# DS 5110 Final Project

Tim Cauley

2022-03-16

## First Steps (Importing, cleaning, eda, joining)

### Importing packages

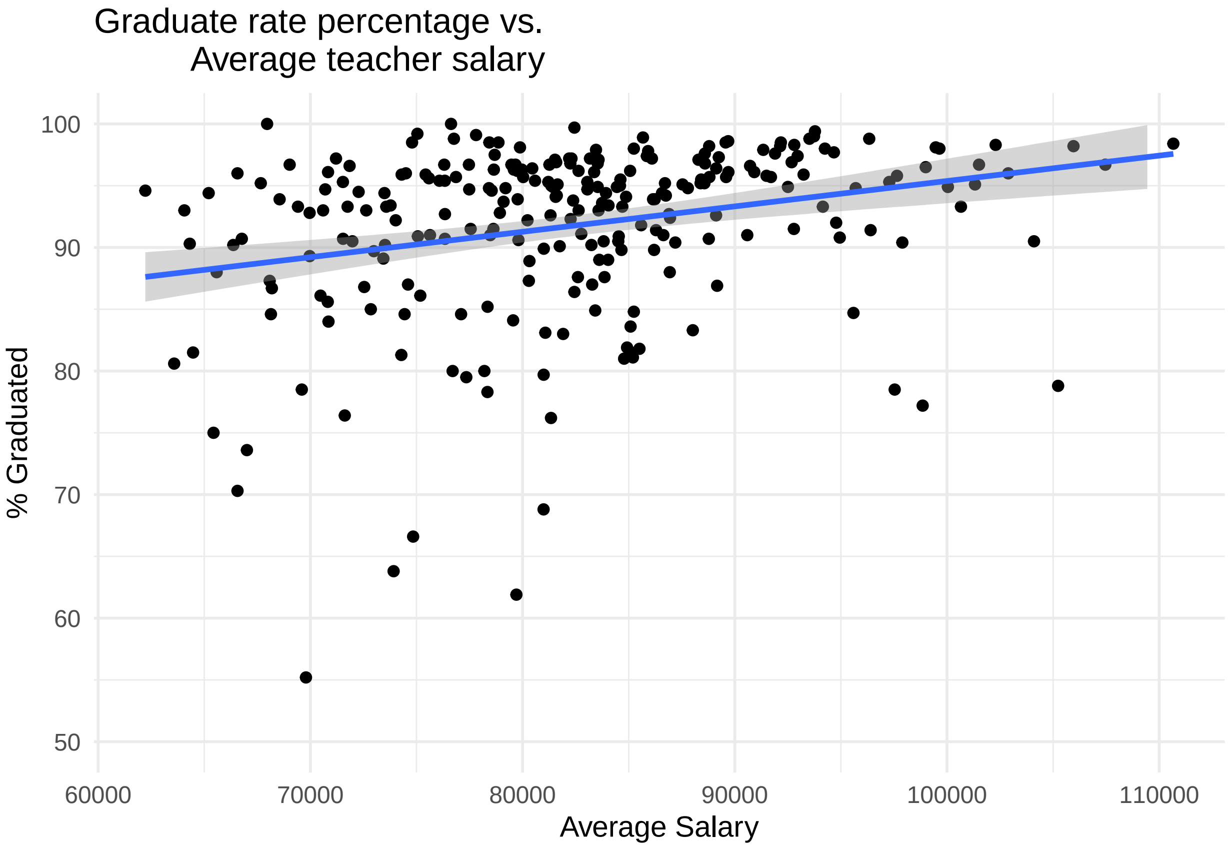
```
library(dplyr)
library(readr)
library(ggplot2)
library(modelr)
library(tidyr)
library(grid)
library(gridExtra)
```

### Importing Data

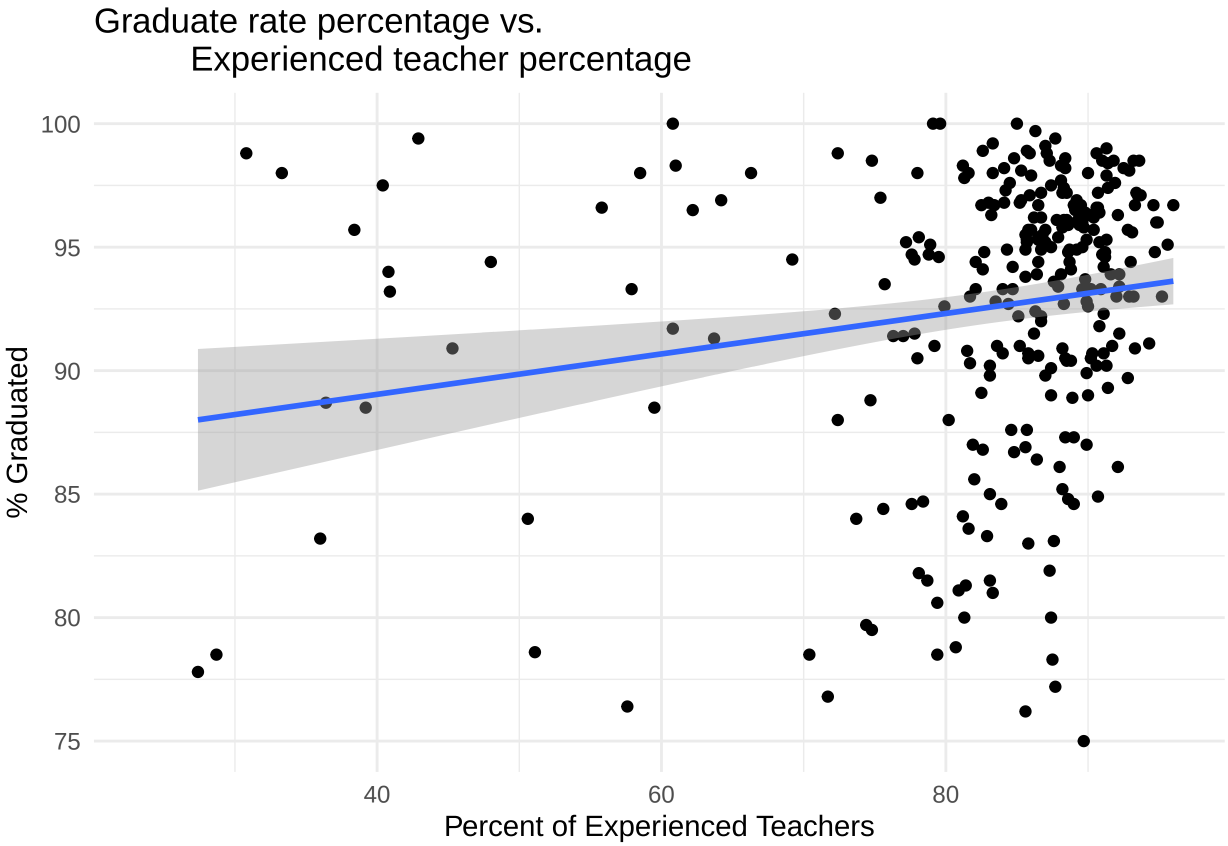
```
sat <- read_csv("/Users/Tim/Downloads/sat_performance.csv")
expend <- read_csv("/Users/Tim/Downloads/PerPupilExpenditures.csv")
salary <- read_csv("/Users/Tim/Downloads/TeacherSalaries.csv")
enroll <- read_csv("/Users/Tim/Downloads/enrollmentbygrade.csv")
ap <- read_csv("/Users/Tim/Downloads/ap_participation.csv")
reten <- read_csv("/Users/Tim/Downloads/staffingretention.csv")
classSize <- read_csv("/Users/Tim/Downloads/ClassSizebyRaceEthnicity.csv")
college <- read_csv("/Users/Tim/Downloads/Gradsattendingcollege.csv")
attendance <- read_csv("/Users/Tim/Downloads/attendance.csv")
attrition <- read_csv("/Users/Tim/Downloads/AttritionReport.csv")
advCourse <- read_csv("/Users/Tim/Downloads/AdvancedCourseCompletion.csv")
gradRate <- read_csv("/Users/Tim/Downloads/gradrates.csv")
art <- read_csv("/Users/Tim/Downloads/artcourse.csv")
eduAge <- read_csv("/Users/Tim/Downloads/EducatorsbyAgeGroupsReport.csv")
discipline <- read_csv("/Users/Tim/Downloads/StudentDisciplineDataReport.csv")
eduGen <- read_csv("/Users/Tim/Downloads/staffracegender.csv")
teachData <- read_csv("/Users/Tim/Downloads/teacherdata.csv")
pop <- read_csv("/Users/Tim/Downloads/finalProject/ClassSizebyGenPopulation.csv")
eth <- read_csv("/Users/Tim/Downloads/finalProject/ClassSizebyRaceEthnicity.csv")
dropout <- read_csv("/Users/Tim/Downloads/finalProject/dropout.csv")
mobile <- read_csv("/Users/Tim/Downloads/finalProject/mobilityrates.csv")
teachProg <- read_csv("/Users/Tim/Downloads/Teacherprogramarea.csv")
daysMissed <- read_csv("/Users/Tim/Downloads/ssdr_days_missed.csv")
selectPop <- read_csv("/Users/Tim/Downloads/selectedpopulations.csv")
```

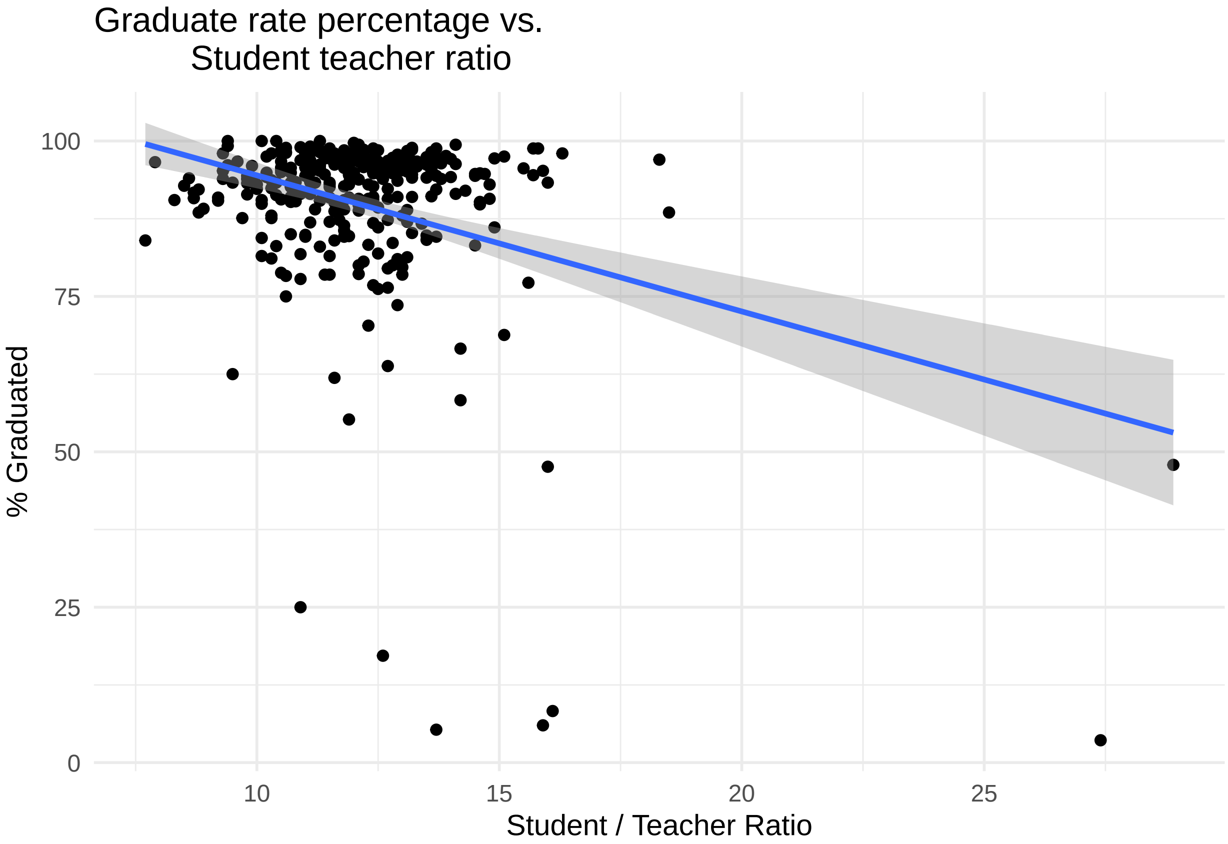
### EDA

**1. Teacher Salary vs. Graduation Rate**



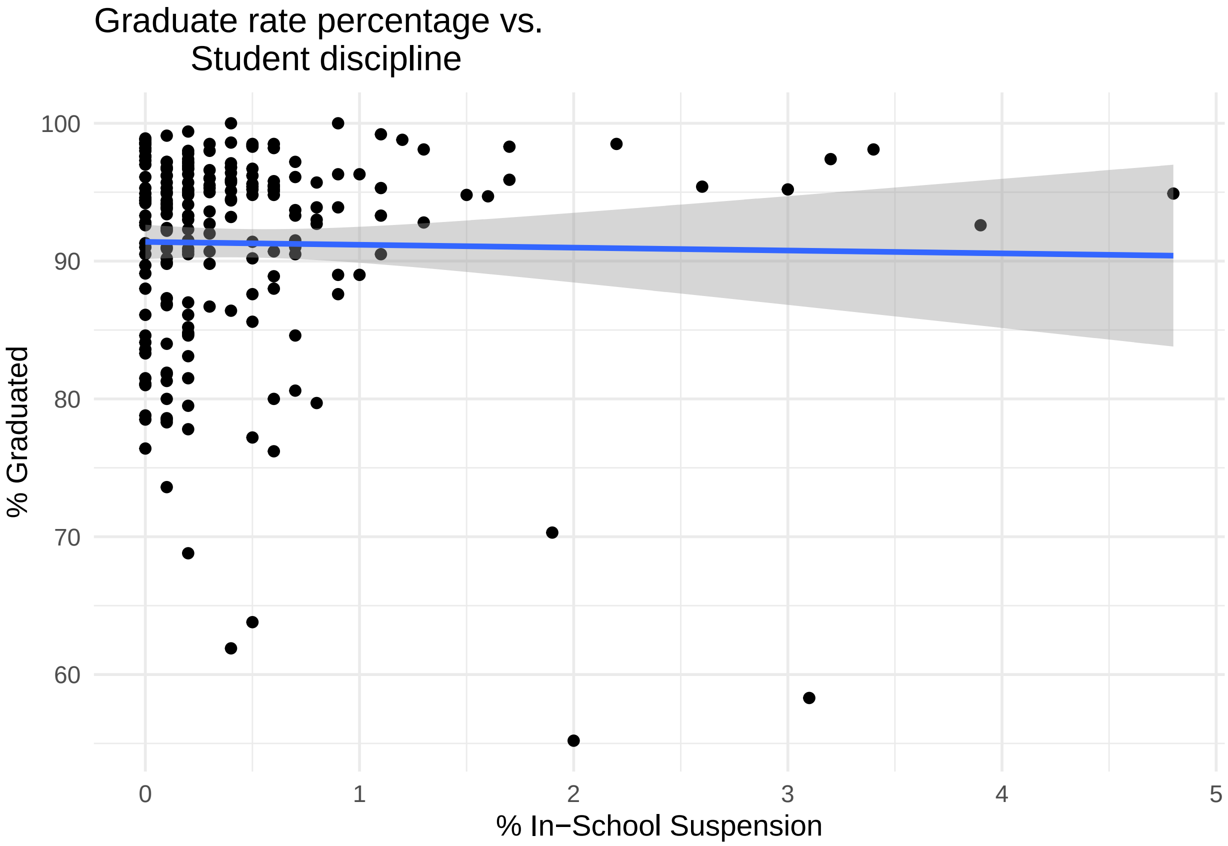
**2. Teacher Data vs. Graduation Rate**





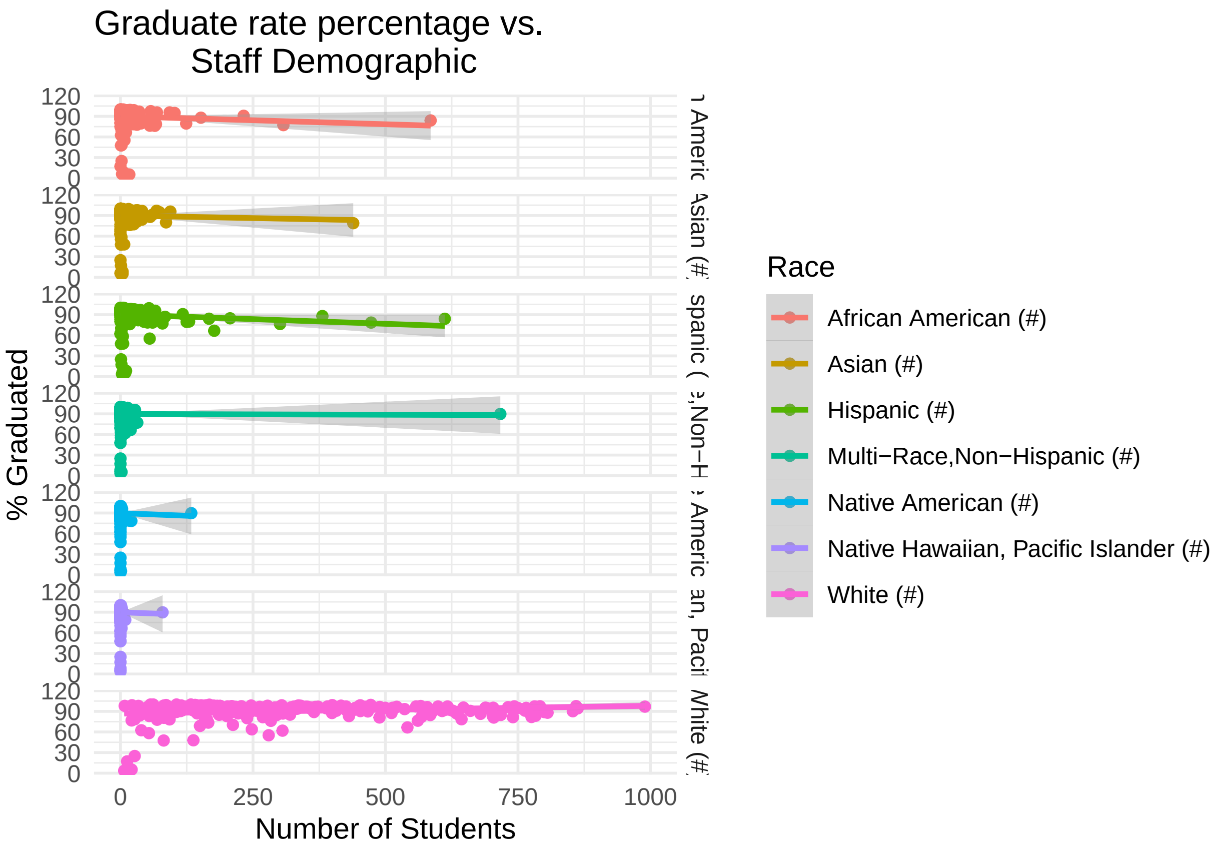
Observation: there may be a correlation; the more experienced teacher, the higher graduation rate. Observation: there may be a correlation; the higher student/teacher ratio, the lower graduation rate.

### 3. Student Discipline vs. Graduation Rate

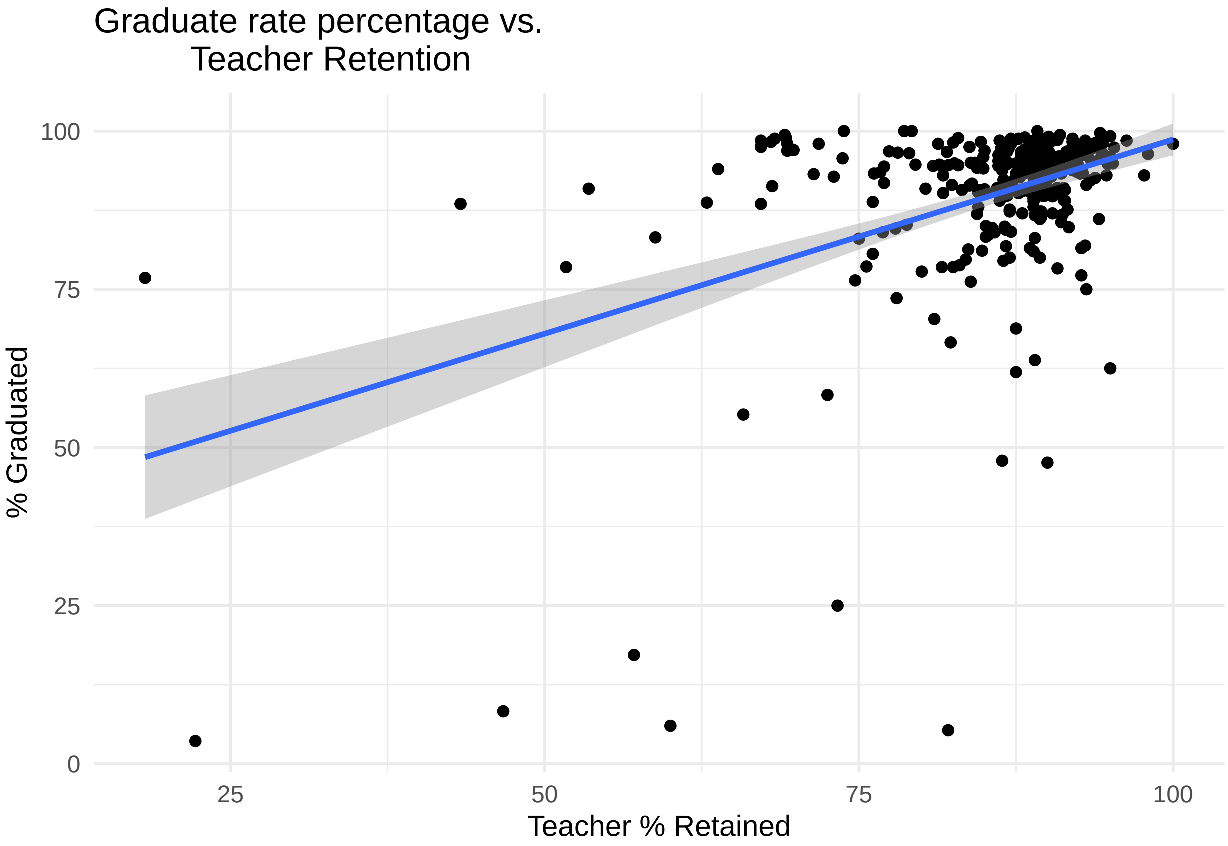


### 4. Demographic vs. Graduation Rate

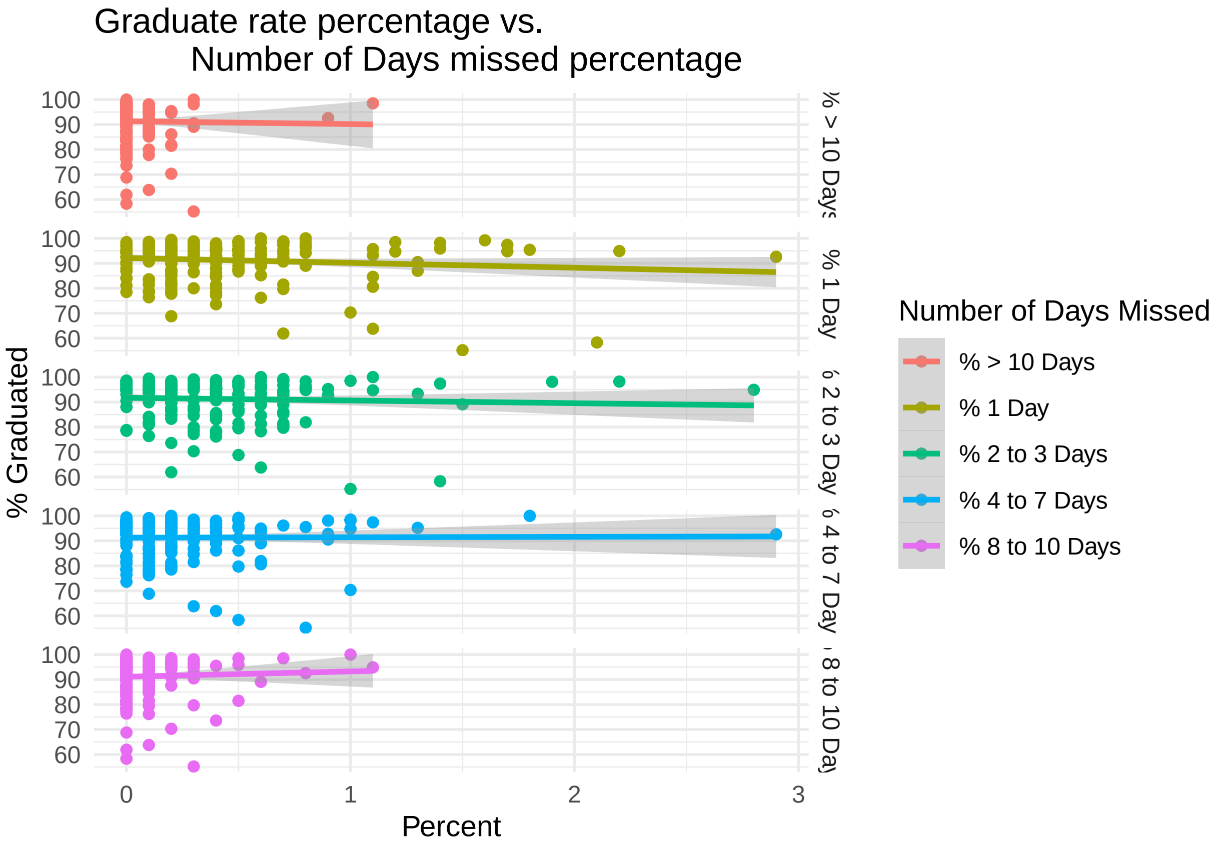
```
## # A tibble: 2,135 x 15
##    `District Name`  `District Code` `# in Cohort` `% Graduated` `% Still in Sc~`
##    <chr>            <chr>                   <dbl>         <dbl>            <dbl>
##  1 Abby Kelley Fos~ 04450000                   82          98.8              0
##  2 Abby Kelley Fos~ 04450000                   82          98.8              0
##  3 Abby Kelley Fos~ 04450000                   82          98.8              0
##  4 Abby Kelley Fos~ 04450000                   82          98.8              0
##  5 Abby Kelley Fos~ 04450000                   82          98.8              0
##  6 Abby Kelley Fos~ 04450000                   82          98.8              0
##  7 Abby Kelley Fos~ 04450000                   82          98.8              0
##  8 Abington         00010000                  163          93.3              2.5
##  9 Abington         00010000                  163          93.3              2.5
## 10 Abington         00010000                  163          93.3              2.5
## # ... with 2,125 more rows, and 10 more variables:
## #   `% Non-Grad Completers` <dbl>, `% H.S. Equiv.` <dbl>,
## #   `% Dropped Out` <dbl>, `% Permanently Excluded` <dbl>,
## #   `District/School Name` <chr>, `Females (#)` <dbl>, `Males (#)` <dbl>,
## #   `FTE Count` <dbl>, Race <chr>, `Number of Students` <dbl>
```


Graduate rate percentage vs.
Staff Demographic
% Graduated
Number of Students
0
250
500
750
1000
120
90
60
30
0
ı Americ
Asian (#
spanic (
e,Non−H
Americ
an, Paci
White (#
Race
African American (#)
Asian (#)
Hispanic (#)
Multi−Race,Non−Hispanic (#)
Native American (#)
Native Hawaiian, Pacific Islander (#)
White (#)

