Predictive Analysis

Mini Project : 4

SUPRASANNA PRADHAN
PGPBABI-O DEC'18, GROUP: 5

TABLE OF CONTENTS

1 THE ASSIGNMENT

Case Study: state assembly Election result of 2017
Problem description

Explain the factors behind the election 2017 .

2017 the election was held in March 2017 for Goa , Uttarakhand, Uttar Pradesh and Manipur, in our case we are considering the data for UP 2017 election for analysis purpose .

2 UNDERSTANDING THE DATA SET

These are the below data sets we brought together from difference sources

1. Election results 2017
   /https://www.eci.gov.in/files/file/4091-state-legislative-assembly-2017/

2. Election result 2012
   My neta data is contains candidate's personal data- link

   http://www.myneta.info/uttarpradesh2017/index.php?action=summary&subAction=candidates_analyzed&sort=candidate#summary

3. Literacy data distract wise

   lhttps://updateox.com/india/district-wise-male-female-literacy-rate-in-india-2011-census/#

3 IMPORTING DATA ELECTION RESULT OF 2017

In this data we are considering the election result of 2017 for state UP only hence we have sorted out our data with station code S24

```
~/My Files/Great Lakes Projects/Project -4/
> #***Importing the election data of 2017***#
> setwd ("~/My Files/Great Lakes Projects/Project -4")
> library(readxl)
> el_data17 <- read_excel("C:/Users/SuprasannaPradhan/Documents/My Files//Great Lakes Projects/Project -4/LA_2017.x
ls",)
> #Sorting only Utter Pradesh for 2017#
> data17<-(subset(el_data17, ST_CODE == "S24"))
> View(data17)
```

The data set has got 5307 observations and 15 variables

```
> str(data17)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       5307 obs. of  15 variables:
 $ ST_CODE             : chr  "S24" "S24" "S24" "S24" ...
 $ ST_NAME             : chr  "Uttar Pradesh" "Uttar Pradesh" "Uttar Pradesh" "Uttar Pradesh" ...
 $ MONTH               : num  3 3 3 3 3 3 3 3 3 3 ...
 $ YEAR                : num  2017 2017 2017 2017 2017 ...
 $ DIST_NAME           : chr  "Saharanpur" "Saharanpur" "Saharanpur" "Saharanpur" ...
 $ AC_NO               : num  1 1 1 1 1 1 1 1 1 1 ...
 $ AC_NAME             : chr  "Behat" "Behat" "Behat" "Behat" ...
 $ AC_TYPE             : chr  "GEN" "GEN" "GEN" "GEN" ...
 $ CAND_NAME           : chr  "NARESH SAINI" "MAHAVEER SINGH RANA" "MOHD. IQBAL" "RANA ADITYA PRATAP SINGH" ...
 $ CAND_SEX            : chr  "M" "M" "M" "M" ...
 $ CAND_CATEGORY       : chr  "GEN" "GEN" "GEN" "GEN" ...
 $ CAND_AGE            : chr  "53" "55" "52" "49" ...
 $ PARTYABBRE          : chr  "INC" "BJP" "BSP" "IND" ...
 $ TOTALVALIDVOTESPOLLED: num  97035 71449 71019 4187 1576 ...
 $ POSITION            : num  1 2 3 4 5 6 7 8 9 10 ...
```

Further we have imported the data from myneta.com
It has been having 4829 observation and 8 variables

```
> #*** Importing data myneta.com***#
> library(installr)
> library(RCurl)
> library(XML)
> library(stringr)
> xml.url<- "http://www.myneta.info/uttarpradesh2017/index.php?action=summary&subAction=candidates_analyzed&sort=ca
ndidate#summary"
> nettable <- readHTMLTable(xml.url,which = 3)
> View(nettable)
> str(nettable)
'data.frame':   4829 obs. of  8 variables:
 $ Sno           : Factor w/ 4829 levels "1","10","100",..: 4828 4829 1 1112 2223 3334 4273 4384 4495 4606 ...
 $ Candidateâ^‡  : Factor w/ 3877 levels "A Hasiv","A Wahid",..: NA 587 1 2 3 4 5 6 7 8 ...
 $ Constituency  : Factor w/ 404 levels "AGRA CANTT. (SC)",..: NA 106 20 133 150 276 207 59 316 392 ...
 $ Party         : Factor w/ 313 levels "Aadhi Aabadi Party",..: NA 202 99 124 290 127 154 124 124 71 ...
 $ Criminal Case : Factor w/ 23 levels "0","1","10","11",..: NA 23 1 1 1 1 1 1 2 1 ...
 $ Education     : Factor w/ 13 levels "10th Pass","12th Pass",..: NA 6 2 1 7 9 10 11 2 10 ...
 $ Total Assets  : Factor w/ 4275 levels "Nil","Rs 1,00,00,000 ~ 1 Crore+",..: NA 4275 2366 3823 2848 1713 2521 910
 1855 3735 ...
 $ Liabilities   : Factor w/ 1151 levels "Liabilities",..: NA 1 886 2 2 2 2 2 2 2 ...
>
```

Few column’s data has come with some special characters , hence we have removed these characters and renamed these columns - these Total assets and liabilities

```
> #Removing the sepcial characters#
> write.csv(neta1, file = "neta.csv")
> neta <-data.frame(neta1)
> neta$Total.Assets <- gsub("Rs","",neta$Total.Assets)
> View(neta)
> neta$new_totalas <- gsub("~.*", "", c(neta$Total.Assets))
> View(neta)
> neta$Liabilities <- gsub("Rs","",neta$Liabilities)
> neta$new_liabilities <- gsub("~.*", "", c(neta$Liabilities))
> View(neta)
> neta_data <- subset(neta, select = -c(6:7))
> str(neta_data)
'data.frame':   4829 obs. of  7 variables:
 $ Candidateâ..    : Factor w/ 3877 levels "A Hasiv","A Wahid",..: NA 587 1 2 3 4 5 6 7 8 ...
 $ Constituency.  : Factor w/ 404 levels "AGRA CANTT. (SC)",..: NA 106 20 133 150 276 207 59 316 392 ...
 $ Party.         : Factor w/ 313 levels "Aadhi Aabadi Party",..: NA 202 99 124 290 127 154 124 124 71 ...
 $ Criminal.Case. : Factor w/ 23 levels "0","1","10","11",..: NA 23 1 1 1 1 1 1 2 1 ...
 $ Education.     : Factor w/ 13 levels "10th Pass","12th Pass",..: NA 6 2 1 7 9 10 11 2 10 ...
 $ new_totalas    : chr  NA "Total Assets" " 3,94,24,827 " " 75,106 " ...
 $ new_liabilities: chr  NA "Liabilities" " 58,46,335 " " 0 " ...
>
```

	Candidateâ..	Constituency.	Party.	Criminal.Case.	Education.	new_totalas	new_liabilities
1	NA	NA	NA	NA	NA	NA	NA
2	Candidateâ	Constituency	Party	Criminal Case	Education	Total Assets	Liabilities
3	A Hasiv	ARYA NAGAR	BSP	0	12th Pass	3,94,24,827	58,46,335
4	A Wahid	GAINSARI	IND	0	10th Pass	75,106	0
5	Aan Shikhar Shrivastava	GOSHAINGANJ	Satya Shikhar Party	0	Graduate	41,000	0
6	Aaptab Urf Aftab	MUBARAKPUR	Islam Party Hind	0	Illiterate	20,000	0
7	Aashi Gaur	KHATAULI	Lok Dal	0	Literate	34,68,543	0
8	Aashif Beg	BARKHERA	IND	0	Not Given	13,30,000	0
9	Aatma Ram	PURQAZI (SC)	IND	1	12th Pass	23,26,259	0

	Candidateâ..
1	NA
2	Candidateâ
3	A Hasiv
4	A Wahid
5	Aan Shikhar Shri
6	Aaptab Urf Aftab
7	Aashi Gaur
8	Aashif Beg
9	Aatma Ram

Renamed the column data of my neta to merge it with election result data of 2017

```
> # Added and Renaming the colnames#
> #as.data.frame(gsub("[[:punct:]]", "", as.matrix(neta_data)))
> colnames(neta_data)[colnames(neta_data)=="Candidate"] <- "CAND_NAME"
> names(neta_data)
[1] "Candidateâ.."    "Constituency."   "Party."          "Criminal.Case."  "Education."      "new_totalas"
[7] "new_liabilities"
> names(neta_data)[1]<-"CAND_NAME"
> names(neta_data)[2]<-"AC_NAME"
> names(neta_data)[3]<-"PART_NAME"
> names(neta_data)[4]<-"CRIM_CASE"
> names(neta_data)[5]<-"EDUCATION"
> names(neta_data)[6]<-"TOTA_ASST"
> names(neta_data)[7]<-"LIABILITES"
> names(neta_data)
[1] "CAND_NAME"  "AC_NAME"    "PART_NAME"  "CRIM_CASE"  "EDUCATION"  "TOTA_ASST"  "LIABILITES"
> View(neta_data)
```

Summary of my neta .com data set

```
> names(neta_data)[1]<-"CAND_NAME"
> names(neta_data)[2]<-"AC_NAME"
> names(neta_data)[3]<-"PARTYABBRE"
> names(neta_data)[4]<-"CRIM_CASE"
> names(neta_data)[5]<-"EDUCATION"
> names(neta_data)[6]<-"TOTA_ASST"
> names(neta_data)[7]<-"LIABILITES"
> names(neta_data)
[1] "CAND_NAME"  "AC_NAME"    "PARTYABBRE" "CRIM_CASE"  "EDUCATION"  "TOTA_ASST"  "LIABILITES"
> summary(neta_data)
        CAND_NAME              AC_NAME       PARTYABBRE     CRIM_CASE            EDUCATION
 Manoj Kumar :  19   AGRA SOUTH     :  26   IND    :1454   0      :3965   Graduate     : 905
 Ajay Kumar  :  18   ALLAHABAD NORTH:  26   BSP    : 400   1      : 440   Post Graduate: 862
 Anil Kumar  :  18   AMETHI         :  24   BJP    : 384   2      : 161   12th Pass    : 830
 Rakesh Kumar:  15   VARANASI CANTT.:  24   SP     : 308   3      :  92   10th Pass    : 558
 Rajesh Kumar:  14   BHADOHI        :  23   RLD    : 276   4      :  54   8th Pass     : 479
 (Other)     :4744   (Other)        :4705   (Other):2006   (Other): 116   (Other)      :1194
 NA's        :   1   NA's           :   1   NA's   :   1   NA's   :   1   NA's         :   1
    TOTA_ASST        LIABILITES
  20000  :  16    0       :3172
  100000 :  14    500000  :  37
 Nil     :  13    100000  :  27
  30000  :  11    200000  :  26
  50000  :  11    1000000 :  24
 (Other) :4763   (Other)  :1542
 NA's    :   1   NA's     :   1
```

Before merging the both data we must check is the variables text and values are same or not , since the data set to be merged based on candidate's name .
We found the candidates' names are written with lower caps and election result of candidates' names mentioned in upper case .Hence we have the neta .com data sets font small to cap.

```
> View(neta_data)
> #Merging the 2017 and my neta dat set #
> library(dplyr)
> neta_data <- neta_data %>% mutate_each(funs(toupper),CAND_NAME )
> neta_data <- neta_data %>% mutate_each(funs(tolower),AC_NAME )
> View(neta_data)
>
```

Merged data set of 2017 election data and my neta data .Now its having 8183 observation and 19 variables

```
> #Merging the 2017 and my neta dat set #
> library(dplyr)
> neta_data <- neta_data %>% mutate_each(funs(toupper),CAND_NAME )
> neta_data <- neta_data %>% mutate_each(funs(tolower),AC_NAME )
> mer_data17 <- merge(data17, neta_data, by='CAND_NAME')
> View(mer_data17)
> final_data17 <- subset(mer_data17, select = -c(16:17))
> View(final_data17)
> str(final_data17)
'data.frame':   8183 obs. of  19 variables:
 $ CAND_NAME            : chr  "A HASIV" "A WAHID" "AAN SHIKHAR SHRIVASTAVA" "AASHI GAUR" ...
 $ ST_CODE              : chr  "S24" "S24" "S24" "S24" ...
 $ ST_NAME              : chr  "Uttar Pradesh" "Uttar Pradesh" "Uttar Pradesh" "Uttar Pradesh" ...
 $ MONTH                : num  3 3 3 3 3 3 3 3 3 3 ...
 $ YEAR                 : num  2017 2017 2017 2017 2017 ...
 $ DIST_NAME            : chr  "Kanpur Nagar" "Balrampur" "Faizabad" "Muzaffarnagar" ...
 $ AC_NO                : num  214 292 276 15 128 13 291 271 271 354 ...
 $ AC_NAME.x            : chr  "Arya Nagar" "Gainsari" "Goshainganj" "Khatauli" ...
 $ AC_TYPE              : chr  "GEN" "GEN" "GEN" "GEN" ...
 $ CAND_SEX             : chr  "M" "M" "M" "F" ...
 $ CAND_CATEGORY        : chr  "GEN" "GEN" "GEN" "GEN" ...
 $ CAND_AGE             : chr  "56" "42" "26" "25" ...
 $ PARTYABBRE           : chr  "BSP" "IND" "SATSHIP" "LD" ...
 $ TOTALVALIDVOTESPOLLED: num  6061 925 862 207 453 ...
 $ POSITION             : num  3 10 9 13 12 11 11 3 4 2 ...
 $ CRIM_CASE            : Factor w/ 23 levels "0","1","10","11",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ EDUCATION            : Factor w/ 13 levels "10th Pass","12th Pass",..: 2 1 7 10 11 10 10 7 7 2 ...
 $ TOTA_ASST            : chr  " 3,94,24,827 " " 75,106 " " 41,000 " " 34,68,543 " ...
 $ LIABILITES           : chr  " 58,46,335 " " 0 " " 0 " " 0 " ...
>
```

