# Data Preparation

Benjamin J Heindl

## Introduction

This report analyzes a small dataset provided, capturing the progress of students in five schools (A, B, C, D, and E) implementing the same math course this semester. The course is comprised of 35 lessons, and there are 30 sections total. The dataset records the number of students who are very ahead, middling, behind, more behind, very behind, or completed the course in each section.

The objective of this analysis is to explore the data and uncover insights that can provide a story (or stories) that can be truthfully and effectively communicated through visualizations. Data exploration and transformation techniques will be used to clean and prepare the data.

This report focuses on identifying the distribution of student performance across the different sections and schools, and identifying any trends in student performance. Factors contributing to these trends will also be investigated. The analysis aims to provide a comprehensive and accurate representation of the data for stakeholders involved in the course's implementation.

## Reading in the Data

```
#replace blank entries as 'NA'.

data_storyteller <- read_csv("data-storyteller.csv", na = c(""))
```

## Inspect data and structure

```
#checking the structure to see which data types need to be adjusted
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 column is listed as a character type and it should be listed as a factor.

```
data_storyteller$School<-factor(data_storyteller$School)

#The Section column is listed as Numeric type and should be listed as factor
data_storyteller$Section <-factor(data_storyteller$Section)

#The six remaining columns are each a count of the students within each category.
#Each of these columns should all be listed as integers
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)
```

## Organizing the Data Structure

```
#Restructuring columns to improved readability.
#Moving "Completed" to the left and making "Section" the first column for better
#section identification and visual appeal.

storyteller2<-data_storyteller[,c(2,1,8,3,4,5,6,7)]

#Preview the top 5 rows.
head(storyteller2)
```

```
## # 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
## # i 2 more variables: ‘More Behind -6-10‘ <int>, ‘Very Behind -11‘ <int>
```

### Missing Data

```
#Checking for any NA values
sum(is.na(storyteller2))
```

```
## [1] 0
```

```
#There are no NA values present in the dataset.

#The dataset has undergone cleaning and is now ready for analysis.
head(storyteller2)
```

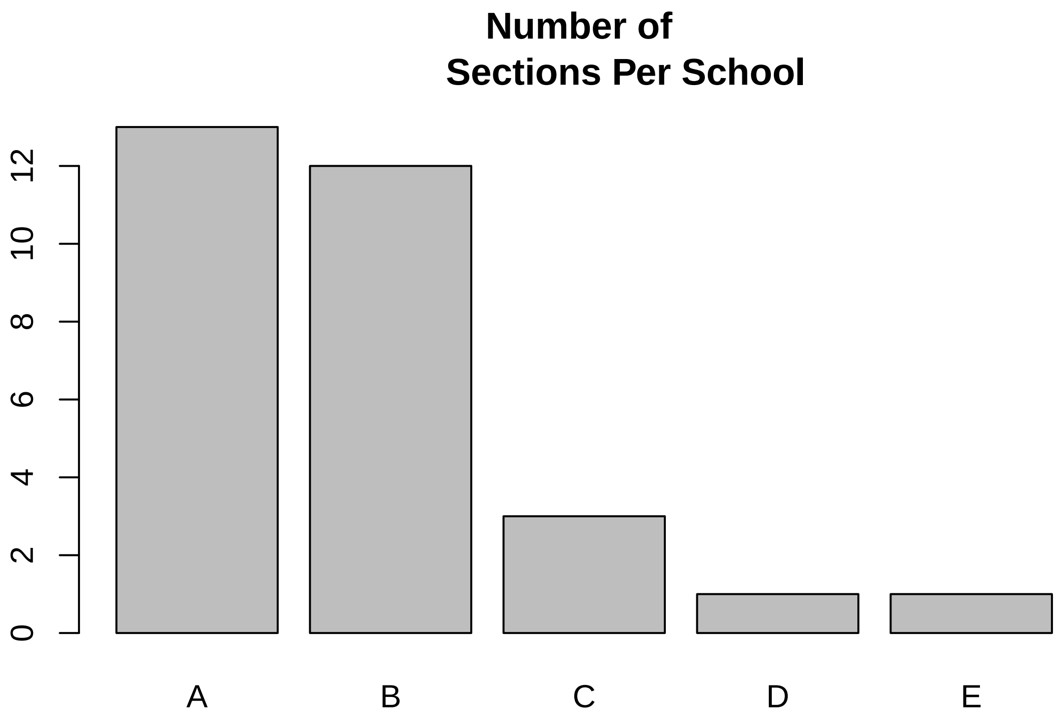
```
## # A tibble: 6 x 8
##   Section School Completed ‘Very Ahead +5‘ ‘Middling +0‘ ‘Behind -1-5‘
##   <fct>   <fct>      <int>           <int>         <int>         <int>
## 1 1       A             10               0             5            54
## 2 2       A              6               0             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
## # i 2 more variables: ‘More Behind -6-10‘ <int>, ‘Very Behind -11‘ <int>
```

## Exploratory Data Analysis and Data Visualization

It can be classified into two groups - ‘Very Ahead’ and ‘Completed’ - to be considered ahead in any manner. On the other hand, three categories are in place to describe ‘behind,’ which may result in responses that are biased towards the overall ‘behind’ category instead of ‘generally ahead.’

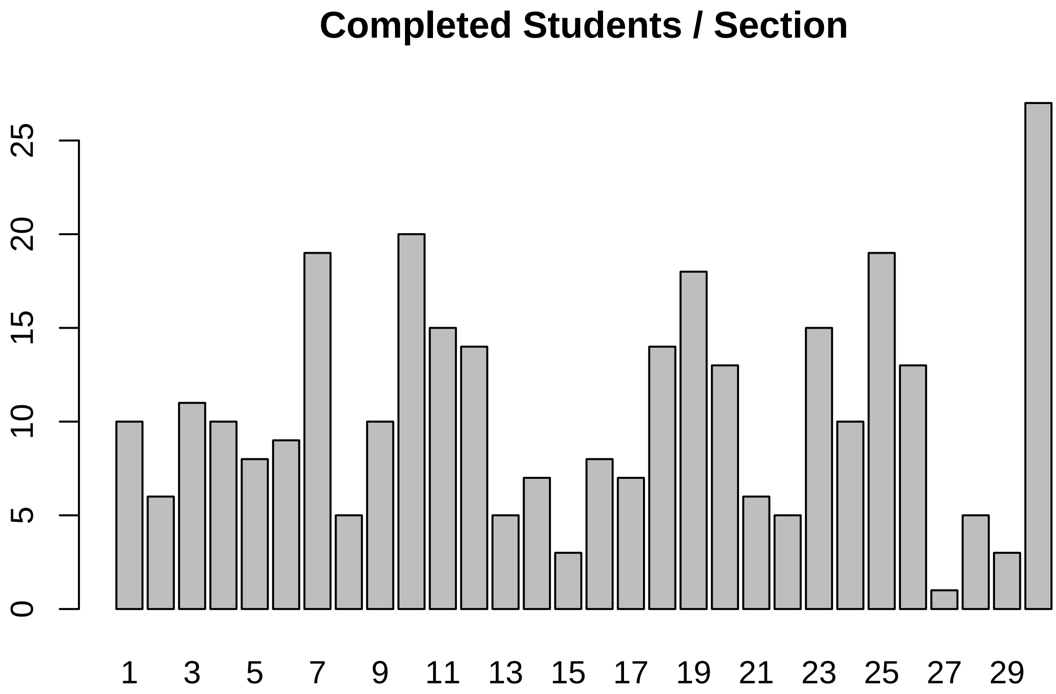
```
#Generate a bar chart that displays the frequency of sections per school.
SectionFreq<-c(length(which(storyteller2$School=='A')),
               length(which(storyteller2$School=='B')),
               length(which(storyteller2$School=='C')),
               length(which(storyteller2$School=='D')),
               length(which(storyteller2$School=='E')))

