Diamonds are forever

Derien Weatherspoon - Group 3

M3 ICA1

Contents

The data

Package `tidyverse` is a set of packages that work in harmony. We will use the `ggplot2` package to produce our data visualizations. This package is part of the `tidyverse` package. As we move forward, we will utilize some of the other packages loaded via `tidyverse`.

```
library(tidyverse)
```

The ggplot2 package comes with a data set called `diamonds`. Let's look at it below. To obtain further details type `?diamonds` in your console window.

```
glimpse(diamonds)
```

```
Rows: 53,940
Columns: 10
$ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23,...
$ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, ...
$ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J,...
$ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS...
$ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4,...
$ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62,...
$ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340,...
$ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00,...
$ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05,...
$ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39,...
```

Variable codebook

The dataset containing the prices and attributes of nearly 54,000 diamonds includes the following variables:

```
*price*
price in US dollars (\$326--\$18,823)

*carat*
weight of the diamond (0.2--5.01)

*cut*
quality of the cut (Fair, Good, Very Good, Premium, Ideal)

*color*
diamond colour, from D (best) to J (worst)

*clarity*
a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

*x*
length in mm (0--10.74)

*y*
width in mm (0--58.9)

*z*
depth in mm (0--31.8)

*depth*
total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)

*table*
width of top of diamond relative to widest point (43--95)
```

The investigation

What influences the price of a diamond?

As a quick primer on diamond pricing, watch the following video on 'the 4C's of Diamonds.' https://www.youtube.com/watch?v=dFiG3ckNCIY&feature=emb_logo

Exploring the data

1. Which of the 4C's do you predict most influences the price of a diamond? The least? Provide a 1-2 sentence explanation of your opinion. ## Since the sparkliness of a diamond is what really matters most when diamonds have similar specs, I'd assume that the cut of a diamond influences prices the most.
2. The data set `diamonds` is stored in R as a tibble. This allows for a convenient way to view the data frame in the console. Type `diamonds` in your console to see.

```
diamonds
```

```
# A tibble: 53,940 x 10
   carat cut       color clarity depth table price     x     y     z
   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
 1 0.23  Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
 2 0.21  Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
 3 0.23  Good      E     VS1      56.9    65   327  4.05  4.07  2.31
 4 0.290 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
 5 0.31  Good      J     SI2      63.3    58   335  4.34  4.35  2.75
 6 0.24  Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
 7 0.24  Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
 8 0.26  Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
 9 0.22  Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
10 0.23  Very Good H     VS1      59.4    61   338  4     4.05  2.39
# ... with 53,930 more rows
```

Let's briefly explore the distribution of some of the variables stored in `diamonds`.

3. a. Variables cut, color, and clarity are all factors. Use the function `levels` to see the levels of each variable. They are sorted from worst to best. Use the `table` command to determine how many cases fall into each level of cut Do the same for color and clarity.

```
table(diamonds$clarity)
```

```
   I1   SI2   SI1   VS2   VS1  VVS2  VVS1    IF
  741  9194 13065 12258  8171  5066  3655  1790
```

```
table(diamonds$cut)
```

```
     Fair      Good Very Good   Premium     Ideal
     1610      4906     12082     13791     21551
```

```
table(diamonds$color)
```

```
    D     E     F     G     H     I     J
 6775  9797  9542 11292  8304  5422  2808
```

b. Because `price` is quantitative, the `table` command won't provide a useful summary of the distribution of observed cases for this variable. Use the `summary` function to get a sense of this variable's distribution.

```
summary(diamonds$price)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18823
```

c. Add a new variable to `diamonds` called `price.per.carat` that represents the price per carat. Are all diamonds priced the same per carat? If not, how much do these rates vary?

```
price.per.carat <- diamonds$price / diamonds$carat
```

Write 1-2 sentences hypothesizing what might cause a diamond to fetch a higher price per carat than others. ## All diamonds do not have the same price per carat. As stated before, I believe cut will influence most. In contrast to (3a) and (3b), which ask about the distribution of a single variable, exercise (3c) is the first to begin looking at the *relationship* between two variables - in this case, a diamond's `price` and `carat`.

Let's investigate relationships further. **What is the relationship between a diamond's price and its cut/color?**

4. a. Remember that a diamond's cut can be too shallow or too steep, and either will cause it to sparkle less dramatically under bright light.