Importing election result 2012 and we have 15 variables here in the data set .

```
> #***Importing the election data of 2012***#
> el_data12 <- read_excel("C:/Users/SuprasannaPradhan/Documents/My Files//Great Lakes Projects/Project -4/LA_2012.
xls",)
> str(el_data12)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       11516 obs. of  15 variables:
 $ ST_CODE      : chr  "S05" "S05" "S05" "S05" ...
 $ ST_NAME      : chr  "Goa" "Goa" "Goa" "Goa" ...
 $ MONTH        : num  1 1 1 1 1 1 1 1 1 1 ...
 $ YEAR         : num  2012 2012 2012 2012 2012 ...
 $ DIST_NAME    : chr  "North Goa" "North Goa" "North Goa" "North Goa" ...
 $ AC_NO        : num  1 1 1 1 1 1 2 2 2 2 ...
 $ AC_NAME      : chr  "Mandrem" "Mandrem" "Mandrem" "Mandrem" ...
 $ AC_TYPE      : chr  "GEN" "GEN" "GEN" "GEN" ...
 $ CAND_NAME    : chr  "LAXMIKANT PARSEKAR" "DAYANAND RAGHUNATH SOPTE" "SAMEER SALGAOCAR" "MENINO FRANCIS MONTEIR
O" ...
 $ CAND_SEX     : chr  "M" "M" "M" "M" ...
 $ CAND_CATEGORY: chr  "GEN" "GEN" "GEN" "GEN" ...
 $ CAND_AGE     : num  56 48 42 51 65 38 56 58 37 61 ...
 $ PARTYABBRE   : chr  "BJP" "INC" "AITC" "GVP" ...
 $ TOTVOTPOLL   : num  11955 8520 4591 295 292 ...
 $ POSITION     : num  1 2 3 4 5 6 1 2 3 4 ...
> data12<-(subset(el_data12, ST_CODE == "S24"))
> names(data12)
 [1] "ST_CODE"       "ST_NAME"       "MONTH"         "YEAR"          "DIST_NAME"     "AC_NO"
 [7] "AC_NAME"       "AC_TYPE"       "CAND_NAME"     "CAND_SEX"      "CAND_CATEGORY" "CAND_AGE"
```

Imported data set of censuses 2011

```
> #***Importing the election data of 2011census***#
> #source https://updateox.com/india/district-wise-male-female-literacy-rate-in-india-2011-census/#
> cens_data11 <- read_excel("C:/Users/SuprasannaPradhan/Documents/My Files//Great Lakes Projects/Project -4/cens_2011.xl
s",)
> str(cens_data11)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       72 obs. of  4 variables:
 $ SL.NO          : num  NA 1 2 3 4 5 6 7 8 9 ...
 $ DISTRICT       : chr  NA "Saharanpur" "Muzaffarnagar" "Bijnor" ...
 $ MALE LITERACY  : num  NA 0.798 0.791 0.787 0.668 ...
 $ FEMALE LITERACy: num  NA 0.633 0.6 0.615 0.496 ...
```

Merging the census 2011 data and election result data of 2012

```
> names(cens_data11)[2]<-"DIST_NAME"
> cen_data11<- subset(cens_data11, select = -c(1))
> View(cen_data11)
> final_data12 <- merge(x = data12, y = cen_data11, by=c('DIST_NAME'),all.x = TRUE)
> View(final_data12)
> str(final_data12)
'data.frame':   7031 obs. of  17 variables:
 $ DIST_NAME      : chr  "Agra" "Agra" "Agra" "Agra" ...
 $ ST_CODE        : chr  "S24" "S24" "S24" "S24" ...
 $ ST_NAME        : chr  "Uttar Pradesh" "Uttar Pradesh" "Uttar Pradesh" "Uttar Pradesh" ...
 $ MONTH          : num  1 1 1 1 1 1 1 1 1 1 ...
 $ YEAR           : num  2012 2012 2012 2012 2012 ...
 $ AC_NO          : num  91 89 91 90 91 88 89 86 94 92 ...
 $ AC_NAME        : chr  "Fatehpur Sikri" "Agra North" "Fatehpur Sikri" "Agra Rural" ...
 $ AC_TYPE        : chr  "GEN" "GEN" "GEN" "SC" ...
 $ CAND_NAME      : chr  "JITENDRA FAUZDAR" "ANJU SINGH CHAUHAN" "OM PRAKASH" "KALI CHARAN SUMAN" ...
 $ CAND_SEX       : chr  "M" "F" "M" "M" ...
 $ CAND_CATEGORY  : chr  "GEN" "GEN" "GEN" "SC" ...
 $ CAND_AGE       : num  44 42 45 47 45 57 28 38 46 37 ...
 $ PARTYABBRE     : chr  "BJP" "VIP" "JaKP" "BSP" ...
 $ TOTVOTPOLL     : num  6927 179 15503 69969 604 ...
```

Preparing the final data which is merged the election data 2017 and 2012 - with left outer join where we are checking the election result 2017 candidate names are true.

```
> library(dplyr)
> #Elet1<- Elet %>% distinct(-AC_NO.x, .keep_all = TRUE)
> #Elet2 <- subset(Elet1, select = -c(8,9,25))
> str(Elet)
'data.frame':   5307 obs. of  19 variables:
 $ DIST_NAME    : Factor w/ 75 levels "Agra","Aligarh",..: 63 63 63 63 63 63 63 63 63 63 ...
 $ AC_NAME      : Factor w/ 403 levels "175-LUCKNOW CANTT.",..: 62 62 62 62 62 62 62 62 62 62 ...
 $ CAND_NAME    : Factor w/ 4110 levels "A HASIV","A WAHID",..: 2097 1690 1944 2922 2193 1442 306 1134 1072
 1544 ...
 $ CAND_SEX     : Factor w/ 4 levels "0","F","M","O": 3 3 3 3 1 3 3 3 3 3 ...
 $ CAND_CATEGORY: Factor w/ 4 levels "0","GEN","SC",..: 2 2 2 2 1 2 2 3 2 3 ...
 $ CAND_AGE     : int  53 55 52 49 0 37 38 39 32 35 ...
 $ PARTYABBRE   : Factor w/ 304 levels "AABHAP","AACP",..: 122 71 97 123 192 60 232 123 123 123 ...
 $ VOTE_2017    : int  97035 71449 71019 4187 1576 1255 1150 1113 945 810 ...
 $ POSITION_2017: int  1 2 3 4 5 6 7 8 9 10 ...
 $ CRIM_CASE    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ EDUCATION    : Factor w/ 13 levels "0","10th Pass",..: 13 8 10 3 1 2 7 4 7 10 ...
 $ TOTA_ASST    : Factor w/ 2790 levels " 1000 "," 10000 ",..: 104 2133 411 2315 2789 2359 2108 1028 415 95
5 ...
 $ LIABILITES   : num  398000 2985082 10214938 286731 0 ...
 $ VOTE_2012    : int  0 70274 0 0 0 0 12903 0 0 418 ...
 $ POSITION_2012: int  0 1 0 0 0 0 4 0 0 19 ...
 $ MALE_LITC    : num  0 0.798 0 0 0 ...
 $ FEMALE_LITC  : num  0 0.633 0 0 0 ...
 $ SEX          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CATEGORY     : num  1 1 1 1 0 1 1 0 1 0 ...
```

Prepared the final data set for our observation – we have total 5307 observations and 19 variables

Thera are 403 candidates were won the election of 2017 in UP, we will try to analyze here what are variables are helped explained the same .

We got almost all data non-continuous data besides Total validated vote polled VOTE_2017

Variable Position_2017 and Positon-2012 is carrying value of equal to or greater than 1, where 1 is marked for the candidate who won the election and greater than 1 they defeated in the election ,whereas the variable gender and category is also part of categorical data we have changed these either 0 or 1 as per the level values

Summary of the final data set

```
> summary(Elet)
     DIST_NAME              AC_NAME               CAND_NAME    CAND_SEX CAND_CATEGORY    CAND_AGE
 Allahabad: 193   Rudauli        :  30   None of the Above: 407   0: 407   0  : 407      Min.   : 0.00
 Gorakhpur: 136   Bilari         :  28   ANIL KUMAR       :  15   F: 484   GEN:3695      1st Qu.:34.00
 Lucknow  : 135   Agra South     :  27   MANOJ KUMAR      :  14   M:4415   SC :1179      Median :43.00
 Varanasi : 135   Allahabad North:  27   RAKESH KUMAR     :  14   O:   1   ST :  26      Mean   :41.45
 Agra     : 130   Amethi         :  25   AJAY KUMAR       :  12                         3rd Qu.:52.00
 Jaunpur  : 130   Varanasi Cantt.:  25   RAJESH KUMAR     :  12                         Max.   :91.00
 (Other)  :4448   (Other)        :5145   (Other)          :4833
   PARTYABBRE     VOTE_2017        POSITION_2017      CRIM_CASE              EDUCATION
 IND    :1476   Min.   :    44.0   Min.   : 1.000   Min.   : 0.0000   0            :1482
 BSP    : 407   1st Qu.:   562.5   1st Qu.: 4.000   1st Qu.: 0.0000   Graduate     : 720
 NOTA   : 407   Median :  1113.0   Median : 7.000   Median : 0.0000   Post Graduate: 650
 BJP    : 388   Mean   : 16509.2   Mean   : 7.636   Mean   : 0.3179   12th Pass    : 631
 SP     : 314   3rd Qu.: 10112.5   3rd Qu.:11.000   3rd Qu.: 0.0000   10th Pass    : 487
 RLD    : 280   Max.   :262741.0   Max.   :30.000   Max.   :36.0000   8th Pass     : 401
 (Other):2035                                                         (Other)      : 936
    TOTA_ASST      LIABILITES          VOTE_2012        POSITION_2012      MALE_LITC
 0        :1482   Min.   :        0   Min.   :     0.0   Min.   : 0.000   Min.   :0.0000
  145000  :  16   1st Qu.:        0   1st Qu.:     0.0   1st Qu.: 0.000   1st Qu.:0.0000
  180000  :  14   Median :        0   Median :     0.0   Median : 0.000   Median :0.0000
  2036761 :  14   Mean   :  1137800   Mean   :  5460.3   Mean   : 2.799   Mean   :0.2227
  226000  :  14   3rd Qu.:        0   3rd Qu.:   575.5   3rd Qu.: 3.000   3rd Qu.:0.7009
  100000  :  12   Max.   :260265000   Max.   :133563.0   Max.   :47.000   Max.   :0.9023
 (Other)  :3755
  FEMALE_LITC          SEX            CATEGORY
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :1.0000
 Mean   :0.1662   Mean   :0.0912   Mean   :0.6963
 3rd Qu.:0.5035   3rd Qu.:0.0000   3rd Qu.:1.0000
```

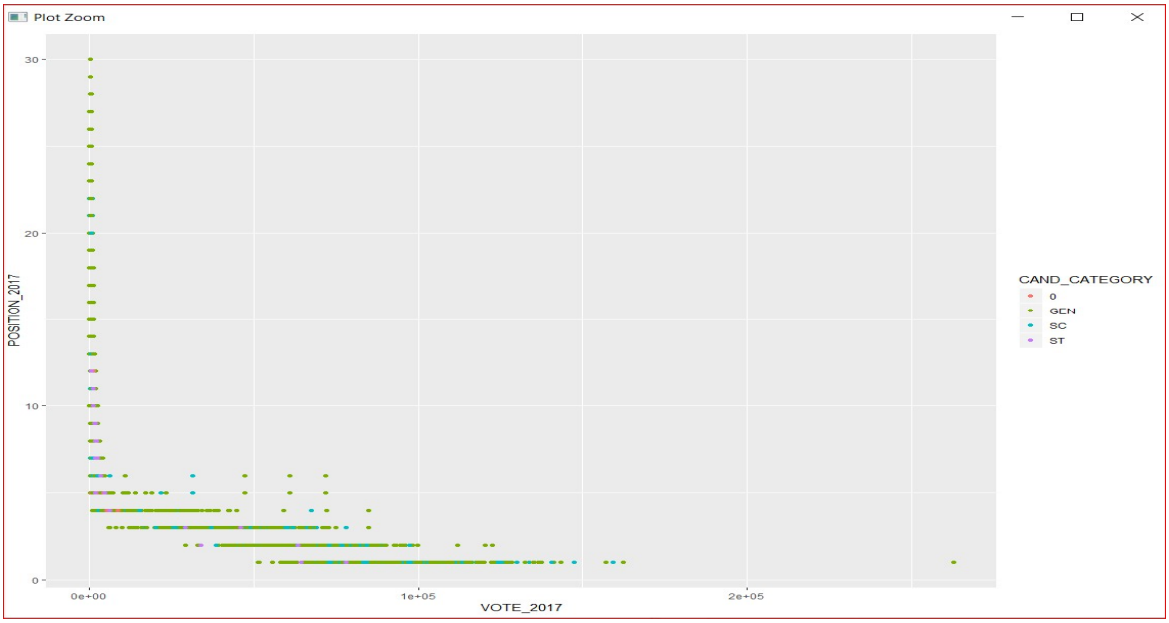
We have created some more numerical dummy variable by splitting main variable of the political parties performed better 2017 election .

