Homework Data Visualization

Jordan S.

Date: 2023-03-01

Hello Reader

This language is markdown. Today I learned a few R topics, including:

- database
- working with date
- ggplot2
- rmarkdown

I therefore make an effort to illustrate what I have learned through project work.

Let’s explore the diamonds (https://ggplot2.tidyverse.org/reference/diamonds.html) together!

The goal of this study is to identify every element that affects the price of diamonds.

Prepare a library and reduce the sample size by 10% of the total sample size before sampling.

```
library(tidyverse)
library(patchwork)
library(dplyr)
set.seed(11)
diamonds <- diamonds %>%
  sample_n(5393)
```

Scatter plot of price vs. carat by color

```
ggplot(diamonds, aes(x = carat, y = price, color = color)) +
  geom_point(alpha = 0.2) +
  scale_color_hue(l = 50, c = 100) +
  theme_minimal() +
  labs(x = "Carat Weight", y = "Price", color = "Diamond Color")
```

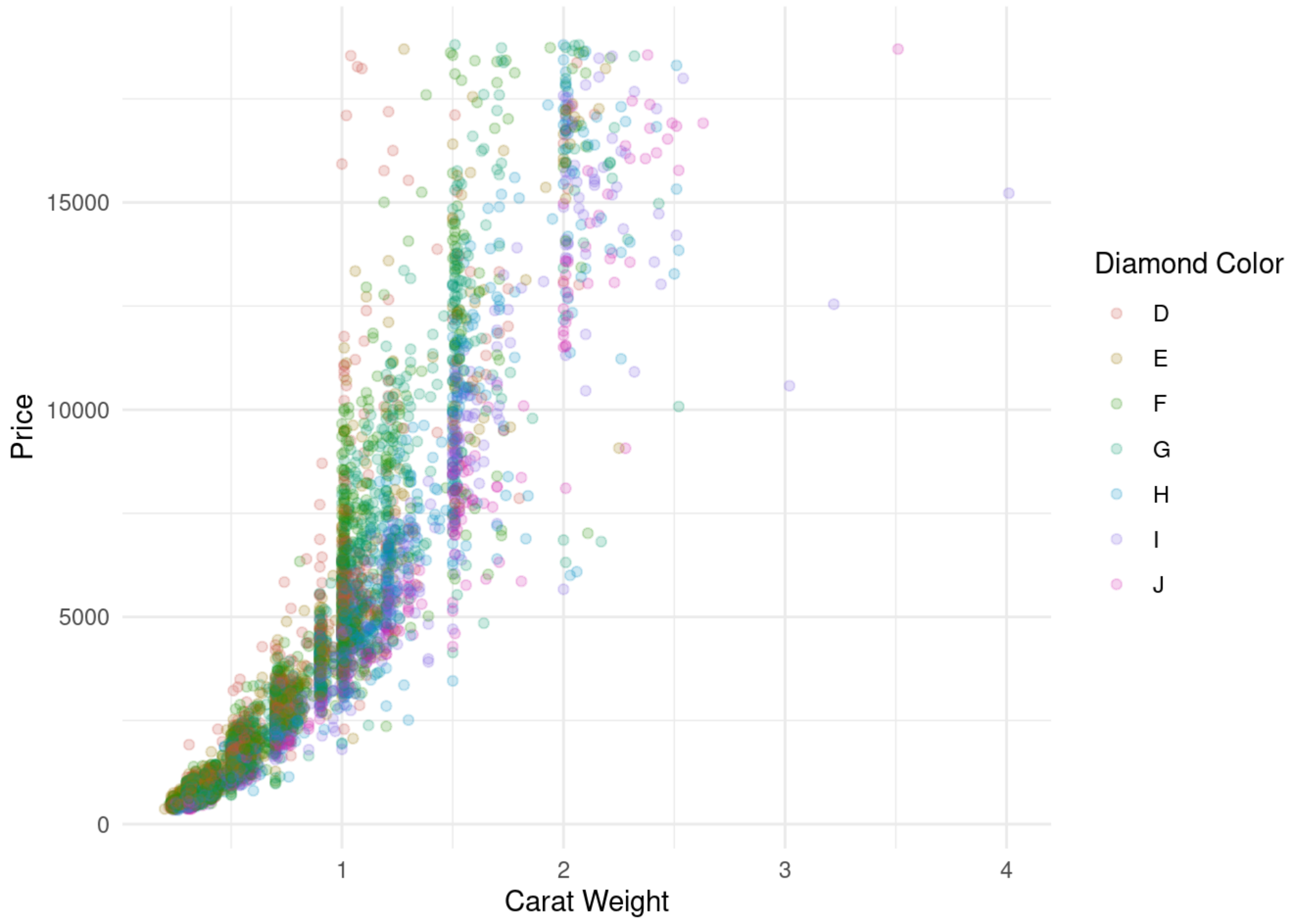

Figure1: Scatter plot of price vs. carat by color

This plot shows how price and carat weight are related to diamond color. We can see that as carat weight and price increase, diamonds become more likely to be in the higher color categories (e.g., D, E, F), which are associated with greater transparency and therefore higher quality. However, there is still a lot of variability in price and color within each carat weight category.

Box plot of price by cut and color

```
ggplot(diamonds, aes(x = cut, y = price, fill = color)) +
  geom_boxplot() +
  scale_fill_hue(l = 50, c = 100) +
  theme_minimal() +
  labs(x = "Diamond Cut", y = "Price", fill = "Diamond Color")
```

