Introduction to Applied Statistical Computing with R

COURSE PROJECT INFORMATION

Richard E.W. Berl

Spring 2019

Last Updated: 2019-04-10

Overview

The course project is designed to have you utilize the skills you've learned in this course and apply them to a research question using real data relevant to your research. Your research question (or questions; you may begin with multiple related questions or come up with follow-up questions as you proceed) should be carefully crafted to fit the scope of this project. For example, "How is forest cover related to atmospheric carbon cycling?" would be much too broad in scope, while a question like "Did general approval of legalized marijuana in Colorado by adults in Fort Collins change from last year to this year?" that only takes one test to answer is too narrow. Aim for a middle ground that will take a few different analyses to investigate from different angles or using different measures.

You should take this project as an opportunity to analyze some of your own data, if you have some, or data similar to those you expect to work with, if you don't. If you are analyzing your own data, your analyses and report could serve as the foundation for a future published paper. With a little attention to professional presentation, the report can also be a nice portfolio item to showcase your skills in data analysis and communication.

However, you are free to use the project for any purpose you'd like; for example, as a side project to explore a different area of interest or to build your skills in using unfamiliar methods. Another good option, if you don't have data of your own, is to attempt to replicate the analyses of a published work in your field for which data are available. Any of these are acceptable approaches for this project. A list of potential sources for many different types of data sets are given later in this document.

A note on self-plagiarism: Though it may sound ridiculous, self-plagiarism is a thing and policies at CSU forbid it. Your project and the analyses performed in it should be your own original work, done specifically for this course.

Requirements

Project Proposal

(Due April 23, 11:59pm)

A couple paragraphs (no more than 300 words), in the style of an abstract or summary for a grant, succinctly describing your research question and how you plan to address it. If you have more than one idea that you may use for your project, desribe each one separately. You are not required to do stick with the topic you describe in your proposal but, if you decide to change topics, it will still need to have a well-defined research question.

Project Report

(Due May 17, 11:59pm)

The written report for your project (1,500 to 3,000 words, not including code, output, figures, and references), which describes the full procedure you used to conduct your analyses in a reproducible format. This means that your report may be slightly different from a traditional academic paper or report, in that it will be in more of a narrative style that describes and shows the analyses you performed as part of a flow of ideas from beginning to end. Code and output should alternate with running commentary on why each test was performed and what was found. Remember that you are telling a story–the structure of your report should reflect that.

For ease of presenting information in this style, you are strongly encouraged to write and submit your report as a knitted R Markdown document (to HTML, or to PDF if you have a version of LaTeX installed). However, you may also insert code chunks, output, and plots into a traditional document using a word processing program such as Word. If you do so, all R code and output still needs to be included, and should be indented from your written text and formatted in a monospaced font (such as Courier New). All code should be fully commented, no matter the format.

If it is helpful, you can install the `wordcountaddin` package (from GitHub, with instructions) to count words in an Rmd file. You'll need to install the `devtools` package first (`install.packages("devtools")`), and both packages may take several minutes to compile and install. You may also need to install the `koRpus` package, if it isn't done automatically. Afterwards, you can find your word count like so:

```
library(wordcountaddin)
word_count()
```

```
## [1] 1337
```

Below is a sample format for your report, with items to include:

Introduction

- The what and why of your project: some background with a couple citations (as necessary) and a description of your research question. Why do you care about this question? Can reuse elements from your Project Proposal.
- A brief overview of the methods and results.
- The source of the data (and a quick description of the methods used to collect it, if your own).

Methods & Results

- A list of packages and any other external resources required to replicate your analyses.
- Loading packages and data.
- Inspection of the data, its format, and the meaning of each variable (some details could go in the Introduction if they make more sense there).
- Data cleaning, wrangling, and tidying, as necessary. Address any problems in your data (missing data, outliers, collinearity, etc.).
- Exploratory analyses (at least one plot required, can be rough).
- Hypothesis tests, regressions, and/or other analyses needed to answer the research question, with ongoing description and observations.
- Polished and labeled publication-ready figures that effectively communicate the most important results to the story.

Discussion

- A summary of findings and their relevance to the research question.
- Can include what you learned, difficulties encountered and changes made.
- Final conclusion on your research question given your results.

References

- List of any references cited (including the source of the data), in a reference style of your choice.

Project Presentation

(Due week of May 13)

A short talk presenting your research question, methods, and findings to the class in the form of a 5-minute lightning talk with a maximum of 5 slides (not including title slide).

These will take place during finals week, at a date and time we arrange as a class. Come prepared with your slides made in PowerPoint, Google Slides, or R Markdown (if you're feeling adventurous). Bring a backup copy in PDF format.

Resources

Forming a research question

- Writing@CSU. Developing a research question. Available: https://writing.colostate.edu/guides/guide.cfm?guideid=25
- The Writing Center, George Mason University. How to write a research question. Available: https://writingcenter.gmu.edu/guides/how-to-write-a-research-question

Writing a reproducible report

- Shalizi, C. Using R Markdown for class reports. Available: http://www.stat.cmu.edu/~cshalizi/rmarkdown/
- Frank, M., & Hartgerink, C. R Markdown for writing reproducible scientific papers. Available: https://libscie.github.io/rmarkdown-workshop/handout.html

Data sets

Packages

- `datasets` package
- `carData` package
- `HistData` package
- `BSDA` package

Collections

- Awesome Public Datasets
- "Data Is Plural" Archive
- Reddit r/datasets

Government/Global

- OpenData Fort Collins
- Colorado Information Marketplace
- Data.gov
- data.Census.gov
- World Bank Open Data
- Integrated Public Use Microdata Series

Polling and News

- Pew Research Center
- Reuters
- FiveThirtyEight
- BuzzFeed News

Nonprofits

- United Nations
- UNESCO Institute for Statistics
- Gapminder
- Data for Democracy
- Southern Poverty Law Center
- ProPublica

Biology, Ecology, and Natural Resources

- Global Biodiversity Information Facility
- Protected Planet
  - `wdpar` package
- Environmental Data Initiative
- MorphoBank

Anthropology and Linguistics

- D-PLACE: Database of Places, Languages, Culture and Environment
  - dplace-data
- Pulotu: Database of Pacific Religions
- WALS: World Atlas of Language Structures
- Glottobank (TBD)

Historical

- John Snow's Cholera Data
- Slave Voyages

Modern/Fun

- Global Database of Events, Language, and Tone
- Yelp Open Dataset
- Million Song Dataset
- Internet Boy Band Database
- So Much Candy Data, Seriously
- Game of Thrones
  - Datasets and Visualizations (json)
  - Character Interaction Networks
  - War of the Five Kings
  - Bayesian Survival Analysis
    - Character Data
- Dungeons & Dragons
  - Character Names
  - Character Biographies
  - Spell Names

Geospatial

- Natural Earth
- USGS EarthExplorer
- TIGER/Line Shapefiles

- Data Basin
- GeoDa Data and Lab

General Repositories

- Google Dataset Search
- Dryad
- figshare
- Harvard Dataverse
- UCI Machine Learning Repository
- Stanford Large Network Dataset Collection
- Open Data on AWS
- data.world
- Kaggle

(pdf / Rmd)