IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019

Homework 2

This is a real job interview question from a data analysis company, and I doubt there is a standard answer to this question. So feel free to explore your story by using the data exploration and transformation techniques appropriately.

----------instruction quote begins----------

Here is a small dataset for you to work with.

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**.

For each section, we record the number of students who are:

- very ahead (more than 5 lessons ahead)
- middling (5 lessons ahead to 0 lessons ahead)
- behind (1 to 5 lessons behind)
- more behind (6 to 10 lessons behind)
- very behind (more than 10 lessons behind)
- completed (finished with the course)

What's the story (or stories) in this data? Find it, and tell it visually and, above all, truthfully.

----------instruction quote ends----------

In this exercise, I want to tell a story with data. So, instead of just exploring the dataset and providing multiple independent visualizations, I have decided to look for answers of the two following questions:

1. **How does the status of students per school differ after ¾ of the semester?**
2. **Which factor(s) affect the different study speeds?**

1. ***Data Import***
   We begin the analysis by importing the dataset. I took a brief look at the dataset in R Studio and identified an existing header. Consequently, I set the import statement with header = TRUE

```
> data.storyteller <- read.csv("C:/Users/steff/Google Drive/3 Masters/3. Semester/2 Data Analytics/Homework/HW2/data-storyteller.csv", header=TRUE )
```

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019

2. *Loading Packages*

I use five packages for my analysis

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(ggpubr)
library(reshape2)
```

3. Analyze the structure of the dataset

I paid special attention to the data type of each attribute and decided to not change any data type at this point.

```
> 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..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 ...
```

Moreover, I did not identify any empty cells:

```
> summary(schooldata)
 School    Section      Very.Ahead..5  Middling..0     Behind..1.5    More.Behind..6.10 Very.Behind..11    Completed
 A:13   Min.   : 1.00   Min.   :0     Min.   : 2.00   Min.   : 4.00   Min.   : 0.000    Min.   : 0.000   Min.   : 1.00
 B:12   1st Qu.: 2.25   1st Qu.:0     1st Qu.: 4.25   1st Qu.:15.25   1st Qu.: 1.000    1st Qu.: 1.250   1st Qu.: 6.00
 C: 3   Median : 5.50   Median :0     Median : 7.50   Median :22.00   Median : 2.000    Median : 5.500   Median :10.00
 D: 1   Mean   : 5.90   Mean   :0     Mean   : 7.40   Mean   :25.13   Mean   : 3.333    Mean   : 6.967   Mean   :10.53
 E: 1   3rd Qu.: 9.00   3rd Qu.:0     3rd Qu.: 9.75   3rd Qu.:34.25   3rd Qu.: 4.750    3rd Qu.:11.500   3rd Qu.:14.00
        Max.   :13.00   Max.   :0     Max.   :19.00   Max.   :56.00   Max.   :12.000    Max.   :24.000   Max.   :27.00
```

However, the header is difficult to read.

```
> head(schooldata)
  School Section Very.Ahead..5 Middling..0 Behind..1.5 More.Behind..6.10 Very.Behind..11 Completed
1      A       1             0           5          54                 3               9        10
2      A       2             0           8          40                10              16         6
3      A       3             0           9          35                12              13        11
4      A       4             0          14          44                 5              12        10
5      A       5             0           9          42                 2              24         8
6      A       6             0           7          29                 3              10         9
```

As each attribute is already described in the instruction quote, I excluded the numbers behind the header to improve clarity.

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019

```
> names <- c("School", "Section", "Veryahead", "Middling", "Behind", "MoreBehind", "VeryBehind", "Completed")
> colnames(schooldata) <- names
>
> head(schooldata)
  School Section Veryahead Middling Behind MoreBehind VeryBehind Completed
1      A       1         0        5     54          3          9        10
2      A       2         0        8     40         10         16         6
3      A       3         0        9     35         12         13        11
4      A       4         0       14     44          5         12        10
5      A       5         0        9     42          2         24         8
6      A       6         0        7     29          3         10         9
>
```

1. **How does the status of students per school differ after ¾ of the semester?**

I restructured the dataset to gain an overview about the number of students, studying at which school -at which section and -on what level:

```
> # Restructure the Dataset
> schooldata_new <- gather(schooldata,status,number,3:8)
>
> #lets look at the new structure
> str(schooldata_new)
'data.frame':   180 obs. of  4 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 ...
 $ status : chr  "Veryahead" "Veryahead" "Veryahead" "Veryahead" ...
 $ number : int  0 0 0 0 0 0 0 0 0 0 ...
>
```

In order to examine the performance of the schools and the sections, I need to pay special attention the status of each observation in the new table "schooldata_new".

I have decided to introduce levels for the attribute status as I can order them in regard to the progress of the students. Then, I convert the data type "character" to "factor" and I assign a number for each of the level. This means, "Completed" =1, "Veryahead" = 2, "Middling" = 3, "Behind" = 4, "MoreBehind" = 5 and "VeryBehind" = 6.

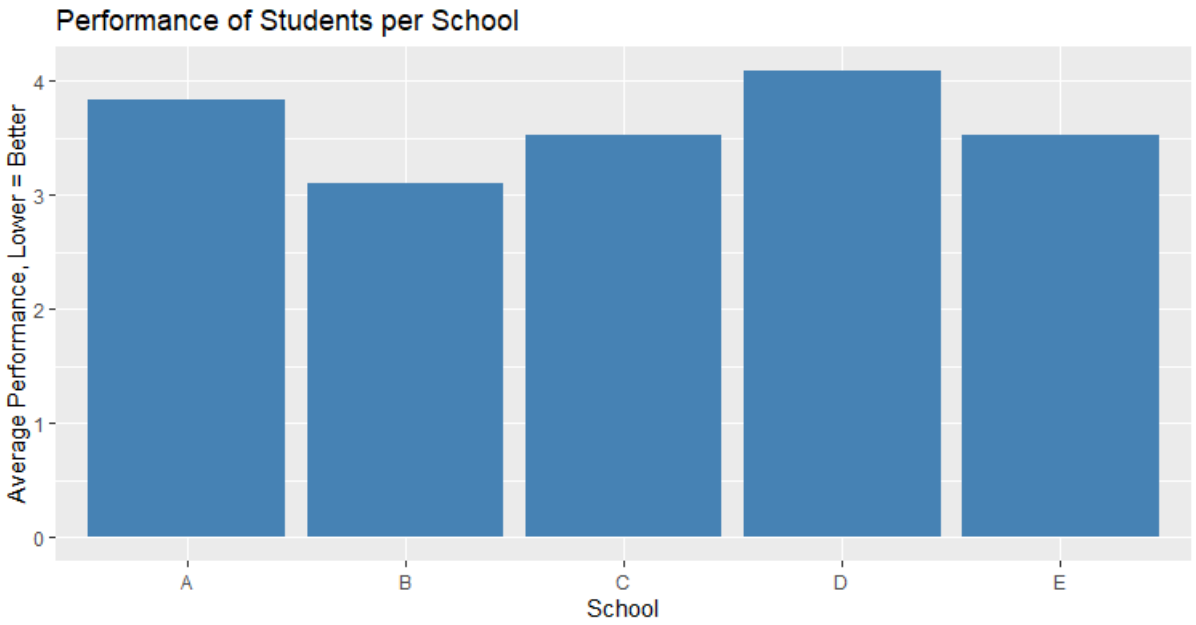
```
> # change the data type of status from character to factor by introducing levels and converting them to numbers
> levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind', 'VeryBehind')
> schooldata_new$status <- factor(schooldata_new$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind', 'VeryBehi
'))
> str(schooldata_new)
'data.frame':   180 obs. of  4 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 ...
 $ status : Factor w/ 6 levels "Completed","Veryahead",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ number : int  0 0 0 0 0 0 0 0 0 0 ...
>
> schooldata_new2 <- schooldata_new %>% mutate_at(vars(status), as.numeric)
> str(schooldata_new2)
'data.frame':   180 obs. of  4 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 ...
 $ status : num  2 2 2 2 2 2 2 2 2 2 ...
 $ number : int  0 0 0 0 0 0 0 0 0 0 ...
```

*To be able to make a statement about the performance/status/speed of each school, we need the average status for each school. Therefore, I extend the table "schooldata_new3" by another column that consists of status * number and calculate the average by summing it up and dividing it by the total number of students per school.*

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019

