Visualizing the Diamonds Data Set

Janna Mangasep

2022-08-26

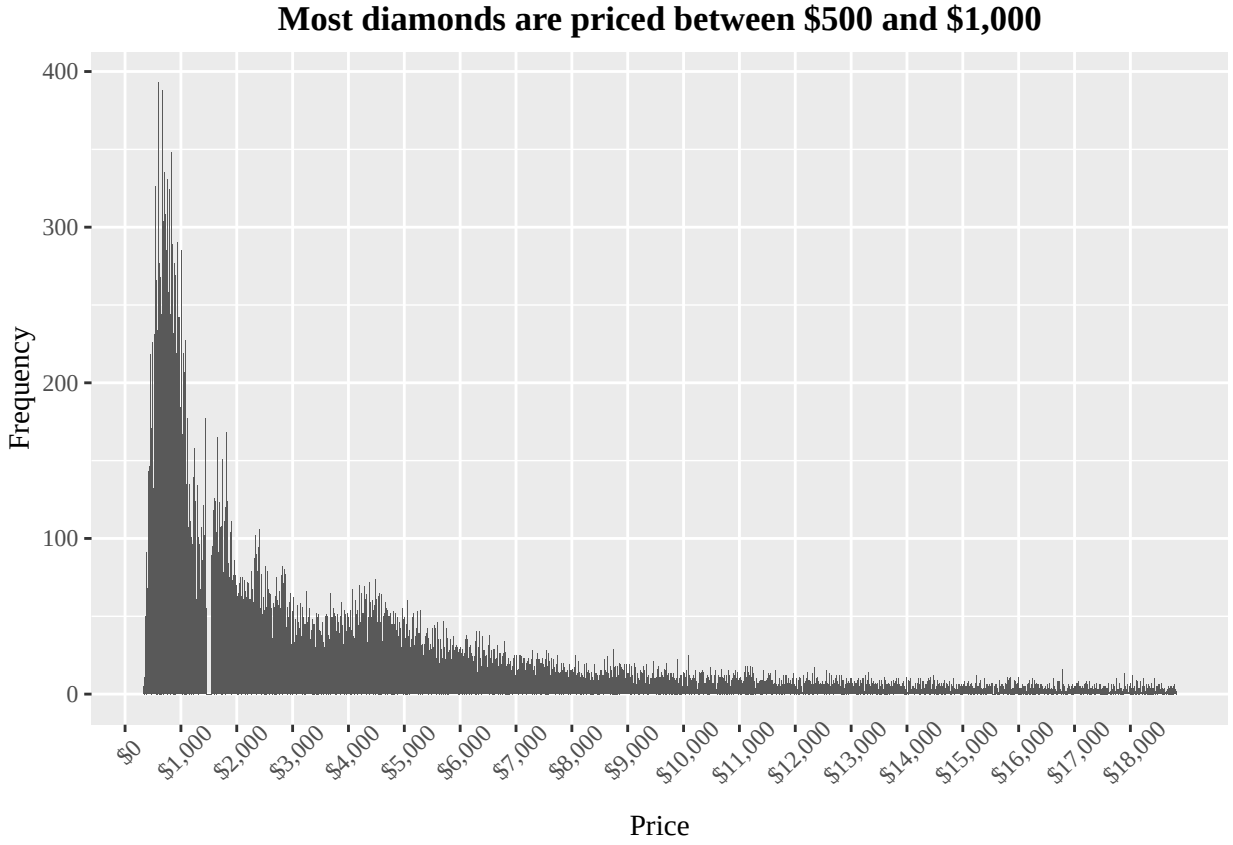
Document Overview

For context, this R Markdown file revisits the work I previously did for my UChicago Harris Class "Data & Programming in R I." All work and outputs are my own and are simply guided by the questions from the assignment. This document aims to explore both variation and covariation in the `diamonds` data set, which is built-in from the `ggplot2` package within `tidyverse` and contains information on 53940 diamonds.

Exploring Data Variation

Price

To begin with, I examine the distribution of the `price` variable within `diamonds` through a histogram. Before doing so, I expected the distribution to be concentrated around the lower dollar prices. As a note, the maximum price in the data is rather extreme at 18823 dollars.

