IST 707 HW2

Victoria Haley

2023-04-14

Import Data

NA's taken care of with importing CSV file wizard

```
library(readr)
setwd("/Users/victoriahaley")
data_storyteller <- read_csv("Downloads/data_storyteller.csv")
```

```
## Rows: 30 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): School
## dbl (7): Section, Very Ahead +5, Middling +0, Behind -1-5, More Behind -6-10...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Examine data

```
str(data_storyteller)
```

```
## spc_tbl_ [30 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ School           : chr [1:30] "A" "A" "A" "A" ...
##  $ Section          : num [1:30] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Very Ahead +5    : num [1:30] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Middling +0      : num [1:30] 5 8 9 14 9 7 19 3 6 13 ...
##  $ Behind -1-5      : num [1:30] 54 40 35 44 42 29 22 37 29 40 ...
##  $ More Behind -6-10: num [1:30] 3 10 12 5 2 3 5 11 8 5 ...
##  $ Very Behind -11  : num [1:30] 9 16 13 12 24 10 14 18 12 5 ...
##  $ Completed        : num [1:30] 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()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Data Cleaning

The 'School' and 'Section' columns should be factors, and columns that describe where each student falls in are discrete and should be integers rather in numeric format.

```
data_storyteller$School <- factor(data_storyteller$School)
data_storyteller$Section <- factor(data_storyteller$Section)
data_storyteller$`Very Ahead +5` <- as.integer(data_storyteller$`Very Ahead +5`)
data_storyteller$`Middling +0`<- as.integer(data_storyteller$`Middling +0`)
data_storyteller$`Behind -1-5` <-as.integer(data_storyteller$`Behind -1-5`)
data_storyteller$`More Behind -6-10` <- as.integer(data_storyteller$`More Behind -6-10`)
data_storyteller$`Very Behind -11`<- as.integer(data_storyteller$`Very Behind -11`)
data_storyteller$Completed <- as.integer(data_storyteller$Completed)
```

Handle missing values

```
is.na(data_storyteller)
```

```
##       School Section Very Ahead +5 Middling +0 Behind -1-5 More Behind -6-10
##  [1,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
##  [2,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
##  [3,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
##  [4,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
##  [5,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
##  [6,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
##  [7,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
##  [8,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
##  [9,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [10,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [11,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [12,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [13,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [14,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [15,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [16,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [17,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [18,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [19,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [20,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [21,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [22,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [23,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [24,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [25,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [26,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [27,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [28,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [29,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
## [30,]  FALSE   FALSE         FALSE       FALSE       FALSE             FALSE
##       Very Behind -11 Completed
##  [1,]           FALSE     FALSE
##  [2,]           FALSE     FALSE
##  [3,]           FALSE     FALSE
##  [4,]           FALSE     FALSE
```

```
##  [5,]           FALSE     FALSE
##  [6,]           FALSE     FALSE
##  [7,]           FALSE     FALSE
##  [8,]           FALSE     FALSE
##  [9,]           FALSE     FALSE
## [10,]           FALSE     FALSE
## [11,]           FALSE     FALSE
## [12,]           FALSE     FALSE
## [13,]           FALSE     FALSE
## [14,]           FALSE     FALSE
## [15,]           FALSE     FALSE
## [16,]           FALSE     FALSE
## [17,]           FALSE     FALSE
## [18,]           FALSE     FALSE
## [19,]           FALSE     FALSE
## [20,]           FALSE     FALSE
## [21,]           FALSE     FALSE
## [22,]           FALSE     FALSE
## [23,]           FALSE     FALSE
## [24,]           FALSE     FALSE
## [25,]           FALSE     FALSE
## [26,]           FALSE     FALSE
## [27,]           FALSE     FALSE
## [28,]           FALSE     FALSE
## [29,]           FALSE     FALSE
## [30,]           FALSE     FALSE
```

```
#there appears to be no missing data, so further investigating is unnecessary
```

See number of sections by school

```
table(data_storyteller$School)
```

```
##
##  A  B  C  D  E
## 13 12  3  1  1
```

```
table(data_storyteller$School)[which.max(table(data_storyteller$School))]
```

```
##  A
## 13
```

School A has the most sections, with only 1 more than B. Schools C, D, and E all have a significantly lower amount of sections.

