Section 7.5.1

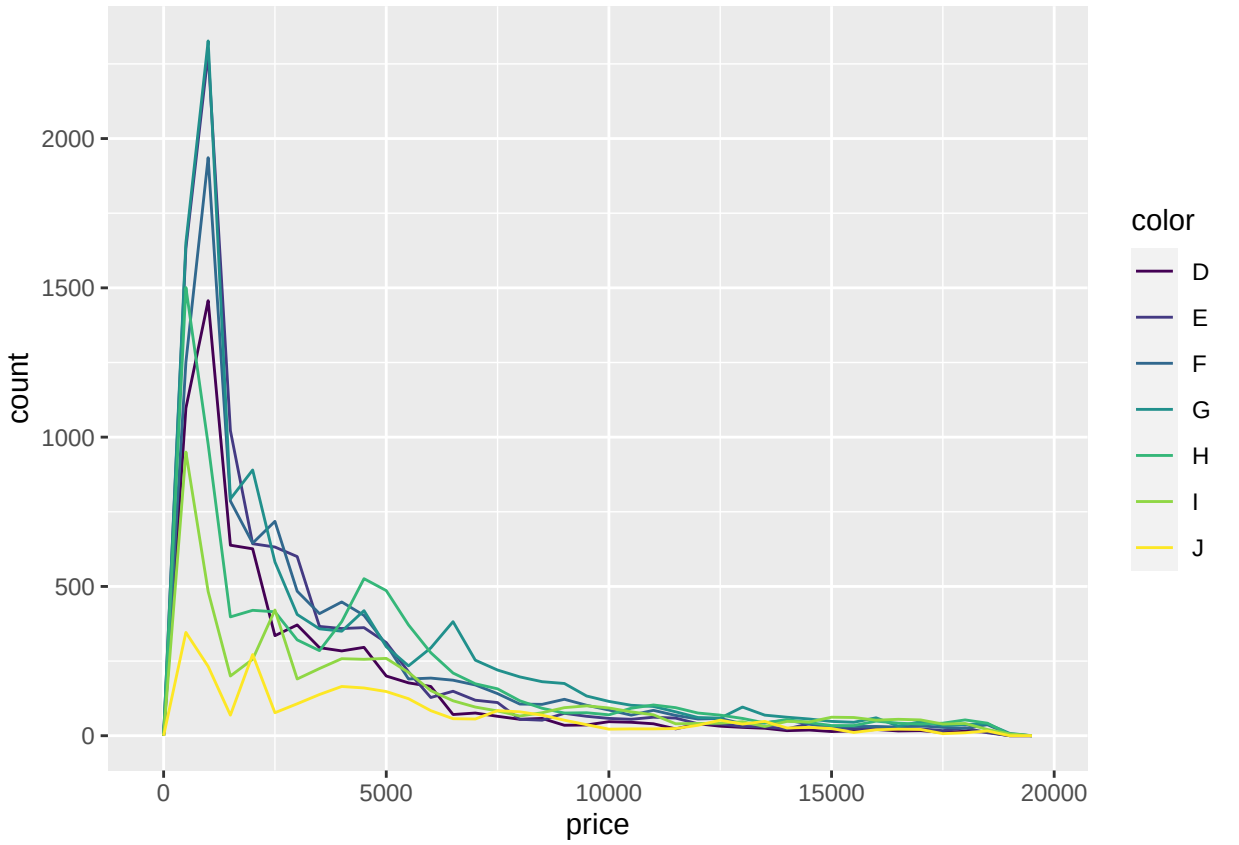
Teresa Burlingame

May 16, 2020

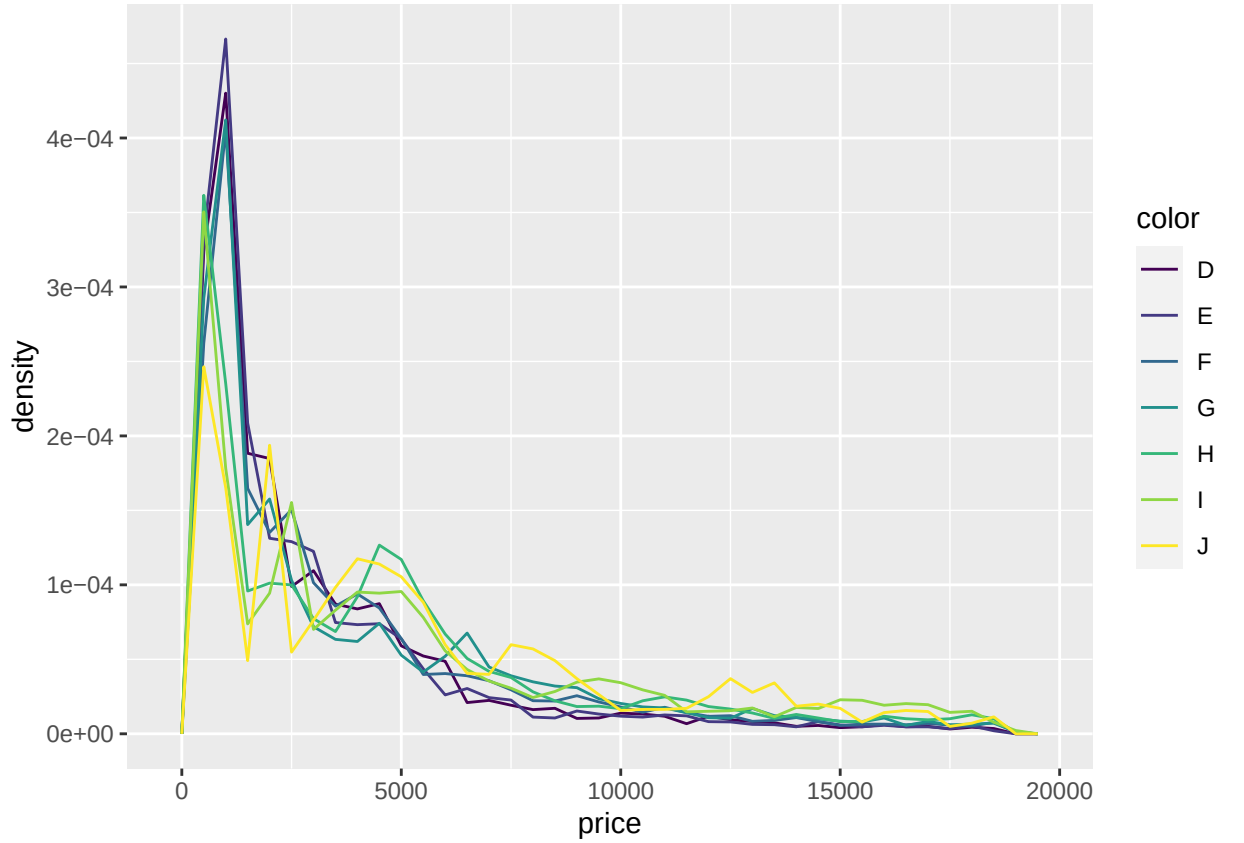
Covariation of a Categorical and a Continuous Variable

The book gave multiple examples of how to deal with a categorical and continuous variable that have covariation. One technique that was given was comparing the **density** of variables rather than the **count**. This can lead to more interpretable visualization when the counts across categorical variables are not very even. The below plots compare the same data of the diamonds data set, but rather than looking at count the second plot looks at density. You can see that colors H and I and J appear to more regulary have diamonds in a higher price point.

```
ggplot(data = diamonds, mapping = aes(x = price)) +
  geom_freqpoly(mapping = aes(colour = color), binwidth = 500)
```

