TDWI Data Visualization Principles and Practices

Previews of TDWI course books offer an opportunity to see the quality of our material and help you to select the courses that best fit your needs. The previews cannot be printed.

TDWI strives to provide course books that are content-rich and that serve as useful reference documents after a class has ended.

This preview shows selected pages that are representative of the entire course book; pages are not consecutive. The page numbers shown at the bottom of each page indicate their actual position in the course book. All table-of-contents pages are included to illustrate all of the topics covered by the course.

COURSE OBJECTIVES

You Will Learn:

- ***Visualization as a communication medium***
- ***Preparing data for visualization***
- ***Components of visualization***
- ***Choosing and using charts and graphs***
- ***Visual exploration and analysis***
- ***Visual design techniques***
- ***Extending visualization with infographics***
- ***Visual storytelling***
- ***Data visualization tools***

TDWI takes pride in the educational soundness and technical accuracy of all of our courses. Please send us your comments—we'd like to hear from you. Address your feedback to:

info@tdwi.org

Publication Date: December 2017

TABLE OF CONTENTS

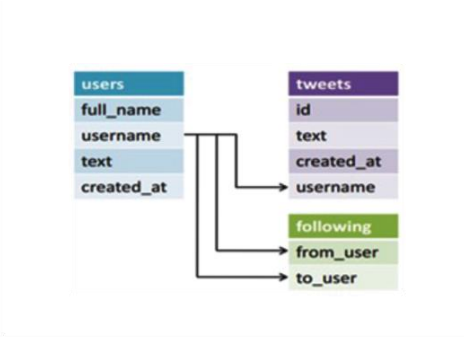
Transforming Data With Intelligence™

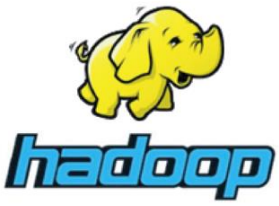
Module 1

Data Visualization Concepts

Data and Visualization

Finding the Right Data





Data and Visualization

Finding the Right Data

MATCHING DATA TO PURPOSE

Before data can be visualized, it must be obtained. Data is selected for visualization based on a business problem or question of interest.

DATA REPOSITORIES

The most obvious sources of data are enterprise repositories—for example, a data warehouse, data mart, or a data lake. These repositories are the primary source of enterprise data, and will be a frequent focus for data visualizations.

Data may be queried or extracted for these sources using a variety of techniques, and filtered based on business-driven criteria such as time or subject.

OTHER SOURCES

Data of interest may also be housed outside an organization—often referred to as external data. External data repositories may include:

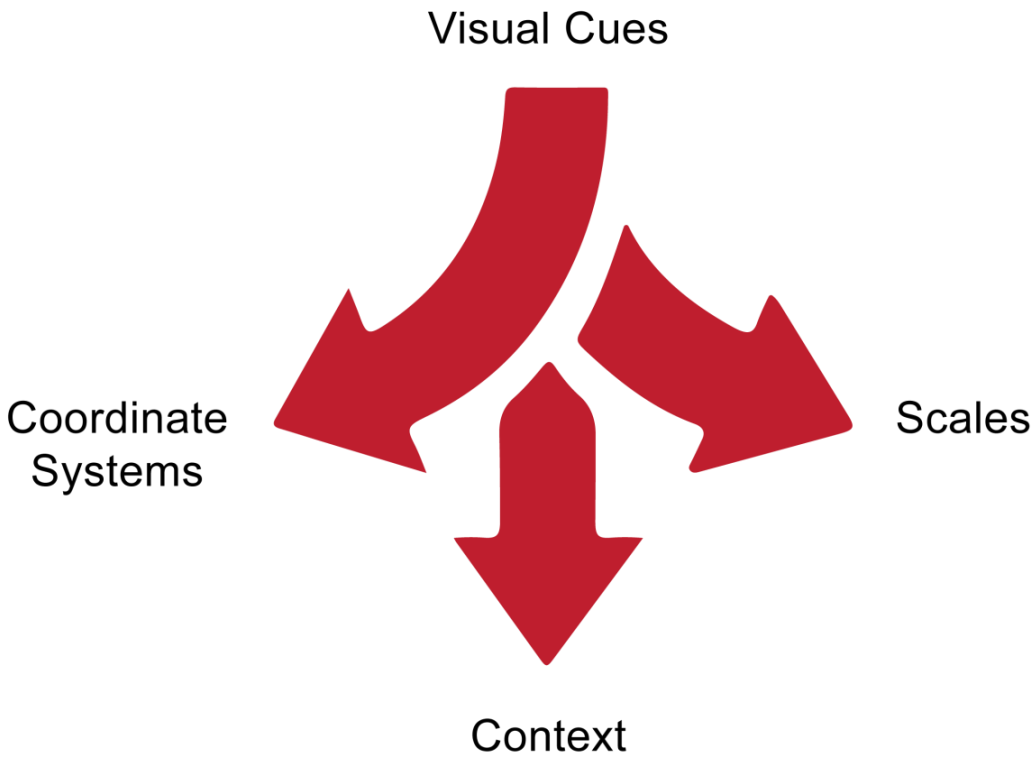
- University data
- General data applications
- Topical data (geographical, sports, census)
- Websites (data scraping)

