DATA VISUALIZATION - WEEK 2

VISUALIZATION DESIGN, DATA TYPES, CHART TYPES, MAKE 10 DIFFERENT CHARTS WITH THE SALES ORDER DATASET USING TABLEAU.

Field Types

When you connect to a new data source, Tableau assigns each field in the data source to either the Dimensions area or the Measures area of the Data pane, depending on the type of data the field contains. We use **Dimensions** to represent **categorical data** and **Measurers** to represent **quantitative data** in tableau and will be using the terms interchangeably during the course.

- If a field contains values that are names, dates, or geographical locations—anything other than numbers—Tableau assigns that field to the **Dimensions** area of the **Data** pane when you first connect to a data source. Tableau treats the values as **discrete**.
- If a field contains numbers, Tableau assigns it to the Measures section.
- If a field has values that are numbers that can be added, averaged, or otherwise aggregated, Tableau assigns that field to the **Measures** area of the **Data** pane when you first connect to a data source. Tableau is assuming that the values are **continuous**.
- If a field is continuous, the background color is **green**. If it is discrete, the background color is **blue**. Background color does not indicate dimension vs. measure—it indicates continuous vs. discrete. By default, dimensions are discrete and measures are continuous, but in fact all four combinations are possible.

Data Types

As a data analysis tool, Tableau classifies every piece of data into one of the four categories namely - String, Number, Boolean date, and datetime. Once data is loaded from source, tableau automatically assigns the data types, but you can also change some of the data types if it satisfies the data conversion rule. Also the user has to specify the data type for calculated fields.

Icon	Data Type	Example
Abc	Text (String) values	Name, address, order date
#	Numerical values	5, 10.33
📅	Date values	05/05/2015
📅🕒	Date & time values	04/04/2014 01:02:03
🌐	Geographic values	Map
T|F	Boolean values	True or false

There are four measurement scales (or types of data): categorical (nominal, ordinal), interval and ratio.

These are simply ways to categorize different types of variables.

- A categorical variable (sometimes called a nominal variable) is one that has two or more categories, but there is no intrinsic ordering to the categories.
- An ordinal variable is similar to a categorical variable. The difference between the two is that there is a clear ordering of the variables.
- An interval variable is similar to an ordinal variable, except that the intervals between the values of the interval variable are equally spaced.
- Ratio scales give us the ultimate–order, interval values, plus the ability to calculate ratios since a "true zero" can be defined.

TYPES OF VISUALIZATIONS:

Comparing categories:

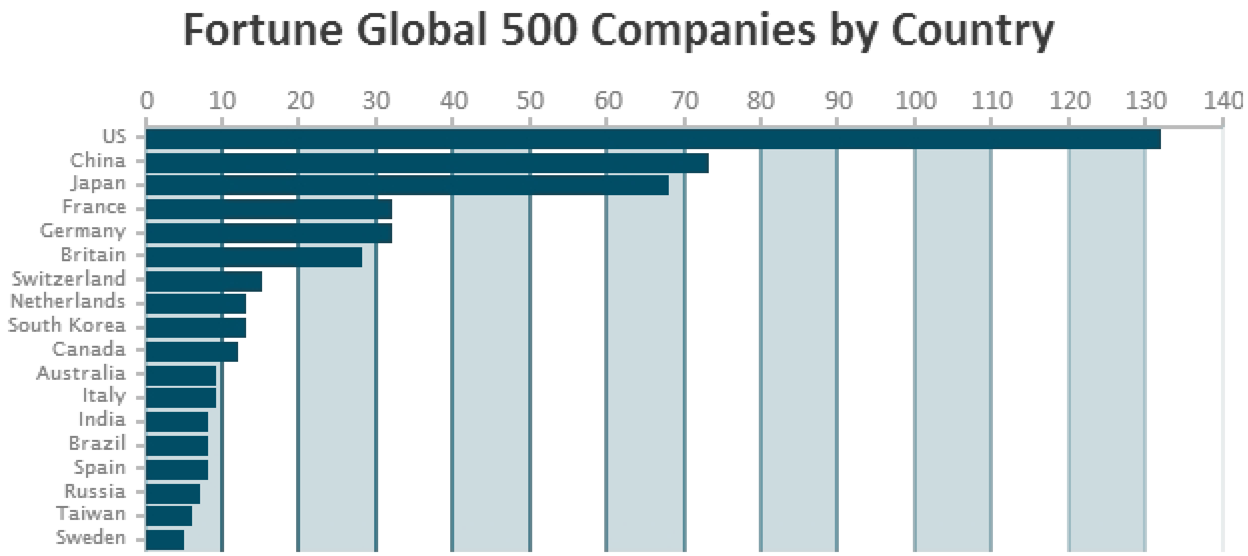
The following examples present chart types that facilitate the comparison of categorical values.

Bar chart (or column chart)

Data variables: 1 categorical, 1 quantitative

Visual variables: Length/height, color-hue

Description: Bar charts convey data through the length or height of a bar, allowing us to draw accurate comparisons between categories for both relative and absolute values. When using length as the visual variable to represent a quantitative value it is important to show the full extent of this property so always start the bar from the zero point on the axis.

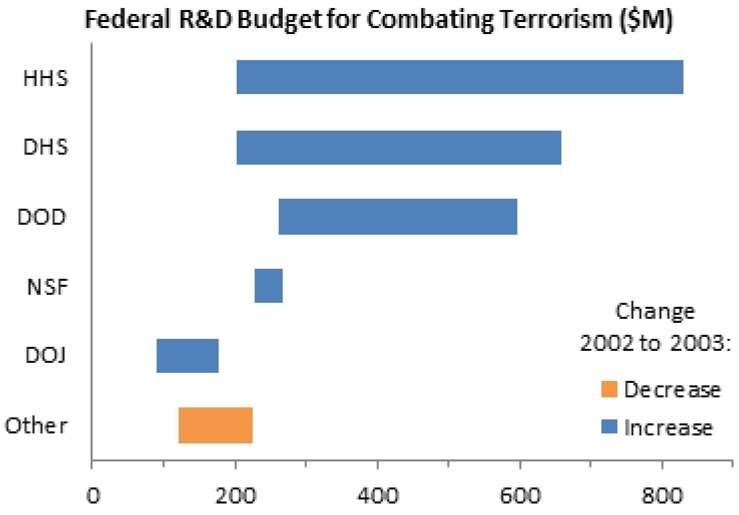


Floating bar (or Gantt chart)

Data variables: 1 categorical-nominal, 2 quantitative

Visual variables: Position, length

Description: A floating bar chart—sometimes labeled a Gantt chart because of similarities in appearance—helps to show the range of quantitative values. It presents a bar stretching from the lowest to the highest values (therefore the starting position is not the zero point). Using such charts enables you to identify the diversity of measurements within a category and view overlaps and outliers across all categories.

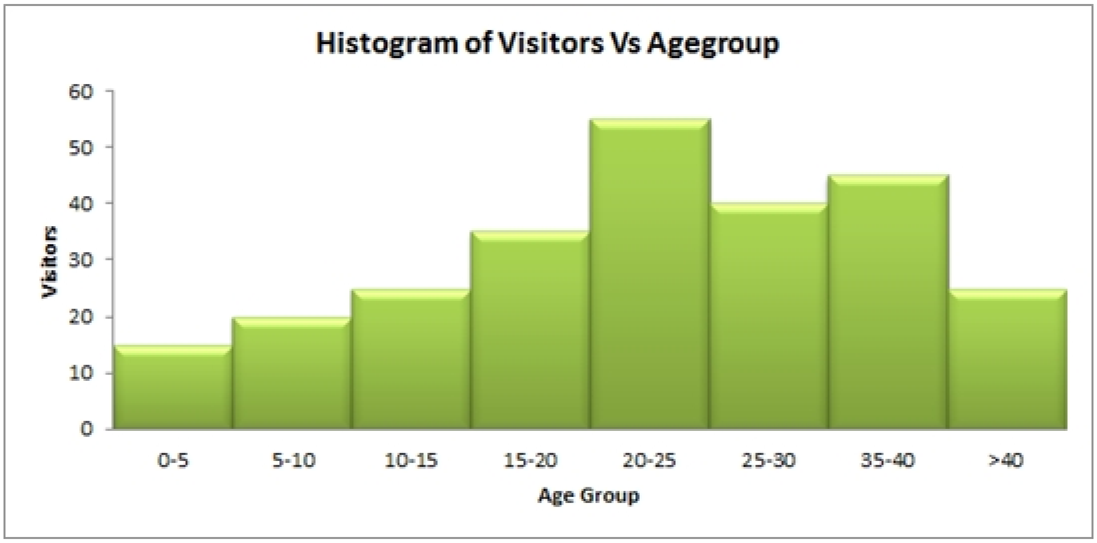


Histogram

Data variables: 1 quantitative-interval, 1 quantitative-ratio

Visual variables: Height, width

Description: Histograms are often mistaken for bar charts but there are important differences. Histograms show distribution through the frequency of quantitative values (y axis) against defined intervals of quantitative values (x axis). By contrast, bar charts facilitate comparison of categorical values. One of the distinguishing features of a histogram is the lack of gaps between the bars, as shown below

