Florida 2016 US Presidential Campaign Contribution Analysis

by Xilin Miao

Introduction

This is an exploration of 2016 US presidential campaigns financial contributions of Florida. Presidential campaign finance data contains much informatino wich is very valuable to reveal the demographic and geographic support for each party and candidate, and it also reflect the current status of the candidates in the race.

The reason why I chose Florida's data is that Florida has been the most pupulated of the "swing states", means where both political parties have a similar level of support. For this exploratory data analysis, the data of 2016 presidential campaigns financial contributions is from Federal Election Commission.

In this analysis, I will try to answer three questions: Firstly, which party and candidates received the most financial support in Florida (from the donations received)? Secondly, is there a difference in donations between genders, and what is the patterns here? Lastly, who made these contributions? which occupation make the most?

Dataset

Download dataset and load data.

```r
# Load the Data
FL <- read.csv("C:/Files/Udacity/Data-Analysis/Projects/Florida 2016 US Presidential Campaign Contributi
colnames(FL) <- c('cmte_id', 'cand_id', 'cand_nm', 'contbr_nm', 'contbr_city',
                  'contbr_st', 'contbr_zip',    'contbr_employer',
                  'contbr_occupation',  'contb_receipt_amt', 'contb_receipt_dt',
                  'receipt_desc', 'memo_cd', 'memo_text', 'form_tp', 'file_num',
                  'tran_id', 'election_tp')
```

Data Preprocessing

Create New Variables

Before analyzing this data set, we should firstly do data processing since some features which I am interested in are not included in such as gender, and donor's geographical location in latigude and longtitude.

```r
# Create party variable

democrat <- c("Clinton, Hillary Rodham",
               "Lessig, Lawrence",
               "O'Malley, Martin Joseph",
               "Sanders, Bernard",
               "Webb, James Henry Jr.")

others <- c("Johnson, Gary", "McMullin, Evan", "Stein, Jill")

FL$party <- ifelse(FL$cand_nm %in% democrat, "Democrat", "Republican")
FL$party[FL$cand_nm %in% others] <- 'Others'
```

```r
# Extract first name from contbr_nm for gender predication
FL$contbr_first_nm <- sub(" .*", "", sub(".*, ", "", FL$contbr_nm))
```

```r
#Create gender dataframe
gender_df <- gender(FL$contbr_first_nm, method = "ssa", c(1932, 2012), countries = "United States")
# "ssa" here means looks up names based from the U.S Social Security Administration baby name data

#Create gender variable
gender_df <- unique(gender_df)
names(gender_df)[1] = "contbr_first_nm"
FL <- inner_join(FL, gender_df, by = 'contbr_first_nm')
```

```r
# Convert zipcode to latitude and longitude by zipcode package
data(zipcode)
FL$contbr_zip <- substr(FL$contbr_zip, 1, 5)
zipcode_FL <- subset(zipcode, state == "FL")[, -c(2,3)]
colnames(zipcode_FL) <- c("contbr_zip", "latitude", "longitude")
FL <- merge(FL, zipcode_FL)
```

```r
#Extract conbution recipt year and month
FL <- FL %>%
  mutate(date = as.Date(contb_receipt_dt, "%d-%b-%y"),
         year = year(date),
         month = month(date),
         year_month = paste(month.abb[month], ",", year))
```

Data Clean

After looking into this data set, we noticed that there are 7608 negative contributions (contb_receipt_amt). These negative values should be the refund, so we should omit these observations. And since there is contribution limits for 2015-2016, there is a limit which is $2,700 per election, per candidate. So the contribution above $2,700 will also be omitted since these contribution will be refunded lastly.

```r
nrow(FL[FL$contb_receipt_amt <= 0,])
```

```
## [1] 7608
```

```r
FL = filter(FL, FL$contb_receipt_amt > 0 & FL$contb_receipt_amt <= 2700)
```

Dataset Overview

```r
str(FL)
```

```
## 'data.frame':    404968 obs. of  31 variables:
##  $ contbr_zip       : chr  "32003" "32003" "32003" "32003" ...
##  $ cmte_id          : chr  "C00575795" "C00575795" "C00575795" "C00580100" ...
##  $ cand_id          : chr  "P00003392" "P00003392" "P00003392" "P80001571" ...
##  $ cand_nm          : chr  "Clinton, Hillary Rodham" "Clinton, Hillary Rodham" "Clinton, Hillary Rod
##  $ contbr_nm        : chr  "SCHWARTZ, JANICE" "HARVEY, ANNEMARIE" "HARVEY, ANNEMARIE" "OROMANER, JEF
##  $ contbr_city      : chr  "FLEMING ISLAND" "FLEMING ISLAND" "FLEMING ISLAND" "FLEMING ISLAND" ...
##  $ contbr_st        : chr  "FL" "FL" "FL" "FL" ...
##  $ contbr_employer  : chr  "RETIRED" "N/A" "N/A" "RETIRED" ...
##  $ contbr_occupation: chr  "HEALTH INSURANCE" "RETIRED" "RETIRED" "RETIRED" ...
##  $ contb_receipt_amt: num  19 50 38 80 37 3.89 50 50 100 75 ...
```

```
##  $ contb_receipt_dt : chr  "17-Aug-16" "27-May-16" "26-Aug-16" "5-Jul-16" ...
##  $ receipt_desc     : chr  "" "" "" "" ...
##  $ memo_cd          : chr  "" "" "" "X" ...
##  $ memo_text        : chr  "" "" "" "" ...
##  $ form_tp          : chr  "SA17A" "SA17A" "SA17A" "SA18" ...
##  $ file_num         : int  1126762 1091720 1126762 1104813 1133832 1094141 1135630 1057553 1133832 1
##  $ tran_id          : chr  "C9249367" "C5262304" "C9583539" "SA18.1467776" ...
##  $ election_tp      : chr  "G2016" "P2016" "G2016" "P2016" ...
##  $ party            : chr  "Democrat" "Democrat" "Democrat" "Republican" ...
##  $ contbr_first_nm  : chr  "JANICE" "ANNEMARIE" "ANNEMARIE" "JEFFREY" ...
##  $ proportion_male  : num  0.0027 0 0 0.9962 0.0025 ...
##  $ proportion_female: num  0.9973 1 1 0.0038 0.9975 ...
##  $ gender           : chr  "female" "female" "female" "male" ...
##  $ year_min         : num  1932 1932 1932 1932 1932 ...
##  $ year_max         : num  2012 2012 2012 2012 2012 ...
##  $ latitude         : num  30.2 30.2 30.2 30.2 30.2 ...
##  $ longitude        : num  -81.7 -81.7 -81.7 -81.7 -81.7 -81.7 -81.7 -81.7 -81.7 -81.7 ...
##  $ date             : Date, format: "2016-08-17" "2016-05-27" ...
##  $ year             : num  2016 2016 2016 2016 2016 ...
##  $ month            : num  8 5 8 7 10 5 4 10 10 9 ...
##  $ year_month       : chr  "Aug , 2016" "May , 2016" "Aug , 2016" "Jul , 2016" ...
```

After processing the data, we have removed observations with either negative amount or amount exceed $2,700. Except the original 18 variables, following main variables were added also:

- party: candidate' party affiliation
- latitude: contributors' geographic latitude
- longitude: contributors' geographic lonitude
- contbr_first_nm: contributors' first name
- gender: contributors' gender
- date: contribution receipt data
- year: contribution receipt year
- month: contribution receipt month
- year_month: contribution receipt date (year month format)

Univariate Plots Section

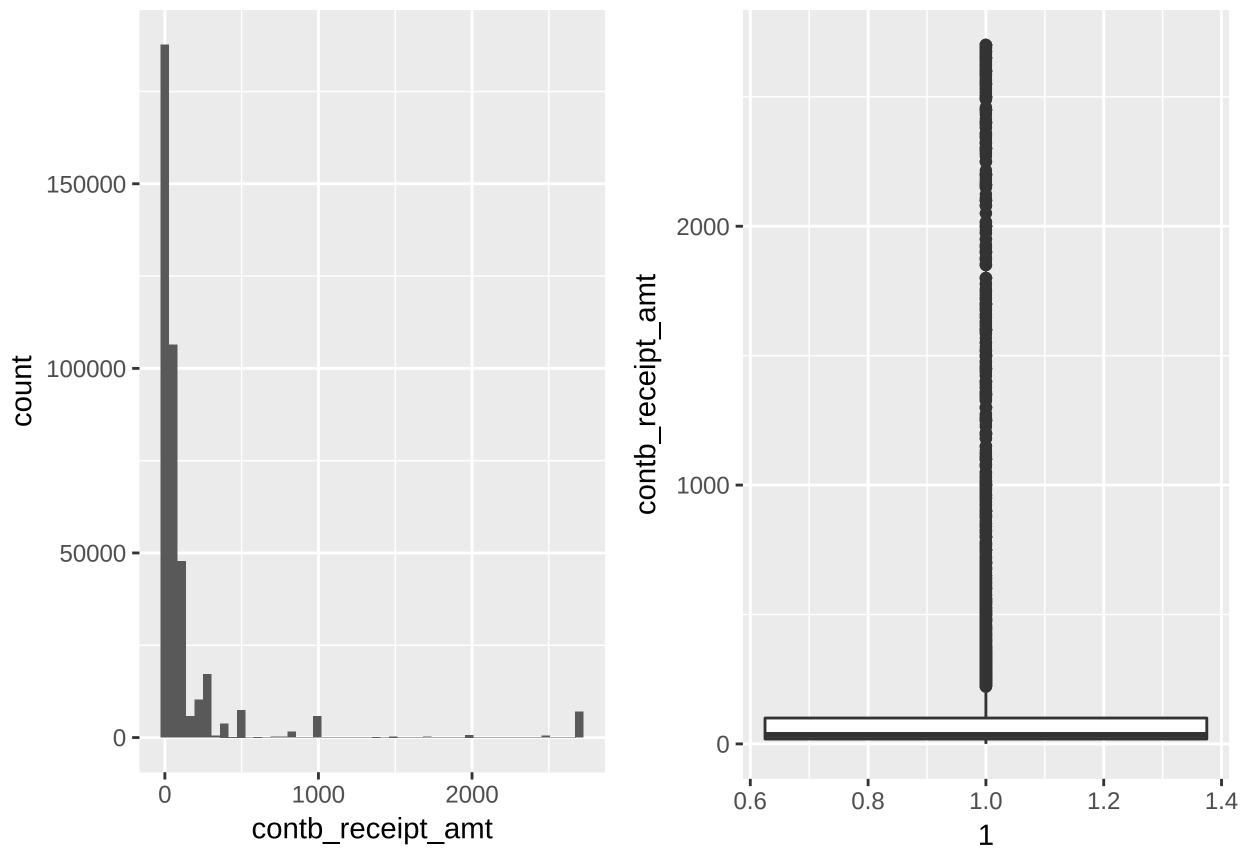
```r
summary(FL$contb_receipt_amt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.15   19.00   35.00  143.13  100.00 2700.00
```

To start, let's have a glance of how the contribution distributed.

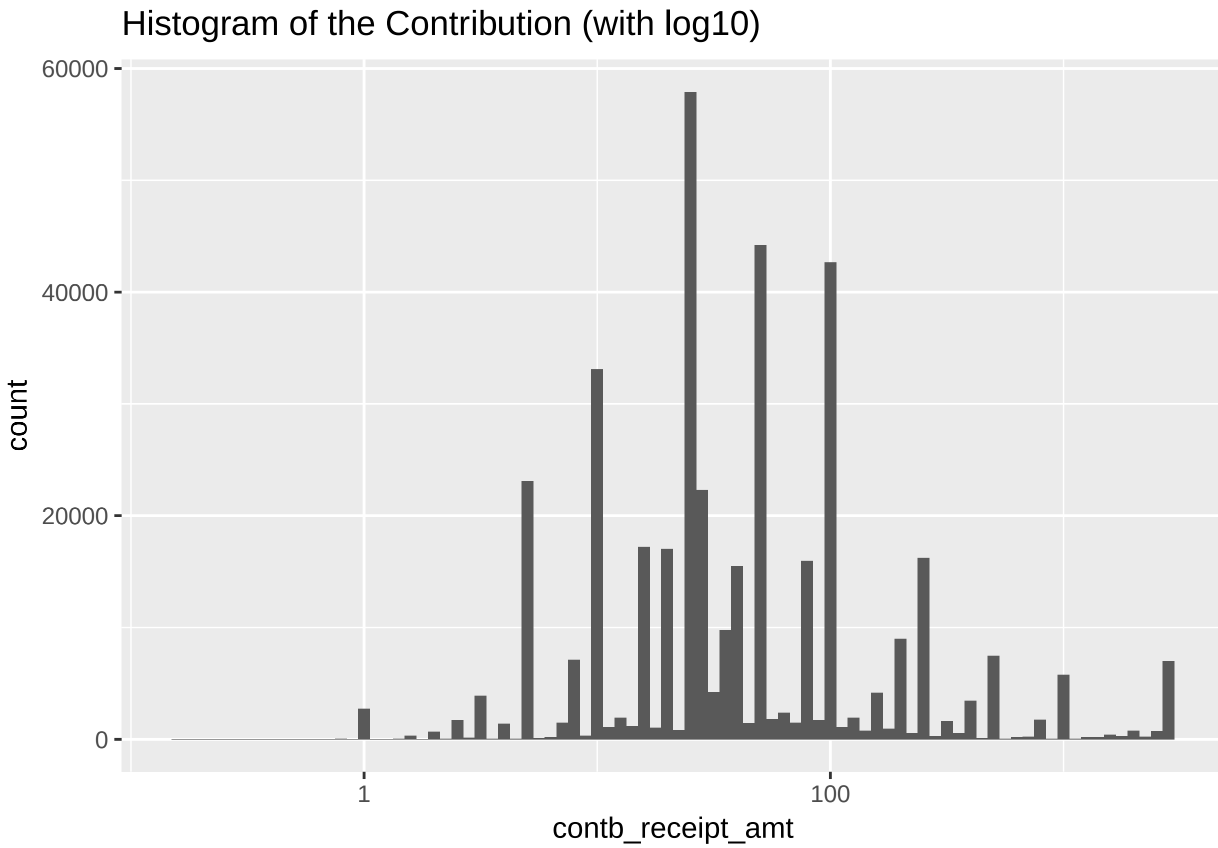
```r
p1 <- ggplot(aes(x = contb_receipt_amt), data = FL) +
  geom_histogram(bins = 50)
p2 <- ggplot(aes(x = 1, y = contb_receipt_amt), data = FL) +
  geom_boxplot()

grid.arrange(p1, p2, ncol = 2)
```



From upper plots we can see that there are so many outliers. To better understand the distribution of the contribution, now transform the plots to log10.

```
ggplot(aes(x = contb_receipt_amt), data = FL) +
  geom_histogram(binwidth = 0.05) +
  scale_x_log10() +
  ggtitle('Histogram of the Contribution (with log10)')
```




```r
tail(sort(table(FL$contb_receipt_amt)), 5)
```

```
##
##     5    10   100    50    25
## 23035 33014 41622 42699 56386
```

```r
summary(FL$contb_receipt_amt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.15   19.00   35.00  143.13  100.00 2700.00
```

The distribution looks normal now and we can see that most donors made small amount of contributions which is among about 50 to 125. Now let's see the contribution distribution across the parties, candidates, genders and occupations.

