NYC Schools Perceptions

RZ

6/29/2021

Import the necessary files and packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(dplyr)
library(purrr)
library(ggplot2)
library(tidyr)
# All data are from 2011 NYC online data
combined <- read_csv('combined.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   DBN = col_character(),
##   school_name = col_character(),
##   boro = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
survey <- read_tsv('2011 data files online/masterfile11_gened_final.txt')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   dbn = col_character(),
##   bn = col_character(),
##   schoolname = col_character(),
##   studentssurveyed = col_character(),
##   schooltype = col_character(),
##   p_q1 = col_logical(),
```

```
##   p_q3d = col_logical(),
##   p_q9 = col_logical(),
##   p_q10 = col_logical(),
##   p_q12aa = col_logical(),
##   p_q12ab = col_logical(),
##   p_q12ac = col_logical(),
##   p_q12ad = col_logical(),
##   p_q12ba = col_logical(),
##   p_q12bb = col_logical(),
##   p_q12bc = col_logical(),
##   p_q12bd = col_logical(),
##   t_q6m = col_logical(),
##   t_q9 = col_logical(),
##   t_q10a = col_logical()
##   # ... with 18 more columns
## )
## i Use `spec()` for the full column specifications.
```

```
survey_75 <- read_tsv('2011 data files online/masterfile11_d75_final.txt')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   dbn = col_character(),
##   bn = col_character(),
##   schoolname = col_character(),
##   studentssurveyed = col_character(),
##   schooltype = col_character(),
##   p_q5 = col_logical(),
##   p_q9 = col_logical(),
##   p_q13a = col_logical(),
##   p_q13b = col_logical(),
##   p_q13c = col_logical(),
##   p_q13d = col_logical(),
##   p_q14a = col_logical(),
##   p_q14b = col_logical(),
##   p_q14c = col_logical(),
##   p_q14d = col_logical(),
##   t_q11a = col_logical(),
##   t_q11b = col_logical(),
##   t_q14 = col_logical(),
##   t_q15a = col_logical(),
##   t_q15b = col_logical()
##   # ... with 14 more columns
## )
## i Use `spec()` for the full column specifications.
```

Filter the data frame to remove unnecessary columns

```
survey_select <- survey %>% select(dbn:aca_tot_11) %>% filter(schooltype=='High School')
survey_75_select <- survey_75 %>% select(dbn:aca_tot_11)
```

Combine the survey data with selected conditions above

```
survey_total <- bind_rows(survey_select, survey_75_select)
survey_total <- survey_total %>% rename(DBN=dbn)
combined_survey <- combined %>% left_join(survey_total, by='DBN')
```

Find the correlation and visualize the correlation

