JohnFieldsHW2knit.R

johnfields

2019-09-14

```
######################################################################
## Name: John Fields
## Class: IST707 - Dr. Ami Gates
## Assignment: Homework #2
## Date: 18 Jul 2019
######################################################################
```

```
# PROCESS & TRANSFORM DATA
```

```
##install.packages("ggplot2")
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse 1
```

```
## v tibble  2.1.1       v purrr   0.3.2
## v tidyr   0.8.3       v dplyr   0.8.0.1
## v readr   1.3.1       v stringr 1.4.0
## v tibble  2.1.1       v forcats 0.4.0
```

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

```
## Set the working directory to the path where the code and datafile are located
setwd("/Users/johnfields/Library/Mobile Documents/com~apple~CloudDocs/Syracuse/IST707/Homework/Week 2/")

## Read in .csv data
filename="datastoryteller.csv"
MyStoryData <- read.csv(filename, header = TRUE, na.strings = "NA")

## Look at the data as a data frame
## I changed Section from int to factor but received errors in some
## of the plots and changed it back to int
MyStoryData
```

```
##    School Section Average Behind MoreBehind VeryBehind Completed
## 1       A       1       5     54          3          9        10
## 2       A       2       8     40         10         16         6
## 3       A       3       9     35         12         13        11
## 4       A       4      14     44          5         12        10
## 5       A       5       9     42          2         24         8
## 6       A       6       7     29          3         10         9
## 7       A       7      19     22          5         14        19
## 8       A       8       3     37         11         18         5
## 9       A       9       6     29          8         12        10
## 10      A      10      13     40          5          5        20
## 11      A      11       8     32          4         10        15
## 12      A      12       2     16          2          3        14
## 13      A      13      10     30          3          8         5
```

```
## 14      B       1       4     22          0          6         7
## 15      B       2       5      7          2          1         3
## 16      B       3       6     31          1          1         8
## 17      B       4       4      7          0          0         7
## 18      B       5       8     14          4          0        14
## 19      B       6       8     11          1          2        18
## 20      B       7       9     21          0          2        13
## 21      B       8      10     23          2          5         6
## 22      B       9      10     21          0          3         5
## 23      B      10       3      8          1          1        15
## 24      B      11       7     19          2          1        10
## 25      B      12      10     17          1          0        19
## 26      C       1       2     15          2          4        13
## 27      C       2       7     20          1          7         1
## 28      C       3       2      4          1          1         5
## 29      D       1       3      8          2          6         3
## 30      E       1      11     56          7         15        27
```

```
str(MyStoryData)
```

```
## 'data.frame':    30 obs. of  7 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 ...
##  $ Average   : int  5 8 9 14 9 7 19 3 6 13 ...
##  $ Behind    : int  54 40 35 44 42 29 22 37 29 40 ...
##  $ MoreBehind: int  3 10 12 5 2 3 5 11 8 5 ...
##  $ VeryBehind: int  9 16 13 12 24 10 14 18 12 5 ...
##  $ Completed : int  10 6 11 10 8 9 19 5 10 20 ...
```

```
#View(MyStoryData)

## Read in data again after spaces removed, column names changed
filename="datastorytellerFIXED.csv"
MyStoryData <- read.csv(filename, header = TRUE, na.strings = "NA")

## Look at the data as a data frame
(head(MyStoryData))
```

```
##   School Section Average Behind MoreBehind VeryBehind Completed SchoolSize
## 1      A       1       5     54          3          9        10      Large
## 2      A       2       8     40         10         16         6      Large
## 3      A       3       9     35         12         13        11      Large
## 4      A       4      14     44          5         12        10      Large
## 5      A       5       9     42          2         24         8      Large
## 6      A       6       7     29          3         10         9      Large
```

```
## Check for missing values
Total <-sum(is.na(MyStoryData))
cat("The number of missing values in StoryTeller data is ", Total )
```

```
## The number of missing values in StoryTeller data is  0
```

```
## Check each numerical variable to see that it is >=0
for(varname in names(MyStoryData)){
## Only check numeric variables
  if(sapply(MyStoryData[varname], is.numeric)){
    cat("\n", varname, " is numeric\n")
```

```
    ## Get median
    (Themedian <- sapply(MyStoryData[varname],FUN=median))
    ##print(Themedian)
    ## check/replace if the values are <=0
    MyStoryData[varname] <- replace(MyStoryData[varname], MyStoryData[varname] < 0, Themedian)
  }

}
```

```
##
##  Section  is numeric
##
##  Average  is numeric
##
##  Behind  is numeric
##
##  MoreBehind  is numeric
##
##  VeryBehind  is numeric
##
##  Completed  is numeric
```

```
(MyStoryData)
```

```
##    School Section Average Behind MoreBehind VeryBehind Completed
## 1       A       1       5     54          3          9        10
## 2       A       2       8     40         10         16         6
## 3       A       3       9     35         12         13        11
## 4       A       4      14     44          5         12        10
## 5       A       5       9     42          2         24         8
## 6       A       6       7     29          3         10         9
## 7       A       7      19     22          5         14        19
## 8       A       8       3     37         11         18         5
## 9       A       9       6     29          8         12        10
## 10      A      10      13     40          5          5        20
## 11      A      11       8     32          4         10        15
## 12      A      12       2     16          2          3        14
## 13      A      13      10     30          3          8         5
## 14      B       1       4     22          0          6         7
## 15      B       2       5      7          2          1         3
## 16      B       3       6     31          1          1         8
## 17      B       4       4      7          0          0         7
## 18      B       5       8     14          4          0        14
## 19      B       6       8     11          1          2        18
## 20      B       7       9     21          0          2        13
## 21      B       8      10     23          2          5         6
## 22      B       9      10     21          0          3         5
## 23      B      10       3      8          1          1        15
## 24      B      11       7     19          2          1        10
## 25      B      12      10     17          1          0        19
## 26      C       1       2     15          2          4        13
## 27      C       2       7     20          1          7         1
## 28      C       3       2      4          1          1         5
## 29      D       1       3      8          2          6         3
## 30      E       1      11     56          7         15        27
```

```
##    SchoolSize
## 1       Large
## 2       Large
## 3       Large
## 4       Large
## 5       Large
## 6       Large
## 7       Large
## 8       Large
## 9       Large
## 10      Large
## 11      Large
## 12      Large
## 13      Large
## 14      Large
## 15      Large
## 16      Large
## 17      Large
## 18      Large
## 19      Large
## 20      Large
## 21      Large
## 22      Large
## 23      Large
## 24      Large
## 25      Large
## 26      Small
## 27      Small
## 28      Small
## 29      Small
## 30      Small
```

```
## Use table to explore the data
(table(MyStoryData$School))
```

