40 years of boxplots

Lisa Stryjewski

December 9, 2010

John Tukey has been credited with introducing the *box and whiskers* plot 1 in his book *Exploratory Data Analysis* [15]. Over time, the box and whiskers plot has been shortened to the *boxplot*, but despite the name simplification the boxplot is still the same five-number summary including:

- the *median*,
- two *hinges*, which are typically the first and third quartiles,
- the upper and lower *extremes*. These are simply the minimum and maximum of a batch of data unless outliers exist. In the case that outliers exist, the extremes are 1.5 times the *inter-quartile* range (3rd quartile - 1st quartile).
- two *whiskers* that connect the quartile to the extreme value on each side, and
- *outliers*. An outlier is a point that is greater than or less than 1.5 times the inter-quartile range.

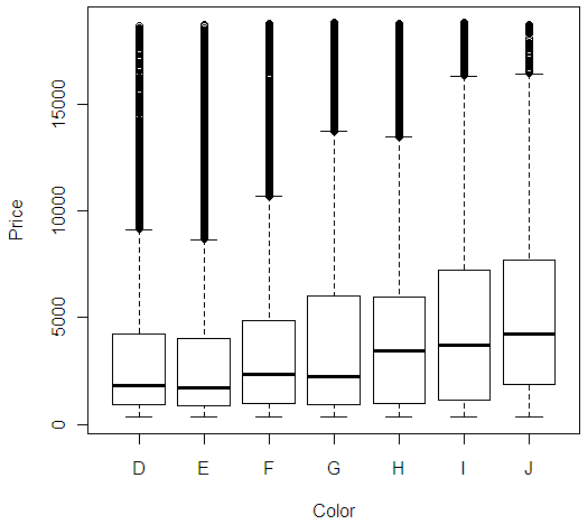

Figure 1: A boxplot of price grouped by color allows easy comparison of medians across multiple color groups.

Since boxplots made their debut in the 1970's, there have been many efforts to improve upon them. There are basically two ways this has been done: adding density curves and extending them to two dimensions. In section 1, we will explore the various methods for adding density to boxplots such as variable-width

plots, vaseplots, violin plots, box-percentile plots, and letter-value boxplots. In section 2 we will investigate two-dimensional boxplots known as rangefinder boxplots, relplots, quelplots, and bagplots.

1 Adding information about density

When boxplots were first introduced they were a very simple and popular method for displaying grouped data. But, like any new process or idea, people began to improve upon the original “box and whiskers” plot almost immediately. As you continue to read through this paper you will realize what a popular idea it was to somehow add information about density to the boxplot. You may wonder what justified this revision or why it seemed so logical to so many. The best way to understand why is to look at Tufte’s [14] version of the boxplot (called the ”midgap” plot) in which he removed the boxes. Figure 2 compares Tukey’s original boxplot to Tufte’s version.

