# Financial Contribution Analysis by Shawar Nawaz

This dataset represents the most important candidates that ran for the 2016 U.S Presidential elections and the contributions that they received in the state of Florida. I decided to use the state of Florida as it has been considered a major swing state in the elections consisting of 29 electoral votes. So, it is obvious to see why the candidates seem to fight so hard for this state.

The dataset not only contains the contribution amounts, but also the city, the job title, the organization they worked for, their zipcodes and the date of the contributions as well which leaves us with a lot of data to help us analyze, who tended to contribute to who based on various features.

The data has been taken from the Federal Election Comittee website.

```
## [1] "C:/Users/shawar/Desktop/DAND P4"
```

# Univariate Plots Section

# Univariate Analysis

## What is the structure of your dataset?

```
## 'data.frame':    426057 obs. of  19 variables:
##  $ row_num          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ cmte_id          : Factor w/ 25 levels "C00458844","C00500587",..: 15 15 6 7 6 15 15 15 15
6 ...
##  $ cand_id          : Factor w/ 25 levels "P00003392","P20002671",..: 23 23 1 12 1 23 23 23 2
3 1 ...
##  $ cand_nm          : Factor w/ 25 levels "Bush, Jeb","Carson, Benjamin S.",..: 23 23 4 20 4
 23 23 23 23 4 ...
##  $ contbr_nm        : Factor w/ 113981 levels "'CALL, ELIZABETH",..: 92995 93000 91255 59125
 17739 82858 82865 87945 82876 73603 ...
##  $ contbr_city      : Factor w/ 1869 levels "","'CALLAHAN",..: 264 1749 1325 601 1673 145 497
362 1578 1839 ...
##  $ contbr_st        : Factor w/ 1 level "FL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contbr_zip       : num  3.38e+04 3.30e+04 3.33e+08 3.20e+08 3.23e+08 ...
##  $ contbr_employer  : Factor w/ 30092 levels "","'SHELL LUMBER",..: 22307 21450 26352 19056 2
5862 20393 22307 13441 20479 18656 ...
##  $ contbr_occupation: Factor w/ 12516 levels "","-","--"," RETIRED AGRICULTURE INSPECTOR",..:
9653 1077 2639 7352 4431 6066 9653 5536 927 238 ...
##  $ contb_receipt_amt: num  68.4 80 15 50 100 ...
##  $ contb_receipt_dt : Factor w/ 723 levels "01-Apr-15","01-Apr-16",..: 204 437 491 106 117 29
418 602 709 2 ...
##  $ receipt_desc     : Factor w/ 46 levels ""," SEE REATTRIBUTION",..: 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/ 209 levels ""," SEE REATTRIBUTION",..: 1 1 21 11 21 1 1 1 1 2
1 ...
##  $ form_tp          : Factor w/ 3 levels "SA17A","SA18",..: 2 2 2 1 2 2 2 2 2 2 ...
##  $ file_num         : int  1146165 1146165 1091718 1077404 1091718 1146165 1146165 1146165 11
46165 1091718 ...
##  $ tran_id          : Factor w/ 424363 levels "A001647F906B94BC7AAD",..: 290615 279692 126484
372156 125779 286015 276990 340354 301688 125579 ...
##  $ election_tp      : Factor w/ 5 levels "","G2016","O2016",..: 2 2 4 4 4 2 2 2 2 4 ...
```

The dataset consists of 426057 rows and 19 columns. (I will be deleting some of these columns as they are not important for my analysis)

4 columns have integers/numbers as data. The rest are have different number of categorical variables in a factor format.

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

The fields that are of interest to me are the the candidate name(cand_nm), zipcodes(contbr_zip), contribution amount(contb_receipt_amt) and date of contribution(contb_receipt_dt) as well the contributors themselves.

I didnt require the ‘receipt_desc’, ‘memo_cd’, ‘memo_text’, ‘form_tp’, ‘file_num’, ‘tran_id’ columns. So I decided to remove them.

```
##  [1] "row_num"           "cmte_id"           "cand_id"
##  [4] "cand_nm"           "contbr_nm"         "contbr_city"
##  [7] "contbr_st"         "contbr_zip"        "contbr_employer"
## [10] "contbr_occupation" "contb_receipt_amt" "contb_receipt_dt"
## [13] "receipt_desc"      "memo_cd"           "memo_text"
## [16] "form_tp"           "file_num"          "tran_id"
## [19] "election_tp"
```

# What other features in the dataset do you think will help support your

# investigation into your feature(s) of interest?

Contributor occupations(contbr_occupation) will help me analyze this dataset better. I will also be adding new columns based on other parameters.

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

Other paramaters that I have created are coordinates data, type of donations, types of contributor occupations, split the dates into day, month and years. I have also distributed the candidates based on their party affiliations.

Creating vectors of candidates based on party affiliations and adding them to the data frame.

# 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?

Looking through the data, I had noticed there were some negative contributions(refunds) in the dataset. I decided to ignore these and work without them.

Looking through the zipcodes, I noticed there were zipcodes in the 5 digit format, as well as the additional digits in some of the cases. I decided to clean the data set to retain only the first 5 digits.

Some of the zipcodes represent locations outside Florida. I got rid of these values as well.

I knew I would have to look deeper into the contributions based on the jobs people had. So I decided to look at the different job titles by grouping them together.

```
## # A tibble: 12,444 × 2
##        contbr_occupation  count
##                    <fctr>  <int>
## 1                RETIRED 134213
## 2           NOT EMPLOYED  33092
## 3  INFORMATION REQUESTED  19853
## 4               ATTORNEY  10042
## 5              PHYSICIAN   6021
## 6                TEACHER   5905
## 7              HOMEMAKER   5657
## 8             CONSULTANT   4109
## 9                  SALES   4042
## 10             PROFESSOR   3903
## # ... with 12,434 more rows
```

As you can see, there were 12380 different types of job titles which would be very cumbersome to handle, so I decided to split the contributors based on the type of employment.

So, I decided to divide them among a business owner or CEO, unemployed or retired and 'other'.

The issue I came across this was that, different terms were used to describe the same thing. For example, a person who had his own business wrote in 'SELF'(sometimes 'SELLF'), 'OWN', 'OWNER' and different versions of these terms.

SO I decided to create a regular expression that finds these patterns and puts the person in the right group.

A similar process was done for unemployed people.

I also created a column to distribute the contributions to various categorical values

I then distributed the different occupations to whether a person was an employee, running his own business( CEOs also included in this group) and unemployed.

I also added the zipcodes library so that i could map the zipcodes in this dataset to those in the zipcodes dataset, thus allowing me to acquire the coordinates data.

The final column I added was to differentiate between the different regions of Florida, whether they were an urban city or a rural area.

Then, I decided to remove all additional columns that I didnt require anymore.

```
##  [1] "zipcode"           "row_num"           "cand_id"
##  [4] "cand_nm"           "contbr_nm"         "contbr_city"
##  [7] "contbr_st"         "contbr_zip"        "contbr_employer"
## [10] "contbr_occupation" "amount"            "contb_receipt_dt"
## [13] "election_tp"       "date"              "day"
## [16] "month"             "year"              "party"
## [19] "emp_type"          "donation_type"     "city"
## [22] "state"             "latitude"          "longitude"
## [25] "location_type"
```

I also decided to make a subset of the final presidential nominees.

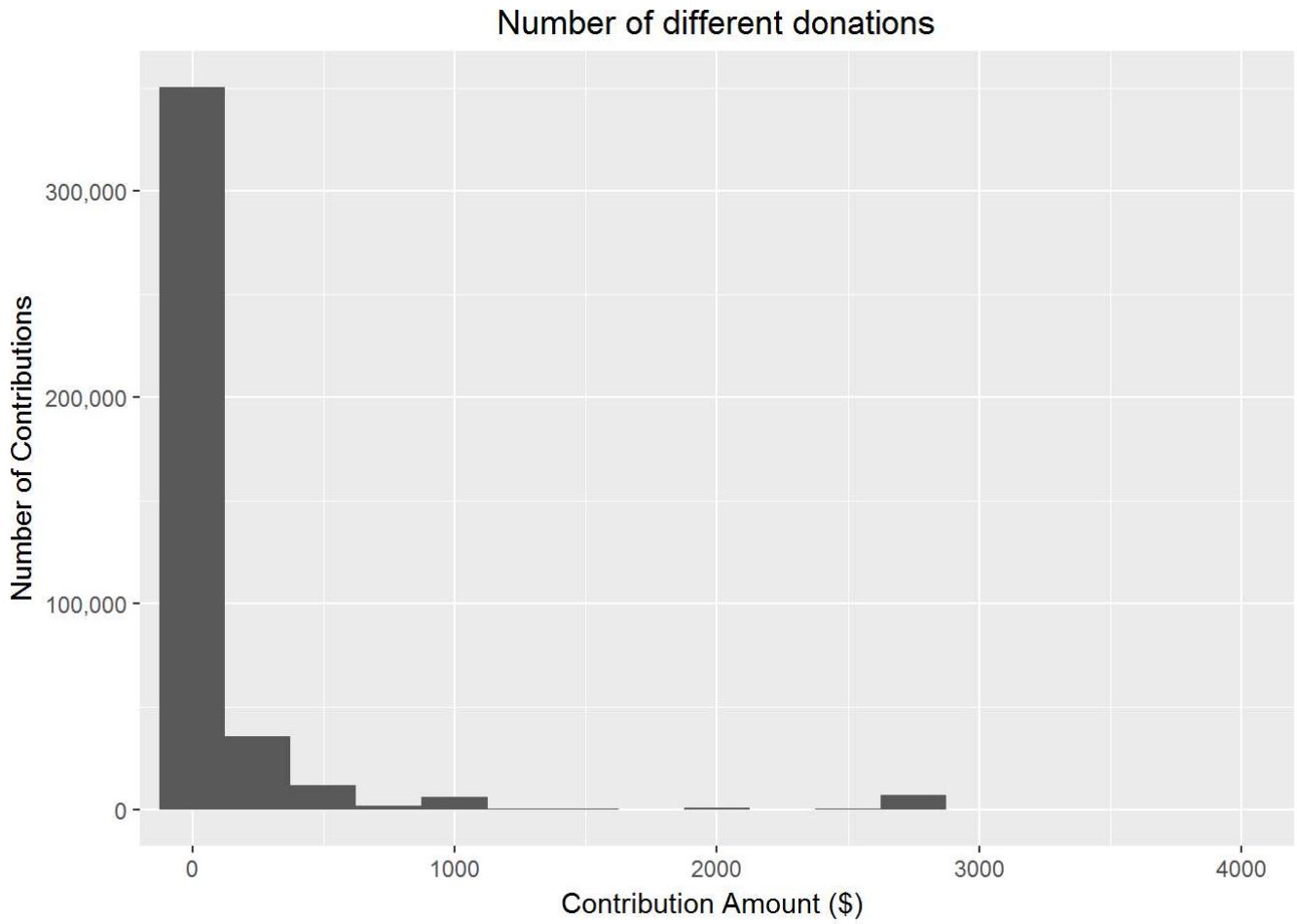
I also decided to ignore the contributions from before the year 2015.

I then took the presidential subset and grouped it together by candidate name. This was done to show the percentages of contributions each received with respect to the other and will be depicted in the conclusions.

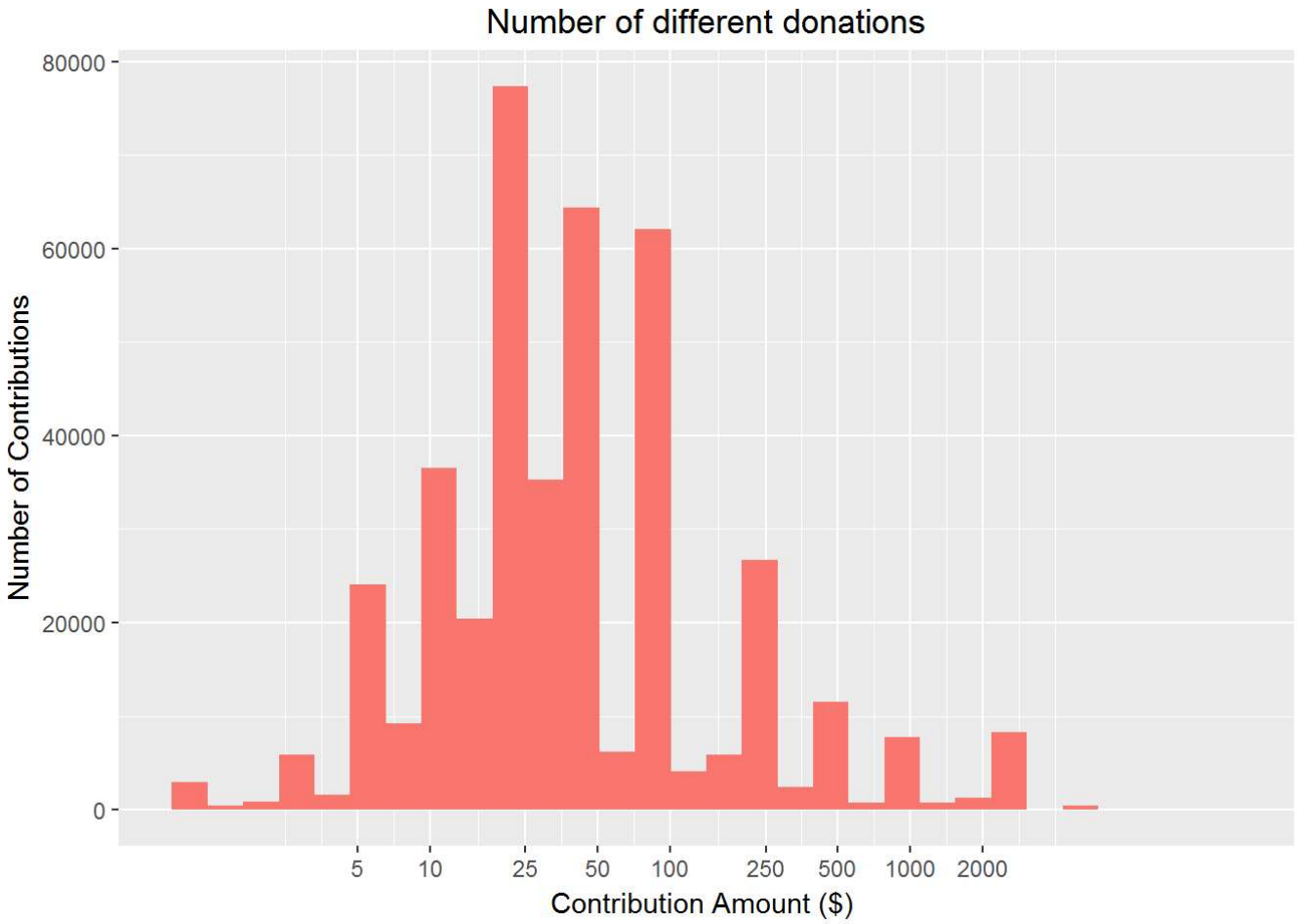
```
## # A tibble: 4 × 7
##                   cand_nm       total     mean median number_contb
##                    <fctr>       <dbl>    <dbl>  <dbl>        <int>
## 1 Clinton, Hillary Rodham 22323771.98 122.6298  25.00       182042
## 2        Trump, Donald J. 12913128.26 169.0864  74.37        76370
## 3           Johnson, Gary   227256.65 288.7632 100.00          787
## 4             Stein, Jill    95726.22 205.8628  50.00          465
## # ... with 2 more variables: percent_number_contb <dbl>,
## #   percent_contb <dbl>
```

Now that the data is clean, it was time to begin the analysis.

The first plot that seemed of interest was the distribution of contributions.




As we can see, the data is positively skewed and some of the data isn't clearly visible. So it is better to convert the x axes to log scale



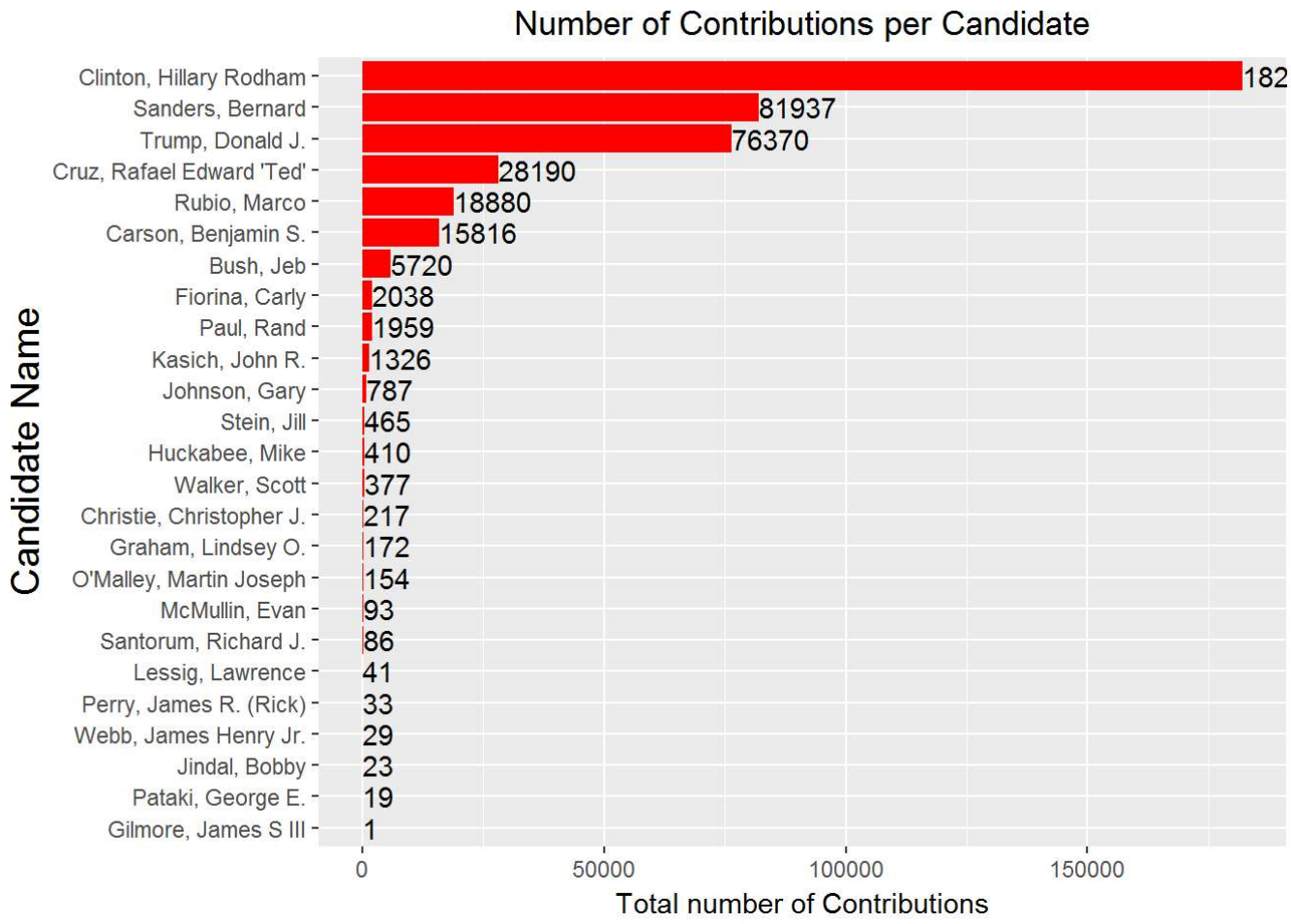

I decided to create a summary of the contributions

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    19.0    35.0   150.7   100.0 20000.0
```

As expected, the mean and median are not close to each other at all, owing to the skew in the data distribution.

I decided to split up the data and look at the number of donations based on each candidate. For that, I had to group the data by each candidate.

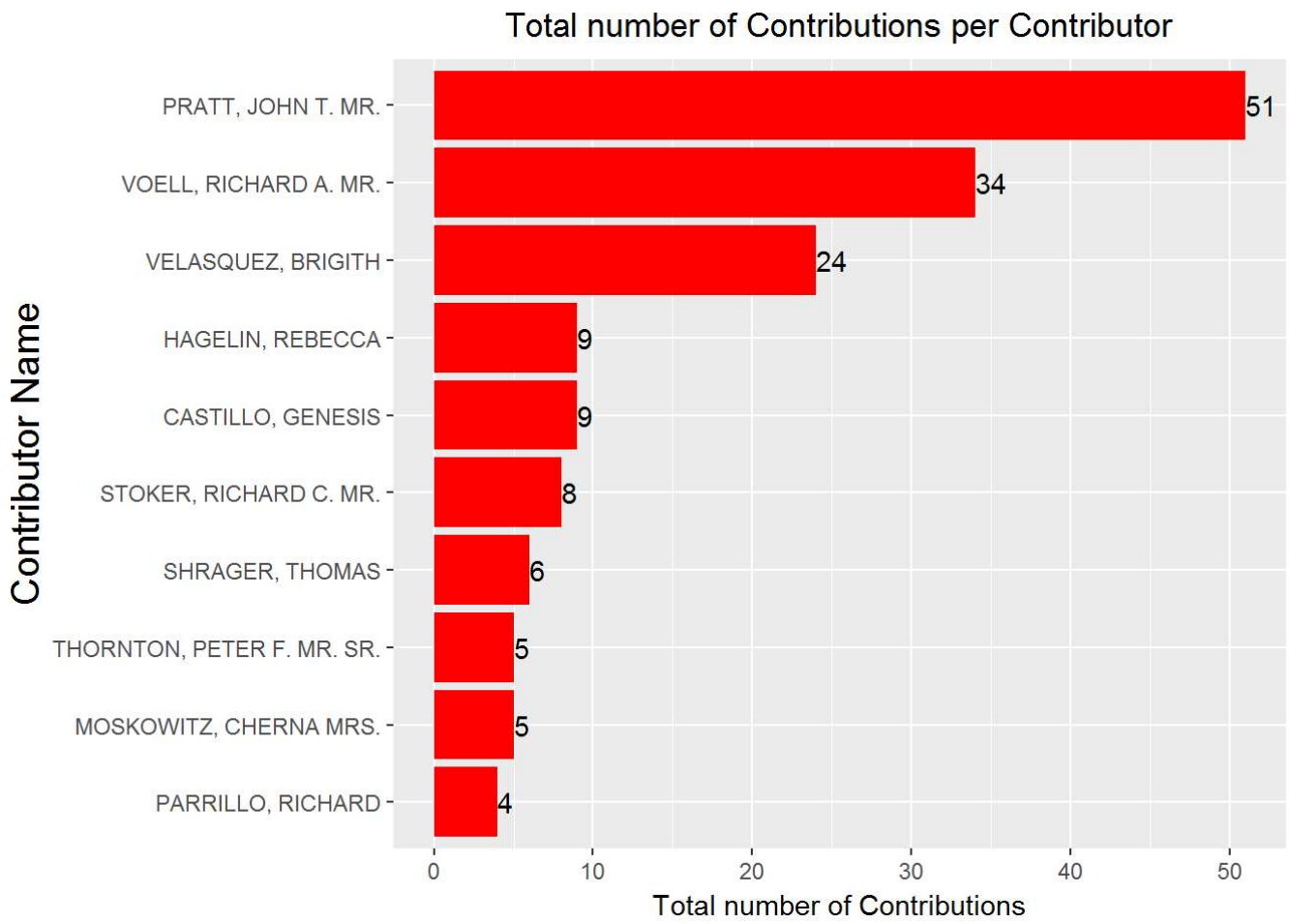
```
## Source: local data frame [25 x 6]
## Groups: cand_nm [25]
##
##                      cand_nm party      total       mean median
##                       <fctr> <chr>      <dbl>      <dbl>  <dbl>
## 1    Clinton, Hillary Rodham     D 22323772.0  122.62979  25.00
## 2           Trump, Donald J.     R 12913128.3  169.08640  74.37
## 3               Rubio, Marco     R  8015192.7  424.53351 100.00
## 4                  Bush, Jeb     R  7029399.0 1228.91591 500.00
## 5  Cruz, Rafael Edward 'Ted'     R  3608352.1  128.00114  50.00
## 6           Sanders, Bernard     D  3481926.7   42.49517  25.00
## 7        Carson, Benjamin S.     R  2034620.1  128.64315  50.00
## 8            Kasich, John R.     R   795640.7  600.03066 250.00
## 9             Fiorina, Carly     R   451328.1  221.45637 100.00
## 10               Paul, Rand     R   394795.5  201.52909  50.00
## # ... with 15 more rows, and 1 more variables: number_contb <int>
```




I wasn't surprised to see both the democrats at the top of the list for number of contributions as they are known to receive a lot of smaller denomination contributions.

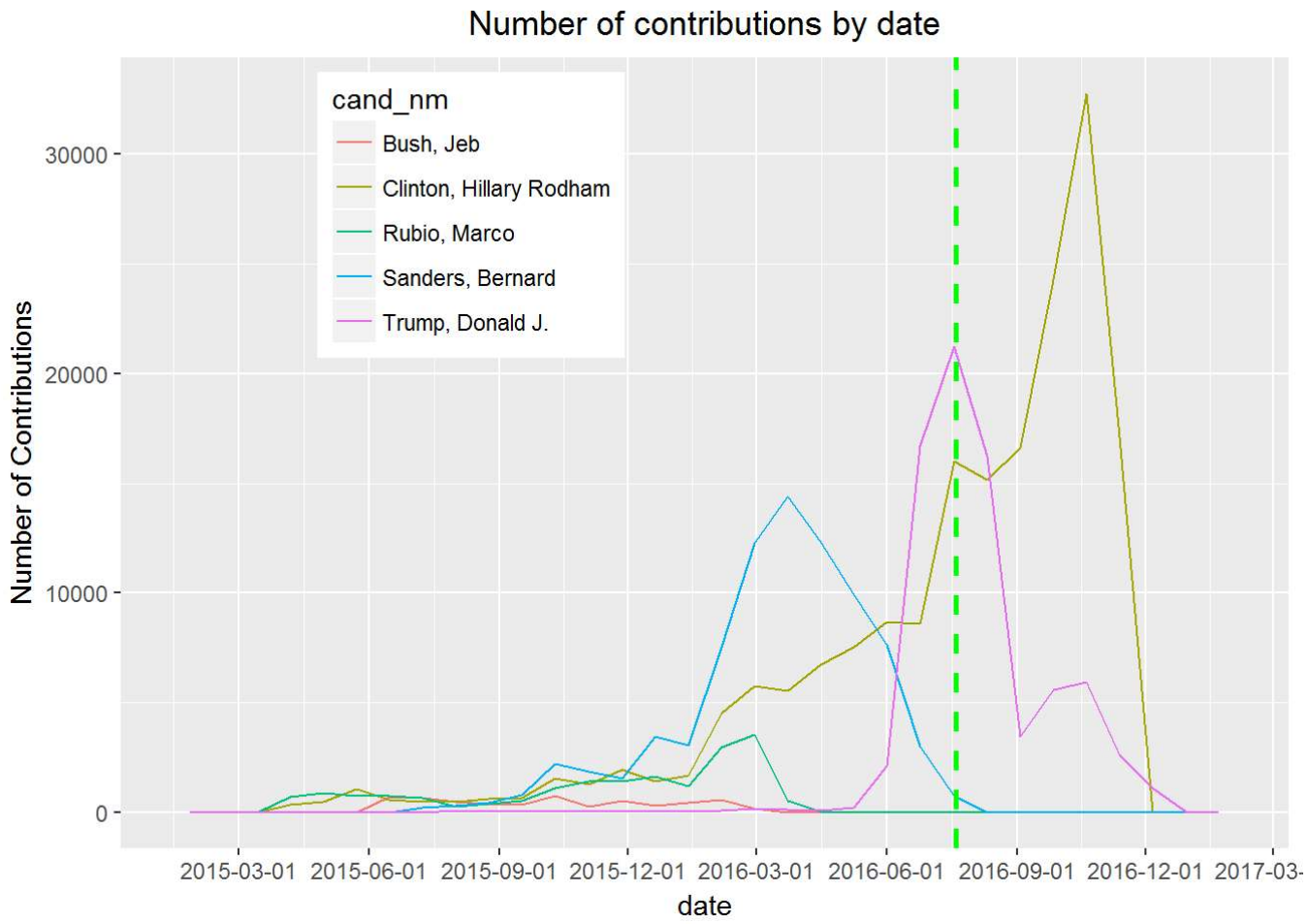
Then I decided to look at the people who had the highest number of contributions. So i had to create another group.

```
## Source: local data frame [121,420 x 6]
## Groups: contbr_nm [112,232]
##
##                     contbr_nm                       contbr_occupation total
##                        <fctr>                                  <fctr> <dbl>
## 1             HAGELIN, REBECCA                               MARKETING 38700
## 2       MOSKOWITZ, CHERNA MRS.                                 RETIRED 30700
## 3            PARRILLO, RICHARD                          CHAIRMAN AND CEO 29700
## 4              SHRAGER, THOMAS                                 ANALYST 27000
## 5  THORNTON, PETER F. MR. SR.                                CHAIRMAN 24300
## 6           PRATT, JOHN T. MR.                                 RETIRED 24050
## 7           VELASQUEZ, BRIGITH INFORMATION REQUESTED PER BEST EFFORTS 23625
## 8       STOKER, RICHARD C. MR.                                 RETIRED 21300
## 9            CASTILLO, GENESIS INFORMATION REQUESTED PER BEST EFFORTS 20991
## 10       VOELL, RICHARD A. MR.                                 RETIRED 20600
## # ... with 121,410 more rows, and 3 more variables: mean <dbl>,
## #   median <dbl>, n <int>
```




I was curious to know from the above graph whether the person who made the highest number of contributions also contributed the most. I go into more detail in a later section

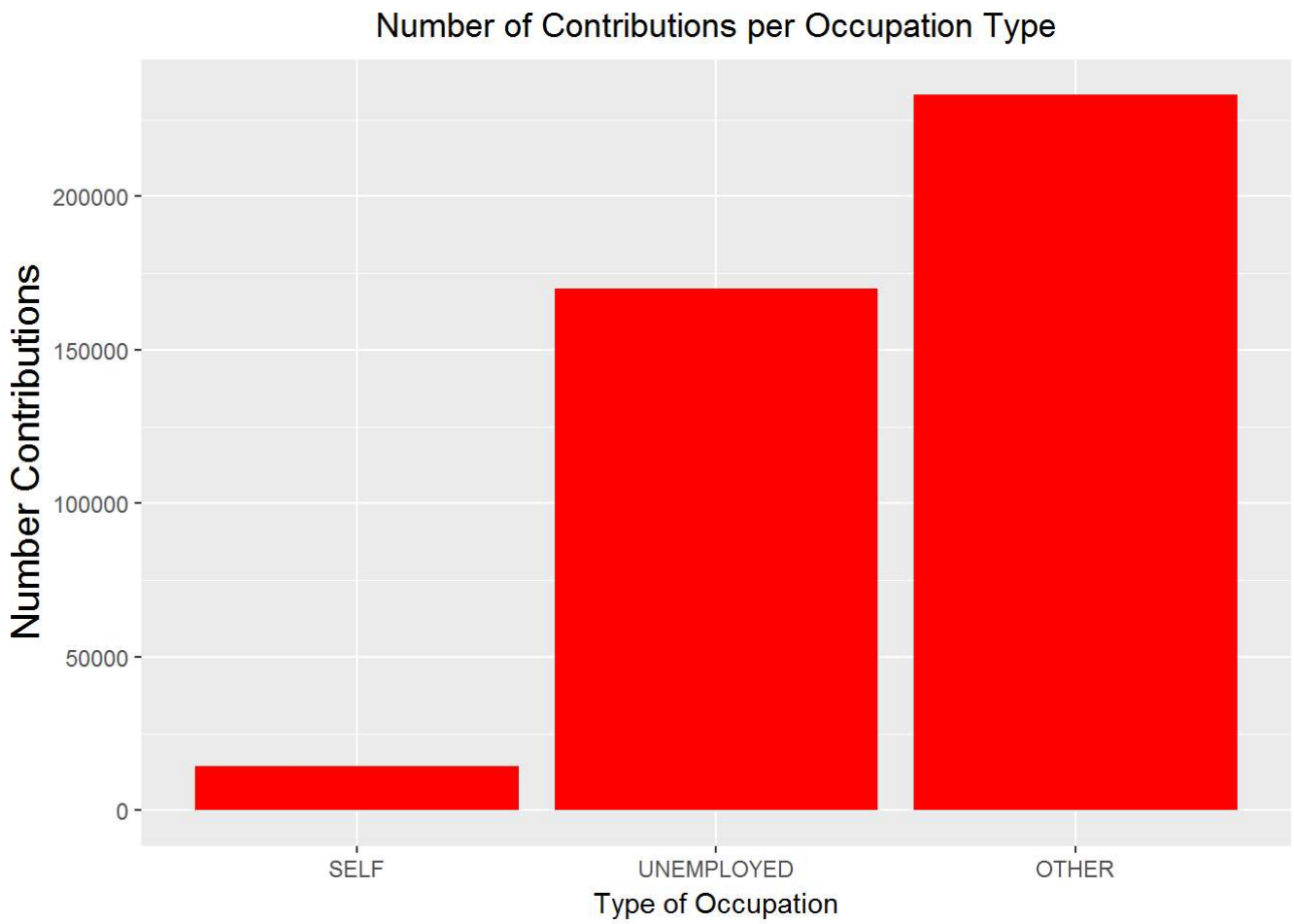
I decided to have a look at how to number of contributions changed over time.




Some of the candidates stopped receiving contributions after a certain date. This is because they dropped out of the presidential race.

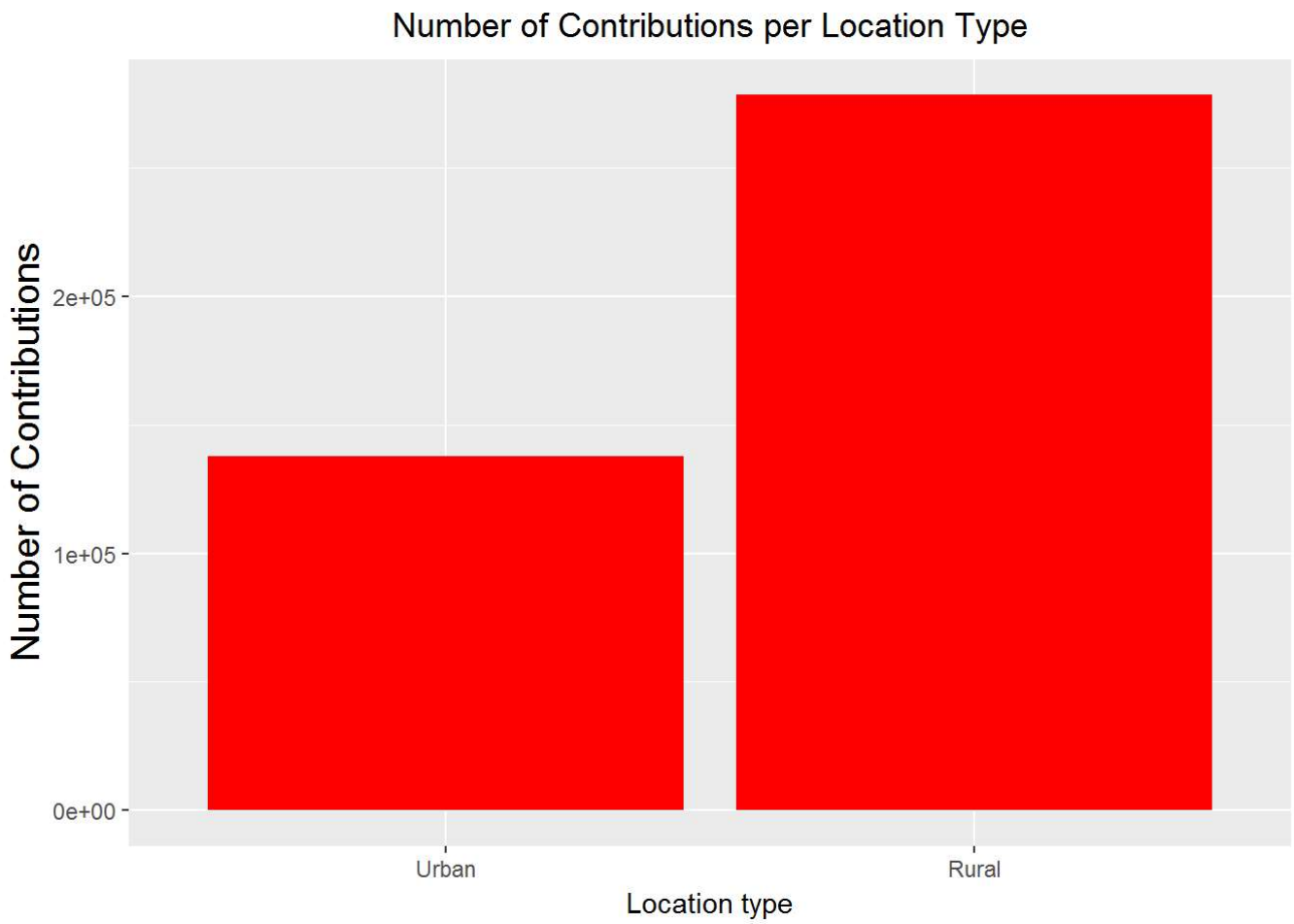
Rubio dropped out around mid march, Jeb Bush around the end of February and Bernie Sanders lost to Hillary Clinton in the primaries. These dates also coincide with the data in the graph above

Finally, for the univariate section, I wanted to check how much people contributed based on their job status.



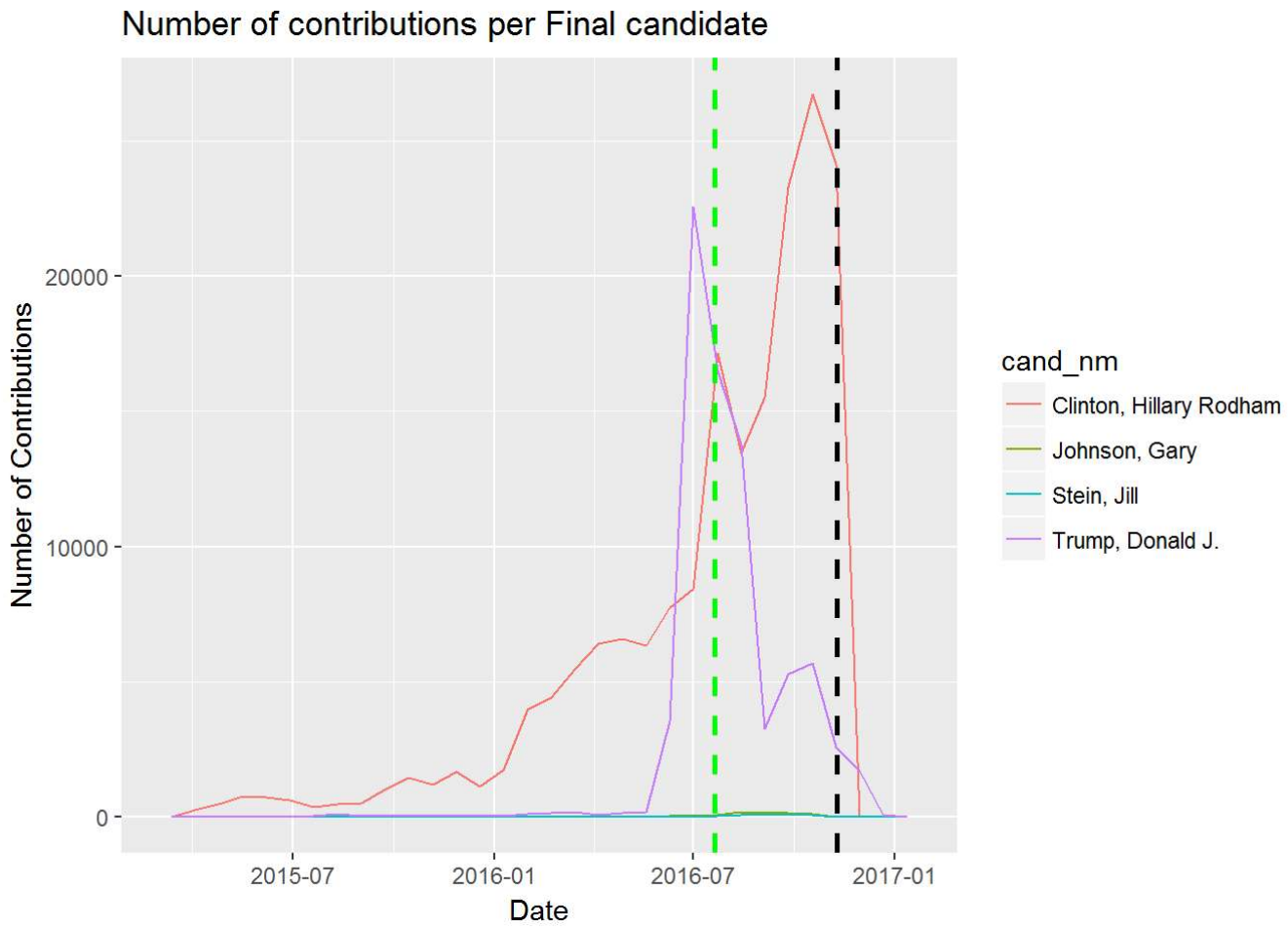

People who ran their own businesses tended to have fewer number of contributions whereas people with regular jobs tended to have the most. This seemed fairly obvious to me since the number of people doing regular jobs far outweigh the number of people running their own business.

Finally I decided to look at contributions based on location type. I had to create a group based on their location and filtered out the data from people who didnt belong to a mainstream party.




Number of contributions is lesser in the urban areas than in the rural areas.

Finally, I wanted to compare the contributions per time for the party nominees.




The green dashed line is the date around the conventions for each party. The black line is election day.

We can see a gradual and constant increase in number of donations for Mrs. Clinton from the beginning of the year, and sudden spikes in and around the convention. There was a sharp drop in contributions immediately after election day for both candidates, which makes sense.

Donald Trump had a uniform number of contributions throughout the year up until the convention where the contributions had a huge spike followed by a huge drop after.

Curious thing is, he still had contributions coming in a whole month after the election.

# Bivariate Plots Section

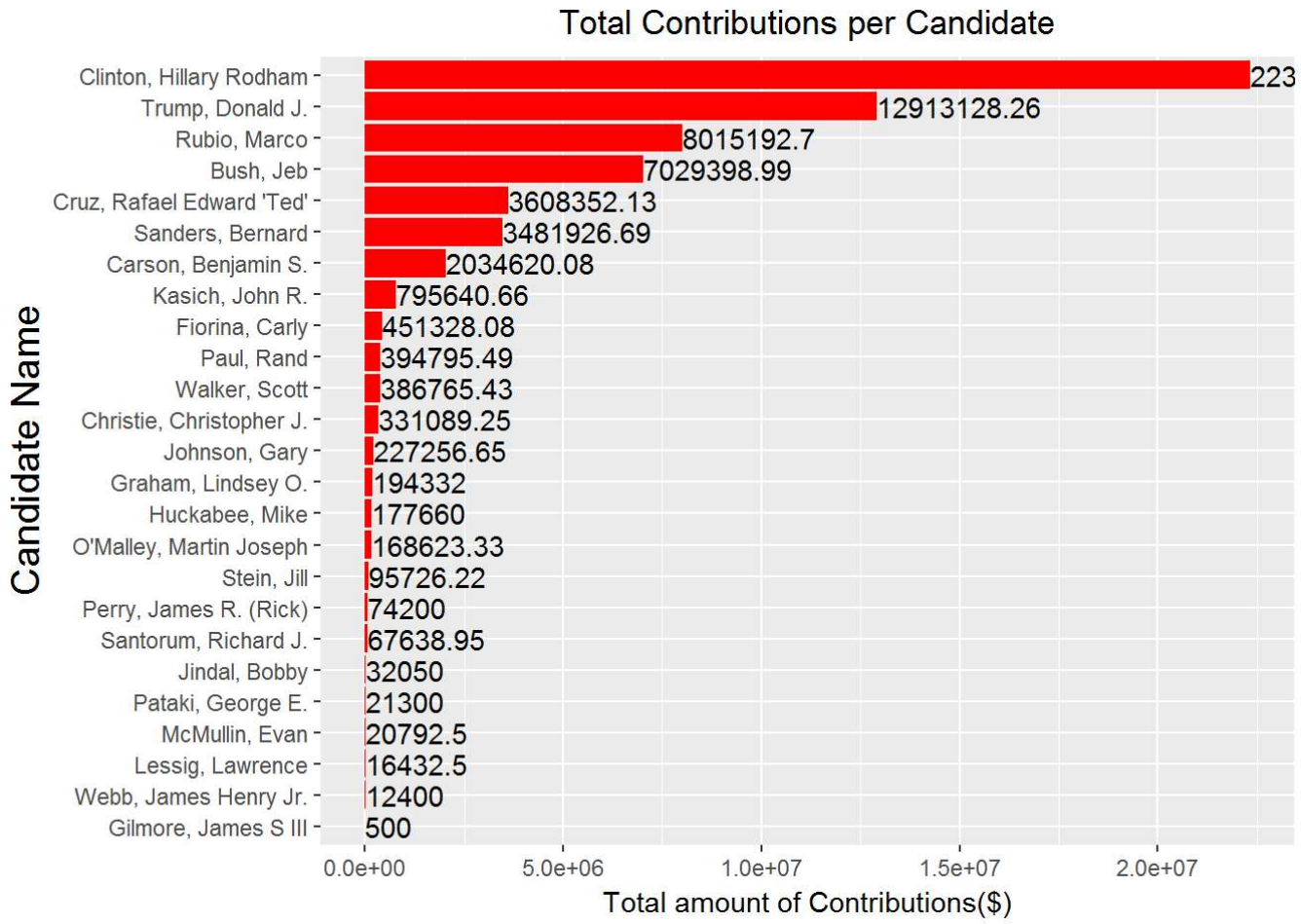
# Bivariate Analysis

## 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

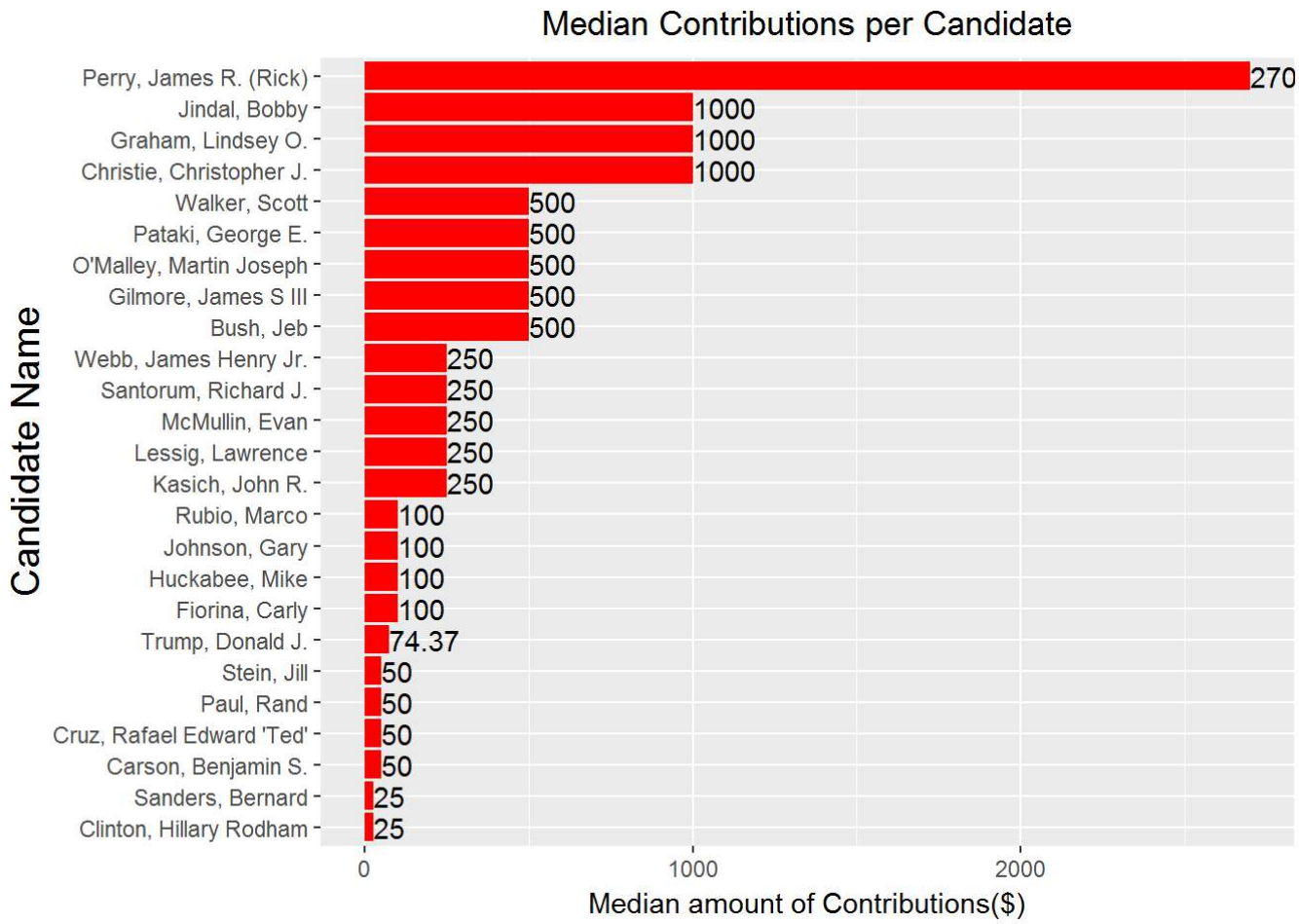
## in the dataset?

Earlier I noticed that the major democratic candidates had more number of contributions. I was expecting them to also top the list in terms of amount of money contributed.



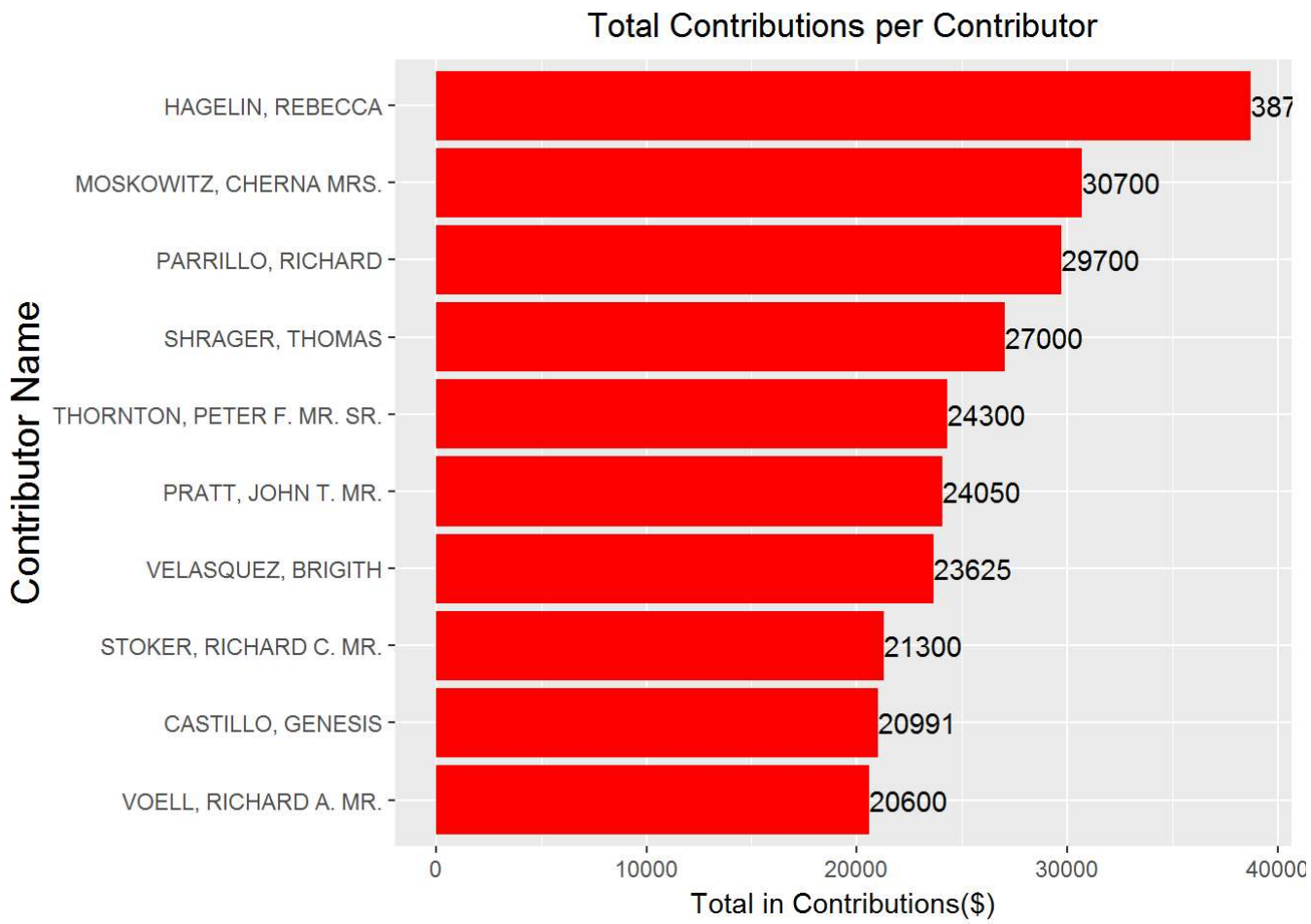

I was a little surprised at first to see that Bernie Sanders was in 6th. So I plotted the median contributions below and noticed that his median contributions were joint lowest.

It is important to note here that Jeb Bush and Marco Rubio are high up on this list despite dropping out of the race early. It is important to note here that Jeb Bush was governer of Florida and Marco Rubio has been a senator in Florida for years.



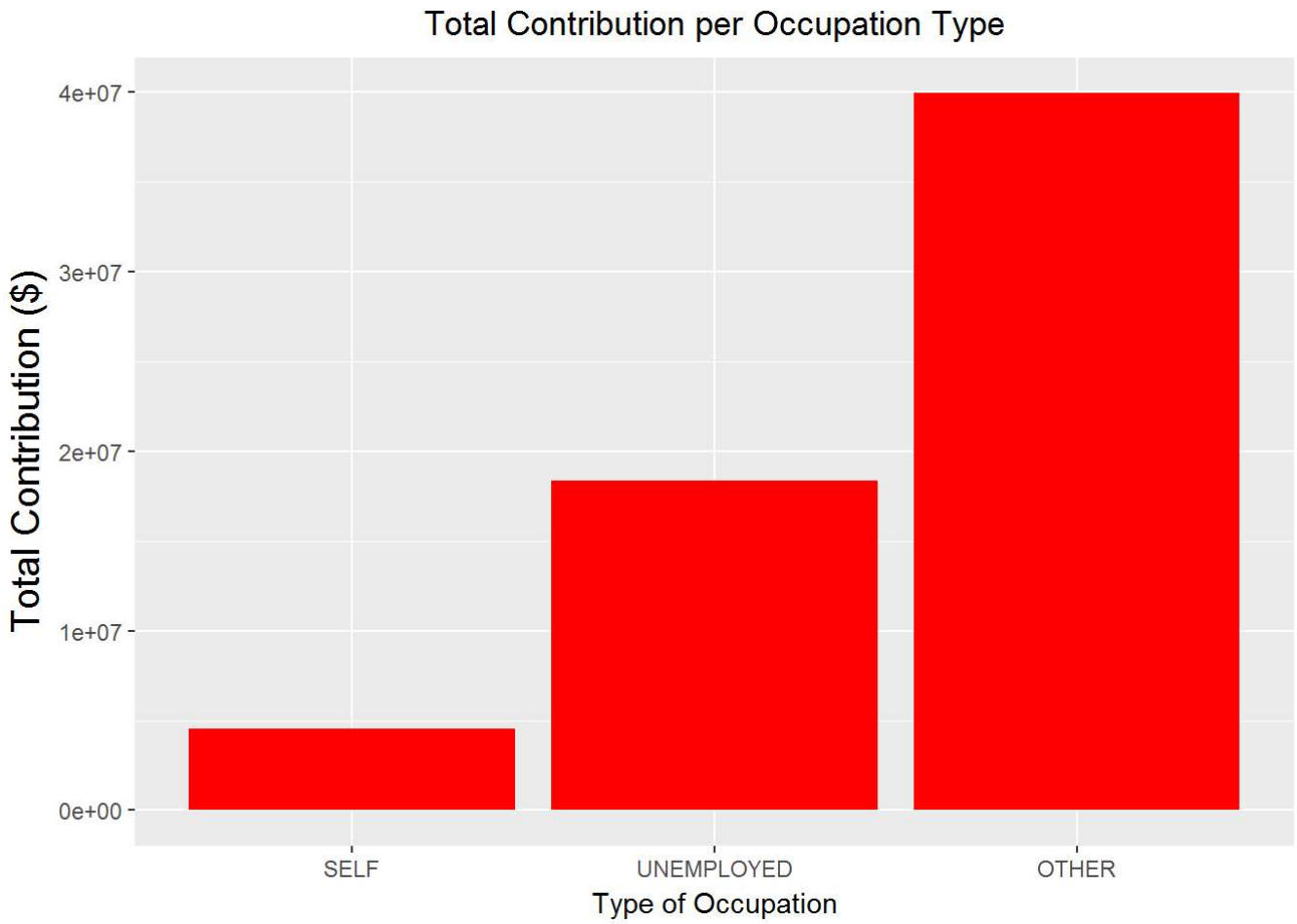

