Data – Basic concepts

Currently there are numerous sources that produce large amounts of data. The trend in recent years has been a progressive increase in this quantity, which is expected to continue. Every day data is generated (by sensors, software, means of communication and others, such as meteorological satellites, weather forecast software or mobile phones) which, if they do not somehow allow us to generate value, it will not be worth keeping them. This value can be obtained through various approaches, such as statistical analysis or machine learning, with visualization being a possibility to provide greater understanding to those who analyze the data, and can support better decisions. This document discusses what can be visualized and presents the main characteristics of the data that are relevant for choosing the visualization techniques to be used in a specific case. Understanding the data and the phenomenon it represents is a fundamental step; If not adequately treated, it may compromise the usefulness and effectiveness of the result. In what follows, a possible classification of the main types of data and datasets is presented, aiming at helping to understand the organization of the introduction of the most common visualization techniques.

1. What may be visualized

When thinking about data, probably the first idea that comes to mind is numerical values, such as height, weight, temperature or pressure. However, there are other types of data that are not numeric (for example days of the week or car brands), or even that do not have a value (such as links between nodes in a transportation network, or text) . Analyzing all of these types of data can provide understanding and support better decisions, but it is critical to their analysis and visualization to know their meaning in the real world (for example, whether they correspond to a person's name or type of product), as well as like their type. The data appears grouped into datasets. Next, some important aspects of a visualization project are presented, regarding the types of data and types of datasets and data preparation, a fundamental step so that visualizations can be produced.

1.1 Data types

Figure 1 presents basic types of data that can be visualized: items, attributes, links, grids and positions. An item is a discrete individual entity, such as a row in a simple data table or a node in a network. For example, items can be people, stocks in NASDAQ, stores of a retail company, countries. They are characterized by their attributes. An attribute is some specific property that can be measured, observed. Attributes can be, for example, height, price, type of product, type of shares or stores. A link is a relationship between items, generally within a network, for example the connection between two bus stops in an urban transport network, the connection between nodes in a graph, or the connection between two people in a social network. A grid indicates how continuous data is sampled in terms of geometric and topological relationships, such as where atmospheric temperature was measured, or X-ray absorption level samples are obtained from the body of a patient undergoing a Computerized Tomography (CT).

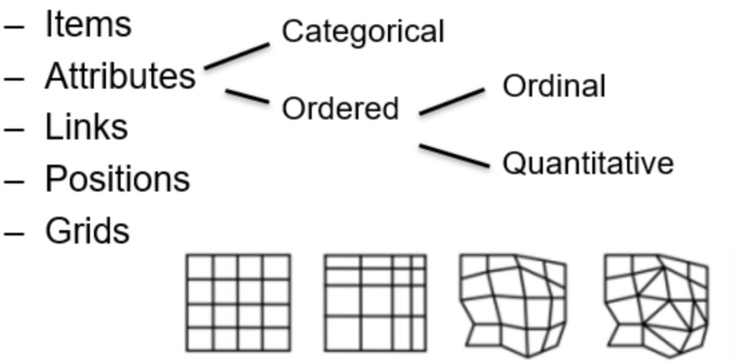


Figure 1. What may be visualized- Data types: items, attributes, links, and grids (Munzner, 2014).

A position provides a location in space, usually two-dimensional (2D) or three-dimensional (3D). It can be, for example, a latitude-longitude pair corresponding to a location on the Earth's surface or three numbers specifying a location within a region of space where X-ray absorption was measured in CT equipment.

1.2 Attribute types

A significant part of design problems in visualization are related to questions about how to visually encode attributes, (e.g., using grephical elements as points or lines, and visual properties as length and color), which is related to human visual perception and the type of attribute to be visually encoded.

Attributes can be categorical or ordered, the latter being divided into ordinal and quantitative (Figure 2). Within the latter, we also distinguish the interval and ratio attributes.

