Mini Project 5 - Project Report by James Peter

Table of Contents

Contents

1. Project objective

- The objective is analyse the previous State Legislative Assembly Election results to find the important factors which influences the prospective of a Candidate to win the poll.
- Collect data from Election commision of India, MyNeta.com and other resources, perform data wrangling to create a good dataset with relevant features which will aid in identifying the important factors which contribute towards a contesting candidates chances to win the election poll.

```
Warning: package 'stringr' was built under R version 3.4.4
```

```
Warning: package 'readr' was built under R version 3.4.4
```

2. Data description and summary

Removing NOTA observations

```
T17 <- AE17 %>% filter(PARTYABBRE!='NOTA')
head(T17)
```

```
# A tibble: 6 x 15
  ST_CODE ST_NAME MONTH  YEAR DIST_NAME AC_NO AC_NAME AC_TYPE CAND_NAME
  <chr>   <chr>   <dbl> <dbl> <chr>     <dbl> <chr>   <chr>   <chr>
1 S05     Goa      3.00  2017 North Goa  1.00 Mandrem GEN     DAYANAND RA~
2 S05     Goa      3.00  2017 North Goa  1.00 Mandrem GEN     LAXMIKANT P~
3 S05     Goa      3.00  2017 North Goa  1.00 Mandrem GEN     SHRIDHAR LA~
4 S05     Goa      3.00  2017 North Goa  1.00 Mandrem GEN     DEVENDRA KR~
5 S05     Goa      3.00  2017 North Goa  1.00 Mandrem GEN     RAJENDRA M.~
6 S05     Goa      3.00  2017 North Goa  1.00 Mandrem GEN     SANJAY KRIS~
# ... with 6 more variables: CAND_SEX <chr>, CAND_CATEGORY <chr>,
#   CAND_AGE <chr>, PARTYABBRE <chr>, TOTALVALIDVOTESPOLLED <dbl>,
#   POSITION <dbl>
```

Observation:

- The data has been filtered to include only the candidates information, excluding the NOTA observation which doesn't represent any contesting candidate.

Summary of 2017 MLA election dataset

```
   ST_CODE                    ST_NAME         MONTH        YEAR
 Length:7243        Goa           : 251   Min.   :3   Min.   :2017
 Class :character   Manipur       : 266   1st Qu.:3   1st Qu.:2017
 Mode  :character   Punjab        :1189   Median :3   Median :2017
                    Uttar Pradesh:4900   Mean   :3   Mean   :2017
                    Uttarakhand   : 637   3rd Qu.:3   3rd Qu.:2017
                                          Max.   :3   Max.   :2017

    DIST_NAME        AC_NO                  AC_NAME     AC_TYPE
 Allahabad: 181   Min.   :  1.0   Rudauli         :  28   GEN:5867
 Ludhiana : 136   1st Qu.: 44.0   Agra South      :  26   SC :1252
 South Goa: 132   Median :114.0   Allahabad North:  26   ST : 124
 Gorakhpur: 127   Mean   :155.1   Bilari          :  26
```

```
 Varanasi : 127   3rd Qu.:262.0   Amethi         :  24
 Lucknow  : 126   Max.   :403.0   Varanasi Cantt.:  24
 (Other)  :6414                  (Other)        :7089
  CAND_NAME         CAND_SEX CAND_CATEGORY    CAND_AGE       PARTYABBRE
 Length:7243        F: 663   GEN:5279      Min.   :25.00   IND    :2131
 Class :character   M:6576   SC :1818      1st Qu.:37.00   BSP    : 592
 Mode  :character   O:   4   ST : 146      Median :45.00   BJP    : 578
                                          Mean   :46.03   INC    : 403
                                          3rd Qu.:54.00   SP     : 334
                                          Max.   :91.00   RLD    : 286
                                                          (Other):2919
 TOTALVALIDVOTESPOLLED    POSITION
 Min.   :    20.0      Min.   : 1.000
 1st Qu.:   452.5      1st Qu.: 3.000
 Median :  1000.0      Median : 7.000
 Mean   : 15230.4      Mean   : 7.154
 3rd Qu.: 14227.5      3rd Qu.:10.000
 Max.   :262741.0      Max.   :30.000
```

3. Data Wrangling and Data selection

- Since the project involved to collect the data unlike other projects, which is most often will be the case when working in real time, have walked through all the steps which was followed to prepare the data for training the models.
- The data collection part being the most time consuming task and is considered one of defining task in performing data analysis which results in direct success of the analysis

3.1 Data extraction and conversion

```
Warning: 1 parsing failure.
row # A tibble: 1 x 4 col     row   col expected actual expected   <int> <int> <chr>    <chr>  actual 1

Warning: 7 parsing failures.
row # A tibble: 5 x 4 col     row   col expected actual expected   <int> <int> <chr>    <chr>  actual 1
... ................. ... ............................ ........ ............................ ...... .
See problems(...) for more details.

Warning: 13 parsing failures.
row # A tibble: 5 x 4 col     row   col expected actual expected   <int> <int> <chr>    <chr>  actual 1
... ................. ... ............................ ........ ............................ ...... .
See problems(...) for more details.
```

Observation:

- The Goa sates Vidhan Sabha election results will be used to explain the process used to collect the data. And the same process have been applied to collect the data for the remaining states.

converting the text data to perform joins

```
can_data_goa$CAND_NAME <- sapply(can_data_goa$CAND_NAME, toupper)
can_data_manipur$CAND_NAME <- sapply(can_data_manipur$CAND_NAME, toupper)
can_data_punjab$CAND_NAME <- sapply(can_data_punjab$CAND_NAME, toupper)
can_data_up$CAND_NAME <- sapply(can_data_up$CAND_NAME, toupper)
can_data_uttarakhand$CAND_NAME <- sapply(can_data_uttarakhand$CAND_NAME, toupper)
```

Subsetting data based on State

```
G17 <- T17 %>% filter(ST_NAME=='Goa')
P17 <- T17 %>% filter(ST_NAME=='Punjab')
M17 <- T17 %>% filter(ST_NAME=='Manipur')
UP17 <- T17 %>% filter(ST_NAME=='Uttar Pradesh')
UKD17 <- T17 %>% filter(ST_NAME=='Uttarakhand')
```

Performing joins to include more features

```
P17 <- P17 %>% left_join(can_data_punjab)
```

```
Joining, by = "CAND_NAME"
```

```
G17 <- G17 %>% left_join(can_data_goa)
```

```
Joining, by = "CAND_NAME"
```

```
M17 <- M17 %>% left_join(can_data_manipur)
```

```
Joining, by = "CAND_NAME"
```

```
UP17 <- UP17 %>% left_join(can_data_up)
```

```
Joining, by = "CAND_NAME"
```

```
UKD17 <- UKD17 %>% left_join(can_data_uttarakhand)
```

```
Joining, by = "CAND_NAME"
```

Choosing Goa state to perform isolated analysis

```
Joining, by = "AC_NAME"
```

```
Warning: Column `AC_NAME` joining factor and character vector, coercing
into character vector
```

Missing values present in data

```
for (col in colnames(G17)){
  cat(col,': ',sum(is.na(G17[,col])))
  cat( '\n')
}
```

```
ST_CODE :  0
ST_NAME :  0
MONTH :  0
```

```
YEAR :  0
DIST_NAME :  0
AC_NO :  0
AC_NAME :  0
AC_TYPE :  0
CAND_NAME :  0
CAND_SEX :  0
CAND_CATEGORY :  0
CAND_AGE :  0
PARTYABBRE :  0
TOTALVALIDVOTESPOLLED :  0
POSITION :  0
Sno :  51
Constituency :  51
Party :  51
Criminal Case :  51
Education :  51
Total Assets :  51
Liabilities :  51
Total_Assets :  52
Liabilities_P :  51
TOTAL VOTES POLLED :  0
TotalElectors :  0
```

Observaton:

- The data set has few missing values which will be filled using revelant method based on the type of parameter

Dataset with added Features

```
 [1] "ST_CODE"               "ST_NAME"
 [3] "MONTH"                 "YEAR"
 [5] "DIST_NAME"             "AC_NO"
 [7] "AC_NAME"               "AC_TYPE"
 [9] "CAND_NAME"             "CAND_SEX"
[11] "CAND_CATEGORY"         "CAND_AGE"
[13] "PARTYABBRE"            "TOTALVALIDVOTESPOLLED"
[15] "POSITION"              "Sno"
[17] "Constituency"          "Party"
[19] "Criminal Case"         "Education"
[21] "Total Assets"          "Liabilities"
[23] "Total_Assets"          "Liabilities_P"
[25] "TOTAL VOTES POLLED"    "TotalElectors"
```

Observation:

- The features in the data set after initial data wrangling

The data of recontesting candidates

```
head(recon_Goa[,c(2,3)])
```

```
# A tibble: 6 x 2
  `Name (Party)`                              `Total Assets in Goa 2017`
  <chr>                                       <chr>
1 Michael Vincent Lobo (BJP)                  "54,59,81,558  \n54 Crore+"
2 Pratapsingh R Rane (INC)                    "50,00,16,663  \n50 Crore+"
3 Pandurang Arjun Madkaikar (BJP)             "32,18,54,849  \n32 Crore+"
4 Atanasio J. Monserrate (United Goans Party) "30,81,18,480  \n30 Crore+"
5 Jennifer Monserrate (INC)                   "30,81,18,480  \n30 Crore+"
6 Kiran Mohan Kandolkar (BJP)                 "9,37,43,482  \n9 Crore+"
```

Performing data cleaning for Candidate name

```
recon_Goa <- recon_Goa %>%
  mutate(CAND_NAME = purrr::map_chr(
    stringr::str_replace(`Name (Party)`, " \\(.*\\)", ""),
    ~ paste(toupper(.))
    ))

head(recon_Goa$CAND_NAME)
```

```
[1] "MICHAEL VINCENT LOBO"      "PRATAPSINGH R RANE"
[3] "PANDURANG ARJUN MADKAIKAR" "ATANASIO J. MONSERRATE"
[5] "JENNIFER MONSERRATE"       "KIRAN MOHAN KANDOLKAR"
```

Observation:

- The recontesting data includes the names of all the contesting candidates who have been elected in the previous election, however the Candidate name has party name in the brackets which needs to be separated before joining based on the candidate name with main dataset.

Performing join to include recontesting feature

```
recon_Goa <- recon_Goa[,c("CAND_NAME","Recontesting")]
G17 <- G17 %>% left_join(recon_Goa)
```

```
Joining, by = "CAND_NAME"
```

Adding features: corepathi, education level, ITR status

```
Goa_winner_corepathi$CAND_NAME <- sapply(Goa_winner_corepathi$CAND_NAME, toupper)
Goa_winner_graduate_above$CAND_NAME <- sapply(Goa_winner_graduate_above$CAND_NAME, toupper)
Goa_winner_filed_ITR$CAND_NAME <- sapply(Goa_winner_filed_ITR$CAND_NAME, toupper)
Goa_can_corepathi$CAND_NAME <- sapply(Goa_can_corepathi$CAND_NAME, toupper)
Goa_can_graduate_above$CAND_NAME <- sapply(Goa_can_graduate_above$CAND_NAME, toupper)
Goa_can_filed_ITR$CAND_NAME <- sapply(Goa_can_filed_ITR$CAND_NAME, toupper)
```

```
G17 <- G17 %>% left_join(Goa_winner_corepathi[,c('CAND_NAME','Winner_corepathi')])
```

```
Joining, by = "CAND_NAME"
```

```
G17 <- G17 %>% left_join(Goa_winner_graduate_above[,c('CAND_NAME','Winner_graduate_above')])
```

```
Joining, by = "CAND_NAME"
```

```
G17 <- G17 %>% left_join(Goa_winner_filed_ITR[,c('CAND_NAME','Winner_filed_ITR')])
```

```
Joining, by = "CAND_NAME"
```

```
G17 <- G17 %>% left_join(Goa_can_corepathi[,c('CAND_NAME','Cand_corepathi')])
```

```
Joining, by = "CAND_NAME"
```

```
G17 <- G17 %>% left_join(Goa_can_graduate_above[,c('CAND_NAME','Cand_graduate_above')])
```

```
Joining, by = "CAND_NAME"
```

```
G17 <- G17 %>% left_join(Goa_can_filed_ITR[,c('CAND_NAME','Cand_filed_ITR')])
```

```
Joining, by = "CAND_NAME"
```

Correcting errors and adding National party indicator

```
# replacing mis-spelled party abbrevation
G17$PARTYABBRE<- str_replace(G17$PARTYABBRE,'AAAP','AAP')

National_parties <- c('BJP', 'BSP', 'INC', 'NCP', 'CPI', 'CPI-M', 'AAP', 'RJD', 'SP', 'AITC')

C_r1 <- function(x){
  if (x %in% National_parties){
      return(1)
  }

  else{
    return(0)
  }
}


G17$National_party_candidate <- sapply(G17$PARTYABBRE, C_r1)
```

Total Variables after Data Wrangling

```
 [1] "ST_CODE"                 "ST_NAME"
 [3] "MONTH"                   "YEAR"
 [5] "DIST_NAME"               "AC_NO"
 [7] "AC_NAME"                 "AC_TYPE"
 [9] "CAND_NAME"               "CAND_SEX"
[11] "CAND_CATEGORY"           "CAND_AGE"
[13] "PARTYABBRE"              "TOTALVALIDVOTESPOLLED"
[15] "POSITION"                "Sno"
[17] "Constituency"            "Party"
```

```
[19] "Criminal Case"            "Education"
[21] "Total Assets"             "Liabilities"
[23] "Total_Assets"             "Liabilities_P"
[25] "TOTAL VOTES POLLED"       "TotalElectors"
[27] "Recontesting"             "Winner_corepathi"
[29] "Winner_graduate_above"    "Winner_filed_ITR"
[31] "Cand_corepathi"           "Cand_graduate_above"
[33] "Cand_filed_ITR"           "National_party_candidate"
```

Observation:

- The total feature set after the data wrangling process which will subset based on important feature that can be used for the creating model.

4. Dataset with most revelant features

Dataset information

```
Classes 'tbl_df', 'tbl' and 'data.frame':   251 obs. of  24 variables:
 $ DIST_NAME               : Factor w/ 121 levels "Agra","Aligarh",..: 85 85 85 85 85 85 85 85 85 85 ..
 $ AC_NAME                 : chr  "Mandrem" "Mandrem" "Mandrem" "Mandrem" ...
 $ AC_TYPE                 : Factor w/ 3 levels "GEN","SC","ST": 1 1 1 1 1 1 1 2 2 2 ...
 $ CAND_NAME               : chr  "DAYANAND RAGHUNATH SOPTE" "LAXMIKANT PARSEKAR" "SHRIDHAR LADU MANJREI
 $ POSITION                : num  1 2 3 4 6 7 8 1 2 3 ...
 $ CAND_SEX                : Factor w/ 3 levels "F","M","O": 2 2 2 2 2 2 2 2 2 2 ...
 $ CAND_CATEGORY           : Factor w/ 3 levels "GEN","SC","ST": 1 1 1 1 1 1 1 2 2 2 ...
 $ CAND_AGE                : num  53 60 69 53 49 70 50 63 62 43 ...
 $ PARTYABBRE              : chr  "INC" "BJP" "MAG" "AAP" ...
 $ TOTALVALIDVOTESPOLLED   : num  16490 9371 678 620 234 ...
 $ TOTAL VOTES POLLED      : num  28071 28071 28071 28071 28071 ...
 $ TotalElectors           : num  31369 31369 31369 31369 31369 ...
 $ Criminal Case           : num  0 0 0 2 0 NA 0 NA 0 0 ...
 $ Education               : Factor w/ 9 levels "10th Pass","12th Pass",..: 5 9 1 5 1 NA 1 NA 5 8 ...
 $ Total_Assets            : atomic  35131000 89813996 27517393 9895320 3100000 ...
  ..- attr(*, "problems")=Classes 'tbl_df', 'tbl' and 'data.frame': 1 obs. of  4 variables:
  .. ..$ row     : int 103
  .. ..$ col     : int NA
  .. ..$ expected: chr "a number"
  .. ..$ actual  : chr "Nil"
 $ Liabilities_P           : num  4550987 2910108 6669749 215000 80000 ...
 $ Recontesting            : num  NA 1 NA NA NA NA NA NA 1 NA ...
 $ National_party_candidate: num  1 1 0 1 1 0 0 0 1 1 ...
 $ Winner_corepathi        : num  1 NA NA NA NA NA NA NA NA NA ...
 $ Winner_graduate_above   : num  1 NA NA NA NA NA NA NA NA NA ...
 $ Winner_filed_ITR        : num  1 NA NA NA NA NA NA NA NA NA ...
 $ Cand_corepathi          : num  1 1 1 NA NA NA NA NA 1 1 ...
 $ Cand_graduate_above     : num  1 1 NA 1 NA NA NA NA 1 NA ...
 $ Cand_filed_ITR          : num  1 1 1 1 1 NA NA NA 1 1 ...
```

