data_explore

Bingfeng Xia

3/29/2017

1 upload data

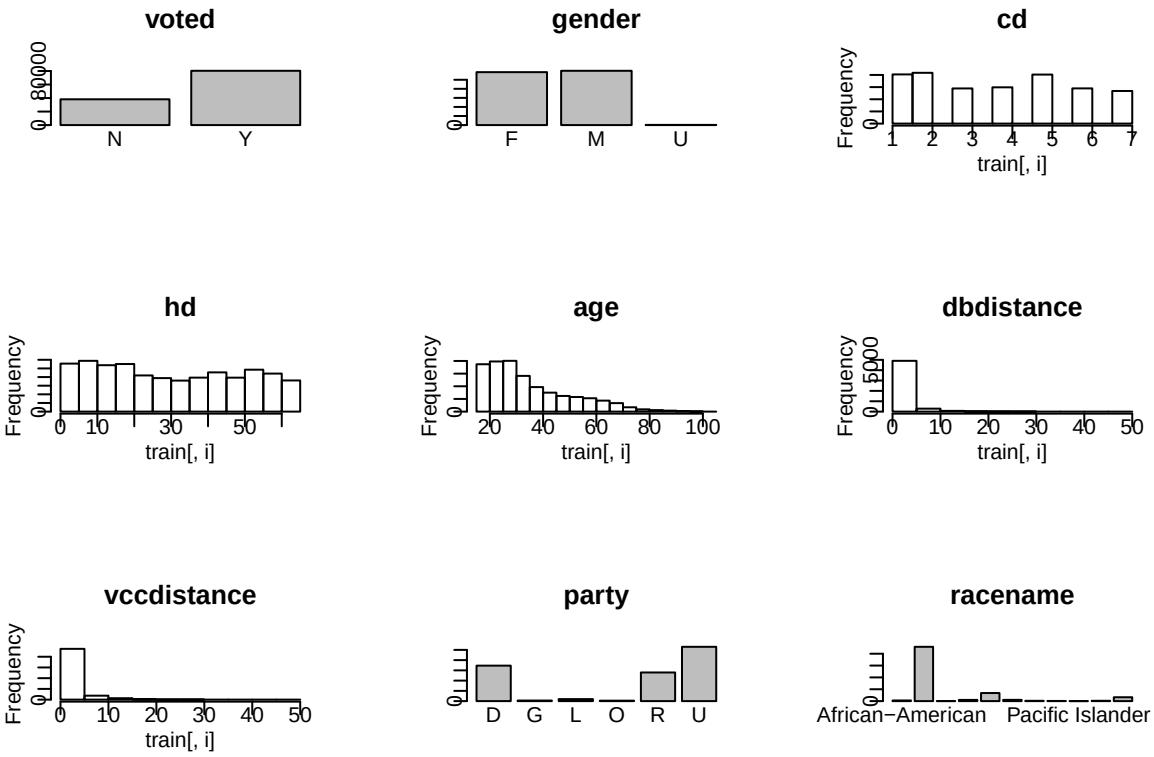
```
train = read.csv(file = "train.csv", header = T)
test = read.csv(file = "test.csv", header = T)
test.new = train
test.new = test
```

