Project 1 Report

CS567: Computational Statistics

Andrew Dunn
Justin Phan Phan
Mark Qin
Sridevi Wagle

Department of Computer Science
Central Washington University
Ellensburg, Wa

March 6, 2020

Contents

1 Introduction

1.1 Brief Introduction

For this project we applied basic statistical methods to the 2019-2020 Democratic Primary election cycle. An election for the President of the United States happens every four years and is the largest political event in the USA. Leading up to the general election, most of the candidates running for president go through a series of state primaries and caucuses. Although these primaries and caucuses are ran differently, they both serve the same purpose. They alow every state to choose their major political parties' nominees for the general election[1]. The campaigning process lasts at least a year for most candidates and costs many millions of dollars in advertising, outreach, and travelling. Since the Republican Presidential Nominee is almost guaranteed to be President Donald Trump, we focused on the Democratic Nominee Primaries. Below you will find the list of objectives used in this project:

1.2 Objectives

- To monitor polling data for the Democratic Party primary candidates
- To read the polling data from source into a R dataframe, filter and plot polling data over time and to run analysis and prediction tools using R script
- To plot the fundraising for Democratic candidates
- To plot the number of delegates awarded for candidates (state wise)

2 Design

2.1 Data Set

All of the polling data used in our analysis was downloaded from FiveThirtyEight, a well known political blog that does predictions for various elections in the USA[2]. They collect polling data from a large number of different pollsters and then combine all into a single *.csv* file that is available for download. This data set is updated on a regular basis with new polling results, and contains many different columns such as pollster name, state, sponsor, sample size, source, poll start and end dates, and poll results as a percentage of votes for each candidate. In addition, FiveThirtyEight rates all pollsters based on their quality and level of bias. Ratings range from A+ down to F. Pollsters with the least bias have the best ratings. Using their ratings, we are able to filter the data set such that only highly rated polls are considered. Other than polling data, FiveThirtyEight presents many other aspects of the election we can analyze such as candidate funding and number of delegates.

2.2 Implementation

For our experiments we implemented a program in R script that can import the data set from FiveThirtyEight and generate various plots of the polling data. Using these plots we can visualize the data and conduct some analysis, looking for trends. Figure 1 below contains an example code snippet that shows how we filter out the important features from the data set.

```
# Load data set from csv file
pdata<-read.csv("data/president_primary_polls.csv", header = TRUE)

# Filter out democrats and needed columns
pdata_dem<-pdata[pdata$party=="DEM",
                 c("question_id", "poll_id", "pollster_id",
                   "fte_grade", "end_date", "candidate_id",
                   "candidate_name", "pct")]

# Replace blank cells with NA
pdata_dem[pdata_dem == ""] = NA

# Replace data strings with date data type
pdata_dem$end_date <- as.Date(pdata_dem$end_date, format = "%m/%d/%y")

# Filter out date range
pdata_dem<-pdata_dem[pdata_dem$end_date >= "2020-01-01",]

# Sort dataframe by poll end date, question id, and candidate id
pdata_dem<-pdata_dem[order(pdata_dem$end_date,
                     pdata_dem$question_id, pdata_dem$candidate_id),]

# Filter out all polls that are rated worse than A-
pdata_dem<-subset(pdata_dem, fte_grade == 'A+' |
                  fte_grade == 'A' | fte_grade == 'A-')
```

Figure 1: Example Code Snippet - Data Filtering

3 Model

3.1 Delegates

Figure 2 shows the number of delegates awarded after each primary/caucus per candidate. The results are displayed for each state up until Super Tuesday. Biden is awarded with the most number number of delegates compared to other candidates. Buttigieg acquired the least number of delegates.

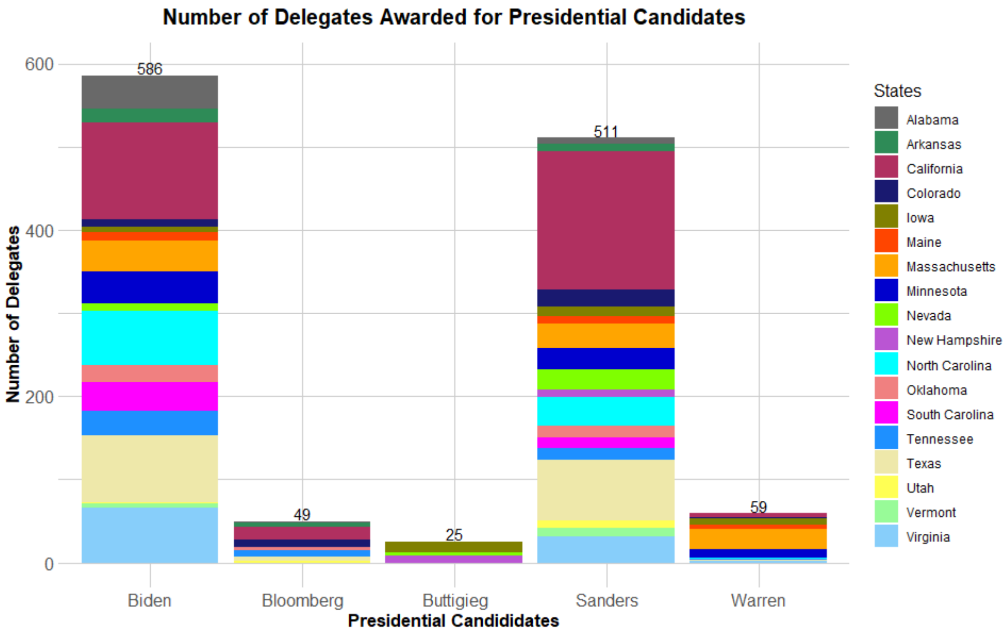

Figure 2: Delegate

3.2 Polling

The poll data is collected and plotted for top five democratic candidates(Biden, Sanders, Warren, Bloomberg and Buttigieg). Data is plotted for both A rated and for all the polls as displayed in figures 3, 4 and 5.

