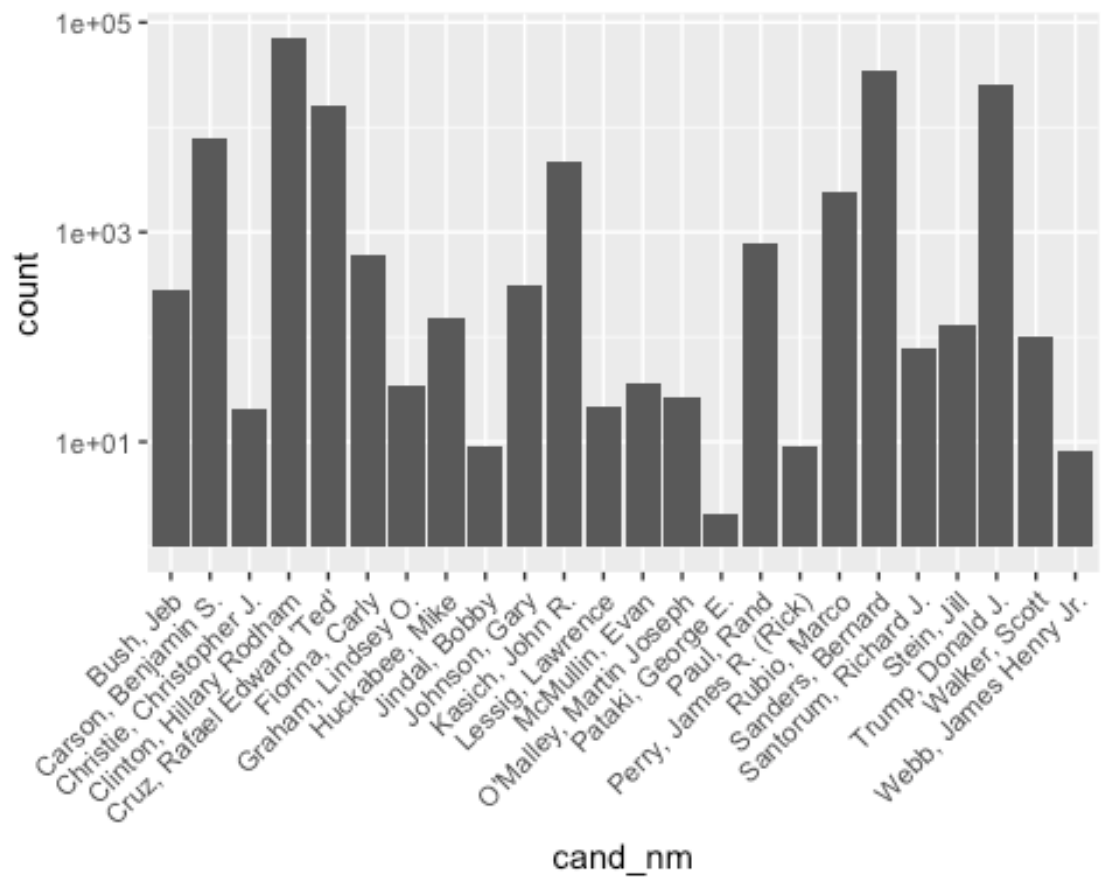
Primary

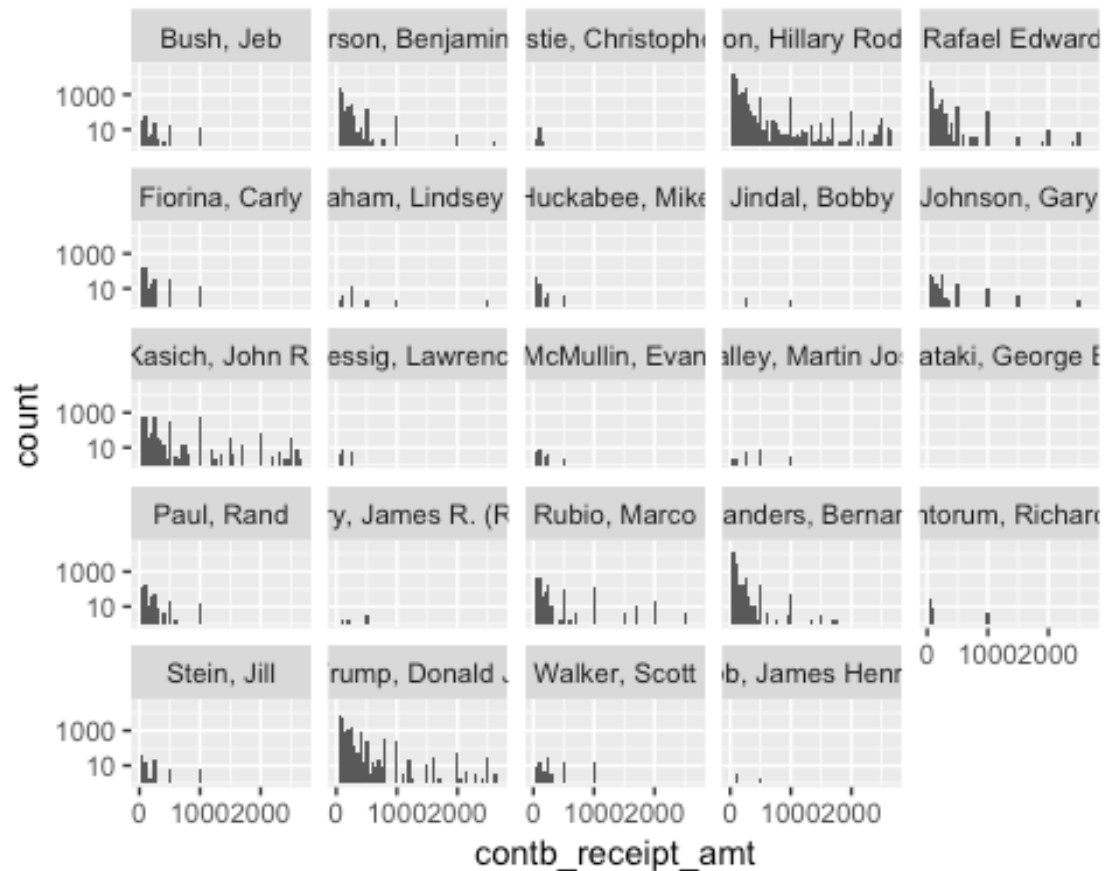




So overall, the majority of donations were under \$100. What about the donations that each party and each candidate received from individual Ohioan (under \$2700)? (Range of number of donations for different parties are significantly different. Therefore, I'm going to use logarithm ($\log_{10}$) to scale the numbers.)



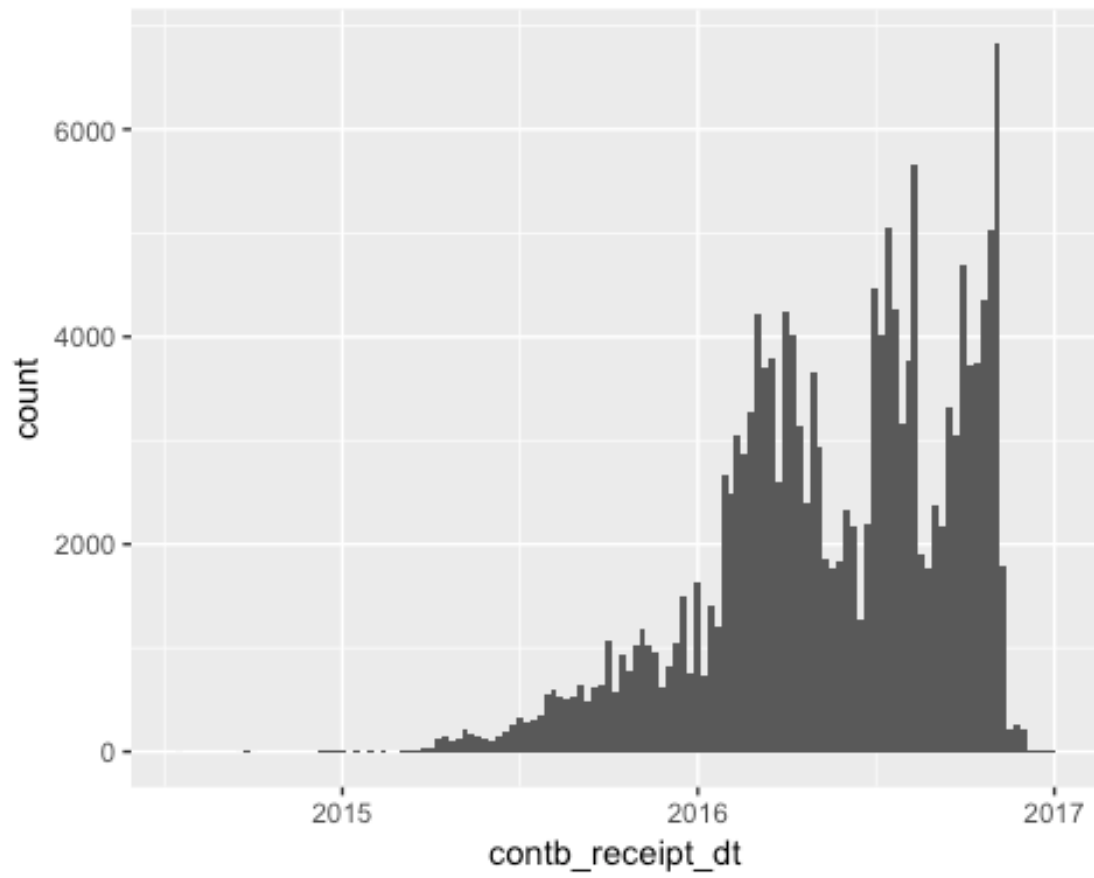




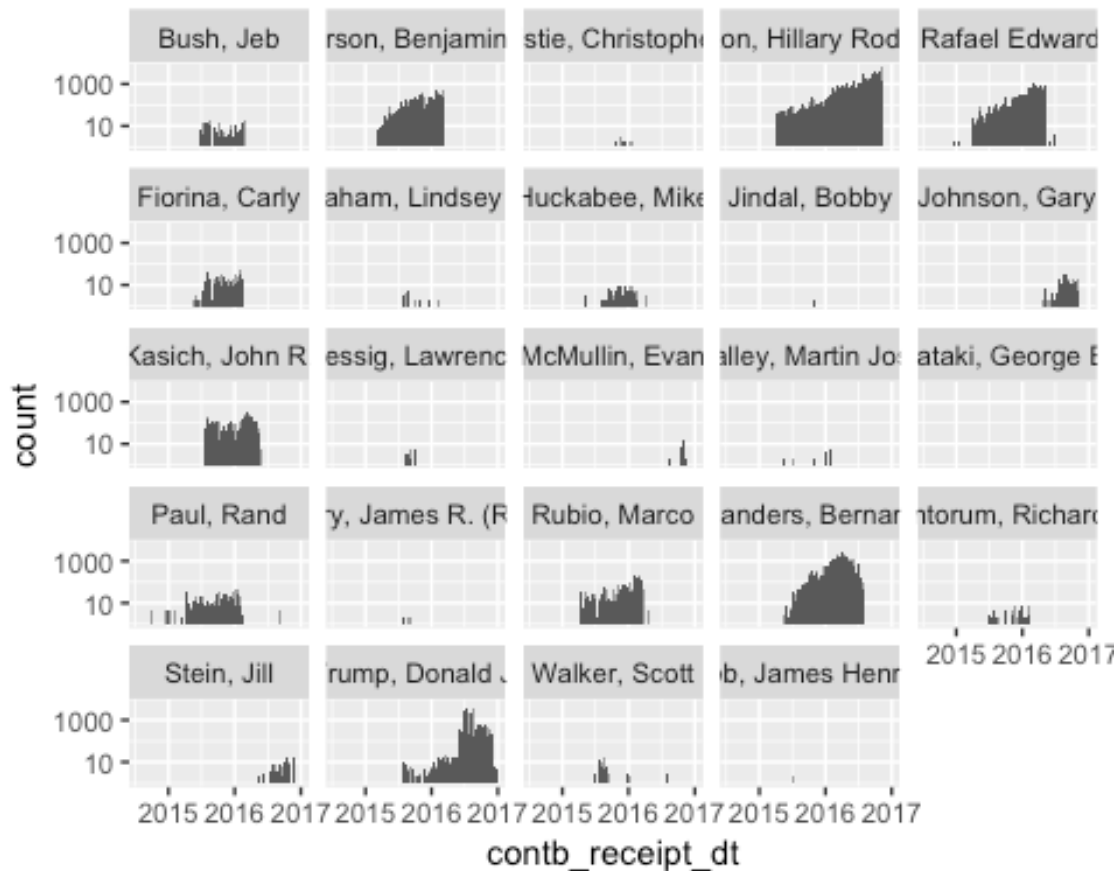
All I can say from these of plots is:

- 1- Overall, Democrats and republicans received the majority of the money
- 2- Clinton, Sanders, Trump, Cruz, Carson, and Kasich are the top 5 candidates in terms of donations to their campaigns.