Identifying NA's in the dataset and imputation

```
DIST_NAME :  0
```

```
AC_NAME :  0
AC_TYPE :  0
CAND_NAME :  0
POSITION :  0
CAND_SEX :  0
CAND_CATEGORY :  0
CAND_AGE :  0
PARTYABBRE :  0
TOTALVALIDVOTESPOLLED :  0
TOTAL VOTES POLLED :  0
TotalElectors :  0
Criminal Case :  51
Education :  51
Total_Assets :  52
Liabilities_P :  51
Recontesting :  224
National_party_candidate :  0
Winner_corepathi :  222
Winner_graduate_above :  239
Winner_filed_ITR :  222
Cand_corepathi :  124
Cand_graduate_above :  175
Cand_filed_ITR :  74
```

```
           10th Pass            12th Pass             5th Pass
                  46                   36                    8
            8th Pass             Graduate Graduate Professional
                  15                   41                   15
           Illiterate              Others         Post Graduate
                   1                   18                   20
                NA's
                  51
```

Filling NA' in Education feature

```
GW17$Education <- as.factor(GW17$Education)
GW17$Education[is.na(GW17$Education)] <- '10th Pass'

summary(GW17$Education)
```

```
           10th Pass            12th Pass             5th Pass
                  97                   36                    8
            8th Pass             Graduate Graduate Professional
                  15                   41                   15
           Illiterate              Others         Post Graduate
                   1                   18                   20
```

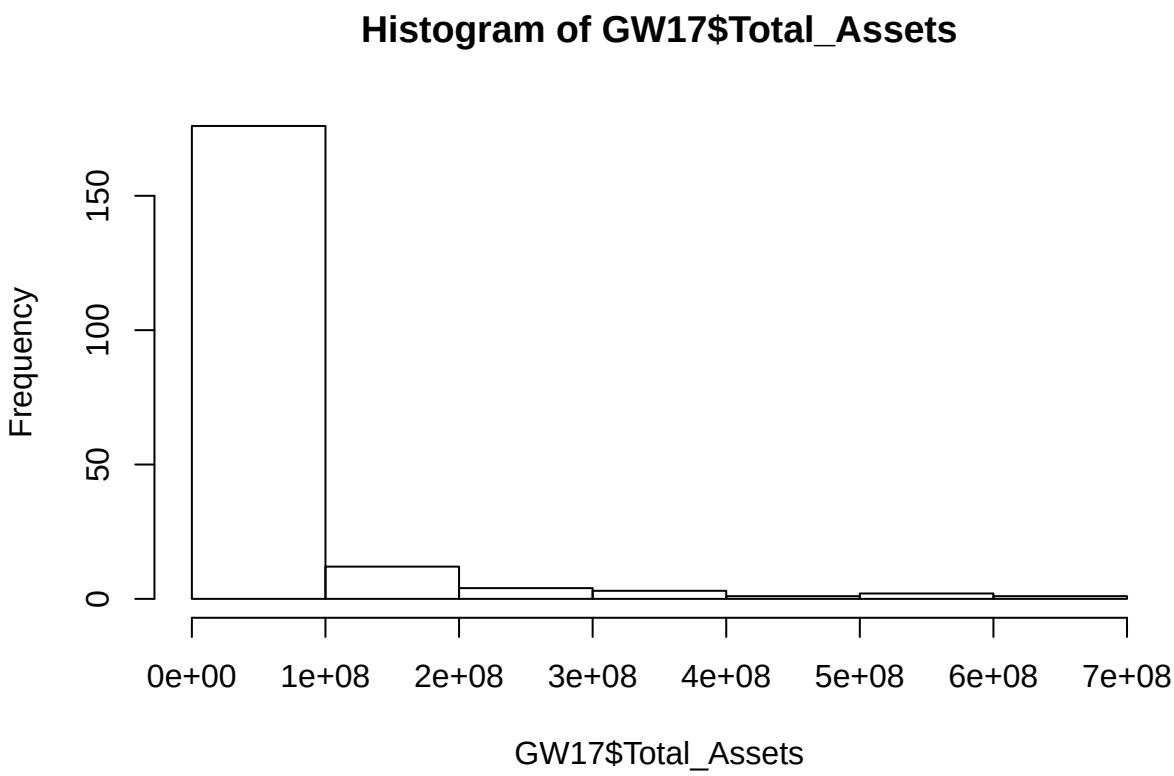
Observation:

- Since 10th pass has more number of count using the mode method to impute the missing values in the education

Imputing Total Assets and Liability feature

Summary of Total Assets variable

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
    20000   4122684  18431376  50065706  51312394 657878880        52
```




```
head(subset(GW17, GW17$National_party_candidate == 1 & is.na(GW17$Total_Assets))[,c('CAND_NAME','PARTYA
```

```
# A tibble: 9 x 2
  CAND_NAME                           PARTYABBRE
  <chr>                               <chr>
1 MANOHAR PANDURANG SHIRODKAR         INC
2 RAJESH VALVAIKAR                    AAP
3 GLENN SOUZA TICLO                   BJP
4 ROSY URSULA D`SOUZA                 AAP
5 OSBERT D`CUNHA                      NCP
6 ANTONIO CAETANO  FERNANDES          INC
7 PEDRO CAITANO PIRES ALIAS PETER PIRES CPI
8 PANDURANG MADKAIKAR                 BJP
9 PRAVIN  ZANTYE                      BJP
```

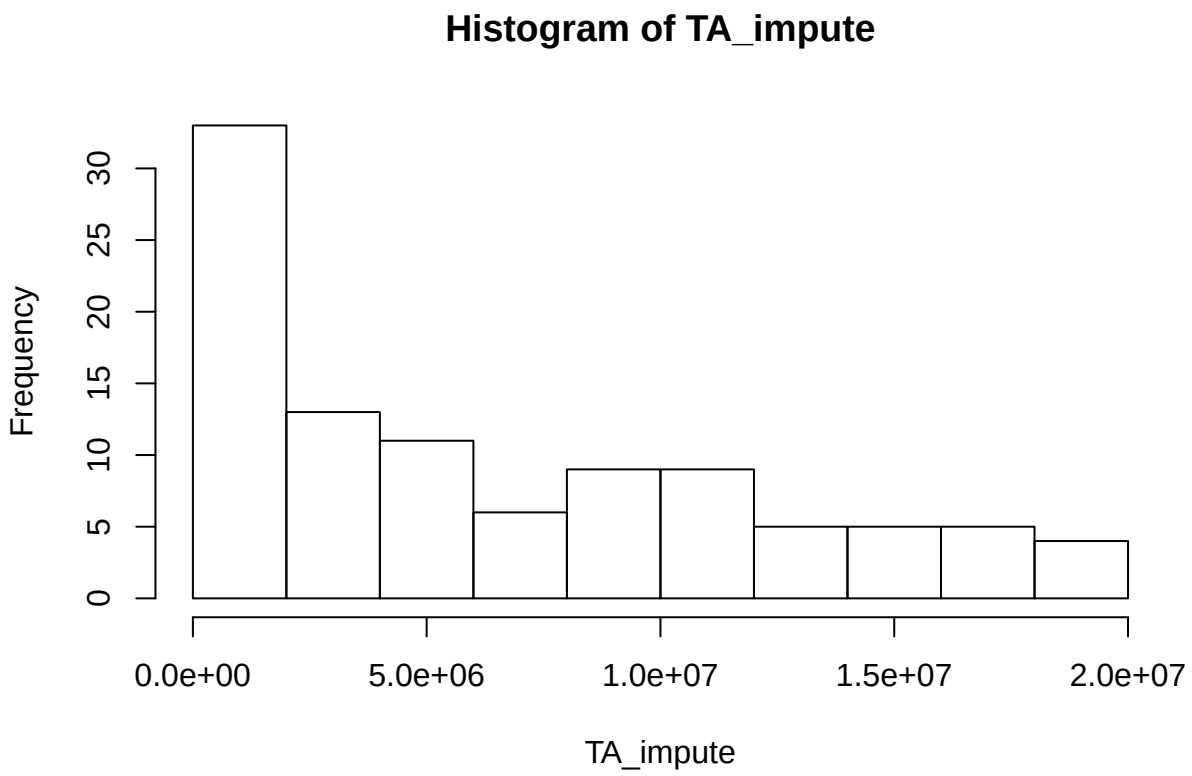
Observatoin:

- The Assets has very huge outliers which are effecting the mean of the feature, so using the mean value would not be ideal, but since also the candidates who have missing assets belong to national candidate who are most likely to rich than the non_national party candidate then using a value between the

median and mean, average of the two, would be considred to impute the missing values.

Choosing dataset below the median

```
TA_impute <- subset(GW17$Total_Assets, GW17$Total_Assets<=18431376)

hist(TA_impute)
```



Observation:

- The data is more observations below the median as expected for Total assets

Summary after threshold selecting

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  20000  1535342  4122684  6420849 10962648 18431376
```

Imputing Total Assets

```
GW17$Total_Assets[is.na(GW17$Total_Assets)] <- 34190676
summary(GW17$Total_Assets)
```

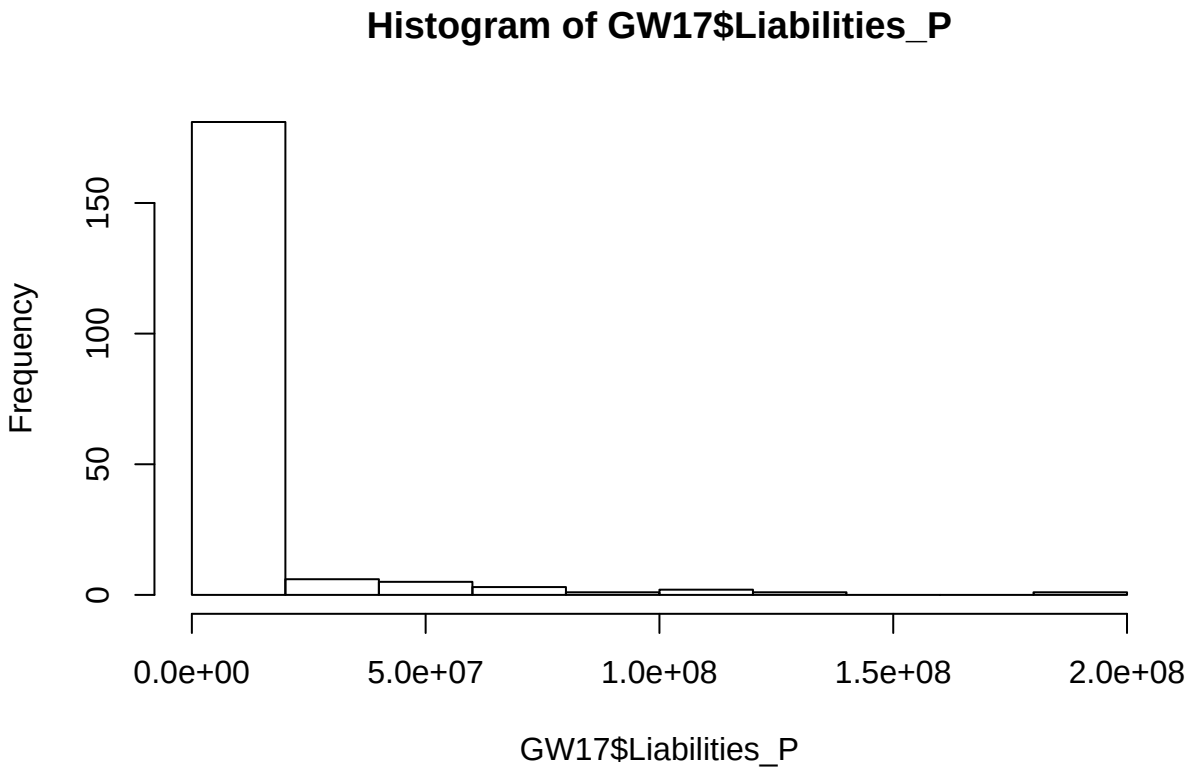
```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   20000   8062397  31265783  46776855  38257656 657878880
```

Observation:

- The summary of the Total_assets after imputation

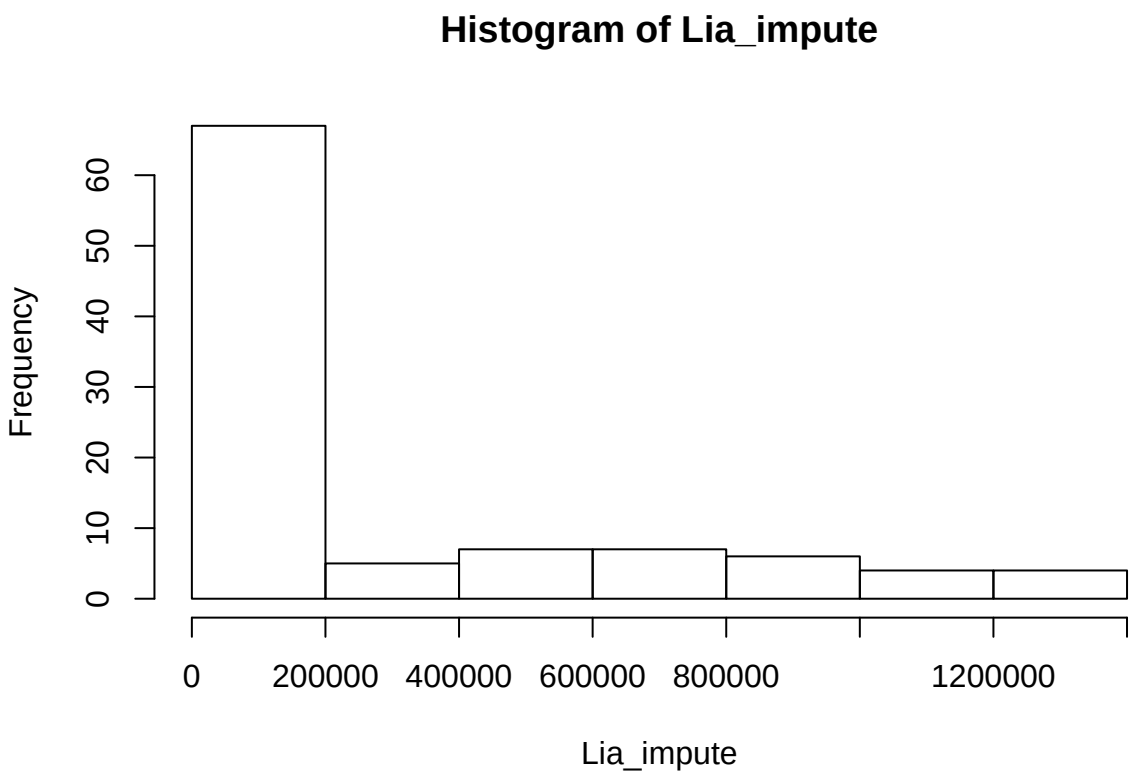
Summary and Distribution of Liability

```
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
      0         0   1388206   8739132   6311804 194070238        51
```



Distribution below the median value

```
Lia_impute <- subset(GW17$Liabilities_P, GW17$Liabilities_P<=1388206)
hist(Lia_impute)
```



Imputing Liability

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0       0       0  255730  474323 1276412 
```

Observation:

- The Liability has more observation with 0 values, that less likely that candidate would have any liability, so imputing liability with zero value would be good.

