R Notebook

```
#Install and load essential packages and libraries
library(dslabs)
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 packages ------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v stringr 1.4.0
## v tidyr   1.0.0     v forcats 0.4.0
## v readr   1.3.1

## -- Conflicts ---------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.6.2

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:dplyr':
##
##     intersect, setdiff, union

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

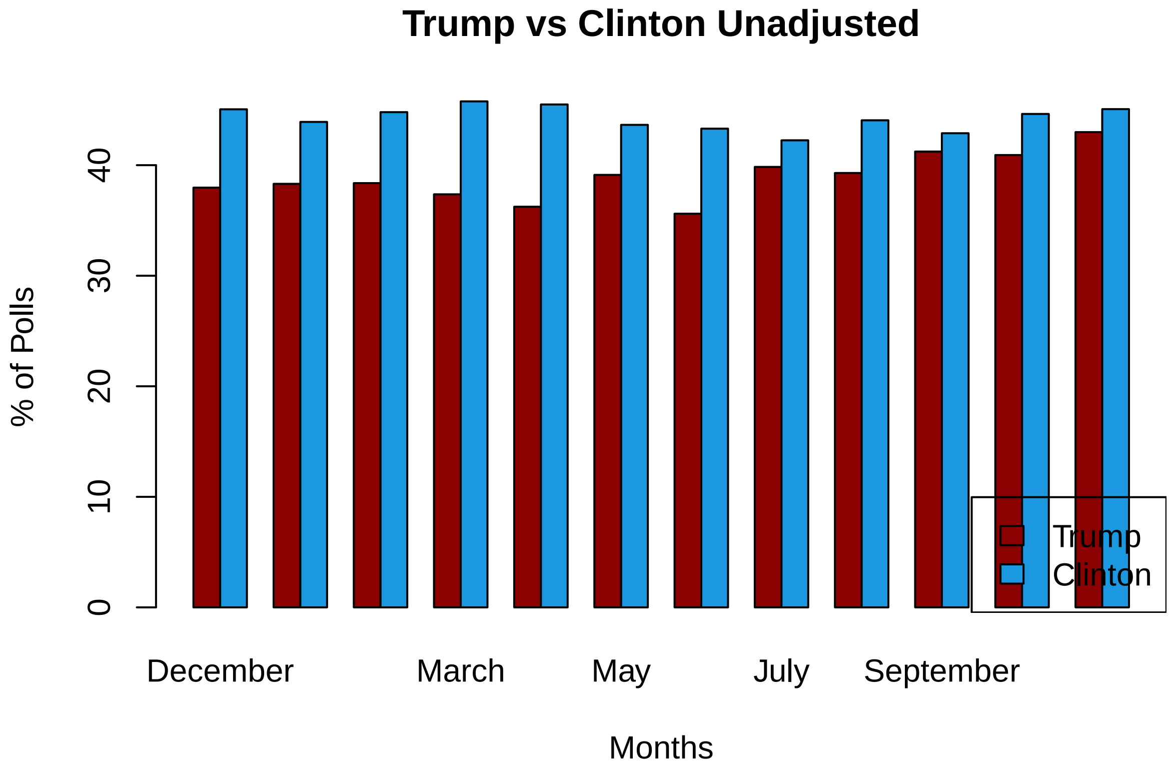
```
# Only the nationwide poll from December 2015 are selected.
poll_us_election_2016 <- polls_us_election_2016%>%
                         filter(state == "U.S." & enddate >="2015-12-01")