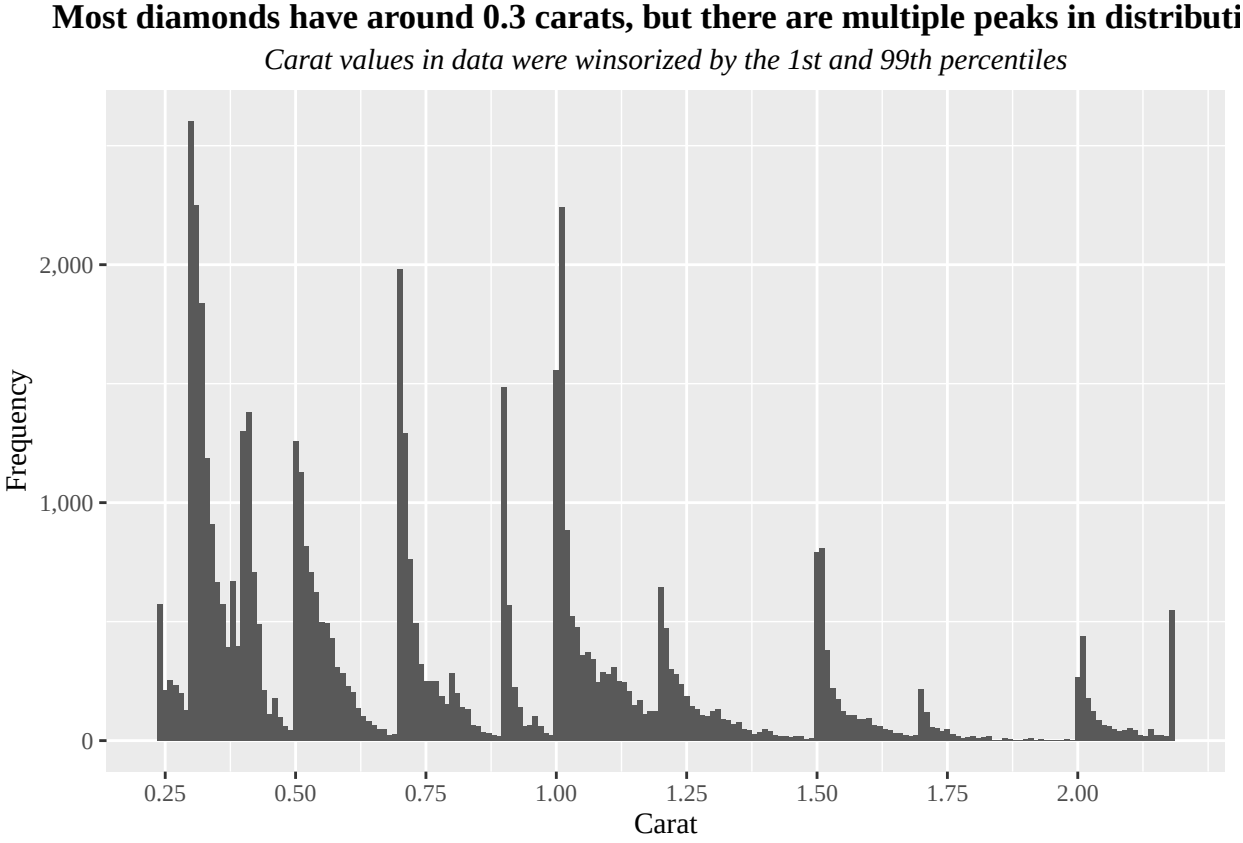


From the histogram above, there are a couple of things to note. First, most observations in `diamonds` have a `price` value between ~500 to ~1,000 dollars (i.e. it is *right-skewed*), and this aligns with my expectations. Also, there is a noticeable lack of observations around the 1,500 dollar mark, which is highly unusual as this is the only clear break in the distribution.

Carat

For the `carat` variable, I expect most of the observations to be concentrated around the lower values. This is because the previous exploration on `price` found most observations to be priced between 500 to 1,000 dollars. Therefore, assuming diamonds' prices are associated with their carats, most diamonds in the data are likely to have lower carats.

Though my predictions for `carat` are similar to those for `price`, it is important to note that the possible values are very different between the two variables. While the latter ranges from 326 dollars to 18823, the former has a much smaller range of 0.2 carats to 5.01 carats. Therefore, I take precautions for proper scaling by winsorizing this variable (i.e., replacing carat outliers with specified quantiles). Please see this documentation page for more information on the `winsorize()` function from `statar`.



The most diamonds in the data appear to have around 0.3 carats. However, there are various, albeit smaller, peaks at other `carat` values (e.g., at 0.5 carats, ~0.7 carats, 1.0 carats, etc.). This does not fit with my expectations, as I expected this to be a smoother distribution, specifically a right-skewed distribution like with the variable `price`. Considering the strong concentration around *lower* prices (meaning from $500 to $1,000 dollars), this is a surprising find as there are variations within different ranges of carat values.