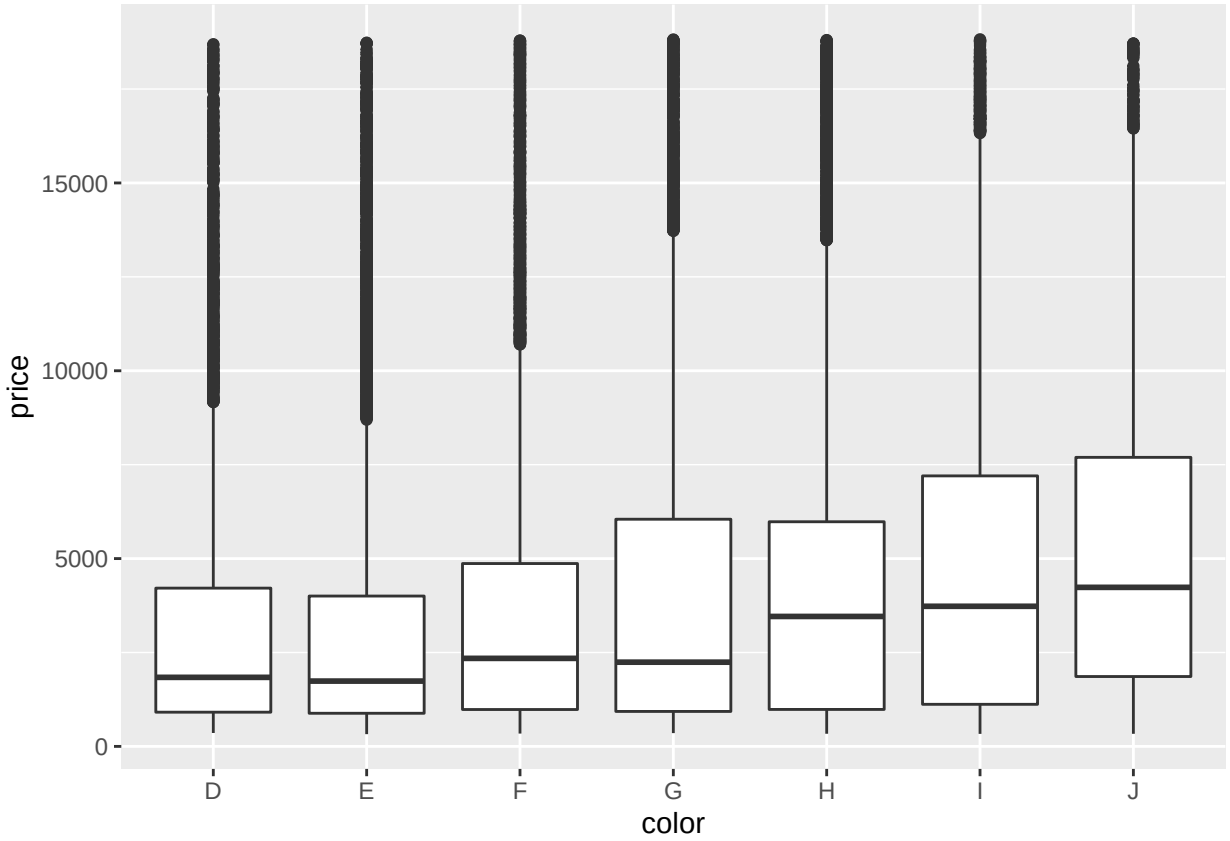


```
ggplot(data = diamonds, mapping = aes(x = price, y = ..density..)) +
  geom_freqpoly(mapping = aes(colour = color), binwidth = 500)
```



The freq_poly plots produced 7 categories, making it difficult to read. Another example would be a box plot. The box shows the 25th, 50th and 75th percentile, with outliers as individual points and points outside of those percentiles as whiskers. Below it is much more clear that H I and J have higher average prices, but they also have wider spreads.

```
ggplot(data = diamonds, mapping = aes(x = color, y = price)) +
  geom_boxplot()
```


price
15000
10000
5000
0
D
E
F
G
H
I
J
color