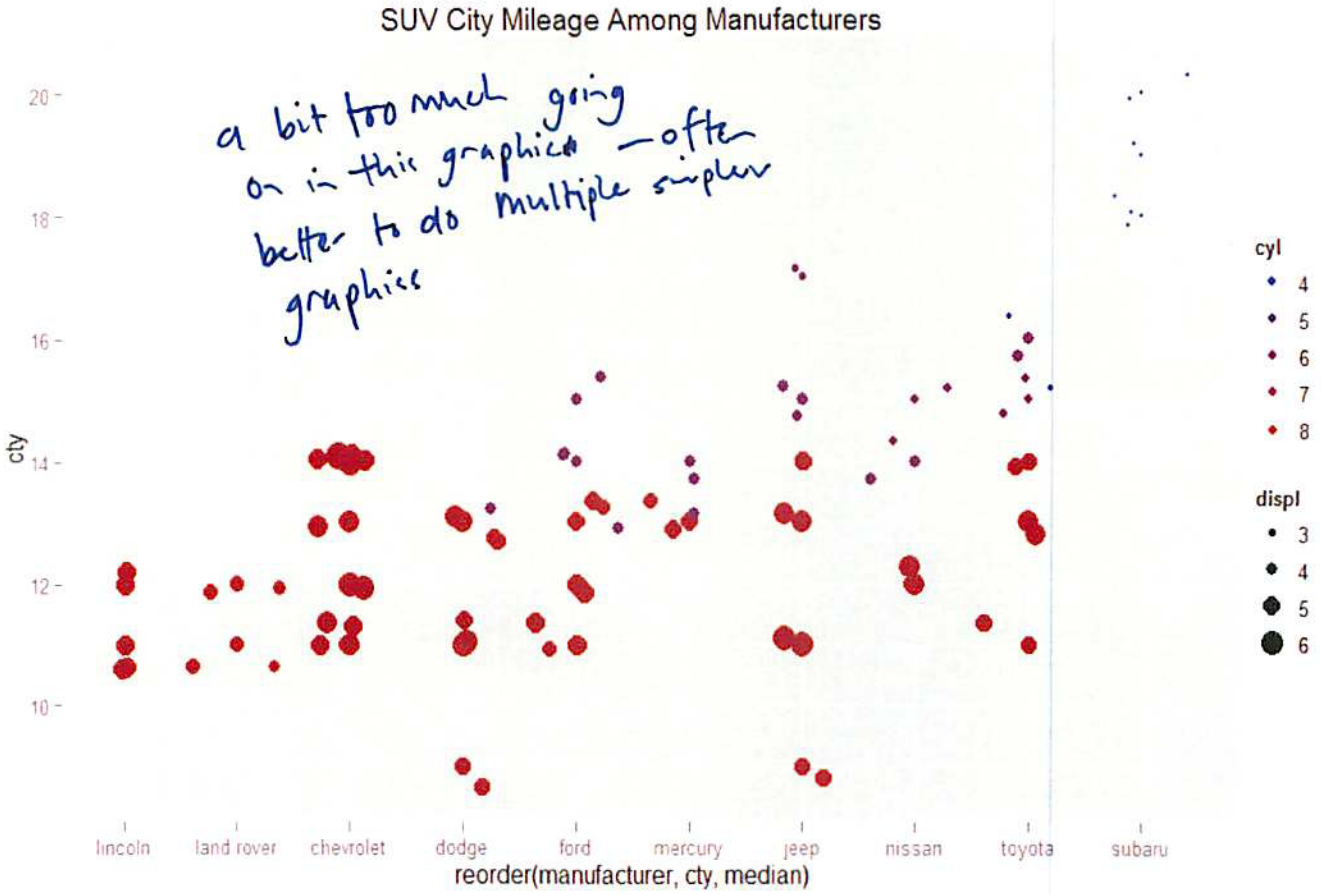
* C 4
S 3
0 4

STAT 405 Homework 1

by

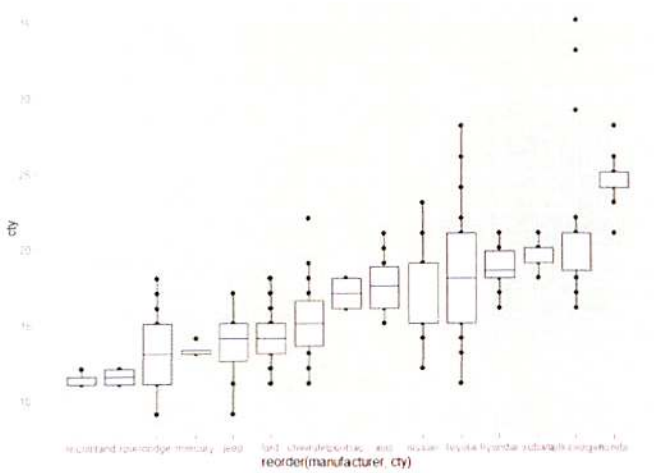


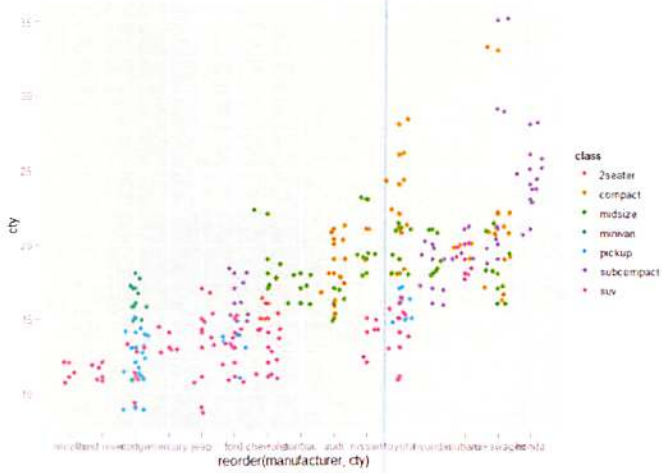

```
> suvs = mpg[mpg$class == "suv", ]
> qplot(reorder(manufacturer, cty, median), cty, data = suvs, size = displ, color =
cyl, main = "SUV City Mileage Among Manufacturers") + geom_jitter()
```

factor(cyl)

good questions

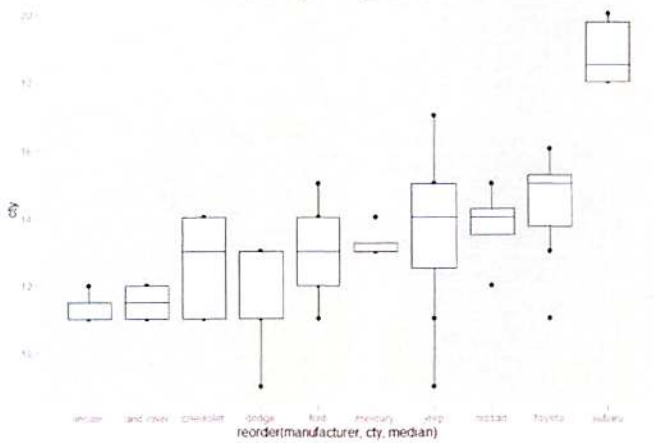
SUVs made by Subaru are more fuel-efficient than those made by other manufacturers. Are these SUVs simply of a higher quality than the others, or are there lurking variables? We can see the trend evidenced by the features of color and size for number of cylinders and engine displacement, respectively—cars with bigger, more powerful engines don't get as many miles to the gallon as those with smaller ones. All Subaru SUVs' four-cylinder, low-displacement engines are puny weaklings compared with the those typical SUV, but they can go farther with the same amount of gas. SUV's with six-cylinder engines tend to have city mileage somewhere between that of those with eight cylinders and those with four cylinders.

