# DS5110:Group-Project

## Importing packages

```r
library(dplyr)
library(readr)
library(ggplot2)
library(tidyr)
library(gridExtra)
library(grid)
library(stringr)
library(tidyverse)
```

## Importing Data

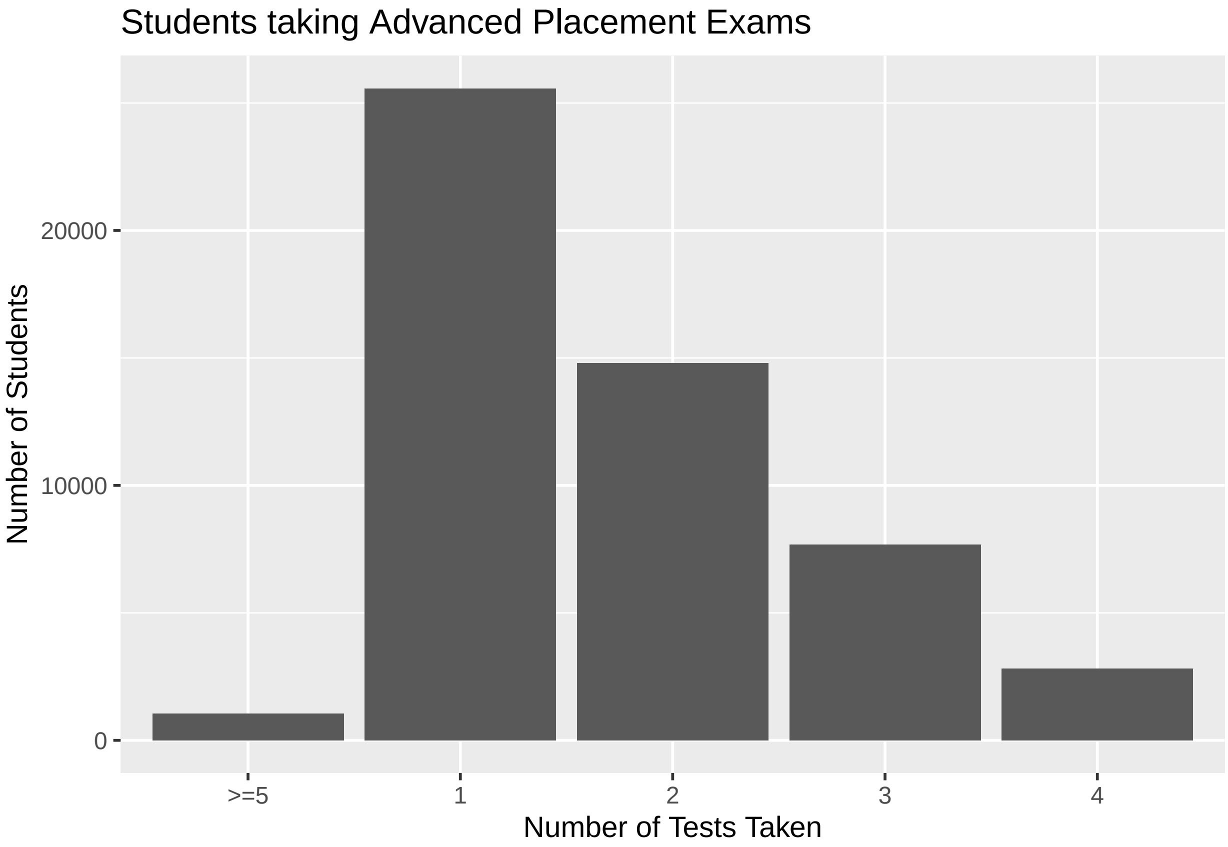
```r
d <- "Datasets"
advCourse <- read_csv(file.path(d,"AdvancedCourseCompletion.csv"))
ap_part <- read_csv(file.path(d,"ap_participation.csv"))
ap_perf <- read_csv(file.path(d,"ap_performance.csv"))
art <- read_csv(file.path(d,"artcourse.csv"))
attendance <- read_csv(file.path(d,"attendance.csv"))
attrition <- read_csv(file.path(d,"AttritionReport.csv"))

classSizeByClass <- read_csv(file.path(d,"ClassSizebyGenPopulation.csv"))
classSizeByRace <- read_csv(file.path(d,"ClassSizebyRaceEthnicity.csv"))

dropOut <- read_csv(file.path(d,"dropout.csv"))

eduAge <- read_csv(file.path(d,"EducatorsbyAgeGroupsReport.csv"))
enrollByGrade <- read_csv(file.path(d,"enrollmentbygrade.csv"))

gradeStaff <- read_csv(file.path(d,"gradestaffing.csv"))
gradRate <- read_csv(file.path(d,"gradrates.csv"))
college <- read_csv(file.path(d,"Gradsattendingcollege.csv"))
mobilityRate <- read_csv(file.path(d,"mobilityrates.csv"))

StudReten <- read_csv(file.path(d,"retention2021.csv"))

sat <- read_csv(file.path(d,"sat_performance.csv"))
selectPop <- read_csv(file.path(d,"selectedpopulations.csv"))

daysMissed <- read_csv(file.path(d,"ssdr_days_missed.csv"))
eduGen <- read_csv(file.path(d,"staffracegender.csv"))
staffReten<- read_csv(file.path(d,"staffingretention.csv"))
discipline <- read_csv(file.path(d,"StudentDisciplineDataReport.csv"))

teachData <- read_csv(file.path(d,"teacherdata.csv"))
teachProg <- read_csv(file.path(d,"Teacherprogramarea.csv"))
```

```
teacherSalary <- read_csv(file.path(d,"TeacherSalaries.csv"))
```

## EDA

### 1. Number of students who took one or more Advanced Placement exams.

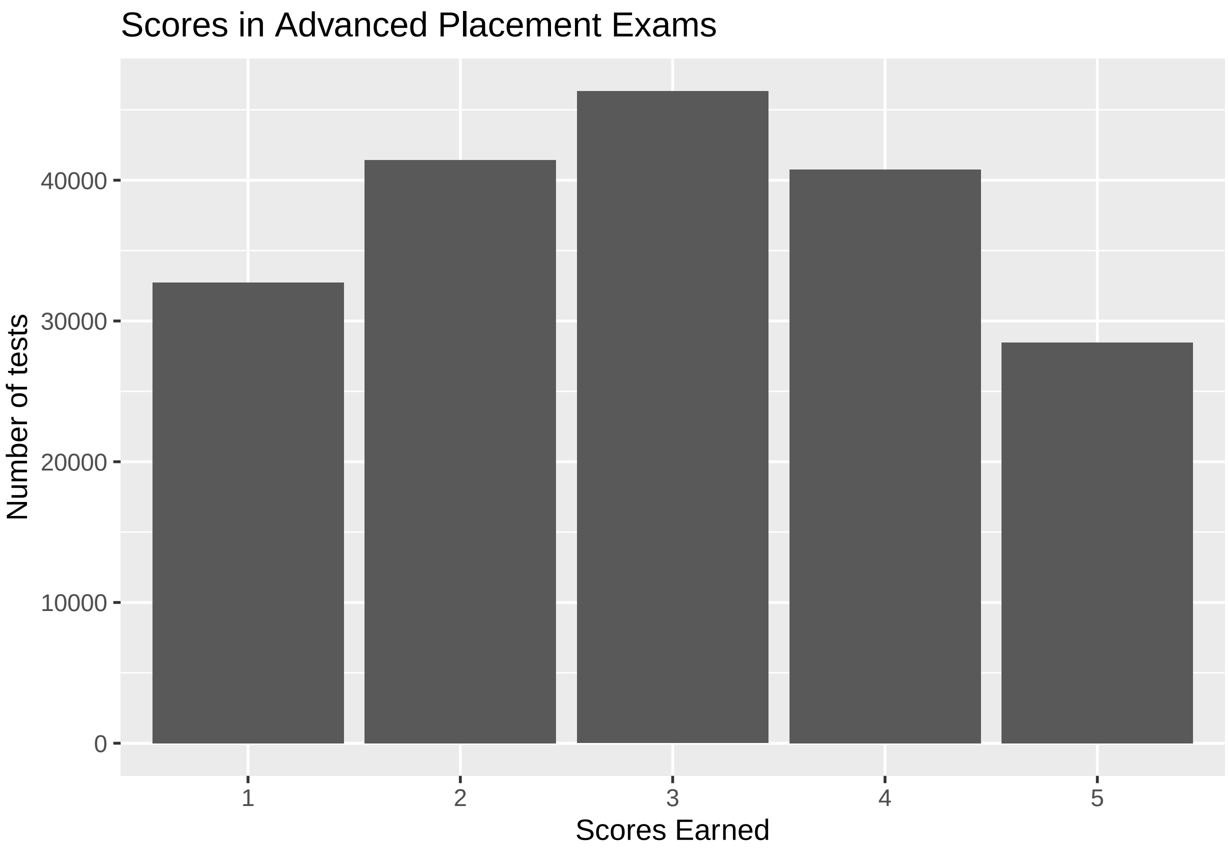
```
ap_part2 <- ap_part |>
  rename(`1`=`One Test` ,`2`=`Two Tests`,`3`=`Three Tests`
                            , `4`=`Four Tests`,`>=5`=`Five or More Tests`) |>
  select(!c(`Tests Taken`,`Tests Takers`)) |>
  filter(`District Code`!="00000000")

pivot_longer(ap_part2, cols=c(`1`,`2`,`3`,`4`,`>=5`), names_to = "TestsTaken", values_to ="TestTakers")
  group_by(TestsTaken) |>
  summarise(TestTakers=sum(TestTakers, na.rm=TRUE)) |>
  ggplot( mapping=aes(x=`TestsTaken`,y=`TestTakers`)) +
  geom_bar(stat = "identity") +
  labs(title="Students taking Advanced Placement Exams",
       x="Number of Tests Taken", y="Number of Students")
```



It shows that most of the students took placements exams **only once** while only less that 2500 students had to take 5 or more exams.

**2. Percentage of tests taken by students with each possible score on the Advanced Placement exam.**

```
ap_perf <- ap_perf |>
  select(!c(`% Score 1-2`,`% Score 3-5`)) |>
  filter(`District Code`!=0)
ap_perf <- ap_perf |> rename(`1`=`Score=1` ,`2`=`Score=2`,`3`=`Score=3`
                             , `4`=`Score=4`,`5`=`Score=5`)

pivot_longer(ap_perf, cols=c(`1`,`2`,`3`,`4`,`5`),
             names_to = "Scores",
             values_to ="# of Tests") |>
  group_by(`Scores`) |>
  summarise(`Total`=sum(`# of Tests`, na.rm=TRUE)) |>
  ggplot( mapping=aes(x=`Scores`,y=`Total`)) +
  geom_bar(stat = "identity") +
  labs(title="Scores in Advanced Placement Exams",
       x="Scores Earned", y="Number of tests")
```

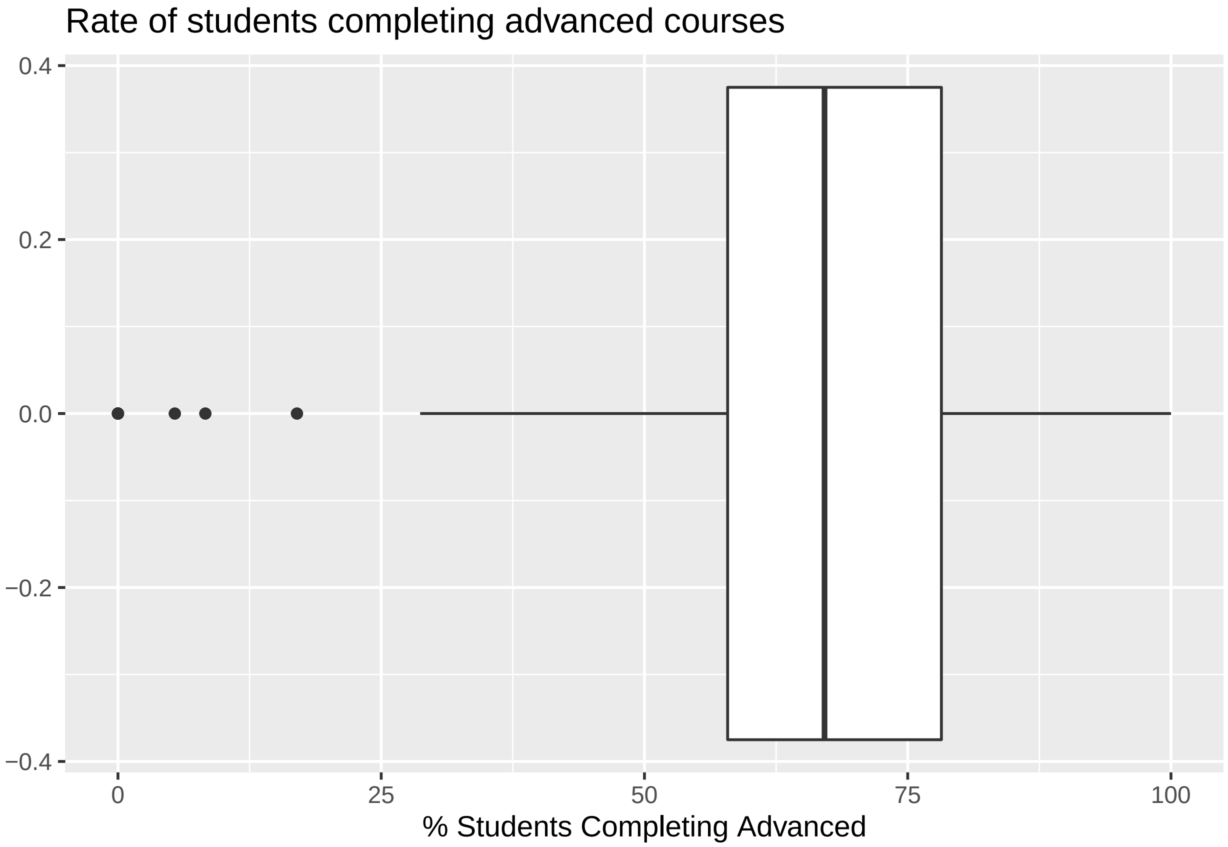


Above plot shows that most of the tests had a score of 3 while least number of tests had a score of 5. Since most of the students had taken the exam only once, it could be possible that most of them had a score of 3.

**3. The rate of Grade 11 and 12 students completing advanced courses by subject area.**

```
advCourse <- advCourse %>%
  select(!c(`Ch 74 Secondary Cooperative Program`,`# Students Completing Advanced`)) |>
  filter(`District Code`!="00000000")
```

```
advCourse |> ggplot( mapping=aes(`% Students Completing Advanced`)) +
  geom_boxplot() +
  labs(title="Rate of students completing advanced courses")
```



It shows that for some of the districts, all students completed the advanced courses while minimum rate was around 25% students. On an average, more than 62.5% of the students in 305 district entries were able to complete the courses.

Districts in which 100% or 0% students completed advanced courses .

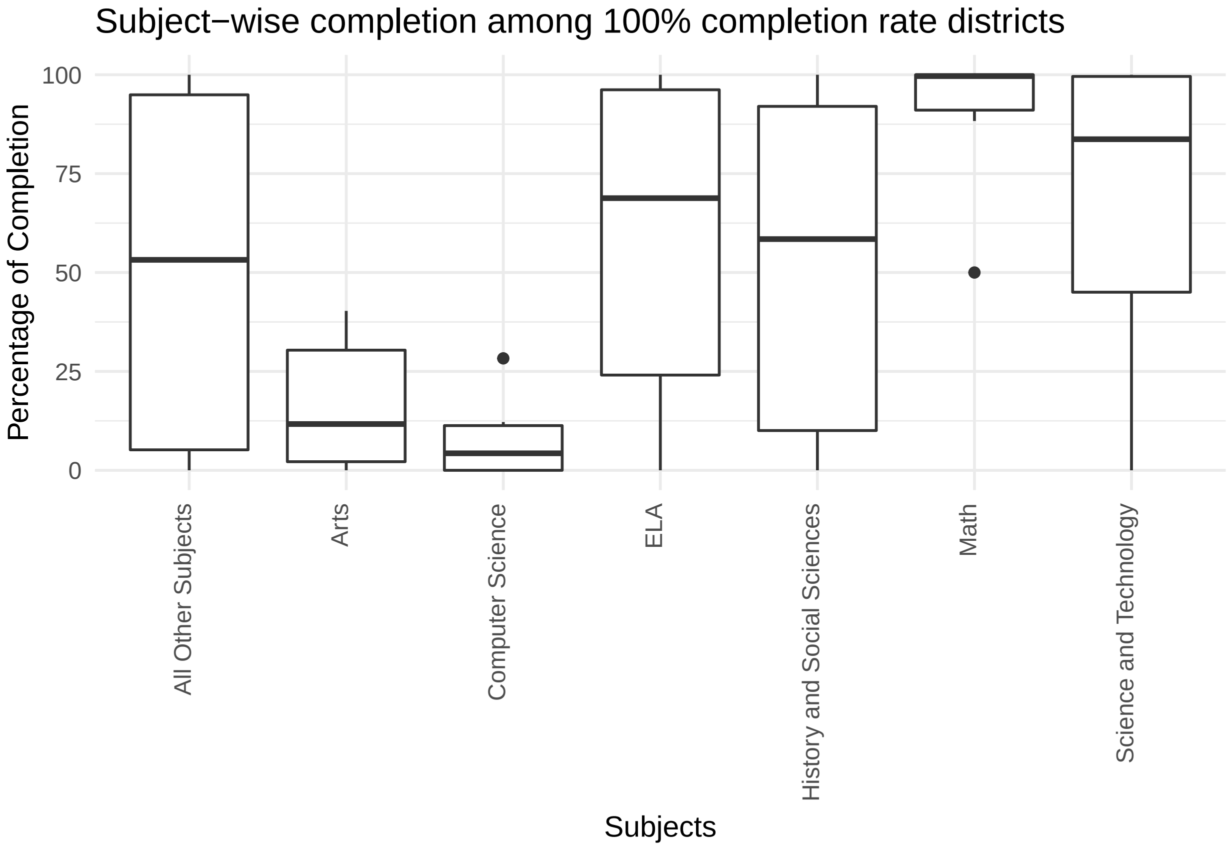
```
advCourse |> select(`District Name`,`District Code`,
                    `% Students Completing Advanced`,
                    `# Grade 11 and 12 Students`) |>
  filter(`% Students Completing Advanced` == 100.0 |
         `% Students Completing Advanced` == 0 ) |>
  rename(`% Completion` =`% Students Completing Advanced`,`Student Count`=`# Grade 11 and 12 Students`)
```

```
## # A tibble: 10 x 4
##    `District Name`              `District Code` `% Completion` `Student Count`
##    <chr>                        <chr>                     <dbl>           <dbl>
##  1 Baystate Academy Charter Publ~ 35020000                  100             102
##  2 Lowell Middlesex Academy Char~ 04580000                    0              44
##  3 Ma Academy for Math and Scien~ 04680000                  100              92
##  4 Martha's Vineyard Charter (Di~ 04660000                  100              22
##  5 Phoenix Academy Public Charte~ 35180000                    0              27
##  6 Phoenix Academy Public Charte~ 35080000                    0              19
##  7 Phoenix Charter Academy (Dist~ 04930000                    0              56
##  8 Pioneer Valley Chinese Immers~ 04970000                  100              68
##  9 Saugus                         02620000                 100             343
## 10 Sturgis Charter Public (Distr~ 04890000                 100             407
```

It can be noticed that the districts with 0% completion has quite low number of students in grade 11 and 12 and except **Martha's Vineyard Charter (District)** and **Pioneer Valley Chinese Immersion Charter (District)** all other schools with 100% completion has comparatively high student count.

