# Unit 2 Live Session

## W203 Instructional Team

## Exploratory Data Analysis




### Class Announcements

1. Lab 1 Assignment 2. Student Introductions (30 second version) 4. Instructor Intro

### 1 Pre Class Exercise Responses

** 1.1 ** Response from: *Jennifer Mahle*

Once the data has been filtered to include only women ages 35-44, there are only 683 observations, which is a pretty small sample size. I realize we have caveats that say we're only extrapolating information for the sample, but it's still not a large sample. This is especially evident when segmenting out the population by "some primary school", "high school graduate" and "college graduate", where the bin with "some primary school" only has 98 women. Looking at the boxplot of number of children by years of education, it's very likely that each of the years of education under 12 have very few data points.

Follow up question: What potential problems might we face if we increase the data range to include more observations?

Type *Markdown* and LaTeX: $\alpha^2$

** 1.2 ** Response from: *Ramiro Cadavid*

The author state that the range of ages chosen was based on the fact that almost all women by age 35 have completed their schooling and fertility, but provide no explanation why the range extends up to 54. Furthermore, the authors do not provide a reason why the two groups built within this range (35-44 and 45 to 54) are grouped in this way. There are reasons to think that women in opposite ends of these ranges might differ significantly in our variables of interest. For example, the [35, 44] group includes "baby boomers" born before 1966 and "generation Xers" born up to 1979 that may hold very different views on education, family size and contraception and different levels of knowledge and access to contraception methods and education.

Follow-up question:

Lets suppose that the assumption Paul has made is wrong and there are some women in this cohort who either, have not finished having children or not, finished getting their education. (but not both for simplicity) What effect could this have on the observed correlation between education and child bearing?

Case 1: Where women delay child bearing for education.

Case 2: Where women delay education to bear children.

In [ ]:

```
```

** 1.3 ** Response from: *Nach Mohan*

The exploratory analysis of Fertility and Schooling by Paul Laskowski uses data from GSS. A subset of data restricting analysis to aged 35 to 44 has been used. Inherent weakness to the analysis is that it does not take into account and adjust for number of factors that could impact the analysis. Given that fertility is an important outcome, it is not clear if current marital status was included as an covariance that could have an impact on the analysis. Similarly schooling, particularly high school graduation and college graduation may depend upon number of factors including affordability and socioeconomic status. Moreover, choice of sample (type of women who were surveyed- non randomized) may also determine responses received for schooling and fertility. Number of variables such as belief, religion, location, age, socio-economic status, occupation, marital status, lifestyle etc have to be controlled for in the analysis to make a true assessment.

Follow-up question:

There are some great examples here choose one and explain how you think it would effect the analysis.

In [ ]:

```
```

**1.4.** Response from: *Katie Mo*

In the histogram of "Years of Education", there seems to also be a sizable uptick for 14 years. This could be explained by women graduating from 2-year degree programs. This bin could be included as another indicator variable to assess the graduation effect with more granularity.

Follow up question:

Can you trust the data points you see with 0-5 years of education? What might explain these responses?

In [ ]:

```

```

## 2 Exploratory Data Analysis Review

### Things to Consider

- Always look for missing data and data with wrong types
- Examine each variable's characteristics and distribution. Are there any strange features of the data?
- Consider transforming the variable, why and how would you do this?
- Tell a story about each variable. In words and in context, what is this graph telling you? Is it suprising/interesting or not in some way ?
- Are there any outliers to the data? Could it be an error or just some rare event?
- Anyone with a reasonable level of programming skill can write a program which pumps out figures and sample characteristics with no context, your job is to provide both!
- No data dumps
- Practice forming a research question about the data population, with your EDA, i.e. this feature in the data is not what I expect, everyone else is ignoring it as an error or uninteresting phenomenon, I want to explore it further and this is why. This is seriously how Nobel prizes get awarded.

## 3 Data Exercise

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.

In [1]:

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

| Variable Name | Description |
|---|---|
| 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 |

**3.1** 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.

In [ ]:

```
```

** 3.2 ** Examine the main output variable, medv. Comment on any unusual values you find, and any features that might be important for statistical modeling.

In [ ]:

```
```

**3.3** 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.

In [ ]:

** 3.4 ** Examine the bivariate relationship between medv and crime. What type of relationship do these variables have?

In [ ]:

**3.5** (As time permits) Continue your exploratory data analysis. Be prepared to share interesting findings with the class.

In [ ]: