An Analysis of Financial Contributions to Presidential Campaigns in Illinois with R - By Tiffany Li

In this analysis, I will explore the financial contribution data in the state of Illinois for the 2016 presidential campaigns, using the the R language. The analysis involves multiple variables such as candidate's name and political party, contributor's city, employer and occupation as well as the contribution amount. Data wrangling techniques are used to process and clean the data before the data exploration process. In the end, visualizations such as bar plot and boxplots reveal some interesting findings regarding these financial contributions in this critical presidential race.

```
# Load the Data
setwd("~/Desktop/Udacity Nanodegree/P4/data set")
getwd()
```

```
## [1] "/Users/xiangli/Desktop/Udacity Nanodegree/P4/data set"
```

```
PC <- read.csv('P1-IL-2.csv', header = TRUE, fill = TRUE)
```

Now that we have successfully loaded the dataset and packages, we can exam the structure of the data.

Also, the description of variables can be found in this link: ftp://ftp.fec.gov/FEC/Presidential_Map/2016/DATA_DICTIONARIES/CONTRIBUTOR_FORMAT.txt (ftp://ftp.fec.gov/FEC/Presidential_Map/2016/DATA_DICTIONARIES/CONTRIBUTOR_FORMAT.txt)

```
# Exam the dataset
str(PC)
```

```
## 'data.frame':    65499 obs. of  19 variables:
##  $ index            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ cmte_id          : Factor w/ 21 levels "C00458844","C00573519",..: 3 5 6 3 6 3 6 6 3 5 ...
##  $ cand_id          : Factor w/ 21 levels "P00003392","P20002671",..: 9 1 11 9 11 9 11 11 9 1
 ...
##  $ cand_nm          : Factor w/ 21 levels "Bush, Jeb","Carson, Benjamin S.",..: 5 4 17 5 17 5 1
7 17 5 4 ...
##  $ contbr_nm        : Factor w/ 21639 levels "AARDEMA, JOHN",..: 9036 6137 11280 10446 11705 12
401 8966 8966 12401 4664 ...
##  $ contbr_city      : Factor w/ 980 levels "ABINGDON","ADAIR",..: 755 459 161 256 696 779 161 1
61 779 132 ...
##  $ contbr_st        : Factor w/ 1 level "IL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contbr_zip       : int  605466525 601564641 606223057 624263133 615542542 600738114 60657622
5 606576225 600738114 622311258 ...
##  $ contbr_employer  : Factor w/ 8580 levels "","--","[BLANK",..: 6699 6665 8546 6281 3374 89 54
38 5438 89 3743 ...
##  $ contbr_occupation: Factor w/ 4574 levels "","  CERTIFIED REGISTERED NURSE ANESTHETIS",..: 256
4 3593 4504 3473 3353 4042 2695 2695 4042 1975 ...
##  $ contb_receipt_amt: Factor w/ 2035 levels "-1,000","-1,040",..: 1492 1208 1492 239 571 230 14
82 991 1492 976 ...
##  $ contb_receipt_dt : Factor w/ 561 levels "01-APR-15","01-APR-16",..: 532 296 101 494 83 313 1
01 101 219 476 ...
##  $ receipt_desc     : Factor w/ 20 levels "","* EARMARKED CONTRIBUTION: SEE BELOW REATTRIBUTIO
N/REFUND PENDING",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_cd          : Factor w/ 2 levels "","X": 1 2 1 1 1 1 1 1 1 2 ...
##  $ memo_text        : Factor w/ 62 levels "","*","* EARMARKED CONTRIBUTION: SEE BELOW",..: 1 8
 3 1 3 1 3 3 1 8 ...
##  $ form_tp          : Factor w/ 3 levels "SA17A","SA18",..: 1 2 1 1 1 1 1 1 1 2 ...
##  $ file_num         : int  1077664 1091718 1077404 1077664 1077404 1077664 1077404 1077404 1077
664 1091718 ...
##  $ tran_id          : Factor w/ 65464 levels "A003EA03DA61F49D7A4E",..: 36972 24381 50338 36835
 50083 36504 50354 50432 37328 24538 ...
##  $ election_tp      : Factor w/ 3 levels "","G2016","P2016": 3 3 3 3 3 3 3 3 3 3 ...
```

By looking at the data types in the str() function, I detected 6 columns whose data types need to be converted. 'index' is just an id so we should change it from integer to character. 'contbr_zip' represents the zipcode of contributor so it should be factor instead of integer. 'contb_receipt_amt' is the amount contributed therefore needs to be represented in numeric form instead of factor. 'contb_receipt_dt' is the date when contribution was received and should be represented in date form. 'file_num' is the file number and should be converted from integer to factor. Lastly, I think character is a better way than factor to represent 'tran_id', which is transaction ID and has unique value for each row.

```
# Convert data types if neccessary
PC$index <- as.character(PC$index)
PC$contbr_zip <- as.factor(PC$contbr_zip)
PC$contb_receipt_amt <- as.numeric(as.character(PC$contb_receipt_amt))
PC$contb_receipt_dt <- as.Date(PC$contb_receipt_dt, format = "%d-%b-%y")
PC$file_num <- as.factor(PC$file_num)
PC$tran_id <- as.character(PC$tran_id)
```

Now that we have all the correct data types, we can begin the analysis.

There are some candidates with very few contributions compared to others. As a result, plots or summary statistics for these candidate might be misleading because of the small data size. Therefore I will exclude these candidate for the rest of the analysis.

```
# Exclude candidates with fewer than 10 contributions
table(PC$cand_nm)
```

```
##
##                 Bush, Jeb       Carson, Benjamin S.
##                       314                       326
##  Christie, Christopher J.   Clinton, Hillary Rodham
##                        21                     28118
## Cruz, Rafael Edward 'Ted'            Fiorina, Carly
##                      6195                        89
##        Graham, Lindsey O.            Huckabee, Mike
##                         3                       161
##             Jindal, Bobby             Johnson, Gary
##                         3                       145
##           Kasich, John R.          Lessig, Lawrence
##                       244                        65
##            McMullin, Evan   O'Malley, Martin Joseph
##                         7                         1
##                Paul, Rand              Rubio, Marco
##                       948                       752
##          Sanders, Bernard      Santorum, Richard J.
##                     26032                         3
##               Stein, Jill          Trump, Donald J.
##                        63                      1628
##             Walker, Scott
##                       381
```

```
PC <- subset(PC, !(PC$cand_nm == "O'Malley, Martin Joseph" |
                     PC$cand_nm == "Graham, Lindsey O." |
                     PC$cand_nm == "Jindal, Bobby" |
                     PC$cand_nm == "Santorum, Richard J." |
                     PC$cand_nm == "McMullin, Evan"))
```

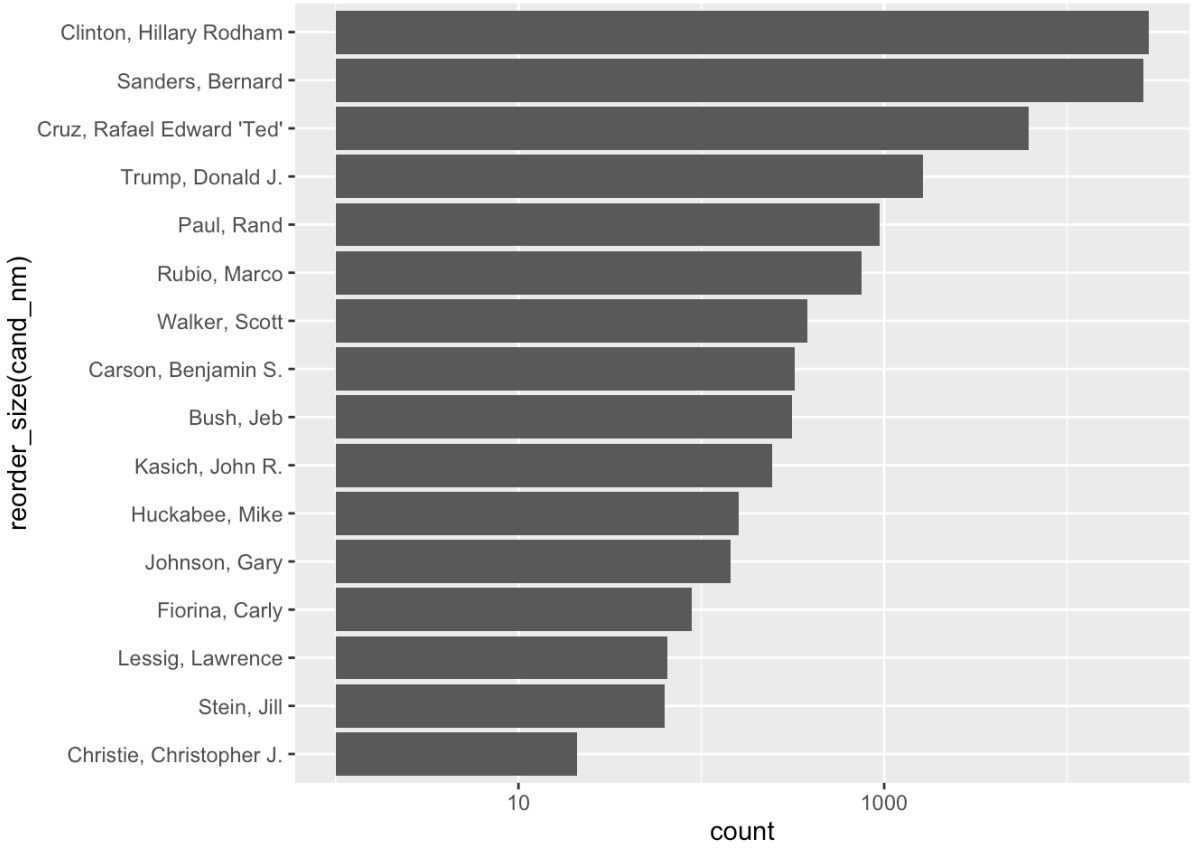
It is also perhaps helpful to add the information about the political parties associated with the candidates. Therefore I will create a new variable for this purpose.

```
# Create a new variable for candidates' associated political parties by creating a new function rn
Party and applying that function
Democratic <- c('Clinton, Hillary Rodham',
                'Lessig, Lawrence',
                'Sanders, Bernard')

rnParty <- function(x) {
if(is.na(x)){
  NA
}else if(x %in% Democratic){
  'D'
}else if(x == 'Johnson, Gary'){
  'L'
}else if(x == 'Stein, Jill'){
  'G'
}else{
  'R'
}
}

PC$cand_nm_party <- apply(PC['cand_nm'],1,rnParty)
```

