WHAT TO LOOK FOR IN DATA

Some activities are instinctive. A baby doesn't need to be taught how to suckle. Most people can use an escalator, operate an elevator, and open a door instinctively. The same isn't true of playing a guitar, driving a car, or analyzing data.

When faced with a new dataset, the first thing to consider is the objective you, your boss, or your client have in analyzing the dataset.

OBJECTIVE

How you analyze data will depend in part on your objective. Consider these four possibilities, three are comparatively easy and one is a relative challenge.

- **Conduct a Specific Analysis** – Your client only wants you to conduct a specific analysis, perhaps like descriptive statistics or a statistical test between two groups. No problem, just conduct the analysis. There's no need to go further. That's easy.
- **Answer a Specific Question** – Some clients only want one thing — answer a specific question. Maybe it's something like "is my water safe to drink" or "is traffic on my street worse on Wednesdays." This will require more thought and perhaps some experience, but again, you have a specific direction to go in. That makes it easier.
- **Address a General Need** – Projects with general goals often involve model building. You'll have to establish whether they need a single forecast, map or model, or a tool that can be used again in the future. This will require quite a bit of thought and experience but at least you know what you need to do and where you need to end up. Not easy but straightforward.
- **Explore the Unknown** – Every once in a while, a client will have nothing specific in mind, but will want to know whatever can be determined from the dataset. This is a challenge because there's no guidance for where to start or where to finish. This blog will help you address this objective.

If your client is not clear about their objective, start at the very ***end***. Ask what decisions will need to be made based on the results of your analysis. Ask what kind of outputs would be appropriate – a report, an infographic, a spreadsheet file, a presentation, or an application. If they have no expectations, it's time to explore.

GOT DATA?

Scrubbing your data will make you familiar with what you have. That's why it's a good idea to know your objective first. There are many kinds of data errors and many things you can do to scrub your data. But, the first thing you have to do is put it into a matrix. Statistical analyses all begin with matrices. The form of the matrix isn't always the same, but most commonly, the matrix has columns that represent variables (e.g., metrics, measurements) and rows that represent observations (e.g., individuals, students, patients, sample units, or dates). Data on the variables for each observation go into the cells. Usually, this is done with spreadsheet software.

Data scrubbing can be cursory or exhaustive. Assuming the data are already available in electronic form, you'll still have to achieve two goals – getting the numbers right and getting the right numbers.

Getting the numbers right requires correcting three types of data errors:

- Alphanumeric substitution, which involves mixing letters and numbers (e.g., 0 and o or O, 1 and l, 5 and S, 6 and b), dropped or added digits, spelling mistakes in text fields that will be sorted or filtered, and random errors.
- Specification errors involve bad data generation, perhaps attributable to recording mistakes, uncalibrated equipment, lab mistakes, or incorrect sample IDs and aliases.
- Inappropriate Data Formats, such as extra columns and rows, inconsistent use of ND, NA, or NR flags, and the inappropriate presence of 0s versus blanks.

Getting the right numbers requires addressing a variety of data issues:

- **Variables and phenomena**. Are the variables sufficient to explore the phenomena in question?
- **Variable scales**. Review the measurement scales of the variables so you know what analyses might be applicable to the data. Also, look for nominal and ordinal scale variables to consider how you might segment the data.
- **Representative sample**. Considering the population being explored, does the sample appear to be representative?
- **Replicates**. If there are replicate or other quality control samples, they should be removed from the analysis appropriately.
- **Censored data**. If you have censored data (i.e., unquantified data above or below some limit), you can recode the data as some fraction of the limit, but not zero.
- **Missing data**. If you have missing data, they should be recoded as blanks or use another accepted procedure for treating missing data.

Data scrubbing can consume a substantial amount of time, even more than the statistical calculations.

WHAT TO LOOK FOR

If statistics wasn't your major in college or you're straight out of college and new to applied statistics, you might wonder where to start looking at a dataset. Here are five places to consider looking.

- Snapshot
- Population or Sample Characteristics
- Change
- Trends and Patterns
- Anomalies

Start with the entire dataset. Look at the highest levels of grouping variables. Divide and aggregate groupings later after you have a feel for the global situation. The reason for this is that the number of possible combinations of variables and levels of grouping variables can be large, overwhelming, each one being an analysis in itself. Like peeling an onion, explore one layer of data at a time until you get to the core.

