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-----------------

ANSWERS

#Load required packages

library(ggplot2)  
library(dplyr)  
library(tidyr)

#Import the data from .csv and list

myschool <-**read**.**csv**("C:/Users/subbuk/Documents/MSDatascience/IST707/storyteller.csv")

head(myschool)

school section very\_ahead middling\_0 behind\_1\_5 more\_behind\_6\_10 very\_behind\_11 completed

1 A 1 0 5 54 3 9 10

2 A 2 0 8 40 10 16 6

3 A 3 0 9 35 12 13 11

4 A 4 0 14 44 5 12 10

5 A 5 0 9 42 2 24 8

6 A 6 0 7 29 3 10 9

str(myschool)

'data.frame': 30 obs. of 8 variables:  
 $ School : Factor w/ 5 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...  
 $ 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 ...

summary(myschool)

School Section Very.Ahead..5 Middling..0 Behind..1.5   
 A:13 Min. : 1.00 Min. :0 Min. : 2.00 Min. : 4.00   
 B:12 1st Qu.: 2.25 1st Qu.:0 1st Qu.: 4.25 1st Qu.:15.25   
 C: 3 Median : 5.50 Median :0 Median : 7.50 Median :22.00   
 D: 1 Mean : 5.90 Mean :0 Mean : 7.40 Mean :25.13   
 E: 1 3rd Qu.: 9.00 3rd Qu.:0 3rd Qu.: 9.75 3rd Qu.:34.25   
 Max. :13.00 Max. :0 Max. :19.00 Max. :56.00   
 More.Behind..6.10 Very.Behind..11 Completed   
 Min. : 0.000 Min. : 0.000 Min. : 1.00   
 1st Qu.: 1.000 1st Qu.: 1.250 1st Qu.: 6.00   
 Median : 2.000 Median : 5.500 Median :10.00   
 Mean : 3.333 Mean : 6.967 Mean :10.53   
 3rd Qu.: 4.750 3rd Qu.:11.500 3rd Qu.:14.00   
 Max. :12.000 Max. :24.000 Max. :27.00

* Make the column names simpler for easy handling.

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

* Convert section to nominal variable for analysis -   
    
  myschool$section <- factor(myschool$section)
* Restructure the data set using tidyr, having one row per observation.

tidyrmyschool <- gather(myschool, key = c(veryAhead5, middling0, behind15, moreBehind610, veryBehind11, completed), value = numofStudents, c(-section, -school))  
  
*(FYI –* *Had a very bad experience with using gather with 3.6.1 version. Got failure for the above* *in office laptop but worked in home computer - because of the version dependency – later version of R doesn’t allow key =c(val1,2,3…) , value, c(-val1, val2).. as it treats key=c(…) as vector while expecting a string and throws error ”* [*Must supply a symbol or a string as argument*](https://stackoverflow.com/questions/57862583/spread-gather-error-must-supply-a-symbol-or-a-string-as-argument)*” . couldn’t find alternate solution, and despite of few uninstalls/reinstalls no luck, ended up using my corp- machine to complete the work. Horrible experience w/R*)

colnames( tidyrmyschool)[3] <- "studentByclass"  
tidyrmyschool$studentByclass <- factor( tidyrmyschool$studentByclass, levels = c("completed",   
 "veryAhead5", "middling0", "behind15", "moreBehind610", "veryBehind11"))

Now, let’s see how many students there are per school and, on average, how many students there are per class?

tidyrmyschool %>% group\_by(school) %>% summarize(numSections = max(as.numeric(section)),   
 numofStudents = sum(numofStudents))

# A tibble: 5 x 3  
 school numSections numofStudents  
 <fct> <dbl> <int>  
 1 A 13 932  
 2 B 12 446  
 3 C 3 85  
 4 D 1 22  
 5 E 1 116

