Data Analysis and Visualization of the Financial Contribution of 2020's Presidential Election Of the United States

May,13,2019

Abstract and Introduction

In this report we will try to explore and analyze the current situation of financial contribution for 2020's presidential election.

Dataset Preparations

The Overview of the Dataset

```
## 'data.frame':    720297 obs. of  18 variables:
##  $ cmte_id          : Factor w/ 17 levels "C00508416","C00580100",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ cand_id          : Factor w/ 17 levels "P00006213","P00006486",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ cand_nm          : Factor w/ 17 levels "Booker, Cory A.",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ contbr_nm        : Factor w/ 192129 levels "39 NEW YORK AVE, LLC",..: 163629 163102 163102 65272
163102 163102 76096 135481 9779 186325 ...
##  $ contbr_city      : Factor w/ 12923 levels "","'CALLAHAN",..: 3927 8908 8908 11658 8908 8908 6463
12457 796 10069 ...
##  $ contbr_st        : Factor w/ 66 levels "AA","AB","AE",..: 6 9 9 9 9 9 11 11 11 11 ...
##  $ contbr_zip       : Factor w/ 93220 levels "","0","10001",..: 87975 68889 68889 70317 68889 68889
73780 76286 85162 82508 ...
##  $ contbr_employer  : Factor w/ 62854 levels "","'CALL PHONOGRAPH LLC",..: 12391 19887 19887 39455
19887 19887 39455 9427 49304 21080 ...
##  $ contbr_occupation: Factor w/ 21289 levels "","#1 NATIONAL REAL ESTATE NETWORK, LLC",..: 11359 14
010 14010 12597 14010 14010 12597 2803 11918 2735 ...
##  $ contb_receipt_amt: num  500 50 5 2700 500 250 500 2700 -150 50 ...
##  $ contb_receipt_dt : Factor w/ 980 levels "1-Apr-17","1-Apr-18",..: 806 840 669 302 636 927 696 71
2 890 870 ...
##  $ receipt_desc     : Factor w/ 98 levels "","* EARMARKED CONTRIBUTION: SEE BELOW  SEEKING REATTRIB
UTION",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_cd          : Factor w/ 2 levels "","X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_text        : Factor w/ 210 levels "","$1,012.50 REFUNDED 4/2019",..: 1 37 37 1 1 37 37 37
191 37 ...
##  $ form_tp          : Factor w/ 3 levels "SA17A","SA18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ file_num         : int  1260664 1324749 1324749 1324749 1324749 1324749 1260664 1260664 1240948
1260664 ...
##  $ tran_id          : Factor w/ 701437 levels "1000003","1000007",..: 701394 100205 100758 100064 1
00813 100047 701425 701221 701047 701407 ...
##  $ election_tp      : Factor w/ 9 levels "","G2016","G2018",..: 9 9 9 9 9 9 9 9 9 9 ...
```

The dataset consist of 720297 observations and 18 variables.

summary of the data

```
##       cmte_id            cand_id                        cand_nm
##  C00580100:605763   P80001571:605763   Trump, Donald J.      :605763
##  C00696948: 28262   P60007168: 28262   Sanders, Bernard      : 28262
##  C00694455: 20678   P00009423: 20678   Harris, Kamala D.     : 20678
##  C00695510: 15654   P00009795: 15654   Booker, Cory A.       : 15654
##  C00699090: 11592   P00010793: 11592   O扮ourke, Robert Beto: 11592
##  C00693234: 11542   P00009621: 11542   Warren, Elizabeth     : 11542
##  (Other)  : 26806   (Other)  : 26806   (Other)               : 26806
##                 contbr_nm           contbr_city       contbr_st
##  BURNHAM, MICHAEL    :   514   NEW YORK   :  9731   CA     :106938
##  DERUYTER, RICHARD   :   392   HOUSTON    :  8589   TX     : 75869
##  FRANK, MICHAELENE   :   304   LOS ANGELES:  5395   FL     : 61754
##  FARHAT, LILLIAN     :   246   DALLAS     :  4896   NY     : 43497
##  RAYUDU, NAGABHUSHANA:   228   SAN ANTONIO:  4619   NJ     : 24153
##  REDMOND, RONALD     :   210   AUSTIN     :  4508   PA     : 21843
##  (Other)             :718403   (Other)    :682559   (Other):386243
##    contbr_zip                  contbr_employer
##  99999  :  2506   RETIRED              :256135
##  4070   :   583   SELF-EMPLOYED        : 68224
##  34990  :   479   INFORMATION REQUESTED: 54156
##  85255  :   408   NOT EMPLOYED         : 23638
##  77024  :   349   ENTREPRENEUR         : 11031
##  92660  :   349   (Other)              :307006
##  (Other):715623   NA's                 :   107
##             contbr_occupation  contb_receipt_amt    contb_receipt_dt
##  RETIRED              :265594   Min.    :  -5600.0   20-Jan-17: 71297
##  INFORMATION REQUESTED: 53505   1st Qu.:     25.0   31-Mar-19: 14922
##  NOT EMPLOYED         : 27072   Median :     37.5   14-Mar-19:  9102
##  PHYSICIAN            : 11831   Mean   :    126.4   8-Dec-16 :  7324
##  ENTREPRENEUR         : 11674   3rd Qu.:     80.0   30-Mar-19:  7082
##  (Other)              :350593   Max.   :2721782.1   15-Dec-16:  6401
##  NA's                 :    28                       (Other)  :604169
##                                              receipt_desc
##                                                     :719536
##  Refund                                             :   572
##  * REDESIGNATED                                     :    32
##  * REDESIGNATION                                    :    32
##  * EARMARKED CONTRIBUTION: SEE BELOW REDESIGNATED BELOW:    28
##  * EARMARKED CONTRIBUTION: SEE BELOW REATTRIBUTED BELOW:     2
##  (Other)                                            :    95
##  memo_cd
##   :264127
##  X:456170
##
##
##
##
##
##                                                        memo_text
##                                                             :637229
##  * EARMARKED CONTRIBUTION: SEE BELOW                        : 80720
##  * EARMARKED CONTRIBUTION THROUGH ACTBLUE ON 03/31/2019     :   273
##  * EARMARKED CONTRIBUTION THROUGH ACTBLUE ON 01/13/2019     :   133
##  NOTE: ABOVE CONTRIBUTION EARMARKED THROUGH THIS ORGANIZATION.:   127
##  * EARMARKED CONTRIBUTION THROUGH ACTBLUE ON 02/28/2019     :    97
##  (Other)                                                    :  1718
##   form_tp          file_num              tran_id        election_tp
```

```
##  SA17A:276151   Min.    :1174081   SA17A.11532:     4   P2020  :639084
##  SA18 :443574   1st Qu.:1193597   SA17A.11602:     4   G2016  : 73022
##  SB28A:   572   Median :1301594   SA17A.11604:     4   G2020  :  8128
##                 Mean   :1269311   SA17A.11611:     4   O2018  :    21
##                 3rd Qu.:1326131   SA17A.11612:     4          :    17
##                 Max.   :1326558   SA17A.11630:     4   P2016  :    12
##                                   (Other)    :720273   (Other):    13
```

In light of the summary above, Donald Trump get the most number of contributions and most contributors are from New York city, while California has the most contributors in all states. There are somethings wrong for the zip codes, 99999 and 4070 are evidently incorret. Most contribution amount are between 25 dollars and 80 dollars, and the average number is higher than this interval, 126.4 dollars. Interestingly, the most common occupation in the contributors is “RETIRED”.

This dataset has not provided the information about the candidiates’ gender and party, so we try to complement this part.

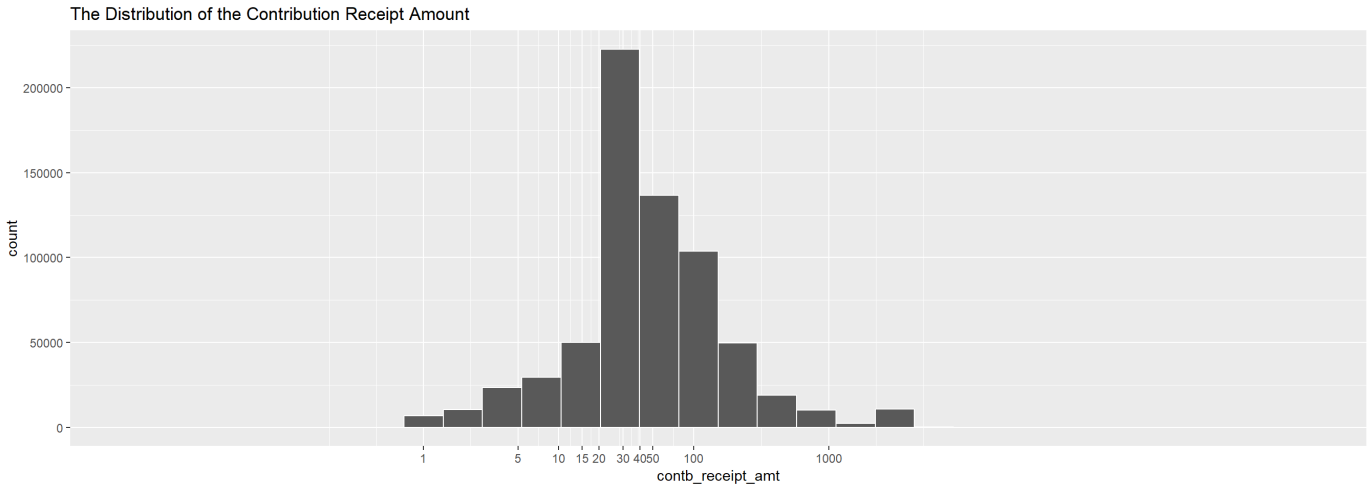
```
##
##         Booker, Cory A.          Buttigieg, Pete          Castro, Julian
##                   15654                     5340                    1158
##        Delaney, John K.           Gabbard, Tulsi    Gillibrand, Kirsten
##                    1492                     2070                    2568
##       Harris, Kamala D.    Hickenlooper, John W.           Inslee, Jay R
##                   20678                     1338                    1726
##       Klobuchar, Amy J. Ojeda, Richard Neece II   O拎ourke, Robert Beto
##                    5015                       47                   11592
##        Sanders, Bernard         Trump, Donald J.      Warren, Elizabeth
##                   28262                   605763                   11542
##   Williamson, Marianne              Yang, Andrew
##                    1611                     4441
```

```
## 'data.frame':    720297 obs. of  20 variables:
##  $ cmte_id          : Factor w/ 17 levels "C00508416","C00580100",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ cand_id          : Factor w/ 17 levels "P00006213","P00006486",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ cand_nm          : Factor w/ 17 levels "Booker, Cory A.",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ contbr_nm        : Factor w/ 192129 levels "39 NEW YORK AVE, LLC",..: 163629 163102 163102 65272
163102 163102 76096 135481 9779 186325 ...
