Dataquest Guided Project: NYC Schools Perceptions

Cindy Zhang

8/22/2020

Contents

Introduction

This is my solution to Dataquest's Guided Project from the first Data Cleaning in R course, which analyzes survey responses on quality perceptions of NYC schools. This analysis focuses on answering two questions:

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?

More details such as the RMD and csv files can be found in the repository in GitHub. More details about the survey response variables can be found here.

Findings

Step 1: Loading the Dataframes

I loaded the combined NYC school characteristics dataset supplied by Dataquest and text files containing the raw results from the 2011 NYC School Survey.

```
combined <- data.frame(read_csv("combined.csv"))
gened <- data.frame(read_tsv("masterfile11_gened_final.txt"))
d75 <- data.frame(read_tsv("masterfile11_d75_final.txt"))
```

`gened` contains the survey results for general education schools while `d75` contains survey results for District 75 schools, which provide special education support for children with special needs such as learning or physical disabilities.

Step 2: Simplifying the Dataframes

The survey dataframes need to be pared down- there are 1942 rows in the `gened` dataframe and 1773 in the `d75` dataframe. I am only interested in the following variables: `dbn` (I will be using this as a key to combine all the dataframes together), the columns containing aggregate scores for each survey group, and some additional characteristics like high school name, location, enrollment size, borough. I did not use all of these columns for this analysis, but including them lets me keep them around for future analysis of this dataset. `highschool` in particular was useful to filter out survey results from schools that do not service high school-age students (for the purposes of this analysis, I only focused on high school students).

Step 3: Creating a Single Dataframe for Analysis

To create a single dataframe, I first combined the two survey dataframes together. I then combined the singular survey results dataframe with the `combined` dataframe using a left join.

```
surveys <- bind_rows(gened, d75)
combined_surv <- combined %>%
  rename(dbn = DBN) %>%
  left_join(surveys, by = "dbn") %>%
  select(1:61)
```

Step 4: Looking for Interesting Correlations

To answer the first question about the relationship between survey responses and academic/demographic metrics, I created a correlation matrix, which calculates a value from -1 to 1 of how correlated a pair of variables are. Quick statistics refresher: the closer to -1 (negative) or 1 (positive), the stronger the correlation. From the matrix, I extracted a tibble, that assigns the variable names to the first column of the tibble.

```
cor_mat <- combined_surv %>%
  select_if(is.numeric) %>%
  cor(use="pairwise.complete.obs")

cor_tib <- cor_mat %>%
  as_tibble(rownames = "variable")
```

Next, I created a more selective tibble based on the variables I was interested in investigating- I filtered for academic/demographics in the columns and survey responses in the rows.

```r
cor_select <- cor_tib %>%
  select(1:55) %>%
  slice(39:54)
```

Finally, I wrote a function that would match the variable names to each correlation coefficient and sorting the coefficients in descending order, largest to smallest value.

```r
cor_func <- function(data, x, y) {
  data %>%
    dplyr::select(x,y) %>%
    dplyr::arrange(desc(!!sym(y)))
}

x_var <- names(cor_select)[1]
y_var <- names(cor_select)[c(6, 11:13, 15:16, 18:23, 25:26)]

cor_rank <- map(y_var, cor_func, data = cor_select, x = x_var)
```

In the following sub-sub sections, I identify academic/demographic variables that had correlation coefficients greater than 0.25, which suggests a relationship with survey responses. I've included the definitions for each part of the survey response variable below for ease of understanding:

- Survey topics:
  - saf = Safety and Respect Aggregate Score
  - aca = Academic Expectations Score
  - eng = Engagement Score
  - com = Communication Score
- Survey audience:
  - _t = Teachers
  - _s = Students
  - _p = Parents
  - _tot = total score from all survey audiences

Each survey response variable is formatted as: "[survey topic]*[survey audience]*[survey year]" (e.g., aca_t_11 refers to the 2011 Academic Engagement Score among teachers).

