IST 707 Homework 2

Due Date: 4/21/2021

Introduction

A particular math course is being taught at five different schools. There are a total of thirty sections across all five schools. There are 35 lessons as part of the math course. The semester is about 3/4 of the way complete. The subsequent report will explore, analyze, and illustrate findings in the data.

Data Preparation

Ensure packages are installed and active.

```
if   ("readr"      %in% installed.packages())   {require(readr)}      else
{install.packages('readr');      require(readr)}      #used for read csv file
```

```
## Loading required package: readr
```

```
if   ("sqldf"      %in% installed.packages())   {require(sqldf)}      else
{install.packages('sqldf');      require(sqldf)}      #used for some sql code
```

```
## Loading required package: sqldf
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
if   ("reshape2"   %in% installed.packages())   {require(reshape2)}   else
{install.packages('reshape2');   require(reshape2)}   #used for restructuring
data
```

```
## Loading required package: reshape2
```

```
if   ("dplyr"      %in% installed.packages())   {require(dplyr)}      else
{install.packages('dplyr');      require(dplyr)}      #used for joining data
```

```
## Loading required package: dplyr
```

```
##
## 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

if   ("ggplot2"    %in% installed.packages())   {require(ggplot2)}    else
{install.packages('ggplot2');    require(ggplot2)}    #used for visualization

## Loading required package: ggplot2
```

Read the data and store it.

```
# make sure that the data-storyteller.csv file is in the current working
directory

storyteller <- as.data.frame(read_csv("data-storyteller.csv"))

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

Change the data types accordingly.

```
# school and section should be factors because they are qualitative and
nominal.

for (colnumber in seq(1, 2)) {storyteller[,colnumber] <-
as.factor(storyteller[,colnumber])}; rm(colnumber)

# the remaining columns should be integer. No credit for partially completed
lessons.

for (colnumber in seq(3, 8)) {storyteller[,colnumber] <-
as.integer(storyteller[,colnumber])}; rm(colnumber)
```

Reorder the columns accordingly.

```
# columns are ordered from greatest to least in terms of lessons completed.
```

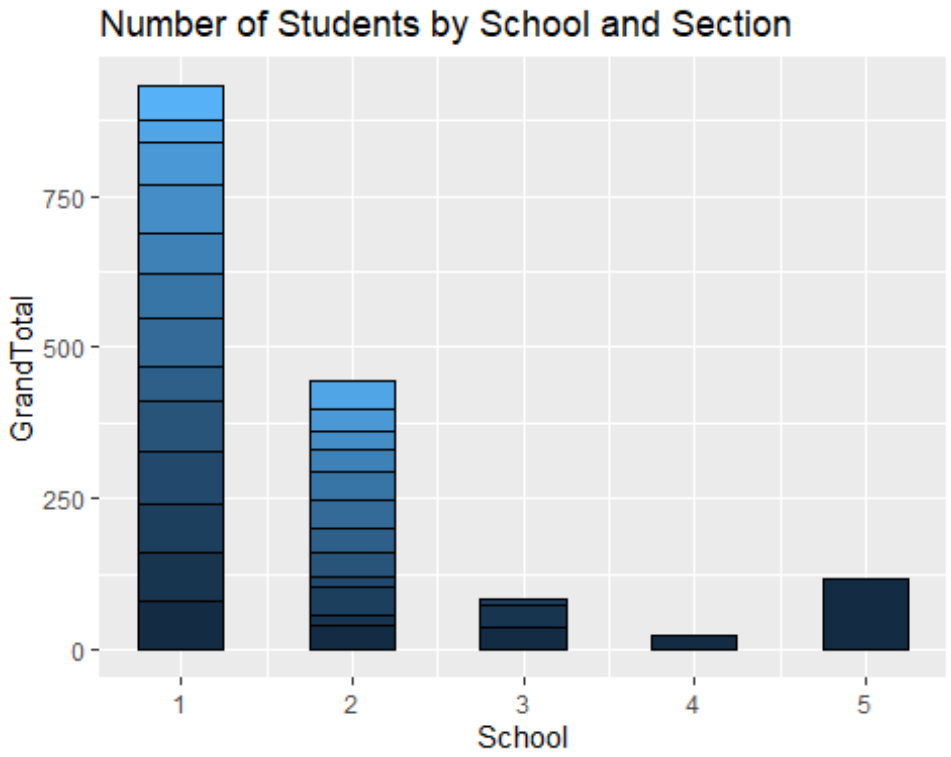
```
storyteller <-

