Diamonds in the rough
A data cleaning and exploration case study

Hadley Wickham

September 17, 2007

This paper presents a dataset containing the prices and other attributes of almost 54,000 diamonds. Simple graphics, scatterplots and histograms, and simple linear models are used to explore and remedy data quality problems, and to reveal interesting and unexpected relationships in the data. A combination of interactive graphics and simple linear models is used to investigate a 3-way interaction when predicting price based on carat, colour and clarity.

1 Introduction

Diamond prices are something everyone can relate to. Diamonds are an expensive, highly desirable, luxury good that students are familiar with (just ask them about "bling"!) We advance previous work (Chu, 1996) by collecting many more observations and using a more exploratory approach, combining both graphics and simple linear models. A large part of this exploration focusses on data quality, a particularly important issue for real-life data.

This dataset was used by the author in a statistical computing class at Iowa State in Spring 2007, with considerable success. Students had some basic familiarity with the data, were generally interested in it, and could find out more details using familiar resources (eg. wikipedia). A number of students discovered features in the data that I hadn't found, and constructed interesting explanations for the phenomena that they discovered.

2 The dataset

The prices and attributes of 53,940 round diamonds were collected by the author from `http://diamondse.info` on 6–7 February 2007. Table 1 lists the variables in the dataset: there is a mixture of continuous (price, carat, and five physical dimensions) and categorical (cut, clarity, and colour) variables. As the focus of this paper is on data cleaning and exploration, the data are provided exactly as extracted from the website, and no effort has been made to fix the errors in the data that we will soon discover.

Variable	Description
price	price in US dollars ($326–$18,823)
carat	weight of the diamond (0.2–5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
colour	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1, SI1, SI2, VS1, VS2, VVS1, VVS2, IF)
x	length in mm (0–10.74)
y	width in mm (0–58.9)
z	depth in mm (0–31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
table	width of top of diamond relative to widest point (43–95)

Table 1: Description of the 10 variables in the diamonds dataset. There were 53,940 observations.

3 Large data: motivation and challenges

In most statistics classes, students are presented with small, well-cleaned datasets and pre-specified analysis goals. This type of data is useful for teaching new statistical methodology, as it frees the student from having to worry about the quality of the data, but it does not resemble a real-life data analysis scenario. In real-life data is often larger, always messier, and typically the analysis goals are not so well defined.

The diamonds dataset is large, messy and has many interesting stories to uncover. However, the size of the data brings with it some challenges for specific methodologies. Below I discuss some particular considerations for the two tools used extensively in this case study: the scatterplot and histogram.

3.1 Scatterplots

For scatterplots, large data is a mixed blessing Unwin et al. (2006). Scatterplots are not efficient in the statistical sense: once the density of points increases past a certain threshold, important trends will be concealed, rather than revealed, due to overplotting. Figure 1 shows a scatterplot of carat vs. price for the diamonds data, and contains over 50,000 points. While we can see the general extent of the point cloud, it is impossible to accurately determine the relationship between carat and price: we can not tell if one point or one thousand points are plotted at a given location.

One way to solve this problem is to use somewhat transparent points. This technique is called alpha blending. Unfortunately, alpha blending is limited by the underlying graphics implementation. Most software uses 8 bit colour, which means that we can only uniquely distinguish up to 256 ($= 2^8$) points plotted at a single location. This limitation is seen on the right hand side of Figure 1: using the smallest alpha of 1/256 gives us more of a feel for the distribution of the points, but it is still not possible to see patterns in the densest parts.

There are striking differences between the two plots in Figure 1. The area of opaque plot which appeared most visually strikingly, the diamonds with high prices, is revealed to contain relatively little data. Alpha blending adds another dimension to the scatterplot: low alpha values

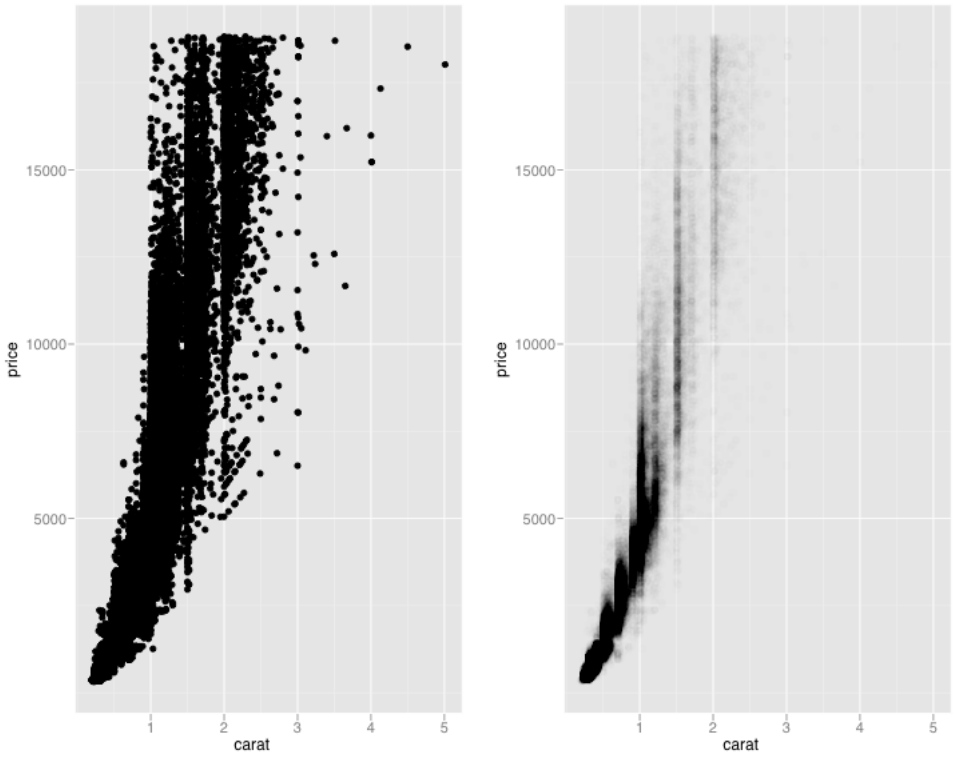

Figure 1: (Left) Scatterplot of carat vs. price, opaque points. You can see the extent of the points, but not their distribution. (Right) The same scatterplot but with alpha value of 0.39%, which means we need 256 points in the same location to have the same darkness as one point on the opaque plot. It is more revealing than the opaque plot, but it is still saturated in many places.

focus on trends, and high alpha values focus on outliers.

Zooming in on a small region of the data is helpful, as in Figure 2. Close inspection shows that carat appears to be discrete at a very fine level. This is because the precision, the smallest non-zero difference between ordered values (Wilhelm, 2005), of carat is low, 0.01 (a precision-range ratio of 481), compared to price, with a precision-range ratio of 18497. Another way to explain this is that, on average, each unique value of carat was recorded 197.5 times, while each unique value of price was only recorded 4.6 times. We also see a gap around $1500, which was not visible in the previous figure. This is a known data collection problem, but has been left in to add another interesting feature for students to discover. (In fact this finding was discovered by a student).