Summarize and compare the prices of Fair (the worst cut) and Ideal (the best cut) diamonds.

```
# use the summary function on price for the particular cuts
price_summary <- diamonds %>% filter(cut == "Fair")
summary(price_summary$price)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    337    2050    3282    4359    5206   18574
```

```
price_sum2 <- diamonds %>% filter(cut == "Ideal")
summary(price_sum2)
```

```
     carat               cut        color       clarity         depth
 Min.   :0.2000   Fair      :    0   D:2834   VS2    :5071   Min.   :43.00
 1st Qu.:0.3500   Good      :    0   E:3903   SI1    :4282   1st Qu.:61.30
 Median :0.5400   Very Good:    0   F:3826   VS1    :3589   Median :61.80
 Mean   :0.7028   Premium   :    0   G:4884   VVS2   :2606   Mean   :61.71
 3rd Qu.:1.0100   Ideal     :21551   H:3115   SI2    :2598   3rd Qu.:62.20
 Max.   :3.5000                      I:2093   VVS1   :2047   Max.   :66.70
                                     J: 896   (Other):1358
     table           price             x               y
 Min.   :43.00   Min.   :  326   Min.   :0.000   Min.   : 0.000
 1st Qu.:55.00   1st Qu.:  878   1st Qu.:4.540   1st Qu.: 4.550
 Median :56.00   Median : 1810   Median :5.250   Median : 5.260
 Mean   :55.95   Mean   : 3458   Mean   :5.507   Mean   : 5.520
 3rd Qu.:57.00   3rd Qu.: 4678   3rd Qu.:6.440   3rd Qu.: 6.445
 Max.   :63.00   Max.   :18806   Max.   :9.650   Max.   :31.800

       z
 Min.   :0.000
 1st Qu.:2.800
 Median :3.230
 Mean   :3.401
 3rd Qu.:3.980
 Max.   :6.030
```

How unexpected! Many of the quantile prices for diamonds rated Fair exceed those rated Ideal (every value of the five-number summary apart from the max price]. This is probably the exact opposite of what you may have expected.

b. Summarize and compare the prices of D (the best color) diamonds and J (the worst color) diamonds.

```
#use the summary function on price for the particular colors.
color_summary <- diamonds %>% filter(color == "D")
summary(color_summary$color)
```

```
   D    E    F    G    H    I    J
6775    0    0    0    0    0    0
```

```
color_sum2 <- diamonds %>% filter(color == "J")
summary(color_sum2)
```

```
     carat               cut        color        clarity         depth      
 Min.   :0.230   Fair      :119   D:   0   SI1     :750   Min.   :43.00  
 1st Qu.:0.710   Good      :307   E:   0   VS2     :731   1st Qu.:61.20  
 Median :1.110   Very Good:678   F:   0   VS1     :542   Median :62.00  
 Mean   :1.162   Premium   :808   G:   0   SI2     :479   Mean   :61.89  
 3rd Qu.:1.520   Ideal     :896   H:   0   VVS2    :131   3rd Qu.:62.70  
 Max.   :5.010                   I:   0   VVS1    : 74   Max.   :73.60  
                                 J:2808   (Other):101                  
     table           price             x                y         
 Min.   :51.60   Min.   :  335   Min.   : 3.930   Min.   : 3.900  
 1st Qu.:56.00   1st Qu.: 1860   1st Qu.: 5.700   1st Qu.: 5.718  
 Median :58.00   Median : 4234   Median : 6.640   Median : 6.630  
 Mean   :57.81   Mean   : 5324   Mean   : 6.519   Mean   : 6.518  
 3rd Qu.:59.00   3rd Qu.: 7695   3rd Qu.: 7.380   3rd Qu.: 7.380  
 Max.   :68.00   Max.   :18710   Max.   :10.740   Max.   :10.540  

       z        
 Min.   :2.460  
 1st Qu.:3.530  
 Median :4.110  
 Mean   :4.033  
 3rd Qu.:4.580  
 Max.   :6.980  
```

Another unexpected result! The summary statistics show that many poorly colored diamonds fetch higher prices than ideally colored diamonds at the same percentile ranking (i.e., the 75th most expensive perfectly-colored diamond is much cheaper than the 75th most expensive poorly-colored diamond). How can this be?