SNAPSHOT

What does the data look like at one point? Usually, it's at a point in time but it could also be a common conditions, like after a specific business activity, or at a certain temperature and pressure. What might you do?

Snapshots aren't difficult. You just decide where you want a snapshot and record all the variable values at that point. There are no descriptive statistics, graphs, or tests unless you decide to subdivide the data later. The only challenge is deciding whether taking a snapshot makes any sense for exploring the data.

The only thing you look for in a snapshot is something unexpected or unusual that might direct further analysis. It can also be used as a baseline to evaluate change.

POPULATION CHARACTERISTICS

It's always a good idea to know everything you can about the populations you are exploring. Here's what you might do.

This is a no-brainer; calculate descriptive statistics. Here's a summary of what you might look at. It's based on the measurement scale of the variable you are assessing.

Scale		Characteristic	Statistic
Grouping Scales	Nominal	Frequency	Counts or Percentages
		Central Tendency	
		Dispersion	
		Shape	
Progression Scales	Ordinal	Frequency	Counts, Percentages, Mode
		Central Tendency	Median, Mean
		Dispersion	Range, Quartiles and other Percentages
		Shape	Histogram, Dot plot, Stem-Leaf Diagram
	Ratio	Frequency	Counts, Percentages, Mode
		Central Tendency	Mean, Median
		Dispersion	Standard Deviation, Quartiles and other Percentages
		Shape	Probability Plot, Skewness, Kurtosis

For grouping (nominal scale) variables, look at the frequencies of the groups. You'll want to know if there are enough observations in each group to break them out for further analysis. For progression (continuous) scales, look at the median and the mean. If they're close, the frequency distribution is probably symmetrical. You can confirm this by looking at a histogram or the skewness. If the standard deviation divided by the mean (coefficient of variation) is over 1, the distribution may be lognormal, or at least, asymmetrical. Quartiles and deciles will support this finding. Look at the measures of central tendency and dispersion. If the dispersion is relatively large, statistical testing may be problematical.

Graphs are also a good way to explore population characteristics. Never calculate a statistic without looking at its visual representation in a graph There are many types of graphs that will let you do that.

Common Types of Graphs for General Data Analysis.					
			Data Scales		
Chart	Used to Show	Chart Axes	Horizontal Axis	Vertical Axis	Additional Axes
Box Plot	Distribution	Rectangular	Categorical, continuous (sample size)	Continuous	
Dot Plot	Distribution	Rectangular	Ordinal, continuous	Ordinal	
Histogram	Distribution	Rectangular	Ordinal, continuous	Ordinal	
Probability Plot	Distribution	Rectangular	Ordinal, continuous	Continuous	
Q-Q Plot	Distribution	Rectangular	Ordinal	Ordinal	
Stem-Leaf Diagram	Distribution	Rectangular	Ordinal	Ordinal, continuous	
Ternary Plot	Mixtures	Triangular	Continuous (percentages)	Continuous (percentages)	Continuous (percentages)
Pie Chart	Mixtures	Circular	Categorical	Continuous (percentages)	
Area Chart	Properties	Rectangular	Ordinal, continuous	Continuous	
Bar Chart	Properties	Rectangular	Categorical	Continuous	
Candlestick Chart	Properties	Rectangular	Continuous	Continuous	
Control Chart	Properties	Rectangular	Continuous	Continuous	
Deviation Plot	Properties	Rectangular	Continuous	Continuous	
Line Chart	Properties	Rectangular	Categorical, ordinal	Continuous	
Map	Properties	Rectangular	Continuous	Continuous	Any
Matrix Plot	Properties	Rectangular	Nominal	Nominal	Text
Means Plot	Properties	Rectangular	Continuous	Continuous	
Spread Plot	Properties	Rectangular	Continuous	Continuous	
Block Diagram	Properties	Cubic	Nominal	Nominal	Nominal
Rose Diagram	Properties	Circular	Ordinal, continuous	Continuous	
Multivariable Plot	Relationships	Rectangular, circular, other	Any	Continuous	Continuous
Bubble Plot	Relationships	Rectangular	Continuous	Continuous	Continuous
Contour Plot	Relationships	Rectangular	Continuous	Continuous	Continuous
Icon Plot	Relationships	Rectangular	Continuous	Continuous	Multivariable plot*
Scatter Plot: 2D	Relationships	Rectangular	Continuous	Continuous	
Scatter Plot: 3D	Relationships	Cubic	Continuous	Continuous	Continuous
Surface Plot	Relationships	Cubic	Continuous	Continuous	Continuous

