Lab 3B - Confound it all!

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

## Finding data in new places

* Since your first forays into doing data science, you've used data from two-sources:
  + Built-in datasets from RStudio.
  + Campaign data from Mobilize Campaign Manager.
* Data can be found in many other places though, especially online.
* In this lab, we'll read an *observational study* dataset from a website.
  + We'll use this data to then explore what factors are associated with a person's lung capacity.

## Our new data

* You can find the data online here:
  + (Right-click and select *Open in New Window*) <http://gh.mobilizingcs.org/ids_labs/extras/webdata/fev.dat.txt>
* Variables that were measured include:
  + Age in years.
  + Lung capacity, measured in liters.
  + The youth's heights, in inches
  + Genders; "1" for males, "0" for females.
  + Whether the participant was a smoker, "1", or non-smoker "0".

## Importing our data

* Rather than *export*-ing the data and then *upload*-ing and *importing*-ing it, we'll pull the data straight from the webpage into R.
* Click on the *Import Dataset* button under the *Environment* tab.
  + Then click on the *From CSV* option.
  + Type or copy/paste the URL into the box and then hit *Update*.
* Before importing, change the following *Import Options*:
  + Name: lungs
  + *Uncheck* the *First Row as Data* box
  + Change *Delimiter* to *Whitespace*

## About the data

* The data come from the *Forced Expiratory Volume (FEV)* study that took place in the late 1970's.
  + The observations come from a sample of 654 youths, aged 3 to 19, in/around East Boston.
  + Researchers were interested in answering the *research question*:

*What is the effect of childhood smoking on lung health.*

## Cleaning your data

* Now that we've got the data loaded, we need to clean it to get it ready for use *(Look at lab 1F for help)*. Specifically:
  + We want to name the variables: "age", "lung\_cap", "height", "gender","smoker", in that order.
  + Change the type of variable for gender and smoker from *numeric* to *character*.
* After changing the variable types for gender and smoker:
  + For gender, use recode to change "1" to "Male" and "0" to "Female".
  + For smoker, use recode to change "1" to "Yes" and "0" to "No".

## Analyzing our data

* Our lungs data is from an observational study.
* **Write down a reason the researchers couldn't use an experiment to test the effects of smoking on children's lungs.**
* Observational studies are often helpful for analyzing how variables are related:
* **Do you think that a person's age affects their lung capacity? Make a sketch of what you think a scatterplot of the two variables would look like and explain.**
* Use the lungs data to create an xyplot of age and lung\_cap.
  + **Interpret the plot and describe why the relationship between the two variables makes sense.**

## Smoking and lung capacity

* Make a plot that can be used to answer the statistical question:

*Do people who smoke tend to have lower lung capacity than those who do not smoke?*

* **Use your plot to answer the question**.
  + **Were you surprised by the answer? Why?**
  + **Can you suggest a possible confounding factor that might be affecting the result?**

## Let's compare

* Create three subsets of the data:
  + One that includes *only* 13 year olds ...
  + One that includes *only* 15 year olds ...
  + and one that includes *only* 17 year olds.
* Make a plot that compares the lung capacity of smokers and non-smokers for each subset.
* **How does the relationship between smoking and lung capacity change as we increase the age from 13 to 15 to 17?**

## Sum it up!

* **Does smoking affect lung capacity? If so, how?**
  + Support your answers with appropriate plots.
  + Explain why you included the variables you used in your plots.