US School Attendence Data Cleaning

Surabhi Chouhan

January 12, 2017

```
library("gdata")
```

```
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

##

## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.

##
## Attaching package: 'gdata'

## The following object is masked from 'package:stats':
##
##     nobs

## The following object is masked from 'package:utils':
##
##     object.size

## The following object is masked from 'package:base':
##
##     startsWith
```

```
library("stringr")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:gdata':
##
##     combine, first, last

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

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

import data and exploring and understanding the data

```
att <- read.xls("D:/Surabhi docs/portfolio/data cleaning/attendance.xls", perl= "C:/Strawberry/perl/bin

colnames(att)
```

```
##  [1] "Table.43..Average.daily.attendance..ADA..as.a.percentage.of.total.enrollment..school.day.lengt
##  [2] "X"
##  [3] "X.1"
##  [4] "X.2"
##  [5] "X.3"
##  [6] "X.4"
##  [7] "X.5"
##  [8] "X.6"
##  [9] "X.7"
## [10] "X.8"
## [11] "X.9"
## [12] "X.10"
## [13] "X.11"
## [14] "X.12"
## [15] "X.13"
## [16] "X.14"
## [17] "X.15"
```

```r
head(att)
```

```
##   Table.43..Average.daily.attendance..ADA..as.a.percentage.of.total.enrollment..school.day.length..an
## 1
## 2
## 3
## 4
## 5
## 6
##                                                                         X
## 1 Total elementary, secondary, and combined elementary/secondary schools
## 2                                            ADA as percent of enrollment
## 3                                                                       2
## 4                                                                    93.1
## 5                                                                    93.8
## 6                                                                    89.9
##      X.1                         X.2    X.3                         X.4
## 1
## 2        Average hours in school day        Average days in school year
## 3                                  3                                  4
## 4 (0.22)                         6.6 (0.02)                         180
## 5 (1.24)                         7.0 (0.07)                         180
## 6 (1.22)                         6.5 (0.05)                         180
##     X.5                          X.6    X.7                          X.8
## 1                                                      Elementary schools
## 2       Average hours in school year        ADA as percent of enrollment
## 3                                  5                                   6
## 4 (0.1)                        1,193  (3.1)                         94.0
## 5 (0.8)                        1,267 (12.3)                         93.8
## 6 (3.4)                        1,163 (22.9)                         91.3
##      X.9                        X.10   X.11                         X.12
## 1                                                       Secondary schools
## 2        Average hours in school day        ADA as percent of enrollment
## 3                                  7                                   8
## 4 (0.27)                         6.7 (0.02)                         91.1
## 5 (1.84)                         7.0 (0.08)                         94.6
## 6 (1.56)                         6.5 (0.05)                         93.2
##     X.13                        X.14   X.15
## 1
## 2        Average hours in school day
## 3                                  9
## 4 (0.43)                         6.6 (0.04)
## 5 (0.38)                         7.1 (0.17)
## 6 (1.57)                         6.2 (0.15)
```

```r
tail(att)
```

```
##                                                                             Table.43..Average.daily.at
## 54
## 55
## 56
## 57
## 58 NOTE: Averages reflect data reported by schools rather than state requirements. School-reported le
## 59
##                                                                X    X.1 X.2
## 54                                                          95.0 (0.57) 6.9
## 55                                                          92.4 (1.15) 6.9
## 56
## 57
## 58
## 59 \\ 2003-04 and 2007-08. (This table was prepared June 2011.)
##       X.3 X.4   X.5   X.6   X.7  X.8    X.9 X.10   X.11 X.12   X.13 X.14
## 54 (0.04) 180 (0.7) 1,246 (8.6) 95.4 (0.41)  6.9 (0.05) 93.0 (1.91)  7.0
## 55 (0.05) 175 (1.3) 1,201 (8.3) 92.2 (1.65)  6.9 (0.05) 92.4 (0.75)  7.0
## 56
## 57
## 58
## 59
##      X.15
## 54 (0.14)
## 55 (0.07)
## 56
## 57
## 58
## 59
```