# Converting the enddate yy-mm-dd format to months
poll_us_election_2016$enddate <- months(as.Date(poll_us_election_2016$enddate))
#The column has been chaanged to endmonth
names(poll_us_election_2016)[which(names(poll_us_election_2016)=="enddate")]<-"poll_endmonth"
```

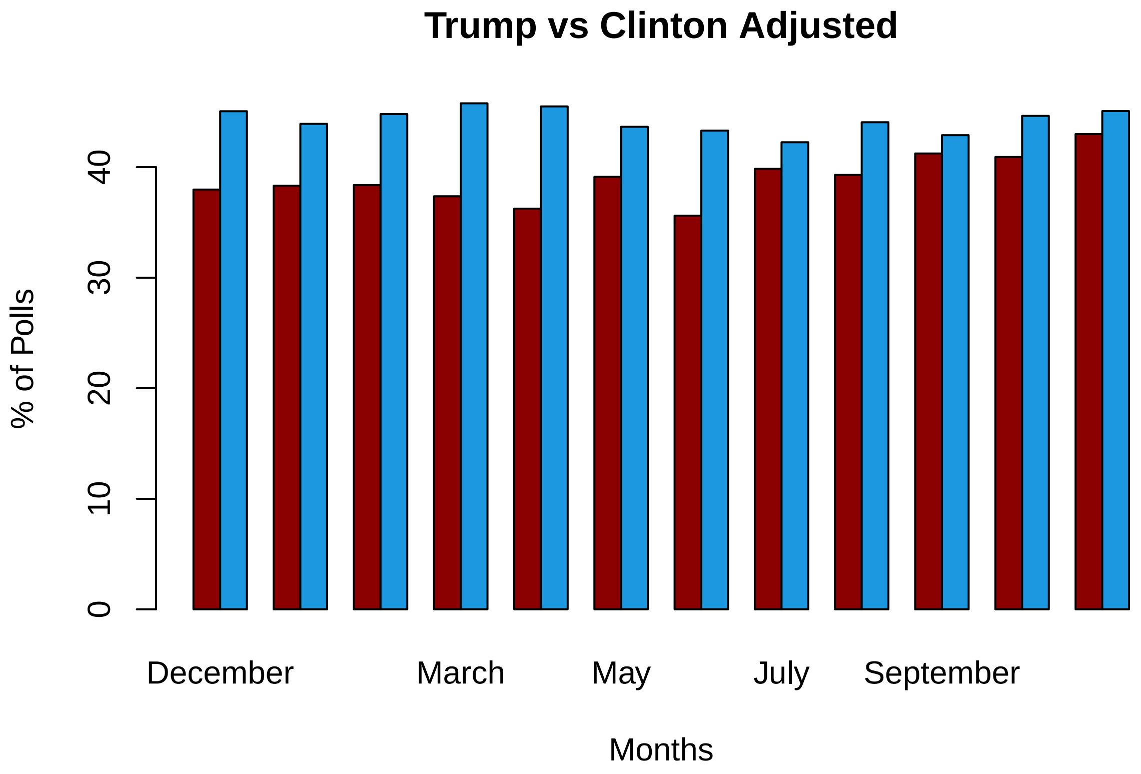
```
poll_trump_clinton <- poll_us_election_2016%>%
   #All columns not necessary in our analysis  are not slected
                  select(-rawpoll_johnson,-rawpoll_mcmullin,
                  -adjpoll_johnson,-adjpoll_mcmullin,-population,-grade)%>%
  #table is arranged by start date
                  arrange(startdate)%>%
  # grouped by endmonth to calcuate average poll% per month for eac candidate,Adjusted and unadjusted
                 group_by(poll_endmonth)%>%
                 mutate(Trump = mean(rawpoll_trump),Clinton = mean(rawpoll_clinton),
                Trump_adj = mean(adjpoll_trump), Clinton_adj = mean(adjpoll_clinton))
# All the duplicated month are removed
Final_table <-poll_trump_clinton[!duplicated(poll_trump_clinton$poll_endmonth),]

#Code for arranging month in order on X-axis
poll_endmonth <- factor(Final_table$poll_endmonth, levels = substr(month.name, 1, 3))
#Barplot code
barplot(cbind(Trump, Clinton) ~ poll_endmonth, data = Final_table,ylab="% of Polls",
        xlab = "Months",main = "Trump vs Clinton Unadjusted", beside = TRUE,
         col = c("#8B0000", "#1b98e0"))

#Code for legend
legend("bottomright",
       legend = c("Trump", "Clinton"),
       fill = c("#8B0000", "#1b98e0"))
```



```
#Barplot code for adjusted values
barplot(cbind(Trump, Clinton) ~ poll_endmonth, data = Final_table,ylab="% of Polls",
        xlab = "Months",main = "Trump vs Clinton Adjusted", beside = TRUE,
        col = c("#8B0000", "#1b98e0"))
```



I went over the data and there were fifteen different variables.I had thoughts about coloring the US map based on every state poll winner.I also had thoughts about visualizing polls based on the grade of polls.I had also thoughts about two animated line plots in one graph with time in X-axis.Finally, I decided I would use normal barplot to compare adjusted and unadjusted poll results based on every month from December to November.

I cleaned the data removing unnecessary columns and converting the end date into months.For example,any polls with end date 2015-12-01 to 2015-12-31 would be named December.Then I calculated the average poll percentage for each month for each candidate and plotted them in barplot.There were polls from different agencies/institute/websites which had some varying results.I thought it would be appropriate to take average of those result and analyze.Instead of average percentage of poll,moving average could have been better.

The question for my visualization would be comparison of poll results every month in adjusted and unadjusted data among the candidates and between the two types of data.I have not omitted any data nor modified original data.My codes are reproducible and anyone should get the result I got if they run this code.I have added enough code so that one can understand what is going on with the code.I have been cautious to add title,legend and names for x- axis and y-axis for the graph.The colors have used are red and blue,this would make easier and intuitive analysis of the graph.The trend of support of public for both the candidates can be compared and measured from graph.We can see Trump is closing the gap from July.Any people who can read are the targeted audience and I believe it is informative and easy to follow through for them.