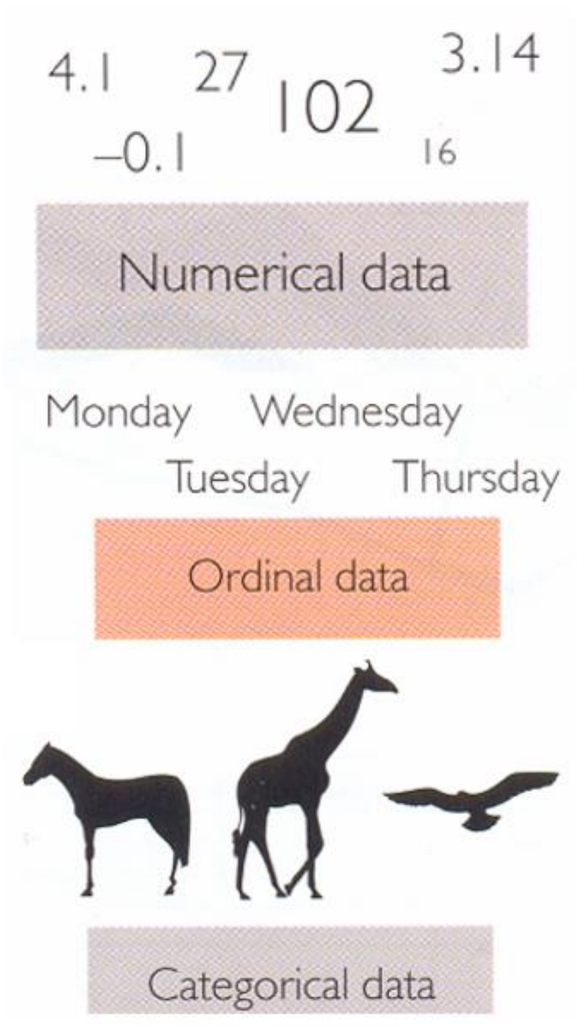


Figure 2. Types of attributes (Spence, 2014)

Categoric

Examples of categorical (or nominal) attributes are those in which the possible values are clearly different, but cannot be intrinsically ordered. This is the case of attributes such as “car brand”, “course taught at a University”, or “order code”. It makes no sense to say that “Toyota” is bigger or smaller than “Ford”, for example. Of course, they can be ordered alphabetically, but then we are ordering the words that represent the brands. The brands themselves have no order.

Ordinal

Ordinal attributes are those in which an order can be established between the different values. This does not necessarily imply that there is a numerical, quantitative value (and, in so-called “pure ordinals”, there is not). For example, clothes can come in size S, M, L or XL. It is known that a size S will be smaller than an M, and so on. We cannot, however, say that an S is twice smaller than an M. We only know that it is smaller (in fact, effective sizes vary from brand to brand and from model to model). Other examples could be “month of the year”, “day of the week”, or “satisfaction with a service”.

Quantitative

Quantitative attributes can be **continuous** or **discrete**. Although the data processed in a computational system are always discrete values, they may correspond to situations in which there is or is not the possibility of existing any intermediate values between two measurements, such as atmospheric temperatures (in which there is the possibility of having values measured with such a resolution as large as possible for the measuring instrument used), being continuous values, or a count of occurrences, in which the values only assume discrete values (e.g. the number of students per class).

Ratio

Ratio attributes are quantitative data in which there is a minimum reference value, typically zero. The difference between this and other quantitative data is that, as this minimum exists, it makes sense to estimate the proportions between different values of the attribute. A ratio attribute is, for example, the height of an object. Nothing has a negative height. Thus, in relation to a book measuring 60 cm high and another measuring 30 cm, it is possible to say without a doubt that the first is twice as long as the second. On the other hand, temperature measured in degrees Celsius is not a ratio attribute, but merely a quantitative one. If one day it is 40 ºC and the other 20 ºC, it is not possible to say that the first day was "twice the temperature" as the second. The difference (20 ºC) can be calculated, but a proportion cannot be established. Otherwise, what should one say when comparing, for example, two days with -10ºC and 10ºC, respectively? Likewise, although height is a ratio attribute, altitude is not (there are negative altitudes).