```
# Find the correlation matrix
cor_mat <- combined_survey %>% select(avg_sat_score, saf_p_11:aca_tot_11) %>% cor(use = "pairwise.compl
# Covert the correlation matrix to tibble which as variable as row names
cor_tib <- cor_mat %>%as_tibble(rownames = "variable")
# Find strong correlations
strong_cors <- cor_tib %>%select(variable, avg_sat_score) %>%filter(avg_sat_score > 0.25 | avg_sat_scor
```

Visualize the avg_sat_score to other strong correlation variables

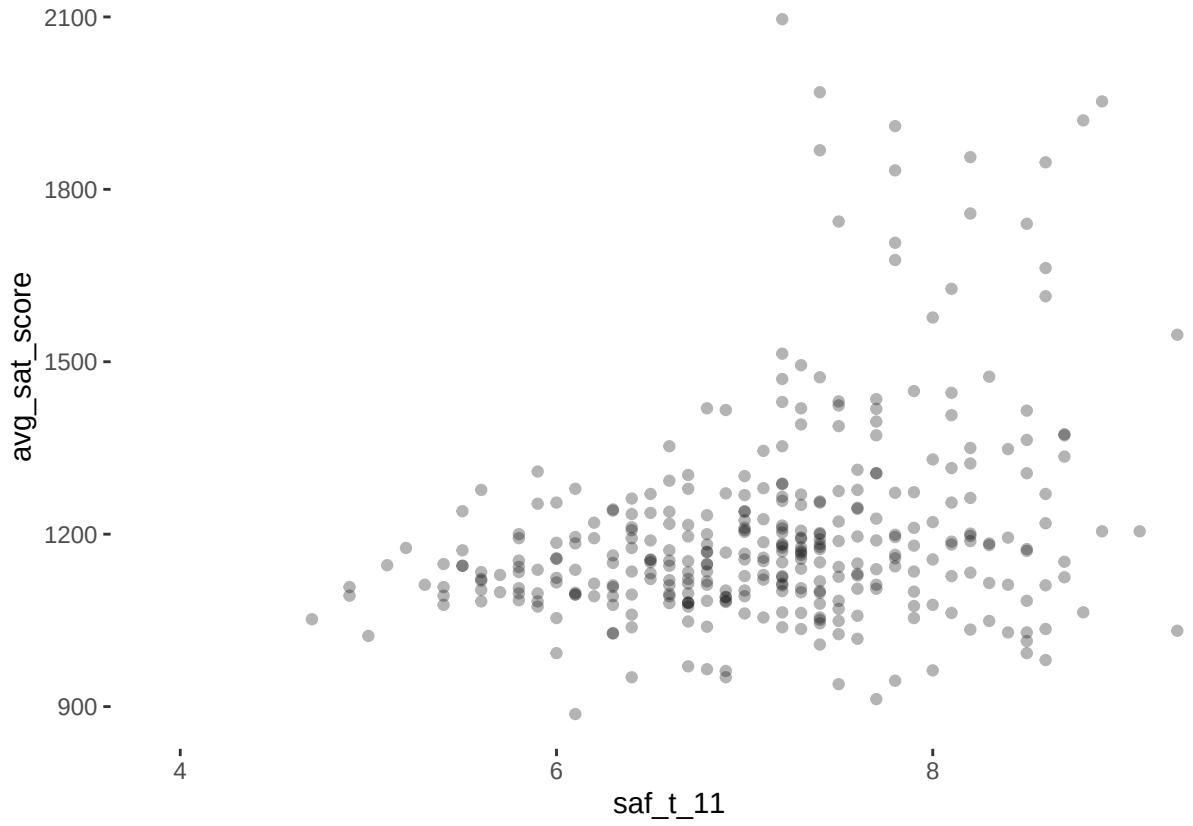
```
create_scatter <- function(x, y) {ggplot(data = combined_survey) + aes_string(x = x, y = y) + geom_poin

x_var <- strong_cors$variable[2:5]
y_var <- "avg_sat_score"

map2(x_var, y_var, create_scatter)
```

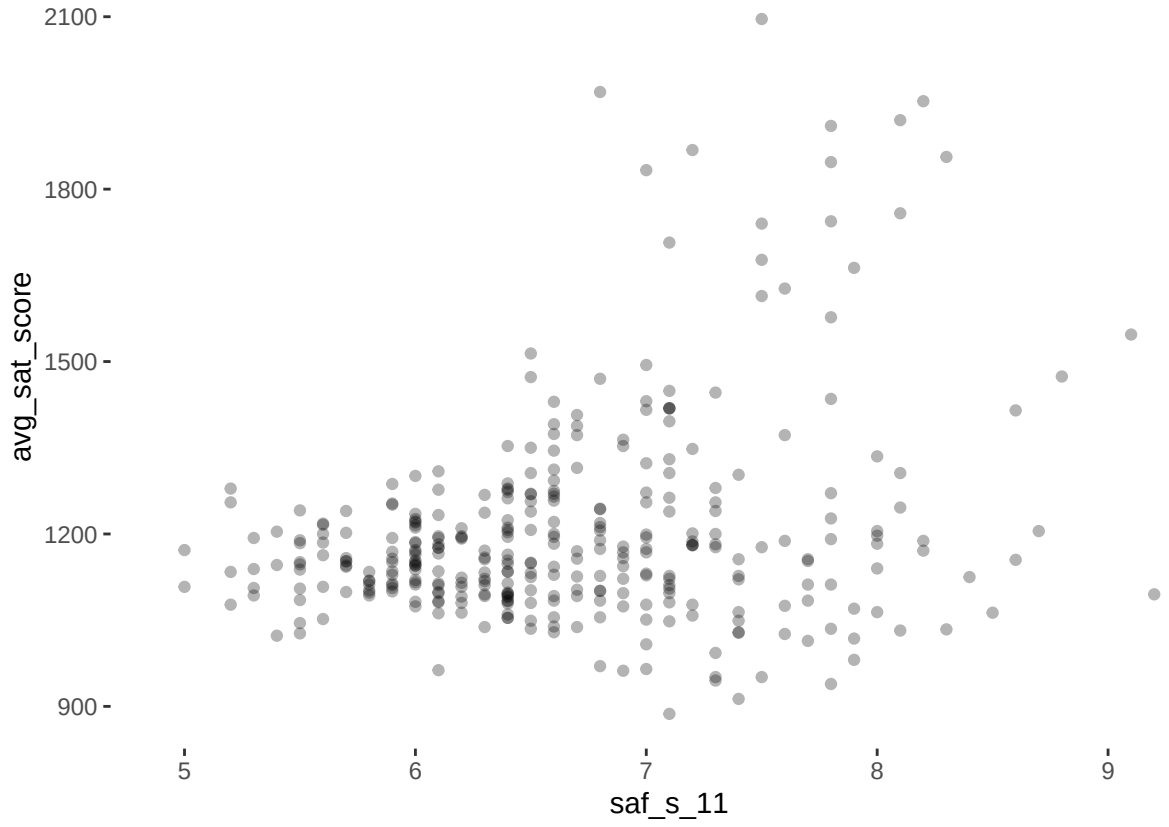
```
## [[1]]
```