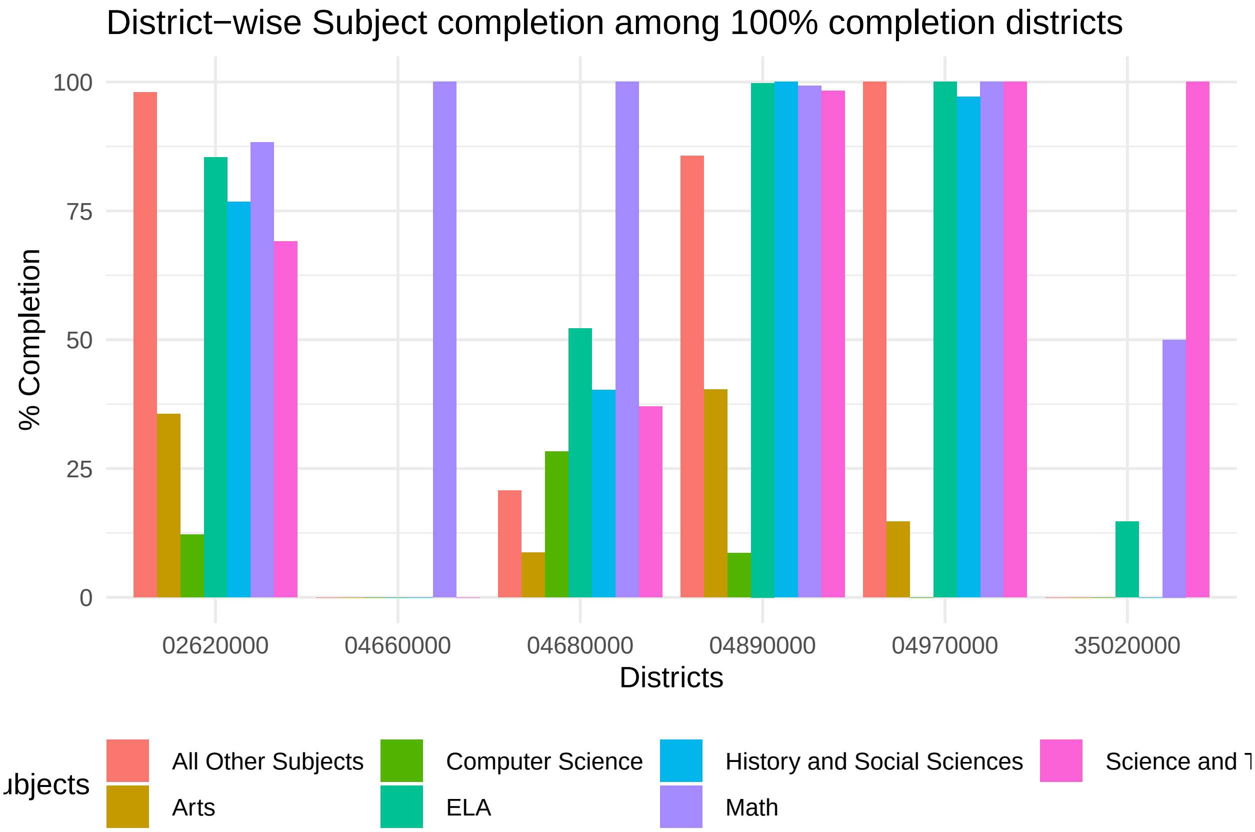
Plotting subject-wise Completions.

```
advCourse100 <- advCourse %>%
  rename(`% Science and Technology`=`% Science and Technology...8`,
         `% Computer Science` = `% Science and Technology...9`) %>%
  pivot_longer( cols=c(`% ELA`,`% Math`,`% Science and Technology`,
                               `% Computer Science`,
                             `% History and Social Sciences`,`% Arts`,
                             `% All Other Subjects`),
             names_to = "Subject",
             values_to ="% Completion") |>
  mutate(Subject=gsub("%","",Subject)) |>
  filter(`% Students Completing Advanced` == 100.0)

advCourse100 |>
  ggplot( mapping=aes(x=factor(`Subject`),y=`% Completion`)) +
  geom_boxplot() +
  labs(title="Subject-wise completion among 100% completion rate districts",
       x="Subjects", y="Percentage of Completion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.1, hjust=1))
```



Clearly, the highest completion has been for Math while surprisingly for Computer Science, it has been the least.

```
advCourse100|>
  ggplot( mapping=aes(x=factor(`District Code`),y=`% Completion`,
                      fill = `Subject`))+
  geom_bar(position="dodge", stat="identity")+
  theme_minimal()+
  theme(legend.position = "bottom") +
  labs(title="District-wise Subject completion among 100% completion districts",
       x="Districts", y="% Completion",
       fill="Subjects")
```



For district code **4660000**, only math has been completed and which is why it seems to join the 100% completion club! Since, we noticed before, it has only 22 students which is really less compared to rest of the other schools that had 100% completion rate. District code **4970000**, has done really well as it only had 68 students as compared to district code **4890000**, that had 407 students!

### 4. Dropouts

```
dropOut |> rename(`# Enrolled 9-12` =`# Enrolled Grades 09 through 12`,
                  `Dropout`=`# Dropout All Grades`) |>
 select(`District Code`,`# Enrolled 9-12`,`Dropout`,`District Name`) |> drop_na() |>
  filter(`District Code`!="00000000") |>
  filter(`Dropout`== max(Dropout))
```

```
## # A tibble: 1 x 4
##   `District Code` `# Enrolled 9-12` Dropout `District Name`
##   <chr>                       <dbl>   <dbl> <chr>
## 1 00350000                    14342     292 Boston
```

Boston has the maximum number of dropouts!

**Analyzing data with the responses.**

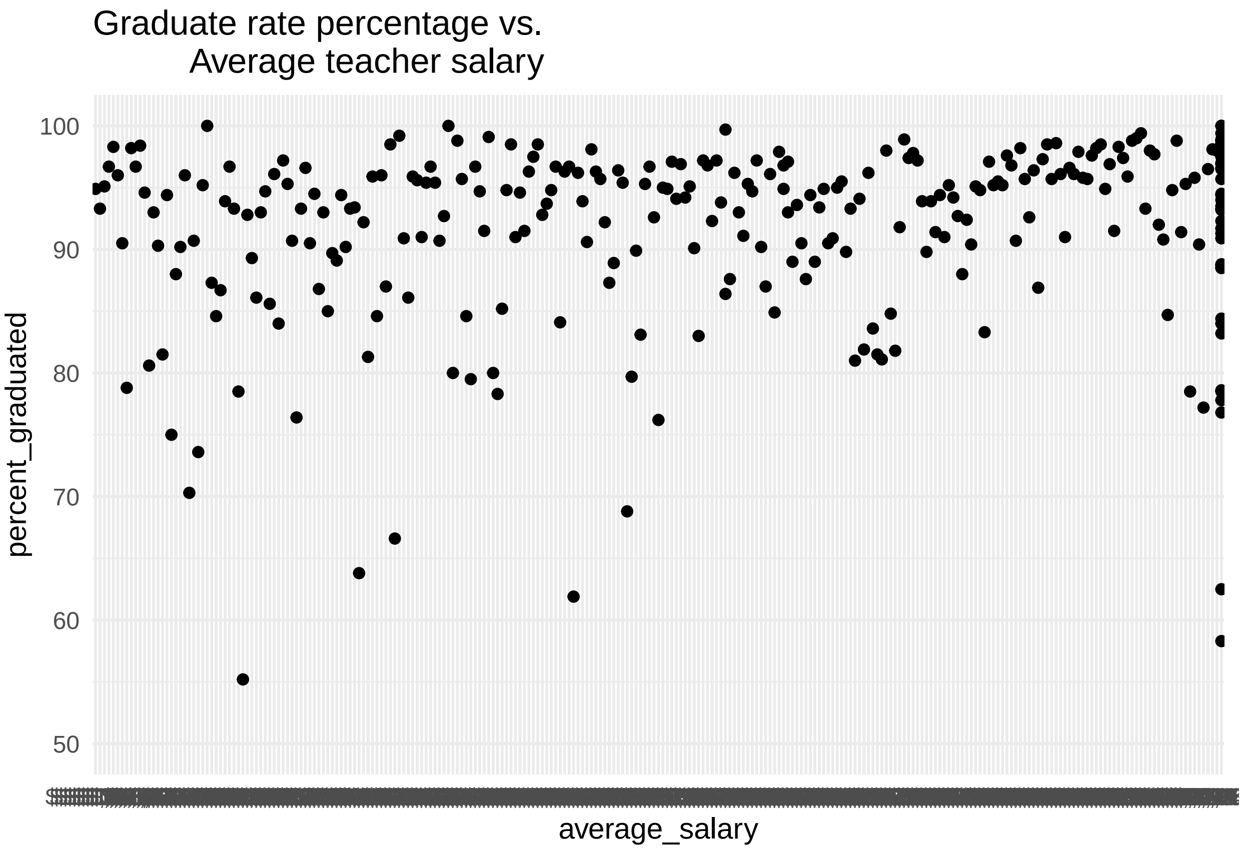
### Tran’s work starts here.

```
eduGen <- eduGen %>%
  rename(
    'District Code' = 'District/School Code')

# daysMissed <- daysMissed %>%
#   rename(
#     'District_code' = 'District Code')
#
# selectPop <- selectPop %>%
#   rename(
#     'District_code' = 'District Code')
#
#
# mobilityRate <- mobilityRate %>%
#   rename(
#     'District_code' = 'District Code')
```

## 1. Teacher Salary vs. Graduation Rate

```
gradRate_teacherSalary <- left_join(gradRate, teacherSalary, by="District Code")

gradRate_teacherSalary <- gradRate_teacherSalary %>%
  rename(
    'percent_graduated' = '% Graduated')

gradRate_teacherSalary <- gradRate_teacherSalary %>%
  rename(
    'average_salary' = 'Average Salary')

# print(gradRate_teacherSalary)
```

```
gradRate_teacherSalary_graph <-
  ggplot(gradRate_teacherSalary, aes(x = average_salary,
                                     y = percent_graduated )) + geom_point() +
  theme_minimal() +
  ggtitle("Graduate rate percentage vs.
          Average teacher salary") + geom_smooth(method=lm) + ylim(50,100)

gradRate_teacherSalary_graph
```

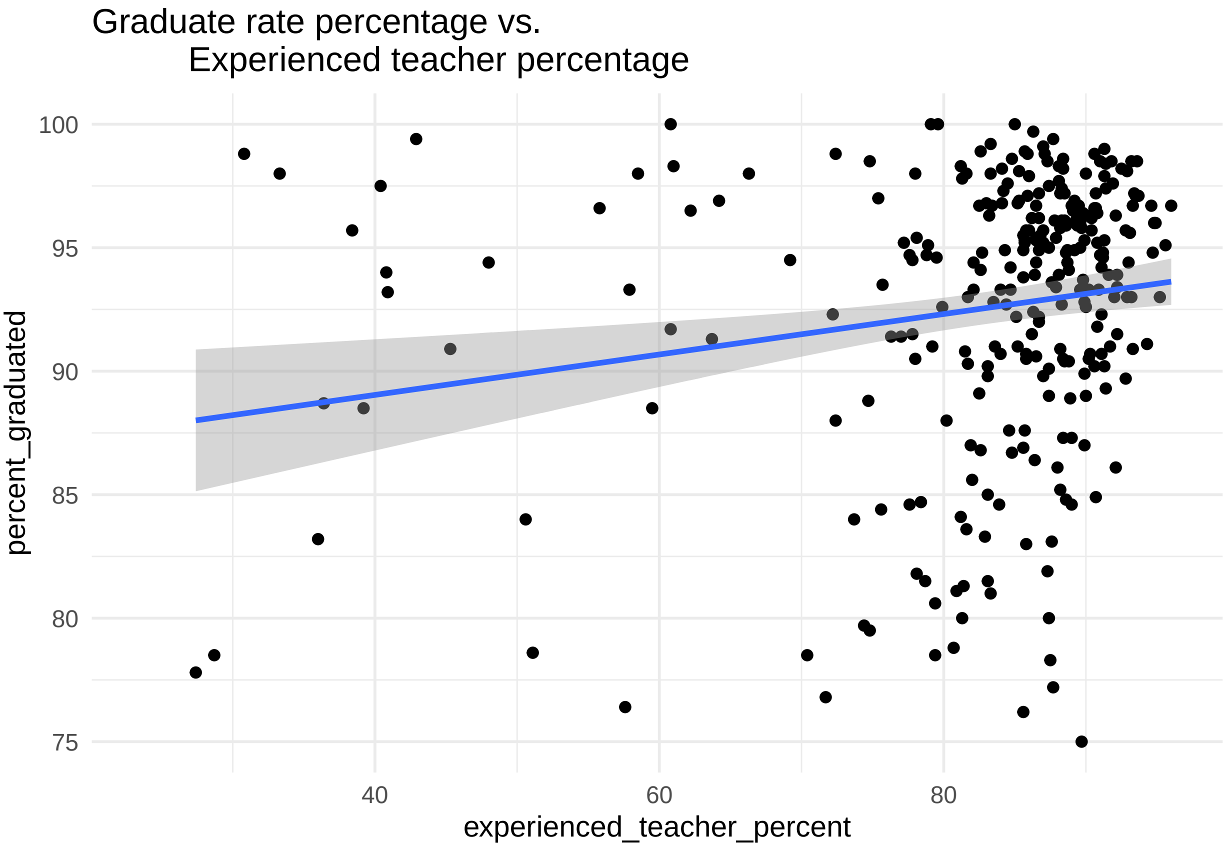


## 2. Teacher Data vs. Graduation Rate

```r
gradRate_teacherData <- left_join(gradRate, teachData, by="District Code")

gradRate_teacherData <- gradRate_teacherData %>%
  rename(
    'percent_graduated' = '% Graduated')

gradRate_teacherData <- gradRate_teacherData %>%
  rename(
    'experienced_teacher_percent' = 'Percent of Experienced Teachers')

gradRate_teacherData <- gradRate_teacherData %>%
  rename(
    'student_teacher_ratio' = 'Student / Teacher Ratio')

# print(gradRate_teacherData)
```

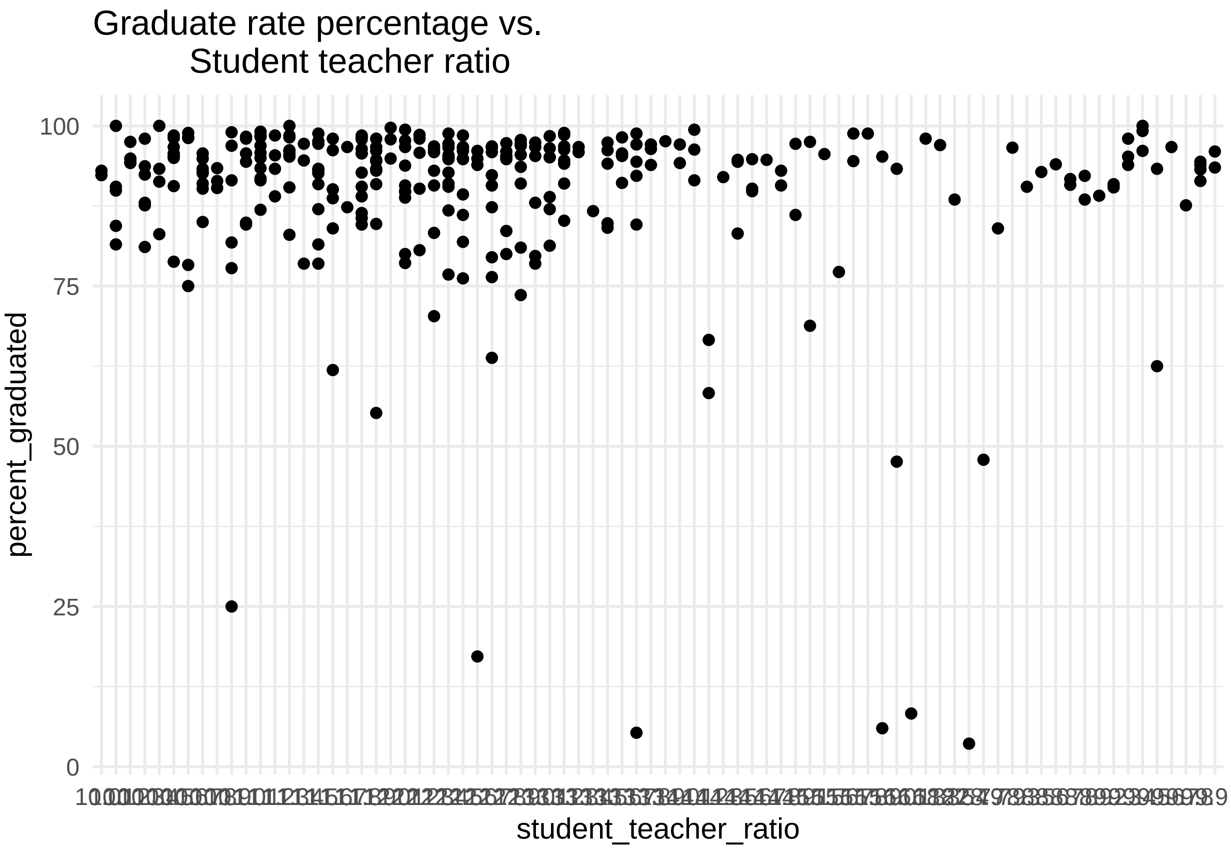
```r
gradRate_teacherData <- gradRate_teacherData %>%
      mutate_at("student_teacher_ratio", str_replace, "to 1", "")

# print(gradRate_teacherData)
```

```r
gradRate_teacherData_graph <-
  ggplot(gradRate_teacherData, aes(x = experienced_teacher_percent,
                                   y = percent_graduated )) + geom_point() +
  theme_minimal() +
  ggtitle("Graduate rate percentage vs.
          Experienced teacher percentage") + geom_smooth(method=lm) +
  ylim(75, 100)

gradRate_teacherData_graph
```



Observation: there is a correlation; the more experienced teacher, the higher graduation rate.

