*C 4
S 4
0 5

Statistics 405 Homework 1

September 1, 2009

Introduction

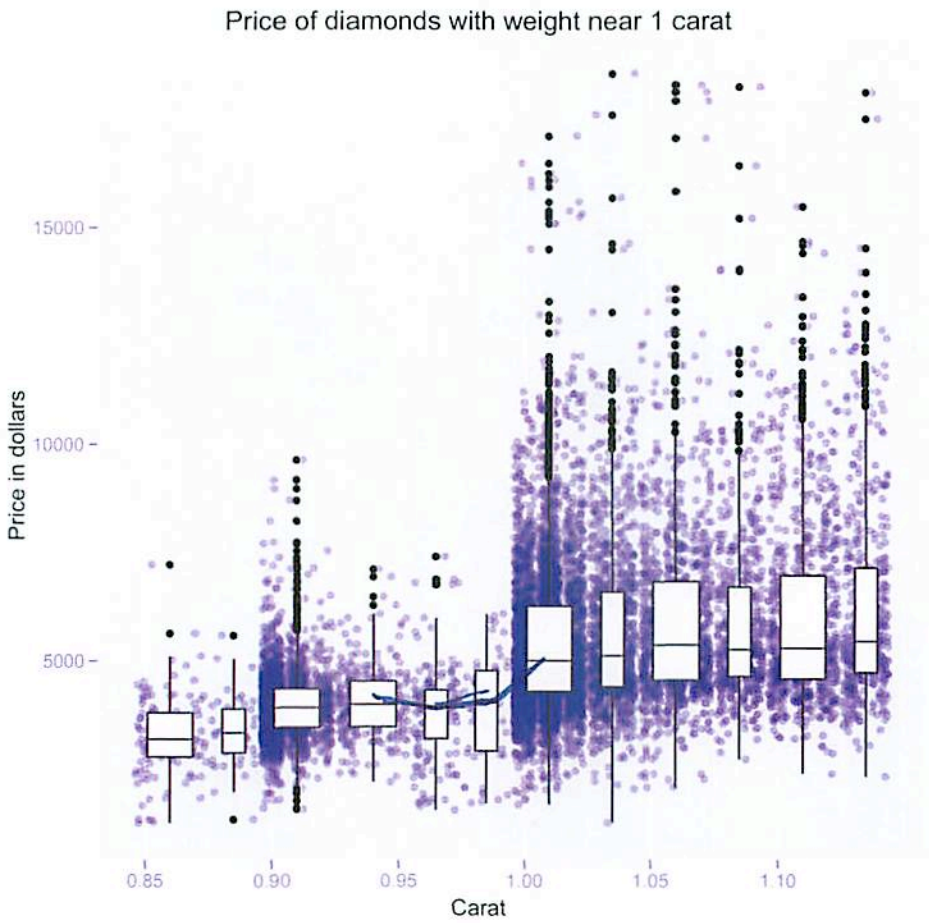
The diamonds dataset includes values for the price, carat weight, cut grade, color, clarity, and physical dimensions of over 50,000 diamonds. The initial analysis in class revealed that for the dataset overall, there is a strong positive relationship between the carat weight and the price of a diamond, which emerges far more clearly than any possible influence of cut or clarity on price.

The wikipedia article on diamonds suggests a topic for further study, pointing out that due to the preference for stones around 1 carat, savvy consumers can get more for their money by purchasing a diamond with weight slightly under 1 carat, or can get a better cut by buying a diamond with weight over 1 carat since stones right at 1 carat are more likely to be poorly cut in order to preserve their weight. With this inspiration, I decided to focus in on the data for diamonds with weight in the neighborhood around 1, and began by creating a new data frame in R:

great idea

```
near_one_carat <- diamonds[diamonds$carat >= .85 &
  diamonds$carat < 1.15, ]
```

