Analysis of 2012 Presidential Campaign Contributions in New York State

Max Edwards

November 16, 2015

Introduction

This is an exploration of 2012 US presidential campaign donations in New York State. The dataset used was downloaded from the Federal Election Committee which oversees the public funding of Presidential elections. Based on reviewing the data dictionary provided by the FEC, I will try to answer a series of questions by using various data exploration techniques in R. Questions to consider would be: Who received the most contributions. Where are they coming from (city/zip)? How much money is a typical donation? What are the differences in donations between Republicans and Democrats? Who are the people that are donating (i.e. occupation) and when do they donate during the campaign cycle?

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Univariate Exploration

I will start out by exploring the entire data set general features then each variable that I think is useful for the purposes of answering the questions above. Let's start with the date variable `contb_receipt_dt` as the timeline prior to the election is really important.

Dataset Features

```
## 'data.frame':    420359 obs. of  18 variables:
##  $ cmte_id          : chr  "C00495820" "C00495820" "C00431445" "C00431445" ...
##  $ cand_id          : chr  "P80000748" "P80000748" "P80003338" "P80003338" ...
##  $ cand_nm          : chr  "Paul, Ron" "Paul, Ron" "Obama, Barack" "Obama, Barack" ...
##  $ contbr_nm        : chr  "NOBIS, JOSEPH FRANK JR." "MCMONAGLE, SEAN" "MASON, TIMOTHY" "WHITE, SARA
##  $ contbr_city      : chr  "TABERG" "EAST AMHERST" "NEW YORK" "NEW YORK" ...
##  $ contbr_st        : chr  "NY" "NY" "NY" "NY" ...
##  $ contbr_zip       : chr  "13471" "14051" "100096503" "10025" ...
##  $ contbr_employer  : chr  "SELF" "BLT CHEMBULK" "SELF-EMPLOYED" "RETIRED" ...
##  $ contbr_occupation: chr  "TRUCK DRIVER" "INTERN" "WRITER" "RETIRED" ...
##  $ contb_receipt_amt: num  1000 100 50 100 300 100 50 100 50 100 ...
##  $ contb_receipt_dt : chr  "06-DEC-11" "06-DEC-11" "06-SEP-11" "27-SEP-11" ...
##  $ receipt_desc     : chr  "" "" "" "" ...
```

```
##  $ memo_cd          : chr  "" "" "" "" ...
##  $ memo_text        : chr  "" "" "" "" ...
##  $ form_tp          : chr  "SA17A" "SA17A" "SA17A" "SA17A" ...
##  $ file_num         : int  779227 779227 756218 756218 756218 756218 756218 756218 756218 756218 ...
##  $ tran_id          : chr  "0925649" "0925702" "C12008058" "C12249568" ...
##  $ election_tp      : chr  "P2012" "P2012" "P2012" "P2012" ...
```

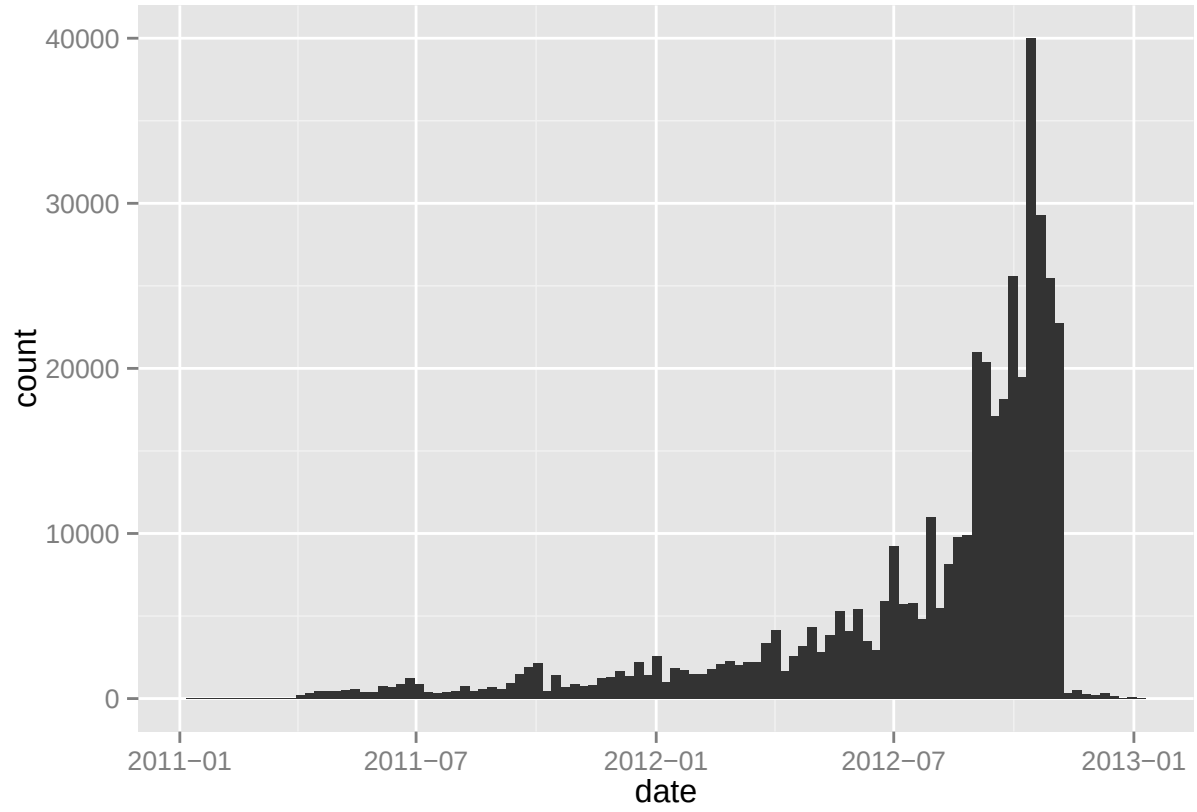
The starting data has 420359 rows and 18 columns. Each row represents a transaction (donation) and each column provides some information about the transaction.

Date Variable

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2011-01-14" "2012-06-23" "2012-09-07" "2012-07-24" "2012-10-15"
##         Max.
## "2012-12-31"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -55.0    22.0    60.0   104.2   136.0   662.0
```

I created date as a Date class and then determined when the Republican primary was finished. Per Mitt Romney's Presidential Campaign Wiki, he formally accepted the nomination on August 30th in Tampa, Florida so I used this as the cutoff date to create a days_from_election variable. This will help create some plots and summarizing data over time. It looks like the first contribution was on 2011-01-14 and the last contribution was on 2012-12-31. Interesting considering the election ended in November.

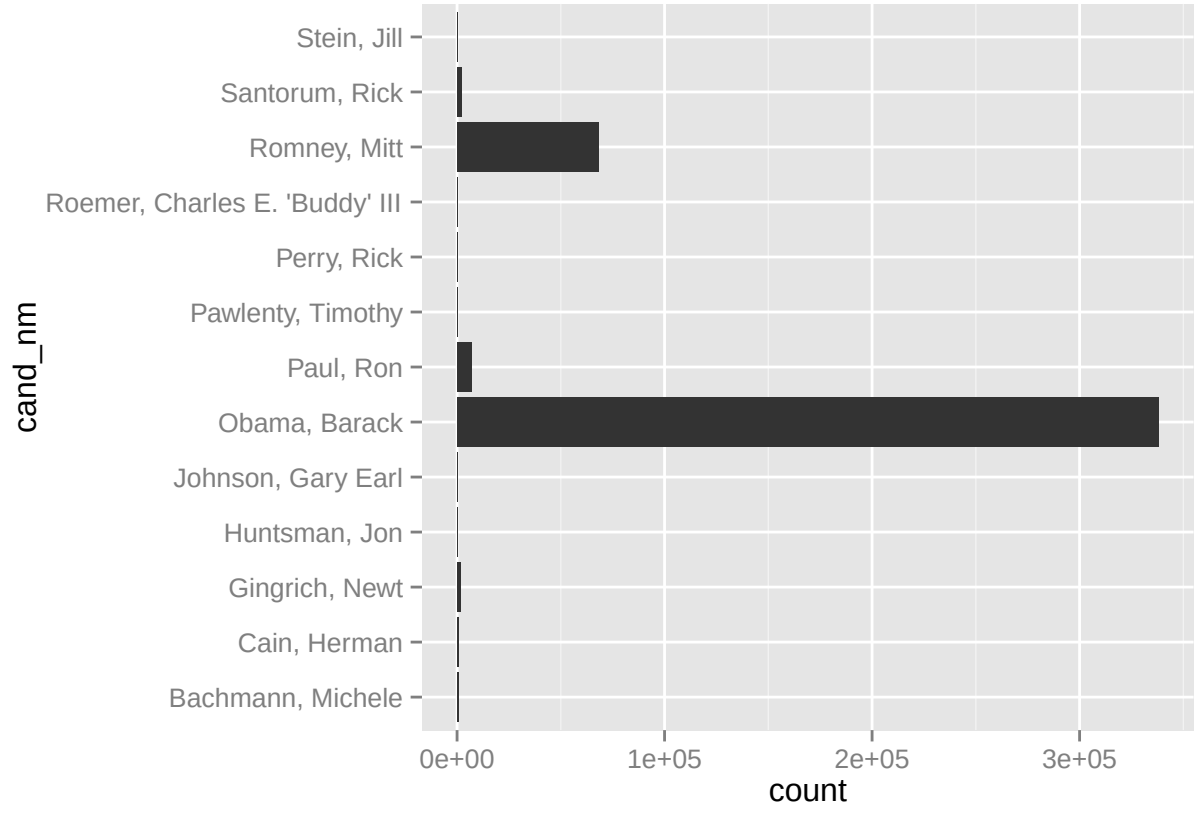


The plot above shows a histogram of the date variable. This provides shows the frequency of the contributions relative to time. As you can see, more contributions occur closer to the actual general election date.

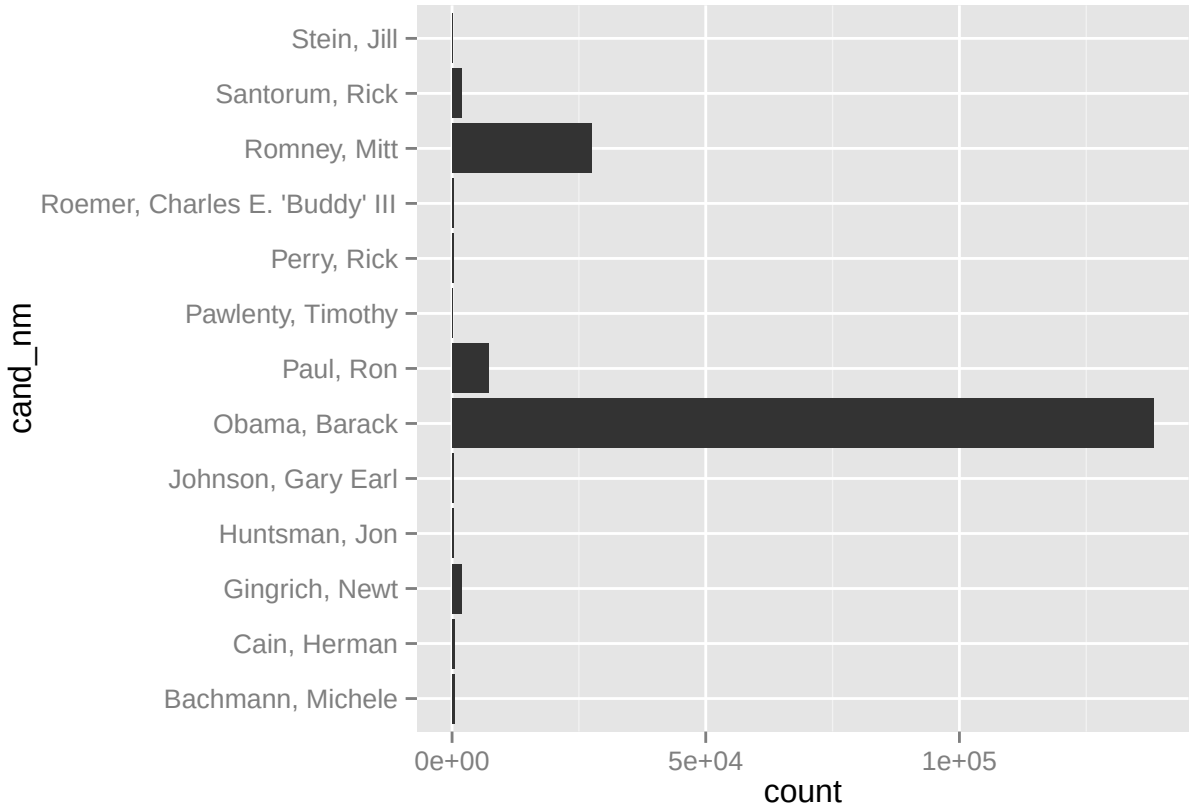
Candidate Names