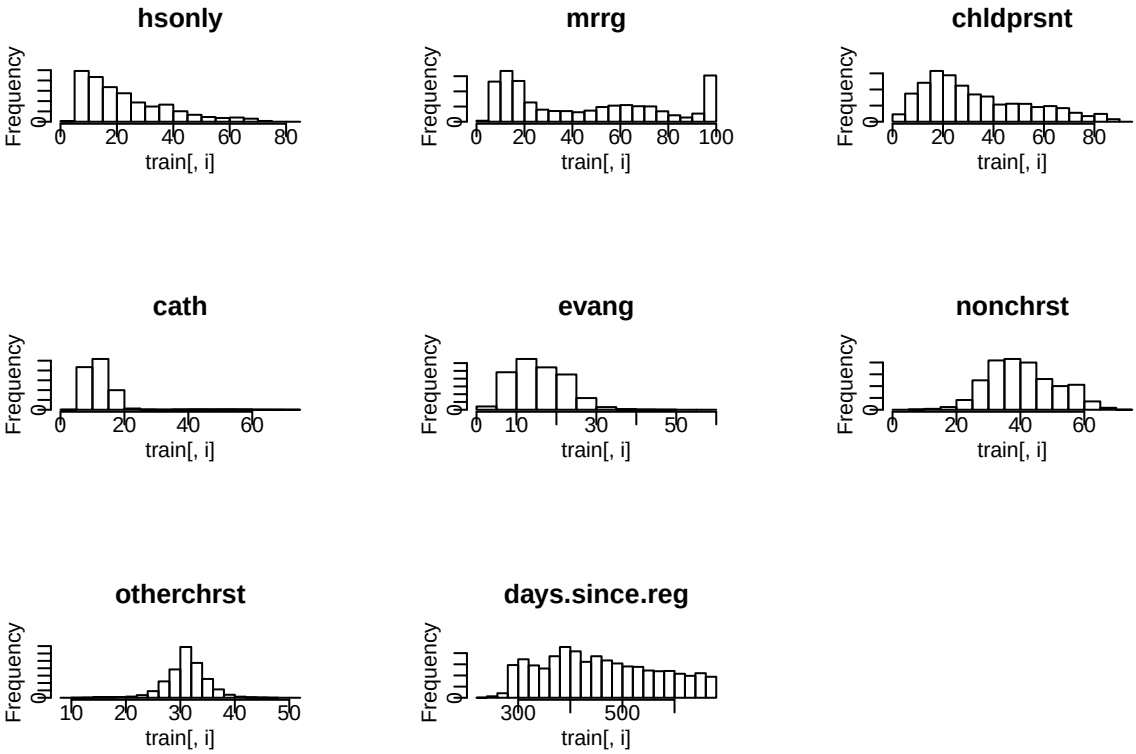
2 data exploration

```
summary(train)
```

```
##  voted     gender          cd              hd             age
##  N:38172   F:58414   Min.   :1.000   Min.   : 1.00   Min.   : 18.00
##  Y:80357   M:59916   1st Qu.:2.000   1st Qu.:14.00   1st Qu.: 23.00
##            U:  199   Median :4.000   Median :29.00   Median : 31.00
##                      Mean   :3.767   Mean   :30.93   Mean   : 35.67
##                      3rd Qu.:5.000   3rd Qu.:48.00   3rd Qu.: 45.00
##                      Max.   :7.000   Max.   :65.00   Max.   :101.00
##                      NA's   :2       NA's   :2
##    dbdistance      vccdistance     party                racename
##  Min.   : 1.77    Min.   : 1.77    D:34641   Caucasian       :91681
##  1st Qu.: 1.96    1st Qu.: 2.03    G:  535   Hispanic        :13656
##  Median : 2.28    Median : 2.44    L: 1940   Uncoded         : 6525
##  Mean   : 2.91    Mean   : 3.24    O:  443   Jewish          : 2107
##  3rd Qu.: 3.06    3rd Qu.: 3.35    R:27958   East Asian      : 2019
##  Max.   :45.99    Max.   :45.99    U:53012   African-American: 1084
##  NA's   :113247   NA's   :113247             (Other)         : 1457
##      hsonly           mrrg         chldprsnt          cath
##  Min.   : 4.00   Min.   : 2.70   Min.   : 0.80   Min.   : 4.60
##  1st Qu.:11.10   1st Qu.:14.90   1st Qu.:17.20   1st Qu.: 8.80
##  Median :18.60   Median :37.40   Median :27.40   Median :11.70
##  Mean   :22.96   Mean   :44.45   Mean   :33.25   Mean   :12.32
##  3rd Qu.:31.80   3rd Qu.:70.40   3rd Qu.:47.70   3rd Qu.:14.40
##  Max.   :84.30   Max.   :99.70   Max.   :90.40   Max.   :74.10
##
##      evang         nonchrst       otherchrst    days.since.reg
##  Min.   : 1.6   Min.   : 4.70   Min.   : 8.70   Min.   :223.0
##  1st Qu.:10.5   1st Qu.:32.90   1st Qu.:29.20   1st Qu.:375.0
##  Median :15.1   Median :39.70   Median :31.30   Median :449.0
##  Mean   :15.8   Mean   :40.75   Mean   :31.13   Mean   :459.4
##  3rd Qu.:20.7   3rd Qu.:47.90   3rd Qu.:33.30   3rd Qu.:546.0
##  Max.   :56.5   Max.   :74.40   Max.   :50.80   Max.   :677.0
##
```

```
par(mfrow = c(3,3))
for (i in 1:17) {
  if(i == 1 || i == 2 || i == 8 || i == 9) {
    plot(train[,i], main= colnames(train)[i], mgp = c(1,0,0))
  }
  else {
    hist(train[,i], main= colnames(train)[i], mgp = c(1,0,0))
  }
}
```






```
#convert dependent variable
train$voted = as.numeric(train$voted)
train$voted[train$voted == 1] = 0
train$voted[train$voted == 2] = 1
#basic logit model
voted.logit1 = glm(voted ~ ., data = train, family = binomial, maxit = 100000)
summary(voted.logit1)
```

```
##
## Call:
## glm(formula = voted ~ ., family = binomial, data = train, maxit = 1e+05)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2632  -1.2304   0.6590   0.8049   1.5564
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.580e+01  5.693e+01  -0.277  0.78141
