HW2\_Alex Hyman

Alex Hyman

10/8/2018

### Data

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.

#### Checking the data

First, I need to read the data into R and check for any data quality issues. This would include missing data and unnecessary data.

#read the data  
data <- read.csv("data-storyteller.csv")  
#rename the columns  
column.names <- c("school", "section", "ahead", "middling", "behind", "more.behind", "very.behind", "completed")  
names(data) <- column.names  
#Re-arranging data  
data <- data[,c("school", "section", "very.behind", "more.behind", "behind", "middling", "completed", "ahead")]  
#Inspect the data  
summary(data)

## school section very.behind more.behind behind   
## A:13 Min. : 1.00 Min. : 0.000 Min. : 0.000 Min. : 4.00   
## B:12 1st Qu.: 2.25 1st Qu.: 1.250 1st Qu.: 1.000 1st Qu.:15.25   
## C: 3 Median : 5.50 Median : 5.500 Median : 2.000 Median :22.00   
## D: 1 Mean : 5.90 Mean : 6.967 Mean : 3.333 Mean :25.13   
## E: 1 3rd Qu.: 9.00 3rd Qu.:11.500 3rd Qu.: 4.750 3rd Qu.:34.25   
## Max. :13.00 Max. :24.000 Max. :12.000 Max. :56.00   
## middling completed ahead   
## Min. : 2.00 Min. : 1.00 Min. :0   
## 1st Qu.: 4.25 1st Qu.: 6.00 1st Qu.:0   
## Median : 7.50 Median :10.00 Median :0   
## Mean : 7.40 Mean :10.53 Mean :0   
## 3rd Qu.: 9.75 3rd Qu.:14.00 3rd Qu.:0   
## Max. :19.00 Max. :27.00 Max. :0

#dropping the ahead column because there are no instances  
data <- data %>% select(-ahead)  
#Checking for missing data  
cat("\nThere are ", sum(!complete.cases(data)), " sections missing data", sep = "")

##   
## There are 0 sections missing data

Because there were not any instances of sections that are “Ahead” of the schedule, the column was removed from the dataset. We should also take note that 25 of the 30 sections are from shools A and B and no sections are missing data.

#### Exploring the data

First, I would like to see how the students are doing in each of the schools. This will provide a simple overview of the distribution of student status for each of the schools. This does not consider the differences in the number of students at each of the schools.

#function to calculate the row percentage  
rowPercentage <- function(df, school = F, schoolCol = NULL, section = F, sectionCol = NULL ){  
 df\_temp <- df %>%   
 select(c("very.behind", "more.behind", "behind", "middling", "completed"))  
 totals <- rowSums(df\_temp)  
 df\_temp <- round(df\_temp / totals \* 100, 2)  
 if (school){  
 df\_temp <- bind\_cols(df[schoolCol], df\_temp)  
 }  
 if (section){  
 df\_temp <- bind\_cols(df[sectionCol], df\_temp)  
 }   
 return(df\_temp)  
}  
  
#compare the totals for the levels  
dataGrouped <- data %>% group\_by(school) %>%  
 select(-section) %>% summarise\_all(sum) %>%  
 rowPercentage(school = T, schoolCol = "school")  
   
knitr::kable(dataGrouped)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| school | very.behind | more.behind | behind | middling | completed |
| A | 16.52 | 7.83 | 48.28 | 12.12 | 15.24 |
| B | 4.93 | 3.14 | 45.07 | 18.83 | 28.03 |
| C | 14.12 | 4.71 | 45.88 | 12.94 | 22.35 |
| D | 27.27 | 9.09 | 36.36 | 13.64 | 13.64 |
| E | 12.93 | 6.03 | 48.28 | 9.48 | 23.28 |

Just on an aggregate level by school, we can see that School D has the largest percentage of children in their school that are very behind in the class, and the school that has the lowest percentage of students that have completed the course.

Additionally, we would like to see how all of the students are spread out by class and by school. The table below is the percentage of students within the course status at that school compared to the entire student population taking the course at all schools

data\_summed <- data %>% group\_by(school) %>% select(-section) %>% summarise\_all(sum) %>%  
 gather("status", "count", -school) %>% mutate(status = factor(status, levels =   
 c("very.behind", "more.behind", "behind", "middling", "completed"), ordered = T))  
  
knitr::kable(round(xtabs(count ~ ., data = data\_summed) / sum(data\_summed$count) \* 100, 2))