       sqldf('select
              School,
             Section,
           Completed,
     `Very Ahead +5`,
       `Middling +0`,
       `Behind -1-5`,
  `More behind -6-10`,
    `Very behind -11`
   from storyteller');

head(storyteller)
```

```
##   School Section Completed Very Ahead +5 Middling +0 Behind -1-5
## 1      A       1        10             0           5          54
## 2      A       2         6             0           8          40
## 3      A       3        11             0           9          35
## 4      A       4        10             0          14          44
## 5      A       5         8             0           9          42
## 6      A       6         9             0           7          29
##   More Behind -6-10 Very Behind -11
## 1                 3               9
## 2                10              16
## 3                12              13
## 4                 5              12
## 5                 2              24
## 6                 3              10
```

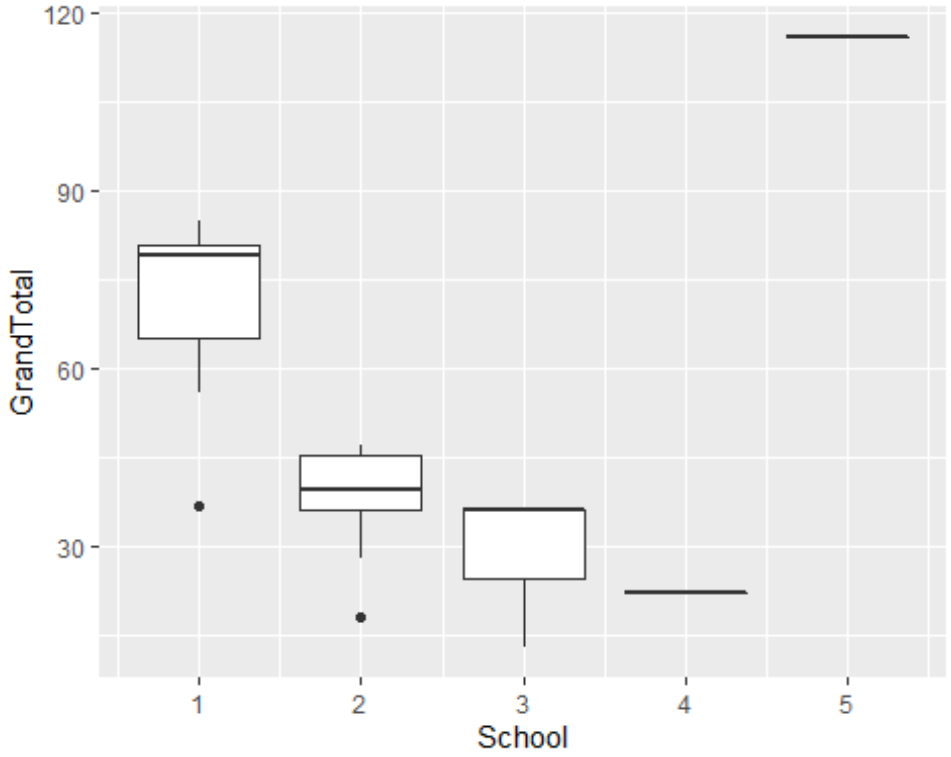
Data Analysis

View number of students in each school and section.

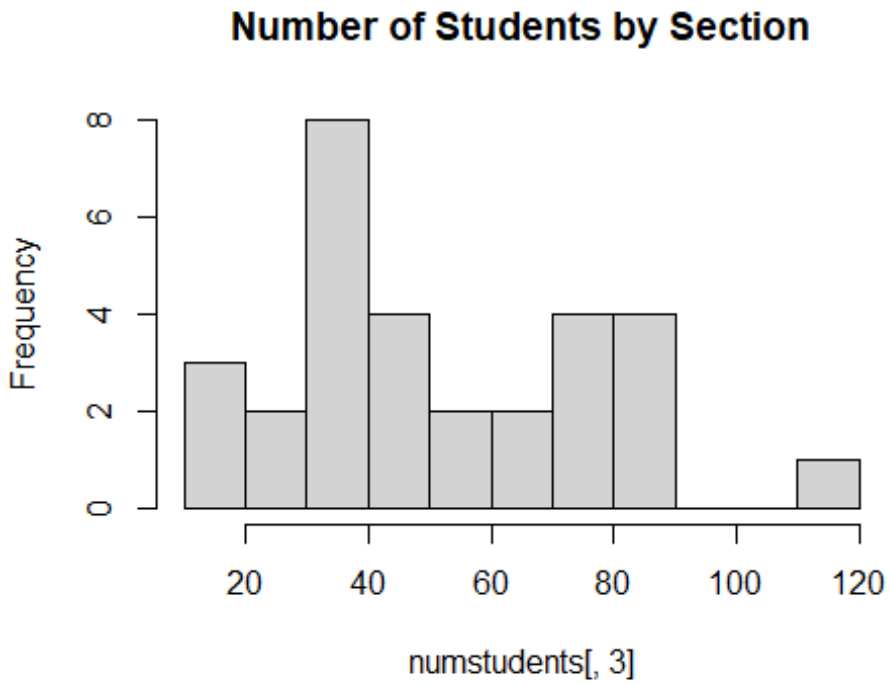
```
# create df

