Exploring Educational Datasets

INFO 4100 Learning Analytics Homework

[[Shuhui Zhu, sz649]]

In this homework, you will conduct an exploratory analysis with a public datasets obtained from [PSLC DataShop](http://pslcdatashop.org). The dataset provides question-level data of students practicing math problems in academic year 2004-05 using the [Assisstments platform](https://www.assistments.org/). On this platform, students can attempt a problem many times to get it right and they can ask for several hints (one at a time) on a problem until the final hint tells them what the answer is.

Learning Objectives:

1. Identify a dataset file format and use the appropriate function to load it.
2. Explore fundamental properties of a dataset using basic functions in R
3. Compute and visualize relationships between variables using correlations, histograms, boxplots, and scatterplots
4. Calculate and visualize sutdent- and question-level quantities and relationships

# Part 1: Loading the Dataset

Before you can load data, you need to figure out the format that it is saved in. The file extension typically corresponds to the format, but this is not always the case. R has functions to load all common data files, most of these functions start with read, e.g. read.csv() for CSVs or read.tsv() for tab-separated values. The **foreign** package adds functions to import many additional data file types. For large data files, consider using the fread function in the **data.table** package: it’s fast and reliable.

The readRDS() and saveRDS() functions allow you to important and export any object in R. This can be a scalar, vector, matrix, data.frame, function, or any other object. Moreover, saving a dataset as an RDS file is much more efficient (smaller file size) than saving it as a CSV. Use the help panel to see examples of how to use any of these functions.

**Question 1:** Load the Assistments dataset (*info4100.data.assisstments.rds*) into R and call it asm.

#######################################  
####### BEGIN INPUT: Question 1 #######  
#######################################  
asm<-readRDS('../Downloads/info4100.data.assisstments.rds')  
#######################################  
#######################################

# Part 2: Exploring the Dataset

It is hard to overstate the importance of understanding the data you are working with. Ideally, you get to understand the data-generating process, how the data was collected. To start off, you should understand what is in the dataset.

**Question 2:** To look at the first few rows of asm, use the head() function.

#######################################  
####### BEGIN INPUT: Question 2 #######  
#######################################  
head(asm)

## studentID itemid correctonfirstattempt attempts hints seconds  
## 1 136 90 1 1 0 58  
## 2 136 91 0 1 3 91  
## 3 136 92 1 1 0 11  
## 4 136 93 1 1 0 10  
## 5 136 94 0 2 0 43  
## 6 136 95 1 1 0 13  
## full\_start\_time full\_finish\_time start\_day  
## 1 01-OCT-04 07.44.43.000000 AM 01-OCT-04 07.45.41.000000 AM 01-OCT-04  
## 2 10-DEC-04 09.27.20.000000 AM 10-DEC-04 09.28.51.000000 AM 10-DEC-04  
## 3 10-DEC-04 09.28.51.000000 AM 10-DEC-04 09.29.02.000000 AM 10-DEC-04  
## 4 10-DEC-04 09.29.02.000000 AM 10-DEC-04 09.29.12.000000 AM 10-DEC-04  
## 5 10-DEC-04 09.29.12.000000 AM 10-DEC-04 09.29.55.000000 AM 10-DEC-04  
## 6 15-OCT-04 07.44.14.000000 AM 15-OCT-04 07.44.27.000000 AM 15-OCT-04  
## start\_time finish\_day finish\_time  
## 1 7.44.43 01-OCT-04 7.45.41  
## 2 9.27.20 10-DEC-04 9.28.51  
## 3 9.28.51 10-DEC-04 9.29.2  
## 4 9.29.2 10-DEC-04 9.29.12  
## 5 9.29.12 10-DEC-04 9.29.55  
## 6 7.44.14 15-OCT-04 7.44.27

#######################################  
#######################################

Based on the first few lines of data, and what we know about the dataset, we can infer the following:

* *studentID* is an identifier for students
* *itemid* is an identifier for math questions
* *correctonfirstattempt* is an indicator of whether a student answered correctly on the first attempt
* *attempts* is the number of answer attempts required
* *hints* the number of hints a student requested
* *seconds* time spent on the question in seconds
* the remaining columns provide start and end times and dates for each question

It also shows us that the dataset is in **long format** (1 row = 1 event) instead of wide format (1 row = 1 individual). However, as you can see from the *attempts* variable, you do not have data on each attempt, but a question-level rollup. The data is at the student-question level, which means that there is one row for each question a student attempted that summarizes interaction with the question (performance indicators and time spent).

