Jacob Dineen

IST 565 Data Mining

Homework 2

10/17/17

The first thing that we need to do with this dataset is understand the scope of the data that we are looking at. As the prompt notes, we are looking at students’ status as it relates to the math course offerings of certain schools. Before beginning analysis, we need to first clean the data. To do so we’ll perform the following actions:

##################################Cleaning Data

#Rename Columns for digestion

colnames(DataStoryTeller) <- c("School", "Section", "VeryAhead", "Middling", "Behind", "MoreBehind", "VeryBehind", "Completed")

#Check Above

colnames(DataStoryTeller)

#Section is Nominal, Change

DataStoryTeller$Section <- as.factor(Section)

#All other columns to Integers

for(i in 3:7)

{

DataStoryTeller[,i] <- as.integer(DataStoryTeller[,i])

}

str(DataStoryTeller)

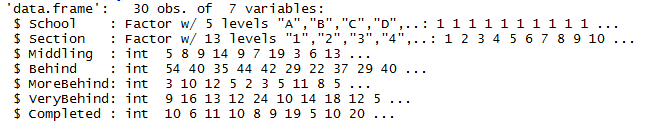
From here, the first thing of notice is that the ‘Very Ahead’ variable contains 0 students. This is significant in our analysis, but will be disruptive in the process of determining central tendency and variance. We will remove this column from our analysis:

#No students are 'Very Ahead'. Remove Column to avoid distortion/noise

DataStoryTeller <- DataStoryTeller [,-3]

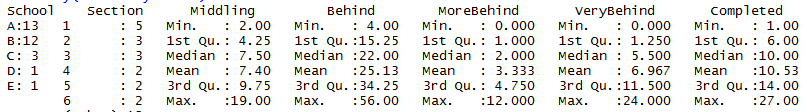
Aggregating is not necessary in this instance, as the source data is already aggregated by school and by section. I will Aggregate amongst specific nominal variables from this dataframe, but I will not disrupt this ‘Master’ dataframe.

Our ‘ready to analyze’ data frame contains 30 observations of 7 variables. 2 of those variables are nominal/categorical, and the remaining 5 variables are integers, represented by the number of students corresponding to a particular status.



Summary statistics are very useful when initializing work with a specific set of data. As seen below:

Summary(DataStoryTeller)



What this shows is a high level view of descriptive statistics. We can see for the two nominal variables, most of the observations tie into schools A & B, and Section 1 is the most represented out of the 13, respectively. In regards to this statement, I think one of the obvious pieces of discernable information from the data is that the prompt states that the semester is about ¾ of the way complete, and yet, the maximum section listed is 13, which, assuming even pace, should have been completed around the midway point of the

When looking at the summary statistics for the numeric variables, less is decipherable due to their current groupings. If this data frame showed per student data, rather than grouping of students represented by counts, we would be able to better understand central tendency and dispersion.

What would first be useful is understanding how many students we are looking at in our data:

Students <- data.frame(colSums(DataStoryTeller[,3:7]))

Students <- transpose(Students)

str(Students)

colnames(Students) <- c("Middling", "Behind", "MoreBehind", "VeryBehind", "Completed")

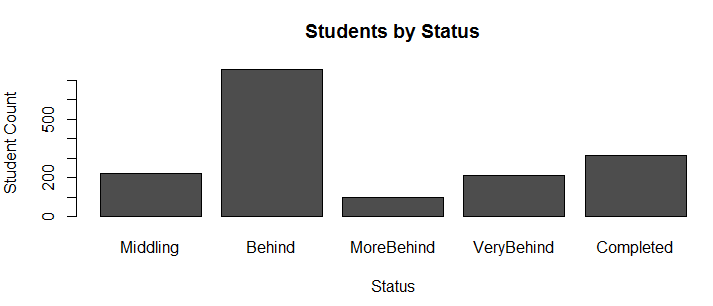
str(Students)

sum(Students)

So our analysis is inclusive of 1601 total students. Starting with Status, we can look at the variance between the statuses:

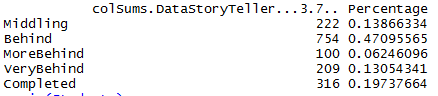
as.matrix(Students)

barplot(as.matrix(Students), main="Students by Status", xlab="Status", ylab="Student Count")



Conversely, if we wanted to show this by a percentage, rather than a count:

Students$Percentage <- Students$colSums.DataStoryTeller...3.7../ sum(Students$colSums.DataStoryTeller...3.7..)



