

DS5110 Homework 2

Kylie Ariel Bemis

22 January 2024

Instructions

Your solutions should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. *Make sure that you answer all parts of the problem.*

Submit your solutions on Canvas by the deadline displayed online. For full credit, your submission must include exactly two files:

- R Markdown (.Rmd)
- Knitted PDF report (.pdf)

Problems must appear in order, and problem numbers must be clearly marked. Any written responses should appear outside of code blocks and use Markdown for text formatting. Code comments are encouraged, but will be ignored for grading purposes. Solutions that are especially difficult to grade due to poor formatting will not receive full credit.

All solutions to the given problems must be your own work. If you use third-party code for ancillary tasks, you **must** cite them.

Part A

Problems 1–2 ask you to find a dataset that is personally interesting to you. It may be a publicly-available dataset, or a dataset for which you have permission to use and share results. There are many places online to find publicly-available dataset, and simply searching Google for your preferred topic plus “public dataset” may provide many hits. Here some additional resources to get you started:

- US Government datasets (<https://catalog.data.gov/dataset>)
- Center for Disease Control (CDC) data (<https://data.cdc.gov>)
- Bureau of Labor Statistics (<https://www.bls.gov/data/>)
- NASA datasets (<https://nssdc.gsfc.nasa.gov>)
- World Bank Open Data (<https://data.worldbank.org>)
- Kaggle Datasets (<https://www.kaggle.com/datasets>)

Problem 1

Import the dataset into R. Perform any preprocessing (tidying and cleaning) necessary for visualizing the data.

Cite the source for the dataset. Describe the data, the variables, and any preprocessing steps you performed.

Problem 2

Visualize something interesting to you from the dataset using `ggplot2`. Comment on what the visualization shows and any key conclusions.

Part B

Problems 3–5 use data on NCAA student-athlete academic performance. Download the data files from “NCAA-D1-APR-2003-14.zip” on Piazza. The files include the codebook and tab-delimited data for team-level Academic Progress Rates (APRs) of Division I student-athletes from 2003-2014.

A team’s APR is calculated out of a maximum score of 1000 points, and takes into account a team’s academic eligibility and retention, to derive an overall cohort rate of academic progress.

Import the dataset into R using the `readr` package, making sure that any missing data codes are imported as NAs.

Problem 3

Create a tidy data frame that includes columns for:

- School ID
- School name
- Sport code
- Sport name
- Year
- APR

All other columns can be discarded.

Use your tidied dataset to visualize the distributions of APRs over time. How does the distribution of APRs change year-to-year from 2004 to 2014?

Hint: The `tidyr::starts_with()` and `stringr::str_sub()` functions may be useful.

Problem 4

We would like to compare APRs between men’s and women’s sports. Transform your tidied dataset to remove mixed sports, and create a column indicating the gender division of each sport. (You may assume sport codes 1-18 are men’s, and 19-37 are women’s.)

Visualize the distributions of APRs over time again, but broken down by gender division. How do the average APRs compare between men’s and women’s sports? Does this relationship hold true across each year from 2004 to 2014?

Hint: The `ifelse()` function may be useful.

Problem 5

We would like to further visualize APR by both gender and specific sports. Process the the sport names to remove the “Men’s” and “Women’s” prefixes so that we can compare men’s and women’s teams within each sport. Then visualize the distribution of APR for both men’s and women’s teams for each sport. Are there sports where men’s and women’s teams have similar APRs?

Hint: The `stringr::str_remove()` function may be useful.