Lab 2G - Getting It Together

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

## Putting data together

* In the labs so far, we've only ever looked at individual data files.
* But often times, we gain additional insights by including additional information from a separate data set.
* In this lab, we will learn how to merge information from our *personality color* data with our *stress/chill* data.
* *Export, upload, import* your *Personality Color* data set and name it colors.
* Then, *export, upload, import* your *Stress/Chill* data set and name it stress.

## Looking at Stress/Chill

* We would like to analyze the research question:

*How do people's personality colors and/or sports participation affect their stress levels?*

* We already have data about *personality color* and a seperate data set about *stress*.
  + What we don't have is a single data set with information from both ... yet.
* We'll start then by strategizing how to merge our data together.

## Deciding how to merge

* Before we merge data, we need to decide *how* we plan to merge it:
* We can *stack* our data sets, that is, take one data set's rows and add them to the bottom of the other data set.
* We can also *join* our data sets horizontally. This is where we take one data set's columns and add them to the end of the other data set's columns based on matching an *ID* variable.
  + The *ID* variable will have entries that we use to *match* observations in both data sets.
* **To answer the statistical question of interest, would it make more sense to *stack* or *join* our colors and stress data?**

## Finding variables in common:

* Look at the names of the variables in each data set.
  + To merge different data sets together, we need to find variables they have in common.
* **Which variables do the data sets have in common?**
* **Which variable would make sense to merge the data sets together with? Why not the others?**

## Caution required

* Whether *stacking* or *joining*, we need to be careful when we merge data:
* When *stacking* data, we need to be absolutely certain that the variables we're stacking represent the exact same measurements.
  + We wouldn't want to stack height in meters and height in inches, for instance (without converting one to the other).
* When *joining* data, we need to make sure that the *id* variable in our primary data set matches to *one and only one* observation in the joining data.
  + Otherwise, R won't know which observation to match to.

## Getting ready

* Our goal is to add the variables from the colors data onto the stress data.
* Start by ensuring that every user.id in the colors data is unique.
  + If there's a duplicate, have your teacher remove the duplicate from the Mobilize Web Response Manager and then re-*export*, *upload*, *import* your colors data.
* **After we add the data from *colors* to *stress*, how many rows should our merged data have? Write this number down.**

## Putting them together

* We can use the merge function to *join* our data sets together using the variables that appear in both sets.
* **Fill in the blanks below to join the information from the colors data onto the stress.**

merge(\_\_\_\_, \_\_\_\_, by = "\_\_\_\_")

* Assign this merged data set the name stress\_colors.
  + Make sure your data has the same number of observations that you wrote down on the previous slide.

## Saving your data:

* View your merged data and make sure nothing appears to be blatantly wrong with it.
* **Why didn't we stack the rows of data instead?**
* **What happens if you swap the order of the data sets in the merge function?**
* Fill in the blank below to save our stress\_color data for later use.

save(stress\_colors, file = "stress\_colors.rda")

* Be sure to look in the *Files* tab to make sure your data was saved.

## Moving on

* In the next lab, we'll begin analyzing our merged data. In the meantime:
* **Make a few plots using variables from the stress data and *facet* or *group* the plots based on variables from the colors data.**
  + **Write down the most interesting discovery you make by just exploring your data. Write out how you found your discovery and interpret what it means for the people in your class.**
* **With our *colors* data, we could answer questions about the *typical* color scores in your class. Why can we no longer answer this question in our stress\_color data?**