Quantitative *vs* Ordinal

Note that a numeric value can correspond to attributes of different types. For example, if a number represents the number of students in a course, its type is quantitative and it makes perfect sense to add two of these numbers; however, if the number represents a course code, it is not quantitative, but just the designation of a category that is a number rather than a text name, and therefore adding two such numbers does not make any sense.

Some confusion often arises regarding ordinal data that are "disguised" as quantitative, as they are represented by numbers: for example, consider a satisfaction survey with questions that must be answered using a Likert-type scale with five levels (as illustrated in Figure 3), which end up being coded with values between 1 and 5 in a table. In this case, it is only possible to guarantee that a person who answers 4 to a question will be more satisfied than if they answered 2, but it is not possible to guarantee that they are twice as satisfied. Although this way of coding satisfaction data is a frequent situation, an alternative coding could be used in A, B, C, D and E, which would help to avoid the “temptation” of calculating statistical parameters and applying methods, which only make sense for quantitative data, such as the mean, standard deviation or Student's t test. In the case considered, the use of the median or non-parametric tests will be appropriate. This issue is also relevant for choosing which visualization techniques to use, since there are techniques that do not make sense to be applied to ordinal data (such as the histogram).

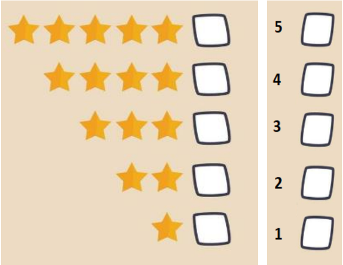
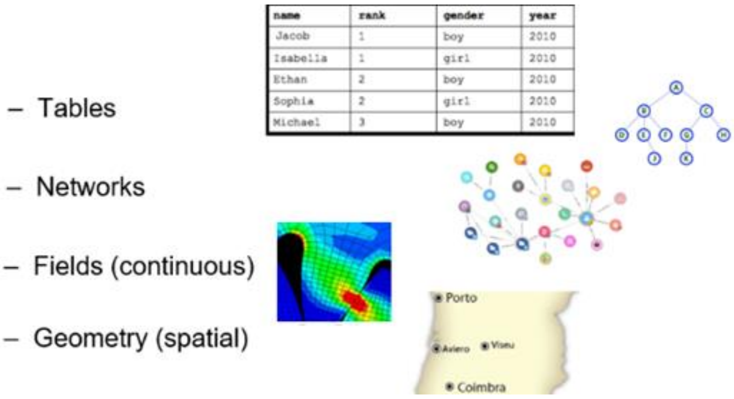


Figure 3. Alternative ways to obtain responses to a satisfaction survey using a Likert-type scale with five levels: stars on the left and, on the right, possible coding of levels in data analysis (1-5).

1.3 Datasets types

A dataset is any collection of information that is the object of analysis. The four basic types of datasets are tabular data, networks, fields, and geometry (spatial data). Other ways to group items include clusters, sets, and lists. In real-world situations, complex combinations of these basic types are common.



name	rank	gender	year
Jacob	1	boy	2010
Isabella	1	girl	2010
Ethan	2	boy	2010
Sophia	2	girl	2010
Michael	3	boy	2010

Figure 3. Types of datasets – tabular data, networks, *fields* and geometry.

