sDiamonds Dataset

A dataset "diamonds-m.csv" containing the prices and other attributes of almost 54,000 diamonds and 10 variables:

id	row id
price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond color, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (IF (best), VVS1, VVS2, VS1, VS2, SI1, SI2, I1 (worst))
popularity	how popular is similar diamond with these features Good, Fair Poor
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	depth from top to bottom [ideal depth = z / mean(x, y)]
table	width of top of diamond relative to widest point

More About The Dataset

The dataset contains information on prices of diamonds, as well as various attributes of diamonds, some of which are known to influence their price (in 2008 $s): the 4 Cs (carat, cut, color, and clarity), as well as some physical measurements (depth, table, x, y, and z).

Carat

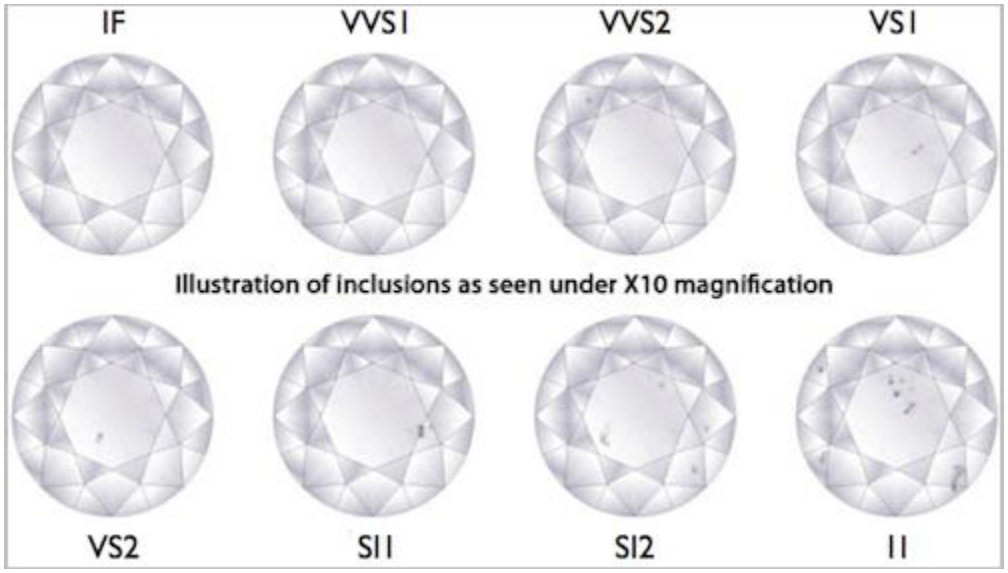
Carat is a unit of mass equal to 200 mg and is used for measuring gemstones and pearls. Cut grade is is an objective measure of a diamond's light performance, or, what we generally think of as sparkle.

Color

The figure below shows color grading of diamonds:

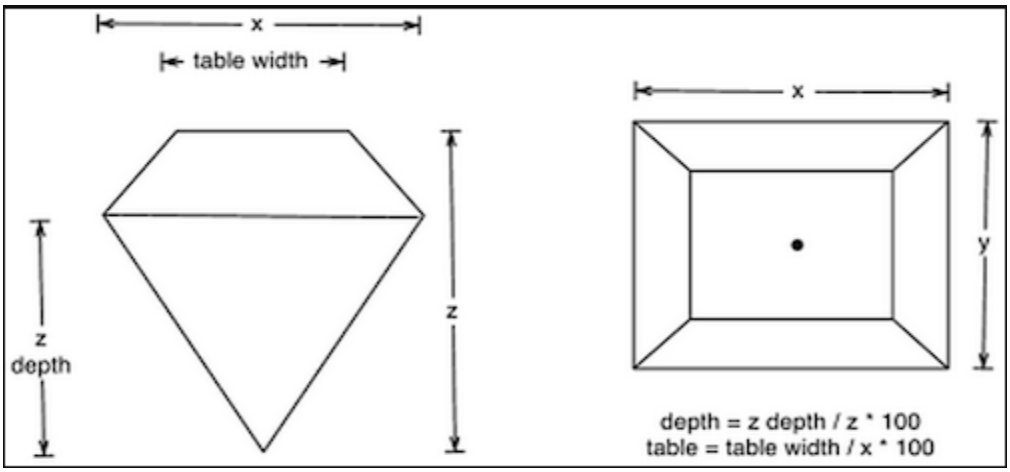
Clarity

The figure below shows clarity grading of diamonds:



Measurements

The figure below shows what these measurements (depth, table, x, y, and z) represent.



Use Python and carry out Exploratory Data Analysis (EDA) / Visual Data Analysis (VDA) on the dataset diamonds-m.csv. In this dataset the column "price" depends on all the other columns. Provide answers to following questions

1. What is structure of the dataset.
2. What are the data type of each columns?
3. What is the length of alpha-numeric columns?
4. What are precision & scale of numeric columns?
5. For each column, find out
   - Number of Null values
   - Number of zeros
   - Provide the obvious errors
   - Identify columns which should not be alpha-numeric. Provide techniques to fix the same.
6. For each numeric column
   - Replace zero values with suitable statistical value of the column. Give reason why
   - Replace null values with lower of mean & median value of the column.
   - Provide the quartile summary along with the count, mean & sum
   - Provide the range, variance and standard deviation
   - Provide the count of outliers and their value. Provide a mechanism to fix the outliers
7. For each non-numeric column
   - Replace null values with suitable statistical value of the column. Give reason why
   - provide frequency distribution table the same
8. Provide suitable mechanism to convert non numeric columns to numeric.
9. Is re-scaling required. If yes, why and what technique would you use for rescaling.
10. Provide histogram for all columns. Provide your interpretation on the same.
11. Provide box & whisker plots for all columns. Provide your interpretation on the same.

12. For numeric columns
    - provide correlation table
    - provide a suitable graph to visual the same
    - State which columns be dropped due to multi-collinearity. Give reasons.
13. Prepare relationship chart showing relation of each numeric column with column "price". Provide your interpretation on the same.
14. Based on feature selection algorithms, identify significant columns of the dataset. Give proper reason why.
15. Refer to formula of "depth" given above, compute a column "computed depth" based on formula given for each row. Identify or flag the records for which difference between "computed depth" & "depth" is greater than 5% Of "depth".
16. **Challenge: Prepare the program in such a way that for any other dataset the above queries (except point 15) are answered without any change in the program except input file name and dependent variable.**

Project Submission

1. Project to be done as per pairs assigned in your class.
2. Prepare the project using a single .py file and submit the Py file itself.
3. Each section and sub-section should be suitably marked and segregated
4. The .py file needs to be submitted using naming format
   "<roll-no>-<name>-<roll-no>-<name>.py"
   e.g. "001-CyrusLentin&032-VipulPatel.py" in Google Classroom
   Only one submission per group is required.
5. The project submission date announced separately.
6. Strict plagiarism check will be performed.
7. Any plagiarism will be severely dealt with. Both the copier and copy-from will be given 0.

Wishing You All The Best!!!

Marks Rubric

Question	Marks
1	5
2	5
3	5
4	5
5.1	5
5.2	5
5.3	5
5.4	5
6.1	5
6.2	5
6.3	5
6.4	5
6.5	5
7.1	5
7.2	5
7.3	5
8	5
9	10
10	10
11	10
12.1	5
12.2	5
12.3	5
13	15
14	10
15	15
16	30
Total	200