Voter Prediction Model - Data Cleaning

Keyan Halperin

```r
#package to open .dta files
#install.packages('readstata13')
require(readstata13)
```

```
## Loading required package: readstata13
```

```
## Warning: package 'readstata13' was built under R version 3.3.2
```

```r
setwd('C:/Users/Keyan/Google Drive/Projects')
anes.data = read.dta13('anes_timeseries_cdf.dta', generate.factors = T, nonint.factors = T)
```

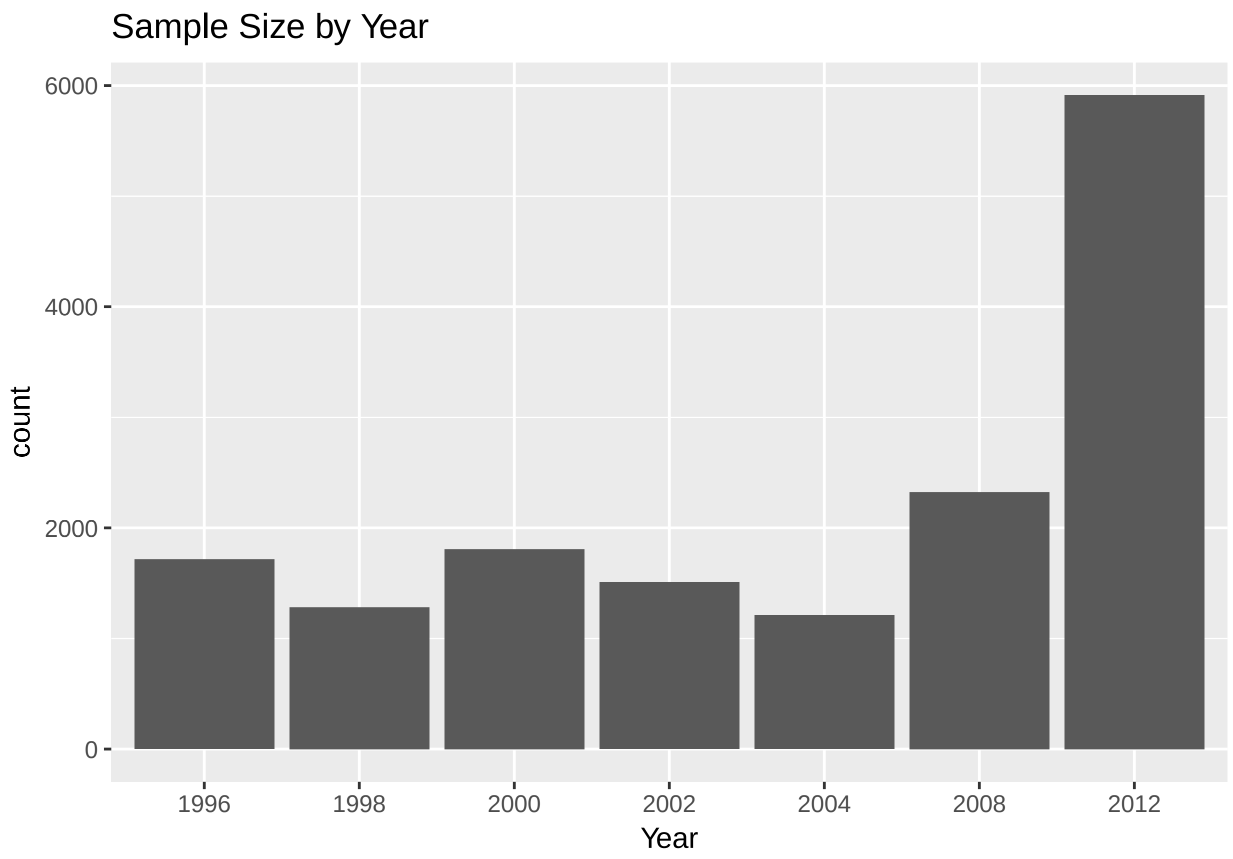
```r
names(anes.data)[2] = 'year'
summary(anes.data$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1948    1970    1984    1983    1998    2012
```

For the sake of simplicity, we will only look at presidential elections since 1996.