Univariate Plots Section

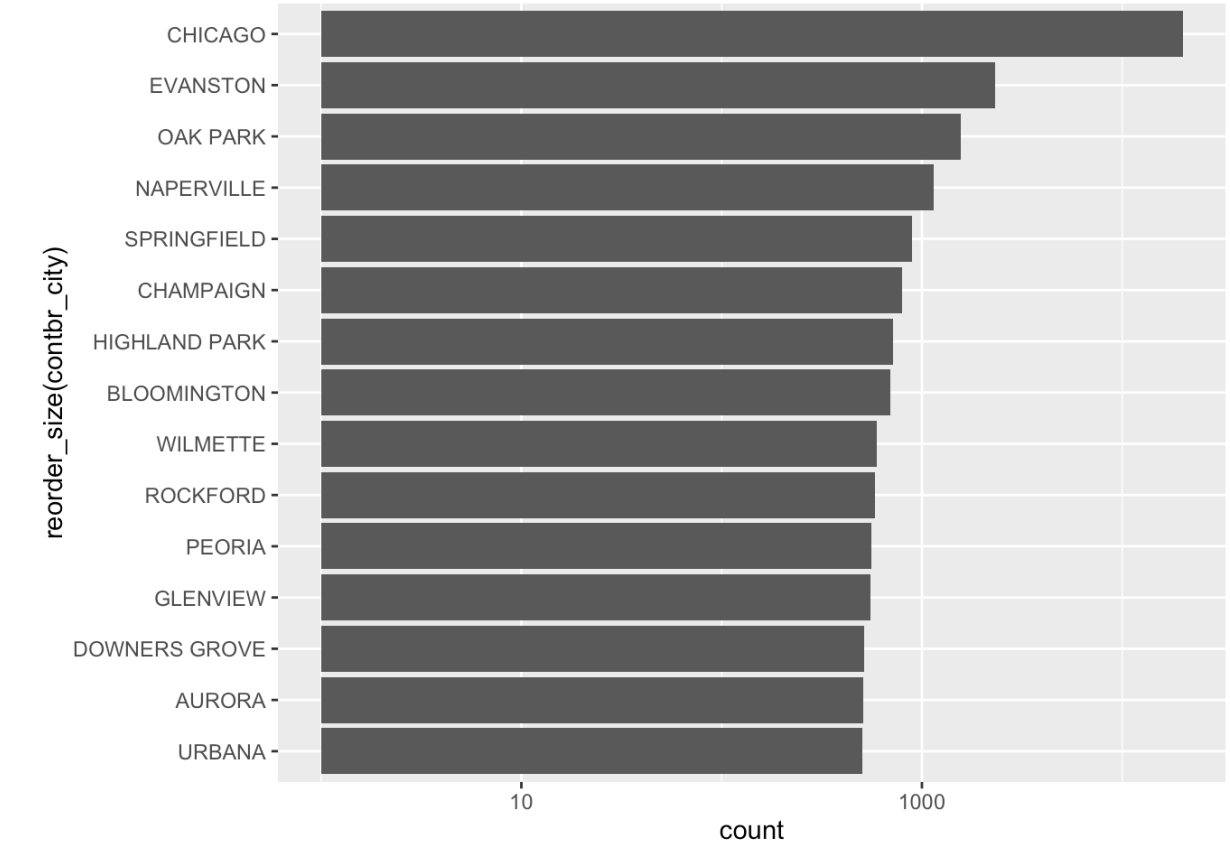
Let's understand the data using visualizations of single variable.

How many contributions (by transaction) did each candidate receive?



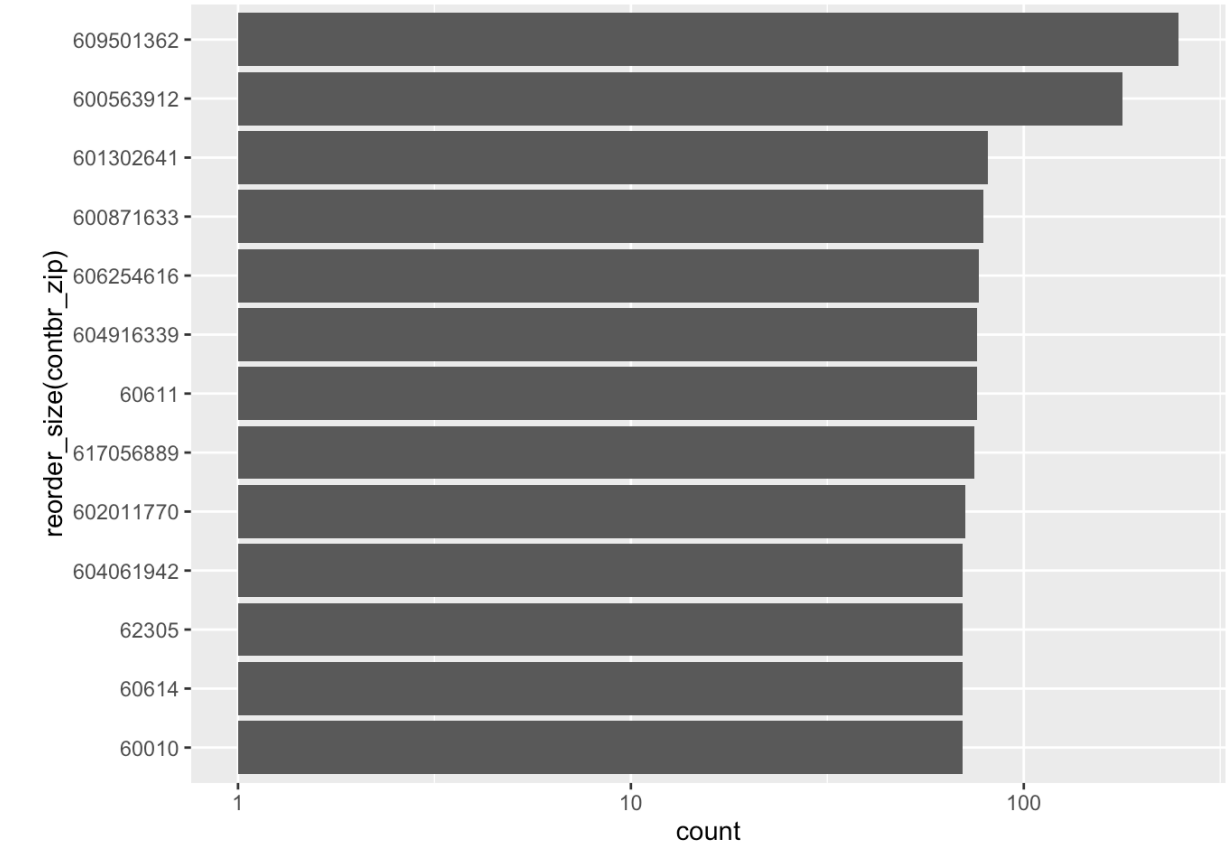
In terms of contribution transactions, Clinton received the most contributions, followed by Sanders, Cruz and Trump. Notice that we added a log10 scale to make the plot easier to read because the results vary significantly across candidates.

How many contribution transactions are from each city in Illinois (only top results shown)?



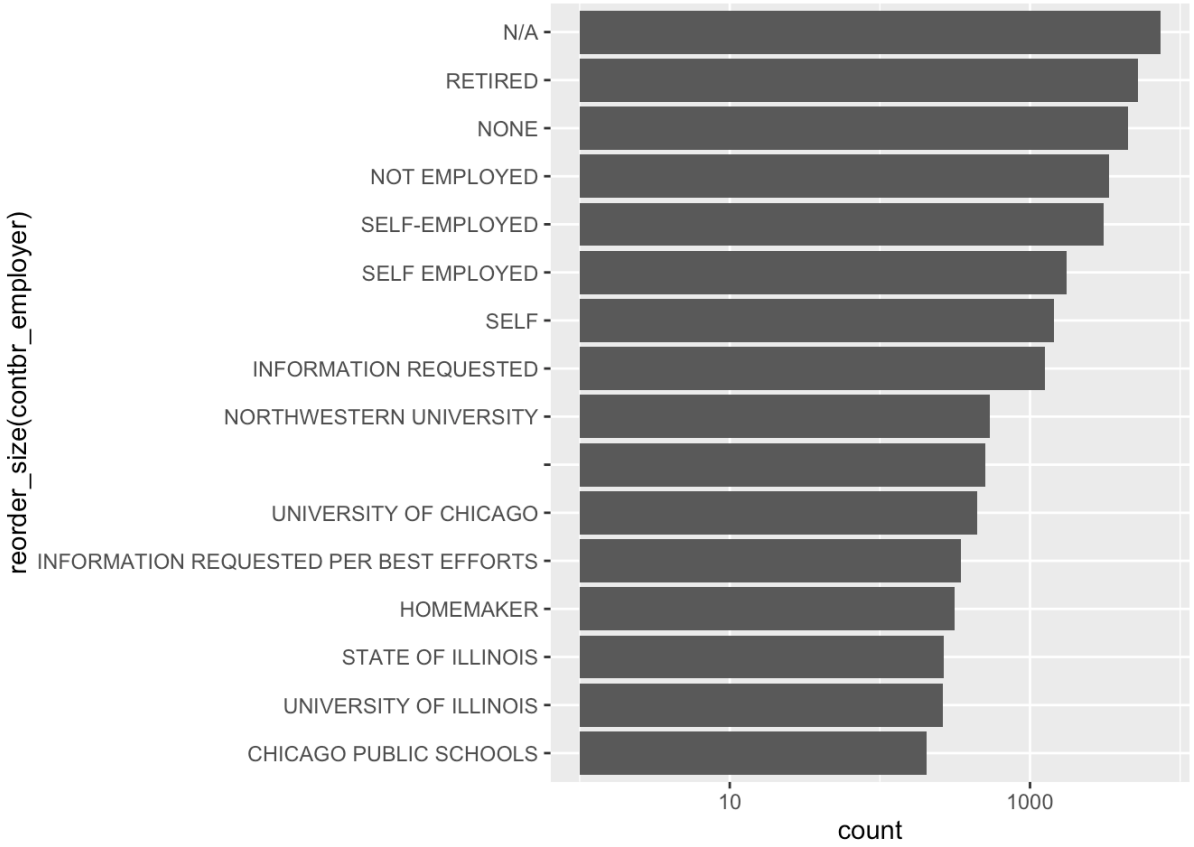
We can observe that Chicago has the most contributions, followed by Evanston and Oak Park.

How many contribution transactions are from each zipcode in Illinois (only top results shown)?



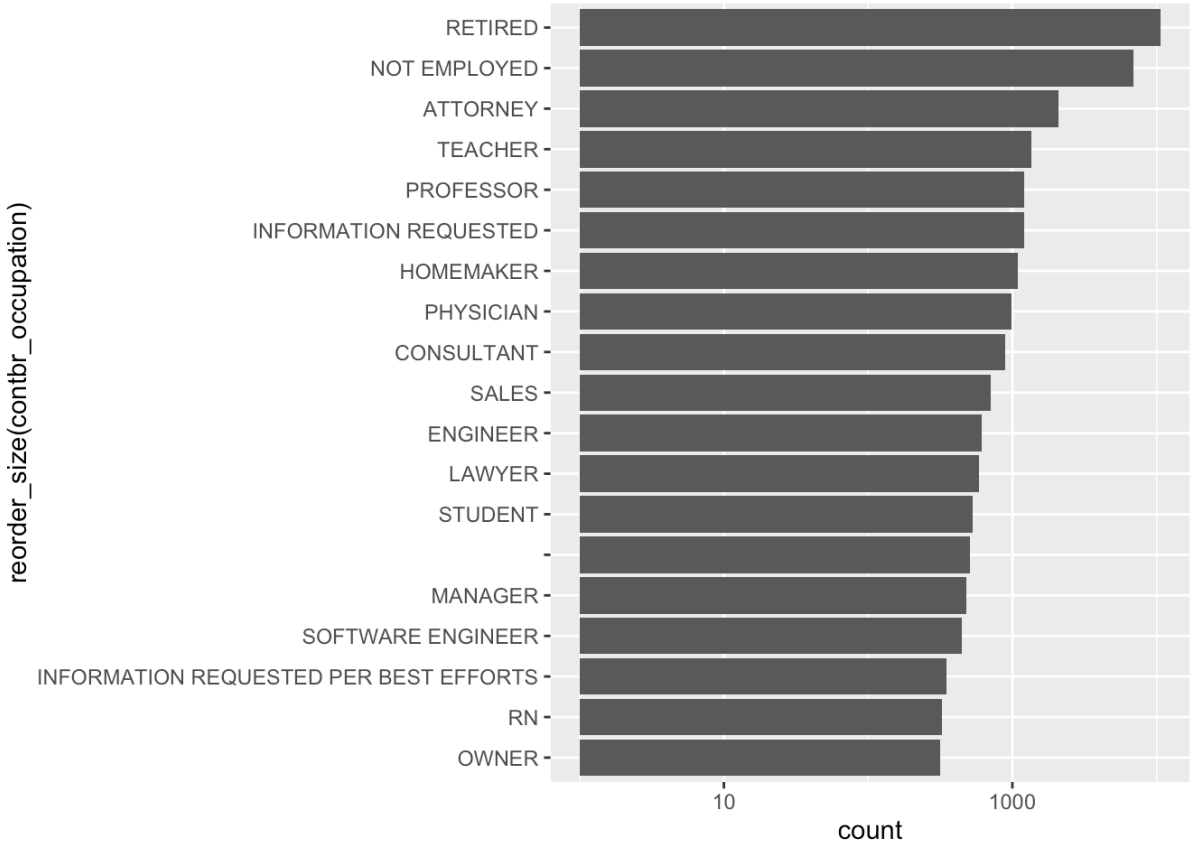
As we can see from the plot, the top three zipcodes with the most contributions are 609501362, 600563912, 601302641.

