DS5110 Homework 5 - Solutions

Kylie Ariel Bemis

10 March 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 three files:

- R Markdown (.Rmd)
- 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 use the "Flash Paper" from Canvas Discussions.

Problem 1

Choose one of the Flash Papers created by your fellow classmates and posted on Canvas Discussions. Cite both the name of the student whose Flash Paper you choose and the original source of the dataset used in the Flash Paper.

Download and import the dataset into R. Perform any preprocessing (tidying and cleaning) that the original author describes. Describe these steps and any challenges encountered (if any).

Problem 2

To the best of your ability, reproduce the figures from the Flash Paper you chose in Problem 1. The data content and visual representation should be as similar as possible. Color schemes and themes do not have to be exactly the same.

You may contact the author of the original Flash Paper; if you do, cite and describe any information you receive from them.

(If you are contacted for information on reproducing figures from your own Flash Paper, you may provide it, but you are not obligated to respond or provide any help.)

Part B

Problems 1–3 use the U.S. Transgender Population Health Survey (TransPop) originally available from https://www.icpsr.umich.edu/web/ICPSR/studies/37938. Use `load()` to import the saved R environment containing the `da37938.0001` data frame. The original dataset includes samples of both transgender and cisgender individuals (not included). The Codebook and User guide describing the dataset are included in the zipped files. You may ignore the sample weights in this homework.

Problem 3

The survey includes several validated scales for measuring constructs related to identity, stress, and health. We would like to use these scales to build a model for predicting satisfaction with life among trans people. Focus your analysis on the following numeric variables described on pages 26-35 of the User Guide:

- Satisfaction with life
- Social well-being
- Non-affirmation of gender identity
- Non-disclosure of gender identity
- Healthcare stereotype threat
- Mental distress/disorder
- Everyday discrimination

Reproducibly partition the full dataset into a training and test sets using a 50/50 split. Then, using the imputed versions of the above variables, identify the best **single** predictor (among these scales) for life satisfaction based on the model's RMSE on the test set.

Problem 4

Using a stepwise model selection strategy (showing and explaining each step), add additional predictors to the model from Problem 3 up to a total of 3 predictors at most. State what are the best 3 predictors (among these scales) for life satisfaction for trans people.

Problem 5

Using only these scales, is it reasonable to build a model for predicting life satisfaction with more than 3 predictors? Why or why not?