I was a little surprised to see Hillary Clinton at the bottom of this graph despite her raising the most money from contributions. Another look at the number of donations graph in the beginning showed that she had many more contributions than the rest.

I then wanted to see who made the most contributions. So I plotted a graph to return the top 10 in highest contributions.

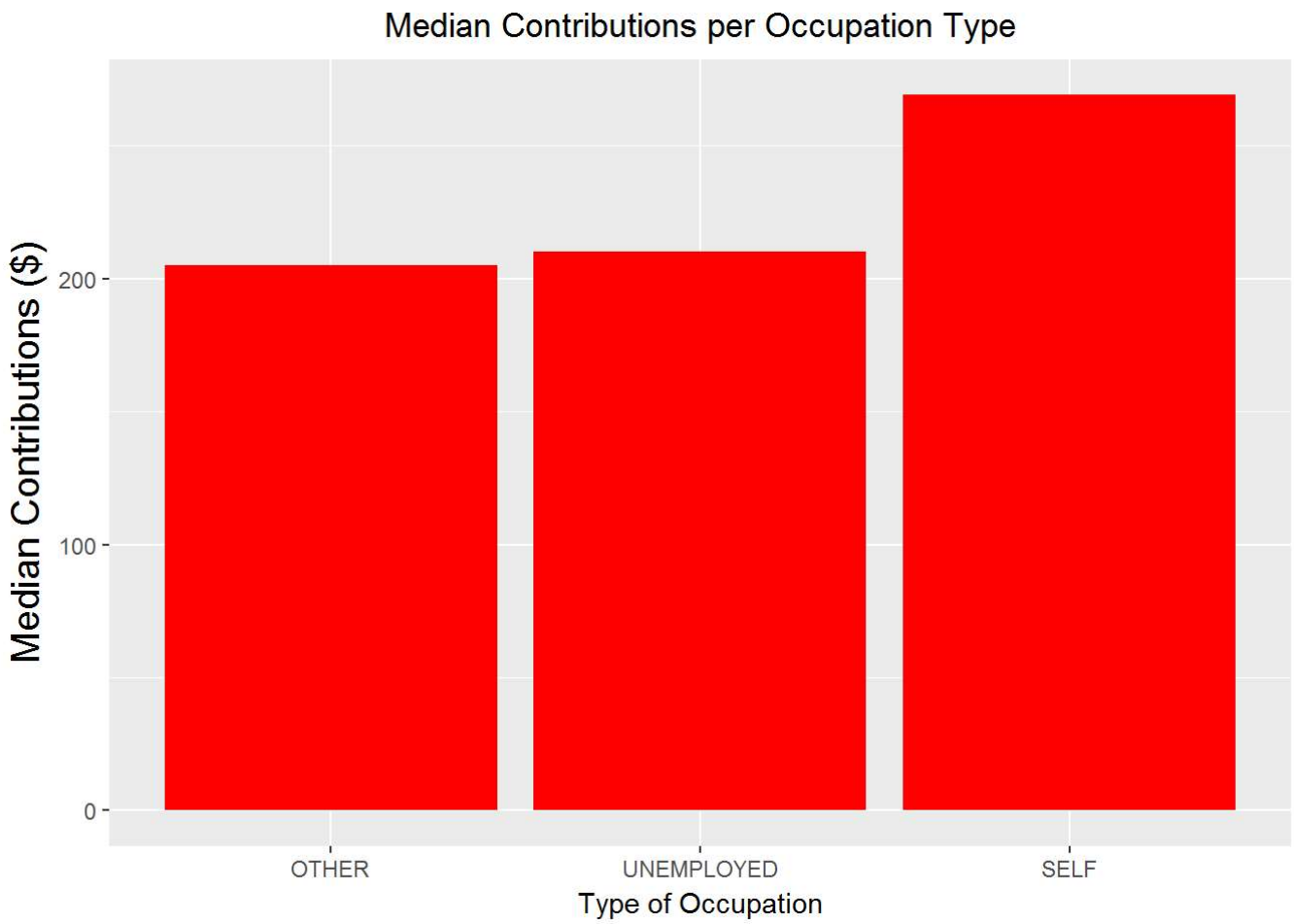



The people who contributed the most amount of money didnt necessarily make the most number of donations. I will go deeper into this in a later plot to see if there was a relationship between the number of contributions vs the amount of money raised.

I then decided to look at the amounts contributed based on occupation type



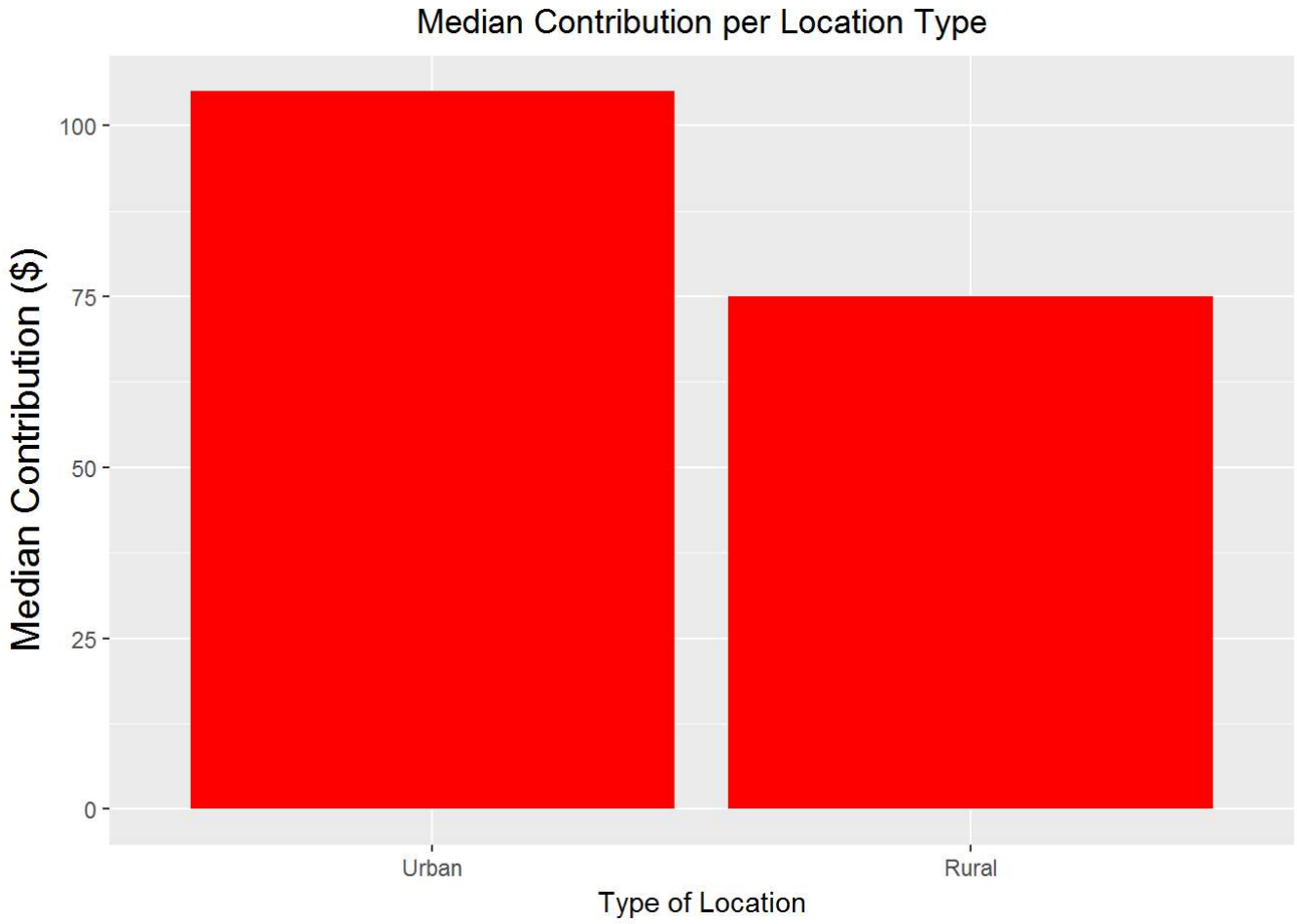

Total contributions seem to be the least for people who were self employed whereas people who worked regular jobs contributed the most. A point to note here is that the number of people who were self employed and contributed were much fewer than the rest as shown from the earlier graph. The median contributions will give us a better idea.




The median contributions from self employed people were more than that of the rest, but what surprised me here was that it was much lesser than I expected.

I then proceeded to have a look at the total contribution based on location type.

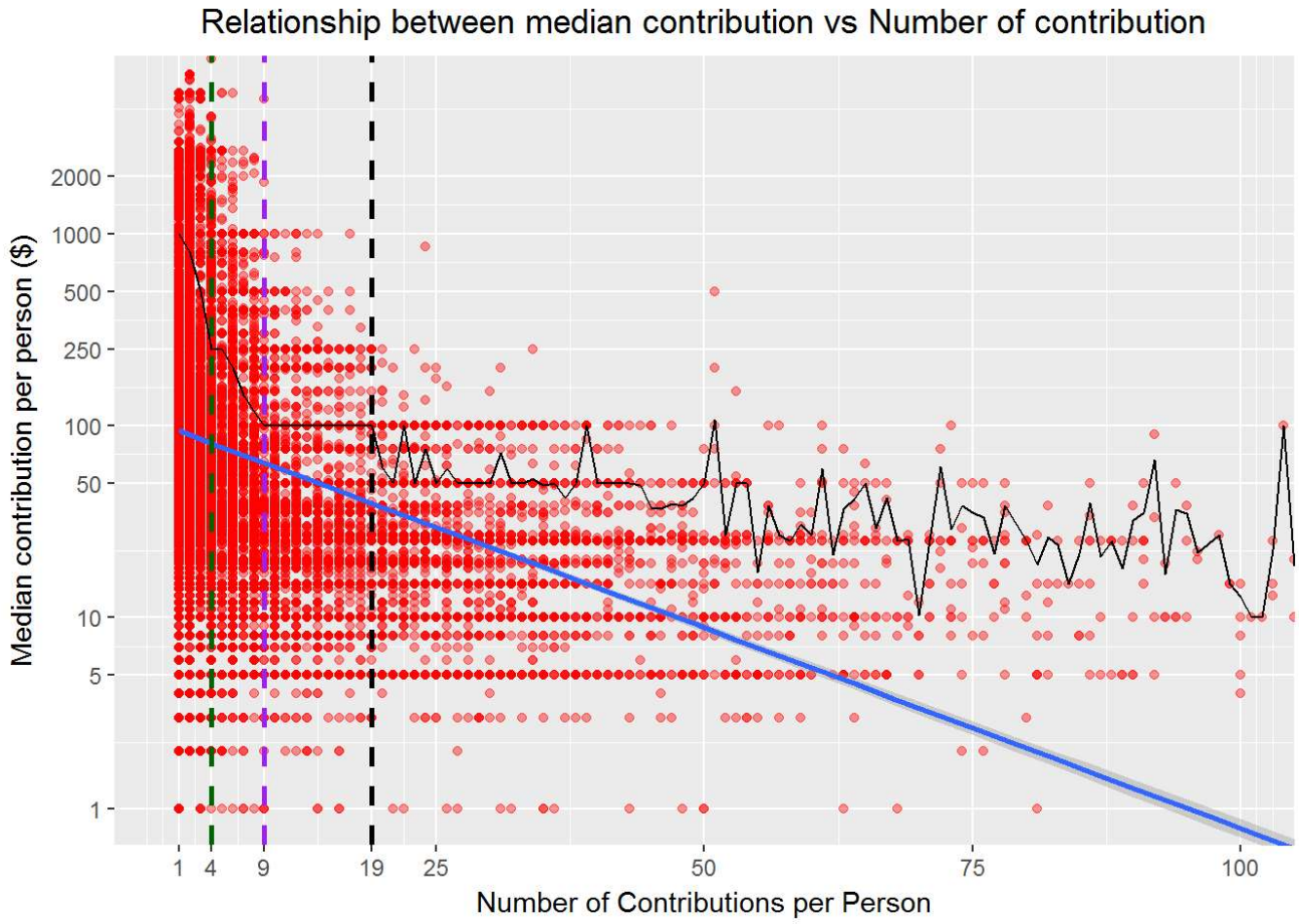

Total Contribution per Location Type
4e+07
3e+07
2e+07
1e+07
0e+00
Total Contribution ($)
Urban
Rural
Type of Location


Median Contribution per Location Type
100
75
50
25
0
Median Contribution ($)
Urban
Rural
Type of Location

The two graphs above didn't really surprise me as I expected median contributions of people living in urban areas to be higher than than of rural areas, but not high enough to bump the total contributions higher than that of rural areas owing to the high number of donations made by people in rural areas.

At this point, I wanted to take only the presidential data and split the contributions between urban and rural areas to compare between the candidates.

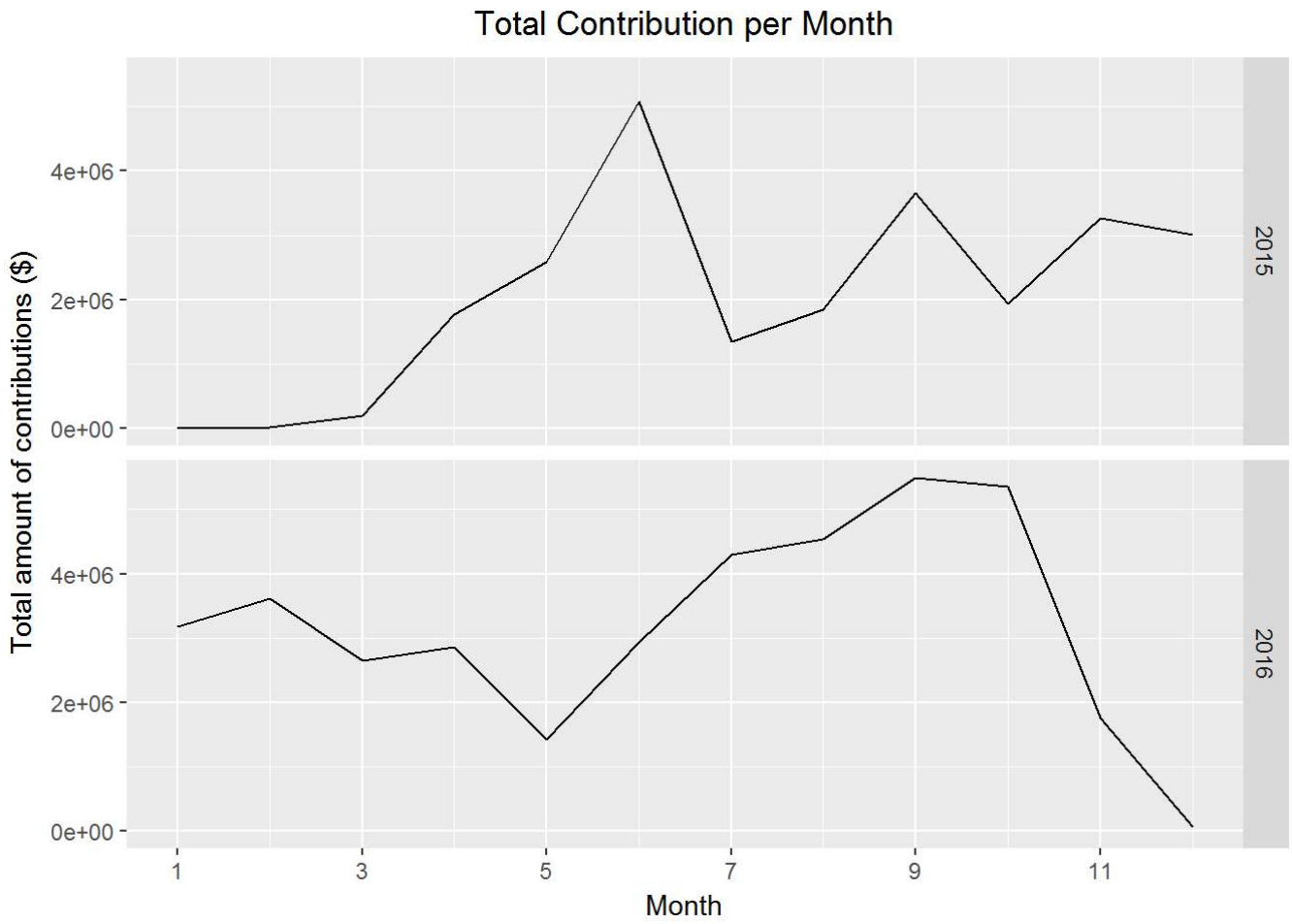
Proceeding to find relationships between the number of contributions vs the amount contributed. I needed the grouped variable I had created earlier cont_by_comp.




This graph shows that people who made large number of contributions tended to make smaller contributions(albeit the relationship isnt as strong as I imagined). The black solid line in the graph depicts the upper quartile for the number of contributions.

I've drawn 3 dotted vertical lines at certain points in the graph which I found interesting. There seems to be a sharp drop in upper quartile donation from people who made 1 contribution to 4 contributions, followed by another sharp drop to people who made 9 contributions. Then it remains steady till people who made 19 contributions and from there on, there are ups and downs but the general trend seems to be a steady and constant decline.

I also wanted to look at how contributions varied between the same months of two different years. So first, I had to group the data by year and month. Then I plotted the data to show the difference in contributions for each year.




June was a good month in both years with a spike in contributions from May.

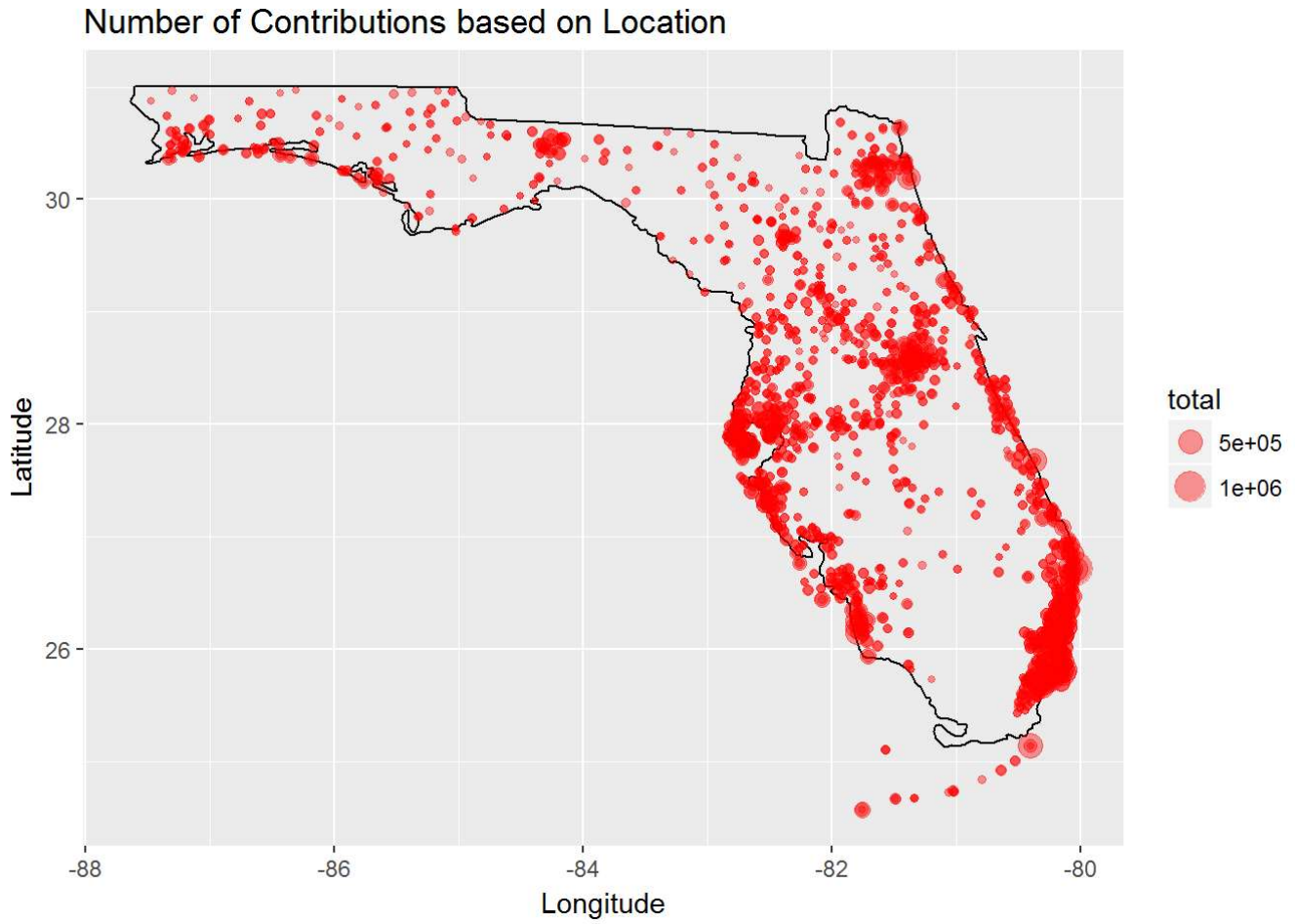
The graph matched my expectation from march to may 2016 with downward trends in total contributions. This was probably due to two major contestants in Marco Rubio and Jeb Bush dropping out of the race.

The graph also seemed to agree with me in thinking that contributions increased as the conventions for both parties were approaching. It also increased up until September as election day was nearing.

What surprised me was the last couple of months before election day where there was a steep drop in total contributions especially considering that Florida was a major swing state for whom both presidential candidates campaigned hard for.

After this, I wanted to check where majority of the contributions came from. So I grouped the contributions by coordinates and plotted the result.

I created a new variable from the maps library called ‘FL’ which has the coordinate data of the state of Florida. Then I grouped the data from my dataset by those coordinates and plotted the data over them.