How many contribution transactions are from each contributor employer (only top results shown)?



We can observe that the employer category with the most contribution (excluding N/A) is retired, followed by none and not employed. Universities such as Northwestern University and University of Chicago also are top in the list.

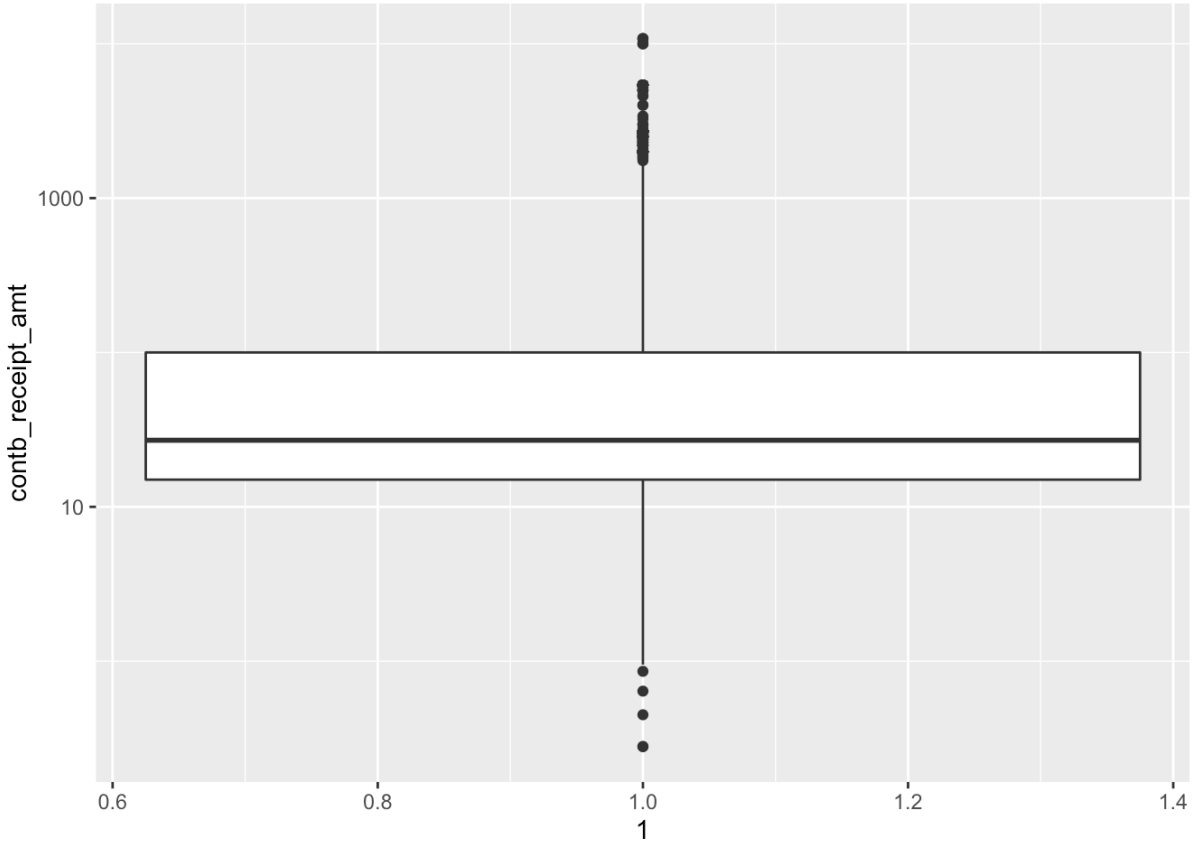
How many contribution transactions are from each occupation (only top results shown)?



Retired shows up as the top occupations, followed by not employed, attorney, teacher and professor.

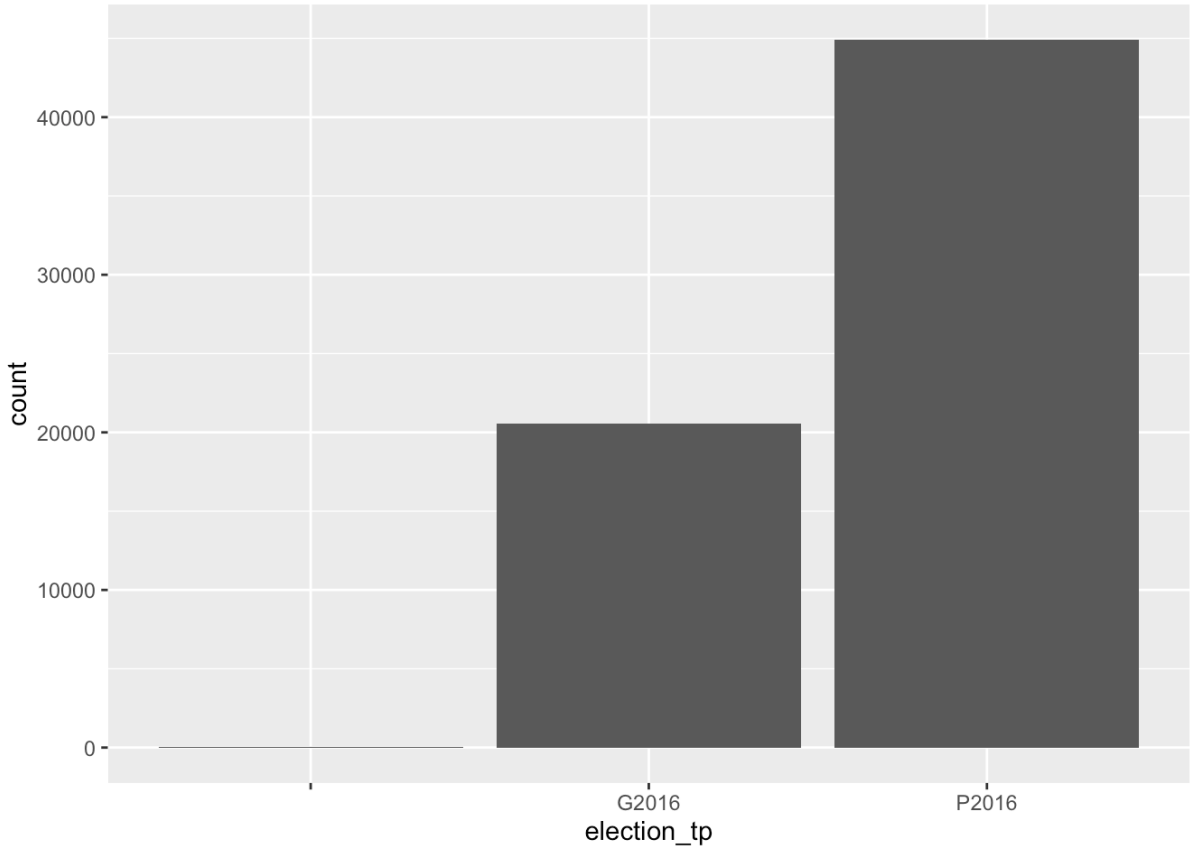
What is first quantile, median and third quantile of the contribution amount?

