Diamond Price Prediction:

About the dataset:

To predict price of given diamond (Regression Analysis).

Data source: https://www.kaggle.com/datasets/soumyakushwaha/gemstone/data

Format: **CSV format**

Columns present in the data:

There are 11 independent columns :

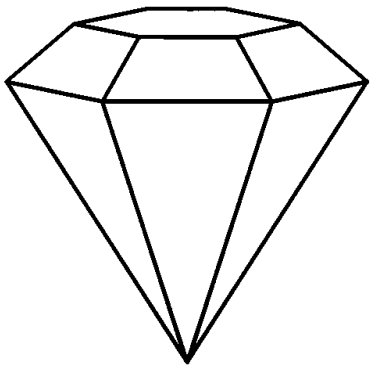
- **id** : unique identifier of each diamond
- **carat** : Carat (ct.) refers to the unique unit of weight measurement used exclusively to weigh gemstones and diamonds.
- **cut** : Quality of Diamond Cut
- **color** : Color of Diamond
- **clarity** : Diamond clarity is a measure of the purity and rarity of the stone, graded by the visibility of these characteristics under 10-power magnification.
- **depth** : The depth of diamond is its height (in millimeters) measured from the culet (bottom tip) to the table (flat, top surface) {depth: Total depth percentage: 100 * z / mean(x, y)}
- table : A diamond's table is the facet which can be seen when the stone is viewed face up.
- **x** : Diamond X dimension
- **y** : Diamond Y dimension
- **x** : Diamond Z dimension

Target variable:

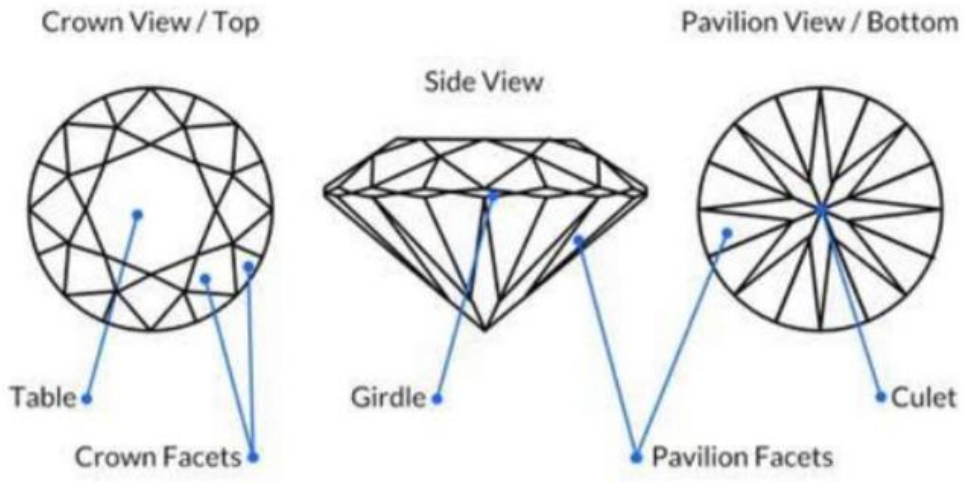
- **price**: Price of the given Diamond.

Size of the data: 193573

There are 3 categorical columns 'color', 'cut', 'clarity' and 7 numerical columns (including target column) 'carat', 'x', 'y', 'z', 'depth', 'table', 'price'.

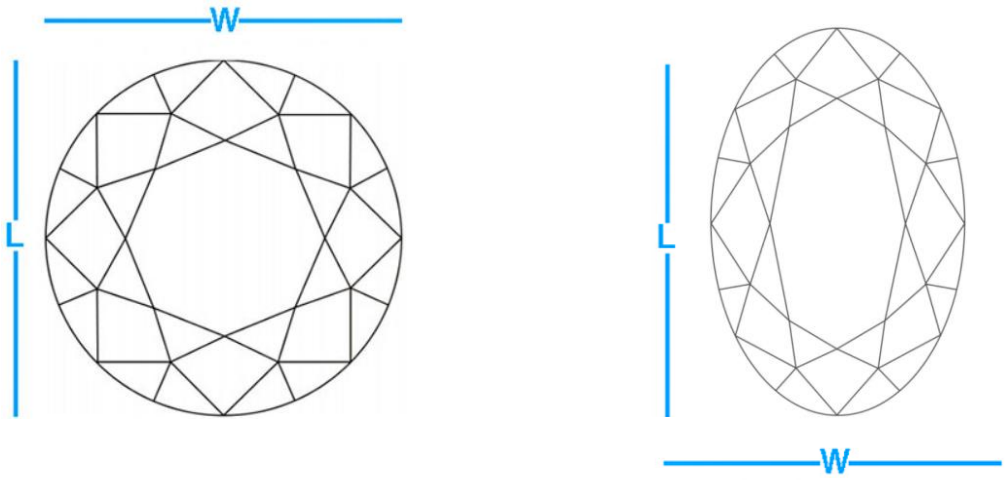


Basic Anatomy of a diamond to understand the dataset clearly.



