P4-Explore and Summarize Data with R Final Project John K. Hancock

John K. Hancock – jkhancock@gmail.com (mailto:jkhancock@gmail.com)

January 16, 2017 (Originally submitted December 26,2016)

2012 NY Presidential Contribution Data Analysis

Overview

This report provides a graphical analysis of donations made to U.S. presidential candidates in 2012 by NY States residents of the state of NY. The report is organized as follows

Introduction to the Data Set - An overview of the Federal Elections Commission dataset for NY state donors to the 2012 Presidential Candidates

Part One - A Univariate Analysis of the Contribution Amount - NY State Donors

Part Two - Bivariate Analysis of The Four Presidential Nominees

Part Three - Multivariate Analysis of the Zip Codes in the dataset

Part Four - Predicting contribution amounts based on number of donors

Final Summary of Findings

Set up the Global Environment for Analysis

Ensure that the environment is looking in the correct directory.

Set up the libraries to be used in this report

Introduction to the Dataset

```
## 'data.frame':    420359 obs. of  19 variables:
##  $ ID               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ cmte_id          : Factor w/ 15 levels "C00410118","C00431171",..: 7 7 3 3 3 3 3 3 3 3 ...
##  $ cand_id          : Factor w/ 14 levels "P00003608","P20002523",..: 12 12 13 13 13 13 13 13
13 13 ...
##  $ cand_nm          : Factor w/ 14 levels "Bachmann, Michele",..: 8 8 7 7 7 7 7 7 7 7 ...
##  $ contbr_nm        : Factor w/ 97577 levels "5146 HIGHBRIDGE ST. L.L.C.",..: 64162 58390 563
63 94096 48193 49147 87432 78127 51203 25945 ...
##  $ contbr_city      : Factor w/ 2206 levels "","  NEW YORK",..: 1964 565 1344 1344 1344 2041 1
344 1344 1484 174 ...
##  $ contbr_st        : Factor w/ 1 level "NY": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contbr_zip       : Factor w/ 50520 levels "","*1046","0",..: 44505 45600 1608 8309 3302 48
338 7480 9748 21633 45343 ...
##  $ contbr_employer  : Factor w/ 36541 levels "","-","--","---",..: 28679 3808 28692 26951 286
92 10712 6248 31618 28692 26951 ...
##  $ contbr_occupation: Factor w/ 16882 levels "","$200 ","(NONUNION) F/T K.B.S.I",..: 15699 78
11 16663 12948 825 1321 668 12948 10983 12948 ...
##  $ contb_receipt_amt: num  1000 100 50 100 300 100 50 100 50 100 ...
##  $ contb_receipt_dt : Factor w/ 644 levels "1-Apr-11","1-Apr-12",..: 565 565 580 419 445 125
520 398 462 602 ...
##  $ receipt_desc     : Factor w/ 17 levels "","REATTRIBUTED",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_cd          : Factor w/ 2 levels "","X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_text        : Factor w/ 152 levels "","$1000 REFUND TO BE ISSUED",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ form_tp          : Factor w/ 3 levels "SA17A","SA18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ file_num         : int  779227 779227 756218 756218 756218 756218 756218 756218 756218 756
218 ...
##  $ tran_id          : Factor w/ 391954 levels "0704020-0001",..: 6679 6680 16441 18254 16053
16993 15242 18068 18761 16457 ...
##  $ election_tp      : Factor w/ 7 levels "","G2008","G2012",..: 7 7 7 7 7 7 7 7 7 7 ...
```

Summary of the Dataset

The data consists of 420,359 observations across 19 variables. The cmte_id variable is for each candidate's presidential campaign committee. These are set up so that the candidates can receive campaign contributions, and it shows that it has 14 levels each representing a candidate. Other variables of interest include, "cand_nm", which is the name of the candidate, "contbr_city", the contributor's city, and "contbr_zip", the contributor's zip code.

What the dataset does not contain is more detailed demographic data about the contributors, e.g. gender, age, race, educational levels, income levels, etc. This means that the analysis in this report will be very shallow. The report won't be able to detail what kind of person donates to presidential campaigns. However, the report will show the locations of the donors and how the number of donations play a part in the total donations.

Part One - A Univariate Analysis

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -60800.0     25.0     50.0    225.9    150.0  60800.0
```

Looking at a summary of the contribution amounts reveal something unusual. It shows the min and max values of -$60,800 and $60,800. Looking into the file we see that the donation was reversed. According to the Federal Elections Committee website, presidential campaign donations are limited to $2700 per individual per election. Primaries and the General election account for two unique elections. For the purpose of this analysis I will create a new dataframe and limit the contb_receipt_amt to $2,700.00 or less. We do not want to include contributions which exceeded the legal limit and was re-designated for other purposes.

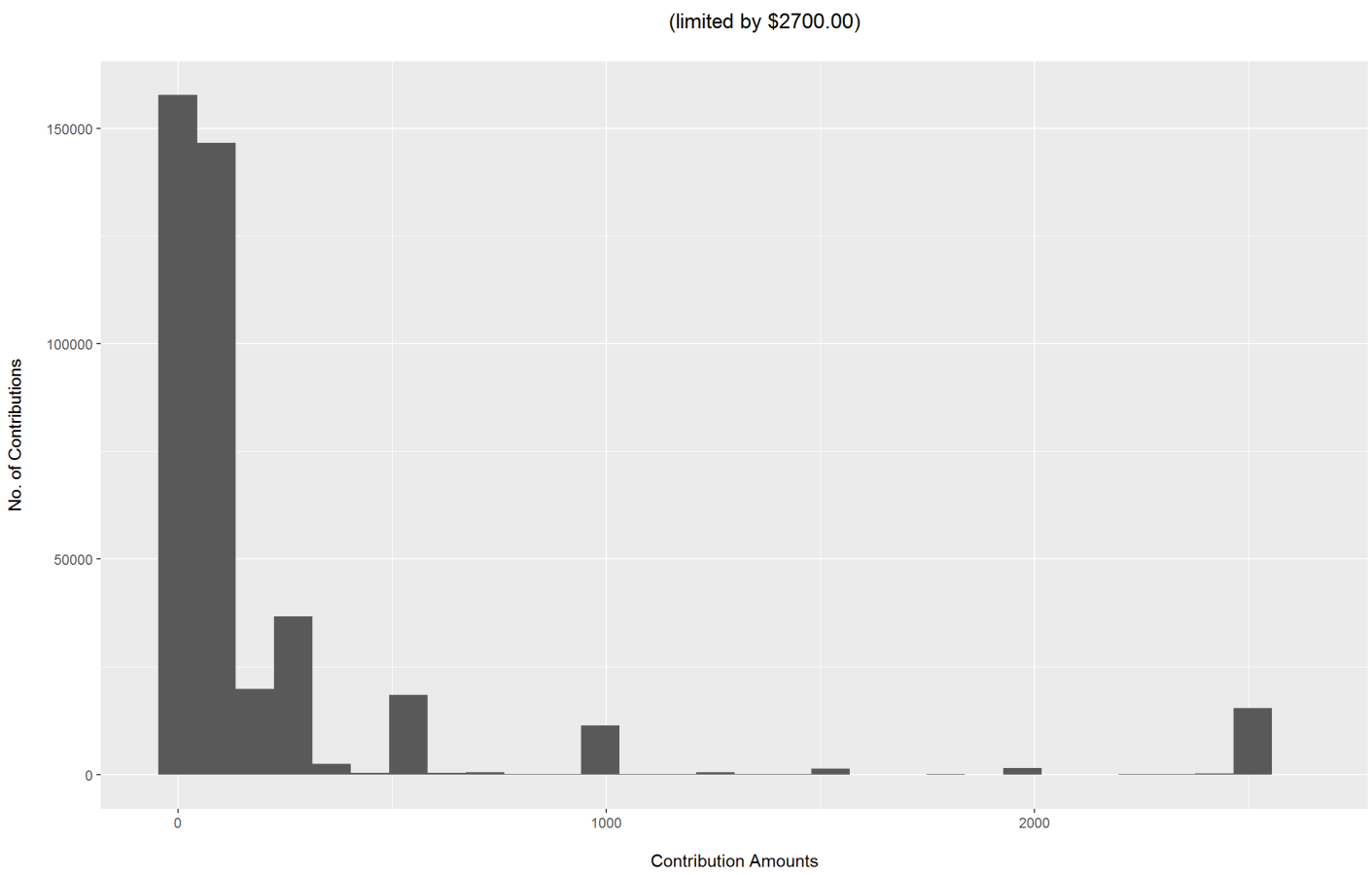
Restricting the analysis to legal campaign contribution amounts show that the number of observations drop to 414,741

```
## 'data.frame':    414741 obs. of  19 variables:
##  $ ID               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ cmte_id          : Factor w/ 15 levels "C00410118","C00431171",..: 7 7 3 3 3 3 3 3 3 3 ...
##  $ cand_id          : Factor w/ 14 levels "P00003608","P20002523",..: 12 12 13 13 13 13 13 13
13 13 ...
##  $ cand_nm          : Factor w/ 14 levels "Bachmann, Michele",..: 8 8 7 7 7 7 7 7 7 7 ...
##  $ contbr_nm        : Factor w/ 97577 levels "5146 HIGHBRIDGE ST. L.L.C.",..: 64162 58390 563
63 94096 48193 49147 87432 78127 51203 25945 ...
##  $ contbr_city      : Factor w/ 2206 levels "","  NEW YORK",..: 1964 565 1344 1344 1344 2041 1
344 1344 1484 174 ...
##  $ contbr_st        : Factor w/ 1 level "NY": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contbr_zip       : Factor w/ 50520 levels "","*1046","0",..: 44505 45600 1608 8309 3302 48
338 7480 9748 21633 45343 ...
##  $ contbr_employer  : Factor w/ 36541 levels "","-","--","---",..: 28679 3808 28692 26951 286
92 10712 6248 31618 28692 26951 ...
##  $ contbr_occupation: Factor w/ 16882 levels "","$200 ","(NONUNION) F/T K.B.S.I",..: 15699 78
11 16663 12948 825 1321 668 12948 10983 12948 ...
##  $ contb_receipt_amt: num  1000 100 50 100 300 100 50 100 50 100 ...
##  $ contb_receipt_dt : Factor w/ 644 levels "1-Apr-11","1-Apr-12",..: 565 565 580 419 445 125
520 398 462 602 ...
##  $ receipt_desc     : Factor w/ 17 levels "","REATTRIBUTED",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_cd          : Factor w/ 2 levels "","X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_text        : Factor w/ 152 levels "","$1000 REFUND TO BE ISSUED",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ form_tp          : Factor w/ 3 levels "SA17A","SA18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ file_num         : int  779227 779227 756218 756218 756218 756218 756218 756218 756218 756
218 ...
##  $ tran_id          : Factor w/ 391954 levels "0704020-0001",..: 6679 6680 16441 18254 16053
16993 15242 18068 18761 16457 ...
##  $ election_tp      : Factor w/ 7 levels "","G2008","G2012",..: 7 7 7 7 7 7 7 7 7 7 ...
```

A new summary of the contribution amounts show that the average contribution was $232.50, and the median contribution was $55.00.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.45   25.00   55.00  232.50  150.00 2600.00
```

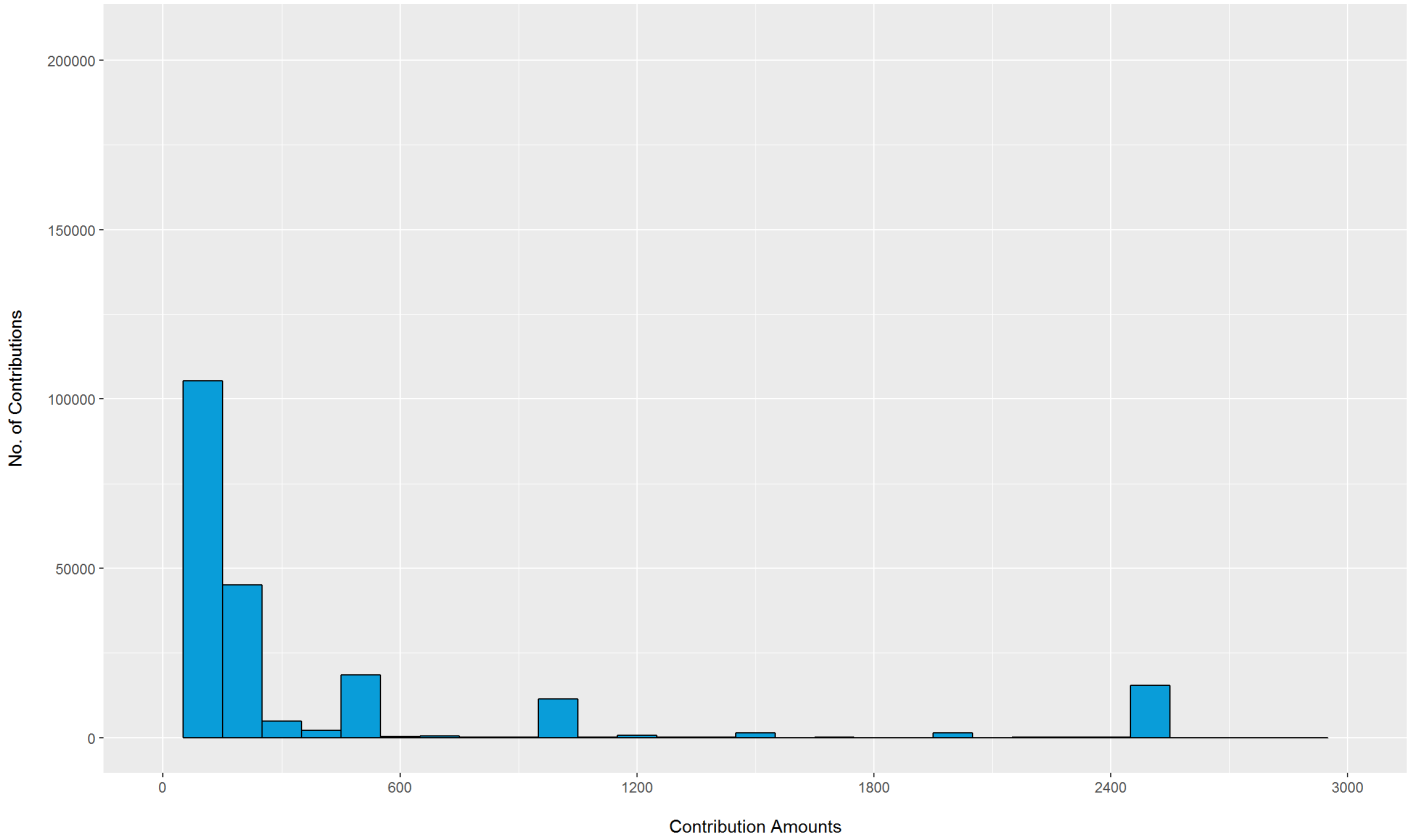
Initial Histogram plot which shows the amounts contributed limited by the legal maximum donation, $2,700.00.



After adjusting the binwidth and the x-axis, I can see that the contribution amounts were skewed with the majority of the contributions being $100 or less. Also note that this histogram shows a number of outlier donations right at the legal limit.

