nych_project_data_prep_midterm.R

ksmit258

Wed Apr 03 08:08:40 2019

```
###Data wrangling/tidying###
###MTH 390Q Midterm Exam###

#Elizabeth Foster
#Kelsey Smith
#Gemma Hoeppner
#Nathanael Whitney
#Larry Breeden


###Loading packages###

suppressWarnings(suppressMessages(library(dplyr)))

###Reading in files###

temp <- file.choose()
nych91 <- read.csv(temp, skip = 2, header = F)
header <- read.csv(temp, header = T)
names(nych91) <- names(header)
rm(header)

temp <- file.choose()
nych93 <- read.csv(temp, skip = 2, header = F)
header <- read.csv(temp, header = T)
names(nych93) <- names(header)
rm(header)

temp <- file.choose()
nych96 <- read.csv(temp, skip = 2, header = F)
header <- read.csv(temp, header = T)
names(nych96) <- names(header)
rm(header)

temp <- file.choose()
nych99 <- read.csv(temp, skip = 2, header = F)
header <- read.csv(temp, header = T)
names(nych99) <- names(header)
rm(header)

temp <- file.choose()
nych02 <- read.csv(temp, skip = 2, header = F)
header <- read.csv(temp, header = T)
names(nych02) <- names(header)
rm(header)

temp <- file.choose()
```

```
nych05 <- read.csv(temp, skip = 2, header = F)
header <- read.csv(temp, header = T)
names(nych05) <- names(header)
rm(header)

temp <- file.choose()
nych08 <- read.csv(temp, skip = 2, header = F)
header <- read.csv(temp, header = T)
names(nych08) <- names(header)
rm(header)

temp <- file.choose()
nych11 <- read.csv(temp, skip = 2, header = F)
header <- read.csv(temp, header = T)
names(nych11) <- names(header)
rm(header)

temp <- file.choose()
nych14 <- read.csv(temp, skip = 2, header = F)
header <- read.csv(temp, header = T)
names(nych14) <- names(header)
rm(header)

temp <- file.choose()
nych17 <- read.csv(temp, skip = 2, header = F)
header <- read.csv(temp, header = T)
names(nych17) <- names(header)
rm(header)

###Selecting variable columns from each and adding column for year###

nych91 <- nych91 %>%
  select(borough, X_d3, X_d4, X_e1, X_e2, X_e3, X_f1, X_f2,
         X_g3, X_g4, X_25a, X_25c, X_26a, X_26c, X_32a,
         X_35a, X_36a, X_36b, X_37a, X_37b, X_38a) %>%
  mutate(year = 1991)

nych93 <- nych93 %>%
  select(borough, X_d3, X_d4, X_e1, X_e2, X_e3, X_f1, X_f2,
         X_g3, X_g4, X_25a, X_25c, X_26a, X_26c, X_32a,
         X_35a, X_36a, X_36b, X_37a, X_37b, X_38a) %>%
  mutate(year = 1993)

nych96 <- nych96 %>%
  select(borough, X_d3, X_d4, X_e1, X_e2, X_e3, X_f1, X_f2,
         X_g3, X_g4, X_25a, X_25c, X_26a, X_26c, X_32a,
         X_35a, X_36a, X_36b, X_37a, X_37b, X_38a) %>%
  mutate(year = 1996)

nych99 <- nych99 %>%
  select(borough, X_d3, X_d4, X_e1, X_e2, X_e3, X_f1, X_f2,
         X_g3, X_g4, X_25a, X_25c, X_26a, X_26c, X_32a,
         X_35a, X_36a, X_36b, X_37a, X_37b, X_38a) %>%
```

```
  mutate(year = 1999)

nych02 <- nych02 %>%
  select(borough, X_d3, X_d4, X_e1, X_e2, X_e3, X_f1, X_f2,
         X_g3, X_g4, X_25a, X_25c, X_26a, X_26c, X_32a,
         X_35a, X_36a, X_36b, X_37a, X_37b, X_38a) %>%
  mutate(year = 2002)

nych05 <- nych05 %>%
  select(borough, X_d3, X_d4, X_e1, X_e2, X_e3, X_f1, X_f2,
         X_g3, X_g4, X_25a, X_25c, X_26a, X_26c, X_32a,