barplot(SectionFreq, names.arg = c('A', 'B', 'C', 'D', 'E'), main='Number of
        Sections Per School')
```



```
#Visualizing the data by creating a plot of the sections and Completed categories
#and then providing a summary of the results

barplot(storyteller2$Completed, main='Completed Students / Section',
        names.arg = c(1:30))
```


Completed Students / Section

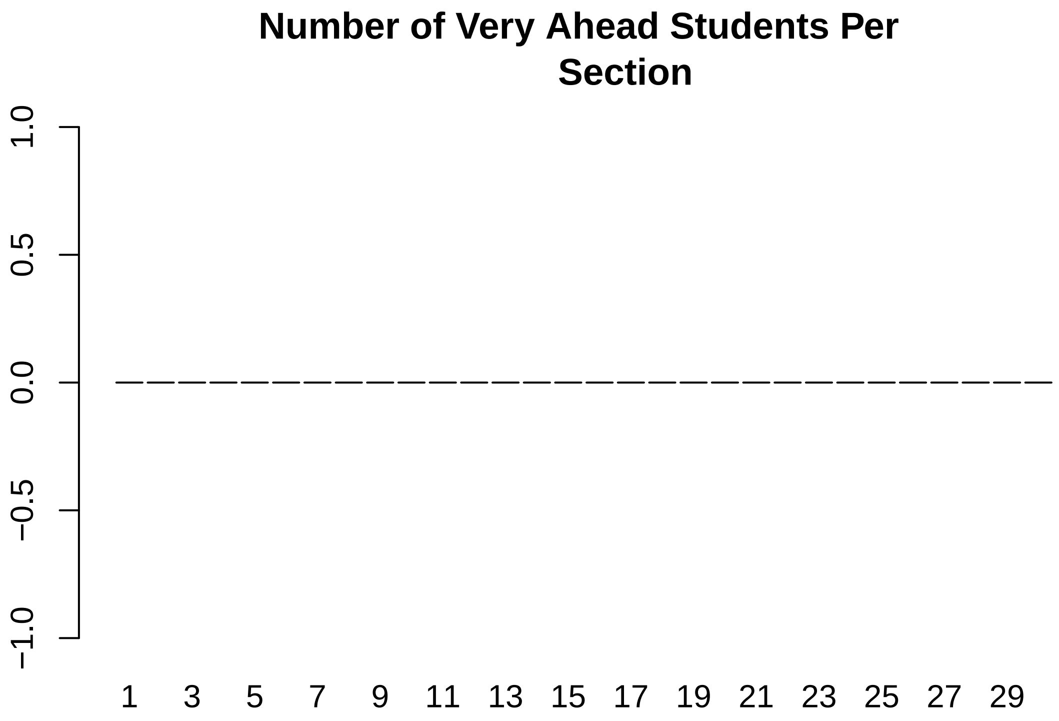
```
summary(storyteller2$Completed)
```

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

## Further Data Cleaning based on Exploratory Data Analysis and Data Visualization

```
#Visualizing the data by creating a plot of the Sections and Very Ahead categories
#and then providing a summary of the results

barplot(storyteller2$`Very Ahead +5`, main='Number of Very Ahead Students Per
        Section', names.arg = c(1:30))
```

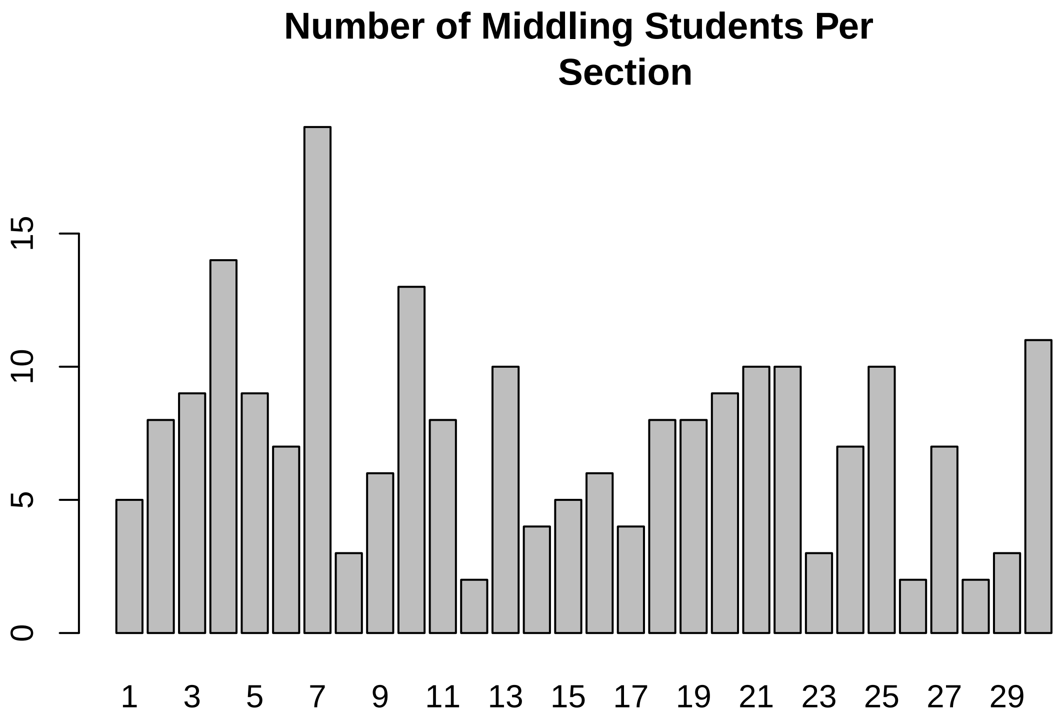