```
GW17$Liabilities_P[is.na(GW17$Liabilities_P)] <- 0
summary(GW17$Liabilities_P)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
        0         0    471011   6963452   4297109 194070238 
```

Checking for negative Wealth

```
subset(GW17[,c('Total_Assets','Liabilities_P')], (GW17$Total_Assets - GW17$Liabilities_P)<0)
```

```
# A tibble: 6 x 2
  Total_Assets Liabilities_P
         <dbl>         <dbl>
1    178767510     194070238
2       594686       1823100
```

```
3     50122325      51963803
4      3672862       7877310
5      4909686       9470389
6      2071716       4229957
```

Observation:

- The check for negative values, means candidate who have debt is less few, and these negative values has not been introduced by the imputation steps, which refelects the actual financial status of the candidates

Candidates count from each party

```
   AAP   APoI    BJP   BMUP    CPI    GFP  GoPrP    GSM   GSRP    GVP
    39      1     36      3      2      4      3      5      8      5
   INC    IND    MAG    NCP   NGRF SaJPCs    SHS    UGP
    37     58     25     17      2      1      3      2
```

Imputing NA's in numeric features

```
GW17 <- GW17 %>% mutate_if(is.numeric, funs(replace(., is.na(.), 0)))

summary(GW17)
```

```
     DIST_NAME     AC_NAME          AC_TYPE    CAND_NAME
 South Goa:132   Length:251         GEN:243   Length:251
 North Goa:119   Class :character   SC :  8   Class :character
 Agra     :  0   Mode  :character   ST :  0   Mode  :character
 Aligarh  :  0
 Allahabad:  0
 Almora   :  0
 (Other)  :  0
    POSITION      CAND_SEX CAND_CATEGORY    CAND_AGE       PARTYABBRE
 Min.   : 1.000   F: 19    GEN:236       Min.   :27.00   IND    :58
 1st Qu.: 2.000   M:232    SC :  8       1st Qu.:41.00   AAP    :39
 Median : 4.000   O:  0    ST :  7       Median :47.00   INC    :37
 Mean   : 4.219                          Mean   :47.99   BJP    :36
 3rd Qu.: 6.000                          3rd Qu.:54.00   MAG    :25
 Max.   :13.000                          Max.   :78.00   NCP    :17
                                                         (Other):39
 TOTALVALIDVOTESPOLLED TOTAL VOTES POLLED TotalElectors   Criminal Case
 Min.   :   20.0       Min.   :16556      Min.   :20948   Min.   :0.0000
 1st Qu.:  254.5       1st Qu.:21333      1st Qu.:26033   1st Qu.:0.0000
 Median : 1479.0       Median :22777      Median :28171   Median :0.0000
 Mean   : 3606.8       Mean   :22793      Mean   :27806   Mean   :0.1833
 3rd Qu.: 6398.5       3rd Qu.:24149      3rd Qu.:30463   3rd Qu.:0.0000
 Max.   :17093.0       Max.   :28522      Max.   :35938   Max.   :4.0000

        Education   Total_Assets       Liabilities_P
 10th Pass    :97   Min.   :    20000   Min.   :        0
 Graduate     :41   1st Qu.:  8062397   1st Qu.:        0
 12th Pass    :36   Median : 31265783   Median :   471011
```

```
 Post Graduate:20   Mean   : 46776855   Mean   :  6963452
 Others       :18   3rd Qu.: 38257656   3rd Qu.:  4297109
 8th Pass     :15   Max.   :657878880   Max.   :194070238
 (Other)      :24
  Recontesting    National_party_candidate Winner_corepathi
 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 :1.0000           Median :0.0000
 Mean   :0.1076   Mean   :0.5219           Mean   :0.1155
 3rd Qu.:0.0000   3rd Qu.:1.0000           3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000           Max.   :1.0000

 Winner_graduate_above Winner_filed_ITR Cand_corepathi
 Min.   :0.00000       Min.   :0.0000   Min.   :0.000
 1st Qu.:0.00000       1st Qu.:0.0000   1st Qu.:0.000
 Median :0.00000       Median :0.0000   Median :1.000
 Mean   :0.04781       Mean   :0.1155   Mean   :0.506
 3rd Qu.:0.00000       3rd Qu.:0.0000   3rd Qu.:1.000
 Max.   :1.00000       Max.   :1.0000   Max.   :1.000

 Cand_graduate_above Cand_filed_ITR
 Min.   :0.0000      Min.   :0.0000
 1st Qu.:0.0000      1st Qu.:0.0000
 Median :0.0000      Median :1.0000
 Mean   :0.3028      Mean   :0.7052
 3rd Qu.:1.0000      3rd Qu.:1.0000
 Max.   :1.0000      Max.   :1.0000
```

Observation:

- The summary of the data after data wrangling and imputation steps

Count of NA's after Imputation

```
for (col in colnames(GW17)){
  cat(col,': ',sum(is.na(GW17[,col])))
  cat( '\n')
}
```