Organizing the data structure so that it is more intuitive

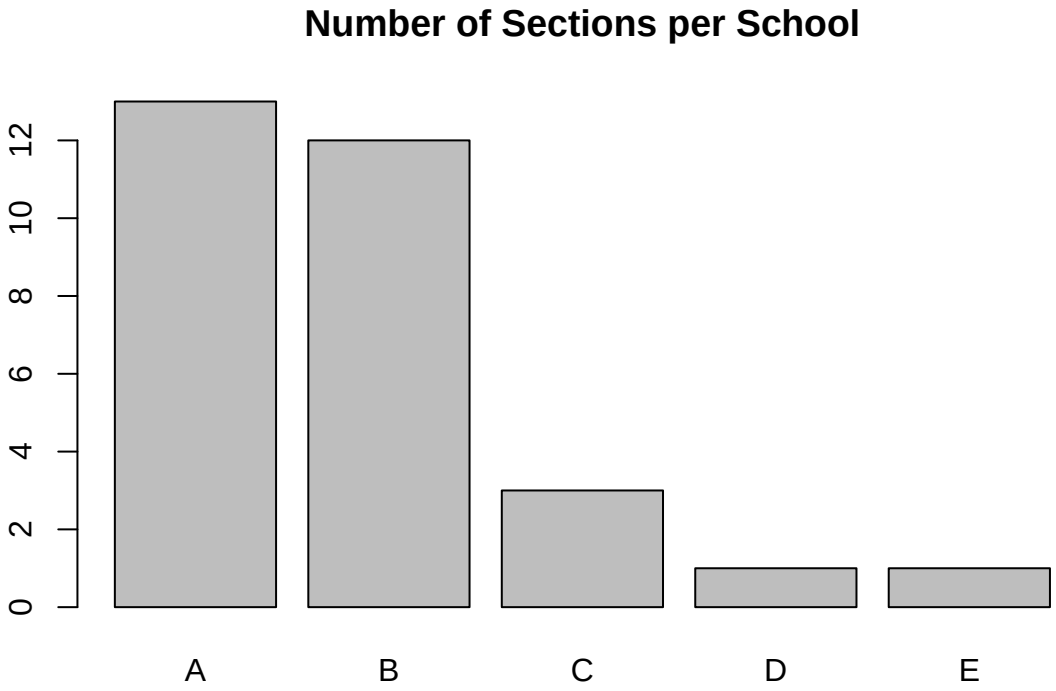
```
storyteller <- data_storyteller[, c(2,1,8,3,4,5,6,7)]
head(storyteller) #displaying top 5 rows
```

```
## # A tibble: 6 x 8
##   Section School Completed `Very Ahead +5` `Middling +0` Behin~1 More ~2 Very ~3
##   <fct>   <fct>      <int>           <int>         <int>   <int>   <int>   <int>
## 1 1       A             10               0             5      54       3       9
```

```
## 2 2       A              6               0             8      40      10      16
## 3 3       A             11               0             9      35      12      13
## 4 4       A             10               0            14      44       5      12
## 5 5       A              8               0             9      42       2      24
## 6 6       A              9               0             7      29       3      10
## # ... with abbreviated variable names 1: `Behind -1-5`, 2: `More Behind -6-10`,
## #   3: `Very Behind -11`
```

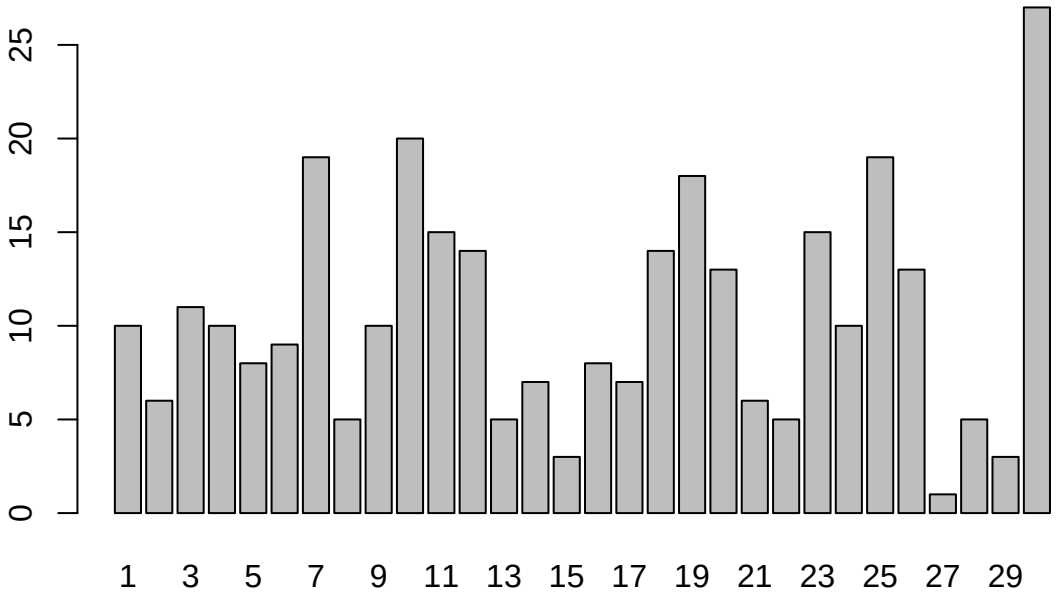
EDA and Visualization

```
#Barplot showing the distributions of sections at each school.
SchoolValues <- c(length(which(storyteller$School=='A')),length(which(storyteller$School=='B')), length
barplot(SchoolValues, names.arg=c("A", "B", "C","D","E"), main='Number of Sections per School')
```