DATA HANDLING

Acquisition of data may require complex processing. Examples include the integration of data from multiple sources, the transformation of data based on business rules and analytics objectives, or the cleansing of data based on quality standards. A variety of interfaces, formats, and processing activities may be required. These data management activities are beyond the scope of this course, but it is important to recognize that they are often required.

Data Visualization Components

The Parts of Data Visuals



Data Visualization Components

The Parts of Data Visuals

VISUAL LANGUAGE COMPONENTS

Recall the earlier statement that data visualization is a language with components and rules. The components of visual communication include visual cues, coordinate systems, and scales. These are the building blocks with which we create charts and graphs.

VISUAL CUES

Visual cues are used to draw the eye of the viewer to specific parts of or relationships in a graph. Visual cues include:

- *Placement*: Where we locate things on a graph with attention to both position and proximity
- *Lines*: Length, angle, and direction of lines in a graph all directly influence visual perception
- *Shapes*: Area and volume of shapes in a graph (circles, squares, etc.) communicate the relative size of the things that they represent
- *Color*: Both hue (red, green, blue...) and saturation (intensity) influence the perceived importance of objects in a graph

COORDINATE SYSTEMS

Common coordinate systems are the means by which quantitative values are plotted to specific locations in a graph. They include:

- *Cartesian systems* place data points on a two-dimensional grid that is defined by an x-axis and a y-axis. Scatter plots and line graphs are examples of graphs using Cartesian coordinates.
- *Polar systems* place data points based on a circular system. The center of the circle is the base point or zero point. The value and characteristics of a data point determine placement as distance from the center point on the radius of the circle and angle from zero to 360 degrees around the circumference of the circle. Radar graphs are a common example of polar coordinates.
- *Geographic systems* place data points on a map. Data point values may be placed according to a coordinate system based on latitude and longitude. Placement may also include altitude for three-dimensional plots. Heat maps are sometimes an example of using a geographic system.

