Analysis of New York City School Survey Data

Oksana Bihun

Contents

1 Introduction

In this project, we analyze New York City (NYC) School Survey data taken from NYC Open Data.

Author's skills demonstrated:

- **Coding in R**: all the subsequent analysis is performed using R programming language. The code is included.
- **R Markdown/Bookdown, R Studio**: this document is formatted using R Markdown/Bookdown in R Studio.
- **Data preparation**: obtaining data from public web sites and combining it into a single dataframe, converting dataframe entries to desired datatypes, and others.
- **Data exploration**: learning the basic properties of the dataframe such as the number of rows and columns, data types, and others.
- **Data cleaning**: disregarding empty data entries when computing correlations and others.
- **Data manipulation**: the relevant dataframes were pivoted and otherwise manipulated to prepare them for plotting and analysis.
- **Data visualization**: simultaneous creation of multiple plots with minimal code. Bar charts and scatter plots were created.
- **Data analysis**: making conclusions from the study of plots and others.
- **Statistical analysis**: the computation and the analysis of a correlation matrix.
- **Written communication**: the results of the study are formulated in layman's terms. The code is documented for the purpose of ease of modifications, if needed. The process of the data analysis is explained in great detail.

Questions answered for NYC high schools:

- 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 The Results of Data Analysis

In this section, we state the results of the data analysis performed in this paper.

2.1 Results of the Correllation Analysis

In this subsection, we formulate the results of the correlation analysis performed in Subsection 5.1. As a reminder, correlation does not imply causation. For example, the fact that the percentage of black students is negatively correlated with teacher's safety scores does not imply that higher percentages of black students cause teachers to feel that the school is less safe. Indeed, a third factor may be at play. For example, it is possible that schools in less safe neighborhoods have higher percentages of black students as well as lower safety scores by teachers. The reasons behind the reported correlations cannot be extracted from the data. Because of this, we report the significant correlations and refrain from making hypotheses about the reasons for the existence of these correlations.

Relationships between race and survey responses

- Schools with higher percentages of white students tend to have higher safety and respect scores by students.
- Schools with higher percentages of Asian students tend to have higher safety and respect scores by teachers.
- Schools with higher percentages of black students tend to have lower safety and respect scores by teachers and students.
- All the other survey responses appear to be unrelated to the race of the students.

Relationships between SAT scores and survey responses

- Schools with higher safety and respect scores by teachers, higher academic expectation scores by students, or higher total safety and respect scores, tend to have higher average SAT scores.
- Schools with higher safety and respect scores by parents, teachers, or students, higher communication scores by students, higher engagement scores by students, higher academic expectation scores by students, or higher total academic expectation scores, tend to have higher percentages of high SAT scores.
- Schools with higher safety and respect scores by teachers, students, or total, or academic expectations scores by students, tend to have higher average SAT critical reading, math, and writing scores.

Relationships between the percentages of special education students and survey responses

- Schools with higher percentages of special education students tend to have lower safety and respect scores by parents, teachers, students, or total.
- Schools with higher percentages of special education students tend to have lower communication scores by students.
- Schools with higher percentages of special education students tend to have lower engagement scores by students.
- Schools with higher percentages of special education students tend to have lower academic expectations by students or total.

Table 1: Significant relationships between some variables and the average SAT scores.

Variable	Relationship with the average SAT score
SAT critical reading, math, and writing scores	Direct
Percentage of white students	Direct
Percentage of students who graduated	Direct
Percentage of English as a foreign language learners	Inverse
Percentage of special education students	Inverse
Percentage of black students	Inverse
Percentage of hispanic students	Inverse
Percentage of students who dropped out	Inverse

Relationships of the Survey Scores among Themselves

- Each survey score is positively correlated with every other survey score except for the following. For example, schools with a higher survey score by parents tend to have higher survey scores by teachers and students, except for the following.
  - Safety and respect scores by parents and communication scores by teachers appear to be unrelated.
  - Communication scores by students and communication scores by teachers appear to be unrelated.

2.2 Results Based on the Scatter Plots of Each Variable Against the Average SAT Scores

The results stated below are based on the scatter plots presented in Subsection 5.2.2. The results are organized in the nearby table. We describe the relationship between a given variable and the average SAT score as "direct" if the increase in the value of the variable is associated with an increase in the average SAT score. We describe the relationship between a given variable and the average SAT score as "inverse" if the increase in the value of the variable is associated with a decrease in the average SAT score.

```
conclusions <- rbind(
c("Variable", "Relationship with the average SAT score"),
c("SAT critical reading, math, and writing scores", "Direct"),
c("Percentage of white students","Direct"),
c("Percentage of students who graduated","Direct"),
c("Percentage of English as a foreign language learners","Inverse"),
c("Percentage of special education students","Inverse"),
c("Percentage of black students","Inverse"),
c("Percentage of hispanic students","Inverse"),
c("Percentage of students who dropped out","Inverse"))

knitr::kable(conclusions, caption="Significant relationships
    between some variables and the average SAT scores.")
```

2.3 Results of the Analysis of Additional Scatter Plots

In Subsection 5.2, we created scatter plots that explore interesting relationships between the variables in more detail. These plots were consistent with the results of the correlation analysis so we do not reiterate the results obtained from the observations of the scatter plots. Subsection 5.2 is nevertheless retained in this document for the following two purposes: (1) to observe the dependency of significantly correlated variables in more detail and (2) to demonstrate author's data visualization skills.

2.4 Results of the Analysis of the Average Ratings by Parents, Students, and Teachers

In Section 5.3, we compared the average survey ratings given by the parents, the students, and the teachers. We discovered that the responses from the parents are higher, on average. Average students' response scores are the lowest, and teachers' are in between.

The subsequent sections detail the process through which the above conclusions were made.

3 Creating a Combined Dataframe: Data Cleaning

The NYC School Survey data is contained in three Excel files and two text files. In this section, we combine this data into a single dataframe.

We begin with reading the Excel and the text files.

```
library(tidyverse)
library(knitr)
library(readxl)

# Reading data dictionary
data_dict <- read_excel("2011_NYC_survey_files/Survey Data Dictionary.xls")
#data_dict[3:nrow(data_dict),]
```

Upon analyzing the data dictionary (uploaded above), we see that the following columns are relevant to the posed questions:

```
relevant_columns <- cbind(c("Field name","dbn",
  "schoolname","schooltype","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"),
  c("Field description","School identification code",
  "School name","School type","Safety and Respect score: parent responses",
  "Communication score: parent responses",
  "Engagement score: parent responses",
  "Academic expectations score: parent responses",
  "Safety and Respect score: teacher responses",
  "Communication score: teacher responses",
  "Engagement score: teacher responses",
  "Academic expectations score: teacher responses",
  "Safety and Respect score: student responses","Communication score: student responses",
  "Engagement score: student responses",
  "Academic expectations score: student responses",
  "Safety and Respect total score",
  "Communication total score",
  "Engagement total score",
  "Academic Expectations total score"))

prmatrix(relevant_columns,
  rowlab=rep("",nrow(relevant_columns)),
  collab=rep("",ncol(relevant_columns)))
```

```
##
##  "Field name" "Field description"
##  "dbn"        "School identification code"
```

```
##  "schoolname" "School name"
##  "schooltype" "School type"
##  "saf_p_11"   "Safety and Respect score: parent responses"
##  "com_p_11"   "Communication score: parent responses"
##  "eng_p_11"   "Engagement score: parent responses"
##  "aca_p_11"   "Academic expectations score: parent responses"
##  "saf_t_11"   "Safety and Respect score: teacher responses"
##  "com_t_11"   "Communication score: teacher responses"
##  "eng_t_11"   "Engagement score: teacher responses"
##  "aca_t_11"   "Academic expectations score: teacher responses"
##  "saf_s_11"   "Safety and Respect score: student responses"
##  "com_s_11"   "Communication score: student responses"
##  "eng_s_11"   "Engagement score: student responses"
##  "aca_s_11"   "Academic expectations score: student responses"
##  "saf_tot_11" "Safety and Respect total score"
##  "com_tot_11" "Communication total score"
##  "eng_tot_11" "Engagement total score"
##  "aca_tot_11" "Academic Expectations total score"
```

Next, we read the two text files provided by the NYC education website. The data in these files is tabulated, thus the use of the `read_tsv` command. We then reduce the number of columns in the corresponding dataframes by selecting only the relevant columns listed above.

```
# Reading file masterfile11_gened_final.txt
survey_1 <-
  read_tsv("2011_NYC_survey_files/masterfile11_gened_final.txt")

# Choosing columns from survey_2 relevant to the
# posed questions
survey_1 <- survey_1 %>%
  select(1,3,7,17:32)

# Reading file masterfile11_d75_final.txt
survey_2 <-
  read_tsv("2011_NYC_survey_files/masterfile11_d75_final.txt")

# Choosing columns from survey_2 relevant to the
# posed questions
survey_2 <- survey_2 %>% select(1,3,7,17:32)
```

Next, we merge dataframes `survey_1` and `survey_2` into a single dataframe by adding the rows of the latter to the former.

```
# Merging survey_1 and survey_2
combined_1 <- survey_1 %>% rbind(survey_2)
```

Next, we read the `combined.csv` file downloaded from here. This file contains the average exam scores and demographic data for each school. The column names of dataframe `combined_2` are printed below. Unfortunately, the meaning of the column names is not explained in the data dictionary on the web site where the data is posted. Hence, we will derive the meaning from the names of the columns, see below.

```
library(tidyverse)
library(knitr)
```

```r
# Reading the data
combined_2 <- read.csv("combined.csv")

colnames(combined_2)
```

```
##  [1] "DBN"
##  [2] "school_name"
##  [3] "Num.of.SAT.Test.Takers"
##  [4] "SAT.Critical.Reading.Avg..Score"
##  [5] "SAT.Math.Avg..Score"
##  [6] "SAT.Writing.Avg..Score"
##  [7] "avg_sat_score"
##  [8] "AP.Test.Takers"
##  [9] "Total.Exams.Taken"
## [10] "Number.of.Exams.with.scores.3.4.or.5"
## [11] "exams_per_student"
## [12] "high_score_percent"
## [13] "avg_class_size"
## [14] "frl_percent"
## [15] "total_enrollment"
## [16] "ell_percent"
## [17] "sped_percent"
## [18] "selfcontained_num"
## [19] "asian_per"
## [20] "black_per"
## [21] "hispanic_per"
## [22] "white_per"
## [23] "male_per"
## [24] "female_per"
## [25] "Total.Cohort"
## [26] "grads_percent"
## [27] "dropout_percent"
## [28] "boro"
## [29] "lat"
## [30] "long"
```

Unfortunately, the Dataquest web site with the `combined.csv` file does not explain the meaning of the variable names. Due to this, we will derive the meaning based on the variable names.

By inspecting the column names of the dataframe `combined_2`, we note that the column with school identification codes is named "DBN". In the `combined_1`, this column is named `dbn`. To combine the above two dataframes into one by matching their school identification codes, we rename the "DBN" column in dataframe `combined_2` to "dbn".

```r
colnames(combined_2)[[1]] <-"dbn"
```

We also notice that both `combined_1` and `combined_2` dataframes contain a column with school names. Upon a full join of these tibbles, this column would be duplicated. Therefore, we remove column `schoolname` from `combined_1` dataframe. We do keep the school code in `dbn` column. We will use this code to match the rows in `combined_1` and `combined_2` dataframes.

```r
combined_1 <- combined_1 %>% select(-schoolname)
```

Next, we combine dataframes `combined_1` and `combined_2` into a single dataframe by performing a full join.

```
combined_full <- combined_2 %>%
left_join(combined_1, by = "dbn")
```

Finally, we filter dataframe `combined_full` to retain only the rows that contain information for high schools.

```
combined_full <- combined_full %>% filter(str_detect(schooltype,"High School"))
```

4 Basic Properties of the Dataframe

In this section, we learn the basic properties of the `combined_full` dataframe that was created in the previous section. We begin with finding out the number of rows and columns as well as the names and the datatypes of the columns.

```
# Learning basic properties of dataframe combined_1
print("A glimpse of the combined_full dataframe:")
glimpse(combined_full)
```

```
## [1] "A glimpse of the combined_full dataframe:"
## Rows: 448
## Columns: 47
## $ dbn                                 <chr> "01M292", "01M448", "01M450", "01~
## $ school_name                         <chr> "HENRY STREET SCHOOL FOR INTERNAT~
## $ Num.of.SAT.Test.Takers              <int> 29, 91, 70, 7, 44, 112, 159, 18, ~
## $ SAT.Critical.Reading.Avg..Score     <int> 355, 383, 377, 414, 390, 332, 522~
## $ SAT.Math.Avg..Score                 <int> 404, 423, 402, 401, 433, 557, 574~
## $ SAT.Writing.Avg..Score              <int> 363, 366, 370, 359, 384, 316, 525~
## $ avg_sat_score                       <int> 1122, 1172, 1149, 1174, 1207, 120~
## $ AP.Test.Takers                      <dbl> 2.5, 39.0, 19.0, 2.5, 2.5, 24.0, ~
## $ Total.Exams.Taken                   <int> NA, 49, 21, NA, NA, 26, 377, NA, ~
## $ Number.of.Exams.with.scores.3.4.or.5 <int> NA, 10, NA, NA, NA, 24, 191, NA, ~
## $ exams_per_student                   <dbl> NA, 1.256410, 1.105263, NA, NA, 1~
## $ high_score_percent                  <dbl> NA, 20.408163, NA, NA, NA, 92.307~
## $ avg_class_size                      <int> 23, 22, 21, 23, 24, 23, 26, 22, 2~
## $ frl_percent                         <dbl> 88.6, 71.8, 71.8, 72.8, 80.7, NA,~
## $ total_enrollment                    <int> 422, 394, 598, 224, 367, NA, 1613~
## $ ell_percent                         <dbl> 22.3, 21.1, 5.0, 4.0, 11.2, NA, 0~
## $ sped_percent                        <dbl> 24.9, 21.8, 26.4, 8.9, 25.9, NA, ~
## $ selfcontained_num                   <int> 35, 10, 19, 0, 36, NA, 0, 0, 0, 1~
## $ asian_per                           <dbl> 14.0, 29.2, 9.7, 2.2, 9.3, NA, 27~
## $ black_per                           <dbl> 29.1, 22.6, 23.9, 34.4, 31.6, NA,~
## $ hispanic_per                        <dbl> 53.8, 45.9, 55.4, 59.4, 56.9, NA,~
## $ white_per                           <dbl> 1.7, 2.3, 10.4, 3.6, 1.6, NA, 44.~
## $ male_per                            <dbl> 61.4, 57.4, 54.7, 43.3, 46.3, NA,~
## $ female_per                          <dbl> 38.6, 42.6, 45.3, 56.7, 53.7, NA,~
## $ Total.Cohort                        <int> 78, 124, 90, NA, 84, 193, 46, 89,~
## $ grads_percent                       <dbl> 55.1, 42.7, 77.8, NA, 56.0, 54.4,~
## $ dropout_percent                     <dbl> 14.1, 16.1, 5.6, NA, 6.0, 18.1, 0~
## $ boro                                <chr> "Manhattan", "Manhattan", "Manhat~
## $ lat                                 <dbl> 40.71376, 40.71233, 40.72978, NA,~
## $ long                                <dbl> -73.98526, -73.98480, -73.98304, ~
## $ schooltype                          <chr> "Middle / High School", "High Sch~
## $ saf_p_11                            <dbl> 7.8, 7.9, 8.7, 8.1, 7.7, 8.3, 8.5~
```

```
## $ com_p_11                          <dbl> 7.7, 7.4, 8.2, 7.0, 7.4, 7.2, 7.9~
## $ eng_p_11                          <dbl> 7.4, 7.2, 8.1, 6.7, 7.2, 7.4, 7.9~
## $ aca_p_11                          <dbl> 7.6, 7.3, 8.4, 7.6, 7.3, 7.5, 8.4~
## $ saf_t_11                          <dbl> 6.3, 6.6, 7.3, 8.5, 6.4, 9.1, 7.6~
## $ com_t_11                          <dbl> 5.3, 5.8, 8.0, 8.2, 5.3, 7.3, 5.6~
## $ eng_t_11                          <dbl> 6.1, 6.6, 8.0, 8.9, 6.1, 8.7, 5.9~
## $ aca_t_11                          <dbl> 6.5, 7.3, 8.8, 8.9, 6.8, 9.1, 7.3~
## $ saf_s_11                          <dbl> 6.0, 6.0, NA, 6.8, 6.4, 8.0, 7.3,~
## $ com_s_11                          <dbl> 5.6, 5.7, NA, 6.1, 5.9, 6.3, 6.4,~
## $ eng_s_11                          <dbl> 6.1, 6.3, NA, 6.1, 6.4, 7.0, 7.0,~
## $ aca_s_11                          <dbl> 6.7, 7.0, NA, 6.8, 7.0, 7.3, 7.7,~
## $ saf_tot_11                        <dbl> 6.7, 6.8, 7.9, 7.8, 6.9, 8.5, 7.8~
## $ com_tot_11                        <dbl> 6.2, 6.3, 7.9, 7.1, 6.2, 7.0, 6.7~
## $ eng_tot_11                        <dbl> 6.6, 6.7, 7.9, 7.2, 6.6, 7.7, 6.9~
## $ aca_tot_11                        <dbl> 7.0, 7.2, 8.4, 7.8, 7.0, 8.0, 7.8~
```

5 Data Analysis

In this section we analyse the data in the `combined_full` dataframe to answer the questions posed in the introduction.

5.1 Correlation Analysis

In this subsection we compute the correlation matrix with correlations between all the numerical variables in tibble `combined_full`. Once we detect pairs of variables with stronger correlations, we will use scatter plots to inspect these relationships closer. We do not delete rows with `NA` entries. Instead, we compute correlations using pairwise complete observations. That is, correlations for each pair of variables are computed using observations with no missing data for that pair. We convert the correlation matrix into a tibble. For each row, we filter the columns of the tibble to retain only the correlations with absolute values higher than 0.25.This is done to determine which pairs of variables demonstrate significant positive or negative correlations.

As a reminder, if two variables have a significant positive correlation, then larger values of the first variable are associated with larger values of the second variable. In the case of significant negative correlation, the relationship is inverse, that is, larger values of the first variable are associated with smaller values of the second variable.

5.1.1 Computing and Plotting the Correlation Matrix

We begin with computing the correlation matrix.

```
# Computing the correlation matrix
cor_mat <- combined_full %>%
select(where(is.numeric)) %>%
cor(use = "pairwise.complete.obs")

# Converting the correlation matrix into a tibble
cor_tib <- cor_mat %>%
as_tibble(rownames = "variable")
```

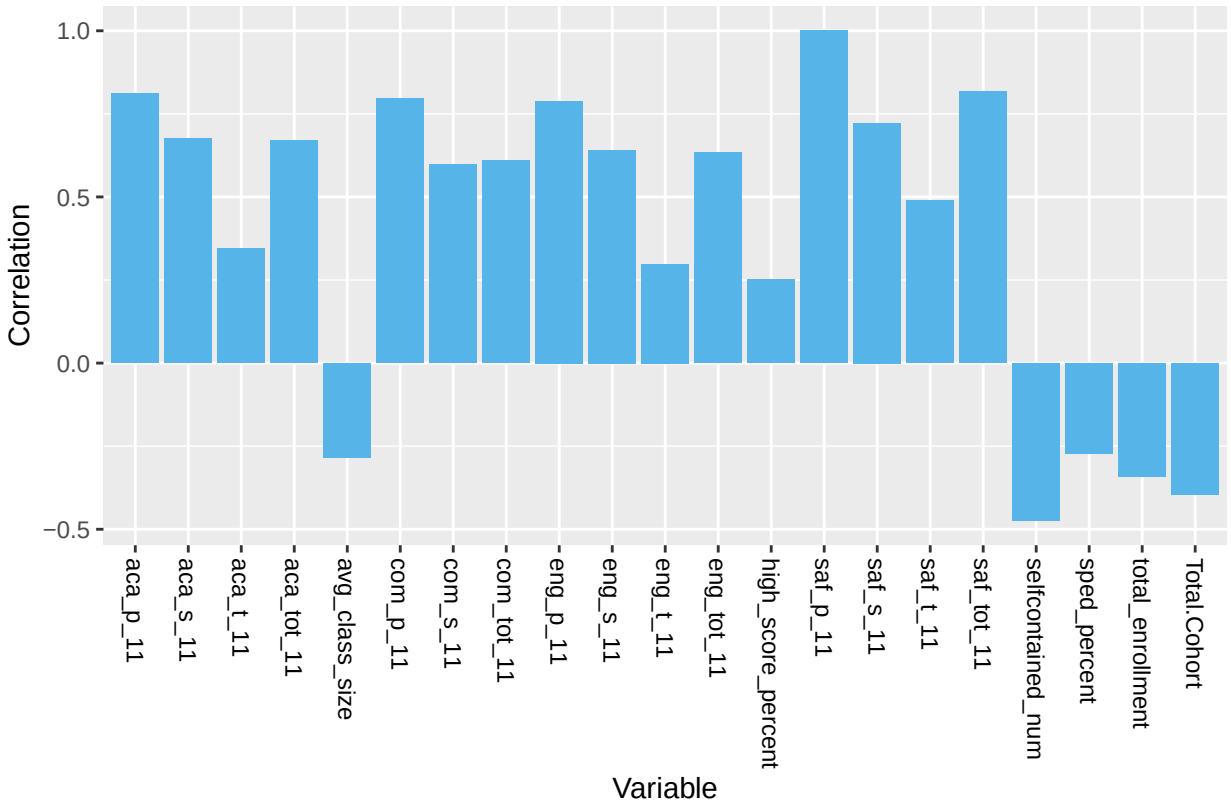
Next, we create function `cor_plot_strong_cor` that, given a column number, extracts significant (absolute value larger than 0.25) correlations from that column and plots those correlations using a bar chart.

```
cor_plot_strong_cor <- function(.data,col_number){

  # Selecting the correlations of the variable with
```

