Hillary 2016 Contributions

Philip Seifi

July 20, 2015

```
## 'data.frame':    38286 obs. of  10 variables:
##  $ contbr_nm        : Factor w/ 22963 levels "AAB, LISA","AAKER, LINDA",..:
13884 13884 13884 13884 13884 13061 7253 4787 11884 5654 ...
##  $ contbr_city      : Factor w/ 2942 levels "","ABERDEEN",..: 673 673 673 67
3 673 673 673 69 69 673 ...
##  $ contbr_st        : Factor w/ 56 levels "AE","AK","AL",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ contbr_zip       : int  96240593 96240593 96240593 96240593 96240593 9265
1001 97160101 91800002 98394900 92651001 ...
##  $ contbr_employer  : Factor w/ 10441 levels "","10-4 SYSTEMS, INC.",..: 230
8 1 2308 2308 2308 8107 9532 9494 9518 9518 ...
##  $ contbr_occupation: Factor w/ 3927 levels "","AC SERVICE TECHNICIAN",..: 2
560 1 2560 2560 2560 576 100 2057 961 961 ...
##  $ contb_receipt_amt: num  25 -100 100 50 50 2700 2700 210 250 2700 ...
##  $ contb_receipt_dt : Factor w/ 83 levels "01/05/2015","01/06/2015",..: 10 5
7 56 58 67 73 82 62 7 73 ...
##  $ receipt_desc     : Factor w/ 2 levels "","Refund": 1 2 1 1 1 1 1 1 1
1 ...
##  $ memo_text        : Factor w/ 24 levels "","*","* EARMARKED CONTRIBUTION:
SEE BELOW",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
##                contbr_nm           contbr_city      contbr_st
##  MUKOKA, DENIS      :   35   NEW YORK     : 3367   CA     : 7325
##  MARSOLAIS, PATRICIA:   29   WASHINGTON   : 2116   NY     : 5895
##  HANNON, STEPHANIE  :   26   LOS ANGELES  : 1138   FL     : 2771
##  CHRISTY, S. M.     :   25   SAN FRANCISCO: 1063   TX     : 2169
##  TAMARO, LANA       :   24   CHICAGO      :  819   DC     : 2108
##  GOOD, CHARLES      :   23   BROOKLYN     :  548   IL     : 1607
##  (Other)            :38124   (Other)      :29235   (Other):16411
##    contbr_zip                    contbr_employer
##  Min.   :      734   N/A                  : 5918
##  1st Qu.:117544909   SELF-EMPLOYED        : 4962
##  Median :334015726   RETIRED              : 1668
##  Mean   :456816095   INFORMATION REQUESTED: 1438
##  3rd Qu.:837093342                        :  518
##  Max.   :998245337   NOT EMPLOYED         :  351
##  NA's   :2           (Other)              :23431
##              contbr_occupation contb_receipt_amt     contb_receipt_dt
##  RETIRED              : 4094   Min.   :-20000.00   12/04/2015: 2620
##  ATTORNEY             : 3446   1st Qu.:    72.58   30/06/2015: 2198
##  INFORMATION REQUESTED: 1545   Median :   250.00   13/04/2015: 1311
##  CONSULTANT           : 1345   Mean   :   994.44   29/06/2015: 1209
##  HOMEMAKER            :  970   3rd Qu.:  2700.00   12/06/2015:  896
##  LAWYER               :  960   Max.   : 20000.00   23/06/2015:  760
##  (Other)              :25926                       (Other)   :29292
##  receipt_desc                                  memo_text
##        :37864                                        :37721
##  Refund:  422   * EARMARKED CONTRIBUTION: SEE BELOW  :  460
##                 * IN-KIND: CATERING, FOOD & BEVERAGES:   24
##                 *                                    :   22
##                 INSUFFICIENT FUNDS                   :   17
##                 * IN-KIND: FOOD & BEVERAGES          :    9
##                 (Other)                              :   33
```

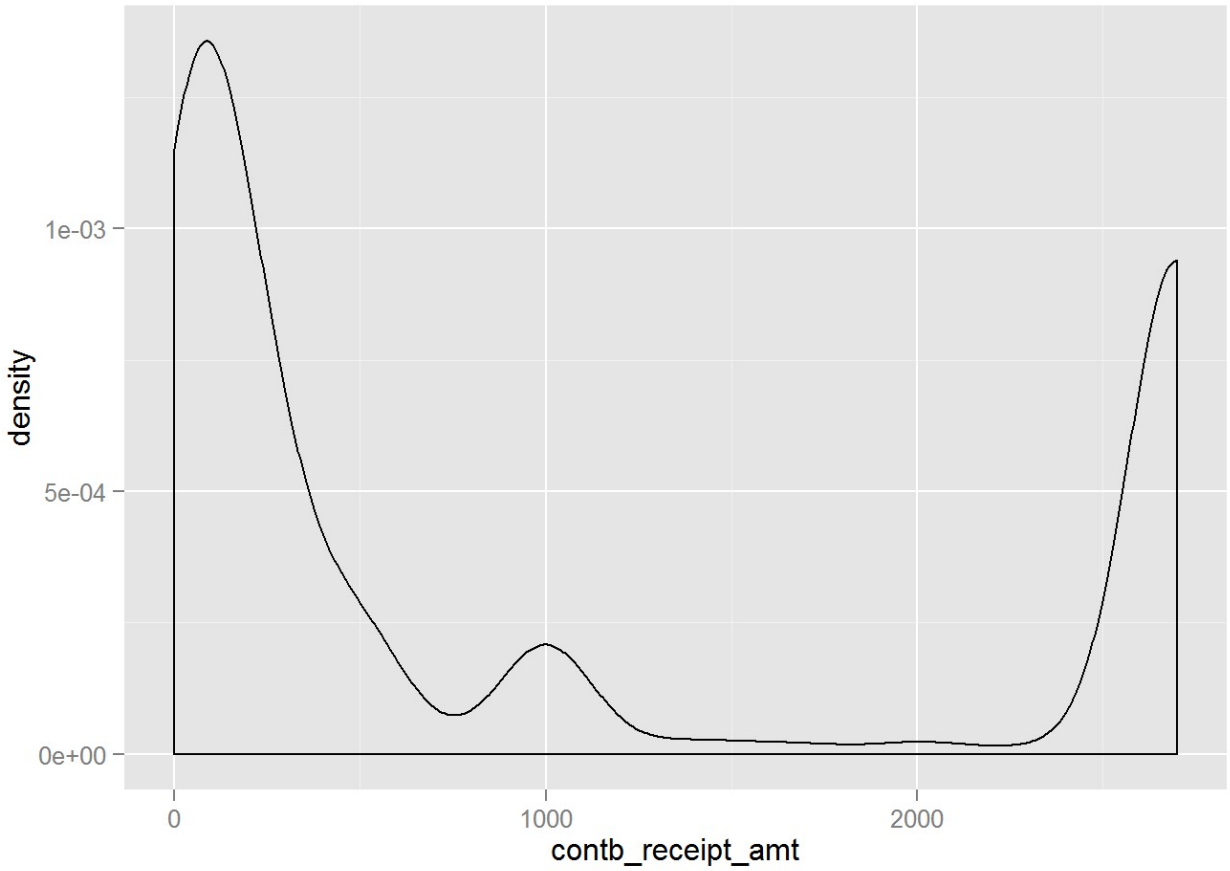
The highest contributing state is California. The highest contributing city is New York.

The median contribution amount is $250.

I am omitting all contributions above the $2,700 limit (Source: http://www.fec.gov/pages/fecrecord/2015/february/contriblimits20152016.shtml (http://www.fec.gov/pages/fecrecord/2015/february/contriblimits20152016.shtml)) as they break the Federal Election Campaign Act and thus have or will be refunded.

To get us started, let's look at a density plot on contribution amounts, omitting refunds (negative contributions).

