IST565 Data Mining

HW2 Instruction

This is a real job interview question from a data analysis company, and I doubt there is a standard answer to this question. So feel free to explore your story by using the data exploration and transformation techniques appropriately.

————-instruction quote begins————-

Here is a small dataset for you to work with.

Each of 5 schools (A, B, C, D and E) is implementing the same math course this semester, with 35 lessons. There are 30 sections total. The semester is about 3/4 of the way through.

For each section, we record the number of students who are:

• very ahead (more than 5 lessons ahead) • middling (5 lessons ahead to 0 lessons ahead) • behind (1 to 5 lessons behind) • more behind (6 to 10 lessons behind) • very behind (more than 10 lessons behind) • completed (finished with the course)

What's the story (or stories) in this data? Find it, and tell it visually and, above all, truthfully.

————instruction quote ends————

```
#Student Name: Alireza Zarrinmehr
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
mydata <-read.csv("/Users/zarrinmehr/Desktop/13- IST.707.M002.FALL22.Applied Machine Learning 16797.123
head(mydata)
```

```
##   School Section Very.Ahead..5 Middling..0 Behind..1.5 More.Behind..6.10
## 1      A       1             0           5          54                 3
## 2      A       2             0           8          40                10
## 3      A       3             0           9          35                12
## 4      A       4             0          14          44                 5
## 5      A       5             0           9          42                 2
## 6      A       6             0           7          29                 3
##   Very.Behind..11 Completed
## 1               9        10
## 2              16         6
```

```
## 3              13        11
## 4              12        10
## 5              24         8
## 6              10         9
```

```
str(mydata)
```

```
## 'data.frame':    30 obs. of  8 variables:
##  $ School           : chr  "A" "A" "A" "A" ...
##  $ Section          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Very.Ahead..5    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Middling..0      : int  5 8 9 14 9 7 19 3 6 13 ...
##  $ Behind..1.5      : int  54 40 35 44 42 29 22 37 29 40 ...
##  $ More.Behind..6.10: int  3 10 12 5 2 3 5 11 8 5 ...
##  $ Very.Behind..11  : int  9 16 13 12 24 10 14 18 12 5 ...
##  $ Completed        : int  10 6 11 10 8 9 19 5 10 20 ...
```

Getting the summary of the data

```
summary(mydata)
```

```
##     School             Section      Very.Ahead..5  Middling..0
##  Length:30          Min.   : 1.00   Min.   :0     Min.   : 2.00
##  Class :character   1st Qu.: 2.25   1st Qu.:0     1st Qu.: 4.25
##  Mode  :character   Median : 5.50   Median :0     Median : 7.50
##                     Mean   : 5.90   Mean   :0     Mean   : 7.40
##                     3rd Qu.: 9.00   3rd Qu.:0     3rd Qu.: 9.75
##                     Max.   :13.00   Max.   :0     Max.   :19.00
##   Behind..1.5    More.Behind..6.10 Very.Behind..11    Completed
##  Min.   : 4.00   Min.   : 0.000    Min.   : 0.000   Min.   : 1.00
##  1st Qu.:15.25   1st Qu.: 1.000    1st Qu.: 1.250   1st Qu.: 6.00
##  Median :22.00   Median : 2.000    Median : 5.500   Median :10.00
##  Mean   :25.13   Mean   : 3.333    Mean   : 6.967   Mean   :10.53
##  3rd Qu.:34.25   3rd Qu.: 4.750    3rd Qu.:11.500   3rd Qu.:14.00
##  Max.   :56.00   Max.   :12.000    Max.   :24.000   Max.   :27.00
```

Among all the schools and sections on Average: • No student is more than 5 lessons ahead • 7 students are 0 to 5 lessons ahead • 25 students are 1 to 5 lessons behind • 3 students are 6 to 10 lessons behind • 6 students are more than 10 lessons behind • 10 students finished with the course Converting Section to the nominal variable:

```
mydata$Section <- factor(mydata$Section)
```

Changing the names for easier handling:

```
names <- c("school", "section", "veryAhead5", "middling0", "behind15", "moreBehind610",
    "veryBehind11", "completed")
colnames(mydata) <- names
```

Gathering the data in a long format to do grouping:

```
tidymydata <- mydata %>% gather(key = "key", value = "value", c(veryAhead5, middling0, behind15, moreBe
    veryBehind11, completed))
```