5. **What influences a diamond's price?** In 1-2 sentences, hypothesize as to why the distribution of poorly cut diamonds trends more expensive than those that are perfectly cut. ## I thought that the cut would influence the price most based on the video, but it looks as if that is not the case. Seems like people really like the color and carat because those are most understandable to people. # Visualizations with ggplot

Visualizations are an excellent tool for exploring the distributions of variables - their patterns, their variability, and how those attributes are related to those of other variables. The exercises below ask you to first recreate a number of visualizations we believe will help to answer our initial question **What influences a diamond's price?** along with several others along the way.

Exercises

5. The code to create each plot below is given. Try to understand what the associated code is doing. As we start delving into Data Visualization you will gain experience and can recreate these plots on your own.

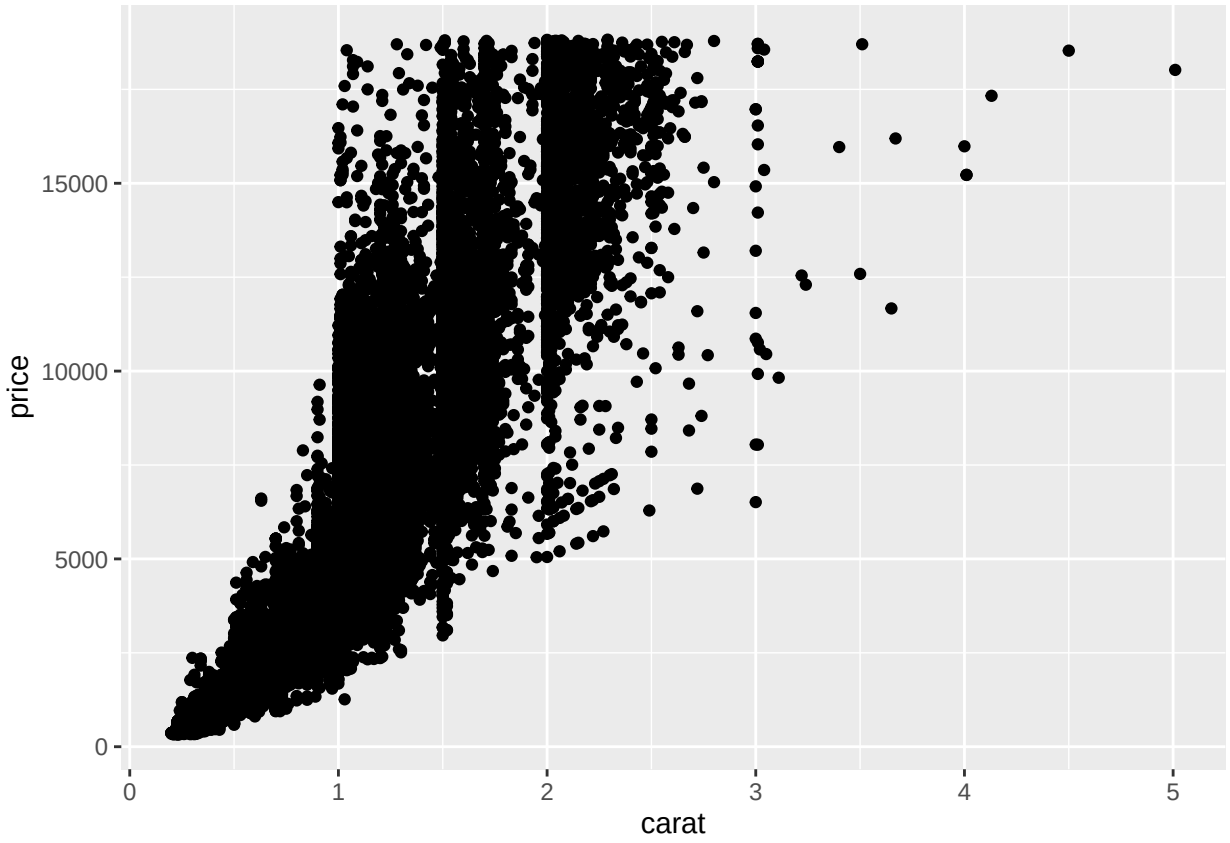
6. Consider Plot 1. Notice how the distribution of `carat` vs. `price` has clustered vertical lines at 1, 1.5, and 2 carats. It seems suspicious that naturally occurring diamonds would appear at these values more often than others (such as 0.9, 1.4, or 1.95). Provide a plausible explanation for why our data on ~54,000 diamonds might display this pattern. ## I'm assuming this is to make the dataset isn't skewed on those specific numbers of 1, 1.5, or 2.

