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 5th 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 could pursue higher education in colleges or universities.

The data are collected for schools in Chicago, IL from the database of the National Center for Education Statistics. The database stores the data 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.

2 Data

The data is collected for examining the total numbers of schools, teachers and students in Chicago, IL over the period of 10 years, from December, 2011 to December, 2011. The required data is retrieved from the Quandl API. Libraries like knitr, ggplot2 and reshape2 are used.

1. **Obtaining the data**

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

   ```r
   # Chunk1 shows the code for downloading the dataset from a
   # secure URL using Quandl API and Auth token.

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

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

2. **Scrubbing data (Cleaning data)**

   Scrubbing and cleaning of the data are done to get only the relevant data needed to obtain the results. Irrelevant data make the analysis difficult. Relevant data could be obtained by just looking into the dataset and selecting the rows and columns from it based on the requirement of the report. All the rows containing column 1 through column 4 are chosen and stored in a new object named cleanData.

   ```r
   # Chunk2 shows the code for extracting the relevant rows
   # and columns.
   # Cleaning the data needed to plot a graph showing the total
   # students, total teachers and total schools in Chicago,
   # Illinois
   cleanData<-myData[,1:4]
   ```

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

   ```r
   # Chunk3 shows the code for changing the columns' name and
   # displaying the data.
   # Changing the columns' name
   colnames(cleanData) <- c("year", "school", "student",
                            "teacher")
   ```

3. **Explore data**

The data are explored by using three functions - class(), str(), summary().

```
# Chunk4 has the code for printing the output of the class
# function.
class(cleanData)
```

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

The function class prints the vector of names of classes an object inherits from. The class of the dataset is data.frame.

```
# Chunk5 has the code for printing the output of the str
# function.
str(cleanData, width=60, strict.width="wrap")
```

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

The function str compactly displays the internal structure of an R object, a diagnostic function and an alternative to summary. Here, the year is in the date format. The school, student and teacher are in the number data type.

```
# Chunk6 has the code for printing the output of the summary
# function.
summary(cleanData)
```

```
##       year                 school         student           teacher
##  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
```

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. For year variable, minimum year is 2001 and maximum year is 2011. The school variable has the maximum of 573 schools and a maximum of 620 schools. The minimum and maximum total number of

3

students was 399013 and 432478, respectively The total number of teachers were 18715 and 24659 as minimum and maximum, respectively.

3 Result

The results of the report are shown by displaying the table and graph:

```
# Chunk 7 has the code for showing the dataset in table format.
# Showing the contents of the cleanData.
cleanData

##           year school student  teacher
## 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
```

Table1 has four columns namely- year, school, student and teacher. The year variable contains the year stating from 2001 through 2011. The school variable contains the total number of schools over 10 years. The student variable has total number of students studying over these years and the teacher variable contain the total number of teachers teaching in the schools from 2001 to 2011.

```
# Chunk8 has the code for melting the dataset and plotting
# the line graph.
# Melting the data
library(reshape2)
moltenData <- melt(cleanData,id.vars="year")

# Plotting the graph 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()
```

Total schools, total students and total teachers

variable
— school
— student
— teacher
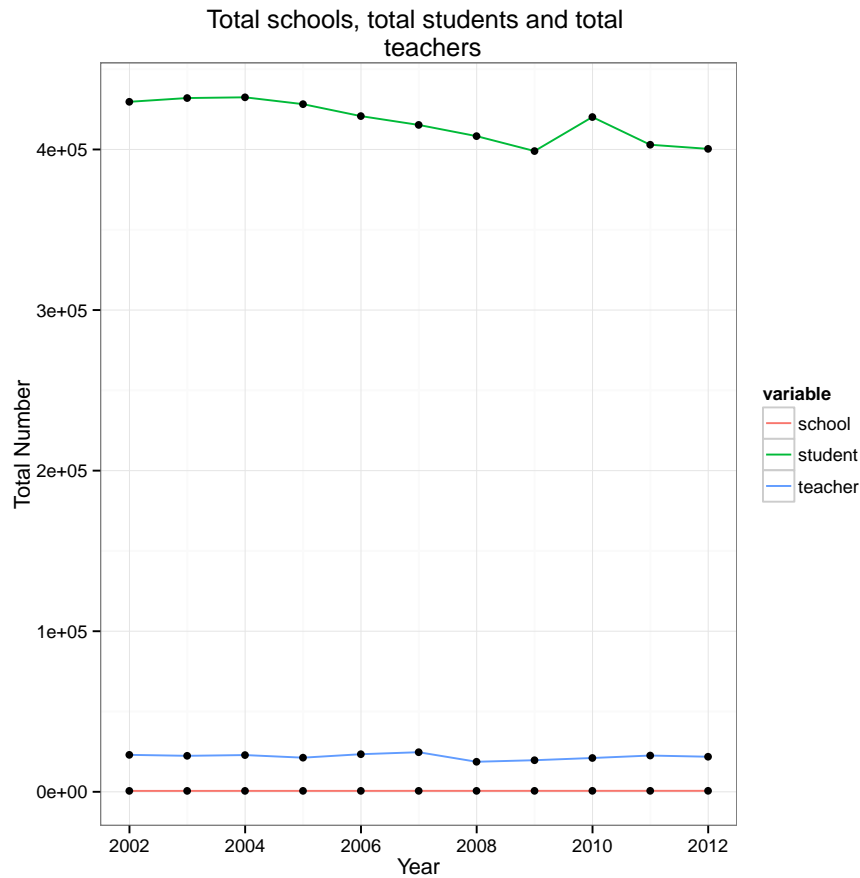
```
# Figure shows a line graph showing the total schools, total
# students and total teachers from 2001 to 2011.
```

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

The graph has three different colored lines. The red, green and blue colored lines are of the school, student and teacher variable, respectively. The X-axis shows the year from December,2001 through December, 2011. The Y-axis shows the total number or count.

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.