Color Writeup

Terrance Luangrath

2025-03-17

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.

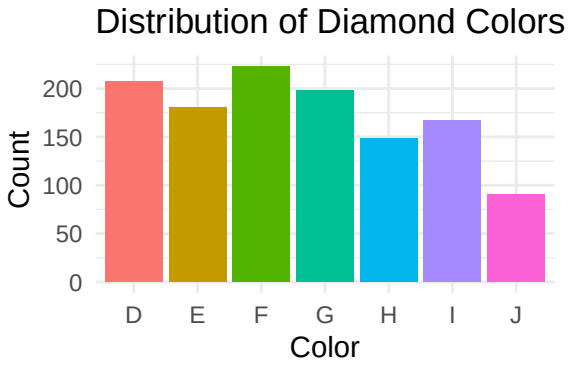


Figure 1: Diamond Color Distribution

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.

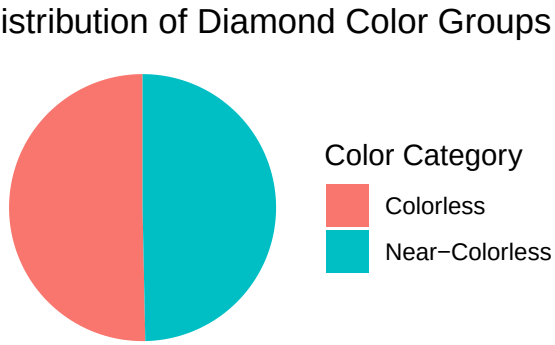


Figure 2: Diamond Color Category visualize through a Pie Chart

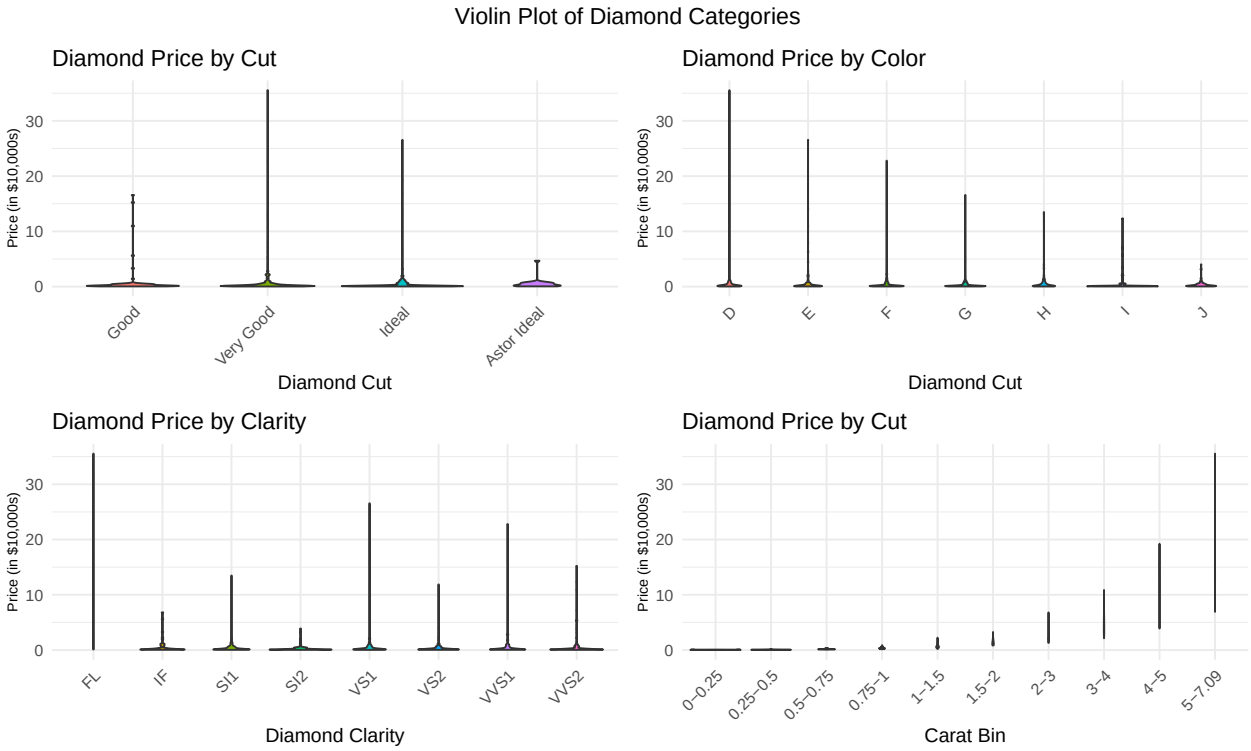
Looking at **??**, the distribution of diamond colors seems to be nearly even, with `Colorless` containing slightly higher amount of diamonds. With a well spread of diamond colors, further analysis is needed to understand the claim and how diamond color would be correlates with the other factors in the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

