Exploring NYC High Schools and Survey Data

Shuangxu Li

December 19, 2019

In this project, I will clean and reorganize data from the New York City Department of Education (DOE) to explore factors that are related to students' SAT test performance, as well as to understand whether parent, teacher, and student perceptions of academic expectations, communication and other factors would affect the average school SAT performance.

Through this analysis, I want to get some general SAT test performance information of public schools in New York city, and want to know whether certain demographic factors (such as gender, income, race and so on) would bring any difference in test results across five boroughs in New York city. In addition, I wonder whether students, teachers, and parents have similar perceptions of NYC school quality and whose survey results are more closer to the actual results. All of these questions are explored through further data visualizations made by Tableau, and you can find them here.

The test performance data, collected in 2012, are public available and can be assessed here, and the survey data can be found here.

This is a guided project from DataQuest.

I'll start by loading the packages that I'll need for this analysis:

```
library(readr)
library(dplyr)
library(stringr)
library(purrr)
library(tidyr)
library(ggplot2)
```

Then I import these documents and will look into each of them one by one.

For the SAT results file, I need information about the average SAT scores for each school. But I find there are only columns for average writing, critical reading and math scores, and they are in character format.

```
glimpse(sat_results)
```

```
## Observations: 478
## Variables: 6
## $ DBN                            <chr> "01M292", "01M448", "01M450",
"01M4…
## $ `SCHOOL NAME`                  <chr> "HENRY STREET SCHOOL FOR
INTERNATIO…
```

```
## $ `Num of SAT Test Takers`          <chr> "29", "91", "70", "7", "44",
"112",…
## $ `SAT Critical Reading Avg. Score` <chr> "355", "383", "377", "414",
"390", …
## $ `SAT Math Avg. Score`             <chr> "404", "423", "402", "401",
"433", …
## $ `SAT Writing Avg. Score`          <chr> "363", "366", "370", "359",
"384", …
```

So I changed them into numeric format and then added them up to get total SAT score (named Avg_sat_score) for each high school.

```
sat_results <- sat_results %>%

    # change data format from character to numeric
    mutate(`Num of SAT Test Takers` =  as.numeric(`Num of SAT Test Takers`),
         `SAT Writing Avg. Score` = as.numeric(`SAT Writing Avg. Score`),
         `SAT Critical Reading Avg. Score` = as.numeric(`SAT Critical Reading
Avg. Score`),
         `SAT Math Avg. Score` = as.numeric(`SAT Math Avg. Score`)) %>%

     # create new column Avg_sat_score by adding other columns up
     mutate(Avg_sat_score = `SAT Writing Avg. Score` + `SAT Critical Reading
Avg. Score` + `SAT Math Avg. Score`)
```

