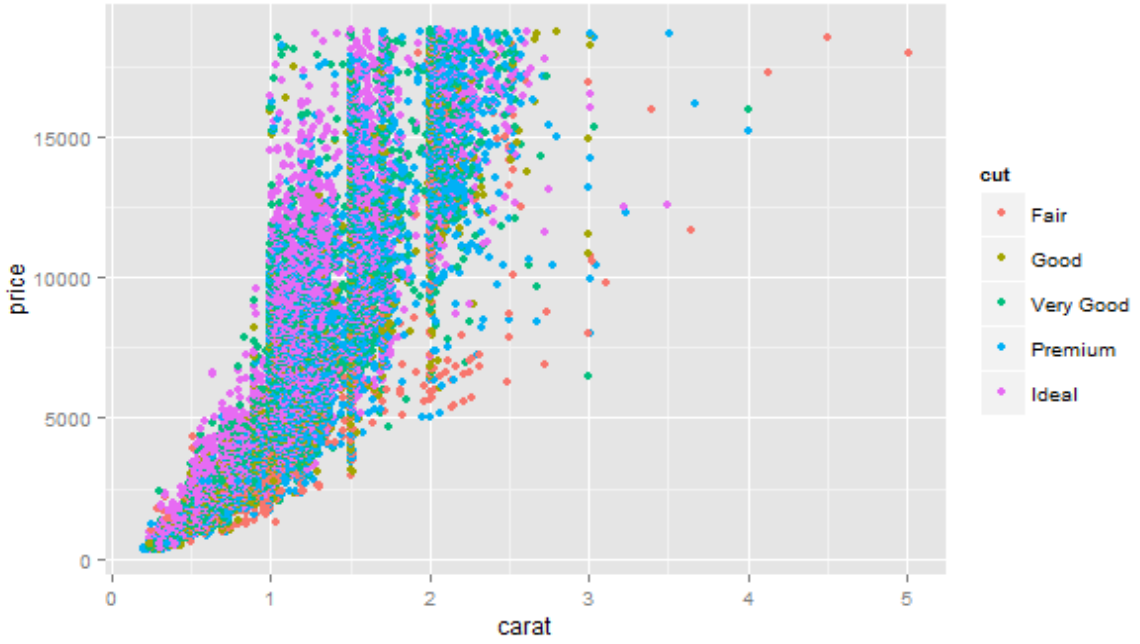
STAT 405
Frank Portman
Homework 1

Summary

The 'diamonds' dataset is included in ggplot2 and contains the prices, cuts, clarities, carats, and other various size measurements of almost 54,000 diamonds, scraped from diamondse.info. An initial naïve analysis confirmed my intuition that price and carat are positively correlated. In order to deal with some of the overplotting issues I encountered, I turned to a boxplot geom. The boxplot revealed that there were many interesting differences between diamonds over and below 1.5 carats. A more thorough exploration highlights the differences precisely and offers some explanations for their existence.

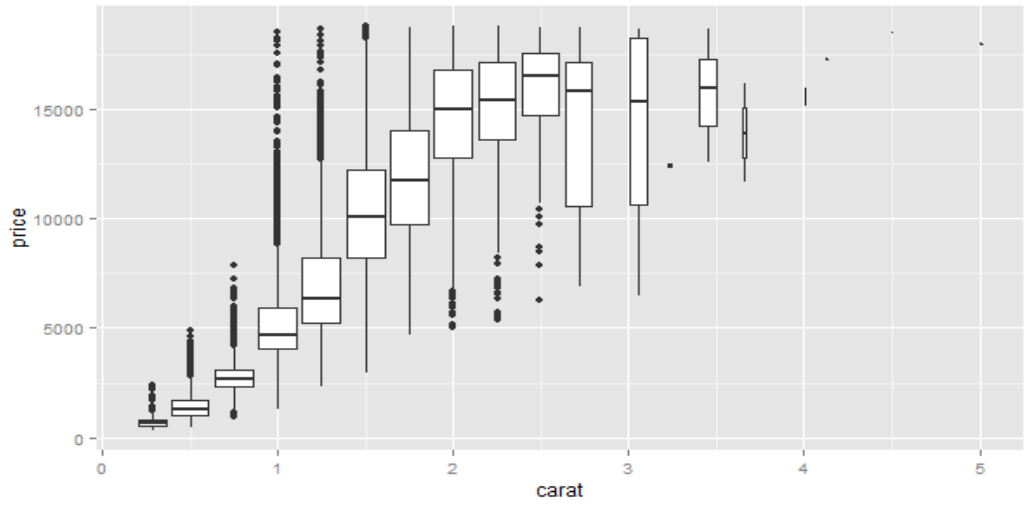
I also explored the 'mpg' dataset, albeit less thoroughly, by considering the effect transmission has on fuel economy. The results were close with what I would expect although with such a small sample size we cannot really make a definitive claim.

