Predictions for United States Presidential Elections 2016

Harshala Rajesh

August 16, 2016

## Introduction

United States of America being one of the Superpowers, the leader of the country will have a significant influence not just for the country but the entire world. The President of the United States is considered one of the world's most powerful people, leading the world's only contemporary superpower. The role includes being the commander-in-chief of the world's most expensive military with the largest nuclear arsenal and leading the nation with the largest economy by real and nominal GDP. The office of the president holds significant hard and soft power both in the United States and abroad.1

The President is elected by Electoral College to a four-year term. Current President Barack Obama will be ending his second four-year term and The United States Presidential Election is scheduled for Tuesday, November 8, 2016. The series of presidential primary elections and caucuses took place between February 1 and June 14, 2016.

Former Secretary of State and New York Senator Hillary Clinton is the Democratic Party's presidential nominee. Businessman and reality television personality Donald Trump is the Republican Party's presidential nominee. Hillary Clinton, if elected will be the first woman to take the Office of the President in United States which makes this year's election very interesting. Opinion Polls are conducted Nationwide and they seem to tell a consistent story as to where the race stands.

The Report is an attempt at predicting the probable President of United States of America based on the Opinion Polls conducted from January 2016 to present day.

## Description of Dataset

Huffington Post, One of the leading news aggregators in America, has been publishing the results of various Polls conducted across the nation. R's XML library will be used to get the live data from Huffington Post - <http://elections.huffingtonpost.com/pollster/2016-general-election-trump-vs-clinton>

The Dataset contains Poll information, Percent Votes for Trump, Clinton, Others, Undecided and Spread. Results of over 30 agents conducting the polls from past 20 months are listed. For the current study, results from January 2016 and onwards will be used.

## Analysis and Cleaning of Variables in the Data Set

Data will be extracted from the website when code is run and results from January 2016 and onwards will be extracted

#Get data from the website  
library(XML)  
  
#Data From Polls published on Huffington Post  
rawHuff <- readHTMLTable('http://elections.huffingtonpost.com/pollster/2016-general-election-trump-vs-clinton')  
Huff <- data.frame(rawHuff[[1]])

Huff <- Huff %>% dplyr::filter(!grepl("2015",Poll))

#### Dataset : 210, 6

Data Extracted Contains Information from Polls conducted by various agencies like Rasmussen, CNN totalling to 26 sources. There are several challenges with this data

* Data is extracted using readHTMLTable function in XML library which has advantage of capturing the latest and greatest poll results
* Data contains information from previous year, hence only Poll results from January 2016 and onwards will be filtered.
* Poll Field in the results is a combination of Poll name, Poll week, Number of people who participated in the poll and the type of voters.
* Fields need to be extracted from the above field

#### Extract poll\_name, start\_date,end\_date,type\_of\_voter and number\_of\_voters from Poll field

# Replace Poll Field to end with "Voters"  
Huff$Poll <- gsub("Adults","Adult Voters",Huff$Poll)  
  
# Extract num\_voters from Poll Column  
num\_voters <- str\_extract(Huff$Poll,word(Huff$Poll,-3))

## Warning in stri\_extract\_first\_regex(string, pattern, opts\_regex =  
## opts(pattern)): empty search patterns are not supported

num\_voters <- as.numeric(gsub(",","",num\_voters))  
  
# Extract Voter Type from Poll Column  
voter\_type <- as.vector(str\_extract(Huff$Poll,word(Huff$Poll,-2)))  
  
# Extract Poll name and clean it.  
poll\_name <- str\_extract(Huff$Poll,word(Huff$Poll,1))  
poll\_name <- gsub("\n","",poll\_name)  
  
# Extract poll week  
patt <- '(\\w+)\\s\*(\\w+)\\s\*\u2013\\s\*(\\w+)\\s\*(\\w+)'  
  
poll\_week <- str\_extract(Huff$Poll,patt)  
  
start\_date <- word(poll\_week,1,2)  
end\_date <- word(poll\_week,-2,-1)  
  
poll\_start\_date <- as.Date(start\_date,"%b %d")  
poll\_end\_date <- as.Date(end\_date,"%b %d")

#### Add the extracted Variable

Extracted Values are added as columns to the dataframe. Based on the poll week, polls need to be sorted on a monthly basis. Percent Votes for each candidate will be converted to numeric values. Finally Poll field will be removed from the dataset. Data will be re-arrange for better readability

# Add 5 new columns to Huff Dataframe  
Huff <- mutate(Huff, num\_of\_voters = num\_voters,type\_of\_voter = voter\_type,  
 poll\_name = poll\_name,poll\_week = poll\_week,  
 poll\_start\_date = poll\_start\_date,poll\_end\_date = poll\_end\_date)  
  