numstudents <-
  as.data.frame(cbind(
  School       = storyteller$School,
  Section      = storyteller$Section,
  GrandTotal   = rowSums(storyteller[3:8])))

# stacked bar

ggplot(numstudents,
aes(fill=Section, y=GrandTotal, x=School)) +
geom_bar(position="stack", stat="identity", width = .5, color = 'black') +
ggtitle("Number of Students by School and Section") +
theme(legend.position = "none")
```




```
# other charts

ggplot(numstudents[, c(1, 3)], aes(y=GrandTotal, x=School, group=School)) +
geom_boxplot()   #box plot
```




```
hist(numstudents[,3], breaks = 10, main = 'Number of Students by Section')
#histogram
```



```
tapply(numstudents$GrandTotal, numstudents$School, FUN = summary)
#summary stats
```

```
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   37.00   65.00   79.00   71.69   81.00   85.00
##
## $`2`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   36.25   39.50   37.17   45.25   47.00
##
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   24.50   36.00   28.33   36.00   36.00
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      22      22      22      22      22      22
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     116     116     116     116     116     116
```

Observations: -school A and school B have the most students -the average number of students per section is higher in school A -school D and school E only have one section. -the section in school E is very large compared to the rest.

Answer some follow up questions.

```
# what percent of students are in school A and B?

paste0(round(sum(numstudents[which(numstudents$School == 1 |
numstudents$School == 2), 3]) / sum(numstudents[, 3]) * 100, 0), '%')
```

```
## [1] "86%"
```

```
# how much higher is the average number of students per section in school A
than the other schools?

melt(round(tapply(numstudents$GrandTotal, numstudents$School, FUN = mean),
0), varnames = 'School', value.name = 'StudentsPerSection')
```

```
##   School StudentsPerSection
## 1      1                 72
## 2      2                 37
## 3      3                 28
## 4      4                 22
## 5      5                116
```

Summarize the data for each school and category.