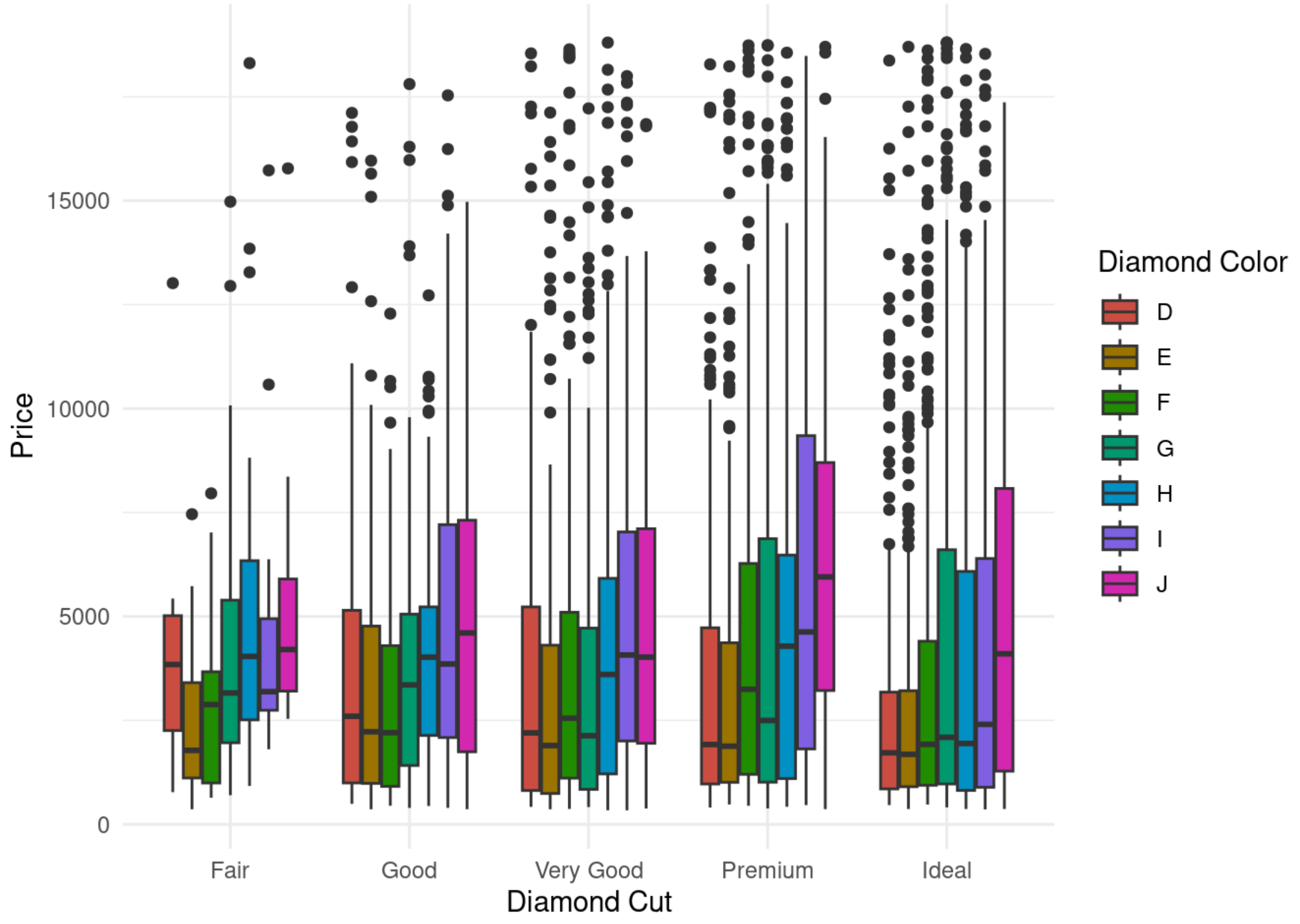

Figure2: Box plot of price by cut and color

The plot suggests that there is a strong relationship between diamond cut and price, as the median price generally increases with higher cut grades. Additionally, there is a clear relationship between diamond color and price, with higher color grades commanding higher prices. There also appears to be an interaction between cut and color, as the relationship between price and cut grade varies depending on the color grade of the diamond. Overall, this plot suggests that both diamond cut and color are important factors that influence the price of a diamond.

Density plot of price by clarity

```
ggplot(diamonds, aes(x = price, fill = clarity)) +
  geom_density(alpha = 0.5) +
  theme_minimal() +
  scale_fill_hue(l = 50, c = 100)
```

