IST707 HW 2

Dan Burke

10/19/2021

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)

Import Data from CSV

```
library(tidyr)
library(ggplot2)

schoolData <- read.csv("C:\\Users\\danbu\\Desktop\\Applied Machine Learning\\Week Three\\HW2\\data-story

colnames(schoolData)
```

```
## [1] "School"            "Section"           "Very.Ahead..5"
## [4] "Middling..0"       "Behind..1.5"       "More.Behind..6.10"
## [7] "Very.Behind..11"   "Completed"
```

```
newColNames <- c("School","Section", "Very Ahead", "Middling","Behind", "More Behind","Very Behind", "Co

#Rename columns for readability
colnames(schoolData)<- newColNames

sdf <- data.frame(schoolData)

schoolA <- sdf[which(sdf$School == "A"),3:8]
schoolB <- sdf[which(sdf$School == "B"),3:8]
schoolC <- sdf[which(sdf$School == "C"),3:8]
schoolD <- sdf[which(sdf$School == "D"),3:8]
schoolE <- sdf[which(sdf$School == "E"),3:8]
```

```
#Find How many and What are the Unique Sections

uniqueSections <- unique(sdf$Section)
uniqueSections
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13
```

```
length(uniqueSections)
```

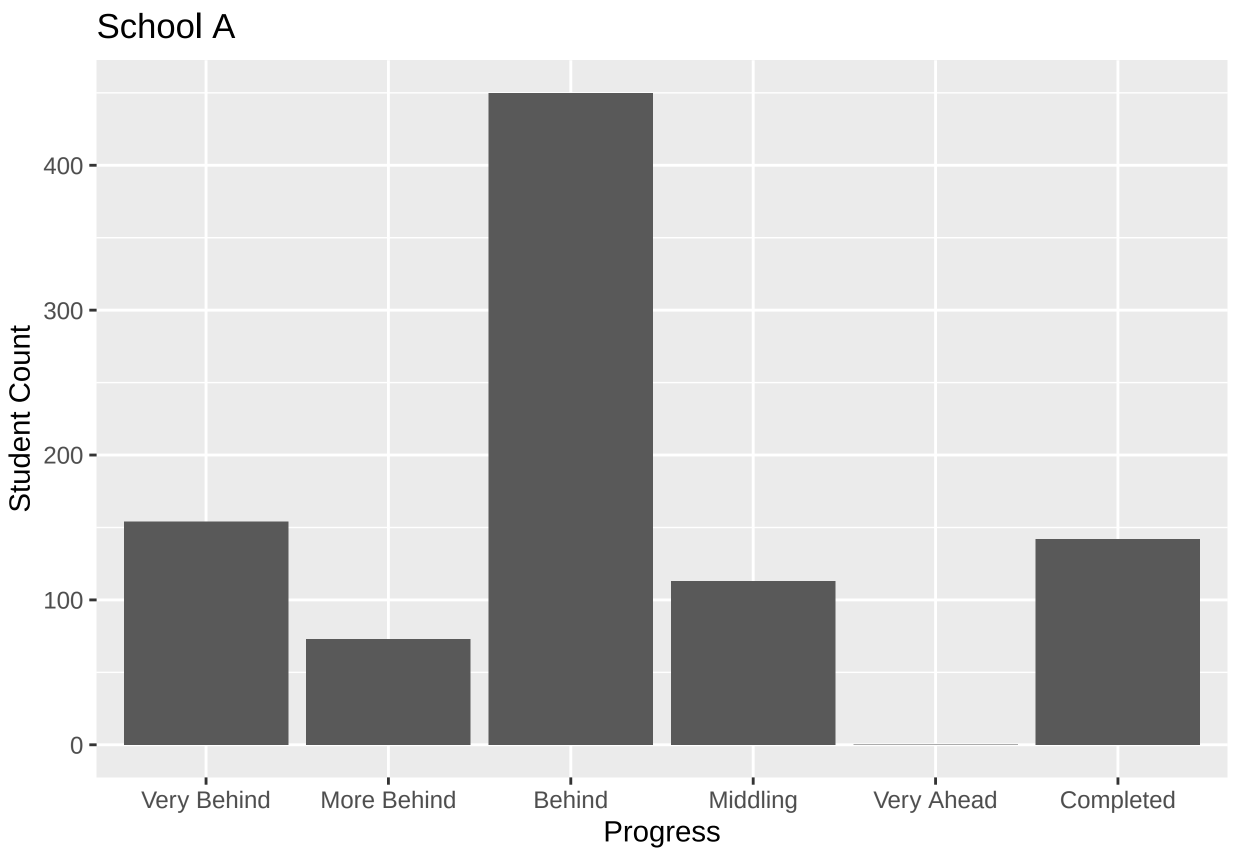
```
## [1] 13
```

```
schoolSums <- data.frame(as.factor(newColNames[3:8]))
colnames(schoolSums) <-c("Progress")
```