```r
# create summary table by number of students for each school / category

summarytable1 <- aggregate(storyteller[,3:8], by=list(storyteller$School),
sum)
summarytable1$GrandTotal <- as.integer(rowSums(summarytable1[,-1]))

# create a summary table by percentage of grand total for each school /
category

summarytable2 <- summarytable1
for (colnumber in seq(2, 7)) {
  newcol <- paste0(round(summarytable2[,colnumber] /
  .GlobalEnv$summarytable2$GrandTotal, 2) * 100, '%')
  summarytable2[,colnumber] <- newcol }
summarytable2$GrandTotal <- replicate(5, '100%')

# combine tables to show both number of students and % of grand total
together

summarytable3 <- rbind(summarytable1, summarytable2)
summarytable3 <- sqldf('select * from summarytable3 order by `Group.1` asc')

summarytable4 <- as.data.frame(rbind(

  paste0(
    summarytable3[1,], ' (', summarytable3[2,], ')'),
  paste0(
    summarytable3[3,], ' (', summarytable3[4,], ')'),
  paste0(
    summarytable3[5,], ' (', summarytable3[6,], ')'),
  paste0(
    summarytable3[7,], ' (', summarytable3[8,], ')'),
  paste0(
    summarytable3[9,], ' (', summarytable3[10,], ')')))

colnames(summarytable4) <- colnames(summarytable1)
summarytable4
```

```
##   Group.1 Completed Very Ahead +5 Middling +0 Behind -1-5 More Behind -6-
10
## 1   1 (1) 142 (15%)        0 (0%)   113 (12%)   450 (48%)           73
(8%)
## 2   2 (2) 125 (28%)        0 (0%)    84 (19%)   201 (45%)           14
(3%)
## 3   3 (3)  19 (22%)        0 (0%)    11 (13%)    39 (46%)            4
(5%)
## 4   4 (4)   3 (14%)        0 (0%)     3 (14%)     8 (36%)            2
(9%)
## 5   5 (5)  27 (23%)        0 (0%)     11 (9%)    56 (48%)            7
```

```
(6%)
##   Very Behind -11 GrandTotal
## 1       154 (17%) 932 (100%)
## 2         22 (5%) 446 (100%)
## 3        12 (14%)  85 (100%)
## 4         6 (27%)  22 (100%)
## 5        15 (13%) 116 (100%)
```

Observations: -there are zero records where the student is very ahead -school B has a large portion of students that are on track or ahead of schedule (47%) -the next highest is school C at 35% -school D has a large portion of stents who are very behind -not concerned because although the % is high, the count is very low

Repeat for School B by section and category.

```
# create summary table by number of students for school B section / category

schoolBtable1 <- storyteller[which(storyteller$School == 'B'),]
schoolBtable1$GrandTotal <- as.integer(rowSums(schoolBtable1[,-c(1,2)]))

# create a summary table by percentage of grand total for school B section /
category

schoolBtable2 <- schoolBtable1
for (colnumber in seq(3, 8)) {
  newcol <- round(schoolBtable2[,colnumber] /
  .GlobalEnv$schoolBtable2$GrandTotal, 2) * 100
  newcol <- paste0(newcol, '%')
  schoolBtable2[,colnumber] <- newcol }
schoolBtable2$GrandTotal <- replicate(dim(schoolBtable2)[1], '100%')

# combine tables to show both number of students and % of grand total
together

schoolBtable3 <- rbind(schoolBtable1, schoolBtable2)
schoolBtable3 <- sqldf('select * from schoolBtable3 order by School, Section
asc')

schoolBtable4 <- as.data.frame(rbind(

  paste0(  schoolBtable3[1,], ' (', schoolBtable3[2,], ')'),     paste0(
schoolBtable3[3,], ' (', schoolBtable3[4,], ')'),
  paste0(  schoolBtable3[5,], ' (', schoolBtable3[6,], ')'),     paste0(
schoolBtable3[7,], ' (', schoolBtable3[8,], ')'),
  paste0(  schoolBtable3[9,], ' (', schoolBtable3[10,], ')'),    paste0(
schoolBtable3[11,], ' (', schoolBtable3[12,], ')'),
  paste0(  schoolBtable3[13,], ' (', schoolBtable3[14,], ')'),   paste0(
schoolBtable3[15,], ' (', schoolBtable3[16,], ')'),
  paste0(  schoolBtable3[17,], ' (', schoolBtable3[18,], ')'),   paste0(
schoolBtable3[19,], ' (', schoolBtable3[20,], ')'),
```