```
##
##              Bachmann, Michele                     Cain, Herman
##                            542                              527
##                 Gingrich, Newt                   Huntsman, Jon
##                           1932                              386
##             Johnson, Gary Earl           McCotter, Thaddeus G
##                            362                                5
##                  Obama, Barack                      Paul, Ron
##                         338040                             7225
##              Pawlenty, Timothy                    Perry, Rick
##                            176                              312
## Roemer, Charles E. 'Buddy' III                   Romney, Mitt
##                            335                            68405
##                 Santorum, Rick                    Stein, Jill
##                           1970                              142
```

There are 14 total candidates in this dataset. The majority of the contributions are to Barack Obama and Mitt Romney is in second. This makes sense as they were the final presidential candiates for their respective party. I'm going to remove Thaddeus McCotter as he only had 5 contributions. Then create a bar chart to observe contribution count by candidate.

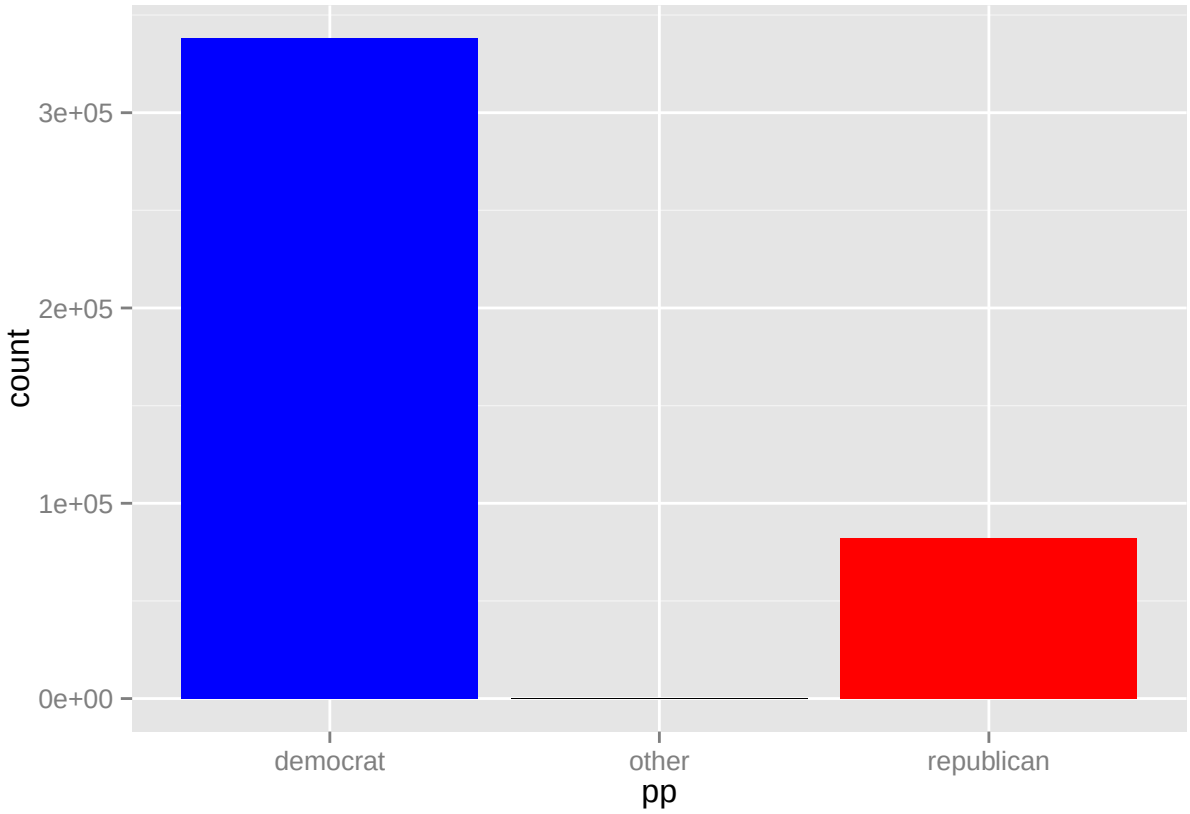


Well, it's not even close. Barack Obama and Mitt Romney received a lot more contributions than any other candidate. This makes sense as Mitt Romney won the primary and Obama was running for re-election as a Democrat. It's hard to discern any differences between the rest except for Ron Paul. Let's look at contributions prior to Mitt Romney winning the Republic Primary.



This plot looks the same but you can atleast see Newt Gingrich and Rick Santorum received more contributions from the others that were not previously mentioned (Obama, Romney, Paul). Time to look at splits by party.

I created the variable `pp` and assigned the appropriate values by candidate. I had to do some digging some of the candidates. Turns out Charles Roemer dropped out of the Republican party and ran under the "Reform Party" and Gary Johnson is a Libertarian. I just assigned these two to "other". Now lets look at a plot.



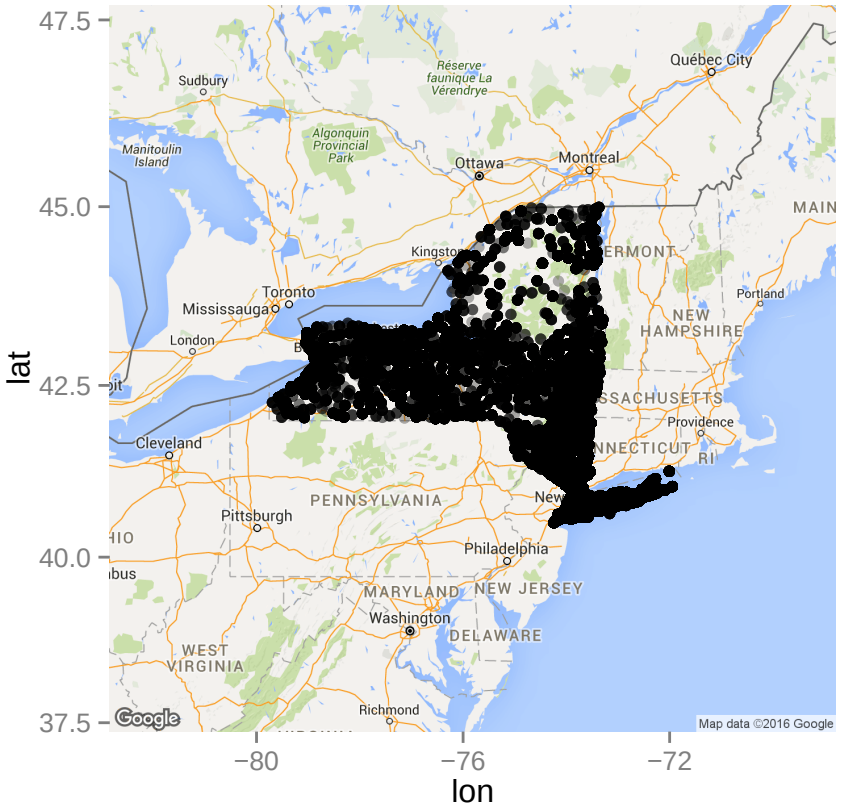
Location variables (`contbr_zip` and `contbr_city`)

First I had to clean up the contbr_zip field as it containd 9 digit zips and some zips less than 5 digits. Additionally, New York States zip code range is between 10001 and 14975. I will just remove these rows. When I looked at city, I noticed there were several variations to how city was entered. Some values contained the state in the city field, some were misspelled, and some were abbreviated. Therefore, I decided zip code is a better location based variable to use when exploring contribution locations as it will be easier to clean.

I know I'm going to want to use ggmap function to make a location based plot so I added in longitude and latitude. The zipcode library is useful to do this. I then joined the zipcode data with the working data.

Now I will use ggmap to create a quick map plot.