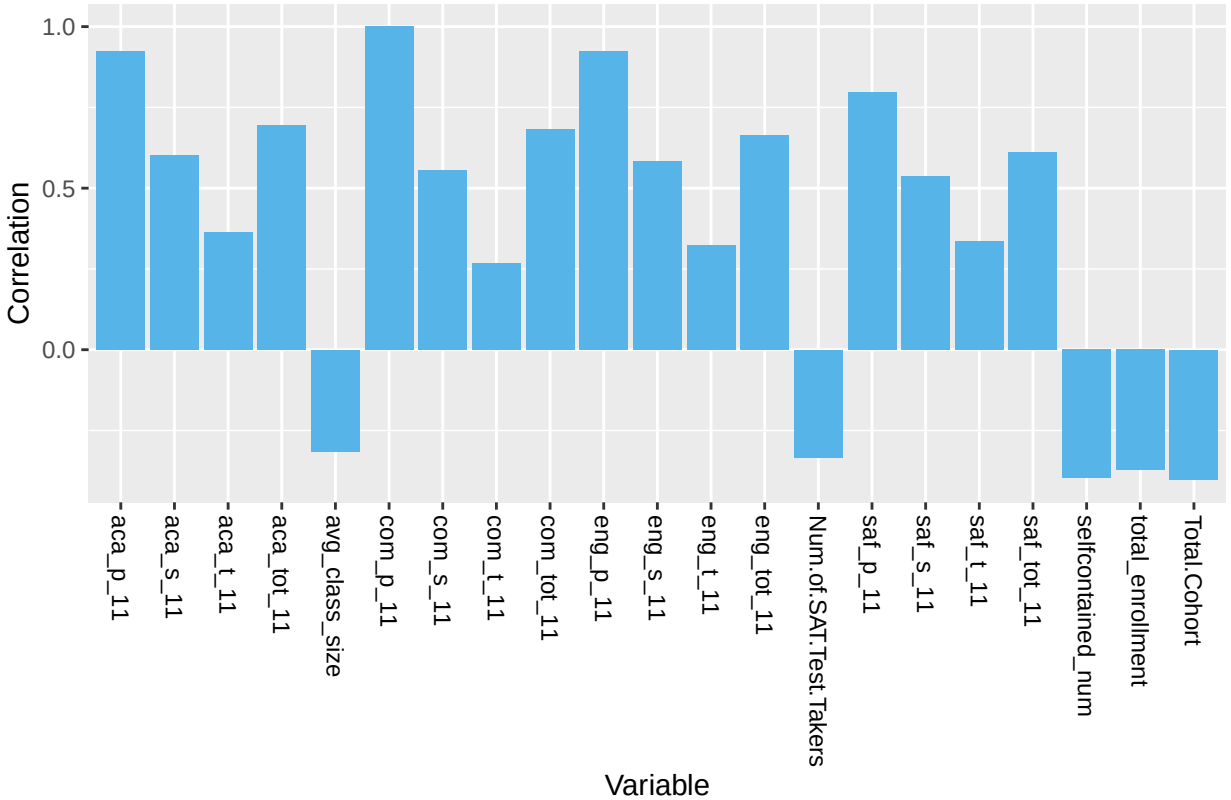
```
  # the number col_number
  # with other variables, filtering only those
  # correlations that are >0.25 in absolute value

  signif_cors <- .data[c(1,col_number)]
  col2_name<-colnames(signif_cors)[2]
  colnames(signif_cors)<-c("col1","col2")
  signif_cors$col2 <- as.numeric(signif_cors$col2)
  signif_cors <- signif_cors %>% filter(abs(col2) > 0.25 )


  # Plotting the last correlations using a bar chart
  plot<-signif_cors %>% ggplot(aes(x=col1,y=col2))+
  geom_bar(position='dodge',stat="identity",fill="#56B4E9")+
  labs(caption = paste("Significant correlations of ",
    col2_name," with other variables"),x="Variable",
    y="Correlation")+
  theme(axis.text.x = element_text(size=9, angle = -90,
  vjust = 0.5, hjust = 0, color = "black") )

  print(plot)
}
```

Next, we plot all the columns of the correlation matrix, with only significant correlations extracted. To this end, we apply the above function `cor_plot_strong_cor` to columns 29 through 44.

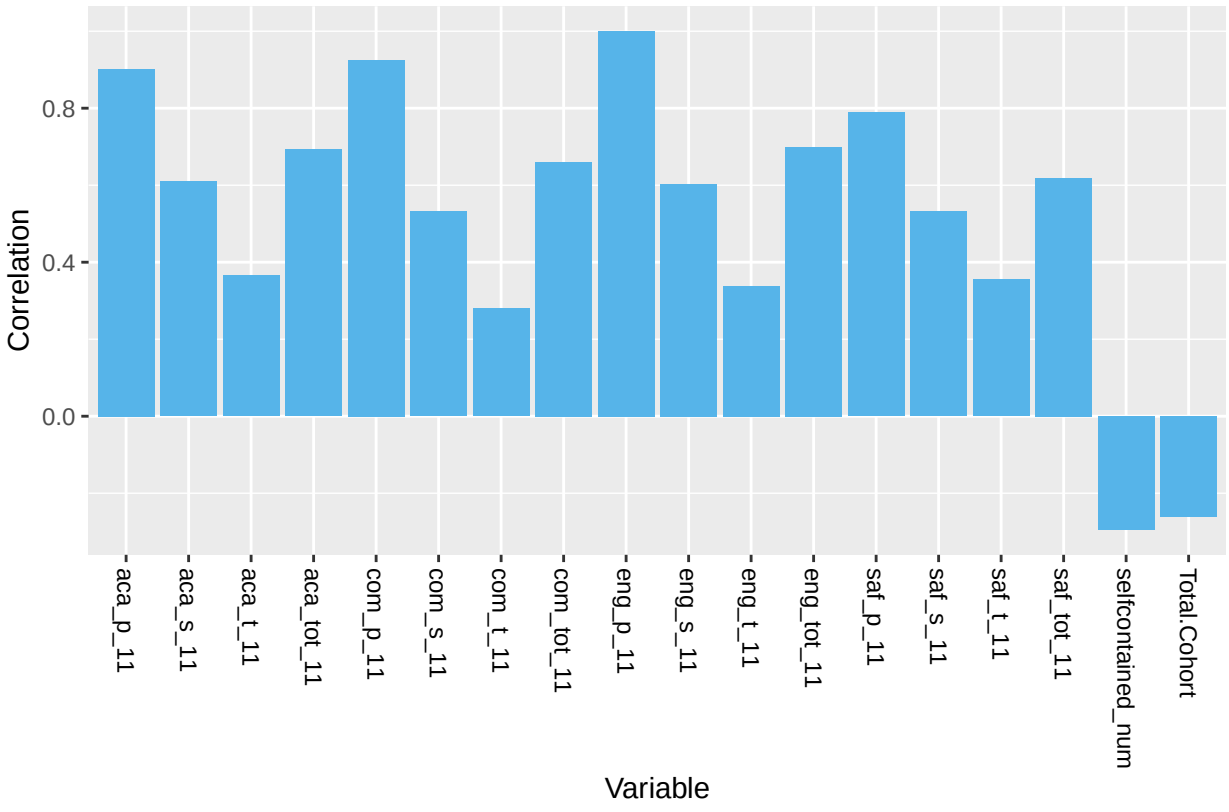
```
#Printing all the correlation bar charts using a loop
for(i in 29:44){
  #Printing the i-th bar chart
  cor_plot_strong_cor(cor_tib,i)
  #Printing a vertical space
  cat("  \n")
  cat("  \n")
  }
```



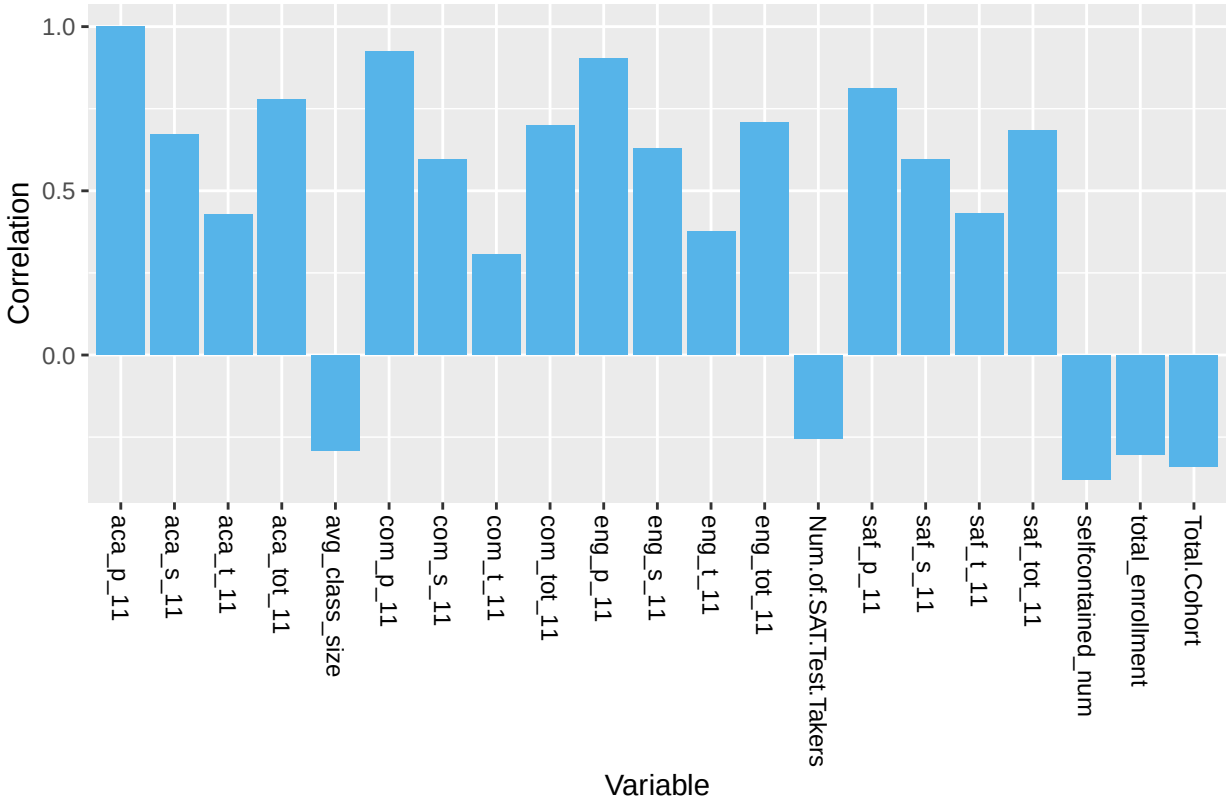
Significant correlations of saf_p_11 with other variables



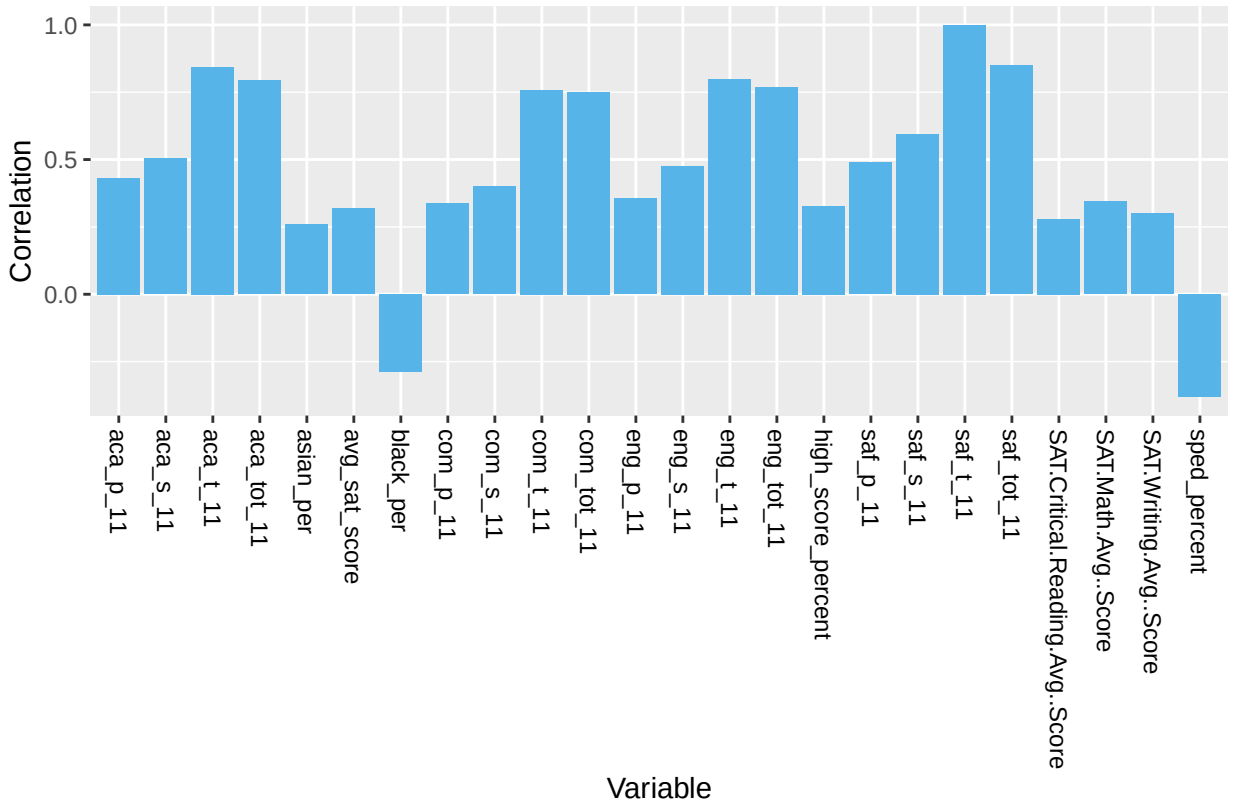
Significant correlations of com_p_11 with other variables



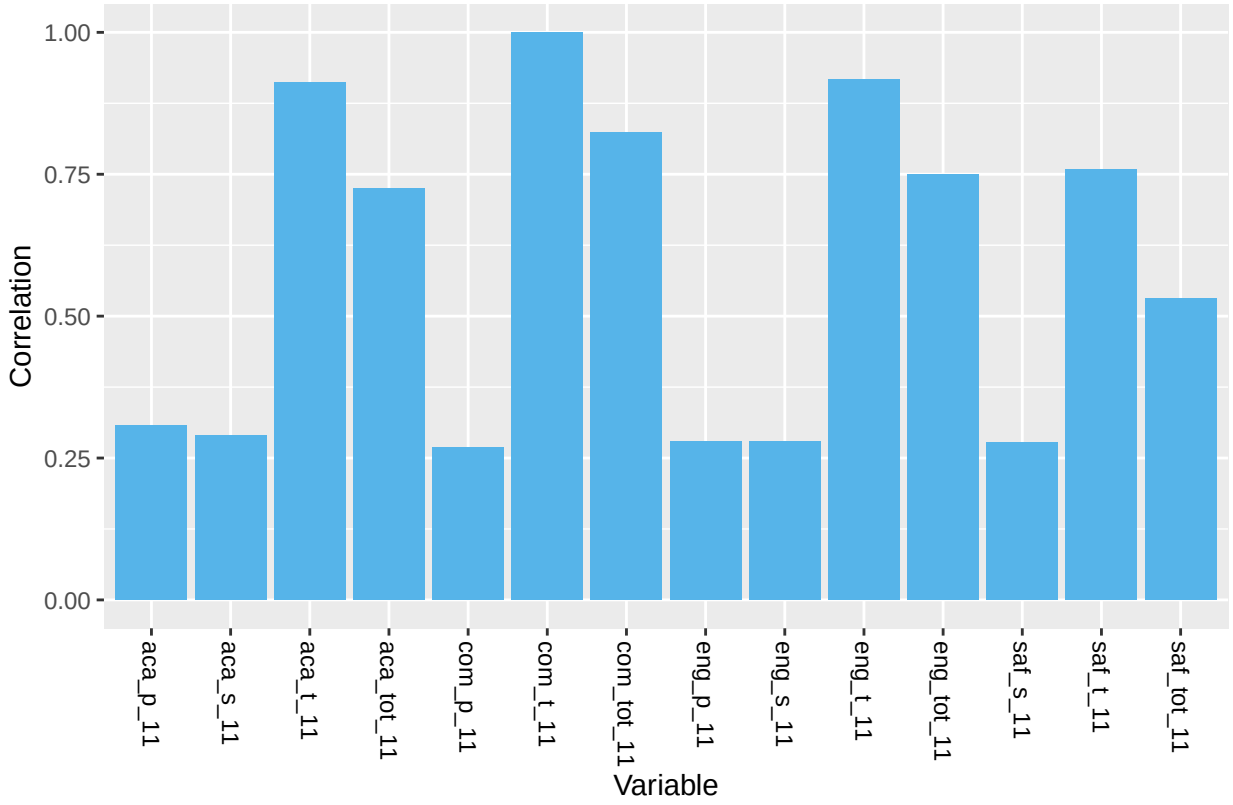
Significant correlations of eng_p_11 with other variables



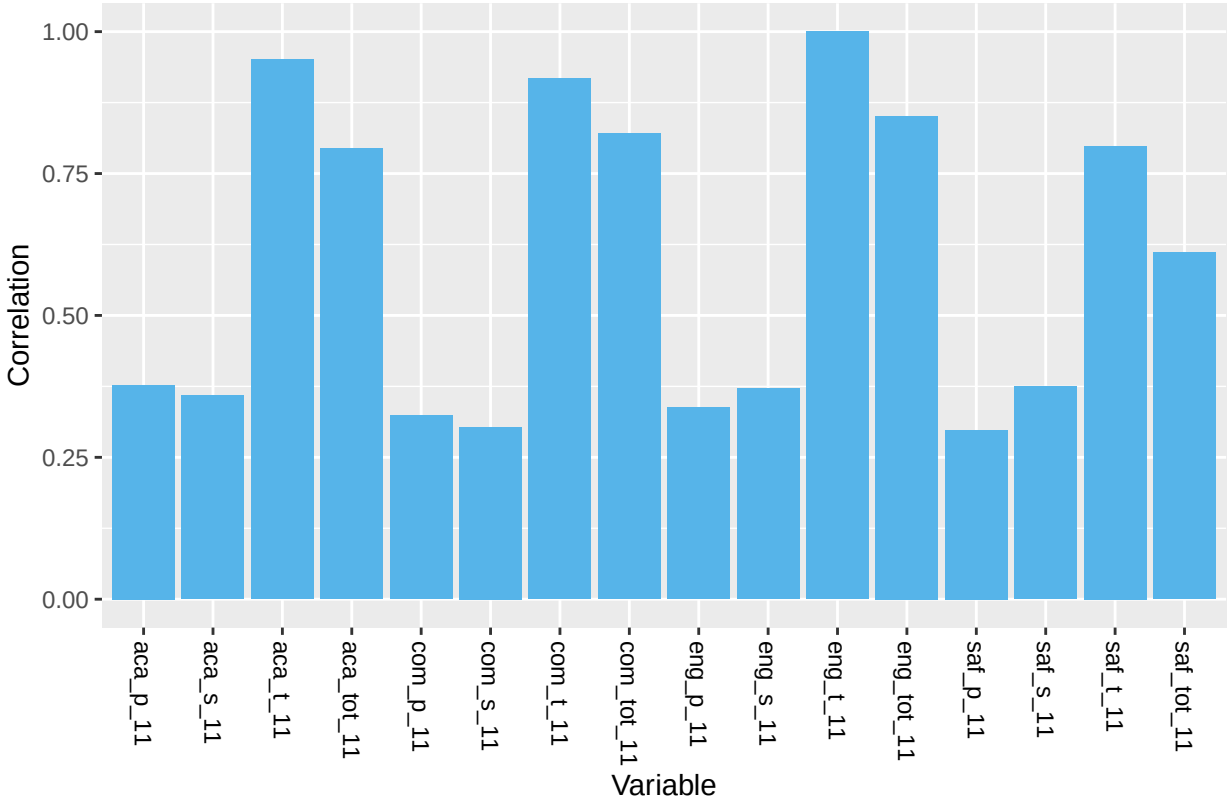
Significant correlations of aca_p_11 with other variables



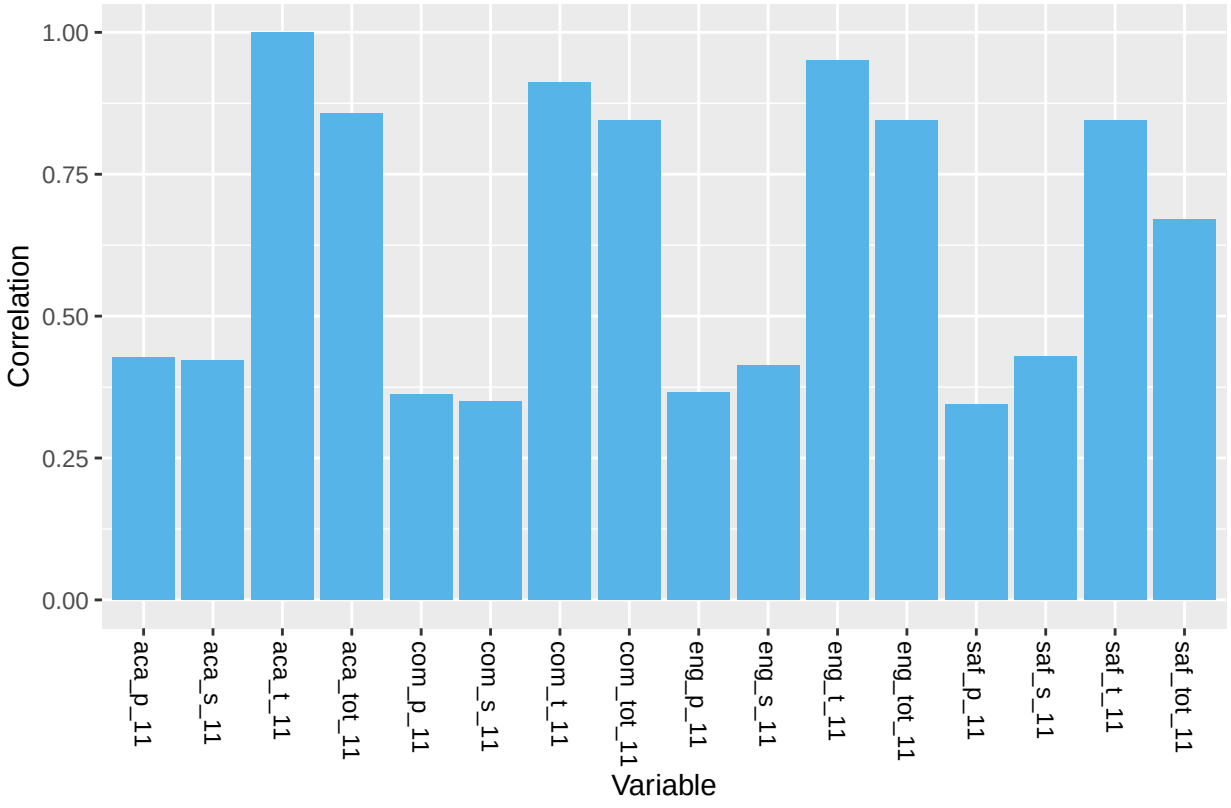
Significant correlations of saf_t_11 with other variables