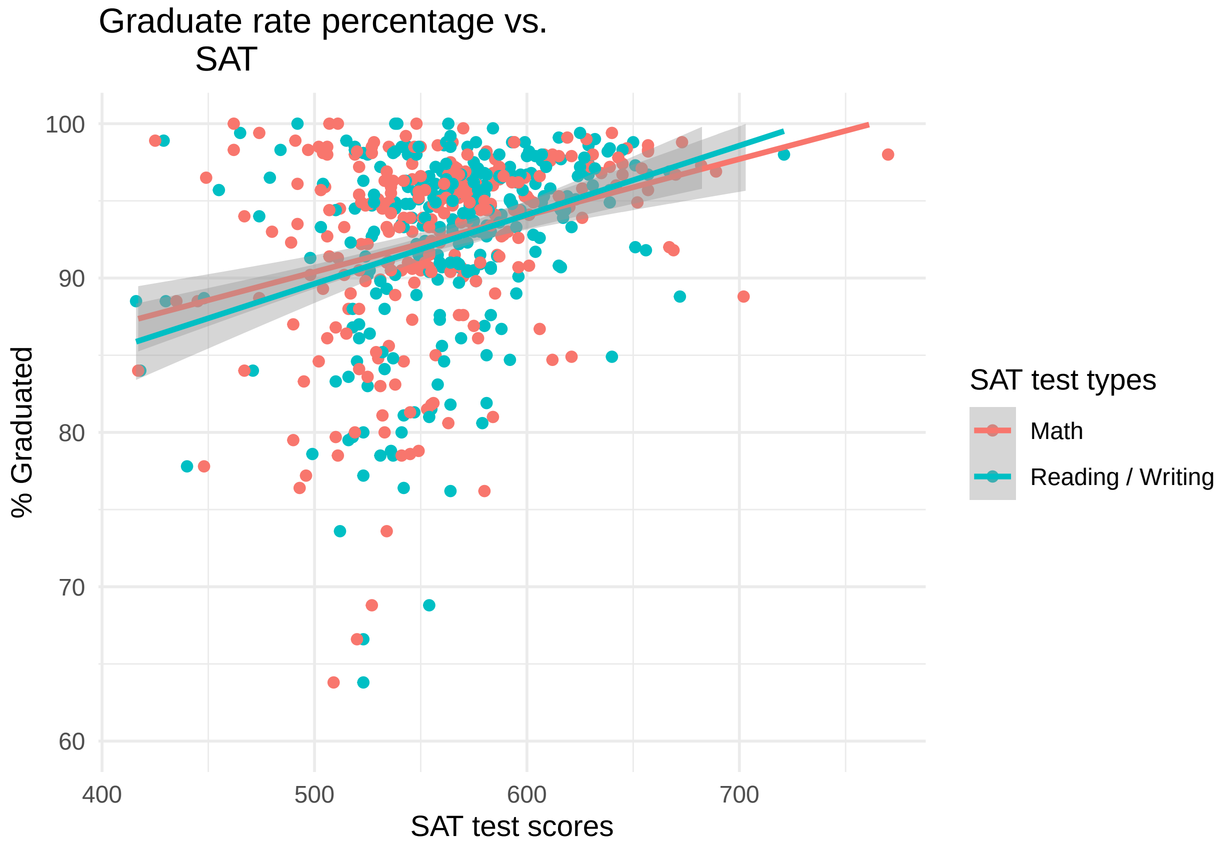
**5. Staffing Retention vs. Graduation Rate**



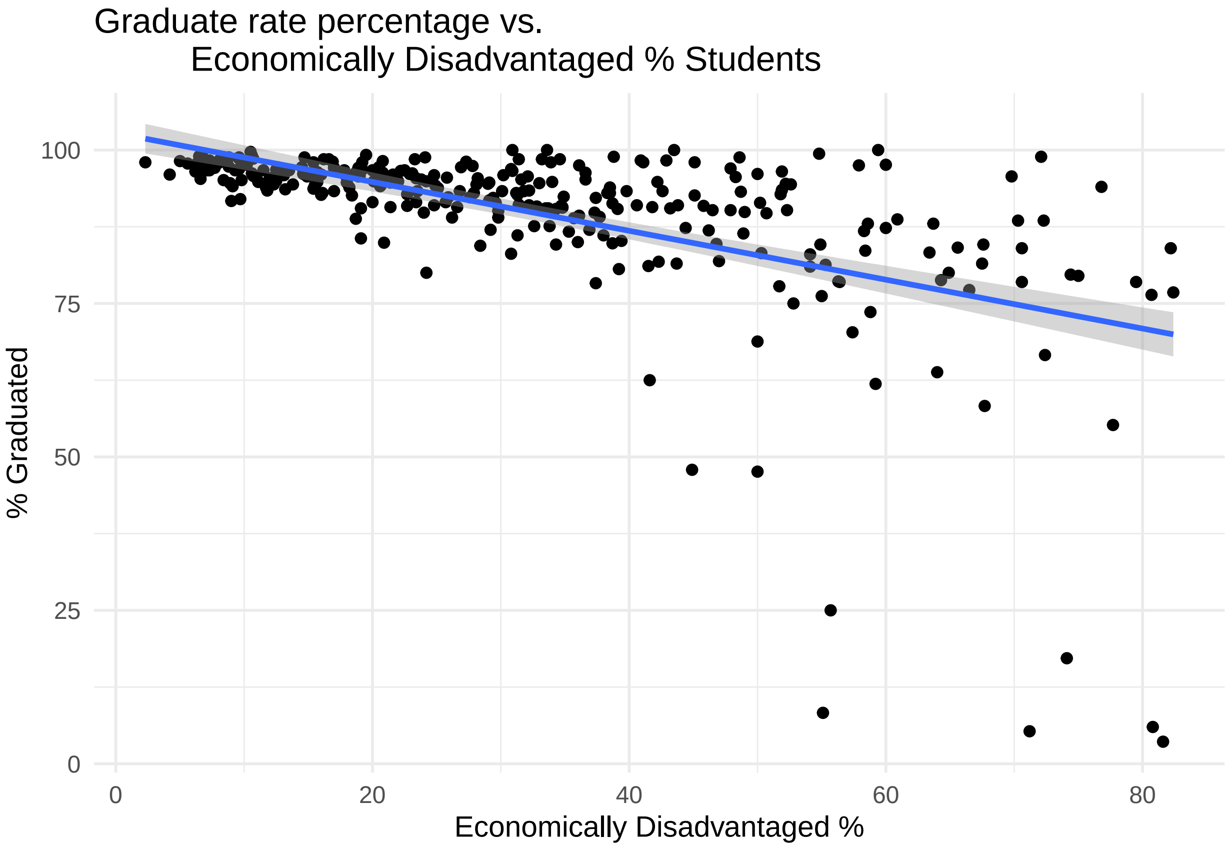
### 6. Graduation Rate vs. Day missed

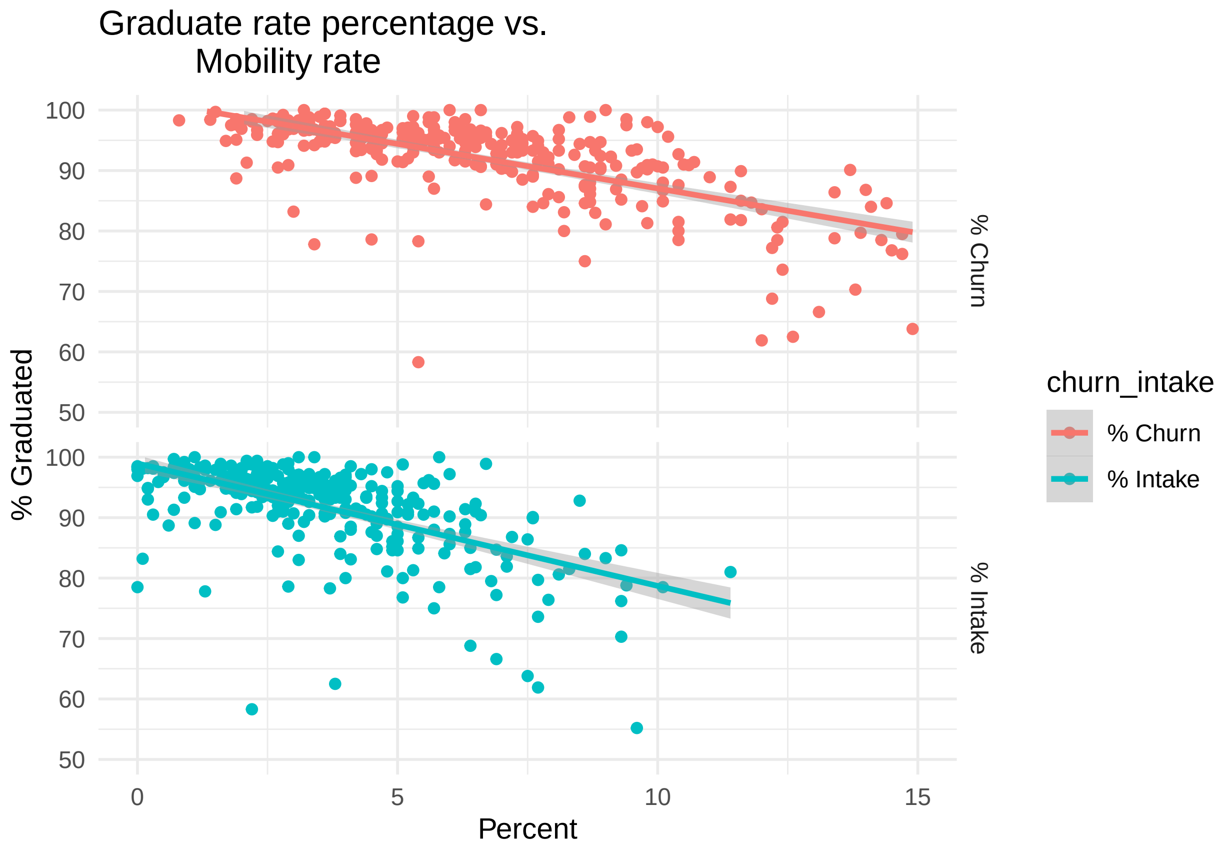


### 7. SAT vs. Graduation Rate



### 9. Graduation rate vs. Students Background





## Cleaning data for joining

```
sat <- sat %>% mutate(`Total Score` = `Reading / Writing` + Math) %>%
  select(!Writing)
expend <- expend %>% select(`District Code`, `Total Expenditures per Pupil`)
expend$`Total Expenditures per Pupil` <- parse_number(expend$`Total Expenditures per
↪  Pupil`)
enroll <- enroll %>%
  mutate(`HS Enrollment` = `9` + `10` + `11` + `12`) %>%
  select(`District Code`, `HS Enrollment`, Total) %>%
  rename(Enrollment = Total)
ap <- ap %>% select(`District Code`, `Tests Takers`)
reten <- reten %>% select(`District Code`, `Teacher % Retained`) %>%
  rename(`Teacher Retention Rate` = `Teacher % Retained`)
salary <- salary %>% select(!`District Name`)
salary$`Average Salary` <- parse_number(salary$`Average Salary`)
salary$`Salary Totals` <- parse_number(salary$`Salary Totals`)
classSize <- classSize %>% select(`District Code`, `Average Class Size`)
college <- college %>%
  select(`District Code`, `Attending Coll./Univ. (%)`)
attendance <- attendance %>% select(`District Code`, `Attendance Rate`, `Average # of
↪  Absences`)
attrition <- attrition %>% select(`District Code`, ALL) %>%
  rename(Attrition = ALL)
advCourse <- advCourse %>%
  select(`District Code`, `% Students Completing Advanced`, `% Math`, `% ELA`) %>%
  rename(`Adv Course % Math` = `% Math`, `Adv Course % ELA` = `% ELA`)
gradRate <- gradRate %>% select(`District Code`, `% Graduated`)
art <- art %>%
  mutate(`% in an Art Course` = `All Grades` / `Total Students` * 100) %>%
  select(`District Code`, `% in an Art Course`)
eduAge <- eduAge %>%
  mutate(`% of Teachers <40` = (`<26 yrs (# )` + `26-32 yrs (#)` + `33-40 yrs (#)`) /
  ↪   `FTE Count` * 100) %>%
  select(`District Code`, `% of Teachers <40`)
discipline <- discipline %>%
  mutate(`% Disciplined` = `Students Disciplined` / `Students` * 100) %>%
  select(`District Code`, `% Disciplined`)
eduGen <- eduGen %>%
  mutate(`% Female Teachers` = `Females (#)` / `FTE Count` * 100) %>%
  select(`District/School Code`, `% Female Teachers`) %>%
  rename(`District Code` = `District/School Code`)
teachData <- teachData %>% select(!`District Name`)
teachData$`Student / Teacher Ratio` <- substr(teachData$`Student / Teacher
↪  Ratio`,1,nchar(teachData$`Student / Teacher Ratio`)-5) %>% parse_number()

pop <- pop %>% select(`District Code`, `English Language Learner %`, `Students with
↪  Disabilities %`, `Economically Disadvantaged %` )

eth <- eth %>% select(-`District Name`, -`Total # of Classes`, -`Average Class Size`,
↪  -`Number of Students`)

dropout <- dropout %>% select(`District Code`, `% Dropout All Grades`)

mobile <- mobile %>% select(!`District Name`)

teachProg <- teachProg %>%
  mutate(`Gen Ed %` = `General Education (#)`/ `FTE Count`) %>%
  select(`District Code`, `Gen Ed %`)

selectPop <- selectPop %>% select(!c(`District Name`, `Free Lunch #`, `Free Lunch %`,
↪  `Reduced Lunch #`, `Reduced Lunch %`, `Economically Disadvantaged #`, `Economically
↪  Disadvantaged %`, `English Language Learner %`, `Students With Disabilities %`))
```

## Joining Data

```
eduData <- inner_join(sat, expend, by = "District Code") %>%
  inner_join(salary, by = "District Code") %>%
  inner_join(enroll, by = "District Code") %>%
  inner_join(ap, by = "District Code") %>%
  inner_join(reten, by = "District Code") %>%
  inner_join(classSize, by = "District Code") %>%
  inner_join(college, by = "District Code") %>%
  inner_join(attendance, by = "District Code") %>%
  inner_join(attrition, by = "District Code") %>%
  inner_join(advCourse, by = "District Code") %>%
  inner_join(gradRate, by = "District Code") %>%
  inner_join(art, by = "District Code") %>%
  inner_join(eduAge, by = "District Code") %>%
  inner_join(discipline, by = "District Code") %>%
  inner_join(eduGen, by = "District Code") %>%
  inner_join(teachData, by = "District Code") %>%
  inner_join(teachProg, by = "District Code") %>%
  inner_join(selectPop, by = "District Code") %>%
  mutate(`Percent of HS in AP` = `Tests Takers` / `HS Enrollment` * 100) %>%
  mutate(`Adjusted Score` = `Total Score` * `% Graduated` / 100)

eduData$`District Code` <- as.double(eduData$`District Code`)

eduData <- eduData %>%
  inner_join(pop, by = "District Code") %>%
  inner_join(eth, by = "District Code") %>%
  inner_join(dropout, by = "District Code") %>%
  inner_join(mobile, by = "District Code")
```

## Aditional EDA

```
summary(eduData)
```

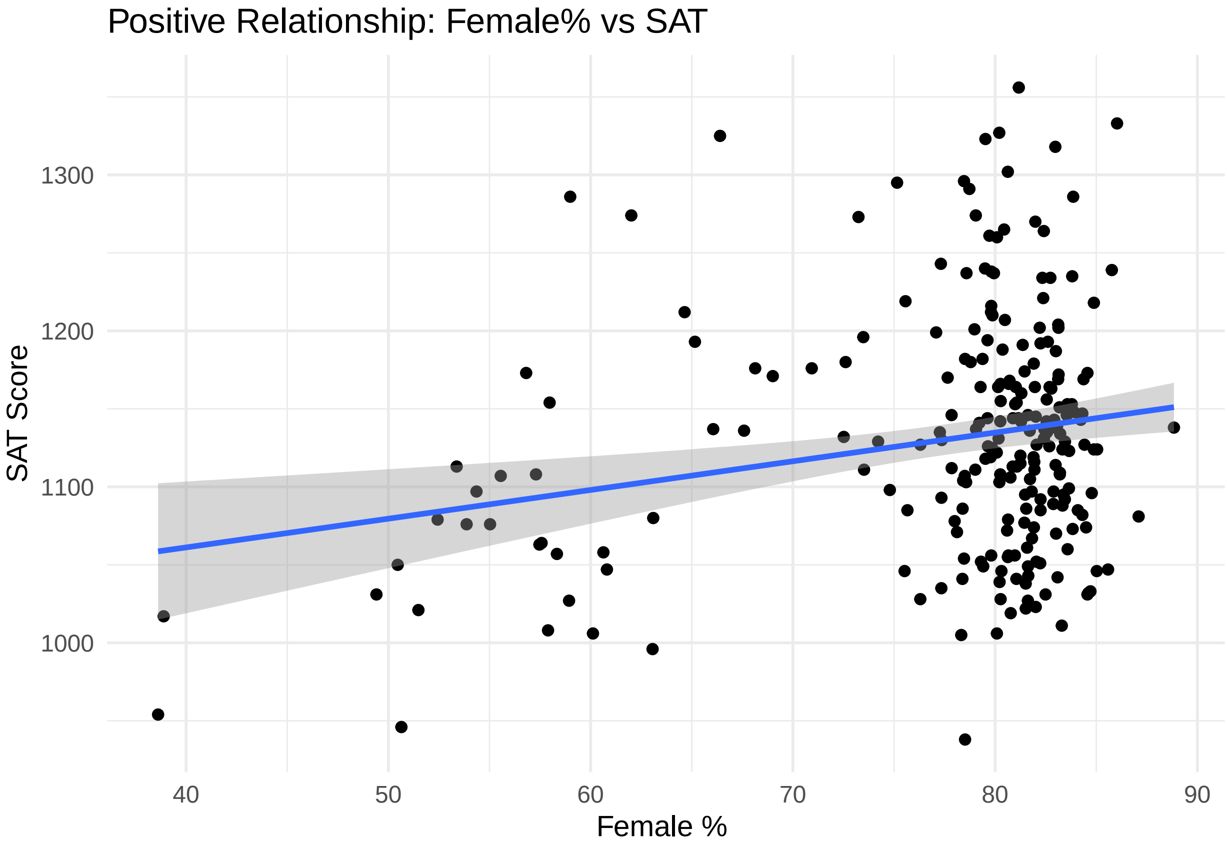
```
##  District Name      District Code      Tests Taken     Reading / Writing
##  Length:252         Min.   :  10000   Min.   :   1.0   Min.   :471.0
##  Class :character   1st Qu.:1377500   1st Qu.:  46.0   1st Qu.:543.0
##  Mode  :character   Median :2515000   Median : 104.0   Median :567.0
##                     Mean   :3556349   Mean   : 164.0   Mean   :568.8
##                     3rd Qu.:6585000   3rd Qu.: 216.5   3rd Qu.:592.0
##                     Max.   :9150000   Max.   :2299.0   Max.   :667.0
##                                                        NA's   :15
##       Math         Total Score   Total Expenditures per Pupil
##  Min.   :462.0   Min.   : 938   Min.   :12216
##  1st Qu.:533.0   1st Qu.:1076   1st Qu.:15528
##  Median :558.0   Median :1126   Median :17022
##  Mean   :561.7   Mean   :1130   Mean   :17746
##  3rd Qu.:585.0   3rd Qu.:1173   3rd Qu.:19303
##  Max.   :689.0   Max.   :1356   Max.   :34027
##  NA's   :15      NA's   :15
##  Salary Totals       Average Salary     FTE Count       HS Enrollment
##  Min.   :  2615403   Min.   : 62224   Min.   :  35.2   Min.   :  102
##  1st Qu.:  9051113   1st Qu.: 76252   1st Qu.: 113.3   1st Qu.:  509
##  Median : 14860386   Median : 82257   Median : 182.2   Median :  798
##  Mean   : 22864194   Mean   : 82381   Mean   : 270.3   Mean   : 1066
##  3rd Qu.: 26101261   3rd Qu.: 87847   3rd Qu.: 307.9   3rd Qu.: 1284
##  Max.   :463706080   Max.   :110665   Max.   :4406.4   Max.   :14021
##
##    Enrollment     Tests Takers     Teacher Retention Rate Average Class Size
##  Min.   :  391   Min.   :   1.00   Min.   :65.80          Min.   : 8.20
##  1st Qu.: 1322   1st Qu.:  73.75   1st Qu.:86.38          1st Qu.:13.50
##  Median : 2214   Median : 149.50   Median :89.00          Median :15.60
##  Mean   : 3278   Mean   : 194.79   Mean   :88.32          Mean   :15.34
##  3rd Qu.: 3847   3rd Qu.: 250.50   3rd Qu.:91.22          3rd Qu.:17.23
##  Max.   :46169   Max.   :3161.00   Max.   :98.00          Max.   :21.70
##
```

```
##  Attending Coll./Univ. (%) Attendance Rate Average # of Absences
##  Min.   :15.20             Min.   :79.90   Min.   : 3.000
##  1st Qu.:56.25             1st Qu.:92.60   1st Qu.: 6.575
##  Median :69.35             Median :94.60   Median : 8.950
##  Mean   :66.37             Mean   :94.06   Mean   : 9.819
##  3rd Qu.:79.10             3rd Qu.:96.00   3rd Qu.:12.325
##  Max.   :88.30             Max.   :98.20   Max.   :33.600
##
##    Attrition      % Students Completing Advanced Adv Course % Math
##  Min.   : 1.400   Min.   : 17.00                 Min.   : 3.90
##  1st Qu.: 4.975   1st Qu.: 58.08                 1st Qu.:44.75
##  Median : 6.650   Median : 66.75                 Median :56.15
##  Mean   : 6.824   Mean   : 66.57                 Mean   :55.71
##  3rd Qu.: 8.125   3rd Qu.: 76.33                 3rd Qu.:65.80
##  Max.   :22.700   Max.   :100.00                 Max.   :99.60
##
##  Adv Course % ELA  % Graduated     % in an Art Course % of Teachers <40
##  Min.   : 0.00    Min.   : 55.20   Min.   : 0.00      Min.   :18.49
##  1st Qu.:10.18    1st Qu.: 90.05   1st Qu.:71.21      1st Qu.:32.16
##  Median :15.90    Median : 94.15   Median :82.36      Median :36.51
##  Mean   :17.99    Mean   : 91.80   Mean   :73.47      Mean   :36.92
##  3rd Qu.:23.12    3rd Qu.: 96.33   3rd Qu.:86.88      3rd Qu.:41.18
##  Max.   :94.30    Max.   :100.00   Max.   :97.67      Max.   :60.61
##
##  % Disciplined    % Female Teachers Total # of Teachers (FTE)
##  Min.   :0.0000   Min.   :38.61     Min.   :  35.2
##  1st Qu.:0.2657   1st Qu.:77.96     1st Qu.: 110.8
##  Median :0.6511   Median :80.63     Median : 178.3
##  Mean   :0.9758   Mean   :77.45     Mean   : 270.0
##  3rd Qu.:1.2869   3rd Qu.:82.58     3rd Qu.: 308.3
##  Max.   :8.4746   Max.   :88.84     Max.   :4595.5
##
##  % of Teachers Licensed Student / Teacher Ratio Percent of Experienced Teachers
##  Min.   : 88.00         Min.   : 8.30           Min.   :57.60
##  1st Qu.: 98.80         1st Qu.:11.10           1st Qu.:83.47
##  Median : 99.55         Median :12.05           Median :87.40
##  Mean   : 99.01         Mean   :12.06           Mean   :86.42
##  3rd Qu.:100.00         3rd Qu.:13.00           3rd Qu.:90.12
##  Max.   :100.00         Max.   :16.00           Max.   :96.00
##
##  Percent of Teachers without Waiver or Provisional License
##  Min.   :71.80
##  1st Qu.:92.88
##  Median :94.95
##  Mean   :93.91
##  3rd Qu.:96.53
##  Max.   :99.30
##
##  Percent Teaching In-Field    Gen Ed %      First Language Not English #
##  Min.   : 83.30            Min.   :0.3050   Min.   :    0.0
##  1st Qu.: 94.20            1st Qu.:0.7880   1st Qu.:   53.5
##  Median : 96.30            Median :0.8392   Median :  148.0
##  Mean   : 95.66            Mean   :0.8128   Mean   :  784.4
##  3rd Qu.: 97.72            3rd Qu.:0.8993   3rd Qu.:  590.2
##  Max.   :100.00            Max.   :1.0000   Max.   :22227.0
##
##  First Language Not English % English Language Learner #
##  Min.   : 0.00                Min.   :    0.00
##  1st Qu.: 3.00                1st Qu.:   17.75
##  Median : 7.35                Median :   58.00
##  Mean   :13.41                Mean   :  366.62
##  3rd Qu.:18.52                3rd Qu.:  202.00
##  Max.   :84.60                Max.   :14038.00
##
##  Students With Disabilities #  Low Income #      Low Income %
##  Min.   :   72.0              Min.   :   67.0   Min.   : 5.80
##  1st Qu.:  246.8              1st Qu.:  319.2   1st Qu.:19.27
##  Median :  415.0              Median :  602.5   Median :32.85
##  Mean   :  634.2              Mean   : 1421.5   Mean   :35.05
##  3rd Qu.:  723.8              3rd Qu.: 1060.5   3rd Qu.:46.30
##  Max.   :10167.0              Max.   :32854.0   Max.   :88.80
##
##  High Needs #...15 High Needs #...16 Percent of HS in AP Adjusted Score
##  Min.   :  164.0   Min.   :19.40     Min.   : 0.07593    Min.   : 658.4
##  1st Qu.:  565.0   1st Qu.:33.60     1st Qu.:13.50227    1st Qu.: 987.3
##  Median :  892.5   Median :45.00     Median :19.68825    Median :1047.6
##  Mean   : 1832.2   Mean   :47.48     Mean   :19.95138    Mean   :1045.5
##  3rd Qu.: 1567.2   3rd Qu.:58.55     3rd Qu.:26.97325    3rd Qu.:1113.4
##  Max.   :37940.0   Max.   :94.10     Max.   :45.28302    Max.   :1314.0
##                                                          NA's   :15
##  English Language Learner % Students with Disabilities %
##  Min.   : 0.000             Min.   : 9.80
##  1st Qu.: 1.200             1st Qu.:16.50
##  Median : 2.500             Median :18.30
##  Mean   : 5.315             Mean   :18.82
##  3rd Qu.: 5.825             3rd Qu.:20.73
##  Max.   :37.400             Max.   :40.60
##
##  Economically Disadvantaged % African American %    Asian %
##  Min.   : 4.20                Min.   : 0.100     Min.   : 0.000
##  1st Qu.:15.57                1st Qu.: 1.475     1st Qu.: 1.200
##  Median :25.85                Median : 2.650     Median : 2.200
##  Mean   :29.14                Mean   : 4.680     Mean   : 5.213
##  3rd Qu.:38.50                3rd Qu.: 5.125     3rd Qu.: 5.925
##  Max.   :82.20                Max.   :60.900     Max.   :41.900
##
##    Hispanic %       White %      Native American %
##  Min.   : 1.30   Min.   : 3.20   Min.   :0.0000
##  1st Qu.: 5.20   1st Qu.:63.45   1st Qu.:0.1000
##  Median : 7.70   Median :78.20   Median :0.1000
##  Mean   :13.29   Mean   :72.50   Mean   :0.2492
##  3rd Qu.:14.05   3rd Qu.:86.22   3rd Qu.:0.3000
##  Max.   :93.80   Max.   :96.50   Max.   :5.6000
##
##  Native Hawaiian, Pacific Islander % Multi-Race, Non-Hispanic %
##  Min.   :0.00000                     Min.   : 0.400
##  1st Qu.:0.00000                     1st Qu.: 2.700
##  Median :0.10000                     Median : 3.700
##  Mean   :0.09286                     Mean   : 4.016
##  3rd Qu.:0.10000                     3rd Qu.: 4.900
##  Max.   :2.40000                     Max.   :12.900
##
##  % Dropout All Grades Churn/Intake Enroll    % Churn          % Intake
##  Min.   :0.000        Min.   :  405       Min.   : 1.400   Min.   : 0.000
##  1st Qu.:0.300        1st Qu.: 1340       1st Qu.: 4.475   1st Qu.: 2.300
##  Median :0.800        Median : 2266       Median : 6.500   Median : 3.300
##  Mean   :1.267        Mean   : 3415       Mean   : 7.031   Mean   : 3.745
##  3rd Qu.:1.800        3rd Qu.: 4005       3rd Qu.: 8.900   3rd Qu.: 4.925
##  Max.   :9.700        Max.   :50831       Max.   :20.100   Max.   :11.400
##
##  Stability Enroll  % Stability
##  Min.   :  400    Min.   :86.10
##  1st Qu.: 1321    1st Qu.:94.60
##  Median : 2194    Median :96.20
##  Mean   : 3297    Mean   :95.87
##  3rd Qu.: 3852    3rd Qu.:97.40
##  Max.   :48444    Max.   :99.20
##
```

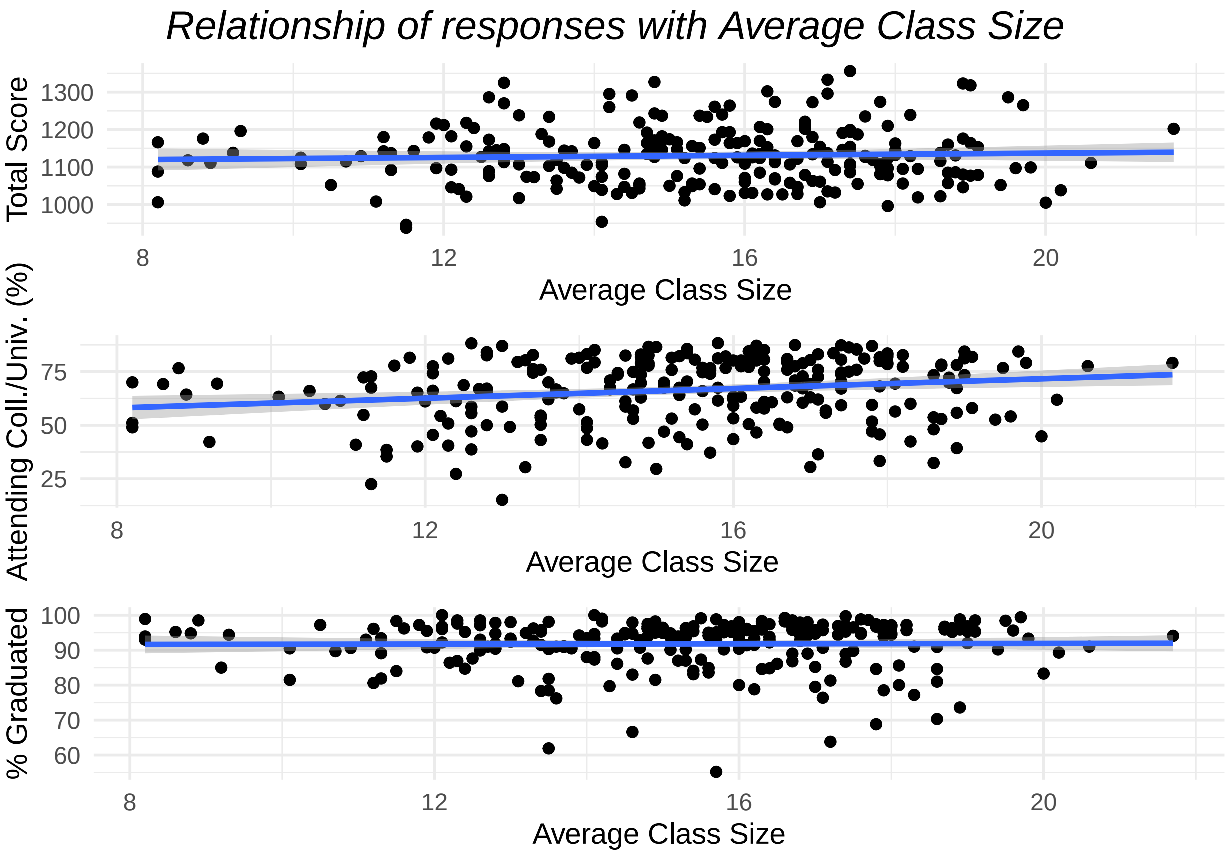
**Inference from summary:** 1) Neither Reading/Writing, nor Math has a perfect score in SAT, same goes for the total score.

