# Homework 2

(Advanced) Data Mining: Algorithms and Applications-Winter 2023

**Due on Feb 13, 11.59PM**

Datasets for this homework can be found at the link below (**umich email only**) and feel free to add this folder to your Google Drive:
`https://drive.google.com/drive/folders/1ehWwunuAo7CE1Vk2JYkUnQMmxh5pph3C?usp=sharing`

- Only submissions through Canvas will be accepted.
- Please don't take screenshots of your codes.
- Please submit your HW2.R or HW2.py along with the document (HW2.pdf) which would include your calculations. In other words, you will be submitting two different files. Please do NOT zip.
- Undergraduate students are responsible for the first four questions.
- Graduate students are responsible for all six questions.

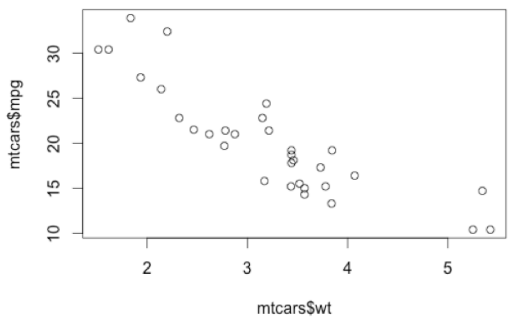
1. Find the distance between objects 1 and 3 by using the formula provided on the slides. Notice that we have mixed type of attributes. (You can scan and submit your handwritten calculation) (25/20 points)

| Object Identifier | test-1(nominal) | test-2 (ordinal) | test-3 (numeric) |
|---|---|---|---|
| 1 | A | excellent | 45 |
| 2 | B | fair | 22 |
| 3 | C | good | 64 |
| 4 | A | excellent | 28 |

2. Write a program in any language which can compute Manhattan and Euclidean distances between any two given vectors with any length. You can pass the length to your function, but please don't limit the dimension to 2. You can test your function on vectors you fill in your code without asking user input. (25/20 points)
3. In the table below, determine whether passing a class has a dependency on attendance by using Chi-square test. Please refer to the formula in the slides. (25/20 points)
   (For the expected value for each cell, multiply the total counts in the rows and columns of the cell and divide by total count.
   For example: Expected value for Attended-Pass=33*31/54 = 18.94. You can scan and submit your handwritten calculation)

| | Passed | Failed | **Total** |
|---|---|---|---|
| Attended | 25 | 6 | **31** |
| Skipped | 8 | 15 | **23** |
| **Total** | **33** | **21** | **54** |

4. In R, there is a built-in data frame called `mtcars`. Please calculate the correlation between mpg and wt attributes of mtcars by using `cor()` function. Then generate scatter plot based on these two attributes. Your scatter plot should be like the one below. You don't need to submit the image, but R script should be submitted (25/20 points)




5. **Grad Students Only** Write an R or Python script which removes or drops the columns which have more than 75% missing values. Then it should replace the missing values in the remaining columns with the median value of the

existing values of that particular column. Download metabolite.csv from Google Drive and use this data set to test your code. Please check the end of this document for some useful R examples and hints. (10 points)

6. **Grad Students Only** Please apply PCA on the processed metabolites data and create a scatter plot by using first two principal components in which points are colored based on the Label column. Please submit your code along with your figure in the same file. (10 points)

(If you are going to use R, you may need to use which(), is.na() functions and consider excluding those columns by name. For that purpose you may investigate %in% and -c(...) type of operations. You can also see examples of subsetting a dataframe below with their outputs. It's also recommended to check `tidyverse` library.) (10 points)

```
# A sample data frame
data <- read.table(header=T, text='
 subject sex size
       1   M    7
       2   F    6
       3   F    9
       4   M   11
 ')


subset(data, subject < 3)
#>   subject sex size
#> 1       1   M    7
#> 2       2   F    6
data[data$subject < 3, ]
#>   subject sex size
#> 1       1   M    7
#> 2       2   F    6


# Subset of particular rows and columns
subset(data, subject < 3, select = -subject)
#>   sex size
#> 1   M    7
#> 2   F    6
subset(data, subject < 3, select = c(sex,size))
#>   sex size
#> 1   M    7
#> 2   F    6
subset(data, subject < 3, select = sex:size)
#>   sex size
#> 1   M    7
#> 2   F    6
data[data$subject < 3, c("sex","size")]
#>   sex size
#> 1   M    7
#> 2   F    6


# Logical AND of two conditions
subset(data, subject < 3  &  sex=="M")
#>   subject sex size
#> 1       1   M    7
data[data$subject < 3  &  data$sex=="M", ]
#>   subject sex size
#> 1       1   M    7


# Logical OR of two conditions
subset(data, subject < 3  |  sex=="M")
```

```
#>   subject sex size
#> 1       1   M    7
#> 2       2   F    6
#> 4       4   M   11
data[data$subject < 3  |  data$sex=="M", ]
#>   subject sex size
#> 1       1   M    7
#> 2       2   F    6
#> 4       4   M   11


# Condition based on transformed data
subset(data, log2(size) > 3 )
#>   subject sex size
#> 3       3   F    9
#> 4       4   M   11
data[log2(data$size) > 3, ]
#>   subject sex size
#> 3       3   F    9
#> 4       4   M   11


# Subset if elements are in another vector
subset(data, subject %in% c(1,3))
#>   subject sex size
#> 1       1   M    7
#> 3       3   F    9
data[data$subject %in% c(1,3), ]
#>   subject sex size
#> 1       1   M    7
#> 3       3   F    9
```