Homework 3

Overview

Objectives

- Identify nominal versus ordinal attributes
- Practice loading a dataset and using ggplot to graph data
- Consider general data mining tasks

Grading

- Uploaded *all* requested files, 5%
- R markdown (and compiled pdf) are properly formatted, 5%
- Part A., 30%
- Part B., 10%
- Part C., 50%

Deliverables

- .pdf file
- .rmd file used to generate the pdf file

Setup

Create a new R Markdown file with the title "Homework 3" and with you as the author (hint: this information should be in your frontmatter at the top of the file). In your R Markdown file, create a section heading for each of the following parts of your homework (see this documentation for how to make a heading):

- Part A
- Part B
- Part C

When you knit your markdown document, these headings should be actual headings (like they are in this document).

Part A. Attribute descriptions

You are given a data file (see assignment attachments on blackboard): `hw03data.csv`. This is a table of ratings for teaching assistants (TA) at some university (this is modified data that originally came from https://archive.ics.uci.edu/ml/datasets.php). There are 6 attributes:

- `id`: unique number to identify each rating
- `instructor`: id number that identifies the instructor
- `classNbr`: id number that identifies the course
- `semester`: 1=summer, 2=fall, 3=winter
- `classSize`: number of students in the class
- `rating`: 1=low, 2=medium, 3=high

For each attribute, describe whether it is nominal or ordinal.

1. `id`
2. `instructor`
3. `classNbr`
4. `semester`
5. `classSize`
6. `rating`

Part B. R practice

Load the tidyverse collection of packages and load `hw03data.csv` in R (as a dataframe or tibble).

Using ggplot2, graph the relationship between class size and rating. The type of graph is up to you (but must be appropriate). Explicitly label your graph axes (e.g., using the `labs` function).

Part C. Thinking about common data mining applications

1. This question asks you to think about clustering. Considering an average kitchen, identify 4 clusters of objects. Name them according to their utility. For example, if I were asked to identify 4 clusters in a sporting goods store, I might choose apparel (e.g., shoes, shirts), containers (e.g., backpacks, coolers), sports equipment (e.g., basketball, tennis racquets), and fishing (e.g., rods, lures).
2. Pick one of the following forms of predictive modeling: regression, classification, or ranking. Describe an application domain where you think it would be useful to apply the chosen predictive modeling approach. What attribute(s) would you be predicting? What attributes would you use to make the prediction?