Why take length, width separately in the dataset?

In round diamonds, the length and width are same but not all diamonds are round. There are oval shaped diamonds as well. To know the exact length and depth of the oval shaped diamonds we take length and depth as different columns.



Understanding the 4Cs of a diamond:

CUT, COLOR, CLARITY, CARAT:

CUT: This refers to the specific proportions of the diamond.

COLOR: There are three basic components of color grading being hue, tone and saturation. If the diamond is transparent, it is expensive.

CLARITY: Clarity refers to the absence of inclusions and blemishes found in diamonds.

CARAT: The weight of diamonds, as in all precious stones, is expressed in "Carats".

Data Cleaning:

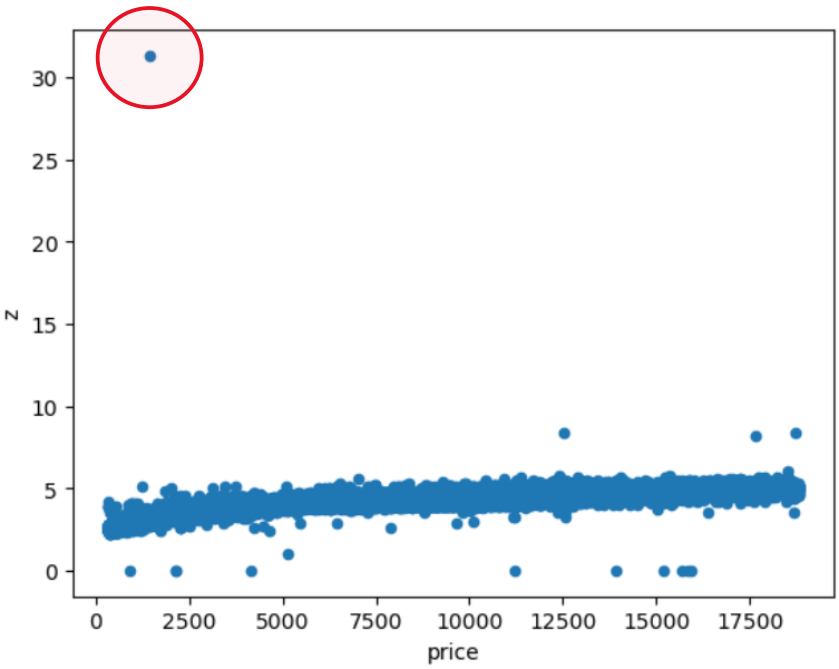
	id	carat	cut	color	clarity	depth	table	x	y	z	price
0	0	1.52	Premium	F	VS2	62.2	58.0	7.27	7.33	4.55	13619
1	1	2.03	Very Good	J	SI2	62.0	58.0	8.06	8.12	5.05	13387
2	2	0.70	Ideal	G	VS1	61.2	57.0	5.69	5.73	3.50	2772
3	3	0.32	Ideal	G	VS1	61.6	56.0	4.38	4.41	2.71	666
4	4	1.70	Premium	G	VS2	62.6	59.0	7.65	7.61	4.77	14453

- **Removing 'id' column: removed the 'id' column as it has all unique values.**

- **Removing rows with '0' value in x, y, z columns- There are few values where there are zeros in the x ,y, z column. As we know there cannot be zero length, zero width or zero depth, it means that there are missing values in the x, y, z columns which means there are only 2 dimensions available in few rows. we will drop them.**

- **Before dropping-193573 ,After dropping – 193563.**

- **Removing outliers: For example, if we look at the relationship between 'z' and 'price' columns, we can see that there is a good linear relationship between these columns but there are data points which can disturb the models. So we will remove them.**



The red circled one is an extreme outlier which can effect the model. So we will remove them.

- **Removed extreme outliers from X, Y, Z columns and capped the 'Depth' and 'Table' column.**