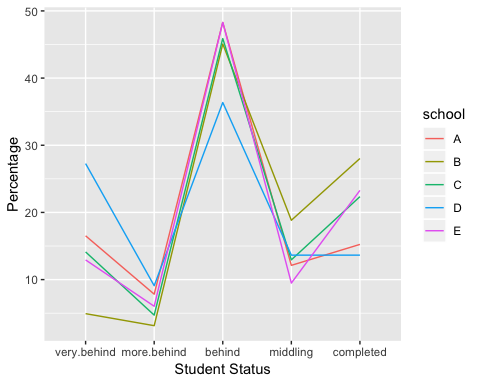
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | very.behind | more.behind | behind | middling | completed |
| A | 9.62 | 4.56 | 28.11 | 7.06 | 8.87 |
| B | 1.37 | 0.87 | 12.55 | 5.25 | 7.81 |
| C | 0.75 | 0.25 | 2.44 | 0.69 | 1.19 |
| D | 0.37 | 0.12 | 0.50 | 0.19 | 0.19 |
| E | 0.94 | 0.44 | 3.50 | 0.69 | 1.69 |

The table shows that school A has the most students and that the largest proprtion of studeint in the cross tabulation are considered to be “behind” and taking the course at school A.

#### Visualizing Schools

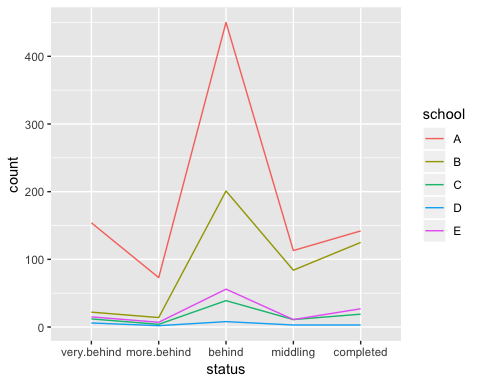
Now we will visualize the percentages for each of the schools. This will essentiall replicate the first table, but be more visually pleasing.

#Attempt to plot each of the schools  
new\_data <- dataGrouped %>% gather("status", "percent", -school) %>%  
 mutate(status = factor(status, levels =   
 c("very.behind", "more.behind", "behind", "middling", "completed"), ordered = T))  
  
#plotting the percentages for each school  
p <- ggplot(new\_data, aes(x = status, y = percent, group = school)) + geom\_line(aes(col = school))  
p <- p + xlab("Student Status") + ylab("Percentage")  
p



As determined earlier, the largest proportion of students are classified as being “behind” in the class at each of the schools. Also, each of the schools follow the general pattern of having a the most students classified as “behind” and then having a dramitic drop in the percentages on the moderate ends of the distribution (more behind and middling), and finally a decently large jump to the extremes of the tails (completed and very behind). To get an idea of the count of students in each category at the school we will create the same graphic, but with the number of students.

data %>% select(-section) %>% group\_by(school) %>%  
 summarise\_all(sum) %>% gather("status", "count", -school) %>%  
 mutate(status = factor(status, levels = c("very.behind", "more.behind",   
 "behind", "middling", "completed"), ordered = T)) %>%   
 ggplot(aes(x = status, y = count, group = school, col = school)) +  
 geom\_line()

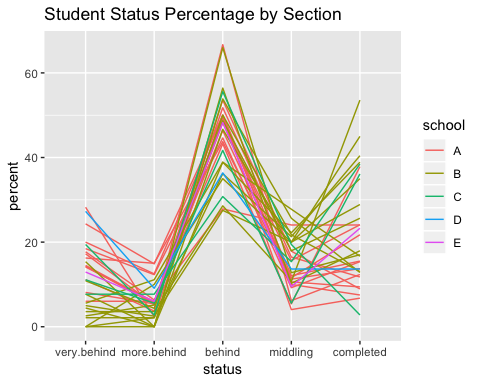


While the shapes are fairly similar, it is apparent that most of the students that are behind in the course are coming from School A (as determined by the second table). If a solution could be found for students in the math course at school A, then most of the problem should be resolved.

#### Visualizing Sections

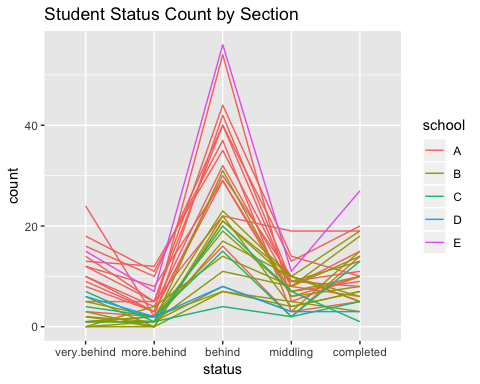
While we determined the pattern of distribution of the students on a school wide basis, we would also want to check if there are any sections that are doing extraordinarily well, or completely slacking. First, we will visualize the distribution for each of the sections as a percentage.

