George_Smith_HW2

George Smith

7/19/2021

installs / load packages

```
#install.packages("caret")
#install.packages("tidyverse")
#install.packages("stargazer")

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(forcats)
library(ggplot2)
library(dplyr)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(modelr)
```

set the working directory and read in the file

```
setwd("c:/Users/GeorgeSmith/Documents")
storyteller<- read_csv("data-storyteller.csv", na = c(""))
```

```
##
## -- 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)
```

```
## spec_tbl_df[,8] [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()
##   .. )
```

Data Cleaning

Note: The School column is of the character type and it should be a factor. Other notes below in comments.

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

#The section column is of the Numeric type and should be 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

#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

#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>
```

EDA and Data Viz

Observation: To be considered 'ahead' in any way, there are two categories:

'Very Ahead' and 'Completed'. There are 3 cateories in place to describe

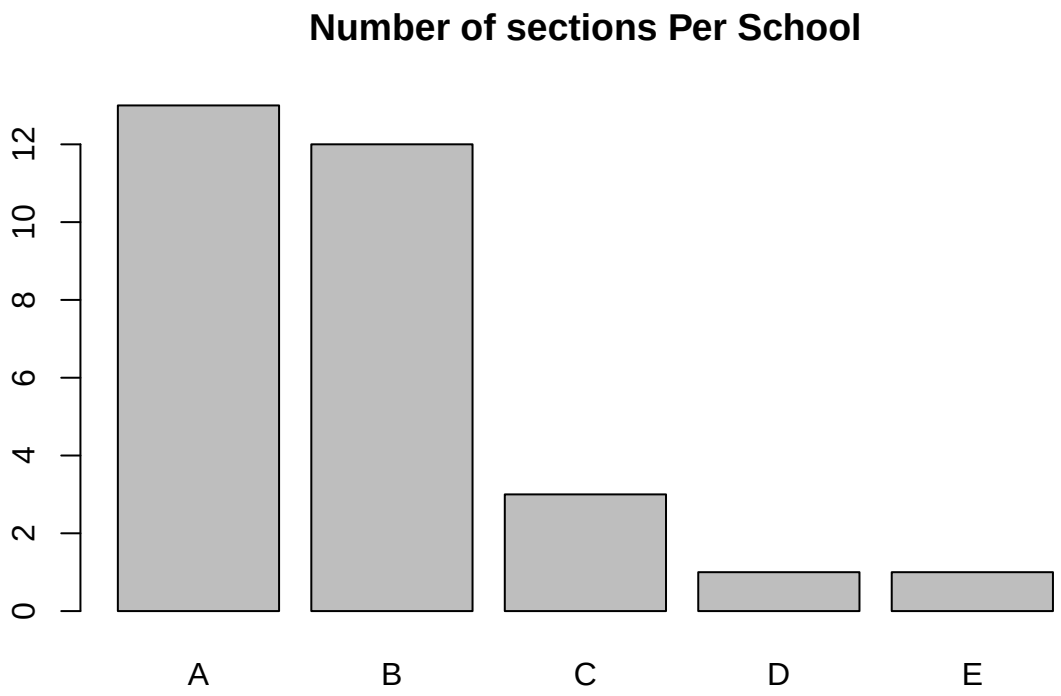
'behind' which may lead the responses to be lopsided in favor of generally

'behind' over 'generally ahead'. What other observations have you made?? Note

them and clean the data appropriately.

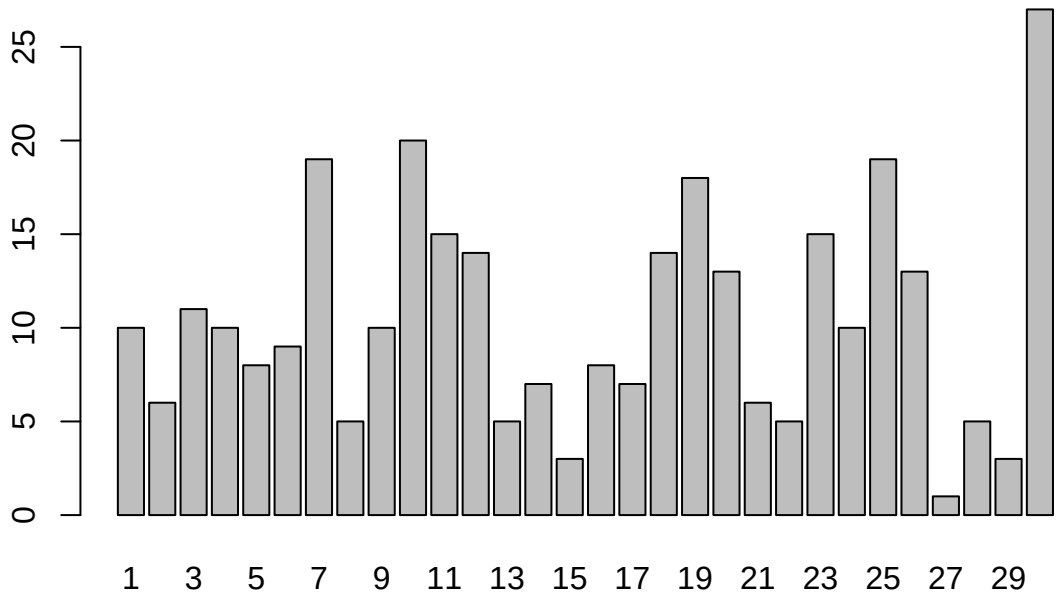
```
# Creating a bar chart to show the number of sections from each school
SchoolValues<-c(length(which(storyteller$School=='A')), length(which(storyteller$School=='B')), length(

