

Chapter 4

DATA VISUALISATION

Learning Objectives and Outcomes

In this unit we shall learn about

- What is Data Visualisation and its importance
- Types of Data Visualisation
- Data Encodings,
- Visual Perception and Retinal variables,
- Mapping Variables to Encodings,
- Visual Encodings

4.1 DATA VISUALIZATION

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software. Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it's processed.

Data visualization is both an art and a science.^[1] It is viewed as a branch of descriptive statistics by some, but also as a grounded theory development tool by others. Increased amounts of data created by Internet activity and an expanding number of sensors in the environment are referred to as "big data" or Internet of things. Processing, analyzing and communicating this data present ethical and analytical challenges for data visualization. Various data visualisation tools are discussed in section 6.2 of this book.

4.2 Data Attributes

When it comes to data attributes, there are two categories: quantitative data and qualitative data. Quantitative data is exactly what it sounds like: a numerical value placed on an ascending scale

(i.e. I am 61 years of age, I drank two bottles of water today). Qualitative data refers to values that cannot be measured numerically, but can be described through language (i.e. I came in 3rd place at the swim meet, since I'm always on the run I prefer a laptop over a desktop). Within these two categories are a total of four subcategories as well:

4.2.1 Quantitative Data

Quantitative Data can take shape as:

1. Ratio (cost \$10, \$20, \$30 or age 10 years old, 20 years old, 30 years old)
2. Data you can perform arithmetic operations on (add, divide, etc)
3. Intervals (temperature -5° , 10° , 25° or time 1am, 5am, 9am)
4. Data with a set value that you cannot perform all arithmetic operations on

Example: You cannot calculate the sum of temperature during a week but you can calculate the average temperature per day and the high/low for each day.

Other data types for Visualisation

Author Stephen Few described eight types of quantitative messages that users may attempt to understand or communicate from a set of data and the associated graphs used to help communicate the message:

Time-series: A single variable is captured over a period of time, such as the unemployment rate over a 10-year period. A line chart may be used to demonstrate the trend.

Ranking: Categorical subdivisions are ranked in ascending or descending order, such as a ranking of sales performance (the measure) by sales persons (the category, with each sales person a categorical subdivision) during a single period. A bar chart may be used to show the comparison across the sales persons.

Part-to-whole: Categorical subdivisions are measured as a ratio to the whole (i.e., a percentage out of 100%). A pie chart or bar chart can show the comparison of ratios, such as the market share represented by competitors in a market.

Deviation: Categorical subdivisions are compared against a reference, such as a comparison of actual vs. budget expenses for several departments of a business for a given time period. A bar chart can show comparison of the actual versus the reference amount.

Frequency distribution: Shows the number of observations of a particular variable for given interval, such as the number of years in which the stock market return is between intervals such as 0-10%, 11-20%, etc. A histogram, a type of bar chart, may be used for this analysis. A boxplot helps visualize key statistics about the distribution, such as median, quartiles, outliers, etc.

Correlation: Comparison between observations represented by two variables (X,Y) to determine if they tend to move in the same or opposite directions. For example, plotting unemployment (X) and inflation (Y) for a sample of months. A scatter plot is typically used for this message.

Nominal comparison: Comparing categorical subdivisions in no particular order, such as the sales volume by product code. A bar chart may be used for this comparison.

Geographic or geospatial: Comparison of a variable across a map or layout, such as the unemployment rate by state or the number of persons on the various floors of a building. A cartogram is a typical graphic used.

4.2.2 Qualitative Data

Ordinal (size small, medium, large or position 1st place, 2nd place, 3rd place)

Data with a fixed ranking with indeterminate distance between the values

Example: A large elephant in India is very different from a large elephant in the Africa, but I don't know exactly how much larger.

(a) Nominal (sports NFL football vs. English football or computers laptop vs. desktop)

Data where you can distinguish between values, but not order them

Based on these classifications, the methods for aggregation and visualization of the data needs to adjust accordingly. For example, if you were to map car manufacturing data like the image below, and your data set included year-to-year manufacturing figures – it makes more sense to stick to an annual order. If you try to sort the values by highest value, your readers will have trouble following the order of the years (1978, 1979, 1980, etc). Ideally, ordinal data should be sorted by its order as opposed to the alphabetical sorting of the names in the values (if you were mapping month-by-month for example).