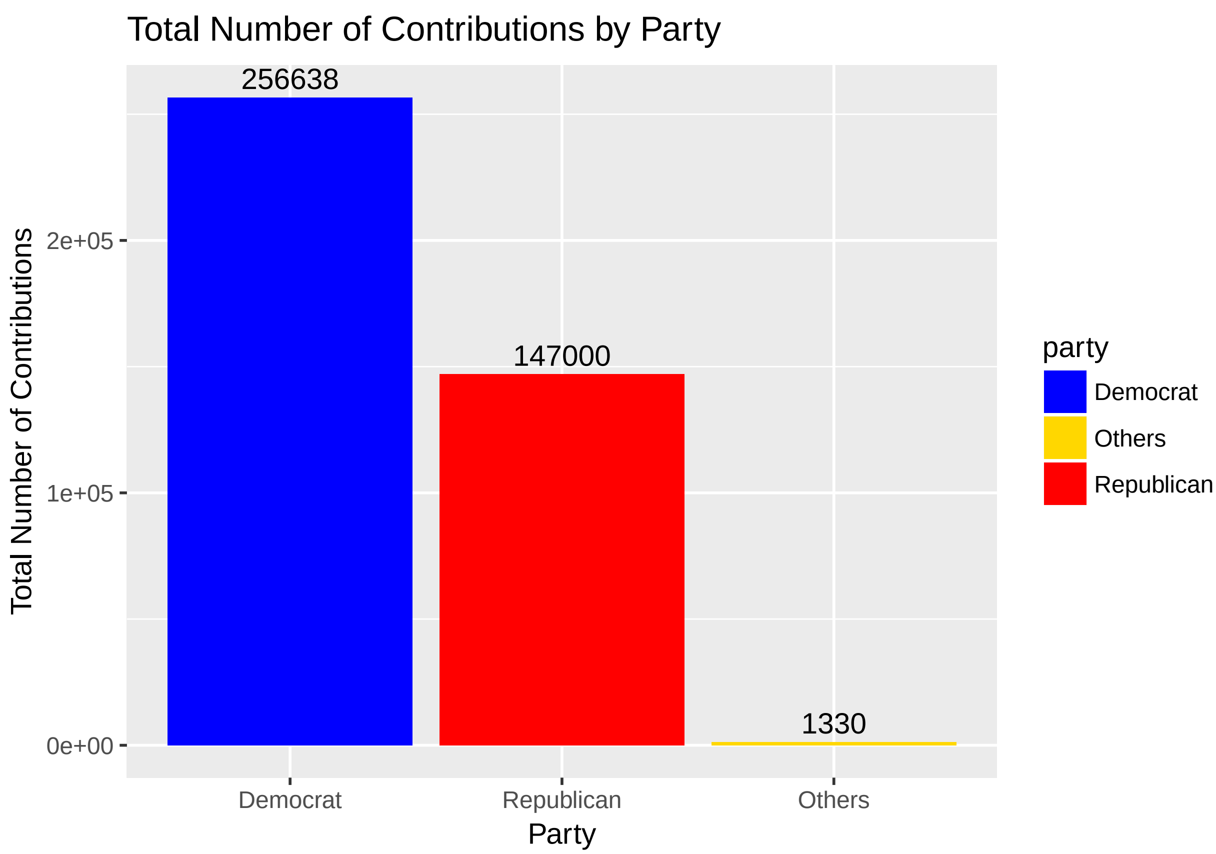
```r
#Create party dataframe with Descending order
party_group <- group_by(FL, party)
FL.contr_by_party <- summarize(party_group,
                               sum_party = sum(contb_receipt_amt),
                               number_of_candidate = length(unique(cand_id)),
                               mean_party = sum_party/number_of_candidate,
                               n = n())

FL.contr_by_party[rev(order(FL.contr_by_party$n)),]
```

```
## # A tibble: 3 x 5
##        party  sum_party number_of_candidate mean_party      n
##        <chr>      <dbl>               <int>      <dbl>  <int>
```

```
## 1   Democrat 24990943.1                     5   4998188.6 256638
## 2 Republican 32633124.7                    17   1919595.6 147000
## 3     Others   337861.5                     3    112620.5   1330
```

```r
#plot of party
ggplot(aes(x = reorder(party, -n), y = n, fill = party), data = FL.contr_by_party) +
  geom_bar(stat = 'identity') +
  geom_text(stat = 'identity', aes(label = n),
            data = FL.contr_by_party, vjust = -0.4) +
  xlab('Party') +
  ylab('Total Number of Contributions') +
  ggtitle('Total Number of Contributions by Party') +
  scale_fill_manual(values = c('blue', 'gold', 'red'))
```




```r
sum(FL.contr_by_party$n)
```

```
## [1] 404968
```

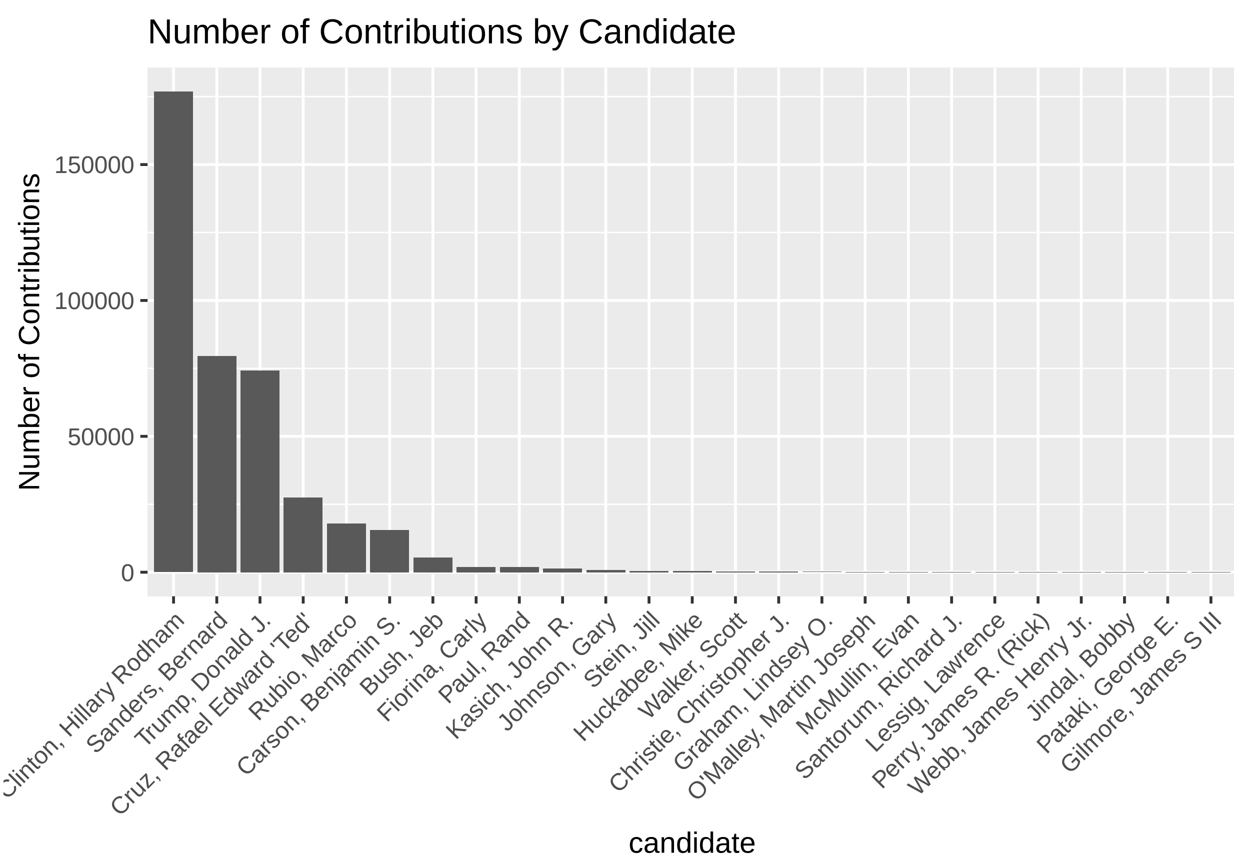
From this we can see that the total number of 2016 presidential campaign finance contributions is near 405K, while the Democratic party took around 257K and almost 1.75 times of the number of donations make to the Republican party, which is 147K, and other parties took very few numbers of donations compared with Democrat party and Republican party, with about 1.3K.

```r
#Create Candidate dataframe with Descending order
cand_group <- group_by(FL, cand_nm)
FL.contr_by_cand <- summarize(cand_group,
                              sum_cand = sum(contb_receipt_amt),
                              n = n())
```

```r
FL.contr_by_cand[rev(order(FL.contr_by_cand$n)),]
```

```
## # A tibble: 25 x 3
##                       cand_nm   sum_cand      n
##                         <chr>      <dbl>  <int>
##  1   Clinton, Hillary Rodham 21451718.3 176790
##  2          Sanders, Bernard  3355944.0  79631
##  3          Trump, Donald J. 12385981.2  74156
##  4 Cruz, Rafael Edward 'Ted'  2775052.6  27436
##  5              Rubio, Marco  6634027.2  17993
##  6       Carson, Benjamin S.  1843217.8  15496
##  7                 Bush, Jeb  6325905.0   5401
##  8            Fiorina, Carly   412640.1   2003
##  9                Paul, Rand   411286.7   1950
## 10           Kasich, John R.   784383.7   1303
## # ... with 15 more rows
```

```r
ggplot(aes(x = reorder(cand_nm, -n), y = n), data = FL.contr_by_cand) +
  geom_bar(stat = 'identity') +
  xlab('candidate') +
  ylab('Number of Contributions') +
  ggtitle('Number of Contributions by Candidate') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

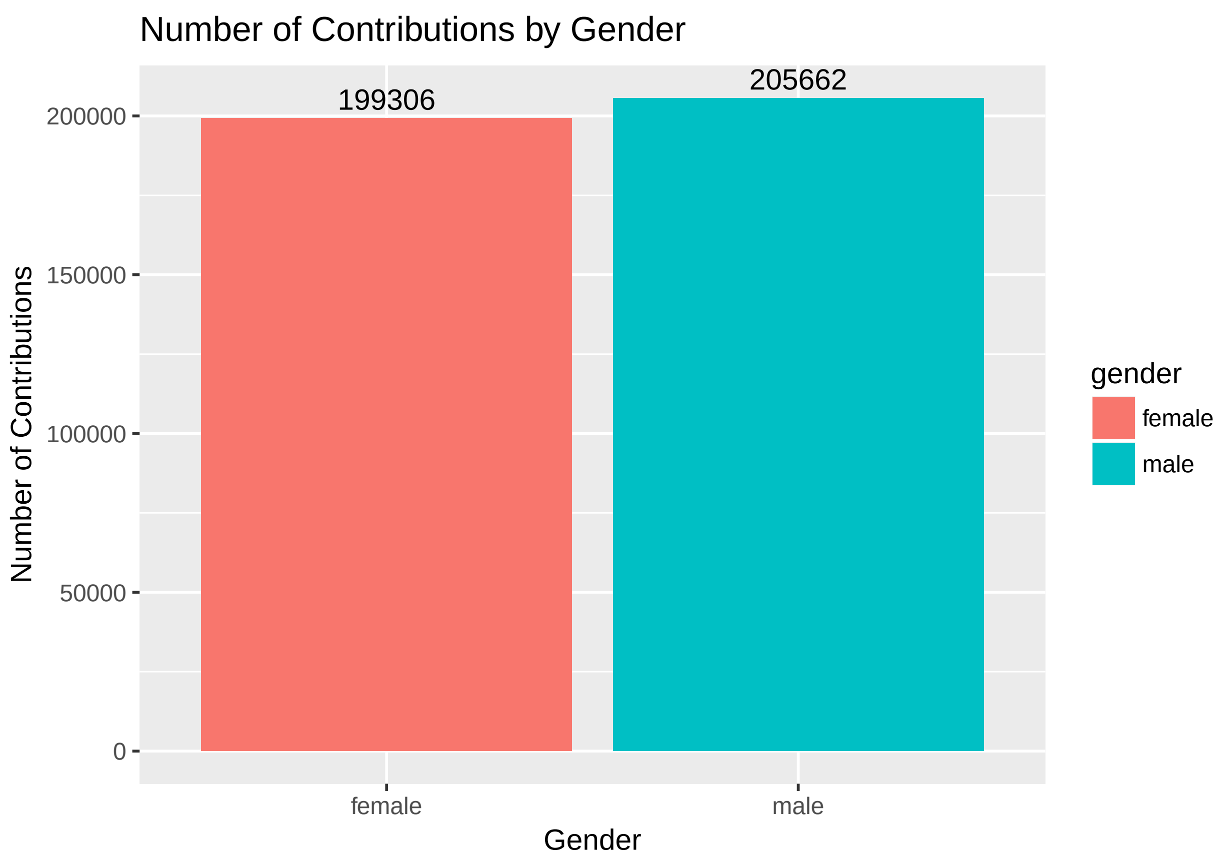


There were intotal 25 candidates, and Clinton, Hillary Rodham took the most number of contributions with about 176.8K, and followed by Sanders, Bernard with about 79.6K and Trump, Donald J. with about 74.2K.

```r
#Create gender dataframe by descending order
gender_group <- group_by(FL, gender)
FL.contr_by_gender <- summarize(gender_group,
                                sum_gen = sum(contb_receipt_amt),
                                mean_gen = mean(contb_receipt_amt),
                                n = n())
FL.contr_by_gender[rev(order(FL.contr_by_gender$n)),]
```

```
## # A tibble: 2 x 4
##   gender  sum_gen mean_gen      n
##    <chr>    <dbl>    <dbl>  <int>
## 1   male 35299939 171.6405 205662
## 2 female 22661991 113.7045 199306
```

```r
#Plot of gender
ggplot(aes(x = gender, y = n, fill = gender),
       data = FL.contr_by_gender, vjust = -0.4) +
  geom_bar(stat = 'identity') +
  geom_text(aes(label = n), stat = 'identity', data = FL.contr_by_gender, vjust = -0.4) +
  xlab('Gender') +
  ylab('Number of Contributions') +
  ggtitle('Number of Contributions by Gender')
```



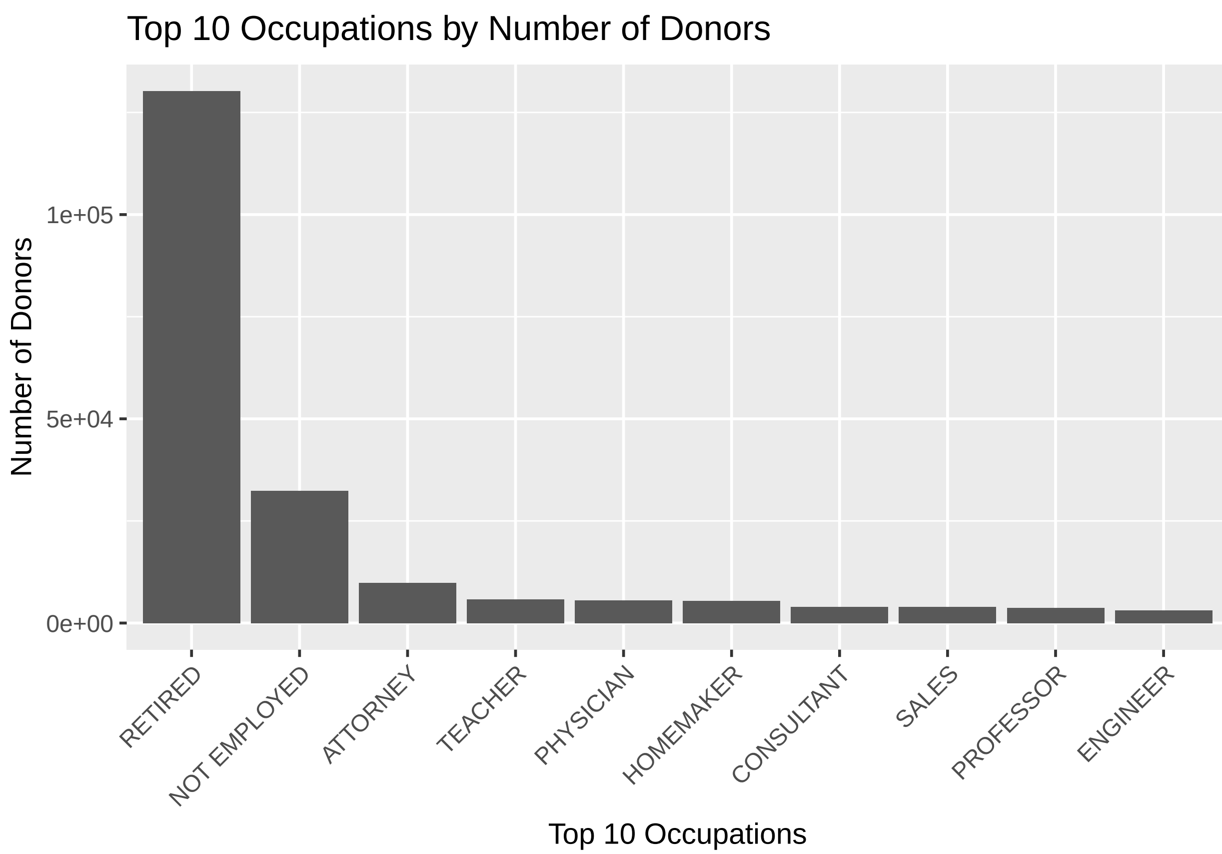
We can see here women and men made comparable donations, with about 199K from women and almost 206K from men. But from the mean we can see there is obvious difference. Let's see what are these donors occupation.

