Project

Navigation :

The purpose of the data project is for you to conduct a reproducible analysis with a data set of your choosing. There are two components to the project, the proposal, which will be graded on a pass/fail basis, and the final report. The outline for each of these are provided in the templates. When submitting the assignments, include the R Markdown file (change the name to include your last name, for example `Bryer-Proposal.Rmd` and `Bryer-Project.Rmd`) along with any supplementary files necessary to run the R Markdown file (e.g. data files, screenshots, etc.). Suggestions for possible data sources are included below, however you are free to use data not listed below. The only requirement is that you are allowed to share the data. Projects will be shared with others on this website so should be presented in a way that other students can reproduce your analysis.

Project Proposal

The proposal can be more informal using bullet points where necessary and include R code and output. You must address the following areas:

- Research question
- What are the cases, and how many are there?
- Describe the method of data collection.
- What type of study is this (observational/experiment)?
- Data Source: If you collected the data, state self-collected. If not, provide a citation/link.
- Response: What is the response variable, and what type is it (numerical/categorical)?
- Explanatory: What is the explanatory variable(s), and what type is it (numerical/categorival)?
- Relevant summary statistics

Example data project proposal (Source Rmarkdown file)

Final Project Format

The final report should be presented in more formal format. Consider your audience to be non data analysts. Fellow data analysts (i.e. students) will be able to access your R Markdown file for details on the analysis. Submit a Zip file with your R Markdown file, the HTML output, and any supplementary files (e.g. data, figures, etc.). You must address the five following sections:

1. **Introduction**: What is your research question? Why do you care? Why should others care?

2. **Data**: Write about the data from your proposal in text form. Address the following points:

    - Data collection: Describe how the data were collected.
    - Cases: What are the cases? (Remember: case = units of observation or units of experiment)
    - Variables: What are the two variables you will be studying? State the type of each variable.
    - Type of study: What is the type of study, observational or an experiment? Explain how you've arrived at your conclusion using information on the sampling and/or experimental design.
    - Scope of inference - generalizability: Identify the population of interest, and whether the findings from this analysis can be generalized to that population, or, if not, a subsection of that population. Explain why or why not. Also discuss any potential sources of bias that might prevent generalizability.
    - Scope of inference - causality: Can these data be used to establish causal links between the variables of interest? Explain why or why not.

3. **Exploratory data analysis**: Perform relevant descriptive statistics, including summary statistics and visualization of the data. Also address what the exploratory data analysis suggests about your research question.

4. **Inference**: If your data fails some conditions and you can't use a theoretical method, then you should use simulation. If you can use both methods, then you

should use both methods. It is your responsibility to figure out the appropriate methodology.

    - Check conditions
    - Theoretical inference (if possible) - hypothesis test and confidence interval
    - Simulation based inference - hypothesis test and confidence interval
    - Brief description of methodology that reflects your conceptual understanding

5. **Conclusion**: Write a brief summary of your findings without repeating your statements from earlier. Also include a discussion of what you have learned about your research question and the data you collected. You may also want to include ideas for possible future research.

Data Sources

You are not to use data sources used in class or the textbooks. Possible data sources include, but are not limited to:

- FiveThirtyEight https://github.com/fivethirtyeight/data
- RStudio data sources http://blog.rstudio.org/2014/07/23/new-data-packages/
- Analyze Survey Data for Free (ASDFree) has many open data sources that can be used http://www.asdfree.com/
- The World Bank Data Catalog http://datacatalog.worldbank.org/
- Google Public Data search engine http://www.google.com/publicdata/directory
- Vanderbilt data sources http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets
- Programme of International Student Assessment (PISA) http://www.oecd.org/pisa/
- Behavioral Risk Factor Surveillance System (BRFSS) http://www.cdc.gov/brfss/
- World Values Survey http://www.worldvaluessurvey.org/wvs.jsp
- American National Election Survey (ANES) http://www.electionstudies.org/
- General Social Survey (GSS) http://www3.norc.org/GSS+Website/
- Integrated Postsecondary Education Data System (IPEDS) https://nces.ed.gov/ipeds/

- U.S. Census and American Community Survey https://cran.r-project.org/web/packages/acs/index.html