Color Writeup

Terrance Luangrath

2025-03-11

Data Description

The data set, `diamonds4.csv`, contains 5 varaibles that describes more than 1000 different diamonds for sale.

- `Carat`: *fill in description*
- `Clarity`: *fill in description*
- `Color`: how colorless a diamond is, the more colorless the diamonds, the rarer it tends to be
- `Cut`: *fill in description*
- `Price`: the price value of the diamond in USD

Variable Analysis: Color

The diamond color refers to how colorless a diamond is. From the Blue Nile website, color is the second most important of the 4Cs of diamond. The more colorless the diamond is the rarer it is. Diamond colors are classify in three main categories,

- Colorless Diamonds
- Near-Colorless Diamonds
- Faint Diamonds

from these three categories, diamond color are grades from the ranges of D (colorless) to K (faintly colored). However, in this dataset, there exists only diamond graded from D to J.

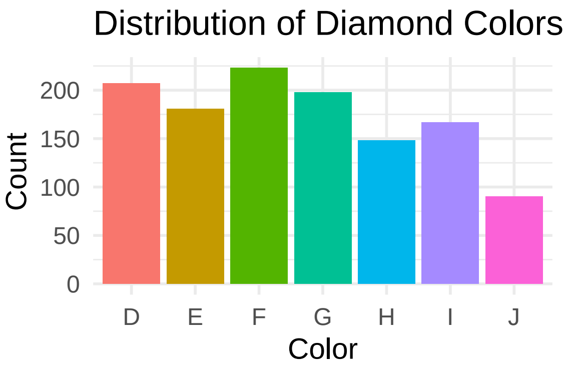


Figure 1: Diamond Color Distribution

Looking at Figure ??, the distribution of diamond color appears fairly consistent, with counts ranging between 150-200 for most color grades. However, color `J` semms to have around 100 diamonds in the data set.

Each diamond color falls into one of the three categories from Colorless to Faintly Colored.

- Colorless diamonds: `D`, `E`, `F` Color Diamonds
- Near-colorless diamonds: `G`, `H`, `I`, `J` Color Diamonds
- Faint color diamonds: `K` Color Diamonds

Since there is no `K` in the data set, it would be assumed there is no faint color diamonds in the data set. To facilitate the analysis, we will create a new column called `color_cateogry` to group each `color` with their respected category.

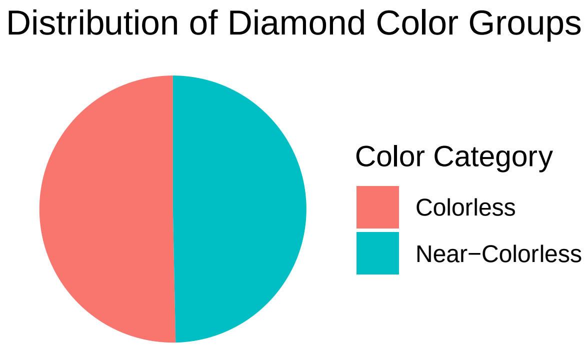


Figure 2: Diamond Color Category visualize through a Pie Chart

Looking at ??, the distribution of diamond colors seems to be nearly even, with `Colorless` containing slightly higher amount of diamonds. With a well spread of diamond colors, further analysis is needed to understand the claim and how diamond color would be correlates with the other factors in the data set.