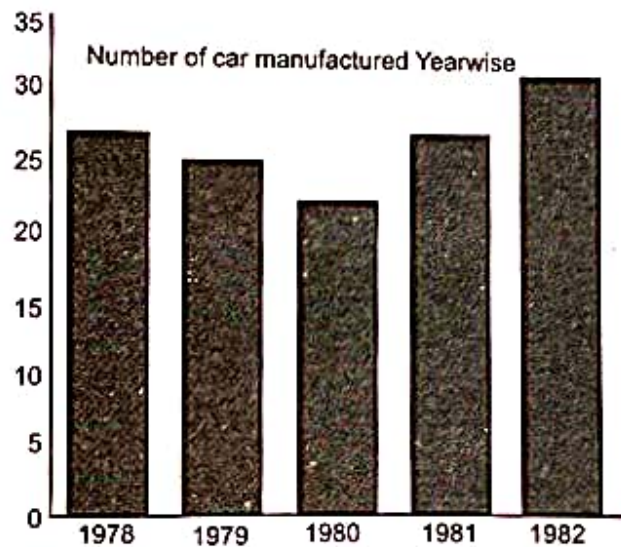


Fig. 4.1 Example of wrong graphical representation for sorting purpose

4.3 IMPORTANCE OF DATA VISUALIZATION

A primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics, plots and information graphics. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message.^[2] Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable.

Because of the way the human brain processes information, using charts or graphs to visualize

large amounts of complex data is easier than poring over spreadsheets or reports. Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments. This subject is further discussed in section 6.2.1 of this book.

Data visualization can also:

1. Identify areas that need attention or improvement.
2. Clarify which factors influence customer behavior.
3. Help you understand which products to place where.
4. Predict sales volumes.
5. Data visualization solutions were initially developed as a business tool for enterprises
6. that could afford to hire business analysts, citizen data scientists and BI experts capable
7. of performing sophisticated discovery and data analysis. Often, these experts functioned
8. (and still do) as internal consulting groups. This model is too expensive, slow and
9. clumsy for midsize businesses, and it should be avoided.

Graphical displays should:

1. show the data
2. induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else
3. avoid distorting what the data has to say
4. present many numbers in a small space
5. make large data sets coherent
6. encourage the eye to compare different pieces of data
7. reveal the data at several levels of detail, from a broad overview to the fine structure
8. serve a reasonably clear purpose: description, exploration, tabulation or decoration
9. be closely integrated with the statistical and verbal descriptions of a data set.

4.4 CONVENTIONAL DATA VISUALIZATION METHODS

Many conventional data visualization methods are often used. They are: table, histogram, scatter plot, line chart, bar chart, pie chart, area chart, flow chart, bubble chart, multiple data series or combination of charts, time line, Venn diagram, data flow diagram, and entity relationship diagram, etc. In addition, some data visualization methods have been used although they are less known compared the above methods. These are described in detail in section 5.6.1 of this book.

The additional methods are: parallel coordinates, treemap, cone tree, and semantic network, etc.^[3].

Parallel coordinates is used to plot individual data elements across many dimensions. Parallel coordinate is very useful when to display multidimensional data. Fig. 4.2 shows parallel coordinates. Treemap is an effective method for visualizing hierarchies. The size of each sub-rectangle represents one measure, while color is often used to represent another measure of data. Figure 4.3 shows a treemap of a collection of choices for streaming music and video tracks in a social network

community. Cone tree is another method displaying hierarchical data such as organizational body in three dimensions. The branches grow in the form of cone. A semantic network is a graphical representation of logical relationship between different concepts. It generates directed graph, the combination of nodes or vertices, edges or arcs, and label over each edge [3].