```r
#Create occupation dataframe with Descending order
occupation_group <- group_by(FL, contbr_occupation)
FL.contr_by_occupation <- summarize(occupation_group,
                                    sum_occupation = sum(contb_receipt_amt),
                                    mean_occupation = mean(contb_receipt_amt),
                                    n = n())

#"INFORMATION REQUESTED" means the donor haven't provide this information, ignored
FL.contr_by_occupation <- subset(FL.contr_by_occupation, contbr_occupation != "INFORMATION REQUESTED" &
FL.contr_by_occupation <- head(arrange(FL.contr_by_occupation,desc(n)), n = 10)
FL.contr_by_occupation[rev(order(FL.contr_by_occupation$n)),]
```

```
## # A tibble: 10 x 4
##    contbr_occupation sum_occupation mean_occupation      n
##                <chr>          <dbl>           <dbl>  <int>
##  1           RETIRED     15571366.7       119.54892 130251
##  2      NOT EMPLOYED      1466685.2        45.30722  32372
##  3          ATTORNEY      3375634.0       344.31191   9804
##  4           TEACHER       325942.4        55.74524   5847
##  5         PHYSICIAN      1206031.2       216.71721   5565
##  6         HOMEMAKER      2042790.3       374.34309   5457
##  7        CONSULTANT       868172.7       218.40822   3975
##  8             SALES       388058.8        97.99464   3960
##  9         PROFESSOR       291716.8        78.92769   3696
## 10          ENGINEER       308134.7        99.91397   3084
```

```r
#plot of occupation
ggplot(aes(x = reorder(contbr_occupation, -n), y = n), data = FL.contr_by_occupation) +
  geom_bar(stat = 'identity') +
  xlab('Top 10 Occupations') +
  ylab('Number of Donors') +
  ggtitle('Top 10 Occupations by Number of Donors') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



After looking into occupations, we can find that retired people take the first place, followed by not employed people, attorney comes to the third.

```r
summary(FL$contb_receipt_dt)
```

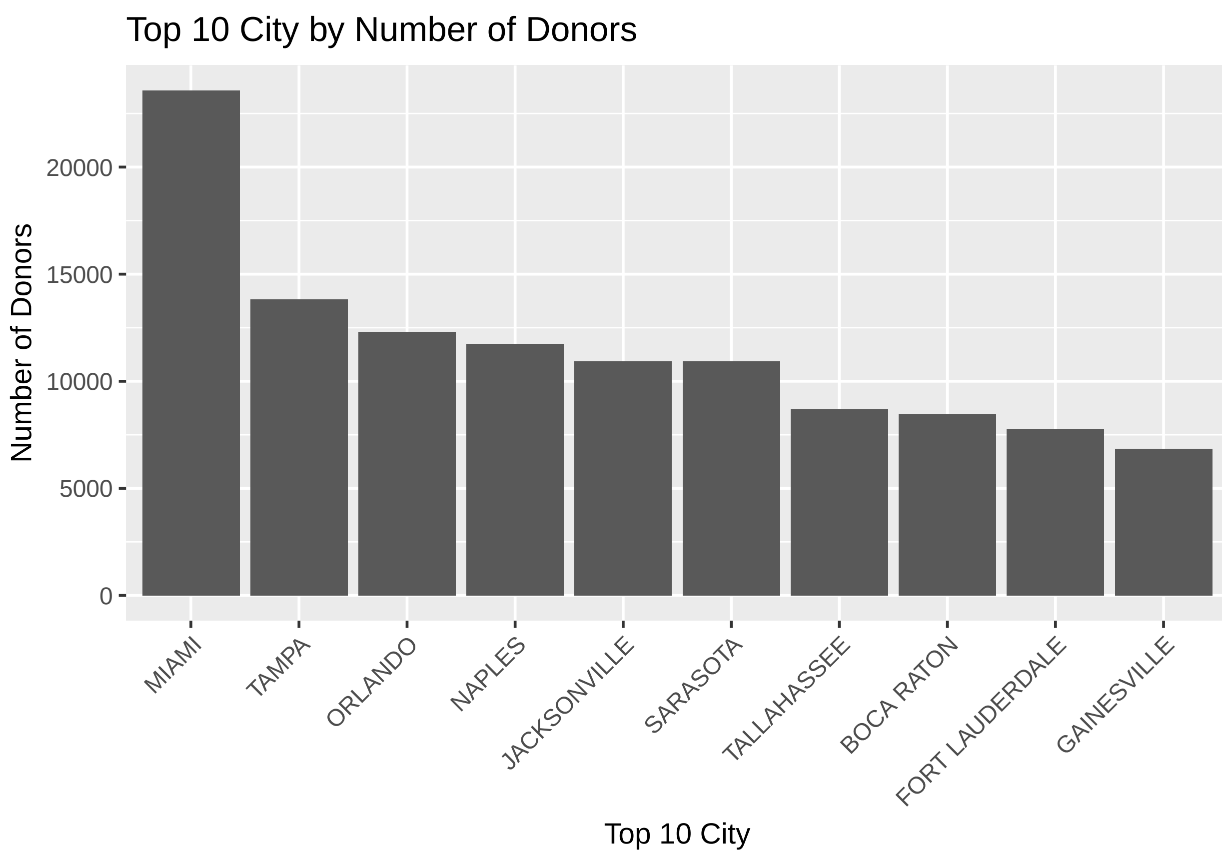
```
##    Length     Class      Mode 
##    404968 character character
```

```r
#Create city dataframe by descending order
city_group <- group_by(FL, contbr_city)
FL.contr_by_city <- summarize(city_group,
                              sum_city = sum(contb_receipt_amt),
                              mean_city = mean(contb_receipt_amt),
                              n = n())
FL.contr_by_city <- head(arrange(FL.contr_by_city,desc(n)), n = 10)
FL.contr_by_city[rev(order(FL.contr_by_city$n)),]
```

```
## # A tibble: 10 x 4
##        contbr_city  sum_city mean_city     n
##              <chr>     <dbl>     <dbl> <int>
##  1           MIAMI 4408990.3 187.05941 23570
##  2           TAMPA 1931246.5 139.74288 13820
##  3         ORLANDO 1584454.4 128.74416 12307
##  4          NAPLES 2408467.2 204.87132 11756
##  5    JACKSONVILLE 1595578.3 146.06173 10924
##  6        SARASOTA 1500683.1 137.38745 10923
##  7     TALLAHASSEE 1313958.0 151.34277  8682
##  8      BOCA RATON 1861869.9 220.15725  8457
```

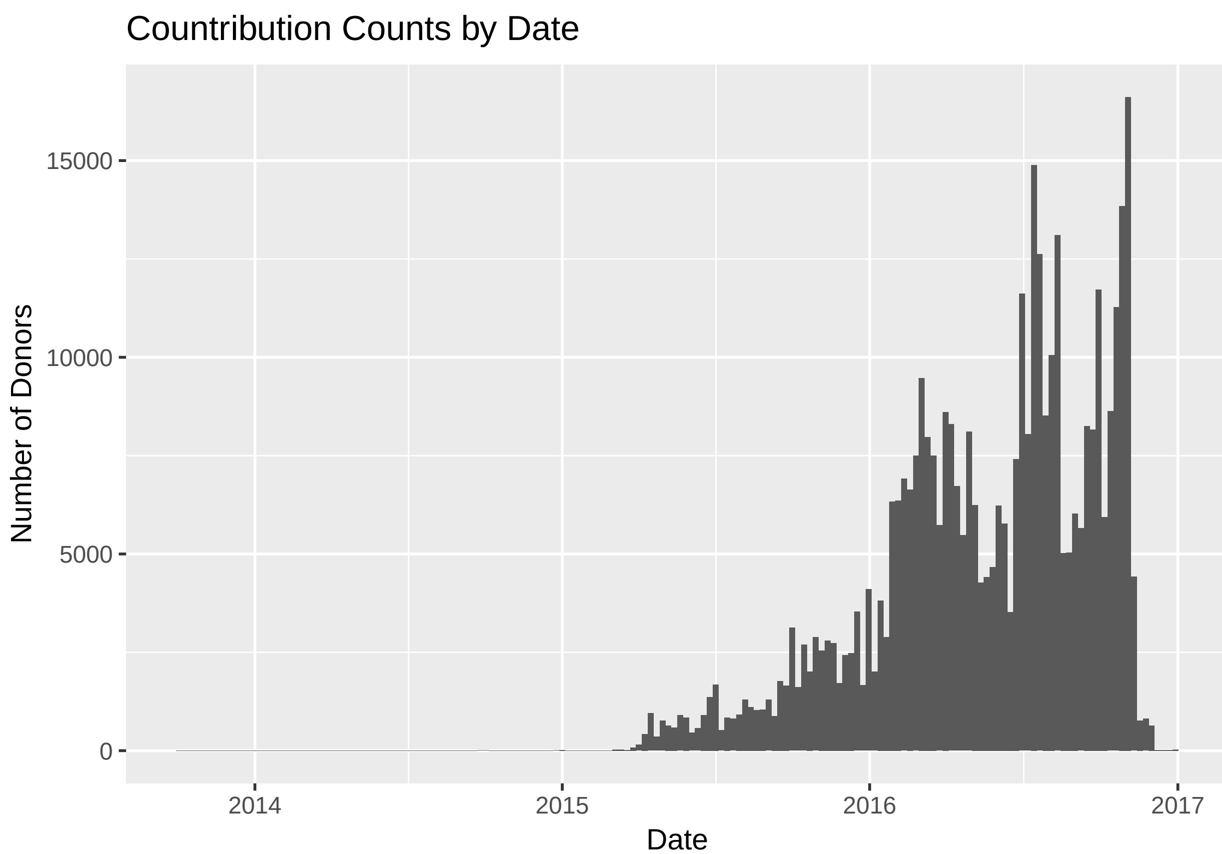
```
## 9 FORT LAUDERDALE 1602231.2 206.47310 7760
## 10     GAINESVILLE  602479.6  87.90189 6854
```

```r
#plot of city
ggplot(aes(x = reorder(contbr_city, -n), y = n), data = FL.contr_by_city) +
  geom_bar(stat = 'identity') +
  xlab('Top 10 City') +
  ylab('Number of Donors') +
  ggtitle('Top 10 City by Number of Donors') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The city with most number of contribution is Miami, with about 23.6K.

```r
#plot of date
ggplot(aes(x = date), data = FL) +
  geom_histogram(position = position_dodge(), binwidth = 7) +
  xlab('Date') +
  ylab('Number of Donors') +
  ggtitle('Countribution Counts by Date')
```



It's interesting to see when people made contributions. this distribution shows that the bimodal with peaking around June 2016 and again when closing to the election.

Univariate Analysis

What is the structure of your dataset?

There are 404968 contributions and 18 variables. The variables that I am more interested in are:

- cand_nm: Candidate Name
- contbr_nm: Contributor name(first name, for gender prediction)
- contvr_zip:Contributor Zipcode
- contbr_occupation: COntributor Occupation
- contb_receipt_amt: COntribution Amount
- contb_receipt_dt: Contribution date

Other observations:

- The median contribution amount is $35.
- Most people contribute small amount of money, around 50 to 125.
- The numbers of contribution of women and men are comparable, while in average men's contribution amout is almost 1.5 times of women's.
- The Democrat received the most number of donations, and it is almost 1.75 times of Republican.
- Hillary Clinton have the most supporters.
- The city with most number of contribution is Miami.
- Retired people make the most number of contributions.

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

The main features of interest in this dataset are party, candidate, and contribution amount. I would like to find the answers to my questions at the beginning and also try to use combination of variables 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?

Features like gender, contributor's occupation, contribution date, contributor's zipcode can help support further insights of the dataset. We can explore more by finding the relationships between these features and the main features of interest above.

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

Yes, 5 new variables are created:

- party: candidates party.
- contbr_first_nm: contributor's first name to predict contributos' gender.
- gender: contributor's gender.
- latitude: contributors' geographic latitude.
- longitude: contributors' geographic lonitude.

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?

In this dataset, the observations with contribution amount value below 0 and 2700 were omitted since negative values are refund and amount value higher than 2700 will be refunded.

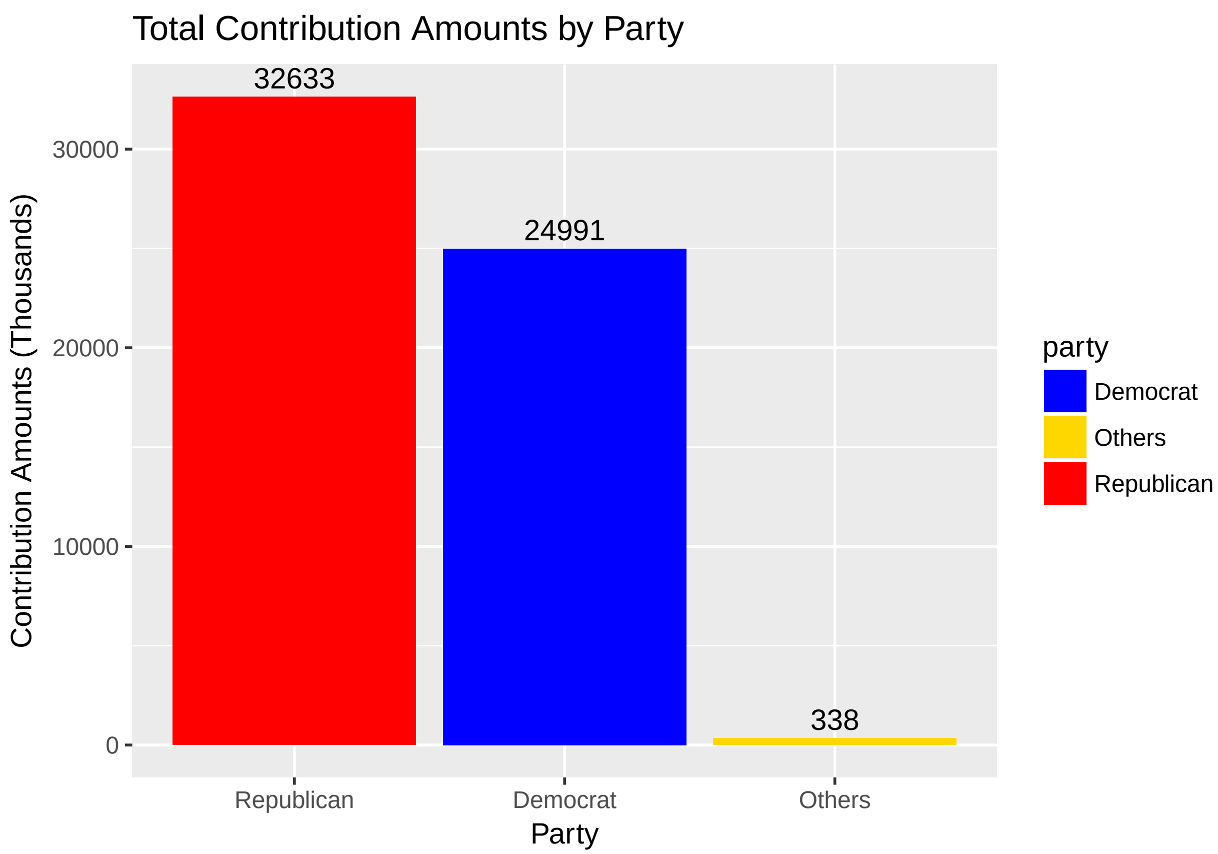
Bivariate Plots Section

Let's see the barplots for contribution amount by party and candidates.

```
FL.contr_by_party[rev(order(FL.contr_by_party$sum_party)),]
```

```
## # A tibble: 3 x 5
##        party  sum_party number_of_candidate mean_party      n
##        <chr>      <dbl>               <int>      <dbl>  <int>
## 1 Republican 32633124.7                  17  1919595.6 147000
## 2   Democrat 24990943.1                   5  4998188.6 256638
## 3     Others   337861.5                   3   112620.5   1330
```

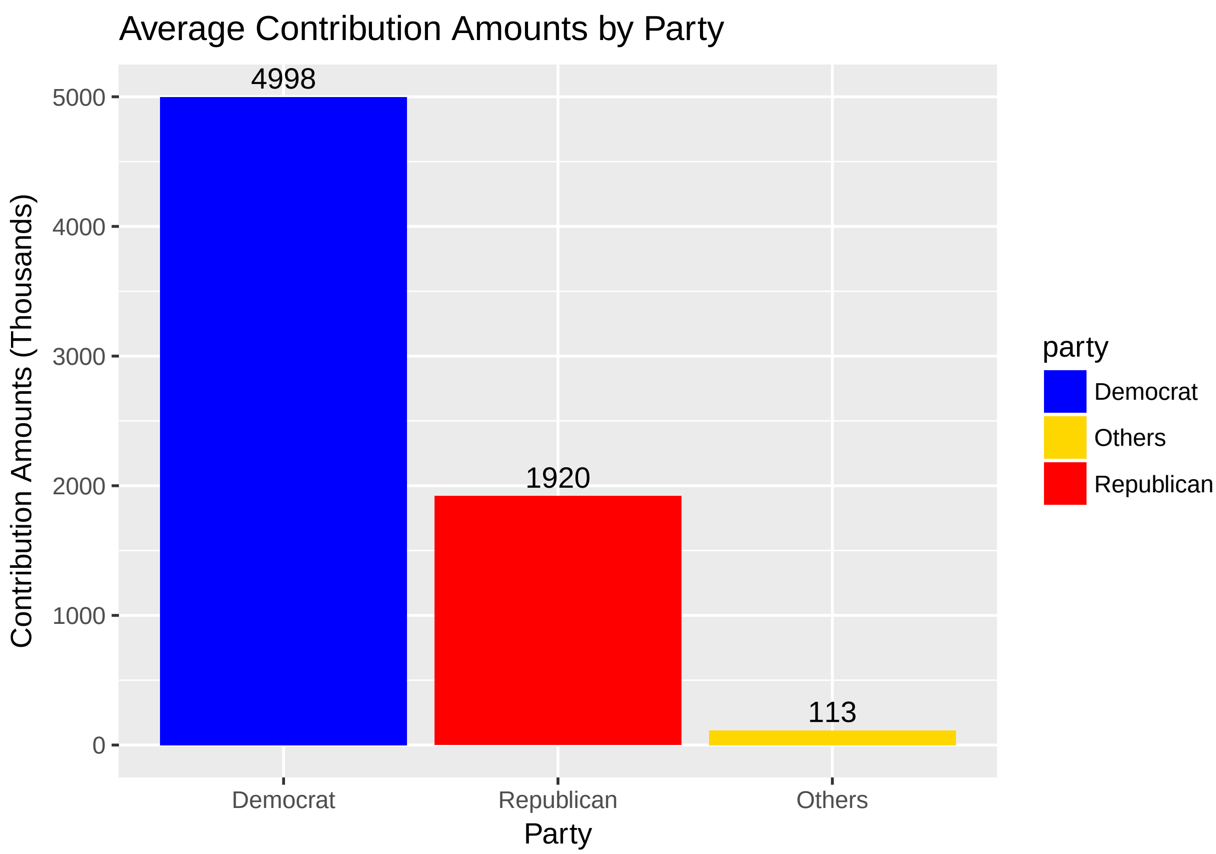
```
#plot of contribution amount by party
ggplot(aes(x = reorder(party, -(sum_party/1000)), y = sum_party/1000, fill = party), data = FL.contr_by
  geom_bar(stat = 'identity') +
  geom_text(stat = 'identity', aes(label = round(sum_party/1000)),
            data = FL.contr_by_party, vjust = -0.4) +
  xlab('Party') +
  ylab('Contribution Amounts (Thousands)') +
  ggtitle('Total Contribution Amounts by Party') +
  scale_fill_manual(values = c('blue', 'gold', 'red'))
```



It's interesting to see shere the contribution money went. Previouly in the contribution number plot, we see Democrat took number almost 1.75 time of Republican. But now from the total amount, we can see Republican is leading now, with amount about $32.6K, almost 1.3 times of Democrat.

Since the nuber of candidates is quite different between Democrat and Republican, with 5 Democrat candidates while 17 for Republican candidates, let's see the average amount plot.

