Methods for Presenting Statistical Information: The Box Plot

Kristin Potter

University of Utah
School of Computing
Salt Lake City, UT
kpotter@cs.utah.edu

Abstract: The display of statistical information is ubiquitous in all fields of visualization. Whether aided by graphs, tables, plots, or integrated into the visualizations themselves, understanding the best way to convey statistical information is important. Highlighting the box plot, a survey of traditional methods for expressing specific statistical characteristics of data is presented. Reviewing techniques for the expression of statistical measures will be increasingly important as data quality, confidence and uncertainty are becoming influential characteristics to integrate into visualizations.

1 Introduction

Understanding datasets is essential to the scientific process. However, discerning the significance of data by looking only at their values is a formidable task. Descriptive statistics are a quick and concise way to extract the important characteristics of a dataset by summarizing the distribution through a small set of parameters. Typically, median, mode, mean, variance, and quantiles are used for this purpose. The main goal of descriptive statistics is to describe quickly the characteristics of the underlying distribution of a dataset through a simplified set of values. Often these parameters provide insights into the data that would otherwise be hidden. In addition, these data summaries facilitate the comparison of multiple datasets.

Methods for visually presenting summary statistics include tables, charts, and graphical plots. Graphical plots are interesting in that they pictorially convey a large amount of information in a concise way that allows for quick interpretation and understanding of the data. There are many graphical ways to present descriptive statistics, so covering all of those methods here would be impractical. This survey will focus on one of the most common techniques for summarizing data, the box plot. In addition to various ways to construct the standard box plot, modifications which increase the amount of information presented in the plot will be discussed.

2 The Box Plot

The box plot has become the standard technique for presenting the *5-number summary* which consists of the minimum and maximum range values, the upper and lower quartiles, and the median. This collection of values is a quick way to summarize the distribution of a dataset. In addition, this reduced representation afforded by the 5-number summary provides a more straightforward way to compare datasets, since only these characteristic values need to be analyzed.

The typical construction of the box plot, which can be seen in Figure 1a, partitions a data distribution into quartiles, that is, four subsets with equal size. A box is used to indicate the positions of the upper and lower quartiles; the interior of this box indicates the *innerquartile range*, which is the area between the upper and lower quartiles and consists of 50% of the distribution. Lines (sometimes referred to as whiskers) are extended to the extrema of the distribution, either minimum and maximum values in the dataset, or to a multiple, such as 1.5, of the innerquartile range [FHI89] to remove extreme outliers. Often, outliers are represented individually by symbols; this type of plot is sometimes referred to as a schematic plot [Tuk77]. Finally, the box is intersected by a crossbar drawn at the median of the dataset. The width and fill of the box, the indication of outliers, and the extent of the range-line are all arbitrary choices depending on how the plot is to be used and the data it is representing.
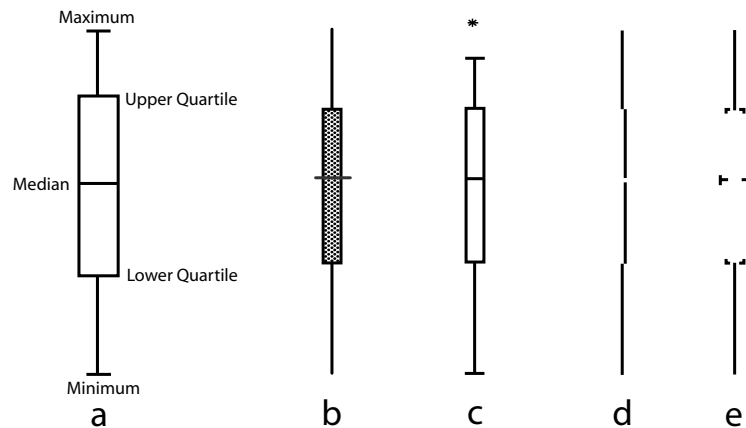

Figure 1: a) The anatomy of a box plot. b-e) Variations of the Box Plot. b) The range-bar chart. c) The box plot. d) The quartile plot. e) The abbreviated box plot.