Demonstrating the gathered data:

```
#tidyrmydata
```

Grouping the data across the columns to show how many students each school has:

```
tidymydata %>% group_by(school) %>% summarize(NumSections = max(as.numeric(section)), NumStudents = sum
```

```
## # A tibble: 5 x 3
##   school NumSections NumStudents
##   <chr>        <dbl>       <int>
## 1 A               13         932
## 2 B               12         446
## 3 C                3          85
## 4 D                1          22
## 5 E                1         116
```

School A has more than twice as many students as school B while having about the same number of sections. School E has more than five times as many students as school D while having the same number of sections. School A has the maximum number of students with 932 students. School D is considered the smallest school with just 22 students.

```
tidymydata %>% group_by(key) %>% summarize(NumStudents = sum(value))
```

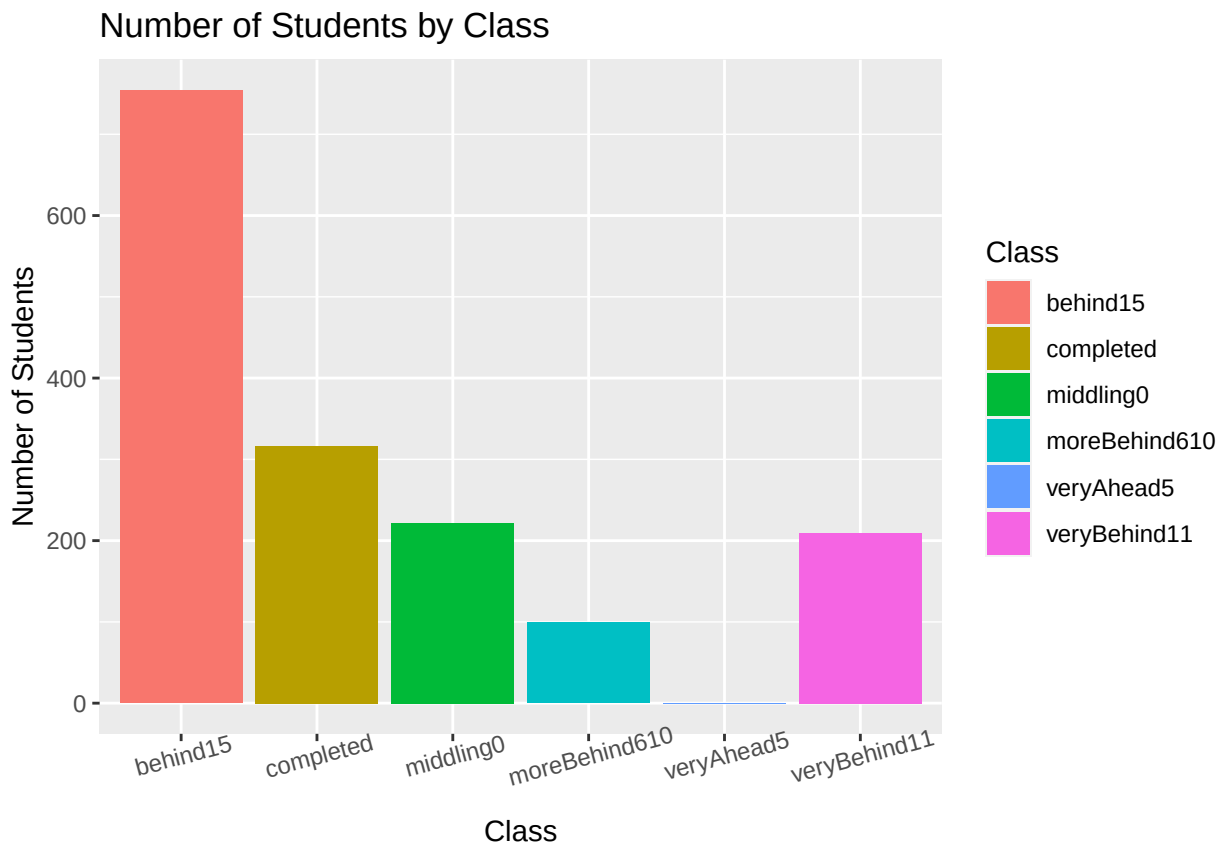
```
## # A tibble: 6 x 2
##   key           NumStudents
##   <chr>               <int>
## 1 behind15              754
## 2 completed             316
## 3 middling0             222
## 4 moreBehind610         100
## 5 veryAhead5              0
## 6 veryBehind11          209
```

Among all the schools and sections: • No student is more than 5 lessons ahead • 222 students are 0 to 5 lessons ahead • 754 students are 1 to 5 lessons behind • 100 students are 6 to 10 lessons behind • 209 students are more than 10 lessons behind • 316 students finished with the course

Even though the second-largest group of students has finished all of their coursework, the program is ineffective because the great majority of participants are falling behind in their classes.

observing how pupils are distributed across the various classes.