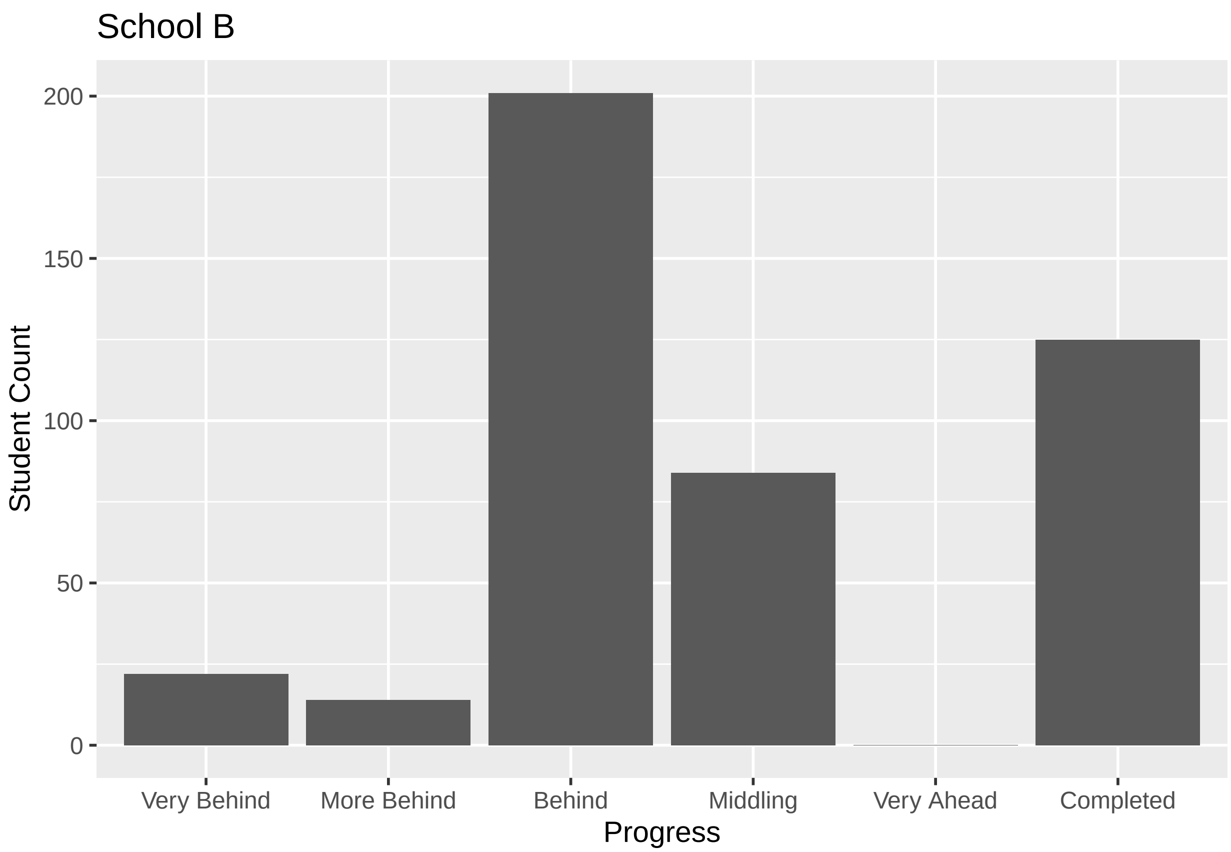
Explore The Data

In this particular scenario we have multiple sections across multiple schools. Lets first look at if there is a general trend with all schools, or if it is perhaps the section material.

