A Jewel Worth Everything to Some! - Stat228 Mini Project 1

Jolie Liu

Contents

Introduction

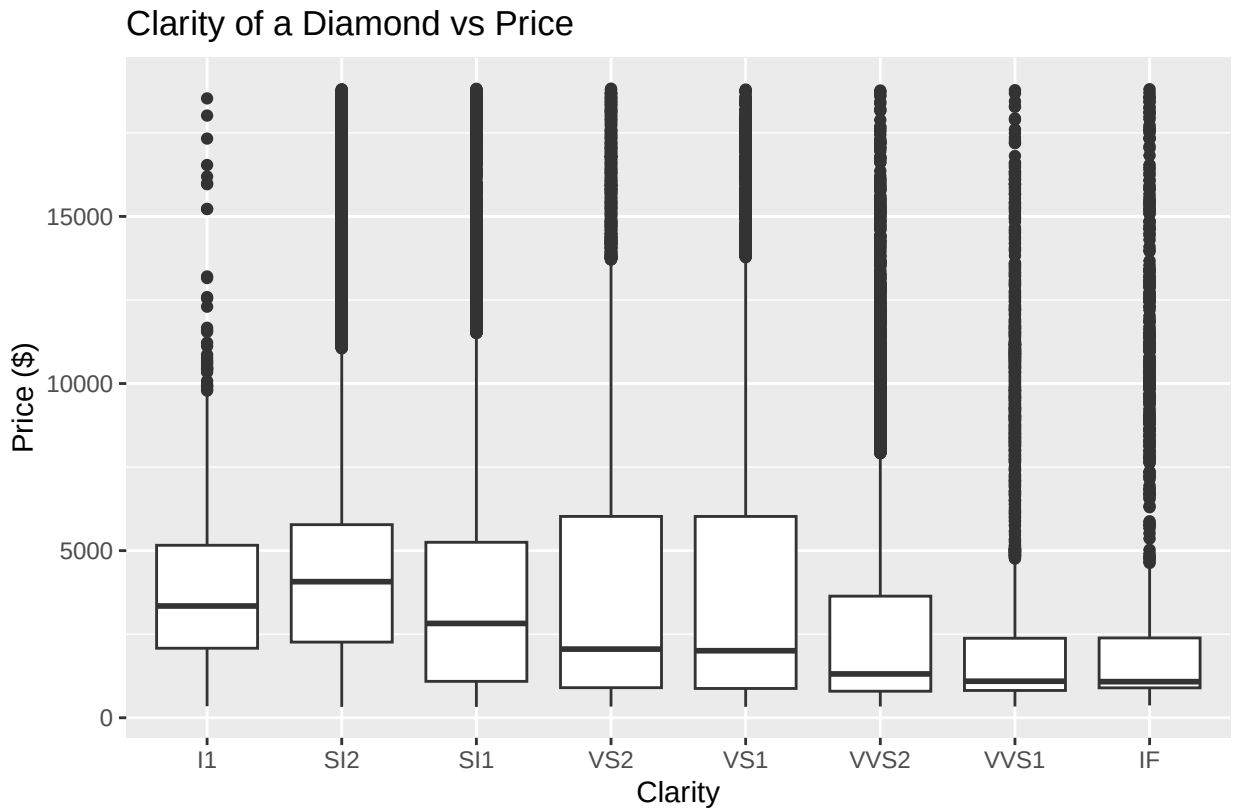
Diamonds, often referred to and implied as one of the most valuable and expensive objects in the world: if you were thirsty in a desert, would rather have a bottle of water or a diamond? Although we know abstractly that diamonds are expensive, I wonder how that specifically plays out comparing different diamonds to one another. What variable(s) are most correlated with the expense of a diamond compared to each other? In the data set "diamonds," we seek to explore the answers to these questions. This data set from R contains 53,940 round cut diamonds and their attributes in the variables: carat (weight), size (in variables x for length,y for width and z for depth), clarity, color and price of each diamond.

