HW2\_IST707

Hana Kim

1/23/2019

## Data-Storytelling

## Introduction

A data analysis is prepared in order to explore and discover useful information from the data set by using appropriate the data exploration and transformation techniques.

## Dataset

Each of 5 schools (A,B,C,D and E) is implementing the same math course this semester, with 35 lessons. There are 30 sections total.  
The semester is about 3/4 of the way through.  
The data set presents the number of students in the categories of Very Ahead, Middling, Behind, More Behind, Very Behind, and Completed in each sections by schools.  
Each 5 schools and its respective sections represent objects in the data set.  
The characteritic of Very Ahead column is students who are more than 5 lesssons ahead.  
The characteristic of Middling column is students who are 5 lessons ahead to 0 lessons ahead.  
The characteristic of Behind column is students who are 1 to 5 lessons behind.  
The characteristic of More Behind column is students who are 6 to 10 lessons behind.  
The characteristic of Very Behind column is students who are more than 10 lessons behind.  
The chracteristic of Completed column is students who are finished with the course.

schooldata <- read.csv("/Users/hana/Documents/data-storyteller (2).csv")   
colnames(schooldata) <- cbind("School", "Section", "Very.Ahead", "Middling", "Behind", "More.Behind","Very.Behind","Completed")   
str(schooldata)

## '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 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Middling : int 5 8 9 14 9 7 19 3 6 13 ...  
## $ Behind : int 54 40 35 44 42 29 22 37 29 40 ...  
## $ More.Behind: int 3 10 12 5 2 3 5 11 8 5 ...  
## $ Very.Behind: int 9 16 13 12 24 10 14 18 12 5 ...  
## $ Completed : int 10 6 11 10 8 9 19 5 10 20 ...

## Data Analysis

There are total of 8 variables, which are School, Section, Very.Ahead, Middling, Behind, More.Behind, Very.Behind, and Completed.  
Each variables has a corresponding data type  
School data type - Factor with 5 levels, A,B,C,D,E  
Section data type - Integer since each sections per schools has a number  
The column variables, Very.Ahead, Middling, Behind, More.Behind, Very.Behind, Complted - Integer which repreents the number of students in the columns

In order to explore and understand the data set, a new column of total which represents the total number of students are added.

schooldata$Total <- schooldata$Very.Ahead+schooldata$Middling+schooldata$Behind+schooldata$More.Behind+schooldata$Very.Behind+schooldata$Completed   
summary(schooldata)

## School Section Very.Ahead Middling Behind   
## A:13 Min. : 1.00 Min. :0 Min. : 2.00 Min. : 4.00   
## B:12 1st Qu.: 2.25 1st Qu.:0 1st Qu.: 4.25 1st Qu.:15.25   
## C: 3 Median : 5.50 Median :0 Median : 7.50 Median :22.00   
## D: 1 Mean : 5.90 Mean :0 Mean : 7.40 Mean :25.13   
## E: 1 3rd Qu.: 9.00 3rd Qu.:0 3rd Qu.: 9.75 3rd Qu.:34.25   
## Max. :13.00 Max. :0 Max. :19.00 Max. :56.00   
## More.Behind Very.Behind Completed Total   
## Min. : 0.000 Min. : 0.000 Min. : 1.00 Min. : 13.00   
## 1st Qu.: 1.000 1st Qu.: 1.250 1st Qu.: 6.00 1st Qu.: 37.50   
## Median : 2.000 Median : 5.500 Median :10.00 Median : 46.50   
## Mean : 3.333 Mean : 6.967 Mean :10.53 Mean : 53.37   
## 3rd Qu.: 4.750 3rd Qu.:11.500 3rd Qu.:14.00 3rd Qu.: 77.75   
## Max. :12.000 Max. :24.000 Max. :27.00 Max. :116.00

The statistical summary of school data does not reveal any meaningful information other than the total number of students in Very.Ahead category is none. In order to draw a insightful summary of each categorical columns, it is necessary to examine the percentage of each categories account for in the total.

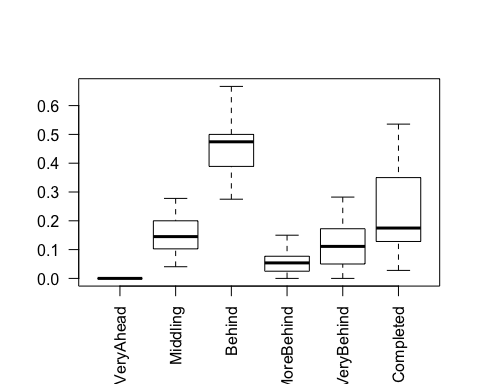
schooldata$Very.Ahead.percent <- schooldata$Very.Ahead / schooldata$Total  
schooldata$Middling.percent <- schooldata$Middling / schooldata$Total  
schooldata$Behind.percent <- schooldata$Behind / schooldata$Total  
schooldata$More.Behind.percent <- schooldata$More.Behind / schooldata$Total  
schooldata$Very.Behind.percent <- schooldata$Very.Behind / schooldata$Total  
schooldata$Completed.percent <- schooldata$Completed / schooldata$Total

With the counts of students converted to a percentage, now the data set is prepared to be presented to draw an overview of each categories.  
To do so, a box plot is presented below. The data set school.data is generated from the schooldata to include only the percentage variables.

school.data <- schooldata[,10:15]  
str(school.data)

## 'data.frame': 30 obs. of 6 variables:  
## $ Very.Ahead.percent : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Middling.percent : num 0.0617 0.1 0.1125 0.1647 0.1059 ...  
## $ Behind.percent : num 0.667 0.5 0.438 0.518 0.494 ...  
## $ More.Behind.percent: num 0.037 0.125 0.15 0.0588 0.0235 ...  
## $ Very.Behind.percent: num 0.111 0.2 0.163 0.141 0.282 ...  
## $ Completed.percent : num 0.1235 0.075 0.1375 0.1176 0.0941 ...

