Presidential campaigns

Pragya Jha

December 17, 2017

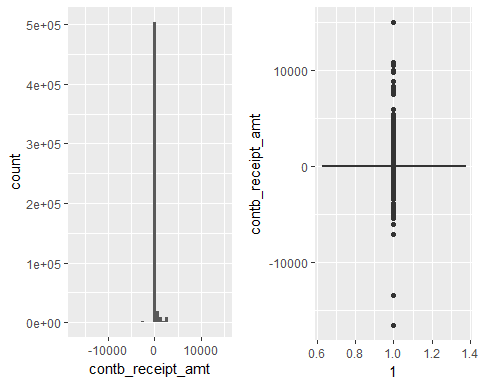
# Introduction

This is an exploration of 2016 US presidential campaign donations in the state of Texas. For this exploration data analysis, I am researching the 2016 presidential campaign finance data from Federal Election Commission. The dataset contains financial contribution transaction till 2016.

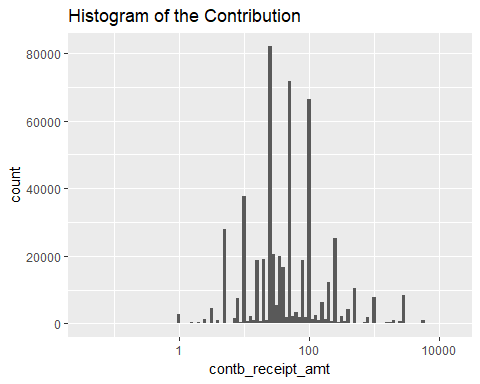
## [1] 548396 18

## 'data.frame': 548396 obs. of 18 variables:  
## $ cmte\_id : Factor w/ 26 levels "C00458844","C00496034",..: 7 16 16 8 7 7 16 16 7 16 ...  
## $ cand\_id : Factor w/ 25 levels "P00003392","P20002671",..: 1 23 23 12 1 1 23 23 1 23 ...  
## $ cand\_nm : Factor w/ 25 levels "Bush, Jeb","Carson, Benjamin S.",..: 4 23 23 20 4 4 23 23 4 23 ...  
## $ contbr\_nm : Factor w/ 129927 levels "'CALL, JANE",..: 78846 104629 104633 67432 47336 126503 99534 93791 58977 93812 ...  
## $ contbr\_city : Factor w/ 2252 levels "","-2111",". FREEPORT. PAY FOR EMPL",..: 427 699 2200 1474 1802 116 1858 965 1918 1248 ...  
## $ contbr\_st : Factor w/ 1 level "TX": 1 1 1 1 1 1 1 1 1 1 ...  
## $ contbr\_zip : Factor w/ 66331 levels "","0","1020",..: 50056 18335 32874 19905 48356 57005 45705 32932 17809 17198 ...  
## $ contbr\_employer : Factor w/ 40164 levels "","'CALL PHONOGRAPH LLC",..: 24038 3596 28912 24933 6082 9778 17644 17644 8244 30753 ...  
## $ contbr\_occupation: Factor w/ 17467 levels "","-"," CERTIFIED REGISTERED NURSE ANESTHETIS",..: 13568 13238 13568 10154 14388 14007 7704 7704 11436 7239 ...  
## $ contb\_receipt\_amt: num 37.1 127.1 80 15 50 ...  
## $ contb\_receipt\_dt : Factor w/ 746 levels "1-Apr-15","1-Apr-16",..: 168 514 6 665 724 605 534 582 317 624 ...  
## $ receipt\_desc : Factor w/ 61 levels "","\* EARMARKED CONTRIBUTION: SEE BELOW REATTRIBUTION/REFUND PENDING",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ memo\_cd : Factor w/ 2 levels "","X": 2 2 2 1 2 2 2 2 2 2 ...  
## $ memo\_text : Factor w/ 291 levels "","$0.10 REFUNDED ON 10/24/2016",..: 40 1 1 31 40 40 1 1 40 1 ...  
## $ form\_tp : Factor w/ 3 levels "SA17A","SA18",..: 2 2 2 1 2 2 2 2 2 2 ...  
## $ file\_num : int 1091718 1146165 1146165 1077404 1091718 1091718 1146165 1146165 1091718 1146165 ...  
## $ tran\_id : Factor w/ 543825 levels "A0009C127D0D344E3885",..: 139500 461349 403810 493948 139256 139041 429428 400817 139706 461688 ...  
## $ election\_tp : Factor w/ 5 levels "","G2016","O2016",..: 5 2 2 5 5 5 2 2 5 2 ...

This dataset contains 548396 contributions and 18 variables. To start, I want to have a glance how the contribution distributed.



I realized that there were so many outliers(extreme high and extreme low values), it was impossible to see details. And there were negative contributions too.