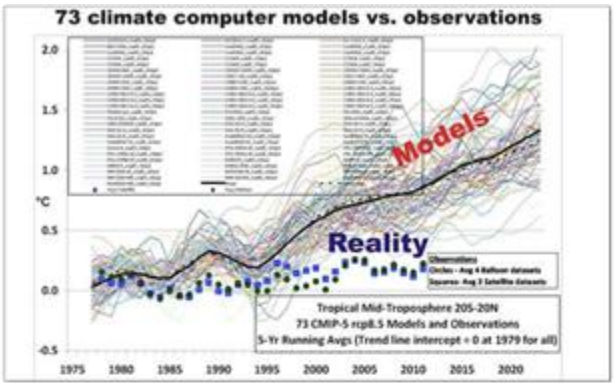
Transforming Data With Intelligence™

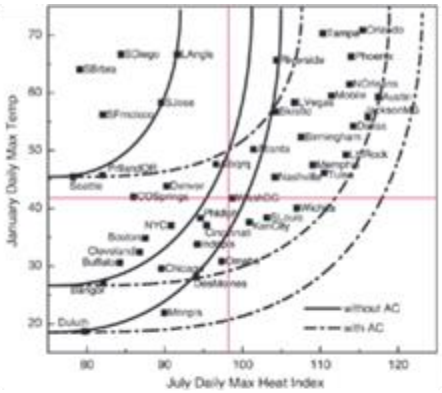
Module 2

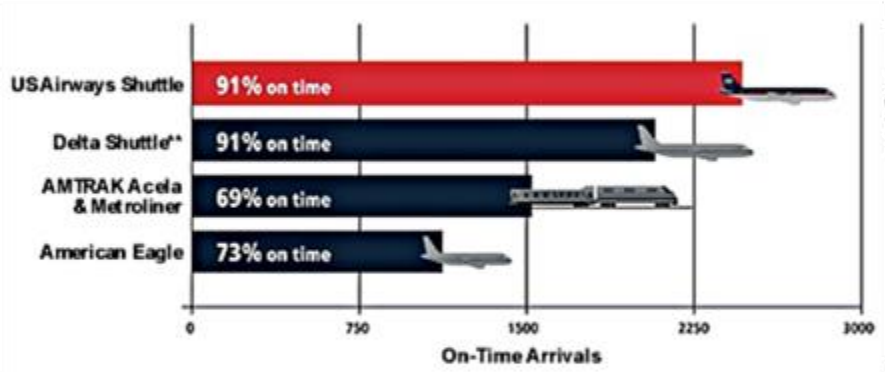
Fundamentals of Visualization

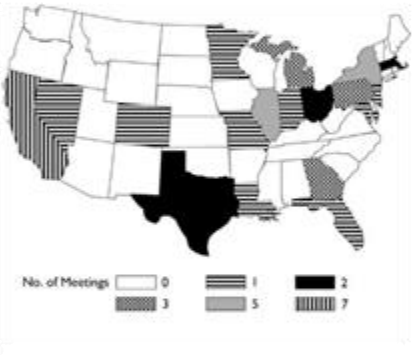
Data Visualization Standards

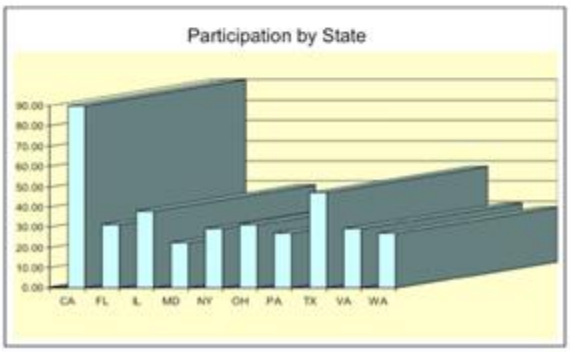
Good Design

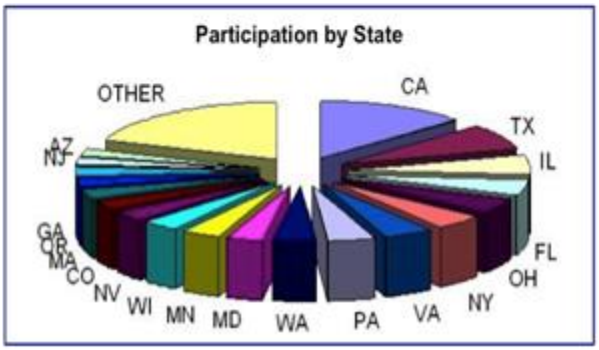












Data Visualization Standards

Good Design

COMMUNICATING WITH VISUALS

The goal of data visualization is good communication. The opportunity to miscommunicate is always present. To avoid miscommunication, pay careful attention to design to be sure that your visuals are readable, clear, and unambiguous.

READABLE

The two examples at the top of the facing page illustrate graphs that are difficult to read.

The climate model is cluttered and confusing with too many lines to follow and a complex legend that consumes more than one-fourth of the visualization space.

The heat index plot is confusing—with lines, data points, and text colliding. The meaning of the lines is uncertain, and the visual uses city abbreviations that are not standard and are subject to interpretation. Both of these images have high potential to confuse and miscommunicate.

CLEAR

The center of the facing page shows two examples of communication that are unclear.

In the on-time arrivals graph, both U.S. Airways Shuttle and Delta Shuttle show 91 percent on-time arrivals but the bars are of different lengths. It takes some study to understand that the bar length represents the total number of arrivals—or is it the total number of on-time arrivals? Next it is uncertain whether the number of arrivals corresponds with the end of the bar or the end of the airplane and train images. Does Delta Shuttle have approximately 2,000 arrivals or 2,500?

