Project Assignment 3: Upload to Courseware by Sunday 2nd Apr by 11:59 pm

Data Management & Exploratory Data Analysis

Prior to beginning this assignment, you should have created a personal codebook that includes the questions, variable labels, and response categories for the variables you intend to use for your project.

During this assignment you will:

1. Setup the RMarkdown file you will use for your project and load the necessary data and packages
2. Conduct data management steps:
    a. Create a data subset with only the variables you will use for your research project
    b. Recode NAs and error codes if necessary
    c. Further subset the data as needed (e.g. subset for particular countries, subset for only respondents who are married, or respondents in a particular age group, etc.)
    d. Create secondary variables if needed (e.g. a composite 'poverty score' or a collapsed education variable that has just two categories: tertiary and non-tertiary, etc.).
3. Generate descriptive statistics for the variables you will use for your project (sample means and standard deviations or sample proportions)
4. Create appropriate bivariate graphs for the associations you will test
5. Summarize what you have learned about your research question and associations thus far.

The detailed instructions are below. After each step, save and backup your work! Knit frequently. And don't forget to add helpful comments to your code.

1. **RMarkdown setup**
    a. Create a new folder on your computer to store all files related to your project work—codebook, RMarkdown files, the data file, etc.
    b. In RStudio, open a new RMarkdown file, edit the header appropriately, delete all other text and sample code below the header.
    c. Type your research question(s) in **bold** text just below the header. State the explanatory and response variable(s) implied by your research question and identify each as categorical or quantitative.
    d. Create the following sections in the RMarkdown file, with an empty code chunk below each section heading:

        ## 1. Load data set(s) and libraries
        ## 2. Create variable subset
        ## 3. Data management I: check for and recode errors and NAs
        ## 4. Data management II: further subset and create secondary variable
        ## 5. Descriptive statistics (sample means, standard deviations, proportions) and univariate displays
        ## 6. Bivariate tables and graphs
        ## 7. Bivariate analysis (hypothesis tests and post-hoc tests)
        ## 8. Moderation
        ## 9. Save

    e. Knit the RMarkdown file to test that your headings render in MS Word
    f. In section 1., enter the code to load in the datafile you will use; use the library() function to load in the descr and stats packages.
    g. Knit the RMarkdown file again to test that the data and packages load.
2. **Data management**
    a. Create a variable subset:

i. Use the section 2 code chunk to create a data subset with only the variables you intend to use for your project. The code for this is in several sample scripts we have used in class. Use those scripts as a guide.
ii. If your research question involves two rounds of Afrobarometer, then you will have two subsets.
iii. Run the code to make sure it works and then test that the RMarkdown file will knit.
iv. Troubleshoot and fix any problems before continuing.

b. Data management 1: In section 3, you will need to decide if NAs need to be recoded and if error codes need to be recoded. These decisions well vary depending on the dataset and specific variables you are using. Read your codebook carefully. If you are unsure, please discuss with your FI or faculty member.
    i. For each ***categorical*** variable, run freq(). Look at the output in the console and the barplot in the plotting window. Are there NAs? Are there error codes?
        - Reference the codebook to understand the meaning of NA for the particular variable. In some datasets NA simply means missing data so there's nothing more to do. In other datasets and for certain questions, NA means the question was skipped. For example, in NESARC, if the respondent doesn't drink at all, there's no reason to ask them how many times per week they drink. In this case, the NA implies 0 drinks per week, so the NA should be recoded into an appropriate category.
        - Reference the codebook to understand the error codes for the variable. All error codes should be re-coded to NA, but make sure you recode error codes to NA AFTER you've dealt with the need to recode NAs!
    ii. For each quantitative variable, use the hist() function to view the distribution and identify error codes. Verify the error codes values in the codebook. Assign NA to each error code. Run the hist() function again to make sure error codes were removed as expected.
    iii. NOTE: treat variables that produce a score, say from 1 to 5, as categorical for the data management step, even though you may treat them as quantitative (e.g. calculate a sample mean score) in later steps.

c. In section 4, further subset the data as needed. You may do one or more of the steps below, depending on your research question and associations you want to test:
    i. Use the subset() function or indexing with [] to further subset your data. For example, if your research only involves couples that are married then create a married subset. If your research question involves one or more African countries from the Afrobarometer, than create a subset for just those countries. If your research question only involves a particular age group, then subset by age.
    ii. Create secondary variables. Secondary variables are variables created from the 'primary' variables in the original dataset. For example, you may want to collapse a religion variable with 13 categories into just four: Christian, Muslim, Traditional, Other. Use sample code to see how this is done.
    iii. Another use of secondary variables is to aggregate responses into a type of score. For example, you could create a 'poverty score' which counts the number of times an Afrobarometer respondent answered that they went without food, medicine, clean water, cash income, or fuel to cook food in the past year. See sample code for creating aggregate variables.

d. IMPORTANT: Verify that counts are greater than 30 for each response category in the variables you aim to use for chi-square, ANOVA, t-test or linear regression. Review the output of the freq() function to verify the counts for each category.

3. Generate descriptive statistics

The type of descriptive statistics that you will calculate depends on the role-type classification for each association you will test. The code for generating descriptive statistics goes in section 5. See code samples from class.

a. C → Q: Report the mean value of the response variable for each category of the explanatory variable. (ii) Report the standard deviation of the response variable for each category of the explanatory variable. The tapply() or by() functions can be used to do this.
b. C → C: Create a table that shows the proportion of the response variable for each category of the explanatory variable. Important: explanatory variable categories are columns and response variable categories are rows. The table() and prop.table() function can be used to do this.
c. Q → Q: Calculate and report the mean and standard deviation for the quantitative explanatory and quantitative response variable separately. Use the mean() and sd() functions to do this.
d. Note that if you are using more than one round of Afrobarometer, than you should report the descriptive statistics for each round separately. Also, if you are testing the association between two variables testing to see if the pattern is different for different countries, then you would also report the descriptive statistics for each country separately.

4. **Create appropriate bivariate graphs for the associations you will test**

   The type of bivariate graph depends on the role-type classification for each association you will test. The code for generating the graphs goes in section 6. See code samples from class.

   IMPORTANT: These graphs need to have an accurate and descriptive main title, appropriate category and/or x-axis labels, and a vertical axis label with the correct unit of measure.

   a. C → Q: (i) Box and whisker plot that shows the distribution of the quantitative response variable for each category of the categorical explanatory variable or (ii) a barplot where the height of the bar is the mean of the response variable for each category of the explanatory variable.
   b. C → C: Barplot where each bar is a category of the explanatory variable and the height of the bars are the proportion of respondents for ONE category of the response variable. If the response variable has multiple categories of interest, then you would create a separate barplot for each.
   c. Q → Q: Scatterplot with a line of best fit. The explanatory variable is on the horizontal axis and the response variable is on the vertical axis.
   d. Most research projects will have between 2 and 5 bivariate graphs.

5. **Summarize what you have learned**

   At this point, you will not have "answered" your research question, but you will have learned quite a bit about the variables you are using and can speculate about possible relationships that may or may not exist between explanatory and response variables. Write two paragraphs summarizing what you have learned so far about your research question and associations. Write this text below the section 6. code chunk.

KNIT AND SAVE YOUR WORK

ORGANIZE YOUR ASSIGMENT SUBMISSION: Upload **the knitted MS Word document** into the Project Assignment 3 submission folder.