Overview

Exploratory data analysis

Stephan Morgenthaler*

Exploratory data analysis, or EDA for short, is a term coined by John W. Tukey for describing the act of *looking at data to see what it seems to say*. This article gives a description of some typical EDA procedures and discusses some of the principles of EDA. © 2009 John Wiley & Sons, Inc. *WIREs Comp Stat* 2009 1 33–44

INTRODUCTION

An exploratory analysis looks at the data from as many angles as possible, always on the lookout for some interesting feature. The data analyst is interested in uncovering facts about the data and may use any procedure of his/her liking to this end. The only limits to such an analysis are those imposed by time constraints and the creativity of the data analyst. EDA is not guided by a desire to confirm the presence of a particular effect, and it is not supported by a statistical model that incorporates a mathematical expression for such an effect.

With such a broad mandate, it is difficult to structure a presentation of EDA. We could follow Tukey's lead and use the type of data as a framework. In Tukey,[1] the first four chapters deal with a single series or multiple series of observations of the same variable. The next five chapters are about regression-like situations, and the next four deal with tables or arrays of observations in which the margins describe different circumstances. Further topics, in particular, questions related to counted data are treated in the last eight chapters. Another possible framework for discussing EDA procedures is the broad attitude underlying the method. These include the *classical mode* of thinking, which uses a specific model and derives procedures that are appropriate for that model, for example, those based on the likelihood function. A second broad attitude is the *exploratory mode*. Flexible, forgiveness, and ease of computation are the main characteristics of such procedures. As an illustration, Tukey used the image of a sailboat that can go anywhere, does not easily tip over, and is easy to sail. A third mode englobes *rough confirmatory* procedures, which are used to identify findings that merit a closer study. If a careful study is undertaken, methods that work well in the face of a variety of realistic circumstances are called for. Tukey called this *mustering strength*. In the remainder of this entry, we will make some general comments on the philosophical foundations and then focus on some procedures that are typical for EDA.

*Correspondence to: stephan.morgenthaler@epfl.ch

Institute of Mathematics, Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland

DOI: 10.1002/wics.002

WHAT IS EDA?

When a statistician is asked to assess a series of observations $y_1, \ldots, y_n$ taken under identical conditions, he or she is most likely to compute the average $\bar{y} = \sum_{i=1}^{n} y_i/n$, the standard deviation $s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$, a Student's t confidence interval for the mean, and maybe a χ^2-based confidence interval for the variance. When modeling the observations as independent realizations drawn from a normal population with unknown mean μ and variance σ^2, these procedures are fully justified. They can also be justified in a weaker, asymptotic or least squares or permutation sense. In most instances such models and their conditions are merely implicit and not explicitly specified.

In the EDA mode, on the other hand, there is no need to consider a model. The data are regarded as a list or batch of numbers, not necessarily representing an underlying population. As the name EDA suggests, one is free to choose any procedure to analyze the data, and the primary aims are to look at the data and to think about the data from many points of view. Informal conclusions are drawn in this manner. Graphical visualization is usually the first order of business. In the case of a single list of numbers, examples of such graphical displays include the drawing of a single axis with the observations indicated by some symbol (called a dot plot), a histogram, a normal quantile plot, or a box plot. These show the data in more or less detail. The choice of procedure may depend on what one has already learned about the data. If a histogram shows two distinctive modes or if the number of observations

is very small, for example, the box plot may not be an appropriate summary of the data. On what scale the data ought to be analyzed is another question of importance. Should we use the numbers as given to us or would the logarithm be an appropriate re-expression? If one decides to summarize the data by a few numbers, selected quantiles offer a better choice than the traditional average and standard deviation.

EDA is sometimes presented as a toolbox. But this aspect is not its essence, it is merely a consequence of the exploratory attitude taken by the data analyst. EDA imposes no restrictions on the procedures to be used, and it is only natural that data analysts come up with new ways to look at data.

From an exploratory analysis, conjectures about the series of observations may emerge, but in the absence of a model and of the concept of an underlying population, no careful checking of the validity of such findings will be done. In order to confirm a finding, new data would have to be gathered in a manner that ensured their relevance and excluded unwanted side effects. In a confirmatory analysis, effective procedures of analysis would have to be applied. The aim in EDA is finding interesting indications in the data, without too much regard to the strength of the evidence.

Those favoring EDA recognize that traditional methods of mathematical statistics with their emphasis on models, sampling distributions, hypothesis testing, and other inferential tools are rarely applicable. Learning from data is mostly done in a context in which one could not apply the traditional paradigm. At the same time, there is no clear boundary between EDA and *statistics*. If anything, EDA adds new tools to the ones based on traditional models and in this sense forms a superset.

Students and users of statistics are often perplexed by the choices open to them. Should one do a regression analysis or would a principal component analysis be more appropriate? Or does the data call for a hierarchical classification? For any given dataset, multiple 'right answers' are possible, and different data analysts will typically approach a dataset with different tools. EDA recognizes this fact. Specifying the nature of the data and explaining how it came about does not dictate how to analyze it. Data analysis is a creative process in which the aims of the data analyst and his attitude and knowledge play a crucial role.

A Brief History

The title of this contribution is taken from an influential book by Tukey,[1] of which a three-volume preliminary edition with the same title had appeared in the early 1970s. These texts present some of the material that Tukey taught in various venues, most prominently in a course that he gave regularly to the students of statistics at Princeton University since 1968.[2]

By the 1950s, statistics had to a very large extent taken the meaning of *inference from the particular to the general*. In his writings and teachings, John W. Tukey found it useful to offer the more accurate *data analysis* in order to describe what statisticians actually did. In a paper published in 1962,[3] he explained that data analysis includes: *procedures for analyzing data, techniques for interpreting the results of such procedures, way of planning the gathering of data, and all the machinery and results of (mathematical) statistics which apply to analyzing data*. Statistics in the sense of making inference from a sample to a population are merely a part of data analysis, not the whole.

The historical roots of both senses of the word statistics-confirmatory versus exploratory as Tukey put it-are of course much deeper than the middle of the 20th century. By the 1950s, the pendulum had swung far in the direction of confirmatory data analysis. Tukey gave it a shove in the opposite direction. He argued for the importance of data exploration, which must necessarily precede model-building and confirmation. His writings (Tukey, collected works, see references) include a number of important articles on the philosophical foundation of data exploration. And, of course, he proposed many ingenious methods, which gained wide acceptance and changed the way statistics was practiced.

