Hillary 2016 Contributions

Philip Seifi

July 20, 2015

```
#
# DATA SOURCE: http://fec.gov/disclosurep/PDownload.do
# FORMAT: ftp://ftp.fec.gov/FEC/Presidential_Map/2016/DATA_DICTIONARIES/CONTRIBUTOR_FORMAT.txt
# ---------------------------------------------------------------
#
# CMTE_ID           COMMITTEE ID                            S - skipped
# CAND_ID           CANDIDATE ID                            S - skipped
# CAND_NM           CANDIDATE NAME                          S - skipped
# CONTBR_NM         CONTRIBUTOR NAME                        S
# CONTBR_CITY       CONTRIBUTOR CITY                        S
# CONTBR_ST         CONTRIBUTOR STATE                       S
# CONTBR_ZIP        CONTRIBUTOR ZIP CODE                    S
# CONTBR_EMPLOYER   CONTRIBUTOR EMPLOYER                    S
# CONTBR_OCCUPATION CONTRIBUTOR OCCUPATION                  S
# CONTB_RECEIPT_AMT CONTRIBUTION RECEIPT AMOUNT             N
# CONTB_RECEIPT_DT  CONTRIBUTION RECEIPT DATE               D
# RECEIPT_DESC      RECEIPT DESCRIPTION                     S
# MEMO_CD           MEMO CODE                               S - skipped
# MEMO_TEXT         MEMO TEXT                               S
# FORM_TP           FORM TYPE                               S - skipped
# FILE_NUM          FILE NUMBER                             N - skipped
# TRAN_ID           TRANSACTION ID                          S - skipped
# ELECTION_TP       ELECTION TYPE/PRIMARY GENERAL INDICATOR S - skipped
#
hrc_in = read.csv('hillary-contribs.csv', header=TRUE, colClasses=c(rep('NULL', 3), rep(NA, 9), 'NULL', NA, rep('NULL', 4)), strip.white=TRUE)

str(hrc_in)
```

```
## 'data.frame':    38286 obs. of  10 variables:
##  $ contbr_nm        : Factor w/ 22963 levels "AAB, LISA","AAKER, LIND
A",..: 13884 13884 13884 13884 13884 13061 7253 4787 11884 5654 ...
##  $ contbr_city      : Factor w/ 2942 levels "","ABERDEEN",..: 673 673 67
3 673 673 673 673 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 9
2651001 97160101 91800002 98394900 92651001 ...
##  $ contbr_employer  : Factor w/ 10441 levels "","10-4 SYSTEMS, INC.",..:
2308 1 2308 2308 2308 8107 9532 9494 9518 9518 ...
##  $ contbr_occupation: Factor w/ 3927 levels "","AC SERVICE TECHNICIA
N",..: 2560 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",..: 1
0 57 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 CONTRIBUTIO
N: SEE BELOW",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(hrc_in)
```

```
##                 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 $2700 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.

```
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

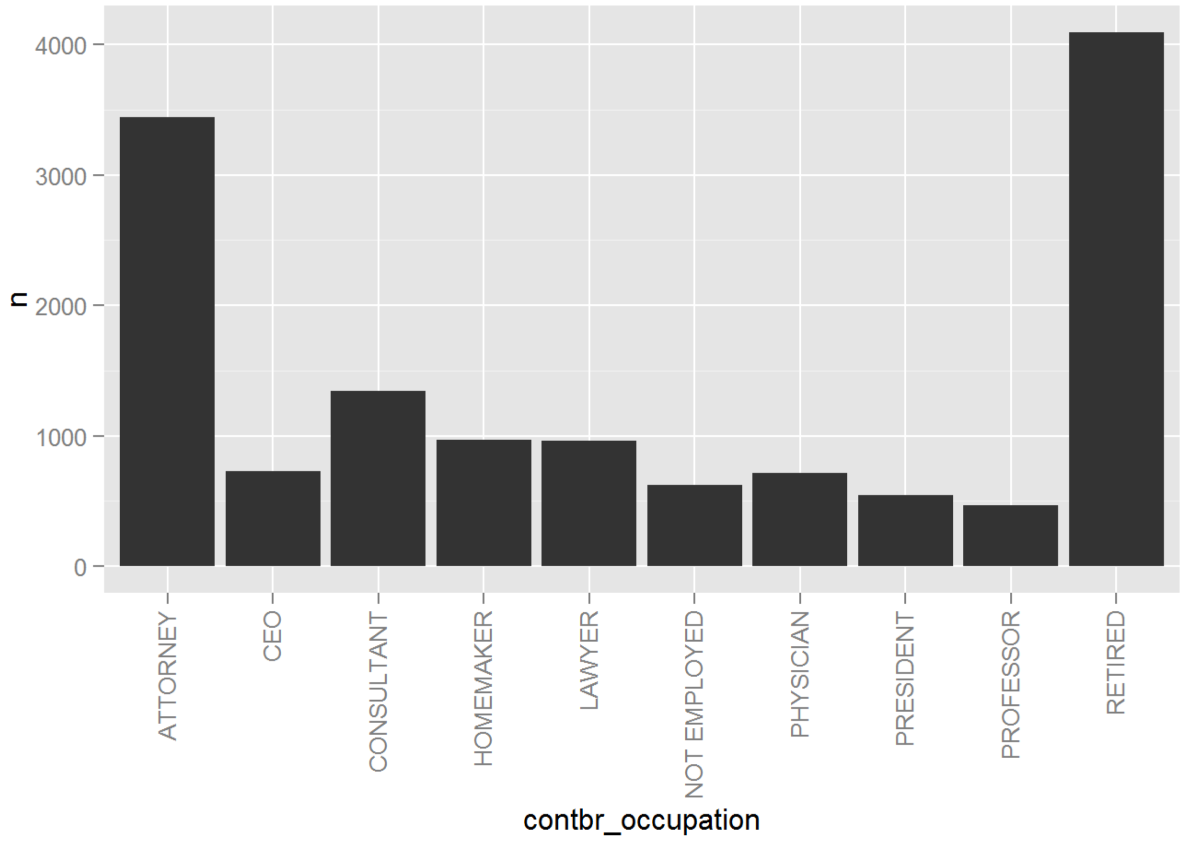
```
hrc_in = filter(hrc_in, contb_receipt_amt <= 2700)

n_distinct(hrc_in[['contbr_occupation']])
```

```
## [1] 3926
```

```
num_by_occup =
    hrc_in %>% filter(contbr_occupation != '', contbr_occupation != 'INFORMA
TION REQUESTED') %>% group_by(contbr_occupation) %>% tally() %>% arrange(des
c(n)) %>% top_n(10)
```