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

```
## Use loop to create all the tables at once
for(i in 1:ncol(MyStoryData)){
  print(table(MyStoryData[i]))
}
```

```
##
##  A  B  C  D  E
## 13 12  3  1  1
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13
##  5  3  3  2  2  2  2  2  2  2  2  2  1
##
##  2  3  4  5  6  7  8  9 10 11 13 14 19
##  3  3  2  2  2  3  4  3  4  1  1  1  1
##
##  4  7  8 11 14 15 16 17 19 20 21 22 23 29 30 31 32 35 37 40 42 44 54 56
##  1  2  2  1  1  1  1  1  1  1  2  2  1  2  1  1  1  1  1  2  1  1  1  1
```

```
##
##  0  1  2  3  4  5  7  8 10 11 12
##  4  6  7  3  2  3  1  1  1  1  1
##
##  0  1  2  3  4  5  6  7  8  9 10 12 13 14 15 16 18 24
##  3  5  2  2  1  2  2  1  1  1  2  2  1  1  1  1  1  1
##
##  1  3  5  6  7  8  9 10 11 13 14 15 18 19 20 27
##  1  2  4  2  2  2  1  4  1  2  2  2  1  2  1  1
##
## Large Small
##    25     5
```

```
(colnames(MyStoryData))
```

```
## [1] "School"     "Section"    "Average"    "Behind"     "MoreBehind"
## [6] "VeryBehind" "Completed"  "SchoolSize"
```

```
(head(MyStoryData))
```

```
##   School Section Average Behind MoreBehind VeryBehind Completed SchoolSize
## 1      A       1       5     54          3          9        10      Large
## 2      A       2       8     40         10         16         6      Large
## 3      A       3       9     35         12         13        11      Large
## 4      A       4      14     44          5         12        10      Large
## 5      A       5       9     42          2         24         8      Large
## 6      A       6       7     29          3         10         9      Large
```

```
(MyStoryData)
```

```
##    School Section Average Behind MoreBehind VeryBehind Completed
## 1       A       1       5     54          3          9        10
## 2       A       2       8     40         10         16         6
## 3       A       3       9     35         12         13        11
## 4       A       4      14     44          5         12        10
## 5       A       5       9     42          2         24         8
## 6       A       6       7     29          3         10         9
## 7       A       7      19     22          5         14        19
## 8       A       8       3     37         11         18         5
## 9       A       9       6     29          8         12        10
## 10      A      10      13     40          5          5        20
## 11      A      11       8     32          4         10        15
## 12      A      12       2     16          2          3        14
## 13      A      13      10     30          3          8         5
## 14      B       1       4     22          0          6         7
## 15      B       2       5      7          2          1         3
## 16      B       3       6     31          1          1         8
## 17      B       4       4      7          0          0         7
## 18      B       5       8     14          4          0        14
## 19      B       6       8     11          1          2        18
## 20      B       7       9     21          0          2        13
## 21      B       8      10     23          2          5         6
## 22      B       9      10     21          0          3         5
## 23      B      10       3      8          1          1        15
## 24      B      11       7     19          2          1        10
## 25      B      12      10     17          1          0        19
```

```
## 26      C       1       2     15          2          4        13
## 27      C       2       7     20          1          7         1
## 28      C       3       2      4          1          1         5
## 29      D       1       3      8          2          6         3
## 30      E       1      11     56          7         15        27
##    SchoolSize
## 1       Large
## 2       Large
## 3       Large
## 4       Large
## 5       Large
## 6       Large
## 7       Large
## 8       Large
## 9       Large
## 10      Large
## 11      Large
## 12      Large
## 13      Large
## 14      Large
## 15      Large
## 16      Large
## 17      Large
## 18      Large
## 19      Large
## 20      Large
## 21      Large
## 22      Large
## 23      Large
## 24      Large
## 25      Large
## 26      Small
## 27      Small
## 28      Small
## 29      Small
## 30      Small
```

```r
## Which variables contain information?  All except Section (identifier)
## and VeryAhead (no completions or bad data)
## Does the Section? The E school likley has bad data since this school
## has 1 section with over 100 students

## The table shows us that we have 5 schools but only 2 have much data
## (School A & B)
## This is important because these larger schools can shift the data
## since they have greater number of students

## Are there outliers or odd values?
## School E has 1 section with 116 students

## The structure (types) of the data
(str(MyStoryData))
```

```
## '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 ...
##  $ Average   : int  5 8 9 14 9 7 19 3 6 13 ...
##  $ Behind    : int  54 40 35 44 42 29 22 37 29 40 ...
##  $ MoreBehind: int  3 10 12 5 2 3 5 11 8 5 ...
##  $ VeryBehind: int  9 16 13 12 24 10 14 18 12 5 ...
##  $ Completed : int  10 6 11 10 8 9 19 5 10 20 ...
##  $ SchoolSize: Factor w/ 2 levels "Large","Small": 1 1 1 1 1 1 1 1 1 1 ...
```

```
## NULL
```

```
## Summary of the data - mean median, sums
summary(MyStoryData)
```

```
##  School    Section         Average          Behind        MoreBehind
##  A:13   Min.   : 1.00   Min.   : 2.00   Min.   : 4.00   Min.   : 0.000
##  B:12   1st Qu.: 2.25   1st Qu.: 4.25   1st Qu.:15.25   1st Qu.: 1.000
##  C: 3   Median : 5.50   Median : 7.50   Median :22.00   Median : 2.000
##  D: 1   Mean   : 5.90   Mean   : 7.40   Mean   :25.13   Mean   : 3.333
##  E: 1   3rd Qu.: 9.00   3rd Qu.: 9.75   3rd Qu.:34.25   3rd Qu.: 4.750
##         Max.   :13.00   Max.   :19.00   Max.   :56.00   Max.   :12.000
##    VeryBehind       Completed     SchoolSize
##  Min.   : 0.000   Min.   : 1.00   Large:25
##  1st Qu.: 1.250   1st Qu.: 6.00   Small: 5
##  Median : 5.500   Median :10.00
##  Mean   : 6.967   Mean   :10.53
##  3rd Qu.:11.500   3rd Qu.:14.00
##  Max.   :24.000   Max.   :27.00
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

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

```
## The following will sum all rows for each "School" and per variable in the data
## Save this new aggregated result as a DF
SumBySchoolDF <- ddply(MyStoryData, "School", numcolwise(sum))
(SumBySchoolDF)
```

```
##   School Section Average Behind MoreBehind VeryBehind Completed
## 1      A      91     113    450         73        154       142
## 2      B      78      84    201         14         22       125
## 3      C       6      11     39          4         12        19
```

```
## 4      D       1       3      8          2          6         3
## 5      E       1      11     56          7         15        27
```

```
## Total the number of students for A - E and sum the columns for each row

(SumBySchoolDF)
```

```
##   School Section Average Behind MoreBehind VeryBehind Completed
## 1      A      91     113    450         73        154       142
## 2      B      78      84    201         14         22       125
## 3      C       6      11     39          4         12        19
## 4      D       1       3      8          2          6         3
## 5      E       1      11     56          7         15        27
```

```
SumOfStudents <- rowSums(SumBySchoolDF[,c("Average", "Behind",
                                          "MoreBehind","VeryBehind","Completed")])

##Create new dataframes for visualizations and calculations
##Calculate the number of students expected to complete (Complete & Average) and
##behind (Behind, More Behind, Very Behind)
(SumOfBehind <- rowSums(SumBySchoolDF[,c("Behind","MoreBehind","VeryBehind")]))
```

```
## [1] 677 237  55  16  78
```

```
(SumOfCompletedAverage <- rowSums(SumBySchoolDF[,c("Average","Completed")]))
```

```
## [1] 255 209  30   6  38
```

```
(SumByBehindCompleteDF <- data.frame(SumBySchoolDF$School,SumOfBehind,SumOfCompletedAverage))
```

```
##   SumBySchoolDF.School SumOfBehind SumOfCompletedAverage
## 1                    A         677                   255
## 2                    B         237                   209
## 3                    C          55                    30
## 4                    D          16                     6
## 5                    E          78                    38
```

```
(SumByBehindCompleteDF$Total <- SumOfBehind+SumOfCompletedAverage)
```

```
## [1] 932 446  85  22 116
```

```
(SumByBehindCompleteDF$PercentCompleteBySchool <- (SumByBehindCompleteDF$SumOfCompletedAverage/SumByBeh
```

```
## [1] 0.2736052 0.4686099 0.3529412 0.2727273 0.3275862
```

```
(TotalBehindBySchool <- (SumBySchoolDF$Behind+SumBySchoolDF$MoreBehind+SumBySchoolDF$VeryBehind))
```

```
## [1] 677 237  55  16  78
```

```
(TotalBehind <- sum(TotalBehindBySchool))
```

```
## [1] 1063
```

```
(TotalCompleteAverageBySchool <- (SumBySchoolDF$Completed+SumBySchoolDF$Average))
```