Part One: 2012 Total Contributions to All Presidential Candidates

(limited by $2700.00)



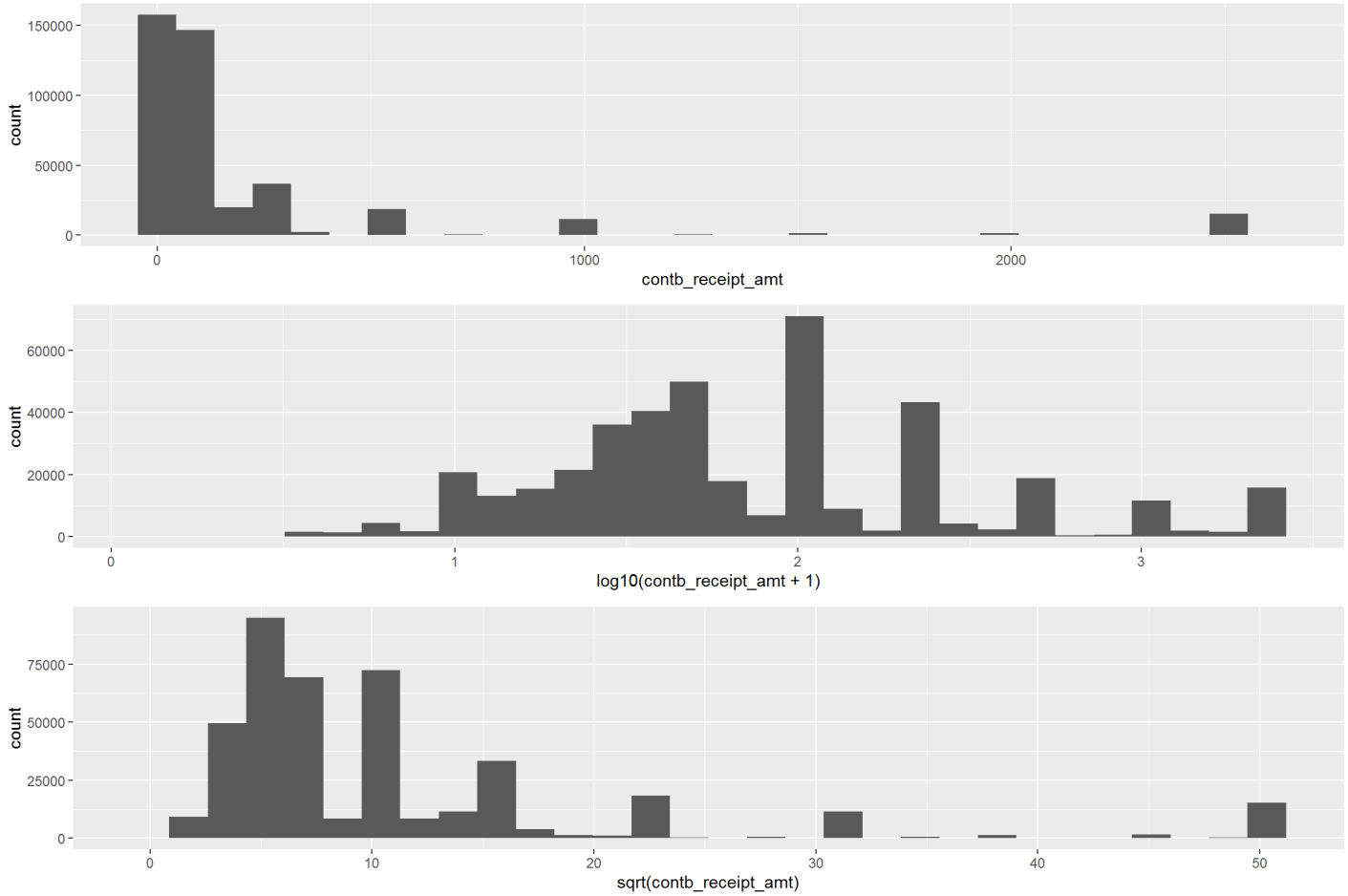
Looking at the plots, we see that the data is over-dispersed. Most of the contriibution amounts are $100 or less.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.45   25.00   55.00  232.50  150.00 2600.00
```

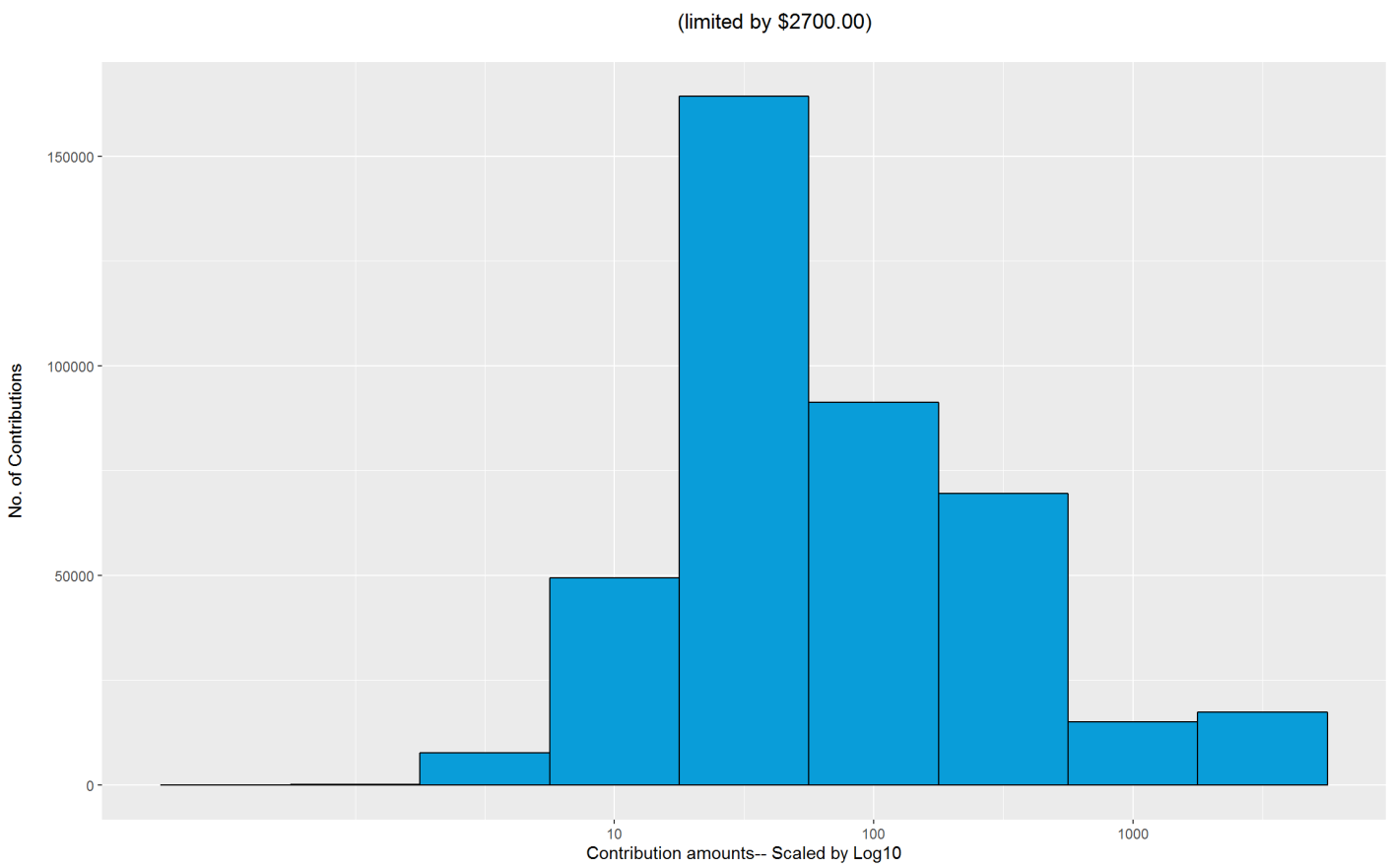
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1614  1.4150  1.7480  1.8700  2.1790  3.4150
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6708  5.0000  7.4160 11.2300 12.2500 50.9900
```

Transforming the scale of the x-axis to a log10 gives us a more normal distribution of the contributions.



By applying a scale_x_log10 layer to the data transforms the contribution receipt data into a normal distribution. We can confirm that approx 300,000 contributions of the 414,741 were less than or equal to $100.



To confirm this, I created a new data frame for contributions of $100 or less and we that there are 299,577 observations.

```
## 'data.frame':    299577 obs. of  19 variables:
##  $ ID               : int  2 3 4 6 7 8 9 10 13 14 ...
##  $ cmte_id          : Factor w/ 15 levels "C00410118","C00431171",..: 7 3 3 3 3 3 3 3 3 7 ...
##  $ cand_id          : Factor w/ 14 levels "P00003608","P20002523",..: 12 13 13 13 13 13 13 13
13 12 ...
##  $ cand_nm          : Factor w/ 14 levels "Bachmann, Michele",..: 8 7 7 7 7 7 7 7 7 8 ...
##  $ contbr_nm        : Factor w/ 97577 levels "5146 HIGHBRIDGE ST. L.L.C.",..: 58390 56363 940
96 49147 87432 78127 51203 25945 23715 65497 ...
##  $ contbr_city      : Factor w/ 2206 levels "","," NEW YORK",..: 565 1344 1344 2041 1344 1344 1
484 174 930 1560 ...
##  $ contbr_st        : Factor w/ 1 level "NY": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contbr_zip       : Factor w/ 50520 levels "","*1046","0",..: 45600 1608 8309 48338 7480 97
48 21633 45343 50071 45177 ...
##  $ contbr_employer  : Factor w/ 36541 levels "","-","--","---",..: 3808 28692 26951 10712 624
8 31618 28692 26951 28692 4519 ...
##  $ contbr_occupation: Factor w/ 16882 levels "","$200 ","(NONUNION) F/T K.B.S.I",..: 7811 166
63 12948 1321 668 12948 10983 12948 16663 4008 ...
##  $ contb_receipt_amt: num  100 50 100 100 50 100 50 100 25 50 ...
##  $ contb_receipt_dt : Factor w/ 644 levels "1-Apr-11","1-Apr-12",..: 565 580 419 125 520 398
462 602 188 565 ...
##  $ receipt_desc     : Factor w/ 17 levels "","REATTRIBUTED",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_cd          : Factor w/ 2 levels "","X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_text        : Factor w/ 152 levels "","$1000 REFUND TO BE ISSUED",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ form_tp          : Factor w/ 3 levels "SA17A","SA18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ file_num         : int  779227 756218 756218 756218 756218 756218 756218 756218 756218 779
227 ...
##  $ tran_id          : Factor w/ 391954 levels "0704020-0001",..: 6680 16441 18254 16993 15242
18068 18761 16457 17262 6633 ...
##  $ election_tp      : Factor w/ 7 levels "","G2008","G2012",..: 7 7 7 7 7 7 7 7 7 7 ...
```

Introducing a new variable - Party Affilitation

There are 14 candidates in the donors dataset: 1 Democrat, 1 Libertarian, 1 Green Party candidate, and 11 Republicans. I am going to create a new variable, “party_affl”, to record the party affiliation for each candidate. The party function below uses a “switch” statement to assign a political party to the candidate.

A quick test shows that the function matches party affilitation to the candidate.

```
## [1] "Democrat"
```

```
## [1] "Libertarian"
```

```
## [1] "Green Party"
```

```
## [1] "Republican"
```

So, we add the variable to the dataframe.

We now see that there is an additional variable to the dataset

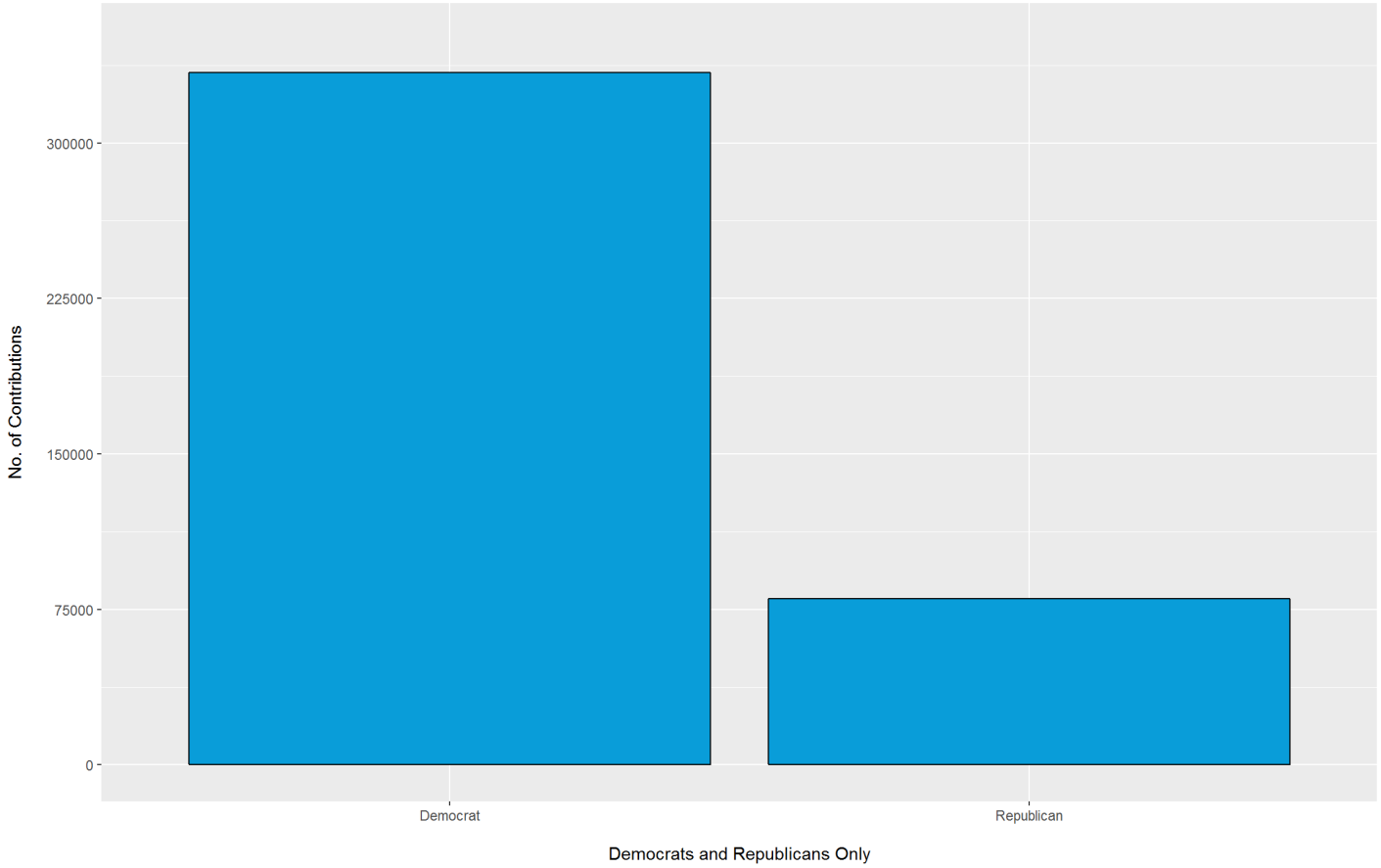
```
## 'data.frame':    414741 obs. of  20 variables:
##  $ ID               : int  353 364 365 366 367 376 377 378 379 384 ...
##  $ cmte_id          : Factor w/ 15 levels "C00410118","C00431171",..: 11 11 11 11 11 11 11 11
11 11 ...
##  $ cand_id          : Factor w/ 14 levels "P00003608","P20002523",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ cand_nm          : Factor w/ 14 levels "Bachmann, Michele",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ contbr_nm        : Factor w/ 97577 levels "5146 HIGHBRIDGE ST. L.L.C.",..: 90069 91053 910
53 91053 91436 92129 93593 93594 93594 93594 ...
##  $ contbr_city      : Factor w/ 2206 levels "","," NEW YORK",..: 1344 238 238 238 1481 252 1662
1662 1662 1662 ...
##  $ contbr_st        : Factor w/ 1 level "NY": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contbr_zip       : Factor w/ 50520 levels "","*1046","0",..: 4894 20260 20260 20260 50134
27281 48401 48401 48401 48401 ...
##  $ contbr_employer  : Factor w/ 36541 levels "","-","--","---",..: 22722 26951 26951 26951 21
252 6719 27459 27459 27459 27459 ...
##  $ contbr_occupation: Factor w/ 16882 levels "","$200 ","(NONUNION) F/T K.B.S.I",..: 11958 12
948 12948 12948 12948 89 13189 13189 13189 13189 ...
##  $ contb_receipt_amt: num  75 25 50 25 75 50 250 250 50 75 ...
##  $ contb_receipt_dt : Factor w/ 644 levels "1-Apr-11","1-Apr-12",..: 311 600 208 586 184 447
331 26 152 321 ...
##  $ receipt_desc     : Factor w/ 17 levels "","REATTRIBUTED",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_cd          : Factor w/ 2 levels "","X": 1 1 1 1 1 1 1 1 1 1 ...
##  $ memo_text        : Factor w/ 152 levels "","$1000 REFUND TO BE ISSUED",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ form_tp          : Factor w/ 3 levels "SA17A","SA18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ file_num         : int  762366 762366 762366 762366 762366 762366 762366 762366 762366 762
366 ...
##  $ tran_id          : Factor w/ 391954 levels "0704020-0001",..: 7621 8164 8853 8035 8244 784
6 8584 8144 8390 8369 ...
##  $ election_tp      : Factor w/ 7 levels "","G2008","G2012",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ party_affl       : chr  "Republican" "Republican" "Republican" "Republican" ...
```