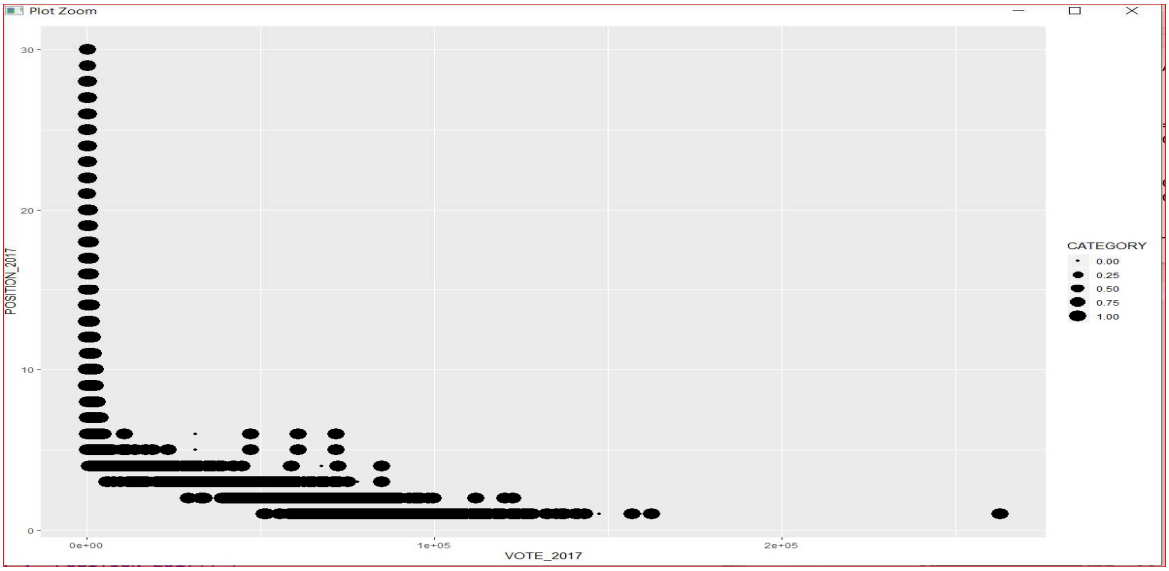
```
> Elet$PARTY_BSP<-ifelse(Elet$PARTYABBRE=="BSP",1,0)
> Elet$PATRY_INC<-ifelse(Elet$PARTYABBRE=="INC",1,0)
> Elet$PARTY_IND<-ifelse(Elet$PARTYABBRE=="IND",1,0)
> Elet$PARTY_NINSHAD<-ifelse(Elet$PARTYABBRE=="NINSHAD",1,0)
> Elet$PARTY_RLD<-ifelse(Elet$PARTYABBRE=="RLD",1,0)
> Elet$PARTY_SBSP<-ifelse(Elet$PARTYABBRE=="SBSP",1,0)
> Elet$PARTY_SP<-ifelse(Elet$PARTYABBRE=="SP",1,0)
> str(Elet)
'data.frame':   5307 obs. of  28 variables:
 $ DIST_NAME    : Factor w/ 75 levels "Agra","Aligarh",..: 63 63 63 63 63 63 63 63 63 63 ...
 $ AC_NAME      : Factor w/ 403 levels "175-LUCKNOW CANTT.",..: 62 62 62 62 62 62 62 62 62 62 ...
 $ CAND_NAME    : Factor w/ 4110 levels "A HASIV","A WAHID",..: 2097 1690 1944 2922 2193 1442 306 1134 1
1544 ...
 $ CAND_SEX     : Factor w/ 4 levels "0","F","M","O": 3 3 3 3 1 3 3 3 3 3 ...
 $ CAND_CATEGORY: Factor w/ 4 levels "0","GEN","SC",..: 2 2 2 2 1 2 2 3 2 3 ...
 $ CAND_AGE     : int  53 55 52 49 0 37 38 39 32 35 ...
 $ PARTYABBRE   : Factor w/ 304 levels "AABHAP","AACP",..: 122 71 97 123 192 60 232 123 123 123 ...
 $ VOTE_2017    : int  97035 71449 71019 4187 1576 1255 1150 1113 945 810 ...
 $ POSITION_2017: int  1 2 3 4 5 6 7 8 9 10 ...
 $ CRIM_CASE    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ EDUCATION    : Factor w/ 13 levels "0","10th Pass",..: 13 8 10 3 1 2 7 4 7 10 ...
 $ TOTA_ASST    : num  104 2133 411 2315 2789 ...
 $ LIABILITES   : num  398000 2985082 10214938 286731 0 ...
 $ VOTE_2012    : int  0 70274 0 0 0 0 12903 0 0 418 ...
 $ POSITION_2012: int  0 1 0 0 0 0 4 0 0 19 ...
 $ MALE_LITC    : num  0 0.798 0 0 0 ...
 $ FEMALE_LITC  : num  0 0.633 0 0 0 ...
 $ SEX          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CATEGORY     : num  1 1 1 1 0 1 1 0 1 0 ...
 $ PARTY_ADAL   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_BJP    : num  0 1 0 0 0 0 0 0 0 0 ...
 $ PARTY_BSP    : num  0 0 1 0 0 0 0 0 0 0 ...
 $ PATRY_INC    : num  1 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_IND    : num  0 0 0 1 0 0 0 1 1 1 ...
 $ PARTY_NINSHAD: num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_RLD    : num  0 0 0 0 0 0 0 1 0 0 0 ...
 $ PARTY_SBSP   : num  0 0 0 0 0 0 0 0 0 0 ...
```

Plotting gained votes position wise with SC, ST and General category

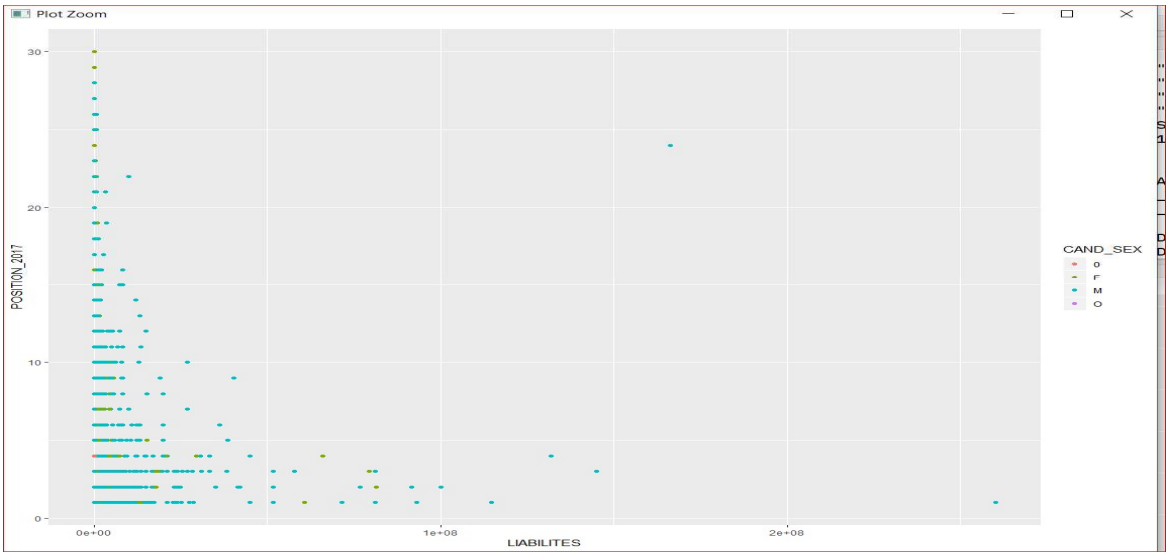
```
Console  Terminal  Jobs
~/My Files/R/R project files/
> qplot(VOTE_2017,POSITION_2017, colour = CAND_CATEGORY , data=Elet)
>
```

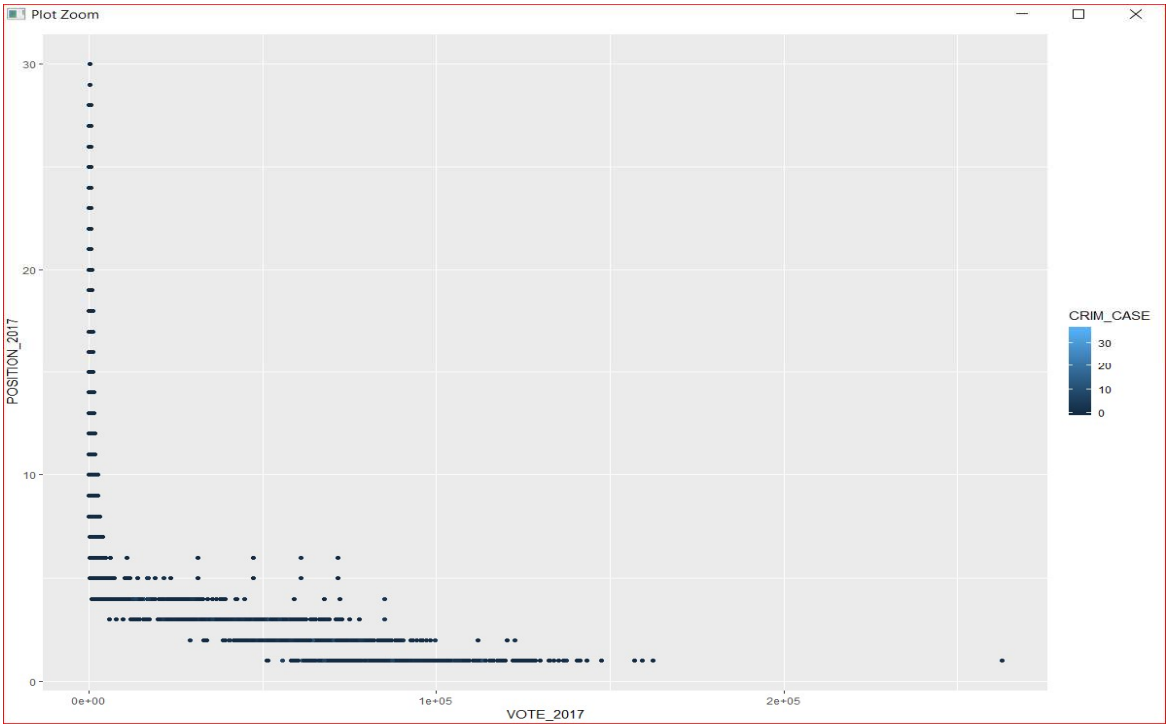
We wanted to find out how many votes are each candidate got with their Position 1 which is explaining us the candidate won the election





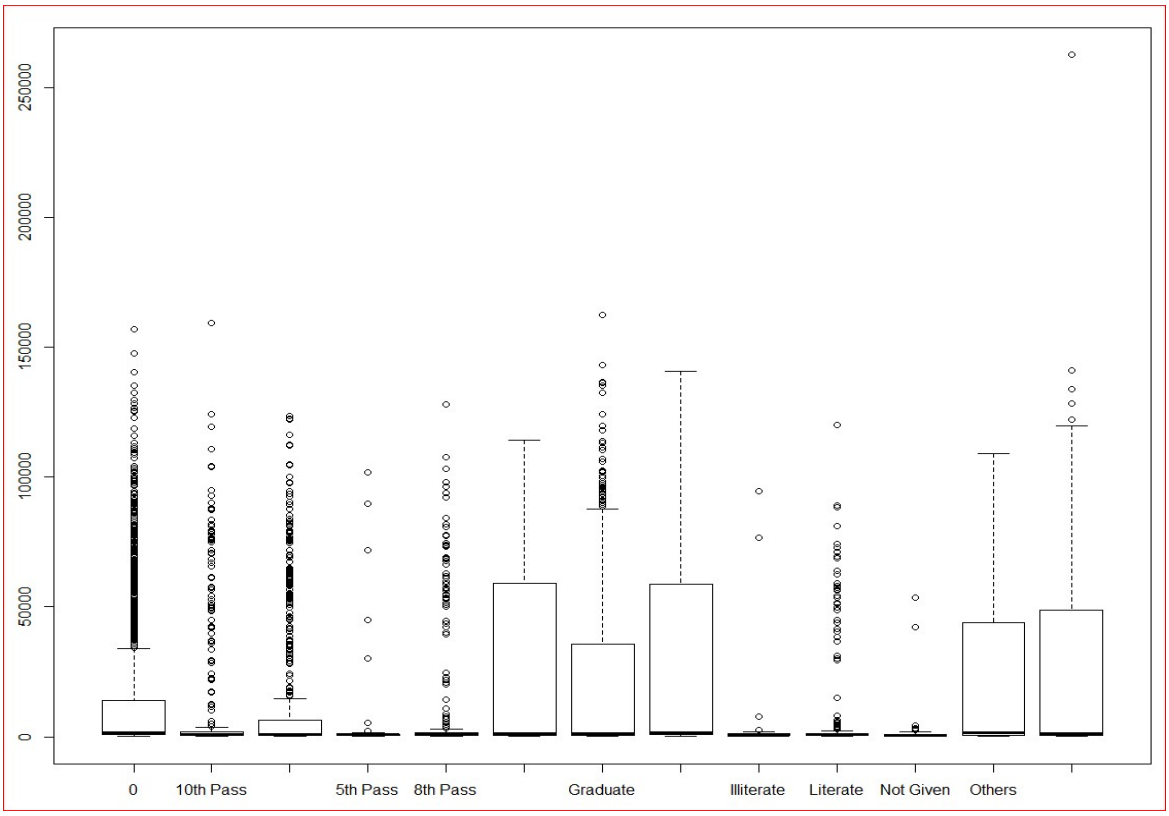

```
> qplot(LIABILITES,POSITION_2017, colour = CAND_SEX , data=Elet)
>
```





Box plot of received vote 2017 impacted with variable education found it is significant

```
> boxplot(Elet$VOTE_2017  ~Elet$EDUCATION,)
> aov.Job<-aov(Elet$VOTE_2017  ~Elet$EDUCATION)
> summary(aov.Job)
                Df    Sum Sq  Mean Sq F value Pr(>F)
Elet$EDUCATION   12 1.992e+11 1.66e+10   19.02 <2e-16 ***
Residuals      5294 4.622e+12 8.73e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> tk.Job<-TukeyHSD(aov.Job)
> tk.Job
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Elet$VOTE_2017 ~ Elet$EDUCATION)
```