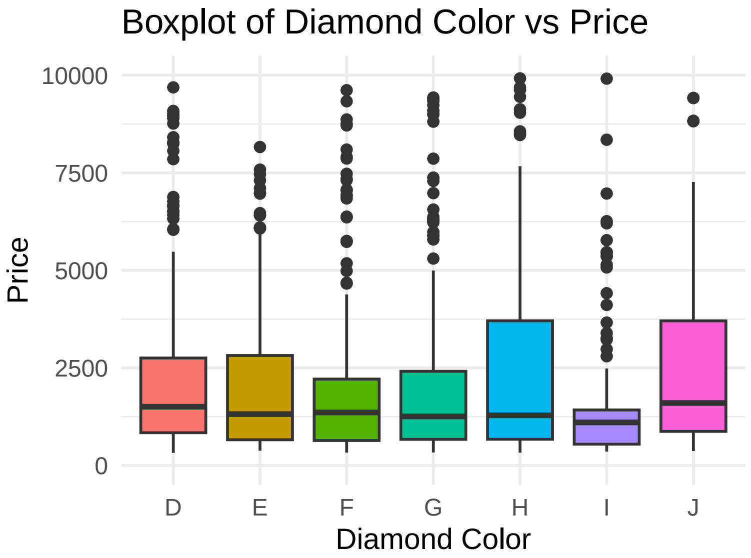
Since there is only two group with a near even split, we will work with each individual colors for the rest of the analysis to get a bigger picture of the data.

Bivariate Analysis

Bivariate Analysis on Prices

Table 1: Bivariate Analysis on Prices

color	min	q1	median	q3	max
D	322	882.00	1781	6048.50	355403
E	376	705.00	1602	5138.00	345397
F	328	704.00	1485	4372.00	227960
G	332	688.75	1374	3871.75	165766
H	326	727.50	1575	6135.00	134856
I	354	574.50	1212	1680.50	123311
J	369	957.50	1803	4504.00	40184

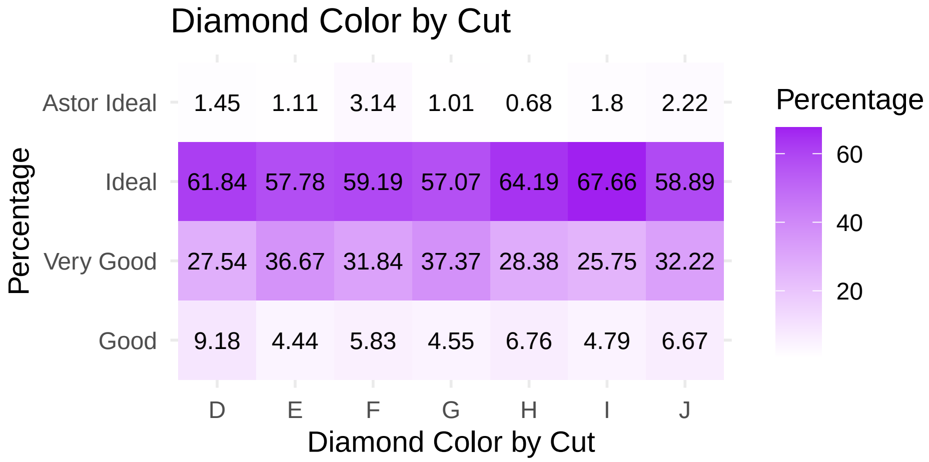


On the Blue Nile site, it mentions that color is the second most important among the 4C's for the price of the diamond. Reviewing the box plot, the colorless diamonds (D, E, and F) tends to have higher prices than the near-colorless diamond (G, H, I, and J). However, some colorless diamond have higher median values than the near-colorless diamond:

- H have higher median values than the colorless diamond F
- J (the lowest color grade) has a higher median value than all the colorless diamond.