Now you will answer a series of questions about this dataset. It is the kind of questions you would provide answers to if you were to write a report about the dataset. Be sure to show your R code and the final answer to the question inside of the input area.

**Question 3:** How many unique students are in the dataset? Tip: Use the unique() function and the length() function.

#######################################  
####### BEGIN INPUT: Question 3 #######  
#######################################  
length(unique(asm$studentID))

## [1] 912

# 912   
#######################################  
#######################################

**Question 4:** How many unique questions are there?

#######################################  
####### BEGIN INPUT: Question 4 #######  
#######################################  
length(unique(asm$itemid))

## [1] 1709

#######################################  
#######################################

**Question 5:** What is the rate of getting it right on the first attempt? Another way to ask this: What proportion of math problems did students get right on the first attempt?

#######################################  
####### BEGIN INPUT: Question 5 #######  
#######################################  
sum(asm$correctonfirstattempt)/length(asm$correctonfirstattempt)

## [1] 0.4047563

#######################################  
#######################################

**Question 6:** What is the rate of asking for any hints? Another way to ask this: For what proportion of math problems did students ask for at least one hint?

#######################################  
####### BEGIN INPUT: Question 6 #######  
#######################################  
sum(asm$hints>0)/length(asm$hints)

## [1] 0.326534

#######################################  
#######################################

**Question 7:** How long do students spend on a question on average? Tip: Use the mean() function.

#######################################  
####### BEGIN INPUT: Question 7 #######  
#######################################  
  
mean(asm$seconds)

## [1] 48.65293

#######################################  
#######################################

**Question 8:** What is the frequency distribution of hints? That is, on how many problems did a student need no hint (0), one hint, two hints, etc.? Tip: Use the table() function.

#######################################  
####### BEGIN INPUT: Question 8 #######  
#######################################  
table(asm$hints)

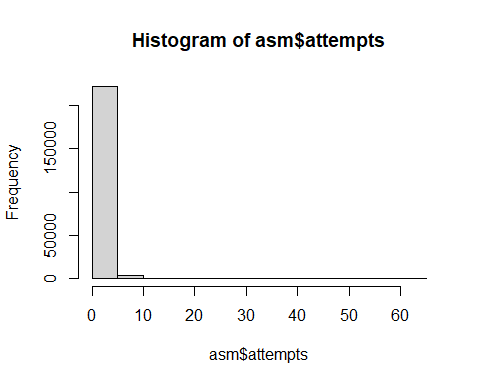
##   
## 0 1 2 3 4 5 6 7 8 9 10   
## 152270 25841 14139 23668 6916 2088 953 74 57 36 17   
## 11 12 13 14 15 16 17 18 19 20 21   
## 5 16 3 2 3 2 1 2 1 1 1   
## 24 28 31   
## 1 1 1

#######################################  
#######################################

# Part 3: Visualizing Distributions of Variables

**Question 9:** How many attempts do students make on questions? To answer, plot the distribution of attempts using a histogram. Tip: Use the hist() function.

