Introduction to Exploratory Data Analysis

Ref: NIST/SEMATECH e-Handbook of Statistical Methods

http://www.itl.nist.gov/div898/handbook/index.htm

The original work in Exploratory Data Analysis (EDA) was done by Tukey (1977) and developed since that time by many others. It is more of an attitude rather than a prescribed set of techniques. The attitude is to “play with the data first” and “visualize the data” and then develop models and conclusions.

The analysis is based on graphical and statistical methods. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

One prominent example of this technique, often in the news, is Data Mining—a process to explore large amounts of data (for example by Google, Amazon, NSA) to try to identify patterns and relationships and then use the conclusions to make predictions. For example, several years ago, Netflix offered a one million dollar prize to the team that could significantly improve their ability to predict movies that customers would enjoy. The techniques described below are the very foundation principles on which these advanced analyses are based on.

Goals of EPA

The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. identify important factors
7. determine optimal values for parameter
8. test underlying assumptions

Insight into the Data

Insight implies detecting and uncovering underlying structure in the data. Such underlying structure may not be encapsulated in the list of items above; such items serve as the specific targets of an analysis, but the real insight and "feel" for a data set comes as the analyst judiciously probes and explores the various subtleties of the data. The "feel" for the data comes almost exclusively from the application of various graphical techniques, the collection of which serves as the window into the essence of the data. Graphics are irreplaceable--there are no quantitative analogues that will give the same insight as well-chosen graphics.

To get a "feel" for the data, it is not enough for the analyst to know what is in the data; the analyst also must know what is not in the data, and the only way to do that is to draw on our own human pattern-recognition and comparative abilities in the context of a series of judicious graphical techniques applied to the data.

Difference between Exploratory Data Analysis and Classical Analysis

These three approaches are similar in that they all start with a general science/engineering problem and all yield science/engineering conclusions. The difference is the sequence and focus of the intermediate steps.

Classical: The focus is on the model--estimating parameters of the model and generating predicted values from the model.

Problem => Data => **Model** => Analysis => Conclusions

Exploratory (EDA): the focus is on the data--its structure, outliers, and models suggested by the data.

Problem => **Data** => Analysis => Model => Conclusions

Bayesian:

Problem => Data => Model => Prior Distribution => Analysis => Conclusions

Gallery of Graphical Techniques

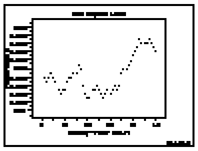
http://www.itl.nist.gov/div898/handbook/eda/section3/eda34.htm

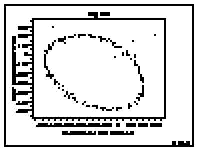
The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

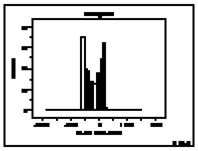
1. Plotting the raw data in an appropriate form. Examples of different plots are shown below.
2. Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data. For large data sets, software programs are used to do these calculations.
3. Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page. These principles described earlier by Tufte are used here.

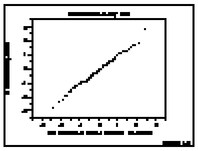
This section provides a gallery of some useful graphical techniques. These are arranged by problem type. The full details can be found using the link above. There are others for more complex problems. (c is a constant; e is the error; t is time).