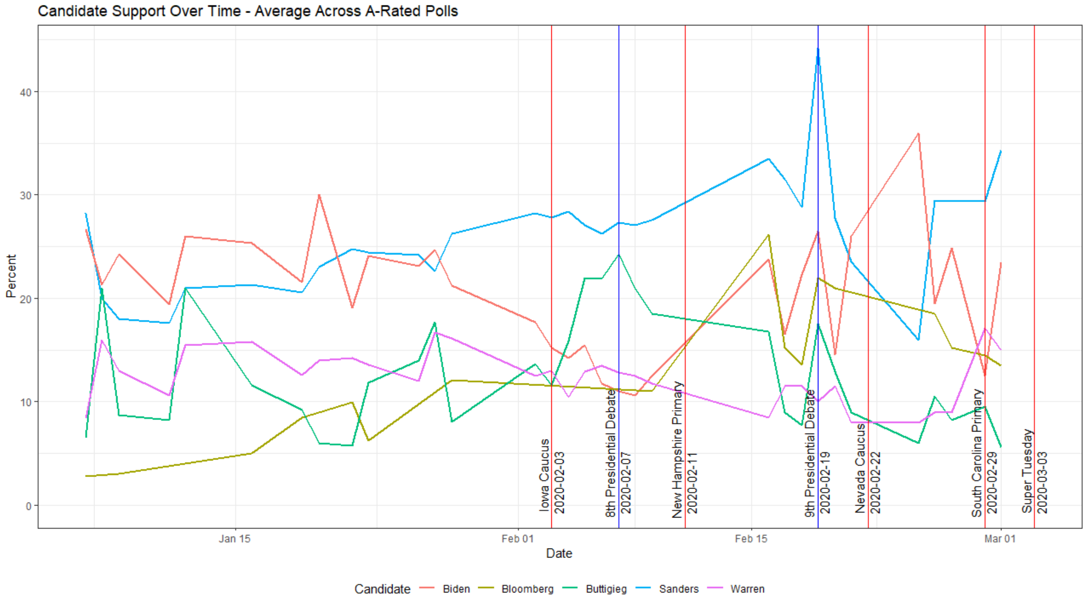

Figure 3: Updated Polling Data

Figure 4 displays the support for candidates collected from A rated polls. At the end of South Carolina Primary , it is observed that Sanders has the highest percentage of support is the highest and Buttigieg has the lowest percentage of support.

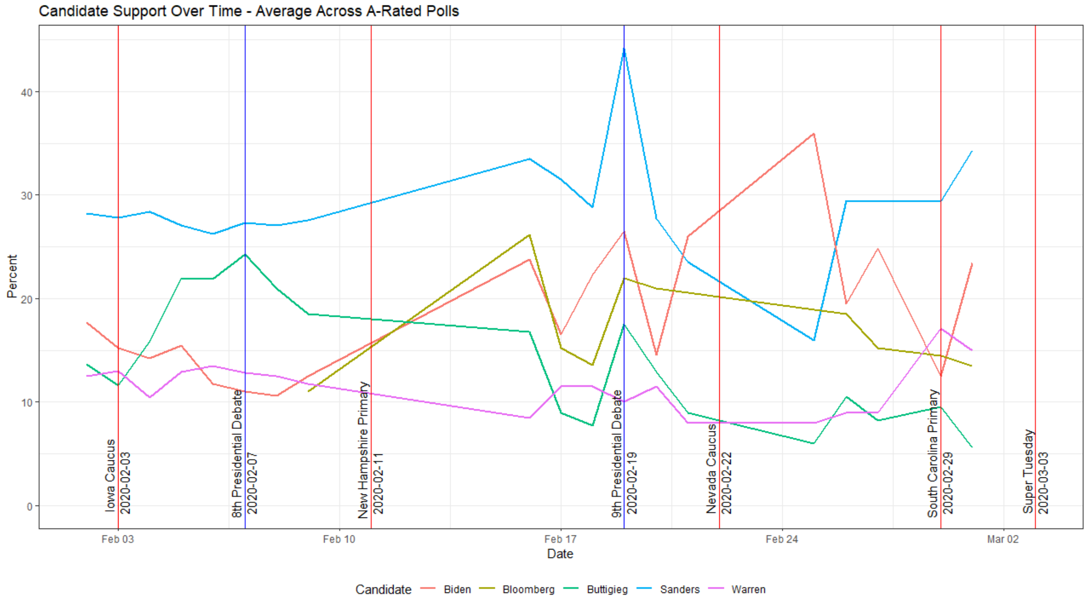

Figure 4: A Rated Polls

Figure 5 shows support for candidates until Super Tuesday from all rated polls. Biden's support shows a drop, making Biden the top candidate with highest percentage of support and Buttigieg with the lowest support percentage.

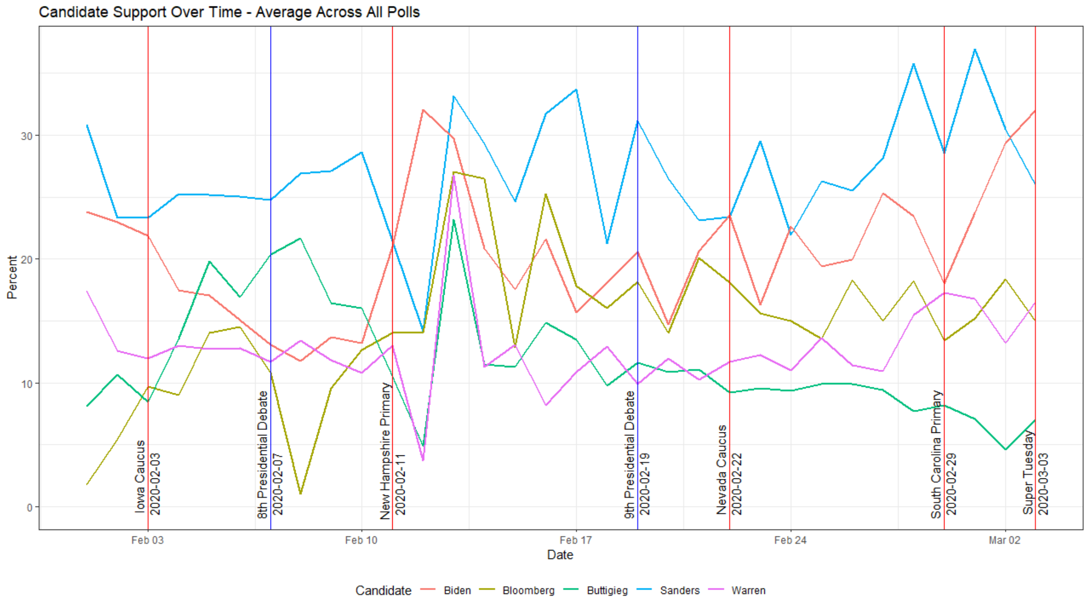

Figure 5: All Rated Polls

Figure 6 shows support for candidates until South Carolina Primary. A slight drop is observed in Sanders' percentage of support after Nevada Caucus. Though there's a steady increase in Warren's support after Nevada caucus and South Carolina Primary, she still has the lowest percentage of support.

