Illinois: Presidential Canadidate Donations

Joseph Pete Thomae

1/14/2020

Intro

I have chosen to explore a dataset which details individual monitary donations to presidential candidates in Illinois. All data was collected from FEC.gov. Original data pulled on 1/21/2020.

Disclaimer: The full dataset is used for all analysis unless specified otherwise. Some analysis requires omissions for readability.

Loading Data and Libraries

```
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
```

```
library(reshape2)
library(varhandle)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:reshape2':
##
##     dcast, melt
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(alr3)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------
--------------------------- tidyverse 1.3.0 --
```

```
## <U+2713> tibble  2.1.3     <U+2713> purrr   0.3.3
## <U+2713> tidyr   1.0.0     <U+2713> stringr 1.4.0
## <U+2713> readr   1.3.1     <U+2713> forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------
--------------------- tidyverse_conflicts() --
## x data.table::between() masks dplyr::between()
## x gridExtra::combine()  masks dplyr::combine()
## x dplyr::filter()       masks stats::filter()
## x data.table::first()   masks dplyr::first()
## x dplyr::lag()          masks stats::lag()
## x data.table::last()    masks dplyr::last()
## x car::recode()         masks dplyr::recode()
## x purrr::some()         masks car::some()
## x purrr::transpose()    masks data.table::transpose()
```

```
library(forcats)
library(openair)
library(knitr)

# Setting the proper directory
getwd()
```

```
## [1] "C:/Users/thomaej/Documents/WGU/4th Semester/R Programming/Project"
```

```
setwd("C:/Users/thomaej/Documents/WGU/4th Semester/R Programming/Project")
#setwd("C:/Users/thoma/Documents/WGU/4th/R Programming/Project")

# Loading the data into R Studio
illinois <- read.csv(("Illinois.csv"), header = T, sep = ",", stringsAsFactors = FALSE)

# Changing this column to date type
illinois$contbr_date <- as.Date(illinois$contbr_date,
                                format = "%d-%b-%y")

# Creating a counter column used for tallying the number of contributions
illinois$counter <- 1

str(illinois)
```

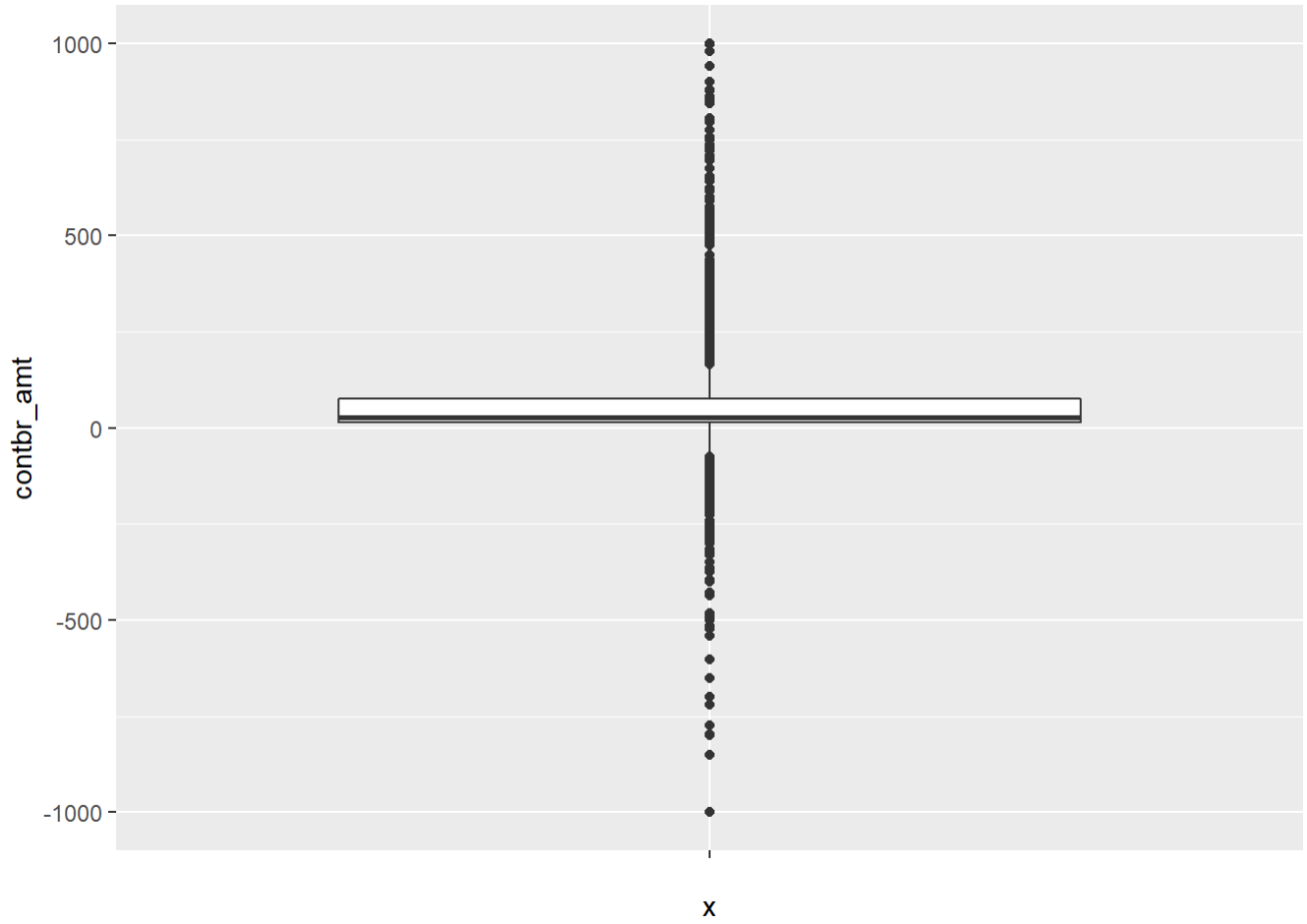
```
## 'data.frame':    77714 obs. of  18 variables:
##  $ cand_id          : chr  "P80001571" "P80001571" "P80001571" "P80001571" ...
##  $ cand_nm          : chr  "Trump, Donald J." "Trump, Donald J." "Trump, Donald J." "Trump, D
onald J." ...
##  $ contbr_nm        : chr  "DEPILLO, KIMBERLY" "HASSAN, RAMZI" "ANDRE, JAMES" "TRZASKA, ALEXA
NDER" ...
##  $ contbr_city      : chr  "JOHNSBURG" "ORLAND PARK" "PARK RIDGE" "INVERNESS" ...
##  $ contbr_st        : chr  "IL" "IL" "IL" "IL" ...
##  $ contbr_zip       : int  60051 60462 60068 60010 60010 60050 60181 60451 61866 61612 ...
##  $ contbr_employer  : chr  "RETIRED" "EDWARDS REALTY COMPANY" "SELF-EMPLOYED" "INFORMATION RE
QUESTED" ...
##  $ contbr_occupation: chr  "RETIRED" "REAL ESTATE DEVELOPER" "CONTRACTOR" "INFORMATION REQUES
TED" ...
##  $ contbr_amt       : num  300 124.8 50 85.1 26.2 ...
##  $ contbr_date      : Date, format: "2017-01-10" "2016-11-25" ...
##  $ receipt_desc     : chr  "" "" "" "" ...
##  $ memo_cd          : chr  "" "X" "" "X" ...
##  $ memo_text        : chr  "" "" "" "" ...
##  $ form_tp          : chr  "SA17A" "SA18" "SA17A" "SA18" ...
##  $ file_num         : int  1174081 1174081 1248056 1174081 1263561 1174081 1301594 1263561 13
01594 1193597 ...
##  $ tran_id          : chr  "SA17A.233535" "SA18.356844" "SA17A.1044727" "SA18.395865" ...
##  $ election_tp      : chr  "P2020" "G2016" "P2020" "P2020" ...
##  $ counter          : num  1 1 1 1 1 1 1 1 1 ...
```

```
# Closer look at donation statistics
summary(illinois$contbr_amt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2800.0    15.0    27.0   106.6    75.0  5600.0
```

```
# Box plot for donations
ggplot(aes(x = " ", y = contbr_amt), data = illinois) +
  geom_boxplot() +
  scale_y_continuous(limits = c(-1000, 1000))
```

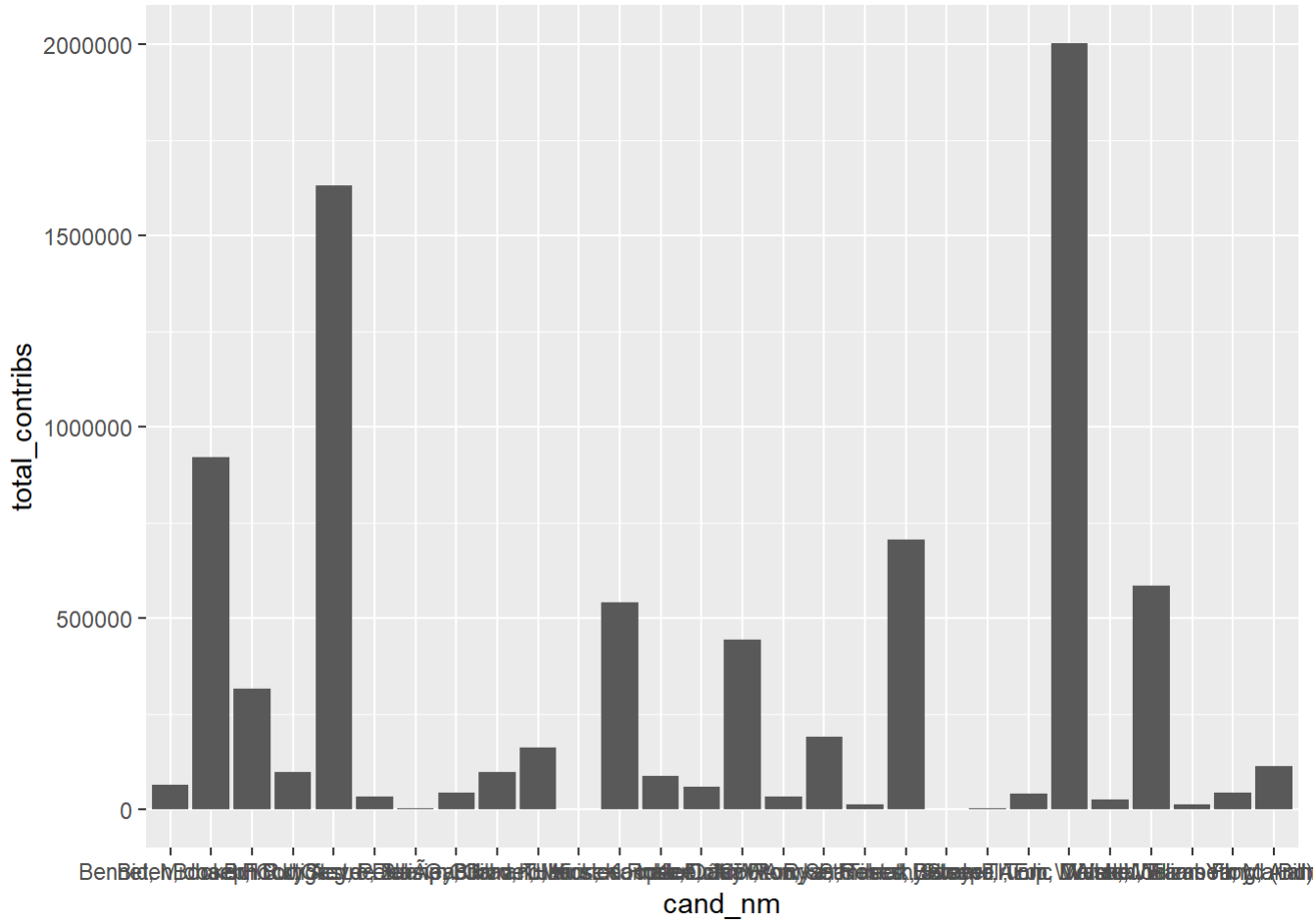
```
## Warning: Removed 1274 rows containing non-finite values (stat_boxplot).
```