Exploration

Diamond Clarity vs Price

The first relationships I want to explore is clarity of a diamond and price. I would assume that diamonds that have a higher level of clarity would be correlated with a higher price since those are attributes of a diamond that make the jewel more desirable by customers.

```
ggplot(diamonds, mapping = aes(x = clarity, y = price))+geom_boxplot() +
  labs(x = "Clarity", y= "Price ($)" )+
  labs(title = "Clarity of a Diamond vs Price",
       caption = "Worst to best clarity is left to right. ")
```



Worst to best clarity is left to right.

```
favstats(price~clarity, data = diamonds)
```

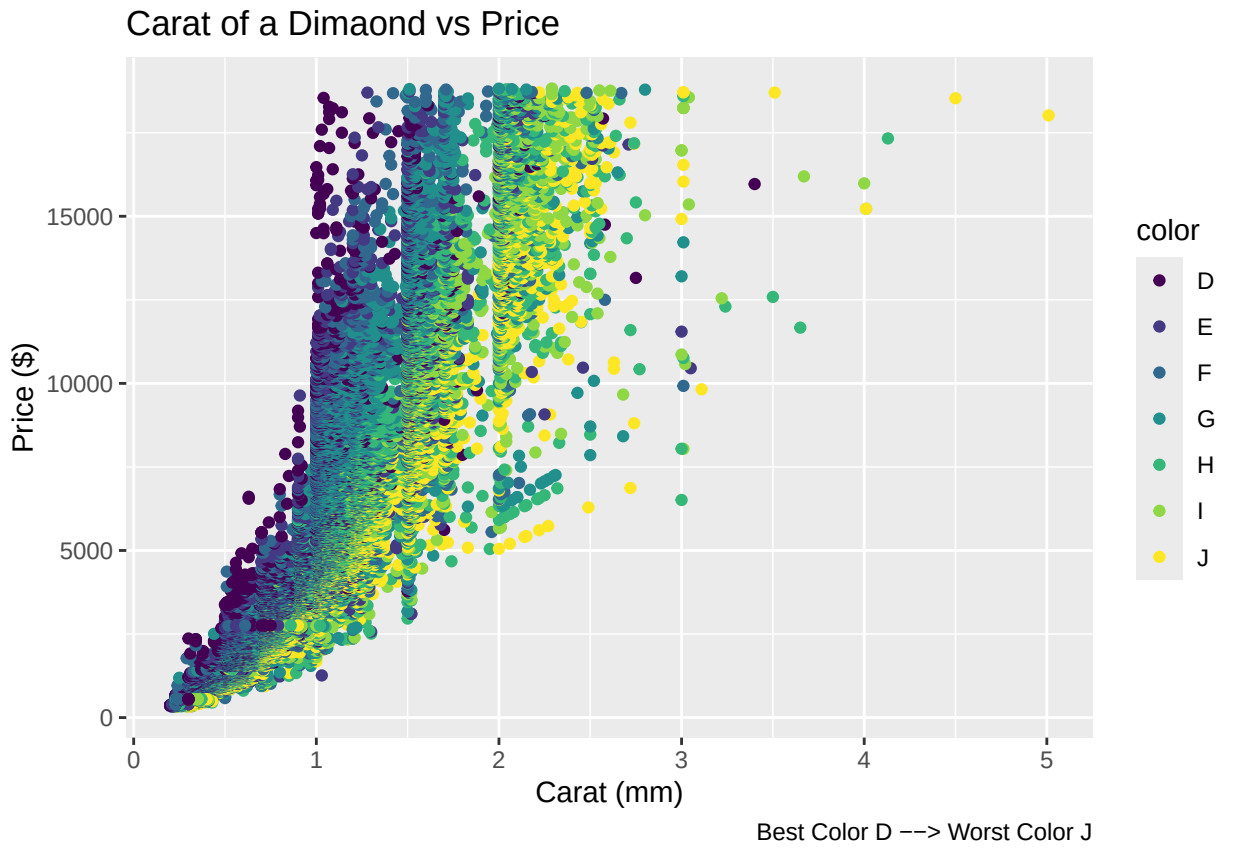
```
##   clarity min      Q1 median      Q3   max     mean       sd     n missing
## 1      I1 345 2080.00   3344 5161.00 18531 3924.169 2806.778   741       0
## 2     SI2 326 2264.00   4072 5777.25 18804 5063.029 4260.459  9194       0
## 3     SI1 326 1089.00   2822 5250.00 18818 3996.001 3799.484 13065       0
## 4     VS2 334  900.00   2054 6023.75 18823 3924.989 4042.303 12258       0
## 5     VS1 327  876.00   2005 6023.00 18795 3839.455 4011.748  8171       0
## 6    VVS2 336  794.25   1311 3638.25 18768 3283.737 3821.648  5066       0
## 7    VVS1 336  816.00   1093 2379.00 18777 2523.115 3334.839  3655       0
## 8      IF 369  895.00   1080 2388.50 18806 2864.839 3920.248  1790       0
```