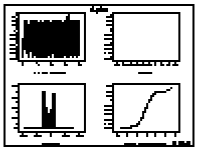
Univariate
y = c + e

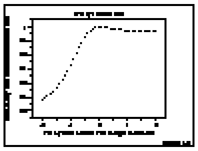

Run Sequence Plot: 1.3.3.25

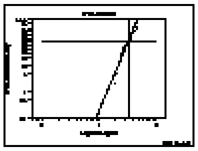

Lag Plot: 1.3.3.15

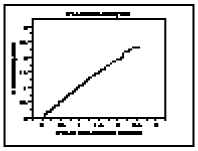

Histogram: 1.3.3.14

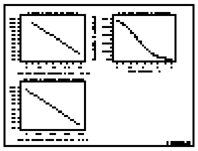

Normal Probability Plot: 1.3.3.21


4-Plot: 1.3.3.32


PPCC Plot: 1.3.3.23



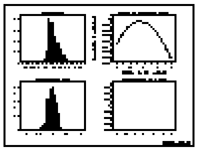




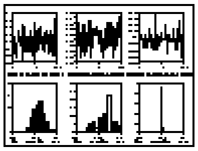
Weibull Plot: 1.3.3.30

Probability Plot: 1.3.3.22

Box-Cox Linearity Plot: 1.3.3.5

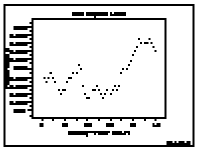


Box-Cox Normality Plot: 1.3.3.6

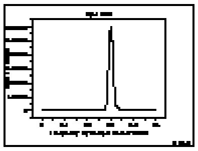


Bootstrap Plot: 1.3.3.4

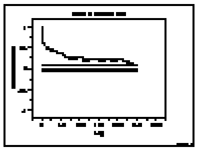
Time Series
y = f(t) + e



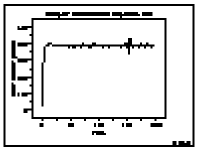
Run Sequence Plot: 1.3.3.25



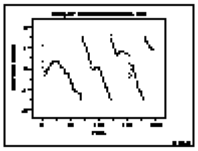
Spectral Plot: 1.3.3.27



Autocorrelation Plot: 1.3.3.1



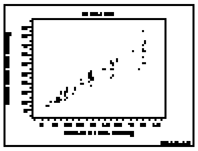
Complex Demodulation Amplitude Plot: 1.3.3.8



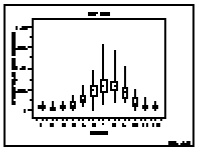
Complex Demodulation Phase Plot: 1.3.3.9

---

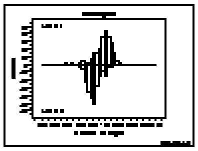
1 Factor
y = f(x) + e



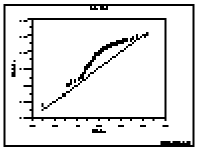
Scatter Plot: 1.3.3.26



Box Plot: 1.3.3.7

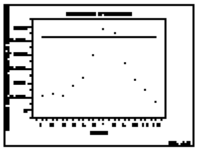


Bihistogram: 1.3.3.2



Quantile-Quantile Plot: 1.3.3.24

Mean Plot: 1.3.3.20



Standard Deviation Plot: 1.3.3.28

Experimental Design and Process Models

In this procedure, one or more process variables (or factors) are changed in the same experiment in order to observe the effect the changes have on one or more response variables. The (statistical) design of experiments (*DOE*) is an efficient procedure for planning experiments so that the data obtained can be analyzed to yield valid and objective conclusions. Well chosen experimental designs maximize the amount of "information" that can be obtained for a given amount of experimental effort.

Process Models

A process model is a mathematical description of a physical process that can predict the outcome based on the individual values of input variable.

It is common to begin with a process model of the `black box' type, with several discrete or continuous input parameters that can be controlled--that is, varied at will by the experimenter--and one or more measured output responses. The output responses are assumed continuous. Experimental data are used to derive an empirical (approximation) model linking the outputs and inputs. These empirical models generally contain first and second-order terms.

Problem for end of section:

An EDA/Graphics Example

Anscombe Example

A simple, classic (Anscombe) example of the central role that graphics play in terms of providing insight into a data set starts with the following data set:

Data

X	Y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

Summary Statistics

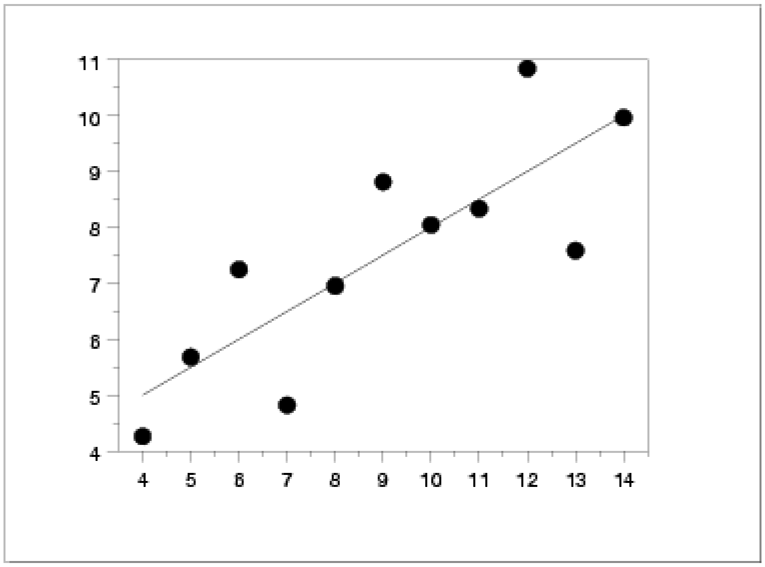
If the goal of the analysis is to compute summary statistics plus determine the best linear fit for Y as a function of X, the results might be given as:

N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.237
Correlation = 0.816

The above quantitative analysis, although valuable, gives us only limited insight into the data.