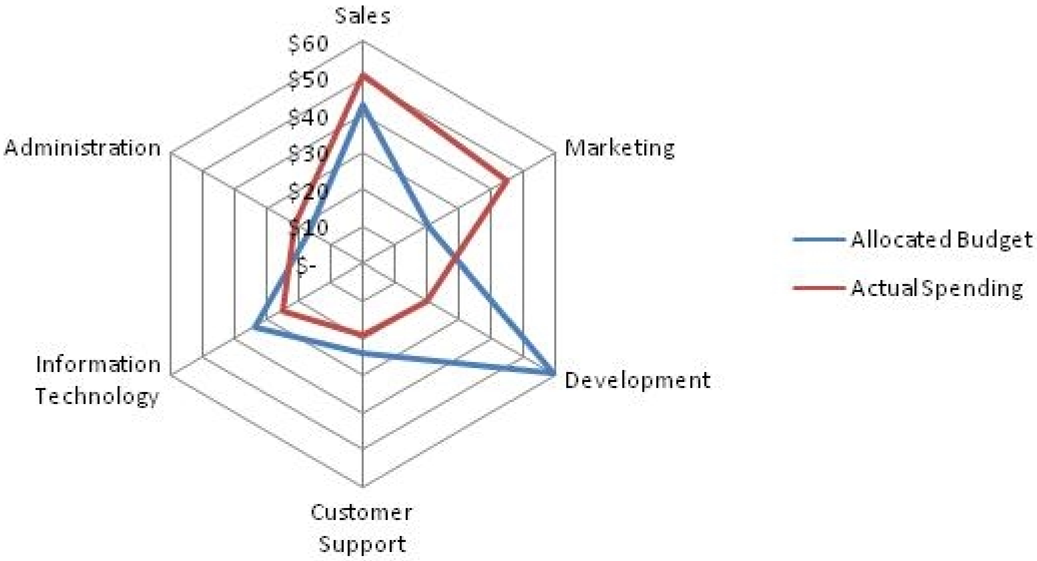


Radical chart

Data variables: 1 or more categorical, 1 categorical-ordinal

Visual variables: Position, color-hue, color-saturation/lightness, texture

Description: A radial chart displays data around a concentric, circular layout.



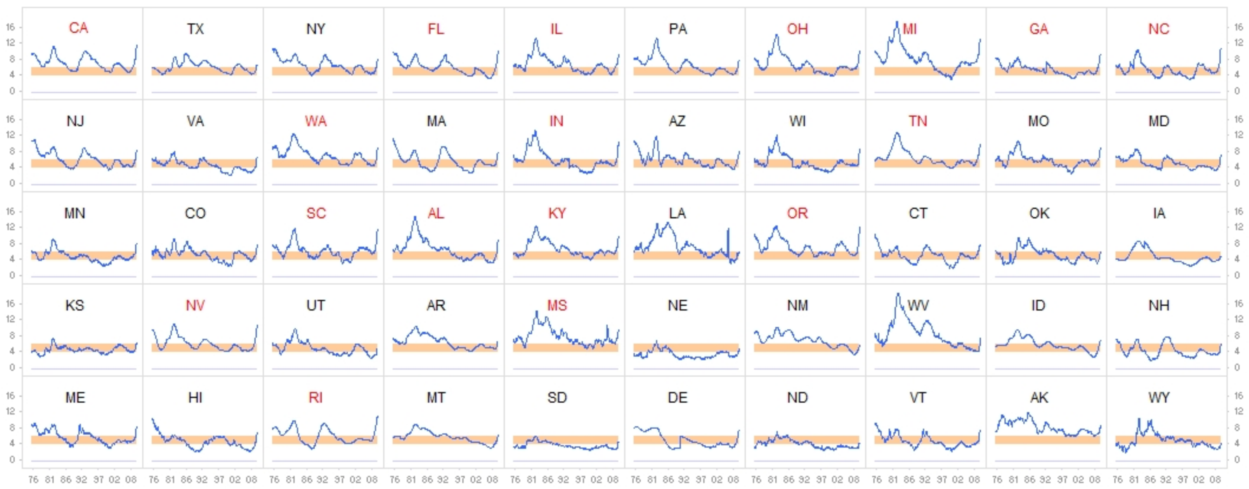
Small multiples (or trellis chart)

Data variables: 1 or more categorical, 1 or more quantitative

Visual variables: Position, any visual variable.

Description: Small multiples are not really a separate chart type but an arrangement approach that facilitates efficient and effective comparisons to be made across a multi panel display of small chart elements. These displays exploit the capacity of our visual system to rapidly scan across a trellis of small similar charts and to be capable of easily and immediately spotting patterns. These are particularly useful for comparing categories that have a broad range of values. They also work very well for showing snapshots of events that change over time.

Monthly Unemployment Rates by State, Jan 1976 - Apr 2009



Source: Bureau of Labor Statistics

Notes: The orange band denotes a "normal" unemployment rate (4%-6%);
State code in red: unemployment rate in April 2009 is higher than the US average

Dot plot

Data variables: 1 or more categorical, 1 or more quantitative

Visual variables: Position, color-hue, symbol

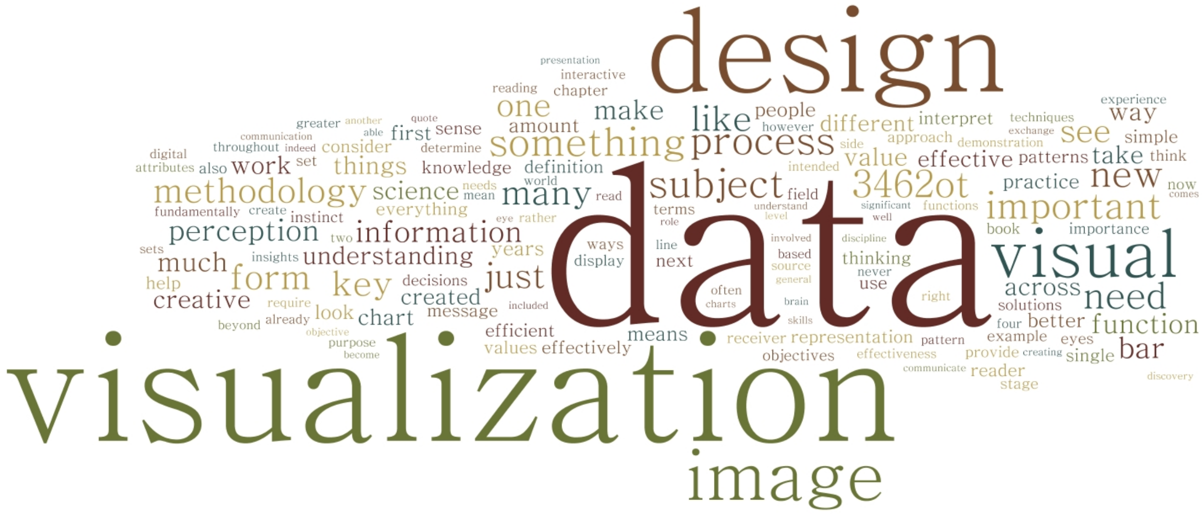
Description: A dot plot compares categorical variables by representing quantitative values with a single mark, such as a dot or symbol. The use of sorting helps you to clearly see the range and distribution of values. You can also combine multiple categorical value series on to the same chart distinguishing them using color or variation in symbol.

Word cloud

Data variables: 1 categorical, 1 quantitative

Visual variables: Size

Description: Word clouds depict the frequency of words used in a given set of text. The font size indicates the quantity of each word's usage. Color is often just used as decoration.