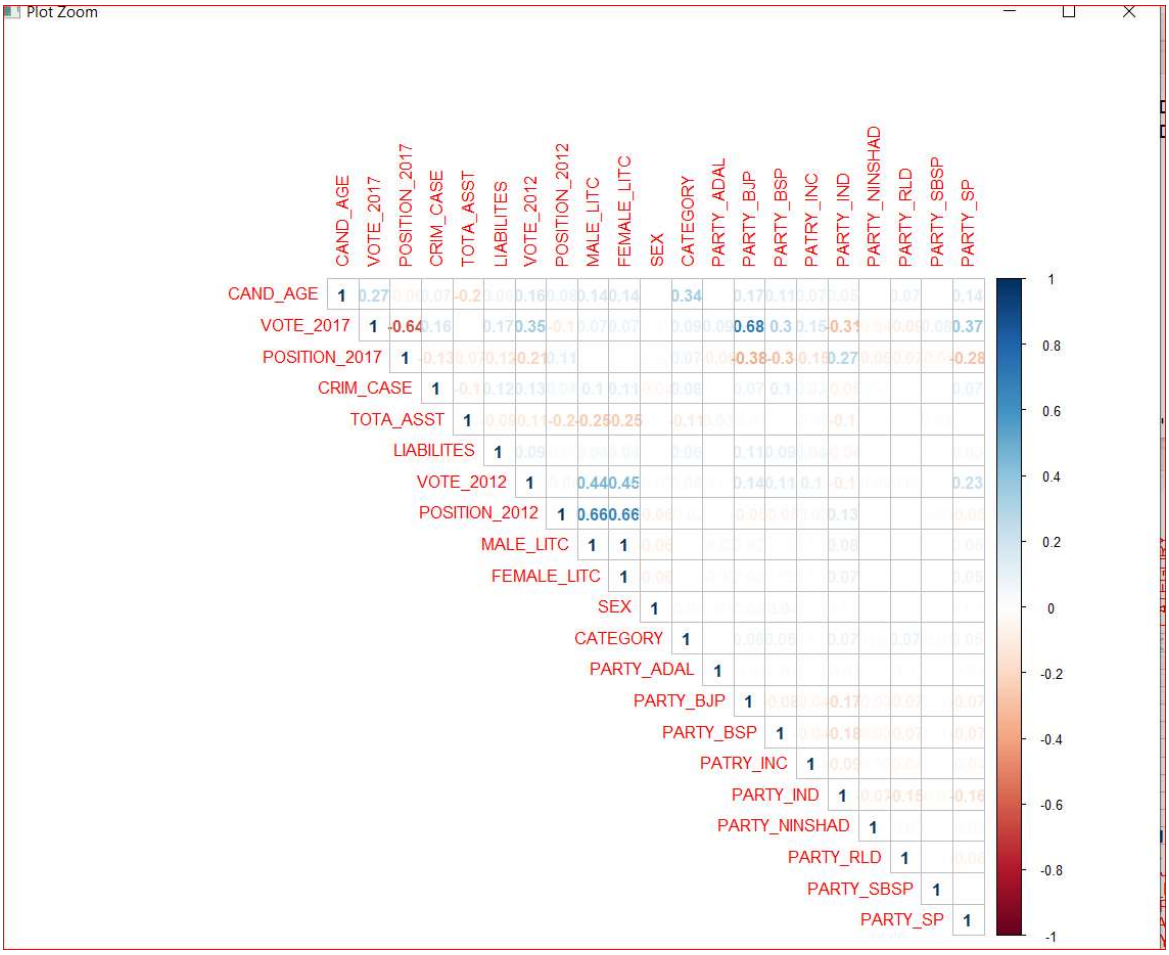



```
> tk.Job<-TukeyHSD(aov.Job)
> tk.Job
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Elet$VOTE_2017 ~ Elet$EDUCATION)

$`Elet$EDUCATION`
                                      diff          lwr         upr     p adj
10th Pass-0                     -6666.1790 -11781.12730  -1551.2307 0.0011351
12th Pass-0                     -2579.4183  -7234.39963   2075.5630 0.8313251
5th Pass-0                     -12432.3262 -23119.82208  -1744.8302 0.0076291
8th Pass-0                      -7955.8776 -13468.21570  -2443.5395 0.0001322
Doctorate-0                      9946.7145  -3161.06827  23054.4973 0.3644790
Graduate-0                       2787.5709  -1661.04886   7236.1906 0.6806470
Graduate Professional-0         11229.3803   5305.06816  17153.6924 0.0000000
Illiterate-0                   -11873.2544 -27377.15213   3630.6434 0.3492933
Literate-0                      -9835.1756 -15825.81645  -3844.5347 0.0000043
```

To check the coloration , we have further considered only numeric variables

```
> data1 <- subset(Elet, select = -c(1:5,7,11))
> str(data1)
'data.frame':   5307 obs. of  21 variables:
 $ CAND_AGE      : int  53 55 52 49 0 37 38 39 32 35 ...
 $ VOTE_2017     : int  97035 71449 71019 4187 1576 1255 1150 1113 945 810 ...
 $ POSITION_2017: int  1 2 3 4 5 6 7 8 9 10 ...
 $ CRIM_CASE     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ TOTA_ASST     : num  104 2133 411 2315 2789 ...
 $ LIABILITES    : num  398000 2985082 10214938 286731 0 ...
 $ VOTE_2012     : int  0 70274 0 0 0 0 12903 0 0 418 ...
 $ POSITION_2012: int  0 1 0 0 0 0 4 0 0 19 ...
 $ MALE_LITC     : num  0 0.798 0 0 0 ...
 $ FEMALE_LITC   : num  0 0.633 0 0 0 ...
 $ SEX           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CATEGORY      : num  1 1 1 1 0 1 1 0 1 0 ...
 $ PARTY_ADAL    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_BJP     : num  0 1 0 0 0 0 0 0 0 0 ...
 $ PARTY_BSP     : num  0 0 1 0 0 0 0 0 0 0 ...
 $ PATRY_INC     : num  1 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_IND     : num  0 0 0 1 0 0 0 1 1 1 ...
 $ PARTY_NINSHAD: num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_RLD     : num  0 0 0 0 0 0 1 0 0 0 ...
 $ PARTY_SBSP    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_SP      : num  0 0 0 0 0 0 0 0 0 0 ...
> library(corrplot)
> datamatrix<-cor(data1)
> corrplot(datamatrix, method ="number", type="upper")
>
```

Checking multicollinearity of the variable , these variables are correlated



4 LINER REGRESSION

We have splinted the data into train and test set with ratio of 70 and 30 , considered 40522 observers -ion of 23 variables for the train data .Test set we go 117 observations.

```
> #Preparing liner regression #
> set.seed(1234)
> pd<-sample(2,nrow(Elet),replace=TRUE, prob=c(0.7,0.3))
>
> train<-Elet[pd==1,]
> val<-Elet[pd==2,]
>
> sum(Elet$POSITION_2017)
[1] 40522
> sum(val$POSITION_2017)
[1] 12097
> sum(train$POSITION_2017)
[1] 28425
```

We have checked variable position and total votes of 2017 , there is coefficient in these two variables the Y is explaining significantly . The R squared values shown .4043 is sign of the variable is explained .40 , also the T values if extreme minus or plus I associated with high coefficients.

```
> Linear.1<-POSITION_2017 ~ VOTE_2017
> OLS.1<-lm(Linear.1, data=train.re)
> summary (OLS.1)

Call:
lm(formula = Linear.1, data = train.re)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6310 -2.6648 -0.6373  1.8395 19.8201

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.223e+00  6.759e-02  136.46   <2e-16 ***
VOTE_2017   -9.993e-05  1.984e-06  -50.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.632 on 3740 degrees of freedom
Multiple R-squared:  0.4043,    Adjusted R-squared:  0.4041
F-statistic:  2538 on 1 and 3740 DF,  p-value: < 2.2e-16
```

Let's do the Multi regression we will check with all variables and we found these are the variable having high significant

Short Name	Description	Short Name	Description
CAND_AGE	Age of Candidate	PARTY_BSP	BSP
VOTE_2017	Validated Vote of 2017	PATRY_INC	INC
CRIM_CASE	Criminal Record	PARTY_IND	IND
TOTA_ASST	Total Assts	PARTY_NINSHAD	NIBSHAD
CATEGORY	Cast	PARTY_RLD	RLD
PARTY_ADAL	ADAL	PARTY_SBSP	SBSP
PARTY_BJP	BJP	PARTY_SP	SP

```
~/My Files/R/R project files/
> OLS.full<-lm( POSITION_2017 ~., data=train.re)
> summary(OLS.full)

Call:
lm(formula = POSITION_2017 ~ ., data = train.re)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7385 -2.1243 -0.3399  1.4264 18.0183

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.062e+00  2.151e-01  37.475  < 2e-16 ***
CAND_AGE       3.173e-02  3.816e-03   8.315  < 2e-16 ***
VOTE_2017     -5.264e-05  5.072e-06 -10.378  < 2e-16 ***
CRIM_CASE     -1.384e-01  4.278e-02  -3.235 0.001226 **
TOTA_ASST     -2.407e-04  6.100e-05  -3.946  8.1e-05 ***
LIABILITES    -2.492e-09  7.158e-09  -0.348 0.727712
VOTE_2012     -1.271e-06  4.447e-06  -0.286 0.775071
POSITION_2012  1.003e-02  1.515e-02   0.662 0.507788
MALE_LITC     -3.305e+00  1.961e+00  -1.685 0.092061 .
FEMALE_LITC    4.811e+00  2.591e+00   1.857 0.063441 .
SEX            3.729e-01  1.909e-01   1.954 0.050830 .
CATEGORY       1.252e+00  1.273e-01   9.842  < 2e-16 ***
PARTY_ADAL    -4.654e+00  1.410e+00  -3.302 0.000971 ***
PARTY_BJP     -4.322e+00  4.905e-01  -8.811  < 2e-16 ***
PARTY_BSP     -4.904e+00  3.204e-01 -15.307  < 2e-16 ***
PATRY_INC     -4.919e+00  4.436e-01 -11.090  < 2e-16 ***
PARTY_IND     -3.207e-01  1.356e-01  -2.366 0.018047 *
PARTY_NINSHAD -4.698e+00  4.660e-01 -10.082  < 2e-16 ***
PARTY_RLD     -3.764e+00  2.517e-01 -14.953  < 2e-16 ***
PARTY_SBSP    -3.902e+00  1.535e+00  -2.541 0.011087 *
PARTY_SP      -4.826e+00  3.840e-01 -12.569  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.318 on 3721 degrees of freedom
Multiple R-squared:  0.5053,    Adjusted R-squared:  0.5026
```

Dropping the irrelevant variables from the liner set .

```
> Linear.f1<-POSITION_2017 ~ VOTE_2017 +CRIM_CASE+TOTA_ASST+SEX +CATEGORY+PARTY_ADAL+   PARTY_BJP+PARTY_BSP+PATRY_I
_NINSHAD+PARTY_RLD+PARTY_SBSP+  PARTY_SP
>
> OLS.f1<-lm(Linear.f1,data=train.re)
> summary(OLS.f1)

Call:
lm(formula = Linear.f1, data = train.re)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8580 -2.1378 -0.3491  1.4905 18.3998

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.245e+00  1.545e-01  59.821  < 2e-16 ***
VOTE_2017     -5.184e-05  5.002e-06 -10.364  < 2e-16 ***
CRIM_CASE     -1.390e-01  4.298e-02  -3.234  0.00123 **
TOTA_ASST     -3.476e-04  5.874e-05  -5.918 3.56e-09 ***
SEX            3.339e-01  1.922e-01   1.737  0.08241 .
CATEGORY       1.523e+00  1.217e-01  12.518  < 2e-16 ***
PARTY_ADAL    -4.361e+00  1.419e+00  -3.072  0.00214 **
PARTY_BJP     -3.910e+00  4.859e-01  -8.047 1.13e-15 ***
PARTY_BSP     -4.564e+00  3.151e-01 -14.484  < 2e-16 ***
PATRY_INC     -4.527e+00  4.418e-01 -10.248  < 2e-16 ***
PARTY_NINSHAD -4.484e+00  4.664e-01  -9.614  < 2e-16 ***
PARTY_RLD     -3.411e+00  2.450e-01 -13.923  < 2e-16 ***
PARTY_SBSP    -3.420e+00  1.547e+00  -2.211  0.02708 *
PARTY_SP      -4.387e+00  3.801e-01 -11.541  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.352 on 3728 degrees of freedom
Multiple R-squared:  0.4942,    Adjusted R-squared:  0.4924
F-statistic: 280.2 on 13 and 3728 DF,  p-value: < 2.2e-16
```

Further we checked VIF , we found Vote 2017 is having value more than 5 ,hence it is sign of multicollinearity , so we have to take it out .

```
> vif(OLS.f1)
    VOTE_2017     CRIM_CASE     TOTA_ASST      CATEGORY    PARTY_ADAL     PARTY_BJP     PARTY_BSP     PATRY_INC
     7.464122      1.053622      1.026680      1.034339      1.073898      5.317238      2.264330      1.409506
PARTY_NINSHAD     PARTY_RLD    PARTY_SBSP      PARTY_SP
     1.011338      1.030036      1.063026      2.696708
>
```