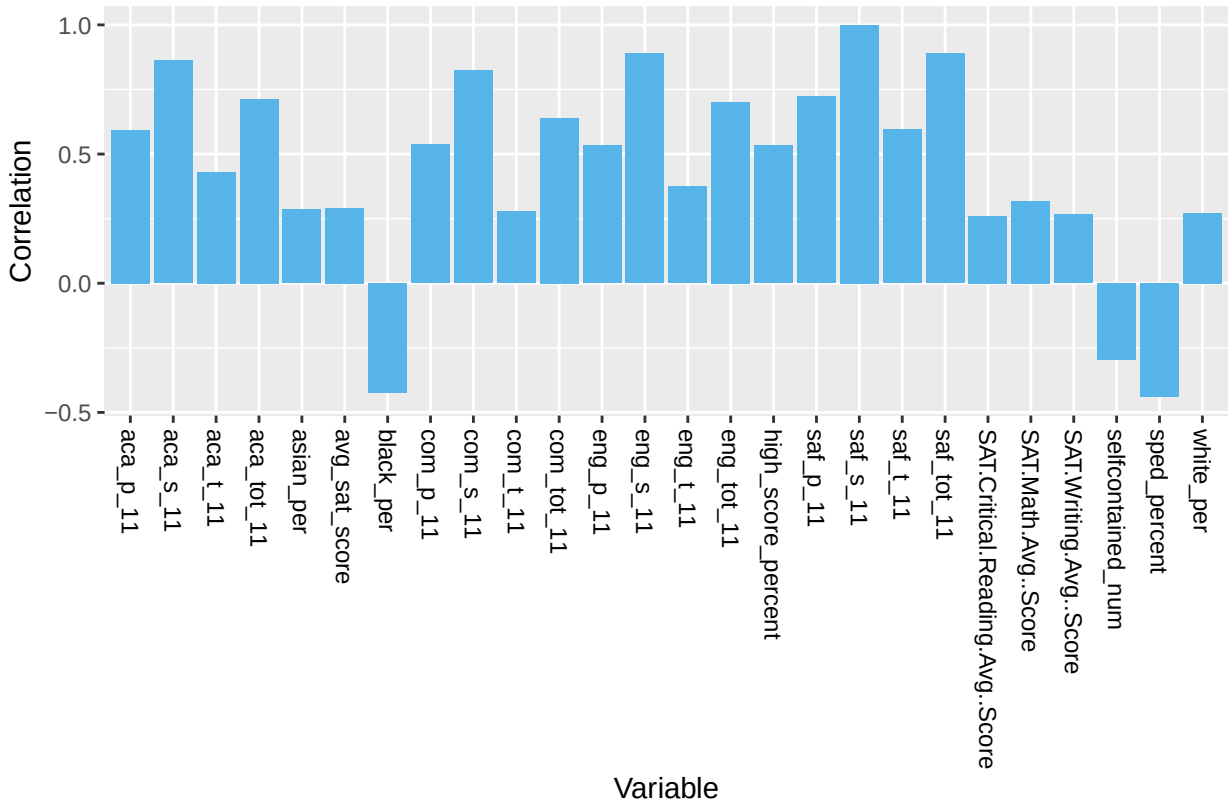
Significant correlations of com_t_11 with other variables



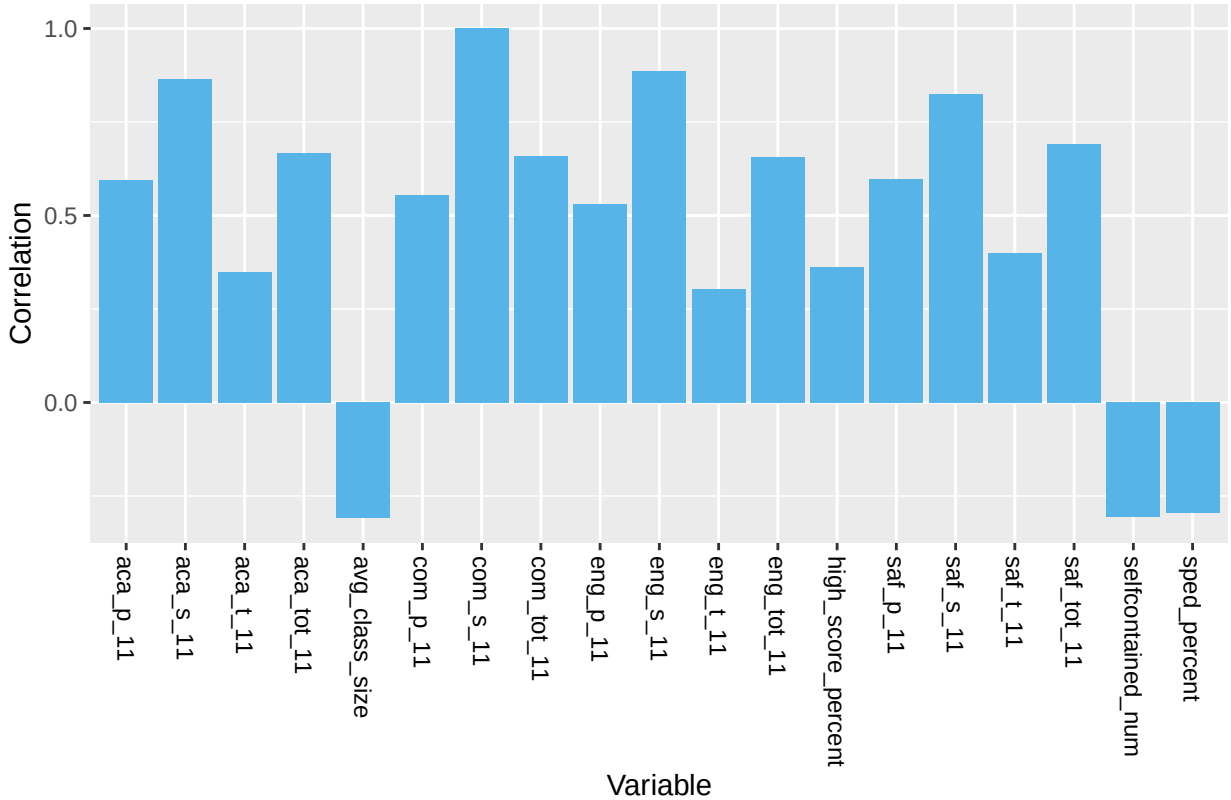
Significant correlations of eng_t_11 with other variables



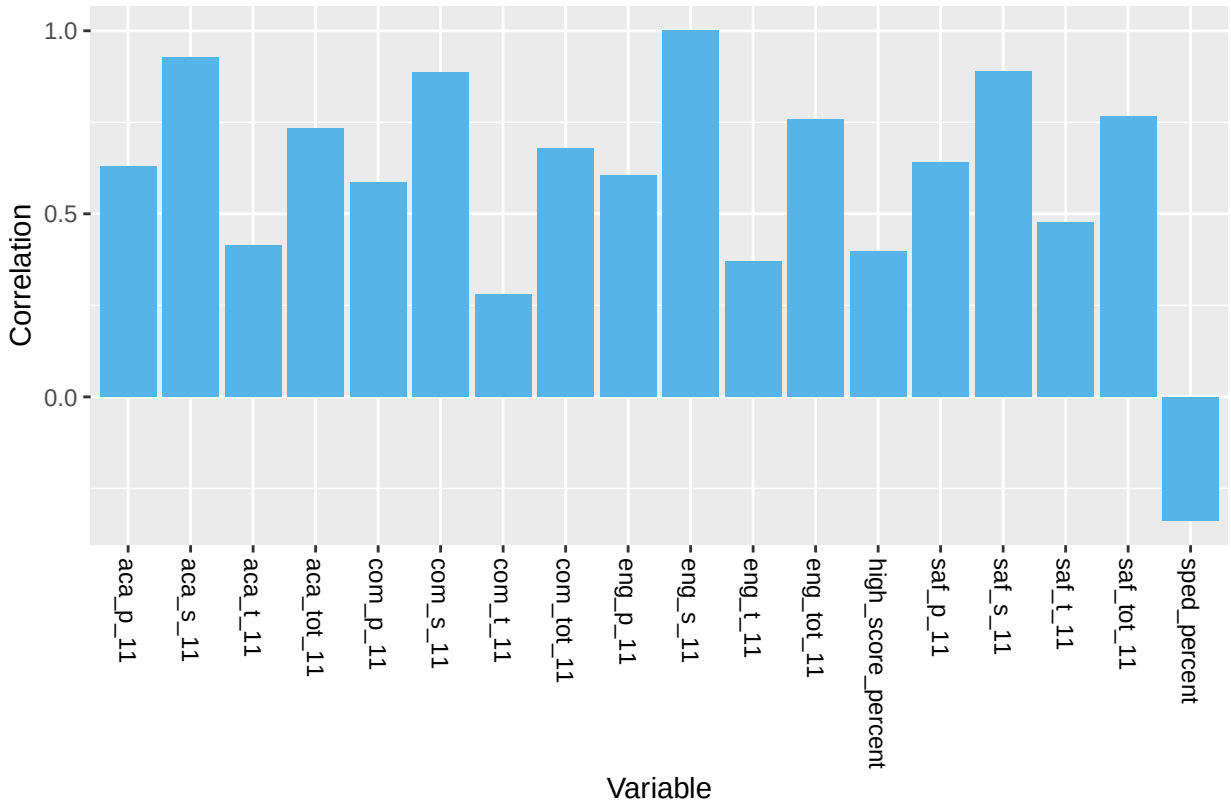
Significant correlations of aca_t_11 with other variables



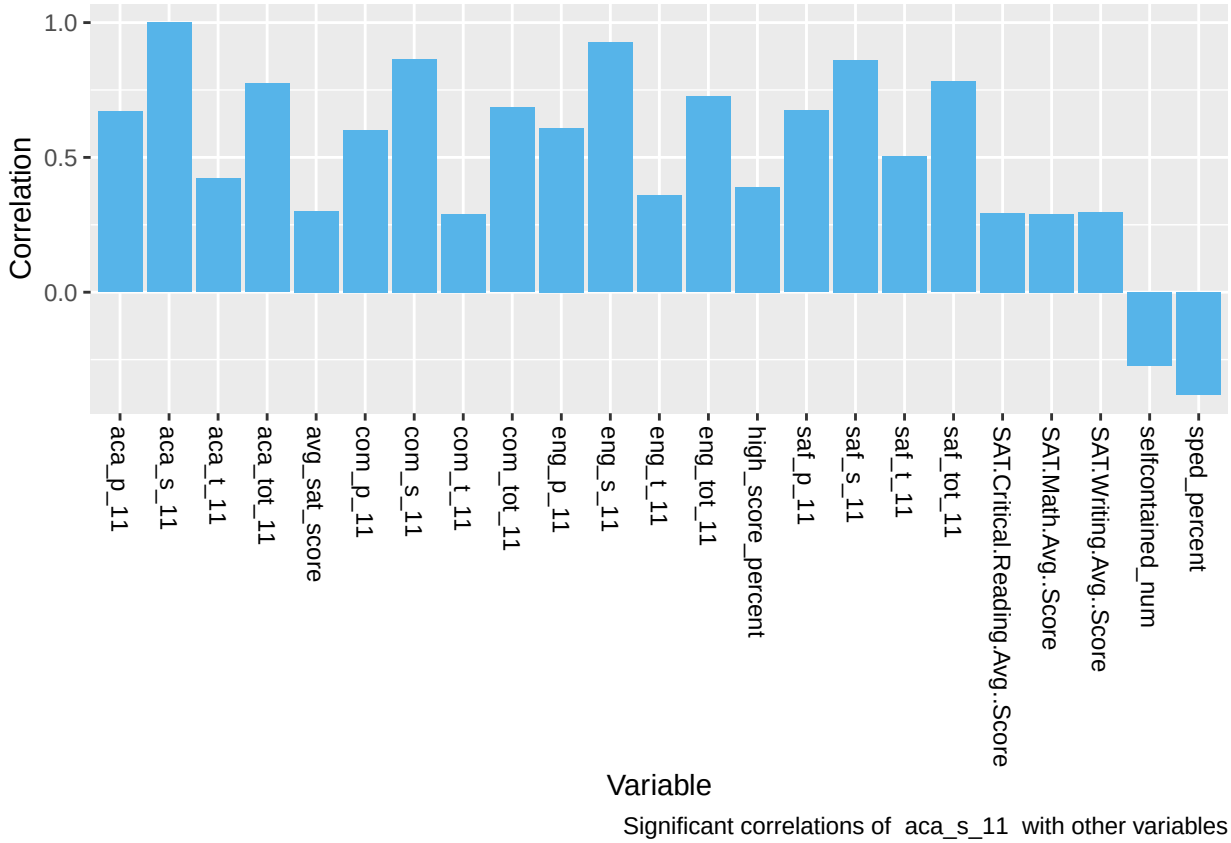
Significant correlations of saf_s_11 with other variables



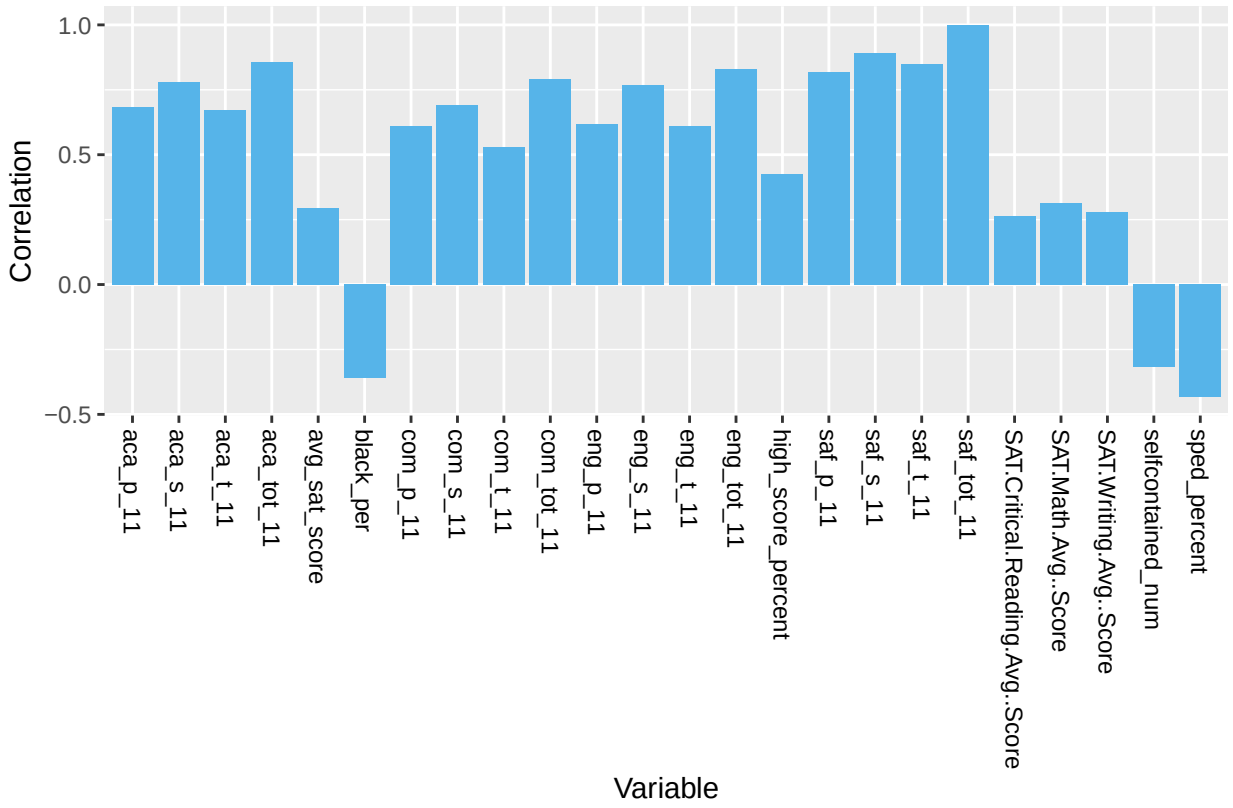
Significant correlations of com_s_11 with other variables



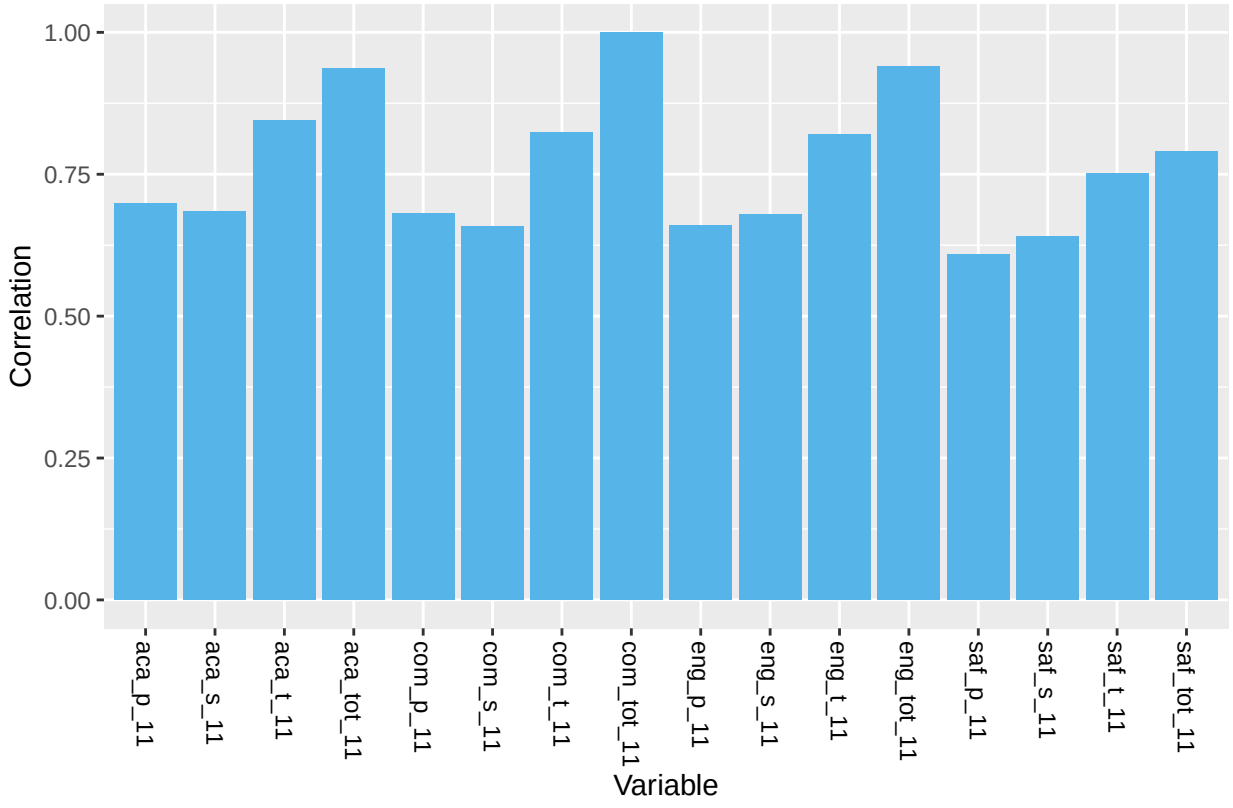
Significant correlations of eng_s_11 with other variables



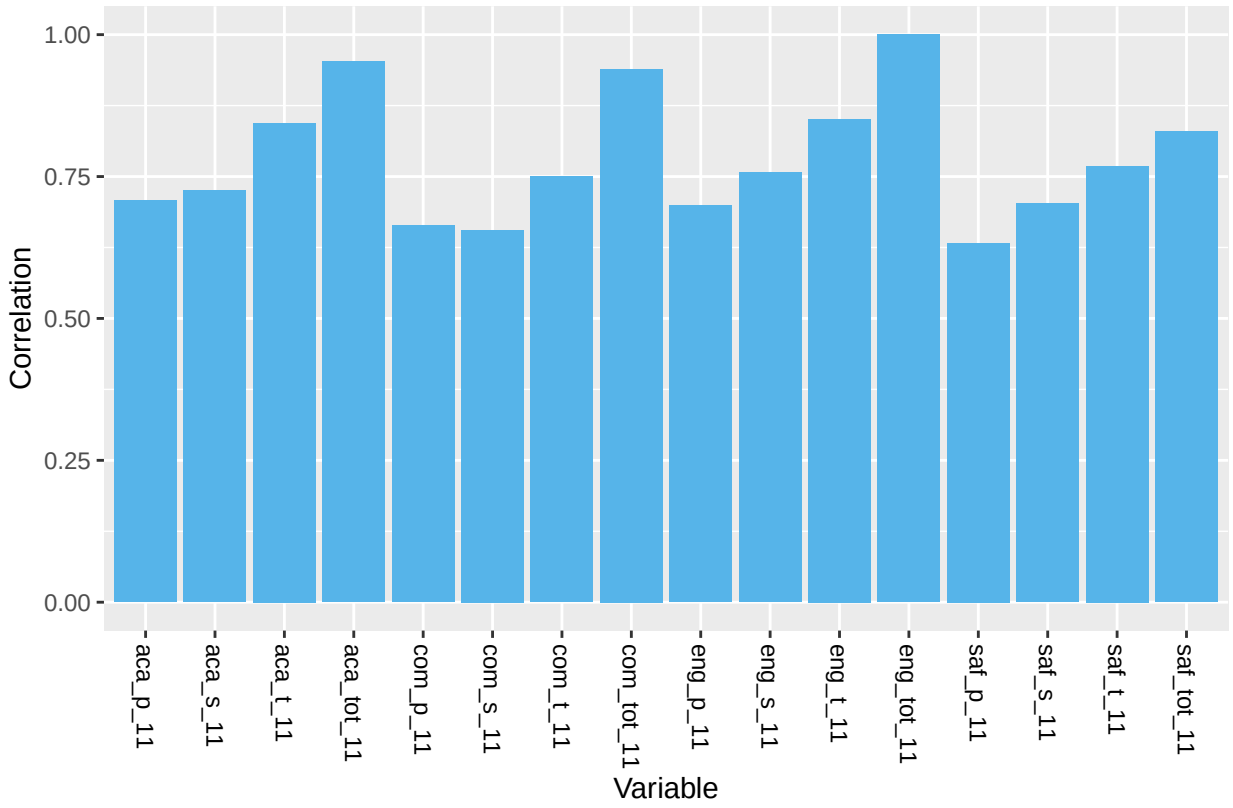
Significant correlations of aca_s_11 with other variables