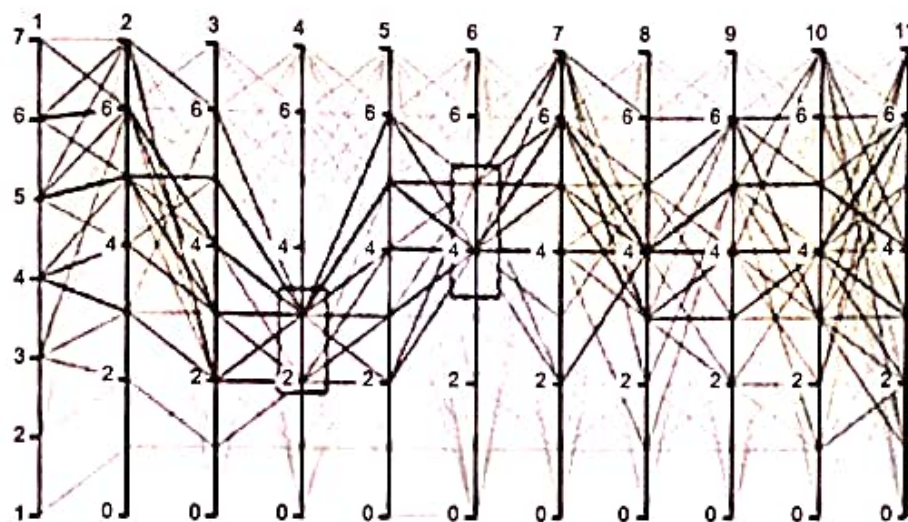


Fig. 4.2. Parallel coordinates [4]

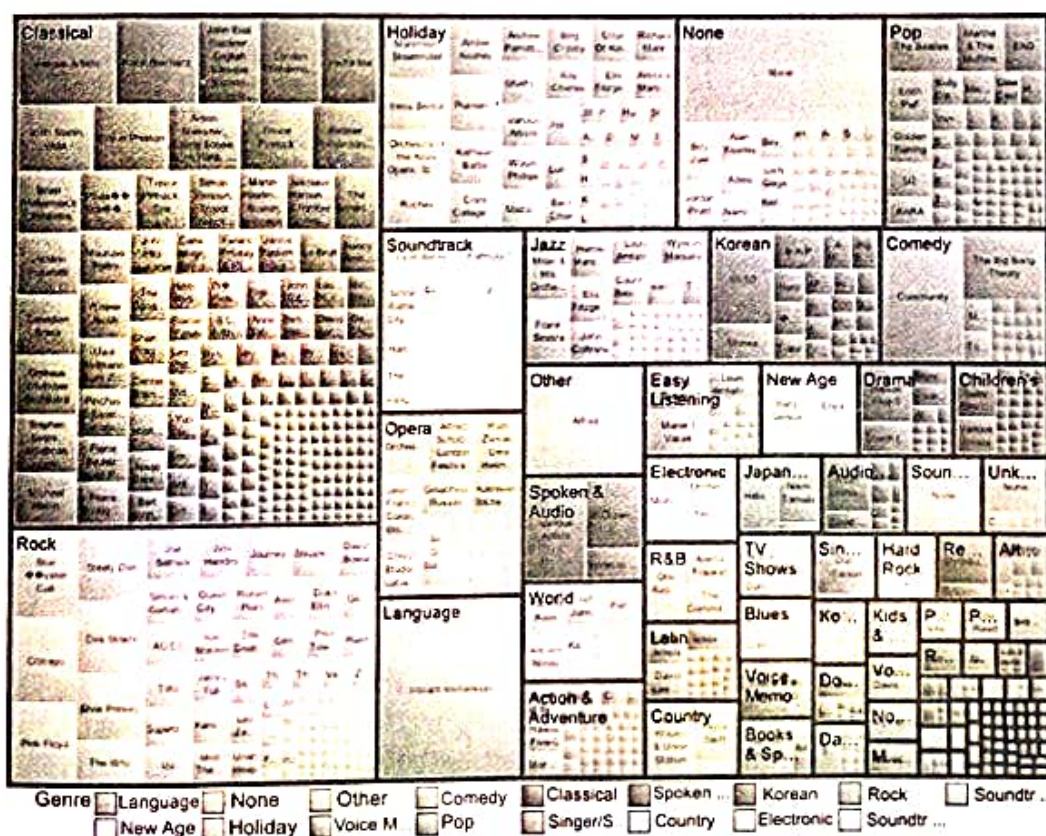

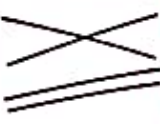






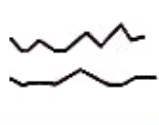





Fig. 4.3. Treemap view of a social network's track selections from a streaming media service [5]

4.4.1 Visual Perception and Data Visualization

A human can distinguish differences in line length, shape, orientation, and colour (hue) readily without significant processing effort; these are referred to as “pre-attentive attributes”. For example, it may require significant time and effort (“attentive processing”) to identify the number of times

the digit "5" appears in a series of numbers; but if that digit is different in size, orientation, or color, instances of the digit can be noted quickly through pre-attentive processing.

