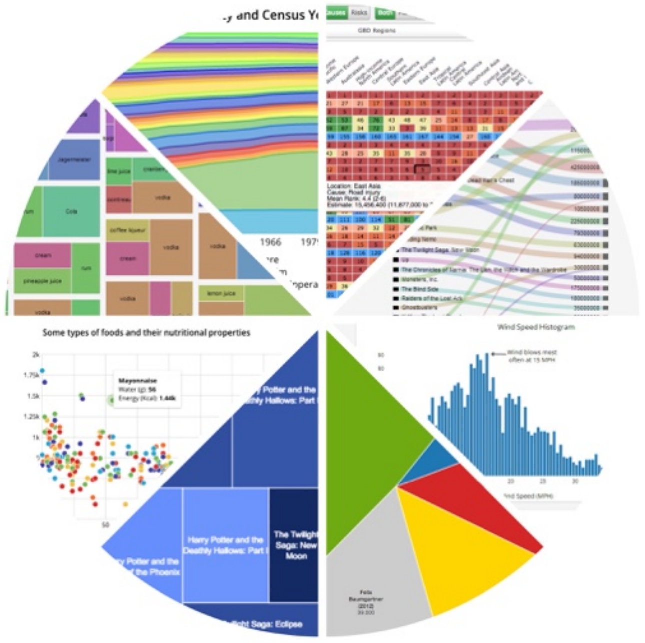
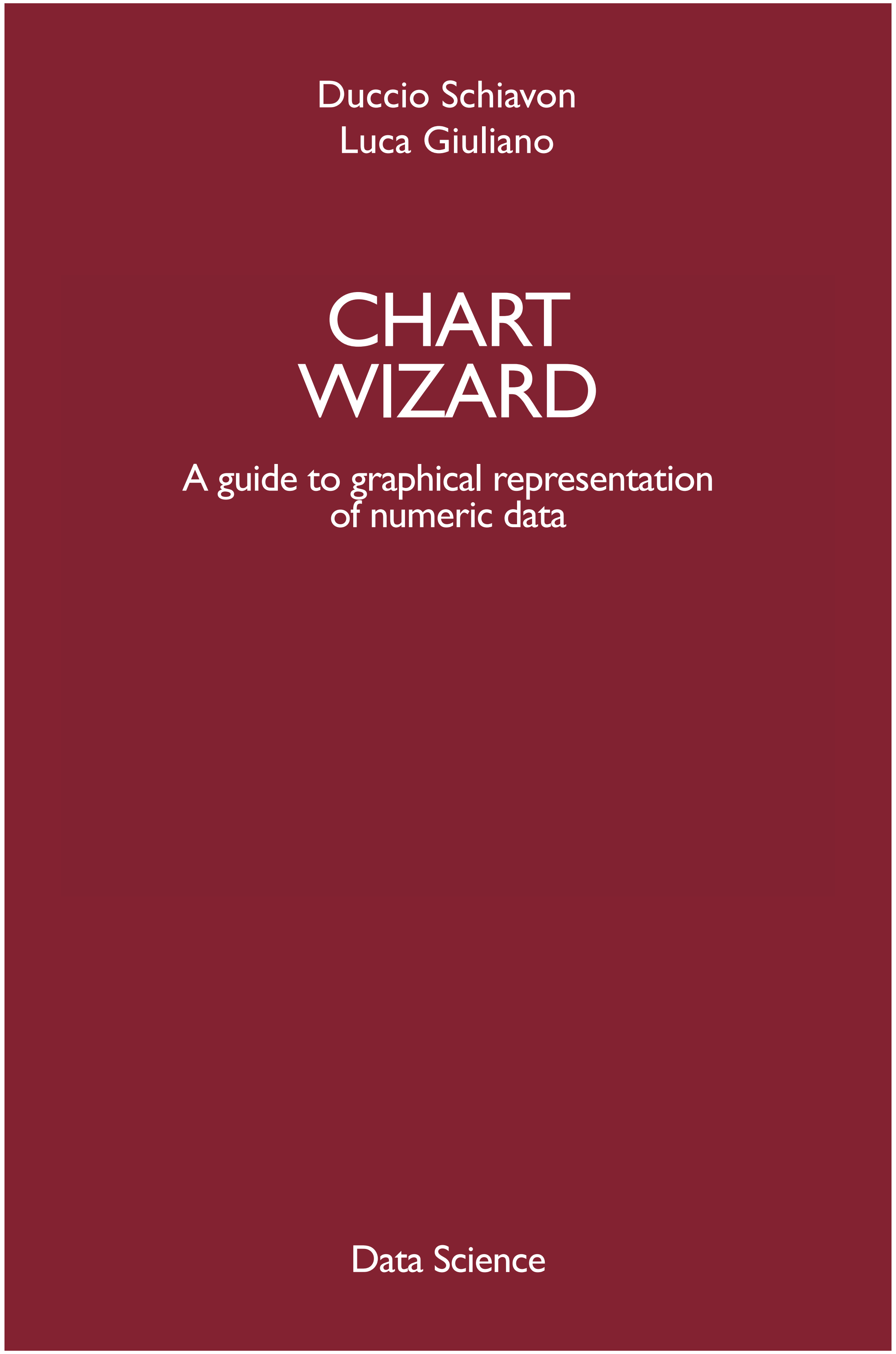
Duccio Schiavon
Luca Giuliano

CHART WIZARD

A guide to graphical representation of numeric data



Data Science



Duccio Schiavon
Luca Giuliano

CHART WIZARD

A guide to graphical representation of numeric data

Data Science

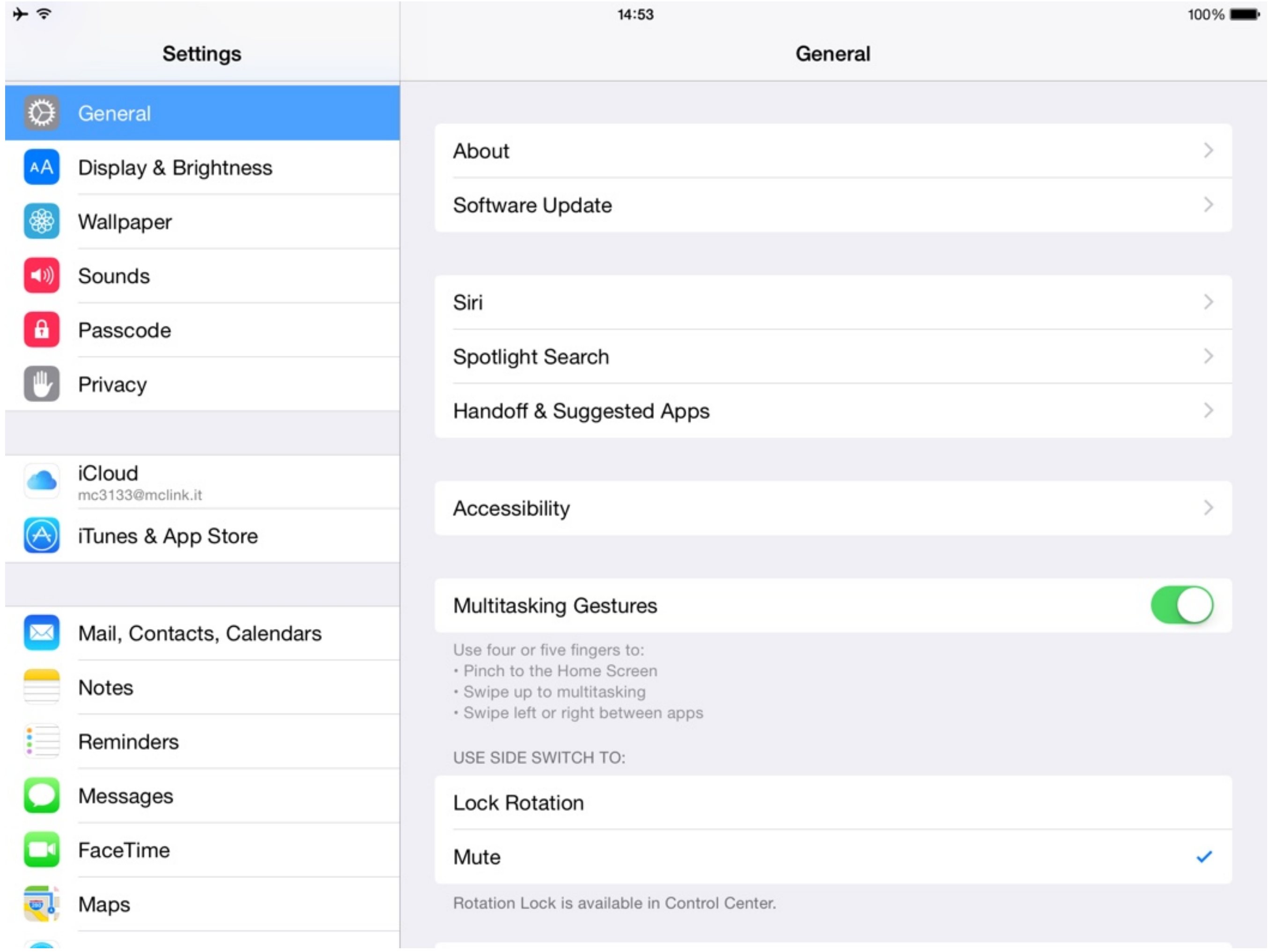
Duccio Schiavon - Luca Giuliano
Chart Wizard. A guide to graphical representation of numeric data
Rome : Department of Statistical Sciences [2015] 70 p.
ISBN 978-88-908757-4-8

To view this book, you must have an iOS device with iBooks 3.0 or later and iOS 5.1 or later, or a Mac with iBooks 1.0 or later and OS X 10.9 or later. Please report any error or inaccuracy to the authors: Duccio Schiavon, info@stat-project.com or Luca Giuliano, luca.giuliano@uniroma1.it. A PDF version of this book is available at: http://www.dss.uniroma1.it/ricerca/pubblicazioni

How to enable the Multitasking Gestures on the iPad

For an optimal reading experience of this interactive book and the use of its links we must first ensure that the Multitasking Gestures option on the iPad is enabled.

Open **Settings** and tap **General**. Scroll down to find "Multitasking Gestures" and set it to ON.



Multitasking Gestures ●

Use four or five fingers to:
• Pinch to the Home Screen
• Swipe up to multitasking
• Swipe left or right between apps

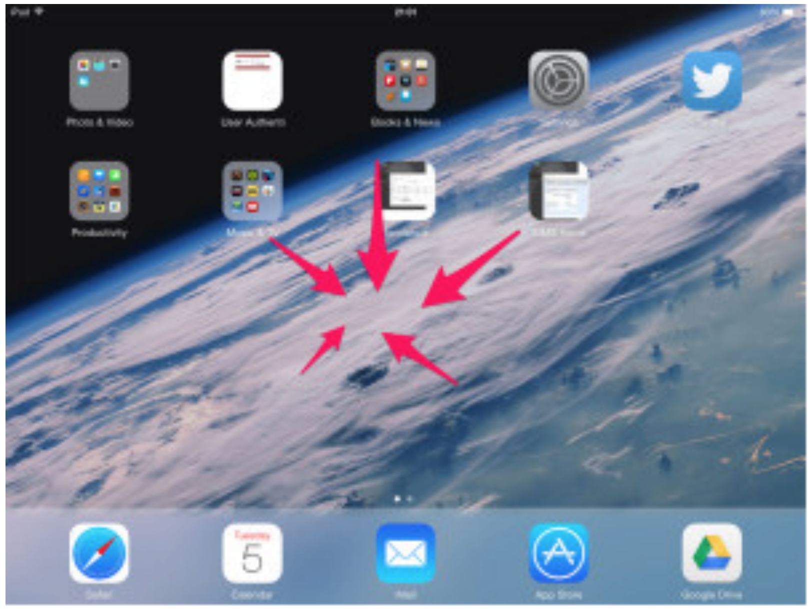
USE SIDE SWITCH TO:

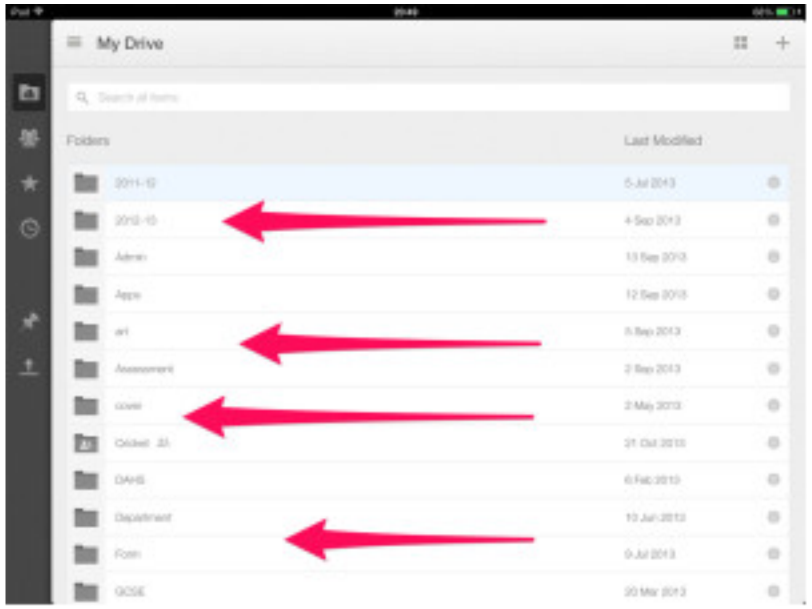
Lock Rotation

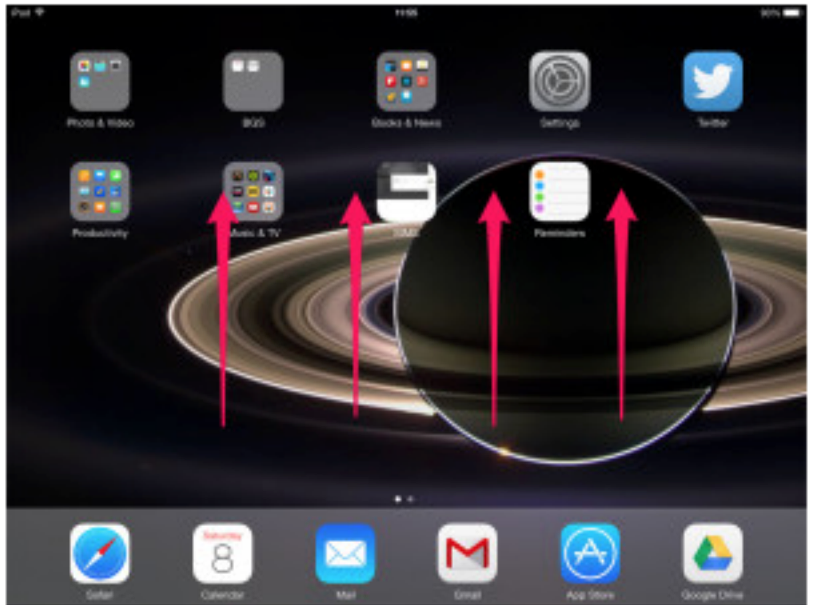
Mute ✓

Rotation Lock is available in Control Center.

Use four or five fingers to swipe horizontally (as illustrated in the figure in the middle, below) to see the active apps. Try scrolling from right to left, since, most likely, you are now displaying the last app that was used.







This gesture allows you to easily return to where you've stopped after following an external link in the Safari browser.

The other – often overlooked – functions of multitasking gestures are:

- Close apps and return to the main screen with a “pinch” gesture, using four or five fingers (figure on the left).
- Bring up the taskbar by swiping up with four or five fingers (figure on the right). To close the taskbar, swipe down.

Chapter 1

Introduction

The idea behind the creation of this chart wizard originates from an Amit Agarwal post – “How to Find the Right Chart Type for your Numeric Data” – published on his Digital Inspiration blog, and which contains a simple but clever diagram he devised.

1

What would you like to show?

Agarwal traced a diagram (fig. 1.1) which has the function to suggest the most suitable chart type in realtion both to the available data and to its representational purpose. This diagram stems from a **central question**:

What would you like to show?

It is a simple question whose answer is by no means so self-evident. In addition to being simple, this question would in fact appear hugely generic and therefore contemplate an unlimited number of solutions. The diagram does not show the analytical framework, the field of application, or what tools are to be used to get the desired representation: there is no fundamental indication that would allow to reduce the number of possible answers. In the introductory note to the post Agarwal himself limits the possibilities of using the diagram, presenting it as a tool designed exclusively for the representation of **numerical data**. However, the note alone is not sufficient by itself to fully clarify the meaning of the diagram: only by observing all the graphical suggestions and by identifying all the items it is possible to understand how to properly interpret Agarwal's diagram.

First of all it can be noted that the limit to the number of possible answers is certainly one defined by **statistics**: the charts suggested by the diagram as ideal for a particular type of analysis are those traditionally employed for the representation of measures and numerical aggregations resulting from statistical processing. Likewise, the terminology used (**distribution**, **composition**, **variable**, **relative difference**, **absolute difference**, etc.) belongs to a predominantly statistical framework, as well as the arrangement in logic flow patterns leading to a final result (representation) has something that is peculiarly statistical.

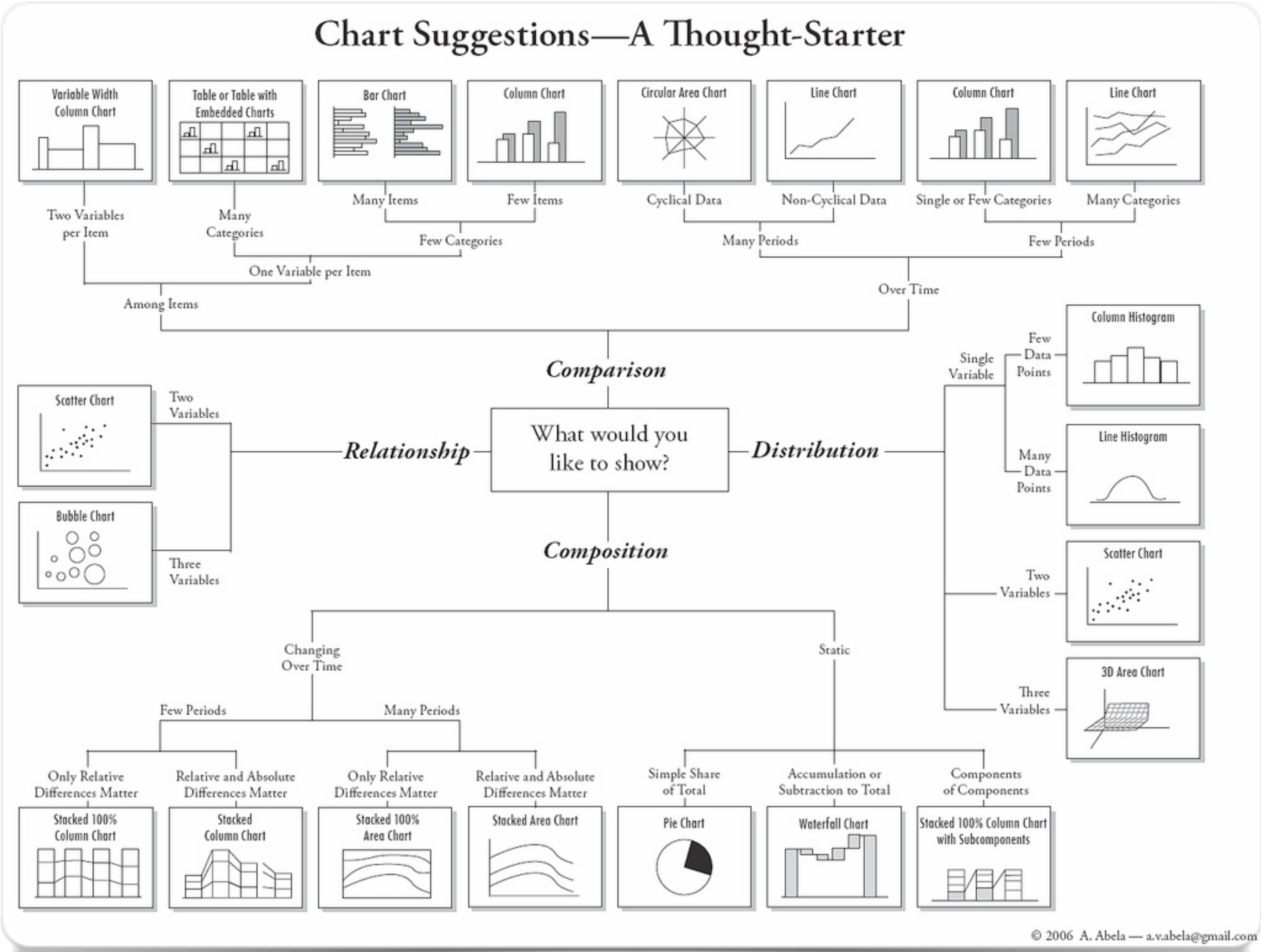


Fig. 1.1 Agarwal diagram

Personally, I really appreciate the simplicity of the diagram. The central, starting question "What would you like to show?" can be answered by choosing among four possibilities ("Relationship", "Comparison", "Distribution", "Composition") following the traditional logic of flow charts. After choosing the first answer, still other questions are asked, from a minimum of 1 to a maximum of 3, depending on the first answer given.

