Plot

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.

Method

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

```
knitr::opts_chunk$set(echo = FALSE)
library(plotly)
```

```
## Loading required package: ggplot2

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## v readr     2.1.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks plotly::filter(), stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
```

Result

We first take a look at the raw data set:

As we can see, there are a few variables with missing values:

```
##     state            startdate           enddate            pollster
##  Length:4208        Length:4208        Length:4208        Length:4208
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     grade             samplesize       population        rawpoll_clinton
##  Length:4208        Min.   :   35.0   Length:4208        Min.   :11.04
##  Class :character   1st Qu.:  447.5   Class :character   1st Qu.:38.00
##  Mode  :character   Median :  772.0   Mode  :character   Median :43.00
##                     Mean   : 1148.2                      Mean   :41.99
##                     3rd Qu.: 1236.5                      3rd Qu.:46.20
##                     Max.   :84292.0                      Max.   :88.00
##                     NA's   :1
##  rawpoll_trump   rawpoll_johnson  rawpoll_mcmullin adjpoll_clinton
##  Min.   : 4.00   Min.   : 0.000   Min.   : 9.0     Min.   :17.06
##  1st Qu.:35.00   1st Qu.: 5.400   1st Qu.:22.5     1st Qu.:40.21
##  Median :40.00   Median : 7.000   Median :25.0     Median :44.15
##  Mean   :39.83   Mean   : 7.382   Mean   :24.0     Mean   :43.32
##  3rd Qu.:45.00   3rd Qu.: 9.000   3rd Qu.:27.9     3rd Qu.:46.92
##  Max.   :68.00   Max.   :25.000   Max.   :31.0     Max.   :86.77
##                  NA's   :1409     NA's   :4178
##  adjpoll_trump    adjpoll_johnson  adjpoll_mcmullin
##  Min.   : 4.373   Min.   :-3.668   Min.   :11.03
##  1st Qu.:38.429   1st Qu.: 3.145   1st Qu.:23.11  
##  Median :42.765   Median : 4.384   Median :25.14  
##  Mean   :42.674   Mean   : 4.660   Mean   :24.51  
##  3rd Qu.:46.290   3rd Qu.: 5.756   3rd Qu.:27.98  
##  Max.   :72.433   Max.   :20.367   Max.   :31.57  
##                   NA’s   :1409     NA’s   :4178   
```