# Compute month variable from poll\_start\_date  
Huff$month <- ifelse(month(Huff$poll\_start\_date)==1,1,  
 ifelse(month(Huff$poll\_start\_date)==2,2,  
 ifelse(month(Huff$poll\_start\_date)==3,3,  
 ifelse(month(Huff$poll\_start\_date)==4,4,  
 ifelse(month(Huff$poll\_start\_date)==5,5,  
 ifelse(month(Huff$poll\_start\_date)==6,6, ifelse(month(Huff$poll\_start\_date)==7,7,  
 ifelse(month(Huff$poll\_start\_date)==8,8,  
 ifelse(month(Huff$poll\_start\_date)==9,9,  
 ifelse(month(Huff$poll\_start\_date)==10,10,  
 ifelse(month(Huff$poll\_start\_date)==11,11,11)  
 ))))))))))  
  
  
Huff$month <- as.factor(Huff$month)  
  
# Remove Poll Column  
Huff$Poll <- NULL  
  
# Convert percent\_votes from char to numeric  
Huff$Trump <- as.numeric(as.character(Huff$Trump))  
Huff$Clinton <- as.numeric(as.character(Huff$Clinton))  
Huff$Other <- as.numeric(as.character(Huff$Other))  
Huff$Undecided <- as.numeric(as.character(Huff$Undecided))  
  
Huff <- Huff[c(8,9,1,2,3,4,12,10,11,7,6,5)]  
  
kable(head(Huff[,1:ncol(Huff)]), format = "markdown")

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| poll\_name | poll\_week | Trump | Clinton | Other | Undecided | month | poll\_start\_date | poll\_end\_date | type\_of\_voter | num\_of\_voters | Spread |
| Morning | Sep 6  Sep 8 | 43 | 44 | NA | 13 | 9 | 2016-09-06 | 2016-09-08 | Likely | 1710 | Clinton +1 |
| ABC/Post | Sep 5  Sep 8 | 43 | 51 | 3 | 4 | 9 | 2016-09-05 | 2016-09-08 | Likely | NA | Clinton +8 |
| UPI/CVOTER | Aug 31  Sep 6 | 47 | 47 | 6 | NA | 8 | 2016-08-31 | 2016-09-06 | Likely | 1262 |  |
| Ipsos/Reuters | Sep 1  Sep 5 | 38 | 40 | 11 | 11 | 9 | 2016-09-01 | 2016-09-05 | Likely | 1084 | Clinton +2 |
| CNN | Sep 1  Sep 4 | 49 | 48 | 4 | 0 | 9 | 2016-09-01 | 2016-09-04 | Likely | 786 | Trump +1 |
| NBC/SurveyMonkey | Aug 29  Sep 4 | 42 | 48 | NA | NA | 8 | 2016-08-29 | 2016-09-04 | Registered | 32226 | Clinton +6 |

Data has to be transformed to obtain certain plots to explain the trend of the candidates across polls and in various months. Data is re-arranged for better readability

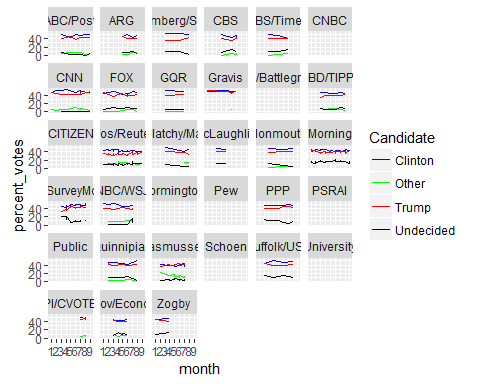
# Transpose data  
  
Huff\_gathered <- Huff %>% gather(Candidate,percent\_votes,Trump:Undecided)  
  
Huff\_gathered$Candidate <- as.factor(Huff\_gathered$Candidate)  
  
Huff\_gathered$Poll <- NULL  
  
Huff\_gathered <- Huff\_gathered[c(1,2,9,10,3,4,5,6,7,8)]  
  
kable(head(Huff\_gathered[,1:ncol(Huff\_gathered)]), format = "markdown")

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| poll\_name | poll\_week | Candidate | percent\_votes | month | poll\_start\_date | poll\_end\_date | type\_of\_voter | num\_of\_voters | Spread |
| Morning | Sep 6  Sep 8 | Trump | 43 | 9 | 2016-09-06 | 2016-09-08 | Likely | 1710 | Clinton +1 |
| ABC/Post | Sep 5  Sep 8 | Trump | 43 | 9 | 2016-09-05 | 2016-09-08 | Likely | NA | Clinton +8 |
| UPI/CVOTER | Aug 31  Sep 6 | Trump | 47 | 8 | 2016-08-31 | 2016-09-06 | Likely | 1262 |  |
| Ipsos/Reuters | Sep 1  Sep 5 | Trump | 38 | 9 | 2016-09-01 | 2016-09-05 | Likely | 1084 | Clinton +2 |
| CNN | Sep 1  Sep 4 | Trump | 49 | 9 | 2016-09-01 | 2016-09-04 | Likely | 786 | Trump +1 |
| NBC/SurveyMonkey | Aug 29  Sep 4 | Trump | 42 | 8 | 2016-08-29 | 2016-09-04 | Registered | 32226 | Clinton +6 |