Then, depending on the path chosen in the Agarwal diagram, the user is provided with a stylized drawing of the (statistical) graph type to be used on the basis of the answers given along the way.

The Chart Wizard is clearly inspired by the Agarwal diagram and tries to be a first attempt at devising a method which may be applied to the most modern consultation and information tools (tablet, mobile device, etc.) and, at the same time, may give the user an easy access to very technical subjects. Of course, much of the effectiveness of the Agarwal diagram lies in its extreme ability to synthesize - a feature that many products of the modern art of **infographics** share.

Like all web-designed representations, the Agarwal diagram isn’t always able to provide an adequately comprehensive explanation. In most cases, this type of web products require the user to make a further effort before finally getting the desired hint.

It would be useful, for example, that the user of the diagram was provided with more comprehensive information on what a pie chart is, rather than simply display a concise reproduction of it.

The Chart Wizard has been designed to respond to a second and a third central question. The **second central question** actually has the function to enrich/integrate the information provided by the Agarwal diagram.

For example, let’s suppose that the user has been suggested to employ a pie chart. Then, a second central question will be:

What is a pie chart?

The Chart Wizard provides a thorough, and at the same time concise answer to the second central question. For each of the possible combinations of the **type of representation/nature of the given data** the Chart Wizard provides, in addition to the name and a picture of the graph, also a brief textual explanation that helps to understand it in a way which its mere model cannot do.

However, the main purpose of the Chart Wizard is to answer a **third central question**: based on the example we have mentioned before, this question is:

How can I make a pie chart?

For me, the answer to this question is the real reason this Chart Wizard was made. While on one side it was necessary to find a method to guide the user through the logical path for the choice of the most suitable graphic for the intended purposes, on the other side the Chart Wizard aims to suggest some tools to be used in the making of the desired graph.

Web-based software

At the basis of the choice of the suggested tools there is a clear intention of limiting the possible options to **web-based softwares** only. The proliferation of web platforms that offer the possibility of creating graphical and high-quality statistical representations is enough to allow almost anyone to get basic and complex charts with just a few clicks and without paying one euro of expensive licenses. Moreover, the functioning of these web tools often rely on data-entry interfaces also web-based and easy to use.

Of course the web, due to its characteristic diversity, involves different uses and operations that change with the changing of the tools employed. It is possible that many web-based softwares, that can be used to create the same representation, may require very different methods of use. In addition to this, the tools differ in terms of the number of options about interactivity, appearance and content customization, "graphic quality” of the representations, and so on.

However, what matters most is the context in which these tools are born and grown: their use is often referred to as ideal in an **open data** context of use, that is in circumstances in which information available to anyone and universally usable formats can be exploited. In some cases, these tools are the result of collaborative work done by users/developers working in open source environments (r-project), or they are the result of academic research (D3.js); in other cases they are tools implemented in actual, data and graphic sharing dedicated web-portals (Many Eyes, today in transition to Watson Analytics, IBM’s new cloud-based service for smart data discovery).

In all of these cases, the solutions are designed mainly to:

1. facilitate a greater dissemination of information;

2. facilitate the creation of low, medium and high complexity charts;

3. ensure a better understanding of the numeric data, thanks to a significant simplification of the graphic details.

This last point brings us to a final consideration: the simplification of graphics requires an often painful but necessary transformation. The increasing availability of numeric data in any areas (marketing, research, communication, etc.) necessarily involves the use of charts that have to be able to clearly show the results of its processing. The tools must be then easy to use, intuitive, and, above all, provide a wide range of possible representations for the same quantitative numerical result.

At the same time the users are requested to make an effort in having a clear idea about what they want to get from these tools, and how, on the basis of the data they have. This situation of constant activity (that is to say “not occasional”) promotes not only the evolution of web-based visualization tools, but also the emergence of **new methods of representation**, through the cognitive contribution of all of the users. As an example, let’s just think of the difficulty that only a few decades ago would have met anyone in trying to graphically describe an event in more than two dimensions (variables), in an effective and sufficiently understandable way.

Today, the creation of a complex, three-dimensional chart with just a few clicks is an operation within the reach of almost everyone. Today anyone can create animated representations of the evolution of historical data just by knowing in what form to use the data through interfaces akin to that of actual spreadsheets (Gapminder).

Today the creation of charts related to very specific contexts, in order to deliver them to the user community as reference tools of representation, is a relatively simple operation compared to not so long ago. The increased knowledge of the tools used, their spread and their concurrent use therefore encourages a creative process that is essential in ensuring an always greater possibility of development and variety of choice.

The Chart Wizard also has this purpose: that is to stimulate the creativity of its users once the concrete feeling of being able to implement new forms of representation from scratch is transmitted to them. It is my hope that the suggestions offered by the Chart Wizard will not just give its users what is necessary to get the desired representation, but mostly will represent an element of charm that will inspire the users to image many other ways, that still don’t exist, to describe their data.

Duccio Schiavon

Chapter 2

Relationship

- Scatterplot
- Binning
- Curve fitting
- Heatmap
- Bubble chart
- Surface plot
- Contour plot
- Correlation matrix
- Matrix plot

2

Relationship between two quantitative variables

The **scatterplot** is a graphical tool through which to link two quantitative variables (continuous or discrete). It is mainly used to infer whether there are relationships of a direct or inverse proportion between the two measures compared. It is also a useful chart if you want to infer, through a single visualization, the distributive nature of the two measures.

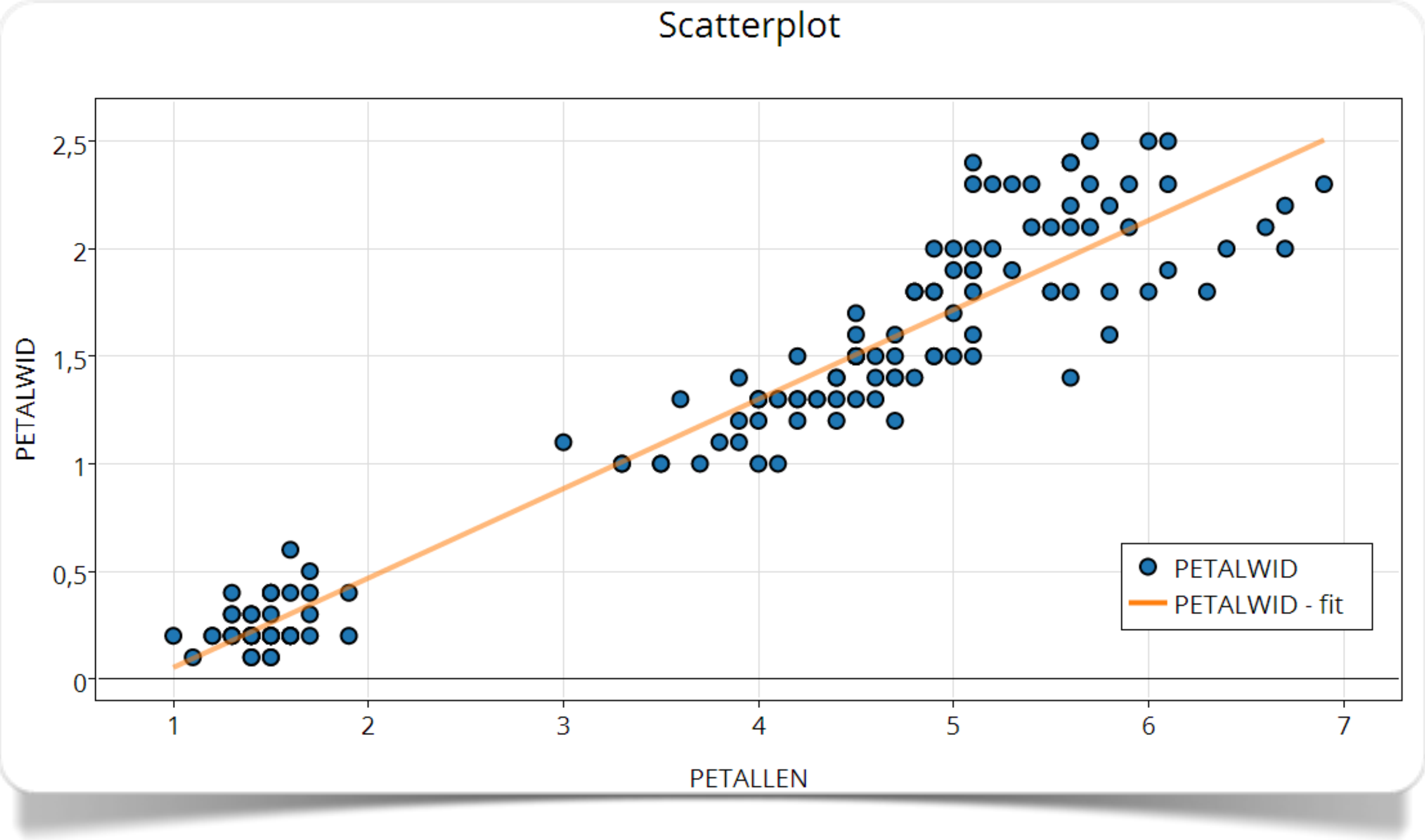


Fig. 2.1 Scatterplot made with plotly

plotly is a web portal for the production of interactive statistical graphs, entirely responsive (the graphs fit the size of the screen, so as to ensure also mobile compatibility), provided with API for interfacing with external environments and a wide range of graphics possibilities. Through the interface you can upload CSV, TSV, Matlab, MS Access, or text files and then export the graphs to PNG, PDF, SVG and EPS format. It is also possible to interface with R through its API.

Binning is a kind of representation that allows to visually identify the most "populated" areas of a **scatterplot**. In technical terms, data binning is a technique of data processing in which the original values that fall in a given minimum interval (bin) are replaced by a single value representative of that range, often represented by the “central” value. For some reference on this specific graphical representation, see the work of Zachary Forest Johnson.

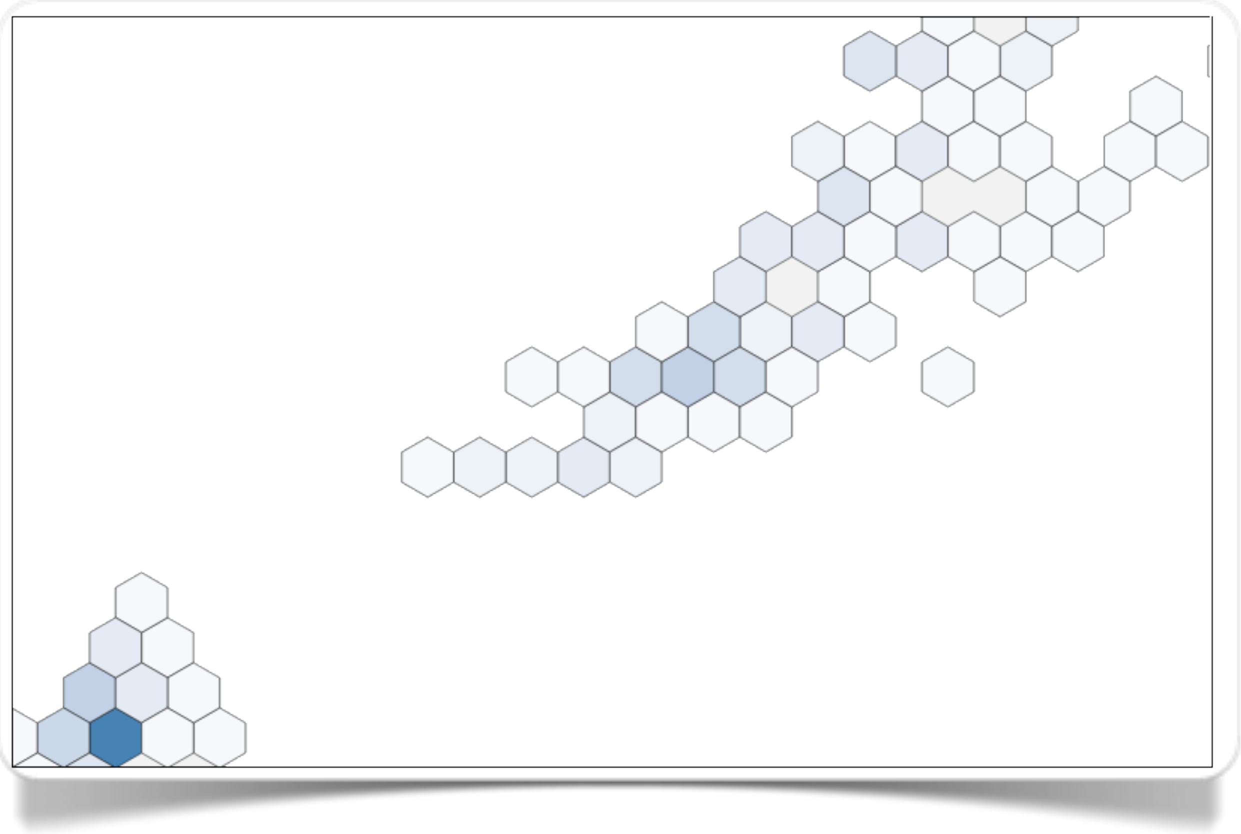


Fig. 2.2 Hexagonal binned scatterplot made with Raw

Raw is an open source platform, whose web interface acts as a generator of representations. The way it works, you load the data by simply copying and pasting it in a text area, then you choose the layout and specify the variables to be used; you can customize some graphic features and finally export the chart to a plain PNG format, or to a vector SVG format too. It is based on the D3.js library.

Curve fitting is especially used in the presence of quantitative variables with many data of **continuous nature**. It is very useful in determining the trends outlined by the relationship of the two variables being compared, and to assess the level of deviation of data points from the **curve fitting** (**variability**).

WolframAlpha is a computational engine that can process the keywords specified by the user and provide a series of numerical information and data. The developers of this search engine are the same people who developed the Mathematica software: this is the reason for its strong focus on computation and statistics. The fitted curve in figure 2.3 was drawn by specifying in its search field the expression:

exponential fit:

```
exponential fit
0.783,0.552,0.383,0.245,0.165,0.097
```

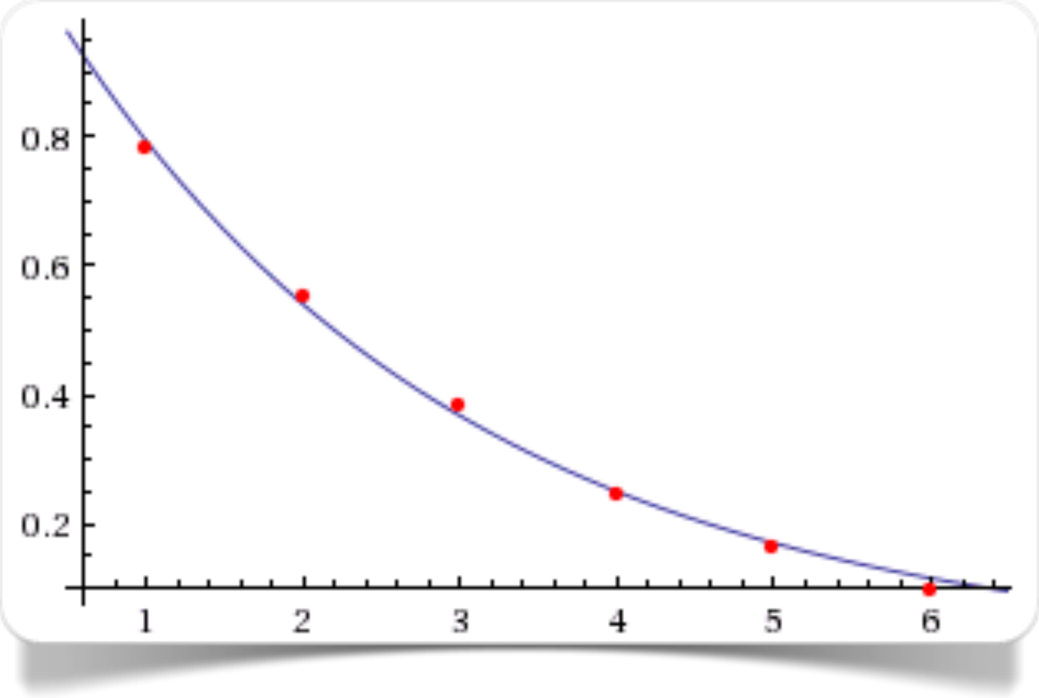


Fig. 2.3 Curve fitting made with WolframAlpha

Among the possible alternatives for the creation of fitted curves we recommend the use of plotly, which, starting from a scatterplot, offers the possibility to use the option FIT DATA to adapt any function to the points shown on the graph.

In figure 2.4 we can see a **scatterplot** where the relationship between life expectancy and duration of gestation in some animal species is represented.

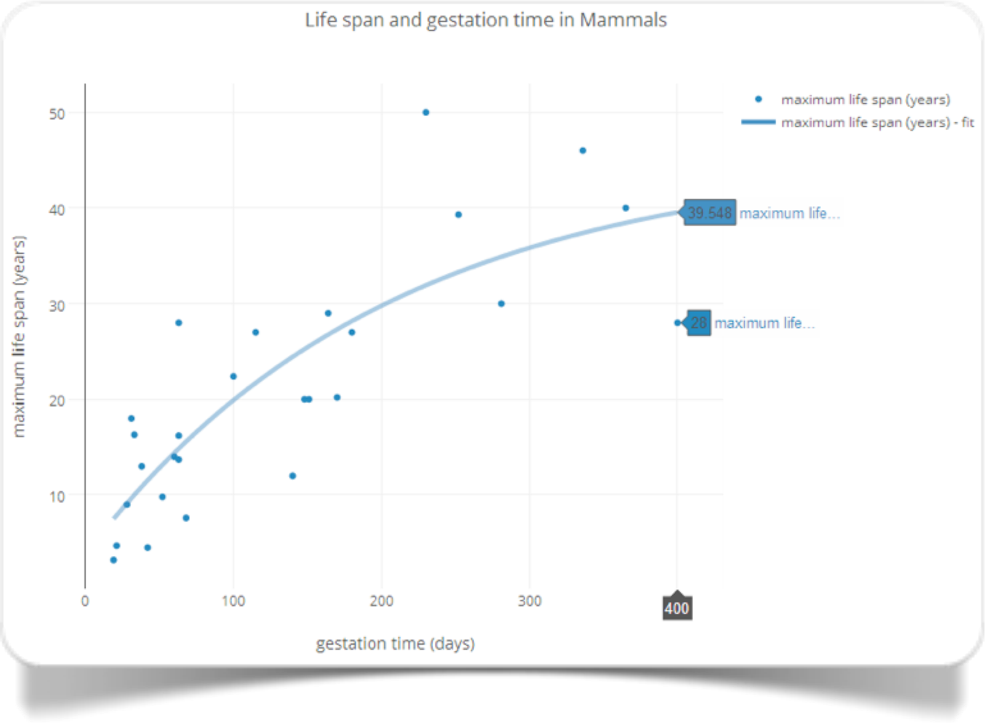


Fig. 2.4 Life expectancy and gestation period in some mammal species (made with plotly)

Species of Animals	maximum life span (years)	gestation time (days)
African-giant-pouched-rat	4.5	42
Arctic-Fox	14	60
Baboon	27	180
Cat	28	63
Chimpanzee	50	230
Cow	30	281
Donkey	40	365
Giraffe	28	400
Goat	20	148
Gorilla	39.3	252
Gray-wolf	16.2	63
Ground-squirrel	9	28
Guinea-pig	7.6	68
Horse	46	336

Fig. 2.5 Data relating to figure 2.4

Relationship between two qualitative variables

The heatmap type plot (Sneath, 1957) is the ideal visual reproduction of a **contingency table** (double-entry table): it compares two categorical variables characterized by a limited number of categories.