2) Among races, at least one district had Hispanic and White students in domination,even though the third quartile of Hispanic students is at 14.05%.
3) At least one district/school has 100% graduate rate and 0% drop rate and yet maximum percentage of students going to College is only 88.30.
4) At least one district/school has % Students Completing Advanced as 100% and none has 100% attendance rate.
5) One school district has 71.80% of their teachers with no license or only a provisional license.
6) None of the school/district has **only** experienced teachers, and in at least one school, 42.4% teachers are not experienced.
7) The data is taken for the COVID time-period(2020-22), yet at least one school had Percent Teaching In-Field as 100%.
8) Even though the schools are in a country where English is the most-commonly spoken language, at least one school has 83.600% students whose first Language is not English.
9) None of the schools has 0% of High Needs or Economically Disadvantaged students.

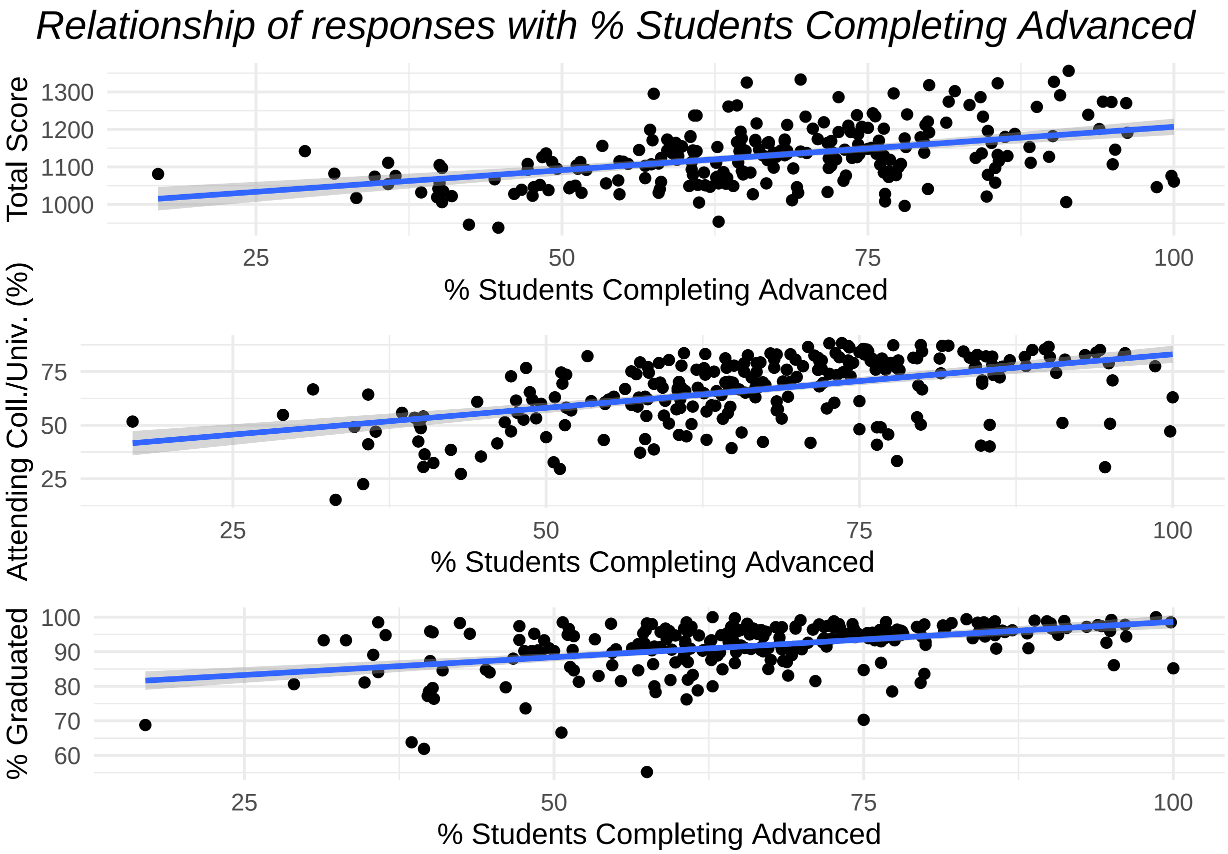
**To see if genders had a relation with Total SAT score.**



### Average Class Size vs SAT, Graduate Rate, and Enrollment in college

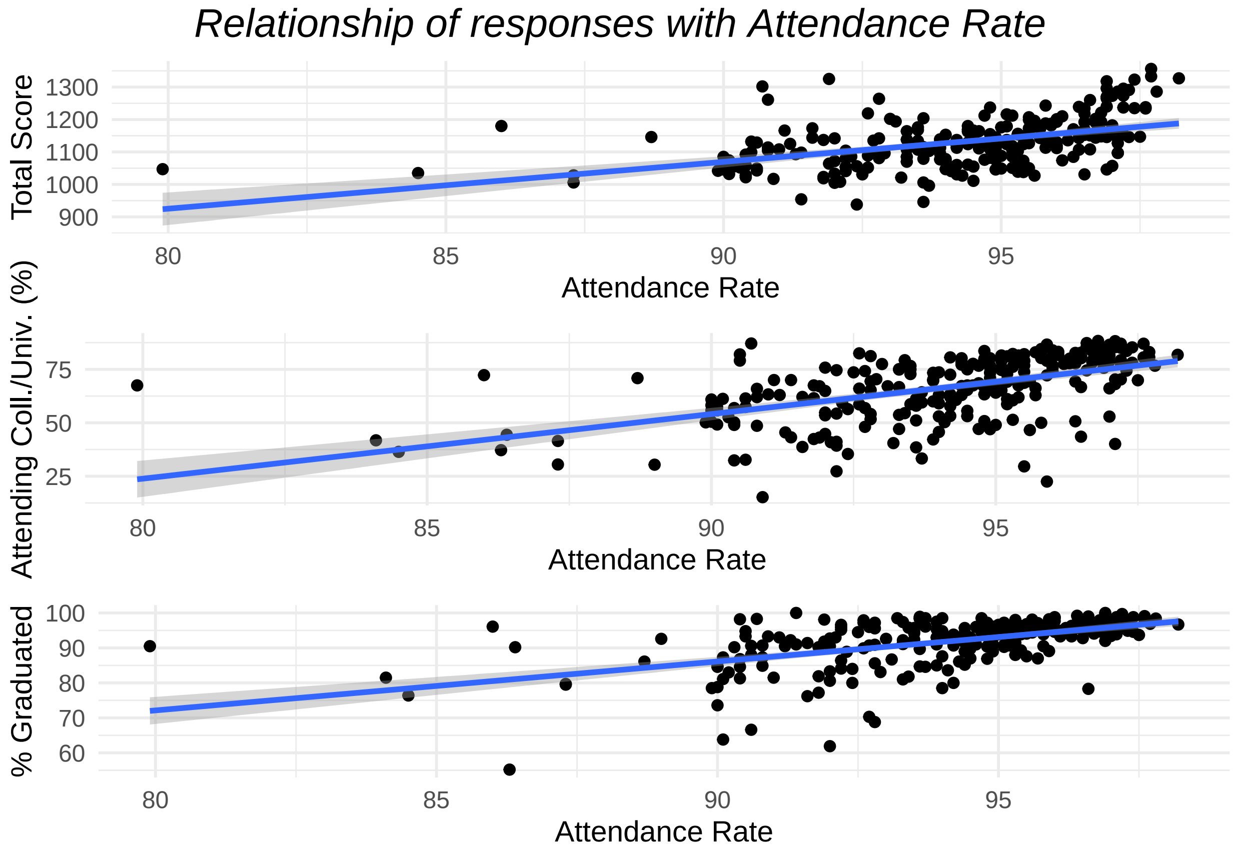


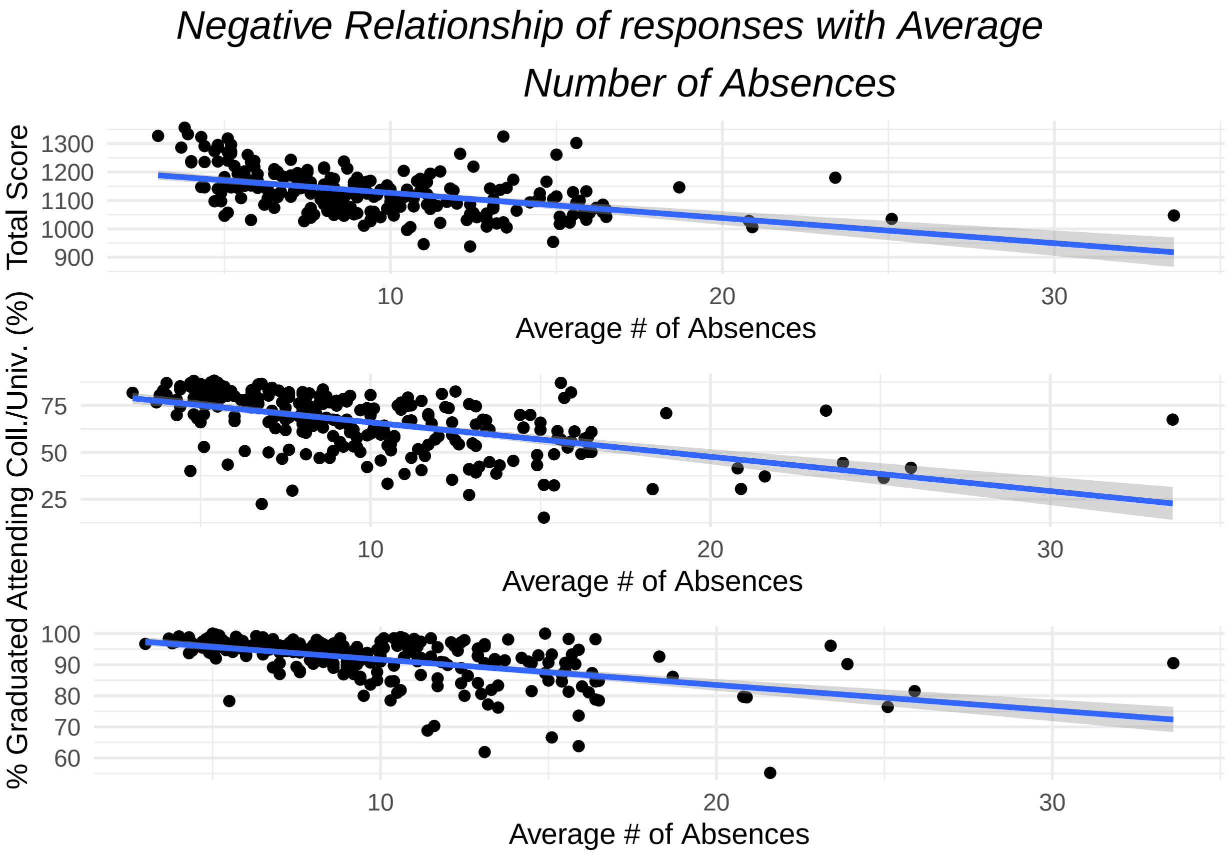
Average Class Size has negative relationship with SAT score and % graduated, while it has a positive relationship with % going to college.



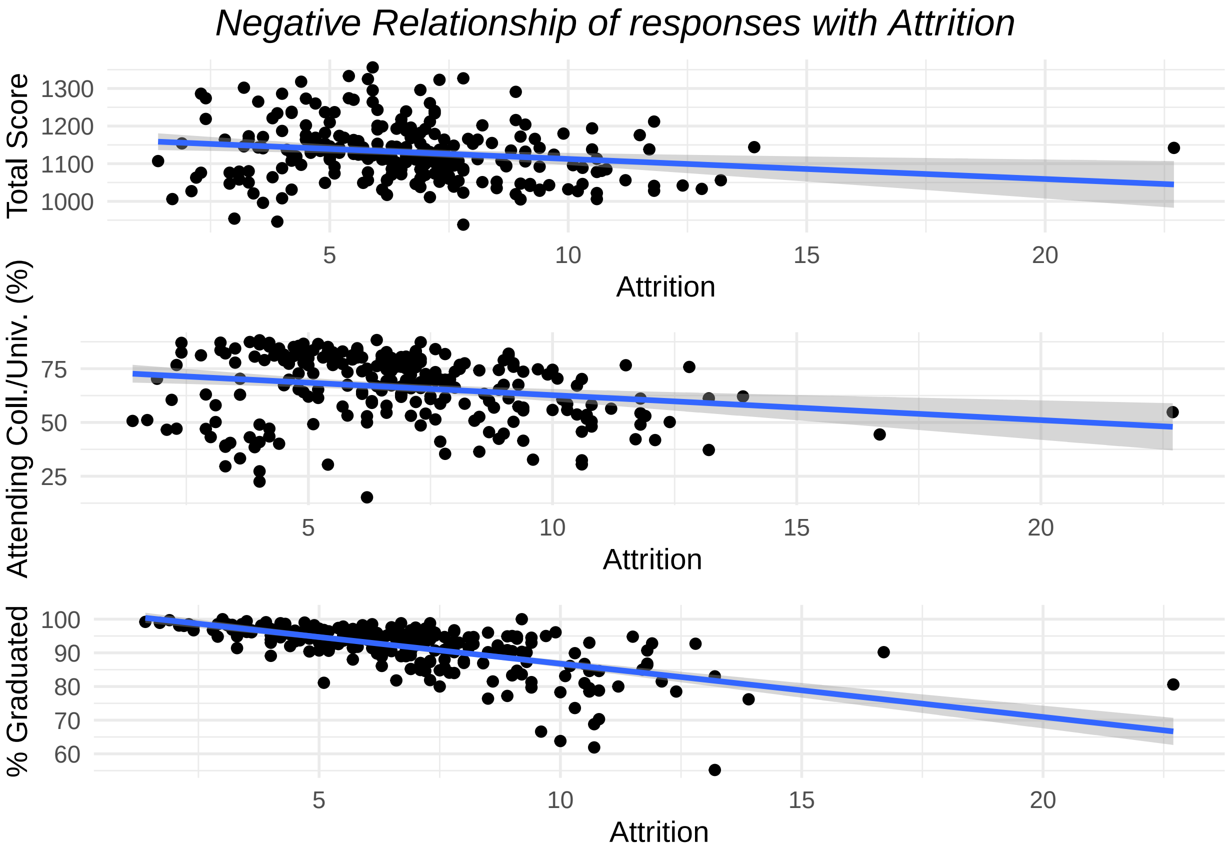
All responses had positive relationship with % Students Completing Advanced.

**Attendance Rate vs Responses**





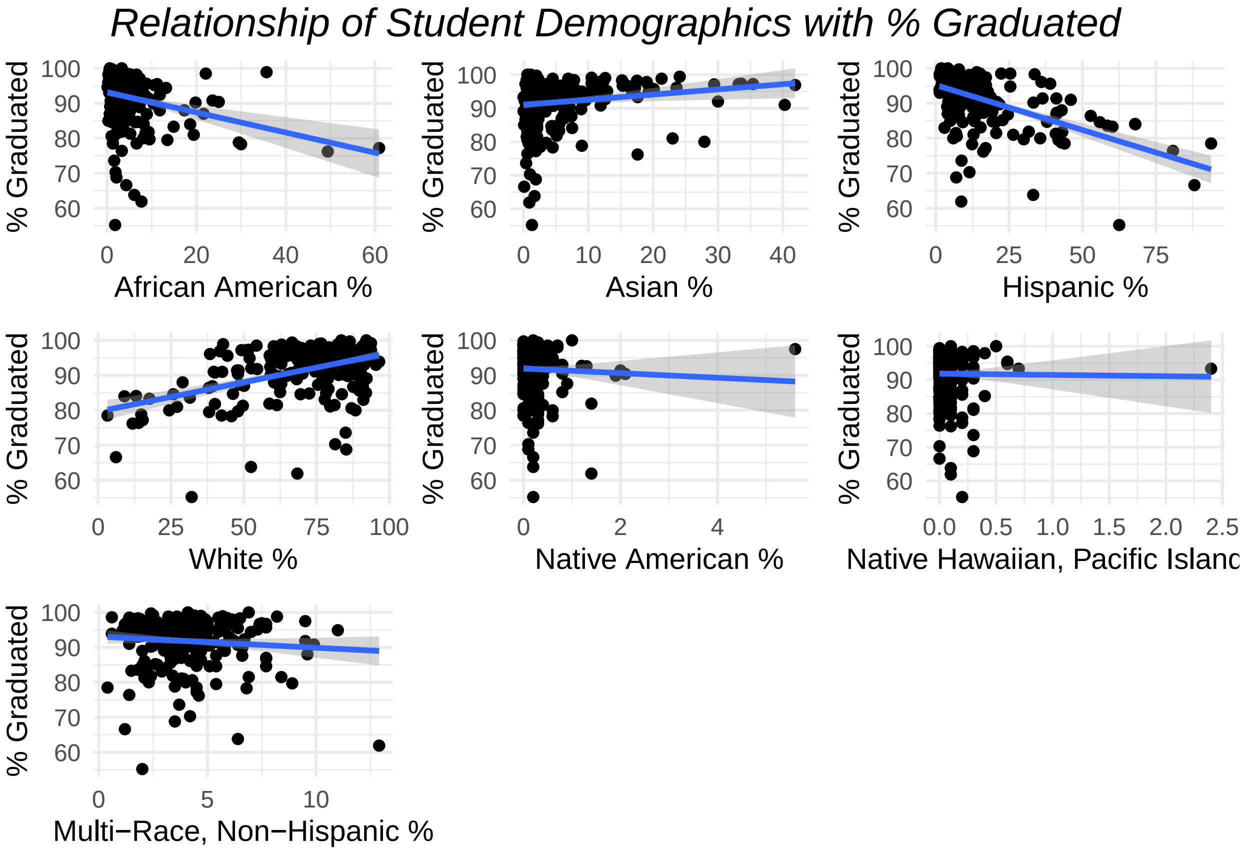
*Negative Relationship of responses with Average Number of Absences*

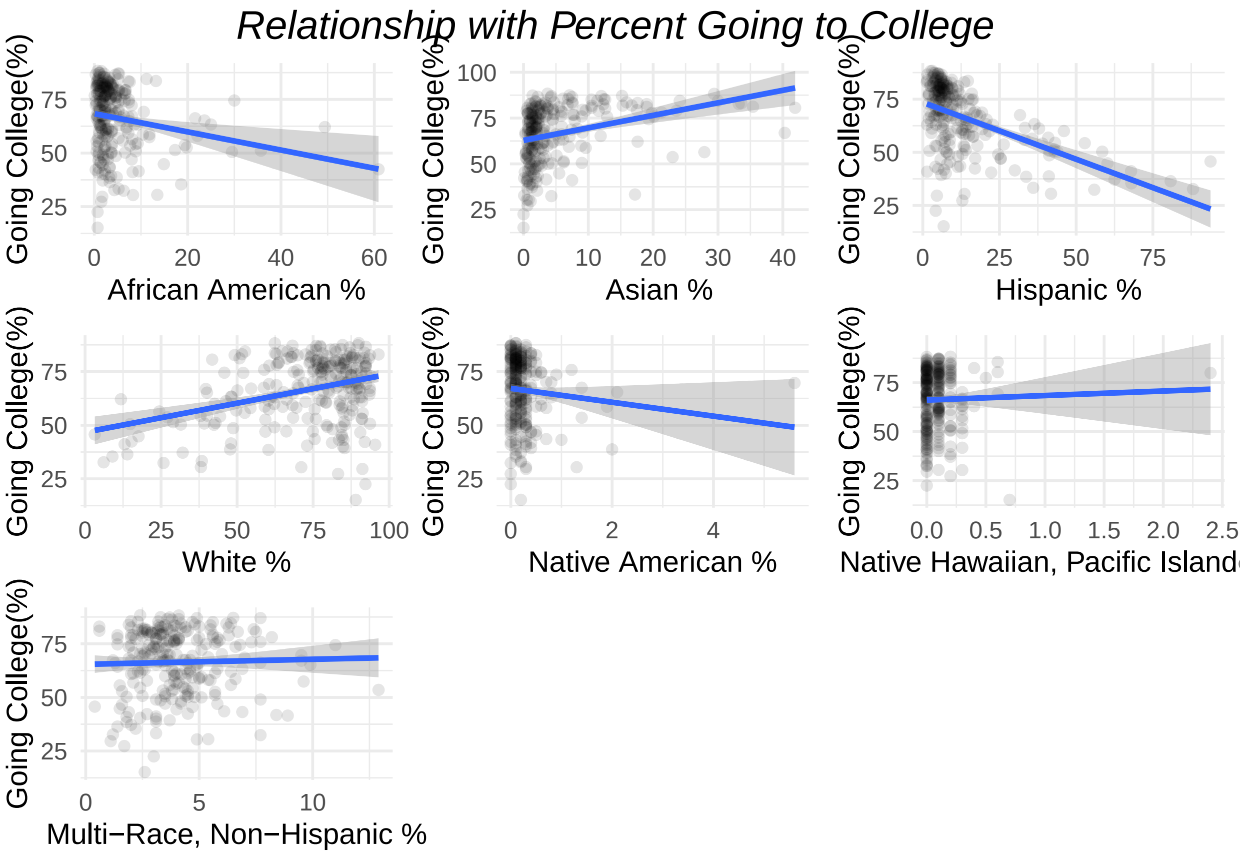


*Negative Relationship of responses with Attrition*

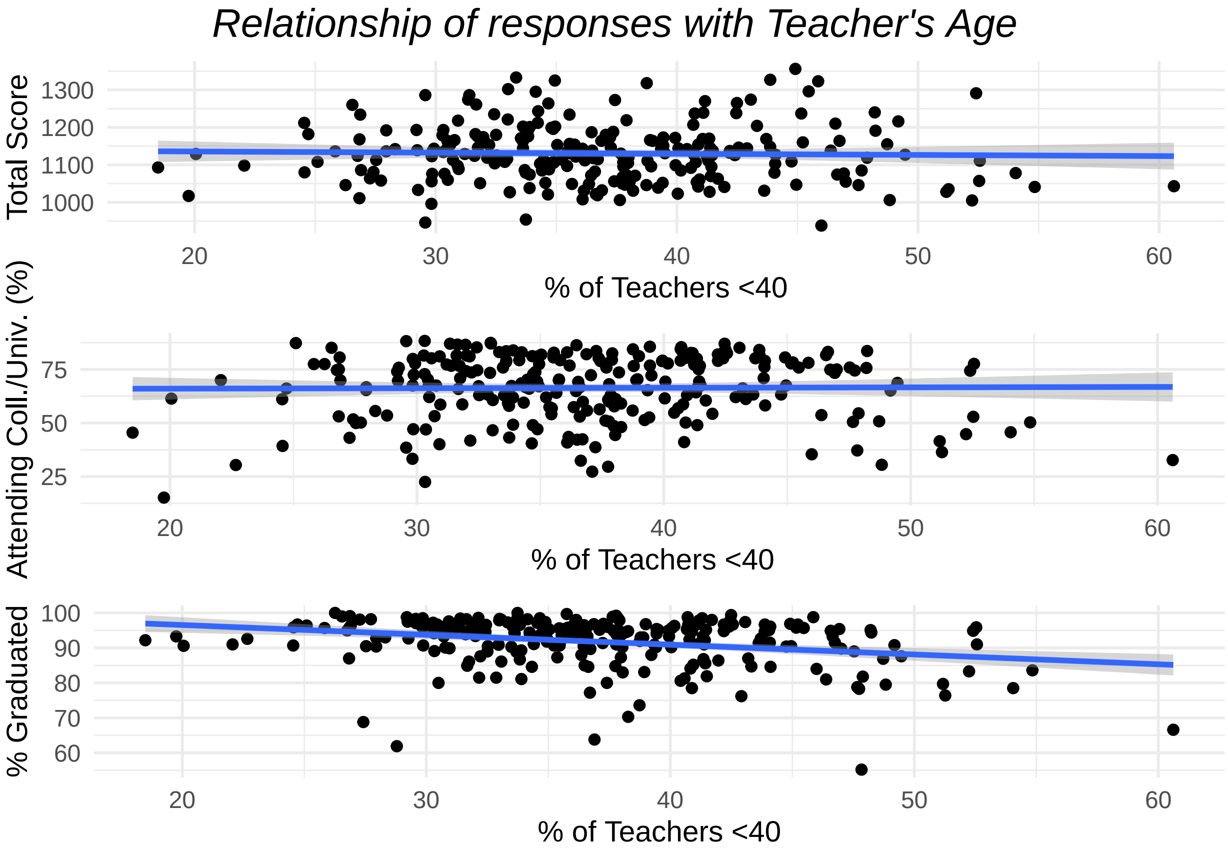
Even though, all of them has negative relationship, the slopes are different i.e., graduation rate drops with larger difference as compared to other responses.