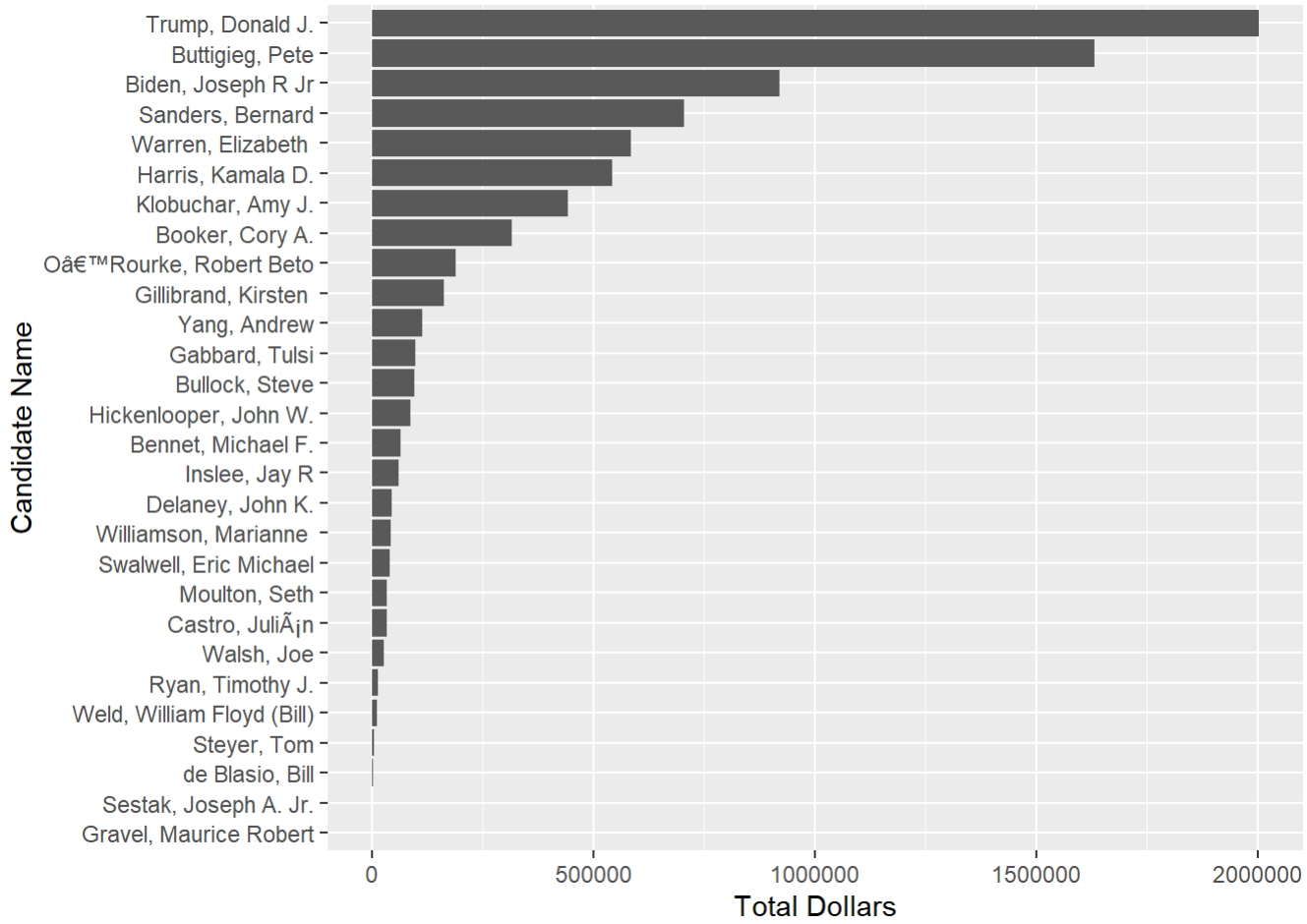
- The vast majority of donations are small amounts, $100 or less. The negative/refunded donations are of some concern to me, I'm not quite sure what to make of them or why they exist.

- Number of contributions to date is exactly as I would expect, far more movement as we draw closer to election year.

```
# Total monitary contributions per candidate
# grouping the data by candidate with running totals for contributions
illinois.candidates <- group_by(illinois, cand_nm)
contribs_by_cand <- dplyr::summarise(illinois.candidates,
                             total_contribs = sum(contbr_amt),
                             n = n())

# Dollar amounts contributed per candidate
ggplot(aes(x = cand_nm, y = total_contribs), data = contribs_by_cand) +
  geom_bar(stat = 'identity')
```

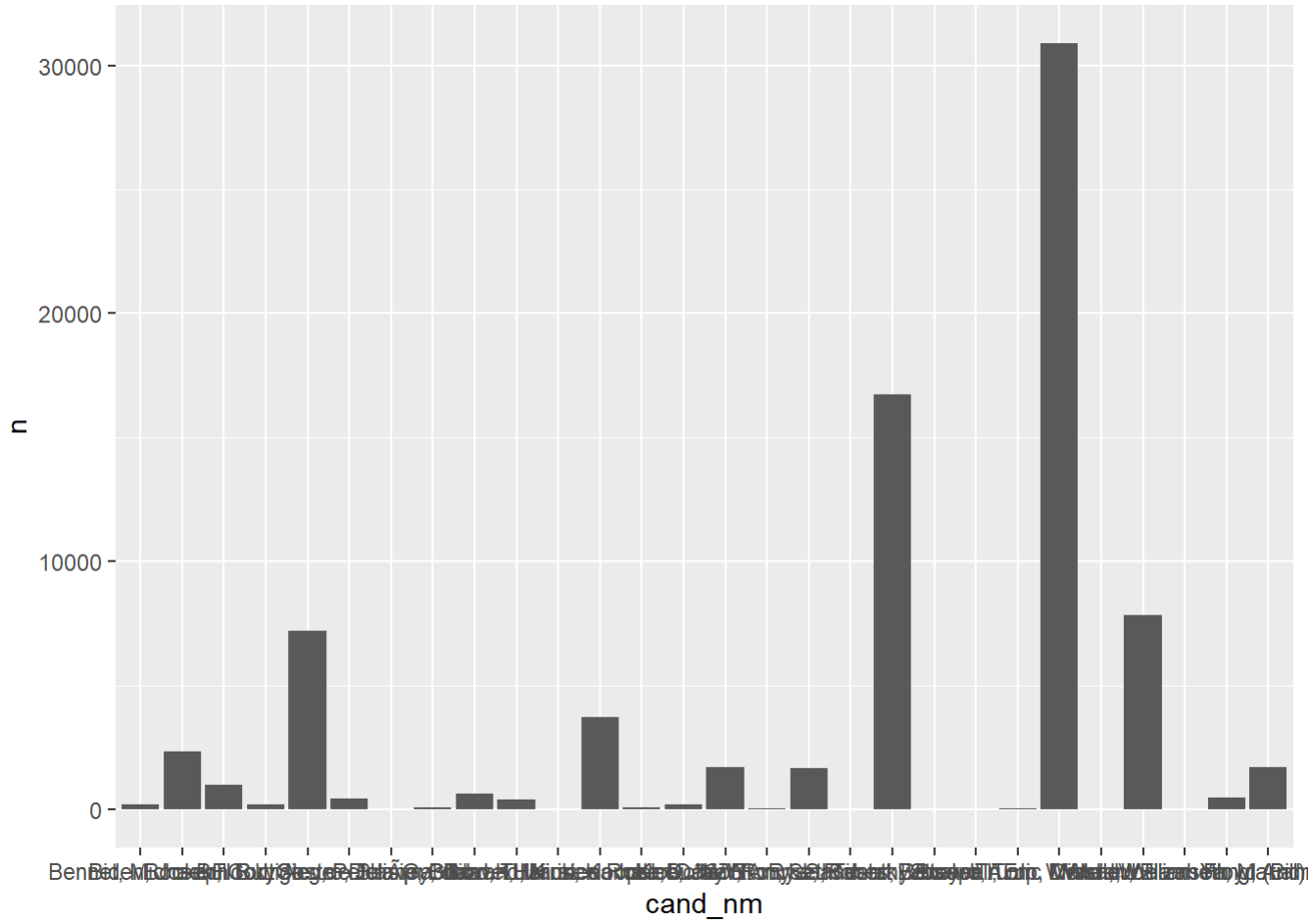



```
# Flipping graph for readability, adding labels
ggplot(aes(x = reorder(cand_nm, total_contribs), y = total_contribs), data = contribs_by_cand) +
  geom_bar(stat = 'identity') +
  coord_flip() +
  labs(x = "Candidate Name", y = "Total Dollars")
```

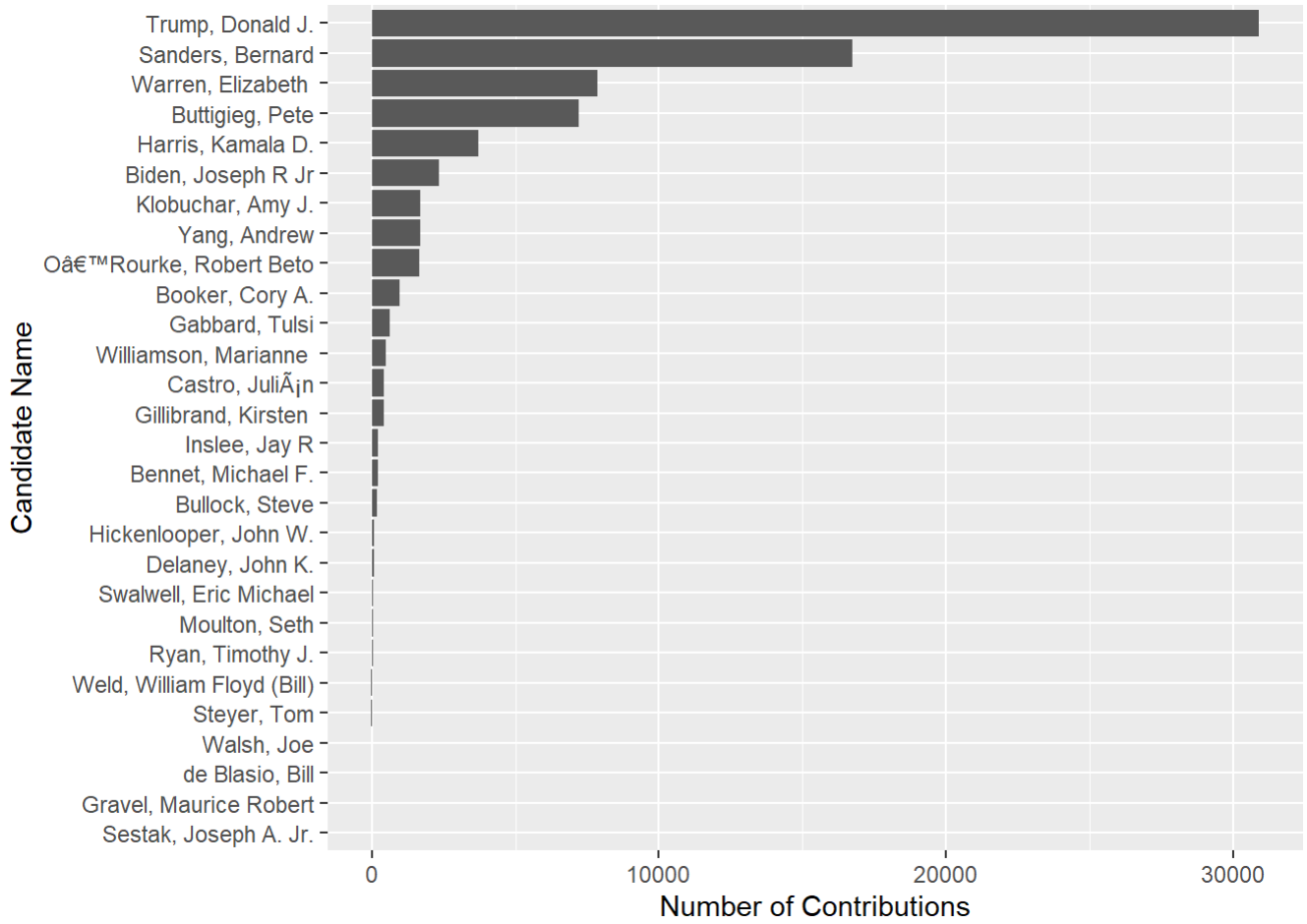


- From the above I can see the top candidates by contribution amount, I've never heard of most candidates, and a few at the top are suprising.

```
# Number of contributions (count) per candidate
ggplot(aes(x = cand_nm, y = n), data = contribs_by_cand) +
  geom_bar(stat = 'identity')
```




```
# Flipping graph for readablity, adding labels
ggplot(aes(x = reorder(cand_nm, n), y = n), data = contribs_by_cand) +
  geom_bar(stat = 'identity') +
  labs(x = "Candidate Name", y = "Number of Contributions") +
  coord_flip()
```

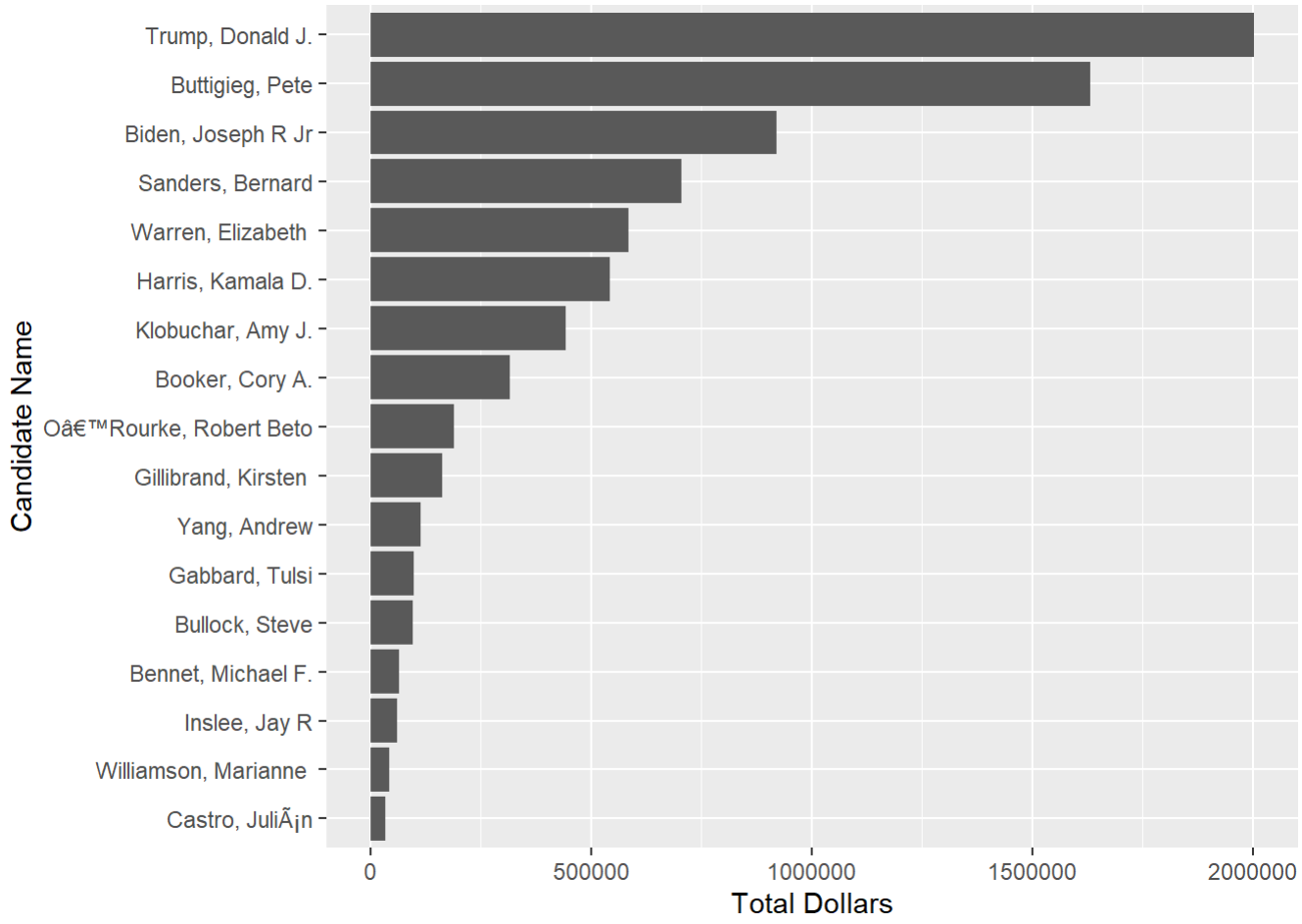


- Total number of contributions can tell a great deal as to how much support a candidate has, Bernie Sanders is high on this graph, suggesting he counts on quantity over quality when collecting donations.