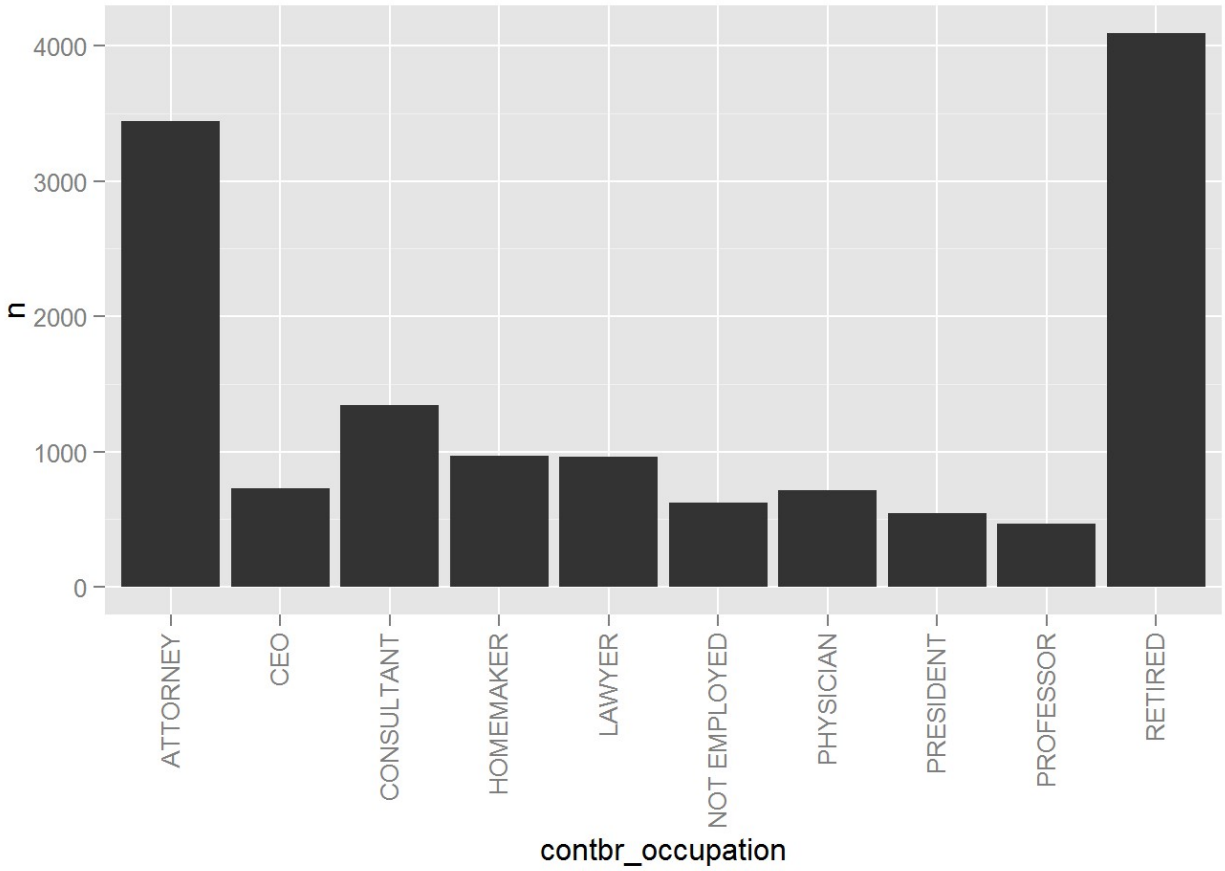
Density



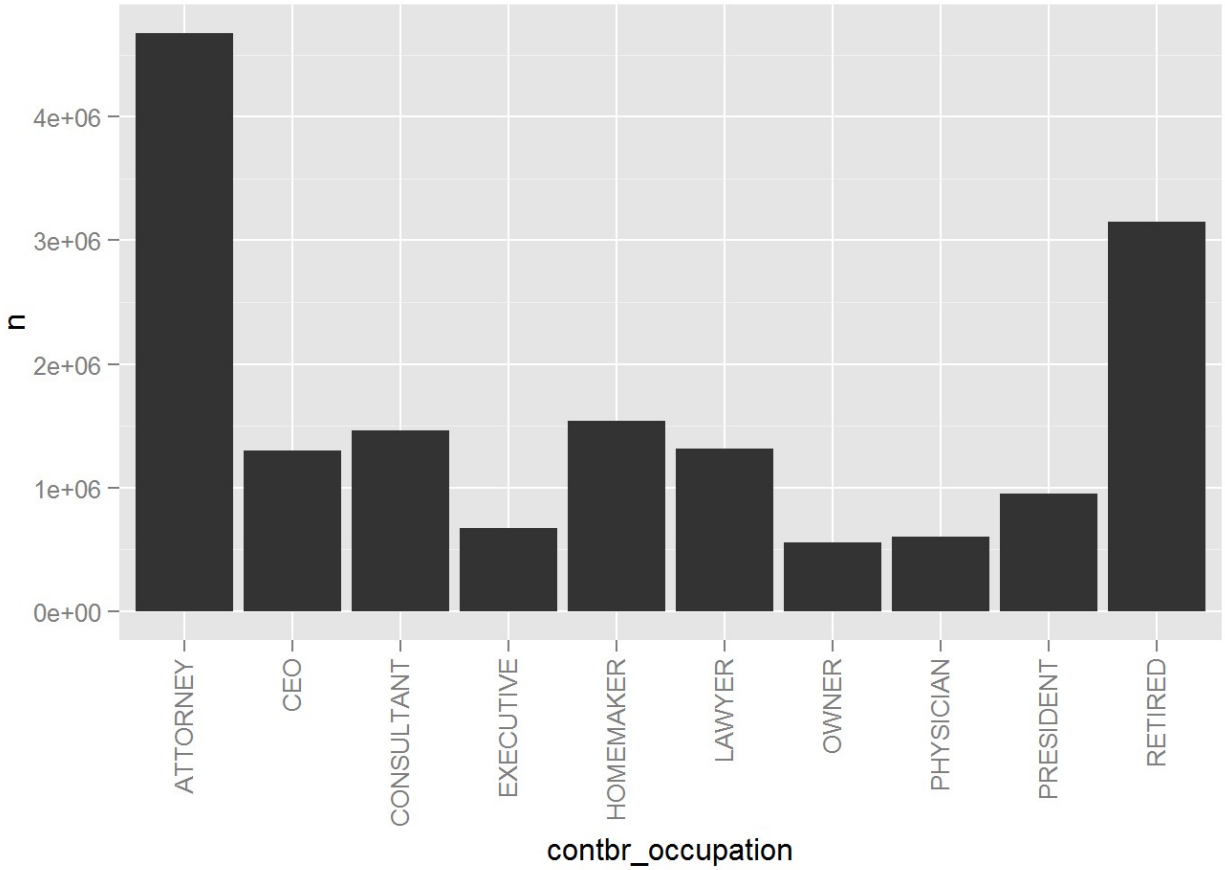
We can observe a U-shaped curve indicating that most contributions fall in one of two groups: small contributions of around $200, and large contributions at the maximum FEC limit of $2,700. There is also a small bump around the $1,000 mark, possibly an artificial, round contribution limit people decide upon in their minds.

My hunch is that the two groups contain very different demographics, so let's go on and explore contributions by occupation.

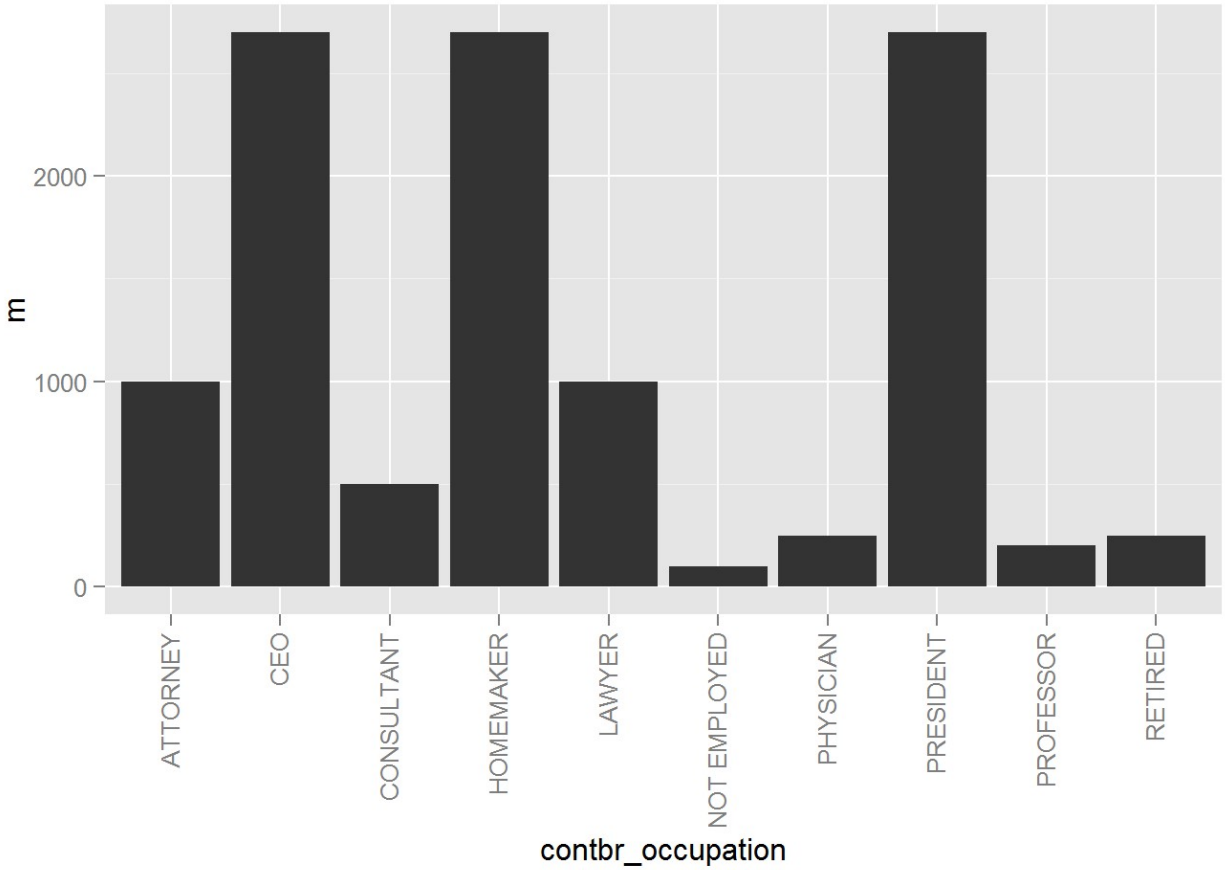
Occupations



Most of the contributors so far are from retired supporters, closely followed by attorneys, and in a distant third, consultants.

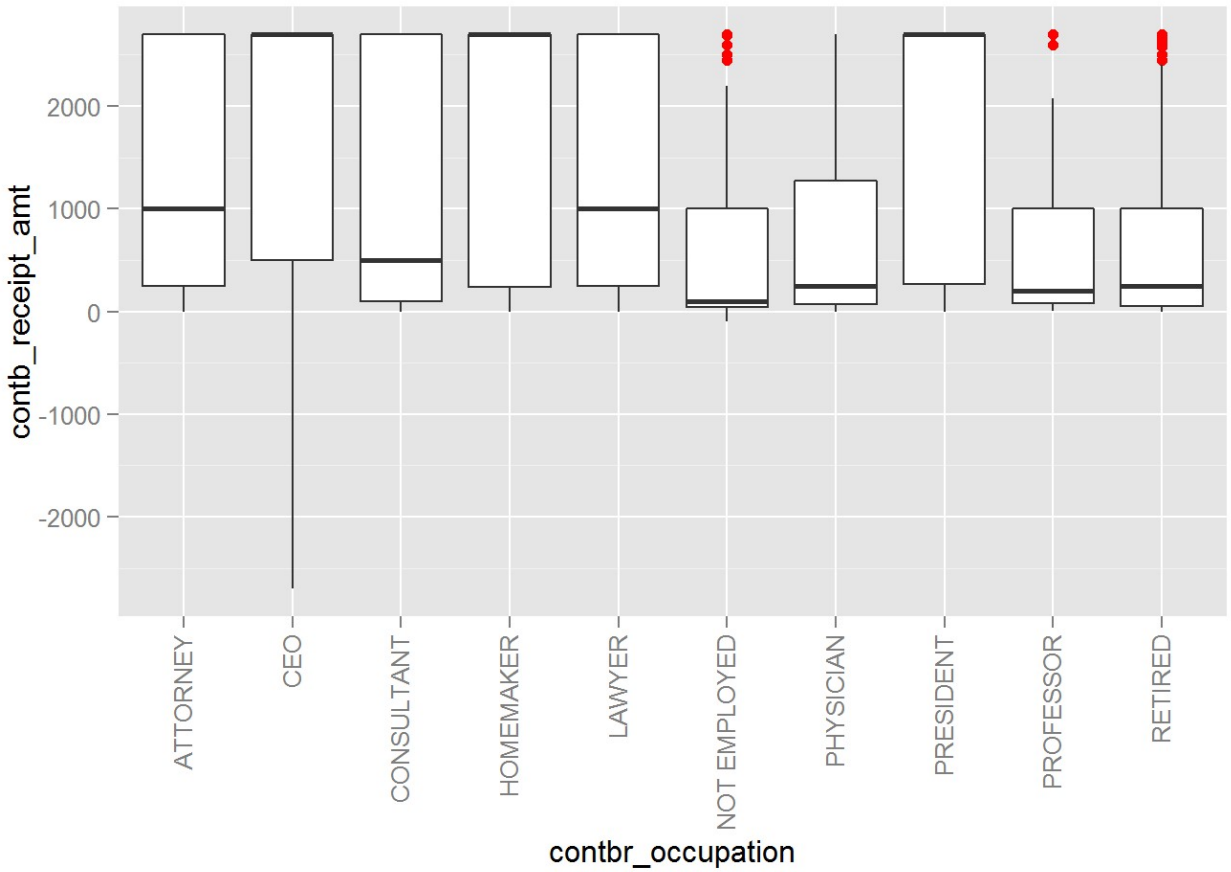


The picture is different if we plot total contribution amounts rather than the number of contributions. Now it's the attorneys who take the first place, followed by retired, homemakers and consultants. This suggests that consultants and the retired make lower average contributions than attorneys and homemakers. Let's confirm this assumption…