##  $ contbr_city      : Factor w/ 12923 levels "","'CALLAHAN",..: 3927 8908 8908 11658 8908 8908 6463
12457 796 10069 ...
##  $ contbr_st        : Factor w/ 66 levels "AA","AB","AE",..: 6 9 9 9 9 9 11 11 11 11 ...
##  $ contbr_zip       : Factor w/ 93220 levels "","0","10001",..: 87975 68889 68889 70317 68889 68889
73780 76286 85162 82508 ...
##  $ contbr_employer  : Factor w/ 62854 levels "","'CALL PHONOGRAPH LLC",..: 12391 19887 19887 39455
19887 19887 39455 9427 49304 21080 ...
##  $ contbr_occupation: Factor w/ 21289 levels "","#1 NATIONAL REAL ESTATE NETWORK, LLC",..: 11359 14
010 14010 12597 14010 14010 12597 2803 11918 2735 ...
##  $ contb_receipt_amt: num  500 50 5 2700 500 250 500 2700 -150 50 ...
##  $ contb_receipt_dt : Factor w/ 980 levels "1-Apr-17","1-Apr-18",..: 806 840 669 302 636 927 696 71
2 890 870 ...
##  $ receipt_desc     : Factor w/ 98 levels "","* EARMARKED CONTRIBUTION: SEE BELOW  SEEKING REATTRIB
UTION",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_cd          : Factor w/ 2 levels "","X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_text        : Factor w/ 210 levels "","$1,012.50 REFUNDED 4/2019",..: 1 37 37 1 1 37 37 37
191 37 ...
##  $ form_tp          : Factor w/ 3 levels "SA17A","SA18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ file_num         : int  1260664 1324749 1324749 1324749 1324749 1324749 1260664 1260664 1240948
1260664 ...
##  $ tran_id          : Factor w/ 701437 levels "1000003","1000007",..: 701394 100205 100758 100064 1
00813 100047 701425 701221 701047 701407 ...
##  $ election_tp      : Factor w/ 9 levels "","G2016","G2018",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ cand_gender      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ cand_party       : Factor w/ 3 levels "Democrats","Independent",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Analysis and Visualization

Univariate Analysis

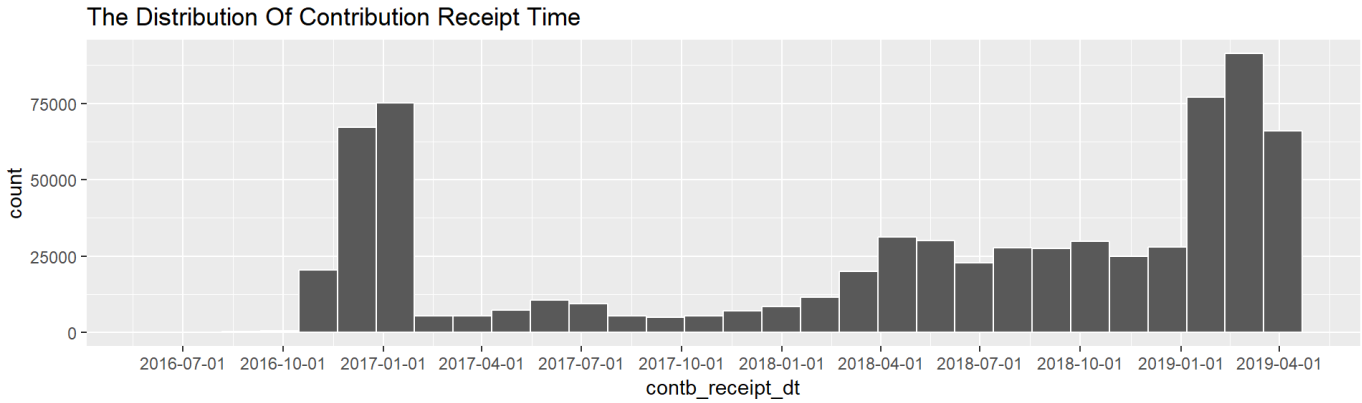
The Distribution of Contribution Amount



After adjusting the x-axis most of data fall in the interval between 1 and 1000. This distribution is a bit left skewed. We can see the most contibutions are between 20 and 40, the second most on count are between 40 dollars to about 100 dollars.

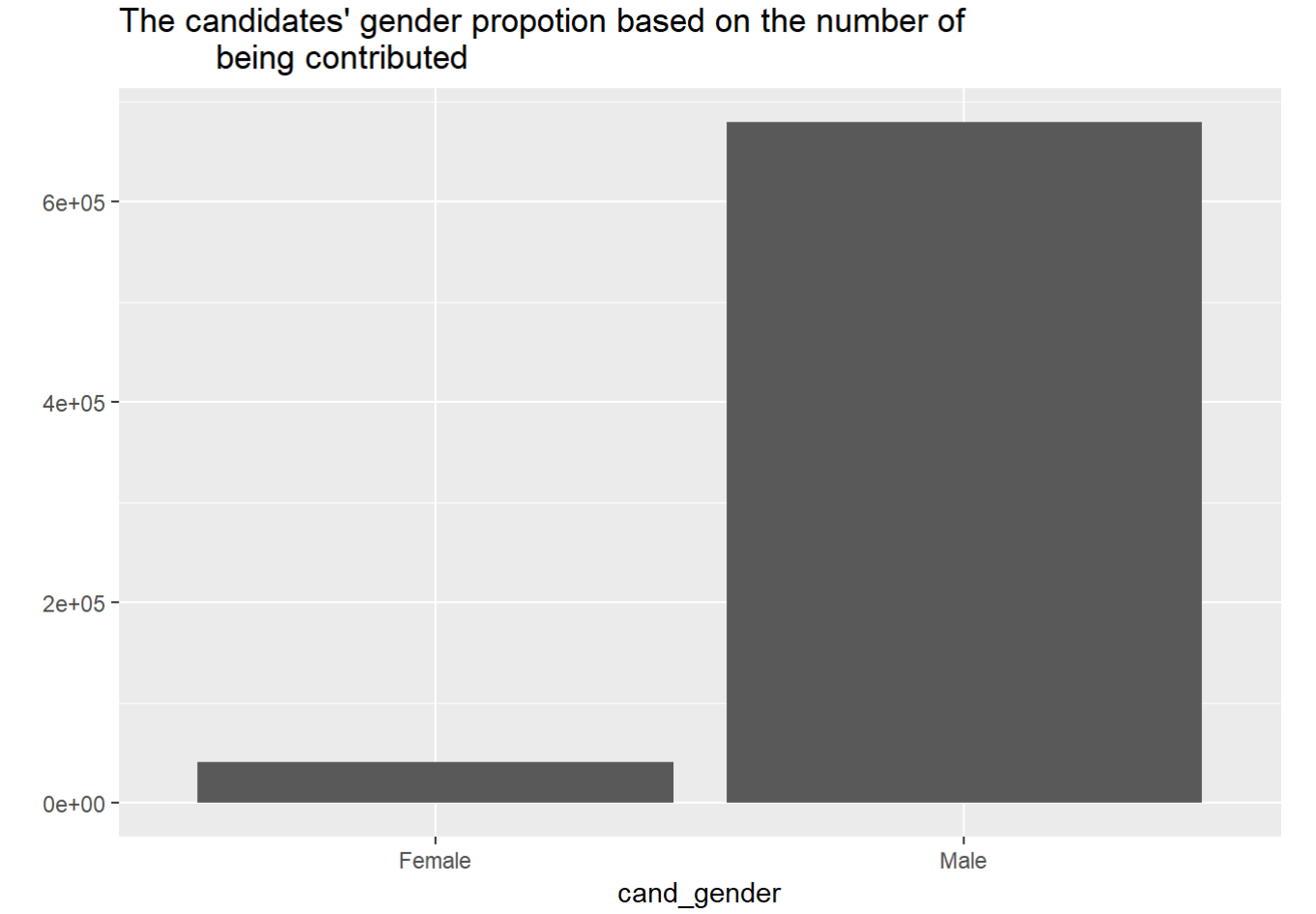
The distribution of the Contribution time

```
## [1] "C"
```



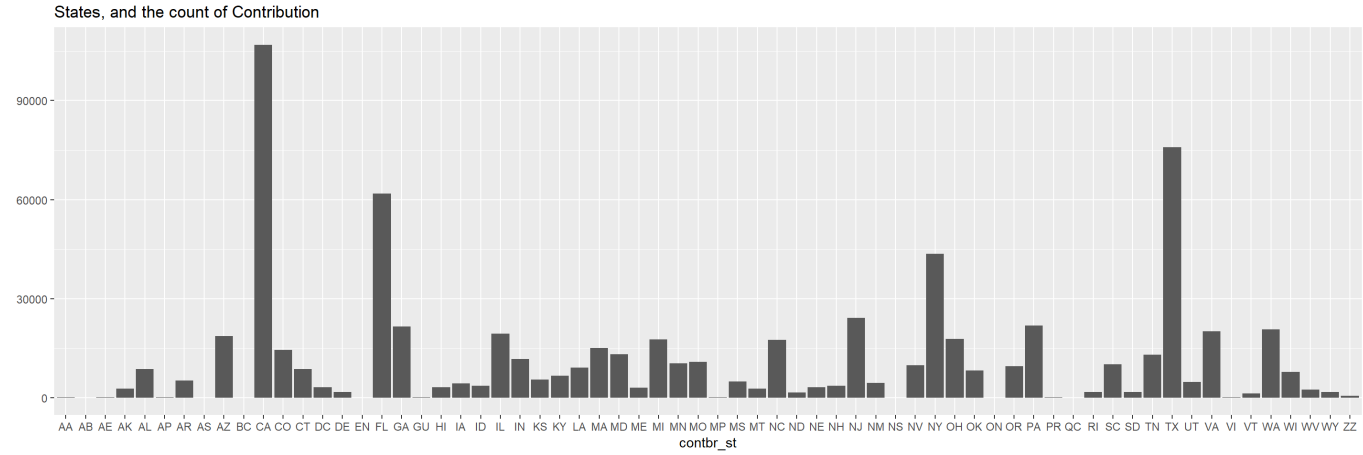
From the figure above we can observe two peaks, one is at the end of 2016 and the beginning of 2017, after that the number of financial contributions fell back to a relatively low level but still increased slowly hereafter. The other peak is at the recent three months the number of contributions have risen to the highest position since the end of 2016.

The candidates' gender propotion based on the number of being contributed



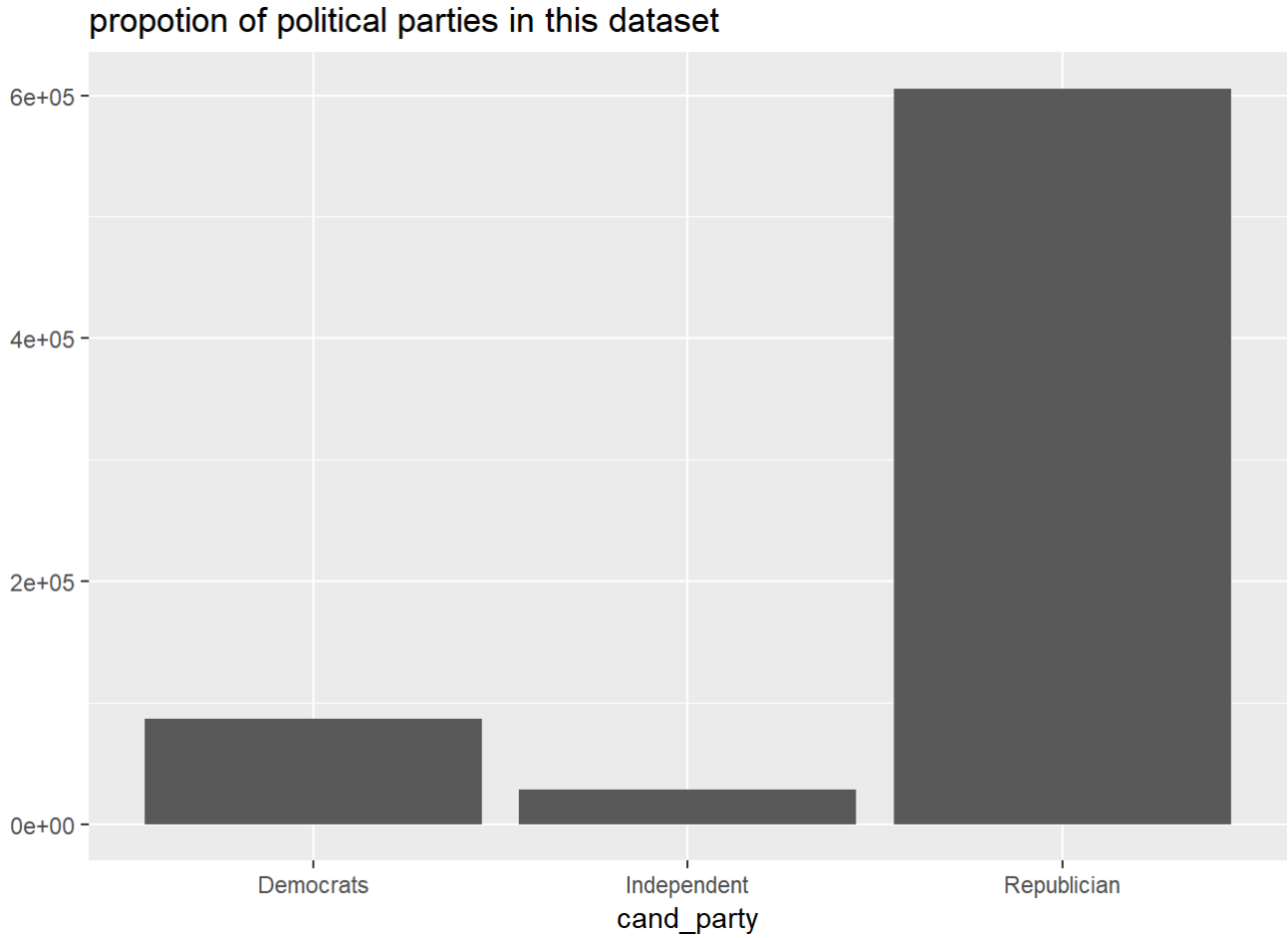
Most of contribution records' candidates are male.

Which state has the most contribution records?



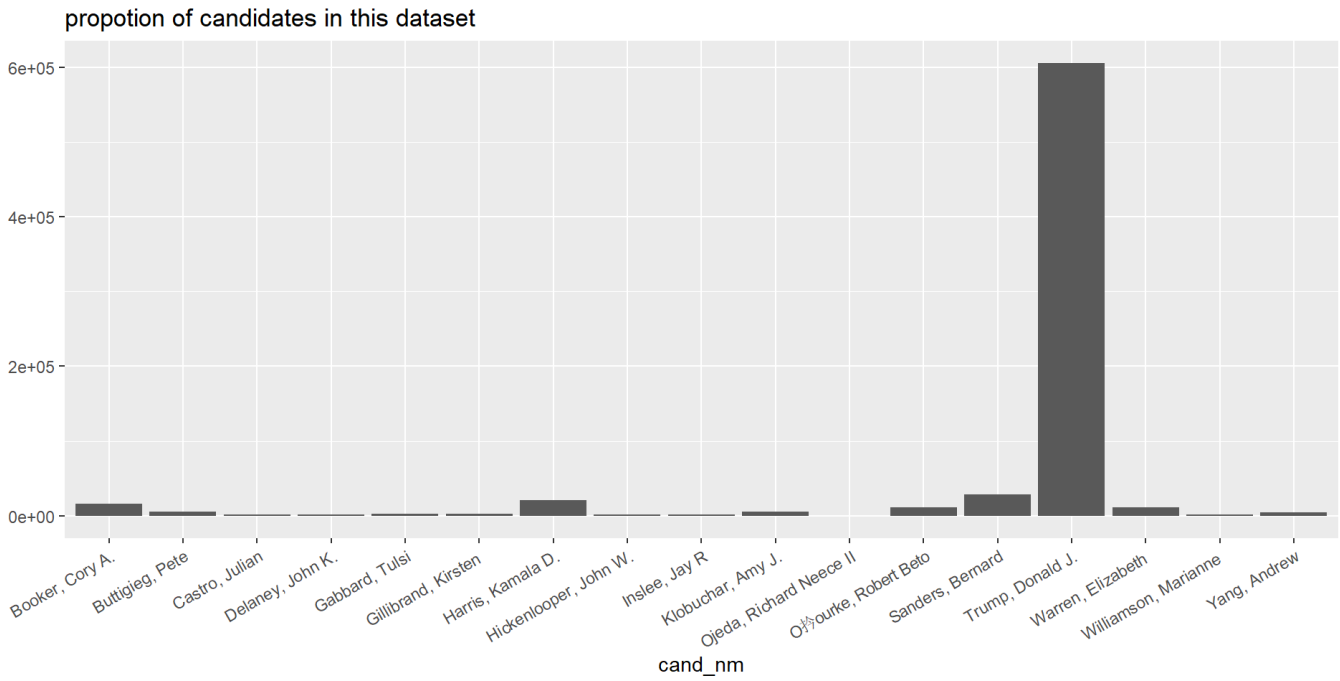
California has the most contribution records in this dataset, followed by Texas, Florida, New York.

propotion of political parties in this dataset



Obvious that most of the contributions are to republician in this dataset. Meanwhile the contribution to independent is the least among three parties.

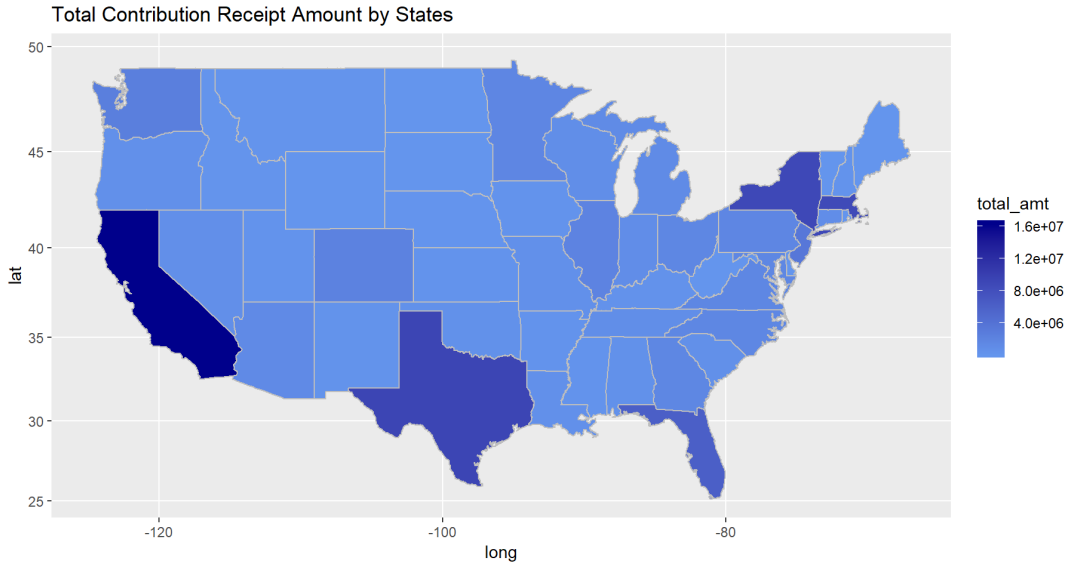
propotion of candidates in this dataset



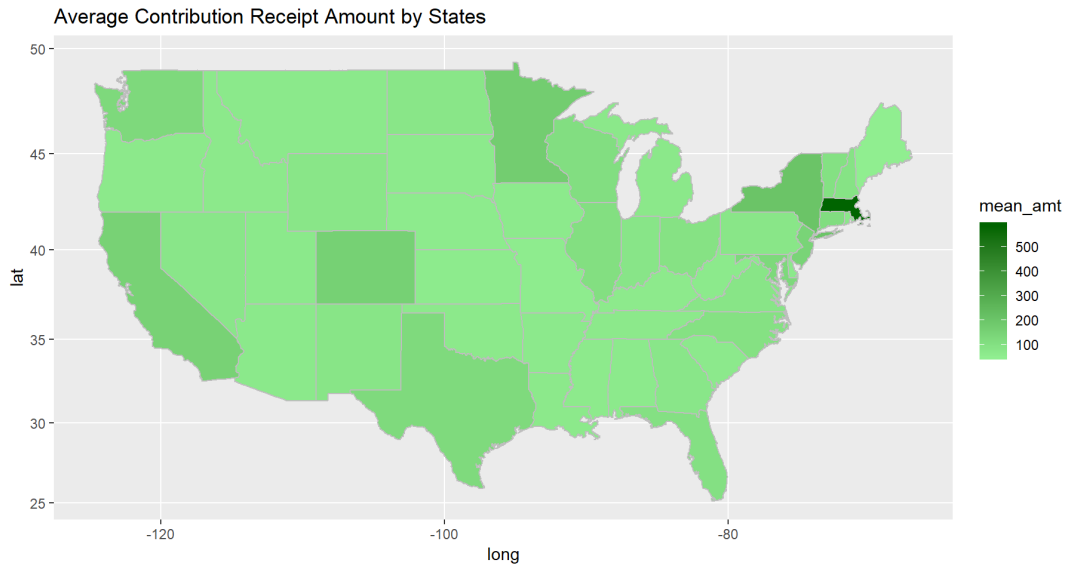
In this dataset only Donald Trump belongs to Republician Party. It is consistent with the data of the count of the contributions to different political parties, the count of the contributions to Trump is much bigger than all other's sum.