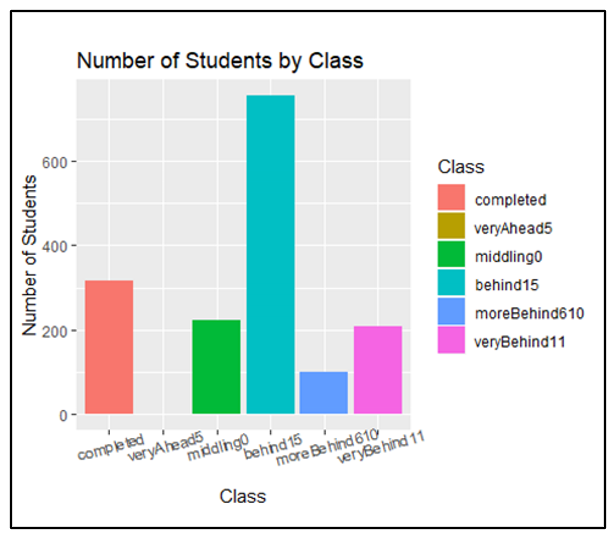
# With same number of sections, School-A has more than double the number of students than #School-B. How about number of students in each classes?  
 tidyrmyschool %>% group\_by(studentByclass) %>% summarise(numofStudents = sum(numofStudents))

# A tibble: 6 x 2  
 studentByclass numofStudents  
 <fct> <int>  
 1 completed 316  
 2 veryAhead5 0  
 3 middling0 222  
 4 behind15 754  
 5 moreBehind610 100  
 6 veryBehind11 209

# Majority of students (47.1 percent) are behind in their math lessons.

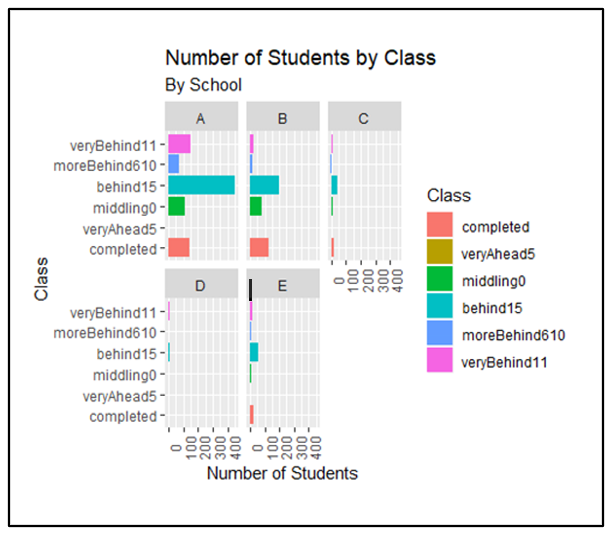
As of now, we assume that three-quarters of the way into the school year, the program is not effective as most students are behind in their courses. However, we can’t dismiss the fact that the second largest group (19.7 percent) of students have completed all their courses. Let’s start breaking down the data set using visualizations.

# Let's look at the distribution of students across the different classes.  
ggplot( tidyrmyschool, aes(x = studentByclass, y = numofStudents, fill = studentByclass)) +   
 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 = 15))



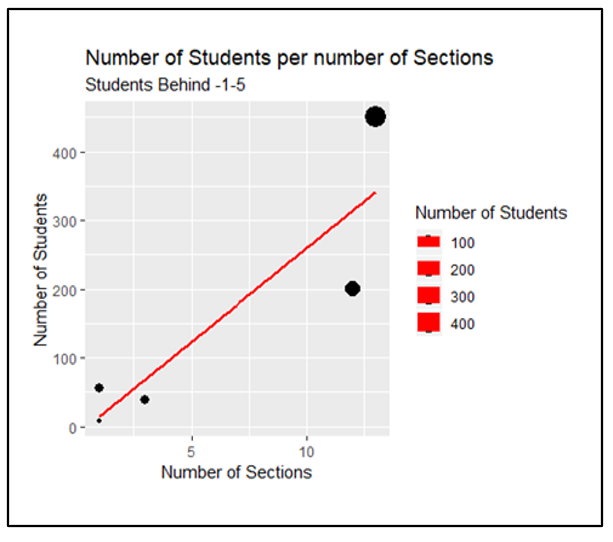
# Change is not prominent , number of students is significantly higher for those# in the Behind -1-5 group. Let's check if same distribution holds among# the five schools.