The hue is indicative of the size of the frequencies of each cell. Higher values (e.g. percentage value) will correpond to more intense colours.

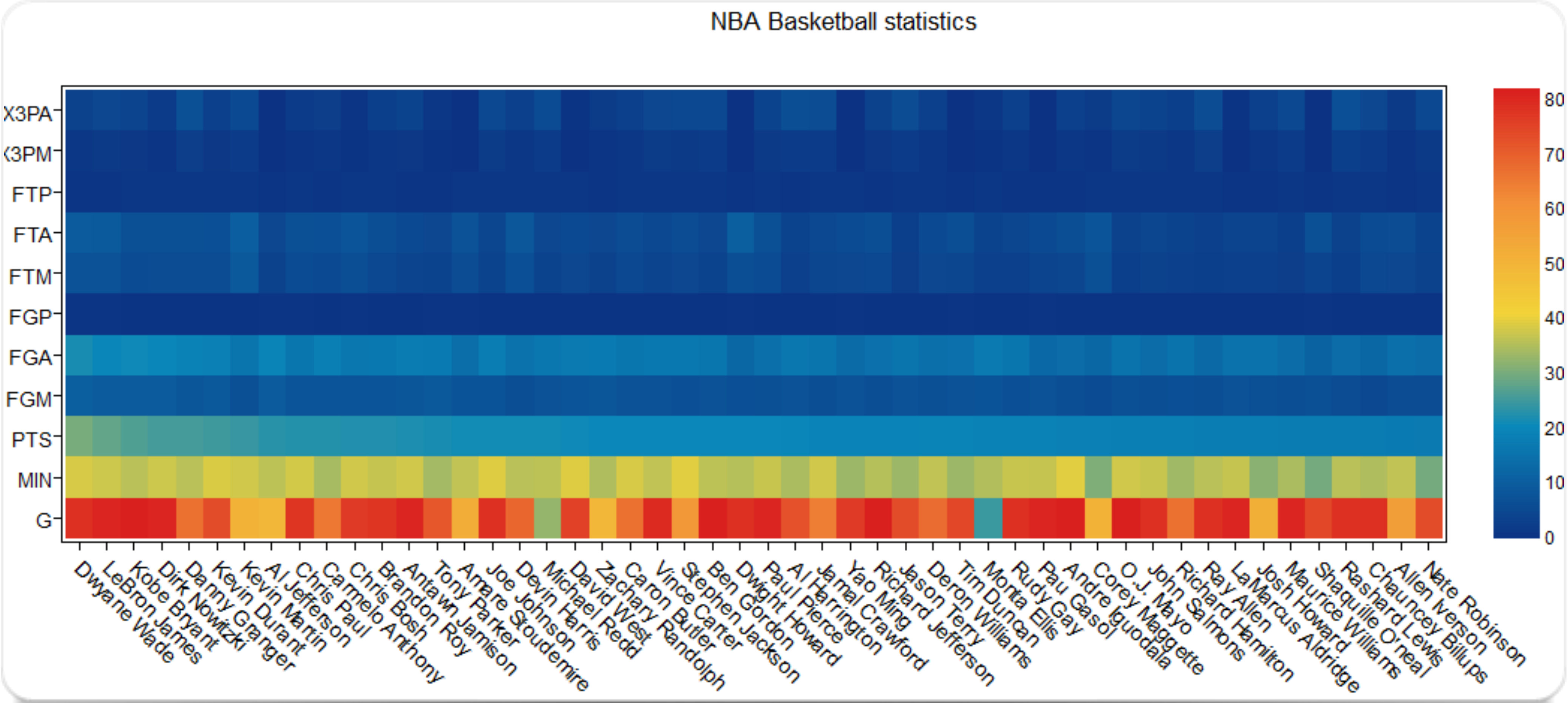


Fig. 2.6 Heatmap made with plotly

A **heatmap** plot can be created with plotly, with a complete control over all of its graphic aspects. Examples like the one in Figure 2.6 can be found in the gallery of the selectable representations.

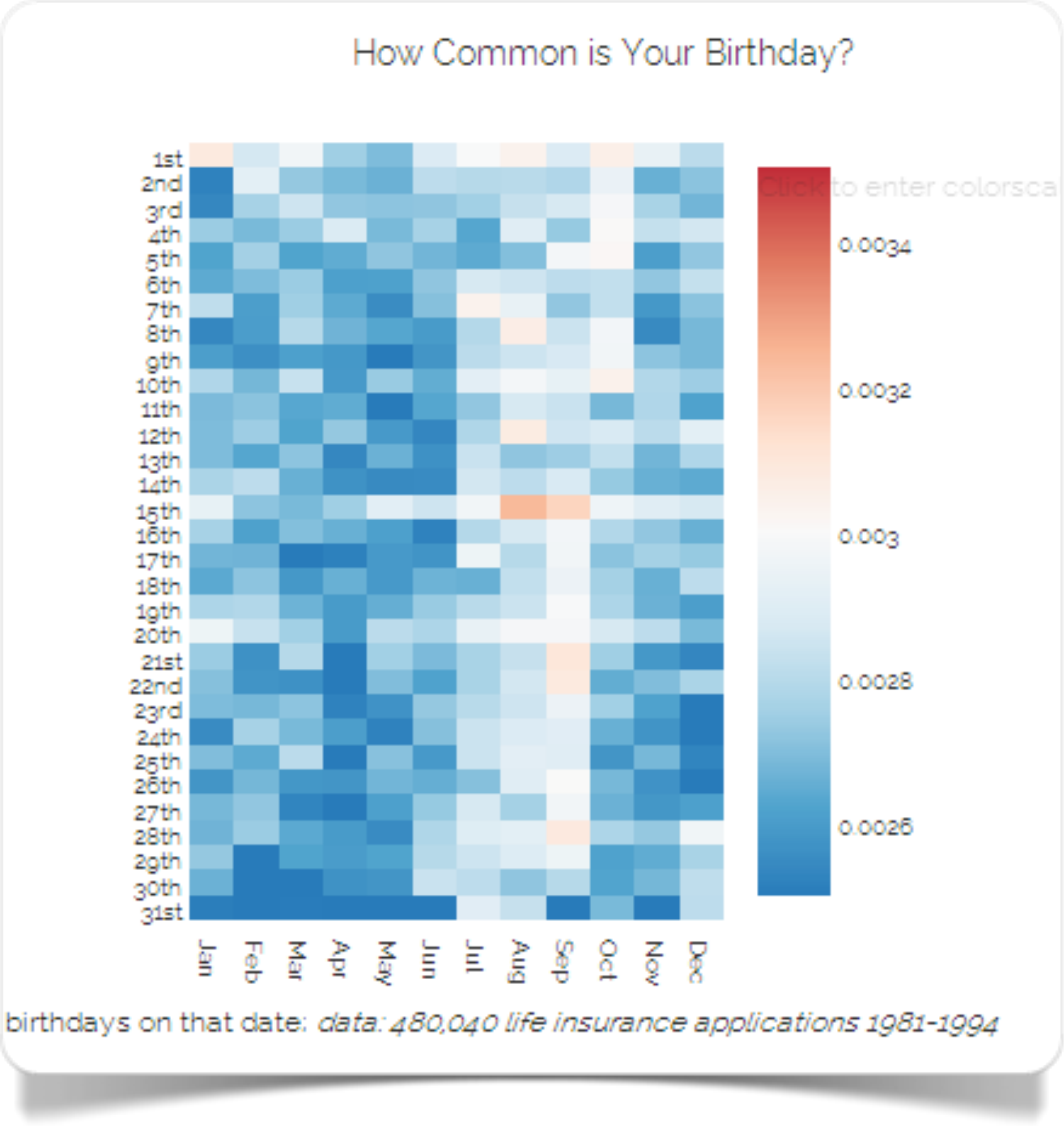


Fig. 2.7 Frequency of birthday by day and month (made by Dreamshot with plotly)

In Figure 2.7 we can observe how the most common birthday, in the population chosen by user Dreamshot who created the chart with plotly, is August 15. The graph shows how births are more frequent during summer.

Alternatively, for the creation of **heatmap** plots, Many Eyes can be used, as Figure 2.8.

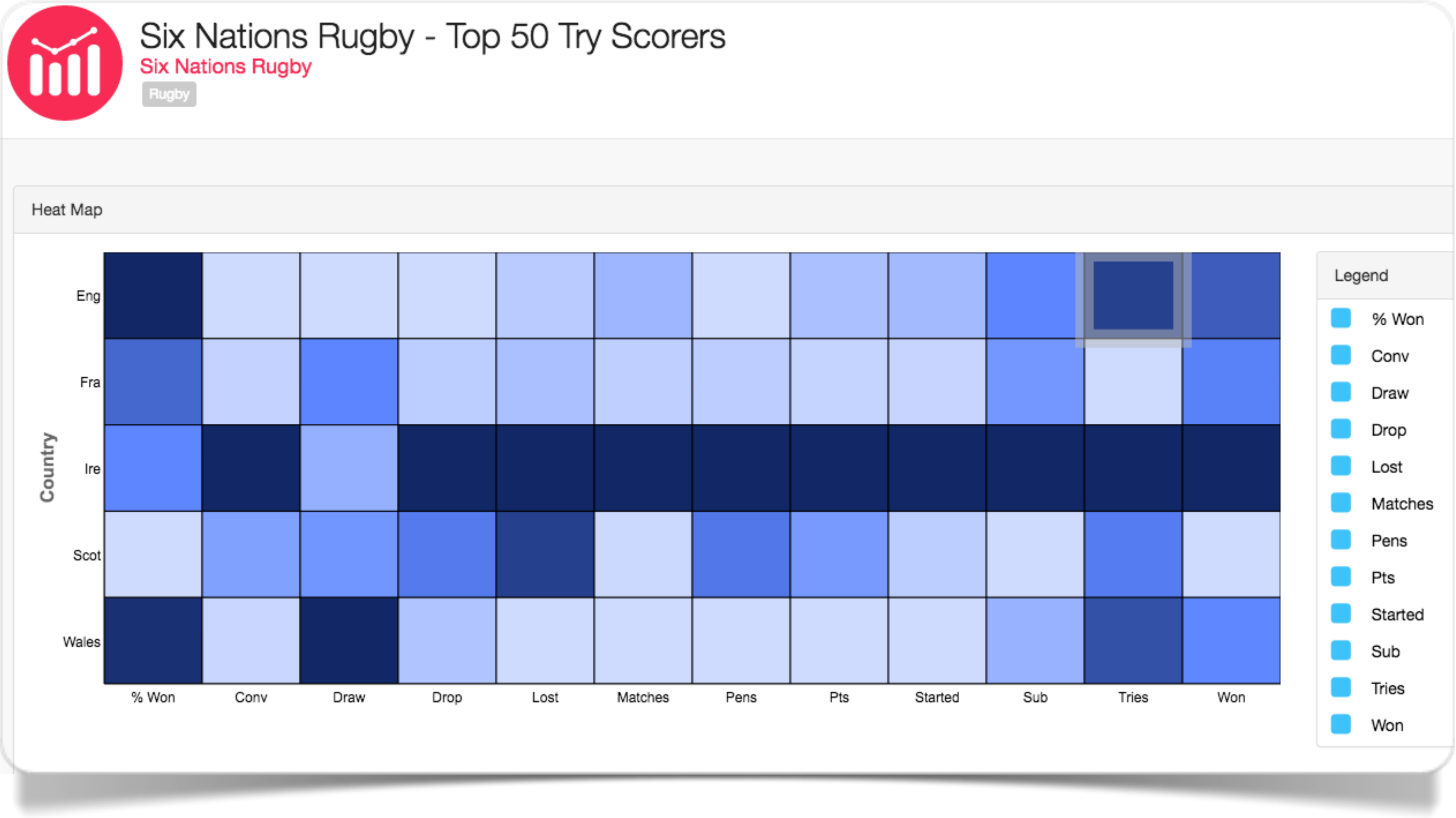
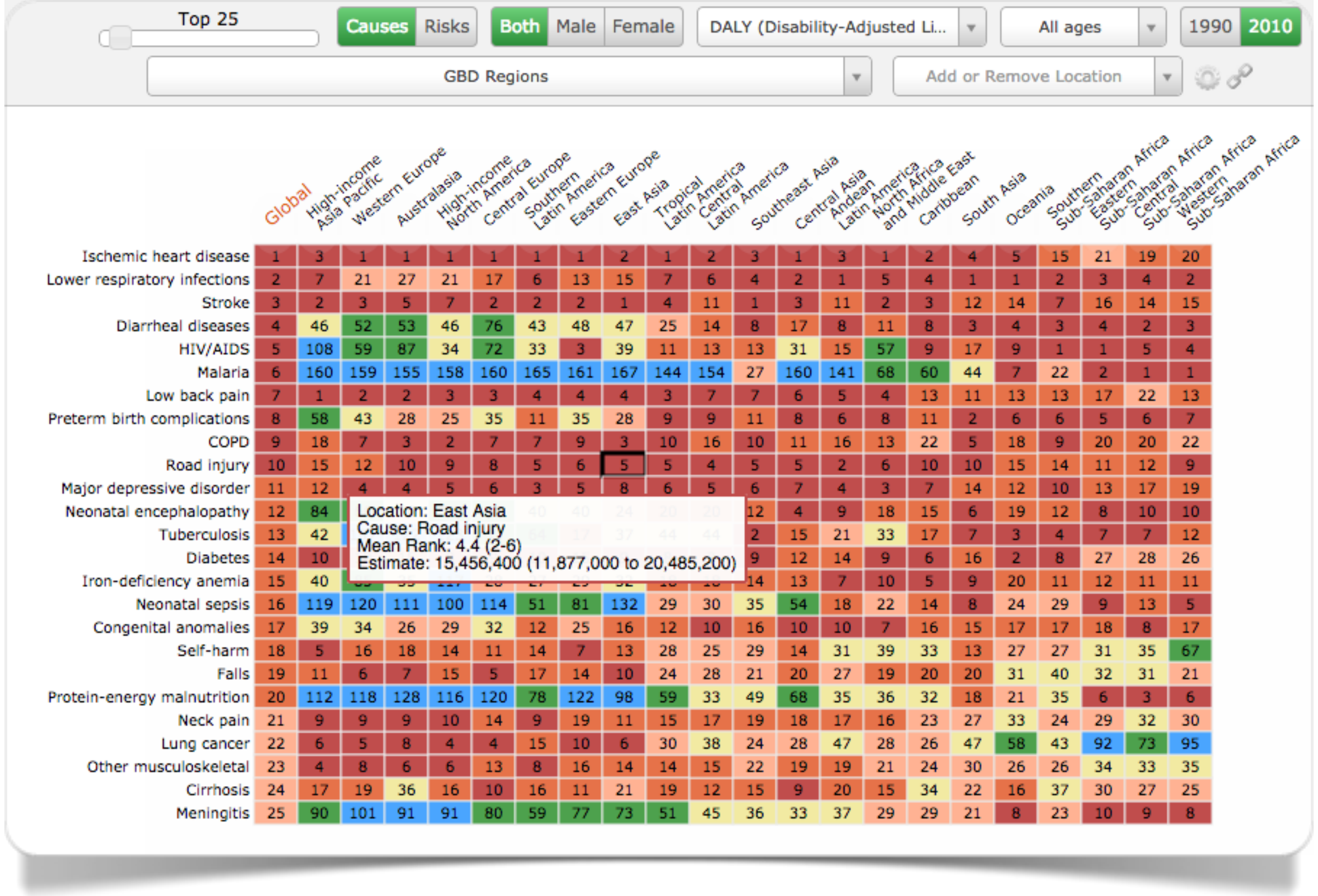


Fig. 2.8 Six Nations Rugby - Top 50 Scores (made with Many Eyes)

In Figure 2.9 we can rank and compare causes and risks in different countries based on deaths, YLLs (Years of Life Lost), YLDs (Years Lived by Disabiliy), and DALYs (Disabiliy-Adjustded Life Years) by age group and sex (Source: Institute for Health Metrics and Evaluation). In the interactive graph you can select several option.



Top 25

Causes | Risks

Both | Male | Female

DALY (Disability-Adjusted Li...

All ages

1990 | 2010

GBD Regions

Add or Remove Location

	Global	High-income Asia Pacific	Western Europe	Australasia	High-income North America	Central Europe	Southern Latin America	Eastern Europe	East Asia	Tropical Latin America	Central Latin America	Southeast Asia	Central Asia	Andean Latin America	North Africa and Middle East	Caribbean	South Asia	Oceania	Southern Sub-Saharan Africa	Eastern Sub-Saharan Africa	Central Sub-Saharan Africa	Western Sub-Saharan Africa
Ischemic heart disease	1	3	1	1	1	1	1	1	2	1	2	3	1	3	1	2	4	5	15	21	19	20
Lower respiratory infections	2	7	21	27	21	17	6	13	15	7	6	4	2	1	5	4	1	1	2	3	4	2
Stroke	3	2	3	5	7	2	2	2	1	4	11	1	3	11	2	3	12	14	7	16	14	15
Diarrheal diseases	4	46	52	53	46	76	43	48	47	25	14	8	17	8	11	8	3	4	3	4	2	3
HIV/AIDS	5	108	59	87	34	72	33	3	39	11	13	13	31	15	57	9	17	9	1	1	5	4
Malaria	6	160	159	155	158	160	165	161	167	144	154	27	160	141	68	60	44	7	22	2	1	1
Low back pain	7	1	2	2	3	3	4	4	4	3	7	7	6	5	4	13	11	13	13	17	22	13
Preterm birth complications	8	58	43	28	25	35	11	35	28	9	9	11	8	6	8	11	2	6	6	5	6	7
COPD	9	18	7	3	2	7	7	9	3	10	16	10	11	16	13	22	5	18	9	20	20	22
Road injury	10	15	12	10	9	8	5	6	5	5	4	5	5	2	6	10	10	15	14	11	12	9
Major depressive disorder	11	12	4	4	5	6	3	5	8	6	5	6	7	4	3	7	14	12	10	13	17	19
Neonatal encephalopathy	12	84	[illegible]	[illegible]	[illegible]	[illegible]	40	40	24	20	20	12	4	9	18	15	6	19	12	8	10	10
Tuberculosis	13	42	[illegible]	[illegible]	[illegible]	[illegible]	64	17	37	44	44	2	15	21	33	17	7	3	4	7	7	12
Diabetes	14	10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	9	12	14	9	6	16	2	8	27	28	26
Iron-deficiency anemia	15	40	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	14	13	7	10	5	9	20	11	12	11	11
Neonatal sepsis	16	119	120	111	100	114	51	81	132	29	30	35	54	18	22	14	8	24	29	9	13	5
Congenital anomalies	17	39	34	26	29	32	12	25	16	12	10	16	10	10	7	16	15	17	17	18	8	17
Self-harm	18	5	16	18	14	11	14	7	13	28	25	29	14	31	39	33	13	27	27	31	35	67
Falls	19	11	6	7	15	5	17	14	10	24	28	21	20	27	19	20	20	31	40	32	31	21
Protein-energy malnutrition	20	112	118	128	116	120	78	122	98	59	33	49	68	35	36	32	18	21	35	6	3	6
Neck pain	21	9	9	9	10	14	9	19	11	15	17	19	18	17	16	23	27	33	24	29	32	30
Lung cancer	22	6	5	8	4	4	15	10	6	30	38	24	28	47	28	26	47	58	43	92	73	95
Other musculoskeletal	23	4	8	6	6	13	8	16	14	14	15	22	19	19	21	24	30	26	26	34	33	35
Cirrhosis	24	17	19	36	16	10	16	11	21	19	12	15	9	20	15	34	22	16	37	30	27	25
Meningitis	25	90	101	91	91	80	59	77	73	51	45	36	33	37	29	29	21	8	23	10	9	8

Location: East Asia
Cause: Road injury
Mean Rank: 4.4 (2-6)
Estimate: 15,456,400 (11,877,000 to 20,485,200)

Fig. 2.9 Causes and risks in different countries (made with d3js.org)