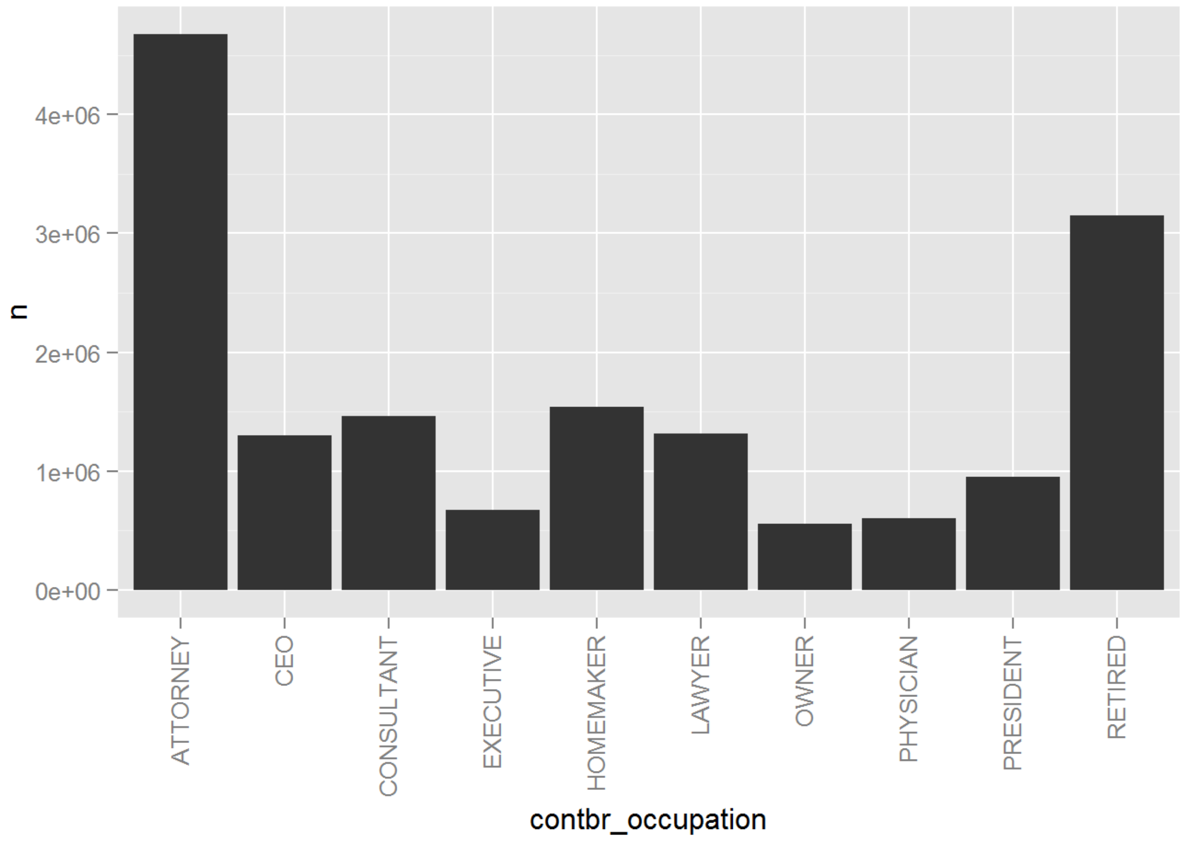
```
## Selecting by n
```



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

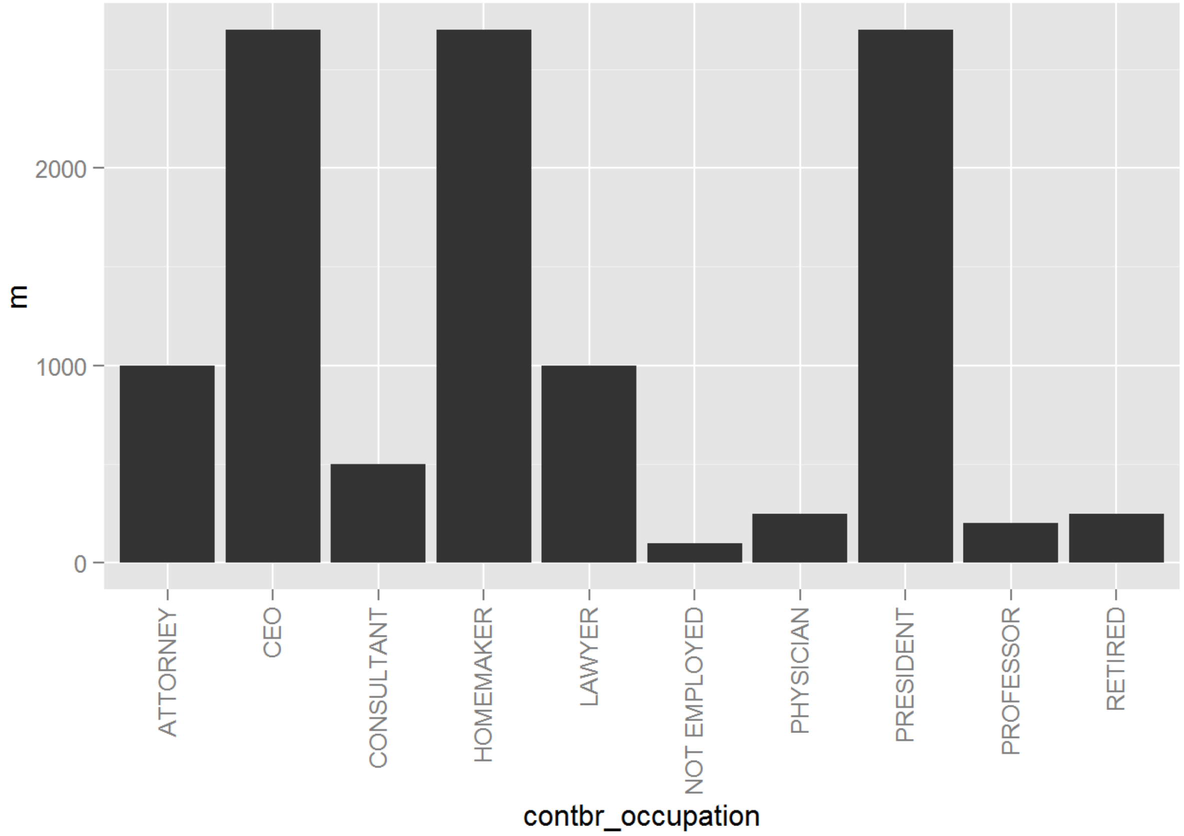
```
val_by_occup =
    hrc_in %>% filter(contbr_occupation != '', contbr_occupation != 'INFORMA
TION REQUESTED') %>% group_by(contbr_occupation) %>% tally(contb_receipt_am
t) %>% arrange(desc(n)) %>% top_n(10)
```

```
## Selecting by n
```



