United States School Statistics: Chicago, Illinois IT497 OSEMN Assignment

Sonali Changkakoti
School of Information Technology
Illinois State University
schangk@ilstu.edu

October 22, 2014

1 Introduction

Schools in the United States comprise of both public as well as private schools. The funding and control of public schools are done by state, local and federal government. Their curricula and staffing are decided by the locally elected school boards. On the other hand, private schools are generally free to determine their own curricula and staffing policies. There are also charter schools, which receive public funding but operate independently. Some states do not have charter school authorization. Around 88% of school-aged students attend public schools, 9% attend private schools and rest 3% are homeschooled.

Education in the United States is compulsory over an age range, which varies from state to state. Formal education is divided into a number of stages. Schools are divided into three groups- elementary school, middle or junior high school and high school. The elementary school are from kindergarten to fifth grade, middle school are from 6th grade to 8th grade and high school are from 9th grade to 12th grade.Children may begin with pre-kindergarten, kindergarten or first grade. The compulsory education is till 12th grade, after that students can pursue higher education in colleges or universities.

The data are collected from the National Center for Education Statistics for the number of schools, students, and teachers in regular schools with membership for the 100 largest cities in the United States, by school operational and charter status and state for school years ending 2001 through 2011. Here, the data from Chicago, IL among the 100 largest cities in the United States has been chosen for analysis.

2 Data

The data is collected for examining the total numbers of schools, teachers and schools in Chicago, IL over the period of 10 years, from December, 2011 to

December, 2011. The required data is reteived from the Quandl API. Libraries like knitr, ggplot2 and reshape2 are used.

1. Obtaing the data

   The dataset is downloaded by using the Quandl API and auth token in a csv format. It can be also access through the URL (https//www.quandl.com/NCES/SCHOOLS_CITIES_CHICAGOILLINOIS) The data is stored in myData object and it will be used for further analysis.

   ```r
   # Loading knitr and Quandl
   library(knitr)
   library(Quandl)

   # To download data from a secure URL using Quandl API
   # and authcode

   myData <- Quandl("NCES/SCHOOLS_CITIES_CHICAGOILLINOIS",
                    authcode="sUMBj-Lb7MRzwnUWXvxe")
   ```

2. Scrubing data (Cleaning data)

   Scrubing and cleaning of the data is done to get only the relevant data needed to obtain the results. Irrelevant data make the analyis difficult. Relevant data can be obtained by just looking into it and determining the requirements of the report. All the rows containing column 1 through column 4 are chosen and stored in a new object named cleanData.

   ```r
   # Cleaning the data needed to plot a graph showing total
   # students, total teachers and total schools in Chicago,
   # Illinois
   cleanData<-myData[,1:4]
   ```

   The columns' name are also changed for convenience and readability. The cleaned data is shown as below:

   ```r
   # Changing the columns' name
   colnames(cleanData) <- c("year", "schools", "students",
                            "teachers")
   # Showing the contents of cleanData
   cleanData

   ##          year schools students teachers
   ## 1  2011-12-31     620   400383 21847.46
   ## 2  2010-12-31     614   402951 22588.93
   ## 3  2009-12-31     610   420193 21062.10
   ```

```
## 4   2008-12-31     600    399013 19674.00
## 5   2007-12-31     597    408311 18715.00
## 6   2006-12-31     600    415293 24659.00
## 7   2005-12-31     588    420787 23417.50
## 8   2004-12-31     588    428221 21261.90
## 9   2003-12-31     581    432478 22876.80
## 10  2002-12-31     574    432027 22419.10
## 11  2001-12-31     573    429684 23012.00
```

3. Explore data

   The data is explored by using three functions - class(), str(), summary().
   The function class prints the vector of names of classes an object inherits
   from.

   ```
   # Class function
   class(cleanData)
   ```

   ```
   ## [1] "data.frame"
   ```

   The function str compactly displays the internal structure of an R object,
   a diagonistic function and an alternative to summary.

   ```
   # Str function
   str(cleanData)
   ```

   ```
   ## 'data.frame': 11 obs. of  4 variables:
   ##  $ year    : Date, format: "2011-12-31" "2010-12-31" ...
   ##  $ schools : num   620 614 610 600 597 600 588 588 581 574 ...
   ##  $ students: num   400383 402951 420193 399013 408311 ...
   ##  $ teachers: num   21847 22589 21062 19674 18715 ...
   ```

   And the summary is a generic function to produce result summaries of the
   results of various model functions. It is especially helpful for seeing basic
   descriptive statistics for all of the variables in a data frame and also the
   variables types.