SOME TYPICAL EDA PROCEDURES

Tukey[1] presented a number of novel tools for the purpose of exploratory data analysis. We will review some of the most popular ones in this section. Readers are invited to look in the corresponding parts of Tukey[1] for a more detailed description and additional comments. The list of well-known procedures contains the stem-and-leaf, hinges, the five-number summary, letter values, the box plots (schematic plots), easy re-expressions, parallel box plots, data = fit PLUS residual, straightening plots, data = smooth PLUS rough, running median smoothers, wandering box plots, delineations (traces), two-way analyses, median polish, two-way plots, plus-one fits (non-additivity), three-way fits, level traces (slicing batches), re-expressions for fractions, re-expression for counts, octave bins, and exploring shapes of distributions.

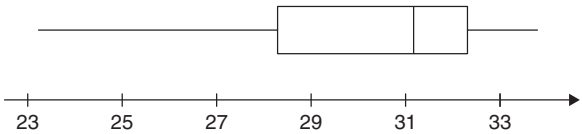
Median

The median M of a batch of numbers is its innermost value. If the batch contains $n = 5$ numbers, the median is the third number counting either from the smallest to the largest number or from the largest to the smallest number. Because of this, we say that the depth of the median is three. For a batch of size n, the median depth is defined as $m = (n + 1)/2$, which for even batch sizes of the form $n = 2k$ leads to a value half-way between the two integers k and $k + 1$. In this case, the median is defined as the average between the numbers with depths k and $k + 1$.

For EDA purposes, the median is preferable to the average. For one, the median is a *resistant* summary of the data. If a small part of the batch values are changed, no matter how much, the median resists and is not much affected. This is obviously not true for the arithmetic average. Experience suggests that most batches contain exceptional values, even if the data were collected to the highest standards. Because of this, resistance is important.

Box Plot

The box plot is one of the most widely used methods of EDA. In its basic form, it is based on the five-number summary, in which a batch of numbers is represented by the two extreme values (E), the two hinges (H), and the median (M). The following example serves as an illustration. The numbers are 23.22, 26.00, 28.29, 31.85, 32.31, 31.01, 33.80, 32.51, 31.17 and are equal to the number of newly constructed apartments (in thousands) in Switzerland between 1978 and 1986. The sorted batch is $23.22 < 26.00 < 28.29 < 31.01 < 31.17 < 31.85 < 32.31 < 32.51 < 33.80$. The median has depth $(9 + 1)/2 = 5$ and is equal to $M = 31.17$. The hinge depth is computed by the same way as the median depth, but starting from the median depth instead of n, that is, $(5 + 1)/2 = 3$. The hinges are obtained by taking the third number from each end of the batch and are equal to $H = (28.29, 32.31)$. The difference between the two hinges is called the *H-spread* and gives an indication for the variation of the batch. The extremes are, of course, easiest, they are simply the first number from each end. In our example, $E = (23.22, 33.80)$ and the resulting box plot is:



The box spans the interval between the hinges with a separator at the median. The whiskers extend from the box to the extremes. The five chosen numbers separate the batch into four intervals, delimited by the lower extreme and the lower hinge, the lower hinge and the median, the median and the upper hinge, and the upper hinge and the upper extreme. By construction, the nine numbers of the batch are distributed roughly in equal proportions between these four parts.

If the batch size is $n = 4k + 1$ with $k = 1, 2, \ldots$, the calculations of the median and the hinges are straightforward. The median depth is $2k + 1$ and the hinge depth is $k + 1$. For other values of n, some interpolation is necessary. In general, the hinge depth is defined as $h = ([m] + 1)/2$, where $[m]$ is the integer part of the median depth. This rounding avoids the complication of having to deal with quarter depths.

Various modifications of the basic box plot have been proposed. It is, for instance, a good idea to mark exceptional numbers within the batch, if they occur. For this purpose, a simple rule has been devised. If one adds $1\frac{1}{2}$ times the H-spread to the upper hinge, an upper fence is found. Similarly, the lower fence is found by subtracting from the lower hinge. Any number outside of the fences is exceptional. In the modified box plot, the whiskers are not drawn all the way to the extremes, but rather to the largest and smallest numbers still within the fences. The observations outside the fences are separately marked and sometimes labeled.

Other, more elaborate generalizations are described in McGill et al.[4]

The box plot is a good example of an EDA procedure. Versatile, forgiving, and very easy to compute. Whenever several batches are to be shown in parallel, the box plot is very effective in displaying the information. Figure 1 shows an example from gas chromatography, which is a technique for separating a mixture of substances into its components. To study the effects of polar stationary phases, the retention was measured on a selection of 127 substances.

Figure 1 shows the effect on the retention when using one of seven polar stationary phases. The effect is defined as the difference between the observed retention and the retention under a standard non-polar stationary phase (see Kováts & Kresz,[5] for the data and further explanations). Inspection of the box plots shows that the effects of polarity is in almost all cases an increase in retention when compared with the standard non-polar stationary phase. The solvent labeled TMO has the largest effect, both in terms of the median and the H-spread. The solvents PCN and TTF follow closely behind. MFT and PSH have the smallest effect.

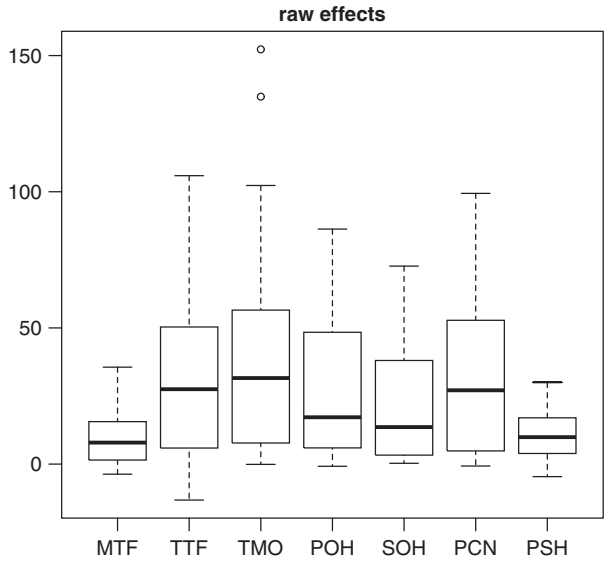


FIGURE 1 | The effect of seven polar stationary phases on the retention of 127 substances. The observations are for a temperature of 130°C. The exceptional substances in the case of the stationary phase labeled TMO are benzyl alcohol and 2-phenylethanol.