The above graph is a box plot. Each category on the x-axis is the level of clarity from best to worst (right to left) and show median, IQR, outliers and fences of prices organized by the clarity. I chose to create a boxplot because we have a categorical variable and numerical variable and it would be useful to see what the price was based on clarity to highlight the relationship between them. What's interesting about this graph is the median of each clarity seems to generally drop lower and lower as we move right to left. This indicates that the better the clarity, the lower the price which is the opposite to my guess! The boxes are very uneven which could indicate a skew, but even when running for the average and the median, the higher the clarity, the lower the price. Perhaps it indicates that there is a confounding variable that is causing diamonds with higher clarity to have lower prices.

Diamond Weight/Carat and Color vs Price

Let's explore further other variables: carat, color, and price.

```
ggplot(diamonds, mapping = aes(x = carat, y = price))+
  geom_point(aes(color = color))+
  labs(title = "Carat of a Dimaond vs Price",
       caption = "Best Color D --> Worst Color J" ) + labs(x = "Carat (mm)",
                                                           y = "Price ($)")
```




```
cor(price~carat, data = diamonds)
```

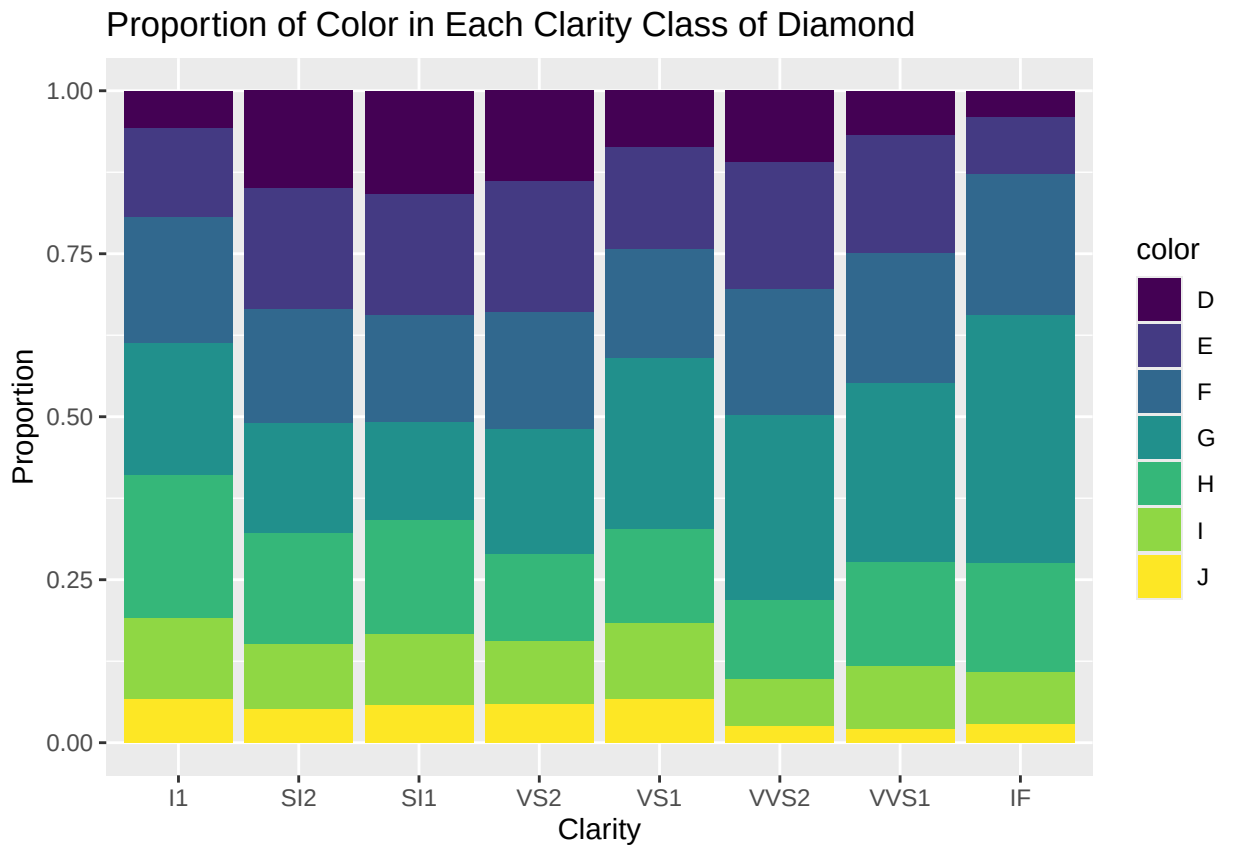
```
## [1] 0.9215913
```

