**NEW JERSEY INSTITUTE OF TECHNOLOGY**

|  |
| --- |
| Default Project |
| Election Prediction of 538’s Forecast |
|  |
| **Garudadri,Sharavan Koushik** |
| **14-Dec-16** |

**THEORY AND UNDERSTANDING BEHIND EACH QUESTION IS WRITTEN AS COMMENTS IN THE \*.R POGRAM FILES**

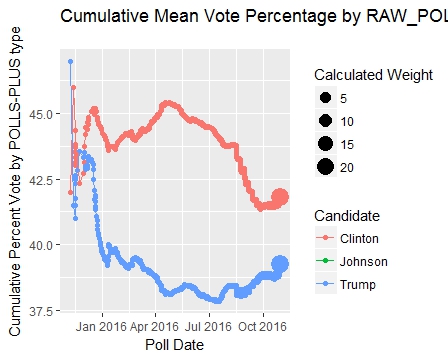
**Question 1:**

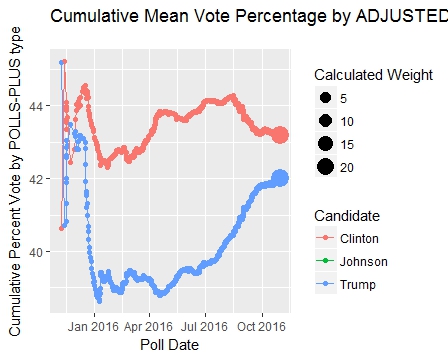
**Required Packages: dplyr,tidyr,stringr,libridate,ggplot2**

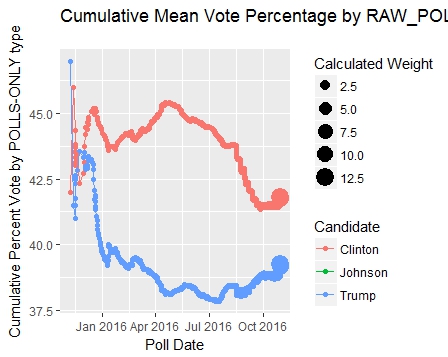
**How to execute program:**

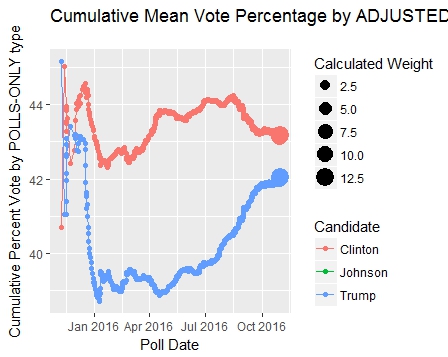
1. Open Rstudio
2. File->Open File , open Polls.R
3. Change the path of the resource files respective to the path in the machine executing the program
4. Execute each line individually or Ctrl-A and Run.
5. Order of Execution of Files
   * 1. Polls.R
     2. Filters.R
     3. Trends\_Over\_Time\_Now\_Cast.R
     4. Trends\_Over\_Time\_Polls\_Only.R
     5. Trends\_Over\_Time\_Polls\_Plus.R

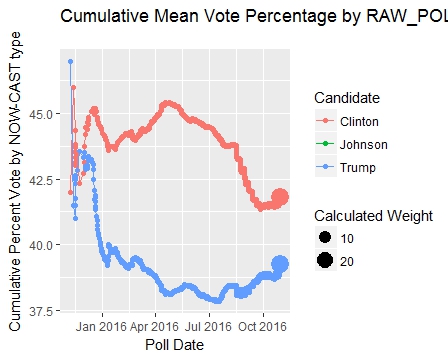
Note: Order of execution is a necessary step , so please execute the files in the above order

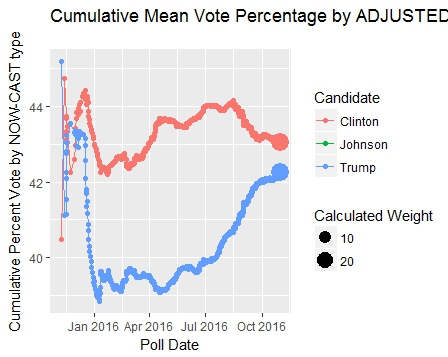
**Output:**

****

****

****

****

****

**Question 2, Question 3:**

**Required Packages: dplyr,tidyr,stringr,libridate,ggplot2,e1071,kLar,ROCR**

**How to execute program:**

1. Open Rstudio
2. File->Open File , open 5Folds.R
3. Change the path of the resource files respective to the path in the machine executing the program
4. Execute each line individually or Ctrl-A and Run.
5. Order of Execution of Files
   * 1. Polls.R
     2. Filters.R
     3. 5Folds.R