Plots & Discussion



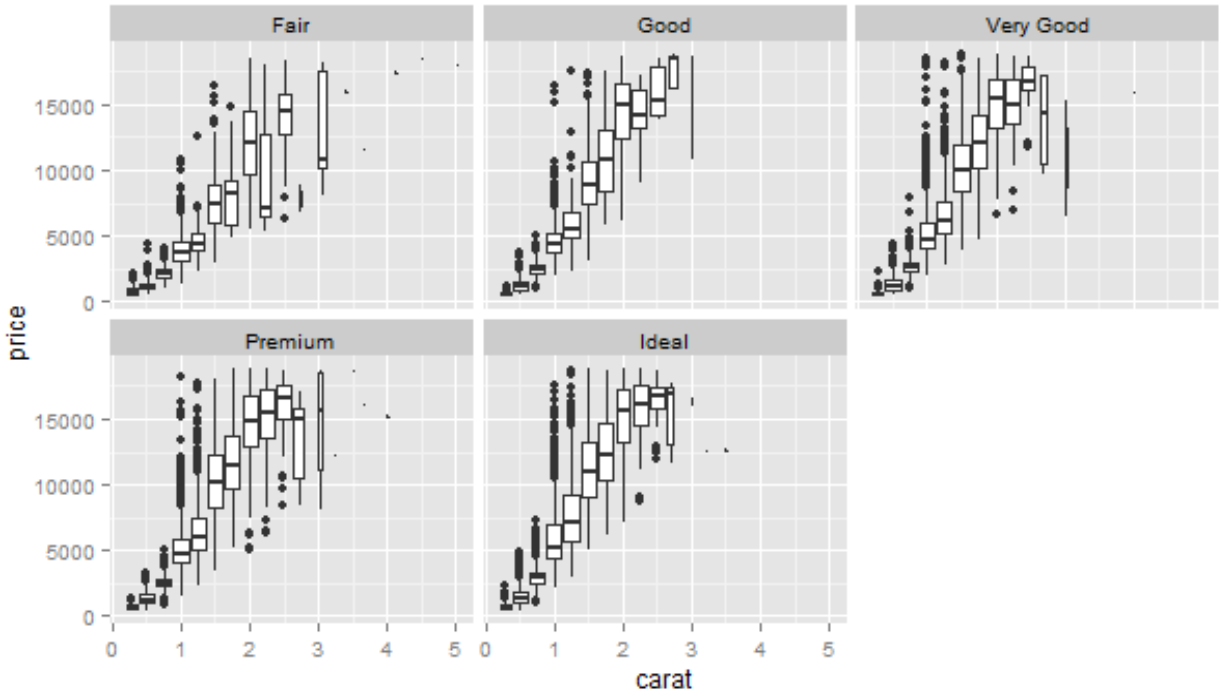

```
qplot(carat, price, color = cut, data = diamonds)
```

Iteration 1: We had already seen in class that price and carat had a strong positive correlation. I decided to explore how 'cut' was related with these two variables. This image suffers from extreme overplotting so I had to consider other options.




```
qplot(carat, price, data = diamonds, geom = "boxplot",
      group = round_any(carat, .25))
```

