Stat 480
Hmwk: Week 7

due March 1, 2007

Review of gplot Chapter

The chapter, "Getting Started with gplot," provided a good introduction of plots that can be made with gplot and what can be done with them. Some of these basic plots include scatterplots, line plots, bar charts, box plots, histograms, and density plots. Aesthetic attributes such as the shapes, sizes, and colors of points can be changed based on a specified variable. Shapes and colors are best used for the levels of categorical variables and sizes are best used for the magnitudes of continuous variables. Plots can be made with more than one variable plotted against the response, where the variables are plotted in different colors or styles. Smoothers and their standard errors can also be added to plots. Different types of plots, such as jitter and box plots or histograms and density plots can be plotted together. Facetting, another tool of gplot, is used to look at side-by-side scatterplots of levels of one variable versus another variable. This is used to look at conditional relationships.

The best thing about this chapter is the large number of examples of plots and the codes to create them. The plots help the reader visualize what is being described and the code provided gives the reader a starting point to explore further on their own with R. The combination of the written description, code and plot makes concepts easy to understand. Also, because I am currently taking a course in time series, I enjoyed the section about line plots and path plots in relation to time series. I had never heard of path plots before and it was interesting to see how they worked and how they differed from the traditional line plots.

I would have liked to have more information in the section about error bars and ribbons. I didn't completely understand what they were and how to use them. It would have been helpful if an example was included here. Also, I think the first place that the term "grob" was used was in the section about error bars and ribbons. I'm not sure what this term means and I couldn't find it defined anywhere in the chapter. One more thing I would have liked to see was a further explanation of the different types of smoothers. Because we've touched on them in my time series course, I know what they are, but I would have liked to see a further explanation of the different types and when they would be useful.

✓

Good.

