Quality of NYC Schools - Survey Analysis

Sandro Mikautadze

Last compiled on 12/07/2022

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**.

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. 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.

ADD INFORMATION ON COMBINED.CSV

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 32, I can say that we can exclude them. You can find the code of the "reduction" in `src/00-data-processing.r`

```
dim(general_reduced)
```

```
## [1] 1646   23
```

```
dim(district_reduced)
```

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

Now we are dealing with a feasible number of variables