##   
## 5 10 100 50 25   
## 28008 37692 65084 69794 80376

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -16600.0 20.0 38.0 138.2 100.0 15000.0

Transforming to log10 to better understand the distribution of the contribution. The distribution looks normal and the data illustrated that most donors made small amount of contributions.

To perform in depth analysis, I decided to omit the negative contributions which I believe they were refund and contributions that exceed $2700 limit, because it breaks Federal Election Campaign Act and will be refunded. This means 5897 contributions are omitted.

## [1] 10093

## [1] 12084

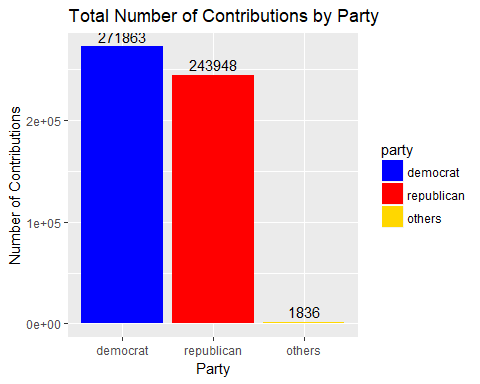
I will need to add more variables such as candidate party affiliate, donors' gender and donors' zipcodes.

The additional variables are:

-party: candidates party affilliation. -contbr\_first\_nm: contributor's first name will be used to predict gender. -gender: contributor's gender. -Latitude: Donor's latitude for map creation. -Longitute: Donor's longitude for map creation.

After adding the variables, I wonder what the contribution distribution looks like across the parties, candidates, genders and occupations.

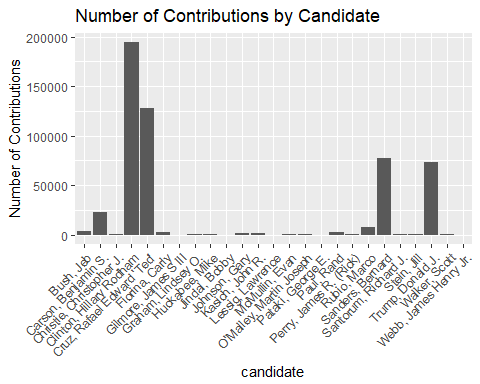
## # A tibble: 3 x 5  
## party sum\_party number\_of\_candidate mean\_party n  
## <chr> <dbl> <int> <dbl> <int>  
## 1 democrat 25422745 5 5084549.0 271863  
## 2 others 442253 3 147417.7 1836  
## 3 republican 47340419 17 2784730.6 243948



## [1] 517647

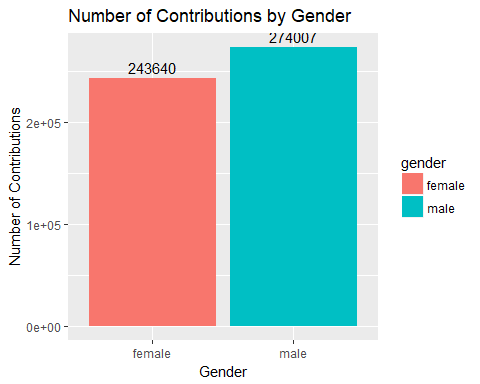
Until November, 2016, total number of donations made to the presidential election near 515K, and the Democratic party took more than 250K.

##   
## Bush, Jeb Carson, Benjamin S.   
## 3160 22471   
## Christie, Christopher J. Clinton, Hillary Rodham   
## 188 194289   
## Cruz, Rafael Edward 'Ted' Fiorina, Carly   
## 127925 2438   
## Gilmore, James S III Graham, Lindsey O.   
## 5 147   
## Huckabee, Mike Jindal, Bobby   
## 705 55   
## Johnson, Gary Kasich, John R.   
## 1309 1158   
## Lessig, Lawrence McMullin, Evan   
## 31 191   
## O'Malley, Martin Joseph Pataki, George E.   
## 208 9   
## Paul, Rand Perry, James R. (Rick)   
## 2943 547   
## Rubio, Marco Sanders, Bernard   
## 8095 77278   
## Santorum, Richard J. Stein, Jill   
## 244 336   
## Trump, Donald J. Walker, Scott   
## 73443 415   
## Webb, James Henry Jr.   
## 57



There were total 25 candidates, Hillary Clinton was the leader in the number of contributions, followed by Rafael Edward, then Bernard Sanders and then Donald Trump.

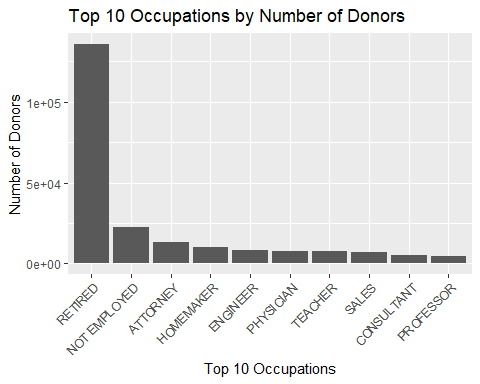
## # A tibble: 2 x 3  
## gender sum\_gen n\_gen  
## <chr> <dbl> <int>  
## 1 female 29100036 243640  
## 2 male 44105382 274007



Who are those donors?