Double variables Analysis and Visualization

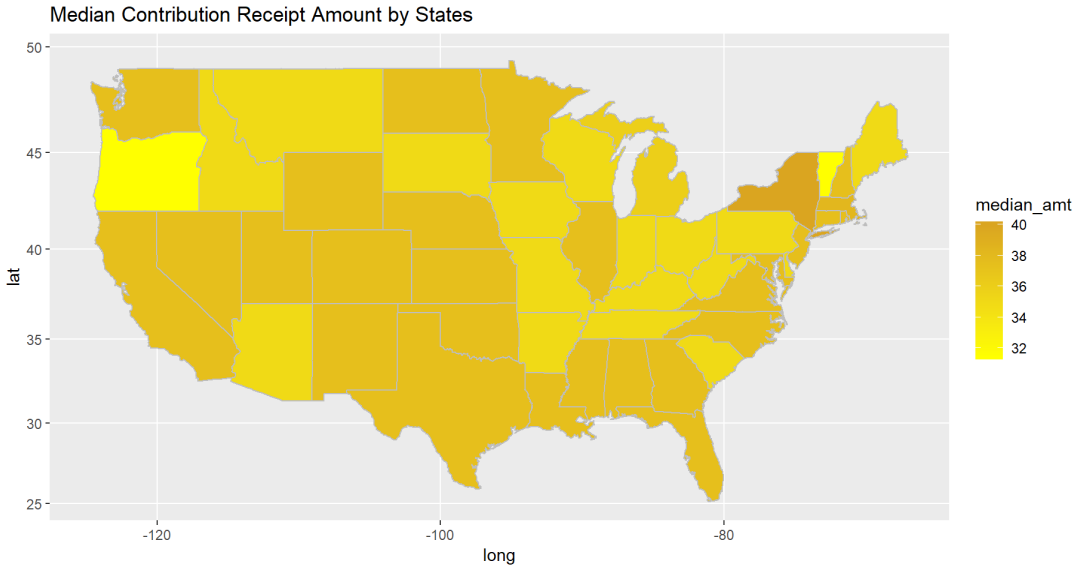
The Total, Average And Median of Contribution Receipt Amount In Every State.



Based on the total number, California contributes the most of fund in all states, the other states with realatively darker color are Texas, New York Massachusetts and Florida.

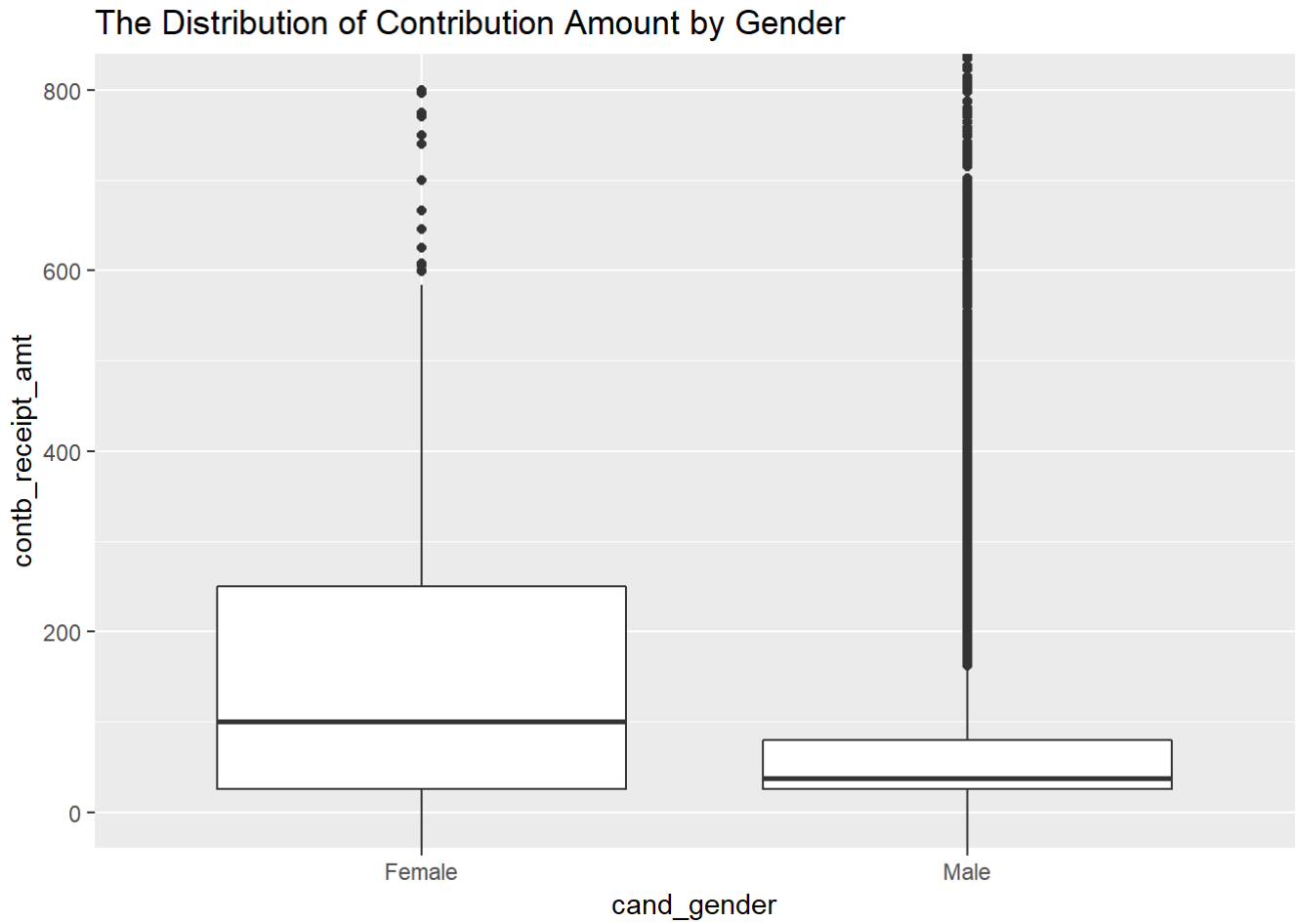


Averagely, the contribution amount in Massachusetts is the highest in all states, other states like California, Washington, Colorado, Minnesota, and New York also worth notice.



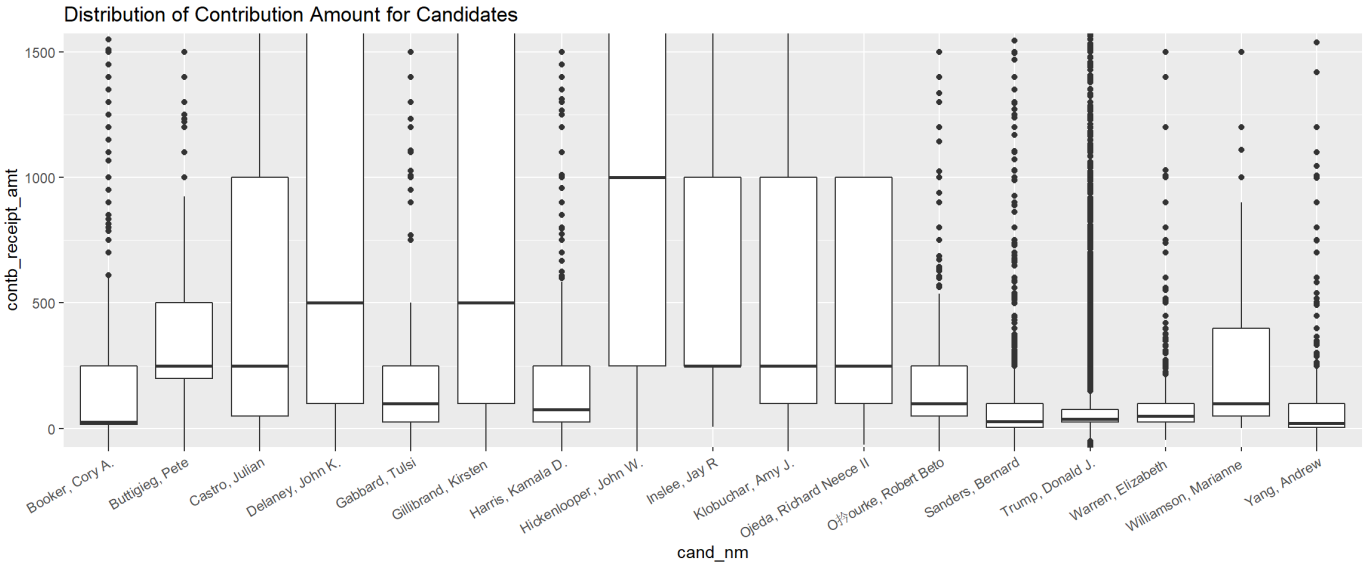
The median contribution amount of every state as above.

Any difference Between Male And Female Candidate On The Contribution Receipt Amount?