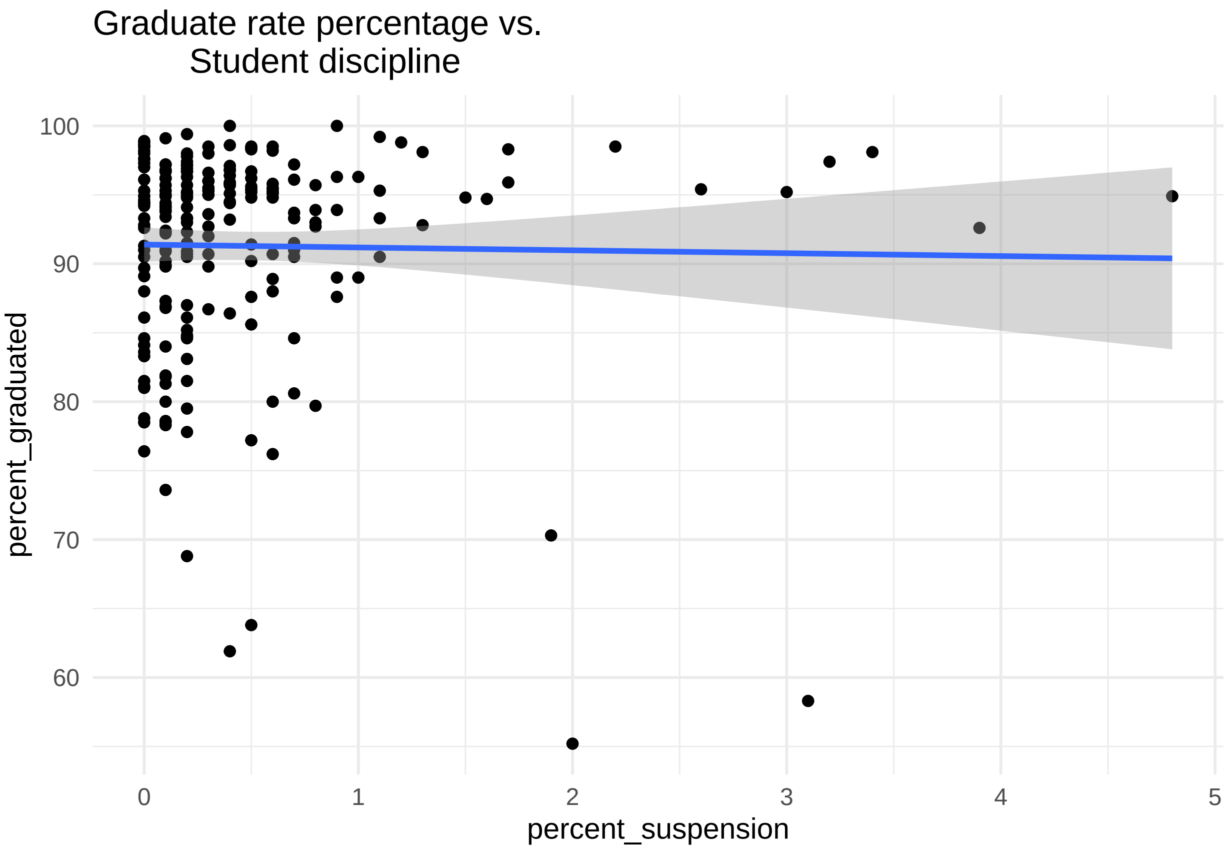
```
StudentTeacherRatio_graph <-
  ggplot(gradRate_teacherData, aes(x = student_teacher_ratio,
                                   y = percent_graduated )) + geom_point() +
  theme_minimal() +
  ggtitle("Graduate rate percentage vs.
          Student teacher ratio") + geom_smooth(method=lm)

StudentTeacherRatio_graph
```



## 3. Student Discipline vs. Graduation Rate

```
gradRate_discipline <- left_join(gradRate, discipline, by="District Code")
gradRate_discipline <- na.omit(gradRate_discipline)

gradRate_discipline <- gradRate_discipline %>%
  rename(
    'percent_graduated' = '% Graduated')

gradRate_discipline <- gradRate_discipline %>%
  rename(
    'percent_suspension' = '% In-School Suspension')

# print(gradRate_discipline)
```

```
gradRate_discipline_graph <-
  ggplot(gradRate_discipline, aes(x = percent_suspension,
                                  y = percent_graduated )) + geom_point() +
  theme_minimal() +
  ggtitle("Graduate rate percentage vs.
          Student discipline") + geom_smooth(method=lm)

gradRate_discipline_graph
```

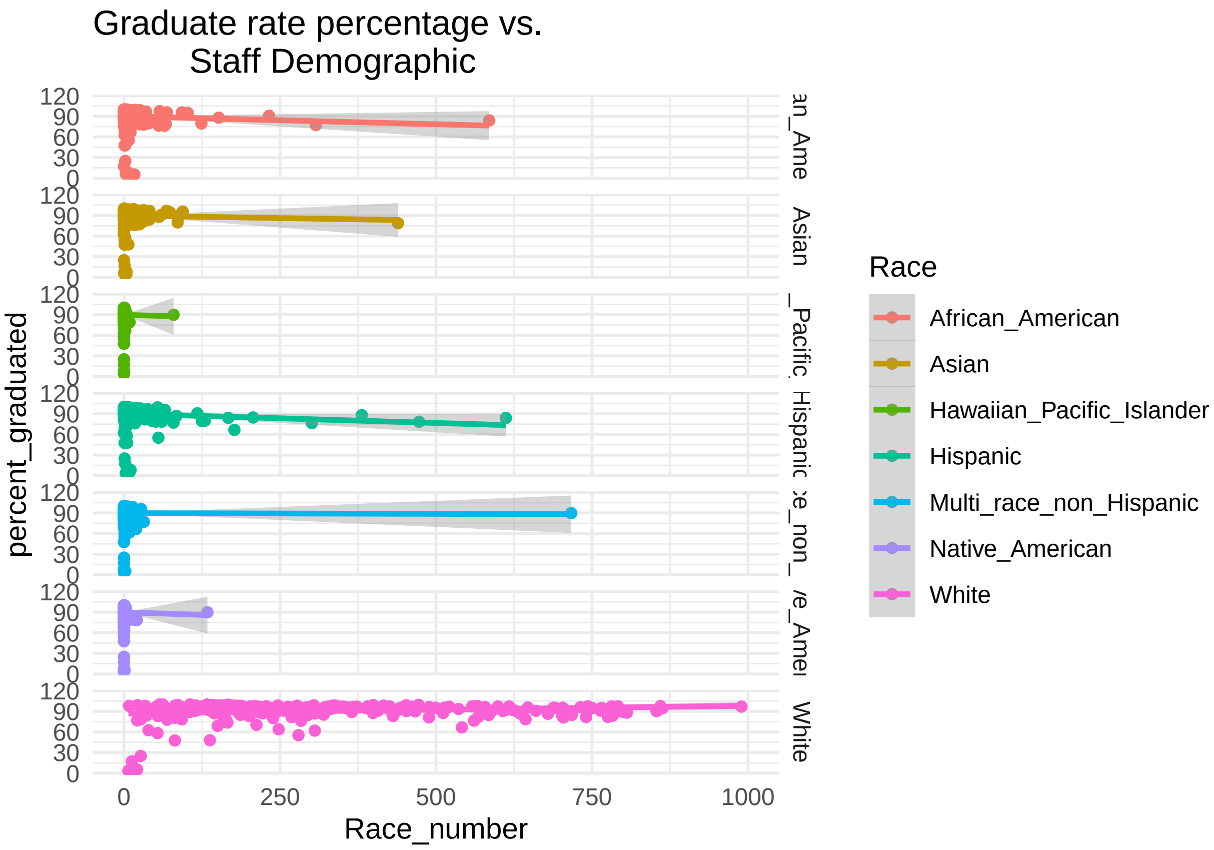


## 4. Demographic vs. Graduation Rate

```
gradRate_demographic <- left_join(gradRate, eduGen, by="District Code")
gradRate_demographic <- na.omit(gradRate_demographic)

gradRate_demographic <- gradRate_demographic %>%
  rename(
    'percent_graduated' = '% Graduated')

gradRate_demographic <- gradRate_demographic %>%
  rename(
    'African_American' = 'African American (#)')
gradRate_demographic <- gradRate_demographic %>%
  rename(
    'White' = 'White (#)')

gradRate_demographic <- gradRate_demographic %>%
  rename(
    'Asian' = 'Asian (#)')

gradRate_demographic <- gradRate_demographic %>%
  rename(
    'Hispanic' = 'Hispanic (#)')

gradRate_demographic <- gradRate_demographic %>%
  rename(
    'Native_American' = 'Native American (#)')

gradRate_demographic <- gradRate_demographic %>%
  rename(
    'Hawaiian_Pacific_Islander' = 'Native Hawaiian, Pacific Islander (#)')

gradRate_demographic <- gradRate_demographic %>%
  rename(
    'Multi_race_non_Hispanic' = 'Multi-Race,Non-Hispanic (#)')

gradRate_demographic_long <- pivot_longer(gradRate_demographic, cols=11:17,
                                          names_to = "Race",
                                          values_to = "Race_number")

print(gradRate_demographic_long)
```

```
## # A tibble: 2,135 x 15
##    `District Name`                `District Code` `# in Cohort` percent_graduat~
##    <chr>                          <chr>                   <dbl>            <dbl>
##  1 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  2 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  3 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  4 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  5 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  6 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  7 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  8 Abington                       00010000                  163             93.3
##  9 Abington                       00010000                  163             93.3
## 10 Abington                       00010000                  163             93.3
## # ... with 2,125 more rows, and 11 more variables: % Still in School <dbl>,
## #   % 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>,
## #   Race_number <dbl>
```

```
gradRate_demographic_graph <-
  ggplot(gradRate_demographic_long, aes(x = Race_number,
                                  y = percent_graduated, color = Race)) +
  geom_point() +
  theme_minimal() +
  ggtitle("Graduate rate percentage vs.
          Staff Demographic") + xlim(0, 1000)+ facet_grid(Race ~.) +
  geom_smooth(method=lm)

gradRate_demographic_graph
```



## 5. Staffing Retention vs. Graduation Rate

```
gradRate_staffingRetention <- left_join(gradRate, staffReten,
                                        by="District Code")
gradRate_staffingRetention <- na.omit(gradRate_staffingRetention)

print(gradRate_staffingRetention)
```

```
## # A tibble: 306 x 19
##    `District Name.~ `District Code` `# in Cohort` `% Graduated` `% Still in Sch~
##    <chr>            <chr>                   <dbl>         <dbl>            <dbl>
##  1 Abby Kelley Fos~ 04450000                   82          98.8              0
##  2 Abington         00010000                  163          93.3              2.5
##  3 Academy Of the ~ 04120000                   59          93.2              6.8
##  4 Acton-Boxborough 06000000                  439          97.3              2.1
##  5 Advanced Math a~ 04300000                  141          98.6              1.4
##  6 Agawam           00050000                  286          90.9              2.1
##  7 Amesbury         00070000                  161          90.1              5.6
##  8 Amherst-Pelham   06050000                  232          91.8              5.2
##  9 Andover          00090000                  460          97.6              1.7
## 10 Argosy Collegia~ 35090000                   60          58.3             25
## # ... with 296 more rows, and 14 more variables: % Non-Grad Completers <dbl>,
## #   % H.S. Equiv. <dbl>, % Dropped Out <dbl>, % Permanently Excluded <dbl>,
## #   District Name.y <chr>, Superintendent Total <dbl>,
## #   Superintendent # Retained <dbl>, Superintendent % Retained <dbl>,
## #   Principal Total <dbl>, Principal # Retained <dbl>,
## #   Principal % Retained <dbl>, Teacher Total <dbl>, Teacher # Retained <dbl>,
## #   Teacher % Retained <dbl>
```

## 6. Graduation Rate vs. Day missed

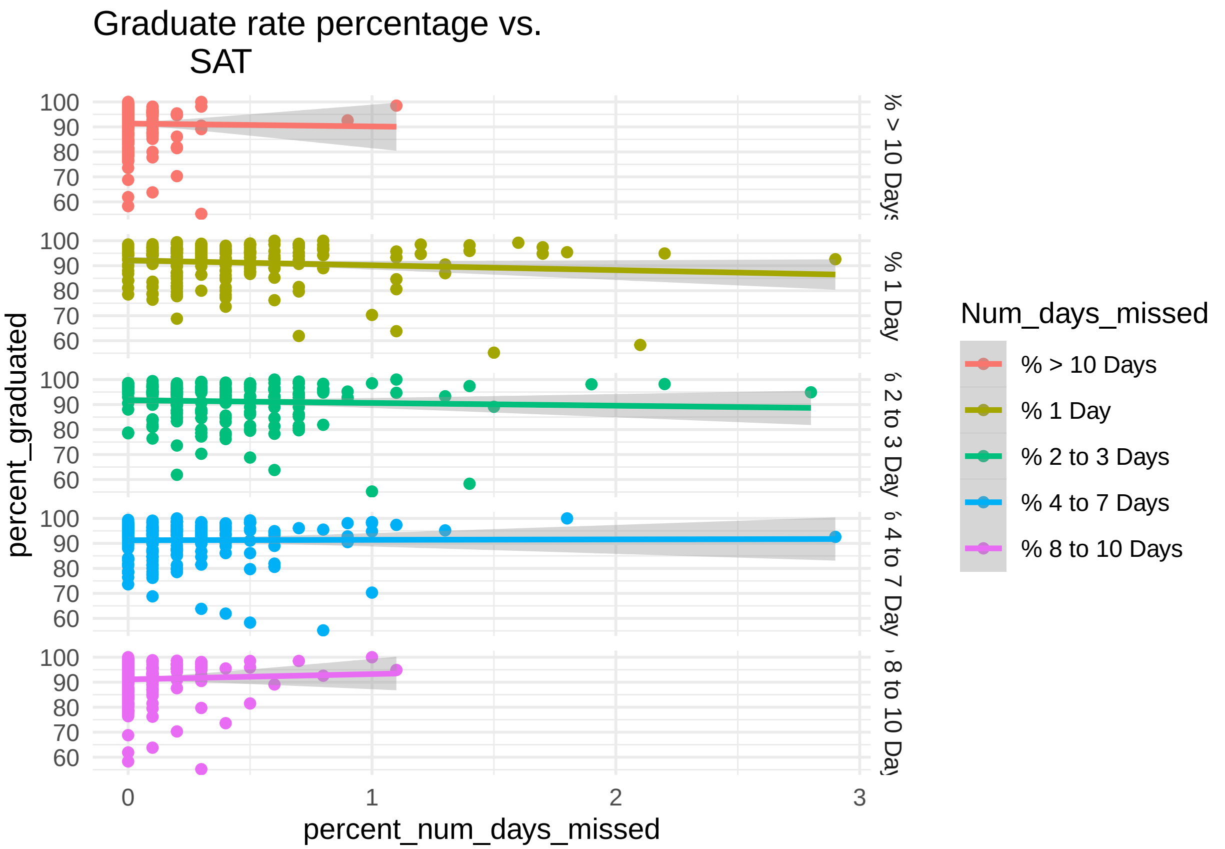
```
daysMissed$District_code <- str_pad(daysMissed$`District Code`, 8, pad = "0")

# gradRateRename <- gradRate %>%
#   rename(
#     'District_code' = 'District Code')

print(daysMissed)
```

```
## # A tibble: 401 x 10
##    `District Name`          `District Code` Students `Students Discip~ `% 1 Day`
##    <chr>                    <chr>              <dbl>             <dbl>     <dbl>
##  1 Abby Kelley Foster Char~ 04450000            1437                 2      NA
##  2 Abington                 00010000            2214                41       0.5
##  3 Academy Of the Pacific ~ 04120000             544                 6       0.6
##  4 Acton-Boxborough         06000000            5320                 8       0.1
##  5 Acushnet                 00030000             940                 6       0.4
##  6 Advanced Math and Scien~ 04300000             974                13       0.5
##  7 Agawam                   00050000            3624                14       0.1
##  8 Alma del Mar Charter Sc~ 04090000             808                17       1.4
##  9 Amesbury                 00070000            2009                 0      NA
## 10 Amherst                  00080000            1103                 0      NA
## # ... with 391 more rows, and 5 more variables: % 2 to 3 Days <dbl>,
## #   % 4 to 7 Days <dbl>, % 8 to 10 Days <dbl>, % > 10 Days <dbl>,
## #   District_code <chr>
```

```
gradRate_daysMissed <- left_join(gradRate, daysMissed, by="District Code")
gradRate_daysMissed <- na.omit(gradRate_daysMissed)

gradRate_daysMissed <- gradRate_daysMissed %>%
  rename(
    'percent_graduated' = '% Graduated')

gradRate_daysMissed_long <- pivot_longer(gradRate_daysMissed, cols=13:17,
                                         names_to = "Num_days_missed",
                                         values_to = "percent_num_days_missed")

print(gradRate_daysMissed_long)
```

```
## # A tibble: 1,065 x 15
##    `District Name.x`              `District Code` `# in Cohort` percent_graduat~
##    <chr>                          <chr>                   <dbl>            <dbl>
##  1 Abington                       00010000                  163             93.3
##  2 Abington                       00010000                  163             93.3
##  3 Abington                       00010000                  163             93.3
##  4 Abington                       00010000                  163             93.3
##  5 Abington                       00010000                  163             93.3
##  6 Academy Of the Pacific Rim Ch~ 04120000                   59             93.2
##  7 Academy Of the Pacific Rim Ch~ 04120000                   59             93.2
##  8 Academy Of the Pacific Rim Ch~ 04120000                   59             93.2
##  9 Academy Of the Pacific Rim Ch~ 04120000                   59             93.2
## 10 Academy Of the Pacific Rim Ch~ 04120000                   59             93.2
## # ... with 1,055 more rows, and 11 more variables: % Still in School <dbl>,
## #   % Non-Grad Completers <dbl>, % H.S. Equiv. <dbl>, % Dropped Out <dbl>,
## #   % Permanently Excluded <dbl>, District Name.y <chr>, Students <dbl>,
## #   Students Disciplined <dbl>, District_code <chr>, Num_days_missed <chr>,
## #   percent_num_days_missed <dbl>
```

```
gradRate_daysMissed_graph <-
  ggplot(gradRate_daysMissed_long, aes(x = percent_num_days_missed,
                                 y = percent_graduated,
                                 color = Num_days_missed)) + geom_point() +
  theme_minimal() +
  ggtitle("Graduate rate percentage vs.
          SAT") + facet_grid(Num_days_missed ~.) + geom_smooth(method=lm)
gradRate_daysMissed_graph
```