Scatter Plot

In contrast, the following simple scatter plot of the data



suggests the following:

1. The data set "behaves like" a linear curve with some scatter;
2. there is no justification for a more complicated model (e.g., quadratic);
3. there are no outliers;
4. the vertical spread of the data appears to be of equal height irrespective of the *X*-value; this indicates that the data are equally-precise throughout and so a "regular" (that is, equi-weighted) fit is appropriate.

Three Additional Data Sets

This kind of characterization for the data serves as the core for getting insight/feel for the data. Such insight/feel does not come from the quantitative statistics; on the contrary, calculations of quantitative statistics such as intercept and slope should be subsequent to the characterization and will make sense only if the characterization is true. To illustrate the loss of information that results when the graphics insight step is skipped, consider the following three data sets [Anscombe data sets 2, 3, and 4]:

X2	Y2	X3	Y3	X4	Y4
10.00	9.14	10.00	7.46	8.00	6.58
8.00	8.14	8.00	6.77	8.00	5.76
13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.77	9.00	7.11	8.00	8.84
11.00	9.26	11.00	7.81	8.00	8.47
14.00	8.10	14.00	8.84	8.00	7.04
6.00	6.13	6.00	6.08	8.00	5.25
4.00	3.10	4.00	5.39	19.00	12.50
12.00	9.13	12.00	8.15	8.00	5.56
7.00	7.26	7.00	6.42	8.00	7.91
5.00	4.74	5.00	5.73	8.00	6.89

Quantitative Statistics for Data Set 2

A quantitative analysis on data set 2 yields

N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.237
Correlation = 0.816

which is identical to the analysis for data set 1. One might naively assume that the two data sets are "equivalent" since that is what the statistics tell us; but what do the statistics not tell us?

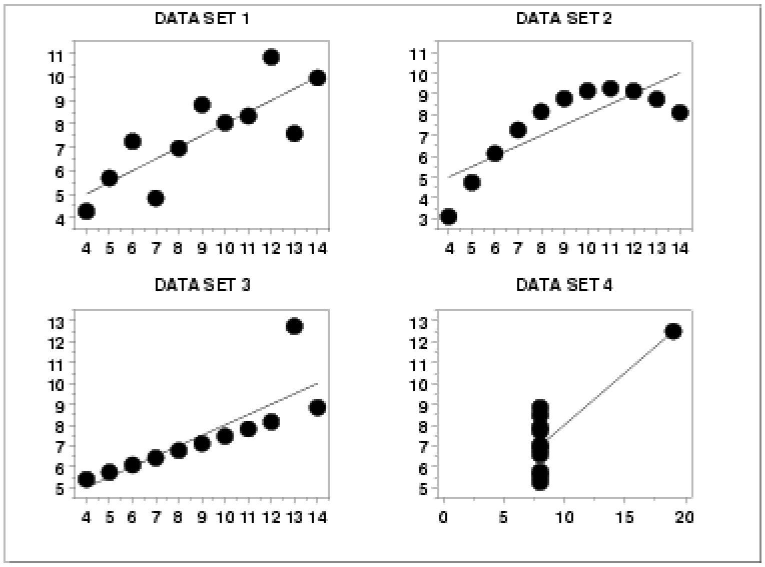
Quantitative Statistics for Data Sets 3 and 4

Remarkably, a quantitative analysis on data sets 3 and 4 also yields

N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.236
Correlation = 0.816 (0.817 for data set 4)

which implies that in some quantitative sense, all four of the data sets are "equivalent". In fact, the four data sets are far from "equivalent" and a scatter plot of each data set, which would be step 1 of any EDA approach, would tell us that immediately.

Scatter Plots



Interpretation of Scatter Plots

Conclusions from the scatter plots are:

1. data set 1 is clearly linear with some scatter.
2. data set 2 is clearly quadratic.
3. data set 3 clearly has an outlier.
4. data set 4 is obviously the victim of a poor experimental design with a single point far removed from the bulk of the data "wagging the dog".

