Introduction

This book contains the infographics course’s lessons held at the Free University of Bolzano Bozen.

This book is in continuous development, please take a look at its version number, which marks important changes.

Disclaimers

This book is designed for novice data analyzers. It contains oversimplifications of reality and many technical details are purposely omitted.

Table of Contents

INTRODUCTION ....................................................1
TABLE OF CONTENTS........................................... 1
1. BASIC NOTIONS ........................................... 2
   1.1. FUNDAMENTAL SCOPES ........................... 2
   1.2. VARIABLES .............................................. 3
   1.3. SCALES .................................................. 3
2. TRADITIONAL GRAPHS .................................... 5
   2.1. ONE NOMINAL VARIABLE ......................... 5
   2.2. ONE SCALE VARIABLE ............................... 6
   2.3. NOMINAL VERSUS NOMINAL VARIABLES ....... 8
   2.4. NOMINAL VERSUS SCALE VARIABLES .......... 9
2.5. SCALE VERSUS SCALE VARIABLES .................. 10
3. SPECIFIC CHARTS ......................................... 12
   3.1. DATA MAP .............................................. 12
   3.2. TIME SERIES .......................................... 14
4. MISLEADING FACTORS ...................................... 18
   4.1. DISTRACTIBLE ELEMENTS ....................... 18
   4.2. ELEMENTS IN THE WRONG PLACE .............. 18
   4.3. WRONG PROPORTIONS BETWEEN DATA AND SIZES 19
   4.4. DIFFERENT UNITS OF MEASURE .................. 20
   4.5. NOT STARTING FROM ZERO ....................... 21
   4.6. TWO AND THREE DIMENSIONAL GROWTH ....... 21
1. Basic notions

1.1. Fundamental scopes

Every time we need to represent anything, being a picturesque view through a painting, numbers through a table or, as it will be the case in this book, numbers through a picture, we must never forget what is the scope of our work. To help us in this task, the following questions can give good hints. They can be used every time we are faced with a representation to point out the bad aspects and to find out possible corrections and improvements.

What is represented?

Representing something without having a prior knowledge of what is going to be represented is clearly pointless, but there are numerous examples of representations built up only to impress the viewer, without any intention of really display something.

The topic which is displayed must be clear in the designer’s mind and also the point which wants to be underlined must be clear. All the image must be dedicated to these two tasks without adding worthless extras and without depriving these two points of any of their aspects for other secondary scopes.

Is something misleading?

The most widespread problem in data visualization are misleading graphs, which deliberately or by ignorance display information in a wrong way to attract the viewer’s attention to the scope of the representation in an unfair way. The typical example of error are wrong proportions, displaying elements of the picture not in relative scale to each other. This may seems a childish trick, easy to discover; however, when it is done using multidimensional pictures it needs a careful eye to be spotted, a can be seen in section 4.

Is something useless?

Too many graphical representations incorporate elements which do not have anything to do with the main scope, with the only task to embellish the graphics. Very often these elements distract the attention, when not moving the viewer’s focus completely away. In some cases they make parts of the representation incomprehensible, as can be the case when fancy fonts are used to write labels and numbers.

Is something missing?

Very often a graphical representation needs also text and numbers. Even though most of it can be stored in the caption, it is necessary that axes contain the units of measure and their scale, that colors contain a proper legend and that, in case useful information is omitted, it be explicitly cited and explained.

A scaleless or unitless graph is useless even when all elements to compare are in scale, any difference which seems huge can instead be irrelevant. A representation with missing relevant details leads immediately the suspect that the designer wanted to hide something.

Last but not least a time indication is fundamental. Data will change in time while the representation stays the same (unless it is fully interactive) and can be looked at even many years after. With a missing time indication the representation is worthless, it can be referred to right now as well as to 3 millennia ago.

Can other information be added?

This question has two aspects. The positive one is checking whether the graph can support other useful details to be stored inside without driving attention away from the main points and without decreasing its comprehensibility. The negative one is making sure that no other useless information is stored inside the graph, better keeping it simple than overloading it with useless elements.
1.2. Variables

We follow Stevens theory\(^1\) and we distinguish the data that we have in three big categories, which are treated in different ways both by statistics as well by graphical representations.

Scale variables, also called numeric variables, are data expressed in a numerical format, be it continuous or discrete, on which mathematical operations make sense. They are usually physical measures, such as time, age, weight, distance in Kilometers, length, number of children, GDP.