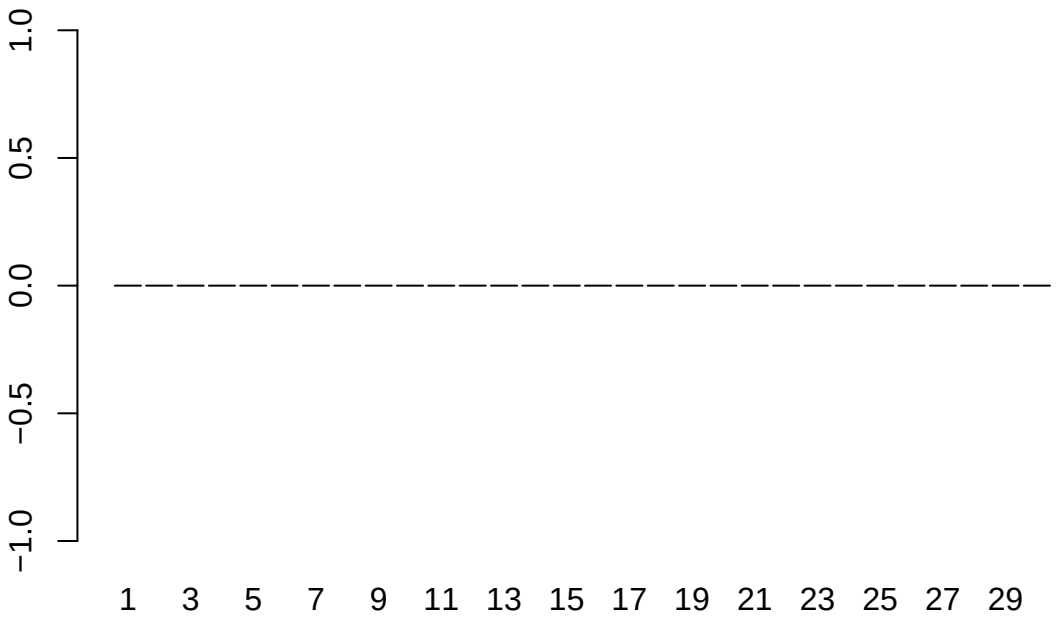


Right off the bat we can see that schools A and B have more sections than C, D, and E combined.

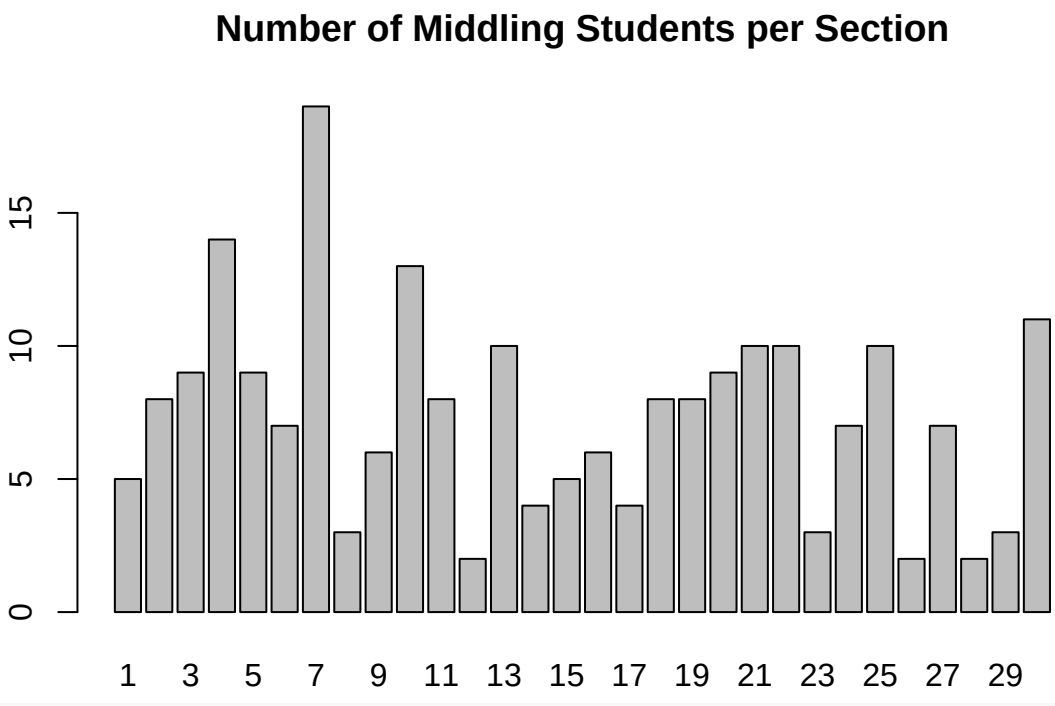
```
#barplot showing the number of students that have completed each section
barplot(storyteller$Completed, names.arg = c(1:30), main='Number of Completed Students per Section')
```