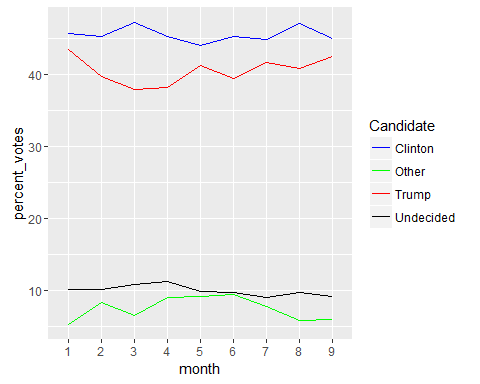
## Exploratory Data Analysis

Line Plots for all the candidates over the months as reported by Various Polls.

ggplot(Huff\_gathered,aes(month,percent\_votes)) +   
 geom\_line(data = Huff\_gathered,aes(group = Candidate,color = Candidate)) +  
 facet\_wrap(~poll\_name) + scale\_colour\_manual(values=c("blue","green","red","black"))



# Trend from January to Present   
  
Huff\_CT\_trend <- aggregate(percent\_votes ~ Candidate + month,Huff\_gathered,mean)  
  
ggplot(Huff\_CT\_trend,aes(month,percent\_votes)) + geom\_line(data = Huff\_CT\_trend,aes(group = Candidate,color = Candidate)) + scale\_colour\_manual(values=c("blue","green","red","black"))

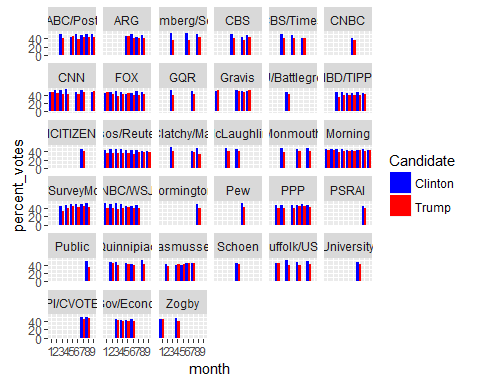


#### Conclusions

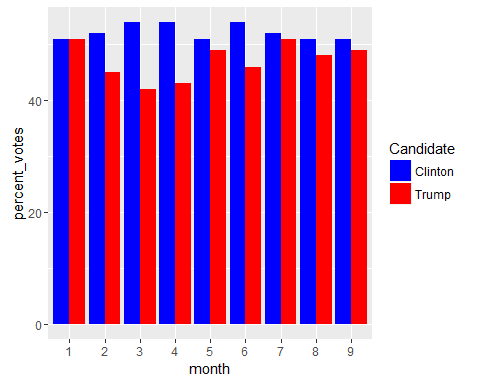
1. "Undecided" and "Other" Candidates do not seem to be in competition as indicated by all of the polls (Plot 1.).
2. Percent Votes from combined polls for each candidate based on month also indicates that the "Undecided" and "Other" Candidates are not in competition from beginning of the year. (Plot 2.)
3. Hence further reporting and predictions will be done for the top runners Hillary Clinton and Donald Trump.

#### Bar Plots for Clinton Vs Trump Based on Polls and well as monthly Trends

Huff\_CT<- Huff\_gathered %>% filter(Candidate == "Clinton" |Candidate == "Trump")  
  
ggplot(Huff\_CT, aes(x=month,y=percent\_votes,fill=Candidate))+  
 geom\_bar(stat="identity",position="dodge") +   
 facet\_wrap(~poll\_name) +   
 scale\_fill\_manual(values=c("blue","red"))



ggplot(Huff\_CT, aes(month,percent\_votes, fill=Candidate)) + geom\_bar(stat="identity",position="dodge") +   
scale\_fill\_manual(values=c("blue","red"))



#### Conclusions

Most of the poll results as well as overall monthly trend indicates that Hillary Clinton has been constantly leading since January.

It is intersting to note that the difference in votes varies on monthly basis this could be mostly due to the issues addressed by the candidates in their rallies. Although the impact of issues on Voter's mood will not be considered in the current report, such a study in future could show intersting trends.

## Predictions

Based on the observations and conclusions predicitions will be done for Hillary Clinton using TimeSeries Modelling.