As we can see, the sample size in 2012 is much bigger than it was in previous years. This is because in 2012, the ANES started doing some of their polling online, which allowed them to survey significantly more people. We will later explore the data in order to determine whether the responses of individuals who were surveyed online significantly differs from the responses of those who were surveyed over the phone or in person.

```r
ggplot2::qplot(as.factor(anes.data$year[anes.data$year >= 1996]), main = 'Sample Size by Year', xlab = 'Year')
```



```
#Only include years of interest
anes.data = anes.data[anes.data$year == 1996 | anes.data$year == 2000 | anes.data$year == 2004 |
                      anes.data$year == 2008 | anes.data$year == 2012, ]

#A much easier way to write that
anes.data = anes.data[anes.data$year %in% seq(1996, 2012, 4),]

summary(anes.data$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1996    2000    2008    2007    2012    2012
```

```r
#Number of NAs
sum(is.na(anes.data))
```

```
## [1] 7729302
```

There are over 7 million NAs, but there are even more missing values since not all missing values are coded as NA, so we'll definitely have to do some data cleaning.

```r
#Data Dimensions
dim(anes.data) #Approximately 13,000 observations and 1,000 variables
```

```
## [1] 12969   952
```

```r
#Summary of the first 20 variables
str(anes.data[,1:20])
```

```
## 'data.frame':    12969 obs. of  20 variables:
##  $ Version : chr  "ANES_cdf-VERSION:2015-May-14" "ANES_cdf-VERSION:2015-May-14" "ANES_cdf-VERSION:2015-May-14" "ANES_cdf-VERSION:2015-M
##  $ year    : num  1996 1996 1996 1996 1996 ...
##  $ VCF0006 : num  1001 1002 1003 1004 1005 ...
##  $ VCF0006a: num  19942539 19920511 19921089 19942448 19920979 ...
##  $ VCF0009x: num  0.83 0.504 0.557 1.681 0.567 ...
##  $ VCF0010x: num  0.83 0.504 0.557 1.681 0.567 ...
##  $ VCF0011x: num  0.83 0.504 0.557 1.681 0.567 ...
##  $ VCF0009y: num  0.83 0.504 0.557 1.681 0.567 ...
##  $ VCF0010y: num  0.83 0.504 0.557 1.681 0.567 ...
##  $ VCF0011y: num  0.83 0.504 0.557 1.681 0.567 ...
##  $ VCF0009z: num  0.83 0.504 0.557 1.681 0.567 ...
##  $ VCF0010z: num  0.83 0.504 0.557 1.681 0.567 ...
##  $ VCF0011z: num  0.83 0.504 0.557 1.681 0.567 ...
##  $ VCF0012 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ VCF0012a: int  0 1 1 0 0 2 1 3 3 3 ...
##  $ VCF0012b: int  4 2 2 2 1 3 4 4 3 4 ...
##  $ VCF0013 : Factor w/ 2 levels "0. No Post-election interview data",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ VCF0014 : Factor w/ 2 levels "0. No Pre-election interview data present",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ VCF0015a: Factor w/ 3 levels "0. Pre IW not abbreviated [1992:'Long' form Pre]",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ VCF0015b: Factor w/ 3 levels "0. Post IW is not abbreviated",..: 1 1 1 1 1 1 1 1 1 1 ...
```

There are almost 1000 variables in the data, but we will only consider variables which are potentially related to how someone will vote and are feasible to be known or determined.

```
variables = c('VCF0006a', 'year',   'VCF0013', 'VCF0017', 'VCF0101', 'VCF0102', 'VCF0104',
              'VCF0106', 'VCF0110', 'VCF0112', 'VCF0114', 'VCF0118', 'VCF0127', 'VCF0128',
              'VCF0138', 'VCF0143', 'VCF0146', 'VCF0147', 'VCF0721', 'VCF0901b', 'VCF0900c',
              'VCF0302', 'VCF0303', 'VCF0702', 'VCF0703', 'VCF0704a', 'VCF9027')

data = anes.data[variables]

#Change variable names
names(data) = c( 'id', 'year', 'post', 'method', 'age', 'age.group', 'gender', 'race', 'educ',
                 'region', 'income', 'work', 'union', 'religion', 'num.children', 'parents.native',
                 'home.own','marital.stat', 'donate', 'state', 'district', 'party.1', 'party.2',
                 'did.vote', 'reg.vote', 'pres.vote', 'previous.vote' )
```

The primary goal of this data cleaning is to recode missing values as NA since that is how missing values should be represented in R. **Although I will go through each variable manually in order to ensure that I do not miss any problems with the data, I will initially demonstrate how I would do it systematically.** Note that DK stands for 'Don't Know', RF means that they refused to respond.

```
#Native Parents
summary(data$parents.native)
```

```
##                                               1. Yes
##                                                10673
##                                                5. No
##                                                 2249
##                                                8. DK
##                                                   21
## 9. NA; RF; no Pre IW; short-form 'new' Cross Section
##                                                   26
```