What you look for in a graph depends on what the graph is supposed to show – distribution, mixtures, properties, or relationships. There are other things you might look for but here are a few things to start with.

For ***distribution*** graphs (box plots, histograms, dot plots, stem-leaf diagrams, Q-Q plots, rose diagrams, and probability plots), look for ***symmetry***. That will separate many theoretical distributions, say a normal distribution (symmetrical) from a lognormal distribution (asymmetrical). This will be useful information if you do any statistical testing later.

For ***mixture*** graphs (pie charts, rose diagrams, and ternary plots), look for ***imbalance***. If you have some segments that are very large and others very small, here may be common and unique themes to the mix that you can explore. Maybe the unique segments can be combined. This will be useful information if you do break out subgroups later.

For ***properties*** graphs (bar charts, area charts, line charts, candlestick charts, control charts, means plots, deviation plots, spread plots, matrix plots, maps, block diagrams, and rose diagrams), look for the ***unexpected***. Are the central tendency and dispersion what you might expect? Where are big deviations?

For ***relationship*** graphs (icon plots, 2D scatter plots, contour plots, bubble plots, 3D scatter plots, surface plots, and multivariable plots), look for ***linearity***.

You might find linear or curvilinear trends, repeating cycles, one-time shifts, continuing steps, periodic shocks, or just random points. This is the prelude for looking for more detailed patterns.

CHANGE

Change usually refers to differences between two time periods but, like snapshots, it could also refer to some common conditions. Change can be difficult, or at least complicated, to analyze because you must first calculate the changes you want to explore. When calculating changes, be sure the intervals of the change are consistent. But after that, what might you do?

First, look for very large, negative or positive changes. Are the percentages of change consistent for all variables? What might be some reasons for the changes.

Calculate the mean and median changes. If the indicators of central tendency are not near zero, you might have a trend. Verify the possibility by plotting the change data. You might even consider conducting a statistical test to confirm that the change is different from zero.

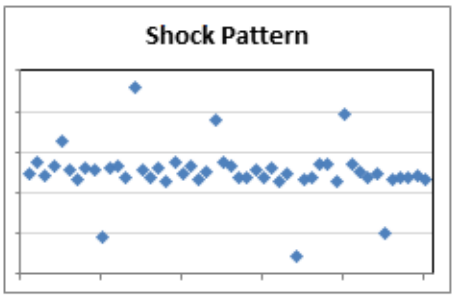
If you do think you have a trend or pattern, there are quite a few things to look for.

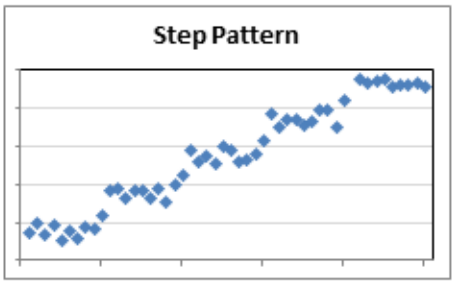
TRENDS AND PATTERNS

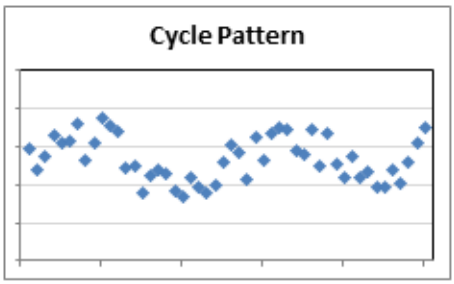
There are at least ten types of data relationships – direct, feedback, common, mediated, stimulated, suppressed, inverse, threshold, and complex – and of course spurious relationships. They can all produce different patterns and trends, or no recognizable arrangement at all.