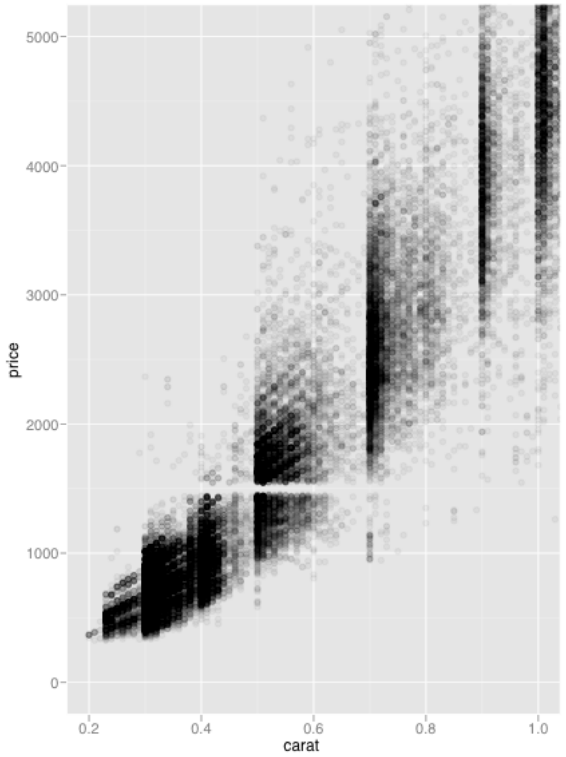

Figure 2: Scatterplot of carat vs. price "zoomed-in" to show smaller region. Alpha of 4% used.

Another approach to remedy the problem of overplotting is to supplement the scatterplot with information from statistical models. For example, if we are interested in predicting x from y, we might add a smoothed conditional mean. A common choice is to augment the plot with a loess smooth, but that doesn't help here: loess is $O(n^2)$ in memory Cleveland et al. (1992), and so can not deal with such a large data set. Other smooth models are still useful, e.g. cubic smoothing splines. Alternatively, if we are interested in inspecting the joint density of x and y, we can supplement the scatterplot with contours from a 2d kernel density estimate. These two approaches are illustrated in Figure 3.

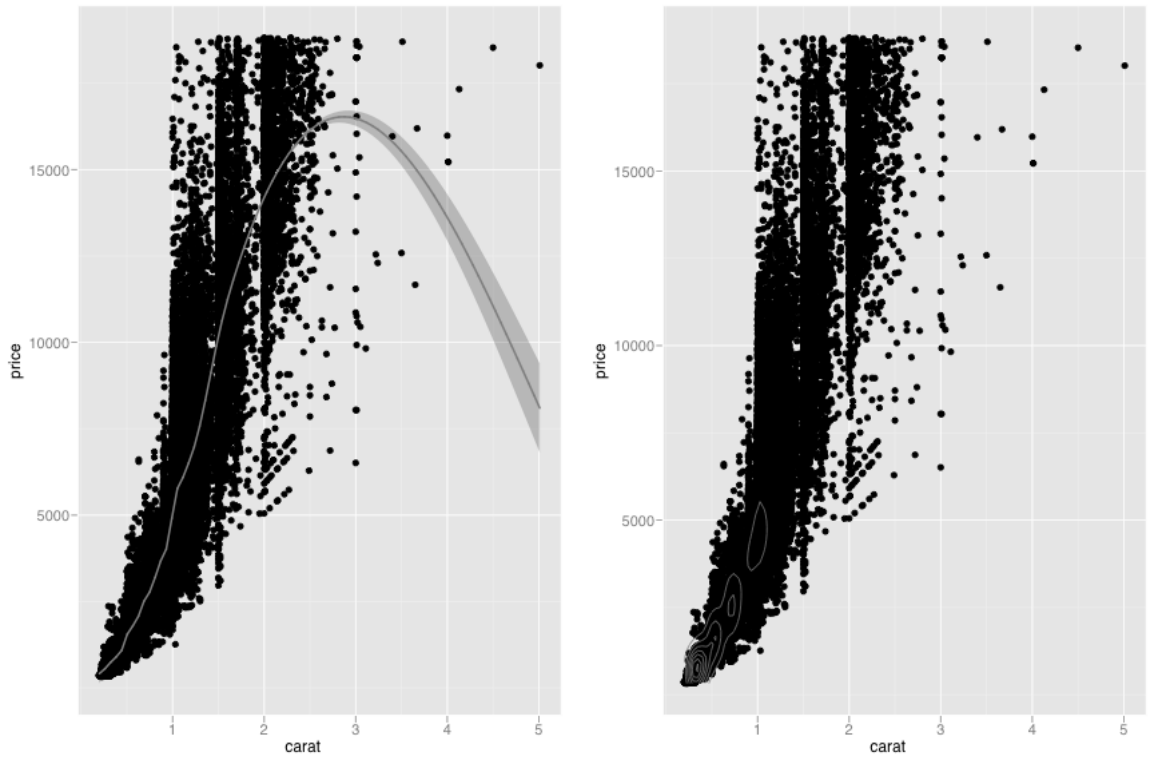

Figure 3: Supplementing the scatterplot of price vs. carat with (left) smoothed prediction of y from x, using a natural spline with 20 degrees of freedom, and (right) a 2d kernel density estimate. This additional model based information shows that overplotting conceals a lot of the pattern in this plot.

3.2 Histograms

Histograms are better at dealing with large amounts of data because they aggregate the data. However, we have an important choice to make when drawing a histogram: what size bins should we use? Usually there will not be one optimum bin size, but a number of different bin sizes that illustrate different stories in the data. As the amount of data increases, the smallest reasonable bin width will decrease, but we will still gain insight from looking at large bin sizes as well. This is similar to the choice of alpha value illustrating high-level (trend) vs low-level (detail) features in the data. This is demonstrated below for the carat variable.

4 Distribution of diamond size

Figure 4 shows a histogram with bin width of 1 carat. This reveals that most diamonds are between 0 and 1 carats, and about half as many again are 1–2 carats. There are few diamonds of 2–3 carats, and we can't see any diamonds for 3–4, 4–5, or 5–6 carats. The fact that the axis extends that far tells us that there are diamonds there, but the bar is too small to see. This is a common problem in most statistical software, apart from the software MANET (Hofmann, 1997) which ensures non-zero bars have a minimum height. If we go back to the data, we can see that there 34 diamonds in the 3–4 carat bin, 5 in the 4–5 bin and 1 in the 5–6 bin.