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -990.0    15.0    27.0   120.1   100.0 10800.0     226
```



The first quantile of contribution is $15, median is $27 and the third quantile is $100.

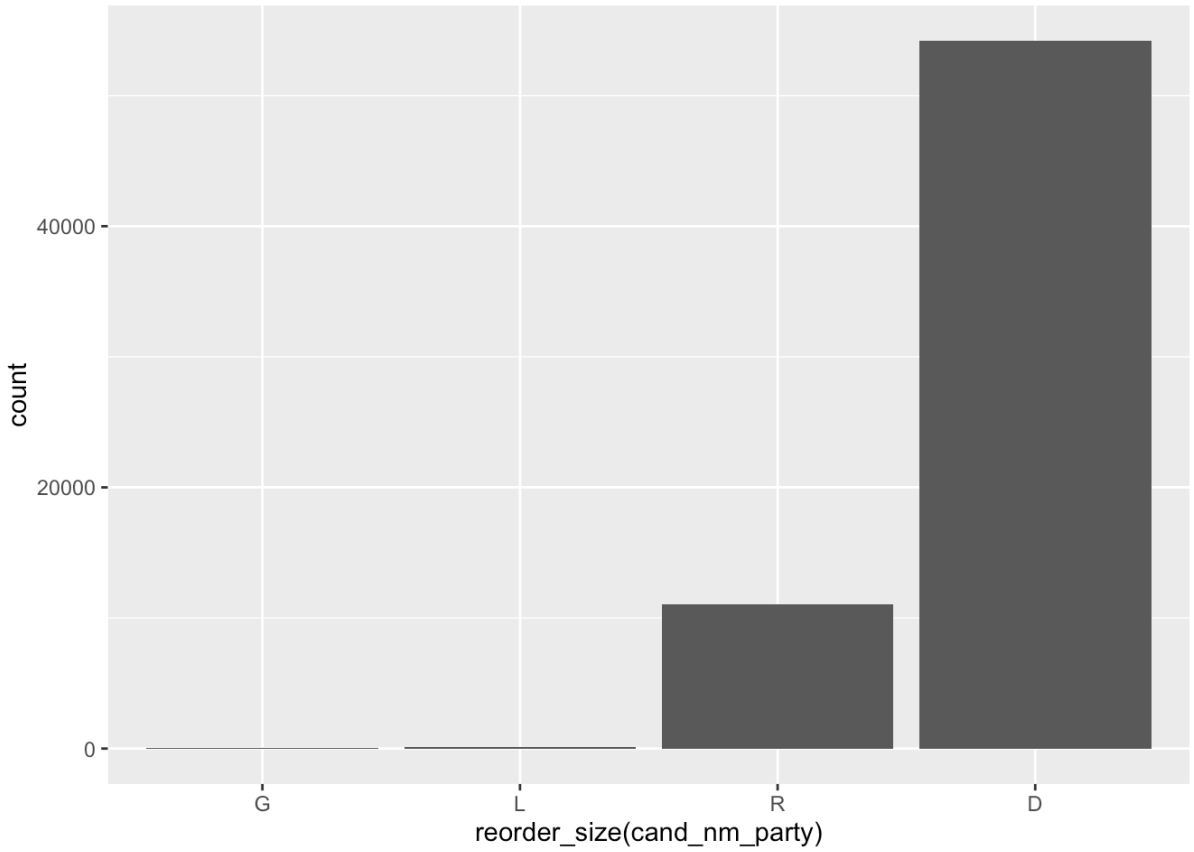
How many contributions were made to primary election and general election?



We can see that 44,923 contributions were made to primary elections and 20,545 contributions were made to general elections. It is apparent that more contributions were made to primary

elections than general elections.

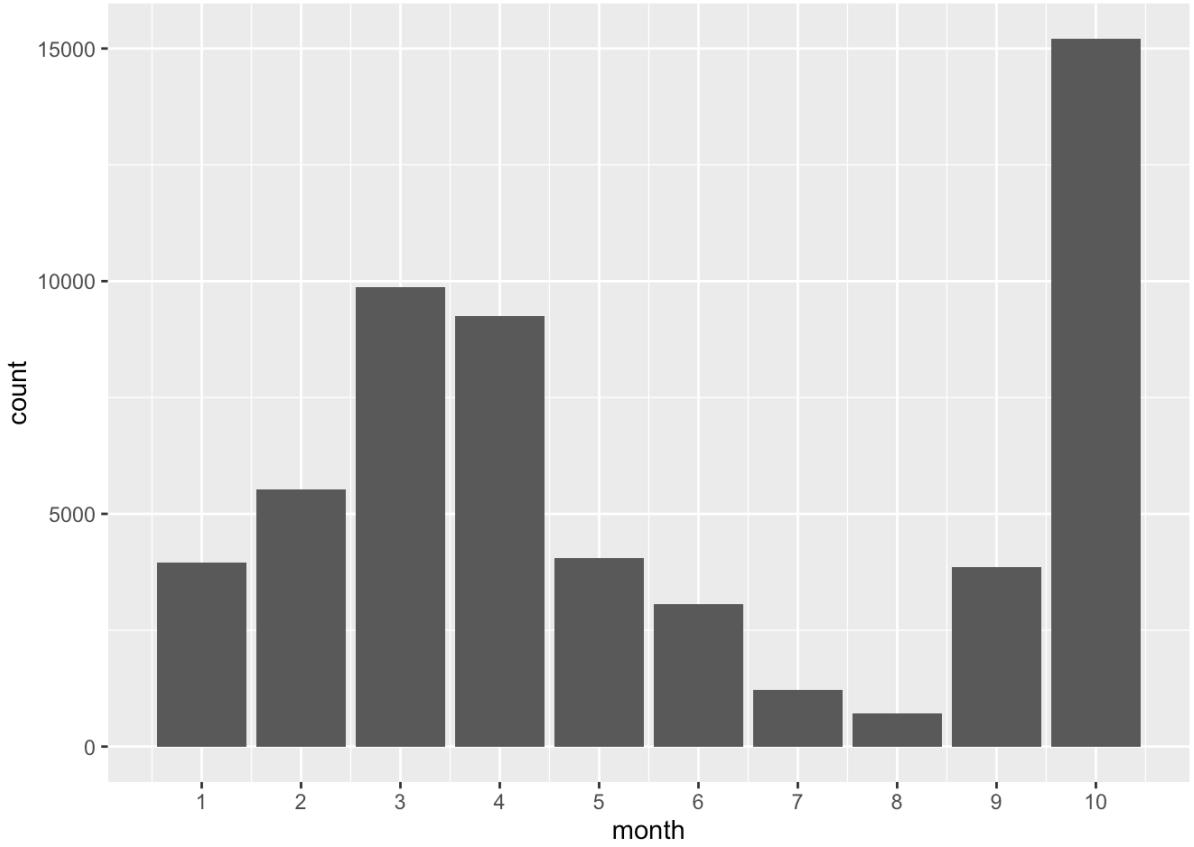
How many contributions were made to different political parties?



We can see that the Democratic party, despite having fewer candidates, received the most contributions, followed by the Republican party. Green party and libertarian party received much fewer contributions by comparison.

Finally, an interesting thing to look at is the date when contribution was made. We can detect trends of when people make more contributions in the election cycle.

To simplify the date variable, we can first extract out the year, month and day information from the contb_receipt_dt variable. Next, we create a plot using the month (only including 2016 data) to observe any trend for contributions over time.



From this graph, it is clear that there is a trend in contributions made over time in 2016. The contributions started small in January and peaked in March. Then the contributions decreased and started increasing again in September, which led to the maximum in October. This makes sense because many candidates dropped out of the race in March, and general election was taken place in November so many contributors made donations in October hoping to make an impact at a critical time.

Univariate Analysis

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

I am mainly interested in the following features: candidate name and contribution receipt amount, since the most interesting topic is analyzing the amount of contributions received by candidates.

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

Features like contributor's city, contributor's zipcode, contributor's employer, contributor's occupation, contribution date, election type, candidate's political party can provide further insights of the dataset. We can find relationships between these features and the features I mentioned in the last question. For example, we can try to determine if candidate's political party has an effect on contribution amount.

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

I created a new variable cand_nm_party, which is the candidate's party information(e.g: Democrat, Republican etc) in order to understand the differences among contributions to