```
> #How does each school performe?
> schoolsummary2 <- schooldata_new3 %>% group_by(School) %>% summarize(status1 = sum(status2)/sum(number))
> ggplot(schoolsummary2, aes(x = School, y = status1)) + geom_bar(stat="identity", fill="steelblue") + ylab("Average Performance, Lower = B
ter") + ggtitle("Performance of Students per School")
```



The visualization shows that the students in school B have made on average the most progress after ¾ of the semester. Contrary, on average the students in School A and school B have made the least progress.

To gain a more detailed look of each school and identify differences in between them, I will visualize the share of students who fall in each status level.

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

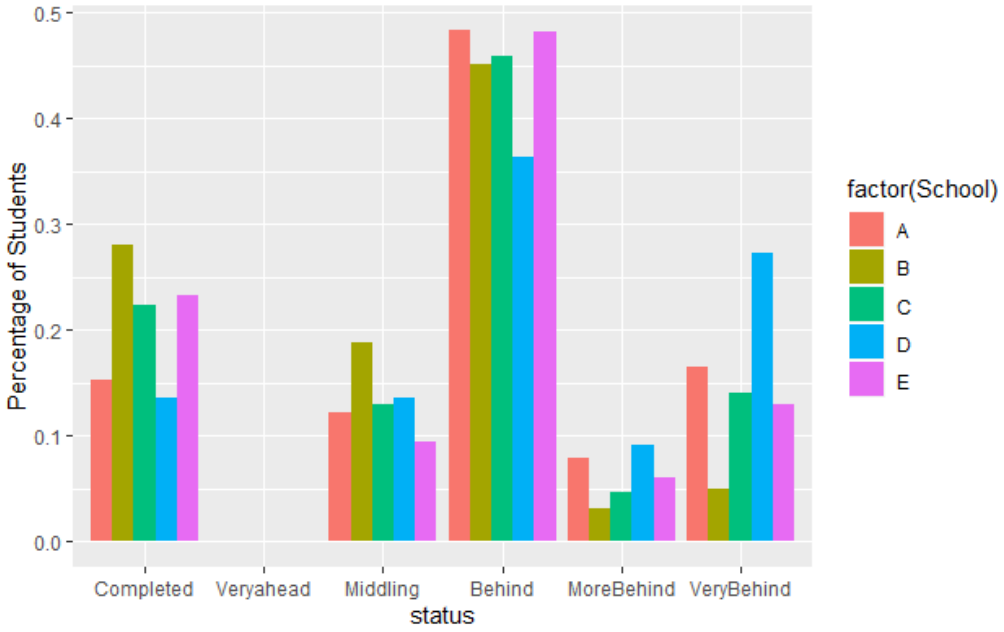
09/09/2019

```
> schoola <- filter(schooldata_new, School == "A")
> schoola$status <- factor(schoola$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind',
VeryBehind'))
> schoola1 <- schoola %>% group_by(School, status) %>% summarize(numStudents = sum(number)/932)
> a<-ggplot(data=schoola1, aes(x=status, y=numStudents)) +
+   geom_bar(stat="identity", fill="steelblue") + ylab("Percentage of Students") + ggtitle("School A") + ylim(0, 0.5
>
> schoolb <- filter(schooldata_new, School == "B")
> schoolb$status <- factor(schoolb$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind',
VeryBehind'))
> schoolb1 <- schoolb %>% group_by(School, status) %>% summarize(numStudents = sum(number)/446)
> b<-ggplot(data=schoolb1, aes(x=status, y=numStudents)) +
+   geom_bar(stat="identity", fill="steelblue") + ylab("Percentage of Students") + ggtitle("School B") + ylim(0, 0.5
>
> schoolc <- filter(schooldata_new, School == "C")
> schoolc$status <- factor(schoolc$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind',
VeryBehind'))
> schoolc1 <- schoolc %>% group_by(School, status) %>% summarize(numStudents = sum(number)/85)
> c<-ggplot(data=schoolc1, aes(x=status, y=numStudents)) +
+   geom_bar(stat="identity", fill="steelblue") + ylab("Percentage of Students") + ggtitle("School C") + ylim(0, 0.5
>
> schoold <- filter(schooldata_new, School == "D")
> schoold$status <- factor(schoold$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind',
VeryBehind'))
> schoold1 <- schoold %>% group_by(School, status) %>% summarize(numStudents = sum(number)/22)
> d<-ggplot(data=schoold1, aes(x=status, y=numStudents)) +
+   geom_bar(stat="identity", fill="steelblue") + ylab("Percentage of Students") + ggtitle("School D") + ylim(0, 0.5
>
> schoole <- filter(schooldata_new, School == "E")
> schoole$status <- factor(schoole$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind',
VeryBehind'))
> schoole1 <- schoole %>% group_by(School, status) %>% summarize(numStudents = sum(number)/116)
> e<-ggplot(data=schoole1, aes(x=status, y=numStudents)) +
+   geom_bar(stat="identity", fill="steelblue") + ylab("Percentage of Students") + ggtitle("School E") + ylim(0, 0.5
>
```

