Quality of NYC Schools - Survey Analysis

Sandro Mikautadze

Last compiled on 13/07/2022

Contents

1 Introduction

This is a TLDR. Enjoy!

1.1 What Is the Project About?

- Do student, teacher and parent perceptions of NYC school quality appear to be related to demographic and academic success metrics?
- Do students, teachers, and parents have similar perceptions of NYC school quality?

2 Data

2.1 Raw Data Analysis

2.1.1 Initial Remarks

In `data\raw-data` 5 files are available: **combined.csv**, **masterfile11_gened_final.txt**, **masterfile11_gened_final.xlsx**, **masterfile11_d75_final.txt** and **masterfile11_d75_final.xlsx**.

INSERT WHERE YOU DOWNLOADED THE FILES FROM.

Without importing the files yet, from the *Survey-Data-Dictionary* file in `data\metadata` we can notice that **masterfile11_gened_final** and **masterfile11_d75_final** differ by a small aspect: **gened** contains information on all community schools, while **d75** from all District 75 schools, that is schools designed to teach and help students with disabilities. As the Dictionary states, "these files display one line of information for each school, by DBN, that includes the response rate for each school, the number of surveys submitted, the size of the eligible survey population at each school, question scores, the percentage of responses selected, and the count of responses selected".

Both files come with two different formats: *.txt* and *.xlsx*. I decide to work working with *.txt*, because the Excel version requires paid software to be visualized (i.e. Microsoft Excel). Having a look at the *.txt* datasets, we can notice that they are actually saved as tsv (tab separated value) files.

The **combined** dataset has been pre-cleaned as an exercise and contains combined information on different NYC schools based on SAT, AP scores and geographical data.

2.1.2 Dataset Loading and Preview

Importing the `readr` package under `tidyverse`, I will save the datasets as `combined`, `general` and `district`, respectively for **combined.csv**, **masterfile11_gened_final.txt** and **masterfile11_d75_final.txt**.

```
dim(combined)
```

```
## [1] 479  30
```

```
dim(general)
```

```
## [1] 1646 1942
```

```
dim(district)
```

```
## [1]   56 1773
```

Looking at the Survey Dictionary we can notice that the first columns indicate somecharacteristics of the school (we'll get into that later). After that, there are some columns that contain aggregate data on the survey. We can identify three groups that responded to the survey: - Students, encoded by `s` - Teachers, encoded by `t` - Parents, encoded by `p`

They were asked questions on 4 main categories: - Safety and Respect, encoded by `saf` - Communication, encoded by `com` - Engagement, encoded by `eng` - Academic expectations, encoded by `aca`

In addition those columns contain at the end a number: 11. We need to be aware of the fact that in the dictionary, that number is 10; so it might represent the year.

EXAMPLE: `eng_p_11` indicates the engagement score collected in 2011 based on the parent responses.

After the above described columns, we have thousands of columns on the precise survey question and answers.

As far as `combined` goes, we mainly have data on SAT scores with some other info on the different groups of people attending the school, the school's position, the class size, etc. Overall, all these pieces of information might come useful, so I decide to perform no cleaning.

2.2 Data Processing

Since we don't really care about the specific survey responses that are present in pretty much all columns but the initial ones, I can say that we can exclude them. Moreover, since it would be great to match performance and perception of school quality to the SAT scores, we can exclude Elementary and Middle Schools from the dataset.

```
unique(general$schooltype)
```

```
## [1] "Elementary School"                "Elementary / Middle School"
## [3] "Middle / High School"             "Middle School"
## [5] "High School"                      "Elementary / Middle / High School"
## [7] "Early Childhood School"           "YABC"
```

We are going to keep only "Middle / High School", "High School", ""Elementary / Middle / High School".

In the d75 dataset the `schooltype` column has a unique value:

```
unique(district$schooltype)
```

```
## [1] "District 75 Special Education"
```

This value might refer to either elementary school of high school. So, for now, we are going to keep every row, and do the necessary cleaning later on.

You can find the code of the "reductions" in `src/00-data-processing.r` under the CLEANING comment.

```
dim(combined_reduced)
```

```
## [1] 479  27
```

```
dim(general_reduced)
```

```
## [1] 483  23
```

```
dim(district_reduced)
```

```
## [1] 56 23
```

Now we are dealing with a feasible number of variables and they are closer to what we really need.

2.2.1 Inspecting NA values

```
colSums(is.na(combined_reduced))
```

```
##                                  dbn                          school_name
##                                    0                                    0
##                num.of.sat.test.takers                        avg_sat_score
##                                   57                                   57
##                       ap.test.takers                    total.exams.taken
##                                    0                                  247
## number.of.exams.with.scores.3.4.or.5                    exams_per_student
##                                  328                                  247
##                     high_score_percent                       avg_class_size
##                                  328                                   44
##                           frl_percent                     total_enrollment
##                                   41                                   41
##                           ell_percent                        sped_percent
##                                   41                                   41
##                      selfcontained_num                           asian_per
##                                   51                                   41
##                             black_per                        hispanic_per
##                                   41                                   41
##                             white_per                            male_per
##                                   41                                   41
##                            female_per                        total.cohort
##                                   41                                   89
##                         grads_percent                      dropout_percent
##                                  111                                  111
##                                 boro                                  lat
##                                  109                                  109
##                                 long
##                                  109
```

```
colSums(is.na(general_reduced))
```

```
##              dbn               bn       schoolname              d75
##                0                0                0                0
## studentssurveyed       highschool       schooltype         saf_p_11
##                0              483                0                0
```

```
##         com_p_11         eng_p_11         aca_p_11         saf_t_11
##                0                0                0                0
##         com_t_11         eng_t_11         aca_t_11         saf_s_11
##                0                0                0                4
##         com_s_11         eng_s_11         aca_s_11       saf_tot_11
##                4                4                4                0
##       com_tot_11       eng_tot_11       aca_tot_11
##                0                0                0
```

```
colSums(is.na(district_reduced))
```

```
##              dbn               bn       schoolname              d75
##                0                0                0                0
## studentssurveyed       highschool       schooltype         saf_p_11
##                0               41                0                0
##         com_p_11         eng_p_11         aca_p_11         saf_t_11
##                0                0                0                0
##         com_t_11         eng_t_11         aca_t_11         saf_s_11
##                0                0                0                2
##         com_s_11         eng_s_11         aca_s_11       saf_tot_11
##                2                2                2                0
##       com_tot_11       eng_tot_11       aca_tot_11
##                0                0                0
```