To better understand this distribution, I look into a sample of the peaks 0.5, 1.0, 1.5, and 2.0 within the entire `diamonds` data set (not winsorized), which I refer to as the "carat peaks". Specifically, I compare the frequency of diamonds at each peak and the frequency of diamonds at the carat value immediately before

each peak (meaning 0.49, 0.99, etc.) to informally test whether sellers round up diamonds' carat values to increase prices.

Table 1: Diamonds per Carat Peak

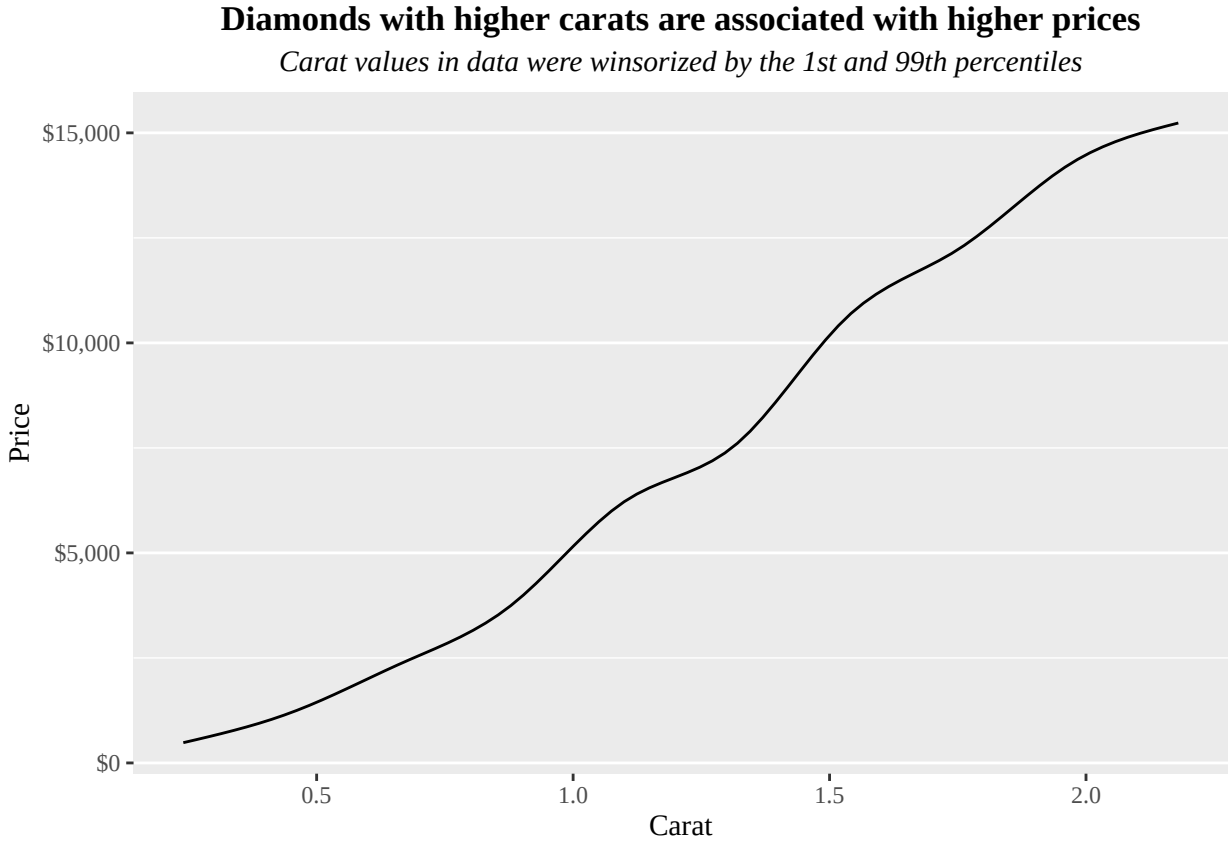
carat	Frequency
0.5	1258
1.0	1558
1.5	793
2.0	265

Table 2: Diamonds per Carat Immediately Before Peak

carat	Frequency
0.49	45
0.99	23
1.49	11
1.99	3

From the tables above, it appears that significantly less diamonds have an "un-rounded" listed carat value. For example, only 23 diamonds are 0.99 carats, but 1,558 diamonds are 1 carat! Again, this discrepancy may come from diamond sellers rounding up the carat to sell diamonds for a greater profit. To investigate this further, I create the plot below to examine the relationship between `price` and `carat`

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



From the plot above, it is apparent that there is a positive relationship between diamonds' carats and their listed prices. Therefore, it is reasonable to assume that diamond sellers would be incentivized to claim that a diamond is of higher carat than it is. This explains the dramatic peaks seen in the previous graph at rounded numbers.