```
## [1] 255 209  30   6  38
```

```
(TotalCompleteAverage <- sum(TotalCompleteAverageBySchool))
```

```
## [1] 538
```

```
(CompleteBehind <- data.frame(Group=c("Total Behind", "Total Complete/Average"),Students = c(TotalBehin
```

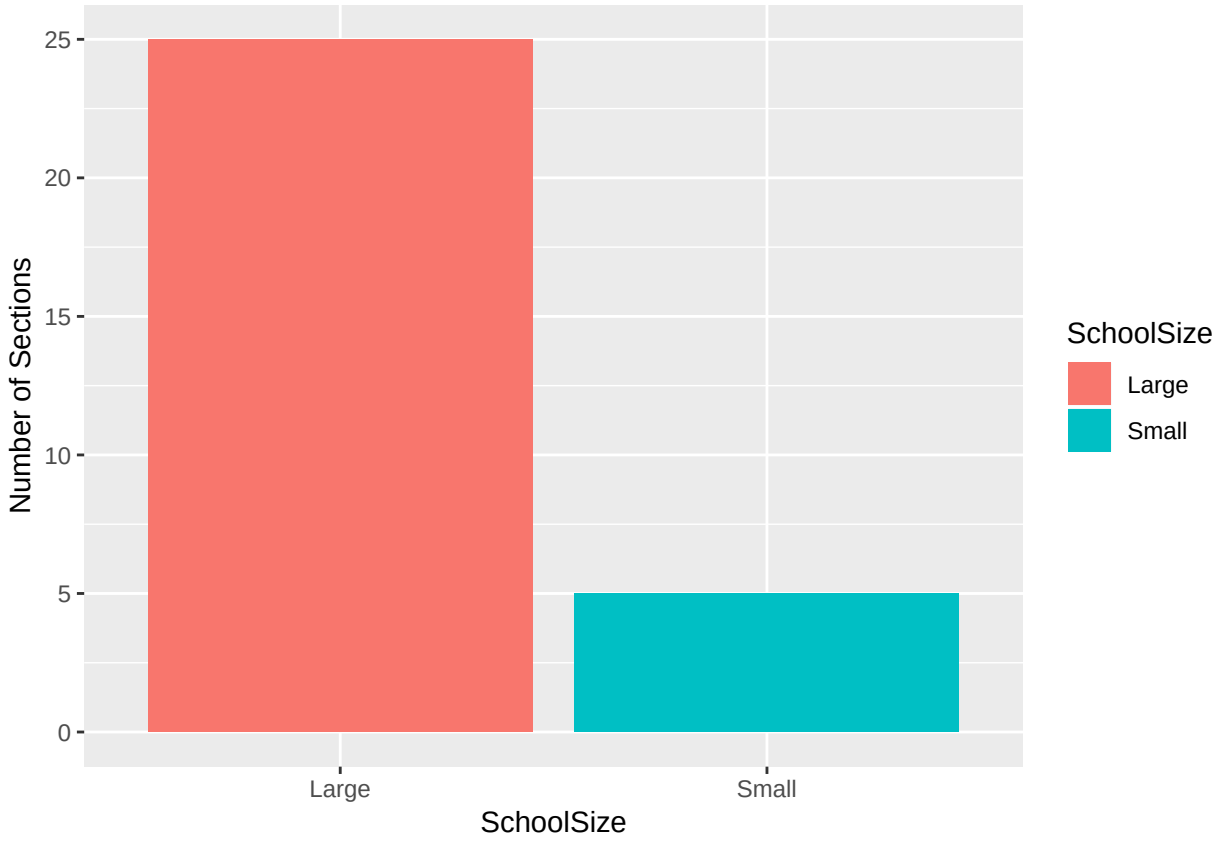
```
##                     Group Students
## 1           Total Behind     1063
## 2 Total Complete/Average      538
```

```
(TotalStudents <- sum(CompleteBehind[1,2]+CompleteBehind[2,2]))
```

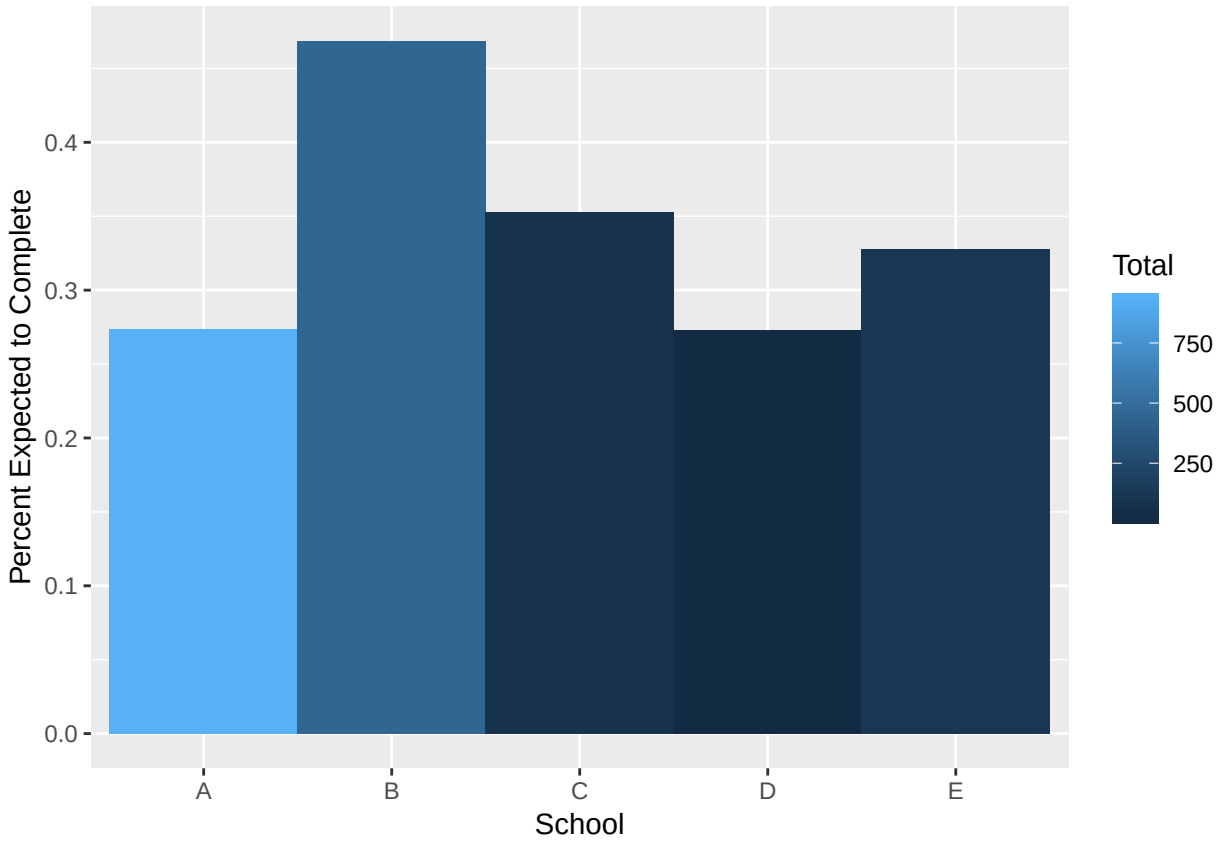
```
## [1] 1601
```

```
# VISUALIZATIONS
```

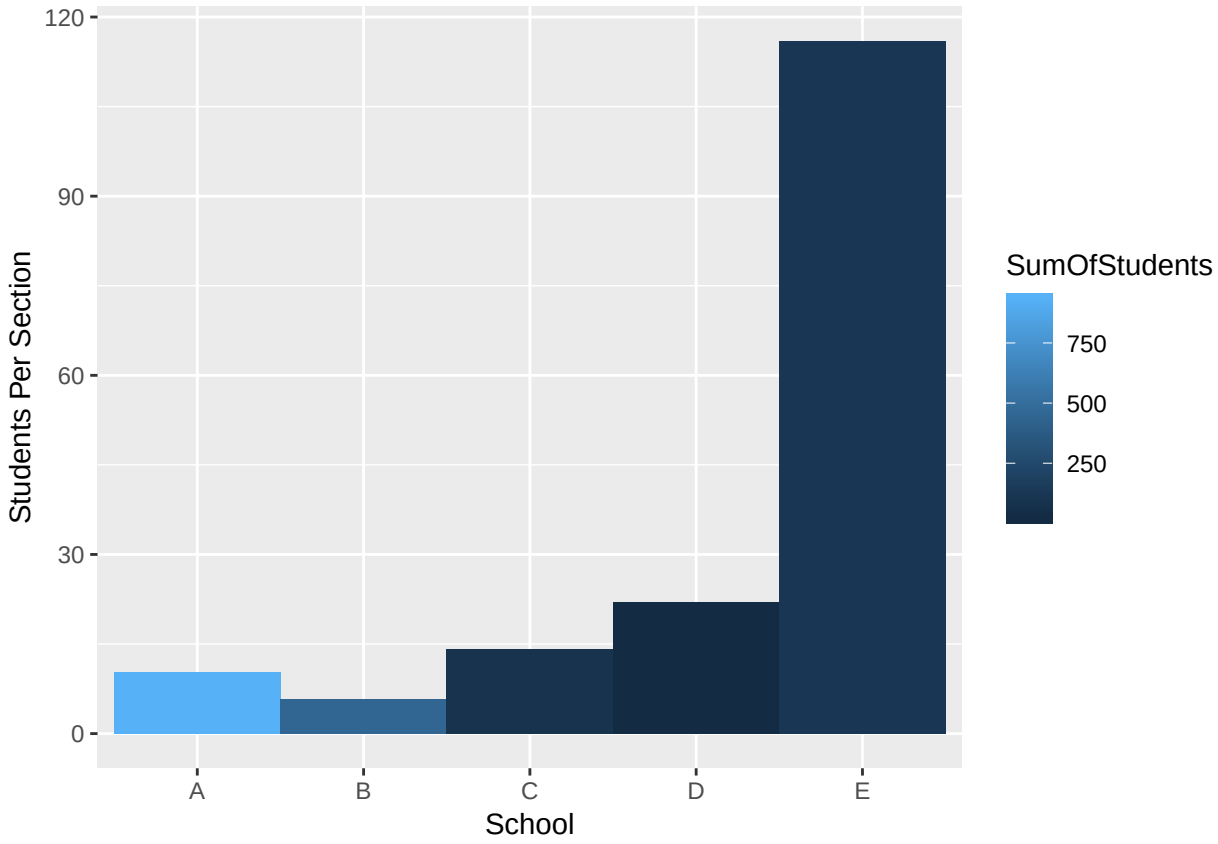
```
##Plot a comparison of the number of sections by school size
BaseGraph <- ggplot(MyStoryData)
(SchoolSizebySection<-BaseGraph + geom_bar(aes(SchoolSize, fill = SchoolSize)) + ylab("Number of Sectio
```