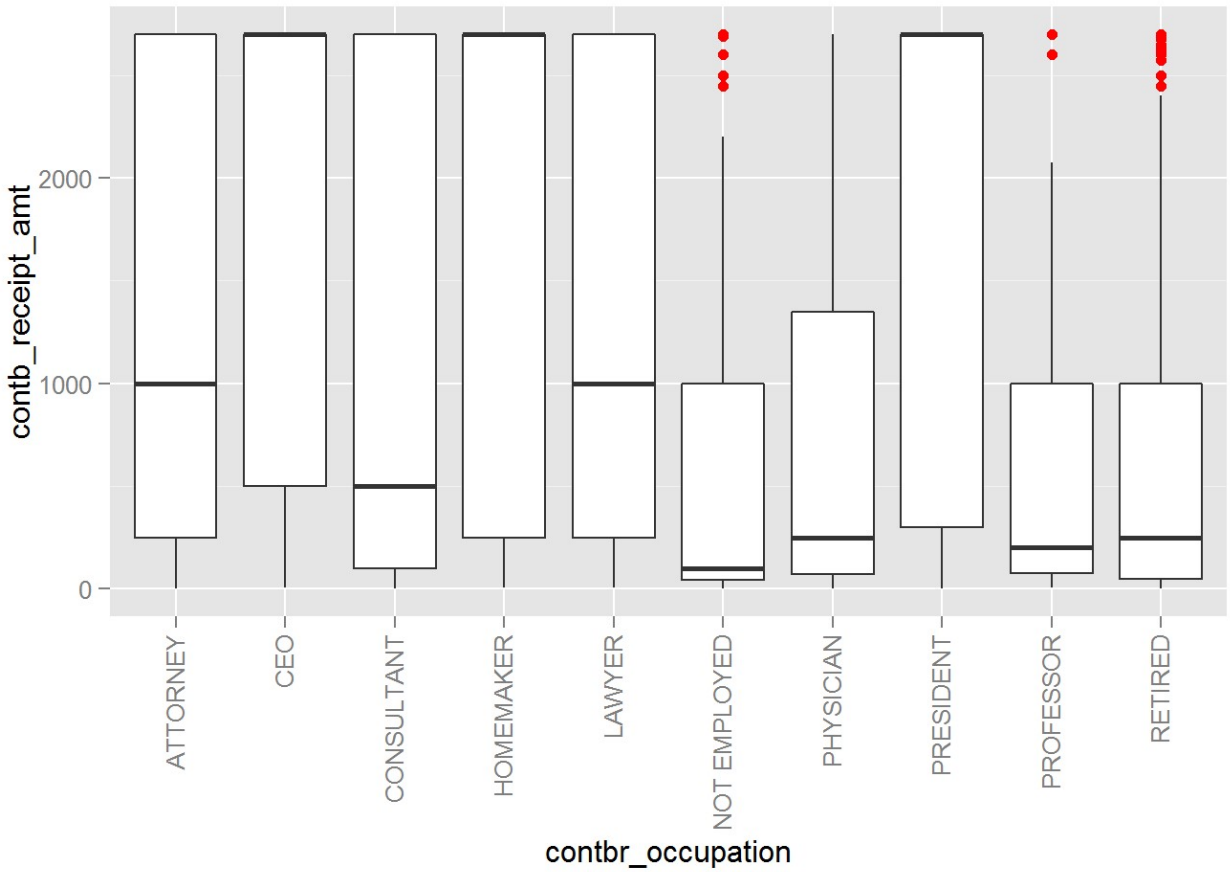


Indeed, it appears that although their total contributions are significant, the median contributions of the retired and consultants are relatively low. Physicians make surprisingly minor contributions given their presumably above-average income. Homemakers, on the other hand, have some of the highest median contributions among top 10 contibutors in this campaign, up there with CEOs and Presidents.

There's clearly a lot of interesting insight to be found here, and a box plot might give us a better overview of this data…



Interesting, but refunds (negative contributions) aren't particularly helpful in this case and make the plot less readable. Let's omitt them and try again…



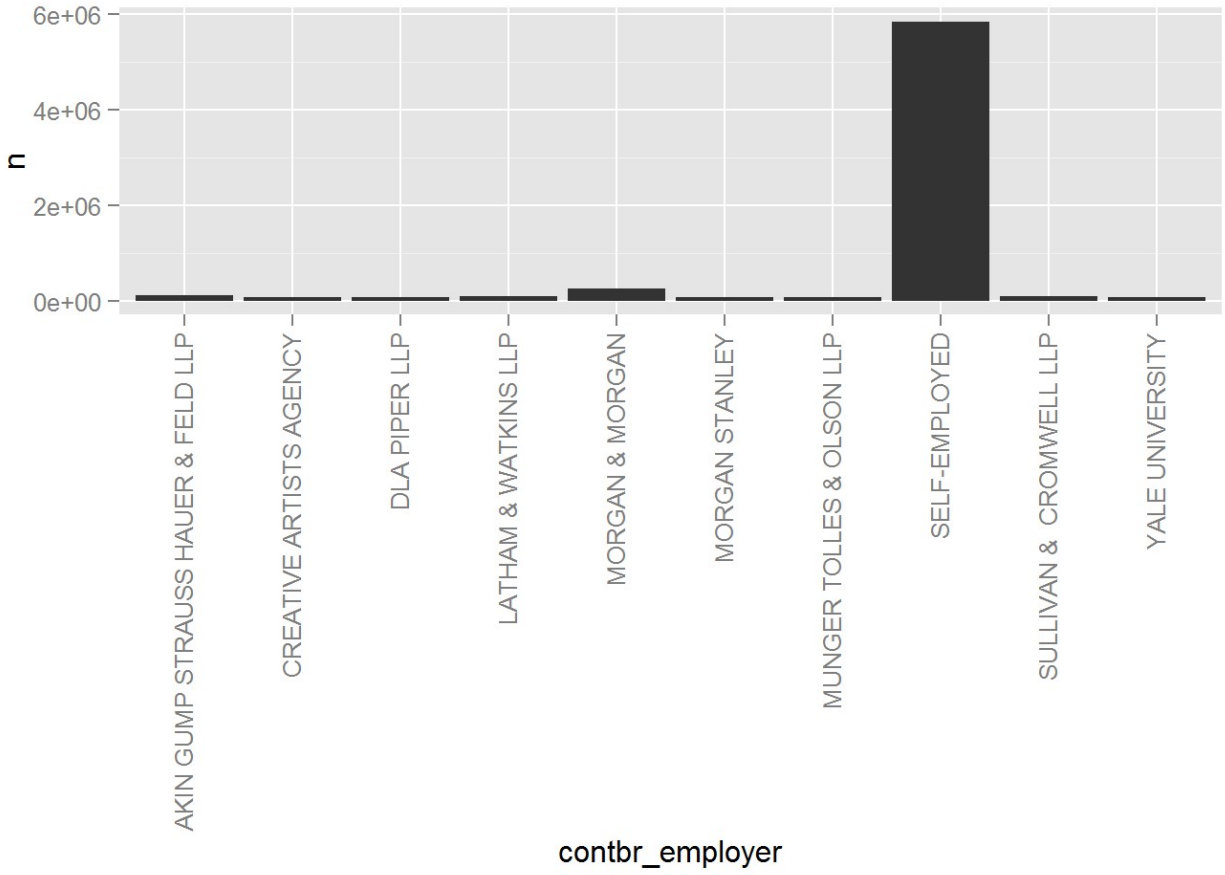
We can see that the average contributions are indeed quite low for the retired, unemployed, professors and physicians. There is significant variation in these groups, however, with some supporters making comparable contributions to those from the other occupations.

We can also observe a lot of variation among attorneys and presidents, some of them contributing up to 3x more than their respective medians. This is likely indicative of the U-shaped distribution of salaries among attorneys (http://qph.is.quoracdn.net/main-qimg-0a0d8f37efe16a83e4f1208aea3b1988?convert_to_webp=true).

Finally, there is equally some negative variation among CEOs. My presumption is that the lower contributions are made by CEOs of startups and other SMEs, who might not have disposable income comparable to that of corporate CEOs classified under the same group.

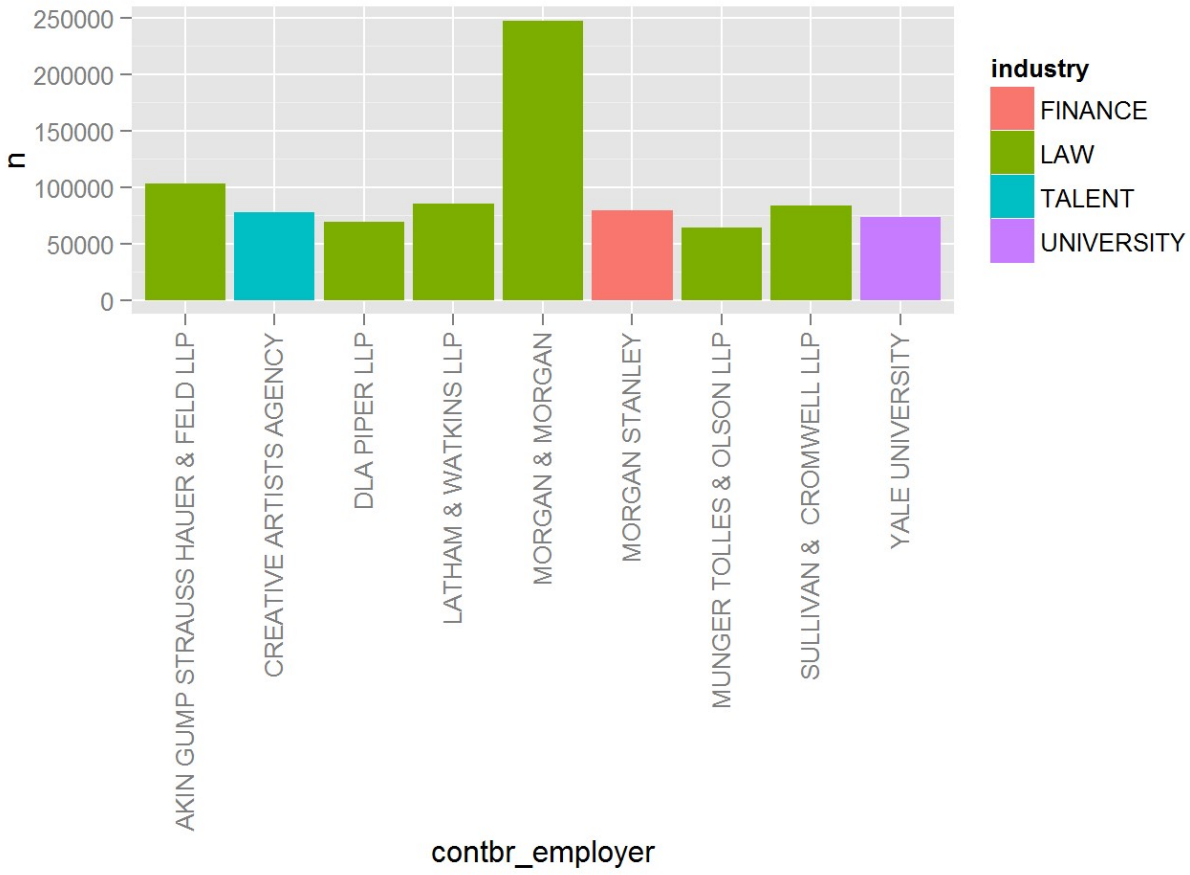
The differences among occupations are certainly interesting, but some of the largest donors in political campaigns tend to be corporations. Let's take a look contribution amounts by employer. I will omitt the retired, and the unemployed, because they vastly outweigh individual companies and organizations, and we've already had a look at these categories earlier.

