CS 4232/5232 – Fall 2023 – Homework 1
Assigned: 09/20/2023
Due: 09/27/2023 at 11:59 p.m.
Maximum Grade: 100 pts.

Objectives: The objectives of this homework are the following:

- Setting up R and RStudio.
- Learn how to write matrices and do matrix multiplication in R
- Learn how to write functions in R.
- Learn how to use the DPlyr package.
- Learn how to plot data.

Instructions:

- This homework is to be done individually.
- Submit **a single HTML file** named: yourlastname_hw1.html **and also the source R markdown file** (with a suffix of .Rmd) with the answers to your homework. To generate this HTML file, you **must use R markdown**. Check the video here for more information https://rmarkdown.rstudio.com/lesson-1.html.
- See the slides posted on Canvas to learn how to install R and RStudio and to learn basic R commands.

Activity 1: (2 pts.) (Software Installation) Do the following tasks:

i. Download R from r-project.org
ii. Download RStudio from rstudio.com
iii. Install these packages: tidyverse, arules, knitr, caret, markdown, reticulate.

If you are using RStudio Server, you might encounter some error messages when trying to install a new package. If, for example, you are trying to install the tidyverse package and you get an error message, then try the following command:

```
install.packages("tidyverse", dependencies = TRUE, IN-
STALL_opts = '--no-lock')
```

After installing these packages, write in your markdown the output of the following R commands (type them in markdown, not in the terminal)

```
inst.packages <- installed.packages()[,1]
all(is.element(c("tidyverse", "arules" , "caret", "markdown", "reticulate",),
inst.packages))
```

The output must be *True*.

Activity 2: (20 pts.) Reading datasets with R

a) Download the adult dataset (https://archive.ics.uci.edu/ml/datasets/adult) from UCI.
b) (1 pt.) Open the *adult.data* containing the training data using a plain text editor, not Word. Notice that the first line in the file does not contain the names of the columns and that your file has the character "?" in some lines, meaning that there are missing

data. Also, notice that all values have a white space in front of them. These observations need to be explicitly addressed in reading our file. What is one line in this file that has missing data? Print it with R.

c) (5 pts.) Read the file *adult.data* containing the training data into the variable 'adult'. For this, you must use the function `read_csv` of the readr package. You cannot use `read.csv`. You must provide the argument `col_names = FALSE`, otherwise, this function assumes that the first line contains the names of the features, which is not the case in our dataset. You must also specify the argument `trim_ws = TRUE` and `na = c('?')`, instructing R to remove the white space around each attribute value in the file and to treat the question mark as an NA.

d) (2 pts.) Print the data frame that results from reading the file. Notice that R correctly treats '?' as NAs. Also notice that the columns have names like "X1" or "X2", which are not acceptable.

e) (5 pts.) Read the file *adult.data* again, but this time, provide a value for the argument col_names = c('age', 'workclass', …[the other names are in the adult.names file]).

f) (2 pts.) Print the data frame that results from reading the file. Notice that now the attributes have the correct names. Also notice that under some of the features it says "chr" or "dbl", meaning that they are character strings or numbers, respectively. We know that almost all attributes, like State-gov, must be factors instead.

g) (5 pts.) Turn these attributes into factors by doing:
adult$workclass <- as.factor(adult$workclass)
and similar commands for the others.

<u>Activity 3</u>: (33 pts.) (Basic R Commands) Using R, perform the following tasks:

a) (4 pts.) Write a function in R that given a vector computes the standard deviation of the elements of that vector. You cannot use the sd function of R or any other function that computes it already.

b) (3 pts.) Use your function to compute the standard deviation of ages of all adults in the pre-processed dataset of Activity 2. Print this value.

c) (4 pts.) Write and execute the command to make a 32561x2 matrix having as its first column all 1s and as its second column the column of the ages in the adult dataset. The first three rows of this matrix are the following:

$$H = \begin{matrix} 1 & 39 \\ 1 & 50 \\ 1 & 38 \\ \cdots & \cdots \end{matrix}$$

d) (2 pts.) Write and execute the command to find H^t, the transpose of the matrix H of task c).

e) (2 pts.) Write and execute the command to multiply the two matrices H^t x H and print the result. Notice that this asks for matrix multiplication, not component-wide multiplication. The resulting matrix must be of size 2x2.

f) (2 pts.) Write and execute the command to find the inverse of the matrix H^t x H.

g) (2 pts.) Write and execute the command to verify that the matrix found in part f is indeed the inverse of H^t x H.