Relationship between three quantitative variables

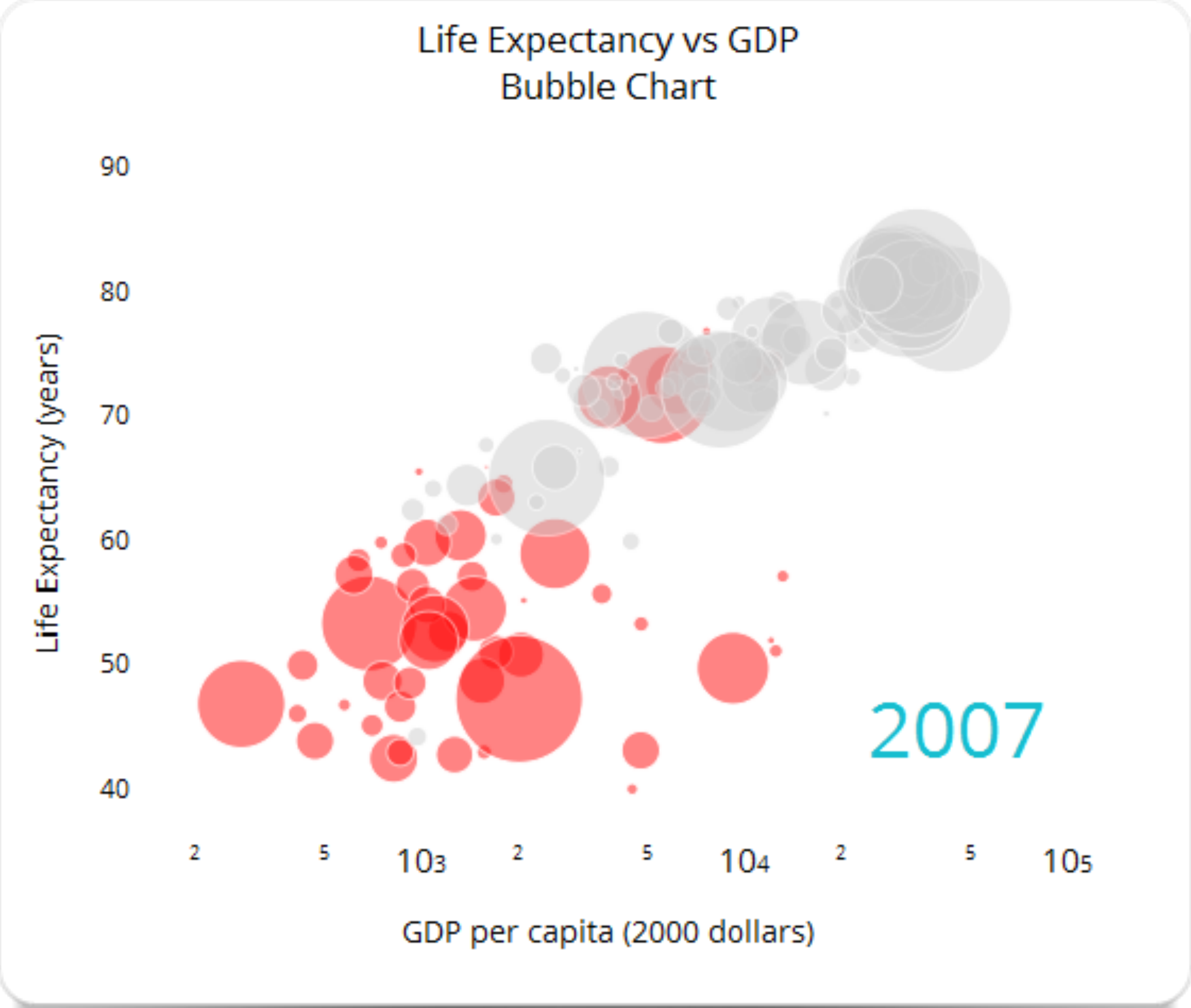


Fig. 2.10 Bubble chart made with plotly

The **scatterplot** (Chambers, 1983), besides allowing to link two quantitative variables to determine whether there is a relationship of direct or inverse proportionality between them, allows to evaluate an optional third variable "of entity" (Z). This variable, which is also numeric, defines the order of magnitude of each data point in the chart. To distinguish it from a simple, two-dimensional scatterplot, this kind of representation is often referred to as bubble chart. Figure 2.10 shows a bubble chart made with plotly. You can make this type of chart also using Raw or Slemma (former Capsidea).

The graph in figure 2.11 was made with Slemma while that of figure 2.12 was made with Raw. The data set is the same; 2.12 graph uses the reduced subset of data given in the scrollable window of Figure 2.13.

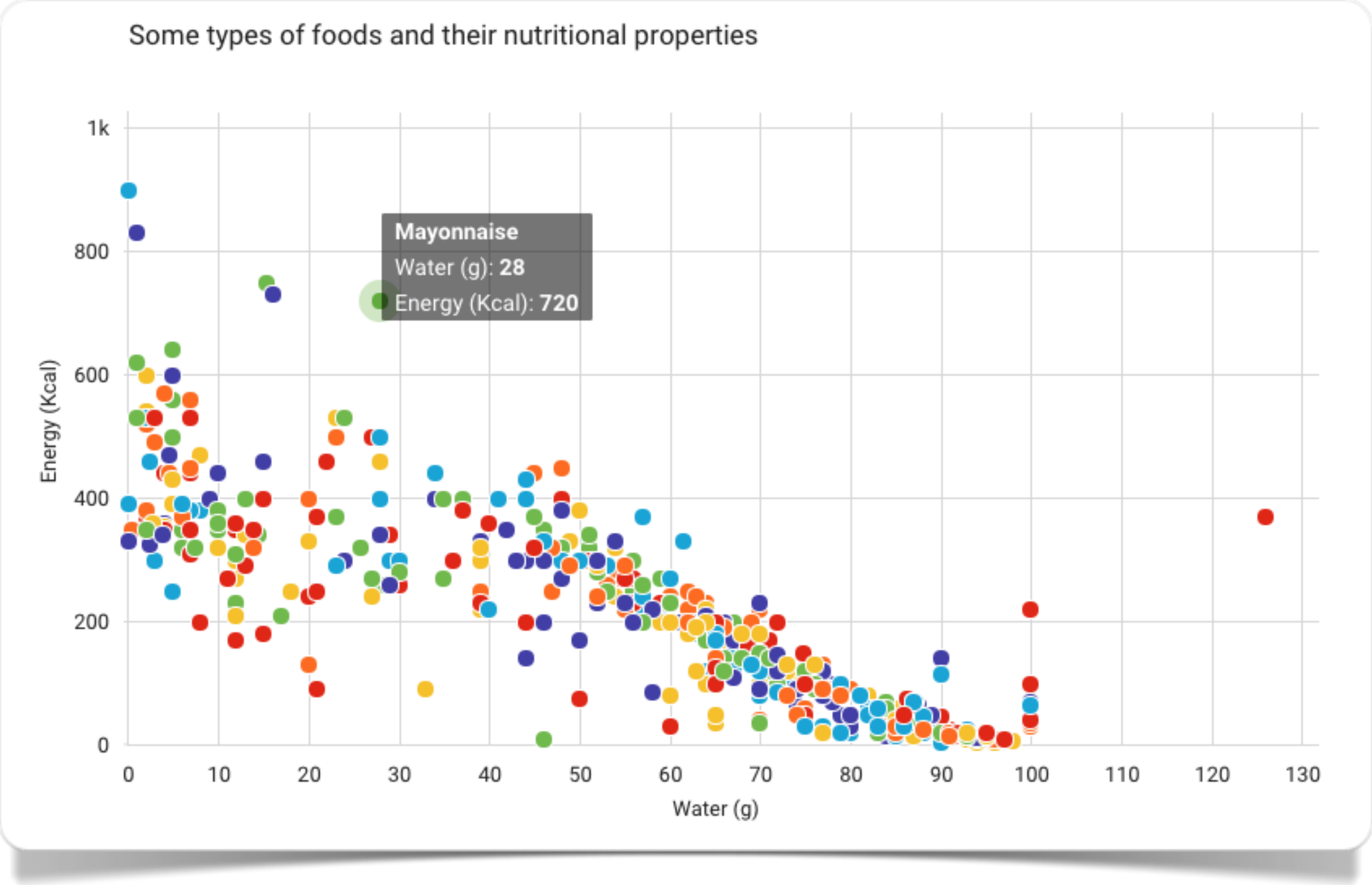


Fig. 2.11 Some types of foods and their nutritional properties (made with Slemma)

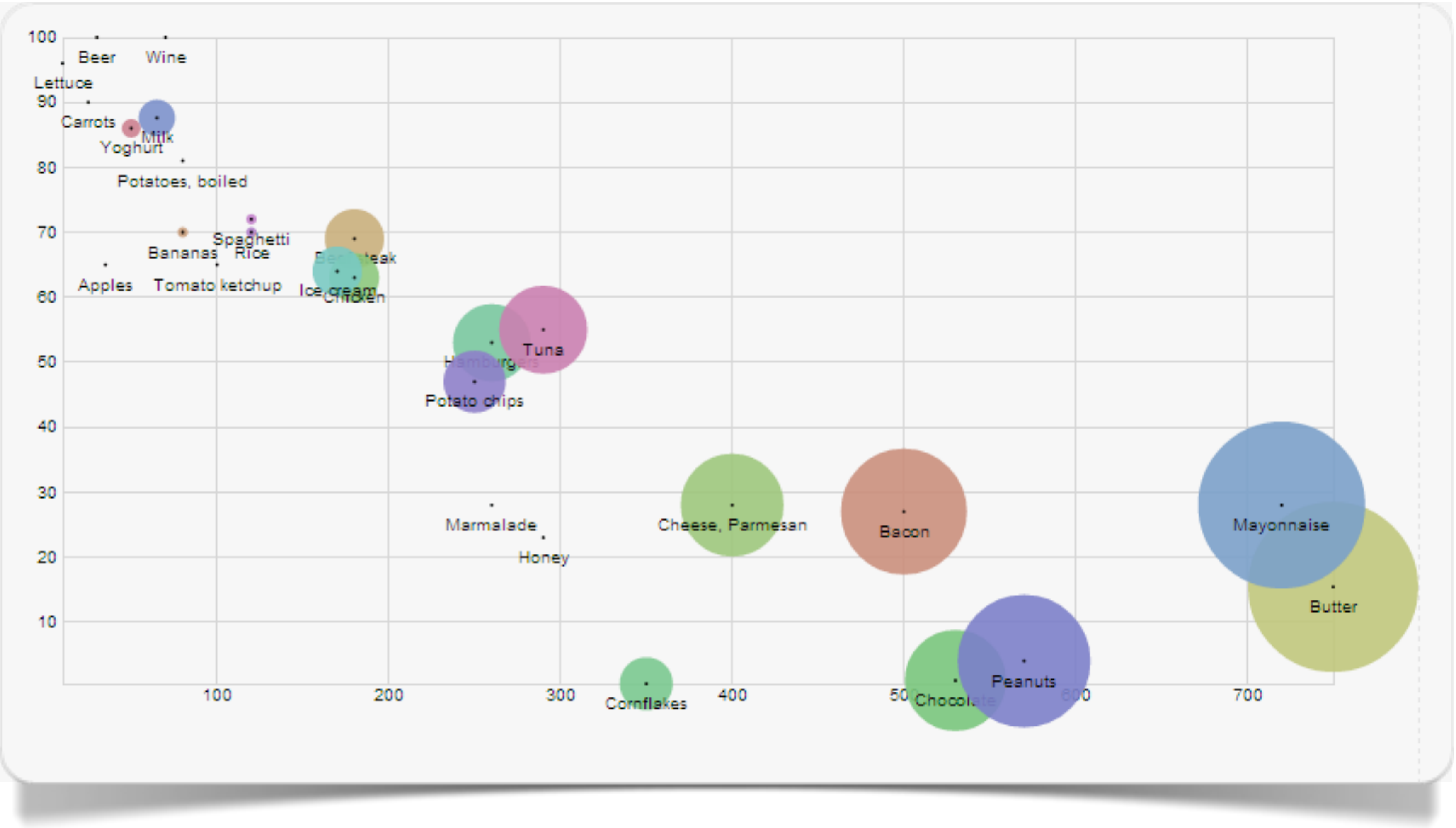
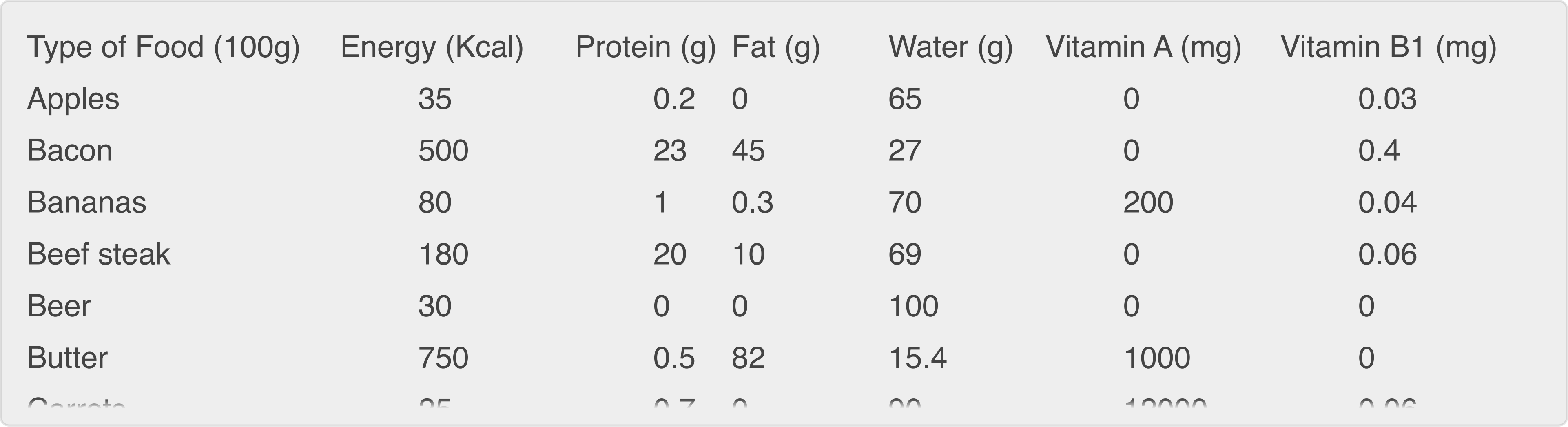


Fig. 2.12 Relationship between Calories (x), Water (y), and Fat (z) per 100 grams of certain foods (made with Raw)



Type of Food (100g)	Energy (Kcal)	Protein (g)	Fat (g)	Water (g)	Vitamin A (mg)	Vitamin B1 (mg)
Apples	35	0.2	0	65	0	0.03
Bacon	500	23	45	27	0	0.4
Bananas	80	1	0.3	70	200	0.04
Beef steak	180	20	10	69	0	0.06
Beer	30	0	0	100	0	0
Butter	750	0.5	82	15.4	1000	0
Carrots	25	0.7	0	90	12000	0.06

Fig. 2.13 Some types of foods and their nutritional properties

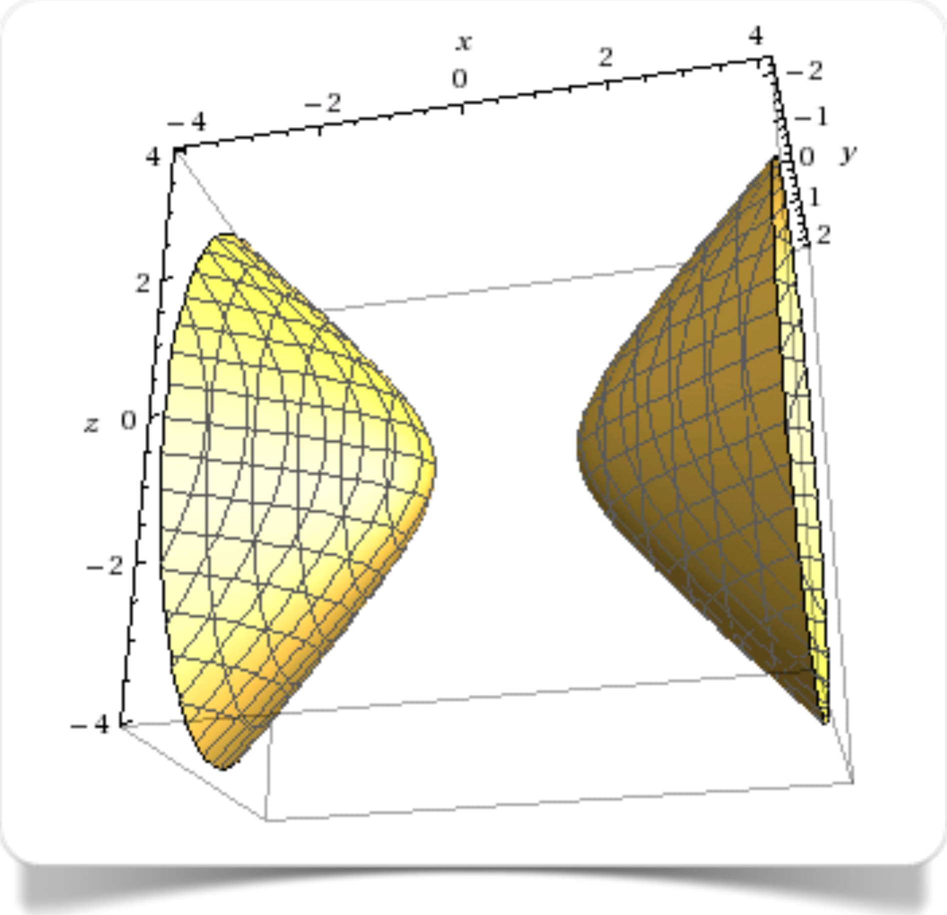


Fig. 2.14 3D graph made with WolframAlpha

The **surface plot** is a particular type of graph that allows a three-dimensional representation of three quantitative variables (most often continuous, but, if necessary, also ordinal). These graphs have the particular merit of taking advantage of different visual elements. As in a topographic map, for example, where the colors and patterns are used to represent the areas that contain the same range of values.

WolframAlpha supports the creation of 3D graphs by simply specifying a formula in the search field: the graph in Figure 2.14 was created by writing this expression in the search field:

```
plot x^2 - 3y^2 - z^2
```

The **contour plot** is the precise implementation of a two-dimensional surface plot. Once set the chosen variables on the horizontal and the vertical axis, the third variable will be represented by lines and curves shown on the Cartesian plane. Each interval defined by the space included between the different curves represents a particular class of variation of the values of the variable Z, marked in turn by a particular colour gradient.

