Diamonds Dataset

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

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

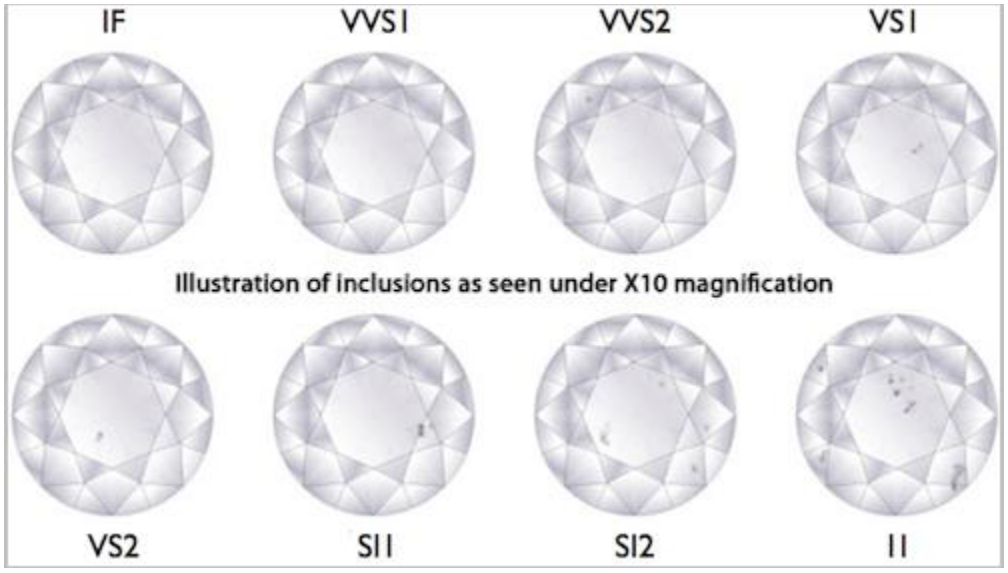
Carat is a unit of mass equal to 200 mg and is used for measuring gemstones and pearls. Cut grade is is an objective measure of a diamond's light performance, or, what we generally think of as sparkle.

Color

The figures below shows color grading of diamonds:

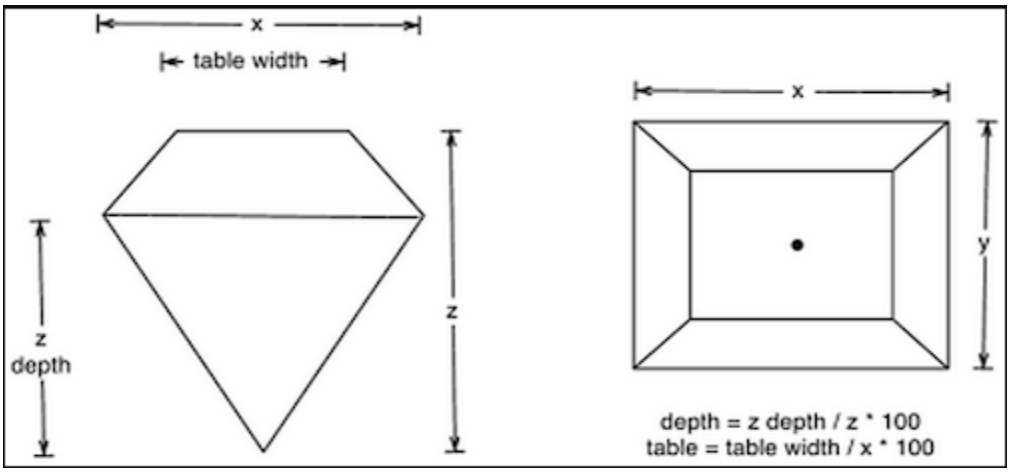
Clarity

The figure below shows clarity grading of diamonds:



Measurements

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



Project – Diamonds Dataset
EDA / VDA / Data Cleaning / Data Imputation

Project Requirements **70 Marks**

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

1. Read Data 10 Marks
   - Read Data
   - Show Structure
   - Basic Summary
2. Data Cleaning & Imputation 10 Marks
   - Check For Nulls In All Columns
   - Check For Zeros In Numeric Columns
   - Check For Outliers in Numeric Columns
   - Check For Undefined Data In Categoric Columns

   Update data with valid values if possible. If not, update invalid data with NaN. Also, Outlieres are just to be identified and no NaN updation is required for Outliers.
3. Machine Learning 1 10 Marks
   - "Price" is dependent on "Carat", "Cut", "Color" and "Clarity"
   - Impute "Price", for Null or Zero or Invalid entries, based on suitable machine learning algorithm; hint use linear regression
4. Machine Learning 2 10 Marks
   - "Cut" is dependent on "Price", "Carat", "Color" and "Clarity"
   - Impute "Cut", for Null or Blank or Invalid entries, based on suitable machine learning algorithm; hint use classification algorithm
5. Exploratory Data Analysis 10 Marks
   - Display Average Price in Crosstab with Carat & Cut
   - Display Average Price in Crosstab with Carat & Clarity
   - Display Average Price in Crosstab with Carat & Color
6. Conformation 10 Marks
   - Refer to formula of "Depth Percentage" above, impute missing or 0 "Depth Percentage", "x", "y", "z" based on formula given.
   - Refer to formula of "Depth Percentage" above, compute "Depth Percentage" based on formula given. Identify record for which "Depth Percentage" is greater than 5
7. Visual Data Analysis 10 Marks
   - Display data distribution for "Price"
   - Display data distribution for "Cut"
   - Display relationship between "Carat" & "Price" also display trend line
   - Display two more visualization which offers significant insights to the dataset.

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

Project Submission

1. Prepare the project using Spyder.
2. The project should be named as
   NMIMS-MBATech-FinalProject-GroupNo-GroupName.py
   Eg NMIMS-MBATech-FinalProject-001-CodeBreakers.py
3. The .py file(s) needs to be submitted via email to assignments@lentins.co.in.
   The email subject line should also be same as the file name
   Eg NMIMS-MBATech-FinalProject-GroupNo-GroupName
4. The project needs to be submitted by 25-Nov-2019 end-of-day.
5. The presentation / evaluation for the project will be held on 26-Nov-2019.

Project Evaluation

Project Code	*Team Effort (same for all team members)*	*40*
Project Viva	*Per Individual (different for all team members)*	*30*

Wishing You All The Best