CS567 Presidential Primaries Project

Andrew Dunn, Justin Phan Phan, Mark Sichong, Sridevi Wagle

Project Description

- Monitor polling data for the Democratic Party primary candidates
- Write an R script with the following functions:
  - Read polling data from source into a dataframe
  - Filter and plot polling data over time
  - Run analysis and prediction tools

Project Goals

- Visualize trends in polling data
- Show how candidate support changes after caucus results
- Make some predictions on who will win the Democratic Nomination

Datasets

- We are using a dataset from fivethirtyeight.com[1]
- FiveThirtyEight is owned by ABC News and was created by Nate Silver
- They are considered a reputable source for polling information
- Their website tracks a large number of different polls
- They do not conduct polls themselves

FiveThirtyEight

Dataset Format

AI7

	F	R	S	AC	AF	AG
1	pollster	start_date	end_date	party	candidate_name	pct
2	Emerson College	2/7/2020	2/8/2020	DEM	Deval Patrick	1.1
3	Emerson College	2/7/2020	2/8/2020	DEM	Joseph R. Biden Jr.	10.8
4	Emerson College	2/7/2020	2/8/2020	DEM	Bernard Sanders	30.4
5	Emerson College	2/7/2020	2/8/2020	DEM	Elizabeth Warren	11.6
6	Emerson College	2/7/2020	2/8/2020	DEM	Pete Buttigieg	20.2
7	Emerson College	2/7/2020	2/8/2020	DEM	Andrew Yang	4.2
8	Emerson College	2/7/2020	2/8/2020	DEM	Amy Klobuchar	13.1
9	Emerson College	2/7/2020	2/8/2020	DEM	Tulsi Gabbard	3.3
10	Emerson College	2/7/2020	2/8/2020	DEM	Tom Steyer	2.2
11	Emerson College	2/7/2020	2/8/2020	DEM	Michael F. Bennet	0.2
12	Suffolk University	2/7/2020	2/8/2020	DEM	Michael F. Bennet	0.2
13	Suffolk University	2/7/2020	2/8/2020	DEM	Joseph R. Biden Jr.	10.4

- Provided in a .csv file format on: https://data.fivethirtyeight.com/
- Updated regularly
- Contains many columns, including:
  - Pollster name, state, sponsor, sample size, and source
  - Poll start and end dates
  - Poll results as a percentage of votes for each candidate
- As of February 11th dataset contains over 16,000 rows
- Poll dates range from the end of 2018 to the present