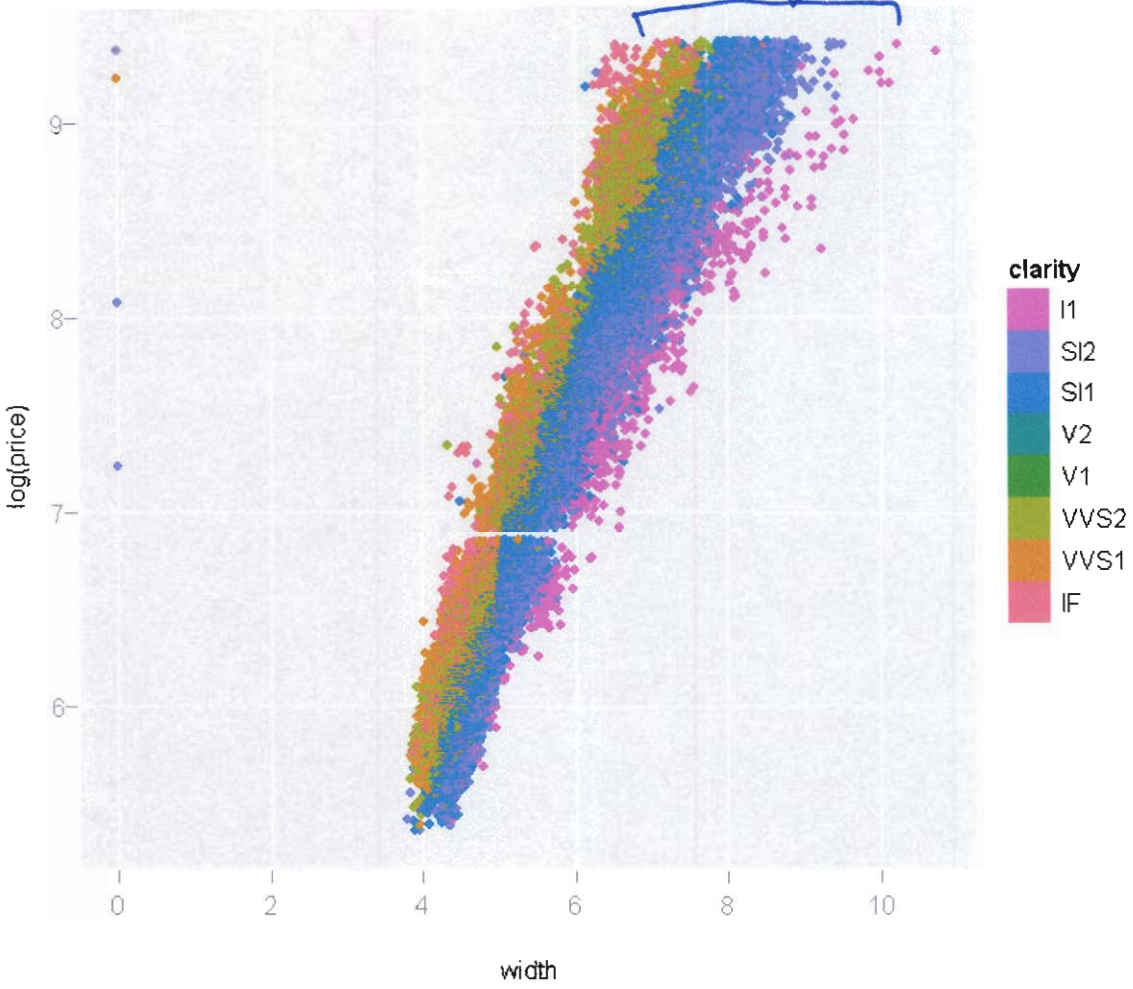
Good points

* C5
S4
O5

Interesting Plots of Diamonds Data

Plot 1
Scatterplot: Log of Price vs. Width, with Clarity Levels in Color
Code: qplot(width, log(price), data=d, colour=clarity)

interesting pattern of colours here.



When exploring how each of the continuous variables effect price, I found that there was an exponential relationship between price and width. After taking the log of price, I discovered the above plot. From looking at the data, I discovered that there were no diamonds with prices between $950 and $1000 (whose log is between 6.85 and 6.9, where the gap shows up in the plot). This gap in the price did not appear with the original data, but showed up when I took the log.
I am not sure why there is a jump in price of this size. I don't know if it is simply because sellers tend to round up to $1000, or if there's more to it than that. There is no other gap in price anywhere near this size in the rest of the data.