```
##Plot percentage of Complete & Average / Total Students
ggplot(SumByBehindCompleteDF, aes(x=SumBySchoolDF.School,y=PercentCompleteBySchool,fill=Total))+
geom_bar(width=1,stat="identity") + ylab("Percent Expected to Complete") + xlab("School")
```

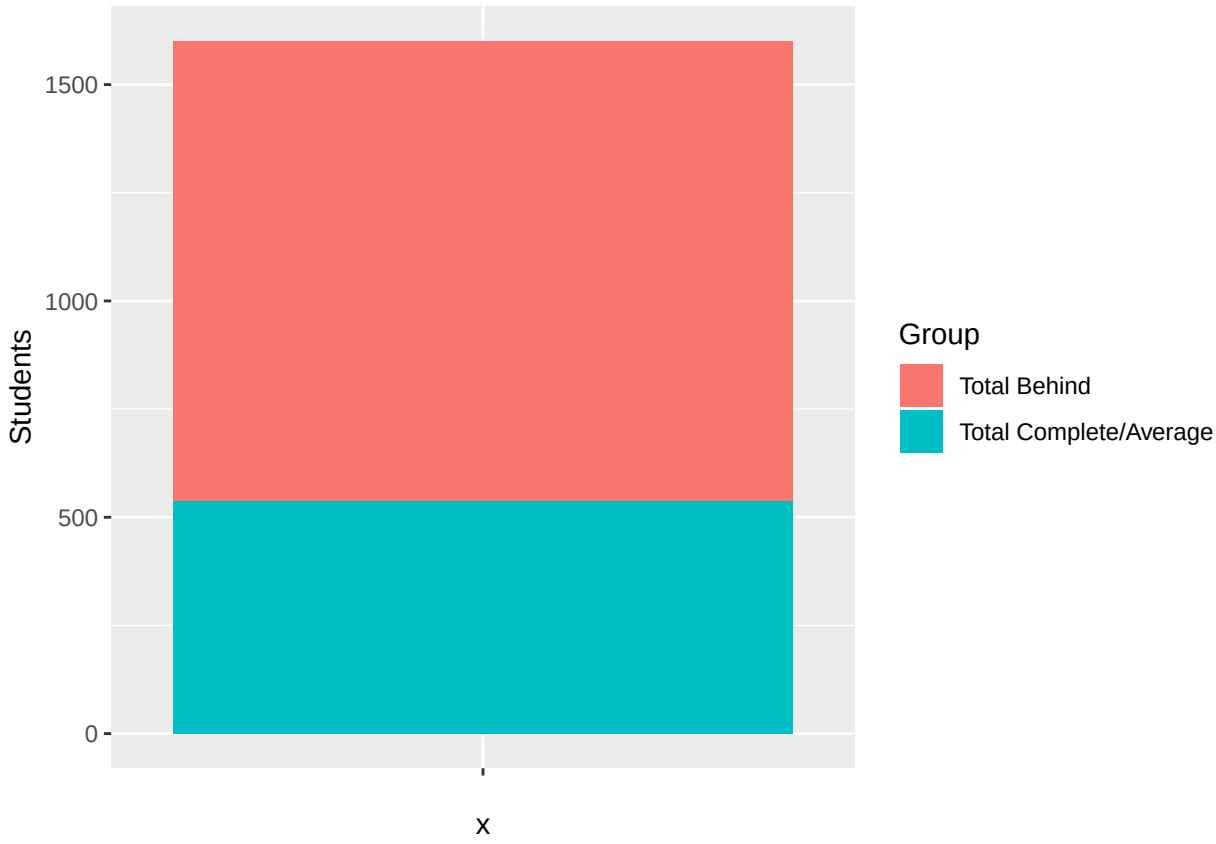



```
##Plot students per section by school and # of students
StudentsPerSection <- data.frame(SumBySchoolDF$School,SumOfStudents/SumBySchoolDF$Section)
ggplot(StudentsPerSection, aes(x=SumBySchoolDF.School,y=SumOfStudents.SumBySchoolDF.Section,fill=SumOfS
  geom_bar(width=1,stat="identity") + ylab("Students Per Section") + xlab("School")
```

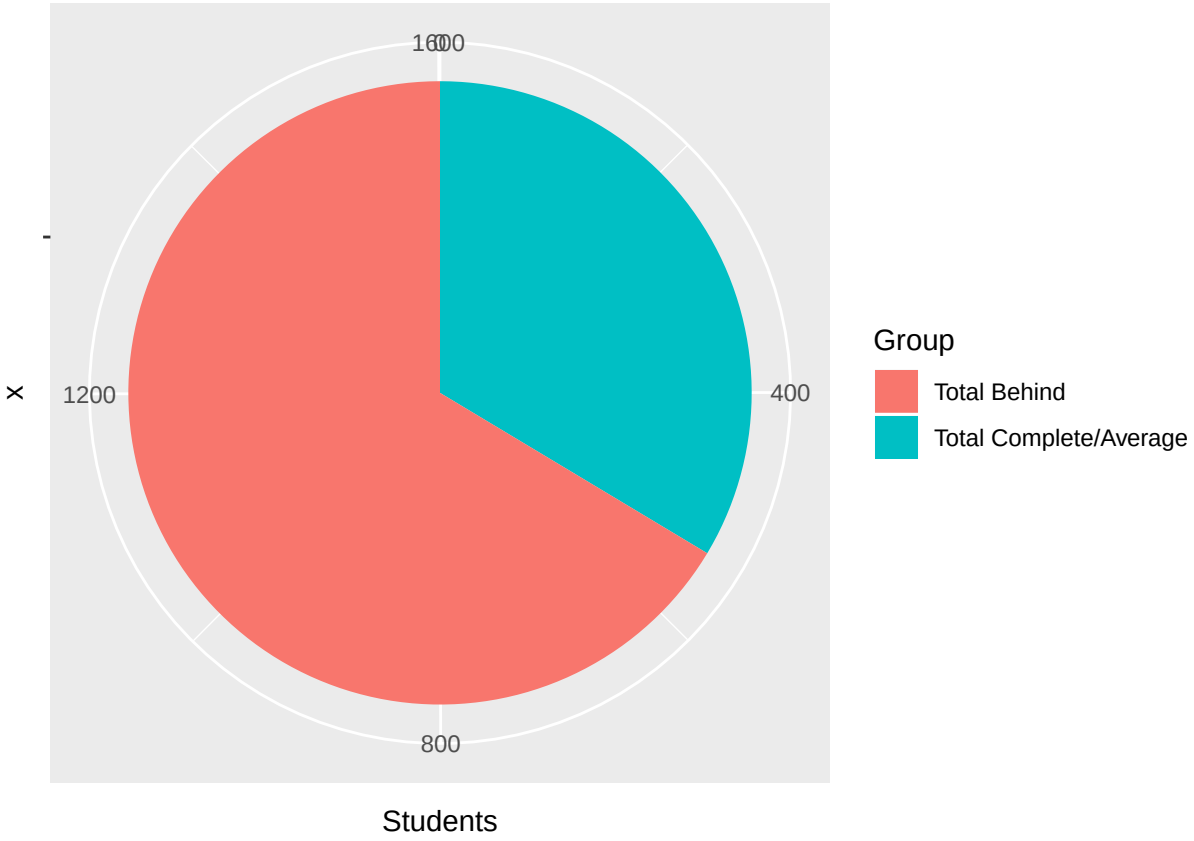