Example Goal Plot

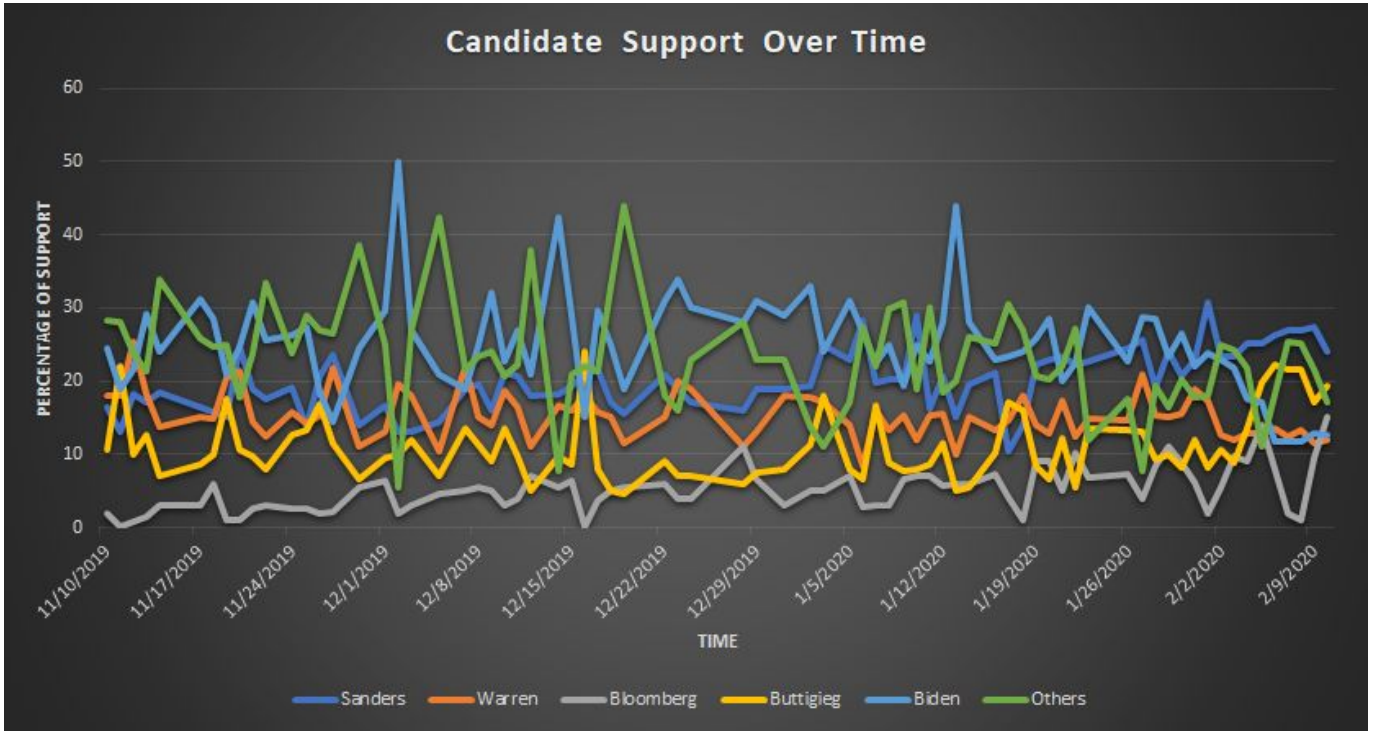

Figure 1 - Candidate support Nov. 10 through Feb. 9

Example of TV Ads Spending[1]

