DataViz

Lab: ggplot2: ggplot 2-dim plots

1. Consider the following from our lab:

```
ggplot(mtcars, aes(x=wt, y=mpg)) +
  geom_point() +
  geom_smooth(method = lm)
```

This uses a “linear model” (lm) for the smoothing method. Is this a good choice? See the discussion at the link below (you only have to look at the methods lm, loess and gam).

https://stats.idre.ucla.edu/r/faq/how-can-i-explore-different-smooths-in-ggplot2/
https://www.reddit.com/r/datascience/comments/6psty1/differences_between_linear_model_lm_and_loess_in_r/

In the above example, they are using the different methods to explore the relationship between hp and displ in the mpg dataset. They also use different regression lines when grouping the data. We did that too:

```
ggplot(mtcars, aes(x=wt, y=mpg, color=cyl, shape=cyl)) +
  geom_point() +
  geom_smooth(method=lm)
```

2. Let’s try something similar with the Diamonds dataset.
str(diamonds) to look at the structure. Notice that there are some ordered factors, including cut and color. One could expect the price to go up with a larger carat size, a better color, and a higher-level cut. All other things being equal, you should expect to pay less for a “fair” cut diamond than for a “good” cut diamond, and less for a “very good” cut diamond than for a “premium” cut diamond, and so forth. So let’s play with that.

- Let’s try a scatter plot just plotting carat size against price. Sure seems to go up. Add a regression line (just “lm” for now. “loess” takes longer.) Yup. Goes up.
- Add grouping—color it by cut. Notice that cut is ordered. You should now have separate regression lines for each cut.
  →Do you notice anything interesting about that graph? Just from the data shown here, and all other things being equal, if you were choosing between a “very good” cut diamond and a “premium” cut diamond, and you wanted to save money, which would you buy?
- Before we examine that, try the same graph with method=”loess”. Look at how the graph curves down as the diamonds get really large. Why might that be? Look at 4-carat diamonds! What does this tell you about really large diamonds?
- So maybe “all other things being equal” isn’t a fair assumption. Maybe all other things aren’t equal. Let’s try to add another dimension that might explain this unusual behavior. What about the color of the diamond? Try making the color of the diamond affect the size of the dots. (Use “lm” again, since “loess” takes too long.

  - Yes, you are warned that using a discrete variable (color) may not be the best choice for the size of the dots. But let’s use it anyway. Run the graph.

- So this is confusing. Let’s change our variables around:
  - Make the color of the dots correspond to the color of the diamonds
  - Make the size of the dots correspond to the cut of the diamonds
  - Run the regression. OMG!! What are all those lines? And what do you think they mean? Hint: you have many groupings of color/cut combinations now!!
- Still not sure about the interplay here—what is really affecting price the most?
  - Try making the x-axis be cut, the y-axis be price, color=color, size=carat
  - Notice that the results are just columns. Why? (Maybe a scatter plot isn’t the best choice for this idea. What might be better for this type of x-variable?)
    →But it does seem to make a point. What is at the top of each column?
    Are the dots getting larger (bigger carats?) as you go up the column (price)
    Are the dots getting higher-quality color as you go up the column (price)?
- Just to throw a monkey wrench into the conversation, recall the graph of diamond buying patterns

  →Are there enough large, high-quality diamonds to make generalizations about them?

Big question: IF you had to choose one visual, of all the ones we looked at in lecture and lab, which would you choose to tell a story about diamonds and prices? Or would it be a selection of graphs?