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
```

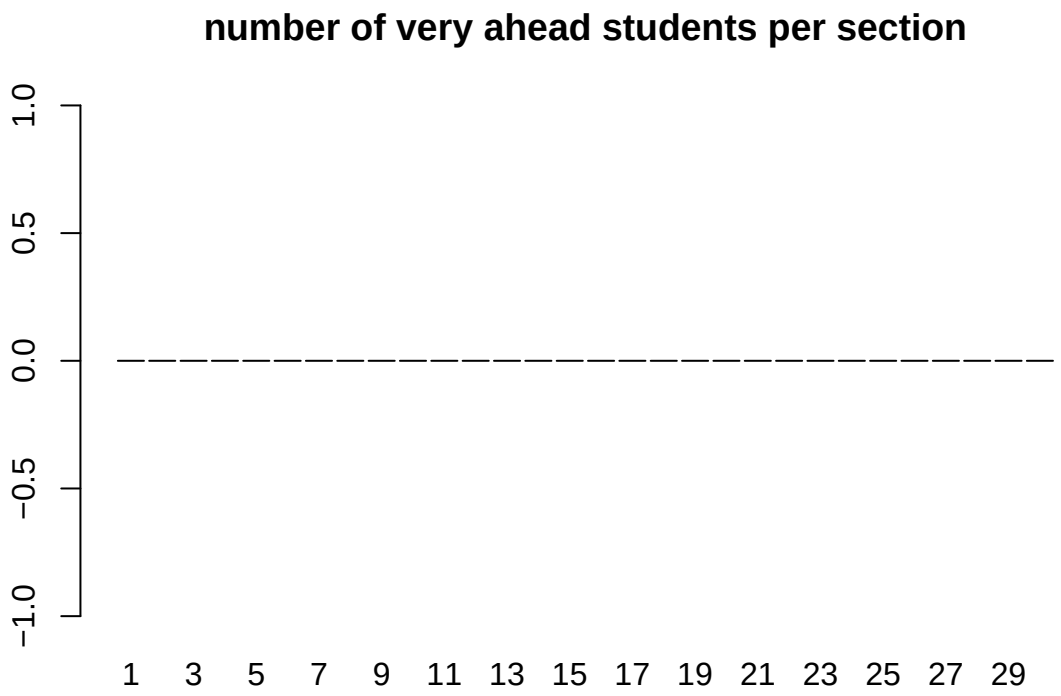
Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 6.00 10.00 10.53 14.00 27.00

Further Data Cleaning based on EDA and Viz

#Plotting section and Very Ahead and summarizing

```
barplot(storyteller$`Very Ahead +5`, main='number of very ahead students per section', names.arg = c(1:3
```

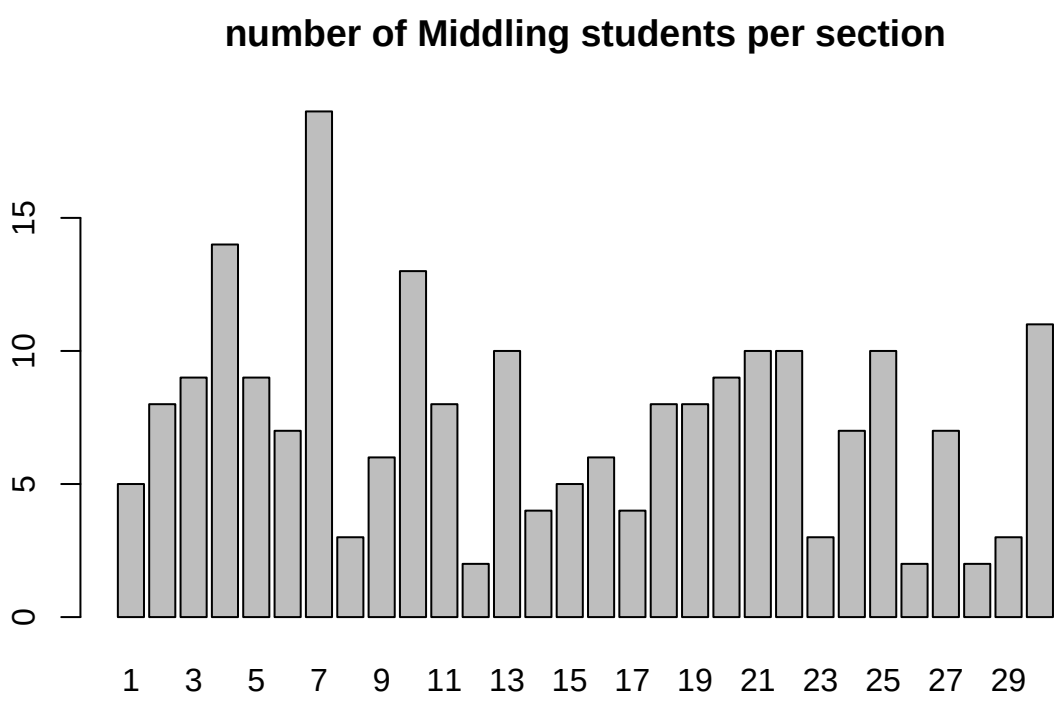