Table of donors by party affiliation

```
##
##    Democrat Green Party Libertarian  Republican
##      333945         142         362       80292
```

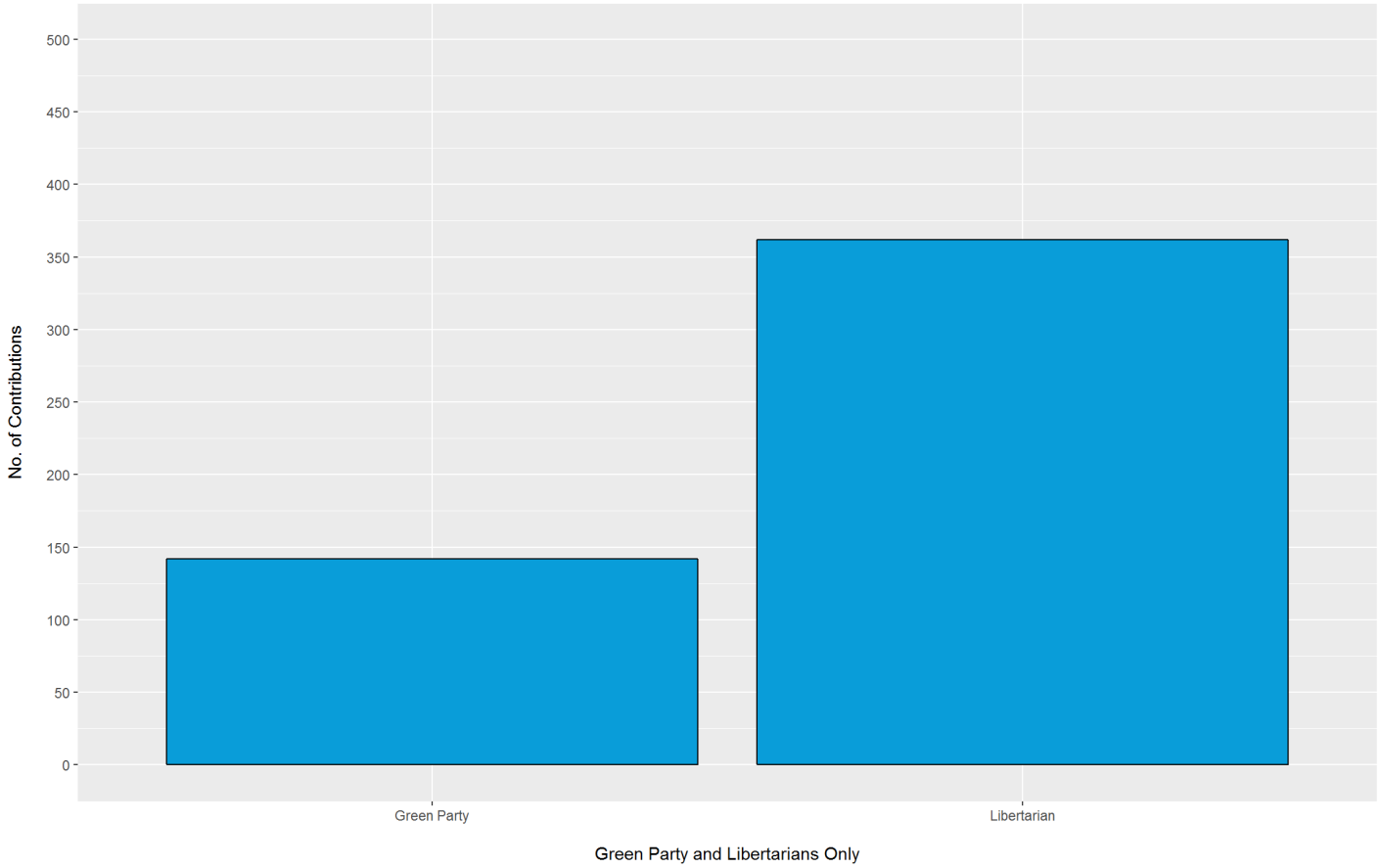
Given that the Democrats and Republicans have far larger donor bases, I will separate them out of the analysis from the other two parties. The plot below shows the political affliation of the donors. Even though there were 11 Republican candidates and only one Democratic candidate, the Democrats had over 4 times the number of donors than the Republicans.

Part One: 2012 Political Party Affiliations



The number of donors to the Green Party and Libertarians was not significant to the dataset. The Libertarians got more than double the amount of the Green Party, but contributions to both parties was not very significant.

Part One: 2012 Political Party Affiliations



Part One Summary of key findings - Univariate Analysis

- **How many NY state residents contributed money to presidential candidates in 2012?**

Using the legal limit of $2,700, a total of 414,741 NY state residents donated to 14 different presidential candidates in 2012. This report cannot say what this total amount of donors represents. This total would have to be compared with the total amount from other states to put into proper perspective.

- **What is the mean and median contributed amounts by NY state residents to presidential candidates in 2012?**

The mean amount was $232.50, and the median amount was $55.

- **Was the data skewed?**

Yes, we saw that the nearly 300,000 of the 414,741 contributions were $100 or less, and we also saw a significant number of donors at the legal limit.

- **What was learned after the contribution amounts were scaled by log10?** *

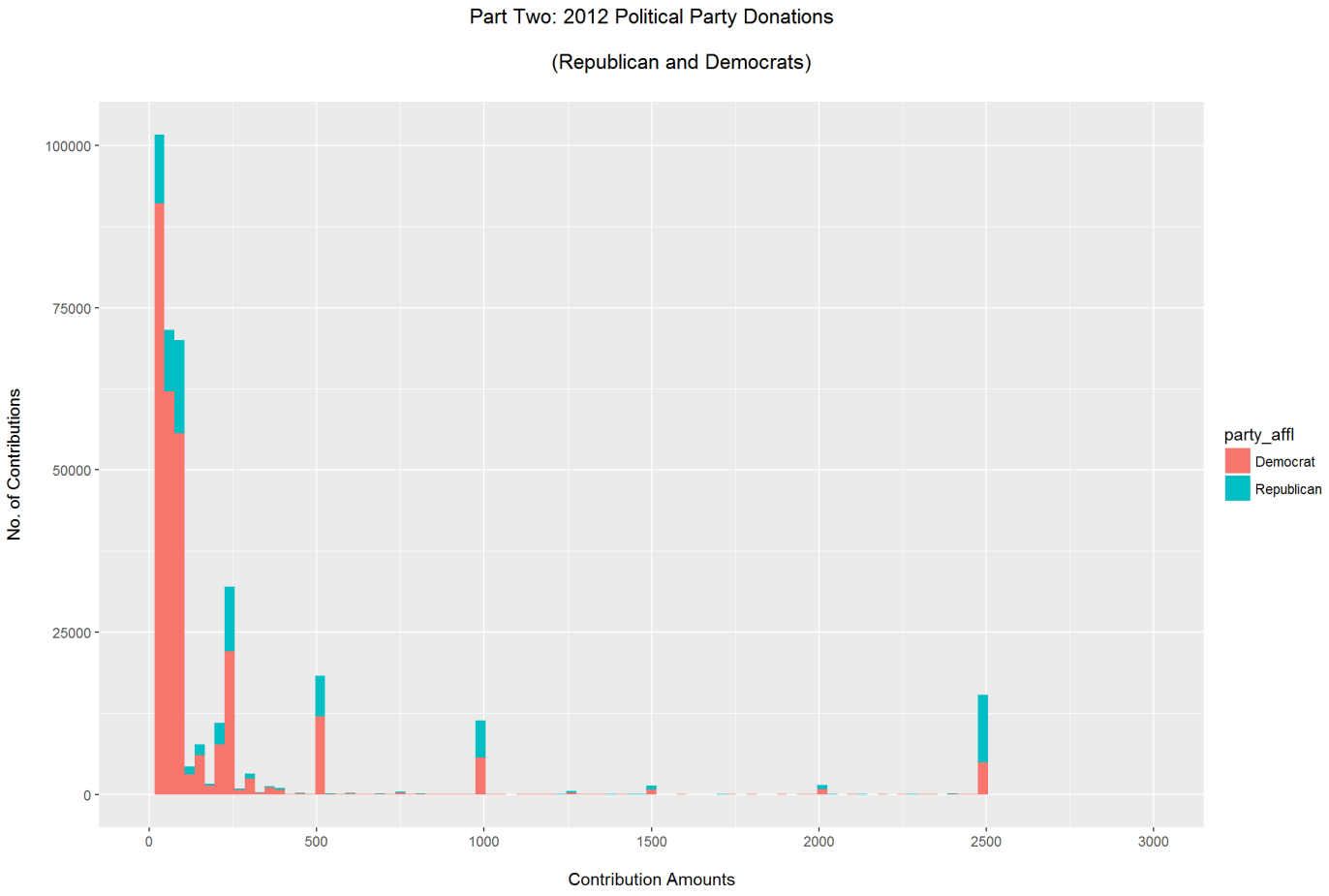
We saw a normal distribution of the data which confirmed that the bulk of the contributions was $100 or less.

- **What was learned about the political affiliation of the donors?** *

The plots show that party affiliation was a major for which party got the most donations. Democrats got 80% of all donations, and the Republicans got 19%. The two other parties got less than 1%.

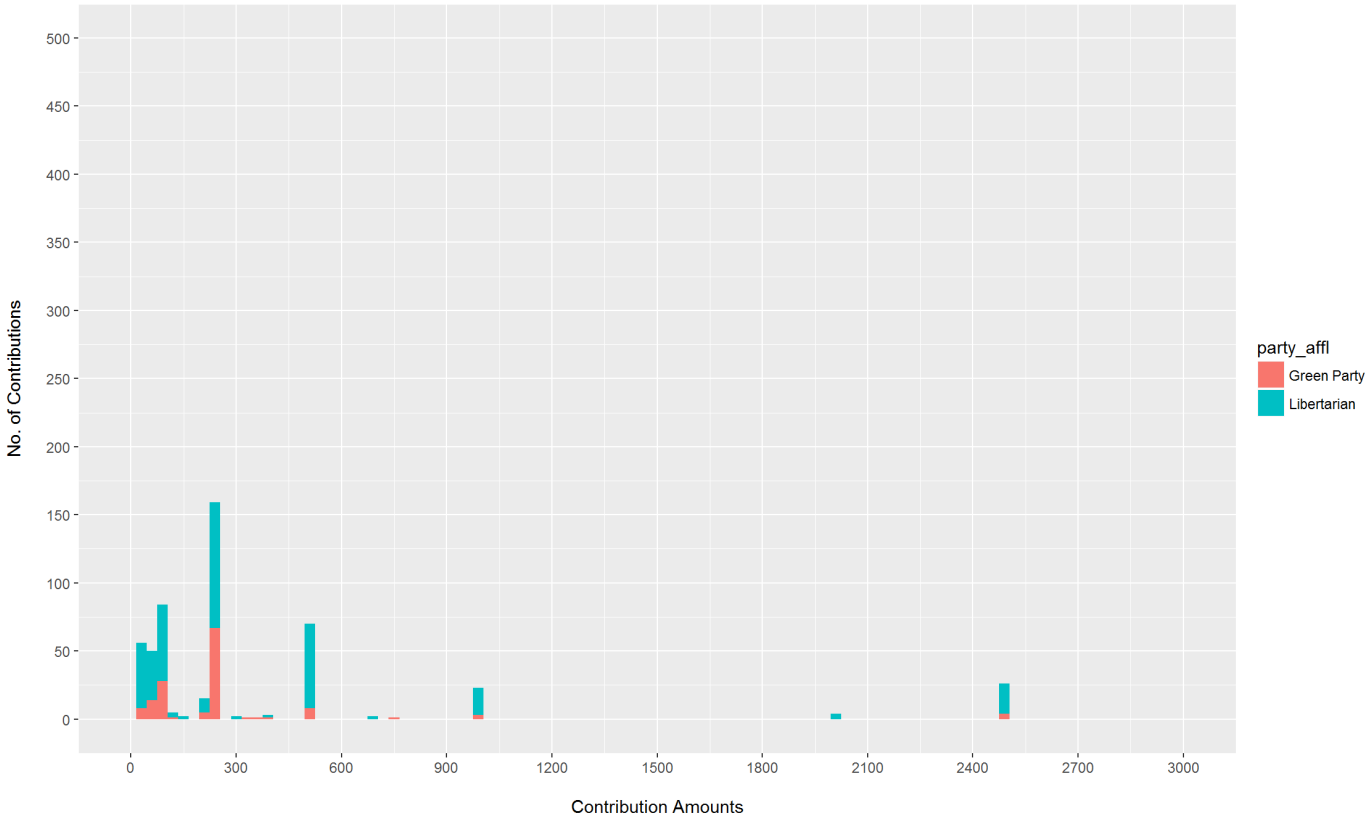
PART TWO - Bivariate Analysis - Party Affiliation and Contributions

Histogram Comparison of the total donation amounts based on party affilitation.