7. Consider Plot 2, which displays the relationship between `carat`, `color`, and `price`, and remember that color is coded across letters D to J, from most desirable to least desirable. We earlier discovered that poorly colored diamonds tend to fetch higher prices than ideally colored diamonds. Write 1-2 sentences describing how this plot helps to explain this counterintuitive finding. ## The least desirable colors have the most carats. This seems counterintuitive to how it should be, but if the diamond had the worst color and the worst carat, then it's likely it wouldn't sell at a given price. This is why the most desirable color can be sold more on a lower carat.

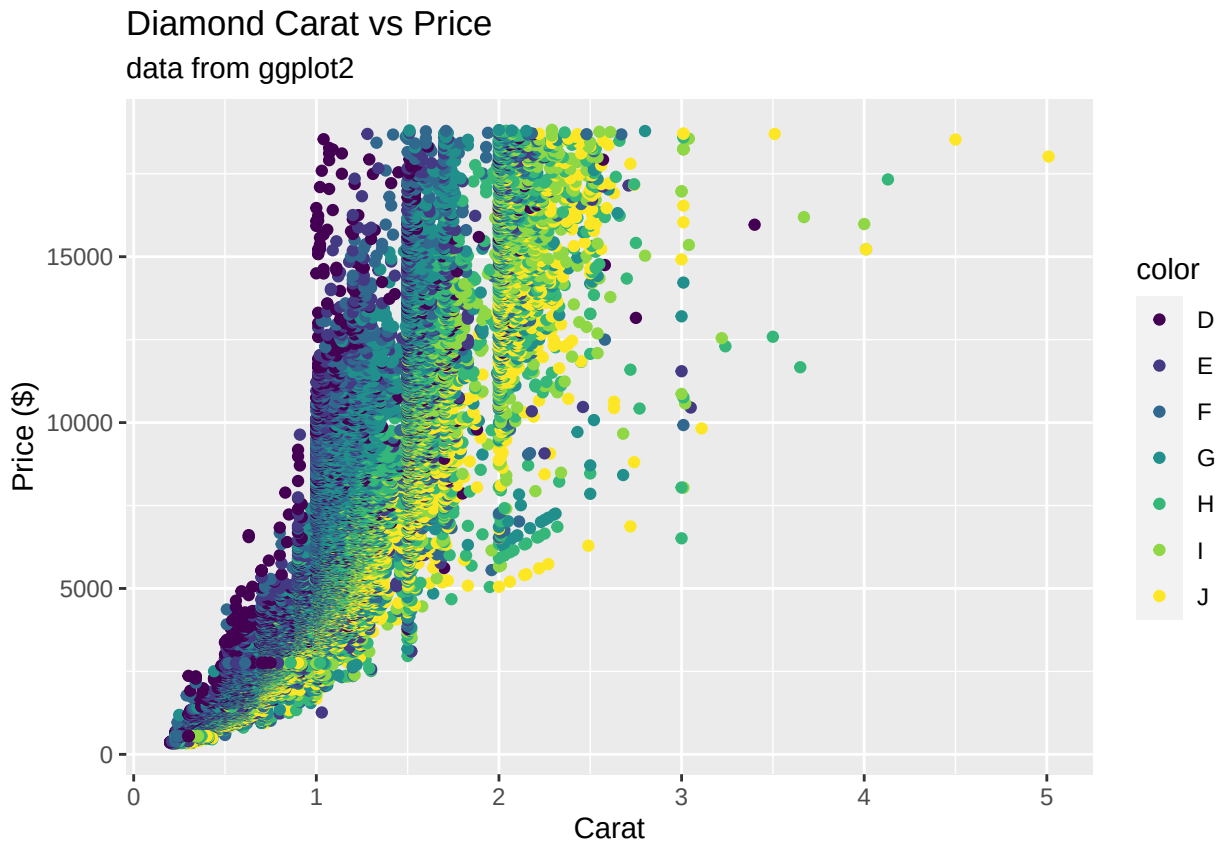
8. You'll notice Plots 3 and 4 display the relationship between the same four variables, `carat`, `color`, `price`, and `cut`. Which of the two plots is more useful when analyzing which factors play a role in the overall price of a diamond? Justify your opinion in 2-3 sentences. ## Seems like plot 4 is just because it is easier to read in my opinion. Although, it doesn't give a discrete relationship since the differences are all in different graphs, the idea is stil there.

