# Math Course Data Story

**Introduction**

The purpose of the report is to conduct exploratory data analysis regarding where students are in completing a math course.

There are no immediate or specific questions to be addressed beyond what story does this data tell.

The following information has been provided regarding the math course being analyzed.

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)

Additional information given about the case suggests that the progress of each student is self-reported by the professor and could be a subjective rather than an exact answer.

**Analysis and Models**

**About the Data**

It is a simple and small data set, and little data cleaning is needed.

**Reading in the data.**

#replacing any possible blank entries as 'NA'.  
  
storyteller <- read\_csv("C:/Users/amcvicar/OneDrive/School/Syracuse/707 Data Analytics/week2\_resources\_2\_2/data-storyteller.csv", na = c(""))

## Parsed with column specification:  
## cols(  
## School = col\_character(),  
## Section = col\_double(),  
## `Very Ahead +5` = col\_double(),  
## `Middling +0` = col\_double(),  
## `Behind -1-5` = col\_double(),  
## `More Behind -6-10` = col\_double(),  
## `Very Behind -11` = col\_double(),  
## Completed = col\_double()  
## )

#checking data types to see what may need changing  
str(storyteller)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 30 obs. of 8 variables:  
## $ School : chr "A" "A" "A" "A" ...  
## $ Section : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Very Ahead +5 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Middling +0 : num 5 8 9 14 9 7 19 3 6 13 ...  
## $ Behind -1-5 : num 54 40 35 44 42 29 22 37 29 40 ...  
## $ More Behind -6-10: num 3 10 12 5 2 3 5 11 8 5 ...  
## $ Very Behind -11 : num 9 16 13 12 24 10 14 18 12 5 ...  
## $ Completed : num 10 6 11 10 8 9 19 5 10 20 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. School = col\_character(),  
## .. Section = col\_double(),  
## .. `Very Ahead +5` = col\_double(),  
## .. `Middling +0` = col\_double(),  
## .. `Behind -1-5` = col\_double(),  
## .. `More Behind -6-10` = col\_double(),  
## .. `Very Behind -11` = col\_double(),  
## .. Completed = col\_double()  
## .. )

**Data Cleaning**

The data cleaning and reformatting steps are as follows:

1. The school column is of the character type is transformed into a factor.
2. The section column is a numeric type and is transformed into a factor.
3. The progress statuses are given as columns with a count of students in each as the data value per school and section.
4. These student counts are updated from a number to integer data type.

storyteller$School <- factor(storyteller$School)  
  
#The section column is of the Numeric type and should be a factor instead  
  
storyteller$Section <- factor(storyteller$Section)  
  
#Each of the remaining columns is a discrete count of the students in each category.  
#As it is not continuous the columns 'Very Ahead', 'Middling', 'Behind', 'More behind', 'Very behind'  
#and 'Completed' should all be integers.  
  
storyteller$`Very Ahead +5` <- as.integer(storyteller$`Very Ahead +5`)  
storyteller$`Middling +0` <- as.integer(storyteller$`Middling +0`)  
storyteller$`Behind -1-5` <- as.integer(storyteller$`Behind -1-5`)  
storyteller$`More Behind -6-10` <- as.integer(storyteller$`More Behind -6-10`)  
storyteller$`Very Behind -11` <- as.integer(storyteller$`Very Behind -11`)  
storyteller$Completed <- as.integer(storyteller$Completed)

**Organizing the Data Structure**

The data structure has been updated to be more user-friendly. The rows are reordered to assume we would be most interested in seeing progress by section before school and reordering the progress statuses to a consistent completed to very behind. This reordering will allow us to visually compare sections across schools and to each other as well as follow a more natural status order.