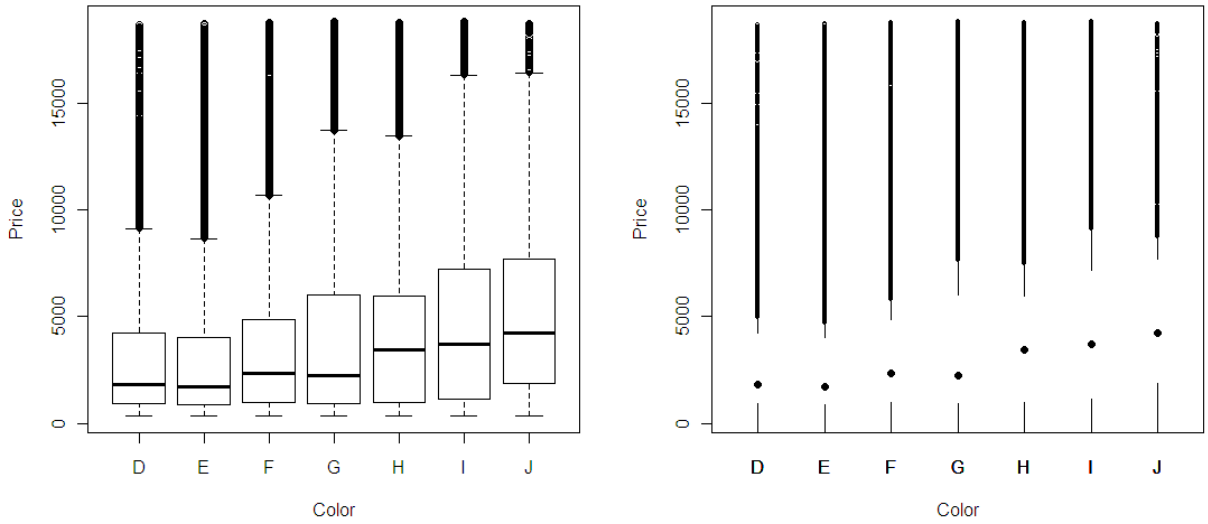

Figure 2: (Left) Tukey’s original boxplot. (Right) Tufte’s version of the boxplot that eliminates the box and thus removes the median and quartiles which boxplots are meant to highlight.

Removing the boxes was done in an attempt to save time and to increase the data-ink ratio. Tufte believes this revision can be made without loss of information. But, the most important information is actually lost in the background. Boxplots were created to emphasize the median, quartiles, and spread of the data. Tufte’s version gives us the impression that data is ”missing” where the data are most concentrated. According to the Gestalt principles, we will ”complete” the figure connecting the lines where there is white space. ”Completing” Tufte’s boxplot, however, will leave us with a straight line with a point for the median, eliminating the quartiles. In fact, Stock and Behrens [13] examined box, line, and Tufte’s midgap plots to judge the bias of estimates of whisker length for each type of plot. They found that estimates of whisker length for midgap plots are less accurate and more biased than estimates for box and line plots. This is evidence that the sides of the boxes are necessary for our perception of the data in a boxplot. Figure 2 compares Tukey’s boxplot with Tufte’s version.

1.1 Variable-Width boxplots

The first attempt to add density to the boxplot was made by Tukey et al. [9]just a few years after the original publication. When the data are not evenly distributed among the groups it is not obvious (from the plot only) where the overall median actually falls. In the Variable Width boxplot, width is used to represent the density, and this is believed to prevent misinterpretation of certain characteristics of the data, in particular the median. In figure 3, J has been subsetted so that the price is greater than $10,000, leaving J with only 440 observations. The remaining groups have between 5,422 and 11,292 observations. This is done only in an attempt to demonstrate how the overall population median can be misinterpreted. From the first plot in figure 3 it appears that the overall median price might be between 5,000 and10,000. However, the actual overall median price is 2,374. Looking at the variable width boxplot in the middle of figure four it is obvious that J has far fewer observations than the other groups. This plot is useful since it suggests that J will not weigh heavily in the computation of the overall median.

In the same paper, Tukey et al. [9] introduced the Notched boxplot, which adds yet another element to the original boxplot by displaying confidence intervals around the medians. Doing so allows one to visually determine whether or not the medians are significantly different between groups. This method has been criticized because we cannot easily compare across panels, according to Cleveland's hierarchy [4]. Figure 3 illustrates the difference between the regular boxplot, variable width boxplot, and the notched boxplot.

