Data Visualization Insights from Diamonds Datasets

Asa B.

2023-07-02

Introduction

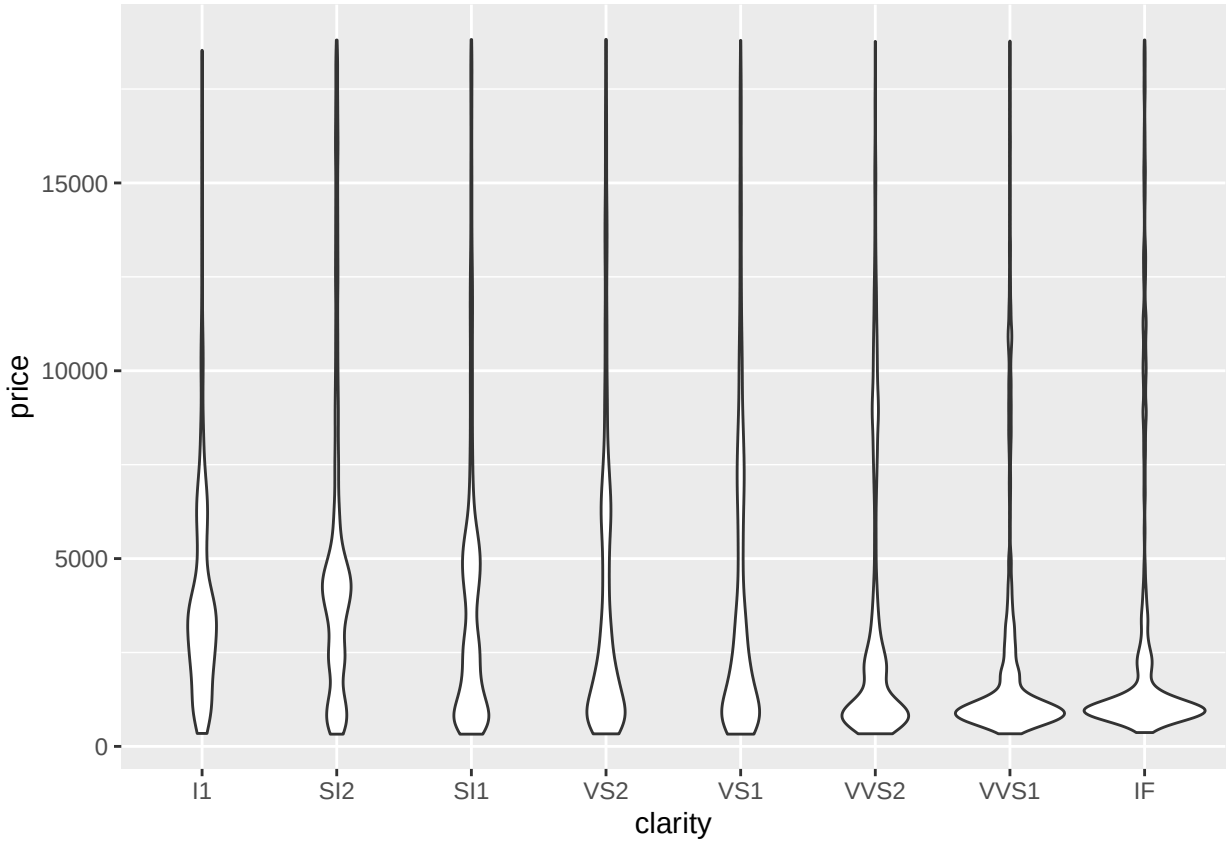
In this assignment, we have to use 'diamonds', which is R's built-in datasets, to find out the patterns underlying the datasets. With the comprehensive knowledge from Data Science Bootcamp class, here're the questions and the results.

Questions and results

Question 1: distribution of price, depends on clarity

Before proceeding with this questions, we have to now that there's diamond clarity grading systems, which is from 'FL' (flawless) to 'I' (Included) grade. We assumed that the FL grade should have the higher price point than the inferior ones. To check out if the assumption made is true, we use the violin chart to answer the questions since the chart has the ability to show the price distribution for each group. The illustration shows something interesting.

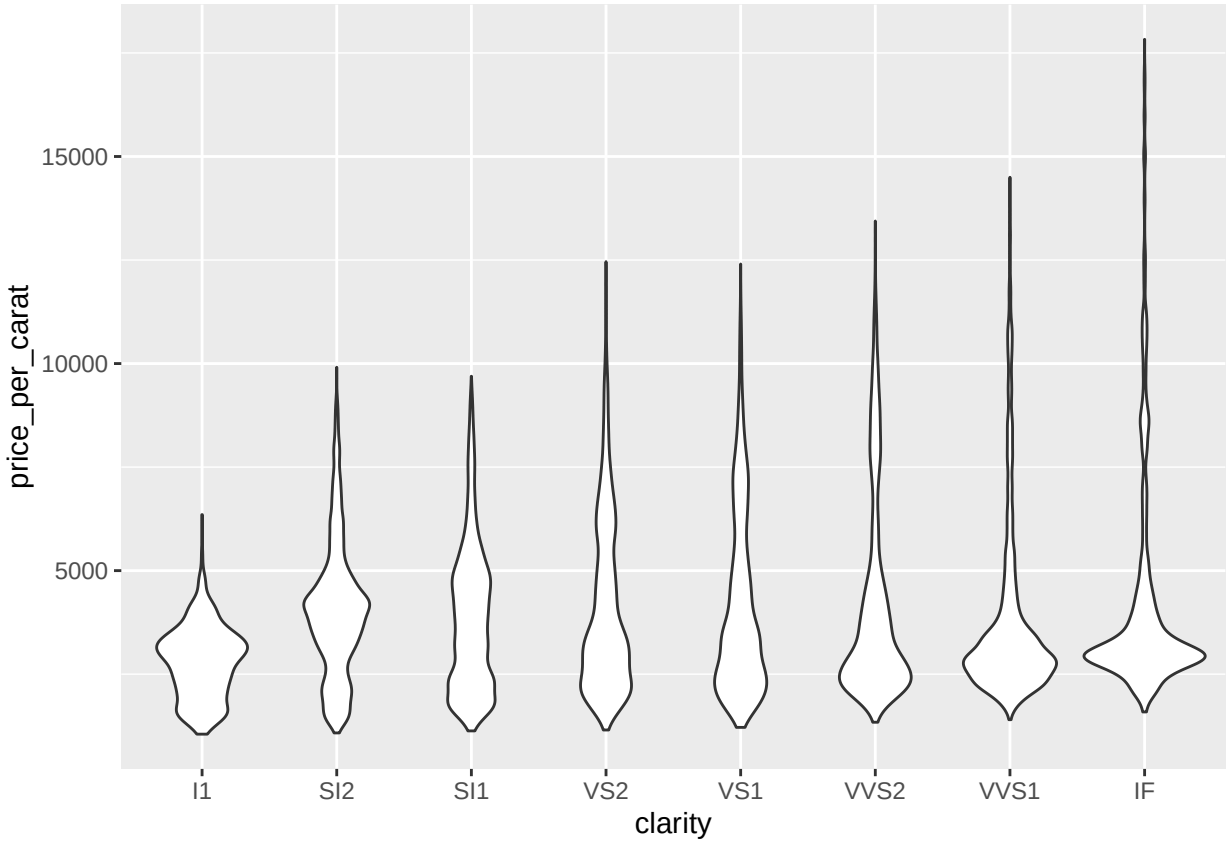
```
# question1: price distribution by clarity
  ggplot(diamonds,aes(clarity,price)) +
    geom_violin()
```



The chart illustrated shows that there's a chance that the inferior grade of clarity could be sold by the higher price point. So, the price of the diamonds is primarily defined by the other factors. We suspects that size (literally it's weight, measured in carats) should be the primary factor so we will find out the distribution of 'price per carat' to the clarity greade, wihch we will calculated as r script as follows.

```
## calculate the price per carat of each diamonds
diamonds <- diamonds %>%
  mutate(price_per_carat = price/carat)

## build the violin chart again using clarity and price per carat
ggplot(diamonds,aes(clarity,price_per_carat)) +
  geom_violin()
```