Nominal variables, also called categorical variables, are data which represent categories. No mathematical operation makes sense on them, even though sometimes numbers are used to identify the categories. Their only scope is dividing the population into categories. For example, country of residence, sex, degree course. Even though scale variable could divide the population into categories, the fact that they have many possible values, even infinite values when the variable is continuous, using them to divide the population into categories would result in a huge number of categories, often with one or zero cases in each category.

The border between nominal and scale variables is however not so straight. And in the midway we find ordinal variables, which define only an order on the thing they are representing. For example, the arrival position in a running competition is clearly ordinal, as it defines the order but no mathematical calculation can be done (we cannot say that the sixth one is 3 times the second one). A typical example is the 5-elements Likert scale used in questionnaires, a set of predetermined answers going from very negative, indicated with 1, to very positive, indicated with 5. It is clear that answering 4 is always more positive than answering 2, but we cannot say that it is twice as positive.

Ordinal variables can always be used as nominal variables, as they have a limited number of values and they divide the population into categories. Sometimes, when they have several values and when the numbers user to represent the order can somehow also induce a mathematical meaning, they are used as scale variables.

1.3. Scales

Unless extravagant solutions are absolutely necessary, there are two scales which are typical in representations of numerical data.

The linear scale is the standard choice, every interval represents a distance which is proportional to its length, with this proportion fixed for the entire graph. It is often a good idea to include also the zero level in this scale, in such a way that the viewer is not mislead into thinking that apparently huge differences are really so big, as in section 4.5. When the zero level is so far away and it is not useful for the problem (for example, because the problem itself concentrates on the differences and not on the absolute values), it is necessary to render very evident the fact that the scale does not start from 0, using graphical tools or text.

The alternative scale is the logarithmic scale, where each interval represents a distance which is exponentially proportional to its length. In the logarithmic scale the logarithm of the values are displayed, but on the axis still the original value is indicated, not its logarithm. Given the features of logarithm, no value can be zero nor negative, as its logarithm does not exist. Traditionally logarithmic scale in base 10 is used, as it is more human comprehensible, but it is equivalent to use any other base such as 2 or e. It is vital to indicate on the axis at least three values, to give the viewer the correct impression that numbers are growing exponentially. In logarithmic scale the usual starting value is 1, as 0 is not representable.

Long-term real growth in US GDP (Log Scale)

GDP adjusted for inflation (2001 dollars) 1871-2010

Trendline (exponential growth rate)

Data from MeasuringWorth.com

VisualizingEconomics.com
2. Traditional graphs

How one or several variables can be graphically visualized depends strongly on the variable type.

2.1. One nominal variable

To represent one nominal variables there are two basic distinct choices and a bunch of variations which are technically identical.

The first choice is the **column plot**, sometimes improperly called histogram. It represents on the horizontal axis the variable’s categories and on the vertical axis the frequency of cases in each category.

![Column Plot Example]

Millions of patent applications in Japan, US, Europe between 1998 and 2007, source WIPO.

The second choice is the **pie chart**. It depicts a circle divided into slices, each one proportional to the number of cases in that category. Comparing it to the column plot we observe that here the visual information on the absolute number of cases is lost, while strong emphasis is put on the percentage of cases.

The third choice is the **radar graph**, which is a good choice when the number of categories is at least five and when they have a circular meaning, such as months, week’s days, hours, angular directions such as N, NE, E, SE, etc. Since it is considered to be a very fancy graph, it is often abused and applied also to variables which are not circular at all, especially in marketing dimensions.

![Radar Graph Example]

Number of shops open per week day (invented data).

Possible variations of the previous two basic type are:

- the **bar plot**, a version of the column plot with horizontal bars;
- the **line plot**, which consists in a column plot where a line connects the top of each column. It must not be confused with a mathematical graph, as on the horizontal axis there is not a numerical variable but

![Line Plot Example]

Line plot of percentage of people who use an App more than 100 times a month with respect to total Apps’ users, source Flurry Analytics
simply categories. Special care must therefore be taken when using it and it is best suited when an ordinal variable appears on the horizontal axis, especially if it is a time related variable;

- the area plot, which is identical to the line plot but has the area below the line colored. This is many times a misleading element, as it gives the impression that values between two categories really exist (they do not, on the horizontal axis we have only categories not a continuous variable);

- all the three dimensional versions of these graphs are also commonly used. However, unless there is a strong necessity for an extra dimension, this is only extra graphics which moves the focus away from the main point and confuses the viewer.

### 2.2. One scale variable

Two graphs are used to represent the distribution of one scale variable, both requiring grouping procedures.