Part Two: 2012 Political Party Donations

(Green Party and Libertarians)

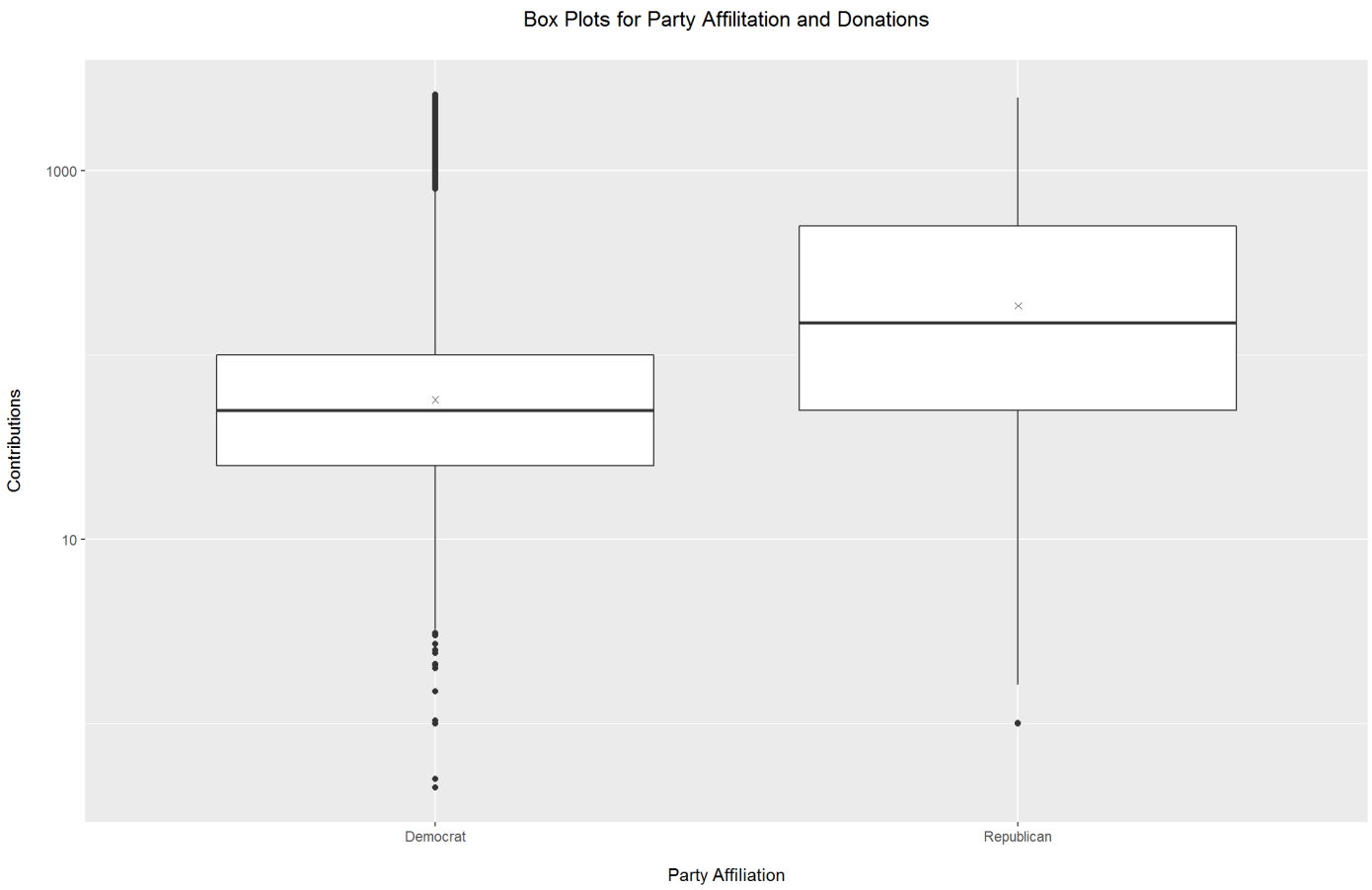


Comparing the two histograms, we can see that the Democrats and Republicans clearly got the most donations in terms of both amounts and number of contributions. The Democratic party collected $51,381,561.00 and the Republicans collected $44,858,263. Meanwhile, the Libertarians only collected $151,141.40 and the Green Party came in last at $40,471.53

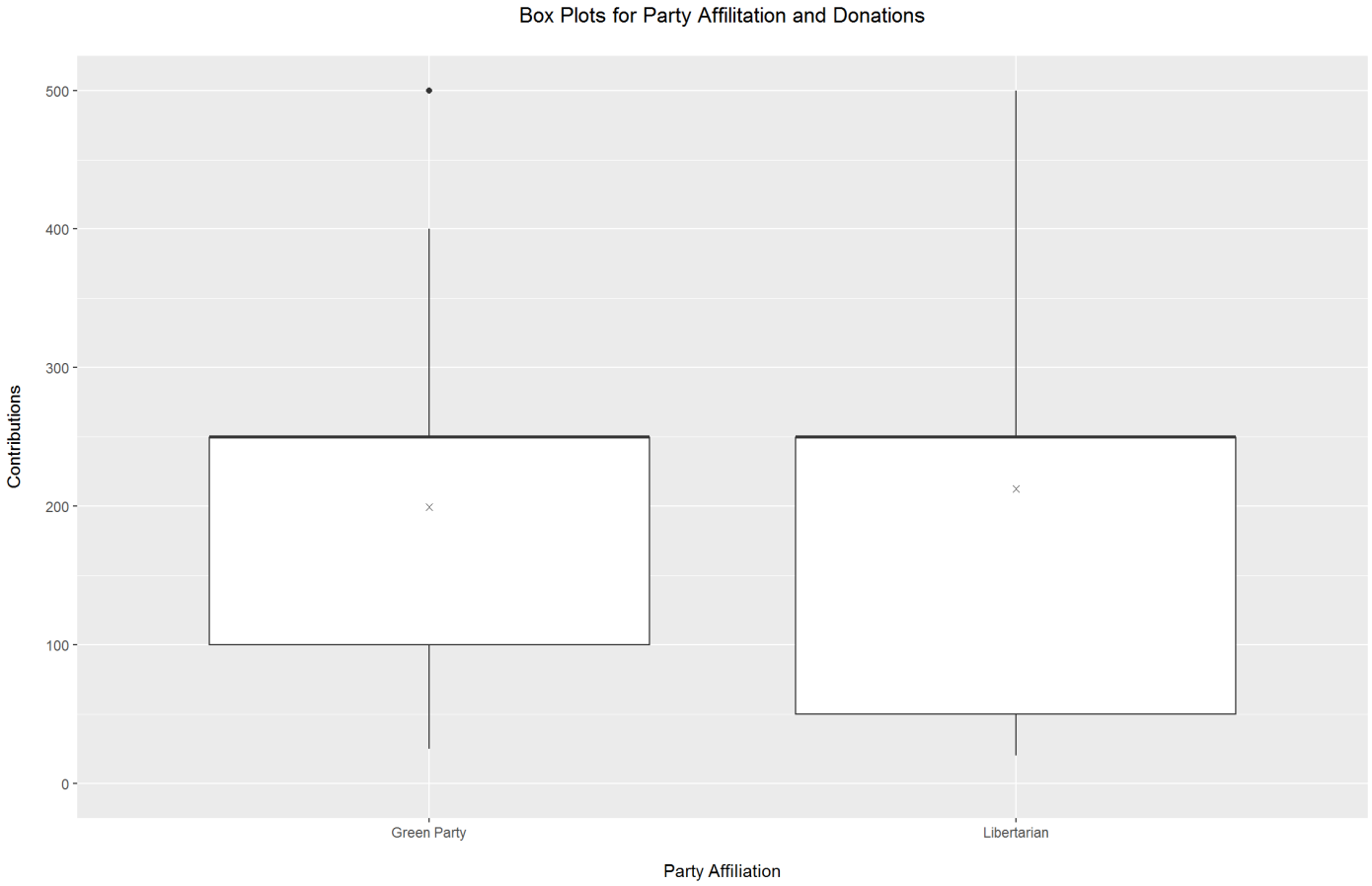
```
## NYDonors$party_affl: Democrat
## [1] 51381561
## ------------------------------------------------------------
## NYDonors$party_affl: Green Party
## [1] 40471.53
## ------------------------------------------------------------
## NYDonors$party_affl: Libertarian
## [1] 151141.4
## ------------------------------------------------------------
## NYDonors$party_affl: Republican
## [1] 44858263
```

Box plots comparison for contributions based on party affiliation

The box plots below compare the contributions for the Democrats and Republicans. This plot shows that the Democrats had a lower average donation and many more outlier donations than did the Republicans.

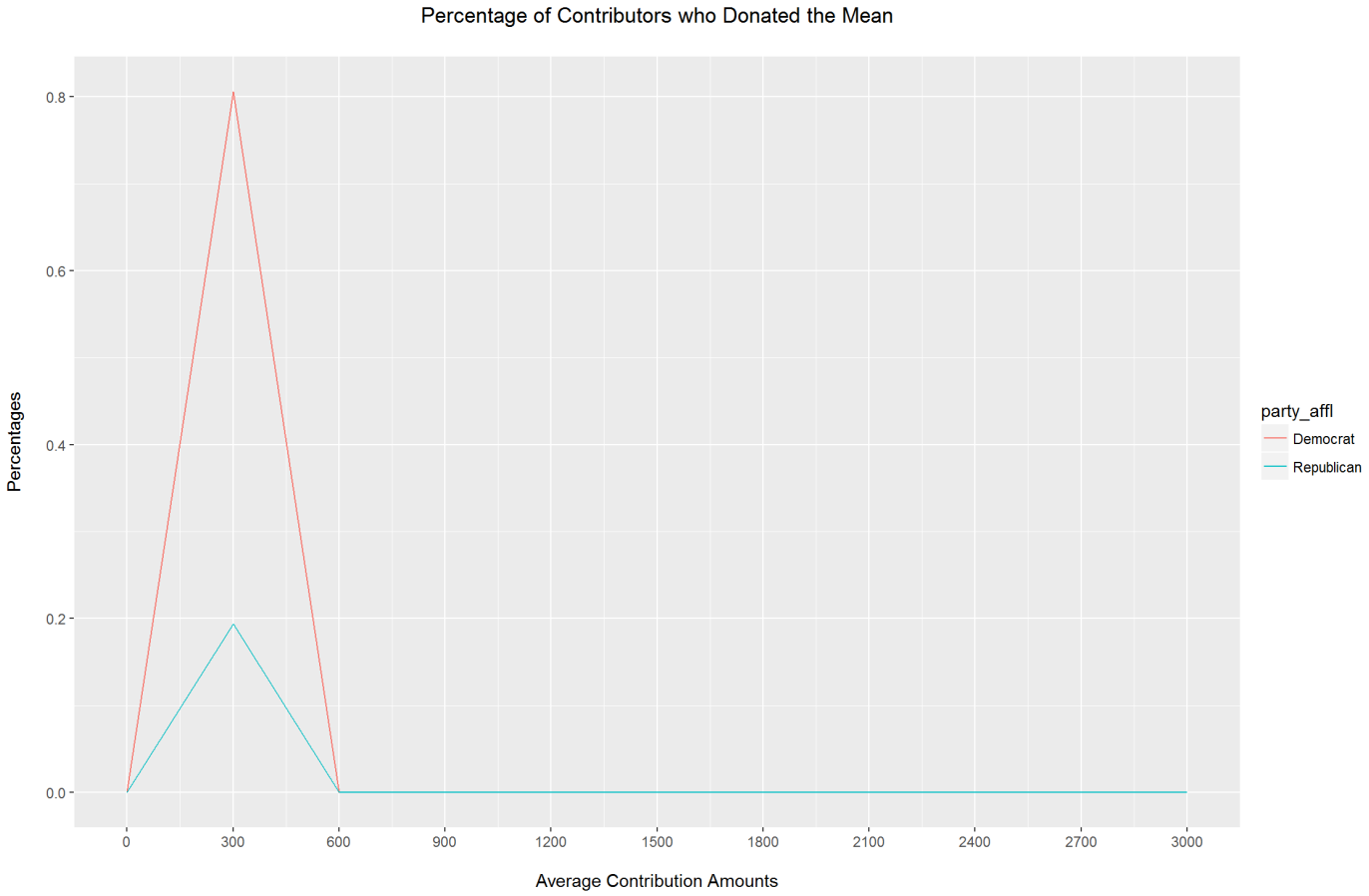


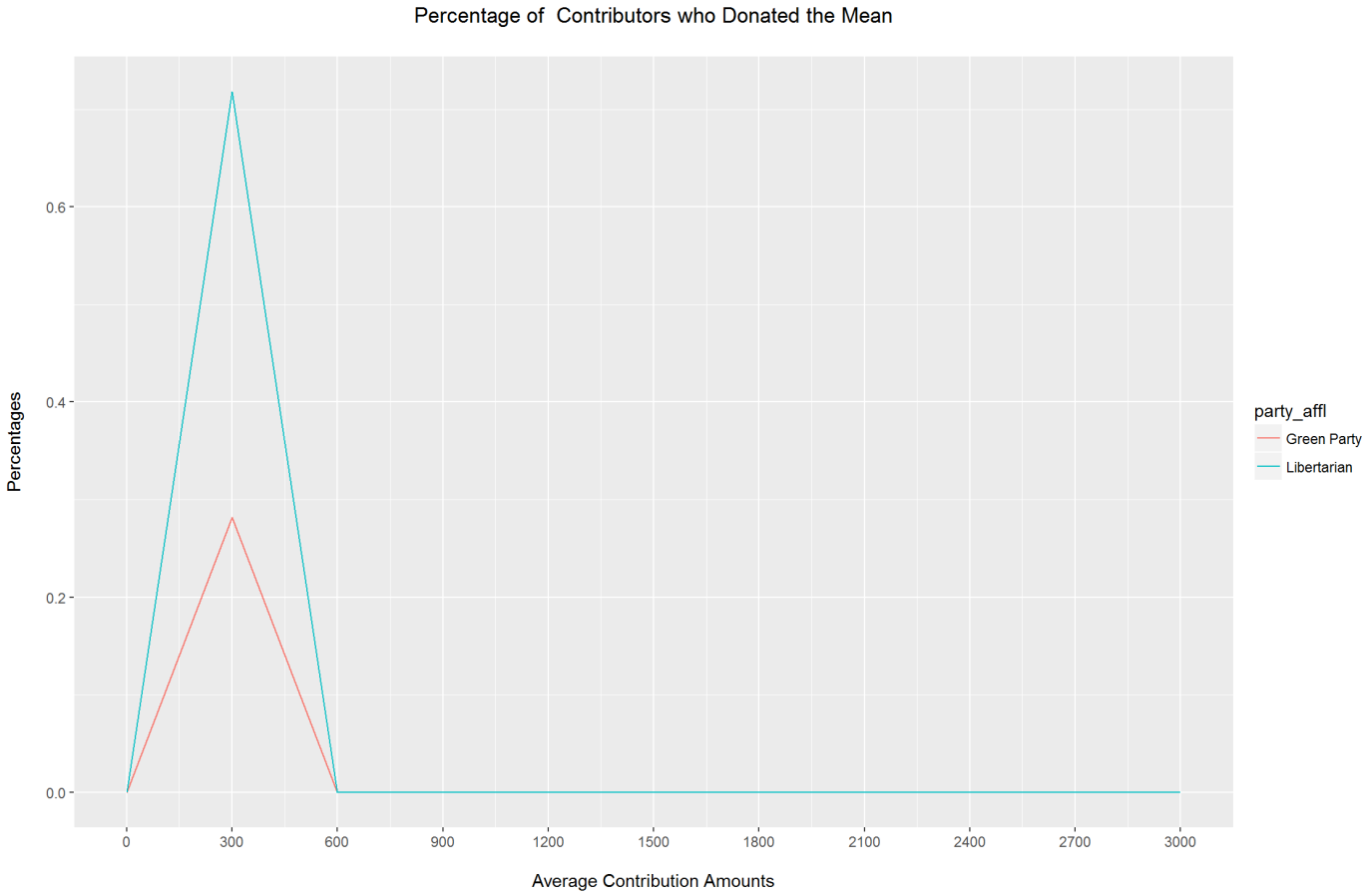
On a much smaller scale, the Green Party and Libertarians had approximately the same average donations with little to no outliers.



Frequency plots – “Star Fleet” logo

The plots below show the percentage of donors who donated the average donations for the Republicans and Democrats. The same plot is done for the Libertarians and the Green Party.



