Data Visualization

Introduction

Carley Dziewicki

January 19, 2019

Overview

Using the `diamonds` data set contained in the `ggplot2` package, this homework asks you to examine factors that affect the price of a diamond. Because the setup chunk above contains the option `echo=TRUE`, whenever you knit your R Markdown file, your document will contain your R commands, the R output and your narrative text. After you have completed the assignment, knit your R Markdown file and submit the knitted HTML file on Moodle by noon Saturday, January 19.

About the Data Set

1. Use R help to learn about the `diamonds` data set. Answer the following questions:
    - How many cases are there in the total data set?
    - How many variables are there?
    - What are the units used for the `price` variable?
    - What are the possible values of the `color` variable? Which is best?

Answer: Cases in Data set: 53940 Variables: 10 Units for price: in US dollars possible values of color: from J to D

```
head(diamonds)
```

```
## # A tibble: 6 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
```

```
?diamonds
```

Creating the Subset for Analysis

Note: As you should have seen above, the `diamonds` data set is very large so for the rest of the homework you will work with a randomly selected subset of 2000 cases created by the following code. You MUST set `eval=TRUE` in the following code to create the subset. Run this chunk by selecting the green arrow at the top of the chunk

```
# Change eval=FALSE to eval=TRUE when you knit this or it won't have any effect
set.seed(295000)
diamond_sub <- diamonds %>%
  filter(carat <= 2.5) %>%
  sample_n(2000)
```