```
> jointdataset <- rbind(schoola1, schoolb1, schoolc1, schoold1, schoole1)
> jointdataset
# A tibble: 30 x 3
# Groups:   School [5]
   School status     numStudents
   <fct>  <fct>            <dbl>
 1 A      Completed        0.152
 2 A      Veryahead        0
 3 A      Middling         0.121
 4 A      Behind           0.483
 5 A      MoreBehind       0.0783
 6 A      VeryBehind       0.165
 7 B      Completed        0.280
 8 B      Veryahead        0
 9 B      Middling         0.188
10 B      Behind           0.451
# ... with 20 more rows
> abcde <- ggplot(jointdataset, aes(x=status, y=numStudents, fill=factor(School)))+
+   geom_bar(position="dodge", stat="identity") + ylab("Percentage of Students")
> abcde
```

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019



Here, you can see the difference of each school. It shows that school B has the highest share of students in status "completed" as well as in status "middling".

The different progress of each school will even get more obvious when comparing the share of students who are currently behind:

```
> schoola_Behind <- sum(schoola1$numStudents[4:6])
> schoolb_Behind <- sum(schoolb1$numStudents[4:6])
> schoolc_Behind <- sum(schoolc1$numStudents[4:6])
> schoold_Behind <- sum(schoold1$numStudents[4:6])
> schoole_Behind <- sum(schoole1$numStudents[4:6])
>
> schoola_Behind
[1] 0.7263948
> schoolb_Behind
[1] 0.5313901
> schoolc_Behind
[1] 0.6470588
> schoold_Behind
[1] 0.7272727
> schoole_Behind
[1] 0.6724138
```

This confirms the first plot about the average of the performance and shows that School B has the lowest share of students that are behind. School a and school d have the highest share of students who lack behind.

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019

2. Which factor(s) affect the different study speeds?

To identify the factors, we need to get a better understanding about each school. We can look at each of the school and summarize the available data:

```
> summary1 <- schooldata_new %>% group_by(School) %>% summarize(numSections = max(Section), numStudents = sum(number))
> summary1
# A tibble: 5 x 3
  School numSections numStudents
  <fct>        <int>       <int>
1 A               13         932
2 B               12         446
3 C                3          85
4 D                1          22
5 E                1         116
```

As you can see, School B has much more students than School D but, as we have seen earlier, the students in School B perform much better than the students in School D. Consequently, the pure number of students per school cannot be the reason for the different progress.

However, the number of sections differ and thus, I will take a closer look at the sections and the size of the sections:

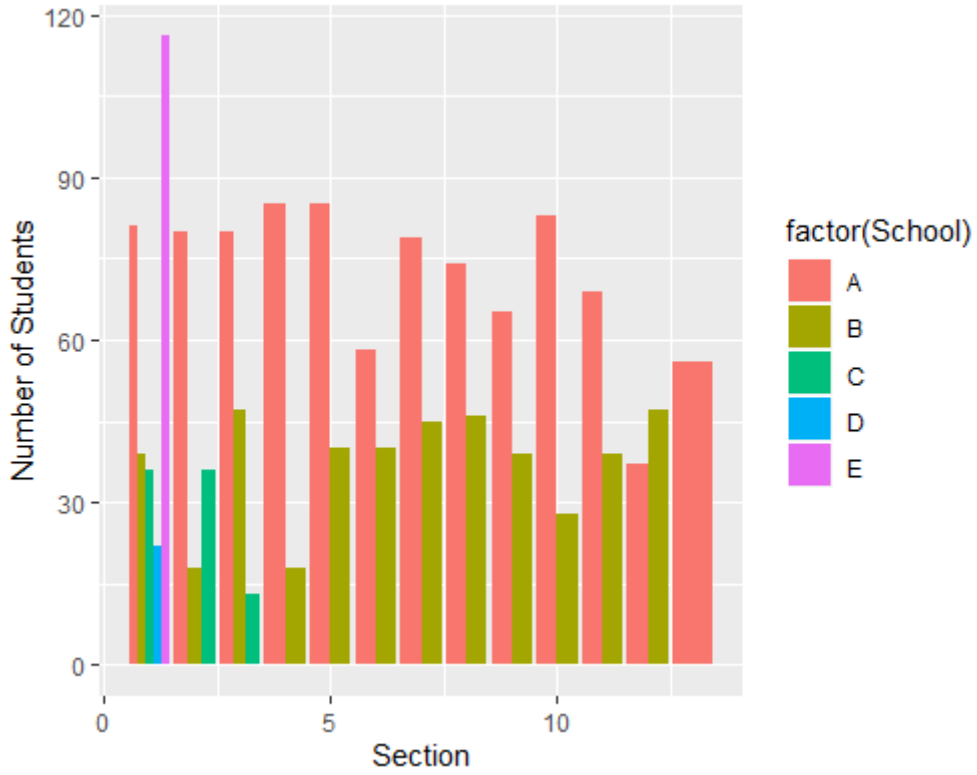
```
> #size of sections
> #individually
> summary3 <- schooldata_new %>% group_by(School, Section) %>% summarize(numStudents = sum(number))
> aa <- filter(summary3, School == "A")
> aaa <- ggplot(aa, aes(x=Section, y=numStudents)) + geom_bar(stat="identity", fill="steelblue") + ylim(0, 90)+ scal
_x_discrete(name ="Section", limits=c(0:15)) + ggtitle("School A")
>
> bb <- filter(summary3, School == "B")
> bbb <- ggplot(bb, aes(x=Section, y=numStudents)) + geom_bar(stat="identity", fill="steelblue") + ylim(0, 90) + sca
e_x_discrete(name ="Section", limits=c(0:15)) + ggtitle("School B")
>
> cc <- filter(summary3, School == "C")
> ccc <- ggplot(cc, aes(x=Section, y=numStudents)) + geom_bar(stat="identity", fill="steelblue") + ylim(0, 90) + sca
e_x_discrete(name ="Section", limits=c(0:15)) + ggtitle("School C")
>
> dd <- filter(summary3, School == "D")
> ddd <- ggplot(dd, aes(x=Section, y=numStudents)) + geom_bar(stat="identity", fill="steelblue") + ylim(0, 90) + sca
e_x_discrete(name ="Section", limits=c(0:15)) + ggtitle("School D")
>
> ee <- filter(summary3, School == "E")
> eee <- ggplot(ee, aes(x=Section, y=numStudents)) + geom_bar(stat="identity", fill="steelblue") + scale_x_discrete(
ame ="Section", limits=c(0:15)) + ggtitle("School E")
```

Instead of showing each school individually, I will compare them in one diagram:

```
> abcde1 <- ggplot(summary3, aes(x=Section, y=numStudents, fill=factor(School)))+
+   geom_bar(position="dodge", stat="identity") + ylab("Number of Students")
> abcde1
```

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019



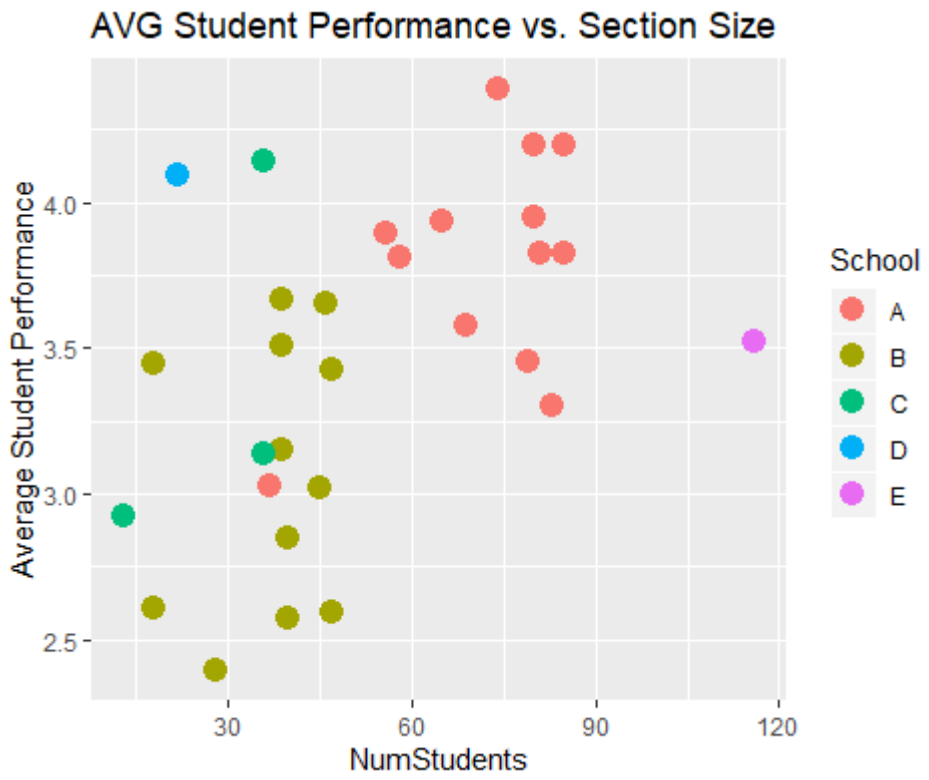
It becomes clear that the number of students per section differ. A and E have the highest numbers have students per section.

Possibly, the number of students per section can have an effect on the performance of the students. Therefore, I will visualize the number of students per section and compare it with the average progress of the section:

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019

```
> ggplot(schoolsummary, aes(x = NumStudents, y = status1, color = School, group = School)) + ylab("Average Student Performance") + ggtitle("AVG Student Performance vs. Se
on Size") + geom_point(lwd = 4)
```

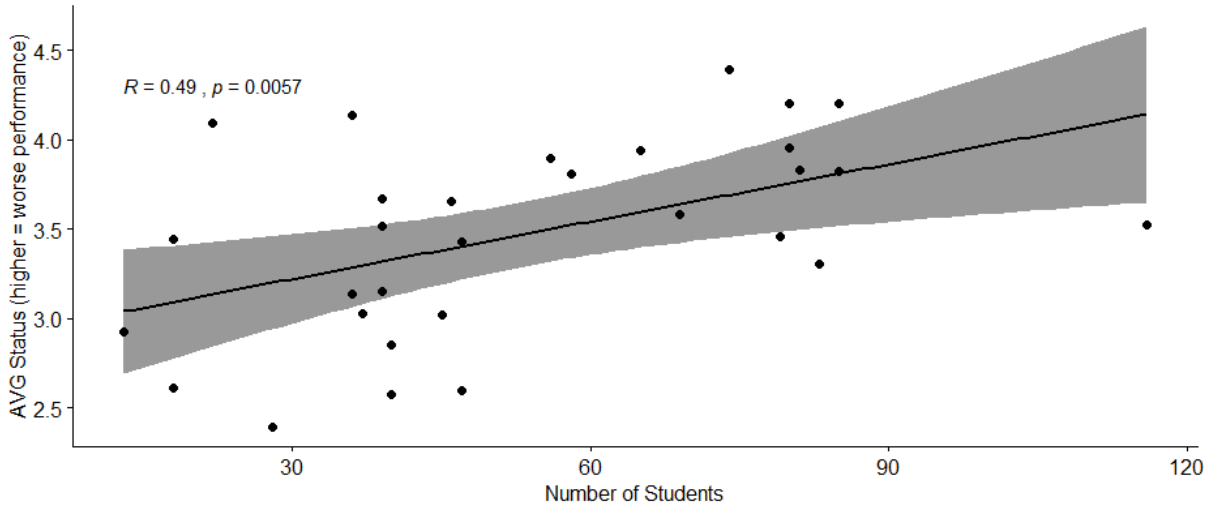


The number of students per section seems to be positive correlated to the average progress of section. To better examine this correlation, I will draft a regression: I will propose the null hypothesis (H0) that the number of students has no effect on the average status. Alternatively, the H1 hypothesis assumes that independent variable "Number of students" has an effect on the average status of the section.