         X_35a, X_36a, X_36b, X_37a, X_37b, X_38a) %>%
  mutate(year = 2005)

nych08 <- nych08 %>%
  select(borough, X_d3, X_d4, X_e1, X_e2, X_e3, X_f1, X_f2,
         X_g3, X_g4, X_25a, X_25c, X_26a, X_26c, X_32a,
         X_35a, X_36a, X_36b, X_37a, X_37b, X_38a) %>%
  mutate(year = 2008)

nych11 <- nych11 %>%
  select(borough, X_d3, X_d4, X_e1, X_e2, X_e3, X_f1, X_f2,
         X_g3, X_g4, X_25a, X_25c, X_26a, X_26c, X_32a,
         X_35a, X_36a, X_36b, X_37a, X_37b, X_38a) %>%
  mutate(year = 2011)

nych14 <- nych14 %>%
  select(borough, X_d3, X_d4, X_e1, X_e2, X_e3, X_f1, X_f2,
         X_g3, X_g4, X_25a, X_25c, X_26a, X_26c, X_32a,
         X_35a, X_36a, X_36b, X_37a, X_37b, X_38a) %>%
  mutate(year = 2014)

nych17 <- nych17 %>%
  select(borough, X_d3, X_d4, X_e1, X_e2, X_e3, X_f1, X_f2,
         X_g3, X_g4, X_25a, X_25c, X_26a, X_26c, X_32a,
         X_35a, X_36a, X_36b, X_37a, X_37b, X_38a) %>%
  mutate(year = 2017)

###Combining into one dataframe###

nych_all <- rbind(nych91, nych93, nych96, nych99, nych02,
                  nych05, nych08, nych11, nych14, nych17)

names(nych_all) <- c("Borough", "Walls_cracks", "Loose_roof", "Broken_windows",
                     "Rotten_windows", "Boarded_windows", "Broken_railings", "Broken_steps",
                     "Floor_wear", "Floor_missing", "Plumbing", "Toilet_breakdowns",
                     "Kitchen_facilities", "Kitchen_functioning", "Heating_breakdowns",
                     "Mice_rats", "Walls_holes", "Floor_holes", "Small_broken_plaster",
                     "Large_broken_plaster", "Water_leak", "Year")

###Replacing "condition not reported" values with NA###
###Replacing borough values with names###
```

```
nych_all[nych_all == 8] <- NA
nych_all$Borough[nych_all$Borough == 1] <- "Bronx"
nych_all$Borough[nych_all$Borough == 2] <- "Brooklyn"
nych_all$Borough[nych_all$Borough == 3] <- "Manhattan"
nych_all$Borough[nych_all$Borough == 4] <- "Queens"
nych_all$Borough[nych_all$Borough == 5] <- "Staten Island"


###Assigning weights###

###Columns 2-10###
nych_all[, 2:10][nych_all[, 2:10] == 9] <- 101
nych_all$Walls_cracks[nych_all$Walls_cracks == 1] <- 8
nych_all$Loose_roof[nych_all$Loose_roof == 1] <- 5
nych_all$Broken_windows[nych_all$Broken_windows == 1] <- 7
nych_all$Rotten_windows[nych_all$Rotten_windows == 1] <- 4
nych_all$Boarded_windows[nych_all$Boarded_windows == 1] <- 5
nych_all$Broken_railings[nych_all$Broken_railings == 1] <- 7
nych_all$Broken_steps[nych_all$Broken_steps == 1] <- 7
nych_all$Floor_wear[nych_all$Floor_wear == 1] <- 3
nych_all$Floor_missing[nych_all$Floor_missing == 1] <- 9
nych_all[, 2:10][nych_all[, 2:10] == 101] <- 1

###Column 11###
nych_all$Plumbing[nych_all$Plumbing == 1] <- 5
nych_all$Plumbing[nych_all$Plumbing == 0] <- 1
nych_all$Plumbing[nych_all$Plumbing == 2] <- 10

###Column 12###
nych_all$Toilet_breakdowns[nych_all$Toilet_breakdowns == 1] <- 5
nych_all$Toilet_breakdowns[nych_all$Toilet_breakdowns == 2] <- 1
nych_all$Toilet_breakdowns[nych_all$Toilet_breakdowns == 3] <- 8
nych_all$Toilet_breakdowns[nych_all$Toilet_breakdowns == 9] <- 0