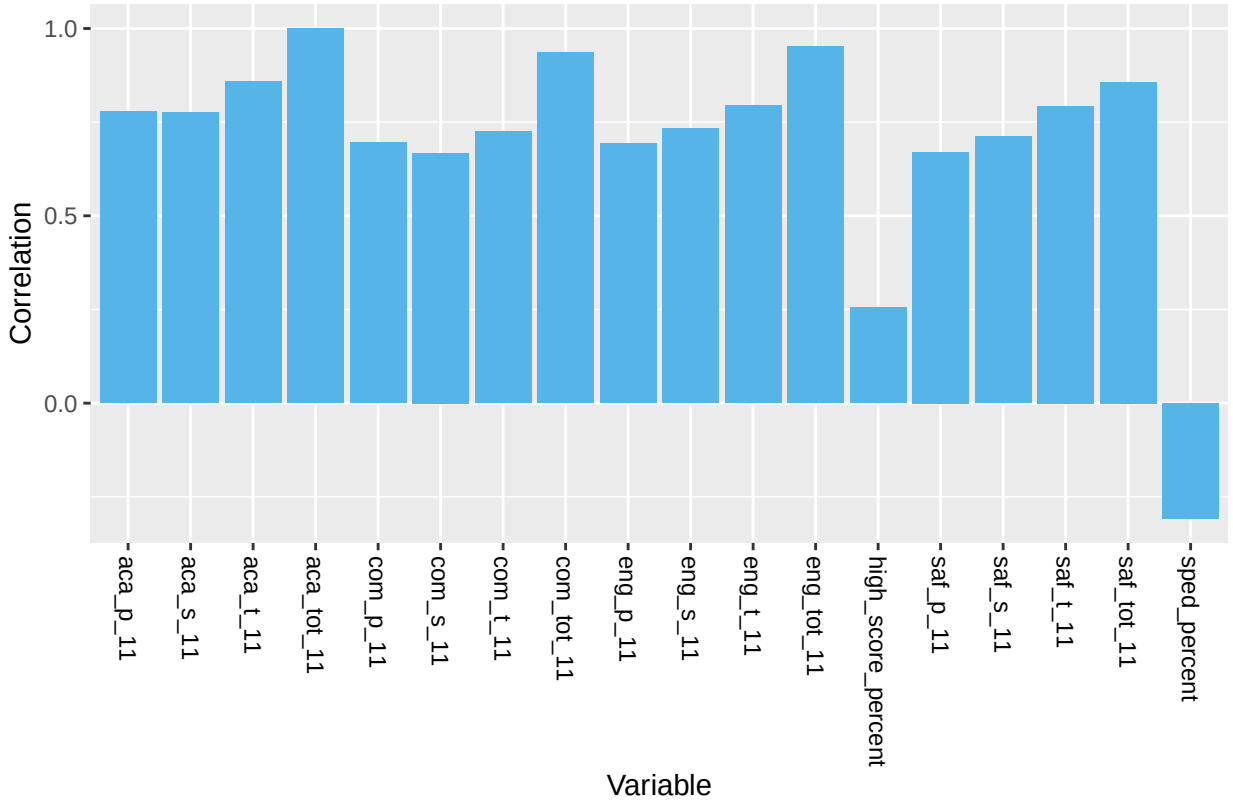
Significant correlations of saf_tot_11 with other variables



Significant correlations of com_tot_11 with other variables



Significant correlations of eng_tot_11 with other variables



Significant correlations of aca_tot_11 with other variables

5.1.2 Conclusions for Parents' Responses

In this subsection, we analyse the correlation bar charts for parents' responses (saf_p_11, com_p_11, eng_p_11, and aca_p_11) in the plots of Section 5.1.1. From the bar chart that displays the correlations for the saf_p_11 variable, we conclude the following. Each among the higher class sizes, higher percentages of students enrolled in special education, and higher total enrollments of the school are associated with lower parent satisfaction with the safety and respect in the school. Higher percentages of high SAT scores are associated with higher parent satisfaction with the safety and respect in the school. Higher safety and respect scores by parents are associated with higher survey scores by all three cohorts (parents, teachers, and students), with the exception of the teachers' communication score. Parents' safety and respect scores appear to be unrelated to teachers' communication scores.

From the bar chart that displays the correlations for the com_p_11 variable, we conclude the following. Each among the higher class sizes, higher number of SAT test takers, and higher total enrollments of the school are associated with lower parent satisfaction with the communication in the school. Higher parents' communication score are associated with higher survey scores by all three cohorts.

From the bar chart that displays the correlations for the eng_p_11 variable, we conclude the following. A higher parents' engagement score is associated with higher survey scores by all three cohorts.

From the bar chart that displays the correlations for the aca_p_11 variable, we conclude the following. Each among the higher class sizes, higher number of SAT test takers, and higher total enrollments of the school are associated with lower parent satisfaction with the academic expectations in the school. A higher parents' academic expectations score is associated with higher survey scores by all three cohorts.

5.1.3 Conclusions for Teachers' Responses

In this subsection, we analyse the correlation bar charts for teachers' responses (saf_t_11, com_t_11, eng_t_11, and aca_t_11) in the plots of Section 5.1.1. From the bar chart that displays the correlations for the saf_t_11 variable, we conclude the following. Each among the higher black students percentage and higher percentages of students enrolled in special education are associated with lower teacher satisfaction with the safety and respect in the school. Higher average SAT, critical SAT reading, math, and writing scores are associated with higher teacher satisfaction with the safety and respect in the school. Higher percentages of Asian students and high SAT scores are also associated with higher teachers' safety and respect scores. Also, higher teachers' safety and respect scores are associated with higher survey scores by all three cohorts.

From the bar chart that displays the correlations for the com_t_11 variable, we conclude the following. Higher communication scores by teachers are associated with higher survey scores by all three cohorts, except for the communication scores by students and safety and respect scores by parents.

From the bar chart that displays the correlations for the eng_t_11 variable, we conclude the following. Higher engagement scores by teachers are associated with higher survey scores by all three cohorts.

From the bar chart that displays the correlations for the aca_p_11 variable, we conclude the following. Higher academic expectations scores by teachers are associated with higher survey scores by all three cohorts.

5.1.4 Conclusions for Students' Responses

In this subsection, we analyse the correlation bar charts for students' responses (saf_s_11, com_s_11, eng_s_11, and aca_s_11) in the plots of Section 5.1.1. From the bar chart that displays the correlations for the saf_s_11 we conclude the following. Each among the higher black students percentage and higher percentages of students enrolled in special education are associated with lower student satisfaction with the safety and respect in the school. Higher critical SAT reading, math, and writing scores are associated with higher student satisfaction with the safety and respect in the school. Higher percentages of white students and high SAT scores are also associated with higher students' safety and respect scores. Also, higher students' safety and respect scores are associated with higher survey scores by all three cohorts.

From the bar chart that displays the correlations for the com_s_11 variable, we conclude the following. Lower communication scores by students are associated with higher average class sizes and higher percentages of special education students. Higher communication scores by students are associated with higher percentages of high SAT scores. Higher communication scores by students are associated with higher survey scores by all three cohorts, except for the communication scores by teachers.

From the bar chart that displays the correlations for the `eng_s_11` variable, we conclude the following. Lower engagement scores by students are associated with higher percentages of special education students. Higher engagement scores by students are associated with higher percentages of high SAT scores. Also, higher engagement scores by students are associated with higher survey scores by all three cohorts.

From the bar chart that displays the correlations for the `aca_s_11` variable, we conclude the following. Lower academic expectation scores by students are associated with higher percentages of special education students. Higher academic expectation scores by students are associated with higher average SAT scores, higher percentages of high SAT scores, and higher average critical reading, math, and writing SAT scores. Also, higher academic expectations scores by students are associated with higher survey scores by all three cohorts.

5.1.5 Conclusions for Total Scores

In this subsection, we analyse the correlation bar charts for total scores (`saf_tot_11`, `com_tot_11`, `eng_tot_11`, and `aca_tot_11`) in the plots of Section 5.1.1. From the bar chart that displays the correlations for the `saf_tot_11` variable, we conclude the following. Each among the higher black students percentage and higher percentages of students enrolled in special education are associated with lower total satisfaction with the safety and respect in the school. Higher SAT average, critical reading, math, and writing scores are associated with higher total satisfaction with the safety and respect in the school. Also, higher total safety and respect scores are associated with higher survey scores by all three cohorts.

From the bar chart that displays the correlations for the `com_tot_11` variable, we conclude the following. Higher total communication scores are associated with higher survey scores by all three cohorts.

From the bar chart that displays the correlations for the `eng_tot_11` variable, we conclude the following. Higher total engagement are associated with higher survey scores by all three cohorts.

From the bar chart that displays the correlations for the `aca_tot_11` variable, we conclude the following. Lower total academic expectation scores are associated with higher percentages of special education students. Higher total academic expectation scores are associated with higher percentages of high SAT scores. Also, higher total academic expectations scores are associated with higher survey scores by all three cohorts.

Overall conclusions based on the analysis of the above correlation plots are stated in Section 2.

5.2 Further Analysis via Scatter Plots

In this section, we create scatter plots to analyze interesting relationships between variables further.

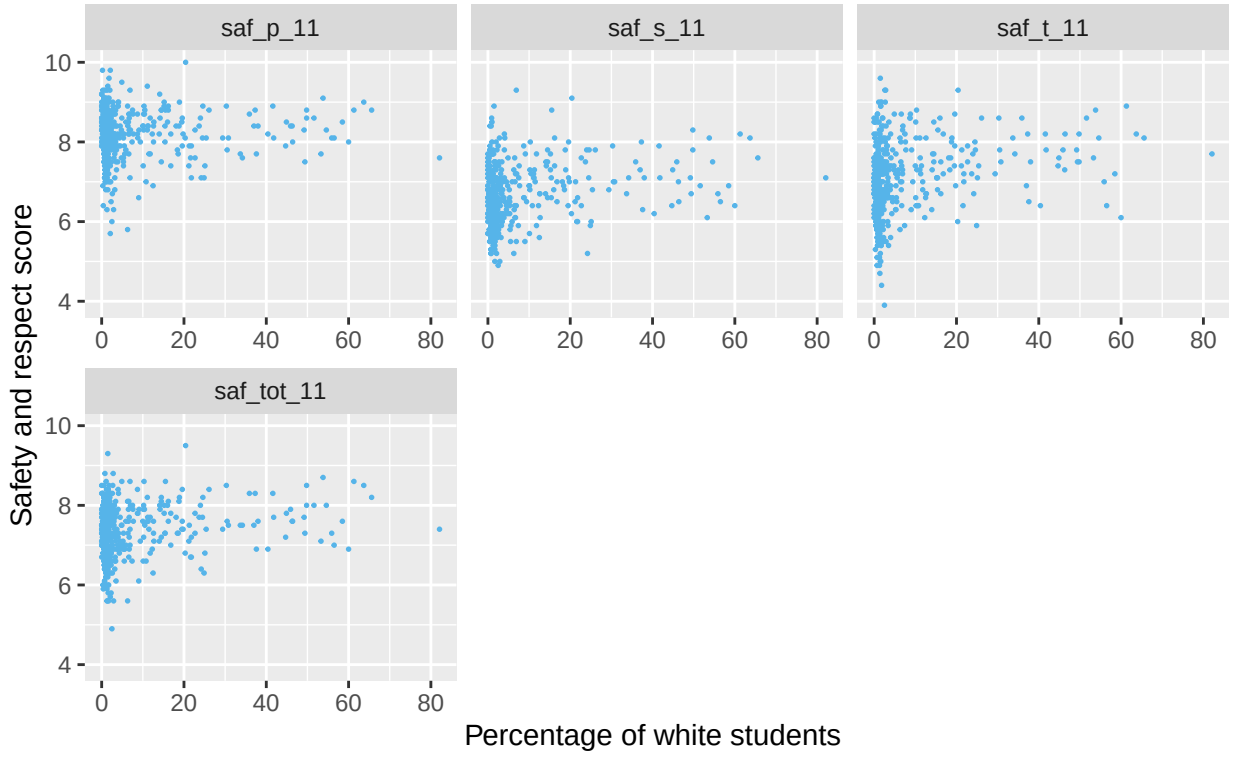
5.2.1 Relationships between Race and the Safety and Respect Scores

We begin with producing scatter plots that illustrate the relationship between the percentage of white students and the safety and respect score.

```
# Pivoting the table combined_full
combined_full_pivoted_saf <- combined_full %>% pivot_longer(cols=c(
    saf_p_11,
    saf_t_11,
    saf_s_11,
    saf_tot_11,
  ),
  names_to="saf_var",
  values_to='saf_score')
```

```
# Plotting the relationships between safety and respect
# scores and the percentage of white students

combined_full_pivoted_saf %>%
ggplot(aes(x= white_per, y=   saf_score))+
```

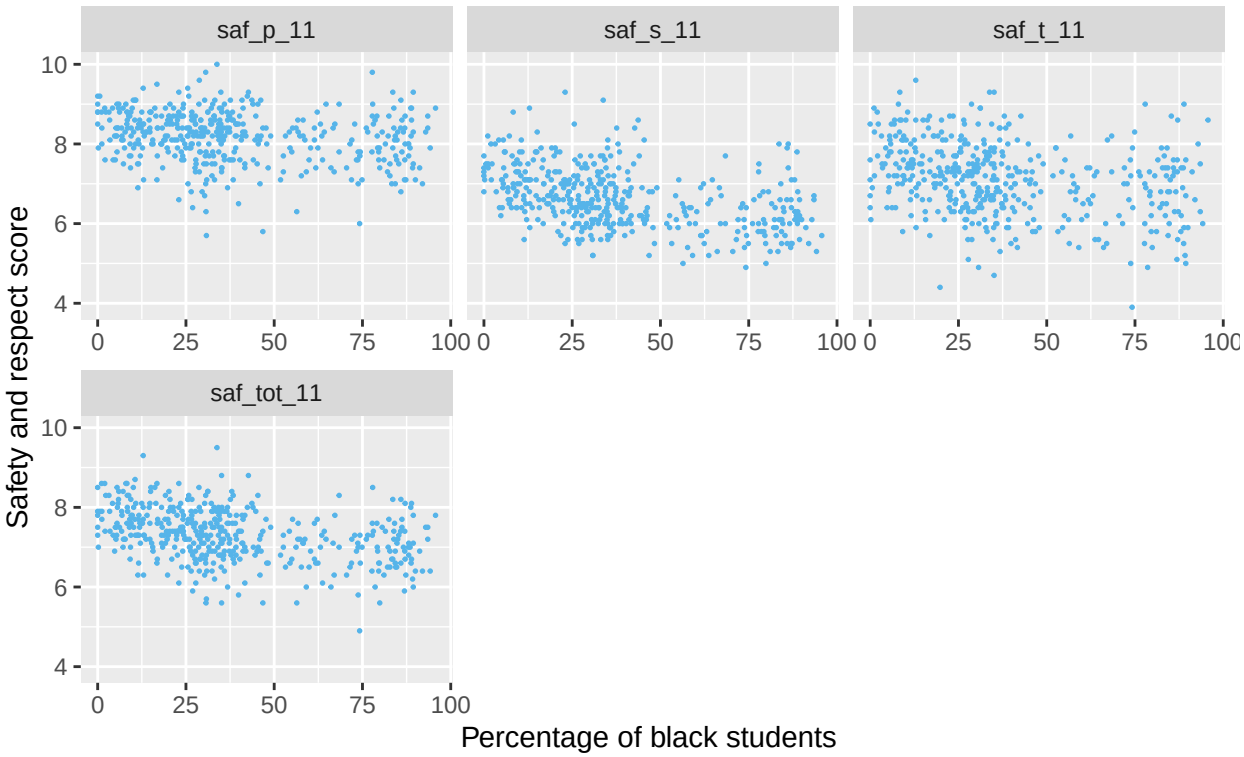
```
geom_point(color="#56B4E9",size=0.3)+
facet_wrap(ncol=3,vars(saf_var), scales = "free_x")+
#ylab("Safety and respect score") +
#xlab("Percentage of white students ") +
labs(caption = paste("The relationship between the percentage
  of white students and the safety and respect
  score."),x="Percentage of white students ",
  y="Safety and respect score")
```



The relationship between the percentage of white students and the safety and respect score.

By observing the last set of plots, we conclude that safety and respect scores between 7 and 9 by the parents occur at many schools with higher percentages of white students. Similar conclusions can be made for the safety and respect scores by teachers, students, and total.

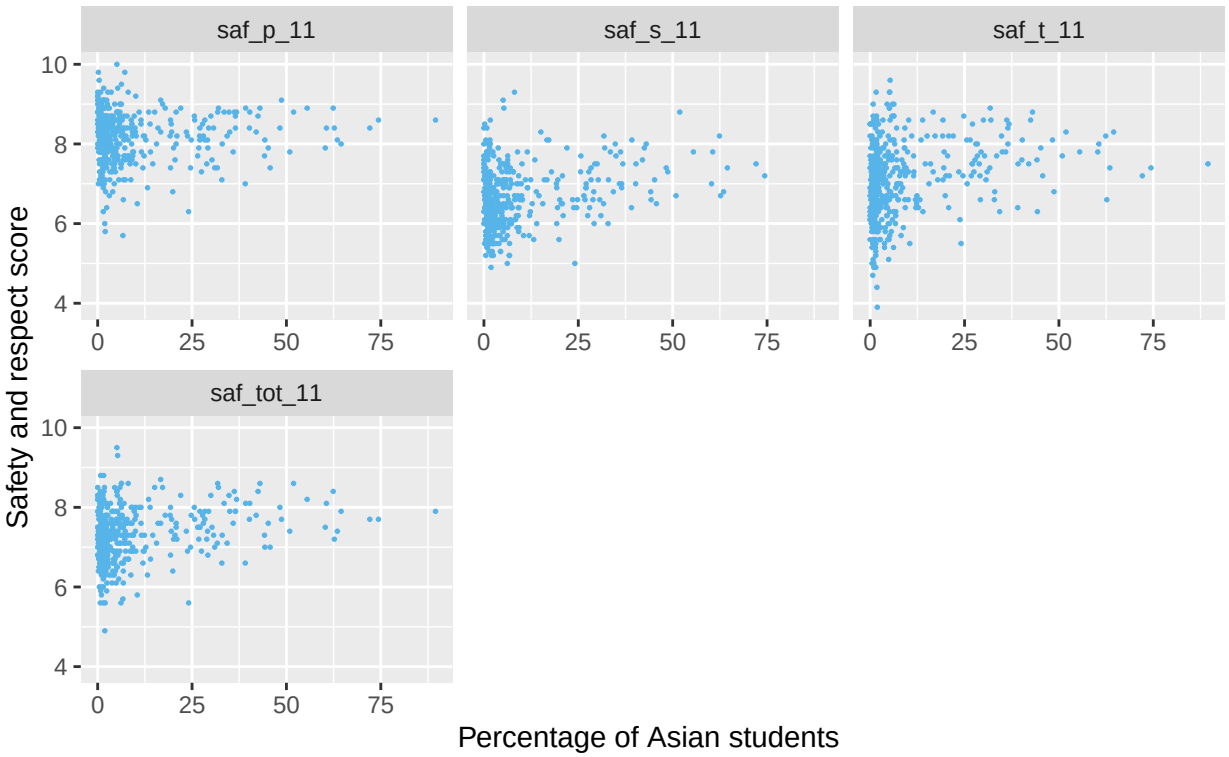
```
# Plotting the relationships between safety and respect
# scores and the percentage of black students

combined_full_pivoted_saf %>%
ggplot(aes(x= black_per, y=   saf_score))+
geom_point(color="#56B4E9",size=0.3)+
facet_wrap(ncol=3,vars(saf_var), scales = "free_x")+
#ylab("Safety and respect score") +
#xlab("Percentage of black students")
labs(caption = paste("The relationship between the percentage
    of black students and the safety and respect
    score."),x="Percentage of black students ",
    y="Safety and respect score")
```



The relationship between the percentage of black students and the safety and respect score.

By visual inspection of the last set of plots, it is difficult to see any relationship between the percentage of black students and the safety and respect score. Student safety and respect scores appear to be slightly lower in the schools with higher percentages of black students. From the correlation analysis, we know that the same holds for the teacher safety and respect scores.

```
# Plotting the relationships between safety and respect
# scores and the percentage of Asian students

combined_full_pivoted_saf %>%
ggplot(aes(x=asian_per, y=   saf_score))+
geom_point(color="#56B4E9",size=0.3)+
facet_wrap(ncol=3,vars(saf_var), scales = "free_x")+
#ylab("Safety and respect score") +
#xlab("Percentage of Asian students ")
labs(caption = paste("The relationship between the percentage
    of Asian students and the safety and respect
    score."),x="Percentage of Asian students ",
    y="Safety and respect score")
```



The relationship between the percentage of Asian students and the safety and respect score.

By observing the lats set of plots, we conclude that safety and respect scores between 7 and 9 by the parents occur at many schools with higher percentages of Asian students. Similar conclusions can be made for the safety and respect scores by teachers, students, and total.

We do not plot the relationships between the safety and respect scores and the percentages of Hispanic students because the correlation analysis indicates that there is no significant correlation between these variables.

5.2.2 Relationships between Average SAT Scores and Other Variables

In this subsection, we create scatter plots that illustrate the relationships between average SAT scores and the variables that have significant correlations with these scores.