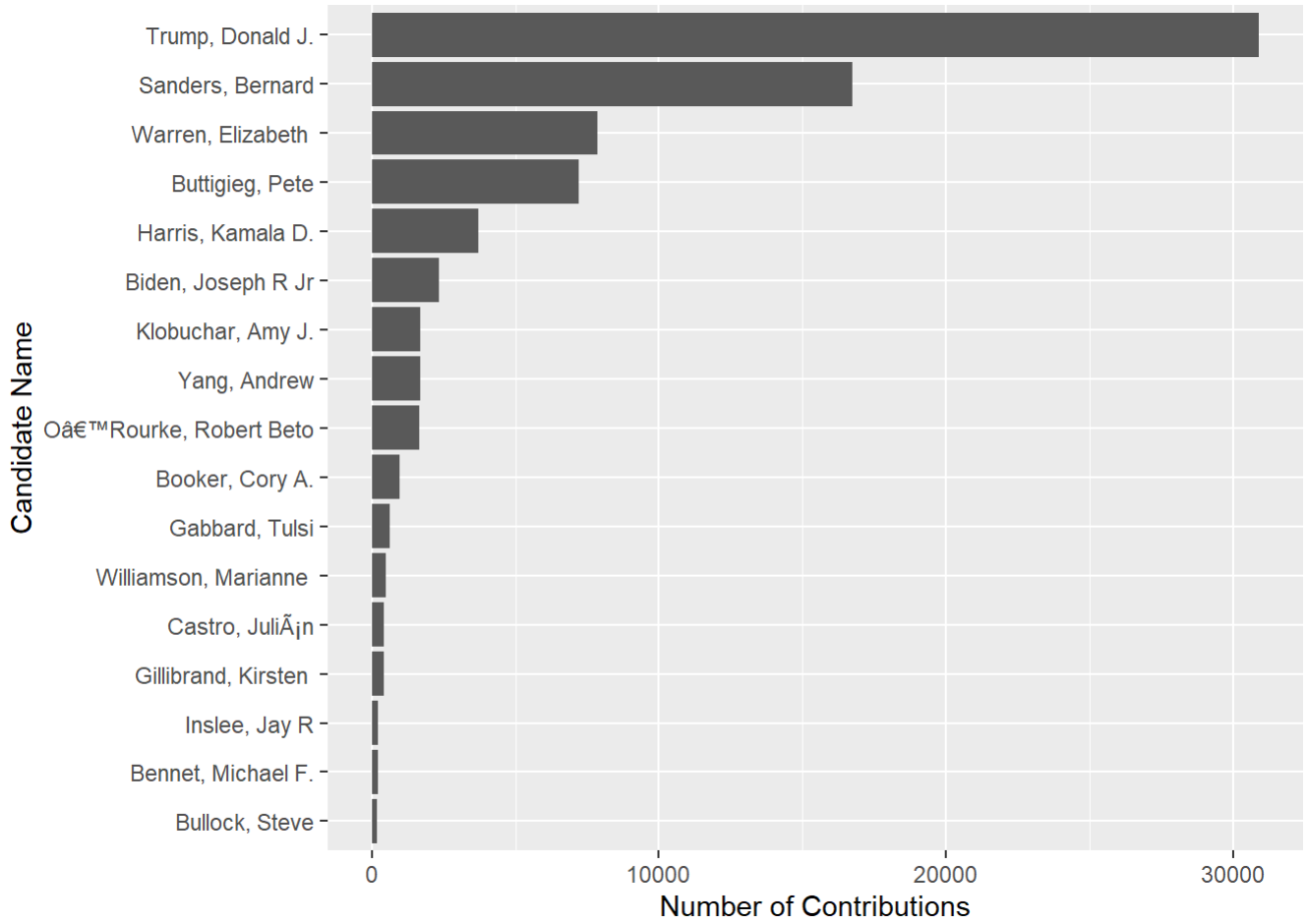
- There are a total of 28 candidates within this dataset, which makes visuizations difficult to work with. The bottom 12 candidates who recieved contributions have totals so low I really don't care to account for them. Limiting the data to candidates with at least 100 contributions will increase readability and omit candidates with no chance (sorry!).

Tip: limiting my data to candidates with a shot (top %50 of totals) below:

```
# Limiting to top %50 of candidates
top.half <- subset(contribs_by_cand, n >= 100)

ggplot(aes(x = reorder(cand_nm, total_contribs), y = total_contribs), data = top.half) +
  geom_bar(stat = 'identity') +
  coord_flip() +
  labs(x = "Candidate Name", y = "Total Dollars")
```




```
ggplot(aes(x = reorder(cand_nm, n), y = n), data = top.half) +
  geom_bar(stat = 'identity') +
  labs(x = "Candidate Name", y = "Number of Contributions") +
  coord_flip()
```

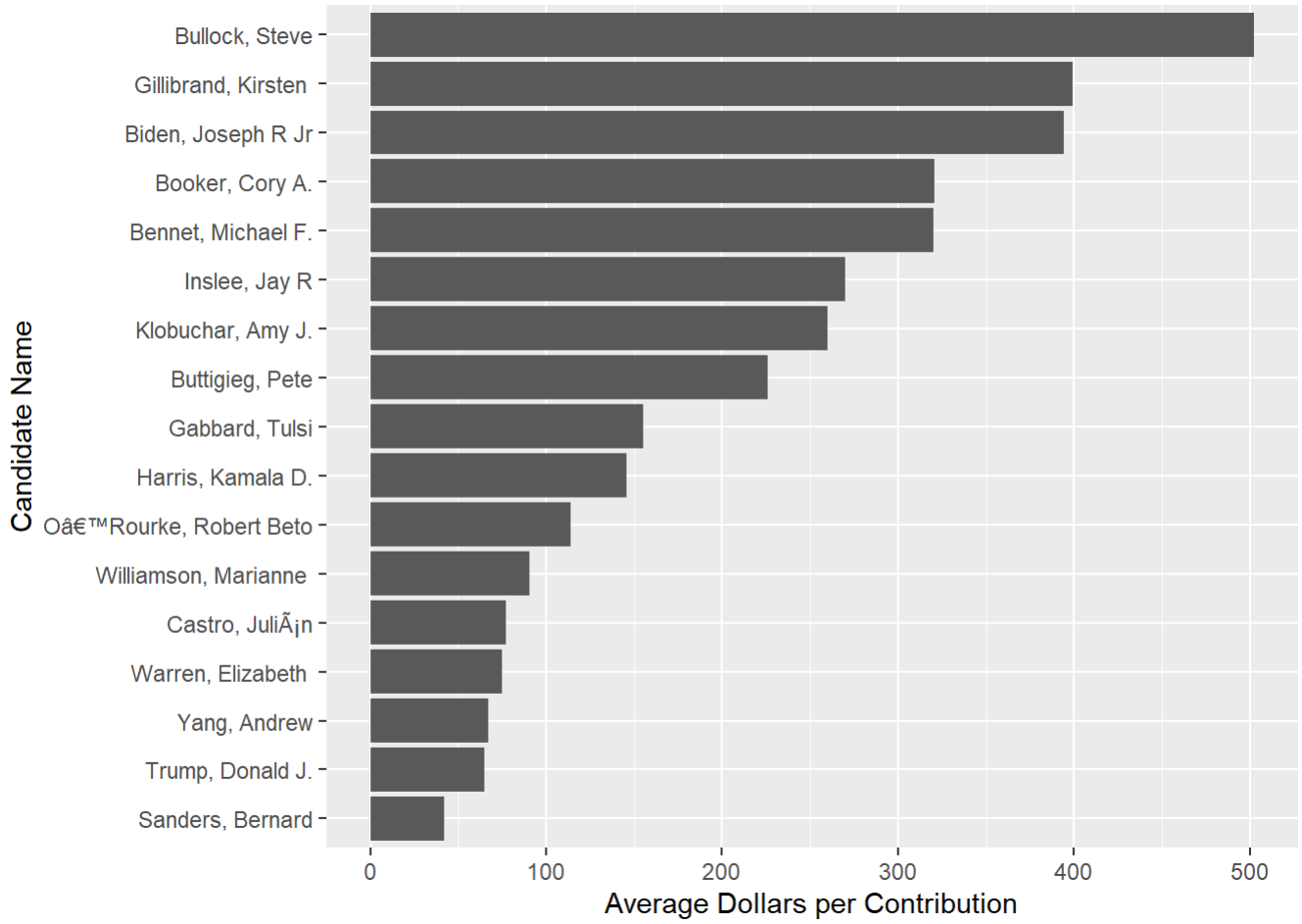


- While total dollar amount for each candidate is important, you can't discount the total number of contributions, which translate to actual votes.

-Example: Steve Bullock has raised nearly as much cash as Andrew Yang, however Andrew Yang's supporters outnumber Steve Bullock's nearly 9:1 based on contributors, which gives him the edge.

- Based on previous findings I'd like to combine these 2 graphs. Total donation amount / # of donations.

```
# Average dollar amount per contribution
ggplot(aes(x = reorder(cand_nm, total_contribs / n), y = total_contribs / n), data = top.half) +
  geom_bar(stat = 'identity') +
  coord_flip() +
  labs(x = "Candidate Name", y = "Average Dollars per Contribution")
```



Exploring average amount per donation

Illinois voter base

- Let’s see what contributors from Illinois do for a living and how they donate.

Tip: Once again I will be segmenting the data into a more usable subset. There are over 3000 unique occupations, with some being typos, errors, or omitted data.

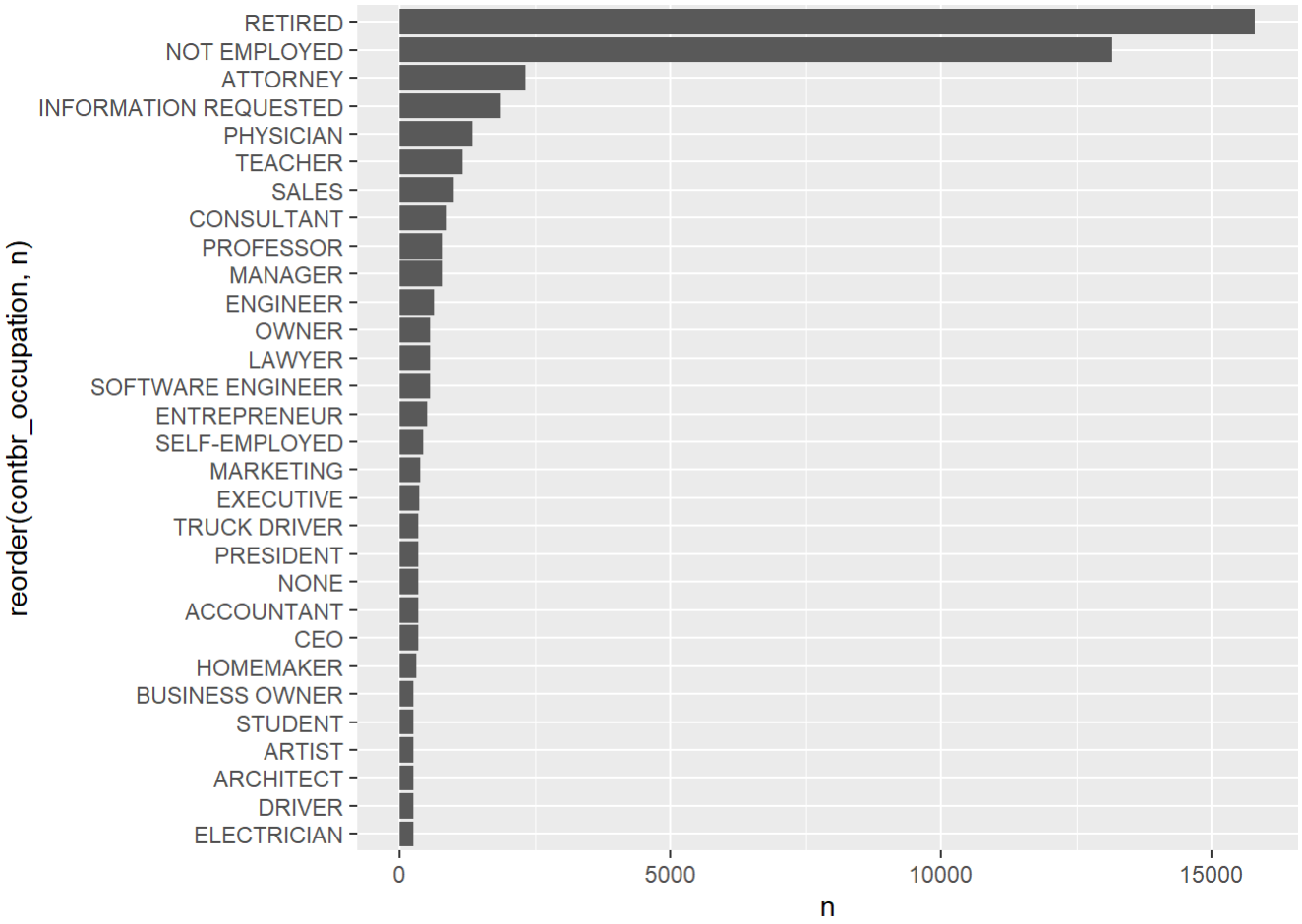
```
summary(unique(illinois$contbr_occupation))
```

```
##    Length     Class      Mode
##      3208 character character
```