```
summary(storyteller2$`Very Ahead +5`)
```

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

```
#Visualizing the data by creating a plot of the Sections and Middling categories
#and then providing a summary of the results


barplot(storyteller2$`Middling +0`, main='Number of Middling Students Per
        Section', names.arg = c(1:30))
```

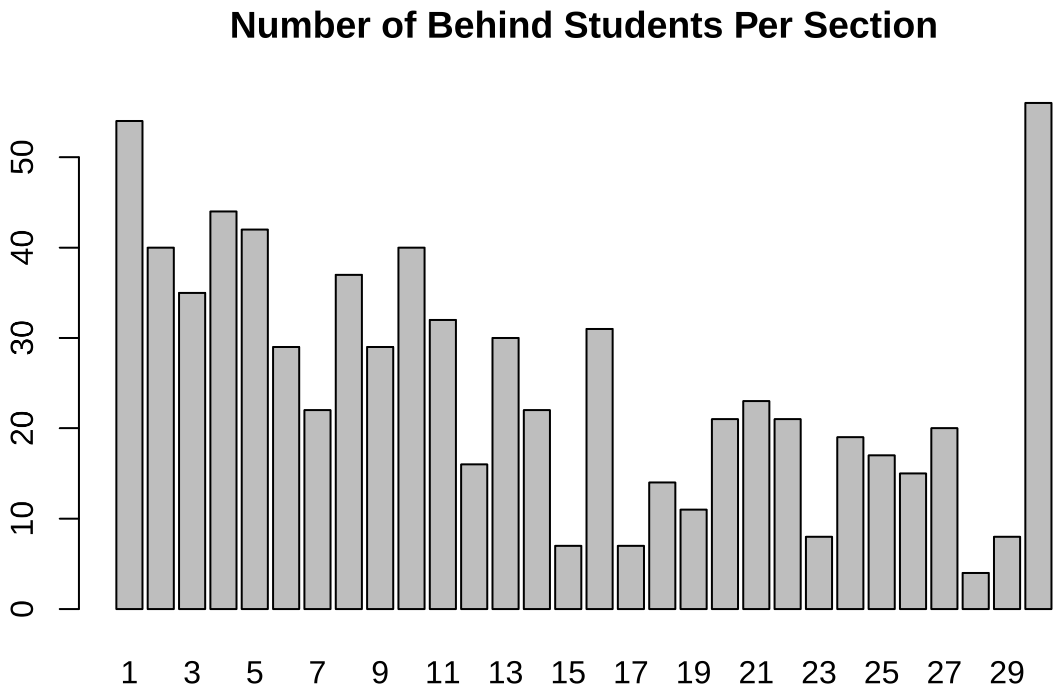


```
summary(storyteller2$`Middling +0`)
```

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

```
#Visualizing the data by creating a plot of the Sections and Behind categories
#and then providing a summary of the results


barplot(storyteller2$`Behind -1-5`, main='Number of Behind Students Per Section',
        names.arg = c(1:30))
```

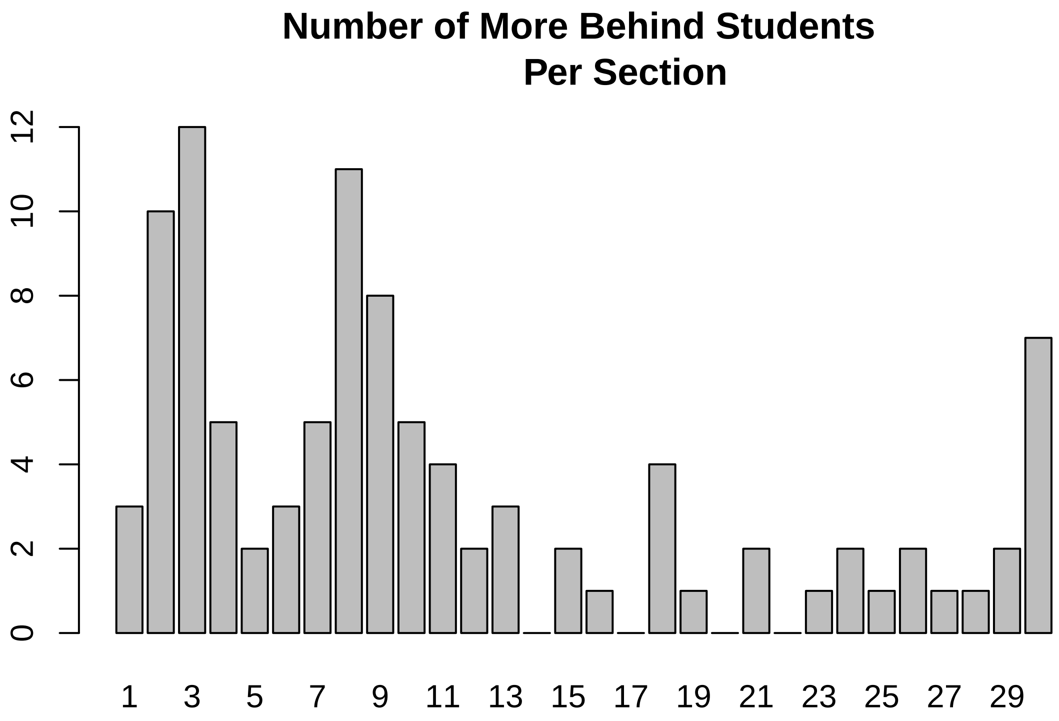


