Final Proposals

Your group will be submitting two components for the final proposal. The proposal itself is a written document that can be done in markdown, Word, etc. Please submit by next Wednesday, 11/5. The second component is your final, clean data set, which should be uploaded to moodle by Friday 11/7.

Content Your final proposals should contain the following content:

1. *Group Members*: List the members of your group
2. *Title*: Your title
3. *Purpose*: Describe the general topic/phenomenon you want to study, as well some focused questions that you hope to answer and specific hypotheses that you intend to assess.
4. *Data*: Describe the data that you plan to use, with specifications of where it can be found (URL) and a short description. Eventually, you will probably want to combine data from multiple sources into one file. We will discuss data management techniques in the coming weeks, but for now you should simply list multiple sources if you have them.
5. *Population*: Specify what the observational units are (i.e. the rows of the data frame), describe the larger population/phenomenon to which you'll try to generalize, and (if appropriate) estimate roughly how many such individuals there are in the population.
6. *Response Variable*: What is the response variable? What are its units? Estimate the range of possible values that it may take on.
7. *Explanatory Variables*: Describe the variables that you'll examine for each observational unit (i.e. the columns of the data frame). Carefully define each variable and describe how each was measured. For categorical variables, list the possible categories; for quantitative variables, specify the units of measurement. You may want to add more variables later on, but you should have at least 5 variables already.

Places to Find Data

Many of your already have a data set picked out, others are still fishing around. Public data sets are available from hundreds of different websites, on virtually any topic. You might not be able to find the exact data that you want, but you should be able to find data that is relevant to your topic. You may also want to refine your research question so that it can be more clearly addressed by the data that you found. But be creative! Go find the data that you want!

Below is a list of places to get started, but this list should be considered grossly non-exhaustive:

- Finding Data on the Internet (`http://www.inside-r.org/howto/finding-data-internet`)
- Gapminder (`www.gapminder.org`)
- Data.gov (`explore.data.gov`)
- StatLib at Carnegie Mellon (`http://lib.stat.cmu.edu/`)
- U.S. Bureau of Labor Statistics (`www.bls.gov`)
- U.S. Census Bureau (`www.census.gov`)

Keep the following in mind as you select your topic and dataset:

- You need to have enough data to make meaningful inferences. There is no magic number of individuals required for all projects. But aim for at least 200 observations and make sure there are at least 20 observations in each category of each of your categorical variables (if you have any).

- An interesting data set will be a mix of continuous and categorical variables.
- When you're looking at a data set, ask yourself, is this a sample from some greater population or is this the population itself? Many of the inferential techniques that we've used are most appropriate for use on sampled data.

0.1 Submitting the data set

- The data must be in CSV format (`.csv`).
- Give the file a descriptive name that clearly communicates the context and distinguishes it from the other groups' data files, e.g., group-X.csv, certainly not Data or even Our Statistics Project Data File. Do not use the following words in your file name: project, data, file, statistics, worksheet. Do not use spaces in filenames.
- Each observation should be on a separate row of the data file, and each variable should be in a separate column.
- Name all variables helpfully and contextually, e.g., use Airport and WaterTemp, not Individuals and Treatments, and certainly not A and B. Similarly, for the category names, use whole words and phrases, not cryptic codes, e.g., use Male and Female, not 1 and 2, or even M and F. A dichotomous variable Female can be coded 0 for male, and 1 for female (and then is self-documenting). A variable `sex` coded 1 and 2 is just asking for trouble.
- That said, try to limit your variable and category names to about a dozen characters. This may take some abbreviation.
- Be sure that are sufficient numbers of individuals in each category of each categorical variable! If there are categories with too few individuals for you to spot any trends or to make meaningful inferences, create an additional version of this variable with fewer, consolidated categories (perhaps including an "Other" or "Miscellaneous" category).
- Check for typos! Manual inspection is OK, but its tedious and its easy to overlook misspellings. Running some simple analyses can more quickly make most data entry errors obvious