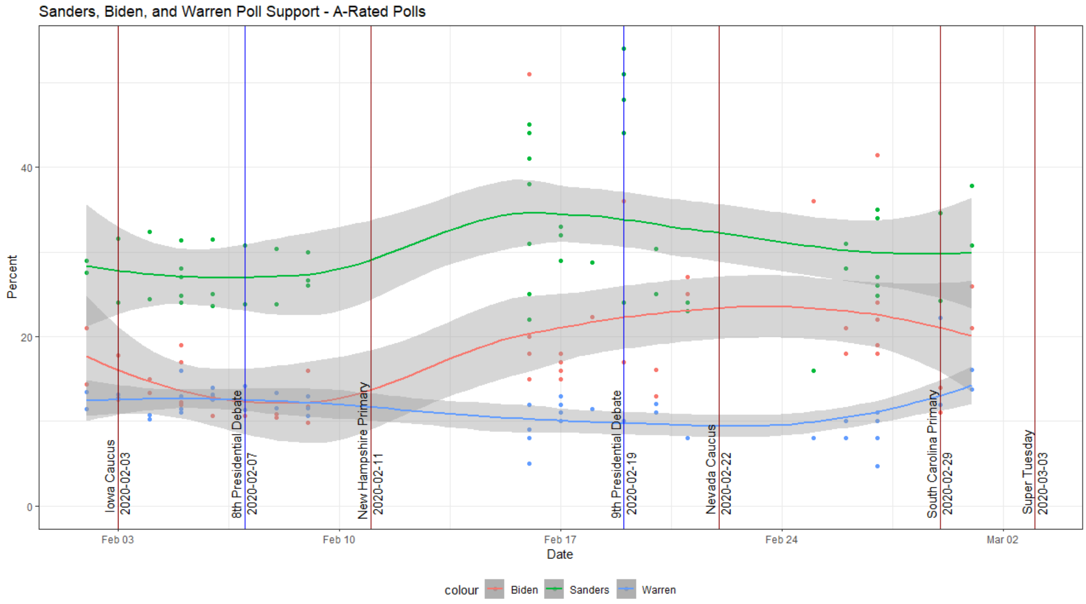

Figure 6: Scatter Plot for A Rated Polls

Figure 7 shows support for candidates until Super Tuesday. There is a steady increase in Sanders' percentage of support after South Carolina Primary. There is a sharp increase in Biden's percentage of support after South Carolina Primary,almost on par with Sanders. Warren's support has been consistently the lowest percentage of support compared to other candidates.

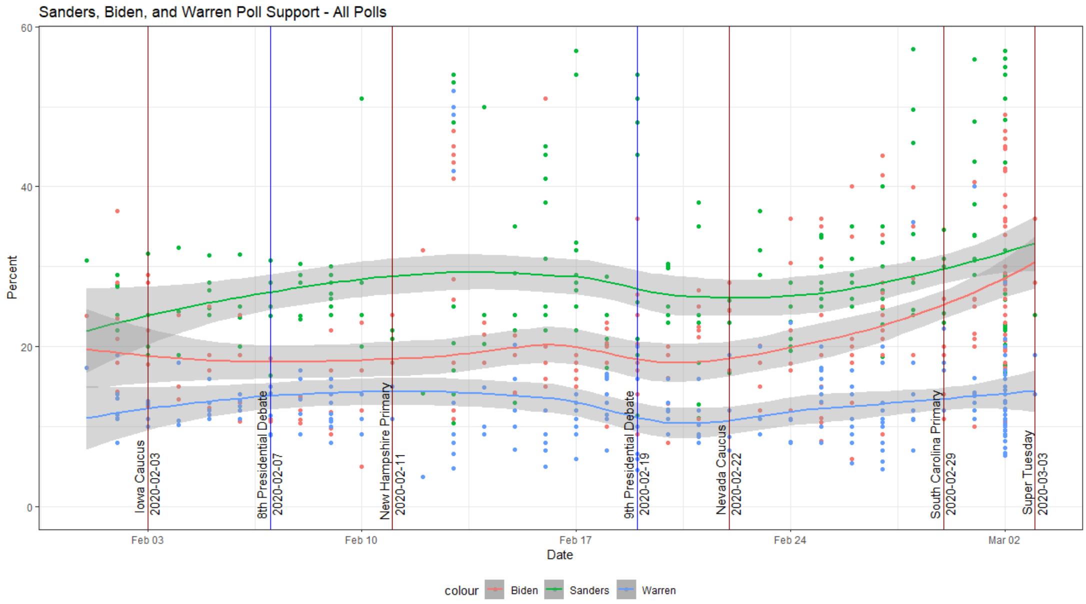

Figure 7: Scatter Plot for All Rated Polls

Figure 8 shows the linear trend for candidate support until Super Tuesday. Overall, starting from Iowa Caucus to Super Tuesday, Sanders' support increases steadily whereas a progressive growth for support is observed in Biden's support percentage. Warren's support percentage shows a slight dip overtime.