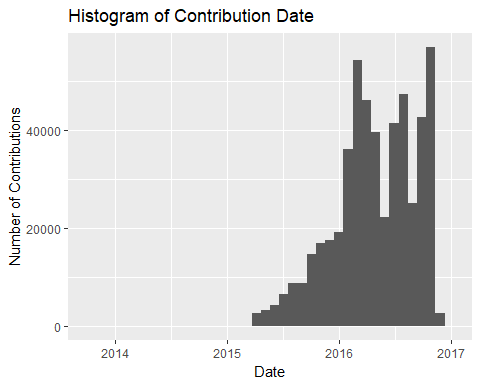
## # A tibble: 10 x 4  
## contbr\_occupation sum\_occu mean\_occu n  
## <fctr> <dbl> <dbl> <int>  
## 1 RETIRED 14463581.2 106.39758 135939  
## 2 NOT EMPLOYED 1115695.9 49.29074 22635  
## 3 ATTORNEY 3970684.4 293.77659 13516  
## 4 HOMEMAKER 3396748.8 327.93482 10358  
## 5 ENGINEER 1178075.9 138.85855 8484  
## 6 PHYSICIAN 1935789.1 244.63404 7913  
## 7 TEACHER 445565.3 57.61123 7734  
## 8 SALES 717011.8 105.16454 6818  
## 9 CONSULTANT 935606.1 179.92425 5200  
## 10 PROFESSOR 470172.9 97.62726 4816

## # A tibble: 10 x 4  
## contbr\_occupation sum\_occu mean\_occu n  
## <ord> <dbl> <dbl> <int>  
## 1 RETIRED 14463581.2 106.39758 135939  
## 2 NOT EMPLOYED 1115695.9 49.29074 22635  
## 3 ATTORNEY 3970684.4 293.77659 13516  
## 4 HOMEMAKER 3396748.8 327.93482 10358  
## 5 ENGINEER 1178075.9 138.85855 8484  
## 6 PHYSICIAN 1935789.1 244.63404 7913  
## 7 TEACHER 445565.3 57.61123 7734  
## 8 SALES 717011.8 105.16454 6818  
## 9 CONSULTANT 935606.1 179.92425 5200  
## 10 PROFESSOR 470172.9 97.62726 4816



When we count the number of donors, retired people take the first place, followed by not employed people, attorney comes to the third, consultant and professor are among the least in terms of number of contributions.

## Min. 1st Qu. Median Mean 3rd Qu.   
## "2013-10-21" "2016-02-07" "2016-04-29" "2016-04-27" "2016-08-12"   
## Max.   
## "2016-12-31"



# Univariate Analysis

## What is the structure of your dataset?

There are 517647 contributions and 18 variables. The variables that interest to me and I will be using are:

-cand\_nm: Candidate Name -contbr\_zip: Contributor Zipcode -contbr\_nm: Contributor name (first name in particular) -contbr\_occupation: Contributor Occupation -contb\_receipt\_amt: Contribution Amount -contb\_receipt\_dt: Contribution date Other observations:

Most people contribute small amount of money.

The democratic party receive the most number of donations. Hillary Clinton have the most supporters.

Retired people make the most number of contributions.

## What is(are) the main features of interest in your dataset?

The main features in the dataset are party, candidate and contribution amount. I'd like to find the answers to my questions at the beginning of this report. I'd also like to try to use combination of variables to build a logistics regression model to predictive a donor's contribution party.

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

Gender, occupation, time of the contribution, location are likely contribute to the contribution amount and contribution party. I think occupation probably contributes most to the average contribution amount, and gender probably contributes most to the contribution party.

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

I created 5 variables:

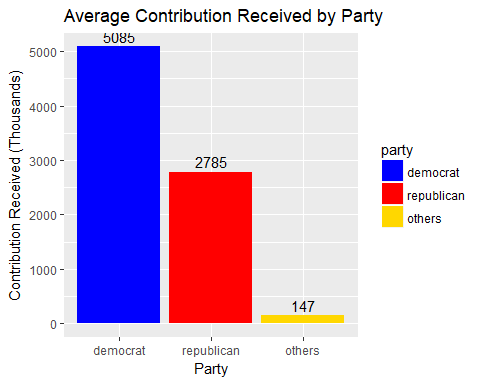
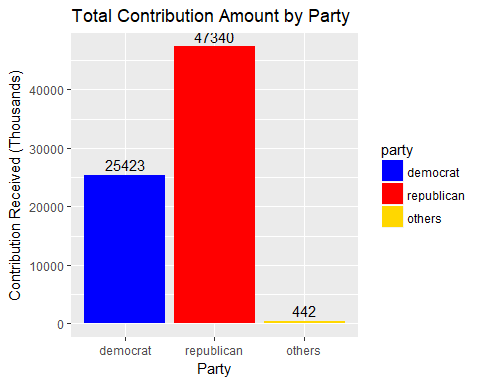
-party: candidates party affilliation. -contbr\_first\_nm: contributor's first name will be used to predict gender. -gender: contributor's gender. -Latitude: Donor's latitude for map creation. -Longitute: Donor's longitude for map creation.