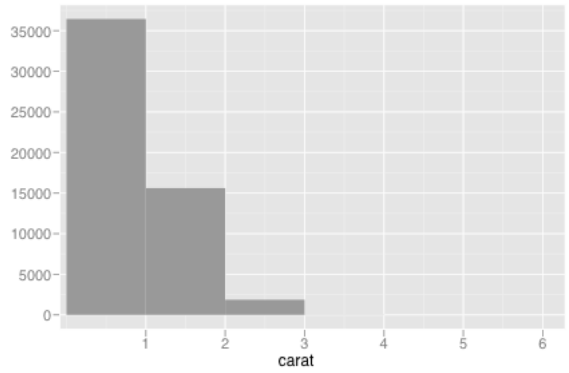

Figure 4: Histogram of carat, with bin width 1, illustrating crude trends in distribution of diamond weights. Note the range of x: while we can not see any bars greater than three carats, they do exist.

The next histograms, Figure 5, have bin widths of 0.1 and 0.01 carats. Note that we have zoomed the x range to (0, 2.5) so we can see what is going on more clearly. This discards 126 observations, 0.23% of the total data. At a bin width of 0.1 carats we can see another story starting to emerge. There is no longer a simple decay, but many bumps and spikes. The spikes 1, 1.5, 2.5, and 3 suggest that there are more diamonds at "nice" numbers. It may be tempting to dismiss these as simple random variation, but we will see that there is an interesting underlying pattern.

With bin width 0.01, the precision of the data, we see a really interesting pattern: it looks like there are popular diamond sizes and the frequency of diamonds larger than these sizes decreases

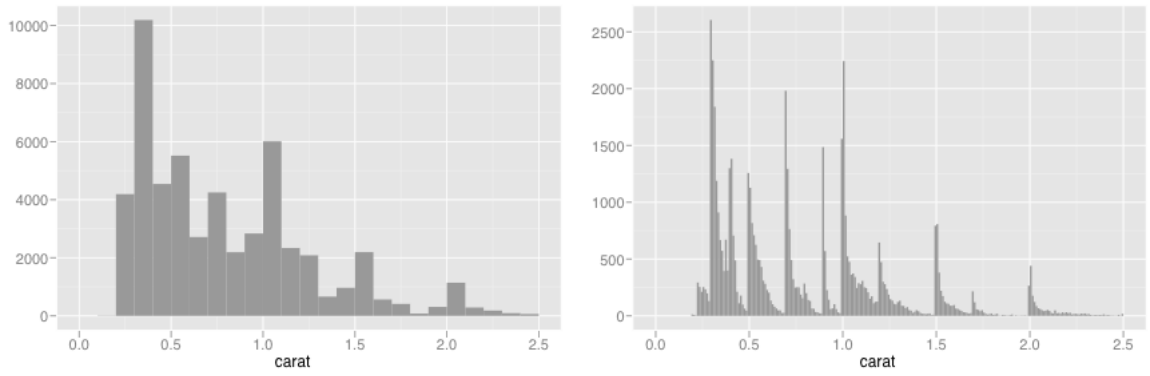

Figure 5: Histogram of carat with bin width 0.1 (left) and 0.01 (right), restricted to carats between 0 and 2.5. On the left, some interesting spikes and hollows start to emerge, which become even more evident on the right, with an interesting pattern of peaks followed by rapid decay.

rapidly. We see a number of peaks that weren't obvious with the coarser bin width, at 0.5, 0.7, 0.9, 1.2 and 1.7 carats. These numbers no longer seem so "nice", so perhaps there is something else going on other than a desire to own a nice round number of carats.

It's interesting to compare these bin widths with those selected by the classical methods. Sturges (1926) suggests a bin width of 0.5, while Freedman and Diaconis (1981), and Scott (1979) both suggest 0.05. Using either of these, as shown in Figure 6, would not have revealed all that the data had to tell us. To make the problem mathematically tractable these methods make strong assumptions about the distribution that are often not true in practice: beware the belief that there is a single optimum bin width.

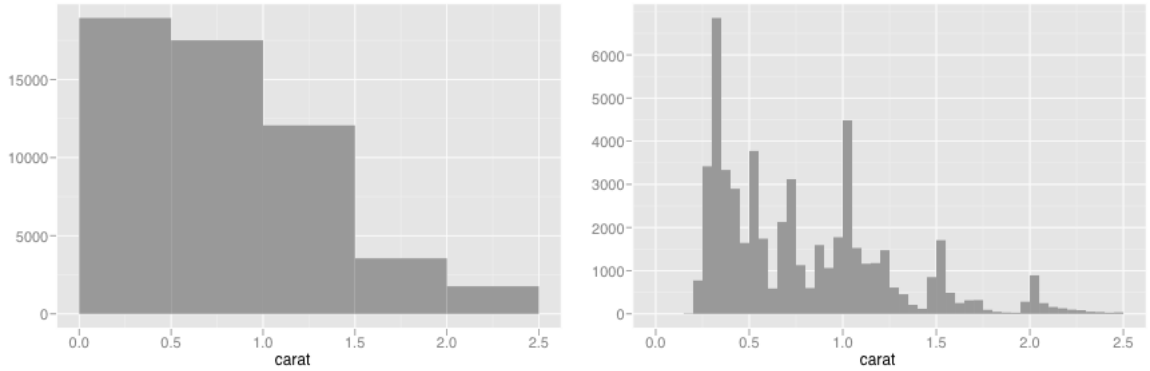

Figure 6: Histogram of with binwidth 0.5 (left) and 0.05 (right), derived from theoretical "optima".

5 Data cleaning

The process of data cleaning is important to teach to students. In real life, data are messy, and usually contain errors. We have three options to deal with these errors: we can remove the entire observation, we can replace the value with a missing value, or we can try and correct the value. In this case study we will use all three methods.

Figure 7 shows pairwise scatterplots of the three diamond dimensions, x, y, and z. There are some clearly outlying points: diamonds that are 3cm long do not sell for less than $20,000! (See Table 1.) What might have caused these errors? One plausible explanation is that the decimal point was misplaced during data entry, and so we will divide any measurement over 30 by 10. There are also a number of zeros. Clearly a diamond must have strictly positive dimension, and so we will replace these zeros with missing values. The results of these modifications are shown in Figure 8. The range of the data is much smaller now, and that allows us to see some more problems. This is typical of the data cleaning process: resolving one set of problems reveals a new set. In the next section we'll attempt to deal with these odd observations by computing a new variable, asymmetry.

