Chapter 05: Statistical summaries

Stan Piotrowski

October 11 2021

Contents

Setup

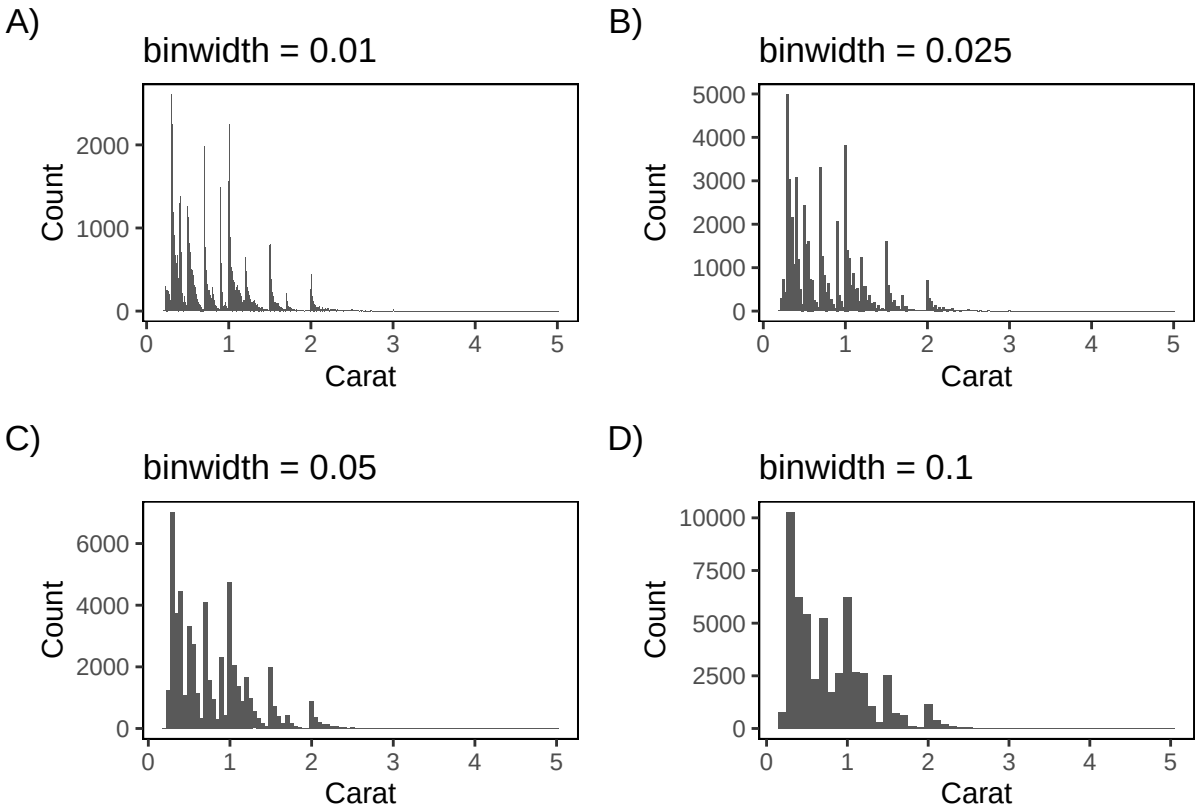
This chapter focuses on displaying statistical summaries of data in geoms to display uncertainty around estimates and distributions.