```
## Warning: bounding box given to google - spatial extent only approximate.

## converting bounding box to center/zoom specification. (experimental)
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=42.74944,-75.821225&zoom=6&size=
```



Thats a start, I'll examine this further in my final plots. Possibly split out by party (diff color dots). I'd expect more red in usptate NY (Republican) and a lot of blue in NYC area (democrat).

Contribution Amounts

```
##      Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -60800.0     25.0     50.0    225.8    150.0  60800.0
```

```
## [1] 94567785
```

The mean transaction is $225.8 and total money contributed was approximately $94.6M. The min and max transactions are interesting - one for -$60,800 (min) and a transaction of $60,800 (max). This seems there was a mistake or possibly a refund.

```
## [1] 4745
```

```
## [1] ""
## [2] "Refund"
## [3] "REDESIGNATION TO GENERAL"
## [4] "REATTRIBUTION TO SPOUSE"
## [5] "REDESIGNATION TO PRIMARY DEBT"
## [6] "REATTRIBUTION / REDESIGNATION REQUESTED CHECK RETURNED BY BANK"
## [7] "REDESIGNATION TO PRIMARY"
```

```
##
##
##                                                          393
## REATTRIBUTION / REDESIGNATION REQUESTED CHECK RETURNED BY BANK
##                                                            2
##                                       REATTRIBUTION TO SPOUSE
##                                                          132
##                                     REDESIGNATION TO GENERAL
##                                                          398
##                                     REDESIGNATION TO PRIMARY
##                                                            3
##                                  REDESIGNATION TO PRIMARY DEBT
##                                                           65
##                                                       Refund
##                                                         3752
```

```
## [1] -3096930
```

```
##          cmte_id   cand_id       cand_nm       contbr_nm contbr_city
## 42171  C00431445 P80003338 Obama, Barack ELMALEH, VICTOR    NEW YORK
## 168657 C00431445 P80003338 Obama, Barack ELMALEH, VICTOR    NEW YORK
##        contbr_st   zip contbr_employer contbr_occupation contb_receipt_amt
## 42171         NY 10017                                               -60800
## 168657        NY 10017         RETIRED           RETIRED              60800
##        contb_receipt_dt receipt_desc memo_cd memo_text form_tp file_num
## 42171         20-JUL-12       Refund                     SB28A   806136
## 168657        19-JUL-12                                  SA17A   806136
##          tran_id election_tp       date days_from_elec       pp     city
## 42171     D47223       P2012 2012-07-20            109 democrat New York
## 168657 C18100331       P2012 2012-07-19            110 democrat New York
##        state latitude longitude
## 42171     NY 40.75216 -73.97231
## 168657    NY 40.75216 -73.97231
```

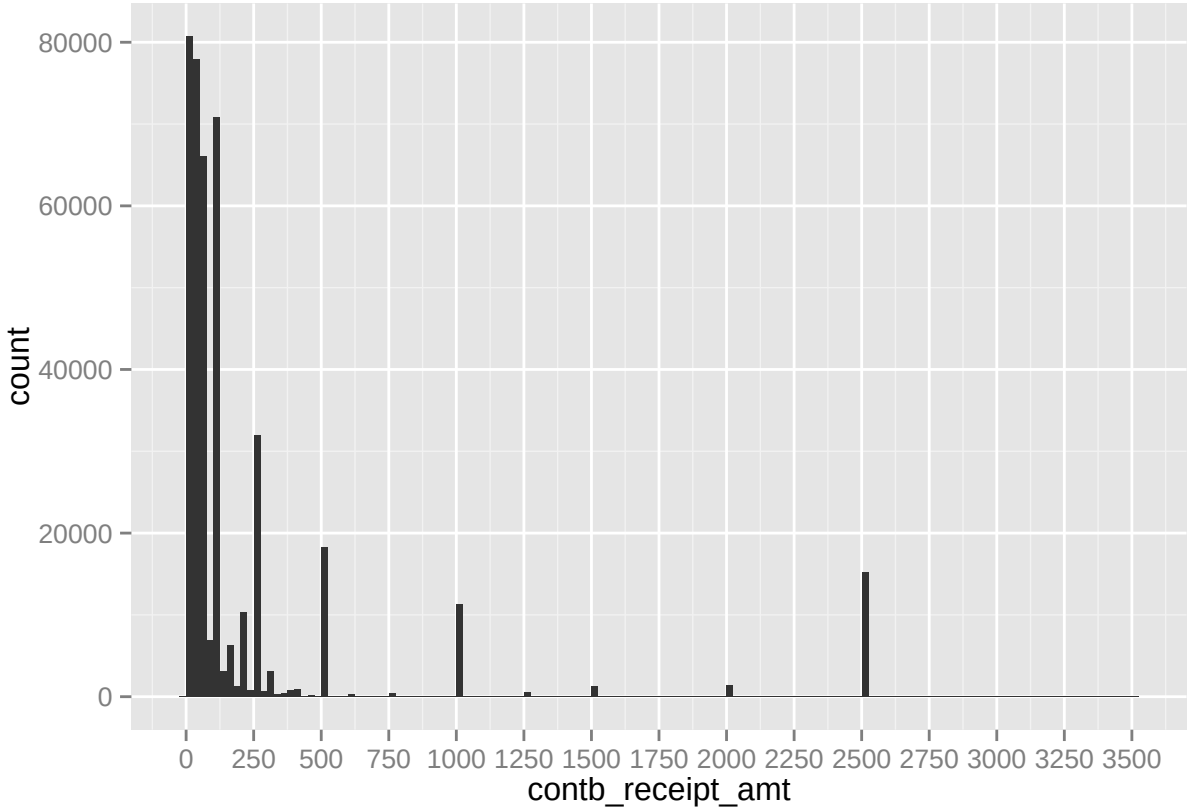
```
## [1] 497
```

```
##
##       (0,5e+03]   (5e+03,1e+04] (1e+04,1.5e+04] (1.5e+04,2e+04]
##          413606               7               0               0
## (2e+04,2.5e+04] (2.5e+04,3e+04] (3e+04,3.5e+04] (3.5e+04,4e+04]
##               0               0               0               1
## (4e+04,4.5e+04] (4.5e+04,5e+04] (5e+04,5.5e+04] (5.5e+04,6e+04]
##               0               0               0               0
## (6e+04,6.5e+04] (6.5e+04,7e+04]
##               1               0
```

```
##
## (-7e+04,-6.5e+04] (-6.5e+04,-6e+04] (-6e+04,-5.5e+04] (-5.5e+04,-5e+04]
##                 0                 1                 0                 0
## (-5e+04,-4.5e+04] (-4.5e+04,-4e+04] (-4e+04,-3.5e+04] (-3.5e+04,-3e+04]
##                 0                 0                 1                 2
## (-3e+04,-2.5e+04] (-2.5e+04,-2e+04] (-2e+04,-1.5e+04] (-1.5e+04,-1e+04]
##                 0                 0                 0                 4
```

```
##   (-1e+04,-5e+03]        (-5e+03,0]
##                23              5211
```

There were 4795 total negative transactions. "Refund" accounted for the majority of the receipt descriptions (3,752 transactions). As for the $60,800, it looks like the same contributor received their money back with a receipt description of "refund". I removed these from the dataset. Also, based on cutting the transaction data, it looks like all but 9 positive transactions fall between 0 and $5000, 497 transactions are for $0, and all but 31 negative transactions fall between $0-5000.



The distribution of the transaction amounts is right skewed. Additionally, most of the transactions are less than or equal to $100. Finally, you can tell $500, $1000, and $2500 were standard dontations. That makes sense as most people wouldn't pick a random number. However, I do find it interesting that $2,500 was really common (compared to $1500 and $2000). Maybe this was a recommended amount by campaign donation soliciters?

Election Type

```
##
##         G2008  G2012  O2012      P  P2008  P2012
##     25     29 222997    348      4     41 195413
```

`election_tp` containd some transactions that are not of interest (i.e. 2008). I removed everything but G2012 (general election contributions) and P2012 (primary contributions).

Univariate Analysis Questions

What is the structure of your dataset?