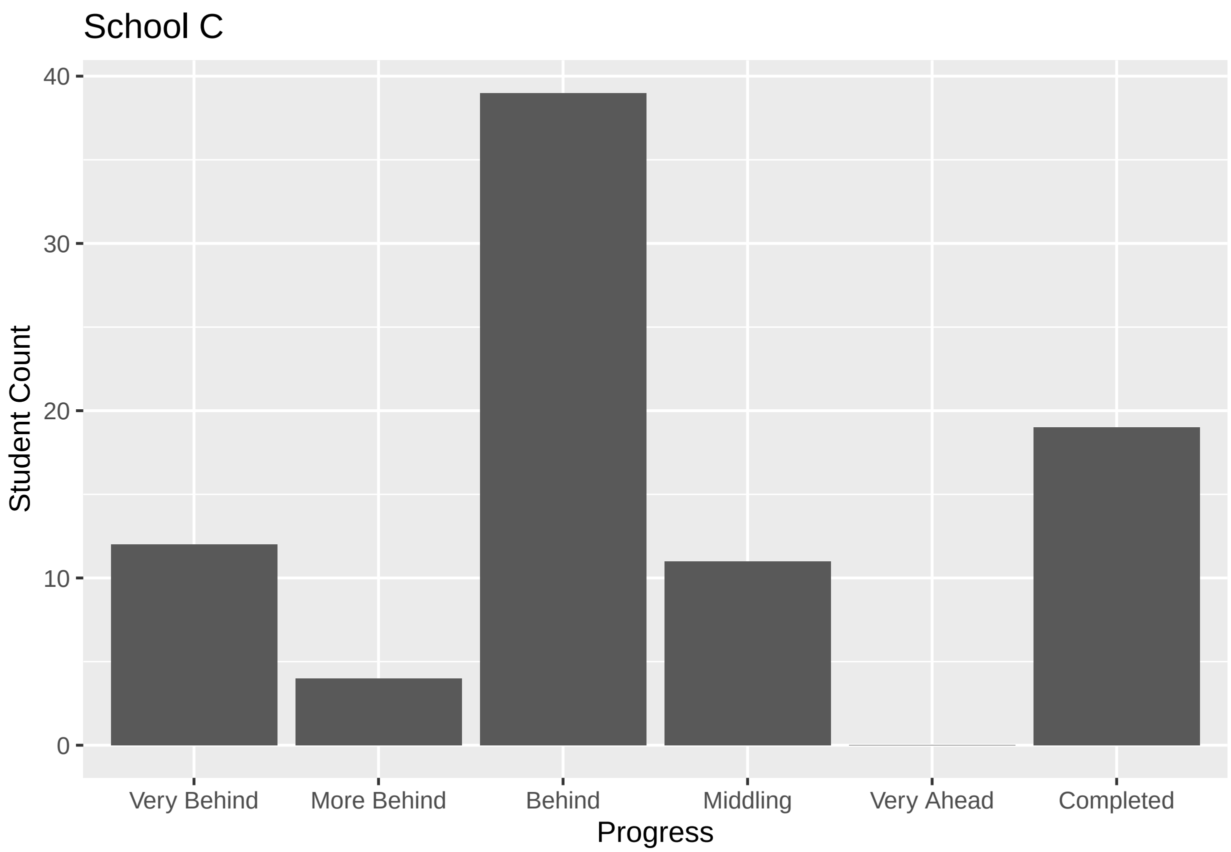
```
#School A
schoolASums <- schoolSums
schoolASums$count <- unlist(colSums(schoolA), recursive = TRUE, use.names = TRUE)

ggplot(schoolASums, aes(x = Progress, y = count)) +
       geom_bar(stat="identity") + xlab("Progress") + ylab("Student Count") + ggtitle("School A") + scal
```



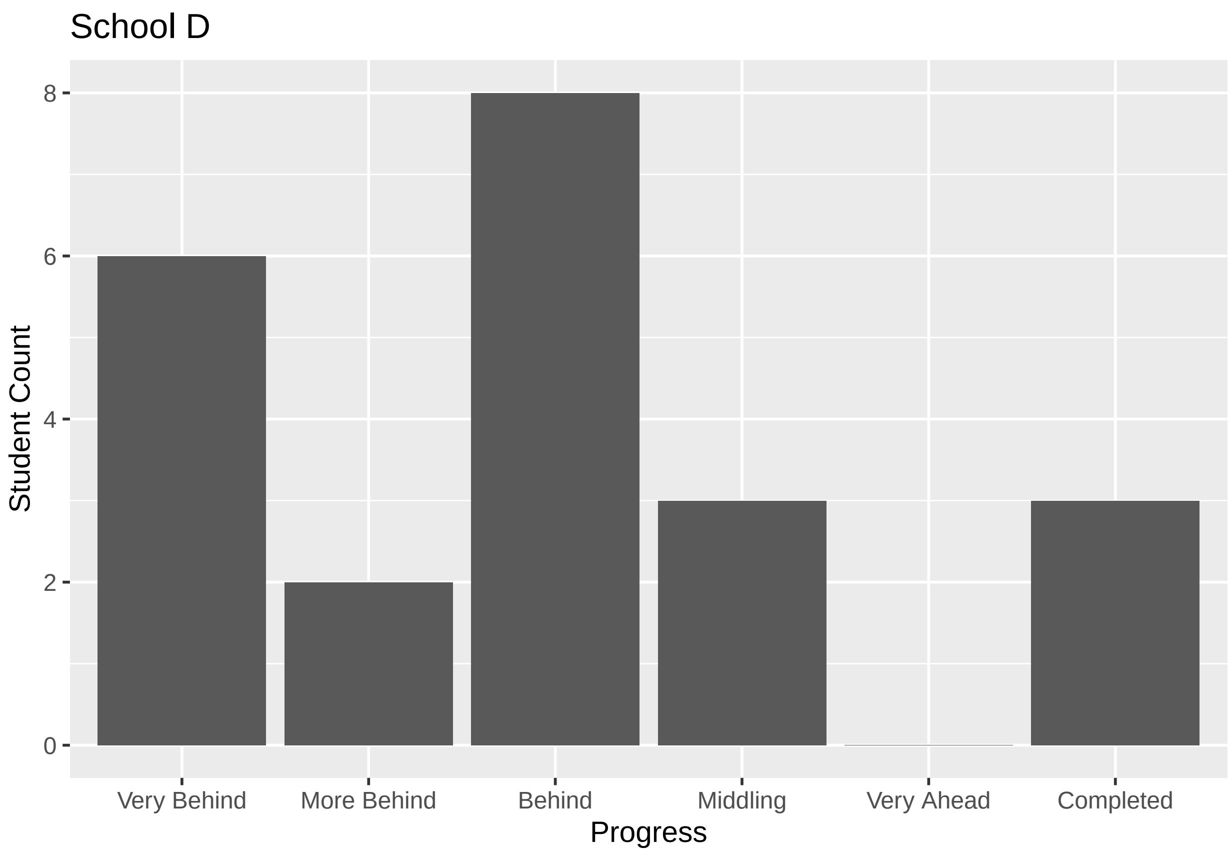
```
#School B
schoolBSums <- schoolSums
schoolBSums$count <- unlist(colSums(schoolB), recursive = TRUE, use.names = TRUE)

ggplot(schoolBSums, aes(x = Progress, y = count)) +
       geom_bar(stat="identity") + xlab("Progress") + ylab("Student Count") + ggtitle("School B")+ scale
```