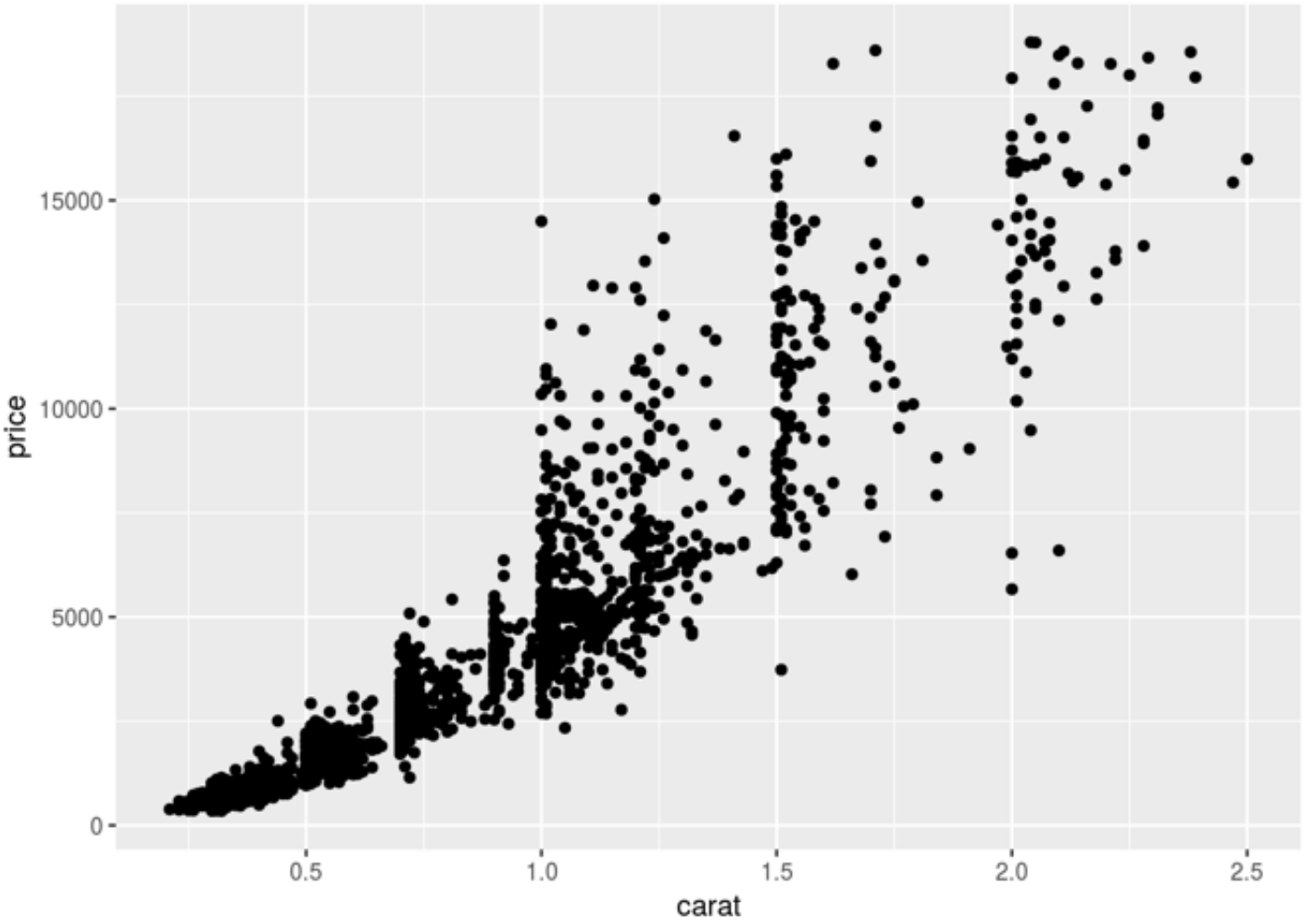
Analysis of the `diamond_sub` dataset

Note: For all of the following, be sure to use appropriate `ggplot2` procedures. Use of other R graphics platforms will not earn credit.

2. Create a scatterplot of diamond price (`y`) versus carat weight (`x`). Comment briefly on the nature of the relationship.