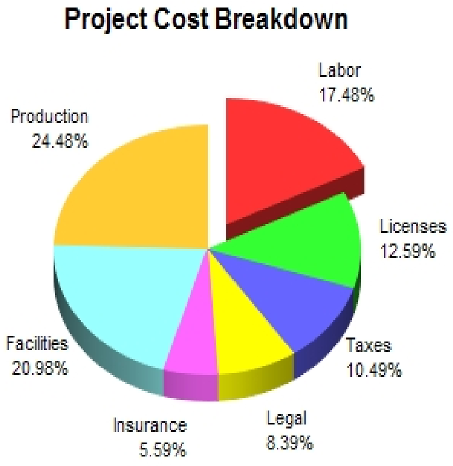
Assessing hierarchies and part-to-whole relationships

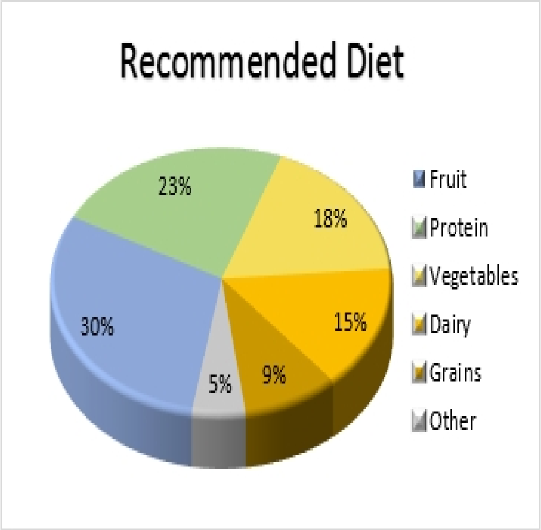
Pie Chart

Data variables: 1 categorical, 1 quantitative

Visual variables: Angle, area, color-hue.

Description: a type of graph in which a circle is divided into sectors that each represent a proportion of the whole.



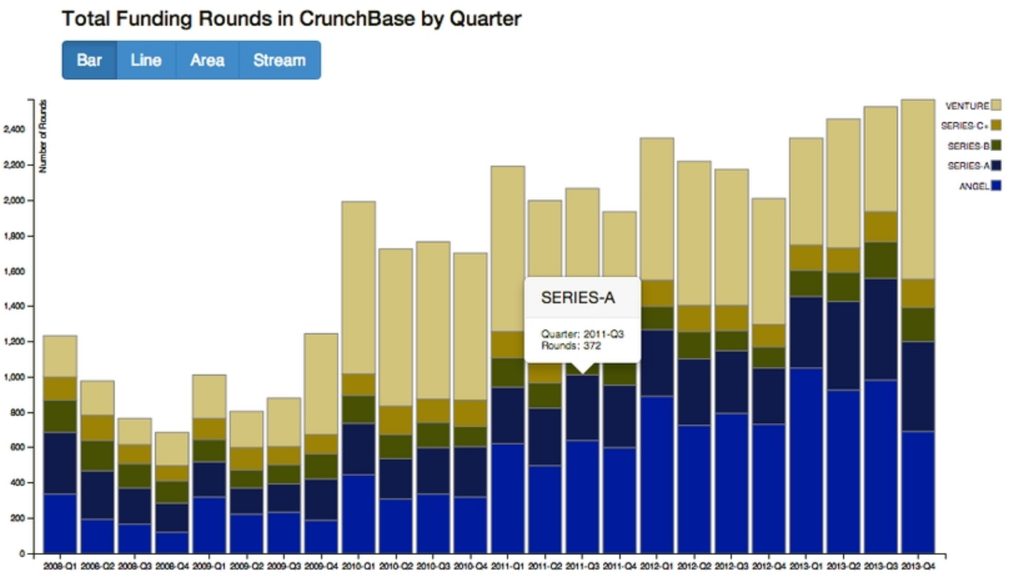


Stacked bar chart (or stacked column chart)

Data variables: 1 or more categorical, 1 quantitative-ratio

Visual variables: Length, color-hue, position, color-saturation/lightness.

Description: Stacked bars are fairly self-explanatory. They can be based on the stacks of absolute values or standardized to show part of a whole breakdown, as in following example. Colors and position differentiate the value categories.



Tree Map

Data variables: 1 or more categorical-nominal, 1 quantitative-ratio

Visual variables: Area, position, color-hue, color-saturation/lightness.

Description: Tree maps take the concept of a whole population and divide up portions of rectangular spaces within to represent organized, clustered constituent units sized according to their relative value. As well as arrangement, various properties of color are typically used to provide additional layers of quantitative or categorical insight.

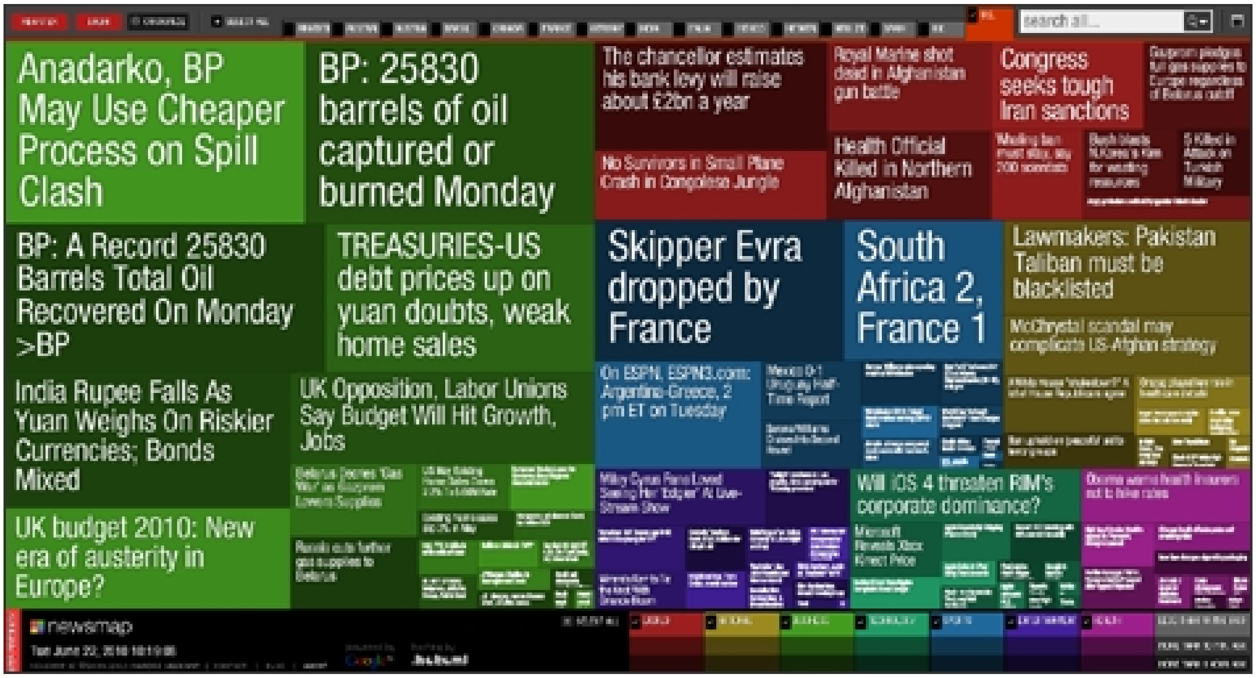


Image from "Newsmap" (http://newsmap.jp/), created by Marcos Weskamp

Circle packing diagram

Data variables: 1 or more categorical, 1 quantitative-ratio

Visual variables: Area, color-hue, position.

Description: As the title suggests, this type of chart seeks to pack together constituent circles into an overall circular layout that represents the whole. Each individual circle represents a different category and is sized according to the associated quantitative value. Other visual variables, such as color and position, are often incorporated to enhance the layers of meaning of the display.