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?

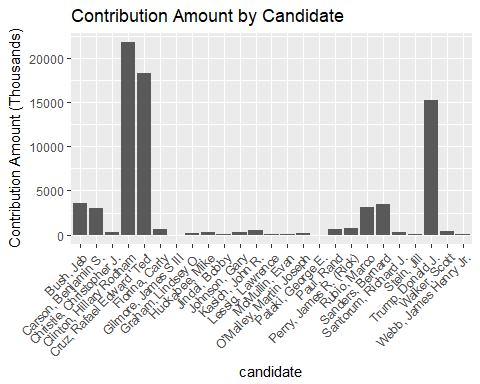
I omitted negative contributions because I believe they were refund, and I omitted contributions that exceed $2700 because because it breaks Federal Election Campaign Act and will be refunded.

# Bivariate plots

## # A tibble: 3 x 5  
## party sum\_party number\_of\_candidate mean\_party n  
## <ord> <dbl> <int> <dbl> <int>  
## 1 democrat 25422745 5 5084549.0 271863  
## 2 others 442253 3 147417.7 1836  
## 3 republican 47340419 17 2784730.6 243948

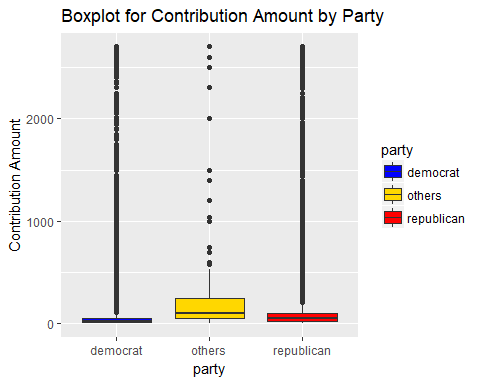


## ptx$cand\_nm  
## Pataki, George E. Gilmore, James S III   
## 6000.00 11800.00   
## Lessig, Lawrence Webb, James Henry Jr.   
## 20027.94 31090.16   
## McMullin, Evan Jindal, Bobby   
## 37237.00 42460.00   
## Stein, Jill O'Malley, Martin Joseph   
## 86921.17 139187.00   
## Graham, Lindsey O. Christie, Christopher J.   
## 146487.00 269355.00   
## Santorum, Richard J. Huckabee, Mike   
## 281288.20 284820.76   
## Johnson, Gary Walker, Scott   
## 318094.81 365142.00   
## Kasich, John R. Fiorina, Carly   
## 516021.00 607165.56   
## Paul, Rand Perry, James R. (Rick)   
## 675935.24 750595.16   
## Carson, Benjamin S. Rubio, Marco   
## 3049869.53 3183709.15   
## Sanders, Bernard Bush, Jeb   
## 3433576.71 3562397.09   
## Trump, Donald J. Cruz, Rafael Edward 'Ted'   
## 15291369.12 18296004.64   
## Clinton, Hillary Rodham   
## 21798863.28



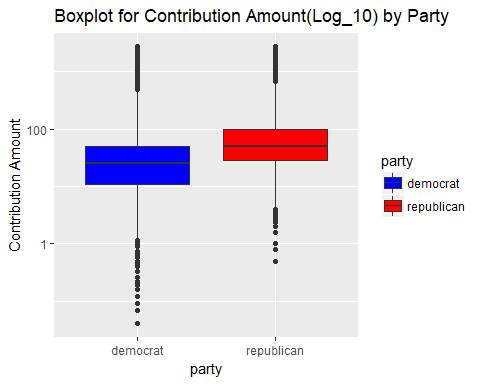
## [1] 73205418

To see contribution patterns between parties and candidates, I start with boxplots.



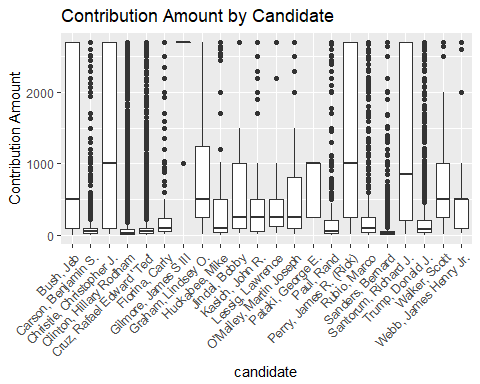
However, it is very hard to compare contributions among all parties at a glance because there are so many outliers. I will apply log scale and remove the 'others' party from now on because my analysis is focused on the Democratic party and the Republican party.