If we look at a summary of `parents.native` for example, we can see that two levels of the variable should be recoded as `NA`. Having looked at a few of these, I have noticed some patterns and will try to systematically loop through all of the variables to recode these missing values.

```
library(stringr)
data2 = data

#Vector of categorical variables
factor.variables = names(data)[sapply(data, is.factor)]

for (i in 1:length(factor.variables)){  #For each variable
```

```r
  var = factor.variables[i]
  lev = levels(data[[var]])
  sel = str_detect(lev, '(NA|DK|RF)') #Detect which levels contain the term NA, DK, or RF

  for (l in lev[sel]){  #For each level that contains NA, DK, or RF

    data2[[var]][(data2[[var]] == l)] <- NA #Recode the observations with that level as NA

  }

}

summary(data2[ ,1:10])
```

```
##        id                 year     
##  Min.   :19920002   Min.   :1996  
##  1st Qu.:20001532   1st Qu.:2000  
##  Median :20081753   Median :2008  
##  Mean   :20066830   Mean   :2007  
##  3rd Qu.:20123618   3rd Qu.:2012  
##  Max.   :20126864   Max.   :2012  
##                                                        
##                                       post      
##  0. No Post-election interview data     : 1202  
##  1. Post-election interview data present:11767  
##                                                        
##                                                        
##                                                        
##                                                        
##                                                        
##                                          method          age       
##  0. All personal                            :7378   56      :  282  
##  1. Telephone pre (personal post or no post): 117   58      :  276  
##  2. Telephone post (personal pre)           : 865   52      :  261  
##  3. All telephone                           : 749   53      :  261  
##  4. All internet (2012: pre and post)       :3860   50      :  260  
##                                                     (Other):11513  
##                                                     NA's    :  116  
##       age.group                 gender     
##  4. 45 - 54:2424   0. NA; no Pre IW:   0
##  3. 35 - 44:2366   1. Male         :5968
##  5. 55 - 64:2320   2. Female       :7001
##  2. 25 - 34:2128
##  6. 65 - 74:1575
##  (Other)   :2040
##  NA's      : 116
##                         race
##  0. Missing, pre-1966 data:   0
##  1. White non-Hispanic    :8177
##  2. Black non-Hispanic    :2152
##  3. Other                 :2542
##  9. Missing, DK/REF/NA    :   0
##  NA's                     :  98
##
##                                                      educ
##  0. DK; NA; no Pre IW; short-form 'new' Cross Section :   0
##  1. Grade school or less (0-8 grades)                 : 401
##  2. High school (12 grades or fewer, incl. non-college:4709
##  3. Some college (13 grades or more but no degree;    :4062
##  4. College or advanced degree (no cases 1948)        :3705
##  NA's                                                 :  92
##
##                                                        region
##  0. NA (1948)                                            :   0
##  1. Northeast (CT, ME, MA, NH, NJ, NY, PA, RI, VT)       :2012
##  2. North Central (IL, IN, IA, KS, MI, MN, MO, NE, ND,   :2875
##  3. South (AL, AR, DE, D.C., FL, GA, KY, LA, MD, MS, NC  :5098
##  4. West (AK, AZ, CA, CO, HI, ID, MT, NV, NM, OR, UT, WA,:2984
##
##
```

It looks like we did a pretty good job! There are a few issues, but they could be easily fixed.

I will now go through each variable manually instead in order to ensure that I do not miss any problems with the data. **Feel free to skip the rest of this file and look at the exploration and modeling phases instead!**

```
#ID
summary(data$id)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 19920000 20000000 20080000 20070000 20120000 20130000
```

```
#Check to make sure there are no duplicates
length(data$id) == length(unique(data$id)) #Looks good
```

```
## [1] TRUE
```

```
#Post-election Interview Data
summary(data$post)
```

```
##      0. No Post-election interview data
##                                    1202
## 1. Post-election interview data present
##                                   11767
```

```
#Drop individuals for whom we do not know how they voted
data = data[data$post != '0. No Post-election interview data', ]
```

```
#Survey Method
summary(data$method)
```

```
##                             0. All personal
##                                        6572
## 1. Telephone pre (personal post or no post)
##                                           0
##             2. Telephone post (personal pre)
##                                         865
##                            3. All telephone
##                                         749
##         4. All internet (2012: pre and post)
##                                        3581
```