Preparing the final modules of liner regression and dropped the variables whose VIF is more than

```
> Linear.f1<-POSITION_2017 ~ CRIM_CASE+TOTA_ASST+CATEGORY+PARTY_ADAL+PARTY_BJP+PARTY_BSP+PATRY_INC+PARTY_NINSHAD+PARTY_RLD+
PARTY_SBSP+     PARTY_SP
> OLS.f1<-lm(Linear.f1,data=train.re)
> summary(OLS.f1)

Call:
lm(formula = Linear.f1, data = train.re)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4081 -2.1649 -0.3551  1.4898 18.7439

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.188e+00  1.551e-01  59.232  < 2e-16 ***
CRIM_CASE     -1.716e-01  4.348e-02  -3.947 8.05e-05 ***
TOTA_ASST     -3.421e-04  5.958e-05  -5.742 1.01e-08 ***
CATEGORY       1.540e+00  1.234e-01  12.481  < 2e-16 ***
PARTY_ADAL    -8.215e+00  1.390e+00  -5.909 3.74e-09 ***
PARTY_BJP     -8.419e+00  2.175e-01 -38.709  < 2e-16 ***
PARTY_BSP     -6.970e+00  2.174e-01 -32.063  < 2e-16 ***
PATRY_INC     -6.953e+00  3.798e-01 -18.308  < 2e-16 ***
PARTY_NINSHAD -4.727e+00  4.724e-01 -10.005  < 2e-16 ***
PARTY_RLD     -3.573e+00  2.480e-01 -14.409  < 2e-16 ***
PARTY_SBSP    -7.220e+00  1.523e+00  -4.740 2.22e-06 ***
PARTY_SP      -7.471e+00  2.388e-01 -31.286  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.4 on 3730 degrees of freedom
Multiple R-squared:  0.4793,    Adjusted R-squared:  0.4778
F-statistic: 312.1 on 11 and 3730 DF,  p-value: < 2.2e-16

> vif(OLS.f1)
    CRIM_CASE     TOTA_ASST      CATEGORY    PARTY_ADAL     PARTY_BJP     PARTY_BSP     PATRY_INC PARTY_NINSHAD
     1.048706      1.026568      1.034015      1.001306      1.035562      1.048340      1.012563      1.008711
    PARTY_RLD    PARTY_SBSP      PARTY_SP
     1.025792      1.002134      1.034345
>
```

We have predicted the fit lines and understand these are the major factors helped the candidates to win the election in 2017

```
> pred.re<-predict(OLS.f1,newdata=val.re, interval="predict")
> pred.re
```

```
>
> head(pred.re)
        fit       lwr       upr
5  8.2343969  1.563957 14.904837
14 9.7744661  3.104978 16.443954
16 2.1017913 -4.621008  8.824590
26 8.8723814  2.201695 15.543068
28 0.9657242 -5.717058  7.648506
29 2.8045451 -3.875577  9.484667
> head(val.re)
   CAND_AGE VOTE_2017 POSITION_2017 CRIM_CASE TOTA_ASST LIABILITES VOTE_2012 POSITION_2012 MALE_LITC FEMALE_LITC SEX
5         0      1576             5         0      2789          0         0             0    0.0000       0.0000   0
14       57       319            14         0      2789          0         0             0    0.0000       0.0000   0
16       46     90318             2         6      1882          0     84623             2    0.7977       0.6330   0
26       37       140            12         0       924          0       588             8    0.7832       0.5916   0
28       47    122574             2         3      2422     150000         0             0    0.0000       0.0000   0
29       47     17350             3         0      2789          0         0             0    0.0000       0.0000   0
   CATEGORY PARTY_ADAL PARTY_BJP PARTY_BSP PATRY_INC PARTY_IND PARTY_NINSHAD PARTY_RLD PARTY_SBSP PARTY_SP
5         0          0         0         0         0         0             0         0          0        0
14        1          0         0         0         0         1             0         0          0        0
16        1          0         0         0         1         0             0         0          0        0
26        0          0         0         0         0         1             0         0          0        0
28        1          0         1         0         0         0             0         0          0        0
29        1          0         0         1         0         0             0         0          0        0
> head(train.re)
  CAND_AGE VOTE_2017 POSITION_2017 CRIM_CASE TOTA_ASST LIABILITES VOTE_2012 POSITION_2012 MALE_LITC FEMALE_LITC SEX
1       53     97035             1         0       104     398000         0             0    0.0000       0.0000   0
2       55     71449             2         0      2133    2985082     70274             1    0.7977       0.6330   0
3       52     71019             3         0       411   10214938         0             0    0.0000       0.0000   0
4       49      4187             4         0      2315     286731         0             0    0.0000       0.0000   0
6       37      1255             6         0      2359          0         0             0    0.0000       0.0000   0
```

5 LOGISTIC REGRESSION

For logistic regression we will be used all these variables, the POSITION _2017 variables values we overwritten with 1 or 0 , value 1 stands for winning candidates , value 0 is stands for other winning candidates .

```
> train.re$POSITION_2017<-ifelse(train.re$POSITION_2017=="1",1,0)
>
```

```
> train.lg<-train.re [,-c(22)]
> val.lg<-val.re [,-c(22)]
> str(train.lg)
'data.frame':   3742 obs. of  21 variables:
 $ CAND_AGE     : int  53 55 52 49 37 38 39 32 35 45 ...
 $ VOTE_2017    : int  97035 71449 71019 4187 1255 1150 1113 945 810 659 ...
 $ POSITION_2017: num  1 0 0 0 0 0 0 0 0 0 ...
 $ CRIM_CASE    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ TOTA_ASST    : num  104 2133 411 2315 2359 ...
 $ LIABILITES   : num  398000 2985082 10214938 286731 0 ...
 $ VOTE_2012    : int  0 70274 0 0 0 12903 0 0 418 0 ...
 $ POSITION_2012: int  0 1 0 0 0 4 0 0 19 0 ...
 $ MALE_LITC    : num  0 0.798 0 0 0 ...
 $ FEMALE_LITC  : num  0 0.633 0 0 0 ...
 $ SEX          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CATEGORY     : num  1 1 1 1 1 1 0 1 0 1 ...
 $ PARTY_ADAL   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_BJP    : num  0 1 0 0 0 0 0 0 0 0 ...
 $ PARTY_BSP    : num  0 0 1 0 0 0 0 0 0 0 ...
 $ PATRY_INC    : num  1 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_IND    : num  0 0 0 1 0 0 1 1 1 1 ...
 $ PARTY_NINSHAD: num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_RLD    : num  0 0 0 0 0 1 0 0 0 0 ...
 $ PARTY_SBSP   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_SP     : num  0 0 0 0 0 0 0 0 0 0 ...
>
```

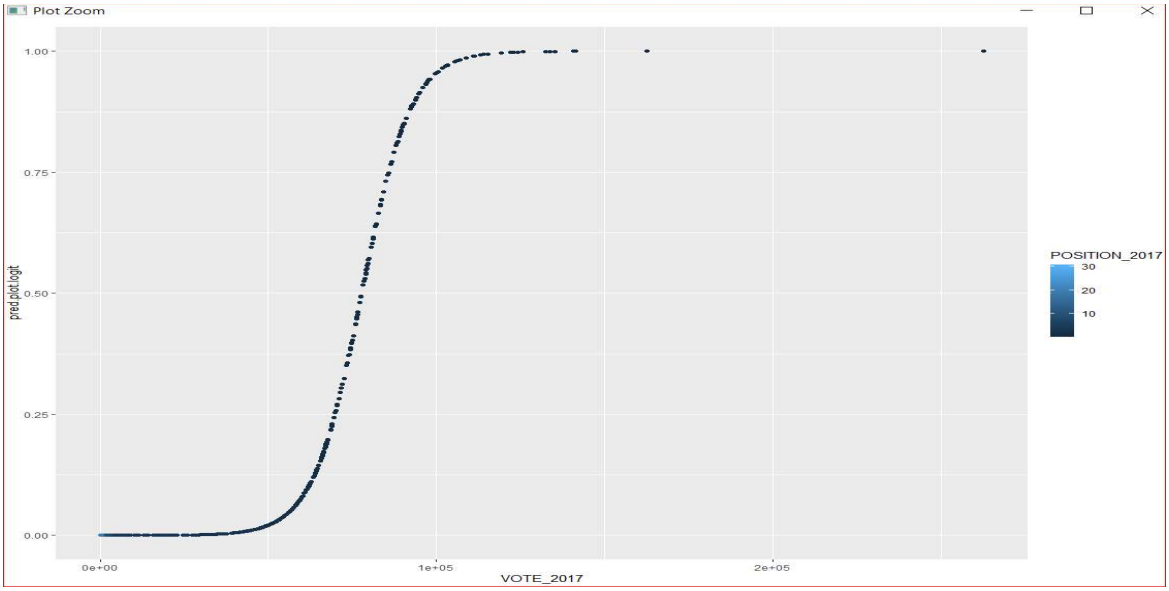
Checking bionomical outputs of variable POSITON _2017 and VOTE_2017

```
>
> # Fit the Sigmoid function
> Logit.1<-POSITION_2017  ~ VOTE_2017
> logit.plot<-glm(Logit.1, data=train.lg, family=binomial())
> summary(logit.plot)

Call:
glm(formula = Logit.1, family = binomial(), data = train.lg)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4278  -0.0076  -0.0070  -0.0068   2.6993

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.074e+01  7.444e-01  -14.43   <2e-16 ***
VOTE_2017    1.381e-04  9.799e-06   14.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1990.35  on 3741  degrees of freedom
Residual deviance:  468.44  on 3740  degrees of freedom
AIC: 472.44

Number of Fisher Scoring iterations: 9
```

Plotting VOTE_2017 and POSITION_2017

```
~/My Files/R/R project files/
> library(ggplot2)
> pred.plot.logit <- predict.glm(logit.plot, newdata=val.lg, type="response")
> qplot( VOTE_2017, pred.plot.logit,data=val.lg, color=POSITION_2017 )
>
```

Here x=1 is the line which separates the given y=0 and y=1 ion a logistic function. It is because y=0 and y=1 makes the perpendicular line and x=1 is the horizontal line which cuts the perpendicular line and separates the line. The logistic function is a sigmoid curve



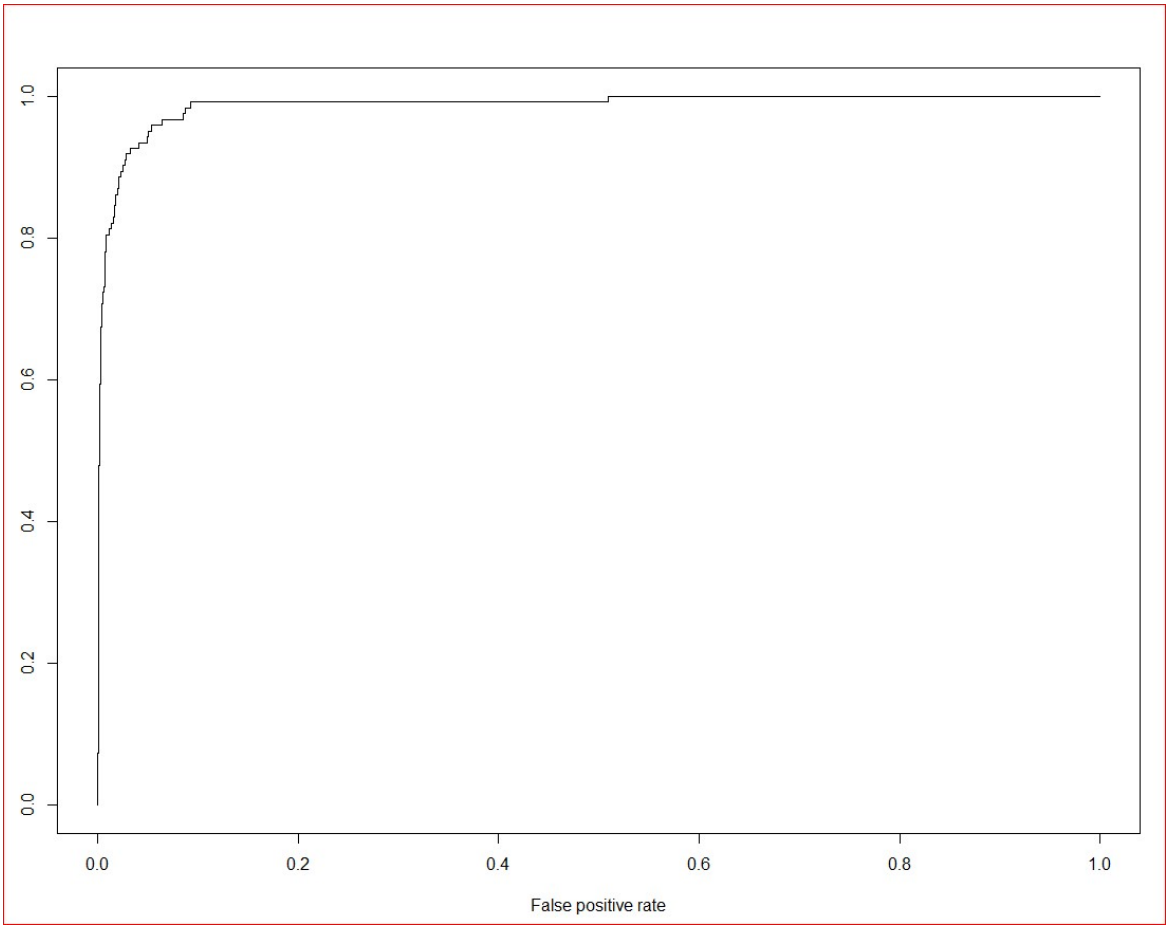
We got only few variables are explaining here Y (VOTE_2017,MALE_LITC,FEMALE_LITC,CRIM_CASE)

```
> #Using all variable#
> Logit_md = glm(POSITION_2017 ~ ., data= train.lg, family = binomial)
> summary(Logit_md)

Call:
glm(formula = POSITION_2017 ~ ., family = binomial, data = train.lg)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.95076  -0.01844  -0.00007  -0.00002   2.87567

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.164e+01  7.127e+02  -0.030   0.9758
CAND_AGE      -1.881e-02  1.395e-02  -1.348   0.1777
VOTE_2017      1.266e-04  1.145e-05  11.057   <2e-16 ***
CRIM_CASE      1.089e-01  6.136e-02   1.775   0.0759 .
TOTA_ASST      2.273e-04  1.621e-04   1.402   0.1608
LIABILITES     6.957e-09  1.048e-08   0.664   0.5067
VOTE_2012     -7.259e-06  8.203e-06  -0.885   0.3762
POSITION_2012 -3.535e-02  4.967e-02  -0.712   0.4766
MALE_LITC      1.061e+01  5.264e+00   2.016   0.0438 *
FEMALE_LITC   -1.320e+01  7.049e+00  -1.872   0.0612 .
SEX           -8.113e-01  4.960e-01  -1.636   0.1019
CATEGORY      -3.266e-01  3.589e-01  -0.910   0.3629
PARTY_ADAL     1.476e+01  7.127e+02   0.021   0.9835
PARTY_BJP      1.387e+01  7.127e+02   0.019   0.9845
PARTY_BSP      1.077e+01  7.127e+02   0.015   0.9879
PATRY_INC      1.134e+01  7.127e+02   0.016   0.9873
PARTY_IND      1.315e+01  7.127e+02   0.018   0.9853
PARTY_NINSHAD -2.118e+00  3.434e+03  -0.001   0.9995
PARTY_RLD      1.382e+01  7.127e+02   0.019   0.9845
PARTY_SBSP     1.174e+01  7.127e+02   0.016   0.9869
```

