Exploring NYC High School Survey Data

David Dunn

May 9, 2023

In this project we will be examining the relationship between responses to surveys from parents, teachers, and students from NYC schools and the scholastic performance of students from those schools. Data was obtained from the New York City Department of Education. The two main questions we seek to answer are: 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?

```r
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
require(readxl)
```

```
## Loading required package: readxl
```

```r
combined <- read_csv("combined.csv")
```

```
## Rows: 479 Columns: 30
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): DBN, school_name, boro
## dbl (27): Num of SAT Test Takers, SAT Critical Reading Avg. Score, SAT Math ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
d75 <- read_tsv("masterfile11_d75_final.txt")
```

```
## Rows: 56 Columns: 1773
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr    (5): dbn, bn, schoolname, studentssurveyed, schooltype
## dbl (1739): d75, highschool, rr_s, rr_t, rr_p, N_s, N_t, N_p, nr_s, nr_t, nr...
## lgl   (29): p_q5, p_q9, p_q13a, p_q13b, p_q13c, p_q13d, p_q14a, p_q14b, p_q1...
##
```

```
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
gened <- read_tsv("masterfile11_gened_final.txt")
```

```
## Rows: 1646 Columns: 1942
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr    (5): dbn, bn, schoolname, studentssurveyed, schooltype
## dbl (1904): d75, highschool, rr_s, rr_t, rr_p, N_s, N_t, N_p, nr_s, nr_t, nr...
## lgl   (33): p_q1, p_q3d, p_q9, p_q10, p_q12aa, p_q12ab, p_q12ac, p_q12ad, p_...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dict <- read_excel("Survey Data Dictionary.xls")
```

```
## New names:
## * `` -> `...2`
```

Looking at the unique values for the schooltype variable so we know how to filter for High Schools.

```
gen_uni <- unique(gened$schooltype)
d75_uni <- unique(d75$highschool)
print(gen_uni)
```

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

```
print(d75_uni)
```

```
## [1]  0 NA
```

We need to simplify the data and drop the variables we don’t need to work with. We can also select for observations that we know or suspect to be high schools. We use str_detect() to find schools that may be listed as Middle / High School as those will still interest us.

```
gened_1 <- gened %>%
    filter(str_detect(`schooltype`, "High", negate = FALSE)) %>%
    select(dbn:aca_tot_11)
d75_1 <- d75 %>%
    filter(is.na(highschool)) %>%
    select(dbn:aca_tot_11)
```

Now that the important observations and variables have been isolated we need to combine the general education and district 75 dataframes. Once we do that we will need to join the new dataframe to the combined dataframe choosing dbn as the key. We choose left_join() in order to

```
gen75 <- bind_rows(gened_1, d75_1) %>%
    rename(DBN = dbn) %>%
    select(-bn, -d75, -studentssurveyed, -highschool)
combined_survey <- combined %>%
    left_join(gen75, by = "DBN")
```

We want to know if student, teacher, and parent perceptions of NYC school quality appear to be related to demographic and academic success. In order to gauge this we need to create a correlation matrix and scatter plots.

```r
cor_mat <- combined_survey %>%
    select(where(is.numeric)) %>%
    cor(use = "pairwise.complete.obs")
```