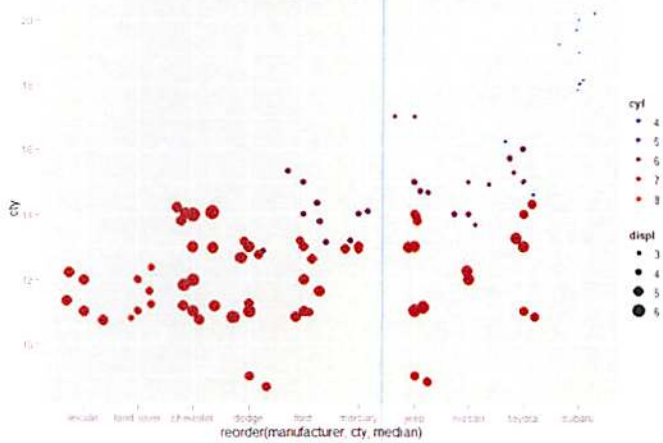




SUV City Mileage Among Manufacturers

SUV City Mileage Among Manufacturers

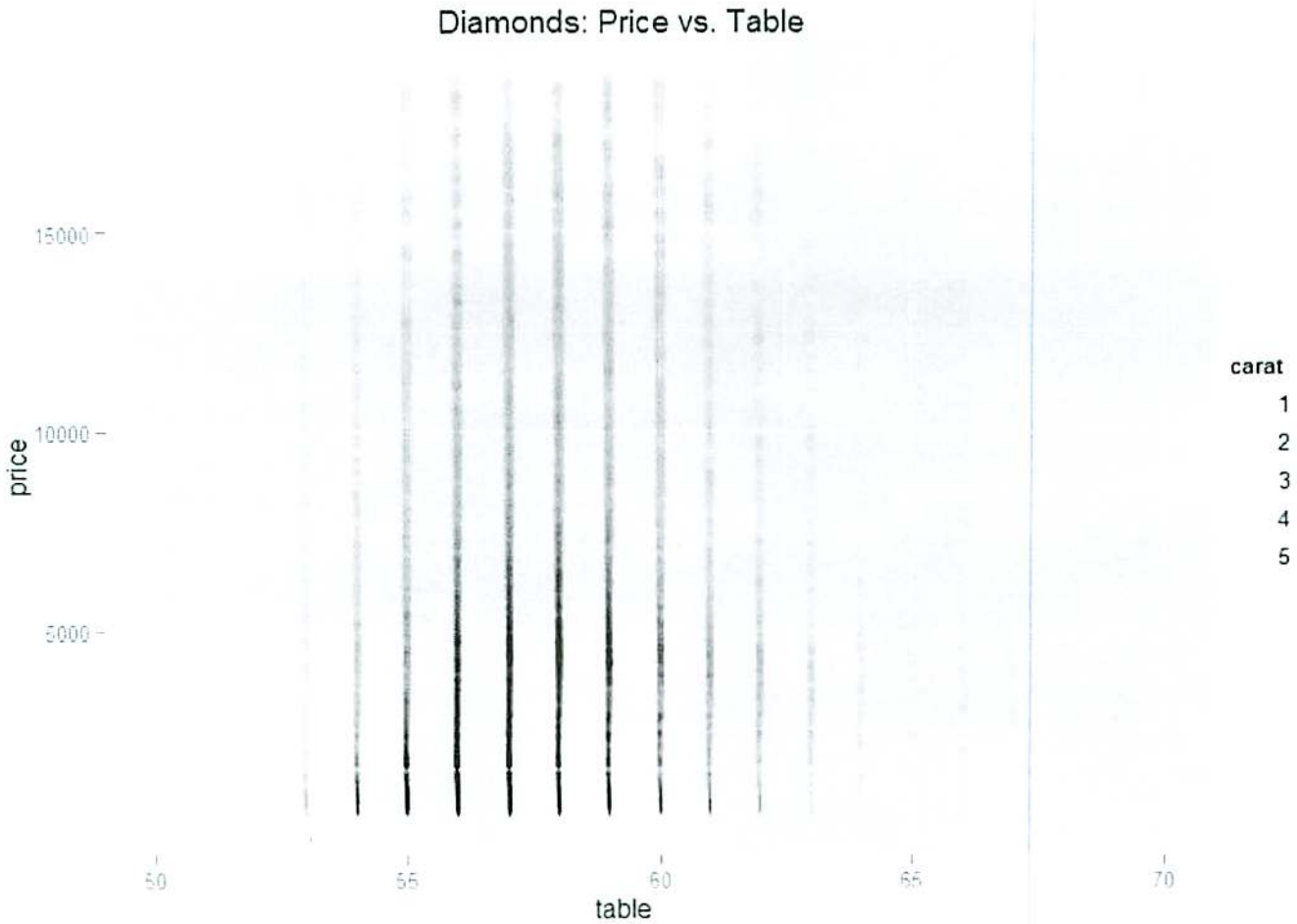





```
1. qplot(reorder(manufacturer, cty), cty, data = mpg) + geom_boxplot()
2. qplot(reorder(manufacturer, cty), cty, data = mpg, color = class) + geom_jitter()
3. qplot(reorder(manufacturer, cty, median), cty, data = suvs) + geom_boxplot()
4. qplot(reorder(manufacturer, cty, median), cty, data = suvs, size = displ, color =
cyl, main = "SUV City Mileage Among Manufacturers") + geom_jitter()
```