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…

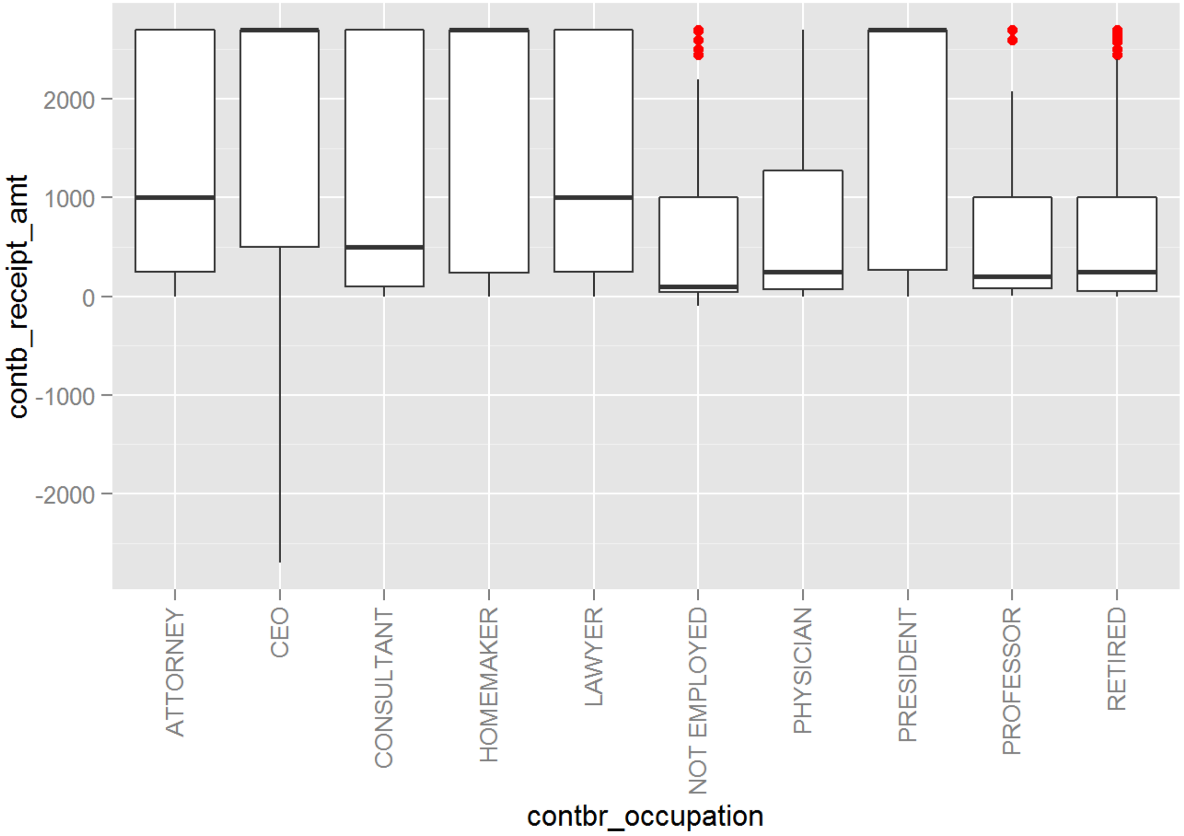
```
median_val_by_occup =
    hrc_in %>% filter(contbr_occupation != '', contbr_occupation != 'INFORMA
TION REQUESTED') %>% group_by(contbr_occupation) %>% summarize(n=n(), m=medi
an(contb_receipt_amt)) %>% arrange(desc(n)) %>% top_n(10, n)
```



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.

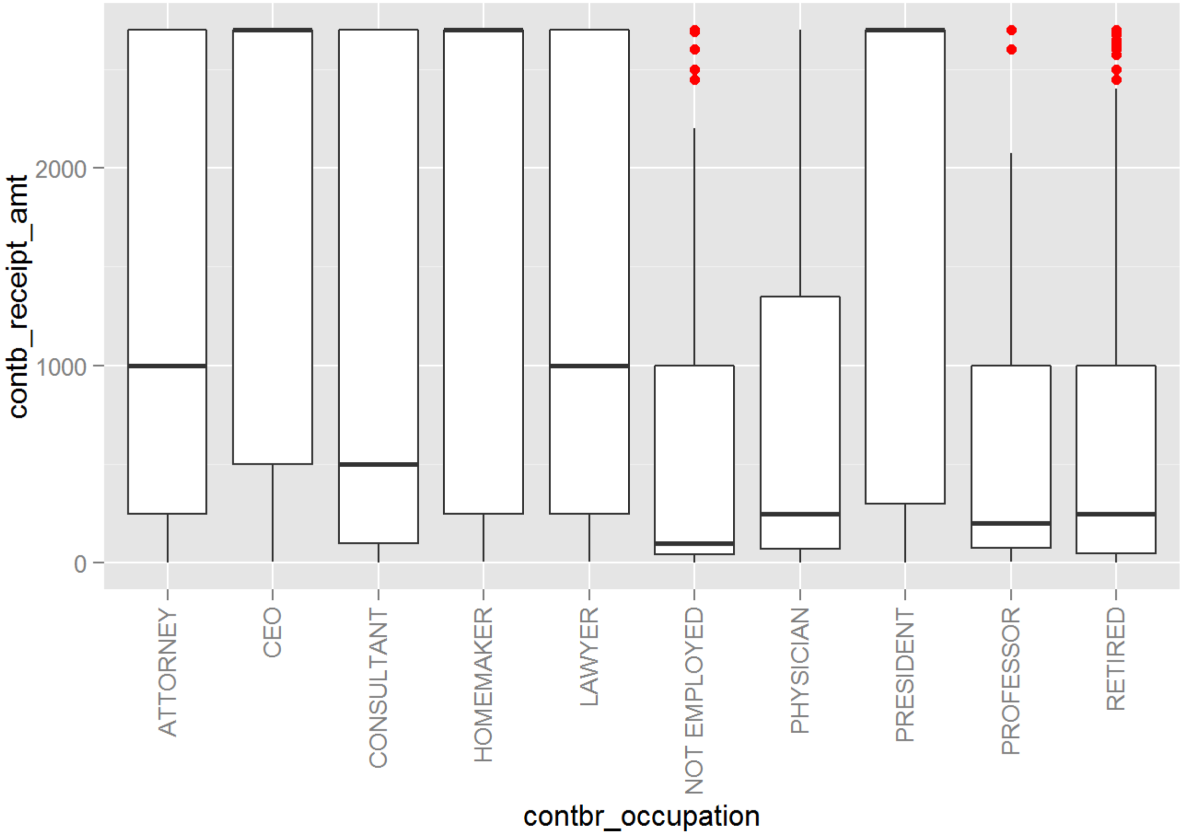
There’s clearly a lot of interesting insight to be found jere, and a box plot might give us a better overview of this data…

```
contbs_top_occups = filter(hrc_in, contbr_occupation %in% num_by_occup[['con
tbr_occupation']])
```



Interesting, but the extreme outliers in the physicians group make the plot difficult to interpret. Refunds (negative contributions) also aren't particularly helpful in this case. Let's omitt both and try again…

```
contbs_top_occups = filter(hrc_in, contbr_occupation %in% num_by_occup[['con
tbr_occupation']], contb_receipt_amt > 0, contb_receipt_amt < 10000)
```



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.

