Quality of NYC Schools - Survey Analysis

Sandro Mikautadze

Last compiled on 22/07/2022

Contents

1 Introduction

This is a TLDR. Enjoy!

1.1 Research Questions

- 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 Cleansing

2.1 Initial Remarks on Raw Data

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**.

These files have been downloaed from the following links:

- https://data.cityofnewyork.us/Education/2011-NYC-School-Survey/mnz3-dyi8 [last visited July 7th, 2022]
- https://data.world/dataquest/nyc-schools-data/workspace/file?filename=combined.csv) [last visited Julty 13th, 2022]

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 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.

2.2 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 some characteristics of the school (we'll get into those later in the report). 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`

In addition, each group was 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`

Those columns contain at the end the 11. We need to be aware of the fact that in the dictionary, that number is 10; so it might represent the year of data collection. This is still non influencing the analysis, so we will discard this detail.

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.

2.3 Raw Data Cleaning

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 "High School" rows.

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

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

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

This value might refer either to elementary school or to high school. In this case the `studentsurveyed` column can help us, because, as written in the dictionary, "This field indicates whether or not this school serves any students in grades 6-12". The values that the column takes are the following:

```
unique(district$studentssurveyed)
```

```
## [1] "Yes" "No"
```

Therefore by keeping only the columns with value "Yes" we will only have high schools, which are what we are interested in.

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] 383  23
```

```
dim(district_reduced)
```

```
## [1] 55 23
```

Now we are dealing with a feasible number of variables and they are closer to what we really need. We can combine the data of the survey in a new dataframe, called `survey`.

```
glimpse(survey)
```

```
## Rows: 438
## Columns: 23
## $ dbn              <chr> "01M448", "01M458", "01M509", "01M515", "01M650", "01~
## $ bn               <chr> "M448", "M458", "M509", "M515", "M650", "M696", "M047~
## $ schoolname       <chr> "University Neighborhood High School", "Forsyth Satel~
## $ d75              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ studentssurveyed <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes~
## $ highschool       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ schooltype       <chr> "High School", "High School", "High School", "High Sc~
## $ saf_p_11         <dbl> 7.9, 8.1, 7.7, 8.3, 9.0, 8.8, 8.9, 7.6, 8.7, 8.0, 7.5~
## $ com_p_11         <dbl> 7.4, 7.0, 7.4, 7.2, 8.4, 8.2, 7.7, 7.0, 8.1, 7.3, 7.1~
## $ eng_p_11         <dbl> 7.2, 6.7, 7.2, 7.4, 8.1, 8.3, 7.9, 6.9, 7.9, 7.1, 6.9~
## $ aca_p_11         <dbl> 7.3, 7.6, 7.3, 7.5, 8.6, 9.1, 8.1, 7.6, 8.3, 7.5, 7.5~
## $ saf_t_11         <dbl> 6.6, 8.5, 6.4, 9.1, 7.6, 8.2, 8.1, 7.3, 8.0, 8.6, 6.6~
## $ com_t_11         <dbl> 5.8, 8.2, 5.3, 7.3, 7.5, 7.4, 6.1, 7.1, 7.7, 8.1, 6.3~
## $ eng_t_11         <dbl> 6.6, 8.9, 6.1, 8.7, 8.3, 7.5, 7.7, 7.8, 7.9, 8.7, 6.8~
## $ aca_t_11         <dbl> 7.3, 8.9, 6.8, 9.1, 8.7, 8.3, 7.2, 7.7, 8.9, 8.9, 7.1~
## $ saf_s_11         <dbl> 6.0, 6.8, 6.4, 8.0, 8.1, 8.3, 7.3, 6.2, 7.4, 7.1, 6.6~
## $ com_s_11         <dbl> 5.7, 6.1, 5.9, 6.3, 6.9, 7.3, 6.3, 5.7, 6.5, 6.5, 6.2~
## $ eng_s_11         <dbl> 6.3, 6.1, 6.4, 7.0, 7.9, 8.0, 7.0, 6.1, 7.3, 7.0, 6.7~
## $ aca_s_11         <dbl> 7.0, 6.8, 7.0, 7.3, 8.4, 8.9, 7.5, 7.2, 7.6, 7.4, 7.5~
## $ saf_tot_11       <dbl> 6.8, 7.8, 6.9, 8.5, 8.3, 8.5, 8.1, 7.0, 7.9, 7.9, 6.9~
## $ com_tot_11       <dbl> 6.3, 7.1, 6.2, 7.0, 7.6, 7.6, 6.7, 6.6, 7.3, 7.3, 6.6~
## $ eng_tot_11       <dbl> 6.7, 7.2, 6.6, 7.7, 8.1, 8.0, 7.5, 6.9, 7.7, 7.6, 6.8~
## $ aca_tot_11       <dbl> 7.2, 7.8, 7.0, 8.0, 8.6, 8.7, 7.6, 7.5, 8.2, 8.0, 7.4~
```

2.4 NA Values Inspection

To better clean the data we can have a look at columns with 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(survey))
```

```
##              dbn               bn       schoolname              d75
##                0                0                0                0
## studentssurveyed       highschool       schooltype         saf_p_11
##                0              424                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                3
##         com_s_11         eng_s_11         aca_s_11       saf_tot_11
##                3                3                3                0
##       com_tot_11       eng_tot_11       aca_tot_11
##                0                0                0
```

