Homework 4

Overview

This assignment may be completed with your group. Each group will submit one copy of their homework responses to blackboard.

You will turn in both the .Rmd file and the knitted .pdf file to blackboard. I have provided a template R markdown file for you to use (on blackboard).

Objectives

This goal of this homework is as follows:

- Practice applying applying data preprocessing techniques using R

Grading

- Uploaded *all* requested files, 5%
- Files are properly/clearly formatted, 5%
  - Proper section headers for each part of your homework.
  - You clearly indicate which question each of your responses are associated with.
- All questions (weighted evenly): 90%

Deliverables

- .Rmd file used to knit your .pdf file
- .pdf file knitted from your R markdown file

I strongly recommend that you read over your knitted .pdf file before submitting your homework to make sure that it is formatted as you expect.

Setup

Download the R markdown template provided to you (on blackboard), and answer all of the questions below. The template is meant to help you organize/format your responses. You will need to add explanatory text and code chunks as necessary (be sure to update your names!).

Part A - Calculating basic statistics

Suppose that the data for an analysis includes the attribute `age`. The following questions ask that you use R to compute some basic properties of the given ages. **For full credit, you *must* write R code to perform any calculations.**

```
age_data <- data.frame(
  age=c(13, 15, 16, 19, 20, 20, 21, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 36, 40, 45, 46, 52, 70)
)
```

1. What is the mean, median, and mode of the ages given by the `age` attribute of `age_data`?

2. Use min-max normalization to transform the age values onto the range [0.0,1.0].; Print out the result.

3. Using the original `age` data, what is the Z-score normalized values for the following ages:

- 19
- 30
- 52
- 70

4. Identify the first, second, and third quartiles (using R).

5. Calculate the interquartile range.

6. Using the interquartile range, identify any outliers.

7. Using `ggplot`, create a boxplot of the ages.

8. Calculate the entropy of ages. Use log base 2. (Hint: if you use a function to calculate entropy, read the function's documentation to see how to configure the units)

Part B - Processing customer data

Load the `movie.csv` data (download from blackboard) into R. These data contain customer records for a fictional movie rental store. The dataset specifies the following attributes for each customer: - `CustomerID` - `Income` - `Age` - `Rentals` (total number of video rentals in the past year) - `AvgPerVisit` - `Incidentals` (whether the customer tends to buy incidental items at the checkout counter) - `Genre` (the customer's preferred movie genre)

For full credit, you *must* write R code to perform any data analysis.

1. For each of the attributes in `movie.csv`, identify whether the attribute is ordinal or nominal. (no R code needed for this question)

2. Use ggplot to graph the distribution of favorite movie genres among customers. What is the most popular genre?

3. Identify which customers (by CustomerID) have missing values in the `Incidentals` attribute.

4. What is the most frequent value for the `Incidentals` attribute?

5. Replace all missing values in the `Incidentals` attribute with the mode.

6. Create a new column in the dataset called `age_group` that discretizes the age attribute based on the following categories:

- youth: 0 to <= 25
- adult: 26 to <= 64
- senior: > 64

There are a number of ways to implement this in R. If you're not sure where to start, the `mutate` function (in the `dplyr` package), `sapply`, or `mapply` functions are good places to start. Here's an example of using the `mapply` function to add a new column to the dataset.

```
# Create a function to determine category/value based on one or more other values
recommend_diehard <- function(fav_genre) {
  if (fav_genre == "Action") {
    return(TRUE)
  } else {
    return(FALSE)
  }
}
data$RecommendDieHard <- mapply(
  data$Genre,
  recommend_diehard
)
```

Use the R documentation (e.g., `?`) to read more about how the `sapply` or `mapply` functions work.

7. Use ggplot to graph a histogram of favorite movie genres by age group (youth, adult, senior). Use the `facet_wrap` function to have ggplot create 3 histograms, one for each age group.