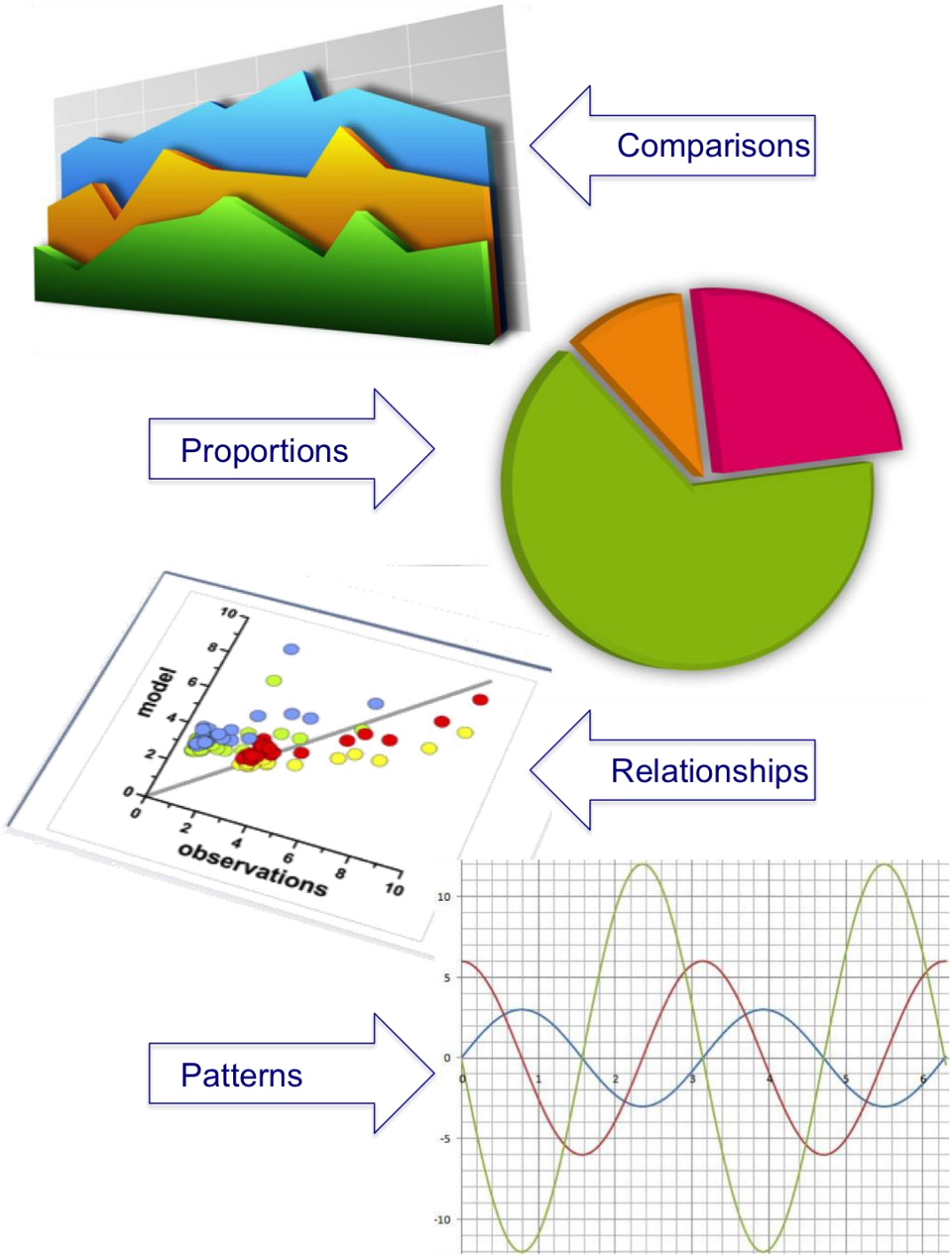
The choropleth map communicates badly because the pattern and shading choices are poor. A quick look suggests that Texas has more of whatever is being measured than California. Only by studying the legend do you see that the Texas shading means two and the California pattern means seven.

UNAMBIGUOUS

The two images at the bottom of the page are ambiguous for similar reasons. Both use 3-D charts to show two-dimensional data—a technique that inhibits rather than advances communication. In the column graph it is difficult to match the top of a column to the y-axis scale, so precision is lost. In the pie chart the relative size of the slices is even more difficult to compare than with a simple two-dimensional wedge. On the left side of the chart it is difficult to match the state abbreviations with the slices they are intended to label.

Visualization with Purpose

What Do You Want to Show?



Visualization with Purpose

What Do You Want to Show?

WHAT DO YOU WANT TO SHOW?

Good visual design begins by knowing what you want to communicate—most commonly comparisons, proportions, relationships, or patterns.

- Comparisons illustrate *what differences* are between things.
- Proportions show *what parts* comprise a whole.
- Relationships show *what dependencies* exist among subjects or variables.
- Distribution patterns show *at what frequencies* data values or conditions occur.
- Location patterns show *what places* are associated with specific data values or conditions.
- Probability patterns show *what the chances* are that a specific value or condition will occur.
- Behavior over time patterns show *what trends* are observed in the data.

Knowing what you want to show is a fundamental first step to choosing the right graphing and visualization techniques to communicate well.