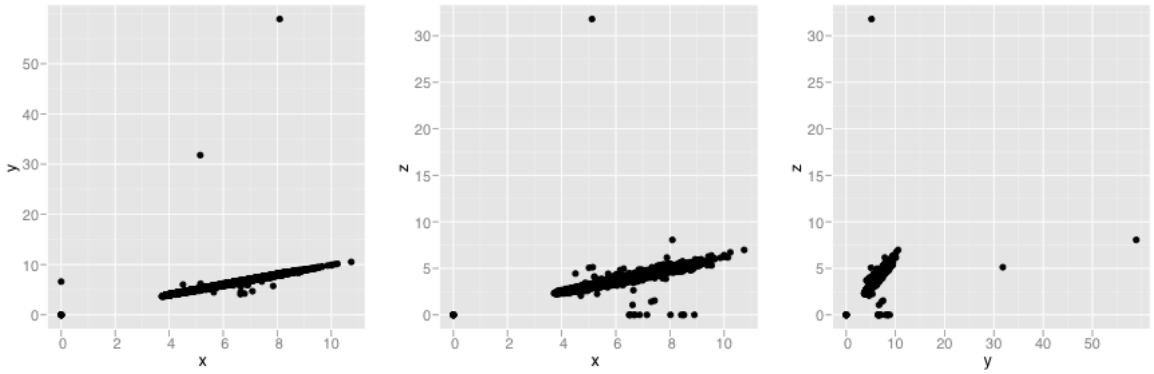

Figure 7: Scatterplot of diamond dimensions. From left to right: x vs. y, x vs. z and y vs. x. Overplotting is not a concern in these plots as we are looking for exceptional points, not the general trend.

6 Symmetry

First we remove the 20 diamonds missing one or more measurements, and then we calculate a variable that describes the asymmetry of a diamond: the standard deviation of the x, y, and z values, after rescaling so that each column has unit variance. This value will be large when the dimensions are different, i.e. far away from the diagonal, and small when the dimensions are similar. Two histograms of this new variable are displayed in Figure 9. It is interesting that there are few diamonds with very low asymmetry, and there are no diamonds with zero asymmetry. This is illustrated rather strikingly in Figure 10: there are no points along the diagonal $x = y$.

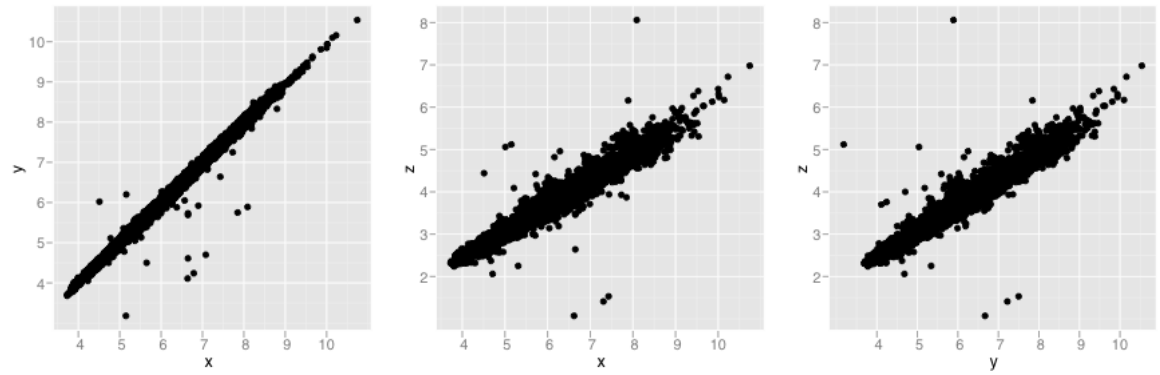

Figure 8: Scatterplot of diamond dimensions, after correction of clearly incorrect values. From left to right: x vs. y, x vs. z and y vs. x

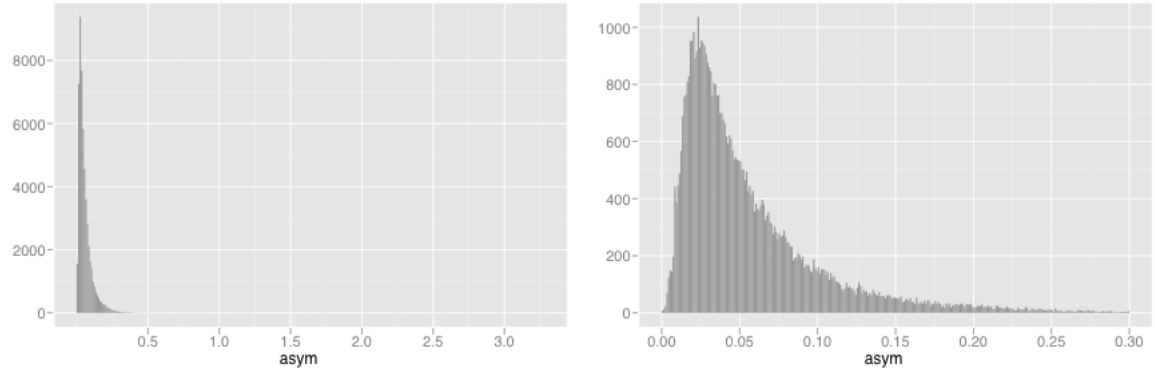

Figure 9: Histogram of asymmetry variable. (Left) bin width of 0.01, full range. (Right) bin width of 0.001, restricted to range (0,0.3). The distribution is unimodal and highly skewed.

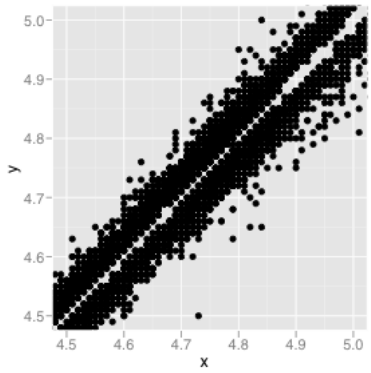

Figure 10: Scatterplot of x vs. y, zoomed in to show only sizes between 4.5 and 5 mm. There are no diamonds with equal x and y dimensions!

We now remove the 267 diamonds with asymmetry greater than 0.3. This is an arbitrary cut-off and it would be worthwhile spending more time exploring other possible cuts. Figure 11 shows the pairwise plots after this cleaning. The high correlation between the three variables is now visible, as is greater variation within the z dimension.

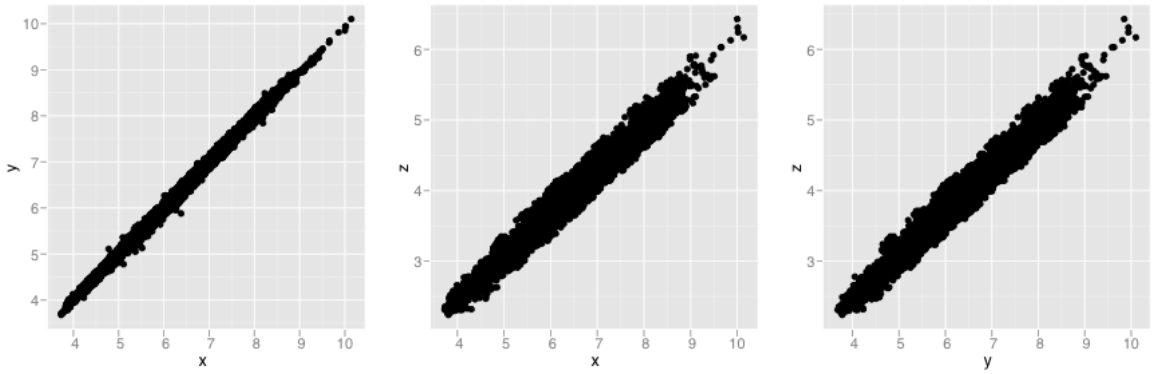

Figure 11: Scatterplots of x, y, and z after removing unusually asymmetric points. All three variables are highly correlated, with greater variation in the z dimension.