```
#plot of contribution amount by party
ggplot(aes(x = reorder(party, -(mean_party/1000)), y = mean_party/1000, fill = party), data = FL.contr_
  geom_bar(stat = 'identity') +
  geom_text(stat = 'identity', aes(label = round(mean_party/1000)),
            data = FL.contr_by_party, vjust = -0.4) +
  xlab('Party') +
  ylab('Contribution Amounts (Thousands)') +
  ggtitle('Average Contribution Amounts by Party') +
  scale_fill_manual(values = c('blue', 'gold', 'red'))
```



After looking at the average amount, we can see that now Democrat is almost 2.6 times of Republican in the average contribution amount.

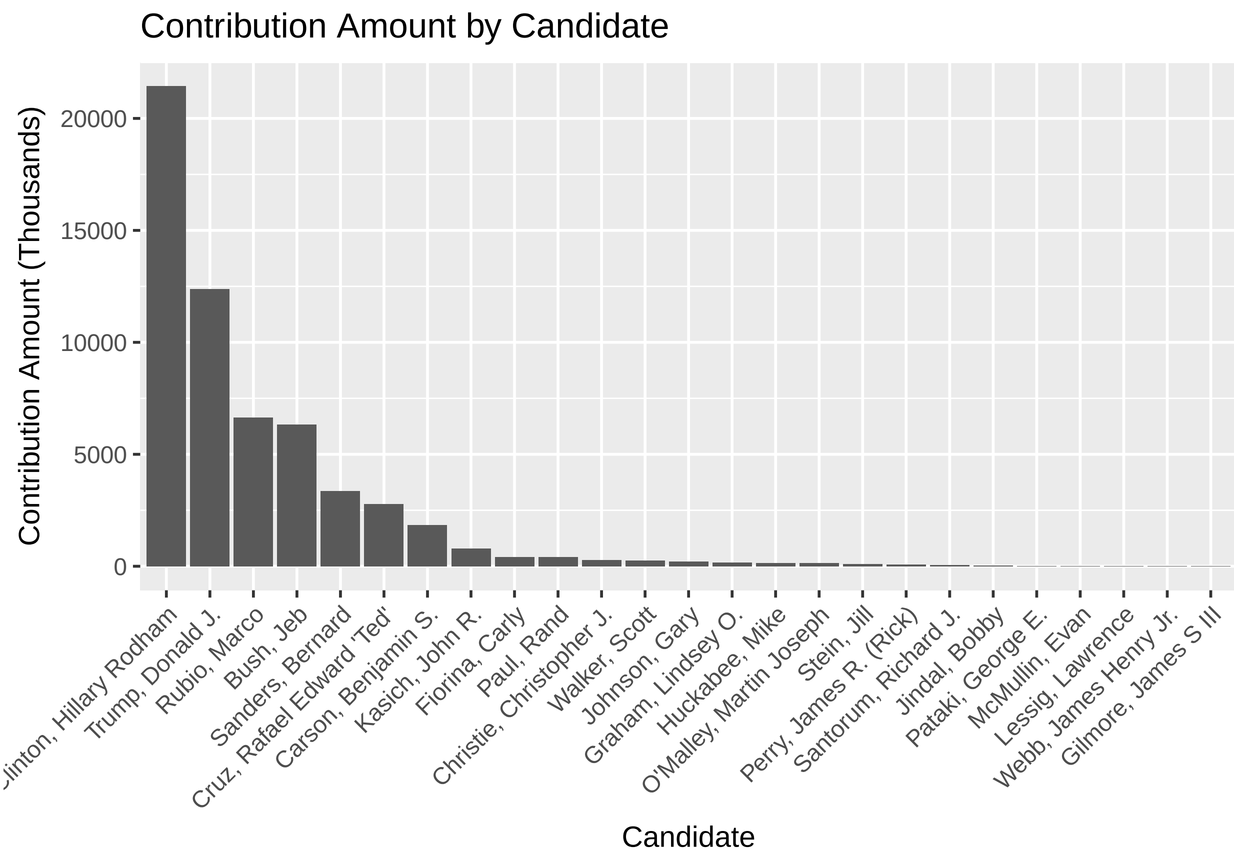
Now let's see the contribution amount distribution by candidates.

```
FL.contr_by_cand[rev(order(FL.contr_by_cand$sum_cand)),]
```

```
## # A tibble: 25 x 3
##                     cand_nm   sum_cand      n
##                       <chr>      <dbl>  <int>
##  1   Clinton, Hillary Rodham 21451718.3 176790
##  2          Trump, Donald J. 12385981.2  74156
##  3             Rubio, Marco  6634027.2  17993
##  4                Bush, Jeb  6325905.0   5401
##  5          Sanders, Bernard  3355944.0  79631
##  6 Cruz, Rafael Edward 'Ted'  2775052.6  27436
##  7       Carson, Benjamin S.  1843217.8  15496
##  8           Kasich, John R.   784383.7   1303
##  9           Fiorina, Carly   412640.1   2003
## 10               Paul, Rand   411286.7   1950
## # ... with 15 more rows
```

```
#plot of contribution amount by candidates
ggplot(aes(x = reorder(cand_nm, -(sum_cand/1000)), y = sum_cand/1000), data = FL.contr_by_cand) +
  geom_bar(stat = 'identity') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab('Candidate') +
  ylab('Contribution Amount (Thousands)') +
```

```r
  ggtitle('Contribution Amount by Candidate')
```



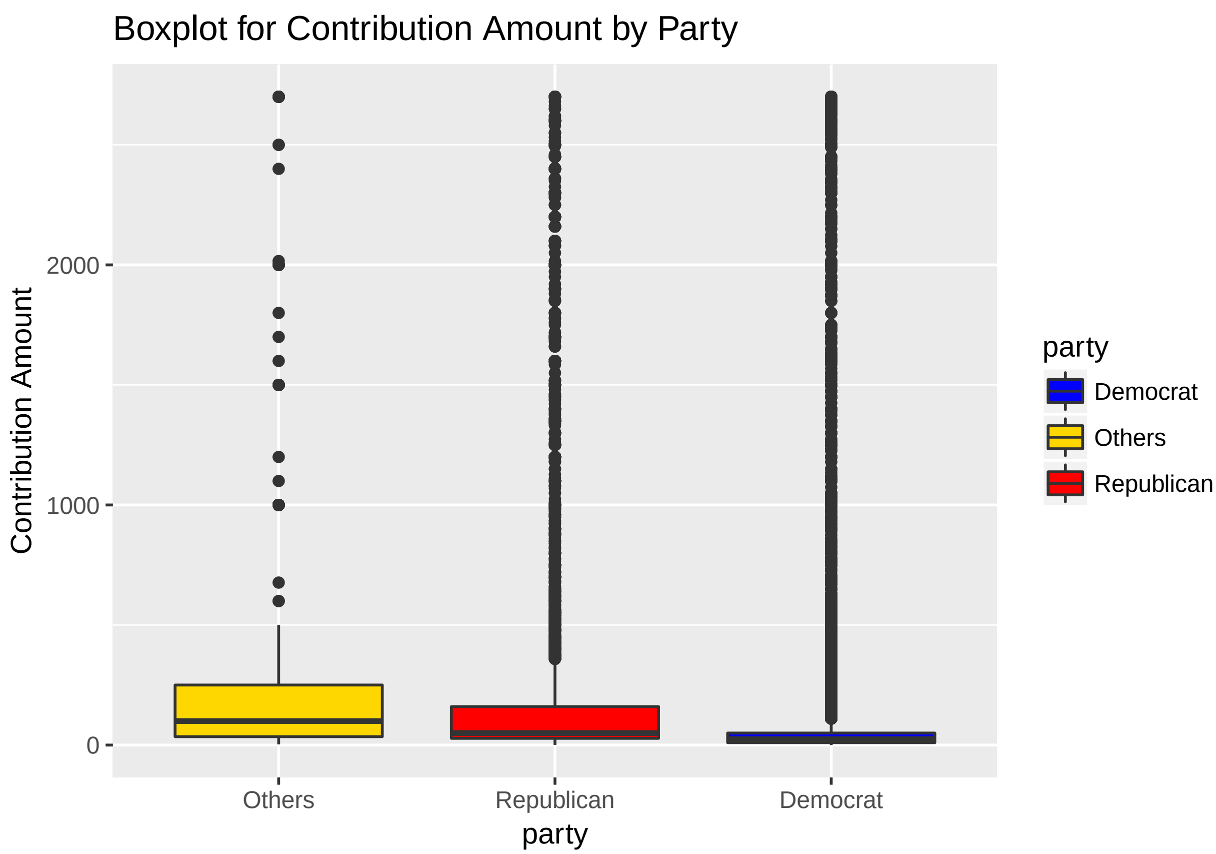

```r
sum(FL$contb_receipt_amt)
```

```
## [1] 57961929
```

Same with the nuber of contributions, Hillary Cliton received the most contribution amount with about $21,451.7K, but followed by Donald Trump with $12,386K, Rubio Marco received about $6,634K as the third most contribution amount.

Now let's see the contribution patterns between parties and candidates.

```r
#boxplot of contribution amount by party
ggplot(aes(x = reorder(party, -contb_receipt_amt), y = contb_receipt_amt, fill = party), data = FL) +
  geom_boxplot() +
  xlab('party') +
  ylab('Contribution Amount') +
  ggtitle('Boxplot for Contribution Amount by Party') +
  scale_fill_manual(values = c('blue', 'gold', 'red'))
```



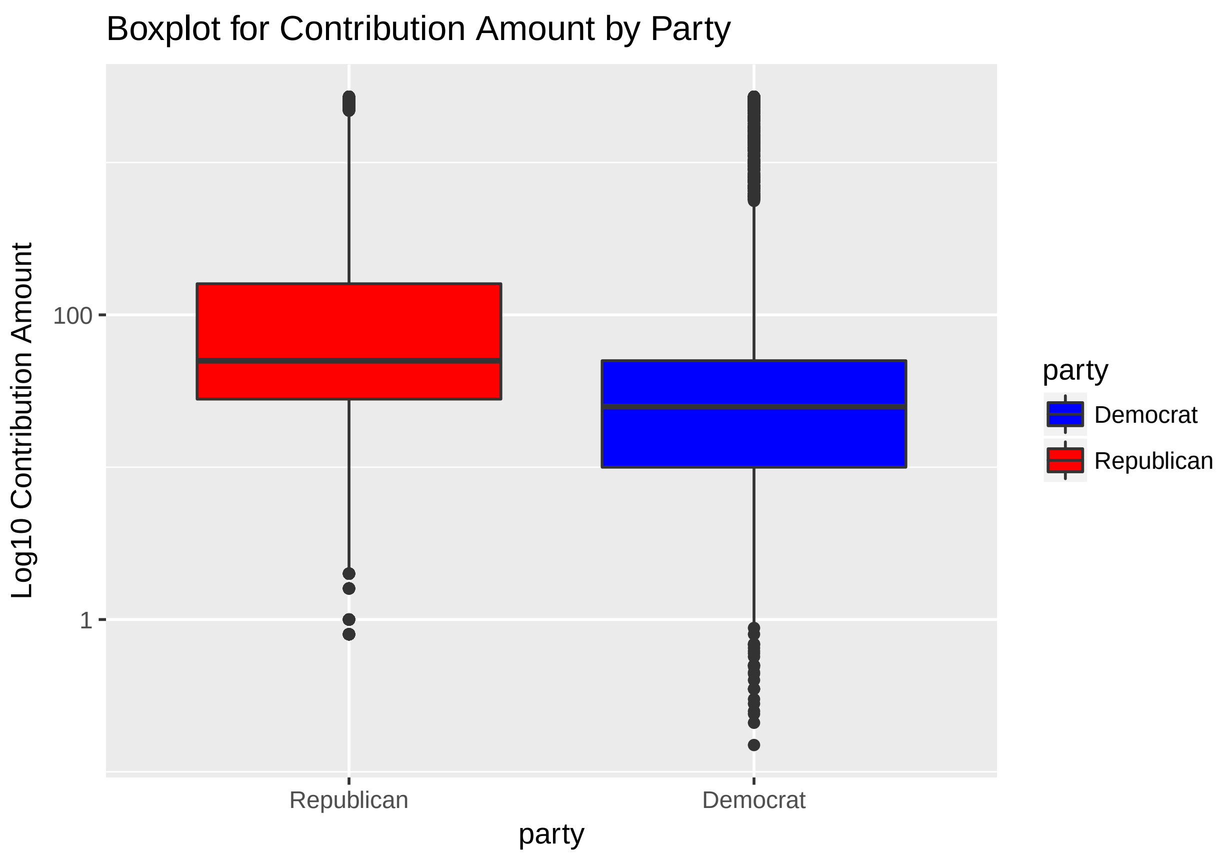
It's hard to compare contributions among all parties since there are too many outliers. And since now I am focuting on Demotrat party and Republican party (both contribution numbers and amount of other parties are not comparable with these 2 main party), the "Ohters" will be removed for now.

```
#remove "Others"
FL <- subset(FL, FL$cand_nm != "Johnson, Gary" & FL$cand_nm != "McMullin, Evan" & FL$cand_nm != "Stein,

by(FL$contb_receipt_amt, FL$party, summary)
```

```
## FL$party: Democrat
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.15   10.00   25.00   97.38   50.00 2700.00
## --------------------------------------------------------
## FL$party: Republican
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.8    28.0    50.0   222.0   160.0  2700.0
```

```
#Add log10: boxplot of contribution amount by party
ggplot(aes(x = reorder(party, -contb_receipt_amt), y = contb_receipt_amt, fill = party), data = FL) +
  geom_boxplot() +
  scale_y_log10() +
  xlab('party') +
  ylab('Log10 Contribution Amount') +
  ggtitle('Boxplot for Contribution Amount by Party') +
  scale_fill_manual(values = c('blue', 'red'))
```



From this boxplot we can see that Republican party has higher median and mean, and Democrat party has more variations, that is contributions to Democrat party has more great and small amount of donations.

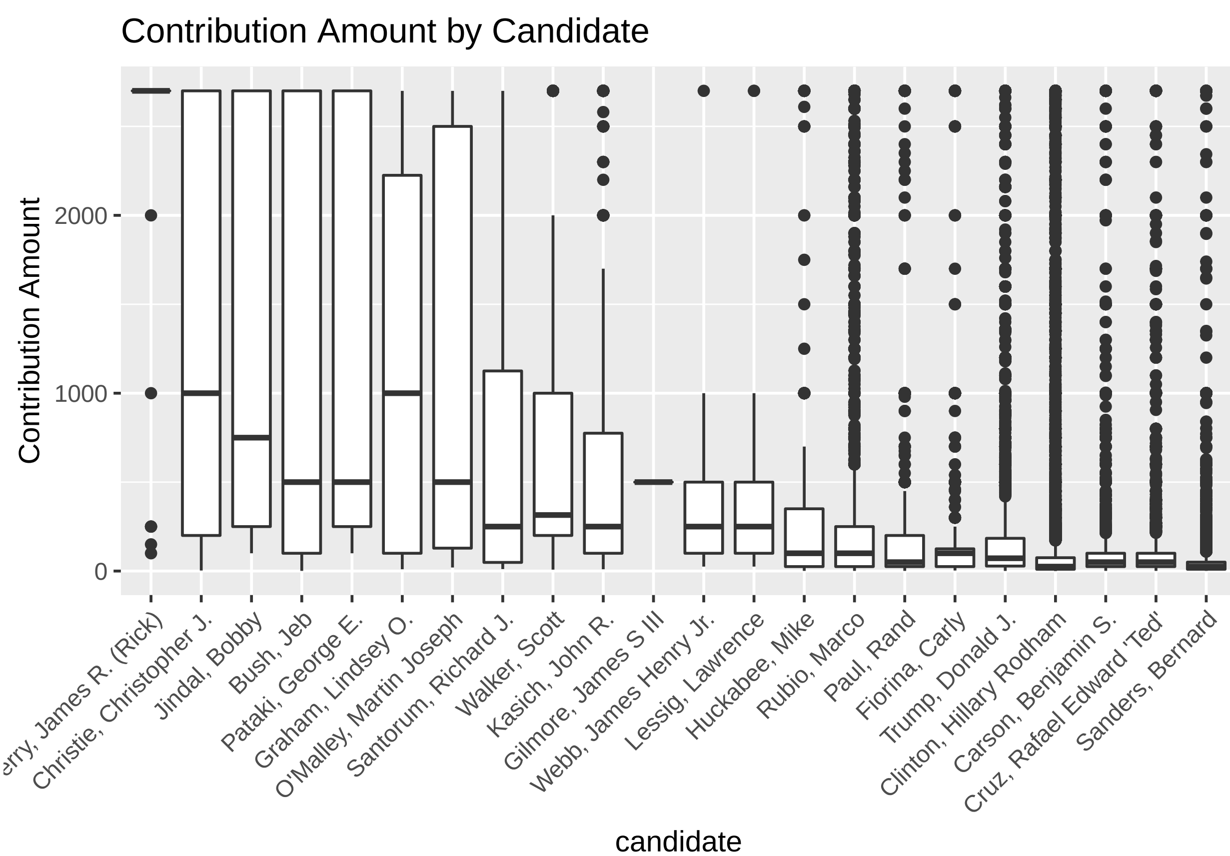
```
#boxplot of contribution amount by candidates in descending order
by(FL$contb_receipt_amt, FL$cand_nm, summary)
```