DBN is a unique identifier for each high school in NYC. I could use DBN variable as a key factor to join other data file. In this Class Size data set, I did not find the DBN variable, but I could generate one by combining the CSD and School Code columns. I also find there are four different program types in it, but I only care about General Education program ('GEN ED'), so I would filter out those rows related to other programs. At last, I will group rows at school-level in order to connect with other files.

```
glimpse(class_size)
```

```
## Observations: 28,724
## Variables: 16
## $ CSD                                       <dbl> 1, 1, 1, 1, 1, 1, 1, 1,
1,…
## $ BORO                                      <chr> "M", "M", "M", "M",
"M", "…
## $ `SCHOOL CODE`                             <chr> "M015", "M015", "M015",
"M…
## $ `SCHOOL NAME`                             <chr> "P.S. 015 ROBERTO
CLEMENTE…
## $ GRADE                                     <chr> "0K", "0K", "01", "01",
"0…
## $ `PROGRAM TYPE`                            <chr> "GEN ED", "CTT", "GEN
ED",…
## $ `CORE SUBJECT \n(MS CORE and \n9-12 ONLY)` <chr> "-", "-", "-", "-", "-
", "…
```

```
## $ `CORE COURSE \n(MS CORE and 9-12 ONLY)`    <chr> "-", "-", "-", "-", "-
", "…
## $ `SERVICE CATEGORY\n(K-9* ONLY)`            <chr> "-", "-", "-", "-", "-
", "…
## $ `DATA SOURCE`                              <chr> "ATS", "ATS", "ATS",
"ATS"…
## $ `NUMBER OF STUDENTS / SEATS FILLED`        <dbl> 13, 17, 19, 16, 16, 17,
11…
## $ `NUMBER OF SECTIONS`                       <dbl> 1, 1, 1, 1, 1, 1, 1, 1,
1,…
## $ `AVERAGE CLASS SIZE`                       <dbl> 13.0, 17.0, 19.0, 16.0,
16…
## $ `SIZE OF SMALLEST CLASS`                   <dbl> 13, 17, 19, 16, 16, 17,
11…
## $ `SIZE OF LARGEST CLASS`                    <dbl> 13, 17, 19, 16, 16, 17,
11…
## $ `SCHOOLWIDE PUPIL-TEACHER RATIO`           <dbl> NA, NA, NA, NA, NA, NA,
NA…
```

```
class_size <- class_size %>%

    # create a new DBN column by combining CSD and Shool Code variables and
add '0' in its 1st position to match the DBN variable
    mutate(DBN = str_c(CSD, `SCHOOL CODE`, sep = "")) %>%
    mutate(DBN = str_pad(DBN, width = 6, side = 'left', pad = "0")) %>%

    # remove unnecessary rows
    filter(GRADE == "09-12", `PROGRAM TYPE` == "GEN ED") %>%

    # group by DBN to obtain school-level information
    group_by(DBN) %>%
    summarize(Avg_class_size = mean(`AVERAGE CLASS SIZE`),
              Avg_largest_class = mean(`SIZE OF LARGEST CLASS`),
              Avg_smallest_class = mean(`SIZE OF SMALLEST CLASS`))
```

For the Demographics data set, I select students who are graduated at year 2011-12, then filter out those unnecessary variables. Many demographic information shown in percentage are characters, I will change them into numbers for further analysis.

```
glimpse(demographics)
```

```
## Observations: 8,818
## Variables: 38
## $ DBN                            <chr> "01M015", "01M015", "01M015",
"01M015"…
## $ `School Name`                  <chr> "P.S. 015 Roberto Clemente", "P.S.
015…
## $ Year                           <chr> "2011-12", "2012-13", "2013-14",
"2014…
## $ `Total Enrollment`             <dbl> 189, 177, 190, 183, 176, 328, 302,
```

```
285…
## $ `Grade PK`                    <dbl> 13, 18, 26, 18, 14, 32, 31, 36, 30,
21…
## $ `Grade K`                     <dbl> 31, 38, 39, 27, 32, 46, 39, 39, 44,
47…
## $ `Grade 1`                     <dbl> 35, 26, 39, 47, 33, 52, 46, 38, 40,
43…
## $ `Grade 2`                     <dbl> 28, 22, 21, 31, 39, 54, 46, 36, 39,
41…
