Project – Diamonds

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	total depth percentage = z / mean(x, y)
table	width of top of diamond relative to widest point

More About The Dataset

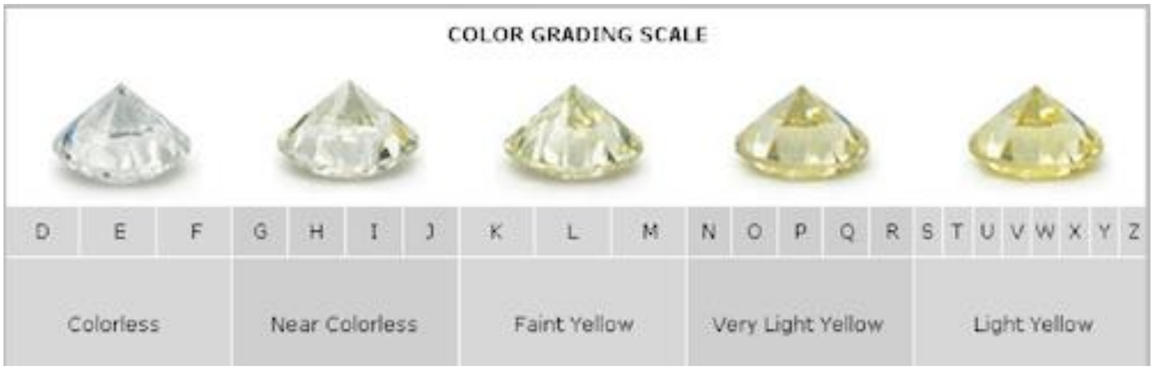
The dataset contains information on prices of diamonds, as well as various attributes of diamonds, some of which are known to influence their price (in 2008 $s): the 4 Cs (carat, cut, color, and clarity), as well as some physical measurements (depth, table, x, y, and z).

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.

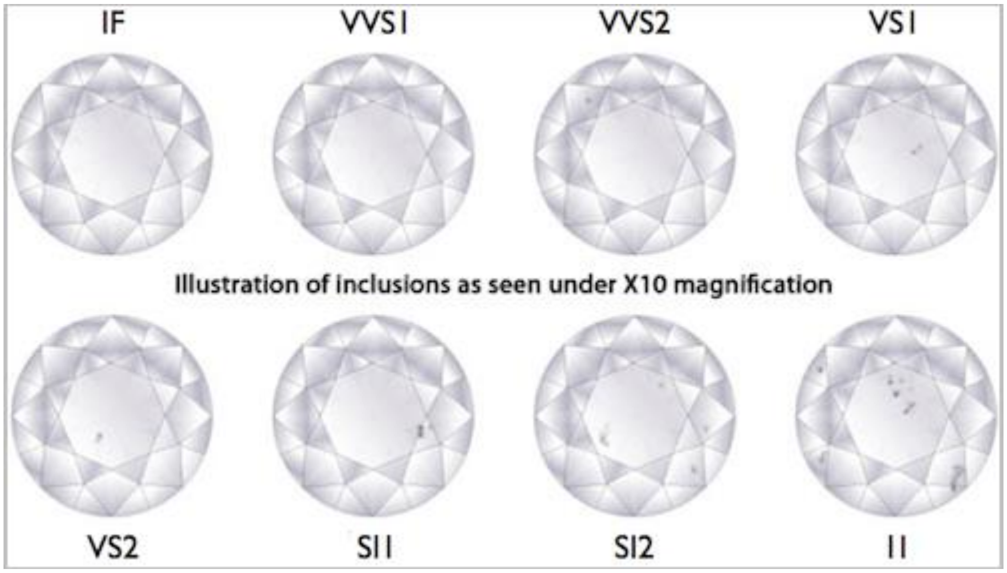
Color

The figure below shows color grading of diamonds:



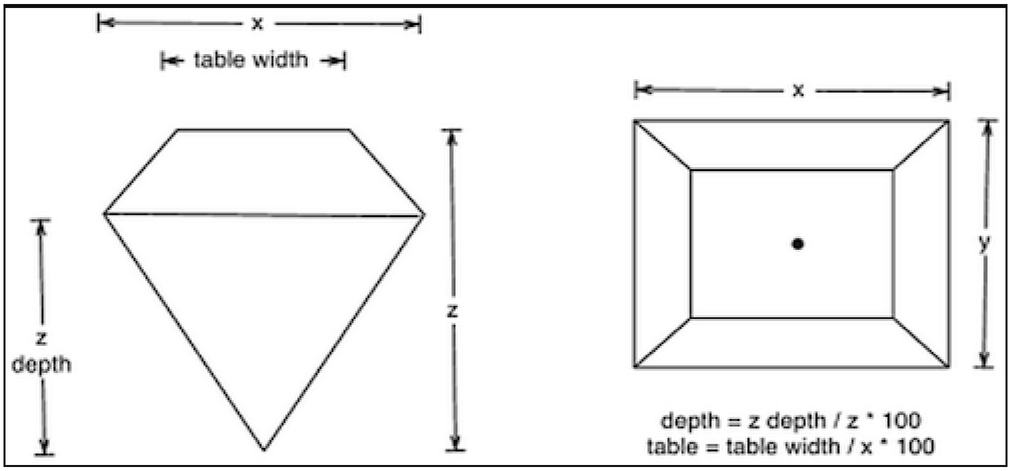
Clarity

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
   - Read Data
   - Show Structure
   - Basic Summary
   - Display Average Price in Crosstab with Carat & Cut
2. Data Cleaning & Imputation
   - Check For Zeros In Numeric Columns (except column "Price"). Convert to Null.
   - Check For Outliers in Numeric Columns (except column "Price"). Convert to Null.
   - Check For Undefined Data In Categoric Columns (except column "Cut"). Convert to Null
   - Check For Nulls In All Columns (except column "Price" & "Cut"). Get final tally of nulls in these columns
   - Suitably impute these Null values.
3. Visual Data Analysis
   - Display data distribution for "Price", "Carat", "Color", "Cut", "Clarity". Interpret the results.
   - Display relationship between "Price" and each of the Cs ie "Carat", "Color", "Cut", "Clarity". Interpret the results.
   - Display any other interesting patterns with help of suitable graphs.
4. Machine Learning 1
   - "Price" maybe dependent on all other columns.
   - Extract data with Null Values in a separate dataframes.
   - Treat main dataframe as your train/test data and dataframe with null as Prediction Data.
   - Predict "Price" for Prediction Data based on linear regression
5. Machine Learning 2 20 Marks
   - "Cut" is dependent on "Price", "Depth", and "Table".
   - Extract data with Null Values in a separate dataframes.
   - Treat main dataframe as your train/test data and dataframe with null as Prediction Data.
   - Predict "Cut" for Prediction Data based on suitable classification algorithm.

Presentation

- The project to be submitted using a single .py file.
- Use “###############”to break .py file into multiple sections.
- The first section introduces you & your team
- Each of the above “Project Requirement” is to be answered in 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 .py file needs to be submitted in “Project Work” assignment of Google Classroom

   Only one submission per group is required
4. The project needs to be submitted by Sat 19-Dec-2020 by 0800 am.
5. The viva / presentation for the project will be held on 19-Dec-2020 during the day.
6. Webex meeting will be set up for each group. Individual groups will be informed of their viva time.
7. Python related questions will be asked based on the not just the project but the full course.

Final Evaluation	**100**
Coding *(how good is the code - same for all team members)*	*20*
Documentation *(inline comments & explanations- same for team members)*	*20*
Accuracy *& Interpretation (are the results correct same for team members)*	*20*
Viva Per Individual *(different for all team members)*	*40*

Wishing You All The Best!!!