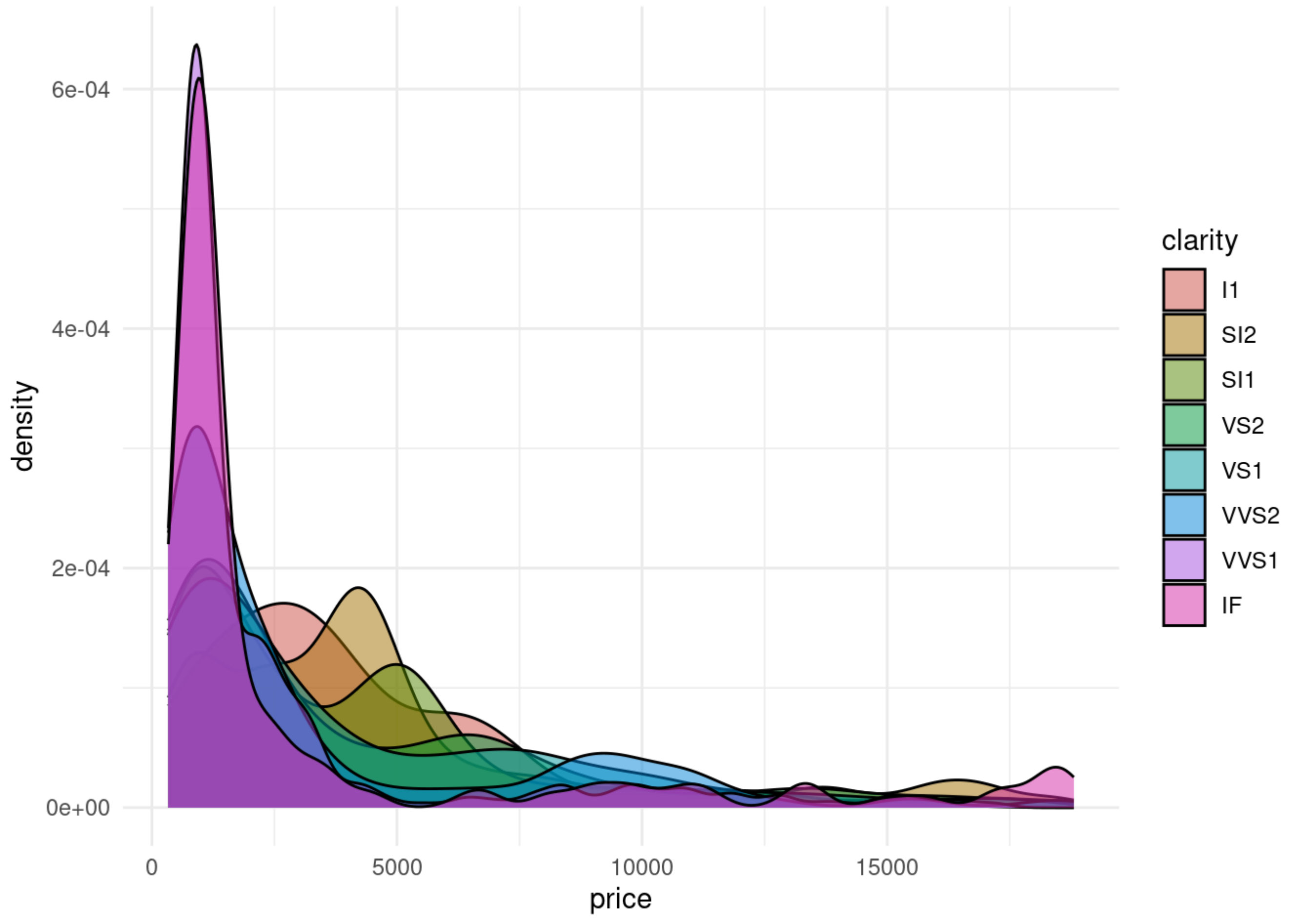

Figure3: Density plot of price by clarity

The plot suggests that there is a relationship between diamond clarity and price, as higher clarity grades tend to command higher prices. Additionally, the plot shows that the distribution of diamond prices varies depending on the clarity grade, with higher clarity grades having a narrower and more concentrated distribution of prices. Overall, this plot suggests that diamond clarity is an important factor that influences the price of a diamond.

Scatter plot of price vs. depth and table

```
ggplot(diamonds, aes(x = depth, y = table, size = carat, color = price)) +
  geom_point(alpha = 0.5) +
  scale_color_gradient(low = "blue", high = "red") +
  theme(legend.position = "bottom")
```