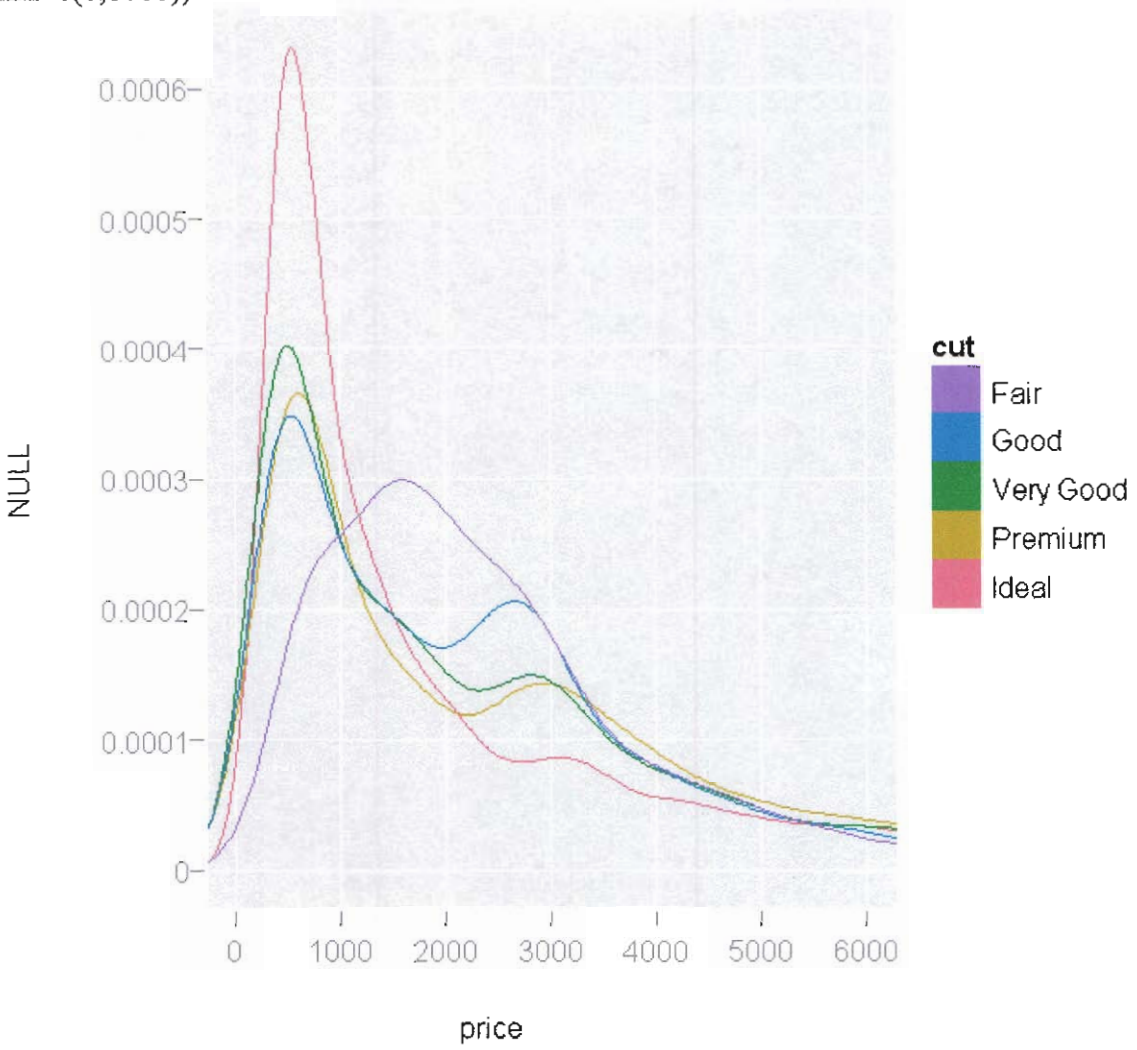
✓

Good spotting. I think its a data collection error by me.