```
## FL$cand_nm: Bush, Jeb
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     100     500    1171    2700    2700
## --------------------------------------------------------
## FL$cand_nm: Carson, Benjamin S.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    25.0    50.0   118.9   100.0  2700.0
## --------------------------------------------------------
## FL$cand_nm: Christie, Christopher J.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       3     200    1000    1411    2700    2700
## --------------------------------------------------------
## FL$cand_nm: Clinton, Hillary Rodham
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.15   10.00   25.00  121.34   75.00 2700.00
## --------------------------------------------------------
## FL$cand_nm: Cruz, Rafael Edward 'Ted'
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    25.0    50.0   101.1   100.0  2700.0
## --------------------------------------------------------
## FL$cand_nm: Fiorina, Carly
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       3      25     100     206     125    2700
## --------------------------------------------------------
## FL$cand_nm: Gilmore, James S III
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     500     500     500     500     500     500
## --------------------------------------------------------
## FL$cand_nm: Graham, Lindsey O.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10     100    1000    1062    2225    2700
## --------------------------------------------------------
## FL$cand_nm: Huckabee, Mike
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    25.0   100.0   388.4   350.0  2700.0
## --------------------------------------------------------
## FL$cand_nm: Jindal, Bobby
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     100     250     750    1334    2700    2700
## --------------------------------------------------------
## FL$cand_nm: Kasich, John R.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10     100     250     602     775    2700
## --------------------------------------------------------
## FL$cand_nm: Lessig, Lawrence
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    25.0   100.0   250.0   422.6   500.0  2700.0
## --------------------------------------------------------
## FL$cand_nm: O'Malley, Martin Joseph
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.0   128.7   500.0  1039.6  2500.0  2700.0
## --------------------------------------------------------
## FL$cand_nm: Pataki, George E.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     100     250     500    1121    2700    2700
## --------------------------------------------------------
## FL$cand_nm: Paul, Rand
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    25.0    50.0   210.9   200.0  2700.0
## --------------------------------------------------------
## FL$cand_nm: Perry, James R. (Rick)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     100    2700    2700    2248    2700    2700
## --------------------------------------------------------
## FL$cand_nm: Rubio, Marco
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    25.0   100.0   368.7   250.0  2700.0
## --------------------------------------------------------
## FL$cand_nm: Sanders, Bernard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.00   25.00   42.14   50.00 2700.00
## --------------------------------------------------------
## FL$cand_nm: Santorum, Richard J.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.99   48.52  250.00  769.51 1125.00 2700.00
```

```
## -------------------------------------------------------------
## FL$cand_nm: Trump, Donald J.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.8    28.0    72.0   167.0   184.0  2700.0
## -------------------------------------------------------------
## FL$cand_nm: Walker, Scott
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.0   200.0   315.0   758.6  1000.0  2700.0
## -------------------------------------------------------------
## FL$cand_nm: Webb, James Henry Jr.
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      25     100     250     450     500    2700
```

```r
ggplot(aes(x = reorder(cand_nm, -contb_receipt_amt), y = contb_receipt_amt), data = FL) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab('candidate') +
  ylab('Contribution Amount') +
  ggtitle('Contribution Amount by Candidate')
```



Remember earilier we saw Hillary Cliton received the most contribution amount and followed by Donald Trump and then Rubio Marco received the third most contribution amount. Now from the boxplot, Perry James R. (Rick) has the higherst median althought very few number and amount of contributions. Donald Trump and Hilary Clition seem to have very low median, but they both has a lot of outliers above the median amount value.

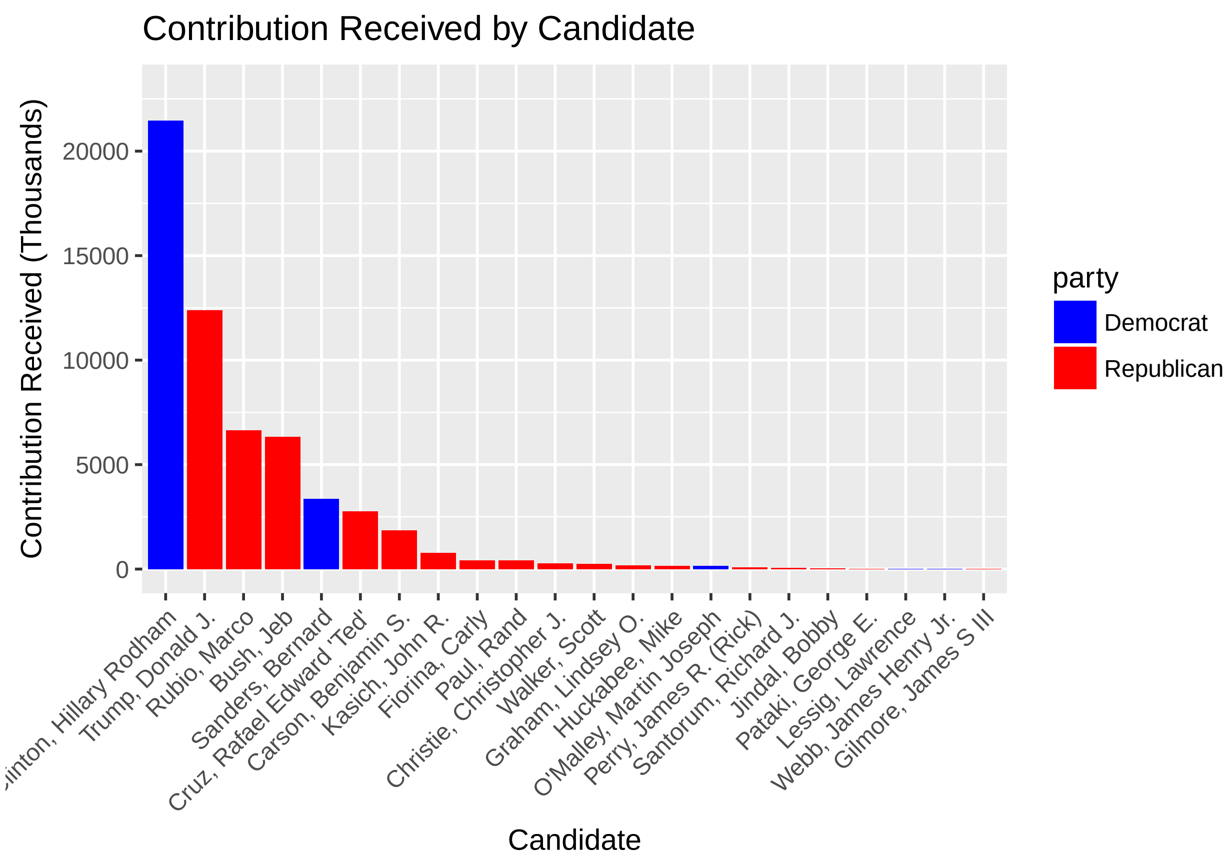
Now we can explore more within parties.

```r
#Create candidate datafram
candidate_group <- group_by(FL, party, cand_nm)
FL.contr_by_candidate <- summarize(candidate_group,
                                   sum_candidate = sum(contb_receipt_amt),
                                   mean_can = mean(contb_receipt_amt),
                                   n = n())

FL.contr_by_candidate[rev(order(FL.contr_by_candidate$sum_candidate)),]
```

```
## # A tibble: 22 x 5
## # Groups:   party [2]
##         party                    cand_nm sum_candidate   mean_can      n
##         <chr>                      <chr>         <dbl>      <dbl>  <int>
##  1   Democrat   Clinton, Hillary Rodham    21451718.3  121.34011 176790
##  2 Republican          Trump, Donald J.    12385981.2  167.02602  74156
##  3 Republican              Rubio, Marco     6634027.2  368.70045  17993
##  4 Republican                 Bush, Jeb     6325905.0 1171.24699   5401
##  5   Democrat          Sanders, Bernard     3355944.0   42.14369  79631
##  6 Republican Cruz, Rafael Edward 'Ted'     2775052.6  101.14640  27436
##  7 Republican       Carson, Benjamin S.     1843217.8  118.94797  15496
##  8 Republican           Kasich, John R.      784383.7  601.98285   1303
##  9 Republican            Fiorina, Carly      412640.1  206.01102   2003
## 10 Republican                Paul, Rand      411286.7  210.91623   1950
## # ... with 12 more rows
```

```r
#plot contribution by candidate
ggplot(aes(x = reorder(cand_nm, -sum_candidate/1000), y = sum_candidate/1000), data = FL.contr_by_candid
  geom_bar(aes(fill = party), stat = 'identity') +
  scale_y_continuous(limits = c(0, 23000)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab('Candidate') +
  ylab('Contribution Received (Thousands)') +
  ggtitle('Contribution Received by Candidate') +
  scale_fill_manual(values = c("blue", "red"))
```



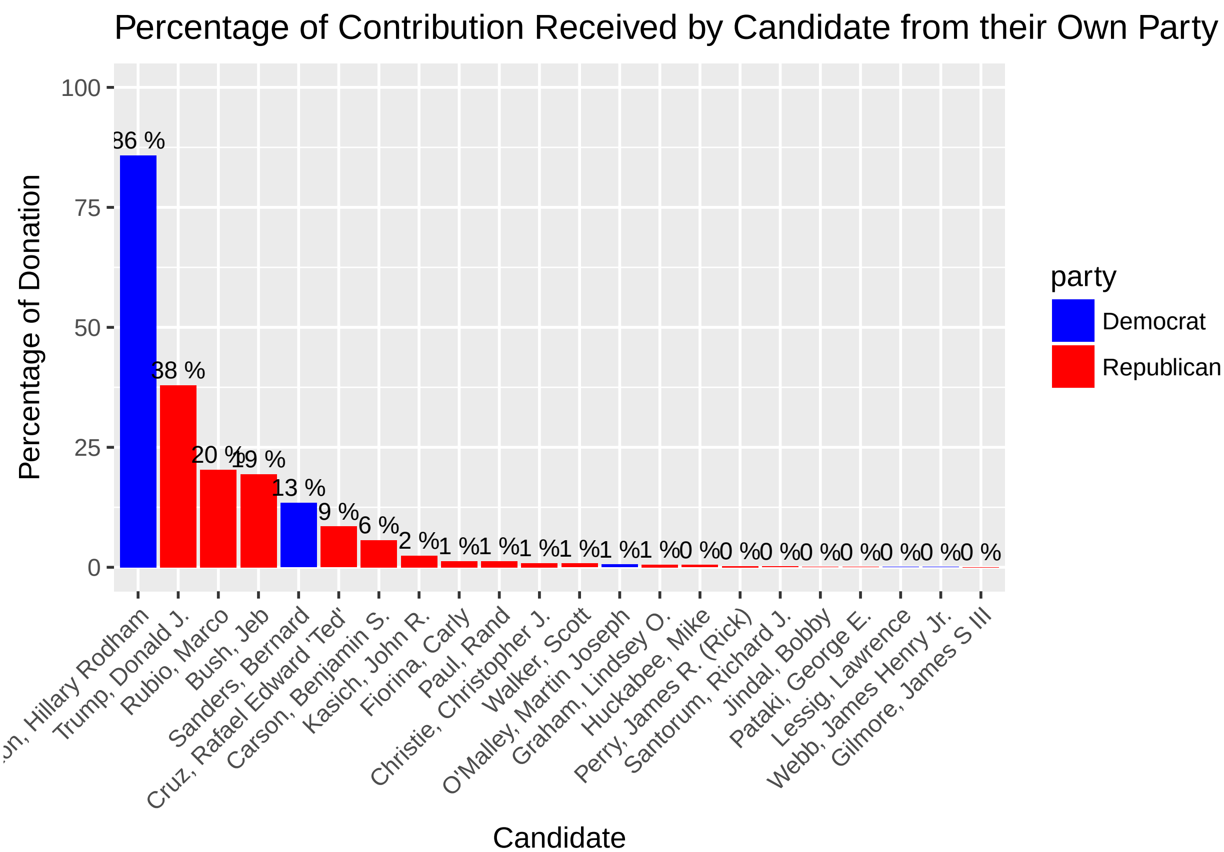

```r
#Create candidate_party dataframe for percentage calculation
can_party <- left_join(FL.contr_by_candidate, FL.contr_by_party, by = 'party')

can_party
```

```
## # A tibble: 22 x 9
## # Groups:   party [?]
##         party                    cand_nm sum_candidate   mean_can    n.x
##         <chr>                      <chr>         <dbl>      <dbl>  <int>
##  1   Democrat   Clinton, Hillary Rodham    21451718.3  121.34011 176790
##  2   Democrat          Lessig, Lawrence       16057.5  422.56579     38
##  3   Democrat   O'Malley, Martin Joseph      152823.3 1039.61449    147
##  4   Democrat          Sanders, Bernard     3355944.0   42.14369  79631
##  5   Democrat      Webb, James Henry Jr.      14400.0  450.00000     32
##  6 Republican                 Bush, Jeb     6325905.0 1171.24699   5401
##  7 Republican       Carson, Benjamin S.     1843217.8  118.94797  15496
##  8 Republican  Christie, Christopher J.      284939.2 1410.59035    202
##  9 Republican Cruz, Rafael Edward 'Ted'     2775052.6  101.14640  27436
## 10 Republican            Fiorina, Carly      412640.1  206.01102   2003
## # ... with 12 more rows, and 4 more variables: sum_party <dbl>,
## #   number_of_candidate <int>, mean_party <dbl>, n.y <int>
```

```r
ggplot(aes(x = reorder(cand_nm, -(sum_candidate/sum_party*100)), y = sum_candidate/sum_party*100), data
  geom_bar(aes(fill = party), stat = 'identity') +
  geom_text(stat='identity', aes(label = paste(round(100*sum_candidate/sum_party,0),'%')),
            size=3, data = can_party, vjust = -0.4)+
  scale_y_continuous(limits = c(0, 100)) +
```

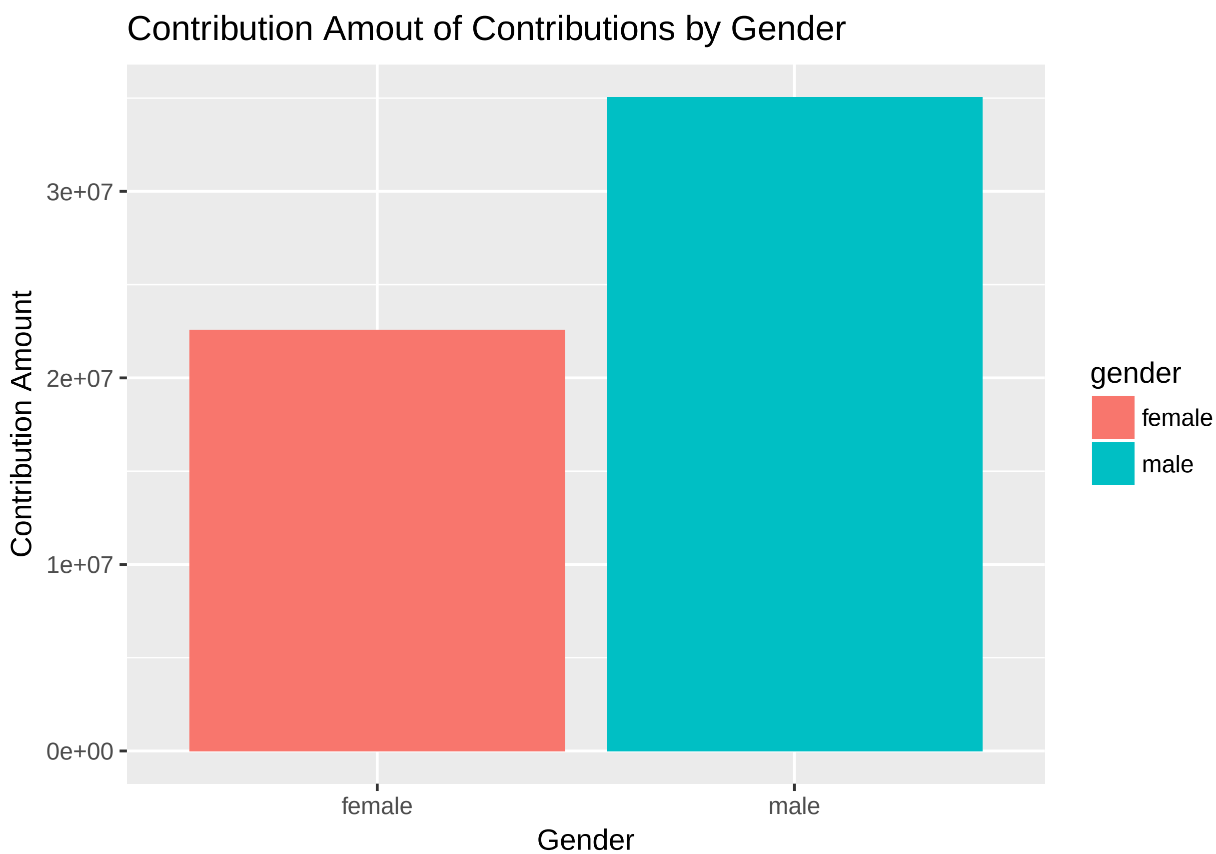
```
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab('Candidate') +
  ylab('Percentage of Donation') +
  ggtitle('Percentage of Contribution Received by Candidate from their Own Party') +
  scale_fill_manual(values = c("blue", 'red'))
```



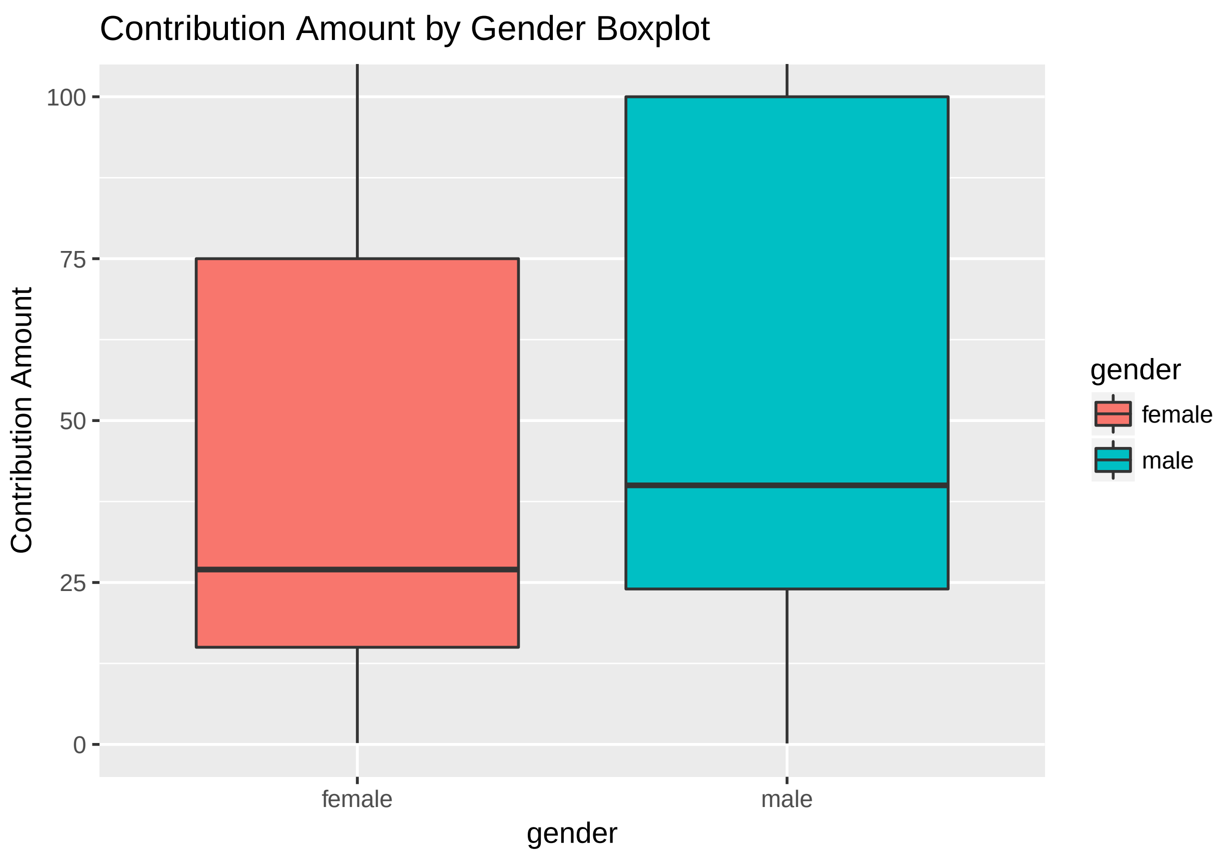
It's obvious to see that the donation within Democrat party was mainly received by Clinton Hillary, around 86%. For Republican party, Donald Trump received the most contributions with about 38%, Marco Rubio and Jeb Bush received similar amount of contibutions for about 20% and 19%. So for each party, we can also say that the majority of the donations were received by few candidates.

Now let's see is there any pattern of contribution amount by gender, compared with earlier contribution number by gender, which has no obvious pattern.