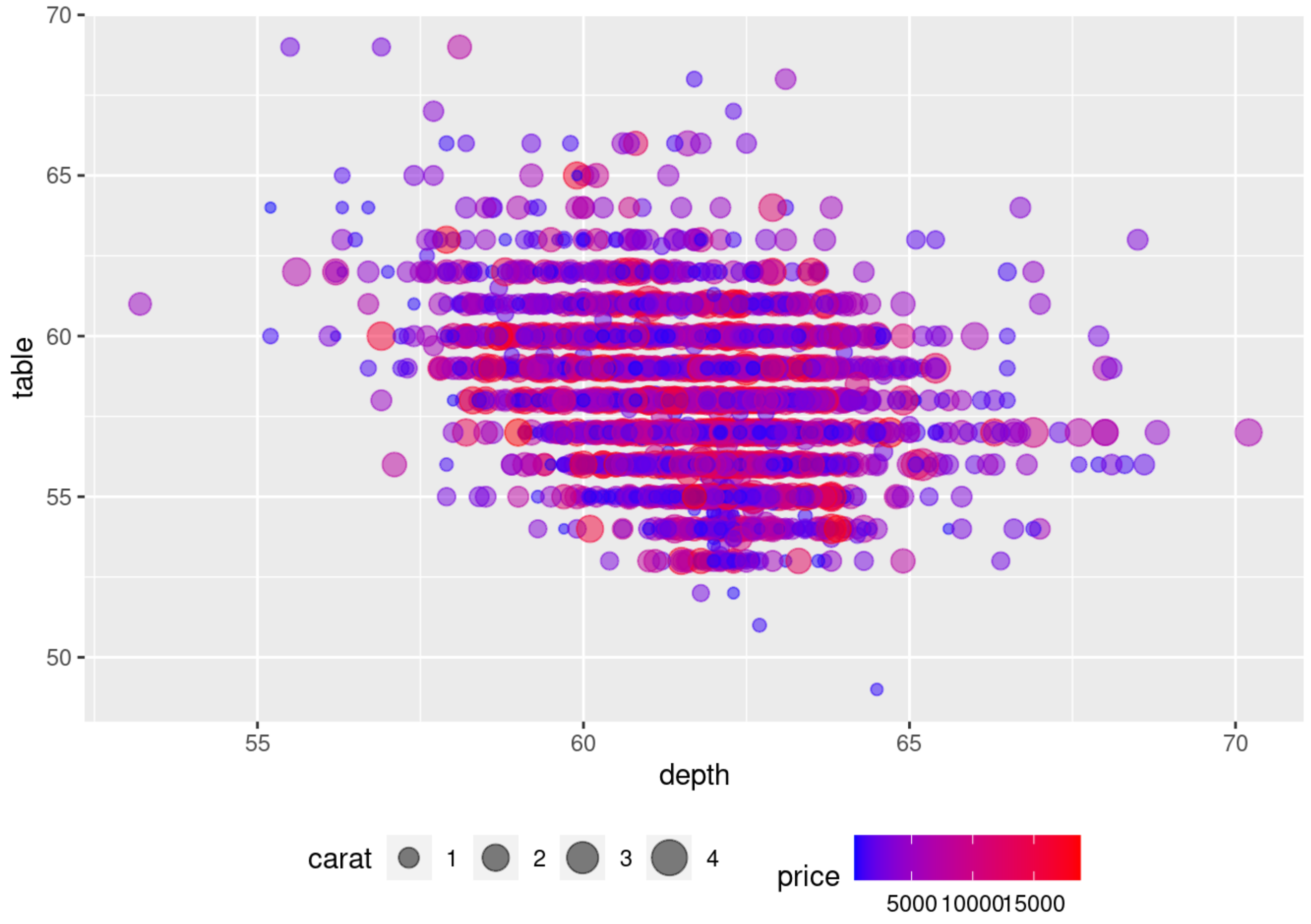

Figure4: Scatter plot of price vs. depth and table

The plot suggests that there is a positive relationship between diamond carat and diamond price, as larger diamonds tend to command higher prices. Additionally, there appears to be a slight relationship between diamond depth and diamond price, with diamonds with depths around 62-63% having higher prices than those with depths outside of this range. The plot also shows that the distribution of diamond prices varies across different combinations of diamond depth and table percentages.

Overall, this plot suggests that diamond carat is the most important factor that influences diamond price, followed by diamond depth and diamond table percentages. The plot also highlights the importance of considering multiple factors simultaneously when trying to understand the relationship between diamond characteristics and price.

Box plot of price by diamond shape

```
diamonds <- diamonds %>%
  mutate(cut_short = case_when(cut == "Fair" ~ "F",
                               cut == "Good" ~ "G",
                               cut == "Very Good" ~ "VG",
                               cut == "Premium" ~ "P",
                               cut == "Ideal" ~ "I",
                               TRUE ~ NA_character_))

diamonds$shape <- ifelse(diamonds$x == diamonds$y & diamonds$x == diamonds$z, "Round",
                         ifelse(diamonds$x == diamonds$y & diamonds$x != diamonds$z, "Square",
                                ifelse(diamonds$x != diamonds$y, "Non-round", NA)))

ggplot(diamonds, aes(x = cut_short, y = price, fill = cut)) +
  geom_boxplot() +
  facet_wrap(~ shape) +
  theme_minimal() +
  labs(x = "Diamonds Cut", y = "Price", fill = "Diamonds Cut")
```