```
  paste0(  schoolBtable3[21,], ' (', schoolBtable3[22,], ')'),   paste0(
schoolBtable3[23,], ' (', schoolBtable3[24,], ')')))

colnames(schoolBtable4) <- colnames(schoolBtable1)
schoolBtable4
```

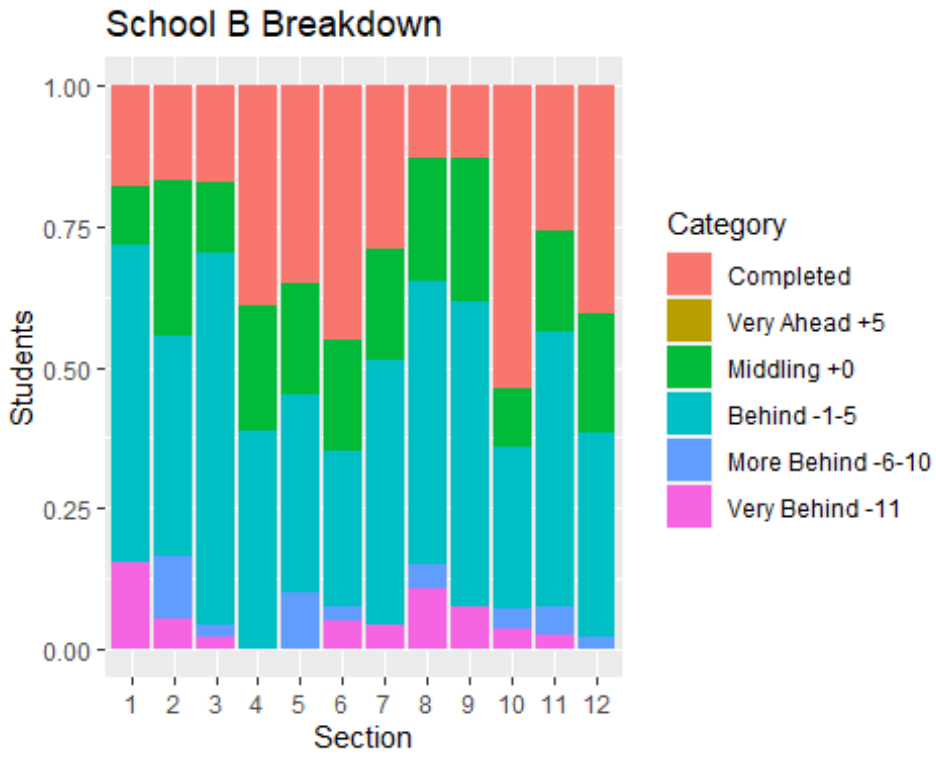
```
##    School Section Completed Very Ahead +5 Middling +0 Behind -1-5
## 1   2 (2)   1 (1)   7 (18%)        0 (0%)     4 (10%)    22 (56%)
## 2   2 (2) 10 (10)  15 (54%)        0 (0%)     3 (11%)     8 (29%)
## 3   2 (2) 11 (11)  10 (26%)        0 (0%)     7 (18%)    19 (49%)
## 4   2 (2) 12 (12)  19 (40%)        0 (0%)    10 (21%)    17 (36%)
## 5   2 (2)   2 (2)   3 (17%)        0 (0%)     5 (28%)     7 (39%)
## 6   2 (2)   3 (3)   8 (17%)        0 (0%)     6 (13%)    31 (66%)
## 7   2 (2)   4 (4)   7 (39%)        0 (0%)     4 (22%)     7 (39%)
## 8   2 (2)   5 (5)  14 (35%)        0 (0%)     8 (20%)    14 (35%)
## 9   2 (2)   6 (6)  18 (45%)        0 (0%)     8 (20%)    11 (28%)
## 10  2 (2)   7 (7)  13 (29%)        0 (0%)     9 (20%)    21 (47%)
## 11  2 (2)   8 (8)   6 (13%)        0 (0%)    