9. Finally, consider Plot 5, which considers the final of the '4 C's' of diamond pricing, clarity. You should notice that the clearest of diamonds (rating IF) include many more ideal cuts than the foggiest of diamonds (rating l1). Hypothesize as to why this might be? ## This is probably likely because a diamond cutter would have to be much more precise when cutting diamonds of that quality at that carat. So the lowest quality cuts can have higher carats usually.

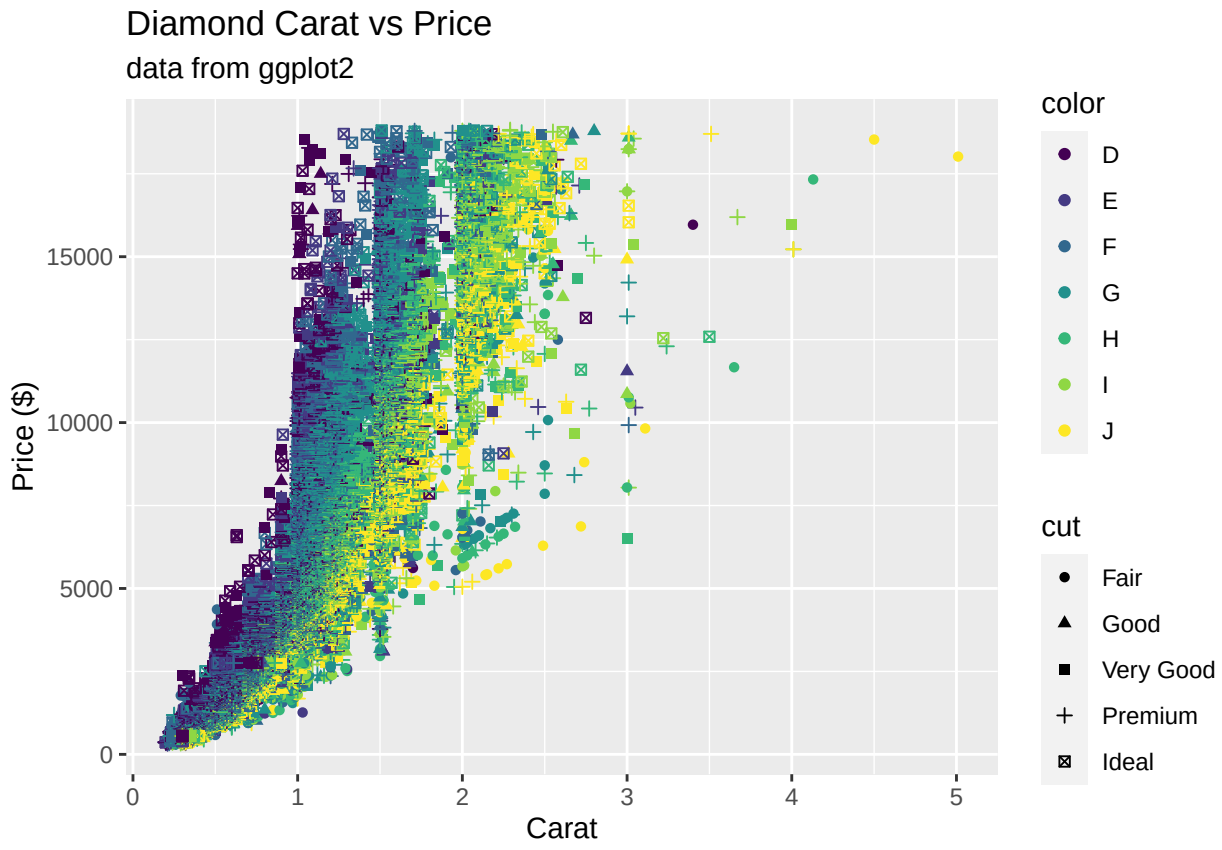
Plot 1



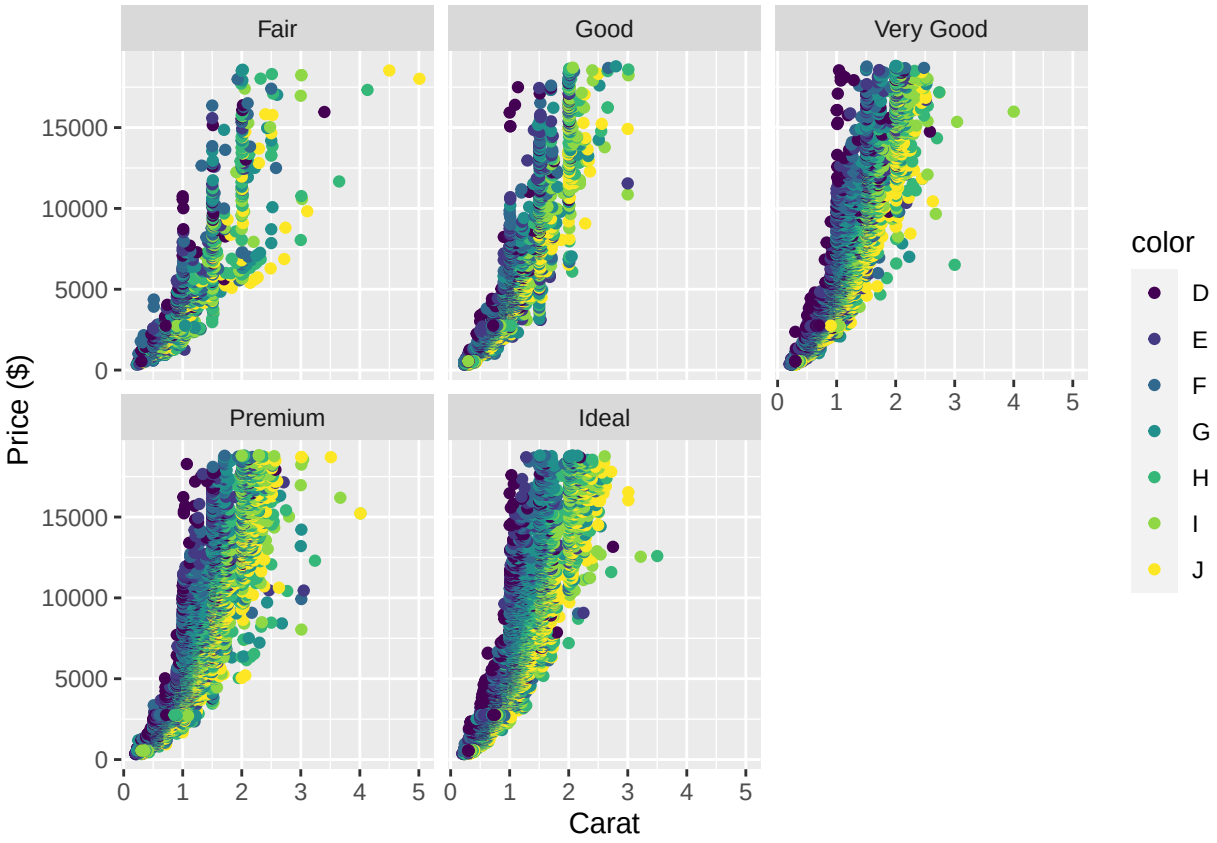
Plot 2


Diamond Carat vs Price
data from ggplot2
Price ($)
15000
10000
5000
0
0
1
2
3
4
5
Carat
color
D
E
F
G
H
I
J