The contributions came from a large number of places all across the state. In a later section, I go deeper into this plot, to check which presidential candidate received more money on the basis of location type (Urban/Rural).

## What was the strongest relationship you found?

plotting of contribution size vs number of contributions seemed to have an inverse relationship to an extent but it wasn't as stong as I had hoped.

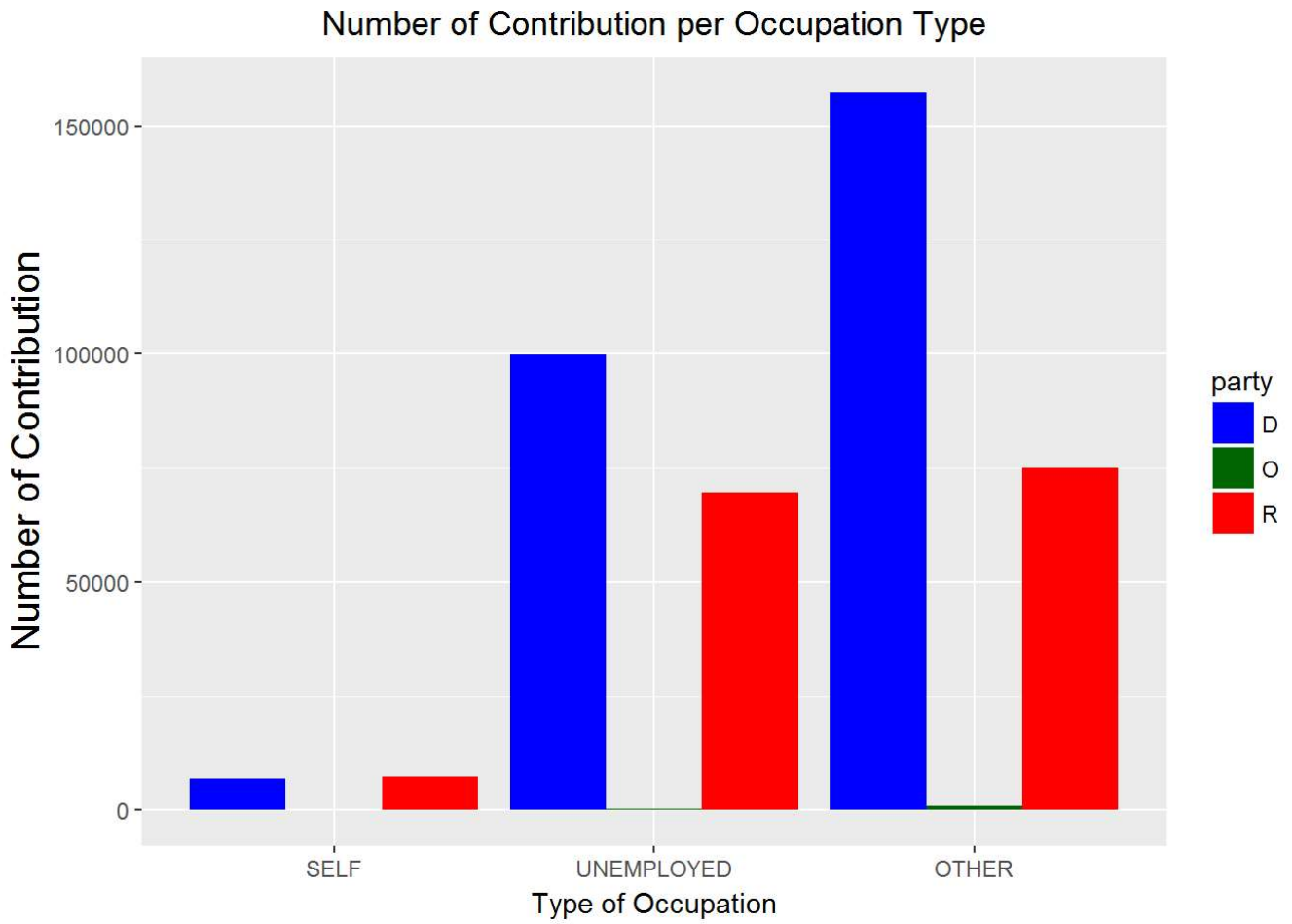
# Multivariate Analysis

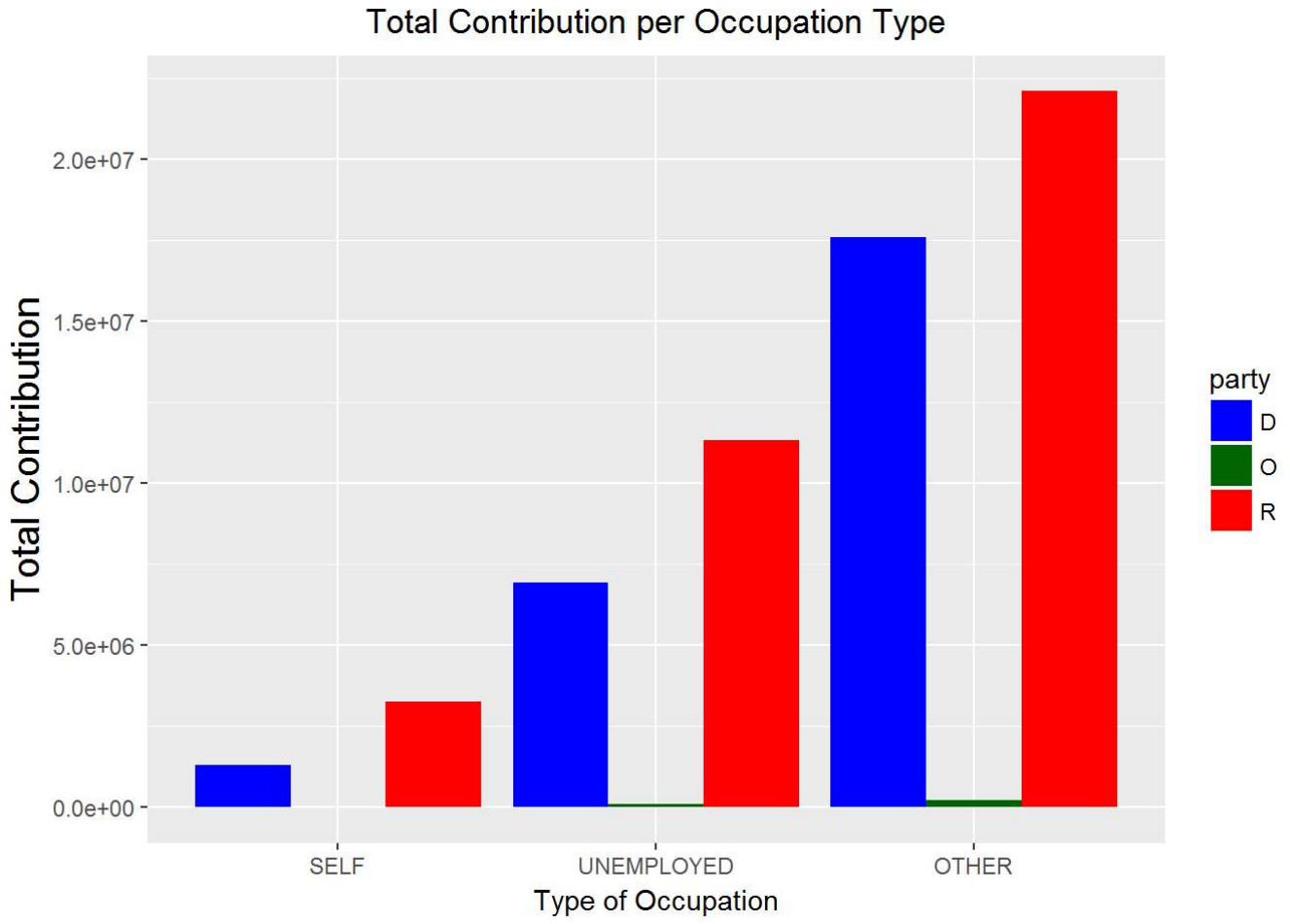
## Talk about some of the relationships you observed in this part of the

## investigation. Were there features that strengthened each other in terms

## of looking at your feature(s) of interest?

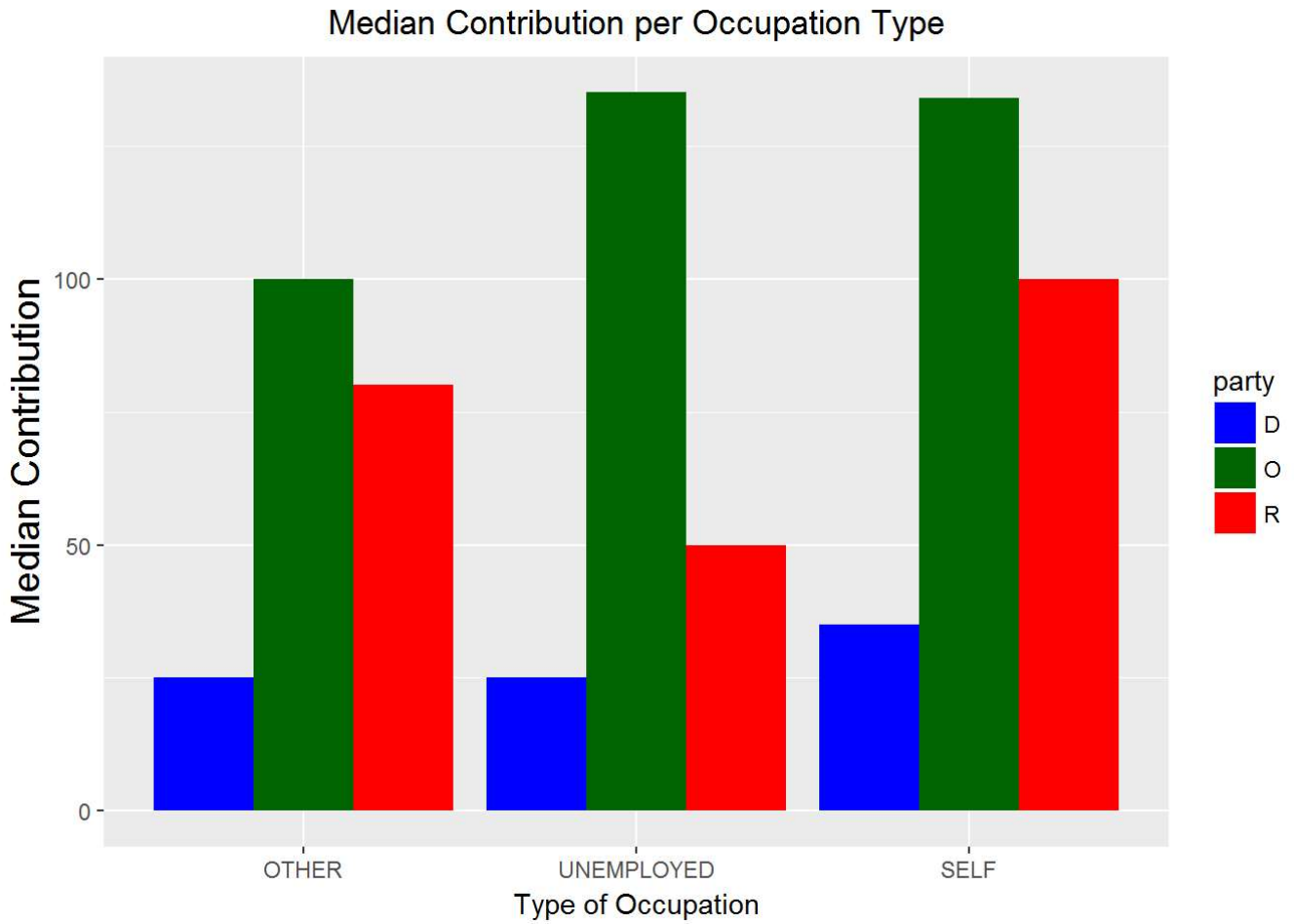
Earlier, we had taken a look at the total contributions per occupation type. I decided to go a bit deeper and see how contributions were distributed by them along with party affiliations.


Number of Contribution per Occupation Type
Number of Contribution
150000
100000
50000
0
SELF
UNEMPLOYED
OTHER
Type of Occupation
party
D
O
R


Total Contribution per Occupation Type
Total Contribution
2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00
SELF
UNEMPLOYED
OTHER
Type of Occupation
party
D
O
R

Not too many of the contributions go towards third party candidates with most of the contribution amount going to the Republican candidates. I wasn't surprised to see that most number of people with regular jobs tended to contribute towards the democratic party.

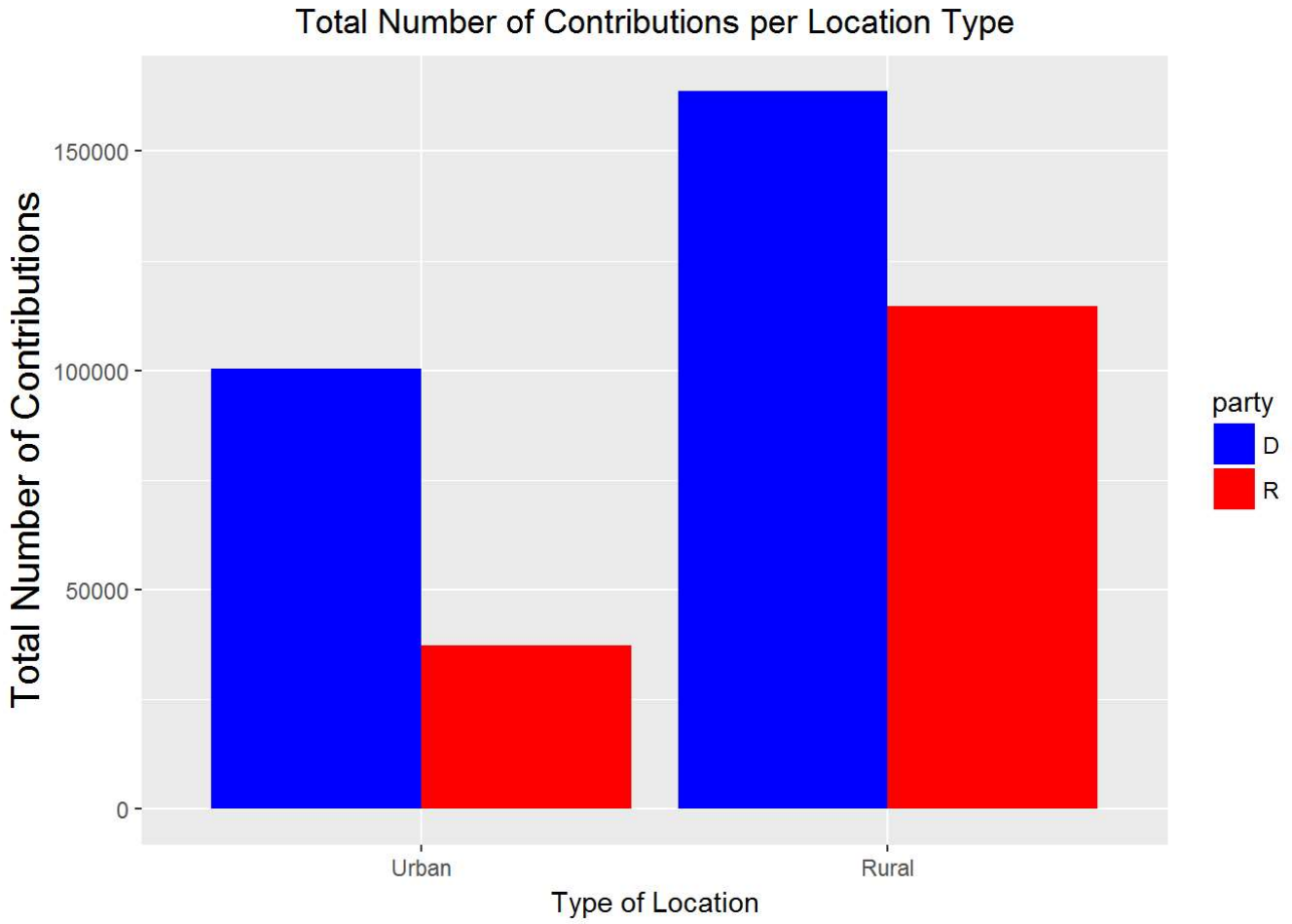
Looking at median contributions would help us paint a better picture



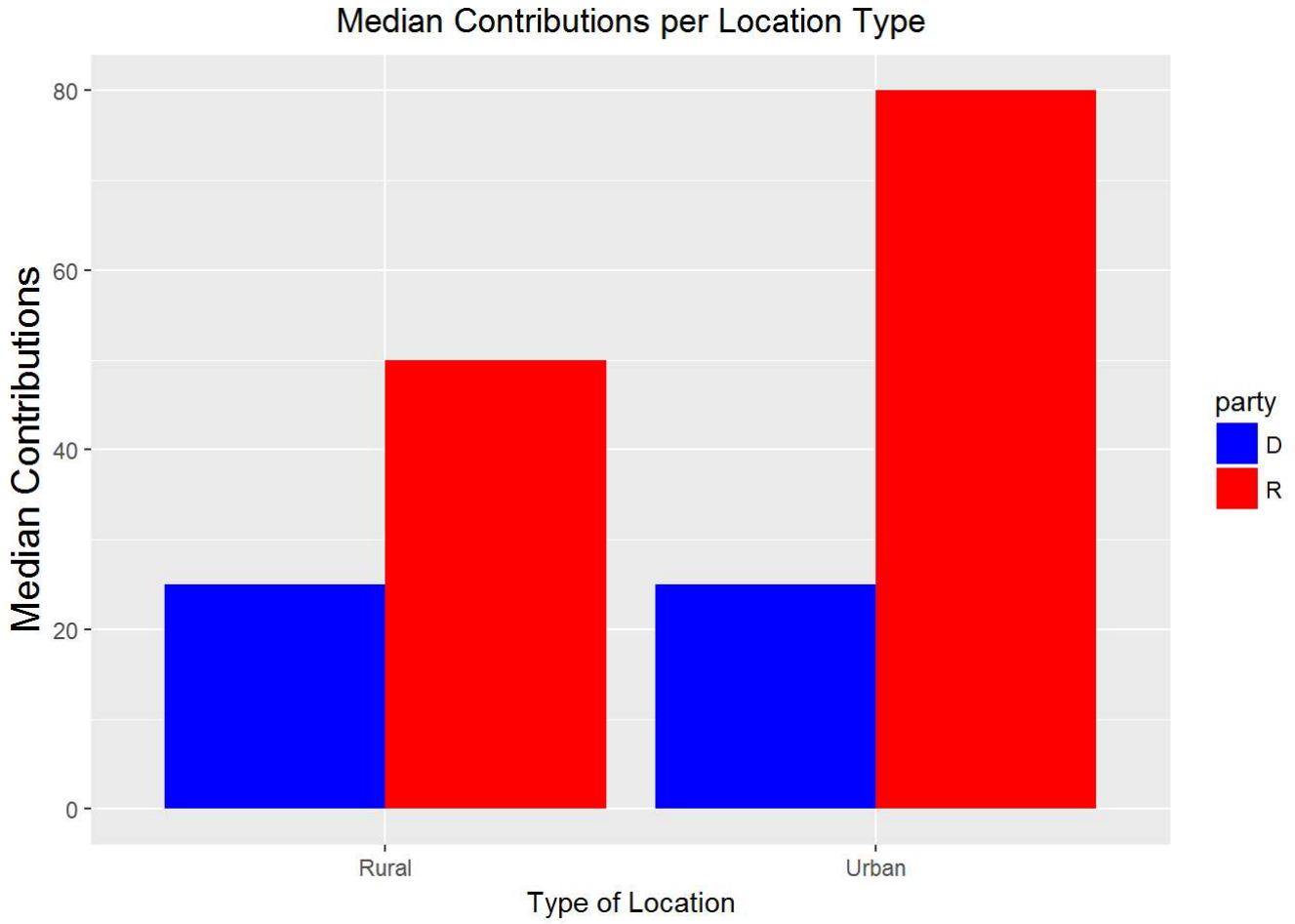

Democrats tended to receive smaller donations whereas third party candidates, albeit getting much lesser number of contributions, tended to get contributions bigger in size.

I decided to then go deeper into contributions per location type.


Total Contribution per Location Type
Total Contribution
2.5e+07
2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00
Urban
Rural
Type of Location
party
D
R

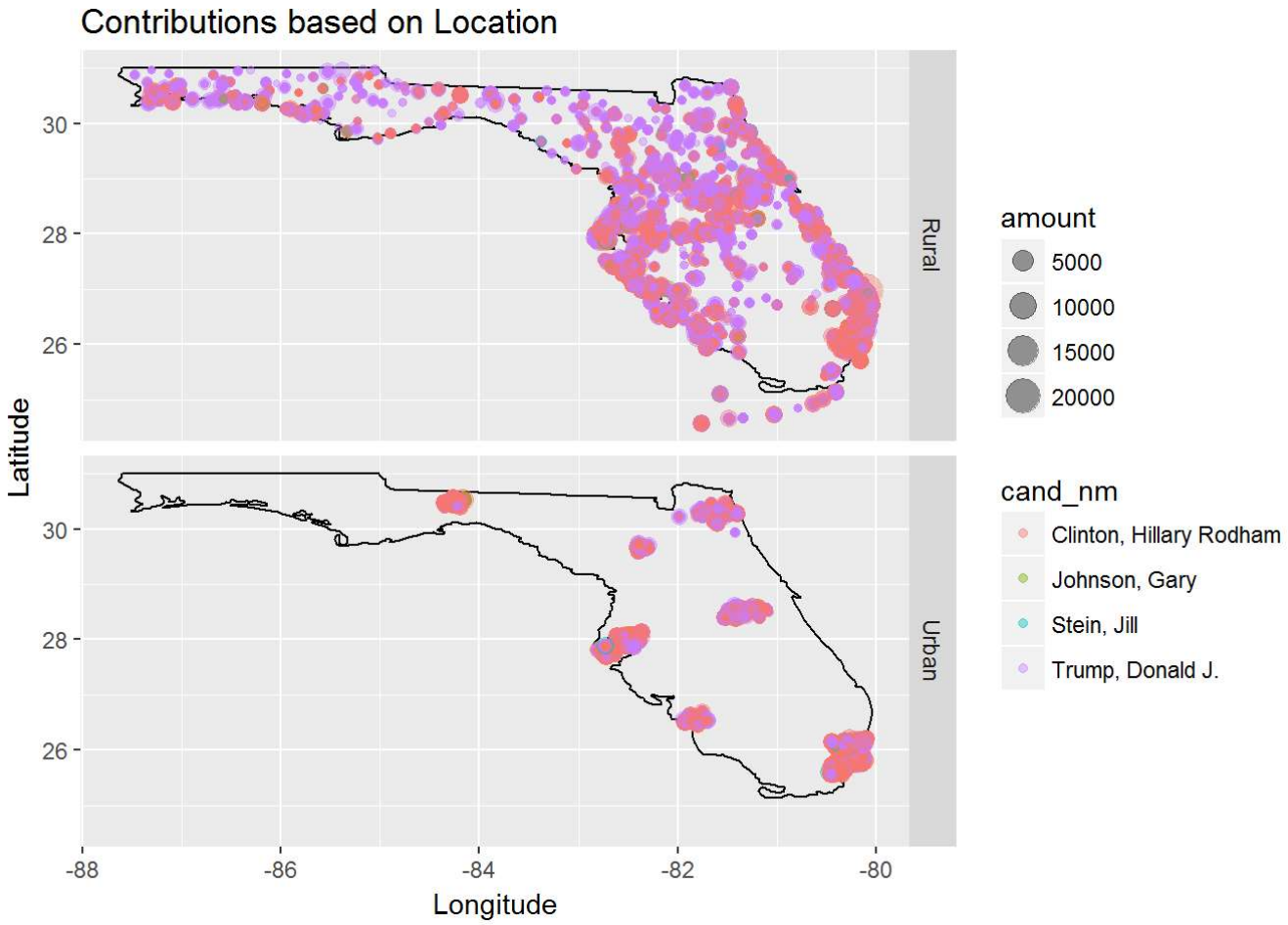

Total Number of Contributions per Location Type
Total Number of Contributions
150000
100000
50000
0
Urban
Rural
Type of Location
party
D
R

In Urban areas, a lot more of the people were contributing money to democratic candidates, therefore we can see that they've contributed more money even thought their median amount(in the following graph) is much lesser.

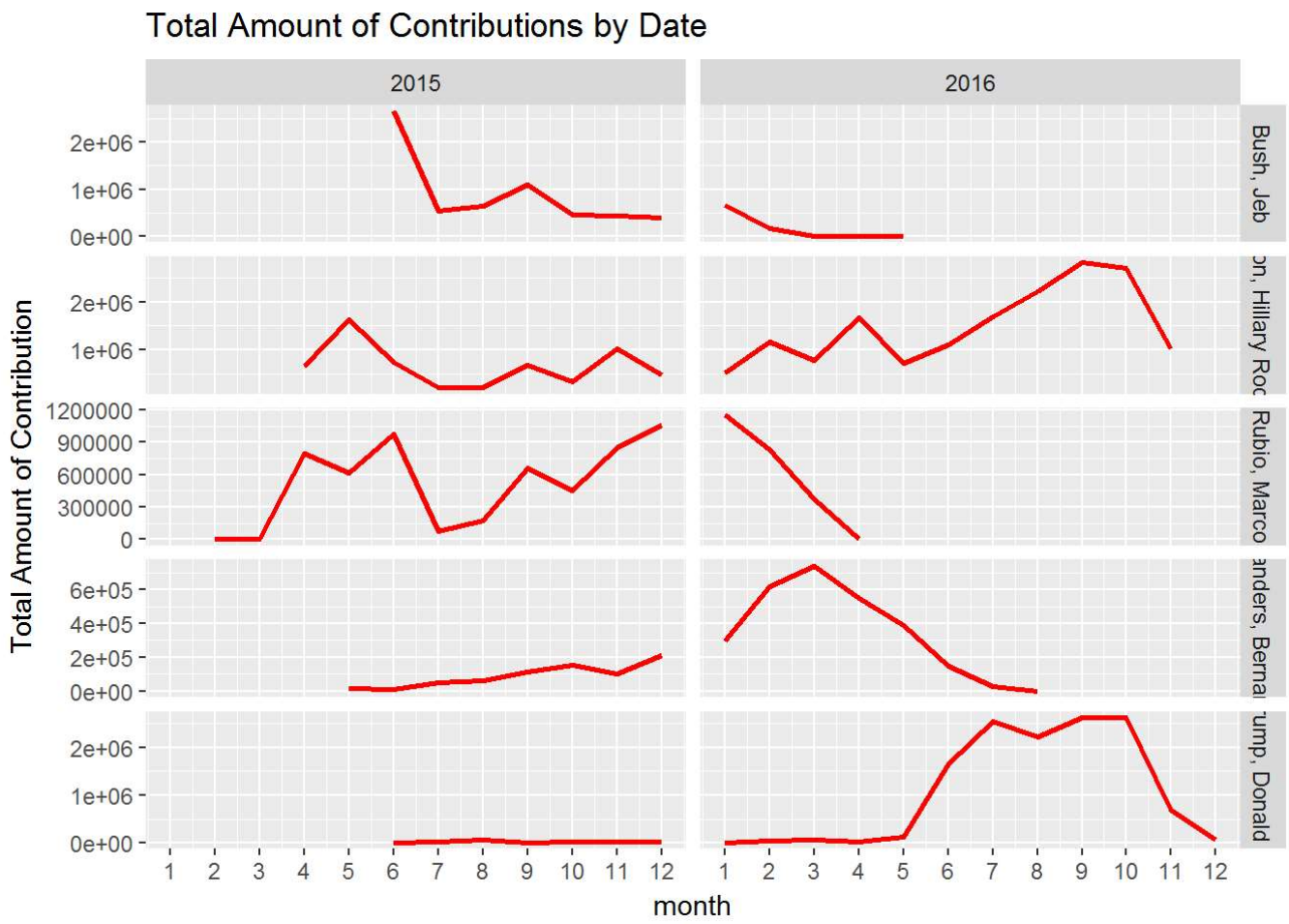



I wanted to have a look at the contributions for the final presidential candidates. So i decided to split the contributions on two maps and check the distribution of contributions in rural and urban areas.

I felt that more people from urban areas would contribute towards Hillary Clinton considering people from cities tend to prefer more progressive candidates.




In an earlier plot, I had shown the number of contributions each candidate received over time. Here, I will plot the total amount of contributions each candidate received over time. For that, I will group the data by the major candidates defined earlier and by the months of each year.




## Were there any interesting or surprising interactions between features?

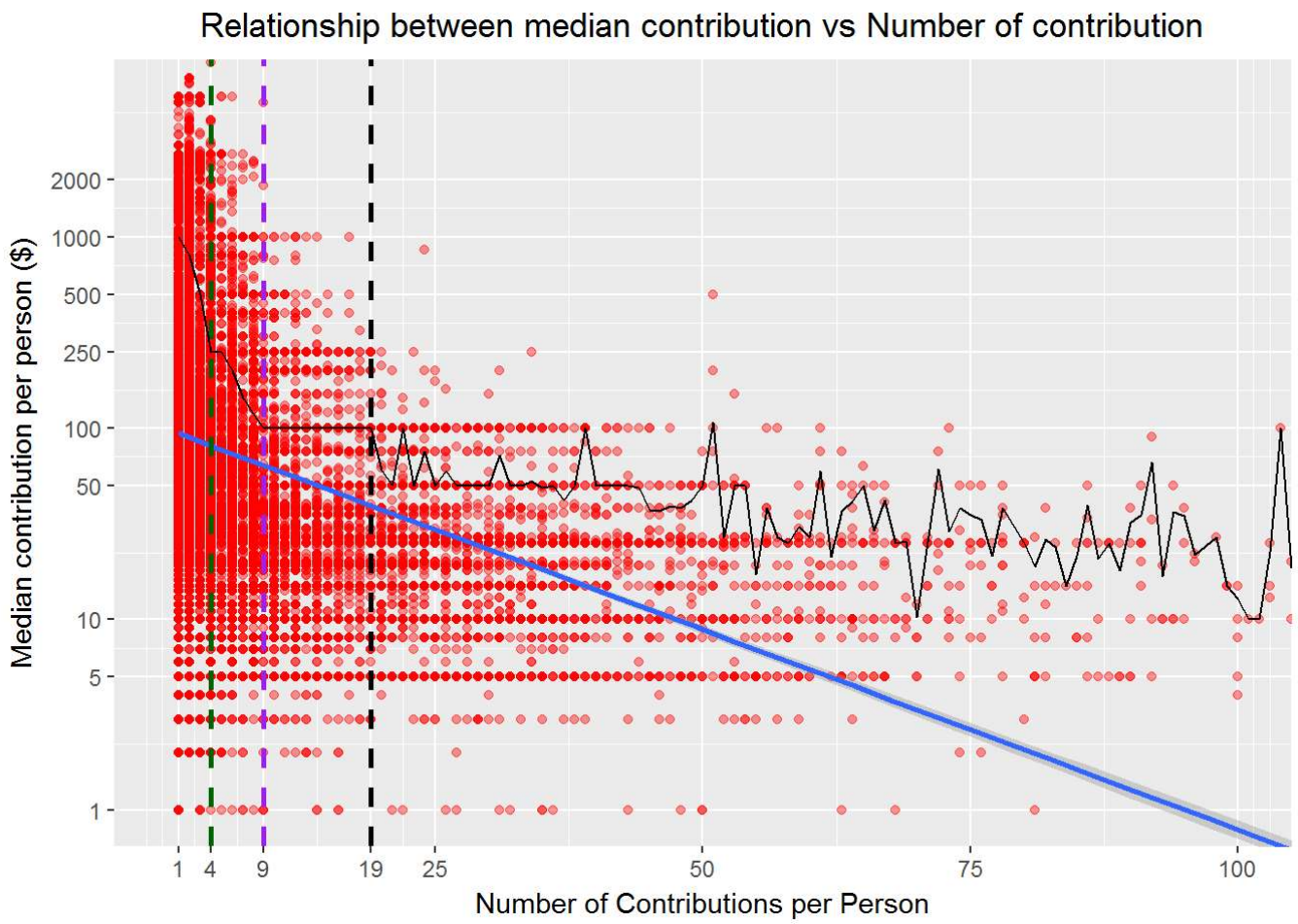
From the map, there seems to be a relation between the location type and the kind of candidate they tended to contribute towards between Mrs. Clinton and Donald Trump.

It does seem that people from the bigger cities tended to contribute towards Hillary Clinton whereas people from rural areas tended to contribute towards Donald Trump.

---

# Final Plots and Summary

## Plot One




# Description One

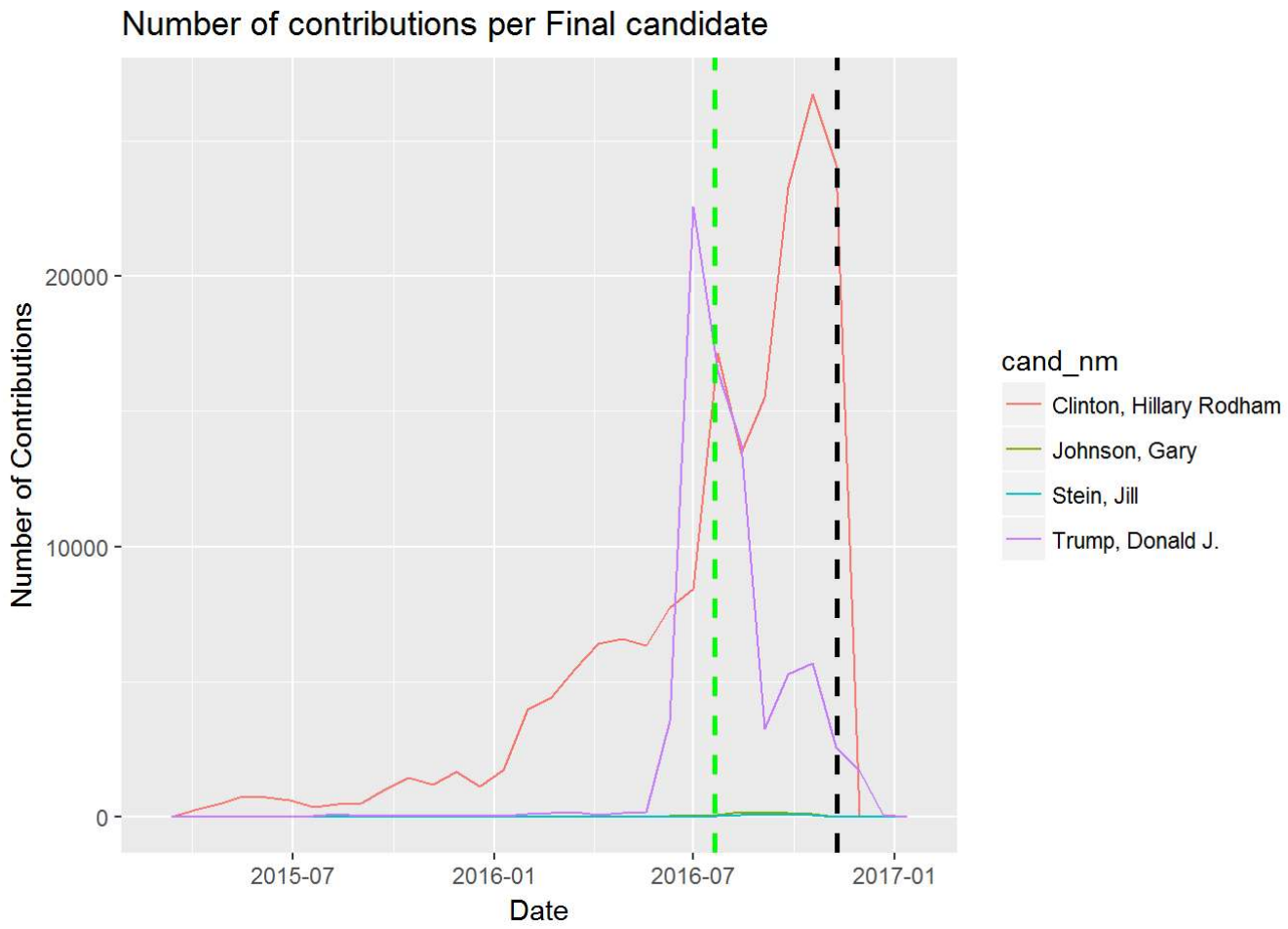
This plot is important because it helps us understand better how people tend to contribute. Initially I assumed that more the number of contributions, greater the total amount of money conributed. But in this case there seems to be a trend opposite of that.

In this dataset, people tended to give larger donations, but few in number, but it in many cases it was larger that the sum of the smaller donations.

The relation is not as strong as I had hoped for as given below.