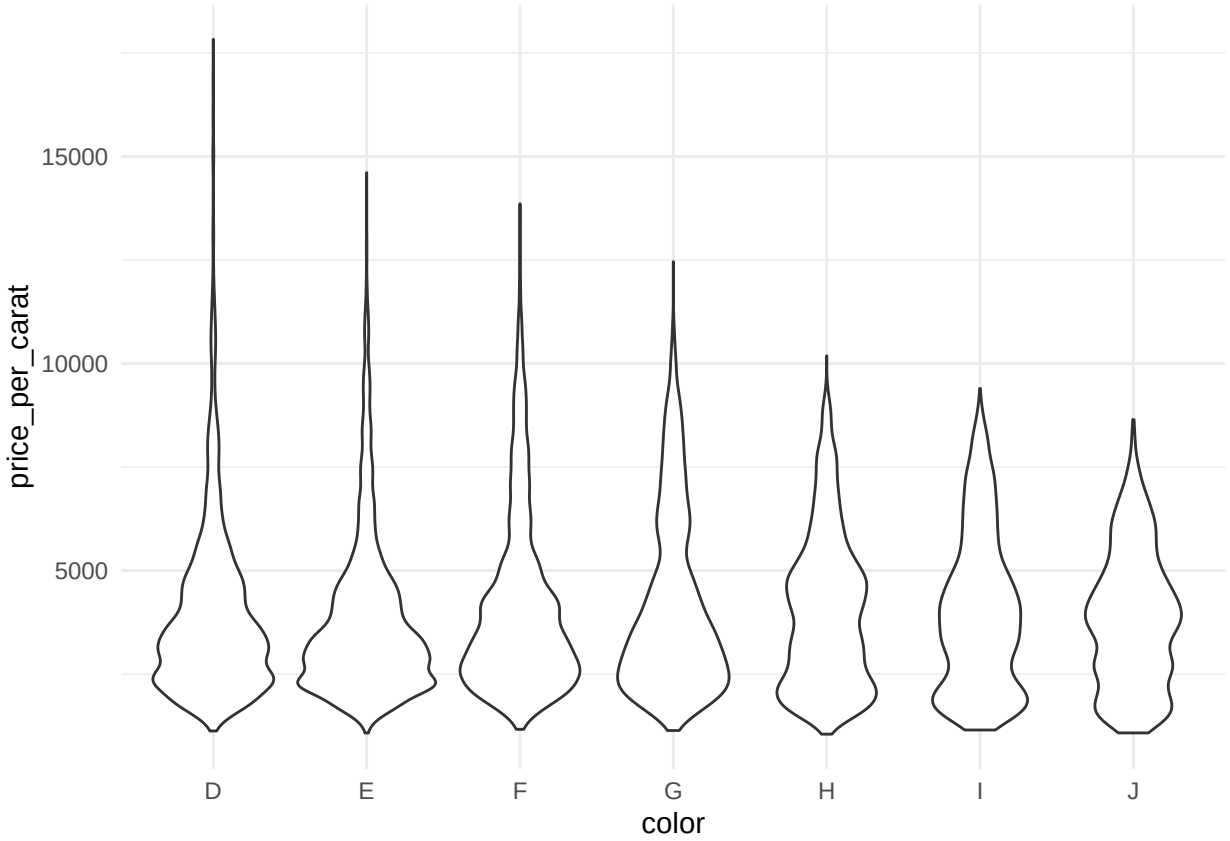


This chart shows that there's more define price per carat for each diamond in higher grades, which we could see from the large clusters of price points around 2500$ per carat mark for the 'IF' grade of diamonds. On the other hand, the price distributions of the inferior grade of diamond clarity if more vary than the higher ones. Moreover, the distribution shows there's a chance that the inferior clarity diamonds has more expensive price per carat.