Employers



The self-employed are clearly, by far, the largest contributors to Hillary Clinton's campaign. This is surprising, given that the self-employed traditionally vote GOP. Of course, without analysis of all contributions in this electoral cycle, it is impossible to tell whether an even larger number of self-employed Americans contribute to Republican candidates, as one would expect, given that Republicans are roughly 50% more likely to be self-employed (Fried, pp. 104-5, 125.)

Let's redraw the same graph omitting the self-employed to have a better look at the individual institutions. I manually colour-coded each of the top employers by their industry.



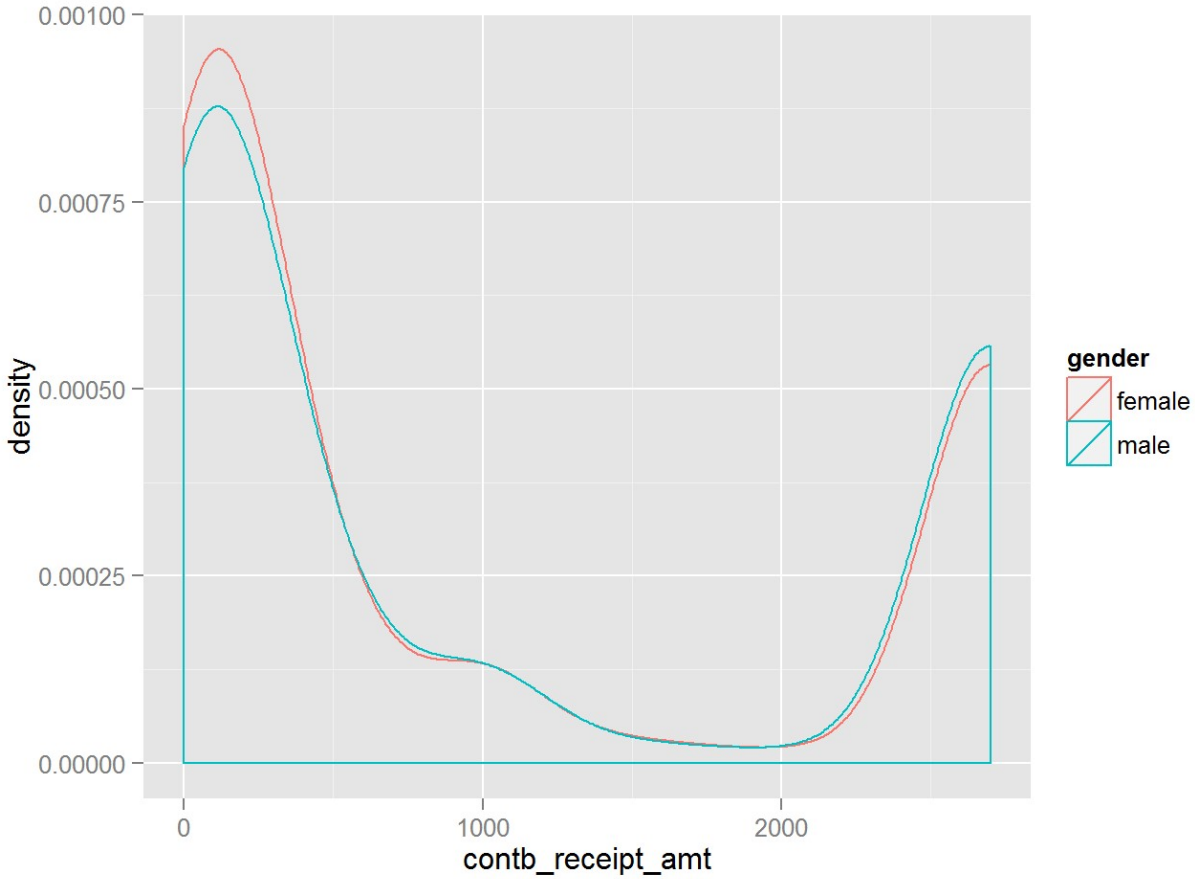
This chart makes it especially clear where much of the money comes from. The largest donor is Morgan & Morgan, a consumer protection and personal injury law firm. Most of the other donors are also legal firms, with the exception of one talent agency, one financial institution and one Ivy League university.

Given that some of the top donors in the 2012 Obama campaign (Source: https://www.opensecrets.org/pres12/ (https://www.opensecrets.org/pres12/)) were major universities, I would have expected to see more purple in the list. I'm also surprised to see Morgan Stanley, one of the top donors in the 2012 Mitt Romney campaign (ibid.). Of course, here again, it is impossible to say whether Morgan Stanley doesn't have an equal, or even larger stake in the campaigns of GOP candidates without analysis of the entire dataset.

Gender

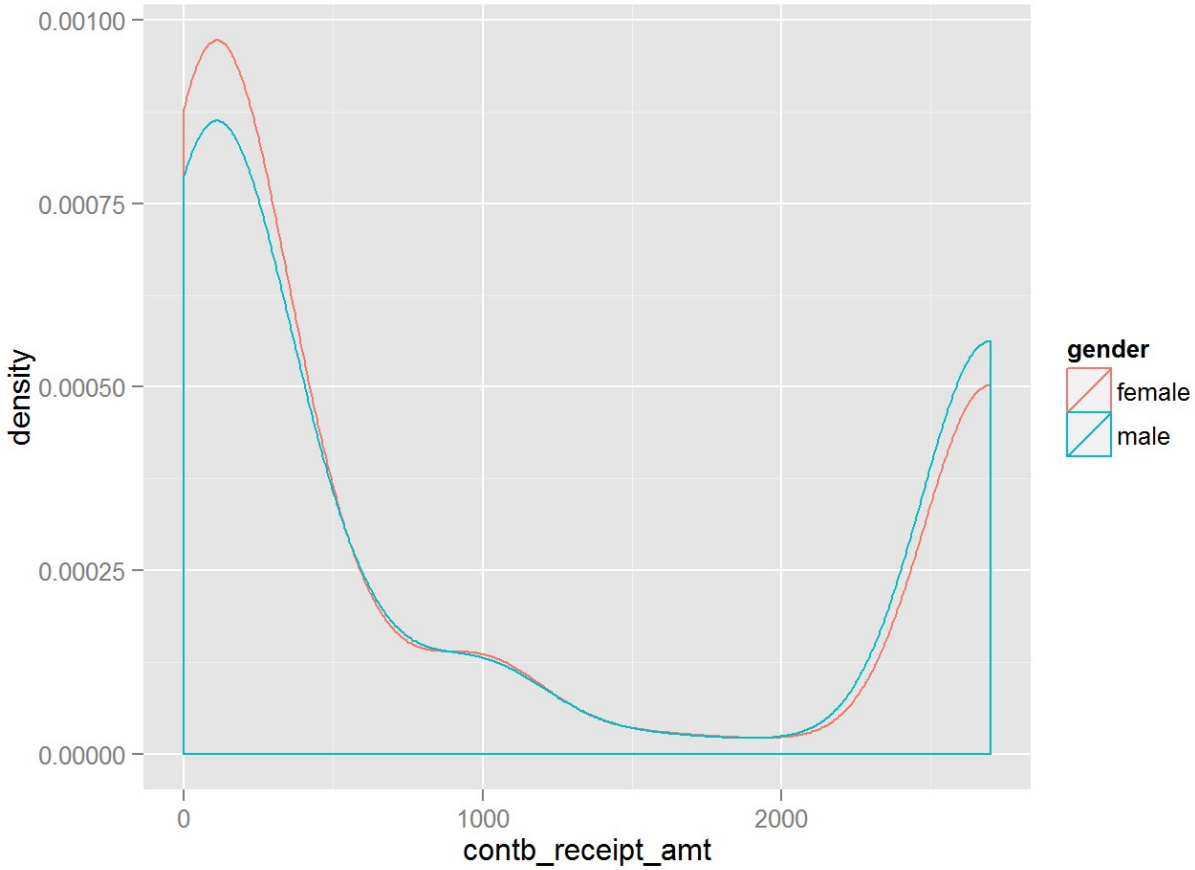
It would be most interesting to analyze contributions by gender, given that the Hillary 2016 campaign actively targets female voters. The FEC dataset does not contain contributor genders, but this information can be prediceted based on first names of the contributors, using the `gender` package and the Social Security Administration gender data.

First let's look at a density distribution by gender.



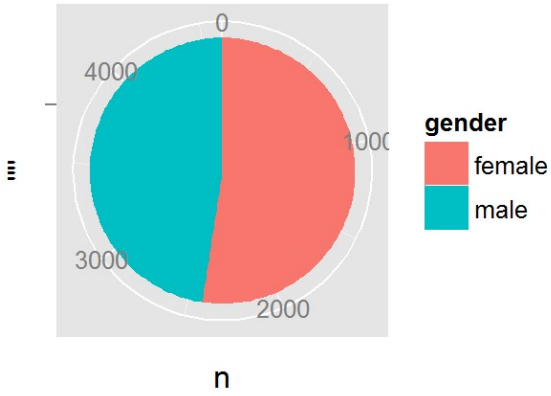
It appears that women tend to make lower contributions that men. This is likely the effect of lower salaries and greater number of homemakers among women (i.g, stay-at-home wifes).

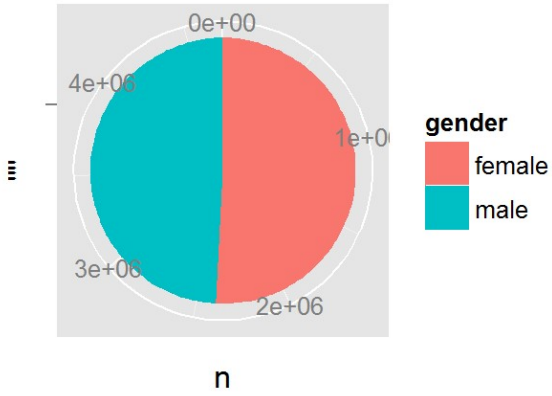
Let’s test this assumption by generating the same plot, but omitting homemakers from the dataset.



My assumptiom was clearly wrong. The new subset in fact shows an even lower number of high-amount contributions among women than before.