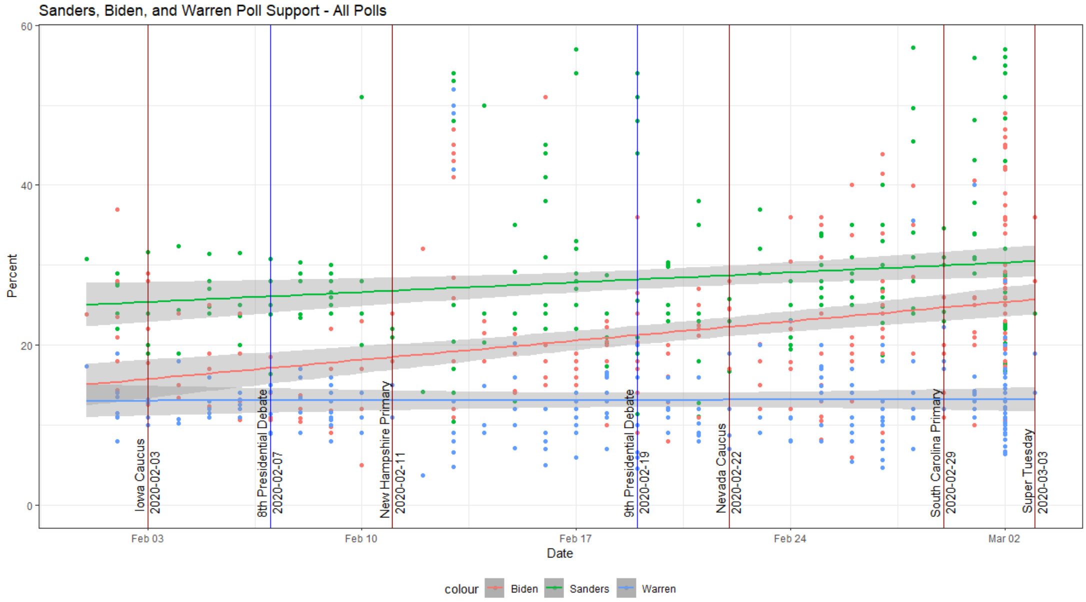

Figure 8: Scatter Plot for All Rated Polls - Linear

3.3 Spending

Figures 9 shows the total amount of money spent on advertisements. The highest amount is spent by Bloomberg on ads. However, after the Iowa caucus, there is a dip in Bloomberg's percentage of support, eventually dropping out of the Presidential Election 2020.

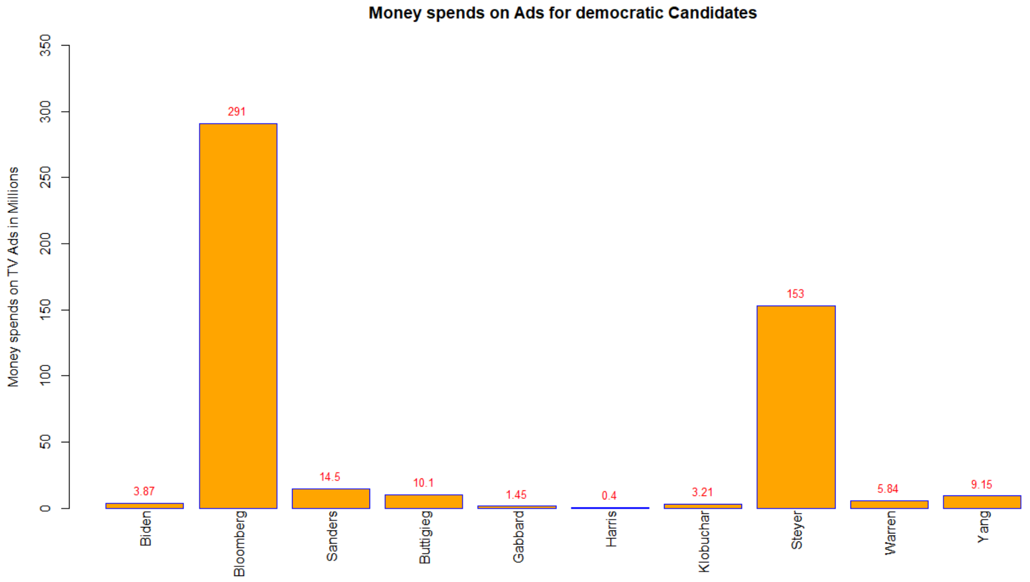

Figure 9: Money spending on Ads

Figures 10 shows the amount of money spent on advertisements in Iowa. The highest amount is spent by Stayer in Iowa. However, Steyer ended up with 0 delegates after the Iowa caucus.

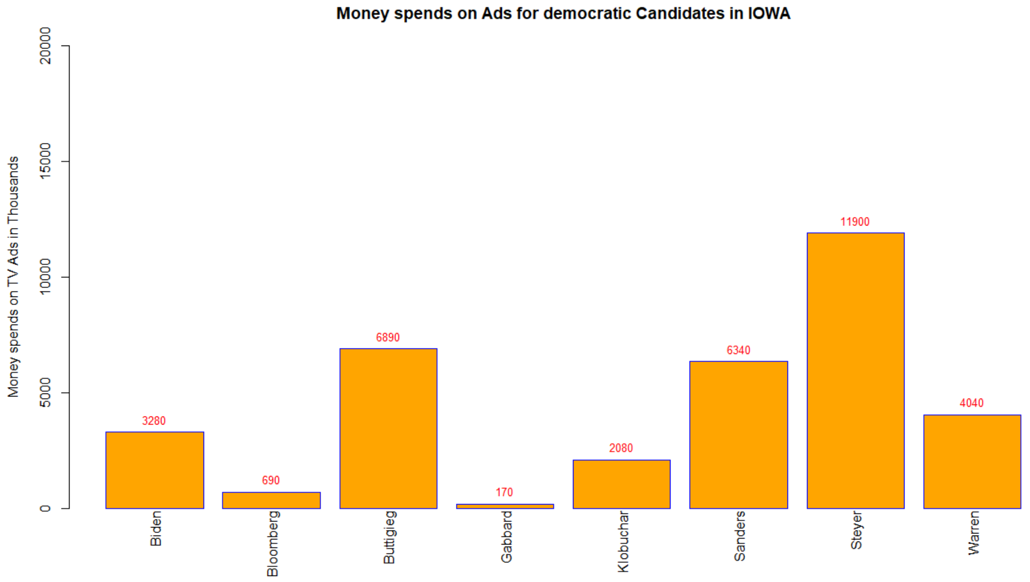

Figure 10: Money spending on Ads in Iowa

Figures 11 shows the amount of money spent on advertisements in New Hampshire. The highest amount is again spent by Stayer and he ended up with 0 delegates after the New Hampshire primary.