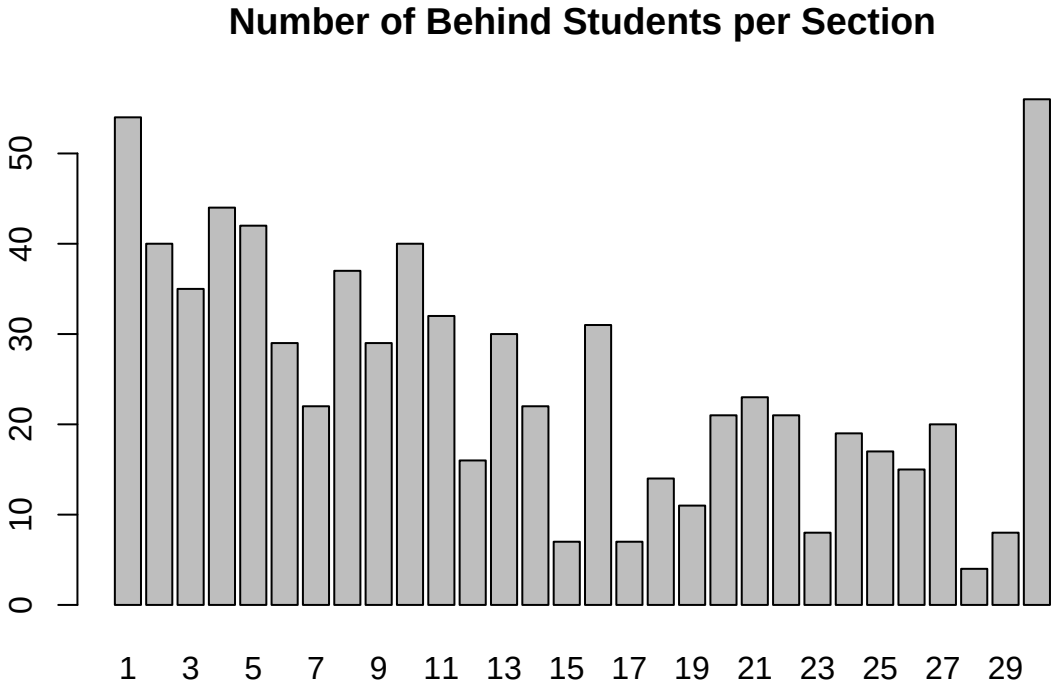

```
#barplot showing the number of students that are very ahead in each section
barplot(storyteller$`Very Ahead +5`, names.arg = c(1:30), main='Number of Very Ahead Students per Sectio
```