```
summary(storyteller$`Very Ahead +5`)
```

```
##    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))
```

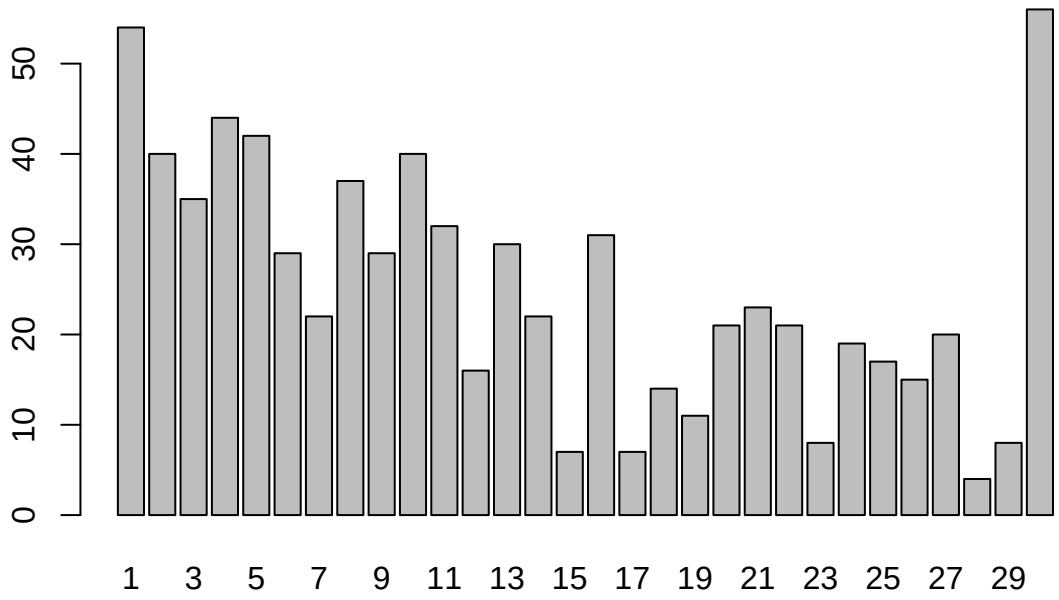



```
summary(storyteller$`Middling +0`)
```

```
##    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))
```

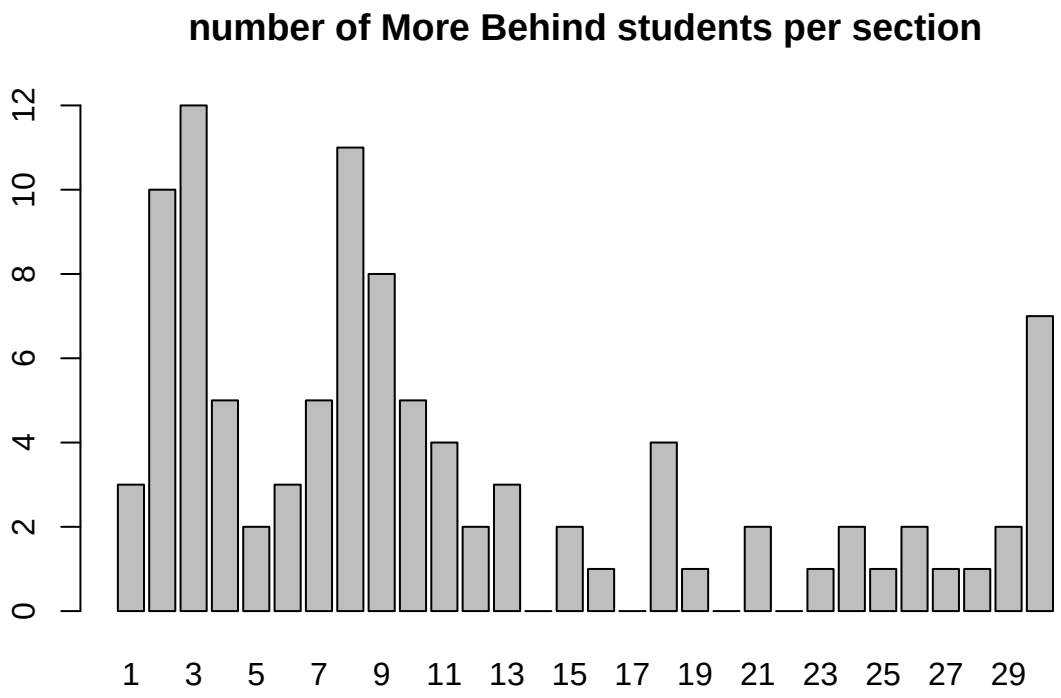