```
DIST_NAME :  0
AC_NAME :  0
AC_TYPE :  0
CAND_NAME :  0
POSITION :  0
CAND_SEX :  0
CAND_CATEGORY :  0
CAND_AGE :  0
PARTYABBRE :  0
TOTALVALIDVOTESPOLLED :  0
TOTAL VOTES POLLED :  0
TotalElectors :  0
Criminal Case :  0
```

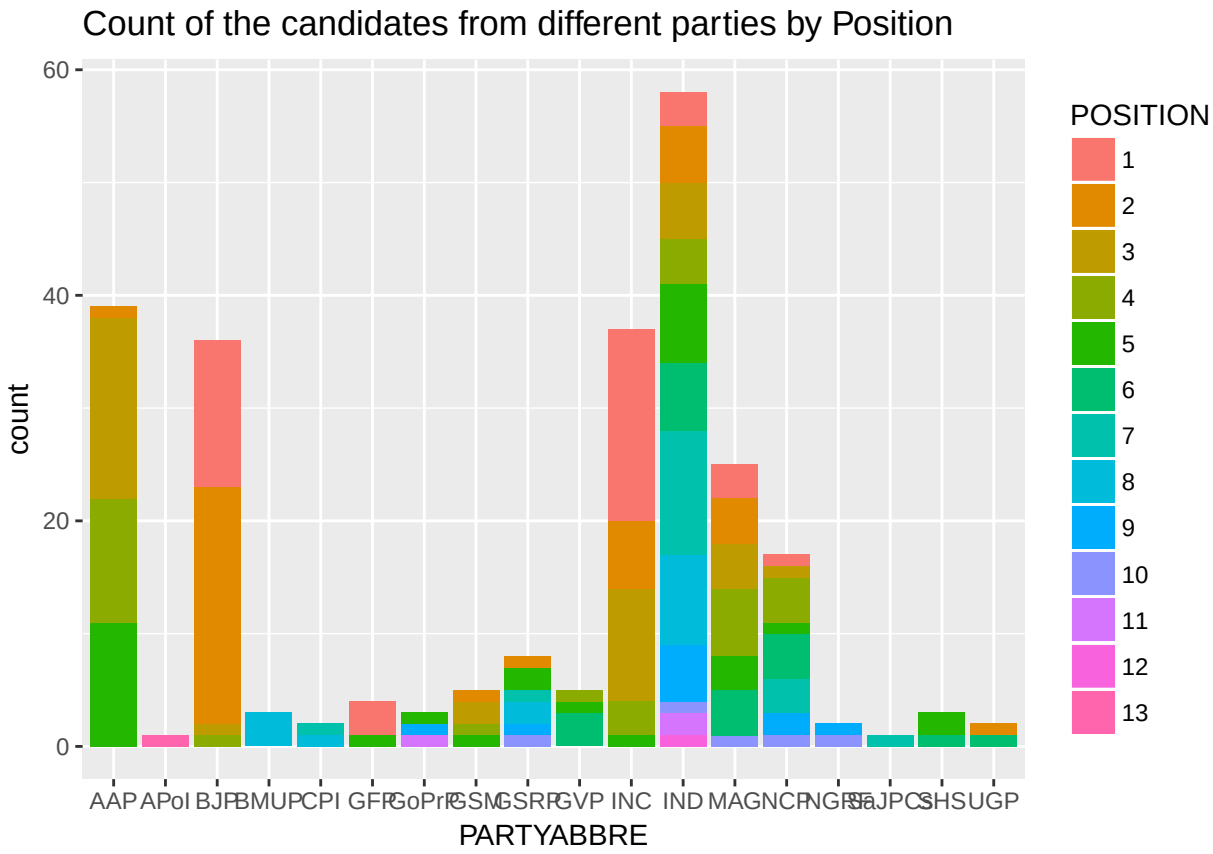
```
Education :  0
Total_Assets :  0
Liabilities_P :  0
Recontesting :  0
National_party_candidate :  0
Winner_corepathi :  0
Winner_graduate_above :  0
Winner_filed_ITR :  0
Cand_corepathi :  0
Cand_graduate_above :  0
Cand_filed_ITR :  0
```

Observation:

- The missing has been imputated and the data contains no missing values.

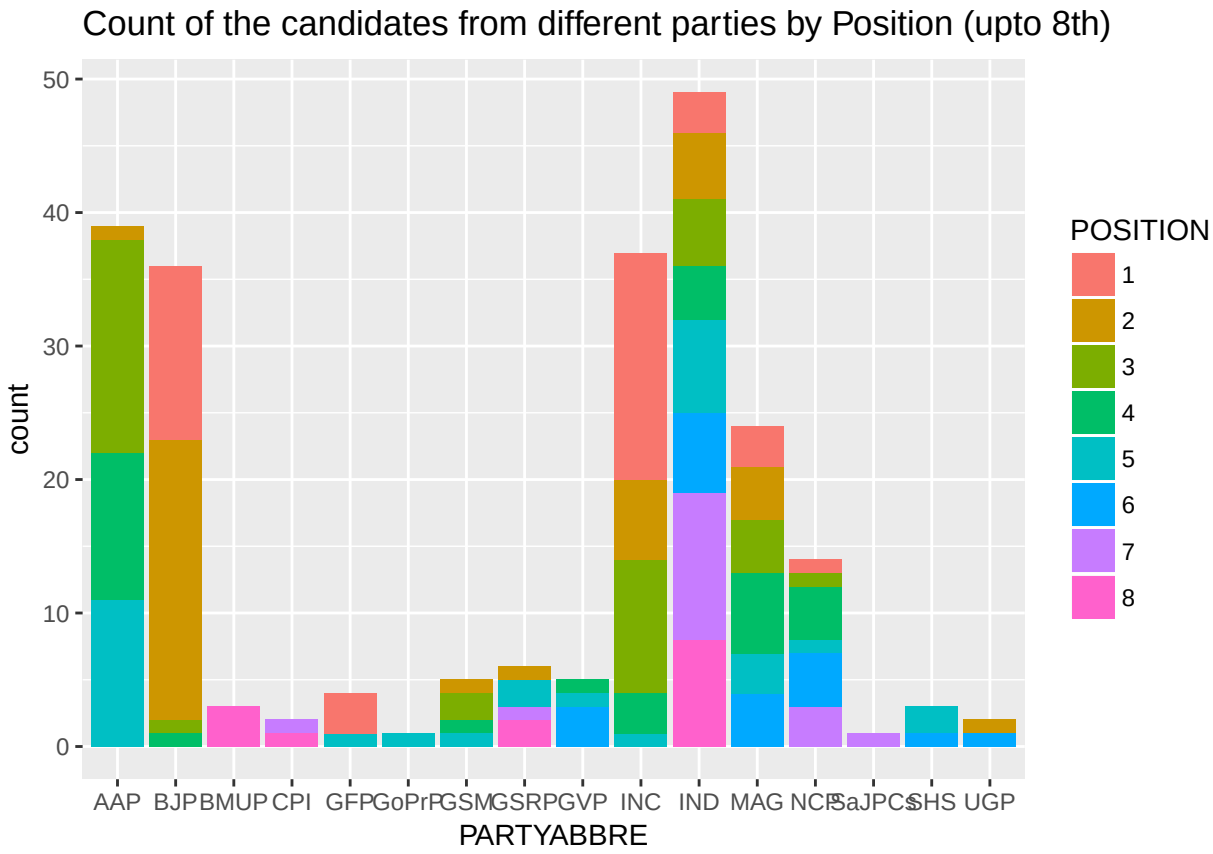
5. Exploratory Data Analysis

Count of the candidates from different parties by Position



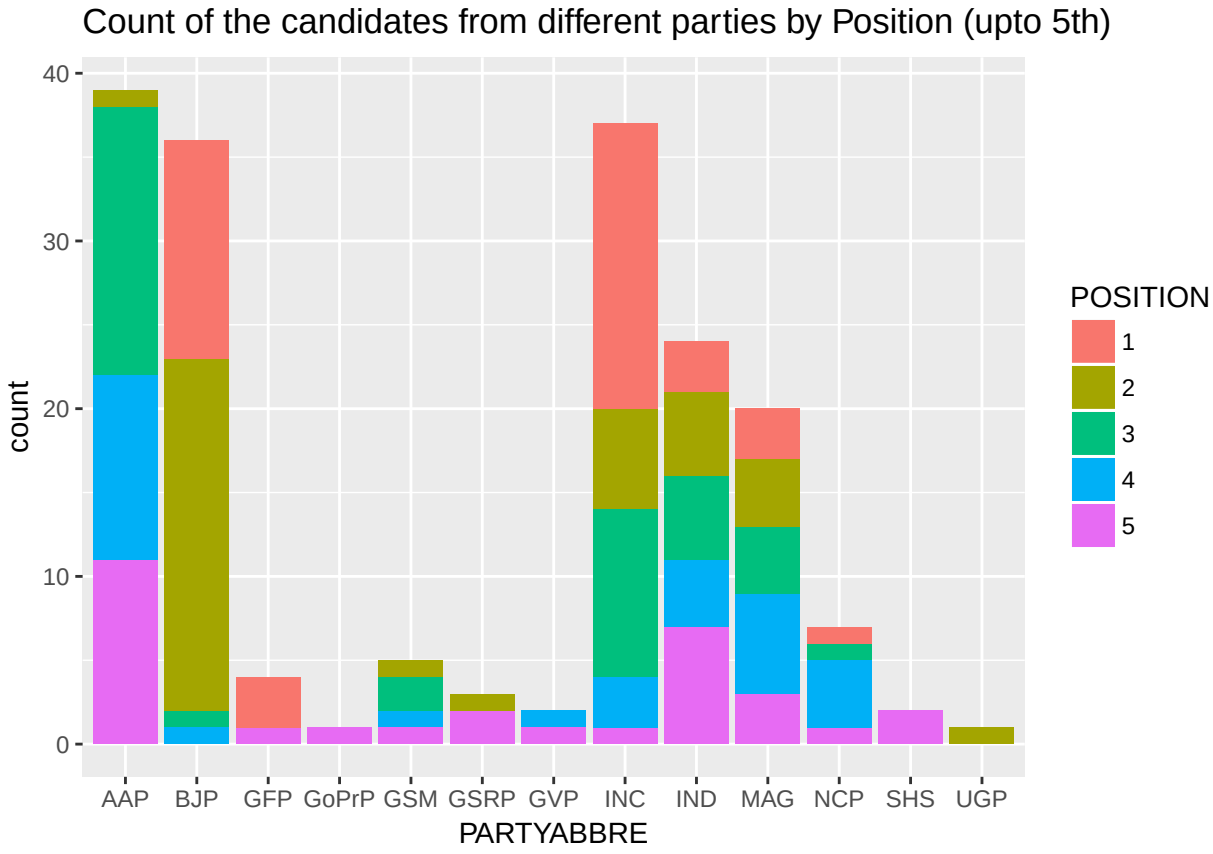
Observation:

- The dataset contains more observation pertaining to the independent candidates, and who seem to have very less of them to have won the elections, and relatively other party candidates have more observation but have less winner count also belong to regional parties.



Observation:

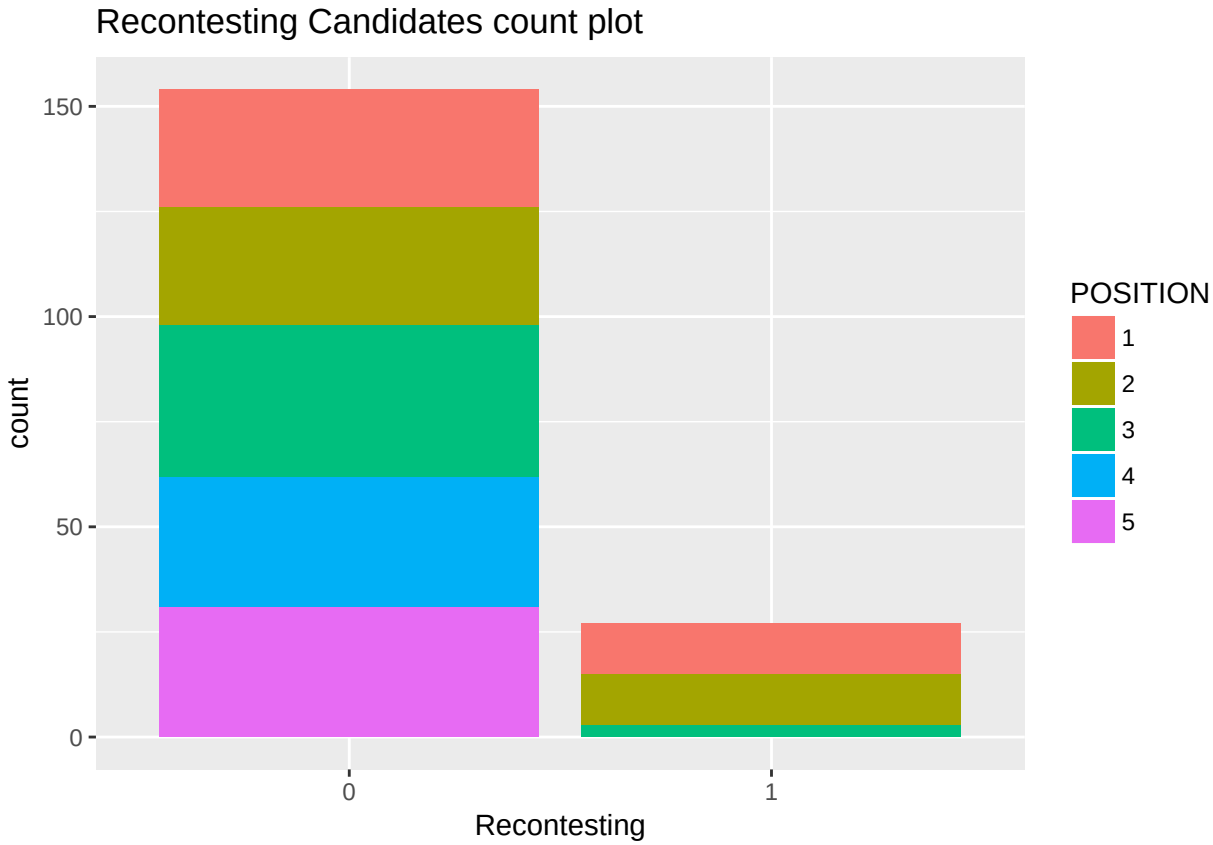
- The data only using the candidates who have secured a position of 8 or above, still has more candidates from independent group.



Observation:

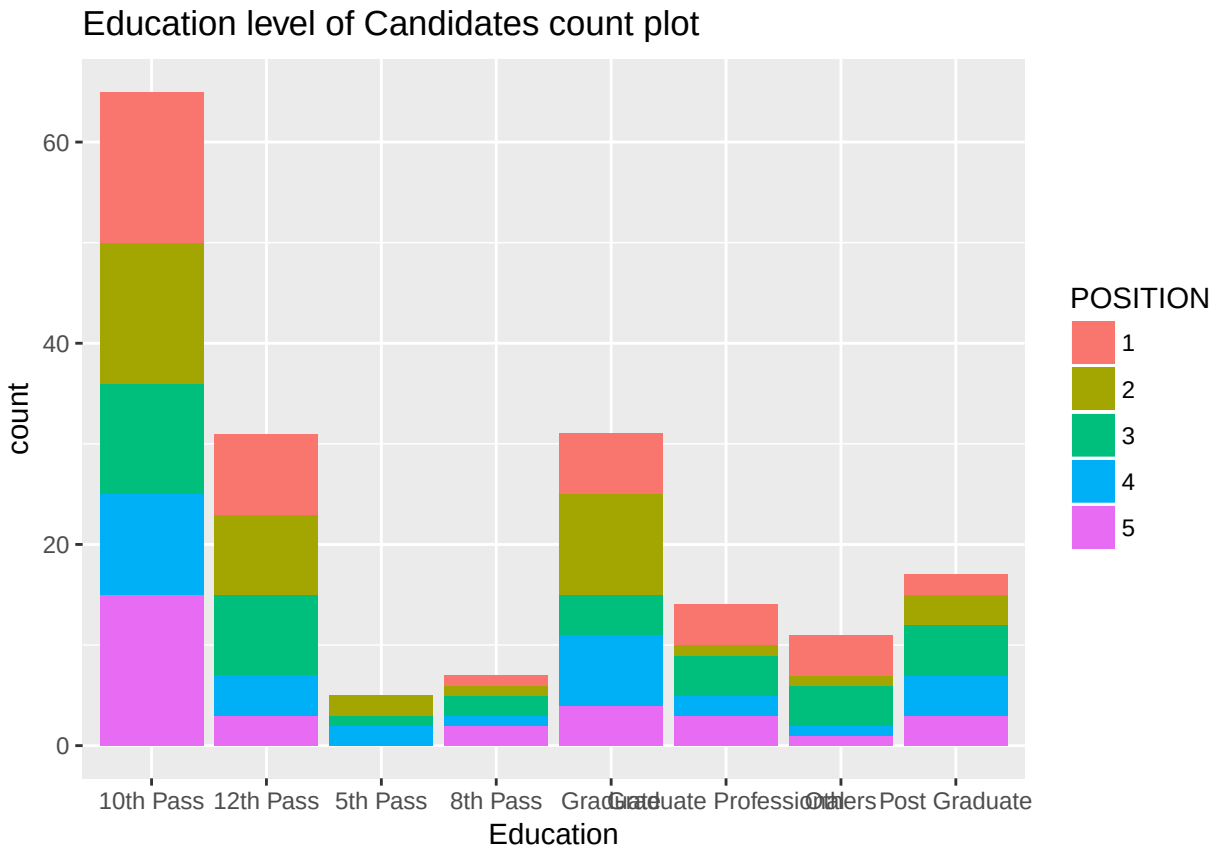
- The data only having candidates who have secured 5th position or above would be ideal, which would also closely target small group to make the data set to reflect those features more who have done fairly well, and by excluding higher position observations provides fewer groups for the model to classify

Count for other features by Position



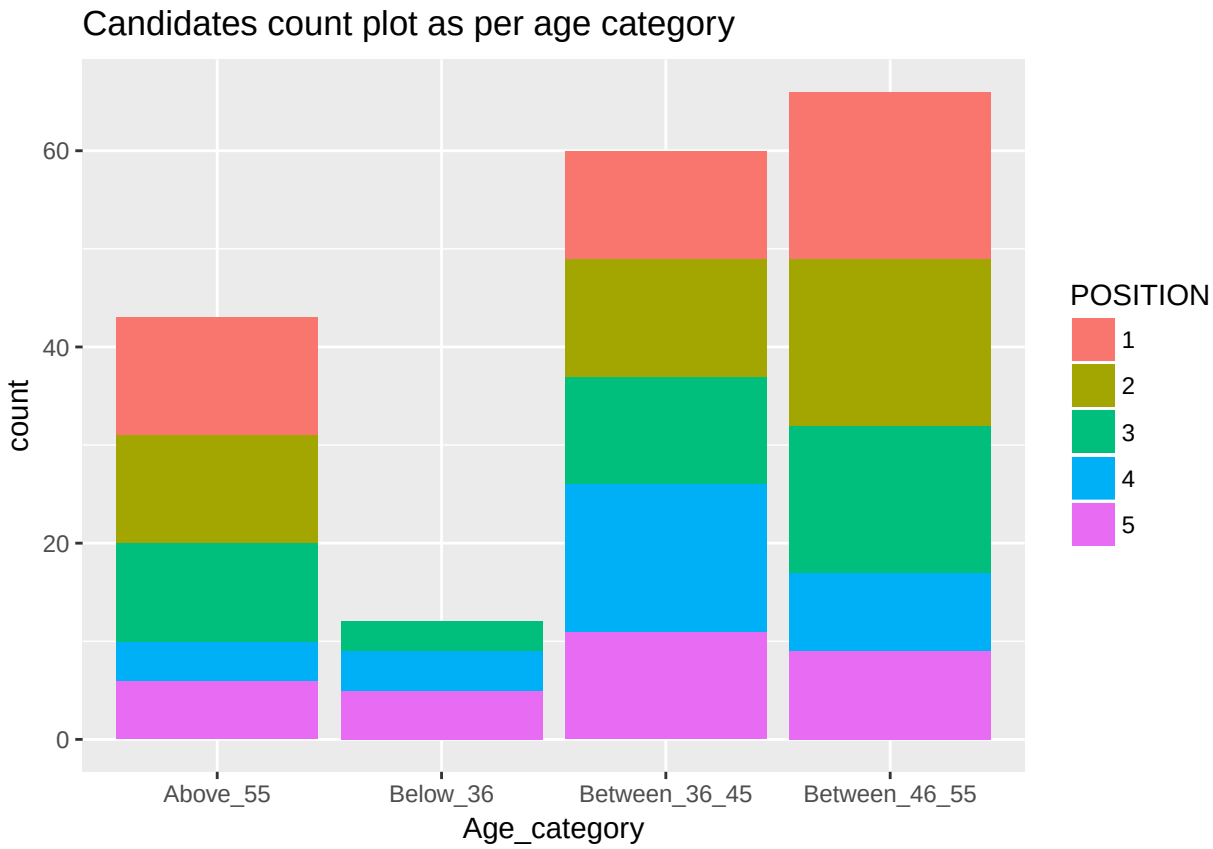
Observation:

- Mostly all of the candidates have recontested and close to half of them have been re-elected.



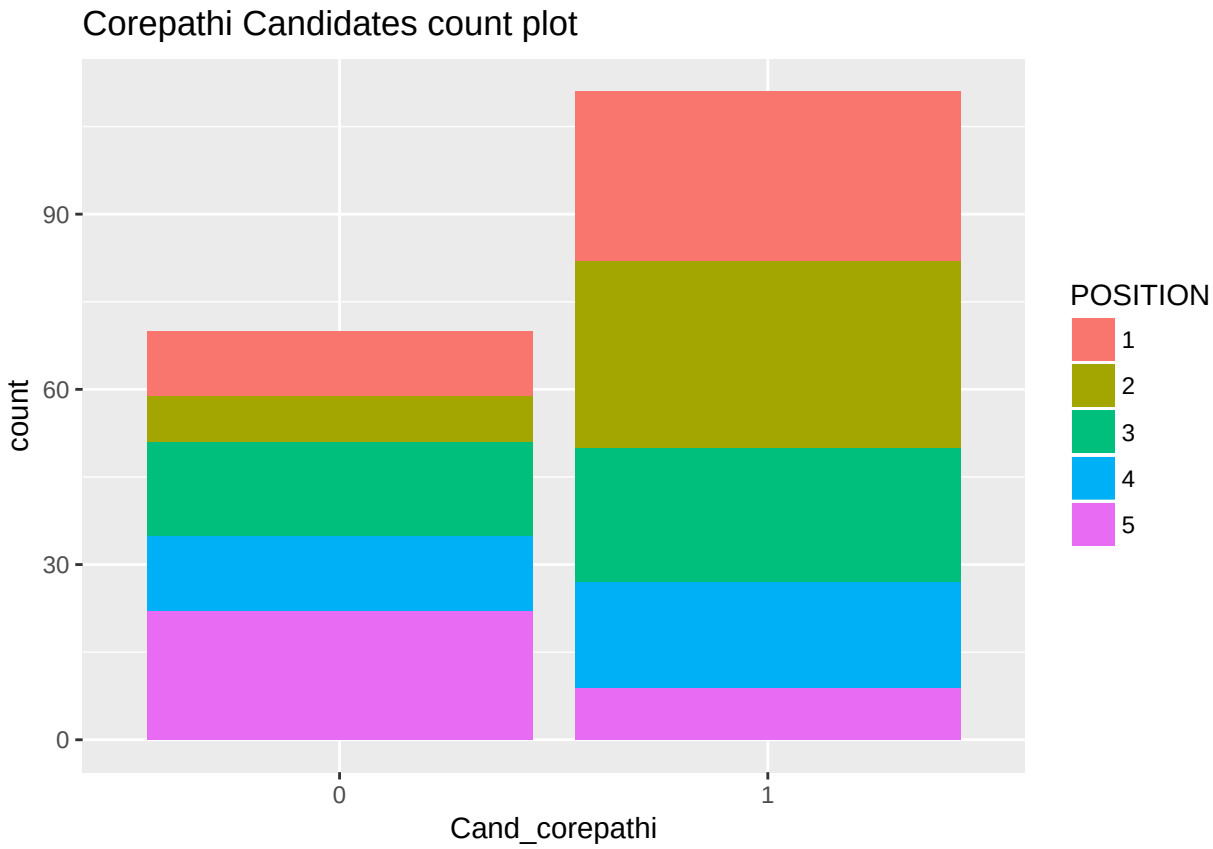
Observation:

- Without performing any data imputation the Goa state didn't have any candidate who were illiterate, and it might not be true with other states which might also have candidates who have not done some education. As Goa has a literacy rate of 88.70% and male literacy rate of 92.65% and female literacy rate of 84.66% as per 2011 population census



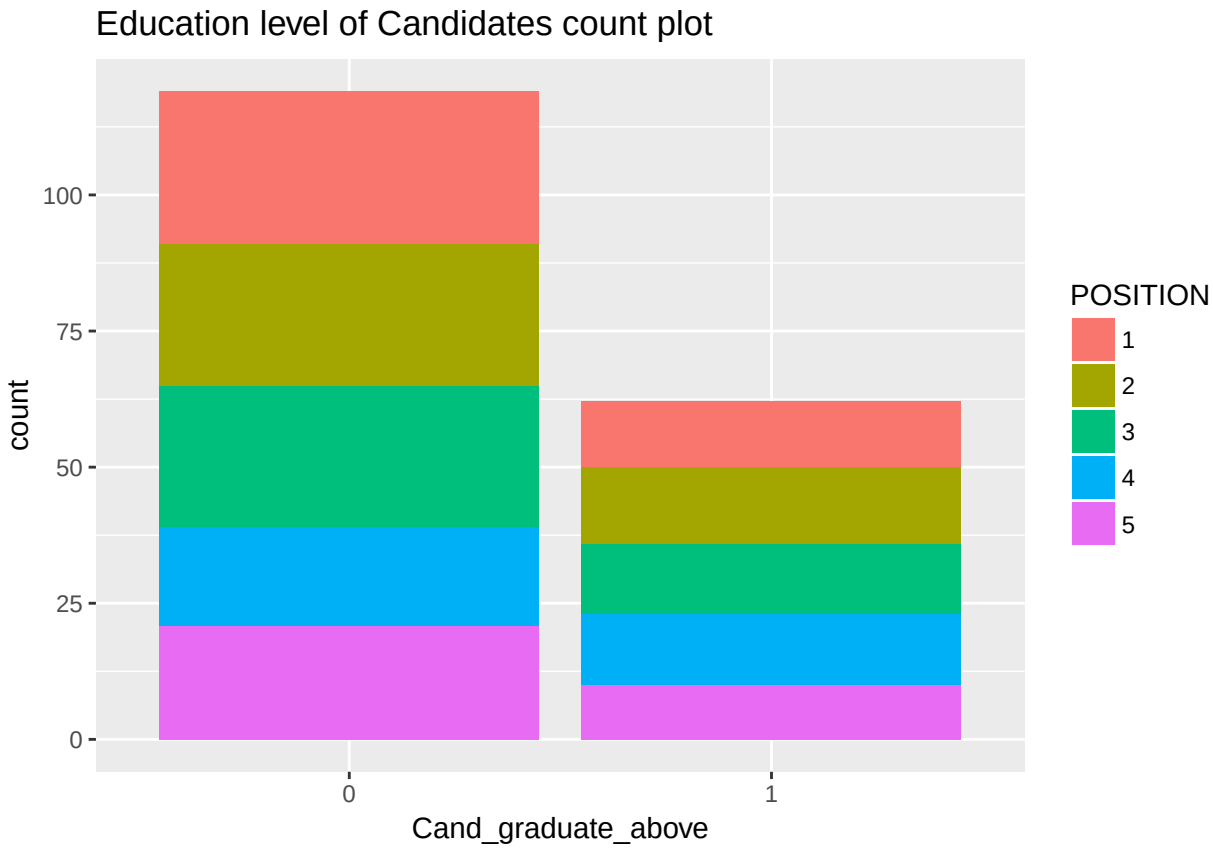
Observation:

- The age seems to play a significant role, considering young candidates have not been able to win and data has more observation belonging to candidates who are above 45 more than half of the other two age category.



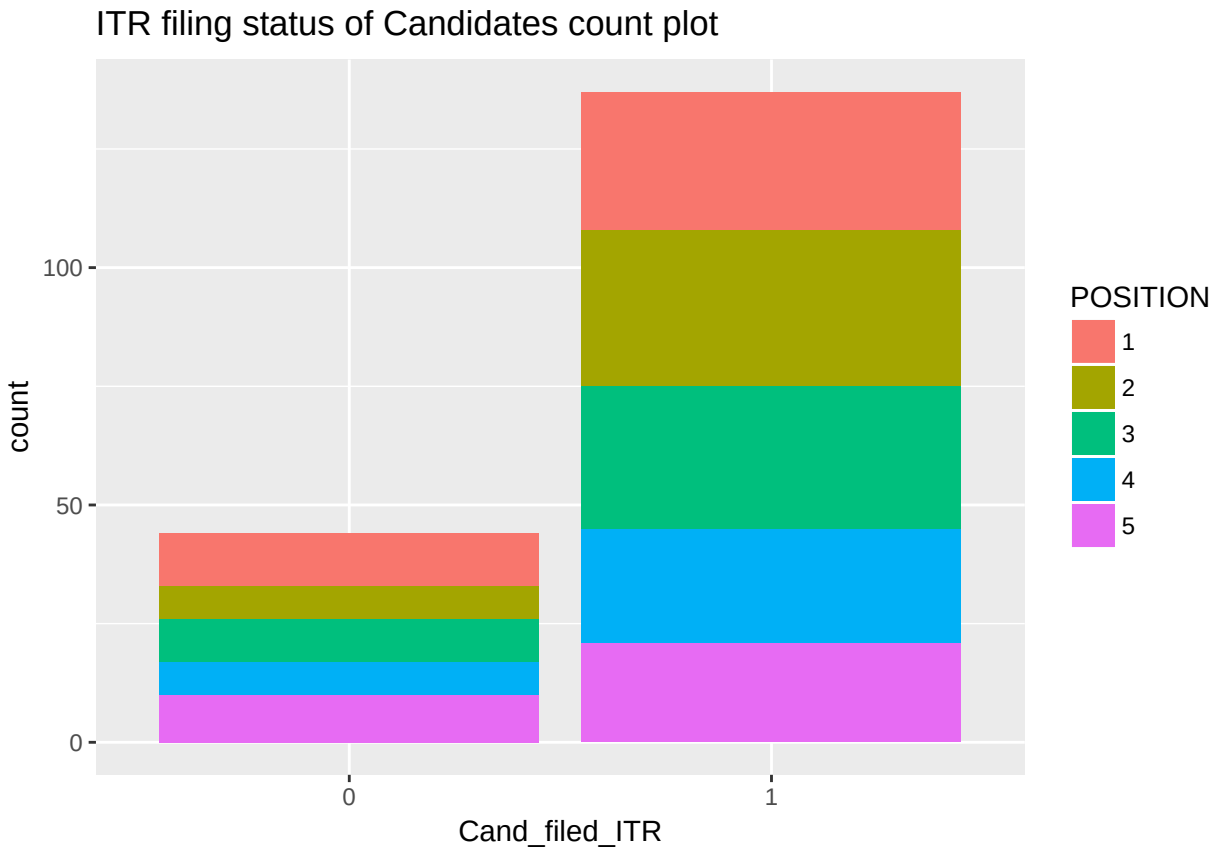
Observation:

- The candidate who are rich have done well in the elections, eventhough candidates who are not corepathis have won but those who have got lower position in more than compared to candidates who have more observations who have atleast managed to secure 2nd position. The corepathis candidates has fewer observations beloning to 5th position and would have been even less if all the postion would have considered.



Observation:

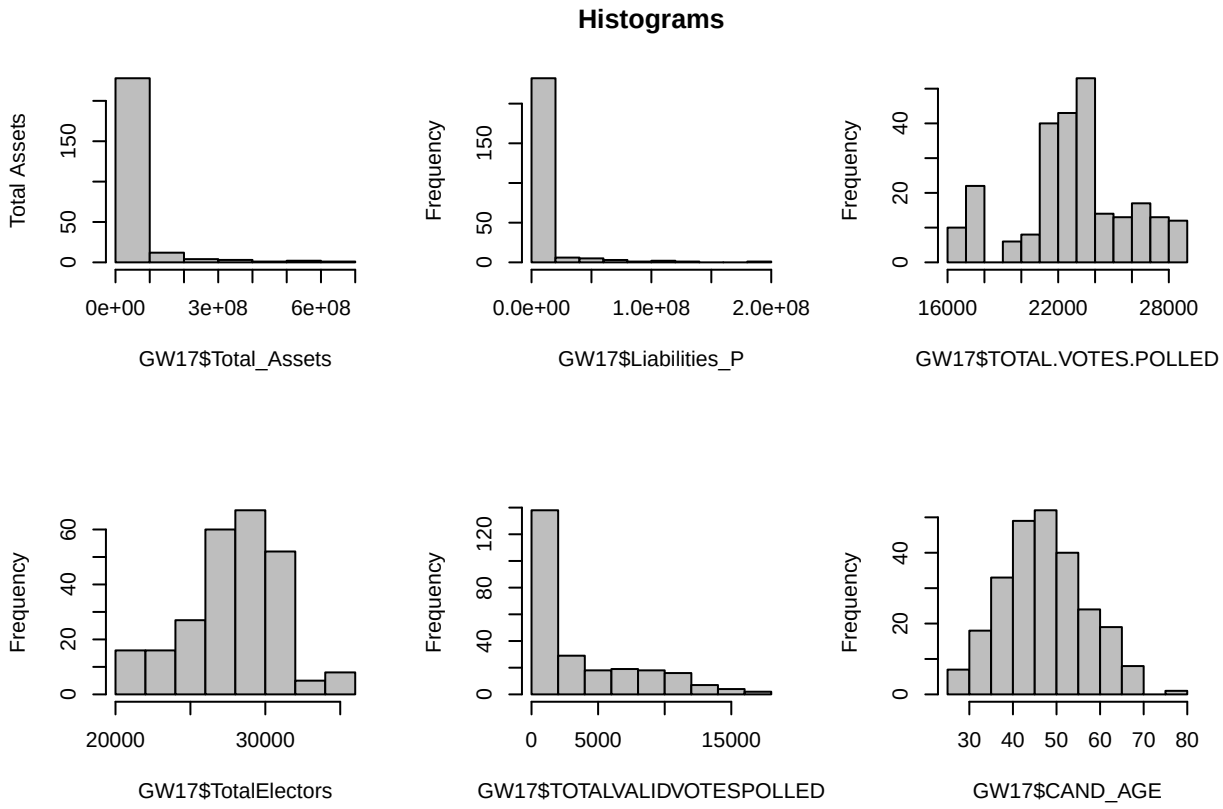
- Candidate having education level of Graduation/above might not be able to distinguish winning capacity of a candidate, as it seem to be equally likely



Observation:

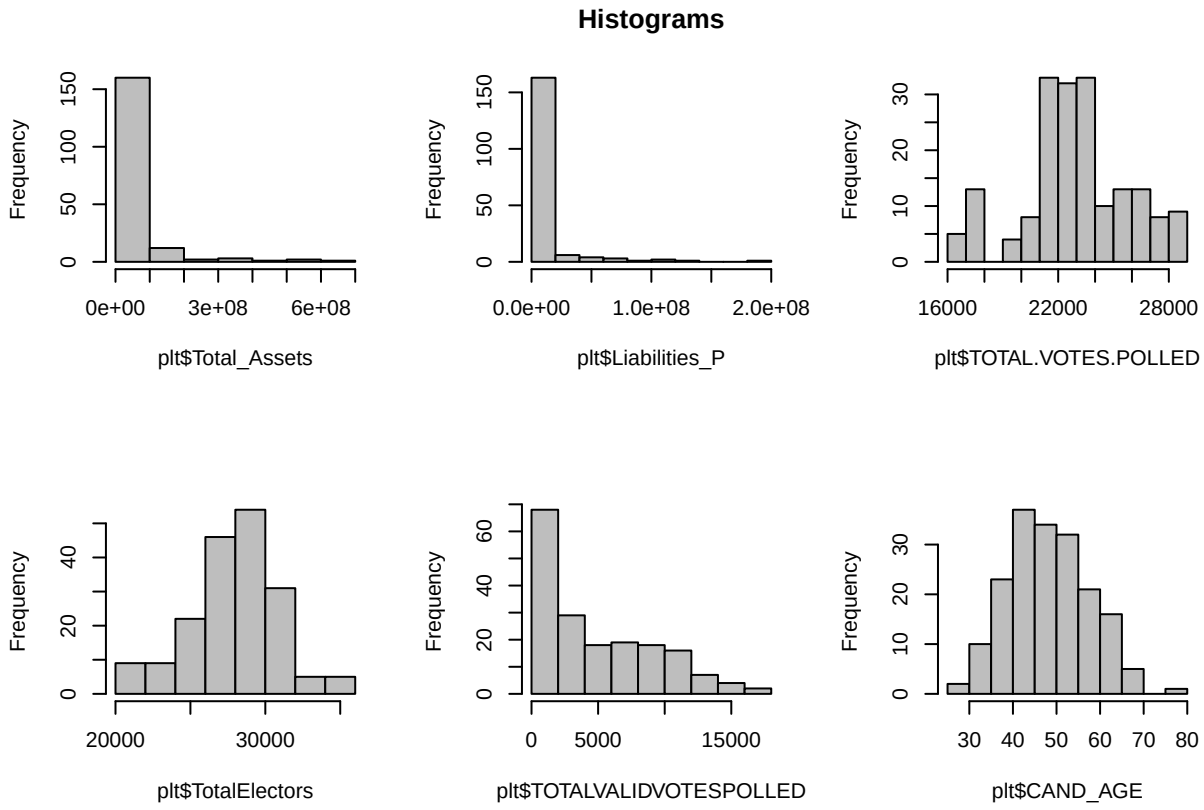
- This is an invisible characteristic of a candidate, unlike the corepathi, education, national party candidate. The electors will not be aware of this status, and less likely to influence the behavior of the electors.

Distribution of numeric features



Observation:

- The histograms from the total dataset including all the positions and candidates who have contested the election

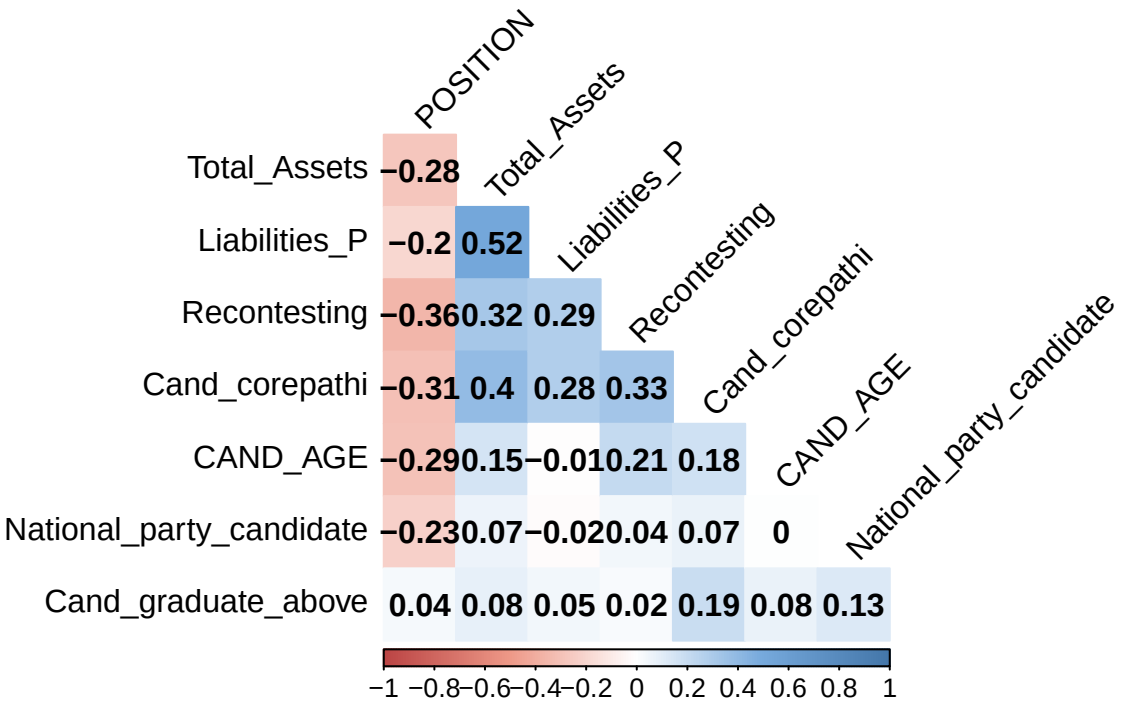


Observation:

- The distribution of the subset where only candidates securing above 6th place
- The distribution resembles closely the distribution in the total dataset, and reducing the observation by position has not changed the distributions significantly

