IST 565 Data Mining

Olivieri

HW 2

1.1

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.

title: "Grades of Students by School"

output: html\_notebook

## Reading the data in

grades <- read.csv('data.csv')

cols <- c("school",

"section", "very\_ahead", "middling", "behind",

"more\_behind", "very\_behind", "completed")

colnames(grades) <- cols

attach(grades)

## Looking at how many sections are completed by school

sections\_by\_school <- aggregate(

section,

by = list(school),

FUN = max

)

sections\_by\_school

| **Group.1**  <fctr> | **x**  <int> |  |  |  |
| --- | --- | --- | --- | --- |
| A | 13 |  |  |  |
| B | 12 |  |  |  |
| C | 3 |  |  |  |
| D | 1 |  |  |  |
| E | 1 |  |  |  |

###

If the semester is nearly 3/4 of the way over then school C, D, and E must be on different schedules because they've only completed 1 section while A and B have completed at least 10 more.

###

current\_course <- max(grades[which(school == "B"), ]$section) / 3 \* 4

30 - current\_course

[1] 14

###

If one of the schools which has the most sections completed, school B, only has 12 sections done. And the semester is already 3/4 of the way through. They will have to do 14 more sections (2 more than they've completed to date) in 1/3 of the time -- 1/4 of the total time of the semester.

This seems odd. Either the final sections of this math course are increasingly easier than the previous 3/4, or they will have to rush the last 1/4 of the semester, or they may not finish. Unless the first 3/4 of the semester are filled with days off for holidays and the previous pace was comparatively slow. Which still would not explain why C, D, and E only have one to three sections done.

(School B was chosen because 12 cleanly makes 3/4 of 16.)

###

max(very\_ahead)

[1] 0

### No kids in any school are 'very ahead'.

# Some quick histograms

getHist <- function(vectorX, bin, xlabel = "x", ylabel = "y") {

qplot(vectorX,

geom="histogram",

binwidth = bin,

fill=I("steelblue"),

col=I("darkorange"),

xlab = xlabel,

ylab = ylabel

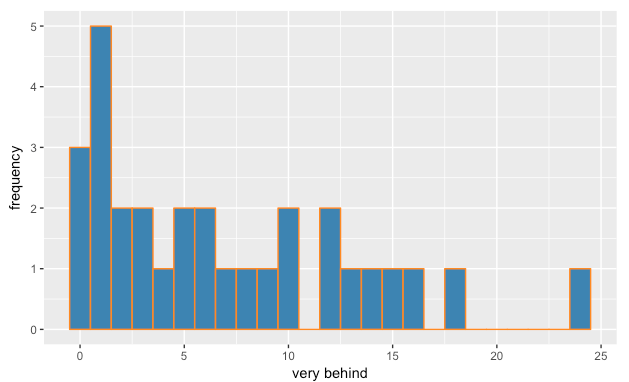
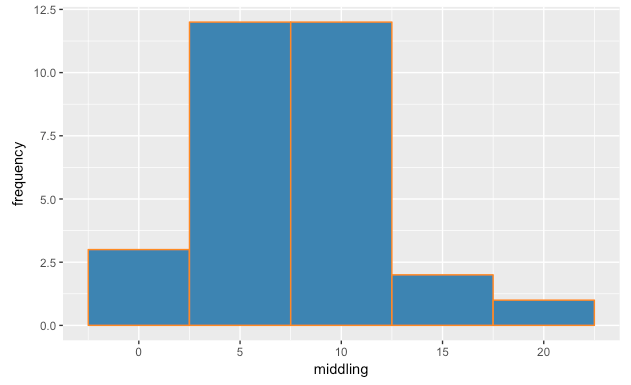
)

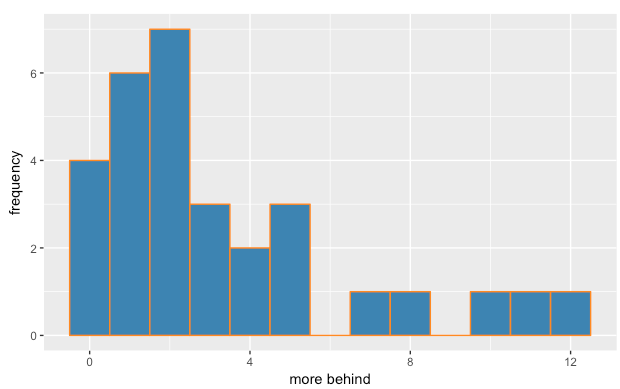
}

getHist(middling, 5, "middling", "frequency")

getHist(very\_behind, 1, "very behind", "frequency")

getHist(more\_behind, 1, "more behind", "frequency")