Age is coded as a factor variable, but in order to convert it to numeric, we need to first convert it to a character. This is because factor variables have a built-in numeric value based on what order the levels are in. As a result, 17 would be converted to a 1, 18 to a 2, 19 to a 3, etc.

```
#Age
head(summary(data$age), 10)
```

```
## 00. NA; DK; RF; no Pre IW                        17
##                       104                         2
##                        18                        19
##                        84                       118
##                        20                        21
##                        155                       130
##                         22                        23
##                        150                       164
##                         24                        25
##                        171                       185
```

```
tail(summary(data$age), 10)
```

```
##                                                       90
##                                                       20
##                                                       91
##                                                        8
##                                                       92
##                                                        2
##                                                       93
##                                                        4
##                                                       94
##                                                        0
##                                                       95
##                                                        0
##                                                       96
##                                                        1
## 97. 97 years old (1952, 1974, 1996 and later: or older)
##                                                        0
##      98. 98 years old (1958-1962, 1966, 1968: or older)
##                                                        0
##        99. 99 years old (1976-1990,1994,2002: or older)
##                                                        0
```

```
#Recode missing values as NA
data$age[data$age == '00. NA; DK; RF; no Pre IW'] <- NA
data$age = as.numeric(as.character(data$age))
summary(data$age) #Looks good
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   17.00   35.00   48.00   48.56   61.00   96.00     104
```

```
#Age Group
summary(data$age.group)
```

```
##          0. NA; DK; RF; no Pre IW                        1. 17 - 24
##                               104                               974
##                        2. 25 - 34                        3. 35 - 44
##                              1909                              2119
##                        4. 45 - 54                        5. 55 - 64
##                              2207                              2152
##                        6. 65 - 74 7. 75 - 99 and over (except 1954)
##                              1472                               830
```

```
data$age.group[data$age.group == '0. NA; DK; RF; no Pre IW'] <- NA
```

```
#Gender
summary(data$gender) #Looks good
```

```
## 0. NA; no Pre IW          1. Male        2. Female
##                0             5447             6320
```

```
#Race
summary(data$race)
```

```
## 0. Missing, pre-1966 data     1. White non-Hispanic
##                         0                      7492
##     2. Black non-Hispanic                  3. Other
##                      1932                      2266
##     9. Missing, DK/REF/NA
##                        77
```

```
data$race[data$race == '9. Missing, DK/REF/NA'] <- NA
```

```
#Education
summary(data$educ)
```

```
##  0. DK; NA; no Pre IW; short-form 'new' Cross Section
##                                                    79
##                  1. Grade school or less (0-8 grades)
##                                                   351
## 2. High school (12 grades or fewer, incl. non-college
##                                                  4215
##     3. Some college (13 grades or more but no degree;
##                                                  3708
##         4. College or advanced degree (no cases 1948)
##                                                  3414
```

Remember that each level of a factor variable is also stored as an integer. So instead of typing out the name of the level e.g. '0. DK; NA; no Pre IW; short-form 'new' Cross Section', you can simply refer to the integer that the level corresponds to. For example:

```
data$educ[as.numeric(data$educ) == 1] <- NA
```

Although it is easier to type, it is not as clear to the reader what exactly is being done.

```
#Region of Residence
summary(data$region) #Looks good
```

```
##                                              0. NA (1948)
##                                                         0
##          1. Northeast (CT, ME, MA, NH, NJ, NY, PA, RI, VT)
##                                                      1819
##     2. North Central (IL, IN, IA, KS, MI, MN, MO, NE, ND,
##                                                      2637
##    3. South (AL, AR, DE, D.C., FL, GA, KY, LA, MD, MS, NC
##                                                      4610
## 4. West (AK, AZ, CA, CO, HI, ID, MT, NV, NM, OR, UT, WA,
##                                                      2701
```

```
#Household Income
summary(data$income)
```

```
## 0. DK; NA; refused to answer; no Pre IW
##                                     814
##                   1. 0 to 16 percentile
##                                    2017
##                  2. 17 to 33 percentile
##                                    1959
##                  3. 34 to 67 percentile
##                                    3871
##                  4. 68 to 95 percentile
##                                    2547
##                 5. 96 to 100 percentile
##                                     559
```

```
data$income[data$income == '0. DK; NA; refused to answer; no Pre IW'] <- NA
```

```
#Employmeny Status
summary(data$work)
```

```
##                                              1. Employed
##                                                     6751
##        2. Not employed: laid off, unemployed, on strike,
##                                                     1396
##                                               3. Retired
##                                                     2374
## 4. Homemaker (since 1972: not working 20 or more hrs/wk;
##                                                      815
##   5. Student (since 1972: not working 20 or more hrs/wk;
##                                                      412
##                                     9. DK; NA; no Pre IW
##                                                       19
```

```
data$work[data$work == '9. DK; NA; no Pre IW'] <- NA
```

```
#Union Membership
summary(data$union)
```

```
##           0. DK; NA; no Pre IW; short-form 'new' Cross
##                                                     52
## 1. Yes, someone (1948: head) in household belongs to a
##                                                   1788
##    2. No, no one in household belongs to a labor union
##                                                   9927
```

```
data$union[as.numeric(data$union) == 1] <- NA
```

```
#Religion
summary(data$religion)
```

```
## 0. DK; NA; refused to answer; no Pre IW; no Post IW;
##                                                  141
##                                        1. Protestant
##                                                 5457
##                         2. Catholic [Roman Catholic]
##                                                 2791
##                                            3. Jewish
##                                                  222
##      4. Other and none (also includes DK preference)
##                                                 3156
```

```
data$religion[data$religion == '0. DK; NA; refused to answer; no Pre IW; no Post IW;'] <- NA
```

```
#Number of Children
summary(data$num.children)
```

```
##                                  0. None
##                                     6036
##                                   1. One
##                                     1216
##                                   2. Two
##                                     1055
##                                 3. Three
##                                      698
##                                  4. Four
##                                        0
##                                  5. Five
##                                        0
##                                   6. Six
##                                        0
##                                 7. Seven
##                                        0
##                         8. Eight or more
##                                        0
## 9. NA; no Pre IW; Panel (1992,1996,2002)
##                                     1207
##                                     NA's
##                                     1555
```