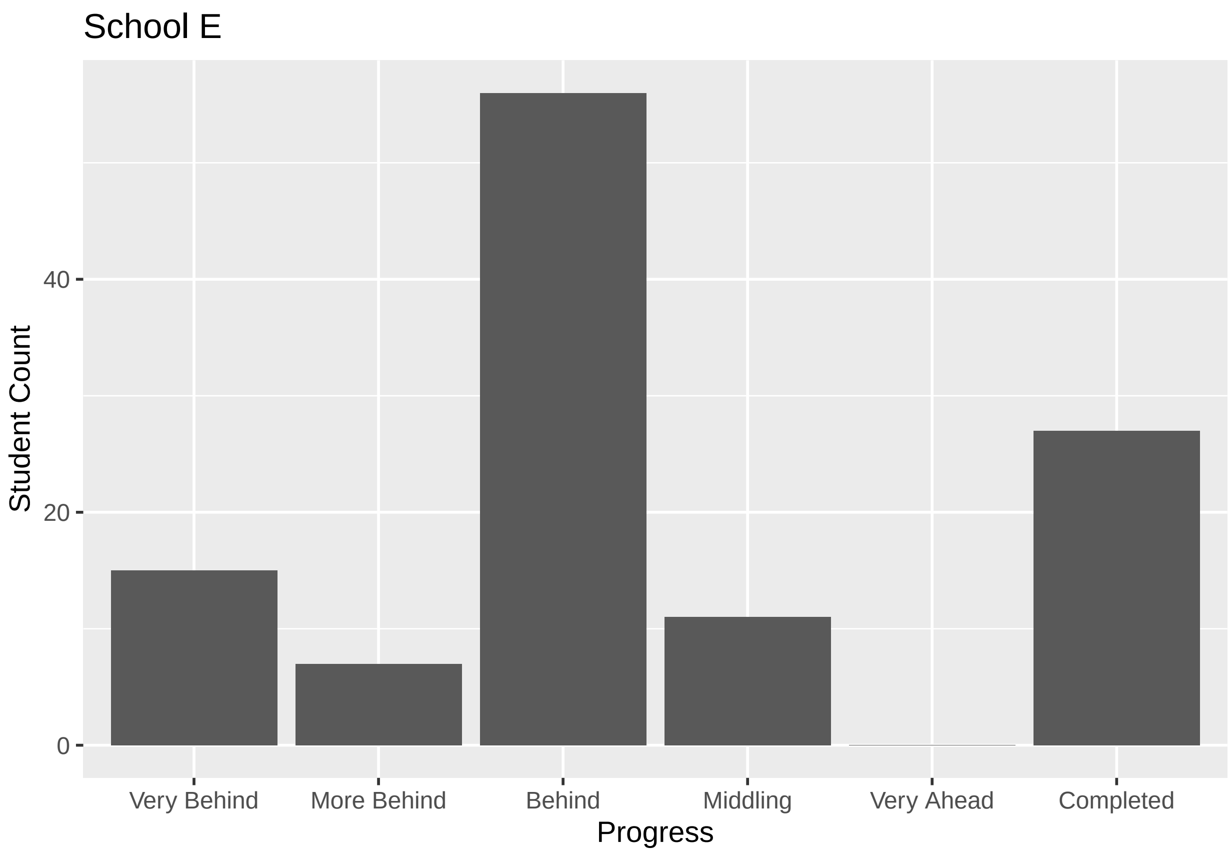
```
#School C
schoolCSums <- schoolSums
schoolCSums$count <- unlist(colSums(schoolC), recursive = TRUE, use.names = TRUE)

ggplot(schoolCSums, aes(x = Progress, y = count)) +
       geom_bar(stat="identity") + xlab("Progress") + ylab("Student Count") + ggtitle("School C")+ scal
```



```
#School D
schoolDSums <- schoolSums
schoolDSums$count <- unlist(colSums(schoolD), recursive = TRUE, use.names = TRUE)

ggplot(schoolDSums, aes(x = Progress, y = count)) +
       geom_bar(stat="identity") + xlab("Progress") + ylab("Student Count") + ggtitle("School D")+ scal
```

