Diamonds Ratings Explained

Introduction

When analysing data, a key part of understanding is *domain knowledge*. What do the variables actually represent? The following is an in-depth explanation of the diamonds data set and what the different variables actually mean.

Description

`diamonds` is a data frame with 53940 rows and 10 variables:

variable	description	value range
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 colour	D (best) to J (worst)
clarity	a measurement of how clear the diamond is	I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)
depth	total depth percentage, calculated by $2 \cdot \frac{z}{(x+y)}$	43–79
table	width of top of diamond relative to widest point	43–95
x	length in mm	0–10.74
y	width in mm	0–58.9
z	depth in mm	0–31.8

Diamond Ratings Explained

There are multiple levels describing the cut, color, and clarity of the diamond and several variables describing the size of the diamonds, known as "the four Cs." The combination of cut, color, carat, and clarity are the key benchmark descriptors of a diamond and provide customers context when shopping for stones and give objectivity to retailers when pricing stones.

x, y, z, table, depth

The x, y, z, depth, and table variables are all descriptive values explaining the size of the diamond. The depth variable is calculated based on the x, y, and z values and defines the total depth percentage. The formula used to create the values in the depth variable for each observation is $2 \cdot \frac{z}{(x+y)}$. The table variable explains the width of the top of the diamond relative to its widest point. For ease of understanding future analysis and to aid in communicating the data, some variables will later be renamed to be more descriptive. Figure 1 visually explains the table and depth percentage.

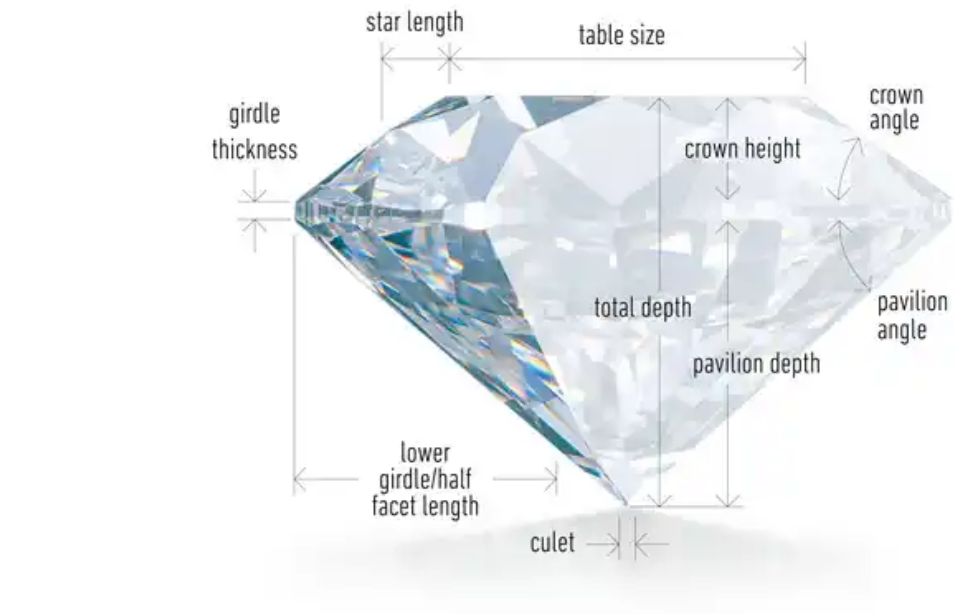

Figure 1: Table and depth percentage

Carat

Other than focusing on length, width, and height, a more simplified method of measuring the size of the stone is by carat. The carat is derived from an ancient unit of measure called the carob seed, which was uniform in size and weight. The carat measurement accounts for the stone's length, width, and height. Figure 2 provides a graphic of carat weights.

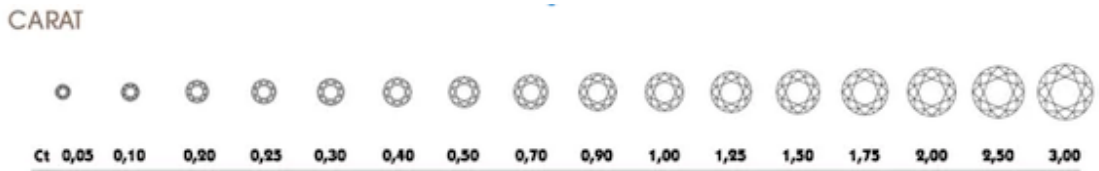

Figure 2: Carat weights

Cut

Cut is a qualitative descriptor ranging in quality from “Fair” to “Ideal,” with “Fair” being the worst cut and “Ideal” as the best cut in the dataset. The cut is a quality of the diamond that creates how light will reflect on the stone’s interior. Diamonds with “Ideal” cuts are highly reflective of light due to the quality of shaping, and diamonds with a “Fair” rating create less brilliance back to the observer based on the external geometry. Figure 3 outlines how light interacts with diamonds based on the cut quality.