#Reordering columns to get a cleaner picture. I.E. 'Completed' being shifted to the other side.  
#And section being a unique identifier is moved to the leftmost column.  
  
storytellerTemp <- storyteller[,c(2,1,8,3,4,5,6,7)]  
  
storyteller <- storytellerTemp  
   
#displaying top 5 rows  
head(storyteller)

## # A tibble: 6 x 8  
## Section School Completed `Very Ahead +5` `Middling +0` `Behind -1-5`  
## <fct> <fct> <int> <int> <int> <int>  
## 1 1 A 10 0 5 54  
## 2 2 A 6 0 8 40  
## 3 3 A 11 0 9 35  
## 4 4 A 10 0 14 44  
## 5 5 A 8 0 9 42  
## 6 6 A 9 0 7 29  
## # ... with 2 more variables: `More Behind -6-10` <int>, `Very Behind -11` <int>

**Missing Data**

There are not missing data values.

#Checking for any NA values  
sum(is.na(storyteller))

## [1] 0

#There are no NA values in this dataset.  
  
#The dataset is cleaned  
head(storyteller)

## # A tibble: 6 x 8  
## Section School Completed `Very Ahead +5` `Middling +0` `Behind -1-5`  
## <fct> <fct> <int> <int> <int> <int>  
## 1 1 A 10 0 5 54  
## 2 2 A 6 0 8 40  
## 3 3 A 11 0 9 35  
## 4 4 A 10 0 14 44  
## 5 5 A 8 0 9 42  
## 6 6 A 9 0 7 29  
## # ... with 2 more variables: `More Behind -6-10` <int>, `Very Behind -11` <int>

The following observation was provided with the dataset:

Observation: To be considered ‘ahead’ in any way, there are two categories: ‘Very Ahead’ and ‘Completed’. There are 3 categories in place to describe ‘behind’, which may lead the responses to be lopsided in favor of generally ‘behind’ over ‘generally ahead’.

Additionally, I would make the following observations:

‘Middling’ could be confusing wording, a description of ‘On Track’ may have been conducive to accurate reporting.

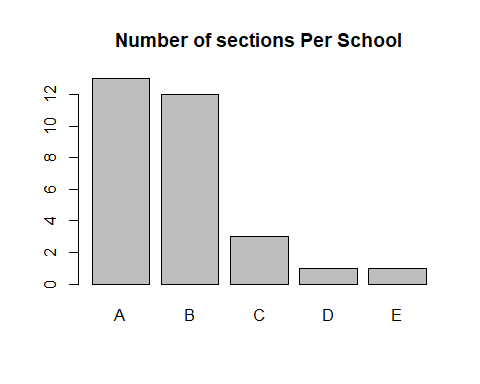
Middling as a segment of on track or 5 ahead suggests a question regarding expectations that may cloud understanding where students are with their studies, are the students expected to be on a specific lesson or are they supposed to be ahead?

There is not an equal segment to completed, such as not started.

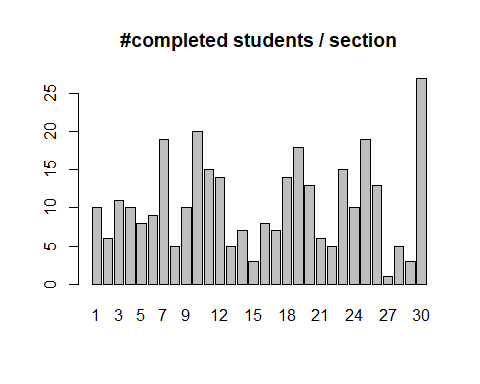
3/4 of the way through a 30 lesson course would indicate that students should be on lesson 22 or 23. The middling segment would be from 22 - 28. The very ahead segment would only be able to apply to students on lessons 27, 28, 29, & 30, assuming it is only the last three lessons, the data is skewed.

Additionally, the potential count of very behind changes as the course progresses. Does being 11 or more sessions behind mean failure, and also signify a milestone at session 11?