```
## Warning: Removed 137 rows containing missing values (geom_point).
```



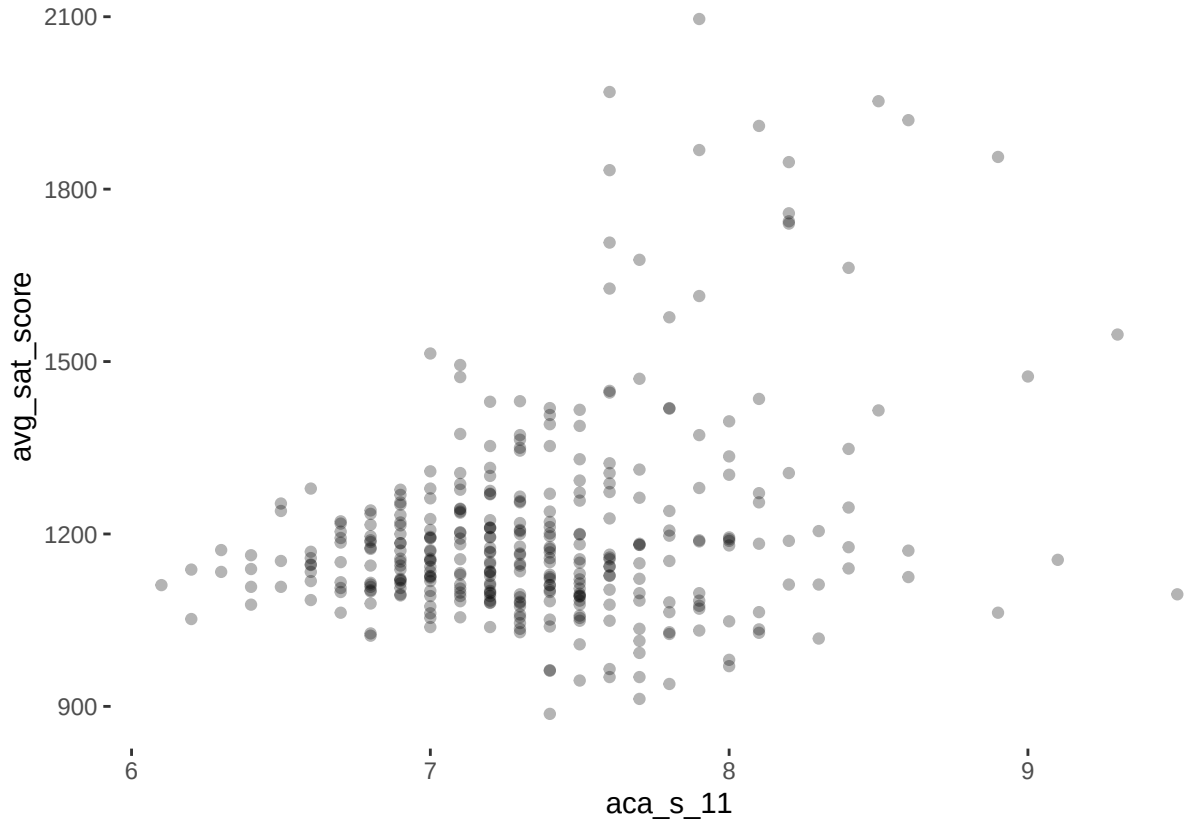