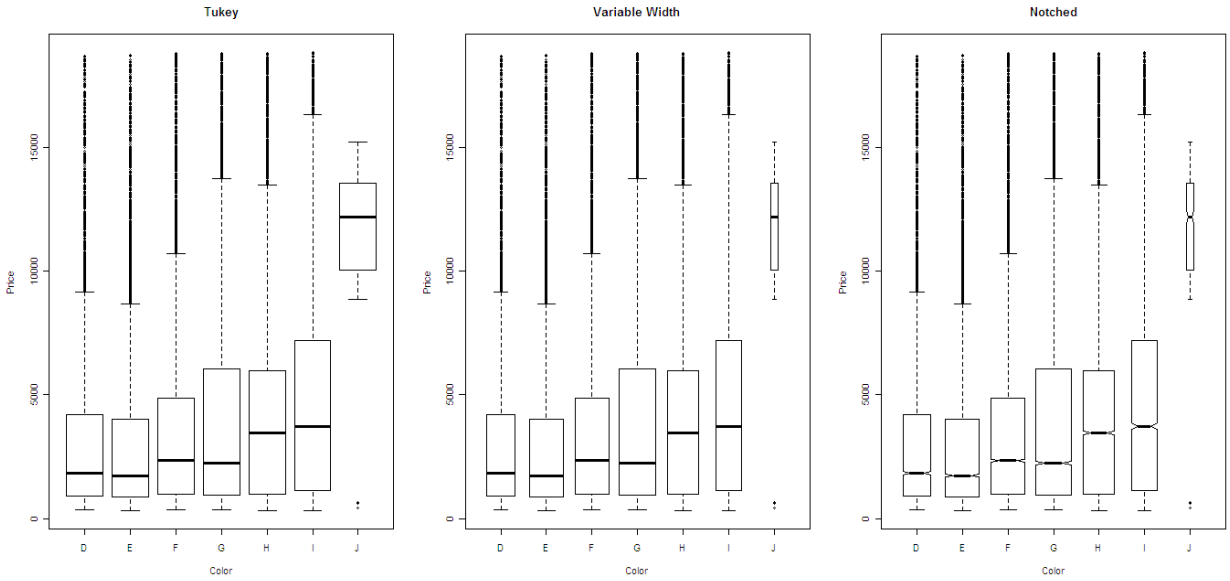

Figure 3: (Left) A regular boxplot: each box has the same width. (Middle) A Variable Width boxplot showing the differences in density among the variables. It is easy to see that there are fewer J diamonds than any other color. (Right) Overlapping of notches between D and E suggest no significant difference in median price. However, non-overlapping of notches between I and J does suggest a significant difference in median prices.

1.2 Vaseplots

Benjamini [2] introduced the *vaseplot*, which is a boxplot where the width of the box at each point is proportional to the estimated density. In a vaseplot, the density estimation is done on the middle part only (the box). The whiskers and outliers remain the same as in the original boxplot. (This differs from violin plots introduced later.) The density curves for vaseplots are generated using non-parametric density

estimation, allowing you to choose the appropriate amount of smoothing. But, since you may choose the "appropriate" amount of smoothing, vaseplots have been criticized for yielding different displays of the same data. Figure 4 represents this.

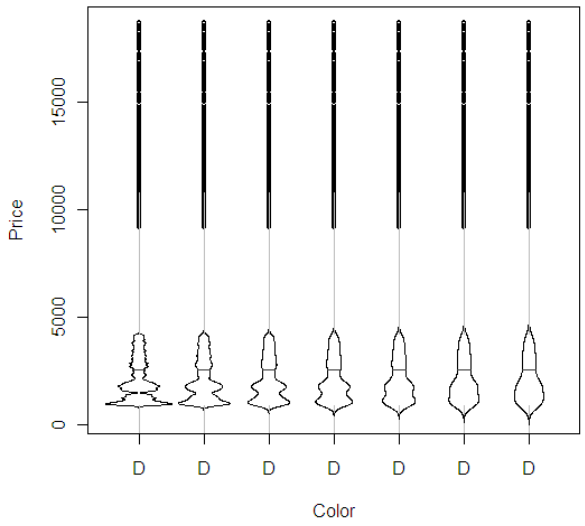

Figure 4: Vaseplots of price for diamonds of color D. The interval length *h* increases from left to right showing varying plots for the same data.

1.3 Violin plots

Violin plots [7] are similar to vaseplots, but there are two main differences between the two: 1) violin plots use *all* the data to plot a density curve. Vaseplots use just the central 50 % of the data, so a vaseplot still has whiskers and outliers. 2) Violin plots use *density trace* estimation [3] whereas the vaseplot uses non-parametric density estimation.

The violin plot allows you to determine the appropriate amount of smoothing, of the curve, by selecting a small or large interval length, *h*. The location density $d(x|h)$ at a point *x* is defined as the fraction of the data values that fall in an interval of length *h* centered about *x*,

$$d(x|h) = \frac{\sum_{i=1}^{n} \delta_i}{nh}$$

δ_i is one when the *i*th data point is in the interval $[x - h/2, x + h/2]$ and zero otherwise. You can see that you can control the "smoothness" of the curve by selecting h appropriately. Small values of *h* yield a very "wiggly" curve, whereas large values of *h* produce very smooth curves. *h* is a taken to be a percentage of the data range. The authors recommend using 15% of the data range, or $h = 0.15$, but this yields a very smooth curve as you can see in figure 5. You can select a smaller value of *h* to prevent over-smoothing, but there is not a single, sure-fire method for choosing *h* that will produce optimal visual results. The amount of smoothing to use is at the discretion of the analyst.