The last variable that I would like to take a look at is the date. I would like to see how the number of contributions changed during the election year:



After primaries more donations were made. The next chart I would like to see is how number of donations were changed for each candidate?



Naturally trump and Clinton received more donations for the general election (after primaries) but Sanders also received many donations before the general election.

Univariate Analysis

What is the structure of your dataset?

Federal Election Commission collects data on every single donation and refund to all candidates for all elections. The data that I'm using provides detailed information on 167259 donations that were made during 2016 presidential election (primary and general) to 24 candidates. There are 18 variables for each contribution. To understand the data better, I also downloaded the "Data Description" that the Federal Election Commission's website provided. The followings are the description of each column:

Abbreviation	Description	Type
CMTE_ID	COMMITTEE ID	S
CAND_ID	CANDIDATE ID	S
CAND_NM	CANDIDATE NAME	S
CONTBR_NM	CONTRIBUTOR NAME	S
CONTBR_CITY	CONTRIBUTOR CITY	S
CONTBR_ST	CONTRIBUTOR STATE	S

CONTBR_ZIP	CONTRIBUTOR ZIP CODE	S
CONTBR_EMPLOYER	CONTRIBUTOR EMPLOYER	S
CONTBR_OCCUPATION	CONTRIBUTOR OCCUPATION	S
CONTR_RECEIPT_AMT	CONTRIBUTION RECEIPT AMOUNT	N
CONTR_RECEIPT_DT	CONTRIBUTION RECEIPT DATE	D
RECEIPT_DESC	RECEIPT DESCRIPTION	S
MEMO_CD	MEMO CODE	S
MEMO_TEXT	MEMO TEXT	S
FORM_TP	FORM TYPE	S
FILE_NUM	FILE NUMBER	N
TRAN_ID	TRANSACTION ID	S

What is/are the main feature(s) of interest in your dataset?

The main features of interest are the number of contributions, recipients of those contributions, the amount of the contributions, recipients of those contributions, and the contributors.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

The occupation of the contributors, the cities and the zip codes that the contributions are coming from.

Did you create any new variables from existing variables in the dataset?

I created two new variables: party and donation_level. Party is the candidates' affiliated political party. Donation level is a categorical variable that shows the level of donation in 5 categories: [0-\$50), [\$50-\$100), [\$100, \$500), [\$500, \$2700), [\$2700,)

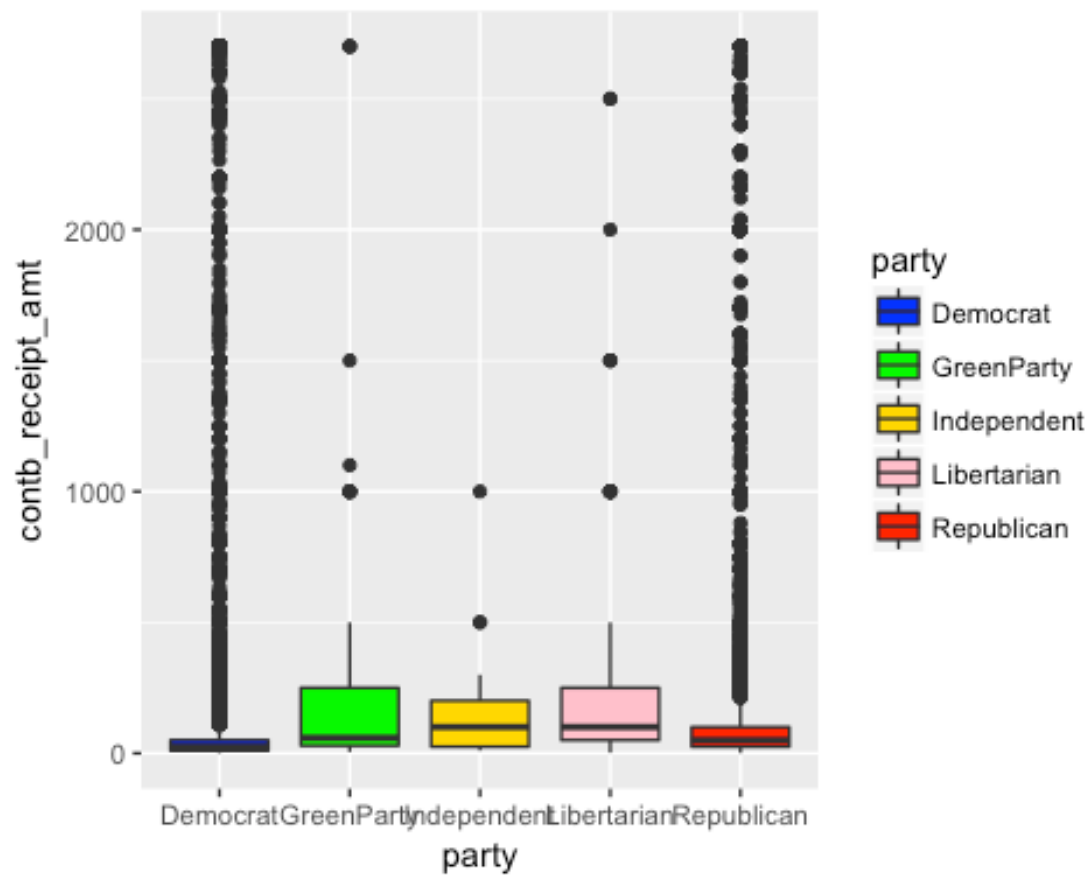
Of the features you investigated, were there any unusual distributions?

Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

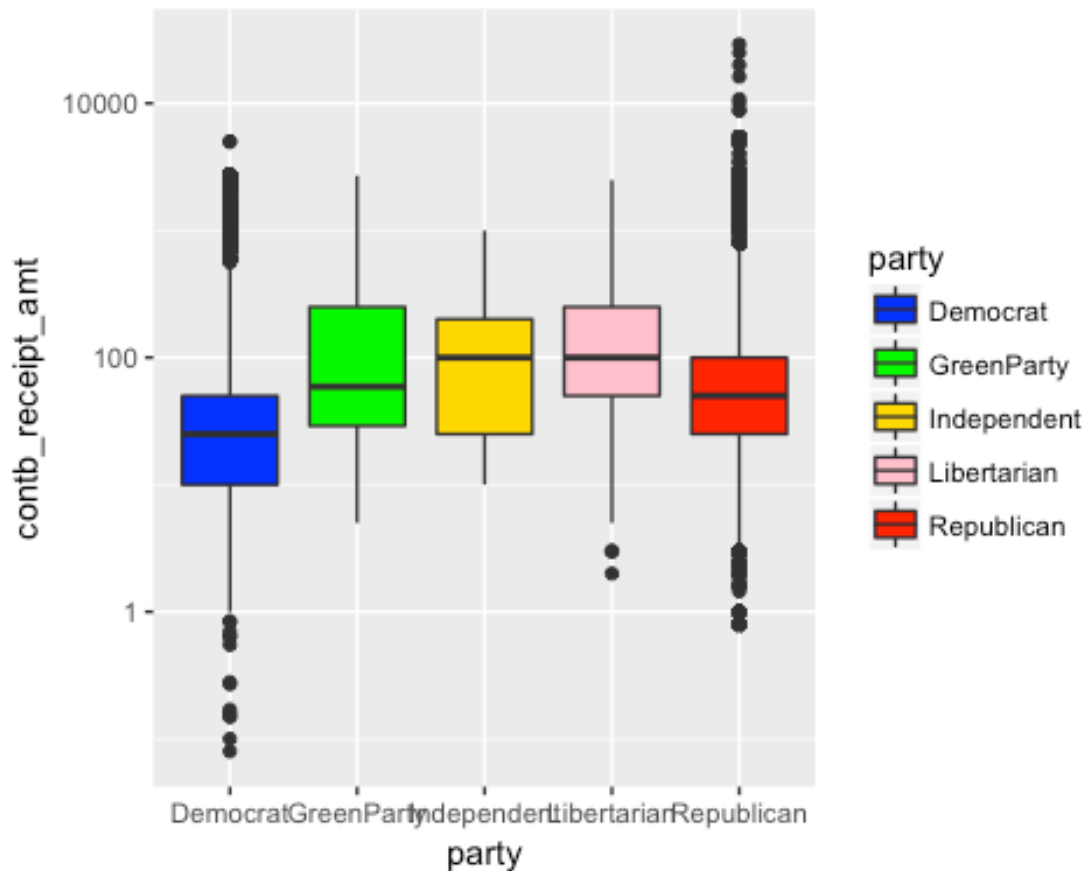
I change the date's data type from 11-Jul-16 to 2016-07-11. This way I can plot the contribution over time. Also I cleaned up the zip code variable to 5 digit format.

Bivariate Plots Section

I would like to see the relationship between contribution amount and party and candidates. So I'm going to take a look at the box plot:



Although I just looked and small donations (under \$2700) still the variety of donation amount are too much. I'm going to take a look at the log of them first then I will use other plot types to explore them further.



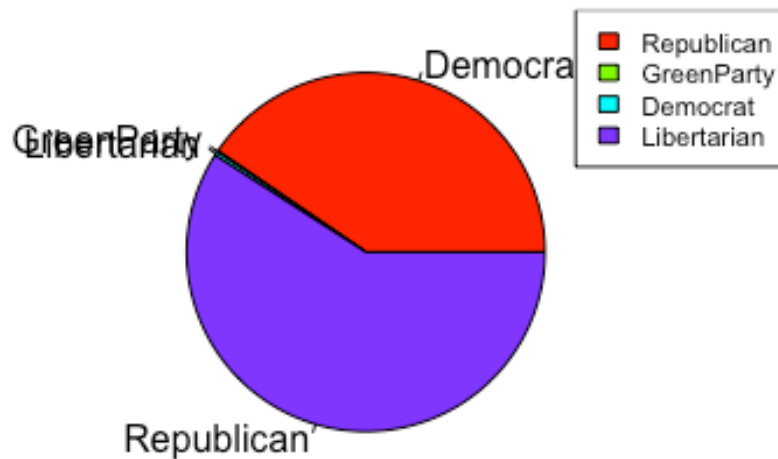
It looks the median contribution amount is lower for democrats compare to others and there are so many outliers for both republicans and democrats.

My next question is How much money each party received overall, during primary and general?