## ptx$party: democrat  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.04 11.00 25.00 93.51 50.00 2700.00   
## --------------------------------------------------------   
## ptx$party: republican  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.5 28.0 50.0 194.1 100.0 2700.0



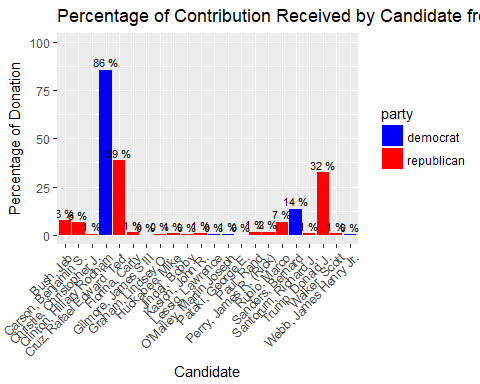
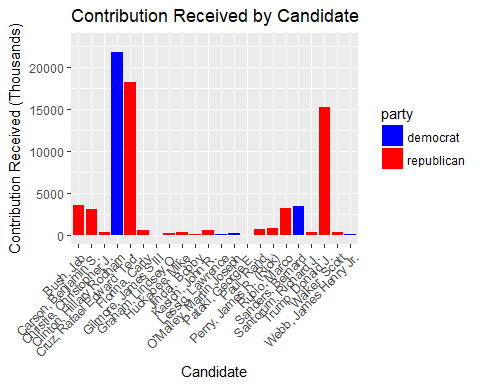
Now it is much better. Although the Republican has the higher median and mean, the Democrat has more variations and the distribution is more spread out. This indicates that the Democrat has more big and small donors.

## ptx$cand\_nm: Bush, Jeb  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 100 500 1127 2700 2700   
## --------------------------------------------------------   
## ptx$cand\_nm: Carson, Benjamin S.  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.5 25.0 50.0 135.7 100.0 2700.0   
## --------------------------------------------------------   
## ptx$cand\_nm: Christie, Christopher J.  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10 100 1000 1433 2700 2700   
## --------------------------------------------------------   
## ptx$cand\_nm: Clinton, Hillary Rodham  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.04 11.00 25.00 112.20 75.00 2700.00   
## --------------------------------------------------------   
## ptx$cand\_nm: Cruz, Rafael Edward 'Ted'  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 25 50 143 100 2700   
## --------------------------------------------------------   
## ptx$cand\_nm: Fiorina, Carly  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.0 50.0 100.0 249.0 232.2 2700.0   
## --------------------------------------------------------   
## ptx$cand\_nm: Gilmore, James S III  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1000 2700 2700 2360 2700 2700   
## --------------------------------------------------------   
## ptx$cand\_nm: Graham, Lindsey O.  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.0 250.0 500.0 996.5 1250.0 2700.0   
## --------------------------------------------------------   
## ptx$cand\_nm: Huckabee, Mike  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 35 100 404 500 2700   
## --------------------------------------------------------   
## ptx$cand\_nm: Jindal, Bobby  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 28 100 250 772 1000 2700   
## --------------------------------------------------------   
## ptx$cand\_nm: Johnson, Gary  
## NULL  
## --------------------------------------------------------   
## ptx$cand\_nm: Kasich, John R.  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.0 50.0 250.0 445.6 500.0 2700.0   
## --------------------------------------------------------   
## ptx$cand\_nm: Lessig, Lawrence  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 27.94 125.00 250.00 646.06 500.00 2700.00   
## --------------------------------------------------------   
## ptx$cand\_nm: McMullin, Evan  
## NULL  
## --------------------------------------------------------   
## ptx$cand\_nm: O'Malley, Martin Joseph  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.0 100.0 250.0 669.2 806.2 2700.0   
## --------------------------------------------------------   
## ptx$cand\_nm: Pataki, George E.  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 250.0 250.0 1000.0 666.7 1000.0 1000.0   
## --------------------------------------------------------   
## ptx$cand\_nm: Paul, Rand  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 25.0 50.0 229.7 201.6 2700.0   
## --------------------------------------------------------   
## ptx$cand\_nm: Perry, James R. (Rick)  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3 250 1000 1372 2700 2700   
## --------------------------------------------------------   
## ptx$cand\_nm: Rubio, Marco  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 30.0 100.0 393.3 250.0 2700.0   
## --------------------------------------------------------   
## ptx$cand\_nm: Sanders, Bernard  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 10.00 27.00 44.43 50.00 2700.00   
## --------------------------------------------------------   
## ptx$cand\_nm: Santorum, Richard J.  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.99 200.00 850.00 1152.82 2700.00 2700.00   
## --------------------------------------------------------   
## ptx$cand\_nm: Stein, Jill  
## NULL  
## --------------------------------------------------------   
## ptx$cand\_nm: Trump, Donald J.  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.8 32.0 80.0 208.2 200.0 2700.0   
## --------------------------------------------------------   
## ptx$cand\_nm: Walker, Scott  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.0 250.0 500.0 879.9 1000.0 2700.0   
## --------------------------------------------------------   
## ptx$cand\_nm: Webb, James Henry Jr.  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.0 100.0 500.0 545.4 500.0 2700.0