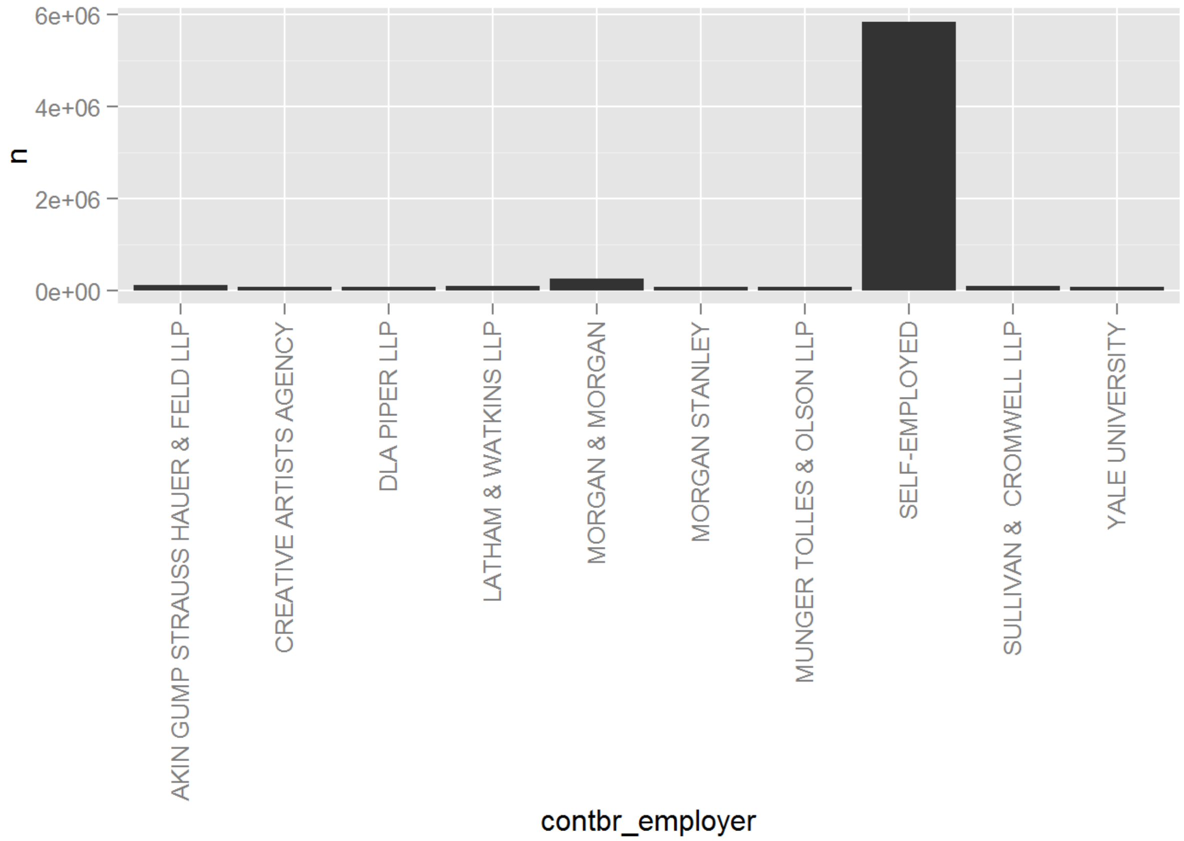
We can also observe a lot of variation among attorneys and presidents, some of them contributing up to 3x more than their respective medians. This is likely indicative of the U-shaped distribution of salaries among attorneys (http://qph.is.quoracdn.net/main-qimg-0a0d8f37efe16a83e4f1208aea3b1988?convert_to_webp=true).

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.

```
val_by_empl =
    hrc_in %>% filter(contbr_employer != '', contbr_employer != 'INFORMATIO
N REQUESTED', contbr_employer != 'N/A', contbr_employer != 'RETIRED', contbr
_employer != 'NOT EMPLOYED') %>% group_by(contbr_employer) %>% tally(contb_r
eceipt_amt) %>% arrange(desc(n)) %>% top_n(10)
```

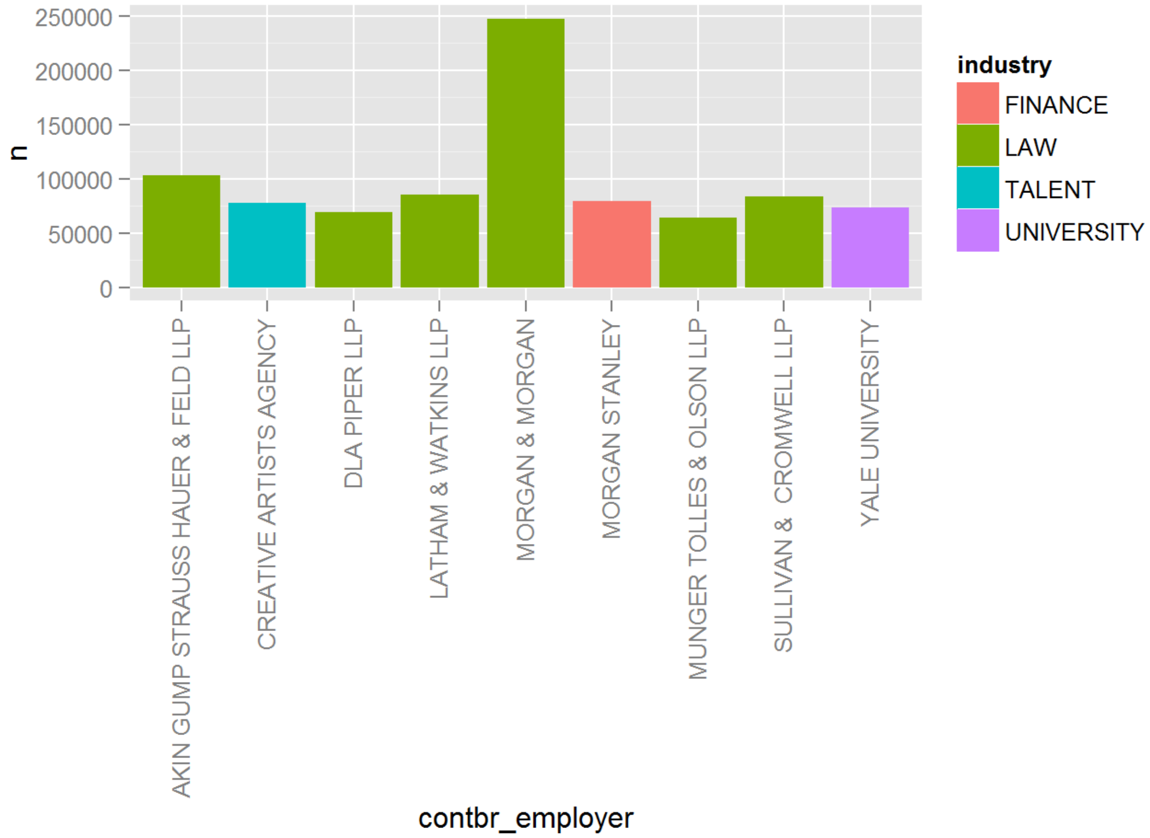
```
## Selecting by n
```



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.

```
val_by_empl_noself =
    val_by_empl %>% filter(contbr_employer != 'SELF-EMPLOYED')
```



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.

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

```
library(RColorBrewer)
library(choroplethr)
library(choroplethrMaps)
library(maptools)
```