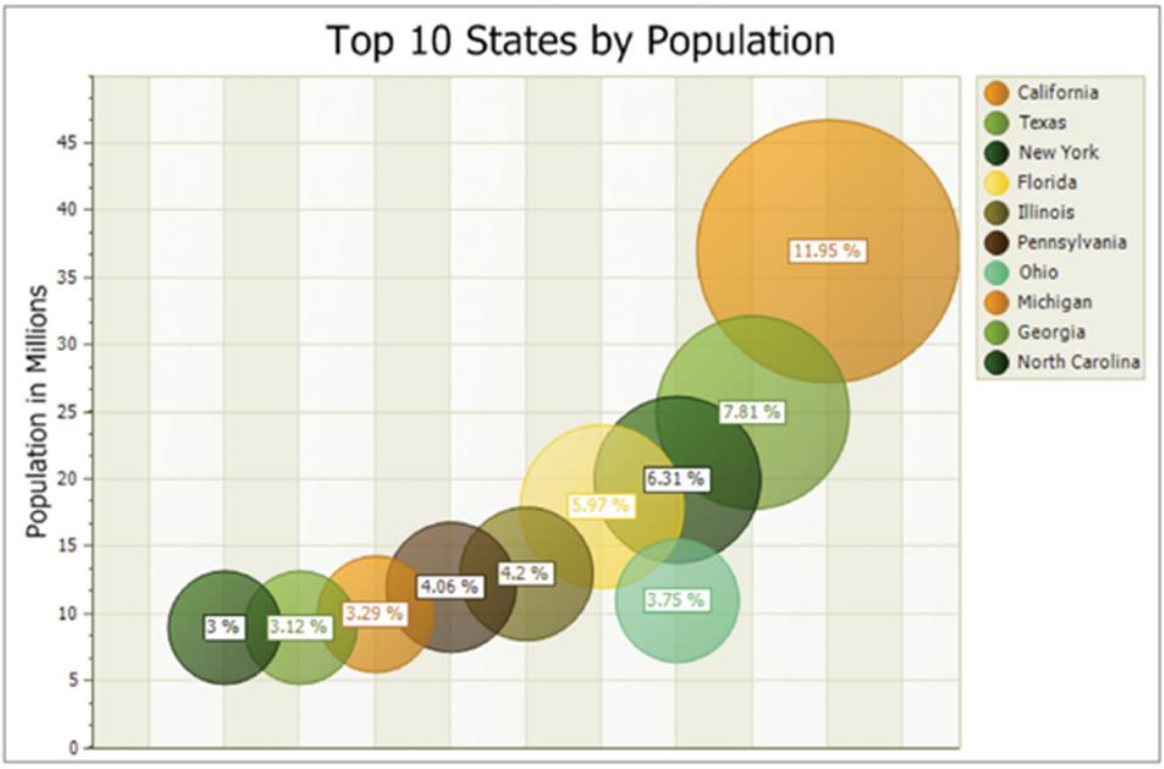
Transforming Data With Intelligence™

Module 3

Data Visualization Techniques

Visualizing Proportions

Proportion Data



Visualizing Proportions

Proportion Data

PARTS TO A WHOLE

Visualizing proportion data focuses on using size or area to show differences or similarities between values or parts to a whole. For proportion analysis we often look for the maximum, minimum, or the overall distribution. Looking at maximum and minimum is simple—sort your data by a particular variable and pick the ends for your maximum and minimum. Distribution looks at frequency of values and also the shape of data.

DATA CHARACTERISTICS

Proportion data typically uses aggregated data, thus some transformation and cleansing may be required to structure the data to be able to analyze value differences or parts to a whole.

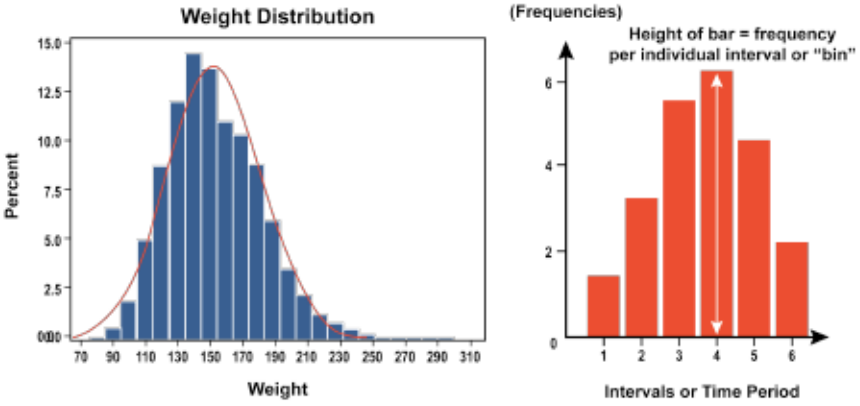
To analyze parts to a whole, the data will need to be aggregated by a category that is a subcategory to the whole. For instance, summarizing sales by store for a region would enable analysis of store sales within a region. Store sales are the parts, and the whole is the region's total sales.