```
##Determine the total number of students completed vs behind

##Visually compare the total likely to complete (Complete+Average)
##and not likely to complete (behind, more behind, very behind)
##The bar and pie graph code is based on examples from sthda.com
(bp <- ggplot(CompleteBehind, aes(x="",y=Students,fill=Group))+
  geom_bar(width=1,stat="identity"))
```

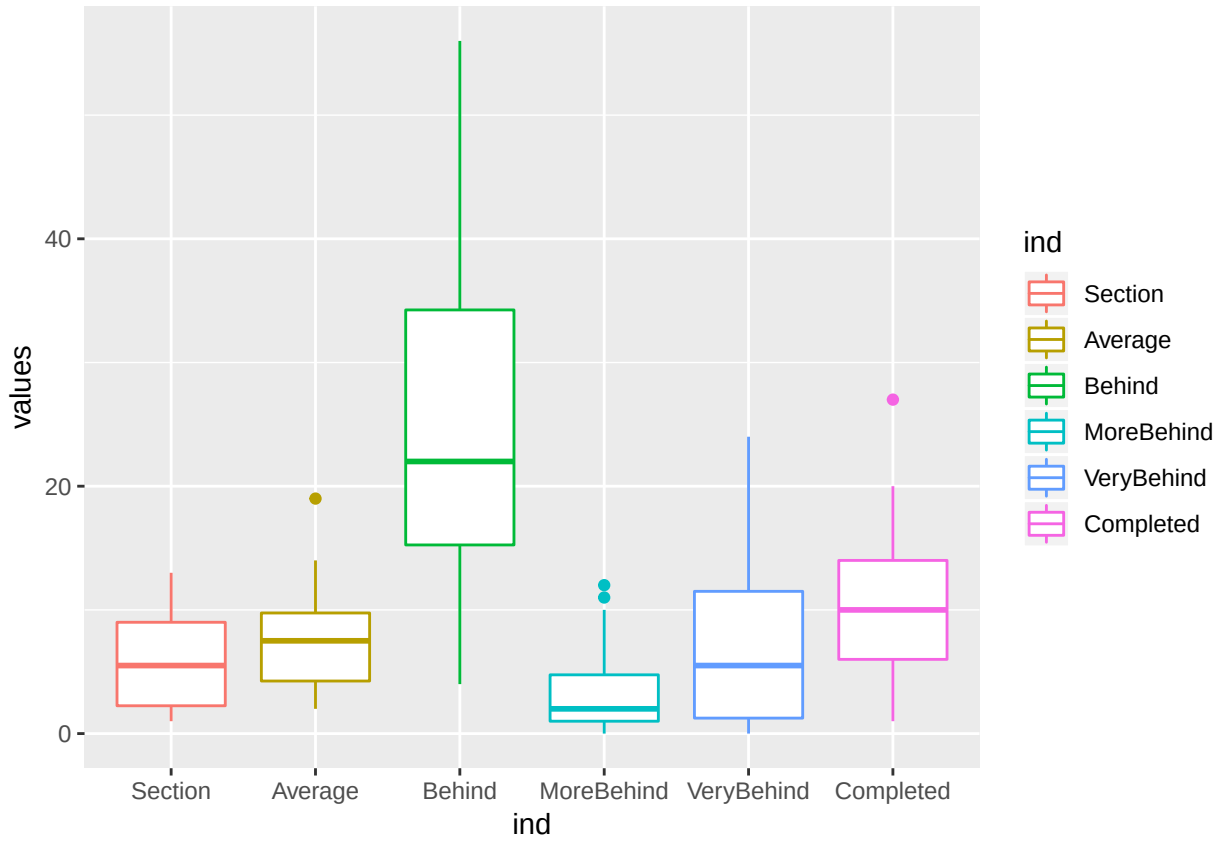



```
(pie <- bp + coord_polar("y", start=0))
```




```
## Use visual EDA - boxplots
## What does this tell us? There is variability from school to school and
##section to section in the completion rates
ggplot(stack(MyStoryData), aes(x = ind, y = values, color=ind)) +
  geom_boxplot()
```

```
## Warning in stack.data.frame(MyStoryData): non-vector columns will be
## ignored
```




```
## School A Boxplot
MyStoryData$School =="A"
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [12]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
JustSchoolA<-subset(MyStoryData, School == "A" )
(JustSchoolA)
```

```
##    School Section Average Behind MoreBehind VeryBehind Completed
## 1       A       1       5     54          3          9        10
## 2       A       2       8     40         10         16         6
## 3       A       3       9     35         12         13        11
## 4       A       4      14     44          5         12        10
## 5       A       5       9     42          2         24         8
## 6       A       6       7     29          3         10         9
## 7       A       7      19     22          5         14        19
## 8       A       8       3     37         11         18         5
## 9       A       9       6     29          8         12        10
## 10      A      10      13     40          5          5        20
## 11      A      11       8     32          4         10        15
## 12      A      12       2     16          2          3        14
## 13      A      13      10     30          3          8         5
##    SchoolSize
## 1       Large
## 2       Large
## 3       Large
```

```
## 4       Large
## 5       Large
## 6       Large
## 7       Large
## 8       Large
## 9       Large
## 10      Large
## 11      Large
## 12      Large
## 13      Large
```

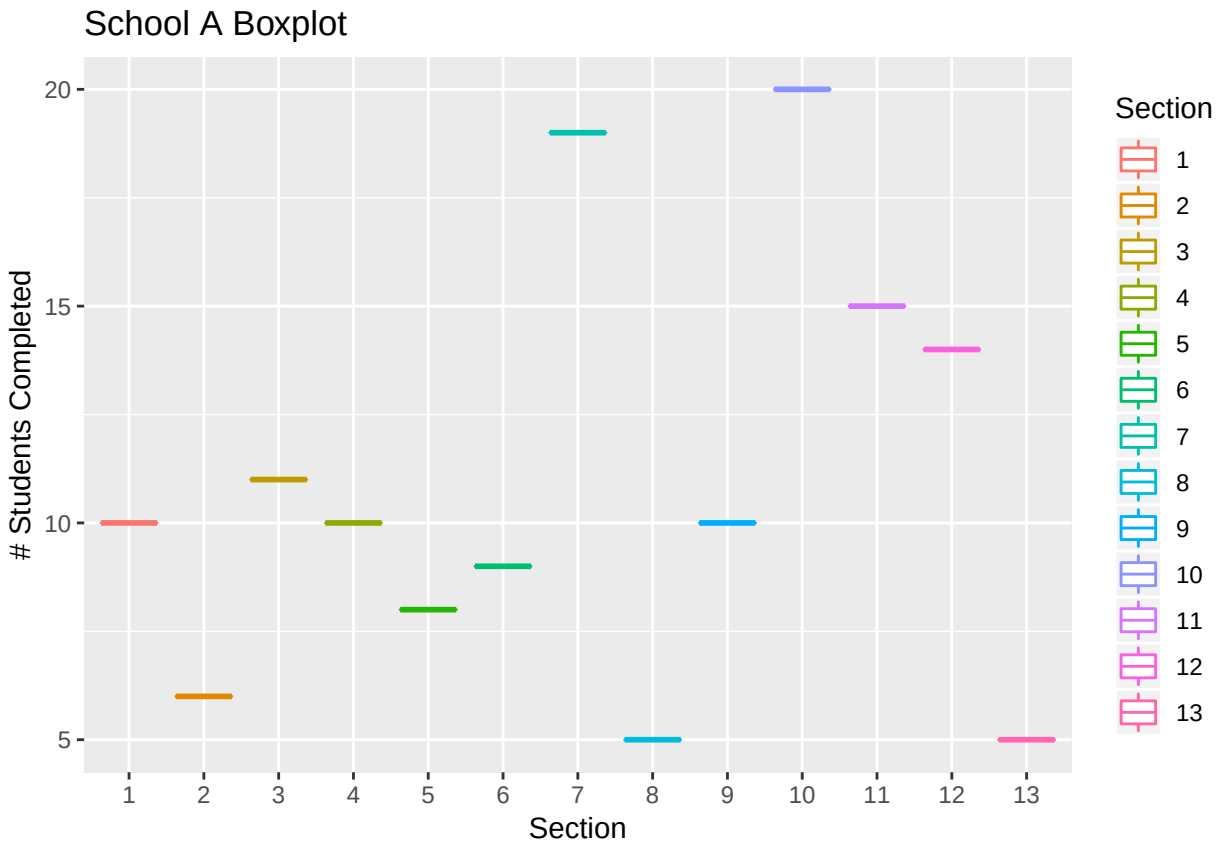
```
(str(JustSchoolA))
```