A scatter plot was a good fit for this data because our variables are both numerical. The price goes on the Y axis because it's our dependent variable and Carat goes on the X axis because it is the independent variable. The variable color is shown by the colors to highlight how the spread of color is depending on weight giving as a categorical variable. Based on the scatter plot of this graph, we can see that Carat and Price seem very correlated, and running for the correlation coefficient we get 0.92. Being so close to 1, this shows that this relationship between carat and price is positive and strongly correlated, meaning as carat increases, so does the price. This plot's points are colored according to the color of the diamond and shows that color is positively correlated with a diamonds price as well. As the weights differ, the most expensive out diamonds out of a particular weight seem to generally be the ones with better color. Perhaps these are the confounding variables that are impacting the negative correlation between clarity and price? Let's check the correlation between color & clarity and carat & clarity.

Color and Diamond Weight/Carat vs Clarity

Checking the correlation between color & clarity and carat & clarity:

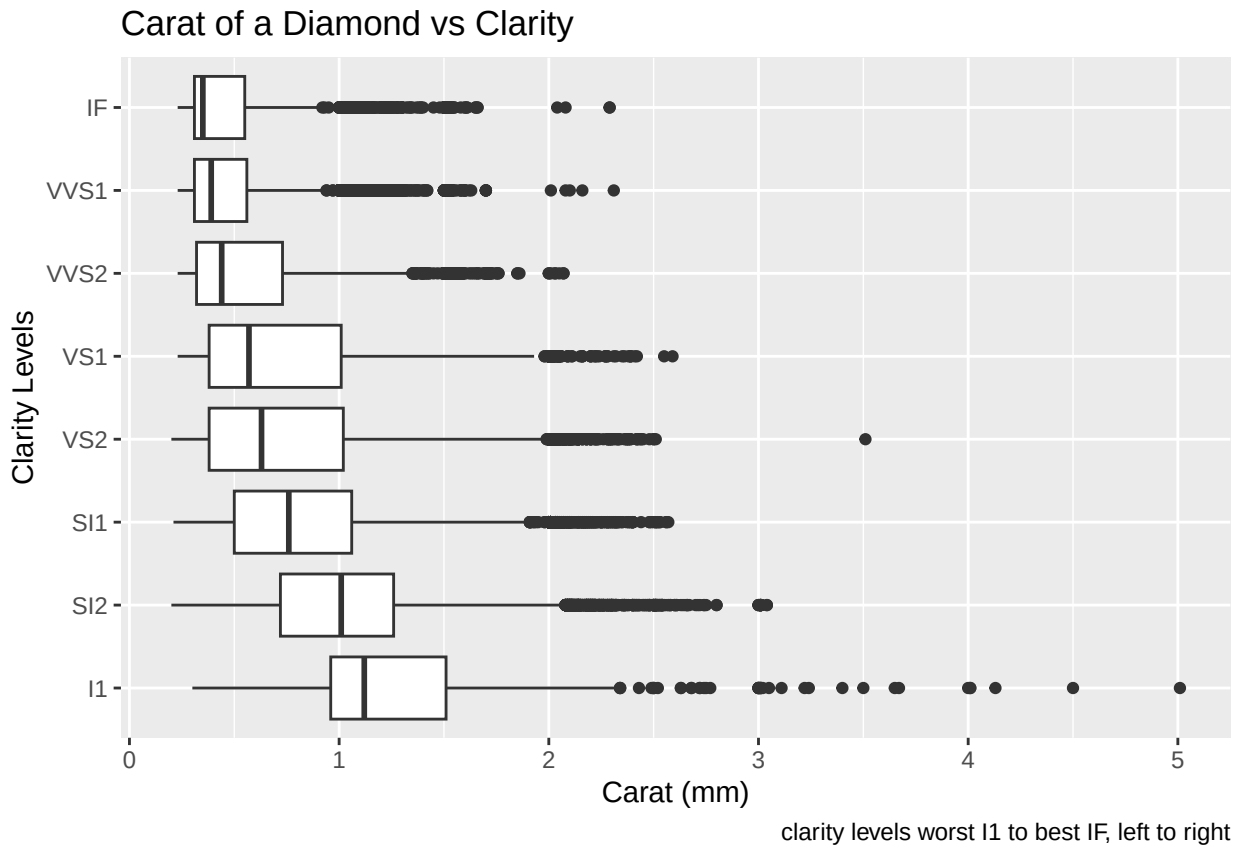
```
ggplot(diamonds, mapping = aes(x = color)) +
    geom_bar(aes(x = clarity, fill = color), position="fill") +
  labs(title = "Proportion of Color in Each Clarity Class of Diamond")+
  labs(x = "Clarity", y = "Proportion")
```