```
#plot of contribution amount by gender
ggplot(aes(x = gender, y = contb_receipt_amt, fill = gender),
       data = FL, vjust = -0.4) +
  geom_bar(stat = 'identity') +
  xlab('Gender') +
  ylab('Contribution Amount') +
  ggtitle('Contribution Amout of Contributions by Gender')
```




```r
#boxplot of contribution amount by gender
ggplot(aes(x = gender, y = contb_receipt_amt, fill = gender), data = FL) +
  geom_boxplot() +
  xlab('gender') +
  ylab('Contribution Amount') +
  ggtitle('Contribution Amount by Gender Boxplot') +
  coord_cartesian(ylim = c(0, 100))
```




```
by(FL$contb_receipt_amt, FL$gender, summary)
```

```
## FL$gender: female
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.21   15.00   27.00  113.44   75.00 2700.00
## --------------------------------------------------------
## FL$gender: male
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.15   24.00   40.00  171.27  100.00 2700.00
```

Now we see from the contribution amount and median and mean by gender, both plots show male contributions are higher than women contributions. On average male donated 1.5 times more than female.

```
# Top candidates
top_candidate <- c("Clinton, Hillary Rodham", "Trump, Donald J.", "Rubio, Marco", "Bush, Jeb", "Sanders

top_candidate
```

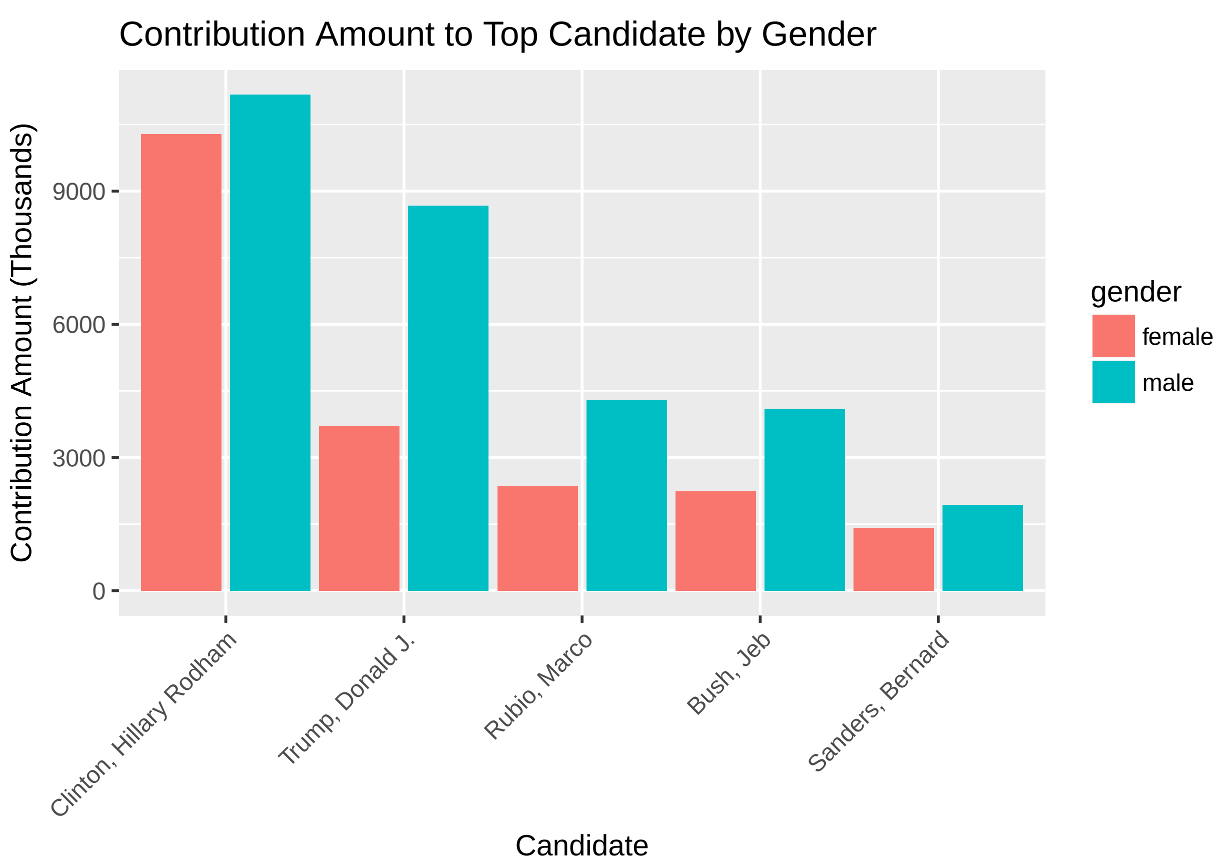
```
## [1] "Clinton, Hillary Rodham" "Trump, Donald J."
## [3] "Rubio, Marco"            "Bush, Jeb"
## [5] "Sanders, Bernard"
```

```
# Create gender_to_top_candidate dataframe for bar plot
FL.gen_to_top_candidate <- FL %>%
  filter(FL$cand_nm %in% top_candidate) %>%
  group_by(cand_nm, gender) %>%
  summarize(sum_gen_can = sum(contb_receipt_amt))
```

```r
FL.gen_to_top_candidate
```

```
## # A tibble: 10 x 3
## # Groups:   cand_nm [?]
##                     cand_nm gender sum_gen_can
##                       <chr>  <chr>       <dbl>
##  1                Bush, Jeb female     2233216
##  2                Bush, Jeb   male     4092689
##  3 Clinton, Hillary Rodham female    10283538
##  4 Clinton, Hillary Rodham   male    11168180
##  5             Rubio, Marco female     2347359
##  6             Rubio, Marco   male     4286668
##  7         Sanders, Bernard female     1419653
##  8         Sanders, Bernard   male     1936291
##  9         Trump, Donald J. female     3716207
## 10         Trump, Donald J.   male     8669774
```

```r
# plot of contribution of candidate by gender
ggplot(aes(x = reorder(cand_nm, -(sum_gen_can/1000)), y = sum_gen_can/1000, fill = gender),
       data = FL.gen_to_top_candidate) +
  geom_bar(stat = 'identity', position = position_dodge(width = 1)) +
  xlab('Candidate') +
  ylab('Contribution Amount (Thousands)') +
  ggtitle('Contribution Amount to Top Candidate by Gender') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



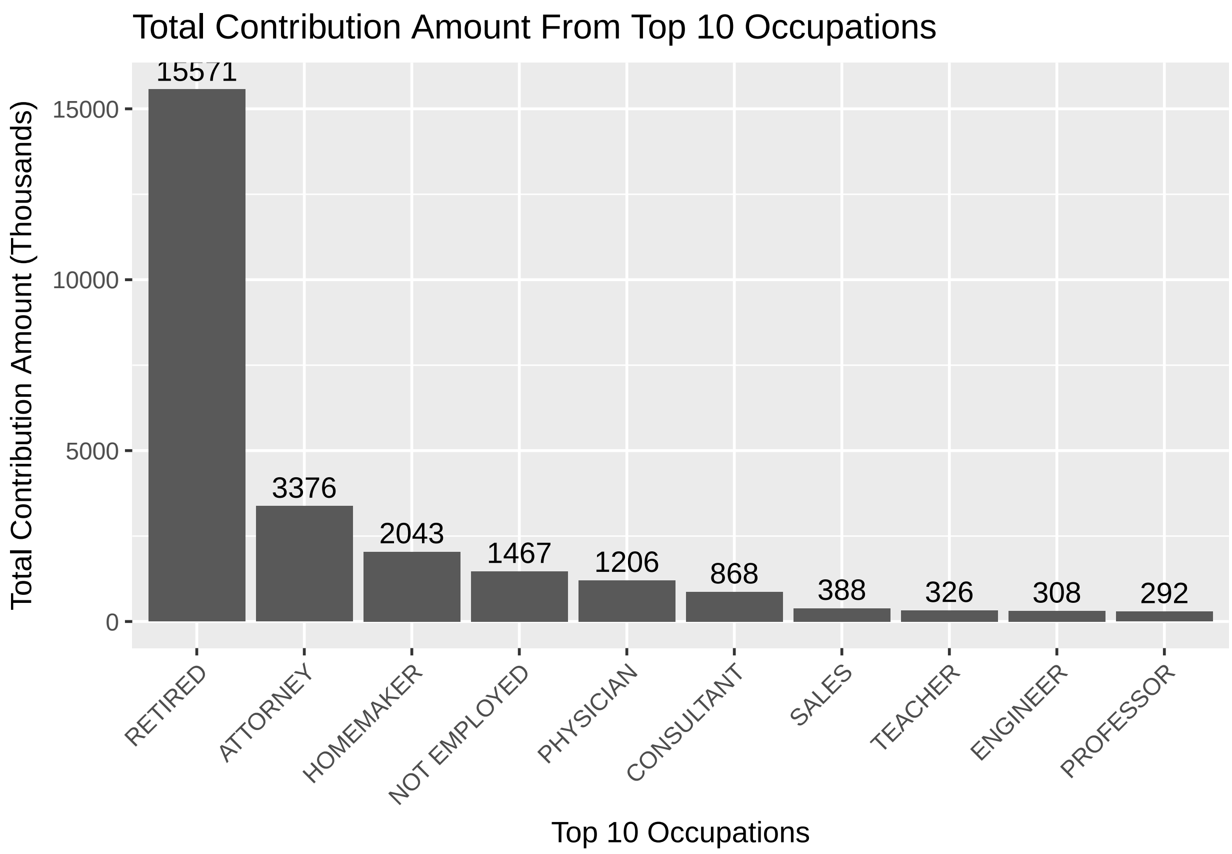
Compared with female's support to Donald Trump, Cliton Hillary has higher proportion of female contributions.

Earlier we noticed that the retired people made the most number of contributions, let's see if there is the same pattern when looking into total contribution amount.