Plot 2

Density plot: Price by Cut

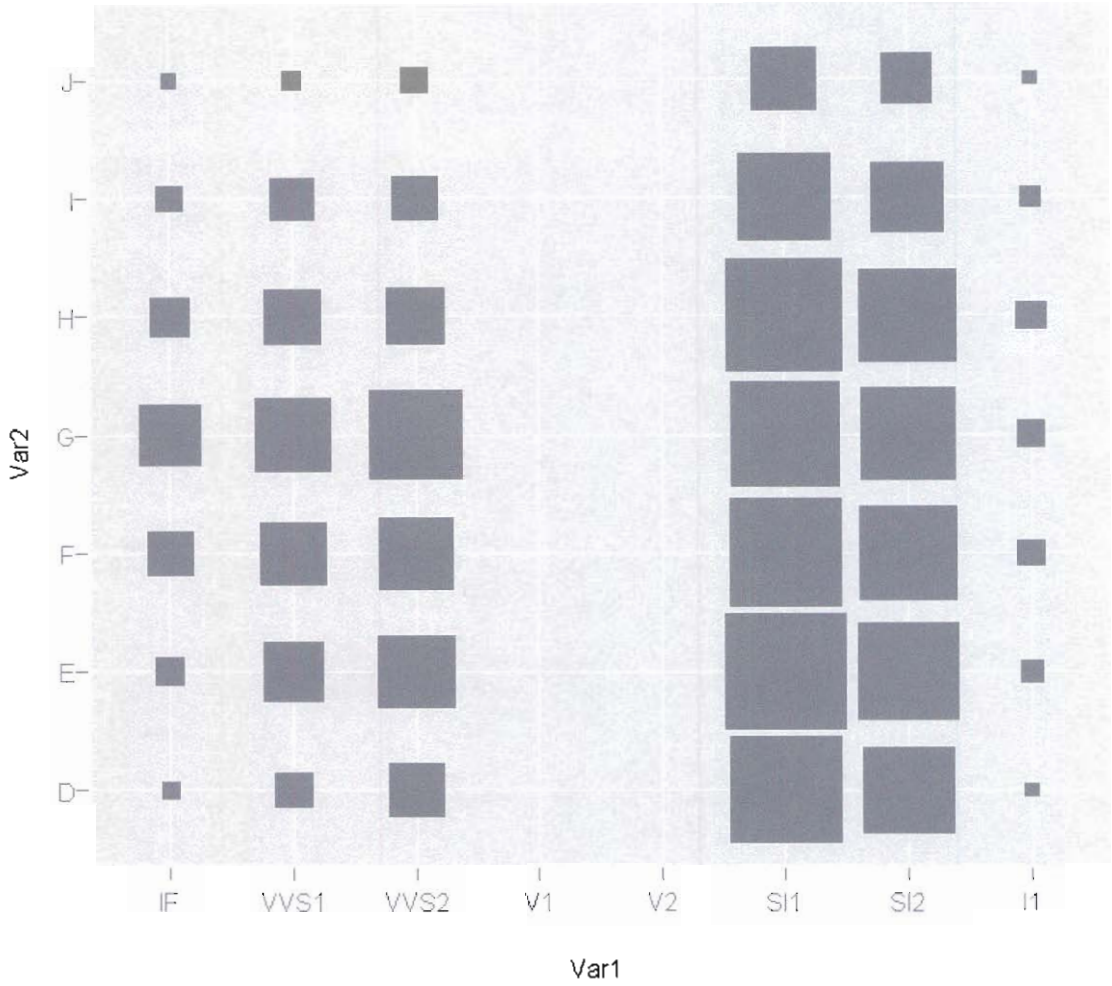
Code: qplot(price, data = d, type = "group", grob = "density", id = cut, colour = cut, xlim=c(0,6000))



From this plot, the peak in the price of observations of the four highest quality cuts (Ideal, Premium, Very Good, and Good) is a lower price than the peak for the lowest quality cut (Fair). Also, there are more, cheaper diamonds of the higher quality cuts, than the lowest quality. The four highest qualities tend to have a similar trend in price, which the trend for Fair cut diamonds is much different. The only thing I can think of that might be causing this is that maybe there are more diamonds sold of the higher quality cuts, so they are cheaper. Fair quality cut diamonds are more rare, so they are more expensive.