```
Warning: package 'corrplot' was built under R version 3.4.4
```

```
corrplot 0.84 loaded
```



Observation:

- The four features has weak to moderate positive correlation with Position are Candidate belonging to national party, Candidate age, Candidate Wealth, Recontesting candidates
- Given that corepathi candidates have managed to win, means more the wealth more likely the positive result
- Since no feature has correlation more 0.5 there are no strongly visible features which might help the model to discriminate well and may not have good predictive power while classifying the candidates

6. Training classification models

6.1 Goa state Election results Model

Features in the dataset

```
 [1] "DIST_NAME"              "AC_NAME"
 [3] "AC_TYPE"                "CAND_NAME"
 [5] "POSITION"               "CAND_SEX"
 [7] "CAND_CATEGORY"          "CAND_AGE"
 [9] "PARTYABBRE"             "TOTALVALIDVOTESPOLLED"
[11] "TOTAL.VOTES.POLLED"     "TotalElectors"
[13] "Criminal.Case"          "Education"
[15] "Total_Assets"           "Liabilities_P"
[17] "Recontesting"           "National_party_candidate"
```

```
[19] "Winner_corepathi"         "Winner_graduate_above"
[21] "Winner_filed_ITR"         "Cand_corepathi"
[23] "Cand_graduate_above"      "Cand_filed_ITR"
[25] "Age_category"
```

Observation:

- The features which will used to traing the model

Data descrition

```
'data.frame':   181 obs. of  25 variables:
 $ DIST_NAME               : Factor w/ 2 levels "North Goa","South Goa": 1 1 1 1 1 1 1 1 1 1 ...
 $ AC_NAME                 : Factor w/ 40 levels "Aldona","Benaulim",..: 14 14 14 14 22 22 22 22 3 3 ..
 $ AC_TYPE                 : Factor w/ 2 levels "GEN","SC": 1 1 1 1 2 2 2 2 1 1 ...
 $ CAND_NAME               : Factor w/ 181 levels "ABHAY RAMCHANDRA PRABHU",..: 31 82 150 34 4 116 171 1
 $ POSITION                : Factor w/ 5 levels "1","2","3","4",..: 1 2 3 4 1 2 3 5 1 2 ...
 $ CAND_SEX                : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ CAND_CATEGORY           : Factor w/ 3 levels "GEN","SC","ST": 1 1 1 1 2 2 2 2 1 1 ...
 $ CAND_AGE                : int  53 60 69 53 63 62 43 40 53 51 ...
 $ PARTYABBRE              : Factor w/ 13 levels "AAP","BJP","GFP",..: 8 2 10 1 10 2 8 1 2 10 ...
 $ TOTALVALIDVOTESPOLLED   : int  16490 9371 678 620 15745 9715 1013 308 10654 9988 ...
 $ TOTAL.VOTES.POLLED      : int  28071 28071 28071 28071 27821 27821 27821 27821 23352 23352 ...
 $ TotalElectors           : int  31369 31369 31369 31369 31360 31360 31360 31360 25958 25958 ...
 $ Criminal.Case           : int  0 0 0 2 0 0 0 0 0 0 ...
 $ Education               : Factor w/ 8 levels "10th Pass","12th Pass",..: 5 8 1 5 1 5 7 1 2 2 ...
 $ Total_Assets            : int  35131000 89813996 27517393 9895320 1535342 25660021 50874189 338392 1
 $ Liabilities_P           : int  4550987 2910108 6669749 215000 0 11709338 7453460 300000 512780 44893
 $ Recontesting            : int  0 1 0 0 0 1 0 0 0 1 ...
 $ National_party_candidate: int  1 1 0 1 0 1 1 1 1 0 ...
 $ Winner_corepathi        : int  1 0 0 0 0 0 0 0 1 0 ...
 $ Winner_graduate_above   : int  1 0 0 0 0 0 0 0 0 0 ...
 $ Winner_filed_ITR        : int  1 0 0 0 0 0 0 0 1 0 ...
 $ Cand_corepathi          : int  1 1 1 0 0 1 1 0 1 1 ...
 $ Cand_graduate_above     : int  1 1 0 1 0 1 0 0 0 0 ...
 $ Cand_filed_ITR          : int  1 1 1 1 0 1 1 1 1 1 ...
 $ Age_category            : Factor w/ 4 levels "Above_55","Below_36",..: 4 1 1 4 1 1 3 3 4 4 ...
```

```
Warning: package 'caret' was built under R version 3.4.4
```

```
Attaching package: 'MASS'
```

```
The following object is masked from 'package:dplyr':

    select
```

```
Warning: package 'biotools' was built under R version 3.4.4
```

```
Warning: package 'rpanel' was built under R version 3.4.4
```

```
Package `rpanel', version 1.1-4: type help(rpanel) for summary information
```