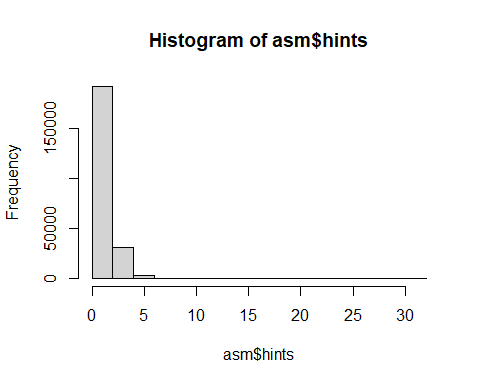
#######################################  
####### BEGIN INPUT: Question 9 #######  
#######################################  
  
hist(asm$attempts)



#######################################  
#######################################

**Question 10:** How many hints do they ask for? To answer, plot the distribution of hints.

#######################################  
####### BEGIN INPUT: Question 10 ######  
#######################################  
  
hist(asm$hints)



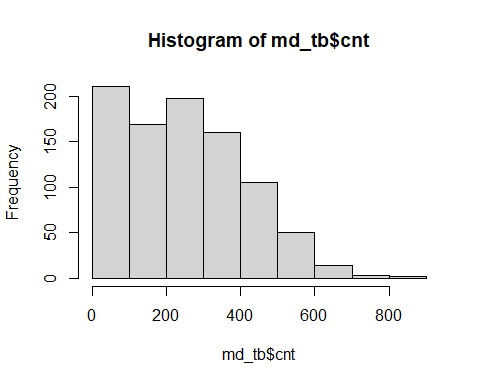
#######################################  
#######################################

**Question 11:** How engaged are students on the platform? To answer, plot the distribution of how many questions each student attempted. Tip: This requires an intermediate step to calculate the frequency distribution of question answering before plotting a histogram.

#######################################  
####### BEGIN INPUT: Question 11 ######  
#######################################  
  
# t<-asm$studentID  
# t  
# hist(table(t))  
   
library(dplyr)  
  
md\_tb <- asm %>%  
 group\_by(studentID) %>%  
 summarise(cnt = n())  
md\_tb

## # A tibble: 912 x 2  
## studentID cnt  
## <int> <int>  
## 1 136 518  
## 2 137 687  
## 3 139 538  
## 4 140 522  
## 5 141 113  
## 6 142 5  
## 7 143 560  
## 8 144 281  
## 9 145 593  
## 10 146 465  
## # ... with 902 more rows

hist(md\_tb$cnt)



#######################################  
#######################################

# Part 4: Exploring Relationships Between Variables

**Question 12:** How is time spent, number of attempts, and getting hints related to one another? To answer, compute the three pair-wise correlations between seconds, attempts, and hints? Tip: Use the cor() function.

#######################################  
####### BEGIN INPUT: Question 12 ######  
#######################################  
cor(asm$hints, asm$seconds)

## [1] 0.1709763

cor(asm$seconds, asm$attempts)

## [1] 0.4345258

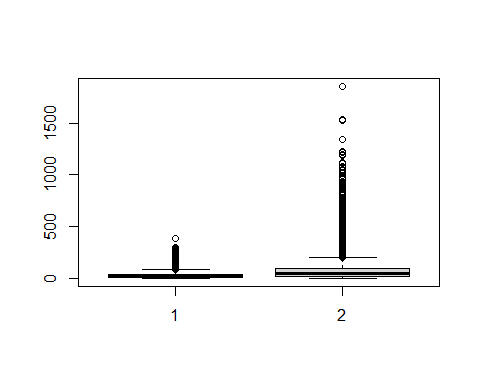
cor(asm$attempts, asm$hints)

## [1] 0.1275844

#######################################  
#######################################

**Question 13:** Do students spend more or less time answering questions that they get right on the first attempt versus those that take multiple attempts? To answer, compare the distributions of time spent for questions that students got right on the first attempts versus those that took them multiple attempts using a boxplot. Tip: Use the boxplot() function; look up the correct syntax in the help panel.

#######################################  
####### BEGIN INPUT: Question 13 ######  
#######################################  
fstAtmp <- subset(asm, correctonfirstattempt == 1)  
mltAtmp <- subset(asm, attempts > 1)  
  
boxplot(fstAtmp$seconds, mltAtmp$seconds)



#######################################  
#######################################

**Question 14:** How is asking for hints related to getting the right answer on the first attempt? To answer, create a cross-tabulation of the frequency distribution of hints against getting it right on the first attempt. Tip: Use the table() function with two parameters Note in the output that 6+2+2 students asked for hints before making an attempt and then got it right on their first attempt.