```
summary(storyteller2$`Behind -1-5`)
```

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

```
#Visualizing the data by creating a plot of the Sections and More Behind
#categories and then providing a summary of the results

barplot(storyteller2$`More Behind -6-10`, main='Number of More Behind Students
        Per Section', names.arg = c(1:30))
```

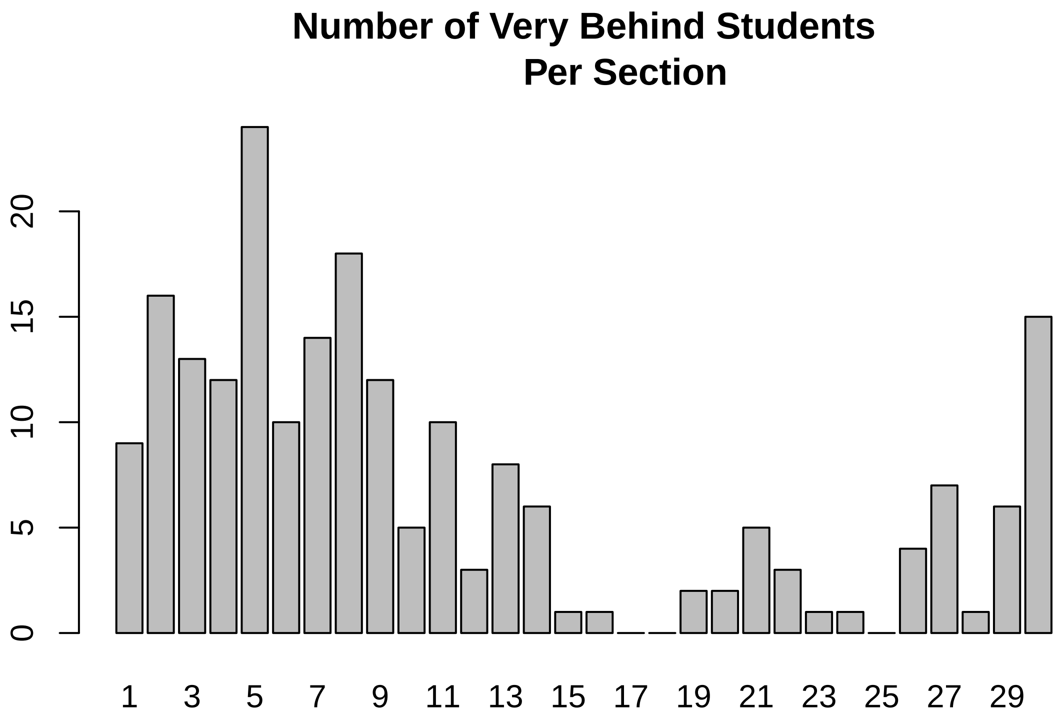


```
summary(storyteller2$`More Behind -6-10`)
```

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

```
#Visualizing the data by creating a plot of the Sections and Very Behind
#categories and then providing a summary of the results

barplot(storyteller2$`Very Behind -11`, main='Number of Very Behind Students
        Per Section', names.arg = c(1:30))
```



```
summary(storyteller2$`Very Behind -11`)
```

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

```
#Calculating the total count of students in each category

StudentCounts<-colSums(storyteller2[,3:8])

sum(StudentCounts)
```

```
## [1] 1601
```

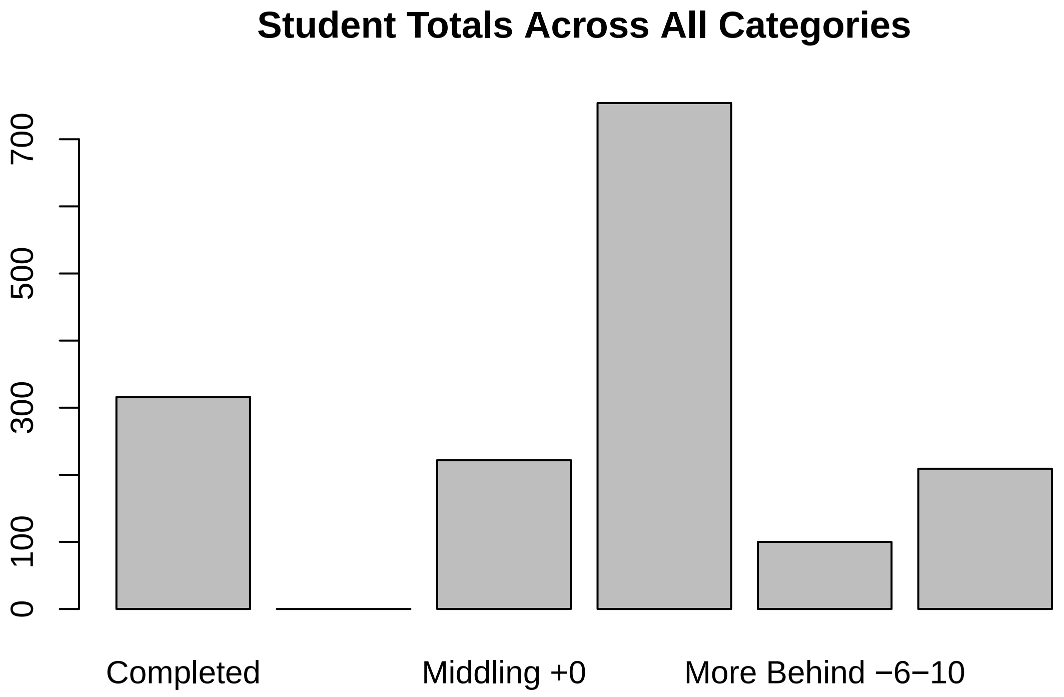
```
#Calculating the total count of students in each section

SectionCounts<-rowSums(storyteller2[,3:8])

#data.frame(SectionCounts)