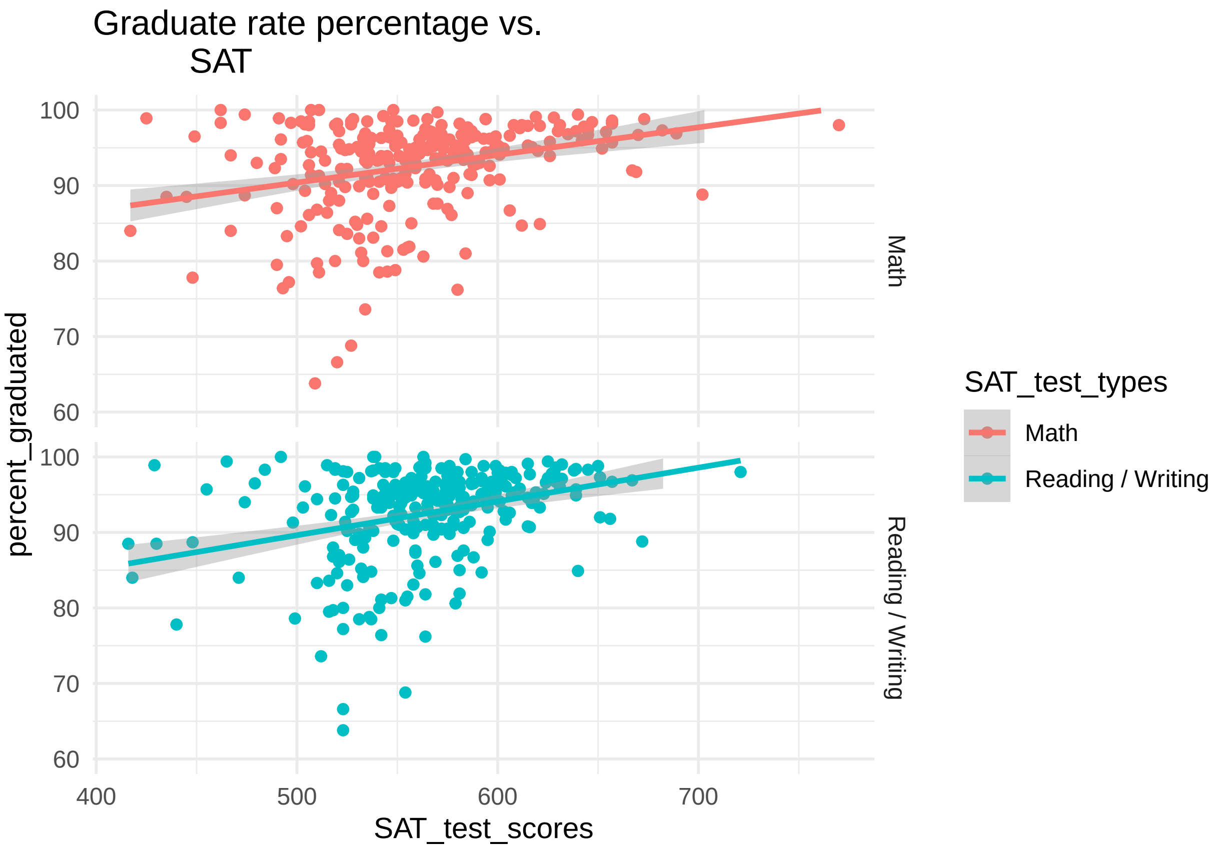
## 7. SAT vs. Graduation Rate

```
gradRate_SAT <- left_join(gradRate, sat, by="District Code")


gradRate_SAT <- gradRate_SAT %>%
  rename(
    'percent_graduated' = '% Graduated')
```

```
gradRate_SAT <- gradRate_SAT %>%
  rename(
    'tests_taken' = 'Tests Taken')

gradRate_SAT_long <- pivot_longer(gradRate_SAT, cols=12:14,
                                  names_to = "SAT_test_types",
                                  values_to = "SAT_test_scores")

gradRate_SAT_long <- gradRate_SAT_long[!(is.na(gradRate_SAT_long$SAT_test_scores)), ]

print(gradRate_SAT_long)
```

```
## # A tibble: 542 x 13
##    `District Name.x`              `District Code` `# in Cohort` percent_graduat~
##    <chr>                          <chr>                   <dbl>            <dbl>
##  1 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  2 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  3 Abington                       00010000                  163             93.3
##  4 Abington                       00010000                  163             93.3
##  5 Acton-Boxborough               06000000                  439             97.3
##  6 Acton-Boxborough               06000000                  439             97.3
##  7 Advanced Math and Science Aca~ 04300000                  141             98.6
##  8 Advanced Math and Science Aca~ 04300000                  141             98.6
##  9 Agawam                         00050000                  286             90.9
## 10 Agawam                         00050000                  286             90.9
## # ... with 532 more rows, and 9 more variables: % Still in School <dbl>,
## #   % Non-Grad Completers <dbl>, % H.S. Equiv. <dbl>, % Dropped Out <dbl>,
## #   % Permanently Excluded <dbl>, District Name.y <chr>, tests_taken <dbl>,
## #   SAT_test_types <chr>, SAT_test_scores <dbl>
```

```
gradRate_SAT_graph <-
  ggplot(gradRate_SAT_long, aes(x = SAT_test_scores,
                                 y = percent_graduated,
                                color = SAT_test_types)) + geom_point() +
  theme_minimal() +
  ggtitle("Graduate rate percentage vs.
          SAT") + facet_grid(SAT_test_types ~.) + geom_smooth(method=lm) + ylim(60,100)
gradRate_SAT_graph
```



## 8. Graduation plan Vs. Graduation rate

```
# gradRate_gradplan <- left_join(gradRate, plansforHSgrad, by="District Code")
#
# gradRate_gradplan <- gradRate_gradplan %>%
#   rename(
#     'percent_graduated' = '% Graduated')
#
# gradRate_gradplan_long <- pivot_longer(gradRate_gradplan, cols=11:20, names_to = "Plan_type", values_
#
# print(gradRate_gradplan_long)
# use cmd/ctrl + shift + c to uncomment
```

```
# gradRate_gradPlan_graph <-
#   ggplot(gradRate_gradplan_long, aes(x = Plan_percentage,
#                                  y = percent_graduated, color = Plan_type)) + geom_point() +
#   theme_minimal() +
#   ggtitle("Graduate rate percentage vs.
#           Plan after high school") + facet_grid(Plan_type ~.) + geom_smooth(method=lm)
# gradRate_gradPlan_graph
```

## 9. Graduation rate vs. Students Background

```
print(selectPop)
```

```
## # A tibble: 401 x 18
##    `District Name`        `District Code` `First Language No~ `First Language N~
##    <chr>                  <chr>                         <dbl>              <dbl>
##  1 Abby Kelley Foster Ch~ 04450000                        952               66.8
##  2 Abington               00010000                        298               14.1
##  3 Academy Of the Pacifi~ 04120000                        173               32
##  4 Acton-Boxborough       06000000                       1117               21.5
##  5 Acushnet               00030000                         10                1.1
##  6 Advanced Math and Sci~ 04300000                        238               24.6
##  7 Agawam                 00050000                        438               12.5
##  8 Alma del Mar Charter ~ 04090000                        349               43.8
##  9 Amesbury               00070000                         74                4
## 10 Amherst                00080000                        300               29.2
## # ... with 391 more rows, and 14 more variables:
## #   English Language Learner # <dbl>, English Language Learner % <dbl>,
## #   Students With Disabilities # <dbl>, Students With Disabilities % <dbl>,
## #   Low Income # <lgl>, Low Income % <lgl>, Free Lunch # <lgl>,
## #   Free Lunch % <lgl>, Reduced Lunch # <lgl>, Reduced Lunch % <lgl>,
## #   High Needs #...15 <dbl>, High Needs #...16 <dbl>,
## #   Economically Disadvantaged # <dbl>, Economically Disadvantaged % <dbl>
```

```r
# gradRateRename1 <- gradRate %>%
#   rename(
#     'District_code' = 'District Code')

selectPop$District_code <- str_pad(selectPop$`District Code`, 8, pad = "0")

print(selectPop)
```

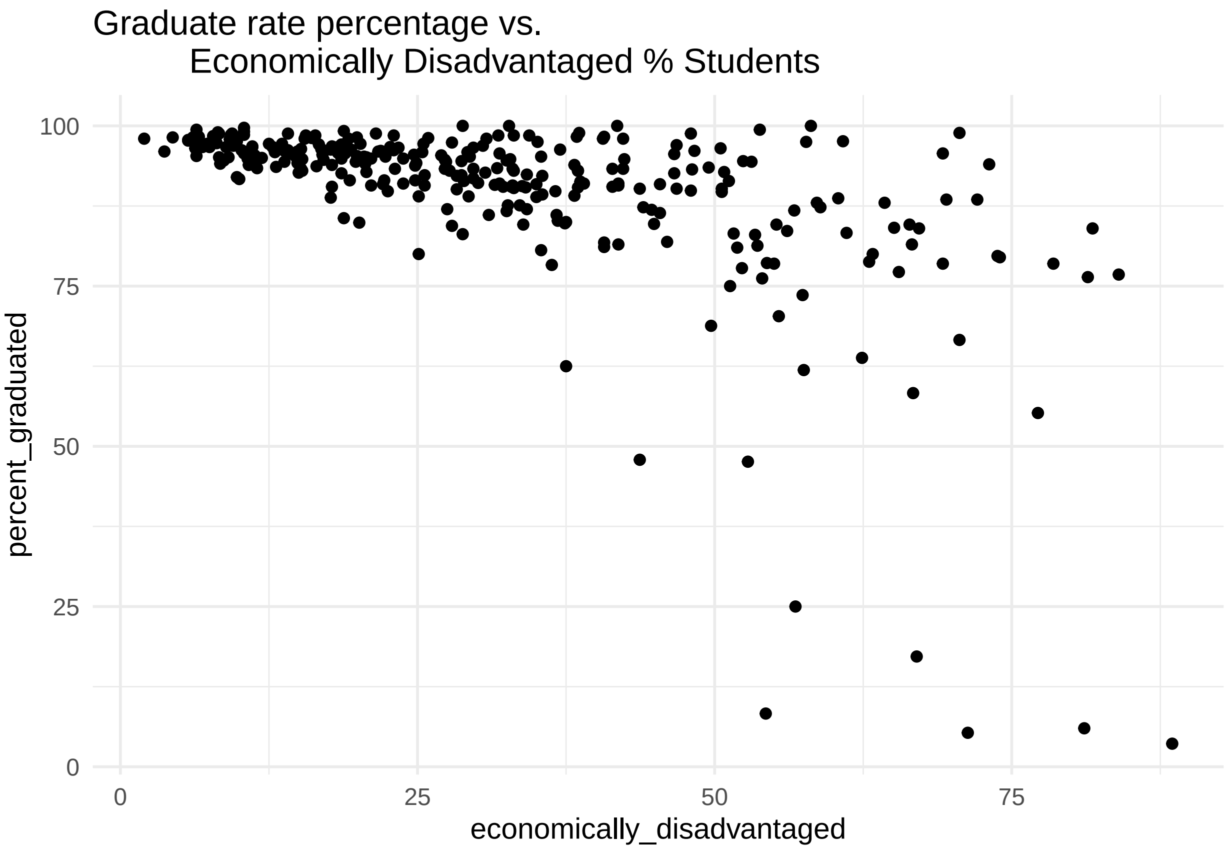
```
## # A tibble: 401 x 19
##    `District Name`        `District Code` `First Language No~ `First Language N~
##    <chr>                  <chr>                         <dbl>              <dbl>
##  1 Abby Kelley Foster Ch~ 04450000                        952               66.8
##  2 Abington               00010000                        298               14.1
##  3 Academy Of the Pacifi~ 04120000                        173               32
##  4 Acton-Boxborough       06000000                       1117               21.5
##  5 Acushnet               00030000                         10                1.1
##  6 Advanced Math and Sci~ 04300000                        238               24.6
##  7 Agawam                 00050000                        438               12.5
##  8 Alma del Mar Charter ~ 04090000                        349               43.8
##  9 Amesbury               00070000                         74                4
## 10 Amherst                00080000                        300               29.2
## # ... with 391 more rows, and 15 more variables:
## #   English Language Learner # <dbl>, English Language Learner % <dbl>,
## #   Students With Disabilities # <dbl>, Students With Disabilities % <dbl>,
## #   Low Income # <lgl>, Low Income % <lgl>, Free Lunch # <lgl>,
## #   Free Lunch % <lgl>, Reduced Lunch # <lgl>, Reduced Lunch % <lgl>,
## #   High Needs #...15 <dbl>, High Needs #...16 <dbl>,
## #   Economically Disadvantaged # <dbl>, Economically Disadvantaged % <dbl>, ...
```

```r
gradRate_selectPop <- left_join(gradRate, selectPop, by="District Code")

gradRate_selectPop <- gradRate_selectPop %>%
  rename(
    'percent_graduated' = '% Graduated')
```

```r
gradRate_selectPop <- gradRate_selectPop %>%
  rename(
    'economically_disadvantaged' = 'Economically Disadvantaged %')


print(gradRate_selectPop)
```

```
## # A tibble: 305 x 27
##    `District Name.x`               `District Code` `# in Cohort` percent_graduat~
##    <chr>                          <chr>                   <dbl>            <dbl>
##  1 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  2 Abington                       00010000                  163             93.3
##  3 Academy Of the Pacific Rim Ch~ 04120000                   59             93.2
##  4 Acton-Boxborough               06000000                  439             97.3
##  5 Advanced Math and Science Aca~ 04300000                  141             98.6
##  6 Agawam                         00050000                  286             90.9
##  7 Amesbury                       00070000                  161             90.1
##  8 Amherst-Pelham                 06050000                  232             91.8
##  9 Andover                        00090000                  460             97.6
## 10 Argosy Collegiate Charter Sch~ 35090000                   60             58.3
## # ... with 295 more rows, and 23 more variables: % Still in School <dbl>,
## #   % Non-Grad Completers <dbl>, % H.S. Equiv. <dbl>, % Dropped Out <dbl>,
## #   % Permanently Excluded <dbl>, District Name.y <chr>,
## #   First Language Not English # <dbl>, First Language Not English % <dbl>,
## #   English Language Learner # <dbl>, English Language Learner % <dbl>,
## #   Students With Disabilities # <dbl>, Students With Disabilities % <dbl>,
## #   Low Income # <lgl>, Low Income % <lgl>, Free Lunch # <lgl>, ...
```

```r
gradRate_selectPop_graph <-
  ggplot(gradRate_selectPop, aes(x = economically_disadvantaged,
                                 y = percent_graduated)) + geom_point() +
  theme_minimal() +
  ggtitle("Graduate rate percentage vs.
          Economically Disadvantaged % Students")
gradRate_selectPop_graph
```



```
print(mobilityRate)
```

```
## # A tibble: 400 x 7
##    `District Name`         `District Code` `Churn/Intake E~ `% Churn` `% Intake`
##    <chr>                   <chr>                      <dbl>     <dbl>      <dbl>
##  1 Abby Kelley Foster Cha~ 04450000                    1437       3.2        2.2
##  2 Abington                00010000                    2215       8.1        4.7
##  3 Academy Of the Pacific~ 04120000                     544       4.2        2.9
##  4 Acton-Boxborough        06000000                    5322       3.7        2.3
##  5 Acushnet                00030000                     942       6.8        3.8
##  6 Advanced Math and Scie~ 04300000                     974       2.8        1.3
##  7 Agawam                  00050000                    3626       9.8        5.2
##  8 Alma del Mar Charter S~ 04090000                     809       3.5        1.5
##  9 Amesbury                00070000                    2010      13.7        7.6
## 10 Amherst                 00080000                    1104      15.8        8.6
## # ... with 390 more rows, and 2 more variables: Stability Enroll <dbl>,
## #   % Stability <dbl>
```