Iteration 2: Before trying to discern specific patterns by cut, I first attempted to see whether a boxplot was helpful in deciphering the carat vs. price data. The boxplot does, in fact, make it easier for us to examine the data and a few interesting facets about the data can be extracted from this iteration. For instance, most of the outliers corresponded to carats below 1.5 and were mostly overpriced.



```
qplot(carat, price, data = diamonds,
      geom = "boxplot", group = round_any(carat, .25)) +
      facet_wrap(~cut)
```

First Plot

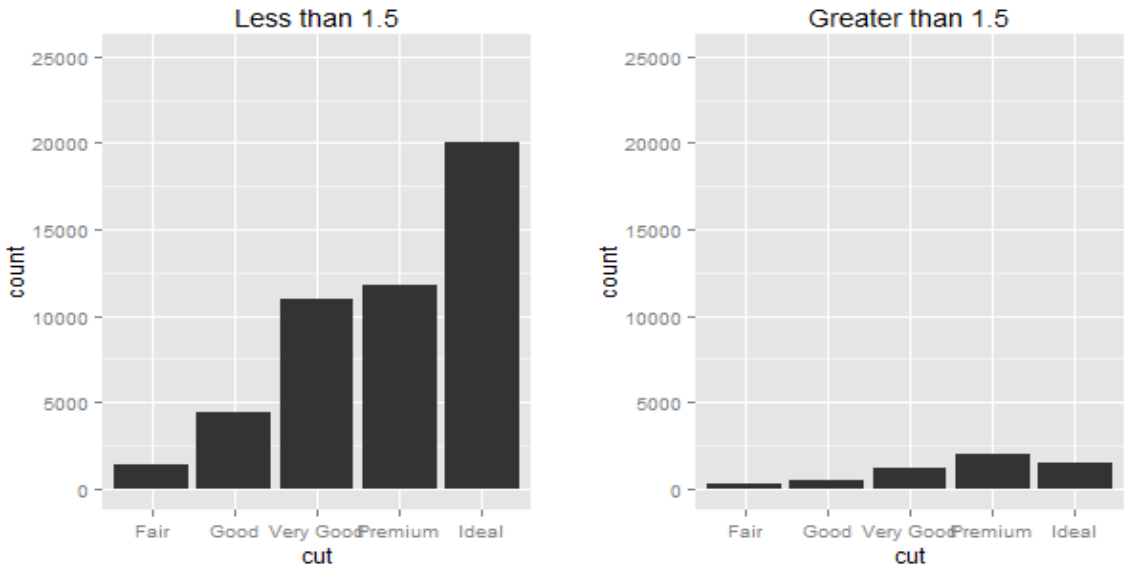
Final Plot: Finally, a facet wrap in conjunction with a boxplot allows us to truly grasp the data. It becomes clear that our observations in 'Iteration 2' are not exclusive to certain cuts. Most cuts, with the exception of ideal, seem to behave erratically around 2-2.5 carats after seeing exponential growth in price. The ideal cut tapers off smoothly at higher carats while the rest display some degree of volatility in their medians.

We have unearthed some fascinating results but it is still unclear as to why most of the outliers are of carats less than 1.5. In order to explore some possibilities, I first subset the diamonds dataset in two ways:

```
greater_than_1.5 <- subset(diamonds, carat > 1.5)
 less_than_1.5 <- subset(diamonds, carat <= 1.5)
```