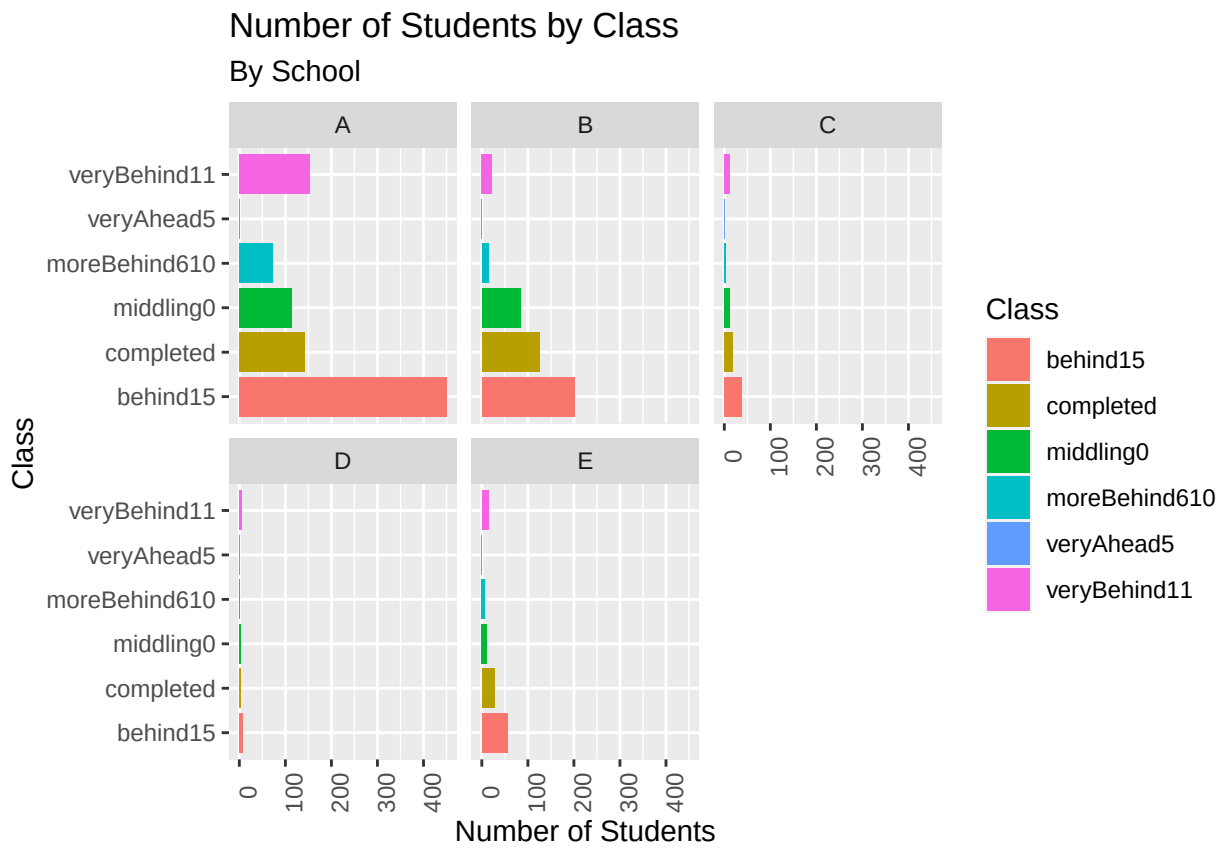
```
ggplot(tidymydata, aes(x = key, y = value, fill = key)) +
    geom_bar(stat = "identity") + labs(x = "Class", y = "Number of Students",
    fill = "Class") + ggtitle("Number of Students by Class") + theme(axis.text.x = element_text(angle =
```