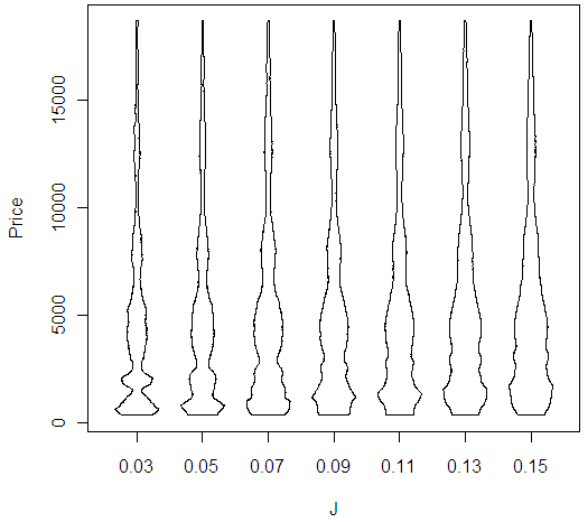

Figure 5: Violin plots of color vs. price of J colored diamonds. The x-axis indicates *h*, the smoothing parameter which is a percentage of the data range. In this plot *h* is increasing from left to right showing varying plots of the same data.

1.4 Box-percentile plots

The box-percentile plot 6, introduced by Esty and Banfield [5], differs from other density boxplots in that it does not use kernel density estimation to plot the density curve. In the box-percentile plot the observed values y_i are ordered from lowest-highest, and then each *y*-value is plotted at a distinct point. Let *w* be the desired maximum width of the box. If y_k less than or equal to the median then it is plotted at height y_k and distance $kw/(n+1)$. If y_k greater than the median then it is plotted at height $(n+1-k)w/(n+1)$. The box-percentile plot does not over or under smooth the density curvel, since each *y*-value is plotted at a distinct point. This eliminates the problem seen with both the vaseplot and the violin plot.

1.5 Letter-Value boxplots

The letter-value boxplot [8] was specifically created to accommodate large datasets. For moderate sized datasets (n $<$ 1,000), estimates about the behavior in the tails are not reliable. Large datasets (n = 10,000 100,000) yield much more reliable estimates about the behavior beyond the quartiles. The simple boxplot is not a good method for large datasets, since there is often too much over-plotting in the outlier region. To overcome this problem, Hofmann, Kafadar, and Wickham (2006) introduced the Letter-Value boxplot. To construct one let $X_{(1),\ldots,X_{(n)}}$ denote the order statistics from a sample of size *n*. The letter values are the order statistics having specific depths (M = median, F = fourths, E = eighths, D = sixteeths, C = thirty-seconds, ...), and are defined as $d_i = (1+d_{i-1})/2$. The i^{th} lower and upper letter values are defined as $L_i = X_{(d_i)}$ and $U_i = X_{(n-d_i+1)}$. The median is plotted as a horizontal line, and the inner-most box is plotted at the first and third quartiles, as in the original boxplot. A smaller box is drawn at the upper and lower eigths, and an even smaller box is drawn at the upper and lower sixteenths. This continues until a stopping point is reached which depends on the sample size.

Figure 7 is a letter-value boxplot of color vs. price. When compared with the original boxplot in figure **??** the letter-value boxplot conveys the data past the quartiles in a more meaningful way.

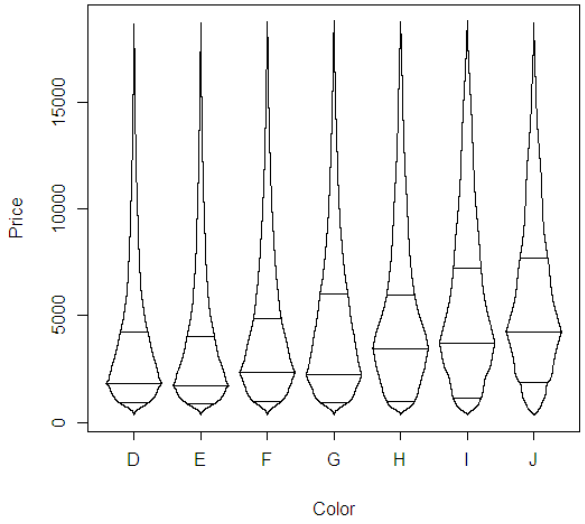

Figure 6: Box-percentile plots of color vs. price.