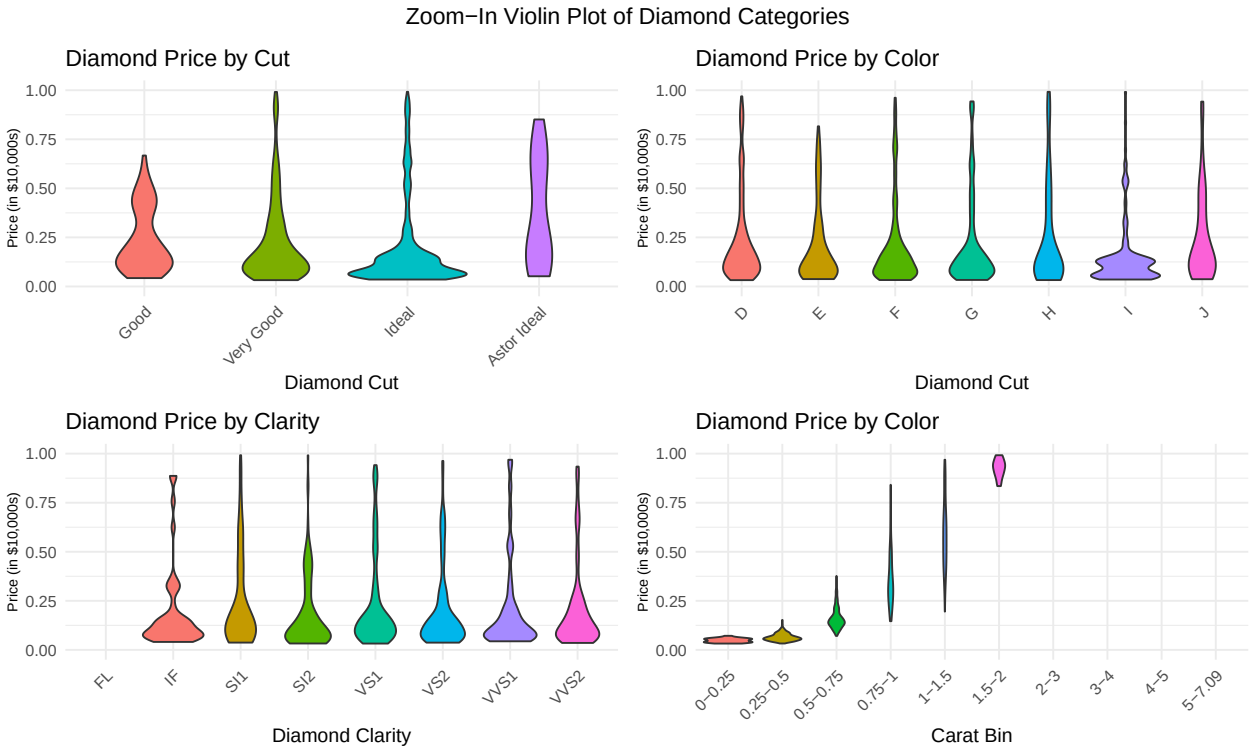
There are several factors that contribute to the diamond pricing, the four main factors that we are looking at is `carat`, `clarity`, `color`, `cut`. In the analysis, we will explore each factors individually and how they contribute to the diamonds price. By using a violin plot, we can visualize the full distribution of the data, including potential outliers, the density ofobservations, and the overall shape of distribution.



In **??**, it is difficult to clearly observe the distribution of the data, this could the presence of many outliers across the different diamond factors. This could be certain factors holding a higher prority than other diamond factors.

