Lab 3. Linking R to the Web

GIS 3 - Geocomputation - Spring 2020 - Lily Cao

Contents

Load necessary libraries

```
suppressMessages(library(knitr))
suppressMessages(library(ggmap))
suppressMessages(library(tidyverse))
```

Load data

```
scores <- read.csv(url("https://raw.githubusercontent.com/lcao21/GIS3/master/Data/scores.csv"), header=
head(scores)
```

```
##   School.ID                                                    School.Name
## 1    02M260                             Clinton School Writers and Artists
## 2    06M211   Inwood Early College for Health and Information Technologies
## 3    01M539 New Explorations into Science, Technology and Math High School
## 4    02M294                                           Essex Street Academy
## 5    02M308                                   Lower Manhattan Arts Academy
## 6    02M545                High School for Dual Language and Asian Studies
##     Borough Building.Code       Street.Address      City State Zip.Code
## 1 Manhattan          M933 425 West 33rd Street Manhattan    NY    10001
## 2 Manhattan          M052   650 Academy Street Manhattan    NY    10002
## 3 Manhattan          M022  111 Columbia Street Manhattan    NY    10002
## 4 Manhattan          M445     350 Grand Street Manhattan    NY    10002
## 5 Manhattan          M445     350 Grand Street Manhattan    NY    10002
## 6 Manhattan          M445     350 Grand Street Manhattan    NY    10002
##   Latitude Longitude   Phone.Number Start.Time End.Time Student.Enrollment
## 1 40.75321 -73.99786   212-695-9114                                     NA
## 2 40.86605 -73.92486 718-935-3660      8:30 AM  3:00 PM                 87
## 3 40.71873 -73.97943   212-677-5190    8:15 AM  4:00 PM               1735
## 4 40.71687 -73.98953   212-475-4773    8:00 AM  2:45 PM                358
## 5 40.71687 -73.98953   212-505-0143    8:30 AM  3:00 PM                383
## 6 40.71687 -73.98953   212-475-4097    8:00 AM  3:35 PM                416
##   Percent.White Percent.Black Percent.Hispanic Percent.Asian
## 1
## 2          3.4%         21.8%            67.8%          4.6%
## 3         28.6%         13.3%            18.0%         38.5%
## 4         11.7%         38.5%            41.3%          5.9%
## 5          3.1%         28.2%            56.9%          8.6%
## 6          1.7%          3.1%             5.5%         88.9%
##   Average.Score..SAT.Math. Average.Score..SAT.Reading.
## 1                       NA                          NA
## 2                       NA                          NA
## 3                      657                         601
## 4                      395                         411
## 5                      418                         428
## 6                      613                         453
##   Average.Score..SAT.Writing. Percent.Tested
## 1                          NA
## 2                          NA
## 3                         601          91.0%
## 4                         387          78.9%
## 5                         415          65.1%
## 6                         463          95.9%
```

I'm using data on average SAT scores (Math, Reading, Writing) by accredited high schools in NYC that I found on Kaggle: https://www.kaggle.com/nycopendata/high-schools. The dataset includes columns for school name, borough, latitude/longitude coordinates, race breakdown, and average scores on each SAT section for the 2014-2015 school year. I read the csv in from my github repository link.

Ultimately, I want to map schools by their average SAT scores. First, though, I need to remove rows where values for the score columns are empty:

Clean data

```r
scores <- scores[!(is.na(scores$Average.Score..SAT.Math.) | is.na(scores$Average.Score..SAT.Reading.)
                   | is.na(scores$Average.Score..SAT.Writing.)), ]
head(scores)
```

```
##   School.ID                                                     School.Name
## 3    01M539 New Explorations into Science, Technology and Math High School
## 4    02M294                                           Essex Street Academy
## 5    02M308                                   Lower Manhattan Arts Academy
## 6    02M545                High School for Dual Language and Asian Studies
## 7    01M292                  Henry Street School for International Studies
## 8    01M696                                 Bard High School Early College
##     Borough Building.Code          Street.Address      City State Zip.Code
## 3 Manhattan          M022     111 Columbia Street Manhattan    NY    10002
## 4 Manhattan          M445        350 Grand Street Manhattan    NY    10002
## 5 Manhattan          M445        350 Grand Street Manhattan    NY    10002
## 6 Manhattan          M445        350 Grand Street Manhattan    NY    10002
## 7 Manhattan          M056        220 Henry Street Manhattan    NY    10002
## 8 Manhattan          M097 525 East Houston Street Manhattan    NY    10002
##   Latitude Longitude Phone.Number Start.Time End.Time Student.Enrollment
## 3 40.71873 -73.97943 212-677-5190    8:15 AM  4:00 PM               1735
## 4 40.71687 -73.98953 212-475-4773    8:00 AM  2:45 PM                358
## 5 40.71687 -73.98953 212-505-0143    8:30 AM  3:00 PM                383
## 6 40.71687 -73.98953 212-475-4097    8:00 AM  3:35 PM                416
## 7 40.71376 -73.98526 212-406-9411    8:30 AM  3:30 PM                255
## 8 40.71896 -73.97607 212-995-8479    9:00 AM  3:50 PM                545
##   Percent.White Percent.Black Percent.Hispanic Percent.Asian
## 3         28.6%         13.3%            18.0%         38.5%
## 4         11.7%         38.5%            41.3%          5.9%
## 5          3.1%         28.2%            56.9%          8.6%
## 6          1.7%          3.1%             5.5%         88.9%
## 7          3.9%         24.4%            56.6%         13.2%
## 8         45.3%         17.2%            18.7%         17.1%
##   Average.Score..SAT.Math. Average.Score..SAT.Reading.
## 3                      657                         601
## 4                      395                         411
## 5                      418                         428
## 6                      613                         453
## 7                      410                         406
## 8                      634                         641
##   Average.Score..SAT.Writing. Percent.Tested
## 3                         601          91.0%
## 4                         387          78.9%
## 5                         415          65.1%
## 6                         463          95.9%
## 7                         381          59.7%
## 8                         639          70.8%
```

After removing those rows, I'm left with 375 schools that have values for all three score columns. To get the total average SAT score (out of 2400), I combined these columns to create a new column called "avg_SAT_score":

Create a new column for average SAT score

```r
scores$avg_SAT_score <- (scores$Average.Score..SAT.Math. + scores$Average.Score..SAT.Reading.
                         + scores$Average.Score..SAT.Writing.)
head(scores)
```

```
##   School.ID                                                    School.Name
## 3    01M539 New Explorations into Science, Technology and Math High School
## 4    02M294                                           Essex Street Academy
## 5    02M308                                   Lower Manhattan Arts Academy
## 6    02M545                High School for Dual Language and Asian Studies
## 7    01M292                  Henry Street School for International Studies
## 8    01M696                                 Bard High School Early College
##     Borough Building.Code          Street.Address      City State Zip.Code
## 3 Manhattan          M022     111 Columbia Street Manhattan    NY    10002
## 4 Manhattan          M445        350 Grand Street Manhattan    NY    10002
## 5 Manhattan          M445        350 Grand Street Manhattan    NY    10002
## 6 Manhattan          M445        350 Grand Street Manhattan    NY    10002
## 7 Manhattan          M056        220 Henry Street Manhattan    NY    10002
## 8 Manhattan          M097 525 East Houston Street Manhattan    NY    10002
##   Latitude Longitude Phone.Number Start.Time End.Time Student.Enrollment
## 3 40.71873 -73.97943 212-677-5190    8:15 AM  4:00 PM               1735
## 4 40.71687 -73.98953 212-475-4773    8:00 AM  2:45 PM                358
## 5 40.71687 -73.98953 212-505-0143    8:30 AM  3:00 PM                383
## 6 40.71687 -73.98953 212-475-4097    8:00 AM  3:35 PM                416
## 7 40.71376 -73.98526 212-406-9411    8:30 AM  3:30 PM                255
## 8 40.71896 -73.97607 212-995-8479    9:00 AM  3:50 PM                545
##   Percent.White Percent.Black Percent.Hispanic Percent.Asian
## 3         28.6%         13.3%            18.0%         38.5%
## 4         11.7%         38.5%            41.3%          5.9%
## 5          3.1%         28.2%            56.9%          8.6%
## 6          1.7%          3.1%             5.5%         88.9%
## 7          3.9%         24.4%            56.6%         13.2%
## 8         45.3%         17.2%            18.7%         17.1%
##   Average.Score..SAT.Math. Average.Score..SAT.Reading.
## 3                      657                         601
## 4                      395                         411
## 5                      418                         428
## 6                      613                         453
## 7                      410                         406
## 8                      634                         641
##   Average.Score..SAT.Writing. Percent.Tested avg_SAT_score
## 3                         601          91.0%          1859
## 4                         387          78.9%          1193
## 5                         415          65.1%          1261
## 6                         463          95.9%          1529
## 7                         381          59.7%          1197
## 8                         639          70.8%          1914
```

The dataset provides information only on the total number of students enrolled for each school and the percentage of them that took the SAT. Thus, I created a new column called "num_tested" that tells us how many students were tested by multiplying the "Percent.Tested" and "Student.Enrollment" columns. Before multiplying, however, I had to convert the "Percent.Tested" column from percentages to decimals:

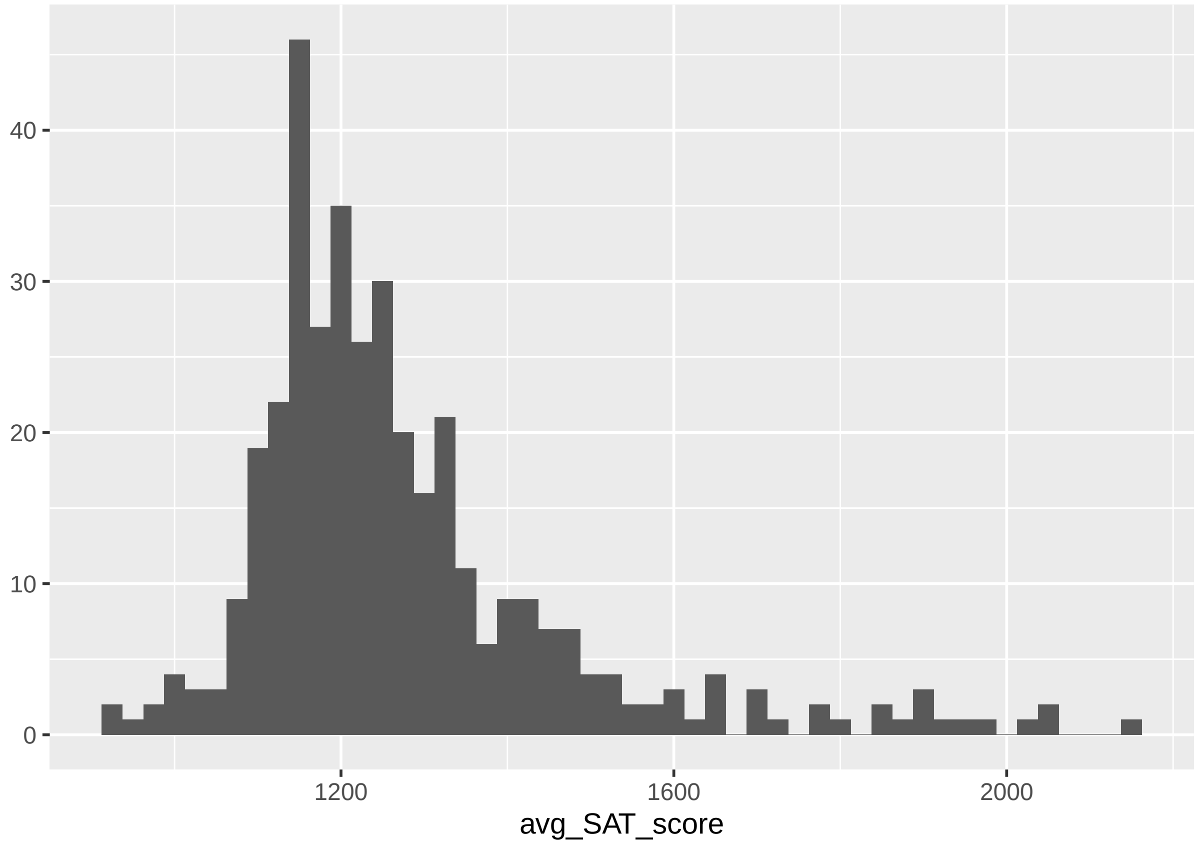
Create a new column for number of students tested

```r
scores$Percent.Tested <- as.numeric(sub("%", "",scores$Percent.Tested,fixed=TRUE))/100
scores$num_tested <- scores$Percent.Tested*scores$Student.Enrollment
head(scores)
```

```
##   School.ID                                                    School.Name
## 3    01M539 New Explorations into Science, Technology and Math High School
## 4    02M294                                           Essex Street Academy
## 5    02M308                                   Lower Manhattan Arts Academy
## 6    02M545                High School for Dual Language and Asian Studies
## 7    01M292                  Henry Street School for International Studies
## 8    01M696                                 Bard High School Early College
##     Borough Building.Code          Street.Address      City State Zip.Code
## 3 Manhattan          M022     111 Columbia Street Manhattan    NY    10002
## 4 Manhattan          M445        350 Grand Street Manhattan    NY    10002
## 5 Manhattan          M445        350 Grand Street Manhattan    NY    10002
## 6 Manhattan          M445        350 Grand Street Manhattan    NY    10002
## 7 Manhattan          M056        220 Henry Street Manhattan    NY    10002
## 8 Manhattan          M097 525 East Houston Street Manhattan    NY    10002
##   Latitude Longitude Phone.Number Start.Time End.Time Student.Enrollment
## 3 40.71873 -73.97943 212-677-5190    8:15 AM  4:00 PM               1735
## 4 40.71687 -73.98953 212-475-4773    8:00 AM  2:45 PM                358
## 5 40.71687 -73.98953 212-505-0143    8:30 AM  3:00 PM                383
## 6 40.71687 -73.98953 212-475-4097    8:00 AM  3:35 PM                416
## 7 40.71376 -73.98526 212-406-9411    8:30 AM  3:30 PM                255
## 8 40.71896 -73.97607 212-995-8479    9:00 AM  3:50 PM                545
##   Percent.White Percent.Black Percent.Hispanic Percent.Asian
## 3         28.6%         13.3%            18.0%         38.5%
## 4         11.7%         38.5%            41.3%          5.9%
## 5          3.1%         28.2%            56.9%          8.6%
## 6          1.7%          3.1%             5.5%         88.9%
## 7          3.9%         24.4%            56.6%         13.2%
## 8         45.3%         17.2%            18.7%         17.1%
##   Average.Score..SAT.Math. Average.Score..SAT.Reading.
## 3                      657                         601
## 4                      395                         411
## 5                      418                         428
## 6                      613                         453
## 7                      410                         406
## 8                      634                         641
##   Average.Score..SAT.Writing. Percent.Tested avg_SAT_score num_tested
## 3                         601          0.910          1859   1578.850
## 4                         387          0.789          1193    282.462
## 5                         415          0.651          1261    249.333
## 6                         463          0.959          1529    398.944
## 7                         381          0.597          1197    152.235
## 8                         639          0.708          1914    385.860
```