## genderM                  -1.354e-01  6.705e-02  -2.019  0.04345 *
## genderU                   1.135e+01  1.833e+02   0.062  0.95065
## cd                        2.610e-02  1.762e-02   1.481  0.13855
## hd                       -5.427e-03  1.965e-03  -2.762  0.00575 **
## age                      -2.628e-03  3.158e-03  -0.832  0.40523
## dbdistance                7.525e-02  2.985e-02   2.521  0.01169 *
## vccdistance              -4.247e-02  2.384e-02  -1.781  0.07487 .
## partyG                   -4.345e-01  5.982e-01  -0.726  0.46764
## partyL                   -4.321e-01  3.394e-01  -1.273  0.20306
## partyO                   -1.374e+00  5.374e-01  -2.557  0.01056 *
## partyR                   -1.459e-01  1.001e-01  -1.457  0.14514
```

```
## partyU                   -6.765e-01  8.382e-02  -8.072 6.94e-16 ***
## racenameCaucasian         3.241e-01  3.243e-01   0.999  0.31770
## racenameCentral Asian     3.869e-01  1.212e+00   0.319  0.74955
## racenameEast Asian        1.830e-01  3.740e-01   0.489  0.62458
## racenameHispanic          2.438e-01  3.370e-01   0.724  0.46936
## racenameJewish            7.012e-01  4.082e-01   1.718  0.08585 .
## racenameMiddle Eastern    5.730e-01  7.233e-01   0.792  0.42830
## racenameNative American  -2.064e-01  8.119e-01  -0.254  0.79930
## racenamePacific Islander -7.531e-01  1.476e+00  -0.510  0.60989
## racenameSouth Asian       8.125e-01  5.543e-01   1.466  0.14270
## racenameUncoded           2.085e-01  3.477e-01   0.600  0.54876
## hsonly                    7.542e-03  3.944e-03   1.912  0.05587 .
## mrrg                      1.261e-02  1.931e-03   6.532 6.48e-11 ***
## chldprsnt                -3.691e-03  1.886e-03  -1.957  0.05032 .
## cath                      1.700e-01  5.693e-01   0.299  0.76517
## evang                     1.399e-01  5.692e-01   0.246  0.80582
## nonchrst                  1.725e-01  5.693e-01   0.303  0.76184
## otherchrst                1.912e-01  5.692e-01   0.336  0.73694
## days.since.reg           -2.011e-03  2.941e-04  -6.839 7.98e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6078.1  on 5281  degrees of freedom
## Residual deviance: 5823.3  on 5251  degrees of freedom
##   (113247 observations deleted due to missingness)
## AIC: 5885.3
##
## Number of Fisher Scoring iterations: 11
```

```
#check collinearity
library(car)
car::vif(voted.logit1)  #imputing
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## gender            1.080112  2        1.019453
## cd                1.101638  1        1.049589
## hd                1.242653  1        1.114743
## age               2.278610  1        1.509507
## dbdistance        3.287136  1        1.813046
## vccdistance       3.402239  1        1.844516
## party             1.129203  5        1.012225
## racename          1.488534 10        1.020089
## hsonly            2.052621  1        1.432697
## mrrg              2.833659  1        1.683347
## chldprsnt         2.106680  1        1.451441
## cath          28484.156511  1      168.772499
## evang         13611.544818  1      116.668525
## nonchrst      32581.357222  1      180.503067
## otherchrst     7210.121691  1       84.912435
## days.since.reg    1.169132  1        1.081264
```