This is a segmented bar chart that has been rescaled to show proportion of color out of one.The segmented barplot was chosen with the segments as colors because it helps illustrate that some low clarity classes such as SI2 and SI2 have a higher proportion of higher color (D and E) and a lower proportion of lower colors(G) compared to other clarity class by comparing proportions rather than just number of different colors in different clarities. Although we see this difference, all in all, the proportions don't really stray that far from one another. They seem pretty even.

Let's continue exploring for confounding variables:

```
 ggplot(diamonds, mapping = aes(x = carat, y = clarity))+geom_boxplot() +
labs(title = "Carat of a Diamond vs Clarity",
     caption = "clarity levels worst I1 to best IF, left to right")+
labs(x = "Carat (mm)", y = "Clarity Levels")
```

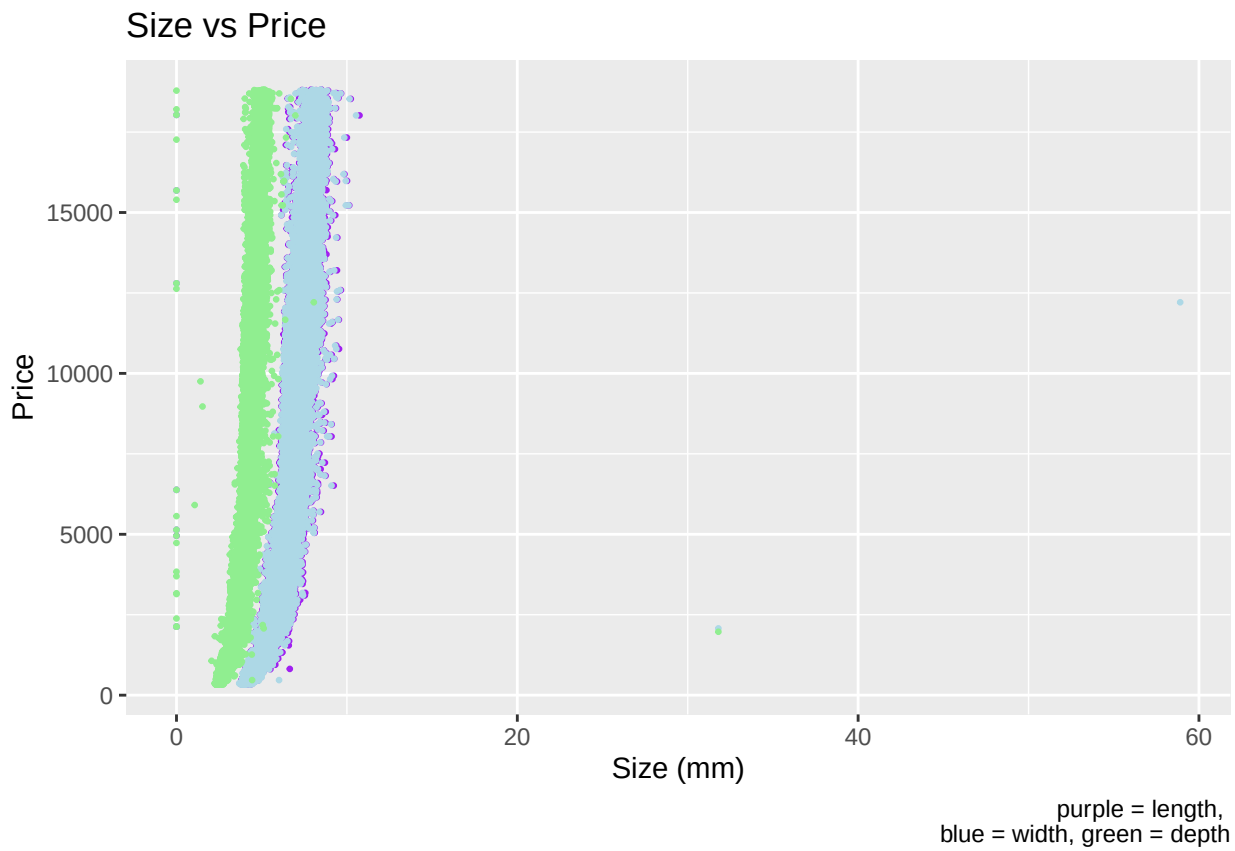


This bar chart was chosen because it shows the spread of the weight of diamonds (numerical variable) in each clarity level(categorical). It seems like the best clarity levels have lower weights and generally shows a correlation: the higher the clarity class, the lower the weight. It seems like we have found our confounding variables!In the first graph, which seemed like it could possibly indicate lower clarity causes higher price was actually a manifestation of the strong correlation of diamond weight and price. Those with lower clarity class had higher weighted diamonds, and therefore would have higher prices than those with higher clarities which had lower weights.

Exploring Diamond Size vs Price

Now that we know weight is the confounding variable and that it has a meaningful strongly positive correlation with price, let's see what dimension exactly is correlated to price the most. The more a diamond weighs, the bigger the diamond is, so we can check dimensions such as length (x), width(y), and depth(z) because we expect these sizes to positively correlate with price as well.