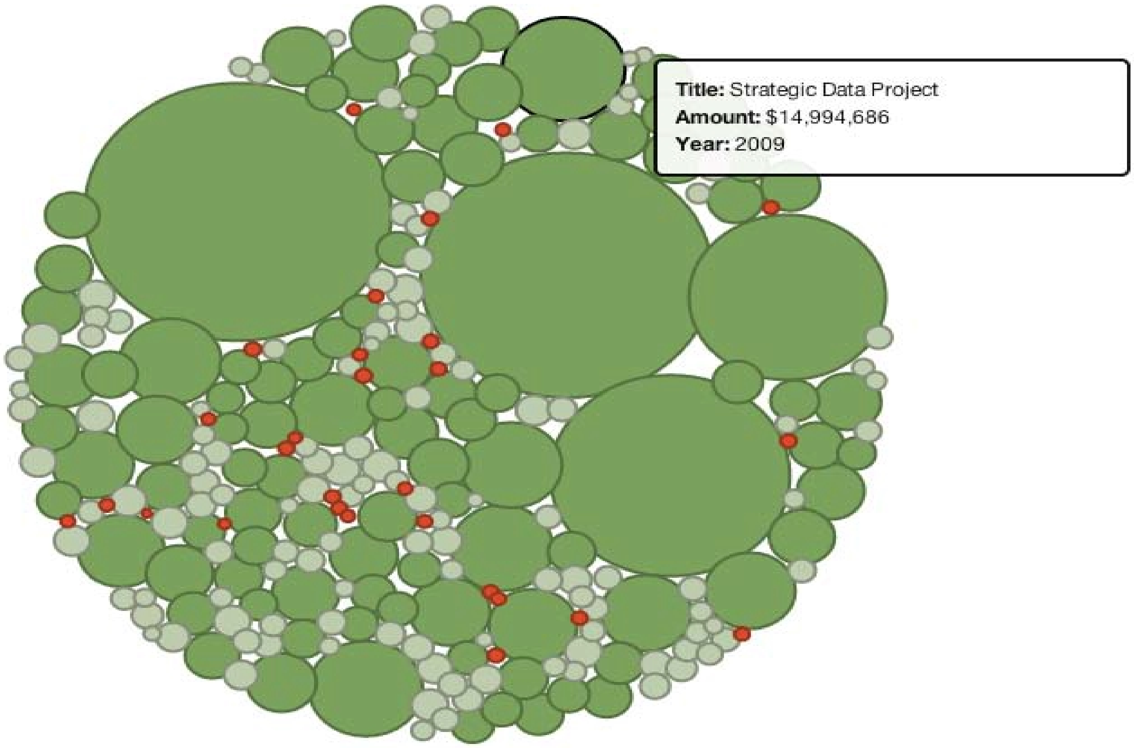


Image from "Gates Foundation Educational Spending" (http://vallandingham.me/vis/gates/), created by Jim Vallandingham

Showing changes over time

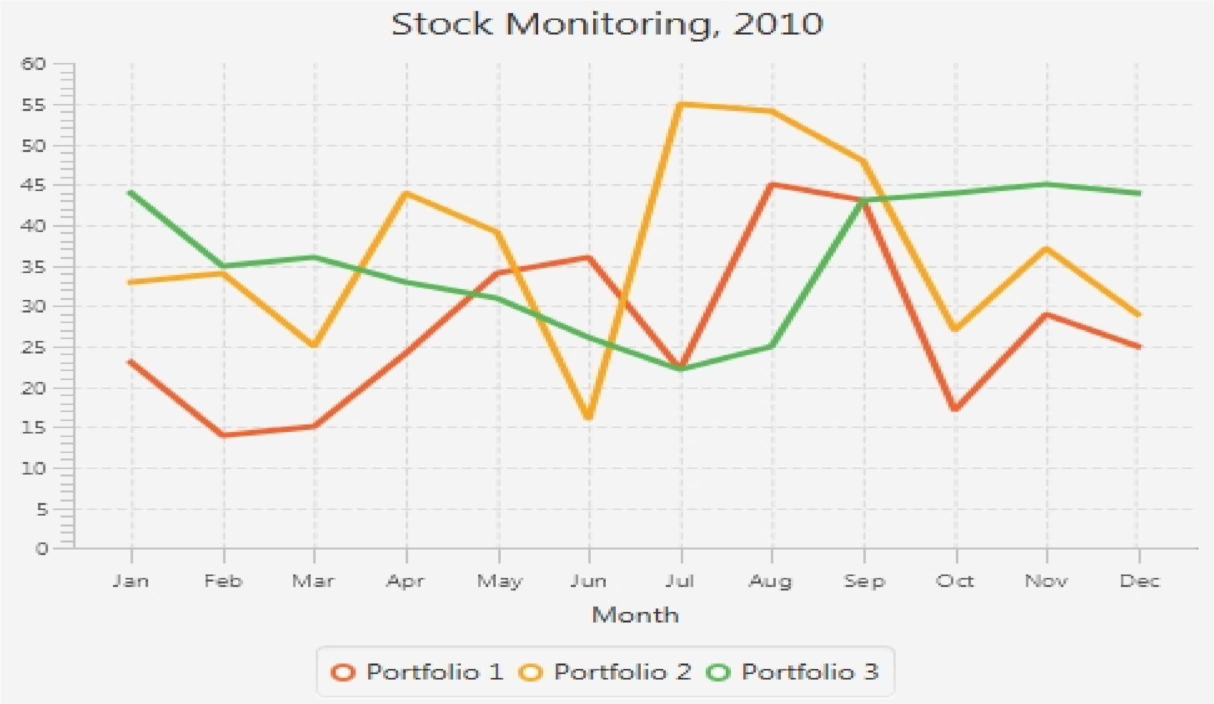
The following examples show alternative ways of graphically showing changes over time:

Line Chart

Data variables: 1 quantitative-interval, 1 quantitative-ratio, 1 categorical.

Visual variables: Position, slope, color-hue.

Description: Line charts are something we should all be familiar with. They are used to compare a continuous quantitative variable on the x axis and the size of values on the y axis. The vertical points are joined up using lines to show the shifting trajectory through the resulting slopes. Line charts can help unlock powerful stories of the relative or (maybe) related transition of categorical values.

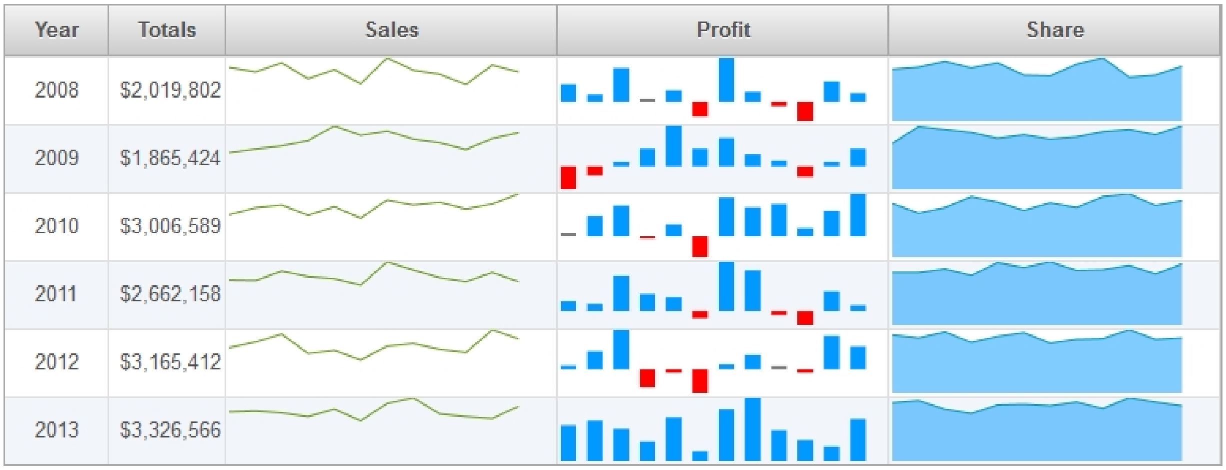


Sparkline

Data variables: 1 or more quantitative-interval, 1 or more quantitative-ratio.

Visual variables: Position, slope.

Description: A sparkline is a very small line chart, typically drawn without axes or coordinates. It presents the general shape of the variation (typically over time) in some measurement, such as sales or profits, in a simple and highly condensed way. Sparklines are small enough to be embedded in text, or several sparklines may be grouped together as elements of a small multiple.



