Color Writeup

Terrance Luangrath

2025-03-07

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

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 catgories, 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.

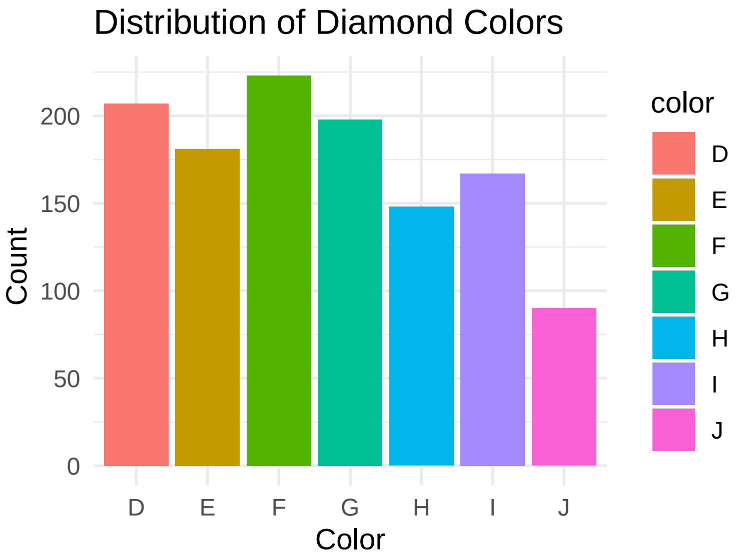

Figure 1: Diamond Color Distribution

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.

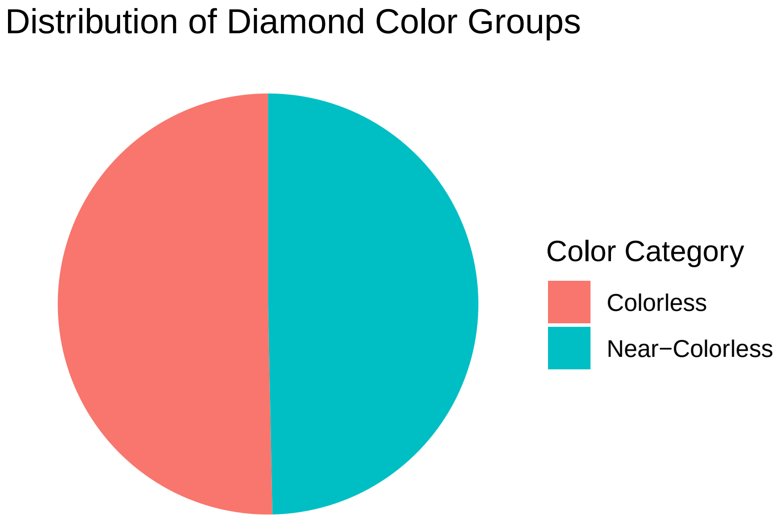

Figure 2: Diamond Color Category visualize through a Pie Chart

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.

Relationship with other variables