#Generating a barplot to visualize the frequency distribution

StudentCounts<-colSums(storyteller2[,3:8])
barplot(StudentCounts, main="Student Totals Across All Categories")
```



Please note that there is a gap between the 'Middling' and 'Completed' categories. As a result, only those students who have finished the program can be considered 'Ahead'. This creates a potential bias in the survey results as respondents must choose between 'Middling' or 'Very Ahead', without an option in between. However, it is reasonable to assume that respondents would provide honest answers and select 'Middling' if appropriate.

It's worth mentioning that the data appears to be heavily skewed towards the 'Behind' category, with a significant margin.

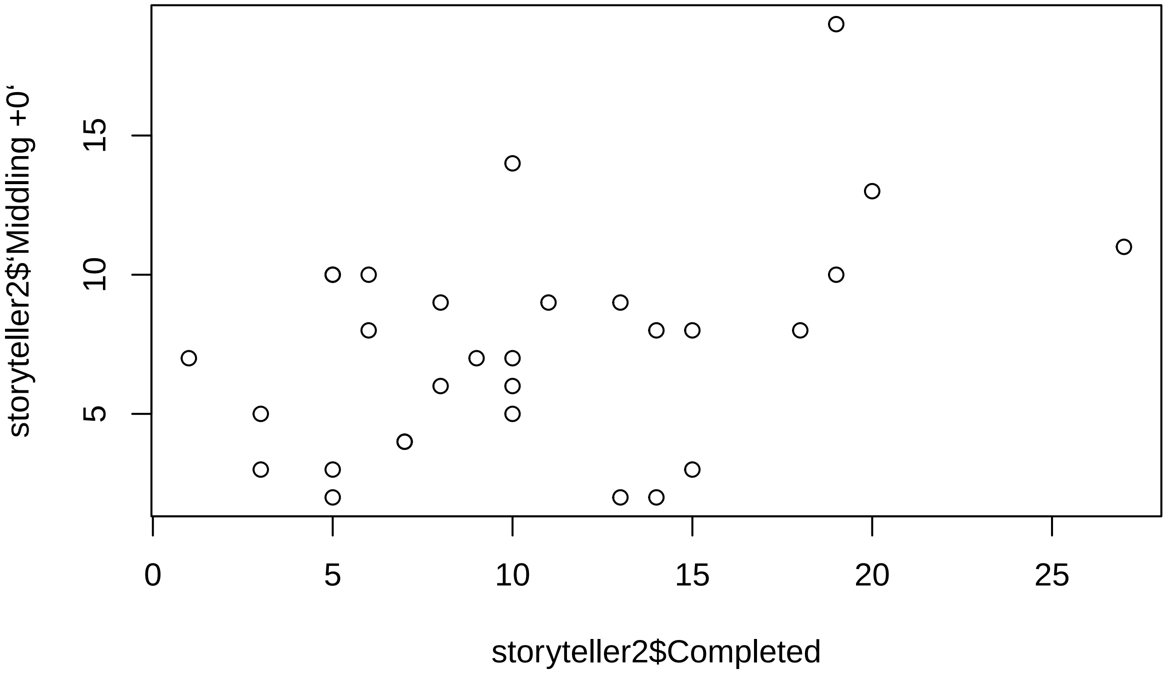
## Additional Exploratory Data Analysis

```
StudentCounts/sum(StudentCounts)
```

```
##         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
```

Observations: - The percentage of students who have completed the program is 19.74%. - There are no students who fall into the 'Very Ahead' category. - The percentage of students in the 'Middling' category is 13.87%. - The majority of students (47.1%) fall into the 'Behind' category. - A significant number of students (13.05%) are categorized as 'Very Behind'. - There is a gradual decrease in the percentage of students as the categories move towards 'Very Ahead'. However, there is a sharp increase in the percentage of students in the ‘Behind’ category compared to the ‘Middling’ category. - Overall, the data shows that a large proportion of students are struggling, with a majority falling into the ‘Behind’ and ‘Very Behind’ categories.

```
plot(storyteller2$Completed, storyteller2$`Middling +0`)
```



We can see that as the value of Completed increases, there is a tendency for the values of Middling +0 to also increase. However, there are also many data points that do not follow this trend, indicating that there may be other factors influencing the performance of the students in this school.

We may want to perform further analysis to understand the relationships between these variables and other variables in the dataset, such as Behind -1-5, More Behind -6-10, and Very Behind -11. Additionally, we may want to consider whether there are any differences in performance between the different schools in the dataset.

Let’s analyze the data by individual schools to gain a better understanding of the data

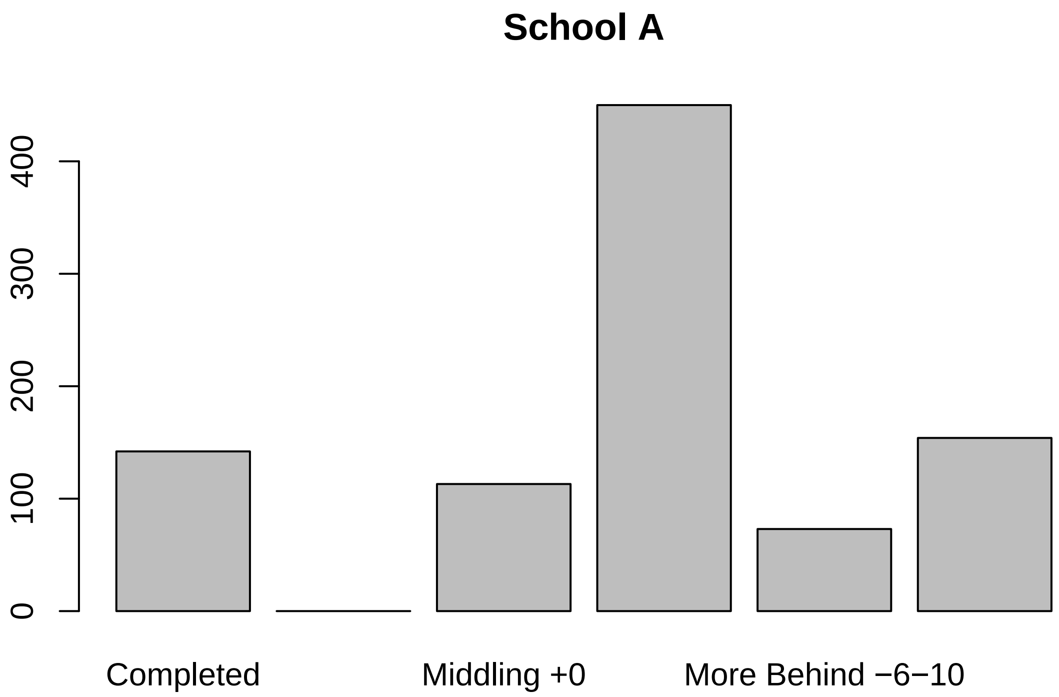
```
#Filtering the data by school and then calculating the total number of students
#in each of the performance categories: Very Ahead +5, Middling +0, Behind -1-5,
#More Behind -6-10, and Very Behind -11.