Analyzing differences or similarities between values may require aggregation of single variables for comparison. A common way to analyze differences or similarities between values is a bubble chart. The facing page shows the top 10 states by population where each bubble reflects the size of population in percent via the size of the bubble.

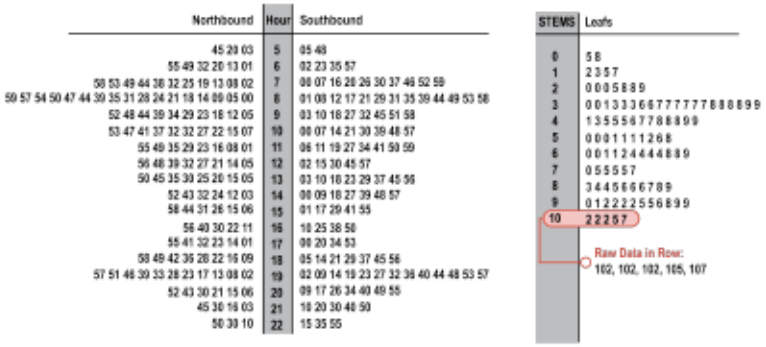
Visualizing Relationships

Distribution

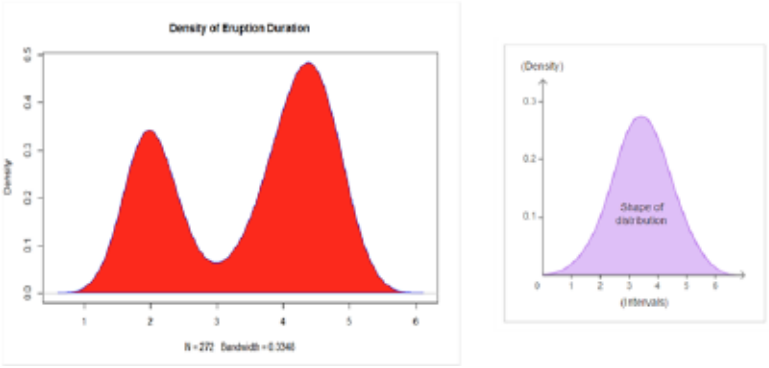
Histogram



Stem and Leaf Plot



Density Plot



Visualizing Relationships

Distribution

DEFINITION

Distribution shows the relationship between the frequency of occurrence of each value of a variable and the total set of values, as well as the mean, median, minimum, and maximum values. Many visualization methods display frequency—how data is spread out over an interval or grouped. We introduced one earlier—the box plot—to demonstrate outliers. There are three main visualization methods that show distribution.

HISTOGRAM

A *histogram* visualizes the distribution of data over a continuous interval or certain time period. Each bar in a histogram represents the tabulated frequency for that interval/bin. The total area of the histogram is equal to the number of data points. Histograms help us estimate where values are concentrated, what the extremes are, and whether there are any gaps or unusual values. They are also useful for giving a rough view of probability distribution.

STEM AND LEAF

Stem and leaf plots organize data by place value to show the distribution of data. Place values are shown ascending downwards on a "stem" column, typically but not always in tens. Data that is within each place value is listed and extends sideways from it as a "leaf."

DENSITY PLOT

A *density plot* depicts the distribution of data over a continuous period. A density plot is a variation of a histogram that uses kernel smoothing to plot values, resulting in a smoother distribution. An advantage that density plots have over histograms is that they show the distribution shape more clearly because they are not affected by the number bars used in a histogram.