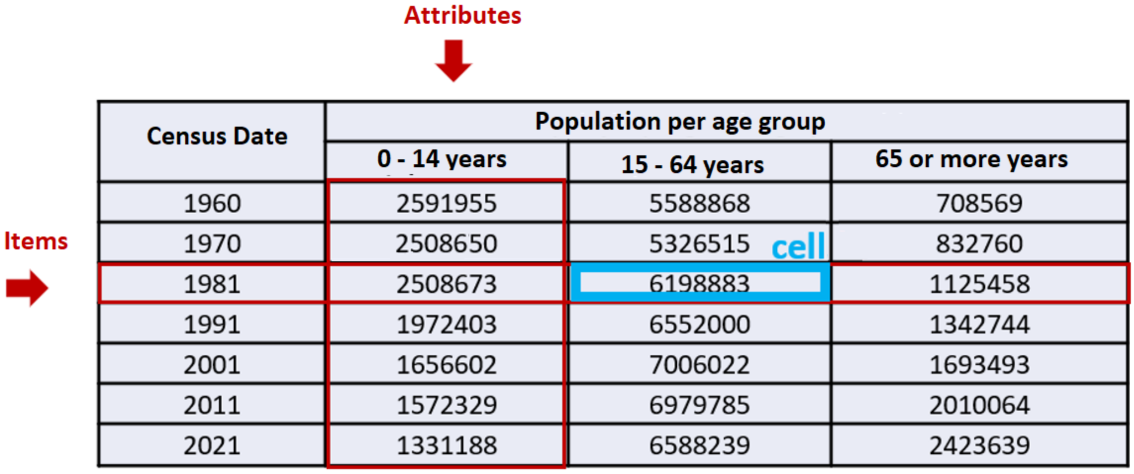
The datasets are composed of different combinations of the five types of data mentioned above: items, attributes, links, grids and positions.

Tabular Data

Often the data is organized in tables with rows of items and columns of attributes, like the simple example shown in Figure 5. Note that there are several terms that can be used to designate the rows and columns of a table. In this document “items” and “attributes” are used, respectively; however, the terms “objects” are frequently used for what is represented in the rows and “variables” or “dimensions” for what is represented in the columns.

Each cell in the table is fully specified by the combination of a row and a column - an item and an attribute - containing a value for that pair.

Temporal data can be considered a particular case in which there are several attributes measured at different moments over time, which deserves special attention since there are numerous situations in which they assume special relevance, such as in Science or Economics, with specific statistics methods, as well as visualization techniques, especially suitable for data of this type.


Attributes
Items
cell

Census Date	Population per age group		
	0 - 14 years	15 - 64 years	65 or more years
1960	2591955	5588868	708569
1970	2508650	5326515	832760
1981	2508673	6198883	1125458
1991	1972403	6552000	1342744
2001	1656602	7006022	1693493
2011	1572329	6979785	2010064
2021	1331188	6588239	2423639

Figure 5. Tabular data: simple table with items, attributes and cells.

Networks and trees

Networks specify relationships between items generally called nodes or nodes; This book adopts the designation of links and nodes for these concepts, respectively. For example, in a metro network (Figure 6, on the left), the nodes are the stations and the links represent the line sections that connect them. The term "graph", used in some domains, also designates a set of nodes with connections between them. There are many other situations in the real world that can be considered networks, such as road networks, computer networks or social networks. Nodes and links can have attributes.

