Ideology and Campaign Finance

Adhitya Venkatraman 12/10/2021

Introduction

In this project, I examine data on the flow of donations in presidential elections from 1980-2016. I ask three key questions:

1. Across presidential elections from 1980-2016, who are 25 the most prolific donors and what are their political views?
2. Do “small” donors contribute less often than “large” donors?
3. What are Corporate Leaders’ donation patterns and how do they vary from the general population?

While my results are limited to the context of the data, I make a number of interesting findings. First, the most prolific donors seem to have more extreme viewpoints and ideologies. Second, there seems to be a negative relationship between the size of contributions and the number of contributions made. This relationship seems to decrease logarithmically. Finally, I find that elites donate more than non-elites on average. I also find that the elites studied may make donations based on the companies they are a part of, and that the distribution of political ideology may be less extreme among elites than it is in the general population. I begin this report by discussing the data, defining key variables and terms, and providing the source of the data. Then, I offer a reflection on the ethical use of this data and my thought process and approach to dealing with those challenges. Then, I conduct my analysis, starting with importing the data, then tidying and cleaning the data, and finally exploring each of the questions above in depth. I conclude by summarizing my findings, acknowledging major limitations, and offering recommendations for future work.

Data Discussion and Source

I gained an interest in this data after learning about how campaign finance and partisanship plays a crucial role in modern politics. Stanford Professor Dr. Adam Bonica’s Database on Ideology, Money in Politics, and Elections offers a resource to understand these dynamics. Dr. Bonica’s work is part of a broader literature focused developing a “comprehensive ideological mapping of political elites, interest groups, and donors” (Bonica). This particular

dataset focuses on contributions to presidential candidates in election cycles from 1980 to 2016. The raw data were collected from public data made available by the Federal Election Commission (FEC), which requires that campaign committees make information about donors public. This includes demographic data on the donor, including their name and employer, as well as the party and committee to which they donated. Trade political action committees that enforce membership or consistent dues are not included because they may not simulate voluntary, actual voting preferences. Dr. Bonica also uses a scoring method to measure the ideology of candidates and contributors on a continuous scale. The scores can be understood as follows:

- a score above 0.5 indicates strong conservatism
- a score between 0 and 0.5 indicates weak conservatism
- a score between -0.5 and 0 indicates weak progressivism
- a score less than -0.5 indicates strong progressivism

These scores are available for most contributors and all candidates. These **CF Scores** (Campaign Finance Scores) are centered around zero, with negative values denoting an ideology that is more progressive, while a positive score suggests a conservative ideology. Put simply, Bonica estimates the views of donors based on who they vote for, while candidate scores are based on who votes for them. His methods builds on the work of Keith Poole and Howard Rosenthal whose NOMINATE scores laid the groundwork for Bonica's methodology. In all, the dataset is comprised of 12,353,166 rows and 46 columns. This data can be found at: https://data.stanford.edu/dime#download-data. An accompanying data dictionary/codebook can be found here: https://dataverse.harvard.edu/file.xhtml?fileId=2865308&version=2.2

I also use data on Fortune 500 CEOs and Executives' contributions collected by Dr. Bonica in the third portion of my analysis. One of his past studies was focused on comparing these corporate elites corporate political action committees. These data include all contributions made these individuals for all campaign types. It is structured the same way as the above dataset. *The only difference is that it contains a separate column for the corporation name and ticker*. This is important because some executives may list their employer as a charity, rather than their corporation, when reporting that information. I will be joining this data onto the presidential data to isolate their behavior in presidential elections. These data are collected in a similar manner to the above data, but are of course focused on public information about the names of CEOs and Executives at Fortune 500 companies. This data can be found here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6R1HAS

Data Ethics

The issue of data ethics is central in campaign finance. Traditionally, the discussion has revolved around the need for greater transparency and for good reason. Without a clear understanding of how politicians are financing their campaigns, corruption may fester. Without provisions for transparency, unethical people in the political process could bribe legislators and threaten the ideal of democracy as a government responsive to the needs of all constituents. Alternatively, perhaps a representative funded by ultra-wealthy donors becomes loyal to a small minority that enables her to win elections, rather than pursuing the interests of the constituents. Politicians might say all the right things during the campaign, but follow the money once they reach office and implement policies that will allow them to raise money for their next campaign. For these reasons, the Untied States Congress implemented reforms that require federal candidates, political parties, and political action committees to disclose all donations and capped the amount any individual can contribute to a political party. The Federal Election Commission (FEC) requires that this data be made publicly available for any citizen to explore. In this sense, campaign finance data could be thought of as a public good that allows citizens to learn more about their representatives and the political landscape. The FEC prevents the commercial use or sale of this data; however, it is of course intended for use in research and educational purposes.

Yet, there might also be ethical concerns about this data being disclosed. On one hand, unless there is clear and informed consent from donors, it may be problematic for names, addresses, or employers to be disclosed. While the committees and FEC strives to make this the case, perhaps there are slippages in some instances where people are just learning about American politics or the websites themselves do not make the disclosure process clear enough. Another concern is that we live in a highly polarized political moment. As we saw on January 6th 2021 at the Capitol Riots, political affiliations can lead people to violence. Public information about political leanings could be used to target individuals of a certain viewpoint. While I personally believe that the solution here is not to restrict information, but instead to reduce the dangerous levels polarization gripping the country, this is a risk that must be acknowledged in the current political climate.

To see how transparent committees and candidates are, I visited a few contribution webpages on both sides of the aisle and looked at the process to make a donation. Across the board, I was encouraged to see that each of them explicitly acknowledged that this information was being requested to make a donation and was required by campaign finance law. Some of them were very clear and stated in large text that the data would be disclosed, while others had it in the fine print below the main content. For the latter group, I

believe it should be required to forefront the fact that this data is disclosed by the party and released to the public. This is important to ensure that all parties are informed and provide their data with proper consent. While the committees do not seem to be doing anything malicious, it is important to recognize that in a digital age, it is much easier to access this data and the ramifications of personal data being public are potentially heightened.

I debated the best way to proceed with this project and ultimately decided that it would be reasonable to use name data here. This is partially because there seem to be robust mechanisms for gaining consent and the data are being used as they are intended. Before giving, it is made clear that there These data are collected for transparency and to enhance public knowledge of where their fellow citizens stand. By learning how different groups exert influence on the political process, this project aims to do just that. However, I do not use location data here that could potentially be used to target individuals. Instead, I focus on simply drawing connections between people and their political preferences. Moreover, I exclusively use data relating to relatively public individuals such as presidential candidates and well-known executives and CEOs. I understand a public individual as someone whose views are generally known by the public, as is certainly the case for the former group and often for the latter. I believe it is important, perhaps in today's political climate more than ever, to create a dialogue and environment where people are not afraid to express their political views and engage in robust political debate. We lose out on critical discourse when we keep our thoughts to ourselves and don't challenge, question, and learn from each other. We certainly have much work to do on this point as a society, but ongoing efforts to normalize a healthy political discourse is essential in this.

Data Import

Before starting the analysis, we can load in the packages we will need.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## x dplyr::lag()    masks stats::lag()
```

```
library(modelr)
```

I begin this analysis by importing our data. There are a few important considerations. First, certain variables, like **cycle**, which refers to the election cycle are listed as numbers, but are really categorical variables. Therefore, we should import this variable as a factor. The data include contributions that were made during primary elections that happen in the middle of a four-year presidential term. Primary elections are designated as happening every two years, so we need factor levels for every other year from 1980 to 2016, inclusive. Line 40 includes the code for importing this variable as a factor. We have a similar issue with the **recipient.party** variable, which refers to the candidate's party affiliation. Here, 100 refers to Democrats, 200 to Republicans, and 328 to Independents. We'll first convert these variables to factors, and then mutate their names to Democrat, Republican, or Independent afterwards. Lines 41 and 42 perform this task.

I will also make the **contributor.state** variable, which designates the state of the donor, a factor. This will allow for more streamlined aggregation along state lines later. I follow the same process as before, defining the variable as a factor in the read_csv() command.

Of note is that there is one listed problem with the parsing process; this is attributed to issues with the version of R being used. Despite reading in the date value in Column 2 correctly, R believes it has come across an error. An examination of the data clearly shows that the correct date is listed. Thus, this parsing issue can be ignored and we can proceed with the analysis.

```
donations<-read_csv('president.csv',
                    col_types = cols(cycle = col_factor(levels =  c('1980','1982','19
                                     recipient.party = col_factor(),
                                     contributor.state = col_factor())) %>%
  mutate(recipient.party = fct_recode(recipient.party, Democrat = "100", Republican =
donations
```

```
## Warning: One or more parsing issues, see `problems()` for details
```

```
## # A tibble: 12,353,166 x 46
##    cycle transaction.id           transaction.type amount date       bonica.cid
##    <ord> <chr>                    <chr>             <dbl> <date>          <dbl>
##  1 1984  comm:1984:111773         24K                 100 1984-09-24          1
```

```
##  2 1986  comm:1986:118145          24K                -100 1985-01-18          1
##  3 2012  e:ind:2012:12120806464849 15                  500 2012-10-25          4
##  4 1980  comm:1980:76130           24K                 250 1979-08-19         26
##  5 1984  comm:1984:111794          24K                 500 1983-08-04         26
##  6 1984  comm:1984:36              24K                1000 1983-01-25         26
##  7 1984  comm:1984:41              24K               -1000 1983-01-25         26
##  8 1984  comm:1984:75              24K                 500 1983-04-25         26
##  9 1988  comm:1988:111113          24K                 250 NA                 26

## 10 1988  comm:1988:232009          18K                 250 NA                 26
## # ... with 12,353,156 more rows, and 40 more variables: contributor.name <chr>,
## #   contributor.lname <chr>, contributor.fname <chr>, contributor.mname <chr>,
## #   contributor.suffix <chr>, contributor.title <chr>,
## #   contributor.ffname <chr>, contributor.type <chr>, contributor.gender <chr>,
## #   contributor.address <chr>, contributor.city <chr>, contributor.state <fct>,
## #   contributor.zipcode <dbl>, contributor.occupation <chr>,
## #   contributor.employer <chr>, is.corp <chr>, recipient.name <chr>, ...
```

Data Cleaning and Tidying

Next, we can begin to clean and tidy our presidential contributions data. This data is largely already tidy. Although the **transaction.id** variable seemingly refers to more than one value, it is included as a primary key to uniquely identify each donation in the dataset; I will actively choose to leave this as is because these different values come together to create a new value: a primary key. Only one variable needs to be corrected: **seat**, which refers to the level of government and the office being sought. To ensure our data is tidy, these values should be separated into two columns, which will be called **election_type** and **seat**. We can use the separate() function to perform this operation because it will split the original **seat** column into two new columns by identifying the separator ":".

```
donations <- donations %>% separate(seat, into = c("election_type","seat"), sep = ":"
```

Then, we will join the data on CEO contributions, which will be called **fortune**, with the **donations** dataset. Because it is at the transaction level, **transaction.id** is a primary key and foreign key for both of these tables. Therefore, we can use it to join. We only need to extract the ticker and corporation name from the **fortune** database, so we can simply select these values along with **transaction.id**. Because we are interested in adding data on the CEO's to our dataset on political contributions, we can simply perform a left join. By adding relevant information to the correct transactions, we will then be able to identify the CEOs

and executives in the dataset without losing any information on the remaining of the sample of contributions. Finally, I will add a binary variable to indicate if the transaction was made by a Fortune 500 CEO/Executive. Because all transactions not performed by one of these individuals will have an “NA” value in the **ticker** column, I can simply add a new column **is.fortune**, that takes on a value of 1 when there is no “NA” in the former and a 0 when there is.

Additionally, some of the values in our **amount** column, which measures the size of each donation are negative. There are not many of these values. While I could not find an explicit reason for this in the data dictionary, election committees sometimes refund donations, and they must provide a record of this (https://www.fec.gov/help-candidates-and-committees/taking-receipts-political-party/refunds-contributions/). Since these donations are not actually used by the committees or candidates, I will filter out these negative values. This is a critical assumption, but one that seems to have a reasonable grounding in evidence.

```
fortune <- read_csv("ceo.csv", col_types = cols(ticker = col_factor(), corpname = col
donations <- donations %>% left_join(fortune) %>% mutate(is.fortune = ifelse(!is.na(t
```

```
## Joining, by = "transaction.id"
```

Finally, I will make a few cleaning transformations to the data on the basis of aesthetics. For example, the **contributor.name** variable has inconsistent case. Standardizing these words to title case seems like the most reasonable choice here. Admittedly, there will be some abbreviations, like “NRA” (National Rifle Association), that will be made into “Nra”, but this should still distinguishable. Moreover, this is better than some abbreviations being capitalized, others being in lower case, and yet others being in sentence case. This thus makes things more readable and standardized.

```
donations$contributor.name <- str_to_title(donations$contributor.name)
```

Research Questions and Analysis

The analysis focuses on three key questions:

1. Across presidential elections from 1980-2016, who are 25 the most prolific donors and what are their political views?
2. Do “small” donors contribute less often than “large” donors?
3. What are Corporate Leaders’ donation patterns and how do they vary from the general population?

In the subsections below, I explore each of these questions in detail.

a) Across presidential elections from 1980-2016, who are 25 the most prolific donors and what are their political views?

I define so-called “prolific” donors in two ways: those who made the highest number of donations and those who donated the greatest amount of money. I explore the most prolific donors under both understandings and offer analysis for each.

Most Frequent Donors

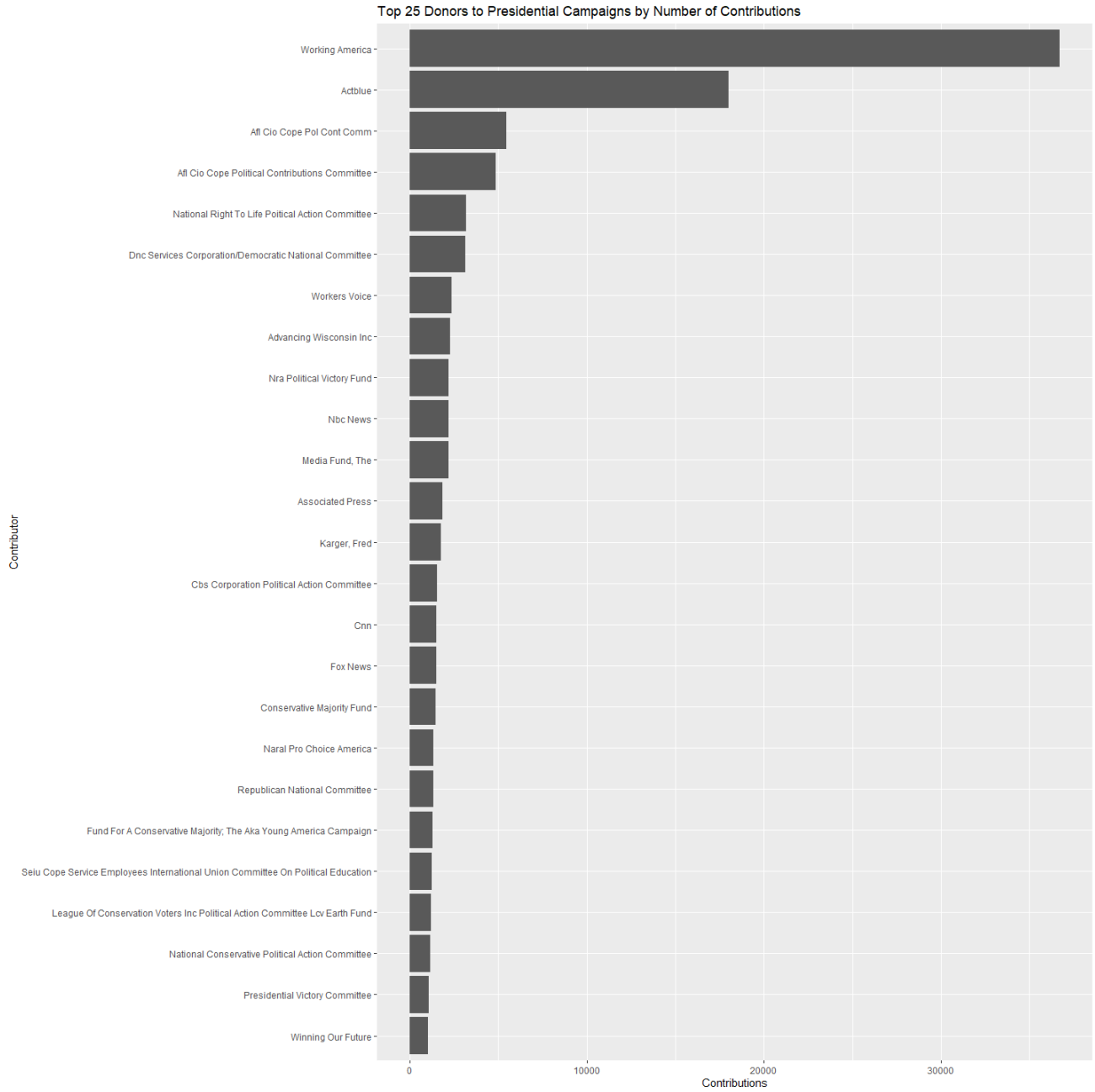
We can begin with the first interpretation. That is, the donors who made the highest number of contributions across the period. This is an interesting measure because it captures which donors are engaged and regularly contributing. I begin by grouping my data on the basis of the **bonica.cid** variable, which is a unique identifier for each contributor. After grouping, we can begin generating information and statistics on each donor. First, because donors self-report their names when filling out their form, some donors change the way they spell their name, include a middle name, or other idiosyncrasies in their donations over time. Therefore, I begin by summarizing the **contributor.name** as the first name value appearing for a given donor. This avoids double-counting the same individual. Keeping in line with the ethical demands of a project like this, the names that could reasonably included in a list like this will be large political action committees, and donations by committee Then, I count the number of contributions made by each donor in a variable called **Contributions**. In addition to **Contributions**, we are also interested in the ideology these groups are most aligned with. So, we can add the **contributor.CFscore** for each donor as well. While each donor has the same CF Score that is computed across all of its donations, we will take the mean the **contributor.CFscore** because it ensures that each contributor is associated with a single instance of the value (rather than several repeated instances of the same number) without altering the value itself; I call this variable **Ideology**. I also calculate the total value of all donations contributed by an individual as **Total_Value_Donations**. Finally, I arrange

our table in descending order of **Contributions**. Finally, because we are interested in just the most prolific donors, I select the 25 donors with the greatest number of contributions.

```
prolific_donors1 <- donations %>% group_by(bonica.cid) %>% summarize(contributor.name
```

We can now visualize these results. First, we can visualize these results using a bar graph mapping **contributor.name** against **Contributions** to highlight the relative differences in donations. Flipping the coordinates allows for better readability, given the length of these donors' names using the coord_flip() command. Turning **contributor.name** into a factor and ordering it by the amount of contributions makes the bar graph further readable as it is clear which donors contributed more than others. In all plots, I use the labs() command to provide labels for them. This plot shows that the most prolific donor during the period is Working America. This is a somewhat intuitive result, given that Working America is the largest non-union association of workers in the US. However, the disparity is rather surprising. With more than 30,000 donations, it made more contributions than the next several donors combined. We also see that two AFL-CIO affiliated organization round out the top four. More broadly, out of the top seven, four are related to workers issues: Working America, the two AFL-CIO organizations, and Workers Voice. This suggest that workers movements have been a seemingly regular source of donations. We also have one individual here: well-known political consultant and former presidential candidate Fred Larger. As expected, plots like this avoid the inclusion of any non-public individuals.

```
prolific_donors1 %>% ggplot() + geom_col(mapping = aes(x= fct_reorder(factor(contribu
```



While Working America has the most donations across the entire period by far, we might be curious about the distribution of these contributions over time. In particular, are they clustered among a few election cycles? We can group the original **donations** dataset by **contributor.name** and **cycle**. Then, we can find the total number of contributions, as we did above, but this time for each cycle. Let's start by generating a table for this data. Then, we can repeat the process for the second-place donor, Act blue. Incredibly, Working America made all but 3 of its 36,723 donations in the 2012 cycle alone. That being said, those three donations were rather large, totaling about a fifth of the donations earned in 2012. By comparison, Act blue contributed mostly in 2012 and 2016. It made just three donations in 2008, but for small amounts. The preponderance of donations in these more recent cycles may be the result of online giving allowing certain organizations to reach a

broader base. This could alternatively point to the incompleteness of the data; as previously mentioned, political action committees that are subsets of union or due-paying organizations are excluded. Perhaps a change in Working America's status in 2012 is the reason why it is included, but does not appear in any previous cycle. Moreover, because organizations may be started in the middle of the timespan of the sample, it may be unreasonable to compare donation habits across election cycles. While a full exploration of these issues is outside the scope of this analysis, I hope to explore these matters in the future.

```
donations %>% group_by(bonica.cid, cycle) %>% summarize(contributor.name = first(cont
```

```
## # A tibble: 2 x 5
## # Groups:   bonica.cid [1]
##   bonica.cid cycle contributor.name Contributions    sum
##        <dbl> <ord> <chr>                   <int>  <dbl>
## 1   90011156 2008  Working America             3 201246
## 2   90011156 2012  Working America         36720 937377
```

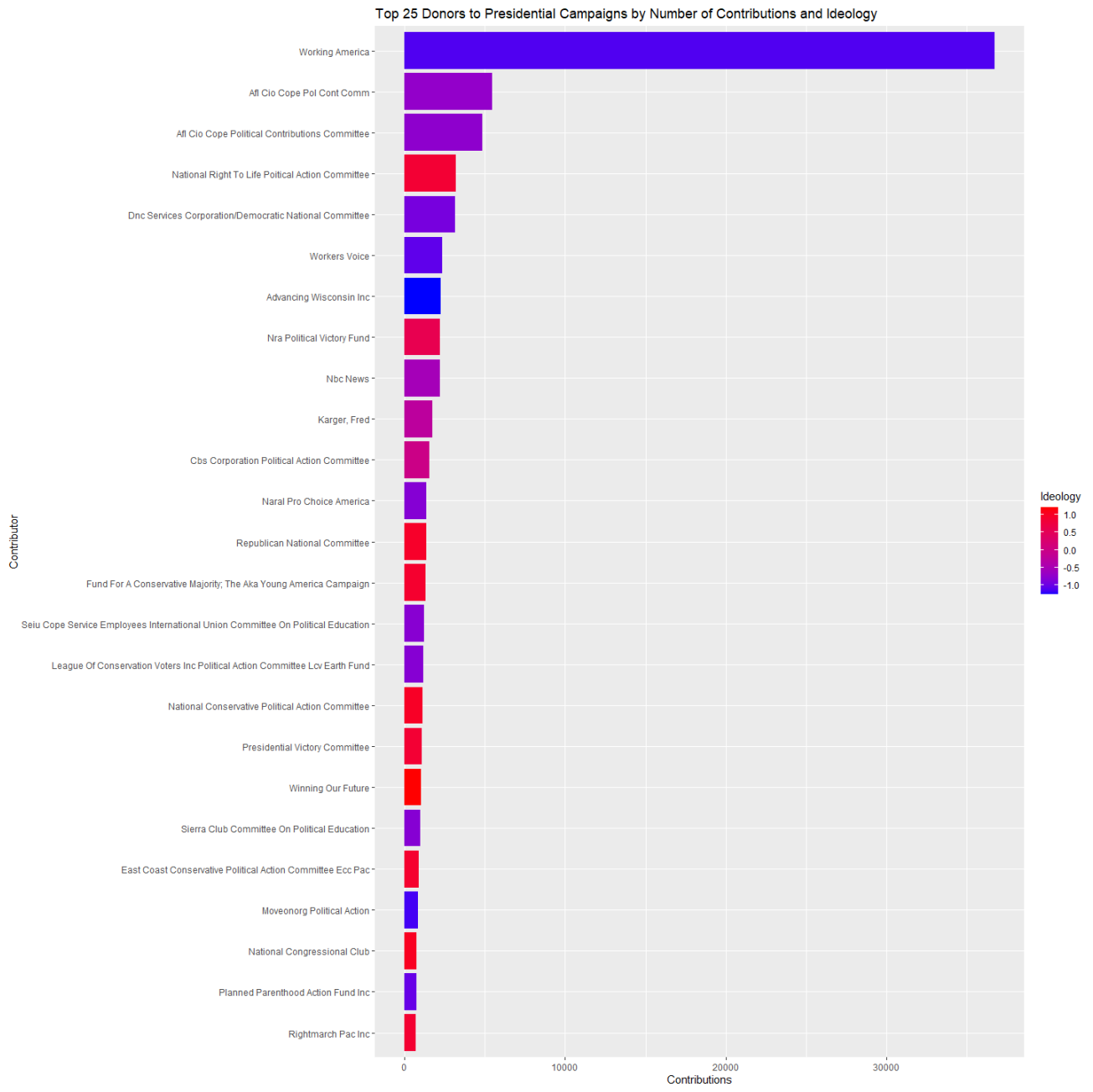
```
donations %>% group_by(bonica.cid, cycle) %>% summarize(contributor.name = first(cont
```

```
## # A tibble: 3 x 5
## # Groups:   bonica.cid [1]
##   bonica.cid cycle contributor.name Contributions      sum
##        <dbl> <ord> <chr>                   <int>    <dbl>
## 1     401224 2008  Actblue                     3    1600
## 2     401224 2012  Actblue                  4427  519474.
## 3     401224 2016  Actblue                 13594 2734348.
```

Next, we can construct a similar plot to the one above, but we will now can color each bar by the donor's ideology as measured by their CF Score. We will follow a similar process as above, generating a table of the top 25 contributors by number of donations made, but this time filtering out those that do not have an ideology CF Score. As previously mentioned, some donors, even large ones such as Actblue, do not have a CF Scores. We then group by the unique identifier and aggregate the statistics we are interested in: total number of contributions, the ideology score, and the names of contributors. We then adjust the fill color to correspond to the **Ideology** variable in our summary table and add a gradient to make progressive-leaning donors blue and conservative-leaning donors red. There are a

number of interesting findings. Firstly, of the top seven donors in this plot, just one is solidly red, perhaps suggesting that conservative funds may not be captured by plots that emphasize the volume of donations received. That being said, several organizations toward the bottom of the plot do lean toward conservative ideology.

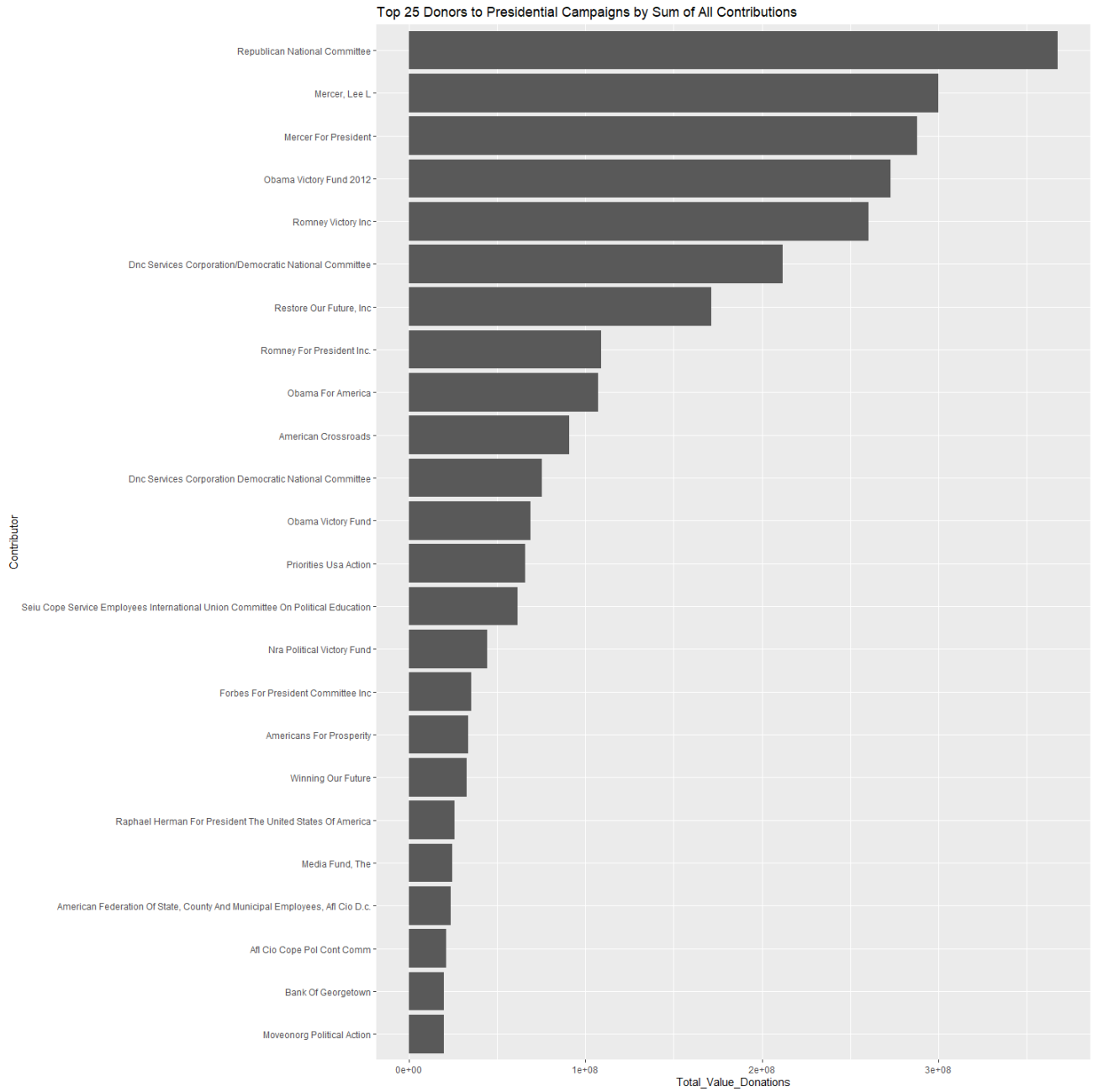
```
prolific_donors1_scores <- donations %>% filter(!is.na(contributor.cfscore)) %>% grou

prolific_donors1_scores %>% ggplot() + geom_col(mapping = aes(x= fct_reorder(factor(c
```



Most Valuable Donors

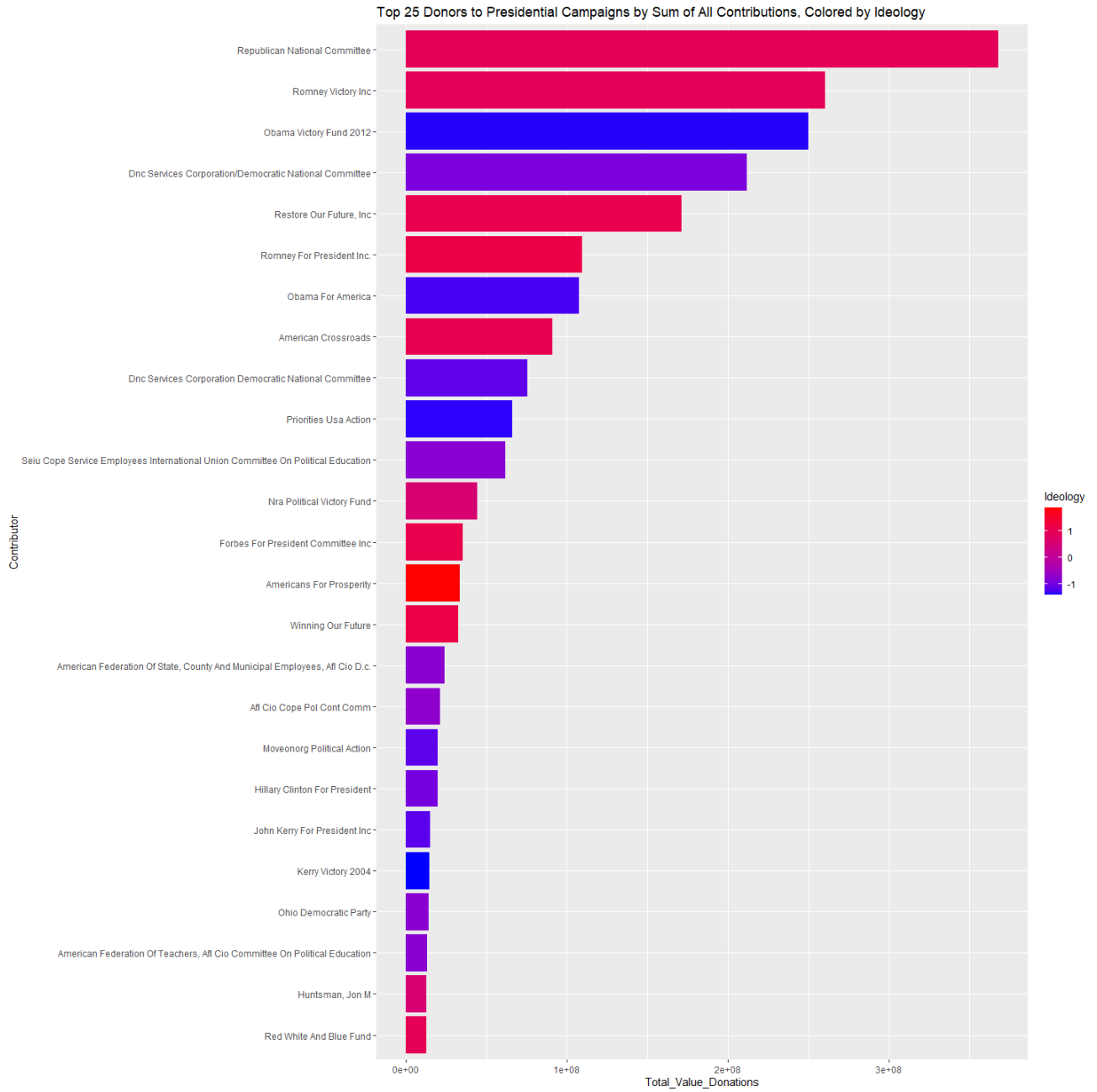
Next, we can change our interpretation of prolific donors to mean those who donated the most across the period. We once again group our data by the unique identifier for donors, but now aggregate the total value of all contributions made by that donor into **Total_Value_Donations**. Then, we can get their name in **contributor.name** and their CF Score in **Ideology**. Once again, we'll limit our table to the top 25 donors to get a sense of which ones are the most important. With this table, we can create an initial bar plot of each donor's total contribution over the length of the sample. As before, converting **contributor.name** into a factor ordered by each donor's value for **Total_Value_Donations** allows for better readability. For an unknown reason, I encountered issues when reordering the factor by **Total_Value_Donations**, as the Obama Victory Fund 2012 would consistently be out of place, so I instead sort my initial table by these sums in descending order and use the row_number() function to create a column of consecutive numbers. I then use these values to achieve my intended ordering for the factors and hard-code the Obama Victory Fund 2012 into the 4th position, which is where it should be. I also flip the axes to have names on the y-axis and the donation sums on the x-axis for greater readability. While there are no shocking disparities here among the top few contributors, there are still a few interesting things to consider. First, the most prolific donor by this measure is the Republican National Committee. Additionally, it is interesting that so many of these donors are clearly from the 2012 election cycle. For example, four of the donor names mention either Romney or Obama. This once again suggests that the most prolific donors are perhaps spurred on by changes that election cycle, whether through online donations or the Citizens United Supreme Court Ruling, which allowed corporations to give as people, opening the doors to greater contribution totals. Interestingly, two of the top three donors are also related to Lee Mercer, who is a failed businessman and former presidential candidate. It is somewhat shocking to see his totals this far up in the data and on par with the amounts raised by the entire Republican National Committee. This data point certainly seems like an error of some kind in the original data, or a major anomaly. I conducted some external research and other sources also seem to suggest that the Mercer for President political action committee did raise a fair amount of money (https://www.campaign money.com/political/campaigns/lee-l-mercer-jr.asp?cycle=08). Perhaps he was transferring money between his personal accounts and his political action committee and creating large numbers, as he is the only person who seems to have donated to his campaign. It is also interesting to see the difference between the Republican National Committee and Democratic National Committee, where the former exceeds the latter by a larger margin than I had expected. Perhaps the Democratic Party has other arms to fundraise, like the DNC Services Corporation (which ranks 11th) so it does not direct as much of its fundraising through the main committee.

```
prolific_donors2 <- donations %>% group_by(bonica.cid) %>% summarize(contributor.name

prolific_donors2 %>% ggplot() + geom_col(mapping = aes(x= fct_rev(fct_relevel(contrib
```



Next, we'll once again create a new table containing the top 25 donors who have CF Scores. We once again aggregate the sum of their donations in **Total_Value_Donations** and put their their CF Score in **Ideology**. Notably, the Mercer organizations fall out of the table, as no CF Score is assigned to those contributions. The plotting process follows from the previous figures, with the obvious change that we will now color the plot by CF Scores from the **Ideology** variable. Once again, a gradient is used to represent changes from conservative to progressive ideology.

```
prolific_donors2_scores <- donations %>% filter(!is.na(contributor.cfscore)) %>% grou

prolific_donors2_scores %>% ggplot() + geom_col(mapping = aes(x= fct_reorder(factor(c
```



Are there any overlapping donors between our two methods? We can answer this question by utilizing a inner join between our two types of tables of prolific donors. This is because an inner join matches on the values shared between two tables. Let's start by examining the overlap between all donors, regardless of ideology.

```
inner_join(prolific_donors1, prolific_donors2)
```

```
## Joining, by = c("bonica.cid", "contributor.name", "Ideology", "Total_Value_Donatio
```

```
## # A tibble: 7 x 6
##    bonica.cid contributor.name     Contributions Ideology Total_Value_Don~ order
##         <dbl> <chr>                        <int>    <dbl>            <dbl> <int>
## 1  3805309992 Afl Cio Cope Pol Co~          5455    -0.71         21240560    23
## 2 33580613966 Dnc Services Corpor~          3155    -0.92        211839345     6
## 3       53553 Nra Political Victo~          2210     0.58         44369368    15
## 4    30000053 Media Fund, The               2186   NaN            24617780    20
## 5        3418 Republican National~          1346     0.96        367765948     1
## 6      523621 Seiu Cope Service E~          1237    -0.8          61620941    14
## 7      507525 Winning Our Future            1020     1.2          32595939    18
```

Here, we see that there are seven shared organizations:

- Afl Cio Cope Pol Cont Comm
- Dnc Services Corporation/Democratic National Committee
- Nra Political Victory Fund
- The Media Fund
- Republican National Committee
- Seiu Cope Service Employees International Union Committee On Political Education
- Winning Our Future

We might consider these the most prolific donating organizations overall, given that their appearance on both lists suggests that they register a high level of success on both criteria.

Now we can repeat this process for only those donors with ideology scores.

```
inner_join(prolific_donors1_scores, prolific_donors2_scores)
```

```
## Joining, by = c("bonica.cid", "contributor.name", "Ideology")
```

```
## # A tibble: 7 x 5
##    bonica.cid contributor.name           Contributions Ideology Total_Value_Don~
##         <dbl> <chr>                             <int>    <dbl>            <dbl>
## 1  3805309992 Afl Cio Cope Pol Cont Comm          5455    -0.71         21240560
## 2 33580613966 Dnc Services Corporation/~          3155    -0.92        211839345
## 3       53553 Nra Political Victory Fund          2210     0.58         44369368
## 4        3418 Republican National Commi~          1346     0.96        367765948
## 5      523621 Seiu Cope Service Employe~          1237    -0.8          61620941
## 6      507525 Winning Our Future                  1020     1.2          32595939
```

```
## 7      341396 Moveonorg Political Action          850    -1.15        19726450.
```

When we limit our search to just those donors whom we have ideological data on, we unsurprisingly have the same list, except that Moveonorg Political Action now replaces The Media Fund. Interestingly, all of these groups have strong views, as measured by their CF Score. While Nra Political Victory Fund is the most moderate (surprisingly) at 0.58, the rest are either strongly progressive or conservative based on the scale discussed in the introduction. This might suggest that the groups donating the most money at the highest rate are extremely passionate about the issues and political landscape. Perhaps their strong views motivate them to literally put their money where their mouth is more than a moderate would.

Thus, an exploration of the data reveals two ways of understanding who the largest donors are in presidential elections from 1980 to 2016. Progressive-leaning organizations and donors seem to dominate measures that focus on the number of donations, specifically in more recent elections. However, when evaluated from the perspective of the sum of all donations, Republican Committees lead the field. When both lists are reconciled against one another, we find that all of the donors both getting a large number of donations and contributing a large total sum have strong liberal or conservative views. Rather than evenly contributing to candidates on both sides, the most prolific donors seem to have an intense support for either side, or perhaps want to influence policy. This could explain their commitment to contributing to political campaigns.

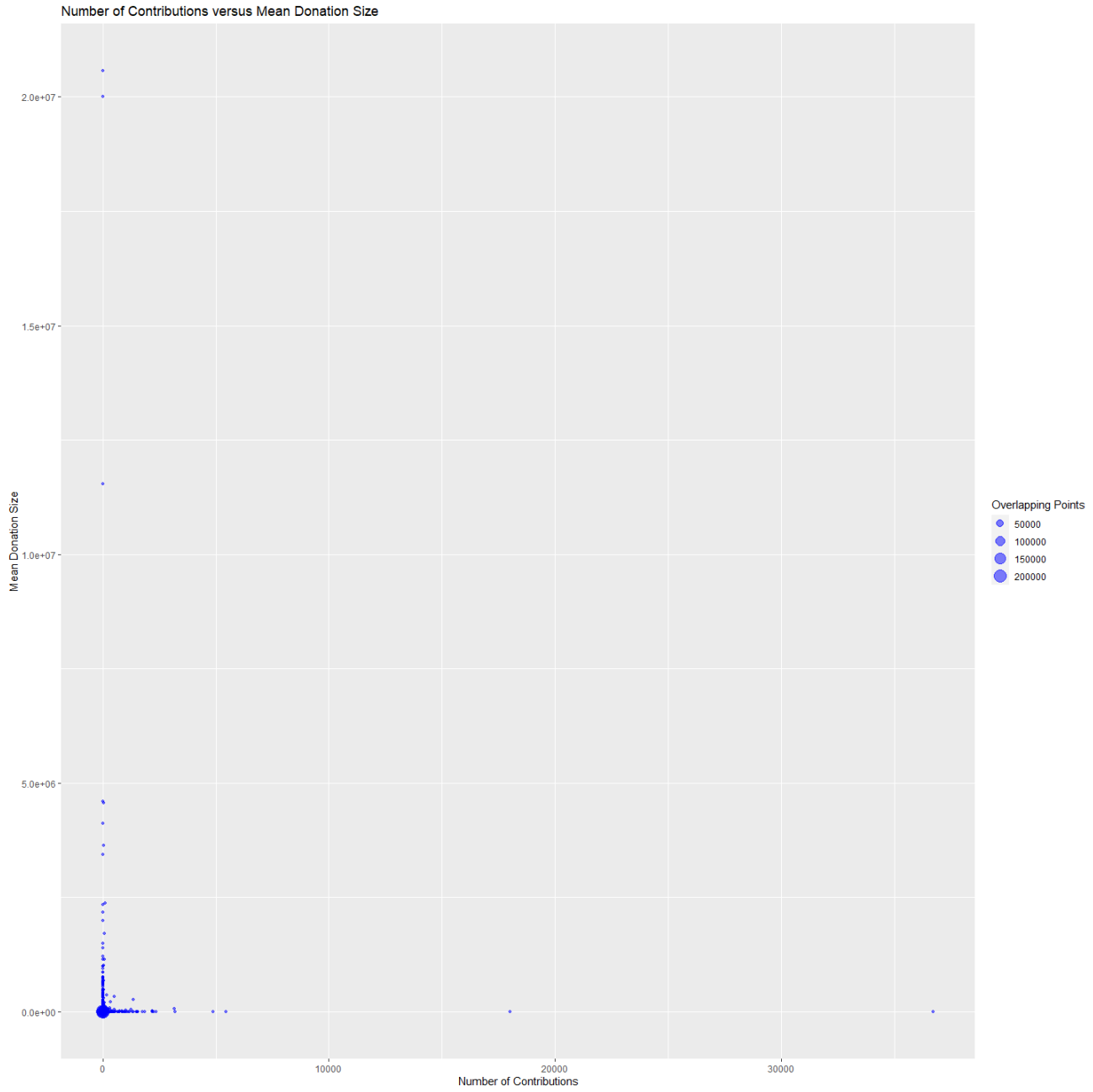
b) Do “small” donors contribute less often than “large” donors?

Over the last few years, much has been made of grassroots campaigns: movements funded by every day people contributing toward a cause they believe in. Such campaigns, however, also require regular and ongoing support from donors. During the 2020 election cycle, candidates from Bernie Sanders to Donald Trump discussed the importance of their grassroots support. But can grassroots support really be reliable on the national stage? To some degree, one might think that those who are unable to donate large sums of money may not have the money to donate on a regular basis. These “everyday people” may have other expenses that take precedence over political contributions. Or, perhaps spreading out opportunities to donate allows people who make smaller donations to remain more engaged and contribute more times to their candidate of choice. To explore this relationship, we can look at the relationship between the average size of a donation and

number of donations made by a given donor. Throughout this section, I use “small donors” to refer to those contributing small amounts of money on average and “large donors” to similarly mean those contributing larger sums on average.

First, we can plot the total number of contributions made by each donor in the sample against the mean size of their donations. Because we are interested in the size and frequency of donations, we can group by each donor and calculate the sum of all their donations and mean size of their donations. These values can then be plotted. We’ll use geom_count() because in traditional scatter plots, overlapping points are plotted on top of one another without it being clear that there are really two points there; this function accounts for that weakness by making size a function of the number of data points present at that location. Because of the amount of data, some points may still overlap, so we can use the alpha parameter to change the transparency of our points to 50%. This will make the plot more readable. While this plot is a bit difficult to gain much information from because of the inclusion of a few outliers, at least one takeaway is that the data are seemingly logarithmic of a form similar to $log(\frac{1}{x})$, though this is certainly not an exact function that would model the data. I include the equation as a reference to the type of curve that could model this data, if it were properly translated and otherwise fit to the data. But this data is still hard to discern. Perhaps zooming into the plot near the origin where much of the data is clustered can help resolve this issue. At present, however, it does seem that Mean Donation Size logarithmically decreases as Contributions increase.

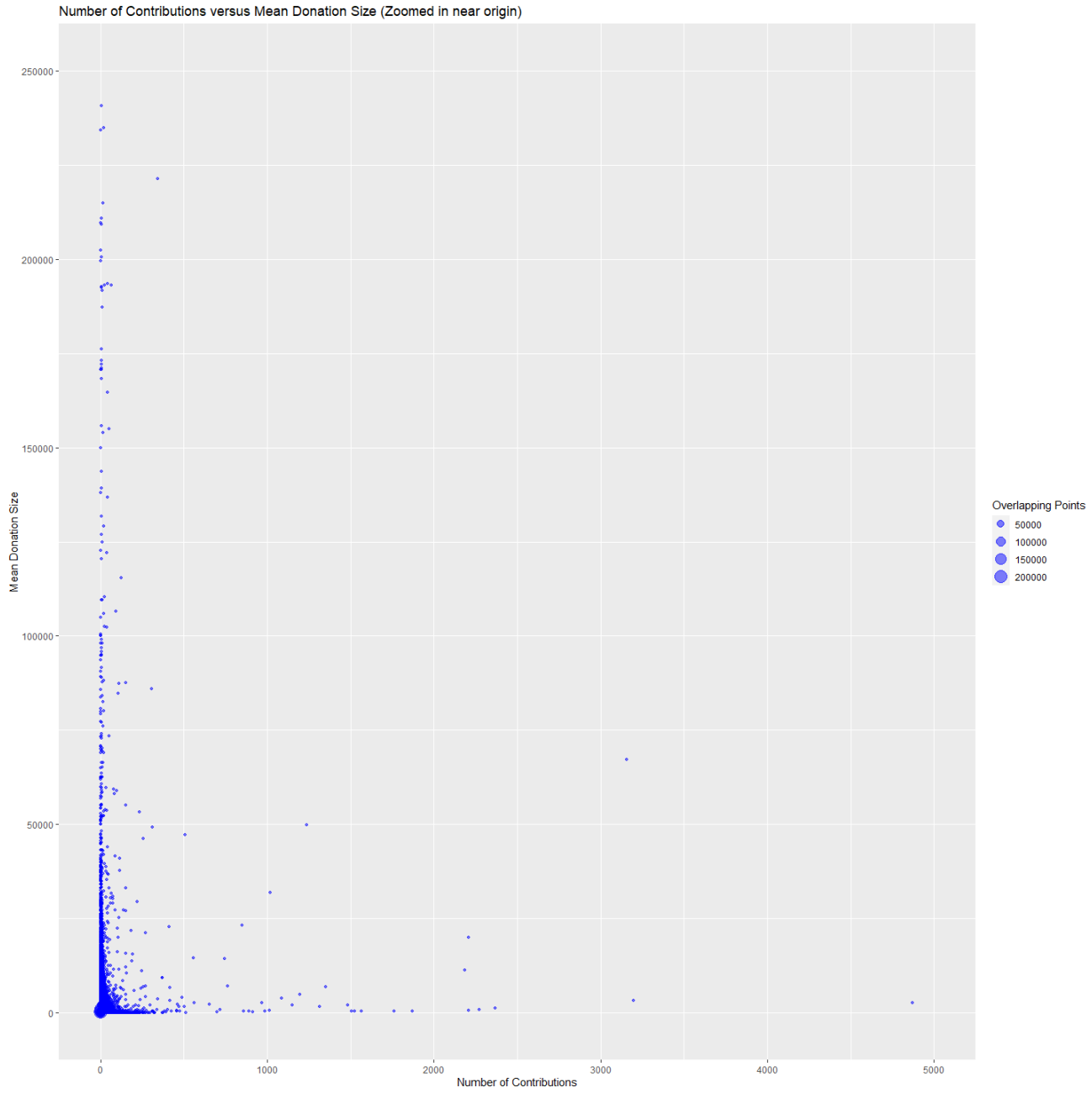
```
donations %>% group_by(bonica.cid) %>% summarize(Contributions = n(), Mean_Donations
```



Restricting the size of the axes allows us to zoom in a bit more on the data closer to the origin. We can use the ylim() and xlim() parameters to perform this. The following plot is the same plot made earlier, but restricted on the x-axis to (0,5000) and on the y-axis to (0,250000). The rest of the function parameters are the same. Here, we can clearly that the data seems to be logarithmic in nature. This supports our earlier intuition that it might be reasonable to perform a log-transformation on our two variables of interest.

```
donations %>% group_by(bonica.cid) %>% summarize(contributor.name = first(contributor
```

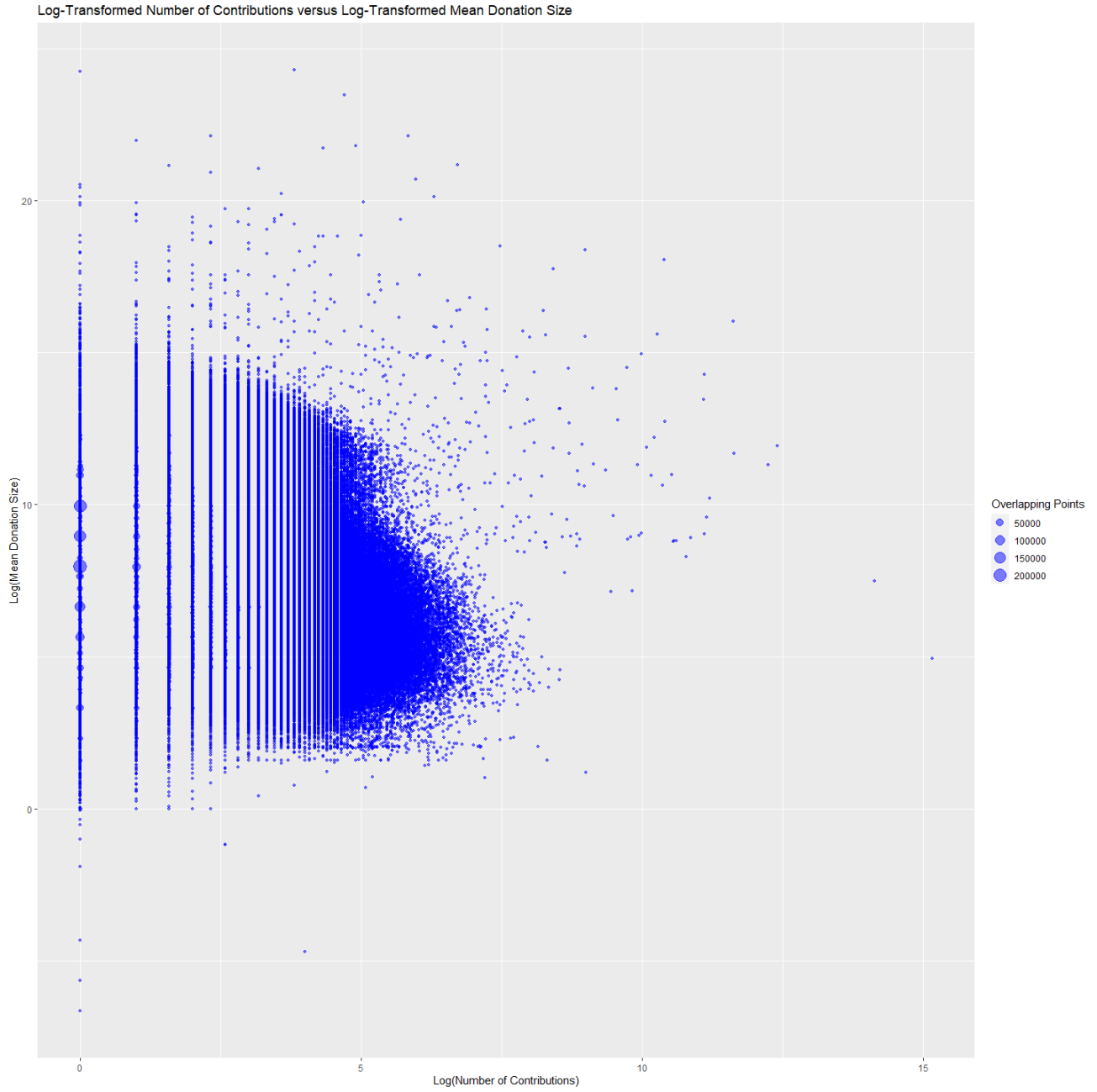
```
## Warning: Removed 70 rows containing non-finite values (stat_sum).
```



Because this data seems to be fairly logarithmic in nature. We can transform **Contributions** and **Mean_Donations** by taking the logarithm (base 2) of both. If we assume a logarithmic relationship exists between these two variables, then we should expect a linear relationship should exist between their log-transformed counterparts **Log_Contributions** and **Log_Mean_Donations**. We can produce a scatter plot of **Log_Contributions** against **Log_Mean_Donations**, once again using the geom_count() function to ensure that overlapping points are captured by the size of a given data point. The rest of the function parameters are the same. Our plot is now much more informative. Notably, there seems to be a “narrowing” effect as the number of contributions increases. By this, I mean that the range of values that the mean value of donations regularly takes on seems to become smaller. This is captured in the somewhat triangular shape of the data. Of course, these

numbers are log-transformed, so we must be careful about interpretations, but the direction of these effects are fairly clear: regular donors seem to contribute a smaller selection of values.

```
log_data <- donations %>% group_by(bonica.cid) %>% summarize(contributor.name = first

log_data %>% ggplot(aes(x = Log_Contributions, y = Log_Mean_Donations)) + geom_count(
```



Now, let's develop a model for these data. We can build a linear model using the lm() command with **Log_Mean_Donations** as our dependent variable and **Log_Contributions** as our independent variable. We'll save this model to **log_model** and can view the

coefficients and their statistical significance using the summary() command. Our model predicts a coefficient of -0.3281636 for the effect of **Log_Contributions** on **Log_Mean_Donations**; the intercept is 7.9482331. The negative coefficient suggests that as the number of contributions increase, the average donation size decreases. This generally supports the hypothesis that smaller donors are likely to donate more often than larger donors. Moreover, the coefficient is statistically significant, suggesting that the model does not ascribe the effect to mere chance. However, this model is certainly not comprehensive; it has an R-squared of just 0.04%, so the model can explain just 4% of the variation in **Log_Contributions**. This may be because key variables are missing. For example, We might want to include other factors like income in such an analysis, since the amount of money at one's disposal might influence willingness to donate and the size of one's donation. However, while our model is certainly far from perfect, it still offers an initial insight into the relationship between **Log_Contributions** and **Log_Mean_Donations**.

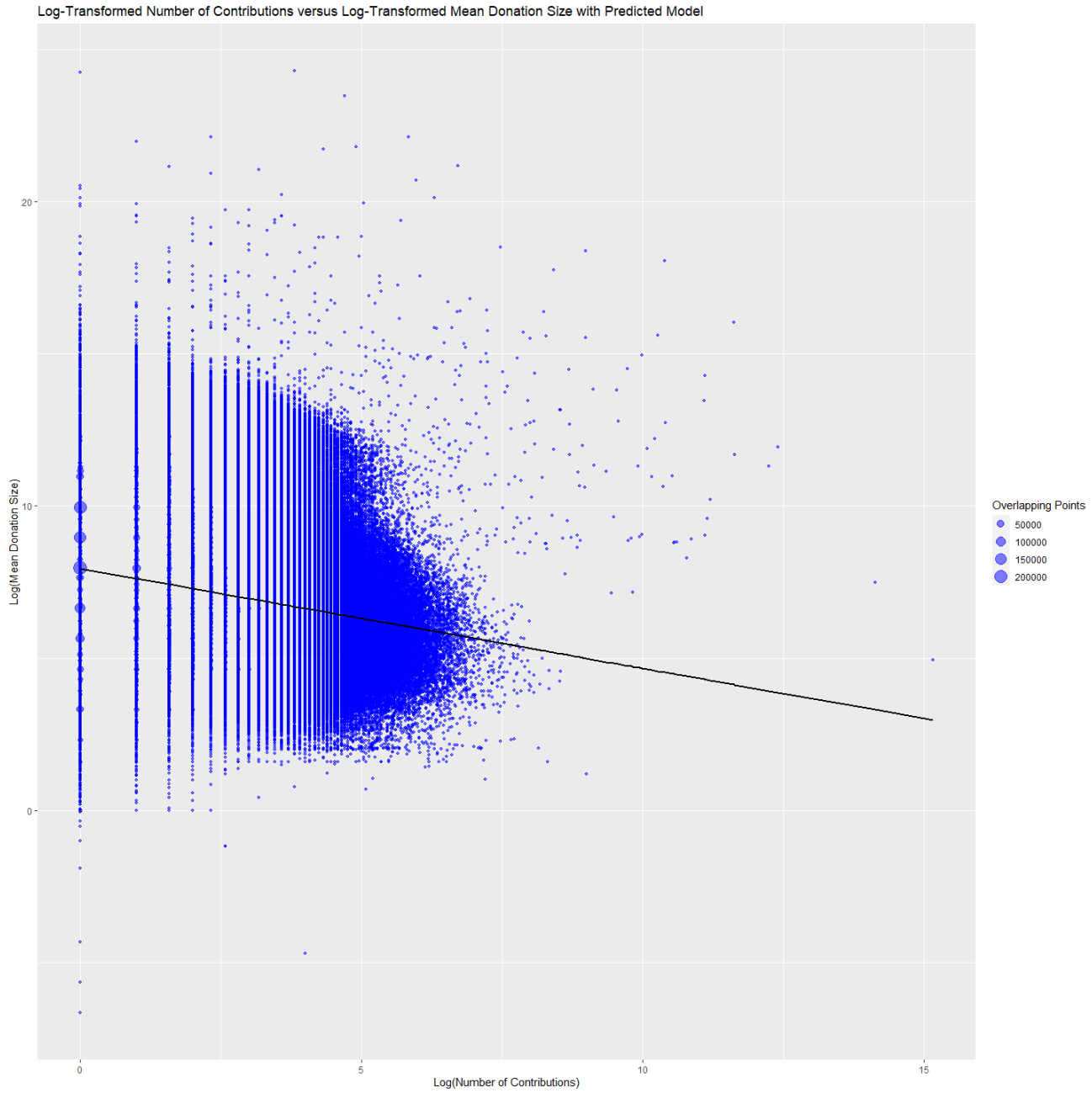
```
log_model <- lm(Log_Mean_Donations ~ Log_Contributions, data = log_data)
summary(log_model)
```

```
##
## Call:
## lm(formula = Log_Mean_Donations ~ Log_Contributions, data = log_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5921  -1.3044   0.0176   1.3618  17.5953
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.9482331  0.0015896  5000.1   <2e-16 ***
## Log_Contributions -0.3281636  0.0008796  -373.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.026 on 3000853 degrees of freedom
## Multiple R-squared:  0.04433,    Adjusted R-squared:  0.04433
## F-statistic: 1.392e+05 on 1 and 3000853 DF,  p-value: < 2.2e-16
```

We can plot the predictions of this model on top of our earlier scatter plot of **Log_Contributions** and **Log_Mean_Donations** to visualize how it fits the data. Using the add_predictions() function, we can add the values predicted by our model to our data. Then, we can create a geom_count() plot of **Log_Contributions** and **Log_Mean_Donations**, just as we did previously with the same parameters. Finally, we

can add a geom_line() to this that takes **Log_Contributions** as its set of x-values and the predictions as its y-value. I make the line a bit more visible using the lwd parameter, which refers to line width. The model line visualizes the predicted relationship between **Log_Contributions** and **Log_Mean_Donations**. The downward sloping line makes clear that as contributions made increases, the size of donations appears to decrease. While this matches our intuition, the model does not necessarily do a great job of estimating the values when the number of contributions is low, as there is such a large range of mean donation sizes. As a result, we would expect the errors, or residuals, to be large in this region. However, it does seem to be a bit more accurate and better fit as this range narrows and the number of contributions increases.

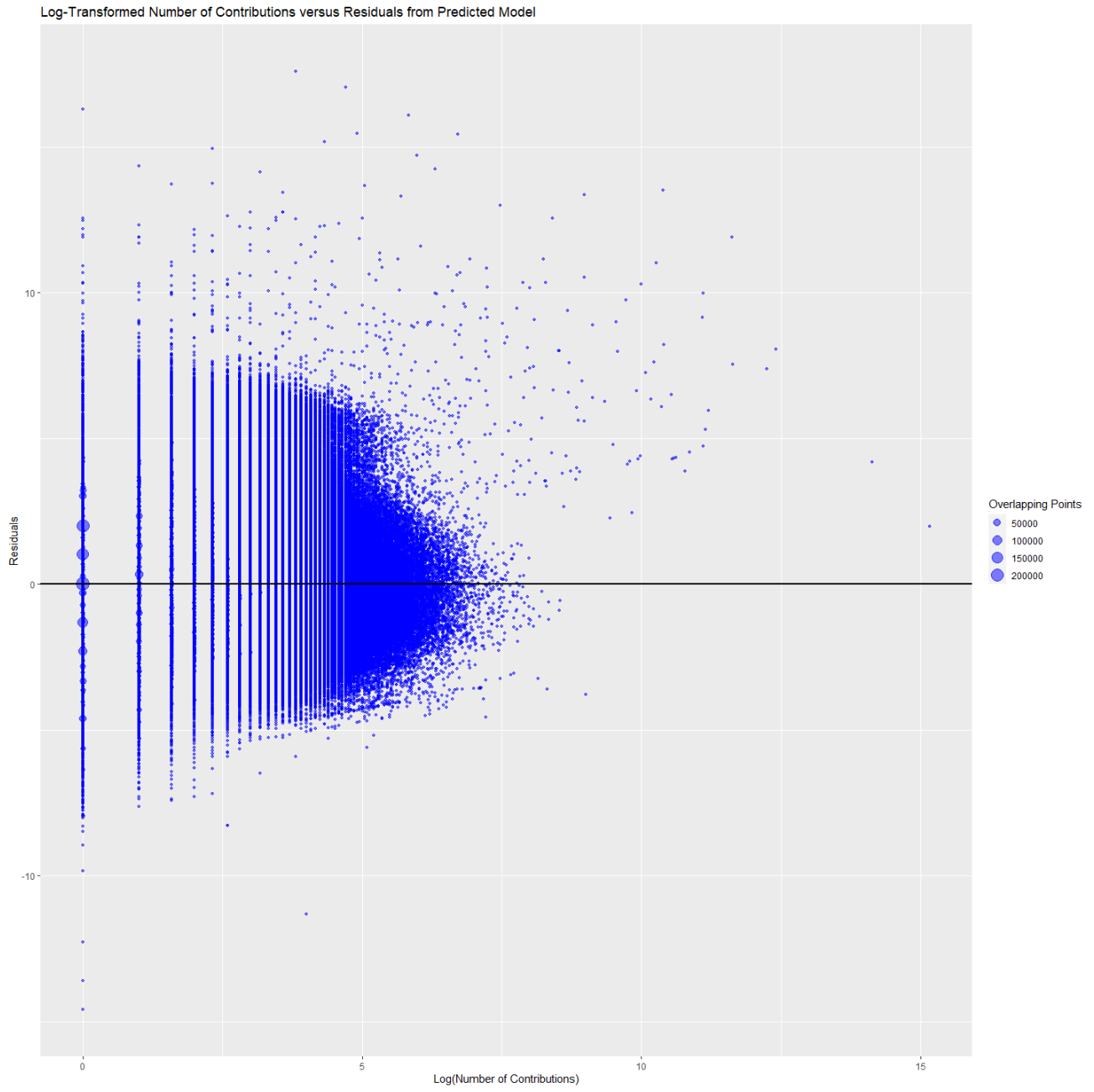
```
log_data %>% add_predictions(log_model) %>% ggplot() + geom_count(aes(x = Log_Contrib
```



Finally, we can plot the residuals of these data against **Log_Contribution** to determine if this is a good linear model for the data. Residuals are the difference between the model's predictions and the actual values. By examining the distribution of residuals greater or less than zero, we can determine if our model systematically over- or under-estimates its predictions. If it does exhibit such bias, perhaps a linear model is not appropriate. We can accomplish this by using the add_residuals function to add a column for the residuals to our data, then plotting a geom_count() scatter plot of **Log_Contributions** against the residuals. The other parameters for this scatter plot are the same as in the previous plots in this section. Then, we can draw a line on the x-axis where y = 0 to make clear the split between positive and negative residuals. The geom_hline() function is helpful for this, as the y-intercept can simply be set to zero and I give my line additional thickness by setting lwd = 1.

The plot reveals that while there are large residuals from the model, the errors do seem to be evenly distributed on either side of the zero line. This is suggests that the residuals are likely independent of one another, which adds some credence to our model. That being said, there are certain regions where the residuals are much greater than in others. Introducing other variables to build better models could reduce these errors and generate better predictions to develop a more keen understanding of donor behavior.

```
log_data %>% add_residuals(log_model) %>% ggplot(x = Log_Contributions, y = resid) +
```



Thus, a preliminary analysis seems to suggest that as the number of contributions increases, the average size of one's donations also falls. This relationship seems to be

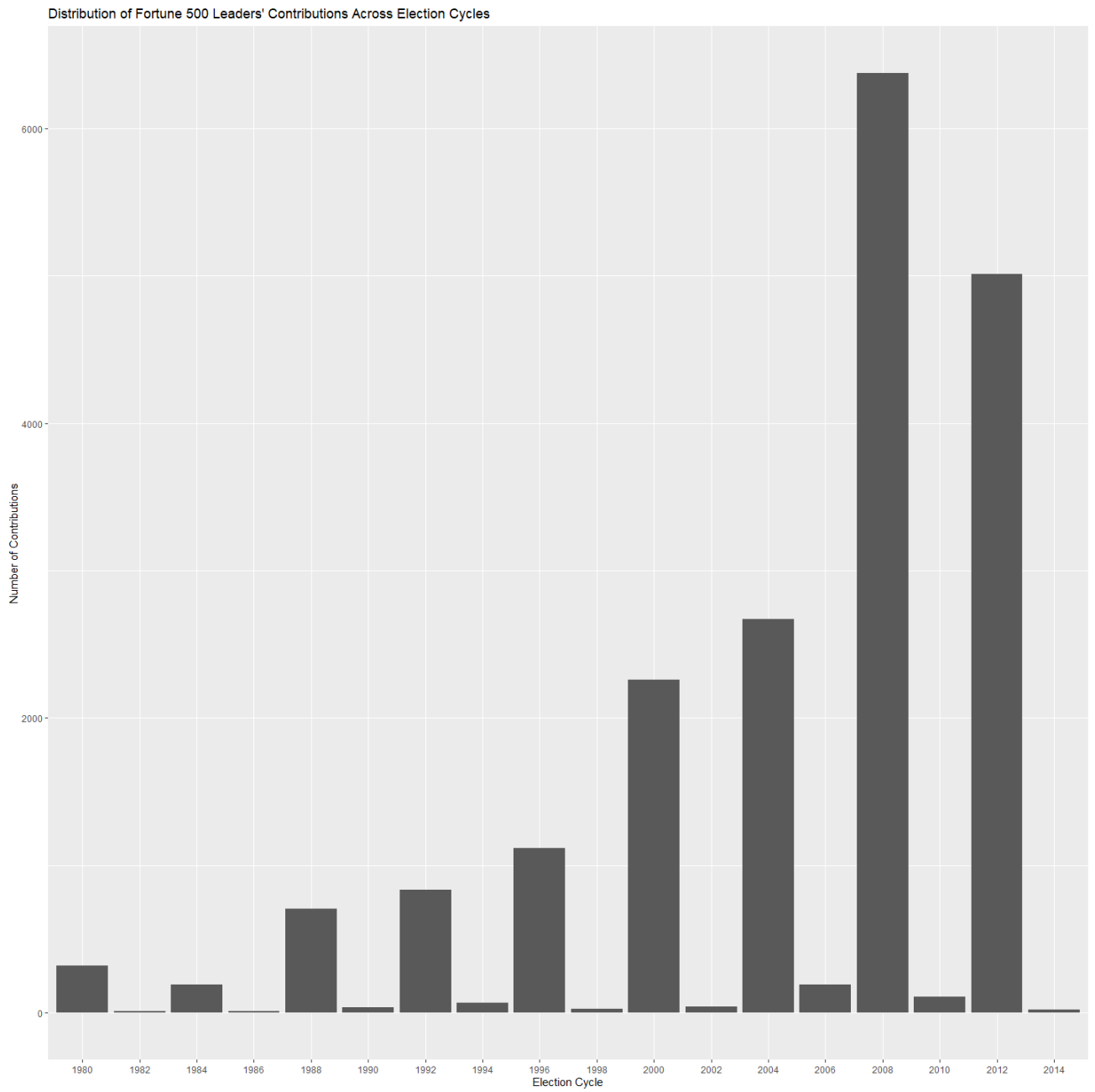
logarithmic in nature. This offers some credence to theories of grassroots organizing and fundraising. However, there are limitations to this analysis that must be accounted for before any clear determination can be made on this issue. I hope to further explore this issue in later analyses.

c) What are Corporate Leaders' donation patterns and how do they vary from the general population?

Corporate influence in American politics is another area that has been increasingly popular as billionaires achieve unheard-of sums of wealth. We might wonder how corporate executives and CEOs factor into election politics through contributions. Recall that corporate leaders are those included in the set of executives and CEO's of Fortune 500 companies as of July 2012. Importantly, this is not a complete list of corporate elites, but a list of those whose contribution records could be identified and who made individual donations that could be linked back to them, rather than contributions to organizations that then donate those funds. The results are thus restricted to transactions that fall under this purview. Moreover, there may be a number of wealthy individuals that do not work at these companies and either inherited their wealth or were not recognized by Fortune. These individuals are not included as corporate leaders, but perhaps would act similarly given the same amount of money or power.

Let's start by generating a plot for the distribution of ideology for these individuals. First, let's filter our **donations** dataset into a dataset with just the corporate leaders. We'll call this tibble **leaders** and can identify leaders using the **is.fortune** variable we previously generated. This allows to simply filter out entries that are marked 0. Next, we can produce a simple plot of their contributions over time. We'll use a bar graph with number of contributions on the y-axis and election cycle on the x-axis. This will give us a sense of the flow of funds over the period. We can simply use geom_bar() because it will record the number of times each cycle value appears in the **leaders** dataset, allowing us to produce our desired plot. There is not too much that is surprising in this plot. Donations are far more frequent during the general election compared to primary elections, as is expected. Also, we observe an increase in donations in 2008 and 2012 consistent with the findings from Part A's discussion on the most prolific donors. However, it is interesting to see that the number of donations have steadily been increasing over the last forty years, despite the dip from 2008 to 2012.

```
leaders <- donations %>% filter(is.fortune == 1)
leaders %>% ggplot() + geom_bar(aes(x = cycle)) + labs(x = 'Election Cycle', y = "Num
```

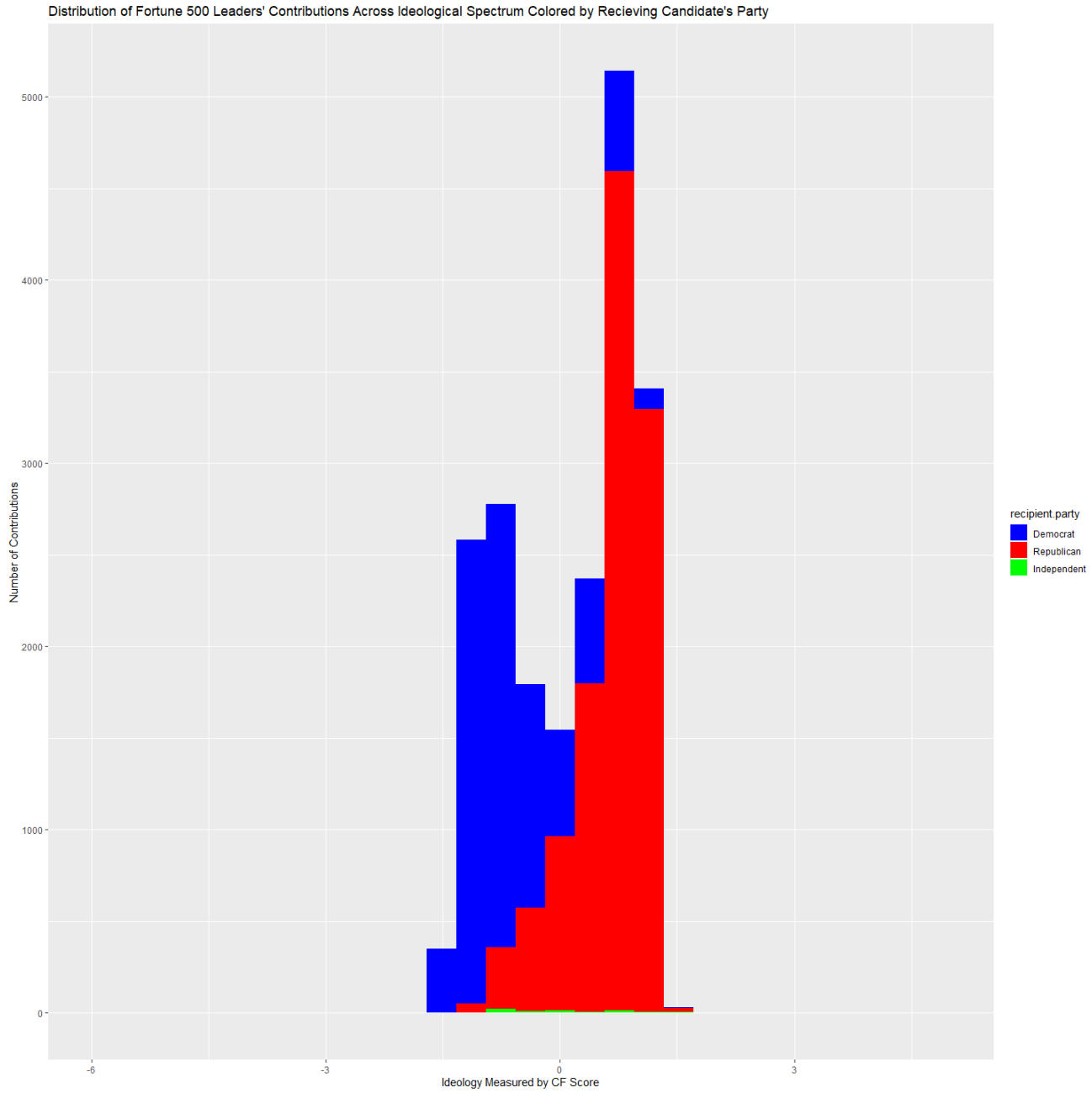


Next, we can compare the distribution of ideology among the elites and the general public. We can begin with the corporate leaders. Here, I plot a histogram of CF Scores from the **leaders** dataset for each transaction. Because we are tracking the number of transactions, the histogram's y-axis represents the number of donations at a given range of ideology scores. For each transaction, we also know the receiving candidate's party. We can use this data to fill our histogram with the appropriate color: blue for Democrats, red for Republicans, and green for Independents. This allows us to capture both their average ideology by position and the parties of those they gave to by color. Importantly, the plot then allows us to see deviations from the average ideology. For easier comparison, I also format the x-axis of the leaders' plot to range from -6 to 5 to be consistent with the next plot of the

general public. To compare this distribution with the general public, I produce the same plot, but for those in the **donations** dataset without Elite status. All other important parameters are the same. The comparison reveals a few key findings. First, the general public has a wider range of ideologies than the elites; this is expected, as there are a fewer number of elites and the backgrounds, values, and preferences of people who are elites may not be random. Moreover, there seems to be many more progressive-leaning donations among the general public compared to the elites. In the general public, this might be because some larger organizations, which are largely funded by working-class people, like Working America skew toward the left. Yet, even though the distribution of elite ideologies is more clustered around the center, more seem to lean on the conservative side. In addition, the general public seems to have some cases where people who generally land on one side of the distribution “flip” and vote for the party traditionally aligned with the other ideology. However, this seems much more commonplace among elites. We see a number of conservative-leaning elites voting for Democrats and progressive-leaning elites voting for Republicans. There are a number of explanations for this. One of the possible reasons is that certain presidents affiliate with one party but are actually reasonably moderate and are able to attract some folks from the other side of the aisle. On the other hand, the parties have also traded off geographies and demographics with one another over the last 40 years. For example, the American South was once dominated by Democrats until the 70’s, but now is dominated by Republicans. Perhaps similar shifts during this period are captured by this plot.

```
leaders %>% ggplot() + geom_histogram(aes(x = contributor.cfscore, fill = recipient.p
```

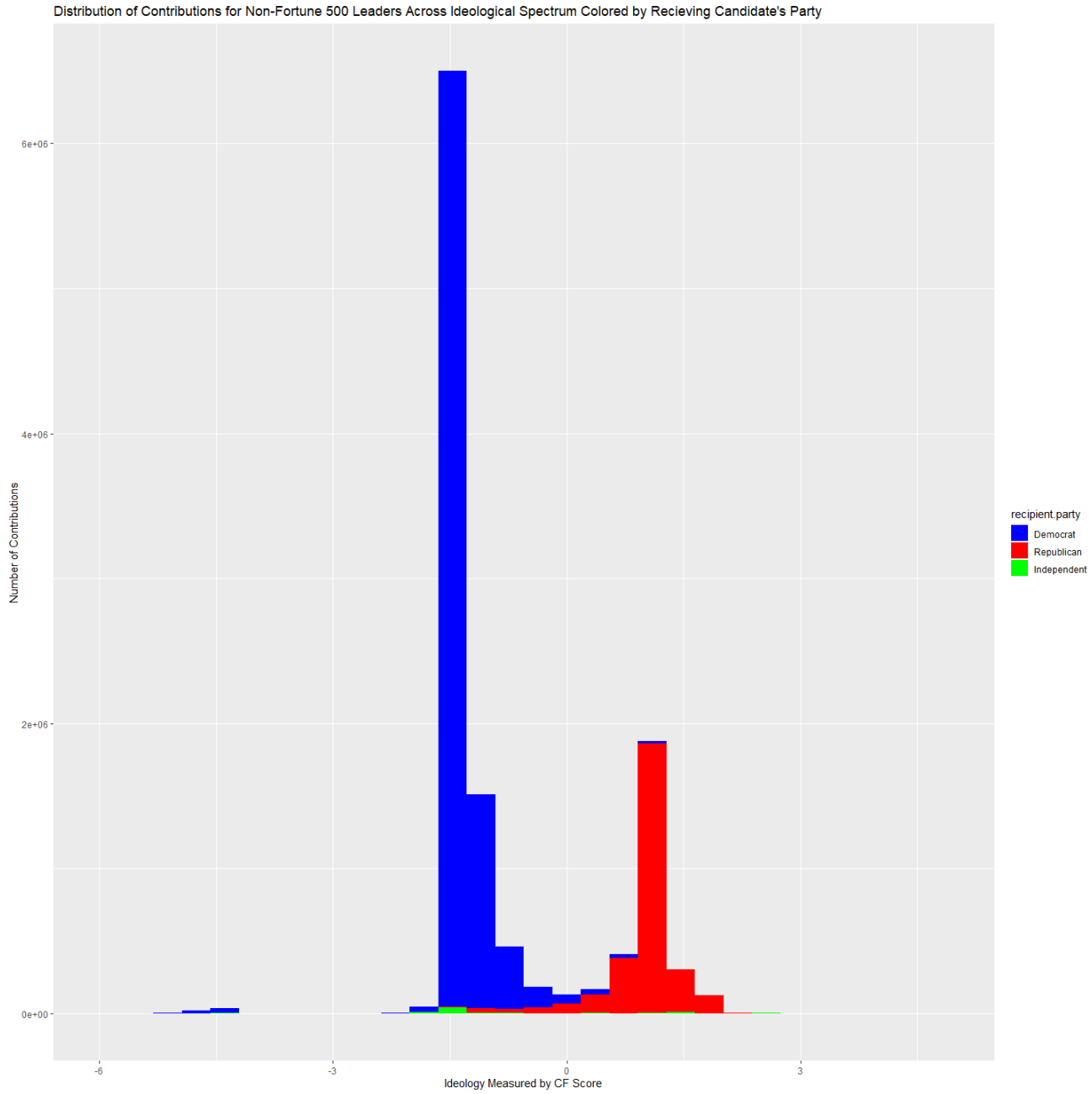
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 6 rows containing missing values (geom_bar).
```




```
donations %>% filter(is.fortune == 0) %>% ggplot() + geom_histogram(aes(x = contribut
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 237345 rows containing non-finite values (stat_bin).
```



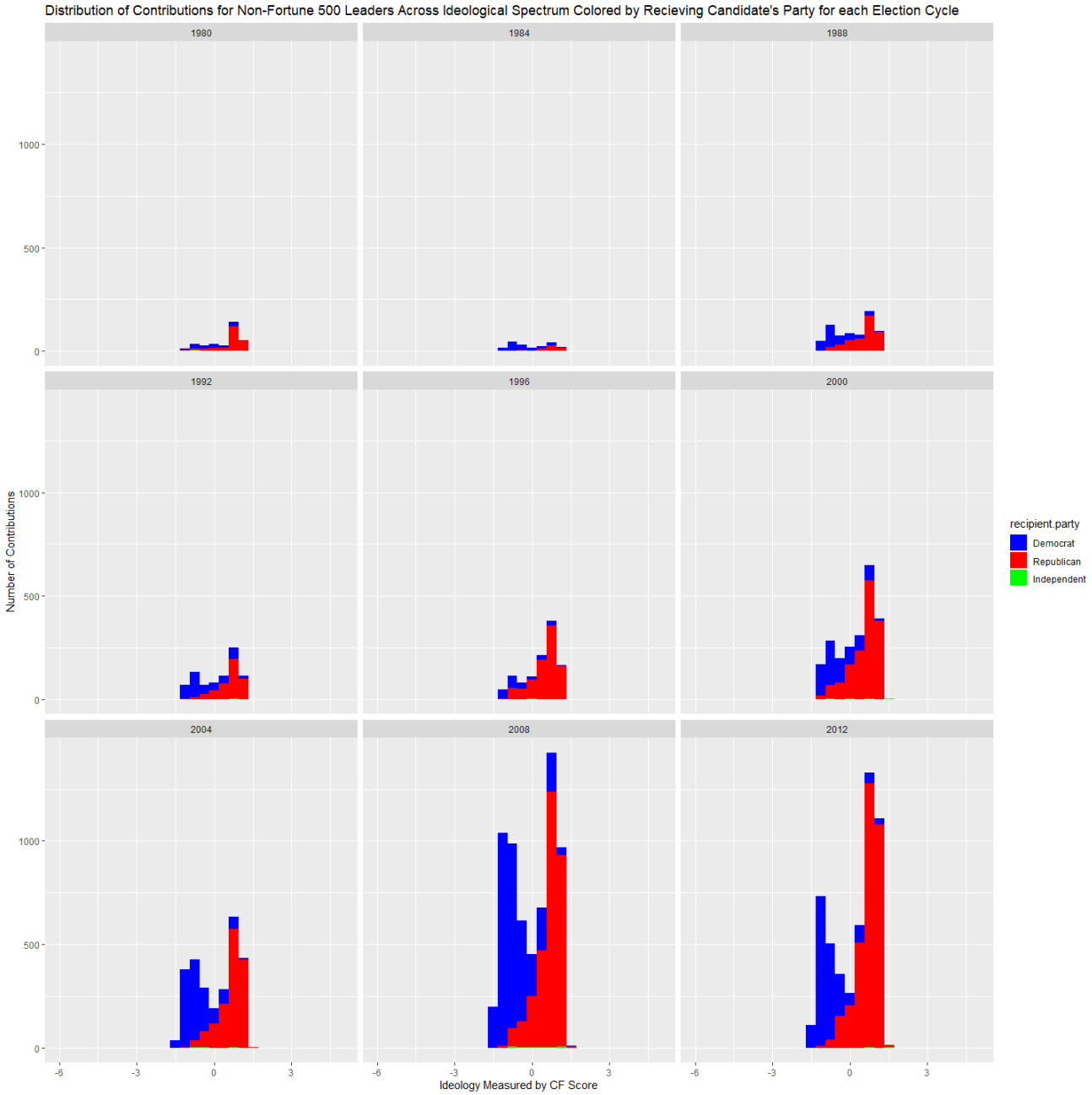
To understand how elites' contribution patters and ideologies shift over time, we can break down our previous plot on elites' ideologies into each of the election cycles to see if there are major shifts or trends over the course of the last forty years. To create this figure, I filter the data to just transactions explicitly from a general election cycle for simplicity, then add a facet_wrap() on the basis for **cycle** to re-create a separate plot for each one. Two plots I want to highlight are 1996 and 2008. In 1996, it seems that a fair number of liberal-leaning elites contributed to a Republican candidate, presumably the nominee Bob Dole. Interestingly, Bob Dole lost in a landslide to Bill Clinton that year, so it is surprising to see that he was able to attract contributions from those on the other side. Of course, elites are only a small portion of the population and should not control elections, but it is still interesting to isolate a show of support from a given faction for Dole. On the other hand,

2008 seems to demonstrate a fair degree of support for the Democratic candidate, Barack Obama, from typically conservative-leaning elites. Perhaps Obama was able to build ties with certain donors or his plan for economic recovery from the initial events of the Great Recession was exciting for corporate elites whose firms had been struggling.

```
leaders %>% filter(cycle %in% c("1980", "1984", "1988", "1992", "1996", "2000", "2004
```

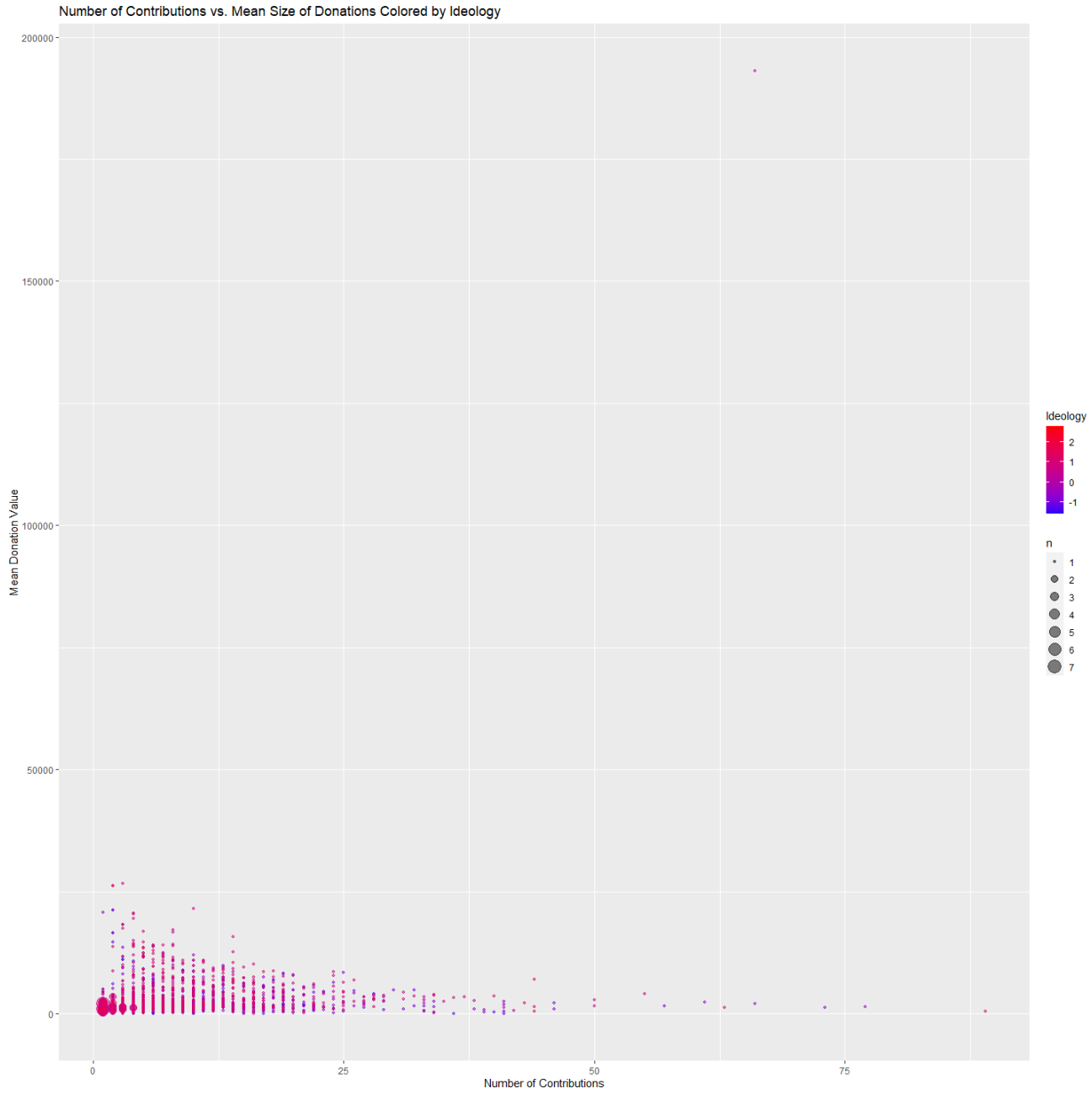
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 52 rows containing missing values (geom_bar).
```



We can also visualize elites' contribution habits by comparing their statistics about their donations. Let's prepare a plot that captures three key variables summarized from the **leaders** dataset: total number of contributions, mean contribution value, and ideology. After grouping on the unique donor identifier and summarizing to compute each of the aforementioned variables, I decided to plot them on a scatter plot, given that all of these data are numeric in nature. I will plot total number of contributions, **Contributions**, on the x-axis and **Mean_Donation_Value** on the y-axis. I color each point based on **Ideology**. As previously mentioned, geom_count() is useful because it determines the size of each point based on the number of overlapping data points. Interestingly, we still see a generally negative trend in **Mean_Donation_Value** as **Contributions** increases. There is an especially large smattering of points near the origin, at both a low frequency and size. Moreover, there does not seem to be any clear trend in the ideologies, which seem to be evenly distributed across the plot.

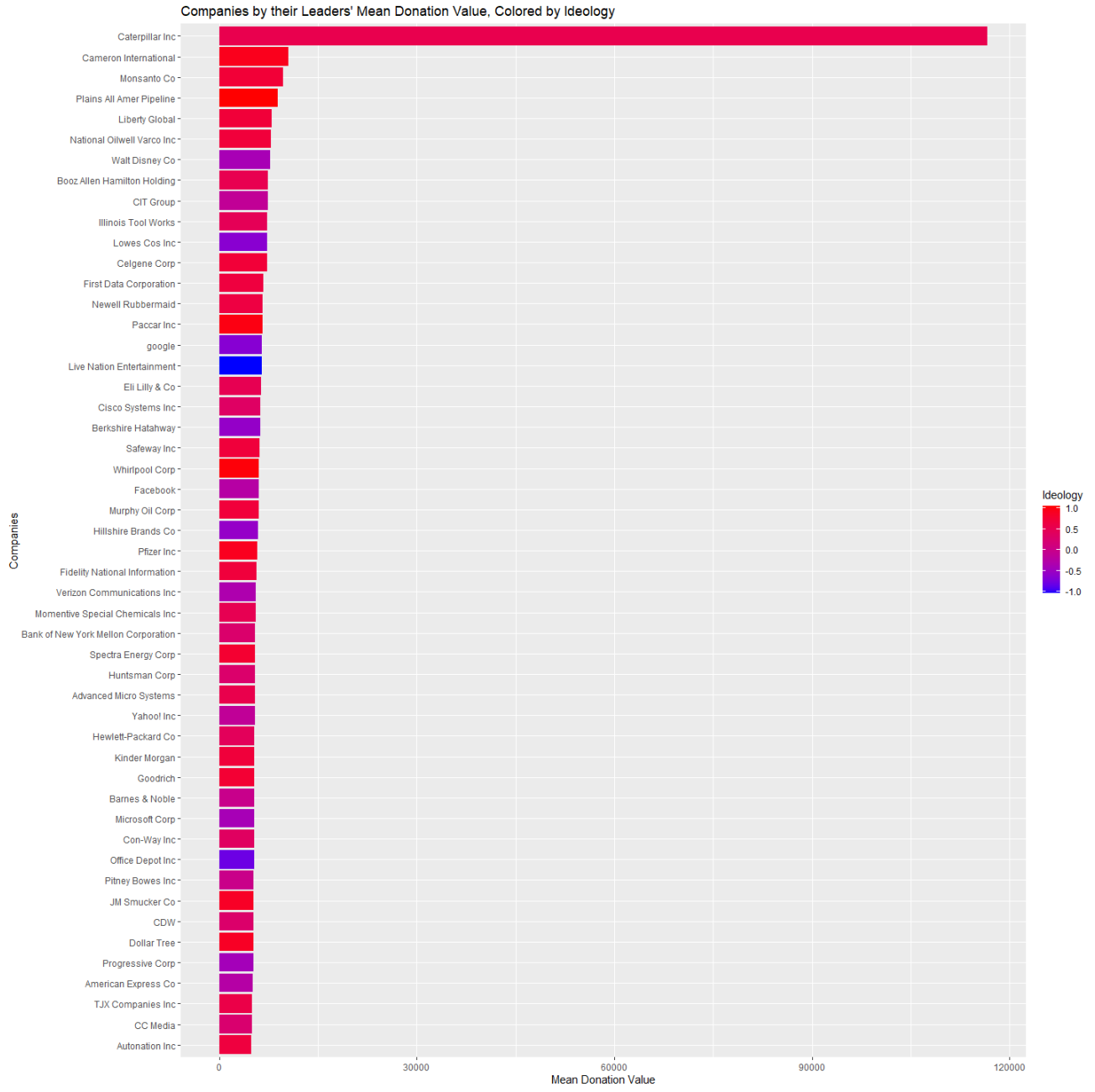
```
leaders %>% group_by(bonica.cid) %>% summarize(contributor.name = first(contributor.n
```



We can also group elites by the companies elites work for and examine trends across firms. This might help us understand if there is a relationship between certain industries and political ideologies. After grouping, I calculate the same statistics as in the previous plot, and plot the mean size of donations made by a given set of coworkers. I turn the company name into a factor and sort it by the average donation size to make the plot more readable from the perspective of size. I then color the plot by the mean ideology of the coworkers and flip the axes of the plot for greater readability. This plot clearly demonstrates that those contributing the largest sums on average are mostly conservative. Caterpillar Inc, a construction equipment manufacturing company leads the field by far in the average size of donations. Using these results and the previous plot to inform our interpretation of the data,

this seems to be a one-time gift, however, so perhaps there was a specific reason for the large donation.

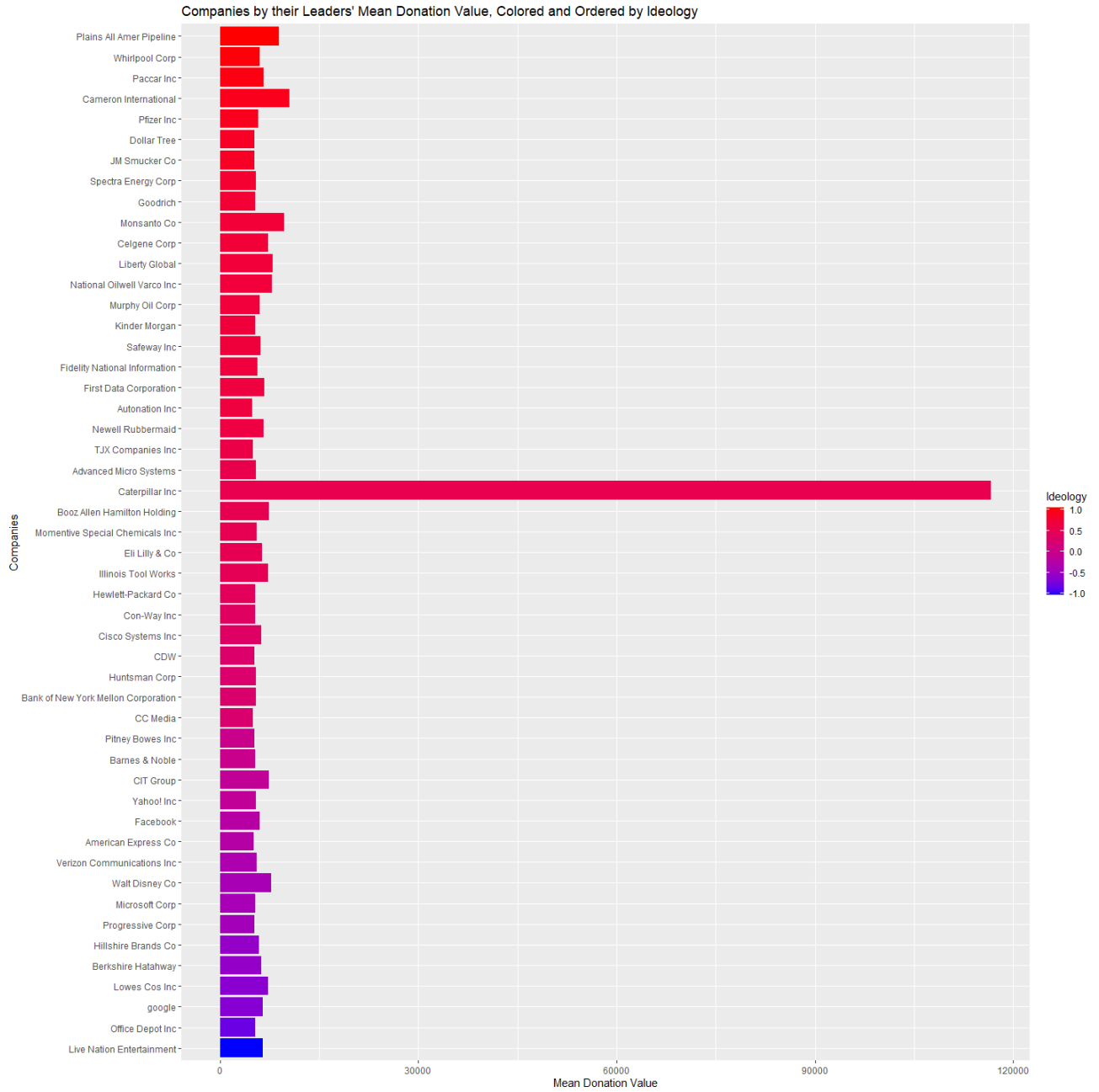
```
leaders %>% group_by(corpname) %>% summarize(Ideology = mean(contributor.cfscore), Co
```



We can then produce the same plot but re-sort the companies by the average ideology of elites to gain a better sense of how the predicted views of elites at each firm compare. We now have a better sense of which companies' elites are more conservative and which ones are more liberal. On the conservative side, it is not surprising to see that the most conservative organization is Plains All American Pipeline, a petroleum manufacturer and Whirlpool, which has been fighting for conservative-backed protectionist policies for some

time. On the other hand, technology companies seem to be divided by age: Facebook, Google, Verizon, Yahoo, and Microsoft all have elites that lean progressive, while Cisco, HP, and Advanced Micro Systems (AMD) all lean conservative. It would be interesting to update these results and see if any of the newer elites' views have shifted as they have come under greater scrutiny from federal regulators.

```
leaders %>% group_by(corpname) %>% summarize(Ideology = mean(contributor.cfscore), Co
```



Finally, we can model how being an elite affects the average size of one's donation relative to being a member of the general public. We can start by grouping our data by the donor unique identifier. Then, we can summarize if each donor is an elite by taking the first value

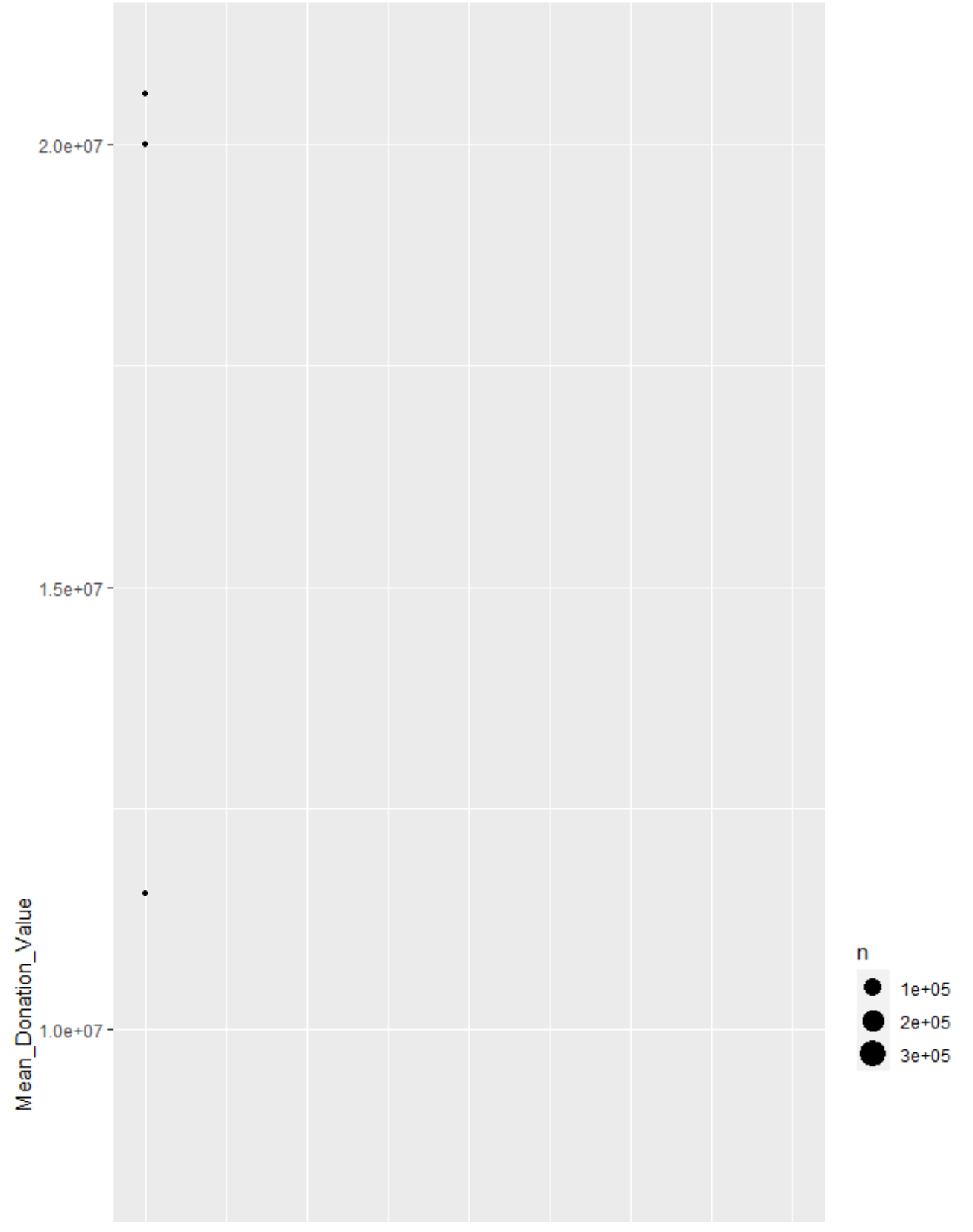
of **is.fortune** of each donor, and because it is consistent, this will simply report the status of the donor; I call this variable **Fortune_500**. We will also summarize the total number of contributions made as **Contributions**, mean donation value as **Mean_Donation_Value** and average CF Score as **Ideology**. Then, using the lm() function, I start by producing a simple linear model with **Mean_Donation_Value** as the dependent variable and **Fortune_500** as the independent variable. Because **Fortune_500** is a binary variable, the model will simply report the average **Mean_Donation_Value** for non-elites as the intercept and the difference between the average **Mean_Donation_Value** for elites and non-elites as the coefficient for **Fortune_500**. I use the summary() command to report the coefficients and find that the intercept is 522.26 while the coefficient for **Fortune_500** is 1954.45. This means that being an elite increases the average size of donations, though this prediction does not us much more than that elites donated more on average than the non-elites. Despite being statistically significant, the R-squared is essentially zero. Because some donors' average contribution extends into the millions, a scatter plot of these predictions is not very informative, though it is included below, with **Fortune_500** on the x-axis and **Mean_Donation_Value** on the y-axis plotted using a geom_count(); the predictions are represented by the two large red dots on the 0 and 1 sides of the x-axis. 0 refers to the general public; 1 refers to elites. However, this plot is generally not very useful and the model itself seems too naive to capture any meaningful effects from being an elite.

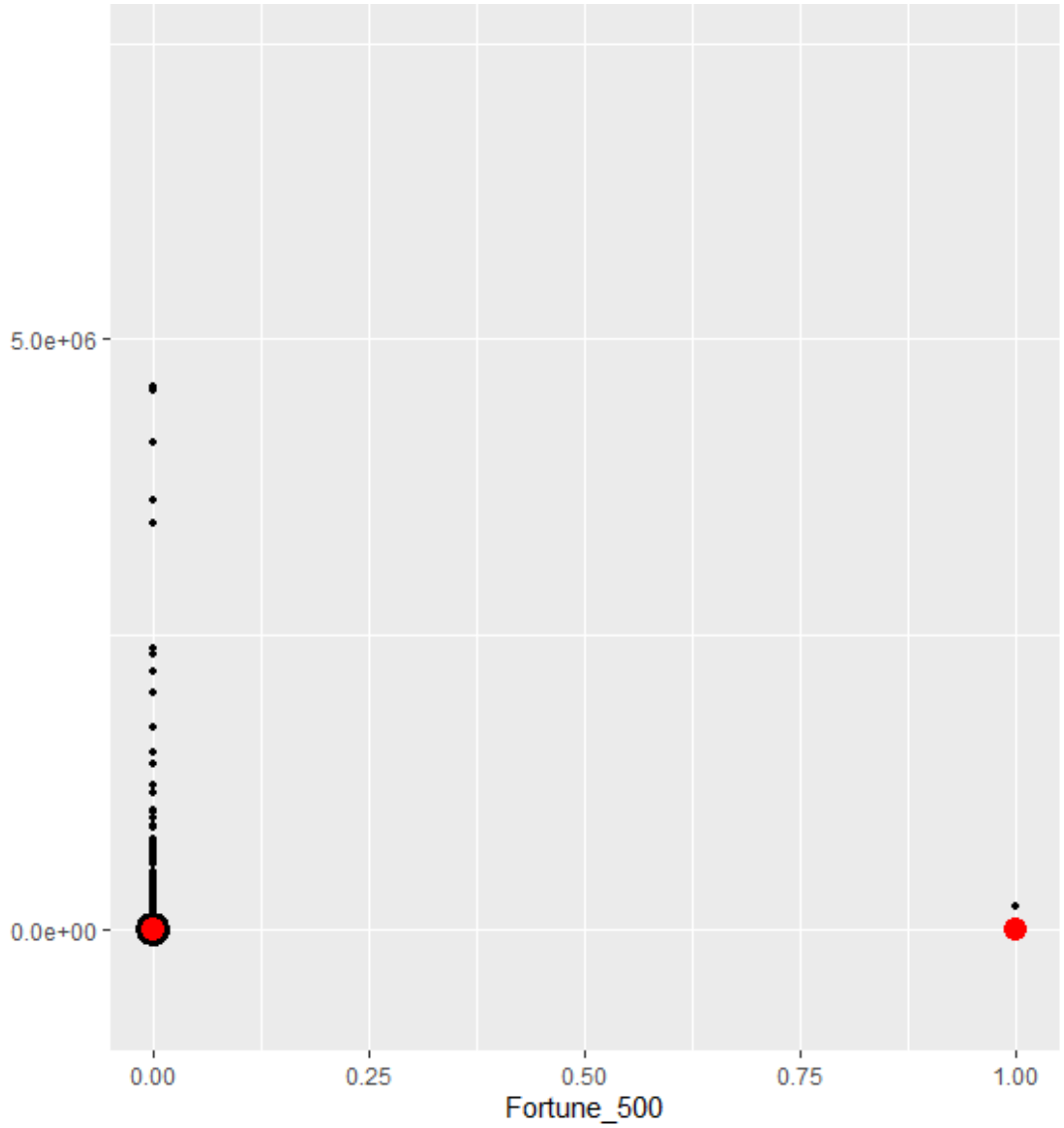
```
leaders2 <- donations %>% group_by(bonica.cid) %>% summarize(Fortune_500 = first(is.f
model_fortune <- lm((Mean_Donation_Value ~ Fortune_500),leaders2)
summary(model_fortune)
```

```
##
## Call:
## lm(formula = (Mean_Donation_Value ~ Fortune_500), data = leaders2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##    -2412     -458     -318      -22 20570963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   522.27      11.03  47.355  < 2e-16 ***
## Fortune_500  1926.49     357.50   5.389 7.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19100 on 3000853 degrees of freedom
```

```
## Multiple R-squared:  9.677e-06,  Adjusted R-squared:  9.344e-06
## F-statistic: 29.04 on 1 and 3000853 DF,  p-value: 7.091e-08
```

```
ggplot(leaders2) + geom_count(aes(x = Fortune_500, y = Mean_Donation_Value)) + geom_p
```





Given that the above model is rather naive, we can attempt to build a more robust model that uses interaction terms between **Fortune_500** and **Contributions** and **Fortune_500** and **Ideology** to predict **Mean_Donation_Value**. While I do not visualize this model, given that it is four-dimensional, I report the coefficients and statistical significant of each using the summary() command. They are listed in the output below. Notably, all of the coefficients are positive, meaning that being an elite, being more conservative, and contributing more all result in greater average donation size. The interactions are also positive. This is interesting in that it rebukes our earlier finding that mean donation size decreased as contributions increased. That being said, this model does not seem to be incredibly robust. There is a large jump in the R-squared to 0.01% over the previous model, which is an improvement, but still a bit unsatisfying. Further work on developing a more thorough, detailed model for donation size should be taken up to enhance our understanding of how donor behavior differs across economic, party, and social lines.

```
model_fortune2 <- lm((Mean_Donation_Value ~ Fortune_500*Contributions + Fortune_500*I
summary(model_fortune2)
```

```
##
## Call:
## lm(formula = (Mean_Donation_Value ~ Fortune_500 * Contributions +
##     Fortune_500 * Ideology), data = leaders2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -34811    -354    -252      68 4568571
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                571.3703     3.7023 154.329  < 2e-16 ***
## Fortune_500                978.2446   149.3745   6.549 5.80e-11 ***
## Contributions                0.9375     0.1458   6.429 1.28e-10 ***
## Ideology                   141.5838     2.8048  50.479  < 2e-16 ***
## Fortune_500:Contributions 114.4672    13.9220   8.222  < 2e-16 ***
## Fortune_500:Ideology      338.9145   138.4684   2.448   0.0144 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5799 on 2853584 degrees of freedom
##   (147265 observations deleted due to missingness)
## Multiple R-squared:  0.001037,   Adjusted R-squared:  0.001035
## F-statistic: 592.4 on 5 and 2853584 DF,  p-value: < 2.2e-16
```

Conclusion

The results of this analysis offer a number of ways to think about politics and campaign finance. On one hand, the first topic regarding which donors are the most prolific offers some insight into the trends in major donors over the last several years. The analysis seems to indicate that the most prolific donors tend to be highly ideological. A limitation of this finding is that this data may leave out key historical donations on the basis of organization type, therefore excluding donors that may be more moderate. More broadly, because these data are exclusively from American presidential elections, any insight they give us is limited to those conditions. They cannot be reasonably applied to other types of elections, or presidential elections in other nations.

Second, the analysis of large and small donors revealed that with the data available, as contributions rose, the mean size of donations fell. This offered some credence to the idea that grassroots organizing can work in the setting of presidential elections within the United States. Once again, any findings are limited by the scope of the data. Moreover, while the model seemed to track the general trend of the data, it still had large residuals that could be reduced by including other variables. After controlling for more factors, perhaps the relationships change or adjust. Indeed it is possible that the relationship we initially observed is weakened or reversed.

Finally, our analysis of elite donation behavior suggests that there may be some important differences in the distribution of ideology between them and the non-elites within the context of the data. Moreover, there seemed to be some industry-driven trends to elites' contribution behavior. The models developed in this section offer some insight into the fact that elites contribute more than non-elites and that other factors, when interacted with elite status may further increase donation size, but these effects seem somewhat tenuous. More broadly, we had an incomplete set of elites, so this behavior cannot be ascribed to every single elite, but just the set included in the data. Moreover, elites contributing to donor organizations may have different preferences and these are not reflected in the data, but are instead pooled with the rest of the non-elites. These challenges make it difficult to pronounce any conclusive theories about elite behavior.

Future work on this data could focus on a number of different areas. I think that a continued discussion about the ethical issues surrounding data on political contributions must be had. Developing norms and ideals about the ways we should handle our democracy in the Information Age is crucial. These discussions should motivate the kind of work pursued by data scientists in this area. Meanwhile, quantitative work, as described above, to better model political attitudes and willingness to donate could also be enhanced by integrating more data from other sources. In particular, income data would be very interesting to examine in relation to how people donate. This might offer insight into the issues that really drive people to contribute a cause when they are truly constrained on the resources they have.