```r
str(att)
```

```
## 'data.frame':    59 obs. of  17 variables:
##  $ Table.43..Average.daily.attendance..ADA..as.a.percentage.of.total.enrollment..school.day.length..a
##  $ X
##  $ X.1
##  $ X.2
##  $ X.3
##  $ X.4
##  $ X.5
##  $ X.6
##  $ X.7
##  $ X.8
##  $ X.9
##  $ X.10
##  $ X.11
##  $ X.12
##  $ X.13
##  $ X.14
##  $ X.15
```

Get rid of all the unnecessary rows

```
rem_row <- c(3,56,57,58,59)
att2 <- att[-(rem_row),]
```

Get rid of unnecessary columns

```
#Does not contain attendence data
rem_col <-  c(3,5,7,9,11,13,15,17)
att3 <- att2[,-(rem_col)]
```

Splitting the data - to separate the data of elementary schools, secondary schools and all schools

```
# Subset just elementary schools: att_elem
att_elem <- att3[,c(1,6,7)]

# Subset just secondary schools: att_sec
att_sec <- att3[,c(1,8,9)]

# Subset all schools: att4
att4 <- att3[,c(1:5)]
```

Replacing the names of columns

```
cnames <- c("state", "avg_attend_pct", "avg_hr_per_day",
            "avg_day_per_yr", "avg_hr_per_yr")

colnames(att4) <- cnames

# Remove first two rows of att4: att5 as it does not contain any useful data
att5 <- att4[-(1:2),]
colnames(att5)
```

```
## [1] "state"          "avg_attend_pct" "avg_hr_per_day" "avg_day_per_yr"
## [5] "avg_hr_per_yr"
```

One of the chafracteristic of this messy data is that in gthe state names, extra characters have been added to make the length of all the state names to be same. We need to remove those extra characters to make the data useful for us

```
# Remove all periods in state column
att5$state <- str_replace_all(att5$state, "\\.", "")

# Remove white space around state names
att5$state <- str_trim(att5$state)
```

```
head(att5)
```

```
##           state avg_attend_pct avg_hr_per_day avg_day_per_yr avg_hr_per_yr
## 4 United States           93.1            6.6            180         1,193
## 5       Alabama           93.8            7.0            180         1,267
## 6        Alaska           89.9            6.5            180         1,163
## 7       Arizona           89.0            6.4            181         1,159
## 8      Arkansas           91.8            6.9            179         1,229
## 9    California           93.2            6.2            181         1,129
```

Some final type conversions

```
# Define vector containing numerical columns: cols
cols <- c(2:5)

# Use sapply to coerce cols to numeric
att5[, cols] <- sapply(att5[, cols], as.numeric)
```

Now that our data is clean, lets take a look at how the data looked like originally and how it looks after cleaning

```
#original messy data
str(att)
```

```
## 'data.frame':    59 obs. of  17 variables:
##  $ Table.43..Average.daily.attendance..ADA..as.a.percentage.of.total.enrollment..school.day.length..a
##  $ X
##  $ X.1
##  $ X.2
##  $ X.3
##  $ X.4
##  $ X.5
##  $ X.6
##  $ X.7
##  $ X.8
##  $ X.9
##  $ X.10
##  $ X.11
##  $ X.12
##  $ X.13
##  $ X.14
##  $ X.15
```

```
#clean data
str(att5)
```

```
## 'data.frame':    52 obs. of  5 variables:
##  $ state         : chr  "United States" "Alabama" "Alaska" "Arizona" ...
##  $ avg_attend_pct: num  22 28 8 6 14 23 29 5 7 11 ...
##  $ avg_hr_per_day: num  7 11 6 5 10 3 11 6 8 10 ...
##  $ avg_day_per_yr: num  10 10 10 11 9 11 2 11 11 11 ...
##  $ avg_hr_per_yr : num  26 45 15 13 36 5 28 19 31 42 ...
```