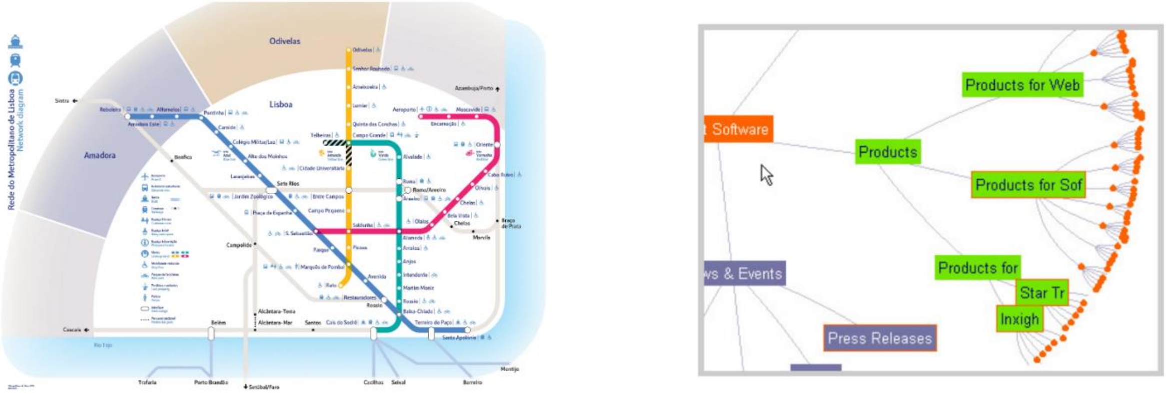


Figure 6. Examples of visual representations of a network (left) and a tree (right)[1]

Trees, which are a particular case of network in which there is a hierarchical relationship between nodes, do not have cycles: each child node has only one parent node connected to it, although each parent node may have several child nodes (Figure 6, right). Company organizational charts or directory trees on a computer are examples of real-life situations that can be well represented by trees. There are appropriate visualization techniques specific to represent this data.

Continuous fields

Fields are sets of cells with attribute values corresponding to measurements obtained directly, or by calculation, from a continuous domain (in which there is the possibility of obtaining measurements on an infinite number of points, which means that it will always be possible to obtain a measurement, between any two other existing measures). Continuous fields represent continuous phenomena that can be measured in the physical world or

[1] https://www.metrolisboa.pt/viajar/mapas-e-diagramas/
http://www.visualcomplexity.com/vc/project.cfm?id=137

simulated using simulation software. Possible examples are the air speed around a mechanical structure in a wind tunnel, real or virtual, the temperature and pressure in a given area of the Earth's atmosphere or the absorption of X-rays in a complementary diagnostic exam obtained from a CT scan. The measurements obtained can be scalar, vector or tensor (for example, atmospheric temperature or X-ray absorption, air speed in the wind tunnel and mechanical deformation of a structure, respectively), corresponding to scalar, vector or tensor fields, and can be obtained and visualized in a space with three dimensions (3D), or with two dimensions (2D), as exemplified through visualizations in Figure 7.

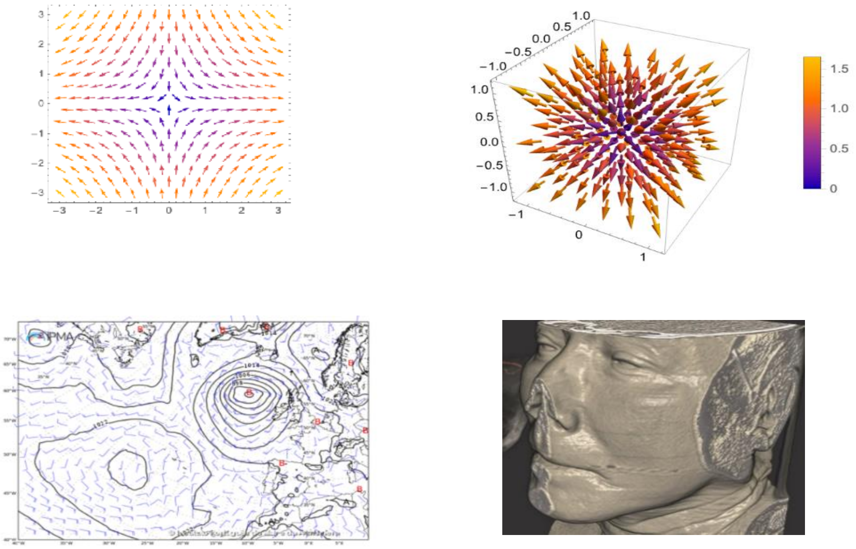


Figure 7. Visualization of continuous fields (from left to right: vector 2D, and scalar 3D)[2]

In these cases, each cell is associated with a specific region of space and the cells may have larger or smaller intervals between them, resulting in a more or less sparse sampling, producing a dataset with lower or higher resolution. Mathematical functions can also be continuous and be represented by continuous field datasets (as illustrated through visualizations in Figure 8).