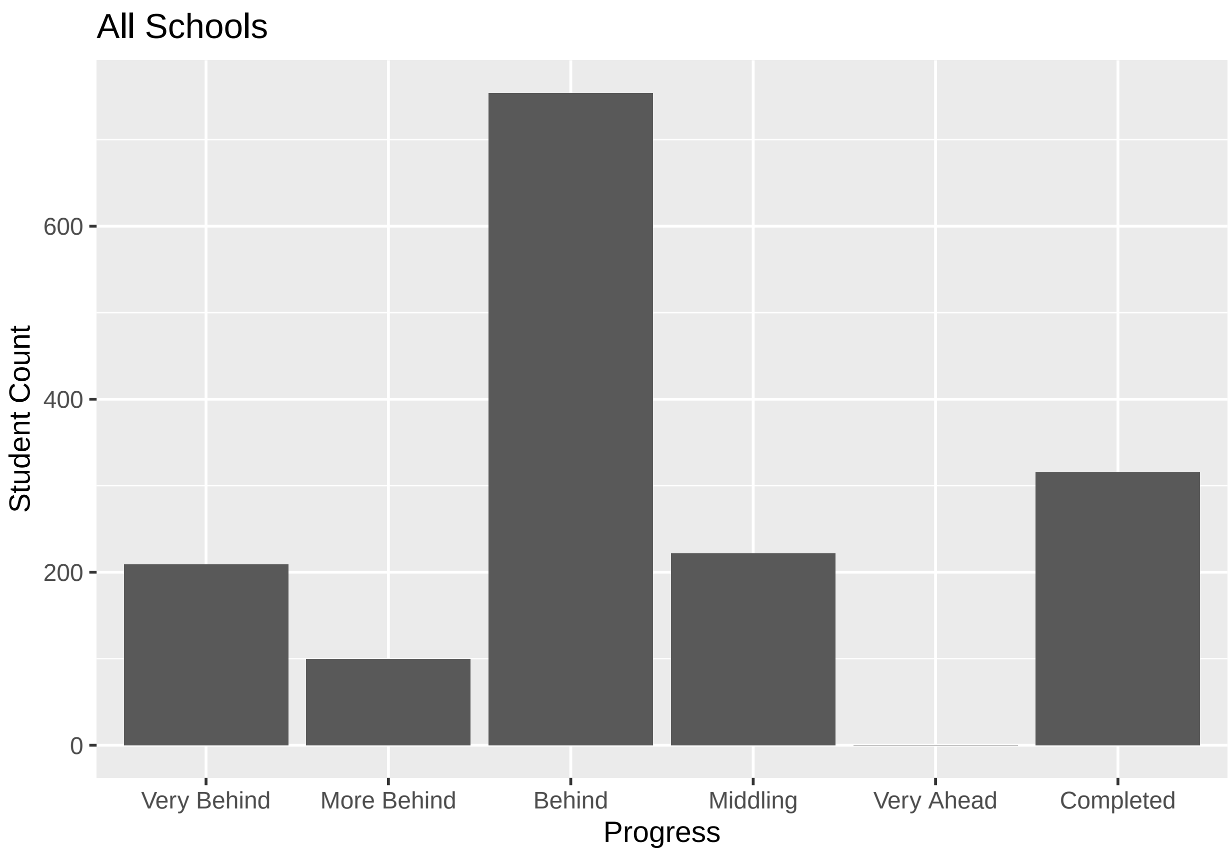


```
#School E
schoolESums <- schoolSums
schoolESums$count <- unlist(colSums(schoolE), recursive = TRUE, use.names = TRUE)

ggplot(schoolESums, aes(x = Progress, y = count)) +
       geom_bar(stat="identity") + xlab("Progress") + ylab("Student Count") + ggtitle("School E")+ scal
```