- Looking at `carat`, as the diamond carat weight increase, the diamond prices also increasing, showing a positive relationship.
- For diamond clarity, while flawless seems to have the highest price, but this is likely skewed due to the fact that there is only 3 diamonds that falls into the flawless categories. As a result, it's difficult to conclude that clarity plays a major role in pricing. Espcially, since clarity does not follow a consistent trends like the other categories.
- Diamond colors follow a consistent trends as the highest color grade is associated with higher prices.
- For the diamond cut, it seems that the `Very Good` cut shows higher prices compared to `Ideal` and `Astor Ideal` cuts. This suggests that other factors such as `carat` and `color`, have more influence over than `cut`.

Since, these violin plots are difficult to visualize full-scale, let's zoom into the a price range of $0 to $10,000 grasp a better bvisualize of the shape of the diamond prices distribution.



Looking at **??**, we can visualize some clear details of the diamond categories.

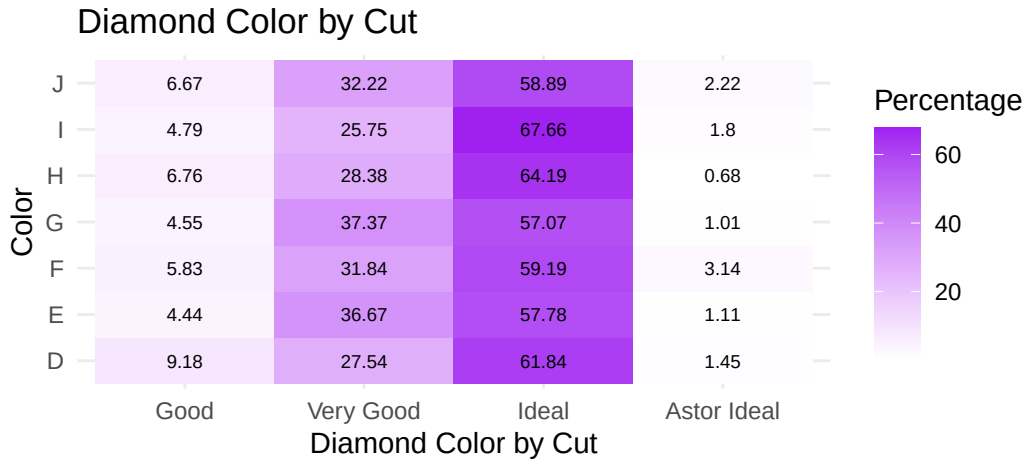
- The plot for diamond cuts shows that `Very Good` and `Astor Ideal` cut have higher price distribution over the other two types of cut. In all the cut catgoires, outliers are present espically `Very Good` cut. This suggests that the price of the diamond can spike regardless of the cut quality.
- The plot for diamond colors shows a positive trend. As the color grade increase, the diamond price also increase. Suggesting that the less color the diamond has, the higher the value is.
- The plot of diamond clarity shows a relatively even plot. Looking at flawless (`FL`) and internally flawless (`IF`), they seems to have the most noticeable outliers, but the rest of clarity shows a similar spread.
- The plot of diamond carat shows a strong positive relationship with price. However, most diamonds are clustered into the smaller bin, but that could be due to the fact, smaller diamonds are more common.

How does this relates to the Blue Nile claims?