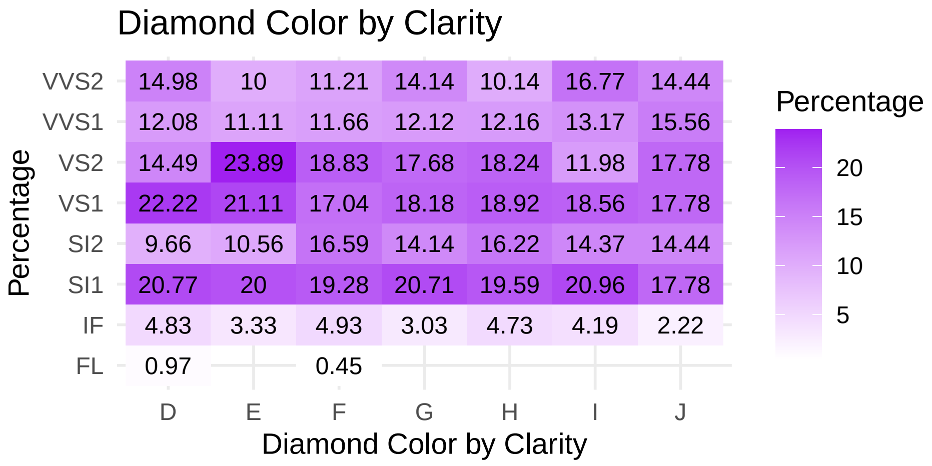
This suggests that color does play a significant factors in the diamond price, but it's not the most important factors to the diamond prices. Other characteristics, such as cut, clarity, and carat, needs to be considered to help prove the claim.

Color vs Other Diamond Characteristcs



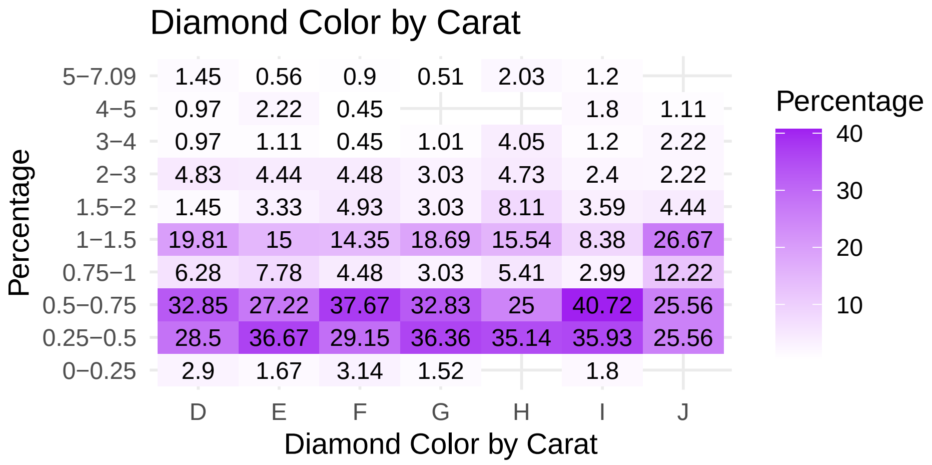
Reviewing the relationship between diamond color grades and cut, it shows that the majority of the diamonds for all the color grades fall under the ideal cut category. This suggets that higher quality are more common amohg diamonds regardless of the color.

The heat map shows that most diamonds are cut as ideal. This would align with the claim that `cut` is the most important factor because the cut is more proitize regardless of the color grades. If `color` was more dominant, then we would have seen color more even distributed across `cut`. Since lower grade cuts are less common than this futher supports our claim.



The higher the color grade, the better the clarity of the diamond tends to be. This suggests that buyers prioritize color over clarity. For buyers that want a lower color grade diamonds, people are more likely to choose a higher clarity as a trade-off.

In relation to the claim, buyers are more likley not to prioritize clarity as much for diamond selection. Since the spread is more even distributed, it suggests that people may choose a better color or cut over a fawless clarity. This aligns with the claim that color plays a more significant role than clarity.



The highest percentage on the heat map is between $(0.25, 0.75]$ carat range. This suggests that most of the diamonds in the data set are in the smaller carat ranges. Since the smaller the carat, the more affordable the diamond, this suggests suggests that most diamonds sold regardless the color grades have a small carat size. This may be because people are more likely to buy a smaller carat size then a large carat size for balancing cost. People who tends to buy lower color grade diamond are more likely to buy a larger carat size.

Going back to the claim, since cut is the most important and carat is the least prioritized, people are more willing to spend on diamond color than carat. This suggests that color is the second most proitize after cut. So, color plays a bigger role in the diamond price then carat.