Figure 12 shows an interesting relationship between the cut of the diamond and asymmetry. It seems that better cuts have lower asymmetry, although the range of asymmetry is similar across all cuts. This would be a good opportunity introduce students to the idea of multivariate outliers: we might have different definitions of what an outlier is depending on the value of cut.

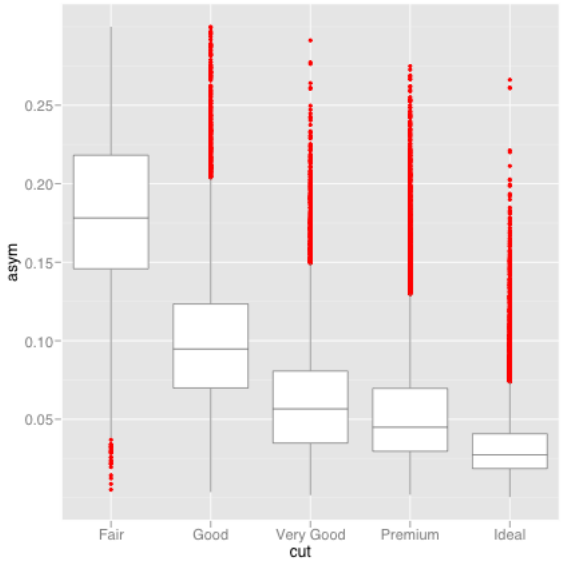

Figure 12: Boxplot illustrating distribution of asymmetry conditional on cut. Better cuts have more symmetrical diamonds.

There are many more ways we could clean up this data. Figures 13 and 14 suggest two possible approaches based on checking derived variables.

Figure 13 investigates the relationship between volume (x $\times$ y $\times$ z) and weight. While all diamonds are not exactly the same shape, they do have the same density, and we see a strong correlation between size and carat. However, it is difficult to detect subtle deviations from this line, so the second figure displays the percentage error in weight when predicting weight from volume using a linear regression. There are a number of diamonds with very high errors that should be investigated further. This is of practical interest: could we use this result to find diamonds that are actually larger than their records indicate? It might be a good way to find a bargain, in a similar way to searching for "plam" pilots on ebay. (If you don't know about this trick, people often misspell the name of the item they are selling. When this happens, it doesn't appear in the regular search, e.g. for palm pilots, and gets fewer bids.)

Figure 14 compares the percent depth supplied in the dataset with percent depth calculated from the diamond measurements (see Table 1). As in the previous case, the figure displays the raw relationship and the percentage error when predicting one from the other. Again there are a number of points with very high errors that should be explored in more detail.

This is hardly an exhaustive cleaning process, but we have already uncovered many interesting features of the data, including clear data entry errors, possible errors and potentially interesting relationships. It really is impossible to completely separate data cleaning from exploration: asking new questions of the data often reveals potential problems that we may not have been looking for.

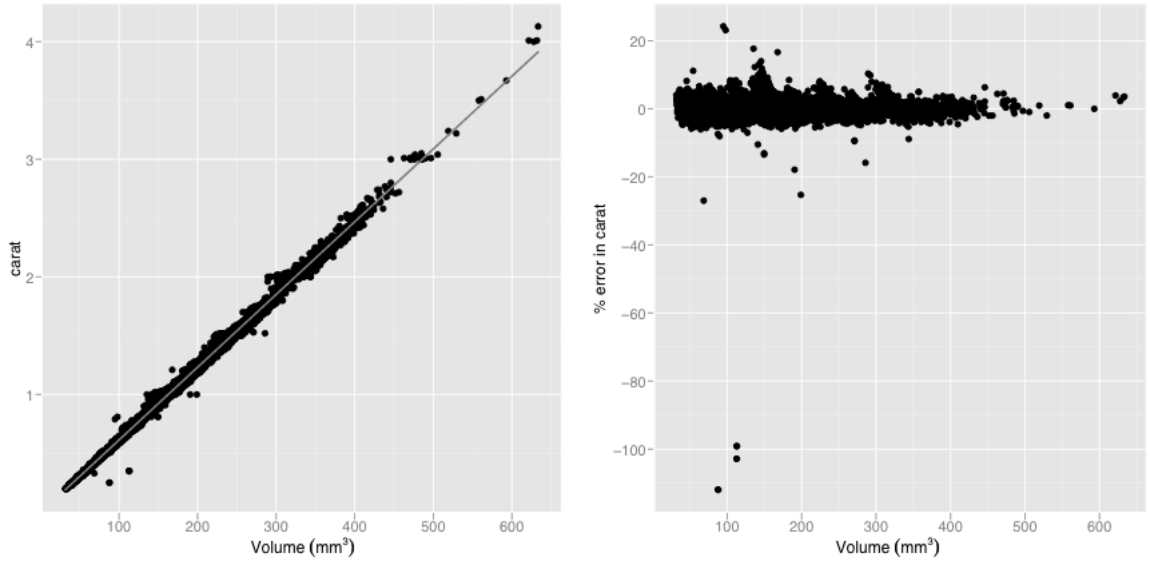

Figure 13: Relationship between volume and weight. (Left) Raw relationship and (Right) percentage error in predicting weight from volume.

7 Higher dimensional relationships

The process of data exploration is greatly aided by interactive and dynamic graphics, which allow us to pose questions and quickly answer them. Many of the findings presented so far were found