Importance of Exploratory Analysis

These points are exactly the substance that provide and define "insight" and "feel" for a data set. They are the goals and the fruits of an open exploratory data analysis (EDA) approach to the data. Quantitative statistics are not wrong per se, but they are incomplete. They are incomplete because they are numeric **summaries** which in the summarization operation do a good job of focusing on a particular aspect of the data (e.g., location, intercept, slope, degree of relatedness, etc.) by judiciously reducing the data to a few numbers. Doing so also **filters** the data, necessarily omitting and screening out other sometimes crucial information in the focusing operation. Quantitative statistics focus but also filter; and filtering is exactly what makes the quantitative approach incomplete at best and misleading at worst.

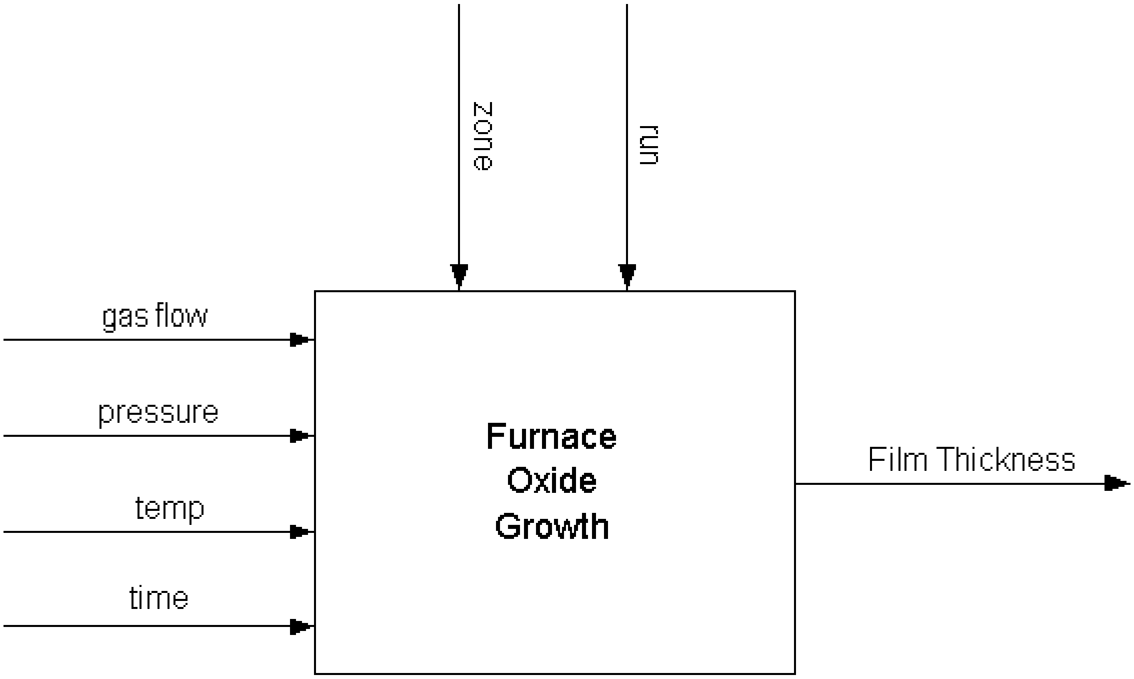
The estimated intercepts (= 3) and slopes (= 0.5) for data sets 2, 3, and 4 are misleading because the estimation is done in the context of an assumed linear model and that linearity assumption is the fatal flaw in this analysis.

The EDA approach of deliberately postponing the model selection until further along in the analysis has many rewards, not the least of which is the ultimate convergence to a much-improved model and the formulation of valid and supportable scientific and engineering conclusions.

Exam Problem Schematic

Process Model

In the picture below we are modeling this process with one output (film thickness) that is influenced by four controlled factors (gas flow, pressure, temperature and time) and two uncontrolled factors (run and zone). The four controlled factors are part of our recipe and will remain constant throughout this study. We know that there is run-to-run variation that is due to many different factors (input material variation, variation in consumables, etc.). We also know that the different zones in the furnace have an effect. A zone is a region of the furnace tube that holds one boat. There are four zones in these tubes. The zones in the middle of the tube grow oxide a little bit differently from the ones on the ends. In fact, there are temperature offsets in the recipe to help minimize this problem.


zone
run
gas flow
pressure
temp
time
Furnace Oxide Growth
Film Thickness