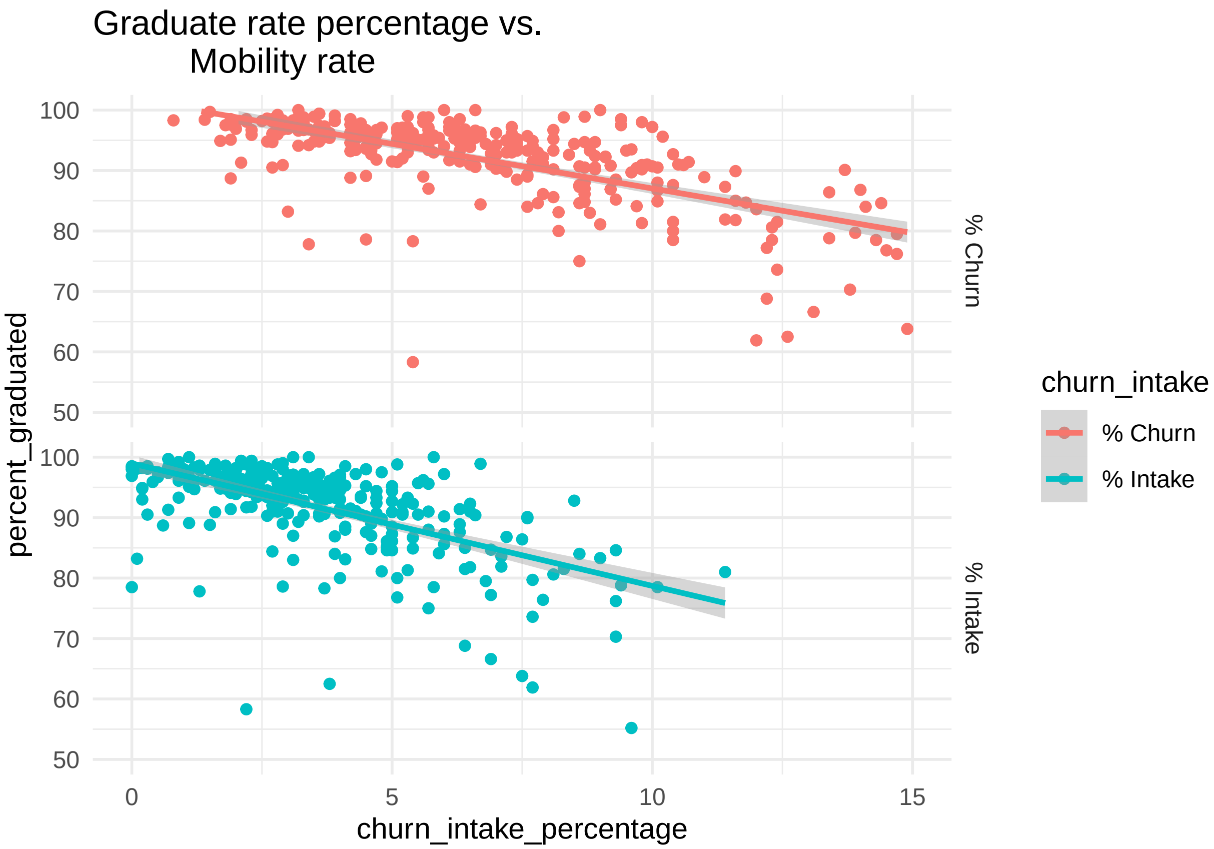
```
# gradRateRename2 <- gradRate %>%
#   rename(
#     'District_code' = 'District Code')

mobilityRate$District_code <- str_pad(mobilityRate$`District Code`, 8, pad = "0")

print(mobilityRate)
```

```
## # A tibble: 400 x 8
##    `District Name`         `District Code` `Churn/Intake E~ `% Churn` `% Intake`
##    <chr>                   <chr>                      <dbl>     <dbl>      <dbl>
##  1 Abby Kelley Foster Cha~ 04450000                    1437       3.2        2.2
##  2 Abington                00010000                    2215       8.1        4.7
##  3 Academy Of the Pacific~ 04120000                     544       4.2        2.9
##  4 Acton-Boxborough        06000000                    5322       3.7        2.3
##  5 Acushnet                00030000                     942       6.8        3.8
##  6 Advanced Math and Scie~ 04300000                     974       2.8        1.3
##  7 Agawam                  00050000                    3626       9.8        5.2
##  8 Alma del Mar Charter S~ 04090000                     809       3.5        1.5
##  9 Amesbury                00070000                    2010      13.7        7.6
## 10 Amherst                 00080000                    1104      15.8        8.6
## # ... with 390 more rows, and 3 more variables: Stability Enroll <dbl>,
## #   % Stability <dbl>, District_code <chr>
```

```
gradRate_mobilityRate <- left_join(gradRate, mobilityRate, by="District Code")

gradRate_mobilityRate <- gradRate_mobilityRate %>%
  rename(
    'percent_graduated' = '% Graduated')

gradRate_mobilityRate_long <- pivot_longer(gradRate_mobilityRate, cols=12:13,
                                           names_to = "churn_intake",
                                          values_to = "churn_intake_percentage")

print(gradRate_mobilityRate_long)
```

```
## # A tibble: 610 x 16
##    `District Name.x`              `District Code` `# in Cohort` percent_graduat~
##    <chr>                          <chr>                   <dbl>            <dbl>
##  1 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  2 Abby Kelley Foster Charter Pu~ 04450000                   82             98.8
##  3 Abington                       00010000                  163             93.3
##  4 Abington                       00010000                  163             93.3
##  5 Academy Of the Pacific Rim Ch~ 04120000                   59             93.2
##  6 Academy Of the Pacific Rim Ch~ 04120000                   59             93.2
##  7 Acton-Boxborough               06000000                  439             97.3
##  8 Acton-Boxborough               06000000                  439             97.3
##  9 Advanced Math and Science Aca~ 04300000                  141             98.6
## 10 Advanced Math and Science Aca~ 04300000                  141             98.6
## # ... with 600 more rows, and 12 more variables: % Still in School <dbl>,
## #   % Non-Grad Completers <dbl>, % H.S. Equiv. <dbl>, % Dropped Out <dbl>,
## #   % Permanently Excluded <dbl>, District Name.y <chr>,
## #   Churn/Intake Enroll <dbl>, Stability Enroll <dbl>, % Stability <dbl>,
## #   District_code <chr>, churn_intake <chr>, churn_intake_percentage <dbl>
```

```
gradRate_mobility_graph <-
  ggplot(gradRate_mobilityRate_long, aes(x = churn_intake_percentage,
                                 y = percent_graduated,
                                 color = churn_intake)) + geom_point() +
  theme_minimal() +
  ggtitle("Graduate rate percentage vs.
          Mobility rate") + facet_grid(churn_intake ~.) + xlim(0, 15) +
  ylim(50, 100) +  geom_smooth(method=lm)
gradRate_mobility_graph
```



### Tran’s work end here

### Cleaning Data

```
sat <- sat %>% mutate(`Total Score` = `Reading / Writing` + Math) %>%
  select(!Writing)

enrollByGrade <- enrollByGrade %>%
  mutate(`HS Enrollment` = `9` + `10` + `11` + `12`) %>%
  select(`District Code`, `HS Enrollment`, Total) %>%
  rename(Enrollment = Total)

ap_part <- ap_part %>% select(`District Code`, `Tests Takers`)

staffReten <- staffReten %>% select(`District Code`, `Teacher % Retained`) %>%
  rename(`Teacher Retention Rate` = `Teacher % Retained`)

classSize <- inner_join(classSizeByClass,classSizeByRace) %>%
  select(!c(`Number of Students`,`Total # of Classes`,
                      `District Name`,`English Language Learner %`,
            `Students with Disabilities %`,`Economically Disadvantaged %`
            ))

college <- college %>%
  rename(`Percent Going to College`=`Attending Coll./Univ. (%)`) %>%
  select(`District Code`, `Percent Going to College`)
```

```r
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`)

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

gradRate <- gradRate %>% select(`District Code`, `% Graduated`, `% Dropped Out`)

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`)

convertPerc <- function(x, na.rm = TRUE) format(round((x / eduGen$`FTE Count`)
                                                      * 100, 3), nsmall = 3)
eduGen <- eduGen %>% mutate_at(c("Females (#)", "African American (#)",
                                 "Asian (#)","Hispanic (#)","White (#)",
                                 "Native American (#)",
                                 "Native Hawaiian, Pacific Islander (#)",
                                 "Multi-Race,Non-Hispanic (#)", "Males (#)"),
                               convertPerc) %>%
  rename(`% Female Teachers`="Females (#)",`% African American Teachers`=
           "African American (#)",`% Asian Teachers`="Asian (#)",
         `% Hispanic Teachers`="Hispanic (#)",`% White Teachers`="White (#)",
         `% Native American Teachers`="Native American (#)",
         `% Native Hawaiian, Pacific Islander Teachers`=
           "Native Hawaiian, Pacific Islander (#)",
         `% Multi-Race,Non-Hispanic Teachers`="Multi-Race,Non-Hispanic (#)",
         `% Male Teachers`="Males (#)") %>%
  select(!c(`District/School Name`,`FTE Count`))

mobilityRate <- mobilityRate %>% select(!c(`District Name`,
                                           `Churn/Intake Enroll`,
                                           `Stability Enroll`))

teachData <- teachData %>% select(!`District Name`)
teachData$`Student / Teacher Ratio` <- substr(
  teachData$`Student / Teacher Ratio`,1,
  nchar(teachData$`Student / Teacher Ratio`)-5) %>% parse_number()
```

```r
selectPop <- selectPop %>% rename(`High Needs %`=`High Needs #...16`) %>%
  select(`District Code`,`First Language Not English %`,
         `English Language Learner %`,`Students With Disabilities %`,
         `High Needs %`,`Economically Disadvantaged %`)
```

### Joining all tables

```r
eduData <- inner_join(sat, enrollByGrade, by = "District Code") %>%
  inner_join(ap_part, by = "District Code") %>%
  inner_join(staffReten, 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(dropOut, by = "District Code") %>%
  inner_join(mobilityRate, 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) %>% drop_na()
```

```r
summary(eduData)
```

```
##  District Name      District Code       Tests Taken     Reading / Writing
##  Length:264         Length:264         Min.   :  10.0   Min.   :416.0
##  Class :character   Class :character   1st Qu.:  50.0   1st Qu.:538.8
##  Mode  :character   Mode  :character   Median : 100.0   Median :564.5
##                                        Mean   : 162.3   Mean   :564.8
##                                        3rd Qu.: 208.8   3rd Qu.:588.0
##                                        Max.   :2299.0   Max.   :721.0
##       Math        Total Score    HS Enrollment       Enrollment
##  Min.   :417.0   Min.   : 835   Min.   :   98.0   Min.   :   98
##  1st Qu.:527.0   1st Qu.:1064   1st Qu.:  447.8   1st Qu.: 1228
##  Median :555.0   Median :1124   Median :  765.0   Median : 2140
##  Mean   :558.4   Mean   :1123   Mean   : 1039.4   Mean   : 3160
##  3rd Qu.:584.0   3rd Qu.:1171   3rd Qu.: 1261.5   3rd Qu.: 3709
##  Max.   :770.0   Max.   :1491   Max.   :14342.0   Max.   :48112
##   Tests Takers     Teacher Retention Rate Average Class Size    Female %
##  Min.   :   1.00   Min.   : 55.60         Min.   : 8.20        Min.   :32.70
##  1st Qu.:  73.75   1st Qu.: 87.20         1st Qu.:13.78        1st Qu.:47.70
##  Median : 147.00   Median : 89.50         Median :16.00        Median :48.80
##  Mean   : 192.44   Mean   : 88.30         Mean   :15.77        Mean   :48.77
##  3rd Qu.: 248.50   3rd Qu.: 91.53         3rd Qu.:17.52        3rd Qu.:49.80
##  Max.   :3161.00   Max.   :100.00         Max.   :45.80        Max.   :74.10
##      Male %      African American %    Asian %         Hispanic %
##  Min.   :25.70   Min.   : 0.000     Min.   : 0.000   Min.   : 0.00
##  1st Qu.:50.10   1st Qu.: 1.600     1st Qu.: 1.200   1st Qu.: 5.30
##  Median :51.20   Median : 3.000     Median : 2.300   Median : 7.90  
##  Mean   :51.13   Mean   : 6.987     Mean   : 5.502   Mean   :14.59  
##  3rd Qu.:52.20   3rd Qu.: 6.350     3rd Qu.: 6.200   3rd Qu.:16.43  
##  Max.   :67.40   Max.   :77.600     Max.   :63.600   Max.   :93.80  
##     White %      Native American % Native Hawaiian, Pacific Islander %
##  Min.   : 0.30   Min.   :0.0000    Min.   :0.00000                    
##  1st Qu.:60.38   1st Qu.:0.1000    1st Qu.:0.00000                    
##  Median :77.40   Median :0.1000    Median :0.10000                    
##  Mean   :68.69   Mean   :0.2481    Mean   :0.09545                    
##  3rd Qu.:85.80   3rd Qu.:0.3000    3rd Qu.:0.10000                    
##  Max.   :96.50   Max.   :5.6000    Max.   :2.40000                    
##  Multi-Race, Non-Hispanic % Percent Going to College Attendance Rate
##  Min.   : 0.400             Min.   :15.20            Min.   :79.90  
##  1st Qu.: 2.600             1st Qu.:58.70            1st Qu.:92.60  
##  Median : 3.700             Median :70.90            Median :94.65  
##  Mean   : 3.927             Mean   :68.09            Mean   :94.18  
##  3rd Qu.: 4.825             3rd Qu.:79.92            3rd Qu.:96.10  
##  Max.   :11.000             Max.   :91.70            Max.   :99.60  
##  Average # of Absences   Attrition      % Students Completing Advanced
##  Min.   : 0.60         Min.   : 1.400   Min.   : 17.00                
##  1st Qu.: 6.50         1st Qu.: 5.000   1st Qu.: 58.90                
##  Median : 8.90         Median : 6.750   Median : 68.05                
##  Mean   : 9.65         Mean   : 7.078   Mean   : 68.15                
##  3rd Qu.:12.15         3rd Qu.: 8.500   3rd Qu.: 78.20                
##  Max.   :33.60         Max.   :31.000   Max.   :100.00                
##  Adv Course % Math Adv Course % ELA  % Graduated     % Dropped Out   
##  Min.   :  8.10    Min.   : 0.00    Min.   : 47.90   Min.   : 0.000  
##  1st Qu.: 45.58    1st Qu.:10.18    1st Qu.: 90.47   1st Qu.: 0.975  
##  Median : 57.10    Median :15.95    Median : 94.40   Median : 2.100  
##  Mean   : 57.47    Mean   :18.73    Mean   : 92.31   Mean   : 3.477  
##  3rd Qu.: 68.45    3rd Qu.:24.73    3rd Qu.: 96.70   3rd Qu.: 4.900  
##  Max.   :100.00    Max.   :94.30    Max.   :100.00   Max.   :35.000  
##  % in an Art Course % of Teachers <40 % Disciplined   
##  Min.   : 0.00      Min.   :11.11     Min.   :0.0000  
##  1st Qu.:67.73      1st Qu.:32.94     1st Qu.:0.2416  
##  Median :81.82      Median :37.42     Median :0.5889  
##  Mean   :72.48      Mean   :39.82     Mean   :0.8471  
##  3rd Qu.:86.88      3rd Qu.:42.93     3rd Qu.:1.2195  
##  Max.   :99.50      Max.   :94.24     Max.   :6.8006  
##  % African American Teachers % Asian Teachers   % Hispanic Teachers
##  Length:264                  Length:264         Length:264         
##  Class :character            Class :character   Class :character   
##  Mode  :character            Mode  :character   Mode  :character   
##                                                                    
##                                                                    
##                                                                    
##  % White Teachers   % Native American Teachers
##  Length:264         Length:264                
##  Class :character   Class :character          
##  Mode  :character   Mode  :character          
##                                               
##                                               
##                                               
##  % Native Hawaiian, Pacific Islander Teachers
##  Length:264
##  Class :character
##  Mode  :character
##
##
##
##  % Multi-Race,Non-Hispanic Teachers % Female Teachers  % Male Teachers
##  Length:264                         Length:264         Length:264
##  Class :character                   Class :character   Class :character
##  Mode  :character                   Mode  :character   Mode  :character
##
##
##
##  Total # of Teachers (FTE) % of Teachers Licensed Student / Teacher Ratio
##  Min.   :   6.0            Min.   : 56.60         Min.   : 7.70
##  1st Qu.: 102.0            1st Qu.: 98.50         1st Qu.:11.07
##  Median : 168.4            Median : 99.40         Median :12.05
##  Mean   : 259.5            Mean   : 97.04         Mean   :12.10
##  3rd Qu.: 299.4            3rd Qu.:100.00         3rd Qu.:13.00
##  Max.   :4595.5            Max.   :100.00         Max.   :28.90
##  Percent of Experienced Teachers
##  Min.   :33.30
##  1st Qu.:82.60
##  Median :86.70
##  Mean   :83.97
##  3rd Qu.:89.90
##  Max.   :96.00
##  Percent of Teachers without Waiver or Provisional License
##  Min.   : 69.10
##  1st Qu.: 91.88
##  Median : 94.70
##  Mean   : 93.02
##  3rd Qu.: 96.50
##  Max.   :100.00
##  Percent Teaching In-Field % Dropout All Grades    % Churn
##  Min.   : 36.40            Min.   : 0.000       Min.   : 0.800
##  1st Qu.: 93.30            1st Qu.: 0.300       1st Qu.: 4.300
##  Median : 96.00            Median : 0.700       Median : 6.400
##  Mean   : 92.77            Mean   : 1.166       Mean   : 6.866
##  3rd Qu.: 97.50            3rd Qu.: 1.425       3rd Qu.: 8.900
##  Max.   :100.00            Max.   :15.600       Max.   :28.600
##     % Intake       % Stability    District_code
##  Min.   : 0.000   Min.   :81.20   Length:264
##  1st Qu.: 2.175   1st Qu.:94.60   Class :character
##  Median : 3.300   Median :96.30   Mode  :character
##  Mean   : 3.651   Mean   :95.89
##  3rd Qu.: 4.825   3rd Qu.:97.42
##  Max.   :16.800   Max.   :99.20
##  First Language Not English % English Language Learner %
##  Min.   : 0.000               Min.   : 0.000
##  1st Qu.: 2.975               1st Qu.: 1.100
##  Median : 7.800               Median : 2.700
##  Mean   :14.479               Mean   : 5.569
##  3rd Qu.:20.725               3rd Qu.: 6.700
##  Max.   :83.400              Max.   :35.700
##  Students With Disabilities %  High Needs %   Economically Disadvantaged %
##  Min.   : 0.00               Min.   : 3.10   Min.   : 2.00
##  1st Qu.:16.00               1st Qu.:29.88   1st Qu.:14.97
##  Median :18.00               Median :39.95   Median :25.55
##  Mean   :18.44               Mean   :43.37   Mean   :29.00
##  3rd Qu.:20.32               3rd Qu.:52.90   3rd Qu.:38.52
##  Max.   :44.10               Max.   :89.00   Max.   :81.80
##  Percent of HS in AP Adjusted Score
##  Min.   : 0.07645    Min.   : 546.1
##  1st Qu.:13.03754    1st Qu.: 965.3
##  Median :19.62341    Median :1044.4
##  Mean   :20.32260    Mean   :1038.4
##  3rd Qu.:26.61466    3rd Qu.:1111.5
##  Max.   :60.03086    Max.   :1461.2
```