Question 2: distribution of price, depends on the color of diamonds

The Jewelry industry defines the gradeing system of the diamonds color alplabetically from 'D' (colorless) to 'Z'(light yellow). We will compared the price per carat of each diamond then illustrated the price distribution by color grade in a violin chart. The results is as follows.

```
# question2: price distribution by color
ggplot(diamonds,aes(color,price_per_carat)) +
  geom_violin() +
  theme_minimal()
```

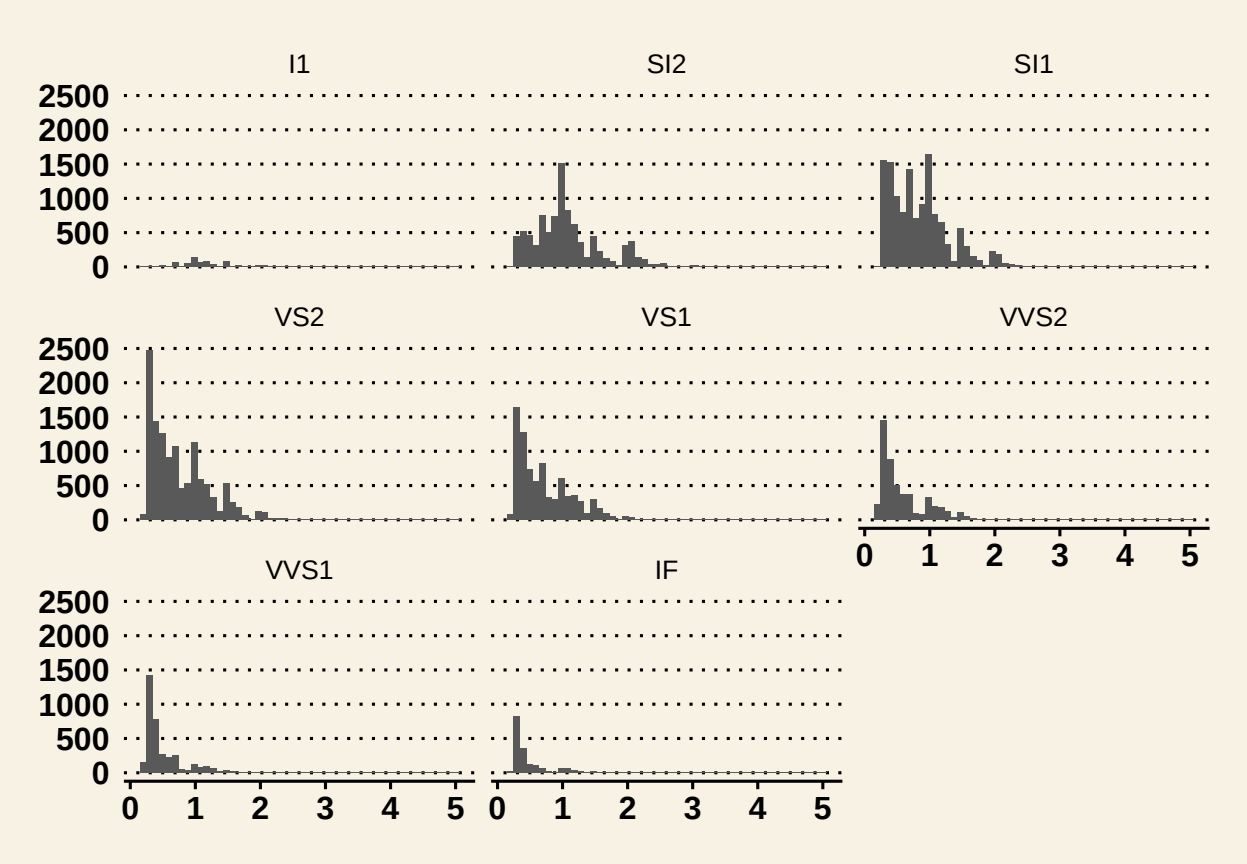


As illustrated, the violin plot tends to show distribution of price per carat as assumed. The clearer the color, the more expensive. Noted that the inferior grade, like I and J for instance, has the more clustered price point than the higher grade ones.