###Column 13###
nych_all$Kitchen_facilities[nych_all$Kitchen_facilities == 2] <- 4
nych_all$Kitchen_facilities[nych_all$Kitchen_facilities == 1] <- 2
nych_all$Kitchen_facilities[nych_all$Kitchen_facilities == 0] <- 1
nych_all$Kitchen_facilities[nych_all$Kitchen_facilities == 3] <- 10

###Column 14###
nych_all$Kitchen_functioning[nych_all$Kitchen_functioning == 2] <- 5
nych_all$Kitchen_functioning[nych_all$Kitchen_functioning == 9] <- 0

###Columns 15-19###
nych_all$Heating_breakdowns[nych_all$Heating_breakdowns == 0] <- 7
nych_all$Mice_rats[nych_all$Mice_rats == 0] <- 6
nych_all$Walls_holes[nych_all$Walls_holes == 0] <- 3
nych_all$Floor_holes[nych_all$Floor_holes == 0] <- 5
nych_all$Small_broken_plaster[nych_all$Small_broken_plaster == 0] <- 2

###Column 20###
nych_all$Large_broken_plaster[nych_all$Large_broken_plaster == 3] <- 0
```

```r
nych_all$Large_broken_plaster[nych_all$Large_broken_plaster == 2] <- 3
nych_all$Large_broken_plaster[nych_all$Large_broken_plaster == 9] <- 0

###Column 21###
nych_all$Water_leak[nych_all$Water_leak == 0] <- 8


###Add column for index value###
nych_all <- mutate(nych_all, index_value = rowSums(nych_all[ , 2:21], na.rm = T))

###Removing rows with NAs###
nych_all_cleaned <- nych_all[complete.cases(nych_all), ]

###Finding general stats###
mean(nych_all_cleaned$index_value)
```

```
## [1] 25.37974
```

```r
median(nych_all_cleaned$index_value)
```

```
## [1] 23
```

```r
min(nych_all_cleaned$index_value)
```

```
## [1] 19
```

```r
max(nych_all_cleaned$index_value)
```

```
## [1] 88
```

```r
#If we were to divide the possible spread of our index value into three portions,
#and then count how many units fell into each segment, that would tell us
#how many units are in the top third, middle third, and bottom third.

#Below I am determining spread, given a max possible index of 122, and min possible
#index of 17.

(122 - 17)/3
```

```
## [1] 35
```

```r
#Below I am establishing the three divided segments, and then determining the
#number of units within each segment.

#17 + 35 = 52
sum(nych_all_cleaned$index_value <= 52)
```

```
## [1] 119013
```

```r
#52 + 35 = 87
sum(nych_all_cleaned$index_value > 52 & nych_all_cleaned$index_value <= 87)
```

```
## [1] 252
```

```r
#87 + 35 = 122
sum(nych_all_cleaned$index_value > 87)
```

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

```
###Finding the proportion (%) within each bracket###
(sum(nych_all_cleaned$index_value <= 52)/nrow(nych_all_cleaned))*100
```

```
## [1] 99.78787
```

```
(sum(nych_all_cleaned$index_value > 52 & nych_all_cleaned$index_value <= 87)/
    nrow(nych_all_cleaned))*100
```

```
## [1] 0.2112924
```

```
(sum(nych_all_cleaned$index_value > 87)/nrow(nych_all_cleaned))*100
```

```
## [1] 0.0008384619
```

```
#Given the ultra small percentages of the housing within the upper two thirds,
#this implies that perhaps we should rework our index strategy so that we see
#more spread, or look at what index values should accurately represent the
#bottom, middle, and top (i.e. perhaps it shouldn't be thirds divided evenly).
```