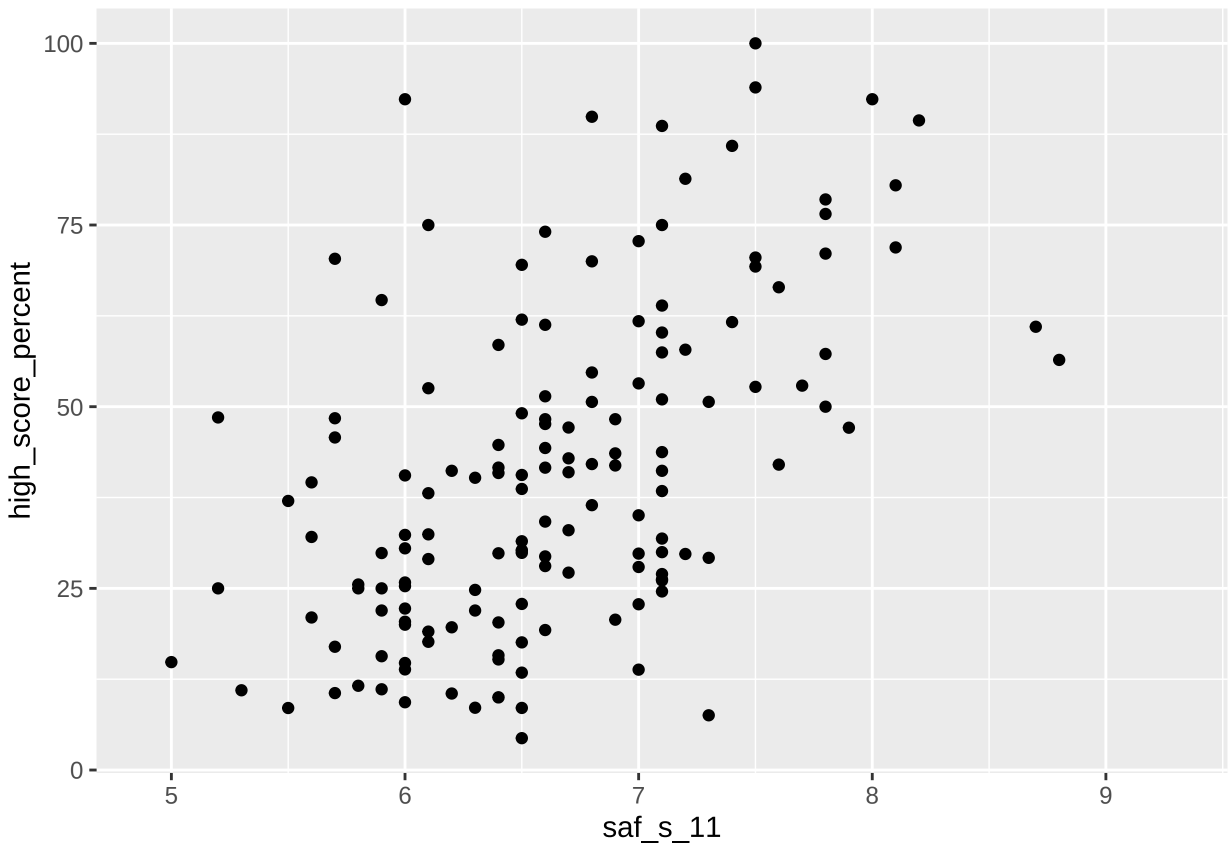
Converting the correlation matrix of all numeric variables we select single variable to compare against all others while filtering for “significant” correlation.

```r
cor_tib <- cor_mat %>%
    as_tibble(rownames = "variable")
high_score_cors <- cor_tib %>%
    select(variable, high_score_percent) %>%
    filter(high_score_percent > 0.25 | high_score_percent < -0.25)
frl_cors <- cor_tib %>%
    select(variable, frl_percent) %>%
    filter(frl_percent > 0.25 | frl_percent < -0.25)
avg_sat_score <- cor_tib %>%
    select(variable, avg_sat_score) %>%
    filter(avg_sat_score > 0.25 | avg_sat_score < -0.25)
aca_tot_cors <- cor_tib %>%
    select(variable, aca_tot_11) %>%
    filter(aca_tot_11 > 0.25 | aca_tot_11 < -0.25)
saf_tot_cors <- cor_tib %>%
    select(variable, saf_tot_11) %>%
    filter(saf_tot_11 > 0.25 | saf_tot_11 < -0.25)
com_tot_cors <- cor_tib %>%
    select(variable, com_tot_11) %>%
    filter(com_tot_11 > 0.25 | com_tot_11 < -0.25)
eng_tot_cors <- cor_tib %>%
    select(variable, eng_tot_11) %>%
    filter(eng_tot_11 > 0.25 | eng_tot_11 < -0.25)
saf_s_cors <- cor_tib %>%
    select(variable, saf_s_11) %>%
    filter(saf_s_11 > 0.25 | saf_s_11 < -0.25)
rr_s_cors <- cor_tib %>%
    select(variable, rr_s) %>%
    filter(rr_s > 0.25 | rr_s < -0.25)
rr_t_cors <- cor_tib %>%
    select(variable, rr_t) %>%
    filter(rr_t > 0.25 | rr_t < -0.25)
rr_p_cors <- cor_tib %>%
    select(variable, rr_p) %>%
    filter(rr_p > 0.25 | rr_p < -0.25)
```

Plotting a few of the variables with high Pearson coefficients to see if they are worth exploring.