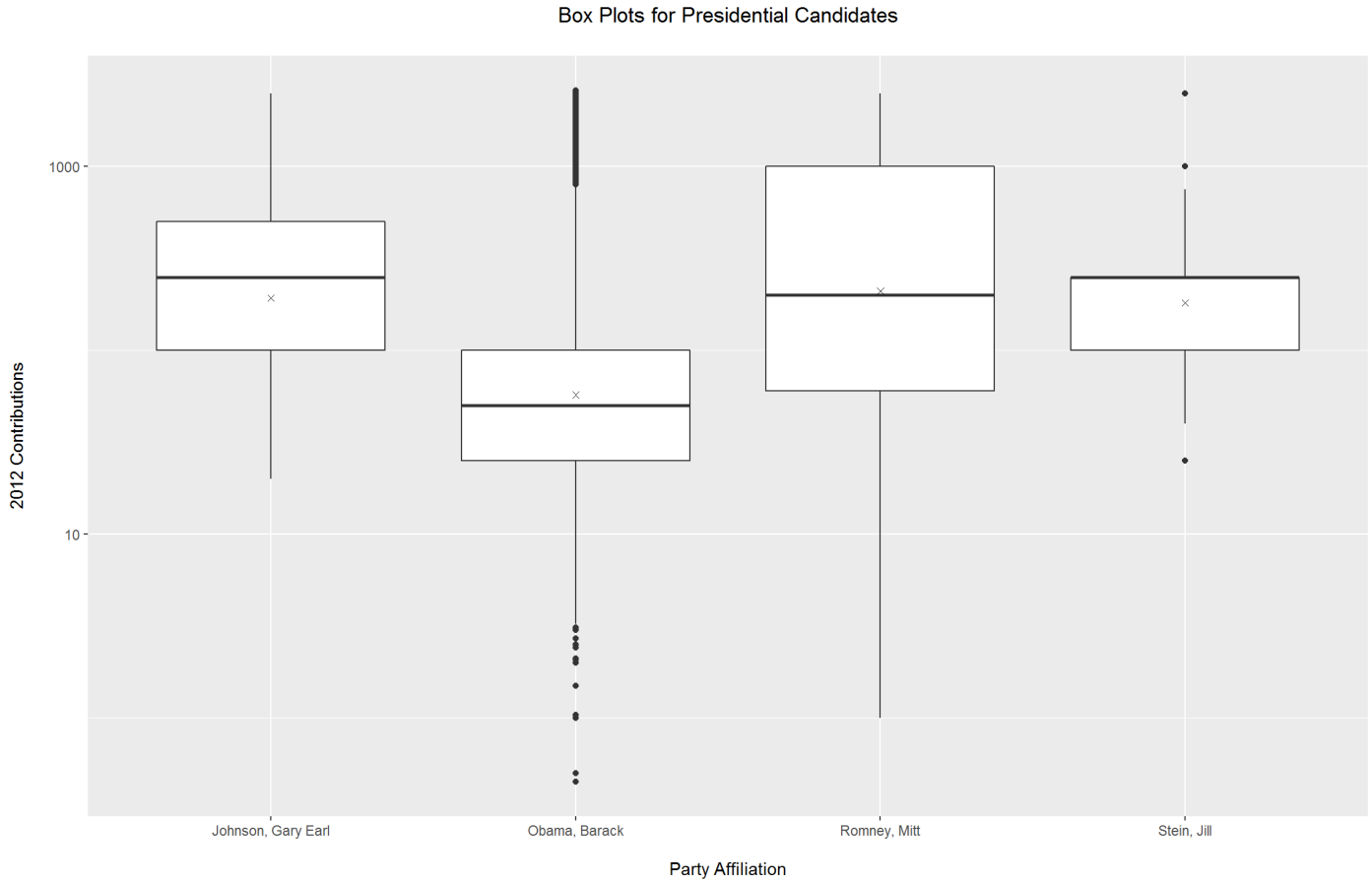


Creating a new data frame for the 4 presidential nominees

In 2012, the nominees for president were Mitt Romney (Republican), Barack Obama (Democrat), Jill Stein (Green Party), and Gary Johnson (Libertarian). We will start by creating a new data frame just for these four candidates

```
##
##              Bachmann, Michele                  Cain, Herman
##                              0                             0
##                 Gingrich, Newt                Huntsman, Jon
##                              0                             0
##             Johnson, Gary Earl           McCotter, Thaddeus G
##                            362                             0
##                  Obama, Barack                   Paul, Ron
##                         333945                             0
##              Pawlenty, Timothy                  Perry, Rick
##                              0                             0
## Roemer, Charles E. 'Buddy' III                 Romney, Mitt
##                              0                         67107
##                 Santorum, Rick                  Stein, Jill
##                              0                           142
```

In the new data frame, the other Republican candidates have been removed. The box plot below shows that Barack Obama (Democrat) had the lowest average contribution amount.



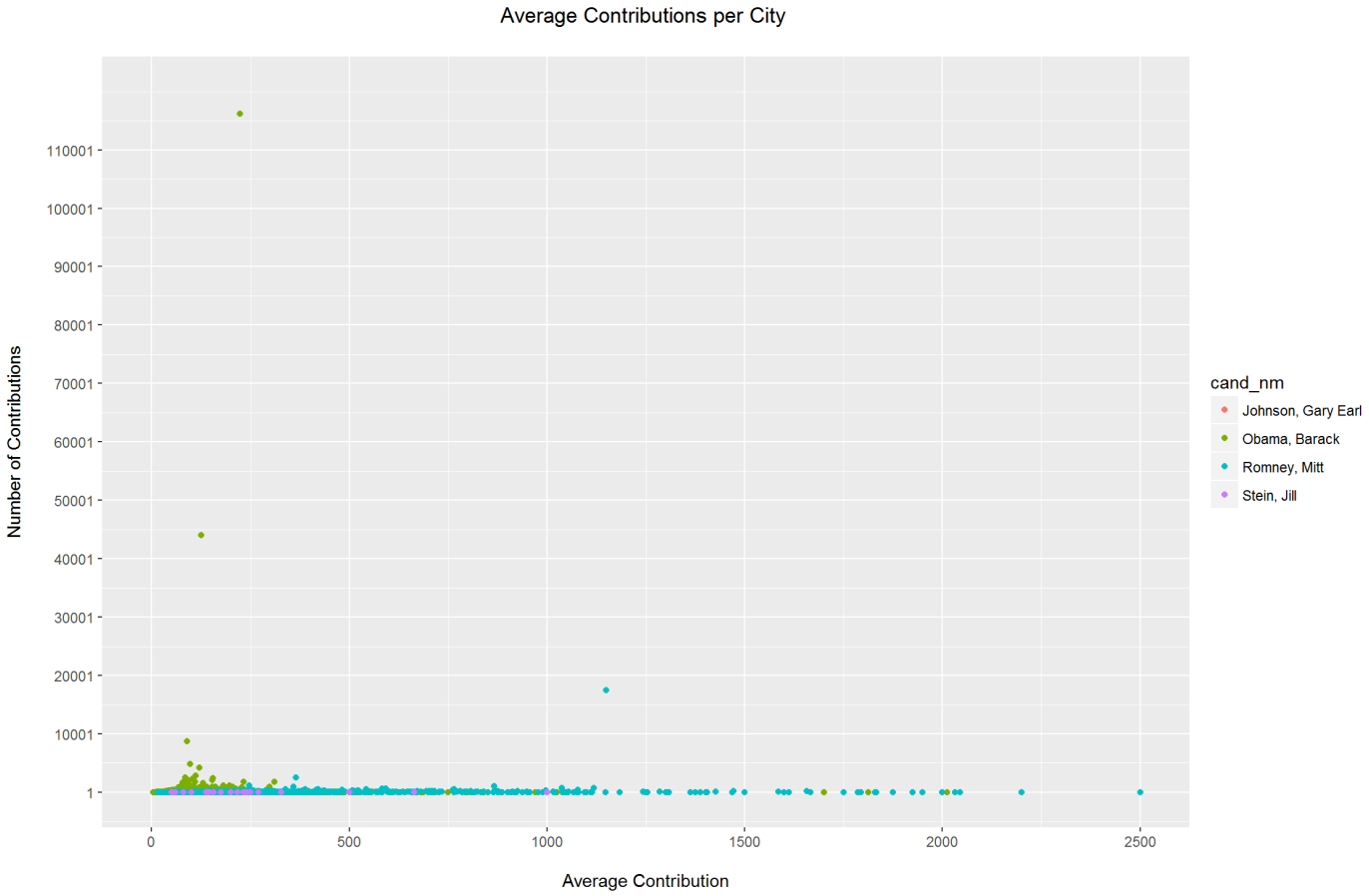
Conditional Means.

Creating a data frame that summarizes data per candidate per city.

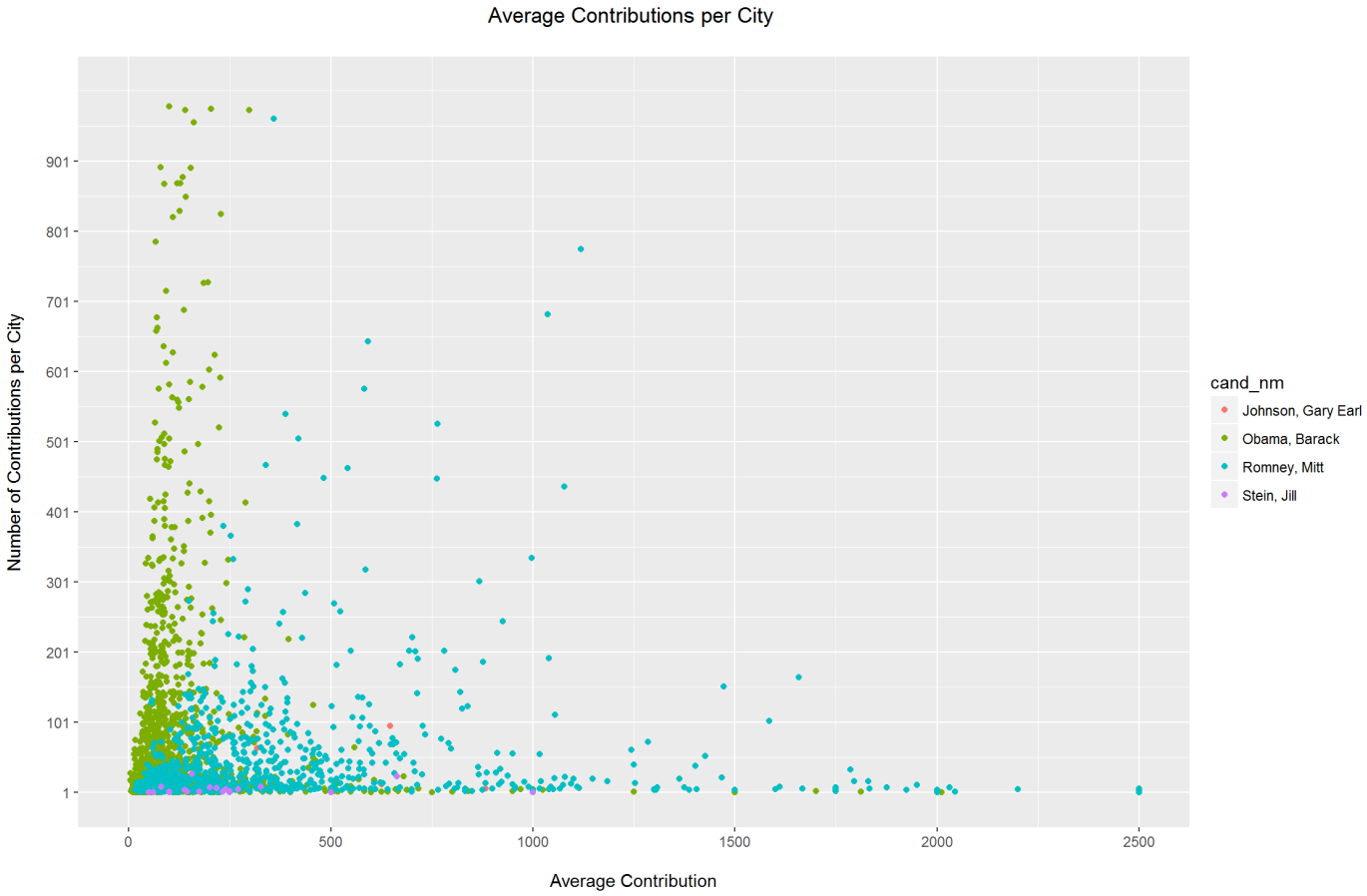
```
## Source: local data frame [6 x 7]
## Groups: cand_nm [1]
##
##              cand_nm contbr_city contribution_average contribution_median
##               <fctr>      <fctr>                <dbl>               <dbl>
## 1 Johnson, Gary Earl     ASTORIA               37.500                37.5
## 2 Johnson, Gary Earl     BEDFORD             1000.000              1000.0
## 3 Johnson, Gary Earl   BOONVILLE              150.000               100.0
## 4 Johnson, Gary Earl     BROOKLY             2500.000              2500.0
## 5 Johnson, Gary Earl    BROOKLYN              317.835                37.5
## 6 Johnson, Gary Earl        CLAY              300.000               300.0
## # ... with 3 more variables: contribution_std_dev <dbl>,
## #   contribution_sum <dbl>, n <int>
```

Scatter plot

When we plot the average contribution against the number of contributions per city and color the points by candidate, we see that the data is highly skewed by the fact that Obama received over 116,000 donations just from New York city alone. In order to get a better understanding of the average donation per city to each candidate, we will adjust the plotting to get a closer examination.