2.1 Origins

The origins of the box plot can be traced to the range-bar chart. Haemer [Hae48] suggested the use of range-bar charts not only for the comparison of ranges of data, but also for ex-

pressing central measures such as median, mean, mode, standard deviation, and tolerance limits through annotations on the chart. This idea was extended to displaying the 5-number summary on the range-bar chart [Spe52], as seen in Figure 1b, by shortening the bar to encompass only the central 50% of the data, using a thin line to indicate the entire range, and a perpendicular line to show the median. This is the first appearance of the form of the box plot we know today. The Tukey box plot became a popular representation and was introduced in 1977 [Tuk77], Figure 1c. This plot truncated the length of the range-line to 1.5 times the length of the innerquartile range. Outliers are indicated by independently marking them on the plot. The look of Tukey's box plot is also refined from that of the range-bar chart. The box fill is removed and the end of the range-line is clearly marked. The visual refinement of the box plot continued with the introduction of the quartile plot [Tuf83], Figure 1d, which sought to reduce visual clutter and maximize the ink-to-paper ratio by removing the box completely, and indicating the innerquartile range by an offset line. The median is simply a break in the innerquartile line. Other versions of this plot indicate the median using a small square and remove the innerquartile line, letting the empty space between the two range-line segments represent the central quartiles. While these plots do reduce the amount of ink used to indicate the 5-number summary, they may also reduce the ease in interpretation of the plot due to the subtle way that the median is indicated, and the similar technique used to show both the range of the data, and the innerquartile range. Furthermore, reducing the amount of area taken up by the innerquartile representation is counterintuitive since this region contains the majority of the data, a fact which the plot should clearly express. The abbreviated box plot [PKR06], as seen in Figure 1e, is another approach which reduces the ink needed to convey the 5-number summary, specifically for the purpose of superimposing further summary statistics on top of the plot. This method maintains the original form of the box plot, but removes the sides of the box, leaving only the corners.

3   Modifications of the Box Plot

One of the major advantages of the box plot is its simplicity of design. Critical information about a dataset is quickly expressed, and the box itself is a signature of the distribution. General characteristics such as the symmetry of the distribution, the location of the central value, and the spread of the observations are immediately apparent. This concise representation allows for the inclusion of additional information about the dataset, and permits the user to customize the plot for specific purposes.

3.1   Density Information

One of the most common types of information added to the box plot is a description of the distribution of the data values. The box plot summarizes the distribution using only 5 values, but this overview may hide important characteristics. For instance, the modality (or number of most often occurring data values) of a distribution is hidden by the box plot, and distinctive distributions with varying modality may be encoded using similar looking
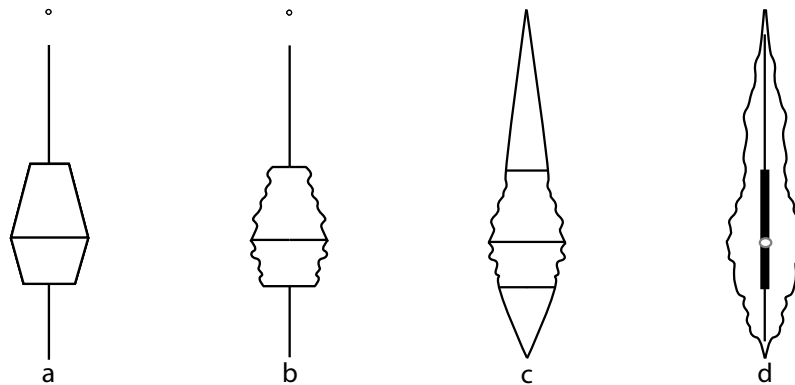
Figure 2: Examples of methods for adding density to the box plot. The a) histplot, b) vaseplot, c) box-percentile plot, and d) violin plot.

box plots. This is especially problematic when no prior information is known about the distributions, as comparing distributions with differing modalities may not be appropriate. One solution to these types of problems is to add into the box plot indications of the density of underlying distribution.

The histplot [Ben88], Figure 2a, is a simple approach for adding density information to a box plot. In the histplot, the density of the distribution is estimated at the median and the two quartiles. The width of the box plot at these locations is then modified to be proportional to the density estimation, and lines are drawn to connect these widths, essentially changing the box of the box plot into a polygon. The histplot adds a quick summary of the density of the central area of the distribution, but it is still possible for important features to be missed. The vaseplot [Ben88], Figure 2b, is a refined version of the histplot which adds in estimated densities for every point between the upper and lower quartiles. A line is drawn between each density estimation point (on both sides), and the polygon of the hist plot is replaced with something that, depending on the distribution, resembles a vase. This modification explicitly shows the density of the central 50% of the data. In addition, confidence intervals can be added to both of these plots by superimposing a light gray shaded bar over the median with the height of the bar signifying confidence.