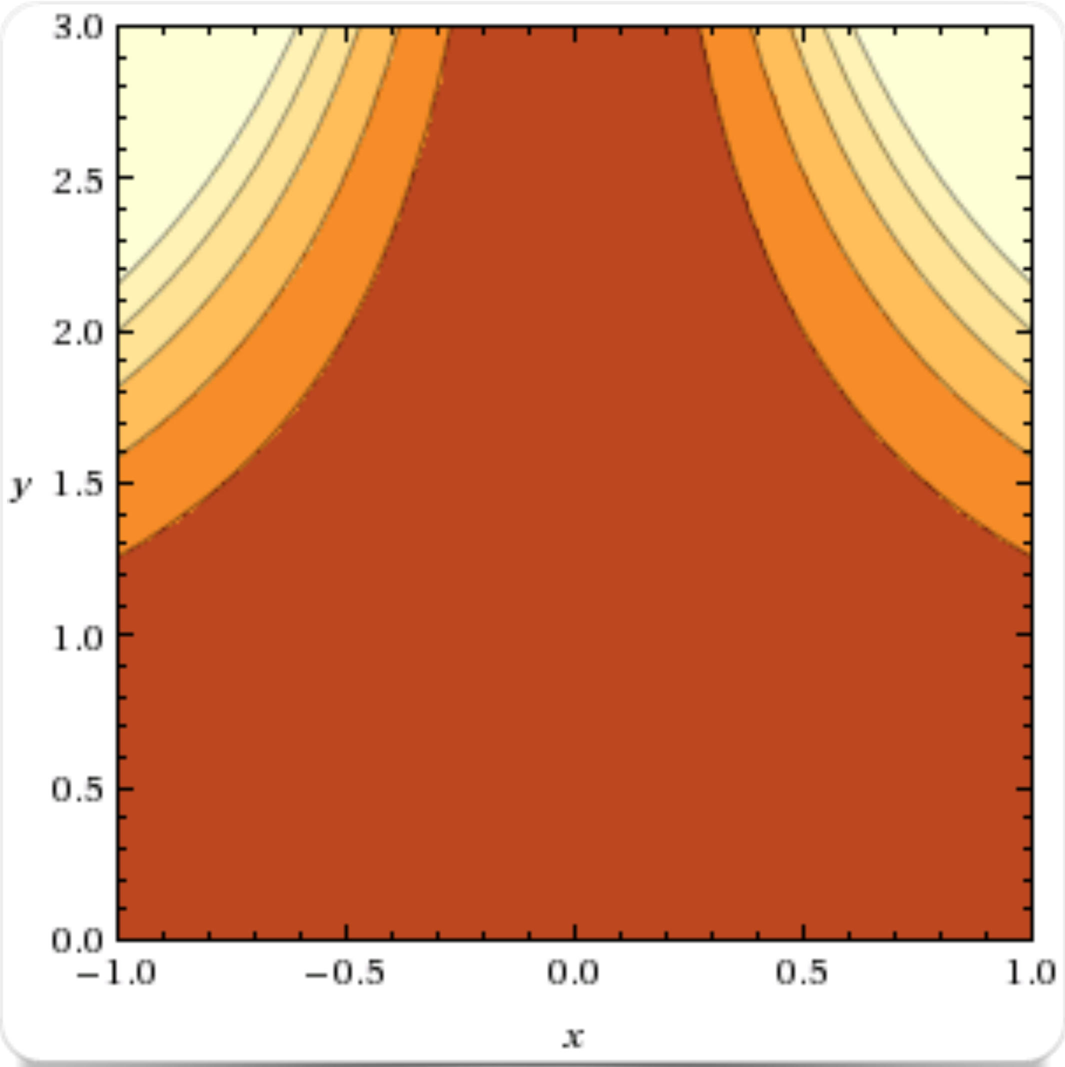
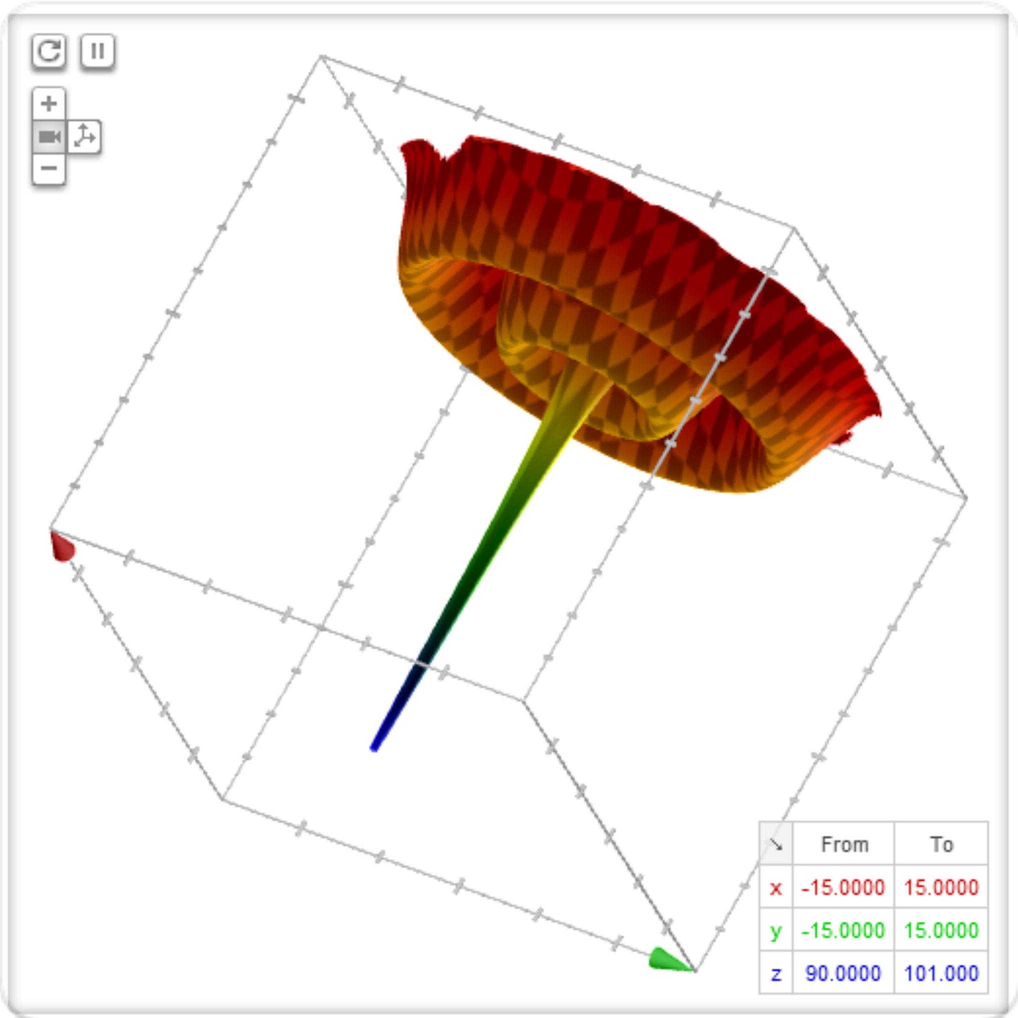


Fig. 2.15 Contour plot made with WolframAlpha



↘	From	To
x	-15.0000	15.0000
y	-15.0000	15.0000
z	90.0000	101.000

Fig. 2.16 3D graph made with Google Search

The contour plot in figure 2.15 was made with WolframAlpha by writing this expression in its search field:

```
x^2 y^3, x=-1..1, y=0..3
```

Usually, contour plots are generated and shown by the software next to the corresponding 3D surface plot.

Google too supports the creation of 3D graphs by simply specifying a formula in the search field: the graph of Figure 2.16 was created by writing in the search field the following expression:

```
100-3/(sqrt(x^2+y^2))+sin(sqrt(x^2+y^2))+sqrt(200-(x^2+y^2)+10*sin(x)+10sin(y))/1000,
            x is from -15 to 15, y is from -15 to 15, z is from 90 to 101
```

The WebGL technology, on which the Google 3D graphics rendering function is based, allows the use of some interesting interactive options, such as zooming or dragging the graph with a rotation along one of the three axes.

Relationship between quantitative variables

The **correlation matrix** is represented by a square matrix NxN, arranged so that the rows correspond to N quantitative variables. In turn, the columns should correspond to the same N quantitative variables, laid out according to the same sequential order in which they are organized by line. The matrix will then be composed of NxN squares, each coloured with a different colour scale depending on the value of the correlation calculated for each possible comparison between pairs of variables.

plotly allows to build a correlation matrix starting from any dataset composed of N variables (columns). plotly automatically calculates the correlation indexes, linking them to a colour on a colour graduated scale.

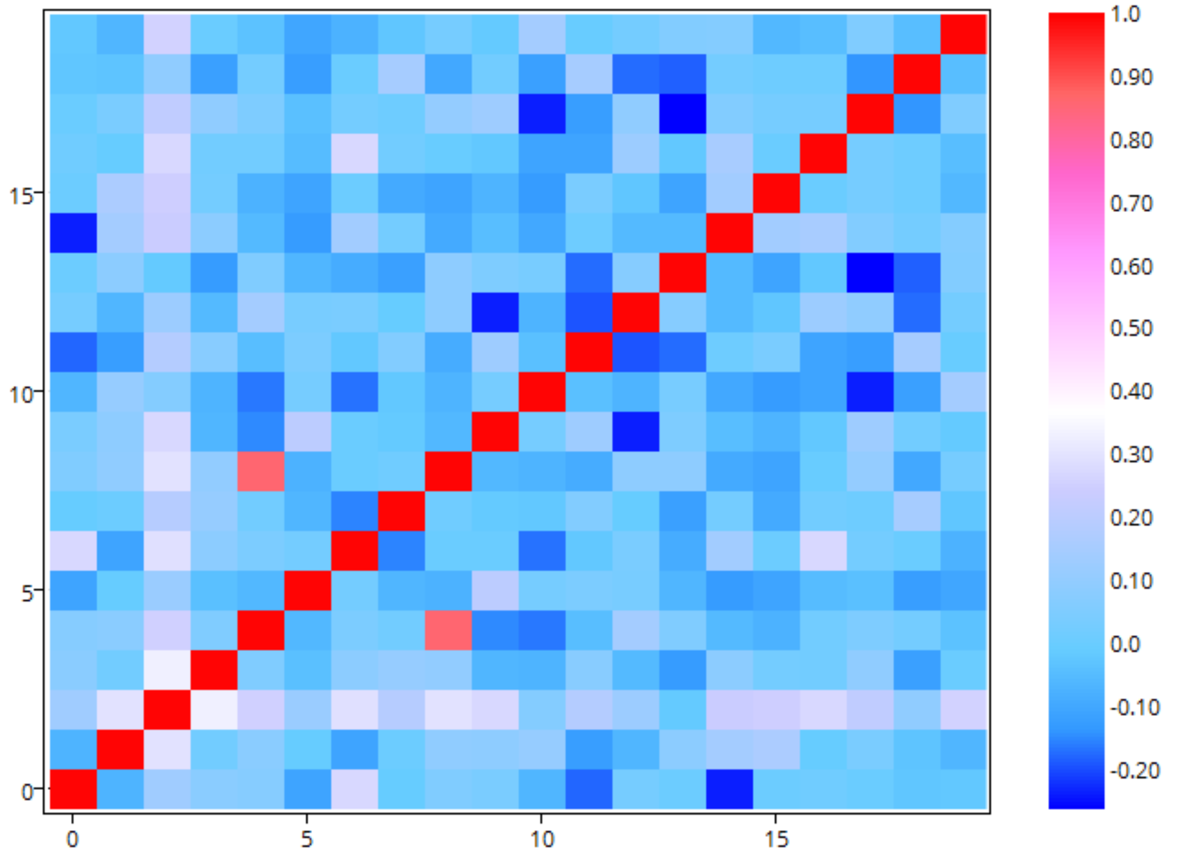
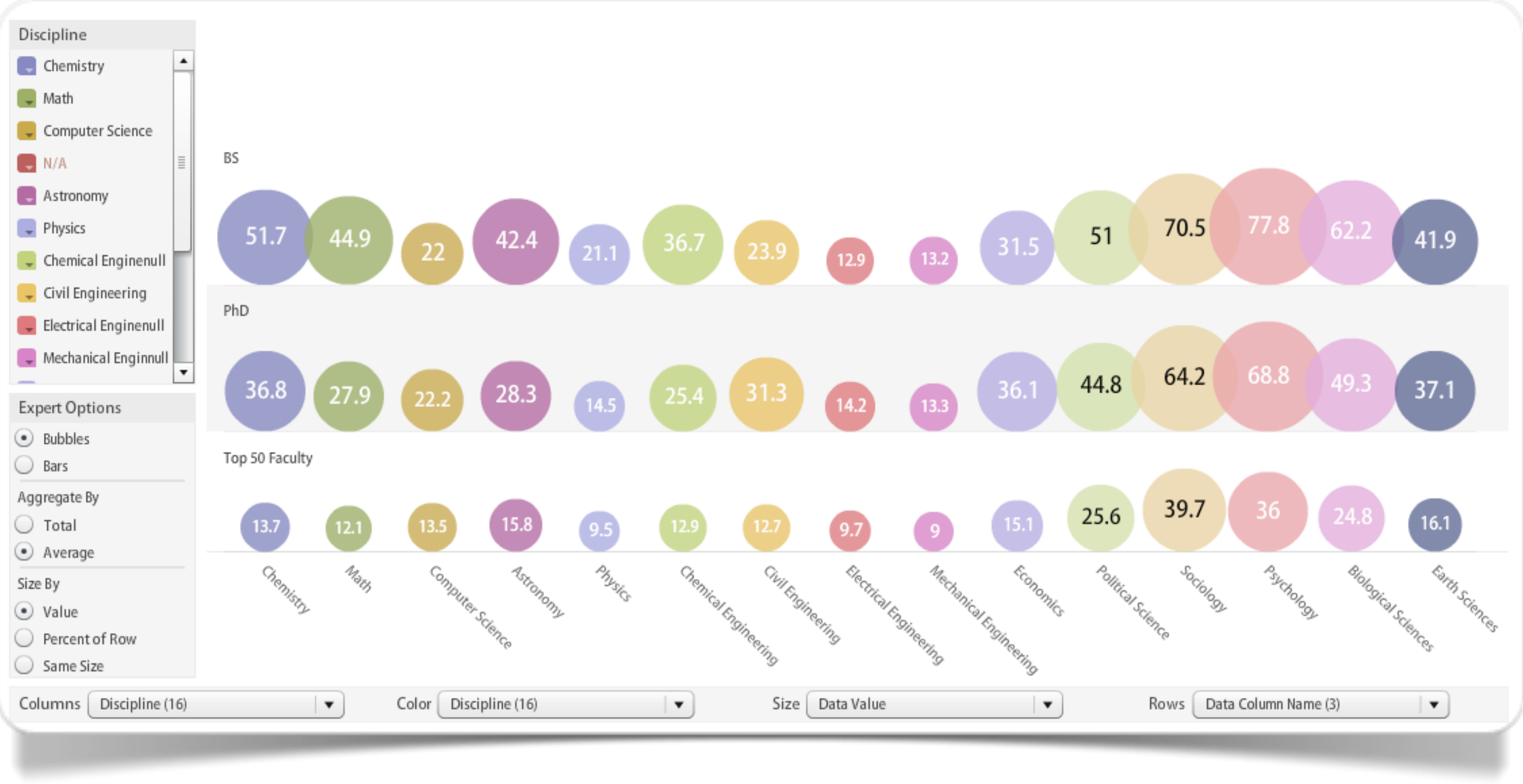


Fig. 2.17 Correlation matrix made with plotly

Relationship between many variables

Graph matrices consist of double entry grids in which at every "intersection" a single graph (scatterplot, pie chart, histogram) is reported, relative to the comparison between pairs of variables.

The Matrix Chart type of display of Many Eyes is used to represent multidimensional data in a grid. In particular, the graphical tool used to represent each cell of the grid can be either a bubble chart or a pie chart.



Expert Options

● Bubbles
○ Bars

Aggregate By

○ Total
● Average

Size By

● Value
○ Percent of Row
○ Same Size

Columns: Discipline (16) Color: Discipline (16) Size: Data Value Rows: Data Column Name (3)

Fig. 2.18 The presence of women in different academic ranks by subject area
(made with Many Eyes; currently not available)

In the case shown in figure 2.18 we see a bubble chart matrix (currently not available with Many Eyes). The data are taken from a report by Donna J. Nelson e Christopher N. Brammer on the presence of women and minorities in academic research and in scientific and technical disciplines. The three groups taken into account in the chart are graduates (BS: Bachelors of Science), PhD students (PhD), and teachers of all grades in 50 key universities in the United States. The diameter of the bubbles is related to the percentage of women in each subject area. It may be noted that there is a significant disparity in the number of women among graduates and doctorates in all disciplines. There are clear differences between the sciences and humanities disciplines such as sociology and psychology.

Chapter 3

Distribution

Data-point plot

Area

Histogram

Categorized histogram

Curve fitting

3

Distribution of a single quantitative variable

A single quantitative variable **data-point plot** allows to transfer any data on a graph by linking it to a marker point. Usually, in the simplest of its forms, the data values are plotted on the vertical axis (Y) while the horizontal axis (X) shows the number of order corresponding to the single values.

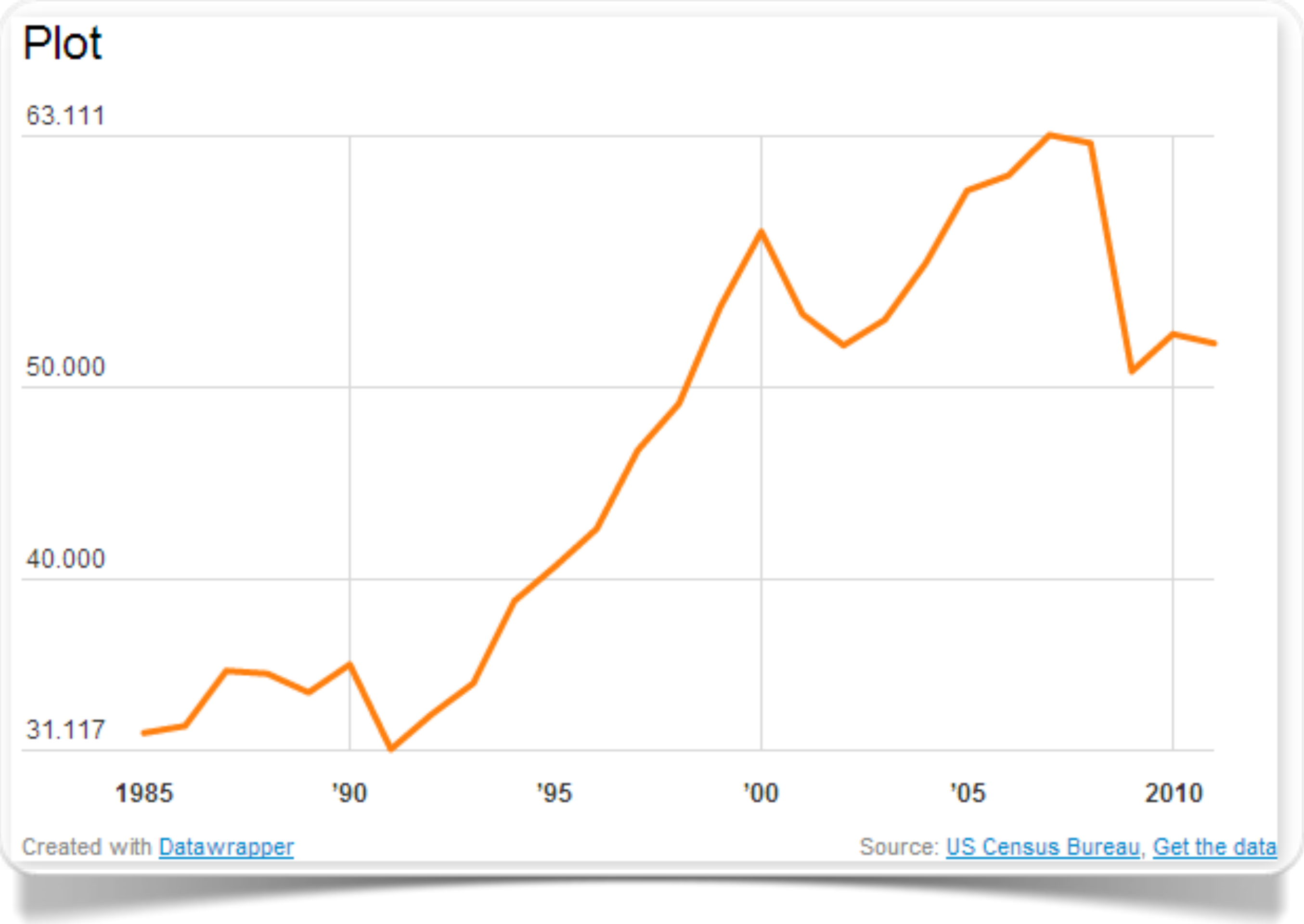


Fig. 3.1 Plot made with Datawrapper

Of all the web tools for the production "on the fly" of interactive and responsive graphs, the easiest to use is certainly Datawrapper. The graph production phases calls for 4 steps: data loading, verification and description, visualization, and finally publishing. The graph in Figure 3.1 represents the time course of a series of univariate data.

The **area** is a graph identical to the plot, with the exception of the area below the line, filled with a colour that indicates the volume.

In Figure 3.2 we can see an ideal use of the area. This being a representation of altitudes, it is self-evident that the effect achieved with the colour filling conveys the figurative idea of the heights of the peaks. The graph in Figure 3.2 was made with plotly.