### Student Background vs Graduation Rate, % Going to College

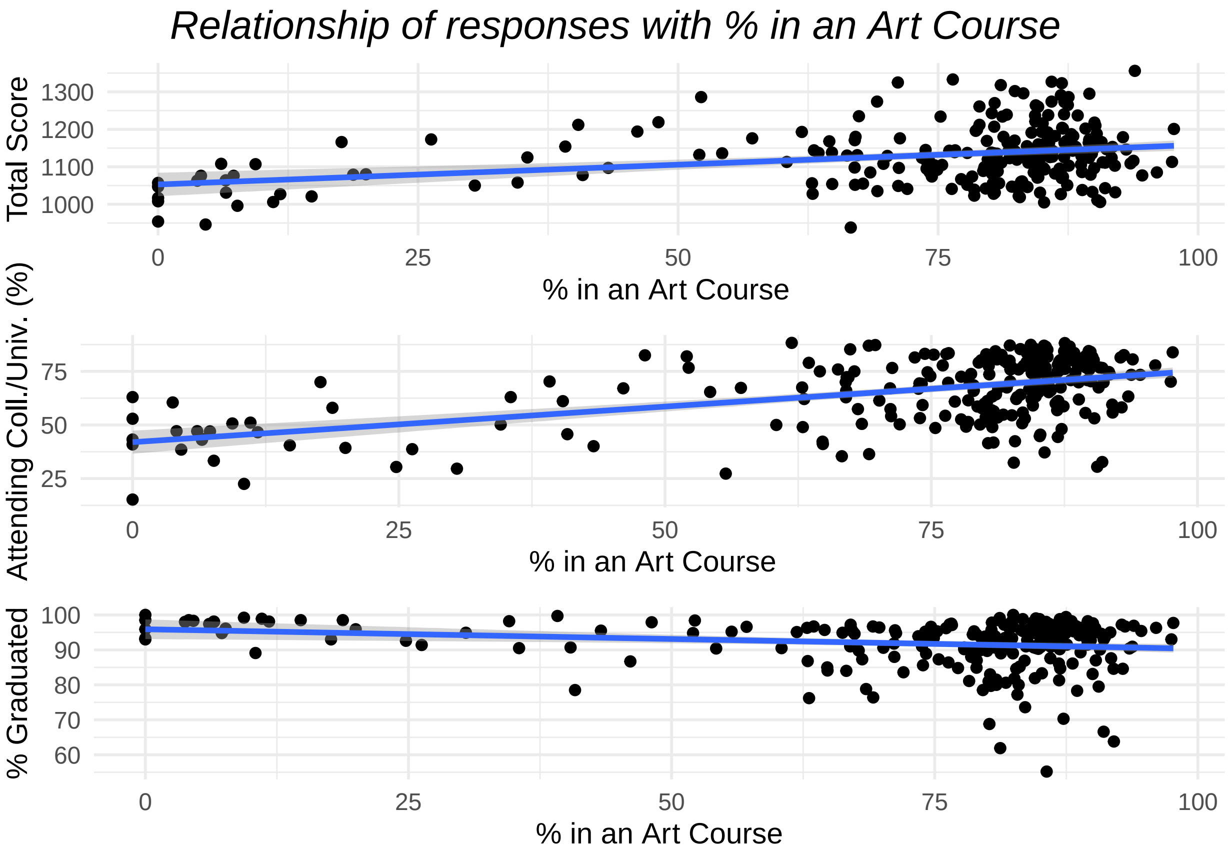


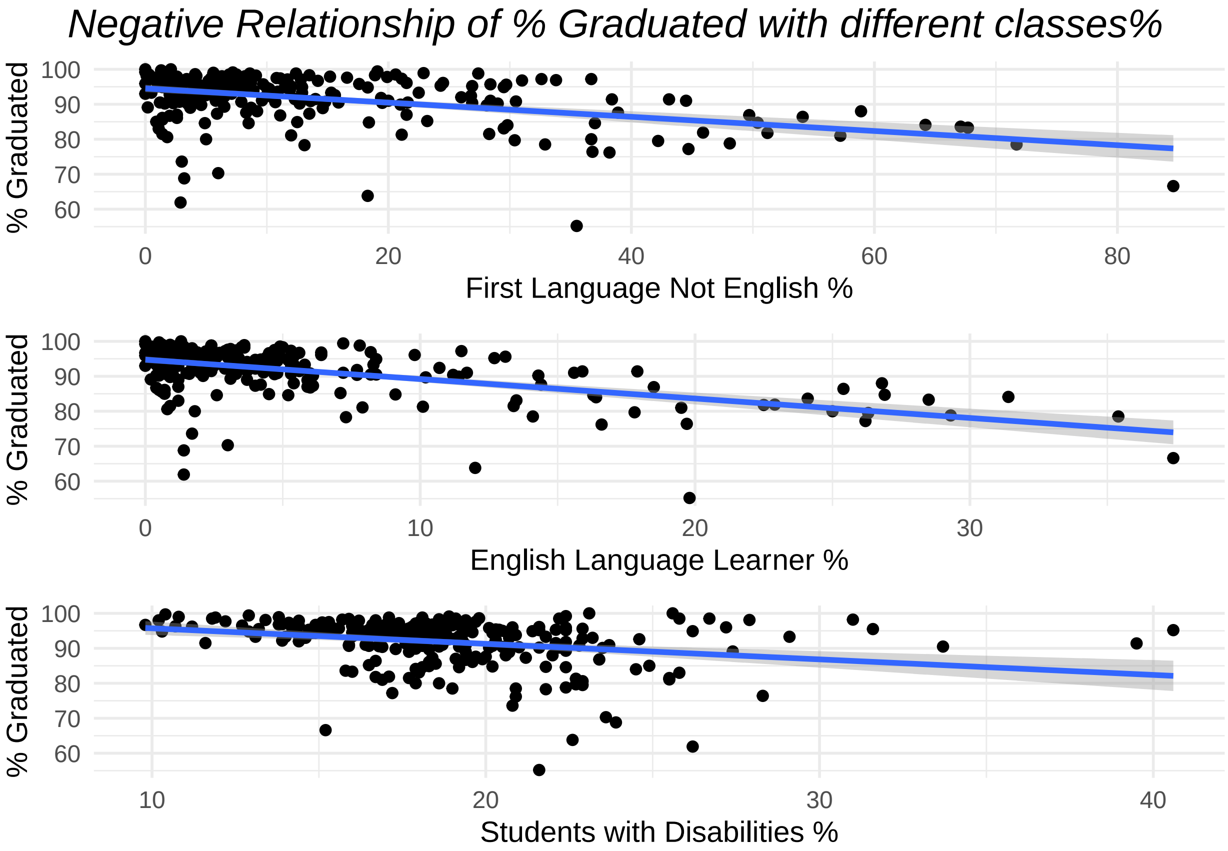

Relationship with Percent Going to College
Going College(%)
75
50
25
0
20
40
60
African American %
Going College(%)
100
75
50
25
0
10
20
30
40
Asian %
Going College(%)
75
50
25
0
25
50
75
Hispanic %
Going College(%)
75
50
25
0
25
50
75
100
White %
Going College(%)
75
50
25
0
2
4
Native American %
Going College(%)
75
50
25
0.0
0.5
1.0
1.5
2.0
2.5
Native Hawaiian, Pacific Islande
Going College(%)
75
50
25
0
5
10
Multi−Race, Non−Hispanic %

### % of Teachers <40 and responses

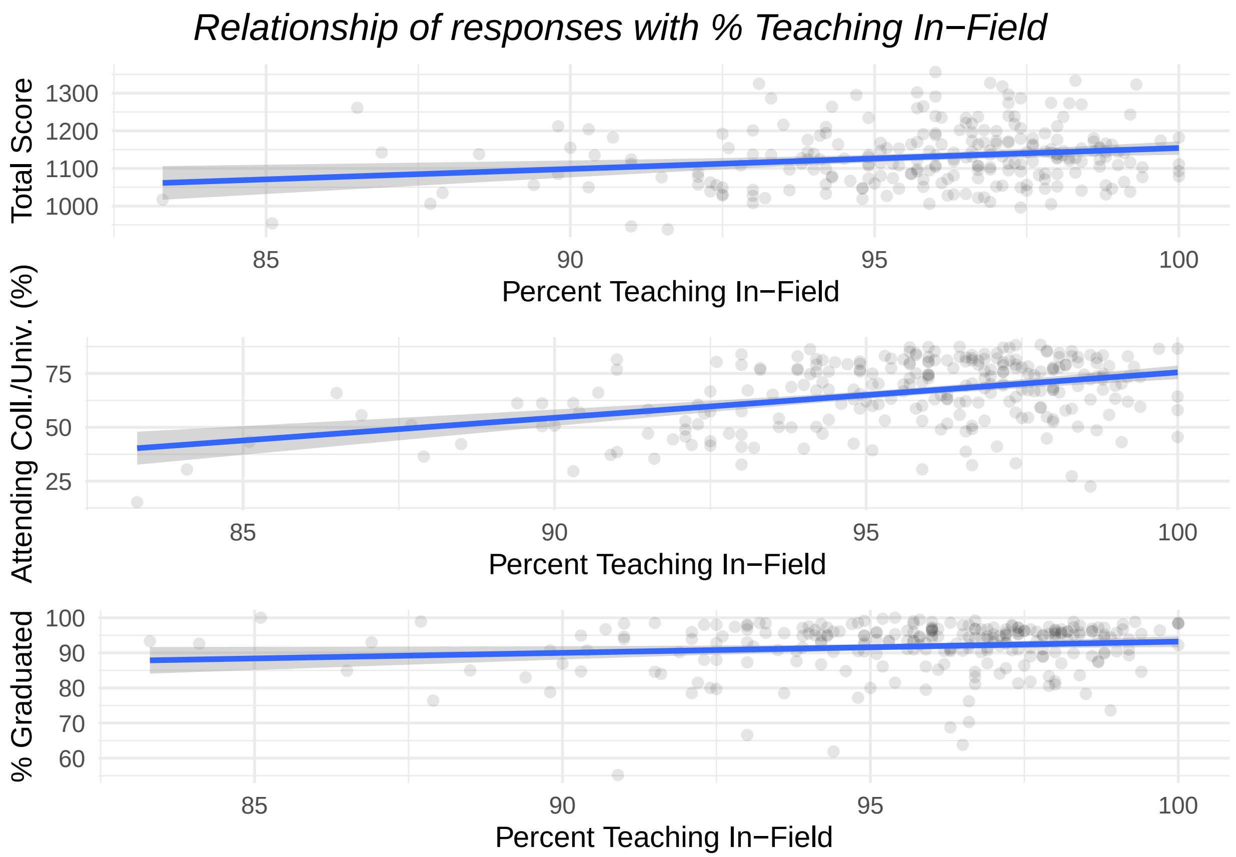


Schools with more % of teacher's age less than 40, had a negative affect on Total SAT score and graduation rate.


Relationship of responses with % in an Art Course
Total Score
1300
1200
1100
1000
0
25
50
75
100
% in an Art Course
Attending Coll./Univ. (%)
75
50
25
0
25
50
75
100
% in an Art Course
% Graduated
100
90
80
70
60
0
25
50
75
100
% in an Art Course


Negative Relationship of % Graduated with different classes%
% Graduated
100
90
80
70
60
0
20
40
60
80
First Language Not English %
% Graduated
100
90
80
70
60
0
10
20
30
English Language Learner %
% Graduated
100
90
80
70
60
10
20
30
40
Students with Disabilities %

### Percent Teaching In-Field vs 3 responses



# Models

## Partitioning the data

```
set.seed(2)

edu_part <- resample_partition(eduData,
                                    p=c(train=0.8,
                                        test=0.1,
                                        valid=0.1))
```

## Calculating correlations

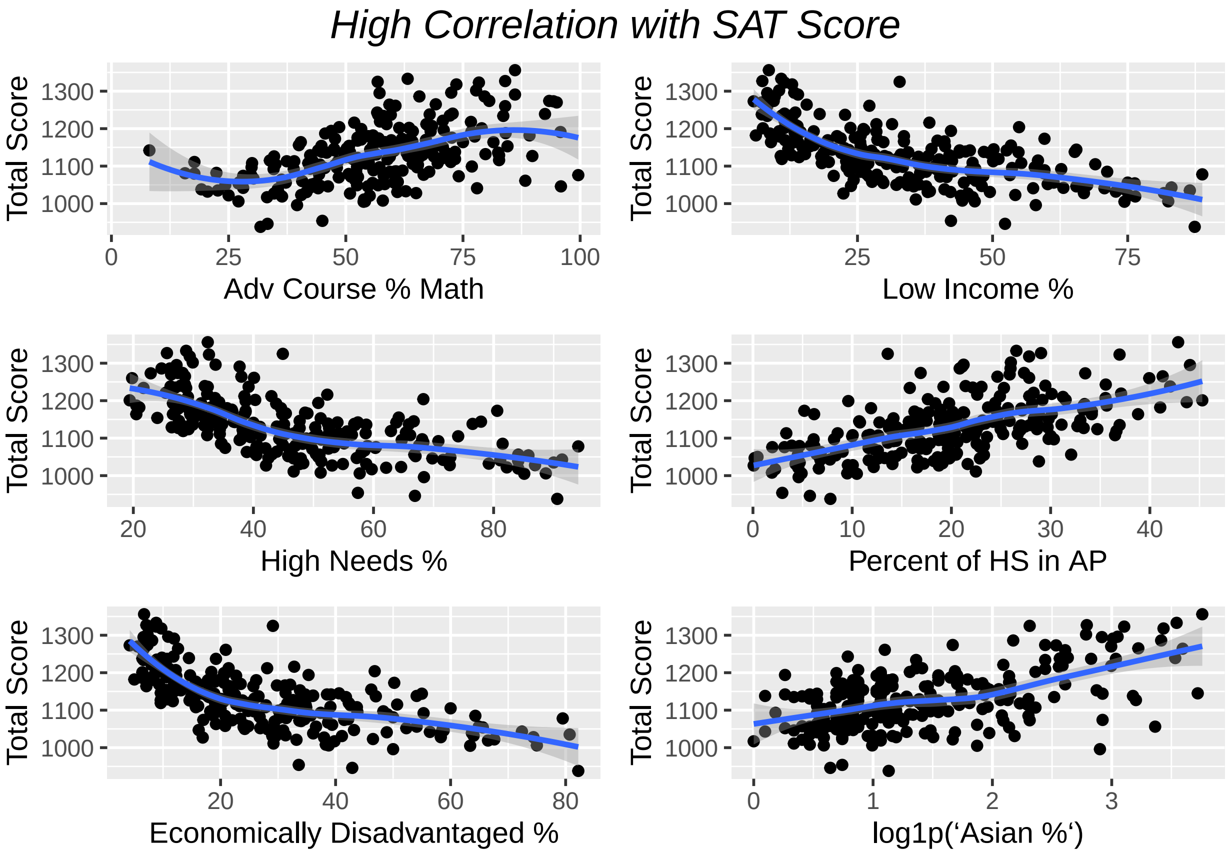
```
temp <- eduData %>%
select(!c(`District Name`, `District Code`, `Reading / Writing`, `Math`, )) %>% na.omit

correlations <- data.frame(abs(cor(temp)))

satCorr <- correlations %>%
  select(Total.Score) %>%
  filter(Total.Score > 0.5)

gradCorr <- correlations %>%
  select(X..Graduated) %>%
  filter(X..Graduated > 0.5)

collCorr <- correlations %>%
  select(Attending.Coll..Univ.....) %>%
  filter(Attending.Coll..Univ..... > 0.5)
```

## SAT

### SAT eda feature selection

```
rownames(satCorr)
```

```
## [1] "Total Score"                    "Attending Coll./Univ. (%)"
## [3] "Adv Course % Math"              "Low Income %"
## [5] "High Needs #"                   "Percent of HS in AP"
## [7] "Adjusted Score"                 "Economically Disadvantaged %"
## [9] "Asian %"
```



```
satEDA <- lm(`Total Score` ~ `Adv Course % Math`  + `Low Income %` + `High Needs #...16`
↪  + `Percent of HS in AP` + `Economically Disadvantaged %` + log1p(`Asian %`), data =
↪  edu_part$train)

summary(satEDA)
```

```
##
```

```
## Call:
## lm(formula = `Total Score` ~ `Adv Course % Math` + `Low Income %` + 
##     `High Needs #...16` + `Percent of HS in AP` + `Economically Disadvantaged %` + 
##     log1p(`Asian %`), data = edu_part$train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -108.301  -31.199   -2.114   27.941  186.097 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    1085.61840   33.16097  32.738  < 2e-16 ***
## `Adv Course % Math`               0.03839    0.27548   0.139    0.889    
## `Low Income %`                   -2.09530    2.37264  -0.883    0.378    
## `High Needs #...16`               0.48387    1.23744   0.391    0.696    
## `Percent of HS in AP`             2.34902    0.51278   4.581 8.62e-06 ***
## `Economically Disadvantaged %`    0.04821    2.26362   0.021    0.983    
## log1p(`Asian %`)                 28.01871    5.52459   5.072 9.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 48.74 on 180 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.6276, Adjusted R-squared:  0.6151 
## F-statistic: 50.55 on 6 and 180 DF,  p-value: < 2.2e-16
```

```
rmse(satEDA, edu_part$test)/1600
```

```
## [1] 0.03287279
```

```
car::vif(satEDA)
```

```
##            `Adv Course % Math`                 `Low Income %` 
##                       1.845319                     161.711616 
##            `High Needs #...16`          `Percent of HS in AP` 
##                      33.419520                       1.957590 
## `Economically Disadvantaged %`               log1p(`Asian %`) 
##                     118.072266                       1.570205 
```

```
# cited this function from prof's 7-Modeling2.Rmd
step1 <- function(response, predictors, candidates, partition)
{
  rhs <- paste0(paste0(predictors, collapse="+"), "+", candidates)
  formulas <- lapply(paste0(response, "~", rhs), as.formula)
  rmses <- sapply(formulas,
                  function(fm) rmse(lm(fm, data=partition$train),
                                    data=partition$valid))
  names(rmses) <- candidates
  attr(rmses, "best") <- rmses[which.min(rmses)]
  rmses
}
```

```
model <- NULL
preds <- "1"
cands <- c('`Adv Course % Math`' , '`Low Income %`', '`High Needs #...16`', '`Percent of
↪  HS in AP`', '`Economically Disadvantaged %`', 'log1p(`Asian %`)')
```

```r
s1 <- step1("`Total Score`", preds, cands, edu_part)

model <- c(model, attr(s1, "best"))
s1
```

```
##            `Adv Course % Math`                 `Low Income %`
##                       52.30379                       62.55912
##            `High Needs #...16`          `Percent of HS in AP`
##                       67.04809                       61.90458
## `Economically Disadvantaged %`               log1p(`Asian %`)
##                       60.99354                       63.05670
## attr(,"best")
## `Adv Course % Math`
##            52.30379
```

```r
preds <- "`Adv Course % Math`"
cands <- c('`Low Income %`', '`High Needs #...16`', '`Percent of HS in AP`',
↪  '`Economically Disadvantaged %`', 'log1p(`Asian %`)')
s2 <- step1("`Total Score`", preds, cands, edu_part)

model <- c(model, attr(s2, "best"))
s2
```

```
##                 `Low Income %`            `High Needs #...16`
##                       57.85240                       60.47527
##          `Percent of HS in AP` `Economically Disadvantaged %`
##                       54.85319                       56.13104
##               log1p(`Asian %`)
##                       56.00554
## attr(,"best")
## `Percent of HS in AP`
##              54.85319
```

```r
preds <- c("`Adv Course % Math`", "`Percent of HS in AP`")
cands <- c('`Low Income %`', '`High Needs #...16`', '`Economically Disadvantaged %`',
↪  'log1p(`Asian %`)')
s3 <- step1("`Total Score`", preds, cands, edu_part)

model <- c(model, attr(s3, "best"))
s3
```

```
##                 `Low Income %`            `High Needs #...16`
##                       60.14335                       61.48609
## `Economically Disadvantaged %`               log1p(`Asian %`)
##                       58.93277                       46.46438
## attr(,"best")
## log1p(`Asian %`)
##         46.46438
```

```r
preds <- c("`Adv Course % Math`", "`Percent of HS in AP`", 'log1p(`Asian %`)')
cands <- c('`Low Income %`', '`High Needs #...16`', '`Economically Disadvantaged %`')
s4 <- step1("`Total Score`", preds, cands, edu_part)

model <- c(model, attr(s4, "best"))
s4
```

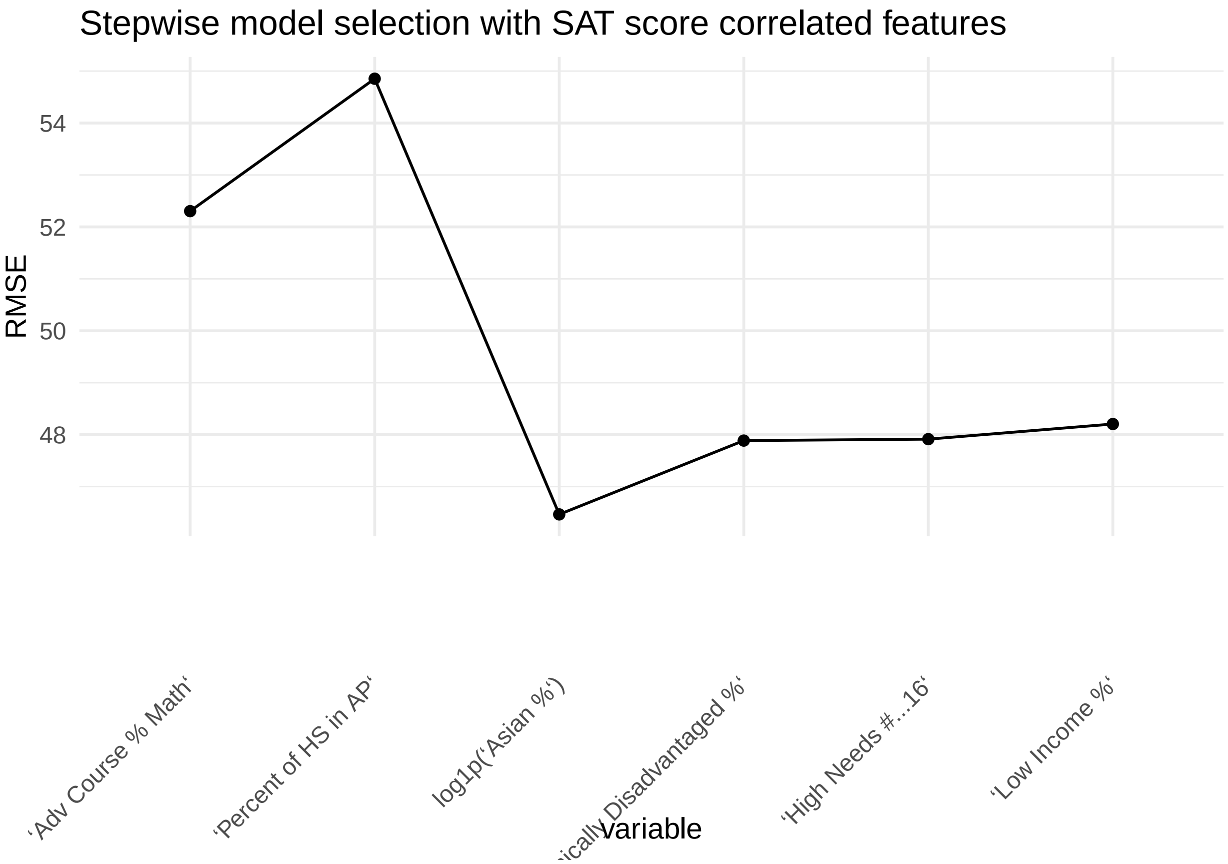
```
##                 `Low Income %`            `High Needs #...16`
##                       48.52851                       49.41032
## `Economically Disadvantaged %`
##                       47.88570
## attr(,"best")
## `Economically Disadvantaged %`
##                        47.8857
```

```r
preds <- c("`Adv Course % Math`", "`Percent of HS in AP`", 'log1p(`Asian %`)',
↪  '`Economically Disadvantaged %`')
cands <- c('`Low Income %`', '`High Needs #...16`')
s5 <- step1("`Total Score`", preds, cands, edu_part)

model <- c(model, attr(s5, "best"))
s5
```

```
##      `Low Income %` `High Needs #...16`
##            48.53682            47.91356
## attr(,"best")
## `High Needs #...16`
##            47.91356
```

```r
preds <- c("`Adv Course % Math`", "`Percent of HS in AP`", 'log1p(`Asian %`)',
↪  '`Economically Disadvantaged %`', '`High Needs #...16`')
cands <- c('`Low Income %`')
s6 <- step1("`Total Score`", preds, cands, edu_part)

model <- c(model, attr(s6, "best"))
s6
```

```
## `Low Income %`
##       48.20481
## attr(,"best")
## `Low Income %`
##       48.20481
```



```
satEDA <- lm(`Total Score` ~ `Adv Course % Math` + `Percent of HS in AP` + log1p(`Asian
↪  %`), data = edu_part$train)

summary(satEDA)
```

```
##
## Call:
## lm(formula = `Total Score` ~ `Adv Course % Math` + `Percent of HS in AP` +
##     log1p(`Asian %`), data = edu_part$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.871  -36.917   -1.478   35.348  191.419
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            971.8867    13.2056  73.596  < 2e-16 ***
## `Adv Course % Math`      0.7130     0.2756   2.587   0.0105 *
## `Percent of HS in AP`    3.4675     0.4902   7.074 3.09e-11 ***
## log1p(`Asian %`)        32.0407     5.3970   5.937 1.43e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.49 on 183 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.5439, Adjusted R-squared:  0.5364
## F-statistic: 72.74 on 3 and 183 DF,  p-value: < 2.2e-16
```

```
rmse(satEDA, edu_part$test)/1600
```

```
## [1] 0.03999947
```

```
car::vif(satEDA)
```

```
##   `Adv Course % Math` `Percent of HS in AP`      log1p(`Asian %`)
##              1.533762              1.484885              1.244015
```

