Sharat_Sripada_HW2

Helper functions

```
get_status_per_school <- function(mydata, school) {
    total <- sum(mydata$Very.Ahead..5, mydata$Middling..0,
mydata$Behind..1.5, mydata$More.Behind..6.10,
    mydata$Very.Behind..11, mydata$Completed)

    # Create a data-frame
    mydf <- data.frame(
        VeryAhead = (sum(mydata$Very.Ahead..5) / total * 100),
        Middle = (sum(mydata$Middling..0) / total * 100),
        Behind = (sum(mydata$Behind..1.5) / total * 100) ,
        More_Behind = (sum(mydata$More.Behind..6.10) / total * 100),
        Very_Behind = (sum(mydata$Very.Behind..11) / total * 100),
        Complete = (sum(mydata$Completed) / total * 100),
        Total = total)

    # Plot a bar-graph
    barplot(c(mydf$VeryAhead,
              mydf$Middle,
              mydf$Behind,
              mydf$More_Behind,
              mydf$Very_Behind,
              mydf$Complete),
        main = paste("Barplot for School-", school, "| Total-students: ",
mydf$Total),
        xlab = 'Categories',
        ylab = '%',
        col = 'blue', space=1,
        cex.names=0.45,
        ylim = c(0, 60),
        names.arg = c(colnames(mydf)[-7]),
        cex.main = 0.75
        )
    return(mydf)
}
```

Load the data

```
data <- read.csv('/Users/ssharat/Documents/Masters@Syracuse/Course-
Related(Study)/IST-707/week2_resources_2_2_2_2_2/data-storyteller.csv')
```

Examine the data

```
str(data)
```

```
## 'data.frame':    30 obs. of  8 variables:
##  $ School          : Factor w/ 5 levels "A","B","C","D",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ Section         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Very.Ahead..5   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Middling..0     : int  5 8 9 14 9 7 19 3 6 13 ...
##  $ Behind..1.5     : int  54 40 35 44 42 29 22 37 29 40 ...
##  $ More.Behind..6.10: int  3 10 12 5 2 3 5 11 8 5 ...
##  $ Very.Behind..11  : int  9 16 13 12 24 10 14 18 12 5 ...
##  $ Completed        : int  10 6 11 10 8 9 19 5 10 20 ...
```

R recognizes the data-type per our requirements accurately. No further work on data-types required here.

Check for non-existing data/values

```
# Check for non-existing values
length(which(is.na(data)))

## [1] 0
```

This shows no missing data.

Exploratory Data analysis

EDA-1: Reality check

Visualization of the overall status of the course across all schools. Now, in a utopian world, we would expect:

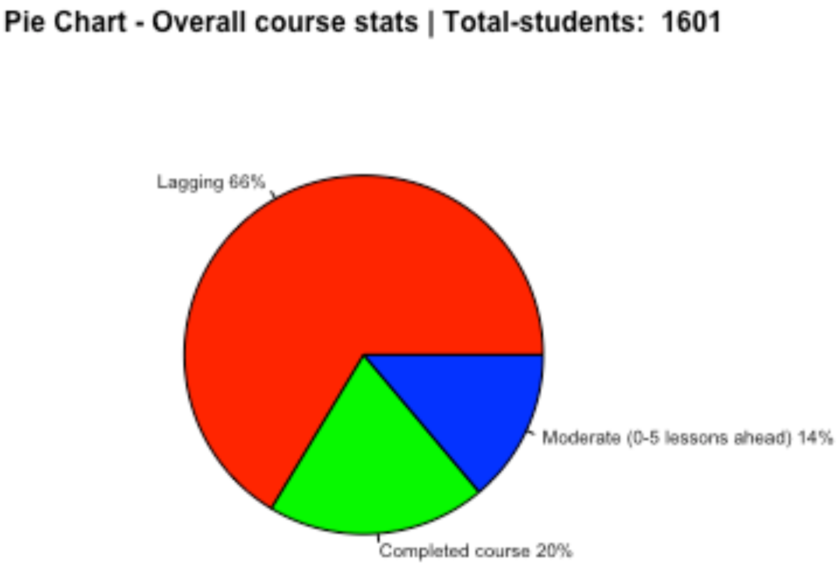
- no student to be lagging behind
- all or most students to be fairly ‘very ahead’ (that is more than 5 lessons ahead)
- AND perhaps a few students to have completed the course