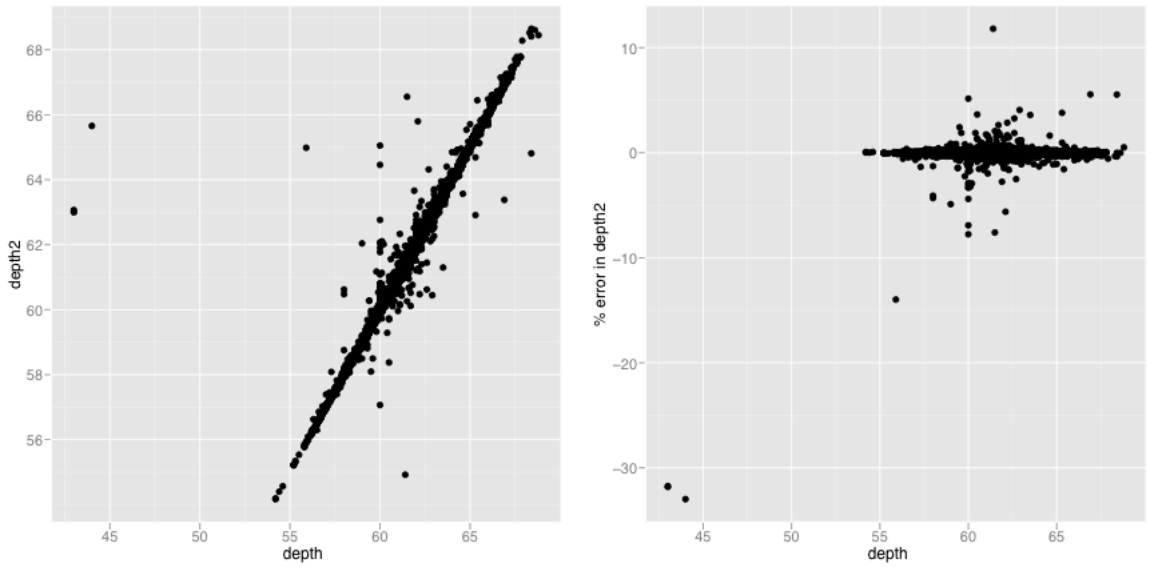

Figure 14: Total depth percentage from website (depth) vs. total depth percentage calculated from diamond dimensions (depth2). (Left) Raw relationship and (Right) percentage error in predicting depth2 from depth.

using interactive and dynamic graphics. For example, the idea of exploring diamond symmetry was inspired by interactively zooming in on a plot of x vs. y and noticing that there were no diamonds along the diagonal. Once something interesting has been found with interactive graphics, it's often useful to tweak and polish the output with static graphics and simple statistics to make the finding as easy as possible to see.

This final section of the paper presents some additional high-dimensional findings initially discovered using the software Mondrian (Theus, 2003). We will attempt to illustrate the process of using linked brushing (Becker and Cleveland, 1987; McDonald, 1982), where we link together multiple views of the data by selecting interesting observations in one view and seeing where they lie in another view. More detailed explanations of linked brushing can be found in Cook and Swayne (2007) and Unwin et al. (2006).

It is particularly interesting to see how the clarity and colour of a diamond affect its price. There is a strong relationship between price and carat, so we will take this into account. The easiest way to do this with interactive graphics is display a scatterplot of price vs. carat, while exploring the effect of carat or colour. In fact, a scatterplot of log(price) vs. log(carat) is even better as the relationship is close to linear. Figure 15 shows the initial set up in Mondrian, with a scatterplot of log(price) vs. log(carat) and a barchart of colour.

To investigate the price-carat relationship conditional on colour, we can brush all diamonds of colour D (the lowest quality) or colour J (the highest quality) by the selecting the appropriate bar. This is shown in Figure 16. It is hard to capture the typical use on paper. Using the keyboard, I repeatedly flicked from D to J while looking at price vs. carat; using animation to highlight any differences. Here we can see that the relationship between price and carat is affected by colour: diamonds of the same size are worth more if they have better colour. This shift seems roughly

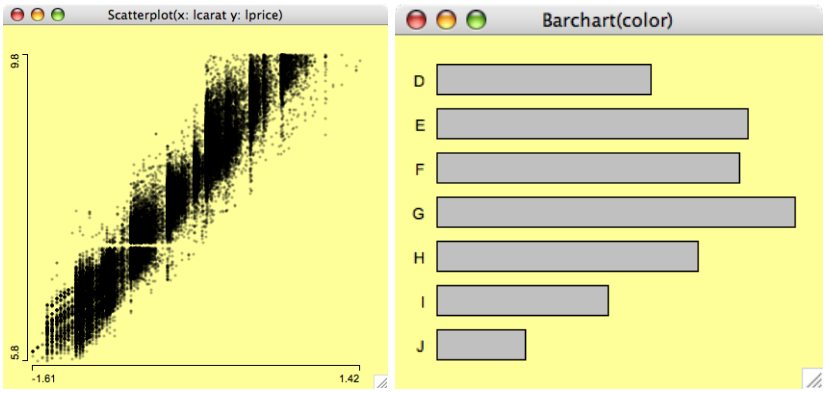

Figure 15: One view of the data in mondrian. (Left) Scatterplot of log(carat) vs. log(price). (Right) Bar chart of colour.

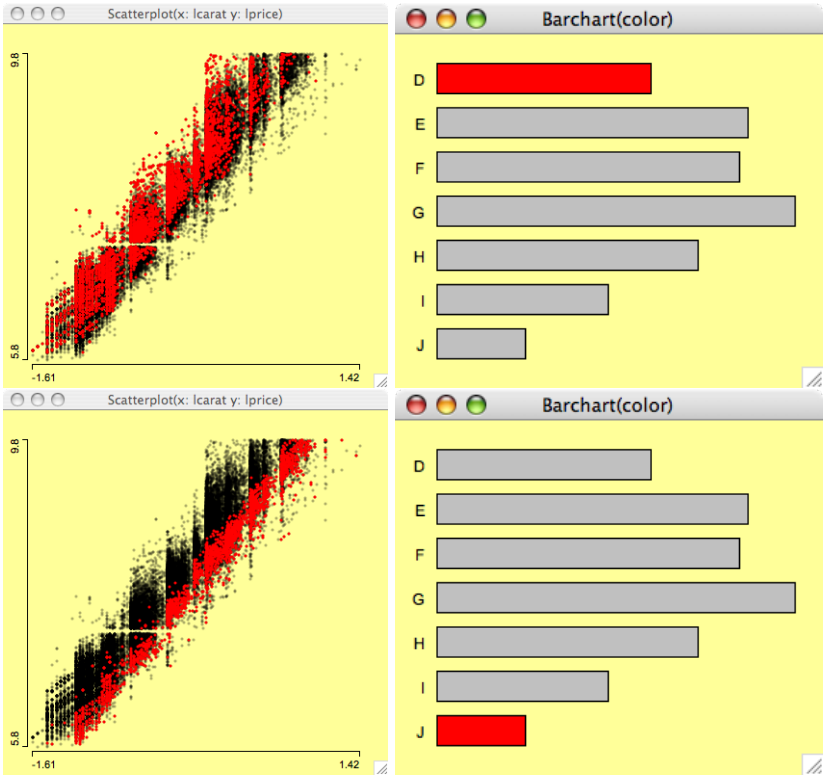

Figure 16: Brushing colour and observing change in scatterplot. Colour D has higher prices for diamonds of the same size with colour J.

constant across the range of carats—it colour modifies intercept, not slope.

We can investigate the relationship between clarity and price in a similar way. What about investigating the effect of both clarity and colour on price? Is the effect of clarity independent of that of colour or do they interact? Is there a three way interaction between colour, clarity and price? One possibility is to draw a fluctuation diagram of clarity and colour, Figure 17 and then brush that. A fluctuation diagram (Hofmann, 2000) is similar to a mosaic plot Hartigan and Kleiner (1981); Hofmann (2003), but displays the joint distribution of two categorical variables, not the conditional. Unfortunately, brushing this fluctuation diagram is not revealing as there are too many combinations to remember and compare.