# Creating a bar chart to show the number of sections from each school  
SchoolValues <- c(length(which(storyteller$School=='A')), length(which(storyteller$School == 'B')), length(which(storyteller$School == 'C')), length(which(storyteller$School == 'D')), length(which(storyteller$School == 'E')))  
  
barplot(SchoolValues, names.arg = c('A', 'B', 'C', 'D', 'E'), main = 'Number of sections Per School')



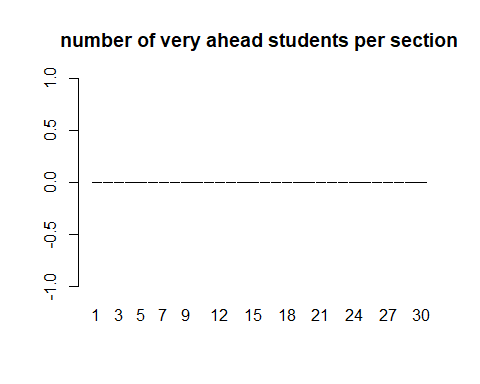
#plotting section and Completed and summarizing the data  
  
barplot(storyteller$Completed, main = '#completed students / section', names.arg = c(1:30))



summary(storyteller$Completed)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 6.00 10.00 10.53 14.00 27.00

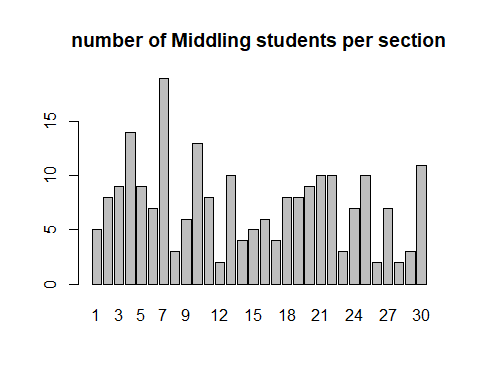
#Plotting section and Very Ahead and summarizing  
  
barplot(storyteller$`Very Ahead +5`, main = 'number of very ahead students per section', names.arg = c(1:30))



summary(storyteller$`Very Ahead +5`)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 0 0 0

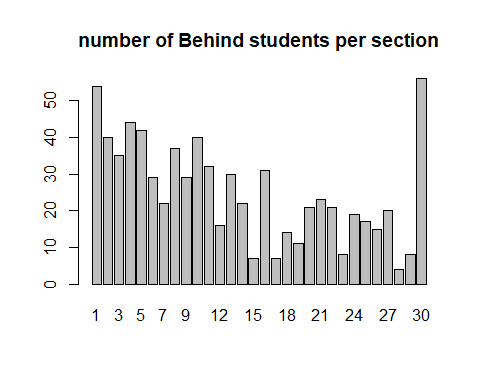
#plotting section and Middling and summarizing  
  
barplot(storyteller$`Middling +0`, main = 'number of Middling students per section', names.arg = c(1:30))



summary(storyteller$`Middling +0`)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.00 4.25 7.50 7.40 9.75 19.00

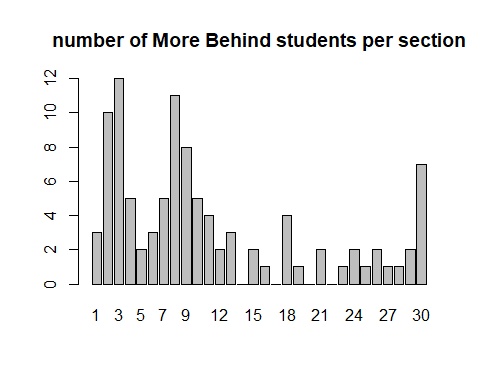
#plotting section and Behind and summarizing  
  
barplot(storyteller$`Behind -1-5`, main = 'number of Behind students per section', names.arg = c(1:30))



summary(storyteller$`Behind -1-5`)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.00 15.25 22.00 25.13 34.25 56.00