The first thing that we can notice is that the `highschool` column in the survey dataframe has 424 NA values, out of 438 observations. This means that that column is pretty much unusable, so we will delete it.

In addition, `combined_reduced` has `number.of.exams.with.scores.3.4.or.5` and `high_score_percent` with 328 NA values, which is more than half of the rows in the dataset. So, it is safe to assume that those columns are useless and we will delete them.

The final dimensions of the cleaned datasets are the following:

```
dim(combined_reduced_2)
```

```
## [1] 479  26
```

```
dim(survey_2)
```

```
## [1] 438  22
```

2.5 Joining the Datasets

Now that the necessary cleaning has been done, we can finally join `survey` and `combined_reduced` into one dataset, that we are going to be using for the analysis.

We are going to apply a `left_join` to `combined_reduced` so that we will have all values for schools of which we have SAT data. We will save it as `school_data_raw`. These are the initial dimensions:

```
dim(school_data_raw)
```

```
## [1] 479  47
```

We can eliminate some redundant columns, such `bn` and `schoolname`. In addition, we now know that we are dealing with high schools, so we can drop `schooltype` and `studentssurveyed`.

We can also notice that there is an duplicated value in the schools

```
sum(duplicated(school_data_raw$dbn))
```

```
## [1] 1
```

So we will remove that duplicate as well.