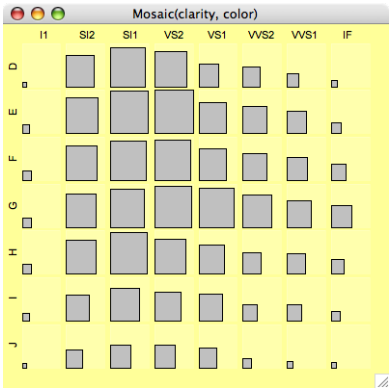

Figure 17: Fluctuation diagram of colour and clarity showing the relative numbers of diamonds for each combination. Most diamonds are of middling clarity and colour.

Instead we turn to a combination of modelling and static graphics. First, we remove the strong linear relationship between log(price) and log(carat) by predicting one from the other and then plotting the residuals. Second, we plot this relationship for every combination of clarity and colour. We will also supplement the scatterplot with a line of best fit, which will help guide our eyes. This large graphic is shown in Figure 18.

This graphic is complex (we are investigating a 3-way interaction), but rewards close inspection. We can compare colours conditional on clarity by looking down the columns, and vice versa by looking across the rows. The first thing that strikes me is clarity I1 (the worst) is quite different to the other clarities, with a negative slope. Remember, we have removed the strong price-carat relationship so this indicates that the price decreases relative to the overall mean, not that price is decreasing with size. We also see increasing slope and intercept as we travel to the top right. This suggests a three way interaction, which we can see clearly as variations in slope in Figure 19, where we have also removed both the clarity and colour main effects.

An alternative approach to these graphics is to go straight to the linear model and inspect the ANOVA table, as shown in Table 2. It isn't very useful for several reasons. First, it is difficult to assess the relative strength of the effects. All the effects are extremely significant, and while the F values provide an ordering, they are not immediately interpretable. Second, The ANOVA table only presents the fact that there are differences, not what those differences are: we need further summaries to elucidate the exact differences. Finally, there is gives no indication of model fit,

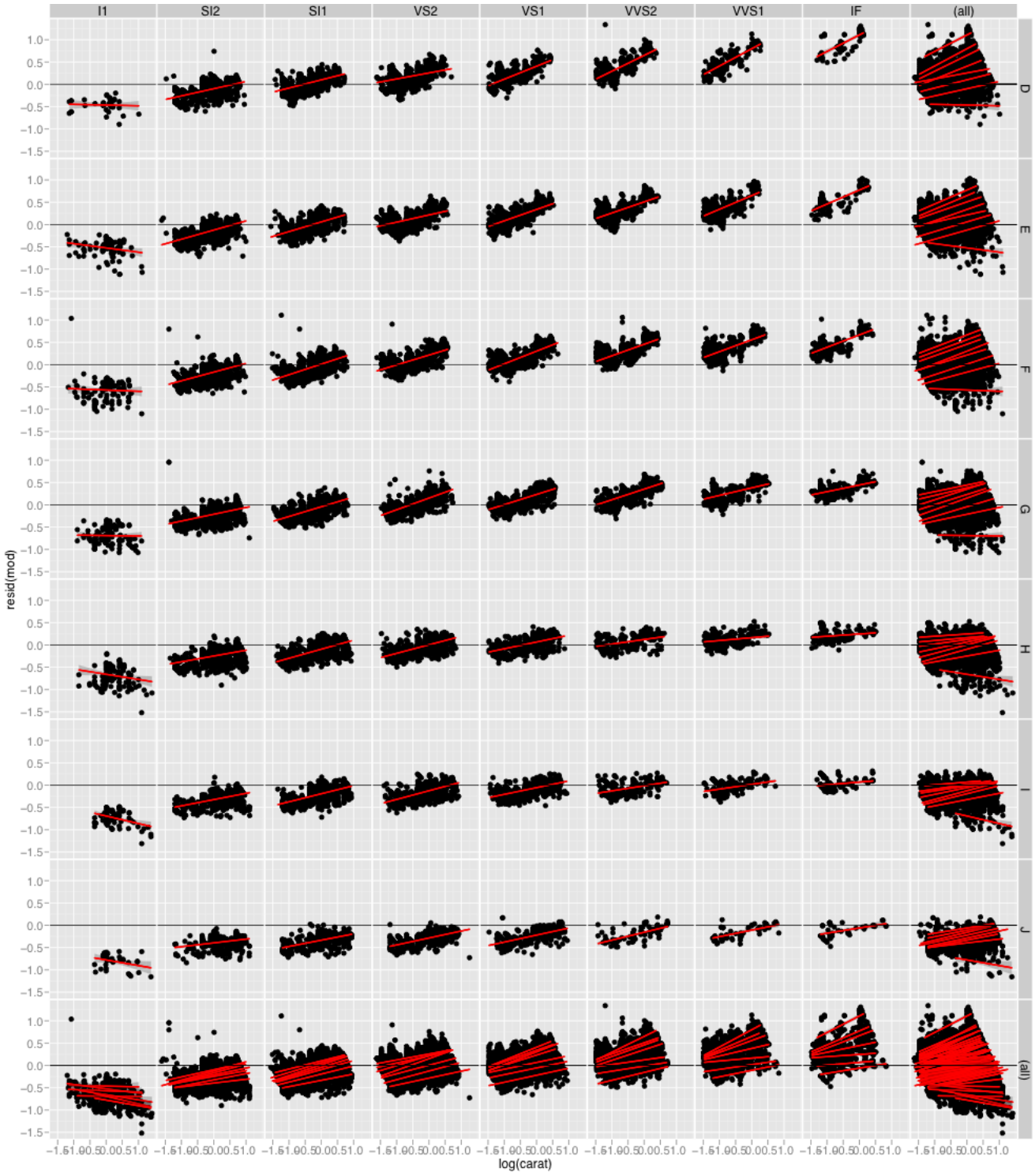

Figure 18: log(carat) vs. log(price) with strong linear relationship removed by linear regression, for each combination of clarity and colour. Line of best fit in red.

