DATASET-2: Diamond Prices

Features Description:

- Price: price in US dollars
- Carat: is the diamond's physical weight measured in metric carats.
- Cut: quality of the cut
- Color: diamond color, from J (worst) to D (best)
- Clarity: a measurement of how clear the diamond is
- X: length in mm
- Y: width in mm
- Z: depth in mm
- Depth: total depth percentage = z / mean(x, y) = 2 * z / (x + y)

Table: width of the top of diamond relative to widest point

Questions:

1. i) What is the shape of the dataset? (Specify rows and columns separately)
   ii) List the column names and their data types?
   ii) Delete 'index' column?

2. Describe the summary statistics, min, max, mean, standard deviation for all numeric columns?

3. List all distinct values and most frequent values in each column 'cut, 'colour' and 'clarity'?

4. Identify and describe any data quality issues or inconsistencies within the data set. What steps would you take to clean and pre-processes the data to ensure its accuracy for further analysis.

5. (i) Convert price in us dollar to rupees? (1 dollar = 80 rupees)

   (ii) Create a new column called 'color_clarity_cut' and values are color+ '_' +clarity+ '_' + cut?
   (Ex: E_ SI2_Ideal , E_ SI1_Premium)

6. Check for any outliers in all numeric columns and then analyze carefully, how they should be addressed.

7. Calculate the correlation (Using heat map) between price and all other numeric columns and list them in descending order and identify the highest and lowest correlation?

8. Draw bar plots, visualize and also indicate any insights can be obtained by taking X-axis vs Y-axis as:
    - Cut vs no.of diamonds
    - Color vs no.of diamonds
    - Clarity vs no.of diamonds

9. Draw a histogram where X-axis-> carat with interval size 0.1 and Y-aixs-> no.of diamonds? and comment on it.

10. Draw a normal probability plot on X or Y or z? Based on the shape and trend of the plot? Is any conclusion can be drawn, if yes what it is?