Project – EDA & VDA Of Diamonds Dataset

Diamonds Dataset

A dataset “diamonds.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 (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y)
table	width of top of diamond relative to widest point

More About The Dataset

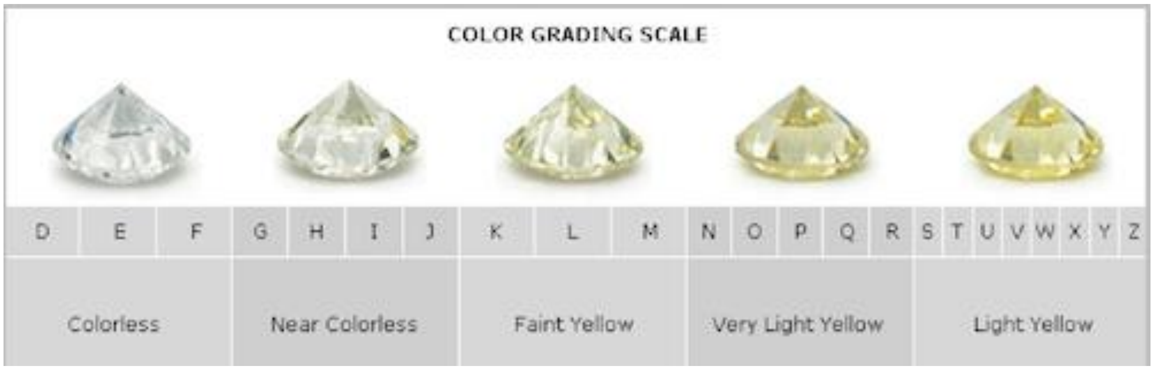
The dataset contains information on prices of diamonds, as well as various attributes of diamonds, some of which are known to influence their price (in 2008 $s): the 4 Cs (carat, cut, color, and clarity), as well as some physical measurements (depth, table, x, y, and z).

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.

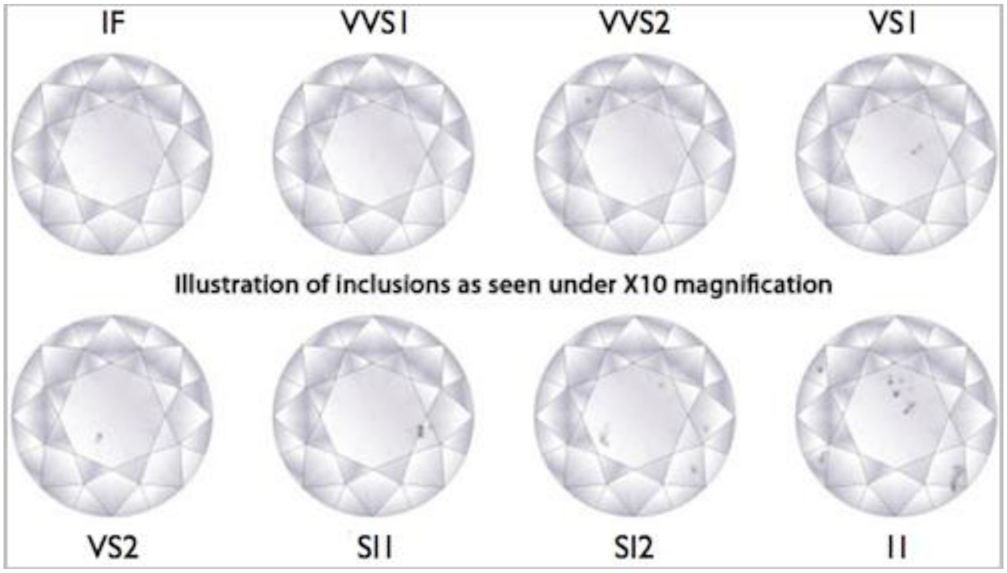
Color

The figure below shows color grading of diamonds:



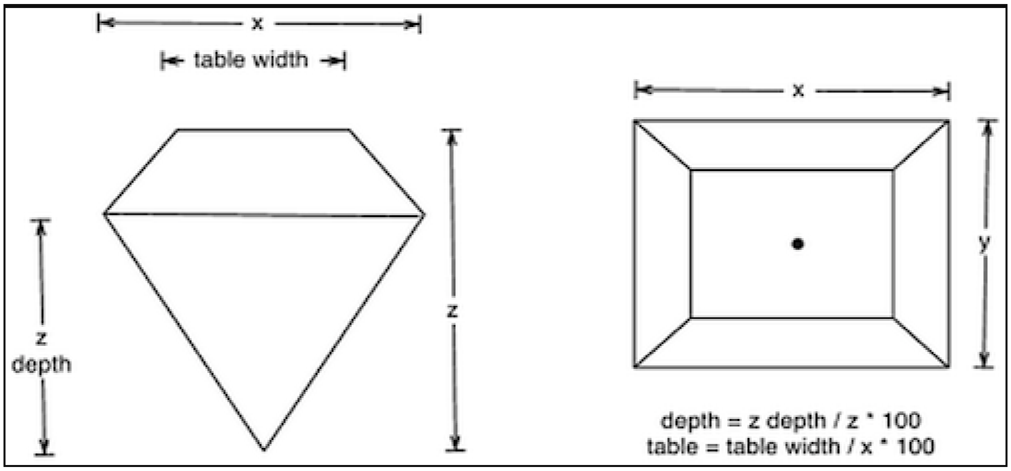
Clarity

The figure below shows clarity grading of diamonds:



Measurements

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



Project Requirements

Please provide the following in EDA, VDA, Linear Regression & Classification to provide relevant insights for the diamonds.csv

1. Read Data 5 Marks
   - Read Data
   - Show Structure
   - Basic Summary
   - Display Average Price in Crosstab with Carat & Cut
   - Refer to formula of "Depth Percentage" above, impute missing or 0 "Depth Percentage", "x", "y", "z" based on formula given.
2. Data Cleaning & Imputation 5 Marks
   - Check For Zeros In Numeric Columns; convert to Null
   - Check For Outliers in Numeric Columns; convert to Null
   - Check For Undefined Data In Categoric Columns; convert to Null
   - Check For Nulls In All Columns; get final tally of nulls in each column tomo
3. Machine Learning 1 10 Marks
   - "Price" is dependent on "Carat", "Cut", "Color" and "Clarity"
   - Impute "Price", for Null values in the column, based on suitable machine learning algorithm
4. Machine Learning 2 10 Marks
   - "Clarity" is dependent on "Price", "Depth", and "Table"
   - Impute "Clarity", for Null values in the column, based on suitable machine learning algorithm
5. Data Validation 5 Marks
   - Refer to formula of "Depth Percentage" above, compute "Computed Depth Percentage" based on formula given each row. Identify or flag the records for which difference between "Computed Depth Percentage" & "Depth" is greater than 5% Of "Depth".
6. Visual Data Analysis 5 Marks
   - Display data distribution for "Price"
   - Display relationship between "Carat" & "Price" also display trend line

   Note - For each visualization, provide reason why the graph used was chosen and the insights provided by the graph.

Project Submission

1. Project to be done in teams of up to 6 participants.
2. Prepare the project using Anaconda Spyder.
3. You may use multiple .py files if you choose.
4. All the zip files should be consolidated into a single zip file

   WeS-MIM-FinalProject-GroupNo-GroupName.zip

   Eg WeS-MIM-FinalProject-001-CodeBreakers.zip
5. The .zip file needs to be submitted via email to assignments@lentins.co.in.

   Only one email per group is required

   The email subject line should also be same as the file name
   Eg WeS-MIM-FinalProject-GroupNo-GroupName
6. The project needs to be submitted by Fri 15-May-2020 on or before 0400 pm.
7. The viva / presentation for the project will be held from 15-May-2019 0600 pm onwards.
8. Zoom meeting will be set up for each group and python related questions will be asked primarily based on the project.

Project Evaluation		**60**
Project Code	*Team Effort (same for all team members)*	*40*
Project Viva	*Per Individual (different for all team members)*	*20*

Final Evaluation		**100**
Project Submission	*As above*	*60*
Internal Assessment	*Attendance & term work*	*40*

Wishing You All The Best!!!