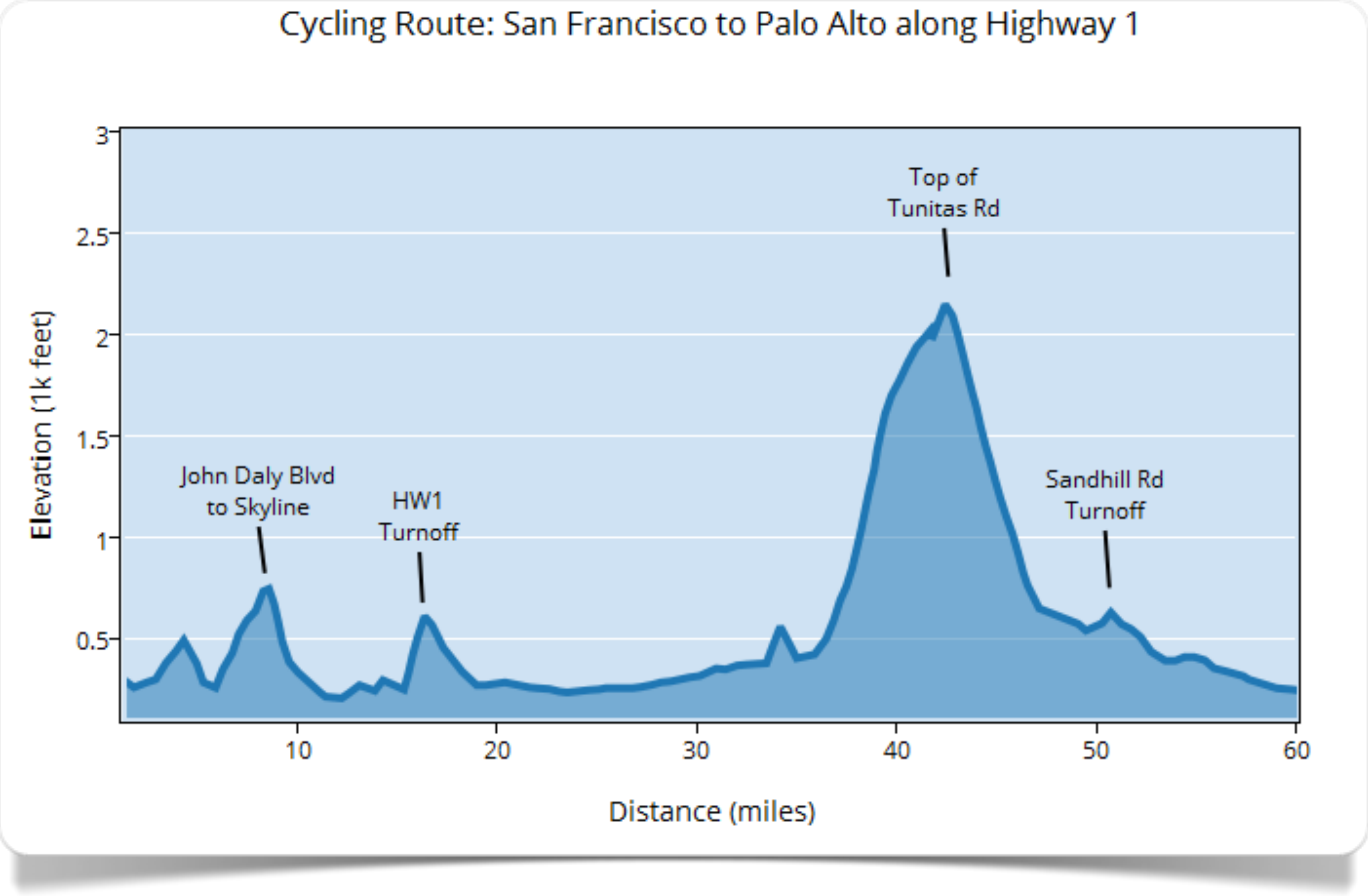


Fig. 3.2 Plot made with plotly

Distribution of a single variable (with limited data)

The histogram (Pearson, 1895) is a bar graph in which each bar represents the frequency in which a number (as in the case of quantitative variables) or a category (as in the case of qualitative variables) occurs within the **variable** taken into consideration. This type of graph is particularly effective when it deals with a limited number of cases. The graph in Figure 3.3 was made with Datawrapper.

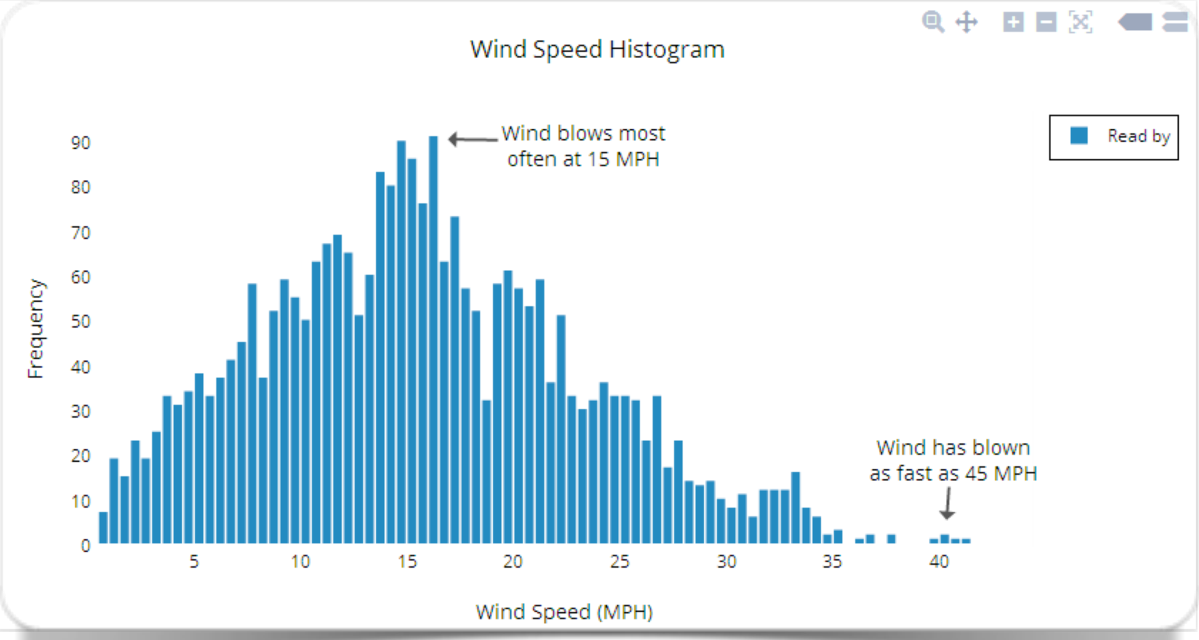
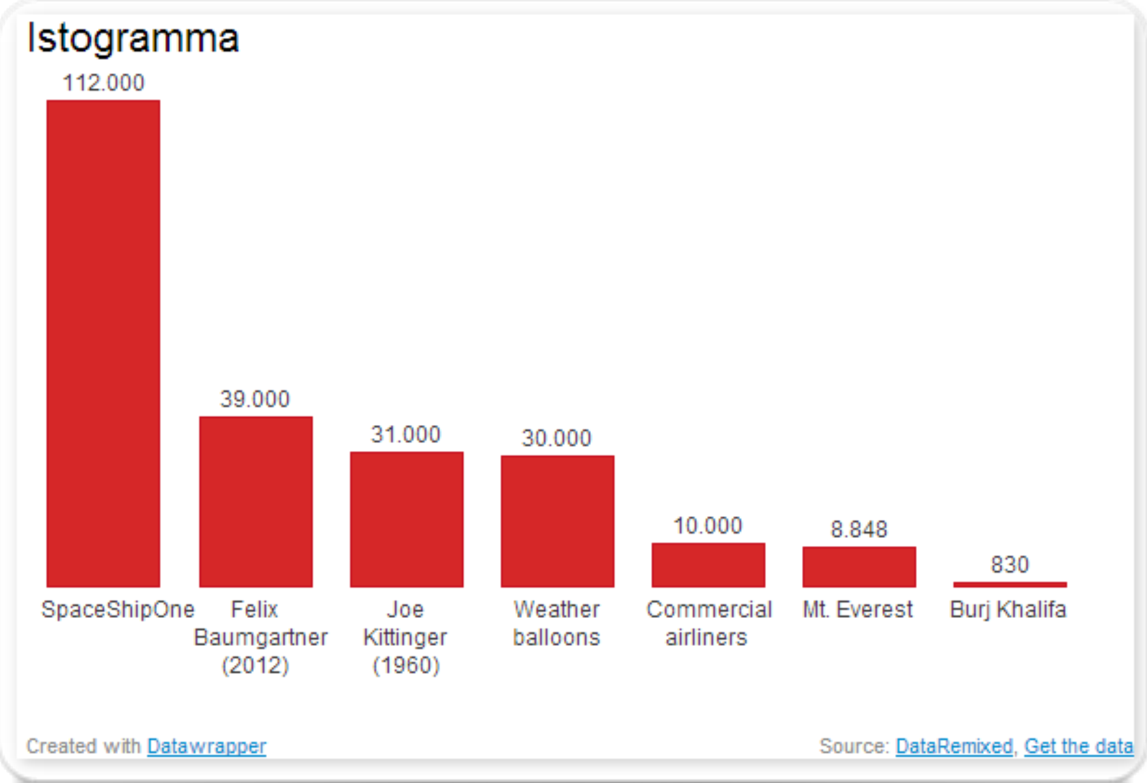


Fig. 3.4 Histogram made with plotly



Created with Datawrapper

Source: DataRemixed, Get the data

Fig. 3.3 Histogram made with Datawrapper

Viable alternatives for the production of histograms are certainly plotly (see figure 3.4), which is characterized by the possibility of writing text comments inside the graph area, or, again, Slemma (see figure 3.5).

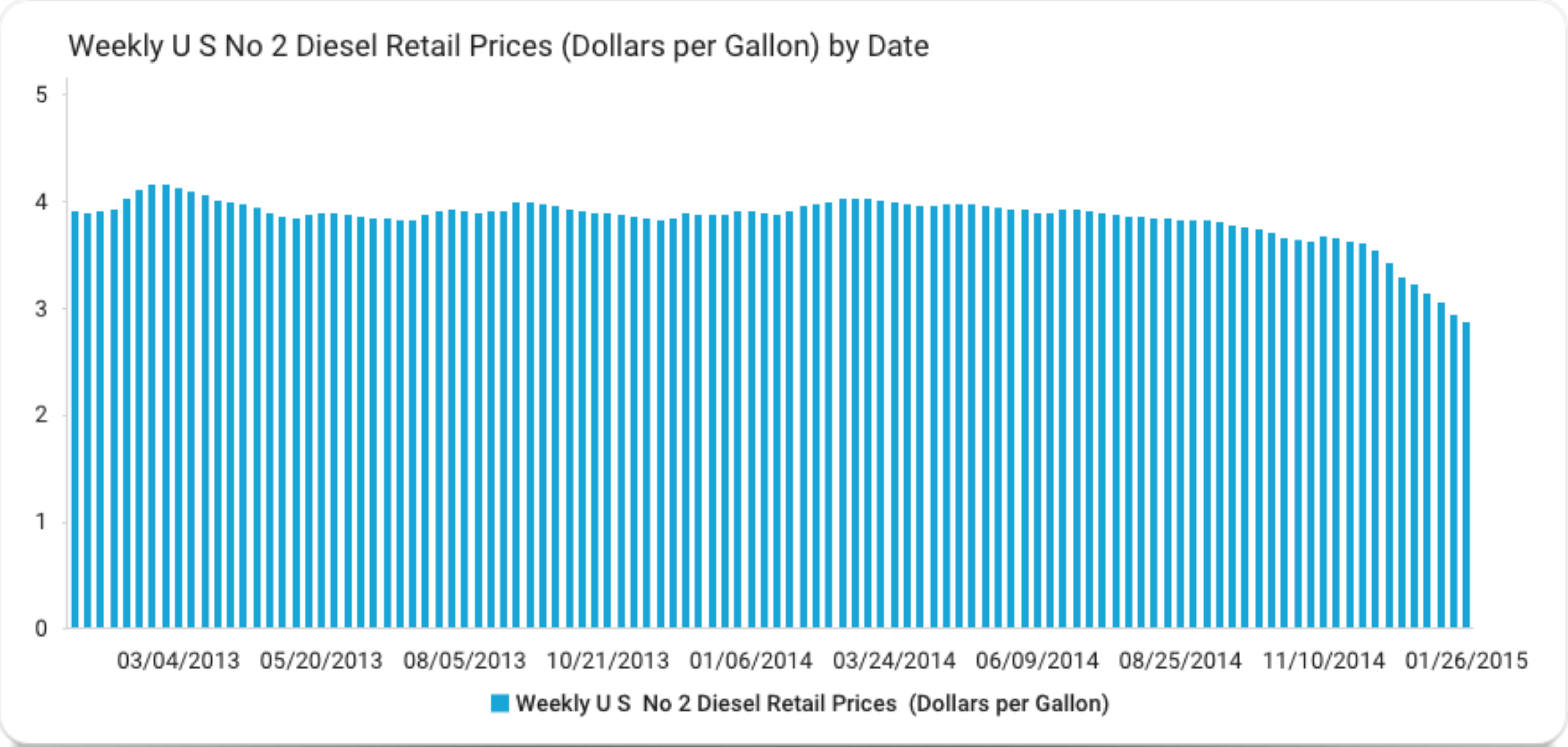
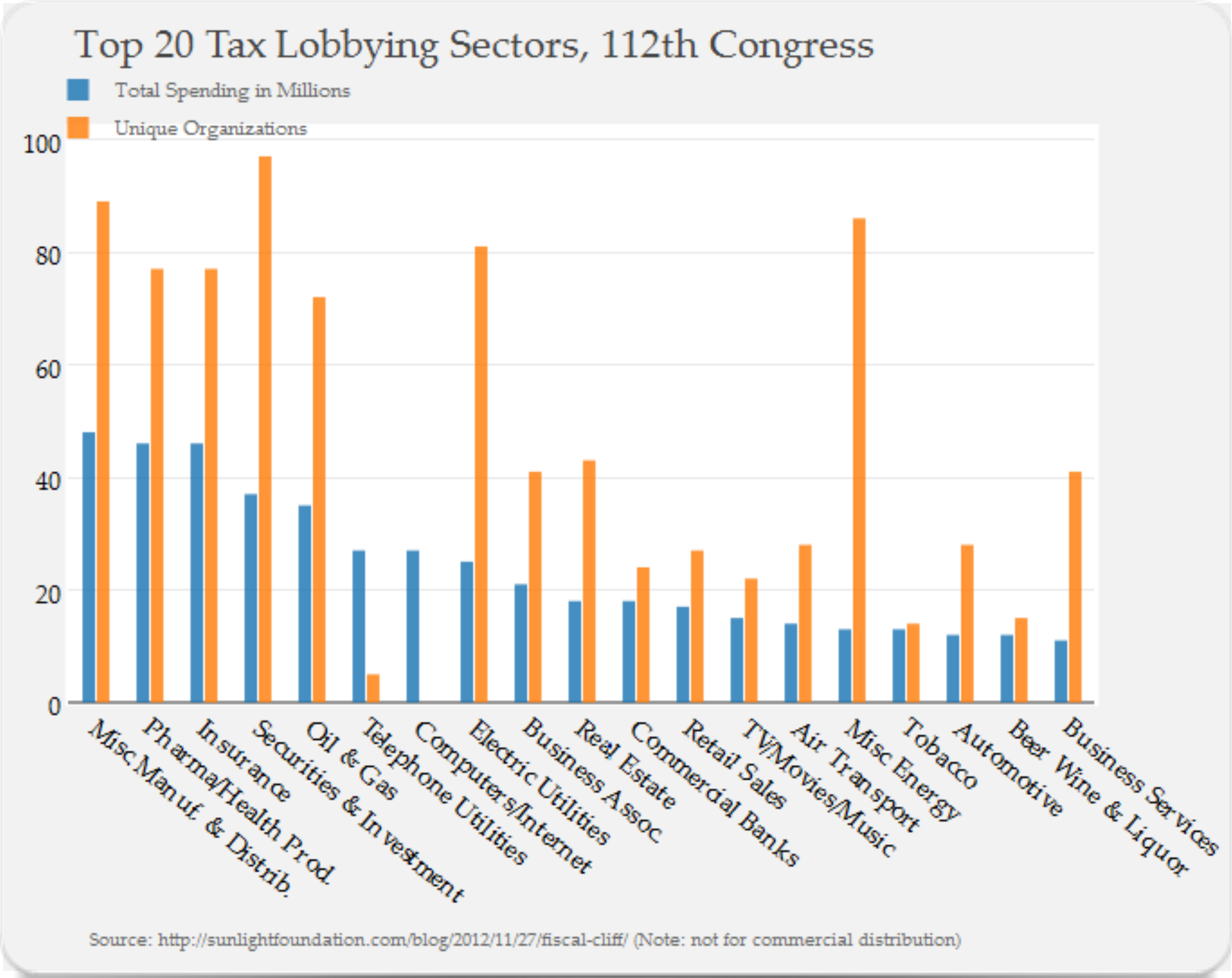


Fig. 3.5 Fuel prices by date (made with Slemma)

Distribution of multiple variables (with limited data)

A **categorized histogram** is a histogram that can represent multiple distributions at the same time. In these cases we usually use a different colour for each one of the individual dimensions involved in the comparison.


Top 20 Tax Lobbying Sectors, 112th Congress

Source: http://sunlightfoundation.com/blog/2012/11/27/fiscal-cliff/ (Note: not for commercial distribution)

Fig. 3.6 Categorized histogram made with plotly

An essential prerequisite for a correct representation of the phenomenon that is to be represented is that the ranges of variation of the dimensions to be compared have similar traits (in scope and limitations), and that the dimensions are reduced in abundance. Through plotly we can generate categorized histograms in a flawless Microsoft Excel-style (Fig. 3.6).

To make categorized histograms with IBM Watson Analytics we have to resort to a **bar chart** type of representation.

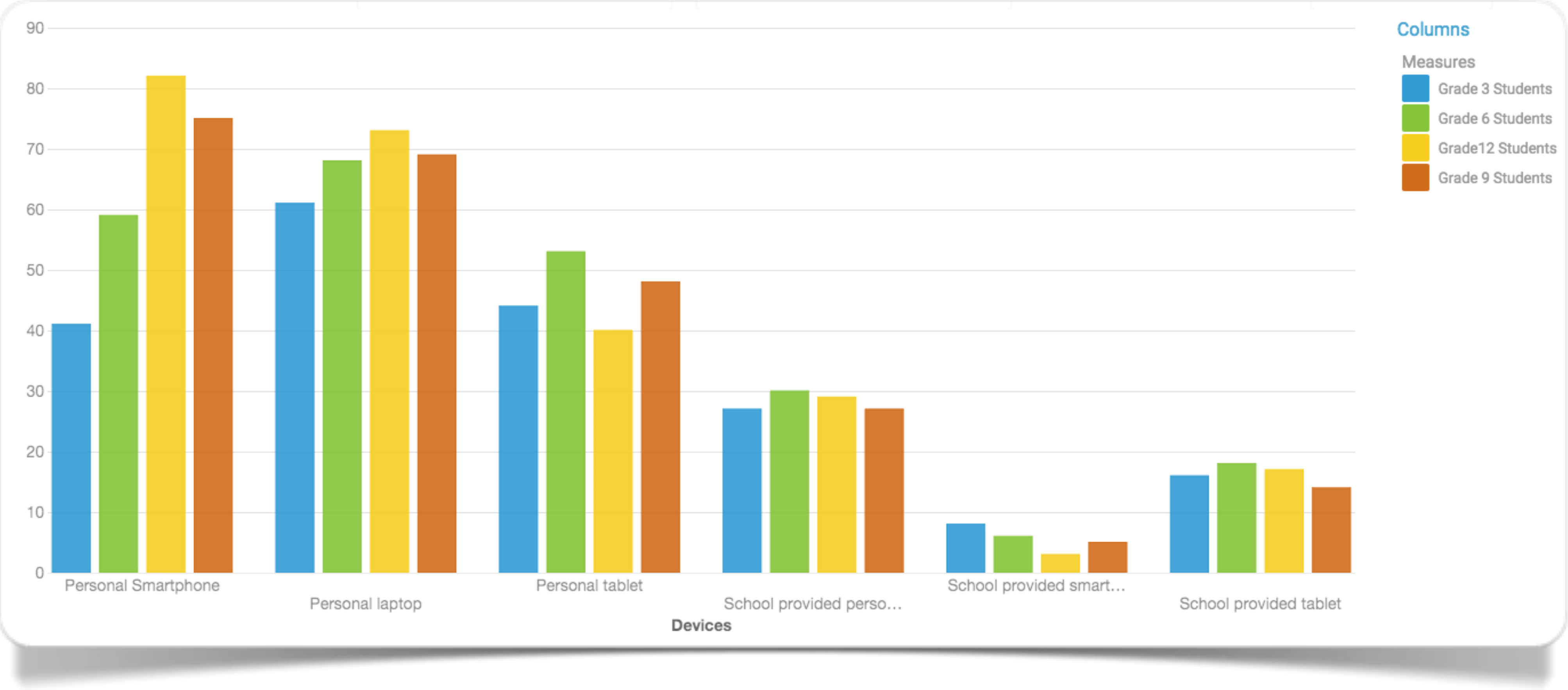


Fig. 3.7 Students access (val.%) to personal or supplied by the school mobile tools for grade level (made with IBM Watson Analytics)

In Figure 3.7 we see a graph of this type that shows on the vertical axis the percentage of students who, at school, use mobile devices: smartphones, tablets or laptops. The graph compares students who have personal devices and students that use those supplied by the school. The data are taken from a report of the Speak UP 2012 National Findings: From Chalkboards to Tablets: The Emergence of K-12 Digital Learner (June 2013). The sample covers 365,000 students from primary to secondary schools for a total of 8,000 schools. The grade levels and the type of access are represented on the horizontal axis.