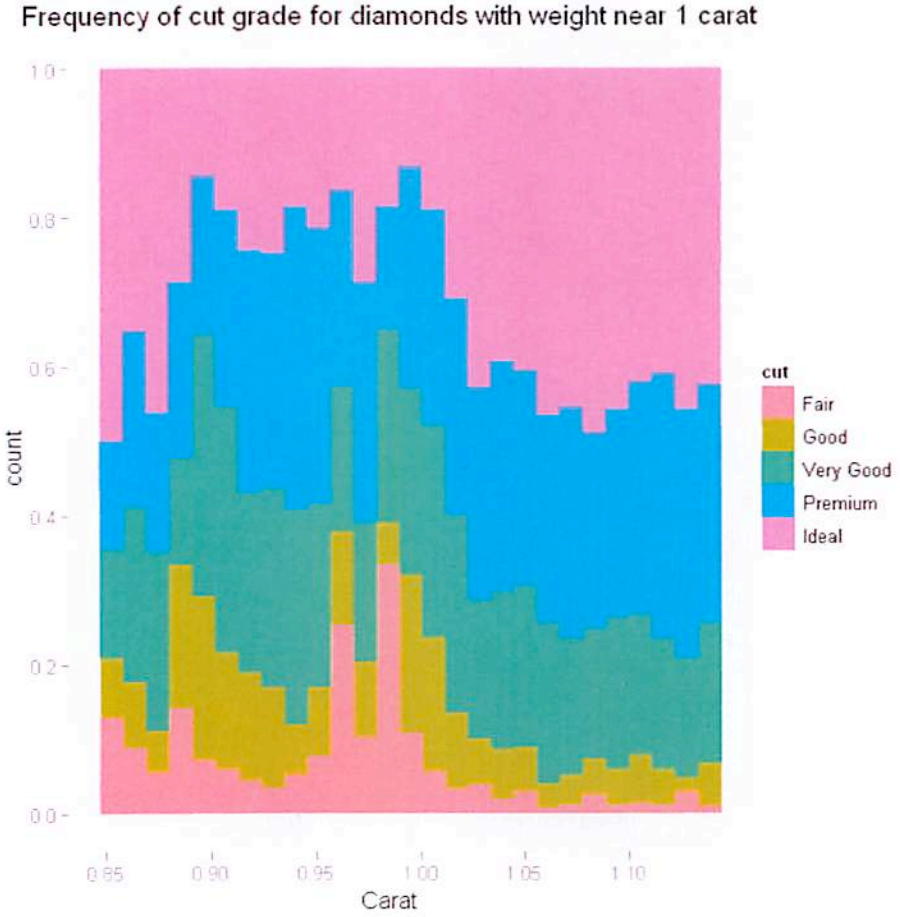
This smaller data frame has observations for 11,514 diamonds, so the techniques for dealing with large data sets are still needed to create effective plots of this subset.




```
qplot(carat, price, data = near_one_carat,
  geom = "blank", group = round_any(carat, .025, floor),
  xlab = "Carat", ylab = "Price in dollars",
  main = "Price of diamonds with weight near 1 carat") +
  geom_jitter(colour = "#4B0082", alpha = .3) +
  geom_boxplot()
```

