Martin\_Alonso\_Wk2

Martin Alonso

October 14, 2018

# HW2 Instructions

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

require(dplyr)

## Loading required package: 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

require(ggplot2)

## Loading required package: ggplot2

require(tidyr)

## Loading required package: tidyr

# First, load and explore the data set.  
dat <- read.csv("Data Storyteller.csv")  
  
head(dat)

## 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(dat)

## '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(dat)

## 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

To simplify the data set, we’ll change the column names. We also want to change ‘section’ into an nominal variable.

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

Now we’ll restructure the data set using tidyr principles, having one row per observation.

tidyDat <- gather(dat, key = c(veryAhead5, middling0, behind15, moreBehind610,   
 veryBehind11, completed), value = numStudents, c(-section, -school))  
colnames(tidyDat)[3] <- "studentClass"  
tidyDat$studentClass <- factor(tidyDat$studentClass, levels = c("completed",   
 "veryAhead5", "middling0", "behind15", "moreBehind610", "veryBehind11"))

First thing we want to know is how many students there are per school and, on average, hov many students there are per class.

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

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

# Despite having almost the same number of sections, school A has more than  
# double the number of students that school B. How about number of students  
# per class?  
tidyDat %>% group\_by(studentClass) %>% summarise(numStudents = sum(numStudents))

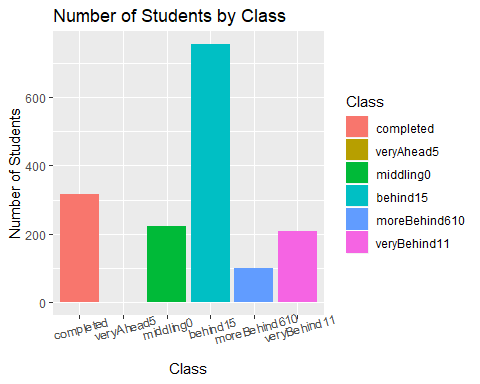
## # A tibble: 6 x 2  
## studentClass numStudents  
## <fct> <int>  
## 1 completed 316  
## 2 veryAhead5 0  
## 3 middling0 222  
## 4 behind15 754  
## 5 moreBehind610 100  
## 6 veryBehind11 209

# We know then, that the majority of students (47.1 percent) are behind in  
# their maths lessons.

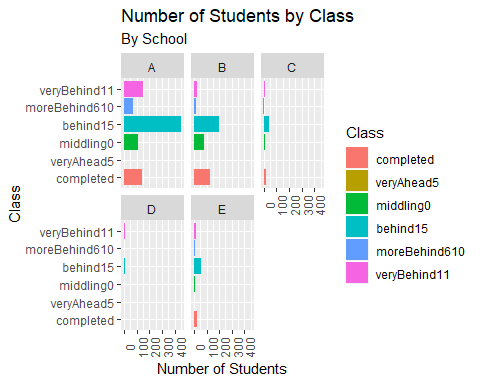
So far, we can assume that three-quarters of the way into the school year, the program is not really effective as the vast majority of students are behind in their courses. However, we can’t dismiss the fact that the second largest group (19.7 percent) of students have completed all their courses.

Let’s start breaking down the data set using visualizations.

# Let's look at the distribution of students across the different classes.  
ggplot(tidyDat, aes(x = studentClass, y = numStudents, fill = studentClass)) +   
 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))



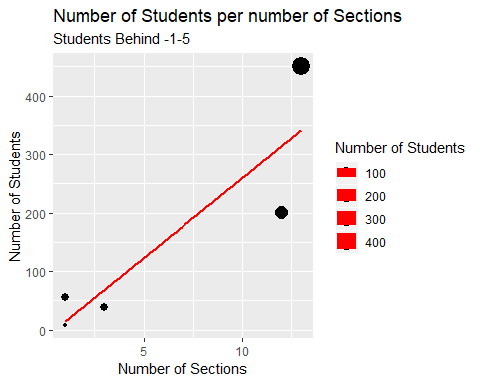
# Not much has changed, it's the same conclusion we've gotten so far.  
  
# We can see that the number of students is significantly higher for those  
# in the Behind -1-5 group. Let's see if the same distribution holds among  
# the five schools.  
tidyDat %>% group\_by(school, studentClass) %>% summarise(numStudents = sum(numStudents)) %>%   
 ggplot(aes(x = studentClass, y = numStudents, fill = studentClass)) + 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))