```
Warning: package 'tkrplot' was built under R version 3.4.4
```

```
Warning: package 'SpatialEpi' was built under R version 3.4.4
```

```
---
biotools version 3.1


dummies-1.5.6 provided by Decision Patterns
```

Splitting the data into training and test set

```
set.seed(123)
training.samples <- plt.copy$POSITION %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data <- plt.copy[training.samples, ]
test.data <- plt.copy[-training.samples, ]
```

Scaling the dataset

```
# Estimate preprocessing parameters
preproc.param <- train.data %>%
  preProcess(method = c("center", "scale"))
# Transform the data using the estimated parameters
train.transformed <- preproc.param %>% predict(train.data)
test.transformed <- preproc.param %>% predict(test.data)
```

```
  DIST_NAME AC_NAME AC_TYPE                                CAND_NAME
1 North Goa Mandrem     GEN                 DAYANAND RAGHUNATH SOPTE
2 North Goa Mandrem     GEN                       LAXMIKANT PARSEKAR
3 North Goa Mandrem     GEN                    SHRIDHAR LADU MANJREKAR
4 North Goa Mandrem     GEN DEVENDRA KRISHNAJI PRABHU PARSEKAR DESAI
5 North Goa  Pernem      SC                 AJGAONKAR MANOHAR TRIMBAK
6 North Goa  Pernem      SC                       RAJENDRA ARLEKAR
  POSITION CAND_SEX CAND_CATEGORY  CAND_AGE PARTYABBRE
1   Winner        M           GEN 0.4230012        INC
2   Second        M           GEN 1.1557254        BJP
3    Third        M           GEN 2.0977994        MAG
4   Fourth        M           GEN 0.4230012        AAP
5   Winner        M            SC 1.4697501        MAG
6   Second        M            SC 1.3650752        BJP
  TOTALVALIDVOTESPOLLED TOTAL.VOTES.POLLED TotalElectors Criminal.Case
1             2.5752548           1.770983      1.189178    -0.3256195
2             0.9814958           1.770983      1.189178    -0.3256195
3            -0.9646408           1.770983      1.189178    -0.3256195
4            -0.9776255           1.770983      1.189178     3.3313384
5             2.4084687           1.682942      1.186118    -0.3256195
6             1.0585085           1.682942      1.186118    -0.3256195
      Education Total_Assets Liabilities_P Recontesting
1      Graduate   -0.1912621    -0.1951578   -0.3854628
2 Post Graduate    0.3456277    -0.2628300    2.5765148
3     10th Pass   -0.2660142    -0.1077770   -0.3854628
4      Graduate   -0.4390317    -0.3739801   -0.3854628
5     10th Pass   -0.5211118    -0.3828470   -0.3854628
6      Graduate   -0.2842503     0.1000629    2.5765148
  National_party_candidate Winner_corepathi Winner_graduate_above
1                0.7412169        2.3046035             3.6751666
```

```
2                0.7412169       -0.4309421            -0.2702328
3               -1.3398921       -0.4309421            -0.2702328
4                0.7412169       -0.4309421            -0.2702328
5               -1.3398921       -0.4309421            -0.2702328
6                0.7412169       -0.4309421            -0.2702328
  Winner_filed_ITR Cand_corepathi Cand_graduate_above Cand_filed_ITR
1        2.3046035      0.7747618           1.4463595      0.5806243
2       -0.4309421      0.7747618           1.4463595      0.5806243
3       -0.4309421      0.7747618          -0.6866555      0.5806243
4       -0.4309421     -1.2818787           1.4463595      0.5806243
5       -0.4309421     -1.2818787          -0.6866555     -1.7104877
6       -0.4309421      0.7747618           1.4463595      0.5806243
   Age_category
1 Between_46_55
2      Above_55
3      Above_55
4 Between_46_55
5      Above_55
6      Above_55
```

Observation:

- The scaled data for training the model

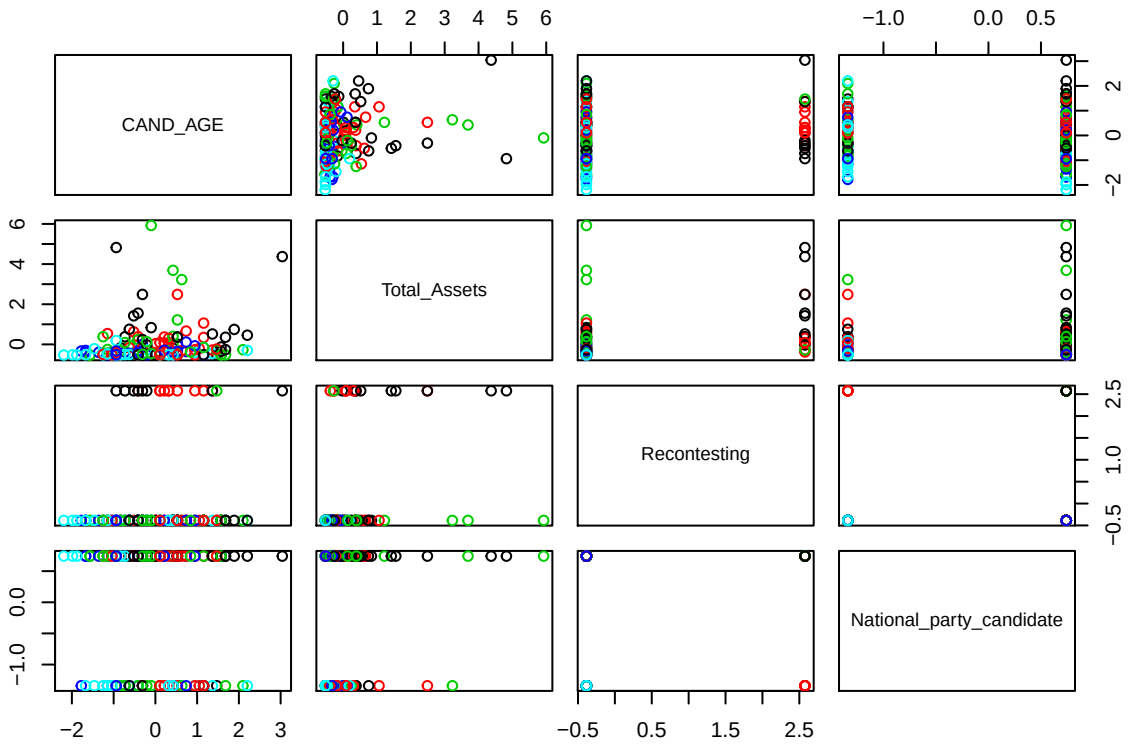
Traing the Linear Discriminant Analysis model

```
model <- lda(POSITION~ ., data = mod_sub)
```

Observation:

- The subset of the features used to guage the performance based on the selected features

Plot based on four important features



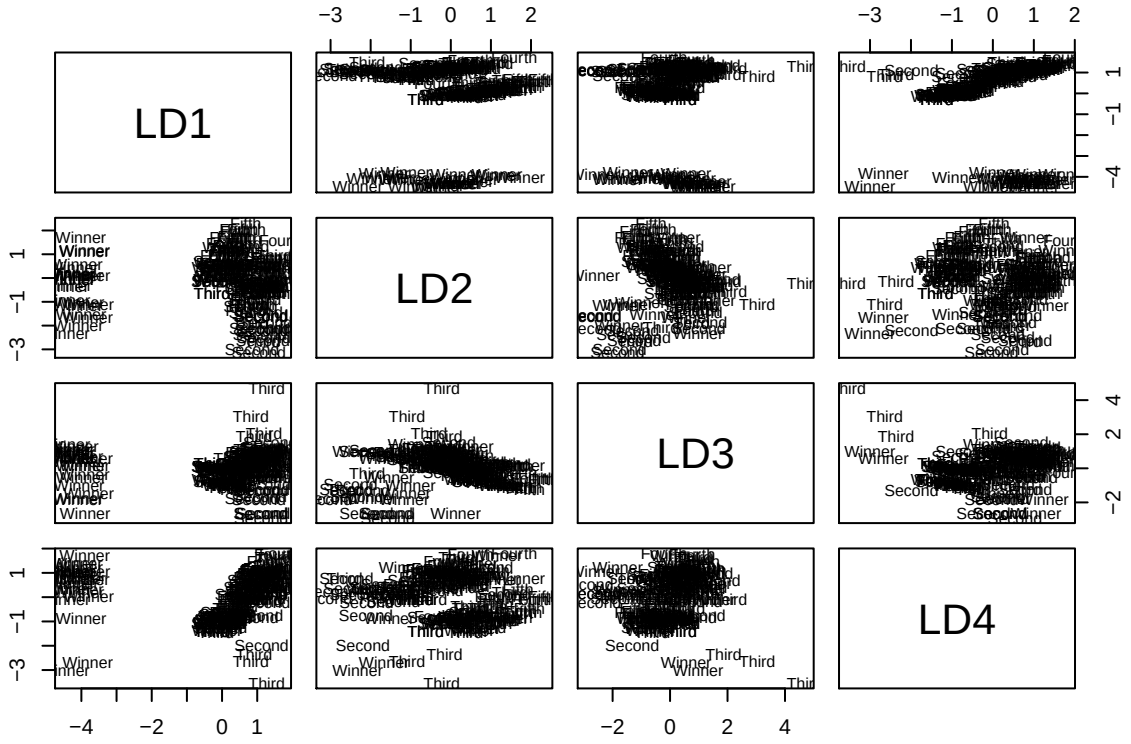
Observation:

- The LDA model will use combination of above features and other to find which combination of features better separates the group, winner group from other runner ups

Model results

```
Call:
lda(POSITION ~ ., data = mod_sub)

Prior probabilities of groups:
   Winner    Second     Third    Fourth     Fifth
0.2191781 0.2191781 0.2191781 0.1712329 0.1712329

Group means:
          CAND_AGE Total_Assets Recontesting National_party_candidate
Winner  0.46225432  0.420685512    0.3550316               0.35100897
Second  0.16131403 -0.004669884    0.5401552               0.02583569
Third   0.03374152  0.201797343   -0.2929010               0.15590500
Fourth -0.34740590 -0.358891264   -0.3854628              -0.09122670
Fifth  -0.49395073 -0.431909340   -0.3854628              -0.59069286
       Cand_corepathi Cand_graduate_above Criminal.Case Winner_corepathi
Winner    0.196331694         -0.02008833    0.13150020        1.5352313
```

```
Second    0.260601710         -0.15340176    0.24578014       -0.4309421
Third     0.003521645          0.04656839    0.01722026       -0.4309421
Fourth    0.116636873          0.25187108   -0.17934123       -0.4309421
Fifth    -0.706019335         -0.08941131   -0.32561955       -0.4309421
       Winner_graduate_above
Winner              0.9627045
Second             -0.2702328
Third              -0.2702328
Fourth             -0.2702328
Fifth              -0.2702328

Coefficients of linear discriminants:
                                    LD1           LD2           LD3
CAND_AGE                  -0.167866966 -0.477016580  0.1802571868
Total_Assets              -0.005535803  0.004068454  0.5858112491
Recontesting               0.033581019 -0.557115198 -0.8664083840
National_party_candidate -0.059602882 -0.498092711  0.4515721076
Cand_corepathi             0.452827164 -0.501097514  0.2320631791
Cand_graduate_above        0.176540265  0.283640964  0.2081837257
Criminal.Case              0.050945174 -0.236088050 -0.0886358790
Winner_corepathi          -1.882577392  0.471185551 -0.0006519706
Winner_graduate_above     -0.045987294 -0.058954033 -0.2042155897
                                    LD4
CAND_AGE                  -0.296805364
Total_Assets              -0.777082319
Recontesting              -0.017631131
National_party_candidate   0.009968697
Cand_corepathi             0.821118753
Cand_graduate_above        0.317133129
Criminal.Case             -0.097208242
Winner_corepathi           0.312029540
Winner_graduate_above     -0.134624143

Proportion of trace:
   LD1    LD2    LD3    LD4
0.8305 0.1095 0.0397 0.0203
```



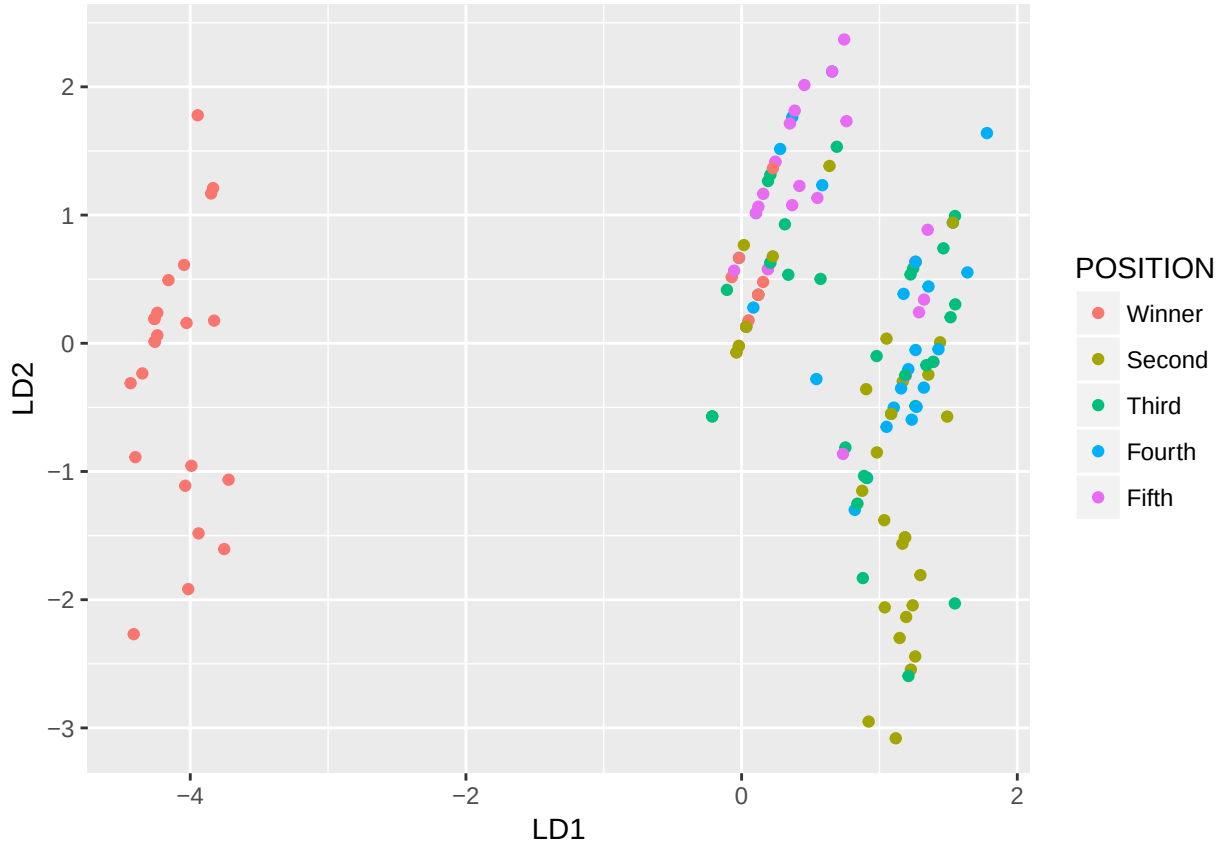
Observation:

- The model classification is not accurate and finding difficult to create proper groupings, and the accuray would not be very high

Model outputs

```
Predicted classes(head)

[1] Winner Third  Second Winner Third  Third
Levels: Winner Second Third Fourth Fifth

Predicted probabilities of class membership:

         Winner       Second        Third       Fourth        Fifth
18 9.999602e-01 2.859432e-05 2.869289e-06 2.426706e-06 5.874002e-06
20 1.245495e-04 3.141039e-01 3.237829e-01 3.058082e-01 5.618051e-02
24 5.281875e-05 9.505913e-01 3.294633e-02 1.367239e-02 2.737108e-03
28 9.998029e-01 2.632357e-06 8.681956e-06 2.685969e-05 1.588836e-04
37 1.163681e-04 2.794788e-01 3.810405e-01 2.822254e-01 5.713901e-02
42 5.546643e-03 1.973257e-01 3.826603e-01 1.433669e-01 2.711005e-01

Linear Discriminants:

          LD1        LD2        LD3        LD4
18 -4.1036812 -1.4884961 -2.0151438 1.06861275
20  1.0687100 -0.6015288  0.4881925 0.79626239
```