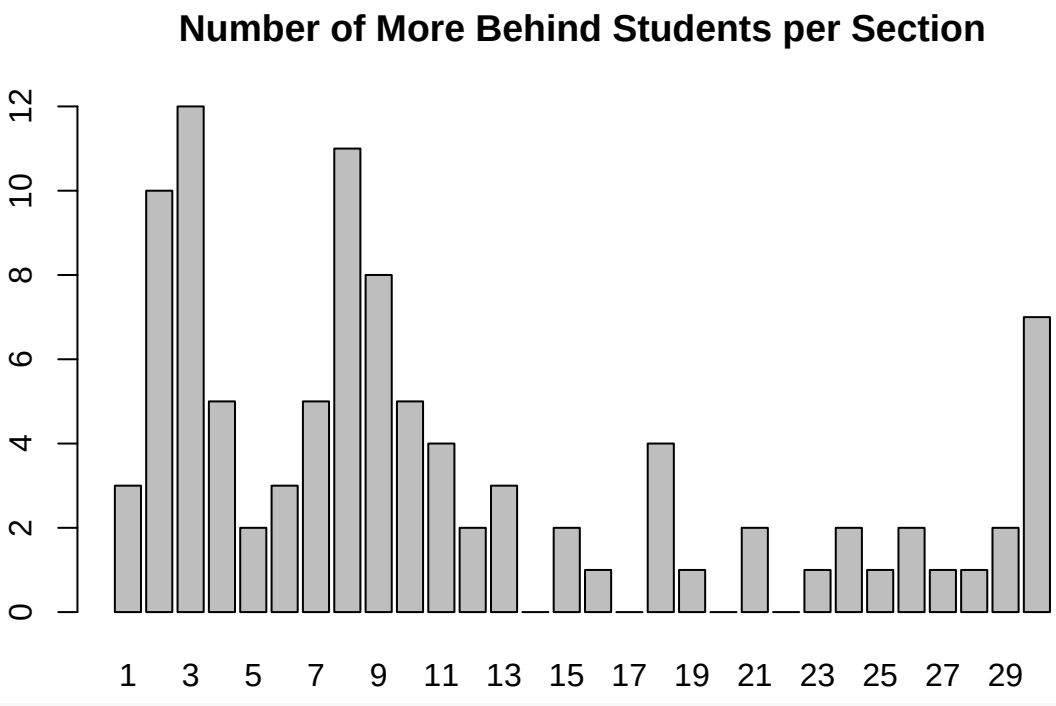

```
#barplot showing the number of students that are middling in each section
barplot(storyteller$`Middling +0`, names.arg = c(1:30), main='Number of Middling Students per Section')
```



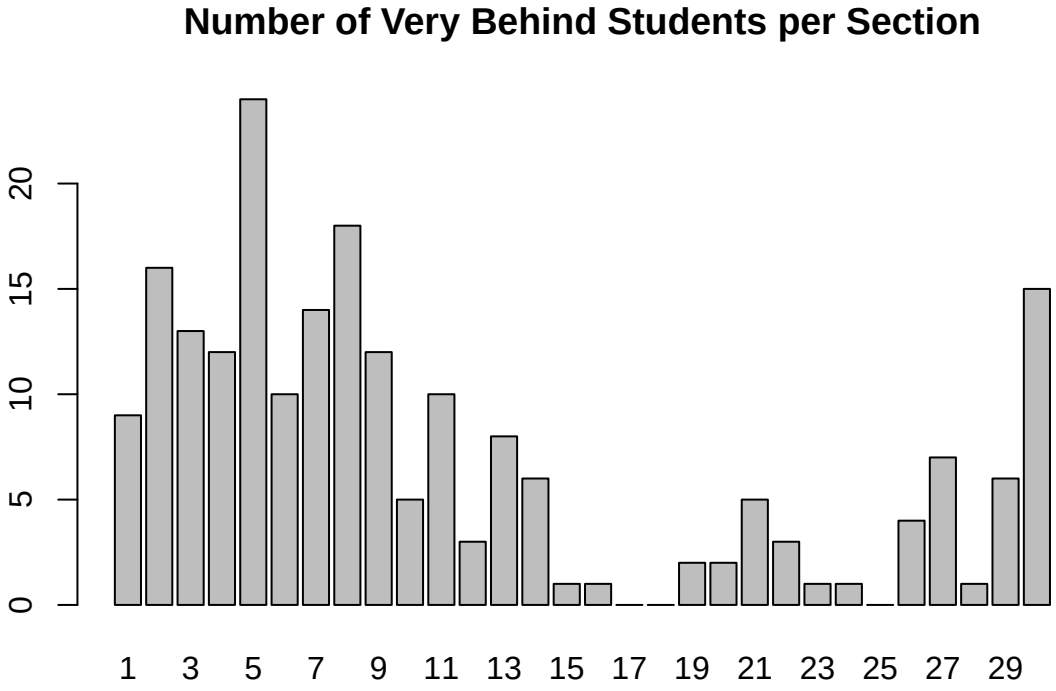

```
#barplot showing the number of students that are behind in each section
barplot(storyteller$`Behind -1-5`, names.arg = c(1:30), main='Number of Behind Students per Section')
```




```
#barplot showing the number of students that are more behind in each section
barplot(storyteller$`More Behind -6-10`, names.arg = c(1:30), main='Number of More Behind Students per S
```




```
#barplot showing the number of students that are very behind in each section
barplot(storyteller$`Very Behind -11`, names.arg = c(1:30), main='Number of Very Behind Students per Se
```



At a glance, there are a few differences that can easily be seen between where students stand in each section.