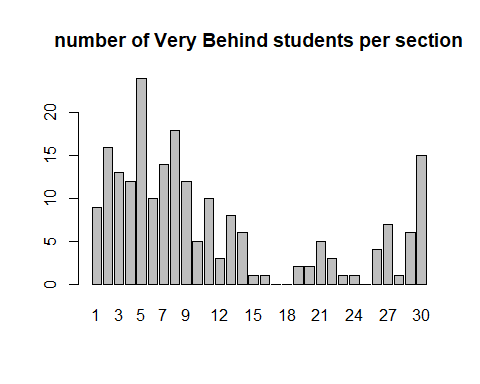
#plotting section and More Behind  
  
barplot(storyteller$`More Behind -6-10`, main = 'number of More Behind students per section', names.arg = c(1:30))



summary(storyteller$`More Behind -6-10`)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 2.000 3.333 4.750 12.000

#Plotting section and Very Behind  
  
barplot(storyteller$`Very Behind -11`, main = 'number of Very Behind students per section', names.arg = c(1:30))



summary(storyteller$`Very Behind -11`)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.250 5.500 6.967 11.500 24.000

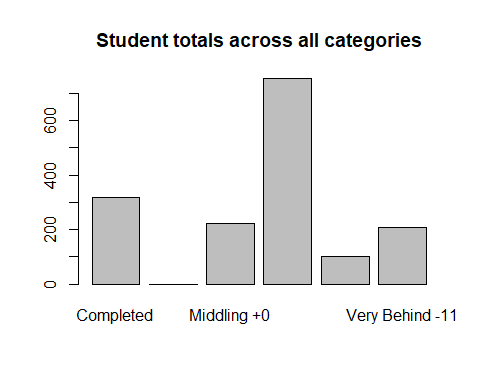
#determining the number of students in each category  
StudentSums <- colSums(storyteller[,3:8])  
  
sum(StudentSums)

## [1] 1601

#determining the amount of students in each section  
  
SectionSums <- rowSums(storyteller[,3:8])  
  
data.frame(SectionSums)

## SectionSums  
## 1 81  
## 2 80  
## 3 80  
## 4 85  
## 5 85  
## 6 58  
## 7 79  
## 8 74  
## 9 65  
## 10 83  
## 11 69  
## 12 37  
## 13 56  
## 14 39  
## 15 18  
## 16 47  
## 17 18  
## 18 40  
## 19 40  
## 20 45  
## 21 46  
## 22 39  
## 23 28  
## 24 39  
## 25 47  
## 26 36  
## 27 36  
## 28 13  
## 29 22  
## 30 116

#Creating a barplot to show distribution  
  
StudentSums <- colSums(storyteller[,3:8])  
barplot(StudentSums, main = "Student totals across all categories")



The data set came with the following observations noted:

- 14% of students are on track in the middling category.

- Nearly 20% of students have completed the program.

- Nearly 50% of students in this program are in the ‘behind’ category alone

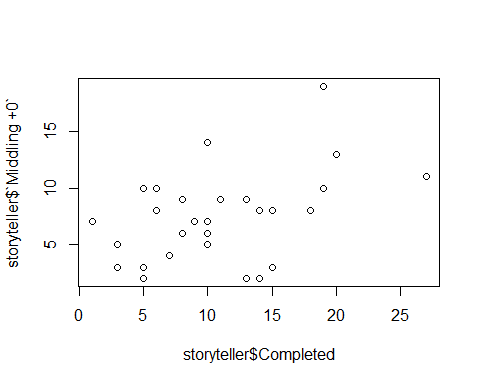
. - Students in the bottom two categories make up 20% of the sample. Meaning ~70% of students are behind in the curriculum.

This data suggests that understanding how many failures to start, false starts, and what lesson each segment is on are critical missing pieces of information

StudentSums/sum(StudentSums)

## Completed Very Ahead +5 Middling +0 Behind -1-5   
## 0.19737664 0.00000000 0.13866334 0.47095565   
## More Behind -6-10 Very Behind -11   
## 0.06246096 0.13054341

plot(storyteller$Completed, storyteller$`Middling +0`)