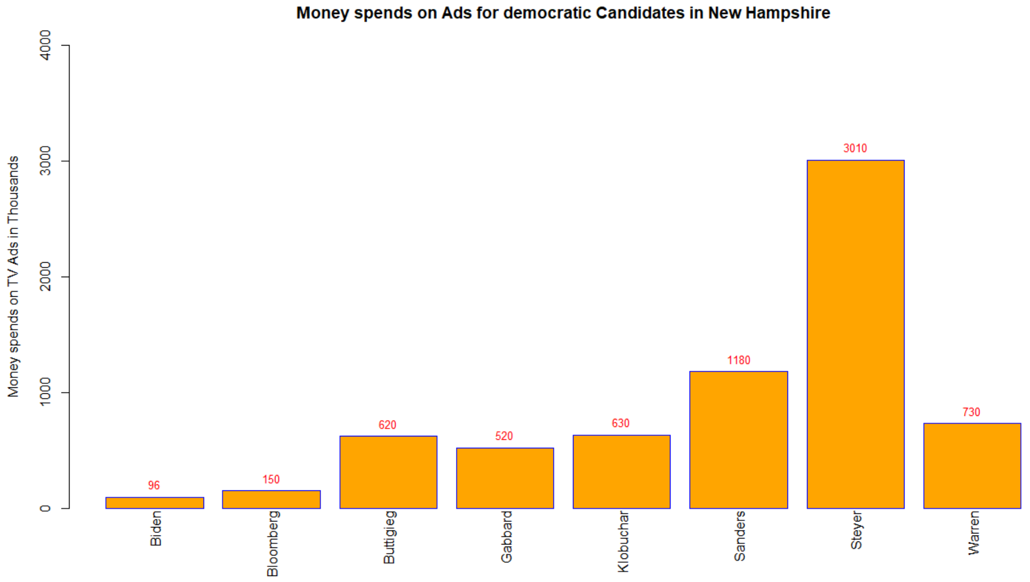

Figure 11: Money spending on Ads in New Hampshire

Figures 12 shows the amount of money spent on advertisements in Nevada. The highest amount is spent by Stayer, again ending up with 0 delegates after the Nevada caucus.

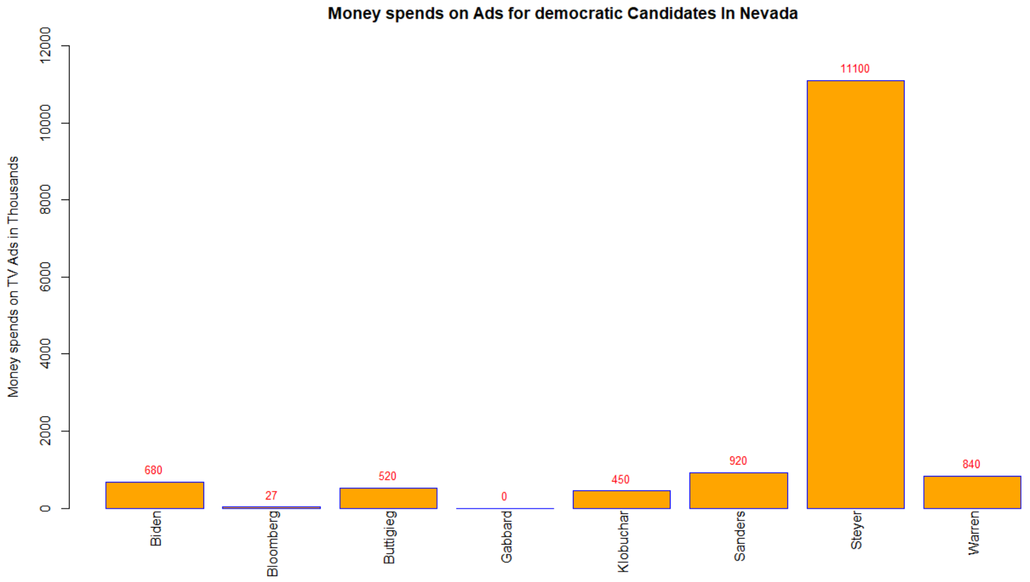

Figure 12: Money spending on Ads in Nevada

3.4 Funding

Figure 13 shows the total funding for presidential elections. Steyer is funded with maximum amount and Gabbard with the least amount for the elections.

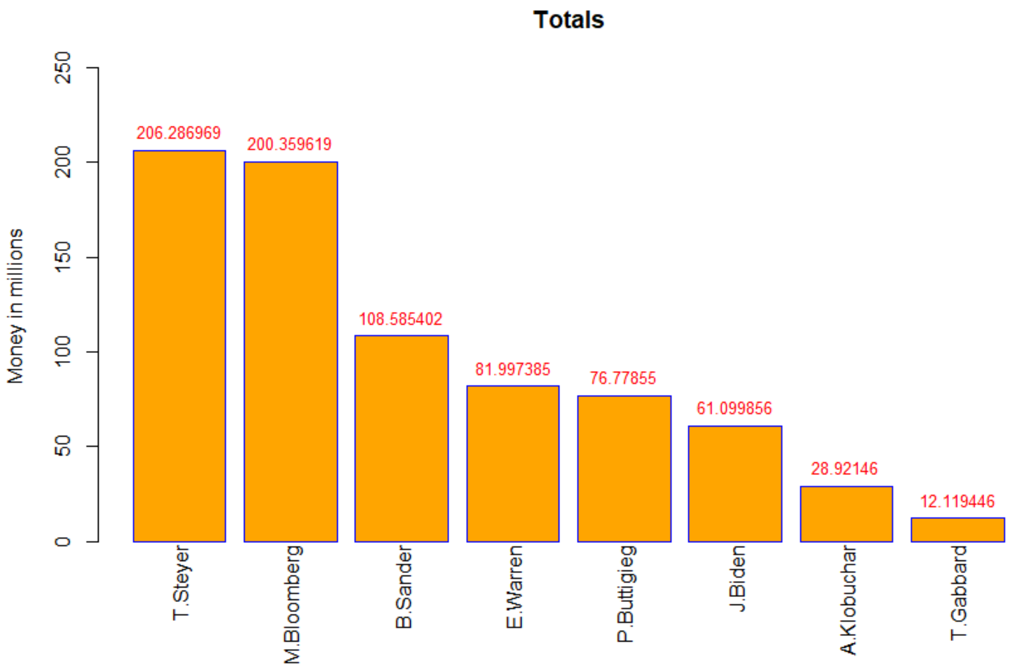

Figure 13: Total Funding

Figure 14 shows the funding from small donors for presidential elections. Sanders is funded with maximum amount and Bloomberg with the least amount from small donors for the elections.

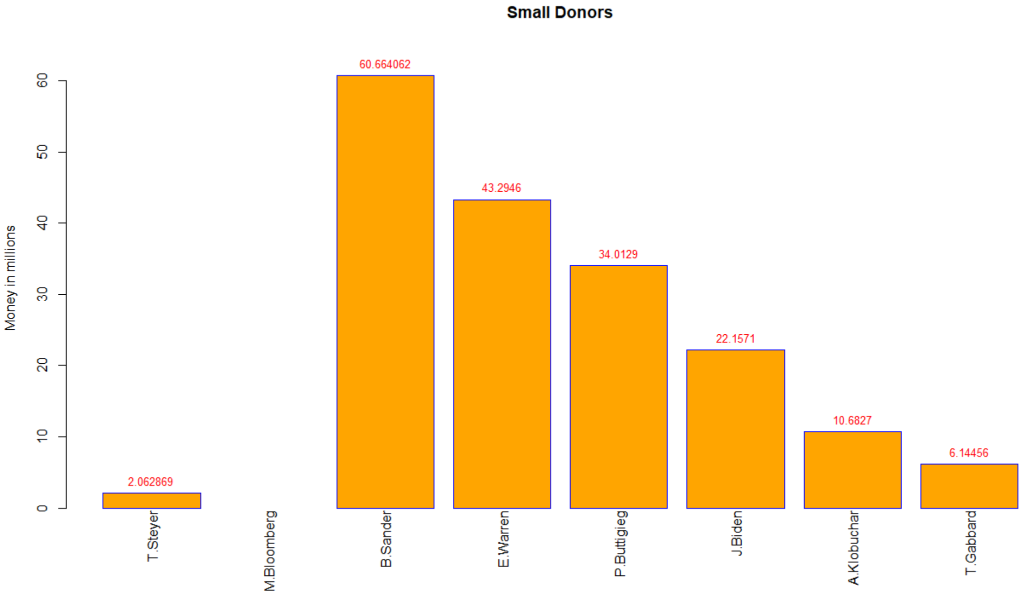

Figure 14: Small Donors Funding

Figure 15 shows the funding from big donors for presidential elections. Buttigieg has maximum funds from big donors and Bloomberg again with the least funding for the elections.

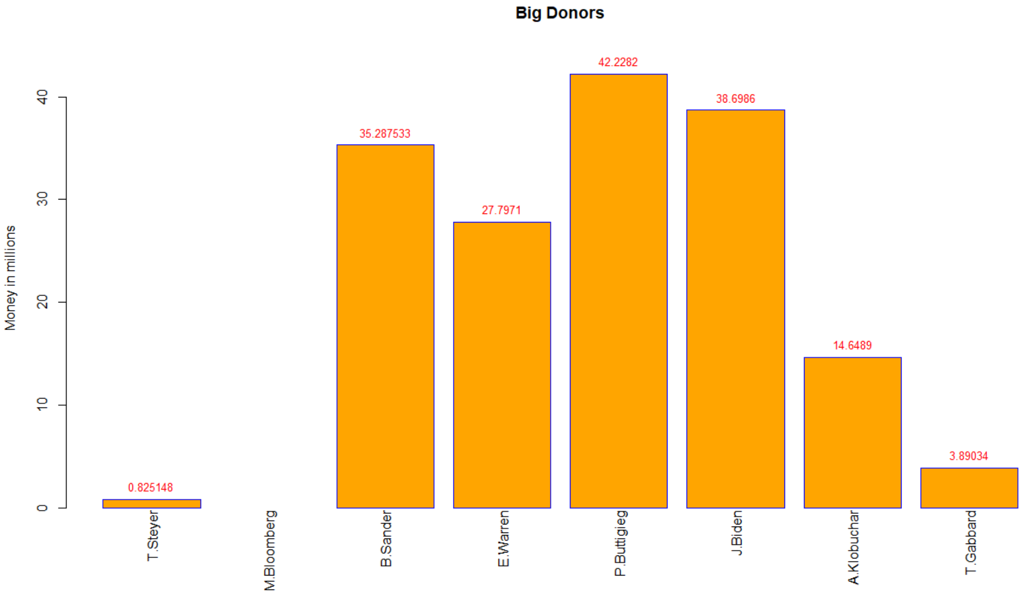

Figure 15: Big Donors Funding

Figure 16 shows self funding for presidential elections. Bloomberg and Steyers has maximum self funding. Rest of the candidates have the least or zero self funding.

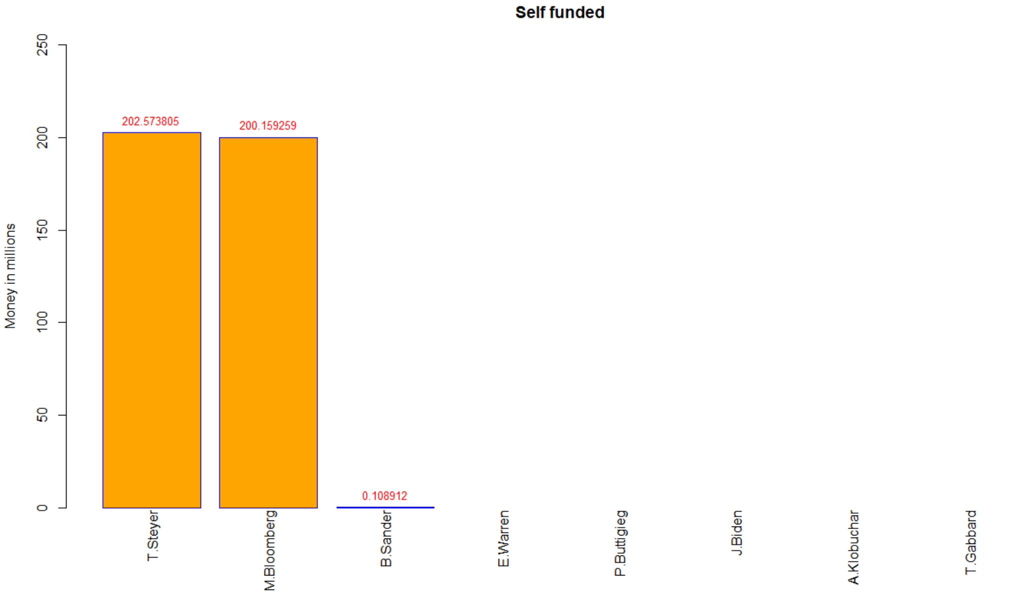

Figure 16: Self Funding

Figure 17 shows transfer funding for presidential elections. Sanders has maximum transfer funding followed by Warren. Steyers, Bloomberg and Buttigieg have the least transfer funding.