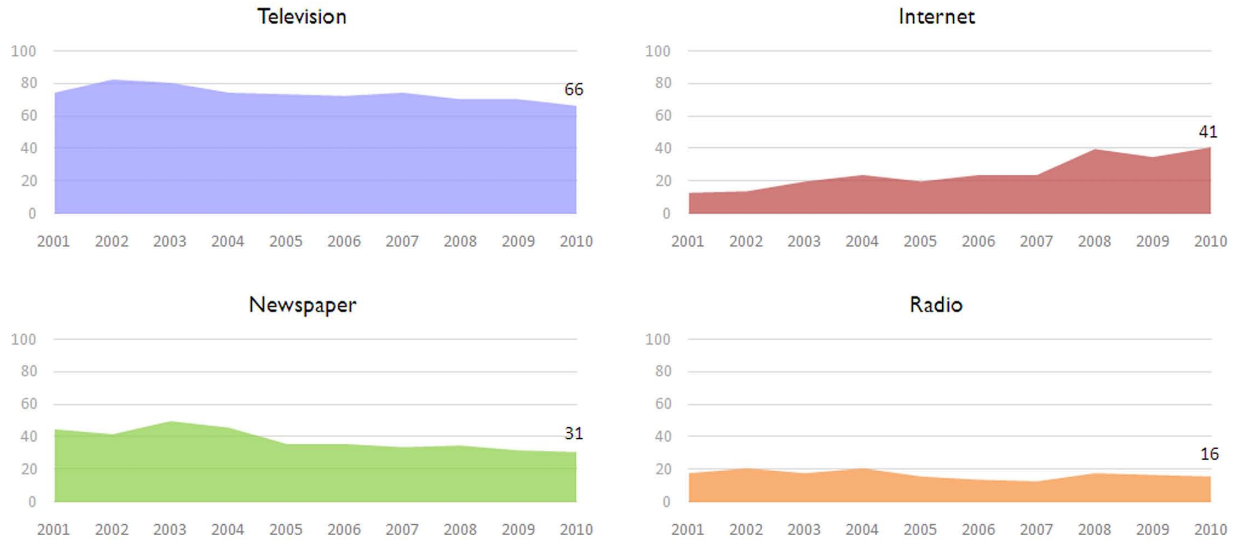
Year	Totals	Sales	Profit	Share
2008	$2,019,802			
2009	$1,865,424			
2010	$3,006,589			
2011	$2,662,158			
2012	$3,165,412			
2013	$3,326,566			

Area chart

Data variables: 1 quantitative-interval, 1 categorical, 1 quantitative-ratio.

Visual variables: Height, slope, area, color-hue.

Description: A number of visual properties are involved in area charts as seen below. The vertical position and connecting slope of the horizon (like a line chart) shows the progression of the values over time and the color area underneath the chart helps to emphasize these changes. Unlike a standard line chart, an area chart should have the y axis starting at zero to ensure the area judgment is being interpreted accurately.



Horizon chart

Data variables: 1 quantitative-interval, 1 categorical, 2 quantitative-ratio.

Visual variables: Height, slope, area, color-hue, color-saturation/lightness.

Description: This is a variation on the area chart, modified to include (and cope with) both positive and negative values. Rather than presenting negative values beneath the x axis, the negative area is mirrored on to the positive side and then colored differently to indicate its negative polarity. The result is a chart that occupies a single row of space, which helps to accommodate multiple stories onto a single display and facilitates comparison to pick out local and global patterns of change over time.

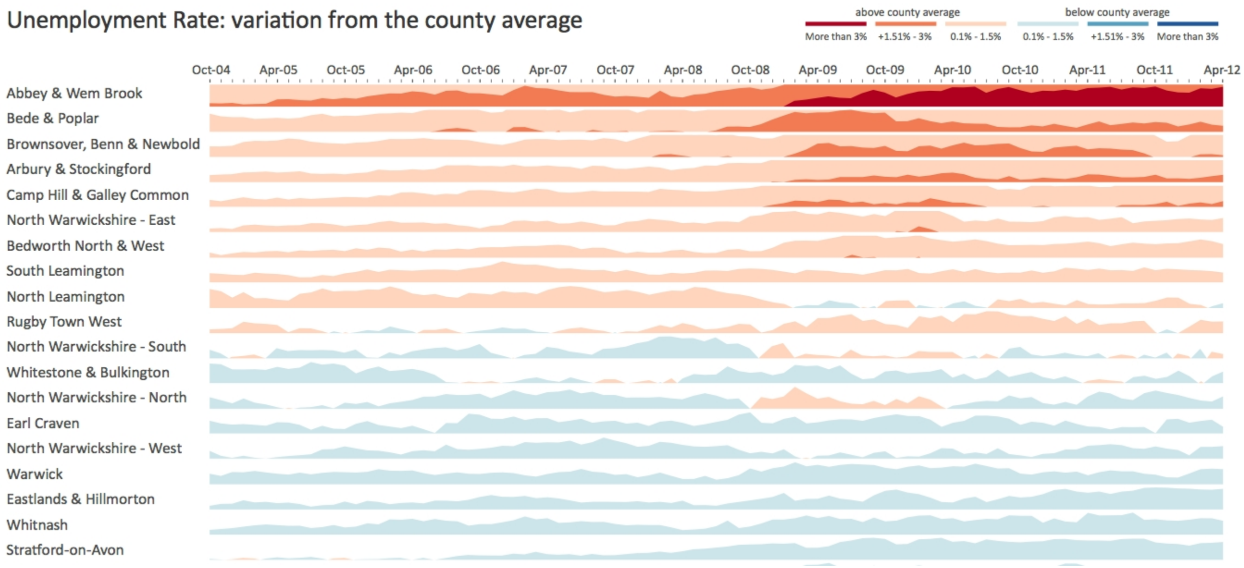


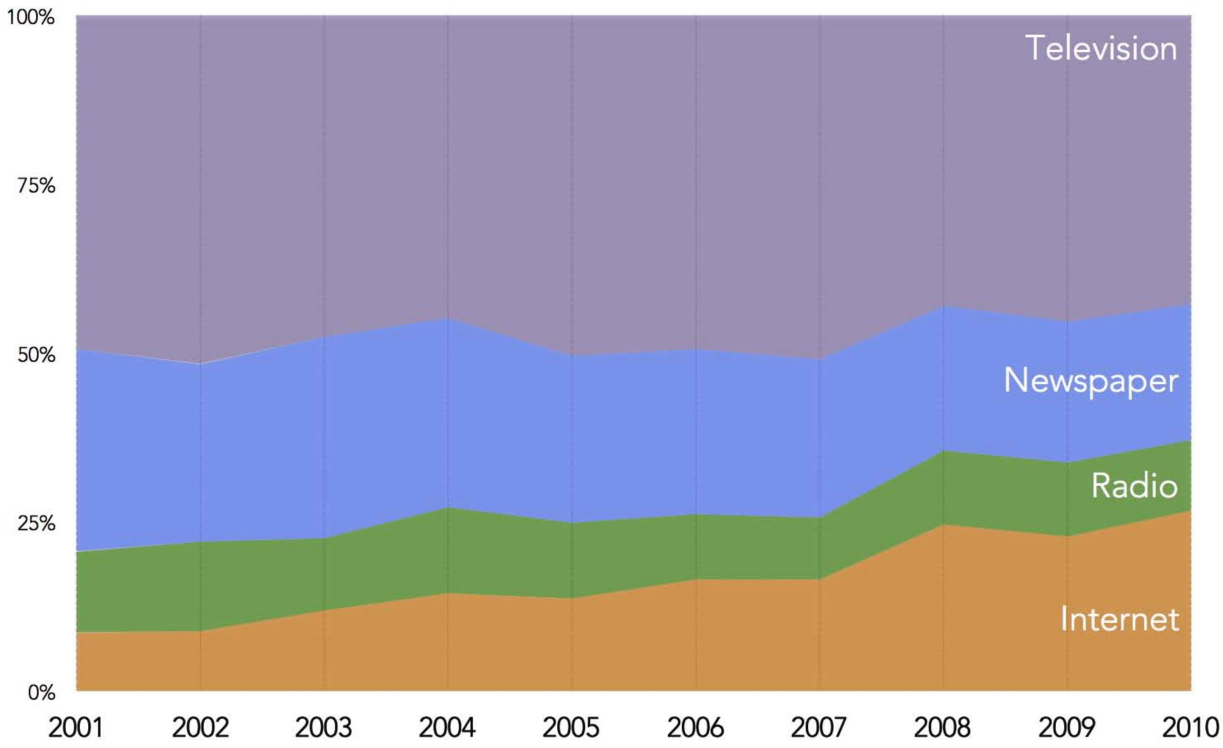
Image from "Unemployment Rate: variation from the county average" (http://warksobservatory.files.wordpress.com/2012/07/unemployment-horizon-chart.pdf), created by Spencer Payne/Warwickshire Observatory

Stacked area chart

Data variables: 1 quantitative-interval, 1 categorical, 1 quantitative-ratio.

Visual variables: Height, area, color-hue.

Description: A stacked area chart provides a compositional view of categories to show their changes over time. As the title suggests, these are based on stacks of area charts differentiated by color and present either absolute aggregates or percentage aggregates. Note that the quantitative values are represented by the height (derived from top and bottom positions) of the area stacks at any given point.

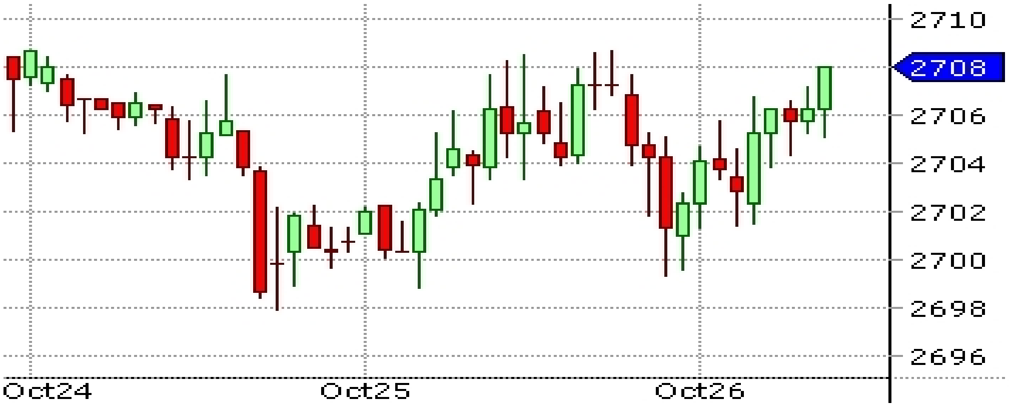


Candlestick chart (or box and whiskers plot, OHLC chart)

Data variables: 1 or more quantitative-interval, 4 quantitative-ratio.

Visual variables: Position, height, color-hue.

Description: The candlestick chart is commonly used in financial contexts to reveal the key statistics about a stock market for a given timeframe (often daily). In the following example, we see stock market changes by day based on the OHLC measures—opening, highest, lowest, and closing prices. The height of the central bar indicates the change from the opening to closing price and the color tells us if this is an increase or decrease.



Plotting connections and relationships

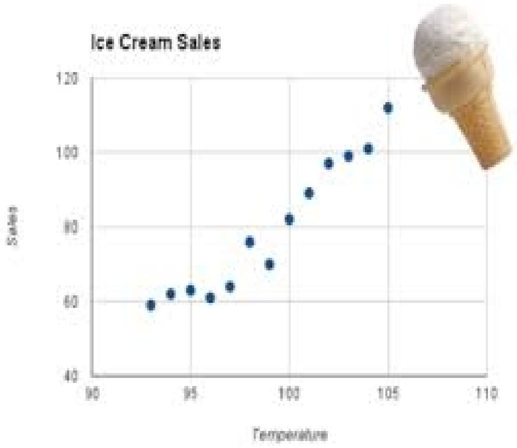
We now look at the different visualization techniques used to plot connections and relationships:

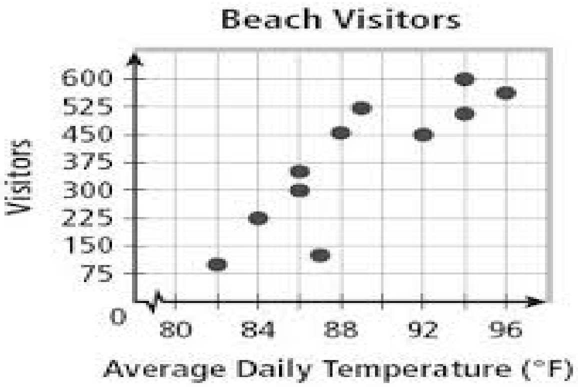
Scatter plot

Data variables: 2 or more quantitative.

Visual variables: Position, color-hue.

Description: A scatter plot is a combination of two quantitative variables plotted on to the x and y axes in order to reveal patterns of correlations, clustering, and outliers. This is a very important chart type, in particular, for when we are familiarizing with and exploring a dataset.





Bubble plot

Data variables: 2 or more quantitative, 2 or more categorical.

Visual variables: Position, area, color-hue.

Description: A bubble plot extends the potential of a scatter plot through multiple encoding of the data mark. In the following example, we see the marks becoming circles of varying size and then colored according to their categorical relationship. Often, you will see a further layer of time-based data applied to convey motion with the plot animated over time.

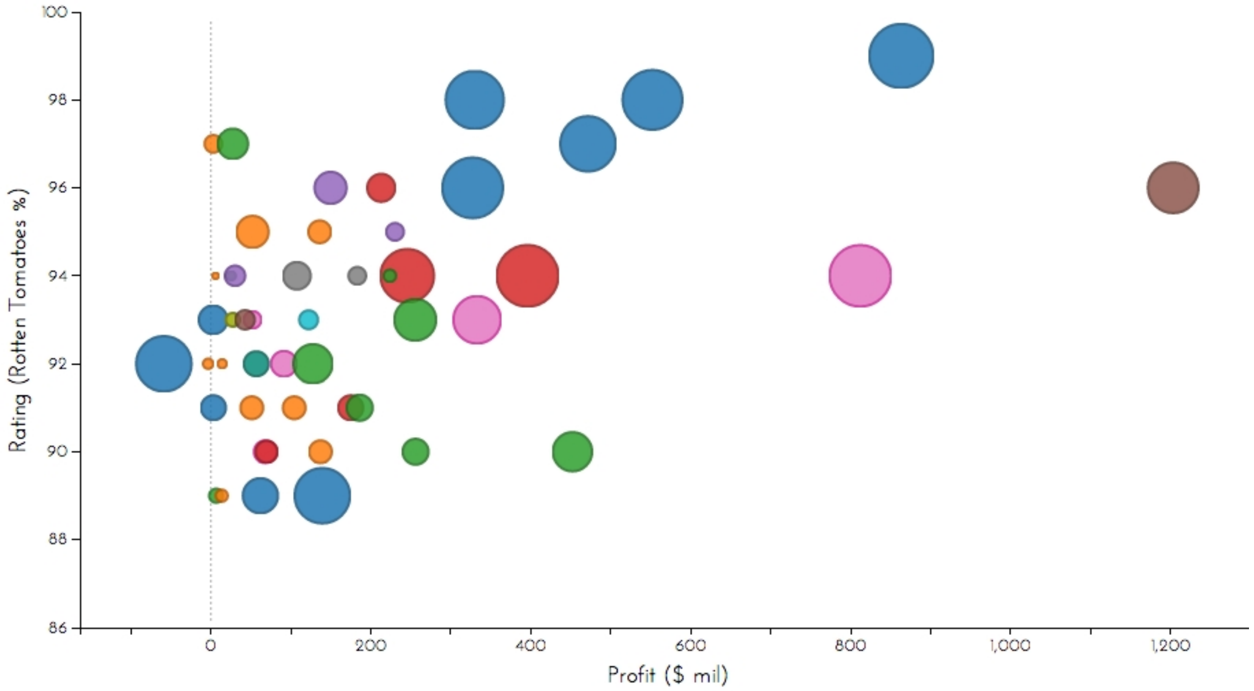


Image from (http://vallandingham.me/vis/movie/)

Scatter plot matrix

Data variables: 2 or more quantitative, 2 or more categorical.

Visual variables: Position, color-hue.