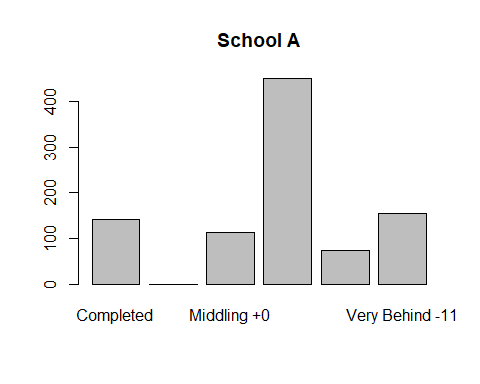


The progress pattern is consistent between schools. The extreme behind group needs to be further segmented to eliminate or understand the group better, and if they are trying or are false starts.

storytellerA <- storyteller[which(storyteller$School == "A"),]  
storytellerB <- storyteller[which(storyteller$School == "B"),]  
storytellerC <- storyteller[which(storyteller$School == "C"),]  
storytellerD <- storyteller[which(storyteller$School == "D"),]  
storytellerE <- storyteller[which(storyteller$School == "E"),]  
  
StudentSumsA <- colSums(storytellerA[3:8])  
StudentSumsA

## Completed Very Ahead +5 Middling +0 Behind -1-5   
## 142 0 113 450   
## More Behind -6-10 Very Behind -11   
## 73 154

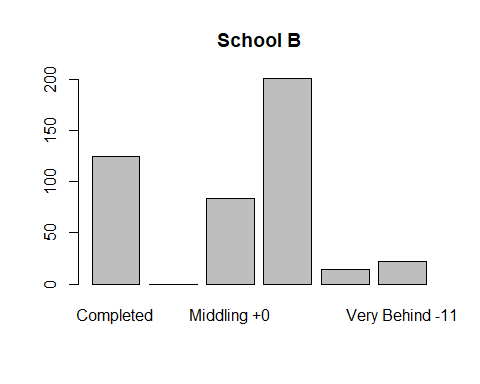
barplot(StudentSumsA, main = "School A")



StudentSumsB <- colSums(storytellerB[3:8])  
StudentSumsB

## Completed Very Ahead +5 Middling +0 Behind -1-5   
## 125 0 84 201   
## More Behind -6-10 Very Behind -11   
## 14 22

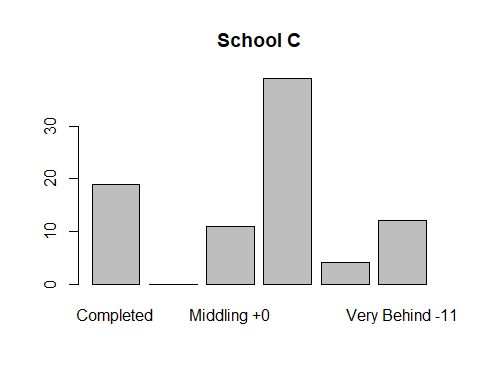
barplot(StudentSumsB, main = "School B")



StudentSumsC <- colSums(storytellerC[3:8])  
StudentSumsC

## Completed Very Ahead +5 Middling +0 Behind -1-5   
## 19 0 11 39   
## More Behind -6-10 Very Behind -11   
## 4 12

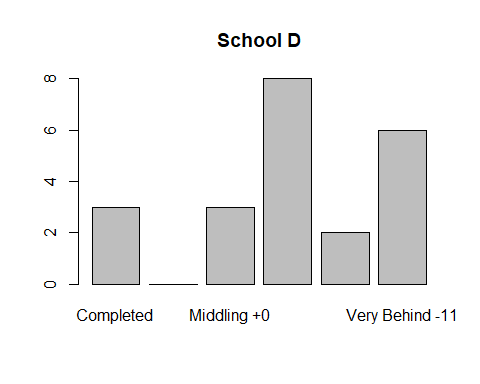
barplot(StudentSumsC, main = "School C")