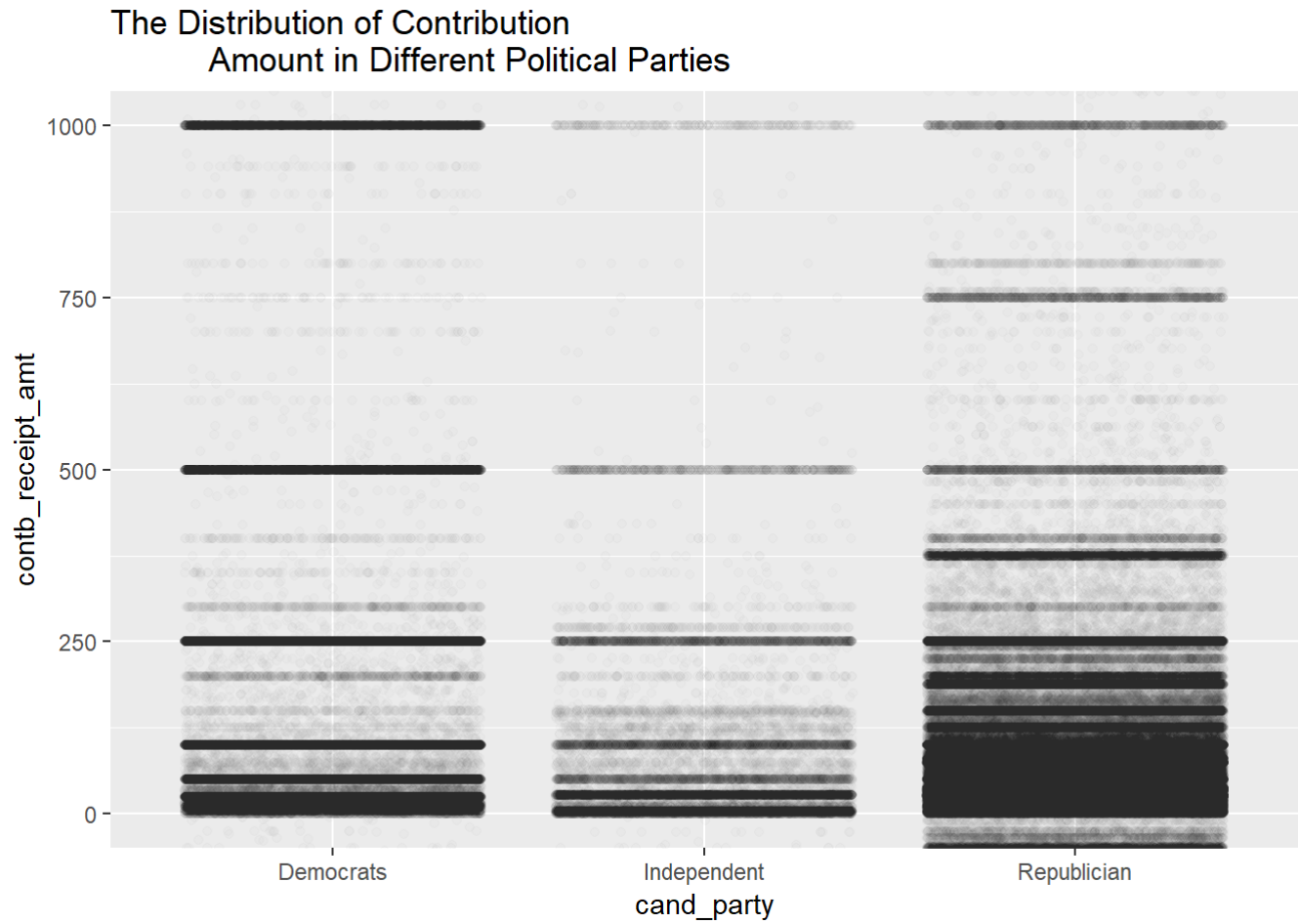
Seems the female candidates get more contributions than the male candidates in general, but there are many outliers in the male.

Distribution of Contribution Amount for Candidates



Some candidates always got large contribution, like Delaney, John K.,Gillbrand,Kristen and Hickenlooper,John W., Although Donald Trump has the most contribution, but most are small-amount.

Any difference Between Political Parties On Contribution Receipt Amount?



I wonder if there is any difference between political parties. From the first figure we can't see the difference between three groups, they all look at the same level, so i adjust the view to zoom in and we get the second figure. it shows the repulician get more fund than democrats and independents, if we don't take the outliers into account, just focus on the main part.

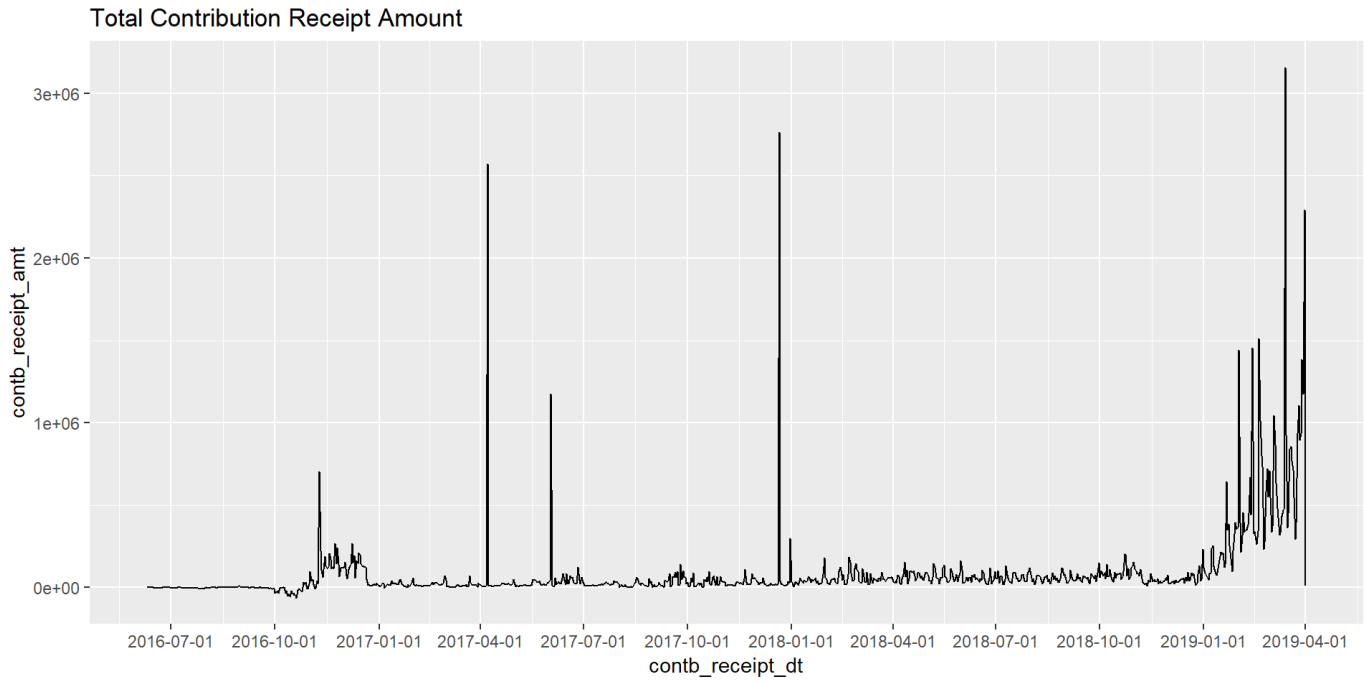
Are There Any Difference On Contributors' Occupations Between Different Political Parties?

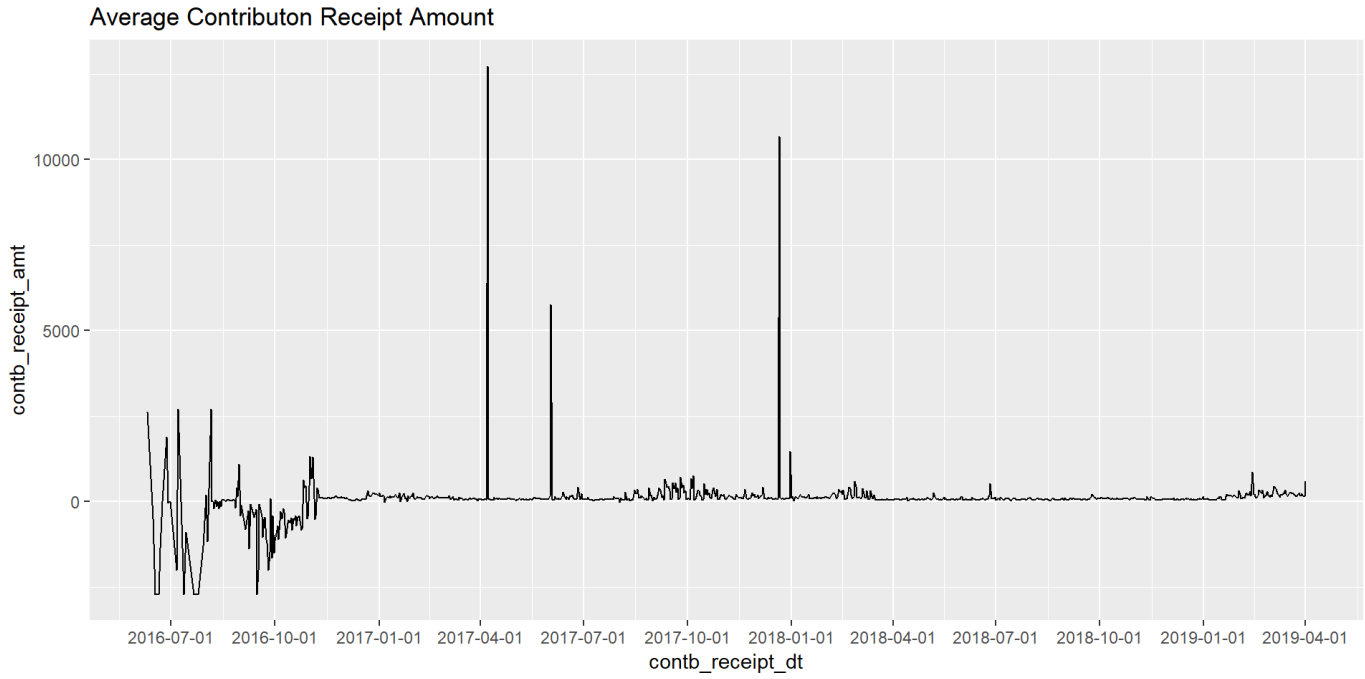
The most common occupation is Retired, or Not employed, for both republician and democrats. if we just see those employed:

The common occupations of contributors to republician:

entrepreneur,homemaker, contractor, business owner, physician, teacher, attorney…
The common occupations of contributors to democrats:
attorney, consultant, professor, physician, software engineer,writer….
The gap is not so clear as expected before. I think other factor like the general political leanings in different states will have an influence.

How Does The Contribution Receipt Amount Change With Time?

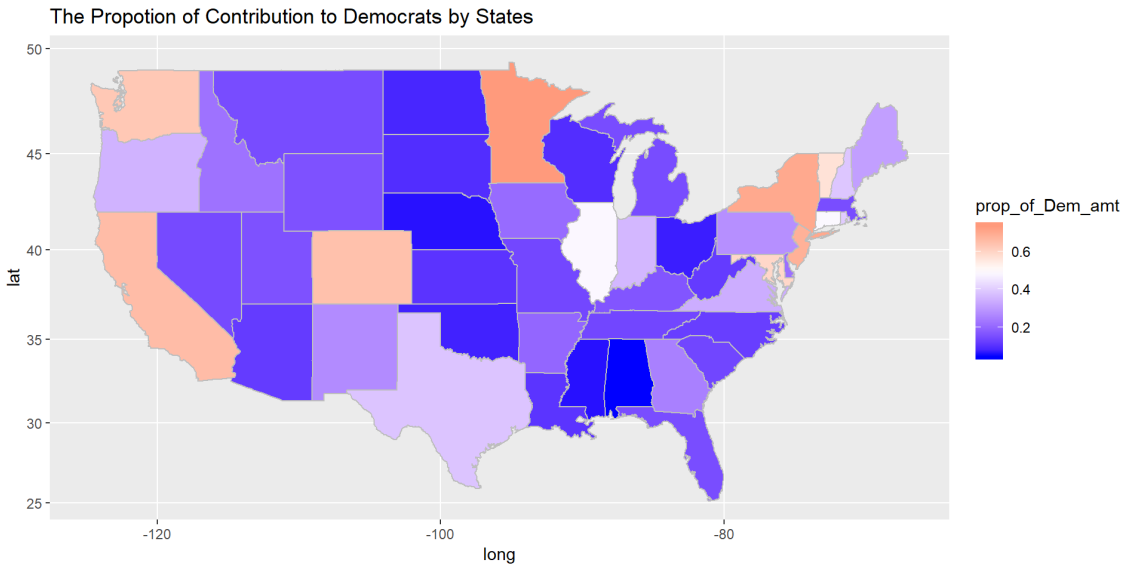




The second figure above is very simillar to the histogram of contribution time distribution, but from the third picture we find the average amount in recent three month is at a low level, this indicates that the big leap in the recent three months is mainly attributed to the increasing of the amount of contributors, not the average contribution amount.

