Re-leveling hd variable

Qianli Sun

1) tranform extra columns of data we found according to new 13 hd levels

```
#read in new training dataset

data_train <- read.csv('train_nafill_v2.csv')
#calculate hd population for re-leveled hd districts
pop <- aggregate(hd.population ~ hd, FUN = mean, data = data_train)[,
"hd.population"]
#calculate area for re-leveled hd districts
area <- aggregate(hd.area ~ hd, FUN = mean, data = data_train)[, "hd.area"]
#calculate housing unit for re-leveled hd districts
housing_unit <- aggregate(hd.housing_unit ~ hd, FUN = mean, data =
data_train)[, "hd.housing_unit"]

group <- c(rep(1, 8), 2, rep(3, 4), rep(4, 8), rep(5, 7), rep(6, 6), rep(7,
7),
           rep(8, 4), rep(9, 2), rep(10, 3), rep(11, 4), rep(12, 9), rep(13,
2))

#output
hd.relevel <- rowsum(cbind(pop, area, housing_unit), group)
hd.relevel <- cbind(hd.relevel, city = as.numeric(hd.relevel[,"pop"] /
hd.relevel[,"area"] > 1000))

#add the hd column for matching in step 3
hd = c(1:13)
hd.relevel = data.frame(cbind(hd, hd.relevel))
```

2) Re-level hd into smaller number of groups

```
##original data
data.train = read.csv('train.csv')
data.test = read.csv('test.csv')

##read in imputed data
train_X = read.csv("train_nafill_X.csv")
test_X = read.csv("test_nafill_X.csv")

#intially have 65 hd levels
unique(train_X$hd)
```

```
##  [1] 31 38 53 30 19  7 13 52 39  5 55 24 40 11  3  1 43 18  2 62 61 59 49
## [24] 34 60 29 35 28  8 14 25 51  6 45 22 56 17 10 58 32 44 65 48 26 42  4
## [47] 37 57  9 47 36 27 15 33 46 41 64 20 16 23 21 63 12 50 54
```

```
typeof(train_X$hd)
```

```
## [1] "integer"
```

```
# re-categorizing the hd variable for training set
train_X$hd[train_X$hd==1 | train_X$hd==2 |train_X$hd==3 | train_X$hd==4
|train_X$hd==5
           |train_X$hd==6|train_X$hd==7 | train_X$hd==8] <- 1
train_X$hd[train_X$hd==9] <- 2
train_X$hd[train_X$hd==10 | train_X$hd==11 |train_X$hd==12 | train_X$hd==13]
<- 3
train_X$hd[train_X$hd==14 | train_X$hd==15 |train_X$hd==16 | train_X$hd==17
|train_X$hd==18
           |train_X$hd==19|train_X$hd==20 | train_X$hd==21] <- 4
train_X$hd[train_X$hd==22 | train_X$hd==23 |train_X$hd==24 | train_X$hd==25
|train_X$hd==26
           |train_X$hd==27|train_X$hd==28] <- 5
train_X$hd[train_X$hd==29 | train_X$hd==30 |train_X$hd==31 | train_X$hd==32
|train_X$hd==33
           |train_X$hd==34] <- 6
train_X$hd[train_X$hd==35 | train_X$hd==36 |train_X$hd==37 | train_X$hd==38
|train_X$hd==39
           |train_X$hd==40 | train_X$hd==41] <- 7
train_X$hd[train_X$hd==42 | train_X$hd==43 |train_X$hd==44 | train_X$hd==45]
<- 8
train_X$hd[train_X$hd==46 | train_X$hd==47] <- 9
train_X$hd[train_X$hd==48 | train_X$hd==49 |train_X$hd==50] <- 10
train_X$hd[train_X$hd==51 | train_X$hd==52 |train_X$hd==53 | train_X$hd==54]
<- 11
train_X$hd[train_X$hd==55 | train_X$hd==56 |train_X$hd==57 | train_X$hd==58
|train_X$hd==59
           |train_X$hd==60|train_X$hd==61|train_X$hd==62|train_X$hd==63] <-
12
train_X$hd[train_X$hd==64 | train_X$hd==65] <- 13
```

```
# re-categorizing the hd variable for test set
test_X$hd[test_X$hd==1 | test_X$hd==2 |test_X$hd==3 | test_X$hd==4
|test_X$hd==5
          |test_X$hd==6|test_X$hd==7 | test_X$hd==8] <- 1
test_X$hd[test_X$hd==9] <- 2
test_X$hd[test_X$hd==10 | test_X$hd==11 |test_X$hd==12 | test_X$hd==13] <- 3
test_X$hd[test_X$hd==14 | test_X$hd==15 |test_X$hd==16 | test_X$hd==17
```

```
|test_X$hd==18
          |test_X$hd==19|test_X$hd==20 | test_X$hd==21] <- 4
test_X$hd[test_X$hd==22 | test_X$hd==23 |test_X$hd==24 | test_X$hd==25
|test_X$hd==26
          |test_X$hd==27|test_X$hd==28] <- 5
test_X$hd[test_X$hd==29 | test_X$hd==30 |test_X$hd==31 | test_X$hd==32
|test_X$hd==33
          |test_X$hd==34] <- 6
test_X$hd[test_X$hd==35 | test_X$hd==36 |test_X$hd==37 | test_X$hd==38
|test_X$hd==39
          |test_X$hd==40 | test_X$hd==41] <- 7
test_X$hd[test_X$hd==42 | test_X$hd==43 |test_X$hd==44 | test_X$hd==45] <- 8
test_X$hd[test_X$hd==46 | test_X$hd==47] <- 9
test_X$hd[test_X$hd==48 | test_X$hd==49 |test_X$hd==50] <- 10
test_X$hd[test_X$hd==51 | test_X$hd==52 |test_X$hd==53 | test_X$hd==54] <- 11
test_X$hd[test_X$hd==55 | test_X$hd==56 |test_X$hd==57 | test_X$hd==58
|test_X$hd==59
          |test_X$hd==60|test_X$hd==61|test_X$hd==62|test_X$hd==63] <- 12
test_X$hd[test_X$hd==64 | test_X$hd==65] <- 13
```

3) Add extra variable from step 1) by matching according to new hd levels

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

train_full_X = left_join(train_X, hd.relevel , by = c("hd"))
## add "voted" column to train data
train_full_relevel = cbind(voted = data.train$voted, train_full_X)

#write the new csv file
write.csv(train_full_relevel, "relevel-train.full.csv",row.names = F)


test_full_X  = left_join(test_X, hd.relevel , by = c("hd"))
## add "id" column to test data
test_full_relevel  = cbind(test_full_X, Id = data.test$Id)
```

```r
#write the new csv file
write.csv(test_full_relevel, "relevel-test.full.csv", row.names = F)
```