Number of children is coded as a categorical variable instead of a numeric one. One way to convert this into a numeric variable would be:

```
#data$num.children2[data$num.children == '0. None'] <- 0
#data$num.children2[data$num.children == '1. One'] <- 1
#data$num.children2[data$num.children == '2. Two'] <- 2
#data$num.children2[data$num.children == '3. Three'] <- 3
```

However, since each level of a factor variable is stored as an integer, a much easier way to do this would be:

```
data$num.children = as.numeric(data$num.children) - 1
data$num.children[data$num.children == 9] <- NA
summary(data$num.children) #Looks good!
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  0.0000  0.0000  0.0000  0.6019  1.0000  3.0000    2762
```

Create a new indicator variable of whether or not the individual has children where ‘Yes’ = Has at least one child, ‘No’ = Does not have any children.

```
#Children Indicator
data$children.ind = ifelse(data$num.children >= 1, 'Yes', 'No')
data$children.ind = as.factor(data$children.ind)
summary(data$children.ind)
```

```
##   No  Yes NA's 
## 6036 2969 2762
```

```
#Native Parents
summary(data$parents.native)
```

```
##                                          1. Yes 
##                                            9722 
##                                           5. No 
##                                            2009 
##                                           8. DK 
##                                              16 
## 9. NA; RF; no Pre IW; short-form 'new' Cross Section 
##                                              20
```

```
#Recode levels 3 and 4 as NA
data$parents.native[as.numeric(data$parents.native) >= 3] <- NA
```

```
#Home Ownership
summary(data$home.own)
```

```
##                     1. Yes, own 
##                            7803 
##                 2. No, not owned 
##                            3921 
##                           8. DK 
##                               8 
## 9. NA; RF; no Pre IW; short form (1992) 
##                              35
```

```
data$home.own[as.numeric(data$home.own) >= 3] <- NA
```

```r
#Marital Status
summary(data$marital.stat)
```

```
##                                                                                                        1. Married
##                                                                                                              5826
##                                                                                                  2. Never married
##                                                                                                              2276
##                                                                                                       3. Divorced
##                                                                                                              1564
##                                                                                                      4. Separated
##                                                                                                               360
##                                                                                                        5. Widowed
##                                                                                                              1131
##                                                           7. Partners; not married (VOLUNTEERED [exc.1986,2012])
##                                                                                                               577
## 8. R not married/partnered, refused to say whether never married, divorced, separated or widowed (1992 only); DK
##                                                                                                                 2
##                    9.  NA; no Pre IW; unmarried at time of IW (1952 only); short-form 'new' Cross-Section (1992)
##                                                                                                                31
```

```r
data$marital.stat[as.numeric(data$marital.stat) >= 7] <- NA
```

```r
#Campaign Donations
summary(data$donate)
```

```
##          0. DK; NA; no Post IW; form III,IV (1972);
##                                                   5
## 1. No (includes 'not asked for money' in 1966,1968)
##                                               10355
##           2. Yes (includes 'tax check-off' in 1976)
##                                                1407
```

```r
data$donate[data$donate == '0. DK; NA; no Post IW; form III,IV (1972);'] <- NA
```

```r
#State of Residence
data$state = as.factor(data$state)
summary(data$state)
```

```
##    99    AK    AL    AR    AZ    CA    CO    CT    DC    DE    FL    GA    HI    IA    ID
##     2     3   243   110   224  1344   262   118    28    47   746   337     7   129    18
##    IL    IN    KS    KY    LA    MA    MD    ME    MI    MN    MO    MS    MT    NC    ND
##  372  307   81   87  280  278  177   15  467  290  159   98   23  310   52
##   NE   NH   NJ   NM   NV   NY   OH   OK   OR   PA   RI   SC   SD   TN   TX
##   56   60  267  130   90  639  419  103  198  389   43  181   16  273 1133
##   UT   VA   VT   WA   WI   WV   WY
##  116  407    8  266  289   50   20
```

```
data$state[data$state == '99'] <- NA
```

```
#District of Residence
data$district = as.factor(data$district)
summary(data$district)
```

```
##    9999    VA09    MN01    CA04    WI04    LA04    MI05    OR04    TX16
##     186     125     124     117     105     103     100      96      96
##    CA19    IN02    IN06    NJ02    AR04    MI04    AL07    TX27    TN02
##      91      84      82      79      75      74      73      73      72
##    FL04    CO02    TX29    OH04    MA03    PA01    FL05    NM02    FL12
##      71      69      69      68      67      67      66      66      64
##    GA01    TX28    TX15    TX21    CA40    IL07    LA02    SC01    WA07
##      64      63      61      61      60      60      60      59      59
##    TX11    AL06    CO01    FL02    TN07    CA11    FL06    GA02    TX20
##      58      56      55      54      53      52      52      52      52
##    UT02    CA08    GA06    LA05    ND01    CA20    TX08    VA03    CO04
##      52      51      51      51      51      50      48      47      46
##    DE01    MS03    NY27    WA09    AZ03    CT03    OK05    TX30    VA07
##      46      46      46      46      45      45      45      45      45
##    AZ05    CA42    CO07    IA04    NM01    WI05    FL27    KS04    OH08
##      44      44      44      44      44      44      43      43      43
##    AL03    MN05    NC12    TX17    FL09    NC01    NY19    SC06    MA01
##      42      42      42      42      41      41      41      41      40
##    NV02    VA05    AZ06    MA08    MD08    MN04    NC04    OR05    PA09
##      40      40      39      39      39      39      39      39      39
##    IA03    NJ11    NY21    OH18    TN09    VA08    IL01    IL02    NJ10
##      38      38      38      38      38      37      36      36      36
## (Other)
##    6065
```

```
data$district[data$district == '9999'] <- NA
```

```
#Political Party 1
summary(data$party.1)
```

```
##                   1. Republican                  2. Independent
##                            2847                            3521
## 3. No preference; none; neither                        4. Other
##                             506                             215
##                     5. Democrat                           8. DK
##                            4545                              73
##                  9. NA; refused
##                              60
```

```
data$party.1[as.numeric(data$party.1) >= 6] <- NA
```

```
#Political Party 2
summary(data$party.2)
```

```
## 0. DK; NA; other; refused to answer; no Pre IW
##                                             90
##               1. Democrats (including leaners)
##                                           6179
##                                2. Independents
##                                           1400
##             3. Republicans (including leaners)
##                                           4098
```

```
data$party.2[data$party.2 == '0. DK; NA; other; refused to answer; no Pre IW'] <- NA
```

```
#Voted in Election
summary(data$did.vote)
```

```
## 0. DK; NA; no Post IW; refused to say if voted;
##                                              23
##                             1. No, did not vote
##                                            2565
##                                   2. Yes, voted
##                                            9179
```

```
data$did.vote[data$did.vote == '0. DK; NA; no Post IW; refused to say if voted;'] <- NA
```

```
#Registered
summary(data$reg.vote)
```

```
## 0. DK/NA if voted; DK/NA whether registered (includes
##                                                    34
##                    1. Not registered, and did not vote
##                                                   1279
##                        2. Registered, but did not vote
##                                                   1275
##                                  3. Voted (registered)
##                                                   9179
```

```r
data$reg.vote[data$reg.vote == '0. DK/NA if voted; DK/NA whether registered (includes'] <- NA
```

```r
#Registered Indicator
#Create indicator variable of whether or not someone is registered to vote
data$registered = ifelse(data$reg.vote == "2. Registered, but did not vote"|
                         data$reg.vote == "3. Voted (registered)",
                         "Yes", "No")

data$registered = as.factor(data$registered)
summary(data$registered)
```

```
##    No   Yes  NA's
##  1279 10454    34
```

```r
#Presidential Vote
summary(data$pres.vote)
```

```
## 0. Did not vote; DK/NA if voted; refused to say if
##                                                3094
##                                         1. Democrat
##                                                5100
##                                       2. Republican
##                                                3573
```

```r
#Presidential Vote Indicator
#Create indicator variable of whether or not someone voted for president
data = data[!is.na(data$did.vote), ]
data$did.vote.pres = ifelse(data$pres.vote == '0. Did not vote; DK/NA if voted; refused to say if',
                            'Did not vote for president', 'Voted for president')

data$did.vote.pres = as.factor(data$did.vote.pres)
summary(data$did.vote.pres)
```

```
## Did not vote for president        Voted for president
##                       3071                       8673
```

```r
#Presidential Vote Binary
#Create variable for whether someone voted for the Republican candidate or the Democratic one
data$pres.vote2 = ifelse(data$pres.vote == '1. Democrat', 'Dem',
                         ifelse(data$pres.vote == '2. Republican', 'Rep', NA))

data$pres.vote2 = as.factor(data$pres.vote2)
summary(data$pres.vote2)
```

```
##  Dem  Rep NA's
## 5100 3573 3071
```

```r
#Previous Vote
summary(data$previous.vote)
```

```
## 0. R did not vote in previous election; R has never voted
##                                                       1821
##                       1. Voted: Democratic Pres. Candidate
##                                                       2112
##                       2. Voted: Republican Pres. Candidate
##                                                       1762
##             3. Voted: DK/NA/Refused which Pres. Candidate
##                                                         95
##                                  5. Voted: Other candidate
##                                                        398
## 9. DK/NA/refused to say if voted in previous presidential
##                                                         68
##                                                       NA's
##                                                       5488
```

```r
data$previous.vote[data$previous.vote == '9. DK/NA/refused to say if voted in previous presidential'] <- NA

#Previous Vote Indicator
#Create indicator variable of whether or not someone voted for president in the previous election
data$previous.did.vote = ifelse(data$previous.vote == '0. R did not vote in previous election; R has never voted',
                                'Did not vote', 'Voted')

data$previous.did.vote = as.factor(data$previous.did.vote)
summary(data$previous.did.vote)
```

```
## Did not vote        Voted         NA's
##          1821         4367         5556
```

```
#Previous Vote Candidate
#Create variable for who someone voted for president in the previous election
data$previous.pres.vote = ifelse(data$previous.vote == '1. Voted: Democratic Pres. Candidate', 'Dem',
                                 ifelse(data$previous.vote == '2. Voted: Republican Pres. Candidate', 'Rep', NA))

data$previous.pres.vote = as.factor(data$previous.pres.vote)
summary(data$previous.pres.vote)
```

```
##  Dem  Rep NA's
## 2112 1762 7870
```

```
#Home Ownership
summary(data$home.own)
```

```
##                              1. Yes, own
##                                     7796
##                        2. No, not owned
##                                     3906
##                                    8. DK
##                                        0
## 9. NA; RF; no Pre IW; short form (1992)
##                                        0
##                                     NA's
##                                       42
```

```
#As you can see in the summary of home.own for example, there are multiple unused levels (i.e. frequency = 0)
#Fortunately, there is an easy function that drops all unused levels
data = droplevels(data)

#Final Data Summary
summary(data)
```

```
##        id                  year
##  Min.   :19920002   Min.   :1996
##  1st Qu.:20001622   1st Qu.:2000
##  Median :20081902   Median :2008
##  Mean   :20068135   Mean   :2007
##  3rd Qu.:20123685   3rd Qu.:2012
##  Max.   :20126864   Max.   :2012
##
##                                       post
##  1. Post-election interview data present:11744
##
##
##
##
##
##
##                                   method          age
##  0. All personal                     :6558   Min.   :17.00
##  2. Telephone post (personal pre)    : 865   1st Qu.:35.00
##  3. All telephone                    : 748   Median :48.00
##  4. All internet (2012: pre and post):3573   Mean   :48.57
##                                              3rd Qu.:61.00
##                                              Max.   :96.00
##                                              NA's   :103
##       age.group          gender                        race
##  4. 45 - 54:2202   1. Male  :5436   1. White non-Hispanic:7482
##  5. 55 - 64:2150   2. Female:6308   2. Black non-Hispanic:1928
##  3. 35 - 44:2116                    3. Other             :2258
##  2. 25 - 34:1901                    NA's                 :  76
##  6. 65 - 74:1472
##  (Other)   :1800
##  NA's      : 103
##                                                     educ
##  1. Grade school or less (0-8 grades)                 : 351
##  2. High school (12 grades or fewer, incl. non-college:4200
##  3. Some college (13 grades or more but no degree;    :3704
##  4. College or advanced degree (no cases 1948)        :3410
##  NA's                                                 :  79
##
##
##                                                       region
##  1. Northeast (CT, ME, MA, NH, NJ, NY, PA, RI, VT)       :1814
##  2. North Central (IL, IN, IA, KS, MI, MN, MO, NE, ND,   :2633
##  3. South (AL, AR, DE, D.C., FL, GA, KY, LA, MD, MS, NC  :4603
##  4. West (AK, AZ, CA, CO, HI, ID, MT, NV, NM, OR, UT, WA,:2694
##
##
##
##                       income
##  1. 0 to 16 percentile  :2012
##  2. 17 to 33 percentile :1954
##  3. 34 to 67 percentile :3868
##  4. 68 to 95 percentile :2547
##  5. 96 to 100 percentile: 559
##  NA's                   : 804
##
##                                                         work
##  1. Employed                                             :6741
##  2. Not employed: laid off, unemployed, on strike,       :1391
##  3. Retired                                              :2371
##  4. Homemaker (since 1972: not working 20 or more hrs/wk;: 811
##  5. Student (since 1972: not working 20 or more hrs/wk;  : 411
##  NA's                                                    :  19
##
##                                                      union
##  1. Yes, someone (1948: head) in household belongs to a:1787
##  2. No, no one in household belongs to a labor union   :9907
##  NA's                                                  :  50
##
##
##
##
##                                              religion     num.children
##  1. Protestant                                  :5451   Min.   :0.0000
##  2. Catholic [Roman Catholic]                   :2788   1st Qu.:0.0000
##  3. Jewish                                      : 221   Median :0.0000
##  4. Other and none (also includes DK preference):3146   Mean   :0.6011
##  NA's                                           : 138   3rd Qu.:1.0000
##                                                         Max.   :3.0000
##                                                         NA's   :2761
##  parents.native             home.own
##  1. Yes:9706    1. Yes, own     :7796
##  5. No :2003    2. No, not owned:3906
##  NA's  :  35    NA's            :  42
```

```
##
##
##
##
##                                                  marital.stat
##  1. Married                                            :5819
##  2. Never married                                      :2270
##  3. Divorced                                           :1558
##  4. Separated                                          : 360
##  5. Widowed                                            :1131
##  7. Partners; not married (VOLUNTEERED [exc.1986,2012]): 573
##  NA's                                                  :  33
##                                                  donate
##  1. No (includes 'not asked for money' in 1966,1968):10334
##  2. Yes (includes 'tax check-off' in 1976)          : 1406
##  NA's                                               :    4
##
##
##
##
##      state         district                                party.1
##  CA     :1340   VA09   :  125   1. Republican                  :2847
##  TX     :1131   MN01   :  124   2. Independent                 :3514
##  FL     : 744   CA04   :  117   3. No preference; none; neither: 502
##  NY     : 636   WI04   :  105   4. Other                       : 214
##  MI     : 466   LA04   :  103   5. Democrat                    :4540
##  (Other):7425   (Other):10985   NA's                           : 127
##  NA's   :   2   NA's   :  185
##                                party.2                    did.vote
##  1. Democrats (including leaners)  :6174   1. No, did not vote:2565
##  2. Independents                   :1385   2. Yes, voted      :9179
##  3. Republicans (including leaners):4096
##  NA's                              :  89
##
##
##
##                                 reg.vote
##  1. Not registered, and did not vote:1279
##  2. Registered, but did not vote    :1269
##  3. Voted (registered)              :9179
##  NA's                               :  17
##
##
##
##                                               pres.vote
##  0. Did not vote; DK/NA if voted; refused to say if:3071
##  1. Democrat                                       :5100
##  2. Republican                                     :3573
##
##
##
##
##                                                    previous.vote
##  0. R did not vote in previous election; R has never voted:1821
##  1. Voted: Democratic Pres. Candidate                     :2112
##  2. Voted: Republican Pres. Candidate                     :1762
##  3. Voted: DK/NA/Refused which Pres. Candidate            :  95
##  5. Voted: Other candidate                                : 398
##  NA's                                                     :5556
##
##  children.ind registered                      did.vote.pres  pres.vote2
##  No  :6024    No  : 1279   Did not vote for president:3071   Dem :5100
##  Yes :2959    Yes :10448   Voted for president       :8673   Rep :3573
##  NA's:2761    NA's:   17                                     NA's:3071
##
##
##
##
##     previous.did.vote previous.pres.vote
##  Did not vote:1821    Dem :2112
##  Voted       :4367    Rep :1762
##  NA's        :5556    NA's:7870
##
##
##
##
```

```r
write.csv(data, "ANES Final Data.csv", row.names = F)
```