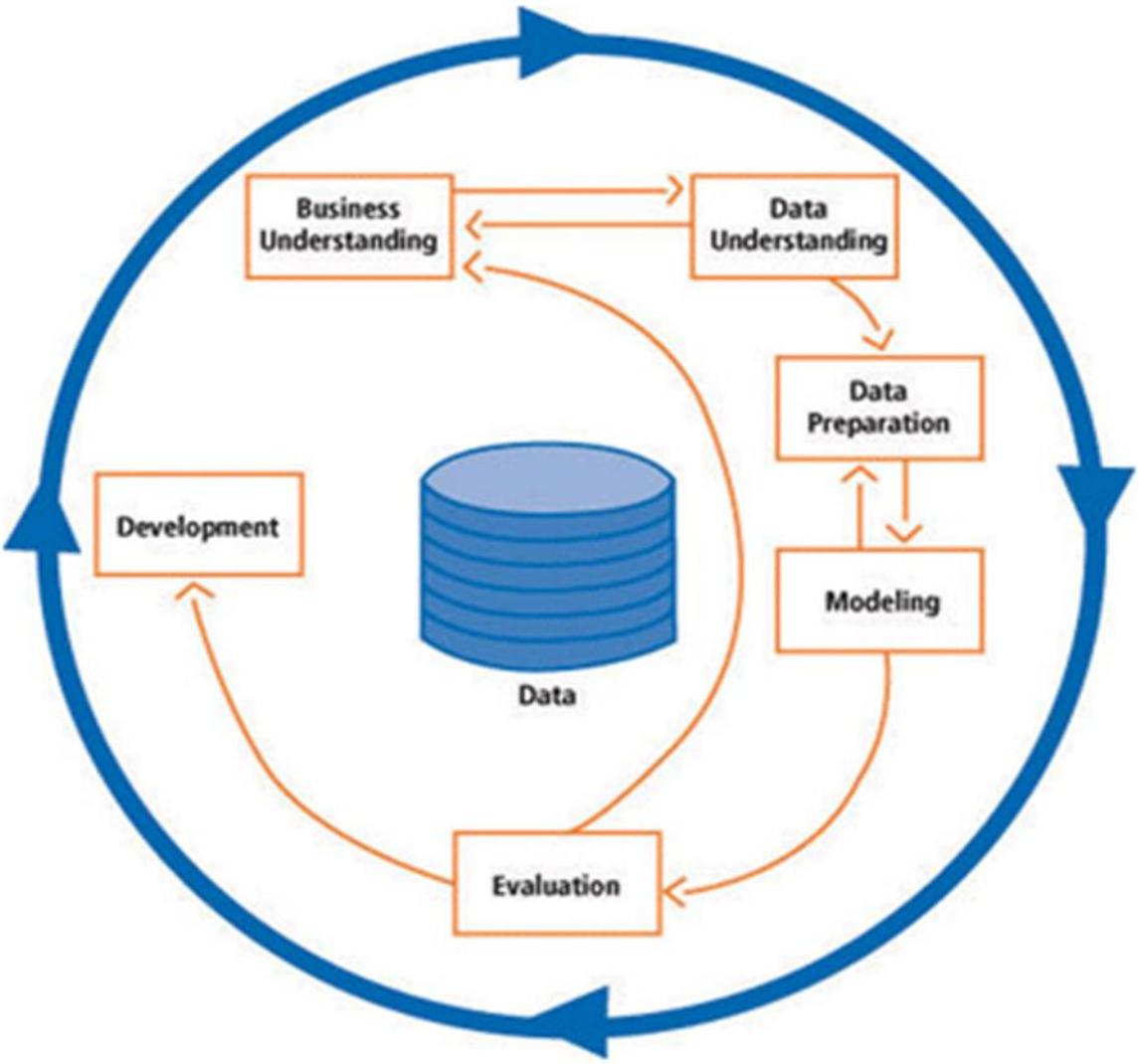
Transforming Data With Intelligence™

Module 4

Visualization and Business Intelligence

Analytics

From Exploration to Models



Analytics

From Exploration to Models

FROM EXPLORATION TO ANALYSIS

Visualization is a valuable part of analytics and it is often used by data scientists as part of the analytical process. The Cross Industry Standard Process for Data Mining (CRISP-DM) is the primary approach used by data scientists and reviewing it illustrates how data visualization assists the process:

Business Understanding—The first phase focuses on understanding the project objectives and requirements and converting this knowledge into an analytics problem definition. Visualization helps explore data used in problem definition.

Data Understanding—The data understanding phase starts with the initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. Data profiling and visualizing data via frequency charts, histograms, and box plots assist with data understanding.

Data Preparation—The data preparation phase covers all activities to construct the final data set from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Visualizing data via frequency charts, histograms, and box plots assists with data validation.

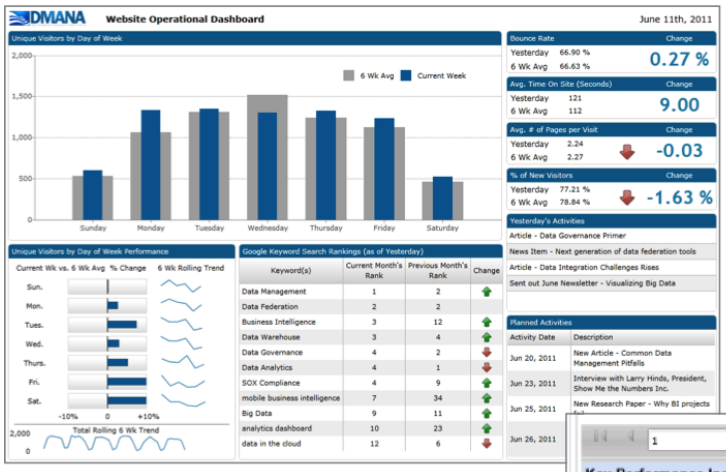
Modeling—In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Visualizing data patterns assists in comparing modeling techniques and results.