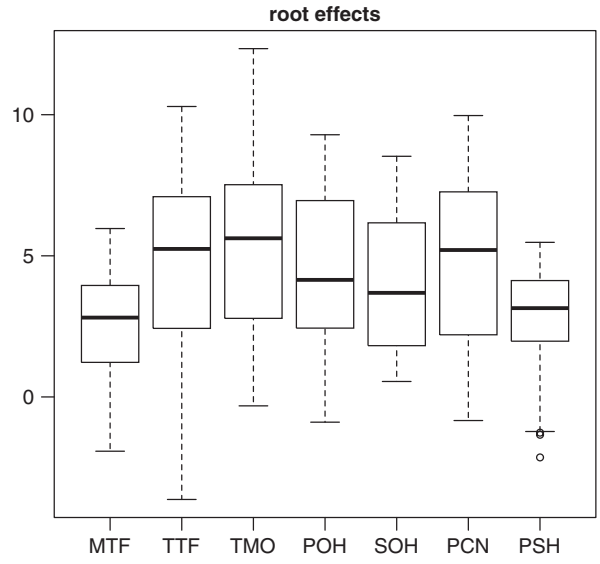


FIGURE 2 | The effect of seven polar stationary phases on the retention of 127 substances. The observations are for a temperature of 130°C. The data have been re-expressed on the logarithmic scale.

Re-Expressions

On what scale should one analyze a batch of numbers. Some simple rules can be given. If the numbers are amounts or otherwise positive by their very nature, it is a good idea to immediately decide to re-express them on a logarithmic scale. Whether the logarithm to base e or to base 10 or to base 2 is used, is of no importance.

Experience shows that for most batches of positive numbers the distribution about the median is far from symmetrical. The asymmetry shows itself in the box plot by the whisker being shorter on one side of the box than the other and the median divider not being in the middle of the box. In terms of the numbers, asymmetry means that one should not think of adding and subtracting. It is better to think of relative (percentage-wise) movements. Adding or subtracting \$6 from a stock price, for example, may mean a gain or a loss of 6%. Or, it may mean a gain or a loss of 20%. The answer depends on the price before the move.

Differences of logarithms automatically reflect relative moves of the untransformed numbers. This is, in most cases, a better scale for a variable that is intrinsically positive. The main benefit of the logarithmic transformation is the translation from a universe where actions are multiplicative to a more familiar one where actions are of the more familiar additive kind.

A welcome side effect is often a symmetrization of the box plot. After re-expression, the box plot has its median divider roughly in the center of the box, and the two whiskers are of about equal length.

A third benefit is the separation of the H-spread from the median. It is often the case that the H-spread increases with the median. This phenomenon can, for example, be seen in Figure 1. The bigger the median, the larger the box and the longer the whiskers. If the batches contain only positive values, the logarithm often corrects this. After re-expression of the numbers, the boxes are about equal size.

The difficulty caused by a linkage between the H-spread and the median should not be underestimated. It makes the comparison between batches difficult. If the boxes are of roughly equal size, the difference between batches can be described as a simple translation or shift.

Other functions than the logarithm can be used for re-expressing numbers. Choosing from among the square, the square root, the logarithm, the inverse of the square root, and the inverse offers a wide spectrum. Any of these can be applied directly to positive numbers, but if the batch also contains negative numbers it is more difficult.

If we denote by $y = f(x)$ the transformed value for $x > 0$, one can extend the definition to negative numbers by the formula $y = f(|x|)\text{sign}(x)$, where $|x|$ is equal to $-x$ for negative values of x and equal to x for positive values. Figure 2 shows the box plots of the retention effects after taking the logarithm. In this example, the box plots do become more symmetrical, but the logarithm seems to go too far and create extended whiskers towards negative values. If we

wanted to achieve more near symmetry, re-expression by the square root could be tried.

Re-expressing variables should always be kept in mind during an exploratory data analysis. Seemingly complex relationships and effects can be described with simple pictures and formulas when the right scale has been chosen.

Median Polish

The median polish is an EDA procedure for two-way tables. Suppose we observe a variable Y under conditions described by two factors. The data of Figure 1 could serve as an example. The variable is the retention effect, and the two factors are the 127 substances and the 7 polar stationary phases. For each combination of substance and stationary phase, once datum has been measured. For printing, the data could be arranged into a table with one of the factors constant along rows and the other constant along columns, and y_{ij} denoting the observation in row i and column j.

For such data, one wants to understand the way in which the variable Y depends on the two factors. Figure 1 shows the dependence on the stationary phases. We could, of course, change our point of view and show the substances instead, as in Figure 3.

How could we go further in analyzing such a two-way table? When the row and column factors are associated with numerical values x_i and z_j, the simplest description would be a linear regression with fitted cell values $\hat{y}_{ij} = b_0 + b_1 x_i + b_2 z_j$. This is easy to understand, only involves three constants and can be generalized by including further terms such as $b_{12} x_i z_j$.

This discussion provides an example of an important principle. The observations y_{ij} are written as a sum of two terms $y_{ij} = \hat{y}_{ij} + r_{ij}$, the fit PLUS residuals. The fit contains a structure that can be described by a mathematical formula, and the residuals contain the remaining structures and indications. Exploration of the dataset proceeds in stages. Features can be removed from the data in layers. First come the most obvious indications, followed by more subtle and hidden structures.

If no numerical values x_i and z_j are at hand to apply the linear regression, one can still use this same idea by replacing $b_1 x_i$ by a row effect c_i and $b_2 z_j$ by a column effect d_j, which leads to $\hat{y}_{ij} = b_0 + c_i + d_j$. There are more constants to be fitted, because we have to construct a numerical variable to go with the rows and columns, but otherwise the models are similar.

The constants b_0, c_i, and d_j are commonly fitted by the least square method, which proceeds as follows.