```
summary(storyteller$`Behind -1-5`)
```

```
##    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 =
```

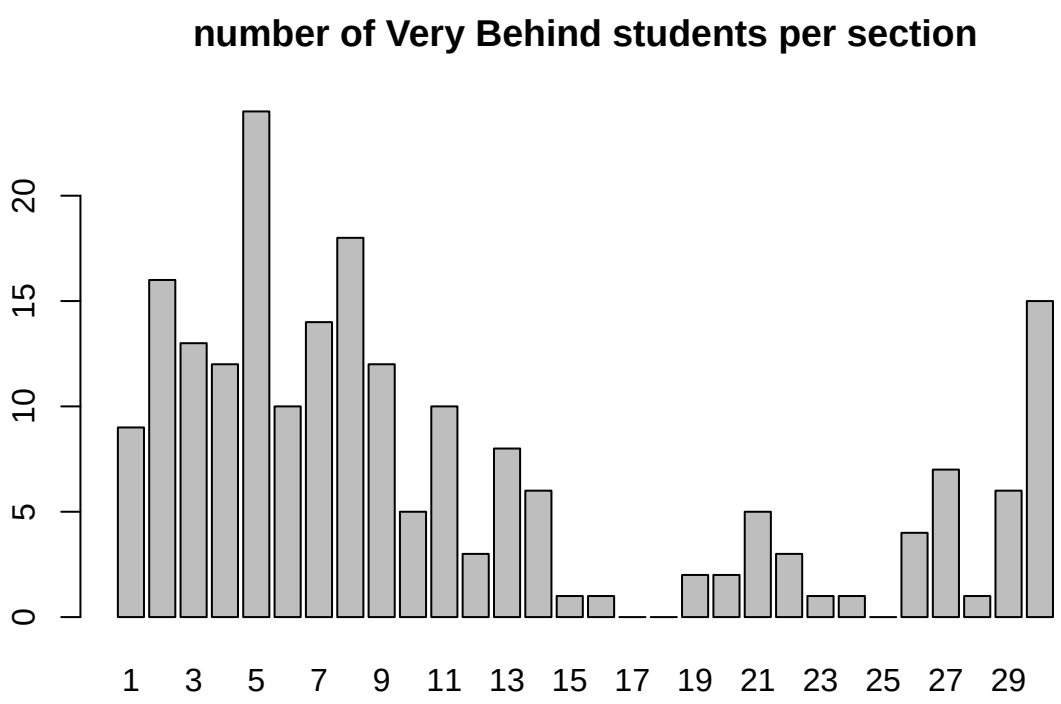



```
summary(storyteller$`More Behind -6-10`)
```

```
##    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
```




```
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

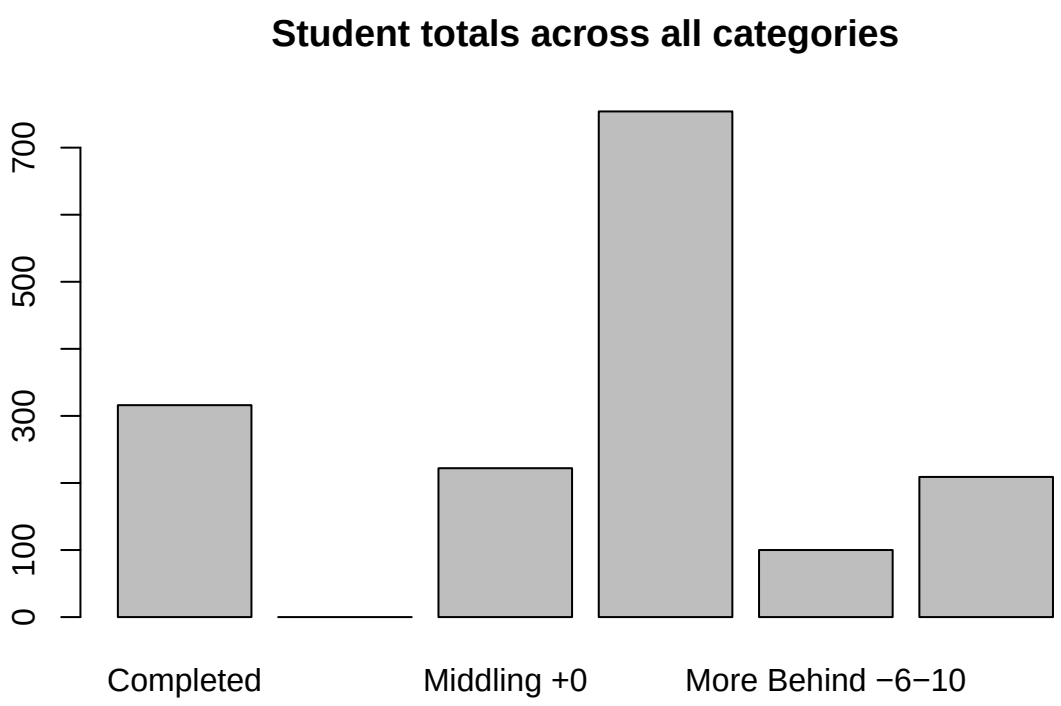
```
SectionSums<-rowSums(storyteller[,3:8])
data.frame(SectionSums)
```

```
##    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")
```



Note: There is a gap between middling and completed, meaning that the only # students who can be described #as 'ahead' have already finished the program. # # There is no category that describes between 'Middling' and 'Very Ahead'. This # means that survey respondents would have to decide between 'middling' or 'Very # Ahead'; should they fall in the middle? This may skew the data towards one # side or the other, however (Assumption) it is reasonable to assume that survey # respondents would stick to a more honest 'Middling Answer' rather than # exaggerating their students' success.