10 (22%)    23 (50%)
## 12  2 (2)   9 (9)   5 (13%)        0 (0%)    10 (26%)    21 (54%)
##    More Behind -6-10 Very Behind -11 GrandTotal
## 1              0 (0%)         6 (15%)  39 (100%)
## 2              1 (4%)          1 (4%)  28 (100%)
## 3              2 (5%)          1 (3%)  39 (100%)
## 4              1 (2%)          0 (0%)  47 (100%)
## 5             2 (11%)          1 (6%)  18 (100%)
## 6              1 (2%)          1 (2%)  47 (100%)
## 7              0 (0%)          0 (0%)  18 (100%)
## 8             4 (10%)          0 (0%)  40 (100%)
## 9              1 (3%)          2 (5%)  40 (100%)
## 10             0 (0%)          2 (4%)  45 (100%)
## 11             2 (4%)         5 (11%)  46 (100%)
## 12             0 (0%)          3 (8%)  39 (100%)
```

Observations: -section 6 is the highest performer with 26 students on track or ahead of schedule (57% of total section) -there are several other sections that stand out as doing better than the rest.

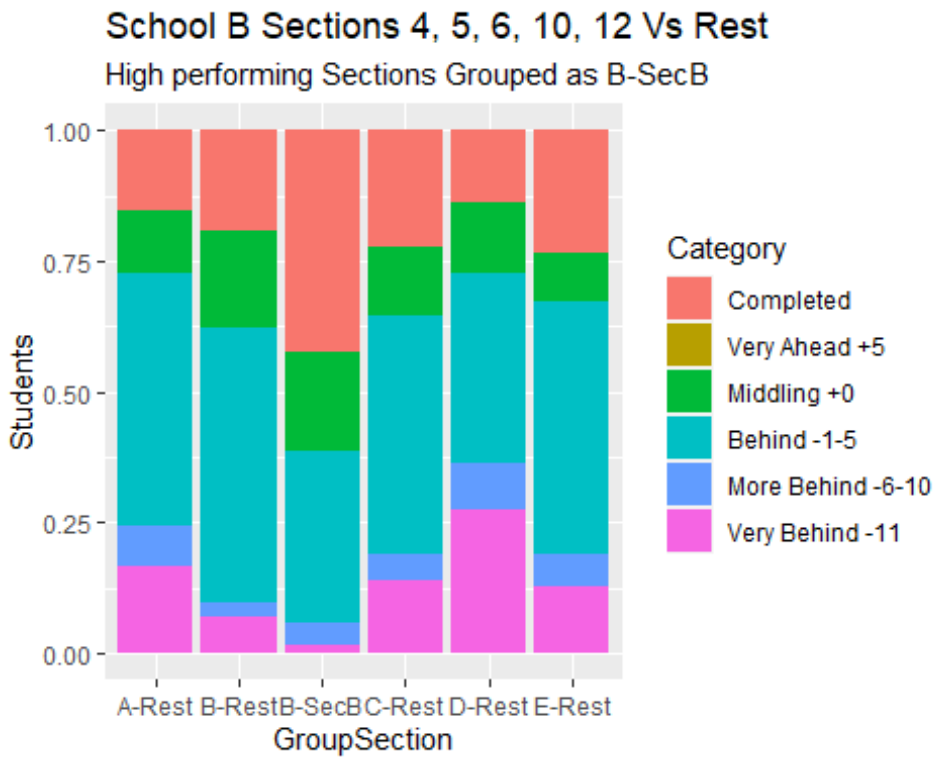
Compare these sections to the others.