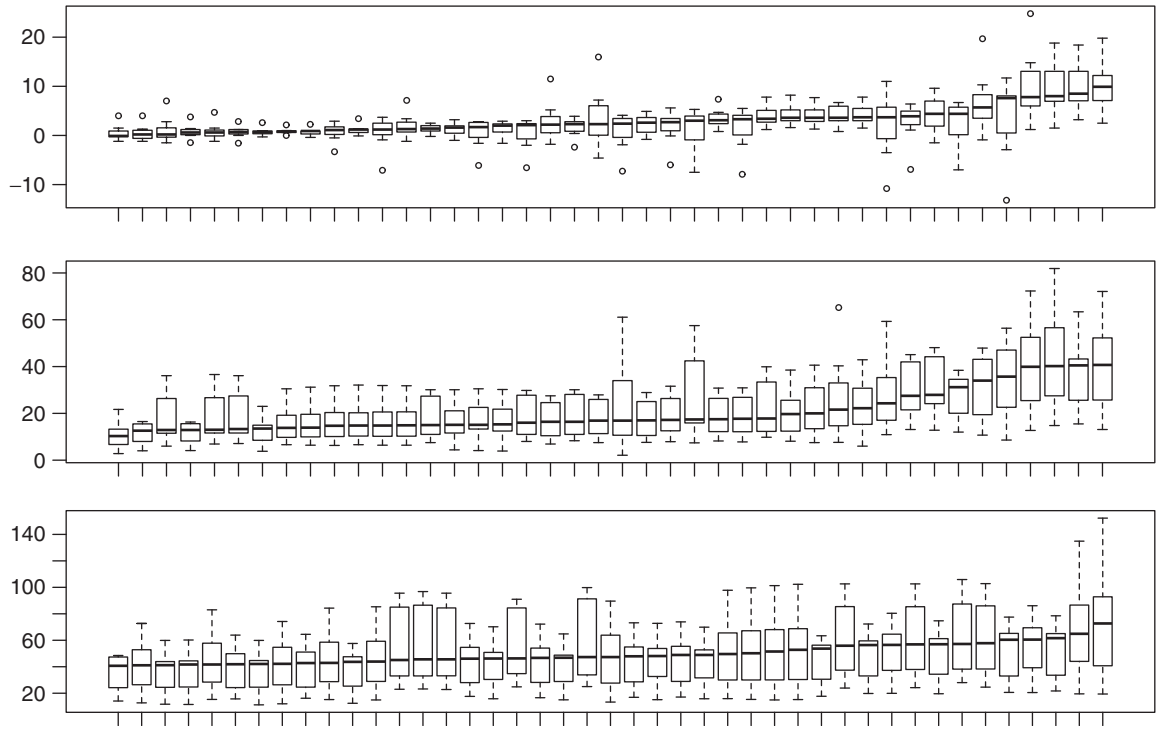


FIGURE 3 | The retention effect for the 127 substances. The substances are ordered by median value. The observations are for a temperature of the 130°C. Note that each box plot provides a summary of only seven numbers. With such small batch sizes, alternative displays that show all the data would be feasible.

Preliminary estimates of the row effects are calculated by taking the mean of the data values in the rows. These means are next subtracted from the rows of the data table. This computation of the mean and subsequent subtraction is called *sweeping out the mean*. The column effect and the residuals are then found by computing the means in the columns and by subtracting these means from the corresponding entries in the data table, that is, sweeping out the means from the columns. The constant b_0 and the final row effects are computed by the mean of the preliminary row effects and by subtracting it from the preliminary effects.

The *median polish* mimicks the least square method, but replaces the mean sweeps with the median sweeps. We can follow what would happen in our retention example by consulting Figures 1 and 3. The preliminary stationary phase effects are equal to the medians in Figure 1. Had we swept these medians from the data table and redone Figure 3 with the modified data, then the medians would be the substance effects. Finally, we would have to sweep the median from the preliminary solvent effects in order to compute the constant b_0 and the final solvent effects.

The mean sweeps lead to a table of residuals with zero mean in each row and each column. No further sweeping is necessary. This is not true for the median sweeps. As a consequence, once the columns have been swept, we may have to restart a new cycle of median sweeping with the rows. The final median polish has been reached when the residual table has zero median in all rows and all columns.

Residuals

To find further indications and to improve the understanding of the data, the residuals have to be looked at next. One can, for example, make a box plot of the residuals and compare it with the box plot of the data. If the fit is effective, one would hope