The most famous is the histogram. It is built deciding a priori a subdivision into intervals of the possible variable’s values and then counting how many cases falls into each interval. The result obviously depends on the choice of intervals, which should be chosen wisely, as too few intervals would result in a useless histogram, while too many intervals would produce an up and down histogram with many 0 cases or 1 case intervals.

Histogram with 5 intervals for 100 cases (invented data)  
Histogram with 50 intervals for 100 cases (same invented data)
Intervals must have the same length. For example, if we build an histogram with an interval 3 times wider than the others it contains the cases which should be in 3 intervals and thus it is high as the sum of three intervals, something which is clearly misleading. Moreover, it is also 3 times wider, which gives the visual wrong impression of an bigger importance for that part.

Less famous, but in some cases much more efficient, is the boxplot. It is a visual representation of the two central quartiles through a rectangle cut by a black line representing the median. It is thus immediately evident where the central 50% of the values’ distribution is. Then it depicts two T-shaped whiskers which usually indicate where the central 90% of the distribution lies. Other values, called outliers or extremes, are indicated with symbols.

One scale variable has also another family of representations which depicts a completely different aspect. Whenever we are interested in displaying the values case by case and not the distribution of all the values, we can use the column plot (with all its variations) or the radar plot with cases instead of categories and variable’s values instead of frequency of cases. This kind of representation is perfectly equivalent, also from the visual power’s point of view, to writing a table with the numeric values and works only when the number of cases is small. Very often in this case the pie chart is totally inappropriate.
<table>
<thead>
<tr>
<th>Country</th>
<th>GDP per capita (USD) in 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>14,991,300,000,000,000</td>
</tr>
<tr>
<td>China</td>
<td>7,203,784,000,000,000</td>
</tr>
<tr>
<td>Japan</td>
<td>5,870,357,000,000,000</td>
</tr>
<tr>
<td>Germany</td>
<td>3,604,061,000,000,000</td>
</tr>
<tr>
<td>France</td>
<td>2,775,518,000,000,000</td>
</tr>
<tr>
<td>Brazil</td>
<td>2,476,651,000,000,000</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>2,429,184,000,000,000</td>
</tr>
<tr>
<td>Italy</td>
<td>2,195,937,000,000,000</td>
</tr>
</tbody>
</table>

### 2.3. Nominal versus nominal variables

The representations of two nominal variables are variants of the one nominal variable graphs. Usually the one with the more categories is displayed as before, while the categories of the other one are represented side by side, line by line using different colors (in case line or radar graphs turns out to be appropriate) or also one above the other, when it makes sense to sum the categories and thus give also an information on the sum for each category of the first variable.

In this case it makes sense to use a 3D graph, where the second nominal variable is represented on the third dimension. However, still the information on the exact proportions is lost.

Average session frequency per month of smartphone usage by operating system, source Flurry Analytics

Sometimes a stacked area graph can be really meaningful. Smartphone shipments by operating system, source www.gartner.com

2.4. Nominal versus scale variables

Also the nominal versus scale representations are variants of the scale variable representation, with the nominal variable’s categories depicted side by side or one over the other. For histograms a special warning is necessary: when the number of cases in each category is strongly different, rescaling the vertical axes for visibility purposes can be misleading, while not doing it can squeeze a distribution to the horizontal axis.
Boxplots of daily log returns at NASDAQ in 2012 for different companies, source www.portfolioprobe.com


2.5. Scale versus scale variables

The traditional way to represent two scale variables is the scatterplot, a graph with two numerical axes where each case is represented with a symbol. Whenever the cases are in horizontal axis values’ order and whenever this order makes sense, a mathematical graph can be built joining the symbols with lines. This lines must not be confused with the regression line (or regression curve in case of a non-linear regression model), which is a statistical tool commonly used in combination with a scatterplot to forecast non-observed values.

Scatterplot of number of medicines used daily per 1000 inhabitants and average cost of daily medicine dose. Cases are regions, source “Relazione sanitaria – Provincia autonoma di Bolzano – 2005”
When the number of scale variables grows to three, surface plot is borrowed from mathematics to display the effect of two variables on a third one. It is a tridimensional version of scatterplot (which would be incomprehensible in three dimensions) with lines and thus works only when the variable on the vertical axis is a function of the ones on the horizontal axes and if the proper rotated view is adopted.