Further Note: The data seems to be centered on the 'Behind' category, and by a

large margin.

EDA (cont.)

```
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
```

More observations:

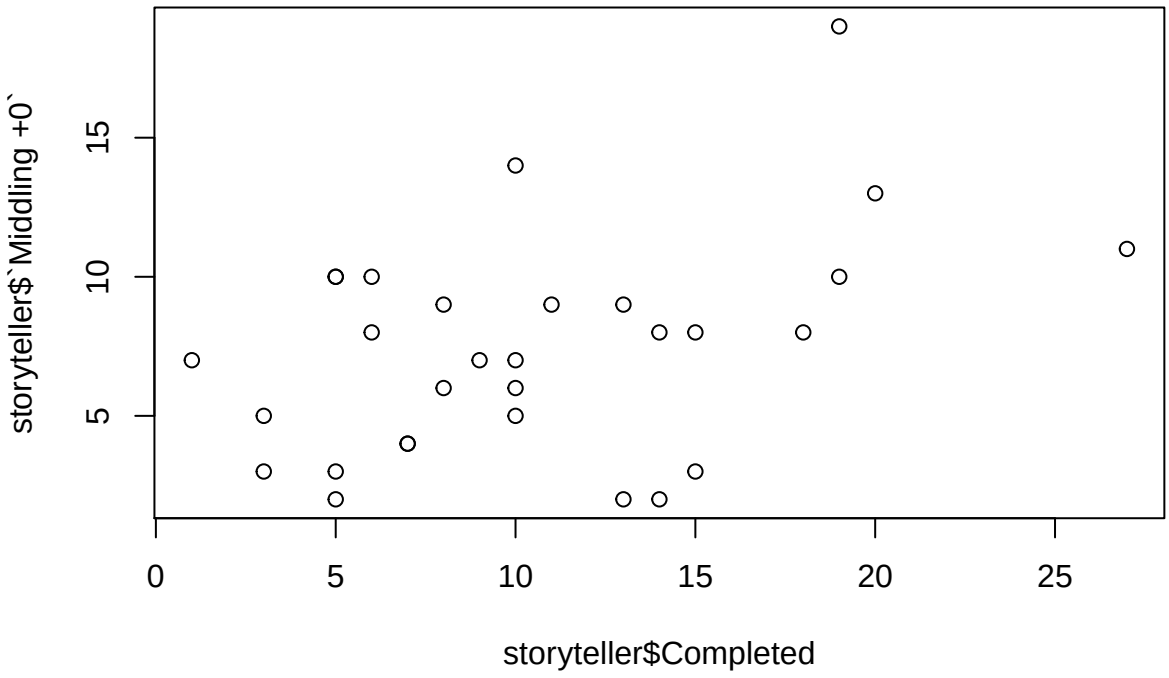
- 14% of students are on track in the middling category.

- 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 curicculum.

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



If we were to assume that students are generally not doing well because the

program is difficult, #there is is still a discrepancy with the number of

students that have completed the course ahead of schedule. Essentially, A

fifth of the students have nothing to work on for the remainder of the time.

So, lets break the data down into individual schools to see if we can better

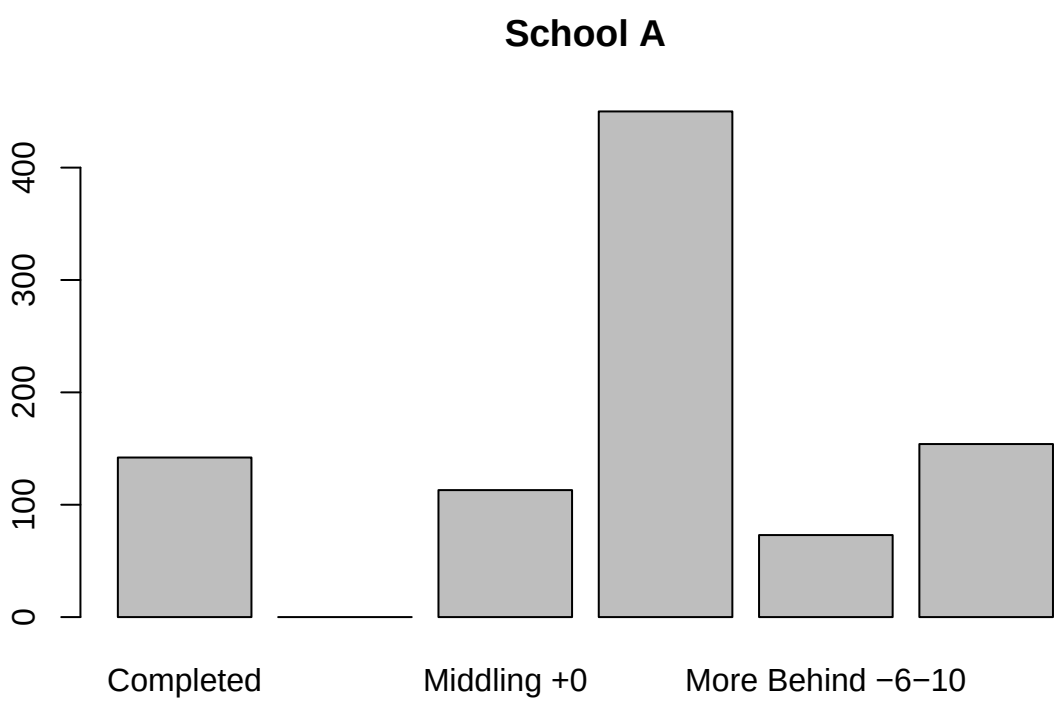
understand the data.

```
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
```