All schools contained within the data show a similarity when ploted via a histogram, most students seem to be found congregating close to behind & completed.

```
#Scatter plot of All Schools with Color as the Section
allsum <- unlist(colSums(sdf[,3:8]), recursive = TRUE, use.names = TRUE)
allSchoolSums <- data.frame(newColNames[3:8], allsum)
colnames(allSchoolSums) <- c("Progress", "Count")

ggplot(allSchoolSums, aes(x = Progress, y = Count)) +
       geom_bar(stat="identity") + xlab("Progress") + ylab("Student Count") + ggtitle("All Schools")+ s
```



After Evaluating each school individually, now we'll example which schools and sections contain the most students who are very behinc, behind and complete. I do this by groupling each school, then normalizing the count of each student, this way we can look at all schools' data holistically and proportionally to their size. This is helpful incase one of these school is less funded than another and will help prevent more data available from one school compared to another.

```
#Adding in a scaling function to normalize between 0 to 1
range01 <- function(x){(x-min(x))/(max(x)-min(x))}

schoolA <- data.frame(sdf[which(sdf$School == "A"),3:8])
schoolB <- sdf[which(sdf$School == "B"),3:8]
schoolC <- sdf[which(sdf$School == "C"),3:8]
schoolD <- sdf[which(sdf$School == "D"),3:8]
schoolE <- sdf[which(sdf$School == "E"),2:8]


scaledA <- as.data.frame(range01(sdf[which(sdf$School == "A"),3:8]))
```

```
scaledA$Section <- sdf[which(sdf$School == "A"),2]
scaledA['School'] = "A"

scaledB <- as.data.frame(range01(sdf[which(sdf$School == "B"),3:8]))
scaledB$Section <- sdf[which(sdf$School == "B"),2]
scaledB['School'] = "B"

scaledC <- as.data.frame(range01(sdf[which(sdf$School == "C"),3:8]))
scaledC$Section <- sdf[which(sdf$School == "C"),2]
scaledC['School'] = "C"

scaledD <- as.data.frame(range01(sdf[which(sdf$School == "D"),3:8]))
scaledD$Section <- sdf[which(sdf$School == "D"),2]
scaledD['School'] = "D"

scaledE <- as.data.frame(range01(sdf[which(sdf$School == "E"),3:8]))
scaledE$Section <- sdf[which(sdf$School == "E"),2]
scaledE['School'] = "E"


normalizedStudents <- rbind(scaledA, scaledB, scaledC, scaledD, scaledE)
```

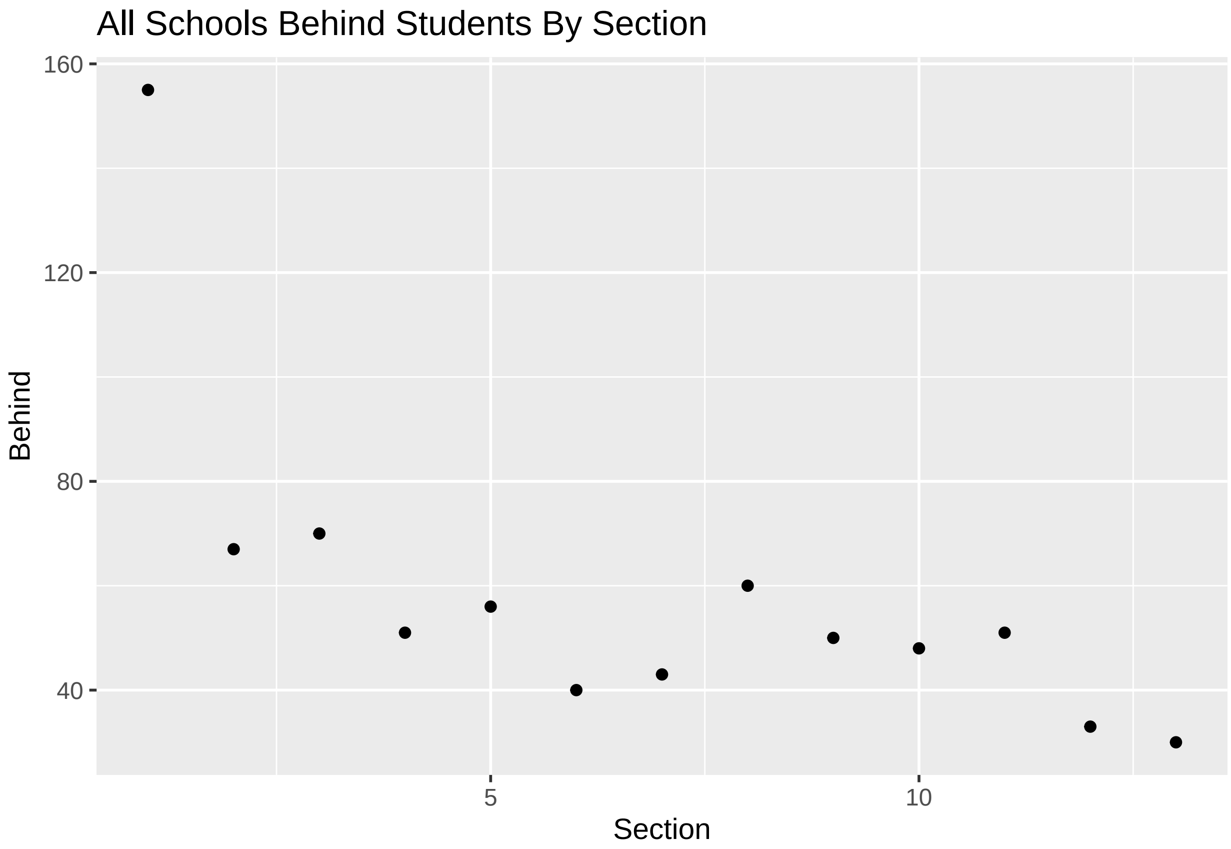
Below I'll aggregate by section in order to produce a bar chart which displays all sections along the X axis.