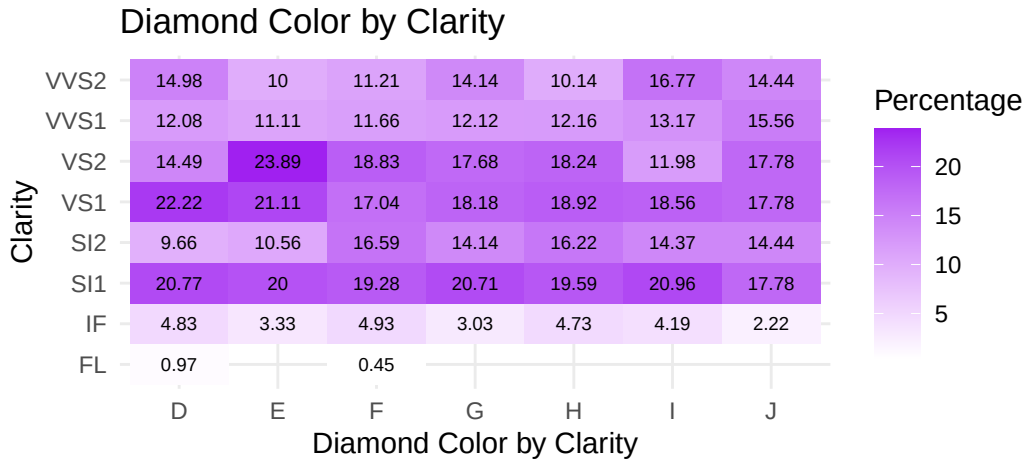
- `Cut` is the most important factor, but in the plot, it shows that `Ideal` cut doesn't have the highest price, but `Very Good` outperform the other cuts. This suggests cut does influence the price, but other factors play a higher significant values than `Cut`.
- The plot helps support the claim that colorless diamonds are worth more than near-colorless diamond. As the color grade shows a trend in the diamond price playing a significant role in the diamond price.
- The price distribution across `Clarity` is relatively even. This shows a lack of clear trends across the clarity levels. Suggesting that people prioritize color or cut over clarity. Thus, clarity is not a sufficient factor in the diamond price.
- `Carat` shows a strong positive relationship with price. This does support the claim that the carat weight affects the diamond price the most. However, carat also interacts with other factors like `Cut` and `Color` reinforcing that people do take in account other factors beside carat.

Color vs Other Diamond Characteristcs

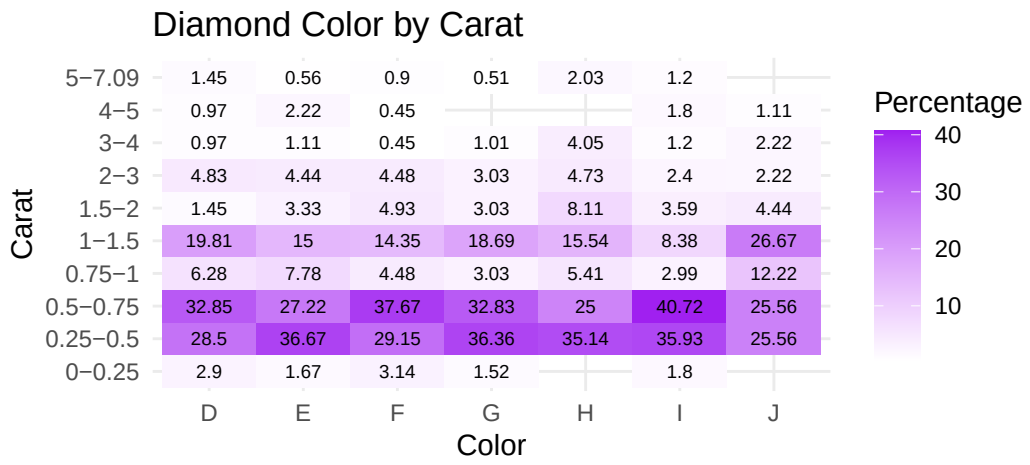
It's important how each categories interact with the diamond price, but how does each categories interact with each other?



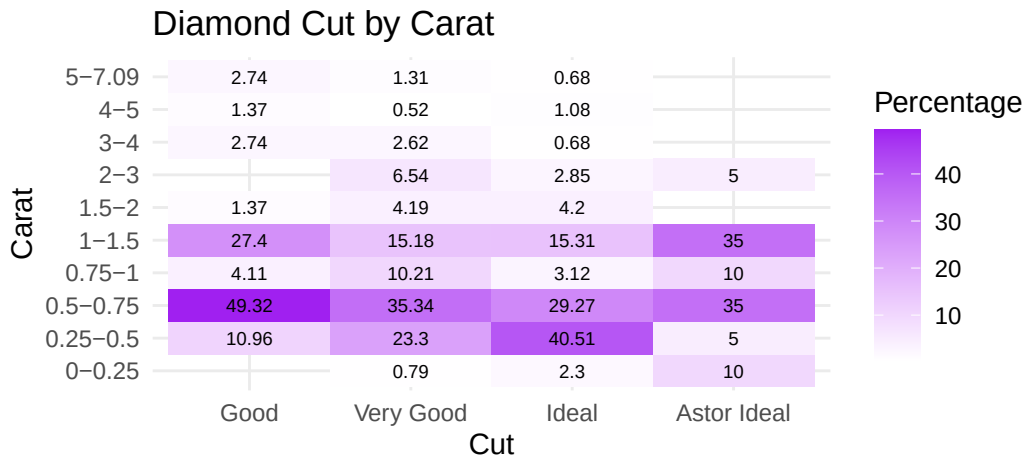
It seems that most of the diamond cuts are `Ideal` with proportion exceeding 50% each color grade. This suggests that `Ideal` cuts are the most common choice across colors. This not necessary mean that cut is less important than color. Rather, it's the more common choice regardless of color grade. This ties with how buyers would decide a balance between color and cut.