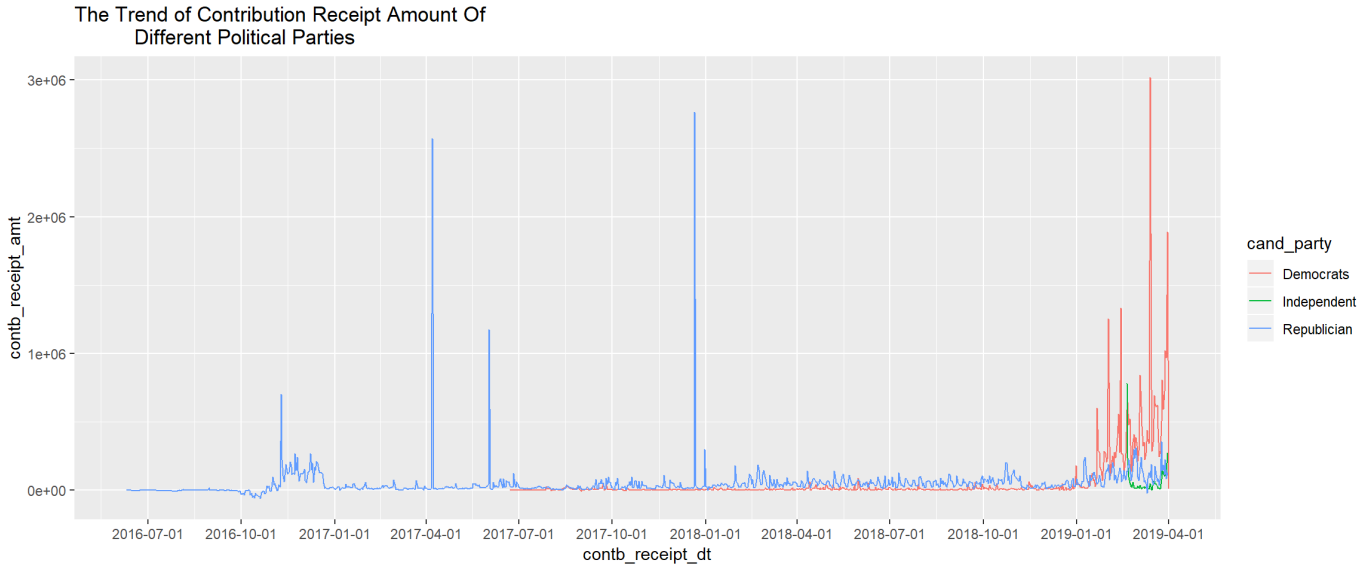
Multivariables Analysis

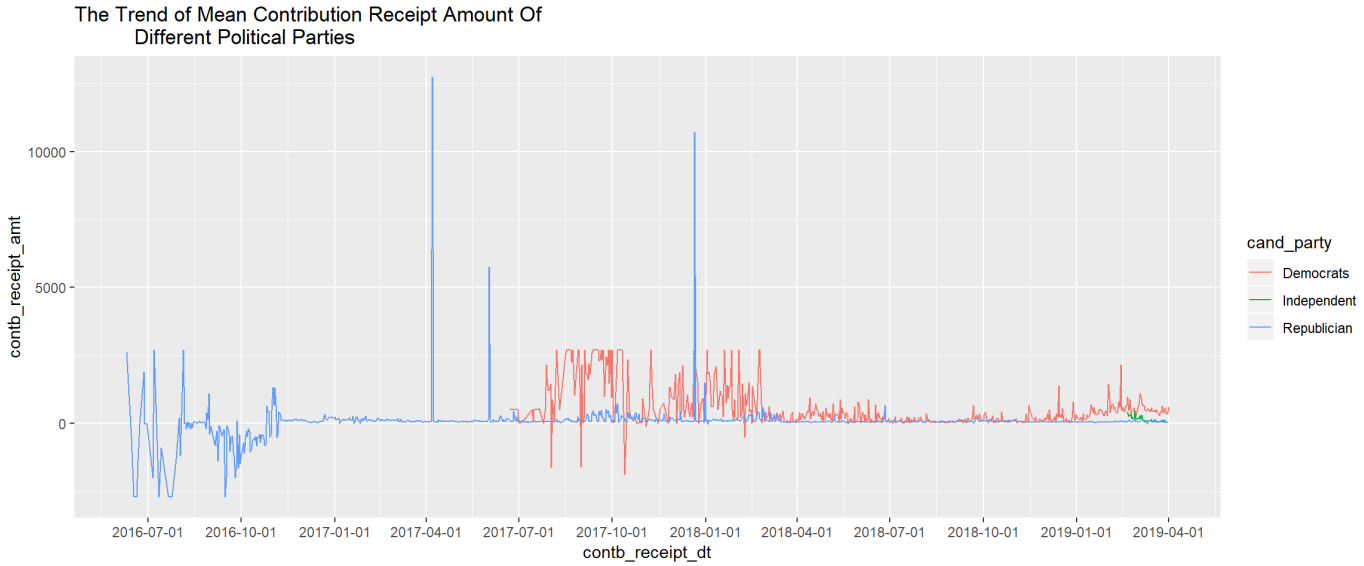
What Is The Political Leaning In Different States?



I try to explore the current political leanings in different states through this dataset, my method is to calculate the propotion of those contribution to Democrats in the total contributions. The result is displyed above, the blue color indicates that propotion of contribution amount to democats is less tha 50%, while the red color means more than 50%.

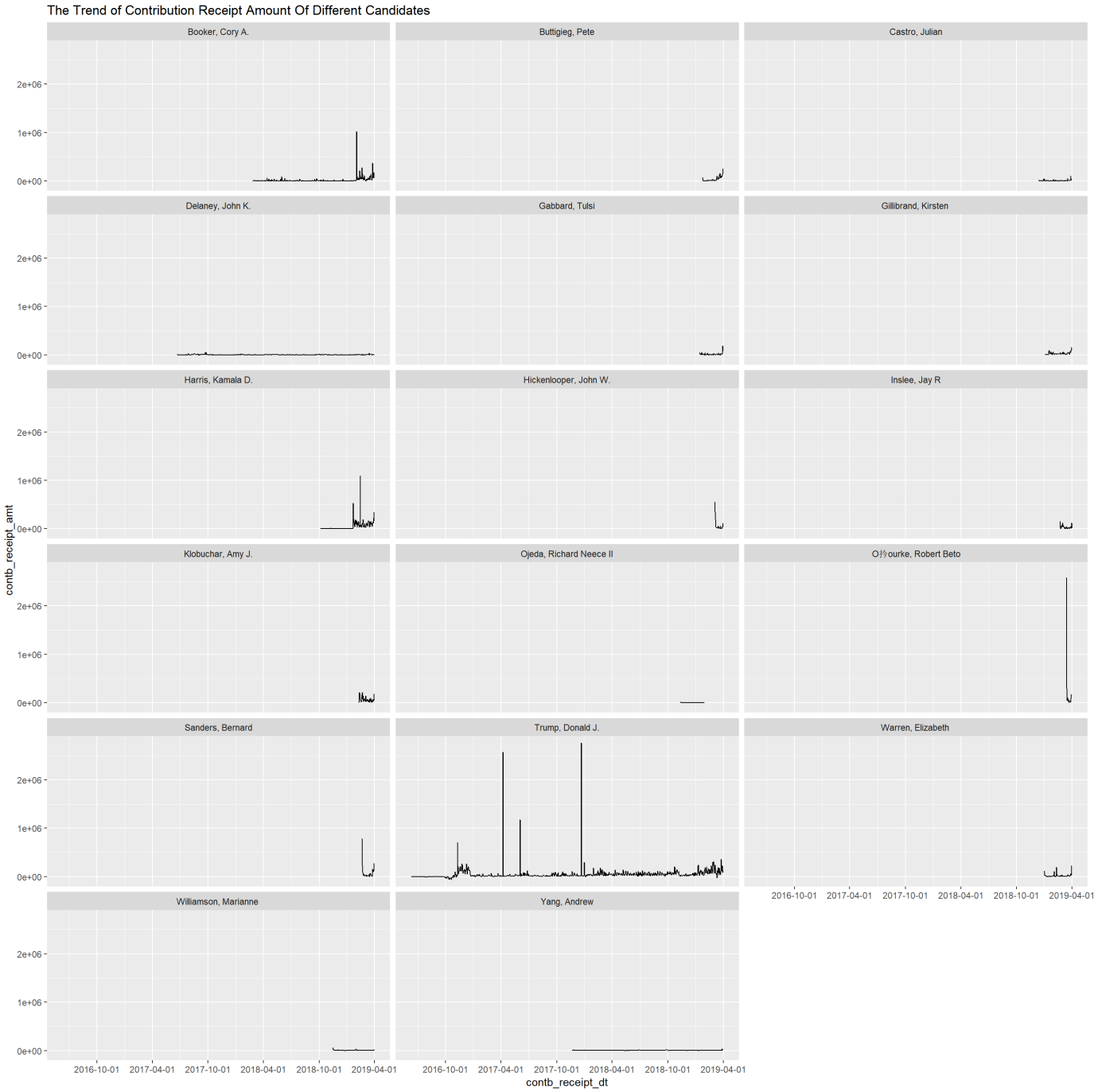
How Do The Different Political Parties' Contribution Receipt Amount Change With Time?