### SAT Feature selection with all available

```
temp <- eduData %>%
  select(!c(`District Name`, `District Code`, `Tests Taken`, `Salary Totals`, `FTE
  ↪  Count`, `HS Enrollment`, Enrollment, `Tests Takers`, `Adjusted Score`, `%
  ↪  Graduated`, `Total # of Teachers (FTE)`, `Reading / Writing`, `Math`, `Attending
  ↪  Coll./Univ. (%)`, `Average # of Absences`))

base.mod <- lm(`Total Score` ~ 1 , data=temp)
all.mod <- lm(`Total Score` ~ . , data= temp)
stepMod <- step(base.mod, scope = list(lower = base.mod, upper = all.mod), direction =
↪  "both", trace = 0, steps = 1000)

shortlistedVars <- names(unlist(stepMod[[1]]))
shortlistedVars <- shortlistedVars[!shortlistedVars %in% "(Intercept)"]

print(shortlistedVars)
```

```
##  [1] "`Low Income %`"                      "`Asian %`"
##  [3] "`% Disciplined`"                     "`Multi-Race, Non-Hispanic %`"
##  [5] "`% Dropout All Grades`"              "`English Language Learner %`"
##  [7] "`Teacher Retention Rate`"            "`% of Teachers <40`"
##  [9] "`Percent of HS in AP`"               "`Percent of Experienced Teachers`"
## [11] "`% Female Teachers`"                 "`Gen Ed %`"
## [13] "`African American %`"                "`% Intake`"
```

```
summary(lm(`Total Score` ~ `Low Income %` + `Asian %` + `% Disciplined` + `Multi-Race,
↪  Non-Hispanic %` + `% Dropout All Grades` + `English Language Learner %` + `Teacher
↪  Retention Rate` + `% of Teachers <40` + `Percent of HS in AP` + `Percent of
↪  Experienced Teachers` + `% Female Teachers` + `Gen Ed %` + `African American %` + `%
↪  Intake`, data = eduData))
```

```
##
## Call:
## lm(formula = `Total Score` ~ `Low Income %` + `Asian %` + `% Disciplined` +
##     `Multi-Race, Non-Hispanic %` + `% Dropout All Grades` + `English Language Learner %` +
##     `Teacher Retention Rate` + `% of Teachers <40` + `Percent of HS in AP` +
##     `Percent of Experienced Teachers` + `% Female Teachers` +
##     `Gen Ed %` + `African American %` + `% Intake`, data = eduData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -117.932  -26.775   -3.725   32.770  126.634
##
## Coefficients:
```

```
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1320.1780    92.9995  14.196  < 2e-16 ***
## `Low Income %`                      -3.0941     0.3345  -9.251  < 2e-16 ***
## `Asian %`                            2.7048     0.4647   5.821 2.03e-08 ***
## `% Disciplined`                    -12.8531     3.7432  -3.434  0.00071 ***
## `Multi-Race, Non-Hispanic %`         8.8337     1.6688   5.294 2.88e-07 ***
## `% Dropout All Grades`               5.8216     3.3952   1.715  0.08780 .
## `English Language Learner %`         2.4912     0.7343   3.392  0.00082 ***
## `Teacher Retention Rate`            -1.8427     0.8278  -2.226  0.02701 *
## `% of Teachers <40`                 -0.7213     0.5049  -1.428  0.15458
## `Percent of HS in AP`                0.7239     0.4649   1.557  0.12090
## `Percent of Experienced Teachers`    1.1610     0.7305   1.589  0.11341
## `% Female Teachers`                 -1.7033     0.5516  -3.088  0.00227 **
## `Gen Ed %`                          63.3327    32.9451   1.922  0.05584 .
## `African American %`                -0.9542     0.4982  -1.915  0.05674 .
## `% Intake`                           4.3827     2.6183   1.674  0.09556 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.15 on 222 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.7254, Adjusted R-squared:  0.7081
## F-statistic:  41.9 on 14 and 222 DF,  p-value: < 2.2e-16
```

```
totalScoreMod <- lm(`Total Score` ~ `Low Income %` + `Asian %` + `% Disciplined` +
↪  `Multi-Race, Non-Hispanic %`  + `Teacher Retention Rate`, data = edu_part$train)

summary(totalScoreMod)
```

```
##
## Call:
## lm(formula = `Total Score` ~ `Low Income %` + `Asian %` + `% Disciplined` +
##     `Multi-Race, Non-Hispanic %` + `Teacher Retention Rate`,
##     data = edu_part$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.482  -27.023   -4.302   26.645  116.071
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1361.1159    72.2982  18.826  < 2e-16 ***
## `Low Income %`                 -2.4706     0.1866 -13.238  < 2e-16 ***
## `Asian %`                       2.6348     0.5437   4.846 2.70e-06 ***
## `% Disciplined`               -19.5543     3.4167  -5.723 4.27e-08 ***
## `Multi-Race, Non-Hispanic %`    8.2980     1.9377   4.282 3.00e-05 ***
## `Teacher Retention Rate`       -2.0077     0.7934  -2.531   0.0122 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.93 on 181 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.6958, Adjusted R-squared:  0.6874
## F-statistic:  82.8 on 5 and 181 DF,  p-value: < 2.2e-16
```

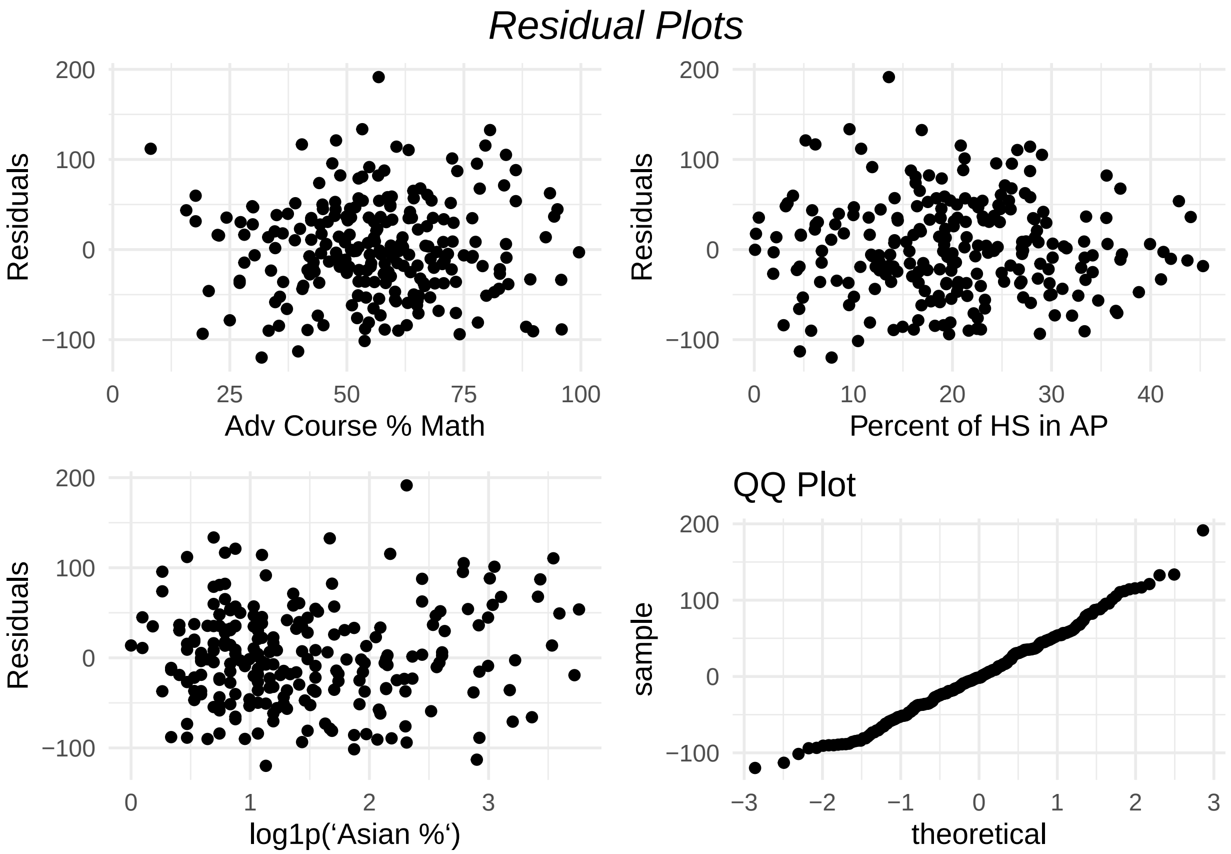
```
rmse(totalScoreMod, edu_part$test)/1600
```

```
## [1] 0.02954361
```

```
car::vif(totalScoreMod)
```

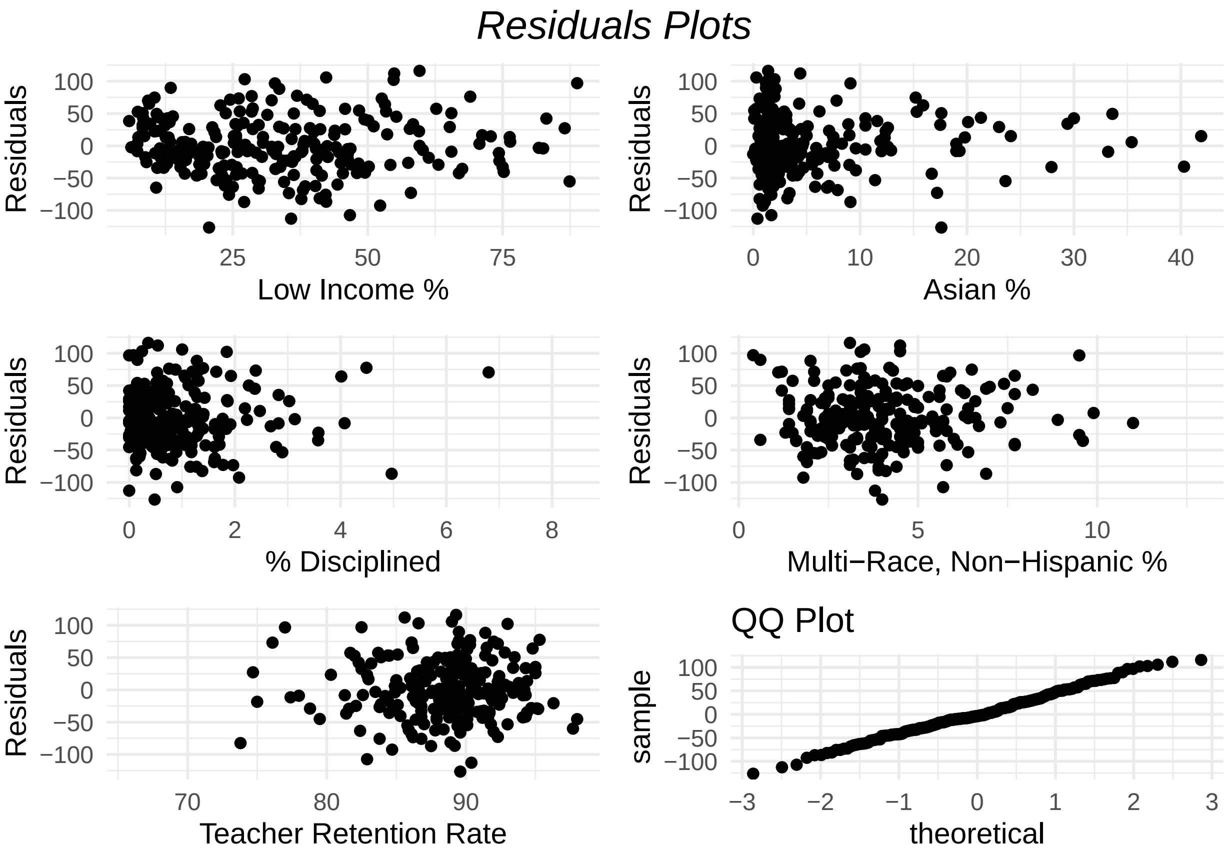
```
##               `Low Income %`                    `Asian %`
##                     1.231735                     1.345503
##              `% Disciplined` `Multi-Race, Non-Hispanic %`
##                     1.132328                     1.118172
##     `Teacher Retention Rate`
##                     1.101235
```

```
g1 <- eduData %>%
  add_residuals(satEDA, "resid") %>%
  ggplot(aes(y=resid,x=`Adv Course % Math`)) +
  geom_point() +
  labs(y="Residuals") +
  theme_minimal()
g2 <- eduData %>%
  add_residuals(satEDA, "resid") %>%
  ggplot(aes(y=resid,x=`Percent of HS in AP` )) +
  geom_point() +
  labs(y="Residuals") +
  theme_minimal()
g3 <- eduData %>%
  add_residuals(satEDA, "resid") %>%
  ggplot(aes(y=resid,x=log1p(`Asian %`)  )) +
  geom_point() +
  labs(y="Residuals") +
  theme_minimal()

g4 <- eduData %>%
  add_residuals(satEDA, "resid") %>%
  ggplot(aes(sample=resid)) +
  geom_qq() +
  theme_minimal()+
  labs(title = "QQ Plot")
gridExtra::grid.arrange(
  g1, g2,g3,g4,
  top = textGrob("Residual Plots",
                                   gp=gpar(fontsize=15,font=3)))
```



### Diagnostics

```
g1 <- eduData %>%
  add_residuals(totalScoreMod, "resid") %>%
  ggplot(aes(y=resid,x=`Low Income %`)) +
  geom_point() +
  labs(y="Residuals") +
  theme_minimal()
g2 <- eduData %>%
  add_residuals(totalScoreMod, "resid") %>%
  ggplot(aes(y=resid,x=`Asian %` )) +
  geom_point() +
  labs(y="Residuals") +
  theme_minimal()
g3 <- eduData %>%
  add_residuals(totalScoreMod, "resid") %>%
  ggplot(aes(y=resid,x=`% Disciplined` )) +
  geom_point() +
  labs(y="Residuals") +
  theme_minimal()
g4 <- eduData %>%
  add_residuals(totalScoreMod, "resid") %>%
  ggplot(aes(y=resid,x=`Multi-Race, Non-Hispanic %` )) +
  geom_point() +
  labs(y="Residuals") +
  theme_minimal()
g5 <- eduData %>%
  add_residuals(totalScoreMod, "resid") %>%
  ggplot(aes(y=resid,x=`Teacher Retention Rate` )) +
  geom_point() +
  labs(y="Residuals") +
  theme_minimal()



g6 <- eduData %>%
  add_residuals(totalScoreMod, "resid") %>%
  ggplot(aes(sample=resid)) +
  geom_qq() +
  theme_minimal()+
  labs(title = "QQ Plot")

gridExtra::grid.arrange(
  g1, g2,g3,g4,g5,g6,
  top = textGrob("Residuals Plots",
                                   gp=gpar(fontsize=15,font=3)))
```

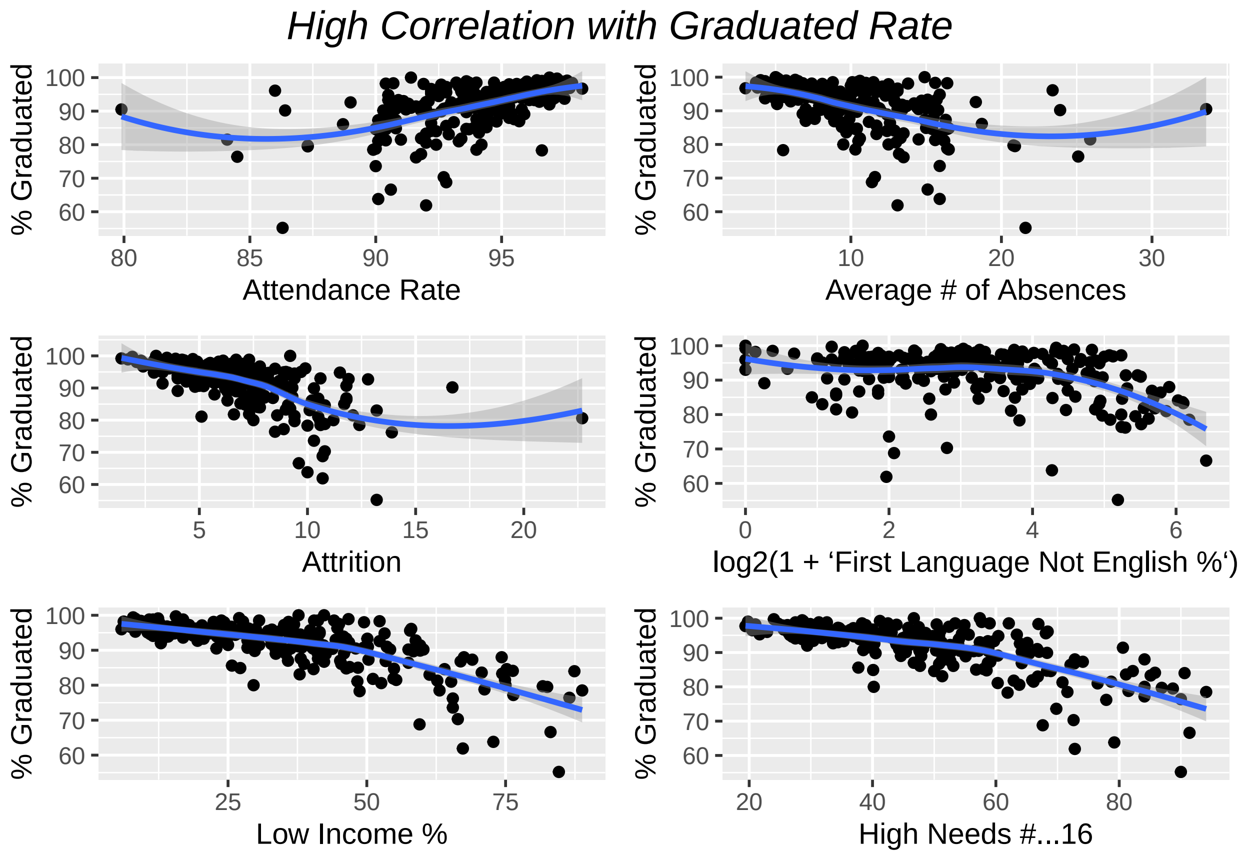


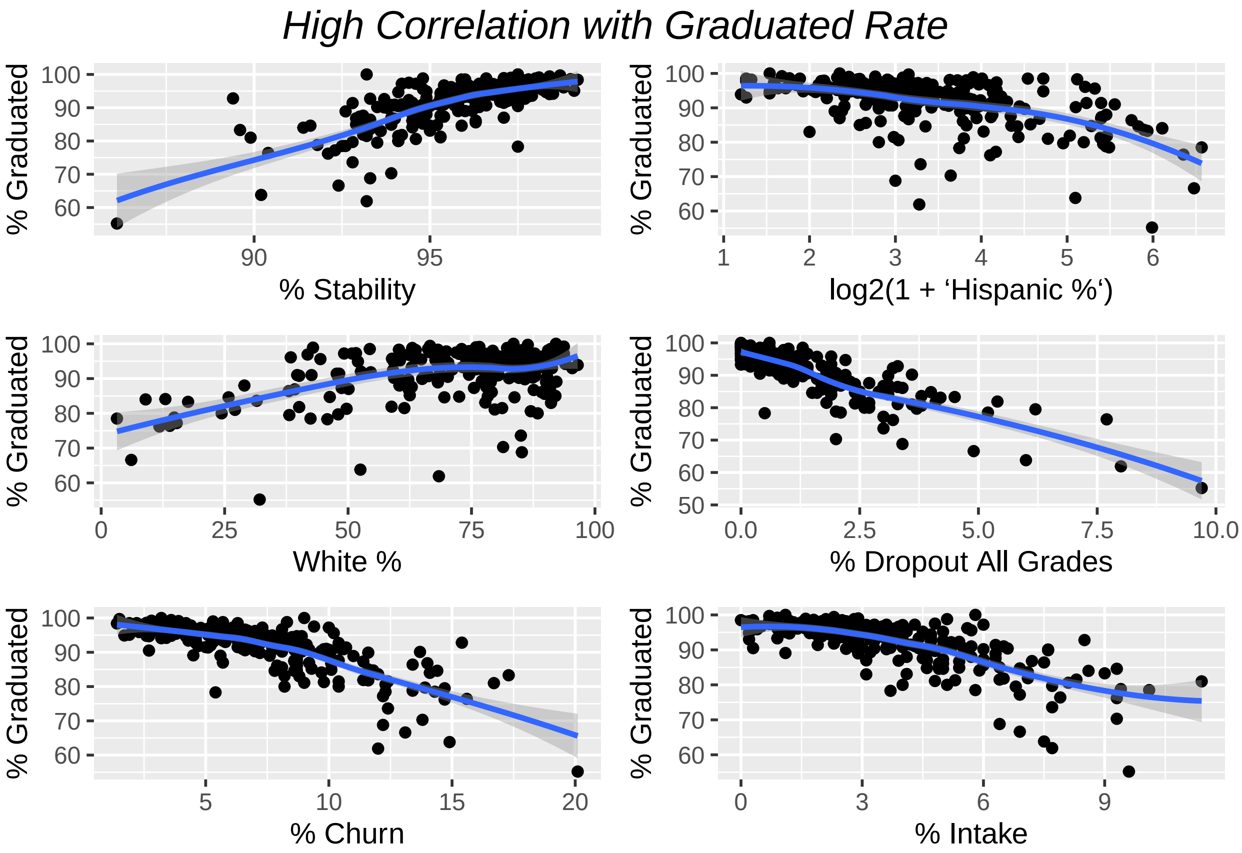
## Graduation rate

### Graduation rate EDA model

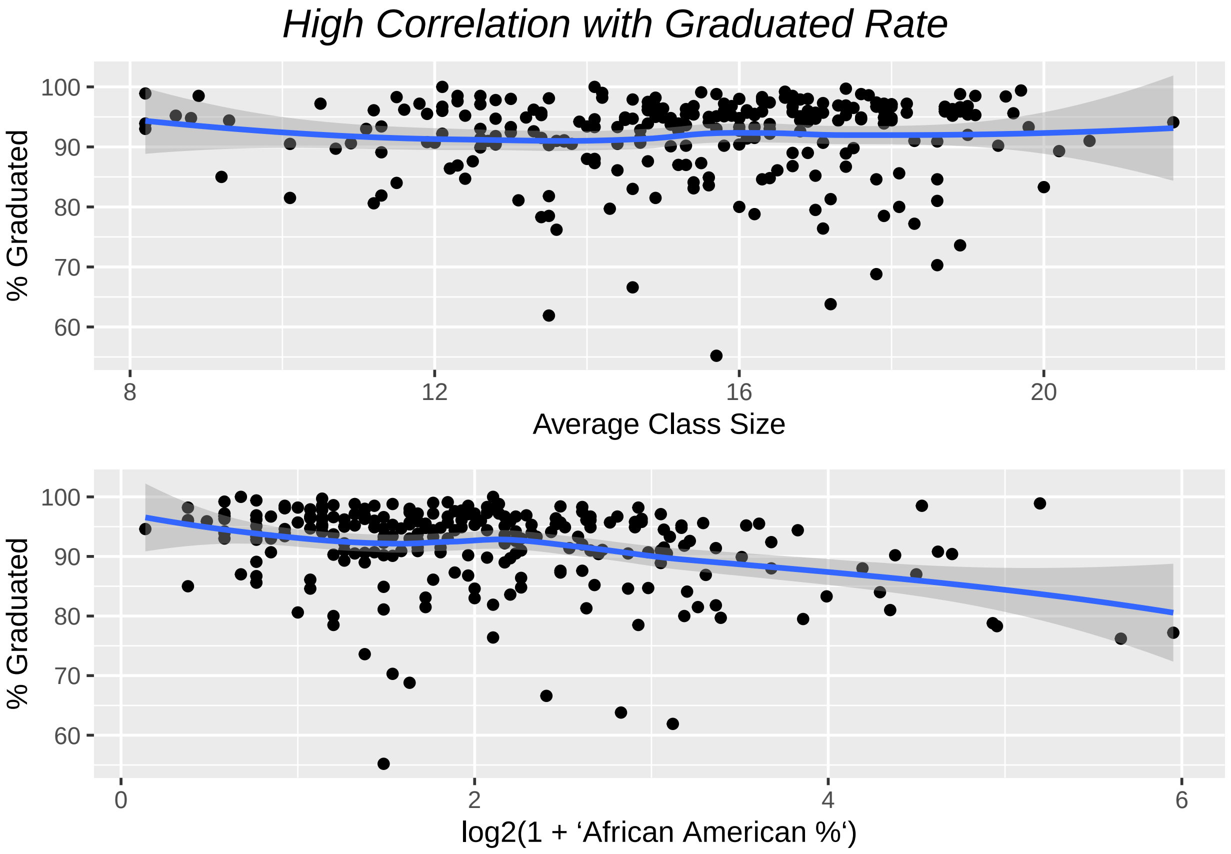
```
rownames(gradCorr)
```

```
##  [1] "Attendance Rate"              "Average # of Absences"
##  [3] "Attrition"                      "% Graduated"
##  [5] "First Language Not English %" "Low Income %"
##  [7] "High Needs #"                 "Adjusted Score"
##  [9] "English Language Learner %"   "Economically Disadvantaged %"
## [11] "Hispanic %"                   "White %"
## [13] "% Dropout All Grades"         "% Churn"
## [15] "% Intake"                     "% Stability"
```





High Correlation with Graduated Rate



High Correlation with Graduated Rate

```
model <- NULL
preds <- "1"
cands <- c('`Attendance Rate`', '`Average # of Absences`', '`Attrition`',
                'log2(1+`First Language Not English %`)', '`Low Income %`',
                '`High Needs #...16`', '`% Stability`', 'log2(1+`Hispanic %`)',
                '`White %`', '`% Dropout All Grades`', '`% Churn`',
                '`% Intake`', 'log2(1+`African American %`)')
s1 <- step1("`% Graduated`", preds, cands, edu_part)

model <- c(model, attr(s1, "best"))
s1
```

```
##                      `Attendance Rate`                `Average # of Absences`
##                               4.918131                               5.056344
##                            `Attrition` log2(1+`First Language Not English %`)
##                               4.834273                               5.503590
##                         `Low Income %`                    `High Needs #...16`
##                               3.926962                               3.778134
##                          `% Stability`                   log2(1+`Hispanic %`)
##                               4.033213                               4.584886
##                              `White %`                 `% Dropout All Grades`
##                               4.956886                               2.895021
##                              `% Churn`                             `% Intake`
##                               3.632539                               3.717872
##           log2(1+`African American %`)
##                               5.598061
## attr(,"best")
## `% Dropout All Grades`
##               2.895021
```

```
preds <- "`% Dropout All Grades`"
cands <- c('`Attendance Rate`', '`Average # of Absences`', '`Attrition`',
                'log2(1+`First Language Not English %`)', '`Low Income %`',
                '`High Needs #...16`', '`% Stability`', 'log2(1+`Hispanic %`)',
                '`White %`', '`% Churn`',
                '`% Intake`', 'log2(1+`African American %`)')
s2 <- step1("`% Graduated`", preds, cands, edu_part)

model <- c(model, attr(s2, "best"))
s2
```

```
##                      `Attendance Rate`                `Average # of Absences`
##                               2.872937                               2.881379
##                            `Attrition` log2(1+`First Language Not English %`)
##                               2.568452                               2.969637
##                         `Low Income %`                    `High Needs #...16`
##                               2.849543                               2.833963
##                          `% Stability`                   log2(1+`Hispanic %`)
##                               2.628827                               2.931460
##                              `White %`                              `% Churn`
##                               2.739402                               2.553979
##                             `% Intake`           log2(1+`African American %`)
##                               2.746075                               2.867987
## attr(,"best")
## `% Churn`
##  2.553979
```

```
preds <- c("`% Dropout All Grades`", "`% Churn`")
cands <- c('`Attendance Rate`', '`Average # of Absences`', '`Attrition`',
                '`First Language Not English %`', '`Low Income %`',
                '`High Needs #...16`', '`% Stability`', 'log2(1+`Hispanic %`)',
                '`White %`', '`% Intake`', 'log2(1+`African American %`)')
s3 <- step1("`% Graduated`", preds, cands, edu_part)

model <- c(model, attr(s3, "best"))
s3
```

```
##              `Attendance Rate`        `Average # of Absences`
##                       2.550188                       2.559937
##                    `Attrition` `First Language Not English %`
##                       2.509251                       2.551464
##                 `Low Income %`            `High Needs #...16`
##                       2.683169                       2.688596
##                  `% Stability`           log2(1+`Hispanic %`)
##                       2.580330                       2.590841
##                      `White %`                     `% Intake`
##                       2.562687                       2.822041
##   log2(1+`African American %`)
##                       2.676882
## attr(,"best")
## `Attrition`
##    2.509251
```

```
preds <- c("`% Dropout All Grades`", "`% Churn`", '`Attrition`')
cands <- c('`Attendance Rate`', '`Average # of Absences`',
                '`First Language Not English %`', '`Low Income %`',
                '`High Needs #...16`', '`% Stability`', 'log2(1+`Hispanic %`)',
                '`White %`', '`% Intake`', 'log2(1+`African American %`)')
s4 <- step1("`% Graduated`", preds, cands, edu_part)

model <- c(model, attr(s4, "best"))
s4
```