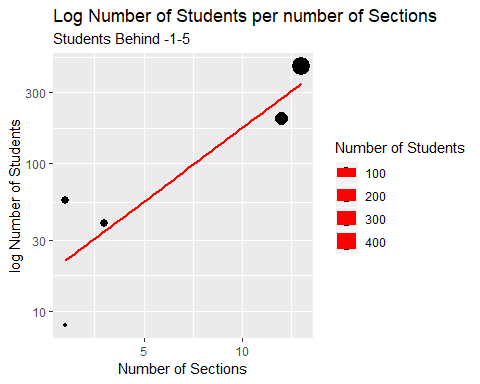
Once we open the data up by school, we start finding much more interesting trends. The first thing that pops out is school A’s large number of students who are behind in their maths courses. With 450 students behind by one to five courses (and 677 behind in total), the program doesn’t appear to be succesful in school A. But, that narrative changes when we turn our attention to the other schools. Were school A has a 72.6 behind rate, the rest of the schools have a 57.7 percent behind rate. Furthermore, school A has a 15.2 completed rate, while the rest of schools have a 26.0 percent completed rate. This seems to point that the problem is not the program in and of itself, but to a problem with school A!

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?

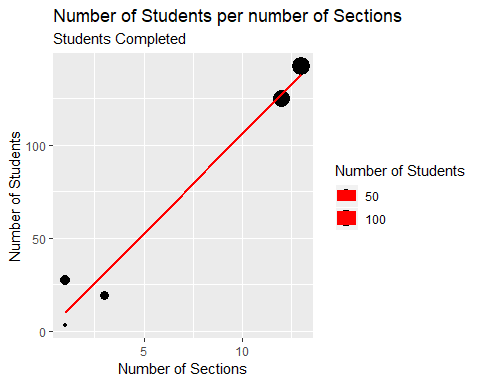
tidyDat %>% filter(studentClass == "behind15") %>% group\_by(school) %>% summarise(numSections = max(as.numeric(section)),   
 numStudents = sum(numStudents)) %>% ggplot(aes(x = numSections, y = numStudents,   
 size = numStudents)) + geom\_point() + geom\_smooth(method = "lm", colour = "red",   
 se = 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 that clear  
# from the graph. Let's convert the y-scale to a logarithm.  
tidyDat %>% filter(studentClass == "behind15") %>% group\_by(school) %>% summarise(numSections = max(as.numeric(section)),   
 numStudents = sum(numStudents)) %>% ggplot(aes(x = numSections, y = numStudents,   
 size = numStudents)) + 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")

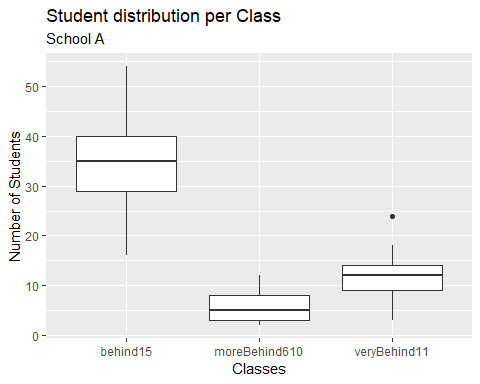


# There's a much clearer relationship but the data set is so small that it  
# may not be wise to conclude that these two variables are correlated.  
  
# What about if we looked at the numbers for students who have completed  
# their maths courses.  
tidyDat %>% filter(studentClass == "completed") %>% group\_by(school) %>% summarise(numSections = max(as.numeric(section)),   
 numStudents = sum(numStudents)) %>% ggplot(aes(x = numSections, y = numStudents,   
 size = numStudents)) + 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")

 So the number of sections does affect the number of students who either complete their courses or are behind. Then again, just because there are more sections does not mean that this causes more students to complete or lag in their courses.

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

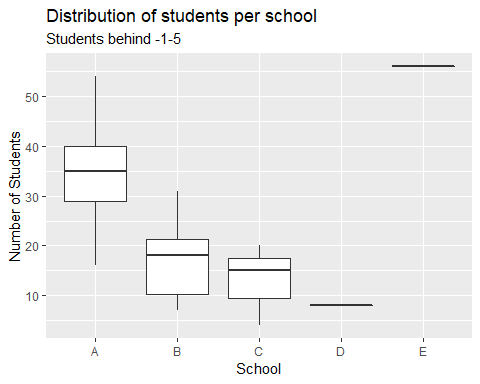
tidyDat %>% filter(school == "A" & studentClass %in% c("behind15", "moreBehind610",   
 "veryBehind11")) %>% ggplot(aes(x = studentClass, y = numStudents)) + geom\_boxplot() +   
 labs(x = "Classes", y = "Number of Students") + ggtitle("Student distribution per Class",   
 subtitle = "School A")



# There seem to be no outliers that are driving School A to have such high  
# numbers of students lagging behind. However, we can say that the largest  
# section has nearly 55 students lagging behind while the smallest section  
# has a bit 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?

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



So, there’s no real explanation why school A is driving up the numbers other than they have more students and sections - something evident by the fact that school E has one section with 56 students lagging behind by 1 to 5 courses. Other than that, there are no outliers nor anything that jumps out when trying to explain why class A has more students lagging above the average.