different parties.

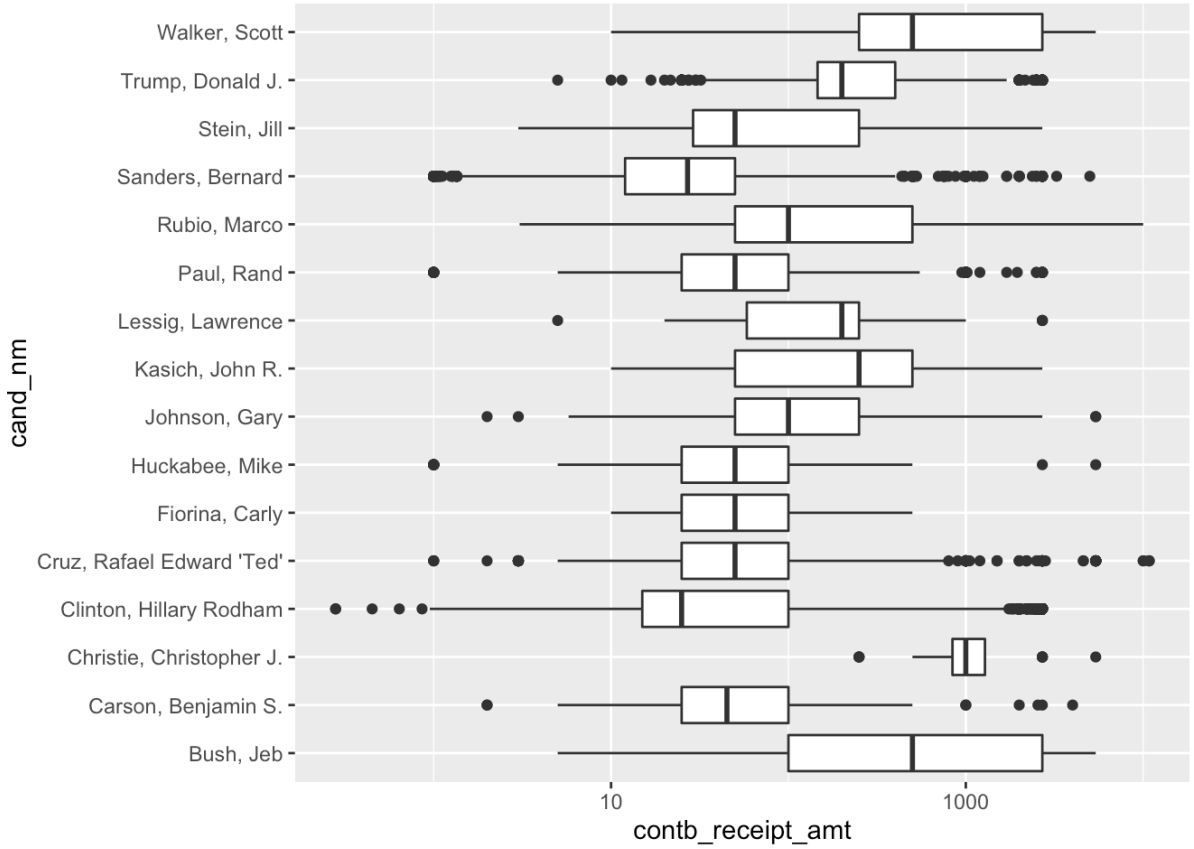
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

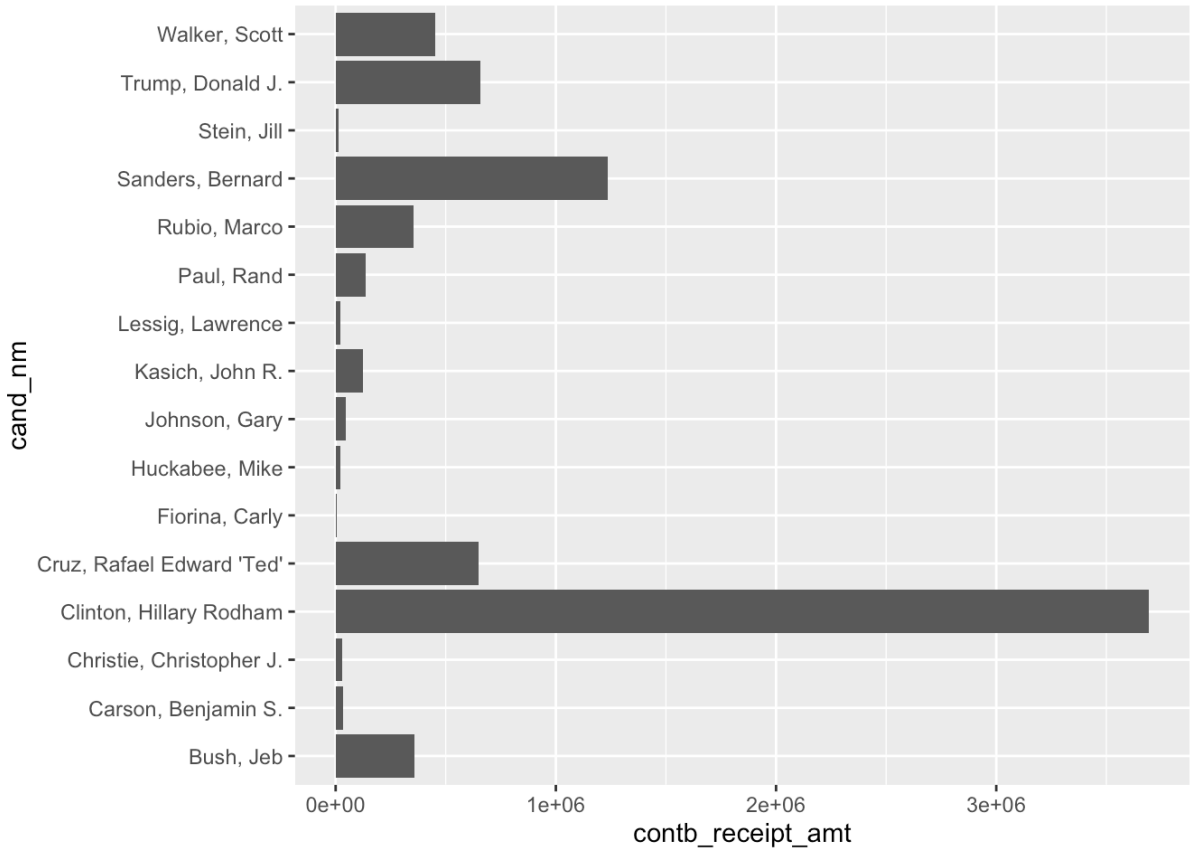
I noticed that the contributions come from many cities but many of them concentrate on one city - Chicago. Similar situation happens to occupation. Additionally, the contributions received by candidates vary significantly, therefore I added a log transformation to make the plot easier to read.

Bivariate Plots Section

We can investigate the dataset by looking at two variables.

First, let's find out the amount of contributions received by each candidate.



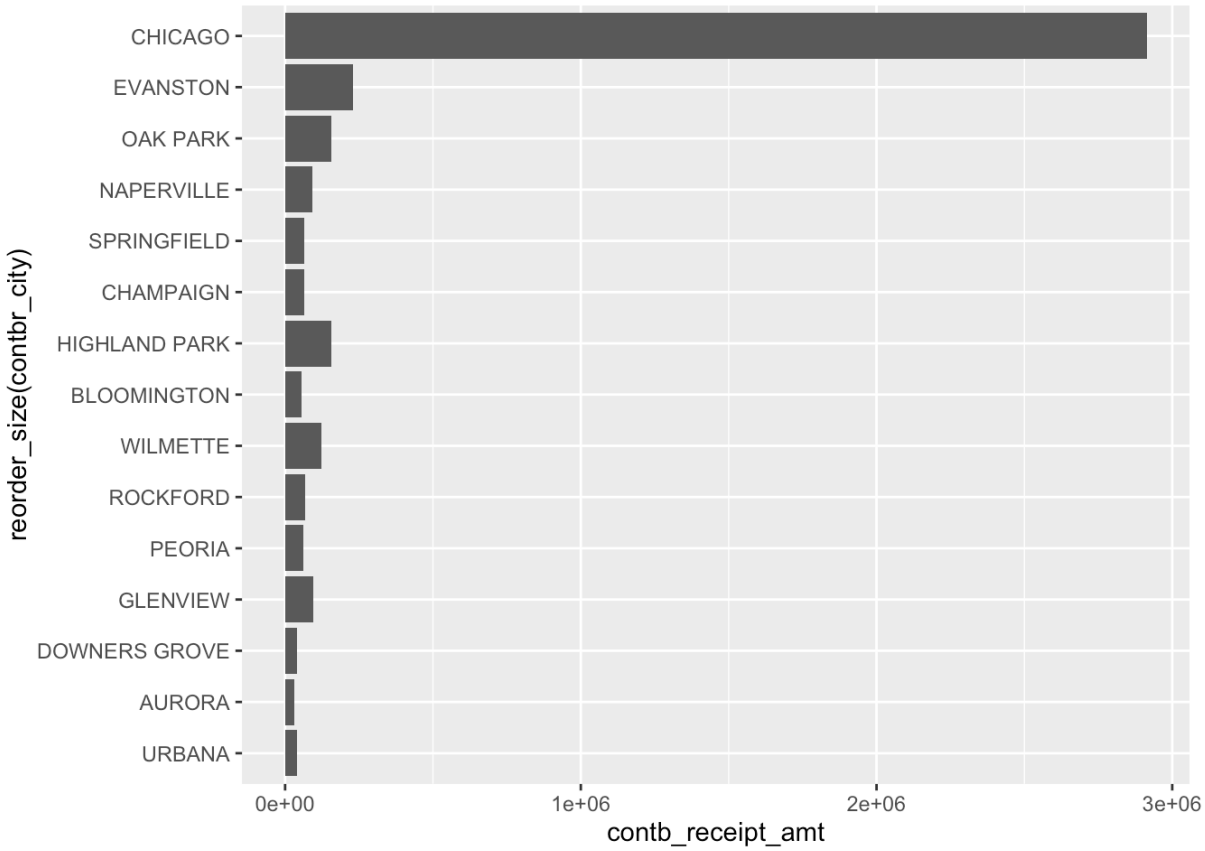


The boxplots of contribution amount are useful, as they show show 25%, 50% and 75% percentile of the contribution amount for each candidate, along with outliers. Christie had the highest median contribution amount and Clinton had the lowest median contribution amount.

Similar to the univariate plot of contribution transactions of each candidate, we are seeing that Clinton received the highest total amount of contribution, followed by Sanders, Cruz and

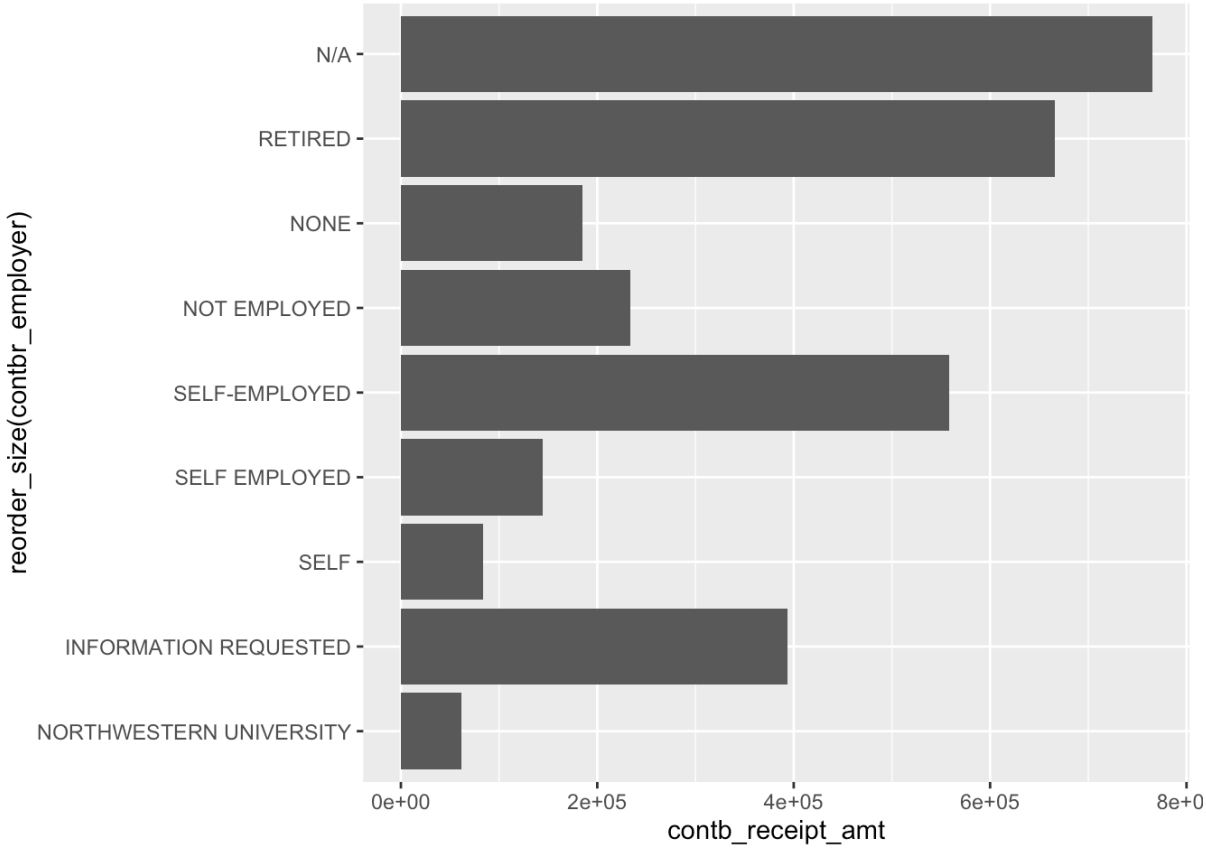
Trump.

What are the total amounts of contribution from each city in Illinois (only top results shown)?



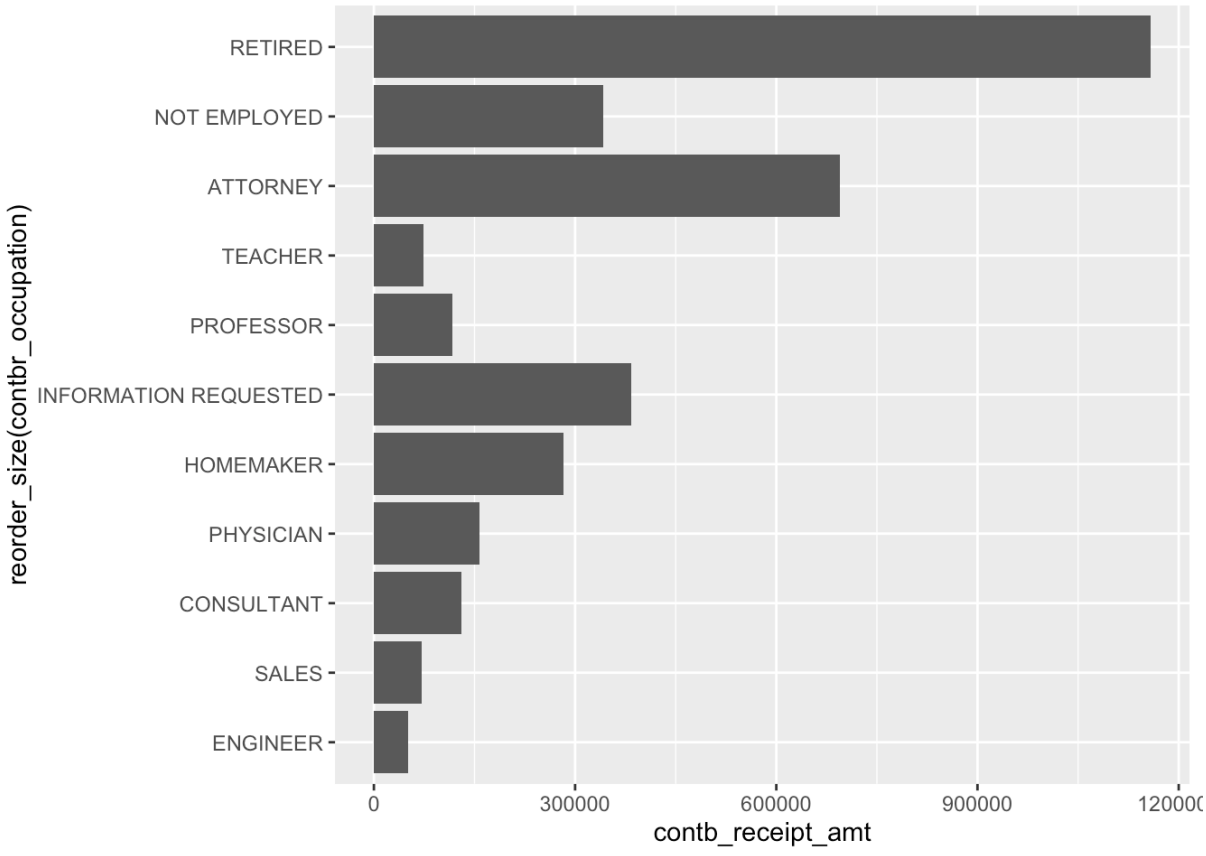
Consistent with our intuition, Chicago has the highest contribution amount since the population is much larger, followed by Evanston, where Northwestern University is located, and Oak Park.

What are the total amounts of contribution from each employer in Illinois (only top results shown)?



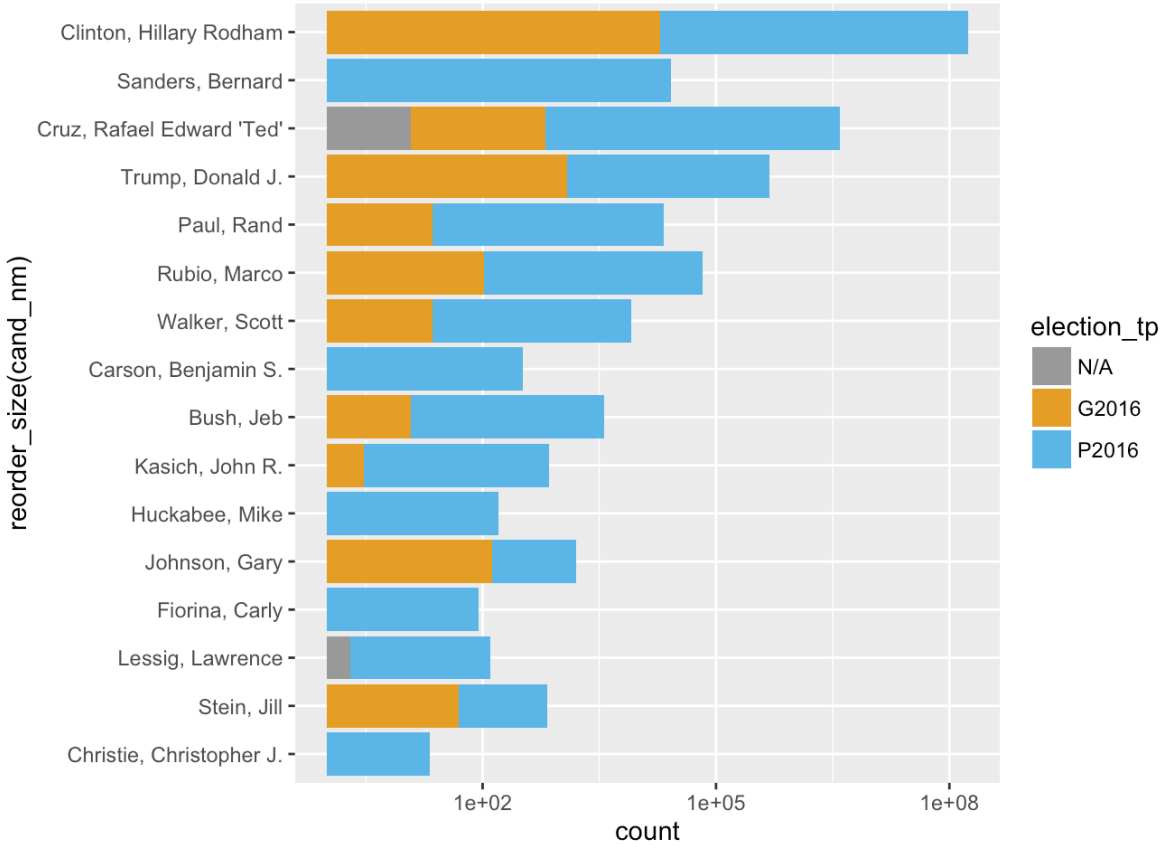
Besides categories like N/A, none, self-employed, contributors from Northwestern University also contributed significant amount.

What are the total amounts of contribution from each occupation in Illinois (only top results shown)?



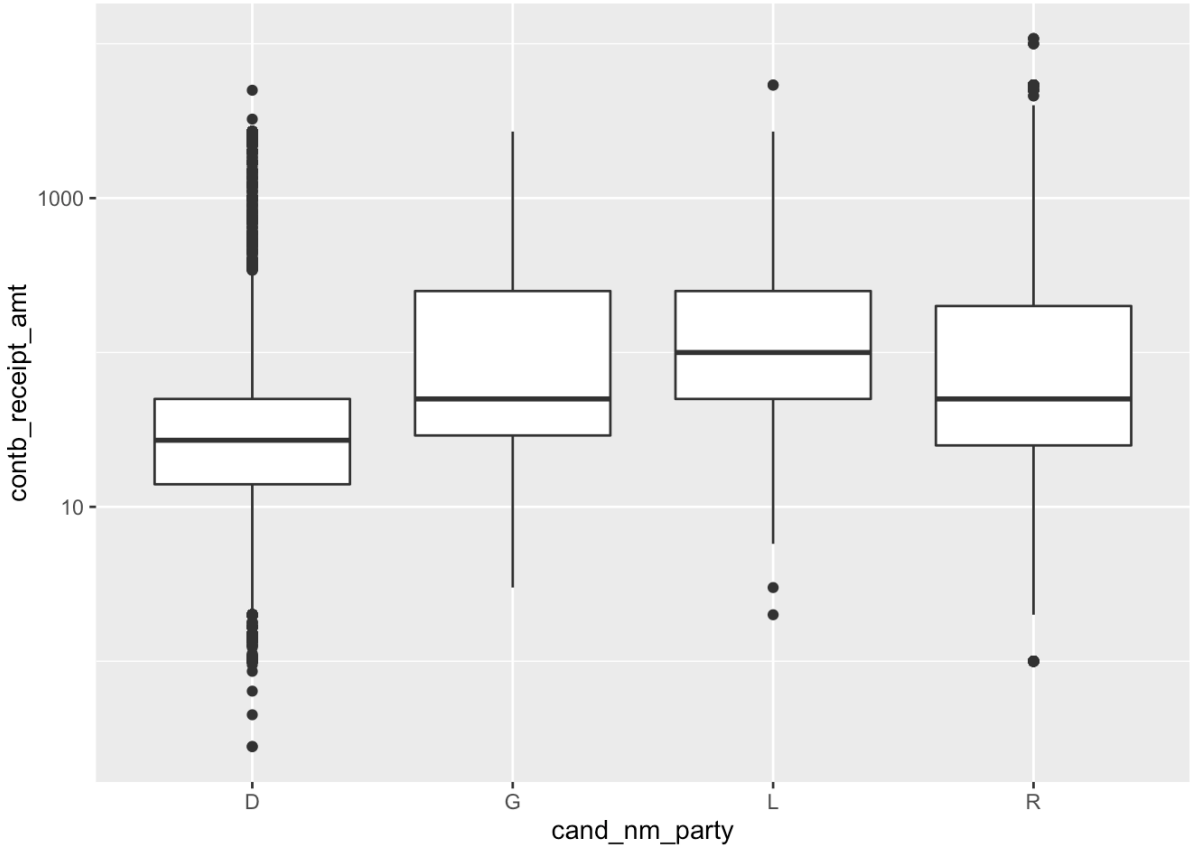
From this graph we can see that the occupation that contributed the most amount is retired, followed by attorney, information requested and not employed.

Another bivariate plot we can create is contributions for each candidate, filled with election type information.



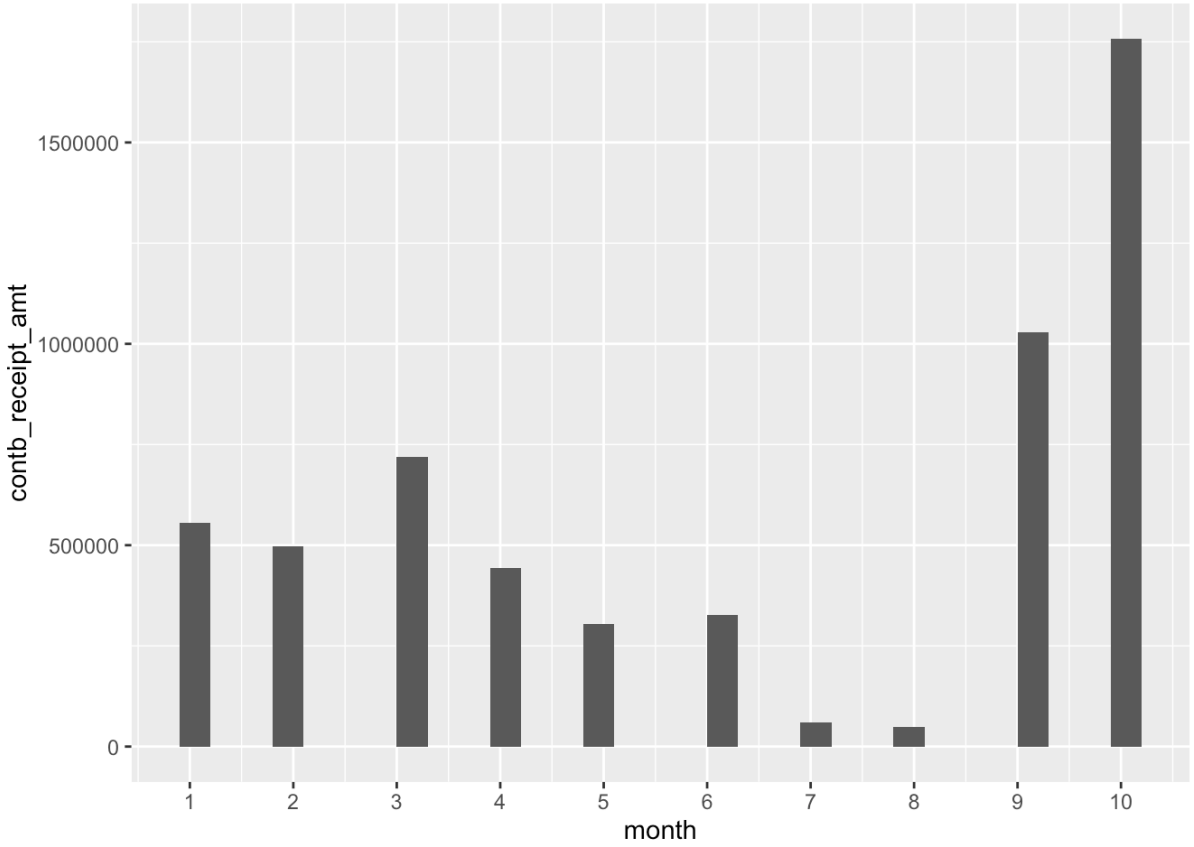
Compared to the previous plot for contributions received by candidates, we obtain additional information about how many contributons were made to primary and general election in this plot.

We can also investigate amount of contributions made to candidates of different political parties using boxplots.