```
## [1] -0.1178756
```

# Plot Two




# Description Two

This graph is very important as it highlights the final presdential candidates and shows how contribution patterns change with time. Its clear to see from this graph that Hillary Clinton steadily increased her contributions over time, whereas Donald Trump started receiving significant contributions much later.

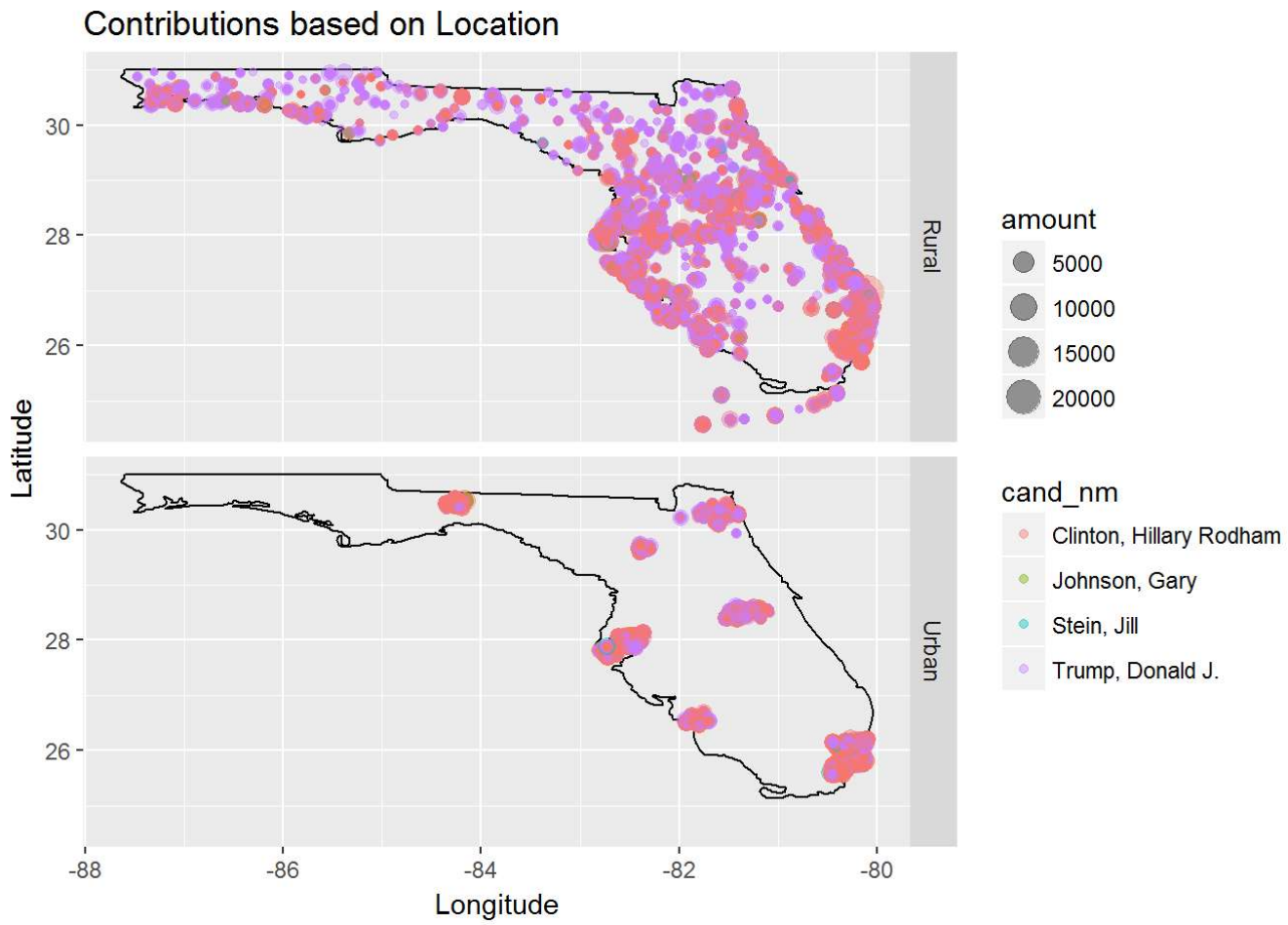
The table below highlights a few important points.