Some data exploration through graphs

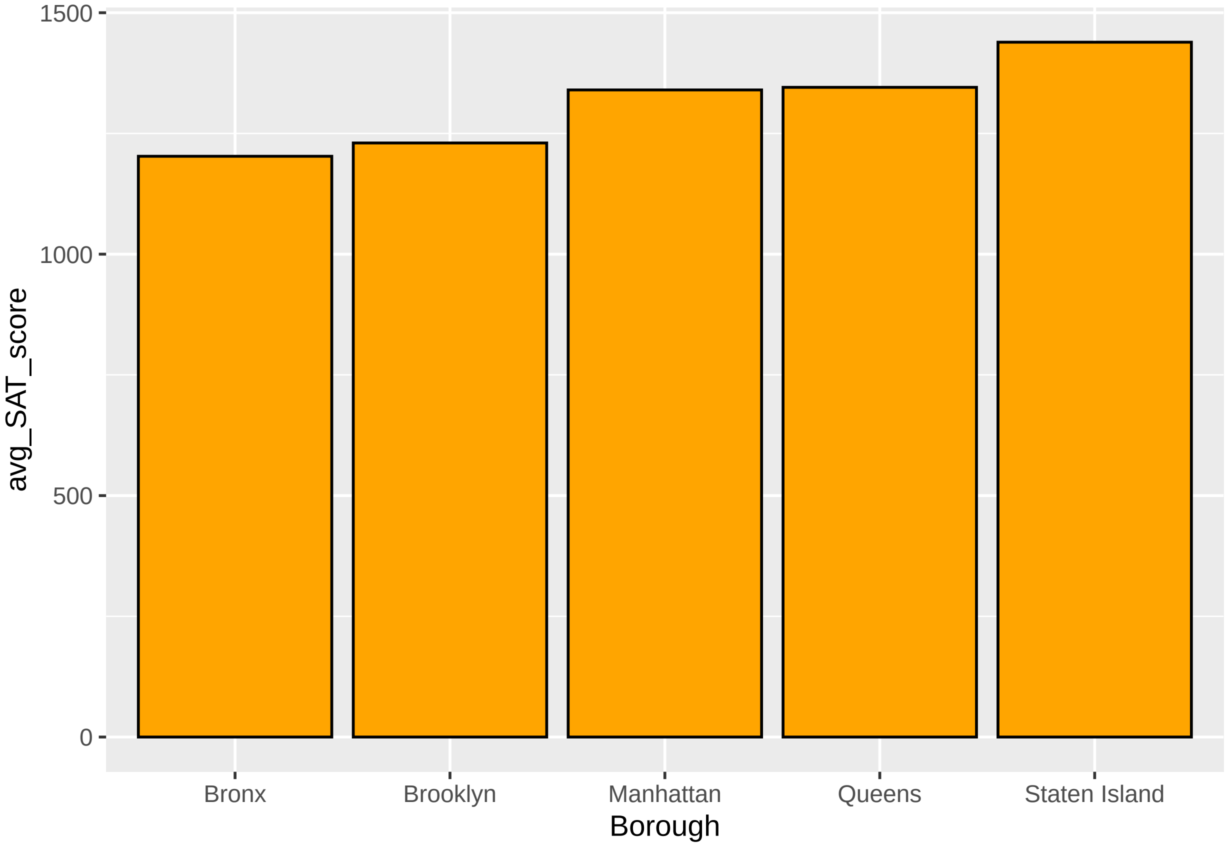
```
qplot(avg_SAT_score, data=scores, geom="histogram", binwidth=25)
```



```
scores %>% group_by(Borough) %>% summarise(avg_SAT_score = mean(avg_SAT_score))
```

```
## # A tibble: 5 x 2
##   Borough       avg_SAT_score
##   <fct>                 <dbl>
## 1 Bronx                 1203.
## 2 Brooklyn              1230.
## 3 Manhattan             1340.
## 4 Queens                1345.
## 5 Staten Island         1439
```

```
ggplot(scores, aes(x=Borough, y=avg_SAT_score)) +
  stat_summary(fun.y="mean", geom="bar", fill="orange", color="black")
```



