Module 1 - EDA

ALY 6040

Jeff Hackmeister

2025-04-16

Introduction

For this Exploratory Data Analysis, I will be working with the Diamonds dataset from the ggplot2 library, which is part of the larger Tidyverse library. According to the document ion available from the ggplot2 library, this dataset contains price and descriptive attributes of almost 54,000 diamonds. I was interested in this dataset because the thought of diamonds for most people invites images of engagement rings and elaborate jewelry, but the vast majority of diamonds (by volume) are sold for use in manufacturing or other less glamorous uses. I was curious to see how the different descriptive aspects of a diamond would influence the price of each diamond.

To begin the analysis, I will load Tidyverse as well as the Patchwork library which allows for a cleaner presentation of multiple plots within this document. Then I will read into R the diamonds dataset.

```
library("tidyverse")
library("patchwork")
data <- diamonds
```

For more details on the the specifics of the the dataset, I will use the str function.

```
str(data)
```

```
tibble [53,940 x 10] (S3: tbl_df/tbl/data.frame)
 $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
 $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
 $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
 $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
 $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

From this we can see that there are 10 variables in the dataset for 53,940 observations. The descriptive variables include numeric values for:

- carat (weight)
- depth (total depth percentage)
- table (width of the top of the diamond relative to the widest point)

- X, Y, and Z (length, width, and depth
- price (measured in US dollars as an integer)

As well as factor variables for:

- cut (quality)
- color
- clarity

To confirm there are no missing values, I will use the colSums function with is.na

```
colSums(is.na(data))
```

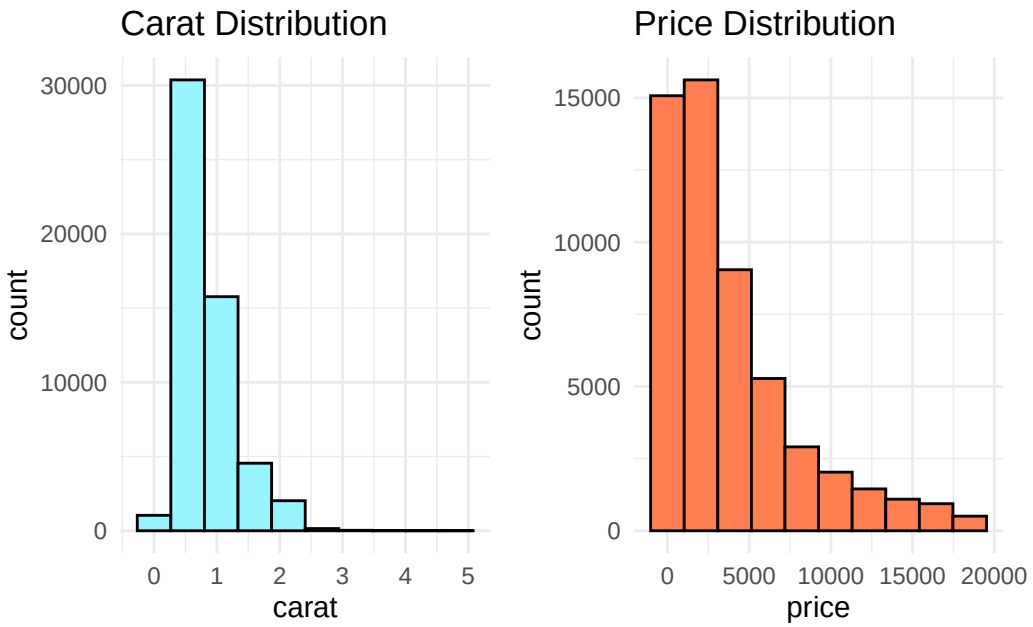
```
  carat     cut   color clarity   depth   table   price       x       y       z
      0       0       0       0       0       0       0       0       0       0
```

which confirms there are no missing values, meaning I will not need to decide on a methodology for replacing null values.

Initial Exploration

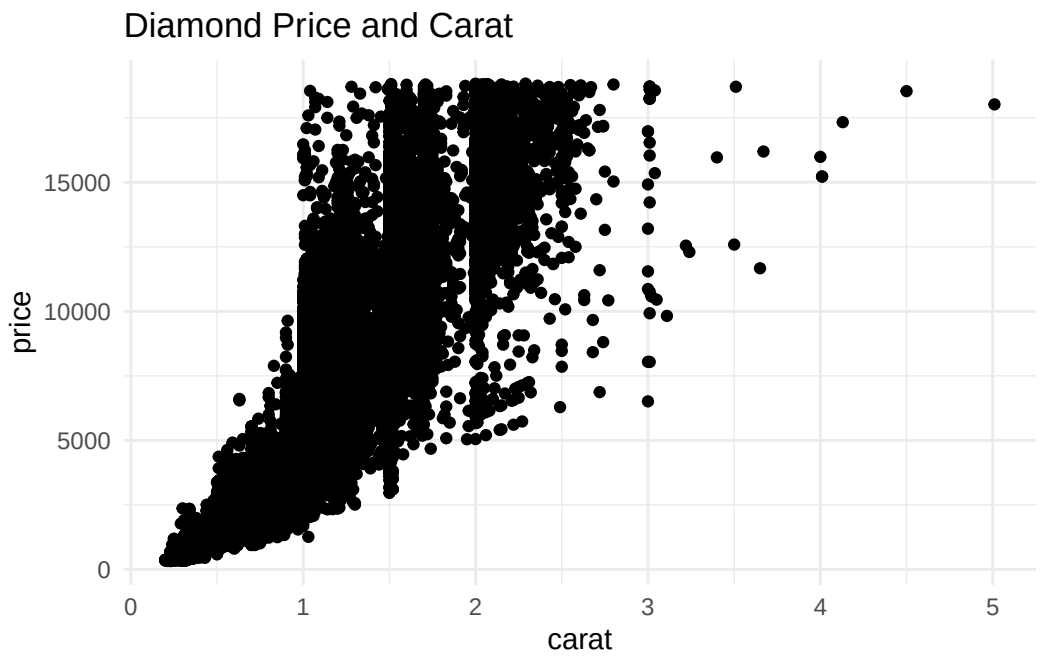
To explore the shape and distribution of the variables, I will create a few simple plots. Starting with histograms for carat and price.