The boxplots show that the Libertarian party had the highest median contribution amount, followed by the Green Party, the Republican party and the Democratic party. However, we must keep in mind that the sample sizes for the Libertarian party and the Green party are much smaller.

Following the analysis with the month variable in univariate analysis, we can continue to explore this variable using total contribution amount.



We see a similar trend compared to what we saw earlier: total amount of contributions moves with the election cycle, with peaks in March and October.

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?

One obvious pattern is that contribution amounts vary significantly across different candidates in terms of total, median, etc. Another pattern is that in general, more contributions go toward primary election rather than general election. More money is dedicated to make the candiates the party nominee perhaps since only one candidate from each party actually competes in the general election. There is also an obvious trend in total contribution amount throughout the election cycle - more contributions were made when more candidates are involved or during critical times in the race.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

The amount of contributions made to candidates of different political party vary quite a lot, as we see in the boxplots. However, even though the candidates from the Democratic party received the lowest median contribution amount, they received the highest number of contributions, which drives up their total amount of contributions received. It is fair to say that Democratic candidates build their campaigns upon small donations in large quantities.

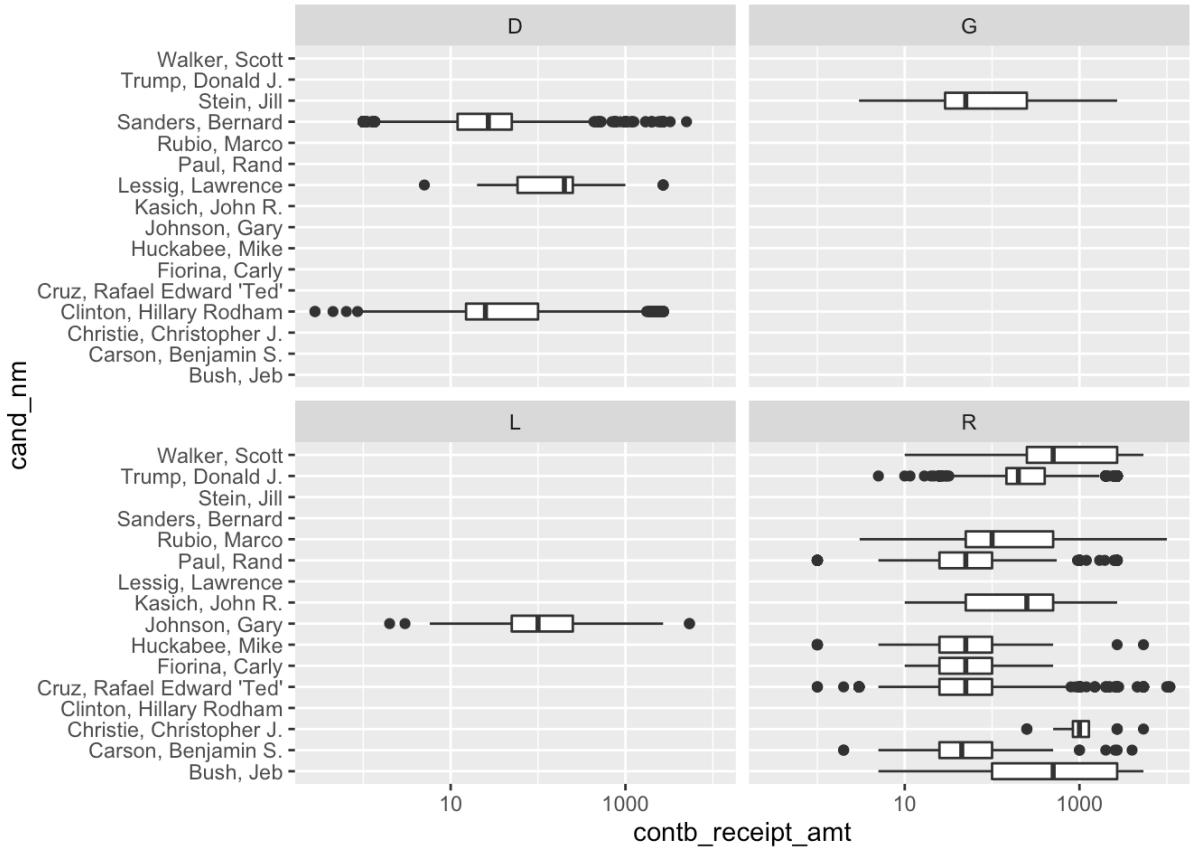
What was the strongest relationship you found?

Since we only have one numeric variable - contribution amount, we cannot find the correlations.

Multivariate Plots Section

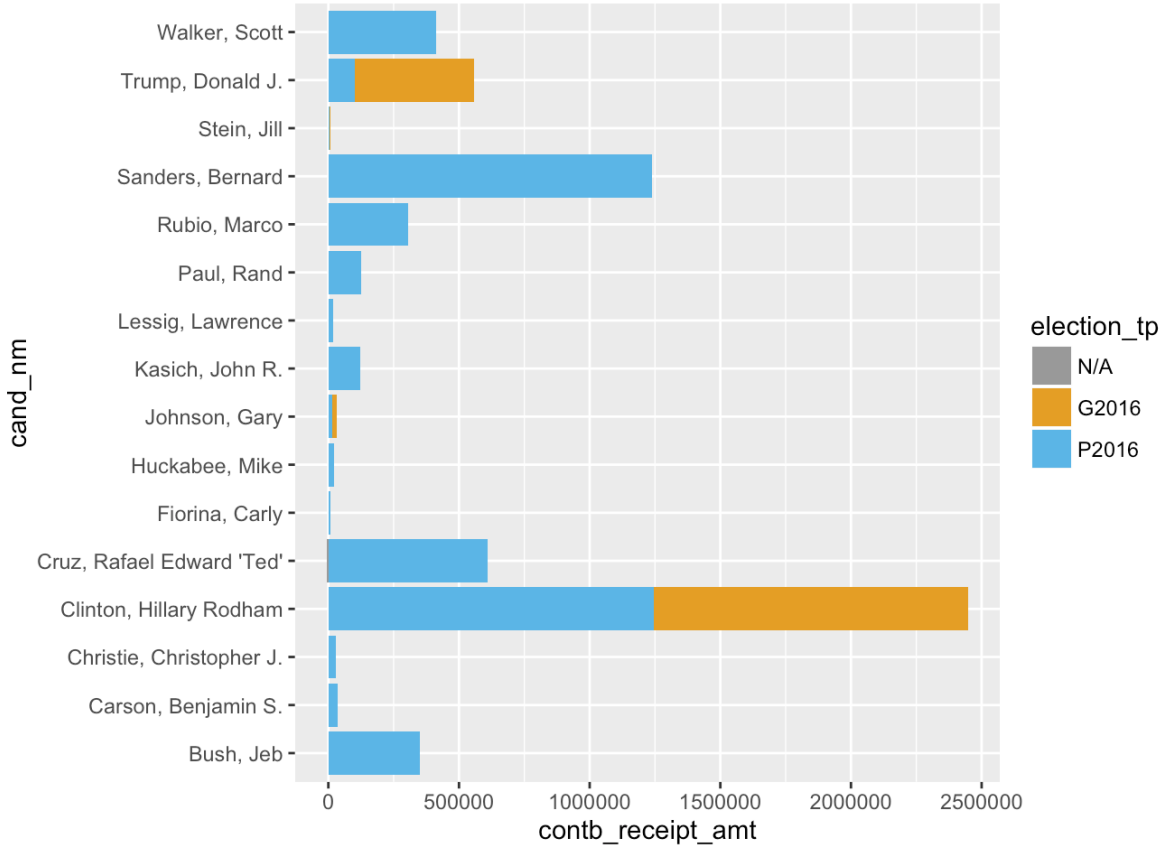
We can also investigate the data set by looking at more than 2 variables

Previously we looked at the amount of contributions received by each candidate, now we can add a facet by political party to the plot.



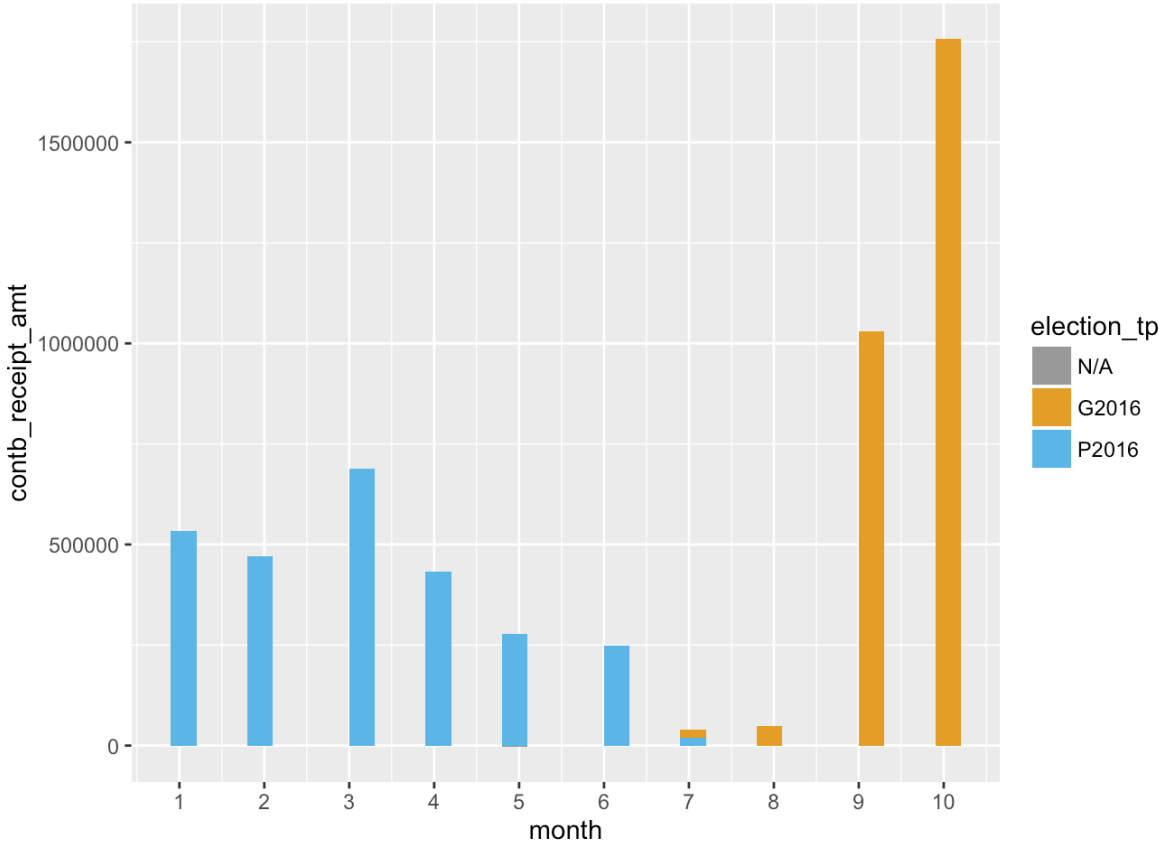
By adding a facet of candidate's political party, we can better compare contribution amount of candidates to others of their own political party. Within Democratic party, Lessig had the highest median contribution amount. Within Republican party, Christie had the highest median contribution amount.

Based on the previous bar plot of total contribution amount received for each candidate, I created a multivariate plot by filling the data with election type information.