```
## 'data.frame':    418410 obs. of  25 variables:
##  $ cmte_id          : chr  "C00495820" "C00495820" "C00431445" "C00431445" ...
##  $ cand_id          : chr  "P80000748" "P80000748" "P80003338" "P80003338" ...
##  $ cand_nm          : chr  "Paul, Ron" "Paul, Ron" "Obama, Barack" "Obama, Barack" ...
##  $ contbr_nm        : chr  "NOBIS, JOSEPH FRANK JR." "MCMONAGLE, SEAN" "MASON, TIMOTHY" "WHITE, SARA
##  $ contbr_city      : chr  "TABERG" "EAST AMHERST" "NEW YORK" "NEW YORK" ...
##  $ contbr_st        : chr  "NY" "NY" "NY" "NY" ...
##  $ zip              : chr  "13471" "14051" "10009" "10025" ...
##  $ contbr_employer  : chr  "SELF" "BLT CHEMBULK" "SELF-EMPLOYED" "RETIRED" ...
##  $ contbr_occupation: chr  "TRUCK DRIVER" "INTERN" "WRITER" "RETIRED" ...
##  $ contb_receipt_amt: num  1000 100 50 100 300 100 50 100 50 100 ...
##  $ contb_receipt_dt : chr  "06-DEC-11" "06-DEC-11" "06-SEP-11" "27-SEP-11" ...
##  $ receipt_desc     : chr  "" "" "" "" ...
##  $ memo_cd          : chr  "" "" "" "" ...
##  $ memo_text        : chr  "" "" "" "" ...
##  $ form_tp          : chr  "SA17A" "SA17A" "SA17A" "SA17A" ...
##  $ file_num         : int  779227 779227 756218 756218 756218 756218 756218 756218 756218 756218 ...
##  $ tran_id          : chr  "0925649" "0925702" "C12008058" "C12249568" ...
##  $ election_tp      : chr  "P2012" "P2012" "P2012" "P2012" ...
##  $ date             : Date, format: "2011-12-06" "2011-12-06" ...
##  $ days_from_elec   : int  336 336 427 406 435 419 460 407 404 426 ...
##  $ pp               : chr  "republican" "republican" "democrat" "democrat" ...
##  $ city             : chr  "Taberg" "East Amherst" "New York" "New York" ...
##  $ state            : chr  "NY" "NY" "NY" "NY" ...
##  $ latitude         : num  43.3 43 40.7 40.8 40.7 ...
##  $ longitude        : num  -75.6 -78.7 -74 -74 -74 ...
```

The dataset after I cleaned it up and added some new columns consisted of 418,410 observations and 25 variables.

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

I believe the main features of interest are `cand_nm` and the `contb_receipt_amt`.

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

`pp` (political party), `latitude`, `longitude`, `contbr_occupation`, `election_tp`, `date`, `days_from_elec`

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