```
p1 <- ggplot(data, aes(x=carat)) +
  geom_histogram(bins=10, fill = "cadetblue1", color = "black") +
  labs(
    title = "Carat Distribution"
  ) +
  theme_minimal()
p2 <- ggplot(data, aes(x=price)) +
  geom_histogram(bins=10, fill = "coral", color = "black") +
  labs(
    title = "Price Distribution"
  ) +
  theme_minimal()
p1 | p2
```



These demonstrate expected relationships, the bulk of the diamonds listed are below 1.5 carars and under $2,500. These are less likely to end up in jewelry but are vital parts of the diamond trade. I also wanted to explore the relationship between these variables with a scatter plot.

```
ggplot(data, aes(x=carat, y=price)) +
  geom_point() +
  labs(title = "Diamond Price and Carat") +
  theme_minimal()
```



This shows some interesting stratification of price at the 1.5, 2, and 2.5 carats - clearly prices are set around those thresholds.

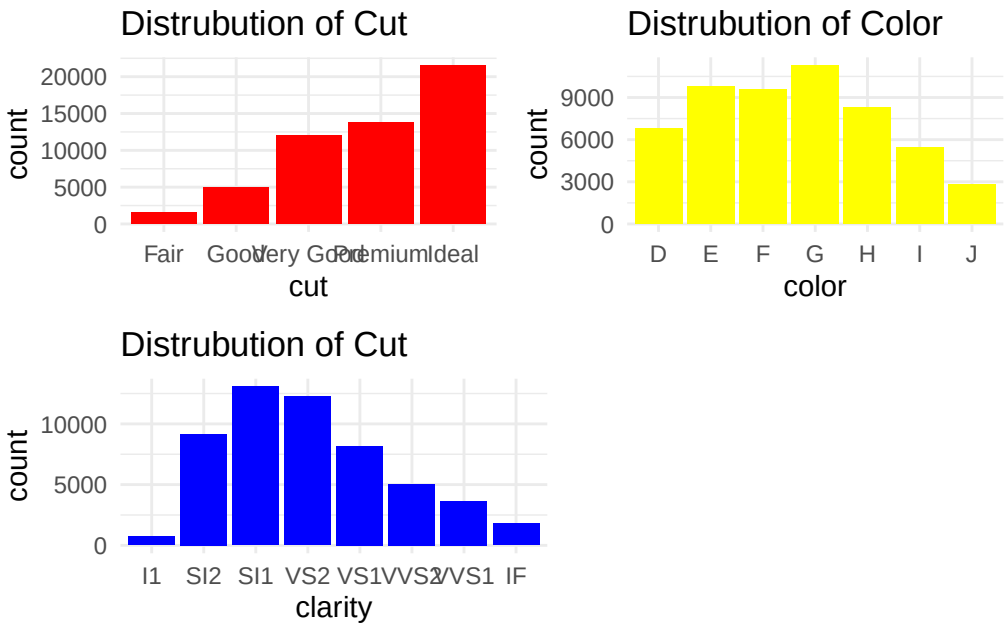
For cut, color and clarity we'll use bar charts to look at their distributions.