We learnt a great amount of information from this plot. In particular, Trump and Clinton are the only candidates with significant contribution amount to their general election campaign. This makes sense considering that they are the two major candidates remained in the general election period. It is also interesting to see that Trump received a lot more contributions during general election than primary election.

To further explore the trends of contributions over time, we can add a filling of election type to the plot we made earlier, where it shows total amount of contributions across months in 2016.



Besides showing us the trend over time, this plot also demonstrates that most of the contributions were made for primary election from January to June, and most of the contributions were made for general election from August to October. Contributions in July were a mixture of primary election and general election donations, as the primary election ended in the middle of July. This confirms our intuition since primary election ended in mid-July and general election campaign started right afterwards.

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?

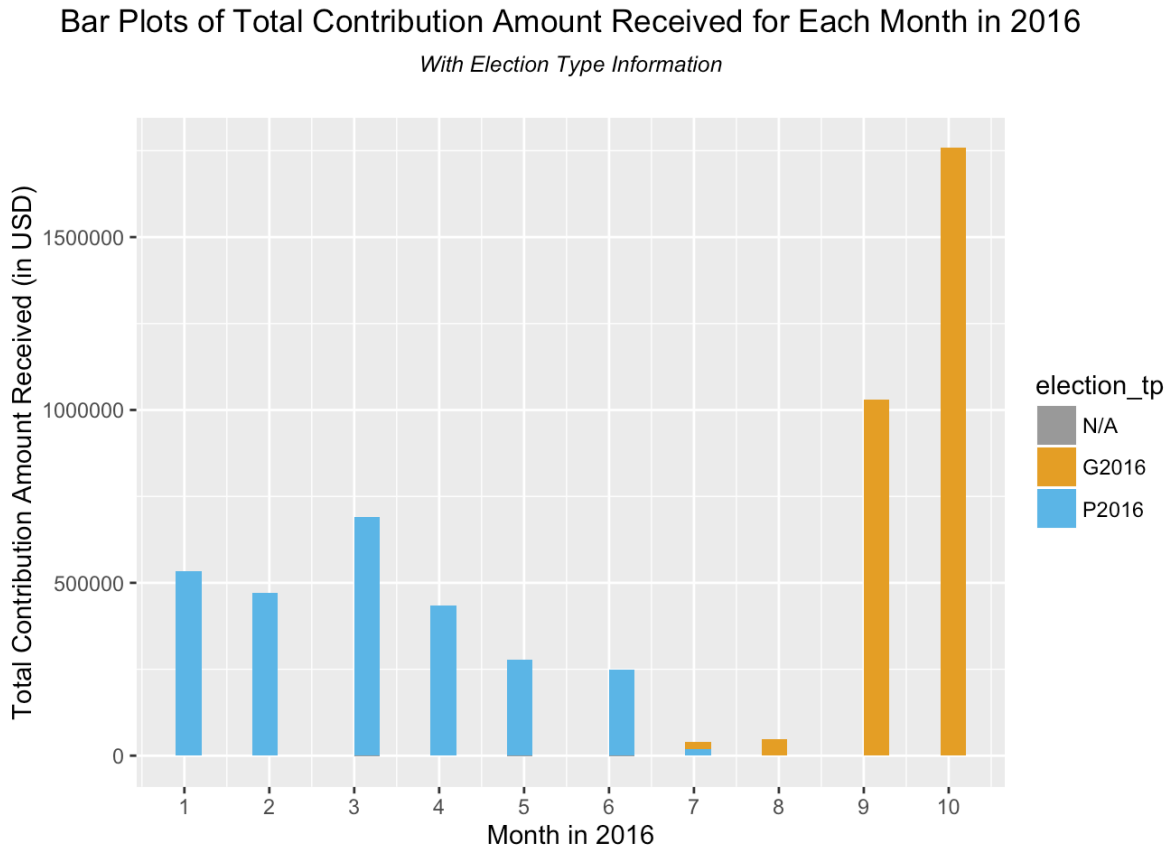
Seperating the bar plots of total contributions received by candidates into different political party and looking at the boxplots of contribution amounts of candidates from different political parties definitely showed us a clear trend that Democratic candidates take smaller donations in a larger quantities, whereas Republican candidates takes bigger donations but in a smaller quantities.

Were there any interesting or surprising interactions between features?

I thought that the interactions between total contribution amount and election type is quite interesting. It confirms our intuition that only the major candidates who become the party nominees would receive donations for general election. We also saw that contributions shifted from primary election to general election in July as the race transitioned from primary election to general election at that time.

Final Plots and Summary

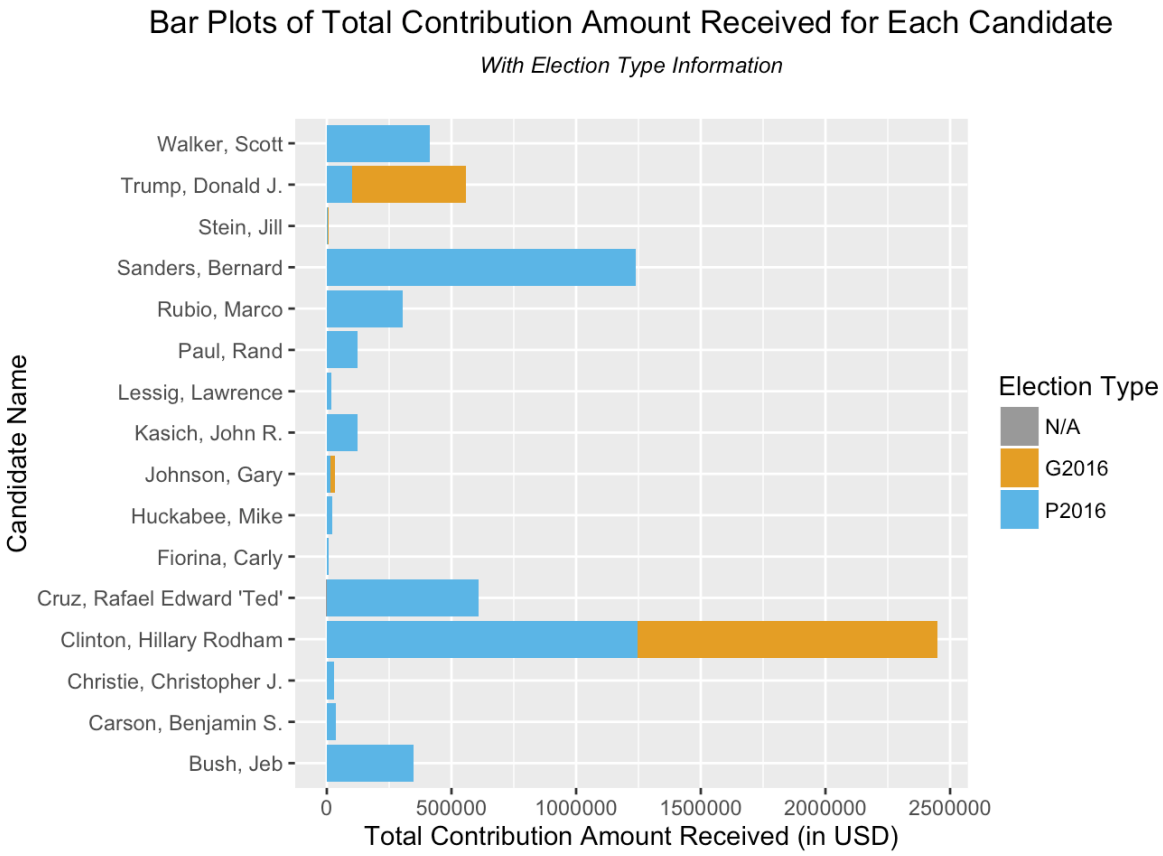
Plot One



Description One

This plot demonstrates the total amount of contributions across different months in 2016, with election type information. The total contribution amount started small in January and peaked in March, then it decreased and started increasing again in September, which led to the maximum in October. This makes sense because many candidates dropped out of the race in March, and general election was taken place in November so many contributors made donations in October hoping to make an impact at a critical time. In addition, we see that most of the contributions