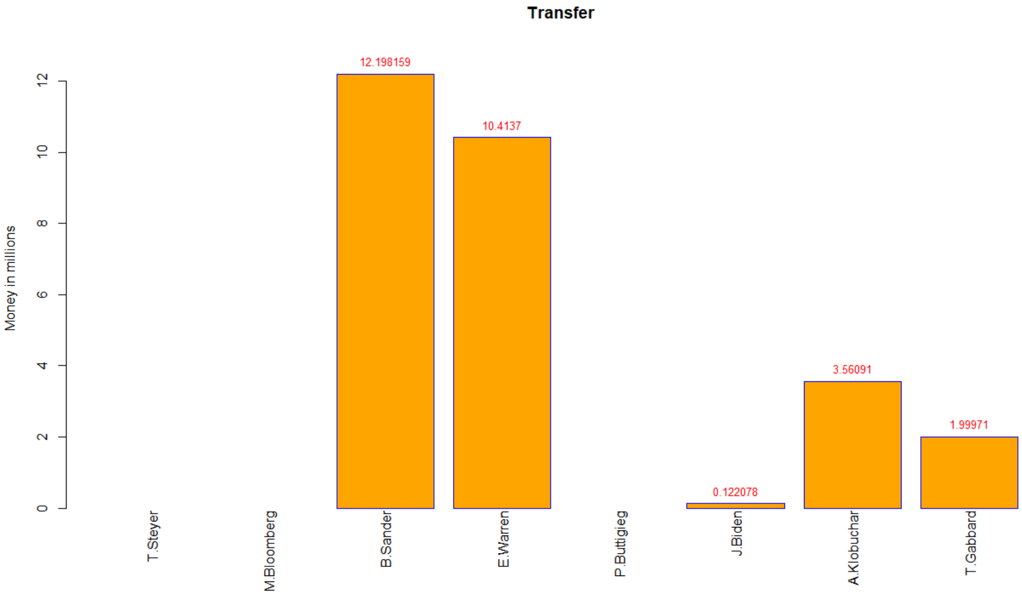

Figure 17: Transfer Funding

Figure 18 shows transfer funding for presidential elections. Steyers has maximum funding from other sources followed by Buttigieg. Klobuchar has the least funding from other sources.

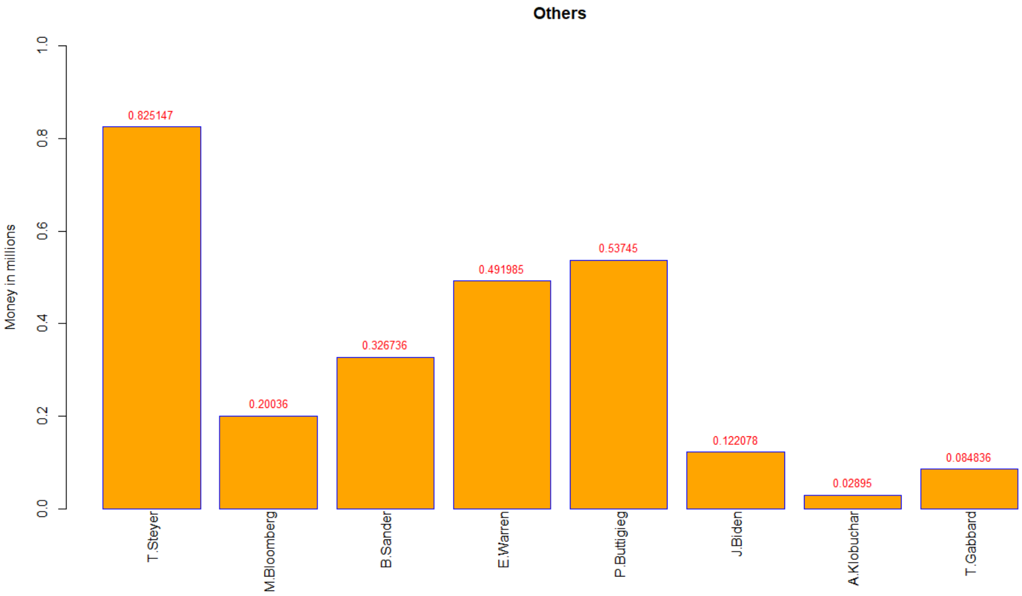

Figure 18: Others Sources Funding

4 Analysis

4.1 Delegates

Figure 2 shows the total number of delegates gathered by each of the major candidates, grouped by state. It is clear based on this graph alone that Biden and Sanders seem to be the two main candidates that have a chance at winning the Democratic Presidential Nomination. Biden appears to be more successful in south eastern states such as North Carolina, South Carolina, and Alabama, while Sanders tends to be more successful in west and mid-western states such as California, Nevada, and Colorado.

4.2 Polling

Looking at figure 3, you will see the average candidate polling support over time. The data is from all A-rated polls collected by FiveThirtyEight. According to this graph, Senator Sanders leads the way. His average support is the highest. Joe Biden's support varies more and is ranked second most of time. Bloomberg, Buttigieg, and Warren are all are fighting for third place, but none of the three clearly come out on top. Pete Buttigieg is one of the youngest candidates, and his achievements in his hometown really changed that city. Looking at poll results however, we can see that his influence and movement is not greatly affecting people outside his city. Starting from middle of February, his support steadily decreased. Comparing with recent news, the graph is correct. Buttigieg, Warren, and Bloomberg have all dropped out. Biden and Sanders will compete for the last position.