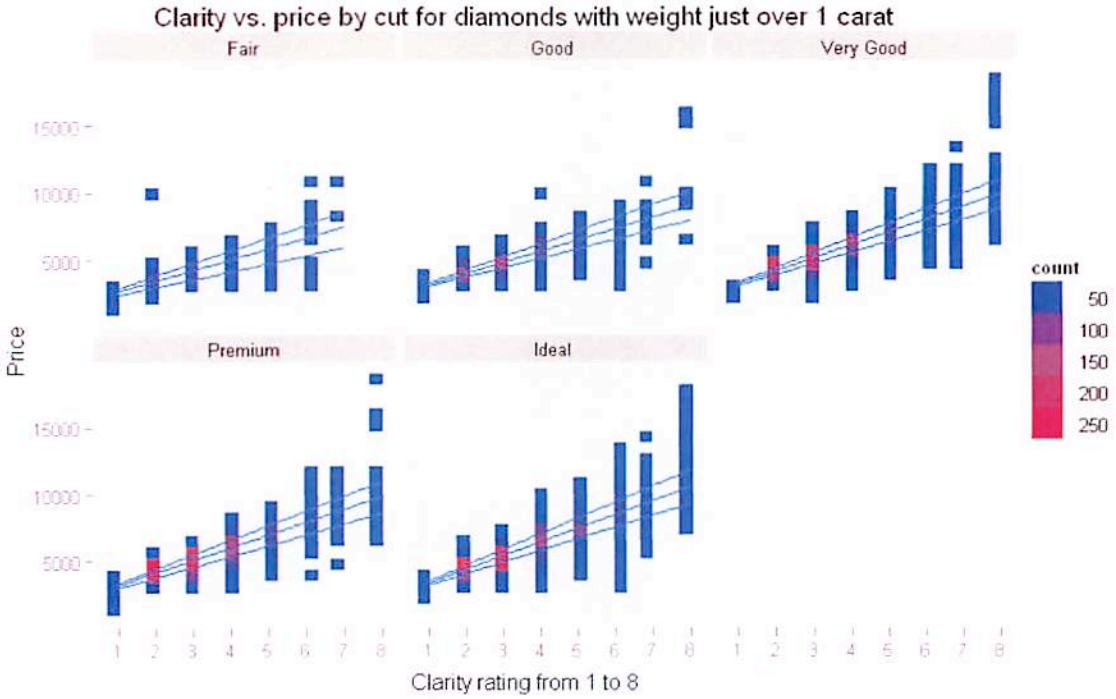
The points in the scatterplot are jittered and made slightly transparent so that overlapping data is visible. The carat values are discretized via rounding so that boxplots can be used to show the trend in the medians and spreads for the data points that fall into each interval. Together these components of the graph make it clear that not only is there an increase in the volume, spread and price at 1 carat, there is also a slight and unexpected *decrease* in the median price around .95 carats which goes against the positive relationship between carat and price for the dataset overall.

it's pretty tiny!




```
qplot(carat, data = near_one_carat, binwidth = .011,
  xlab = "Carat", ylab = "Frequency",
  fill = cut,
  main = "Frequency of cut grade for diamonds with weight near 1 carat",
  position = "fill")
```

This plot utilizes stacked bars to show the proportions of the various cut grades for diamonds with carat weight around 1, revealing peaks in the frequency of lower quality cuts for diamonds with weight between .9 and 1 carat, followed by a decrease in the proportion of lower quality cuts above 1. The data supports the advice on wikipedia that diamonds with weight a bit over 1 are more likely to have a higher quality cut than diamonds with a weight of one carat, and additionally indicates that the proportion of quality cuts for diamonds with weight a bit less than 1 is even lower than for diamonds with weight 1.

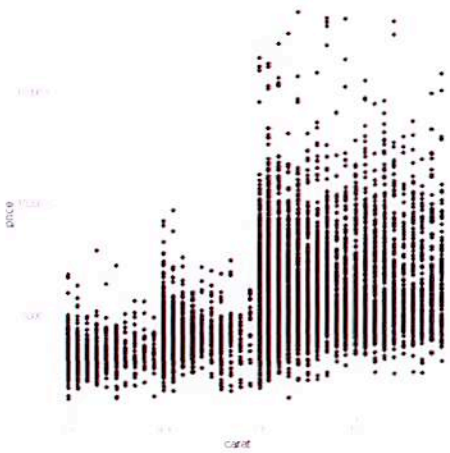



```
just_over_one_carat <- diamonds[diamonds$carat >= 1 & diamonds$carat < 1.10, ]
just_over_one_carat$clarity_rating <- (just_over_one_carat$clarity == "I1") +
  (just_over_one_carat$clarity == "SI2") * 2 +
  (just_over_one_carat$clarity == "SI1") * 3 +
  (just_over_one_carat$clarity == "VS2") * 4 +
  (just_over_one_carat$clarity == "VS1") * 5 +
  (just_over_one_carat$clarity == "VVS2") * 6 +
  (just_over_one_carat$clarity == "VVS1") * 7 +
  (just_over_one_carat$clarity == "IF") * 8
qplot(clarity_rating, price, data = just_over_one_carat, geom="bin2d", bins=20,
  xlab = "Clarity rating from 1 to 8", ylab="Price",
  main = "Clarity vs. price by cut for diamonds with weight just over 1 carat") +
  facet_wrap(~ cut) + geom_quantile()
```