```
##
## [[2]]
```

```
## Warning: Removed 139 rows containing missing values (geom_point).
```



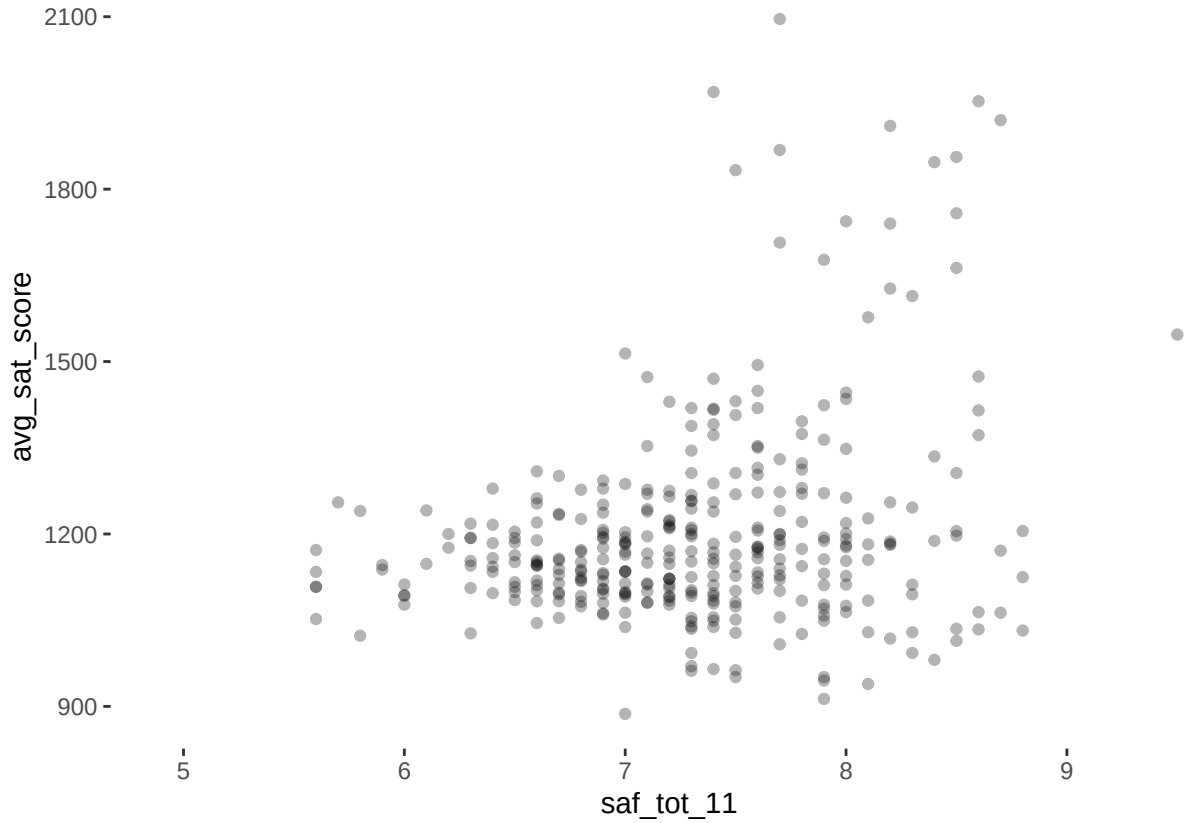