## Now let's examine within parties

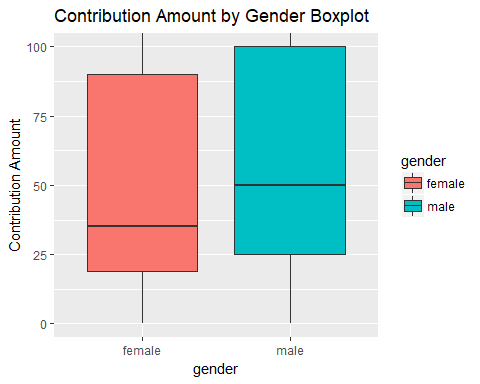
## # A tibble: 22 x 5  
## # Groups: party [2]  
## party cand\_nm sum\_can mean\_can n  
## <chr> <fctr> <dbl> <dbl> <int>  
## 1 republican Pataki, George E. 6000.00 666.6667 9  
## 2 republican Gilmore, James S III 11800.00 2360.0000 5  
## 3 democrat Lessig, Lawrence 20027.94 646.0626 31  
## 4 democrat Webb, James Henry Jr. 31090.16 545.4414 57  
## 5 republican Jindal, Bobby 42460.00 772.0000 55  
## 6 democrat O'Malley, Martin Joseph 139187.00 669.1683 208  
## 7 republican Graham, Lindsey O. 146487.00 996.5102 147  
## 8 republican Christie, Christopher J. 269355.00 1432.7394 188  
## 9 republican Santorum, Richard J. 281288.20 1152.8205 244  
## 10 republican Huckabee, Mike 284820.76 404.0011 705  
## # ... with 12 more rows



Within each party, majority of the donations were received by only few candidates. For Democratic party, Hillary Clinton and Bernard Sanders take almost 100% of all donations to the Democratic party, and of which, 86% went to Hillary Clinton. For the Republican party, Rafael Edward led the way taking 39% of all donations to the Republican party. Rafael Edward, Donald Trump, Marco Rubio, Ted Cruz all together taking 86% of all donations to the Republican party, the remaining 14% were shared by the other Republican candidates.

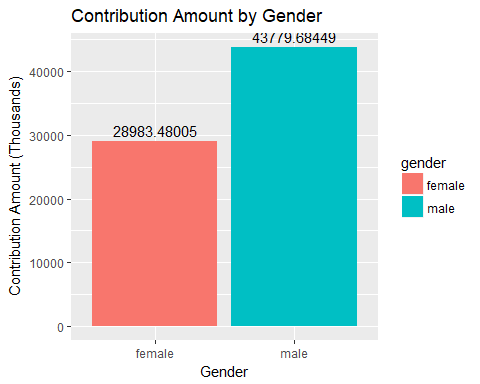
From the above charts, we are able to see who were the top candidates in each party in Massachusetts. I will examine the following candidates who received at least 8% of total donations in their party in details later.

## [1] "Clinton, Hillary Rodham" "Sanders, Bernard"   
## [3] "Trump, Donald J." "Rubio, Marco"   
## [5] "Cruz, Rafael Edward 'Ted'"

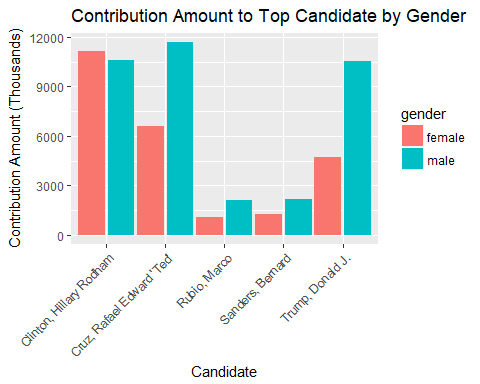


## ptx$gender: female  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.04 19.00 35.00 119.24 90.00 2700.00   
## --------------------------------------------------------   
## ptx$gender: male  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.12 25.00 50.00 160.52 100.00 2700.00

## # A tibble: 2 x 3  
## gender sum\_gen n  
## <chr> <dbl> <int>  
## 1 female 28983480 243066  
## 2 male 43779684 272745