Question 3: Relationship between clarity and size

Based on question1, we're suspected if it's possible that the industries might have to decided to keep the diamond size over their clarity (bigger the diamond, lower the clarity). We will built the multiple histogram chart shows the count of diamond's size by the clarity grade. We try to mimic the 'Wall Street Journal (WSJ)' style of charts.

```
# question3: Relationship between clarity and size
ggplot(diamonds,aes(carat)) +
  geom_histogram(bins=50) +
  facet_wrap(~clarity) +
  theme_wsj()
```



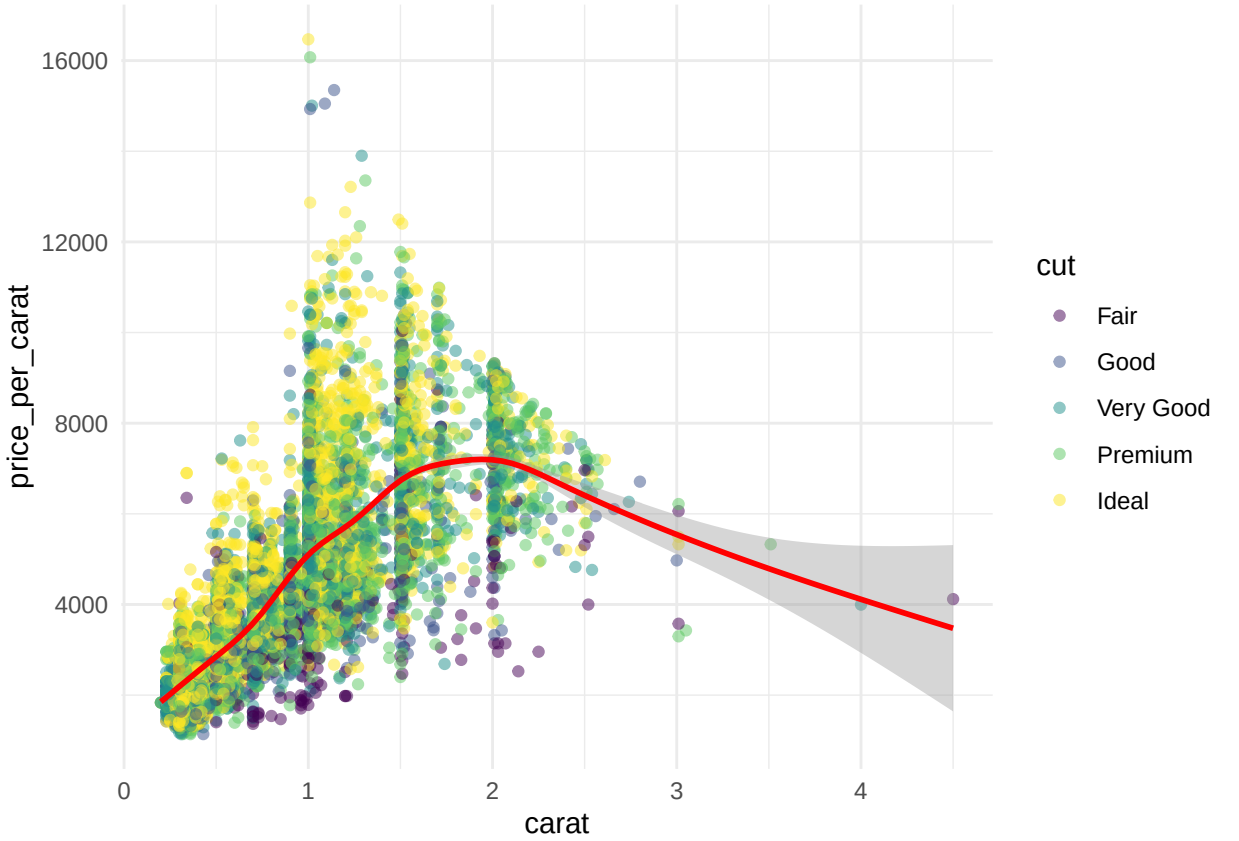
The results shows that the assumptions made is 'reasonable' since we found that the inferior clarity grade, like SI, has the more vary size distribution. Whilst the higher grade of dismonds tends to have the smaller size of the diamonds, mostly 0.2 carats each. The results shows that the most inferior grade of samples, the I1 grade, has the very little observations.