```
ggplot(df_1, aes(x = size, y = price)) +
  geom_point(color = "purple", size = 0.5)+
  geom_point(data = df_2, color = "lightblue", size = 0.5) +
  geom_point(data = df_3, color = "lightgreen", size = 0.5)+
  labs(title = "Size vs Price", caption = "purple = length,
       blue = width, green = depth")+ labs(x = "Size (mm)", y = "Price" )
```



In this Size vs Price Scatter plot, we are comparing two numerical variables, and by choosing these variables, we can see their correlation easier. The different dimensions are different colors so that we can differentiate how each dimension may be correlated differently to price. Here we can see the blue and purple almost completely overlapping. This should make sense as they represent length and width. The diamonds in this data set are round cut, so their lengths and widths should be extremely similar. It seems like the depth (in green), length (blue) and width (purple) are all also positively correlated to the price.

```
cor(price~x, data = diamonds)
```

```
## [1] 0.8844352
```

```
cor(price~y, data = diamonds)
```

```
## [1] 0.8654209
```

```
cor(price~z, data = diamonds)
```

```
## [1] 0.8612494
```

They all have similar correlation coeffecients and therefore show similarly strongly positive correlations!

Conclusion

We conclude that the price of a roundcut diamond is strongly positively correlated mostly with it's size and weight - mostly that a higher weight is strongly positively correlated with price - but also a being higher in any of it's dimensions in length, width and depth are also strongly positively correlated with a higher price. At first it seemed like perhaps diamond price was negatively correlated with clarity, but upon further inspection to the spread of weight based on clarity of diamonds, it seems to actually illustrate and bolster evidence of the strongly positive correlation between size & weight and diamond price.