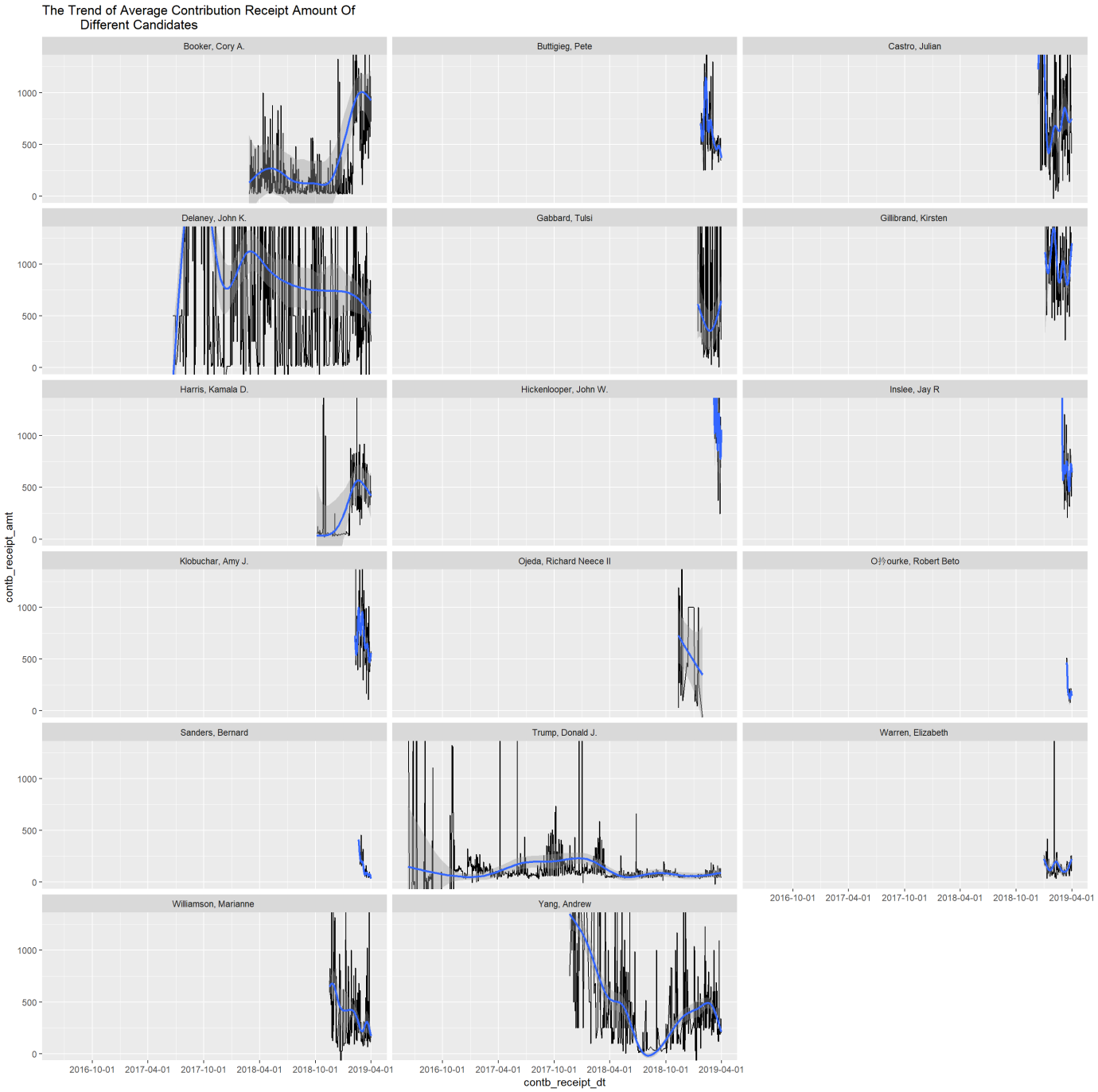


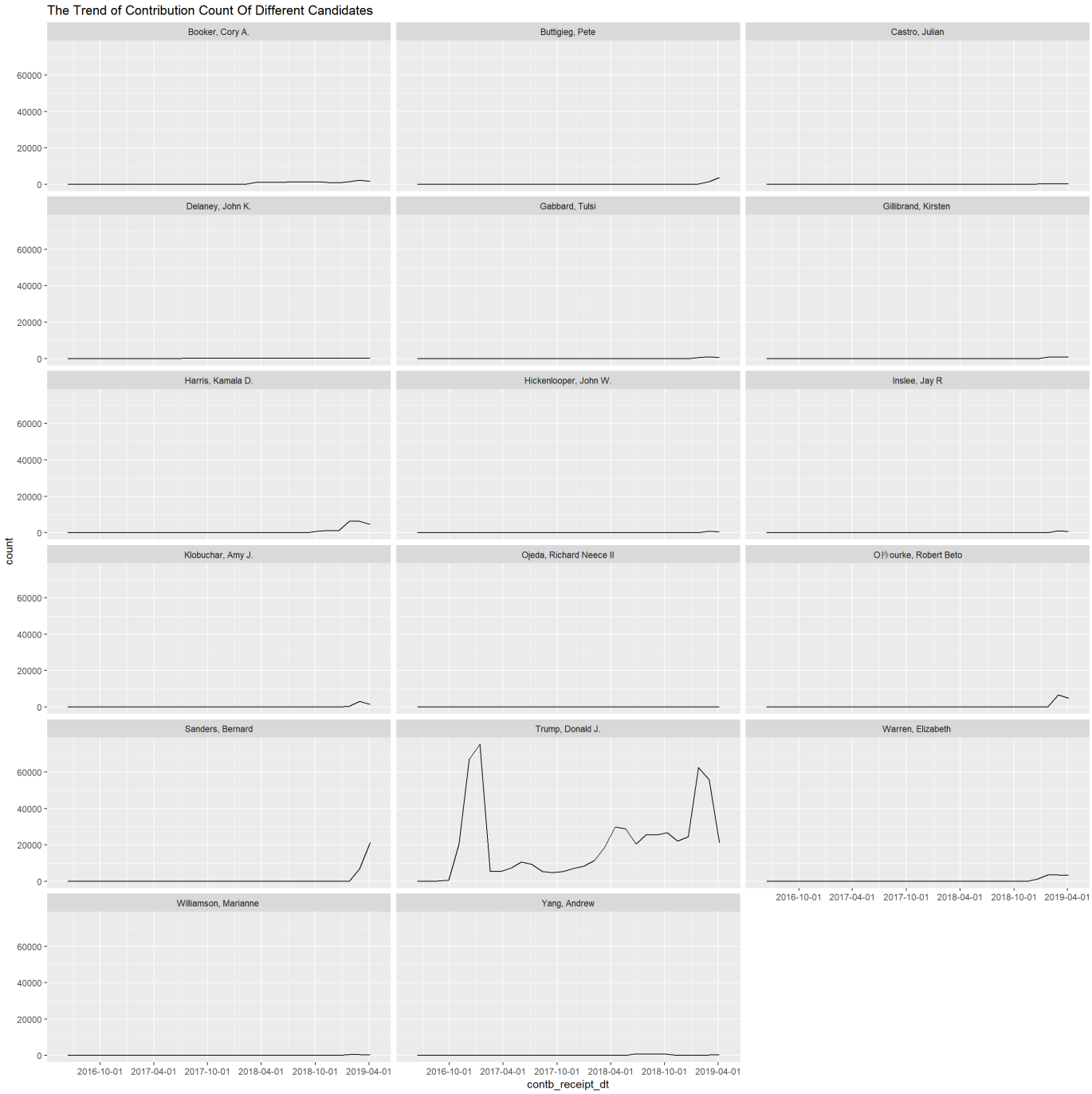


We have analyzed the increase in recent 3 months, and now from the figure displayed above, this looks like it is a Democrats' theme, high total, but relatively low mean value.The contributions to democrats take a jumping growth recently, while those to republician and independents remian "ordinary". I doubt whether this phenomenon is mainly from California, since they have very simillar features on the data: mainly for democrats, big total, but the average is not very high.

The Trend of Contribution Receipt Amount Of Different Candidates

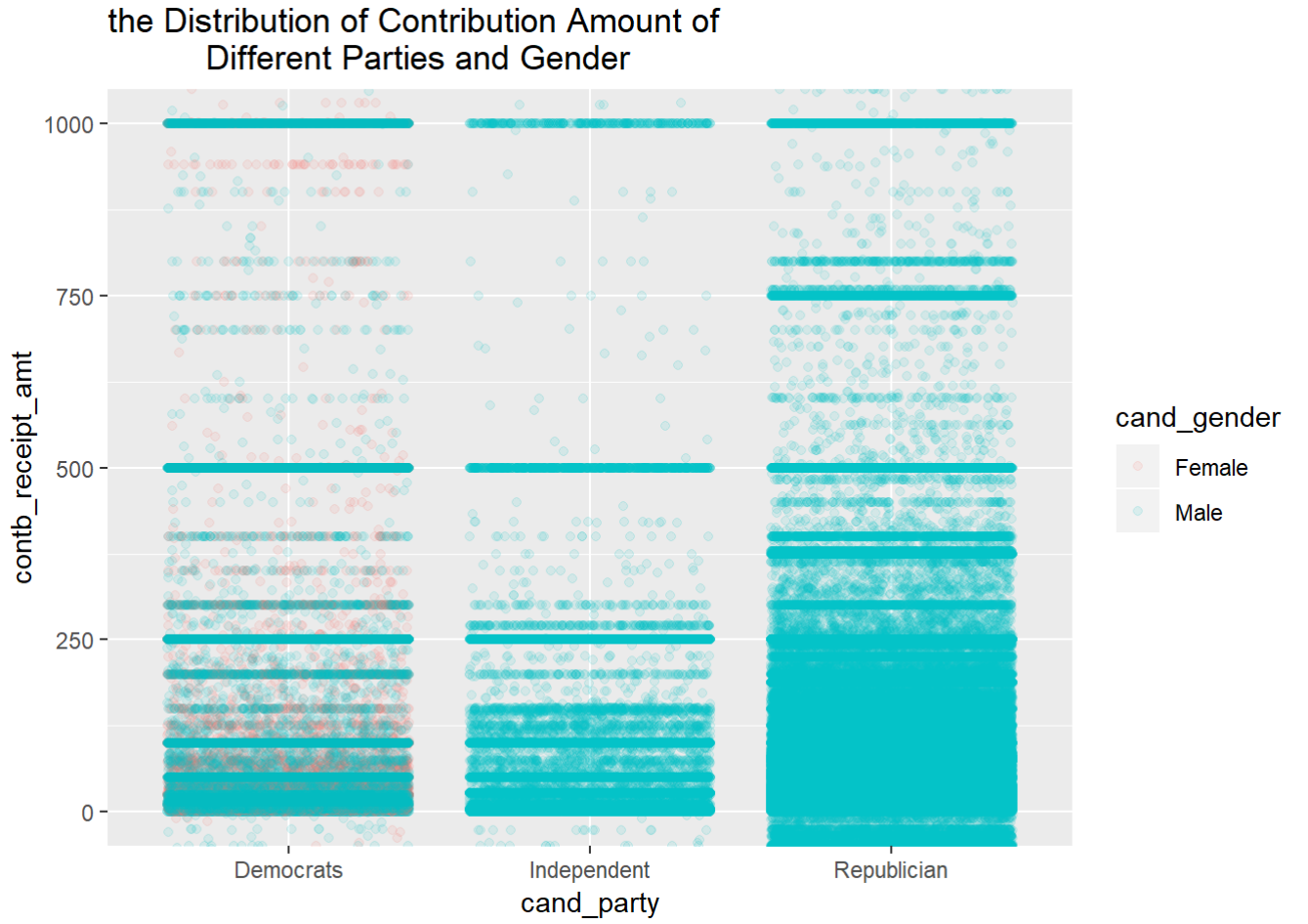

The Trend of Contribution Receipt Amount Of Different Candidates
Booker, Cory A.