Mapping the top 50 highest average SAT math scores by school

Get the top 50 schools by math score

```
top_scores <- scores %>% top_n(n = 50, wt = scores$Average.Score..SAT.Math.)
head(top_scores)
```

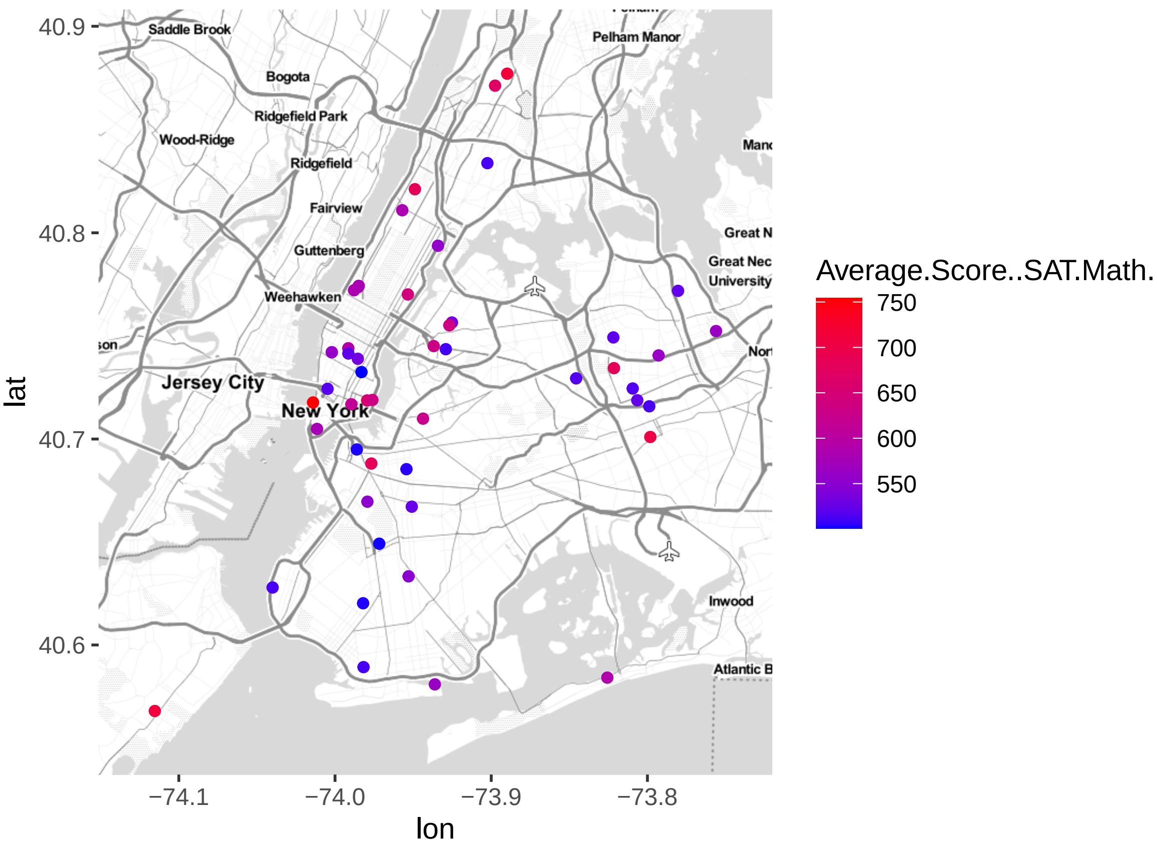
```
##   School.ID                                                    School.Name
## 1    01M539 New Explorations into Science, Technology and Math High School
## 2    02M545                High School for Dual Language and Asian Studies
## 3    01M696                                 Bard High School Early College
## 4    02M407                          Institute for Collaborative Education
## 5    02M418                                         Millennium High School
## 6    02M411                              Baruch College Campus High School
##     Borough Building.Code          Street.Address      City State Zip.Code
## 1 Manhattan          M022     111 Columbia Street Manhattan    NY    10002
## 2 Manhattan          M445        350 Grand Street Manhattan    NY    10002
## 3 Manhattan          M097 525 East Houston Street Manhattan    NY    10002
## 4 Manhattan          M475    345 East 15th Street Manhattan    NY    10003
## 5 Manhattan          M824         75 Broad Street Manhattan    NY    10004
## 6 Manhattan          M874     55 East 25th Street Manhattan    NY    10010
##   Latitude Longitude Phone.Number Start.Time End.Time Student.Enrollment
## 1 40.71873 -73.97943 212-677-5190    8:15 AM  4:00 PM               1735
## 2 40.71687 -73.98953 212-475-4097    8:00 AM  3:35 PM                416
## 3 40.71896 -73.97607 212-995-8479    9:00 AM  3:50 PM                 545
## 4 40.73249 -73.98305 212-475-7972    8:00 AM  3:00 PM                 482
## 5 40.70492 -74.01151 212-825-9008    8:30 AM  3:00 PM                 659
## 6 40.74405 -73.99148 212-683-7440    8:20 AM  2:50 PM                 451
##   Percent.White Percent.Black Percent.Hispanic Percent.Asian
## 1         28.6%         13.3%            18.0%         38.5%
## 2          1.7%          3.1%             5.5%         88.9%
## 3         45.3%         17.2%            18.7%         17.1%
## 4         56.5%         14.1%            14.9%          5.8%
## 5         32.8%          7.6%            18.2%         38.4%
## 6         22.8%          6.2%            14.9%         54.8%
##   Average.Score..SAT.Math. Average.Score..SAT.Reading.
## 1                      657                         601
## 2                      613                         453
## 3                      634                         641
## 4                      501                         550
## 5                      577                         560
## 6                      592                         526
##   Average.Score..SAT.Writing. Percent.Tested avg_SAT_score num_tested
## 1                         601          0.910          1859   1578.850
## 2                         463          0.959          1529    398.944
## 3                         639          0.708          1914    385.860
## 4                         541          0.786          1592    378.852
## 5                         567          0.940          1704    619.460
## 6                         531          0.943          1649    425.293
```

