NYC School Perceptions

Mari Osmundson

September 2019

This project uses data from two main sources: 1) the City of New York’s data website, found at <https://data.cityofnewyork.us/Education/2011-NYC-School-Survey/mnz3-dyi8>. The two files were imported from text. They contain information about the 2011 parent/teacher/student survey results from the NYC school district. 2) A cleaned dataset named “combined” was imported using read.csv, courtesy of Dataquest: <https://data.world/dataquest/nyc-schools-data/workspace/file?filename=combined.csv>. I also added average household income data for each NYC borough from a third source: <https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork,bronxcountybronxboroughnewyork,kingscountybrooklynboroughnewyork,newyorkcountymanhattanboroughnewyork,queenscountyqueensboroughnewyork,richmondcountystatenislandboroughnewyork/PST045218>.

I loaded the tidyverse package in R Notebook.

I filtered masterfile11\_gened\_final and masterfile11\_d75\_final by column V6, highschool, to retain only high schools. I selected relevant columns and saved the results to two dataframes: gened\_filter and d75\_filter.

gened\_filter <- masterfile11\_gened\_final %>%  
 filter(`V6` != "0") %>%  
 select(`V1`, `V2`, `V3`, `V4`, `V7`, `V17`:`V32`)  
  
d75\_filter <- masterfile11\_d75\_final %>%  
 filter(`V6` != "0") %>%  
 select(`V1`:`V3`, `V17`:`V32`)

Removed two more columns from gened\_filter in order to merge it with d75\_filter using rbind().

gened\_filter <- gened\_filter %>%  
 select(-`V4`, -`V7`)

Merged the two filtered dataframes into one.

gened\_d75 <- rbind(gened\_filter, d75\_filter)

Changed column name to DBN to use it as the key for the left join.

gened\_d75 <- (`DBN` = `V1`)

I joined the dataframes using DBN as the key, then belatedly realized the column names didn’t transfer when I imported the original text files into R Studio. I utilized colnames() to transfer all column names to combined\_join, then ran the code again to form the join.

combined\_join <- combined %>%  
 left\_join(gened\_d75, key = `DBN`)  
  
colnames(gened\_d75) <- c("DBN", "bn", "schoolname", "saf\_p\_11", "com\_p\_11", "eng\_p\_11", "aca\_p\_11", "saf\_t\_11", "com\_t\_11", "eng\_t\_11", "aca\_t\_11", "saf\_s\_11", "com\_s\_11", "eng\_s\_11", "aca\_s\_11", "saf\_tot\_11", "com\_tot\_11", "eng\_tot\_11", "aca\_tot\_11")  
  
combined\_join <- combined %>%  
 left\_join(gened\_d75, key = `DBN`)

The survey results needed to become numeric types in order to run correlations.

combined\_join <- combined\_join %>%  
 mutate\_at(33:48, as.numeric)

I created a correlation tibble to examine any relationship between average SAT scores and the survey variables.

combined\_matrix <- combined\_join %>%  
 select(avg\_sat\_score, saf\_p\_11:aca\_tot\_11) %>%  
 cor(use = "pairwise.complete.obs")  
  
combined\_tibble <- combined\_matrix %>%  
 as.tibble(rownames = "variable")

I dropped NAs from the borough column (boro) and removed NAs from calculations of each variable’s mean.

boro\_group <- combined\_join %>%  
 drop\_na(boro) %>%  
 group\_by(boro) %>%  
 summarize(average\_SAT\_score = mean(avg\_sat\_score, na.rm = TRUE),  
 average\_class\_size = mean(avg\_class\_size, na.rm = TRUE),  
 average\_frl\_percent = mean(frl\_percent, na.rm = TRUE),  
 average\_ell\_percent = mean(ell\_percent, na.rm = TRUE),  
 average\_sped\_percent = mean(sped\_percent, na.rm = TRUE),  
 average\_asian\_percent = mean(asian\_per, na.rm = TRUE),  
 average\_black\_percent = mean(black\_per, na.rm = TRUE),  
 average\_hispanic\_percent = mean(hispanic\_per, na.rm = TRUE),  
 average\_white\_percent = mean(white\_per, na.rm = TRUE),  
 average\_grads\_percent = mean(grads\_percent, na.rm = TRUE),  
 average\_dropout\_percent = mean(dropout\_percent, na.rm = TRUE),  
 average\_safety\_score = mean(saf\_tot\_11, na.rm = TRUE),  
 average\_comm\_score = mean(com\_tot\_11, na.rm = TRUE),  
 average\_engagement\_score = mean(eng\_tot\_11, na.rm = TRUE),  
 average\_expectations\_score = mean(aca\_tot\_11, na.rm = TRUE))