There are four patterns to look for:

- Shocks
- Steps
- Shifts
- Cycles.







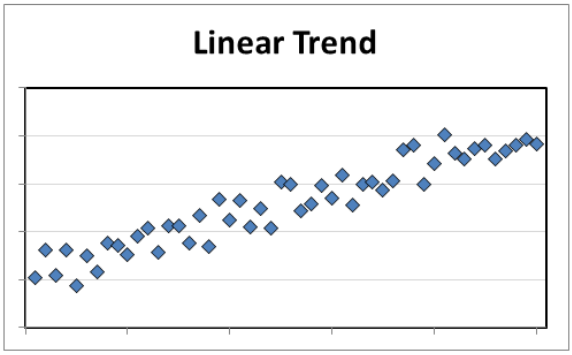
Shocks are seemingly random excursions far from the main body of data. They are outliers but they often reoccur, sometimes in a similar way suggesting a common, though sporadic cause. Some shocks may be attributed to an intermittent malfunction in the measurement instrument. Sometimes they occur in pairs, one in the positive direction and another of similar size in the negative direction. This often appears in business data when reporting cutoff dates are missed.

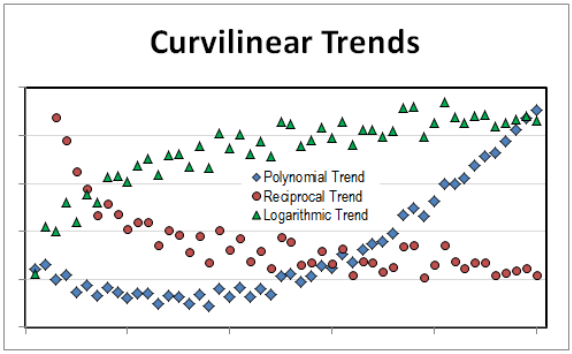
Steps are periodic increases **or** decreases in the body of the data. Steps progress in the same direction because they reflect a progressive change in conditions. If the steps are small enough, they can appear to be, and be analyzed as, a linear trend.

Shifts are increases **and/or** decreases in the body of the data like steps, but shifts tend to be longer than steps and don't necessarily progress in the same direction. Shifts reflect occasional changes in conditions. The changes may remain or revert to the previous conditions, making them more difficult to analyze with linear models.

Cycles are increases **and** decreases in the body of the data that usually appear as a waveform having fairly consistent amplitudes and frequencies. Cycles reflect periodic changes in conditions, often associated with time, such as daily or seasonal cycles. Cycles cannot be analyzed effectively with linear models. Sometimes different cycles add together making them more difficult to recognize and analyze.

Simple trends can be easier to identify because they are more familiar to most data analysts. Again, **graphs** are the best place to look for trends.

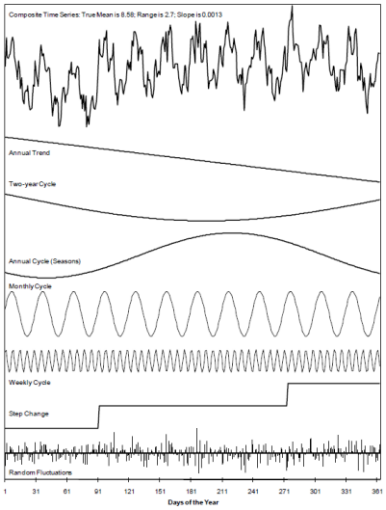




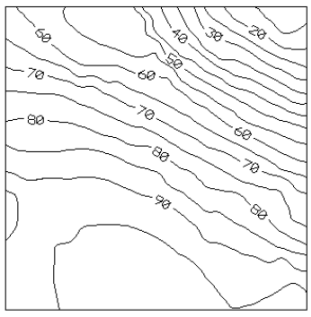
Linear trends are easy to see; the data form a line. ***Curvilinear*** trends can be more difficult to recognize because their path is more complex. With some experience and intuition, however, they can be identified. ***Nonlinear*** trends can look similar to curvilinear trends or be far more complex. They require more complicated nonlinear models to analyze. Curvilinear trends can be analyzed with linear models with the use of transformations.

