Activity 8

Arushi Singh

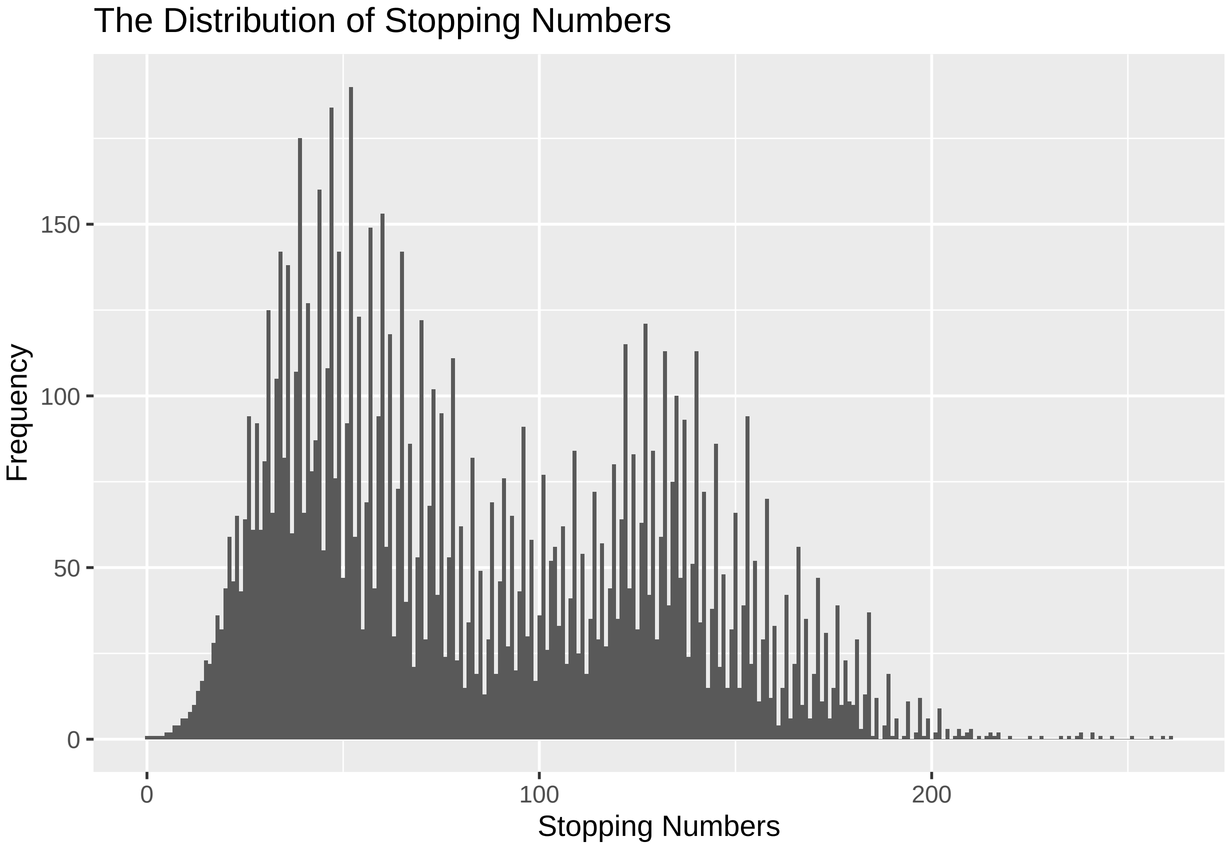
2022-11-04

Collatz Conjector (Section 1)

The Collatz Conjecture is one of the most famous problems in math, and currently remains unsolved. The conjecture asks if repeating two simple operations will eventually transform every positive integer into 1. It can be presented as:

```
f(n)= { f(n/2) if n is even,
        f(3n+1) if n is odd,
        stop    if n is one }
```

The function is recursive, and calls on itself continuously until the stopping condition is fulfilled. Prof Neil wants to see the distribution of the "stopping numbers" for the first 10000 positive integers. A stopping number is defined as the number of times we recursively invoke the function to get to 1.