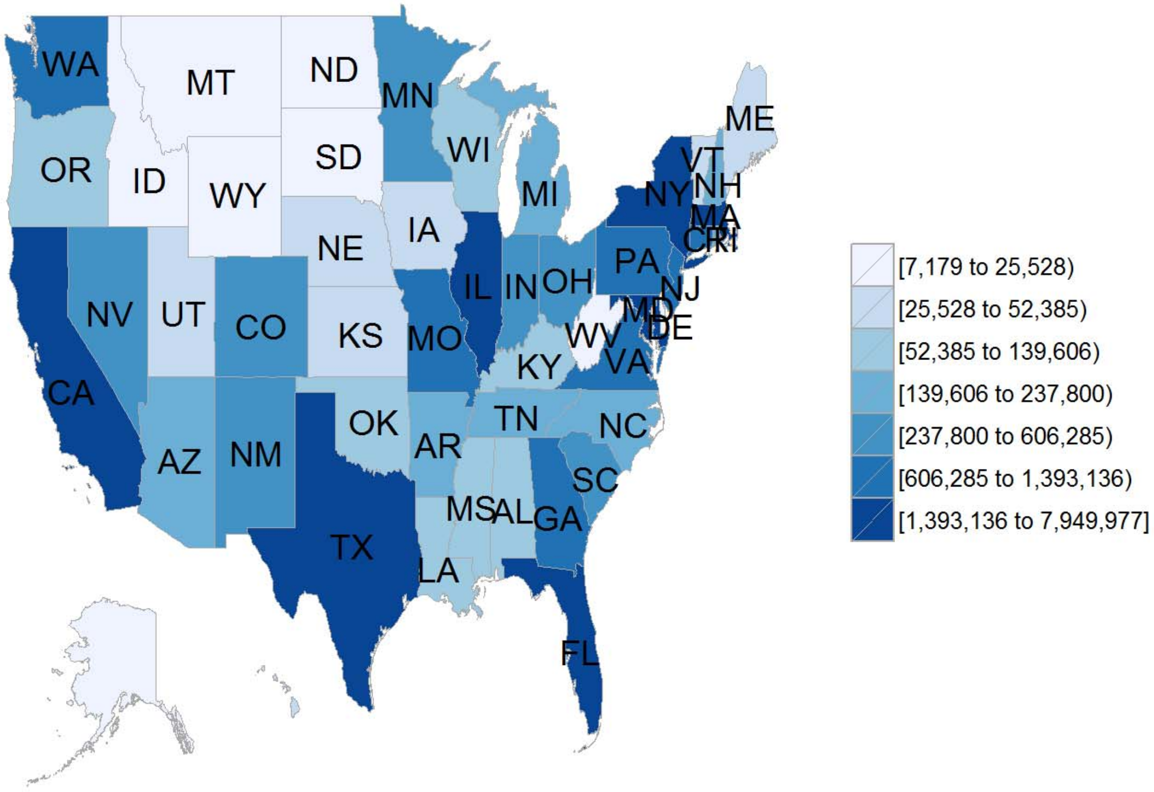
```
## Loading required package: sp
## Checking rgeos availability: TRUE
```

```
val_by_state =
    hrc_in %>% filter(contbr_st != '', contbr_st != 'INFORMATION REQUESTE
D', contbr_st != 'N/A') %>% group_by(contbr_st) %>% summarize(value = sum(co
ntb_receipt_amt)) %>% arrange(desc(value))

val_by_state$region = tolower(state.name[match(val_by_state$contbr_st, stat
e.abb)])

val_by_state = na.omit(val_by_state)
```

```
## Warning in self$bind(): The following regions were missing and are being
## set to NA: district of columbia
```



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

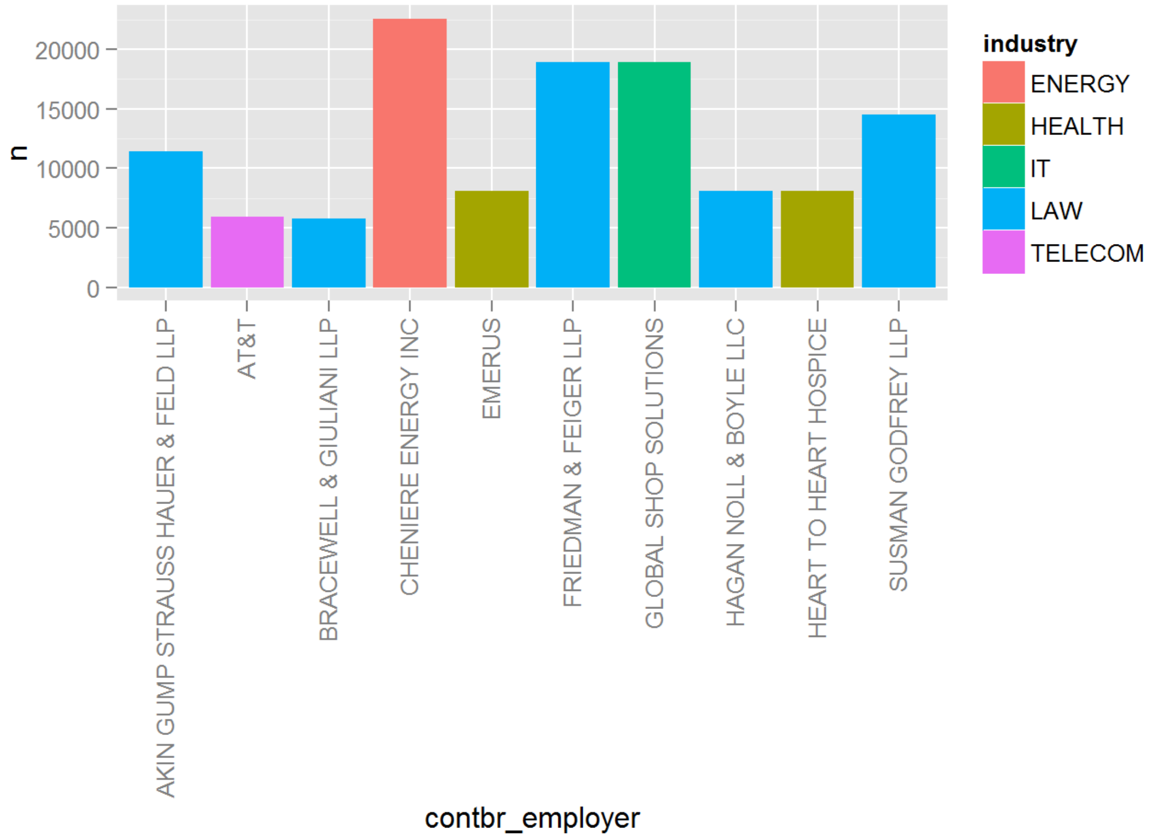
Let’s redraw the employer bar chart using a subset of contributions coming from Texas…

```
hrc_in_texas = filter(hrc_in, contbr_st == 'TX')
```

```
#fixing name inconsistencies that severely skewed the results
hrc_in_texas$contbr_employer = gsub("INC.", "INC", hrc_in_texas$contbr_emplo
yer)
hrc_in_texas$contbr_employer = gsub(", ", " ", hrc_in_texas$contbr_employer)

texas_val_by_empl =
    hrc_in_texas %>% filter(contbr_employer != '', contbr_employer != 'INFOR
MATION REQUESTED', contbr_employer != 'N/A', contbr_employer != 'RETIRED', c
ontbr_employer != 'NOT EMPLOYED', contbr_employer != 'SELF-EMPLOYED') %>% gr
oup_by(contbr_employer) %>% tally(contb_receipt_amt) %>% arrange(desc(n)) %
>% top_n(10)
```

```
## Selecting by n
```