2.6 Final Dataset

Therefore our final cleaned dataset, named `school_data` is the following:

```
glimpse(school_data)
```

```
## Rows: 478
## Columns: 44
## $ X                      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, ~
## $ dbn                    <chr> "01M292", "01M448", "01M450", "01M458", "01M509~
## $ school_name            <chr> "HENRY STREET SCHOOL FOR INTERNATIONAL STUDIES"~
## $ num.of.sat.test.takers <int> 29, 91, 70, 7, 44, 112, 159, 18, 130, 16, 62, 5~
## $ avg_sat_score          <int> 1122, 1172, 1149, 1174, 1207, 1205, 1621, 1246,~
## $ ap.test.takers         <dbl> 2.5, 39.0, 19.0, 2.5, 2.5, 24.0, 255.0, 2.5, 2.~
## $ total.exams.taken      <int> NA, 49, 21, NA, NA, 26, 377, NA, NA, NA, NA, NA~
## $ exams_per_student      <dbl> NA, 1.256410, 1.105263, NA, NA, 1.083333, 1.478~
## $ high_score_percent     <dbl> NA, 20.408163, NA, NA, NA, 92.307692, 50.663130~
## $ avg_class_size         <int> 23, 22, 21, 23, 24, 23, 26, 22, 21, 16, 23, 15,~
## $ frl_percent            <dbl> 88.6, 71.8, 71.8, 72.8, 80.7, NA, 23.0, 69.8, 1~
## $ total_enrollment       <int> 422, 394, 598, 224, 367, NA, 1613, 218, 617, 17~
## $ ell_percent            <dbl> 22.3, 21.1, 5.0, 4.0, 11.2, NA, 0.2, 3.2, 0.2, ~
## $ sped_percent           <dbl> 24.9, 21.8, 26.4, 8.9, 25.9, NA, 2.7, 6.9, 0.8,~
## $ selfcontained_num      <int> 35, 10, 19, 0, 36, NA, 0, 0, 0, 10, 4, 2, 17, 3~
## $ asian_per              <dbl> 14.0, 29.2, 9.7, 2.2, 9.3, NA, 27.8, 0.5, 15.1,~
## $ black_per              <dbl> 29.1, 22.6, 23.9, 34.4, 31.6, NA, 11.7, 45.4, 1~
## $ hispanic_per           <dbl> 53.8, 45.9, 55.4, 59.4, 56.9, NA, 14.2, 49.5, 1~
## $ white_per              <dbl> 1.7, 2.3, 10.4, 3.6, 1.6, NA, 44.9, 4.1, 49.8, ~
## $ male_per               <dbl> 61.4, 57.4, 54.7, 43.3, 46.3, NA, 49.2, 39.9, 3~
## $ female_per             <dbl> 38.6, 42.6, 45.3, 56.7, 53.7, NA, 50.8, 60.1, 6~
## $ total.cohort           <int> 78, 124, 90, NA, 84, 193, 46, 89, 139, 25, 102,~
## $ grads_percent          <dbl> 55.1, 42.7, 77.8, NA, 56.0, 54.4, 100.0, 55.1, ~
## $ dropout_percent        <dbl> 14.1, 16.1, 5.6, NA, 6.0, 18.1, 0.0, 6.7, 0.7, ~
## $ boro                   <chr> "Manhattan", "Manhattan", "Manhattan", NA, "Man~
## $ lat                    <dbl> 40.71376, 40.71233, 40.72978, NA, 40.72057, NA,~
## $ long                   <dbl> -73.98526, -73.98480, -73.98304, NA, -73.98567,~
## $ d75                    <int> NA, 0, NA, 0, 0, 0, NA, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ saf_p_11               <dbl> NA, 7.9, NA, 8.1, 7.7, 8.3, NA, 9.0, 8.8, 8.9, ~
## $ com_p_11               <dbl> NA, 7.4, NA, 7.0, 7.4, 7.2, NA, 8.4, 8.2, 7.7, ~
## $ eng_p_11               <dbl> NA, 7.2, NA, 6.7, 7.2, 7.4, NA, 8.1, 8.3, 7.9, ~
## $ aca_p_11               <dbl> NA, 7.3, NA, 7.6, 7.3, 7.5, NA, 8.6, 9.1, 8.1, ~
## $ saf_t_11               <dbl> NA, 6.6, NA, 8.5, 6.4, 9.1, NA, 7.6, 8.2, 8.1, ~
## $ com_t_11               <dbl> NA, 5.8, NA, 8.2, 5.3, 7.3, NA, 7.5, 7.4, 6.1, ~
## $ eng_t_11               <dbl> NA, 6.6, NA, 8.9, 6.1, 8.7, NA, 8.3, 7.5, 7.7, ~
## $ aca_t_11               <dbl> NA, 7.3, NA, 8.9, 6.8, 9.1, NA, 8.7, 8.3, 7.2, ~
## $ saf_s_11               <dbl> NA, 6.0, NA, 6.8, 6.4, 8.0, NA, 8.1, 8.3, 7.3, ~
## $ com_s_11               <dbl> NA, 5.7, NA, 6.1, 5.9, 6.3, NA, 6.9, 7.3, 6.3, ~
## $ eng_s_11               <dbl> NA, 6.3, NA, 6.1, 6.4, 7.0, NA, 7.9, 8.0, 7.0, ~
## $ aca_s_11               <dbl> NA, 7.0, NA, 6.8, 7.0, 7.3, NA, 8.4, 8.9, 7.5, ~
## $ saf_tot_11             <dbl> NA, 6.8, NA, 7.8, 6.9, 8.5, NA, 8.3, 8.5, 8.1, ~
## $ com_tot_11             <dbl> NA, 6.3, NA, 7.1, 6.2, 7.0, NA, 7.6, 7.6, 6.7, ~
## $ eng_tot_11             <dbl> NA, 6.7, NA, 7.2, 6.6, 7.7, NA, 8.1, 8.0, 7.5, ~
## $ aca_tot_11             <dbl> NA, 7.2, NA, 7.8, 7.0, 8.0, NA, 8.6, 8.7, 7.6, ~
```