What we can see from the above two visuals is that almost half of the students, 47%, are behind, and 19% more are either more behind, or very behind. That means that over 66% of the total students in the math course are 1-10+ lessons behind the curriculum. 20% have completed the represented section.

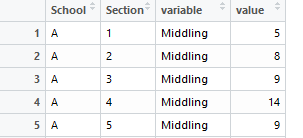
From here, it gets a little bit more adhoc- What I choose to do was melt the data into variable and value columns for easier visual representation.

#Melting Data

Melt <- melt(DataStoryTeller)

Melt

Example of melted DF:

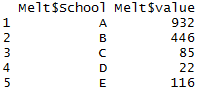


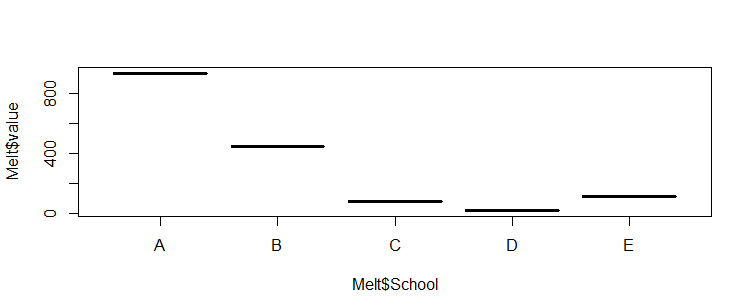
From here we can look at distribution by school:

#Showing Number of Students Per School Taking Math course

aggregate(Melt$value ~ Melt$School, data = Melt, FUN = sum)

plot(aggregate(Melt$value ~ Melt$School, data = Melt, FUN = sum))

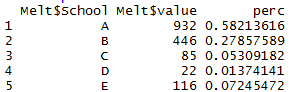


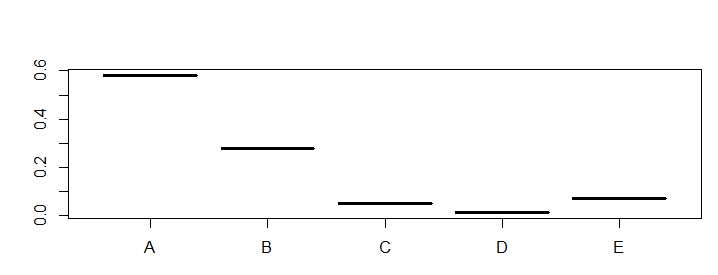
So it easy to see that most of the students in this data set represent school A and school B. If we wanted to look at a percentage split:

meltsps$perc <- meltsps$`Melt$value`/sum(meltsps$`Melt$value`)

meltsps

plot(meltsps$`Melt$School`, meltsps$perc)





So 86% of the students are from schools A and B. I think this leads us to a question of distribution amongst the schools – Are there more students at schools A and B, or are students required to take the course at certain schools, and not at others?

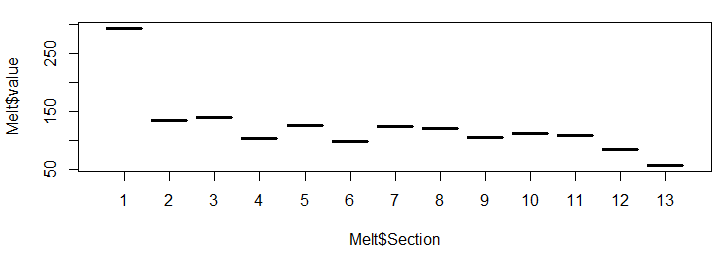
If we want to look at distribution by section, we can see:

#Showing Number of Students Per Section

aggregate(Melt$value ~ Melt$Section, data = Melt, FUN = sum)

plot(aggregate(Melt$value ~ Melt$Section, data = Melt, FUN = sum))

#Section 1 is much more represented than the other sections

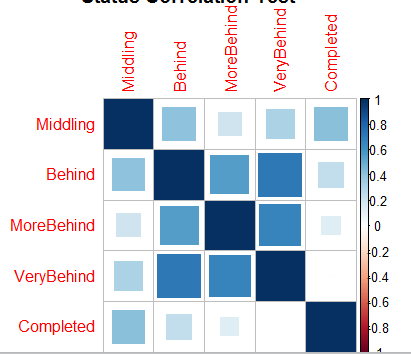


To touch back to the above chart, this distribution shows us the number of students who have reached a certain section, regardless of status. This right here shows us a linear (maybe exponential initially) downtrend in the student count in relation to the section. As the course went on, fewer and fewer students reached the next section- I’ll touch upon this in my summary.