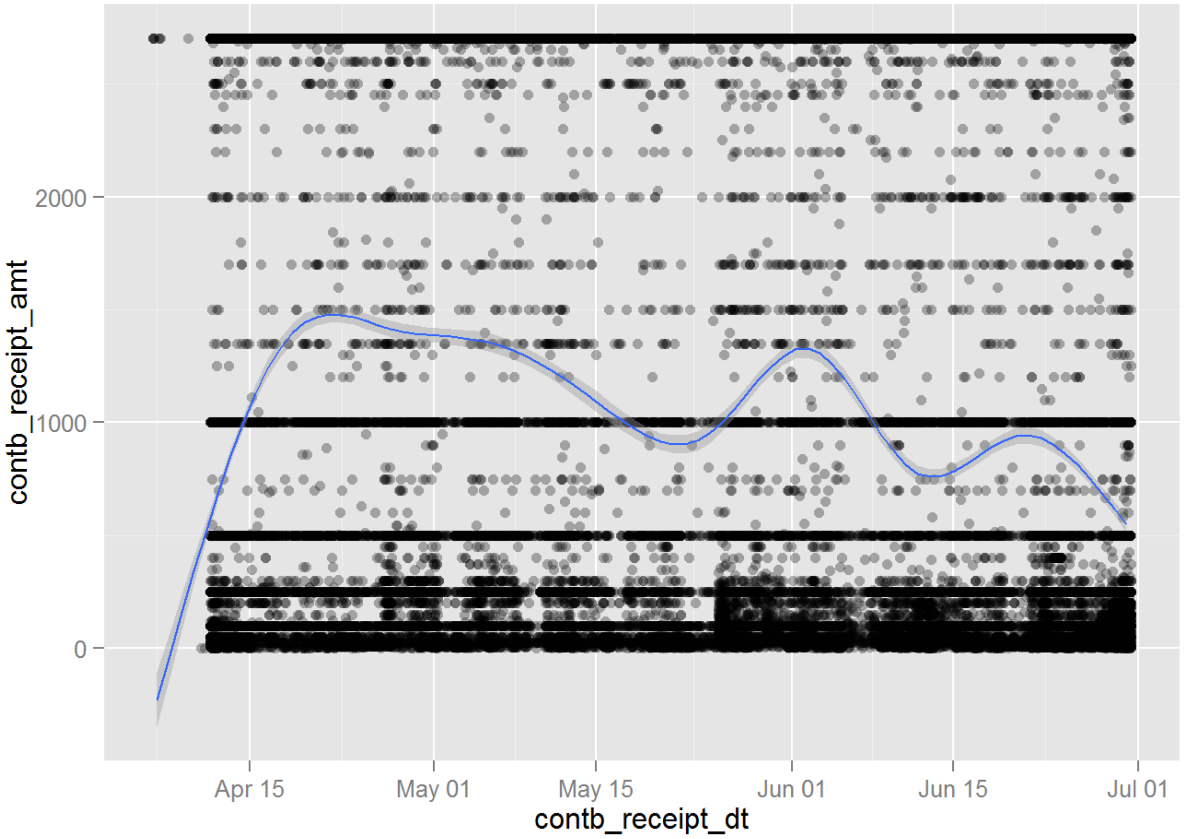


Nothing obvious in this graph, other than law firms are not the dominants sources of contributions in Texas, compared to the national totals. It is also interesting to note that the top contributor is an energy company, a sector which is not represented among the top 10 nationwide.

Contributions from Texas are certainly worth looking into, possibly in other datasets. But now, let's take a look at how contributions change throughout the campaign.

```
hrc_in$contb_receipt_dt = as.Date(hrc_in$contb_receipt_dt, "%d/%m/%Y")
hrc_in_positive = filter(hrc_in, contb_receipt_amt > 0, !is.na(contb_receipt
_dt))
#omitting an extreme outlier from the graph
hrc_in_positive = filter(hrc_in_positive, contb_receipt_amt < 20000)
```

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using
gam with formula: y ~ s(x, bs = "cs"). Use 'method = x' to change the smooth
ing method.
```

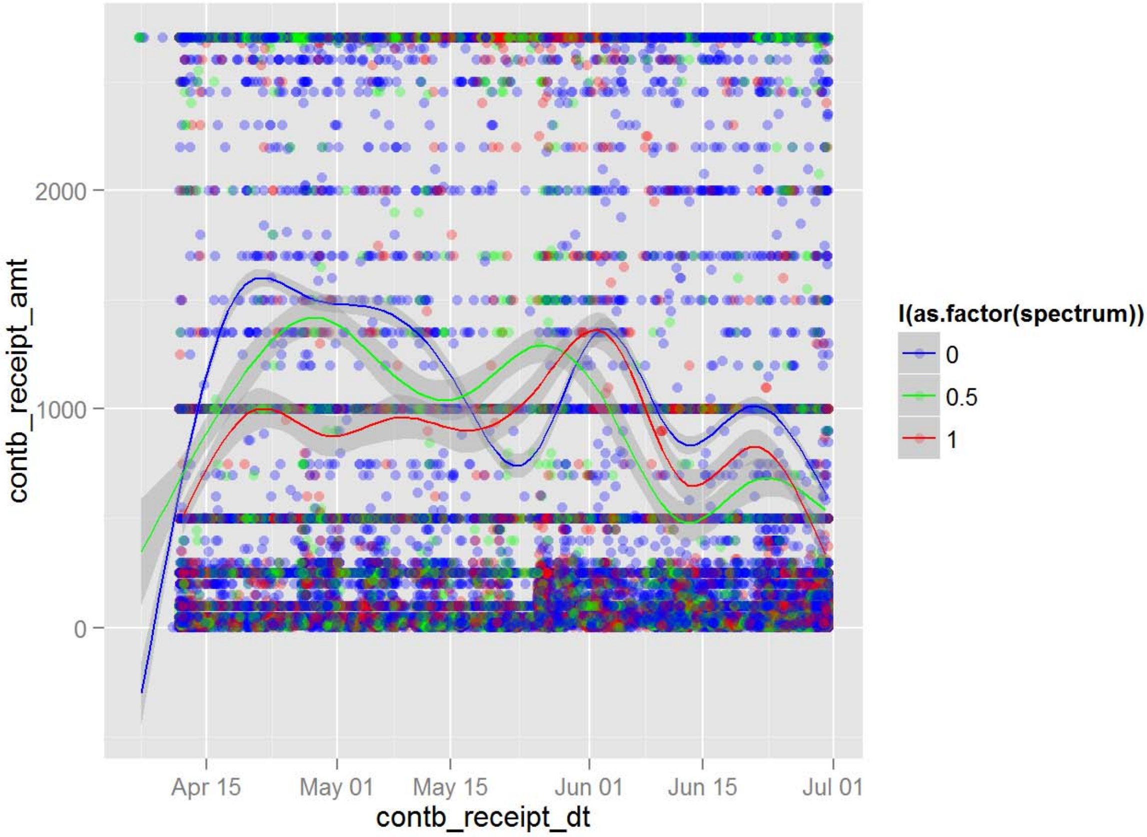


We can see that outide of the occasional outlier, the contribution amounts appear to be remarkably consistent. A an increase of small contributions can be observed in the last few months. This could be seasonal, or a natural evolution as the electoral cycle progresses. Let’s add some colours… red for Republican states, blue for Democrat states, and green for swing states.

```
#adding spectrum colour. 0 for blue states, 1 for red states, 0.5 for swing
states
#based on https://en.wikipedia.org/wiki/Red_states_and_blue_states#/media/Fi
le:Red_state,_blue_state.svg (summary of results of the 2000, 2004, 2008, an
d 2012 presidential elections)
state_spectrum = data.frame(contbr_st = c("AK","AL","AR","AZ","CA","CO","C
T","DC","DE","FL","GA","GU","HI","IA","ID", "IL","IN","KS","KY","LA","MA","M
D","ME","MH","MI","MN","MO","MS","MT","NC","ND","NE","NH","NJ","NM","NV","N
Y","OH","OK","OR","PA","PR","PW","RI","SC","SD","TN","TX","UT","VA","VI","V
T","WA","WI","WV","WY"), spectrum = c(1,1,1,1,0,0.5,0,0,0,0.5,1,0,0,0,1,0,1,
1,1,1,0,0,0,0,0,0,1,1,1,1,1,1,0,0,0,0.5,0,0.5,1,0,0,0,0,0,1,1,1,1,1,0.5,0,0,
0,0,1,1))
hrc_in_positive = merge(hrc_in_positive, state_spectrum, by="contbr_st")
head(hrc_in_positive)
```