I used the get_stamenmap function from the ggmap library; it accesses a tile server for Stamen Maps (http://maps.stamen.com/) and downloads map tiles to format a map. I found it to be easier to use than geocode() and get_map() (from the same library and used in the lab) and more sophisticated. To plot the ggmap object, I used ggmap(). There are various map types (one of the function's argument), and I tested out 3 different kinds.

Create the map

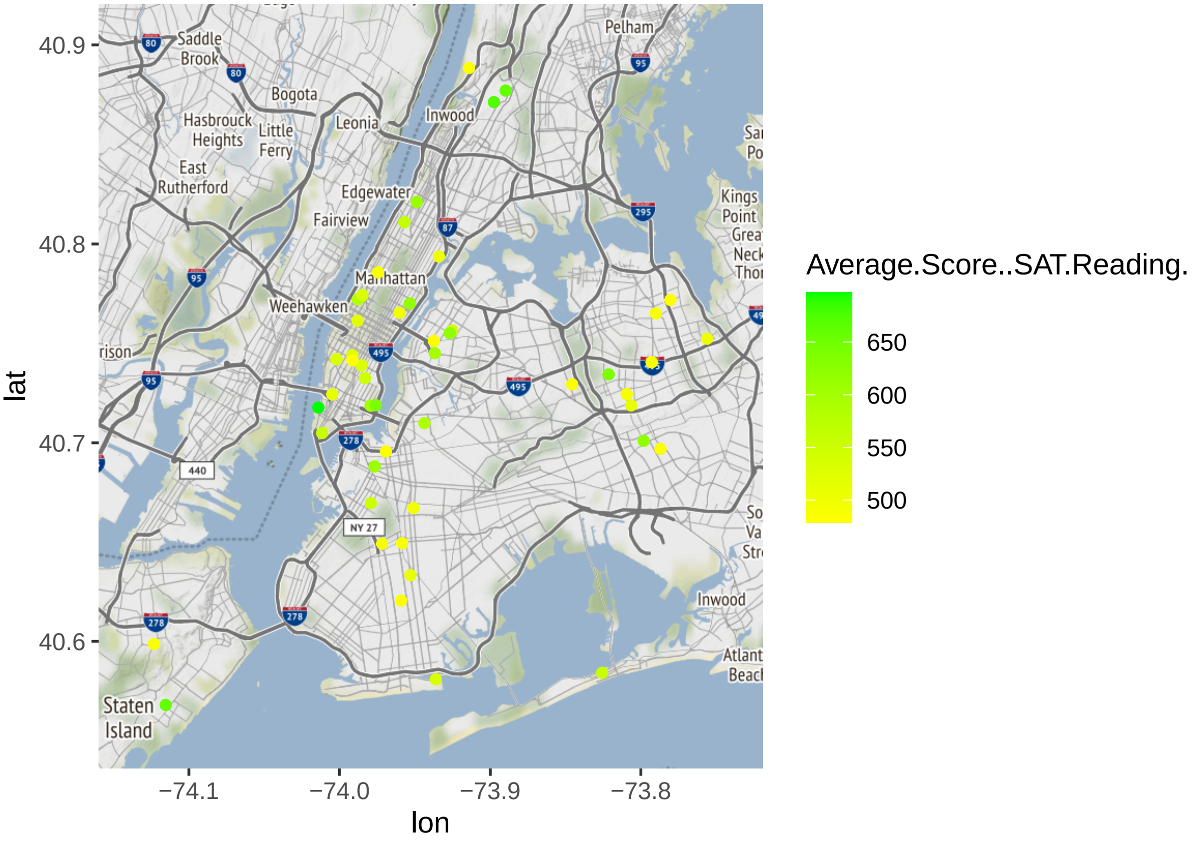
```r
top_height <- max(top_scores$Latitude) - min(top_scores$Latitude)
top_width <- max(top_scores$Longitude) - min(top_scores$Longitude)
top_borders <- c(bottom = min(top_scores$Latitude) - 0.1 * top_height,
                 top = max(top_scores$Latitude)  + 0.1 * top_height,
                 left = min(top_scores$Longitude) - 0.1 * top_width,
                 right = max(top_scores$Longitude) + 0.1 * top_width)

map <- get_stamenmap(top_borders, zoom = 11, maptype = "toner-lite")
ggmap(map) +
  geom_point(data = top_scores, mapping = aes(x = Longitude, y = Latitude,
  col = Average.Score..SAT.Math.)) +
  scale_color_gradient(low = "blue", high = "red")
```