Exploring Data Covariation

After covering variation within the variables `price` and `carat`, I now look into covariation in the data. Upon examining the mean prices associated with each diamond `cut` value (see table below), the assignment implored us to find the most important variable for predicting diamond price in the data set.

Table 3: Average Diamond Price per Cut Type

cut	Mean Price
Fair	$4,358.76
Good	$3,928.86
Very Good	$3,981.76
Premium	$4,584.26
Ideal	$3,457.54

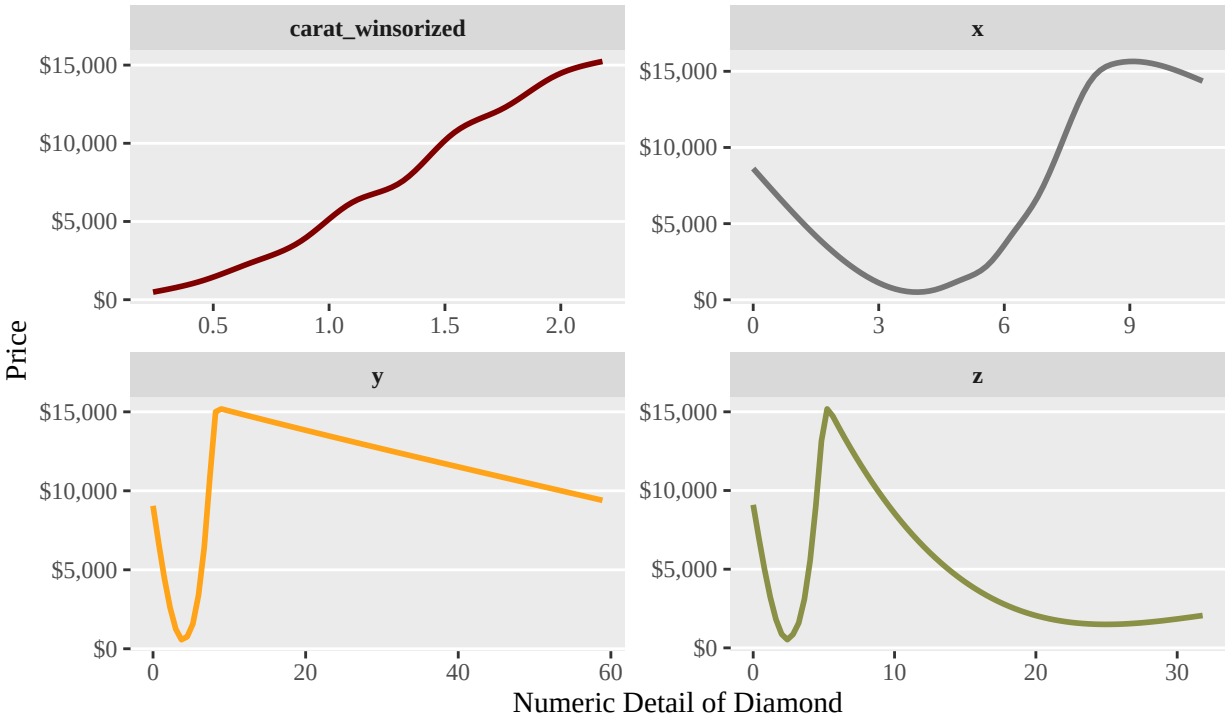
Fair-cut diamonds had the *second-highest* mean price in the data, despite fair being the worst `cut` value for a diamond. This signals omitted-variable bias, so I look into all other candidate predictors for `price` aside from this variable.

With `geom_smooth()`, I plot `price` with the following NUMERIC variables: `carat`, length `x`, width `y`, and depth `z`. I look at these numeric variables as there is a wide range of values in `price` (as seen from the histogram on price variation), so using a categorical variable like `color` or `clarity` will *not* be useful for prediction.