```
## # A tibble: 4 × 7
##                   cand_nm       total     mean median number_contb
##                    <fctr>       <dbl>    <dbl>  <dbl>        <int>
## 1 Clinton, Hillary Rodham 22323771.98 122.6298  25.00       182042
## 2        Trump, Donald J. 12913128.26 169.0864  74.37        76370
## 3           Johnson, Gary   227256.65 288.7632 100.00          787
## 4             Stein, Jill    95726.22 205.8628  50.00          465
## # ... with 2 more variables: percent_number_contb <dbl>,
## #   percent_contb <dbl>
```

Hillary Clinton is the top row and Donald Trump is the second.

It shows how Hillary Clinton had a much higher number of contributions(70.1%) of all contributions) among the final presidential candidates as well the highest amount in contributions(62.77%), which makes it all the more impressive that Donald Trump won that state, both in the primaries and in the presidential elections.

# Plot Three




# Description Three

This plot shows a trend I was really interested in looking at. It shows contribution trends between the two major presidential candidates based on location type.

The urban map shows a clear difference in contributions between Donald Trump and Hillary Clinton, in that Hillary Clinton seems to be getting the lion's share of the contributions.

The difference is not so clear in the rural map. It seems like Hillary Clinton has many more contributions in number, but it is important to remember from this representation and earlier ones, that Hillary Clinton had a majority of smaller sized donations. The table below will give us a better understanding of the map.