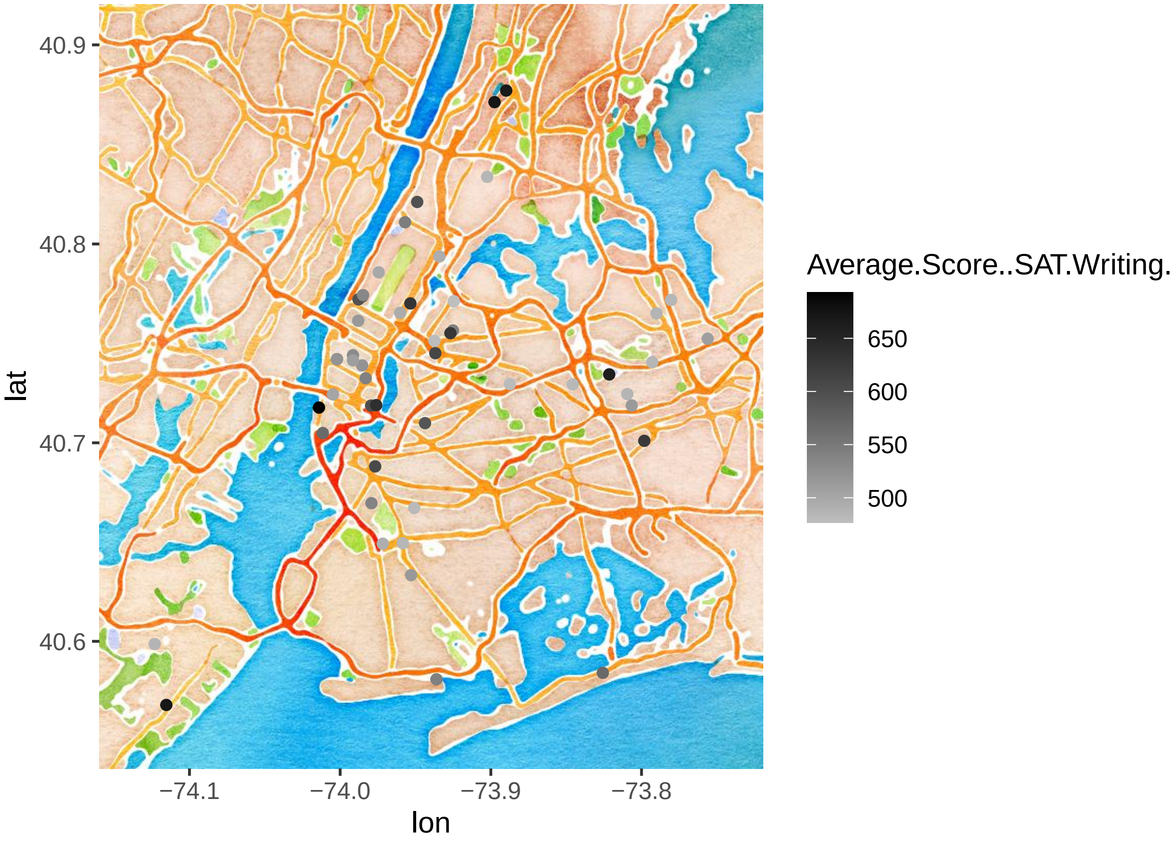


I repeated the same process for the average SAT reading and writing scores:

Mapping the top 50 highest average SAT reading scores by school



Mapping the top 50 highest average SAT writing scores by school

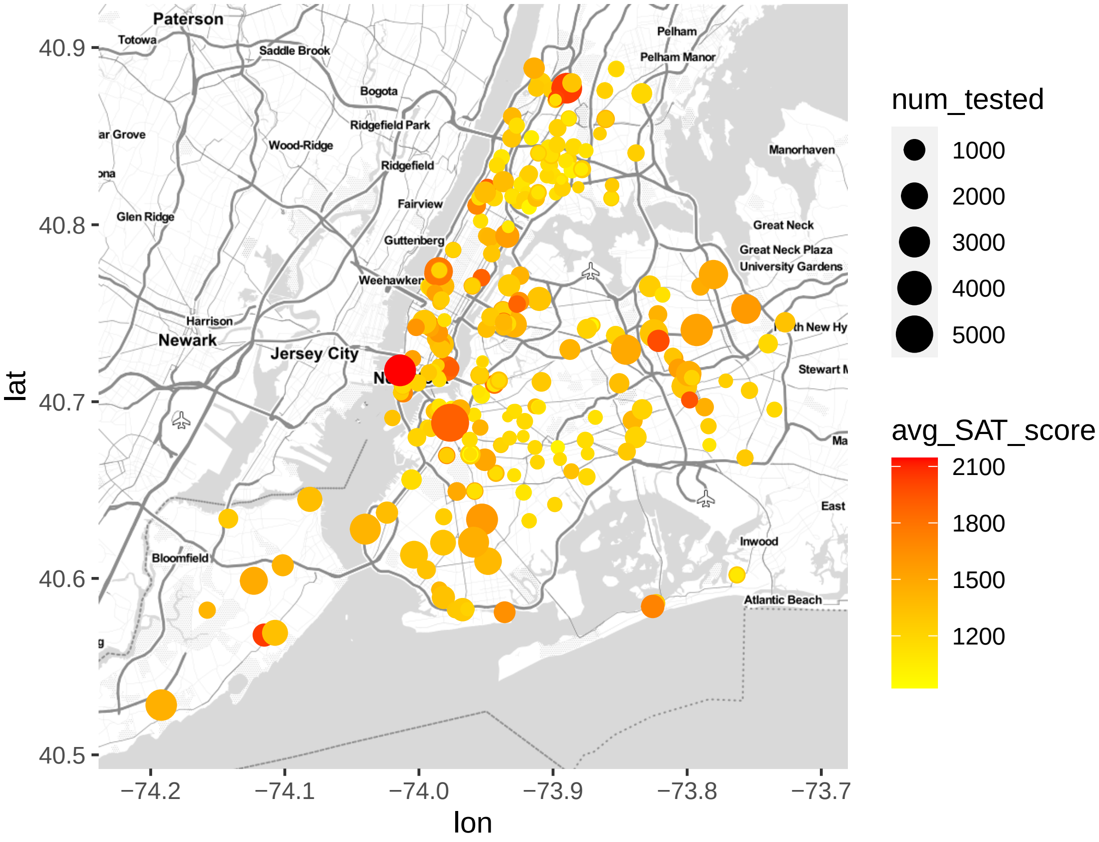


Mapping avg. SAT score by high school and number of students tested

```
height <- max(scores$Latitude) - min(scores$Latitude)
width <- max(scores$Longitude) - min(scores$Longitude)
borders <- c(bottom  = min(scores$Latitude)  - 0.1 * height,
             top     = max(scores$Latitude)  + 0.1 * height,
             left    = min(scores$Longitude) - 0.1 * width,
             right   = max(scores$Longitude) + 0.1 * width)

map_b <- get_stamenmap(borders, zoom = 11, maptype = "toner-lite")

ggmap(map_b) +
    geom_point(data = scores, mapping = aes(x = Longitude, y = Latitude,
                                            col = avg_SAT_score, size = num_tested)) +
    scale_color_gradient(low = "yellow", high = "red")
```

num_tested
1000
2000
3000
4000
5000
avg_SAT_score
2100
1800
1500
1200
lat
40.9
40.8
40.7
40.6
40.5
lon
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Paterson
Totowa
Saddle Brook
Bogota
Ridgefield Park
Wood-Ridge
Ridgefield
Glen Ridge
Fairview
Guttenberg
Weehawken
Harrison
Newark
Jersey City
Pelham
Pelham Manor
Manorhaven
Great Neck
Great Neck Plaza
University Gardens
Stewart M
Bloomfield
Inwood
East
Atlantic Beach