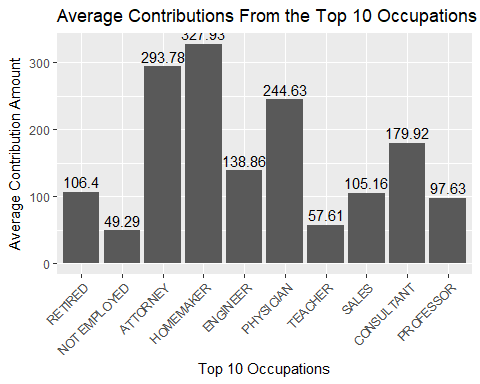
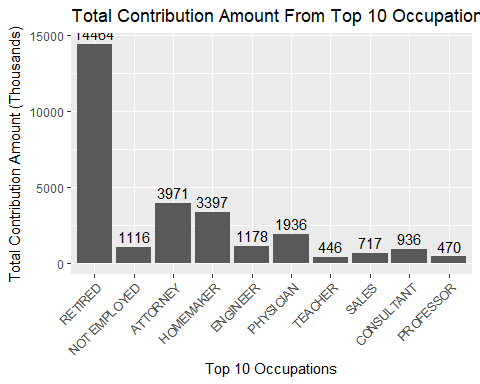


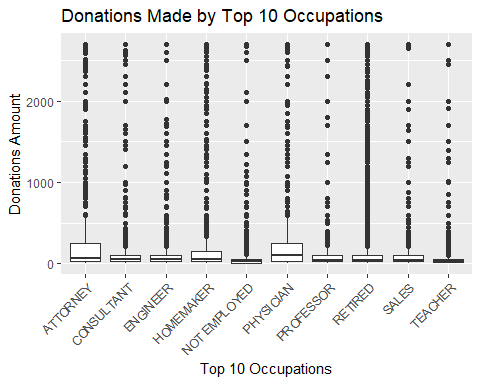
## # A tibble: 10 x 3  
## # Groups: cand\_nm [?]  
## cand\_nm gender sum\_gen\_can  
## <fctr> <chr> <dbl>  
## 1 Clinton, Hillary Rodham female 11181372  
## 2 Clinton, Hillary Rodham male 10617491  
## 3 Cruz, Rafael Edward 'Ted' female 6606053  
## 4 Cruz, Rafael Edward 'Ted' male 11689952  
## 5 Rubio, Marco female 1081671  
## 6 Rubio, Marco male 2102038  
## 7 Sanders, Bernard female 1250888  
## 8 Sanders, Bernard male 2182689  
## 9 Trump, Donald J. female 4715652  
## 10 Trump, Donald J. male 10575717

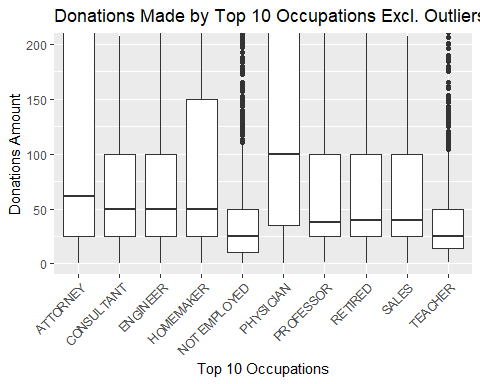


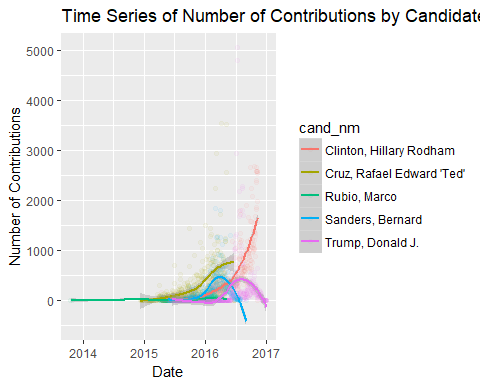
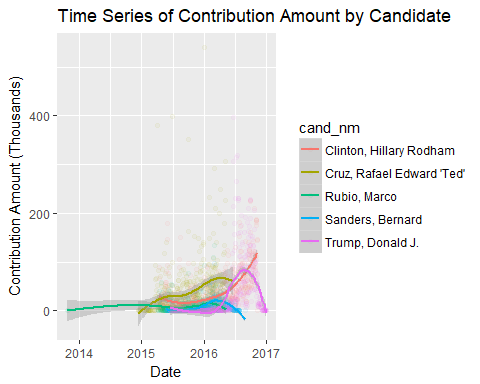
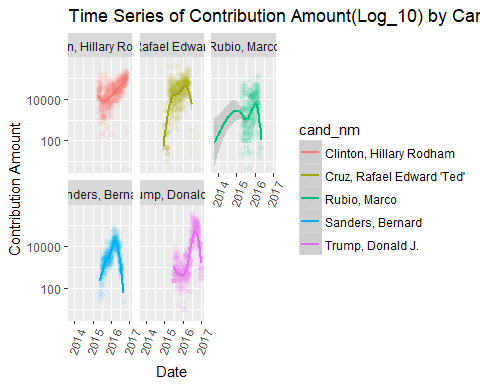
## How about total contribution amount and average contribution amount cross top 10 occupations?