Taking a look at schools on an individual level

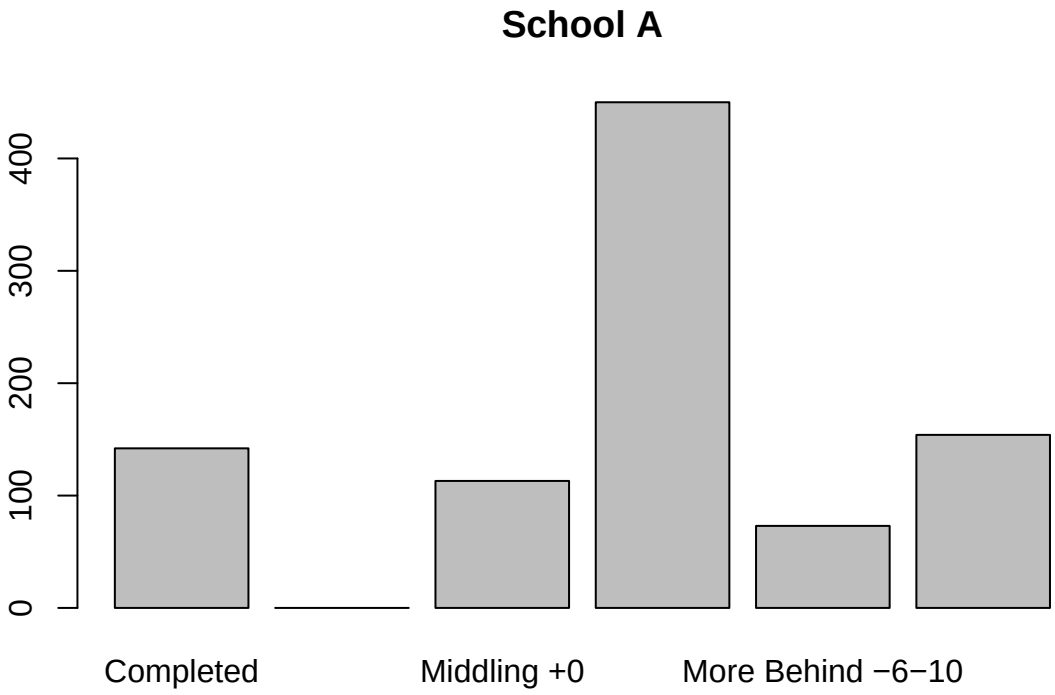
```
storytellerA <- storyteller[which(storyteller$School == "A"),]
storytellerB <- storyteller[which(storyteller$School == "B"),]
storytellerC <- storyteller[which(storyteller$School == "C"),]
```

```
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
```

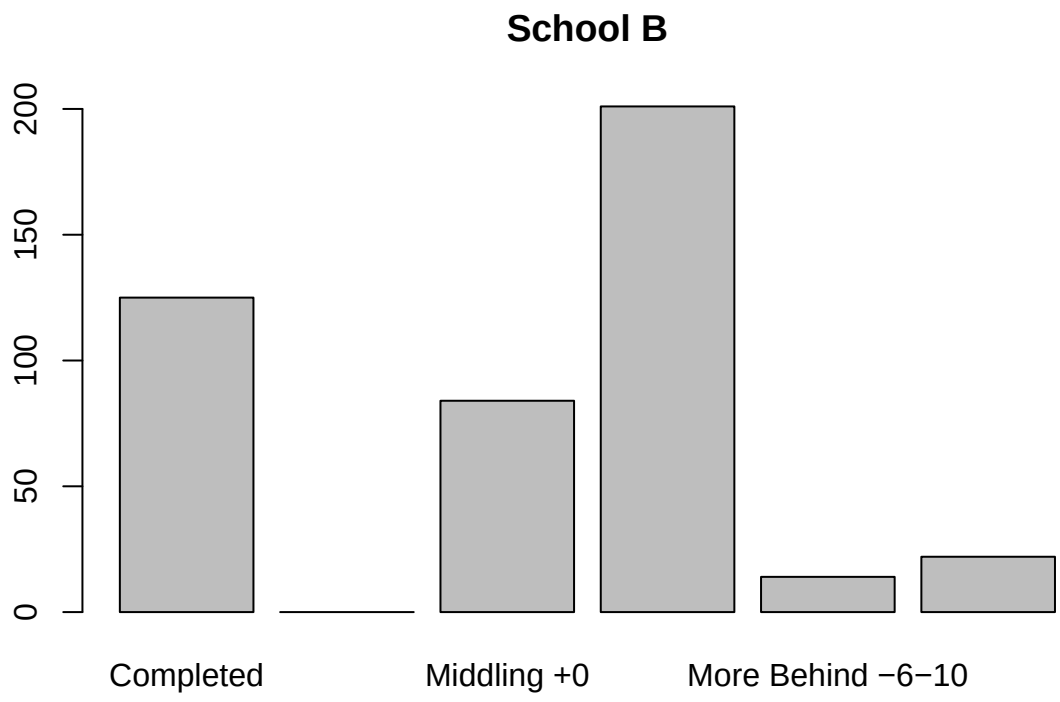
```
barplot(StudentSumsA, main = "School A")
```




```
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
```