schoolA<-storyteller2[which(storyteller2$School == "A"),]
schoolB<-storyteller2[which(storyteller2$School == "B"),]
schoolC<-storyteller2[which(storyteller2$School == "C"),]
schoolD<-storyteller2[which(storyteller2$School == "D"),]
schoolE<-storyteller2[which(storyteller2$School == "E"),]
```

```
StudentCountsA<-colSums(schoolA[3:8])
StudentCountsA
```

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

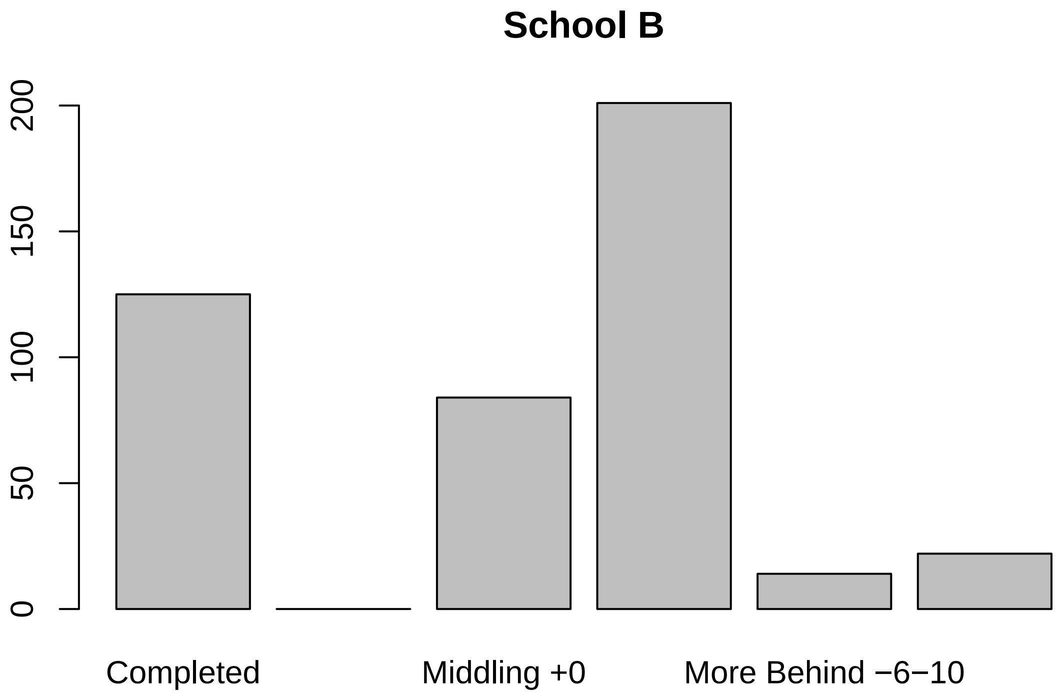
```
barplot(StudentCountsA, main = "School A")
```



```
StudentCountsB<-colSums(schoolB[3:8])
StudentCountsB
```

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

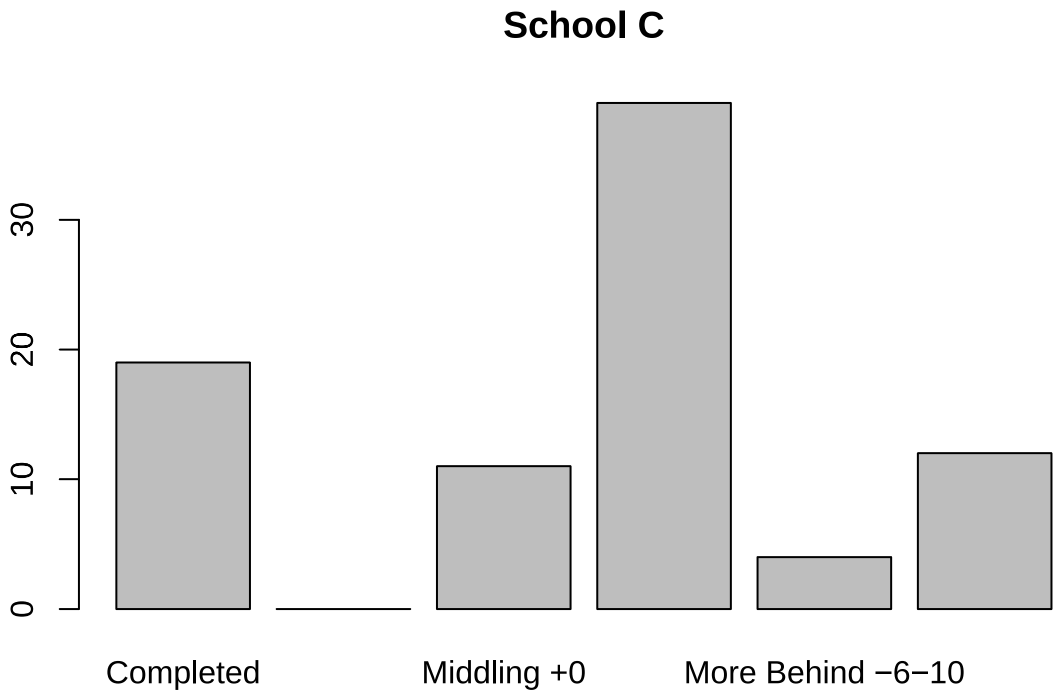
```
barplot(StudentCountsB, main = "School B")
```