```
24  0.8856867 -3.0496086 -1.6337345 0.05797513
```



Observation:

- The discrimination of the classes based on first two LDAs
- The model is able to discriminate well between the winner and loosing candidates. The model is considering the candidates who have not managed to win to be closely similar to each other and distinct from the winners

Model Accuracy

```
mean(predictions$class==test.transformed$POSITION)
```

```
[1] 0.4
```

Cross table of the prediction and actual label

```
table(predictions$class, test.transformed$POSITION)
```

```
        Winner Second Third Fourth Fifth
  Winner      6      0     0      0     0
  Second      0      2     2      1     0
  Third       1      3     4      0     4
```

```
Fourth      0      3      0      0      0
Fifth       1      0      1      5      2
```

Observation:

- As expected the accuray of the model is not that good, the model is not able to discriminate well among the data. But the it is able to distinguish between the winner, 1st position, and other remaining positions. The model is treating the candidates who have not won to be closely related in characteristic, then analysing which are the features that winners have that other remaining position would allow to identify important features which help to distinguish winners from candidates who have not managed to win. And one of the feature which can be identified is whether a candidate belongs to national party, then it is hightly probably that candidate would win and second would be if the candidate is recontesting feature
- From the table it is clear that no one group has high correctly predicted results, except the winner group to some extent, all are scattered between the classes and winner has been predicted correctly for 6 observation out of the 8 observation in the sample single observation, which gives and accuracy for the winner group of 0.75 higher than among the other groups
- Using the cross validation might help to improve the accuracy

Training the model with cross validation

```
        Winner Second Third Fourth Fifth
Winner      23      0     0      0     0
Second       0     10     5      2     1
Third        5     13    12      6     1
Fourth       0      5     7     11     6
Fifth        4      4     8      6    17
```

Observation:

- The Accuracy for the within winner group is 0.71 which is less than what the model of able to achieve without cross validation on the test set
- The cross validation is able to do some what better for the other groups, but the results are not great. Given it has been not able to classify well the features doesn't have enough predictive which directly contributes in correctly classifying the winners

Removing winner related features which hightly correlated

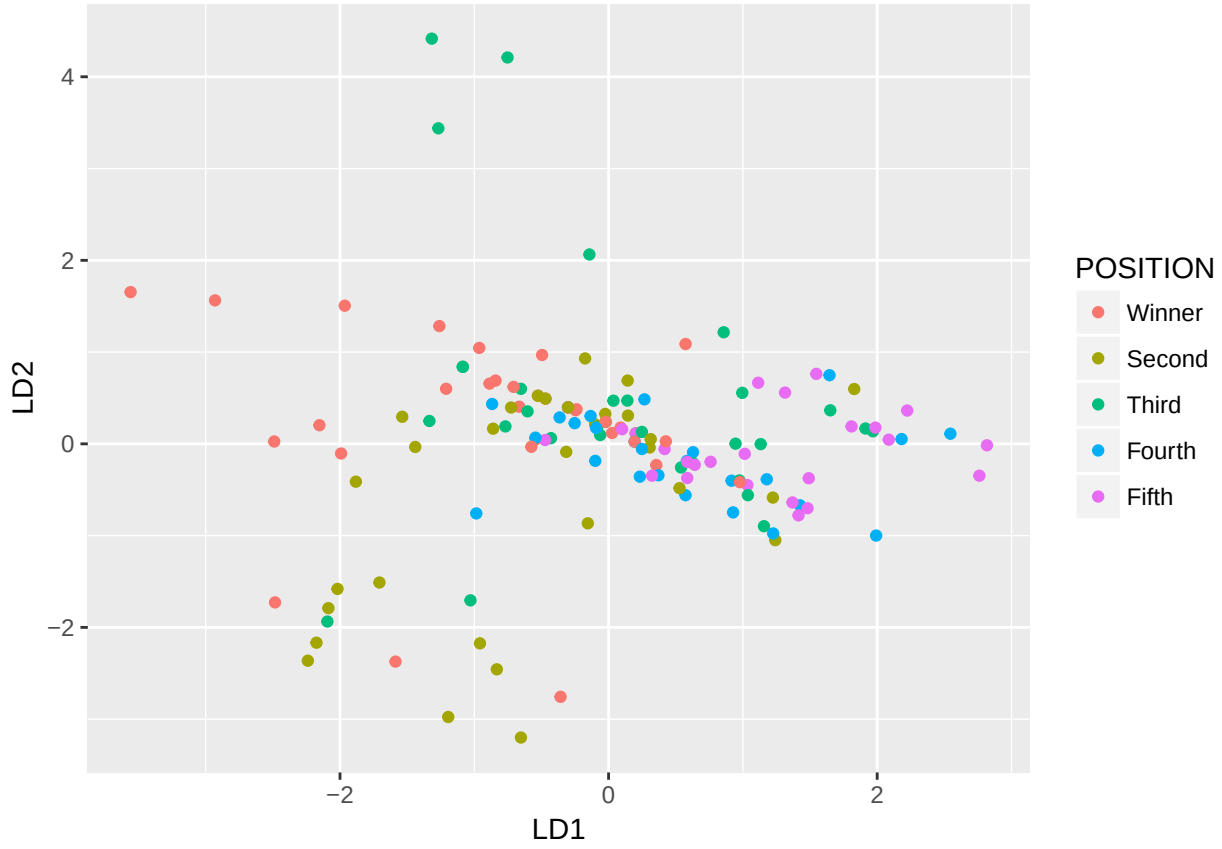
```
# The winner related features are removed which are highly correlated with
# winner results and it shadows which would not be available while predicting
# in real scenario
plt.copy <- plt[,-c(1,2,3,4,6,7,9,10,11,12,14,19,20,21,25)]
```

Model results

```
Call:
lda(POSITION ~ ., data = train.transformed)

Prior probabilities of groups:
   Winner    Second     Third    Fourth     Fifth
```

```
0.2191781 0.2191781 0.2191781 0.1712329 0.1712329

Group means:
          CAND_AGE Criminal.Case Total_Assets Liabilities_P Recontesting
Winner  0.46225432    0.13150020  0.420685512   0.340013947    0.3550316
Second  0.16131403    0.24578014 -0.004669884   0.004829101    0.5401552
Third   0.03374152    0.01722026  0.201797343   0.087052510   -0.2929010
Fourth -0.34740590   -0.17934123 -0.358891264  -0.308152743   -0.3854628
Fifth  -0.49395073   -0.32561955 -0.431909340  -0.244673571   -0.3854628
       National_party_candidate Cand_corepathi Cand_graduate_above
Winner               0.35100897    0.196331694         -0.02008833
Second               0.02583569    0.260601710         -0.15340176
Third                0.15590500    0.003521645          0.04656839
Fourth              -0.09122670    0.116636873          0.25187108
Fifth               -0.59069286   -0.706019335         -0.08941131
       Cand_filed_ITR
Winner    -0.06375098
Second     0.07944352
Third      0.07944352
Fourth     0.12240187
Fifth     -0.24417604

Coefficients of linear discriminants:
                                LD1        LD2          LD3         LD4
CAND_AGE                 -0.5347336  0.3012039 -0.261436696  0.02007683
Criminal.Case            -0.1799143 -0.1800657  0.002715231 -0.51071213
Total_Assets             -0.1317400  0.6599204 -0.187573683 -0.21206255
Liabilities_P            -0.1069876  0.3936591 -0.237756596  0.29376317
Recontesting             -0.4472566 -0.8460172 -0.443419453  0.21228593
National_party_candidate -0.5071986  0.3314490  0.227600604 -0.04366676
Cand_corepathi           -0.5405755 -0.4865225  1.098262781  0.35439150
Cand_graduate_above       0.1643864  0.0536216  0.289139766  0.64651789
Cand_filed_ITR            0.4609319  0.2089770 -0.294072706 -0.91613148

Proportion of trace:
   LD1    LD2    LD3    LD4
0.6705 0.1925 0.1223 0.0148
```

Model accuracy




```
[1] 0.2571429
```

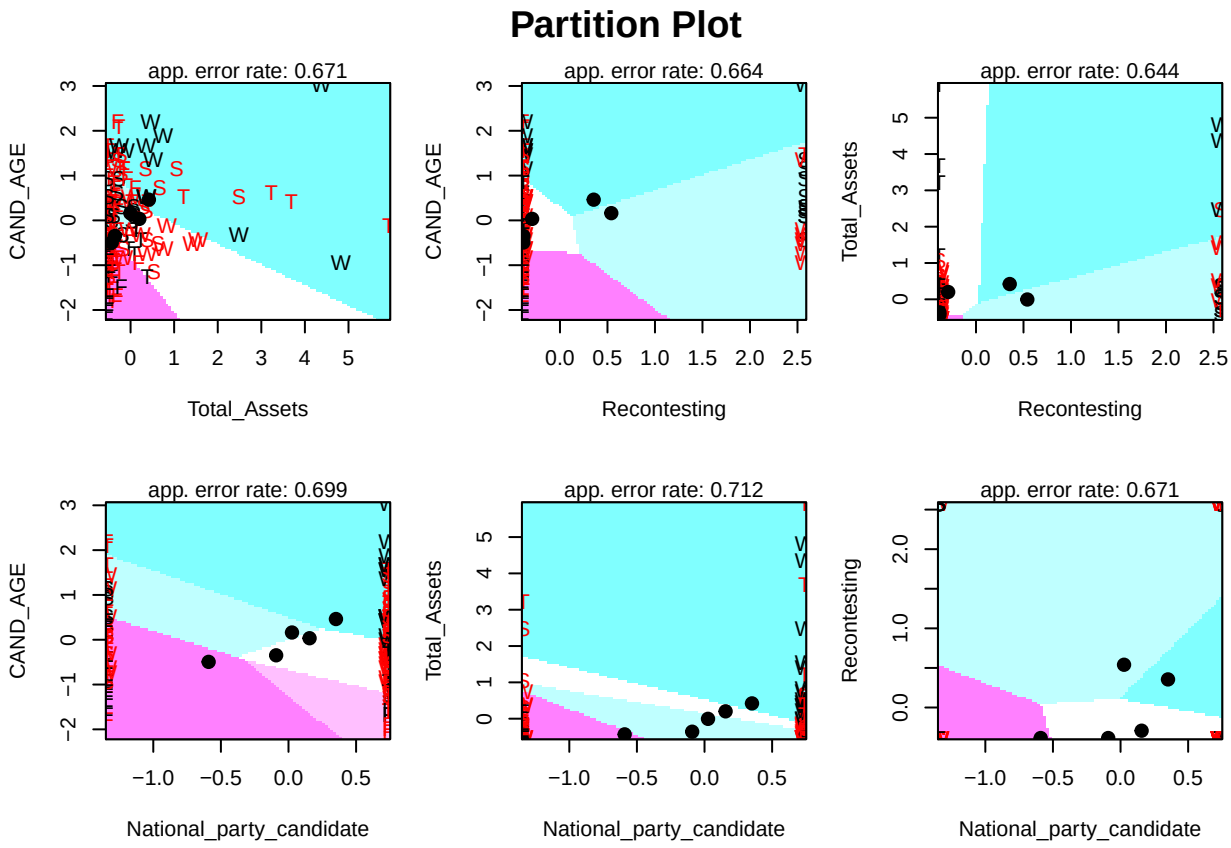
	Winner	Second	Third	Fourth	Fifth
Winner	1	1	2	0	1
Second	4	1	2	1	0
Third	0	3	2	1	1
Fourth	2	3	0	1	0
Fifth	1	0	1	3	4

Observation:

- The accuracy has droped as the winner related features was shadowing the winners which resulted that the model was able to distinguish only the winner group from the others and finding difficult to distinguish other groups from each other. Though at first time winner related features was included thinking that it is a good indicator.
- But after thinking it through why only the winner is differentiated from others, then winner_corepathi, winner_graduate_above and winner_filed_ITR was actually a statistic after the results has been announced, which would not be known before the results. And would have been a error to build a model including those features.

Plot for pairs of features and error rates given by LDA

```
Warning: package 'klaR' was built under R version 3.4.4
```



Observation:

- No two pairs is able to get an accuracy of even 0.6 from the choosen subset of features, checking using more features to improve the model would be to see if model is able to find some combination of features to correctly discriminate the groups
- The model is not able to distinguish the winner from others and within group accuracy is also very low, the features are not to able to signify the relation which a winning candidate would have, given the data it is a general information of the candidate which fails to lend any significant indicator other than it being a statistic of the election, which convesy the proportion of the results in different category
- A more granular data like sentiment of the electors, GDP growth of the district, state under the previous elected candidate would be required to get some actual sense of the voting for a particular candidate
- But as per initial exploratory data analysis if three key features which can be attributed to have an appeal to the electors would be Recontesting candidate, National party candidate, Political experience (Candidate age), and additional candidate wealth somehow indirectly makes a candidate to have power who can fulfill the needs of the electors

6.2 Running the model for the Punjab state

```
  DIST_NAME  AC_NAME AC_TYPE        CAND_NAME POSITION CAND_SEX
1 Pathankot Sujanpur     GEN     DINESH SINGH        1        M
2 Pathankot Sujanpur     GEN       AMIT SINGH        2        M
3 Pathankot Sujanpur     GEN      NARESH PURI        3        M
4 Pathankot Sujanpur     GEN      NATHA SINGH        4        M
5 Pathankot Sujanpur     GEN KULBHUSHAN SINGH        5        M
```

```
6 Pathankot Sujanpur     GEN    KARNAIL CHAND        6        M
  CAND_CATEGORY CAND_AGE PARTYABBRE TOTALVALIDVOTESPOLLED
1           GEN       54        BJP                 48910
2           GEN       40        INC                 30209
3           GEN       47        IND                 28675
4           GEN       62      RMPOI                 10581
5           GEN       44       AAAP                  2831
6            SC       58        BSP                  1083
  TOTAL.VOTES.POLLED TotalElectors Criminal.Case     Education
1             125616        159005             0     12th Pass
2             125616        159005             0     12th Pass
3             125616        159005             0      Graduate
4             125616        159005             0     10th Pass
5             125616        159005             0 Post Graduate
6             125616        159005             0      5th Pass
  Total_Assets Liabilities_P Recontesting National_party_candidate
1     61992943       1630462            1                        1
2    105210992        825636            0                        1
3    278907766        909434            0                        0
4      9504078        876400            0                        0
5     11567613        241543            0                        1
6     38590000       1401750            0                        1
  Winner_corepathi Winner_graduate_above Winner_filed_ITR Cand_corepathi
1                1                     0                1              1
2                0                     0                0              1
3                0                     0                0              1
4                0                     0                0              0
5                0                     0                0              1
6                0                     0                0              1
  Cand_graduate_above Cand_filed_ITR  Age_category
1                   0              1 Between_46_55
2                   0              1 Between_36_45
3                   1              1 Between_46_55
4                   0              1      Above_55
5                   1              1 Between_36_45
6                   0              1      Above_55
```

Punjab state wrangled data set

```
plt.copy <- subset(punjab, punjab$POSITION<=5)
plt.copy$POSITION <- as.factor(plt.copy$POSITION)

plt.copy$POSITION <- factor(plt.copy$POSITION, levels = c(1:5), labels = c("Winner", "Second", "Third",

colnames(plt.copy)
```

```
 [1] "DIST_NAME"               "AC_NAME"
 [3] "AC_TYPE"                 "CAND_NAME"
 [5] "POSITION"                "CAND_SEX"
 [7] "CAND_CATEGORY"           "CAND_AGE"
 [9] "PARTYABBRE"              "TOTALVALIDVOTESPOLLED"
[11] "TOTAL.VOTES.POLLED"      "TotalElectors"
[13] "Criminal.Case"           "Education"
```