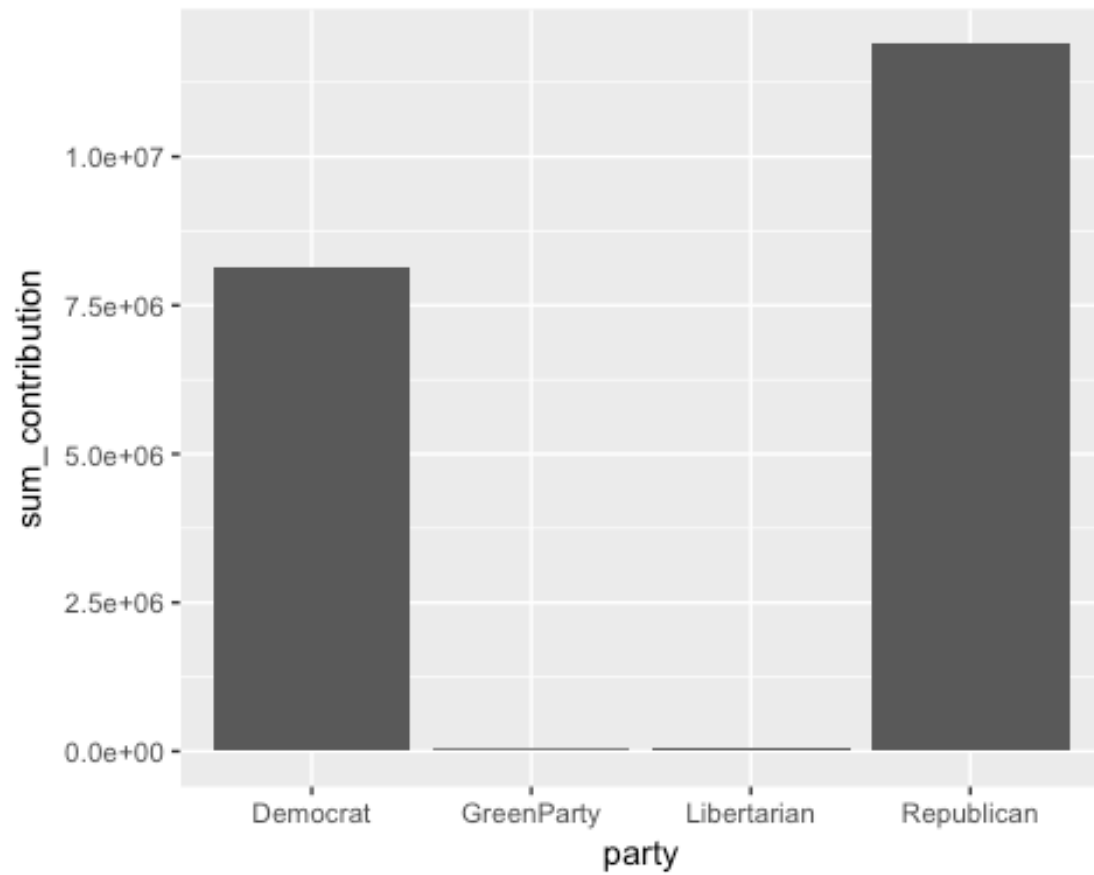
##	party	mean_cand_nm	median_cand_nm	sum_contribution	Percent_c
ontb					
## 1	Democrat	77.74677	25	8139309.53	40.436
4706					
## 2	GreenParty	236.57364	59	30518.00	0.151
6149					
## 3	Libertarian	207.32124	100	65098.87	0.323
4142					
## 4	Republican	202.03768	50	11888502.97	59.062
6391					
##	n	mean_cand_nm_p	median_cand_nm_p	sum_contribution_p	
## 1	104690	74.50947	25	4514081.5	
## 2	129	181.19355	240	5617.0	
## 3	314	470.61667	250	8471.1	
## 4	58843	187.84267	50	8838561.1	
##	Percent_contb_p	n.x	mean_cand_nm_g	median_cand_nm_g	sum_contribution_g
## 1	22.42616799	60584	79.91294	25	3627
328.16					
## 2	0.02790552	31	76.77922	30	5

912.00					
## 3	0.04208482	18	191.70393	100	57
127.77					
## 4	43.91038536	47053	207.58024	80	2777
216.01					
##	Percent_contb_g	n.y			
## 1	18.02073605	45391			
## 2	0.02937109	77			
## 3	0.28381343	298			
## 4	13.79733910	13379			

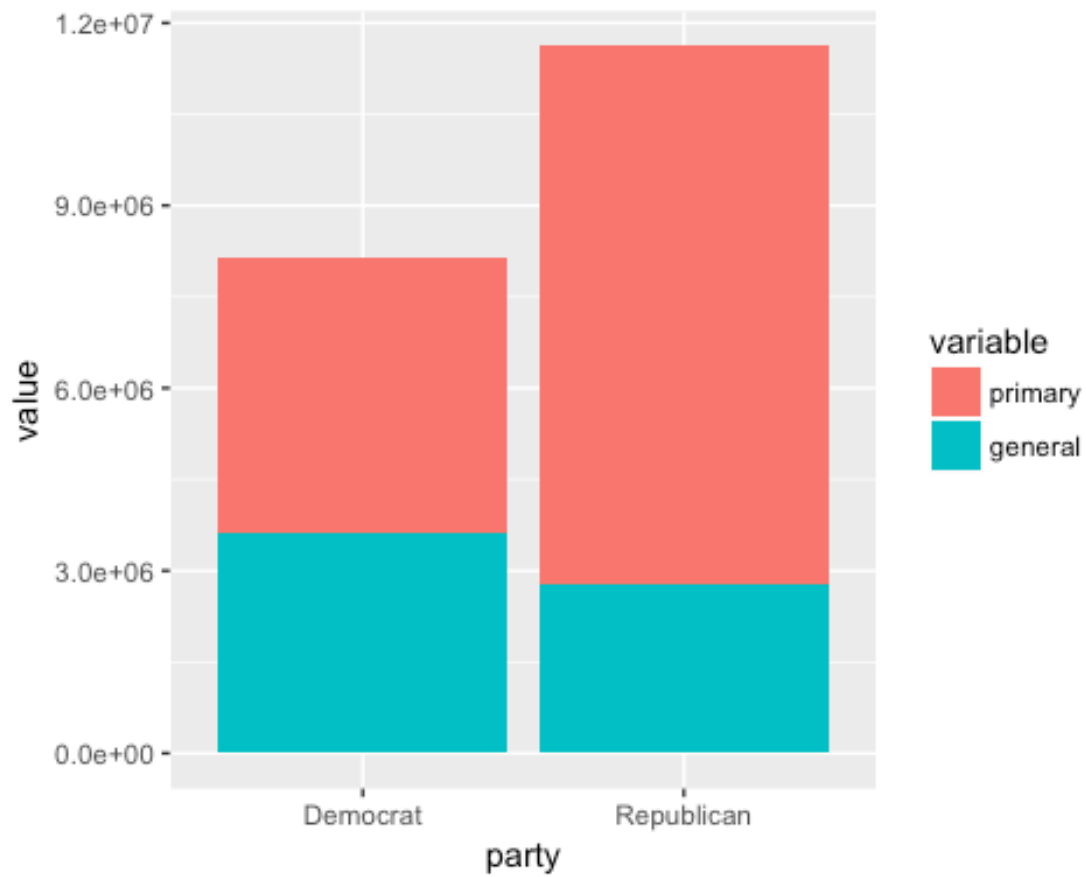
Pie Chart of Campaign Contributions



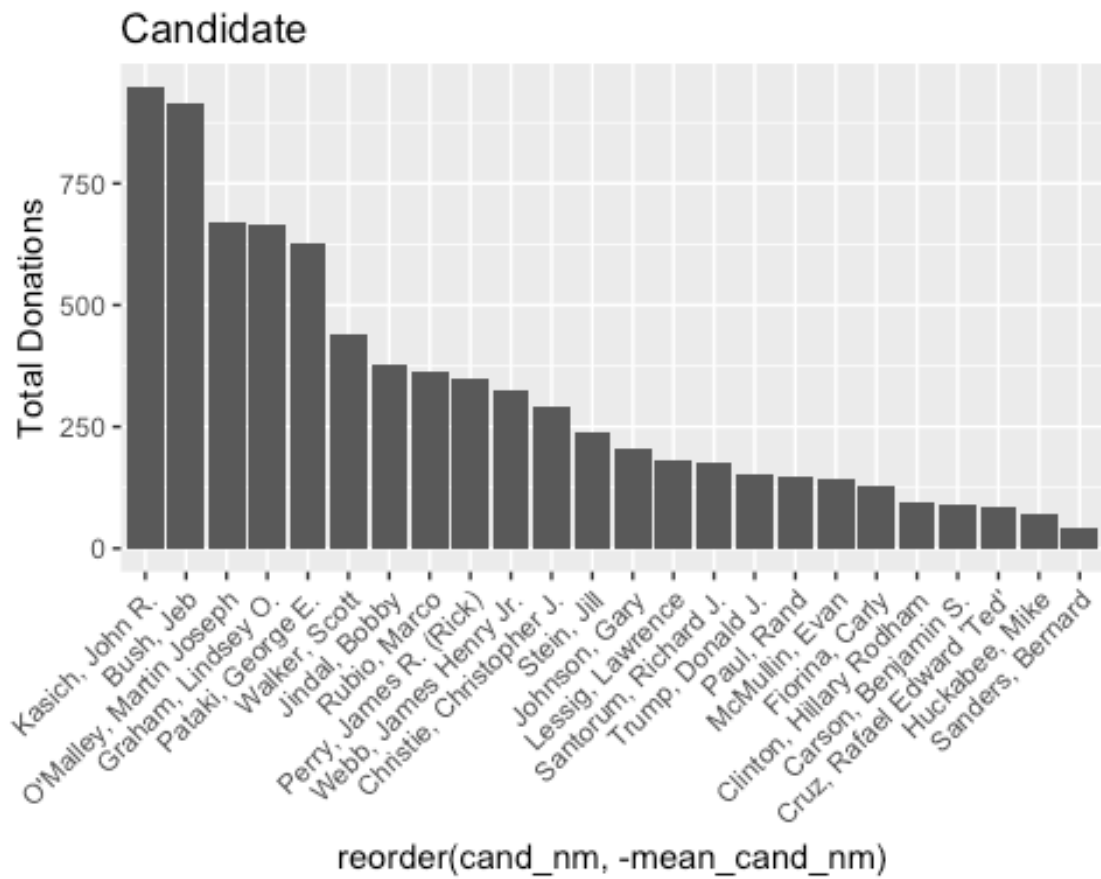
The pie chart clearly shows that the two major political parties received the majority of the monetary contributions. However, since the amount of contribution is extremely different between the political parties, the other 3 values are almost hidden. In order to see them better, I'm going to make a bar chart with log10 of values:

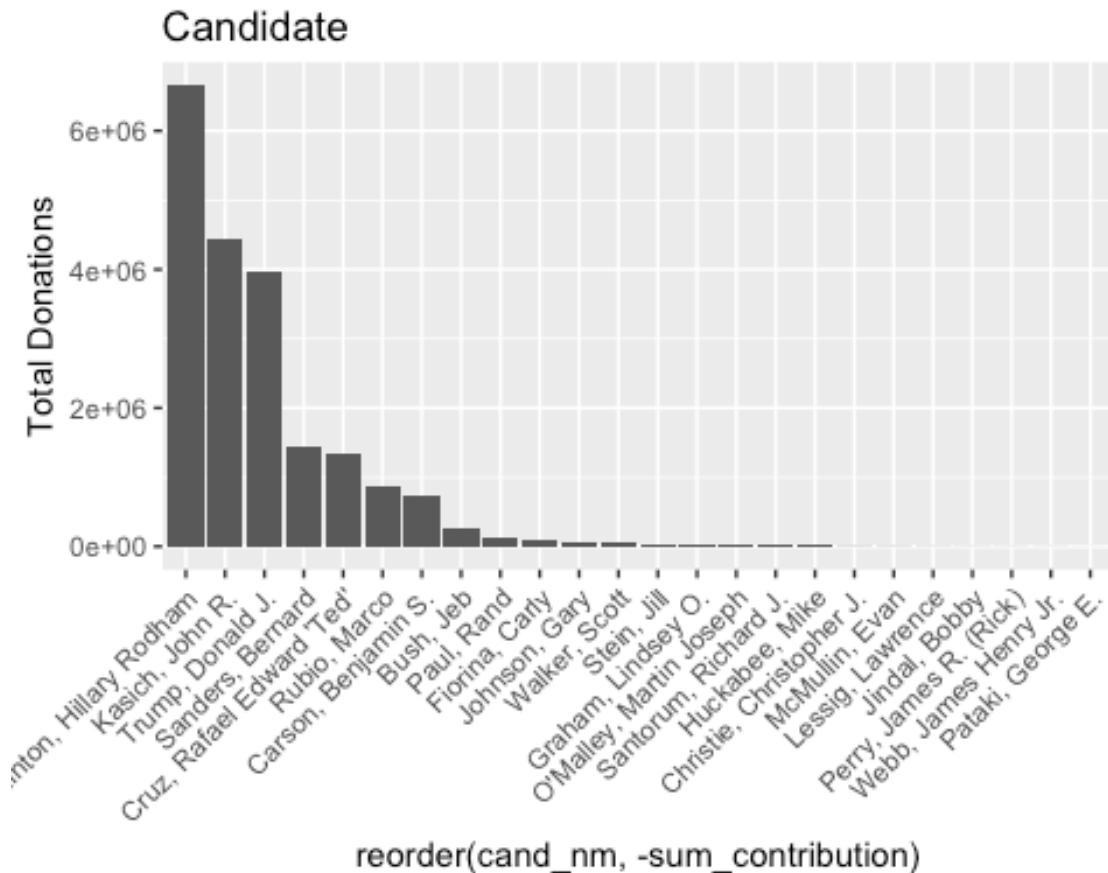


The bar charts highlights the difference even more. Since the other parties did not receive that much money, I'm going to compare only Republicans and Democrats for primary and general election.



Now I want to know how different is the amount of money that each candidate received from Ohioans over all? What about the average donation amount?



