IT497 OSEMN Assignment

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

October 21, 2014

United States School Statistics (100 Largest Cities): Chicago, Illinois

1 Introduction

Schools in the United States comprise of both public as well as private schools. Public schools are available universally. 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. 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. Schools are divided into three groups- elementary school, middle or junior high school and high school. The American school year traditionally begins at the end of August, after a summer recess.

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

We have to collect data 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. We will get the required data from the Quandl API. We will be using various libraries available in R like RCurl, ggplot2 and reshape2.

1. Obtaing the data.

   There are many ways to download data into R. We will be using RCurl. It is a laborious but a good way to download data from a secure URL using getURL command in the RCurl package. We will also use read.csv and textConnection command, which are available in base R.

```
# Loading RCurl
library(knitr)
library(Quandl)

# To download data from a secure URL using RCurl

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

2. Cleaning data

   Cleaning data is the most time taking job. We will scrub and clean our data to get only the relevant data needed to obtain the results. Irrelevant data makes the analyis difficult. By scanning the data, we came to know that we only need column 1 through column 4. Therefore, we will choose and store all rows containing only these columns in new variable called cleanData. For further analysis, we will be using cleanData variable only. We will also change the colunms' name for convenience.

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

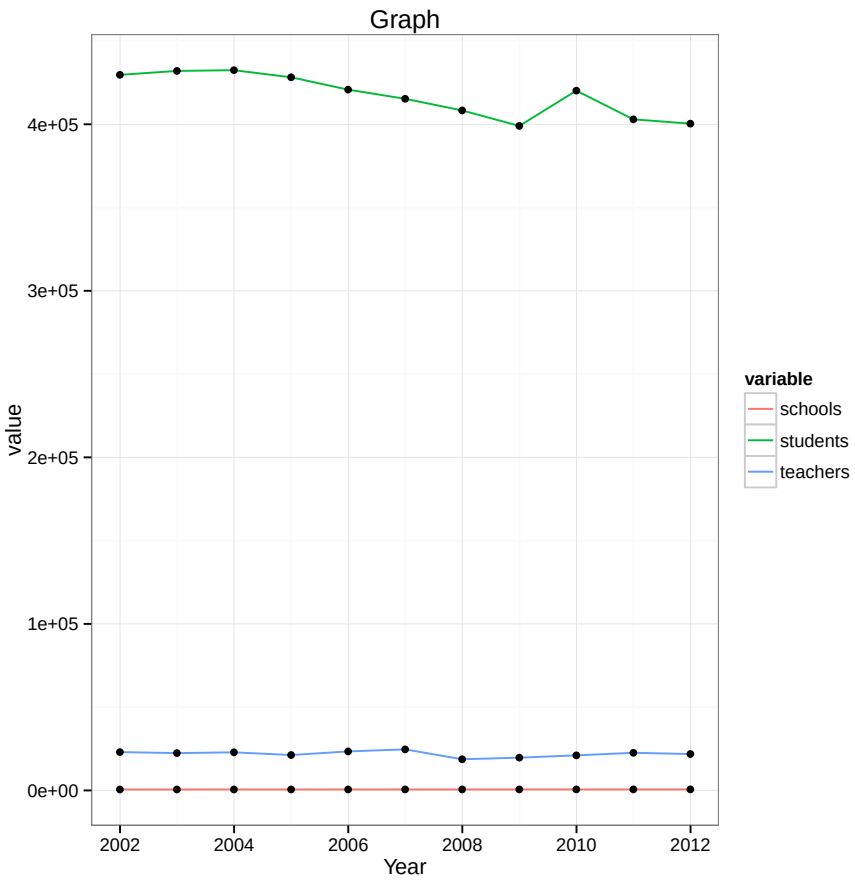
   We will explore the data by using three commands - class(), str(), summary(). The function class prints the vector of names of classes an object inherits from. The function str compactly displays the internal structure of an R object, a diagonistic function and an alternative to summary. And the summary is a generic function to produce result summaries of the results of various model functions.

```
# Class
class(cleanData)

## [1] "data.frame"

# Str
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 ...

# Summary
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

We will plot a line graph from the data using the ggplot2 package. For doing this, first we have to melt to data using the melt command in the reshape2 package.

```
# Melting the data
library(ggplot2)
library(reshape2)
moltenData <- melt(cleanData,id.vars="year")
ggplot(moltenData, aes(as.Date(year,"%e %b %Y"),value))+
  geom_line(aes(color = variable))+
   geom_point()  + xlab("Year") +
  labs(title = "Graph")+ theme_bw()
```



The graph is havong three lines showing