were made for primary election from January to June, and most of the contributions were made for general election from August to October. Contributions in July were a mixture of primary election and general election donation, as the primary election ended in the middle of July. This confirms our intuition since primary election ended in mid-July and general election campaign started right afterwards.

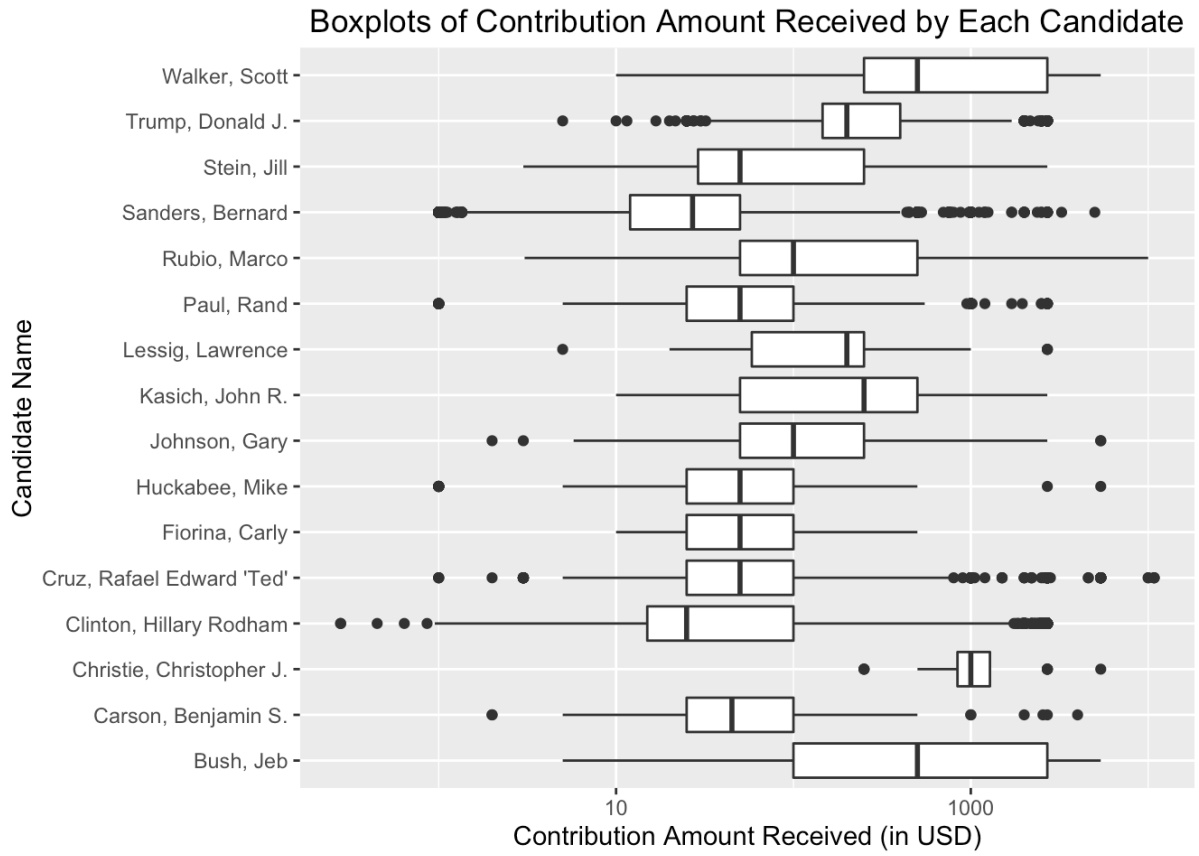
Plot Two



Description Two

This plot demonstrates the total amount of contributions received for each candidate, filled with election type information (general or primary). It shows that Clinton received the highest total amount of contributions and Clinton and Trump were the only candidates with major general election contributions which is consistent with our intuition.

Plot Three



Description Three

The boxplots show 25%, 50% and 75% percentile of the contribution amount for each candidate, along with outliers. We can see that Christie had the highest median amount of contributions amongest all candidates.

Reflection

By doing univariate, bivariate and multivariate analysis, we gained a lot of insights about the political contributions made to 2016 presidential candidates in Illinois, including which candidate received the most contributions and total amount of contributions, summary statistics of contributions for each candidate, differences in contributions across election types, political parties and time.

One major finding is that Democratic candidates received smaller donations but in a larger quantities, whereas the opposite is true for Republican candidates. Another important finding is that contribution amount varies over time as more contributions were made when more candidates are involved or during critical times in the race.

I discovered that bar plots work very well for this particular analysis since a lot of the variables are categorical.

However, one limitation of this analysis is that since most of the variables are categorical, finding correlations is difficult. It was also challenging to transform visualizations in order to present in a manner that convey the message clearly and concisely.

A way to improve this analysis in the future is to incorporate more demographic information of the contributor, if possible. For example, we can conjecture the gender of the contributor based on the first name, then analyze if the contribution behavior between males and females are different, such as the contribution amount and the candidates they tend to contribute to. If I can use more time, I would also add a choropleth map with the zipcode information to enhance the visualization.