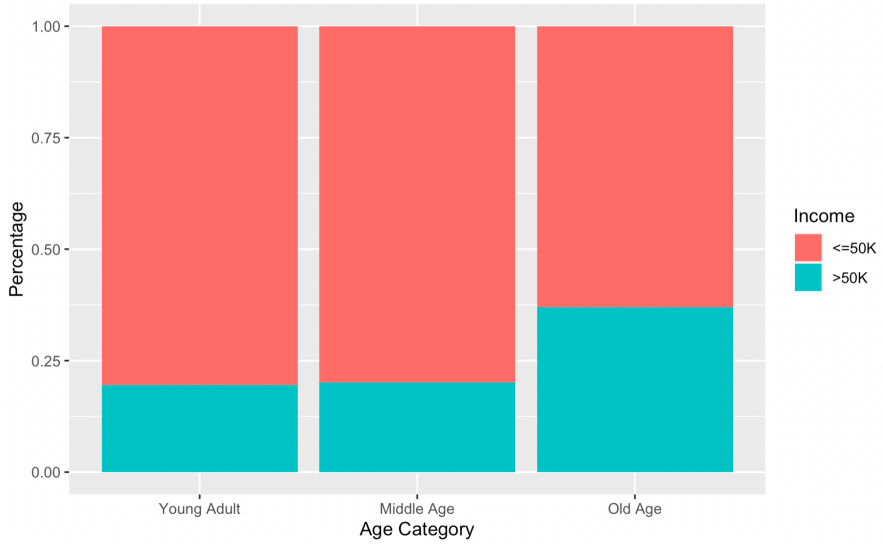
h) (2 pts.) Without using ggplot, write and execute the code to make a barplot having the response attribute (the one with values '<=50K' and '>50K') on the x-axis and the counts of the number of people with that income on the y-axis. Your plot needs to have properly labeled axes and a title.
i) (4 pts.) Repeat exercise f) using ggplot.
j) (4 pts.) Write code to replace all the "Never-worked" values with "Unemployed". You cannot use dplyr.
k) (4 pts.) Repeat exercise h) but using dplyr.

Activity 4: (36 pts.) (DPlyr) Using the adult dataset that results from the pre-processing done in Activity 2, perform the following tasks:

a) (4 pts.) Using DPlyr, write the code to select all the rows of the dataset corresponding to individuals who never married and are younger than 45. Then run it and print the top 5 rows.
b) (4 pts.) Repeat task a) but using only base R, i.e., without using DPlyr or similar packages.
c) (4 pts.) Using DPlyr, write the code to print the values of only the age, workclass, and marital-status columns. Run it to print the top 5 rows.
d) (4 pts.) Repeat task c) but using only base R, i.e., without using DPlyr or similar packages.
e) (4 pts.) Using DPlyr, write the code to add to the adult dataset the column 'Age Category', containing 'Young Adult' for those rows with age < 45, 'Middle Age' for those rows with 45 <= age <= 65, and 'Old Age' otherwise. `case_when` will be useful here.
f) (4 pts.) Repeat task e) but using only base R, i.e., without using DPlyr or similar packages.
g) (4 pts.) Using DPlyr, write code to calculate the standard deviation of the age attribute.
h) (4 pts.) Using DPlyr, select all rows that have NA in the workclass attribute.
i) (4 pts.) Repeat task h) but using only base R, i.e., without using DPlyr or similar packages.

Activity 5: (9 pts.) (Plotting) Using the adult dataset that results from the pre-processing done in Activity 2, do the following tasks:

a) (5 pts.) Write R code to replicate the figure below.



b) (4 pts.) What conclusions do you draw about the dataset from the figure above?

To help you with these exercises, check the Introduction to R slides that I posted on Canvas. Also, check the R cheat sheets posted under the "Tutorials and Installation Instructions for R/R Cheat Sheets." These are very nice-looking files containing a summary of all the commands of each of the packages above. I highly recommend them.

Important Note: As I mentioned in the first lab, **you cannot have all your code in a single R markdown chunk**. You need to break down your markdown file into meaningful chunks. For example, one chunk can be a function definition, another chunk a function call, another function the definition of a set of related variables, etc.