You can find the cleaned dataset in `data/clean-data/school-data.csv`.

3 Data Analysis

The questions are the following:

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

For demographics metrics, in the dataset we have observations on the latitude, longitude and borough and race (*btw, this is an American thing... that's really racist. No wonder America is one of the most racist countries in the world!*). For our purposes the borough is enough, as the latitude and longitude provide overly detailed information on the position of the school. Here is some information on the boroughs that might be useful to understand their differences.

```
##             Area Median Household Income (USD) Mean Household Income (USD)
## 1          Bronx                         34156                          46
## 2       Brooklyn                         41406                         298
## 3      Manhattan                         64217                          46
## 4         Queens                         53171                         298
## 5 Staten Islands                         66985                         121
##   Percentage in Poverty (%)
## 1                      27.1
## 2                      21.9
## 3                      17.6
## 4                      12.0
## 5                       9.8
```

For academic metrics, in the dataset we have many data points on SAT results and class information. To properly address the question, we will consider as a metric the average SAT score. I also decide to keep other details like:

- `frl_percent`: percentage of a school's students eligible for receiving school lunch at a discount based on household income
- `ell_percent`: percentage of a school's students who are learning to speak English
- `sped_percent`: percentage of a school's students who receive specialized instruction to accommodate special needs such as learning or physical disabilities

They could provide more insights.

I decide to remove some other columns that might be useless (see code for more info).

I will also create new columns:

- `avg_p`, `avg_t` and `avg_s` to indicate the average score on the different questions type that each group answered to
- `avg_saf`, `avg_com`, `avg_eng` and `avg_aca` to indicate the average score on the average satisfaction on the different categories by all groups.

Our processed dataset for the question is now the following

```
glimpse(school_data_question)
```

```
## Rows: 478
## Columns: 35
## Rowwise:
## $ dbn             <chr> "01M292", "01M448", "01M450", "01M458", "01M509", "01M~
## $ school_name     <chr> "HENRY STREET SCHOOL FOR INTERNATIONAL STUDIES", "UNIV~
## $ avg_sat_score   <int> 1122, 1172, 1149, 1174, 1207, 1205, 1621, 1246, 1856, ~
## $ avg_class_size  <int> 23, 22, 21, 23, 24, 23, 26, 22, 21, 16, 23, 15, 23, 21~
## $ frl_percent     <dbl> 88.6, 71.8, 71.8, 72.8, 80.7, NA, 23.0, 69.8, 18.0, 66~
## $ ell_percent     <dbl> 22.3, 21.1, 5.0, 4.0, 11.2, NA, 0.2, 3.2, 0.2, 8.0, 2.~
## $ sped_percent    <dbl> 24.9, 21.8, 26.4, 8.9, 25.9, NA, 2.7, 6.9, 0.8, 32.2, ~
## $ asian_per       <dbl> 14.0, 29.2, 9.7, 2.2, 9.3, NA, 27.8, 0.5, 15.1, 1.7, 3~
## $ black_per       <dbl> 29.1, 22.6, 23.9, 34.4, 31.6, NA, 11.7, 45.4, 15.1, 32~
## $ hispanic_per    <dbl> 53.8, 45.9, 55.4, 59.4, 56.9, NA, 14.2, 49.5, 18.2, 59~
## $ white_per       <dbl> 1.7, 2.3, 10.4, 3.6, 1.6, NA, 44.9, 4.1, 49.8, 6.3, 4.~
## $ male_per        <dbl> 61.4, 57.4, 54.7, 43.3, 46.3, NA, 49.2, 39.9, 31.3, 42~
## $ female_per      <dbl> 38.6, 42.6, 45.3, 56.7, 53.7, NA, 50.8, 60.1, 68.7, 57~
## $ grads_percent   <dbl> 55.1, 42.7, 77.8, NA, 56.0, 54.4, 100.0, 55.1, 96.4, 7~
## $ dropout_percent <dbl> 14.1, 16.1, 5.6, NA, 6.0, 18.1, 0.0, 6.7, 0.7, 4.0, 2.~
## $ boro            <chr> "Manhattan", "Manhattan", "Manhattan", NA, "Manhattan"~
## $ saf_p_11        <dbl> NA, 7.9, NA, 8.1, 7.7, 8.3, NA, 9.0, 8.8, 8.9, 7.6, 8.~
## $ com_p_11        <dbl> NA, 7.4, NA, 7.0, 7.4, 7.2, NA, 8.4, 8.2, 7.7, 7.0, 8.~
## $ eng_p_11        <dbl> NA, 7.2, NA, 6.7, 7.2, 7.4, NA, 8.1, 8.3, 7.9, 6.9, 7.~
## $ aca_p_11        <dbl> NA, 7.3, NA, 7.6, 7.3, 7.5, NA, 8.6, 9.1, 8.1, 7.6, 8.~
## $ saf_t_11        <dbl> NA, 6.6, NA, 8.5, 6.4, 9.1, NA, 7.6, 8.2, 8.1, 7.3, 8.~
## $ com_t_11        <dbl> NA, 5.8, NA, 8.2, 5.3, 7.3, NA, 7.5, 7.4, 6.1, 7.1, 7.~
## $ eng_t_11        <dbl> NA, 6.6, NA, 8.9, 6.1, 8.7, NA, 8.3, 7.5, 7.7, 7.8, 7.~
## $ aca_t_11        <dbl> NA, 7.3, NA, 8.9, 6.8, 9.1, NA, 8.7, 8.3, 7.2, 7.7, 8.~
## $ saf_s_11        <dbl> NA, 6.0, NA, 6.8, 6.4, 8.0, NA, 8.1, 8.3, 7.3, 6.2, 7.~
## $ com_s_11        <dbl> NA, 5.7, NA, 6.1, 5.9, 6.3, NA, 6.9, 7.3, 6.3, 5.7, 6.~
## $ eng_s_11        <dbl> NA, 6.3, NA, 6.1, 6.4, 7.0, NA, 7.9, 8.0, 7.0, 6.1, 7.~
## $ aca_s_11        <dbl> NA, 7.0, NA, 6.8, 7.0, 7.3, NA, 8.4, 8.9, 7.5, 7.2, 7.~
## $ avg_p           <dbl> NA, 7.35, NA, 7.30, 7.35, 7.45, NA, 8.50, 8.55, 8.00, ~
## $ avg_t           <dbl> NA, 6.60, NA, 8.70, 6.25, 8.90, NA, 7.95, 7.85, 7.45, ~
## $ avg_s           <dbl> NA, 6.15, NA, 6.45, 6.40, 7.15, NA, 8.00, 8.15, 7.15, ~
## $ avg_saf         <dbl> NA, 6.6, NA, 8.1, 6.4, 8.3, NA, 8.1, 8.3, 8.1, 7.3, 8.~
## $ avg_com         <dbl> NA, 5.8, NA, 7.0, 5.9, 7.2, NA, 7.5, 7.4, 6.3, 7.0, 7.~
## $ avg_eng         <dbl> NA, 6.6, NA, 6.7, 6.4, 7.4, NA, 8.1, 8.0, 7.7, 6.9, 7.~
## $ avg_aca         <dbl> NA, 7.3, NA, 7.6, 7.0, 7.5, NA, 8.6, 8.9, 7.5, 7.6, 8.~
```