As noted in the introductory paragraph, I added each borough’s average household income to the dataframe.

boro\_avg\_income\_2017\_dollars <- c(36593, 52782, 79781, 62008, 76244)  
boro\_group$boro\_avg\_income <- boro\_avg\_income\_2017\_dollars

I created a new function for scatterplots.

create\_scatter <- function(x,y) {  
ggplot(data = boro\_group) +  
 aes\_string(x = x, y = y) +  
 geom\_point(alpha = 0.3)  
}  
  
x\_var <- names(boro\_group)[1]  
y\_var <- names(boro\_group)[2:17]  
  
boro\_variable\_comp <- map2(x\_var, y\_var, create\_scatter)  
  
print(boro\_variable\_comp)

There is a weak positive correlation between average SAT score and teachers’ safety scores, students’ safety scores, total safety score, and academic expectations of students.

combined\_select <- combined\_tibble %>%  
 select(variable, avg\_sat\_score) %>%  
 filter(avg\_sat\_score > 0.25 | avg\_sat\_score < -0.25)

I used gather() to organize the data by survey response type (parent, teacher, student, total) for each school.

comb\_resp\_gather <- combined\_join %>%  
 gather(key = "responder", value = score, saf\_p\_11:aca\_tot\_11)

Subset the responder column into response type (i.e. parent) and question (i.e. saf).

comb\_resp\_gather <- comb\_resp\_gather %>%  
 mutate(response\_type = str\_sub(responder, 4, 6 ),  
 question = str\_sub(responder, 1, 3))

Changed the abbreviated letters in response\_type to full words for clarity.

comb\_resp\_gather <- comb\_resp\_gather %>%  
 mutate(response\_type = ifelse(response\_type == "\_p\_", "parent",  
 ifelse(response\_type == "\_t\_", "teacher",  
 ifelse(response\_type == "\_s\_", "student",  
 ifelse(response\_type == "\_to", "total", "NA")))))

Changed the question column’s abbreviations into words for clarity.

comb\_resp\_gather <- comb\_resp\_gather %>%  
 mutate(question = ifelse(question == "saf", "safety+respect",  
 ifelse(question == "com", "communication",  
 ifelse(question == "eng", "engagement",  
 ifelse(question == "aca", "acad. expectations", "NA")))))

I experimented with some plots. I grouped the dataframe by response\_type and ran a bar chart plot to see the spread of averages. Teachers had the highest average score and students the lowest.

response\_group <- comb\_resp\_gather %>%  
 group\_by(response\_type) %>%  
 summarize(avg\_score = mean(score, na.rm = "TRUE"))  
  
response\_group\_plots <- ggplot(data = response\_group) +  
 aes(x = response\_type, y = avg\_score) +  
 geom\_bar(stat = "identity")  
print(response\_group\_plots)

I grouped by the question column and ran another bar chart plot on the averages. There was not great variation in the average scores among the question types. Safety and respect scored the highest, academic expectations scored the lowest.

question\_group <- comb\_resp\_gather %>%  
 group\_by(question) %>%  
 summarize(avg\_score = mean(score, na.rm = "TRUE"))

question\_group\_plots <- ggplot(data = question\_group) +  
 aes(x = question, y = avg\_score) +  
 geom\_bar(stat = "identity")  
print(question\_group\_plots)

In the following plot, teachers had both significantly higher scores in all four question types, as well as a greater distribution of scores within each question type. Students provided the lowest scores for each question.



It is interesting that in the next plot’s distribution, parents and students scored safety and respect higher than the other questions. Teachers showed a greater distribution of scores and higher scores than parents and students.

