Problem description is documented in the Github in the below link.

<https://github.com/harshalas/US2016PresidentialElections/blob/master/Proposal.docx>

***Deeper Dive into Dataset***

Live data is extracted from Huffington Post link below, whenever the code is run,

<http://elections.huffingtonpost.com/pollster/2016-general-election-trump-vs-clinton>

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

· Based on the poll week, polls need to be sorted on a monthly basis.

· Data had to be transformed to obtain certain plots to explain the trend of the candidates across polls and in various months.

· Columns - Spread, type\_of\_voter, number\_of\_voters have not been used for the present analysis, however they could be used in the future.

· Also, electoral votes along with historical winners from previous election would have been a good information for prediction but that is not being used as part of current analysis.

· Current Analysis predicts based solely on the Polls i.e, Whom Voters would elect if the elections were held today.

· The moods of voters is highly dependent on the issues addressed by Presidential Candidates in the Campaign rallies and a great swing is seen in the polls right after the rally. Hence monthly trend is a significant predictor.

· Although Polls from Popular Sources would give a better idea, current analysis makes use of all the available polls published by Huffington Post in order to rule out favoring of a particular candidate by a given source.

***Preliminary Exploration***

Preliminary exploration of data indicates that “Other” and “Undecided” candidates have a very low votes based on the polls and have continued the trend from January 2016. Therefore, for Current predictions, only the two major candidates Hillary Clinton, Donald Trump will be monitored.

After the initial analysis, there is no need for any change of approach initially proposed . Only change would be the two forerunners will be considered for predictions instead of 4 . Initial trend indicates Hillary Clinton to be leading in most of the polls.