We begin with filtering the correlation matrix so it includes only the significant (>0.25) correlations between the average SAT scores and other variables.

```
# Selecting the correlations of avg_sat_score with
# with other variables, filtering only those
# correlations that are >0.25 in absolute value

signif_cors <- cor_tib %>% select(variable,avg_sat_score) %>%
  filter(abs(avg_sat_score) > 0.25)
```

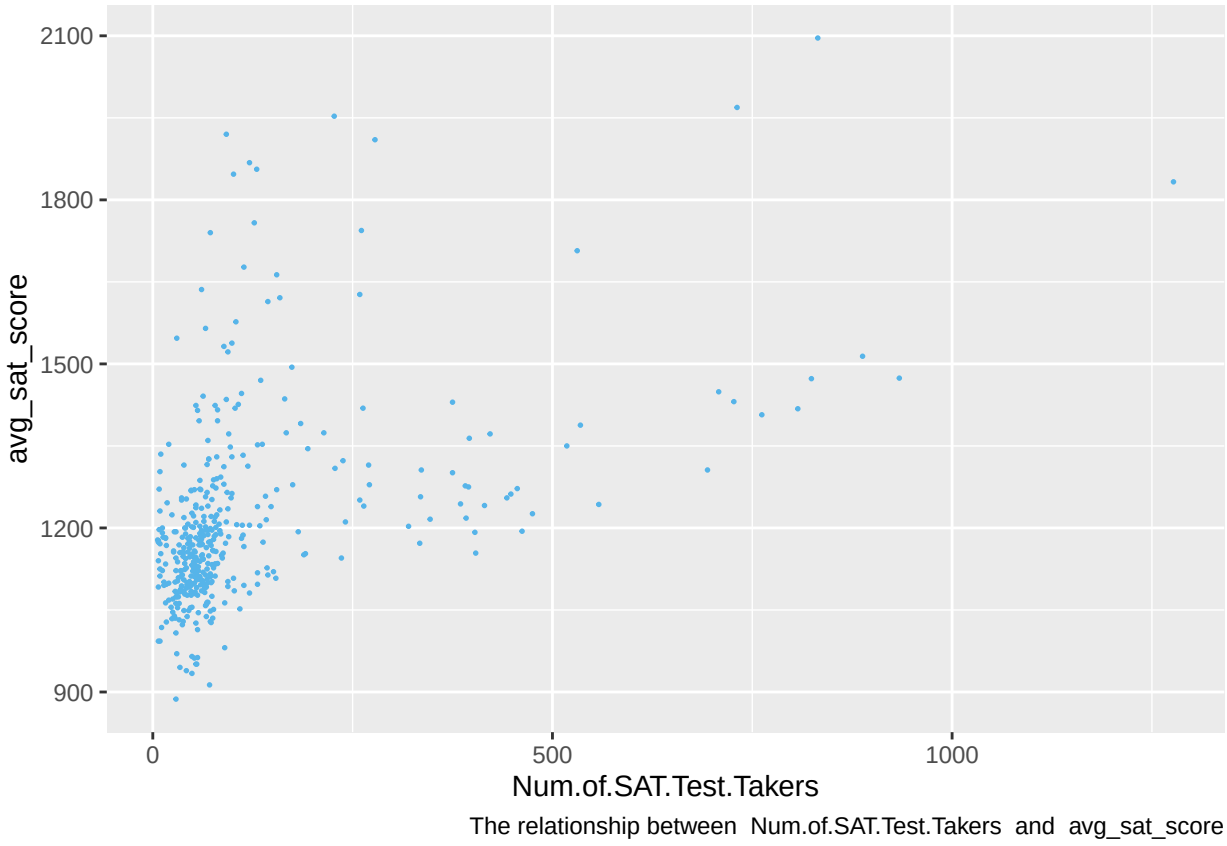
Next, we create a function `create_scatter` that produces a scatter plot of two given variables in the `combined_full` dataframe.

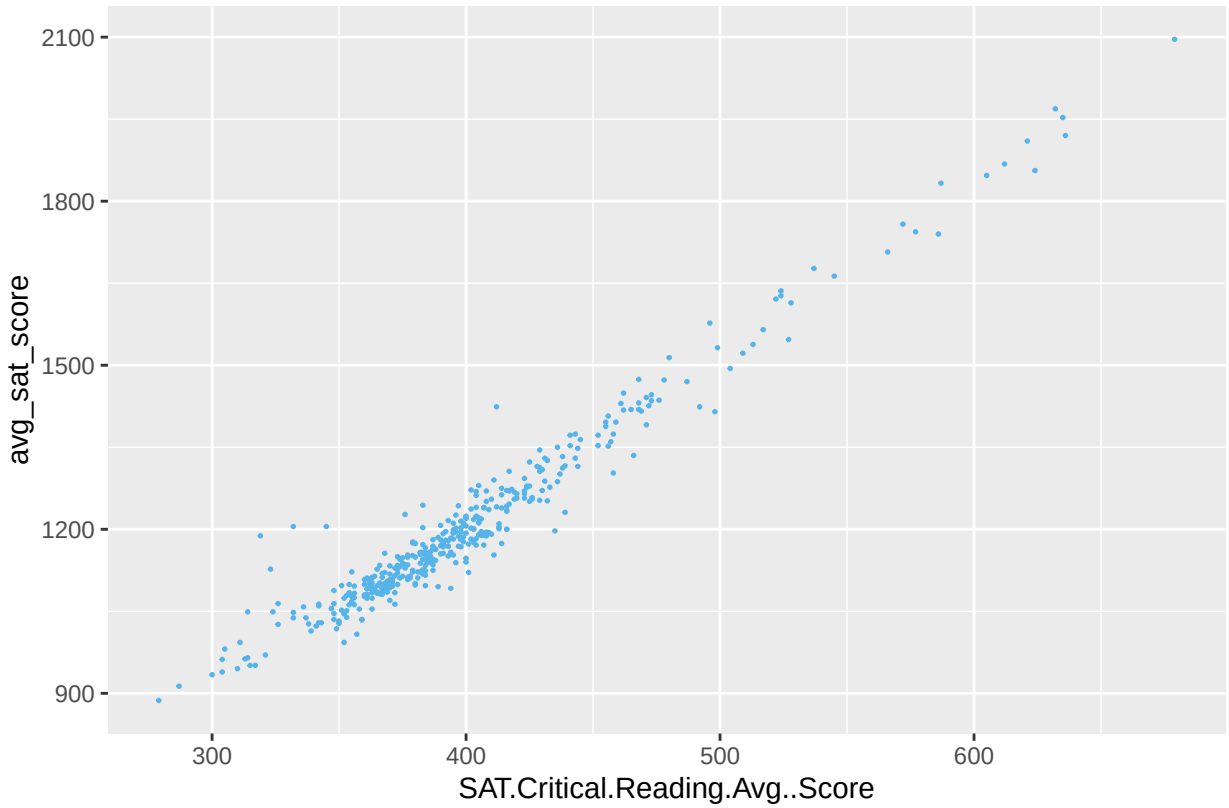
```
create_scatter <- function(xvar,yvar){
  combined_full %>%
  ggplot()+aes_string(x=xvar,y=yvar)+
    geom_point(color="#56B4E9",size=0.3)+
```

```
    #xlab(xvar)+
    #ylab(yvar)
   labs(caption = paste("The relationship between ", xvar,
      " and ", yvar),x=xvar,
    y=yvar)
}
```

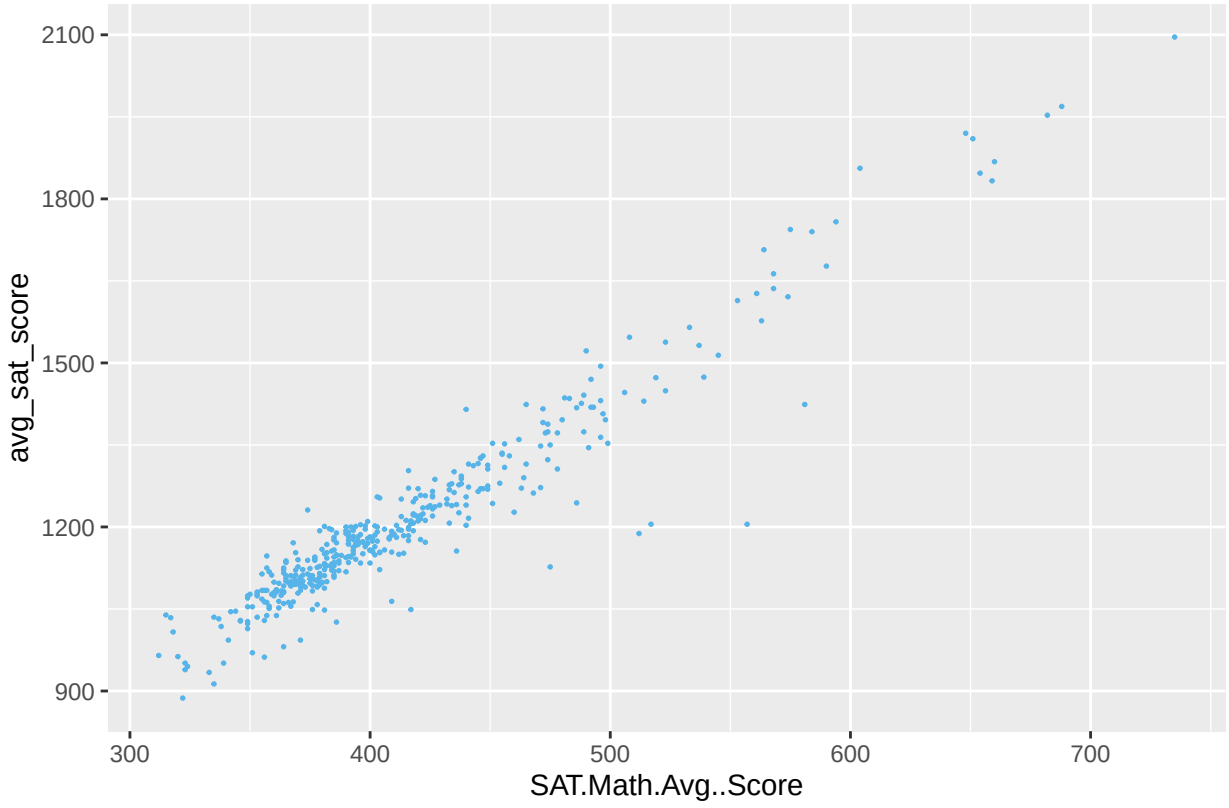
Finally, we plot the relationships between `avg_sat_score` and all the other variables such that their correlations with `avg_sat_score` are larger than 0.25.

```
xvar <- signif_cors$variable
yvar <- "avg_sat_score"

map2(xvar,yvar,create_scatter)
```

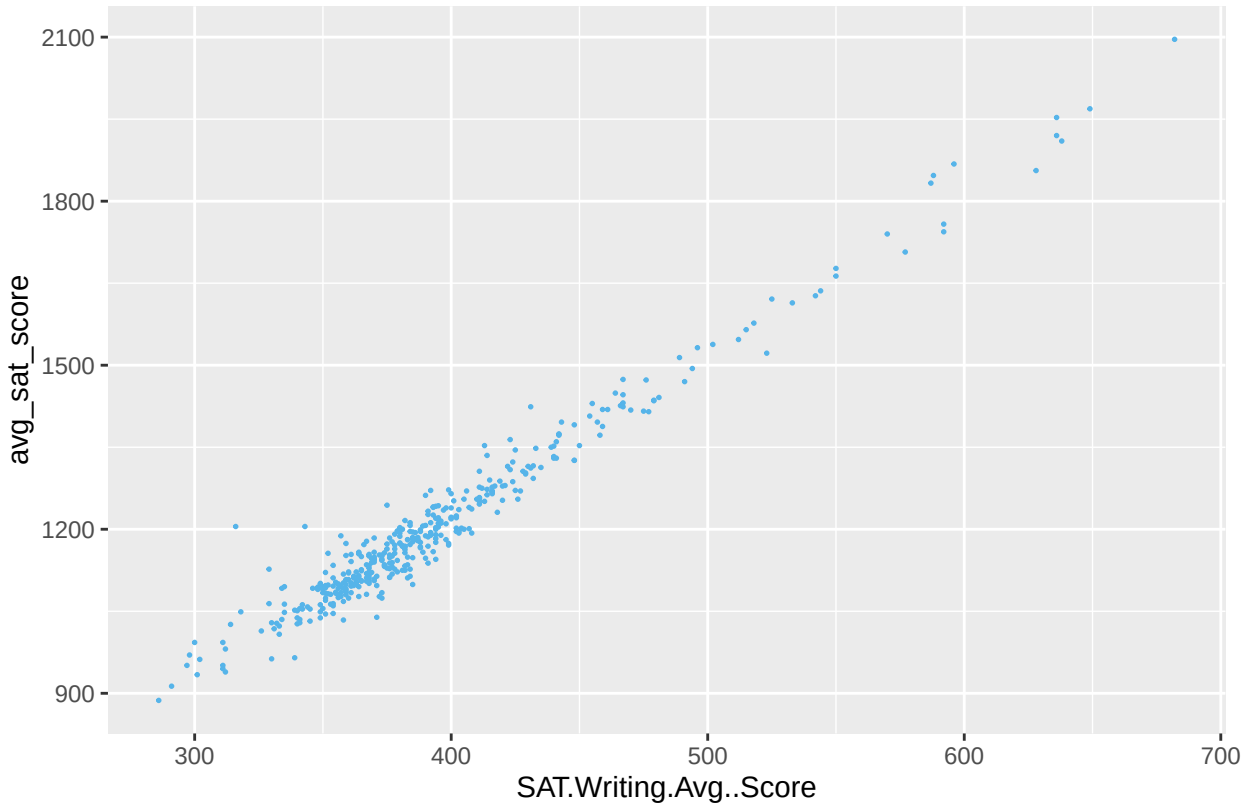




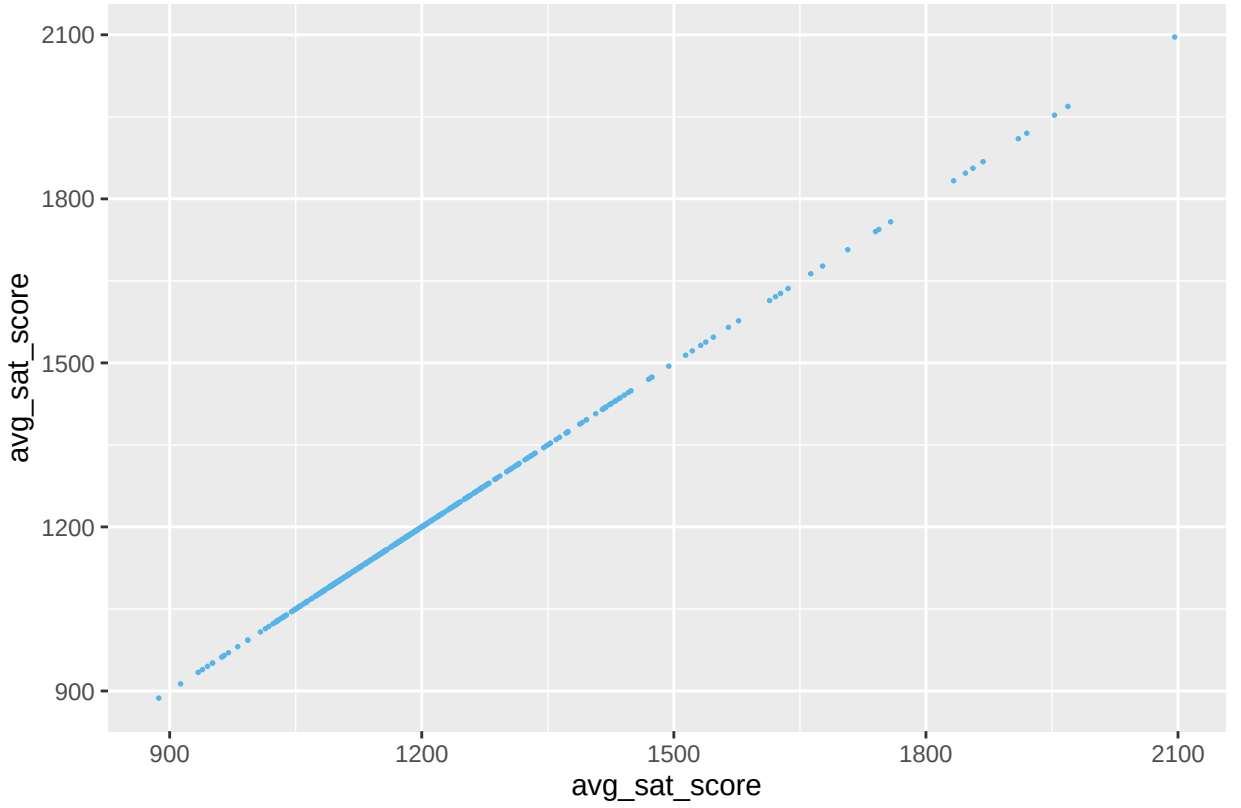
The relationship between SAT.Critical.Reading.Avg..Score and avg_sat_score



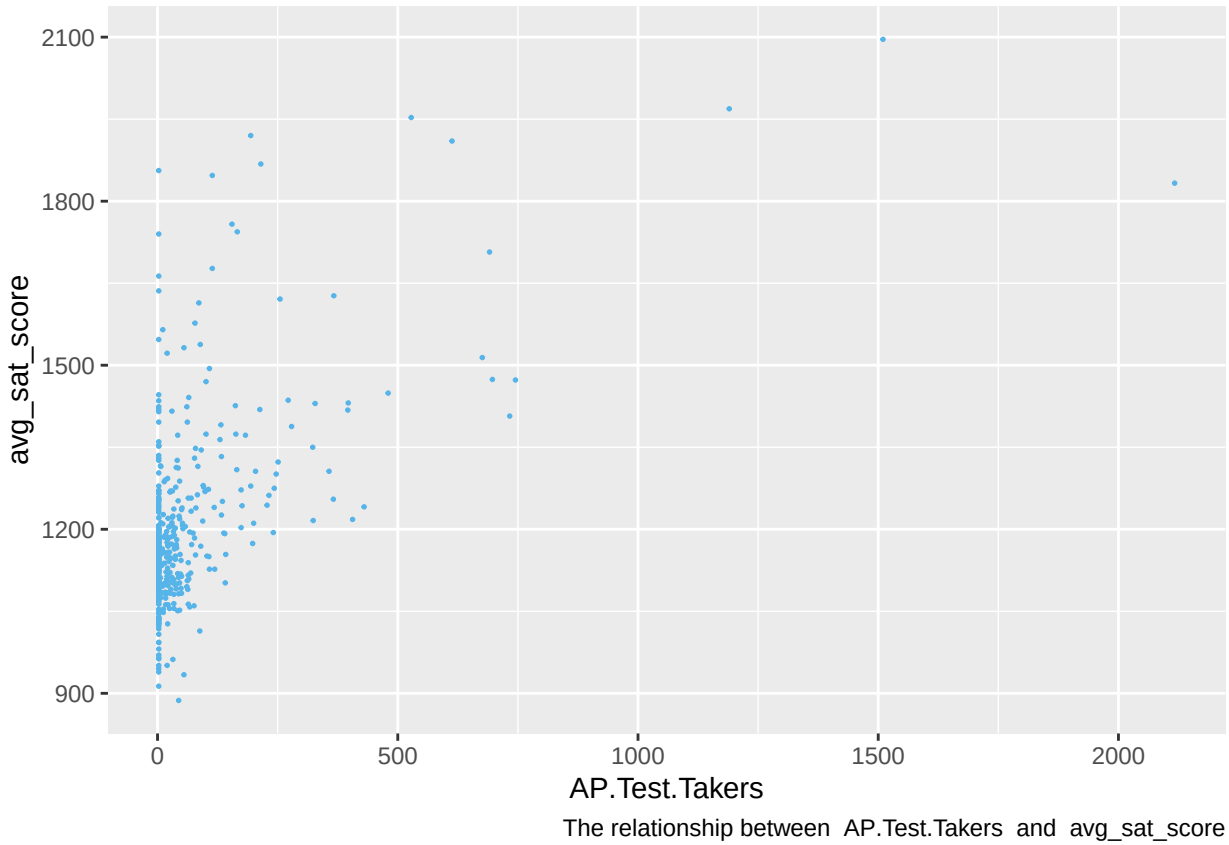
The relationship between SAT.Math.Avg..Score and avg_sat_score



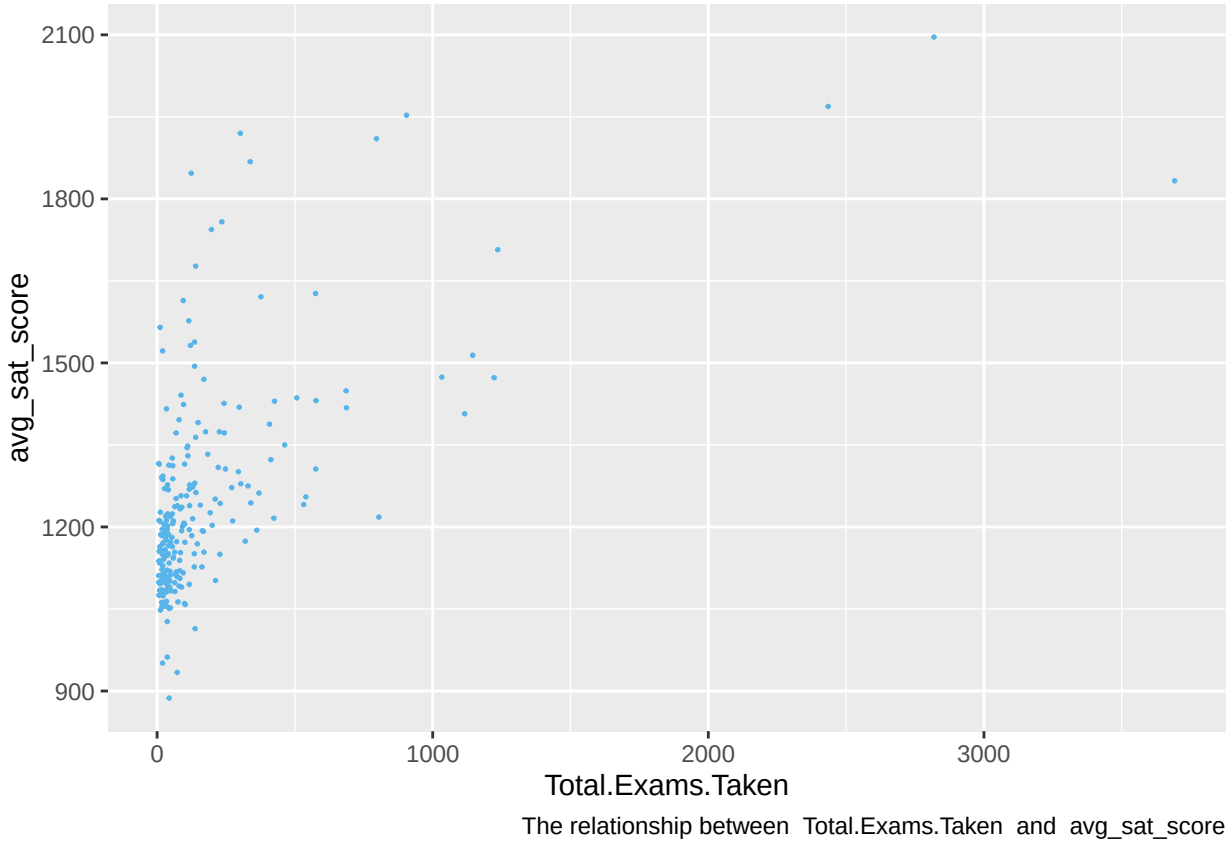
The relationship between SAT.Writing.Avg..Score and avg_sat_score