Description: Similar to the small multiples chart that we saw earlier, a scatter plot matrix takes advantage of the eye's rapid capability to spot patterns across multiple views of the same type of chart. In the following case, we have a panel of multiple combined scatter plots:

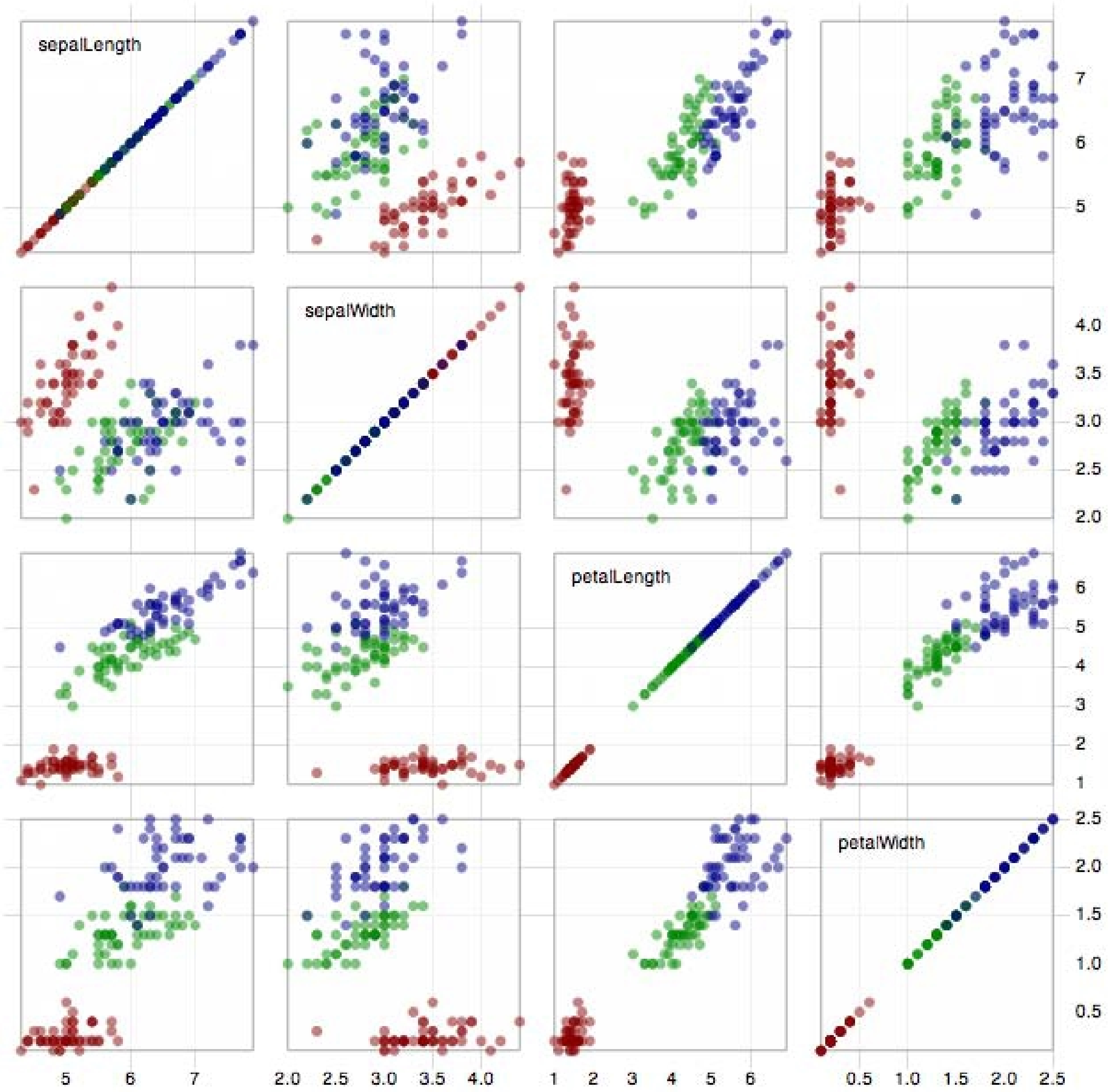


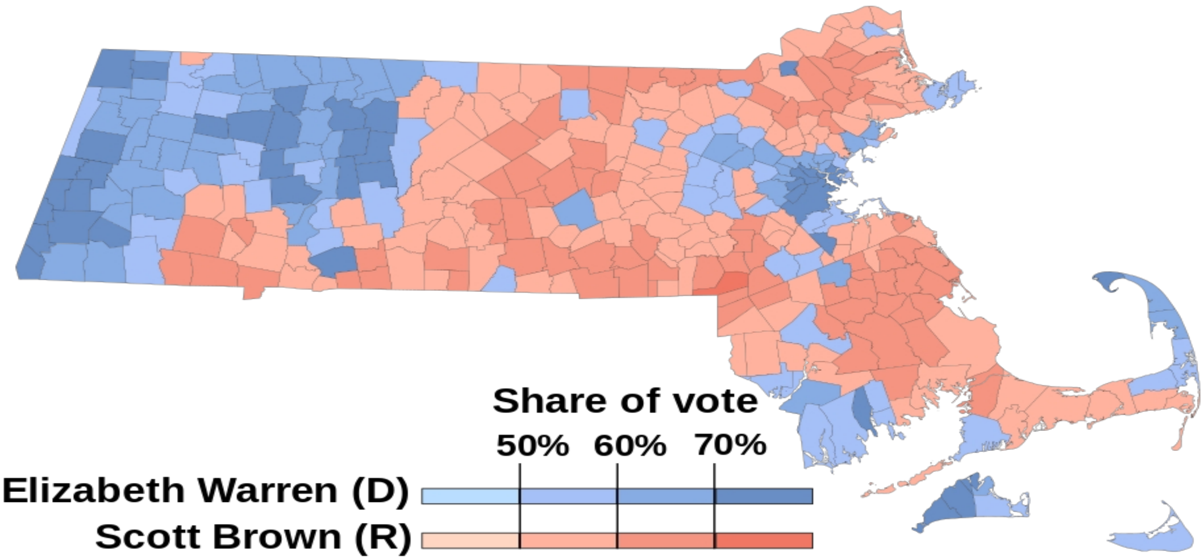
Image from "Scatterplot Matrix" (http://mbostock.github.com/d3/ex/splom.html), created by Mike Bostock

Heatmap (or matrix chart)

Data variables: Multiple categorical, 1 quantitative-ratio.

Visual variables: Position, color-saturation.

Description: With further similarities to small multiples, heat maps enable us to perform rapid pattern matching to detect the order and hierarchy of different quantitative values across a matrix of categorical combinations. The use of a color scheme with decreasing saturation or increasing lightness helps create the sense of data magnitude ranking.

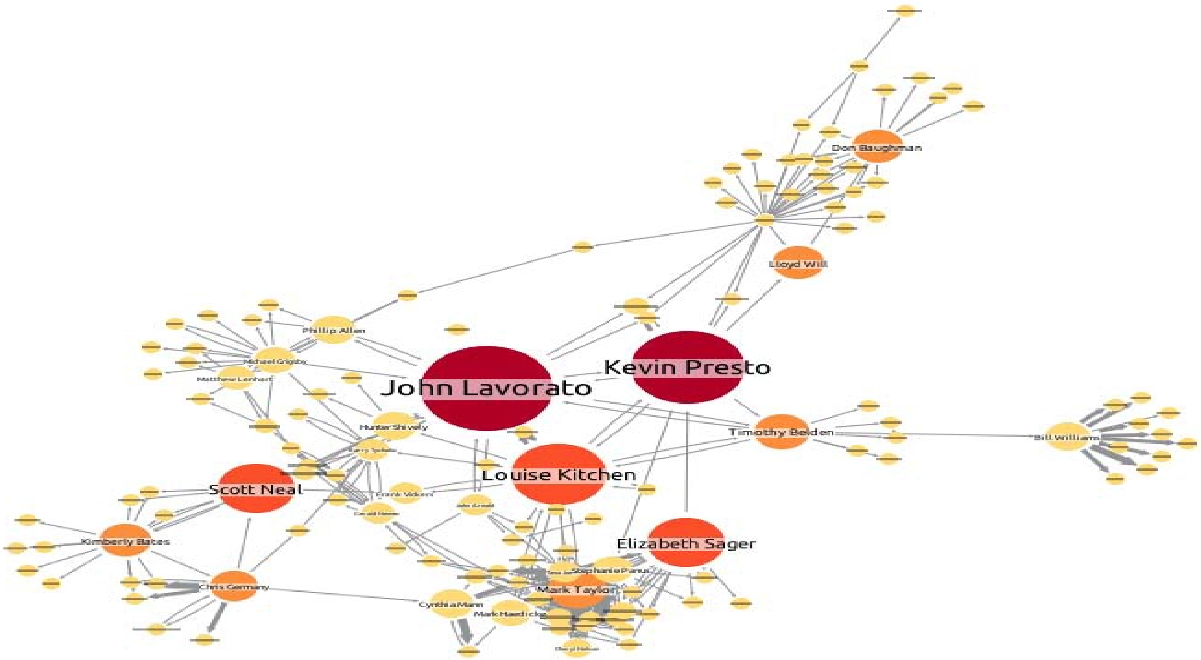


Network diagram (or force-directed/node-link network)

Data variables: Multiple categorical-nominal, 1 or more quantitative-ratio.