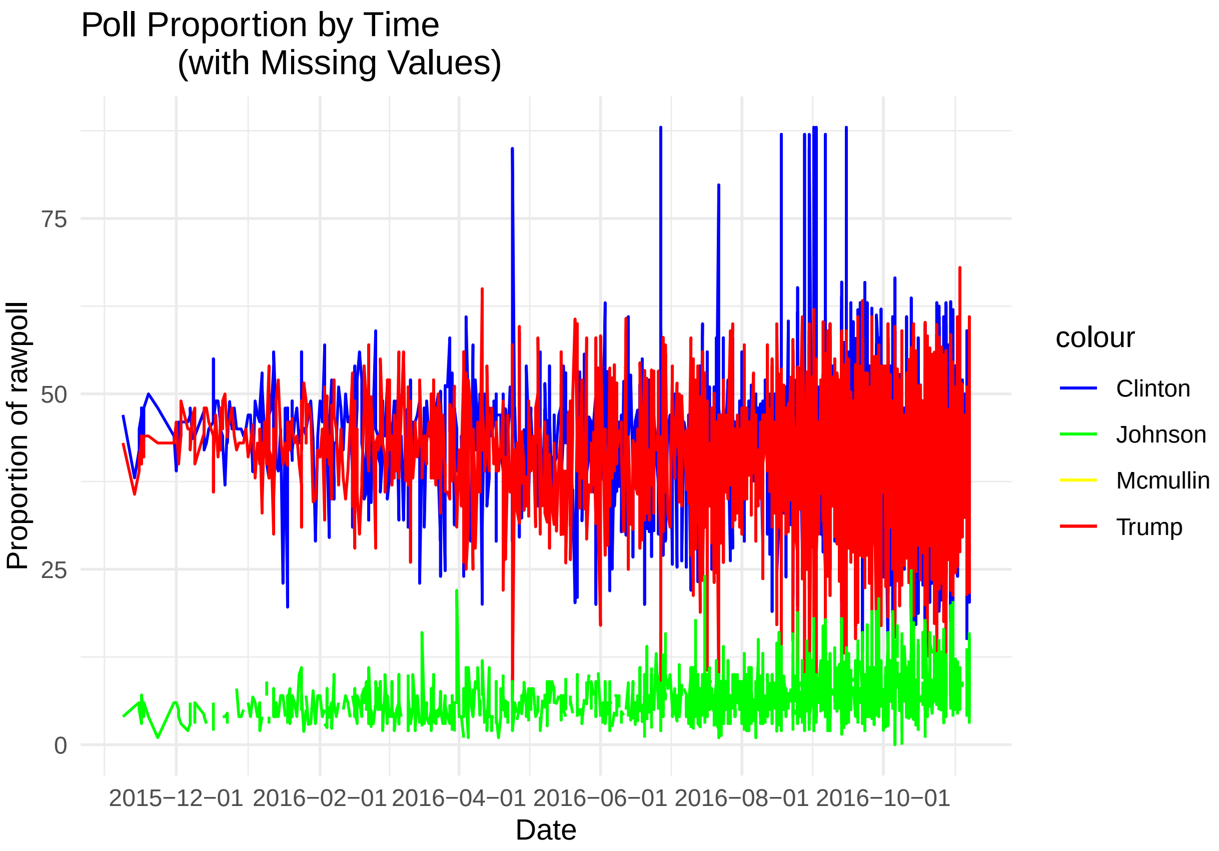
```
##   start_time   end_time clinton trump johnson mcmullin
## 1 2016-11-03 2016-11-06   47.00 43.00    4.00       NA
## 2 2016-11-01 2016-11-07   38.03 35.69    5.46       NA
## 3 2016-11-02 2016-11-06   42.00 39.00    6.00       NA
## 4 2016-11-04 2016-11-07   45.00 41.00    5.00       NA
## 5 2016-11-03 2016-11-06   47.00 43.00    3.00       NA
## 6 2016-11-03 2016-11-06   48.00 44.00    3.00       NA
```

```
##      start_time   end_time clinton trump johnson mcmullin
## 3872 2015-11-07 2015-11-08      42    47      NA       NA
## 3968 2015-11-09 2015-11-13      50    36      NA       NA
## 4104 2015-11-11 2015-11-15      37    48      NA       NA
## 4138 2015-11-12 2015-11-15      48    38      NA       NA
## 3847 2015-11-10 2015-11-16      41    40      NA       NA
## 3857 2015-11-10 2015-11-16      41    44      NA       NA
```