## $ `Grade 3`                     <dbl> 25, 26, 16, 19, 23, 52, 48, 45, 35,
43…
## $ `Grade 4`                     <dbl> 28, 23, 26, 17, 17, 46, 50, 47, 40,
35…
## $ `Grade 5`                     <dbl> 29, 24, 23, 24, 18, 46, 42, 44, 42,
40…
## $ `Grade 6`                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,…
## $ `Grade 7`                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,…
## $ `Grade 8`                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,…
## $ `Grade 9`                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,…
## $ `Grade 10`                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,…
## $ `Grade 11`                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,…
## $ `Grade 12`                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,…
## $ `# Female`                    <dbl> 92, 91, 93, 84, 83, 181, 154, 141,
132…
## $ `% Female`                    <chr> "48.7%", "51.4%", "48.9%", "45.9%",
"4…
## $ `# Male`                      <dbl> 97, 86, 97, 99, 93, 147, 148, 144,
138…
## $ `% Male`                      <chr> "51.3%", "48.6%", "51.1%", "54.1%",
"5…
## $ `# Asian`                     <dbl> 12, 15, 9, 8, 9, 51, 44, 41, 30,
27, 1…
## $ `% Asian`                     <chr> "6.3%", "8.5%", "4.7%", "4.4%",
"5.1%"…
## $ `# Black`                     <dbl> 63, 63, 72, 65, 57, 81, 69, 56, 47,
55…
## $ `% Black`                     <chr> "33.3%", "35.6%", "37.9%", "35.5%",
"3…
## $ `# Hispanic`                  <dbl> 109, 93, 104, 107, 105, 158, 150,
148,…
## $ `% Hispanic`                  <chr> "57.7%", "52.5%", "54.7%", "58.5%",
"5…
## $ `# Other`                     <dbl> 1, 3, 2, 1, 3, 10, 6, 10, 8, 3, 8,
```

```
13,…
## $ `% Other`                       <chr> "0.5%", "1.7%", "1.1%", "0.5%",
"1.7%"…
## $ `# White`                       <dbl> 4, 3, 3, 2, 2, 28, 33, 30, 27, 16,
16,…
## $ `% White`                       <chr> "2.1%", "1.7%", "1.6%", "1.1%",
"1.1%"…
## $ `# Students with Disabilities` <dbl> 52, 55, 65, 64, 57, 66, 79, 89, 82,
80…
## $ `% Students with Disabilities` <chr> "27.5%", "31.1%", "34.2%", "35.0%",
"3…
## $ `# English Language Learners`  <dbl> 22, 21, 19, 17, 16, 33, 26, 25, 18,
13…
## $ `% English Language Learners`  <chr> "11.6%", "11.9%", "10.0%", "9.3%",
"9.…
## $ `# Poverty`                     <dbl> 189, 177, 190, 183, 176, 328, 228,
213…
## $ `% Poverty`                     <chr> "100.0%", "100.0%", "100.0%",
"100.0%"…
```

```
demographics <- demographics %>%

    # only include students who are graduated at 2011-12
    filter(Year == "2011-12" & "Grade 9" != "NA") %>%

    # select columns that are useful or I am interested in
    select(DBN, `School Name`,`Total Enrollment`, `% Poverty`, `% English
Language Learners`, `% Students with Disabilities`,
         `% Asian`, `% Black`, `% Hispanic`, `% White`,
         `% Male`, `% Female`) %>%

    # change these percentage variables from character into numeric format
    mutate(`% Poverty` = parse_number(`% Poverty`)/100,
         `% English Language Learners` = parse_number(`% English Language
Learners`)/100,
         `% Students with Disabilities` = parse_number(`% Students with
Disabilities`)/100,
         `% Asian` = parse_number(`% Asian`)/100,
         `% Black` = parse_number(`% Black`)/100,
         `% Hispanic` = parse_number(`% Hispanic`)/100,
         `% White` = parse_number(`% White`)/100,
         `% Male` = parse_number(`% Male`)/100,
         `% Female` = parse_number(`% Female`)/100)
```

What I need in the Graduation file is the graduation rate and dropped-out rate. I will select 4 year duration and cohort year starting at 2008, since this corresponds to the graduation year 2011-12.

```
glimpse(graduation)
```

```
## Observations: 21,647
## Variables: 24
## $ DBN                          <chr> "01M292", "01M292", "01M292",
"01M292", "…
## $ `School Name`                <chr> "ORCHARD COLLEGIATE ACADEMY", "ORCHARD
CO…
## $ Category                     <chr> "All Students", "All Students", "All
Stud…
## $ `Cohort Year`                <dbl> 2013, 2012, 2011, 2010, 2009, 2008,
2007,…
## $ Cohort                       <chr> "4 year August", "4 year August", "4
year…
## $ `Cohort #`                   <dbl> 36, 44, 73, 61, 85, 70, 77, 78, 64,
36, 4…
## $ `Toal Grads #`               <chr> "25", "24", "46", "26", "49", "36",
"45",…
## $ `% of cohort`                <chr> "69.4", "54.5", "63.0", "42.6",
"57.6", "…
## $ `Total Regents #`            <chr> "23", "20", "41", "26", "44", "30",
"29",…
## $ `% of cohort 1`              <chr> "63.9", "45.5", "56.2", "42.6",
"51.8", "…
## $ `% of grads`                 <chr> "92.0", "83.3", "89.1", "100.0",
"89.8", …
## $ `Advanced Regents #`         <chr> "0", "1", "0", "1", "0", "0", "0",
"0", "…
## $ `% of cohort 2`              <chr> "0.0", "2.3", "0.0", "1.6", "0.0",
"0.0",…
## $ `% of grads 1`               <chr> "0.0", "4.2", "0.0", "3.8", "0.0",
"0.0",…
## $ `Regents without Advanced#`  <chr> "23", "19", "41", "25", "44", "30",
"29",…
## $ `% of cohort 3`              <chr> "63.9", "43.2", "56.2", "41.0",
"51.8", "…
## $ `% of grads 2`               <chr> "92.0", "79.2", "89.1", "96.2",
"89.8", "…
## $ `Local #`                    <chr> "2", "4", "5", "0", "5", "6", "16",
"7", …
## $ `% of cohort 4`              <chr> "5.6", "9.1", "6.8", "0.0", "5.9",
"8.6",…
## $ `% of grads 3`               <chr> "8.0", "16.7", "10.9", "0.0", "10.2",
"16…
## $ `Still Enrolled #`           <chr> "3", "10", "18", "18", "28", "18",
"22", …
## $ `% of cohort 5`              <chr> "8.3", "22.7", "24.7", "29.5", "32.9",
"2…
## $ `Dropout #`                  <chr> "7", "10", "7", "17", "8", "13", "5",
"11…
## $ `% of cohort 6`              <chr> "19.4", "22.7", "9.6", "27.9", "9.4",
"18…
```

```
graduation <- graduation %>%

    # select cohort year starting at 2008
    filter(`Cohort Year` == "2008" & Cohort %in% c("4 year August","4 year
June")) %>%

    # unselect hose unnecessary rows
    select(DBN, `School Name`,Cohort, `Toal Grads #`,`% of cohort`, `% of
cohort 6`) %>%

    # change them into inutive names
    rename('Total_Grads_%'  = `% of cohort`, 'Dropped_Out_%' =`% of cohort
6`) %>%

    # change data type
    mutate(`Toal Grads #` = as.numeric(`Toal Grads #`),
         `Total_Grads_%` = as.numeric(`Total_Grads_%`),
         `Dropped_Out_%` = as.numeric(`Dropped_Out_%`)) %>%

    # obtain school-level information
    group_by(DBN) %>%
    summarize(`Toal Grads #` = sum(`Toal Grads #`),
              `Total_Grads_%` = mean(`Total_Grads_%`),
              `Dropped_Out_%` = mean(`Dropped_Out_%`)) %>%

    # get a percentage format
    mutate(`Total_Grads_%` = `Total_Grads_%`/100,
            `Dropped_Out_%` = `Dropped_Out_%`/100)
```