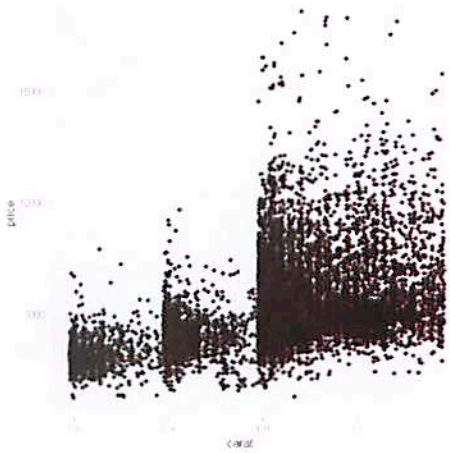
Once again I narrowed my data set, now focusing in on diamonds with carat weight in the range from 1 to 1.1. By faceting the plot by cut, I was able to show the expected positive association between clarity and price for diamonds of similar weight and identical cut. For this plot, I used the heat map effect to make the density of the data points visisble, and created a numeric clarity rating from 1 (worst) to 8 (best) to allow the addition of the quantile lines, which make the uniformly positive trend of clarity and price much clearer.

I think qplot(clarity, price, data=just...) + facet_wrap(~cut) would have been much easier & probably more informative

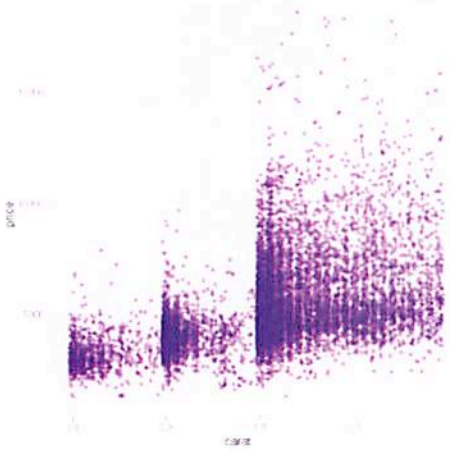
Iterations in creating my first plot



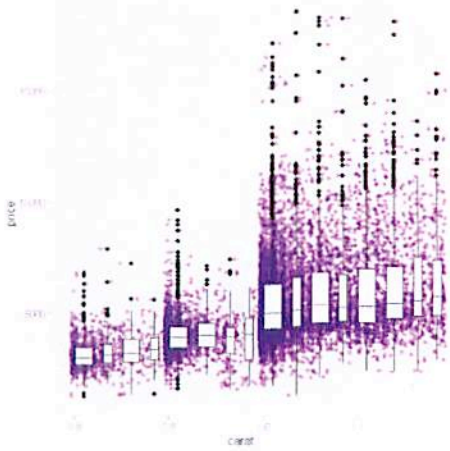
My initial plot of carat vs. price is overplotted.



Adding a jitter helps, but the data still seems to be overlapping in a dark mass.



Now the scatterplot looks nicer, but patterns in the data aside from a general positive assocation are hard to identify.



Adding boxplots shows the trend in the median once the rounding is tweaked to the appropriate level.

The following sequence of commands was used to create these plots:

```
near_one_carat <- diamonds[carat >=  .85 & carat < 1.15, ]
qplot(carat, price, data = near_one_carat)
qplot(carat, price, data = near_one_carat, geom = "jitter")
qplot(carat, price, data = near_one_carat, geom = "blank") +
  geom_jitter(colour = "#4B0082", alpha = .3)
qplot(carat, price, data = near_one_carat, geom = "blank",
```

```
group = round_any(carat, .025, floor)) +
geom_jitter(colour = "#4B0082", alpha = .3) +
geom_boxplot()
```

Conclusion

To study this data further, a variable categorizing the diamonds as either real or synthetic would be useful. Since synthetic diamonds often have better color and clarity than natural diamonds at a lower price, faceting the data based on this factor might help reveal the separate pricing trends for the two varieties. Data from various points in time would also be enriching, in particular, information from both before and after the 2004 antitrust conviction of De Beers for fixing prices. Clearly, price fixing would be a lurking variable affecting price which we were unable to consider here.