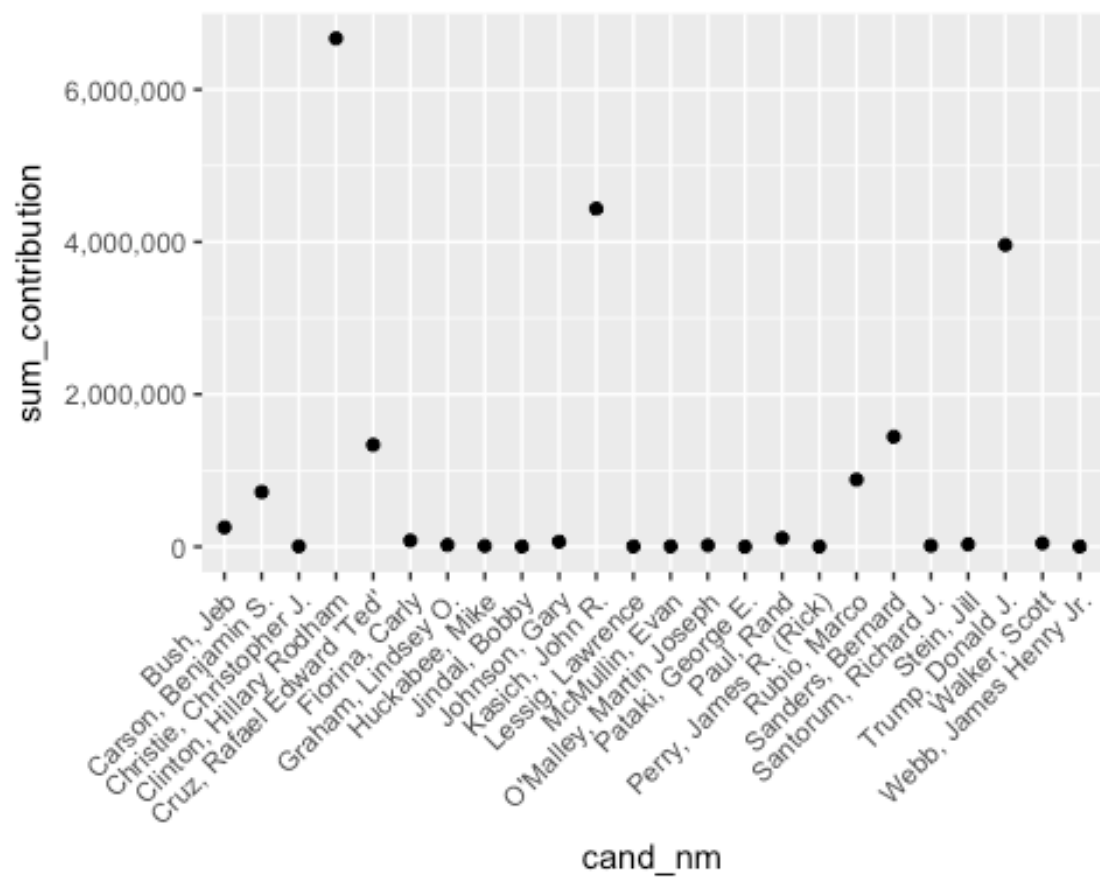


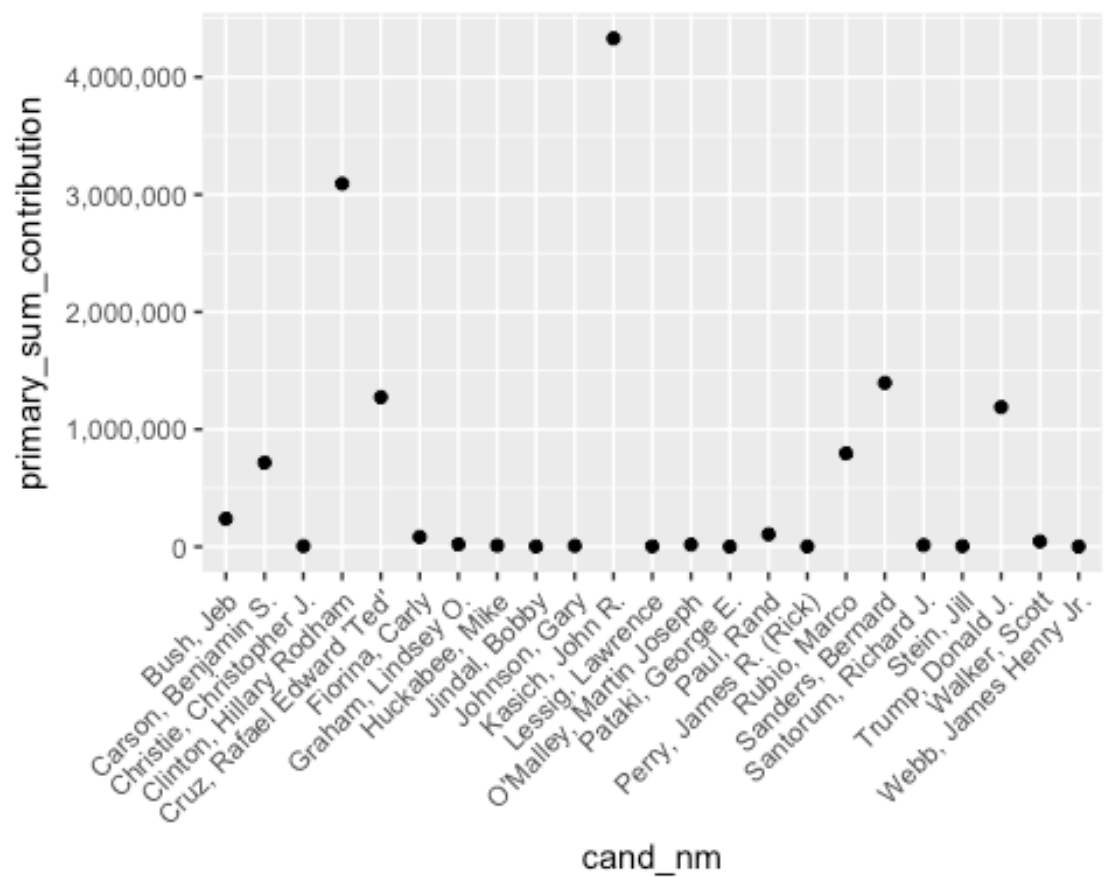
It looks like three candidates received way more money than the others. Overall Ohioan donated more money to Hillary Clinton's campaign than any other candidate. Apparently Kasich and Bush supporters donated more money on average per donation but O'Malley and Pataki are not far behind. George Pataki received the least amount of money from them, which is really interesting since Pataki's supporters donated on average more money than many other candidates. In contrast, Sanders has the forth place when it comes to the total contribution but he has the lowest mean contribution!

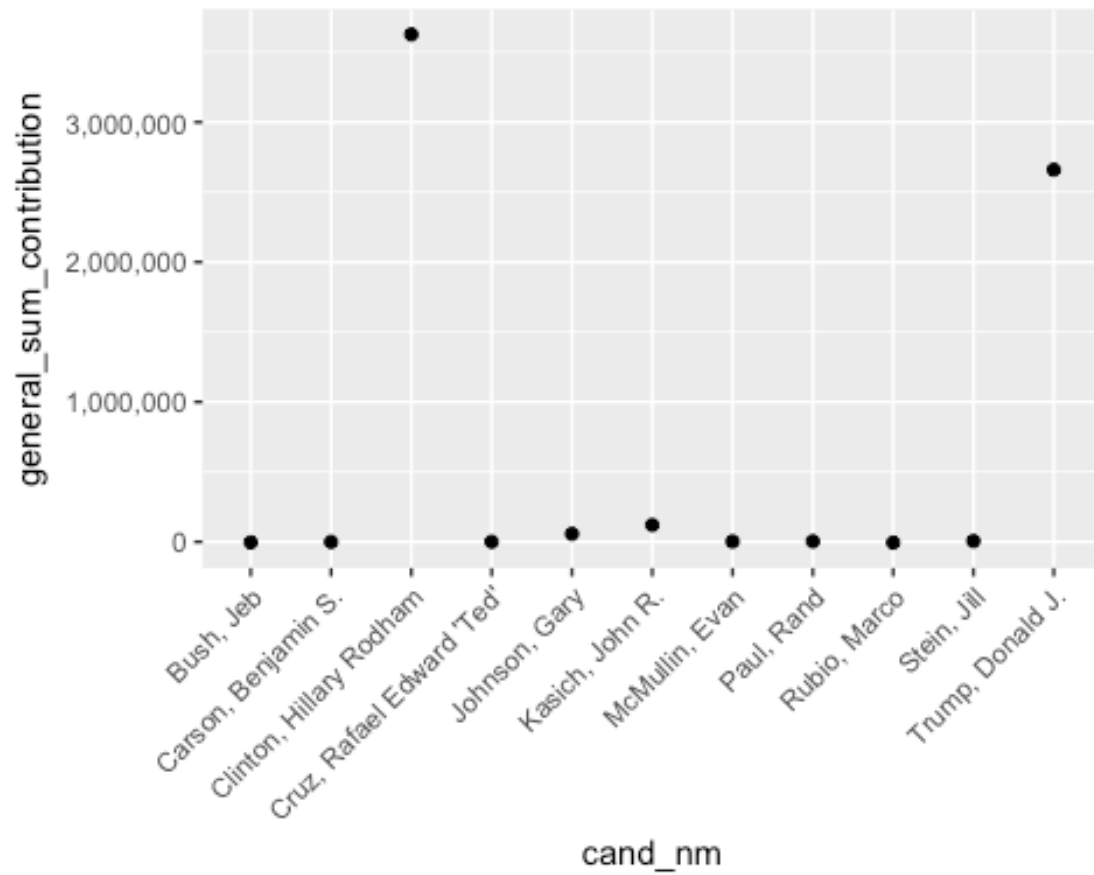
Another question is who from each party received the most money? To answer this question, we only need to find the candidate from Republican Party who received he most money: Clinton received the most money overall, therefore she got the most money among democrats as well. Libertarian, Independent, and Green Party had only one candidate so there is nothing to compare. So now I'm going to find among republicans, who received the most monetary contribution?

```
## [1] Kasich, John R.
## 24 Levels: Bush, Jeb Carson, Benjamin S. ... Webb, James Henry Jr.
```

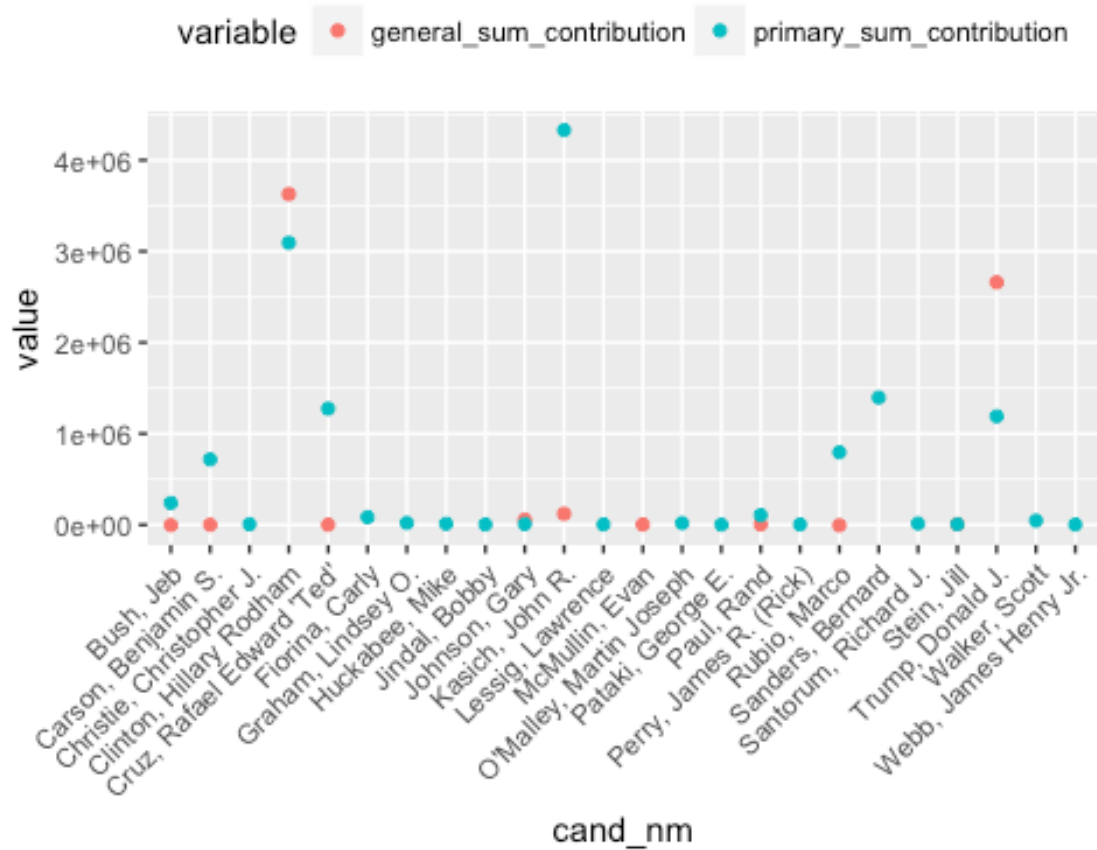
Kasich received the most amounts of contributions, which makes sense since he is the governor of Ohio and Pataki received the least amount of money among republicans. But another thing came to my attention: Kasich received all those money during the primary election, while contribution to Trump and Clinton's campaign continued since they were nominees. So I would like to see primary and general election contribution stats separately:



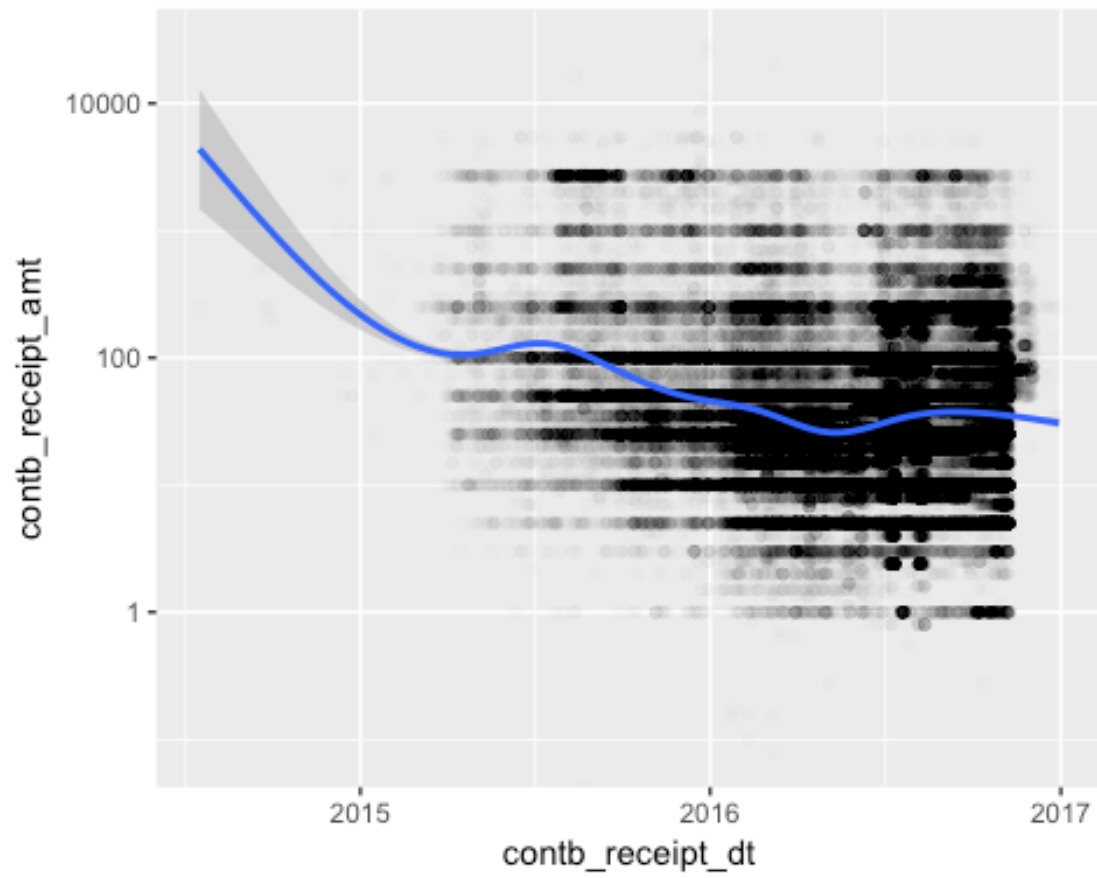




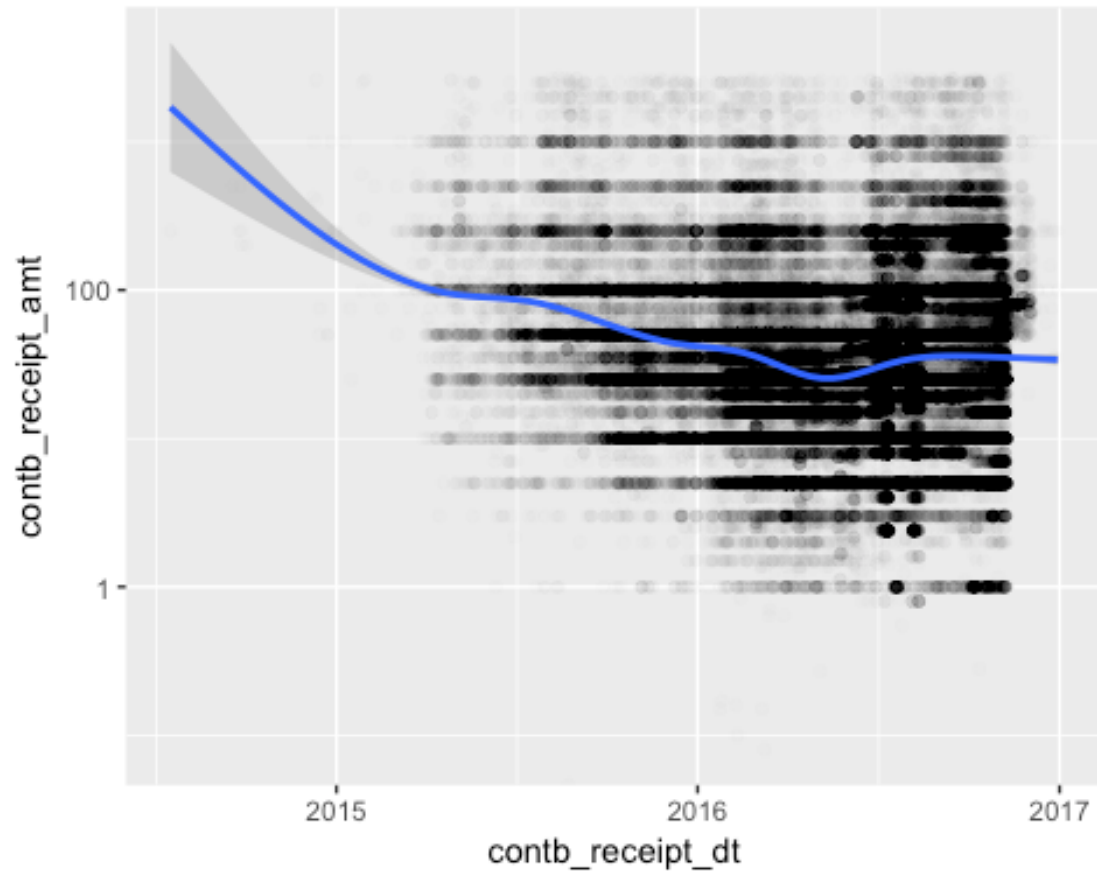
I don't think these plots going to help me! The scales are different on Y-axis and some candidates did not receive any contribution after primaries. What if I put genral and primary election together in one plot?



Now I can see and compare them better. Interesting point is even after primaries, people continued to donate to some republican or democrat candidates beside trump and Clinton.

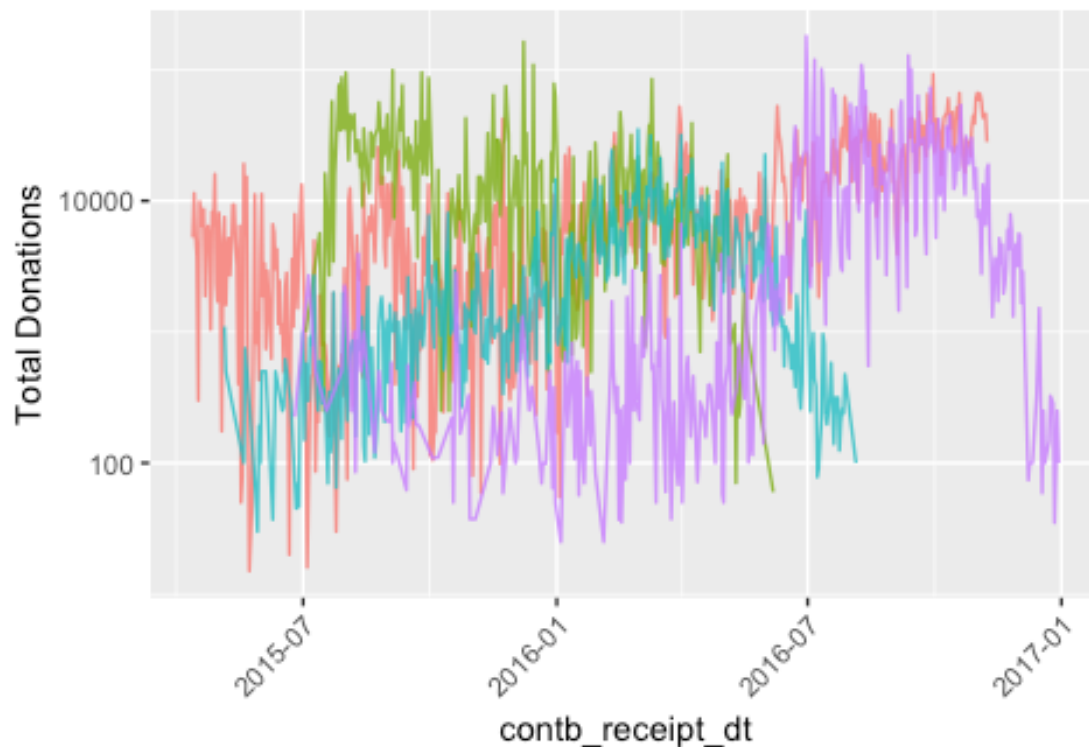


The refunds and corporation donations make it difficult to detect a pattern here. So I'm going to take look at individual donations.



it looks with time individual contributors donated smaller and smaller amount of money. Also it is interesting to see how donating to different candidates has changed during the lection. Specifically, I would like to see how it changed for the top 4 donation recipients.

nd_nm Clinton, Hillary Rodham Kasich, John R. Sanders, Bernard Tru



Another variable that I would like to investigate is the contributor's occupation for individual donors. It looks for individual donors retired and attorneys donated the highest amount of money! I'm a little surprised by almost half a million donation by not employed individuals!

```
## # A tibble: 5 × 3
```

	contbr_occupation	occup_sum_contribution	n
	<fctr>	<dbl>	<int>
## 1	RETIRED	3355186.2	42965
## 2	INFORMATION REQUESTED	882538.2	8289
## 3	ATTORNEY	541662.6	3172
## 4	NOT EMPLOYED	466603.8	10349
## 5	PHYSICIAN	436258.2	2973

And finally I would like to see how contributions has changes accross Ohio based on zipcodes:

```
## # A tibble: 5 × 4
```

	contbr_zip	zip_sum_contribution	mean_sum_contribution	n
	<chr>	<dbl>	<dbl>	<int>
## 1	44122	433207.0	225.9817	1917
## 2	44022	401964.9	439.7866	914
## 3	45243	388591.0	394.9096	984
## 4	45208	387420.6	311.4313	1244
## 5	43209	370421.5	248.2718	1492