I added `pp` (political party), `latitude`, `longitude`, `days_from_elec`.

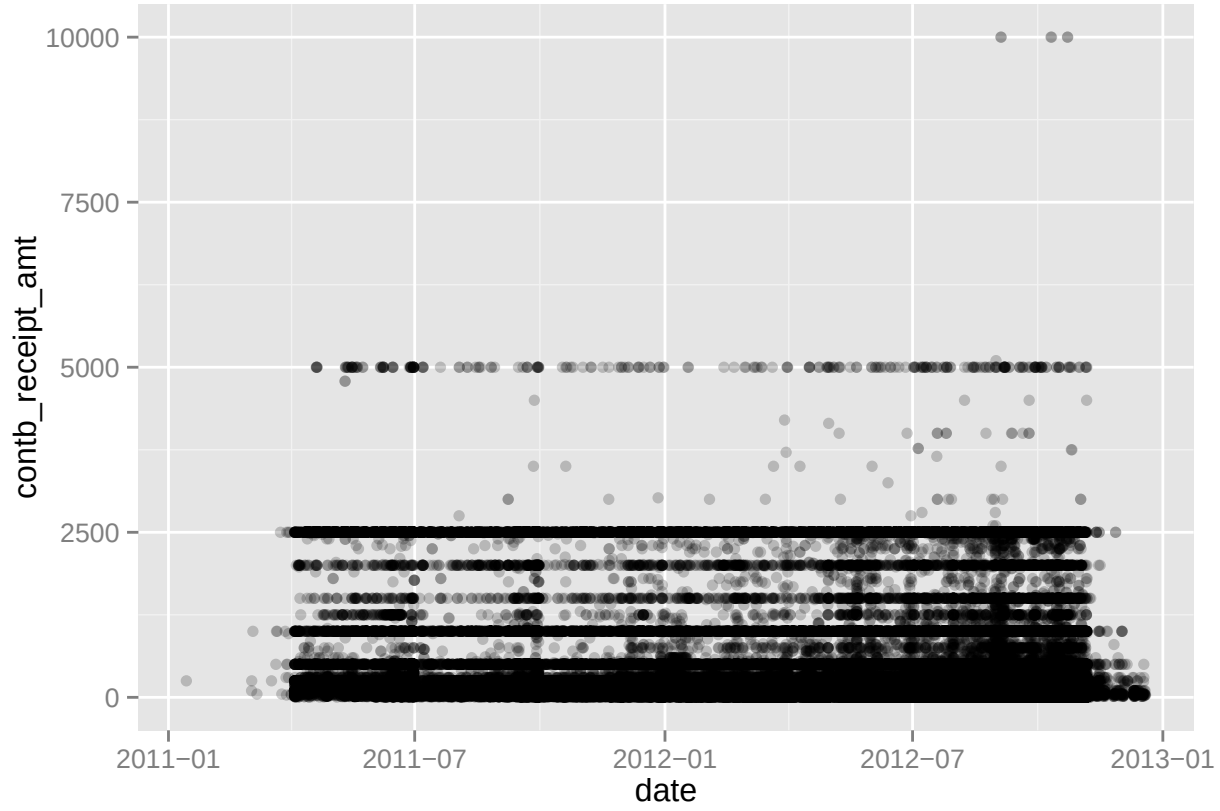
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

There weren't any distrubitions that were that surprising or what I would consider unusual. I cleaned the `contbr_zip` field to contain only 5 digit zips and joined with the zipcode data filtered by NY state. This created the latitude and longitude fields which will be useful for mapping. I also cleaned out `election_tp` to show transactions only for general and primary elections. This helped as it removed unwanted data and will allow for easier grouping (only 2 unique values).

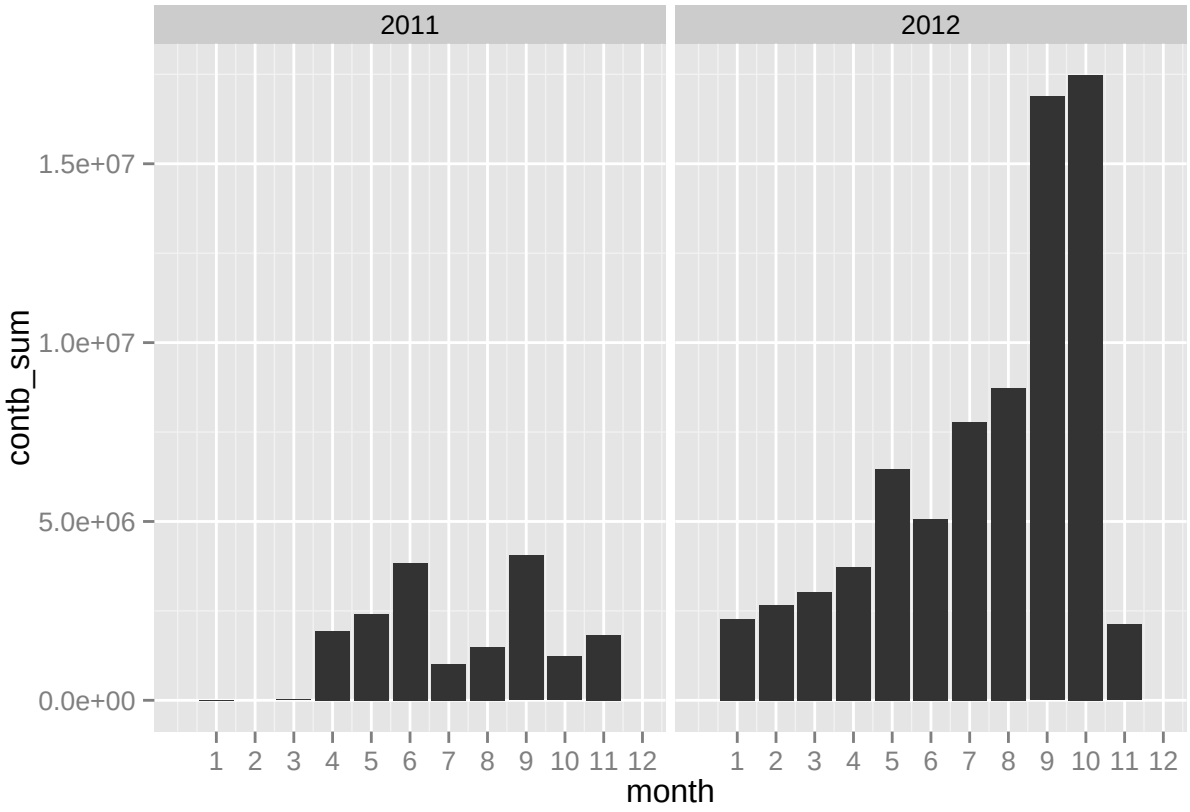
Bivariate Exploration

I want to look at contributions over time. I already looked at the count of contributions but now I want to look at the amount of money over time.

Time Series of Contributions Amounts

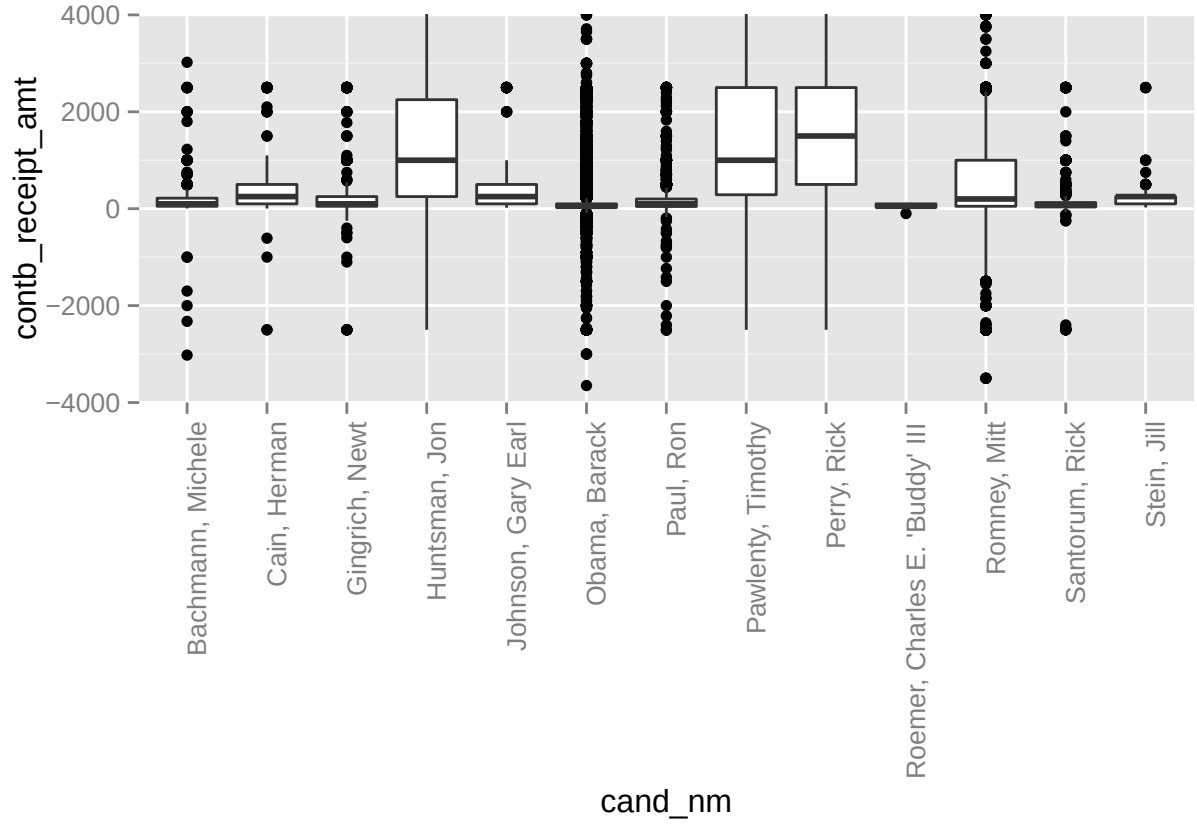


I filtered the data to remove negative transactions and any transactions over $10k. This plot is shows the different levels of contributions but it's really hard to tell the amount of money in a certain discrete time period. If you look close, you do see an increase in the contibutions in late 2012. Let's look by month for each year.



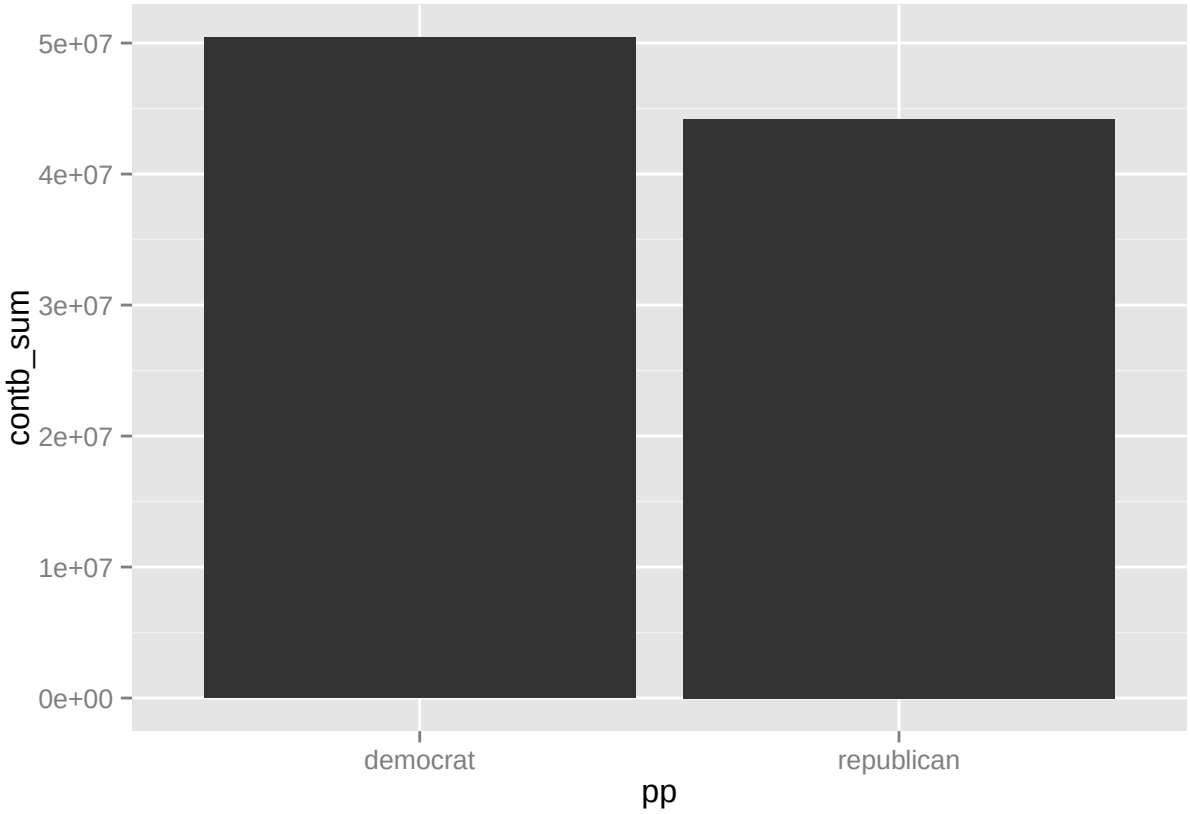
As expected, there is an increasing trend in contributions a significant amount contriubted in Sept/Oct 2012.

Contributions Amonts Received by Candidate



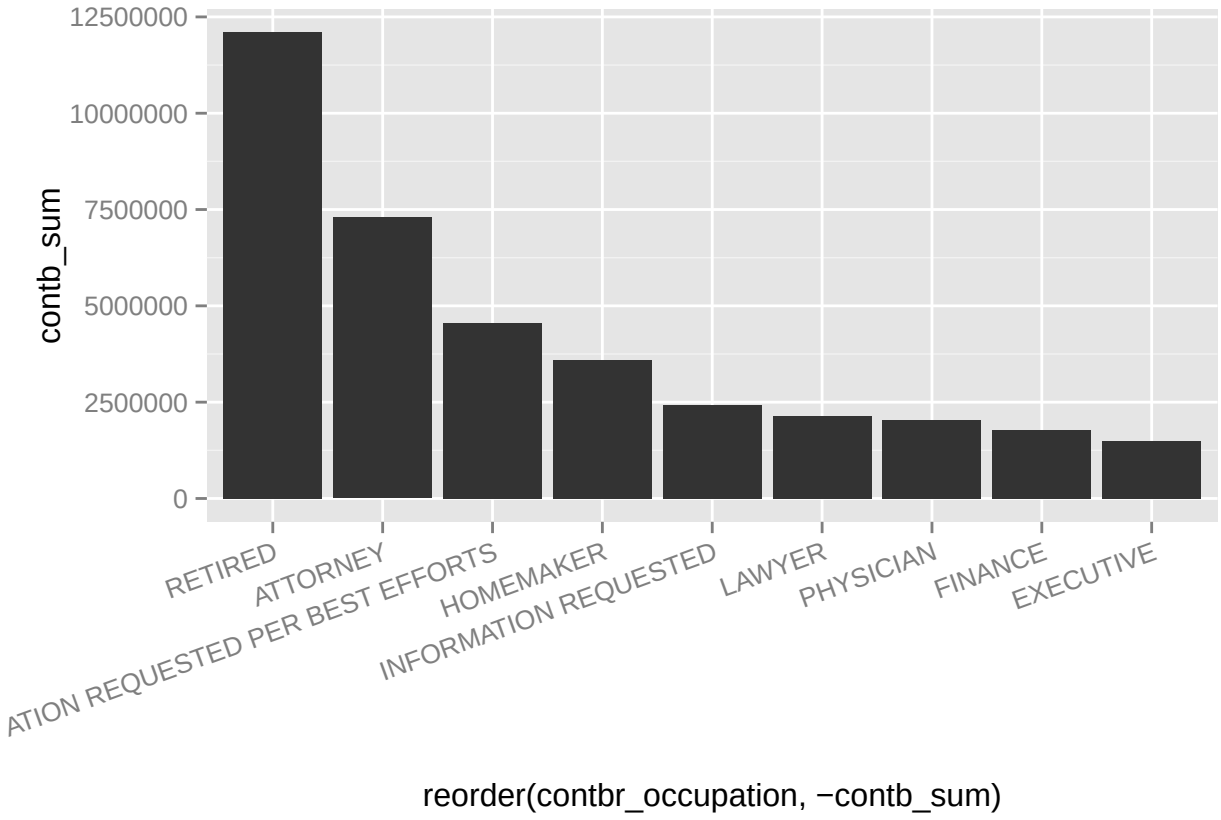
What I find most interesting here how compact the contributions at low dollar amounts were for Barack Obama even tho he had highest contbribution total. You can really see this in the boxplot as his box portion is essentially centered what looks like around $0-$50. Mit Romney on the other has a box longer dispersed box indicating there was a little more variance in his contributions versus Obama.

Contributions by party



More money was donated to Obama (Democrat) compared to total Republican donations.

Contributions by Occupation



Well, retired people are donating the most money. That is interesting. Glad my social security contributions is being put to something other than those one arm bandits (slot machines)!

Conbributions by city

