MAT5314 Project 1: Data Visualization

Teng Li(7373086)
Shiya Gao(300381032)
Chuhan Yue(300376046)
Yang Lyu(8701121)

Introduction

A data set of the 2016 US election polls was given. In this project we aim to understand the data structure by creating various visualizations.

The data set was published by FiveThirtyEight to illustrate the reliability and quality of each pollster to which a letter grade ranging from A+ to D- was given.

Method

We use various R packages to present the data set and to plot the graphs.

Result

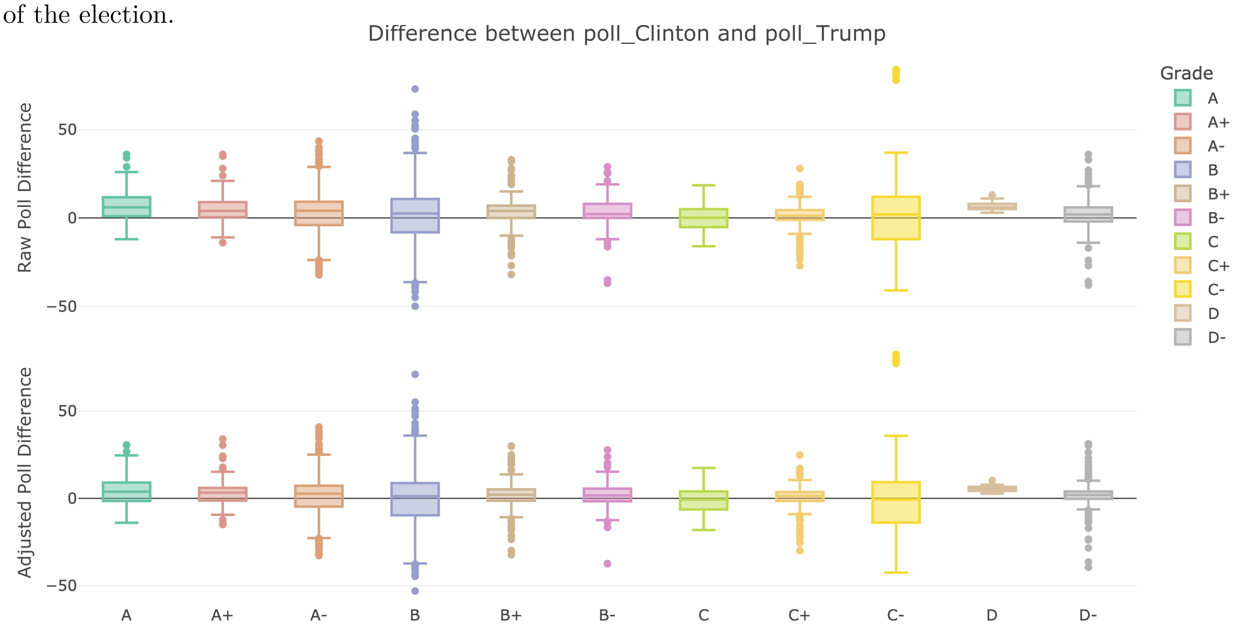
We first created a data variable definition table to give an initial understanding of the data. As one can see, there were a few variables with missing values:

Table 1: Data Variable Definition

Variables	Size	Type	Example	Number.Unique	Number.Missing	Comment
state	4208	character	U.S., New Mexico, Virginia	57	0	The name of the state (or national) where the election is held
startdate	4208	character	2016/11/3, 2016/11/1, 2016/11/2	352	0	Start date of poll
enddate	4208	character	2016/11/6, 2016/11/7, 2016/11/5	345	0	End date of poll
pollster	4208	character	ABC News/Washington Post, Google Consumer Surveys, Ipsos	196	0	Organization name that conducts or analyzes opinion polls
grade	4208	character	A+, B, A-	11	429	Grade assigned by Fivethirtyeight to pollster
samplesize	4208	integer	2220, 26574, 2195	1767	1	Sample size of polls for each pollster
population	4208	character	lv, rv, a	4	0	Type of population being polled
rawpoll_clinton	4208	numeric	47, 38.03, 42	1312	0	Poll Percentage for Hillary Clinton
rawpoll_trump	4208	numeric	43, 35.69, 39	1385	0	Poll Percentage for Donald Trump
rawpoll_johnson	4208	numeric	4, 5.46, 6	585	1409	Poll Percentage for Gary Johnson
rawpoll_mcmullin	4208	numeric	NA, 24, 27.6	17	4178	Poll Percentage for Evan Mcmullin
adjpoll_clinton	4208	numeric	45.20163, 43.34557, 42.02638	4200	0	Adjusted percentage for Hillary Clinton
adjpoll_trump	4208	numeric	41.7243, 41.21439, 38.8162	4204	0	Adjusted percentage for Donald Trump
adjpoll_johnson	4208	numeric	4.626221, 5.175792, 6.844734	2211	1409	Adjusted percentage for Gary Johnson
adjpoll_mcmullin	4208	numeric	NA, 24, 27.70142	31	4178	Adjusted percentage for Evan Mcmullin

Note that the poll results for Johnson and McMullin had lots of missing values. In particular, Johnson had 33.48% raw poll result and 33.48% adjusted poll result missing, and McMullin had 99.29% and 99.29% missing. Due to the fact that these two candidate didn't make to the final election, we chose to ignore their data in some of the analysis.