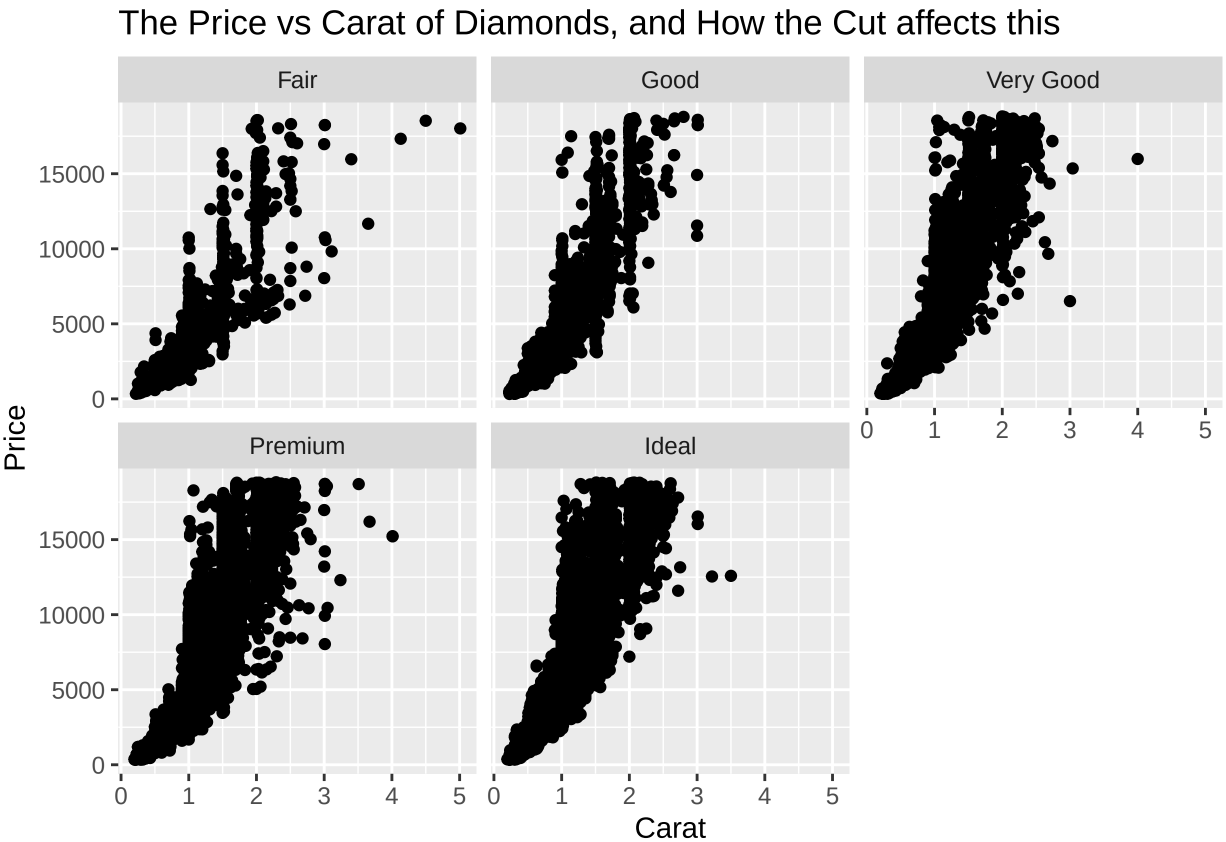
The histogram of the stopping points shows the distribution of the number of stopping points for all positive integers up to 10000. As shown by the data visualization, the stopping numbers tend to be around 45 the most and around 125 a decent amount as well. All of the stopping points are under 300.

Diamonds Analysis (Section 2)

In the next phase of this course, we explored data visualizations in R, particularly using the ggplot2 package. The diamonds data set was a hub of exploration and experimentation. The diamonds data set contains data on over 50,000 diamonds. The carat, cut, color, clarity, depth, table, price, length, width and height are all recorded.Below, I have formed two data visualizations and a table, showing different aspects of the diamonds data set, each showing statistics related to the price of diamonds, and which attributes included in the diamonds data set affect price.

Graph 1

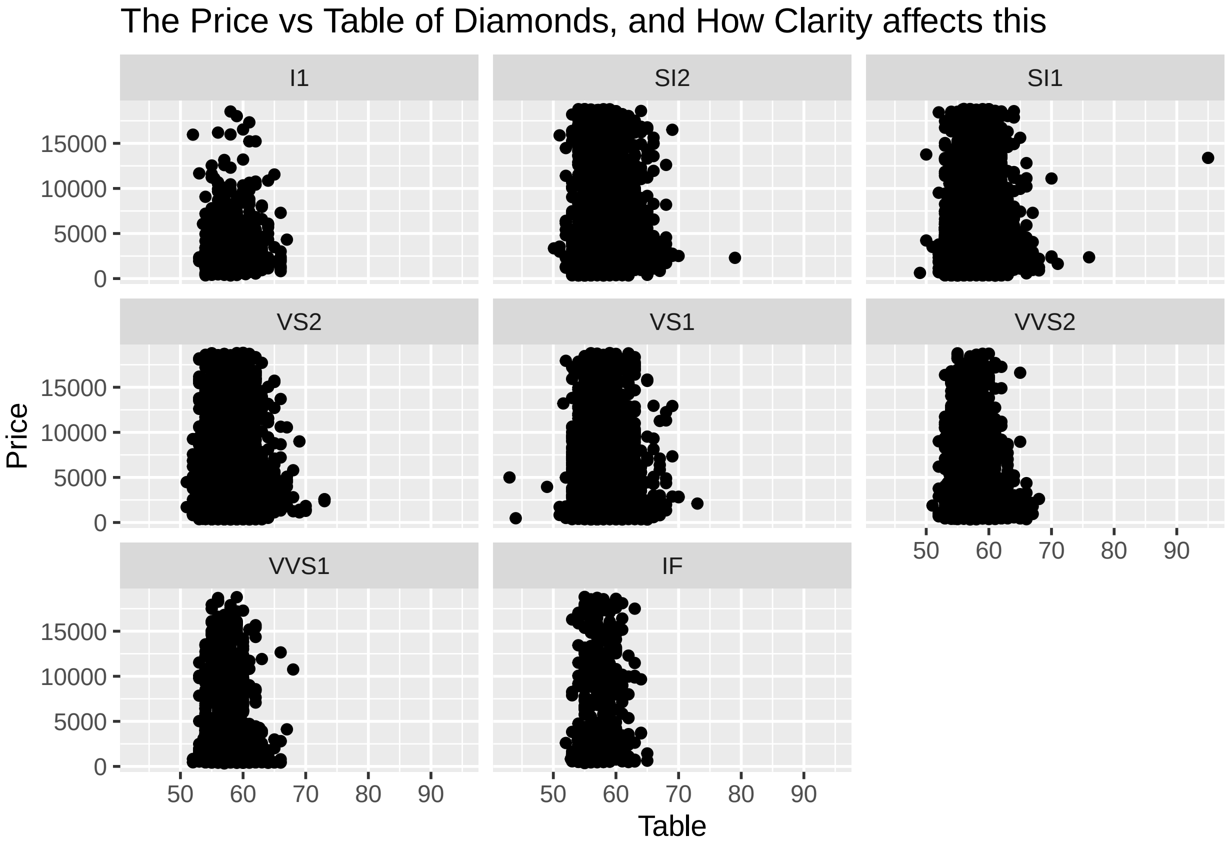
The following data visualization was formed to show the relationship between three values of the diamonds data set: carat, cut, and price. The carat of a diamond is a measure of its weight. Since diamonds usually have small weights in grams, carat is used to ease communication of a diamonds' characteristics. The cut of a diamond is a rating of the quality of its cut, ranging from fair to ideal.