Let’s do a reality check.

```
very_ahead <- sum(data$Very.Ahead..5)
complete_students <- sum(data$Completed)
middling_students <- sum(data$Middling..0)
lagging_students <- sum(data$Behind..1.5) + sum(data$More.Behind..6.10) +
                    sum(data$Very.Behind..11)
total_students <- sum(very_ahead, complete_students, middling_students,
                      lagging_students)

slices <- c(lagging_students, complete_students, middling_students)
lbls <- c( "Lagging", "Completed course", "Moderate (0-5 lessons ahead)")
pct <- round(slices / sum(slices) * 100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls, "%", sep="")
```

```
# Plot a pie-chart
pie(slices, labels = lbls, col = rainbow(length(lbls)),
   main = paste("Pie Chart - Overall course stats | Total-students: ",
total_students),
   cex = 0.5,
   cex.main = 0.75)
```



Initial data exploration or analysis shows, 66% of all students are lagging in one of the following, they are:

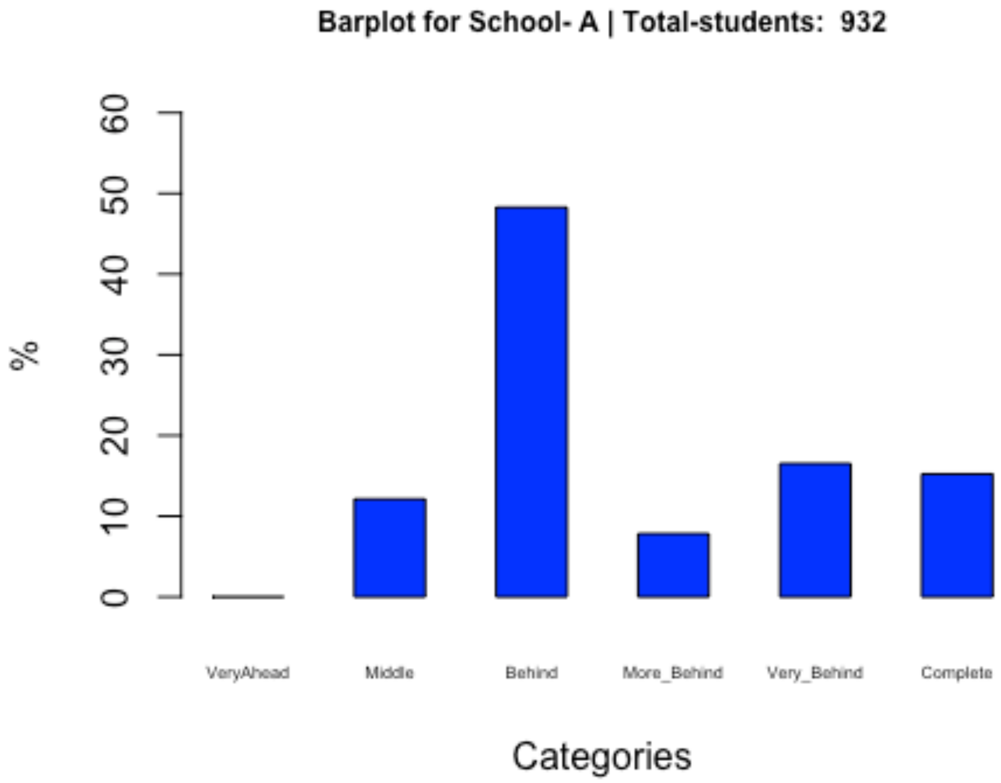
- behind by 1 to 5 lessons
- behind by 6 to 10 lessons
- OR behind by more than 10 lessons