**Note: Order of execution is a necessary step , so please execute the files in the above order.**

**Approach**

Two different approaches:

**Approach 1:**

* Calculated Weighted Mean & Mean for 3 types of Filters/Categories
* Days passed since the forecast ended from today()[package : lubridate]
* A new factor called weight is calculated using

**WT**<- poll\_wt\* sqrt(samplesize)/dayspassed

* Each contender is now given 2 weights based on the factor WT ,

**ContendorWT** <- WT\* AdjustedPollwt

* Contender mean weight = mean(adjPollwt) of each contender
* **Contender Weighted mean weight** = Cumsum(ContenderWT/WT)
* Winner of a region is calculated based on contender with max **ContenderWT**

Predictors/Descriptors:

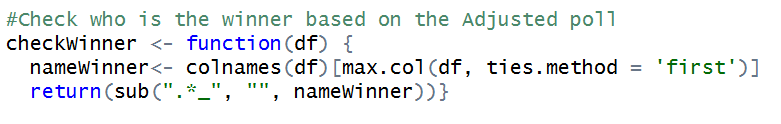
* ContenderWT
* State
* Grade

**Approach 2 :**

* From the Trends over time plots it is clear that the competition between Trump,Clinton is precisely more higher compared to remaining contenders,so the details of the remaining contenders is discarded.
* To make the results more stationary over each month of Trump and Clinton,

the difference between their Adjusted\_Poll weights is calculated ,and to verify the spread across all the states , it is then multipllied by each instance of Poll\_wt.

* The winner & loser of the region is calculated based a custom method CheckWinner() to which the Adjusted Poll weights are given as arguments



Predictors /Descriptors:

* Grade
* State
* Spread\_Across

**Classification Algorithms**

Three different types of classification algorithms are used

1. Logistic Regression
2. Space Vector Machine(SVM)
3. Naïve Bayes

Both the approaches mentioned above are implemented for all the three classification algorithms,but only the Approach 2 results are documented

**Output Plots [ACCURACY,PRECISION,TPR/FPR]for all the Folds(5-Folds) cross validation test for all the three Classifiers**

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

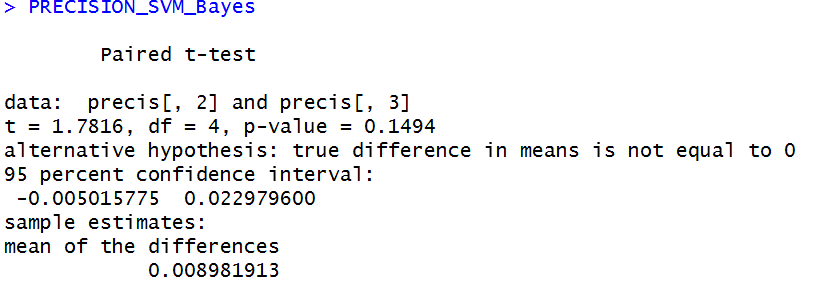
**Question 4**

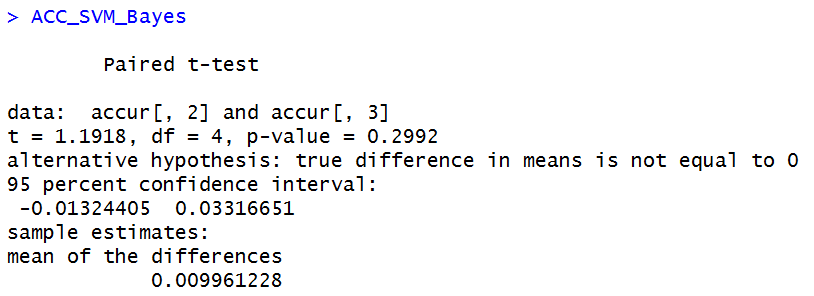
**Required Packages: dplyr,tidyr,stringr,libridate,ggplot2,e1071,kLar,ROCR**