The number of students in the Behind -1-5 group is noticeably higher, as can be shown. Examining each school separately to check if the distribution is the same:

```
tidymydata %>% group_by(school, key) %>% summarise(numStudents = sum(value)) %>%
    ggplot(aes(x = key, y = numStudents, fill = key)) + geom_bar(stat = "identity") +
    facet_wrap(~school) + labs(x = "Class", y = "Number of Students", fill = "Class") +
    ggtitle("Number of Students by Class", subtitle = "By School") + coord_flip() +
    theme(axis.text.x = element_text(angle = 90))
```

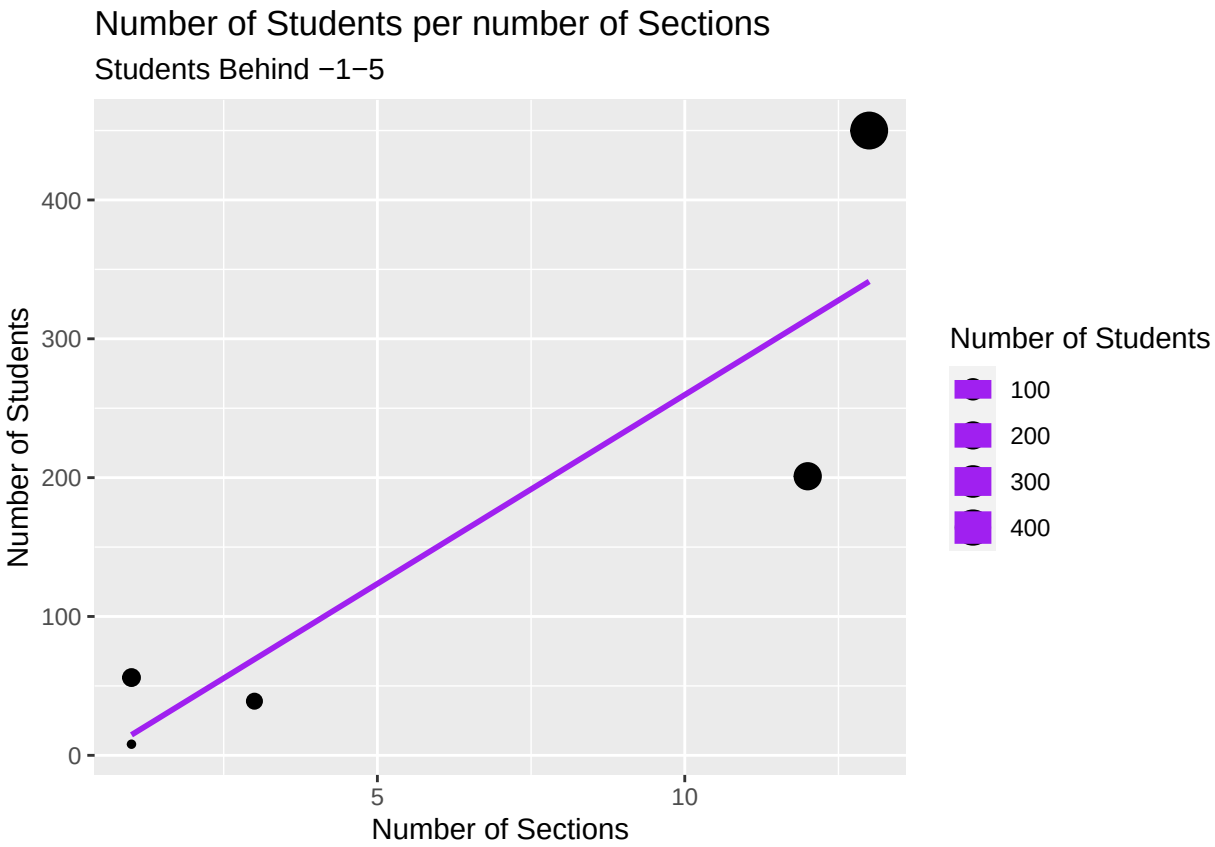
```
## `summarise()` has grouped output by 'school'. You can override using the
## `.groups` argument.
```