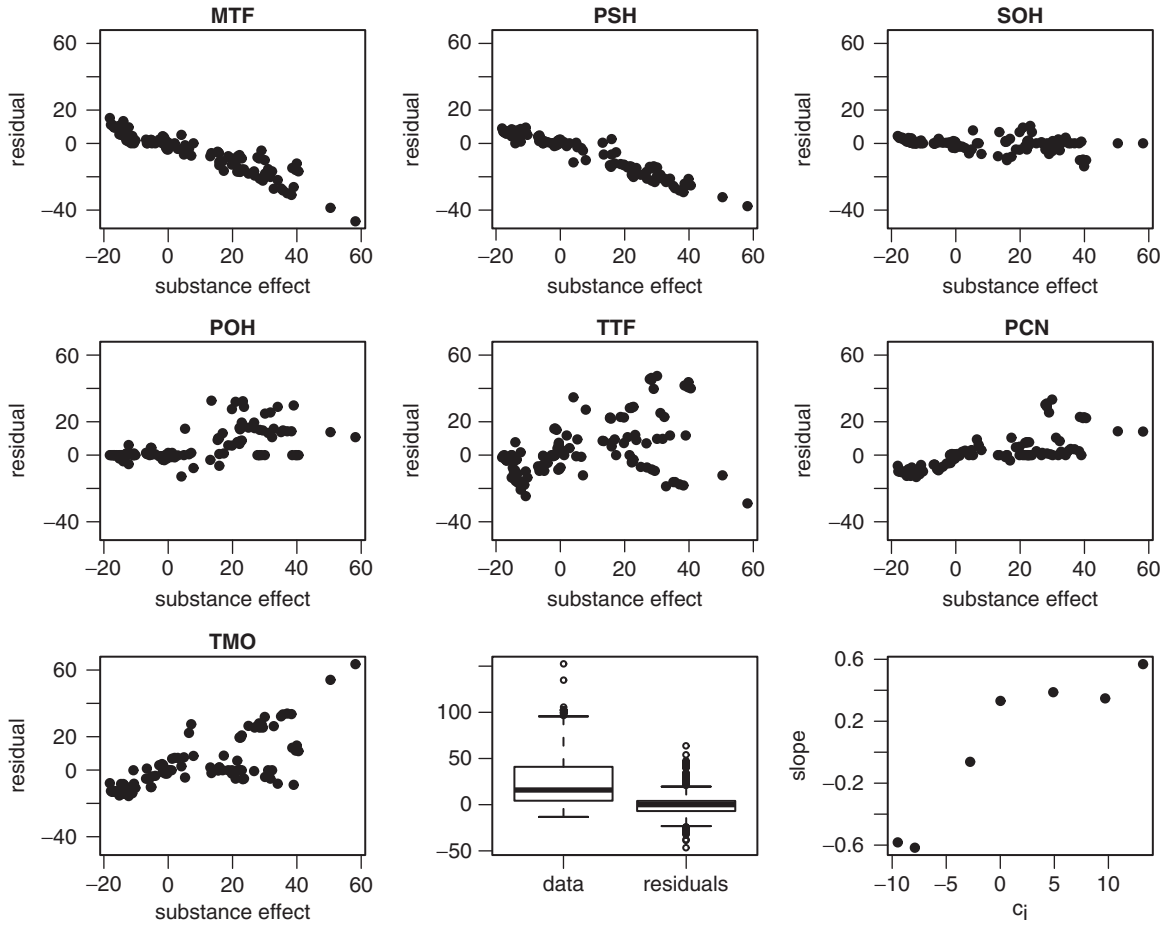


FIGURE 4 | The residuals obtained from a median polish are plotted against the substance effects, r_{ij} versus d_j. The polar stationary phases are ordered by effect size. The last plot shows in parallel the box plot of the original retention data and the box plot of the residuals after removal of the median effects.

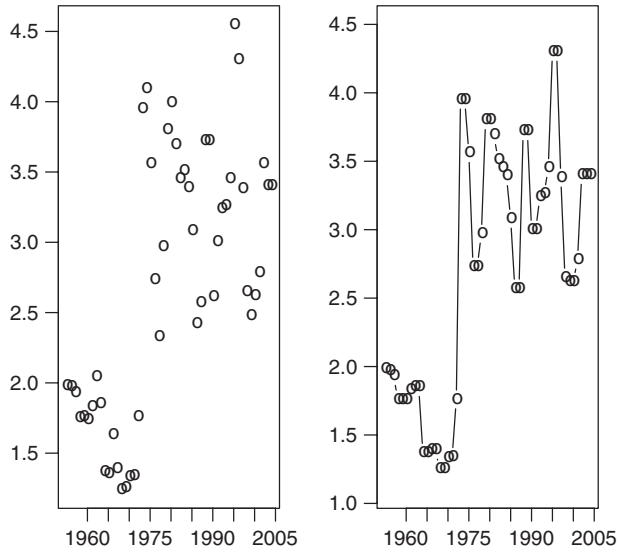


FIGURE 5 | The left-hand panel shows the raw data, the farm prices for wheat between 1955 and 2005 (see http://www.ers.usda.gov/Data/Feedgrains/). The right-hand panel shows the smoothed sequence.

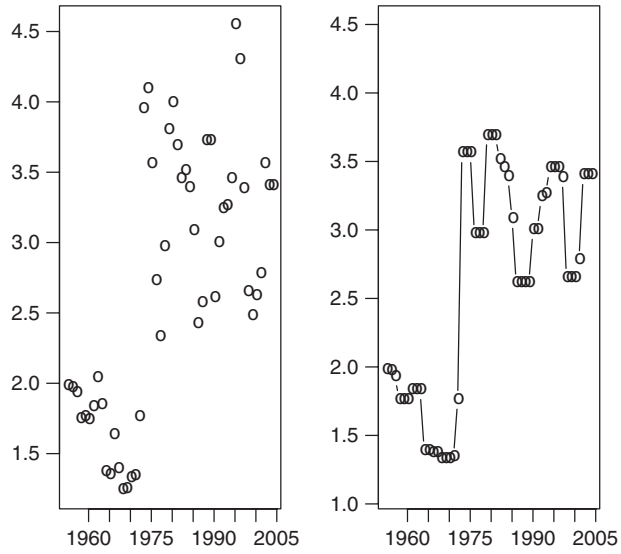


FIGURE 6 | The left-hand panel shows the raw data, the right-hand panel shows the smooth obtained by the smoother *3PR*. Compare this smooth with the one in Figure 5.

for a much reduced H-spread in the residuals. One can also plot the residuals of each row against the column effects, that is, r_{ij} versus d_j for each row i. And of course, we can equally well do the plotting of r_{ij} versus c_i for all j.

Figure 4 shows one of these plots for the retention data. As our eye moves across these seven charts, we notice a clear indication. The residuals depend linearly on the substance effects, with a slope that starts out being negative for MTF, but flattens out and finally becomes positive for PCN and TMO. The last panel shows that the residual H-spread decreases substantially following a full median sweep. The rightmost panel in the third row shows the slopes in the residual plots against the row effects c_i. The behavior of the residuals we observe in this example is not atypical and deserves to be explored further.

Non-Additivity

Recall the median polish fit $\hat{y}_{ij} = b_0 + c_i + d_j$. Figure 4 suggests that if we fit to the residuals a straight line in the column effects d_j with a slope that depends on the stationary phase effects c_i, we may improve our fit and further reduce the residuals. Such a term amounts to slope$(c_i) \times d_j$. The last panel of Figure 4 suggests a simple solution. Why not take slope$(c_i) = e \times c_i$? Adding this to the median polish results in $\hat{y}_{ij} = b_0 + c_i + d_j + ec_id_j$, which is nicely symmetric in the row and column effects. This model has been proposed in Tukey[6] and is called one-degree-of-freedom-for-non-additivity. The name is apt, because one actually needs to fit a single additional constant e.

To determine e, one can proceed by fitting the median polish and computing the residuals. Next, plot the residuals r_{ij} versus the product c_id_j. If this results in a scatter of points roughly centered around a straight line, find its slope e. This plot has also a diagnostic quality. Whenever it shows just a random blob of points, adding a term of the form c_id_j to the median polish is not helpful.

Running Median

For smoothing observations taken at regular intervals, the running median is the basic EDA tool of choice. The underlying idea is very simple and consists in applying the median to sequences of three subsequent observations. If the input sequence is $x_1, \ldots, x_n$, the output at index $i = 2, \ldots, n-1$ is $y_i = \text{median}(x_{i-1}, x_i, x_{i+1})$. The first and last values are computed as $y_1 = \text{median}(x_1, y_2, 3y_2 - 2y_3)$. The analogous end value rule for the last value is $y_n = \text{median}(x_n, y_{n-1}, 3y_{n-1} - 2y_{n-2})$. Note that $3y_2 - 2y_3 = y_2 - 2(y_3 - y_2) = \widehat{y}_{-1}$ is a linear extrapolation of the smoothed sequence.