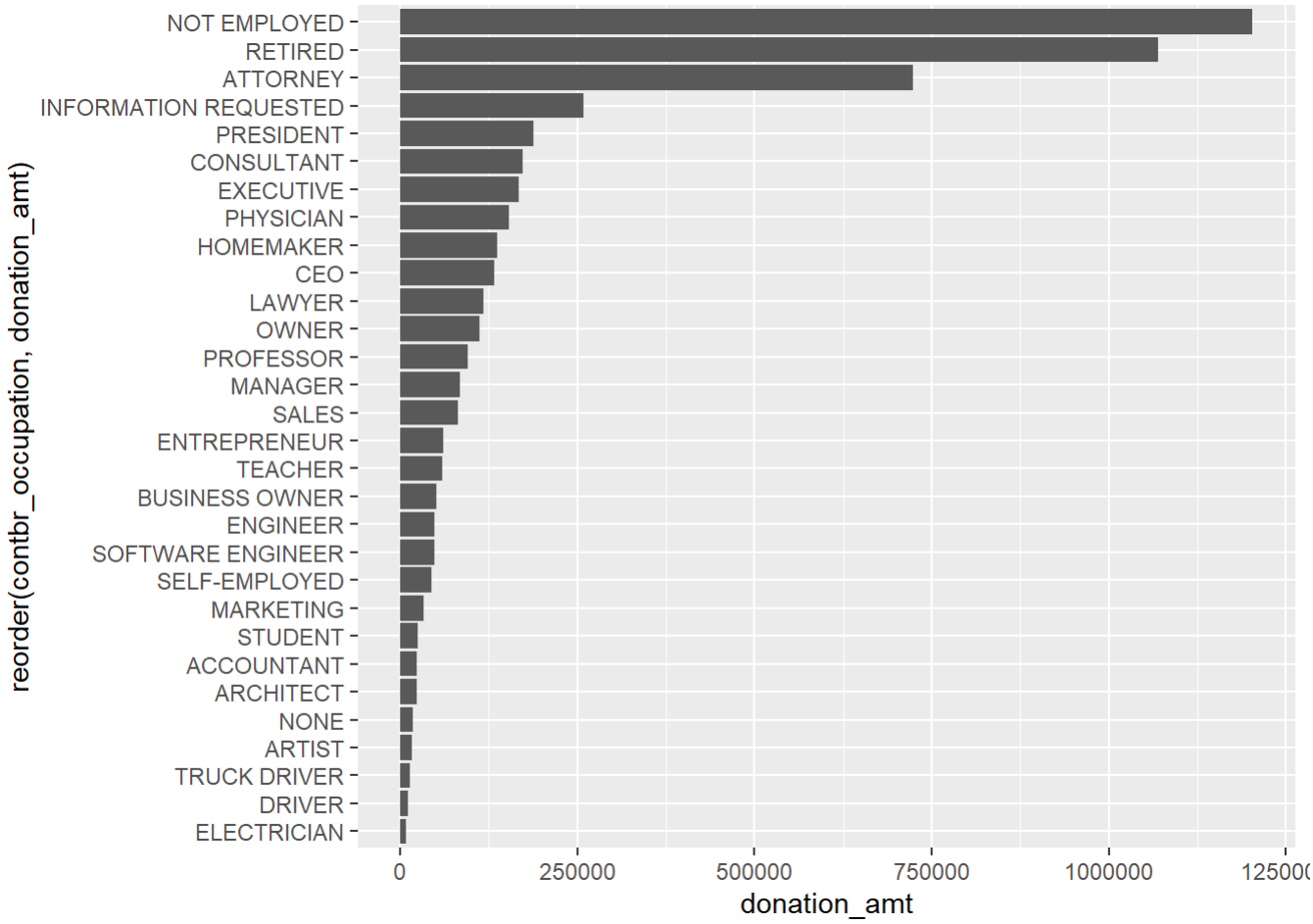
```
# 3208 unique occuaptions, far too many to work with, this will be limited

# Grouping by occupation as well as amount donated per occupation
illinois.occupations <- group_by(illinois, contbr_occupation, na.rm=TRUE)
contribs_by_occupation <- dplyr::summarise(illinois.occupations,
                                           donation_amt = sum(contbr_amt),
                                           n = n())

# Limiting to top 30 occupations by occurance
illinois.top.30 <- subset(contribs_by_occupation, n > 245)

summary(illinois.top.30)
```

```
##  contbr_occupation   na.rm          donation_amt           n          
##  Length:30          Mode:logical   Min.   :   7889   Min.   :  246.0  
##  Class :character   TRUE:30        1st Qu.:  26587   1st Qu.:  339.5  
##  Mode  :character                  Median :  71162   Median :  468.5  
##                                    Mean   : 172734   Mean   : 1554.7  
##                                    3rd Qu.: 149606   3rd Qu.:  847.0  
##                                    Max.   :1202970   Max.   :15807.0
```

```
# Top 30 occupations by occurance
ggplot(aes(x = reorder(contbr_occupation, n), y = n), data = illinois.top.30) +
         geom_bar(stat = 'identity') +
         coord_flip()
```




```
# Top 30 occupations by occurance with donations
ggplot(aes(x = reorder(contbr_occupation, donation_amt), y = donation_amt), data = illinois.top.
30) +
        geom_bar(stat = 'identity') +
        coord_flip()
```

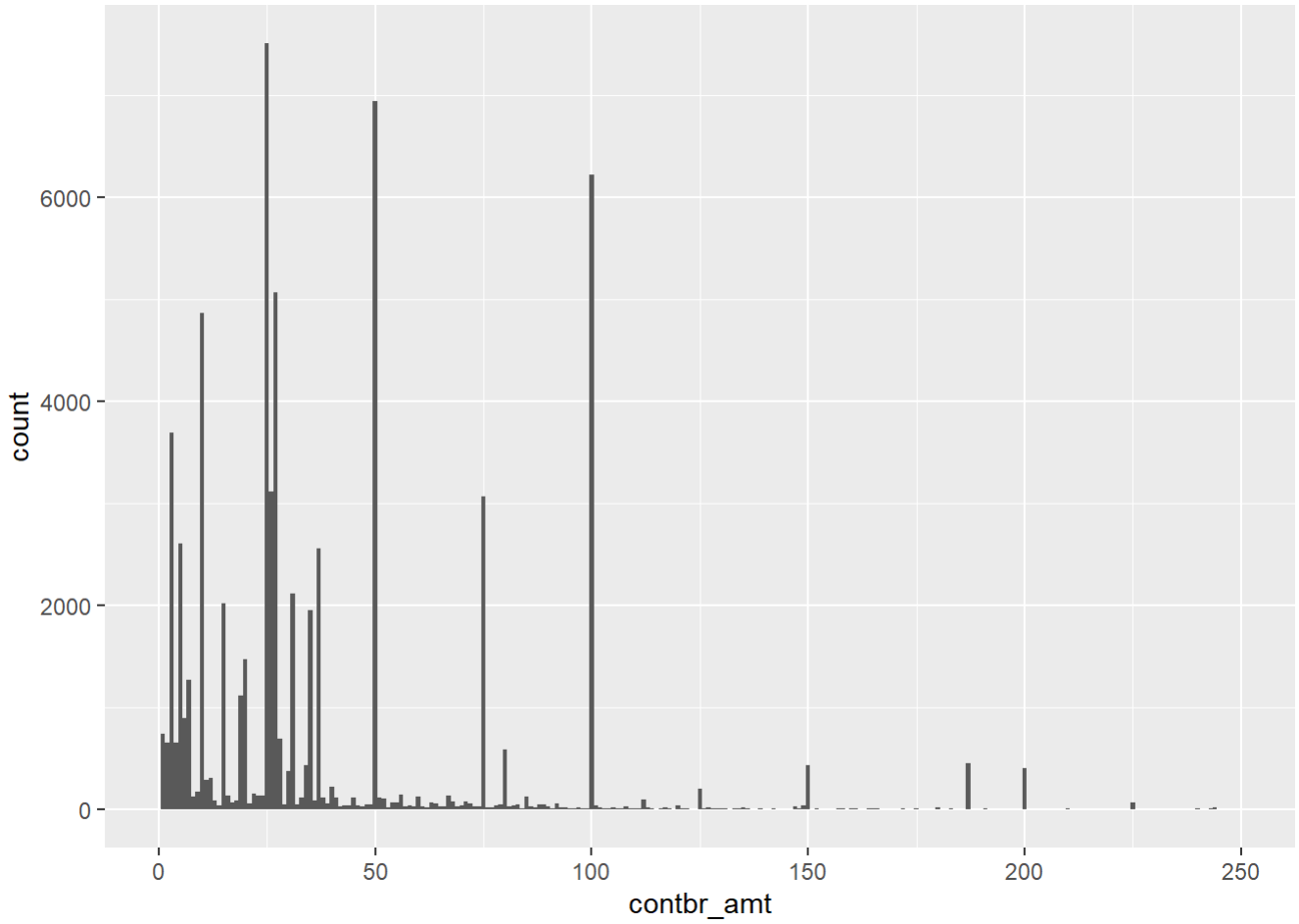


- Only relevant point here is that most donors are either unemployeed, retired, or unreported. also a suprising number of attornies here.

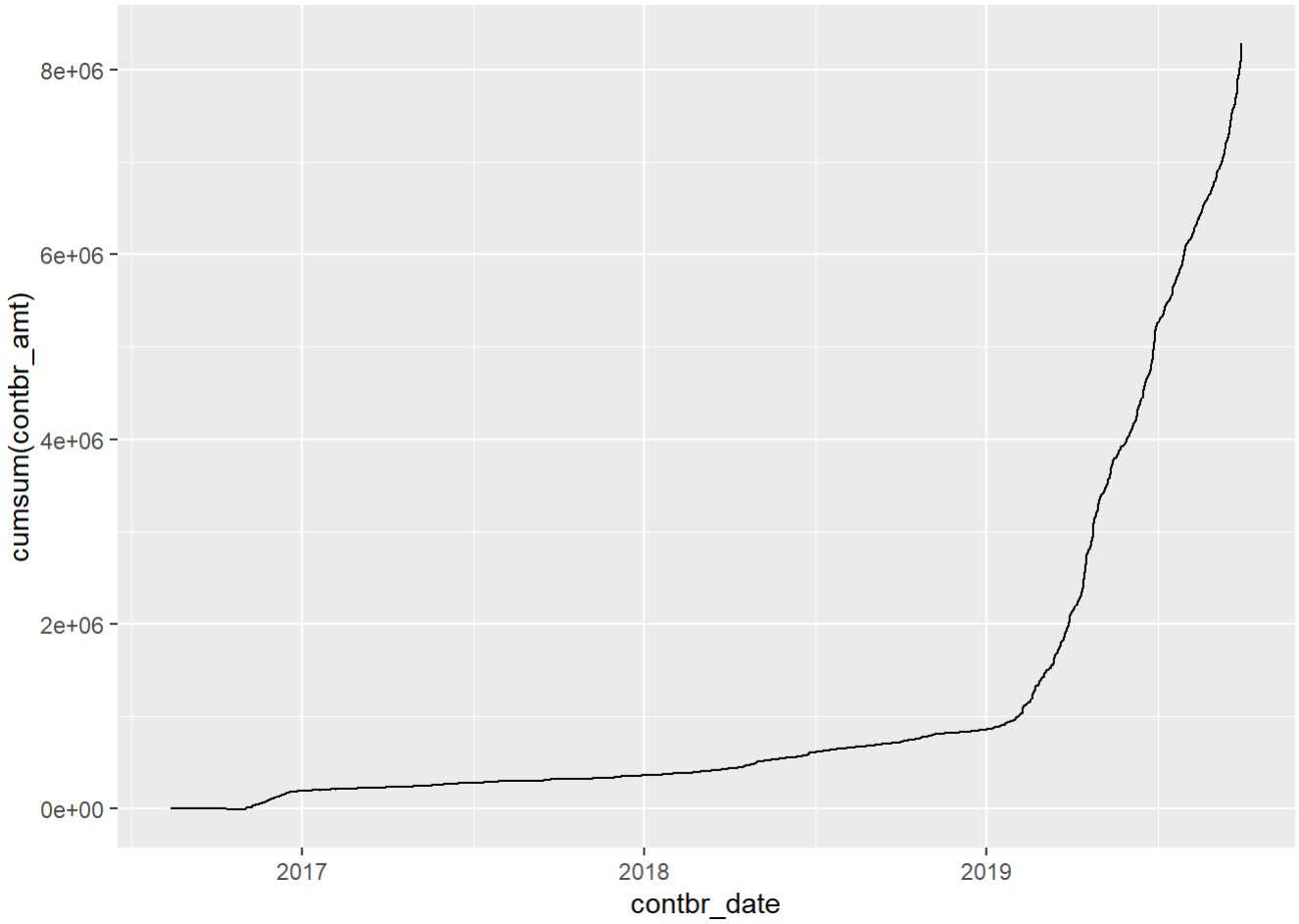
```
# Donation amount by count
ggplot(aes(x = contbr_amt), data = illinois) +
  geom_histogram(binwidth = 1) +
  xlim(c(0, 250))
```

```
## Warning: Removed 5937 rows containing non-finite values (stat_bin).
```