Summary

So it is very obvious that both the students and the schools themselves are behind in the curriculum, which I believe points to a flaw in the process itself. We either have a case where the students are deterring the course from moving at its intended pace, or we have an intended pace that is disruptive to the students’ ability to complete the sections.

It appears that Schools C, D and E scrapped the course early on in the semester after realizing that students were falling behind at a disproportionate rate. A correlation test on the statuses show us that there is a very strong relationship between student counts of for behind, more behind, and very behind- This leads me to believe that students who fell behind a little bit were never able to catch up. Their data can probably be removed from this dataset, but there weren’t enough observations to drastically impact the results anyways.



Because there aren’t any extreme values or outliers present, and we aren’t really modeling anything here, I don’t believe that transformation techniques are needed for our numeric variables.

#Via Log

Melt$log <- log(Melt$value)

plot(Melt$value, Melt$log)

#Via Zscore

Melt$zscore <- scale(Melt$value, center = TRUE, scale = TRUE)

Melt$zscore

#Via MinMax

Melt$minmax <- (Melt$value - min(Melt$value, na.rm = TRUE))/(max(Melt$value, na.rm = TRUE)-min(Melt$value, na.rm = TRUE))

Minmax

plot(Melt$value, Melt$log)

plot(Melt$value, Melt$minmax)

plot(Melt$value, Melt$zscore)

All in all, I think it is discernable that the course should probably be restructured as soon as possible for whomever is still using it. The data tells us that it was too ambitious and students could just not keep up with the course plan.

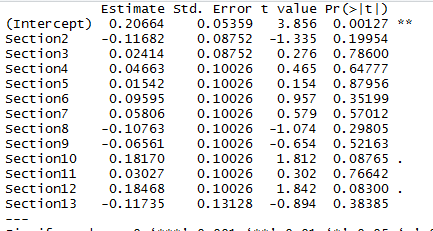
I don’t really see a way of turning this into a regression problem, except for one. I thought maybe using Section as a dependent variable and a ratio of completion percentage would help to predict future completion rates per Section:

DataStoryTeller$TotalStudents <- rowSums(DataStoryTeller[,3:7])

DataStoryTeller$CompPerc <- DataStoryTeller$Completed/DataStoryTeller$TotalStudents

lm(formula = CompPerc~ Section, data = DataStoryTeller)

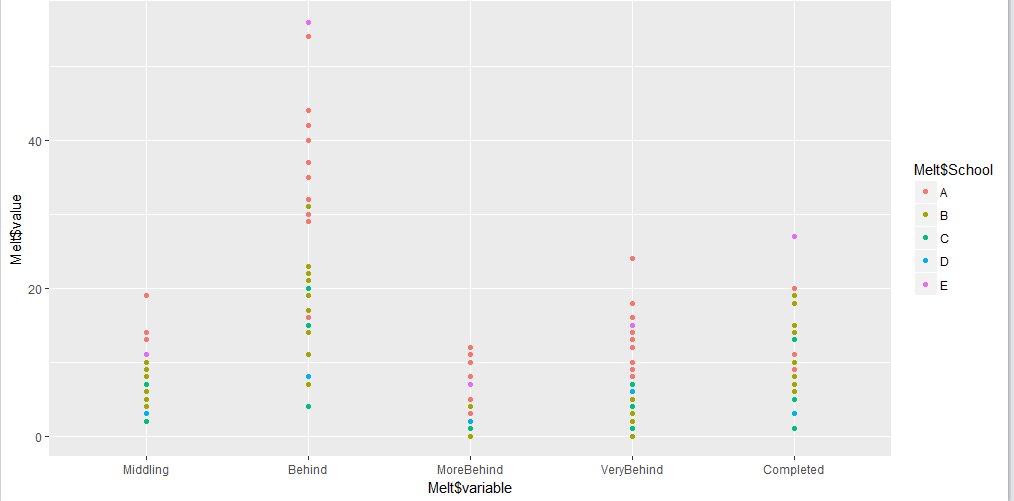
summary(lm(formula = CompPerc~ Section, data = DataStoryTeller))



As expected, the skew of distribution towards Section1 is impacting a linear model, and transforming the data per the above method heeds no advancement.

Lastly, I found this graph to be interesting:

ggplot(Melt, aes(x=Melt$variable, y=Melt$value)) + geom\_point(aes(color=Melt$School))



The Y axis is on an interval of students per section, X axis representing status, and the color representing school. We can see that most of the sections had class sized below 20. We can also that A and B are the most represented schools. This is kind of like a consolidated view of everything that was said above.

Note – I found out about RMD after completing this assignment.