   ```
   # Summary function
   summary(cleanData)
   ```

   ```
   ##       year                 schools         students          teachers
   ##  Min.   :2001-12-31   Min.   :573.0   Min.   :399013   Min.   :18715
   ##  1st Qu.:2004-07-01   1st Qu.:584.5   1st Qu.:405631   1st Qu.:21162
   ##  Median :2006-12-31   Median :597.0   Median :420193   Median :22419
   ##  Mean   :2006-12-31   Mean   :595.0   Mean   :417213   Mean   :21958
   ##  3rd Qu.:2009-07-01   3rd Qu.:605.0   3rd Qu.:428953   3rd Qu.:22944
   ##  Max.   :2011-12-31   Max.   :620.0   Max.   :432478   Max.   :24659
   ```

3 Result

The three variables- schools, students and teachers are plotted against year (X-axis) in a line graph from the data obtained using the ggplot2 package. The melt command is used to reshape the data from wide to long format. The year variable is not melted as it identify the dataset's subject.

```r
# Melting the data
library(reshape2)
moltenData <- melt(cleanData,id.vars="year")
# Plotting the grapg using ggplot2
library(ggplot2)
ggplot(moltenData, aes(as.Date(year,"%e %b %Y"),value))+
    geom_line(aes(color = variable))+
    geom_point()  + xlab("Year") + ylab("Total Number")+
    labs(title = "Total schools, total students and total
    teachers")+theme_bw()
```
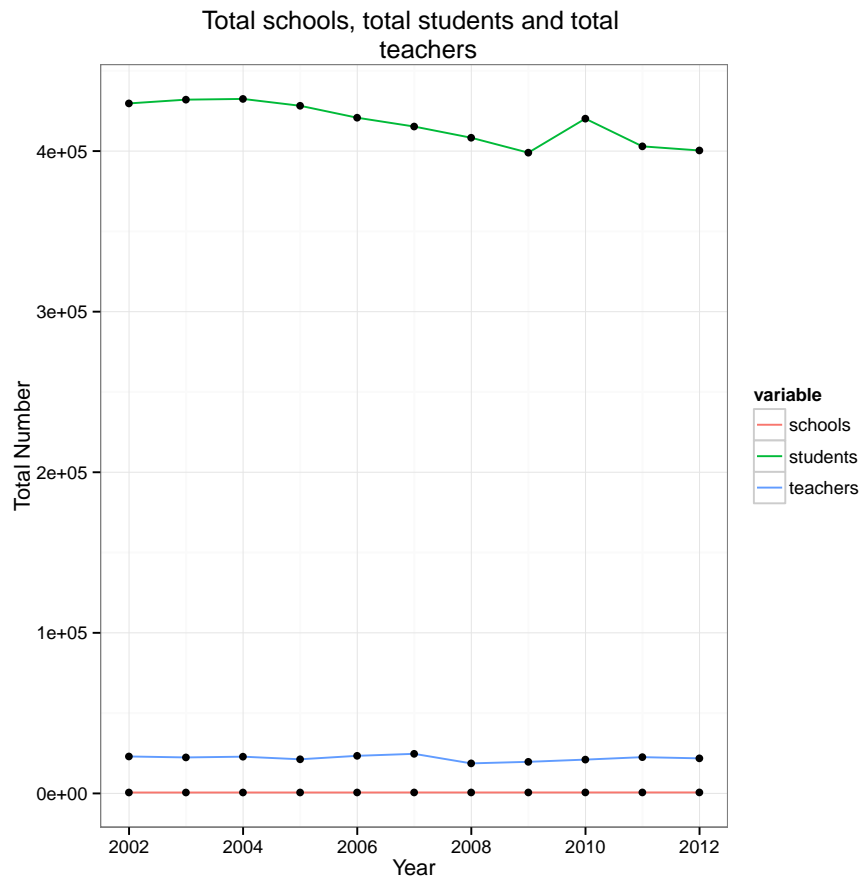
The graph has three different colored lines. The red,green and blue colored lines are of school, students and teachers variable, respectively. The total number of schools in 2001 was 573. It has some rise and fall in the total number. Finally, it became 620 in the year 2011.

The total number of students in 2001 was around 429,684, which decreased to around 399,013 in 2008. It again increased in 2009 to around 420,193 and finally decreased to 400,383 in 2011.

The total number of teachers in 2001 was 23,012, which reached its highest peak of around 24,659 in 2006. Later it decreased to around 18,715 in 2007. It again increased to around 22,589 in 2010 and finally fell to around 21,847 in 2011.