Question 4: Relationship between size, and price per carat

It's the fact that bigger the size of diamonds, bigger the price tag. However, if we calculated the price per carat, is it possible that price per carat is bigger.

```
# question4: Relationship between size and price per carat
ggplot(sample_frac(diamonds,0.2),
       aes(carat,price_per_carat,color=cut)) +
  geom_point(alpha=0.5) +
  geom_smooth(col="red") +
  theme_minimal()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

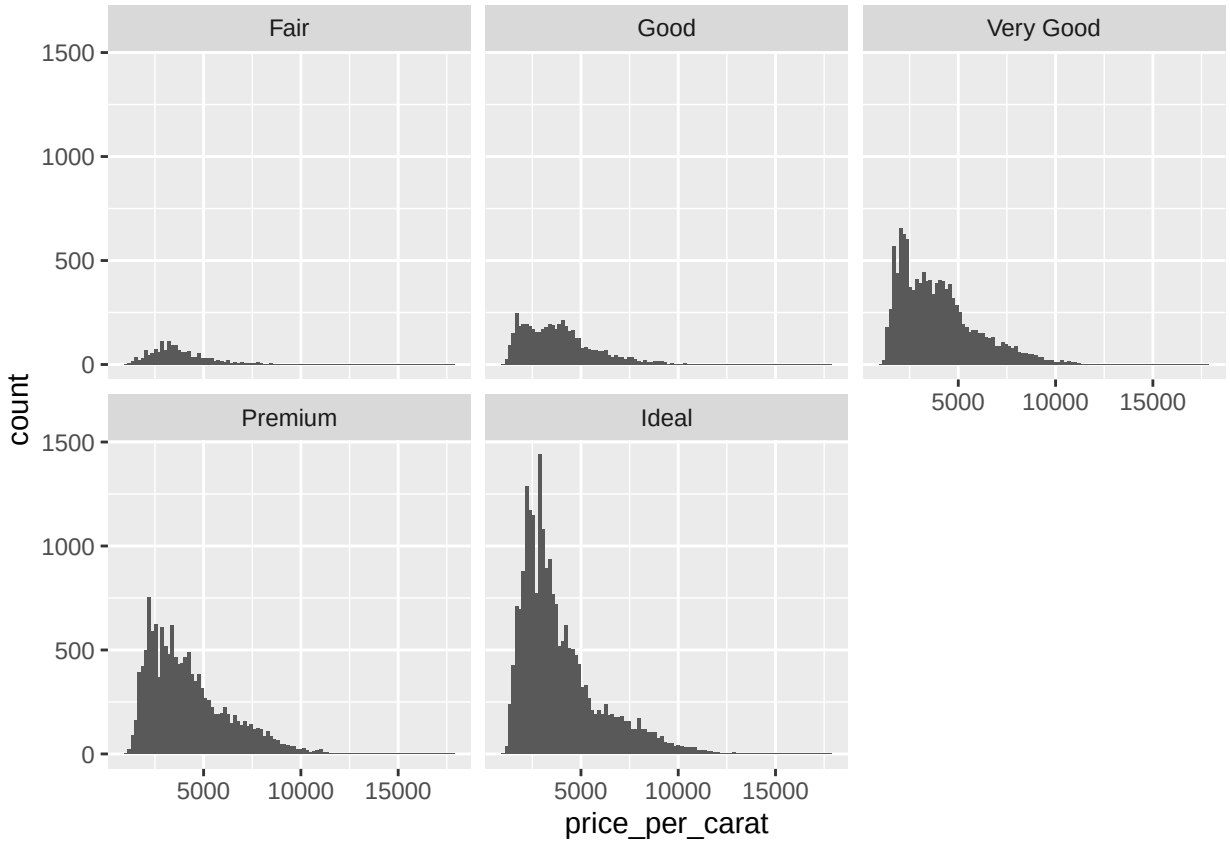


The results shows that when the diamond size is bigger, the size per carat tends to be more expensive. However, based on the sampling we've got, it shows that price per carat tends to be cheaper when the size is pass 2 carat mark.

Question 5: Most marketable cut, and their price points

From the perspective of entrepreneur, they might need to know what options of diamonds is most marketable. In this study, we will find out the most marketable options from price per carat based on the cut.

```
# question5: Most marketable price points
ggplot(diamonds,aes(price_per_carat)) +
  geom_histogram(bins=100) +
  facet_wrap(~cut)
```



The results shows that the ideal price points should be around 2,500$ per carat, the price point is slightly vary depends on the cut, but still on the same price marks.