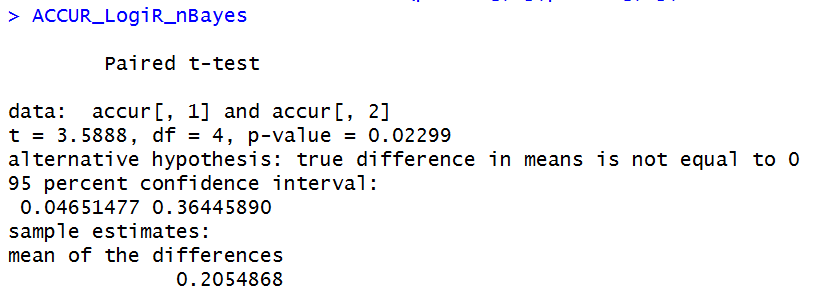
**How to execute program:**

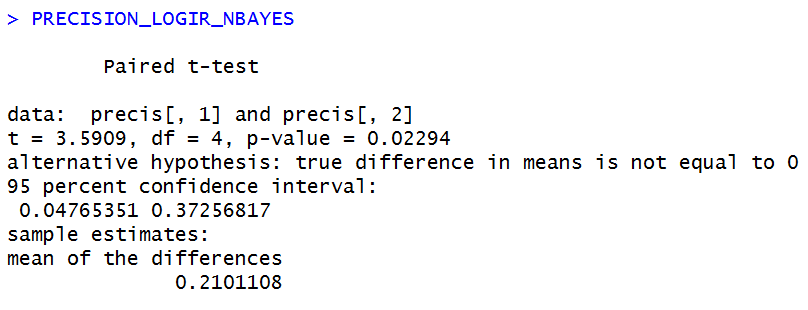
1. Open Rstudio
2. File->Open File , open Tests.R
3. Change the path of the resource files respective to the path in the machine executing the program
4. Execute each line individually or Ctrl-A and Run.

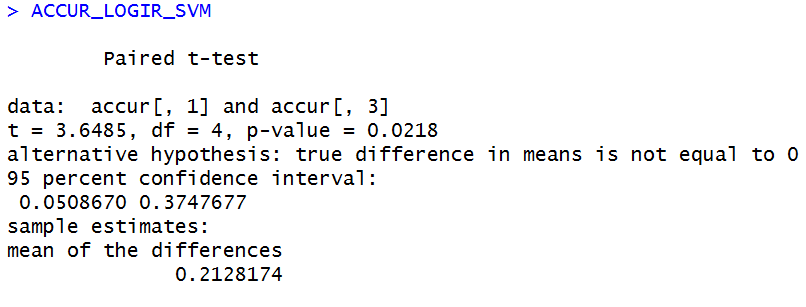
**Since there are three classifiers , a t-test is used here with significance of 95%**

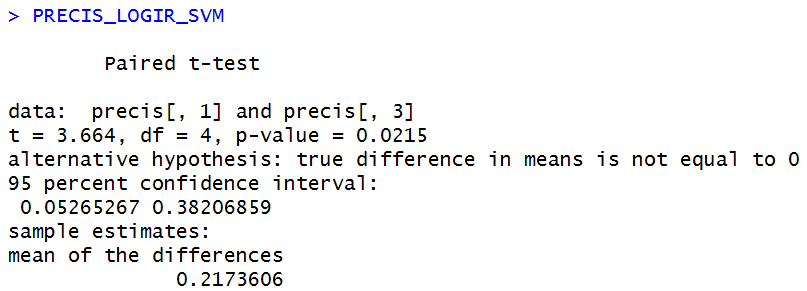












**To display the above T-TEST results in the Rstudio , please run the following snipping of code**

Accuracy comparision between Logistic Regression & Naïve Bayes

* ACCUR\_LogiR\_nBayes

Precision comparision between Logistic Regression & Naïve Bayes

* PRECISION\_LOGIR\_NBAYES

Accuracy comparision between SVM & Naïve Bayes

* ACC\_SVM\_Bayes

Precision comparision between SVM & Naïve Bayes

* PRECISION\_SVM\_Bayes

Accuracy comparision between Logistic Regression & SVM

* ACCUR\_LOGIR\_SVM

Precision comparision between Logistic Regression & SVM

* PRECIS\_LOGIR\_SVM

To conclude, according to the following table **SVM Model [Third Column]** is less precise than **Naïve bayes [Second Column]** , than **Logistic Regression [ First Column].** **Model SVM & Naïve Bayes are not significantly more or less accurate than one another.**