Exercises

5.4.1 Exercises

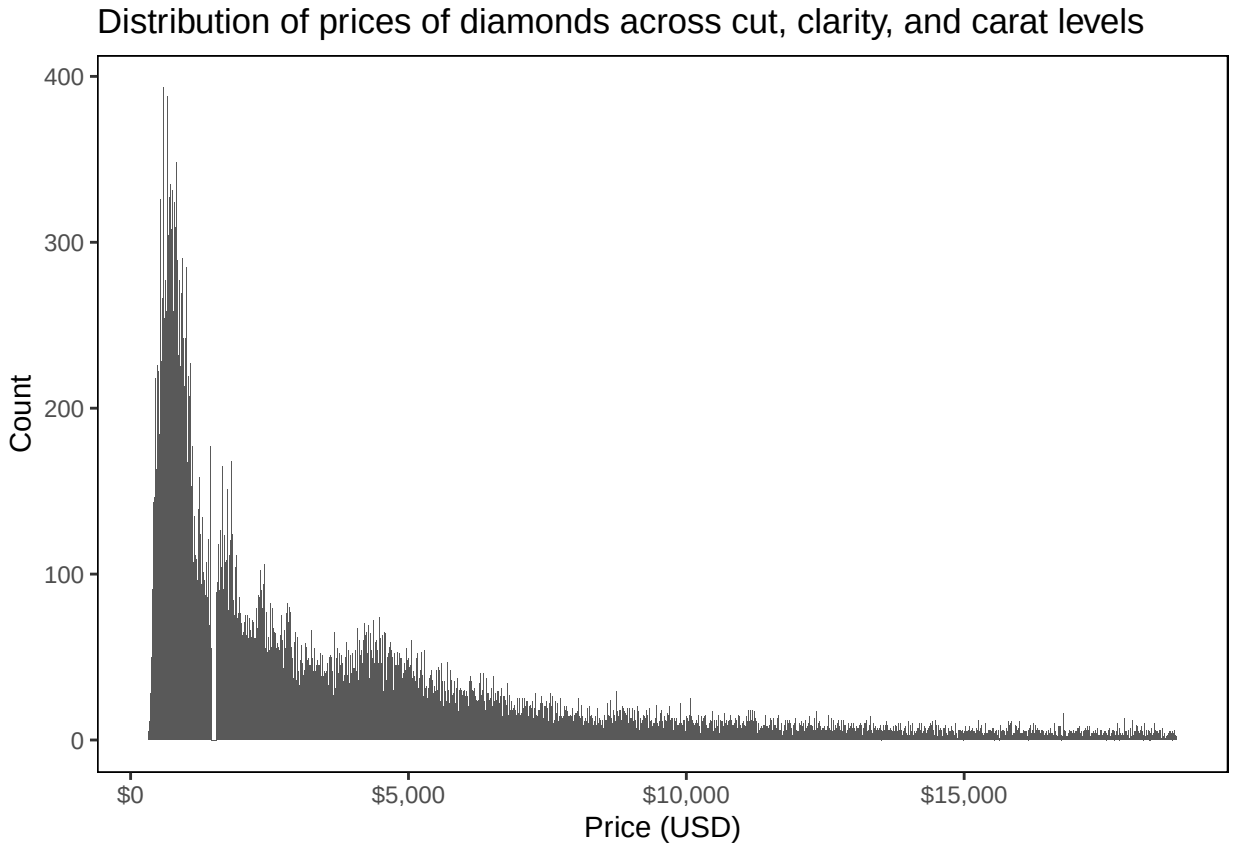
1) *What binwidth tells you the most interesting story about the distribution of* `carat`*?*

Here, I'll create a series of plots and vary the binwidth for each to see how binning the data differently can lead to alternative interpretations. Looking at the series of plots below, we can see that most of the diamonds appear to be around whole number carats.



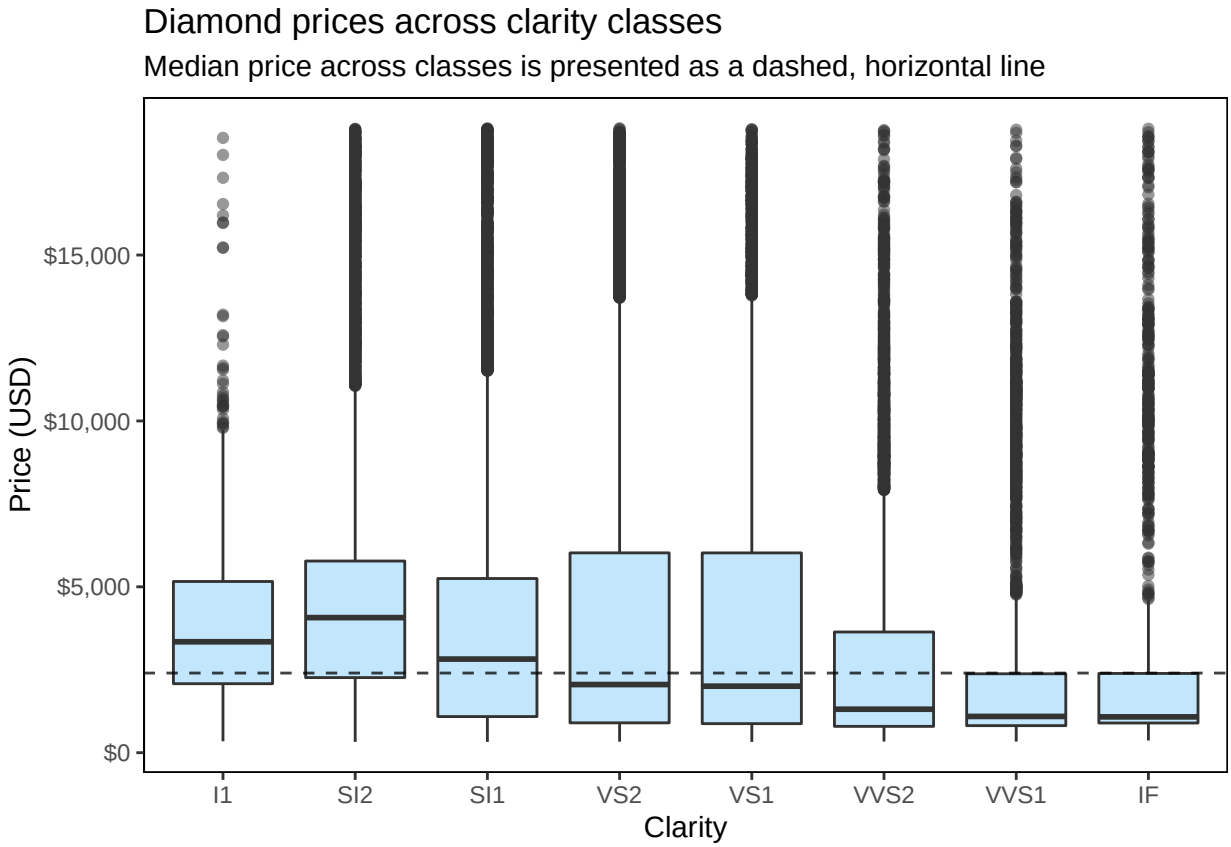
2) *Draw a histogram of* `price`*. What interesting patterns do you see?*