```
## 'data.frame':    13 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 ...
##  $ Average   : int  5 8 9 14 9 7 19 3 6 13 ...
##  $ Behind    : int  54 40 35 44 42 29 22 37 29 40 ...
##  $ MoreBehind: int  3 10 12 5 2 3 5 11 8 5 ...
##  $ VeryBehind: int  9 16 13 12 24 10 14 18 12 5 ...
##  $ Completed : int  10 6 11 10 8 9 19 5 10 20 ...
##  $ SchoolSize: Factor w/ 2 levels "Large","Small": 1 1 1 1 1 1 1 1 1 1 ...
```

```
## NULL
```

```
## Change Section to a factor
JustSchoolA$Section<-as.factor(JustSchoolA$Section)

ggplot(JustSchoolA, aes(x = Section, y = Completed, color=Section)) +
  geom_boxplot() + ggtitle("School A Boxplot") + ylab("# Students Completed")
```




```
## School B Boxplot
MyStoryData$School =="B"
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [23]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
```

```
JustSchoolB<-subset(MyStoryData, School == "B" )
(JustSchoolB)
```

```
##    School Section Average Behind MoreBehind VeryBehind Completed
## 14      B       1       4     22          0          6         7
## 15      B       2       5      7          2          1         3
## 16      B       3       6     31          1          1         8
## 17      B       4       4      7          0          0         7
## 18      B       5       8     14          4          0        14
## 19      B       6       8     11          1          2        18
## 20      B       7       9     21          0          2        13
## 21      B       8      10     23          2          5         6
## 22      B       9      10     21          0          3         5
## 23      B      10       3      8          1          1        15
## 24      B      11       7     19          2          1        10
## 25      B      12      10     17          1          0        19
##    SchoolSize
## 14      Large
## 15      Large
## 16      Large
## 17      Large
```

```
## 18      Large
## 19      Large
## 20      Large
## 21      Large
## 22      Large
## 23      Large
## 24      Large
## 25      Large
```

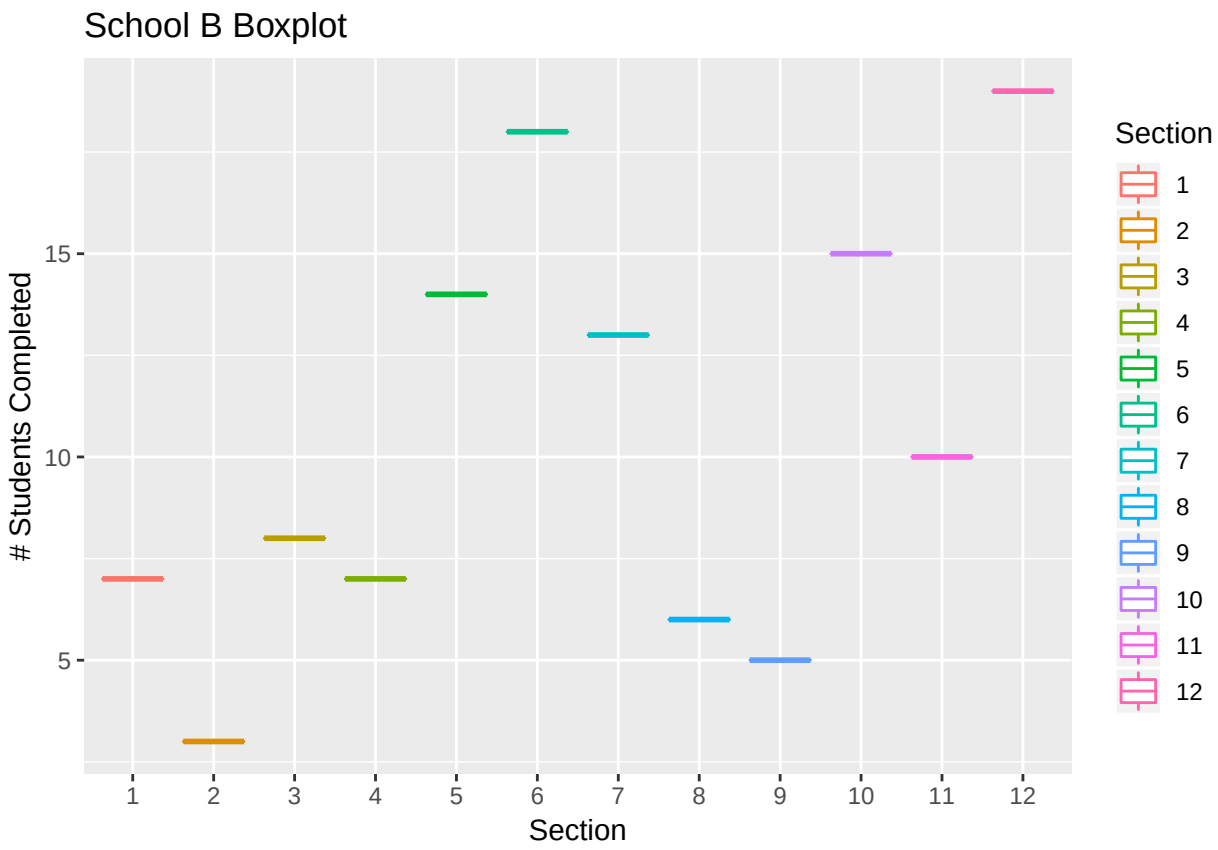
```
(str(JustSchoolB))
```

```
## 'data.frame':    12 obs. of  8 variables:
##  $ School    : Factor w/ 5 levels "A","B","C","D",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Section   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Average   : int  4 5 6 4 8 8 9 10 10 3 ...
##  $ Behind    : int  22 7 31 7 14 11 21 23 21 8 ...
##  $ MoreBehind: int  0 2 1 0 4 1 0 2 0 1 ...
##  $ VeryBehind: int  6 1 1 0 0 2 2 5 3 1 ...
##  $ Completed : int  7 3 8 7 14 18 13 6 5 15 ...
##  $ SchoolSize: Factor w/ 2 levels "Large","Small": 1 1 1 1 1 1 1 1 1 1 ...
```

```
## NULL
```

```
## Change Section to a factor
JustSchoolB$Section<-as.factor(JustSchoolB$Section)

ggplot(JustSchoolB, aes(x = Section, y = Completed, color=Section)) +
  geom_boxplot() + ggtitle("School B Boxplot") + ylab("# Students Completed")
```



```
## School C Boxplot
MyStoryData$School =="C"
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [23] FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE
```