Visual variables: Position, connection, area, color-hue.

Description: We can look quite daunting through their visual complexity and apparent clutter (indeed, often they are described as "hairballs"). Their intention and value is to facilitate exploration of complex data frameworks based on the existence or quantifiable strength of relationships, connections, and logical organization. The typical purpose of these graphs is to enable the viewer to get a sense of patterns—picking out the elements that are of interest, observing clusters and gaps, dominant nodes and sparse connections.



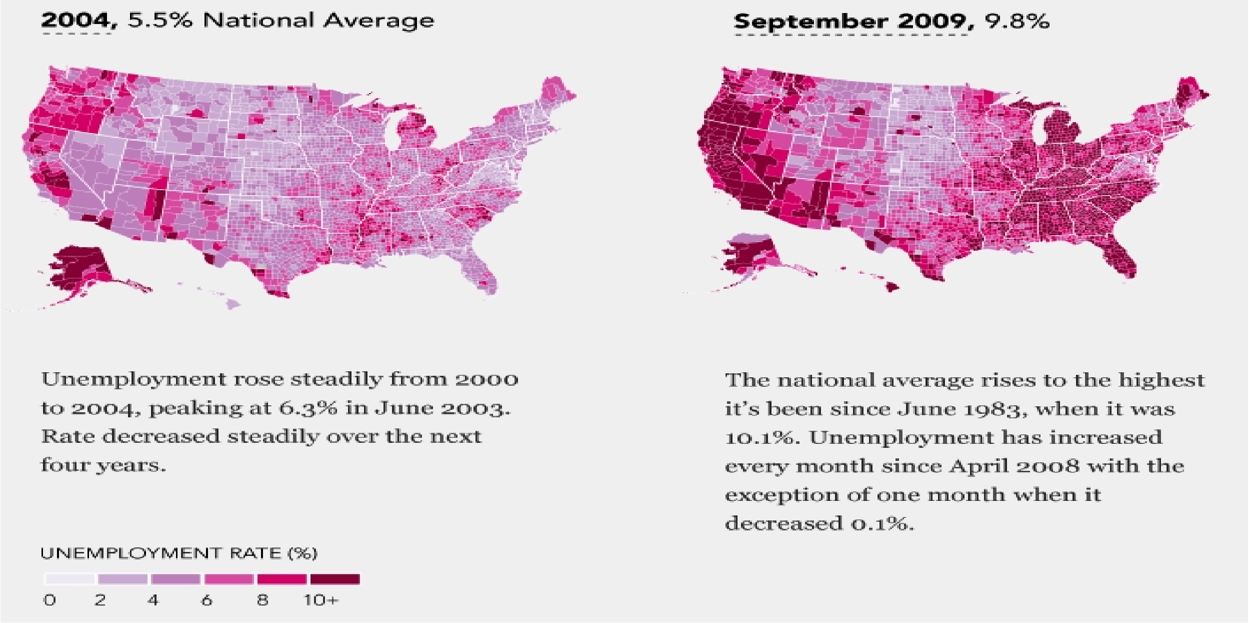
Mapping geo-spatial data

Choropleth map

Data variables: 2 quantitative-interval, 1 quantitative-ratio.

Visual variables: Position, color-saturation/lightness.

Description: Choropleth maps color the constituent geographic units (such as states or counties) based on quantitative values using a sequential or diverging scheme of saturation/lightness.



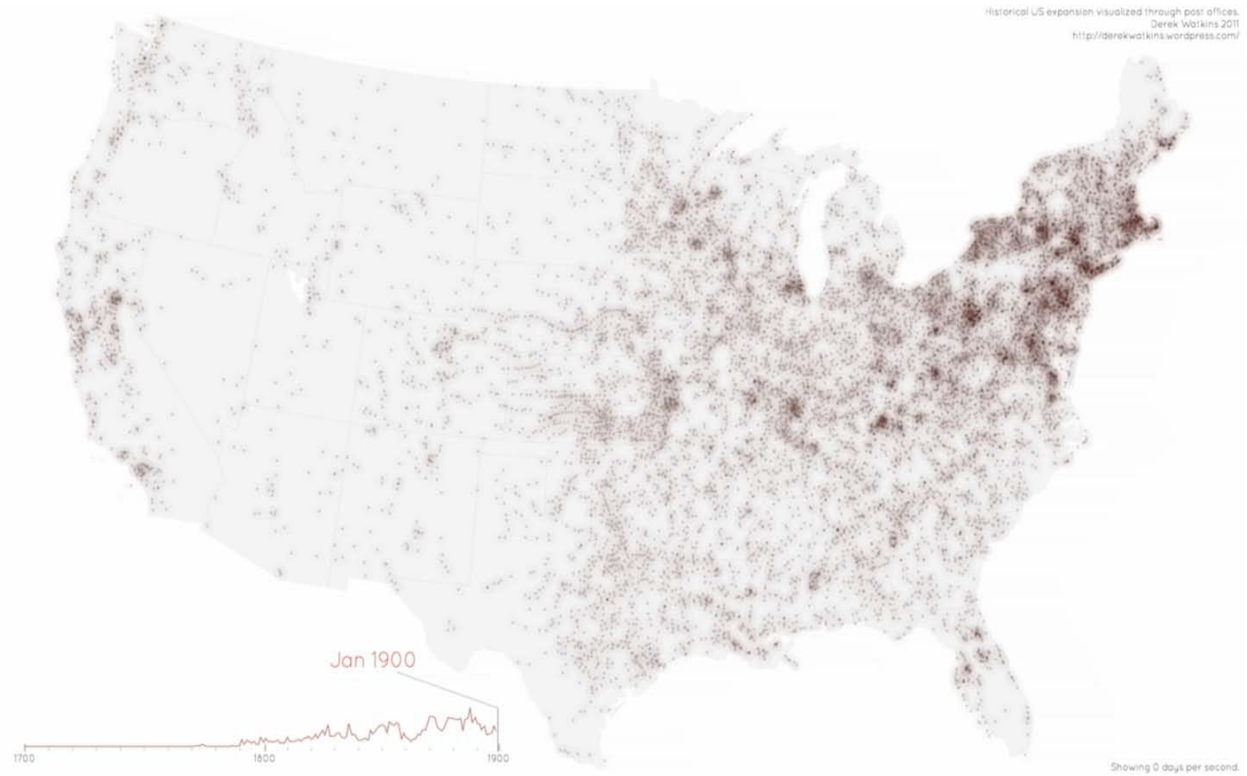
Unemployment rose steadily from 2000 to 2004, peaking at 6.3% in June 2003. Rate decreased steadily over the next four years.

The national average rises to the highest it's been since June 1983, when it was 10.1%. Unemployment has increased every month since April 2008 with the exception of one month when it decreased 0.1%.

Image (cropped) from "Unemployment, 2004 to present" (http://projects.flowingdata.com/america/unemployment/raw.html), created by Nathan Yau

Dot plot map

Data variables: 2 quantitative-interval.
Visual variables: Position.
Description: A dot plot map essentially displays a geographical scatter plot of records, combining the longitude and latitude to position marks on the map. In the following example, we also see this data being gradually plotted over time to reveal a story of geographical spread.



Historical US expansion visualized through post offices.
Derek Watkins 2011
http://derekwatkins.wordpress.com/

Image from "Posted: Visualizing US Expansion Through Post Offices" (http://blog.dwtkns.com/2011/posted/), created by Derek Watkins

Readings: Some of the useful links you can go over to understand visualization, used in the course material and just see some of the visualizations being done.

1. http://www.visualisingdata.com/
2. https://public.tableau.com/s/resources?edition=public&version=9.3.5&__full-version=9300.16.0726.1843&platform=windows&status=buy&qt-overview_resources=1#qt-overview_resources
3. http://queue.acm.org/detail.cfm?id=1805128
4. http://vallandingham.me/vis/movie/
5. http://bl.ocks.org/mbostock/4063663
6. http://www.scribblelive.com/blog/2012/07/27/45-ways-to-communicate-two-quantities/
7. http://www.storytellingwithdata.com/
8.