[2] https://reference.wolfram.com/language/howto/PlotAVectorField.html;
https://www.mevislab.de/
https://www.math.umd.edu/~jmr/241/surfaces.htm;
https://www.ipma.pt/pt/otempo/prev.numerica/

When a continuous field dataset is obtained through sampling at completely regular intervals, as in the case of a CT scan, the cells are organized according to a uniform grid, there is no need to explicitly store the grid geometry. However, there are cases where sampling is not uniform (for example, it is not done in a regular way, or it is done along a curvilinear shape), which implies storing more information about the geometry and topology. In either case, it is necessary to consider the mathematical issues of sampling, how often to take measurements, what type of interpolation will be possible and how values can be shown between the sampled points in a way that does not lead to error. These issues are studied in areas such as Signal Processing and Statistics, but must be considered when visualizing data of this type, being fundamental in Scientific Data Visualization (SciVis).

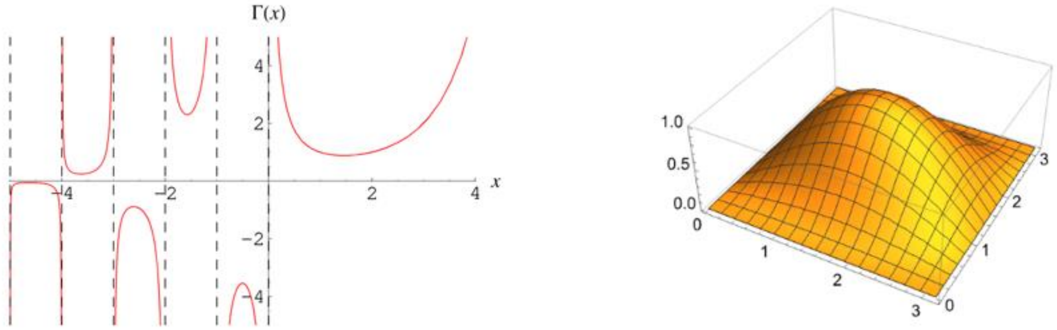


Figure 8. Examples of visualizing one and two variables functions (from left to right)[3]

Geometry and spatial data

This type of dataset contains information in the form of items with explicit spatial positions: points, one-dimensional lines or curves, 2D surfaces or regions, or 3D volumes. Being intrinsically spatial, they are generally used for tasks that require understanding the shape (as is also common in the case of continuous fields). These datasets do not necessarily have attributes and generally become interesting in Visualization only when they are obtained in a way that requires consideration of design choices (such as isobar or isotherm lines on a

[3] https://mathworld.wolfram.com/GammaFunction.html
https://reference.wolfram.com/language/howto/PlotFunctionsOfTwoVariables.html

weather chart or burned forest contours generated from a field, in the first case, and from satellite images in the second). In most cases, this data serves as a reference on which other data is visualized, such as, for example, the geometry of an airplane wing on which the pressure exerted by the air, or a map on which a drought index are represented as illustrated in Figure 9). The datasets corresponding to sets of positions on the Earth's surface, which are generally called georeferenced data, were among the first to be visualized and are currently produced in large quantities by sensors, being especially interesting in Information Visualization; also, for this type of data there are specific adequate visualization techniques.