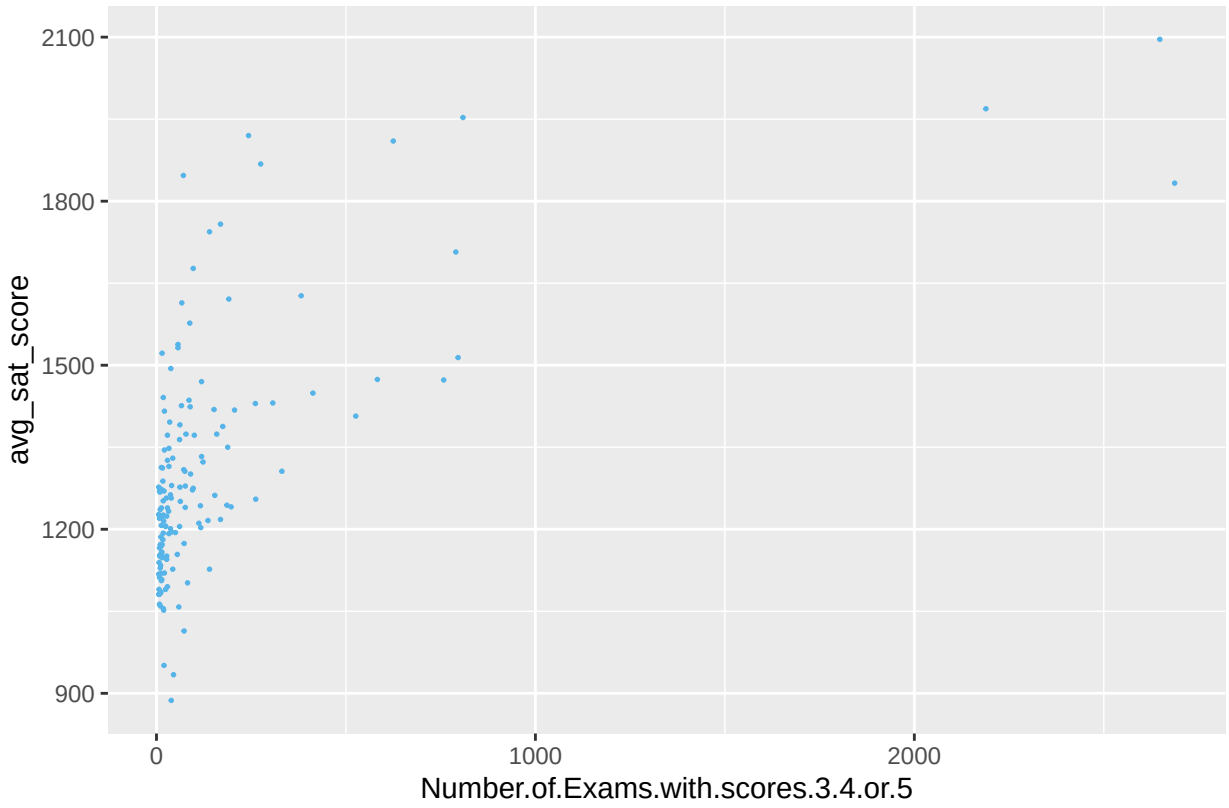
The relationship between avg_sat_score and avg_sat_score



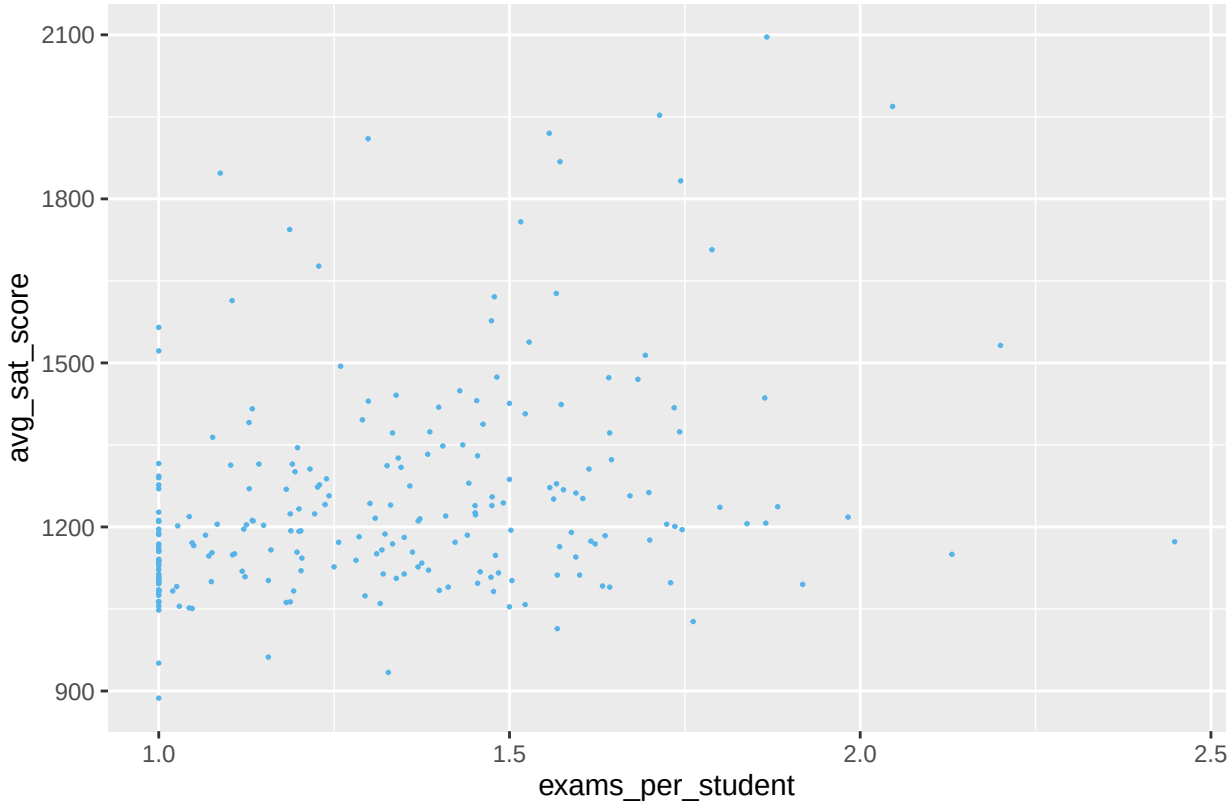
The relationship between AP.Test.Takers and avg_sat_score



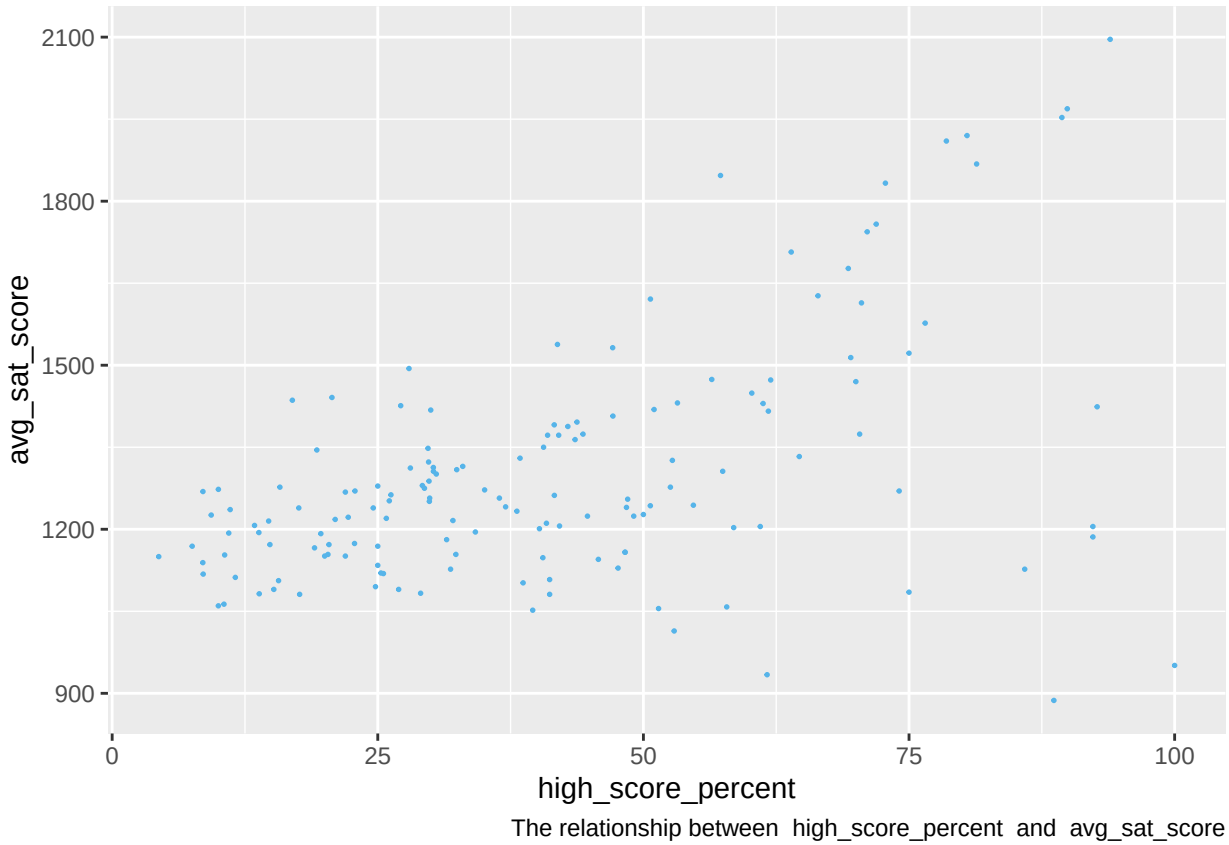
The relationship between Total.Exams.Taken and avg_sat_score



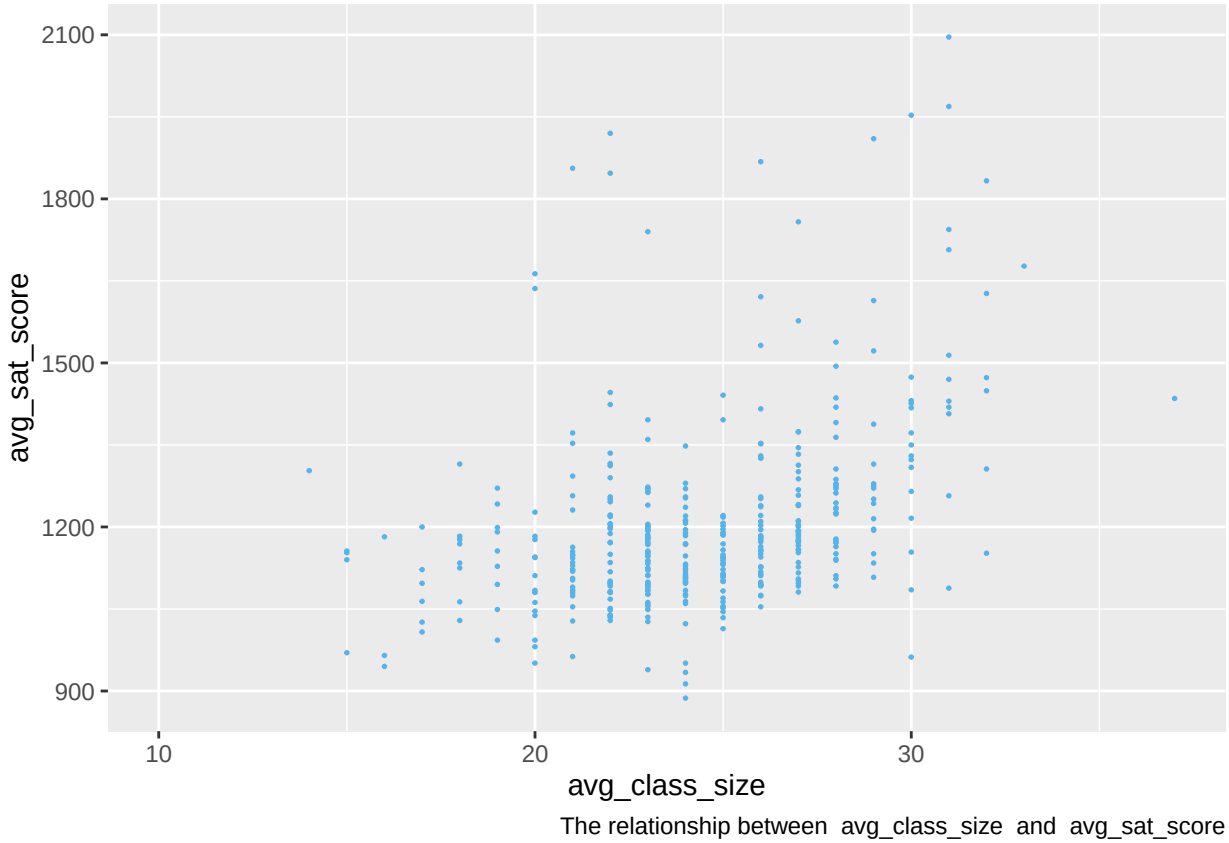
The relationship between Number.of.Exams.with.scores.3.4.or.5 and avg_sat_score



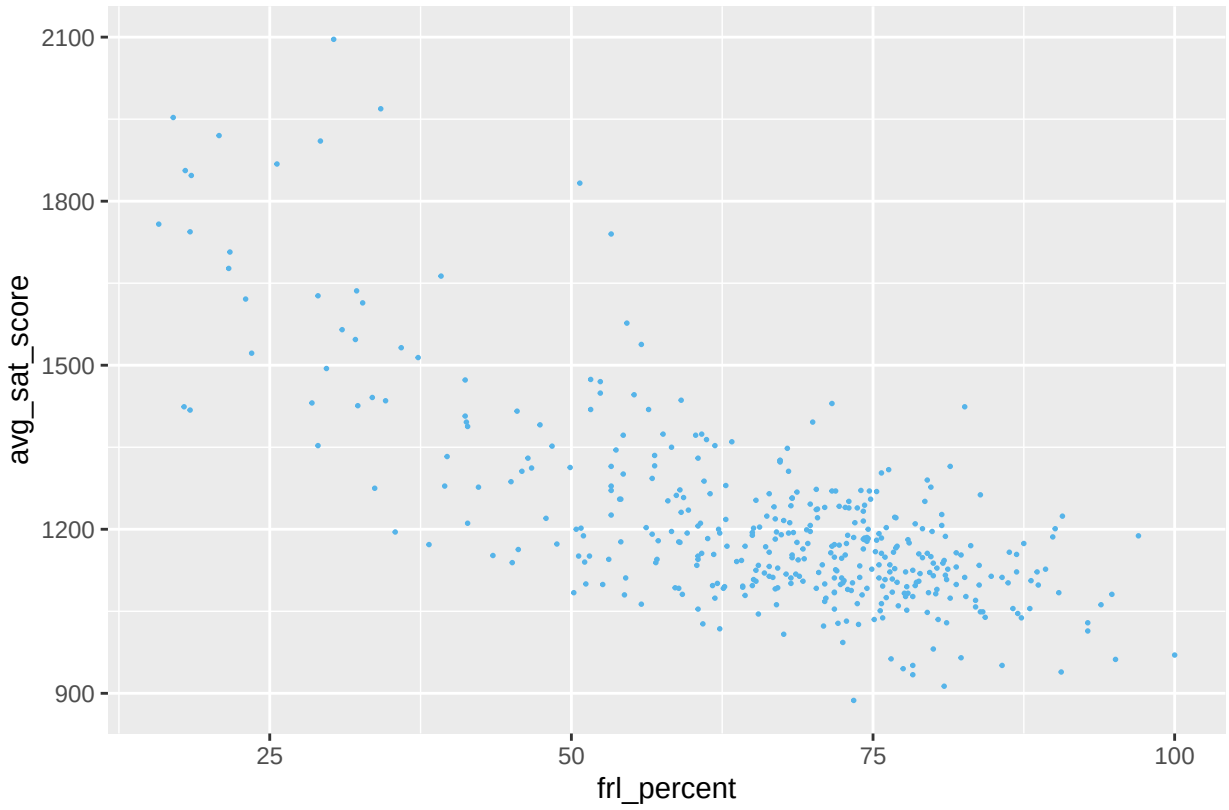
The relationship between exams_per_student and avg_sat_score



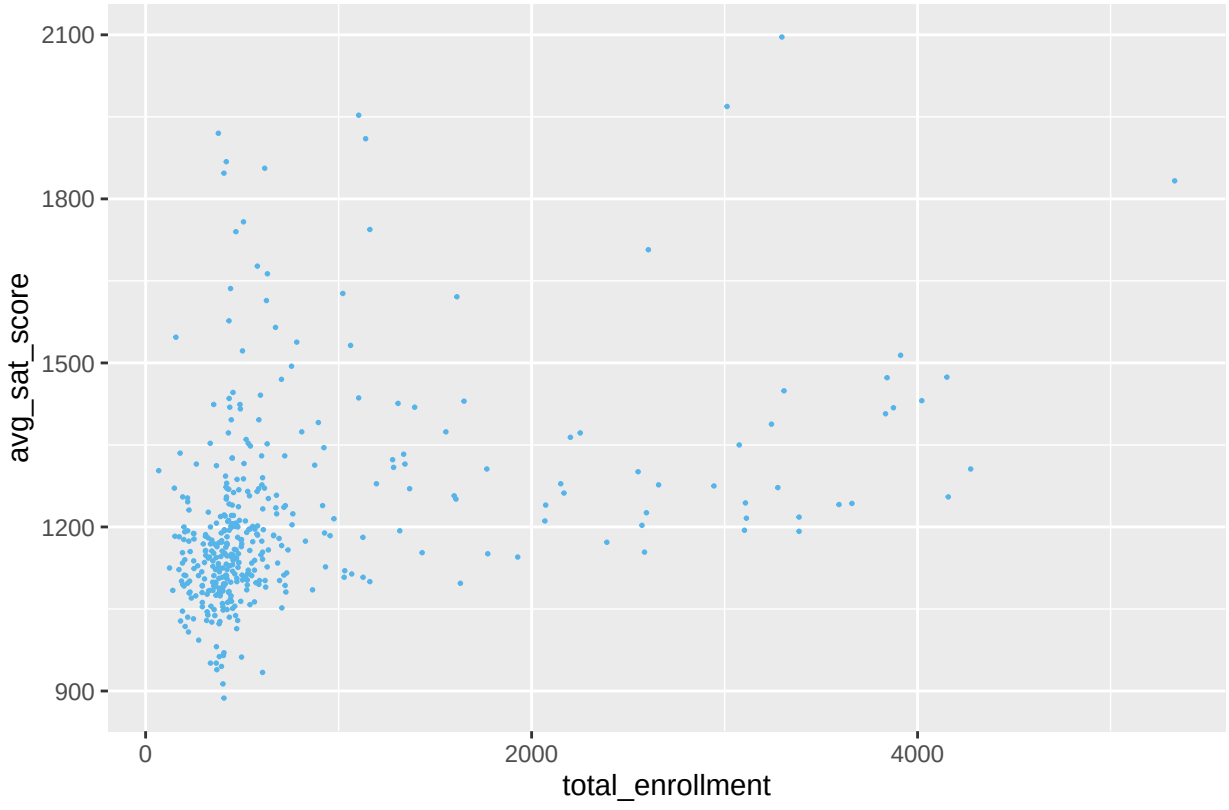
The relationship between high_score_percent and avg_sat_score



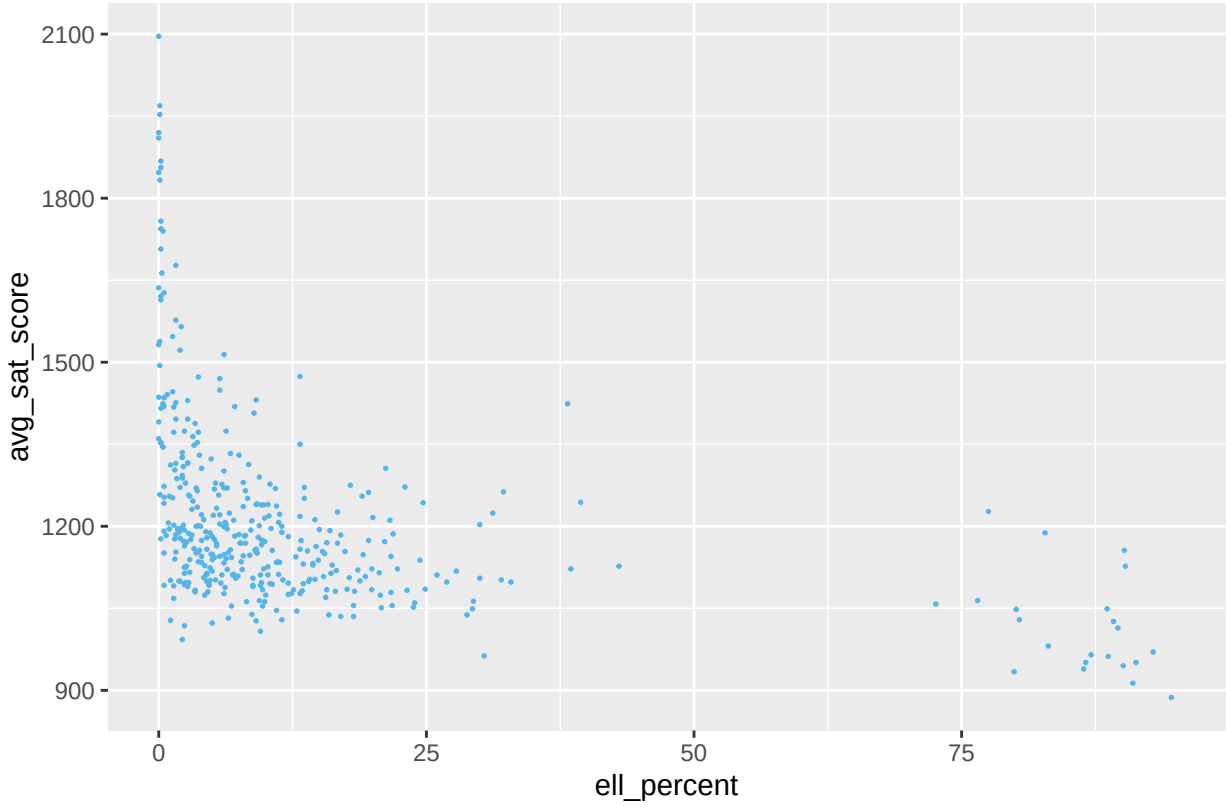
The relationship between avg_class_size and avg_sat_score