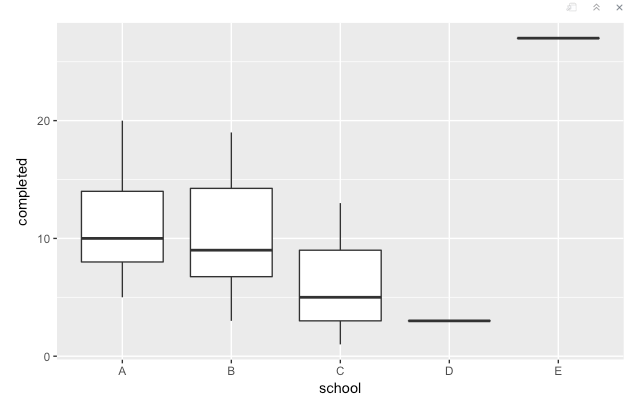


###

Not much can be derived from these, there aren't strong patterns here. One stand-out is that one section of one school reported over 20 kids (almost 25) were 'very behind.'

###

ggplot(grades, aes(x = school, y = completed)) + geom\_boxplot()

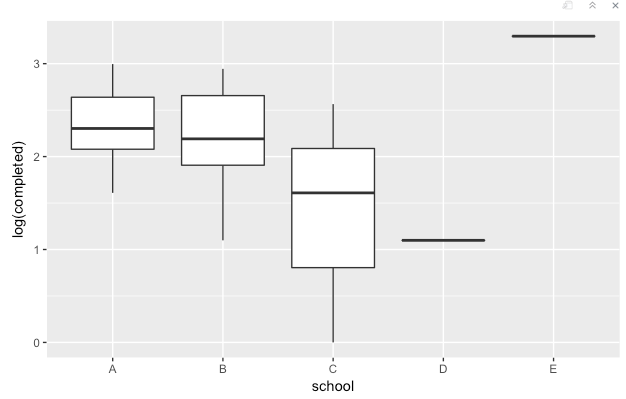


##

This is a strange result. If we remember from earlier, School E only has one section finished and they have over 20 kids who have completed the entire course.

##

ggplot(grades, aes(x = school, y = log(completed))) + geom\_boxplot()



##

Changing the ‘completed’ value to log evens out the schools that have finished fewer sections but shows school E is still an outlier.

##

behind\_plot <- ggplot(grades, aes(x = section, y = value, color = variable)) +

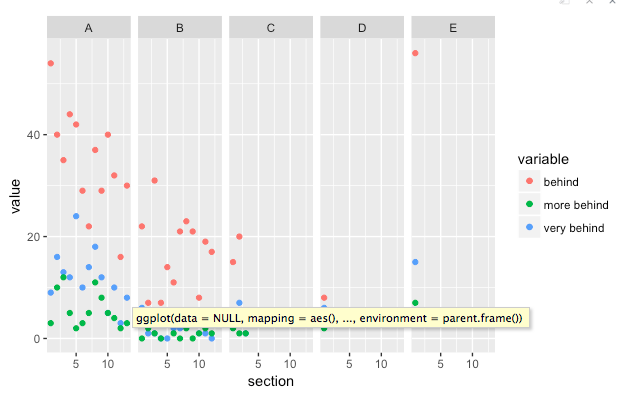
geom\_point(aes(y = very\_behind, col = "very behind")) +

geom\_point(aes(y = more\_behind, col = "more behind")) +

geom\_point(aes(y = behind, col = "behind")) +

facet\_grid(. ~ school)

behind\_plot



##

This shows a gradual downtrend among students who are behind each week as the sections progress.

##

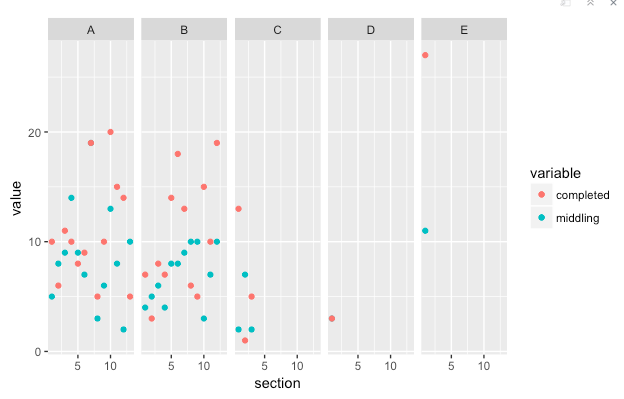
doing\_well\_plot <- ggplot(grades, aes(x = section, y = value, color = variable)) +

geom\_point(aes(y = middling, col = "middling")) +

geom\_point(aes(y = completed, col = "completed")) +

facet\_grid(. ~ school)

doing\_well\_plot



##

There is some random noise in this plot, but a noticeable uptrend among middling students at school B.

##