Answer: The plot seems to have might variability in each of the carats and most seem to be a .5 number carat as opposed to between .5 and 1.

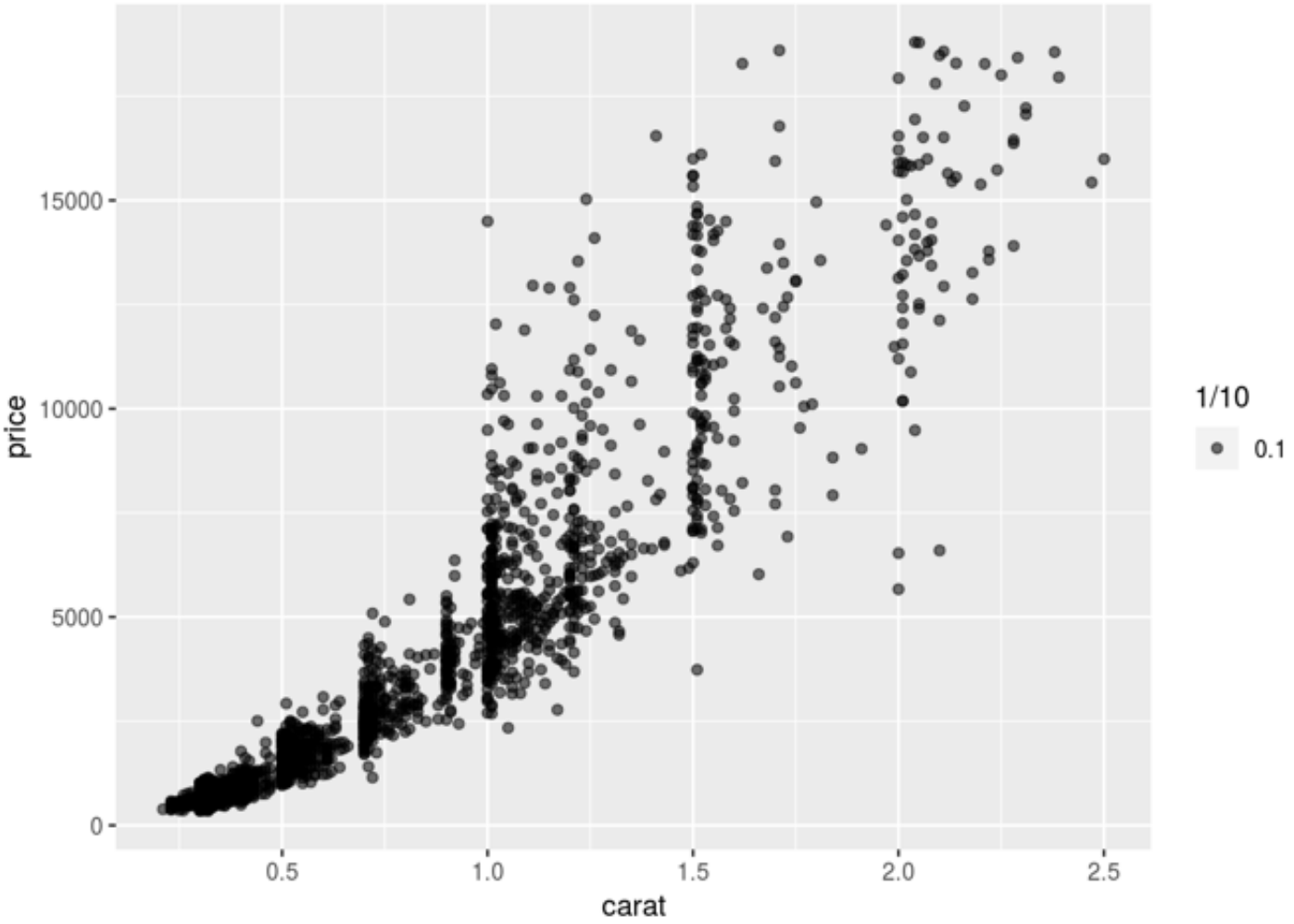
```
ggplot(data=diamond_sub) + geom_point(mapping=aes(x=carat,y=price))
```