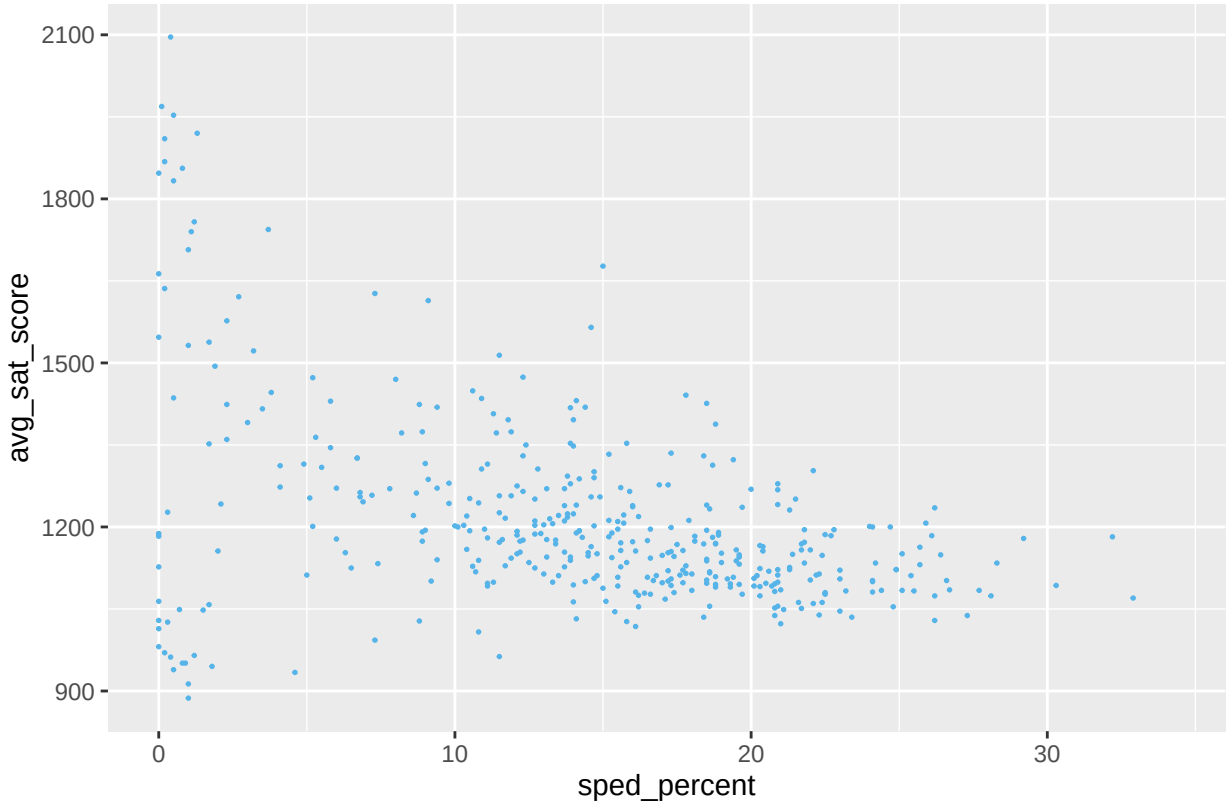
The relationship between frl_percent and avg_sat_score



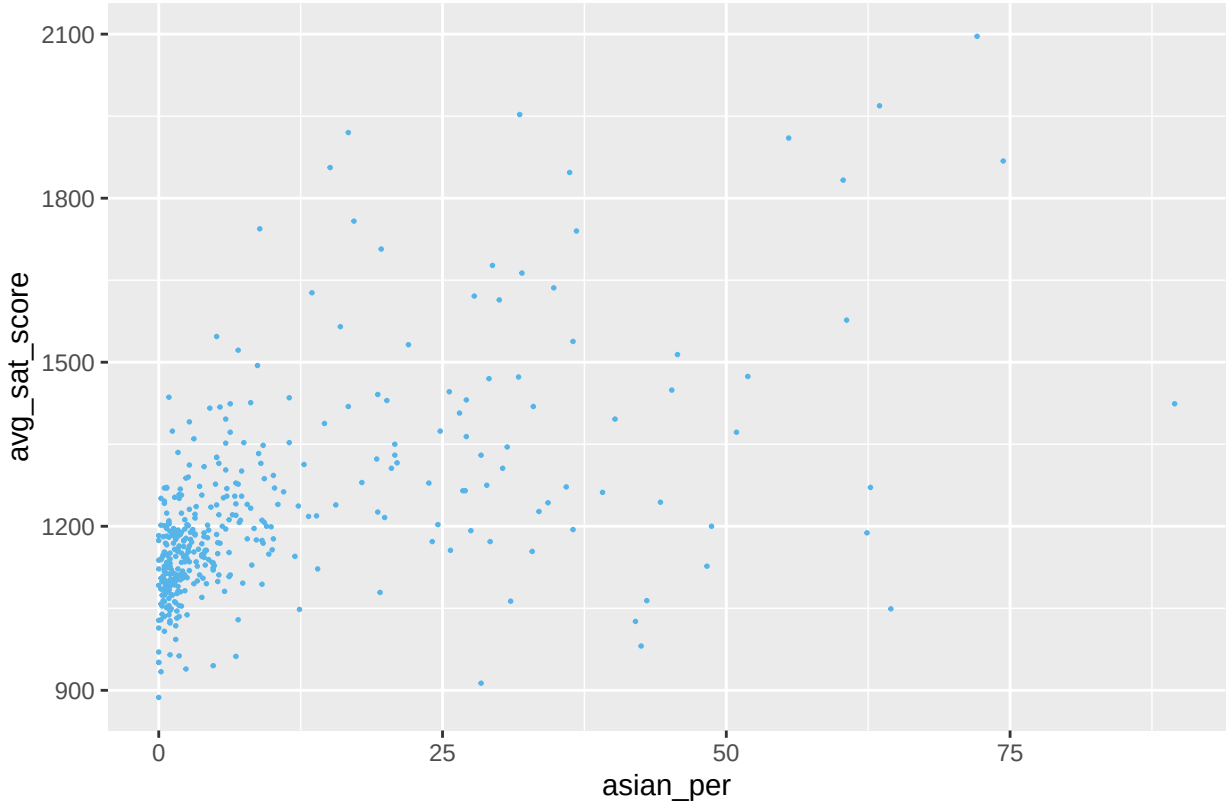
The relationship between total_enrollment and avg_sat_score



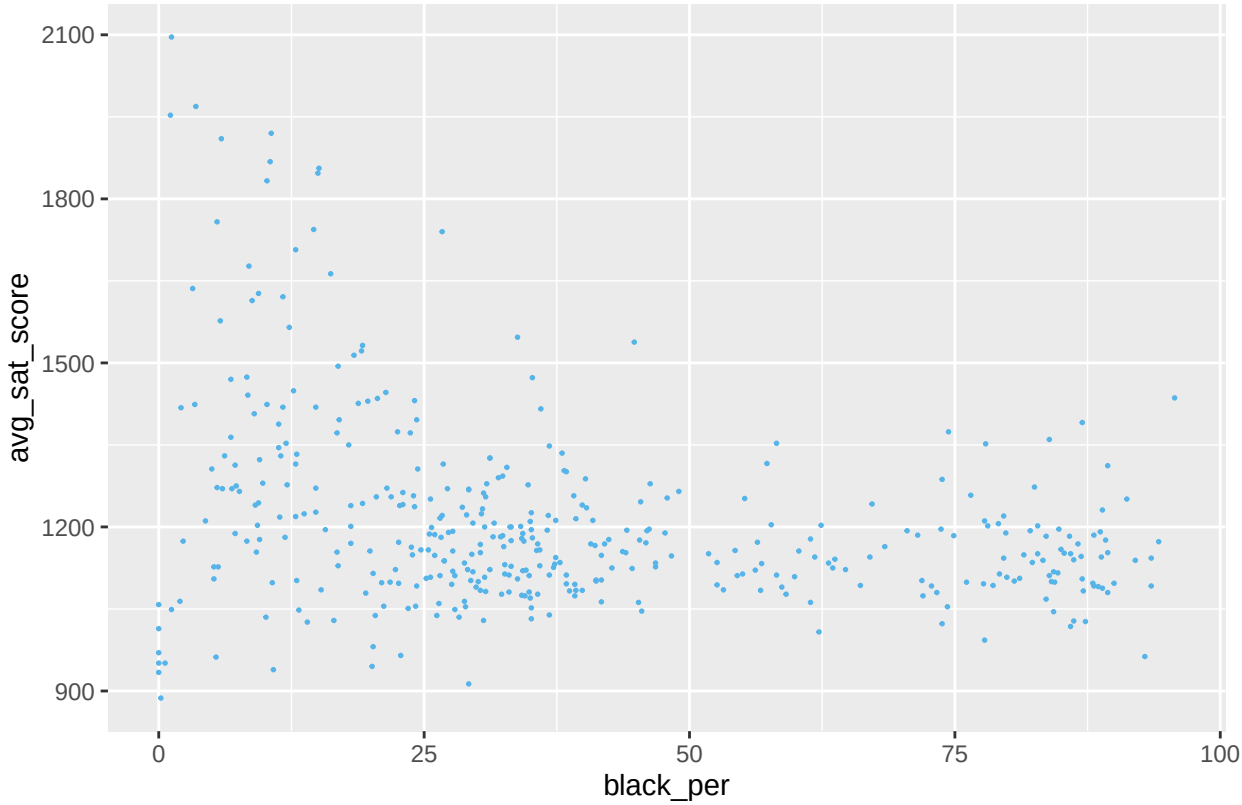
The relationship between ell_percent and avg_sat_score



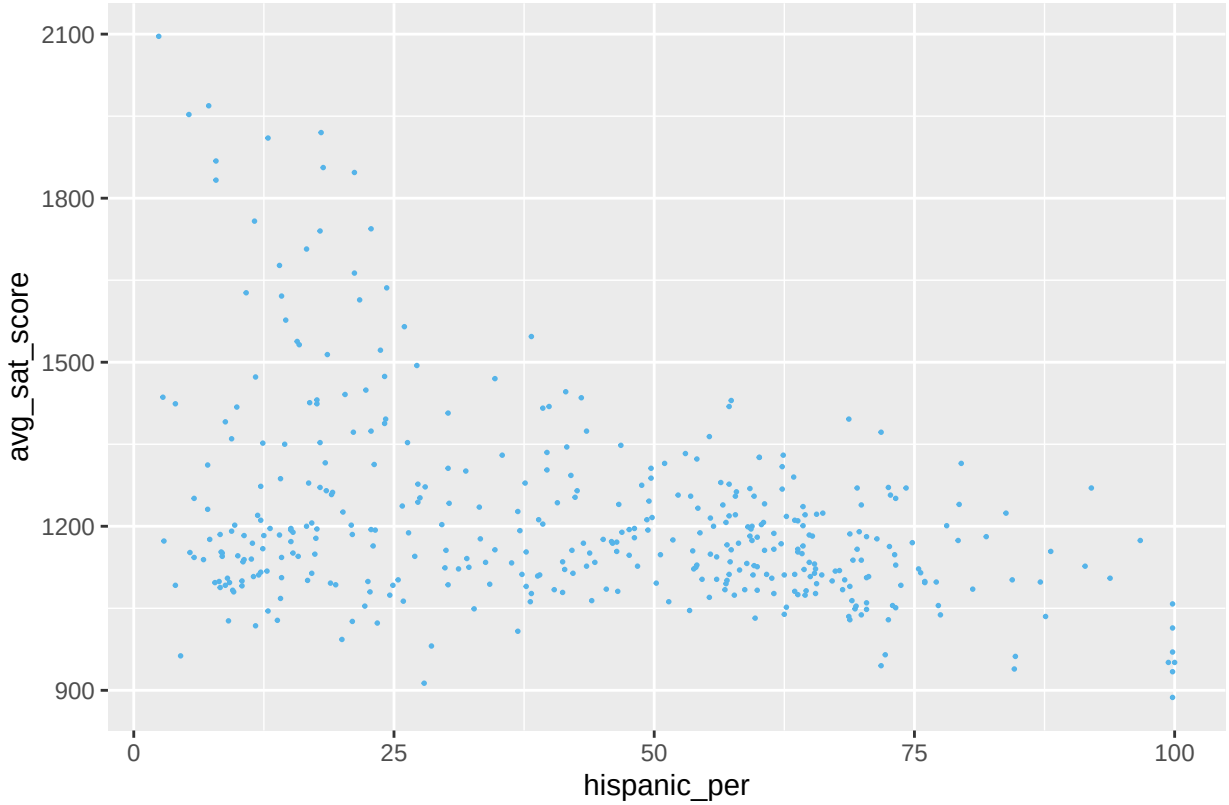
The relationship between sped_percent and avg_sat_score



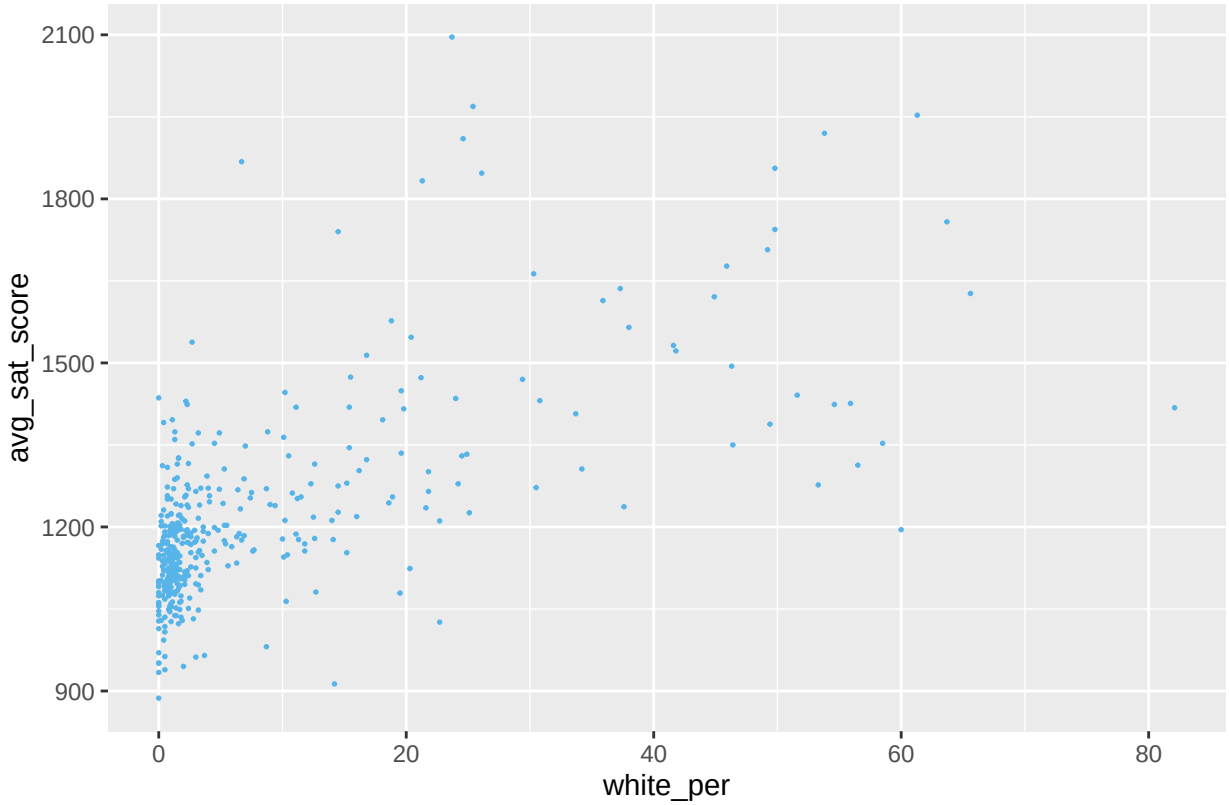
The relationship between asian_per and avg_sat_score



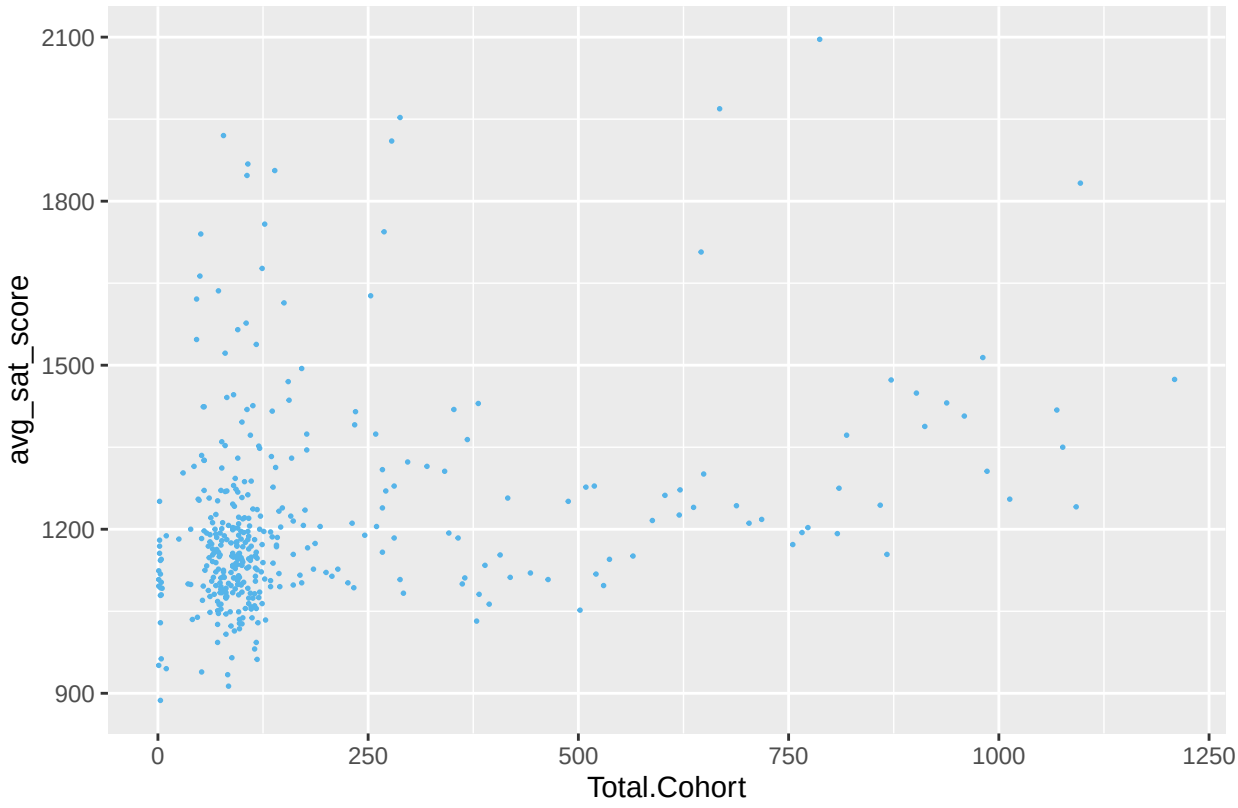
The relationship between black_per and avg_sat_score



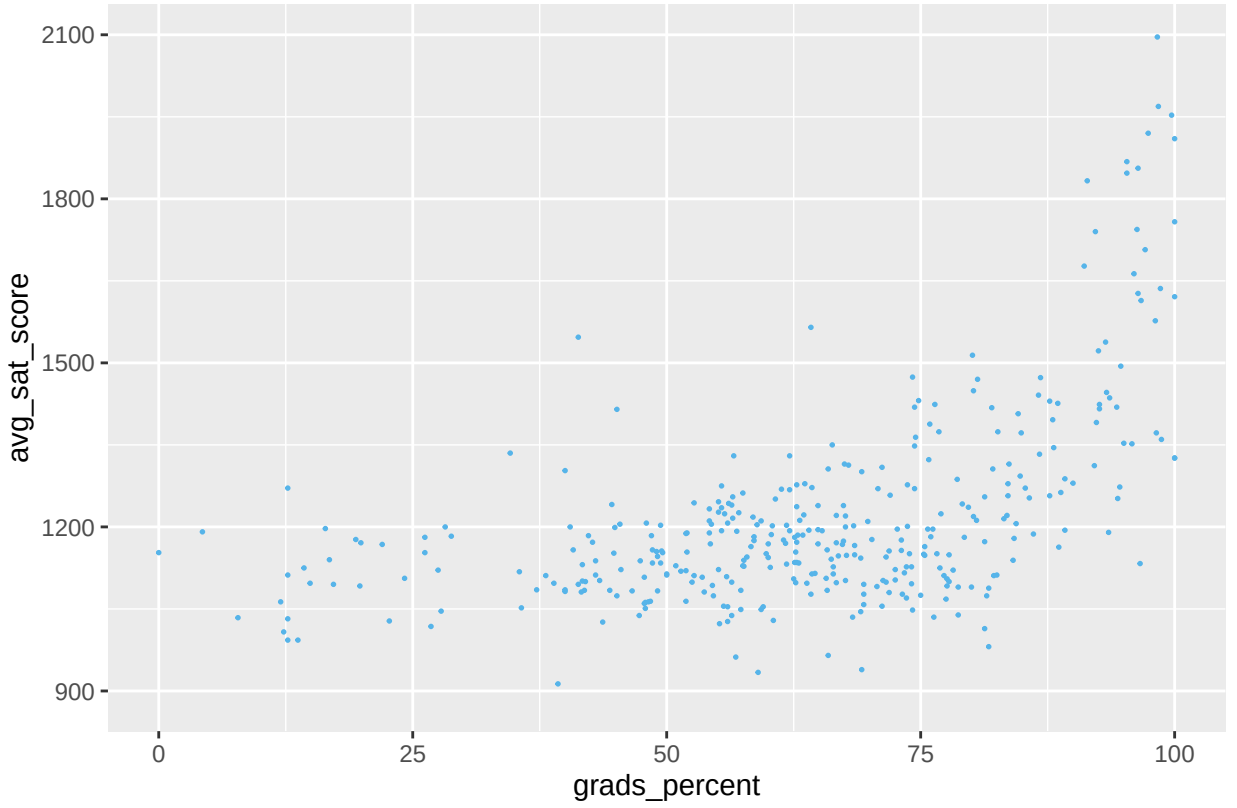
The relationship between hispanic_per and avg_sat_score