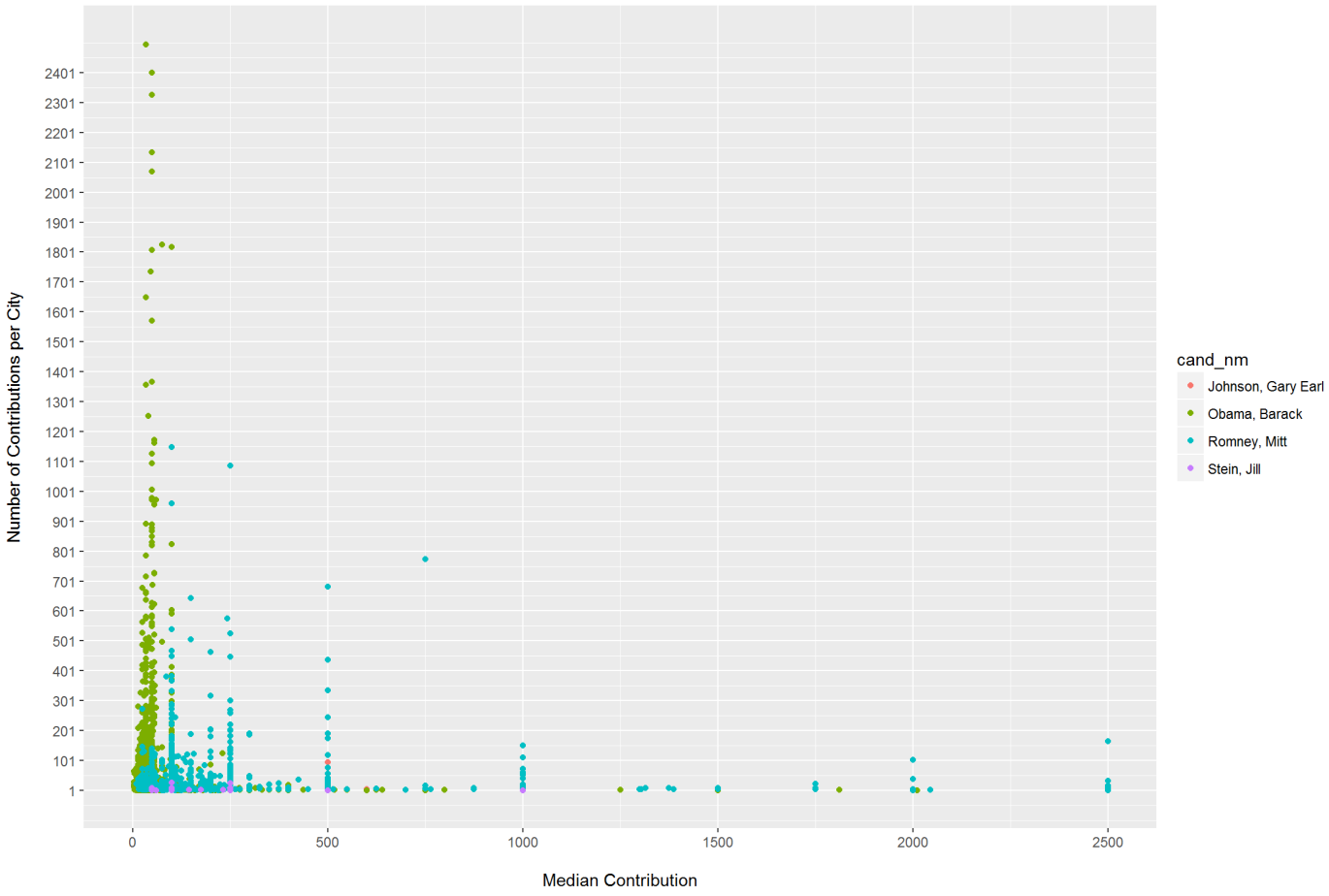


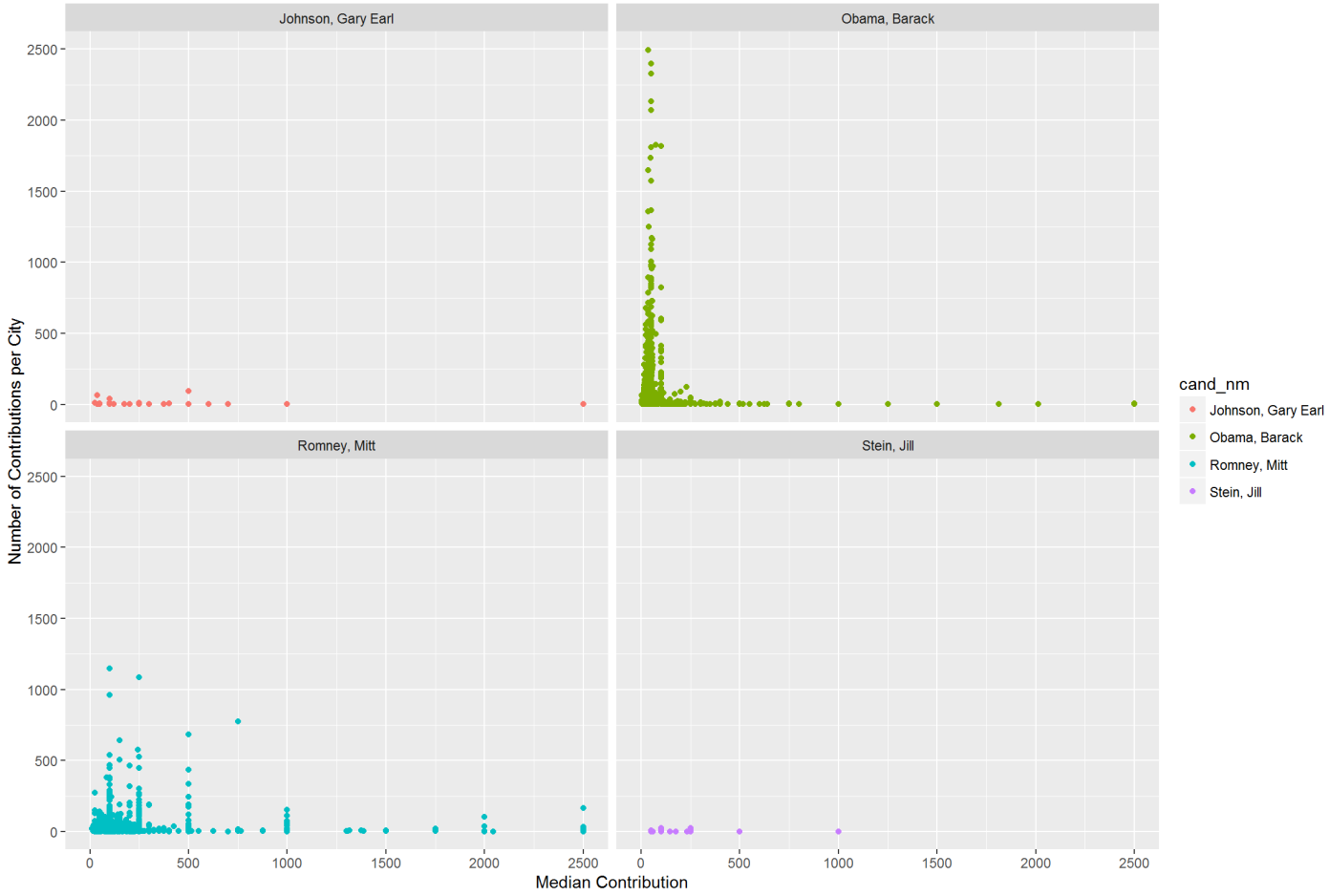
Adjusted contributions per city based on average. This new scatter plot shows that the Obama had lower average contributions per city, but he had a higher number of contributors per city. The reverse is true for Mitt Romney who had a much higher average contribution per city, but a lower number of donations per city. Jill Stein had both a low average per city contribution and a low number of contributions per city, and Gary Johnson had very low numbers of donors per city..



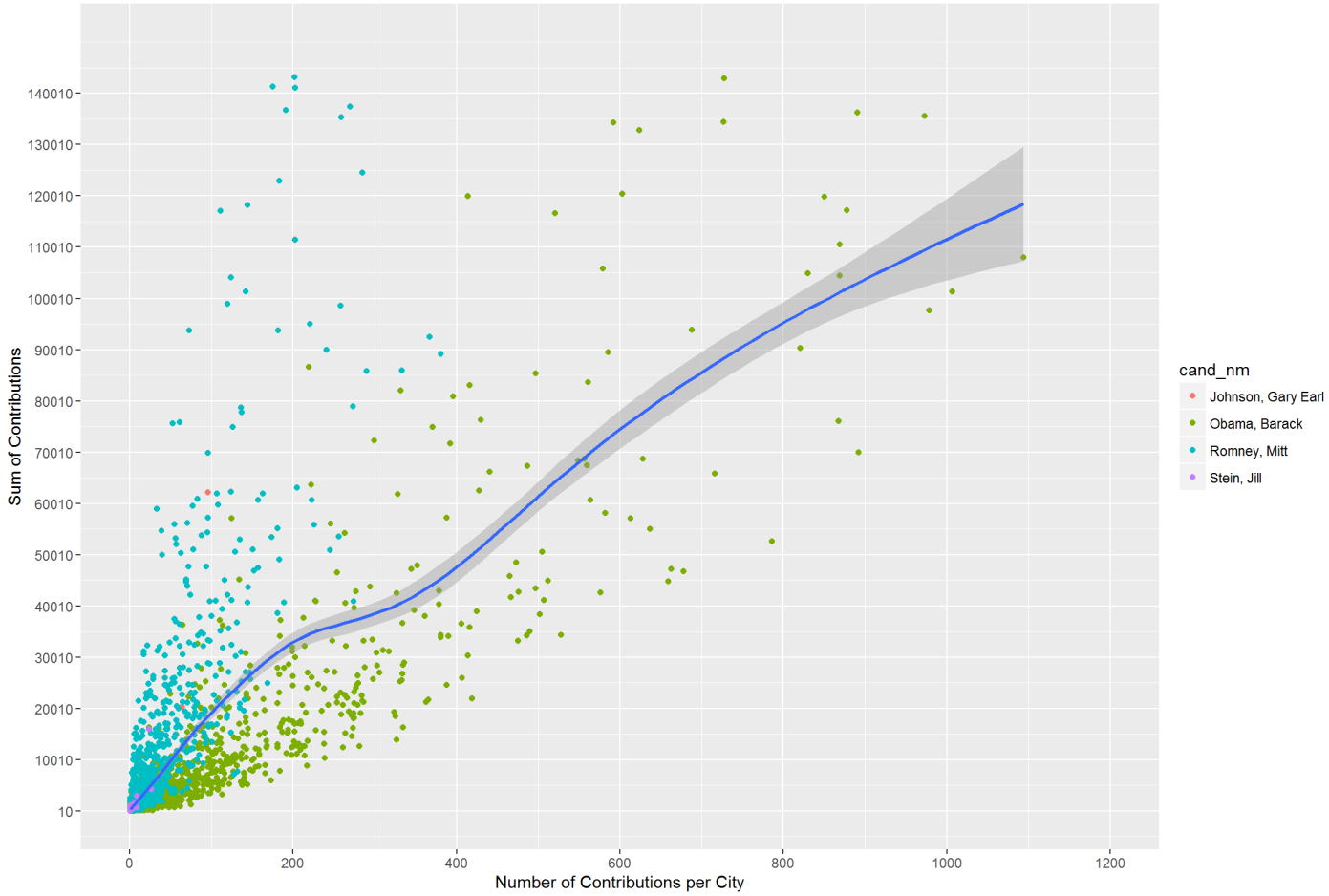
We can see this more clearly by doing a facet wrap on the candidates names

Contributions per city based on median





Correlation between Number of Contributions per city and the total amount of contributions. Is there a correlation between the total amount of donations per city and the number of donations per city?




```
##
##  Pearson's product-moment correlation
##
## data:  NYDonors.Contributions_by_Candidates_City$n and NYDonors.Contributions_by_Candidates_C
ity$contribution_sum
## t = 96.502, df = 3223, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8527658 0.8705298
## sample estimates:
##      cor
## 0.861912
```

There's a strong correlation between the number of donations and the total amounts donated.

Part Two Summary of key findings - Bivariate Analysis

- **Does party affiliation have an impact on the total donation amounts?**

Yes. Donations to the Democrats and the Republicans totaled: $96,239,824, and donations to the other two parties was $191,612.93

- **What is the mean and median contribution amounts to each party in 2012?**

Democrats: $153.86 Republican: $558.69 Green Party: $285.01 Libertarian: $417.52

- **What presidential candidate had the lowest average donation?** Barack Obama had the lowest average, but he also had the most outlier donations as the box plots showed.
- **What presidential candidate had the lowest average donation per city?** This new scatter plot shows that the Obama had lower average contributions per city, but he had a higher number of contributors per city. The reverse is true for Mitt Romney who had a much higher average contribution per city, but a lower number of donations per city. Jill Stein had both a low average per city contribution and a low number of contributions per city, and Gary Johnson had very low numbers of donors per city.
- **Was there a correlation between the number of contributions per city and the total amount of contributions?** The cor.test function showed that there was a strong correlation between the number of donations to a candidate and the total sum.
- **How many NY state residents contributed money to presidential candidates in 2012?**

Using the legal limit of $2,700, a total of 414,741 NY state residents donated to 14 different presidential candidates in 2012. This report cannot say what this total amount of donors represents. This total would have to be compared with the total amount from other states to put into proper perspective.

- **What is the mean and median contributed amounts by NY state residents to presidential candidates in 2012?**

The mean amount was $232.50, and the median amount was $55.

- **Was the data skewed?**

Yes, we see from the histogram that the vast majority of the contributions were $100 or less.

- **What was learned after the contribution amounts were scaled by log10?** *

We saw a normal distribution of the data which shows that the bulk of the contributions (over 150,000) was around $100

- **What is the political affiliation of the donors?** *

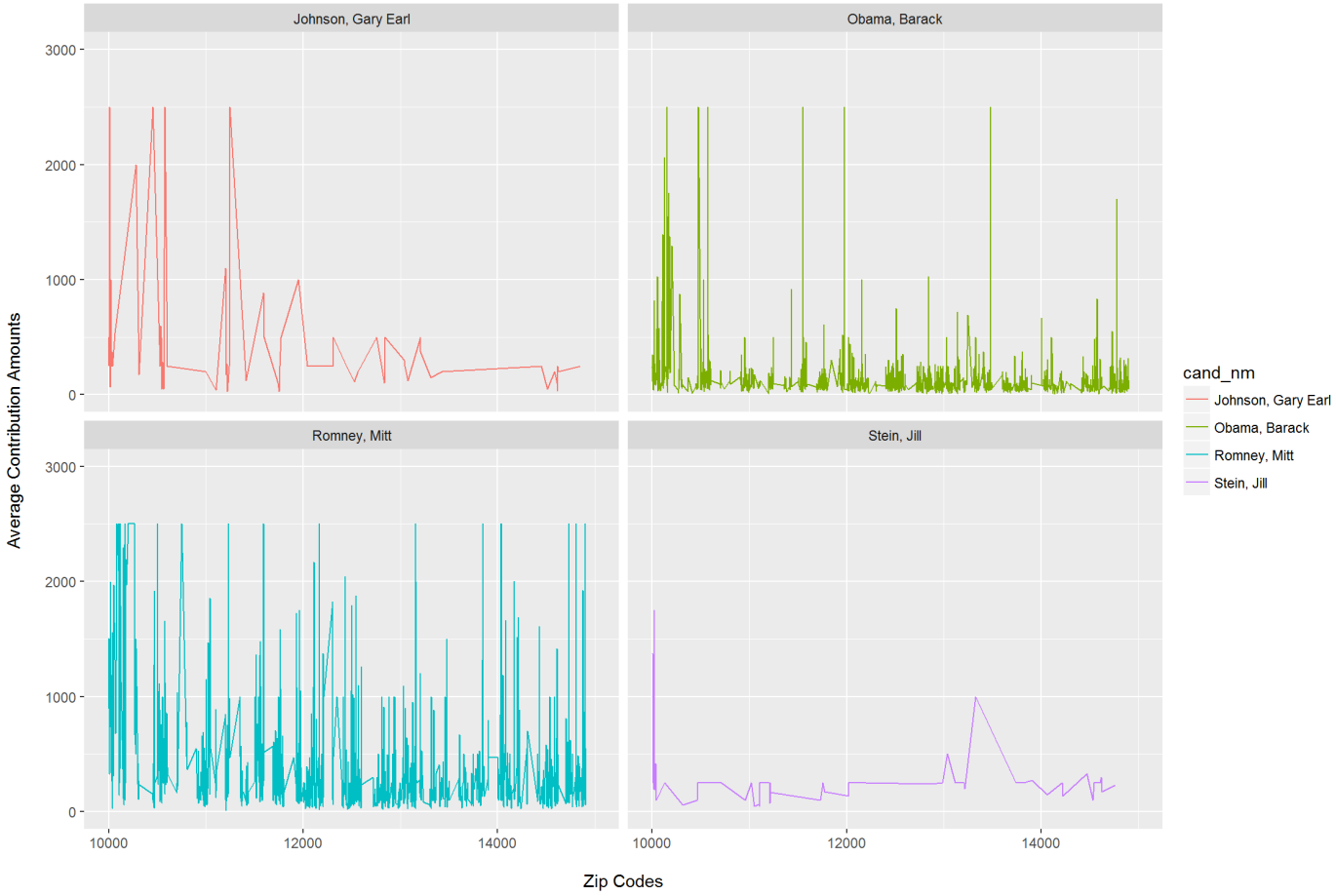
We see that 333,945 donated to the Democrats, 142 to the Green Party, 362 to the Libertarians, and 80,292 to the Republican. On a percentage basis, 80.5% of the donations went to the Democrats, 19.4% went to the Republicans, and less than 1% of the donations went to the Libertarians and Greens. The plots show that party affiliation is a major driver of which party gets the most donations.