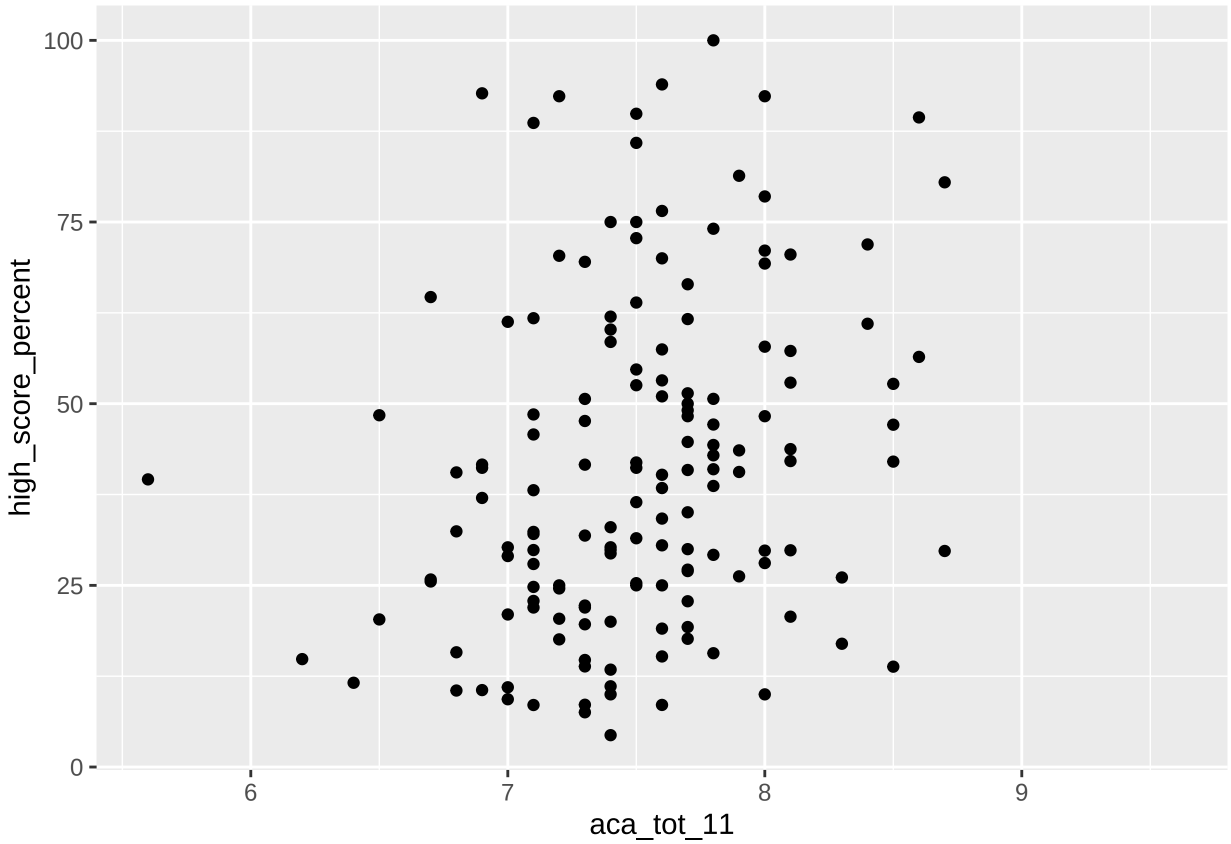
```r
ggplot(data = combined_survey,
       aes(x=saf_s_11, y=high_score_percent)) +
        geom_point()
```

```
## Warning: Removed 329 rows containing missing values (`geom_point()`).
```



```
ggplot(data = combined_survey,
       aes(x=aca_tot_11, y=high_score_percent)) +
        geom_point()
```