Pattern	Example	Pattern	Example
High, low and in between		Non-intersecting and intersecting	
Going up, going down and remaining flat		Symmetrical and skewed	
Steep and gradual		Wide and narrow	
Steady and fluctuating		Clusters and gaps	
Random and repeating		Tightly and loosely distributed	
Straight and curved		Normal and abnormal	

Visualizations are not only static; they can be interactive. Interactive visualization can be performed through approaches such as zooming (zoom in and zoom out), overview and detail, zoom and pan, and focus and context or fish eye^[6]. The steps for interactive visualization are as follows^[3]:

1. **Selecting:** Interactive selection of data entities or subset or part of whole data or whole data set according to the user interest.
2. **Linking:** It is useful for relating information among multiple views.
3. **Filtering:** It helps users adjust the amount of information for display. It decreases information quantity and focuses on information of interest.
4. **Rearranging or Remapping:** Because the spatial layout is the most important visual mapping, rearranging the spatial layout of the information is very effective in producing different insights.

We have already identified a process to determine what data type it is you have (nominal, ordinal, interval, ratio) and the axis to map it on. Now we need to figure out how to best visually display that data using colours, shapes, sizes and position.

For proper perspective on the subject, in 1984 William S. Cleveland and Robert McGill published a landmark piece of research on graphical perception that articulated the standards that many data visualizations abide by today. Their research, which was published in the *Journal of American*

Statistical Association, concluded that everyone has different perceptions of visualizations but there are a few simple steps that everyone can follow. Cleveland and McGill tested a series of visual encoding theories through experimentation and established a series of guidelines based on which visual marker is more accurate vs. less accurate.

4.4.2 Mapping of Data visualisation

For all data to be mapped to a visualization, these are your basic options of display:

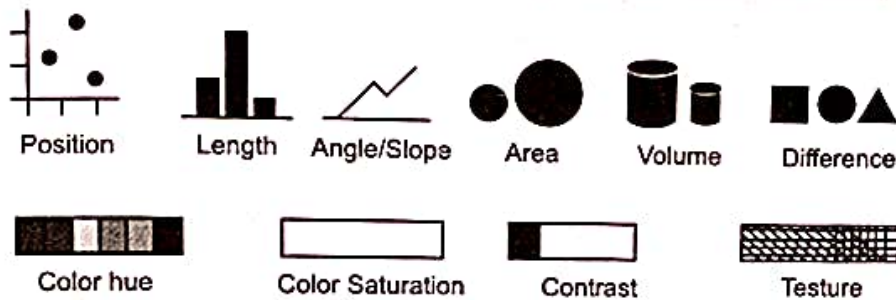


Fig. 4.4: Data Mapping : Basic Options of Display

4.5 RETINAL VARIABLES

Simone Garlandini Sara Irina Fabrikant in their article "Evaluating the Effectiveness and Efficiency of Visual Variables for Geographic Information Visualization" (typically multivariate) spatial data into a two-, three- or four-dimensional visuospatial display. This process is typically performed by applying scientific (i.e., Systematic, Transparent, and Reproducible) cartographic design methods, as well as aesthetic expressivity visual means. He lists seven basic visual variables and presents effects of varying the perceptual properties of the visual variables in order to derive meaningful representations.

4.5.1 Seven Variables for Visulisation

1. two planar variables
2. five so-called "retinal"

There are two planar variables (the x and y position on the map plane).

Five "Retinal" Variables

1. Size,
2. Color value,
3. Color hue,
4. Shape, and
5. Orientation,

Which we (and perhaps vision researchers) would probably translate as "pre-attentive" (Bertin, 1967/83). Although Bertin (1967/83) lists these variables individually, effective map representables.