What about total contributions by gender? A couple of pie charts can be a good way to visualize this…



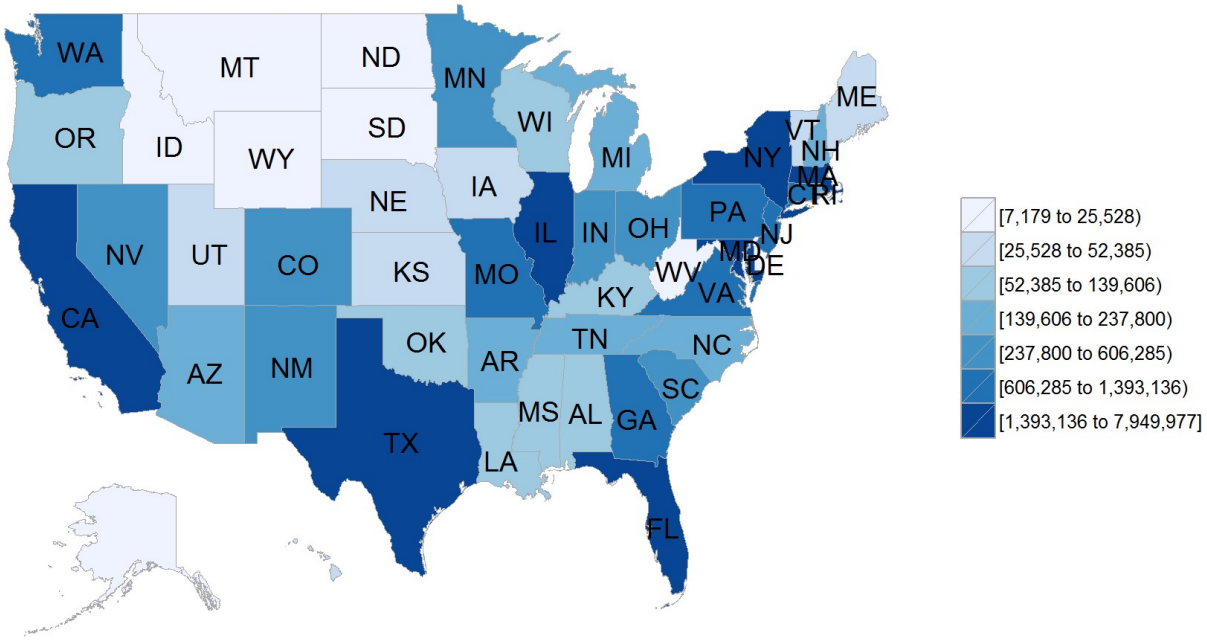


The first pie chart is the number of contributions by gender. The second, the total value by gender. It appears that although the number of contributions is larger for female supporters, the total value is comparable.

This confirms the findings from the density plot that women tend to make smaller contributions than men, but also alligns with Hillary's appeal to female voters.

Next, let's make a choropleth map of states by total contributions to see where the money flows from…

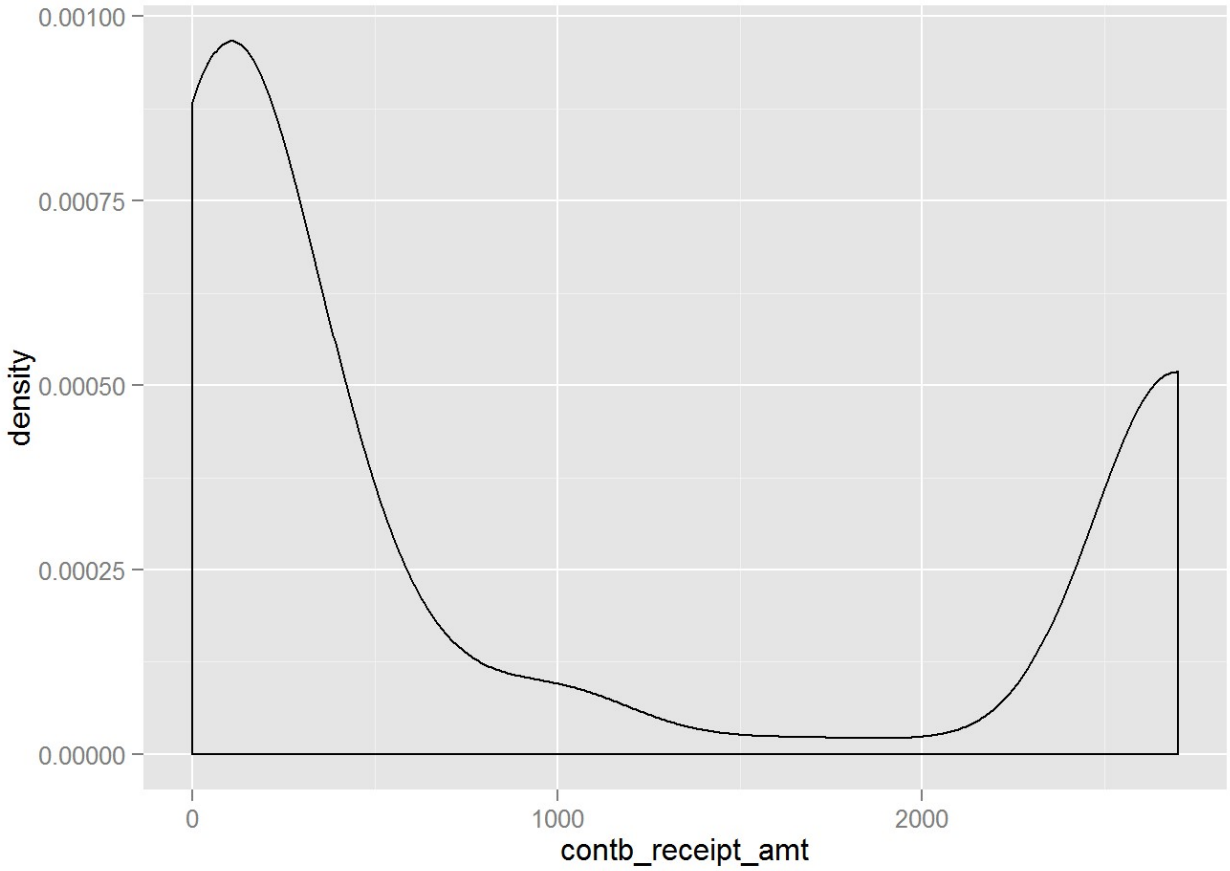
Origin regions



Again, some surprising results worth investigating. Why is Texas, a predominantly Republican state, one of the top sources of contribution for Hillary?

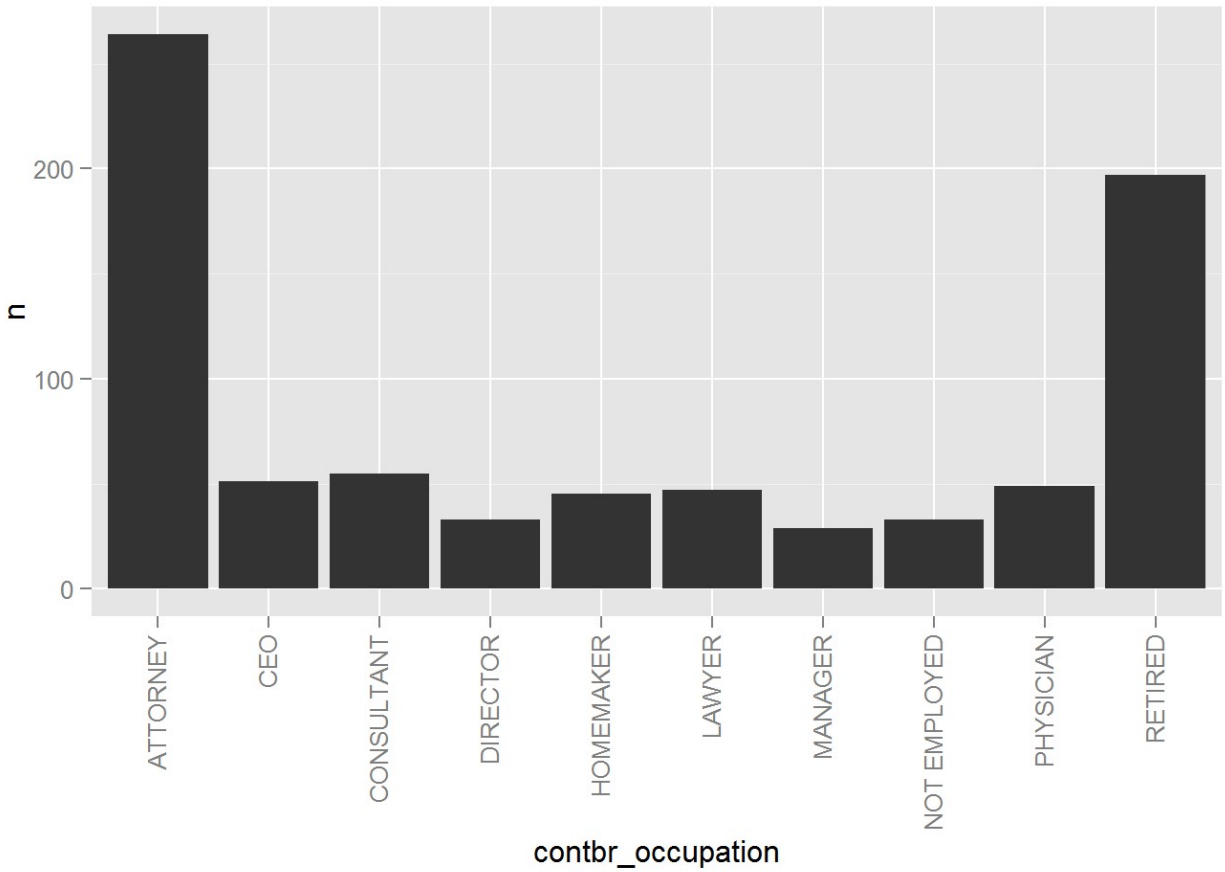
Let’s start with a density plot on contribution amounts, omitting refunds (negative contributions), coming from the state of Texas.