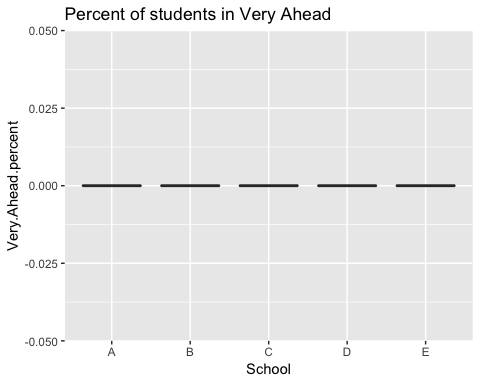
boxplot(school.data, las=2, names = c("VeryAhead","Middling","Behind","MoreBehind","VeryBehind","Completed"))



The box plot show that the total number of students in Behind category account for about fifty percent, all other categories account for less then twenty percent.

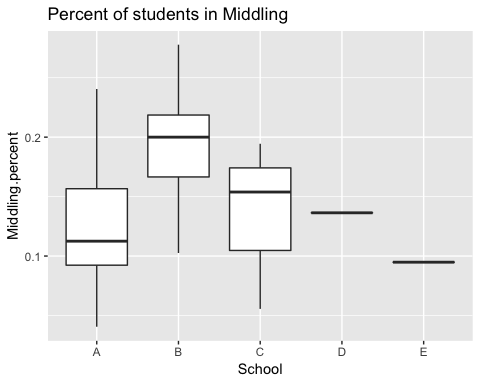
Now a represendtation of each categories by School is needed to explore how students are doing per each school

library(ggplot2)  
ggplot(schooldata) + geom\_boxplot(aes(x=School, y=Very.Ahead.percent)) + labs(title = "Percent of students in Very Ahead")



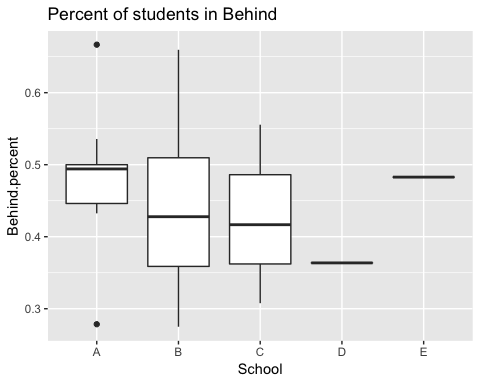
Not suprisingly, the all of the boxplot is flat. This is expected as seen in the schooldata summarmy where zero students are accounted in this column.

ggplot(schooldata) + geom\_boxplot(aes(x=School, y=Middling.percent)) + labs(title = "Percent of students in Middling")



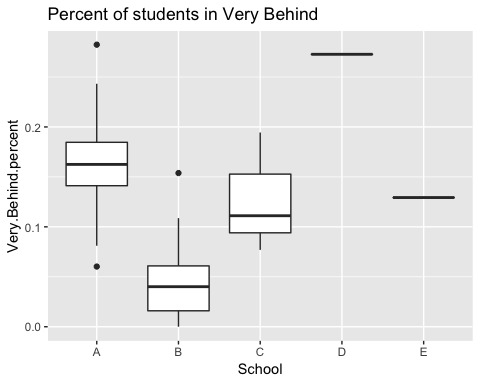
The box plot shows that 11% of students from School A, 20% of students from School B, 15% of students from School C, 14% students from School D, and 9% of students from School E are is in Middling column.

ggplot(schooldata) + geom\_boxplot(aes(x=School, y=Behind.percent)) + labs(title = "Percent of students in Behind")



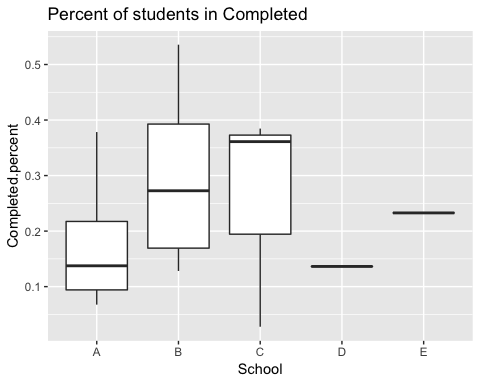
Similarly, the box plot is shown for Behind Column for each school

ggplot(schooldata) + geom\_boxplot(aes(x=School, y=Very.Behind.percent)) + labs(title = "Percent of students in Very Behind")



The box plot is shown for Very Behind Column for each school

ggplot(schooldata) + geom\_boxplot(aes(x=School, y=Completed.percent)) + labs(title = "Percent of students in Completed")



attach(schooldata)  
schoolaggregate <- aggregate(cbind(Very.Ahead,Middling,Behind,Very.Behind,Completed),by=list(School=School),FUN=sum)  
schoolaggregate

## School Very.Ahead Middling Behind Very.Behind Completed  
## 1 A 0 113 450 154 142  
## 2 B 0 84 201 22 125  
## 3 C 0 11 39 12 19  
## 4 D 0 3 8 6 3  
## 5 E 0 11 56 15 27

## Conclusion

The summary data of schoolaggregate is genereated in order to show the total number of students in the columns.