#Getting percentage of each section and keeping the school and section name  
data %>% rowPercentage(school = T, schoolCol = "school", section = T, sectionCol = "section") %>%  
 gather("status", "percent", c(-school, -section)) %>% mutate(status = factor(status, levels =   
 c("very.behind", "more.behind", "behind", "middling", "completed"), ordered = T),   
 school\_section = paste(school, "-", section, sep = "")) %>%   
 ggplot(aes(x = status, y = percent)) + geom\_line(aes(group = school\_section, color = school)) +  
 ggtitle("Student Status Percentage by Section")



The above plot shows that each of the ections follow generally the same trend of having most students in the “behind” category and decreasing in either direction of “behind”. Then on the tails, the students are generally at a higher percentage than the in-between category. Now we will visualize the count of the students in each category in each of the sections.

data %>%   
 gather("status", "count", c(-school, -section)) %>% mutate(status = factor(status, levels =   
 c("very.behind", "more.behind", "behind", "middling", "completed"), ordered = T),   
 school\_section = paste(school, "-", section, sep = "")) %>% ggplot(aes(x = status,   
 y = count, group = school\_section, color = school)) + geom\_line() +  
 ggtitle("Student Status Count by Section")

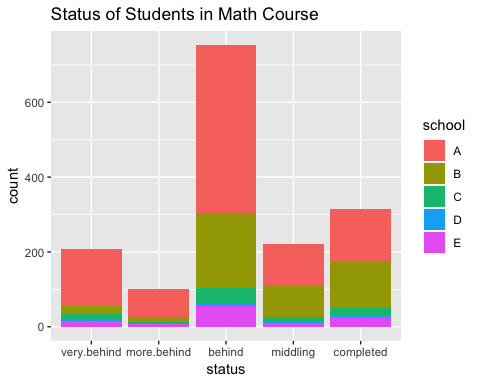


It appears that the only section offered at school E is a fairly large section. We can also see that the sections at school B are generally smaller than the sections at school A. While the sections are fairly similar in their distribution, we can see that on section at school A has figured out a way to have fewer students categorized as “behind”, and having a fairly large count of students in the “middling” or “completed” category.

#### Student level

The school average on the student level gives an important perspective on how the schools are performing on a class-wide basis, but it is important to get a better idea of the spread of students. If half of the class is completing the course and the other half is very behind, we would be failing half of the students in the school. Below will be a breakdown of how many students are in which category, colored by the school.

#tidying by school and status  
school\_df <- data %>% select(-section) %>% gather("status", "count", -school) %>%  
 mutate(status = factor(status, levels = c("very.behind", 'more.behind',   
 'behind', 'middling', 'completed'), ordered = T))  
  
#Creating a plot that has the count in each level by school  
p.count <- ggplot(school\_df, aes(x = status, y = count)) + geom\_col(aes(fill = school))  
p.count <- p.count + ggtitle("Status of Students in Math Course")  
p.count

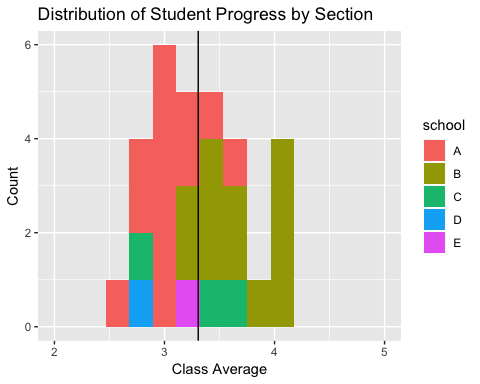


#### Adjusting sections size

The data was inspected on a school-wide level in the last section, but it could prove some more insight if the data was inspected on a sectional basis. This will be accomplished by multiplying each of the columns by 1, 2, 3, 4, or 5, with 1 being mapped for ‘very behind’ students and 5 being mapped to completed students. These values will be averaged by the count of students to determine what the average status is for each of the sections.