```
# create a lookup table that will be used to lookup the grand total of
students

storytellerlookup                 <- storyteller
storytellerlookup$GrandTotal      <- as.integer(rowSums(storytellerlookup[,-
c(1,2)]))
storytellerlookup$lookupID        <- paste(storytellerlookup$School,
storytellerlookup$Section, sep = '-')
storytellerlookup                 <-
storytellerlookup[,which(colnames(storytellerlookup) %in% c('GrandTotal',
'lookupID'))]
```

```r
# create a collapsed table that will be used to create stacked bar
visualization

storytellerstacked                <- melt(storyteller, id = c('School',
'Section'), variable.name = 'Category', value.name =  'Students')
storytellerstacked$lookupID       <- paste(storytellerstacked$School,
storytellerstacked$Section, sep = '-')
storytellerstacked                <- left_join(storytellerstacked,
storytellerlookup, on = 'lookupID')

## Joining, by = "lookupID"

storytellerstacked$pctTotal       <- round((storytellerstacked$Students /
storytellerstacked$GrandTotal)*100, 0)
storytellerstacked$GroupSection   <- ifelse(storytellerstacked$School == 'B'
& storytellerstacked$Section %in% c(4, 5, 6, 10, 12), 'SecB', 'Rest')
storytellerstacked$GroupSection   <- paste(storytellerstacked$School,
storytellerstacked$GroupSection, sep = '-')
SchoolBBreakdown                  <-
storytellerstacked[which(storytellerstacked$School == 'B'),]

# create a stacked bar chart showing school B and the percent of students
breakdown by section

  ggplot(SchoolBBreakdown, aes(fill=Category, y=Students, x=Section)) +
  geom_bar(position="fill", stat="identity") +
  ggtitle("School B Breakdown")
```



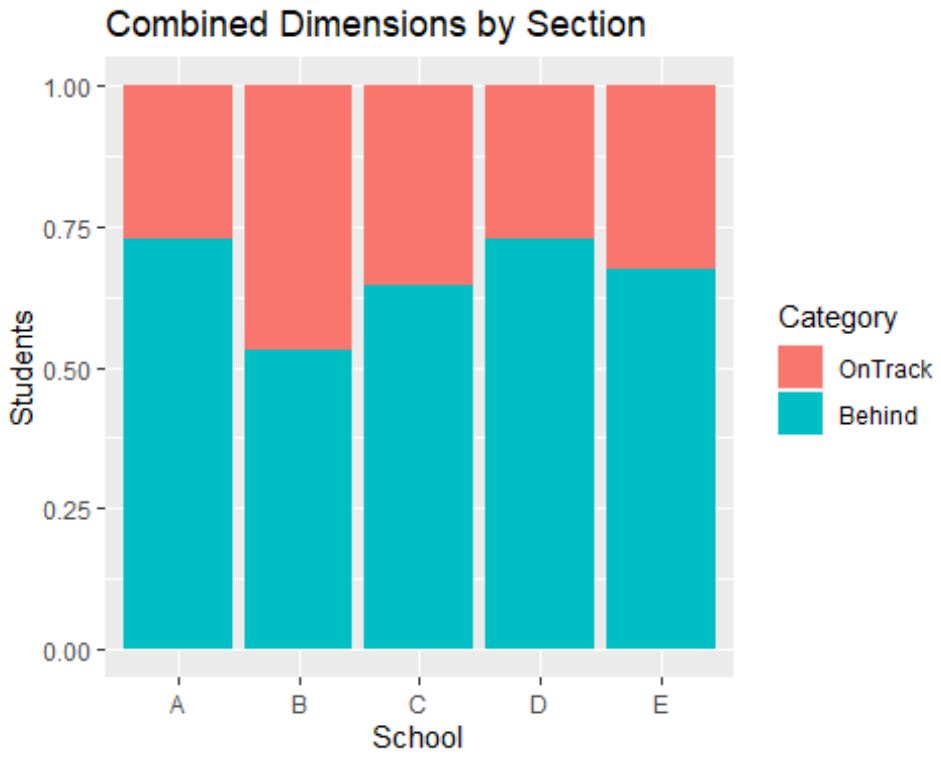