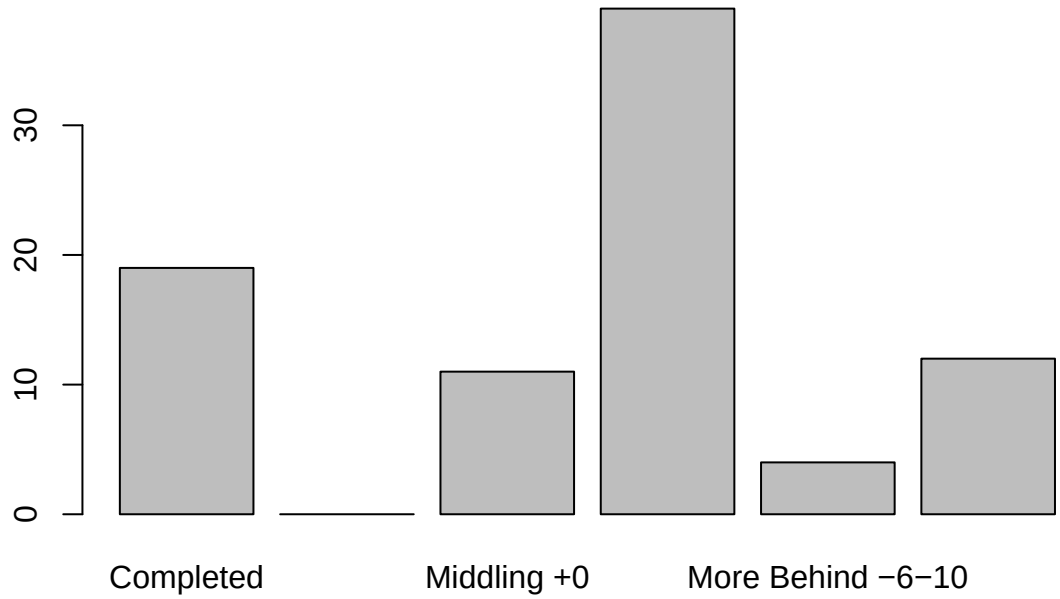
```
barplot(StudentSumsB, main = "School B")
```




```
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
```

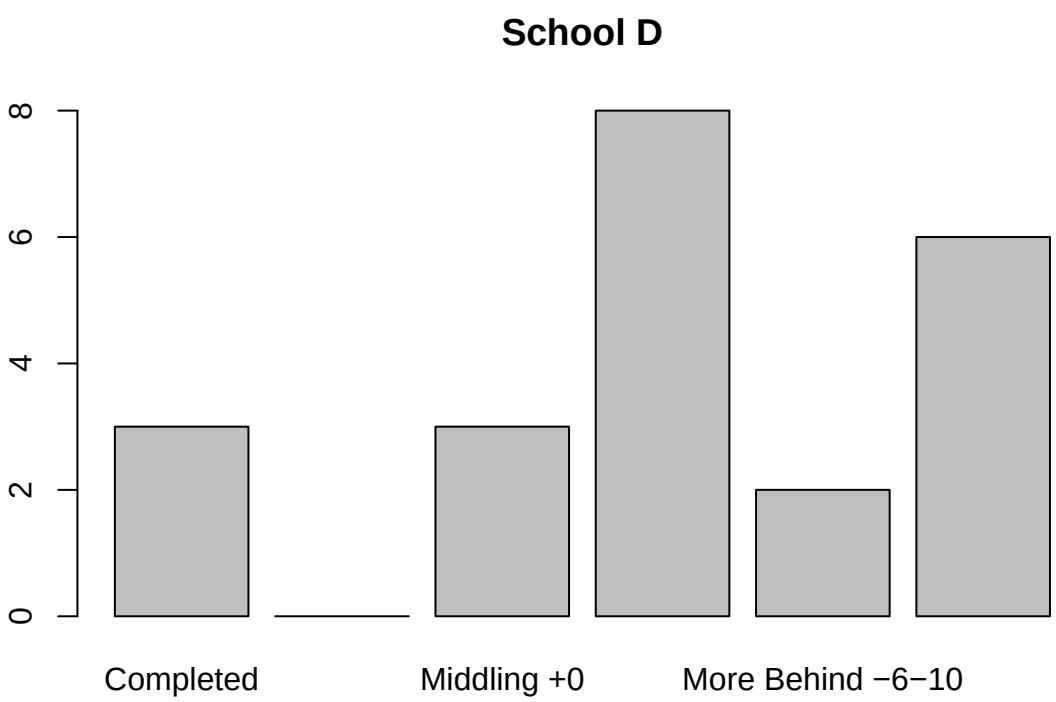
```
barplot(StudentSumsC, main = "School C")
```



```r
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
```

