1.1 Prepare Datasets

Published | Assign To | Edit | ⋮

Download this step-by-step PDF (http://static.trogu.com/documents/classes/523/2025/fall2025/1.1_prepare_datasets_step-by-step_F2025.pdf) to complete this assignment. The PDF also includes links to videos. Some links in the PDF have changed but we will go over the changes in class as we do the exercise together.
Below is some important software information and the list of files to upload at the end of the assignment.

Project 1 is a 4-week project composed of three parts:

This page is Part 1: 1.1 Prepare datasets
2 clean comma-separated value files (CSV)
3 plots (PDF)
1 R file (R)

Part 2: 1.2 Generate graphs (line, bar, scatterplot) (https://sfsu.instructure.com/courses/61861/assignments/697651)
Part 3: 1.3 Clean up and format final graphs in Illustrator (https://sfsu.instructure.com/courses/61861/assignments/697652)

Note: for spreadsheets, you must use Excel, not Google Sheets, because by editing in a web browser, HTML can introduce "dirty" characters like "/" that can cause errors later in the visualizations. The clean dataset file should always be checked with a text-only editor. Download all the necessary programs, including Excel (https://its.sfsu.edu/service/office365students), BBEdit (https://www.barebones.com/products/bbedit/) (Mac) or Notepad++ (https://notepad-plus-plus.org/downloads/) (PC), and R (https://www.r-project.org/).

Some notes on the plots and info for dowloading and installing R and RStudio.

The goal of this project is to visualize the disparity in breast cancer mortality between

White females and Black females, and show that the rate is almost 50% higher for Black, despite them having a lower incidence rate (fewer cases per 100K). We will not look at causes in this assignment, but it turns out that while lack of health care plays a big role, genetics also does.

After creating the two dataset files, plot a matrix for each using R, showing every possible combination of pairs of variables (columns). There is a bit of coding, just a few short lines. Plot also a scatterplot of just the white and black death rates.

Note: although R (or any other program) won't be required as the only program to use in the class, this little exercise will show that it's very reliable to just get a base graph plotted, and therefore I highly recommend it. Illustrator can then be used to clean up the base graph.

Use the links below to download and install all the necessary software for this assignment:

BBEdit (Mac)

• Install the *text-only editor* **BBEdit (https://www.barebones.com/products/bbedit/download.html)** — all the demos use this particular program so it will be easier to match demos with your work.

https://www.barebones.com/products/bbedit/download.html (https://www.barebones.com/products/bbedit/download.html)

Notepad++ (PC)

https://notepad-plus-plus.org/downloads/ (https://notepad-plus-plus.org/downloads/)

Excel (MS Office)

• Install **Excel (https://its.sfsu.edu/service/office365students)**, it's free from the university and part of the full Microsoft Office 365 suite:

https://its.sfsu.edu/service/office365students

(https://its.sfsu.edu/service/office365students)

Note; work in Google Sheets will not be accepted, you must install Microsoft Excel!

R

• Download and install R (R Project: https://www.r-project.org (https://www.r-project.org/) — direct links below):
From one of the servers: https://cran.r-project.org/mirrors.html (https://cran.r-project.org/mirrors.html)
For example Iowa State University: https://mirror.las.iastate.edu/CRAN/ (https://mirror.las.iastate.edu/CRAN/)
The R website is very old school looking — just make sure you download the correct version for your system, Mac or PC.
Once on the mirror site, download the latest version for Mac. Check you OS for compatibility but the two versions below (Mac or PC) should work.

https://cran.r-project.org/bin/macosx/ (https://cran.r-project.org/bin/macosx/)

Latest release: R-4.5.1-arm64.pkg (https://cran.r-project.org/bin/macosx/big-sur-arm64/base/R-4.5.1-arm64.pkg) (Please note that this version only works on silicon Macs: M1,2,...)

For older Intel-based mac: R-4.5.1-x86_64.pkg (https://cran.r-project.org/bin/macosx/big-sur-x86_64/base/R-4.5.1-x86_64.pkg)

Or the latest version for Windows: https://cran.r-project.org/bin/windows/ (https://cran.r-project.org/bin/windows/)

Click on base (https://cran.r-project.org/bin/windows/base/), then click on: Download R-4.3.1 for Windows (https://mirror.las.iastate.edu/CRAN/bin/windows/base/R-4.5.1-win.exe)

RStudio

After installing R (RStudio will remind you) download and install also RStudio (https://posit.co/download/rstudio-desktop/), which is a graphic interface that runs on

top of R. Download the free RStudio Desktop version: **https://posit.co/download/rstudio-desktop/ (https://posit.co/download/rstudio-desktop/)**

Go to the **step-by-step PDF (http://static.trogu.com/documents/classes/523/2025/fall2025/1.1_prepare_datasets_step-by-step_F2025.pdf)** for file and plot instructions. (See also note at top of this page). Upload 6 files by the deadline:

1. lastName_brecan_75_17.csv
2. lastName_brecan_wb_2019.csv
3. lastName_7517_plot1.pdf
4. lastName_wb_19_plot2.pdf
5. lastName_wb_19_scatterplot3.pdf
6. lastName_brecan.R

Points	10
Submitting	a file upload

Due	For	Available from	Until
Aug 27 at 8:30am	Everyone	-	-

+ **Rubric**