```
> ggscatter(schoolsummary, x = "NumStudents", y = "status1",
+           add = "reg.line", conf.int = TRUE,
+           cor.coef = TRUE, cor.method = "pearson",
+           xlab = "Number of Students", ylab = "Status (higher = worse performance)")
```

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019



The scatter plot unveils a positive correlation of the number of Students and the average status per section. A high number of students per section is associated with a worse average performance of that section.

The p-value indicates that the number of students is statistically significant. The P value provides information about the probability of finding the observed or more extreme results when the null hypothesis (H0) is true. As p = 0.0057, I will reject the null hypothesis and accept the alternative hypothesis that there is a relationship

However, the strong variation indicates that only the number of students per section does not sufficiently explain the average of sections' students status. This becomes clear when looking at the average status of school E. The school has only 1 section with over 100 students but the average student status is still better than the sections of other schools with a lower number of student.

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019

Appendix

Appendix 1: R Script

```
# import data set
data.storyteller <- read.csv("C:/Users/steff/Google Drive/3 Masters/3. Semester/2 Data Analytics/Homework/HW2/data-storyteller.csv",
header=TRUE)

# define new name
schooldata <- data.storyteller

#use of dplyr package
library(dplyr)
library(ggplot2)
library(tidyr)
library(ggpubr)
library(reshape2)

# analyze the structure of the dataset
str(schooldata)
head(schooldata)
summary(schooldata)
#no missing values

#Change header names
names <- c("School", "Section", "Veryahead", "Middling", "Behind", "MoreBehind", "VeryBehind", "Completed")
colnames(schooldata) <- names

# Restructure the Dataset
schooldata_new <- gather(schooldata,status,number,3:8)

#lets look at the new structure
str(schooldata_new)
schooldata_new

# change the data type of status from character to factor by introducing levels and converting them to numbers
levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind', 'VeryBehind')
schooldata_new$status <- factor(schooldata_new$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind',
'VeryBehind'))

schooldata_new2 <- schooldata_new %>% mutate_at(vars(status), as.numeric)
str(schooldata_new2)

schooldata_new3 <- schooldata_new2 %>% mutate(status2 = status * number)

#How does each school performe?
schoolsummary2 <- schooldata_new3 %>% group_by(School) %>% summarize(status1 = sum(status2)/sum(number))
ggplot(schoolsummary2, aes(x = School, y = status1)) + geom_bar(stat="identity", fill="steelblue") + ylab("Average Performance, Lower
= Better") + ggtitle("Performance of Students per School")


#More details: How does each school per section performe?
scholsummary <- schooldata_new3 %>% group_by(School, Section) %>% summarize(NumStudents = sum(number), status1 =
sum(status2)/sum(number))
schoolsummary


# how many sections does each school have and how many students in total?
```

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019

```
summary2 <- schooldata_new %>% group_by(School) %>% summarize(numSections = max(Section), numStudents = sum(number))
head(summary2)


#size of sections
#individually
summary3 <- schooldata_new %>% group_by(School, Section) %>% summarize(numStudents = sum(number))
aa <- filter(summary3, School == "A")
aaa <- ggplot(aa, aes(x=Section, y=numStudents)) + geom_bar(stat="identity", fill="steelblue") + ylim(0, 90)+ scale_x_discrete(name
="Section", limits=c(0:15)) + ggtitle("School A")

bb <- filter(summary3, School == "B")
bbb <- ggplot(bb, aes(x=Section, y=numStudents)) + geom_bar(stat="identity", fill="steelblue") + ylim(0, 90) + scale_x_discrete(name
="Section", limits=c(0:15)) + ggtitle("School B")

cc <- filter(summary3, School == "C")
ccc <- ggplot(cc, aes(x=Section, y=numStudents)) + geom_bar(stat="identity", fill="steelblue") + ylim(0, 90) + scale_x_discrete(name
="Section", limits=c(0:15)) + ggtitle("School C")

dd <- filter(summary3, School == "D")
ddd <- ggplot(dd, aes(x=Section, y=numStudents)) + geom_bar(stat="identity", fill="steelblue") + ylim(0, 90) + scale_x_discrete(name
="Section", limits=c(0:15)) + ggtitle("School D")

ee <- filter(summary3, School == "E")
eee <- ggplot(ee, aes(x=Section, y=numStudents)) + geom_bar(stat="identity", fill="steelblue") + scale_x_discrete(name ="Section",
limits=c(0:15)) + ggtitle("School E")