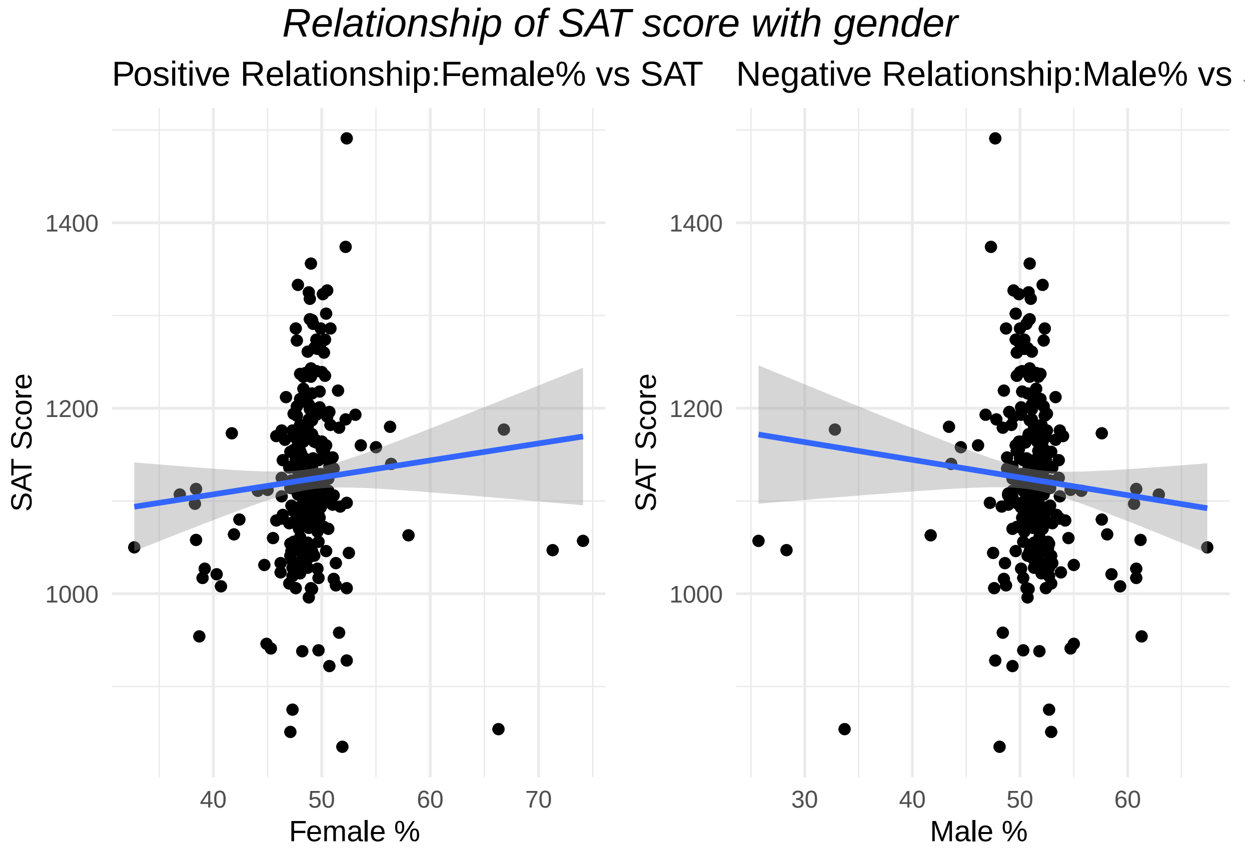
**Inference from summary:** 1) Neither Reading/Writing, nor Math has a perfect score in SAT, same goes for the total score.

2) Among races, at least one district had Hispanic and White students in domination,even though the third quartile of Hispanic students is at 16.43%.
3) Even though on average Male % students is more than Female % among districts, the minimum and maximum of gender % is higher for Females.
4) At least one district/school has 100% graduate rate and 0% drop rate and yet maximum percentage of students going to College is only 91.70.
5) At least one district/school has % Students Completing Advanced as 100% and none has 100% attendance rate.
6) The school/district with minimum % of Teachers Licensed, has almost half of the teachers unlicensed.
7) None of the school/district has **only** experienced teachers, and in at least one school, 67% teachers are not experienced.
8) The data is taken for the COVID time-period(2020-21), yet at least one school had Percent Teaching In-Field as 100%.
9) Even though the schools are in a country where English is the most-commonly spoken language, at least one school has 83.400% students whose first Language is not English.
10) None of the schools has 0% of High Needs or Economically Disadvantaged students.

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

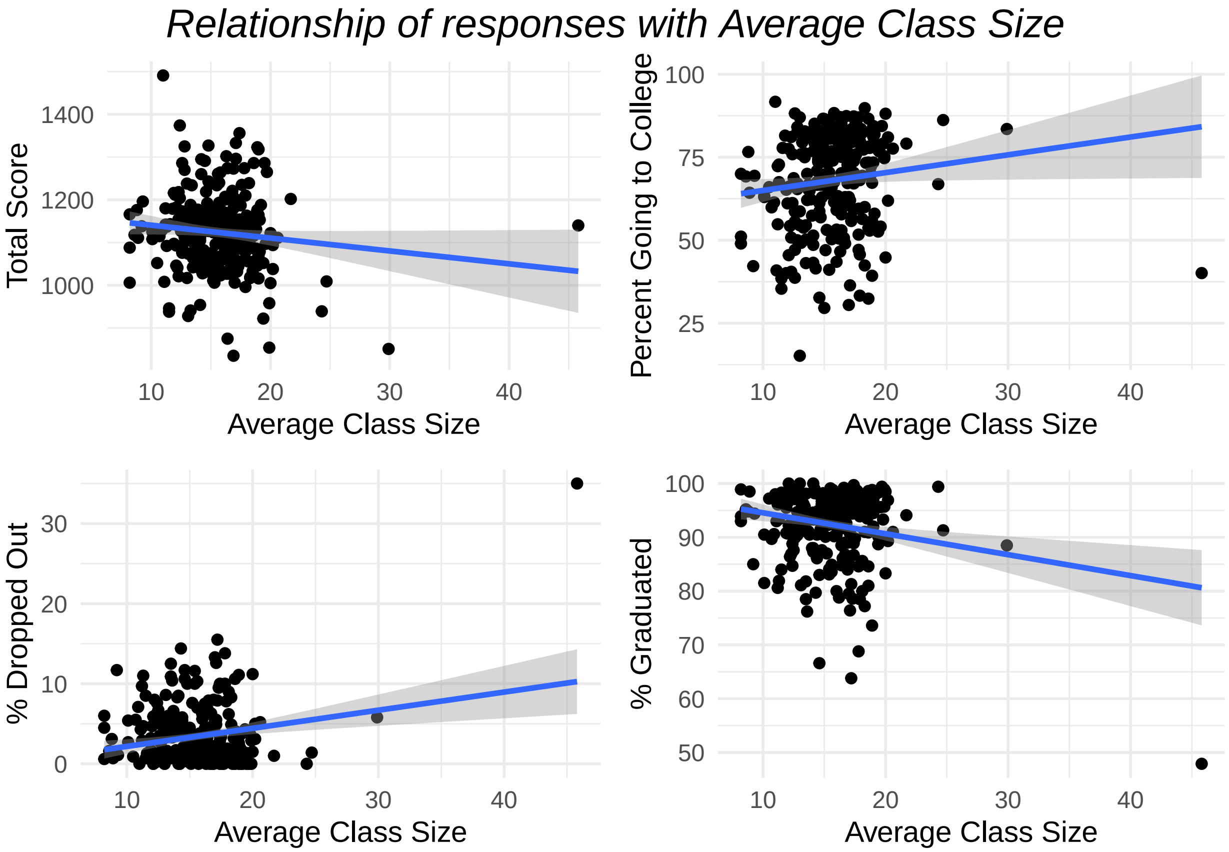
```
g1 <- eduData %>% ggplot( mapping=aes(x=`Female %`,y=`Total Score`))+
  geom_point() +
  geom_smooth(method=lm) +
  labs(title="Positive Relationship:Female% vs SAT",
       x="Female % ", y="SAT Score") +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`Male %`,y=`Total Score`))+
  geom_point() +
  geom_smooth(method=lm) +
  labs(title="Negative Relationship:Male% vs SAT ",
       x="Male % ", y="SAT Score") +
  theme_minimal()
```

```
gridExtra::grid.arrange(
  g1, g2,nrow=1 ,top = textGrob("Relationship of SAT score with gender",
                                   gp=gpar(fontsize=15,font=3)))
```



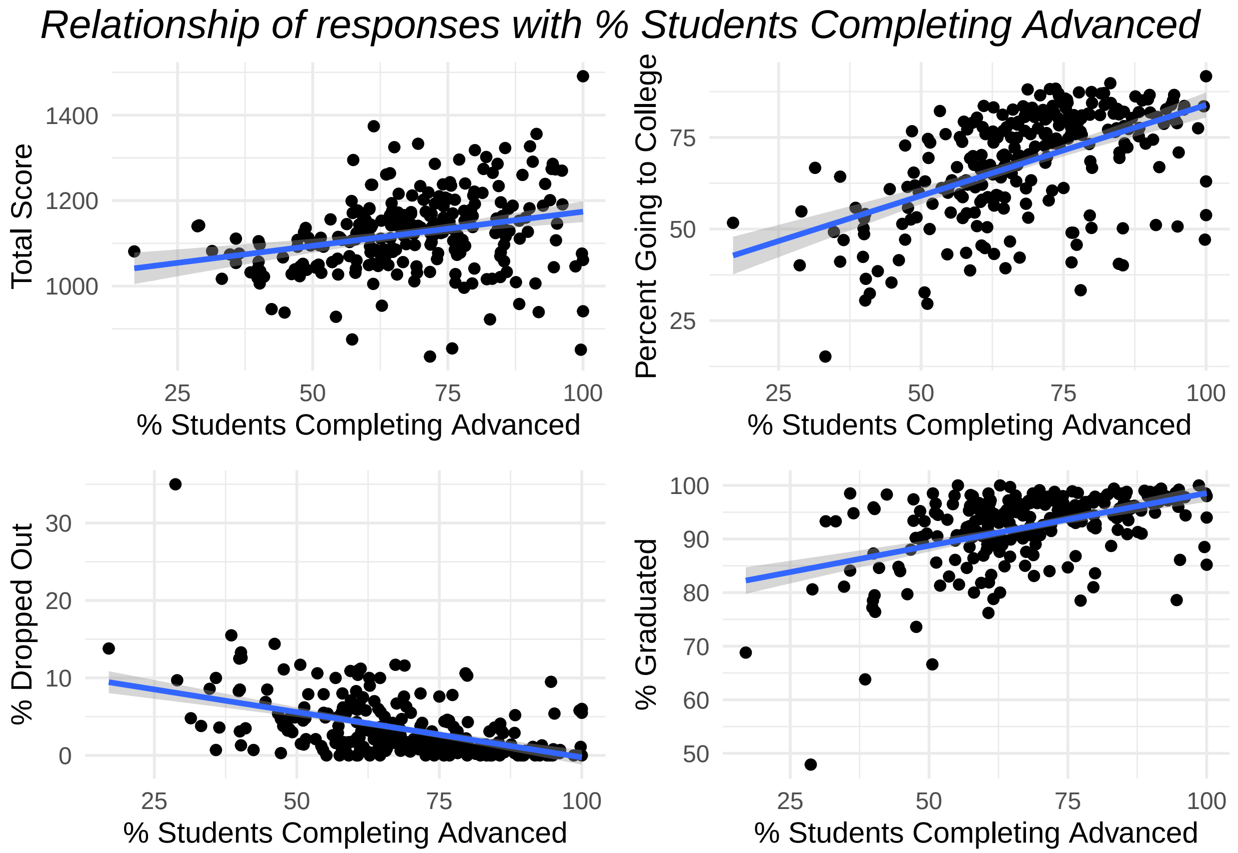
### Average Class Size vs SAT, Graduate Rate, drop rate and Enrollment in college

```
g1 <- eduData %>% ggplot( mapping=aes(x=`Average Class Size`,y=`Total Score`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`Average Class Size`,
                                      y=`Percent Going to College`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g3 <- eduData %>% ggplot( mapping=aes(x=`Average Class Size`,
                                      y=`% Dropped Out`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g4 <- eduData %>% ggplot( mapping=aes(x=`Average Class Size`,y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

gridExtra::grid.arrange(
  g1, g2,g3,g4 ,
  top = textGrob("Relationship of responses with Average Class Size",
                                                gp=gpar(fontsize=15,font=3)))
```



There happens to be an outlier, which is in fact a correctly reported value (~45), so we won't remove it. Average Class Size has negative relationship with SAT score and % graduated, while it has a positive relationship with % going to college and %dropped out.

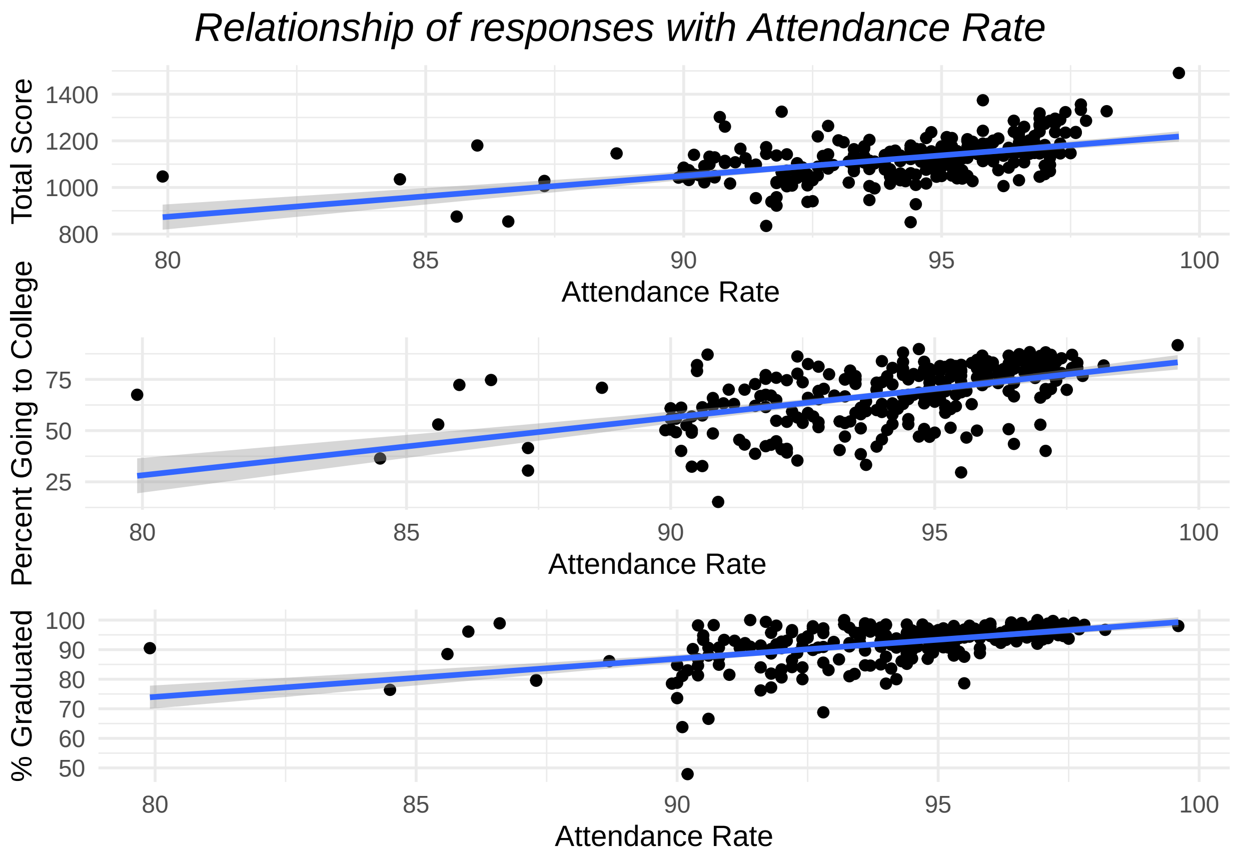
```
g1 <- eduData %>% ggplot( mapping=aes(x=`% Students Completing Advanced`,
                                      y=`Total Score`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`% Students Completing Advanced`,
                                      y=`Percent Going to College`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g3 <- eduData %>% ggplot( mapping=aes(x=`% Students Completing Advanced`,
                                      y=`% Dropped Out`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g4 <- eduData %>% ggplot( mapping=aes(x=`% Students Completing Advanced`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

gridExtra::grid.arrange(
  g1, g2,g3,g4 ,
  top = textGrob("Relationship of responses with % Students Completing Advanced",
                                     gp=gpar(fontsize=15,font=3)))
```



All responses had positive relationship with % Students Completing Advanced except %dropped out, which is expected since it's inverse of %graduated so, here onwards we will consider only one of them.

