Assignment

In this assignment, we will analyze diamond prices from using a dataset from Kaggle. You will complete this assignment using only (generic) Python. No additional Python modulaes like Pandas, Numpy, SciPy are allowed. You will get 0 for this assignment if you use any of these modules.

For the dataset, we use diamond prices dataset from Kaggle:
`https://www.kaggle.com/shivam2503/diamonds`

Dataset Description: From the Kaggle website: "A data frame with 53940 rows and 10 variables"

1. price: price in US dollars ($326–$18,823)
2. carat: weight of the diamond (0.2–5.01)
3. cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
4. color: diamond colour, from J (worst) to D (best)
5. clarity: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
6. x: length in mm $(0 - 10.74)$
7. y: width in mm $(0 - 58.9)$

8. z: depth in mm $(0 - 31.8)$
9. depth: total depth percentage $(43 - 79)$ computed as

$$z/\text{mean}(x, y) = 2 * z/(x + y)$$

10. table: width of top of diamond relative to widest point $(43 - 95)$

You will use a subset of this dataset depending on the last digit of your BU ID as follows:

1. analyze diamonds with cut "Fair" if your BU ID ends on 0 or 1 (Group 1)
2. analyze diamonds with cut "Good" if your BU ID ends on 2 or 3 (Group 2)
3. analyze diamonds with cut "Very Good" if your BU ID ends on 4 or 5 (Group 3)
4. analyze diamonds with cut "Premium" if your BU ID ends on 6 or 7 (Group 4)
5. analyze diamonds with cut "Ideal" if your BU ID ends on 8 or 9 (Group 5)

Question 1:

1. load the "diamonds" csv file as a list of lines using Python and construct a sublist for you group
2. how many entries are there?
3. compute the average weight (in carats) for your group
4. compute the average price for diamonds in your group (round to the 4-th decimal point)

Question 2: In this question, we will investigate 2 ways to compute averages. Suppose you have N diamonds. Let w_1 denote the weight of diamond 1, w_2 denote the weight of diamond 2, ..., w_N denote the weight of diamond N. Let $p_1, \ldots, p_N$ denote the corresponding prices. Consider the following 2 methods to compute "average" prices:

method "a":

$$\mu_a = \frac{1}{N} \cdot \left[\frac{p_1}{w_1} + \frac{p_2}{w_2} + \cdots + \frac{p_N}{w_N} \right]$$

method "b":

$$\mu_b = \frac{p_1 + p_2 + \cdots + p_N}{w_1 + w_2 + \cdots w_N}$$

1. compute average prices per carat using both methods

2. which average price is lower?

3. compute the maximum price per carat:

$$\text{max price per carat} = \max(p_1/w_1, \ldots, p_N/w_N)$$

4. compute the minimum price per carat

$$\text{min price per carat} = \min(p_1/w_1, \ldots, p_N/w_N)$$

5. compute the median price per carat

$$\text{median price per carat} = \text{median}(p_1/w_1, \ldots, p_N/w_N)$$

Question 3: For each of the two methods to compute price per carat, what combination of other parameters (color, clarity, depth, etc.) gave you

1. highest value

2. lowest value

Question 4: Using your prices per carat, compute the price you will be paying for 102 carat diamond sold recently by Sotheby. How close is your price to the real price? Why?
`https://www.cnbc.com/video/2020/09/10/sothebys-will-auction-off-a-102-carat-diamond-that-could-set-a-new-record.html`