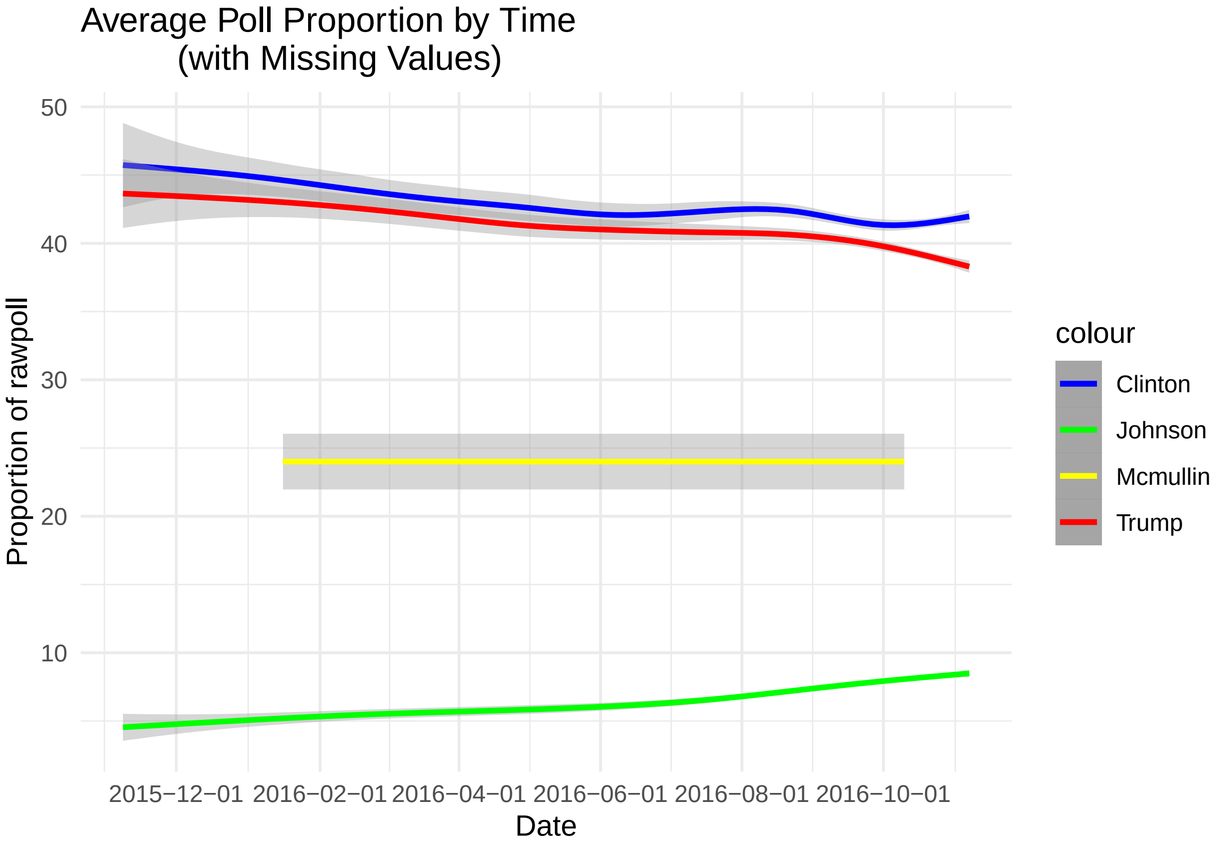
Without removing the missing values:

```
## Warning: Removed 2 rows containing missing values (‘geom_line()‘).
```

```
## Warning: Removed 1895 rows containing missing values (‘geom_line()‘).
```



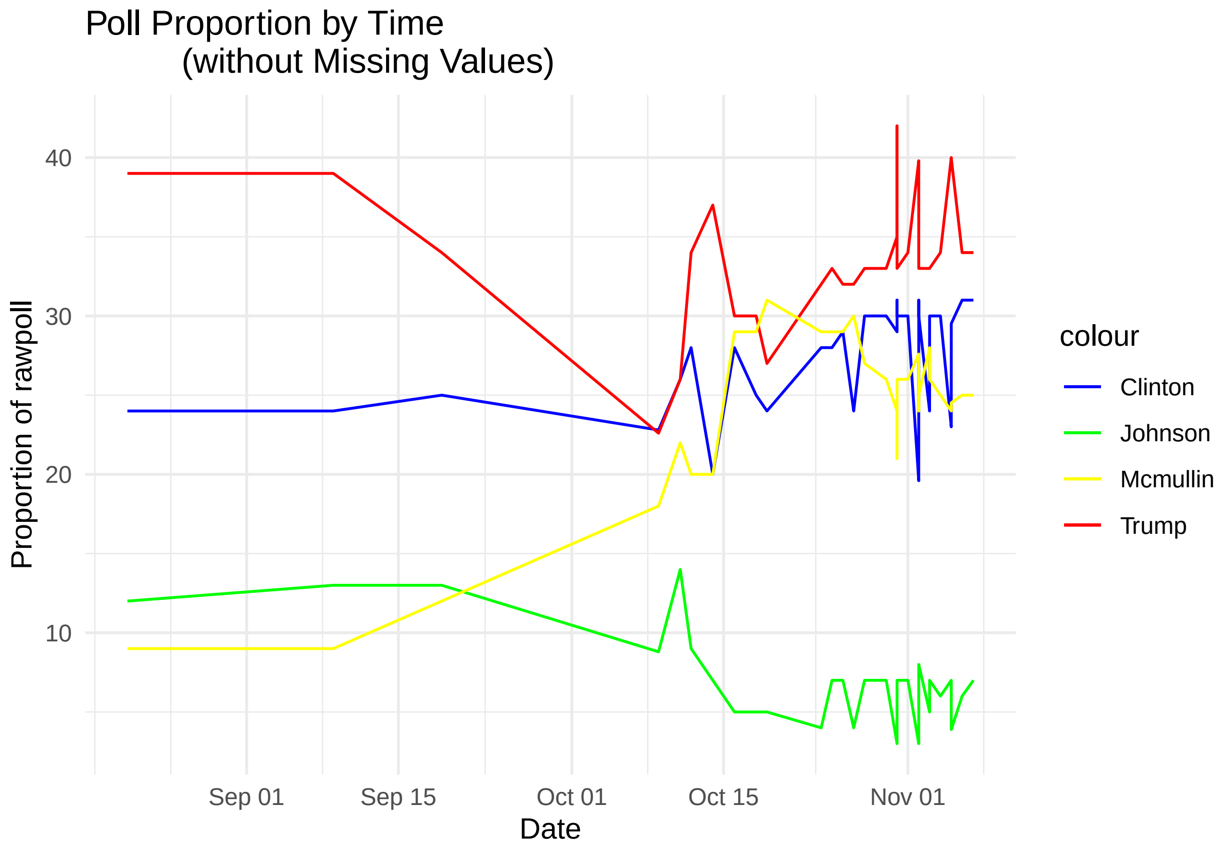
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 1409 rows containing non-finite values (`stat_smooth()`).

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 4178 rows containing non-finite values (`stat_smooth()`).
```



Remove the missing values, count the row number of 4 candidates cleaned data:

```
## [1] 30
```

Using the data with missing values deleted, a trend chart showing the change in the proportion of votes obtained by all four candidates over time is made.



3 candidate's (Clinton, Trump and Johnson) trend by time with cleaned data.

```
##   start_time   end_time clinton trump johnson
## 1 2016-11-03 2016-11-06   47.00 43.00    4.00
## 2 2016-11-01 2016-11-07   38.03 35.69    5.46
## 3 2016-11-02 2016-11-06   42.00 39.00    6.00
## 4 2016-11-04 2016-11-07   45.00 41.00    5.00
## 5 2016-11-03 2016-11-06   47.00 43.00    3.00
## 6 2016-11-03 2016-11-06   48.00 44.00    3.00
```

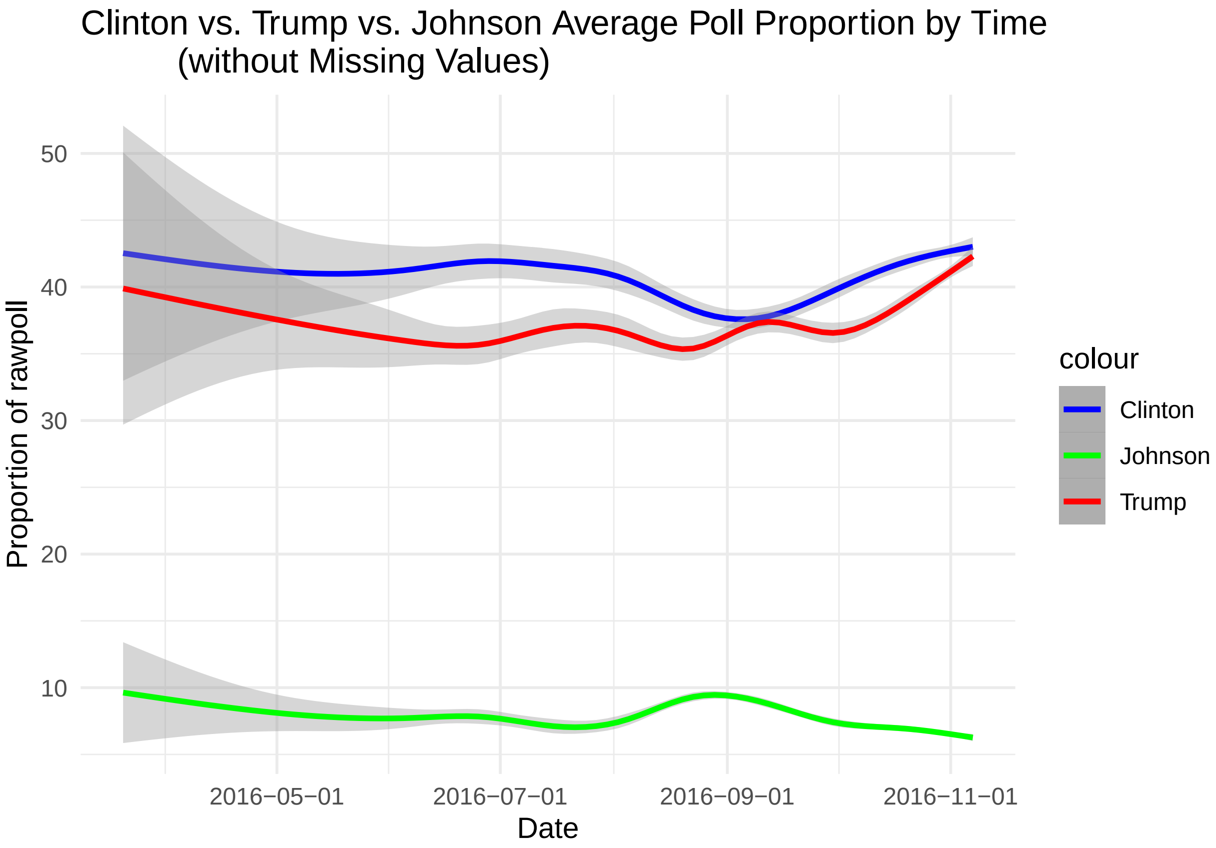
3 candidates data removing the missing values:

```
##   start_time   end_time clinton trump johnson
## 1 2016-11-03 2016-11-06   47.00 43.00    4.00
## 2 2016-11-01 2016-11-07   38.03 35.69    5.46
## 3 2016-11-02 2016-11-06   42.00 39.00    6.00
## 4 2016-11-04 2016-11-07   45.00 41.00    5.00
## 5 2016-11-03 2016-11-06   47.00 43.00    3.00
## 6 2016-11-03 2016-11-06   48.00 44.00    3.00
```

```
##      start_time   end_time clinton trump johnson
## 3496 2016-03-17 2016-03-20      42    34      11
## 3536 2016-03-28 2016-03-30      49    41      10
## 3562 2016-04-23 2016-04-25      49    40       8
## 2906 2016-05-02 2016-05-04      28    48       6
## 3390 2016-05-06 2016-05-09      42    38       4
## 2999 2016-05-13 2016-05-15      41    33      14
## [1] 2799
```

Using the data with missing values deleted, a trend chart showing the change in the proportion of votes obtained by 3 candidates (Clinton, Trump and Johnson) over time is made.

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



2 candidate's (Clinton, Trump) trend by time with cleaned data.

```
##   start_time   end_time clinton trump
## 1 2016-11-03 2016-11-06   47.00 43.00
## 2 2016-11-01 2016-11-07   38.03 35.69
## 3 2016-11-02 2016-11-06   42.00 39.00
## 4 2016-11-04 2016-11-07   45.00 41.00
## 5 2016-11-03 2016-11-06   47.00 43.00
## 6 2016-11-03 2016-11-06   48.00 44.00
```

2 candidates data removing the missing values:

```
##   start_time   end_time clinton trump
## 1 2016-11-03 2016-11-06   47.00 43.00
## 2 2016-11-01 2016-11-07   38.03 35.69
## 3 2016-11-02 2016-11-06   42.00 39.00
## 4 2016-11-04 2016-11-07   45.00 41.00
## 5 2016-11-03 2016-11-06   47.00 43.00
## 6 2016-11-03 2016-11-06   48.00 44.00
```

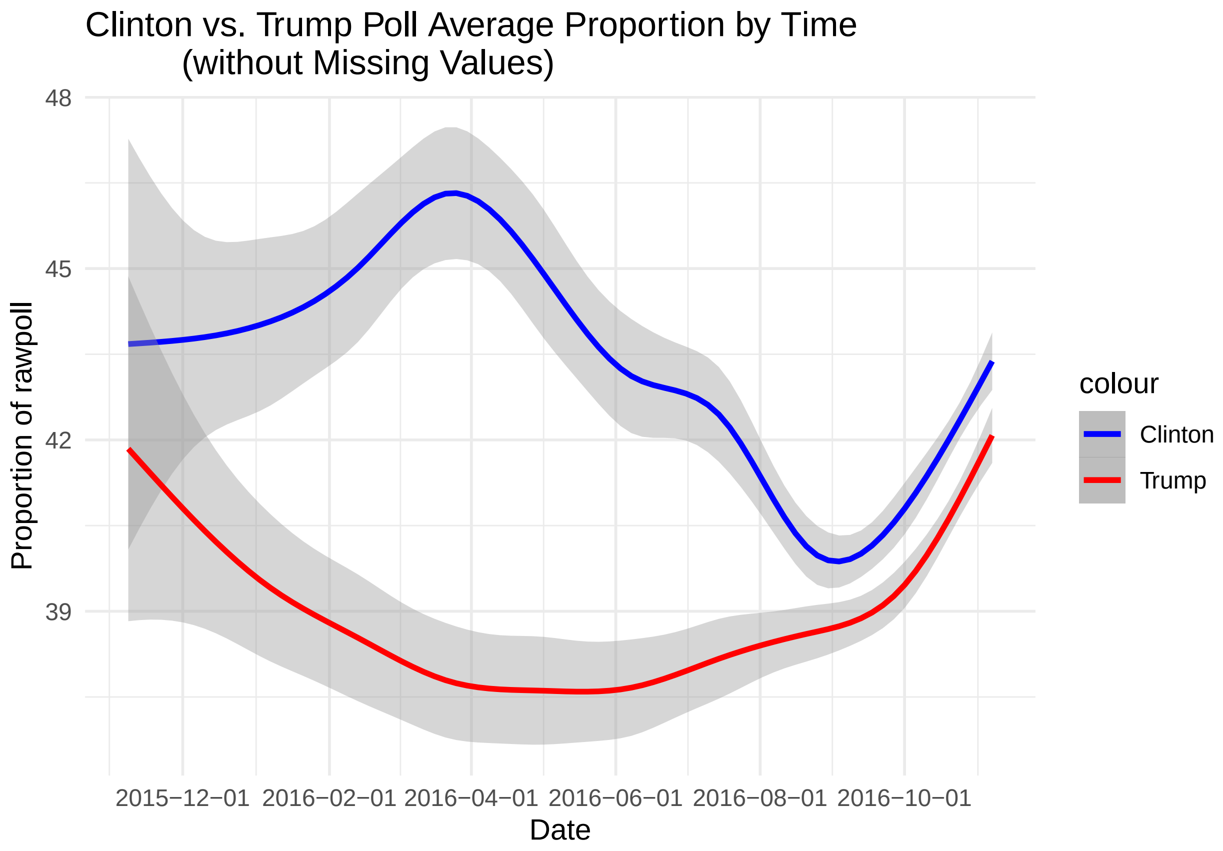
```
##      start_time   end_time clinton trump
## 3872 2015-11-07 2015-11-08      42    47
## 3968 2015-11-09 2015-11-13      50    36
## 4104 2015-11-11 2015-11-15      37    48
## 4138 2015-11-12 2015-11-15      48    38
## 3847 2015-11-10 2015-11-16      41    40
## 3857 2015-11-10 2015-11-16      41    44
```

```
##      start_time   end_time clinton trump
## 649  2016-11-01 2016-11-07   23.23 47.41
## 651  2016-11-01 2016-11-07   52.12 18.86
## 677  2016-11-01 2016-11-07   53.62 10.62
## 691  2016-11-01 2016-11-07   40.69 23.10
## 709  2016-11-01 2016-11-07   43.62 28.72
## 4206 2016-11-01 2016-11-07   21.33 35.05
```

```
## [1] 4208
```

Using the data with missing values deleted, a trend chart showing the change in the proportion of votes obtained by 2 candidates (Clinton vs. Trump) over time is made.

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```


Clinton vs. Trump Poll Average Proportion by Time
(without Missing Values)
Proportion of rawpoll
Date
colour
Clinton
Trump