Campaign Contributions:
An Analysis of Individual Contributors.

Adam Sabra

1 Introduction

Campaign contributions play a fundamental role in the American political system. When most people consider campaign contributions, they will generally think of Super PACs since they contribute the most amount of money to candidates. This, however, overlooks individual contributions from every corner of the country.

With public data at our disposal from the Federal Election Commission (FEC,) we can observe individual contribution data from their website.

2 Exploratory Data Analysis

The table with the data in question is not perfect and needed some steps for cleaning with the most important notes being:

- Each entry was separated by the vertical bar character, |.
- The labels of the columns were not present.
- The transaction date column needed to reformatted.

Luckily, the accompanying CSV file with the data set contained the labels of the columns. For the transaction dates, a bit more work needed to be done.

If read automatically, R would assume that the transaction date is a number, simply because it is. The structure of the date is as follows: MMDDYYYY. This structure is important. If the month is less than 10, the date's month would begin with a 0, which means R decides to omit it because it reads the date in as an integer, making the conversion to a Date object difficult.

This problem can be alleviated by loading in a sample of the data, manipulating the transaction date's class, and use the set of classes for loading in a larger subset of data for exploration - which utilizes 1 million samples in comparison to the complete data set with over 34 million observations. On my laptop, this took around a minute and a half. By using any more rows, the load time becomes exponentially larger.

Even with 1 million samples, it is still not representative of the bigger picture - especially given the fact that this data can be structured as a time series. By taking the first million out of 34, we are only observing a small window of time in contributions. We will observe how these faults hinder our analysis in the subsequent sections.

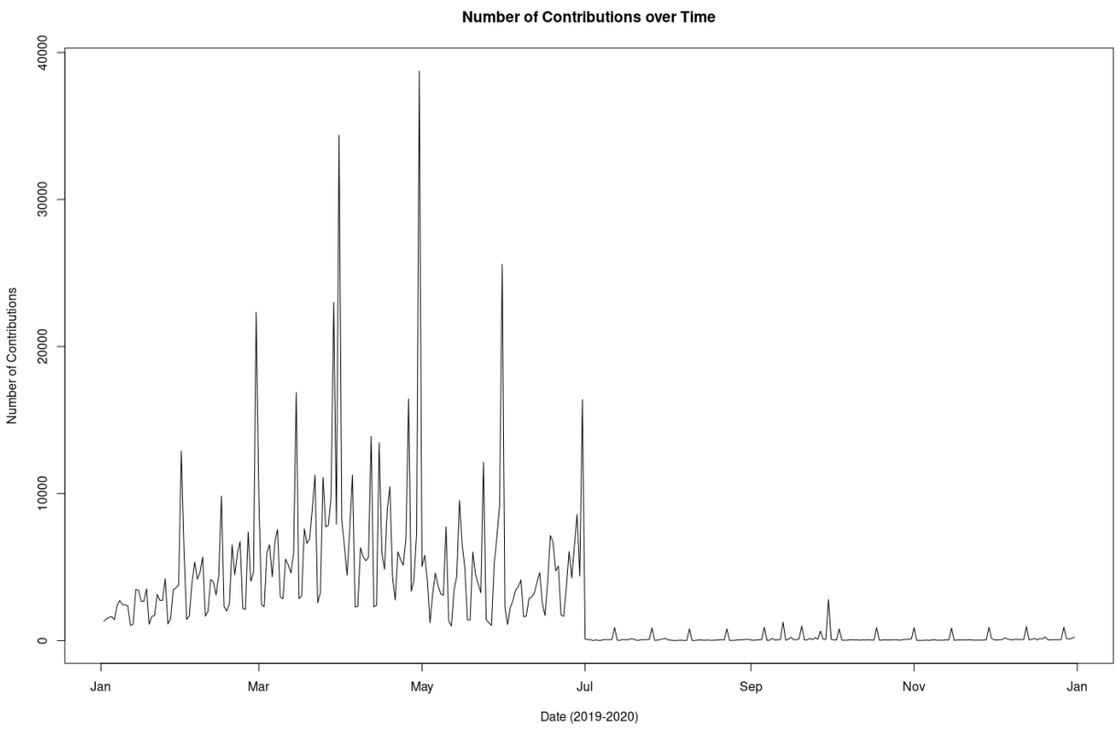
It is also important to note that while the code is segmented throughout the paper, it is intended to be run as one single script.

```
setwd('School/STA141B/HW1/')
library(openxlsx)
library(plyr)
library(dplyr)
library(tidyr)
# For color palette
library(wesanderson)
# Load in necessary variables and libraries
path = 'itcont.txt'
headers <- read.csv('indiv_header_file.csv')
# From Piazza - predetermining the classes of the columns
samp <- read.delim(path, sep = '|', header = FALSE,
                   nrows = 1000, col.names = colnames(headers))
classes <- sapply(samp, class)
classes[14] <- 'character'
# Working with 1 million samples
data <- read.delim(path, sep = '|', header = FALSE, nrows = 1e6,
                   col.names = colnames(headers),
                   colClasses = classes)
# Set color scheme.
scheme <- wes_palette('FantasticFox1', n = 51,
                      type = 'continuous')
# Convert transaction dates to Date object
data$TRANSACTION_DT <- as.Date(data$TRANSACTION_DT, '%m%d%Y')
```

3 A Time Series Analysis

As mentioned earlier, this data can be molded into a time series. However, in this report, the analysis is only of the first 1 million samples. After a certain date, we should expect a significant drop off in contribution count. With a powerful enough computer, this code could be scaled to analyze all 34 million observations.

```r
# Viewing contributions by date between each state
dates = data %>%
  group_by(TRANSACTION_DT) %>%
  # Filter through 2019-2020 dates
  filter(TRANSACTION_DT > '2019-01-01' & TRANSACTION_DT < '2020-12-31') %>%
  count(TRANSACTION_DT)
# View time series of dates:
plot(dates, type = 'l',
     main = 'Number of Contributions over Time',
     xlab = 'Date (2019-2020)',
     ylab = 'Number of Contributions')
```



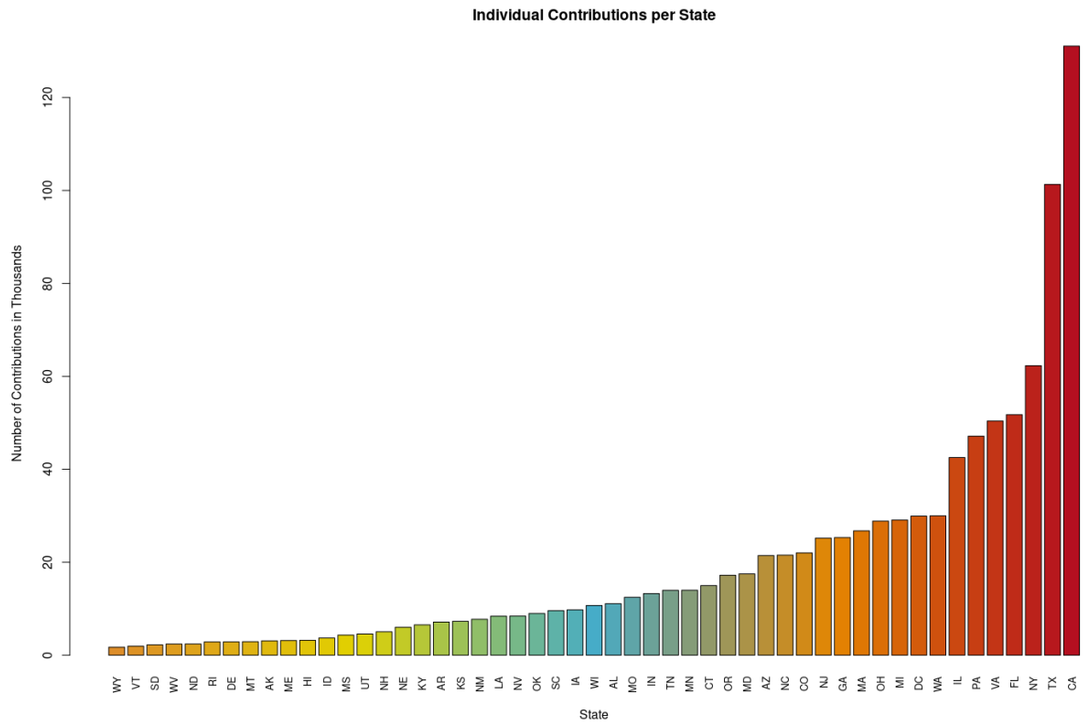
As we can see above, our assumption is correct. This implies that the first million observations predominately take place within the first seven months of 2019.

4 A State Analysis

A natural assumption is to break down and compare the individual contributions on a state-by-state level. When grouping by each state and finding their relative frequency, it is important to note that there are 62 states - with 61 of them being valid and the other being empty.

But, how could there be 61 valid state entries? When removing the typical 50 we are already aware of and DC, the remaining 10 constitute of a mixture of Commonwealth/Territories, Army States, Foreign States, and, of course, the empty state. The number of contributions by these 10 states are negligible in magnitude when compared to the 50 states and DC. Because of this, the main focus of this exploration will be the 50 states and DC.

```r
# State exploration:
# Including DC in our observations:
state.abb = append(state.abb, 'DC')
state.abb = state.abb[order(state.abb)]
states = table(data$STATE)
# Filter down to 50 states using inbuilt dataset in R and order states
states = states[names(states) %in% state.abb]
ordered.states = states[order(states)]
# Plot state data.
barplot((ordered.states / 1000), las = 3, cex.names = 0.8,
        main = 'Individual Contributions per State',
        xlab = 'State',
        ylab = 'Number of Contributions in Thousands',
        col = scheme)
```

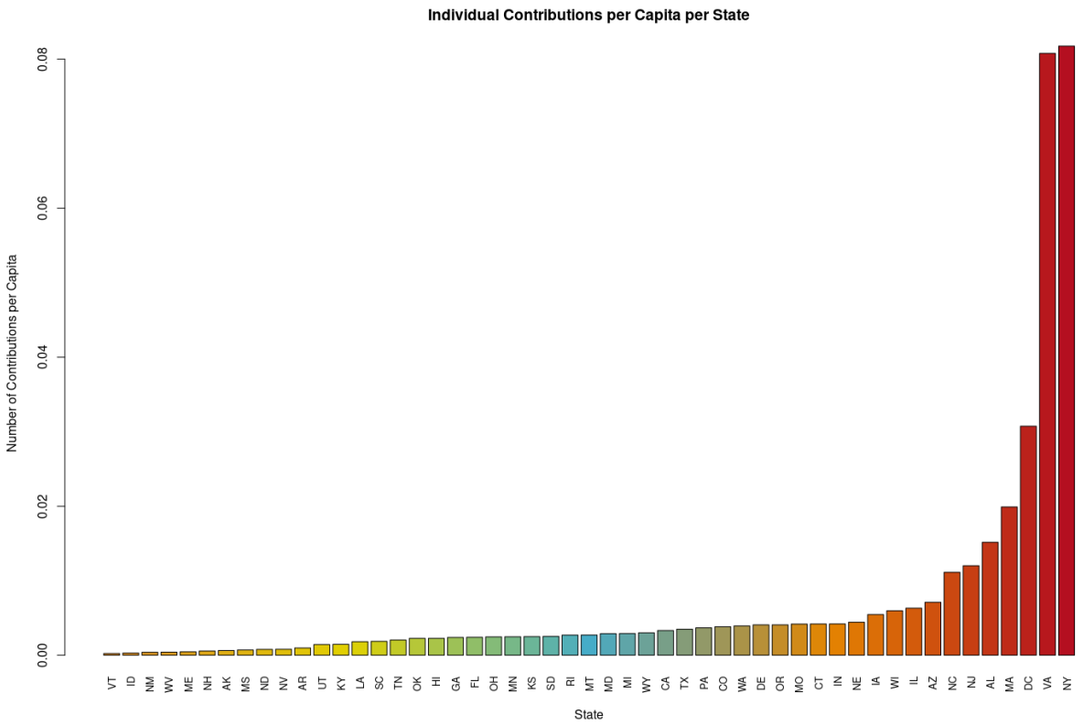


The pure number of votes alone shows for an exponential disparity among certain states. However, sheer volume alone is not representative of contributions' effects on elections. In fact, it should be expected that California and Texas have the largest turnout simply because they are the largest states. To try and combat this, visualizing the number of contributions per state per capita will be used.

To find the state per capita, we take our number of observations and divide it by the population of each state. The United States Census provides each of the 50 states' population and it is what will be used here. As expected, the excel sheet containing our information needed some cleaning as well. Important notes about the excel file are as follows:

- The column names were not in read in place due to irregular formatting.
- State names all contain a period in front of them (i.e. ".Texas" as opposed to "Texas".)
- The Contributions data set uses abbreviations (TX) for the states where as the Census data uses the state's actual name (Texas.)

```
# Finding state per capita
pop = read.xlsx('nst-est2019-01.xlsx')
# Change column name for States
names(pop)[1] <- 'States'
# Remove periods from State names (e.g. turn .Texas into Texas)
remove.period <- function(x){
  x = substr(x, 2, nchar(x))
  return(x)
}
pop$States <- sapply(pop$States, remove.period)
state.name <- append(state.name, 'District of Columbia')
state.name <- state.name[order(state.name)]
# Return population of states in 2019 alphabetically
pop.2019 <- pop$X13[pop$States %in% state.name]
abb.population <- structure(pop.2019, names = state.abb)
# Ensure order is the same
per.capita <- states[sort(names(abb.population))] / abb.population[sort(names(abb.population
    ))]
# Sort for visual aestetics
per.capita <- per.capita[order(per.capita)]
# Plotting barplot
barplot(per.capita, las = 3, cex.names = 0.8,
        main = 'Individual Contributions per Capita per State',
        xlab = 'State',
        ylab = 'Number of Contributions per Capita',
        col = scheme)
```

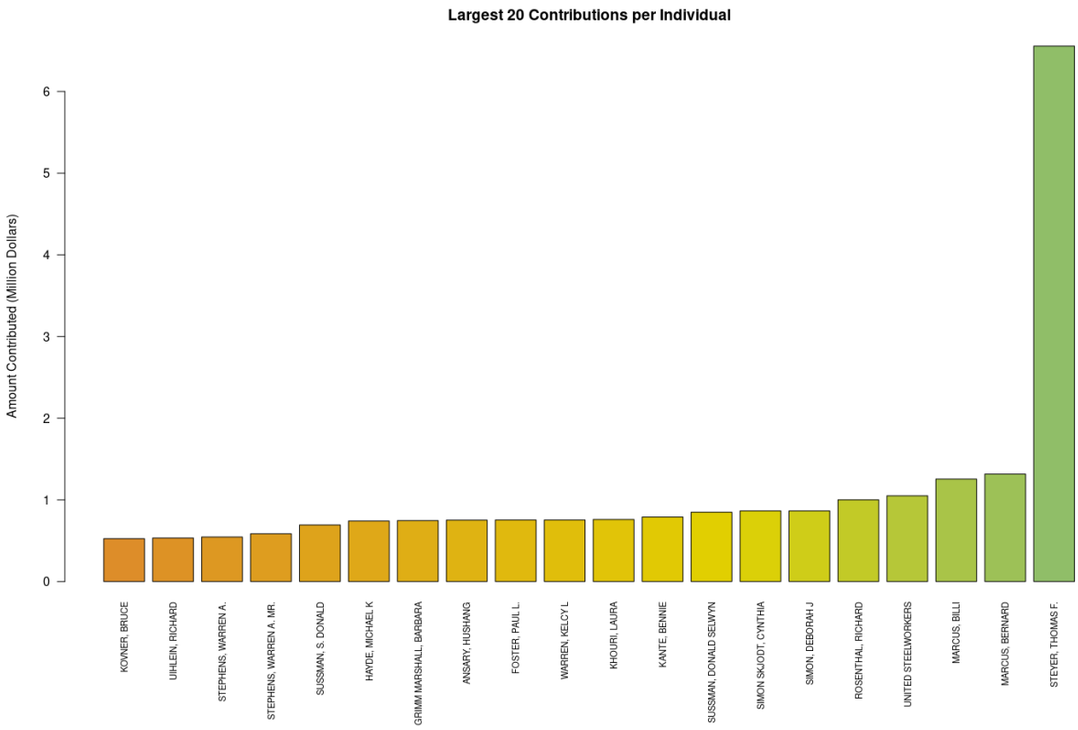


As we can see from the results above, New York and Virginia had the largest contribution count per capita, likely making them key states for fundraising. This further solidifies the notion that the number of contributions alone is not representative of the overall campaign process.

5 Who's in the Lobby?

Being that this data set's main focus is on looking at individual contributions, it is only suitable to analyze the individuals themselves. However, there are nearly infinite variations of names that could arise in a million contributions. There needs to be a straight forward method to seeing who, as an individual, has a deep passion - and even deeper pockets - for politics. To find our hometown lobbyist, I grouped by all the names and summed up the column representing the amount of money contributed. From there, we can take the top 20 contributors and see how much they were willing to shell out.

```
# Individual Study
# Group by name, sum transaction amount
# then arrange by quantity.
# https://github.com/rstudio/cheatsheets/blob/master/data-transformation.pdf
names = data %>%
  group_by(NAME) %>%
  summarise(totalContribution = sum(TRANSACTION_AMT)) %>%
  arrange(totalContribution)
# Taking 20 lowest and highest amounts contributed
lobbying <- names %>% top_n(20)
# Plotting highest 20 contributions
# https://rquicktips.wordpress.com/2012/10/05/how-do-i-prevent-my-tick-mark-labels-from-
    being-cut-off-and-running-into-the-x-label/
par(mar=c(10, 4.1, 4.1, 2.1))
barplot(lobbying$totalContribution / 1e6,
        las = 2, cex.names = 0.7,
        names.arg = lobbying$NAME,
        main = 'Largest 20 Contributions per Individual',
        ylab = 'Amount Contributed (Million Dollars)',
        col = scheme)
```



This analysis is not perfect, however. We should remember that these are the largest contributors within the first half of 2019. Also, when looking at the third and fourth bar to the left, we see that the contributions are by the same person. However, in one instance, his title (Mr.) is in his name, whereas in the other, it is not. Basic regular expressions could be used to clean through the names by omitting titles such as Mister, Miss, and so on. However, this does not neglect the large sums contributed by each individual. Even with the top 20 contributors, there is a significant disparity between Thomas F. Steyer (who contributed the most) and Bruce Kovner (who contributed the "least") by a factor of over 8 times!

6 Conclusion

With a more powerful computer, the analysis performed above can bring much deeper insights into the sorts of money campaigns can come into in this country. Despite Super PACs dominating the political sphere with their massive outreach and influence, the top 20 contributors alone in the first leg of 2019 make up over $10 million worth of contributions. Our time series is the analysis that suffered most from the lack of computational resources. This was expected and confirmed in section 3. Because of this, it would also be safe to assume that the state-by-state data could also look different than the results in this report as we cannot use the first half of 2019 to predict another year's worth of contributions.