I've also produced a scatter plot plotting section along the X Axis and behind along the Y axis. I've also used color to denote which point belongs to which school (Key located on the right hand side of chart).This will aid in identifying if a particular section or set of sections are where students are falling behind.

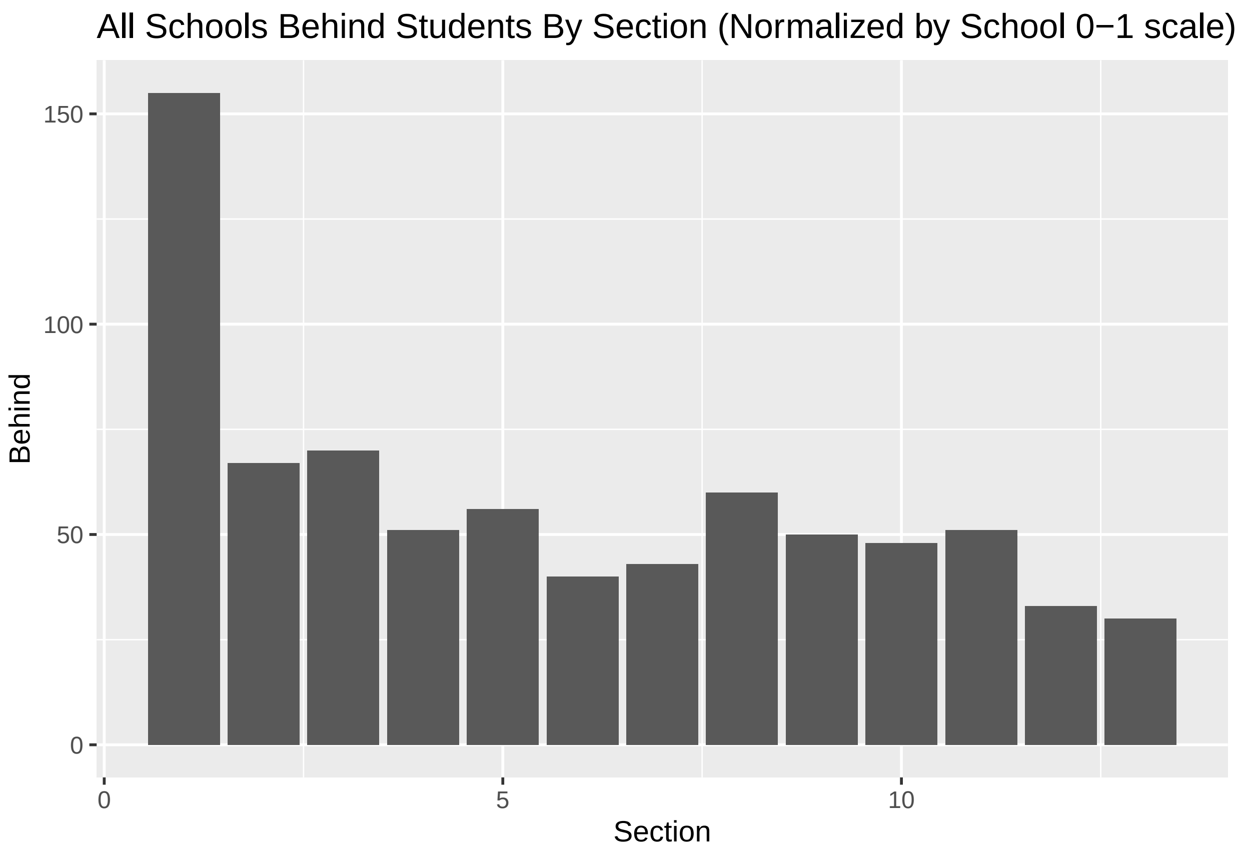
```
#Now We need to Aggregate by Section
require(dplyr)
```

```
## Loading required package: dplyr

## Warning: package 'dplyr' was built under R version 4.1.1

##
## 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(dplyr)

aggBySectionBehind <- sdf %>% group_by(Section) %>% summarise(Behind = sum(Behind))

#Plot Scatter Plot X Axis is Section, Y Axis is Behind Count, Color is School

ggplot(aggBySectionBehind, aes(x=Section, y=Behind)) + geom_point() + ggtitle("All Schools Behind Studen
```



```
ggplot(aggBySectionBehind, aes(x = Section, y = Behind)) +
       geom_bar(stat="identity") + xlab("Section") + ylab("Behind") + ggtitle("All Schools Behind Studen
```