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

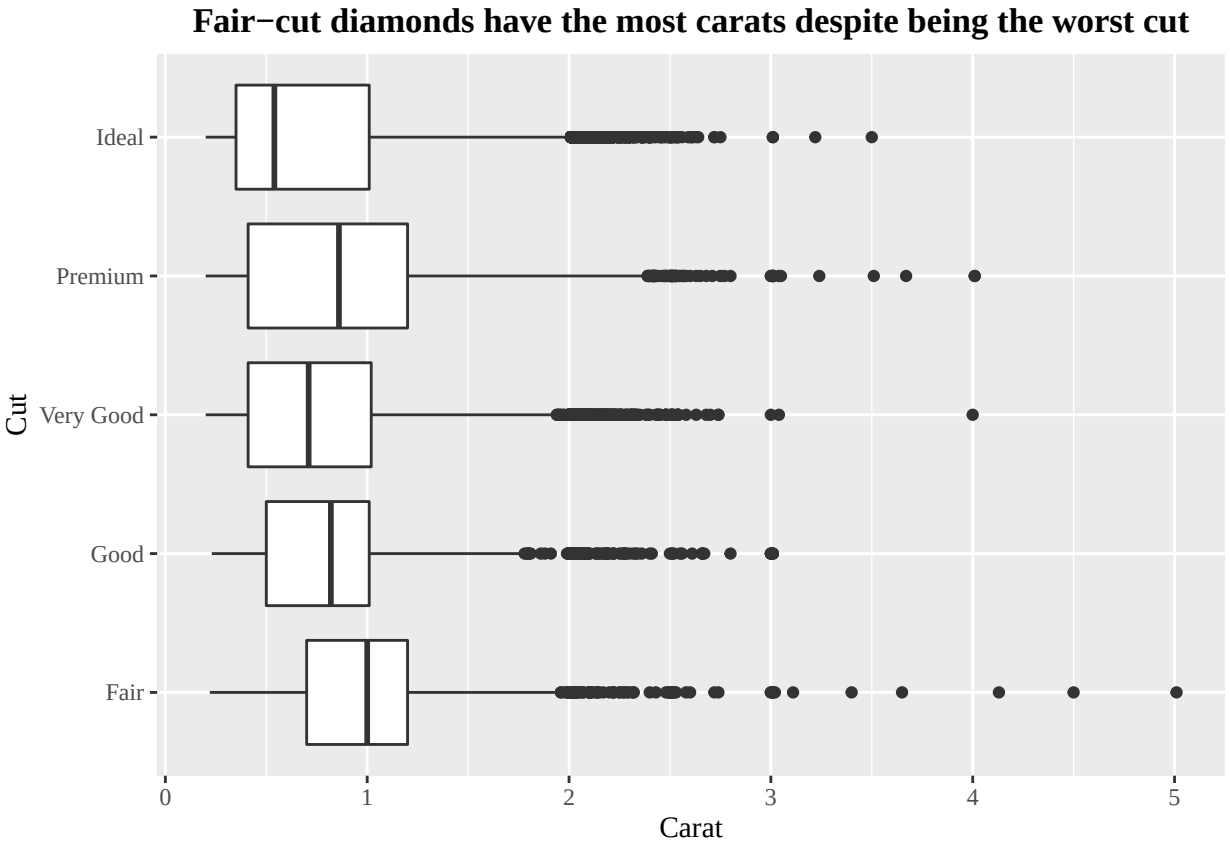
Only carat* holds a significant relationship to the price of a diamond

Length (x), width (y), and depth (z) each have an unclear relationship with price



**Carat winsorized by 1st and 99th percentiles*

From the graphs above, it is apparent that the `carat` of a diamond is the *most* reliable predictor for diamond price. No measurements of the diamond have a clear relationship with diamond price, so I consider width, depth, and length of a diamond as poor predictors. Knowing the significance of the `carat` variable, I look into its relationship with `cut`, as suggested by the assignment.



As seen above, the continuous `carat` variable is correlated with the categorical variable `cut` variable in that *fair* cuts have the *highest* carat values while *ideal* cuts have the *lowest*. Fair-cut diamonds have the highest median carat (1) out of the cut types, AND they have the most variation in *extreme* carat values (i.e. 3 or more carats). Other than premium cut diamonds (which have the second-highest median carat value), it seems that the "better" the cut, the LOWER the carat value.

Final Thoughts on the Table of Average Diamond Prices per Cut

The table is misleading as it omits the crucial variable of `carat`. It implies that "poorer" cut diamonds (e.g. Fair) are actually more expensive than the best cut diamonds (e.g. Ideal). However, this table fails to show that it is *not* the cut that determines `price`. Rather, it is `carat` that determines `price`, and the diamonds in the data happen to have fair-cut diamonds with large carat values.

This is likely due to the fact that `cut` is related to aesthetics (i.e., brilliance and reflection), but an ideal cut does *not* beget a high carat value. Therefore, the table is misleading due to its omission of the crucial `carat` value for a diamond's price.