Distribution of a single quantitative variable (multiple data)

Curve fitting lend itself to many uses. One of these is certainly that of representing in a "simplified" way one or more frequency distributions. The line-fitting allows to highlight some important aspects of the individual distributions: for example, through them it is possible to perceive the presence of asymmetries or subsamples from different populations.

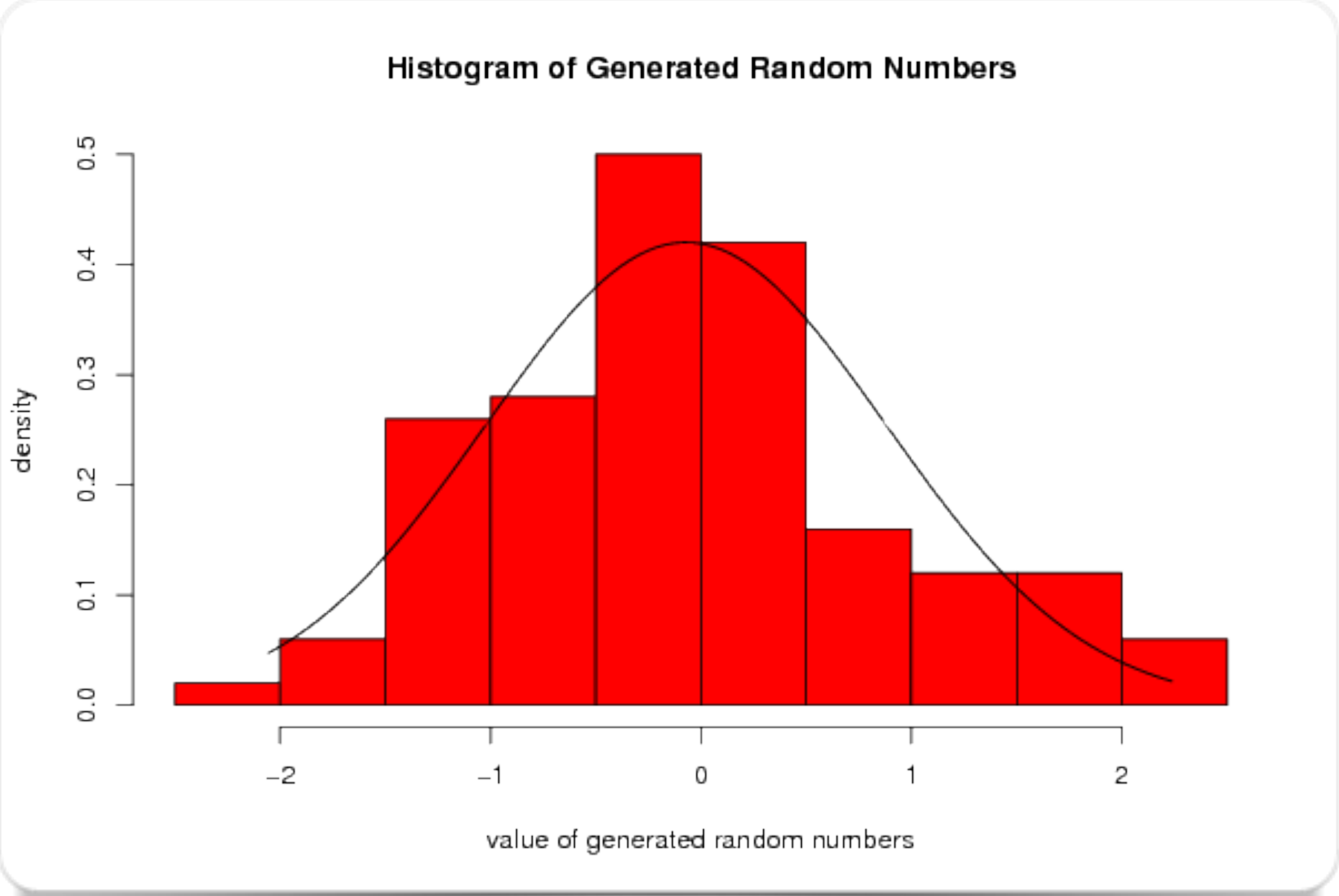


Fig. 3.8 Curve fitting made with Wessa

This example of curve fitting (Fig. 3.8, click on *Compute*) was made using Wessa. In particular, the histogram and the curve were made following a random data generation with normal distribution.

For this example R *MASS* and *msm* libraries were used.

Chapter 4

Distribution and composition

Pie chart

Treemap

Circle packing

Stacked bar chart

Stacked area chart

Stack Graph

4

Distribution and composition of a single categorical variable

The **pie chart** (Playfair, 1801) is one of the most intuitive graphical representations: it depicts the frequency distribution of a categorical **variable** (whether nominal or ordinal) when the available categories are limited in number. A fundamental and intuitive condition for the representation to be considered reliable is that the sum of the frequencies (percentages) of all available categories is equal to 100%.

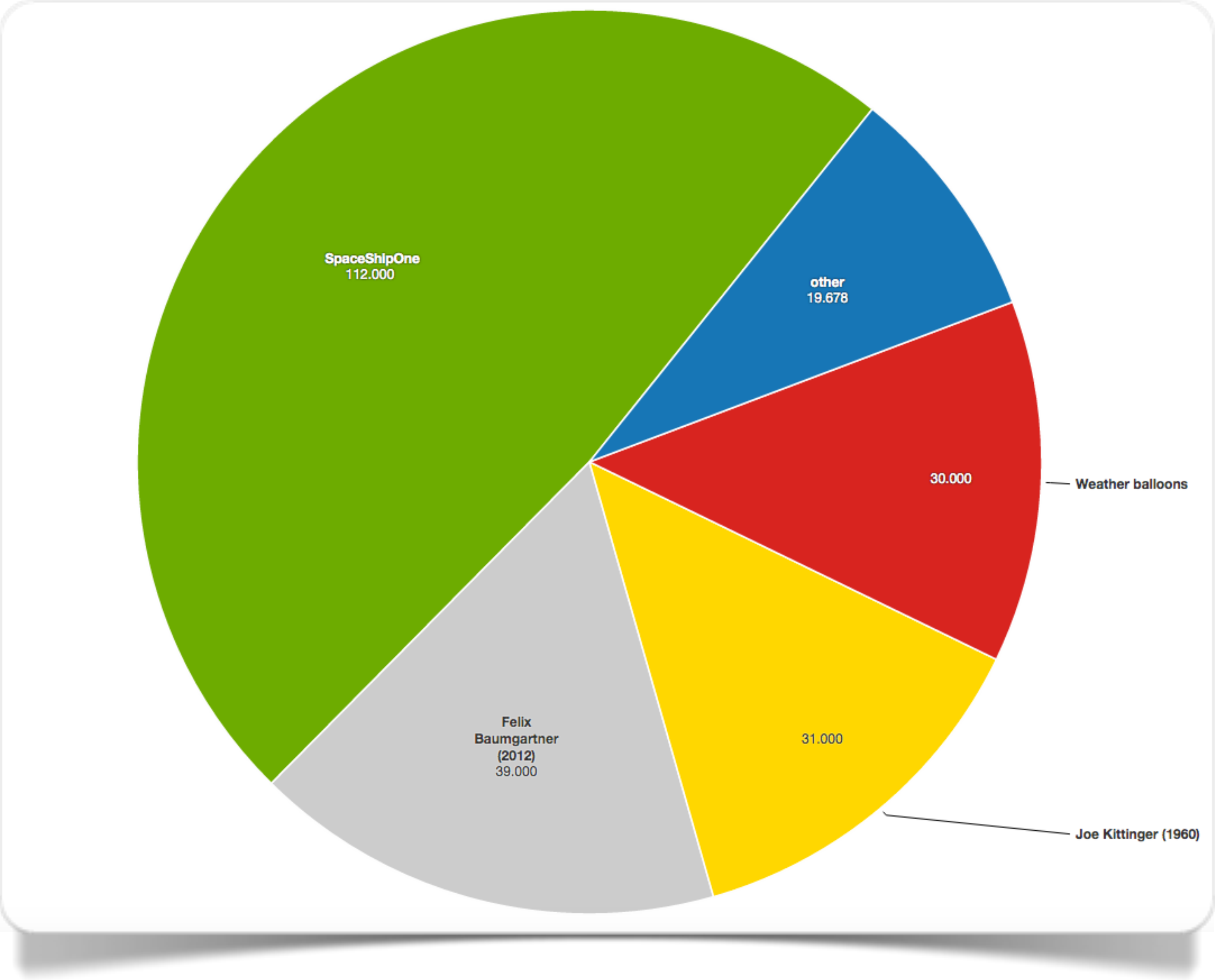
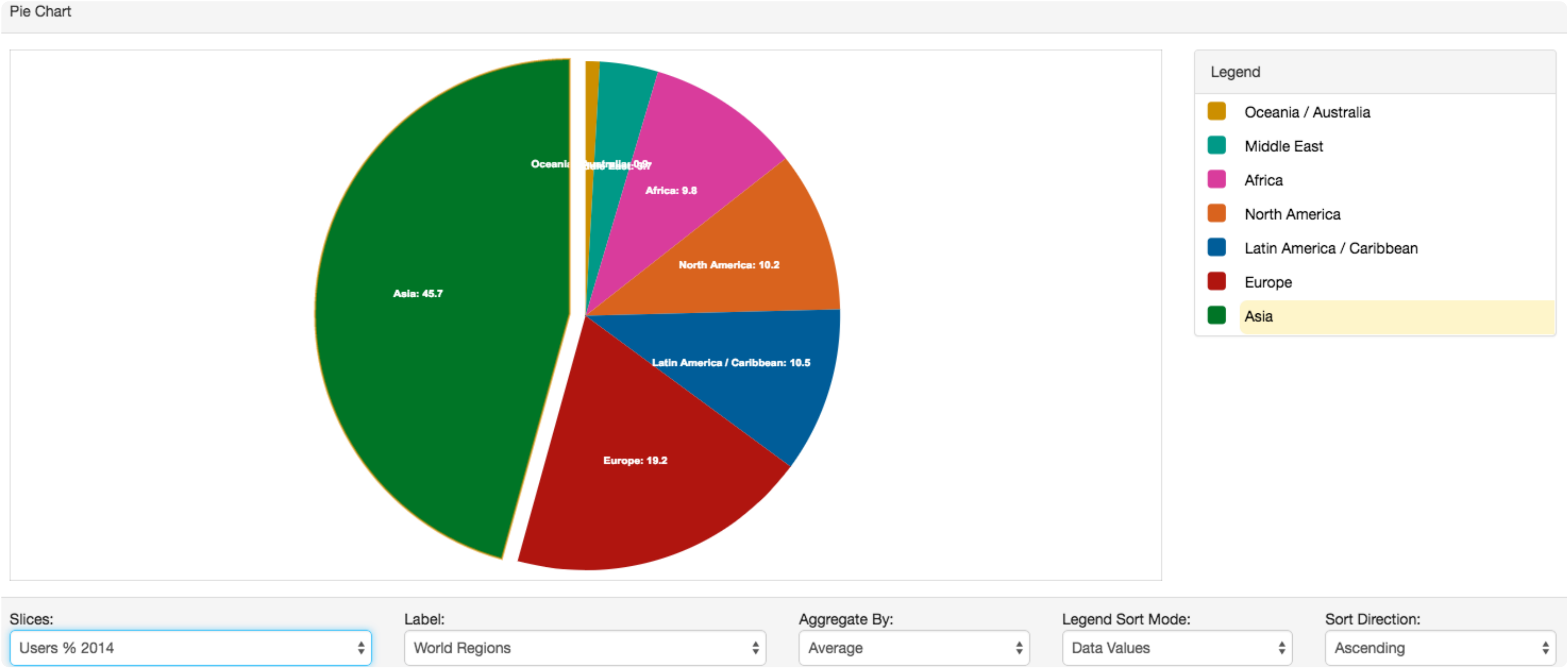


Fig. 4.1 Pie chart made with Datawrapper

Among Datawrapper available representations there are the pie charts, packed with many options for colour, label and size customization.

To make pie-charts with Many Eyes we have to use the Pie Chart type of visualization. Among the possible options, the *Slice size* function lets you update the data on the basis of categorical information (eg., the year of reference).

In the gallery 4.1 we can see the interactive pie chart that allows to compare the distribution of Internet users in 2014 and 2000 (source: Internet World Stats). As we can see, in fourteen years the situation has changed significantly: North America and Europe, grouped together, have given way to Asia.



Slices: Users % 2014

Label: World Regions

Aggregate By: Average

Legend Sort Mode: Data Values

Sort Direction: Ascending

Gallery 4.1 Internet users in 2014 and 2000

Internet users by world region in 2014 (scrolling graph – made with Many Eyes)

The **treemap** (Shneiderman, 2009) is an alternative to the pie chart: the function is the same (ie the representation of a frequency **distribution**), however, the treemap stands out for its ability to represent sub-distributions in a hierarchical way. Each "quadrant" is equivalent to a category, which in turn may represent the sum of the units belonging to a limited set of sub-categories.

To treemaps chart with IBM Watson Analytics has a number of useful, interactive options. The example in Figure 4.2 allows us to compare the incom that the highest-grossing films in the last ten years made in their first year of release.

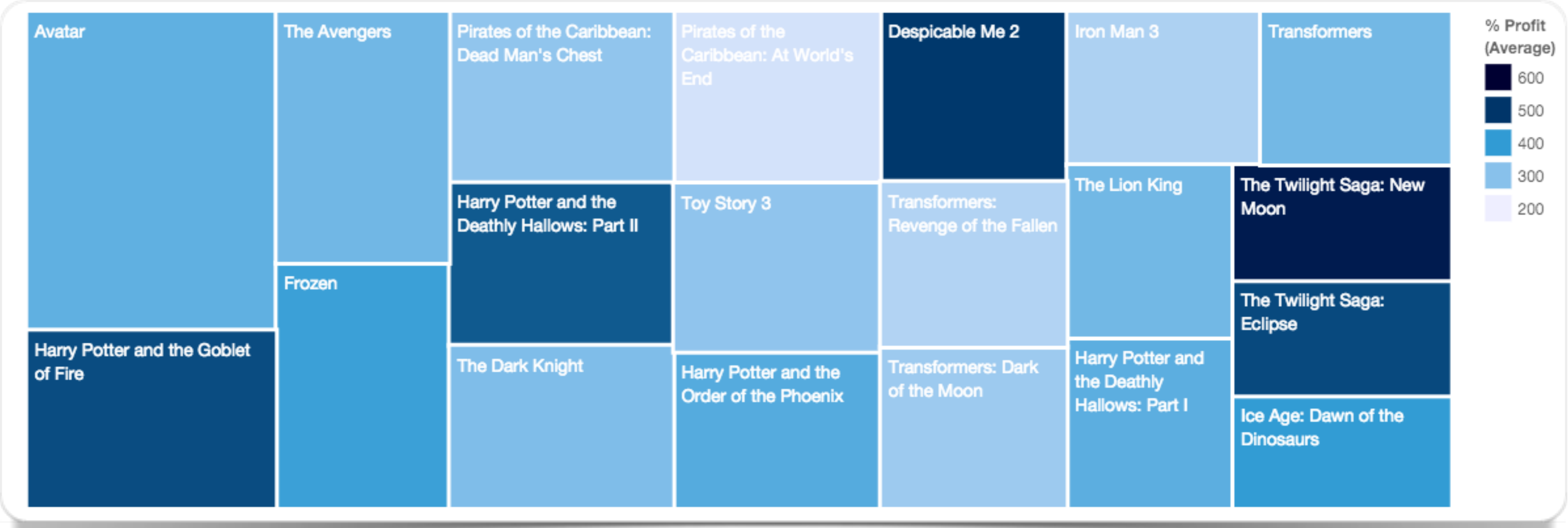


Fig. 4.2 Most Profitable Movies - Last 10 Years (made with IBM Watson Analytics)

The largest block is occupied by the highest-income film of all: *Avatar*. The intense blue colouring, which decreases getting closer to the white colouring, tells us that *The Twilight Saga: New Moon* is the film with the highest return on investment (% Profit).

The treemap graph of Watson Analytics allows many other options. In addition to the displaying of information by moving the mouse over the area of interest, it is possible to display data for the best profitable movie genres (Fig. 4.3).

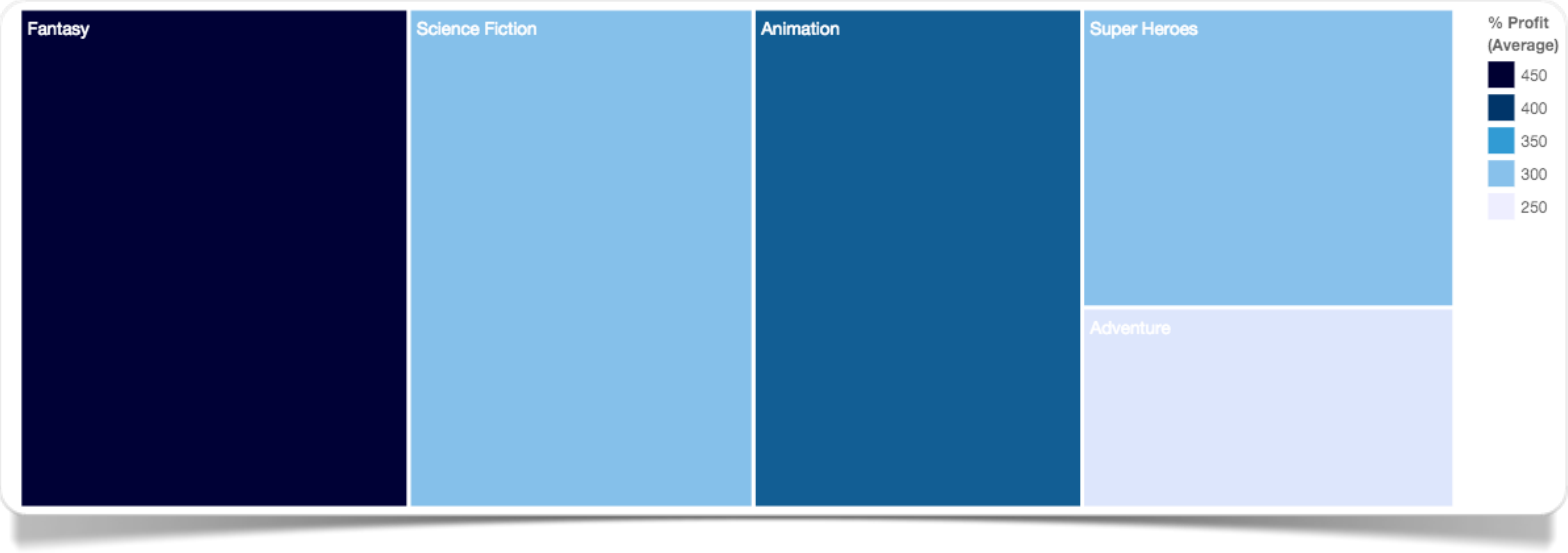


Fig. 4.2 Most Profitable Movies by genre - Last 10 Years (made with IBM Watson Analytics)

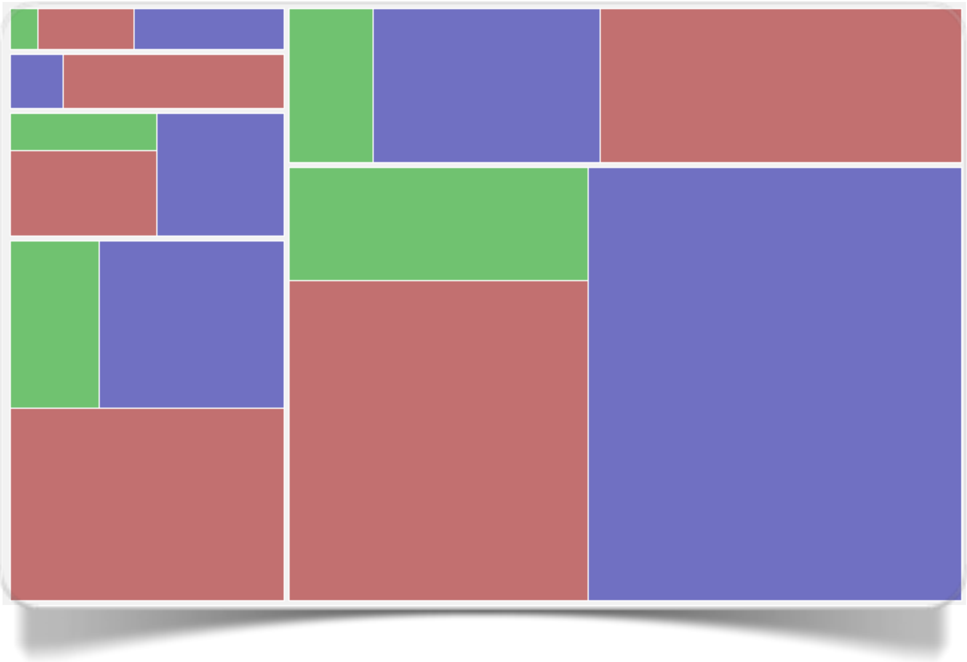


Fig. 4.4 Treemap made with Raw

With Raw it is possible to make treemaps through the use of a single, simple interface (fig. 4.3). Unlike many other similar tools, Raw does not save the data on a database.