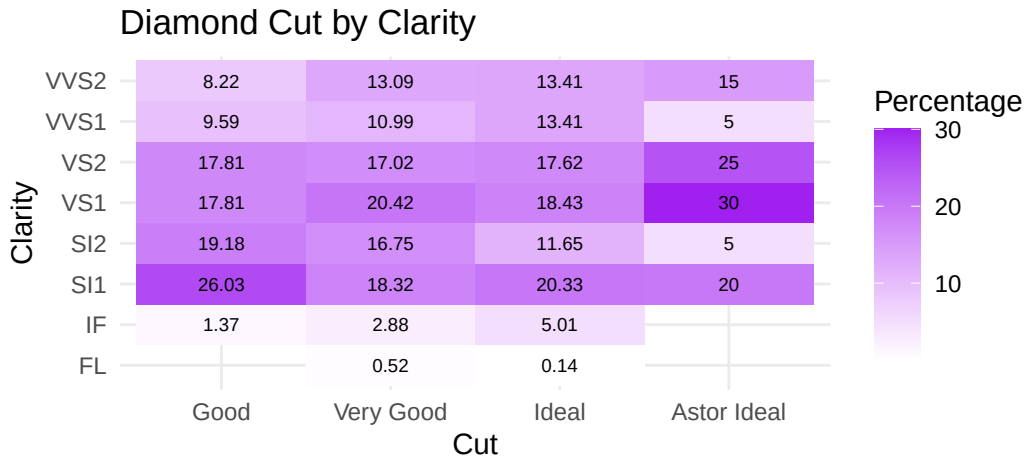
Reviewing the heatmap, most diamonds are concentrated around `VS1`, `VS2`, and `SI1` across all color grades. The distribution across those clarity categories are fairly consistent with no trend. This indicates that clarity is not that important and people would buy a higher color grade diamond than a higher clarity grade. This suggests a trade-off between color and clarity rather than solely focusing on clarity.



The majority of the diamond are concentrated from the 0.25 to 0.75 carat range. This suggests that the carat weight are commonly around the carat range listed before. In the colorless category, it's seems buyers would decide a colorless diamond over a heavier carat, but some buyers may opt for a heavier diamond with a lower color grades. It's suggesting that people would want these factors does not work separately from each other, but rather they works together in pricing and selection.



In Figure **??**, diamonds with carat weights below 1.5 are majority represented across all cut categories. This support the idea that carat weight is decided by buyer preference and budget considerations.In the earlier heatmap (Figure **??**), `Very Good` cuts were more common, but here, `Ideal` and `Astor Ideal` cuts are more prevalent. This suggest that cut may be prioritzed over carat weight when making purchasing decision.



This heatmap shows that most diamond cuts fall within the mid-range clarity levels. This suggests that buyers who prorize a higher quality cut may also prefer a decent clarity grade. Overall, cut takes precedence over clarity in the process, since buyers are willing to accept a lower clarity grades if they get a higher quality cut. This reinforces the idea that cut is the most influential factor when choosing a diamond.

Reviewing all the bi-analysis done, we detmerine

- Carat is the important when deciding the price of the diamond. Buyers tend to choose a lower carat weight than a heaver carat if that means saving some extra money.
- Color is the second most important factors in the selection process.
- Clarity is seems to be the less influential when selecting the diamond, as most buyers would choose a mid-level clarity grades.
- Overall, the heatmap shows the influence on all 4Cs, suggesting buyers often trade-offs between them based on personal preferences and budget considerations.