```
# create a stacked bar chart showing the best school B sections versus the
rest of the school sections

  ggplot(storytellerstacked, aes(fill=Category, y=Students, x=GroupSection))
+
  geom_bar(position="fill", stat="identity") +
  ggtitle("School B Sections 4, 5, 6, 10, 12 Vs Rest",
  subtitle = 'High performing Sections Grouped as B-SecB')
```



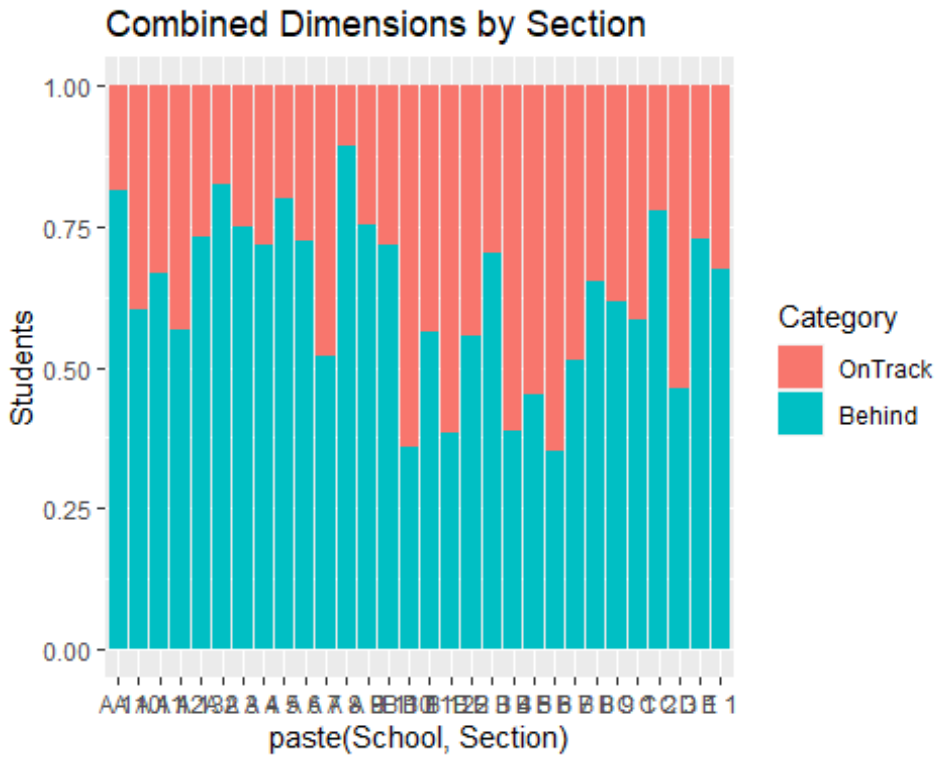
Observations: -it is clear that school B sections 4, 5, 6, 10, 12 are performing better than the others -for these sections, over 50% of the students are on track or ahead of schedule -most is coming from completed, although middling is still slightly higher than average

Experiment with dimensionality reduction

```
# create a new dataset with the combined dimensions
storytellernew <- storyteller
storytellernew$OnTrack <- storytellernew$`Very Ahead +5` +
storytellernew$`Middling +0` + storytellernew$Completed
storytellernew$Behind <- storytellernew$`Behind -1-5` + storytellernew$`More
Behind -6-10` + storytellernew$`Very Behind -11`
storytellernew <- storytellernew[, which(colnames(storytellernew) %in%
c('School', 'Section', 'OnTrack', 'Behind'))]
storytellernew <- melt(storytellernew, id = c('School', 'Section'),
variable.name = 'Category', value.name =  'Students')

  #visualize by school
  ggplot(storytellernew,
  aes(fill=Category, y=Students, x=School)) +
  geom_bar(position="fill", stat="identity") +
  ggtitle("Combined Dimensions by Section")
```




```
#visualize by section
ggplot(storytellernew,
aes(fill=Category, y=Students, x=paste(School, Section))) +
geom_bar(position="fill", stat="identity") +
ggtitle("Combined Dimensions by Section")
```



Conclusion

My analysis confirms that there is something that some sections are doing better than others. It is unlikely that this is due to random variation. The reason for this will need to be further investigated. A good next step could be to meet with teachers and students to gather information, and collect more data if needed. Some potentially useful data could be quality of the teacher, student demographics, location of the school, difficulty of the course, past sections of the course, data on similar courses, etc.