#Getting class average for student profress  
adj\_data <- data %>% mutate(totalStudents = rowSums(.[3:7])) %>%  
 mutate(more.behind = more.behind \* 2, behind = behind \* 3,  
 middling = middling \* 4, completed = completed \* 5) %>%  
 mutate(avg = rowSums(.[3:7] / totalStudents)) %>%  
 select(c(school, section, avg, totalStudents))  
  
#plotting a histogram of the data  
p1 <- ggplot(adj\_data, aes(avg, fill = school)) + geom\_histogram(bins = 15) + xlim(2, 5)  
p1 <- p1 + geom\_vline(xintercept = mean(adj\_data$avg)) + ylab("Count") + xlab("Class Average")  
p1 <- p1 + ggtitle("Distribution of Student Progress by Section")  
p1

## Warning: Removed 5 rows containing missing values (geom\_bar).



The above histogram provides some additional insight into general performance by school, as well as performance on a sectional basis. School B appears to be the best performing school, with all but two sections performing above the mean. School A generally underperforms, except for two classes. School D’s one section is not performing well at all, and School E’s section performs just about on the average.

To see which sections are the most behind, let’s look at the five lowest performing section on average.

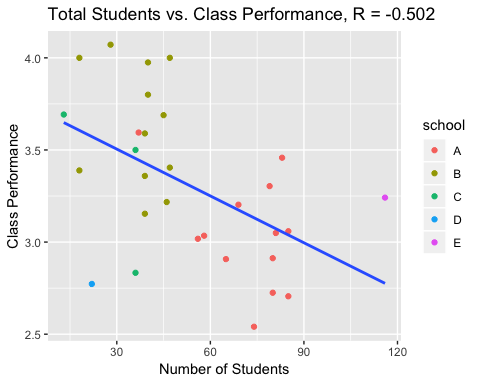
head(adj\_data[order(adj\_data$avg), ], 5)

## school section avg totalStudents  
## 8 A 8 2.540541 74  
## 5 A 5 2.705882 85  
## 2 A 2 2.725000 80  
## 29 D 1 2.772727 22  
## 27 C 2 2.833333 36

#### Correlation between class size and performance

Single attribute analysis is interesting, but we would also like to see if there are any relationships between two variables. A correlation and plot between class size and section performance could inform us if there is a relationship between the two variables.

plot.title <- paste("Total Students vs. Class Performance, R = ", round(cor(adj\_data$total, adj\_data$avg), 3), sep = "")  
p2 <- ggplot(adj\_data, aes(x = totalStudents, y = avg)) + geom\_point(aes(color = school))  
p2 <- p2 + ggtitle("Number of Students vs. Performance") + xlab("Number of Students")  
p2 <- p2 + ylab("Class Performance") + geom\_smooth(method = 'lm', se = F)  
p2 <- p2 + ggtitle(plot.title)  
p2



The plot shows that there a negative correlation between performance and the number of students. As the number of students in a section increases, the performance of the class generally decreases. The performance of section from School E seems to be an outlier in the data, which seems to be hurting our overall correlation of the tewo variables.

### Summary

As the semester is ending, each of the math courses should be wrapping up. While the evaluation categories were appropriate for a majority of the school year, sections can no longer be categorized as having students ahead in their class; these students would now be classified as being complete with the course.

The five different schools also varied widely in the number of sections that offered the course, and the number of students taking the course in each section. Of the 30 different sections, 25 of them came from either school A or school B. Of the remaining three schools, the fewest students took the course from school D. Additionally, the section that had the most students taking the course came from the only section offered at school E. The school that had the most students take the course was school A.

On a performance basis, school B generally performed the best; they had the smallest percentage of students categorized as “very behind” and had the largest percentage of students categorized as “completed”. On an analysis of the extremes, school D as a whole performed the worst as they had the largest percentage of students “very behind” and the smallest percentage of students “completed”.

Overall, most students taking the course were categorized as being “behind” and the fewest number of students were categorized as “more behind”. On the extremes, there were more students that had completed the course than determined to be “very behind”. For all of the categories, school A represented the largest proportion of students, except for the “completed” category. School B had the most students that had completed the course.

On a section-wide analysis, it becomes more evident that school B is the highest performing school as they had all but two of their twelve sections performing above the mean of all class performances, including the five top performing classes. School A had three worst performing sections (sections 8, 5, 2), followed by the lone section from school D, and the section 2 from school C. It is noteworhty that the three worse performing sections from school A had more than twice the students in the course than the sections from schools C and D.

In an attempt to find a relationship between the number of students in a class and the class performance, a scatter plot was created and a correlation was cacluated. While there are not enough instances to have a statistically significan conclusion, there does appear to be a negative correlation between performance and class size. The larger the class size, the worse a class performs. However, this tend could be due to numerous other variables, including school SAT, acceptance rate, and the type of school (engineering, liberal arts, etc.).