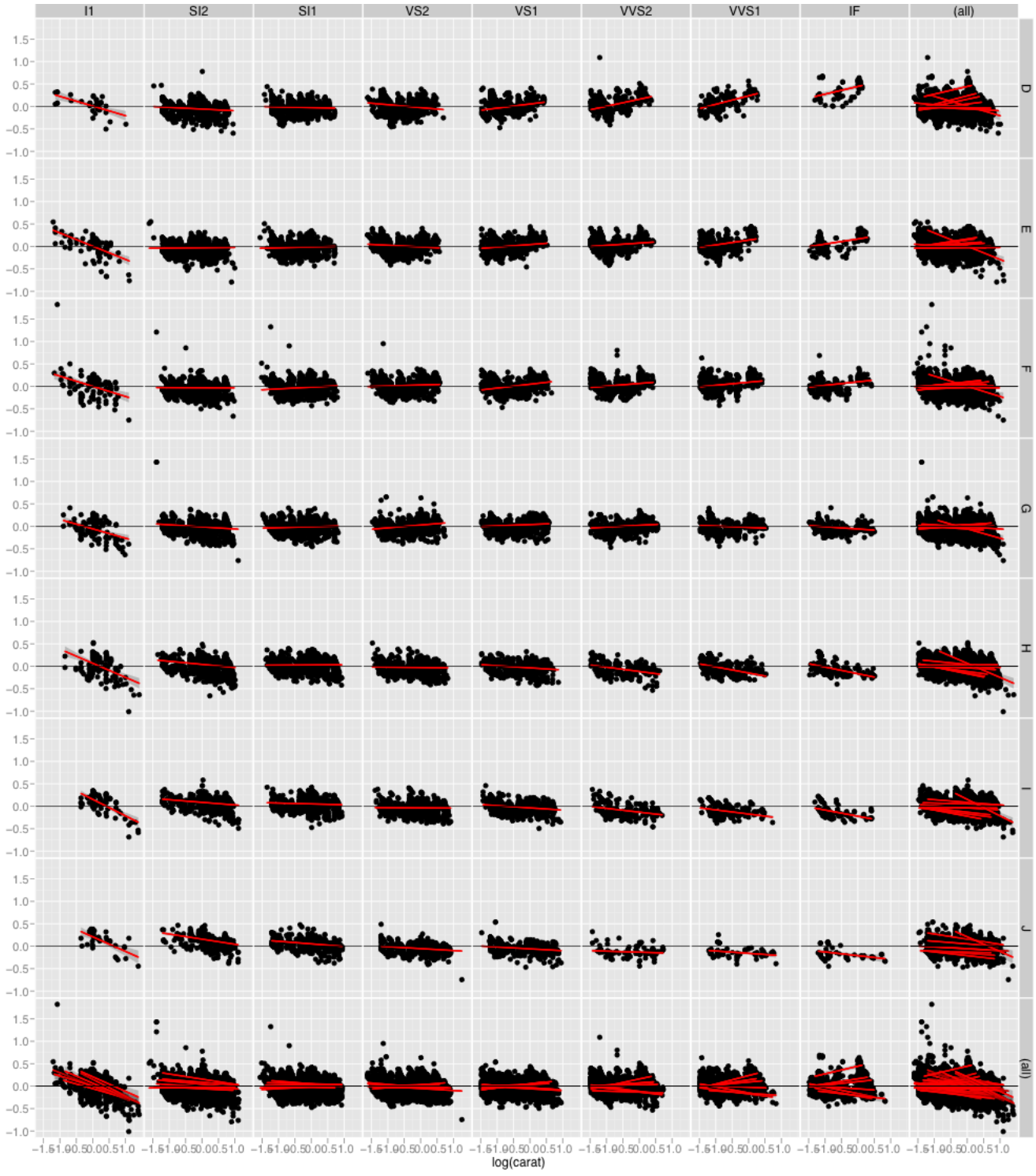

Figure 19: log(carat) vs. log(price) with strong main effects of price, clarity and colour removed by linear regression, for each combination of clarity and colour. Line of best fit in red.

and we would need to other methods to determine this.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
log(carat)	1	51581.43	51581.43	3130101.67	< 2.2e-16
clarity	7	1735.15	247.88	15041.96	< 2.2e-16
color	6	891.73	148.62	9018.78	< 2.2e-16
log(carat):clarity	7	9.87	1.41	85.55	< 2.2e-16
log(carat):color	6	2.57	0.43	25.99	< 2.2e-16
clarity:color	42	91.63	2.18	132.39	< 2.2e-16
log(carat):clarity:color	42	20.50	0.49	29.62	< 2.2e-16
Residuals	53541	882.31	0.02		

Table 2: ANOVA table for linear model predicting log(price) from log(carat), clarity and colour with all interactions. All terms are highly significant.

8 Conclusion

This paper presented a brief exploratory analysis of the prices and characteristics of nearly 54,000 diamonds. Many interesting patterns were discovered, including an unusual distribution of diamond weights, much incorrect measurement data, the relationship between symmetry and diamond cut, and how carat, colour and clarity predict the price of a diamond. Many of these patterns violate the typical assumptions of statistical analyses.

The analysis is technically achievable by students with limited statistical skills, focussing on a small range of graph types and simple linear models, with an emphasis on asking questions and exploring answers to generate still more questions. The combination of scatterplots, histograms and simple linear models is very powerful, and limited primarily by the curiosity of the practitioner.

There are many interesting relationships that remain to be found in this data, and we hope others will discover and report their findings.

References

R. A. Becker and W. S. Cleveland. Brushing Scatterplots. *Technometrics*, 29(2):127–142, 1987.

S. Chu. Diamond ring pricing using linear regression. *Journal of Statistics Education*, 4(3), 1996. URL `http://www.amstat.org/publications/jse/v4n3/datasets.chu.html`.

W. Cleveland, E. Grosse, and W. Shyu. Local regression models. In J. Chambers and T. Hastie, editors, *Statistical Models in S*, chapter 8. Wadsworth & Brooks/Cole, 1992.

D. Cook and D. F. Swayne. *Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and GGobi*. Springer, 2007.

D. Freedman and P. Diaconis. On the histogram as a density estimator: L2 theory. *Zeitschrift fuer Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57:453–476, 1981.

J. A. Hartigan and B. Kleiner. Mosaics for contingency tables. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 268–273, Fairfax Station, VA, 1981. Interface Foundation of North America, Inc.

H. Hofmann. Graphical stability of data analysing software. *Classification and Knowledge Organization. Proceedings of the 20th Annual Conference of the Gesellschaft für Klassifikation e.V.*, pages 36–43, 1997.

H. Hofmann. Exploring categorical data: Interactive mosaic plots. *Metrika*, 51(1):11–26, 2000.

H. Hofmann. Constructing and reading mosaicplots. *Computational Statistics and Data Analysis*, 43(4):565–580, 2003.

J. A. McDonald. *Interactive graphics for data analysis*. PhD thesis, Stanford University, 1982.

D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(605-610), 1979.

H. Sturges. The choice of a class-interval. *Journal of the American Statistical Association*, 21 (65-66), 1926.

M. Theus. Interactive data visualiating using Mondrian. *Journal of Statistical Software*, 7(11): 1–9, 2003.

A. Unwin, M. Theus, and H. Hofmann. *Graphics of Large Datasets*. Springer, 2006.

A. Wilhelm. Interactive statistical graphics: the paradigm of linked views. In *Handbook of statistics 24: Data mining and data visualisation*, 2005.