Week 2 Live Session

w203 Instructional Team

Fall 2017

1. **Slack Channel**
   If you are not on the course-wide slack channel yet, please

   1. make sure you have joined the ucbischool.slack.com slack team

   2. enter your slack ID into the short answer poll

2. **Lab 1 Group Formation**
   In Lab 1, you will work as part of team to generate an exploratory analysis. Your instructors have compiled four different datasets for you to choose from. Please select a topic you find interesting, but be prepared to move to another topic to balance the teams.

   (a) Corruption

   (b) Broadband

   (c) Forrest Fires

   (d) Cancer

   In the week 4 live session, we will ask each team to share their report (maximum 15 minute presentation). Please plan for each team member to present a portion of your report.

3. **Fertility and Education EDA:** A selection of your critiques

   (a) Victoria Eastman (Section 5) - The reasons for the initial sub setting are understandable and justifiable but it seems to me that there are so few observations left... There isn't a ton that can be done, because we only have the data that is available, but maybe reconsidering the initial assumptions or including a larger group would add value to the less explainable portions of the data

   (Follow-up Question: What are the risks in reducing the minimum age / increasing the maximum age?)

   (b) Paul Durkin (Section 6) - I am concerned about the initial choice of using a population range of 35-44. There were two assumptions given for this range; first women have typically had all their children by 35 and second that the majority have completed their education at this point. I am mainly concerned about the first point here, that women have usually finished having children by 35. There is a recent CDC report (here: https://www.cdc.gov/nchs/products/databriefs/db260.htm) showing that 33%, 23% and 6% of women with 0, 1 or more children respectively, and in the age range 35-44, expect to have another child...

   (Follow-up question: If it's true that many women in the study have not finished having children or have not finished their education, how might this be reflected in the observed relationship between fertility and education?)

   (c) Jack Workman (Section 5) - One potential weakness is the lack of consideration of women that attended college but did not graduate. In the "Number of Children by Educational Attainment" boxplot on page 11 of the Exploratory Analysis PDF, the categories are "Some Primary School", "High School Graduate", and "College Graduate" with the year bins being 0-11, 11-15, 16-Inf. These bins could be effectively hiding an important trend in those women with education between 12-15 years.

   (Follow-up question: How do you decide when to bin levels of a variable together?)

(d) Lucy Xie (Section 6) - One weakness of this analysis is the decision to include data points where years of educations is less than the minimum years set by compulsory education laws. As the report suggests, the respondents may have misunderstood or miscalculated their years of education; this seems likely, as they could easily assume the question was asking for years of education BEYOND secondary school. It is worth returning to the Fert_all data set and examining the total responses with <5 years of education, then determining whether this is a misleading question by comparing the distribution to the overall population (information should be available with U.S. Department of Education).

(Follow-up Question: What action should you take in R, if you suspect these data points do not really represent total years of education?)

4. **Crime and House Prices:** An in-class EDA exercise

The file Boston_w203.csv contains data on neighborhoods in the Boston area. You are given the following codebook.

crim	- per capita crime rate by town
zn	- proportion of residential land zoned for lots over 25,000 sq.ft.
indus	- proportion of non-retail business acres per town
chas	- Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	- nitrogen oxides concentration (parts per 10 million)
rm	- average number of rooms per dwelling
age	- proportion of owner-occupied units built prior to 1940
dis	- weighted mean of distances to five Boston employment centres
rad	- index of accessibility to radial highways
tax	- full-value property-tax rate per $10,000
ptratio	- pupil-teacher ratio by town
black	- $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
lstat	- lower status of the population (percent)
medv	- median value of owner-occupied homes in $1000

You are to begin an exploratory analysis with the objective of understanding how the price of a home relates to neighborhood characteristics, with an emphasis on crime.

```
Boston = read.csv("Boston_w203.csv")
```

1. Generate a scatterplot matrix for all metric variables. Take a few minutes to draw as many insights as you can about the relationships in the data.
2. Examine the main output variable, medv. Comment on any unusual values you find, and any features that might be important for statistical modeling.
3. Examine the main independent variable of interest, crim. What transformation could you apply to this variable to aid in visualizing it? Comment on any unusual features you find.
4. Examine the bivariate relationship between medv and crime. What type of relationship do these variables have?
5. (As time permits) Continue your exploratory data analysis. Be prepared to share interesting findings with the class.