Visualizing the Diamonds Data Set

Janna Mangasep

2022-08-26

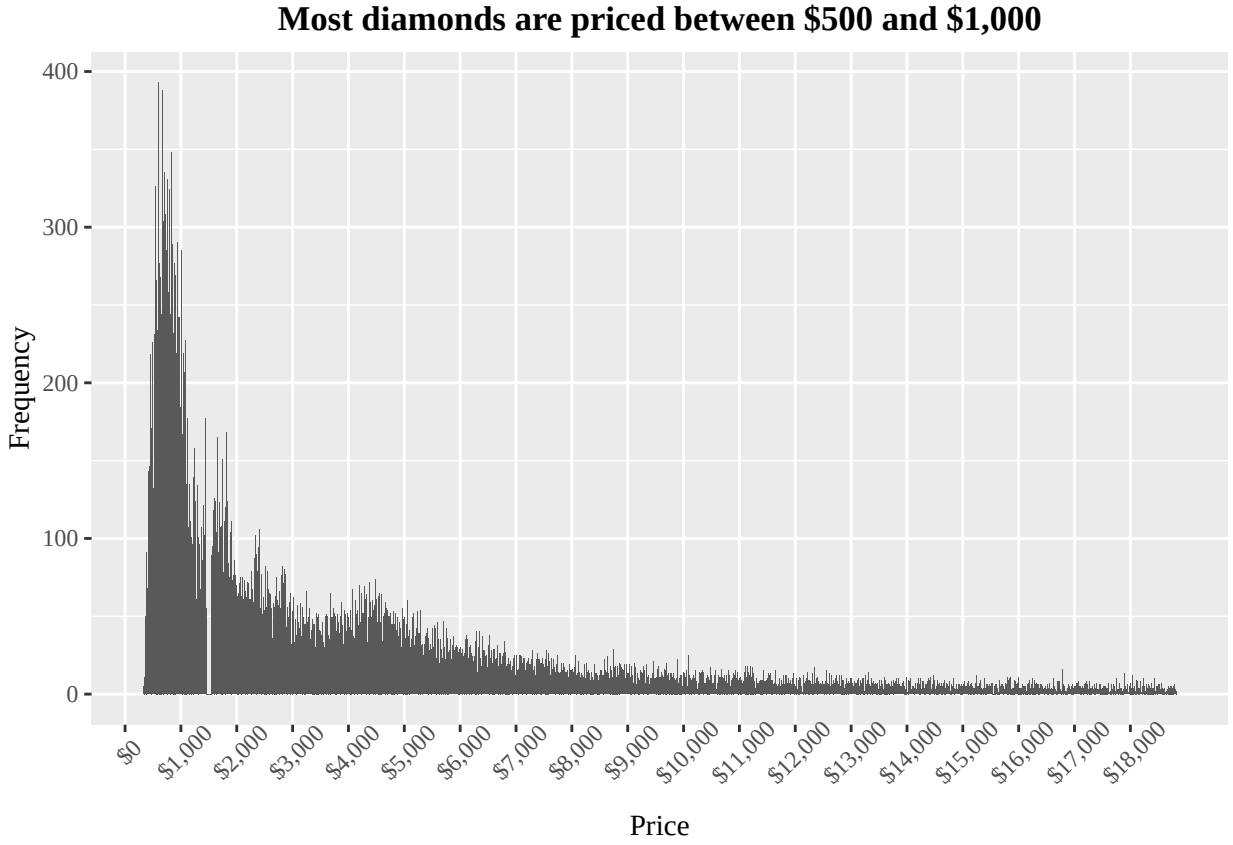
Document Overview

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.

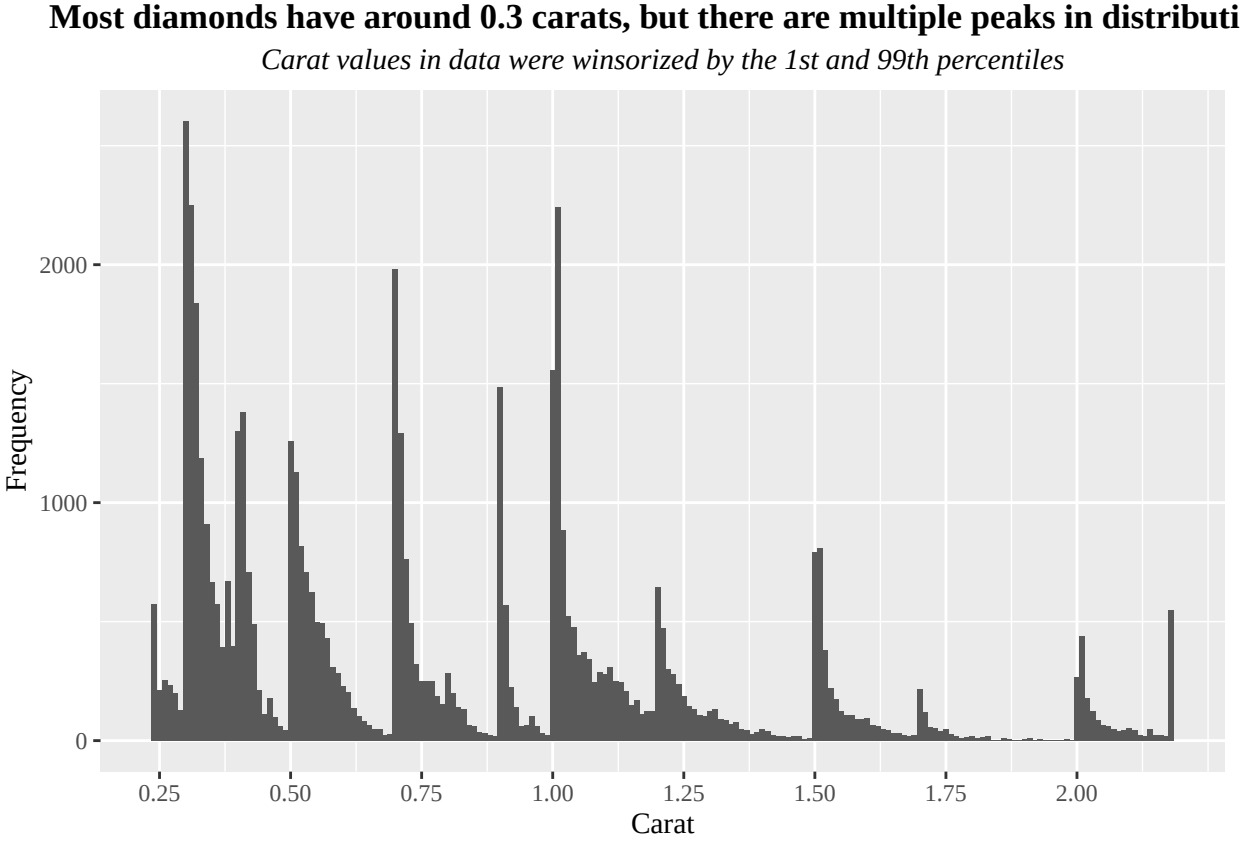


From the histogram above, there are a couple of things to note. First, most observations in `diamonds` have a `price` value between ~500 to ~1,000 dollars (i.e. it is *right-skewed*), and this aligns with my expectations. Also, there is a noticeable lack of observations around the 1,500 dollar mark, 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! This discrepancy may come from diamond sellers rounding up the carat to sell diamonds for a greater profit. I create the plot below to support this argument, finding the relationship between `price` and `carat`, as higher carats meaning higher prices would justify my hypothesis.