```
##   contbr_st       contbr_nm contbr_city contbr_zip
## 1         AK   HAWKS, DENISE      SEWARD  996641169
## 2         AK   HAWKS, DENISE      SEWARD  996641169
## 3         AK   HARRIS, HOLLY     DOUGLAS  998245337
## 4         AK  SCHOLLE, MARIE   FAIRBANKS  997081011
## 5         AK MENDEL, ALLISON   ANCHORAGE  995013223
## 6         AK  ANDREE, JUDITH      JUNEAU  998019760
##                           contbr_employer contbr_occupation
## 1 SOUTHERN ARIZONA VA HEALTH CARE SYSTEM         PHYSICIAN
## 2 SOUTHERN ARIZONA VA HEALTH CARE SYSTEM         PHYSICIAN
## 3                           EARTHJUSTICE          ATTORNEY
## 4                        LOCKHEED MARTIN    SAFETY OFFICER
## 5            MENDEL COLBERT & ASSOCIATES          ATTORNEY
## 6                                    N/A           RETIRED
##   contb_receipt_amt contb_receipt_dt receipt_desc memo_text spectrum
## 1                50       2015-04-12                               1
## 2                50       2015-06-16                               1
## 3               102       2015-06-14                               1
## 4               250       2015-04-12                               1
## 5              1350       2015-05-31                               1
## 6                50       2015-05-16                               1
```

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using
gam with formula: y ~ s(x, bs = "cs"). Use 'method = x' to change the smooth
ing method.
```

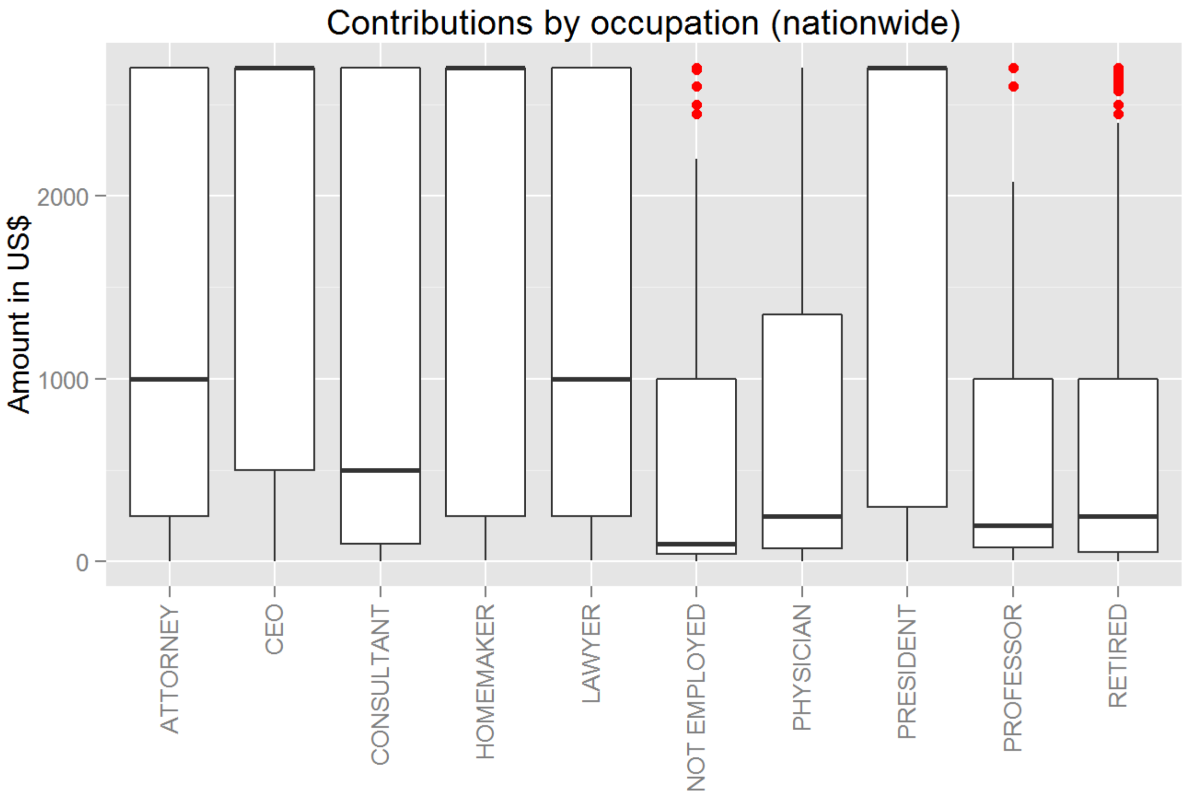


Although the median contributions are consistently higher from Democrat states, the difference isn't very large. Of note are the last two weeks of May, when median Republican and swing state contributions were higher than those from Democrat states.

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)



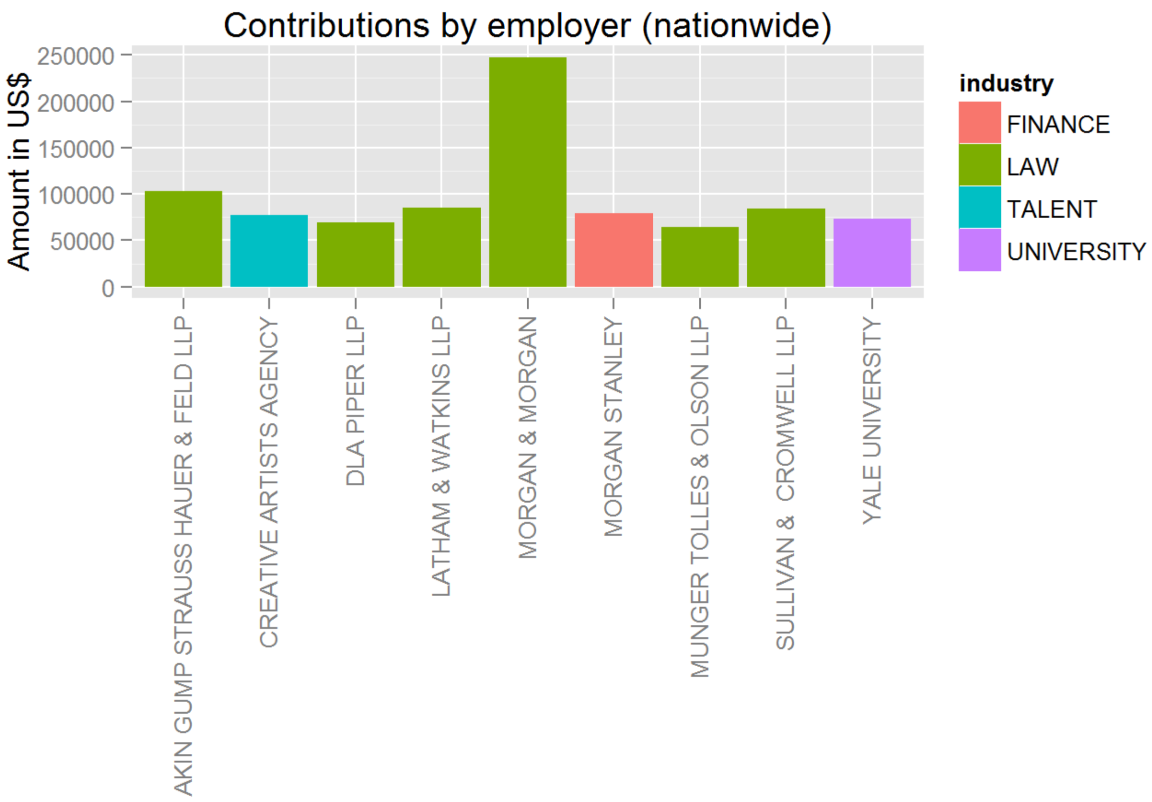
The average contributions to the campaign differ substantially across occupations.