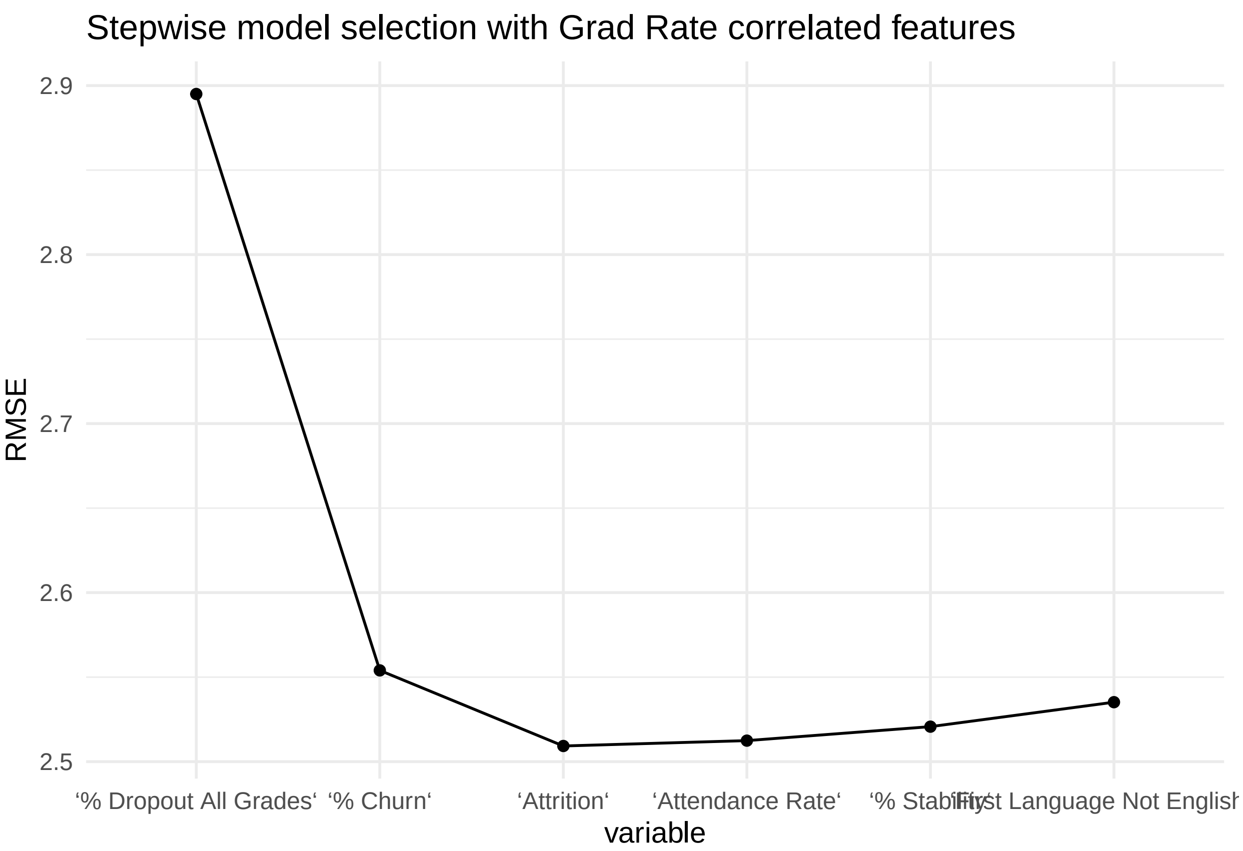
```
##              `Attendance Rate`        `Average # of Absences`
##                       2.512426                       2.523718
## `First Language Not English %`                 `Low Income %`
##                       2.530239                       2.651984
##            `High Needs #...16`                  `% Stability`
##                       2.662974                       2.524281
##           log2(1+`Hispanic %`)                      `White %`
##                       2.557340                       2.533134
##                     `% Intake`   log2(1+`African American %`)
##                       2.740519                       2.666985
## attr(,"best")
## `Attendance Rate`
##          2.512426
```

```
preds <- c("`% Dropout All Grades`", "`% Churn`", '`Attrition`', '`Attendance Rate`')
cands <- c('`Average # of Absences`',
                '`First Language Not English %`', '`Low Income %`',
                '`High Needs #...16`', '`% Stability`', 'log2(1+`Hispanic %`)',
                '`White %`', '`% Intake`', 'log2(1+`African American %`)')
s5 <- step1("`% Graduated`", preds, cands, edu_part)

model <- c(model, attr(s5, "best"))
s5
```

```
##        `Average # of Absences` `First Language Not English %`
##                       2.583497                       2.531443
##                 `Low Income %`            `High Needs #...16`
##                       2.782944                       2.785733
##                  `% Stability`           log2(1+`Hispanic %`)
##                       2.520697                       2.574982
##                      `White %`                     `% Intake`
##                       2.535775                       2.729656
##   log2(1+`African American %`)
##                       2.677647
## attr(,"best")
## `% Stability`
##      2.520697
```

```
preds <- c("`% Dropout All Grades`", "`% Churn`", '`Attrition`', '`Attendance Rate`', '`%
 ↪  Stability`' )
cands <- c('`Average # of Absences`',
                '`First Language Not English %`', '`Low Income %`',
                '`High Needs #...16`', 'log2(1+`Hispanic %`)',
                '`White %`', '`% Intake`', 'log2(1+`African American %`)')
s6 <- step1("`% Graduated`", preds, cands, edu_part)

model <- c(model, attr(s6, "best"))
s6
```

```
##        `Average # of Absences` `First Language Not English %`
##                       2.584488                       2.535171
##                 `Low Income %`            `High Needs #...16`
##                       2.781980                       2.790295
##           log2(1+`Hispanic %`)                      `White %`
##                       2.576433                       2.535816
##                     `% Intake`   log2(1+`African American %`)
##                       2.747295                       2.682208
## attr(,"best")
## `First Language Not English %`
##                       2.535171
```



```
grad_ead_model <- lm(`% Graduated` ~ `% Dropout All Grades` + `% Churn`,
↪  data=edu_part$train)
summary(grad_ead_model)
```

```
##
## Call:
## lm(formula = `% Graduated` ~ `% Dropout All Grades` + `% Churn`,
##     data = edu_part$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7146  -1.4645   0.1963   1.6727  12.4051
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               99.8627     0.6250 159.772  < 2e-16 ***
## `% Dropout All Grades`  -3.0105     0.2616 -11.510  < 2e-16 ***
## `% Churn`                -0.6190     0.1095  -5.652 5.48e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.552 on 198 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7484
## F-statistic: 298.4 on 2 and 198 DF,  p-value: < 2.2e-16
```

```
rmse(grad_ead_model, edu_part$test)
```

```
## [1] 2.918455
```

### Grad Rate Model with all available

```
temp <- eduData %>%
  select(!c(`District Name`, `District Code`, `Tests Taken`, `Salary Totals`, `FTE
  ↪  Count`, `HS Enrollment`, Enrollment, `Tests Takers`, `Adjusted Score`, `Total
  ↪  Score`, `Total # of Teachers (FTE)`, `Reading / Writing`, `Math`, `Attending
  ↪  Coll./Univ. (%)`, `Average # of Absences`))

base.mod <- lm(`% Graduated` ~ 1 , data=temp)
all.mod <- lm(`% Graduated` ~ . , data= temp)
stepMod <- step(base.mod, scope = list(lower = base.mod, upper = all.mod), direction =
↪  "forward", trace = 0, steps = 1000)

shortlistedVars <- names(unlist(stepMod[[1]]))
shortlistedVars <- shortlistedVars[!shortlistedVars %in% "(Intercept)"]

print(shortlistedVars)
```

```
## [1] "`% Dropout All Grades`"         "`% Intake`"
## [3] "`Economically Disadvantaged %`" "Attrition"
## [5] "`Average Class Size`"           "`African American %`"
## [7] "`% Churn`"
```

```
summary(lm(`% Graduated` ~ `% Dropout All Grades` + `% Intake` + `Economically
↪  Disadvantaged %` + Attrition + `Average Class Size` + `African American %` + `%
↪  Churn`, data = eduData))
```

```
## 
## Call:
## lm(formula = `% Graduated` ~ `% Dropout All Grades` + `% Intake` + 
##     `Economically Disadvantaged %` + Attrition + `Average Class Size` + 
##     `African American %` + `% Churn`, data = eduData)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -13.3304  -1.4834   0.1201   1.6019  10.9938 
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    103.48740    1.43502  72.116  < 2e-16 ***
## `% Dropout All Grades`          -2.87393    0.22739 -12.639  < 2e-16 ***
## `% Intake`                      -0.85877    0.29737  -2.888  0.00423 ** 
## `Economically Disadvantaged %`  -0.09060    0.01958  -4.626 6.04e-06 ***
## Attrition                       -0.29569    0.10456  -2.828  0.00507 ** 
## `Average Class Size`            -0.17405    0.07892  -2.205  0.02837 *  
## `African American %`            -0.06038    0.03209  -1.882  0.06108 .  
## `% Churn`                        0.39585    0.22869   1.731  0.08472 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.147 on 244 degrees of freedom
## Multiple R-squared:  0.8026, Adjusted R-squared:  0.797 
## F-statistic: 141.8 on 7 and 244 DF,  p-value: < 2.2e-16
```

```
gradRateMod <- lm(`% Graduated` ~ `% Dropout All Grades` + `% Intake` + `Economically
↪  Disadvantaged %`, data = edu_part$train)

rmse(gradRateMod, edu_part$test)/100
```

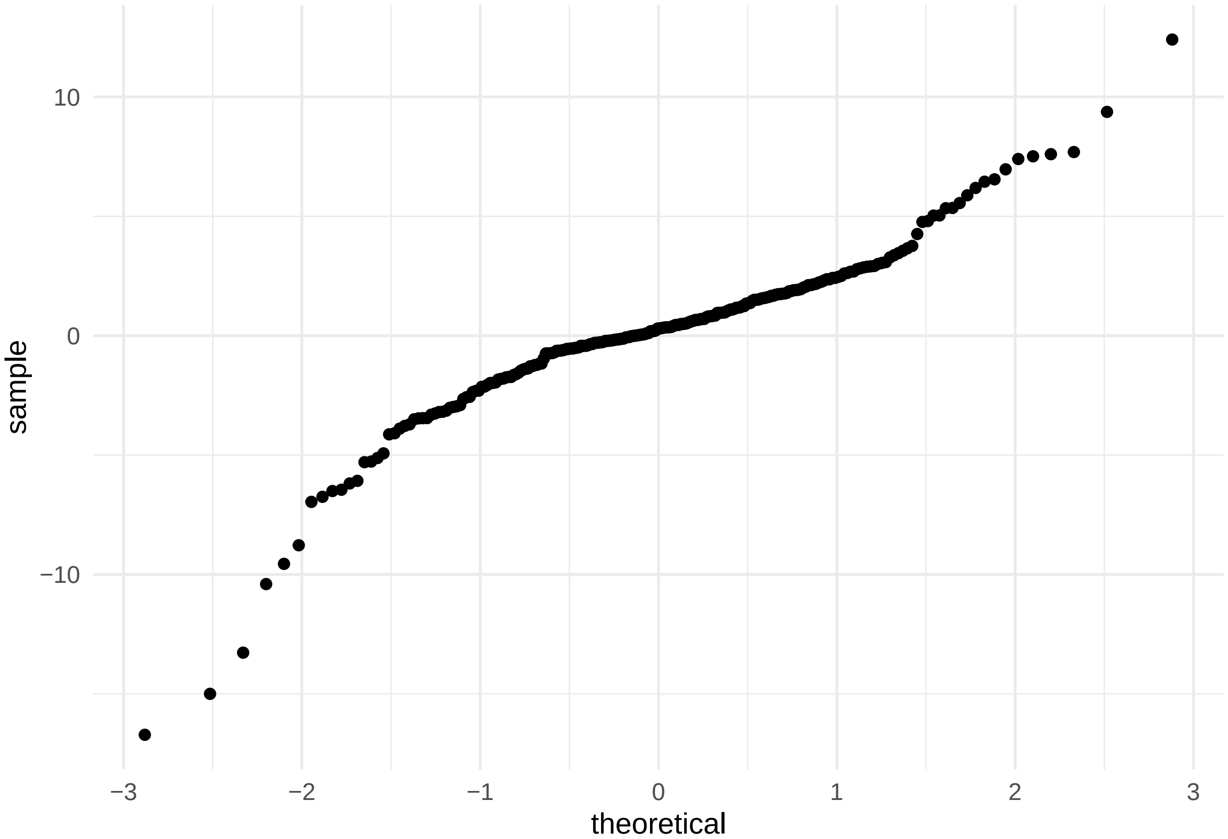
```
## [1] 0.027882
```

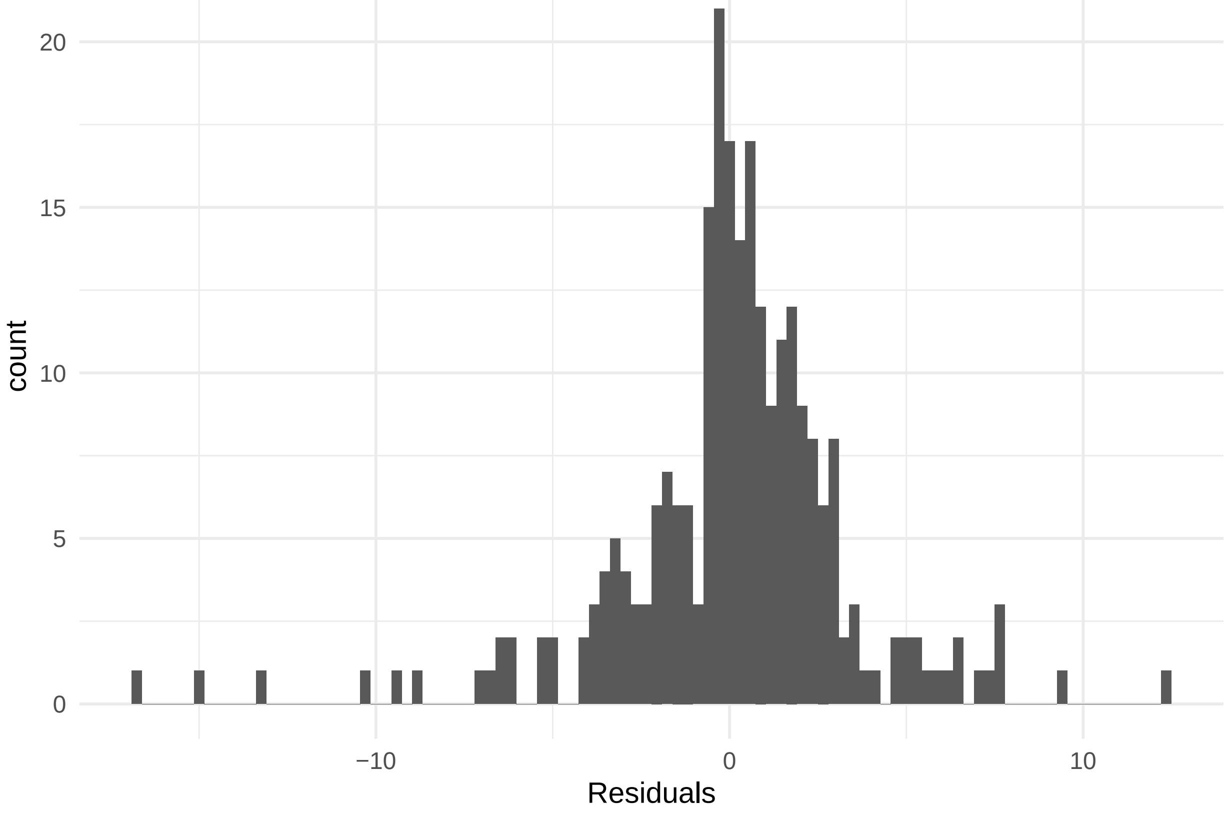
```
summary(gradRateMod)
```

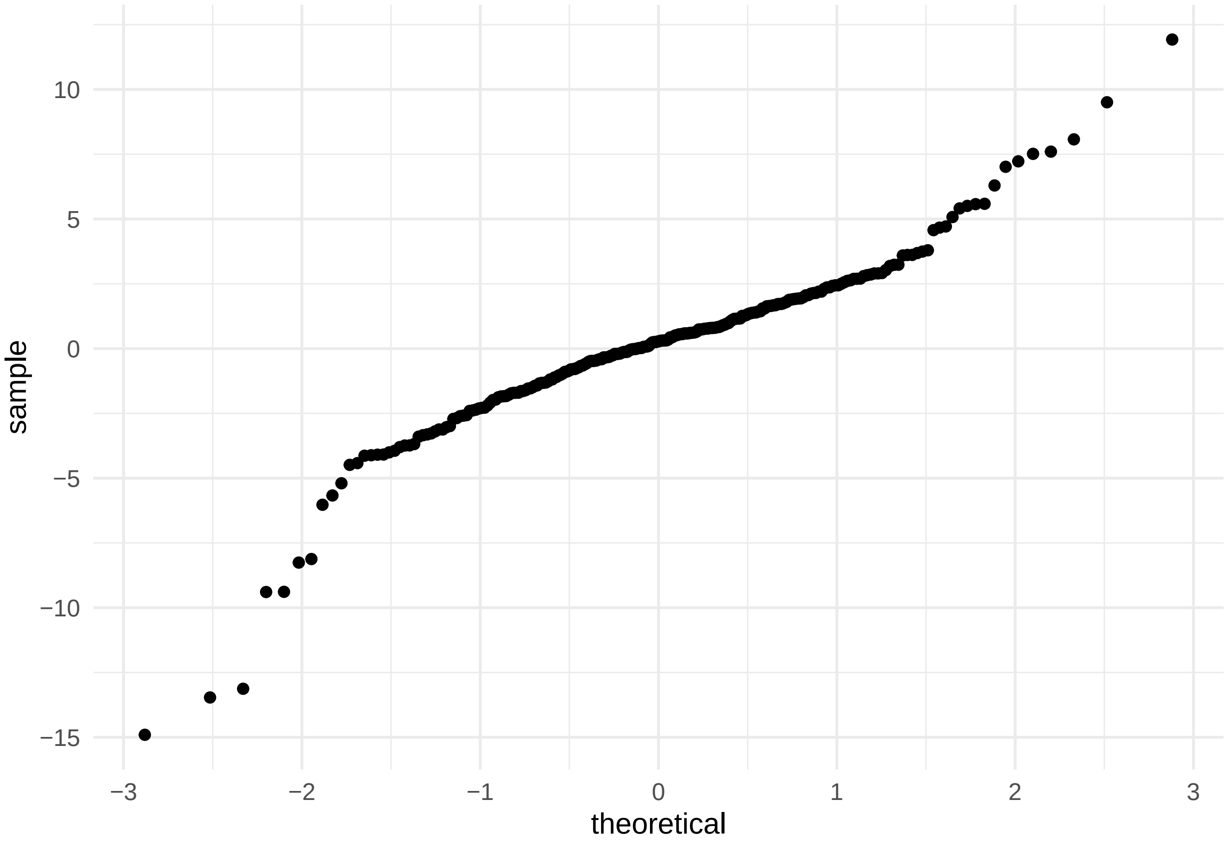
```
##
## Call:
## lm(formula = `% Graduated` ~ `% Dropout All Grades` + `% Intake` +
##     `Economically Disadvantaged %`, data = edu_part$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9034  -1.5412   0.2517   1.6701  11.9253
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   100.23790    0.52567 190.686  < 2e-16 ***
## `% Dropout All Grades`         -2.75253    0.23701 -11.613  < 2e-16 ***
## `% Intake`                     -0.72376    0.15657  -4.622 6.85e-06 ***
## `Economically Disadvantaged %`  -0.07969    0.02002  -3.980 9.70e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.336 on 197 degrees of freedom
## Multiple R-squared:  0.7813, Adjusted R-squared:  0.778
## F-statistic: 234.6 on 3 and 197 DF,  p-value: < 2.2e-16
```

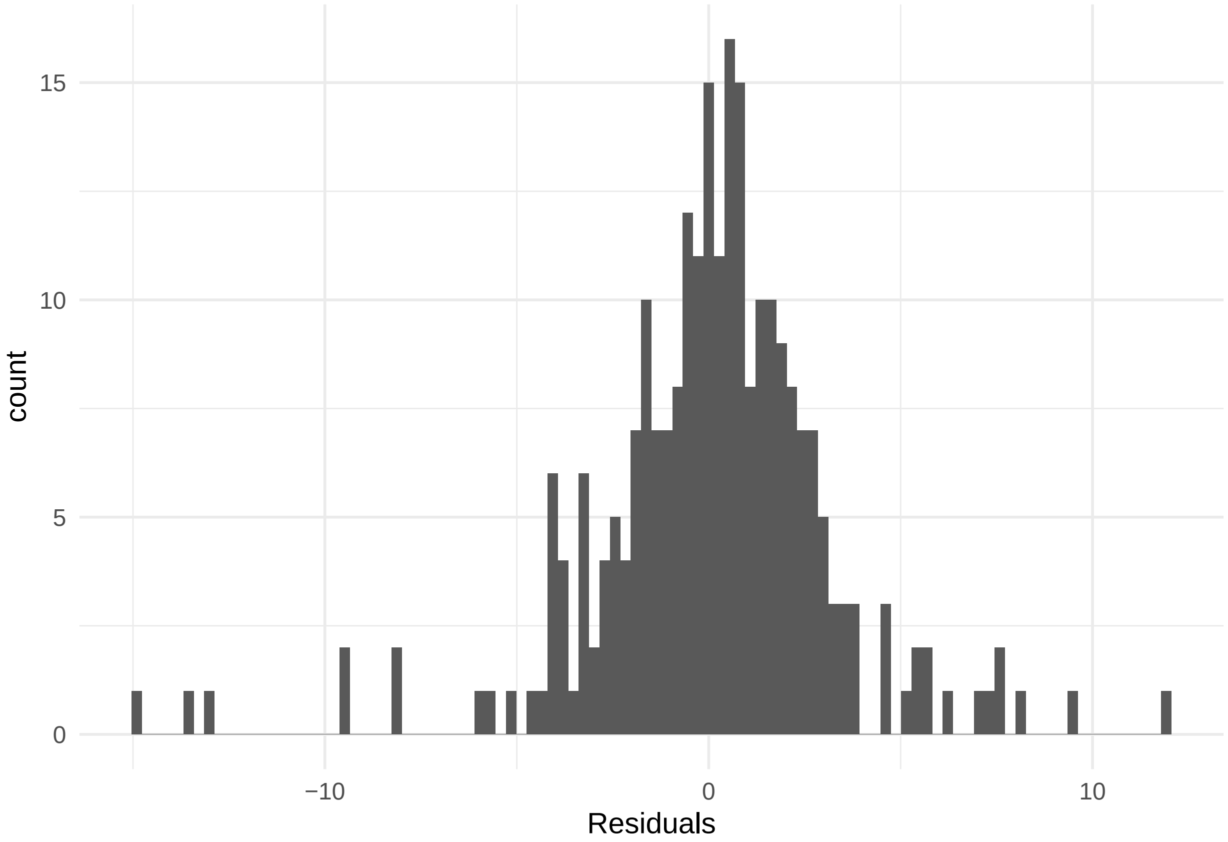
```
car::vif(gradRateMod)
```

```
##         `% Dropout All Grades`                     `% Intake`
##                       2.121131                       2.086534
## `Economically Disadvantaged %`
##                       2.253888
```