The visual above displays the relationship between the carat, price, and cut of the diamonds in the diamonds data set in ggplot2 in side-by-side graphs. On the x-axis, carat is plotted and on the y-axis, price is plotted. Although some of the data is concentrated in certain areas, this version of this data visualization is more accessible than if we used color to display the cut. The side-by-side graphs show that cut does not necessarily affect the price of the diamond, as all of the graphs have somewhat similar concentrations of data and there is no drastic difference between any of them. However, carat definitely affects the price, as shown by the positive correlation of the data in each graph.

Graph 2

The next data visualization depicts the relationship between price, table, and clarity of a diamond. The table of a diamond is the facet that can be seen when the diamond is placed face up, which indicates how well it can reflect light, as well as the diamonds' brilliance. The clarity of a diamond is a measure of the purity of the stone. The less flaws a diamond has, the higher its clarity rating is.



The visual above displays the relationship between the table of a diamond, price, and clarity of the diamonds in the diamonds dataset using side-by-side scatter plots, distinct by the clarity category. As seen by this visualization, table affects the price of a diamond more than clarity, as the relationships between price and table remain relatively consistent among each clarity quality.

Table 1

The following table displays a summary of diamond prices by cut, to have further quantitative insight into how price is affected by cut.

Through this table, you are able to learn about the price of diamonds in relation to cut. You are to see how price differs by cut with this thorough number summary that contains the minimums, quintile measurements, medians, maximums, means, and standard deviations of price and compare them by cut. We can see that there really isn't much of a difference between the different cuts, excluding Fair. The price data summary of the diamonds with the Fair cut are consistently higher than the other cut categories. This could be due

Table 1: Summary Table of Price (in USD) of Diamonds by Cut

Cut	Minimum	Quintile 1	Quintile 2	Median	Quintile 3	Quintile 4	Maximum	Arithmetic Mean	Standard Deviation	Count
Fair	337	1790.6	2805.0	3282.0	3947.4	6090.4	18574	4358.76	3560.39	1610
Good	327	876.0	2176.0	3050.5	3888.0	5834.0	18788	3928.86	3681.59	4906
Very Good	336	760.0	1892.4	2648.0	3751.6	6288.0	18818	3981.76	3935.86	12082
Premium	326	924.0	2100.0	3185.0	4408.0	7485.0	18823	4584.26	4349.20	13791
Ideal	326	803.0	1243.0	1810.0	2529.0	5613.0	18806	3457.54	3808.40	21551

to a there being less diamonds with the Fair clarification in the set, or could show that despite a lower cut classification, people are still able to sell these diamonds at a high cost.

What have I learned so far? (Section 3)

Stat 184 has been a challenging class for me personally. After I contracted Covid during the second week of the semester, and the lecture-less classes the week after, I had a hard time really understanding what the code I was writing meant. Now, I feel I have a better handle on coding, and on how to handle similar situations. I feel that we have gone very in depth with learning about different data visualizations, and I liked that we learned more about the theory behind good data visualizations before creating them. It helped me frame my thought process in a helpful way. I feel very confident in my ability to create data visualizations and tables after this class. In addition, I feel like I've started to think about coding in general very differently than I did before. This class has taught me the importance of planning in advance and I am incredibly thankful for that, as its helped me in my other coding classes as well. All in all, I have learned a lot in this class!