#######################################  
####### BEGIN INPUT: Question 14 ######  
#######################################  
table(asm$hints, asm$correctonfirstattempt)

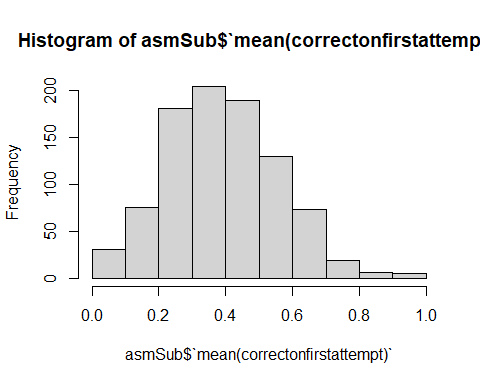
##   
## 0 1  
## 0 60765 91505  
## 1 25835 6  
## 2 14137 2  
## 3 23666 2  
## 4 6916 0  
## 5 2088 0  
## 6 953 0  
## 7 74 0  
## 8 57 0  
## 9 36 0  
## 10 17 0  
## 11 5 0  
## 12 16 0  
## 13 3 0  
## 14 2 0  
## 15 3 0  
## 16 2 0  
## 17 1 0  
## 18 2 0  
## 19 1 0  
## 20 1 0  
## 21 1 0  
## 24 1 0  
## 28 1 0  
## 31 1 0

#######################################  
#######################################

# Part 5: Students and Questions as the Unit of Analysis

**Question 15:** How are students scoring on average? To answer, plot the student-level distribution (i.e. 1 value per student) of answering correctly on the first attempt. The plot should be a histogram with the proportion answering correctly on the x-axis and the corresponding number of students on the y-axis. Tip: You first need to compute the proportion of questions that each student got right on first attempt; there are several ways to do this. You can load the tidyverse package and use group\_by and summarise; or you can use sapply(); or load the data.table package; whatever you do, don’t use a *for* loop unless you really cannot solve it otherwise.

#######################################  
####### BEGIN INPUT: Question 15 ######  
#######################################  
  
library(tidyverse)  
asmSub <- asm %>% group\_by(studentID) %>% summarise(mean(correctonfirstattempt))  
hist(asmSub$`mean(correctonfirstattempt)`)

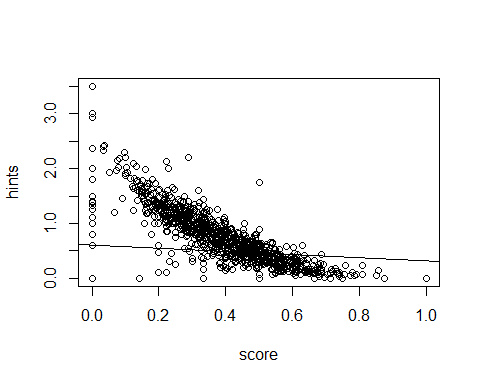


#######################################  
#######################################

**Question 16:** Do students who score better ask for more hints? To answer, plot the student-level relationship (i.e. 1 value per student) between average correctness on first attempt (on the x-axis) and the average number of hints (on the y-axis) using a scatter plot. Tip: Use the plot() function. Bonus: Try adding a straight line to fit the data using the lm() and abline() functions.

Answer: No, students who score better do not ask for more hints.

#######################################  
####### BEGIN INPUT: Question 16 ######  
#######################################  
  
asmScrHts <- asm %>% group\_by(studentID) %>% summarise(score = mean(correctonfirstattempt), hints = mean(hints))  
with(asmScrHts,plot(score, hints))  
abline(lm(asmScrHts$score ~ asmScrHts$hints)) #



#######################################  
#######################################

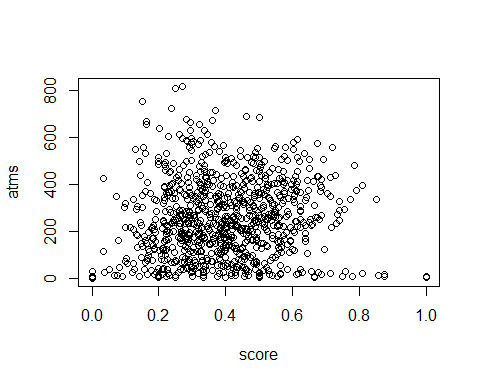
**Question 17:** Are students who attempt more questions (i.e. get more practice) more likely to answer correctly on the first attempt? To answer, report the correlation and make a scatterplot.