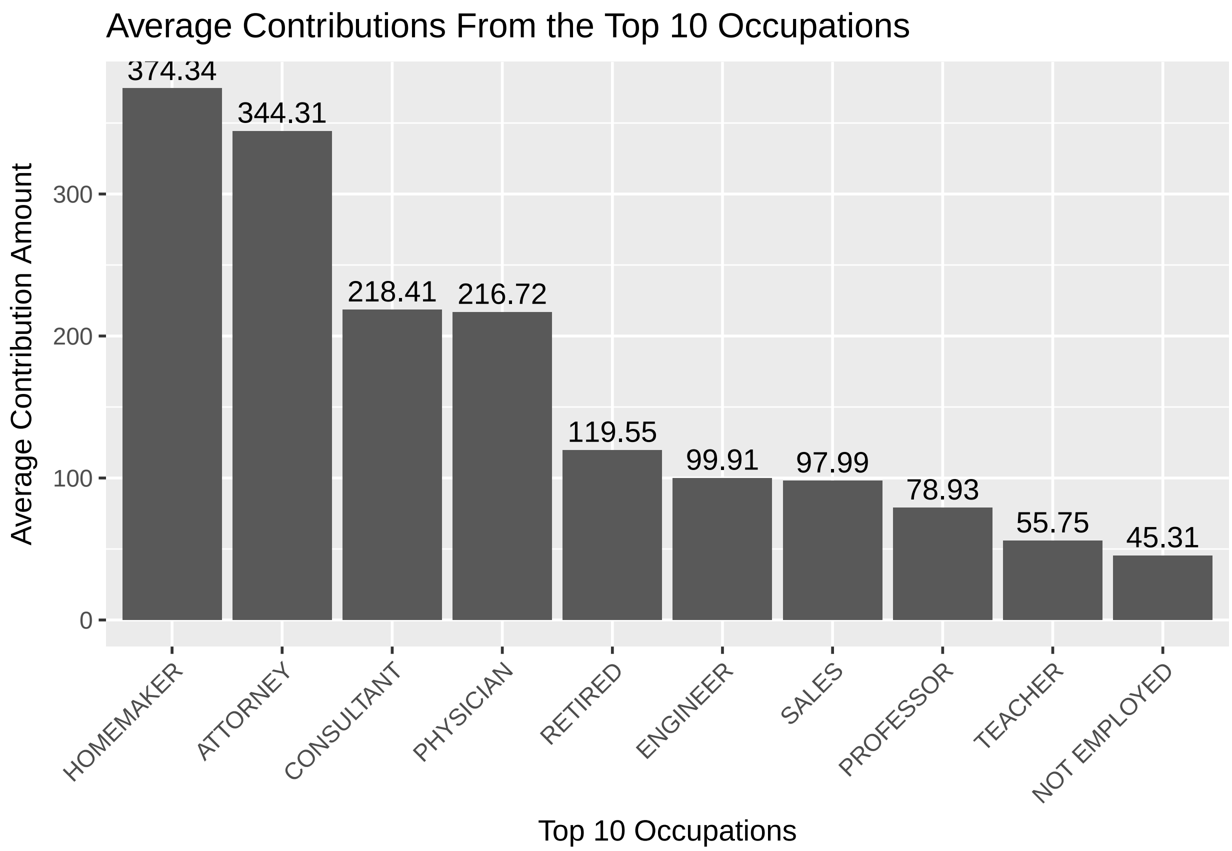
```
FL.contr_by_occupation
```

```
## # A tibble: 10 x 4
##    contbr_occupation sum_occupation mean_occupation      n
##                <chr>          <dbl>           <dbl>  <int>
##  1           RETIRED     15571366.7       119.54892 130251
##  2      NOT EMPLOYED      1466685.2        45.30722  32372
##  3          ATTORNEY      3375634.0       344.31191   9804
##  4           TEACHER       325942.4        55.74524   5847
##  5         PHYSICIAN      1206031.2       216.71721   5565
##  6         HOMEMAKER      2042790.3       374.34309   5457
##  7        CONSULTANT       868172.7       218.40822   3975
##  8             SALES       388058.8        97.99464   3960
##  9         PROFESSOR       291716.8        78.92769   3696
## 10          ENGINEER       308134.7        99.91397   3084
```

```
ggplot(aes(x = reorder(contbr_occupation, -(sum_occupation/1000)), y = sum_occupation/1000), data = FL.
  geom_bar(stat = 'identity') +
  geom_text(stat = 'identity', aes(label = round(sum_occupation/1000)), data = FL.contr_by_occupation, 
  xlab('Top 10 Occupations') +
  ylab('Total Contribution Amount (Thousands)') +
  ggtitle('Total Contribution Amount From Top 10 Occupations') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
ggplot(aes(x = reorder(contbr_occupation, -(mean_occupation)), y = round(mean_occupation,2)), data = FL
  geom_bar(stat = 'identity') +
  geom_text(stat = 'identity', aes(label = round(mean_occupation,2)), data = FL.contr_by_occupation, vj
  xlab('Top 10 Occupations') +
  ylab('Average Contribution Amount') +
  ggtitle('Average Contributions From the Top 10 Occupations') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

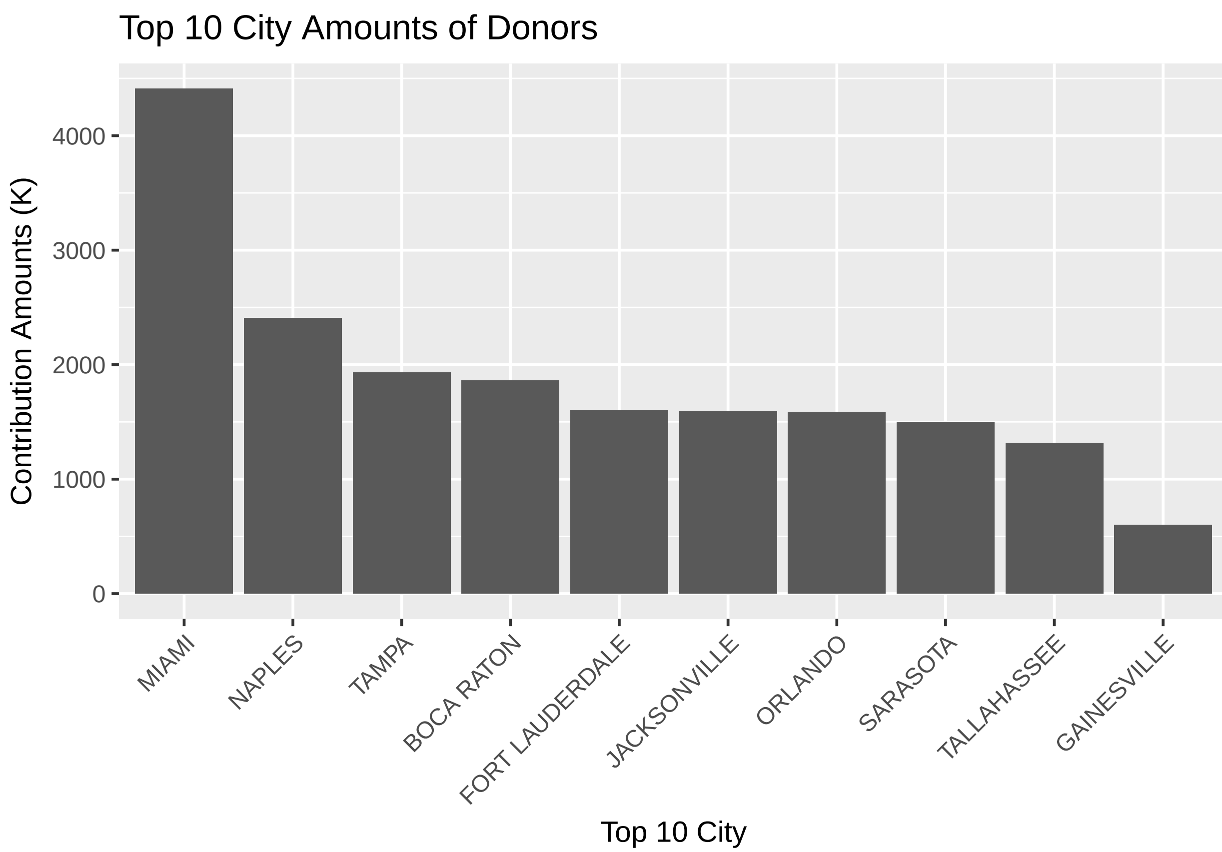


Retired people contributed the most, but when lookint into average contributions, homemaker comes to the first and attorney takes the second place. The unemployed people contribute the least on average among these 10 top occupations.

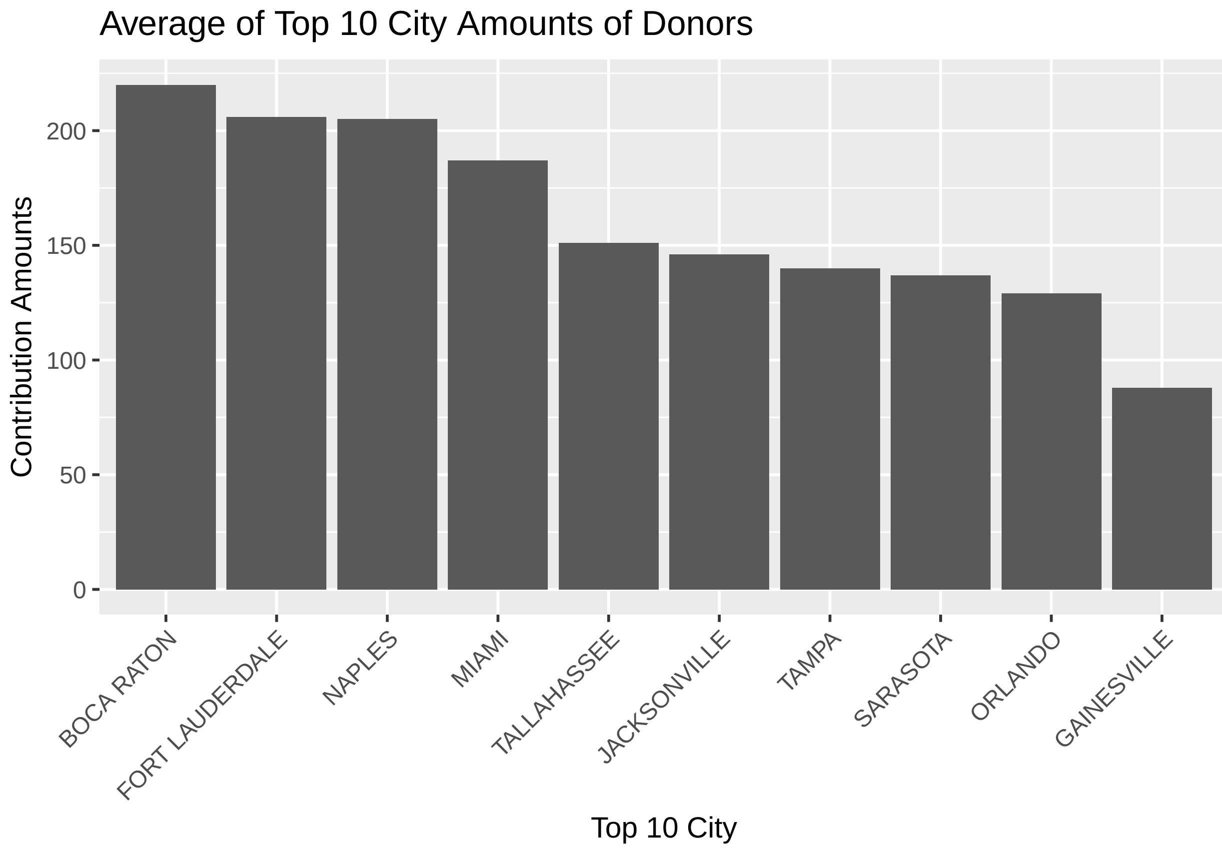
```r
FL.contr_by_city
```

```
## # A tibble: 10 x 4
##        contbr_city  sum_city mean_city     n
##              <chr>     <dbl>     <dbl> <int>
##  1           MIAMI 4408990.3 187.05941 23570
##  2           TAMPA 1931246.5 139.74288 13820
##  3         ORLANDO 1584454.4 128.74416 12307
##  4          NAPLES 2408467.2 204.87132 11756
##  5    JACKSONVILLE 1595578.3 146.06173 10924
##  6        SARASOTA 1500683.1 137.38745 10923
##  7     TALLAHASSEE 1313958.0 151.34277  8682
##  8      BOCA RATON 1861869.9 220.15725  8457
##  9 FORT LAUDERDALE 1602231.2 206.47310  7760
## 10     GAINESVILLE  602479.6  87.90189  6854
```

```r
#plot of contribution by city
ggplot(aes(x = reorder(contbr_city, -sum_city), y = round(sum_city/1000)), data = FL.contr_by_city) +
  geom_bar(stat = 'identity') +
  xlab('Top 10 City') +
  ylab('Contribution Amounts (K)') +
  ggtitle('Top 10 City Amounts of Donors') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
#plot of contribution by city
ggplot(aes(x = reorder(contbr_city, -mean_city), y = round(mean_city)), data = FL.contr_by_city) +
  geom_bar(stat = 'identity') +
  xlab('Top 10 City') +
  ylab('Contribution Amounts') +
  ggtitle('Average of Top 10 City Amounts of Donors') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Miami is still the top cicy in conbribtions but if talking about the average, Boca Raton has the highest contibution amount on average.

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?

- Republican party received almost 1.3 times contribution of Democrat party. But since there were 17 Republican candidates and 5 Democrat candidates, on average Democrat is almost 2.6 times of Republican.
- Hillary Cliton received the most contribution amount, which is 1.73 times of the second position Donald Trump received.
- For each party, the majority of contributions are received by a few candidates.
- Although there are comparable number of contributions for female and male, male donated more than female both in total and on average.
- Retired people contribute the most in total amount, while on average homemaker contribute most, followed by attorney.
- Miami contributes the most in total amount in Florida, but on average Boca Raton contributes the most.

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

One interesting finding is that on average of the amount, homemaker contributed the most. Although in total their contributions are not the most, and this is because the number of contributions is not much.

What was the strongest relationship you found?

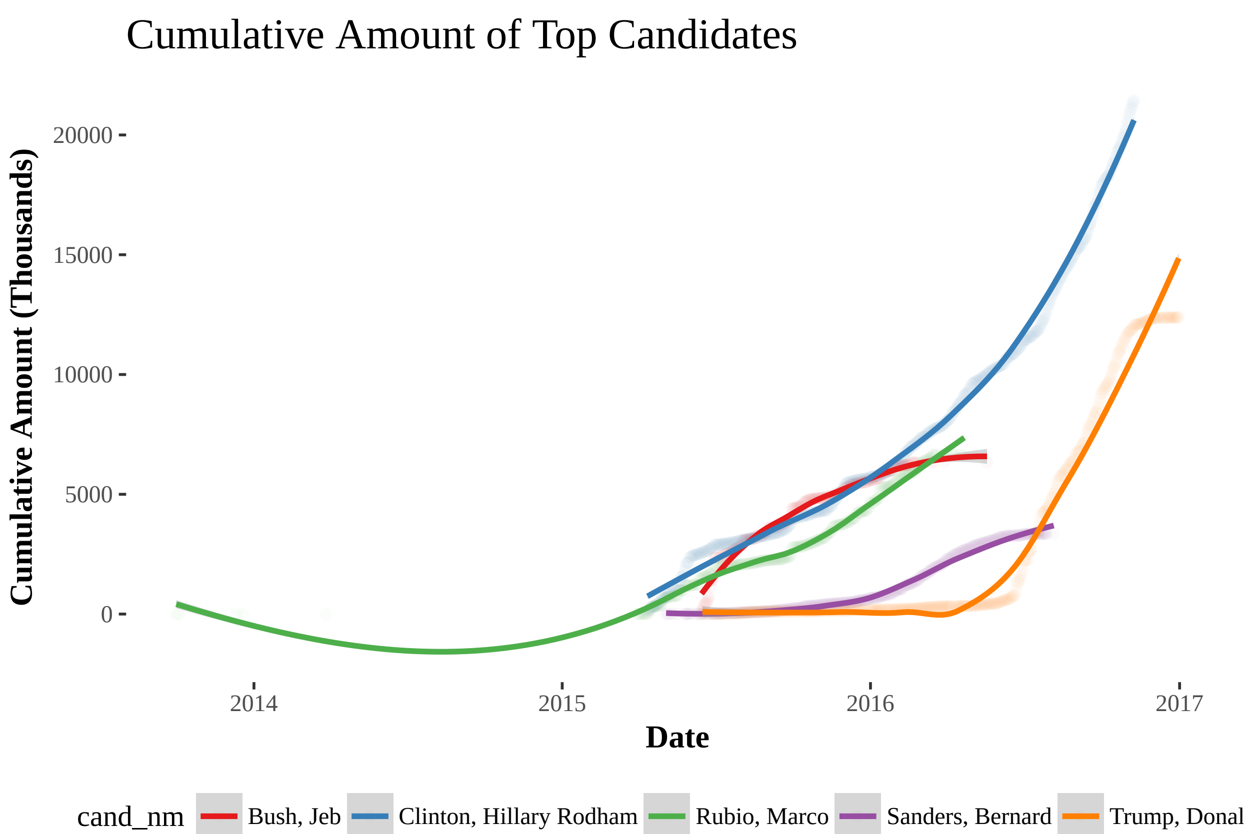
Since now only contribution amount is numeric variable, we cannot find the correlations.

Multivariate Plots Section

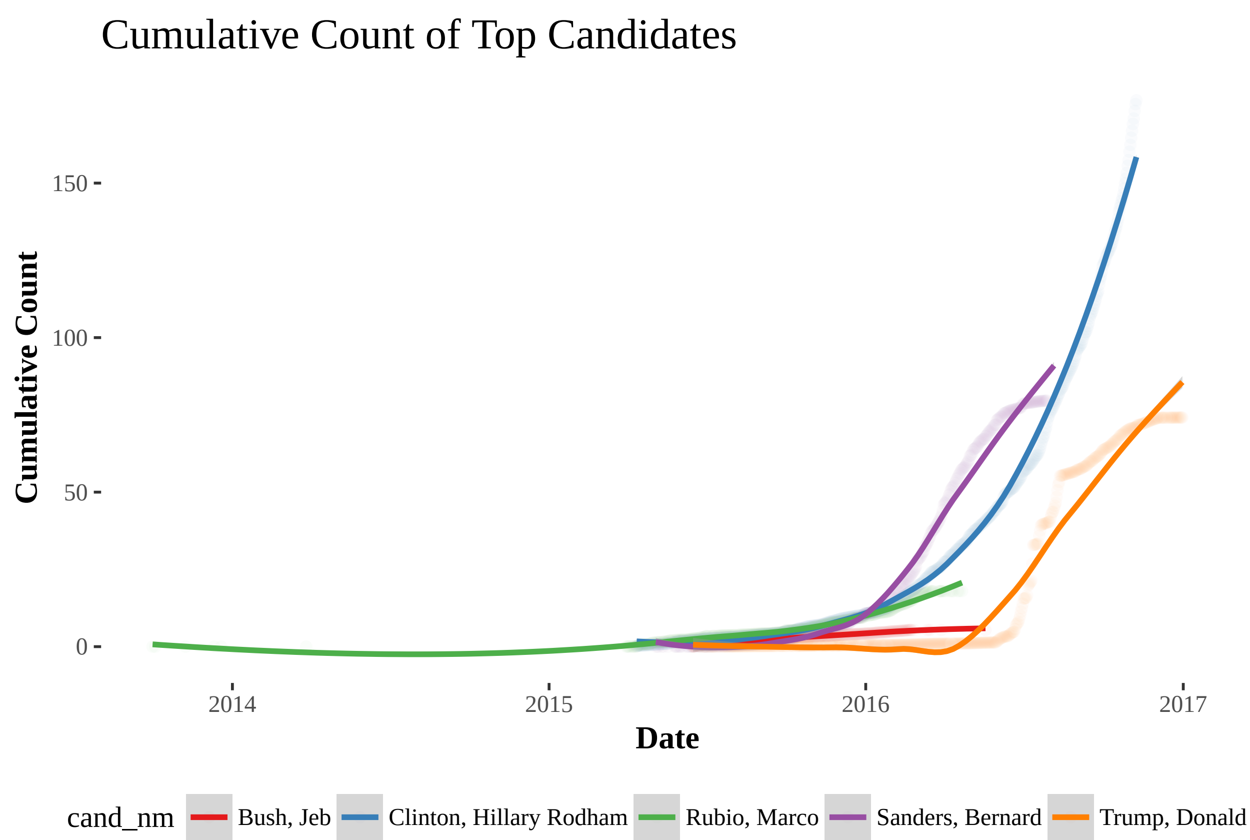
Now let's explore the top candidates' contribution trend over time.

```
#Cumulative dataframe
candidate_cum <- FL %>%
  filter(cand_nm %in% top_candidate) %>%
  group_by(cand_nm, date) %>%
  summarize(n = n(),
            total = sum(contb_receipt_amt)) %>%
  mutate(cumn = cumsum(n),
         cum_total = cumsum(total))
```

```
#Cumulative plot (Amount)
ggplot(aes(x = date, y = cum_total/1000, color = cand_nm),
       data = candidate_cum) +
  geom_point(alpha = 1/40) +
  geom_smooth(method = "loess") +
  theme_tufte() +
  xlab("Date") +
  ylab("Cumulative Amount (Thousands)") +
  ggtitle("Cumulative Amount of Top Candidates") +
  theme(plot.title = element_text(size = 16),
        axis.title = element_text(size = 12, face = "bold"),
        legend.position = "bottom") +
  scale_color_brewer(type = "qual", palette = "Set1")
```




```r
#Cumulative plot (number)
ggplot(aes(x = date, y = cumn/1000, color = cand_nm),
       data = candidate_cum) +
  geom_point(alpha = 1/40) +
  geom_smooth(method = "loess") +
  theme_tufte() +
  xlab("Date") +
  ylab("Cumulative Count") +
  ggtitle("Cumulative Count of Top Candidates") +
  theme(plot.title = element_text(size = 16),
        axis.title = element_text(size = 12, face = "bold"),
        legend.position = "bottom") +
  scale_color_brewer(type = "qual", palette = "Set1")
```