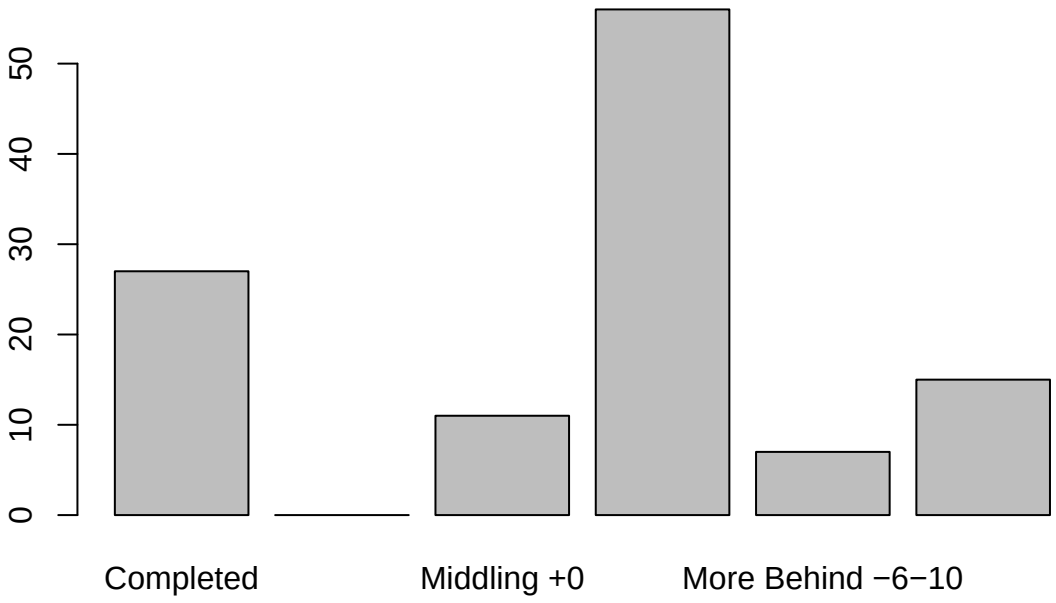
```r
barplot(StudentSumsD, main = "School D")
```




```r
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
```

```r
barplot(StudentSumsE, main = "School E")
```



Observation and Report

Looking at the data, we can see that sections are not evenly offered at each school. Data from Schools A and B heavily outweigh the rest, and likely skews the information. I'm curious as to why these particular schools offer more sections. However, with the information that is provided, it appears to suggest that the success of each student does not correlate with the number of sections offered at a given school. Looking at the graphs in the "Taking a look at schools on an individual level" section shows that students each school are more likely to be "Behind" on assignments in their classes. All schools seem to have a similar distribution of students in the "Middling" to "More Behind" range, while all schools except for A and D have more students that have "Completed" their class rather than being "Very behind." Seeing that the semester is 3/4 of the way through, the information shown here is troubling but is also lacking a lot of context. For instance, are these numbers based off of true late work (student's fault) or could it be due to teachers not submitting grades on time? Knowing what is causing these students to be falling behind at this point of the school year is crucial for the data analysis company to know so that they are able to plan the best course of action to get these schools on track to providing quality education to our future generations, if that is the purpose of these schools to consult with the data analysis company.