```
p3 <- ggplot(data, aes(x=cut)) +
  geom_bar(fill = "red") +
  labs(title = "Distrubution of Cut") +
  theme_minimal()

p4 <- ggplot(data, aes(x=color)) +
  geom_bar(fill = "yellow") +
  labs(title = "Distrubution of Color") +
  theme_minimal()

p5 <- ggplot(data, aes(x=clarity)) +
  geom_bar(fill = "blue") +
  labs(title = "Distrubution of Cut") +
  theme_minimal()

(p3 | p4) / (p5 | plot_spacer())
```



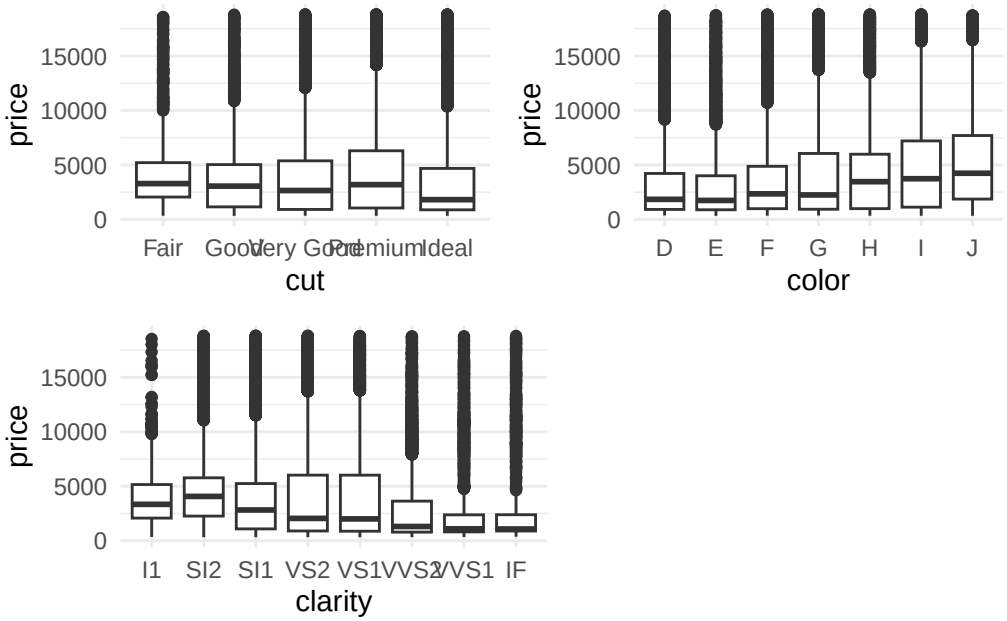
Her we can see a relatively normal distribution for color and clarity, but a large tilt in cut towards "ideal". To explore the relationship between these variables and price, I used a series of boxplots.

```
p6 <- ggplot(diamonds, aes(x = cut, y = price)) +
  geom_boxplot() +
  theme_minimal()

# Boxplot of price by color
p7 <- ggplot(diamonds, aes(x = color, y = price)) +
  geom_boxplot() +
  theme_minimal()

# Boxplot of price by clarity
p8 <- ggplot(diamonds, aes(x = clarity, y = price)) +
  geom_boxplot() +
  theme_minimal()

# Arrange in a 2x2 grid with one empty space
(p6 | p7) / (p8 | patchwork::plot_spacer())
```



These plots show that while the interquartile ranges are fairly tight across all variables and categories, price is not dictated by these variables alone as seen by the large numbers of exterior values.

Takeaways

From this EDA, I have found that is no missing data from the dataset. Additionally, while there are no traditional outliers, it is clear that while carat, cut, color, and clarity all impact the price of a diamond, there are likely other variables not included in this dataset that cause larger disparities in price. Going forward with this study, I would do more statistical tests to determine the strength of the relationship between these variables and could proceed to produce a price prediction model to estimate the sales price of a diamond given the descriptive variables available.

References

[1] Prices of over 50,000 round cut diamonds — diamonds. (n.d.). Ggplot2.Tidyverse.org. https://ggplot2.tidyverse.org/reference/diamonds.html

[2] The Composer of Plots. (n.d.). Patchwork.data-Imaginist.com. https://patchwork.data-imaginist.com/

[3] Quarto Reference. Retrieved from https://quarto.org/docs/reference/formats/pdf.html