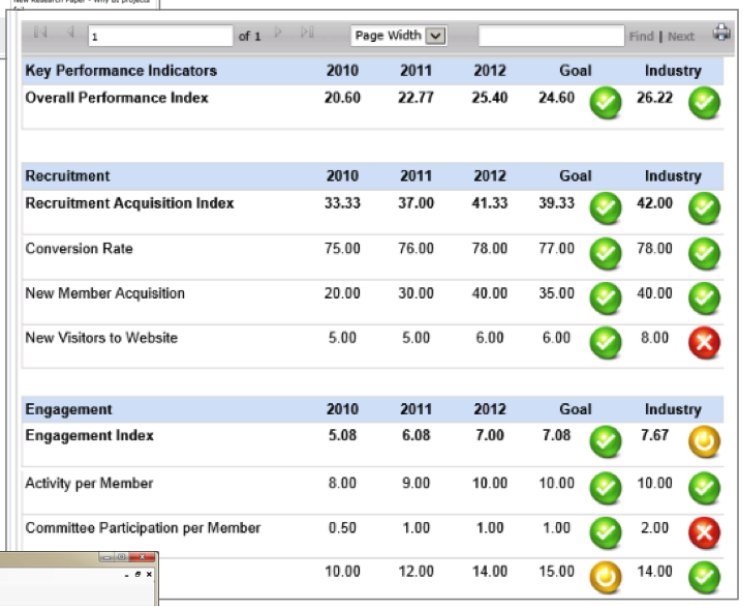
Evaluation—At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Visualizing the final model results supports explanation to multiple stakeholders.

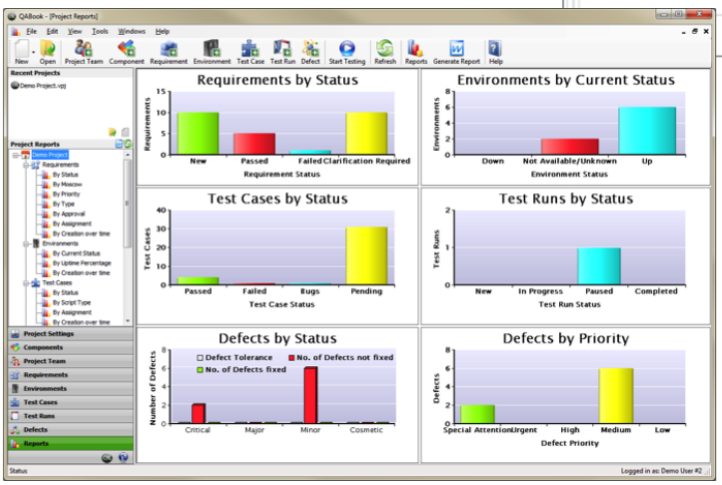
Deployment—Models are deployed into production, monitored, and revised as needed. Using visualization to track model performance via trends and time series is typical after a model is deployed.

Visual Reporting

Dashboards, Scorecards, and Reports







Visual Reporting

Dashboards, Scorecards, and Reports

PUBLISHING VISUALIZATIONS

Many data visualizations are published on a periodic basis as content in reports, scorecards, and dashboards. For each of these uses the designer must consider placement, proximity, and other visual cues for a visual space that includes multiple charts, graphs, tables, etc.

Any type of graph, chart, or table may be used in reports, scorecards, and dashboards. Common practices include:

- Reports using the most familiar types of graphs—line, bar, column, and pie charts—that are easily understood by a diverse group of people.
- Scorecards using tables with embedded graphics, including specialty graphs that we'll discuss in a moment.
- Dashboards using gauges, bar graphs, column graphs, and specialty graphs. Although gauges are common, they are not necessarily good visualization practice. Most gauges are in the form of radial bar graphs. Rectangular bars convey the same information in less space and in a way that is easier for visual comparison.

Some good practices when organizing multiple graphs for reporting include:

- Reports should be organized with careful attention to sequence. Putting summary graphs ahead of details gives context for more detailed information. Grouping graphs—for example, all of the revenue graphs together, then all of the expense graphs—helps with continuity and comparisons.
- Scorecards should have the most important performance indicators at the top. It is also helpful to group metrics with lagging indicators immediately followed by contributing leading indicators.
- Dashboards should have a limited number of graphs and other visuals. Seven or fewer graphical objects is a good rule of thumb.