```
## Source: local data frame [4 x 6]
## Groups: cand_nm [2]
##
##                  cand_nm location_type    total     mean median      n
##                   <fctr>         <chr>    <dbl>    <dbl>  <dbl>  <int>
## 1 Clinton, Hillary Rodham         Rural 11941191 107.8474  25.00 110723
## 2 Clinton, Hillary Rodham         Urban 10382581 145.5795  25.00  71319
## 3        Trump, Donald J.         Rural  9944754 166.2030  68.37  59835
## 4        Trump, Donald J.         Urban  2968374 179.5207  80.00  16535
```

As we can see, there is clear daylight in urban areas where Hillary Clinton has many more donations(71,319) than Donald Trump(16,535). In fact, she has almost 4 times the number of contributions. She is also beating him in terms of total amount of contribution ($11,941,191 vs $2,968,374).

The rural areas are much more interesting. Here again, Hillary Clinton has many more contribtuions (110,723) than Donald Trump (59,835). But the difference isnt as much as in the urban case.

# Conclusion and Reflection:

1. I couldn't build a linear model for this data set, as there weren't many parameters by which I could've created quantitative relationships. Age data or salary range data could've helped understand this dataset better.
2. The top two candidates in the election were also the candidates that received the most amount of money. However, there isn't emough data to confidently suggest that cadidates who received more contributions did better than candidates who didn't.
3. A good estimate of how much each candidate/contributor received/contributed is the median of the values. Most of the data is highly skewed with the presence of a few outliers.
4. A method of improving this analysis in the future would be to add the votes received by each candidate in both, the primaries and the presidential elections. Then, we could check to see if contributions, in terms of their number or size, had any effect on how people voted for the candidates.