Compared with the classical running mean, this procedure has added advantage of resistance. Isolated unusual observations are made to disappear. Figure 5 shows an example.

Because the median is a nonlinear function, filtering a sequence by a running median of three observations has surprising properties. Repeated application of the mean of three consecutive observations, for example, will converge to an output sequence that becomes increasingly smooth until it is finally arranged along a straight line. This is not true for the running median. The smoothed sequence from Figure 5 is obtained by repeating the running median computation until no further changes occur. This procedure is denoted by the tag *3R*, where R stands for repeated.

Smoothing in this manner has, however, also some bad surprises in store. The smooth sequence produced by the running median may not be smooth enough to the eye. It is, however, easy to produce a higher degree of smoothing, without destroying the advantages of the running median procedure. One can, for example, apply the running median and then take the output and process it with a running mean.

Another noticable feature of sequences smoothed by a running median is the all too frequent occurrence of two consecutive equal values at the top of peaks or at the bottom of valleys. This may sometimes be justified, but in other instances it is simply a feature created by the smoother. Thus smoothing by medians of length three demands particular attention to pairs of equal values. In Tukey,[1] a solution to this problem based on splitting the pairs and using the end value rule was suggested. The resulting smoother is called *3RSS*. Typically, after splitting, the pairs of equal values are replaced by longer stretches of constancy, but is not easy to understand and to explain. An alternative solution was suggested to the author by John Tukey in 1992.[7] It consists in looking ahead to the next observation when computing the running median.

The straightforward running median has $y_2 = \text{median}(x_1, x_2, x_3)$. When keeping an eye on couples of equal values, we change this to

$$y_2 = \begin{cases} x_2\,, & \text{if } x_1 = x_2 \\ \text{median}(x_1, x_2, x_4)\,, & \text{if } x_1 \neq x_2 \text{ and } x_2 = x_3 \\ \text{median}(x_1, x_2, x_3)\,, & \text{if } x_1 \neq x_2 \text{ and } x_2 \neq x_3 \end{cases}$$

The only change implied by this equation happens when x_2 is the first of either a couple or a triple of equal values. In this case, the original formula is modified by skipping x_3. The analogous equation is applied to compute y_{n-1}. The new equation for $i = 3, \ldots, n-2$ is

$$y_i = \begin{cases} x_i\,, & \text{if } x_i = x_{i+1} \text{ and } x_i = x_{i-1} \\ x_i\,, & \text{if } x_i = x_{i+1} \text{ and } x_i = x_{i+2} \\ \text{median}(x_{i-1}, x_i, x_{i+2})\,, & \text{if } x_i = x_{i+1} \text{ and } x_i \neq x_{i-1} \text{ and } x_i \neq x_{i+2} \\ \text{median}(x_{i-2}, x_i, x_{i+1})\,, & \text{if } x_i \neq x_{i+1} \text{ and } x_i = x_{i-1} \\ \text{median}(x_{i-1}, x_i, x_{i+1})\,, & \text{if } x_i \neq x_{i+1} \text{ and } x_i \neq x_{i-1} \end{cases}$$

The first two cases apply if x_i is either the first or second of a triple of equal values. The third case applies to the situation where x_i is the first of a couple of equal values. Here x_{i+1} is skipped in computing the running median. The fourth case applies when x_i is the second member of a couple or the third of a triple of equal values and leads to the skipping of x_{i-1}. The usual formula of the running median, finally, is applied in the last case, where x_i is surrounded by different values. The repeated version of this smoother is called *3PR*, where *P* stands for pairs. The end values can be smoothed as before. Figure 6 shows the result when applied to our data.

The Shape of Distributions

Figure 7 shows several pictures of the distribution of the daily log-returns ($r_i = \ln[d_i/d_{i-1}]$) of the DowJones index (d_i). The first is the traditional histogram, the second applies the square root transformation to the histogram counts, which means that 2500 on the vertical axis of the histogram becomes 50 for the rootogram, and the third shows a kernel density estimate, again transformed through the square root. The last picture plots the log-returns sorted from smallest to largest against the quantiles of a standard normal distribution.

The re-expression by square roots is useful in making the variation in the height of the bars more nearly uniform, but above all it gives a visually more pleasing picture and brings out the details that one is left guessing about in the raw histogram. The shape of the distribution is sharply peaked in the middle with fairly long tails to either small negative or large positive returns. A smudgen of asymmetry is visible.

The shape of a distribution of numbers is any feature that goes beyond the center and spread. An archetype of a shapeless pattern, entirely determined