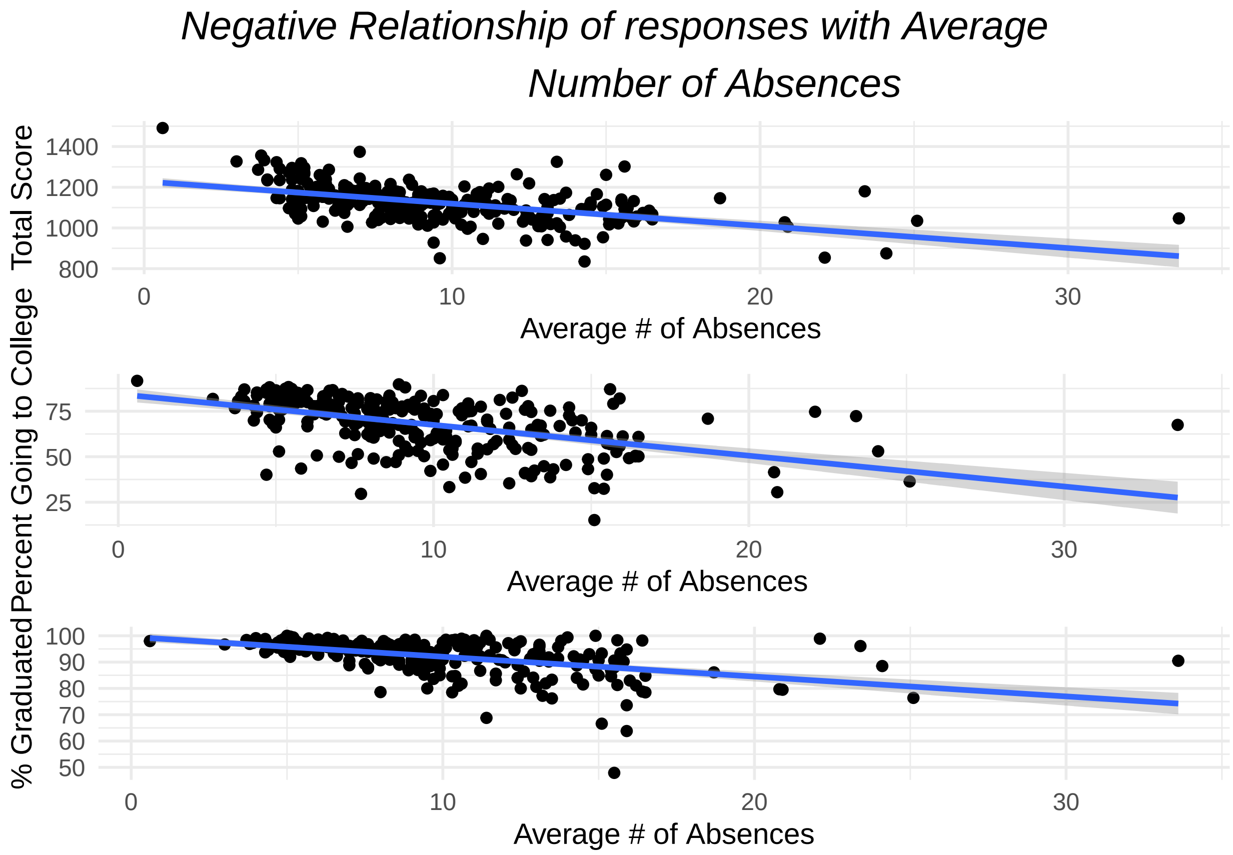
### Attendance Rate vs Responses

```
g1 <- eduData %>% ggplot( mapping=aes(x=`Attendance Rate`,
                                      y=`Total Score`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`Attendance Rate`,
                                      y=`Percent Going to College`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g3 <- eduData %>% ggplot( mapping=aes(x=`Attendance Rate`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

gridExtra::grid.arrange(
  g1,g2,g3,
  top = textGrob("Relationship of responses with Attendance Rate",
                                   gp=gpar(fontsize=15,font=3)))
```

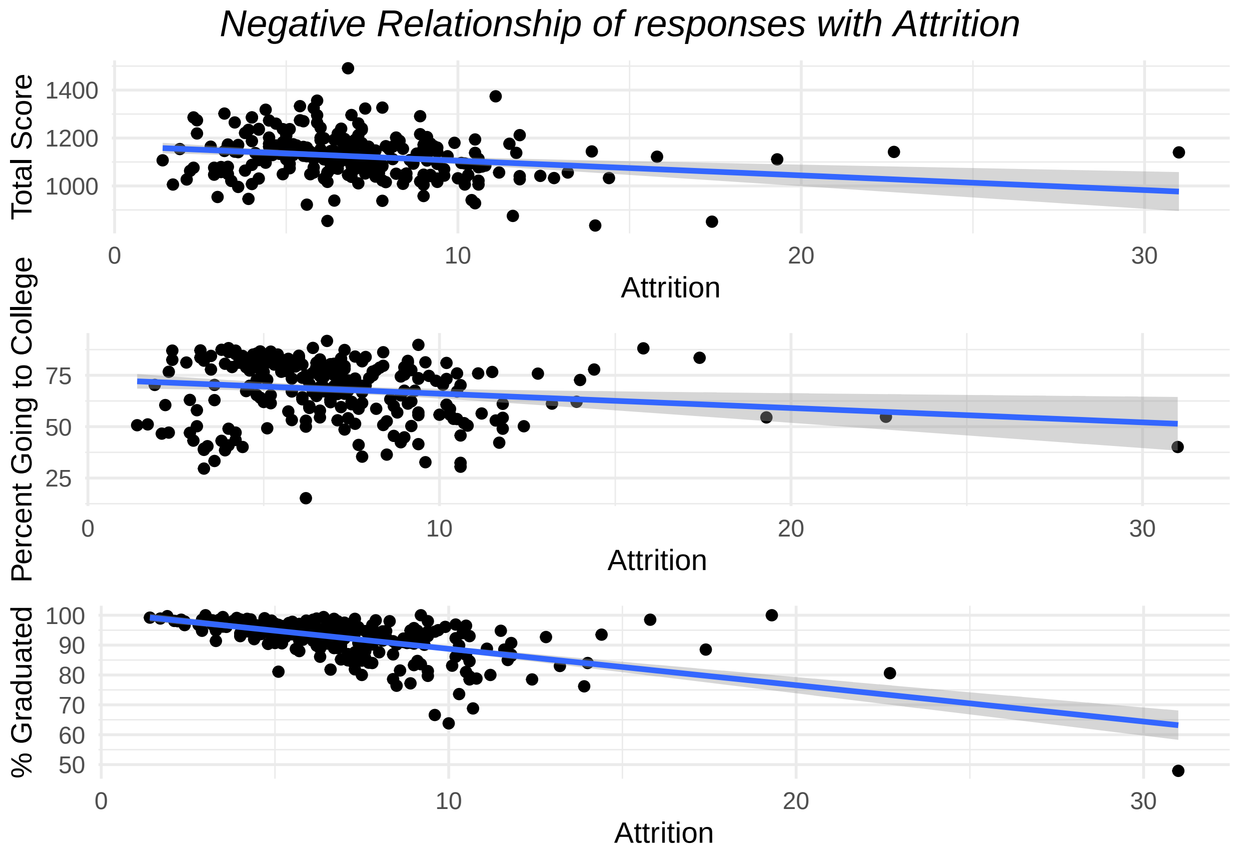


```
g1 <- eduData %>% ggplot( mapping=aes(x=`Average # of Absences`,
                                      y=`Total Score`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`Average # of Absences`,
                                      y=`Percent Going to College`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()
```

```
g3 <- eduData %>% ggplot( mapping=aes(x=`Average # of Absences`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

gridExtra::grid.arrange(
  g1,g2,g3,
  top = textGrob("Negative Relationship of responses with Average
                Number of Absences", gp=gpar(fontsize=15,font=3)))
```



```
g1 <- eduData %>% ggplot( mapping=aes(x=`Attrition`,
                                      y=`Total Score`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`Attrition`,
                                      y=`Percent Going to College`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g3 <- eduData %>% ggplot( mapping=aes(x=`Attrition`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

gridExtra::grid.arrange(
  g1,g2,g3,
  top = textGrob("Negative Relationship of responses with Attrition",
                gp=gpar(fontsize=14,font=3)))
```

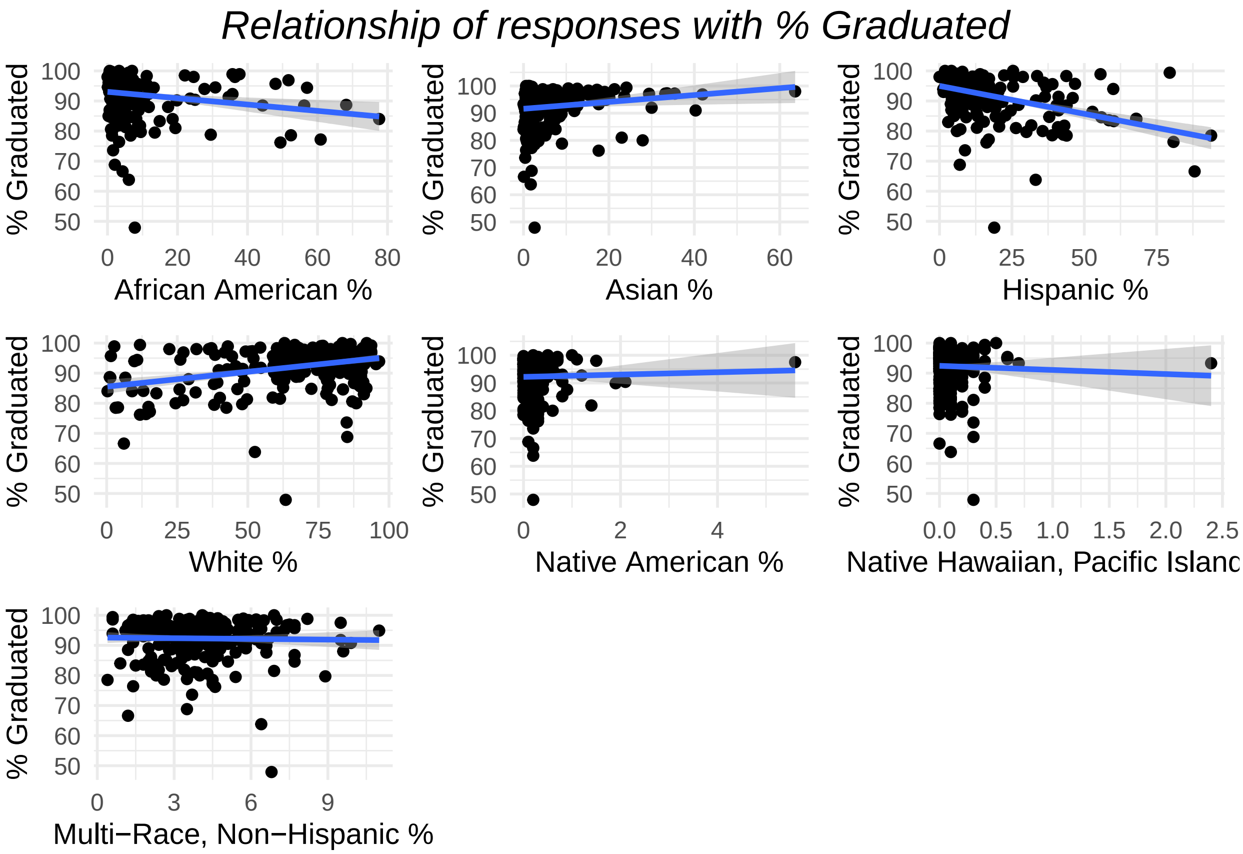


Even though, all of them has negative relationship, the slopes are different i.e., graduation rate drops with larger difference as compared to other responses.

### Student Background vs Graduation Rate

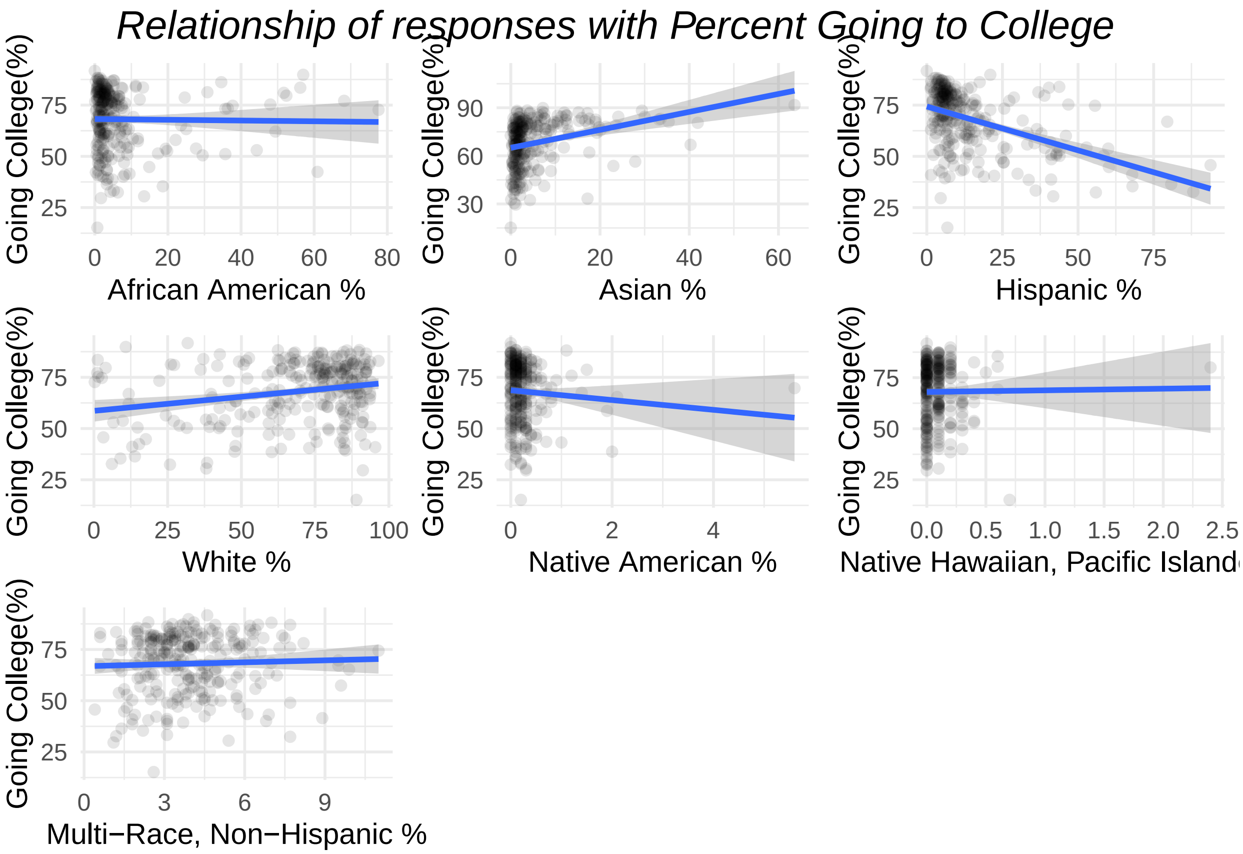
```
g1 <- eduData %>% ggplot( mapping=aes(x=`African American %`,
                                       y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()#-

g2 <- eduData %>% ggplot( mapping=aes(x=`Asian %`,
                                       y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g3 <- eduData %>% ggplot( mapping=aes(x=`Hispanic %`,
                                       y=`% Graduated`))+
```

```r
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()#-

g4 <- eduData %>% ggplot( mapping=aes(x=`White %`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g5 <- eduData %>% ggplot( mapping=aes(x=`Native American %`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g6 <- eduData %>% ggplot( mapping=aes(x=`Native Hawaiian, Pacific Islander %`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()#-

g7 <- eduData %>% ggplot( mapping=aes(x=`Multi-Race, Non-Hispanic %`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

gridExtra::grid.arrange(
  g1,g2,g3,g4,g5,g6,g7,
  top = textGrob("Relationship of responses with % Graduated",
                 gp=gpar(fontsize=15,font=3)))
```



While “African American”, “Hispanic” and “Native Hawaiian,Pacific Islander” students have negative relationship with graduation rate, “Multi-Race, Non-Hispanic” students have somewhat constant graduation rate. Something else to notice is that population of “white” with higher graduation rate is closer to 100% while for others,higher graduation rate is closer to 0%.

