GE2324 Assignment 2

Due: 24-Mar-2022 (Week10 Thursday) 20:00

Each 24 hours late submission halves the score. 0 marks after 3 days.

Submission:

Upload your answers to Canvas under Assignments. Your submission should be one single file in the format of .pdf. Your answers could be created by handwriting, as well as typing and tables using computer tools.

Source of Answers:

Students should keep the sources (e.g. files of .doc / .docx / .xls, papers of handwriting and drawing etc.), that would be required to upload during this semester when we perform random sampling for further investigation.

Question 1 [55 marks].

Index	Red	Green	Blue
1	93	123	144
2	89	89	94
3	103	136	156
4	115	147	166
5	135	159	174
6	153	164	172
7	162	173	182
8	188	188	193
9	218	215	217
10	142	147	155
11	48	49	52
12	66	67	71
13	33	31	33
14	125	121	123
15	12	11	12
16	20	19	20

K-means clustering is useful in computer and engineering fields. For instance, in computer graphics, K-means clustering could be used to reduce the number of different colors in a stored image. In this question, we'll consider the color image on the top, which initially contains 16 different colors.

Each used color is represented as a tuple of 3 numbers (range 0..255), for its Red, Green and Blue components respectively. 0 represent the total absence of such color, while 255 represent 100% of such color. For instance, yellow, which is formed by mixing 100% red, 100% green and 0% blue is stored as {Red: 255, Green: 255, Blue: 0}. The 16 used colors of the image are tabulated (see above table).

Now we want to use K-means clustering to reduce the number of colors from 16 to 3. Derive the Red/Green/Blue values for the 3 colors **with intermediate steps** (0 marks if no steps are given). Apply **Euclidean distance** when deriving distances between colors and centroids. Also, **use real numbers** for calculation and answers *(3 decimal places)* despite that the original numbers are integers.

a) Use color indexes 2, 5 and 8 from the table above as the initial centroids.
b) Then rework on the clustering all over again (with intermediate steps), this time use color indexes 2, 3 and 4 respectively as the initial centroids. Do you get the same result?

Question 2 [45 marks].

In this question, we'll consider the information regarding a number of Ginseng species.

Age *(years)*	Length *(cm)*	Water content percentage	Price *(per 100g)*
5	10	10	150
12	25	10	180
6	10	8	200
9	30	12	200
8	15	12	220
7	15	13	250
20	42	18	300
18	40	21	350
36	120	25	400

Answer the following questions with intermediate steps and tables shown whenever appropriate. *Note that 0 marks will be given if no intermediate steps/tables are provided.*

a) Using the formula of Pearson correlation, derive whether correlation exist between Length and Water content percentage (if yes, in what way).

b) Which attribute is a better indicator / predictor of price? Age or Water content? Justify your answer using Spearman Rank correlation.

c) Calculate the Kendall's Tau between Length and Price.

- END -