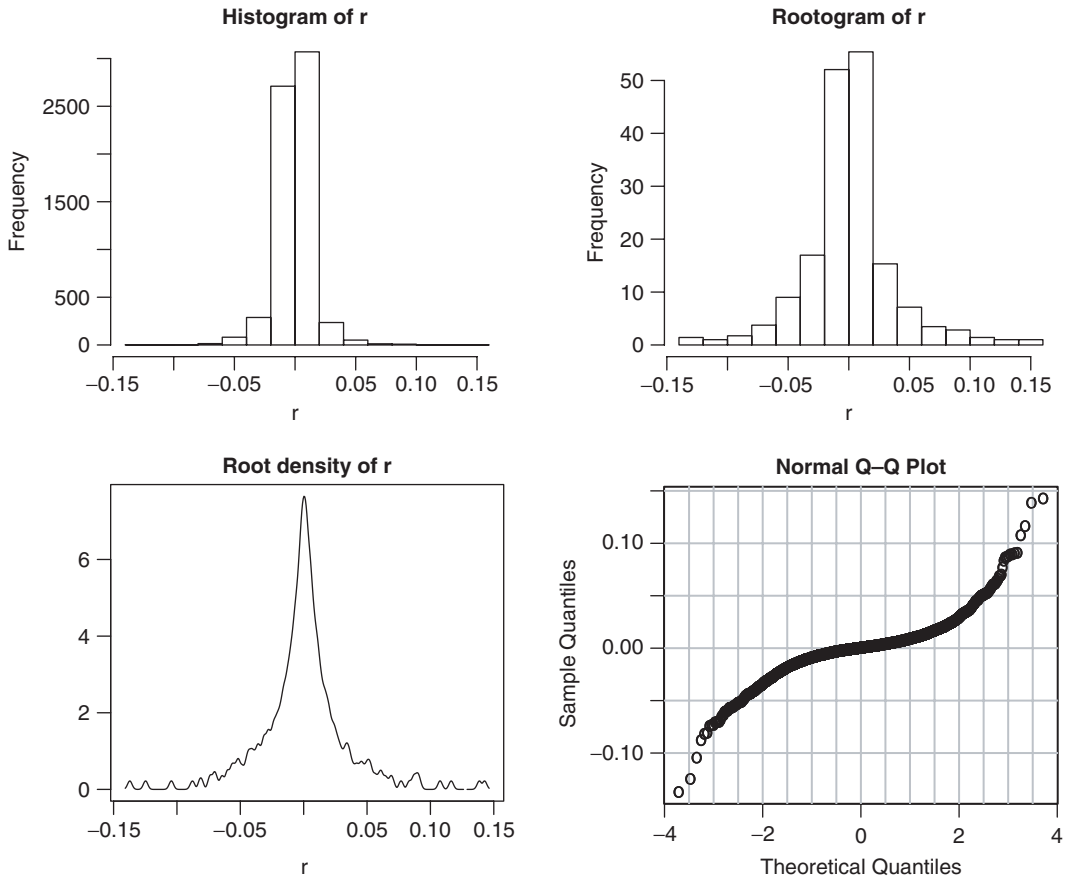


FIGURE 7 | Four ways of depicting the distribution of a batch of numbers.

by its center and spread, is the normal or Gaussian distribution. In some sense, the quantile plot in the last panel of Figure 7 is the most useful of the four possibilities, because it favors a direct comparison with the normal archetype. If the log-returns had the normal shape, one would see (roughly) a straight line. In the example, this is clearly not the case.

Quantiles

The median is an example of a quantile. The median is such that half the values are smaller and half are larger. Because 50% of the numbers are smaller or equal, we call the median as 0.5-quantile. In a similar way, α-quantiles are defined for any value $0 < \alpha < 1$. This definition of quantiles is not of practical use. Which value of α would we, for example, attribute to the smallest of a batch of n numbers? Which value of α_i would we attribute to the ith smallest?

Any formula of the type $\alpha_i = (i - a)/(n + 1 - 2a)$, where $0 \leq a \leq 1/2$ makes sense. The formula satisfies $\alpha_{n+1-i} = (n + 1 - i - a)/(n + 1 - 2a) = 1 - \alpha_i$, which is a basic symmetry requirement. It means that the first and the last, the second and the second last, and so on, are treated symmetrically. If $a = 0$, we take the sorted numbers of the batch as endpoints that divide the real line into $n + 1$ intervals. If $a = 1/2$, we imagine the real numbers divided into n pieces, with the sorted numbers in the 'middle' of these intervals.

The exact choice of a depends on our intentions. When $a = 1/3$ and $\alpha_i = (3i - 1)/(3n + 1)$, one gets a fairly good agreement of the batch quantiles with the medians of population quantiles in cases where the batches are samples from continuous distributions. These are the numbers we used in constructing the QQ-plot in the fourth panel of Figure 7. The theoretical quantile corresponding to the ith smallest number is computed as $z_{\alpha_i} = \Phi^{-1}(\alpha_i)$, where Φ is the standard normal cumulative distribution function.

g-h Technology

How could one describe the deviation from the normal shape? The QQ-plot gives an idea. Let Y be the variable of which we have a batch of numbers. The QQ-plot shows that Y can be written as a

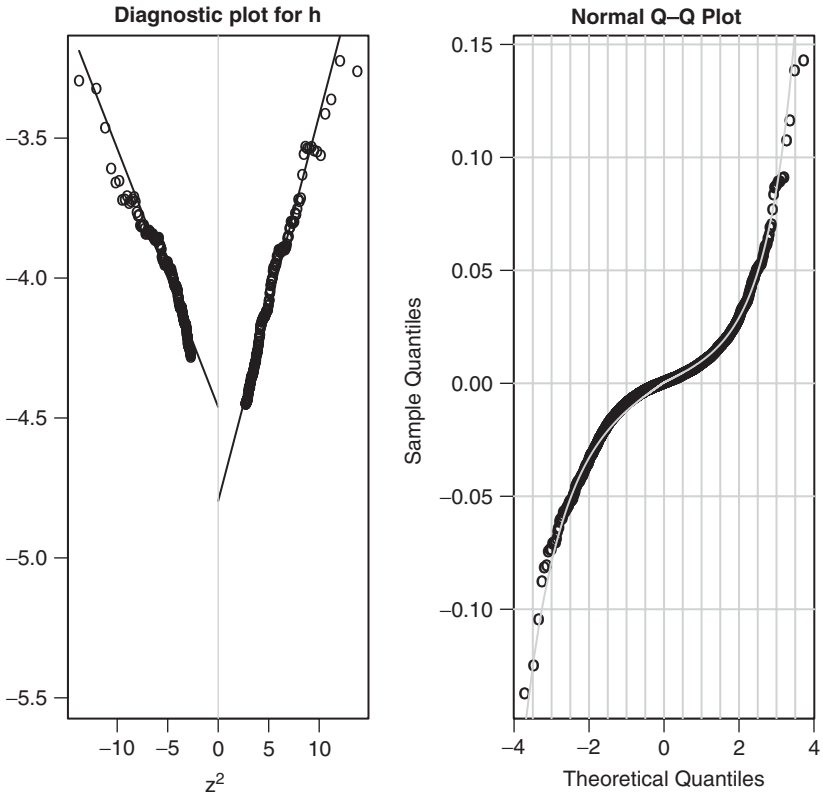


FIGURE 8 | The left-hand panel shows the plot of the logarithm of the ratio of half-spreads versus $z^2_{1-\alpha}$ and $-z^2_\alpha$, respectively. Using $-z^2_\alpha$ allows one to separate the two tails. Only $0 < \alpha < 0.05$ are shown. The resulting values for the parameters are $s = 0.011$, $h = 0.185$ (lower tail) and $s = 0.0083$, $h = 0.275$ (upper tail). The right-hand panel shows the QQ-plot from Figure 7 again, this time with the fitted h-distribution superposed.

smooth function of a variable with a standard normal distribution Z of the form

$$Y = m + sZV(Z)\,,$$

where m is the center, s is the spread, and the function $V(Z)$ serves as shape modifier. If $V(Z)$ is constant and equal to 1, the distribution of Y is normal, because Y differs from Z only by center and spread. If $V(0)$ is equal to 1 and $V(Z)$ increases as Z moves away from $Z = 0$, the shape is distorted and heavy-tailed distributions are created. Tukey's h-distribution is an example of this principle. It is based on the choice $V(Z) = \mathrm{e}^{hZ^2/2}$ for $h \geq 0$, so that

$$Y = m + sZ\mathrm{e}^{hZ^2/2}\,.$$

In an exploratory analysis of a batch of numbers, we determine the center m by the median. The values of s and h can easily be found with the help of half-spreads. Let $0 < \alpha < 1/2$. If we take the difference of the $(1-\alpha)$-quantile and the α-quantile, we obtain the α-spread. The difference of the $(1-\alpha)$-quantile and the median is an (upper) half-spread. Similarly, the distance between the median and the α-quantile is a (lower) half-spread. Because the normal distribution of Z is symmetric around zero, we have $-z_\alpha = z_{1-\alpha} > 0$. For Y, the computation of an upper half-spread leads to $sz_{1-\alpha}\mathrm{e}^{hz^2_{1-\alpha}/2}$. It follows that

$$\log\left(\frac{\text{half-spread of } Y}{\text{half-spread of } Z}\right) = \log(s) + hz^2_{1-\alpha}/2\,.$$

The left-hand side can be computed from the batch of numbers. The plot of it against $z^2_{1-\alpha}$ will then reveal s and h. Figure 8 shows the example of the financial data.

In this analysis, we have made use of the possibility of setting separate parameters s and h for the two tails. This takes care of the asymmetry. The higher value for h in the upper tail indicates that the positive returns are slightly heavier tailed. As a compensation, the spread parameter s is reduced. The h-distribution allows us to obtain a very good description for the quantiles of the log-returns.

It would also have been possible to modify the shape distorting function for the purpose of adding asymmetry. The real parameter g in

$$Y = m + sZ\frac{\mathrm{e}^{gZ} - 1}{gZ}\mathrm{e}^{hZ^2/2}$$

serves this purpose.

CONCLUSION

Processing data, noticing oddities, finding connections, and so on are fundamental to daily life and to science. Exploratory data analysis happens when one takes a broad view of this activity. EDA gives less weight to traditional views of statistics, which are centered on randomness, stochastic models, and population parameters, and which lead to questions involving the precision of estimates or the significance of a finding. EDA recognizes limitations to this traditional paradigm and puts more emphasis on data exploration using any procedure deemed appropriate by the analyst.

Which areas of the modern world of data analysis have been influenced most by EDA? I would mention computational statistics, data visualization, data mining, and machine learning. Many applied areas rich in data, in particular, those connected to molecular genetics, also have an EDA flavor. Tukey[1] relied almost exclusively on hand calculation and hand-made graphics, and emphasized the importance of this almost intimate manner of performing an analysis of data. But shortly after 1977, the personal computer was invented, which led to the computational revolution we are all familiar with. Today, almost all data analyses are performed with the help of computers. The principles of EDA remain true nonetheless. The most common use of any statistical package ought to be data exploration, and all software systems should be constructed on EDA's triple pillars of flexibility, forgiveness, and ease of computation.

REFERENCES

1. Tukey JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.
2. Tukey JW. *Statistics: Lecture Notes for Course 411*. Department of Statistics, Princeton University, 1980.
3. Tukey JW. The future of data analysis (Corr: V33 P812). *Annal Math Stat* 1962, 33:1–67.
4. McGill R, Tukey JW, Larsen WA. "Variations of Box Plots". *Am Stat* 1978, 32(1):12–16.
5. Kováts E, Kresz R. Wrong gas/liquid partition data by gas chromatography. *J Chromatogr A* 2006, 1113:206–219.
6. Tukey JW. One degree of freedom for non-additivity. *Biometrics* 1949, 5:232–242.
7. Tukey JW. The 3PR Smoother (Oral Communication). Department of Statistics, Princeton University, 1992.

FURTHER READING

Brillinger DR ed. *The Collected Works of John W. Tukey*, Time Series (1949–1964), Vol. I. Wadsworth Publishing Co Inc: Pacific Grove, CA; 1984, 689.
Brillinger DR ed. *The Collected Works of John W. Tukey*, Time Series (1965–1984), Vol. II. Wadsworth Publishing Co Inc: Pacific Grove, CA; 1984, 583.
Cleveland WS ed. *The Collected Works of John W. Tukey*, Graphics (1965–1985), Vol. V. Wadsworth Publishing Co Inc: Pacific Grove, CA; 1988, 528.
Cox, DR ed. *The Collected Works of John W. Tukey*, Factorial and ANOVA (1949–1962), Vol. VII. Wadsworth Publishing Co Inc: Pacific Grove, CA; 1992, 266.
Du Toit SHC, Steyn AGW, Stumpf RH. *Graphical Exploratory Data Analysis*. Springer-Verlag Inc: New York; 1986, 314.
Hoaglin DC, Mosteller F, Tukey JW ed. *Understanding Robust and Exploratory Data Analysis*. Wiley: New York; 1983, 447.
Hoaglin DC, Mosteller F, Tukey JW ed. *Exploring Data Tables, Trends, and Shapes*. Wiley: New York; 1985, 527.
Hoaglin DC, Mosteller F., Tukey JW ed. *Fundamentals of Exploratory Analysis of Variance*. Wiley: New York; 1991, 430.
Jambu M. *Exploratory and Multivariate Data Analysis*. Academic Press: New York; 1991, 474.
Jones LV ed. *The Collected Works of John W. Tukey*, Vols III and IV. Wadsworth Publishing Co Inc: Pacific Grove, CA; 1986, 1122.
Mallows CL ed. *The Collected Works of John W. Tukey*, More Mathematical (1938–1984) Vol VI. Wadsworth Publishing Co Inc: Pacific Grove, CA; 1990, 661.

Velleman PF, Hoaglin DC. *Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press: Pacific Grove, CA; 1981, 354.
Wainer H. The Suspended Rootogram and Other Visual Displays: An Empirical Validation, Vol. 28. *The American Statistician*; 1974, 143–145.
Wainer H. *Graphic Discovery: A Trout in the Milk and other Visual Adventures*, Princeton University Press: Princeton, NJ; 2005, 192.