Since the final two candidates in Election 2016 are Clinton and Trump, we plotted box plots of the difference of their poll results, one for the raw poll and one for the adjusted poll. We saw that there's little difference between the distribution of the raw and the adjusted data. However, the mean of each grade of the adjusted poll result was a little closer to zero than that of the raw poll result. This indicated that the adjustment that FiveThirtyEight made was an improvement because the raw poll difference of each grade was mostly above zero, which clearly showed that the poll result was more in favour of Clinton yet Trump was the final winner of the election.



We notice that there are 57 pollsters (almost 30% of the total number of pollsters) whose grades are missing in this data set. And we cannot just delete them, because it will cause a lot of missing data in other columns. We suppose that there are two possible reasons for these missing data: one is that there are some errors of data in the original file; the other is that fivethirtyeight has not rated these pollsters yet. So we searched online and found a more detailed and authoritative file about the pollsters' grade from the fivethirtyeight website. Here is the link: https://projects.fivethirtyeight.com/pollster-ratings/.

According to the fivethirtyeight website, we found that 26 pollsters with no grade in the origin data set actually have the grades like "A/B", "B/C", "C/D", "B", "B-"; otherwise, the rest 31 pollsters without grades haven't been scored yet. Based on these information, we updated the "grade" column of the origin data set. We replace "NA" with the actual grades and none.

Table 2: Updating the missing data

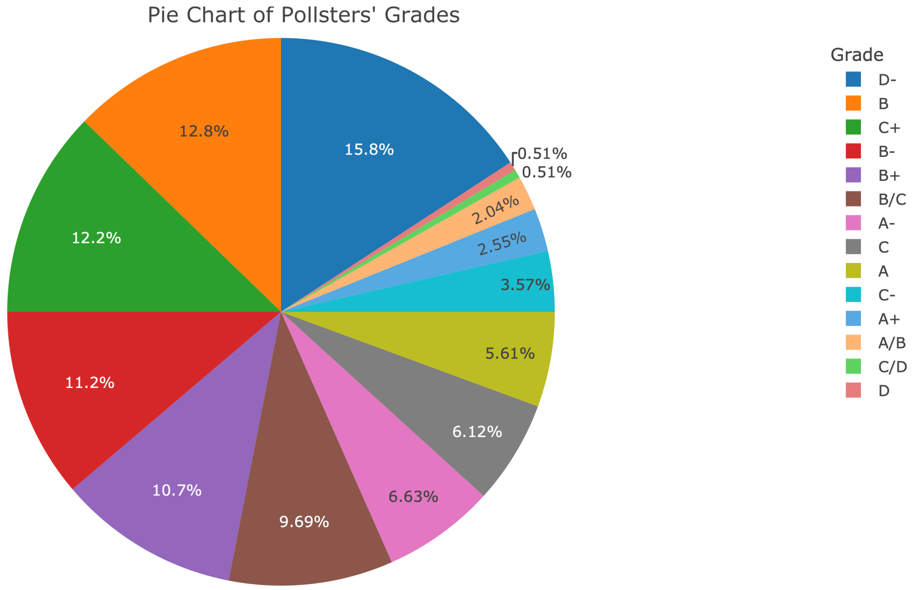
	pollster	grade
23	Remington	B
27	Morning Consult	B-
35	Saguaro Strategies	B/C
37	Insights West	B/C
44	BK Strategies	B/C
59	Data Orbital	A/B
65	Starboard Communications	B/C

	pollster	grade
69	Strategic National	B/C
84	Bendixen & Amandi International	B/C
86	Associated Industries of Florida	B/C
97	Centre College	B/C
98	Public Religion Research Institute	A/B
101	Praecones Analytica	B/C
106	Craciun Research	B/C
108	University of Colorado	B/C
112	Baldwin Wallace University	B/C
122	University of Wyoming	C/D
131	HighGround	B/C
133	Michigan State University	A/B
140	Echelon Insights	A/B
152	Meredith College	B/C
155	Mercyhurst University	B/C
167	Strategy Research	B/C
176	Hickman Analytics	B/C
184	Data Targeting	B/C
196	Ogden & Fry	B/C

Because some pollsters are repeated in different rows of the data set, so we want to verify that each pollster corresponds to only one kind of grade. The result is as follow:

```
## the column of number_of_grades only contains one type of value: 1
```

From the “Pie Chart of Pollsters’ Grades” below, we could see that the unrated pollsters make up the largest percentage, about 15.8% of the total; pollsters with grades “B” and “C+” account for the second and third most, 12.8% and 12.2% respectively; “D” grade has the smallest percentage of pollsters, only about 0.51%. Besides, B-level grades, including “B+”, “B” and “B-”, are around 34.7%, almost one-third of the total; C-level and A-level grades are about 21.89% and 14.79% respectively. For those without explicit grades, whose grades are “A/B”, “B/C” and “C/D”, they account only for 12.24%, “B” grade pollsters make up the majority of this part especially, almost 9.69%.



Discussion

Conclusion