CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Fall 2021: Homework 1 (10 points)

Due date: Sun, September 12 2021, 11:59:59 PM Chicago Time

Please read all of the parts of the homework carefully before attempting any question. If you detect any ambiguities in the instructions, please let me know right away instead of waiting until after the homework has been graded.

1. Exercises (4 points)

1.1 Tan, Chapter 1 (2 points divided evenly among the questions)

Besides the lecture, make sure you read Chapter 1. After doing so, answer the following questions at the end of the chapter: 1, 3.

1.2 Tan, Chapter 2 (2 points divided evenly among the questions)

Besides the lecture, make sure you read Chapter 2, sections 2.1 – 2.3. After doing so, answer the following questions at the end of the chapter: 2, 3, 7, 12.

2. Programming Problem

Please label your answers clearly, see Template.Rmd R notebook for an example (Template.Rmd R notebook is available in "Blackboard → Assignment and Projects → Homework 0". Rename Template.Rmd to firstname.lastname.Rmd.)

Each answer must be preceded by the R markdown as shown in the Homework 0 R notebook (### Part 2.1-A-ii, for example). Failure to clearly label the answers in the submitted R notebook will lead to a loss of 2 points. Round up all decimal numbers to two significant digits.

2.1 Exploratory data analysis (6 points divided evenly among the constituent parts)

This exercise relates to the College data set, which can be found in the file College.csv. It contains 19 attributes (or dimensions) for 777 different universities and colleges in the US. The variables are

- Name: Name of the college or university
- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled

- Top10perc : New students from top 10 % of high school class
- Top25perc : New students from top 25 % of high school class
- F.Undergrad : Number of full-time undergraduates
- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

Feel free to use any R package that you think will make benefit you. For instance, the dplyr() package can be used to answer some of the questions below.

(a) Read the data from College.csv into a data frame. Consider the first column to be the "row name" and read it in as such. (Hint: See the row.names parameter to read.csv() method.). Print out the first five rows and the following columns from the data frame: Columns 1, 5, 8, 10.

(b) Count (without looping) the number of private and public colleges in the dataset. Your code should be structured as follows:

```
private <- ...
public <- ...
```

Your output must be of the following form:

```
There are 565 private colleges, and 121 public colleges in the dataset
```

(c) Create a new data frame that contains all of the rows of the original data frame but only the following columns: Private, Apps, Accept, Enroll, PhD, perc.alumni, S.F.Ratio, and Grad.Rate. Print the top six observations of the new data frame.

(d) For the data frame created in (c), prepare two histograms as described below. Each histogram should have a labeled X- and Y-axes, and be appropriately titled. Graph the histograms using the hist() function in base R or the ggplot() package. The ggplot() package has a reasonably steep learning curve, so if you have not used it before, you should continue with hist(). If you want to challenge yourself and learn ggplot(), then this may be the opportunity.
(i) Prepare a histogram of PhD holders in private colleges.
(ii) Prepare a histogram of PhD holders in public colleges.
(iii) **Extra credit (1 point)**: Use color to make your graph as attractive as possible, i.e., color each histogram bin. Figure out the bins programmatically. Hint: First create a histogram, but do not plot it (see the help or manual page for hist()). Examine the histogram object and figure how which member field will contains the information for the bins. Then, use rainbow() to create a color palette.

(e) From the data frame created in (c):
(i) print the top 5 colleges that have the **minimum** graduation rates
(ii) print the top 5 colleges that have the **maximum** graduation rates

(Hint: To sort the data frame on certain attributes, see https://www.statmethods.net/management/sorting.html)

(f) Install package psych.
(i) From that package, using the pairs.panel() function, draw a correlation plot for the data frame you created in (c) using the following attributes: PhD, S.F.Ratio, and Grad.Rate. Based on the correlation plot, answer the following questions:
(ii) Which two attributes have the highest correlation? Explain the reason behind this positive correlation (i.e., does this correlation make sense? Why?).
(iii) Which two attributes exhibit the lowest correlation? Explain the reason behind this negative correlation (i.e., does this correlation make sense? Why?).

(g) Using the data frame in (a), and the boxplot() function, produce **side-by-side** boxplots that help answers the following question: Which alumni donate more to their colleges --- those who go to public schools or those who go to private schools?

To do this, you will use the boxplot() function, but you will group the attribute you are interested in by a control group. The control group here is the attribute Private (which is 'Yes' or 'No'), and the attribute of interest here is perc.alumni. The format is boxplot(x, data=), where x is a formula and data= denotes the data frame providing the data. An example of a formula is y~group where a separate boxplot for numeric variable y is generated for each value of group. Label the X- and Y-axes appropriately and provide a main= parameter to the boxplot() command for a graph title. You should see two boxplots if all works.

As an added resource, take a look at https://www.statmethods.net/graphs/boxplot.html

(h) Using the data frame in (a), create a cumulative distribution function for the attribute Expend. (Hint: see ecdf()).
(iii) Plot the ecdf. Put a grid on the plot after plotting it by issuing the grid() command. This will allow you to answer the questions below better. Using this graph, answer the following questions:
(i) What is the median expenditure per student?
(ii) 80% of the students pay less than how many dollars?