Diamonds 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

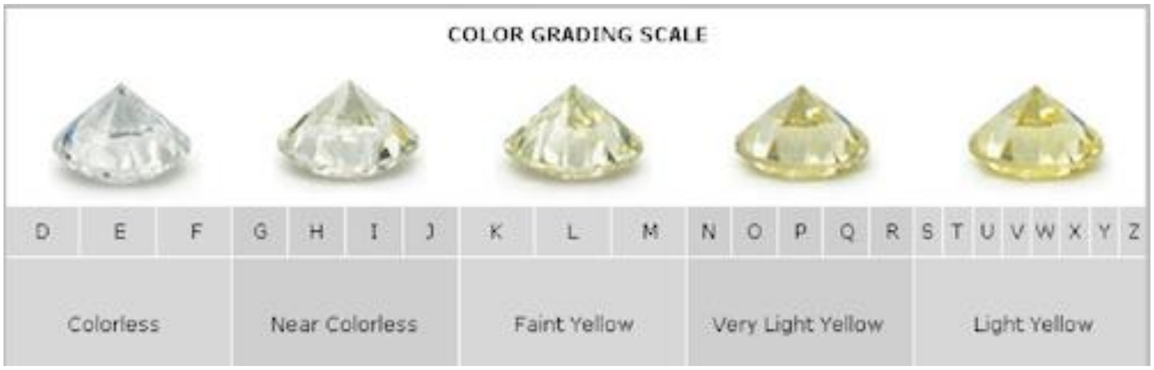
The dataset contains information on prices of diamonds, as well as various attributes of diamonds, some of which are known to influence their price (in 2008 $s): the 4 Cs (carat, cut, color, and clarity), as well as some physical measurements (depth, table, x, y, and z).

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.

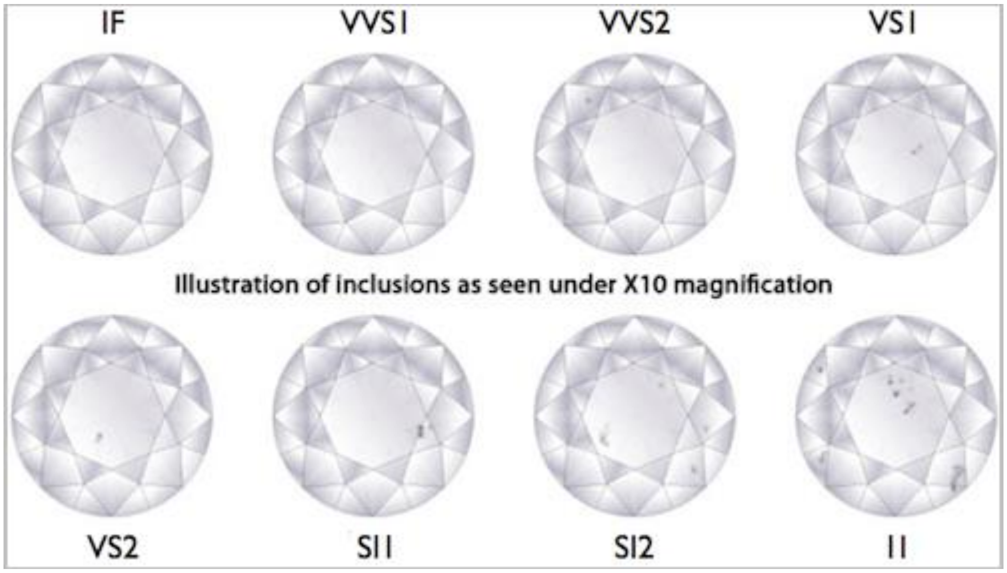
Color

The figure below shows color grading of diamonds:



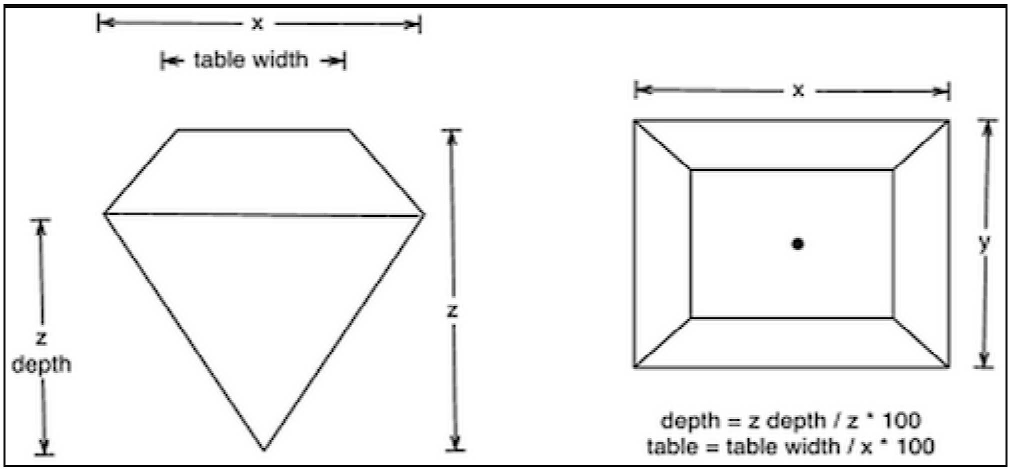
Clarity

The figure below shows clarity grading of diamonds:



Measurements

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



Ramniranjan Jhunjhunwala College
MSC DSAI
EDA / VDA Project

Use Python and carry out EDA / VDA on Iris.csv and 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. Identify significant columns of the dataset.
6. For each column, find out
   - Number of Null values
   - Number of zeros
7. For each column
   - Provide the obvious errors
8. For each numeric column
   - Replace null values with median value of the column.
9. For each numeric column
   - Replace zero values with suitable statistical value of the column. Give reason why
10. For each numeric column
    - Provide the quartile summary along with the count, mean & sum
11. For each numeric column
    - Provide the range, variance and standard deviation
12. For each numeric column
    - Provide the count of outliers and their value
13. Are there any class or categoric variables? If yes,
    - provide frequency distribution table & chart for the same
14. For all numeric columns
    - Provide histogram
15. For all numeric variables
    - Provide box & whisker plot
16. For all numeric variables
    - Provide correlation table & graph
17. Prepare relationship chart showing relation of each numeric column with all other numeric columns.
18. Find out the difference between the Actual Depth & Ideal Depth.

Challenge: Prepare the program in such a way that for any other dataset the above queries (except point 18) are answered without any change in the program except input file name.

Presentation

- The first section introduces your team.
- Each of the above query is answered on a separate section within the program.
- The last section should describe your experience of creating this project.

Project Submission

1. Project to be done as per groups assigned in your class.
2. Prepare the project using .py file.
3. The single .py files should be consolidated into a single zip file
   RJ-MSC-DS-GroupNo-GroupName.zip
   Eg RJC-MSC-DS-01-CodeMasters.zip
4. The .zip file needs to be submitted in "Challenge Project Work" assignment of Google Classroom
   Only one submission per group is required
5. The project needs to be submitted by 26-Jan-2020 end-of-day.
6. Project mentoring Zoom meeting will be set up for each group on a suitable date after submission of the project.

Wishing You All The Best!!!