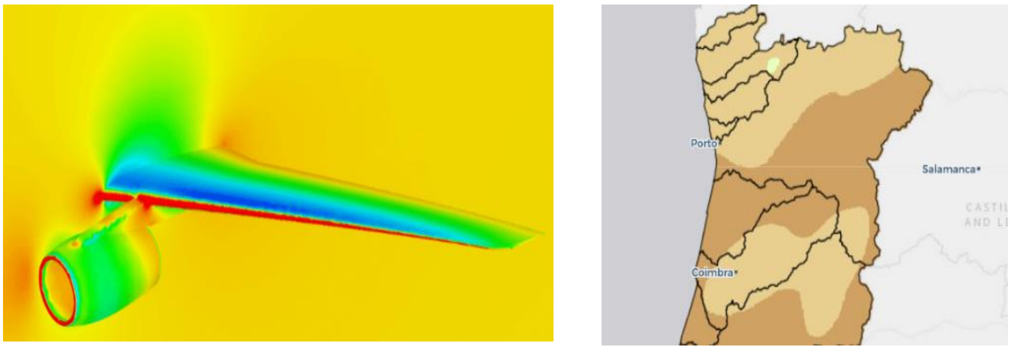


Figure 9. Examples of using geometry to visualize other datasets[4]*:* pressure around an airplane wing (left) and the PDSI (*Palmer Drought Severity Index*) in Portugal (right)

2. Data preparation

Before making any decision regarding how to visualize a dataset, it will be necessary to ensure that the data is in the appropriate form to be visualized. Data preparation is a fundamental phase in a visualization process; if it is not properly carried out, it compromises the usefulness of the result, since not even an excellent visualization project can overcome poor quality data. However, the work that this preparation involves is often underestimated in terms of the detail required and can involve a large amount of work that was not foreseen at the outset, ending up unrecognized. This phase can easily consume a very significant part of the resources

[4] https://www.simscale.com/docs/tutorials/tutorial-compressible-flow-simulation-around-a-wing/
https://www.ipma.pt/pt/oclima/observatorio.secas/

needed in a visualization project, whatever its size. Therefore, data integrity is essential, even more so when the volume of data increases and it becomes necessary, for example, to update, aggregate or use it for calculations and it is sensible to consider that if you do not have access to data already properly prepared, it will take much more time than anticipated to prepare them. In more sophisticated organizations, data may already be made available in a suitable form to be directly used in a visualization project, but in many situations, this does not happen and it becomes necessary to prepare it before proceeding with visualization.

The following are the main types of actions that can be carried out in the data preparation phase: first, cleaning and then transformation, which may involve deletion, obtaining derived data, coding, aggregation or normalization.

2.1 Data cleaning

The data generally contains typing errors or inconsistencies, does not comply with a standard, or does not make sense considering the reality it represents, so it needs to be "cleaned" before starting to analyze it in more detail. Some cases could be, for example, finding in a service user register, an address in the city of Lusboa, a user with a height of 2.85m, or a telephone number with an incorrect number of digits. However, sometimes values that appear abnormal may not be so and care must be taken to confirm before correcting. Basically, only with knowledge of the domain to which the data refers is it possible to clean it. Care must also be taken to keep an updated record of all corrections made so that it is possible to identify any problems that have occurred.

Outlier detection

There are numerous techniques that allow obtaining an indication of when a value requires special attention. One of the most common is the detection of outliers, values that "fall outside" the most usual range for a given attribute. A simple ways to find an outlier is**:**

- Calculate the quartiles of the data: order them and find the value that leaves behind 25% (Q1), 50% (Q2 or median) and 75% (Q3) of the various values in the set;
- Calculate interquartile range (IQR = Q3-Q1);
- Identify as outliers all values that are more than 1.5 x IQR above Q3 or below Q1.

Other more sophisticated forms may be better applied to specific cases.

Missing data

Another common problem is that datasets are incomplete, with items in which some of the attributes are missing. There are five recurring ways of dealing with this missing data.

One possible way is to simply delete the record in which the value of an attribute is missing. We can also not delete it, but assign a sentinel value to the attribute in question. This will be a value that is guaranteed not to be a possible value for this attribute (a height of -1 meter, for example). This will help us, when creating the visualization, to be sure that this value was missing and to be able to show it (or not) appropriately.

A little more sophisticated way would be to assign to the missing attribute the average value of all values of that attribute in the rest of the dataset. The underlying logic is that this will be a typical value and therefore a suitable replacement.