3. Modify your plot in 2) setting the `alpha` transparency value to 1/10. Why is this useful for large datasets?

Answer:

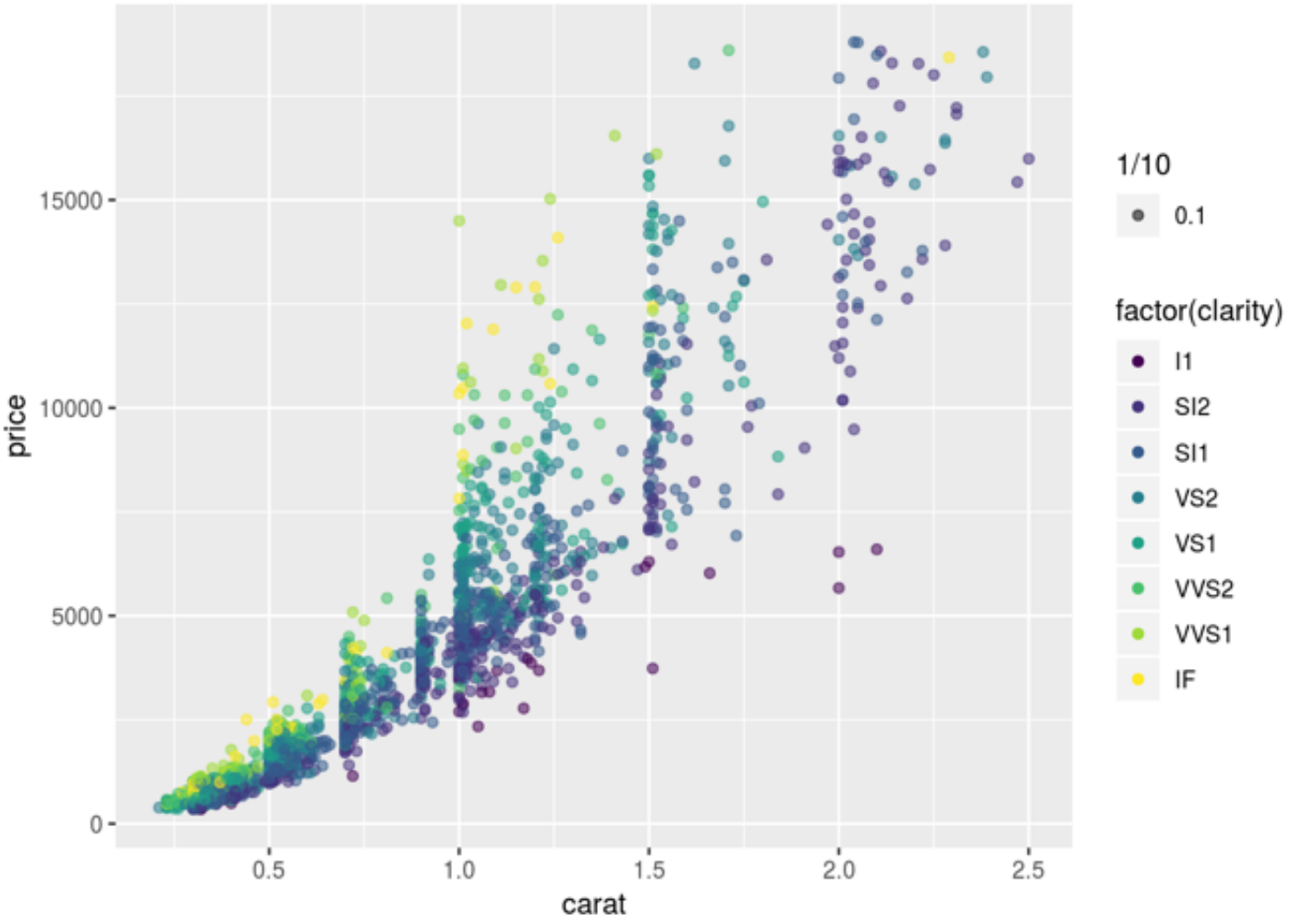
```
ggplot(data=diamond_sub) + geom_point(mapping=aes(x=carat,y=price, alpha=1/10))
```