EDA-2: Course standings grouped by school

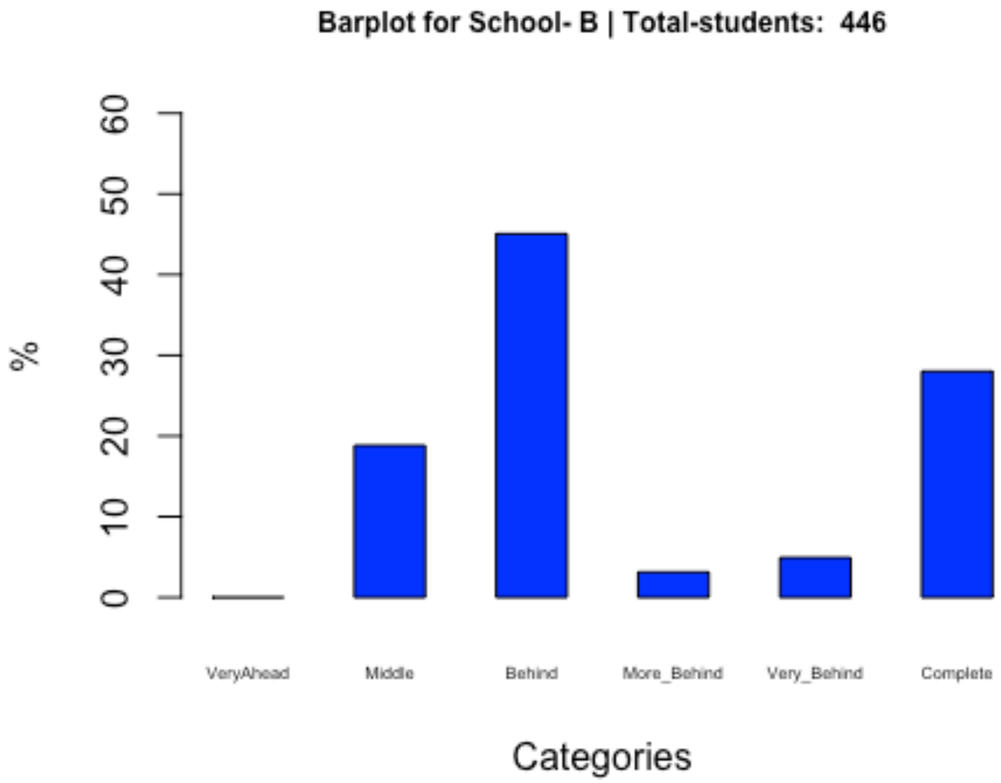
Diving deeper, to uncover trends of how the course is progressing across each school - A, B, C, D and E

```
# Get data per school
stats_by_school <- data.frame()
```