## # A tibble: 10 x 4  
## contbr\_occupation sum\_occu mean\_occu n  
## <ord> <dbl> <dbl> <int>  
## 1 RETIRED 14463581.2 106.39758 135939  
## 2 NOT EMPLOYED 1115695.9 49.29074 22635  
## 3 ATTORNEY 3970684.4 293.77659 13516  
## 4 HOMEMAKER 3396748.8 327.93482 10358  
## 5 ENGINEER 1178075.9 138.85855 8484  
## 6 PHYSICIAN 1935789.1 244.63404 7913  
## 7 TEACHER 445565.3 57.61123 7734  
## 8 SALES 717011.8 105.16454 6818  
## 9 CONSULTANT 935606.1 179.92425 5200  
## 10 PROFESSOR 470172.9 97.62726 4816



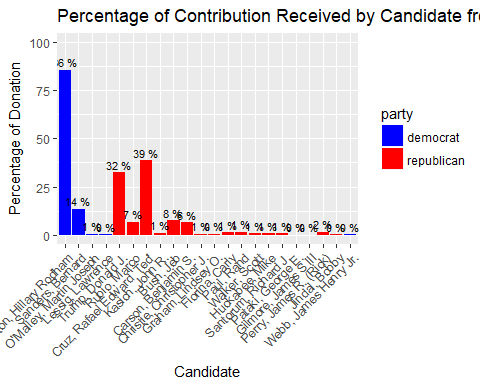




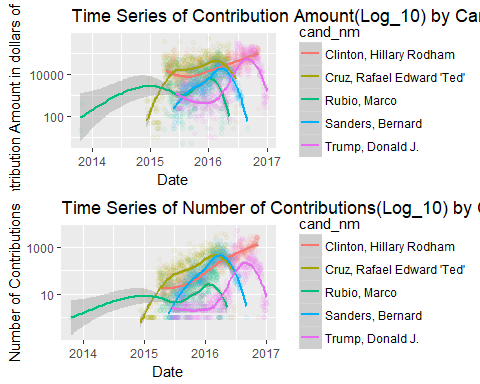
 

# Final Plots and Summaries

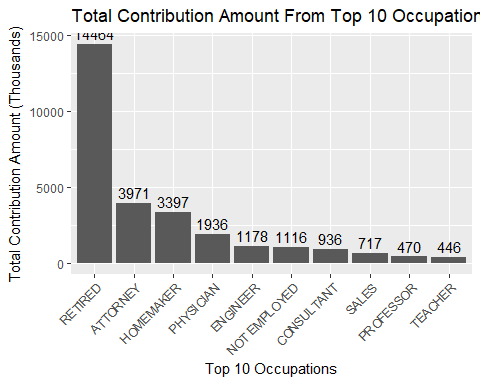
Most donations went to a few candidates. In Texas financial donations to the presidential campaign were distributed unevenly. Especially in Democrat, 86% of the donations for Democrat went to Hillary Clinton. "In the 2016 presidential election, it would appear, the question is not whether Democrats can turn Texas blue; it is whether Donald Trump will do so. ... Polling has found him underperforming in a number of red states, and it is natural that Texas should be among them."



## Time Series of Top Candidates



Hillary Clinton dominated the contribution amount and number of contributions donors.From the plots, it looks like Bernard Sanders and Donald Trump drastically reduced the contribution towards the end of elections and Hillary Clinton continued to contribute.



The top occupation that most generously contributed was 'retired'. The total contribution across occupations differed substantially.

Tecahers and Professors contributed very less (surpising was that they contributed lesser than "not employed")

# Reflection

# Challenges

The original dataset did not contain gender information, to analyze the relationship between gender and donations, I added gender column using R's gender package which used to predict gender from donor's first name. It was a challenge to learn and do R in general especially the ggplots as sometimes when proper fomatting is not done, nothing appears on the graphs.

The ggplot2 and dplyr packages are the most important packages for this project.

# Issues

Gender information was not there. Latitude and Longitude had to be added.

# Conclusion

By analyzing Texas financial donation data, I found several interesting characteristics:

It is no doubt that Texas is one of the red states although it was not much apparent in 2016 elections. Few candidates collected the most donations. We need to have the right political knowledge to know the exact situation. This analysis might prove out to be useful fo those who are abreast with the current political knowledge. Also, contribution is just one factor, there might be several other factors which influenced the elections.