4. Modify your plot in 2), mapping the `clarity` of the diamond to the color aesthetic of the point. Comment on what this plot says about the relationship between `clarity` and the two other variables?

Answer: This plot is showing that the more clarity the higher the price and the higher the carat the higher the price.

```
ggplot(data=diamond_sub) + geom_point(mapping=aes(x=carat,y=price, alpha=1/10, color=
factor(clarity)))
```



5. Once again, modify your plot in 2) mapping the `clarity` of the diamond to the shape aesthetic of the point. Explain why you get warning messages for this plot.

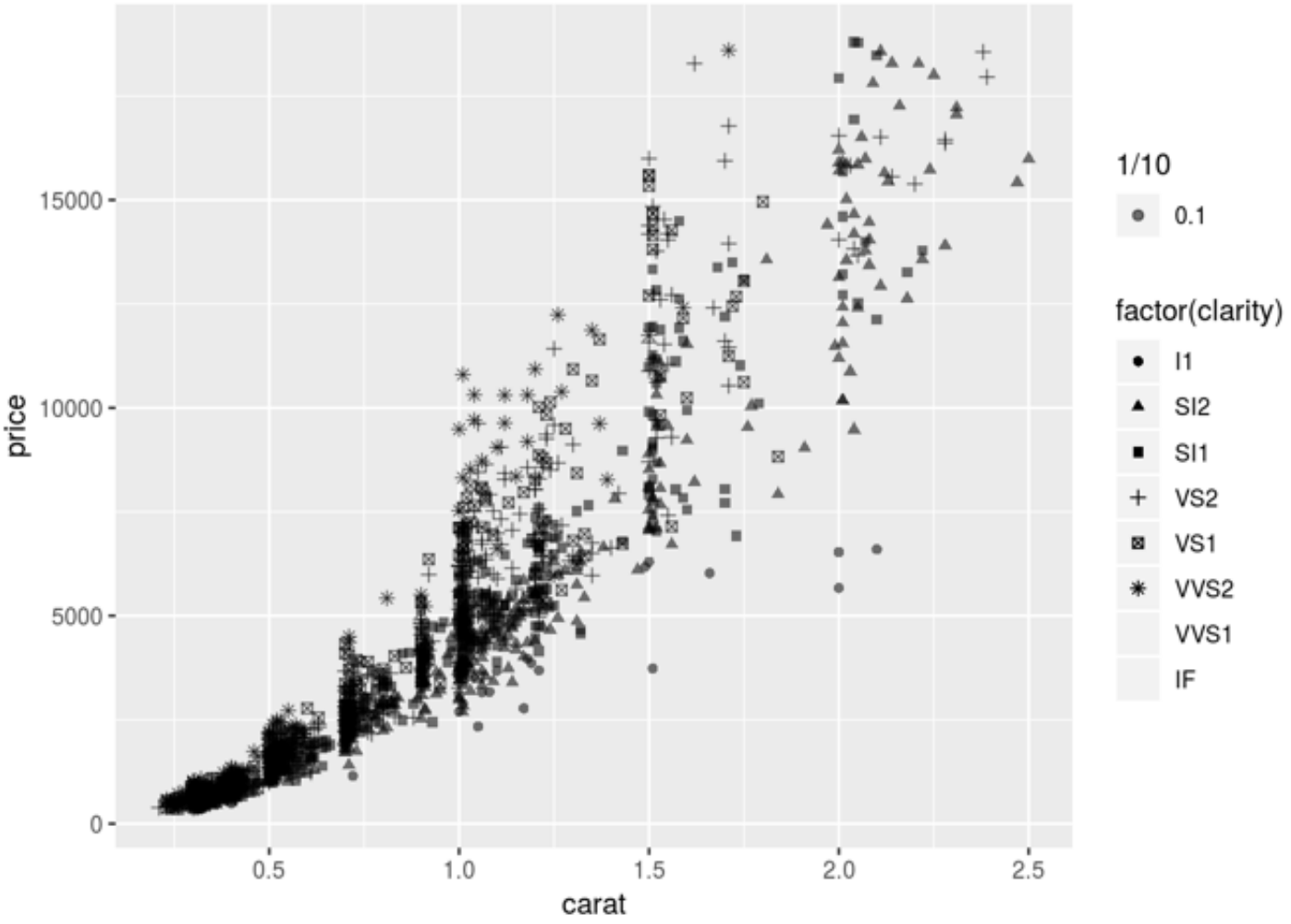
Answer: There are too many levels of clarity and not enough shapes

```
ggplot(data=diamond_sub) + geom_point(mapping=aes(x=carat,y=price, alpha=1/10, shape=
factor(clarity)))
```

```
## Warning: Using shapes for an ordinal variable is not advised
```

```
## Warning: The shape palette can deal with a maximum of 6 discrete values because
## more than 6 becomes difficult to discriminate; you have 8. Consider
## specifying shapes manually if you must have them.
```

```
## Warning: Removed 202 rows containing missing values (geom_point).
```