```
StudentCountsC<-colSums(schoolC[3:8])
StudentCountsC
```

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

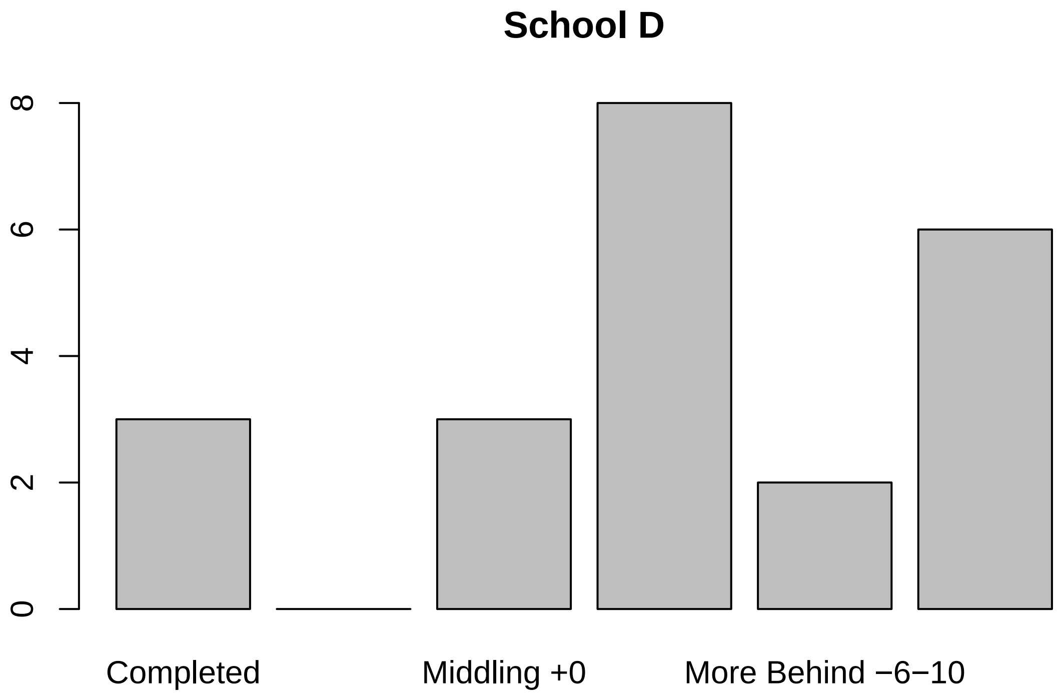
```
barplot(StudentCountsC, main = "School C")
```



```
StudentCountsD<-colSums(schoolD[3:8])
StudentCountsD
```

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

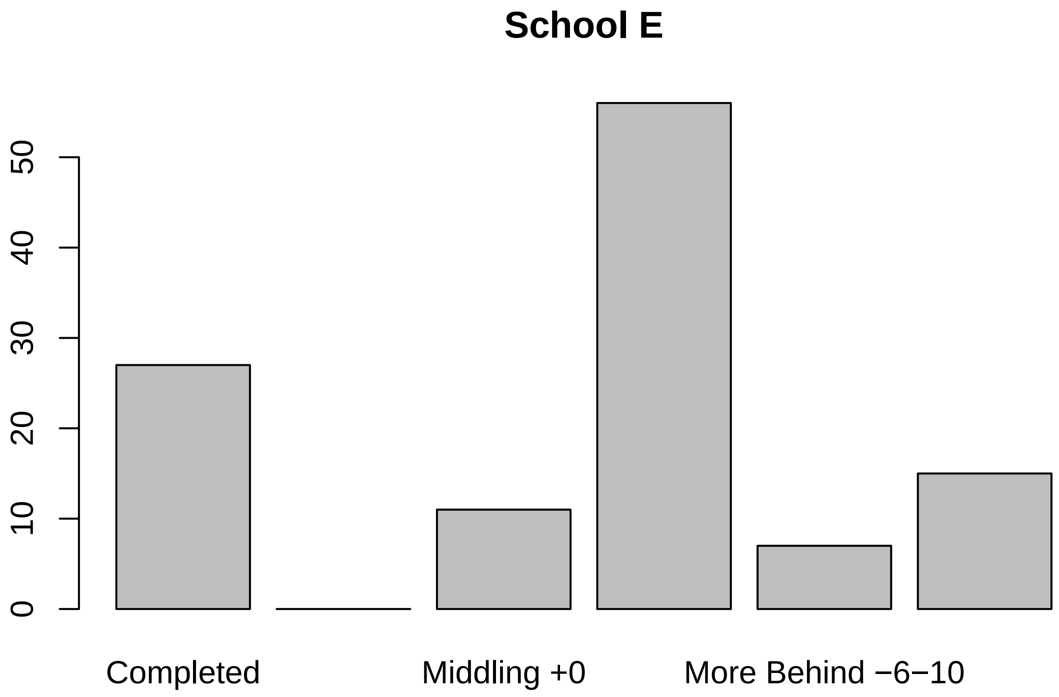
```
barplot(StudentCountsD, main = "School D")
```



```
StudentCountsE<-colSums(schoolE[3:8])
StudentCountsE
```

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

```
barplot(StudentCountsE, main = "School E")
```



## Initial Observations and Remarks

It’s clear that School A has the highest number of students who have completed the course and the highest number of students that are behind (defined as those falling in the -1 to -10 range). This suggests that School A may have a larger number of students and/or a more demanding curriculum that leaves less time for studies.

School B has the second-highest number of students who have completed the course but the lowest number of students that are behind. This suggests that School B may have a smaller number of students or a more flexible curriculum that allows for more time to work on schoolwork.

Schools C, D, and E each have a much lower number of students who have completed the course and tend to have a higher proportion of students that are behind schedule. This could suggest that these schools have fewer resources or less emphasis on time to work on the course in their curriculum.

Overall, these observations could be used to guide further investigation into the factors that contribute to student success in this course, such as class size, curriculum, resources, and instructor support.