I started out by plotting cty versus manufacturer, hoping to find the company that makes the most fuel-efficient cars. This was not very useful because most manufacturers make several different classes of car, and we know that different classes vary greatly in mileage. I then made the same plot and set color to class, helping matters slightly. I tried using the boxplot and jitter to help make the graphs clearer, without success. There were too many variables for boxplot to work nicely. I got rid of the color feature for class and used facetting with class instead, at which point I decided to focus solely on SUVs, the class with the most entries in the dataset and made by more manufacturers than any other.

Obviously Subaru's line of SUVs got better mileage than the others', so I looked for answers. Using color for number of cylinders and size for engine displacement (as well as jitter) revealed some clues—Subaru SUVs have smaller engines than most other SUVs, and cars with smaller engines tend to have better fuel economy.



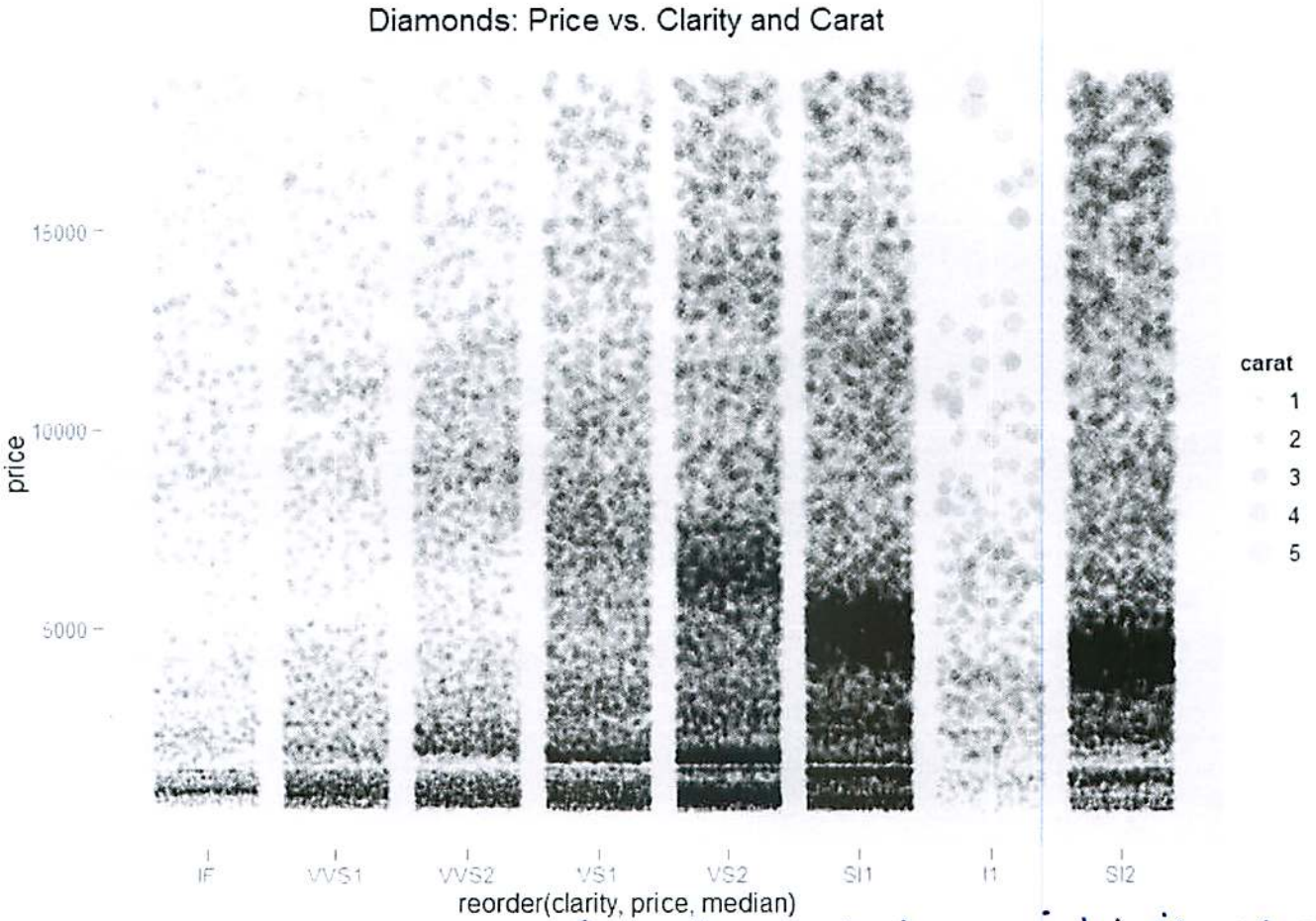

```
> qplot(table, price, data = diamonds, alpha = I(1/75), size = carat, main
= "Diamonds: Price vs. Table") + xlim(c(50, 70))
```

Does a diamond's table, the ratio of the diameter of its top surface to the diameter at its widest point, affect its price? In the above graph, the darkest portions of the bars for tables of 57 and 58 extend higher than those of their neighbors. We interpret that thediamonds with tables close to 57 or 58 tend to be priced slightly higher than smaller or bigger tables. Perhaps a table of 57-58 is the most desirable shape for a diamond, and sellers price such ideal diamonds accordingly. ✓

↑ if you look at diamond website there is some support for this hypothesis

Aside from showing the trend that bigger diamonds tend to be more expensive, the size feature, used with carat, shows that for a given price, diamonds with different tables tend to be roughly the same weight. However, we have not investigated the relationship between table and other variables, such as color, cut, or clarity. There might be another relationship among them that helps to explain the greater concentration of more expensive diamonds at table 57-58.

but need to be wary of confounding relationship with carat



↑ clarity already has an intrinsic order

```
> qplot(reorder(clarity, price, median), price, data = diamonds, geom = "jitter",
size = carat, alpha = I(1/10), main = "Diamonds: Price vs. Clarity and Carat")
```

The size feature with carat is helpful in elucidating the relationship between price and clarity; without it, all we would know is that some levels of clarity are much more common than others. The size feature lets us see that among diamonds at a given price, especially a very expensive price, clearer diamonds tend to be smaller than cloudier ones. The points near the top of the columns farther to the right tend to be bigger than the points at the top of the columns farther to the left. This is especially evident when looking at the I1(cloudiest) and IF(clearest) columns—a 5-carat I1 diamond costs the same as some 1 to 2-carat IF diamonds, even though carat is a much stronger predictor of price.

I like the idea but it's v. difficult to compare size with so much overplotting