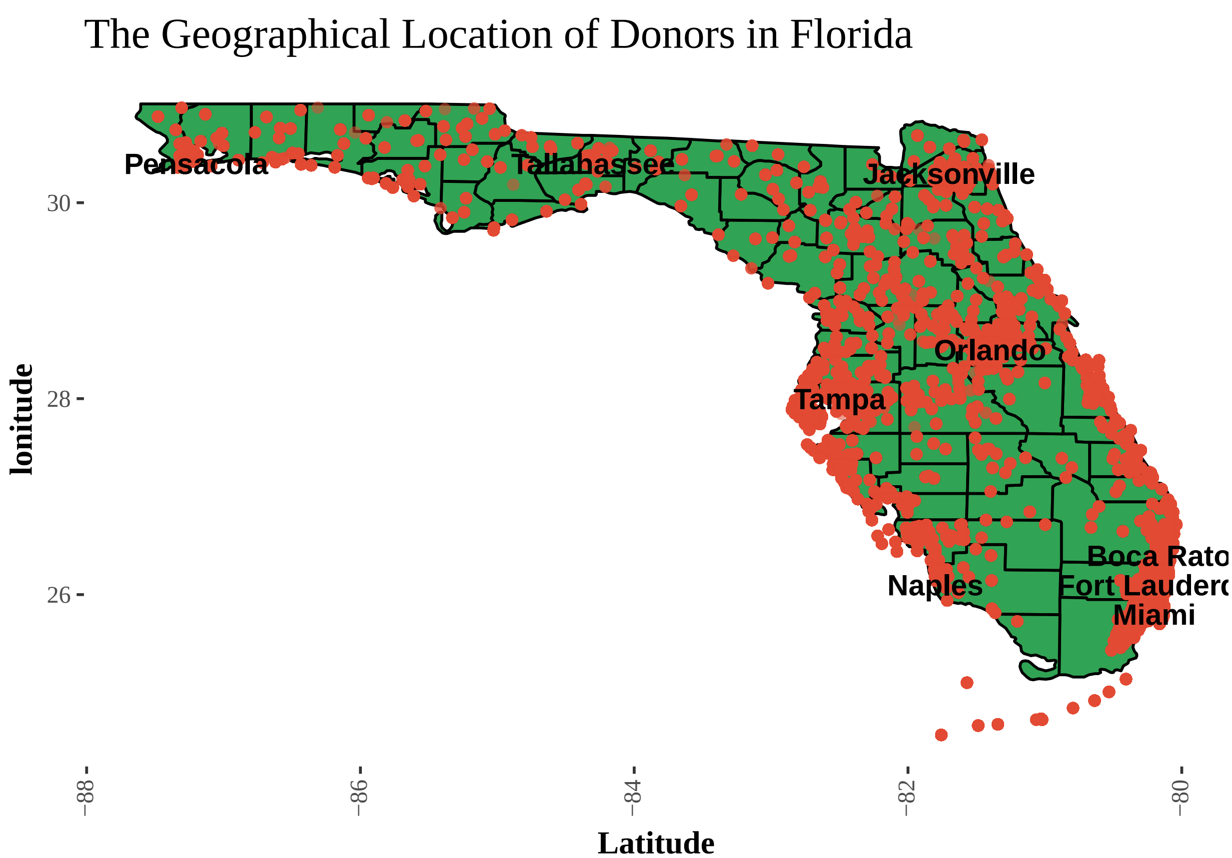
We can notice that Cliton Hillary and Donald Trump both have steep trend in cumulative donation amount and count, but the thrend of Donald Trump are both under Cliton Hillary's trends, that means Cliton Hillary received higher contributions earlier than Donald Trump. Bernard Sanders's donation count trend is even more steep than Cliton Hillary but the donation amount trend is not, this tell us most supporters for Bernard Sanders are smaller donors.

Let's see where does the donation come from.

```
#Dataframe for main cities longitude and latitude
map_FL = map_data('county', 'florida')
main_area <- data.frame(
  city = c("Miami",
           "Naples",
           "Tampa",
           "Boca Raton",
           "Fort Lauderdale",
           "Jacksonville",
           "Orlando",
           "Pensacola",
           "Tallahassee"),
  longitude = c(-80.2, -81.8, -82.5, -80.1, -80.1, -81.7, -81.4, -87.2, -84.3),
  latitude  = c(25.8, 26.1, 28.0, 26.4, 26.1, 30.3, 28.5, 30.4, 30.4)
)

ggplot() +
  geom_polygon(data = map_FL, aes(x = long, y = lat, group = group),
               colour = "black", fill = "#31a354") +
```

```r
geom_point(data = FL,
           aes(x = longitude, y = latitude, alpha = 1/50),
           color = "#e34a33") +
geom_text(data = main_area,
          aes(longitude, latitude, label = city),
          size = 4, fontface = "bold", colour = "black") +
theme_tufte() +
xlab("Latitude") +
ylab("lonitude") +
ggtitle("The Geographical Location of Donors in Florida") +
theme(legend.position = "none",
      plot.title = element_text(size = 16),
      axis.title = element_text(size = 12, face = "bold"),
      axis.text.x = element_text(angle = 90, hjust = 1, vjust = .4))
```



This is only the donors' geographical distribution. We can see the major cities have more contributors.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation.

- While closer to the election day, more big pocket donors supported both Hillary Clinton and Donald Trump.

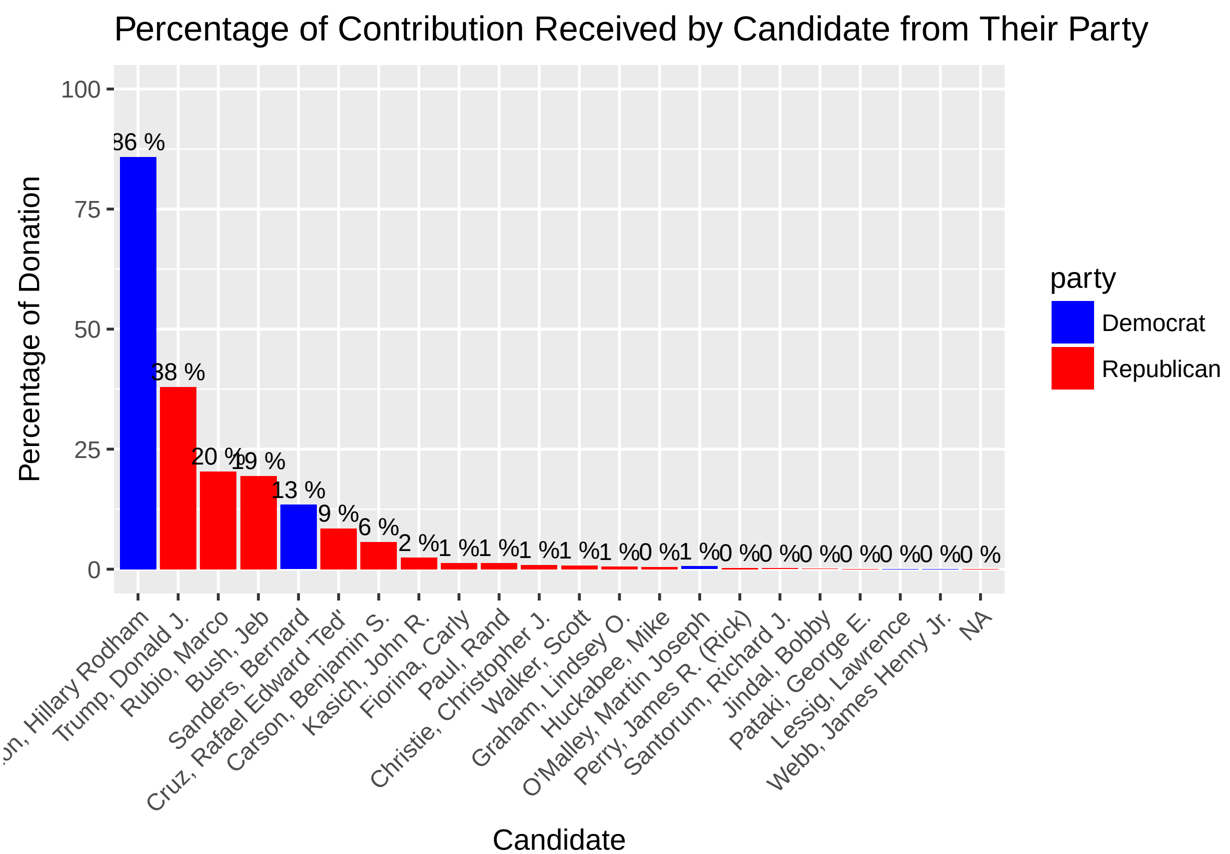
- Clinton Hillary received more donations count and amount than Donald Trump, and also got support earlier than Donald Trump.

Were there any interesting or surprising interactions between features?

For a certain period of time, Bernard Sanders gained more popularity than Hillary Clinton. But the amount of donations is not. That means supporters of Bernard Sanders are more small pocket donors.

Final Plots and Summary

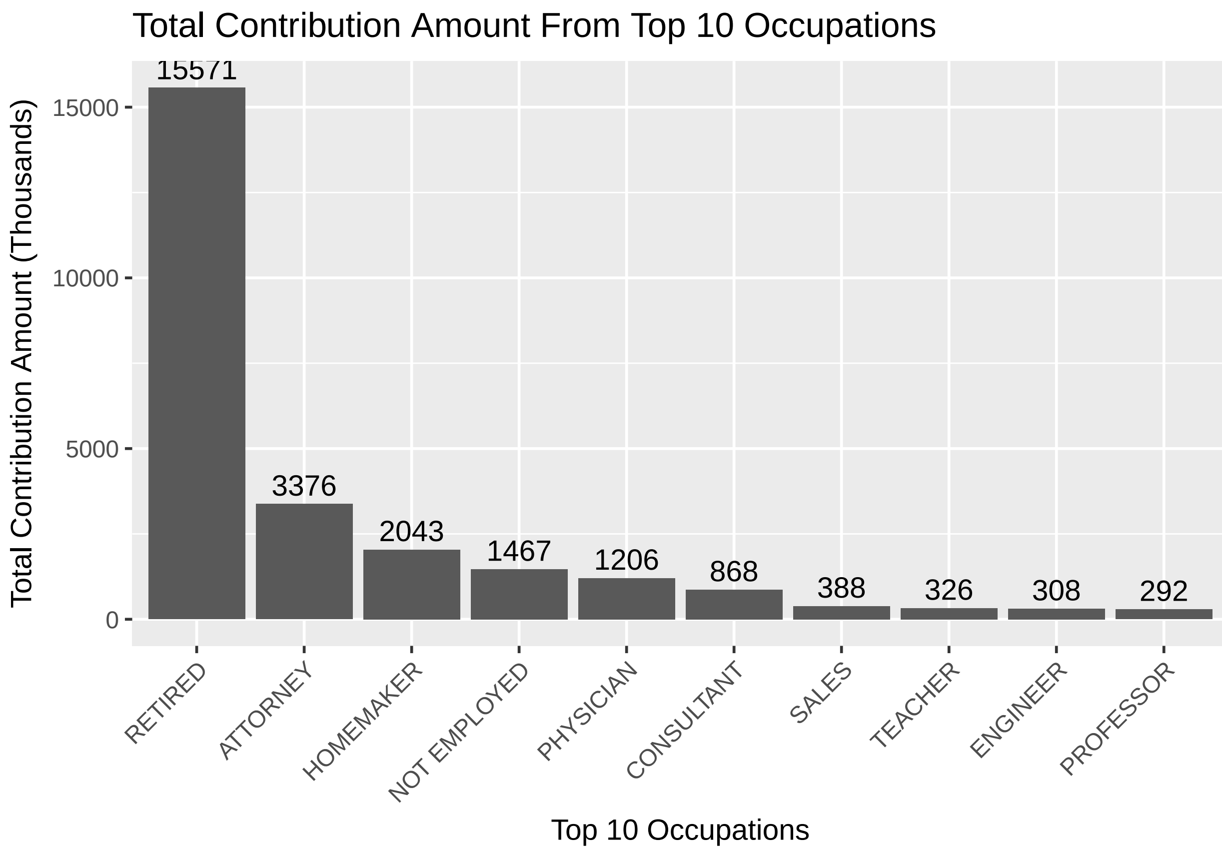
Plot One (Most Donations went toward a few candidates)



Description One

This plot demonstrates the contributions distribution for each candidate within each party. Actually it is uneven, most of the contributinos were received by a few candidates in each party, especially in Democrat, 99% of the donations for Democrat went to two candidates and Hillary Cliton took up to 86%. Republican got a few more candidates to take the majority of the contribution, and Donald Trump took the most among this party with 38%, Marco Rubio and Jeb Bush took comparable portion with 20% and 19% respectively.

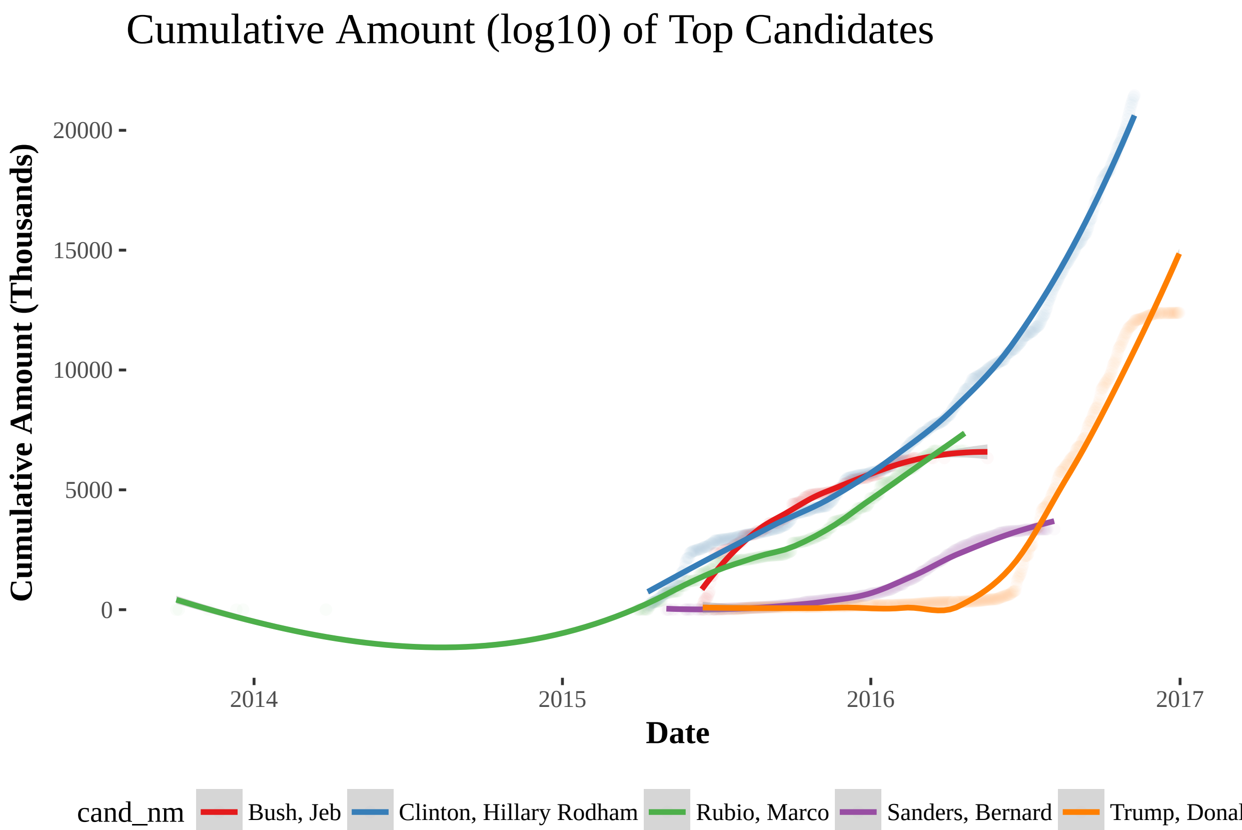
Plot Two (Contribution by Occupation)



Description Two

The total contribution across occupations differ substantially. And the most contribution amount is from retired people, which is surprised to me because if I was asked which occupation contributed the most to presidential candidates in Florida in 2016 before exploring the data, I would have guessed 'lawyers' or 'CEOs', but actually it's not. And Teachers, Engineers and professors are the last three among these top 10 occupations, we may need more knowledge of industry political background to explain these findings.

Plot Three



Description Three

Hillary Clinton dominated the contribution amount especially when close to the election day. And the closer to the election, also the more money came to Donald Trump. But since the date when Donald Trump started to receive higher contributions was later than Hillary Clinton, so even later when close to the election day they got almost the same slope of the cumulative amount trend, but Hillary Cliton still received higher total amount when comparing at the same time.

Reflection

Challenges and Struggles

Throughout the analysis, I had to deal with several issues:

- The original dataset did not contain gender information. In order to analyze the relationshihp between gender and donations, I have added gender column with "gender" package which is the prediction base on donors' first name.
- I have also added latitude and longitude columns with "zipcode" package to see clearer of donor's geographic location.
- Some observations need to be removed which with contribution amount either exceed $2700 or the negative amount values. This is because the contribution limits is 2700 and negative values are refunds.

- It's not that easy for me to deal with every dataframe and plot since I am still practicing myself to be more familar with them.

Success

I have learned a lot from this project. The ggplot2 and dplyr packages are so important in this analysis, and there are also some other powerful packages I haven't used before, such as gender and zipcode. It's really a very good practice.

Conclusion

After analyzing Florida 2016 presidential election financial donation data, I have some interesting findings:

- Few candidates received the most donations.
- Republican received higher amount of contributions than Democrat while they got comparable number of contributors.
- Male contributed more than female while the number of male contributors and female contributors are comparable.
- The retired people are the largest contribution group.
- The closer to the election day, the more mondy came to both Hillary Cliton and Donald Trump. But intotal Hillary Cliton received the most contributions.

Future Work

This is only the analysis from Florida, it would be interesting to also analyze campaign finance data for other "swing" states, for example Ohio, to see if there same findings and what's different. And some "blue" or "red" states' data are also worth to analyze since they should have different patterns.

Reference

https://stackoverflow.com/questions/4066607/reading-a-csv-file-with-repeated-row-names-in-r https://www.rdocumentation.org/packages/choroplethrZip/versions/1.5.0 https://stat.ethz.ch/pipermail/r-help/2006-February/088987.html https://stackoverflow.com/questions/16961921/plot-data-in-descending-order-as-appears-in-data-f
https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/ggmap/ggmapCheatsheet.pdf