

Homework 2: Exploratory Analysis (Part 2)

(3 book questions, 4 data questions, 1 bonus)

Due 1/22

1. Book work

This homework begins with reading Chapter 2 of your textbook. (This will help with your lab due next week also.) As we mentioned in class, you can skim over the pieces referring to mean and standard deviation as we will come back to those in a week or so. For now, focus on histograms and boxplots (and the ingredients that go into making them).

Questions: 2.35, 2.36 and 2.80

In addition to these questions, you are to complete the following lab-like assignment involving an analysis of the Registrar's data.

2. Data analysis

We begin by loading the “student view” data set we discussed in class. As usual, you load it with a call to the `source` command:

```
source("http://www.stat.ucla.edu/~cocteau/stat13/data/student.R")  
  
ls()
```

You should see a data set (data frame) called `student` in your workspace. The data are actually a little different from what appears in lecture. Have a look at the first 10 observations.

```
student[1:10,]
```

Each row is a different student. You should see variables named `id` (a recoding of each student's ID), `lev` (their degree program, graduate, undergraduate or professional), `num_classes` (the number of classes they took last quarter), `num_upperdiv`

(the number of those classes that were upper division), `min_start` (the hour of their earliest class), `time_spent` (the time spent in class, in minutes), and `days_on_campus` (the number of days per week they are required to be on campus).

```
names(student)
dim(student)
```

Question 1: How many students are in the data set? Describe the type of each variable measured on each student (quantitative, categorical, and so on).

Have a look at the variables (well, not `id` as that is not data about students, but instead a unique number assigned to each one for the purposes of recordkeeping) using some of the tools we described in class. For example, consider a table of the number of students in each level (graduate, undergraduate and professional); make a barplot of the days students are on campus; and so on. With each variable, think about a sensible graphical or numerical summary; use R to construct that summary; and then think about whether or not the results match your experience. Given that your experience is probably as an undergraduate, you might want to subset the data first and just consider undergrads. For that you could use a command like the following.

```
undergrads = subset(student, student$lev=="U")
```

This will create a new data set called `undergrads`; the following commands show that all the data are still there and how many observations you selected (the number of undergrads enrolled at UCLA last quarter).

```
names(undergrads)
dim(undergrads)
```

Question 2: Write up your results for two variables, commenting on what you see and whether or not it matches your expectations.

The registrar did not provide us with each student's year at UCLA. We can perhaps infer beginning from advanced students based on the proportion of upper division classes they are taking. The next two questions relate to this.

Question 3: Create a new variable called `prop` that represents the proportion of classes undergraduates take that are upper division. Make a histogram of `prop`. What does this say? (Hint: Recall how we formed the variable `bmi` in lab.)

We can now create a variable that is `TRUE` if a student is taking more than 50% upper division courses and `FALSE` otherwise. Here is the command that would do this.

```
adv = prop > 0.5
table(adv)
```

The two commands above make a new variable called `adv` that has the values we want (look at it if you like by typing its name and hitting enter) and then creates a table summarizing the number of students in each category. (What proportion of students are “advanced” according to this definition?) The following will give us a table that breaks down the number of days per week students of each category have to be on campus.

```
table(adv,undergrads$days_on_campus)
```

Question 4: Make a mosaic plot from the table above and describe what you see. What does it mean?

Bonus: In Question 4 we examined the relationship between two variables. Considering the data we have, come up with another relationship that you think would be interesting and “present it” in some way; either through a numerical summary or a graph (boxplots, histograms, a mosaic plot). Use either the data set `student` (perhaps comparing grads, undergrads and professional students in some way) or the data set `undergrads`.