FirstRMD

Jack Billings

2022-11-04

The Collatz Conjecture

The Collatz Conjecture is a theory that repeating the two arithmetic equations, {if the number is even, divide by 2} and {if the number is odd, multiply by three and add one}, on any number it will eventually result in One. It remains a theory as we cannot compute this for an infinite amount of numbers but so far it is yet to be disproved.

Stopping Numbers

The stopping number for a certain integer is the amount of steps it takes for the function to arrive at zero using the Collatz Conjecture.

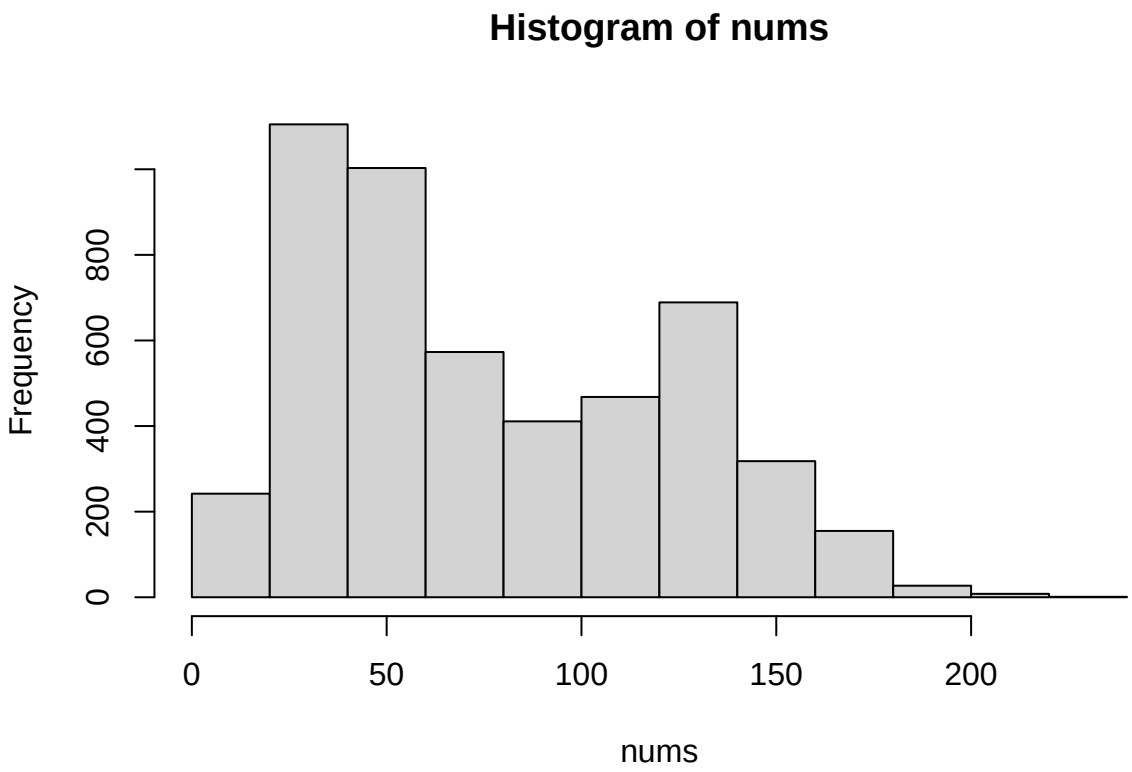
Create a set of rules that mirrors the Collatz conjecture and a program that returns the amount of steps taken from an input variable to the stop.

The rules for the Collatz Conjecture as stated above are :

- if the number is even, divide by 2
- if the number is odd, multiply by three and add one

The Stopping numbers can be calculated by creating a program which runs the Collatz Conjecture and adding in a step variable which increases each time one of the rules is applied. By running the conjecture until the output is 1 we can take the result of the step variable to be the stopping variable.

A histogram of the numbers 1 to 50000 using the rules above would look like:

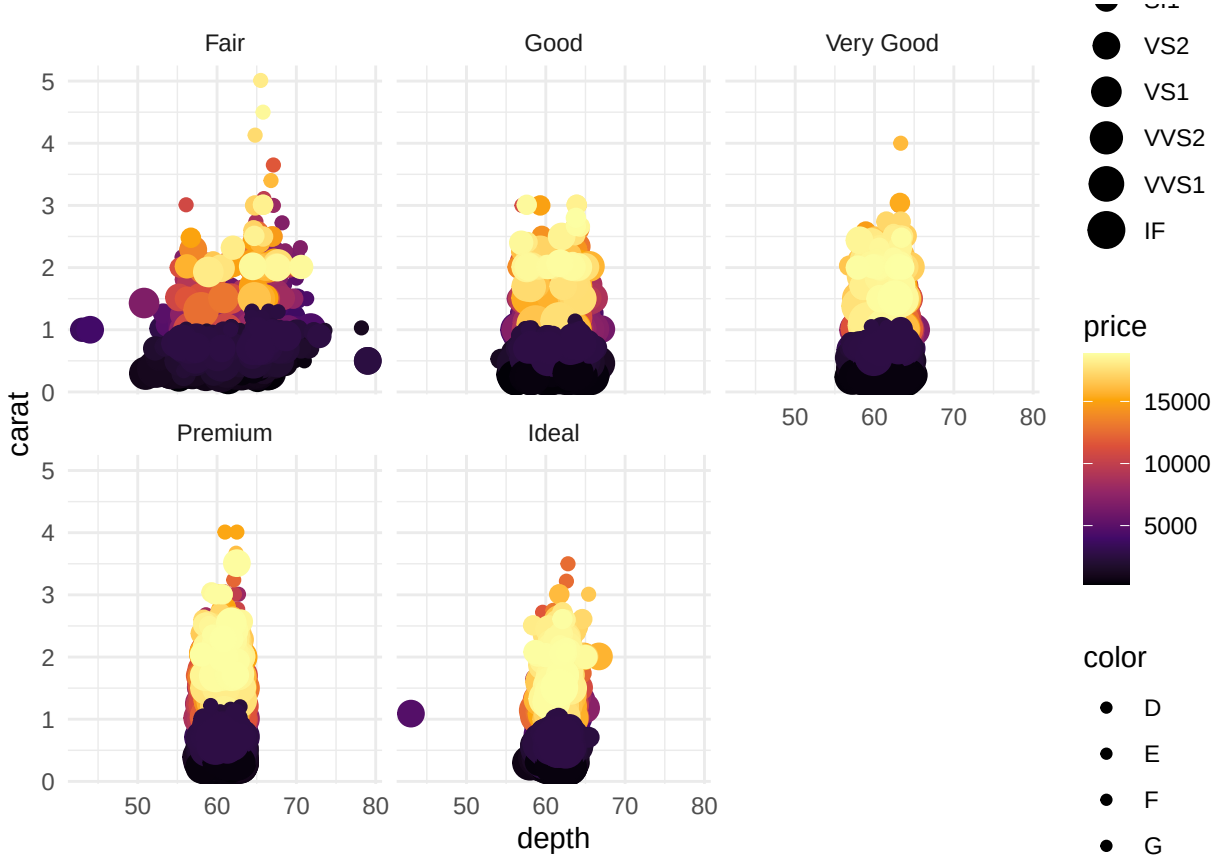


Data Visualization

Graphs

Much can be learned from a well designed table. The two prominent theorists in this area are Tufte and Kosslyn for creating the six principles of design and the eight principles for effective graphing respectively. The objective of these principles are to be able to take the data from a table or other source and translate it into a easily readable graph for a nontechnical viewer to understand.

The graph below is working from the ggplot table "diamonds" and equates the carat, cut, clarity, and depth to the price as shown by color. The x and y axes are determined by depth(size) and carat respectively, with the color indicating price on a scale from purple (low) to yellow (high). The different facets are determined by the cut of the diamond, as each cut should be valued differently since it determines the overall baseline quality. A simplification of the relationship shown is that the higher carat and depth of a diamond roughly equates higher price. This visualization is intended to show a trend in price as related to the different aspects of a diamond, showing which are more impactful than others.



Tables

To visualize data as a table it is required to be tidy. Tidy data is defined by R as:

- Every column is a variable.
- Every row is an observation.
- Every cell is a single value.

This results in the most readable set of standardized data and also prepares it well for other types of visualization which could be performed. The next step for keeping it as a table is to create a summary table which effectively describes the data to the degree specified, usually including a count, min, median, max, mean, and standard deviation.

The table below also visualizes diamonds, but describes differences in width rather than price. In a brief overview, fair cut has the greatest average size but least number of cases, showing that quality and size serve similar purposes in relation to price. Aside from premium, the higher the cut the lower the average size. This can be extrapolated by only viewing count and mean, so the quintiles included allow the viewer to see the trends in the data set over smaller sections as well.

```
## # A tibble: 5 x 11
##   cut       count minimum first~1 secon~2 median third~3 fourt~4 maximum arith~5
##   <ord>     <int>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Fair       1610       0    5.50    5.98   6.1     6.23    7      10.5     6.18
```

```
## 2 Good       4906       0    4.74    5.63   5.99    6.23    6.6     9.38    5.85
## 3 Very Good 12082       0    4.62    5.4    5.77    6.16    6.68    9.94    5.77
## 4 Premium   13791       0    4.66    5.62   6.06    6.4     6.94   58.9     5.94
## 5 Ideal     21551       0    4.46    4.96   5.26    5.72    6.57   31.8     5.52
## # ... with 1 more variable: `arithmetic standard deviation` <dbl>, and
## #   abbreviated variable names 1: firstQuintile, 2: secondQuintile,
## #   3: thirdQuintile, 4: fourthQuintile, 5: `arithmetic mean`
```

Course Takeaways

Coming into this class I was already fairly confident in my coding skills, and R is just a derivative of python so I was expecting to mostly relearn. Instead we focused much more on the theory behind data wrangling and visualization, which I am much less comfortable with than my technical skills. Overall this class taught me how to take the data I gather and create something presentable for consumers and supervisors both. Before this course I was perfectly fine with taking a screenshot of a raw table output and sticking it on a slide, which would require me to describe it from scratch on the spot while presenting, but using what I learned in this class I now know how to go about creating a translation between data and viewer to benefit both.