Average SAT Score: Positive Correlations

```r
avg_sat_score <- as.data.frame(cor_rank[1])
pandoc.table(avg_sat_score, style = "rmarkdown", emphasize.strong.cells=which(avg_sat_score[2] > 0.25, a
```

variable	avg_sat_score
saf_t_11	0.303
aca_s_11	0.2859
saf_tot_11	0.2804

variable	avg_sat_score
saf_s_11	0.2717
aca_tot_11	0.1773
eng_s_11	0.167
com_s_11	0.1632
aca_t_11	0.1372
saf_p_11	0.1132
eng_tot_11	0.09558
com_t_11	0.09367
com_tot_11	0.08815
eng_t_11	0.0488
aca_p_11	0.03305
eng_p_11	0.03144
com_p_11	-0.09085

Average SAT score has the highest positive correlation with Safety and Respect aggregate scores from teachers, suggesting that the higher the SAT score, the higher the Safety and Respect score from teachers. Other survey response variables with positive correlations with average SAT score include Academic Expectations score from students, Safety and Respect score from all audiences, and Safety and Respect score from students.

Percentage of High AP Scores: Strong Positive Correlations

```
high_score_percent <- as.data.frame(cor_rank[2])
pandoc.table(high_score_percent, style = "rmarkdown", emphasize.strong.cells=which(high_score_percent[2
```

variable	high_score_percent
saf_s_11	0.5329
saf_tot_11	0.4259
eng_s_11	0.3967
aca_s_11	0.3896
com_s_11	0.3634
saf_t_11	0.3269
aca_tot_11	0.2554
saf_p_11	0.251
eng_tot_11	0.222
aca_t_11	0.1986
com_tot_11	0.1914
aca_p_11	0.1453
eng_t_11	0.1167
eng_p_11	0.1069
com_p_11	0.09029
com_t_11	0.07796

Percentage of high AP scores (among students who took AP exams) is very strongly correlated with Safety and Respect scores from all audiences, individually and as a whole. This academic variable is strongly correlated with other student survey responses, which corresponds with the trend that students that score highly on AP exams likely attend prestigious high schools that achieve high survey scores from students. `high_score_percent` also has a strong positive correlation with the total Academic Expectation score from all audiences, again reinforcing the trend that prestigious high schools in NYC have high survey scores.

Average Class Size: Negative Correlations

```
avg_class_size <- as.data.frame(cor_rank[3])
pandoc.table(avg_class_size, style = "rmarkdown", emphasize.strong.cells=which(avg_class_size[2] < -0.25
```

variable	avg_class_size
com_t_11	0.1432
saf_t_11	0.05215
aca_t_11	0.0409
eng_t_11	0.0368
com_tot_11	-0.1069
eng_tot_11	-0.1227
aca_tot_11	-0.1381
saf_tot_11	-0.1533
eng_p_11	-0.1883
aca_s_11	-0.2114
saf_s_11	-0.2285
eng_s_11	-0.2406
saf_p_11	-0.2845
aca_p_11	-0.2911
com_s_11	-0.31
com_p_11	-0.3156

Average class size is negatively correlated with several survey response variables among parents and students, most strongly with Communication scores. This corresponds with the idea that smaller class sizes allow for more effective teaching and communication; thus, high schools with smaller average class sizes scored higher among parents and students while schools with larger average class sizes scored lower.

Percentage of Special Ed Students: Strong Negative Correlations

```
sped_percent <- as.data.frame(cor_rank[6])
pandoc.table(sped_percent, style = "rmarkdown", emphasize.strong.cells=which(sped_percent[2] < -0.25, a
```

variable	sped_percent
com_t_11	-0.09224
com_p_11	-0.1314
eng_t_11	-0.139
eng_p_11	-0.1439
com_tot_11	-0.1977
aca_t_11	-0.2087
aca_p_11	-0.2293
eng_tot_11	-0.2426
saf_p_11	-0.2715
com_s_11	-0.2941
aca_tot_11	-0.3076

variable	sped_percent
eng_s_11	-0.3382
aca_s_11	-0.38
saf_t_11	-0.3812
saf_tot_11	-0.4329
saf_s_11	-0.4401

Percentage of Special Ed students has a strong negative correlation with several survey response variables among parents and students. In other words, schools with high percentages of special ed students tended to score lower on all survey topics, especially among students.