```
## Source: local data frame [10 x 4]
##
##          city  contb_amt  count contb_amt_per_trans
##         (chr)      (dbl)  (int)               (dbl)
## 1    New York 47054168.9 140810            334.1678
## 2    Brooklyn  6514077.2  48709            133.7346
## 3   Scarsdale  1547555.9   3082            502.1272
## 4         Rye  1183237.7   1833            645.5197
## 5     Buffalo  1097558.4   5679            193.2661
## 6       Bronx  1042731.4   9841            105.9579
## 7  Bronxville   938321.1   1978            474.3787
## 8   Rochester   877127.6   6309            139.0280
## 9   Larchmont   775525.6   2350            330.0109
## 10 Great Neck   590778.1   1646            358.9174
```

No surpirse that NYC toppped the list on the contb amount. However, they did not top the list on a per transaction basis. Rye holds that distinction for the cities in the top 10. I know the area quite well and Rye is certainly a wealthy area so this makes sense.

Bivariate Analysis Questions

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?

Contributions over time increased and substantially increased in the final months prior to the general election. Not a huge surprise.

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

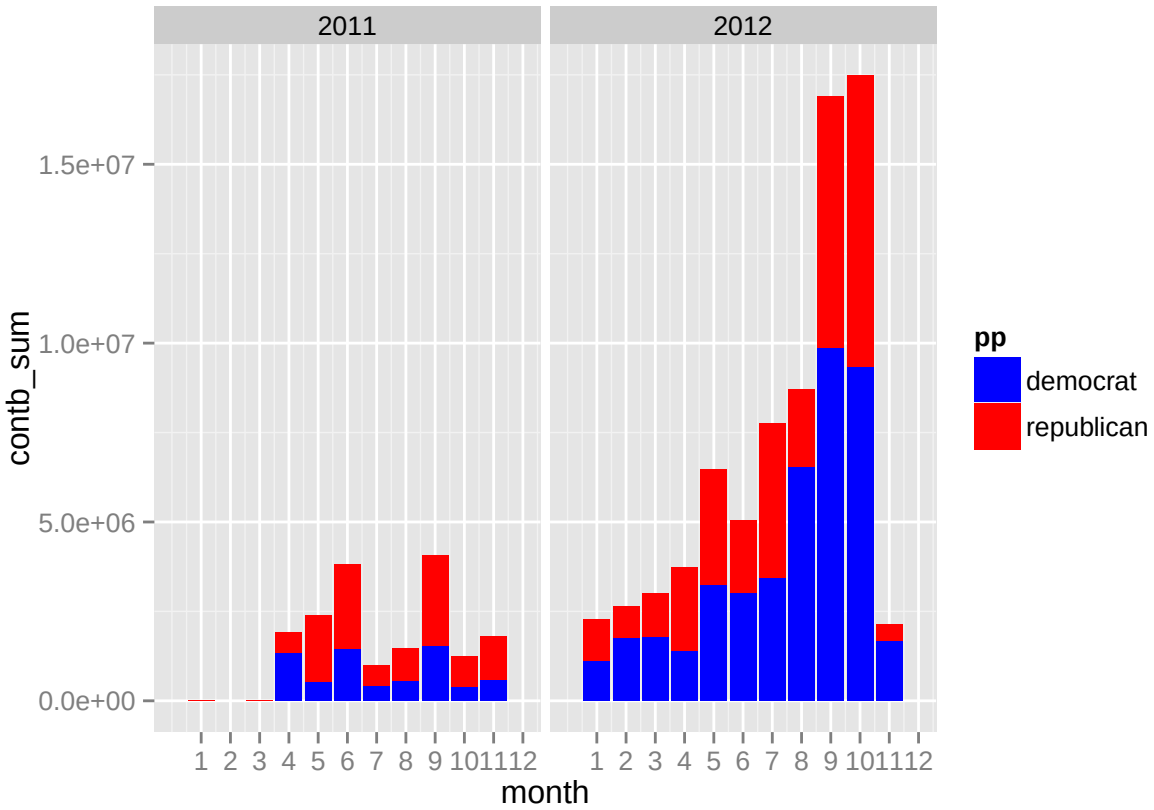
Contributions vary more in size with the Republican candidates. Obama on the other hand, generated most of his contributions via low dollar amounts. Additionally, people that are retired donated the most money.

What was the strongest relationship you found?

The most relevant and strongest relationship is the increase in contributions over time. Again, this is no surprise.

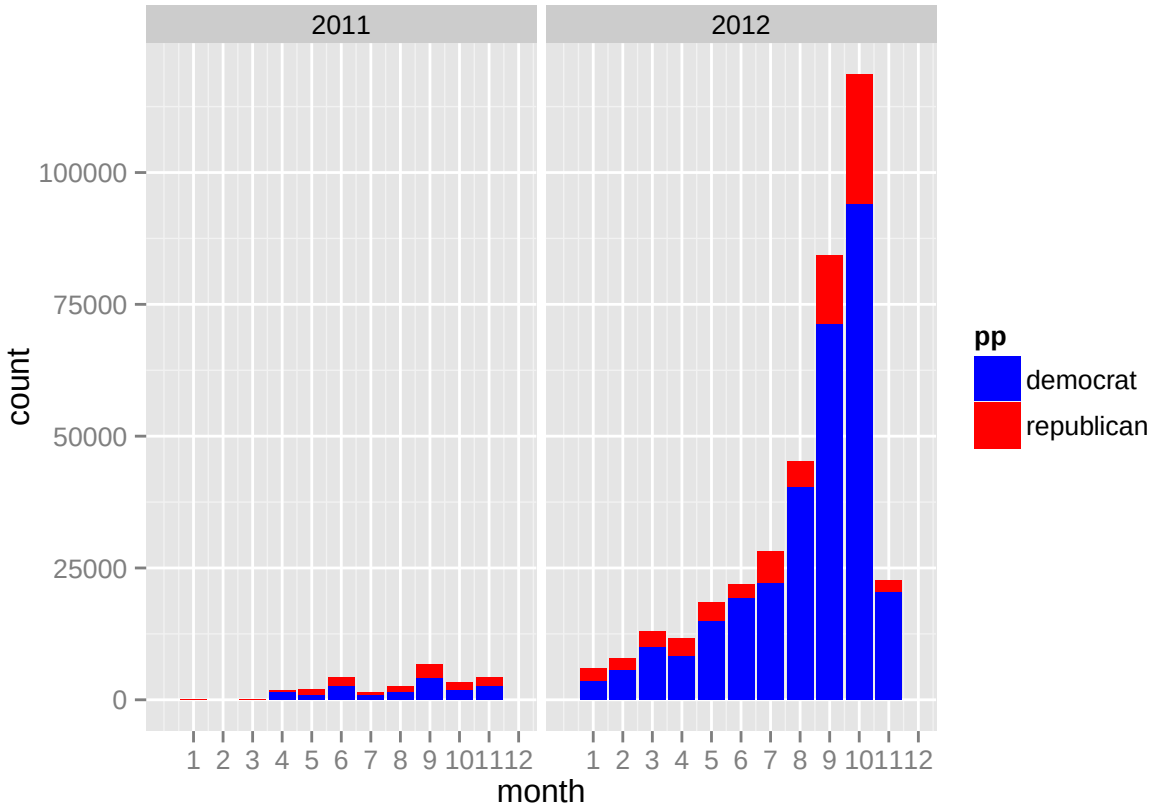
Multivariate Exploration

Contributions Amounts over time by party



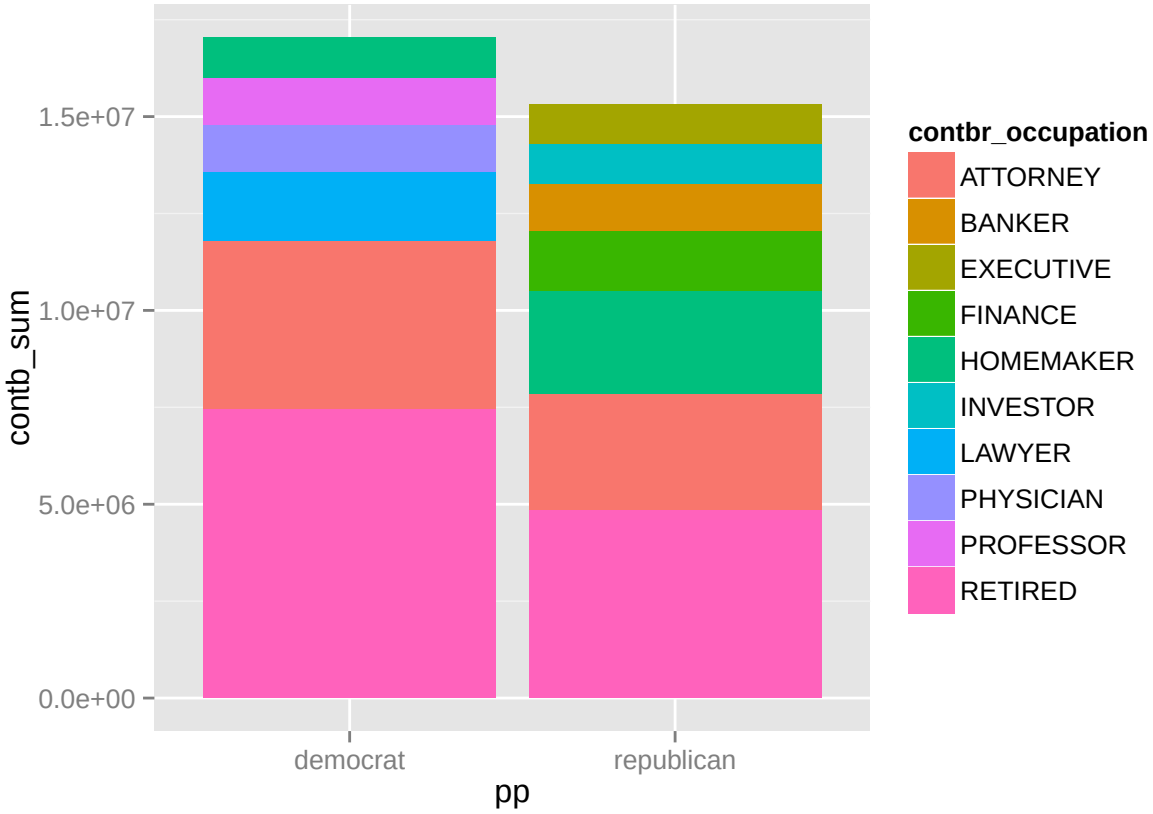
This splits out the contribution amount over time by party. Interesting, however lets look at the same thing except for number of contributions.

Contribution number over time by party



Wow. This shows that Obama (the only Democrat) really dominated the amount of contributions. Even while the Republic Primary was going on.

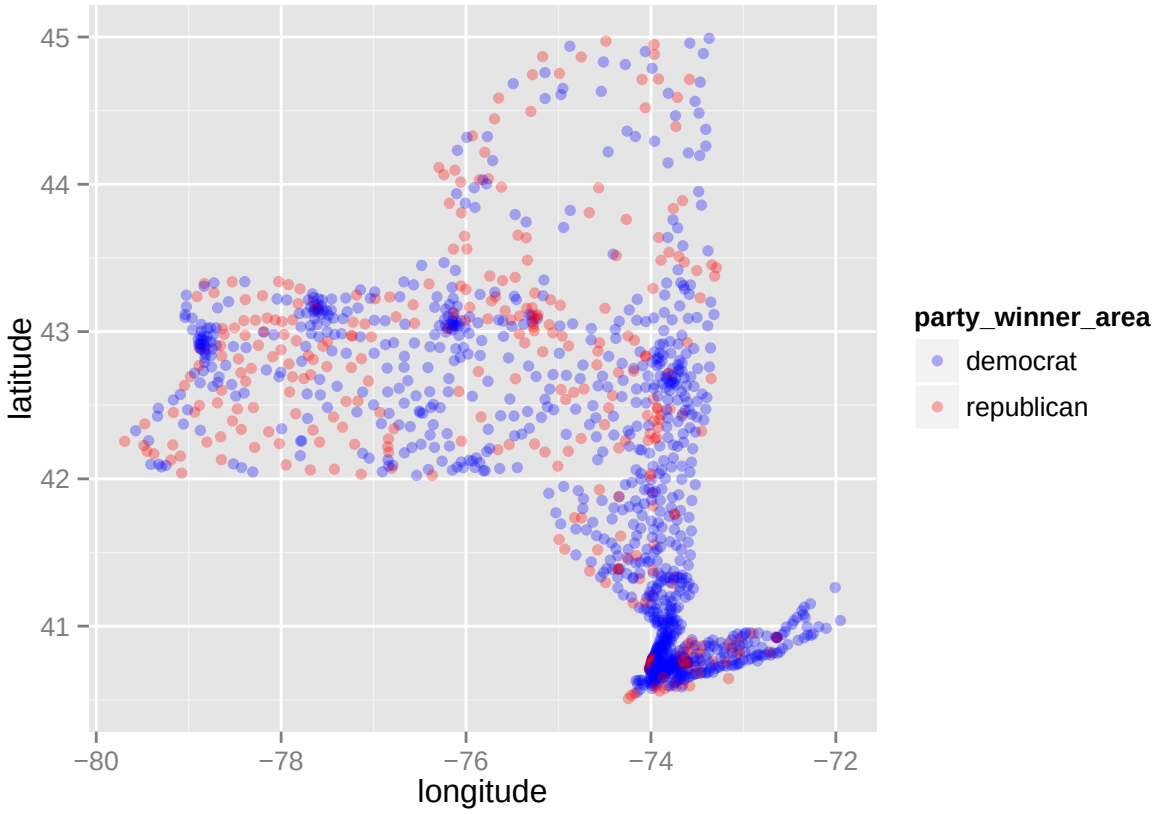
Contribution by Occupation to which Party



I filtered by occupation and party and then filtered the sum conbributions. This gives you the general idea of who donated to which party. Republicans both share the top 2 with retired people and attorneys. Republicans receive more their money from bankers, finance, and exectuives while Democrats receive more of their money from Investors, Physicians, and Professors.

Contribution by area by Party

This shows data split out by party by zipcode. One of my final plots will show the contribution amounts on a map by party. For now, this just shows an outline of NY. You can see clustering around the major cities. If you know NY, you'll know which city is which without the map.



Multivariate Analysis Questions