tidyrmyschool %>% group\_by(school, studentByclass) %>% summarise(numofStudents = sum(numofStudents)) %>%   
 ggplot(aes(x = studentByclass, y = numofStudents, fill = studentByclass)) + 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))

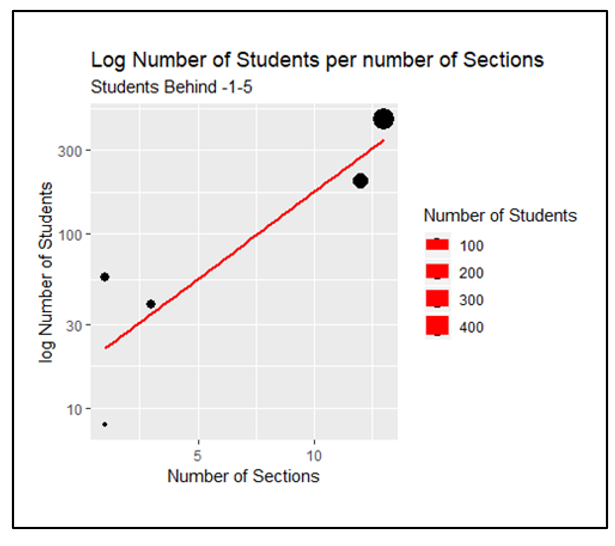


By comparing data by school, we notice interesting trends. School-A has highest number of students behind in their math courses. With 450 students behind by one to five courses (and 677 behind in total), the program doesn’t appear to be successful in School-A. School-A having a 72.6 behind rate, the rest of the schools have a 57.7 percent behind rate and it has a 15.2 completed rate, while the rest of schools have a 26.0 percent completed rate. This raises many follow-up questions. Is there a correlation between the number of sections and number of students behind? How are students distributed by section in School-A? Are there any outliers?

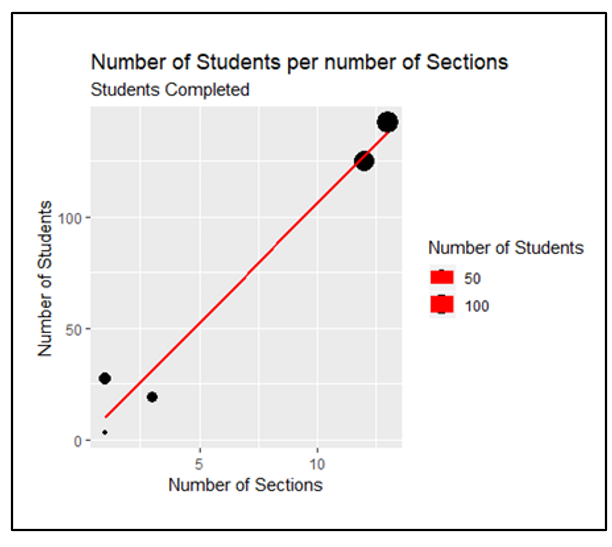
tidyrmyschool %>% filter(studentByclass == "behind15") %>% group\_by(school) %>% summarise(numSections = max(as.numeric(section)),   
 numofStudents = sum(numofStudents)) %>% ggplot(aes(x = numSections, y = numofStudents,   
 size = numofStudents)) + geom\_point() + geom\_smooth(method = "lm", colour = "red",   
 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")



# There seems to be an apparent relationship between the number of sections and number of #students (r-squeared = 0.796). But this is not clear from the graph. Let's convert the y-scale to #a logarithm.  
 tidyrmyschool %>% filter(studentByclass == "behind15") %>% group\_by(school) %>% summarise(numSections = max(as.numeric(section)),   
 numofStudents = sum(numofStudents)) %>% ggplot(aes(x = numSections, y = numofStudents,   
 size = numofStudents)) + geom\_point() + geom\_smooth(method = "lm", colour = "red",   
 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")