#aggregation
abcde1 <- ggplot(summary3, aes(x=Section, y=numStudents, fill=factor(School)))+
  geom_bar(position="dodge", stat="identity") + ylab("Number of Students")
abcde1

summary2 <- schooldata_new %>% group_by(School, status) %>% summarize(numStudents = sum(number))
head(summary2)

schoola <- filter(schooldata_new, School == "A")
schoola$status <- factor(schoola$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind', 'VeryBehind'))
schoola1 <- schoola %>% group_by(School, status) %>% summarize(numStudents = sum(number)/932)
a<-ggplot(data=schoola1, aes(x=status, y=numStudents)) +
  geom_bar(stat="identity", fill="steelblue") + ylab("Percentage of Students") + ggtitle("School A") + ylim(0, 0.5)

schoolb <- filter(schooldata_new, School == "B")
schoolb$status <- factor(schoolb$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind', 'VeryBehind'))
schoolb1 <- schoolb %>% group_by(School, status) %>% summarize(numStudents = sum(number)/446)
b<-ggplot(data=schoolb1, aes(x=status, y=numStudents)) +
  geom_bar(stat="identity", fill="steelblue") + ylab("Percentage of Students") + ggtitle("School B") + ylim(0, 0.5)

schoolc <- filter(schooldata_new, School == "C")
schoolc$status <- factor(schoolc$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind', 'VeryBehind'))
schoolc1 <- schoolc %>% group_by(School, status) %>% summarize(numStudents = sum(number)/85)
c<-ggplot(data=schoolc1, aes(x=status, y=numStudents)) +
  geom_bar(stat="identity", fill="steelblue") + ylab("Percentage of Students") + ggtitle("School C") + ylim(0, 0.5)

schoold <- filter(schooldata_new, School == "D")
schoold$status <- factor(schoold$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind', 'VeryBehind'))
schoold1 <- schoold %>% group_by(School, status) %>% summarize(numStudents = sum(number)/22)
d<-ggplot(data=schoold1, aes(x=status, y=numStudents)) +
  geom_bar(stat="identity", fill="steelblue") + ylab("Percentage of Students") + ggtitle("School D") + ylim(0, 0.5)

schoole <- filter(schooldata_new, School == "E")
schoole$status <- factor(schoole$status, levels = c("Completed", "Veryahead", 'Middling', 'Behind', 'MoreBehind', 'VeryBehind'))
```

IST 407/707 Data Analytics & Data Mining
Steffen Tim Knoedler
sknodler@syr.edu

09/09/2019

```
schoole1 <- schoole %>% group_by(School, status) %>% summarize(numStudents = sum(number)/116)
e<-ggplot(data=schoole1, aes(x=status, y=numStudents)) +
  geom_bar(stat="identity", fill="steelblue") + ylab("Percentage of Students") + ggtitle("School E") + ylim(0, 0.5)


#aggregation
jointdataset <- rbind(schoola1, schoolb1, schoolc1, schoold1, schoole1)
jointdataset

#Percentage of Students per status per School
abcde <- ggplot(jointdataset, aes(x=status, y=numStudents, fill=factor(School)))+
  geom_bar(position="dodge", stat="identity") + ylab("Percentage of Students")
abcde


#how many are behind?
schoola_Behind <- sum(schoola1$numStudents[4:6])
schoolb_Behind <- sum(schoolb1$numStudents[4:6])
schoolc_Behind <- sum(schoolc1$numStudents[4:6])
schoold_Behind <- sum(schoold1$numStudents[4:6])
schoole_Behind <- sum(schoole1$numStudents[4:6])

schoola_Behind
schoolb_Behind
schoolc_Behind
schoold_Behind
schoole_Behind


#AVG stuent Performance vs section size
ggplot(schoolsummary, aes(x = NumStudents, y = status1, color = School, group = School)) + ylab("Average Student Performance") +
ggtitle("AVG Student Performance vs. Section Size") + geom_point(lwd = 4)


#Regression
ggscatter(schoolsummary, x = "NumStudents", y = "status1",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Number of Students", ylab = "Status (higher = worse performance)")
```