CS 422: Data Mining

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

Fall 2024: Homework 1
(10 points)

Due date: Tuesday, September 03 2024, 11:59:59 PM Chicago time.

1 Exercises (1 points)

1.1 Describe in 2-3 sentences what you expect to achieve from this course. Add this description as a PDF file into the ZIP archive that you will upload.

2 Practicum problems (9 points)

2.1 Problem 1

This exercise will ensure that you are comfortable with the mechanics of creating and submitting homeworks in Canvas for the autograder.

In the Homework 1 page on Canvas, you will see a file called template.r. Use this file as a template for your homeworks. Rename this file to a new R script called *firstname-lastname.r*, where *firstname* is your first name, and *lastname* is your family name (surname). Interspersed in this file are various "TODO:" lines, these are lines where you will add your R code.

Load the *cars* dataset; the *cars* dataset is a built-in dataset in R and gives the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s. To access it, simply type "attach(cars)" in your program.

1(a) Write a function to return the following rows (all columns) from the cars dataframe: 1, 5, 8, 20.

1(b) Write a function to plot the data in the *cars* dataset as a scatterplot. On the X-axis is the variable *speed* and on the Y-axis is the variable *dist*. The plot should label the X-axis as "speed (mph)" and the Y-axis as "distance (ft)". The plot should be entitled "scatterplot of the cars dataset". (The title of the plot and labels on the x- and y-axis **must** be lowercased.)

1(c) Write a function to plot the data in the *cars* dataset as a line graph. On the X-axis is the variable *speed* and on the Y-axis is the variable *dist*. The plot should label the X-axis as "speed (mph)" and the Y-axis as "distance (ft)". The plot should be entitled "line graph of the cars dataset". (The title of the plot and labels on the x- and y-axis **must** be lowercased.)

1(d) Write a function to print a summary of the *cars* dataset using the R *summary()* method. The function should return the summary.

1(e) Write a function to return a list containing two elements: the maximum speed and the minimum distance as shown in the *summary()* command.

1(f) Write a function to calculate of ratio of the speed to the distance. Using R vectorization, write one line of code that will calculate this ratio and save it in a vector called ratio. **Do not use for loops!**

Add this vector as a new column to the cars dataframe, and return the dataframe. (From the resources I gave you to learn R, you should have encountered how to add a new column to an existing dataframe. One line of code will do this.)