One interesting pattern that arose when visualizing the histogram of `price` is there appears to be a break around $1,500 or so. This could indicate that there simply aren't any scored diamonds at this particular price point, or it could be due to potential quality issues in the data set.



3) *How does the distribution of* `price` *vary with* `clarity`

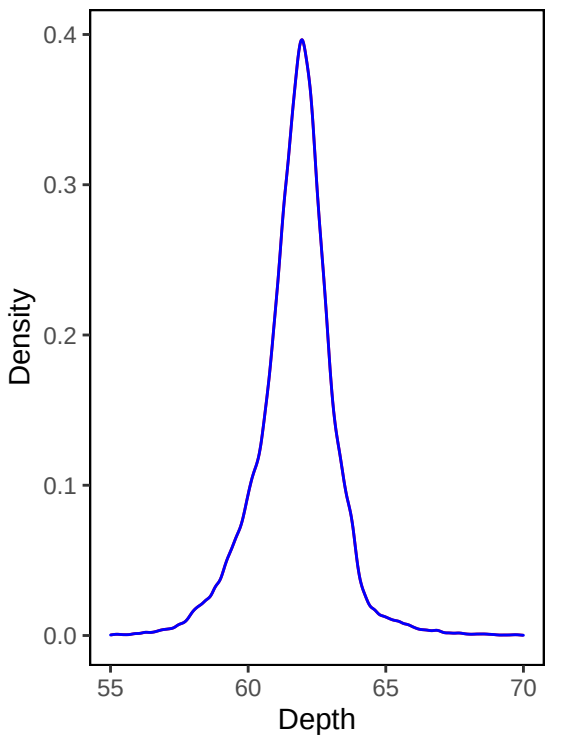
Below I'll plot a series of boxplots to visualize the distribution of `price` with `clarity`. I've plotted the median price across all clarity classes as a dashed horizontal line. In general, the median values tend to decrease with increasing diamond clarity, but the number of outliers generally tends to increase with diamond clarity.



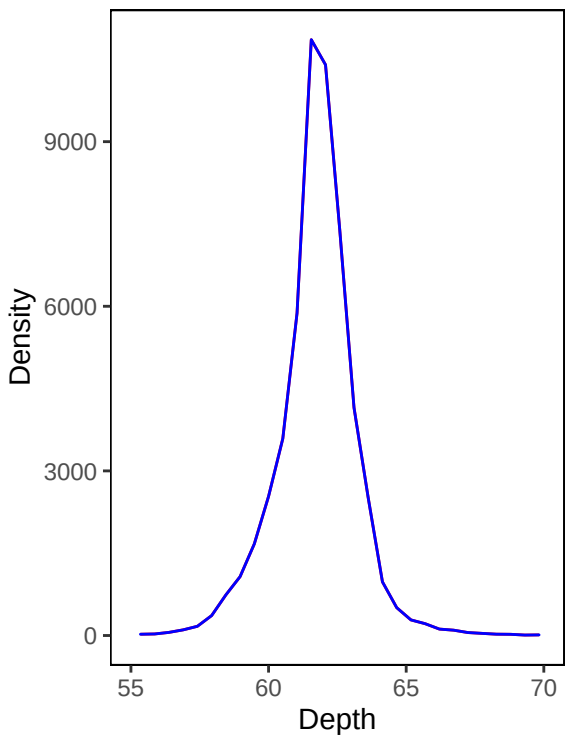
4) *Overlay a frequency polygon and density plot of depth. What computed variable do you need to map to y to make the two plots comparable? (You can either modify geom_freqpoly() or geom_density().*

In order to overlay the plots, we can take two approaches: a) modify the `geom_freqpoly()` call with the argument `stat = "density"` to calculate the density, which is calculated as the observations within each bin, divided by the total number of observations, multiplied by the binwidth; b) modify the `geom_density()` call with the argument `stat = "count"` to modify the statistical transformation to present the number of observations for each bin instead of calculating the density.

A)


geom_freqpoly(stat = 'density')
Density
0.4
0.3
0.2
0.1
0.0
55
60
65
70
Depth

B)


geom_density(stat = 'bin')
Density
9000
6000
3000
0
55
60
65
70
Depth