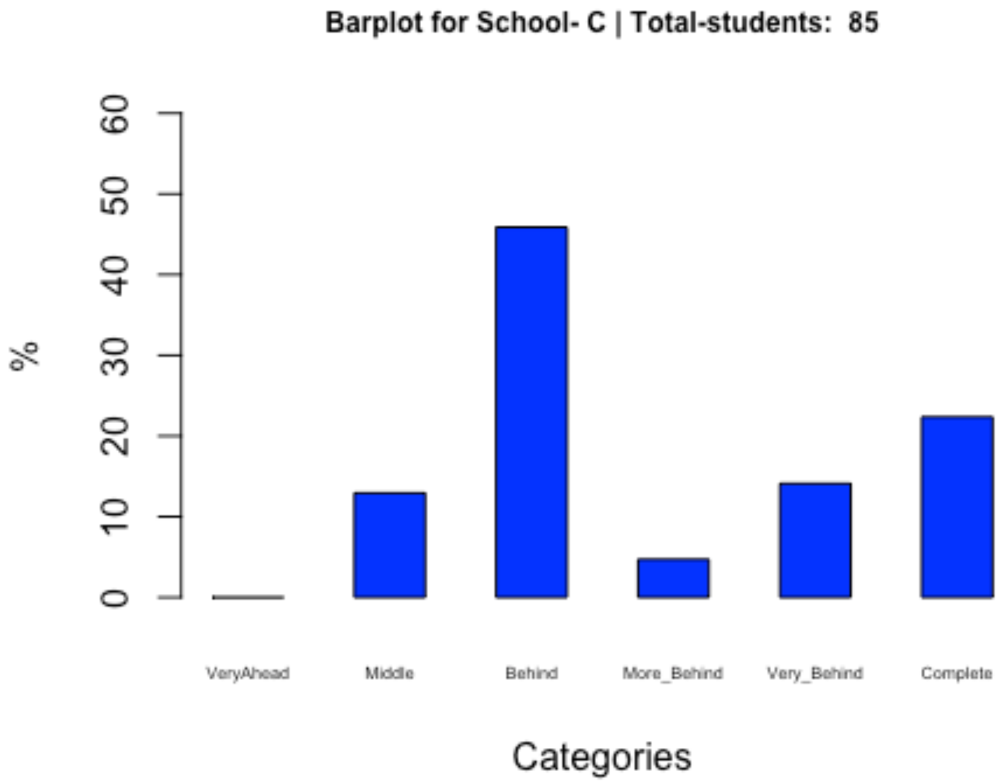
```
# For School-A:
data_school <- data[data$School=='A',]
stats_by_school <- rbind(stats_by_school, get_status_per_school(data_school,
'A'))
```



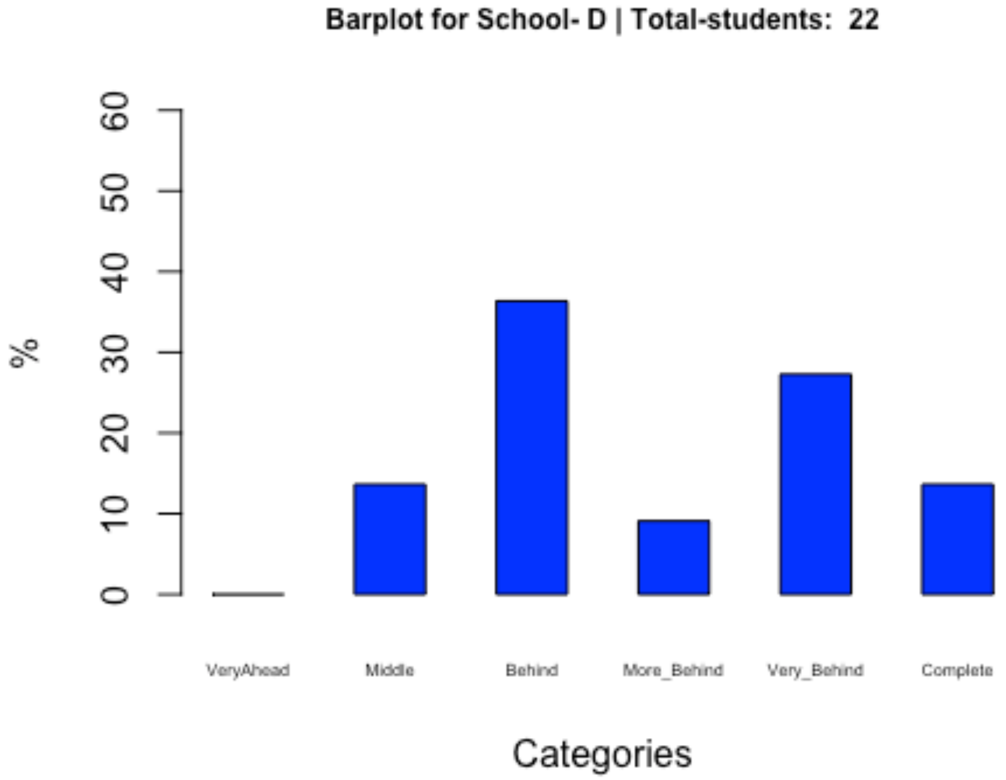
```
# For School-B:
data_school <- data[data$School=='B',]
stats_by_school <- rbind(stats_by_school, get_status_per_school(data_school,
'B'))
```



