STAT228 Mini-Project 2: Messy to Meaningful

A Tutorial on Data Wrangling

Due Friday, April 11, 5:00 PM

Change log

Date	Change	Description
2025-03-25	Initial release w/ updated deadline	Deadline extended several days to accomodate religious holiday.

Overview

In this mini-project, you'll contribute to your growing data science portfolio by **writing a tutorial on data wrangling in R**[1].

Your goal is to effectively communicate the entire data wrangling and cleaning process in the context of a real dataset. While the goal of this project is to convey *how* one may manipulate/organize data, the chosen dataset should still provide context in the sense that the tutorial flows from beginning to end, and each step is well-motivated.

The target audience for your tutorial is someone who is familiar with R, but new to data wrangling, e.g., a friend who hasn't taken STAT228 or a potential employer. Writing for a beginner might feel challenging, but that's the point! Teaching someone else is one of the best ways to show you really understand something. If you can explain your process clearly and simply, it means you truly get it — and that's exactly what this project is about. :)

What to include

Your tutorial should include the following:

- **Introduction:** text that introduces data wrangling and the dataset that you'll use
- **Data wrangling process with exercises and data visualizations:** text, code, and data visualizations that walk through the data wrangling process. You do **not** have to cover every function covered in class; however the tutorial should be more complex that a laundry list of of `mutate()`s and `filter()`s calls. Create something that is insightful and well-crafted.
- **Conclusion:** text summarizing the data wrangling process and key takeaways

You will compose this blog post tutorial as an R Studio project. You will then upload a zipped file to moodle containing the following:

1. An `.Rproj` file
2. An `.Rmd` blog post tutorial
3. The knitted `.html` file

[1] Special thanks to Professor Lauren Trichtinger for providing the basis of this project.

4. Any external files (e.g., data files, images, etc) (if applicable)

Be sure to disable any unnecessary messages that appear when loading packages, warning messages from plots, etc. by using appropriate chunk options. Pretend you're submitting this post as part of a job application – you want it crisp and clean!

Optional: make the tutorial interactive with the learnr R Package

A particularly useful way of writing a tutorial is to use the `learnr` R package to make an interactive tutorial. An interactive tutorial is not required for this project; however it can make your project stand out if you're thinking about using this a portfolio piece for a future job application.

Submission instructions

Like mini-project 1, you will compose this blog post as a self-contained R Studio project.

1. Write your blog post tutorial in R Markdown and knit to `HTML` (this is a blog post after all :)).
2. Upload a zipped folder of your RStudio project folder to Moodle. If you run into issues with moodle, send me your folder via email *before* the deadline.

Important: your `.Rmd` must knit without errors to ensure reproducibility.

Due date

The mini-project is due Friday, 4/11 at 5 PM and counts for 15% of your grade.

How will I be graded?

The rubric is available in this google sheet.

Late Work

Late submissions for the project will incur a penalty of 10% per day.

Where should I find my data?

Like mini-project 1, you are free to use whatever data you want as long as it's different than the one you used in mini-project 1 and isn't a dataset we've analyzed heavily in class or homeworks.

Here a few suggestions:

R packages

- `babynames`: history of baby names from the Social Security Administration
- `Lahman`: comprehensive historical archive of major league baseball data. Read more about this package here.
- `fueleconomy`: fuel economy data from the EPA, 1985–2015
- `fivethirtyeight`: provides access to data sets that drive many articles on FiveThirtyEight. Read more about this package here.

Other data sources

- TidyTuesday – A lot of **great** and interesting datasets from a wide range of topics **- I strongly recommend at least considering this source!**
- Kaggle Datasets – Need to set up an account to download datasets, but it's simple and free

- Awesome Public Datasets
- Anything else you're excited about.

Where can I find examples of data wrangling tutorials?

There are many data wrangling tutorials on the web, in your textbook, and elsewhere. Here are some examples (certainly not an exhaustive list):

- Transitioning into the tidyverse (This is longer than yours needs to be!)
- 10 Tricks for tidyverse in R
- A Beginner's Guide to Tidyverse (Content is slightly different from what we've covered, but format might be useful.)
- An Introduction to Tidyverse (Again, longer than yours needs to be.)

Tips

- Start early
- Knit your notebook early and often
- Before submitting, knit your R notebook on another computer. Confirm all external datasets are in the Rproject and the file knits without errors
- Lower your stress by submitting early