Percentage of Asian Students: Positive Correlations

```
asian_percent <- as.data.frame(cor_rank[7])
pandoc.table(asian_percent, style = "rmarkdown", emphasize.strong.cells=which(asian_percent[2] > 0.25, a
```

variable	asian_per
saf_s_11	0.2853
saf_t_11	0.2595
saf_tot_11	0.2397
aca_s_11	0.1602
eng_s_11	0.1493
com_s_11	0.1377
aca_t_11	0.09524
aca_tot_11	0.07204
eng_tot_11	0.04024
saf_p_11	0.03384
com_t_11	0.03056
eng_t_11	0.02395
com_tot_11	0.02139
aca_p_11	-0.07264
eng_p_11	-0.08095
com_p_11	-0.139

Percentage of Asian students is positively correlated with Safety and Respect scores among students and teachers, suggesting schools with higher percentages of Asian students scored higher on Safety and Respect.

Percentage of Black Students: Strong Negative Correlations

```
black_percent <- as.data.frame(cor_rank[8])
pandoc.table(black_percent, style = "rmarkdown", emphasize.strong.cells=which(black_percent[2] < -0.25,
```

variable	black_per
eng_p_11	-0.05404
com_t_11	-0.06177

variable	black_per
com_p_11	-0.06239
aca_p_11	-0.07324
eng_t_11	-0.08308
aca_t_11	-0.1279
com_tot_11	-0.1348
eng_tot_11	-0.153
aca_tot_11	-0.1572
aca_s_11	-0.1928
saf_p_11	-0.1956
com_s_11	-0.2062
eng_s_11	-0.2418
saf_t_11	-0.2872
saf_tot_11	-0.3602
saf_s_11	-0.4244

In sharp contrast to the previous sub-sub section, Percentage of Black students has a strong negative correlation with Safety and Respect scores among all audiences. These strong negative correlation suggest that schools with a high percentage of black students score lower on Safety and Respect, which corresponds with trends of school gentrification along racial divides in NYC.

Variables with Weak Correlations

The following demographic variables had weak correlations (correlation coefficient scores between -0.25 and 0.25) with survey response variables: `frl_percent` (percent of students that are on free/reduced lunch), `ell_percent` (percent of students that are English language learners), percentage of Hispanic students, percentages of male and female students, graduation percentages, and dropout percentages. These demographic variables have a weak relationship with survey response variables and likely do not influence student, teacher, and parent perceptions of NYC schools.

Step 5: Reshaping the Data Based on Differences in Student,Parent, and Teacher Perceptions

To answer the second question about the similarity between student, teacher, and parent perceptions of NYC schools, I used `pivot_longer` to reshape the data so that the survey response variable and the corresponding scores each have their own column:

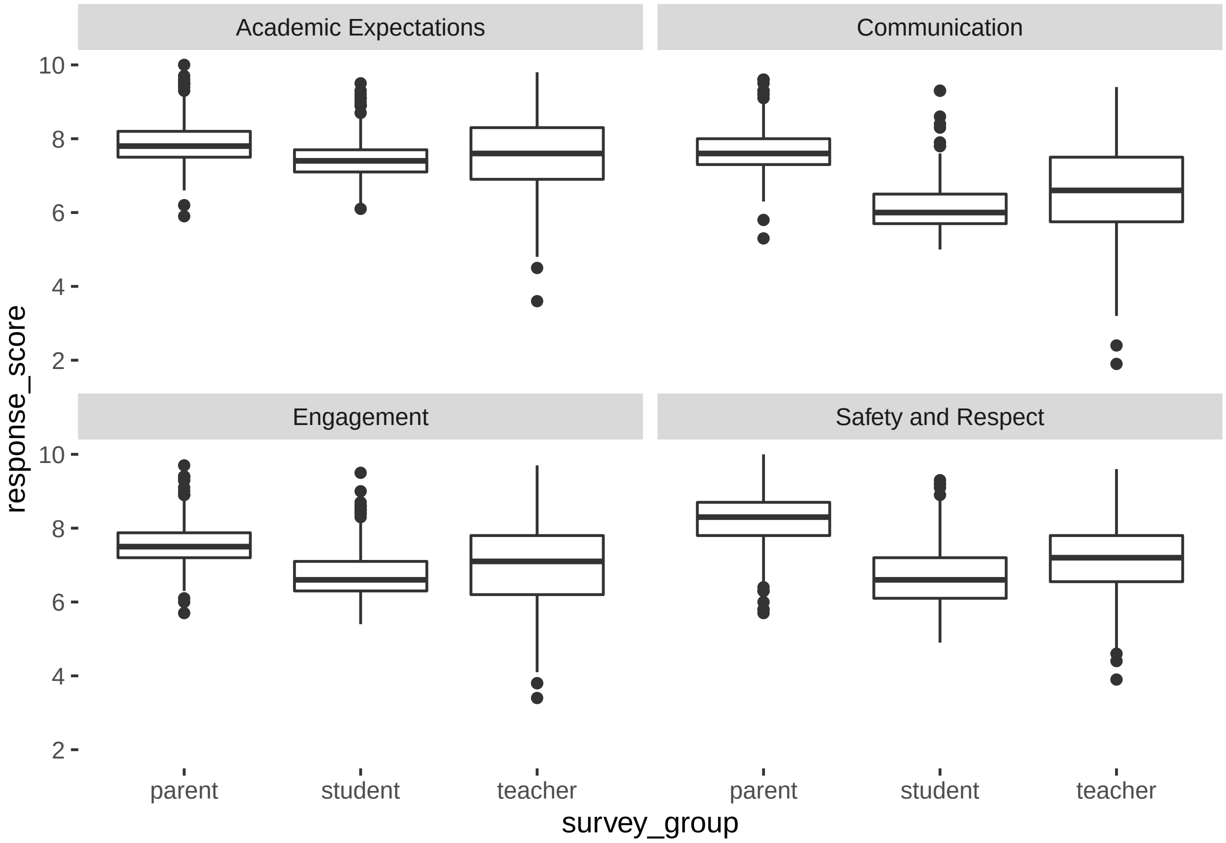
```
combined_surv_long <- combined_surv %>%
  pivot_longer(cols = c(46:57),
               names_to = "response_type",
               values_to = "response_score")
```