For the High School Directory file, I am interested in the Boroughs where high schools are located at and the specific coordinates information in location column.

```
glimpse(hs_directory)
```

```
## Observations: 435
## Variables: 64
## $ dbn                                <chr> "21K540", "15K429", "24Q530",
"05M36…
## $ school_name                        <chr> "John Dewey High School",
"Brooklyn …
## $ borough                            <chr> "Brooklyn", "Brooklyn", "Queens",
"M…
## $ building_code                      <chr> "K540", "K293", "Q520", "M043",
"Q46…
## $ phone_number                       <chr> "718-373-6400", "718-694-9741",
"718…
## $ fax_number                         <chr> "718-266-4385", "718-694-9745",
"718…
## $ grade_span_min                     <dbl> 9, 6, 9, 9, 9, 9, 9, 9, 9, 9, 9,
9, …
```

```
## $ grade_span_max                  <dbl> 12, 12, 12, 12, 12, 12, 12, 12,
12, …
## $ expgrade_span_min               <dbl> NA, NA, NA, NA, NA, NA, NA, NA,
NA, …
## $ expgrade_span_max               <dbl> NA, NA, NA, NA, NA, NA, NA, NA,
NA, …
## $ bus                             <chr> "B1, B4, B64, B82", "B103, B45,
B57,…
## $ subway                          <chr> "D to Bay 50th St ; F to Ave X ;
N t…
## $ primary_address_line_1          <chr> "50 Avenue X", "284 Baltic
Street", …
## $ city                            <chr> "Brooklyn", "Brooklyn", "Long
Island…
## $ state_code                      <chr> "NY", "NY", "NY", "NY", "NY",
"NY", …
## $ postcode                        <dbl> 11223, 11201, 11101, 10027,
11691, 1…
## $ website                         <chr> "http://johndeweyhighschool.org",
"w…
## $ total_students                  <dbl> 1937, 275, 503, 309, 412, 260,
155, …
## $ campus_name                     <chr> "N/A", "N/A", "Middle College
Campus…
## $ school_type                     <chr> NA, "Consortium School",
"Consortium…
## $ overview_paragraph              <chr> "We offer an innovative form of
educ…
## $ program_highlights              <chr> "Computer Science Institute,
Medical…
## $ language_classes                <chr> "Chinese, French, Italian,
Russian, …
## $ advancedplacement_courses       <chr> "Art History, Biology, Calculus
AB, …
## $ online_ap_courses               <chr> NA, "Art History, Calculus AB",
NA, …
## $ online_language_courses         <chr> NA, "Arabic, Bengali, Chinese,
Chine…
## $ extracurricular_activities      <chr> "Anime, Asian-American, ASPIRA of
Ne…
## $ psal_sports_boys                <chr> "Basketball, Cross Country,
Football…
## $ psal_sports_girls               <chr> "Basketball, Cross Country,
Football…
## $ psal_sports_coed                <chr> NA, NA, NA, NA, NA, "Cricket",
NA, N…
## $ school_sports                   <chr> "We also offer a variety of
after-sc…
## $ partner_cbo                     <chr> "Jewish Board of Family and
Children…
```

```
## $ partner_hospital                  <chr> "Coney Island Hospital Center,
JASA …
## $ partner_highered                  <chr> "Kingsborough Community College,
Med…
## $ partner_cultural                  <chr> "Theatre Development Fund (TDF),
Cen…
## $ partner_nonprofit                 <chr> "National Academy Foundation
(NAF), …
## $ partner_corporate                 <chr> NA, NA, "Shearman & Sterling",
NA, "…
## $ partner_financial                 <chr> "Citigroup, Ernst & Young ,
Federal …
## $ partner_other                     <chr> "National Aeronautics Space
Administ…
## $ addtl_info1                       <chr> "Community Service Requirement",
"Sa…
## $ addtl_info2                       <chr> NA, "Extended Day Program,
Uniform R…
## $ start_time                        <time> 08:13:00, 08:45:00, 08:00:00,
08:35…
## $ end_time                          <time> 15:05:00, 15:10:00, 15:30:00,
15:45…
## $ se_services                       <chr> "This school will provide
students w…
## $ ell_programs                      <chr> "ESL; Transitional Bilingual
Program…
## $ school_accessibility_description <chr> "Not Functionally Accessible",
"Not …
## $ number_programs                   <dbl> 8, 1, 1, 1, 1, 1, 1, 1, 6, 1, 1,
1, …
## $ priority01                        <chr> "Priority to Brooklyn students or
re…
## $ priority02                        <chr> "Then to New York City
residents", "…
## $ priority03                        <chr> "For K56B only: Open only to
student…
## $ priority04                        <chr> NA, "Then to New York City
residents…
## $ priority05                        <chr> NA, NA, NA, NA, NA, NA, "Then to
New…
## $ priority06                        <chr> NA, NA, NA, NA, NA, NA, NA, NA,
NA, …
## $ priority07                        <chr> NA, NA, NA, NA, NA, NA, NA, NA,
NA, …
## $ priority08                        <chr> NA, NA, NA, NA, NA, NA, NA, NA,
NA, …
## $ priority09                        <chr> NA, NA, NA, NA, NA, NA, NA, NA,
NA, …
## $ priority10                        <chr> NA, NA, NA, NA, NA, NA, NA, NA,
NA, …
```

```
## $ `Location 1`                        <chr> "50 Avenue\nX Brooklyn, NY
11223\n(4…
## $ `Community Board`                   <dbl> 13, 6, 2, 9, 14, 13, 3, 9, 7, 3,
11,…
## $ `Council District`                  <dbl> 47, 33, 26, 7, 31, 47, 36, 18, 6,
1,…
## $ `Census Tract`                      <dbl> 308, 69, 179, 219, 100802, 306,
291,…
## $ BIN                                 <dbl> 3194998, 3006401, 4003442,
1059723, …
## $ BBL                                 <dbl> 3071850020, 3004020001,
4002490001, …
## $ NTA                                 <chr> "Gravesend", "DUMBO-Vinegar Hill-
Dow…
```

```
hs_directory <- hs_directory %>%

    rename(DBN = dbn) %>%

    # select variables I will need
    select(DBN, school_name, borough, `Location 1`) %>%

    # split Location text by '\n\' and choose the last part
    mutate(lat_long = str_split(`Location 1`, "\n", simplify = TRUE)[,3]) %>%

    # split it further by ',' and assign them into latitude and longitude
columns
    mutate(lat = str_split(lat_long, ",", simplify = TRUE)[,1],
           long = str_split(lat_long, ",", simplify = TRUE)[,2]) %>%

    # get rid of the parenthesis
    mutate(Latitude = str_sub(lat,2,-1), Longitude = str_sub(long, 1,-2)) %>%

    # change data type
    mutate_at(vars(Latitude, Longitude), as.numeric) %>%

    select(DBN, school_name, borough, Latitude, Longitude)
```