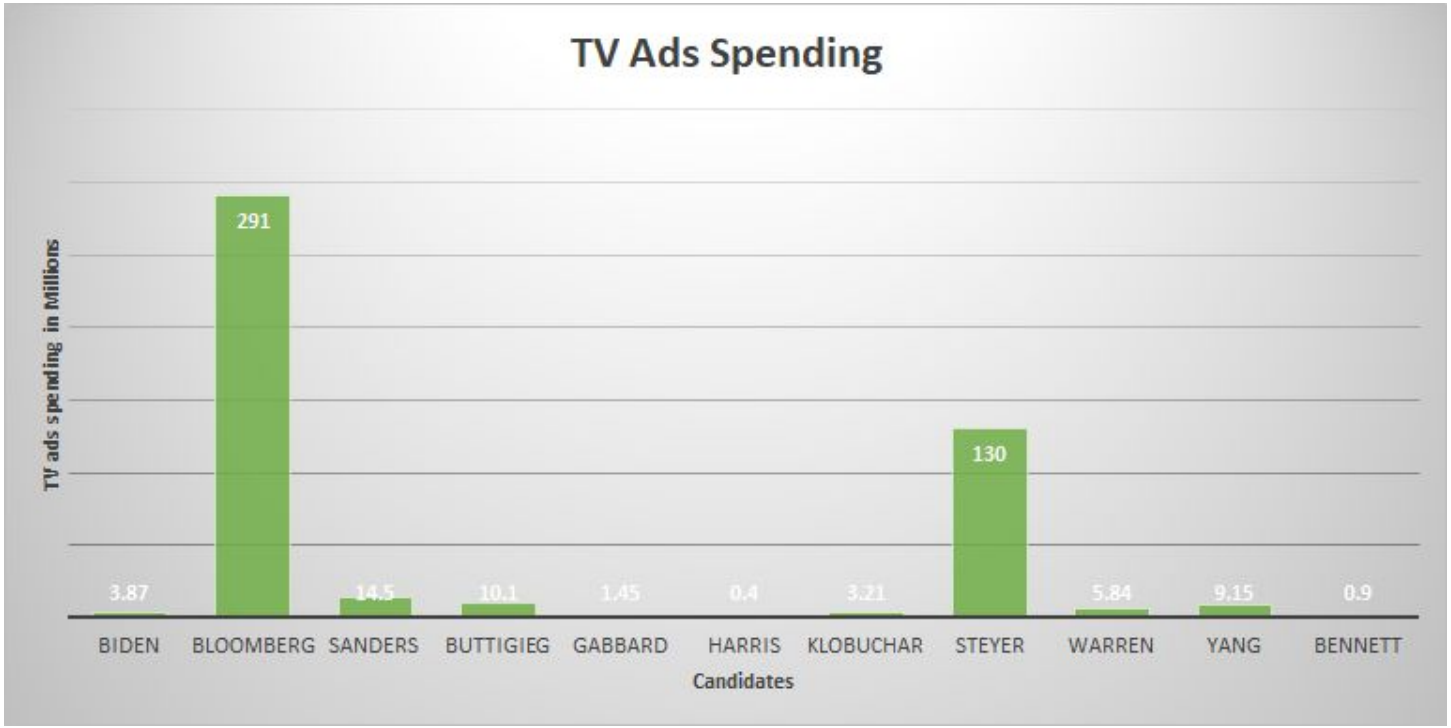

Figure 2 - TV Ad spending in millions of dollars by candidate

Example Prototype Plot

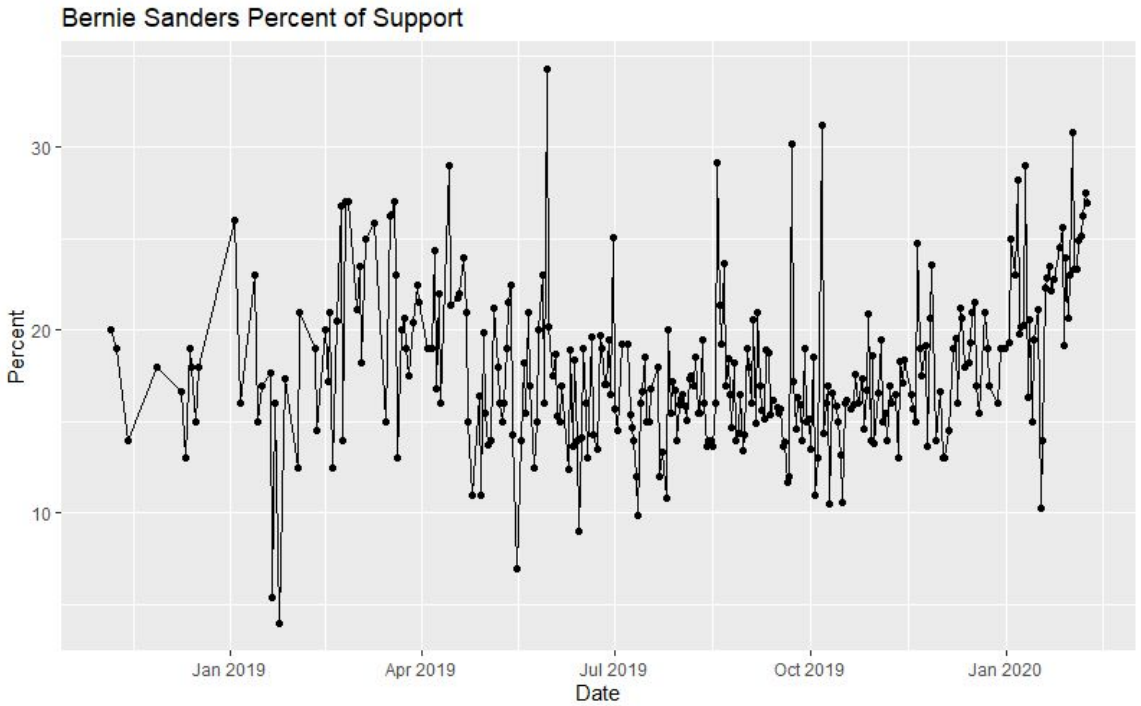

Figure 3 - Bernie Sanders average polls over time

R-Script Prototype

```r
library(ggplot2)

setwd("C:\\Users\\Coder\\OneDrive\\Documents\\CS 567 Computational Stats\\CS567_Pres_Primary")

pdata<-read.csv("data/president_primary_polls.csv", header = TRUE)
View(pdata)

sanders<-pdata[pdata$candidate_id == 13257,
               c("question_id", "end_date", "candidate_id", "candidate_name", "pct")]

sanders$end_date <- as.Date(sanders$end_date , format = "%m/%d/%y")
sanders<-sanders[order(sanders$end_date ),]
View(sanders)

sanders_pct<-setNames(aggregate(sanders[, 5], list(sanders$end_date), mean), c("end_date", "pct"))
View(sanders_pct)

# df[order(df$State,df$Mortality.Rate,df$Hospital.Name),]
sanders_pct_sorted<-sanders_pct[order(sanders_pct$end_date),]
View(sanders_pct_sorted)

ggplot(data=sanders_pct, aes(x=end_date, y=pct, group=1)) +
  geom_line()+
  geom_point()+
  ggtitle("Bernie Sanders Percent of Support")+
  labs(y="Percent", x="Date")
```

Figure 4 - R-Script prototype

Conclusion

- Dataset used is from fivethirtyeight.com
- The dataset is formatted as a CSV file
- Prototype currently only plots a single candidate
- Future version will plot all candidates

References

[1] Rakich, Nathaniel, et al. “FiveThirtyEight.” *FiveThirtyEight*, fivethirtyeight.com/