2.5.1. Bubble charts

When the two variables on the horizontal axes do not share the same unit of measure, as for example population and GDP, it is more effective to use another representation, known as bubble chart. In this representation one of the scale variables is displayed increasing the size of bubbles, either the diameter or the area (even though a two dimensional distortion comes into play, see section 4.6). Moreover, colors can be used to represent a fourth nominal or scale variable.

Website www.gapminder.org has a tool to display several bubble charts which move with time, thus introducing also another dimension. However, it works only on their data.

In order to produce bubble charts on personal data Excel or Google Developers (https://developers.google.com/chart/interactive/docs/gallery/bubblechart) can be used.
3. Specific charts

Several fields make use of particular variables for which data have well-known proprieties, which can be used to display particular graphs which turns out to be more appropriate and more understandable by expert eyes than standard ones.

3.1. Data map

When one of the involved variable is a geographic location it can theoretically be converted into two scale variables using its geographic coordinates. However, instead of using methods for three or more scale variables, we can take profit of the existence of a well-known map, immediately recognizable by the views, and depict the chart on the map itself. If the other variable is a scale variable, we can build a color scale, while if the other variable does not exist we can build practically a scatterplot using symbols to mark the geographic locations but drawn above a map.

![Map of Cancer Mortality Rates by State Economic Area](source_annals INTERNAL MEDICINE Vol. 133, N. 2, Pages 161-162)

Other elements can be added to data maps, using the fact that we can display a proportion using the element’s size as well as a direction using the map’s orientation. A typical example is the flows of traffic from various points of the map. Whenever displaying relations among elements of the map is displayed on the map itself, we must unavoidably cut those elements which are too small, as otherwise the number of relations is 2 to the power of the number of elements and can easily make the entire picture incomprehensible.
3.1.1. Statsilk software

A very efficient free software which gives the opportunity to load a map below and data from a spreadsheet, displaying them over the map in a variety of ways including side graphs is developed by Statsilk [www.statsilk.com](http://www.statsilk.com) and comes in three packages:

- StatPlanet, a web-based and desktop software for visualizing spatial data through interactive maps, graphs and charts
- StatTrends, a web-based and desktop software for visualizing any kind of data through interactive graphs and charts
- StatWorld/StatPlanet, a web-based and desktop software for exploring world statistics from various international organizations through interactive visualizations.

These programs are interactive, meaning that graphs’ proprieties can be changed with a simple click and web-based, meaning that they can work through a browser connected to Statsilk’s website.
3.2. Time series

3.2.1. Financial chart

Financial data use scale variables time, low price, high price, open price and close price and the main interest is usually depicting prices as a function of time. They have the property that low is always smaller than all the others and high is always larger than all the others. Therefore they can be represented using a candlesticks chart, where time is on the horizontal axis, the central rectangle has values open and close as extremes (and is white or green when open is the base, black or red then close is the base) and the vertical line goes from low to high. In this way five different scale variables can be efficiently represented together. Sometimes traded volume or other indicators is depicted as a daily column plot below the chart, using the same horizontal time line.

Special care must be taken when producing or observing a financial graph for missing values, either because the security is not traded (market closed or security suspended) or because of data errors; in cases like this it is preferable to leave a blank space if just one of the four prices is missing.
A similar representation is typical for weather, as is shares the high and low of financial data, as well as the presence of other indicators such as precipitation and humidity.

3.2.2. Gantt chart

A very alternative way to display a starting and ending time of an event which moves ahead (in space or moves on with its planned activities) is using one dimension as time and the other one as moving space, connecting the starting and ending points with a line. The steepness of the line indicates how fast is that part proceeding.

When dealing with project’s planning, this graph uses bars and takes the name of Gantt chart, in honor of Henry Gantt.
3.2.3. Narrative graph

Figurative map of French army in Russian campaign 1812-1813 drawn by C.J. Minard in 1869 is the fundamental example of a narrative graph. This graph manages to depict in a very efficient way on the same chart several scale variables: the coordinates of the army’s position including all possible branches, the army’s size and the weather temperature. Moreover, it includes useful elements such as the main battles and the crossed rivers. Everything which is relevant for the campaign is reported, with the only exception of the timeline which is only sketched for the retreat and is not in scale (as the horizontal axis is in geographical and not temporal scale).

Other elements can be represented on a figurative map, bearing in mind that if the horizontal axis is dedicated to the timeline then the geographic information must be squeezed in the vertical axis only or inserted through other means.
Figurative map of the men losses of the French army in the Russian campaign 1812-1813, source C.J. Minard 1869.