We can tell how many pupils at school A are falling behind in their math classes. 450 students are one to five courses behind schedule (and 677 behind in total). While the remaining schools have lower rates of behind-ness. In addition, school A's completion rate is 15.2, whereas the completion rate for the other schools is 26.0 percent. This seems to indicate that the issue is not with the program per se, but rather with school A! This prompts a lot more inquiries. Do the number of sections and the number of pupils that are falling behind correlate? In school A, how are students divided up into sections?

```
tidymydata %>% filter(key == "behind15") %>% group_by(school) %>% summarise(numSections = max(as.numeri
    numStudents = sum(value)) %>% ggplot(aes(x = numSections, y = numStudents,
    size = numStudents)) + geom_point() + geom_smooth(method = "lm", colour = "purple",
    se = FALSE) + labs(x = "Number of Sections", y = "Number of Students", size = "Number of Students")
    ggtitle("Number of Students per number of Sections", subtitle = "Students Behind -1-5")
```

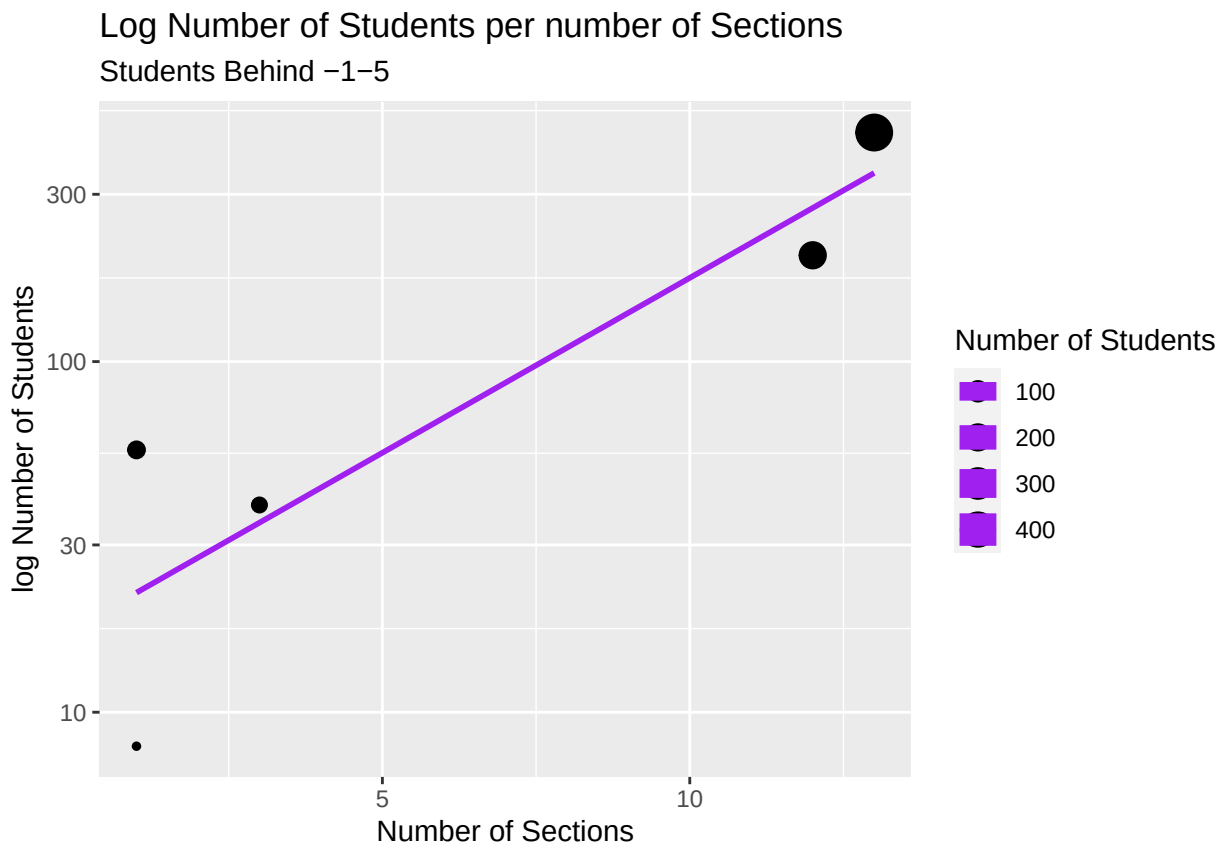
```
## `geom_smooth()` using formula 'y ~ x'
```