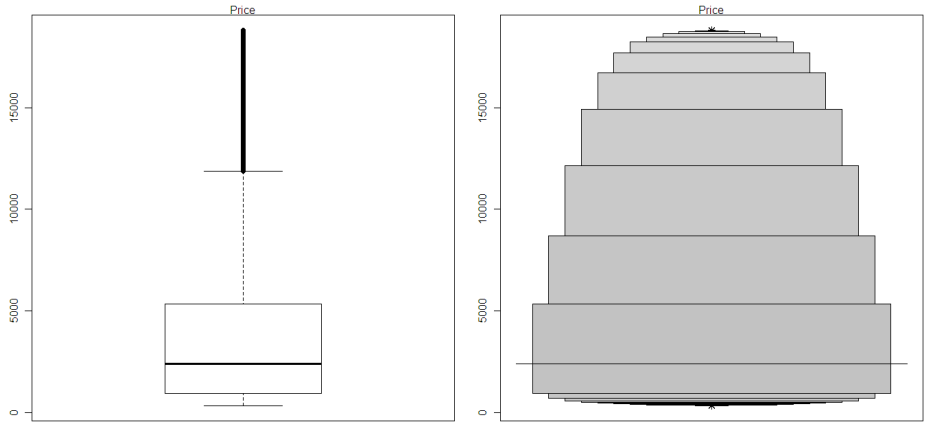

Figure 7: Letter-value boxplots of color vs. price display the data past the quartiles in a more meaningful manner than does the original boxplot.

1.6 Summary

Density boxplots contain more information than regular boxplots, but are they necessarily better? The answer depends on the characteristics of your data. Here are some things to keep in mind when choosing the appropriate boxplot:

- The variable width boxplot should be used when the groups are not evenly distributed. This will not show you the *shape* of the density, however. The variable width boxplot is the simplest of the density-type boxplots.
- The vaseplot can basically be used whenever the violin plot is used. If outliers exist, however, the vaseplot is a better choice since it will not stretch out the density curve to meet one outlying point. Both the vaseplot and violin plot can be over or under smoothed leading to different potrayals of the same information.
- The box-percentile plot can be used to prevent the smoothing problem seen with both the vaseplot and violin plot.
- Letter-value boxplots should be used when the dataset is large ($n > 10,000$), because they better convey information about the data beyond the quartiles.

Knowing when to use the appropriate type of boxplot can ensure that your information is conveyed as accurately as possible.

2 Two-Dimensional boxplots

In addition to density boxplots, bivariate boxplots are an extension of the original. A bivariate boxplot is typically superimposed over a scatterplot of two variables allowing you to see the location and spread of the data.

2.1 Rangefinder boxplots

Becketti and Gould [1] introduced the rangefinder plot, a bivariate extension of the boxplot that uses scatterplots and scatterplot matrices. First, a scatterplot of x vs. y is plotted, and then six lines are superimposed on the scatterplot. Two of the lines form a cross in the center of the rangefinder plot at the cross median values. Two vertical lines are drawn at the interquartile range of the variable on the x axis; these are drawn where the whiskers would end for the y variable. And, two horizontal lines are drawn at the interquartile range of the variable on the y axis; these are drawn where the whiskers would end for the x variable. The rangefinder boxplot contains all the information of the original boxplot for both variables on one plot. Figure 8 is a rangefinder boxplot of displacement vs. city miles per gallon. The dataset used in figure 8 and throughout section two is a subset of the EPA's fuel economy data and can be found in the R package ggplot2 (Wickham, 2009).

2.2 Relplots and Quelplots

Goldberg and Iglewicz [6] introduced two bivariate versions of the boxplot: the relplot and the quel plot 9. Unlike the rangefinder boxplot (1987), the relplot cannot be drawn by hand due to its computationally intense nature. The rangefinder boxplot uses two simultaneous univariate estimators to draw a rectangular shape around a scatterplot of the data. Golberg and Iglewicz claim, however, that the rangefinder method fails to capture the relationship between the two variables. Instead, they focus on just one estimator, a modified biweight M estimator, and model the data based on the bivariate Gaussian distribution. The confidence