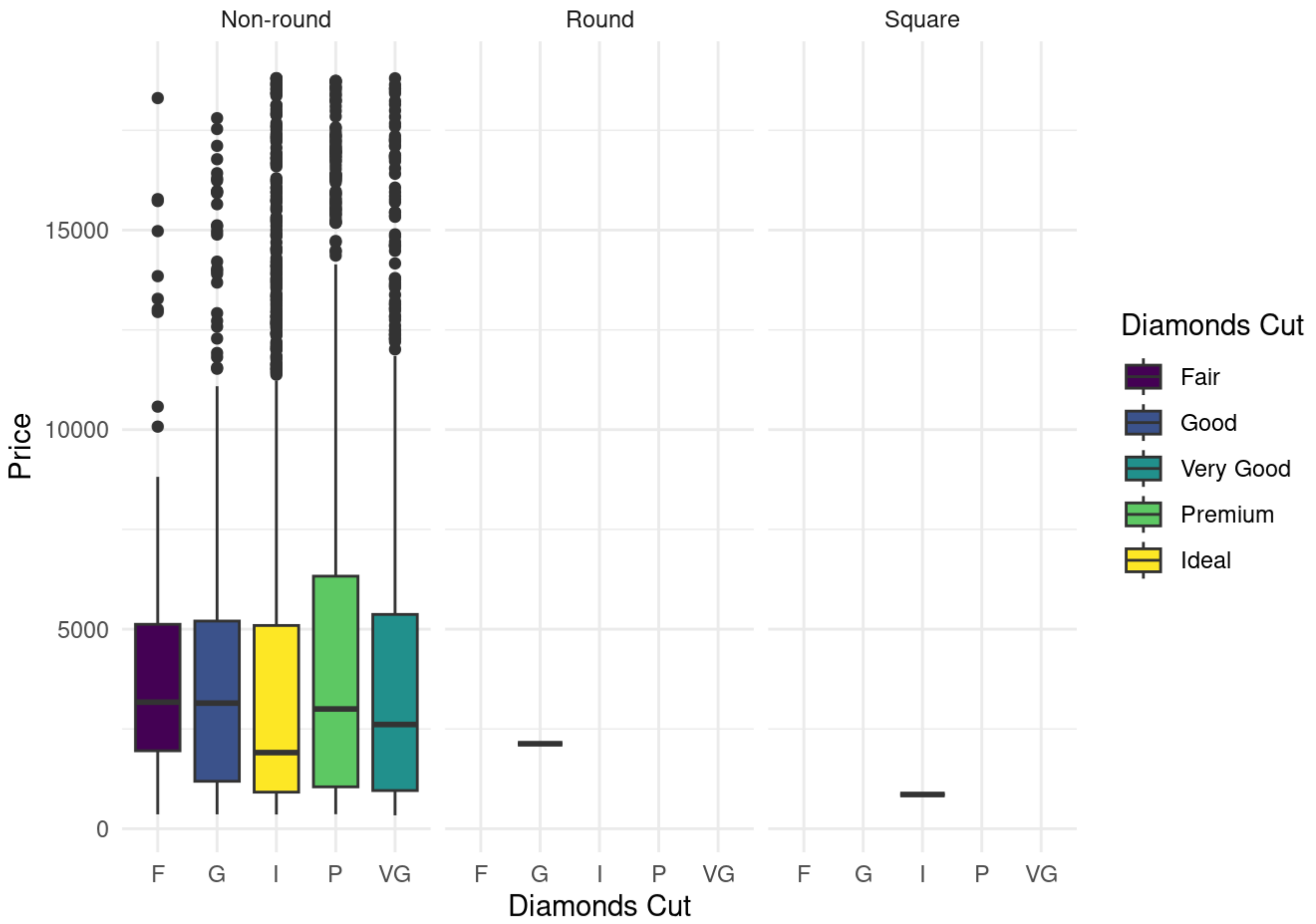

Figure5: Box plot of price by diamond shape

The chart reveals that the distribution of prices for diamonds with an Ideal cut is the most concentrated and has the narrowest range. On the other hand, the distribution of prices for diamonds with a Fair cut is the most dispersed and has the widest range. This indicates that the cut of a diamond has a significant impact on its price and that diamonds with an Ideal cut tend to have a higher price than those with a Fair cut. The chart also highlights some interesting differences between the shapes of diamonds. For example, the boxplots for Non-round diamonds tend to be skewed to the right, indicating a higher concentration of more expensive diamonds in this group. The boxplots for Square diamonds are also slightly skewed to the right, but to a lesser extent than Non-round diamonds. Finally, the boxplots for Round diamonds are relatively symmetric, with a generally consistent distribution of prices within each cut category.

Conclusion

The five charts provided insights into the factors that influence diamond prices, including carat weight, color, clarity, cut, depth, and table percentages. As carat weight and price increase, diamonds become more likely to be in the higher color categories, which are associated with greater transparency and higher quality. Both diamond cut and color are important factors that influence the price of a diamond, with higher cut and color grades commanding higher prices. Diamond clarity is also an important factor that influences the price of a diamond, with higher clarity grades tending to command higher prices. Overall, diamond carat is the most important factor that influences diamond price, followed by diamond depth and diamond table percentages. Additionally, the cut of a diamond has a significant impact on its price, with diamonds having an Ideal cut tending to have a higher price than those with a Fair cut. Finally, the shapes of diamonds also impact their prices, with Non-round diamonds having a higher concentration of more expensive diamonds compared to Square and Round diamonds.