10
0
−10
sample
−3
−2
−1
0
1
2
3
theoretical


20
15
10
5
0
count
−10
0
10
Residuals


sample
10
5
0
−5
−10
−15
−3
−2
−1
0
1
2
3
theoretical

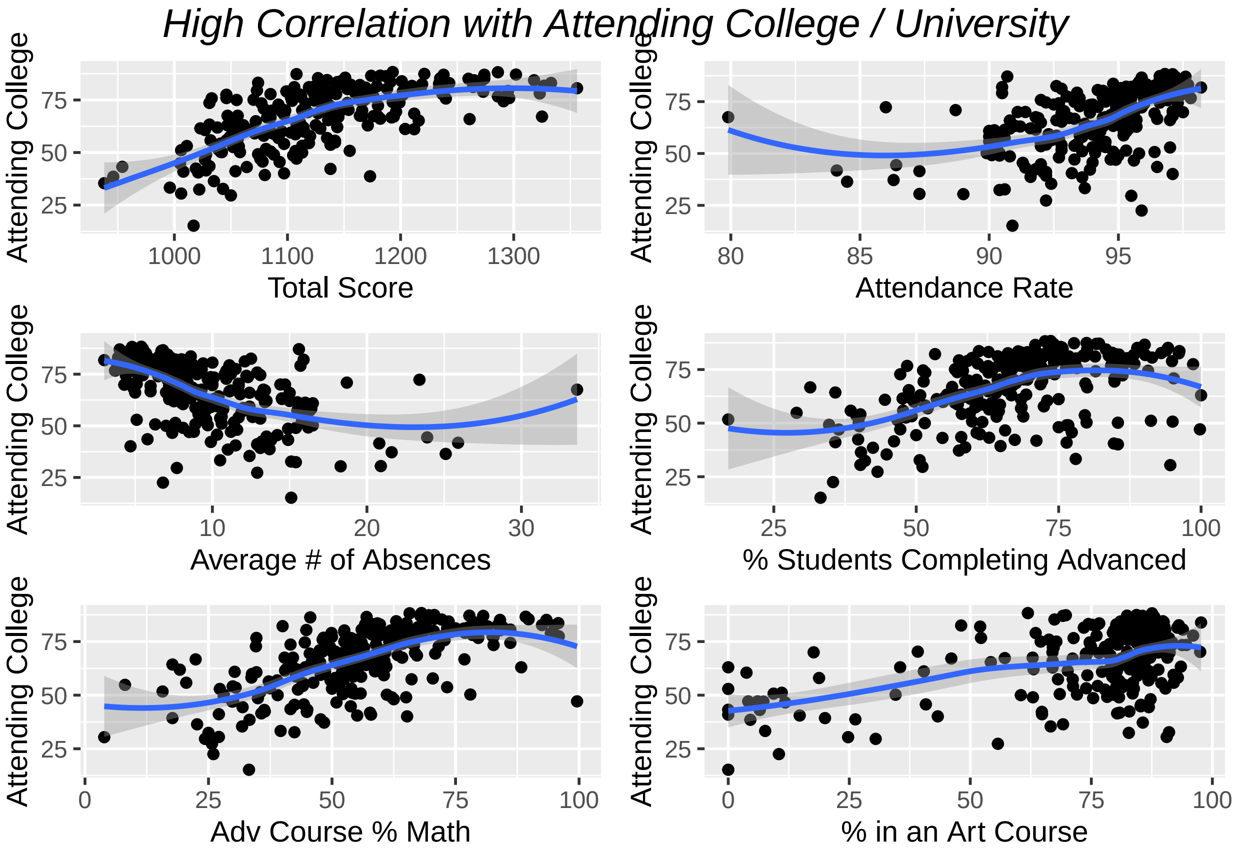

count
15
10
5
0
−10
0
10
Residuals

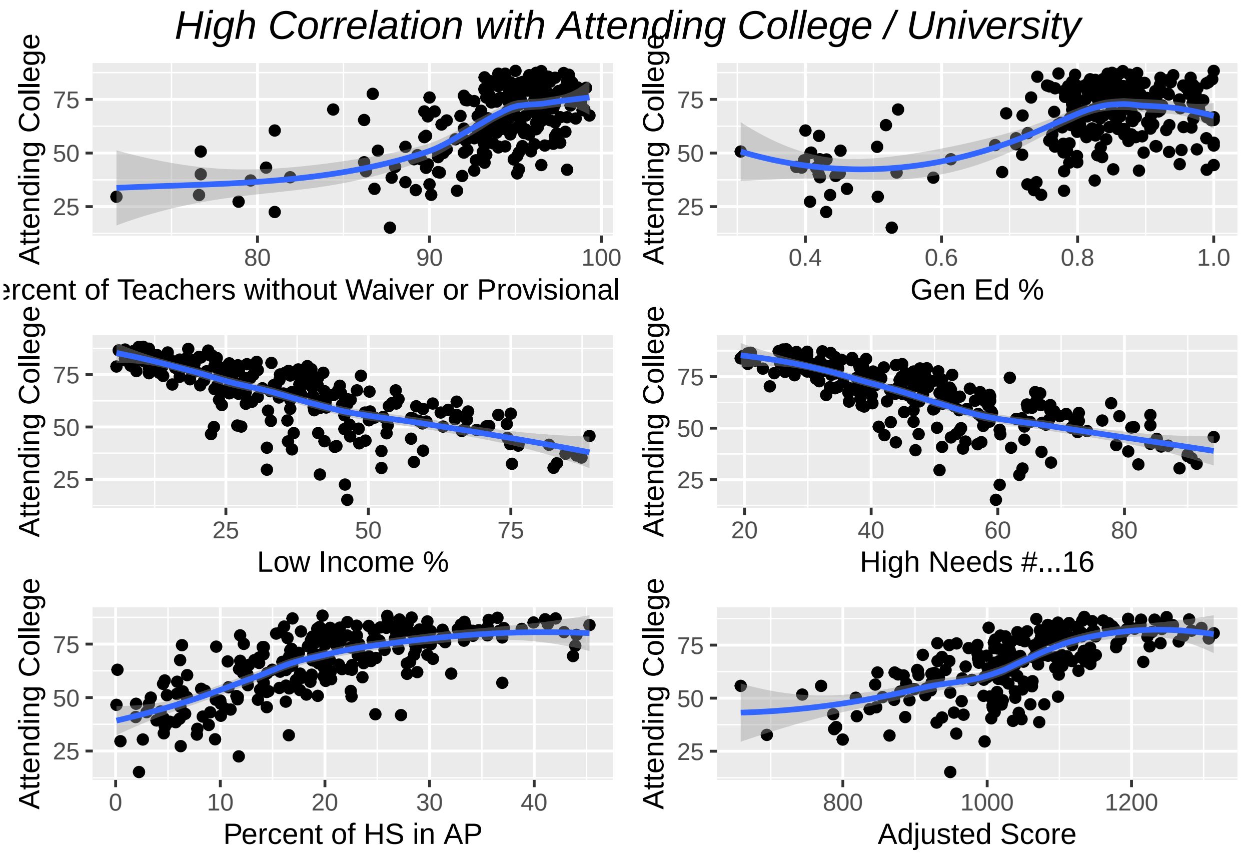
## Percent going to college

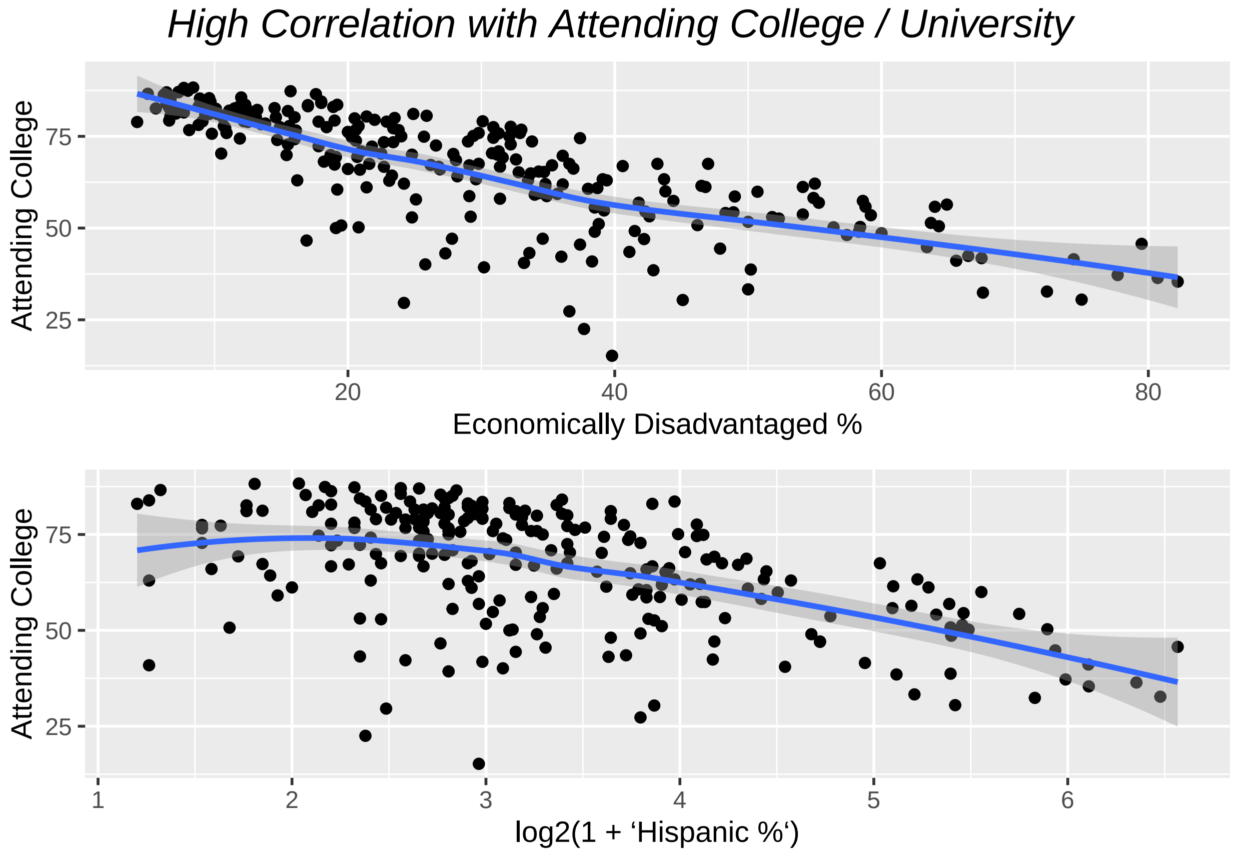
### Percent going to college EDA model

```
rownames(collCorr)
```

```
##  [1] "Total Score"
##  [2] "Attending Coll./Univ. (%)"
##  [3] "Attendance Rate"
##  [4] "Average # of Absences"
##  [5] "% Students Completing Advanced"
##  [6] "Adv Course % Math"
##  [7] "% in an Art Course"
##  [8] "Percent of Teachers without Waiver or Provisional License"
##  [9] "Gen Ed %"
## [10] "Low Income %"
## [11] "High Needs #"
## [12] "Percent of HS in AP"
## [13] "Adjusted Score"
## [14] "Economically Disadvantaged %"
## [15] "Hispanic %"
```




High Correlation with Attending College / University
Attending College
75
50
25
80
90
100
ercent of Teachers without Waiver or Provisional
Attending College
75
50
25
0.4
0.6
0.8
1.0
Gen Ed %
Attending College
75
50
25
25
50
75
Low Income %
Attending College
75
50
25
20
40
60
80
High Needs #...16
Attending College
75
50
25
0
10
20
30
40
Percent of HS in AP
Attending College
75
50
25
800
1000
1200
Adjusted Score


High Correlation with Attending College / University
Attending College
75
50
25
20
40
60
80
Economically Disadvantaged %
Attending College
75
50
25
1
2
3
4
5
6
log2(1 + ‘Hispanic %‘)

```
model <- NULL
preds <- "1"
cands <- c('`Total Score`'  , '`Attendance Rate`' , '`Average # of Absences`' ,
                '`% Students Completing Advanced`' , '`Adv Course % Math`' ,
                '`% in an Art Course`' , '`% in an Art Course`' , '`Percent of Teachers
                ↪   without Waiver or Provisional License`' ,
                '`Gen Ed %`' , '`Low Income %`' , '`High Needs #...16`' ,
                '`Percent of HS in AP`' , '`Adjusted Score`' , '`Economically
                ↪   Disadvantaged %`', 'log2(1 + `Hispanic %`)')
s1 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s1, "best"))
s1
```

```
##                                               `Total Score`
##                                                   11.530270
##                                           `Attendance Rate`
##                                                   11.191387
##                                     `Average # of Absences`
##                                                   11.251328
##                            `% Students Completing Advanced`
##                                                   11.301456
##                                         `Adv Course % Math`
##                                                   11.584559
##                                        `% in an Art Course`
##                                                   12.100420
##                                        `% in an Art Course`
##                                                   12.100420
## `Percent of Teachers without Waiver or Provisional License`
##                                                   10.701734
##                                                  `Gen Ed %`
##                                                   12.336662
##                                              `Low Income %`
##                                                    7.199538
##                                         `High Needs #...16`
##                                                    7.895609
##                                       `Percent of HS in AP`
##                                                    9.789755
##                                            `Adjusted Score`
##                                                    9.216976
##                              `Economically Disadvantaged %`
##                                                    7.175310
##                                      log2(1 + `Hispanic %`)
##                                                   10.211288
## attr(,"best")
## `Economically Disadvantaged %`
##                        7.17531
```

```
preds <- "`Economically Disadvantaged %`"
cands <- c('`Total Score`'  , '`Attendance Rate`' , '`Average # of Absences`' ,
                '`% Students Completing Advanced`' , '`Adv Course % Math`' ,
                '`% in an Art Course`' , '`% in an Art Course`' , '`Percent of Teachers
                ↪   without Waiver or Provisional License`' ,
                '`Gen Ed %`' , '`Low Income %`' , '`High Needs #...16`' ,
```

```r
                '`Percent of HS in AP`' , '`Adjusted Score`' , 'log2(1 + `Hispanic %`)')
s2 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s2, "best"))
s2
```

```
##                                               `Total Score` 
##                                                    6.986060 
##                                           `Attendance Rate` 
##                                                    7.151598 
##                                     `Average # of Absences` 
##                                                    7.132073 
##                            `% Students Completing Advanced` 
##                                                    6.924454 
##                                         `Adv Course % Math` 
##                                                    7.113582 
##                                        `% in an Art Course` 
##                                                    6.003286 
##                                        `% in an Art Course` 
##                                                    6.003286 
## `Percent of Teachers without Waiver or Provisional License` 
##                                                    5.651191 
##                                                  `Gen Ed %` 
##                                                    6.583532 
##                                              `Low Income %` 
##                                                    7.379616 
##                                         `High Needs #...16` 
##                                                    8.412189 
##                                       `Percent of HS in AP` 
##                                                    6.878095 
##                                            `Adjusted Score` 
##                                                    7.045213 
##                                      log2(1 + `Hispanic %`) 
##                                                    7.246287 
## attr(,"best")
## `Percent of Teachers without Waiver or Provisional License` 
##                                                    5.651191
```

```r
preds <- c("`Economically Disadvantaged %`", "`Percent of Teachers without Waiver or
↪   Provisional License`")
cands <- c('`Total Score`'  , '`Attendance Rate`' , '`Average # of Absences`' ,
                '`% Students Completing Advanced`' , '`Adv Course % Math`' ,
                '`% in an Art Course`' , '`% in an Art Course`' ,
                '`Gen Ed %`' , '`Low Income %`' , '`High Needs #...16`' ,
                '`Percent of HS in AP`' , '`Adjusted Score`' , 'log2(1 + `Hispanic %`)')
s3 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s3, "best"))
s3
```

```
##                    `Total Score`                `Attendance Rate` 
##                         5.572271                         5.615093 
##          `Average # of Absences` `% Students Completing Advanced` 
##                         5.603352                         5.559186 
##             `Adv Course % Math`            `% in an Art Course` 
##                        5.751164                        5.494709 
##            `% in an Art Course`                      `Gen Ed %` 
##                        5.494709                        5.705164 
##                  `Low Income %`             `High Needs #...16` 
##                        5.933657                        6.276938 
##           `Percent of HS in AP`                `Adjusted Score` 
##                        5.690070                        5.623835 
##          log2(1 + `Hispanic %`) 
##                        5.844701 
## attr(,"best")
## `% in an Art Course` 
##             5.494709 
```

```
preds <- c("`Economically Disadvantaged %`", "`Percent of Teachers without Waiver or
↪  Provisional License`", "`% in an Art Course`")
cands <- c('`Total Score`'  , '`Attendance Rate`' , '`Average # of Absences`' ,
                '`% Students Completing Advanced`' , '`Adv Course % Math`' ,
                '`Gen Ed %`' , '`Low Income %`' , '`High Needs #...16`' ,
                '`Percent of HS in AP`' , '`Adjusted Score`' , 'log2(1 + `Hispanic %`)')
s4 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s4, "best"))
s4
```

```
##                   `Total Score`               `Attendance Rate` 
##                        5.447488                        5.462822 
##         `Average # of Absences` `% Students Completing Advanced` 
##                        5.457332                        5.299078 
##             `Adv Course % Math`                      `Gen Ed %` 
##                        5.389467                        5.335064 
##                  `Low Income %`             `High Needs #...16` 
##                        5.591492                        5.986228 
##           `Percent of HS in AP`                `Adjusted Score` 
##                        5.303476                        5.457505 
##          log2(1 + `Hispanic %`) 
##                        5.620249 
## attr(,"best")
## `% Students Completing Advanced` 
##                         5.299078 
```

```
preds <- c("`Economically Disadvantaged %`", "`Percent of Teachers without Waiver or
↪  Provisional License`", "`% in an Art Course`", "`% Students Completing Advanced`")
cands <- c('`Total Score`'  , '`Attendance Rate`' , '`Average # of Absences`' ,
                '`Adv Course % Math`' ,
                '`Gen Ed %`' , '`Low Income %`' , '`High Needs #...16`' ,
                '`Percent of HS in AP`' , '`Adjusted Score`' , 'log2(1 + `Hispanic %`)')
s5 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s5, "best"))
s5
```

```
##           `Total Score`       `Attendance Rate` `Average # of Absences` 
##                5.333482                5.304777                5.303432 
##     `Adv Course % Math`              `Gen Ed %`          `Low Income %` 
##                5.376611                5.100123                5.336616
##     `High Needs #...16`   `Percent of HS in AP`        `Adjusted Score`
##                5.630512                5.187120                5.327966
##  log2(1 + `Hispanic %`)
##                5.433785
## attr(,"best")
## `Gen Ed %`
##   5.100123
```

```
preds <- c("`Economically Disadvantaged %`", "`Percent of Teachers without Waiver or
↪  Provisional License`", "`% in an Art Course`", "`% Students Completing Advanced`",
↪  "`Gen Ed %`")
cands <- c('`Total Score`'  , '`Attendance Rate`' , '`Average # of Absences`' ,
                '`Adv Course % Math`' ,
                '`Low Income %`' , '`High Needs #...16`' ,
                '`Percent of HS in AP`' , '`Adjusted Score`' , 'log2(1 + `Hispanic %`)')
s6 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s6, "best"))
s6
```

```
##           `Total Score`       `Attendance Rate` `Average # of Absences`
##                5.114968                5.116683                5.116049
##     `Adv Course % Math`          `Low Income %`     `High Needs #...16`
##                5.083527                5.180812                5.442762
##   `Percent of HS in AP`        `Adjusted Score`  log2(1 + `Hispanic %`)
##                5.121749                5.126516                5.137812
## attr(,"best")
## `Adv Course % Math`
##            5.083527
```

```
preds <- c("`Economically Disadvantaged %`", "`Percent of Teachers without Waiver or
↪  Provisional License`", "`% in an Art Course`", "`% Students Completing Advanced`",
↪  "`Gen Ed %`", "`Adv Course % Math`")
cands <- c('`Total Score`'  , '`Attendance Rate`' , '`Average # of Absences`' ,
                '`Low Income %`' , '`High Needs #...16`' ,
                '`Percent of HS in AP`' , '`Adjusted Score`' , 'log2(1 + `Hispanic %`)')
s7 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s7, "best"))
s7
```

```
##           `Total Score`       `Attendance Rate` `Average # of Absences`
##                5.082219                5.087174                5.086187
##          `Low Income %`     `High Needs #...16`   `Percent of HS in AP`
##                5.099398                5.376610                5.059528
##        `Adjusted Score`  log2(1 + `Hispanic %`)
##                5.087224                5.100789
## attr(,"best")
## `Percent of HS in AP`
##              5.059528
```

```
preds <- c("`Economically Disadvantaged %`", "`Percent of Teachers without Waiver or
↪  Provisional License`", "`% in an Art Course`", "`% Students Completing Advanced`",
↪  "`Gen Ed %`", "`Adv Course % Math`", "`Percent of HS in AP`")
```

```
cands <- c('`Total Score`'  , '`Attendance Rate`' , '`Average # of Absences`' ,
                '`Low Income %`' , '`High Needs #...16`' ,
                '`Adjusted Score`' , 'log2(1 + `Hispanic %`)')
s8 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s8, "best"))
s8
```

```
##           `Total Score`       `Attendance Rate` `Average # of Absences`
##                5.098500                5.075122                5.075628
##          `Low Income %`     `High Needs #...16`        `Adjusted Score`
##                5.081886                5.346903                5.091001
##  log2(1 + `Hispanic %`)
##                5.003283
## attr(,"best")
## log2(1 + `Hispanic %`)
##               5.003283
```

```
preds <- c("`Economically Disadvantaged %`", "`Percent of Teachers without Waiver or
↪   Provisional License`", "`% in an Art Course`", "`% Students Completing Advanced`",
↪   "`Gen Ed %`", "`Adv Course % Math`", "`Percent of HS in AP`", "log2(1 + `Hispanic
↪   %`)")
cands <- c('`Total Score`'  , '`Attendance Rate`' , '`Average # of Absences`' ,
                '`Low Income %`' , '`High Needs #...16`' ,
                '`Adjusted Score`' )
s9 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s9, "best"))
s9
```

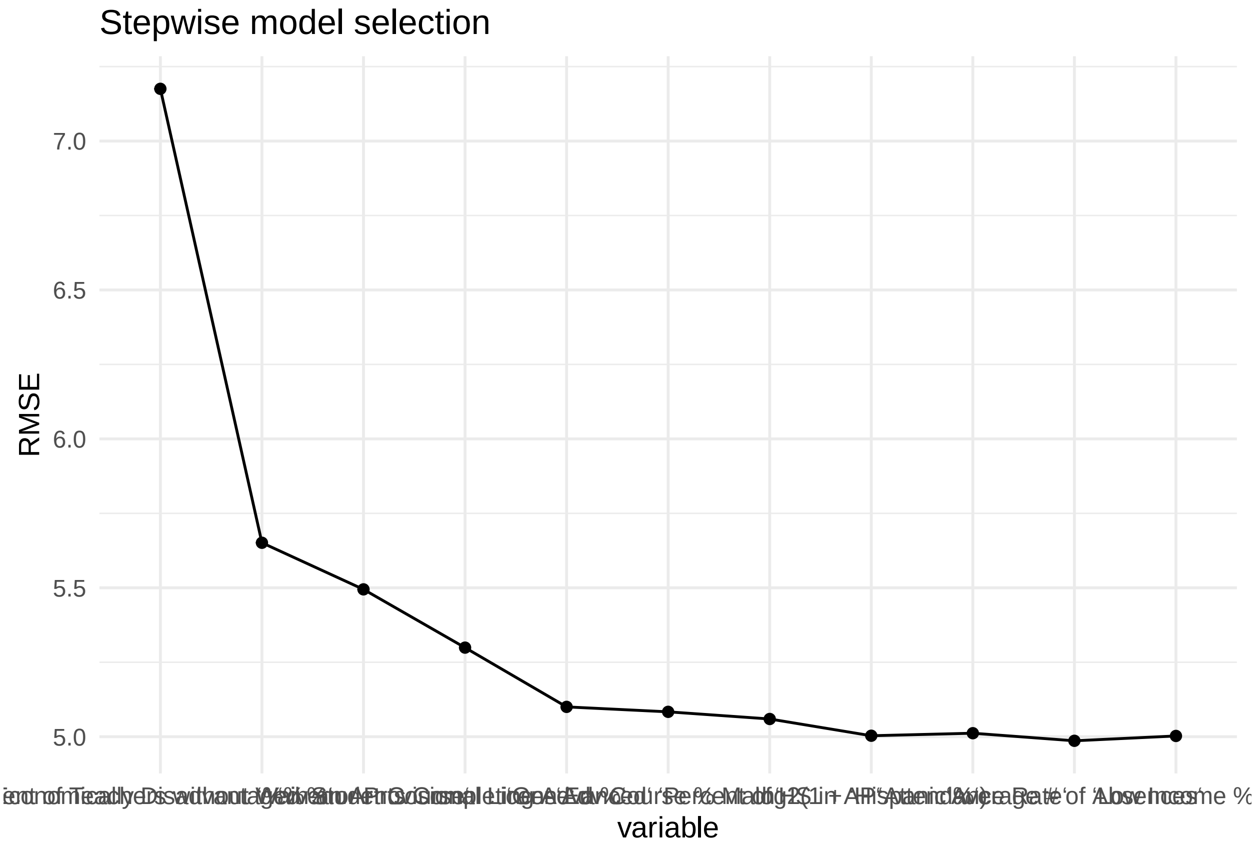
```
##           `Total Score`       `Attendance Rate` `Average # of Absences`
##                5.040602                5.011750                5.012006
##          `Low Income %`     `High Needs #...16`        `Adjusted Score`
##                5.022106                5.366384                5.026348
## attr(,"best")
## `Attendance Rate`
##           5.01175
```

```
preds <- c("`Economically Disadvantaged %`", "`Percent of Teachers without Waiver or
↪   Provisional License`", "`% in an Art Course`", "`% Students Completing Advanced`",
↪   "`Gen Ed %`", "`Adv Course % Math`", "`Percent of HS in AP`", "log2(1 + `Hispanic
↪   %`)", '`Attendance Rate`')
cands <- c('`Total Score`'  , '`Average # of Absences`' ,
                '`Low Income %`' , '`High Needs #...16`' ,
                '`Adjusted Score`' )
s10 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s10, "best"))
s10
```

```
##           `Total Score` `Average # of Absences`          `Low Income %`
##                5.038727                4.986275                5.034581
##     `High Needs #...16`        `Adjusted Score`
##                5.376055                5.025512
## attr(,"best")
## `Average # of Absences`
##                4.986275
```

```
preds <- c("`Economically Disadvantaged %`", "`Percent of Teachers without Waiver or
↪  Provisional License`", "`% in an Art Course`", "`% Students Completing Advanced`",
↪  "`Gen Ed %`", "`Adv Course % Math`", "`Percent of HS in AP`", "log2(1 + `Hispanic
↪  %`)", '`Attendance Rate`','`Average # of Absences`')
cands <- c('`Total Score`'  ,
                '`Low Income %`' , '`High Needs #...16`' ,
                '`Adjusted Score`' )
s11 <- step1("`Attending Coll./Univ. (%)`", preds, cands, edu_part)

model <- c(model, attr(s11, "best"))
s11
```

```
##        `Total Score`      `Low Income %` `High Needs #...16`    `Adjusted Score`
##             5.035559            5.002538            5.316699            5.024726
## attr(,"best")
## `Low Income %`
##       5.002538
```



```
coll_ead_model <- lm(`Attending Coll./Univ. (%)` ~ `Economically Disadvantaged %`+
↪  `Percent of Teachers without Waiver or Provisional License`+ `% in an Art Course`+
↪  `Gen Ed %`+ `Adv Course % Math`+log2(1 + `Hispanic %`), data=edu_part$train)
summary(coll_ead_model)
```

```
##
```

```
## Call:
## lm(formula = `Attending Coll./Univ. (%)` ~ `Economically Disadvantaged %` +
##     `Percent of Teachers without Waiver or Provisional License` +
##     `% in an Art Course` + `Gen Ed %` + `Adv Course % Math` +
##     log2(1 + `Hispanic %`), data = edu_part$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.4999  -3.7571  -0.1024   4.8520  12.4832
##
## Coefficients:
##                                                              Estimate Std. Error
## (Intercept)                                                   8.03812   12.55461
## `Economically Disadvantaged %`                               -0.58333    0.04697
## `Percent of Teachers without Waiver or Provisional License`   0.39606    0.14382
## `% in an Art Course`                                          0.11241    0.03003
## `Gen Ed %`                                                   21.81720    5.06045
## `Adv Course % Math`                                           0.12384    0.03308
## log2(1 + `Hispanic %`)                                        1.53075    0.65188
##                                                              t value Pr(>|t|)
## (Intercept)                                                    0.640 0.522764
## `Economically Disadvantaged %`                               -12.419  < 2e-16
## `Percent of Teachers without Waiver or Provisional License`    2.754 0.006450
## `% in an Art Course`                                           3.744 0.000239
## `Gen Ed %`                                                     4.311 2.58e-05
## `Adv Course % Math`                                            3.744 0.000238
## log2(1 + `Hispanic %`)                                         2.348 0.019870
##
## (Intercept)
## `Economically Disadvantaged %`                               ***
## `Percent of Teachers without Waiver or Provisional License` **
## `% in an Art Course`                                         ***
## `Gen Ed %`                                                   ***
## `Adv Course % Math`                                          ***
## log2(1 + `Hispanic %`)                                       *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.558 on 194 degrees of freedom
## Multiple R-squared:  0.8221, Adjusted R-squared:  0.8166
## F-statistic: 149.4 on 6 and 194 DF,  p-value: < 2.2e-16
```

```
rmse(coll_ead_model, edu_part$test)/100
```

```
## [1] 0.06973732
```

### College % model using all features

```
temp <- eduData %>%
  select(!c(`District Name`, `District Code`, `Tests Taken`, `Salary Totals`, `FTE
  ↪  Count`, `HS Enrollment`, Enrollment, `Tests Takers`, `Adjusted Score`, `Total
  ↪  Score`, `Total # of Teachers (FTE)`, `Reading / Writing`, `Math`, `% Graduated`,
  ↪  `Average # of Absences`))
```

```
base.mod <- lm(`Attending Coll./Univ. (%)` ~ 1 , data=temp)
all.mod <- lm(`Attending Coll./Univ. (%)` ~ . , data= temp)
stepMod <- step(base.mod, scope = list(lower = base.mod, upper = all.mod), direction =
↪  "both", trace = 0, steps = 1000)

shortlistedVars <- names(unlist(stepMod[[1]]))
shortlistedVars <- shortlistedVars[!shortlistedVars %in% "(Intercept)"]

print(shortlistedVars)
```

```
##  [1] "`High Needs #...16`"
##  [2] "`% Female Teachers`"
##  [3] "`White %`"
##  [4] "`Gen Ed %`"
##  [5] "`Adv Course % Math`"
##  [6] "`Economically Disadvantaged %`"
##  [7] "`Total Expenditures per Pupil`"
##  [8] "`Native American %`"
##  [9] "`% of Teachers <40`"
## [10] "`Percent of Teachers without Waiver or Provisional License`"
## [11] "`First Language Not English %`"
## [12] "`English Language Learner %`"
## [13] "`Percent of HS in AP`"
## [14] "`% of Teachers Licensed`"
## [15] "`Native Hawaiian, Pacific Islander %`"
## [16] "`% Churn`"
```

```
summary(lm(`Attending Coll./Univ. (%)` ~ `High Needs #...16` + `% Female Teachers` +
↪  `White %` + `Gen Ed %` + `Adv Course % Math` + `Economically Disadvantaged %` +
↪  `Total Expenditures per Pupil` + `Native American %` + `% of Teachers <40` + `Percent
↪  of Teachers without Waiver or Provisional License` + `First Language Not English %` +
↪  `English Language Learner %` + `Percent of HS in AP` + `% of Teachers Licensed` +
↪  `Native Hawaiian, Pacific Islander %` + `% Churn`, data = eduData))
```

```
##
## Call:
## lm(formula = `Attending Coll./Univ. (%)` ~ `High Needs #...16` +
##     `% Female Teachers` + `White %` + `Gen Ed %` + `Adv Course % Math` +
##     `Economically Disadvantaged %` + `Total Expenditures per Pupil` +
##     `Native American %` + `% of Teachers <40` + `Percent of Teachers without Waiver or Provisional L
##     `First Language Not English %` + `English Language Learner %` +
##     `Percent of HS in AP` + `% of Teachers Licensed` + `Native Hawaiian, Pacific Islander %` +
##     `% Churn`, data = eduData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7105  -3.5471  -0.3165   3.6587  13.1081
##
## Coefficients:
##                                                               Estimate
## (Intercept)                                                 -1.313e+01
## `High Needs #...16`                                         -1.935e-01
## `% Female Teachers`                                          3.680e-01
## `White %`                                                   -1.608e-01
## `Gen Ed %`                                                   1.417e+01
## `Adv Course % Math`                                          8.245e-02
## `Economically Disadvantaged %`                              -4.148e-01
## `Total Expenditures per Pupil`                              -2.592e-04
## `Native American %`                                          2.230e+00
## `% of Teachers <40`                                          1.777e-01
## `Percent of Teachers without Waiver or Provisional License`  2.797e-01
## `First Language Not English %`                              -2.225e-01
## `English Language Learner %`                                 3.039e-01
## `Percent of HS in AP`                                        9.188e-02
## `% of Teachers Licensed`                                     4.096e-01
## `Native Hawaiian, Pacific Islander %`                        3.037e+00
## `% Churn`                                                   -3.234e-01
##                                                             Std. Error t value
## (Intercept)                                                  2.852e+01  -0.460
## `High Needs #...16`                                          1.296e-01  -1.493
## `% Female Teachers`                                          8.697e-02   4.232
## `White %`                                                    4.521e-02  -3.557
## `Gen Ed %`                                                   4.281e+00   3.311
## `Adv Course % Math`                                          2.879e-02   2.863
## `Economically Disadvantaged %`                               1.224e-01  -3.388
## `Total Expenditures per Pupil`                               1.517e-04  -1.709
## `Native American %`                                          8.543e-01   2.611
## `% of Teachers <40`                                          6.691e-02   2.656
## `Percent of Teachers without Waiver or Provisional License`  1.285e-01   2.177
## `First Language Not English %`                               9.812e-02  -2.267
## `English Language Learner %`                                 1.936e-01   1.570
## `Percent of HS in AP`                                        5.937e-02   1.547
## `% of Teachers Licensed`                                     2.736e-01   1.497
## `Native Hawaiian, Pacific Islander %`                        1.979e+00   1.535
## `% Churn`                                                    2.306e-01  -1.402
##                                                             Pr(>|t|)
## (Intercept)                                                 0.645688
## `High Needs #...16`                                         0.136783
## `% Female Teachers`                                         3.32e-05 ***
## `White %`                                                   0.000454 ***
## `Gen Ed %`                                                  0.001074 **
## `Adv Course % Math`                                         0.004569 **
## `Economically Disadvantaged %`                              0.000827 ***
## `Total Expenditures per Pupil`                              0.088819 .
## `Native American %`                                         0.009620 **
## `% of Teachers <40`                                         0.008460 **
## `Percent of Teachers without Waiver or Provisional License` 0.030492 *
## `First Language Not English %`                              0.024296 *
## `English Language Learner %`                                0.117768
## `Percent of HS in AP`                                       0.123101
## `% of Teachers Licensed`                                    0.135722
## `Native Hawaiian, Pacific Islander %`                       0.126157
## `% Churn`                                                   0.162195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.749 on 235 degrees of freedom
## Multiple R-squared:  0.8671, Adjusted R-squared:  0.8581
## F-statistic: 95.85 on 16 and 235 DF,  p-value: < 2.2e-16
```

```
collegeMod <- lm(`Attending Coll./Univ. (%)` ~ `White %` + `Gen Ed %` + `Adv Course %
↪  Math` + `Economically Disadvantaged %` + `Total Expenditures per Pupil` + `Native
↪  American %` + `% of Teachers <40` + `Percent of Teachers without Waiver or
↪  Provisional License` + `First Language Not English %`, data = edu_part$train)

summary(collegeMod)
```