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 employer (nationwide)

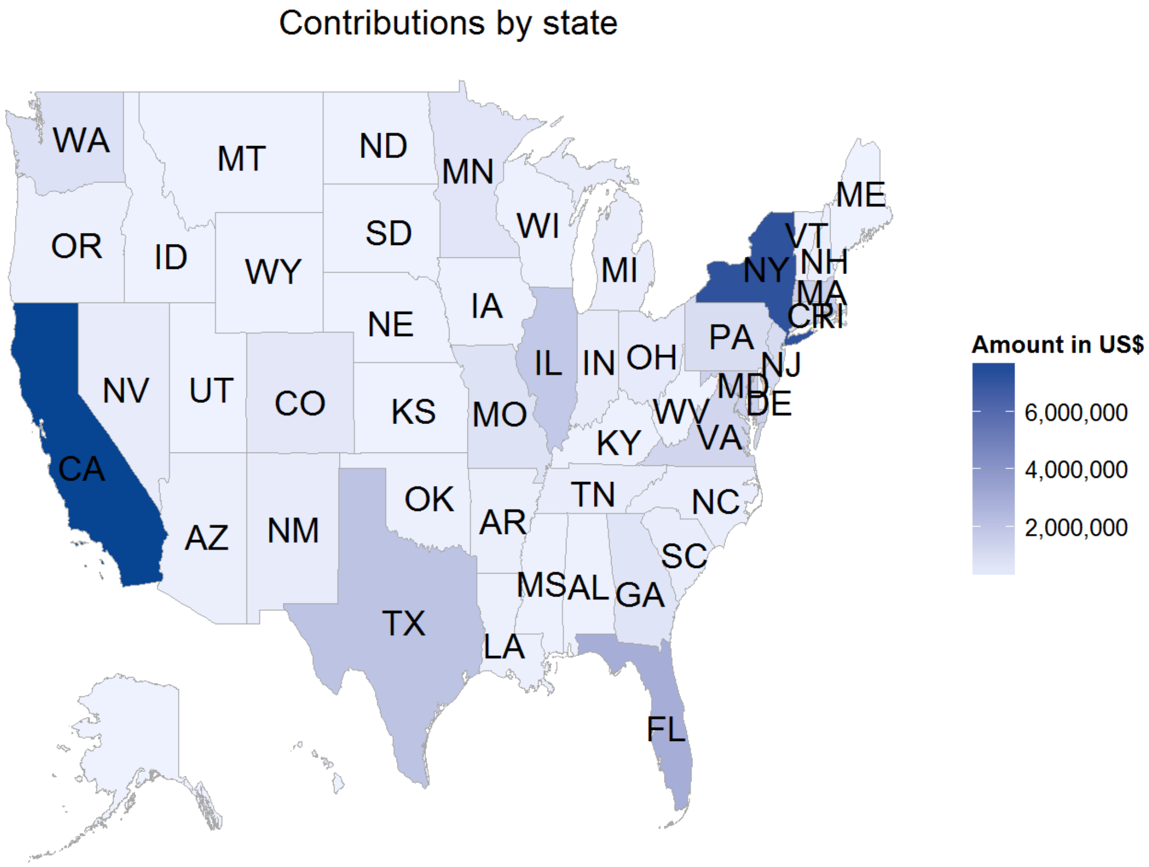


This chart makes it especially clear where a lot of the money flows from. The largest donor is Morgan & Morgan, a consumer protection and personal injury law firm. Most of the other donors are also legal firms, with the exception of one talent agency, one financial institution and one Ivy League university.

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, it is impossible to say whether Morgan Stanley doesn't have an equal, or even larger stake in campaigns of GOP candidates without analysis of the entire dataset.

3. Contributions by state

```
## Warning in self$bind(): The following regions were missing and are being
## set to NA: district of columbia
```



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

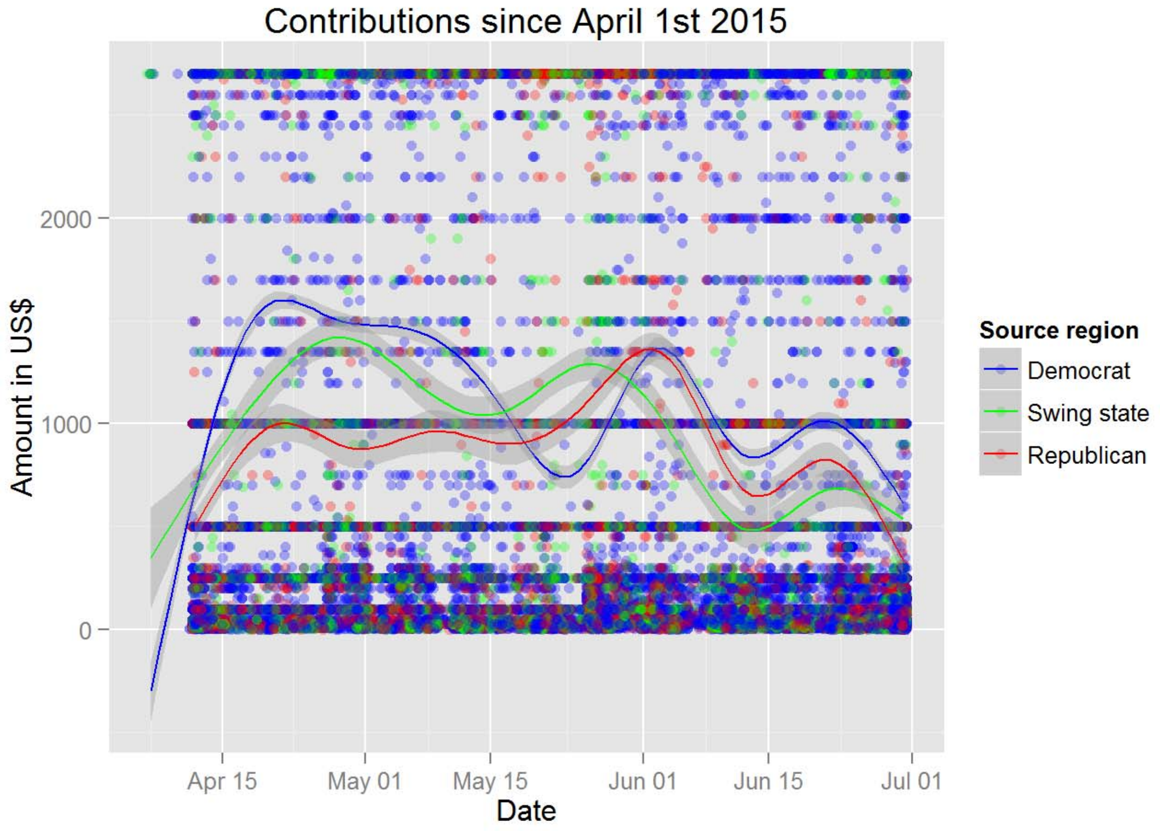
Why is Texas, a predominantly Republican state, one of the top sources of contributions for Hillary Clinton?

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

It may be worthwhile to explore the Texas subset further, at a more granular level, mapping contributions by ZIP code.

4. Contribution time series

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using
gam with formula: y ~ s(x, bs = "cs"). Use 'method = x' to change the smooth
ing method.
```



Plotting contributions from April 1st till July 1st, gives us yet another perspective on the data.

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 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.

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 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.)

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.