```r
#Calculating the proportion of students from schools B and D relative to the
#total number of students
sum(StudentCountsB)/sum(StudentCounts)
```

```
## [1] 0.2785759
```

```
sum(StudentCountsD)/sum(StudentCounts)
```

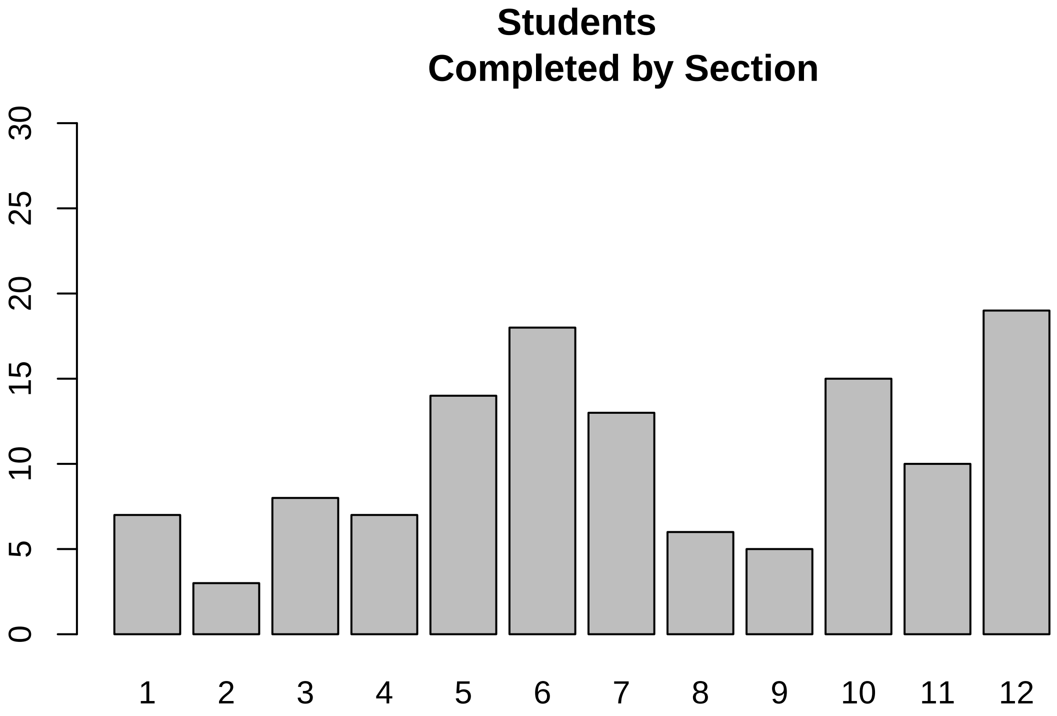
```
## [1] 0.01374141
```

School B has around 27.86% of the total students in the dataset, while School D has only around 1.37% of the total students.

Note: Other factors such as the location, demographics, and socio-economic status of each school's students may play a significant role in the differences observed between schools. Further analysis and exploration may be necessary to understand the underlying factors driving these differences.

```
#Analyzing the data to identify high-quality sections

barplot(schoolB$Completed, names.arg =c(1:12),ylim=c(0,30), main = "Students
        Completed by Section")
```



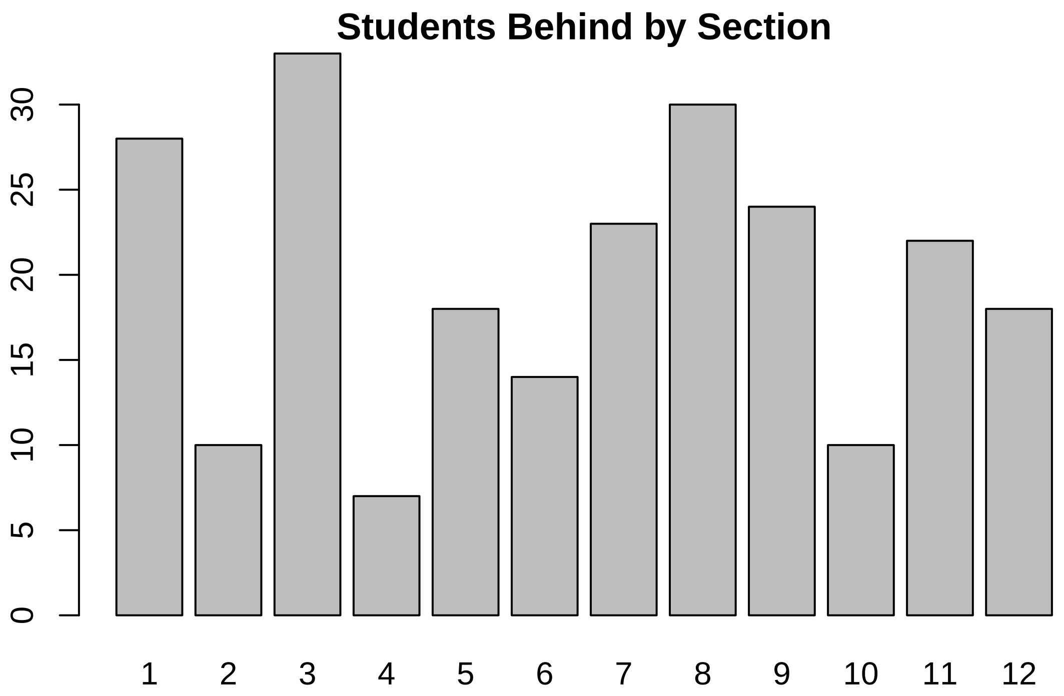
In school B, sections 6, 10, and 12 exhibit the most encouraging outcomes.

```
#Summing up the values in columns 3 to 8 (the counts for each level of completion)
#for each row in schoolB, which represents the total number of students in each
#section

rowSums(schoolB[,3:8])
```

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

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



The barplot suggests that there is a significant proportion of students who are struggling in school B, particularly in the "Behind" category. This information could be used to inform interventions or targeted support programs for these students. Additionally, it may be worth investigating the reasons behind the higher number of students in the "Behind" category in certain sections (such as sections 1, 3, and 8) and exploring ways to improve the educational experience for those students.

There are variations in the performance of the five schools. Schools A and B have the highest number of completed students, whereas Schools C, D, and E have a relatively low number of completed students. Additionally, Schools A and B have more students who are ahead or in the middle, while Schools C, D, and E have more students who are behind or very behind.

Analyzing the data by section within School B revealed that sections 6, 10, and 12 had the highest number of completed students, which suggests that these sections may be of higher quality. On the other hand, analyzing the data by section within School B also revealed that there are some sections with a high number of students who are behind or very behind, indicating that there may be issues with the quality of teaching in those sections.

Overall, the data analysis highlights the need for further investigation into the factors that contribute to the variations in student performance between the schools and the sections within the schools. It is also important to consider additional data such as teacher performance and student demographics to gain a more complete understanding of the factors that contribute to student success.