Students get more practice does not necessarily perform better on the first attempt.

#######################################  
####### BEGIN INPUT: Question 17 ######  
#######################################  
asmScrAtm <- asm %>% group\_by(studentID) %>% summarise(score = mean(correctonfirstattempt), atms = n())  
sprintf('The correlation is: %f', cor(asmScrAtm$score, asmScrAtm$atms))

## [1] "The correlation is: 0.100778"

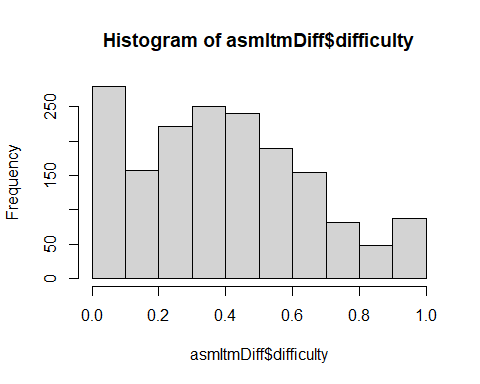
with(asmScrAtm, plot(score, atms))



#######################################  
#######################################

**Question 18:**: How difficult are the questions? To answer, plot the question-level distribution (i.e. 1 row per question) of the proportion of students who get it right on the first attempt using a histogram. This quantity is also called “item difficulty”. Tip: Use the same general approach as for the student-level questions.

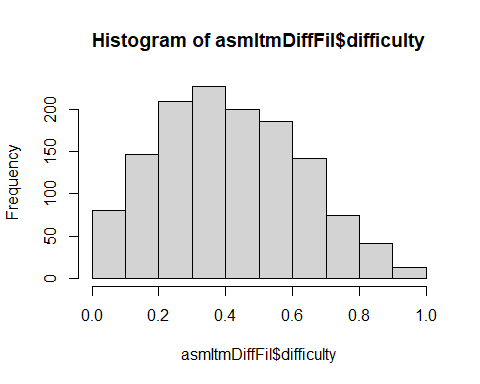
#######################################  
####### BEGIN INPUT: Question 18 ######  
#######################################  
asmItmDiff <- asm %>% group\_by(itemid) %>% summarise(difficulty = mean(correctonfirstattempt))  
hist(asmItmDiff$difficulty)



#######################################  
#######################################

**Question 19:** For just those questions that were answered at least 10 times (i.e. where we have enough data to get a reliable estimate), how difficult are the questions? Plot a histogram like before and note that you should see that the filering reduces the spikes at 0 and 1). Tip: Use the subset() function.

#######################################  
####### BEGIN INPUT: Question 19 ######  
#######################################  
asmItmDiff <- asm %>% group\_by(itemid) %>% summarise(difficulty = mean(correctonfirstattempt), times = n())  
asmItmDiffFil <- subset(asmItmDiff, times>=10)  
hist(asmItmDiffFil$difficulty)



#######################################  
#######################################

# Self-reflection

**Briefly summarize your experience on this homework. What was easy, what was hard, what did you learn?**

* Writing up the code was easy, but to understand the question and relate it back to education purpose takes a second thought. It is hard to think all aspects, for code and application.
* I was a bit confused on *correctonfirstattempt* and *attempt*. It would be more helpful if the distinction between the two could be given.

# Submit Homework

This is the end of the homework. Please **Knit to Word**. The resulting file has to show both the R code and R output. Upload it on the EdX platform before the due date.