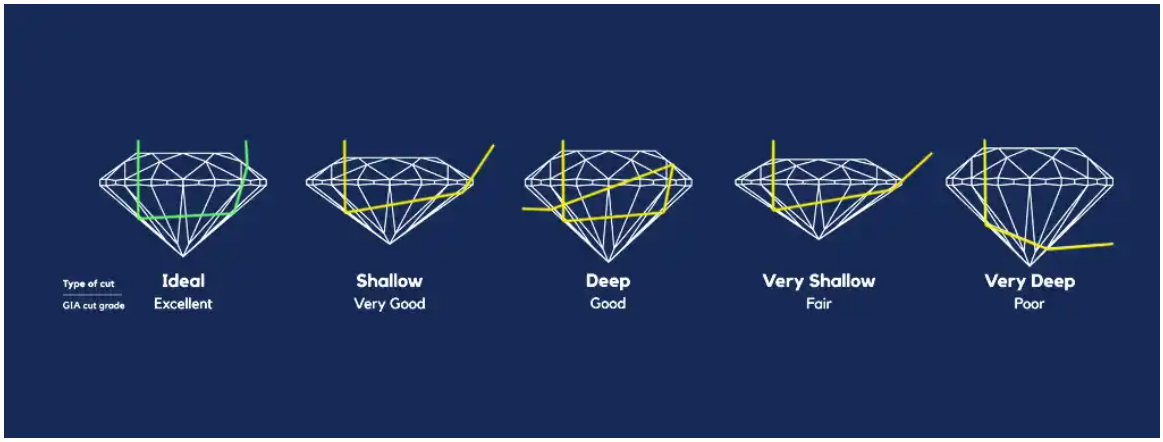

Figure 3: Cut quality

Color

The color variable is also represented through qualitative descriptions of color ranging from “D” as the worst and “J” as the best, following an alphabetic sequence. However, diamonds can contain ratings beyond “J.” Figure 4 below illustrates all color ratings for diamonds. As evident from Figure 4, the colors in the dataset are not representative of all color ratings. Additionally, the diamonds in the dataset represent the most clear rating available. Therefore, it is likely that the diamonds in the dataset are for jewelry and not industrial use.

Clarity

Clarity is slightly more complex in that it does not follow a numeric or alphabetic sequence to rank each of the colors of diamonds, but a sequence from best to worst within the dataset is “IF”, “VVS1”, “VVS2”, “VS1”, “VS2”, “SI1”, “SI2”, “I1”. Each of the ratings is, however, an acronym describing the clarity of the diamond. For example, a clarity rating of ’IF” is an acronym for internally flawless. Understanding that diamonds of lower grade are not seen in the dataset further reinforces the notions that the diamonds contained in the dataset are of higher quality. Figure 5 provides a visual representation of the appearance of each clarity rating a diamond can possess.

Figure 4: Color ratings

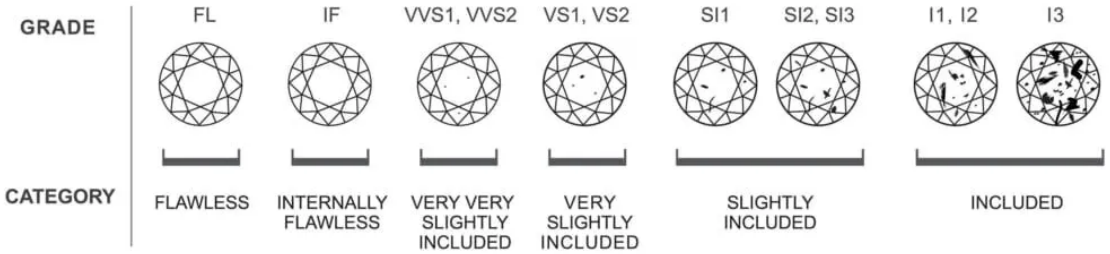

Figure 5: Clarity