Project Guidelines

Guidelines for Final Project for Intro to R

This project is supposed to get you used to working with R in a way that would be conducive for collaborating or creating reproducible analyses.

1. Please identify a dataset to analyze. Any dataset will do, as long as you can perform the following requirements and you **do not violate any privacy restrictions** for the data. Also please ensure that the instructors have access to your data so that they can evaluate your project.

Thus send us a link to your data if it is hosted somewhere or make sure to turn in a data file (csv, excel etc.) on CoursePlus. If the data is not hosted publicly somewhere, you must turn it in on Drop Box with your other files.

You are also free to create your own data if you wish, but please ensure that it is large enough to perform the rest of the following requirements.

Options for places to find data are:

- from Kaggle: https://www.kaggle.com/datasets (https://www.kaggle.com/datasets)
- from the US Government: https://data.gov/ (https://data.gov/)
- from the CDC: https://data.cdc.gov/ (https://data.cdc.gov/)
- from healthdata.gov: https://healthdata.gov/ (https://healthdata.gov/)
- from Baltimore City: https://data.baltimorecity.gov/ (https://data.baltimorecity.gov/)
- from R itself: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html)
- from the `tidytuesdayR` package: https://github.com/thebioengineer/tidytuesdayR (https://github.com/thebioengineer/tidytuesdayR)
- from the `catdata` package: https://CRAN.R-project.org/package=catdata (https://CRAN.R-project.org/package=catdata)
- from the `dataplay` package: https://github.com/avahoffman/dataplay (https://github.com/avahoffman/dataplay)
- from https://github.com/awesomedata/awesome-public-datasets (https://github.com/awesomedata/awesome-public-datasets), which links to many more sources (be sure to click “view raw” if the preview doesn’t show up)

To use the data that comes with R, enter `datasets::` and press *tab* in RStudio to see the names of the datasets - for example `datasets::ability.cov` will load the ability dataset.

You are not limited to these options for finding your data.

2. Please **describe** what your data looks like and **where** you got it. Identify what the variables and samples are. Describe how the data was originally created.

(.5 points total - 0.25 points for describing data, 0.25 points for describing where you got your data)

3. Perform at least **three different** data subsetting, cleaning, or manipulation methods that were described in this course on your data. Examples are: renaming the columns, recoding values, reshaping the data, filtering the data etc.

(3 points - per successful method)

4. Please **describe** what you did to clean/subset/wrangle/manipulate your data and why.

(1 point)

5. Make **two different** kinds of visualizations of your data using `ggplot2` .

(2 points total: 1 point for each of 2 **different** kinds of plots completed)

6. Perform a **simple analysis** of your data. This can involve summarizing the data to describe aspects about it (quartiles, means, range etc.) or you may perform a simple statistical test.

(2 points total, 1 point for attempt)

7. Describe what analysis you performed and why. Provide some simple **interpretation** about what your analysis might indicate about your data. You ***will not*** be graded based on the validity of your **statistical interpretation**, but rather the implementation and description of what you did.

(1 point total - 0.5 points for describing analysis, 0.5 points for interpretation)

8. Add `sessionInfo()` to the end of your analysis so that you and others can see what version of R and packages you used.

(0.5 points)

9. All steps and descriptions should be written in an RMarkdown file that is rendered to an html file. Code should be included in code chunks. Please make sure that your file knits properly to create an html file. **Submit at least the Rmarkdown (.Rmd) file**.

Please see the project example on our website for an example project: Source code Rmd (ProjectExample/Final_Project_Example.Rmd) and the output html (ProjectExample/Final_Project_Example.html).

Bonus: Create a function as part of your analysis. If you do this correctly, it can make up for lost points on other sections. (1.5 points)

10 points total plus bonus

Grading Rubric:

Item	Description	points
Describe Data Source	Describe what your data looks like. Identify what the variables and samples are. Describe how the data was originally created.	0.25
Describe Data	Describe where you got your data	0.25
Wrangling - cleaning, subsetting, manipulation (ex renaming, recoding, reshaping, filtering)	Perform at least three different methods	3
Describe wrangling and reason	Please describe what you did to clean/subset/wrangle/manipulate your data and why.	1
Data Viz	Make 2 different kinds of plots	2
Data Analysis	Perform a simple analysis of your data. This can involve summarizing the data to describe aspects about it (quartiles, means, range etc.) or you may perform a simple statistical test.	2
Describe Analysis	Describe what analysis you performed and why.	0.5
Interpret Analysis	Provide some simple interpretation about what your analysis might indicate about your data. You will not be graded based on the validity of your statistical interpretation, but rather the implementation and description of what you did.	0.5
Session info	Need to have session info	0.5
Bonus	Create a function as part of your analysis. If you do this correctly, it can make up for lost points on other sections. (1.5 points)	1.5 extra