In a slightly more informed way, we can assign the value of that attribute in the nearest item to the missing attribute. This makes sense in the case where there is a correlation between the values of the various attributes considered. Imagine, for example, a dataset whose items are people and two attributes are their height and the size of the shoes they wear. It is known that, although the match is not perfect, taller people wear bigger shoes. Imagine, now, that we do not know the shoe size of a given person, who is 1.75m tall. It is not unreasonable to assume that this will be the same as someone else having also 1.75m. It may not be the exact value, but it probably won't be far from it. In more complex datasets, with more attributes, the similarity between items will be stronger and this method will have better results.

Finally, a more general way will be through value assignment, in which, using more or less sophisticated statistical methods, it is decided which value to use. In general, the choice of which method to use depends on the domain in question. In different cases, different methods will make more or less sense.

2.2 Data transformation

After “cleaning” the data, this phase focuses on organizing it in a way suitable for the visualization project. Examples of some possible transformations are:

Elimination - taking into account the objectives of the analysis, some of the observations may not be necessary or deviate from the pattern of the rest (such as outliers) or some attributes may only be needed to calculate the derived data, so they can be eliminated (for example if it is only important to analyze the evolution of average temperatures or daytime and monthly temperature ranges in a set of locations, it will probably not be worth maintaining the maximum and minimum daytime temperatures in the locations). However, it is good practice to investigate the appropriateness and impact of deletion beforehand.

Coding - if the data includes answers to open-ended questions it will be necessary to analyze the answers and create a coding (for example, Lisbon, Porto or Algarve region as preferred national tourist destinations).

Aggregation - if the level of detail is excessive for the analysis in question, it may be necessary to aggregate the data (for example, North, Center and South, in the case of national destinations, or by season of the year and not by week).

Standardization - some may be necessary to obtain coherent data when coming from multiple sources or to conform to conventions (e.g., using degrees Celsius or Fahrenheit to measure temperature values).

In addition to these basic transformations, there are other relevant issues that must be considered before making decisions regarding how to visualize the data, such as **dimensionality reduction** (i.e., what to do when there are too many attributes). Finally, the importance of this preliminary step for the success of any visualization project is paramount.

Derived attributes

It is often necessary to create derived attributes, calculated from those originally existing in the dataset, and add them as new attributes. This is the case of calculating average temperatures per week or month, or the thermal amplitude of the various days considered in a dataset.

The calculation of derived attributes often arises from a deep reflection on what you really want to visualize. Consider a dataset that consists of a person measuring their weight daily over the course of a few months. The constant attribute in the dataset is, therefore, the weight. However, what this person really wants to see is not the weight itself, but whether they are gaining weight or losing weight. That said, there is a derived measure whose visualization and analysis is much more suitable for this objective: not the weight itself, but the variation in weight compared to the first day. This will allow a direct assessment of the increase or decrease in weight, explicitly visually coded.

In short, it is important to reflect on the questions that really need to be answered. Often, the attributes originally present in the dataset are not the ones that are used in the visualizations.

Summary

In this document what can be visualized is discussed and the main characteristics of the data that are relevant for choosing the visualization techniques to use in each case are introduced.

A possible classification of the main types of data and datasets that can serve as a basis for organizing the presentation of visualization techniques is presented. The basic types of data that can be visualized - items, attributes, links, positions and grids – are also described as well as the main types of datasets in which data can be organized - tabular data, networks and trees, continuous fields, geometry and spatial data; examples of visualizations of these types of datasets are given. Finally, the importance of the issue of data preparation is stressed, and some of the most used methods are briefly addressed.

Main bibliography

Camões, J. (2018). *Data at Work: Best practices for creating effective charts and information graphics in Microsoft® Excel®*, O'Reilly.

Fisher, D. and Meyer, M. (2018). *Making data Visible*, O'Reilly.

Kirk, A. (2019). *Data Visualisation A Handbook for Data Driven Design* (2nd Ed.). Sage.

Munzner, T. (2014). *Visualization Analysis and Design*. A K Peters/CRC Press.

Spence, R. (2014). *Information Visualization, An Introduction*. Springer.