Seems the model is not great ,it may be overfitting , we got false classifications = 1425+99

Hence, we may end with the conclusion that module is performing 98%, we cannot process further its overfilled module .

```
> Logit_pd = predict(Logit_md,newdata = test.lg,type = "response")
> table(test.lg$POSITION_2017,Logit_pd>0.5)

    FALSE TRUE
  0  1425   17
  1    24   99
> (1412+99)/nrow(na.omit(test.lg))
[1] 0.9654952
>
> library(ROCR)
> ROCRpred = prediction(Logit_pd, test.lg$POSITION_2017)
> as.numeric(performance(ROCRpred, "auc")@y.values)
[1] 0.9875117
> perf = performance(ROCRpred, "tpr","fpr")
> plot(perf)
>
```



6 PREPARING K-NEAREST NEIGHBORS

We have prepared the data set for KNN Modeling with splitting some dummy variable of education

It is 26 variable with 5307 variables

```
> #KNN Modeling #
> data2 <- Elet [,-c(1:5,7,11)]
> data2$POSITION_2017<-ifelse(data2$POSITION_2017=="1",1,0)
> str(data2)
'data.frame':   5307 obs. of  26 variables:
 $ CAND_AGE      : int  53 55 52 49 0 37 38 39 32 35 ...
 $ VOTE_2017     : int  97035 71449 71019 4187 1576 1255 1150 1113 945 810 ...
 $ POSITION_2017 : num  1 0 0 0 0 0 0 0 0 0 ...
 $ CRIM_CASE     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ TOTA_ASST     : num  104 2133 411 2315 2789 ...
 $ LIABILITES    : num  398000 2985082 10214938 286731 0 ...
 $ VOTE_2012     : int  0 70274 0 0 0 12903 0 0 418 ...
 $ POSITION_2012 : int  0 1 0 0 0 0 4 0 0 19 ...
 $ MALE_LITC     : num  0 0.798 0 0 0 ...
 $ FEMALE_LITC   : num  0 0.633 0 0 0 ...
 $ SEX           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CATEGORY      : num  1 1 1 1 0 1 1 0 1 0 ...
 $ PARTY_ADAL    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_BJP     : num  0 1 0 0 0 0 0 0 0 0 ...
 $ PARTY_BSP     : num  0 0 1 0 0 0 0 0 0 0 ...
 $ PATRY_INC     : num  1 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_IND     : num  0 0 0 1 0 0 0 1 1 1 ...
 $ PARTY_NINSHAD : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_RLD     : num  0 0 0 0 0 0 1 0 0 0 ...
 $ PARTY_SBSP    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_SP      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ EDU_HighSchool: num  0 0 0 0 0 0 0 0 0 0 ...
 $ EDU_UG        : num  0 0 0 1 0 0 0 0 0 0 ...
 $ EDU_PG        : num  1 0 0 0 0 0 0 0 0 0 ...
 $ EDU_Graduate  : num  0 1 0 0 0 0 0 0 0 0 ...
 $ EDU_Other     : num  0 0 0 0 0 0 0 0 0 0 ...
>
> library(caTools)
> set.seed(110)
> spl = sample.split(data2$POSITION_2017, SplitRatio = 0.7 )
> train_knn = subset(data2, spl == T)
> test_knn = subset(data2, spl == F)
> dim(train_knn)
[1] 3715   26
```

Pitching data set for train and test

```
> library(caTools)
> set.seed(110)
> spl = sample.split(data2$POSITION_2017, SplitRatio = 0.7 )
> train_knn = subset(data2, spl == T)
> test_knn = subset(data2, spl == F)
> dim(train_knn)
[1] 3715   26
> dim(test_knn)
[1] 1592   26
>
```

Scaling the data set for analyzation

```
> scale(train_knn)
       CAND_AGE    VOTE_2017 POSITION_2017  CRIM_CASE     TOTA_ASST    LIABILITES   VOTE_2012 POSITION_2012
1     0.70811360  2.633619407      3.488622 -0.2357693 -1.7688902976 -0.1112592547 -0.32781526   -0.50809556
3     0.64580121  1.779855723     -0.286569 -0.2357693 -1.4410566573  1.2574028436 -0.32781526   -0.50809556
4     0.45886405 -0.413361318     -0.286569 -0.2357693  0.5921526298 -0.1267722042 -0.32781526   -0.50809556
5    -2.59444288 -0.499046165     -0.286569 -0.2357693  1.0983192275 -0.1667477904 -0.32781526   -0.50809556
6    -0.28888458 -0.509580380     -0.286569 -0.2357693  0.6391385587 -0.1667477904 -0.32781526   -0.50809556
7    -0.22657220 -0.513026152     -0.286569 -0.2357693  0.3711051915 -0.1667477904  0.44249729    0.22112020
10   -0.41350936 -0.524183888     -0.286569 -0.2357693 -0.8601397181 -0.1654930246 -0.30286055    2.95567932
11    0.20961451 -0.529139235     -0.286569 -0.2357693  1.0983192275 -0.1667477904 -0.32781526   -0.50809556
14    0.95736314 -0.540296971     -0.286569 -0.2357693  1.0983192275 -0.1667477904 -0.32781526   -0.50809556
15    1.14430030  2.546326534      3.488622 -0.2357693  0.5889490437  1.7568083093 -0.32781526   -0.50809556
16    0.27192689  2.413188493     -0.286569  4.1182463  0.1297683749 -0.1667477904  4.72420047   -0.14348768
17   -0.22657220  1.593094917     -0.286569 -0.2357693  1.0983192275 -0.1667477904 -0.32781526   -0.50809556
```

Checking K value with -3 (Value of K can be chosen basis the sqrt of n)

```
~/My Files/R/R project files/
> library(class)
> pred = knn(train_knn[-1], test_knn[-1], train_knn[,1], k = 3)
> table.knn = table(test_knn[,1], pred)
>
> accuracy.knn = sum(diag(table.knn))/sum(table.knn)
> accuracy.knn
[1] 0.08668342
> loss.knn = table.knn[2,1]/(table.knn[2,1] + table.knn[1,1])
> loss.knn
[1] 0.01136364
>
```

K value of 19

```
> library(class)
> pred = knn(train_knn[-1], test_knn[-1], train_knn[,1], k = 19)
> table.knn19 = table(test_knn[,1], pred)
>
> accuracy.knn19 = sum(diag(table.knn19))/sum(table.knn19)
> accuracy.knn19
[1] 0.1086683
> loss.knn = table.knn19[2,1]/(table.knn19[2,1] + table.knn19[1,1])
> loss.knn
[1] 0.007575758
```

K value of 2

```
> library(class)
> pred = knn(train_knn[-1], test_knn[-1], train_knn[,1], k =2)
> table.knn2 = table(test_knn[,1], pred)
>
> accuracy.knn2 = sum(diag(table.knn2))/sum(table.knn2)
> accuracy.knn2
[1] 0.06595477
> loss.knn = table.knn2[2,1]/(table.knn2[2,1] + table.knn2[1,1])
> loss.knn
[1] 0.01538462
>
```

7 PREPARING NAÏVE BAYES CLASSIFICATION

Extracting data set for test and train with ration of 70 and 30

```
> #Preapring NB module#
> set.seed(119)
> spx = sample.split(data2$POSITION_2017, SplitRatio = 0.7 )
> train_nb = subset(data2, spx == T)
> test_nb = subset(data2, spx == F)
> dim(train_nb)
[1] 3715   26
> dim(test_nb)
[1] 1592   26
>
```

A training model is created by using naïve Bayes function. The model is used to predict the survival status of a random sample

```
> library(e1071)
> NB = naiveBayes(x = train_nb[-1], y = train_nb$POSITION_2017)
> ypred.NB = predict(NB, newdata = test_nb[-1])
> ypred.NB
   [1] 1 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
  [45] 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
  [89] 0 1 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0
 [133] 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 0 0 0 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0
 [177] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1
 [221] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 1
 [265] 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0
 [309] 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
 [353] 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
 [397] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
 [441] 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [485] 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0
 [529] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
 [573] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0
 [617] 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0
 [661] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

a vector containing outcomes for each sample. 'nb' is a string specifying that the classification model is naïve Bayes

```
> tab.NB = table(test_nb[,1], ypred.NB)
> tab.NB
    ypred.NB
       0   1
  0  112   0
  25  26   0
  26  20   1
  27  26   1
  28  22   0
  29  19   0
  30  37   2
  31  29   0
  32  35   3
  33  28   3
  34  40   3
  35  33   2
  36  47   5
  37  33   2
  38  39   2
  39  41   5
  40  54   9
  41  35   6
  42  37  12
```

Accuracy of NB

```
> accuracy.NB = sum(diag(tab.NB))/sum(tab.NB)
> accuracy.NB
[1] 0.07035176
>
```

8 SOURCE CODE

Project PM.R

```
> #***Importing the election data of 2017***#
> setwd ("~/My Files/Great Lakes Projects/Project -4")
> library(readxl)
Warning message:
package 'readxl' was built under R version 3.5.3
> el_data17 <- read_excel("C:/Users/SuprasannaPradhan/Documents/My Files//Great Lakes Projects/Pr
oject -4/LA_2017.xls",)
> #Sorting only Utter Pradesh for 2017#
> data17<-as.data.frame(subset(el_data17, ST_CODE == "S24"))
> write.csv(data17, file = "data17.csv")
> View(data17)
> str(data17)
'data.frame':   5307 obs. of  15 variables:
 $ ST_CODE             : chr  "S24" "S24" "S24" "S24" ...
 $ ST_NAME             : chr  "Uttar Pradesh" "Uttar Pradesh" "Uttar Pradesh" "Uttar Pradesh"
 ...
 $ MONTH               : num  3 3 3 3 3 3 3 3 3 ...
 $ YEAR                : num  2017 2017 2017 2017 2017 ...
 $ DIST_NAME           : chr  "Saharanpur" "Saharanpur" "Saharanpur" "Saharanpur" ...
 $ AC_NO               : num  1 1 1 1 1 1 1 1 1 ...
 $ AC_NAME             : chr  "Behat" "Behat" "Behat" "Behat" ...
 $ AC_TYPE             : chr  "GEN" "GEN" "GEN" "GEN" ...
 $ CAND_NAME           : chr  "NARESH SAINI" "MAHAVEER SINGH RANA" "MOHD. IQBAL" "RANA ADITYA PR
ATAP SINGH" ...
 $ CAND_SEX            : chr  "M" "M" "M" "M" ...
 $ CAND_CATEGORY       : chr  "GEN" "GEN" "GEN" "GEN" ...
 $ CAND_AGE            : chr  "53" "55" "52" "49" ...
 $ PARTYABBRE          : chr  "INC" "BJP" "BSP" "IND" ...
 $ TOTALVALIDVOTESPOLLED: num  97035 71449 71019 4187 1576 ...
 $ POSITION            : num  1 2 3 4 5 6 7 8 9 10 ...
> #*** Importing data myneta.com***#
```

```
> library(installr)
Loading required package: stringr

Welcome to installr version 0.21.0

More information is available on the installr project website:
https://github.com/talgalili/installr/

Contact: <tal.galili@gmail.com>
Suggestions and bug-reports can be submitted at: https://github.com/talgalili/installr/issues

                      To suppress this message use:
                      suppressPackageStartupMessages(library(installr))

Warning messages:
1: package 'installr' was built under R version 3.5.3
2: package 'stringr' was built under R version 3.5.3
> library(RCurl)
Loading required package: bitops
Warning messages:
1: package 'RCurl' was built under R version 3.5.2
2: package 'bitops' was built under R version 3.5.2
> library(XML)
Warning message:
package 'XML' was built under R version 3.5.2
> library(stringr)
> xml.url<- "http://www.myneta.info/uttarpradesh2017/index.php?action=summary&subAction=candidate
s_analyzed&sort=candidate#summary"
> nettable <- readHTMLTable(xml.url,which = 3)
> View(nettable)
> str(nettable)
'data.frame':   4829 obs. of  8 variables:
```