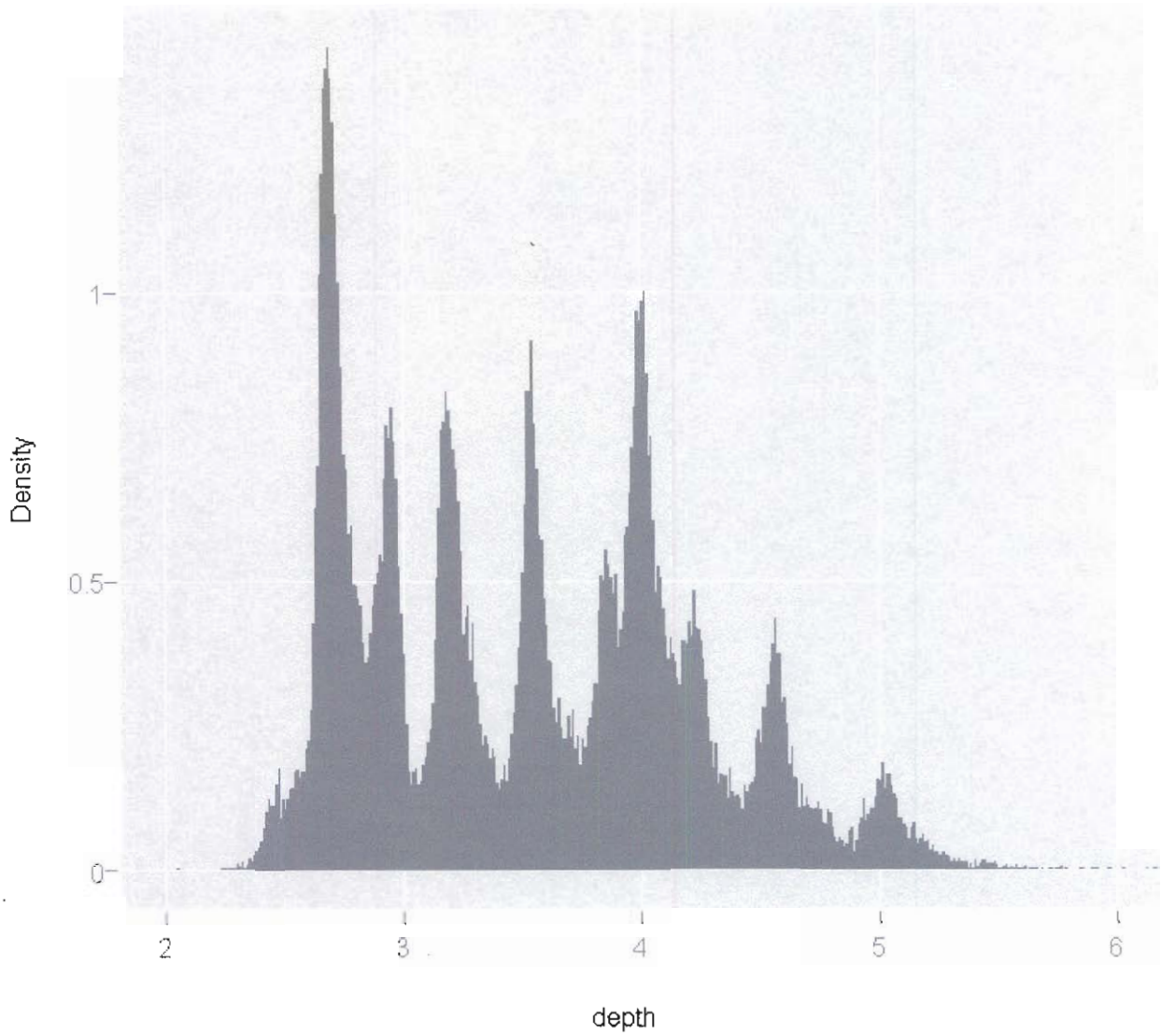
Or maybe it's related to the other variables in the data set.

Plot 3
Fluctuation Diagram: Color vs. Clarity
Code: ggfluctuation(table(d$color, d$clarity))



It is interesting that the boxes for the clarity levels V1 and V2 are so small that they can't be seen. There must be very few diamonds with clarities of V1 and V2 of any color. Also, I noticed that for the higher clarities of diamonds (IF, VVS1, and VVS2), the most common color is G, which is on the high end of the nearly colorless scale. Very few diamonds are that are virtually flawless are also colorless, which is probably why they're more expensive. Also, many colorless diamonds appear to internal flaws (SI1) that cannot be seen with the naked eye.

Plot 4
Histogram: Depth
Code: qplot(depth, data=d,type="histogram",breaks=seq(0,32,by=.01),xlim=c(2,6))



By decreasing the bin sizes as small as I could, I noticed that there are peaks in depth of the diamond. I have a feeling that these peaks are caused by rounding the value of depth to two decimal places.

– there's too few of them to be caused by that. I think the peaks correspond to the popularity of different carat sizes.