PART THREE - MULTIVARIATE ANALYSIS

In this section, we will normalize the Zip codes. In the current dataset, some of the zip codes are Zip+5, and others are not. To normalize, we take the first five digits of the zip code and convert it to an integer.

Line plots

The follwing four line plots compare the average contribution amounts per zip codes for each of the four presidential candidates. For the state of New York, zip codes range from 10001 to 14925.



These plots show that Mitt Romney had the highest average campaign contributions across all Zip codes in the state.

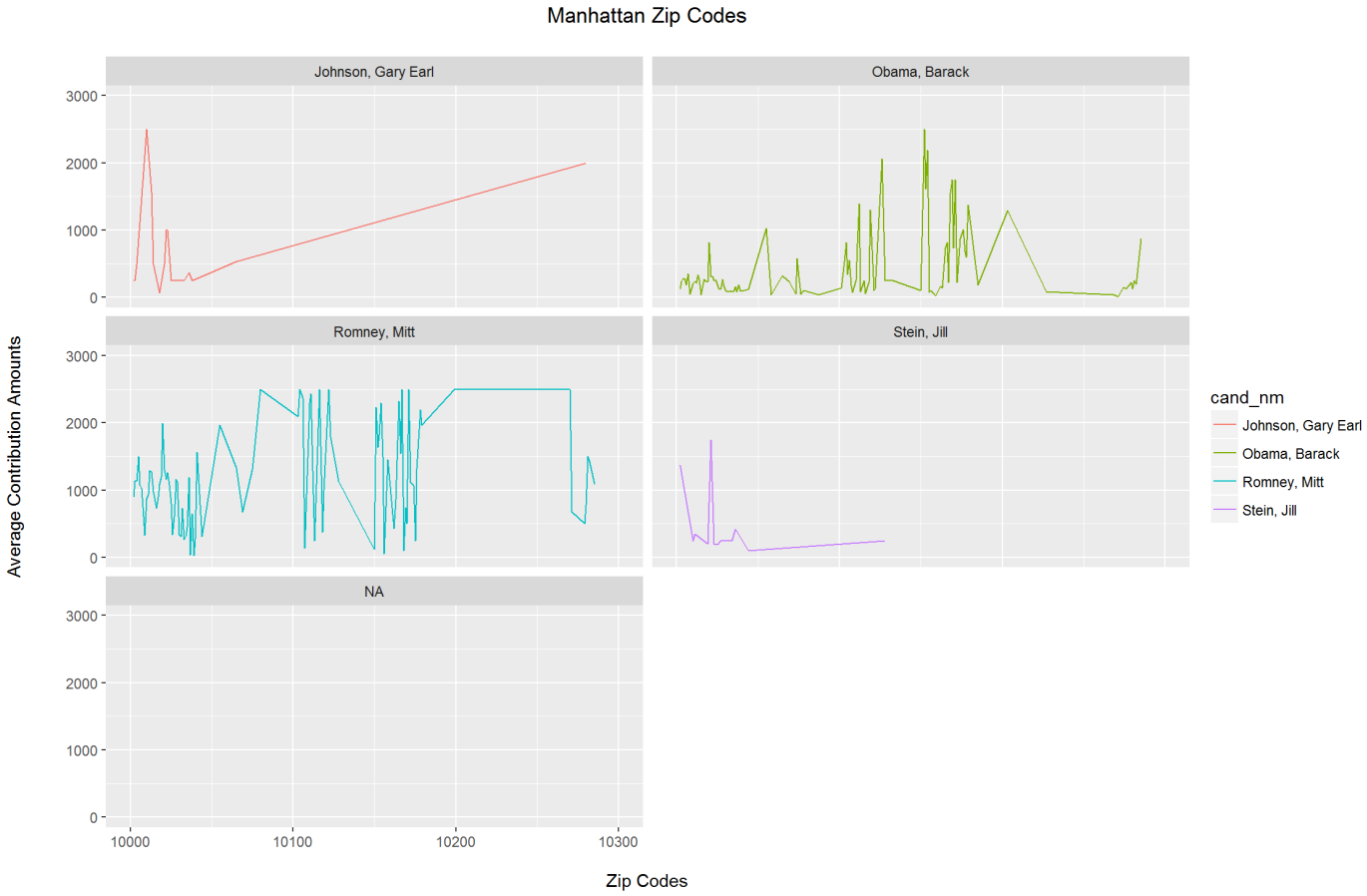
Adding a new variable for each range of zip code.

Putting the Zips into Ranges for the presidential nominees. The following code puts the zips into buckets per county in NY state. These ranges will be in a new variable called Zips.ranges into the

New York City - Manhattan analysis.

```
##
## (10001,10301] (10301,10451] (10451,10501] (10501,10509] (10509,10901]
##        135654          3455          9305          2219         38959
## (10901,10910] (10910,11001] (11001,11004] (11004,11201] (11201,12007]
##             0          8817           478         15385        102250
## (12007,12008] (12008,12010] (12010,12015] (12015,12017] (12017,12018]
##            35           309            90             0           117
## (12018,12019] (12019,12025] (12025,12031] (12031,12064] (12064,12108]
##           226           394           168          1946          1793
## (12108,12167] (12167,12401] (12401,12501] (12501,12701] (12701,12801]
##          2151          8087          3146          9745          1211
## (12801,12809] (12809,12851] (12851,12901] (12901,12914] (12914,12922]
##           383          1176          2138           115            99
## (12922,13020] (13020,13021] (13021,13028] (13028,13032] (13032,13040]
##          1324           238           313           189           468
## (13040,13053] (13053,13054] (13054,13065] (13065,13124] (13124,13143]
##           344             0           257          2755           467
## (13143,13305] (13305,13324] (13324,13601] (13601,13732] (13732,13737]
##          3397           432          2851          1326            76
## (13737,14001] (14001,14005] (14005,14008] (14008,14009] (14009,14029]
##          3069            56            26             8           182
## (14029,14041] (14041,14048] (14048,14098] (14098,14410] (14410,14414]
##           356           214          2324          7104            78
## (14414,14415] (14415,14424] (14424,14529] (14529,14805] (14805,14814]
##             0           511          3216          9949           144
```

The first range on the list above covers Manhattan. The plot below shows



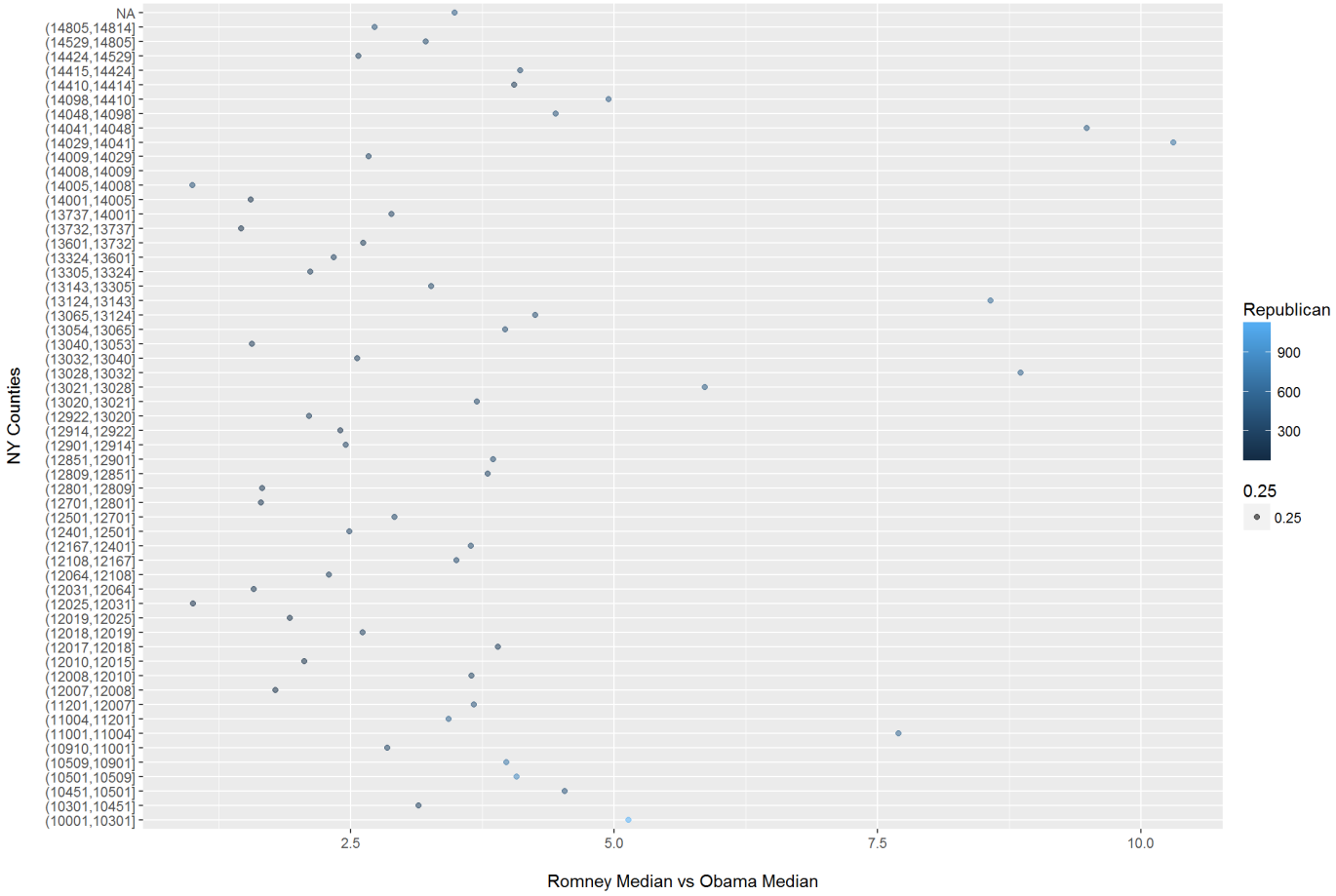
Summarizing the data by party affiliation and the Zips ranges

Let’s summarize the data based on party affiliation and zip codes for each county in NY state. The table shows the mean, median, standard deviation, sum, and the numer of donations (“n”) per NY county.

```
##   party_affl              Zips.ranges  contribution_average
##  Length:158         (10001,10301]:  4   Min.   :  43.83
##  Class :character   (10301,10451]:  4   1st Qu.:  97.26
##  Mode  :character   (10451,10501]:  4   Median : 197.15
##                     (10509,10901]:  4   Mean   : 252.27
##                     (10910,11001]:  4   3rd Qu.: 304.03
##                     (Other)      :135   Max.   :2500.00
##                     NA's         :  3
##  contribution_median contribution_std_dev contribution_sum
##  Min.   :  12.5      Min.   :   0.0       Min.   :      60
##  1st Qu.:  50.0      1st Qu.: 104.6       1st Qu.:    2100
##  Median : 100.0      Median : 195.5       Median :   17384
##  Mean   : 156.2      Mean   : 290.0       Mean   :  589724
##  3rd Qu.: 200.0      3rd Qu.: 456.1       3rd Qu.:  159728
##  Max.   :2500.0      Max.   :1020.8       Max.   :25995897
##                      NA's   :7
##        n
##  Min.   :     1
##  1st Qu.:    10
##  Median :   127
##  Mean   :  2541
##  3rd Qu.:   811
##  Max.   :117700
##
```

Re-Shape the data

The plot below shows a ratio between the median Romney contribution amount to that of Obama across all counties in NY state. There were 7 counties where Romney's median was at least 5 times that of Obama.



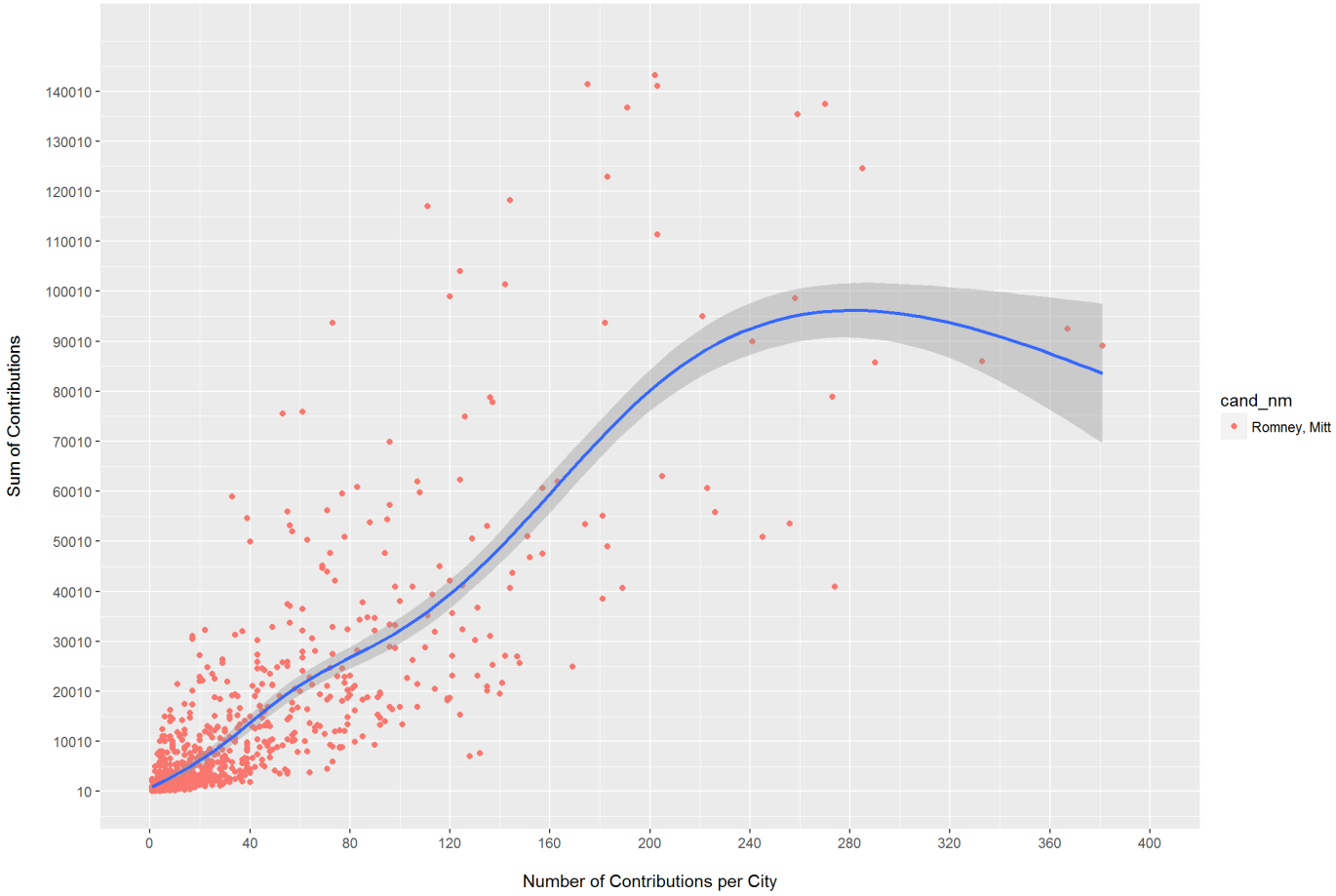
Correlation

As stated in Part two, we saw that there was a direct correlation between the number of donations that Obama got and the total amount donated to him. IOW, Obama relied on a large base of donors to match the donation amounts that Romney received from a relatively smaller base. In the plot below, we see that the more donations Obama received in New York city, the higher his total donations. I fit in a linear regression line in the model.

Obama

The same is not true for Romney. His total donations from New York city rose even though his total donors stayed relatively the same.