```
> #taken out the Sl No#
> neta1 <- subset(nettable, select = -c(1))
> str(neta1)
'data.frame':   4829 obs. of  7 variables:
 $ Candidateâˆ‡  : Factor w/ 3877 levels "A Hasiv","A Wahid",..: NA 587 1 2 3 4 5 6 7 8 ...
 $ Constituency  : Factor w/ 404 levels "AGRA CANTT. (SC)",..: NA 106 20 133 150 276 207 59 316 3
2 ...
 $ Party         : Factor w/ 313 levels "Aadhi Aabadi Party",..: NA 202 99 124 290 127 154 124 12
 71 ...
 $ Criminal Case : Factor w/ 23 levels "0","1","10","11",..: NA 23 1 1 1 1 1 1 2 1 ...
 $ Education     : Factor w/ 13 levels "10th Pass","12th Pass",..: NA 6 2 1 7 9 10 11 2 10 ...
 $ Total Assets  : Factor w/ 4275 levels "Nil","Rs 1,00,00,000 ~ 1 Crore+",..: NA 4275 2366 3823
848 1713 2521 910 1855 3735 ...
 $ Liabilities   : Factor w/ 1151 levels "Liabilities",..: NA 1 886 2 2 2 2 2 2 2 ...
> #Removing the sepcial characters#
> neta <-data.frame(neta1)
> neta$Total.Assets <- gsub("Rs","",neta$Total.Assets)
> neta$new_totalas <- gsub("~.*", "", c(neta$Total.Assets))
> neta$Liabilities <- gsub("Rs","",neta$Liabilities)
> neta$new_liabilities <-gsub("~.*", "", c(neta$Liabilities))
> write.csv(neta, file = "neta_backup.csv")
> neta <- read.csv("C:/Users/SuprasannaPradhan/Documents/My Files//Great Lakes Projects/Project -
4/neta.csv",)
> View(neta)
>
```

```
~/My Files/Great Lakes Projects/Project -4/
> View(neta)
> #Removing the sepcial characters#
> neta <-data.frame(neta1)
> neta$Total.Assets <- gsub("Rs","",neta$Total.Assets)
> neta$new_totalas <- gsub("~.*", "", c(neta$Total.Assets))
> neta$Liabilities <- gsub("Rs","",neta$Liabilities)
> neta$new_liabilities <-gsub("~.*", "", c(neta$Liabilities))
> write.csv(neta, file = "neta_backup.csv")
> neta <- read.csv("C:/Users/SuprasannaPradhan/Documents/My Files//Great Lakes Projects/Project -
4/neta.csv",)
> View(neta)
> str(neta)
'data.frame':   4829 obs. of  9 variables:
 $ Candidateâ..    : Factor w/ 3877 levels "A Hasiv","A Wahid",..: NA 587 1 2 3 4 5 6 7 8 ...
 $ Constituency.  : Factor w/ 404 levels "AGRA CANTT. (SC)",..: NA 106 20 133 150 276 207 59 316
 392 ...
 $ Party.         : Factor w/ 313 levels "Aadhi Aabadi Party",..: NA 202 99 124 290 127 154 124 1
24 71 ...
 $ Criminal.Case. : Factor w/ 23 levels "0","1","10","11",..: NA 23 1 1 1 1 1 1 2 1 ...
 $ Education.     : Factor w/ 13 levels "10th Pass","12th Pass",..: NA 6 2 1 7 9 10 11 2 10 ...
 $ Total.Assets   : Factor w/ 4275 levels " 1,00,00,000 ~ 1 Crore+",..: NA 4275 2365 3822 2847 17
12 2520 909 1854 3734 ...
 $ Liabilities    : Factor w/ 1151 levels " 0 ~"," 1,00,00,000 ~ 1 Crore+",..: NA 1151 885 1 1 1
 1 1 1 1 ...
 $ new_totalas    : Factor w/ 4275 levels " 1000 "," 10000 ",..: NA 4275 2619 3761 2691 1296 2382
 534 1610 3841 ...
 $ new_liabilities: Factor w/ 1151 levels " 0 "," 100 ",..: NA 1151 876 1 1 1 1 1 1 1 ...
> neta_data <-as.data.frame(subset(neta, select = -c(6:7)))
> View(neta_data)
```

```
~/My Files/Great Lakes Projects/Project -4/
> # Added and Renaming the colnames#
> #as.data.frame(gsub("[[:punct:]]", "", as.matrix(neta_data)))
> colnames(neta_data)[colnames(neta_data)=="Candidate"] <- "CAND_NAME"
> names(neta_data)
[1] "Candidateâ.."    "Constituency."   "Party."          "Criminal.Case."  "Education."
[6] "new_totalas"     "new_liabilities"
> names(neta_data)[1]<-"CAND_NAME"
> names(neta_data)[2]<-"AC_NAME"
> names(neta_data)[3]<-"PARTYABBRE"
> names(neta_data)[4]<-"CRIM_CASE"
> names(neta_data)[5]<-"EDUCATION"
> names(neta_data)[6]<-"TOTA_ASST"
> names(neta_data)[7]<-"LIABILITES"
> names(neta_data)
[1] "CAND_NAME"  "AC_NAME"    "PARTYABBRE" "CRIM_CASE"  "EDUCATION"  "TOTA_ASST"  "LIABILITES"
> summary(neta_data)
       CAND_NAME              AC_NAME       PARTYABBRE     CRIM_CASE
 Manoj Kumar :  19   AGRA SOUTH     :  26   IND    :1454   0      :3965
 Ajay Kumar  :  18   ALLAHABAD NORTH:  26   BSP    : 400   1      : 440
 Anil Kumar  :  18   AMETHI         :  24   BJP    : 384   2      : 161
 Rakesh Kumar:  15   VARANASI CANTT.:  24   SP     : 308   3      :  92
 Rajesh Kumar:  14   BHADOHI        :  23   RLD    : 276   4      :  54
 (Other)     :4744   (Other)        :4705   (Other):2006   (Other): 116
 NA's        :   1   NA's           :   1   NA's   :   1   NA's   :   1
       EDUCATION       TOTA_ASST        LIABILITES
 Graduate     : 905    20000  :  16    0       :3172
 Post Graduate: 862    100000 :  14    500000  :  37
 12th Pass    : 830    Nil    :  13    100000  :  27
 10th Pass    : 558    30000  :  11    200000  :  26
 8th Pass     : 479    50000  :  11    1000000 :  24
```

```
 NA's        :   1   NA's    :   1   NA's     :   1
> View(neta_data)
> #Merging the 2017 and my neta dat set #
> library(dplyr)
> neta_data <- neta_data %>% mutate_each(funs(toupper),CAND_NAME )
> neta_data <- neta_data %>% mutate_each(funs(tolower),AC_NAME )
> str(neta_data)
'data.frame':   4829 obs. of  7 variables:
 $ CAND_NAME : chr  NA "CANDIDATE<U+2207>" "A HASIV" "A WAHID" ...
 $ AC_NAME   : chr  NA "constituency" "arya nagar" "gainsari" ...
 $ PARTYABBRE: Factor w/ 313 levels "Aadhi Aabadi Party",..: NA 202 99 124 290 127 154 124 124
 ...
 $ CRIM_CASE : Factor w/ 23 levels "0","1","10","11",..: NA 23 1 1 1 1 1 1 2 1 ...
 $ EDUCATION : Factor w/ 13 levels "10th Pass","12th Pass",..: NA 6 2 1 7 9 10 11 2 10 ...
 $ TOTA_ASST : Factor w/ 4275 levels " 1000 "," 10000 ",..: NA 4275 2619 3761 2691 1296 2382 53
 1610 3841 ...
 $ LIABILITES: Factor w/ 1151 levels " 0 "," 100 ",..: NA 1151 876 1 1 1 1 1 1 1 ...
>
```

```
> mer_data17<- merge(x = data17,y = neta_data,by=c('CAND_NAME'),all.x=TRUE)
> final_data17 <- subset(mer_data17, select = -c(16:17))
> write.csv(final_data17, file = "final_data17bk.csv")
> final_data17<- read.csv("C:/Users/SuprasannaPradhan/Documents/My Files//Great Lakes Projects/Pr
oject -4/final_data17.csv",)
> View(final_data17)
> str(final_data17)
'data.frame':   5307 obs. of  20 variables:
```

```
9 2108 1028 415 955 ...
 $ LIABILITES          : num  398000 2985082 10214938 286731 0 ...
> #***Importing the election data of 2012***#
> el_data12 <- read_excel("C:/Users/SuprasannaPradhan/Documents/My Files//Great Lakes Proj
oject -4/LA_2012.xls",)
> data12<-(subset(el_data12, ST_CODE == "S24"))
> names(data12)
 [1] "ST_CODE"       "ST_NAME"       "MONTH"         "YEAR"          "DIST_NAME"
 [6] "AC_NO"         "AC_NAME"       "AC_TYPE"       "CAND_NAME"     "CAND_SEX"
[11] "CAND_CATEGORY" "CAND_AGE"      "PARTYABBRE"    "TOTVOTPOLL"    "POSITION"
> #colnames(data12)[colnames(data12)=="ST_NAME"] <- "ST_NAME2012"
> View(data12)
> str(data12)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       7031 obs. of  15 variables:
 $ ST_CODE      : chr  "S24" "S24" "S24" "S24" ...
 $ ST_NAME      : chr  "Uttar Pradesh" "Uttar Pradesh" "Uttar Pradesh" "Uttar Pradesh" ...
 $ MONTH        : num  1 1 1 1 1 1 1 1 1 1 ...
 $ YEAR         : num  2012 2012 2012 2012 2012 ...
 $ DIST_NAME    : chr  "Saharanpur" "Saharanpur" "Saharanpur" "Saharanpur" ...
 $ AC_NO        : num  1 1 1 1 1 1 1 1 1 1 ...
 $ AC_NAME      : chr  "Behat" "Behat" "Behat" "Behat" ...
 $ AC_TYPE      : chr  "GEN" "GEN" "GEN" "GEN" ...
 $ CAND_NAME    : chr  "MAHAVEER SINGH RANA" "NARESH" "UMAR ALI KHAN" "AJAY CHAUHAN" ...
 $ CAND_SEX     : chr  "M" "M" "M" "M" ...
 $ CAND_CATEGORY: chr  "GEN" "GEN" "GEN" "GEN" ...
 $ CAND_AGE     : num  50 47 34 29 39 39 31 35 28 35 ...
 $ PARTYABBRE   : chr  "BSP" "INC" "SP" "BJP" ...
 $ TOTVOTPOLL   : num  70274 69760 47366 23623 2251 ...
 $ POSITION     : num  1 2 3 4 5 6 7 8 9 10 ...
```

```
> #Preparing the final subset fo the data and #Creting dummy variable# #
> Elet <- subset(final_data, select = -c(1,2,3,4,5,7,9))
> Elet$TOTA_ASST <- as.numeric(Elet$TOTA_ASST)
> Elet$SEX<-ifelse(Elet$CAND_SEX =="F",1,0)
> Elet$CATEGORY <-ifelse(Elet$CAND_CATEGORY =="GEN",1,0)
> #Cosierding only elected candidate#
> #Elet1<-as.data.frame(subset(Elet,POSITION.x  == "1"))
> library(dplyr)
> #Created some more dummy varibales#
> Elet$PARTY_ADAL<-ifelse(Elet$PARTYABBRE=="ADAL",1,0)
> Elet$PARTY_BJP<-ifelse(Elet$PARTYABBRE=="BJP",1,0)
> Elet$PARTY_BSP<-ifelse(Elet$PARTYABBRE=="BSP",1,0)
> Elet$PATRY_INC<-ifelse(Elet$PARTYABBRE=="INC",1,0)
> Elet$PARTY_IND<-ifelse(Elet$PARTYABBRE=="IND",1,0)
> Elet$PARTY_NINSHAD<-ifelse(Elet$PARTYABBRE=="NINSHAD",1,0)
> Elet$PARTY_RLD<-ifelse(Elet$PARTYABBRE=="RLD",1,0)
> Elet$PARTY_SBSP<-ifelse(Elet$PARTYABBRE=="SBSP",1,0)
> Elet$PARTY_SP<-ifelse(Elet$PARTYABBRE=="SP",1,0)
```

```
~/My Files/Great Lakes Projects/Project -4/
> #install.packages(c("SDMTools","pROC", "Hmisc"))
> library(SDMTools)
> library(pROC)
> library(Hmisc)
> Elet$EDU_HighSchool<-ifelse(Elet$EDUCATION  =="10th Pass",1,0)
> Elet$EDU_UG<-ifelse(Elet$EDUCATION  =="12th Pass",1,0)
> Elet$EDU_HighSchool<-ifelse(Elet$EDUCATION  =="5th Pass",1,0)
> Elet$EDU_HighSchool<-ifelse(Elet$EDUCATION  =="8th Pass",1,0)
> Elet$EDU_PG<-ifelse(Elet$EDUCATION  =="Doctorate",1,0)
> Elet$EDU_Graduate<-ifelse(Elet$EDUCATION  =="Graduate",1,0)
> Elet$EDU_Graduate <-ifelse(Elet$EDUCATION  =="Graduate Professional",1,0)
> Elet$EDU_PG<-ifelse(Elet$EDUCATION  =="Post Graduate",1,0)
> Elet$EDU_Other <-ifelse(Elet$EDUCATION  =="Literate",1,0)
> Elet$EDU_Other <-ifelse(Elet$EDUCATION  =="Illiterate",1,0)
> Elet$EDU_Other <-ifelse(Elet$EDUCATION  =="Others",1,0)
> Elet$EDU_Other <-ifelse(Elet$EDUCATION  =="Not Given",1,0)
> data2 <- Elet [,-c(1:5,7,11)]
> str(data2)
'data.frame':   5307 obs. of  26 variables:
 $ CAND_AGE      : int  53 55 52 49 0 37 38 39 32 35 ...
 $ VOTE_2017     : int  97035 71449 71019 4187 1576 1255 1150 1113 945 810 ...
 $ POSITION_2017 : int  1 2 3 4 5 6 7 8 9 10 ...
 $ CRIM_CASE     : num  0 0 0 0 0 0 0 0 0 0
```