```
##
## [[3]]
```

```
## Warning: Removed 139 rows containing missing values (geom_point).
```




```
##
## [[4]]
```

```
## Warning: Removed 137 rows containing missing values (geom_point).
```



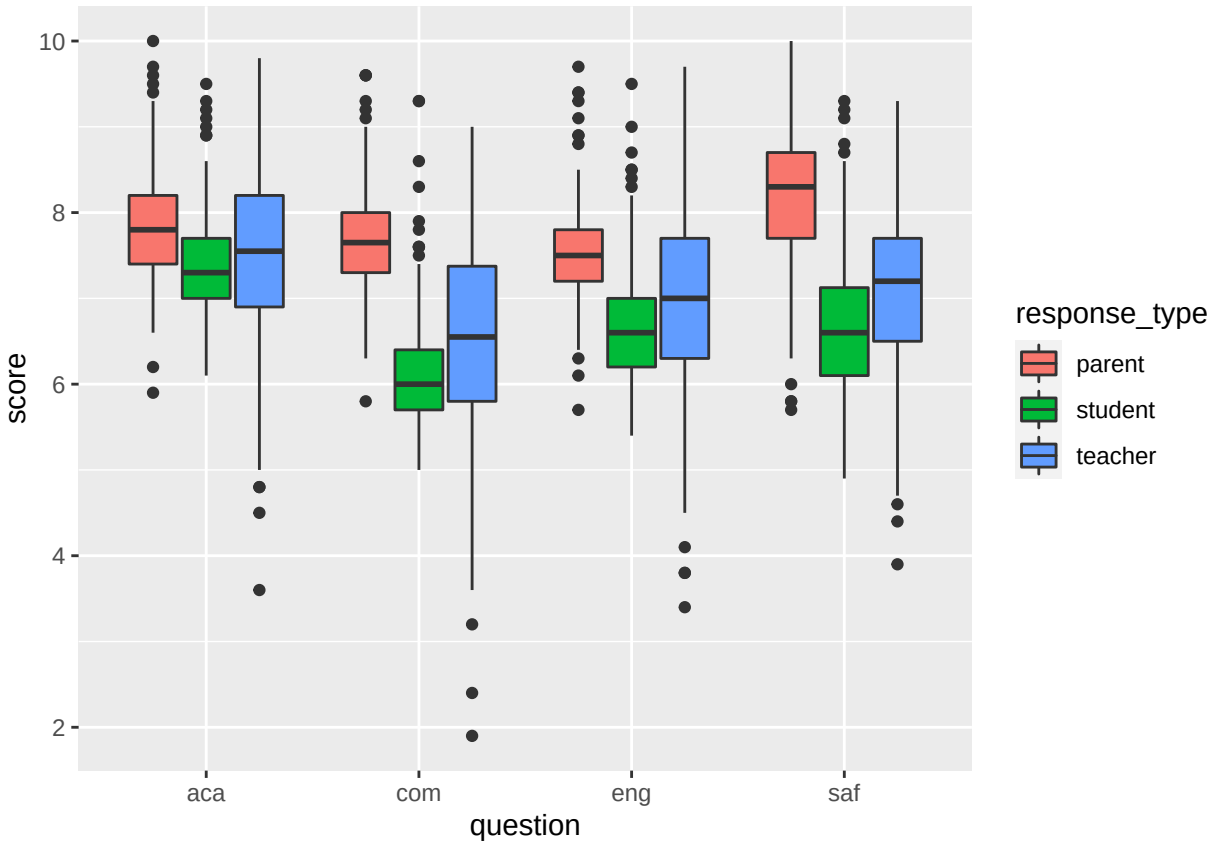
Reshape the data so that you can investigate differences in student, parent, and teacher responses to survey questions.

```
combined_survey_gather <- combined_survey %>%
  pivot_longer(cols = saf_p_11:aca_tot_11,
               names_to = "survey_question",
               values_to = "score")
# Extract values from the string
combined_survey_gather <- combined_survey_gather %>%
  mutate(response_type = str_sub(survey_question, 4, 6)) %>%
  mutate(question = str_sub(survey_question, 1, 3))

combined_survey_gather <- combined_survey_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")))))

# Make a box plot to see if there appear to be differences in how the three groups of responds (parents
combined_survey_gather %>%
  filter(response_type != "total") %>%
  ggplot(aes(x = question, y = score, fill = response_type)) +
  geom_boxplot()
```

```
## Warning: Removed 1268 rows containing non-finite values (stat_boxplot).
```


10
8
6
4
2
score
aca
com
eng
saf
question
response_type
parent
student
teacher