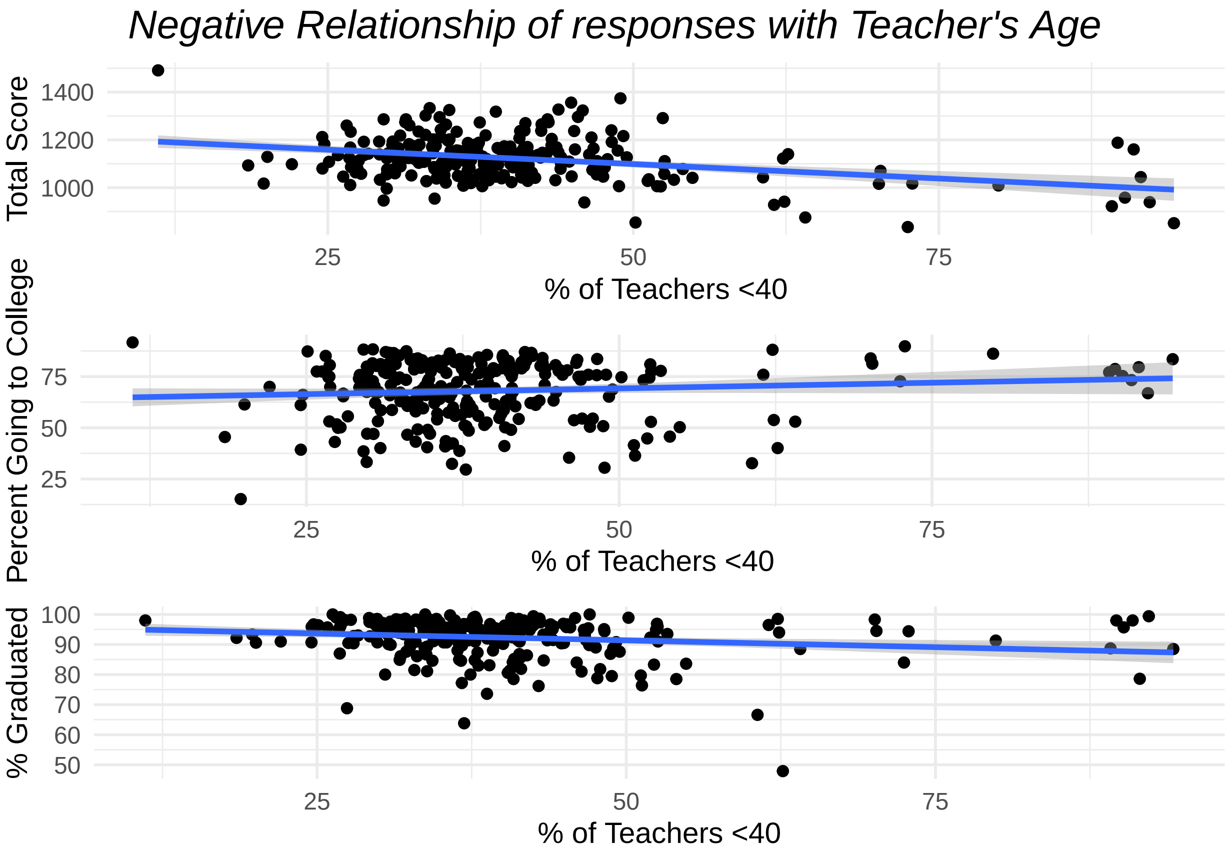
```
g1 <- eduData %>% ggplot( mapping=aes(x=`African American %`,
                                      y=`Percent Going to College`))+
  geom_point(alpha=0.1) +
  geom_smooth(method=lm) +
  labs(y="Going College(%)") +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`Asian %`,
                                      y=`Percent Going to College`))+
  geom_point(alpha=0.1) +
  geom_smooth(method=lm) +
  labs(y="Going College(%)") +
  theme_minimal()

g3 <- eduData %>% ggplot( mapping=aes(x=`Hispanic %`,
                                      y=`Percent Going to College`))+
  geom_point(alpha=0.1) +
  geom_smooth(method=lm) +
  labs(y="Going College(%)") +
  theme_minimal()#-

g4 <- eduData %>% ggplot( mapping=aes(x=`White %`,
                                      y=`Percent Going to College`))+
  geom_point(alpha=0.1) +
  geom_smooth(method=lm) +
  labs(y="Going College(%)") +
  theme_minimal()

g5 <- eduData %>% ggplot( mapping=aes(x=`Native American %`,
                                     y=`Percent Going to College`))+
  geom_point(alpha=0.1) +
  geom_smooth(method=lm) +
  labs(y="Going College(%)") +
  theme_minimal()#-

g6 <- eduData %>% ggplot( mapping=aes(x=`Native Hawaiian, Pacific Islander %`,
                                     y=`Percent Going to College`))+
  geom_point(alpha=0.1) +
  geom_smooth(method=lm) +
  labs(y="Going College(%)") +
  theme_minimal()

g7 <- eduData %>% ggplot( mapping=aes(x=`Multi-Race, Non-Hispanic %`,
                                     y=`Percent Going to College`))+
  geom_point(alpha=0.1) +
  geom_smooth(method=lm) +
  labs(y="Going College(%)") +
  theme_minimal()

gridExtra::grid.arrange(
  g1,g2,g3,g4,g5,g6,g7,
  top = textGrob("Relationship of responses with Percent Going to College",
                 gp=gpar(fontsize=15,font=3)))
```



When plotted against "Going to College(%)",the only differences were that "African American" showed no significant change,and "Native American" had a negative relationship,while "Multi-Race, Non-Hispanic" and "Native Hawaiian, Pacific Islander" students have slightly positive relationship.

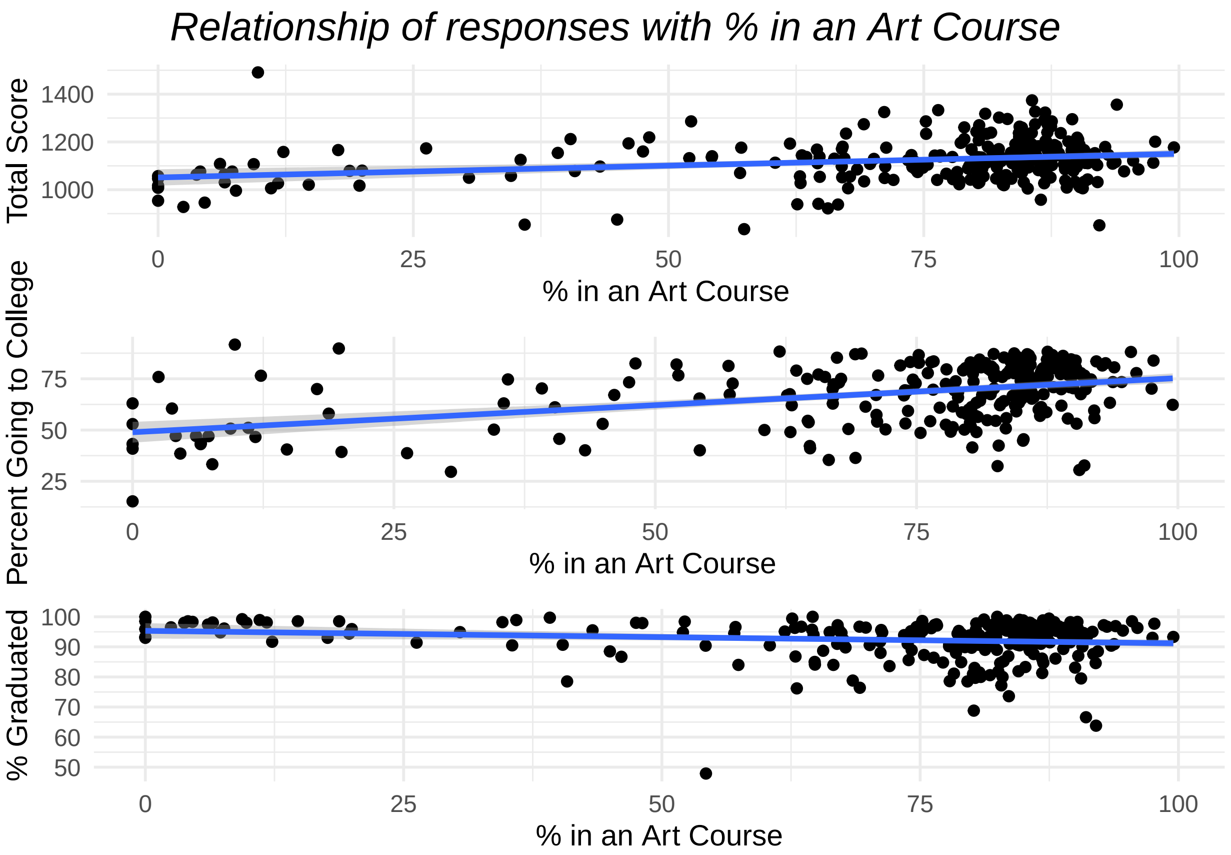
### % of Teachers <40 and responses

```
g1 <- eduData %>% ggplot( mapping=aes(x=`% of Teachers <40`,
                                      y=`Total Score`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`% of Teachers <40`,
                                      y=`Percent Going to College`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g3 <- eduData %>% ggplot( mapping=aes(x=`% of Teachers <40`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

gridExtra::grid.arrange(
  g1,g2,g3,
  top = textGrob("Negative Relationship of responses with Teacher's Age",
                gp=gpar(fontsize=15,font=3)))
```

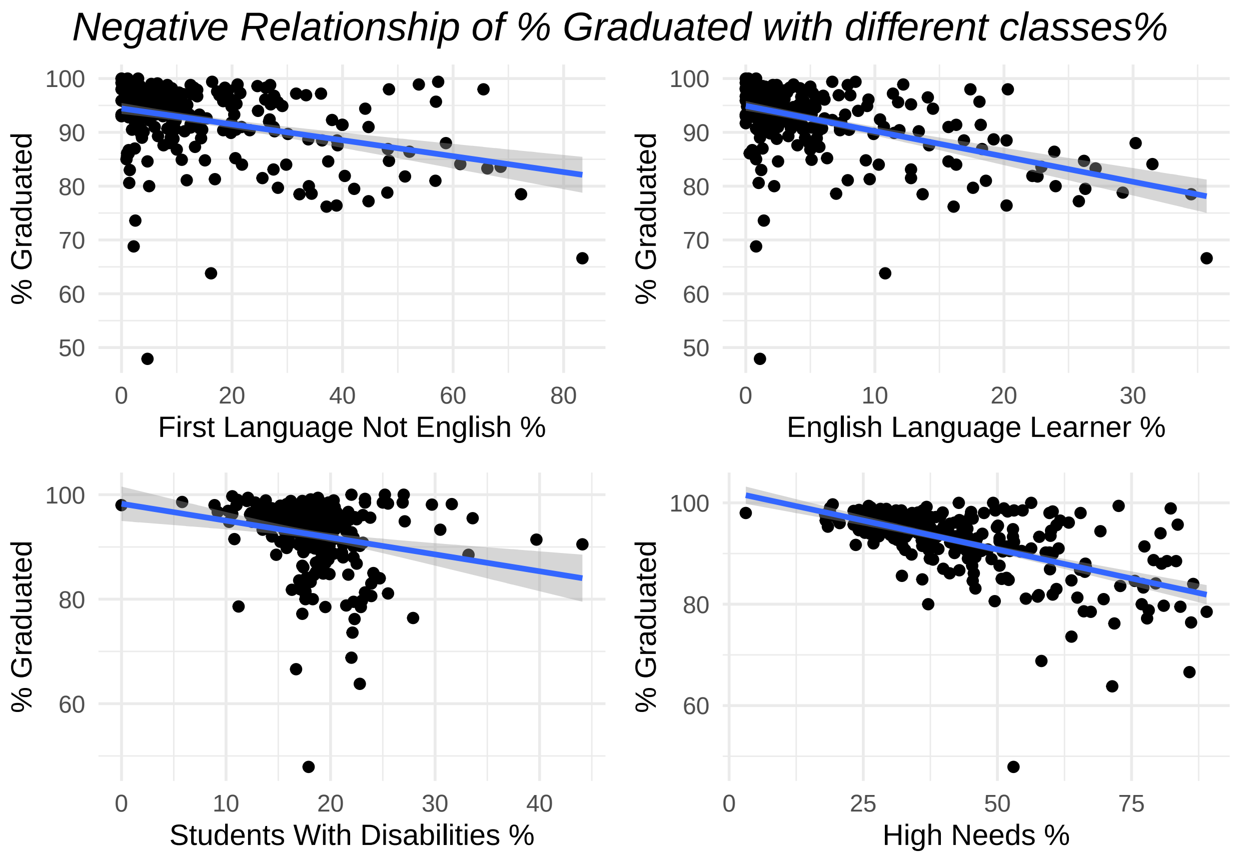


Schools with more % of teachers's age less than 40, had a negative affect on Total SAT score and graduation rate while a positive affect on percentage of students going to college.

```
g1 <- eduData %>% ggplot( mapping=aes(x=`% in an Art Course`,
                                      y=`Total Score`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`% in an Art Course`,
                                      y=`Percent Going to College`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g3 <- eduData %>% ggplot( mapping=aes(x=`% in an Art Course`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

gridExtra::grid.arrange(
  g1,g2,g3,
  top = textGrob("Relationship of responses with % in an Art Course",
                gp=gpar(fontsize=15,font=3)))
```



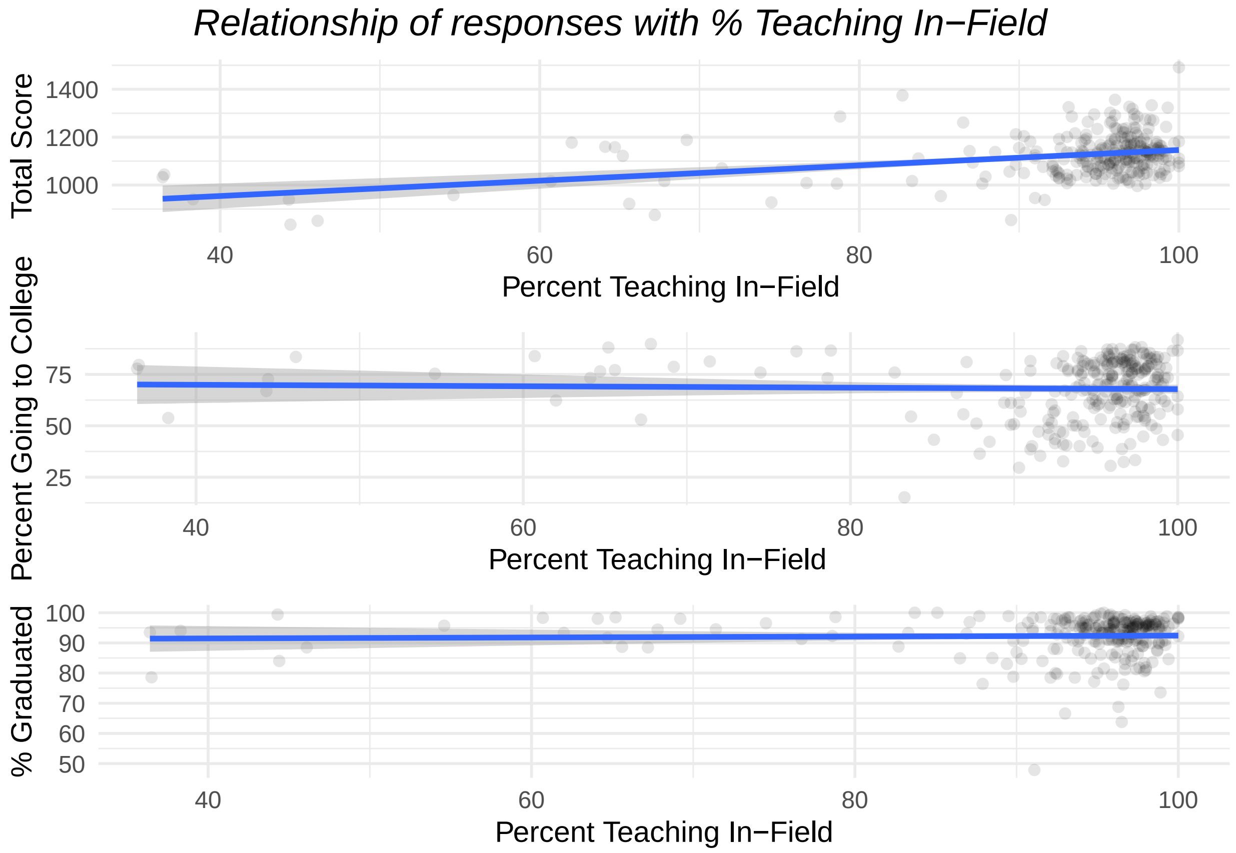
```
g1 <- eduData %>% ggplot( mapping=aes(x=`First Language Not English %`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`English Language Learner %`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g3 <- eduData %>% ggplot( mapping=aes(x=`Students With Disabilities %`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

g4 <- eduData %>% ggplot( mapping=aes(x=`High Needs %`,
                                      y=`% Graduated`))+
  geom_point() +
  geom_smooth(method=lm) +
  theme_minimal()

gridExtra::grid.arrange(
  g1,g2,g3,g4,
  top = textGrob("Negative Relationship of % Graduated with different classes%",
                 gp=gpar(fontsize=15,font=3)))
```



```
# English Language Learner %
# Students With Disabilities %
# High Needs %
# First Language Not English %
```

**Percent Teaching In-Field vs 3 responses**

```
# Percent Teaching In-Field
g1 <- eduData %>% ggplot( mapping=aes(x=`Percent Teaching In-Field`,
                                      y=`Total Score`))+
  geom_point(alpha=0.1) +
  geom_smooth(method=lm) +
  theme_minimal()

g2 <- eduData %>% ggplot( mapping=aes(x=`Percent Teaching In-Field`,
                                      y=`Percent Going to College`))+
  geom_point(alpha=0.1) +
  geom_smooth(method=lm) +
  theme_minimal()

g3 <- eduData %>% ggplot( mapping=aes(x=`Percent Teaching In-Field`,
                                      y=`% Graduated`))+
  geom_point(alpha=0.1) +
  geom_smooth(method=lm) +
  theme_minimal()

gridExtra::grid.arrange(
  g1,g2,g3,
  top = textGrob("Relationship of responses with % Teaching In-Field",
                 gp=gpar(fontsize=14,font=3)))
```



Percentage going to college has significantly small negative relationship.