```
  2.254199e-01   4.341961e-02   2.264853e-01   1.147375e-01
> OLS.full<-lm( POSITION_2017 ~., data=train.re)
> summary(OLS.full)

Call:
lm(formula = POSITION_2017 ~ ., data = train.re)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9865 -2.1497 -0.3274  1.4733 18.6768

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.062e+00  2.174e-01  37.087  < 2e-16 ***
CAND_AGE       3.374e-02  3.809e-03   8.859  < 2e-16 ***
VOTE_2017     -4.939e-05  5.122e-06  -9.643  < 2e-16 ***
CRIM_CASE     -1.678e-01  4.881e-02  -3.438 0.000593 ***
TOTA_ASST     -2.766e-04  6.226e-05  -4.443 9.12e-06 ***
LIABILITES    -3.196e-08  1.015e-08  -3.148 0.001660 **
VOTE_2012     -2.262e-06  4.549e-06  -0.497 0.619106
POSITION_2012  1.113e-02  1.540e-02   0.723 0.469865
MALE_LITC     -3.250e+00  1.985e+00  -1.637 0.101632
FEMALE_LITC    4.813e+00  2.629e+00   1.831 0.067207 .
SEX            5.748e-01  1.959e-01   2.935 0.003360 **
CATEGORY       1.335e+00  1.275e-01  10.477  < 2e-16 ***
PARTY_ADAL    -5.305e+00  1.233e+00  -4.304 1.72e-05 ***
PARTY_BJP     -4.611e+00  4.960e-01  -9.297  < 2e-16 ***
PARTY_BSP     -5.082e+00  3.190e-01 -15.928  < 2e-16 ***
PATRY_INC     -5.134e+00  4.569e-01 -11.235  < 2e-16 ***
PARTY_IND      4.948e-01  1.275e-01   3.845 0.000124 ***
```

```
 $ MALE_LITC      : num  0 0.798 0 0 0 ...
 $ FEMALE_LITC    : num  0 0.633 0 0 0 ...
 $ SEX            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CATEGORY       : num  1 1 1 1 0 1 1 0 1 0 ...
 $ PARTY_ADAL     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_BJP      : num  0 1 0 0 0 0 0 0 0 0 ...
 $ PARTY_BSP      : num  0 0 1 0 0 0 0 0 0 0 ...
 $ PATRY_INC      : num  1 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_IND      : num  0 0 0 1 0 0 0 1 1 1 ...
 $ PARTY_NINSHAD  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_RLD      : num  0 0 0 0 0 0 1 0 0 0 ...
 $ PARTY_SBSP     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_SP       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ EDU_HighSchool: num  0 0 0 0 0 0 0 0 0 0 ...
 $ EDU_UG         : num  0 0 0 1 0 0 0 0 0 0 ...
 $ EDU_PG         : num  1 0 0 0 0 0 0 0 0 0 ...
 $ EDU_Graduate   : num  0 1 0 0 0 0 0 0 0 0 ...
 $ EDU_Other      : num  0 0 0 0 0 0 0 0 0 0 ...
> set.seed(122)
> pd1<-sample(2,nrow(data2),replace=TRUE, prob=c(0.7,0.3))
> train.lg<-data2[pd==1,]
> test.lg<-data2[pd==2,]
>
```

```
 [ reached getOption("max.print") -- omitted 1267 rows ]
> head(pred.re)
        fit       lwr       upr
1   3.801380 -2.942362 10.545123
2   1.590153 -5.119924  8.300231
3   3.706111 -3.005649 10.417871
4  10.012423  3.312770 16.712075
9  10.728895  4.027724 17.430065
10  8.912360  2.210670 15.614050
> head(val.re)
   CAND_AGE VOTE_2017 POSITION_2017 CRIM_CASE TOTA_ASST LIABILITES VOTE_2012 POSITION_2012
1        53     97035             1         0       104     398000         0             0
2        55     71449             2         0      2133    2985082     70274             1
3        52     71019             3         0       411   10214938         0             0
4        49      4187             4         0      2315     286731         0             0
9        32       945             9         0       415          0         0             0
10       35       810            10         0       955       9000       418            19
   MALE_LITC FEMALE_LITC SEX CATEGORY PARTY_ADAL PARTY_BJP PARTY_BSP PATRY_INC PARTY_IND
1     0.0000      0.0000   0        1          0         0         0         1         0
2     0.7977      0.6330   0        1          0         1         0         0         0
3     0.0000      0.0000   0        1          0         0         1         0         0
4     0.0000      0.0000   0        1          0         0         0         0         1
9     0.0000      0.0000   0        1          0         0         0         0         1
10    0.8500      0.6267   0        0          0         0         0         0         1
   PARTY_NINSHAD PARTY_RLD PARTY_SBSP PARTY_SP
1              0         0          0        0
2              0         0          0        0
```

```
  2.487964e-01   8.630414e-02
> str(train.lg)
'data.frame':   3707 obs. of  26 variables:
 $ CAND_AGE      : int  0 37 38 39 45 34 57 60 46 38 ...
 $ VOTE_2017     : int  1576 1255 1150 1113 659 617 319 94375 90318 65328 ...
 $ POSITION_2017 : num  0 0 0 0 0 0 0 1 0 0 ...
 $ CRIM_CASE     : num  0 0 0 0 0 0 0 0 6 0 ...
 $ TOTA_ASST     : num  2789 2359 2108 1028 2789 ...
 $ LIABILITES    : num  0 0 0 0 0 ...
 $ VOTE_2012     : int  0 0 12903 0 0 0 0 0 84623 0 ...
 $ POSITION_2012 : int  0 0 4 0 0 0 0 0 2 0 ...
 $ MALE_LITC     : num  0 0 0.84 0 0 ...
 $ FEMALE_LITC   : num  0 0 0.646 0 0 ...
 $ SEX           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CATEGORY      : num  0 1 1 0 1 1 1 1 1 1 ...
 $ PARTY_ADAL    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ PARTY_BJP     : num  0 0 0 0 0 0 0 1 0 0 ...
 $ PARTY_BSP     : num  0 0 0 0 0 0 0 0 0 1 ...
```

```
 $ EDU_Other     : num  0 0 0 0 0 0 0 0 0 0 ...
> # Fit the Sigmoid function
> Logit.1<-POSITION_2017  ~ VOTE_2017
> logit.plot<-glm(Logit.1, data=train.lg, family=binomial())
> summary(logit.plot)

Call:
glm(formula = Logit.1, family = binomial(), data = train.lg)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4648  -0.0107  -0.0098  -0.0095   2.6024

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.006e+01  6.701e-01  -15.01   <2e-16 ***
VOTE_2017    1.310e-04  8.976e-06   14.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
Number of Fisher Scoring iterations: 9

> library(ggplot2)
> pred.plot.logit <- predict.glm(logit.plot, newdata=test.lg, type="response")
> qplot( VOTE_2017, pred.plot.logit,data=test.lg, color=POSITION_2017 )
> #Using all variable#
> Logit_md = glm(POSITION_2017 ~ ., data= train.lg, family = binomial)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(Logit_md)

Call:
glm(formula = POSITION_2017 ~ ., family = binomial, data = train.lg)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6599  -0.0186   0.0000   0.0000   3.0474

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.237e+01  7.009e+02  -0.032   0.9745
CAND_AGE      -1.264e-02  1.351e-02  -0.935   0.3496
VOTE_2017      1.178e-04  1.080e-05  10.903   <2e-16 ***
```

```
> Logit_pd = predict(Logit_md,newdata = test.lg,type = "response")
> table(test.lg$POSITION_2017,Logit_pd>0.5)

    FALSE TRUE
  0  1466   16
  1    21   97
> (1412+99)/nrow(na.omit(test.lg))
[1] 0.944375
> library(ROCR)
> ROCRpred = prediction(Logit_pd, test.lg$POSITION_2017)
> as.numeric(performance(ROCRpred, "auc")@y.values)
[1] 0.9928206
> perf = performance(ROCRpred, "tpr","fpr")
> plot(perf)
> #KNN Modeling #
> data2 <- Elet [,-c(1:5,7,11)]
> data2$POSITION_2017<-ifelse(data2$POSITION_2017=="1",1,0)
> str(data2)
'data.frame':   5307 obs. of  26 variables:
 $ CAND_AGE      : int  53 55 52 49 0 37 38 39 32 35 ...
 $ VOTE_2017     : int  97035 71449 71019 4187 1576 1255 1150 1113 945 810 ...
 $ POSITION_2017 : num  1 0 0 0 0 0 0 0 0 0 ...
 $ CRIM_CASE     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ TOTA_ASST     : num  104 2133 411 2315 2789 ...
 $ LIABILITES    : num  398000 2985082 10214938 286731 0 ...
 $ VOTE_2012     : int  0 70274 0 0 0 0 12903 0 0 418 ...
 $ POSITION_2012 : int  0 1 0 0 0 0 4 0 0 19 ...
 $ MALE_LITC     : num  0 0.798 0 0 0 ...
 $ FEMALE_LITC   : num  0 0.633 0 0 0 ...
 $ SEX           : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
 $ EDU_PG        : num  1 0 0 0 0 0 0 0 0 0 ...
 $ EDU_Graduate  : num  0 1 0 0 0 0 0 0 0 0 ...
 $ EDU_Other     : num  0 0 0 0 0 0 0 0 0 0 ...
> #Diving data set for trian and test#
> library(caTools)
> set.seed(110)
> spl = sample.split(data2$POSITION_2017, SplitRatio = 0.7 )
> train_knn = subset(data2, spl == T)
> test_knn = subset(data2, spl == F)
> dim(train_knn)
[1] 3715   26
> dim(test_knn)
[1] 1592   26
> #Scaling the data #
> sum(data2$POSITION_2017)
[1] 403
> sum(test_knn$POSITION_2017)
[1] 121
> sum(train_knn$POSITION_2017)
[1] 282
> scale(train_knn)
       CAND_AGE    VOTE_2017 POSITION_2017  CRIM_CASE     TOTA_ASST    LIABILITES
1     0.70811360  2.633619407      3.488622 -0.2357693 -1.7688902976 -0.1112592547
3     0.64580121  1.779855723     -0.286569 -0.2357693 -1.4410566573  1.2574028436
4     0.45886405 -0.413361318     -0.286569 -0.2357693  0.5921526298 -0.1267722042
5    -2.59444288 -0.499046165     -0.286569 -0.2357693  1.0983192275 -0.1667477904
6    -0.28888458 -0.509580380     -0.286569 -0.2357693  0.6391385587 -0.1667477904
7    -0.22657220 -0.513026152     -0.286569 -0.2357693  0.3711051915 -0.1667477904
```

```
  2.146076e-01   0.000000e+00   2.310200e-01   2.786699e-01   3.190615e-01   3.257677e-01
  EDU_Graduate      EDU_Other
  2.427045e-01   9.663954e-02
> library(class)
> pred = knn(train_knn[-1], test_knn[-1], train_knn[,1], k = 3)
> table.knn = table(test_knn[,1], pred)
> accuracy.knn = sum(diag(table.knn))/sum(table.knn)
> accuracy.knn
[1] 0.08417085
> loss.knn = table.knn[2,1]/(table.knn[2,1] + table.knn[1,1])
> loss.knn
[1] 0.01086957
> library(class)
> pred = knn(train_knn[-1], test_knn[-1], train_knn[,1], k = 19)
> table.knn19 = table(test_knn[,1], pred)
> accuracy.knn19 = sum(diag(table.knn19))/sum(table.knn19)
> accuracy.knn19
[1] 0.1067839
> loss.knn = table.knn19[2,1]/(table.knn19[2,1] + table.knn19[1,1])
> loss.knn
[1] 0.007575758
> library(class)
> pred = knn(train_knn[-1], test_knn[-1], train_knn[,1], k =2)
> table.knn2 = table(test_knn[,1], pred)
> accuracy.knn2 = sum(diag(table.knn2))/sum(table.knn2)
> accuracy.knn2
[1] 0.07600503
> loss.knn = table.knn2[2,1]/(table.knn2[2,1] + table.knn2[1,1])
> loss.knn
[1] 0.01538462
>
```

```
~/My Files/Great Lakes Projects/Project -4/
> #Preapring NB module#
> set.seed(119)
> spx = sample.split(data2$POSITION_2017, SplitRatio = 0.7 )
> train_nb = subset(data2, spx == T)
> test_nb = subset(data2, spx == F)
> dim(train_nb)
[1] 3715   26
> dim(test_nb)
[1] 1592   26
> train_nb$POSITION_2017<-as.factor(train_nb$POSITION_2017)
> test_nb$POSITION_2017<-as.factor(test_nb$POSITION_2017)
> library(e1071)
> NB = naiveBayes(x = train_nb[-1], y = train_nb$POSITION_2017)
> ypred.NB = predict(NB, newdata = test_nb[-1])
> ypred.NB
   [1] 1 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
  [45] 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
  [89] 0 1 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0
 [133] 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 0 0 0 1 1 0 0 0 1 1 0 0 0 0 0 0
 [177] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 1 1 0 0
```

```
Levels: 0 1
> tab.NB = table(test_nb[,1], ypred.NB)
> tab.NB
    ypred.NB
       0   1
  0  112   0
  25  26   0
  26  20   1
  27  26   1
  28  22   0
  29  19   0
  30  37   2
  31  29   0
  32  35   3
  33  28   3
  34  40   3
  35  33   2
  36  47   5
  37  33   2
  38  39   2
  39  41   5
```

```
  77   2   0
  78   1   1
  80   0   1
  81   1   0
  82   1   0
> accuracy.NB = sum(diag(tab.NB))/sum(tab.NB)
> accuracy.NB
[1] 0.07035176
>
>
>
```