# 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.
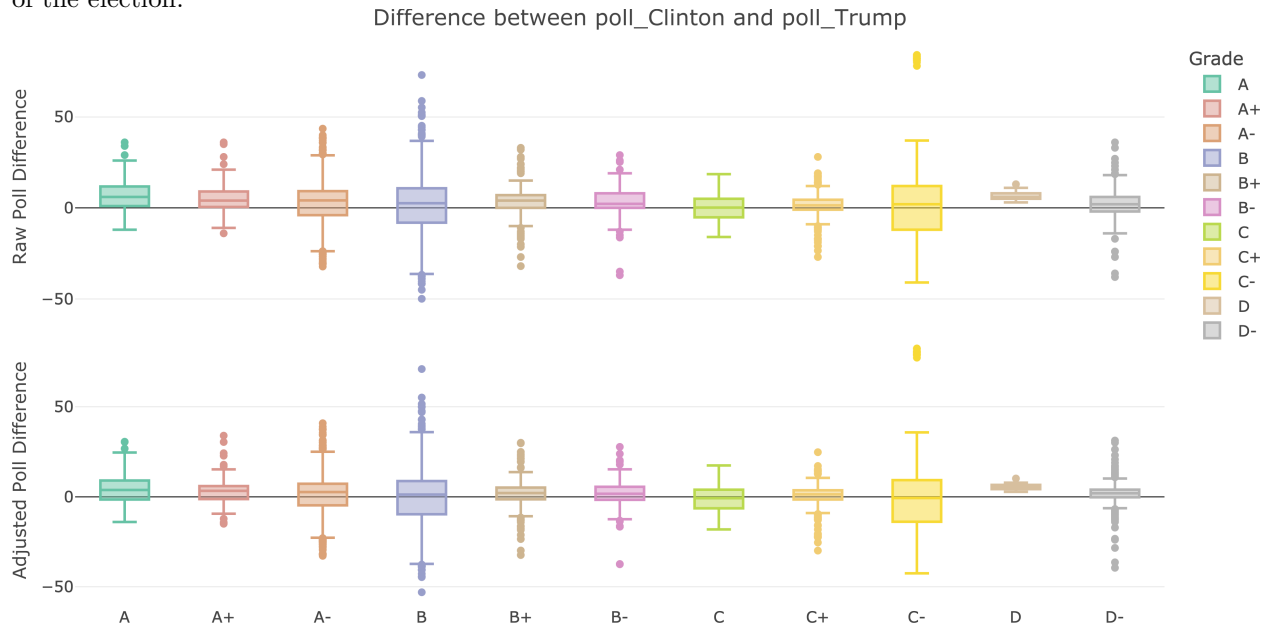
## Result

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 | 2016-11-03, 2016-11-01, 2016-11-02 | 352 | 0 | Start date of poll |
| enddate | 4208 | character | 2016-11-06, 2016-11-07, 2016-11-05 | 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.



Difference between poll_Clinton and poll_Trump

# Discussion

# Conclusion