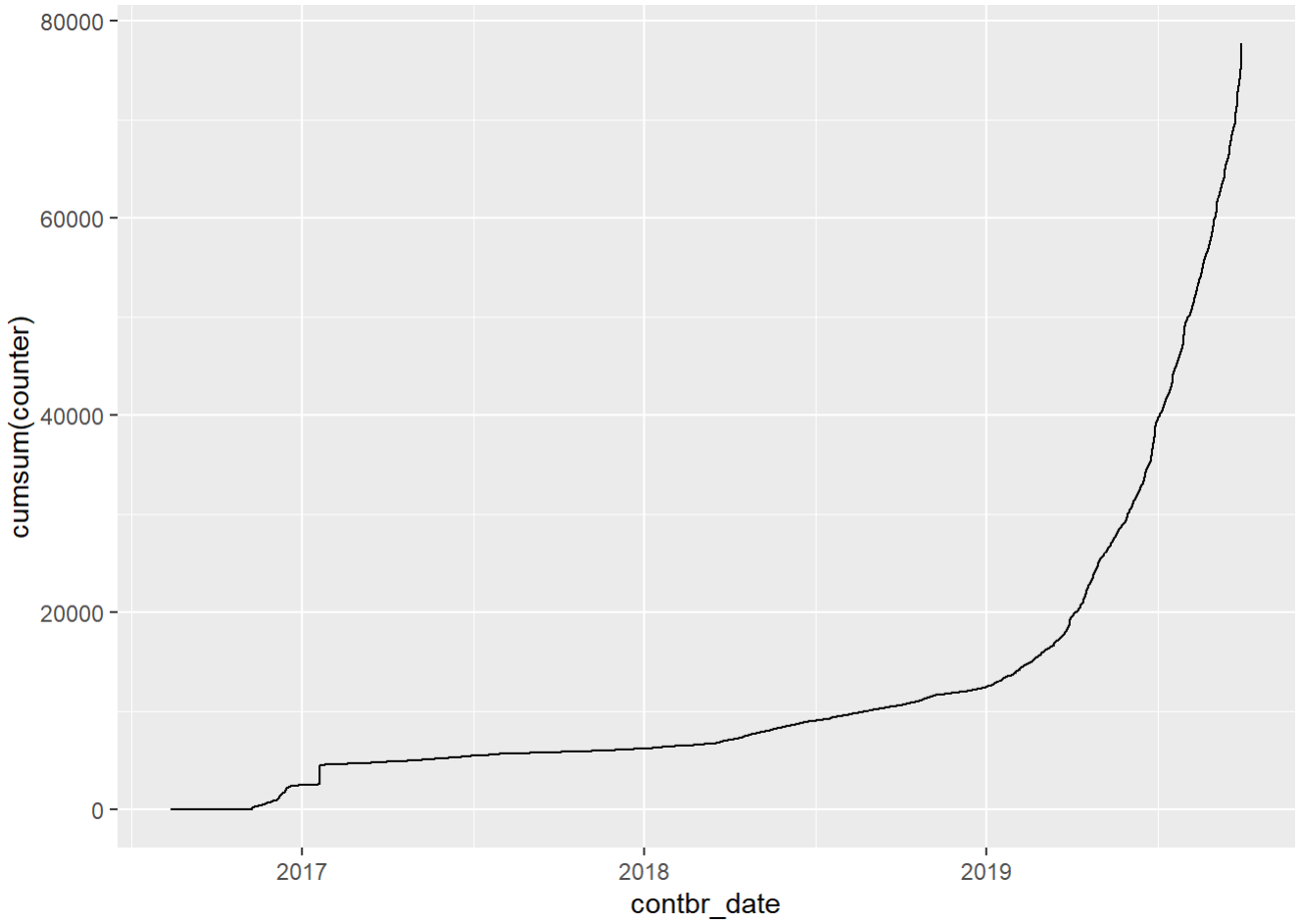
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



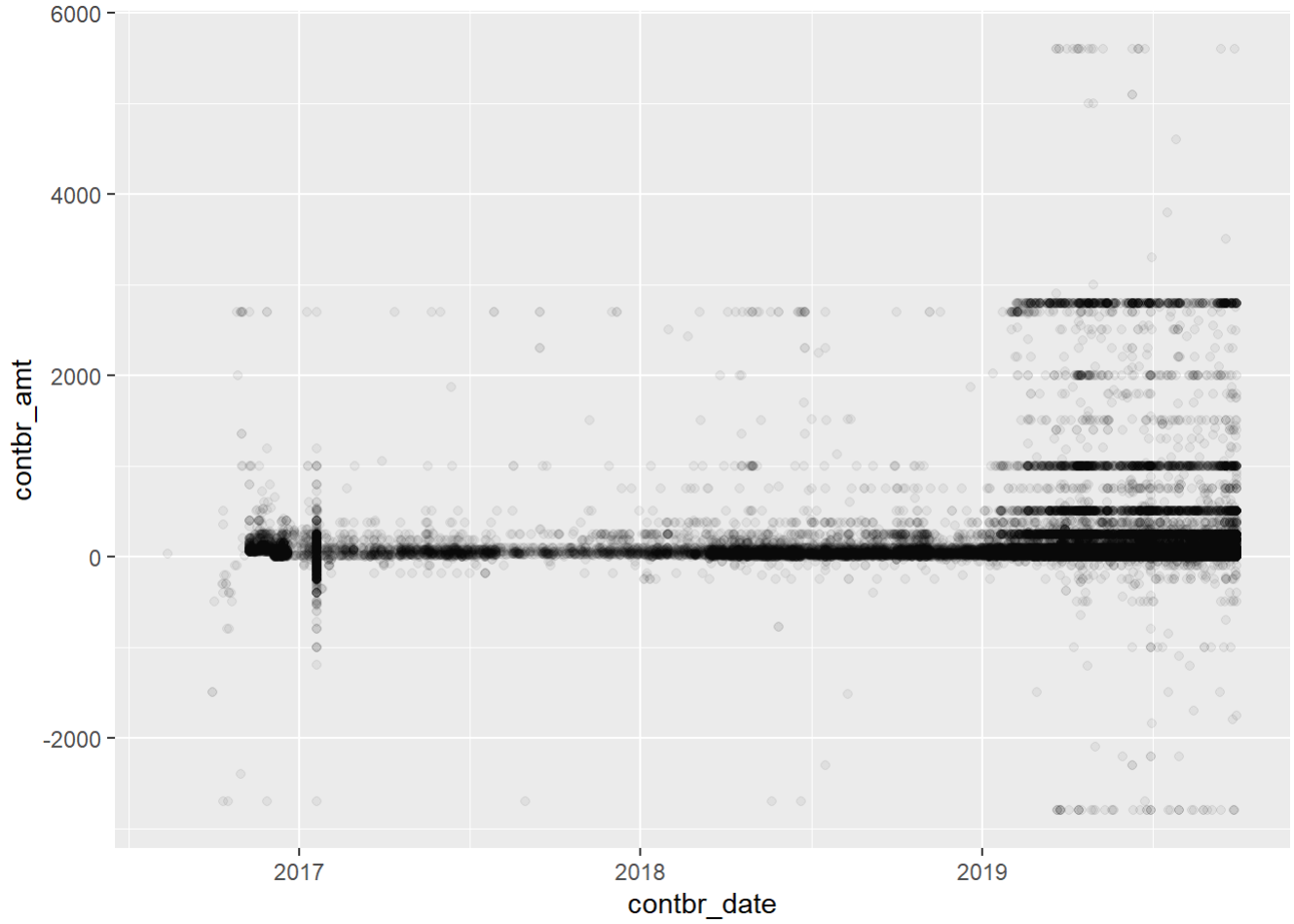

```
# Contributions over time
illinois.by.year <- illinois[order(as.Date(illinois$contbr_date, format = "%d-%b-%y")),]

# Running total of contribution amounts over time, monitary value
ggplot(aes(x = contbr_date, y = cumsum(contbr_amt)), data = illinois.by.year) +
  geom_line()
```




```
# Running total of number of contributions over time, not monitary value
ggplot(aes(x = contbr_date, y = cumsum(counter)), data = illinois.by.year) +
  geom_line()
```




```
# All contributions over time
ggplot(aes(x = contbr_date, y = contbr_amt), data = illinois) +
  geom_point(alpha = 1/20)
```



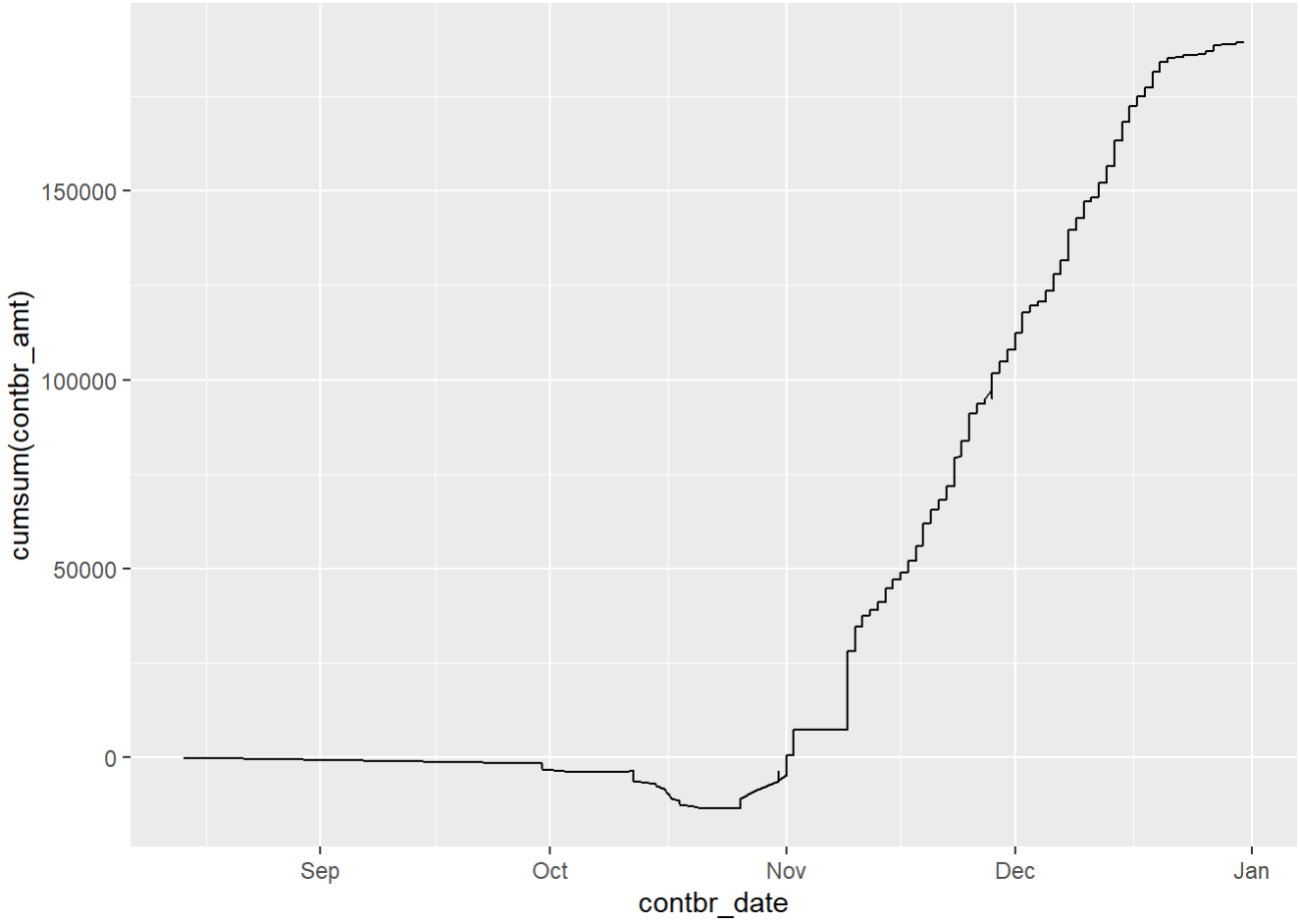

```
# Something potentially at early 2017
```

- Most donations happen at predicable levels every 25 dollars. Donations 100 and below seem to be where the vast majority of the data lies.

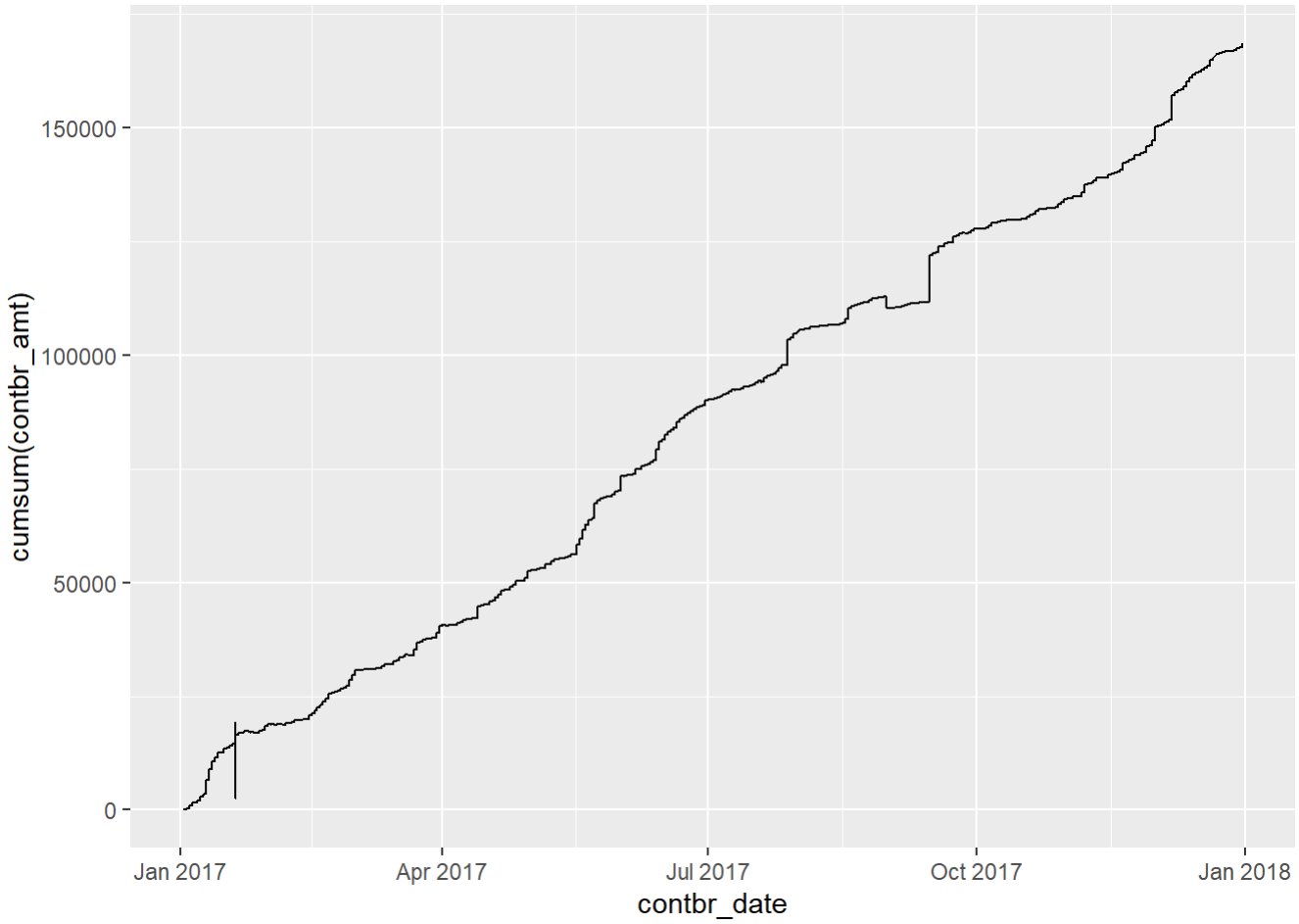
- As expected, the number of donations increase exponetially as time moves closer to the election.

- Hopefully I can find what's behind the 2017 anomoly shown in the final graph above.

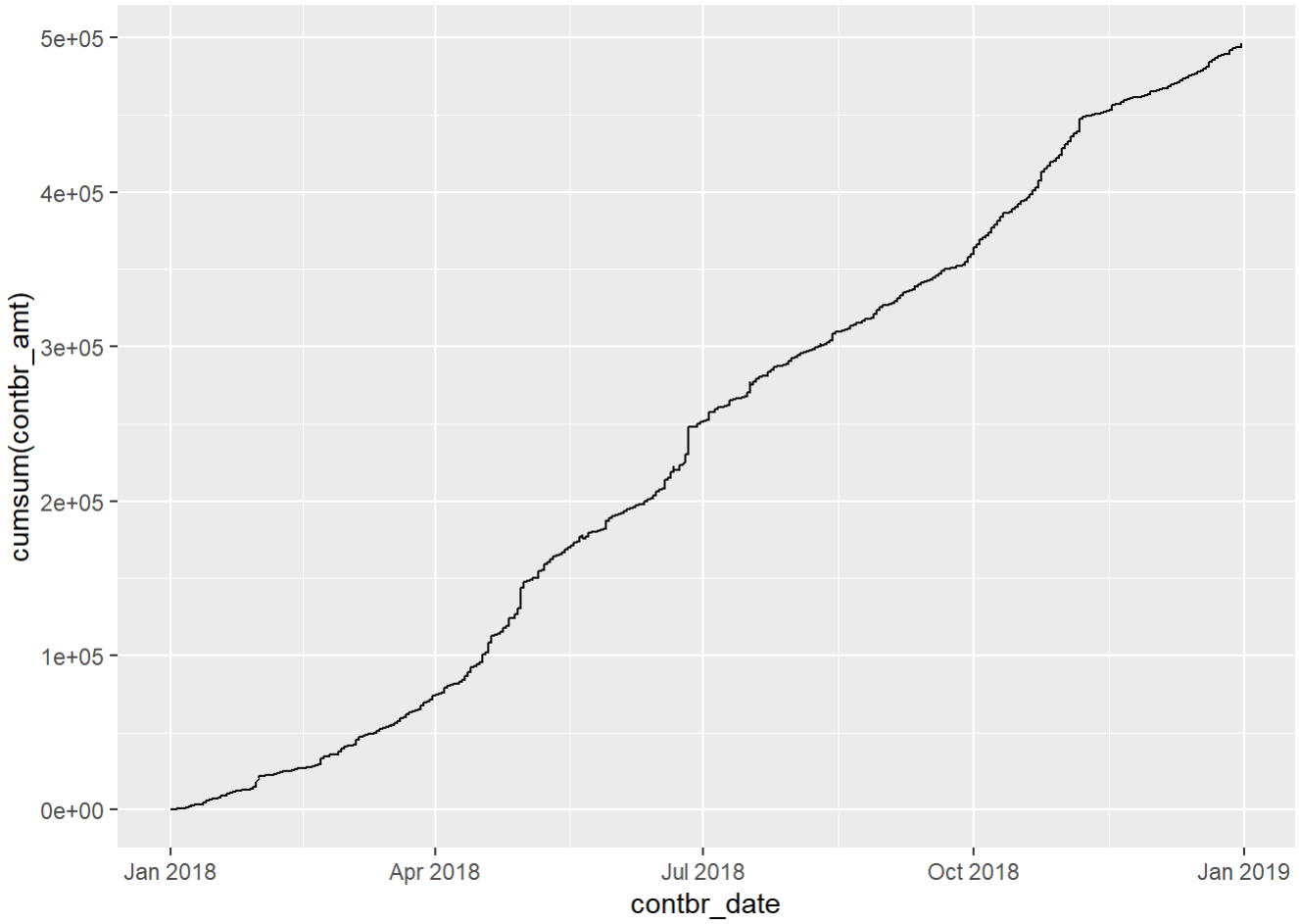
```
illinois <- illinois[order(as.Date(illinois$contbr_date, format = "%d-%b-%y")),]

# Subsetting all data by year
illinois.2016 <- subset(illinois, contbr_date < "2017-01-01")
# Interesting anomoly here in 2017
illinois.2017 <- subset(illinois, contbr_date < "2018-01-01" & contbr_date > "2016-12-31")
illinois.2018 <- subset(illinois, contbr_date < "2019-01-01" & contbr_date > "2017-12-31")
illinois.2019 <- subset(illinois, contbr_date < "2020-01-01" & contbr_date > "2018-12-31")

# Sorting the yearly data by date
illinois.2016 <- illinois.2016[order(as.Date(illinois.2016$contbr_date, format = "%d-%b-%y")),]
illinois.2017 <- illinois.2017[order(as.Date(illinois.2017$contbr_date, format = "%d-%b-%y")),]
illinois.2018 <- illinois.2018[order(as.Date(illinois.2018$contbr_date, format = "%d-%b-%y")),]
illinois.2019 <- illinois.2019[order(as.Date(illinois.2019$contbr_date, format = "%d-%b-%y")),]

# 2016
ggplot(aes(x = contbr_date, y = cumsum(contbr_amt)), data = illinois.2016) +
  geom_line()
```



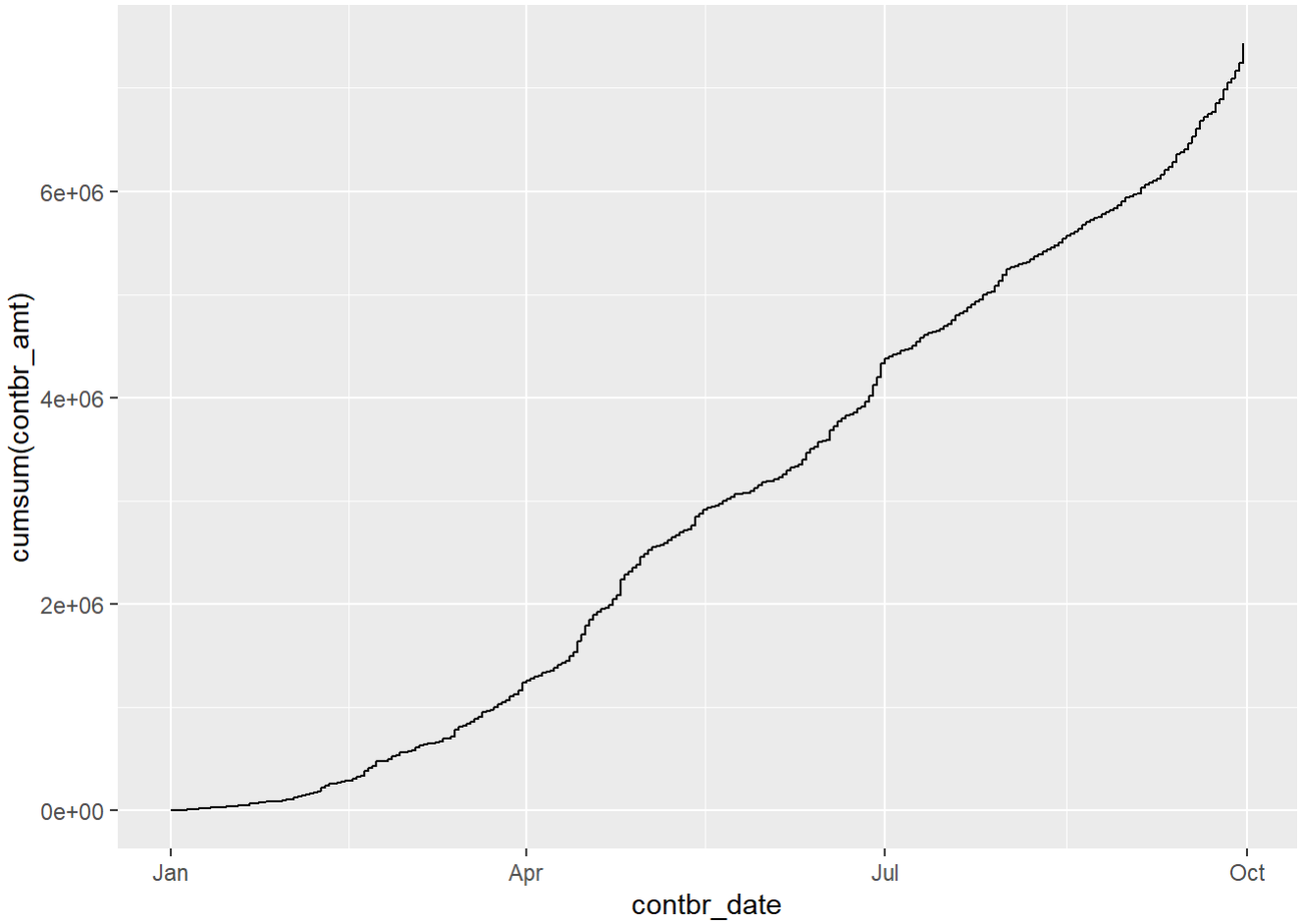

```
# 2017
ggplot(aes(x = contbr_date, y = cumsum(contbr_amt)), data = illinois.2017) +
  geom_line()
```




```
#2018
ggplot(aes(x = contbr_date, y = cumsum(contbr_amt)), data = illinois.2018) +
  geom_line()
```

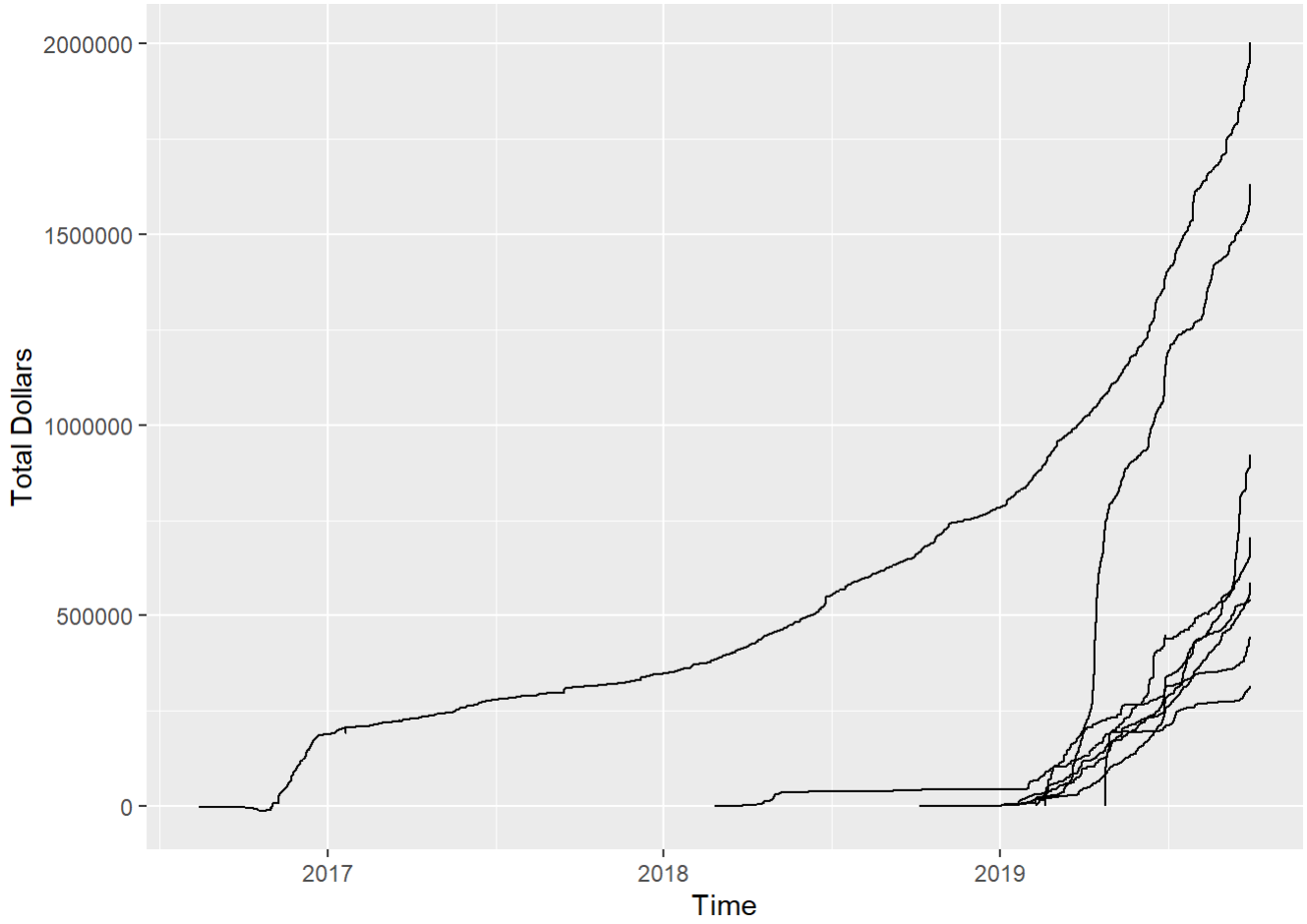



```
#2019
ggplot(aes(x = contbr_date, y = cumsum(contbr_amt)), data = illinois.2019) +
  geom_line()
```

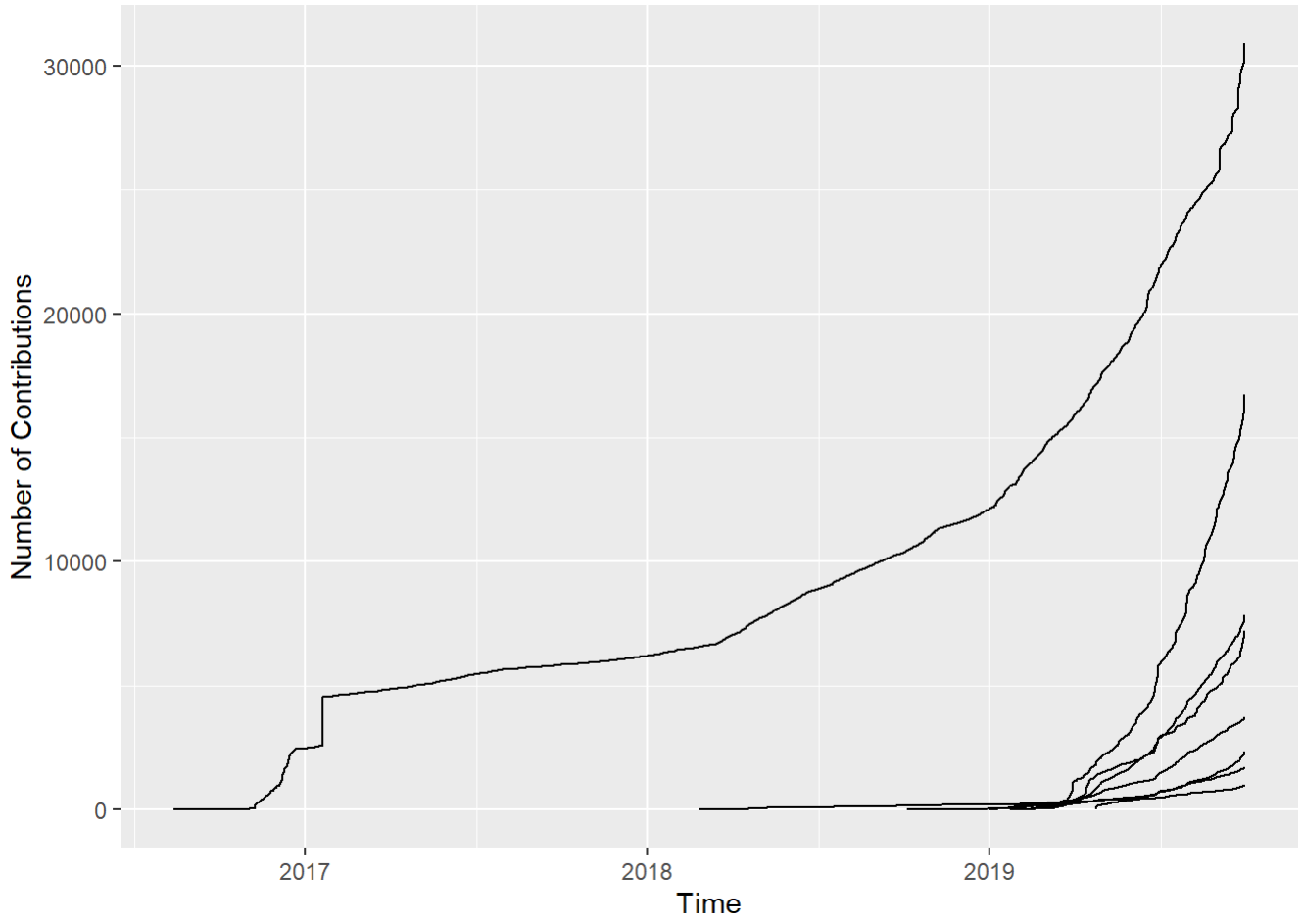


- The anomoly we see early in our 2017 graph happens to be inauguration day. There are a large amount of donations made on that single day alone, no doubt due to the President Trump’s supporters who hope to propel him to a 2020 win.

```
# Expanding upon the contributions over time data

# Filtering the candidates into individual datasets
# This is the only way I could figure out how to do the following, there is likely a more effici
ent method
trum <- filter(illinois, cand_nm == "Trump, Donald J.")
butt <- filter(illinois, cand_nm == "Buttigieg, Pete")
bide <- filter(illinois, cand_nm == "Biden, Joseph R Jr")
sand <- filter(illinois, cand_nm == "Sanders, Bernard")
warr <- filter(illinois, cand_nm == "Warren, Elizabeth ")
harr <- filter(illinois, cand_nm == "Harris, Kamala D.")
klob <- filter(illinois, cand_nm == "Klobuchar, Amy J.")
book <- filter(illinois, cand_nm == "Booker, Cory A.")

# Dollar amount of donations over time
ggplot() +
  geom_line(data = trum, aes(x = contbr_date, y = cumsum(contbr_amt))) +
  geom_line(data = butt, aes(x = contbr_date, y = cumsum(contbr_amt))) +
  geom_line(data = bide, aes(x = contbr_date, y = cumsum(contbr_amt))) +
  geom_line(data = sand, aes(x = contbr_date, y = cumsum(contbr_amt))) +
  geom_line(data = warr, aes(x = contbr_date, y = cumsum(contbr_amt))) +
  geom_line(data = harr, aes(x = contbr_date, y = cumsum(contbr_amt))) +
  geom_line(data = klob, aes(x = contbr_date, y = cumsum(contbr_amt))) +
  geom_line(data = book, aes(x = contbr_date, y = cumsum(contbr_amt))) +
  xlab('Time') +
  ylab('Total Dollars')
```



```
# Number of donations over time
ggplot() +
  geom_line(data = trum, aes(x = contbr_date, y = cumsum(counter))) +
  geom_line(data = butt, aes(x = contbr_date, y = cumsum(counter))) +
  geom_line(data = bide, aes(x = contbr_date, y = cumsum(counter))) +
  geom_line(data = sand, aes(x = contbr_date, y = cumsum(counter))) +
  geom_line(data = warr, aes(x = contbr_date, y = cumsum(counter))) +
  geom_line(data = harr, aes(x = contbr_date, y = cumsum(counter))) +
  geom_line(data = klob, aes(x = contbr_date, y = cumsum(counter))) +
  geom_line(data = book, aes(x = contbr_date, y = cumsum(counter))) +
  xlab('Time') +
  ylab('Number of Contributions')
```




```
# Filtering to only include the top 14 candidates by monitary contributions
illinois.top14 <- filter(illinois, cand_id == c("P80001571", "P60007168", "P00009621",
                                 "P00010298", "P00009423", "P80000722", "P80006117", "P00006486"
,
                                 "P00010793", "P00009795", "P00009183", "P00009910", "P00009092"
,
                                 "P00009290", "P00010454", "P00011833", "P00011999"))
```

```
## Warning in cand_id == c("P80001571", "P60007168", "P00009621", "P00010298", :
## longer object length is not a multiple of shorter object length
```

- One of the most interesting things I found in the data, expanded upon in final plots.

Final Plots and Summary:

```
# Contributions vs Number of Contributions
m1 <- ggplot(aes(x = reorder(cand_nm, total_contribs), y = total_contribs), data = top.half) +
  geom_bar(stat = 'identity', color = I('black'), fill = I('yellow')) +
  coord_flip() +
  labs(x = "Candidate Name", y = "Total Dollars")

m2 <- ggplot(aes(x = reorder(cand_nm, n), y = n), data = top.half) +
  geom_bar(stat = 'identity', color = I('black'), fill = I('yellow')) +
  labs(x = "Candidate Name", y = "Number of Contributions") +
  coord_flip()

grid.arrange(m1, m2, ncol = 1)
```

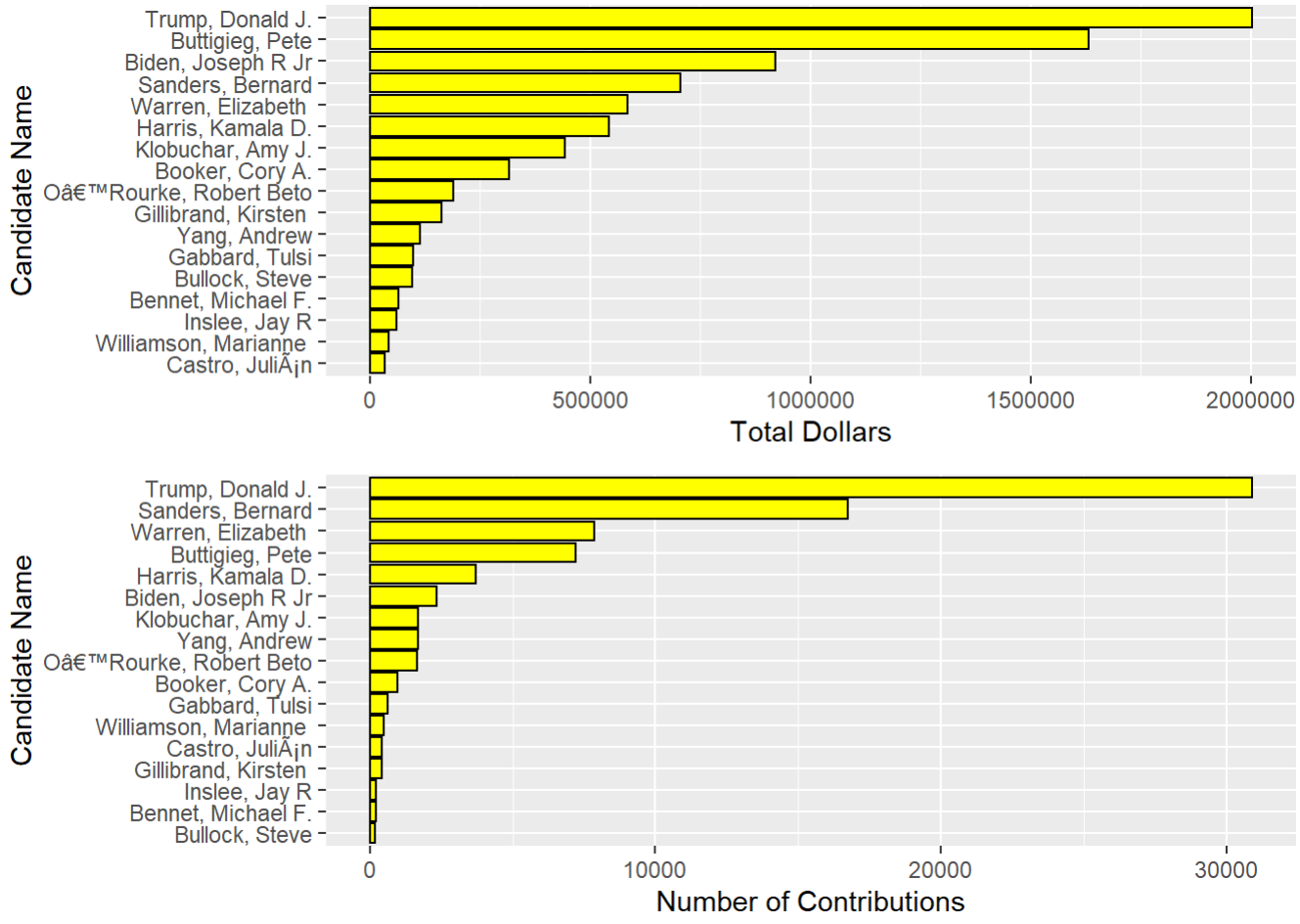


Figure 1

- Initial findings indicated donations and donation amounts were the big factors in this data, which is what the above attempts to convey. We are able to pick out who the frontrunners and real contenders are based on the donations they have recieved from Illinoisians.

```
# Most Prevalent Occupations and their Donations
occu <- ggplot(aes(x = reorder(contbr_occupation, n), y = n), data = illinois.top.30) +
         geom_bar(stat = 'identity', color = I('black'), fill = I('blue')) +
         labs(y = "Number of Contributions", x = "Occupation") +
         coord_flip()

dona <- ggplot(aes(x = reorder(contbr_occupation, donation_amt), y = donation_amt), data = illin
ois.top.30) +
         geom_bar(stat = 'identity', color = I('black'), fill = I('blue')) +
         labs(y = "Donation Amount in Dollars", x = "Occupation") +
         coord_flip()

grid.arrange(occu, dona, ncol = 2)
```

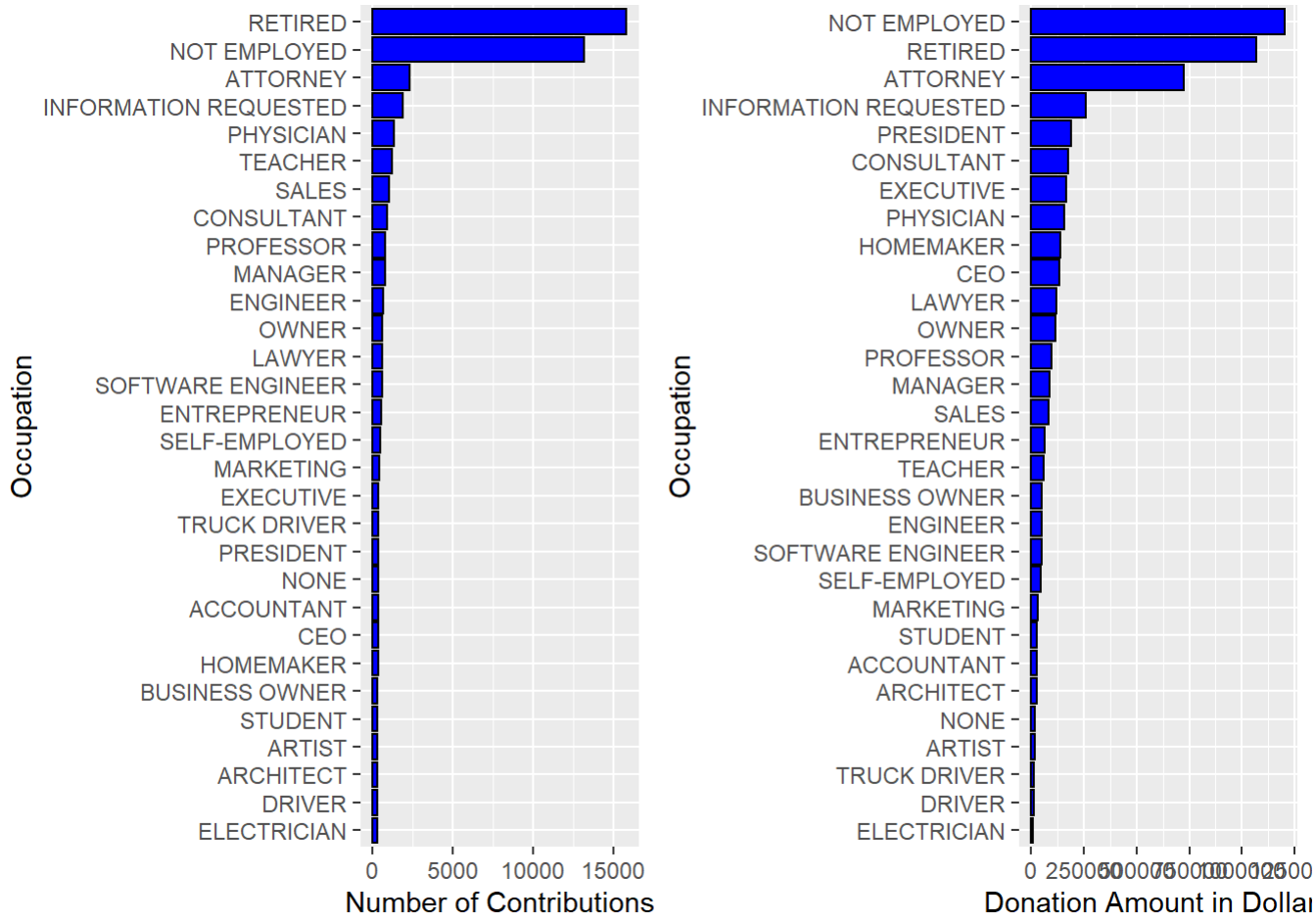


Figure 2

- Further expanding upon the donations given I wanted to know who they were coming from. As the above graph depicts, most donors are retired, unemployeed, or did not provide an occupation. While the employed are overshadowed by the retired in this dataset I must add a side note. If this data were to be properly cleaned, with occupations falling into broader fields rather than individual occupations, I believe we would be able to see trends that aren't represented by the data in its current state..

```
# Monitary Contributions to the Top 8 Candidates over time
ggplot() +
  geom_line(data = trum, aes(x = contbr_date, y = cumsum(contbr_amt), color = 'Donald Trump', si
ze = 'Donald Trump')) +
  geom_line(data = butt, aes(x = contbr_date, y = cumsum(contbr_amt), color = 'Pete Buttigieg',
 size = 'Pete Buttigieg')) +
  geom_line(data = bide, aes(x = contbr_date, y = cumsum(contbr_amt), color = 'Joseph Biden', si
ze = 'Joseph Biden')) +
  geom_line(data = sand, aes(x = contbr_date, y = cumsum(contbr_amt), color = 'Bernard Sanders',
size = 'Bernard Sanders')) +
  geom_line(data = warr, aes(x = contbr_date, y = cumsum(contbr_amt), color = 'Elizabeth Warren'
, size = 'Elizabeth Warren')) +
  geom_line(data = harr, aes(x = contbr_date, y = cumsum(contbr_amt), color = 'Kamala Harris', s
ize = 'Kamala Harris')) +
  geom_line(data = klob, aes(x = contbr_date, y = cumsum(contbr_amt), color = 'Amy Klobuchar', s
ize = 'Amy Klobuchar')) +
  geom_line(data = book, aes(x = contbr_date, y = cumsum(contbr_amt), color = 'Cory Booker', siz
e = 'Cory Booker')) +
  scale_color_manual(values = c('Donald Trump' = 'red', 'Pete Buttigieg' = 'blue', 'Joseph Bide
n' = 'darkblue',
                                'Bernard Sanders' = 'black', 'Elizabeth Warren' = 'orange', 'Kam
ala Harris' = 'brown',
                                'Amy Klobuchar' = 'pink', 'Cory Booker' = 'green')) +
  scale_size_manual(guide = 'none', values = c('Donald Trump' = 1.5, 'Pete Buttigieg' = 1.5, 'Jo
seph Biden' = 1.5,
                                'Bernard Sanders' = 1.5, 'Elizabeth Warren' = 1.5, 'Kamala Harri
s' = 1.5,
                                'Amy Klobuchar' = 1.5, 'Cory Booker' = 1.5)) +
  xlab('Time') +
  ylab('Total Dollars') +
  labs(color = 'Candidates')
```

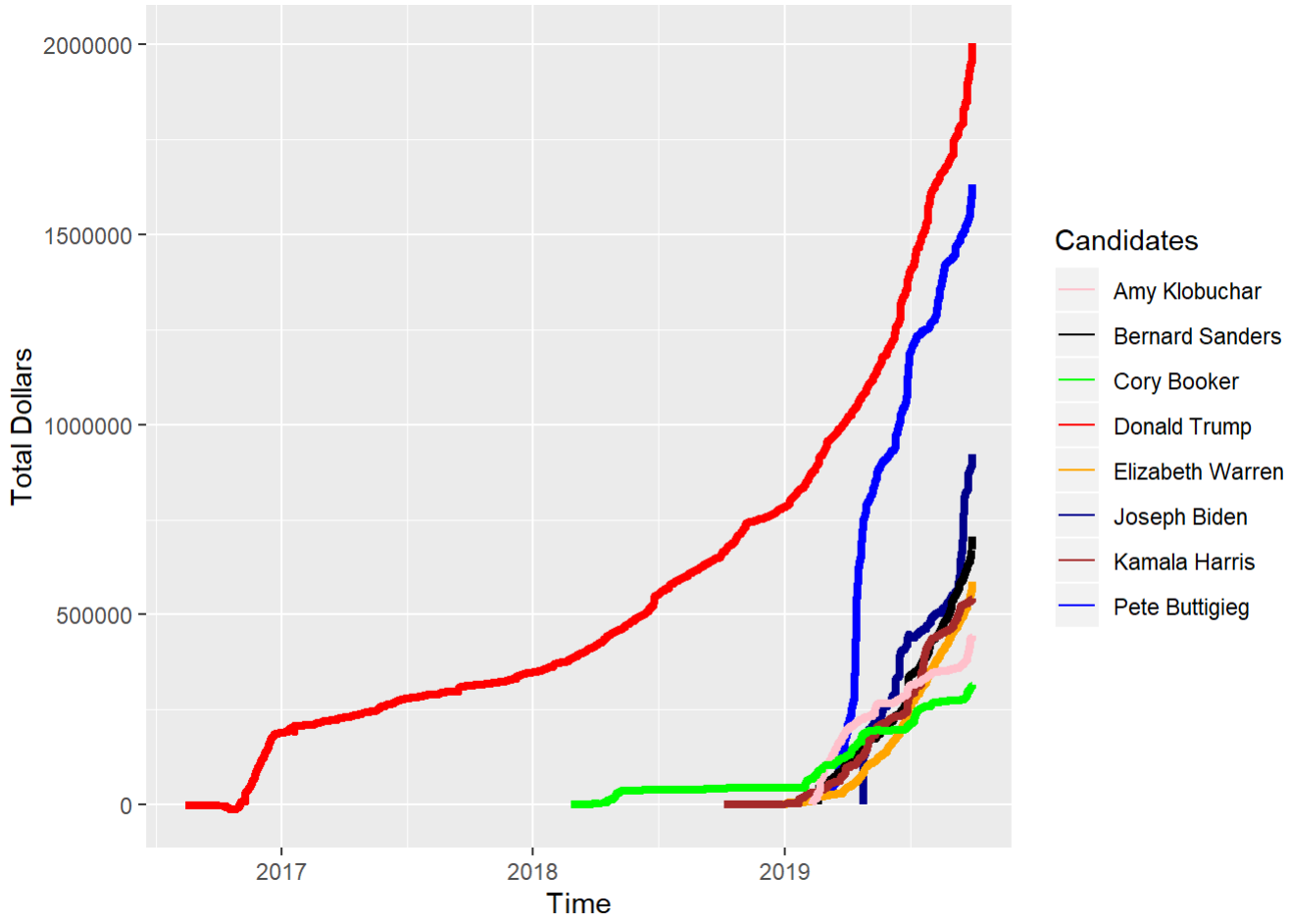


Figure 3

- Last but not least my final plot, and personally the most insightful point I found. I belive this graph clearly shows the advantage a sitting president has over his opponents, almost 2 full years of additional contributions. Contribution amount plays a large role in the success of a candidate in modern times, with 4 of the last 5 presidents winning their respective re-elections. This statistic cannot be taken lightly.

Reflection:

The Illinois dataset contains over 77,000 records of contributions made to 28 candidates. After initial exploration I concluded the most telling data points would revolve around contributions, their amounts and quantities.

The trends I found were mostly straight forward and exactly as I would expect, given a few. The most popular candidates in the media were clearly represented as the favorites by donation amounts as well. Some less popular candidates will perform better simply due to the fact that this data is from the state of Illinois, and all states will have outlier candidates which are only popular to a particular state or surrounding area. I quickly noticed the top half of the candidates, both in total dollars and number of contributions, were of the most interest and I focused my attention there.

This model does have its share of limitations. This data is constantly being updated as more and more donations are made each day, possibly to even more candidates. It is also limited in prediction power due to the existance of the other 49 states and territories

donations would be coming from. I would be interested in comparing my final plot, detailing contributions over time for top candidates, to google search trends for the same candidates over time. How do people learn about these candidates? What do they see or hear that compells them to donate to a particular candidate? Can we tie these contributions to specific events in the news or media? These would be interesting points to explore.