Buttigieg, Pete
Castro, Julian
Delaney, John K.
Gabbard, Tulsi
Gillibrand, Kirsten
Harris, Kamala D.
Hickenlooper, John W.
Inslee, Jay R
Klobuchar, Amy J.
Ojeda, Richard Neece II
O'ourke, Robert Beto
Sanders, Bernard
Trump, Donald J.
Warren, Elizabeth
Williamson, Marianne
Yang, Andrew
contb_receipt_amt
0e+00
1e+06
2e+06
2016-10-01
2017-04-01
2017-10-01
2018-04-01
2018-10-01
2019-04-01
contb_receipt_dt


The Trend of Average Contribution Receipt Amount Of Different Candidates
Booker, Cory A.
Buttigieg, Pete
Castro, Julian
Delaney, John K.
Gabbard, Tulsi
Gillibrand, Kirsten
Harris, Kamala D.
Hickenlooper, John W.
Inslee, Jay R
Klobuchar, Amy J.
Ojeda, Richard Neece II
O'Rourke, Robert Beto
Sanders, Bernard
Trump, Donald J.
Warren, Elizabeth
Williamson, Marianne
Yang, Andrew
contb_receipt_amt
0
500
1000
2016-10-01
2017-04-01
2017-10-01
2018-04-01
2018-10-01
2019-04-01
contb_receipt_dt



We notice some features about the contribution to different parties and candidates from above 3 figures.
The contributions for Trump is keeping steady generally from end of 2016, although sometimes he will get a bit amount of contribution.
The contributions for Democrats candidates are mostly occured at relatively recent time since their announcement for 2020 US presidential campaign, but the features for these contributions are distinctive: mainly occured recently, averagely high(compared to those for republician), and numbers is currently small.

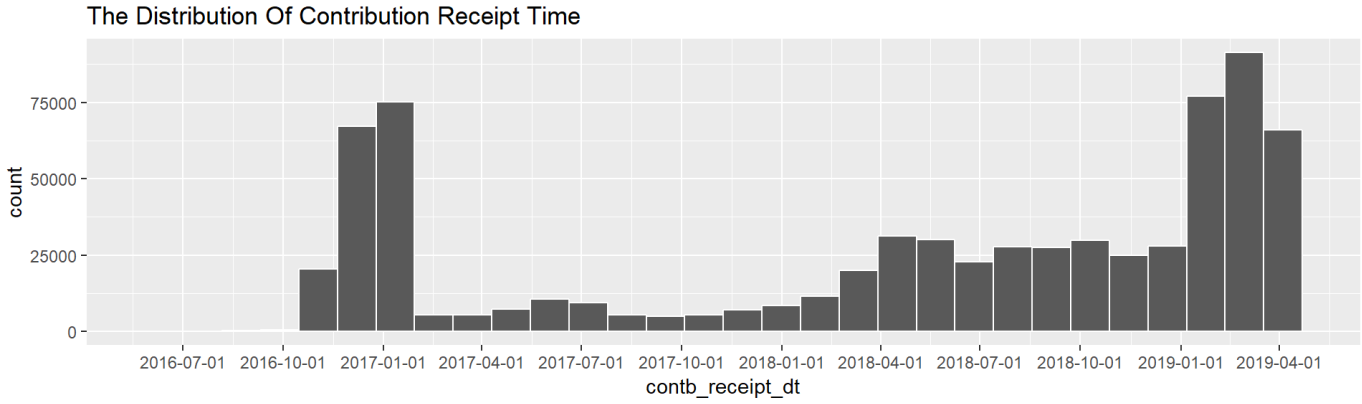
Does Gender Really Have An Influence On Contribution Receipt Amount?



try to explore whether gender will influence the contributions, but the result show the correlation between contribution and gender is very vague, from the figure we can only tell the contribution amount to female candidates are more are more discrete than those to male candidates. And the amount for male candidates is much bigger, and more assemble at the relative lower level, which make the male candidate's contribution amount lower than female in the boxplot.

Final Plot

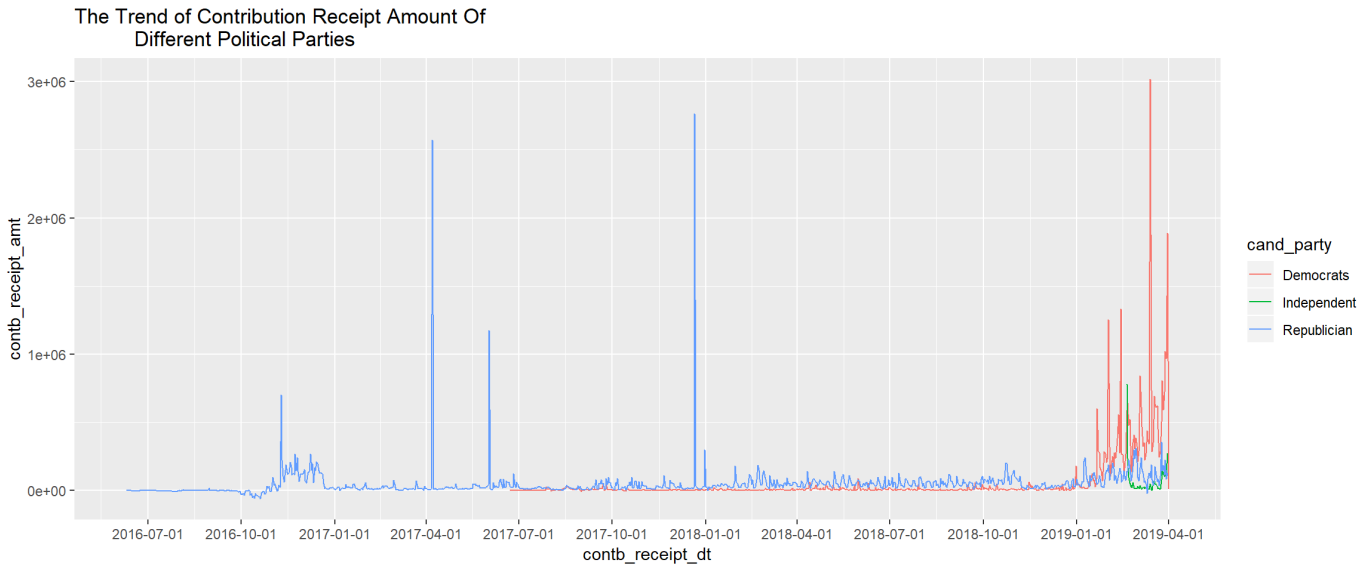
Plot One



Description One

I choose this figure as one of the final plot since it may imply the trend of the Contribution Changing. From this graph we can observe the in the last 3 months the number of contribution has risen a lot, as 2020 is coming, this situation may continue.

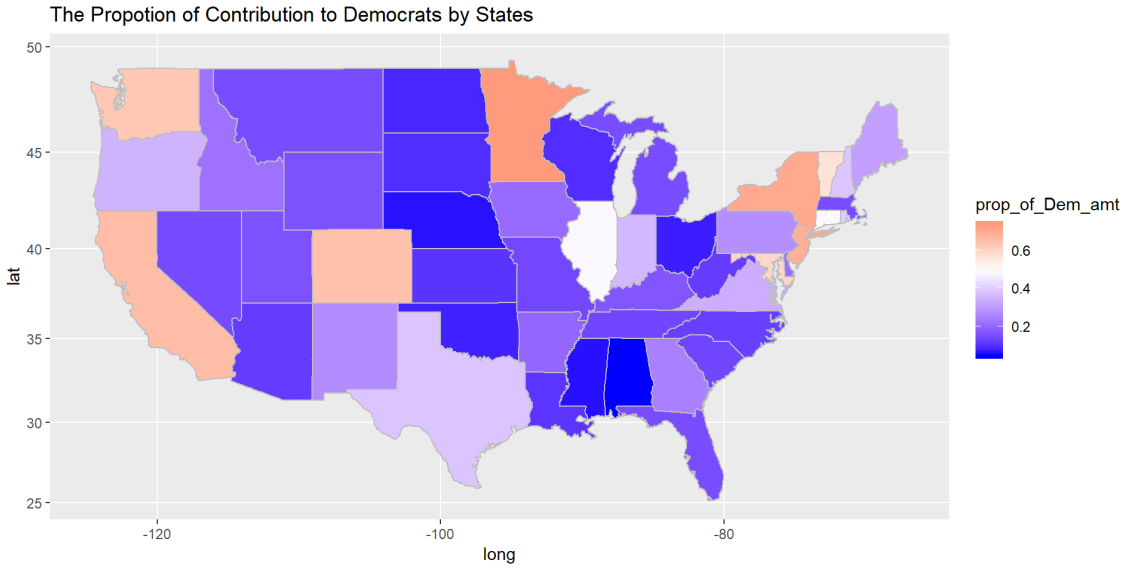
Plot Two



Description Two

This figure is more detailed than the plot one. We can read from this figure is the drastic increase on the contribution receipt amount for democrats, this is an important information we should notice when we talk about the prospect of politic.

Plot Three



Description Three

As the common sense, people always make contributions to those political parties or candidate they back, so it is a way to explore people's general political leanings through the financial contribution data. This image implies the information and the extent about the political leanings in different states.

Reflections

The Difficulties in this project:
For I was not famillar with R at first, i came across many techniques problem with R, that's a really tough process for me. In this project, I spent most of time in looking and solving the techniques problem.
What I achieved:

Learning how to code in R, how to handle the difficulties,obtain the skill to draw more complicated graph, improve the thoughts for data analysis.

The limitations and prosepct of this analysis.

Although we have a very big dataset with more than 720,000 obeservations, the limitations still exist. Our dataset can reflect the present situation, but the future is uncertain, for example, we note republician dominate currently, but democrats may reverse it in the future. So the further observation is needed. Besides, the political contribution is one way to reflect one’s political stance, but to what extent it can refelct one’s political stance still wait for research, for instance, we can explore the correlation between votes and contribution for this issue.