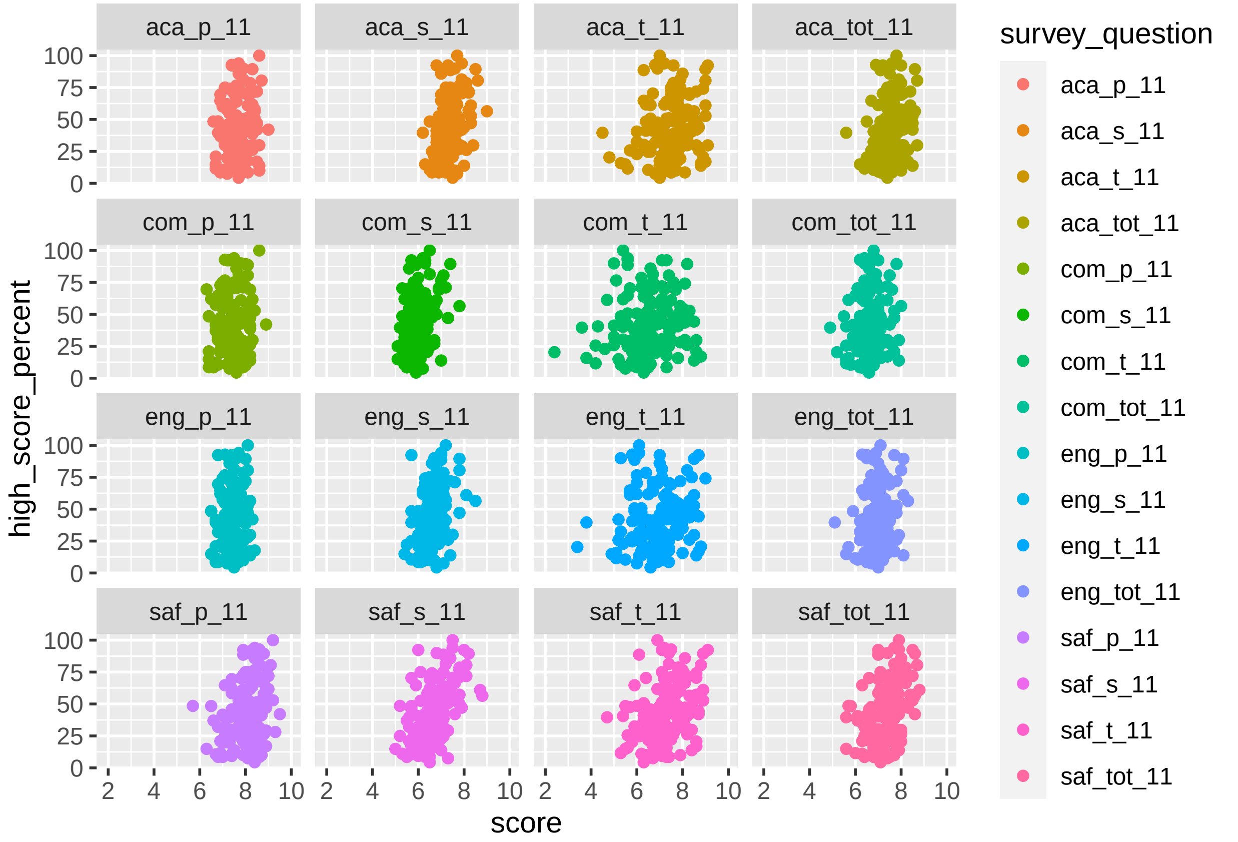
```
## Warning: Removed 328 rows containing missing values (`geom_point()`).
```



We are pivoting the dataframe longer in order to prepare the survey data for plotting. All 16 of the survey response variables will pivot into the survey_question variable while the values for them will pivot into the score variable.

```
combined_survey_longer <- combined_survey %>%
    pivot_longer(cols = c(saf_p_11, com_p_11, eng_p_11, aca_p_11, saf_t_11, com_t_11, eng_t_11, aca_t_1
                 names_to = "survey_question",
                 values_to = "score")
ggplot(data = combined_survey_longer,
       aes(x = score, y = high_score_percent, color = survey_question)) +
        geom_point() +
        facet_wrap(~survey_question)
```

```
## Warning: Removed 5252 rows containing missing values (`geom_point()`).
```



Using recursion to parse the survey_question variable and extract the substrings "p", "s", and "t" in order to create a new variable "response_type". Still working out the syntax, need to revisit.

```
#combined_survey_longer_2 <- combined_survey_longer %>%
#    mutate(reponse_type = str_sub(survey_question, 4, 6)) %>%
#    mutate(response_type = if_else(response_type == "_p_", "parent",
#                   if_else(response_type == "_s_", "student",
#                   if_else(response_type == "_t_", "teacher",
#                   if_else(response_type == "_tot_", "total", "NA")))))
```

Cleaner approach to the above operation. This parses out 3 substrings separated by "_" which gives 3 new variables which we are calling "metric" for survey question type, "response_type" for type of person who responded, and "year" for the year the survey was conducted. Since all values of year are the same we don't need this so we select it out.

```
combined_survey_longer_1 <- combined_survey_longer %>%
    separate(col = `survey_question`,
             into = c("metric", "response_type", "year"),
             sep = "_") %>%
    select(-year)
```

Creating a summary grouping by type of respondents and survey metric i.e. academic, communications, safety, engagement. We can see that in general parents gave higher ratings on average and specifically the greatest disparity was between student and parent ratings of safety. Teachers ratings were slightly higher than students with the biggest disparity also being safety.

```
summary <- combined_survey_longer_1 %>%
    group_by(response_type, metric) %>%
```

```r
    summarize(mean(score, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'response_type'. You can override using the
## `.groups` argument.
```

```r
print(summary)
```

```
## # A tibble: 16 x 3
## # Groups:   response_type [4]
##    response_type metric `mean(score, na.rm = TRUE)`
##    <chr>         <chr>                        <dbl>
##  1 p             aca                           7.84
##  2 p             com                           7.65
##  3 p             eng                           7.54
##  4 p             saf                           8.20
##  5 s             aca                           7.41
##  6 s             com                           6.15
##  7 s             eng                           6.68
##  8 s             saf                           6.68
##  9 t             aca                           7.51
## 10 t             com                           6.51
## 11 t             eng                           6.99
## 12 t             saf                           7.13
## 13 tot           aca                           7.58
## 14 tot           com                           6.77
## 15 tot           eng                           7.07
## 16 tot           saf                           7.33
```