A new chart of history, source Lectures on History and General Policy by J. Priestley 1788.
4. Misleading factors

4.1. Distractible elements

Any element which is not directly related to the focus of the representation should be avoided, especially when it is only an embellishment of the chart. The typical example is inserting pictures which do not help comprehensibility inside bar plots.

As can be seen in these two pictures, the elements added in the column bars do not help the comprehensibility at all. They do not represent, as they could be, neither the years nor the three analyzed variables in the second graph, where it could be much better displaying a picture related to the variable itself. Their only task is to distract the viewer in the first one with the help of a three-dimensional object (see section 4.6) and in the second one from the non-zero start of vertical axis (see section 4.5).

![Revenue Growth Forecast!](Source Beautiful Evidence by E. Tufte.)

![Operating Revenues, Net Income (Loss), Exploration & Development Expenditures](Source Day Mines 1974 Annual Report.)

In this picture instead a fancy left to right perspective effect has been introduced. Its effect on a better comprehensibility of the data is none, while it makes comparing the bars a very difficult task and gives the impression of a much larger increase.

![Graph Example](Black columns represent New York State aid to localities in billions of USD, total column represent total budget, source New York Times 1 Feb 1976.)

4.2. Elements in the wrong place

The position of elements must be planned with special care, as the viewer can induce that it contains a special meaning. In case it does not, this must be evident. For example, in one of the versions of the worldwide famous billion dollar gram, which depicts some of the world’s expenditures on a two dimensional scale. The graph is without any doubt very impressive and efficient in comparing some
expenses to others. However, apart from the impact on the viewer, there is no need to use two dimensions, one dimension is enough as the variable to be displayed is only one. The only advantage in using two dimensions is the possibility to easily slip insert some rectangles inside other ones. However, this automatically induce into the viewer the idea that a rectangle is part of the outside one, as it is correctly depicted for Online Advertising as part of Worldwide Advertising Spend. However, the viewer is expected to see the same for Iraq War, but the three rectangles are not in the total one, giving the impression of a larger expense.

4.3. Wrong proportions between data and sizes

Respecting proportions between numerical data and graphical elements’ sizes is one of the basic principles of any visual representation. In the billion dollar gram the proportion to be respected is between numerical data and areas, but the Facebook area, which correspond to 15 billion USD, should be less than one third of Bill Gates area, which evidently is not as can be estimated with a simple ruler. The picture below shows instead a billion dollar gram where no evidently wrong proportion or no wrongly placed element seem to appear.
4.4. Different units of measure

Data with a different unit of measure may never be put one aside the other, unless their units of measure is very clear or unless they are properly rescaled. Apart from evident units of measure’s difference, in the examples below the last category contains a shorter sample’s unit, which is evident only in the second picture. In both cases a rescaling would have been very appropriate.
4.5. Not starting from zero

Starting the axis from a number which is not zero is a legitimate choice whenever it is necessary to emphasize small differences (and using logarithmic scale is not appropriate), but it must be very evident and possibly underlined. In the second example of section 4.1 the three column plots not only do not start from zero without any notice nor number on the vertical axes, but they start from a largely negative number without any necessity, since the variables represented by the first and the third plots cannot have negative values, and the three plots have a different scale and zero level. This also hides the fact that the first column of the second plot is negative.

4.6. Two and three dimensional growth

Whenever a single scale variable is represented using a two dimensional image special care must be taken in the proportions. The viewer is faced up with two different scales: the vertical size and the area size. Only if the horizontal base is kept constant, it is possible to respect both the proportions among data and vertical sizes and areas. If also the horizontal base of the graphical element is rescaled, then the area becomes obviously proportional to the square of the data, as can be seen in the man in the first picture of section 4.1, and in the two pictures below. In the first one, the fact that the element is also a three dimensional barrel introduces an even further misleading effect, as the perceived volume is the cube of the linear size.

A three dimensional growth on an abstract element is displayed in the next picture, where a column of a standard column is exploded into a three dimensional object which clearly gives the impression of being much taller and with a larger volume than the two dimensional ones.

Much subtler is the growth effect of the next second graph, which depicts the volume of water and fresh water compared to Earth’s volume. The volume’s proportions are probably correct, but what is misleading here is that the vast majority of water is scattered on 20 Km of Earth’s surface depth, which is exactly the place where water is needed. So it should compared to the surface of Earth, eventually multiplied by 20 Km to have a volume, and not to the entire Earth’s volume as water will never flow deeply inside Earth. Looking at any terrestrial globe immediately shows that water is far from scarce on Earth’s surface.