Romney



To examine this further, let's create two new dataframes just for Obama. The second dataframe summarises his campaign donations.

Now, we'll run a correlation test to see if the contribution sum correlates to the number of contributions, and the results show that it does.

```
##
##  Pearson's product-moment correlation
##
## data:  NYDonors_Obama.Contributions_by_County$contribution_sum and NYDonors_Obama.Contributio
ns_by_County$n
## t = 143.01, df = 1785, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9551287 0.9625880
## sample estimates:
##       cor
## 0.9590242
```

- **Which presidential nominee had the highest average contribution per zip code?** Mitt Romney had the highest average contribution amount per zip code according the line plot even though he had less total number of donors.
- **Which presidential nominee had the highest average contribution per Manhattan zip codes (range from 10000 to 10300)?** Again, Mitt Romney had the highest average contribution per Manhattan zip code.

- **Is there a correlation between the number of contributions and the total amount of contributions?** For Obama, yes, there was a strong correlation between the number of donations and the total amount of donations. However for Romney, there wasn't a correlation.

Part Three Summary of key findings - Multivariate Analysis

- **Which presidential nominee had the highest average contribution per zip code?** Mitt Romney had the highest average contribution amount per zip code according the line plot even though he had less total number of donors.
- **Which presidential nominee had the highest average contribution per Manhattan zip codes (range from 10000 to 10300)?** Again, Mitt Romney had the highest average contribution per Manhattan zip code.
- **Is there a correlation between the number of contributions and the total amount of contributions?** For Obama, yes, there was a strong correlation between the number of donations and the total amount of donations. However for Romney, there wasn't a correlation.

PART FOUR – Predicting contribution amounts based on number of donors.

Finally, we will build a prediction model based on the number of contributors and the Zip codes to see if we can predict the contribution sum. The R-squared value (.9) of our model shows a tight fit to the regression line.

```
##
## Calls:
## m2: lm(formula = I(contribution_sum) ~ I(n) + Zips, data = NYDonors_Obama.Contributions_by_Co
unty)
##
## ===================================
##   (Intercept)      -12330.054***
##                     (1505.767)
##   I(n)                207.591***
##                       (1.453)
##   Zips                 0.150*
##                       (0.071)
## -----------------------------------
##   R-squared                 0.9
##   adj. R-squared            0.9
##   sigma                 41745.9
##   F                     10242.2
##   p                         0.0
##   Log-likelihood        -21534.6
##   Deviance        3107274513914.9
##   AIC                   43077.2
##   BIC                   43099.2
##   N                      1786
## ===================================
```

```
##        fit      lwr      upr
## 1 438292.9 356144.3 520441.5
```

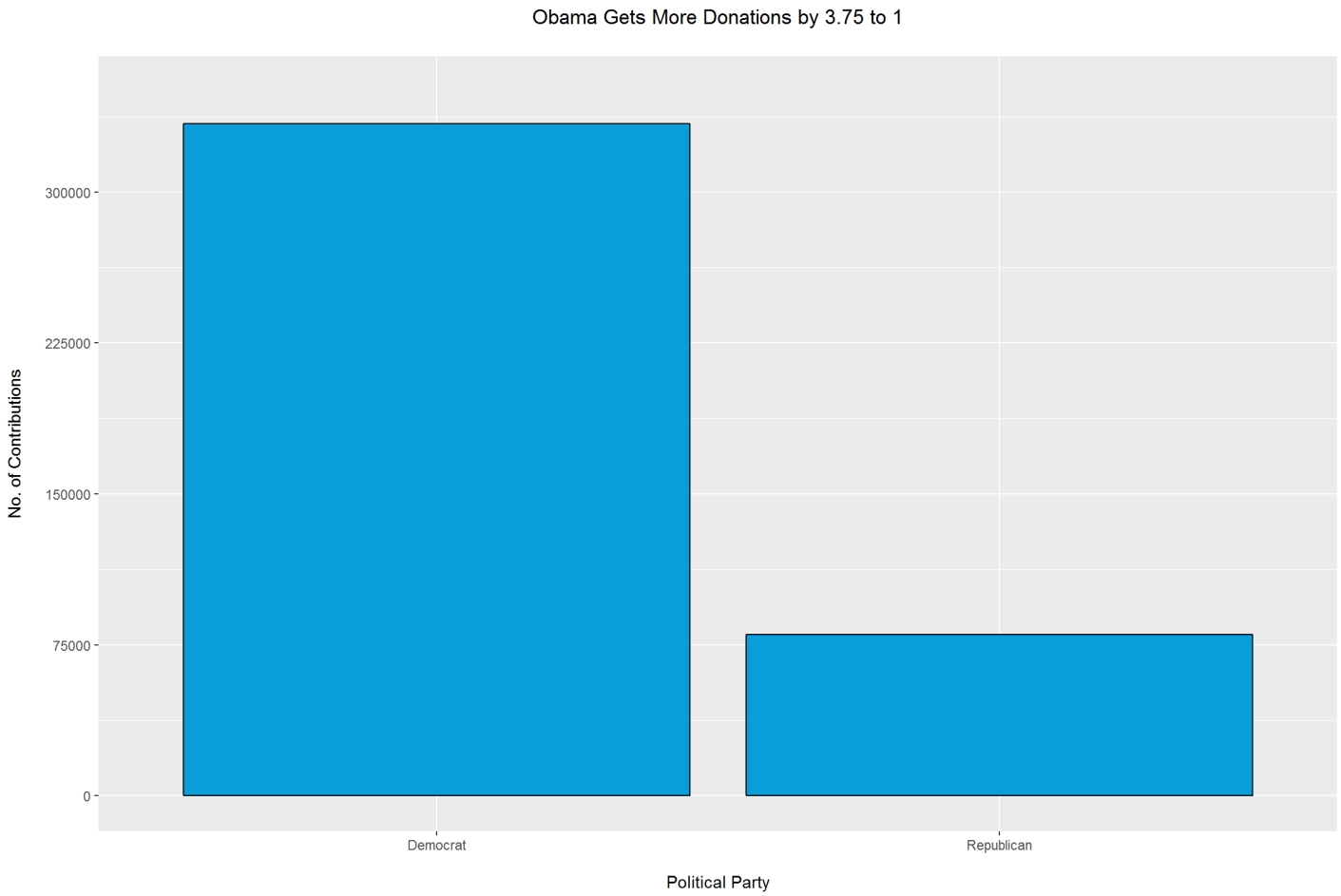
Part Four – Three key plots

Unforunately, the model does not accurately predict the contribution sum. Entering the number of contributors (2160) and the Zip code (10036)into the model does not predict the contribution sum for that zip code which is outside the confidence interval.

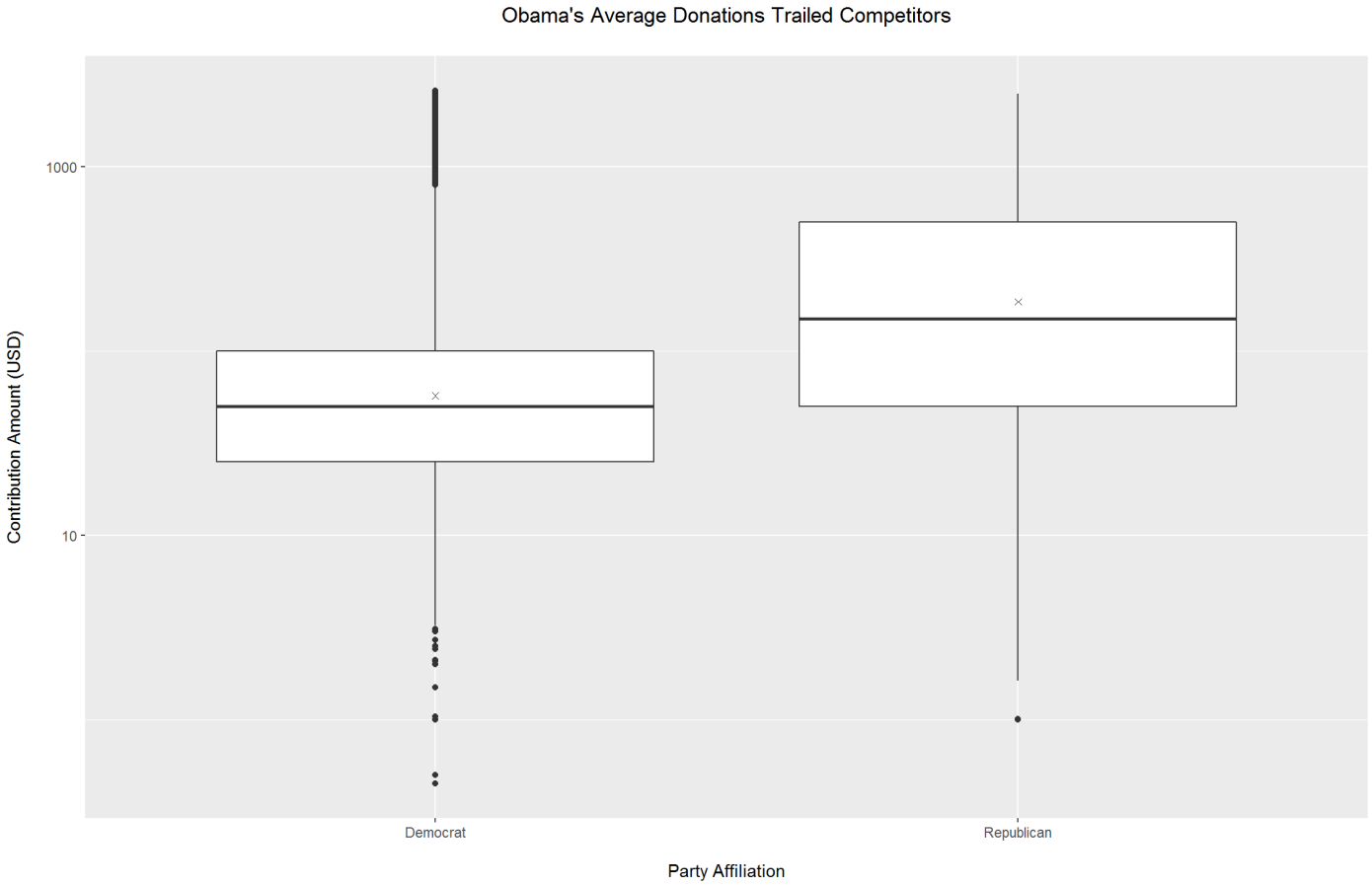
Final Summary and Three Key Plots

Barack Obama

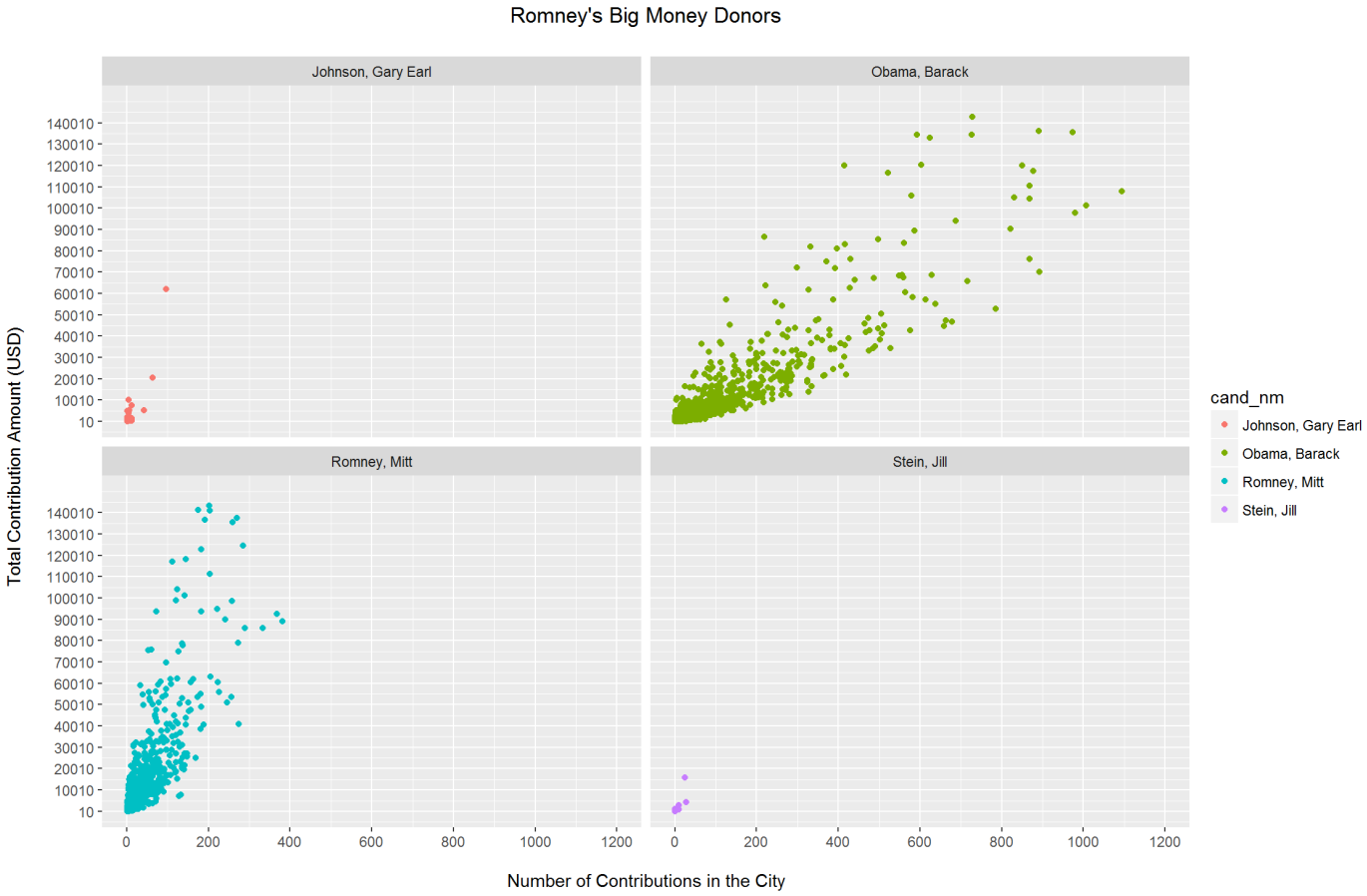
1. Even though Barack Obama was the only Democrat in the 2012 presidential race and there were 11 Republicans in the same race, he still had more donors than all other Republicans combined by a ratio of 3.75 to 1.



2. However Obama's lead in total donors did not raise his average donations. Republicans had a higher donor average while Obama relied on an extreme number of outlier donations.



3. Mitt Romney, Obama's chief rival for the presidency, and the other 10 Republicans vying for the presidency got just over 80,000 donors, but Romney's average donations were higher than Obama's which means that Romney relied upon bigger donations from a smaller donation base. Mitt Romney received $41,603,237 in donations from NY state from an entire donor base of about 80,000. Barack Obama, the lone Democrat in the race, received $51,381,561 from a donor base of over 300,000.



Summary

The most consistent point that these plots show is that a Republican nominee for president can get nearly the same amount of donations from NY state than the Democratic nominee even with far less donors. The Republicans rely on a smaller donor base than do the Democrats to achieve these results. The second most important summary is that party affiliation drives donations. Minor political parties like the Libertarians and Green parties were unable to come close to the donors nor donation amounts of the Republicans and Democrats.

Final Reflections

The dataset was not very robust in that there was little to no details about the donors. For example, there was no demographic data about the donors, age, income, gender, education level, etc. Without such data, the analysis was very shallow. The course provided me with the tools to contruct the plots. However, I was disappointed with not being able to get the prediction model to work. A future project would be to try to get the model to work by gathering more data about the donors.