There are also more complex trends involving different dimensions, including:

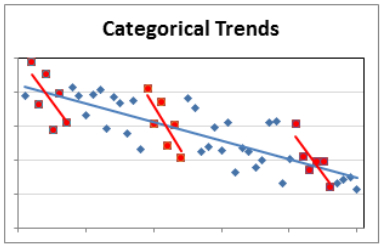
- Temporal
- Spatial
- Categorical
- Hidden
- Multivariate

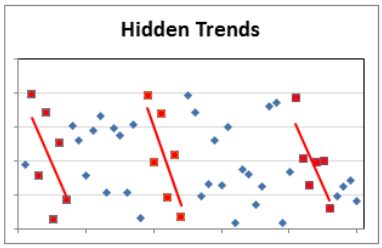


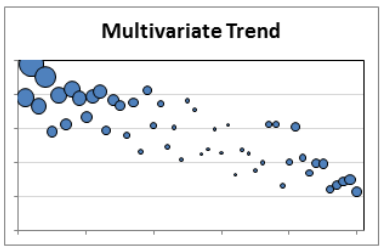
Temporal Trends can be more difficult to identify because Time-series data can be combinations of shocks, steps, shifts, cycles, and linear and curvilinear trends. The effects may be seasonal, superimposed on each other within a given time period, or spread over many different time periods. Confounded effects are often impossible to separate, especially if the data record is short or the sampled intervals are irregular or too large.



Spatial Trends present a different twist. Time is one-dimensional; at least as we now know it. Distance can be one-, two-, or three-dimensional. Distance can be in a straight line ("as the crow flies") or along a path (such as driving distance). Defining the location of a unique point on a two-dimensional surface (i.e., a plane) requires at least two variables. The variables can represent coordinates (northing/easting, latitude/longitude) or distance and direction from a fixed starting point. At least three variables are needed to define a unique point location in a three-dimensional volume, so a variable for depth (or height) must be







added to the location coordinates. Looking for spatial patterns involves interpolation of geographic data using one of several available algorithms, like moving averages, inverse distances, or geostatistics.

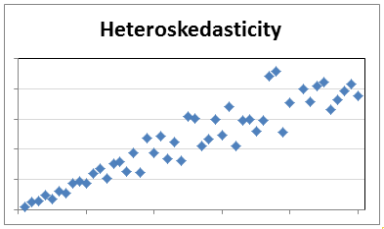
Categorical Trends are no more difficult to identify than any trend except you have to break out categories to do it, which can require a lot of data and be a lot of work. One thing you might see when analyzing categories is Simpson's paradox. The paradox occurs when trends appear in categories that are different from the overall group. ***Hidden Trends*** are trends that appear only in categories and not the overall group. You may be able to detect linear trends in categories without graphs if you have enough data in the categories to calculate correlation coefficients within each.

Multivariate Trends add a layer of complexity to most trends, which are bivariate. Still, you look for the same things, patterns and trends, only you have to examine at least one additional dimension. The extra dimension may be an additional axis or some other way of representing data, like icon type, size, or color.

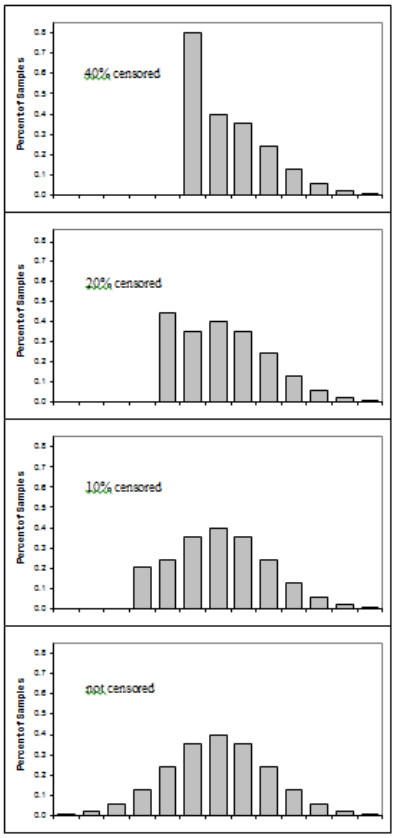
ANOMALIES

Sometimes the most interesting revelations you can garner from a dataset are the ways that it *doesn't* fit expectations. Three things to look for are"

- Censoring
- Heteroskedasticity
- Outliers.