I decided to look at the comparison over contributions by party. I did this mainly because I know most of the Republican contributions went to Romney while all the Democratic contributions went to Obama. It was easier to explore the data using a binary variable instead of splitting out by candidate. I will split out by candidate in my final plot though (only the top 5 though).

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?

I observed the amount of contributions over time compared to the conbtribution amounts over time by party. You can see that the conbribution amounts look relatively even (favoring Obama slightly) but the total contributions heavily favored Obama. This isn't surprising to me given the demographics of the different parties and the fact that NY is always a blue state.

Were there any interesting or surprising interactions between features?

I wasn't really surprised by much of the analysis. NY state is really predictable. It's a blue state and the data has really shown that. The biggest surprise I see is the clustering of red in the area I'm from (Utica, NY). I never would of guessed that.

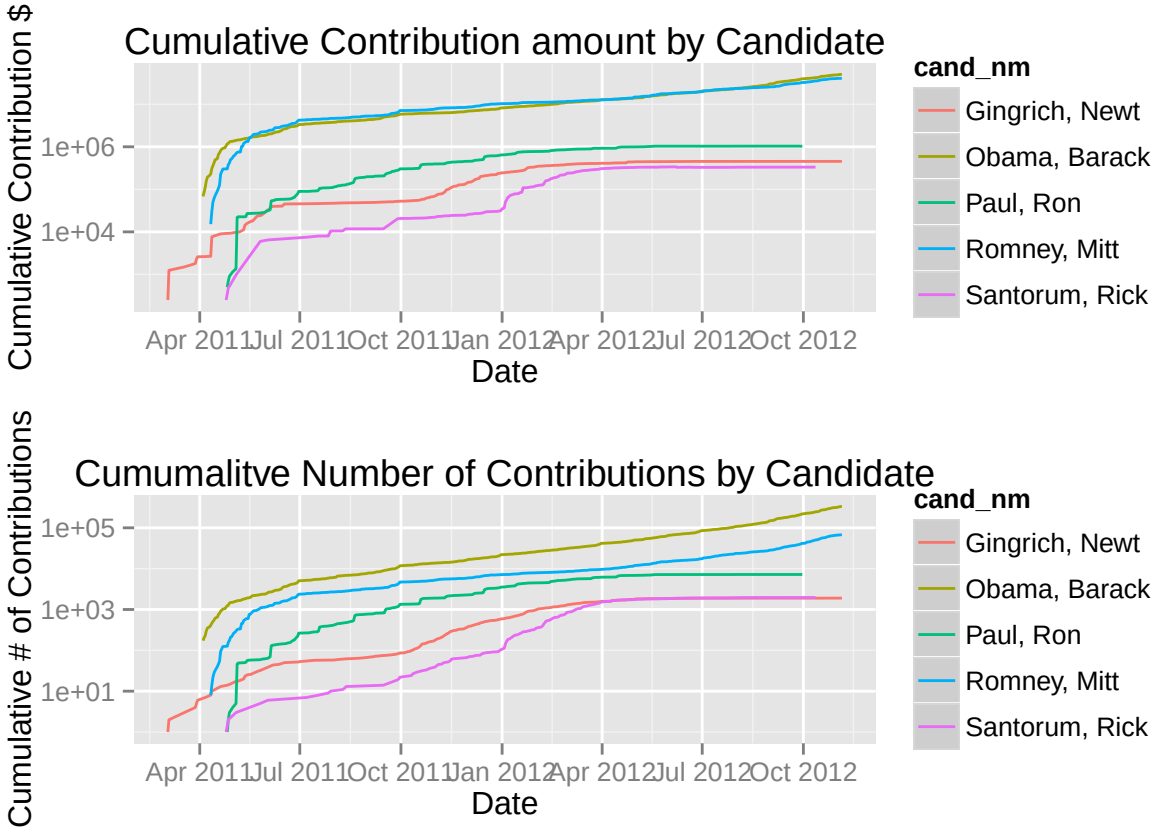
Final Plots and Summary

I created three plots that I thought summarized the most interesting things I found while exploring the data. * Plot One. Contributions over time * Plot Two. Contibution winners by area (zipcode) * Plot Three. Contributions by party by occupation

Plot One - Contributions over time

The main thing I'm interested in is contributions over time. As we were able to learn in data exploration phase, Obama received a lot of lower dollar contributions while the Republicans - mainly Romney at the end - received the higher dollar lower volume contributions. The following plot is my attempt to display this information. I only included the Republicans that were in it to the end of the Republican primary to keep things clean.

You really can see a lot here. Obama really dominated in the number of contributions but Romney kept it pretty close in the total amount of contributions. Pretty cool. Note that the Y values are in log_scale so even though Romney looks really close to Obama at the end in terms of contribution amount, he still received $10M less than Obama in NY. The point of this plot is to show the disparity in contbitution amount versus number of contributions. Also of note, you can see the rise of Rick Santorum in early 2012 when he was trailing in the primaries. I remember he kind of came out of no where to be a real contender in the Republican primary and this aligns with my memory!



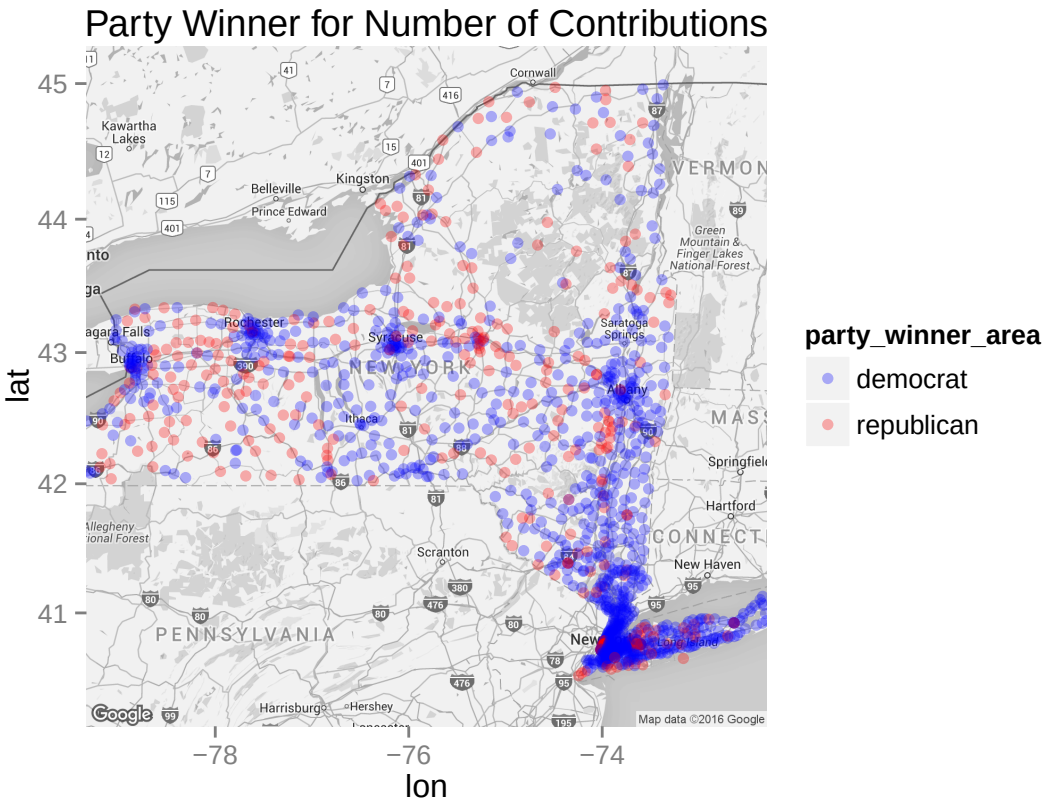
Plot Two - Contributions by area (zipcode)

Time for the Map. I created this in the multivariate exploration section but now I'm adding the overlay of a map extracted using the get_map function and then ggmap. As you can see, there are clusters of blue in the

cities with red sprinkled throughout the state. Again this is no surprise.

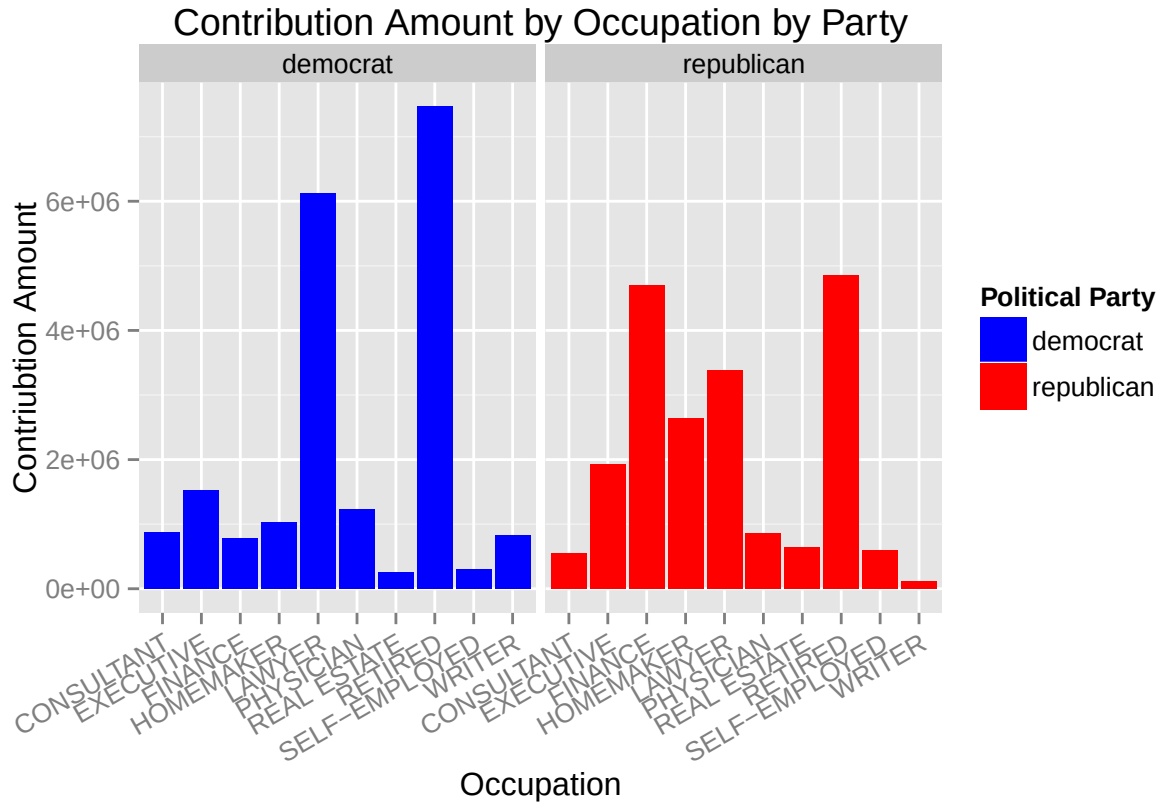
```
## Warning: bounding box given to google - spatial extent only approximate.