From here we proceed constructing the correlation matrix keeping only values with correlation $|r| \geq 0.2$.

```
cor_df
```

```
## # A tibble: 16 x 2
##    variable        avg_sat_score
##    <chr>                   <dbl>
##  1 avg_sat_score           1    
##  2 avg_class_size          0.395
##  3 frl_percent            -0.724
##  4 ell_percent            -0.392
##  5 sped_percent           -0.438
##  6 asian_per               0.567
##  7 black_per              -0.308
##  8 hispanic_per           -0.372
##  9 white_per               0.651
## 10 grads_percent           0.546
## 11 dropout_percent        -0.481
## 12 saf_t_11                0.309
## 13 saf_s_11                0.277
## 14 aca_s_11                0.293
## 15 avg_s                   0.237
## 16 avg_saf                 0.301
```

Still using the previous dataset, we can plot some graphs. You can visualize the graphs in the correct format in the `output-graphics` folder.

4 Findings

4.1 Do student, teacher and parent perceptions of NYC school quality appear to be related to demographic and academic success metrics?

From `sat-borough` graph we can notice the average SAT score in each borough. In increasing order, we have Bronx (1153), Brooklyn (1178), Manhattan (1291), Queens (1292), Staten Island (1383). First thing to notice: the poorer the borough, the lower the average SAT score in that borough. Refer to table in Section 3 for more info.

Inspecting graph `response-borough-per-category` we can identify some patterns.

- "Safety and respect" appears the less constant category on average and no true intuitive relationship can be found at first sight.
- Communication increases in mean as the borough has a higher average SAT score.
- Engagement has a pretty equal mean and distribution for the first four boroughs, but the mean increases in Staten Islands. This pattern can also slightly be identified in communication.
- Academic expectation seems to have a common mean across all boroughs.

Staten Island, though, seems to be the borough with the most skewed data. For academic expectation most responses aggregate around the mean. Whereas for communication and engagement 50% of the data differs by one point (which is a lot).

In addition, looking at the the graphs starting with `sat-vs` we can notice that there is some correlation between SAT score and teacher safety perception (0.309), student academic expectation perception (0.293) and student safety perception (0.277).

This is not really useful but draws some interesting insights that would require further investigation.

Overall, even though some patterns can be found, I believe that they are not that meaningful for "poorer" boroughs. Instead, there seems to be some slight shift between Staten Island and the others in school perception. In the end, the change is so small that drawing further conclusions might not be appropriate.

So, to answer the question, the data suggests some slight relationship and to better address the question some further investigation would be necessary.

4.2 Do students, teachers, and parents have similar perceptions of NYC school quality?

We can have a look at the graph called `response-borough-per-group`. The first impression is that they have completely different opinions. Overall, parents appear more satisfied than teachers, who seem more satisfied than students.

The answer to the question seems easy: no.

But in addition, we can find some trends. As the borough gets less poor, parent satisfaction decreases in mean (note that Staten Island has most of its scores concentrated around the mean here as well!). Instead, teacher's satisfaction increases! This might be because of better working conditions (?) but it would require more data on that. On the other hand, students perceptions seem pretty stable (and low), regardless of the borough and some score fluctuations.

Should we conclude that students hate school. . . ?!

4.3 Final Personal Remark

I have not drawn any causal connection. I just stated relationships and facts. It is not my aim to seek motivations behind the correlation of events. If you feel like doing it, be free to do it.

5 Methods of Improvement

Too many to write down :)