The box-percentile plot [EB03], Figure 2c, is another method for adding the empirical cumulative distribution of the dataset into the box plot. In this type of plot, both sides of the box plot are used to plot the percentile of the distribution at each point. Thus, for each position in the plot, the width of the box is proportional to the percentile of that data value, up to the 50th percentile, at which point the width is switched to being proportional to minus the percentile. The sides of the plot are symmetric and the 25th, median, and 75th percentiles are marked with a line. The advantages of this plot are that there is no question as to how it should be drawn, it covers the entire range of data, and it does not use any arbitrary choices for its creation. Additionally, the plot is straightforward enough to be understandable by untrained readers, but includes details for trained readers.

The violin plot [HN98], Figure 2d, combines the standard box plot with a density trace to exploit the information contained in both types of diagrams. The box plot is used to show the innerquartile range, however, it is modified in two ways. The first modification changes the box plot by making the box solid black and replacing the median line with a circle; this allows for quick identification of the median and easy comparisons. The second modification removes the individual symbols for outlying data values since the outliers are contained in the density trace and individual points would clutter the diagram. A density trace is added as an alternative density estimator to the histogram and gives a smoother indication of frequency by allowing the intervals in which density is calculated to overlap, in contrast to the histogram. The density trace [CCKT83] at value y, $f(y)$ is defined by

$$f(y) = \frac{1}{hn} \sum_{i=1}^{n} W(\frac{y - y_i}{h}), \quad \text{where} \quad W(u) = \{ \begin{array}{l} 1 \text{ if } |u| \leq \frac{1}{2}, \\ 0 \text{ otherwise} \end{array}$$

where n is the sample size, and h is the interval width. The trace is added to the violin plot as two symmetric curves on either side of the box plot, making the density and magnitude easy to see. The main factor that controls the look of the density trace is the size of the interval width h. There is no specific size that works best in every situation, but an h value around 15% of the data range often produces good results; and the h values should stay between 10 and 40% of the data range to maintain a pleasing smoothness of the density trace curve.
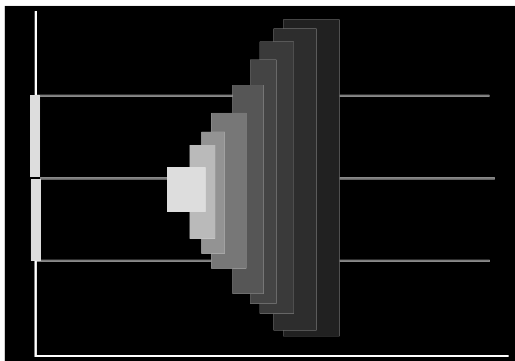

Figure 3: The sectioned density plot.

The sectioned density plot [CC06], Figure 3, exploits characteristics of the human visual system to present, in implied 3D, shape information of a data distribution and trends in variance and central tendency. The human visual system is capable of using occlusion and intensity variation as cues to spatial depth. The sectioned density plot uses these cues to display the distribution of a dataset in order to create the illusion of 3D. To create a sectioned density plot, the data is partitioned into fixed-width intervals, the number of which is variable. Each of these intervals is plotted onto a black background. From lowest density to highest, each interval is plotted using a rectangle shifted slightly to the left, occluding

the previous interval, and filled with a monotonically increasing intensity. The 5-number summary is incorporated into these plots by using the coordinate axis to show the range of the values, indicating the upper and lower quartiles with thin rectangles superimposed on the axis, and the median as a break in the range line. Each of these values is extended through the graph as a thin white line.