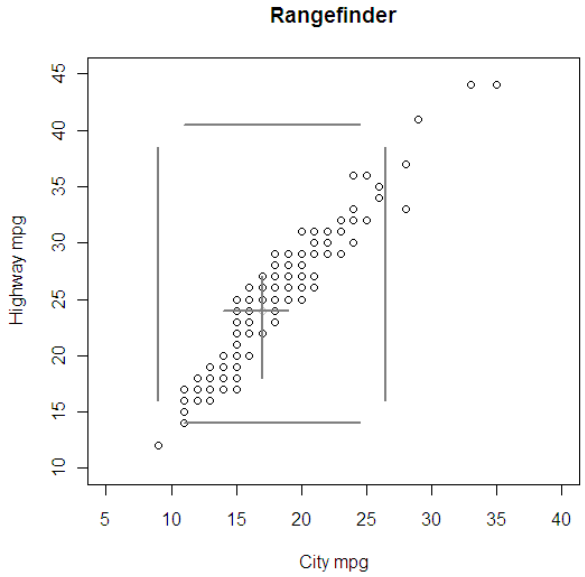

Figure 8: A rangefinder boxplot of city mpg vs. highway mpg.

limits for bivariate Gaussian data are confidence ellipses, which will be used in the two-dimensional boxplot. Four estimators are needed to draw the elliptical fence and hinge: location, scale (standard deviations of x and y), correlation, and a constant D that regulates the distance from the fence to the hinge. The hinge is centered on the median of the standardized errors, E_m, which are the Mahalanobis distances of each point from the center (T_x^*, T_y^*); T_x^* and T_y^* are the means of x and y respectively. The elliptical hinge and fence are both drawn using an algorithm involving the parametric equation of an ellipse.

Ellipses assume symmetric data, so to avoid this restrictive assumption, generalized ellipses, or quels can be used. A quel is made up of four separate quarter ellipses designed taking into account the proportion of the total standard deviation due to residuals in the positive direction of both the major and minor axes; doing so adds two degrees of asymmetry. Two quels are plotted; an inner quel (hinge) and an outer quel (fence). The inner quel is comparable to the box, and the outer quel is comparable to the whiskers in the original boxplot. Outliers lie outside of the fence.

2.3 Bagplots

Rousseeuw, Ruts, and Tukey [12] introduced another bivariate extension of the boxplot known as the *bagplot*. The bagplot consists of a scatterplot of the data, a bag that encloses 50% of the data points, and a fence that separates inliers from outliers. Unlike relplots and quelplots, the bag and fence of the bagplot are both polygons, not ellipses or quels. The polygons are constructed by the algorithm BAGPLOT which uses three algorithms from previous work; one, called LDEPTH computes the location depth of an arbitrary point [10], another, ISODEPTH, computes the depth median [10], and the third, HALFMED, constructs the vertices of a depth contour [11]. Rousseeuw and Ruts believe the bagplot is superior to the relplot and quelplot because the bagplot does not assume an elliptical shape and is thus model-free. 9