I added two new variables that designates each survey response according to the audience (e.g., student) and topic (e.g., Safety and Respect) in the original response variable formatting.

```
combined_surv_add <- combined_surv_long %>%
  mutate(survey_initial = str_sub(response_type, 4,6)) %>%
  mutate(survey_group = ifelse(survey_initial == "_p_", "parent", ifelse(survey_initial == "_t_", "teach
  mutate(topic_initial = str_sub(response_type, 1,3)) %>%
  mutate(survey_topic = ifelse(topic_initial == "saf", "Safety and Respect", ifelse(topic_initial == "co
```

Finally, I created box plots measuring each survey audience's scores for each of the four topics.

```
ggplot(data=combined_surv_add) +
  aes(x=survey_group, y=response_score)+
  geom_boxplot()+
  facet_wrap(~survey_topic, nrow=2)+
  theme(panel.background=element_rect(fill="white"))
```



Based on the boxplots:

- *Academic Expectations*: the range of scores among parents and students is similar, but there are more low outlier scores among teachers (also a larger distribution of scores).
- *Communication*: parents gave higher scores than students and teachers and have a somewhat even distribution across the quartiles and outliers, suggesting a favorable perception of communication among parents. Distribution of survey scores among students is smaller and the quartile scores are lower than parents and teachers, but there are more high score outliers. Distribution of survey scores is the largest among teachers and has more low score outliers.
- *Engagement*: similar trends as with communication; parents have overall favorable perceptions of engagement, students less favorable, and teachers more varied.
- *Safety and Respect*: while distribution of scores was higher among parents, there are more low outlier scores. Student score distribution is lower but has more high score outliers, while the opposite was the case for teachers.

Conclusion

Certain academic and demographic variables are strongly correlated with student, teacher, and parent perceptions of NYC schools. These correlations correspond with widely-observed trends, including but not limited to: * Perceptions of a lack of safety and respect among schools with high percentages of black students and special ed students. * Perceptions of ineffective communication among schools with large average class sizes. * Perceptions of strong academic expectations and safety and respect among schools with high average SAT scores and high percentages of students with high AP scores.

There are also noticeable differences in perceptions between students, parents, and teachers. Parents have overall favorable views on all four topics while students have less favorable views, especially in regards to communication. Teachers have the largest range of perceptions on all four topics and any outlier scores tend to be lower.

Despite the work done to inspect and clean the data for this analysis, there are some important caveats to consider. There is probably a survey response bias in the dataset- prestigious, well-funded schools likely have a greater response rate than underserved schools. Student and parent survey responses should also be scrutinized further, as the pattern of small score distribution among all four topics is a little suspect.

I plan to revisit this dataset in the future to see if more nuanced insights can be gleaned.

Thanks to everyone who's been reading so far and stay tuned for more projects!