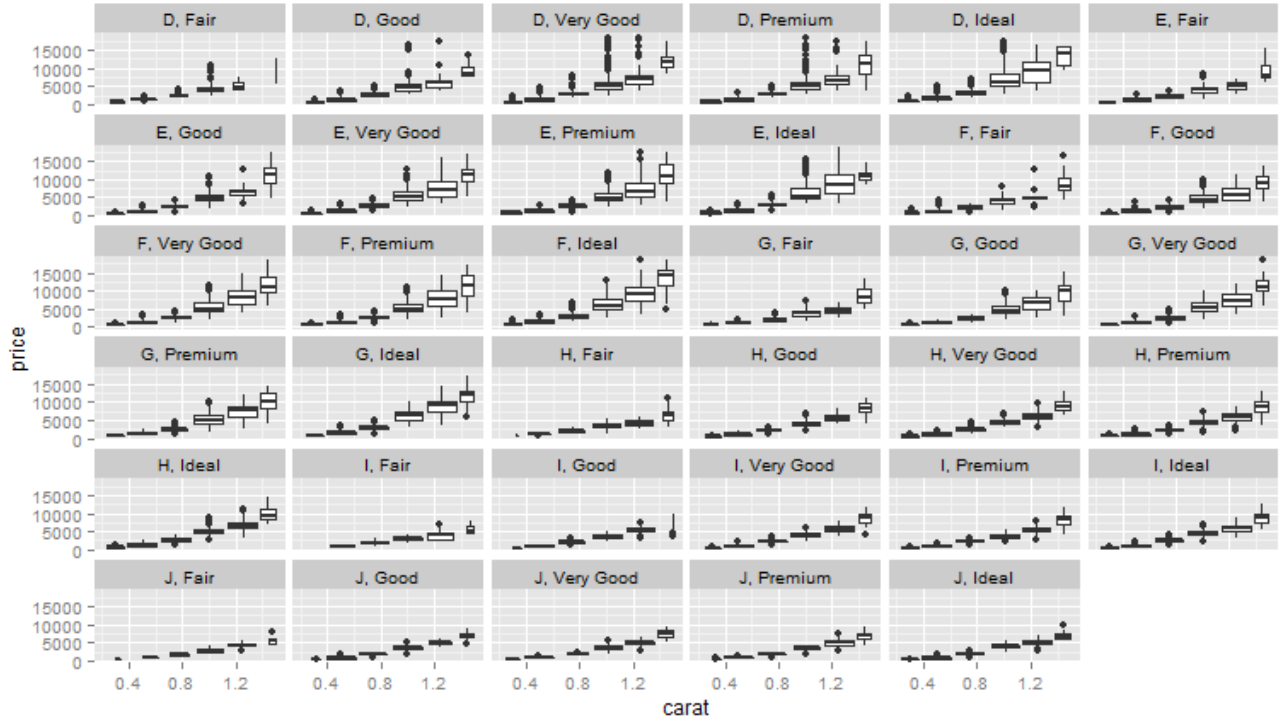
I then considered whether the lower carat diamonds had better cuts at higher frequency. That might explain why the higher priced outliers belonged to the lower carat subset.

```
p1 <- qplot(cut, data = less_than_1.5) + ylim(0,25000) +
ggtitle("Less than 1.5")

p2 <- qplot(cut, data = greater_than_1.5) + ylim(0, 25000) +
ggtitle("Greater than 1.5")

grid.arrange(p1, p2, nrow = 1)
```



This graph isn't very informative but it does confirm my suspicions that smaller carat diamonds had higher quality cuts. With this in mind I decided to explore this route further.

```
qplot(carat, price, data = less_than_1.5,
      geom = "boxplot", group = round_any(carat, .25)) +
      facet_wrap(color~cut)
```