For the Survey file, there are two files, one is survey results about general education, the other one is about District 75 program, a program that provides highly specialized instructional support for students with significant challenges. I will retain rows related to High School and relevant survey scores. Since they have the same variables, I then combine both files by stacking one file on another by rows.

```
survey_gened <- survey_gened %>%
  filter(schooltype == 'High School') %>%
  select(dbn,schoolname,saf_p_11:aca_tot_11)

survey_d75 <- survey_d75 %>%
```

```
  select(dbn,schoolname,saf_p_11:aca_tot_11)

# combine `survey` and `survey_d75` data frames
survey <- survey_gened %>%
  bind_rows(survey_d75) %>%
  rename(DBN = dbn)
```

After doing some cleaning work on each data file, I want to check whether there are any duplicated DBNs in each of them. I create a list to include all of 6 data set, then apply the check function to them at once. Luckily, no duplicated DBN is found :)

```
# Create a list of the six data frames named ny_schools.
ny_schools <- list(sat_results, class_size, demographics, graduation,
hs_directory, survey)
names(ny_schools) <- c("sat_results", "class_size", "demographics",
"graduation", "hs_directory", "survey")

## Return a list of rows from each data frame that contain duplicate values
of DBN.
duplicate_DBN <- ny_schools %>%
  map(mutate, is_dup = duplicated(DBN))  %>%
  map(filter, is_dup == "TRUE")
```

Since SAT test results are the dependent variable in my analysis, so I use sat_results as the base file to left join other ones by the key factor DBN. The new file named combined_db.

```
combined_db <- sat_results %>%
  left_join(class_size, by = "DBN") %>%
  left_join(demographics, by = "DBN") %>%
  left_join(graduation, by = "DBN") %>%
  left_join(hs_directory, by = "DBN") %>%
  left_join(survey, by = "DBN")
```

The combined file include many columns related to school names and each of them contains certain null values. I check each of them and choose the one with least null values.

```
# check the null values in the combined data set
colSums(is.na(combined_db))
```

```
##                             DBN                     SCHOOL NAME
##                               0                               0
##          Num of SAT Test Takers SAT Critical Reading Avg. Score
##                              57                              57
##             SAT Math Avg. Score          SAT Writing Avg. Score
##                              57                              57
##                  Avg_sat_score                 Avg_class_size
##                              57                              48
##               Avg_largest_class             Avg_smallest_class
##                              48                              48
##                     School Name               Total Enrollment
##                              49                              49
```

```
##                         % Poverty     % English Language Learners
##                                49                              49
##    % Students with Disabilities                          % Asian
##                                49                              49
##                           % Black                     % Hispanic
##                                49                              49
##                           % White                          % Male
##                                49                              49
##                         % Female                   Toal Grads #
##                                49                              60
##                   Total_Grads_%                   Dropped_Out_%
##                                60                              60
##                     school_name                         borough
##                                109                             109
##                        Latitude                       Longitude
##                                109                             109
##                      schoolname                        saf_p_11
##                                104                             104
##                        com_p_11                        eng_p_11
##                                104                             104
##                        aca_p_11                        saf_t_11
##                                104                             104
##                        com_t_11                        eng_t_11
##                                104                             104
##                        aca_t_11                        saf_s_11
##                                104                             106
##                        com_s_11                        eng_s_11
##                                106                             106
##                        aca_s_11                     saf_tot_11
##                                106                             104
##                      com_tot_11                     eng_tot_11
##                                104                             104
##                      aca_tot_11
##                                104
```

```
# unselect school name columns with more null values
combined_db <- combined_db %>%
  select(- 'School Name',-school_name, -schoolname) %>%
  rename(School_name = `SCHOOL NAME`, Borough = borough) %>%
  mutate(School_name = str_to_title(School_name, locale = "en"))
```