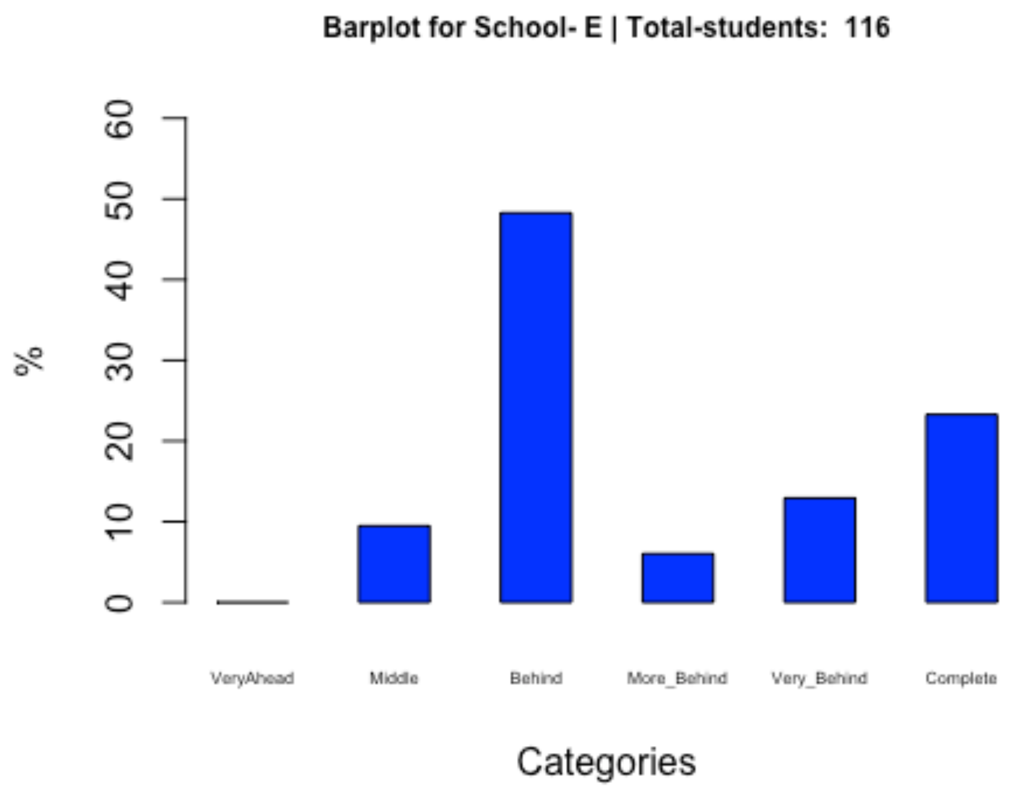
```
# For School-C:
data_school <- data[data$School=='C',]
stats_by_school <- rbind(stats_by_school, get_status_per_school(data_school,
'C'))
```



```r
# For School-D:
data_school <- data[data$School=='D',]
stats_by_school <- rbind(stats_by_school, get_status_per_school(data_school,
'D'))
```



```
# For School-E:
data_school <- data[data$School=='E',]
stats_by_school <- rbind(stats_by_school, get_status_per_school(data_school,
'E'))
```



```
# Rename rownames with School-Names
rownames(stats_by_school) <- c('School-A', 'School-B', 'School-C', 'School-
D', 'School-E')
stats_by_school

##          VeryAhead    Middle   Behind More_Behind Very_Behind Complete
Total
## School-A         0 12.124464 48.28326    7.832618   16.523605 15.23605
932
## School-B         0 18.834081 45.06726    3.139013    4.932735 28.02691
446
## School-C         0 12.941176 45.88235    4.705882   14.117647 22.35294
85
## School-D         0 13.636364 36.36364    9.090909   27.272727 13.63636
22
## School-E         0  9.482759 48.27586    6.034483   12.931034 23.27586
116
```

A few notable trends:

- high-percentage of students are lagging behind by 1 to 5 lessons
- no student is 0 to 5 lessons ahead although there are several students who already completed the course (so, ‘Very.Ahead’ is a measurement that can probably go!)

