

# Homework 2 Solutions

Stats 32: Introduction to R for Undergraduates

Due: Thursday, April 18, 2024, 11:59 am PT

## Instructions

You may (in fact, are encouraged to) use the Internet (including AI assistants, like ChatGPT) to search up any information to help you with this assignment, though you must cite any external (i.e. non-course related) resources that you use. Similarly, *after attempting this assignment by yourself*, you may collaborate with other students in the course, but you must each write your own code and acknowledge all students with whom you collaborated *for each problem* (you don't need to cite by subpart). However, you may not post on Internet forums (e.g. Stack Exchange) for help with this assignment; doing so is considered an Honor Code violation. You also may not copy verbatim any significant amount of code from the Internet (including AI assistants, like ChatGPT), even with citation. Feeding in the problems directly into AI assistants (or substantively paraphrased version) is also not permitted.

Please provide your code responses to each problem in the `.Rmd` file in the R code chunks directly below each subpart, inserting additional R code chunks if needed. Any text response can go right underneath the corresponding question.

On Gradescope, please submit a single `.pdf` file created by knitting the document with your responses. Problem 0 will provide guidance on how to do this.

Credit is given based on the approach and code, not necessarily the final answer.

---

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr    1.5.1
## v ggplot2    3.4.4      v tibble     3.2.1
## v lubridate  1.9.3      v tidyr      1.3.1
## v purrr      1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

## Problem 1

On Canvas, `droughts.csv` contains the percentage area of each county in California that was in each of 5 possible categories of drought — D0, D1, D2, D3, and D4 — and not in drought (None) at the end of each week in the years 2000 through 2020. Assume each observation is taken on the date given in `MapDate`.

- (3 points) Download and read in the `droughts.csv` file on Canvas, as a tibble. Use `dplyr::select()` to drop the columns `ValidStart` and `ValidEnd`, and store the result in a variable called `droughts`. Create a column called `SevereOrWorse` which corresponds to the percentage land area of the county in categories D2, D3, and D4 of drought.

- (b) (3 points) Verify, using `group_by()` and `summarise()`, that for each `MapDate` and county pair, there is exactly one entry.
- (c) (2 points) For each variable in `droughts`, justify whether it is categorical or quantitative. It may help to look up what a FIPS code is.
- (d) (4 points) What was the average percent area of Santa Clara County that was in severe or worse drought in 2020? Hint: You may want to create a new column called `Year` that specifies the year of the observation based on `MapDate`. You can do this with `math` (modular arithmetic or truncation), or by extracting the first 4 characters in `MapDate`. The question is asking you to average over all weeks in 2020.
- (e) (2 points) Download and read in `USA_counties.csv` from Canvas and provide some numerical summary information (minimum, median, maximum) about the `SQMI` variable.
- (f) (4 points) Generate a tibble that contains the percentage land area of California in severe or worse drought for each `MapDate`. Order the tibble by `MapDate`, from most recent to least recent, and store it in a variable called `CA_severe_percent`. Then print out the first few rows of this tibble. Hint: You will need the information from `SQMI` in the previous part.
- (g) (2 points) Repeat part (f), but for the percentage land area of the 9 Bay Area counties (Marin County, Napa County, Sonoma County, Solano County, Alameda County, Contra Costa County, Santa Clara County, San Mateo County, and San Francisco County) in severe or worse drought for each `MapDate`. Store this tibble in a variable called `Bay_Area_severe_percent`.
- (h) (4 points) In what proportion of weeks from 2000-2020 did the Bay Area have a higher percentage of its land area under severe or worse drought than California as a whole?
- (i) (1 point) Save `CA_severe_percent` from above as `CA_severe_percent.csv` in a folder where you won't lose it.