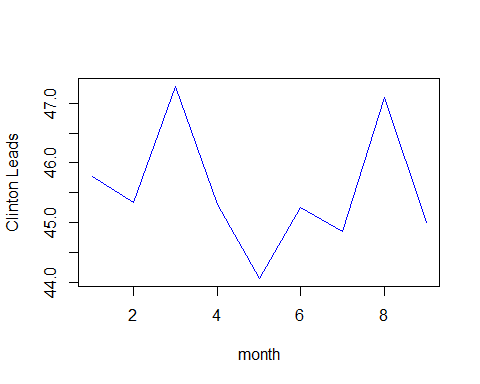
Data needs to be stored in Time-Series format

Huff$ClintonLeads <- Huff$Clinton - Huff$Trump  
Huff\_clinton <- select(Huff,month,Clinton)  
Huff\_clintonLeads <- select(Huff,month,ClintonLeads)  
  
Huff\_agg <- aggregate(Clinton~month,data= Huff\_clinton,mean)  
Huff\_ts\_agg <- ts(Huff\_agg$Clinton)

Plot of the time series data is generated. This gives the trend over time

#### Time Series Plots for Clinton monthly Trends

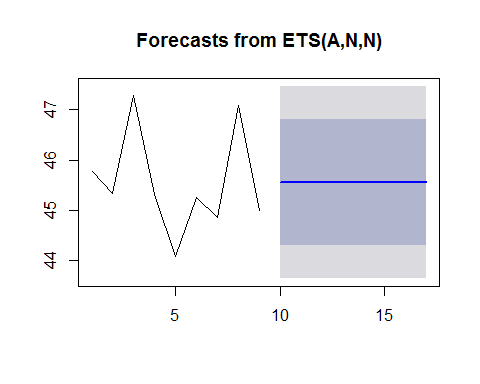
plot.ts(Huff\_ts\_agg,col="blue",xlab = "month", ylab = "Clinton Leads")



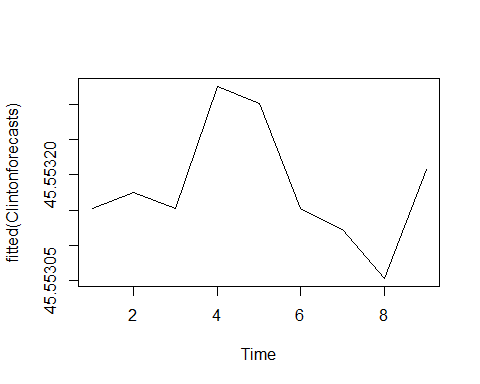
Looking at the plot, there seems to have been fluctuations for Clinton earlier during the year, this could probably be because there was race among democratic candidates. However after May , Clinton has been having Upward trend. The fluctuation in the upward trend could be due to the issues she has been addressing during the rallies. This is a potential factor that could be used for predictions in the future(this report does not use the issues addressed by candidates)

Predictions are done using forecast() function.

Clintonforecasts <- forecast(Huff\_ts\_agg,h=8,level=c(80,95))  
  
plot(Clintonforecasts)



plot(fitted(Clintonforecasts))



summary(Clintonforecasts)

##   
## Forecast method: ETS(A,N,N)  
##   
## Model Information:  
## ETS(A,N,N)   
##   
## Call:  
## ets(y = object, lambda = lambda, allow.multiplicative.trend = allow.multiplicative.trend)   
##   
## Smoothing parameters:  
## alpha = 1e-04   
##   
## Initial states:  
## l = 45.5532   
##   
## sigma: 0.9778  
##   
## AIC AICc BIC   
## 23.37161 25.37161 23.76606   
##   
## Error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 2.079566e-05 0.9778377 0.776689 -0.0454914 1.694566 0.5390525  
## ACF1  
## Training set -0.2058676  
##   
## Forecasts:  
## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## 10 45.55315 44.3 46.8063 43.63663 47.46968  
## 11 45.55315 44.3 46.8063 43.63663 47.46968  
## 12 45.55315 44.3 46.8063 43.63663 47.46968  
## 13 45.55315 44.3 46.8063 43.63663 47.46968  
## 14 45.55315 44.3 46.8063 43.63663 47.46968  
## 15 45.55315 44.3 46.8063 43.63663 47.46968  
## 16 45.55315 44.3 46.8063 43.63663 47.46968  
## 17 45.55315 44.3 46.8063 43.63663 47.46968

From the forecast above, it is very much apparent that Clinton will maintain her lead in next couple of months up until the election which will be result of final election as well.

This report can be further used to study the opinion polls conducted by various agencies and get the mood of general public. This could help the Presidential candidates to address the issues identified in the polls. Since this report is a live up-to-date timeline series, strategists of the respective parties could identify the changes and act immediately to reverse any adverse trends.

## Acknowledgements

I am grateful to my mentor **Joel Bangalan** for being a constant support throught out my Data Science learning Journey and Suggesting Capstone Project and Time Series modelling method.