3.2 Additional Descriptive Statistics

Often there are instances in which the 5-number summary is not enough information, however adding a density plot is not feasible or necessary. For instance, when doing a comparison of multiple datasets, adding the density distribution of each dataset may clutter the plot, however, it would be useful to have information such as the relative number of observations. Additional information may also reduce the possibility of the user making false conclusions.
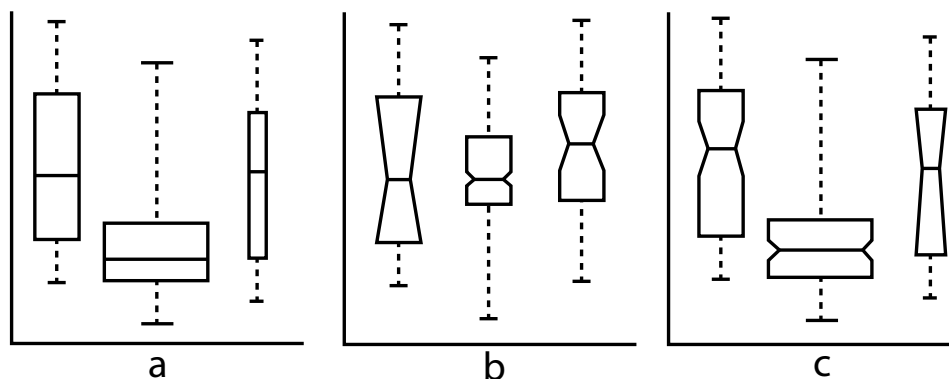

Figure 4: Variations of the box plot. a) Variable width box plot. b) Notched box plot. c) Variable width notched box plot.

McGill et al. [MTL78] suggested a few minor modifications of the original box plot to address these issues. The first variant is the variable width box plot which can be seen in Figure 4a. This plot uses the width of the box to proportionally encode the size of the dataset. The addition of this size clue easily alerts the viewer to distinctions in the number of observations in each dataset and can help the viewer avoid misinterpretation. The second variant proposed is the notched box plot as shown in Figure 4b. In this plot, notches are added to the box plot to roughly indicate the significance of differences between values or the confidence level of the data. The last proposed plot is the variable width notched plot which combines the information contained in the previous two plots and can be seen in Figure 4c.

One of the drawbacks of the simplicity of the box plot is that the box plot can hide distinguishing features of a distribution, and possibly encode very different distributions in

similar plots. The addition of density information tried to solve this problem, but it is not always feasible to add in this (possibly) large amount of data. An alternative to using density information is to use statistics that describe specific characteristics of a distribution.

An example of adding descriptive statistics to the box plot is the addition of skew and kurtosis measures. Skew and kurtosis are statistics which describe the symmetry and peakiness of the distribution and can indicate modality. One method for adding these measures into a box plot thickens the sides of the box when these measures indicate skew or high kurtosis in a specific direction [CM05], Figure 5a. The topmost plot indicates that the distribution is skewed toward the right. A bimodal distribution would be skewed in both directions, and this is shown in the center plot in which both ends of the box plot are thickened. Finally, a distribution that is centrally peaked has the median line of the box plot thickened, as shown in the bottom plot. This technique quickly conveys an indication of these statistics and can be used to distinguish between differing distributions.
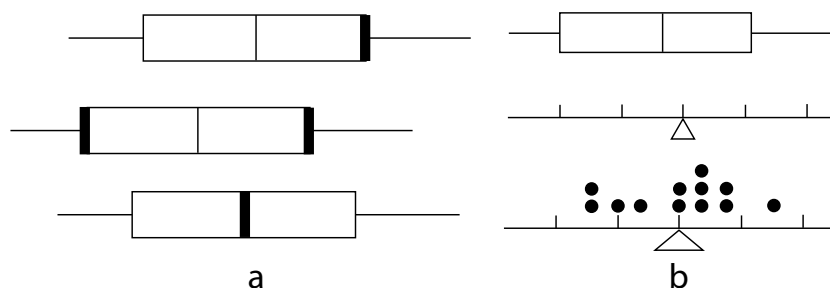

Figure 5: a) Box plots with varying skew and kurtosis. From top to bottom: right-skewed, bimodal, and centrally peaked distributions. b) A Beam and Fulcrum diagram. A dot plot is added to the bottom figure to indicate frequency and the size of the fulcrum base shows the width of a confidence interval.