The number of sections and the number of students are related. We logarithmize the y-scale for further investigation.

```
tidymydata %>% filter(key == "behind15") %>% group_by(school) %>% summarise(numSections = max(as.numeri
    numStudents = sum(value)) %>% ggplot(aes(x = numSections, y = numStudents,
    size = numStudents)) + geom_point() + geom_smooth(method = "lm", colour = "purple",
    se = FALSE) + scale_y_log10() + labs(x = "Number of Sections", y = "log Number of Students",
    size = "Number of Students") + ggtitle("Log Number of Students per number of Sections",
    subtitle = "Students Behind -1-5")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

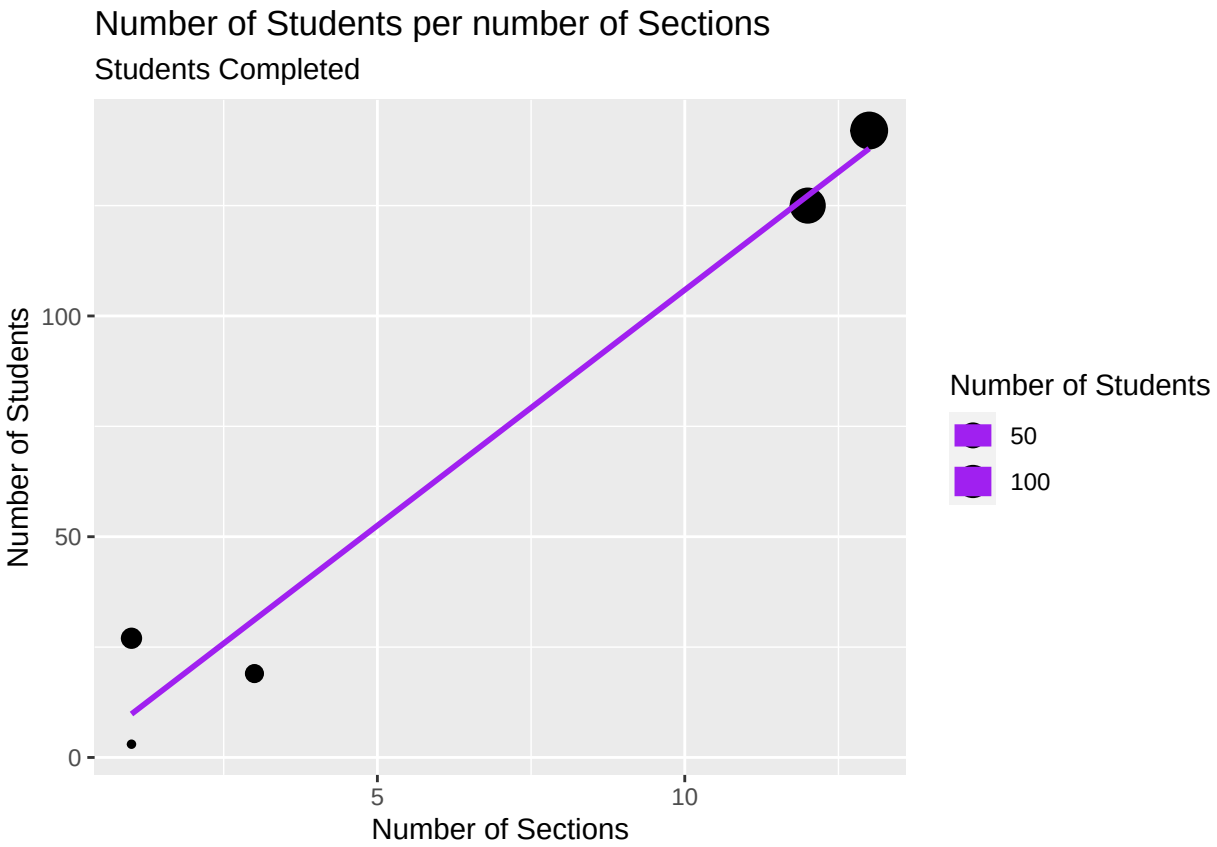


It might not be advisable to conclude that these two variables are correlated because the sample size is so small.

Visualizing the statistics for students who have finished their math courses:

```
tidymydata %>% filter(key == "completed") %>% group_by(school) %>% summarise(numSections = max(as.numer
    numStudents = sum(value)) %>% ggplot(aes(x = numSections, y = numStudents,
    size = numStudents)) + geom_point() + geom_smooth(method = "lm", colour = "purple",
    se = FALSE) + labs(x = "Number of Sections", y = "Number of Students", size = "Number of Students")
    ggtitle("Number of Students per number of Sections", subtitle = "Students Completed")
```

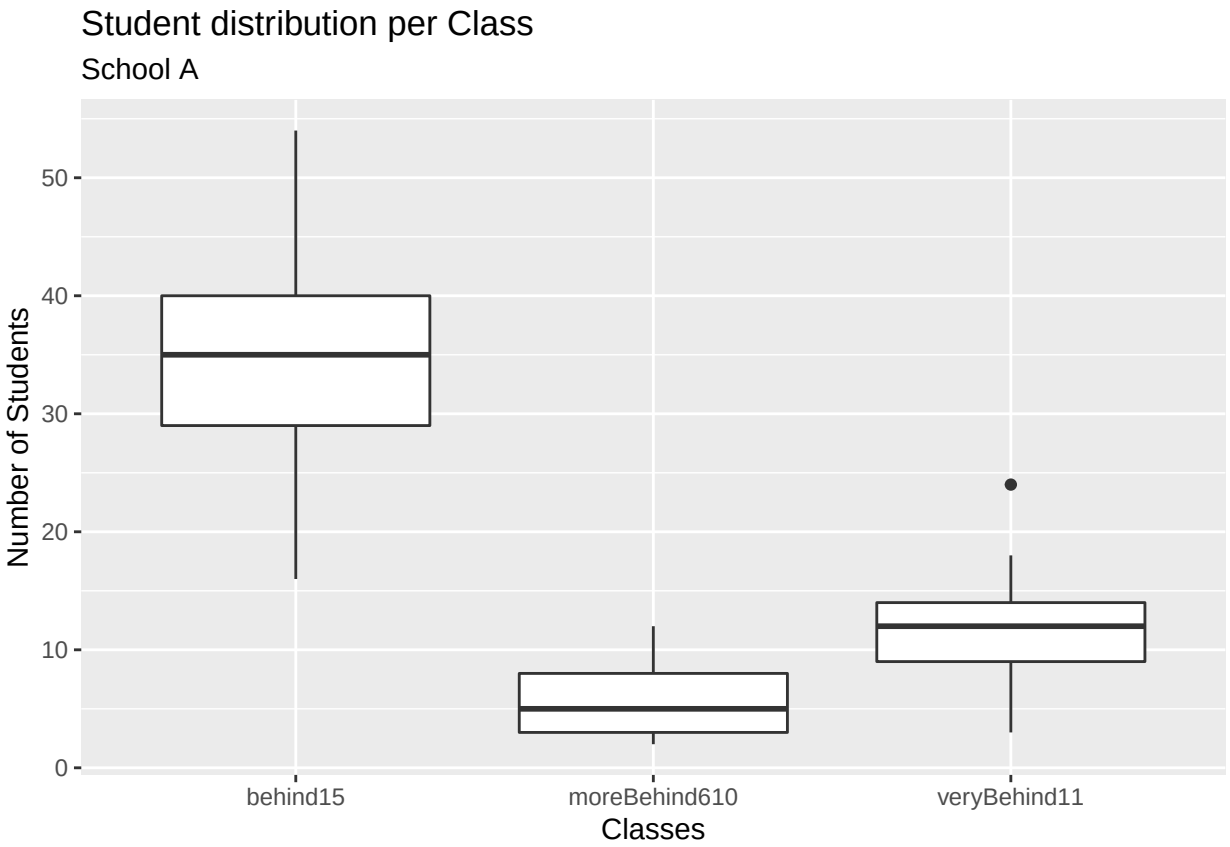
```
## `geom_smooth()` using formula 'y ~ x'
```



Thus, the number of sections does have an impact on the proportion of students who pass their courses or fall behind. However, just because there are more sections doesn't necessarily mean that more students will fail or fall behind in their courses.

examining the distribution of children who are behind in all school sections A.

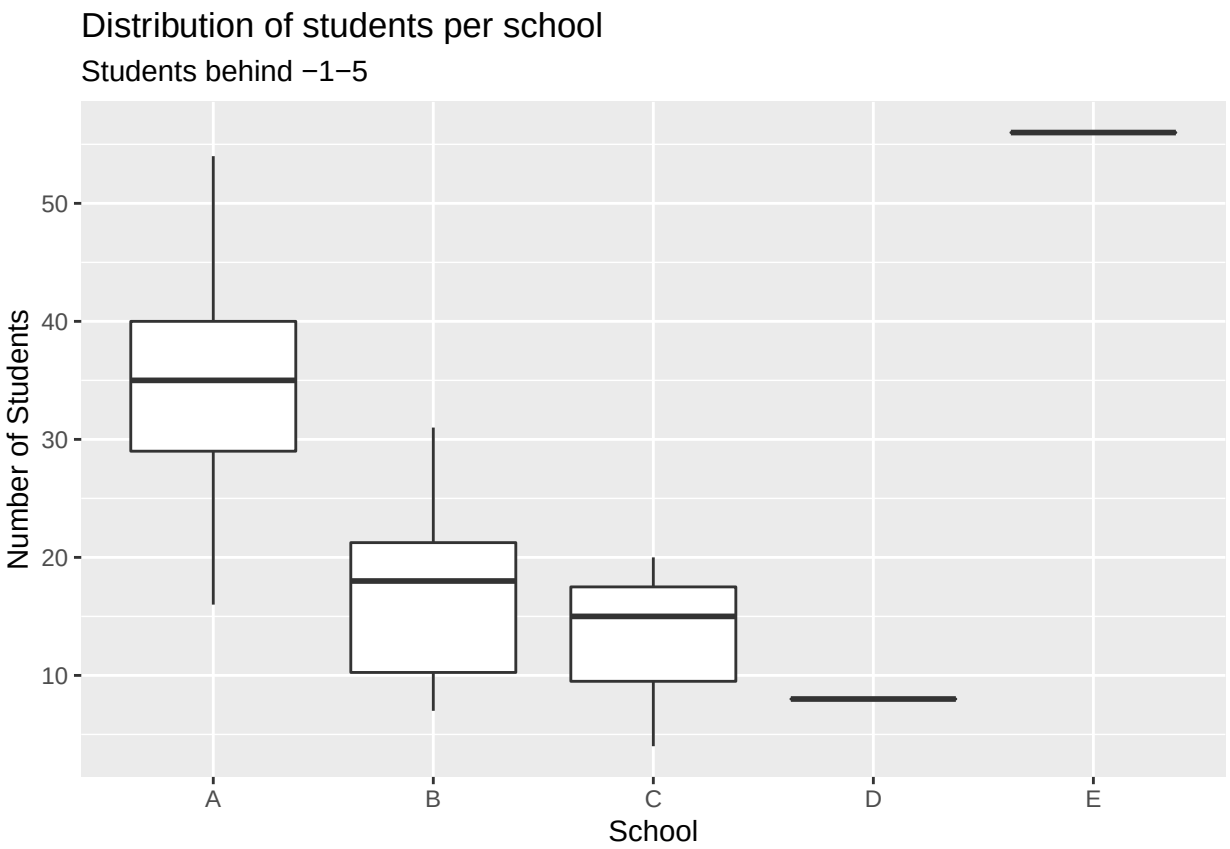
```
tidymydata %>% filter(school == "A" & key %in% c("behind15", "moreBehind610",
    "veryBehind11")) %>% ggplot(aes(x = key, y = value)) + geom_boxplot() +
    labs(x = "Classes", y = "Number of Students") + ggtitle("Student distribution per Class",
    subtitle = "School A")
```



There don't appear to be any outliers that are causing so many students at School A to be falling behind. The largest section, however, has nearly 55 students falling behind, while the smallest section has just a little bit more than 15. This clarifies that the number of sections is what's driving school A's high statistics, with a median of 35 students and an average of 34.6 students.

Comparing the figures across all schools:

```
tidymydata %>% filter(key %in% c("behind15")) %>% ggplot(aes(x = school,
    y = value)) + geom_boxplot() + labs(x = "School", y = "Number of Students") +
    ggtitle("Distribution of students per school", subtitle = "Students behind -1-5")
```



56 pupils in School E's one section are behind in one to five courses. No outliers or obvious patterns stand out when attempting to explain why class A has more pupils performing below average. There isn't a good reason for school A to be increasing the numbers other than the fact that they have more pupils and sections.