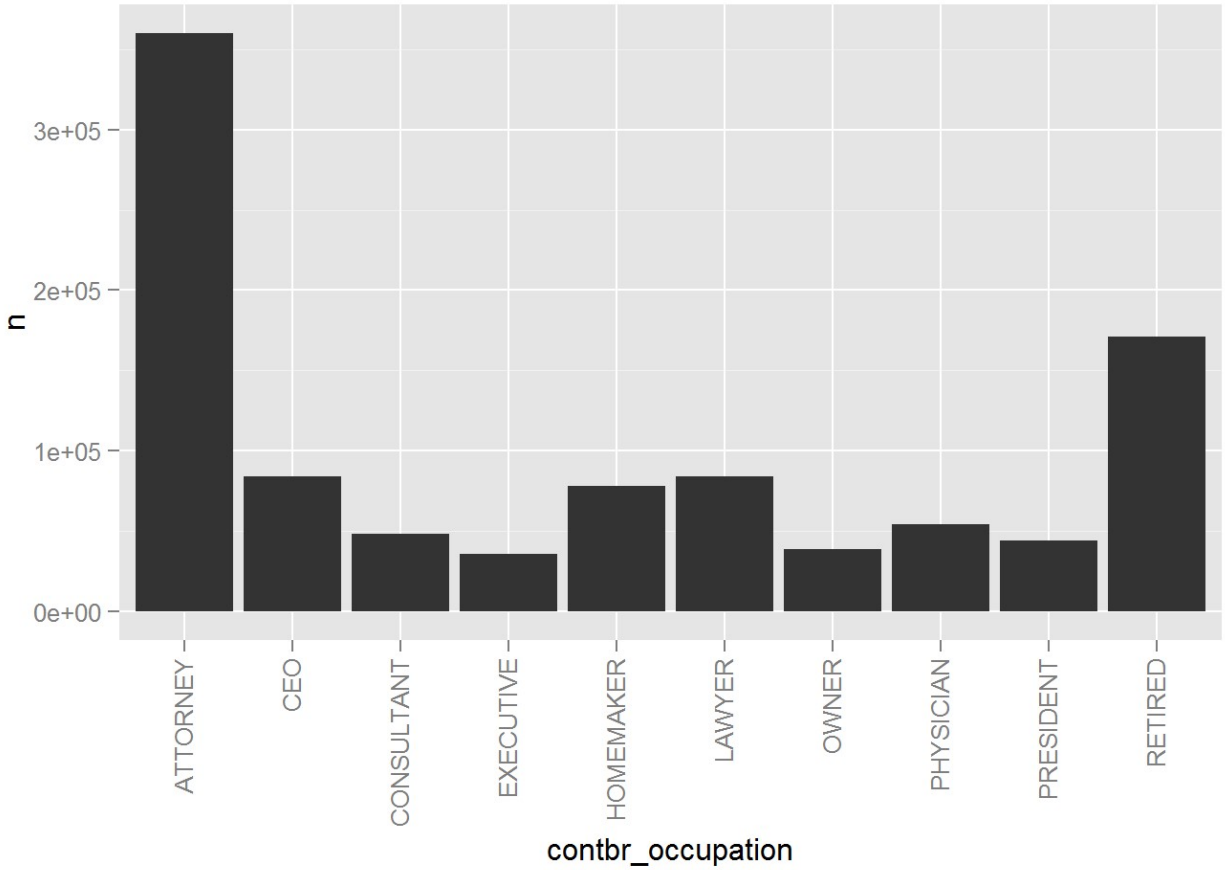
Drilling into Texas



Let’s redraw the occupation bar charts using a subset of contributions coming from Texas…

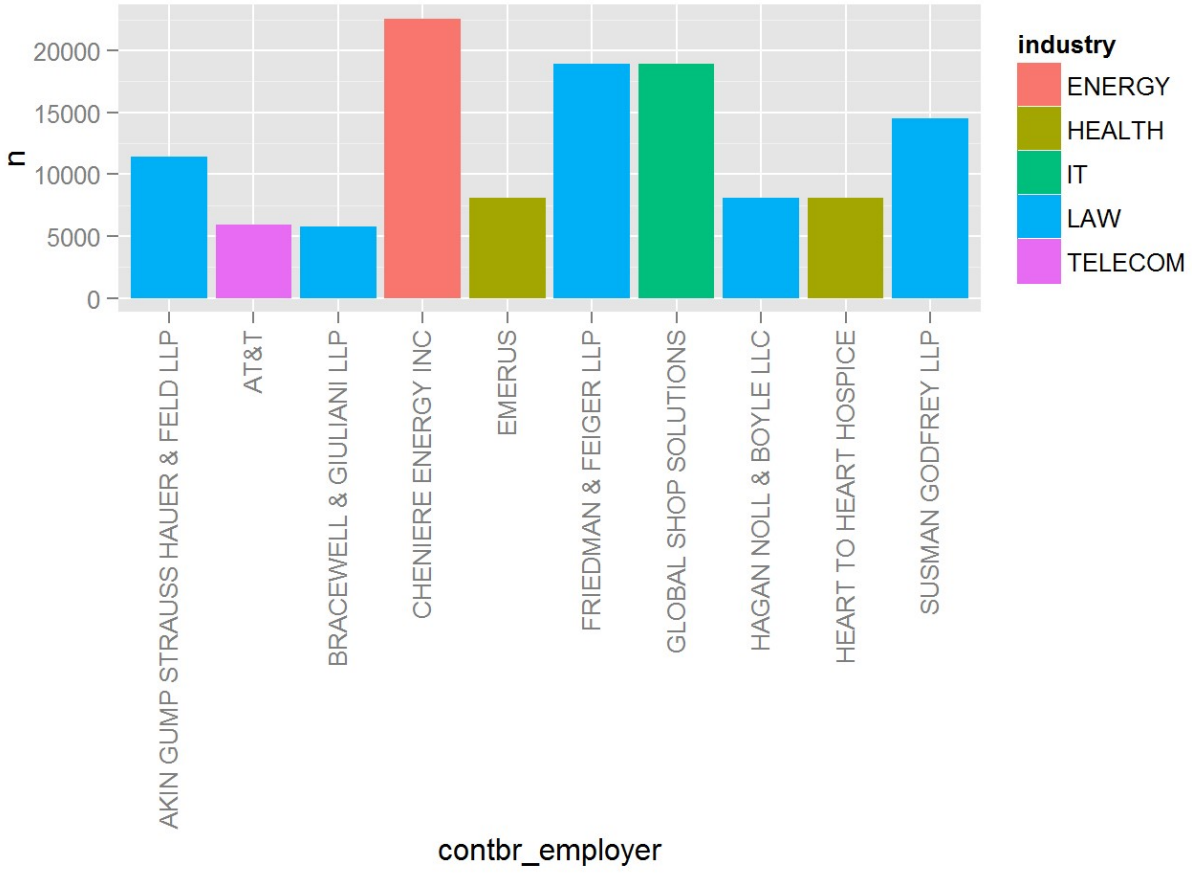
Nothing suprising to see here. The distribution is similar to that of contributions nationwide, with the exception of a less noticeable bump at the $1,000 level and a slightly lower peak at the $2,700 FEC limit.


n
200
100
0
ATTORNEY
CEO
CONSULTANT
DIRECTOR
HOMEMAKER
LAWYER
MANAGER
NOT EMPLOYED
PHYSICIAN
RETIRED
contbr_occupation



Attorneys appear to be an even more dominant occupation in Texas than in the rest of the country, both in number and total amount of contributions.

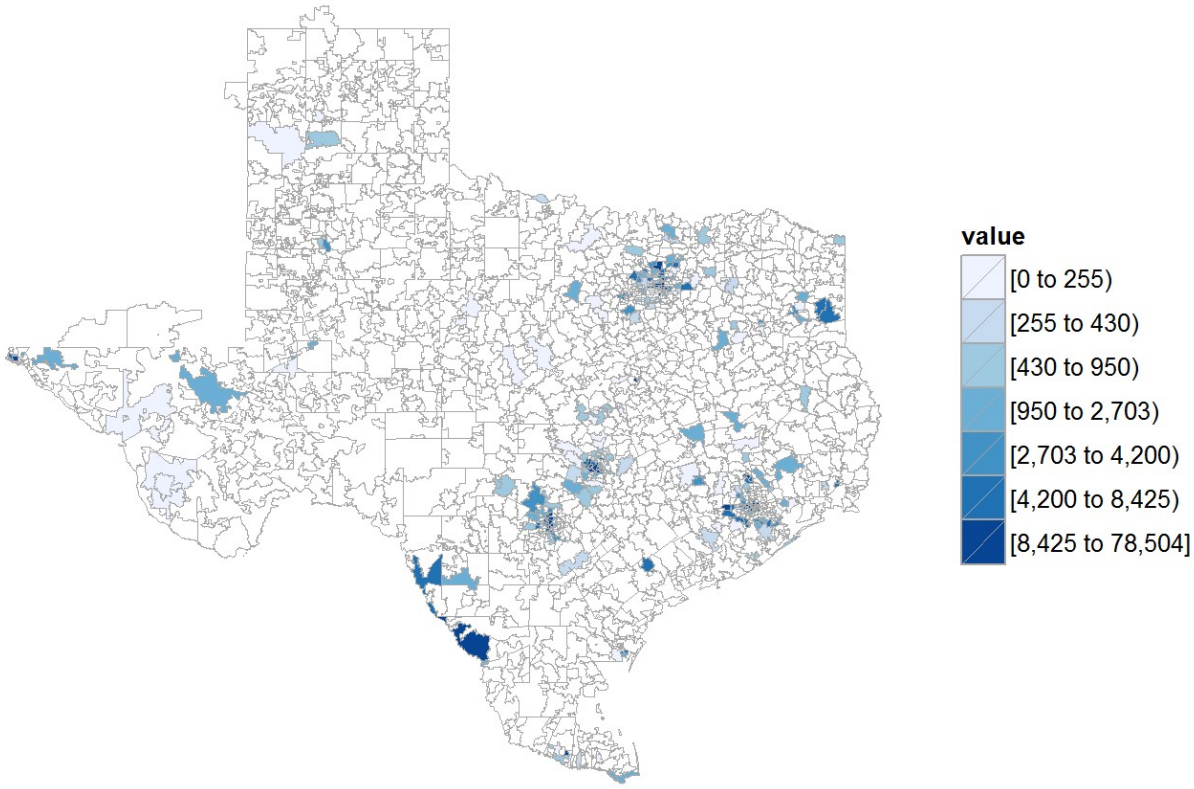
Let’s see if a plot of employers confirms this finding.



Surprisingly, after all, law firms are not the dominants sources of contributions in Texas, compared to the national totals. Could this be caused by a smaller average size of law firms in the state?

It is also interesting to note that the top contributor is an energy company, a sector which is not represented among the top 10 nationwide.