By calculating the correlations between Avg_sat_score with all other columns, I will know which factor may has high influence on the test results. I then filter those variables with high correlations values more than 0.5 and less than -0.5.

```
combined_db %>%
  select_if(is.numeric) %>%

  # create a correlation matrix for all numeric variables
  cor(use = "pairwise.complete.obs") %>%
```

```
  # change the matrix into tibble format
  as_tibble(rownames = "variable") %>%

  # select correlations between Avg_sat_score and other variables
  select(variable, Avg_sat_score) %>% #

  # filter the ones with high correlations
  filter(Avg_sat_score > 0.5 | Avg_sat_score < -0.5) %>%

  # oder the correlation values from high to low
  arrange(desc(Avg_sat_score))
```

```
## # A tibble: 8 x 2
##   variable                        Avg_sat_score
##   <chr>                                   <dbl>
## 1 Avg_sat_score                           1
## 2 SAT Writing Avg. Score                  0.981
## 3 SAT Critical Reading Avg. Score         0.975
## 4 SAT Math Avg. Score                     0.953
## 5 % White                                 0.648
## 6 Total_Grads_%                           0.547
## 7 % Asian                                 0.544
## 8 % Poverty                              -0.682
```

For visualizations, I need to reshape certain columns (social, racial and sex factors) with percentage into two columns, one for factor and the other one for corresponding percentage. As regards survey data, the response score to each question represent separate column, so I also need to pivot these data. After that, I extract information from the names (e.g. saf_p_11) to identity the response question category (Academic Expectations, communication, Engagement or Safety and Respect) and response group (parent, student and teacher) for further analysis.

```
# for social factors such as income, language learning and physical
conditions
combined_db <- combined_db %>%
    # retain original columns with different names
    mutate(`% Poverty_n` = `% Poverty`,  `% English Language Learners_n`  =
`% English Language Learners`,
           `% Students with Disabilities_n` = `% Students with Disabilities`)
%>%
    # pivot columns
    gather(key = `Socio_indicator`, value = `% Socio_indicator`, `% Poverty`:
`% Students with Disabilities`)

# for racial factors (White, Asian, Black and Hispanic)
combined_db <- combined_db %>%
    # retain original columns with different names
    mutate(`% Asian_n`=`% Asian`, `% Black_n`=`% Black`, `% Hispanic_n`=`%
```

```
Hispanic`, `% White_n`=`% White`) %>%
    # pivot the columns
    gather(key = `Race`, value = `% Race`, `% Asian`:`% White`)

# for gender factors (male and female)
combined_db <- combined_db %>%
    # retain original columns with different names
    mutate(`% Male_n`=`% Male`, `% Female_n`=`% Female`) %>%
    # pivot columns
    gather(key = `Gender`, value = `% Gender`, `% Male`:`% Female`)

# for survey results
combined_db <- combined_db %>%
    # pivot columns
    gather(key = Response_category , value = Response_score ,
saf_p_11:aca_tot_11) %>%
    # extract information about response category
    mutate(Response_type = str_sub(Response_category,1,3)) %>%
    # extract information about response group
    mutate(Response_by = str_sub(Response_category,5,6))

# indicate the response category
combined_db <- combined_db %>%
  mutate(Response_type = if_else(Response_type == 'saf', 'Safety and
Respect',
                                 if_else(Response_type == 'com',
'Communication',
                                         if_else(Response_type == 'eng',
'Engagement',
                                                 if_else(Response_type ==
'aca', 'Academic Expectations', 'NA')))))
# indicate the response group
combined_db <- combined_db %>%
  mutate(Response_by = if_else(Response_by == 'p_', 'Parent',
                               if_else(Response_by == 't_', 'Teacher',
                                       if_else(Response_by == 's_',
'Student',
                                               if_else(Response_by == 'to',
'Total','NA')))))
```

Finally, I have cleaned up the data and it is ready for next step. I will do the following data visualization by using Tableau, and you can find them on my Tableau Public page (click here).