## converting bounding box to center/zoom specification. (experimental)
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=42.74944,-75.821225&zoom=7&size=
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```



Plot Three - Contributions by Occupation by Party

I think exploring contributions based on the occupation is interesting. That's why I chose it for my last plot. I decided to seperate out the top occupation contributors as there were way too many unique values due to the varying ways people recorded their occupation. However, this captures the essense of the most general occupations such as exectuive, finance, physician, and retired. Overall, you can see that Retired people donated more to the democrats while people in Finance contributed more to Republicans. What was surprising was the difference in the lawyer contributions as Obama received significantly more from that category.



Reflection

To explore the data, I first examined the structure of the dataset. I quickly learned that most of my analysis would revolve around the contribution amounts and number of contributions as there wasn't a lot of continuous data. I also saw there was a date variable which I could convert to Date class and then perform some analysis over time. I examined each univariate that I thought was useful and cleaned or created a few variables such as political party, month, year, latitude, longitude, and days from election.

I used these newly crafted and cleaned variables to examine the relationships of the data. I learned that contributions over time increase and that a significant amount of contributions are provided in the last two months prior ot the general election. I found that retired population contributes a singificant amount and confirmed that NY State is infact a "blue" state (for 2012) by mapping party winners by zipcode. Finally and most importantly, I gleaned that there was a significant difference in the amount of contributions between Obama and the Republicans. People who donated to Republicans (mainly Romney) tended to donate larger amounts while people who donated to Obama tended to donate smaller amounts.

If I were to model the data, I'd imagine the pure volume of contributions would be a good indicator for who will win the state in the general election. I think it would be a good exercise cycle through dataset for each state and look at which candidate received the most amount of contributions. I would be interested to see if that simple heuristic was an accurate representation of who actually won the state in the 2012 election.

References

- http://fec.gov/disclosurep/PDownload.do

- http://kbroman.org/knitr_knutshell/pages/Rmarkdown.html
- http://www.inside-r.org/packages/cran/ggmap/docs/get_map
- https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf