Assignment 1 - FEC Individual Contribution Data Analysis

Yifu Wu (916619585) *

1 Reading the Data Sets

The data set we used this time is a tabular data from the US Federal Election (FEC), which contains information for contributions to the presidential election given by individuals.

In order to analysis the data, we need to first read the data into R. To identify the structure of the data, I read the first five lines, determining that there were no column names (header) in the file and the separator was the "|" character (vertical bar). The first challenging part was the error in line 647 when I read first 1000 lines. However, after splitting each line into several character, I could not find any error since all lines had 21 characters. Then I counted number of fields in each of the lines, finding that the apostrophe caused several NAs in the result, which means that I should add quote argument into the code.

The second challenge part was the error in line 3166 when I tried to read first 10000 lines. I used the same method as before and found that for line 3166, the number of fields did not equals to 21. This was because the "#" character served as the comment character, so I turn off the interpretation of comments altogether by adding the comment.char argument.

The third challenge part was when converting the transaction date to a Date object, I had to prefix values with a leading 0 character since R just recognized this column as the integer. Based on the data type in the official description file, I specified each column using colClasses argument.

After finishing debugging process, I imported the associated CSV file from the "Header file" link and matched each column with the correct column name. The only thing we should care about was to make sure that the header names were character vectors of a data.frame.

The last challenging part was to find the population data and combine it with the itcont data. I found the state population data at https://worldpopulationreview.com/states. This Web page contained the 2020 state population and the annual population growth rate. Other than 50 states in the US, this data also included Puerto Rico and District of Columbia. Since I also needed the 2016 state population, I estimated it by assuming the annual population growth rates from 2016 to 2020 were the same, because the US society was under steady development, which leaded to a steady population. To connect population data with itcont data, I first matched the full state name in the population data with state abbreviation, and then selected rows in itcont data which state also appeared in the population data. When plotting the number/amount of contributions by state per capita, I just needed to merge two datasets by state abbreviations.

To verify the data were correct, I first ran the whole dataset to make sure there was no error message. Then I got the factor level of each column and compared with the official description file to make sure there was no wrong data.

*Department of Statistics, UC Davis and ifuwu@ucdavis.edu

2 Exploring and Interpreting the Data

In this section, I made several plots and tables to explore the dataset. Before summarizing the key features and findings, I needed to explain the reason why I chose such a special part. When I tried to plot the number of contribution by day, I found that most of the contributions (more than 99.9%) were ranging from 2019-01-01 to 2020-10-01 (current time), others were future contributions and past contributions. The contribution amount in this time period also made up more than 99.9% of the total amount in 2020 election cycle, which indicated that most of the contribution for the 2020 presidential election were made in the recent two years. So I decided to focus on this part of the data. Another issue was that because of the limitation of the population data, I only explored the data in those 52 places in the US, since I only cared about the contribution in the US, and also those places contributed the most part of the total contribution(99.7%). The itcont data also contained a column which indicated whether a specific contribution was not included in the itemization total. After explored this part of the data, I found that those only made up a very small portion of the contribution, so I did not omit those data. I also needed to deal with the 2016 election data, since the situation was rather similar with 2020 data, I just simply applied the same processing steps to the 2016 election data. In this section, for the single year plots, I tried to use barplot since it could make it clear to see the number/amount of contribution versus date or state. However, for the comparison plots, the dot plot made much sense, since the different color of dots could show the difference between two years clearly.

2.1 Explore the number of contribution in 2020 Election Data

For the 2020 election data, I first discovered the relation between number of contribution and the date. When I first tried to plot the number of contribution by day, it was hard to observe since there were too many dates, and the number of contribution for each day varied a lot. Therefore, I plotted the number of contribution by month, which was shown in Figure 1 below. From the barplot we know that from 2019-01 to 2020-06, the number of contribution was increasing month by month and peaked in 2020-06. There was a severe drop between 2019-12 and 2020-02, and there were two possible reason, one was due to the outbreak of covid-19, the other was that loss of data in this time period. But I am not sure what was the real reason, more information was needed to verify the speculation.

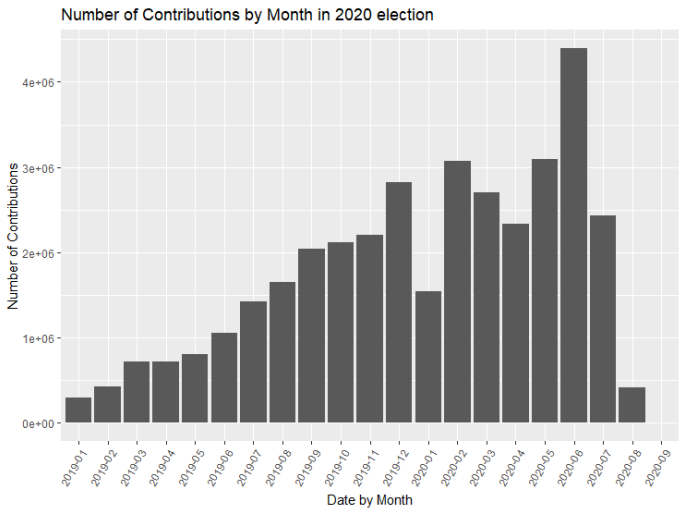

Figure 1: Number of Contributions by Month in the 2020 Presidential Election

Then I explored the relation between number of contribution and state, which was shown in the Figure 2 below. From the plot I knew that the number of contribution varied a lot between states, and I could also know that CA, NY, TX and FL were the top four states, that was because those four states had most population.

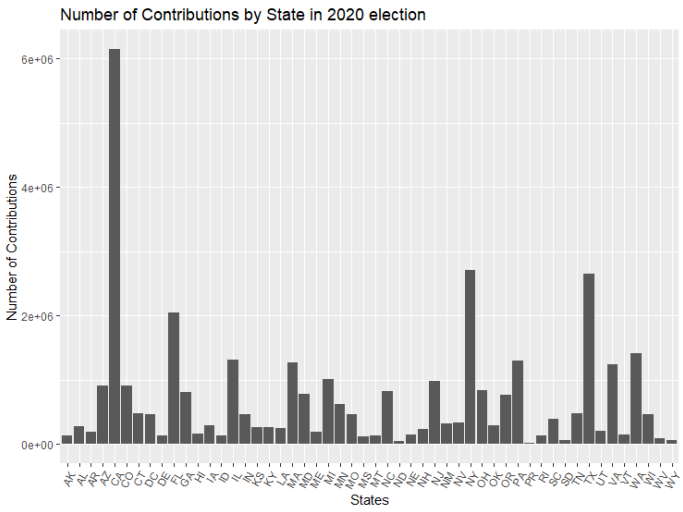

Figure 2: Number of Contributions by State in the 2020 Presidential Election

I also discovered the number of contribution by state per capita in 2020 election. For most states, the number of contribution per capita were around 0.1 - 0.2. However, DC had it more than 0.6, I could speculate that since it was the political center of the US, most people there cared about and had some connection to the election, which caused the result. While in PR, which is an unincorporated territory of the US located in the northeast Caribbean Sea. Maybe most people there did not care as much as the 50 state in the US about the election, which might lead to the lowest number per capita.

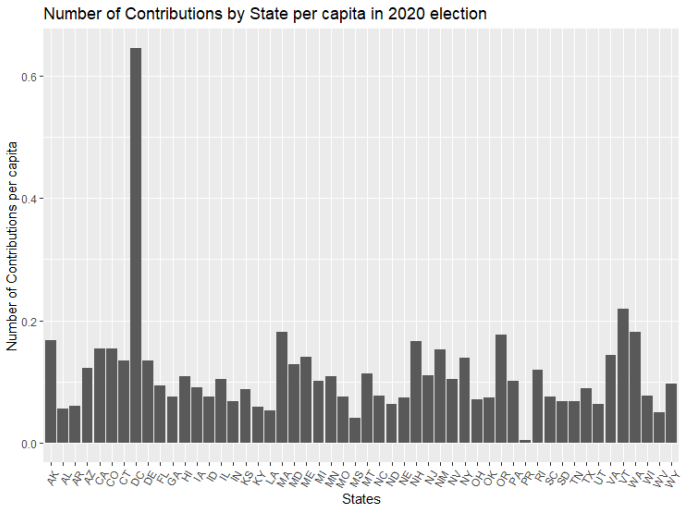

Figure 3: Number of Contributions by State per capita in the 2020 Presidential Election

2.2 Explore Other Aspects in 2020 Election Data

Other than the number of contribution, I also wanted to explore the relationship between amount of contribution and month. From the plot I knew that from 2019-01 to 2020-02, the amount of contribution was increasing month by month and peaked in 2020-02. Although there was a drop between 2020-03 and 2020-05, the amount of contribution in 2020-06 was still the second highest among two years. Combined with the number of contribution, I could say that although some months did not have a large number of contribution, they still had a huge amount of contribution. Therefore, we should use both plots to determine which months had more contributions.

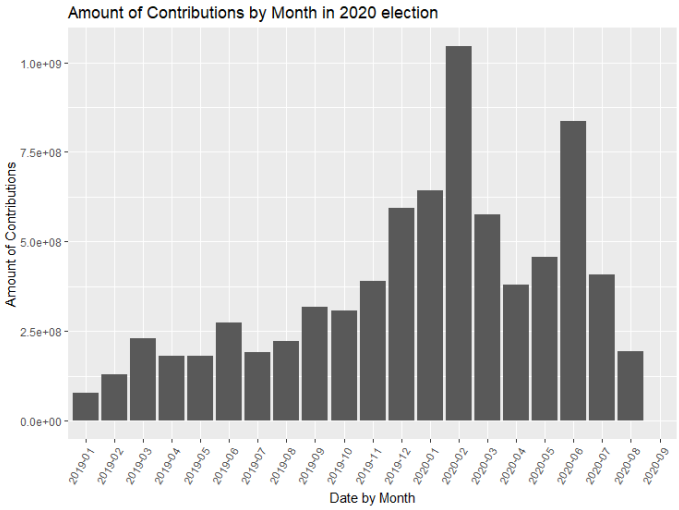

Figure 4: Amount of Contributions by Month in the 2020 Presidential Election

When exploring the entity type of the 2020 election, I found that although the number of individual who made contribution was much more than candidate, organization, and committee, the amount of contributions were not much higher than other entity. I could conclude that although a huge number of individual would like to contribute to the election, each of them could contribute little amount of money. However, for other entities like candidates or committees, each of them could contribute a large amount of money. It also indicated that election was a costly activity for candidates.

Entity Type	**Number of Contribution**	**Amount of Contribution**
Individual	36194103	5613846546
Candidate	13834	1442077475
Organization (not a committee and not a person)	13130	448459880
Political Action Committee	3119	98406294
Candidate Committee	642	1122677
Committee	614	21515599
Not recorded	51	73325
Party Organization	26	53413

Table 1: Number and Amount of Contributions in Different Entity Type in 2020 Presidential Election

Since there were too many transaction type in the data, I chose the top 6 transaction type in both number and amount of contribution. The detailed description of the transaction type codes was in the Appendix. We may got the same conclusion as above. Although transaction type 15C (Contribution from candidate) had the lowest number of contribution among 6 type, the amount of contribution was on the top three of all transaction type. It clearly indicated that each candidate would pay a lot of money for election. Another interesting phenomenon was that transaction type 24T (Earmarked contribution passed to intended recipient from intermediary's treasury) had the second highest number of contribution, but the amount was relative low. One possible reason was that although plenty of people chose or had to contribute by intermediary, they still did not fully trust the intermediary because intermediary were not authorized to raise funds [Ref 1]. So people tended to contribute small amount of money through the intermediary.

Transaction Type	**Number of Contribution**	**Amount of Contribution**
15E	16754612	1443636122
24T	10443272	999542881
15	8455362	2200023265
22Y	315275	76750163
10	223580	1359077426
15C	12904	1442627561

Table 2: Top 6 Number and Amount of Contributions in Different Transaction Type in 2020 Presidential Election

2.3 Comparison between 2016 and 2020 Presidential Election

In this section, I would make comparison of number and amount of contribution between 2016 and 2020 presidential election. By summarizing both data set, I knew that the number of contribution was 20385563 in 52 regions during 2015-01 to 2016-12, the amount of contribution was 6066741135 in 52 regions during 2015-01 to 2016-12. While during 2019-01 to 2020-09 the number of contribution was 36225519, and the amount of contribution was 7625555209. Although the data for 2020 election might not be the final version, it was still much higher than 2016 election. Then I compared the number and amount of contribution by state (and also per capita) in 2016 and 2020 election.

The comparison of number of contribution by state between 2016 and 2020 election was shown in Figure 5 below. From the plot we know that for most regions the number of contribution in 2020 is higher than 2016. Also we knew that the difference between each two regions in 2016 were rather similar with the difference in 2020.

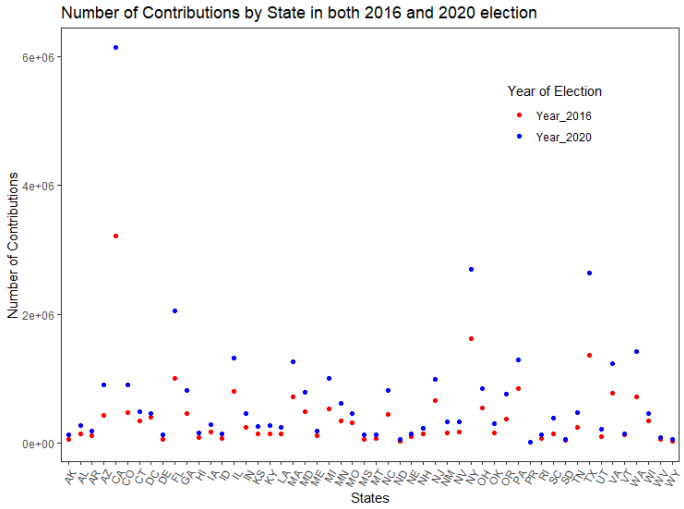

Figure 5: Comparison of Number of Contributions by State in the 2016 and 2020 Presidential Election

The comparison of amount of contribution by state between 2016 and 2020 election was shown in Figure 6 below. Similar with the analysis above, most of the regions had higher amount of contribution during 2020 election. But for several regions like DC, FL, etc. had a higher amount in 2016, which indicated that although more people were willing to contribute in 2020, people tended to contribute less in 2020, which can be proved in the Figure 8 below.

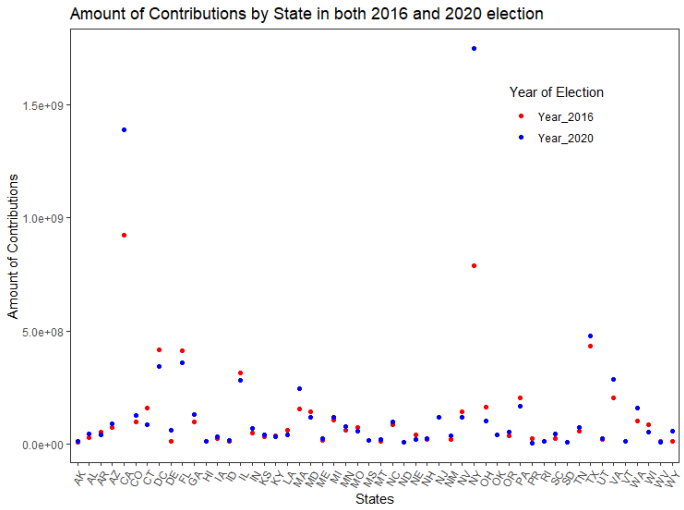

Figure 6: Comparison of Amount of Contributions by State in the 2016 and 2020 Presidential Election

The comparison of number of contribution by state per capita between 2016 and 2020 election was shown in Figure 7 below. We could see that all regions in 2020 had a significant higher number of contribution per capita compared with 2016. Also we knew that the difference between each two regions in 2016 were rather similar with the difference in 2020. The conclusion is rather similar with Figure 5 above since the population in 2016 was estimated based on 2020 population with a constant growth rate.

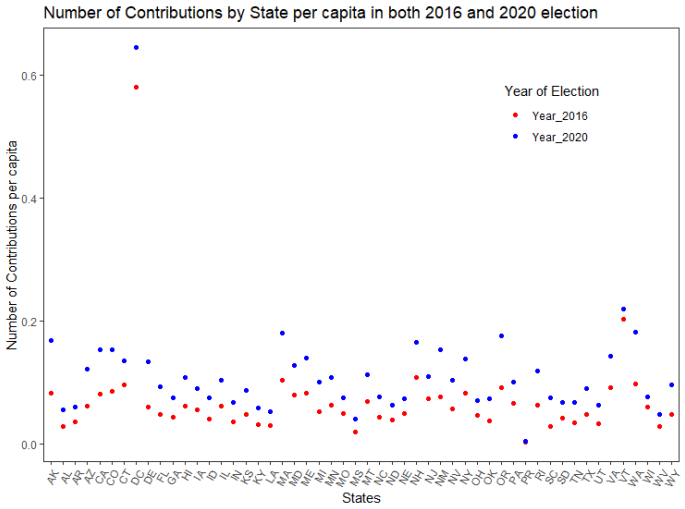

Figure 7: Comparison of Number of Contributions by State per capita in the 2016 and 2020 Presidential Election

The comparison of amount of contribution by state per capita between 2016 and 2020 election was shown in Figure 8 below. We could know that for most regions, the amount per capita was rather similar, except for DC, DE, NY and WY. For DC, people was willing to pay less for the election, which could prove the conclusion in Figure 6 above. While for DE, NY, and WY, people had positive attitude to contribute the election.

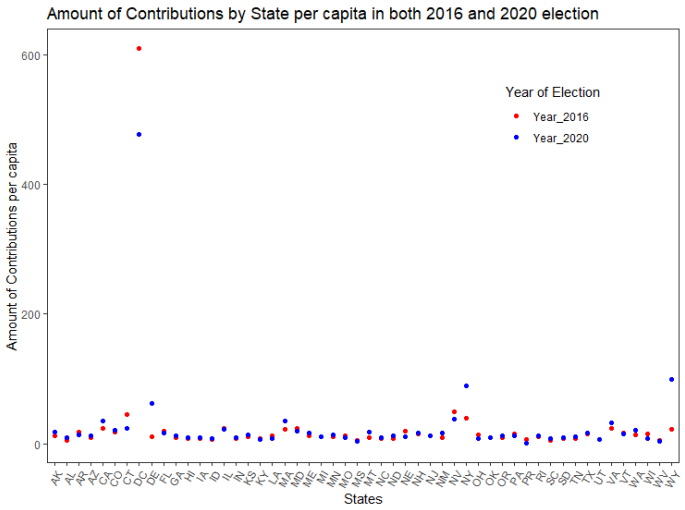

Figure 8: Comparison of Amount of Contributions by State per capita in the 2016 and 2020 Presidential Election

3 Appendix

This part will show the description of the transaction type code:

10 : Contribution to Independent Expenditure-Only Committees (Super PACs), Political Committees with non-contribution accounts (Hybrid PACs) and nonfederal party "soft money" accounts (1991-2002) from a person (individual, partnership, limited liability company, corporation, labor organization, or any other organization or group of persons)

15 : Contribution to political committees (other than Super PACs and Hybrid PACs) from an individual, partnership or limited liability company

15C : Contribution from candidate

15E : Earmarked contributions to political committees (other than Super PACs and Hybrid PACs) from an individual, partnership or limited liability company

22Y : Contribution refund to an individual, partnership or limited liability company

24T : Earmarked contribution passed to intended recipient from intermediary's treasury (treasury out)

4 References

1. https://www.fec.gov/updates/earmarked-contributions/