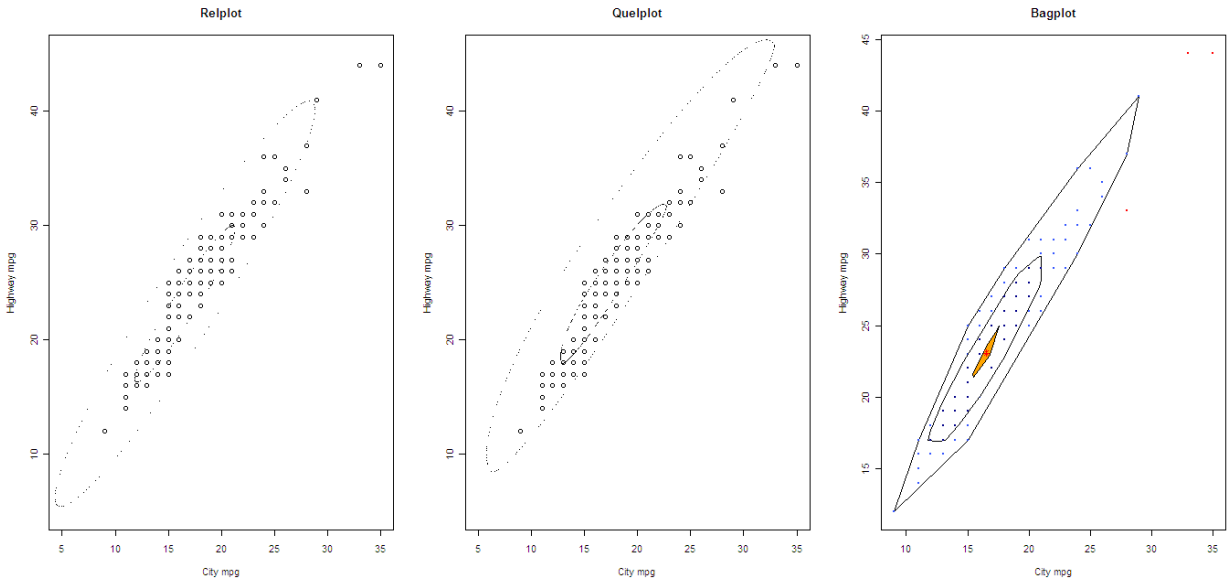

Figure 9: (Left) A relplot: The hinge and fence are symmetrical confidence ellipses. (Middle) A quelplot: Both the hinge and fence are made up of four quarter-ellipses. (Right) A bagplot of city mpg vs. hwy mpg made up of two polygons.

3 Summary

Over the past forty years boxplots have been adjusted to accomodate our needs, especially as computers became more widely available. Computers allowed analysts to create more complicated plots that require cumbersome coding, such as the quleplot. Just because we have the computing power to create complicated plots, however, does not mean that we should underestimate the value of the original boxplot. Boxplots were originally meant to summarize a batch of data in a simple fashion, and adding too much information can detract from that meaning. An original Tukey boxplot can serve as a starting point in the analysis of grouped data. Then, if you are further interested in the distribution among groups you can use a density boxplot, but doing so does not come without cost. The original boxplot is so simple to understand because we use two elementary tasks (Cleveland, 1984) to extract information from them: 1.) judging position along a common scale, and 2.) judging length. Both of these elementary tasks are at the top of the Cleveland's perceptual hierarchy (Cleveland, 1984). The lengths that we are judging are from the median to the quartiles, from the quartiles to the whiskers, from the whiskers to the outliers, and each combination thereof. The density boxplots are slightly more conceptually complicated since they require us to judge the area that represents the density. Judging area is near the bottom of Cleveland's perceptual hierarchy (1984) making it harder to understand the information in the plot accurately.

Bivariate boxplots allow you to see the location and spread of two variables, but they are not all created equally. All of the bivariate boxplots mentioned in section 2 contain the same information, but they are perceived differently. The rangefinder boxplot is the easiest to construct and the easiest to perceive. The elementary task is judging the length of the vertical and horizontal lines, or the range of the x and y variables. The relplot, quelplot, and bagplot all force you to judge the area enclosed by the fence and hinge making them perceptually more difficult.

Both density boxplots and bivariate boxplots are two very popular extensions of the original, and despite their weaknesses they both have some great advantages. The next forty years will likely bring even more modification to the boxplot, since they have been so fundamental to statistical graphics in the past.

References

[1] S. Becketti and W. Gould. Rangefinder box plots: A note. *The American Statistician*, 41:149, 1987.

[2] Y. Benjamini. Opening the box of a boxplot. *The American Statistician*, 42(4):pp. 257–262, 1988.

[3] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey. *Graphical Methods for Data Analysis*. CA: Wadsworth, 1983.

[4] W. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79:531–554, 1984.

[5] W. Esty and J. D. Banfield. The box-percentile plot. *Journal of Statistical Software*, 8:1–14, 2003.

[6] K. M. Goldberg and B. Iglewicz. Bivariate extensions of the boxplot. *Technometrics*, 34(3):pp. 307–320, 1992.

[7] J. L. Hintze and R. D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):pp. 181–184, 1998.

[8] H. Hofmann, K. Kafadar, and H. Wickham. Letter value boxplots. 2007.

[9] T. J. W. McGill, R. and W. Larsen. Variations of box plots. *The American Statistician*, 32:12–16, 1978.

[10] P. J. Rousseeuw and I. Ruts. Bivariate location depth. *Applied Statistics*, 45:516–526, 1996.

[11] P. J. Rousseeuw and I. Ruts. Constructing the bivariate tukey median. *Statistica Sinica*, 8:827–839, 1998.

[12] R. I. Rousseeuw, P. J. and J. W. Tukey. The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):pp. 382–387, 1999.

[13] W. A. Stock and J. T. Behrens. Box, line, and midgap plots: Effects of display characteristics on the accuracy and bias of estimates of whisker length. *Journal of Educational Statistics*, 16(1):pp. 1–20, 1991.

[14] E. Tufte. *The Visual Display of Quantitative Information*. Chesire: Graphics Press, 1983.

[15] J. W. Tukey. *Exploratory Data Analysis*. Reading, Mass: Addison-Wesley Publishing Co., 1977.