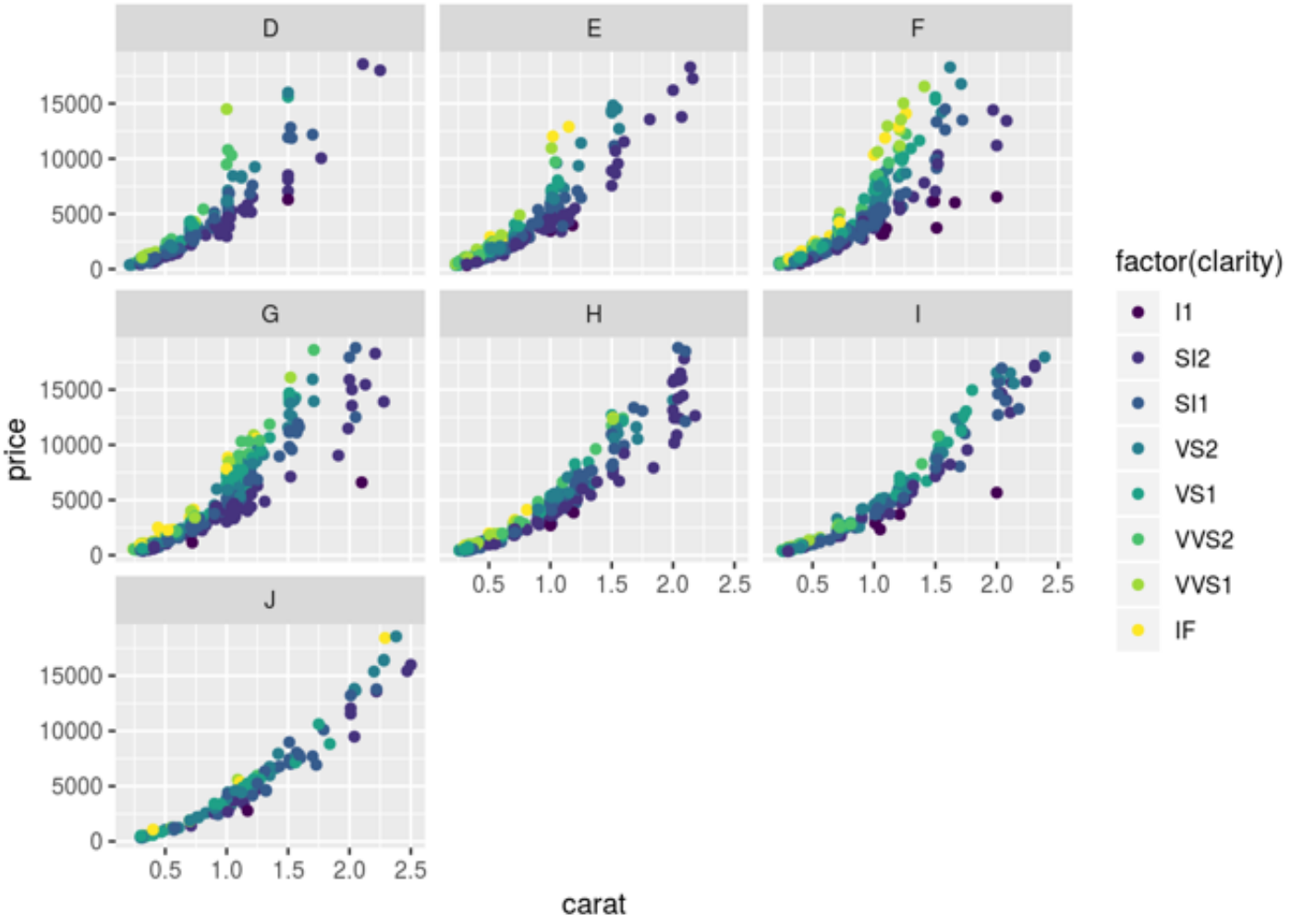
6. Which of the two plots above do you prefer, the one that maps `clarity` to color or shape? Justify your choice.

Answer: I prefer the plot that maps clarity to color. It is easier to see the pattern then the shapes

7. Examine the relationship between `price` and `carat` weight in more detail by creating one plot for each diamond `color`. Describe what you can learn from this analysis.

Answer: From an analysis like this you can look to see if there are different trends in the price and carat based on the color. It would be hard to compare the colors directly to one anohter but you can see individual variability. Such as F is more variable then I and J. Also I and J dont have as many if any clarity of IF. So maybe color and clarity are associated.