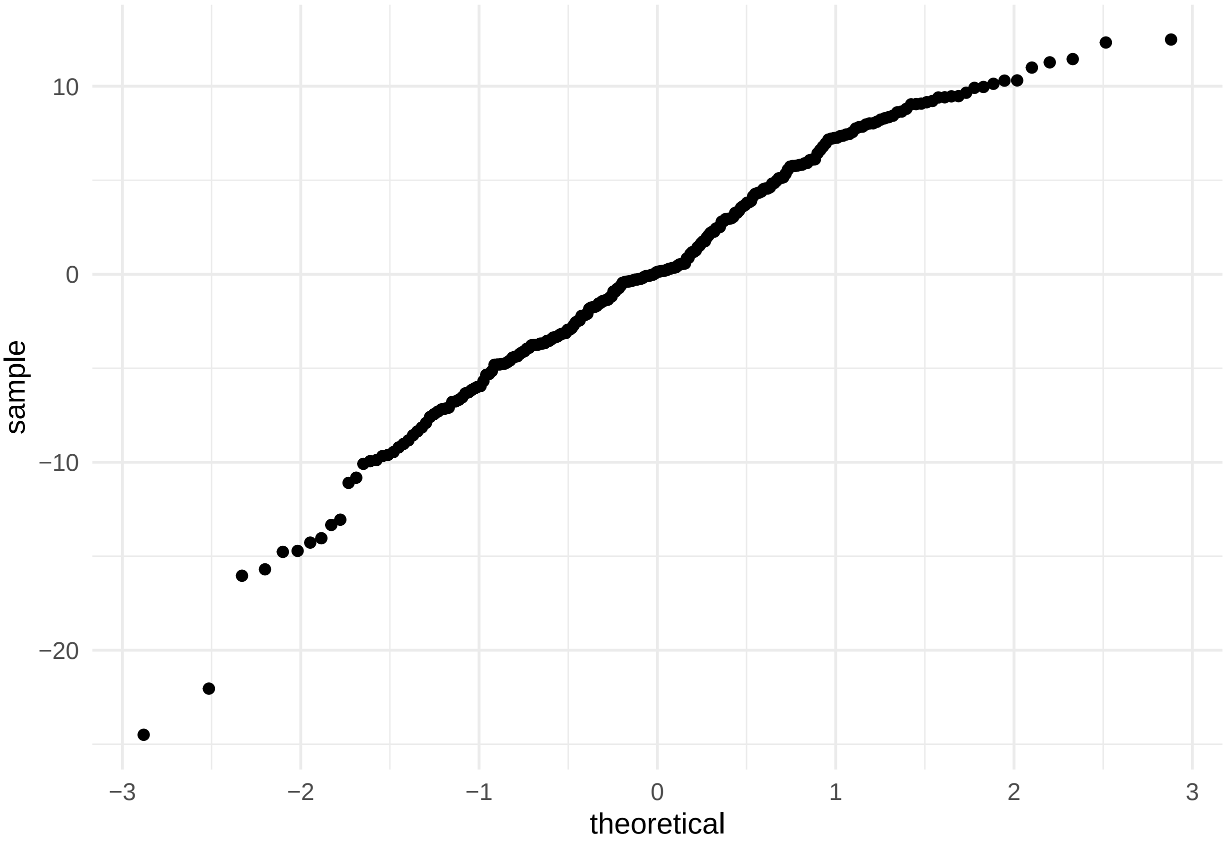
```
## 
## Call:
## lm(formula = `Attending Coll./Univ. (%)` ~ `White %` + `Gen Ed %` + 
##     `Adv Course % Math` + `Economically Disadvantaged %` + `Total Expenditures per Pupil` + 
##     `Native American %` + `% of Teachers <40` + `Percent of Teachers without Waiver or Provisional Li
##     `First Language Not English %`, data = edu_part$train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -20.8081  -4.3347   0.1277   4.3722  15.3439 
## 
## Coefficients:
##                                                                 Estimate
## (Intercept)                                                   27.5716391
## `White %`                                                     -0.1911560
## `Gen Ed %`                                                    28.4832304
## `Adv Course % Math`                                            0.1039088
## `Economically Disadvantaged %`                                -0.6204821
## `Total Expenditures per Pupil`                                -0.0006251
## `Native American %`                                            4.0221167
## `% of Teachers <40`                                            0.2301134
## `Percent of Teachers without Waiver or Provisional License`    0.4826204
## `First Language Not English %`                                -0.1554236
##                                                               Std. Error t value
## (Intercept)                                                   15.2263862   1.811
## `White %`                                                      0.0548503  -3.485
## `Gen Ed %`                                                     3.8837985   7.334
## `Adv Course % Math`                                            0.0325913   3.188
## `Economically Disadvantaged %`                                 0.0419313 -14.798
## `Total Expenditures per Pupil`                                 0.0001563  -3.999
## `Native American %`                                            1.6413024   2.451
## `% of Teachers <40`                                            0.0803157   2.865
## `Percent of Teachers without Waiver or Provisional License`    0.1387868   3.477
## `First Language Not English %`                                 0.0712041  -2.183
##                                                               Pr(>|t|)    
## (Intercept)                                                   0.071746 .  
## `White %`                                                     0.000610 ***
## `Gen Ed %`                                                    6.19e-12 ***
## `Adv Course % Math`                                           0.001673 ** 
## `Economically Disadvantaged %`                                 < 2e-16 ***
## `Total Expenditures per Pupil`                                9.08e-05 ***
## `Native American %`                                           0.015163 *  
## `% of Teachers <40`                                           0.004636 ** 
## `Percent of Teachers without Waiver or Provisional License`   0.000627 ***
## `First Language Not English %`                                0.030271 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.257 on 191 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.833
## F-statistic: 111.9 on 9 and 191 DF,  p-value: < 2.2e-16
```

```
rmse(collegeMod, edu_part$test)/100
```

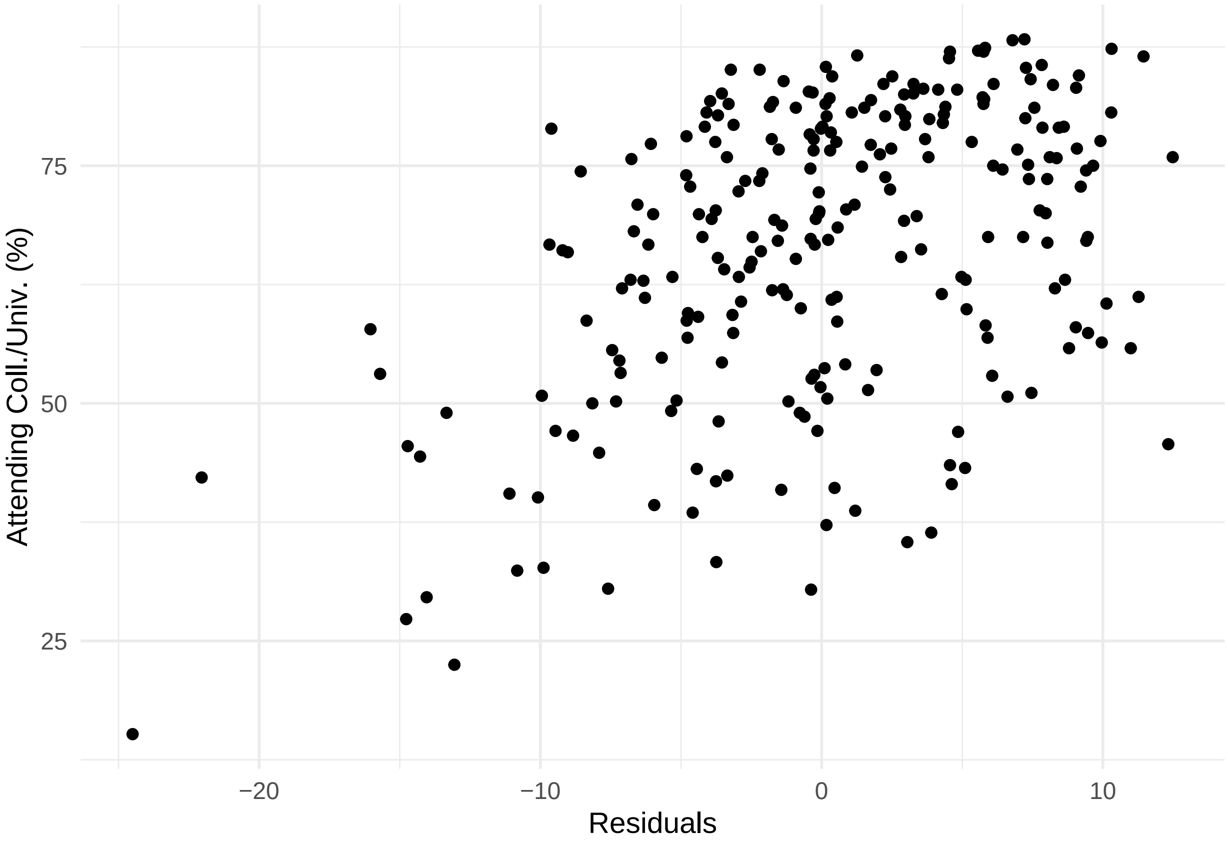
```
## [1] 0.05867334
```

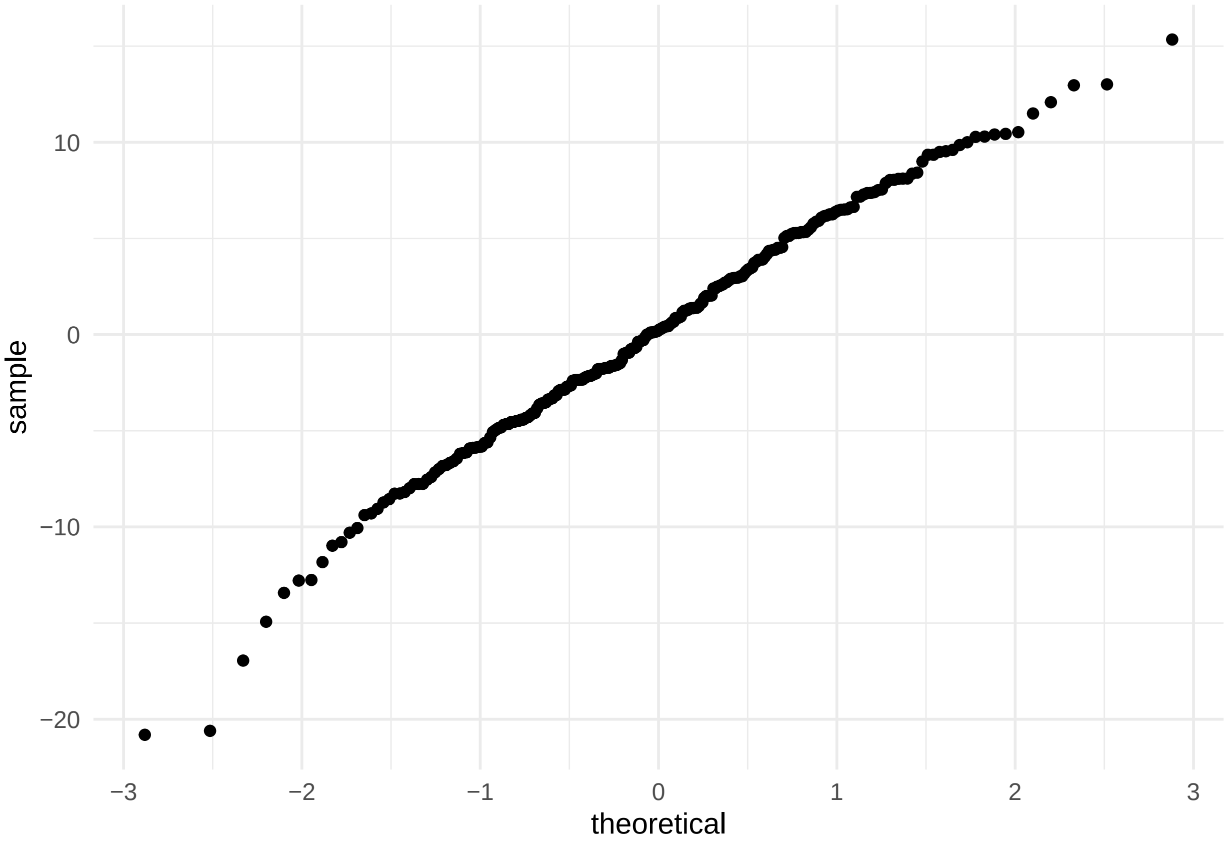
```
car::vif(collegeMod)
```

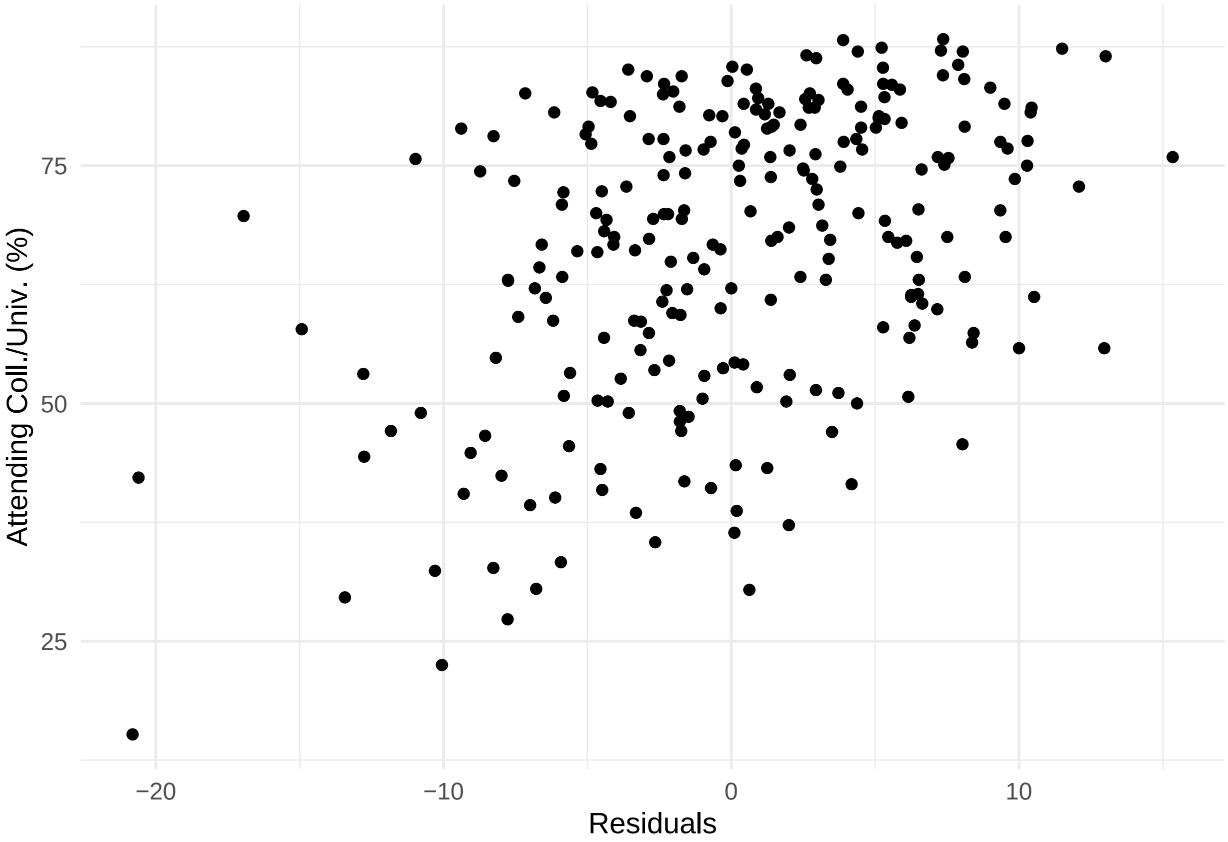
```
##                                                    `White %`
##                                                     5.670571
##                                                   `Gen Ed %`
##                                                     1.842293
##                                          `Adv Course % Math`
##                                                     1.717061
##                               `Economically Disadvantaged %`
##                                                     2.809041
##                               `Total Expenditures per Pupil`
##                                                     1.495016
##                                          `Native American %`
##                                                     1.164458
##                                          `% of Teachers <40`
##                                                     1.534741
## `Percent of Teachers without Waiver or Provisional License`
##                                                     1.968482
##                               `First Language Not English %`
##                                                     5.039596
```



**Diagnostics**




10
0
−10
−20
sample
−3
−2
−1
0
1
2
3
theoretical


75
50
25
Attending Coll./Univ. (%)
−20
−10
0
10
Residuals

# Comparing models

## SAT

EDA Features

```
summary(satEDA)
```

```
## 
## Call:
## lm(formula = `Total Score` ~ `Adv Course % Math` + `Percent of HS in AP` + 
##     log1p(`Asian %`), data = edu_part$train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -119.871  -36.917   -1.478   35.348  191.419 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           971.8867    13.2056  73.596  < 2e-16 ***
## `Adv Course % Math`     0.7130     0.2756   2.587   0.0105 *  
## `Percent of HS in AP`   3.4675     0.4902   7.074 3.09e-11 ***
## log1p(`Asian %`)       32.0407     5.3970   5.937 1.43e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 53.49 on 183 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.5439, Adjusted R-squared:  0.5364 
## F-statistic: 72.74 on 3 and 183 DF,  p-value: < 2.2e-16
```

```
rmse(satEDA, edu_part$valid)/1600
```

```
## [1] 0.02904024
```

```
AIC(satEDA)
```

```
## [1] 2024.989
```

All available features

```
summary(totalScoreMod)
```

```
## 
## Call:
## lm(formula = `Total Score` ~ `Low Income %` + `Asian %` + `% Disciplined` + 
##     `Multi-Race, Non-Hispanic %` + `Teacher Retention Rate`, 
##     data = edu_part$train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -126.482  -27.023   -4.302   26.645  116.071 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  1361.1159    72.2982  18.826  < 2e-16 ***
## `Low Income %`                 -2.4706     0.1866 -13.238  < 2e-16 ***
## `Asian %`                       2.6348     0.5437   4.846 2.70e-06 ***
## `% Disciplined`               -19.5543     3.4167  -5.723 4.27e-08 ***
## `Multi-Race, Non-Hispanic %`    8.2980     1.9377   4.282 3.00e-05 ***
## `Teacher Retention Rate`       -2.0077     0.7934  -2.531   0.0122 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.93 on 181 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.6958, Adjusted R-squared:  0.6874
## F-statistic:  82.8 on 5 and 181 DF,  p-value: < 2.2e-16
```

```
rmse(totalScoreMod, edu_part$valid)/1600
```

```
## [1] 0.02950406
```

```
AIC(totalScoreMod)
```

```
## [1] 1953.247
```

## Graduation Rate

EDA Features

```
summary(grad_ead_model)
```

```
##
## Call:
## lm(formula = `% Graduated` ~ `% Dropout All Grades` + `% Churn`,
##     data = edu_part$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7146  -1.4645   0.1963   1.6727  12.4051
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             99.8627     0.6250 159.772  < 2e-16 ***
## `% Dropout All Grades`  -3.0105     0.2616 -11.510  < 2e-16 ***
## `% Churn`               -0.6190     0.1095  -5.652 5.48e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.552 on 198 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7484
## F-statistic: 298.4 on 2 and 198 DF,  p-value: < 2.2e-16
```

```
rmse(grad_ead_model, edu_part$valid)/100
```

```
## [1] 0.02553979
```

```
AIC(grad_ead_model)
```

```
## [1] 1084.882
```

All available features

```
rmse(gradRateMod, edu_part$valid)/100
```

```
## [1] 0.02733896
```

```
summary(gradRateMod)
```

```
##
## Call:
## lm(formula = `% Graduated` ~ `% Dropout All Grades` + `% Intake` +
##     `Economically Disadvantaged %`, data = edu_part$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9034  -1.5412   0.2517   1.6701  11.9253
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    100.23790    0.52567 190.686  < 2e-16 ***
## `% Dropout All Grades`          -2.75253    0.23701 -11.613  < 2e-16 ***
## `% Intake`                      -0.72376    0.15657  -4.622 6.85e-06 ***
## `Economically Disadvantaged %`  -0.07969    0.02002  -3.980 9.70e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.336 on 197 degrees of freedom
## Multiple R-squared:  0.7813, Adjusted R-squared:  0.778
## F-statistic: 234.6 on 3 and 197 DF,  p-value: < 2.2e-16
```

```
AIC(gradRateMod)
```

```
## [1] 1060.678
```

## College %

EDA features

```
summary(coll_ead_model)
```

```
##
## Call:
## lm(formula = `Attending Coll./Univ. (%)` ~ `Economically Disadvantaged %` +
##     `Percent of Teachers without Waiver or Provisional License` +
##     `% in an Art Course` + `Gen Ed %` + `Adv Course % Math` +
##     log2(1 + `Hispanic %`), data = edu_part$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.4999  -3.7571  -0.1024   4.8520  12.4832
##
## Coefficients:
##                                                              Estimate Std. Error
## (Intercept)                                                   8.03812   12.55461
## `Economically Disadvantaged %`                               -0.58333    0.04697
## `Percent of Teachers without Waiver or Provisional License`   0.39606    0.14382
## `% in an Art Course`                                          0.11241    0.03003
## `Gen Ed %`                                                  21.81720    5.06045
## `Adv Course % Math`                                          0.12384    0.03308
## log2(1 + `Hispanic %`)                                       1.53075    0.65188
##                                                             t value Pr(>|t|)
## (Intercept)                                                   0.640 0.522764
## `Economically Disadvantaged %`                              -12.419  < 2e-16
## `Percent of Teachers without Waiver or Provisional License`   2.754 0.006450
## `% in an Art Course`                                          3.744 0.000239
## `Gen Ed %`                                                    4.311 2.58e-05
## `Adv Course % Math`                                           3.744 0.000238
## log2(1 + `Hispanic %`)                                        2.348 0.019870
##
## (Intercept)
## `Economically Disadvantaged %`                              ***
## `Percent of Teachers without Waiver or Provisional License` **
## `% in an Art Course`                                        ***
## `Gen Ed %`                                                  ***
## `Adv Course % Math`                                         ***
## log2(1 + `Hispanic %`)                                      *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.558 on 194 degrees of freedom
## Multiple R-squared:  0.8221, Adjusted R-squared:  0.8166
## F-statistic: 149.4 on 6 and 194 DF,  p-value: < 2.2e-16
```

```
rmse(coll_ead_model, edu_part$valid)/100
```

```
## [1] 0.05104855
```

```
AIC(coll_ead_model)
```

```
## [1] 1335.346
```

All available features

```
summary(collegeMod)
```

```
##
## Call:
## lm(formula = `Attending Coll./Univ. (%)` ~ `White %` + `Gen Ed %` +
##     `Adv Course % Math` + `Economically Disadvantaged %` + `Total Expenditures per Pupil` +
##     `Native American %` + `% of Teachers <40` + `Percent of Teachers without Waiver or Provisional L
##     `First Language Not English %`, data = edu_part$train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.8081  -4.3347   0.1277   4.3722  15.3439
##
## Coefficients:
##                                                               Estimate
## (Intercept)                                                 27.5716391
## `White %`                                                   -0.1911560
## `Gen Ed %`                                                  28.4832304
## `Adv Course % Math`                                          0.1039088
## `Economically Disadvantaged %`                              -0.6204821
## `Total Expenditures per Pupil`                              -0.0006251
## `Native American %`                                          4.0221167
## `% of Teachers <40`                                          0.2301134
## `Percent of Teachers without Waiver or Provisional License`  0.4826204
## `First Language Not English %`                              -0.1554236
##                                                             Std. Error t value
## (Intercept)                                                 15.2263862   1.811
## `White %`                                                    0.0548503  -3.485
## `Gen Ed %`                                                   3.8837985   7.334
## `Adv Course % Math`                                          0.0325913   3.188
## `Economically Disadvantaged %`                                0.0419313 -14.798
## `Total Expenditures per Pupil`                               0.0001563  -3.999
## `Native American %`                                          1.6413024   2.451
## `% of Teachers <40`                                          0.0803157   2.865
## `Percent of Teachers without Waiver or Provisional License`  0.1387868   3.477
## `First Language Not English %`                               0.0712041  -2.183
##                                                             Pr(>|t|)
## (Intercept)                                                 0.071746 .
## `White %`                                                   0.000610 ***
## `Gen Ed %`                                                  6.19e-12 ***
## `Adv Course % Math`                                         0.001673 **
## `Economically Disadvantaged %`                                < 2e-16 ***
## `Total Expenditures per Pupil`                              9.08e-05 ***
## `Native American %`                                         0.015163 *
## `% of Teachers <40`                                         0.004636 **
## `Percent of Teachers without Waiver or Provisional License` 0.000627 ***
## `First Language Not English %`                              0.030271 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.257 on 191 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.833
## F-statistic: 111.9 on 9 and 191 DF,  p-value: < 2.2e-16
```

```
rmse(collegeMod, edu_part$valid)/100
```

```
## [1] 0.05943887
```

```
AIC(collegeMod)
```

```
## [1] 1319.325
```