Figure 4 contains the same A-rated polling data as figure 3, but is focused around the various primary and caucus events leading up to super tuesday. For most of the time series displayed, Senator Sanders is ranked first in support. Although he lost some support following the 9th Presidential Debate, the polls bounced back leading up to the South Carolina Primary. Joe Biden's support rate dropped starting from Feb. 25 following the widely televised "dog-faced pony tail soldiers" insult that he spoke to a college girl. However support his rate rebounded at Feb. 29 following the South Carolina Primary. During this time frame Bloomberg trades blows with Biden before settling in around third place with Warren. Finally, Buttigieg's support appears to steadily decrease before sharply declining, leading up to him dropping out of the race on Feb. 29. Note that at the time of writing for this report, there was no data available for A-rated polls on super tuesday.

Due to the lack of polling data from super tuesday, we also plotted the averages of all polls, regardless of their rating, which you can see in figure 5. This gives us a little more data, showing poll results up to super tuesday. In this graph we can see that Biden overtakes Sanders in average support right before super tuesday. Again Bloomberg and Warren are tied in third place, with Buttigieg coming last.

Figures 6, 7, and 8 display the polling data as a scatter plot with regression lines. The shaded grey areas represent the 95% confidence interval's for each candidate. These plots only contain the the primary candidates left in the race at the time of analysis. As you can see from the scatter plots, there is a large variation in polling results, even among the A-rated polls. This large variation makes accurate predictions very difficult. In addition, there is a difference in trends between figures 6 and 7. In figure 6, only the A-rated polls are plotted. Here it appears that both Sanders and Biden's level of support is declining after the South Carolina Primary. In contrast, figure 7 contains polling data from all polls and shows the opposite trend. Now both Sanders and Biden show an increasing level of support leading up to and after the South Carolina Primary, with Biden's support increasing faster than Sanders. Figure 8 shows the same set of polls but displays a linear regression. This graph shows the same general trend as figure 7, where both Sanders and Biden have an ever increasing level of support. The discrepancies between polls make accurate predictions based on simple regression lines near impossible.

4.3 Spending

Next we will look at the ad spending from each candidate and see if it impacted their level of support. Figure 3 shows that prior to Iowa caucus, Sanders has about 25% percent support and he spent around 6.3 million on ads there (figure 10). His support dropped around 5% before the caucus. After the caucus, his support steadily increased, showing that this event did help him regain traction. Pete also spent about 6.8 million, and his support did raise by about 10% after Iowa caucus. Biden spent around 3.2 million but his support drops around 7% after the Iowa caucus. Warren spent about 4 million and her supports dropped around 4%. Based on this, we would conclude that the candidates who spent more money did get a little more gain from the ads than those who spent less. Bloomberg is an outlier because he spent just under 0.7 million but his support after the Iowa caucus only dropped by 1%.

In New Hampshire (figure 11), Joe Biden spent just 0.1 million in ads, however his support was increased close to Bernie Sander's who spend around 1 million in the same state to stay at 30% support. Bloomberg spent about 0.15 million in ads, and his support raised up to 25%. Warrens support however drops about 8% even though she spent about 0.73 million in ads, which is significantly more than Bloomberg and Biden.

Finally, in Nevada (figure 12), Biden spent 0.7 million in ads but his support drops by 5%. Bloomberg only spent 0.027 million dollars here, and his support also drops. Sanders spent almost 1 million dollars in Nevada only for his support to drop by 5% before rebounding. Warren spent about 0.84 million dollars on ads in Nevada and her support increases slightly afterwards. From this data we can conclude that ad spending tends to increase support for candidates who spend lots of money on ads, but it is not guaranteed.

4.4 Funding

In this last section we will look at funding for each candidate and see if it impacted their level of support. Figure 13 shows the total funding gathered by each candidate in millions of dollars. Both Steyer and Bloomberg have gathered more than 200 million dollars of funding for their campaigns, greatly overshadowing all other candidates. Since Steyer and Bloomberg have both since dropped out of the race due to very poor performance in the primaries, we can infer that that the large source of funding did not greatly improve their general level of support. Figures 14 through 18 break down this total funding into different sources. Steyer and Bloomberg are both multi-billionaires and thus the majority of their funding comes from self funding. Sanders comes in third place at 108 million dollars of total funding, the majority of which came from small donors. Both Warren and Buttigieg gathered more funding for their campaigns than Biden, however they have both since dropped out of the race. Biden and Sanders are the two remaining candidates in the race. Out of the two, Biden has collected significantly less funds than Sanders and yet he is performing very competitively. There does not seem to be a direct connection between the level of funding and the level of support for a candidate.

5 Conclusions

In this project, we focused on applying statistical methods to try and foresee the results of 2019-2020 Democratic Primary election cycle. First, we collected data from FiveThirtyEight and then used the R language to process that data. We generated the plots and established models from different aspects to try to reveal the real situation in the Democratic Primary's. We not only concentrated on the number of delegates and supporting rate of candidates, but also investigated the ad spending and funding of candidates.

From the models we built, it's obvious that Bernie Sanders and Joe Biden are ranked first and second most often. Bloomberg, Warren and Buttigieg somehow have advantages at one specific point, but these advantages did not last long. Combining all the graphs and plots we have from our model, it is clear that Bernie Sanders and Joe Biden will stand out.

According to recent news, Bernie Sanders and Joe Biden have huge advantages over other candidates. Warren, Bloomberg and Buttigieg have since dropped out. They quit for various reasons, but the main reason is that they had virtually no chance in beating Biden or Sanders to nomination.

Our model represents today's situation, however in the political landscape things can change very quickly. From our model, it looks like the future election competition between Biden and Sanders will be very fierce. From the polls that we analysed, both Sanders and Biden have the same tendency for increasing support over the last couple months. Although Sanders has been polling higher than Biden, this doesn't mean he will be the one who can compete with President Trump. We have yet to get solid results back from super tuesday, which may have a large impact on the polls.

We found that ad spending may have an impact on how a candidate polls, but differences of funding between the top eight candidates did not seem to correlate with their performance. We attempted to do some basic predictions using regression lines, but the large variance in polls coupled with unpredictable events that can create large swings in support make predictions unreliable.

References

[1] Presidential election process.

[2] P. B. Jr., C. Malone, G. Skelley, M. Koerth, N. Paine, J. Dubin, T. Sawchik, and N. Rakich. Fivethirtyeight.