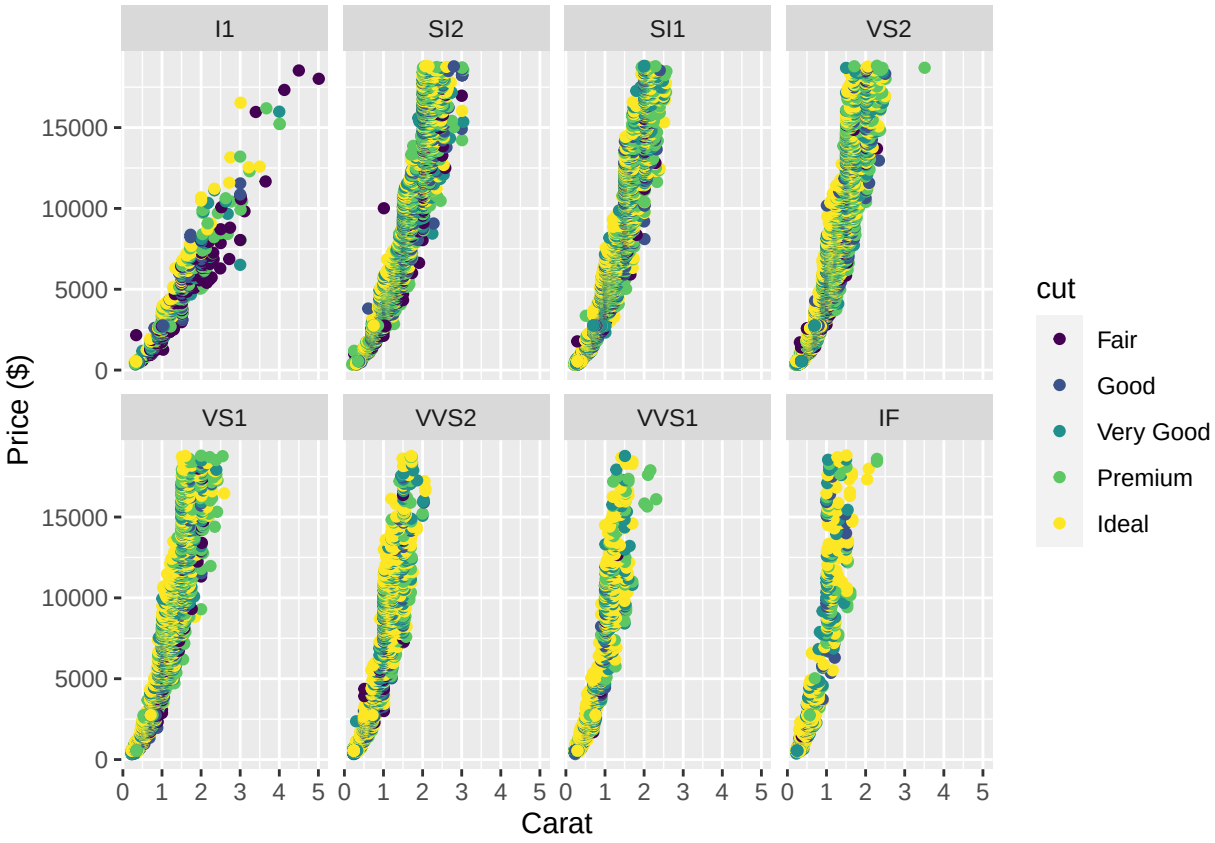
Plot 3


Diamond Carat vs Price
data from ggplot2
Price ($)
15000
10000
5000
0
0
1
2
3
4
5
Carat
color
D
E
F
G
H
I
J
cut
Fair
Good
Very Good
Premium
Ideal

Plot 4


Fair
Good
Very Good
Premium
Ideal
Price ($)
Carat
color
D
E
F
G
H
I
J

Plot 5



Wrap up

In this in-class activity, we used visualization to investigate the relationships between two or more variables. You likely noticed that apparent relationships between two variables (for instance, that poorer colored diamonds tended to be more expensive) were sometimes influenced by the interaction with a third, unobserved variable. Such unobserved variables are often considered *confounding variables*.

In this case, a diamond's `carat` (or size) is very strongly associated with its `price`. Additionally, the larger a diamond grows, the more unlikely it is to maintain ideal color or clarity, and the more difficult it is for a jeweler to cut it perfectly. This helps to explain our initial finding that poorly colored or cut diamonds were more expensive - often, they were just bigger. This is one example of the power of data visualization, which helps make intelligible patterns in data that might not otherwise make sense.

References

1. Grolemund, G., & Wickham, H. (2019). R for Data Science. https://r4ds.had.co.nz/