GE2324 Assignment 2, 2024/25B

Due: **31-Mar-2025** (Week 11 Monday) 20:00
Each 24 hours late submission halves the score.
(That is, no deduction of marks within the first 24 hours). 0 marks after 3 days.

Question 1 [55 marks].

	R	G	B	INDEX
	8	52	73	1
	168	198	217	2
	68	115	141	3
	4	21	10	4
	187	186	166	5
	239	239	243	6
	57	77	67	7
	138	170	192	8
	107	145	165	9
	191	212	227	10
	30	48	34	11
	139	144	122	12
	215	210	192	13
	165	167	148	14
	28	85	119	15
	91	108	94	16

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 (1 decimal place) despite that the original numbers are integers.

a) Use color indexes 5, 9 and 13 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 1, 2 and 3 respectively as the initial centroids. Do you get the same result?

Question 2 [45 marks].

In this question, we'll consider a dataset related to car attributes and their prices.

Car Age (years)	**Distance Driven (thousands of km)**	**Number of Previous Owners**	**Price (in thousands of dollars)**
5	50	1	20
3	30	2	25
8	80	3	15
2	20	1	30
6	60	2	18
4	40	1	22
7	70	3	16
1	10	1	35
9	90	4	12

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 exists between Distance Driven and Number of Previous Owners (if yes, in what way).

b) Which attribute is a better indicator/predictor of price? Car Age or Number of Previous Owners? Justify your answer using Spearman Rank correlation.

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