```
JustSchoolC<-subset(MyStoryData, School == "C" )
(JustSchoolC)
```

```
##    School Section Average Behind MoreBehind VeryBehind Completed
## 26      C       1       2     15          2          4        13
## 27      C       2       7     20          1          7         1
## 28      C       3       2      4          1          1         5
##    SchoolSize
## 26      Small
## 27      Small
## 28      Small
```

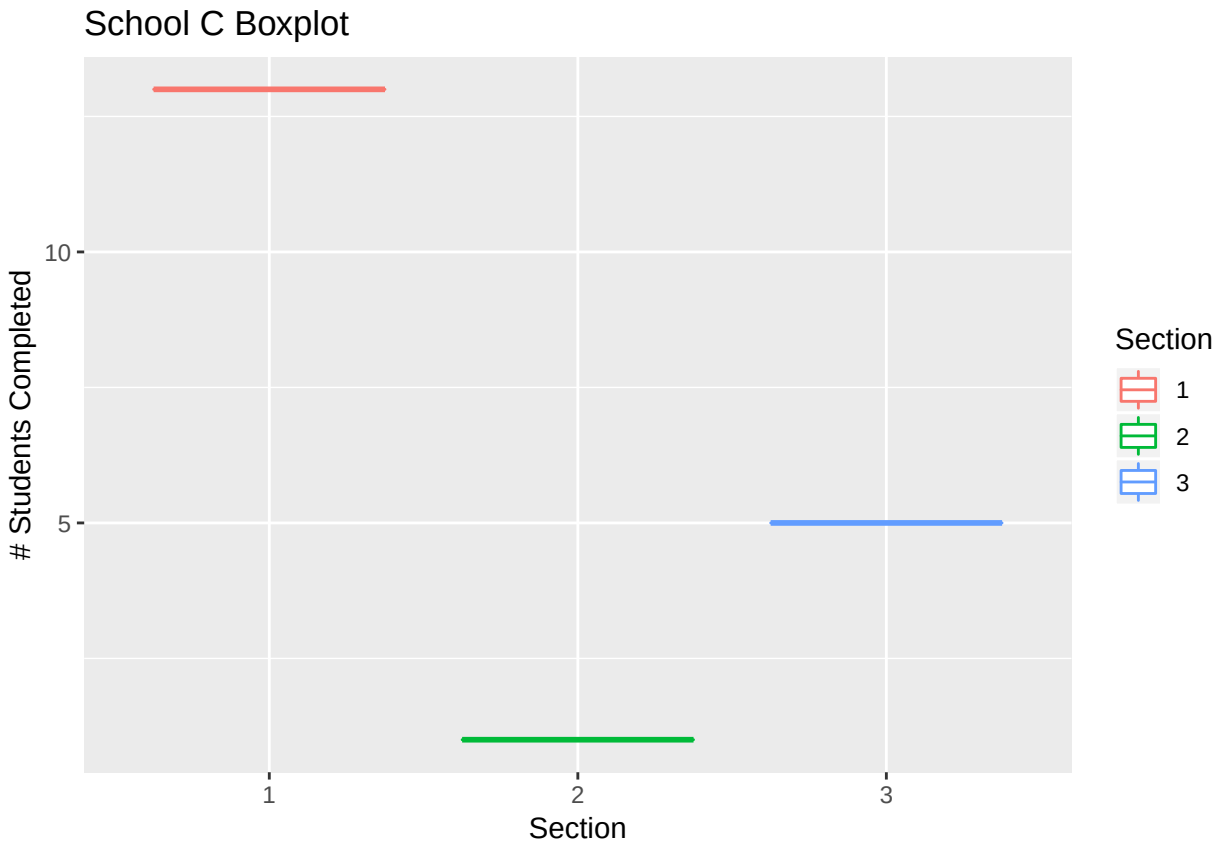
```
(str(JustSchoolC))
```

```
## 'data.frame':    3 obs. of  8 variables:
##  $ School    : Factor w/ 5 levels "A","B","C","D",..: 3 3 3
##  $ Section   : int  1 2 3
##  $ Average   : int  2 7 2
##  $ Behind    : int  15 20 4
##  $ MoreBehind: int  2 1 1
##  $ VeryBehind: int  4 7 1
##  $ Completed : int  13 1 5
##  $ SchoolSize: Factor w/ 2 levels "Large","Small": 2 2 2
```

```
## NULL
```

```
## Change Section to a factor
JustSchoolC$Section<-as.factor(JustSchoolC$Section)

ggplot(JustSchoolC, aes(x = Section, y = Completed, color=Section)) +
  geom_boxplot() + ggtitle("School C Boxplot") + ylab("# Students Completed")
```




```
## School D Boxplot
MyStoryData$School =="D"
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [23] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
```

```
JustSchoolD<-subset(MyStoryData, School == "D" )
(JustSchoolD)
```

```
##    School Section Average Behind MoreBehind VeryBehind Completed
## 29      D       1       3      8          2          6         3
##    SchoolSize
## 29      Small
```

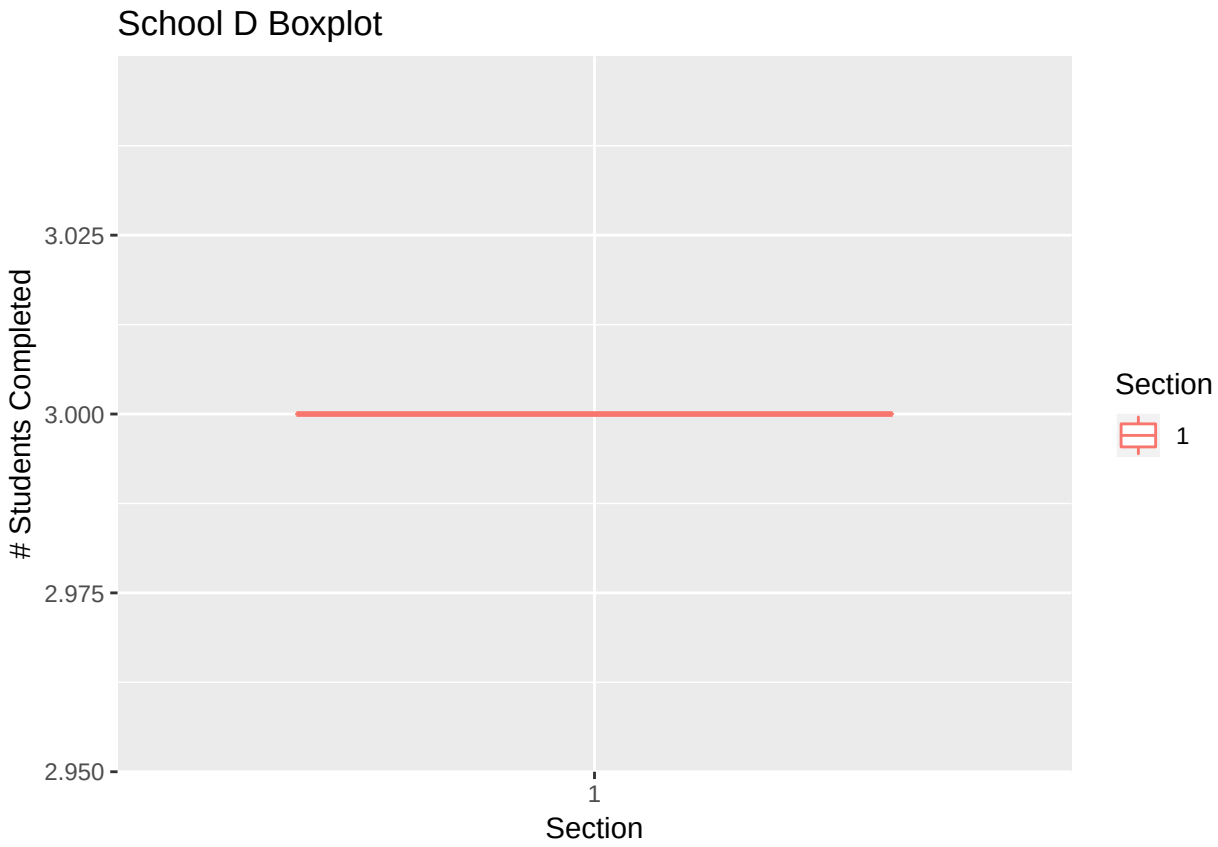
```
(str(JustSchoolD))
```

```
## 'data.frame':    1 obs. of  8 variables:
##  $ School    : Factor w/ 5 levels "A","B","C","D",..: 4
##  $ Section   : int 1
##  $ Average   : int 3
##  $ Behind    : int 8
##  $ MoreBehind: int 2
##  $ VeryBehind: int 6
##  $ Completed : int 3
##  $ SchoolSize: Factor w/ 2 levels "Large","Small": 2
```

```
## NULL
```

```
## Change Section to a factor
JustSchoolD$Section<-as.factor(JustSchoolD$Section)

ggplot(JustSchoolD, aes(x = Section, y = Completed, color=Section)) +
  geom_boxplot() + ggtitle("School D Boxplot") + ylab("# Students Completed")
```




```
## School E Boxplot
MyStoryData$School =="E"
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
```

```
JustSchoolE<-subset(MyStoryData, School == "E" )
(JustSchoolE)
```

```
##    School Section Average Behind MoreBehind VeryBehind Completed
## 30      E       1      11     56          7         15        27
##    SchoolSize
## 30      Small
```

```
(str(JustSchoolE))
```

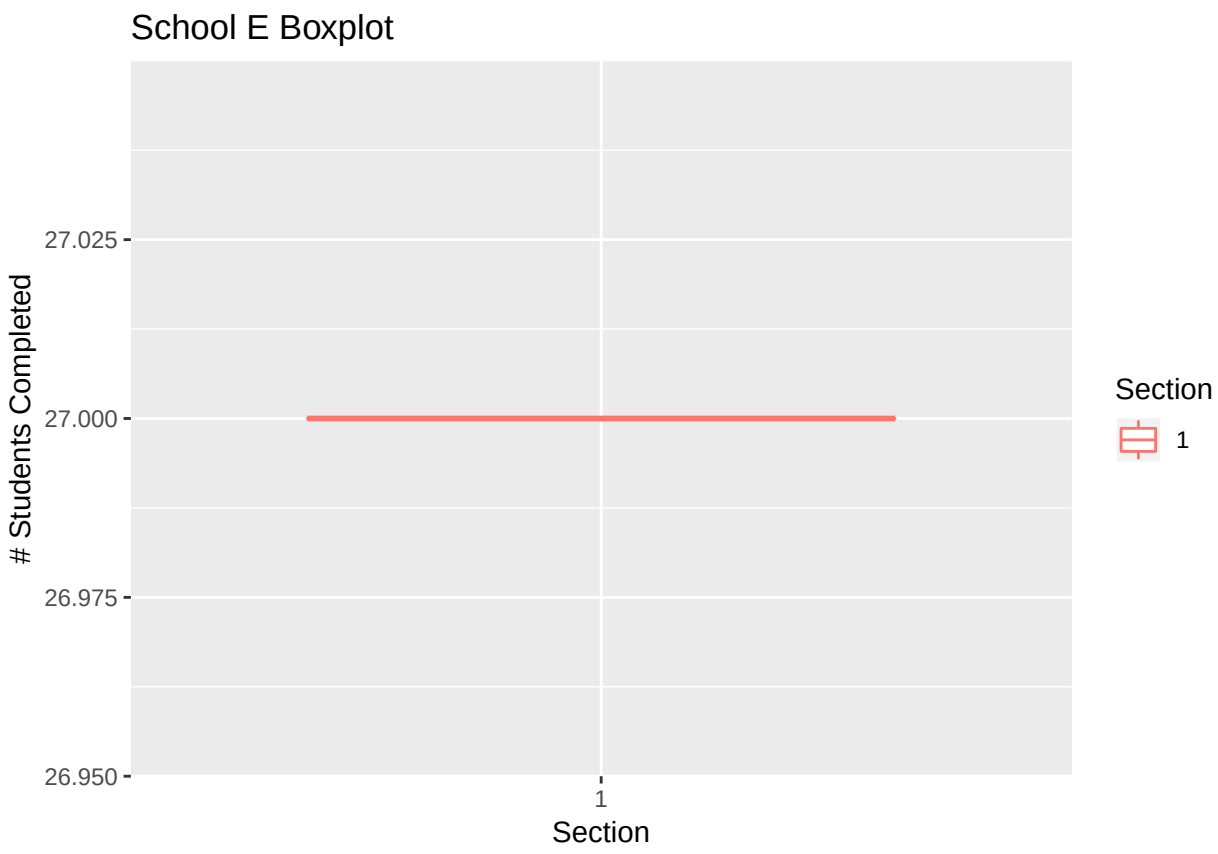
```
## 'data.frame':    1 obs. of  8 variables:
##  $ School    : Factor w/ 5 levels "A","B","C","D",..: 5
##  $ Section   : int 1
##  $ Average   : int 11
##  $ Behind    : int 56
##  $ MoreBehind: int 7
```

```
##  $ VeryBehind: int 15
##  $ Completed : int 27
##  $ SchoolSize: Factor w/ 2 levels "Large","Small": 2
```

```
## NULL
```

```
## Change Section to a factor
JustSchoolE$Section<-as.factor(JustSchoolE$Section)

ggplot(JustSchoolE, aes(x = Section, y = Completed, color=Section)) +
  geom_boxplot() + ggtitle("School E Boxplot") + ylab("# Students Completed")
```



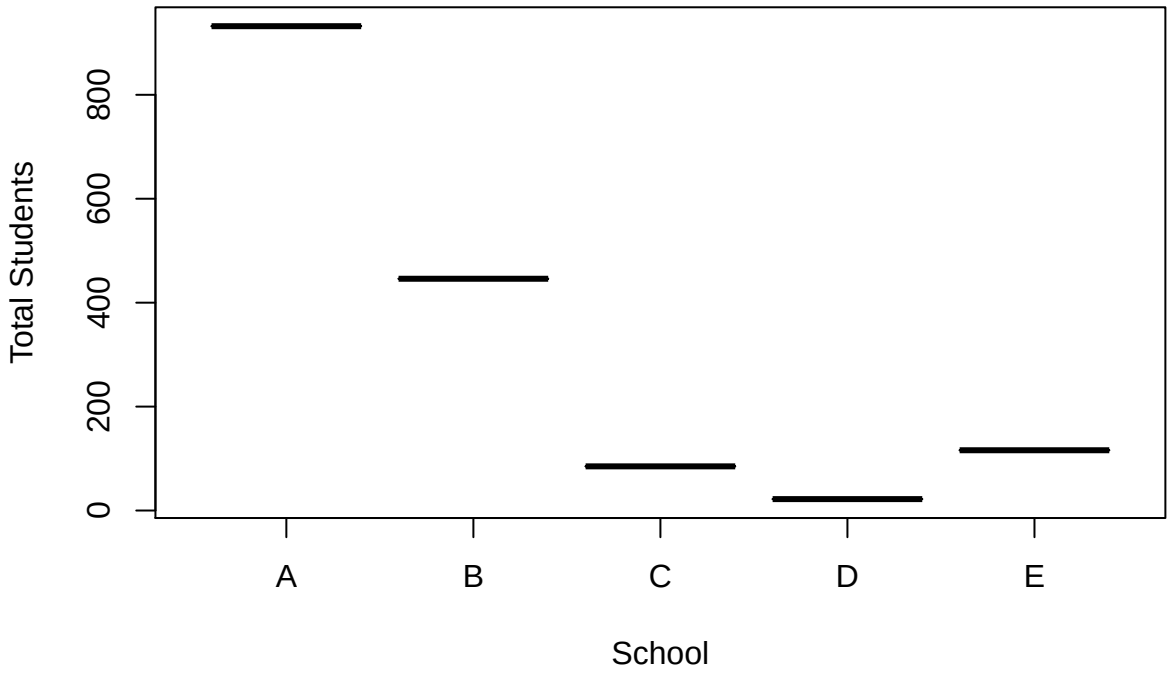

```
## The overall analysis goal is to understand the completion rate of students

# There was an error so changed to SumBySchoolDF$School from StudentsSumPerSchool$School
# Original code TotalPerSchool <- data.frame("School" = StudentsSumPerSchool$School,
#"Total" = SumOfStudents)
TotalPerSchool <- data.frame("School" = SumBySchoolDF$School,
                             "Total" = SumOfStudents)

(TotalPerSchool)
```

```
##   School Total
## 1      A   932
## 2      B   446
## 3      C    85
## 4      D    22
## 5      E   116
```

```
plot(TotalPerSchool$School,TotalPerSchool$Total,xlab="School",ylab="Total Students")
```




```
## This plot shows the % completion rate by school
ggplot(SumBySchoolDF, aes(x=School,y=Completed/TotalPerSchool$Total,fill=SumOfStudents))+
  geom_bar(width=1,stat="identity") + ylab("Percent Complete") + xlab("School")
```