The beam and fulcrum display [DT00] is a complementary diagram to the box plot and this combination can be seen as the two diagrams at the top of Figure 5b. In this type of display, the range is represented as a line (or beam) and the fulcrum, represented as a triangle, is placed at the mean. On each side of the fulcrum, tick marks are used to show standard deviation points. As seen in Figure 5b, bottom, a dot plot can be added to the beam and fulcrum display to show the frequency of data values, and the size of the fulcrum base can be modified to express the width of a confidence interval. The benefits of such a diagram when presented alongside a box plot are that the user is able to quickly pick out non-normal distributions (i.e., when the mean and median are not equal), see where the data are distributed with respect to the standard deviation scale ($\pm\sigma, \pm2\sigma, ...$) , and easily find outliers, (i.e., data points outside 3 standard deviations). It is also a useful learning tool, students can easily understand that the mean balances the distribution.
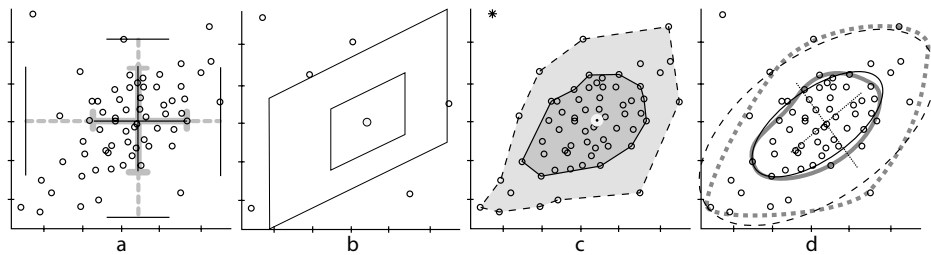
Figure 6: Bivariate extensions of the box plot. a) Rangefinder box plot. b) The Bagplot. c) The 2D box plot. d) The Quelplot and Relplot.

3.3 Bivariate Extensions

Standard implementations of the box plot focus on univariate data distributions. The 5-number summary is a useful descriptor of not only univariate, but also bivariate data distributions. The main challenge in extending the box plot for use with higher dimensional data is how to translate the 5-number summary values, which are vector values in the bivariate case, into visual metaphors with meaningful spatial positions, while maintaining the simplicity of the original box plot. A bivariate box plot can show not only the location and a summary of the data distribution, but also skew, spread and correlation.

A rangefinder box plot [BG87], as seen as the solid back lines in Figure 6a, is a simple extension of the box plot into 2D. To create a rangefinder box plot, all data values are plotted as points on a 2D graph (this is often called a scatterplot). For each variable, the 5-number summary is calculated, a line segment is drawn along the innerquartile range and perpendicular lines are placed at the adjacent values of the variable, where the 1D box plot would terminate. The intersection of the two central line segments is the cross-median value. This idea was further improved upon, as shown as the thick gray lines in Figure 6a, to emphasize the quartiles rather than the range, by moving the perpendicular lines from the adjacent values to the upper and lower quartile positions and extending whisker lines to the extrema value of the variable [Len88]. These extensions of the box plot into 2D are an unobtrusive expression of the summary of each variable, but the *correlation* between the two variables is not visible.

Other techniques for extending the box plot into 2D all use the notion of a *hinge* that encompasses 50% of the data and a *fence* that separates the central data from potential outliers. The distinctions between each of these methods are the way the contour of the hinge and fence are represented, and the methods used to calculate the contours.

The two-dimensional box plot [Ton05], as seen in Figure 6b, computes a robust line through the data by dividing the data into three partitions, finding the median value of the two outer partitions and using these points as the line. Depending on the relationship between the slope of the line and each variable, the quartile and fence lines are drawn either parallel to the robust line, or parallel to the variable's coordinate axis. The lines not comprising the outer-fence and the inner-hinge boxes are removed.

104

The bagplot [RRT99] uses the concept of halfspace depth to construct a bivariate version of the box plot, as seen in Figure 6c. The halfspace depth $ldepth(\theta|Z)$ of some point θ is the smallest number of data points $z_i \in Z = z_1, z_2, ..., z_n$ contained in any closed halfplane with a boundary line through θ. The depth region D_k, which is a convex polygon, is the set of all θ with $ldepth(\theta|Z) > k$ and $D_{k+1} \subset D_k$. To construct the bagplot a scatterplot of the data is first created. The *depth median* is then found which is the θ with the highest $ldepth(\theta|Z)$, if there is only one such θ, otherwise it is the center of gravity of the deepest region. This point, which is at the center of the plot, is represented as a cross. The *bag* is a dark gray region in the plot encompassing 50% of the data. The *fence* separates the outliers of the dataset, but is not drawn, and the *loop* is a light gray region of the plot that contains points outside of the bag, but inside of the fence. Outliers are highlighted as black stars. Options for reducing the visual clutter of the bagplot are to not plot data points contained in the bag, and to not fill the regions contained in the bag and the loop, but instead surround the bag with a solid line and the loop with a dashed line. In addition, a confidence region can be added into the bagplots as a *blotch* drawn around the depth median.