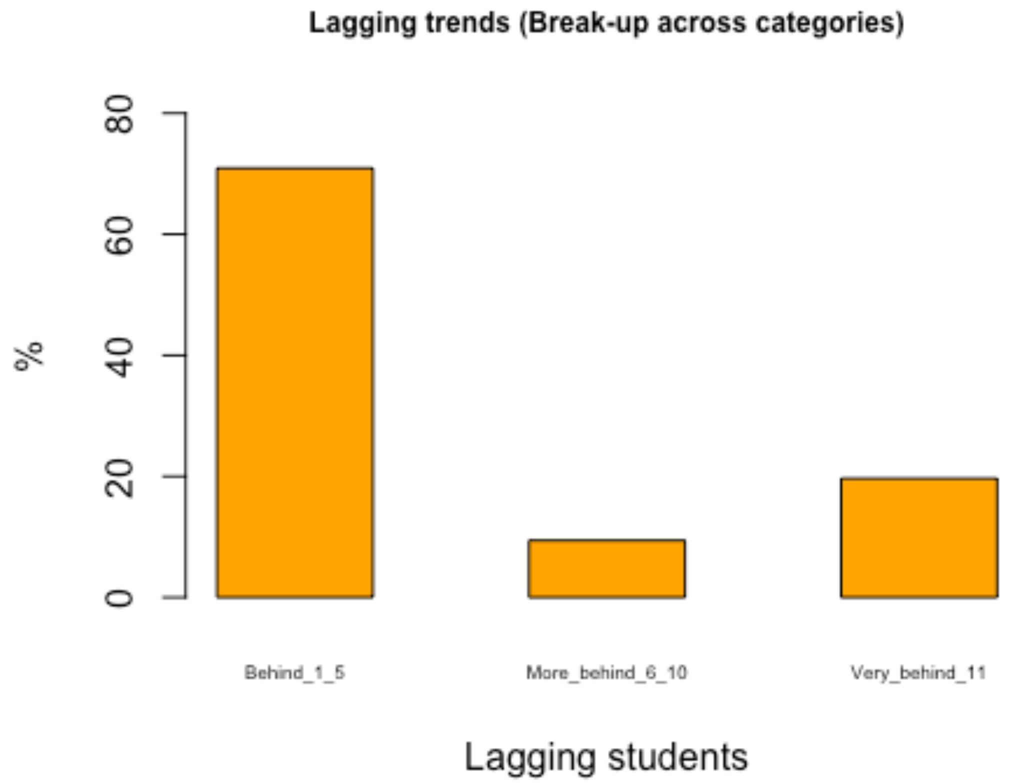
EDA-3: Lagging trends

From EDA-2, it is apparent that a large percentage of students are lagging behind by 1 to 5 lessons. The study in this section would present first, a percentage break-up between students lagging behind their lessons in different degrees viz. moderately behind or signifcantly behind (more than 6 lessons behind).

Subsequently, we conservatively predict or estimate students not likely to complete the course in the current semester.

```r
lag_behind_1_5 <- sum(data$Behind..1.5) / lagging_students * 100
lag_more_behind_6_10 <- sum(data$More.Behind..6.10) / lagging_students * 100
lag_very_behind_11 <- sum(data$Very.Behind..11) / lagging_students * 100

barplot(c(lag_behind_1_5, lag_more_behind_6_10, lag_very_behind_11),
names.arg = c('Behind_1_5', 'More_behind_6_10', 'Very_behind_11'),
main = 'Lagging trends (Break-up across categories)',
ylim = c(0,80),
col = 'orange',
cex.names=0.5,
xlab = 'Lagging students',
ylab = '%',
space=1,
cex.main = 0.75
)
```



Unless, something miraculously changes a high proportion of students in:

- More.Behind..6.10
- Very.Behind..11

Are likely, to NOT complete the course in the current semester.