StudentSumsD <- colSums(storytellerD[3:8])  
StudentSumsD

## Completed Very Ahead +5 Middling +0 Behind -1-5   
## 3 0 3 8   
## More Behind -6-10 Very Behind -11   
## 2 6

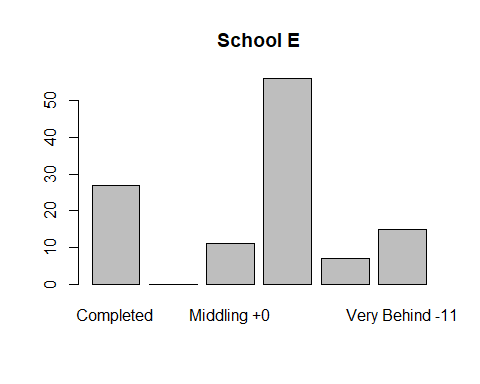
barplot(StudentSumsD, main = "School D")



StudentSumsE <- colSums(storytellerE[3:8])  
StudentSumsE

## Completed Very Ahead +5 Middling +0 Behind -1-5   
## 27 0 11 56   
## More Behind -6-10 Very Behind -11   
## 7 15

barplot(StudentSumsE, main = "School E")



**Initial Observations and Remarks**

Observations received with the dataset:

* Schools B and D do not follow the same picture that was shown by the data combined.
* A, C, and E follow generally the same pattern as shown by figure 1
* B shows many students ahead of the curriculum, completing the program.
* The majority of students that are behind are behind by 1-5 assignments, and very few are in the lowest two categories.
* This is the most positive picture that the data shows out of the schools.
* School D is the opposite with a small percentage of students having completed all assignments and a large percentage of students being considered ‘Very Behind.’

Comparing the two schools as a representation of the population of program takers …

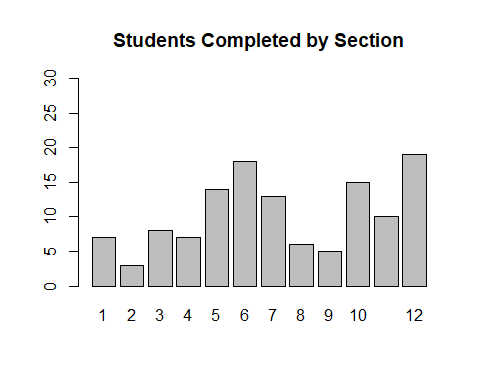
sum(StudentSumsB)/sum(StudentSums)

## [1] 0.2785759

sum(StudentSumsD)/sum(StudentSums)

## [1] 0.01374141

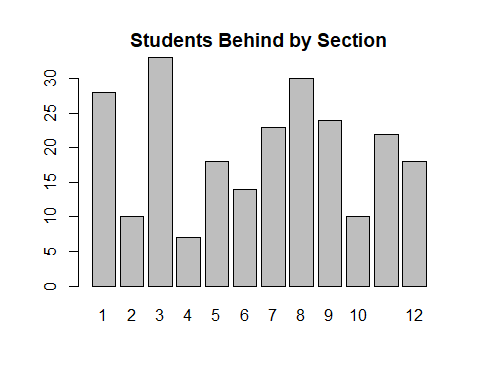
#looking at the data to determine quality sections  
  
barplot(storytellerB$Completed, names.arg =c(1:12),ylim=c(0,30), main = "Students Completed by Section")



rowSums(storytellerB[,3:8])

## [1] 39 18 47 18 40 40 45 46 39 28 39 47

barplot(storytellerB$`Behind -1-5`+storytellerB$`More Behind -6-10`+storytellerB$`Very Behind -11`, names.arg =c(1:12), ylim=c(0,30), main = "Students Behind by Section")



**Conclusion**

The progress segments give a high level view of where students are in their progress but an overall deeper dive is frustrated.

The progress segments may be valuable to alert for bucket metrics such as how many students have completed the course, how many are on track or better, how many need attention, and how many are in the unrecoverable failure zone, but any deeper understanding is frustrated.

It would have been preferable to receive more granular and precise information on what lessons students where on and what lesson they should be on to be on track.

The data results confirm that there is a difference in progress that can be related to specific school sections, as well as varying overall trends at schools being driven by how the segments are implemented at the school or the students, enrolled or how enrollment is handled and monitored.