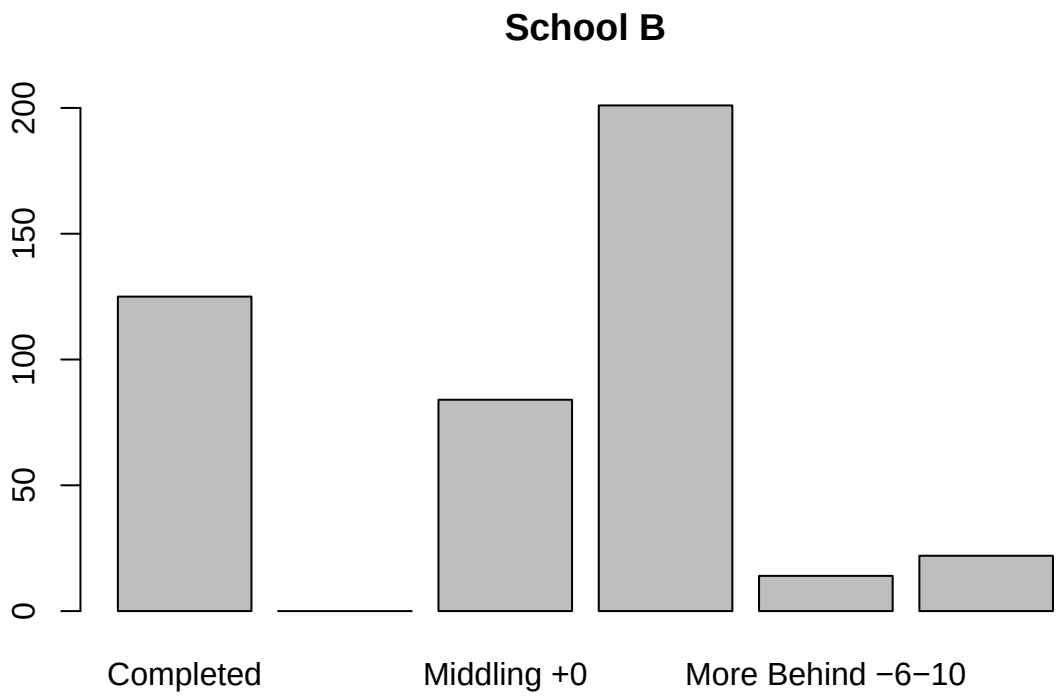
```
barplot(StudentSumsA, main = "School A")
```




```
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
```

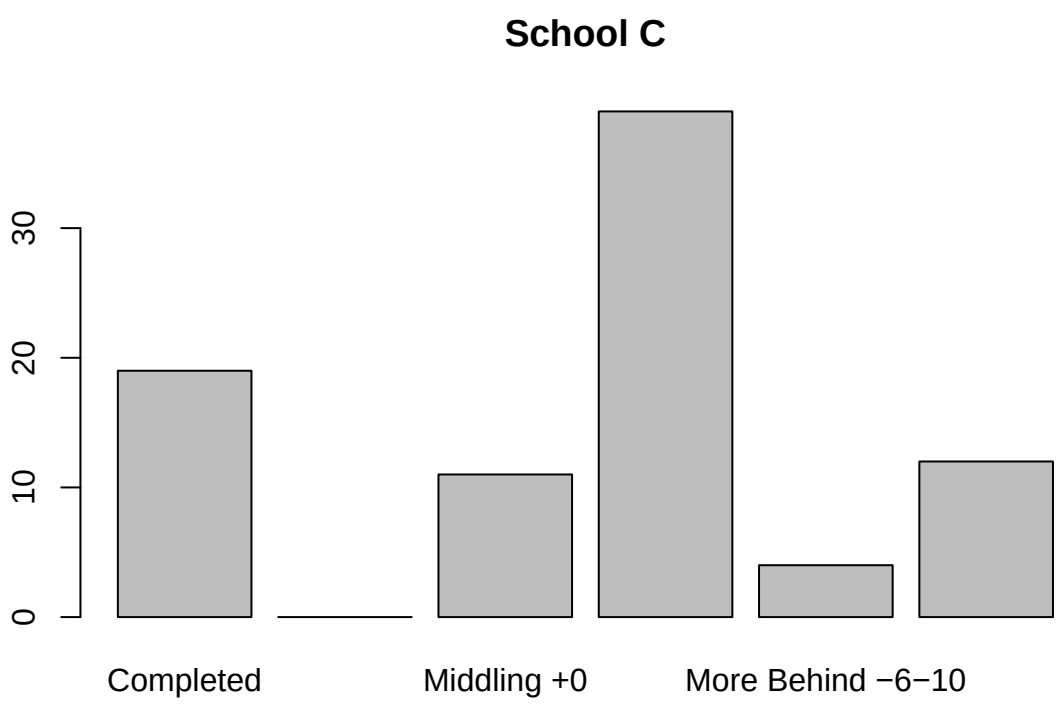
```
barplot(StudentSumsB, main = "School B")
```




```
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
```