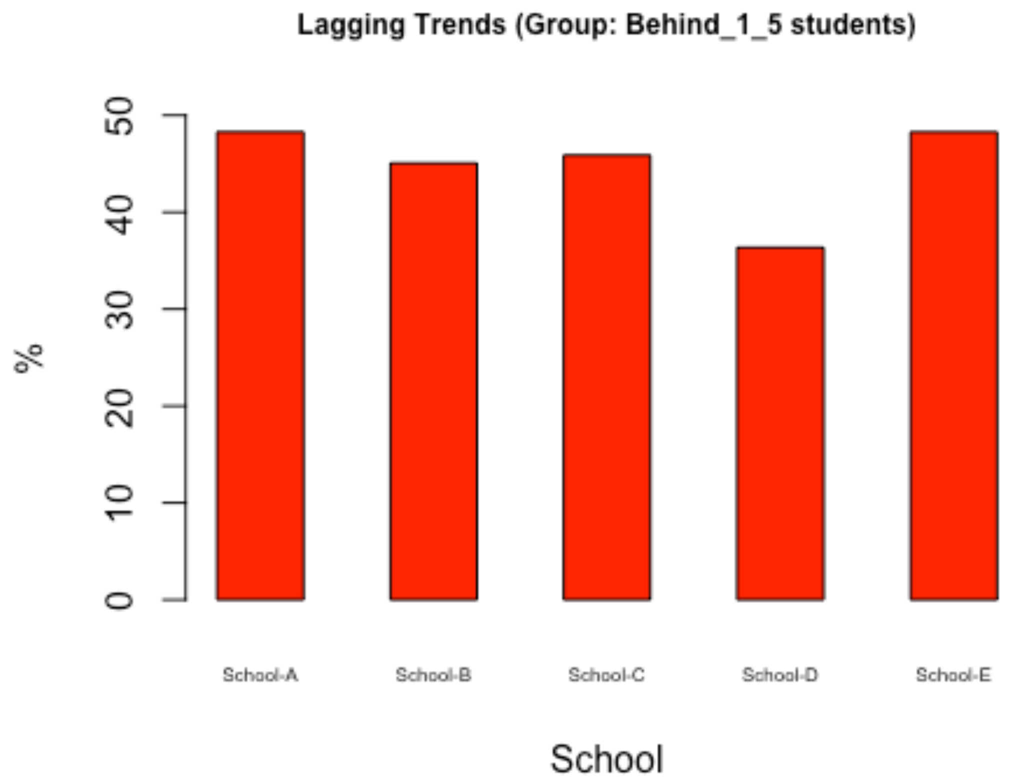
That count of such students (based on past outcomes, we could at some point dampen or acurately predict this):

```
students_incomplete <- sum(data$More.Behind..6.10) +
sum(data$Very.Behind..11)
pct_students_incomplete <- students_incomplete / total_students * 100
print(pct_students_incomplete)

## [1] 19.30044
```

Finally, turning attention to the group ‘Behind..1.5’, we see nearly similar numbers across all schools

```
barplot(stats_by_school$Behind,
names.arg = c(rownames(stats_by_school)),
main = 'Lagging Trends (Group: Behind_1_5 students)',
ylim = c(0,50),
col = 'red',
cex.names=0.5,
xlab = 'School',
ylab = '%',
space=1,
cex.main = 0.75
)
```



Conclusion

If one were analyzing this data from the perspective of estimating how many students are likely to complete the Math course this semester, there are sufficient trends at the time (at about 3/4 of the semester) indicating the number of students that were lagging behind.

As next steps, student counsellors should perhaps be working:

- with Professors and students in undertanding challenges with the Math course (and perhaps have compartive studies with other courses in the current semester or trends of the same course in the past).
- with some of the 20% of the students who have already completed the course to extend assistance

- to urge students lagging behind 1 to 5 lessons to sufficiently catch up and not slip further
- with students lagging behind more than 6 lessons in formalizing an alterate strategy to complete the course