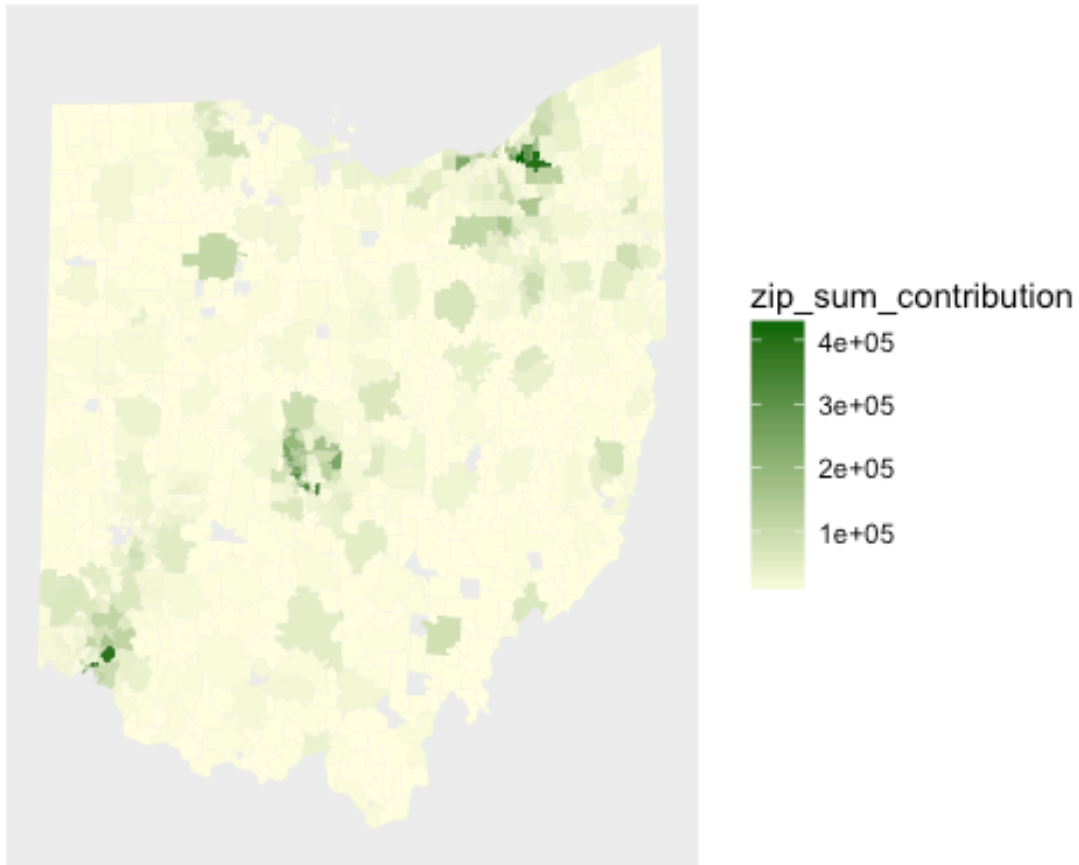
```
## # A tibble: 5 × 4
##   contbr_zip zip_sum_contribution mean_sum_contribution    n
##   <chr>      <dbl>                <dbl> <int>
## 1    43214    159490.3                77.12297 2068
## 2    44122    433207.0                225.98174 1917
## 3    44107    127138.3                 76.08515 1671
## 4    43220    254866.4                166.36189 1532
## 5    43221    309142.0                204.05412 1515
```

```
## # A tibble: 5 × 4
##   contbr_zip zip_sum_contribution mean_sum_contribution    n
##   <chr>      <dbl>                <dbl> <int>
## 1    43659      2700                2700.000     1
## 2    45262      2700                2700.000     1
## 3    43759      7500                1875.000     4
## 4    45254     11075                1845.833     6
## 5    43945      3400                1700.000     2
```

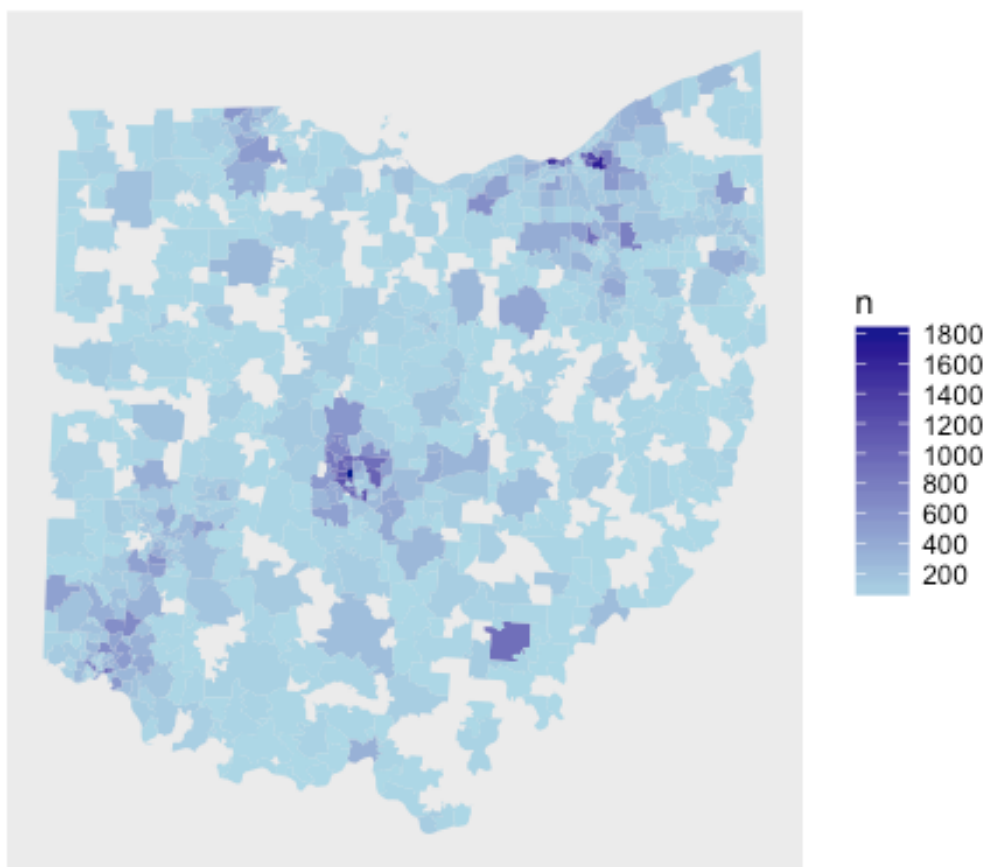
There are a couple of interesting facts about donations based on zip codes:

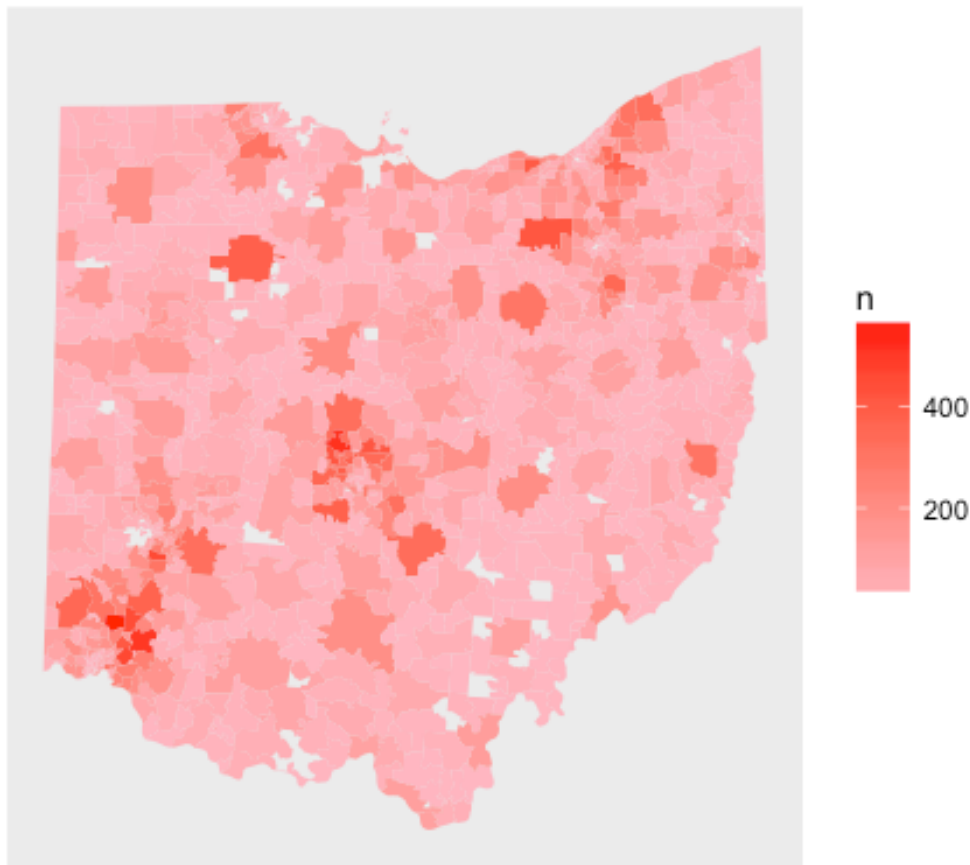
- 1- zip code 44122 (Cleveland area) the highest amount of total donations, which is interesting because neither average donation nor number of donations in this zip code are not the highest in the State.
- 2- The highest amount of individual contribution happened in two zip code with only 1 donations. I looked up those two: 43659 is very small zip code in Toledo and 45262 is a P.O.Box. The next highest mean belongs to another small zip code in east of Ohio with only 4 donations. There are 73 zip codes with only 1 donations. While I did not investigate all 73 of them, the hands full that I looked up were either one-block zip codes or P.O.Boxes. The highest number of donations belongs to zip code 43214 in Columbus with 2068 donations. I would like to see the distribution of donation based on sum of donations and number of donations.

```
## OGR data source with driver: ESRI Shapefile
## Source: "Ohio_zip_map", layer: "zip_code_area"
## with 1014 features
## It has 3 fields
## Integer64 fields read as strings: AREA_ID
```



This maps show three significantly darker green blue areas: Columbus, Cincinnati, and Cleveland, which is not surprising since these three are largest city in Ohio. But I don't get much more than this from it. So I'm going to compare donations to Democrats and Republicans across Ohio.





Both parties received more money from high density urban areas. Republicans also received a high number of donations in lower-density zip codes as well.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

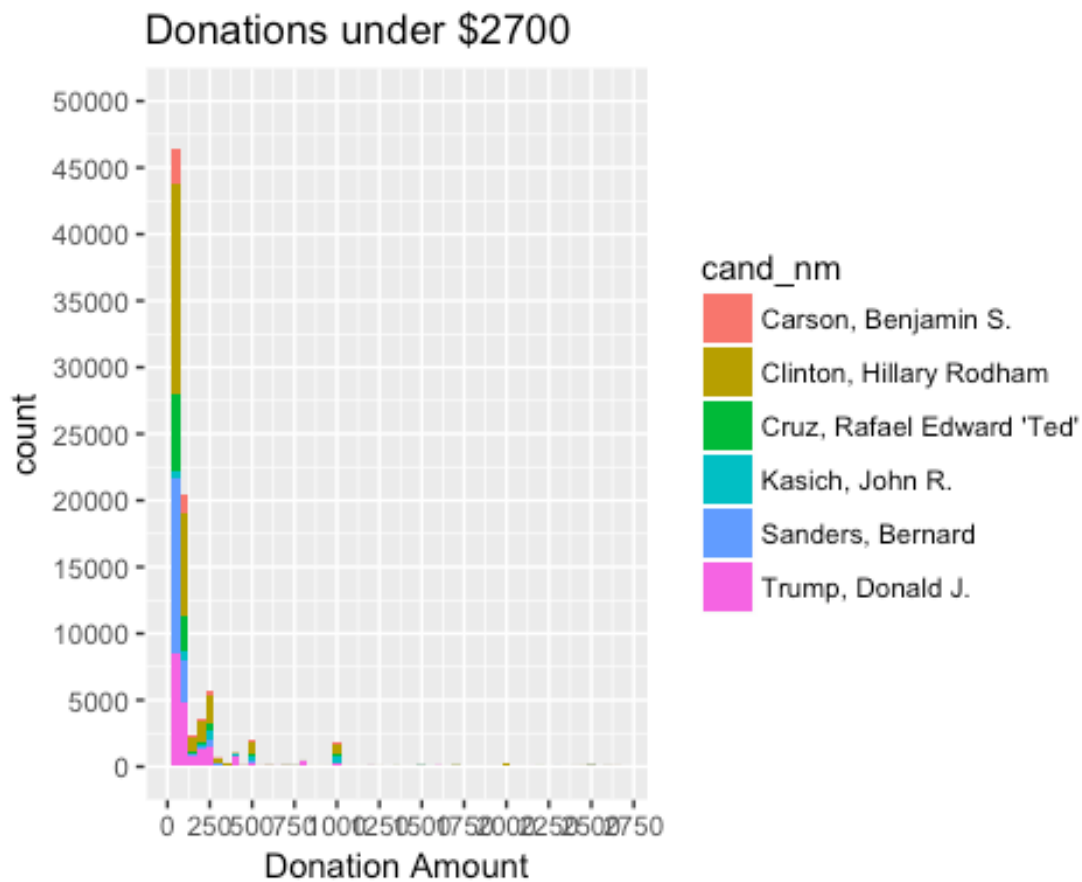
58% of the total contribution in Ohio went to the Republican Party. Since the number of Republican Party's candidates were 3 times more than Democrats (5 Democrats and 16 Republicans) this difference makes sense. However, in each party few candidates received the majority of the contributions. Another observation is for the primary election, Democrats received more donations than the Republicans, but for general elections Republicans received the majority of the donations. Finally, it looks like retired people donated the majority of the money during the 2016 presidential election.

Did you observe any interesting relationships between the other features?