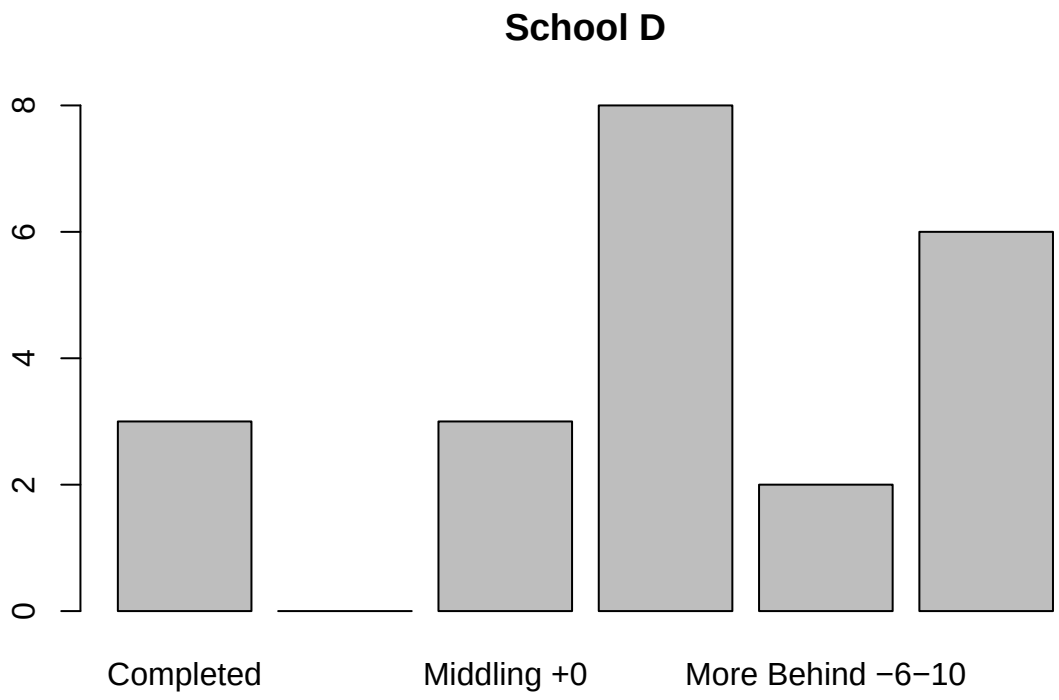
```
barplot(StudentSumsC, main = "School C")
```




```
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
```

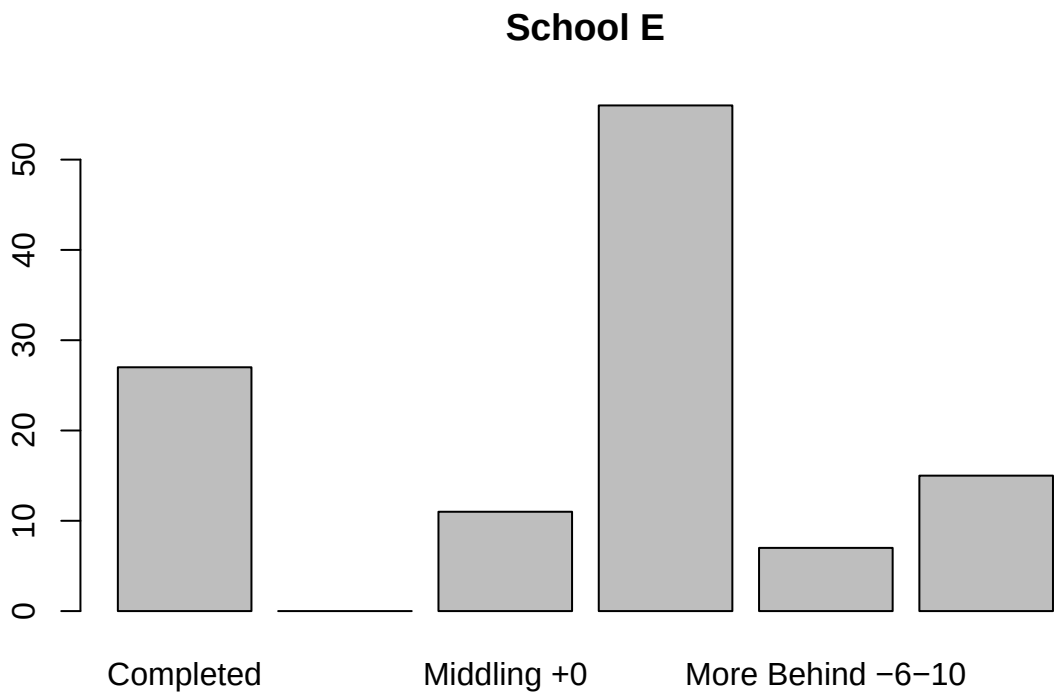
```
barplot(StudentSumsD, main = "School D")
```




```
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

Observe:

- 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
```