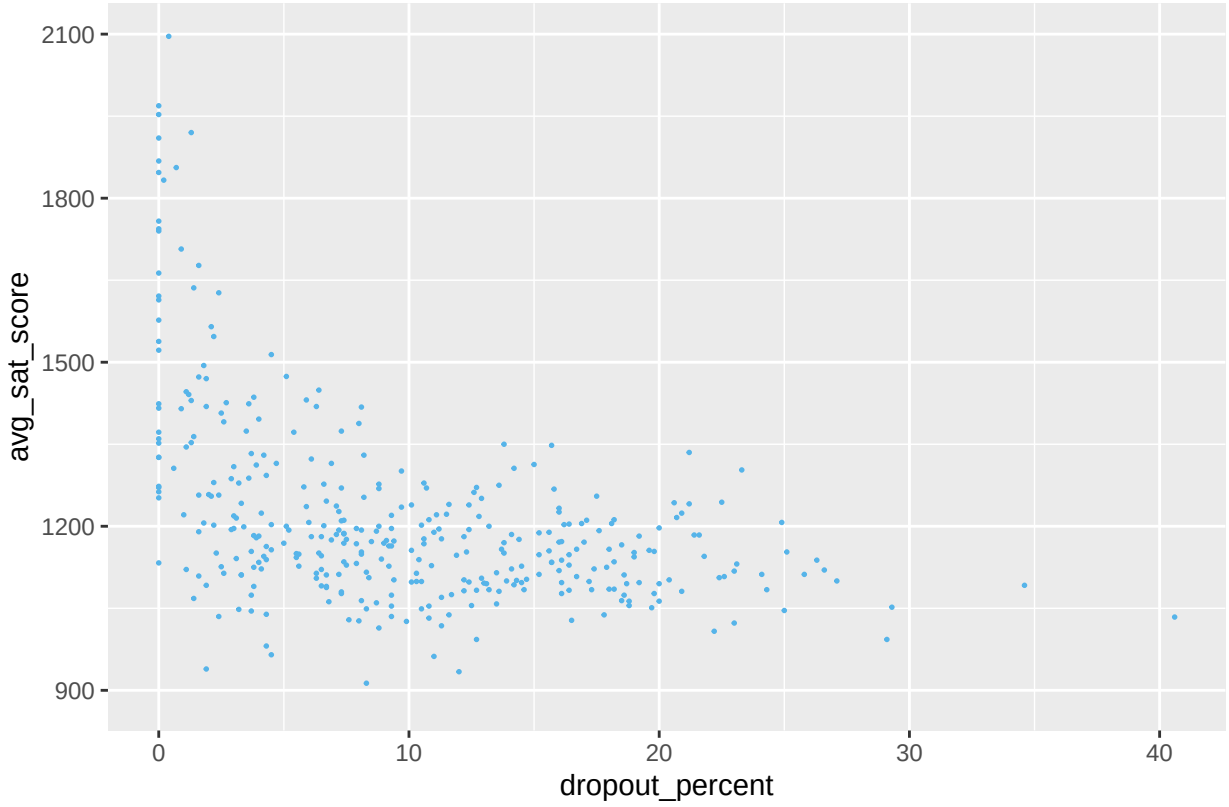
The relationship between white_per and avg_sat_score



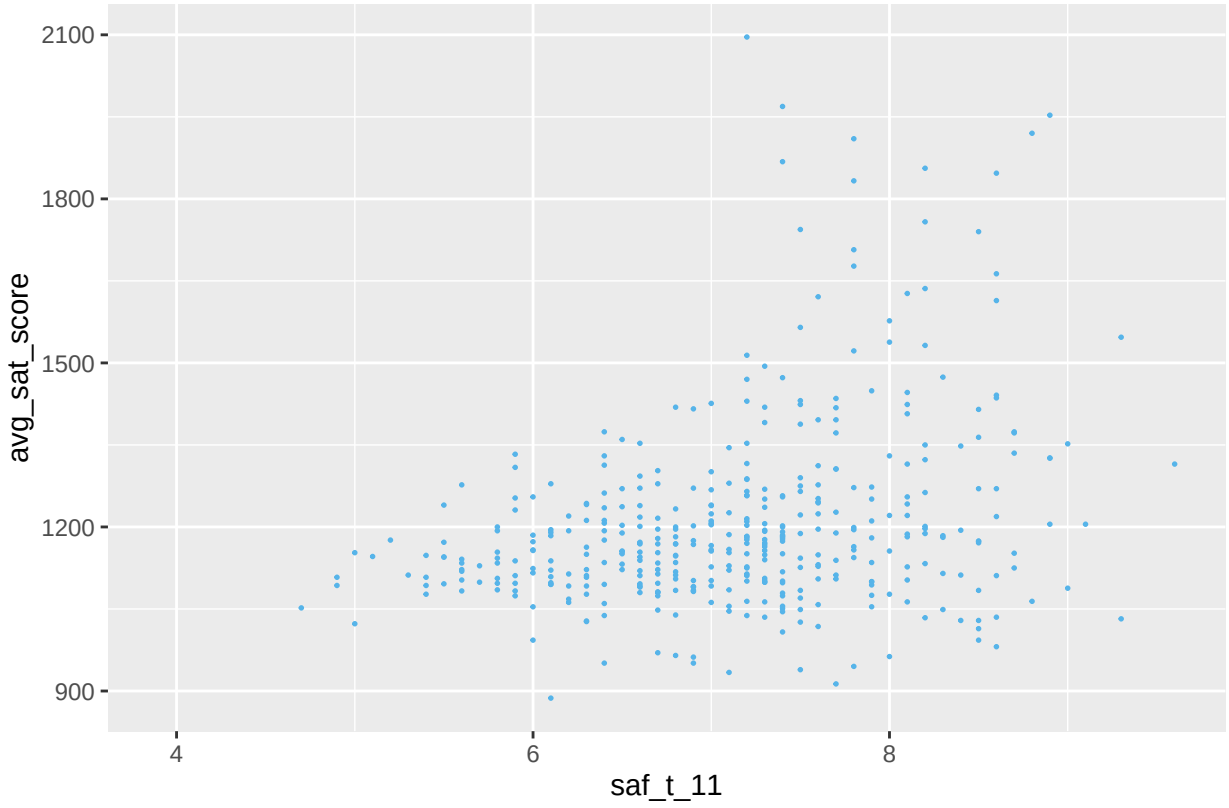
The relationship between Total.Cohort and avg_sat_score



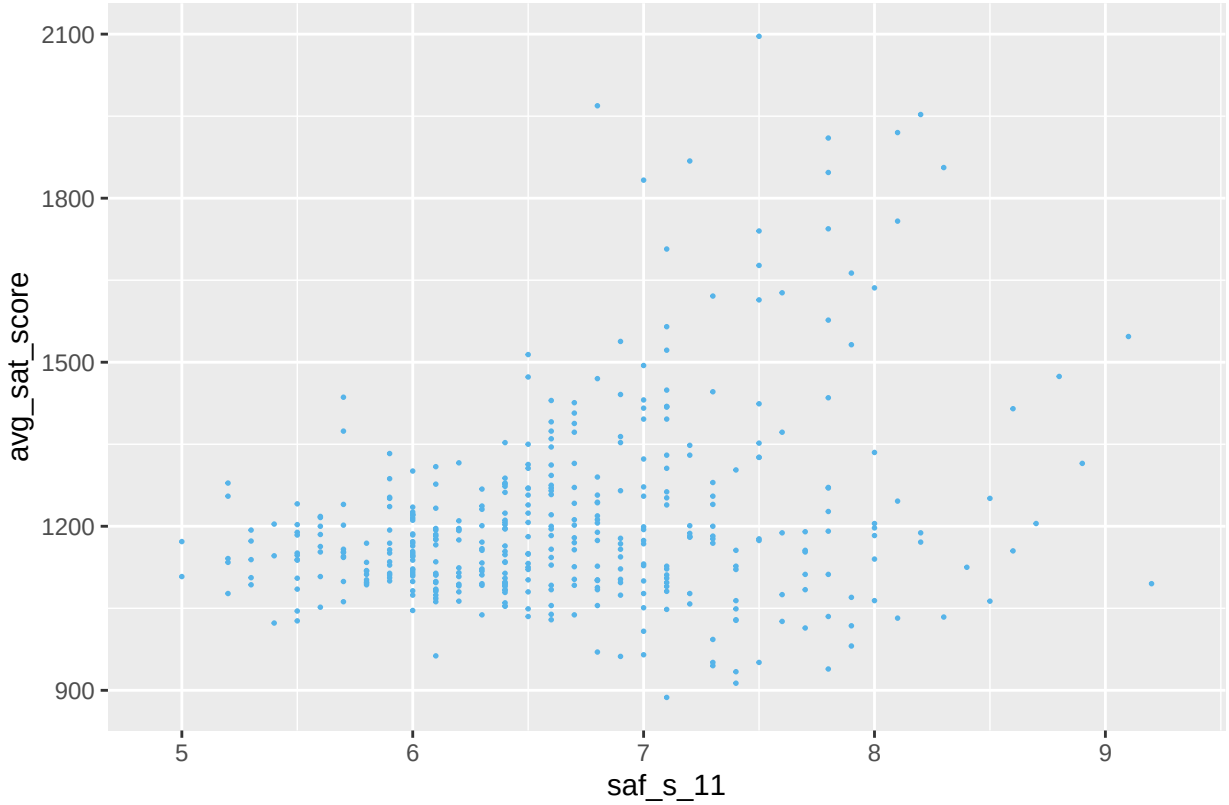
The relationship between grads_percent and avg_sat_score



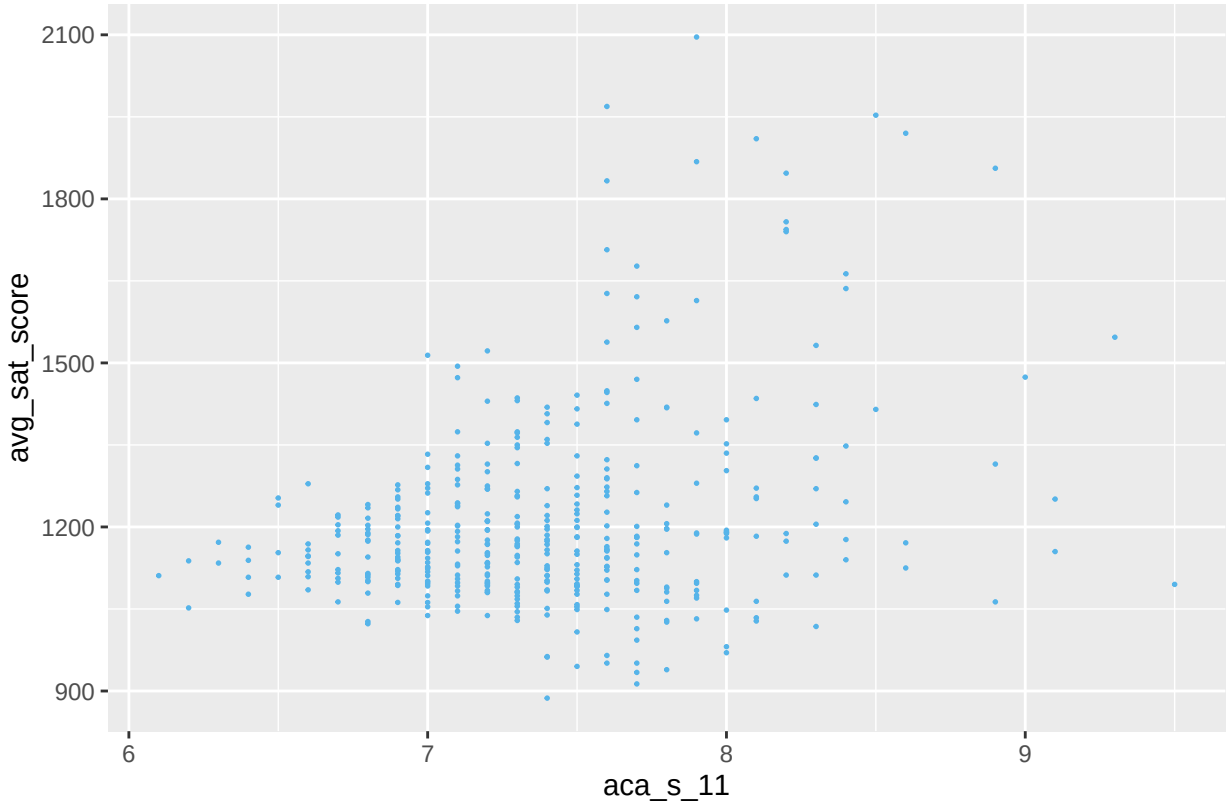
The relationship between dropout_percent and avg_sat_score



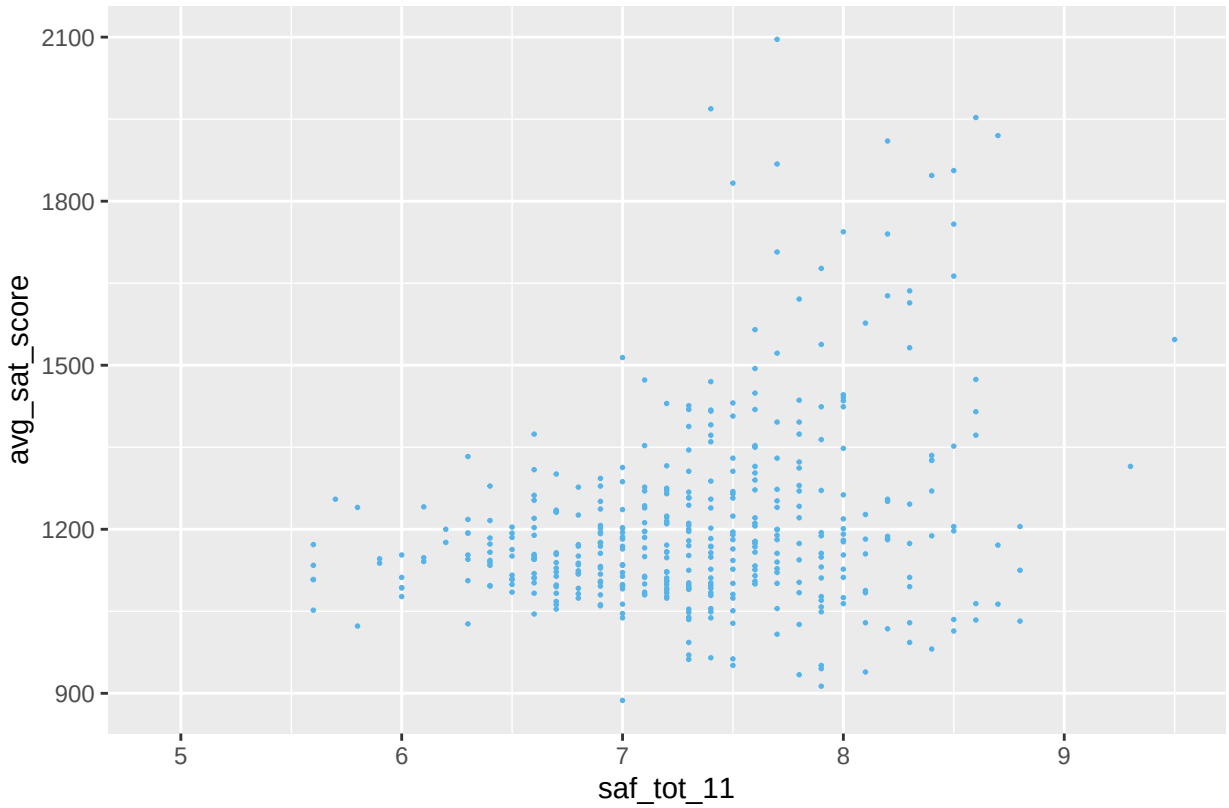
The relationship between saf_t_11 and avg_sat_score



The relationship between saf_s_11 and avg_sat_score



The relationship between aca_s_11 and avg_sat_score



The relationship between saf_tot_11 and avg_sat_score

```
## [[1]]
##
## [[2]]
##
## [[3]]
##
## [[4]]
##
## [[5]]
##
## [[6]]
##
## [[7]]
##
## [[8]]
##
## [[9]]
##
## [[10]]
##
## [[11]]
##
## [[12]]
##
## [[13]]
##
## [[14]]
```

```
##
## [[15]]
##
## [[16]]
##
## [[17]]
##
## [[18]]
##
## [[19]]
##
## [[20]]
##
## [[21]]
##
## [[22]]
##
## [[23]]
##
## [[24]]
##
## [[25]]
##
## [[26]]
```

Our conclusions from the last set of plots are stated in Section @ref{ResultsSection}.

5.3 Analysis of the Differences between the Responses from Parents, Teachers, and Students

In this subsection, we compute the average scores for parents', teacher's, students', and total survey responses and compare them. We begin with creating a tibble with the above mentioned average responses.

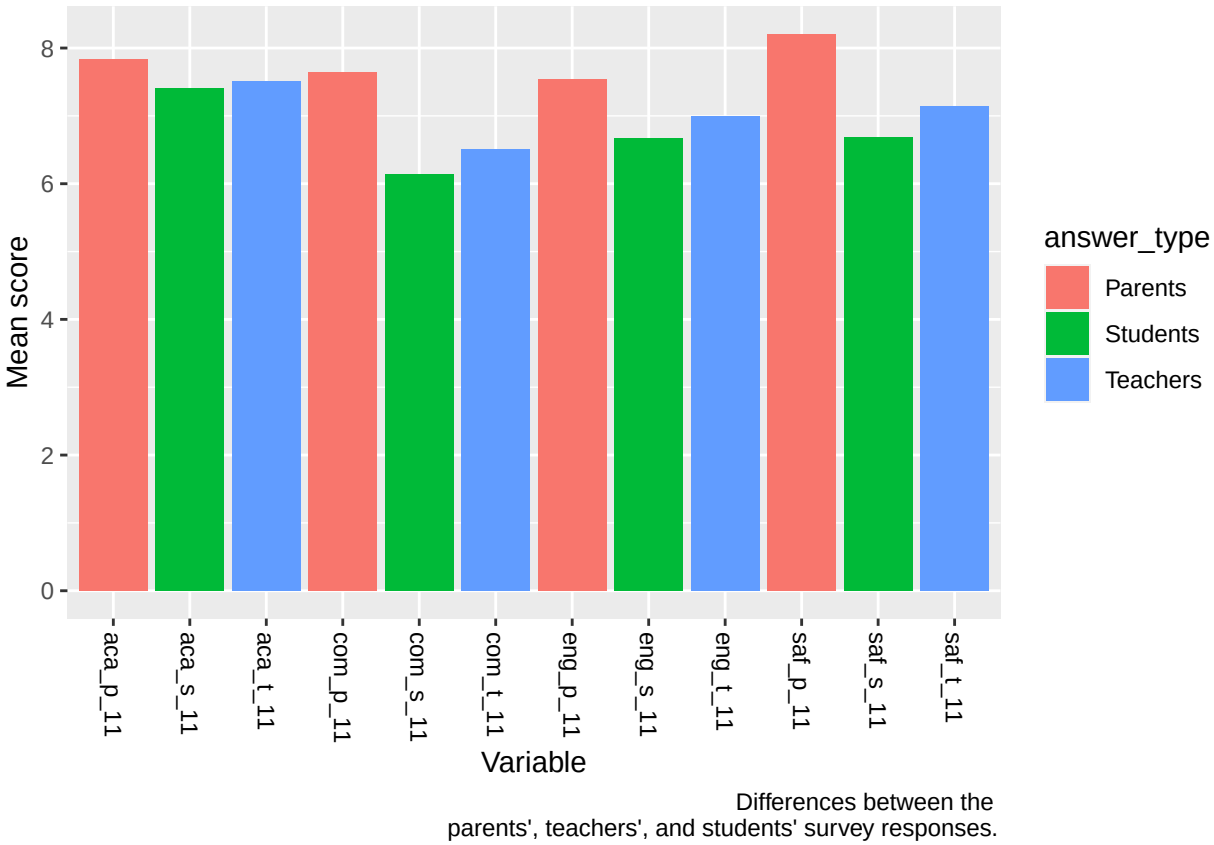
```
responses_pivoted<-combined_full[32:43]%>%
  pivot_longer(cols=c(
    saf_p_11,
    saf_t_11,
    saf_s_11,
    com_p_11,
    com_t_11,
    com_s_11,
    eng_p_11,
    eng_t_11,
    eng_s_11,
    aca_p_11,
    aca_t_11,
    aca_s_11
  ),
  names_to="variable",
  values_to='score')

mean_scores <- responses_pivoted %>% group_by(variable) %>%
  summarise(mean_score=mean(score,na.rm = TRUE))
```

We then prepare tibble `mean_scores` for graphing. In particular, we create column `answer_type` that contains the name of the group that responded to a particular question.

```
mean_scores <- mean_scores %>% mutate(
  answer_type=case_when(
    str_detect(variable,"_p_") ~ "Parents",
    str_detect(variable,"_t_") ~ "Teachers",
    str_detect(variable,"_s_") ~ "Students",
    str_detect(variable,"_tot_") ~ "Total"
  ),
  question_type=case_when(
    str_detect(variable,"saf_") ~ "Safety and respect",
    str_detect(variable,"com_") ~ "Communication",
    str_detect(variable,"eng_") ~ "Engagement",
    str_detect(variable,"aca_") ~ "Academic expectations"
  )
)
```

Finally, we create a plot that illustrates the differences between the parents', teachers', students', and total responses.

```
 plot<-mean_scores %>% ggplot(aes(x=variable,y=mean_score,fill=answer_type))+
  geom_bar(position='dodge',stat="identity")+
  labs(x="Variable",
    y="Mean score", caption="Differences between the
    parents', teachers', and students' survey responses.")+
  theme(axis.text.x = element_text(size=9, angle = -90,
  vjust = 0.5, hjust = 0, color = "black") )

  print(plot)
```



Differences between the parents', teachers', and students' survey responses.

From the last bar chart, we conclude that the responses from the parents are higher, on average. Average students' response scores are the lowest, and teachers' are in between.

6 Conclusions

In this paper, we analyzed survey data relating to NYC schools. We studied the relationship between race and survey responses. We analyzed the difference between parents', students', and teachers' responses. Most importantly, we identified all the significant correlations between data variables and survey responses. To perform the analysis, the data was appropriately prepared, explored, cleaned, and visualized. Statistically meaningful parameters such as correlation and means were calculated. The information extracted from the data proved interesting; it is reported in Section 2.