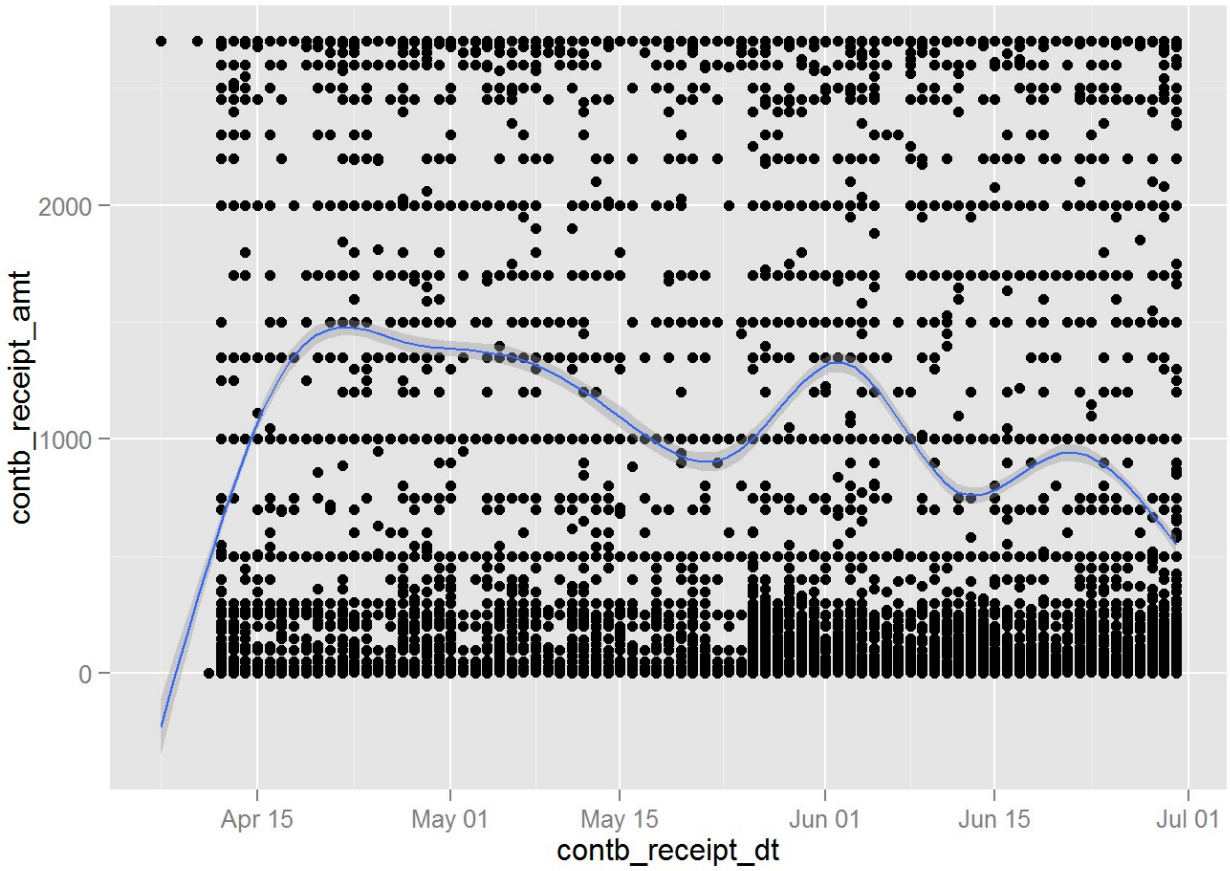
Let’s redraw the choropleth, but this time zoom onto the state of Texas and investigate individual ZIP codes.



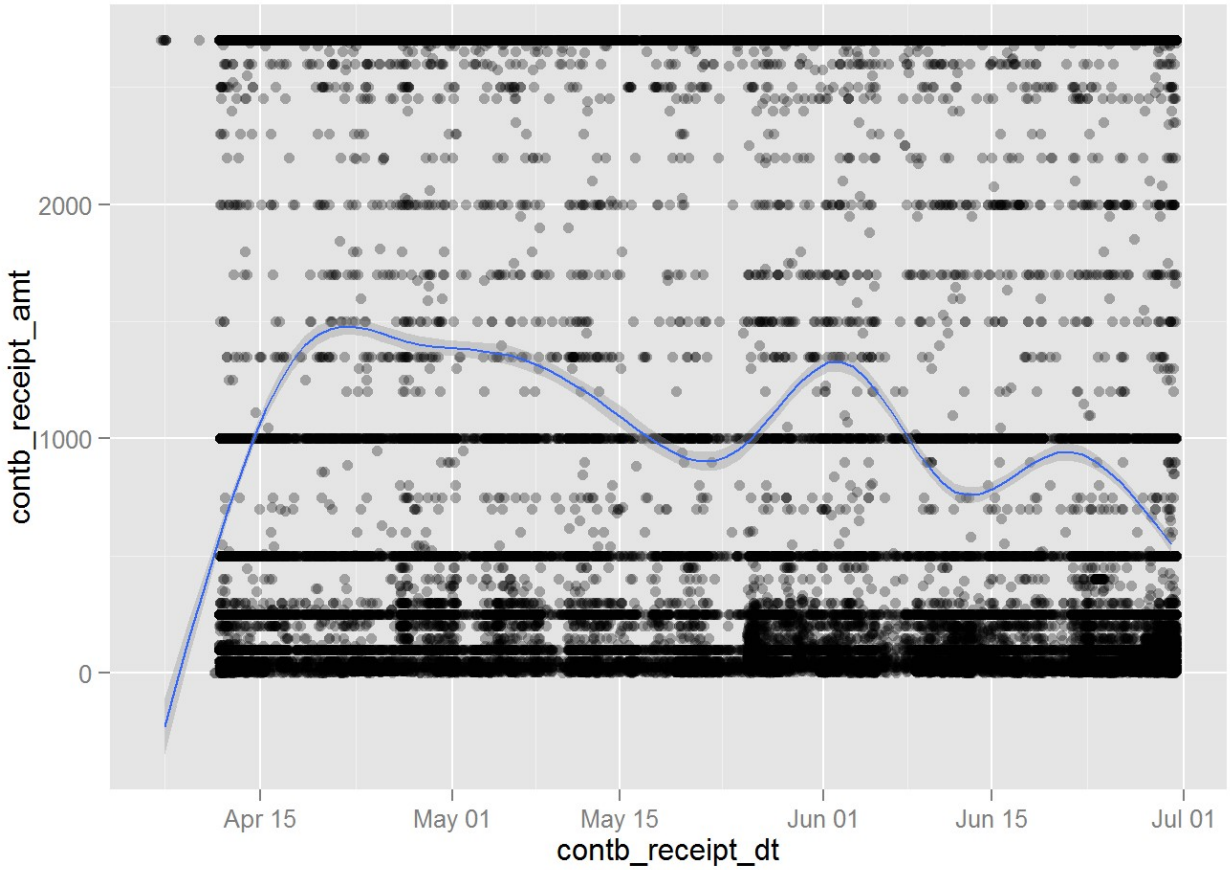
The more fine-grained choropleth of contributions from Texas shows that most contributions come from a few specific regions of the state. This is not surprisingly, but any further conclusions require a better knowledge of the demographics of Texas than I possess, and ideally data on political affiliations of the individual regions.

Contributions from Texas are certainly worth looking into, but now, let’s take a look at how contributions change throughout the campaign.

Time series



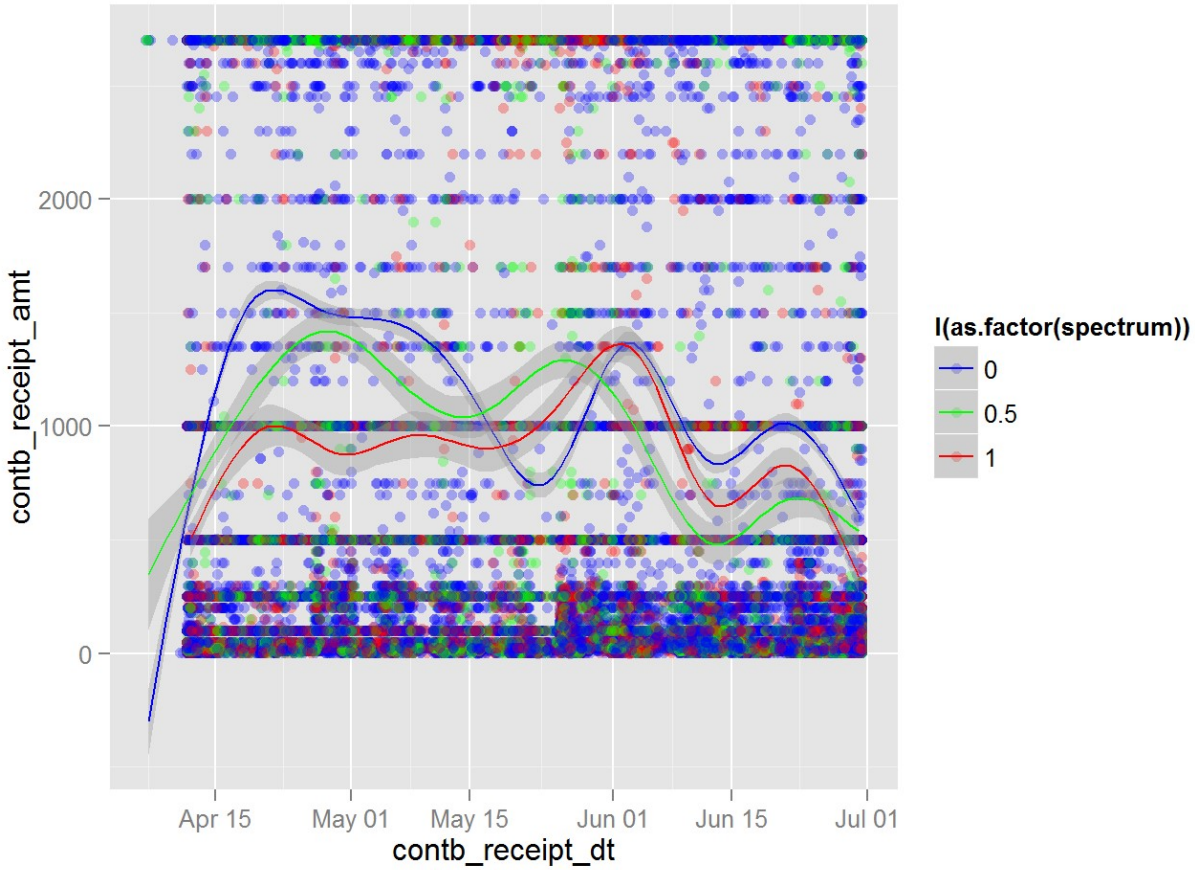
There's some overplotting in the lower levels of the contributions amount gradient. The data also gives the false perception that all contributions are made at the same time of each day. Let's add a little bit of transparency and position jitter to avoid falling to false conclusions.



We can see that outide of the occasional outlier, the contribution amounts appear to be remarkably consistent. A an increase of small contributions can be observed in the last few months. This could be seasonal, or a natural evolution as the electoral cycle progresses.

Of note are also frequent contributions at the $250, $500, $1,000 levels, likely a psychological effect of people choosing round values for their donations.

Let’s add some colours… red for Republican states, blue for Democrat states, and green for swing states.