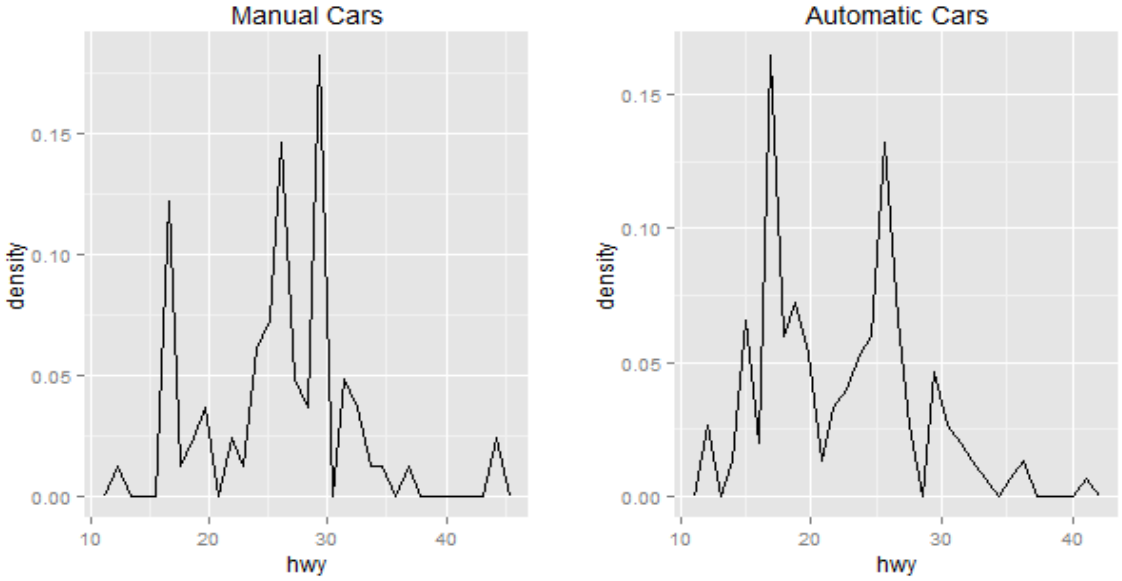
Second Plot

While this plot may seem cluttered and uninformative at first glance, a closer look yields some intriguing results. We already know that all cuts of diamond displayed high priced outliers at low carats. The above plot, however, shows us that only diamonds that had D or E color (the two best) had significant amounts of high outliers. This result lies closely with my intuition since I would assume better colors correspond to higher prices. However, a deeper exploration into how color and price are related could certainly disprove my theory. Also, since most diamonds were of higher quality color, the significant amount of outliers in the D and E colors could be due to sheer sample size.

Next, I considered the 'mpg' dataset and attempted to determine whether or not there was a significant difference between fuel economy between transmission types. Since the dataset is relatively small (only 234 observations) I decided to just group cars into two types: manual and auto.

```
mpg.manual <- subset(mpg, trans == "manual(m5)" | trans ==
"manual(m6)")

mpg.auto <- subset(mpg, trans != "manual(m5)" & trans !=
"manual(m6)")

p1 <- qplot(hwy, ..density.., data = mpg.manual, geom = "freqpoly") +
ggtitle("Manual Cars")
```

```
p2 <- qplot(hwy, ..density.., data = mpg.auto, geom = "freqpoly") +
ggtitle("Automatic Cars")

grid.arrange(p1, p2, nrow = 1)
```



Third Plot

Plotting with density on y-axis is extremely helpful here because the automatic cars subset had roughly twice as many observations. From the plots, it can quickly be ascertained that manual cars are mostly in the high 20's – low 30s highway miles per gallon while automatic cars have two large peaks, one at around 18 miles per gallon and another at 26 miles per gallon. However, we cannot be too sure of our result. Firstly, with such a small sample size it isn't clear whether or not we see this relationship strictly due to "chance". Alongside that, our findings might be attributed to the class of car associated with each type of transmission. If manual cars were mostly 'compact' and other small classes, my claim from the above plot would lose cogency since it is unclear whether the differences in mpg are actually attributed to the transmission. Some more exploration on which cars tend to be manual would be beneficial in this situation.

Conclusion

We were able to easily see that price and carat had a strong positive relationship in the 'diamonds' dataset. After some deeper analysis, I was also able to unearth that prices tended to level off at around 2-2.5 carats after exponentially growing before. Finally, after some investigation of why outliers tended to be overpriced, I found that it was mainly because the outliers tended to be of ideal cut and D color (the best cut and color). It would be worth looking at the relationship between price and color to determine whether my methodology and conclusion has value.

In the 'mpg' dataset, I was able to see that manual cars tended to be more fuel efficient than automatic cars. However, the small number of observations limits the validity of my claim. On top of that, the higher fuel efficiency of manual cars might be attributed to the classes of cars which tend to be manual (among other factors) as opposed to the transmission itself. An analysis of transmission type vs. car class would certainly yield some interesting results and help determine the significance of my findings.