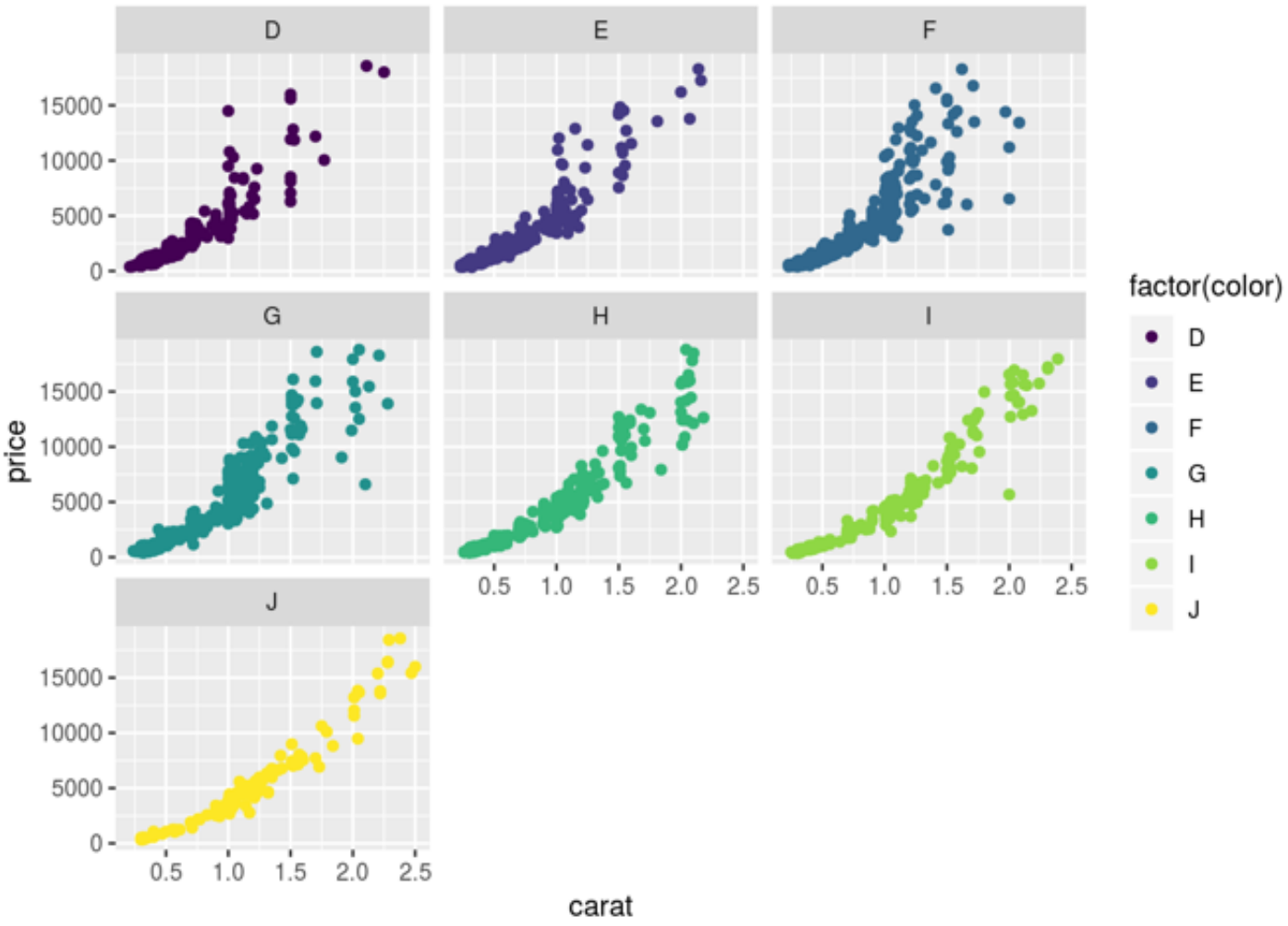
```
#ggplot(data=diamond_sub) + geom_point(mapping=aes(x=carat,y=price, alpha=1/10, shape
=factor(clarity)))
ggplot(diamond_sub, aes(x=carat, y=price, color=factor(clarity))) + geom_point() + fa
cet_wrap(~color)
```



8. Just for fun, map diamond `color` to the aesthetic color so that the points on each plot are a different color.

Answer:

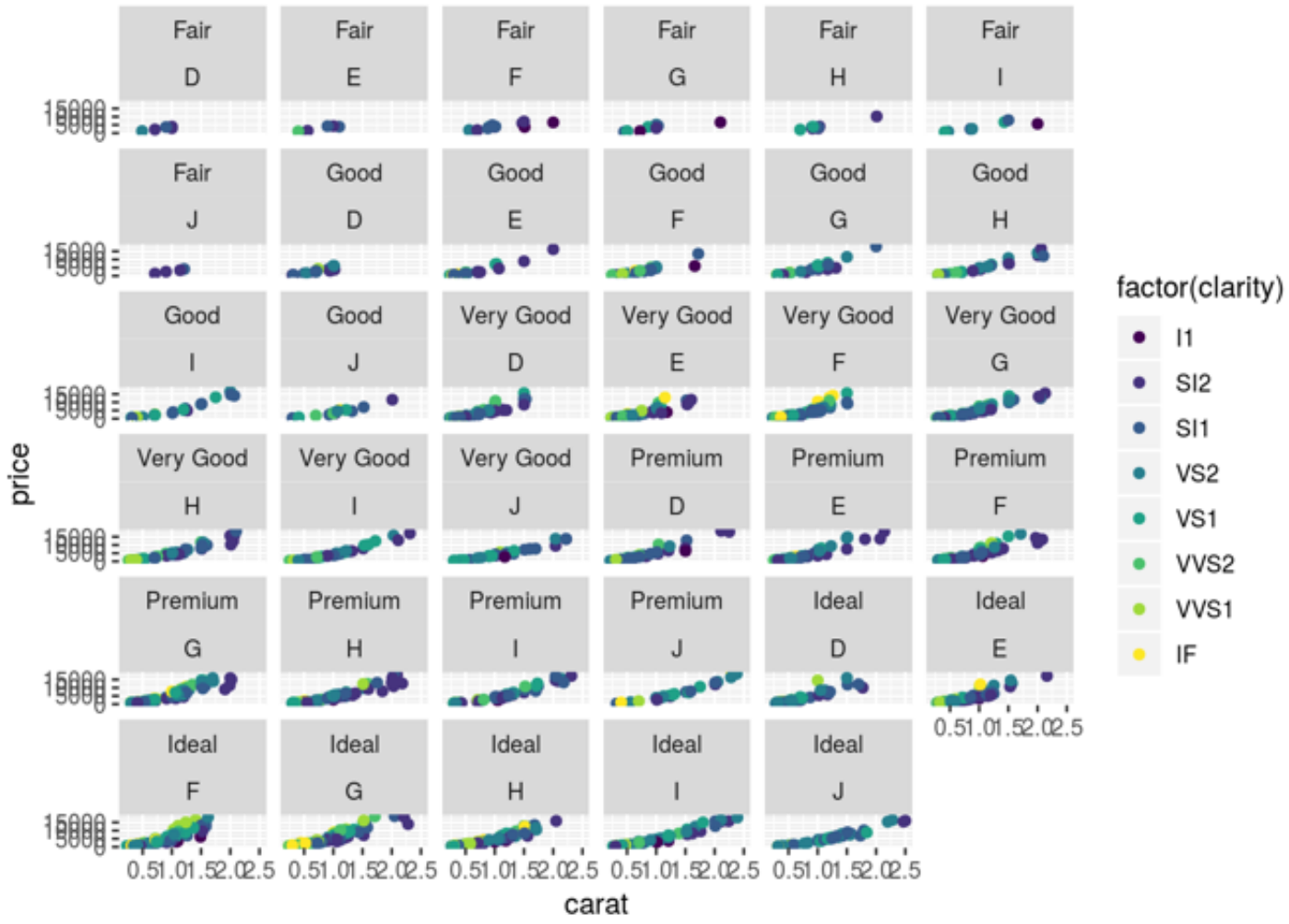
```
ggplot(diamond_sub, aes(x=carat, y=price, color=factor(color))) + geom_point() + face
t_wrap(~color)
```



9. Extend your plot in 7) to obtain a separate scatterplot of the relationship between `price` and `carat` weight for every combination of diamond `cut` and `color`. (Hint: You should have 35 individual scatterplots.)

Answer:

```
ggplot(diamond_sub, aes(x=carat, y=price, color=factor(clarity))) + geom_point() + fa
cet_wrap(cut~color)
```



10. Challenge: Create a brand new scatterplot of the relationship between `price` and `carat` weight but in this plot use aesthetics to color all points correponding to diamonds with depth (`z`) greater than 3.5mm a different color than those with smaller depths. What is striking about this plot?

Answer:

```
head(diamond_sub)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.73 Ideal     G     IF       62.3    56  4142  5.72  5.78  3.58
## 2  1.02 Premium   G     VS1      62      59  6238  6.45  6.39  3.98
## 3  0.32 Ideal     D     SI1      62      55   756  4.47  4.37  2.74
## 4  0.45 Ideal     G     VVS1     61.5    57  1297  4.91  4.95  3.03
## 5  0.3  Very Good H     VVS2     62      59   581  4.27  4.34  2.67
## 6  1.05 Premium   I     I1       61.8    56  2339  6.53  6.48  4.02
```

```
ggplot(diamond_sub, aes(x=carat, y=price, z=z, color=(z>3)))+ geom_point()
```