School D has a small number of students, meaning that the data from school D

has a greater probability of not being representative of the situation and may

be acconted for by random chance. School B is a sample that is approximately

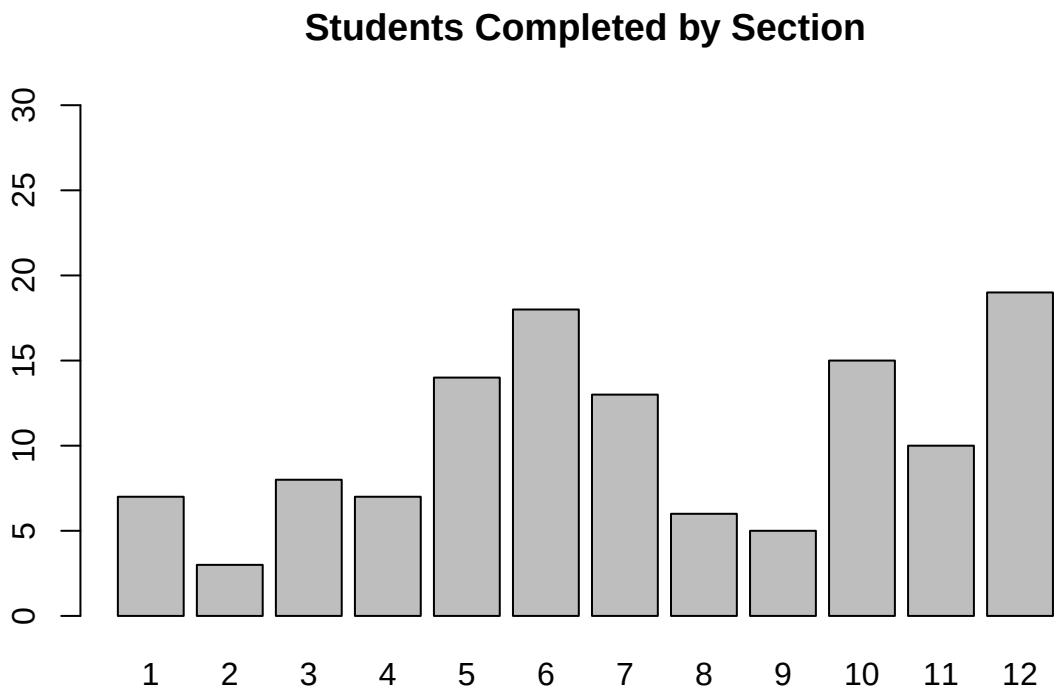
one quarter of the overall population, making it more representative of the

situation. It also shows that school B is doing something that the rest of the

schools are not in order to get students through the course.

#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"
```

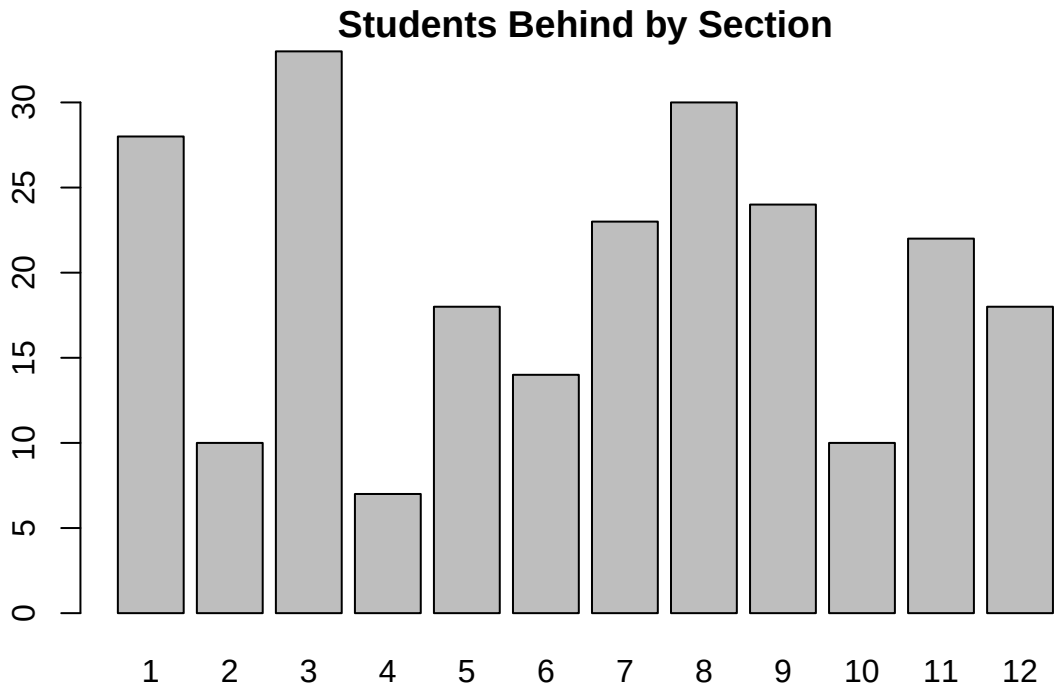



```
# Sections 6, 10, and 12 have the most promising results within school B.

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`, name
```


Students Behind by Section
0
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10
11
12