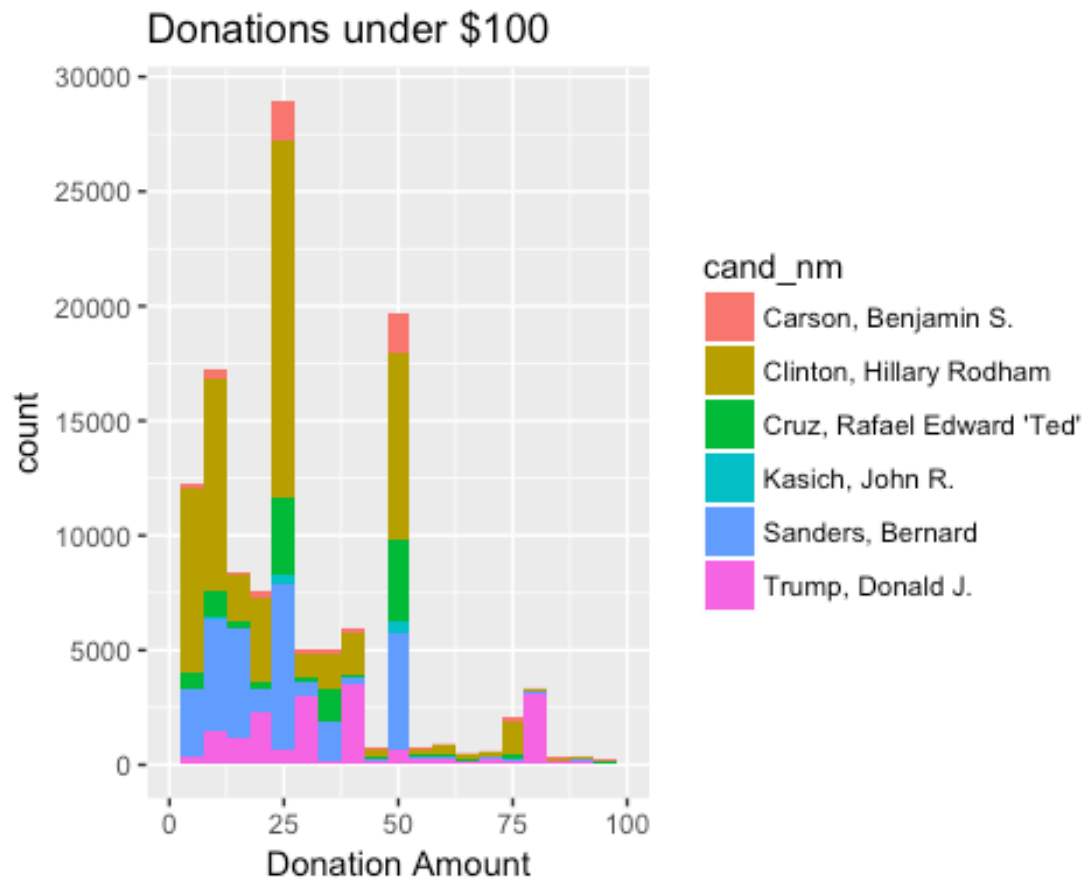
I found it really interesting that the unemployed folks donated half a million to the campaigns!

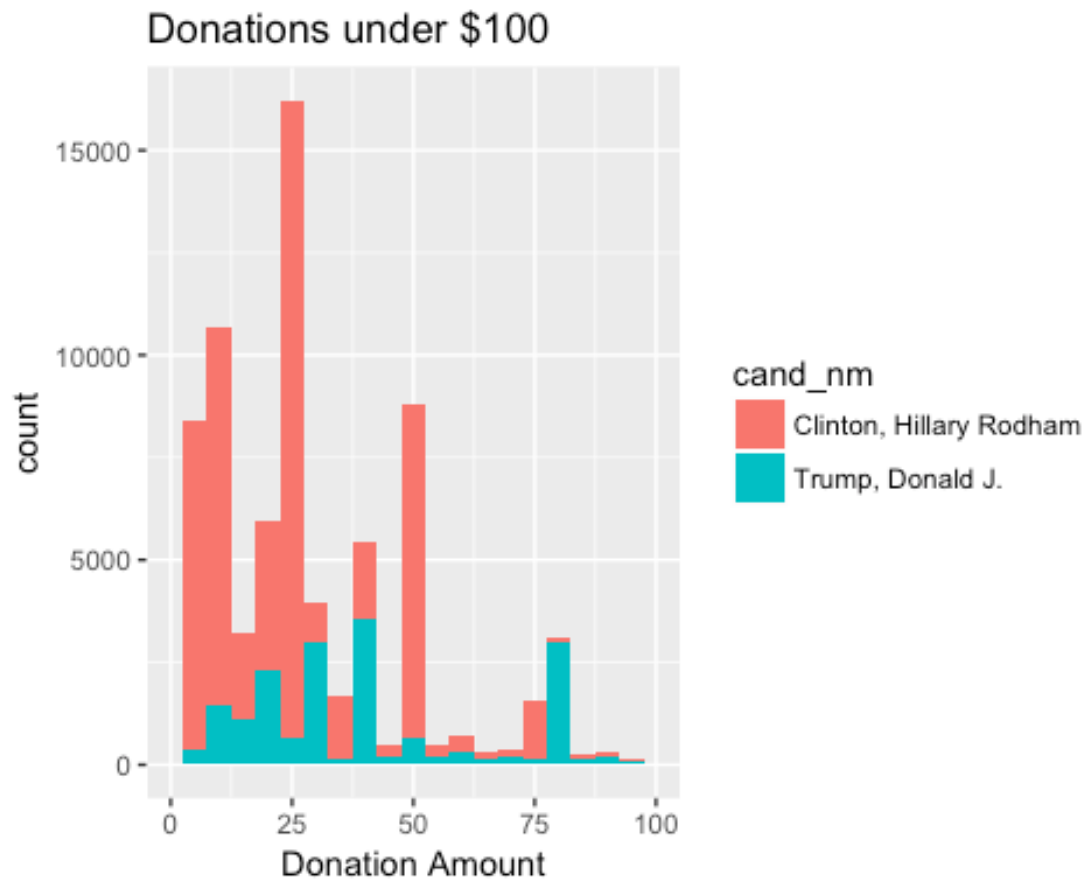
Multivariate Plots Section

Based what I learned about the donations, I wonder how the size of donation is different for each candidate.

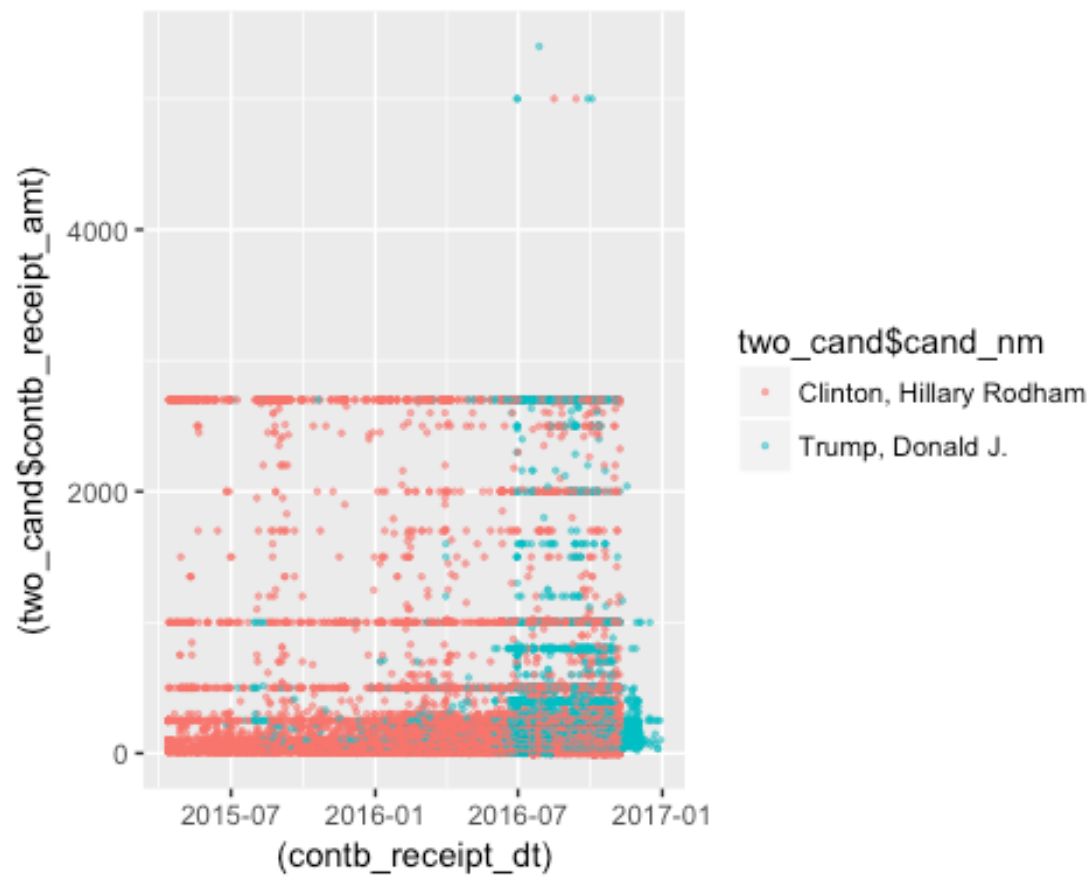


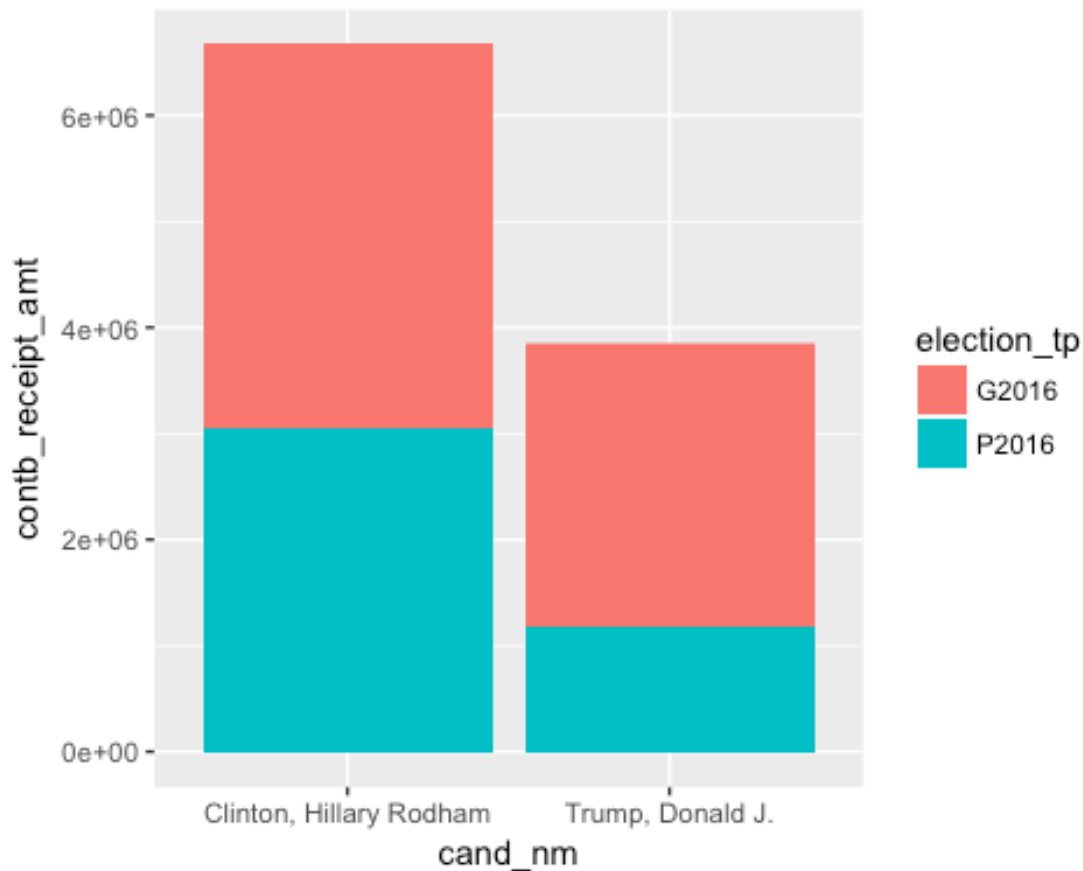
The most common donation amount is \$50 for all candidates. I'm going to make the charts more specific. I would like to see these contributions for top 6 as well as Clinton and Trump, for donations under \$100.