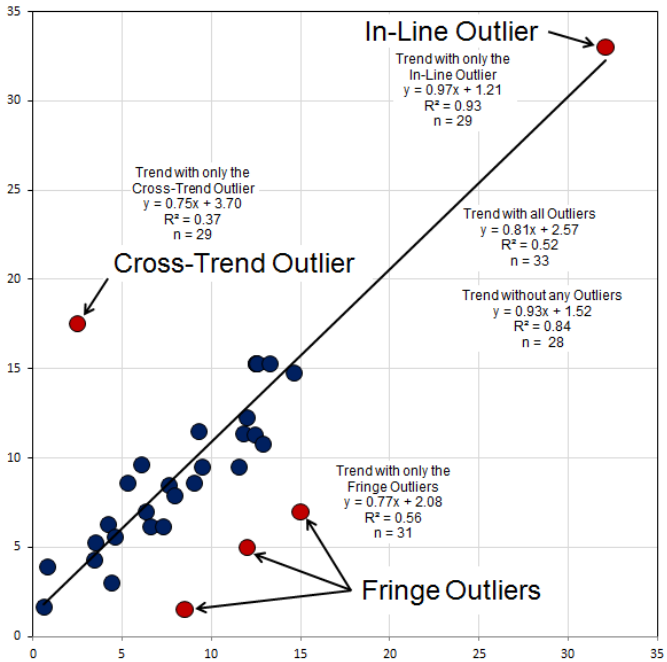


Censoring is when a measurement is recorded as <value or >value, indicating that the measurement instrument was unable to quantify the real value. For example, the real value may be outside the range of a meter, counts can't be approximated because there are too many or too few, or a time can only be estimated as before or after. Censoring is easy to detect in a dataset because they should be qualified with < or >.



Heteroskedasticity is when the variability in a variable is **not** uniform across its range. This is important because homoscedasticity (the opposite of heteroskedasicity) is assumed by probability statements in parametric statistics. Look for differing thicknesses in plotted data.

Influential observations and outliers are the data points that don't fit the overall trends and patterns. Finding anomalies isn't that difficult; deciding why they are anomalous and what to do with them are the really tough parts. Here are some examples of the types of outliers to look for.



How and Where to Look

That's a lot of information to take in and remember, so here's a summary you can refer to in the future if you ever need it.

And when you're done, be sure to document your results so others can follow what you did.

			Where to Look				
			Individual Values	Descriptive Statistics	Graphs	Tests	Procedures
What to Look for	Snapshot		All data defining the snapshot				
	Changes		Differences	Appropriate statistics defining central tendency	Bivariate plot		
	Population Characteristics			Appropriate statistics defining frequency, central tendency, dispersion, and shape for measurement scale	Graphs to show distribution, mixtures, and properties	Hests and ANOVA	
	Anomalies	Censoring	< or > data qualifiers		Histogram and other frequency distribution graphs		Replacement or imputation
		Hetero-skedasicity		Appropriate statistics defining dispersion	Bivariate plot	Levene's test	Trans-formations
		Outliers		Appropriate statistics defining central tendency and dispersion	Distribution and bivariate plot	Outlier tests	Exclusion, Replacement or inclusion
	Patterns	Shocks		Appropriate statistics defining central tendency and dispersion	Control charts and other bivariate plots	Outlier tests	Regression
		Steps			Control charts and other bivariate plots		Regression
		Shifts			Control charts and other bivariate plots		Regression
		Cycles			Control charts and other bivariate plots		Trans-formations
	Trends	Linear			Bivariate plot		Correlation and regression
		Curvilinear			Bivariate plot	Trans-formations	Correlation and regression
		Temporal			Time-series plot, correlograms	Trans-formations	Regression, ARIMA
		Spatial			Variograms, contour map		Geostatistics
		Categorical			Bivariate plot		Within group correlations
		Hidden			Bivariate plot		Within group correlations
		Multivariate			Bubble plot, 3-D plot, ternary plot		Multiple correlation and regression

Stats with Cats

January 2019