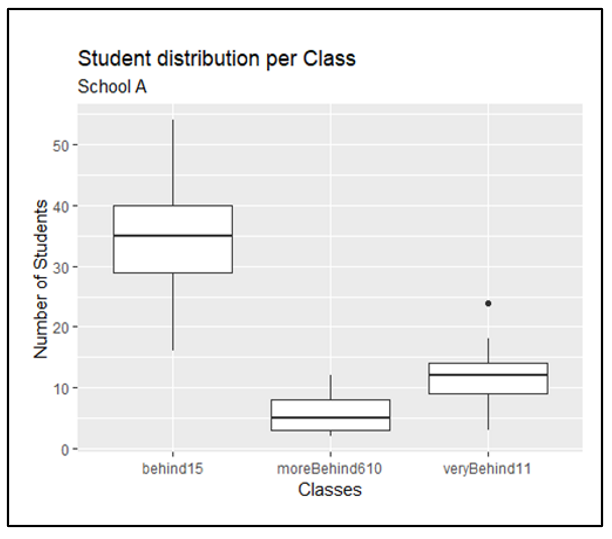
#There's much clearer relationship but data set is so small & not enough to conclude correlation# What about if we looked at the numbers for students who have completed# their math courses.  
 tidyrmyschool %>% filter(studentByclass == "completed") %>% group\_by(school) %>% summarise(numSections = max(as.numeric(section)),   
 numofStudents = sum(numofStudents)) %>% ggplot(aes(x = numSections, y = numofStudents,   
 size = numofStudents)) + geom\_point() + geom\_smooth(method = "lm", colour = "red",   
 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")



The number of sections does affect the number of students who either complete their courses or behind. Also more sections does not mean to influence more students to complete or lag in their courses.

Let’s look at the distribution for students behind for all sections of School-A.

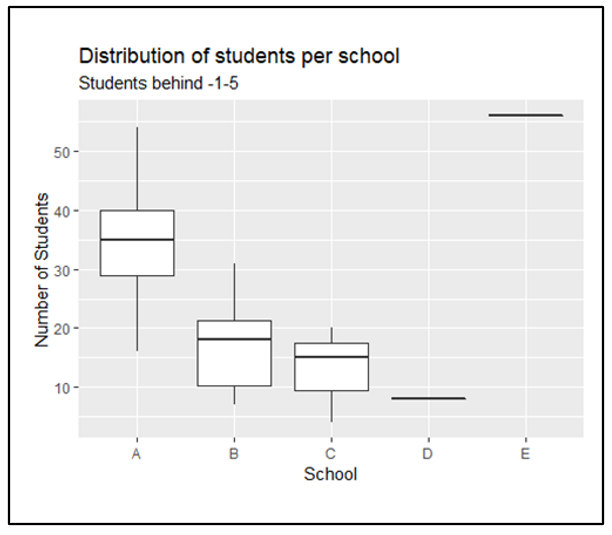
tidyrmyschool %>% filter(school == "A" & studentByclass %in% c("behind15", "moreBehind610",   
 "veryBehind11")) %>% ggplot(aes(x = studentByclass, y = numofStudents)) + geom\_boxplot() +   
 labs(x = "Classes", y = "Number of Students") + ggtitle("Student distribution per Class",   
 subtitle = "School A")



# No outliers driving School-A to have such high numbers of students lagging. However, largest# section has nearly 55 students lagging while the smallest section little over 15. With a median #of 35 students and an average of 34.6 students, this explains that the number of sections is #driving School-A's high numbers.

How do the numbers compare when looking at all schools?

tidyrmyschool %>% filter(studentByclass %in% c("behind15")) %>% ggplot(aes(x = school,   
 y = numofStudents)) + geom\_boxplot() + labs(x = "School", y = "Number of Students") +   
 ggtitle("Distribution of students per school", subtitle = "Students behind -1-5")



School-A is driving up the numbers but we don’t have clear evidence other than they have more students and sections, School-E has one section with 56 students lagging behind by 1 to 5 courses, but School-E has most number of students completed Math.