Ok! It looks \$25 donation is the most common donation. Clinton Also received the smallest donations as well. Did the donation amount's pattern changed during the campaign year for these two candidates?





The time series plot shows how trump's donations increased when he got the Republican Parties nomination. The stacked bar makes it really clear that while Trump received significantly less money during the primaries than Clinton, for the general election he was able to gather more donations.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of

looking at your feature(s) of interest? Were there any interesting or surprising

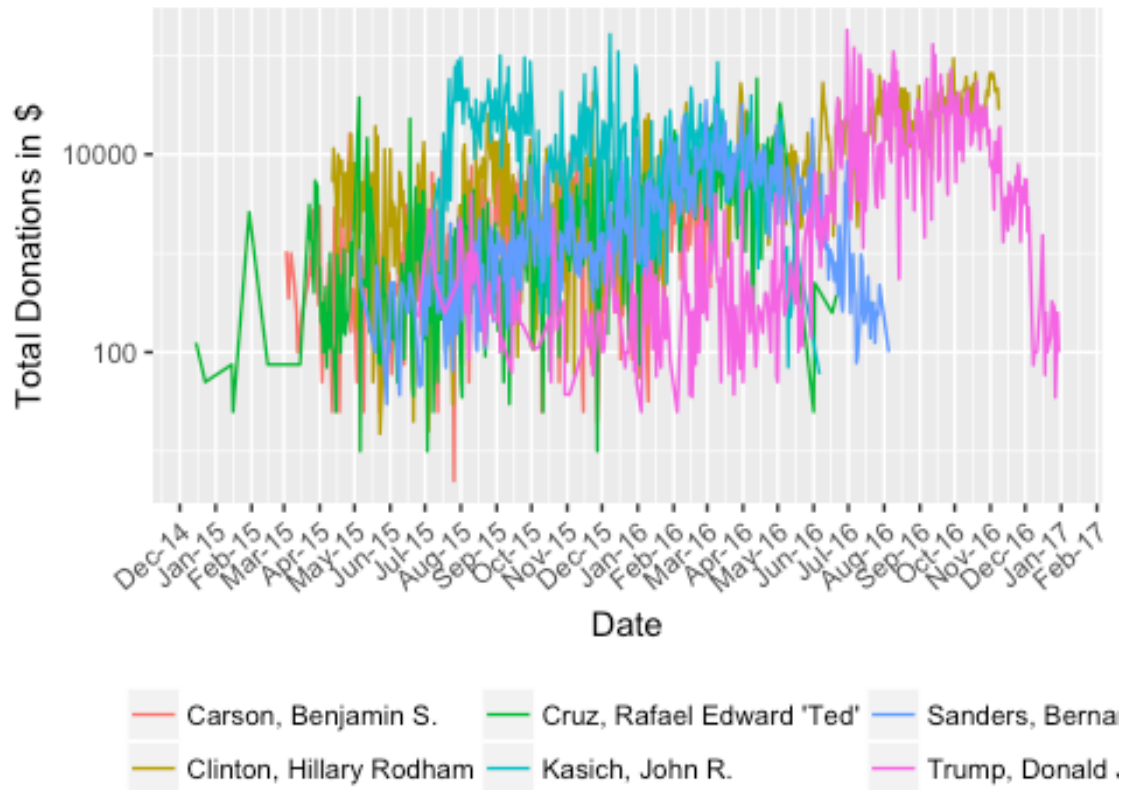
interactions between features?

Almost all of the relationships were predictable: more donations will be made closer to primary and general election day. Two major parties will receive more money than others as well as two nominees from those parties.

Final Plots and Summary

Plot One

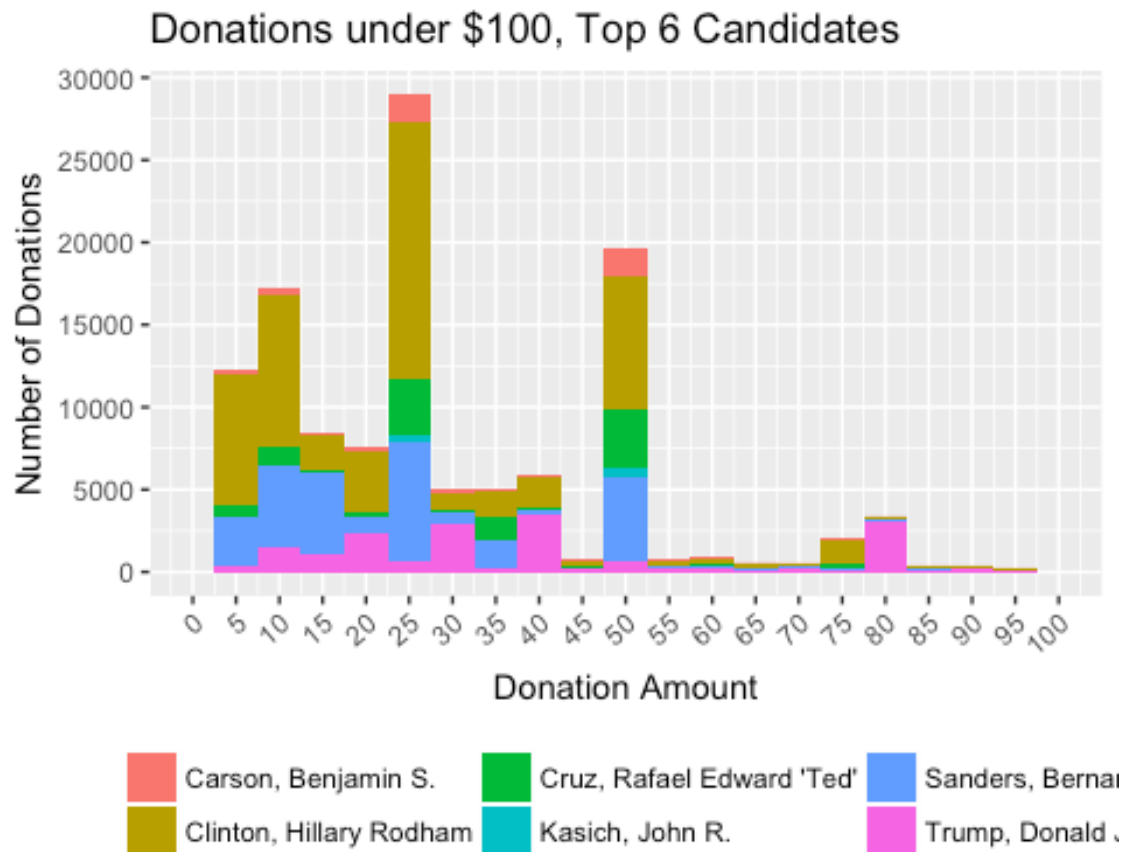
Candidate Donations Over Time, Top 6 Candidates



Description One

This plot shows the monetary contribution over time to the top 6 candidates in terms of over all donation. Sanders's campaign received money even after 2016 Democratic National Convention and nomination of Clinton as Democrats candidate. It might be because his supporters were hoping to convince him to run as an independent candidate. This plot also shows interesting spikes in Trump and Kasich's campaign in some points. For Trump is it easy to explain: most of the donations happened after he became Republican's nominee. For Kasich, the most noticeable spike happened on mid December.

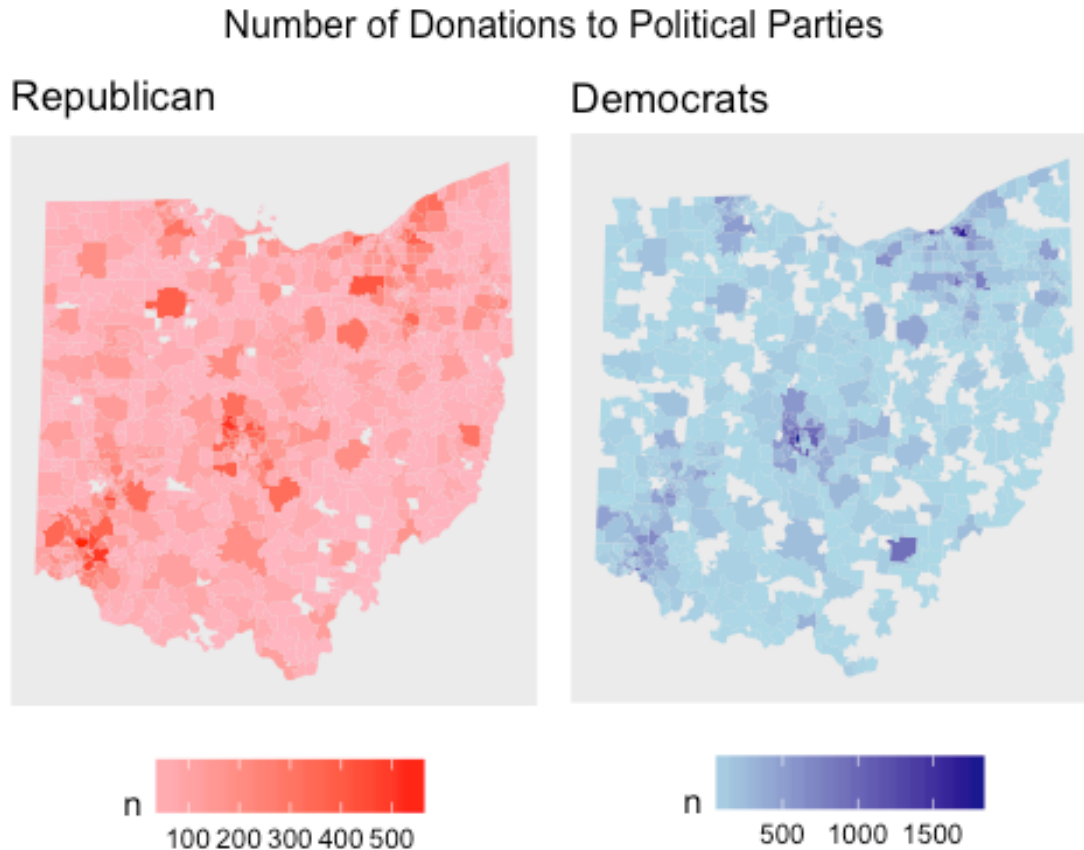
Plot Two



Description Two

This plot shows the under \$100 donation for the top 6 donation recipients candidates. I found this plot very interesting since it shows clearly what sanders said during his campaign: average donation to his campaign was \$27. I think majority of his supporter donated somewhere around \$25 so the mode of donations to his campaign got so close to the average.

Plot Three



Description Three

I like this visualization because it shows how different areas donated differently to Democrats and Republicans. While the darkness of the colors may represent almost the same amount of contribution for both parties in some areas, but it is important to pay attention to the legend: the scales are extremely different. I tried to force equal breaks for these two maps but unfortunately it did not work.

Reflection

Ohio 2016 presidential campaign contributions data has information on more than 160000 campaign donations July 2014 to December 2016. While I enjoyed working with such a large dataset and on a topic that I really interested in, since this was my first big project in R faced some challenges.

Challenge and My Solutions

- The first big problem that I faced was a simple data cleaning for the zip codes. After a lot of search I found `substring()` will solve part of my problem in just 1 line of code.
- I wanted to add a party variable to my dataset so I can summarize the data and visualize it based on political parties. I spend a good amount of time googling and learning about `ifelse()` and `%in%` which helped me to add that variable.
- But the challenging part of data cleaning was removing the refunded contributions. I looked into many Stack Overflow post, watched many YouTube tutorials, looked into similar analysis and finally

combination of all of those and with a hint from my husband I was able to use the zoo package with trial and error! But it worked and I cleaned out the refunded records.

- The most interesting and exciting part of visualization was the mapping. Since I have many years of experience of using mapping software such as ArcGIS, I was really excited to try mapping in R. However I found it really challenging. I used a shape file of Ohio's zip codes and every single step took me a great deal of time. I spend two days on reading different resources and finally a combination of online resources and Udacity's awesome forum mentors helped me to make a map.
- Another problem that I faced was many times I had to go back to data cleaning and exclude some records for all analysis or for some of the plots. I think I should have study the data in more depth at the beginning to avoid this going back and forth.
- I spend a good amount of time on thinking about the variables that I have and their possible relationships as well as what is the best way to present them. Trying different plot types, sub setting the data and grouping it and making new variables were the strategies that I used.
- Using ggplot2 to the extend that I did was really awesome. I feel confident that now I can use this package really well.

Future Work

The first thing that comes to mind is probably adding different data from reliable sources will take this analysis to the next level and makes it easy to accept or reject after election analysis regarding the socio-economic group of people who donated and voted for Stein, Trump, Clinton, and Sanders. Census data will be a good resource for this and poverty, race, gender, and level of education are the additional data that will help with this analysis.