The relplot and the quelplot [GI92] use concentric ellipses to delineate between the hinge and fence regions. Both the relplot and quelplot can be seen in Figure 6d. The relplot uses full ellipses which assume symmetric data and are constructed using a robust estimator such as the minimum volume ellipsoid. In the figure, the relplot is shown as ellipses drawn in thin black lines. The quelplot divides the ellipses into four quarters aligned on the major and minor axes, and computed using an M-estimator. The quelplot is shown in the figure as thick, gray lines. The quelplot can show skewed data, since each quarter ellipse can be tranformed individually.

4 Conclusion

The box plot is a standard technique for presenting a summary of the distribution of a dataset. Its use has become prevalent in all forms of scientific inquiry, and understanding its construction, origins, and modifications can help not only with interpretation of the information presented by the box plot, but also in its creation and use. The concise representation provides not only insights to the important characteristics of a distribution, but permits the addition of information which enables the customization of the box plot to specific scenarios. Overall, the simplicity of the box plot makes it an elegant method for the presentation of scientific data.

5 Acknowledgments

References

[Ben88]   Yoav Benjamini. Opening the Box of a Boxplot. *American Statistician*, 42(4):257–262, November 1988.

[BG87]   Sean Becketti and Willian Gould.   Rangefinder Box Plots.   *American Statistician*, 41(2):149, May 1987.

[CC06]   Dale J. Cohen and Jon Cohen.   The Sectioned Density Plot.   *American Statistician*, 60(2):167–174, May 2006.

[CCKT83]   John M. Chambers, William S. Cleveland, Beat Kleiner, and Paul A. Tukey. *Graphical Methods for Data Analysis*. Wadsworth, 1983.

[CM05]   Chamnein Choonpradub and Don McNeil. Can the Box Plot be Improved? *Songklanakarin Journal of Science and Technology*, 27(3):649–657, 2005.

[DT00]   David P. Doane and Ronald L. Tracy.   Using Beam and Fulcrum Displays to Explore Data. *American Statistician*, 54(4):289–290, November 2000.

[EB03]   Warren W. Esty and Jeffery D. Banfield. The Box-Percentile Pot. *Journal of Statistical Software*, 8(17), 2003.

[FHI89]   Michael Frigge, David C. Hoaglin, and Boris Iglewicz. Some Implementations of the Box Plot. *The American Statistician*, 43(1):50–54, February 1989.

[GI92]   Kenneth M. Goldberg and Boris Iglewicz. Bivariate Extensions of the Box Plot. *American Statistician*, 34(3):307–320, August 1992.

[Hae48]   Kenneth W. Haemer. Range-Bar Charts. *American Statistician*, 2(2):23, April 1948.

[HN98]   Jerry L. Hintze and Ray D. Nelson. Violin Plots: A Box Plot-Density Trace Synergism. *American Statistician*, 52(2):181–184, May 1998.

[Len88]   Russel V. Lenth. Comment on Rangefinder Box Plots. *American Statistician*, 42(1):87–88, February 1988.

[MTL78]   Robert McGill, John W. Tukey, and Wayne A. Larsen. Variations of Box Plots. *American Statistician*, 32(1):12–16, February 1978.

[PKR06]   Kristin Potter, Joe Kniss, and Richard Riesenfeld. Visual Summary Statistics. *To Appear*, 2006.

[RRT99]   Peter J. Rousseeuw, Ida Ruts, and John W. Tukey. The Bagplot: A Bivariate Boxplot. *American Statistician*, 53(4):382–387, November 1999.

[Spe52]   Mary Eleanor Spear. *Charting Statistics*. McGraw-Hill Book Company, INC., 1952.

[Ton05]   Phattrawan Tongkumchum. Two-Dimensional Box Plot. *Songklanakarin Journal of Science and Technology*, 27(4):860–866, 2005.

[Tuf83]   Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[Tuk77]   John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.