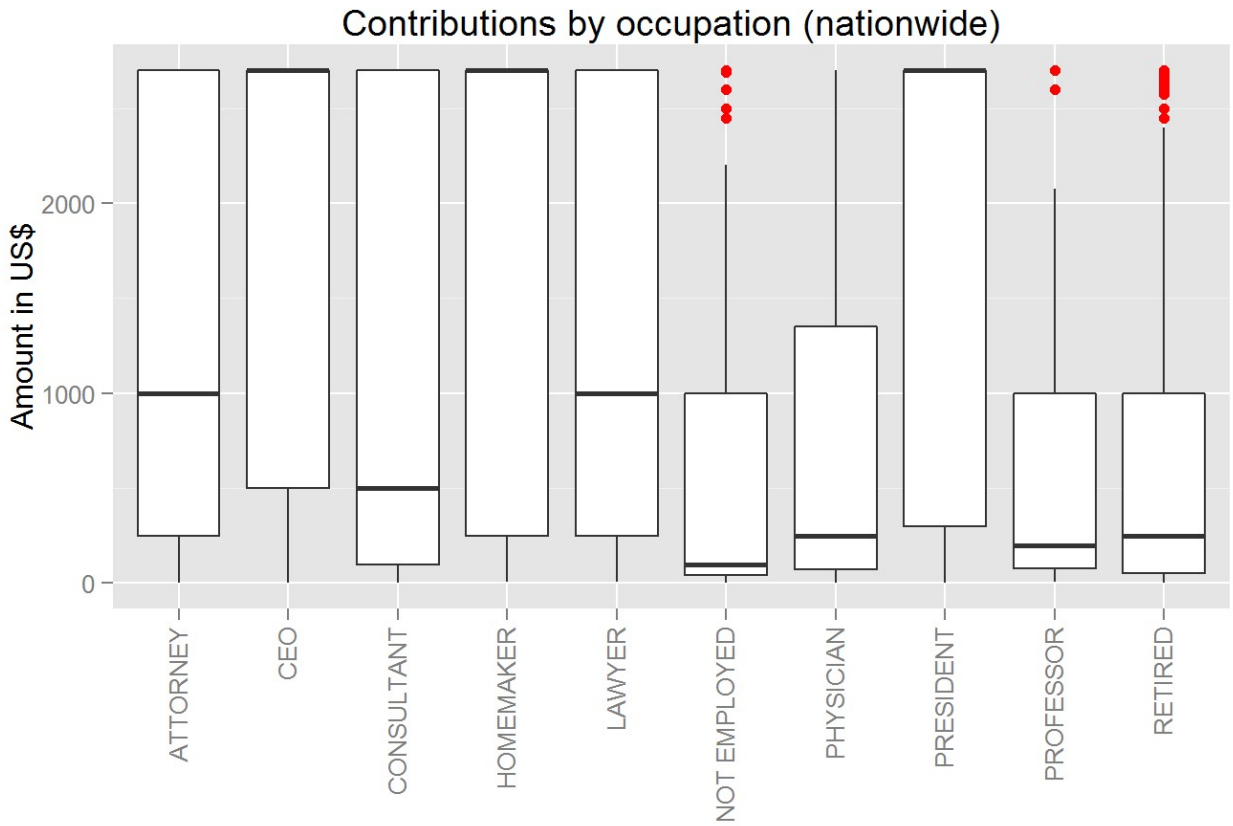


Although the median contributions are consistently higher from Democrat states, the difference isn't very large. Of note are the last two weeks of May, when median Republican and swing state contributions were higher than those from Democrat states.

Final plots and summary

Exploratory Data Analysis of the Hillary Rodham Clinton 2016 campaign from April to July 2016 identified several avenues for further investigation.

1. Contributions by occupation (nationwide)



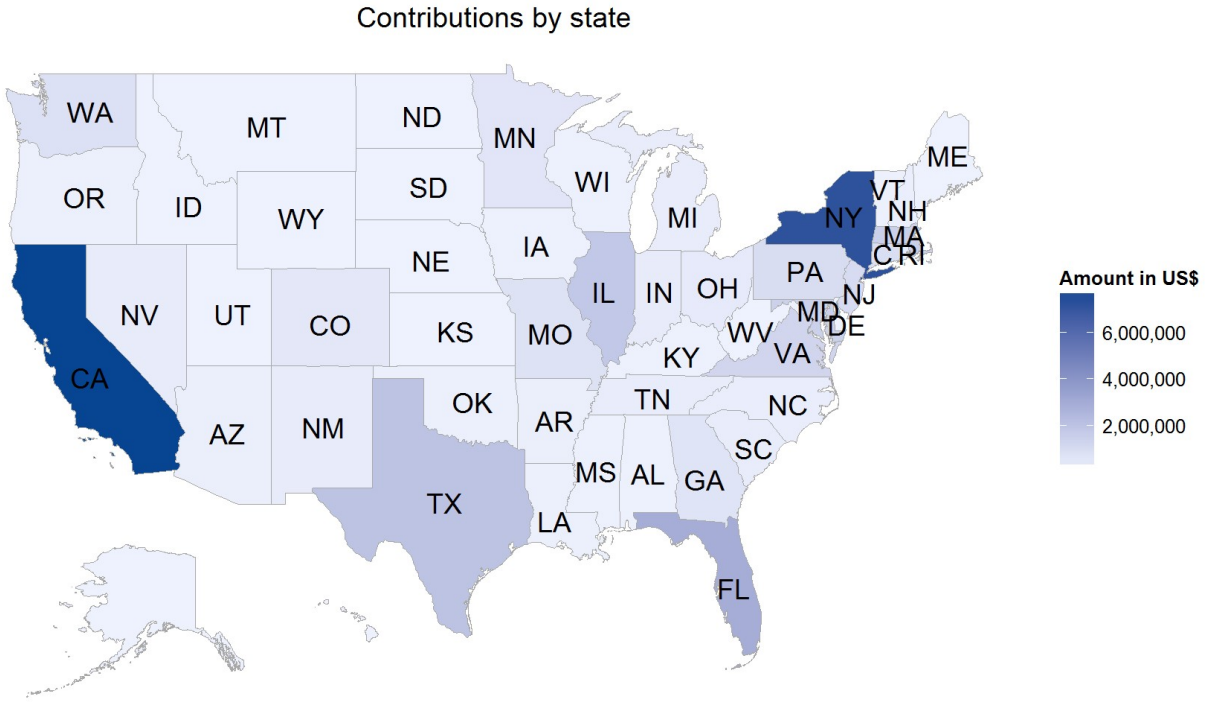
The median contributions to the campaign differ substantially across occupations.

The retired, unemployed, professors and physicians all have very low median contributions, with some outliers in the $2000-2700 range.

CEOs, homemakers and presidents have a median equal to the contribution limit. Attorneys, lawyers, and consultants do not make high median contributions but there is very significant variance. This is likely because some of these contributors are employed by major law and consulting firms who may even endorse and cover their contributions, whereas others are part of a smaller partnership and cover the contribution out of pocket.

The variance may also be indicative of the U-shaped distribution of salaries in the legal professions (http://qph.is.quoracdn.net/main-qimg-0a0d8f37efe16a83e4f1208aea3b1988?convert_to_webp=true).

2. Contributions by state



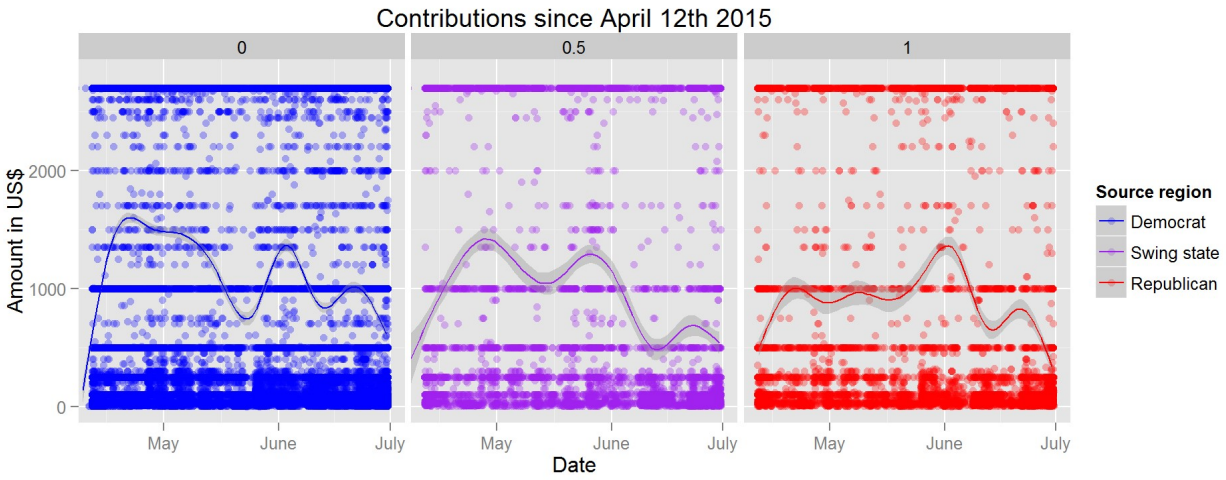
The choropleth map of total amount of contributions by state is largely unsurprising, with one exception.

Why is Texas, a predominantly Republican state, one of the top sources of contributions for Hillary Clinton?

I tried to investigate this question by exploring contributions by employer using a Texas subset of the data, but discovered no easy explanation.

Generating a more granualar choropleth of contributions from Texas, using the ZIP code information showed that most contributions come from a few select areas of the state. This is not surprising, but any further conclusions require a better knowledge of the demographics of Texas than I possess, and ideally data on political affiliations of the individual regions.

3. Contribution time series



Plotting contributions from April 12th, when Hillary Clinton announced her candidacy, till July 1stm, faceted by origin states' political affiliations gives us yet another perspective on the data.

Total contributions appear to be remarkably consistent. A an increase in small contributions (<$250) can be observed in the last few months. This could be seasonal, or a natural evolution as the electoral cycle progresses.

Although the median contributions are consistently higher from Democrat states, the difference isnât very large. Of note are the last two weeks of May, when median Republican and swing state contributions were higher than those from Democrat states.

Reflection

The Hillary 2016 campaign contributions dataset contains over 38,000 entrie from April till July 2015. Although the elections are still far away, analysis of existing data can give us some indication of most important contribution sources, and allows us to predict the demographics Hillary Clinton should approach as her fundraising progresses.

I started by understanding the individual variables by studying the official dataset format (ftp://ftp.fec.gov/FEC/Presidential_Map/2016/DATA_DICTIONARIES/CONTRIBUTOR_FORMAT.txt), then conducting basic descriptive statistical analysis.

To begin with, I chose to omitt all entries above the $2700 contribution limit (http://www.fec.gov/pages/fecrecord/2015/february/contriblimits20152016.shtml) as they break the Federal Election Campaign Act (http://www.fec.gov/law/feca/feca.pdf) and thus were or will be refunded.

I then used these findings and my domain expertise to explore contribution numbers and amounts across a number of variables including occupation, employer, and state.

Some findings where not surprising. Contributions to Hillary's campaign tend to come from California and the city of New York, they tend to be made by the retired, and law firms are some of the major employers behind these contributors.

Other findings were more surprising, however:

- Physicians, who are among the best paid in the country (http://www.bls.gov/oes/current/oes_nat.htm), make very small contributions compared to other occupations.
- Texas, a predominantly Republican state (https://en.wikipedia.org/wiki/Politics_of_Texas), is among the top 10 sources of contributions to Hillary Clinton's campaign.
- The self-employed are clearly, by far, the largest contributors to Hillary Clinton's campaign. This is surprising, given that Republicans are roughly 50% more likely to be self-employed (Fried, pp. 104â“5, 125.). Of course, without analysis of all contributions in this electoral cycle, it is impossible to tell whether an even larger number of self-employed Americans contribute to GOP candidates.

It would be most interesting to explore this dataset further by combining it with data from Hillary Clinton's 2008 presidential campaign. Have Hillary's contribution sources changed since 8 years ago? Could we predict current and future contributors using data from previous elections? These are all important questions that could be studied using a combined dataset.