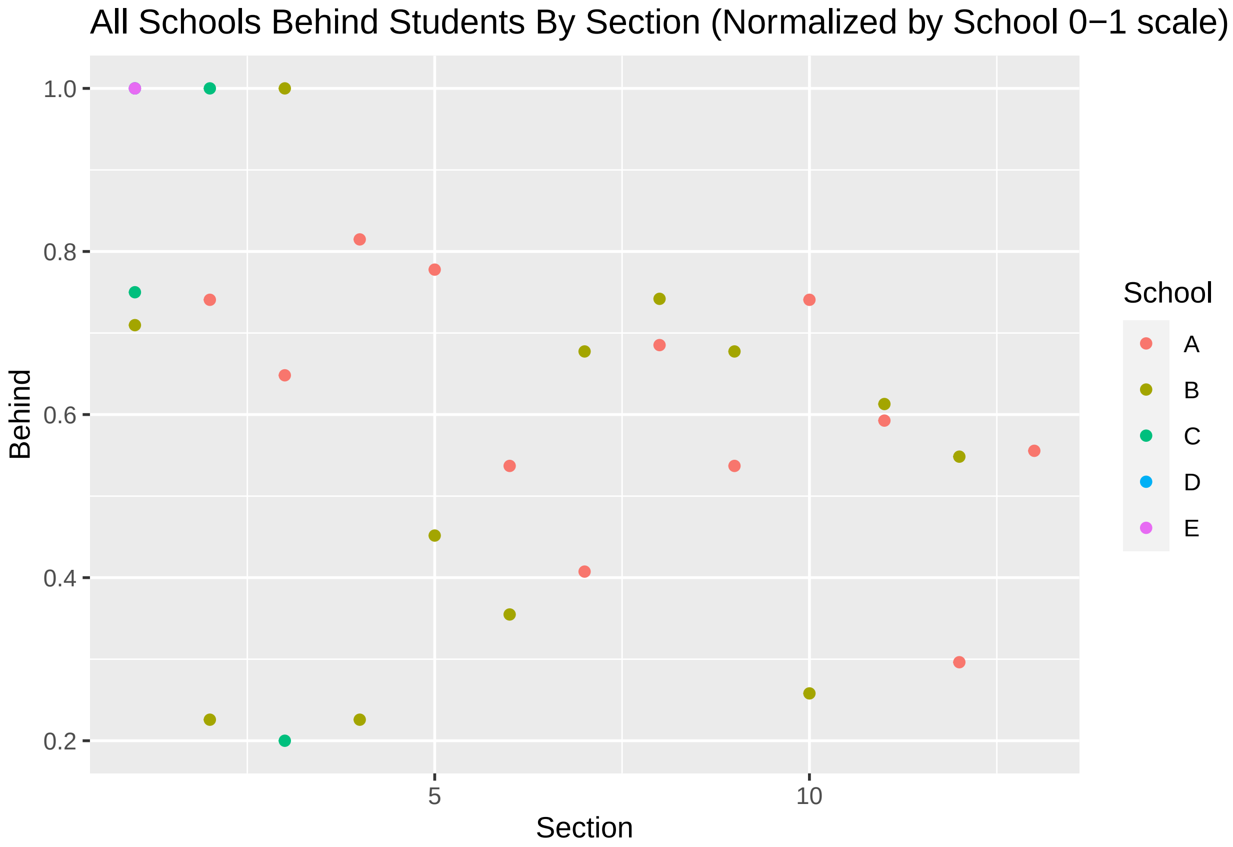
```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.1.1
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
ggplot(normalizedStudents, aes(x=Section, y=Behind, color=School)) + geom_point() + ggtitle("All Schools
```



After viewing the bar chart (X=Section, Y= Behind Count) along with the colorized scatter plot (in conjunction with the intitial bar charts at the begining of my exploration) it appears evident that across all schools students are struggling with early course material with the chart heavily skewed to the left.

However, after examining the colorized scatter plot (normalized section counts), it appears to tell a different story. A story where of the data provided, school E, B, C have a greater abount of students behind within early sections, while schools A possesses students behind more flatly across all sections. School B posseses some similarity to School A with the exception of some students struggling within earlier sections.