```
[15] "Total_Assets"             "Liabilities_P"
[17] "Recontesting"             "National_party_candidate"
[19] "Winner_corepathi"         "Winner_graduate_above"
[21] "Winner_filed_ITR"         "Cand_corepathi"
[23] "Cand_graduate_above"      "Cand_filed_ITR"
[25] "Age_category"
```

subset of features for training the model

Model output for punjab state

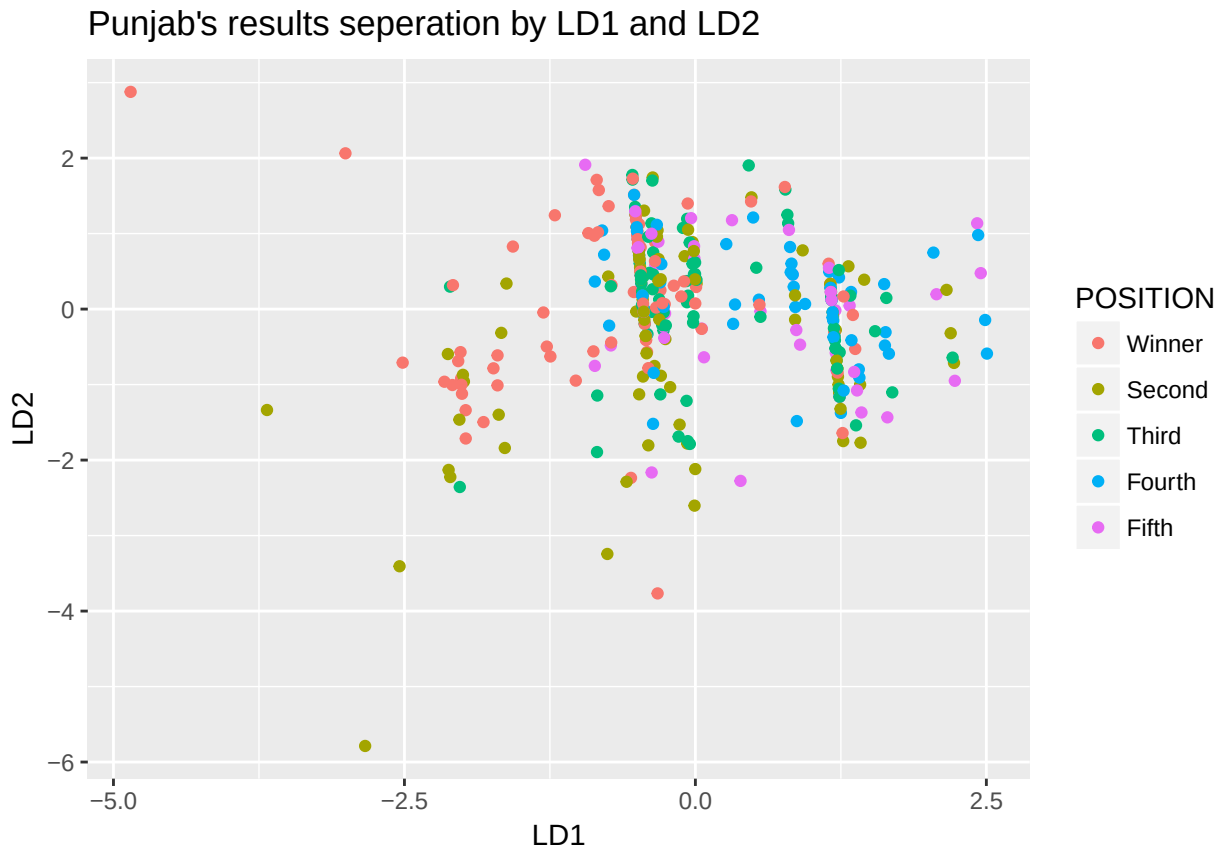
```
model <- lda(POSITION~ ., data = train.transformed)

model
```

```
Call:
lda(POSITION ~ ., data = train.transformed)

Prior probabilities of groups:
   Winner    Second     Third    Fourth     Fifth
0.2425068 0.2316076 0.2288828 0.1580381 0.1389646

Group means:
          CAND_AGE Criminal.Case Total_Assets Liabilities_P Recontesting
Winner -0.13621166  0.1781105568   0.17748413   0.187722138    0.2248641
Second  0.30051635 -0.0008084923   0.16366665   0.002721011    0.2883583
Third  -0.12591579 -0.1482837490  -0.03326242  -0.056540365   -0.1540166
Fourth -0.10938455 -0.0059242972  -0.25976065  -0.111527172   -0.2962261
Fifth   0.06863094 -0.0585034031  -0.23230572  -0.112168423   -0.2824480
       National_party_candidate Cand_corepathi Cand_graduate_above
Winner              0.466987672     0.08810825          0.08781124
Second             -0.063862706     0.14094440          0.08075499
Third               0.001702397     0.08356371         -0.05554315
Fourth             -0.395571230    -0.24214049         -0.11622587
Fifth              -0.261440055    -0.25092375         -0.06416961
       Cand_filed_ITR
Winner    -0.02594495
Second     0.07642842
Third      0.09147751
Fourth    -0.10410190
Fifth     -0.11438267

Coefficients of linear discriminants:
                                 LD1         LD2         LD3         LD4
CAND_AGE                  0.03351795 -0.62862062  0.28216437  0.67197711
Criminal.Case            -0.20375454  0.26689964  0.52310822 -0.26267423
Total_Assets             -0.09695203 -0.60342652 -0.49450542 -0.09112014
Liabilities_P            -0.11129051  0.60118068  0.49451917  0.01608864
Recontesting             -0.58739459 -0.31944979  0.28940614 -0.52844914
National_party_candidate -0.78939306  0.30775828 -0.09953747  0.59008877
Cand_corepathi           -0.40695066 -0.06446029 -0.73678882 -0.27634909
Cand_graduate_above      -0.14347573 -0.04716205  0.61010707  0.23325189
Cand_filed_ITR            0.63870790  0.11137074 -0.29075627  0.09388917
```

```
Proportion of trace:
   LD1    LD2    LD3    LD4
0.6983 0.1883 0.1005 0.0129
```

Testing the model with test data

```
predictions <- model %>% predict(test.transformed)

lda.data <- cbind(train.transformed, predict(model)$x)
ggplot(lda.data, aes(LD1, LD2)) +
  geom_point(aes(color = POSITION)) + ggtitle("Punjab's results seperation by LD1 and LD2")
```



Observation:

- There is not clear separation of the candidates belonging to different, as the result is same as Goa states results as the same feature set is used. To get more predictive features would require features which are more specific to each state and more granular data for each district and candidate

Accuracy of the model

```
mean(predictions$class==test.transformed$POSITION)
```

```
[1] 0.2333333
```

```
table(predictions$class, test.transformed$POSITION)
```

```
         Winner Second Third Fourth Fifth
  Winner      7      9     8      5     5
  Second     10      5     2      0     1
  Third       4      4     6      6     1
  Fourth      1      3     5      3     5
  Fifth       0      0     0      0     0
```

Observation:

- The accuracy is about 0.23, and cross table shows larege misclassification. It more looks like random allocation of observation, in this scenario a random guessing if adopted based on feature by painstakingly going through the results we might get the similar or if lucky a better accuracy

6.3 Running the model for the Uttar Pradesh state

```
   DIST_NAME AC_NAME AC_TYPE               CAND_NAME POSITION CAND_SEX
1 Saharanpur   Behat     GEN             NARESH SAINI        1        M
2 Saharanpur   Behat     GEN      MAHAVEER SINGH RANA        2        M
3 Saharanpur   Behat     GEN              MOHD. IQBAL        3        M
4 Saharanpur   Behat     GEN RANA ADITYA PRATAP SINGH        4        M
5 Saharanpur   Behat     GEN               KAMRAN ALI        6        M
6 Saharanpur   Behat     GEN                     ARUN        7        M
  CAND_CATEGORY CAND_AGE PARTYABBRE TOTALVALIDVOTESPOLLED
1           GEN       53        INC                 97035
2           GEN       55        BJP                 71449
3           GEN       52        BSP                 71019
4           GEN       49        IND                  4187
5           GEN       37      BhaSP                  1255
6           GEN       38        RLD                  1150
  TOTAL.VOTES.POLLED TotalElectors Criminal.Case             Education
1             252563        336576             0         Post Graduate
2             252563        336576             0 Graduate Professional
3             252563        336576             0              Literate
4             252563        336576             0             12th Pass
5             252563        336576             0             10th Pass
6             252563        336576             0              Graduate
  Total_Assets Liabilities_P Recontesting National_party_candidate
1     10780291        398000            0                        1
2     55663295       2985082            1                        1
3    140103795      10214938            0                        1
4     65920214        286731            0                        0
5       682859             0            0                        0
6      5477510             0            0                        0
  Winner_corepathi Winner_graduate_above Winner_filed_ITR Cand_corepathi
1                1                     1                1              1
2                0                     0                0              1
3                0                     0                0              1
4                0                     0                0              1
5                0                     0                0              0
6                0                     1                1              1
```

```
  Cand_graduate_above Cand_filed_ITR  Age_category
1                   1              1 Between_46_55
2                   1              1 Between_46_55
3                   0              1 Between_46_55
4                   0              0 Between_46_55
5                   0              0 Between_36_45
6                   1              1 Between_36_45
```

```
 [1] "DIST_NAME"                "AC_NAME"
 [3] "AC_TYPE"                  "CAND_NAME"
 [5] "POSITION"                 "CAND_SEX"
 [7] "CAND_CATEGORY"            "CAND_AGE"
 [9] "PARTYABBRE"               "TOTALVALIDVOTESPOLLED"
[11] "TOTAL.VOTES.POLLED"       "TotalElectors"
[13] "Criminal.Case"            "Education"
[15] "Total_Assets"             "Liabilities_P"
[17] "Recontesting"             "National_party_candidate"
[19] "Winner_corepathi"         "Winner_graduate_above"
[21] "Winner_filed_ITR"         "Cand_corepathi"
[23] "Cand_graduate_above"      "Cand_filed_ITR"
[25] "Age_category"
```

Training the UP state model with few important features

```
data_sub <- plt.copy[,c('POSITION','CAND_AGE','Criminal.Case','Total_Assets','Liabilities_P','Recontesti
head(data_sub)
```

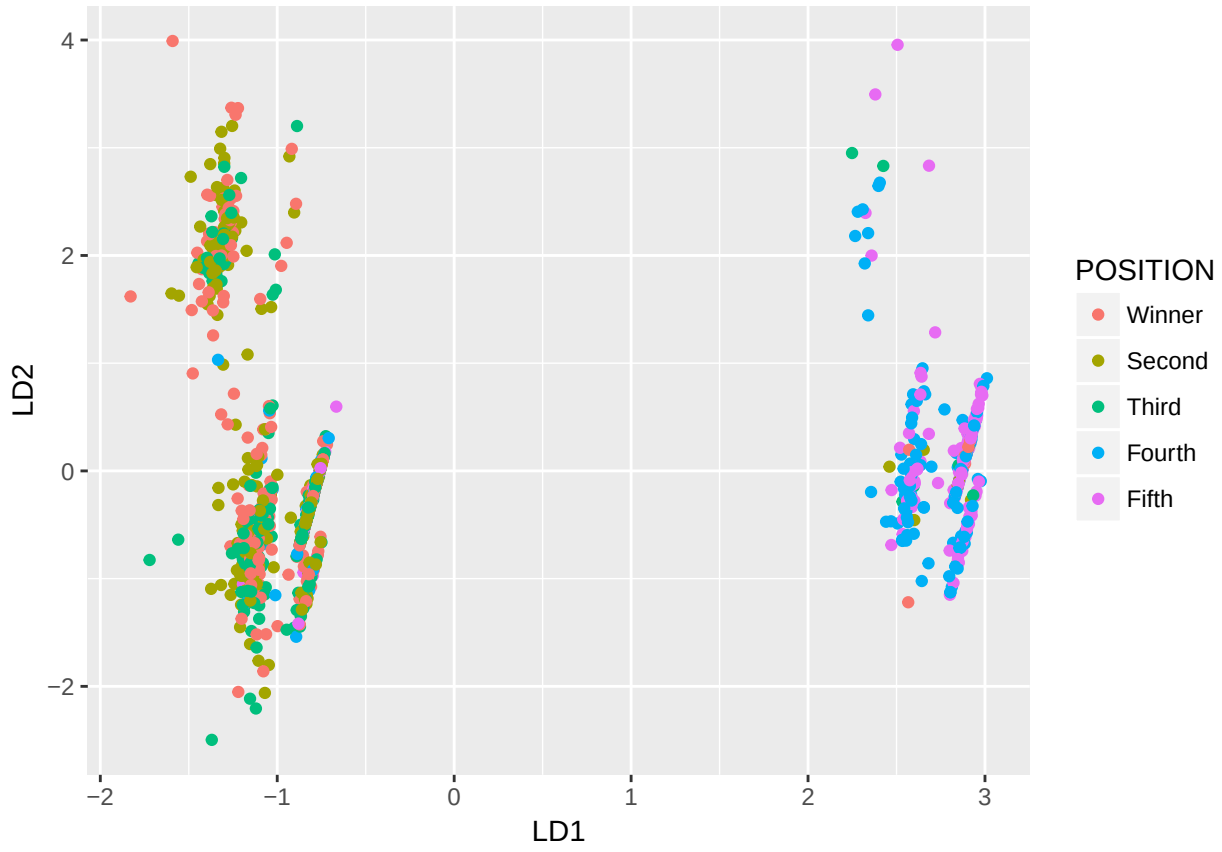
```
   POSITION CAND_AGE Criminal.Case Total_Assets Liabilities_P Recontesting
1    Winner       53             0     10780291        398000            0
2    Second       55             0     55663295       2985082            1
3     Third       52             0    140103795      10214938            0
4    Fourth       49             0     65920214        286731            0
14   Winner       60             0     65599100      13797000            1
15   Second       46             6     45048336             0            0
   National_party_candidate Cand_corepathi Cand_graduate_above
1                         1              1                   1
2                         1              1                   1
3                         1              1                   0
4                         0              1                   0
14                        1              1                   1
15                        1              1                   0
   Cand_filed_ITR
1               1
2               1
3               1
4               0
14              1
15              1
```

```
model <- lda(POSITION~ ., data = train.transformed)

model
```

```
Call:
```

```
lda(POSITION ~ ., data = train.transformed)

Prior probabilities of groups:
   Winner    Second     Third    Fourth     Fifth
0.2364650 0.2332803 0.2308917 0.1536624 0.1457006

Group means:
          CAND_AGE Criminal.Case Total_Assets Liabilities_P Recontesting
Winner  0.24571638 -0.0047184232   0.13652531    0.13440235   0.12394828
Second  0.05393184  0.0879956256   0.03536288    0.01332483   0.34448097
Third  -0.06050693  0.0210809629   0.05808414   -0.01428079  -0.08834821
Fourth -0.15906285  0.0009558626  -0.07961465   -0.04011842  -0.26435436
Fifth  -0.22149538 -0.1676464881  -0.28627385   -0.17452126  -0.33390267
       National_party_candidate Cand_corepathi Cand_graduate_above
Winner                0.5141664     0.17953089         0.110935142
Second                0.5437209     0.21556174         0.035637710
Third                 0.5119661     0.09661808         0.003784338
Fourth               -1.1130009    -0.25405223        -0.133462525
Fifth                -1.3425063    -0.52167992        -0.102343043
       Cand_filed_ITR
Winner      0.1383421
Second      0.1205508
Third       0.1120650
Fourth     -0.1903163
Fifth      -0.3944088

Coefficients of linear discriminants:
                                    LD1          LD2         LD3         LD4
CAND_AGE                     0.037223586  0.225091274 -0.82380420 -0.1113789
Criminal.Case               -0.035302762 -0.004621701  0.24895880 -0.3732480
Total_Assets                 0.011063792 -0.304958557 -0.12911096 -0.2906243
Liabilities_P               -0.031202323  0.238600354 -0.38363927 -0.0857605
Recontesting                -0.071434419  0.965118386  0.28784707  0.1722600
National_party_candidate    -1.644624714 -0.219751942  0.05845383  0.2596369
Cand_corepathi              -0.153855527  0.183422062  0.29716821 -0.6960336
Cand_graduate_above          0.019253740  0.145963364 -0.42278083  0.6075881
Cand_filed_ITR              -0.009159933 -0.425225538 -0.08687692 -0.1712943

Proportion of trace:
   LD1    LD2    LD3    LD4
0.9739 0.0138 0.0085 0.0038
```

```
predictions <- model %>% predict(test.transformed)

lda.data <- cbind(train.transformed, predict(model)$x)
ggplot(lda.data, aes(LD1, LD2)) +
  geom_point(aes(color = POSITION))
```




```
[1] 0.3910256
```

	Winner	Second	Third	Fourth	Fifth
Winner	23	20	19	4	1
Second	11	20	7	1	0
Third	36	30	37	3	6
Fourth	2	2	7	11	7
Fifth	2	1	2	29	31

Interpretation of UP states model results:

- The model accuracy is way better than what was achieved for other states, and as only few important features were used and UP state more observation the model was able to find some combination of LDA to attain the accuracy of 0.39
- The LDA1 was able to discriminate among the position largely
- And it is still not a good model which discriminates well the positions (results), of the candidates and the features which have been used to train the model are not the high influencing factors for a candidate to be able to win the election

7. Project conclusion

- The data which are basic ingredients for building predictive models and it is very important to have data which represents the underlying relations with the dependent variable

- Given politics is a very complex subject and predicting election results would require a complex model with even more high quality granular data
- It might require to build multiple models and complex intrepretation to improve the accuracy than what has been achieved using the above model and data