The **circle packing** is a particular geometric study that is based on the occupation of areas or surfaces with a series of circles of equal or variable radius, with the main goal being to avoid overlaps. It is only since a relatively short time that this type of study is used to create statistical representations similar in purpose to the treemap: in this case, the arrangement of the graph is such that both the categorical variables for higher hierarchies and those relating to the lower hierarchies respect the principle at the base of the circle packing.

Through Raw it is possible to make circle packing inside circles (see Wikipedia). To make them in Raw we can select as many categorical variables as the hierarchies we want to arrange in circles. The size of each circle is defined not by the frequency of each combination of hierarchies but by the specification of a numerical **variable**.

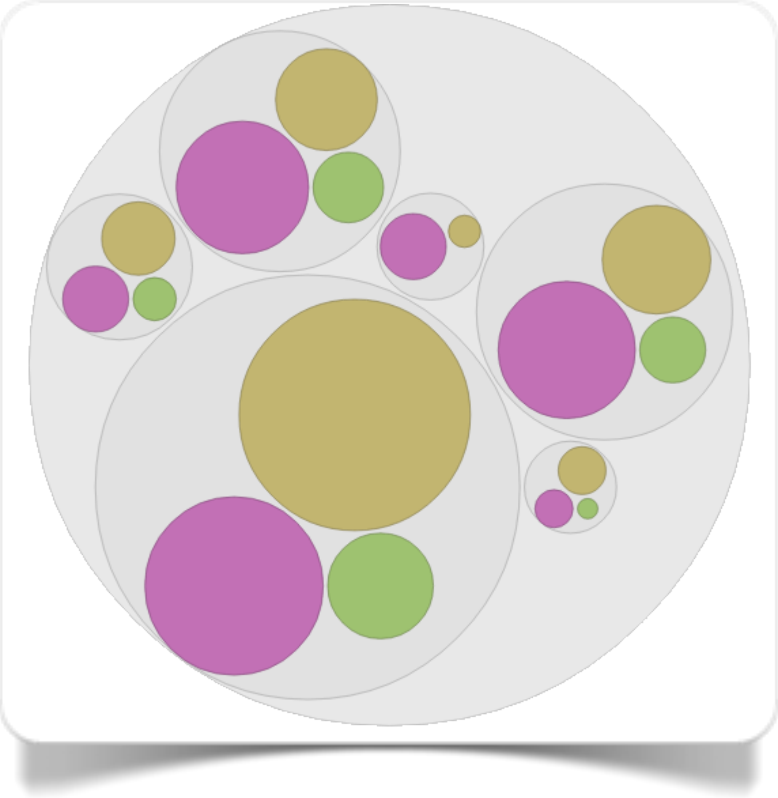


Fig. 4.5 Circle packing made with Raw

D&C of a single categorical variable in a segmented chart (absolute values)

The **stacked column chart** is the ideal tool to visualize the distribution of the occurrences of each category of a specific categorical variable (qualitative) across the different levels of a second categorical variable (X).

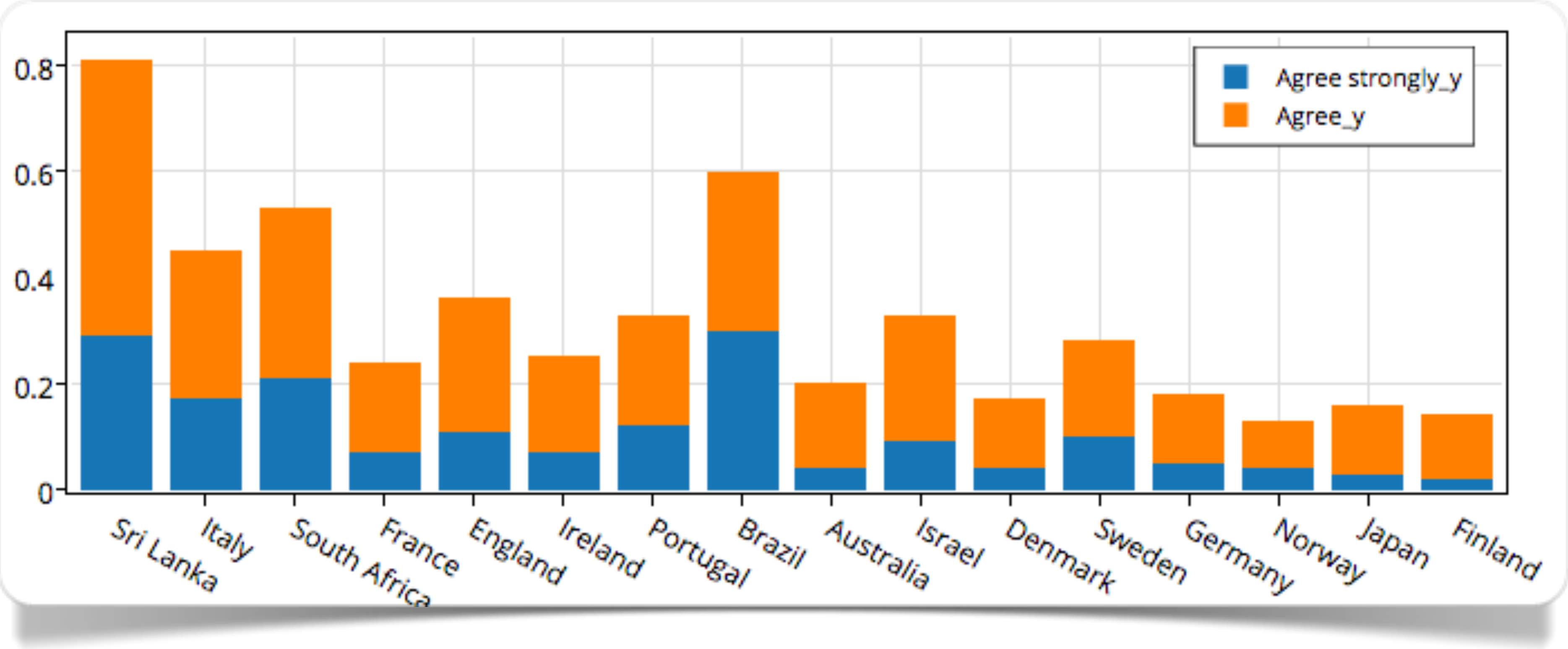


Fig. 4.6 Stacked column chart made with plotly

plotly allows to make stacked column charts using the Bar Charts type of visualization (Fig. 4.6).

Also Datawrapper offers the ability to create stacked columns charts as in the example (Fig. 4.7).

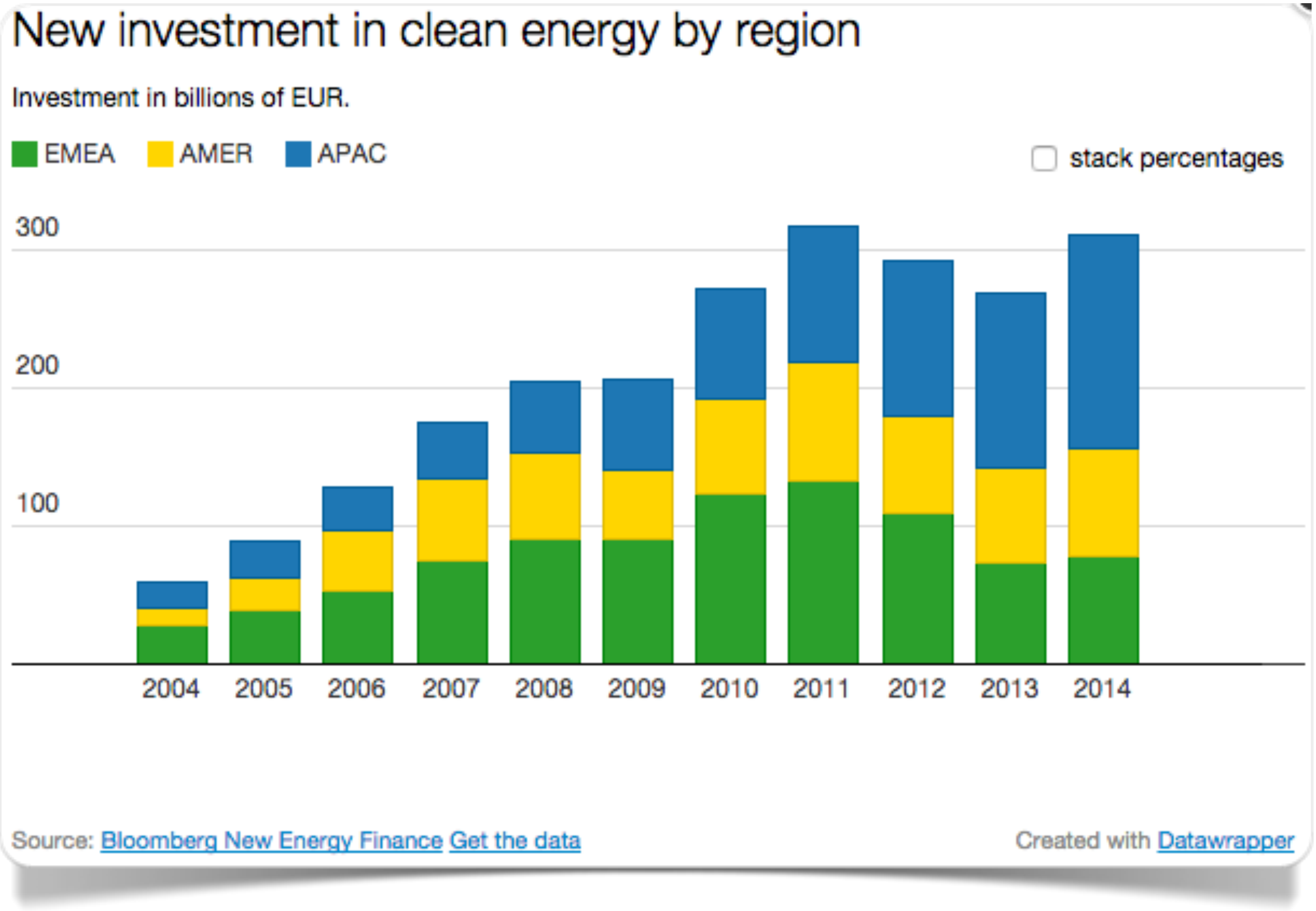


Fig. 4.7 Stacked column chart made with Datawrapper

The **stacked area chart** is a corrected version of the stacked column chart. The difference lies in its typical “continuous” representation of values along the horizontal axis. The stacked areas find their ideal application when the horizontal axis is given a temporal dimension: the characteristic "continuous" course of the lines that outline the areas allows to reveal in an optimal way possible trends and developments over time.

The **Stack Graphs** are among the most aesthetically effective graphics available in Slemma. In figure 4.8 we can see the distribution of population by county in a time series from 1841 to 2011 (source: Central Statistic Office Ireland). The amount of the population is represented on the vertical axis, while the years of the survey are represented on the horizontal axis. In the interactive chart it is possible to select the visualization of a single county or the percentage distribution of the population in the counties for each year. This allows, for example, to see more clearly the effect of urbanization, with the growth of the counties of Dublin and Cork at the expense of the others.

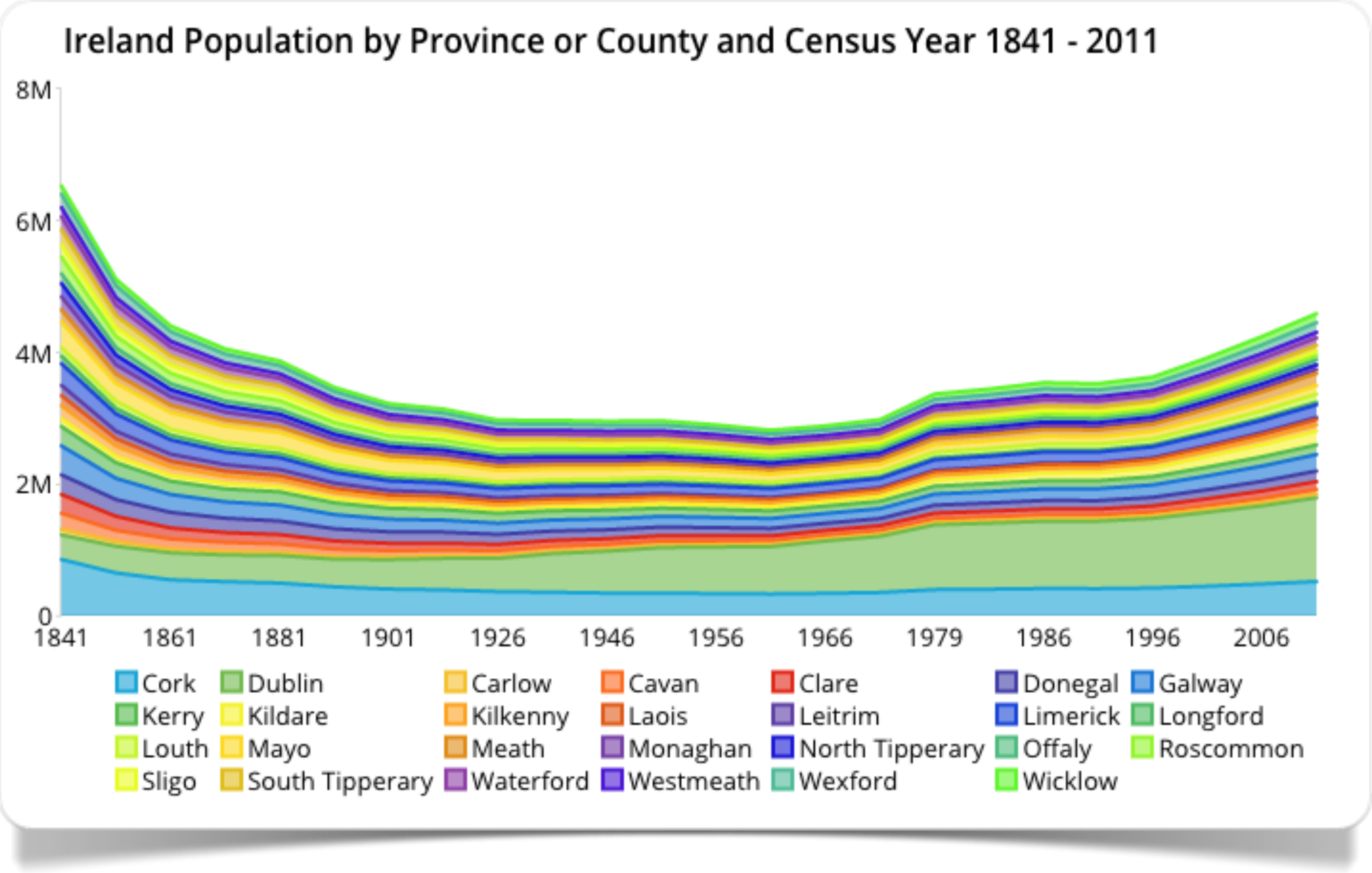


Fig. 4.8 Ireland resident population by county from 1841 to 2011 (made with Slemma)

D&C of a single categorical variable in a segmented chart (% val.)

The **stacked bar chart** is the ideal tool to represent the frequency **distribution** (of percentages) of each of the categories of a specific categorical (qualitative) variable across the different levels of a second categorical variable (X). The Bar Charts graph creation option available in plotly allows to easily make stacked bar charts.

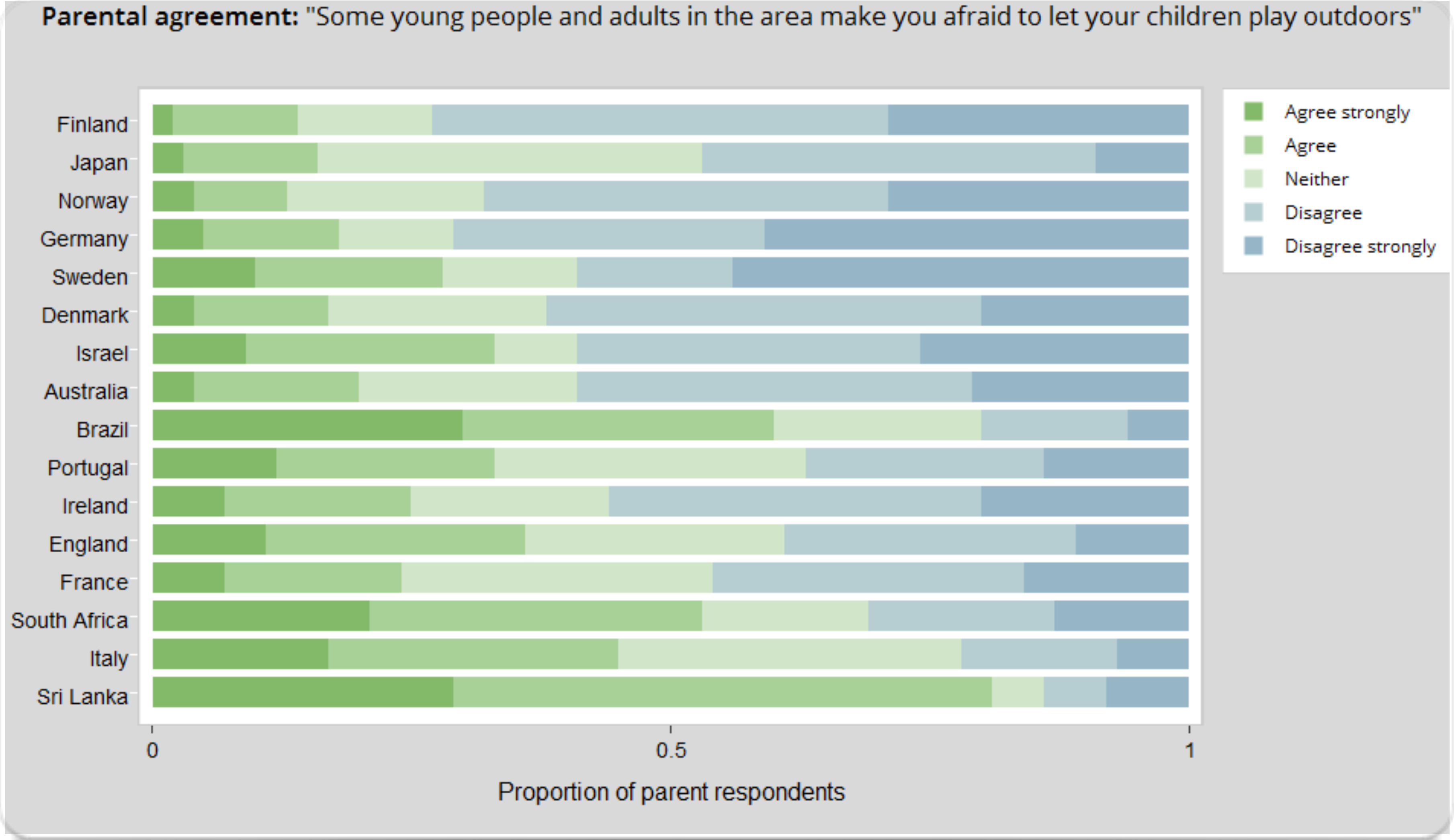


Fig. 4.9 Stacked bar chart made with plotly

This is done by selecting the categorical variable as the X or Y axis of the graph (in fig. 4.9 the variable “countries” was selected as the Y axis) and making a multiple selection of numeric variables (percentages) for a final sum of 100.

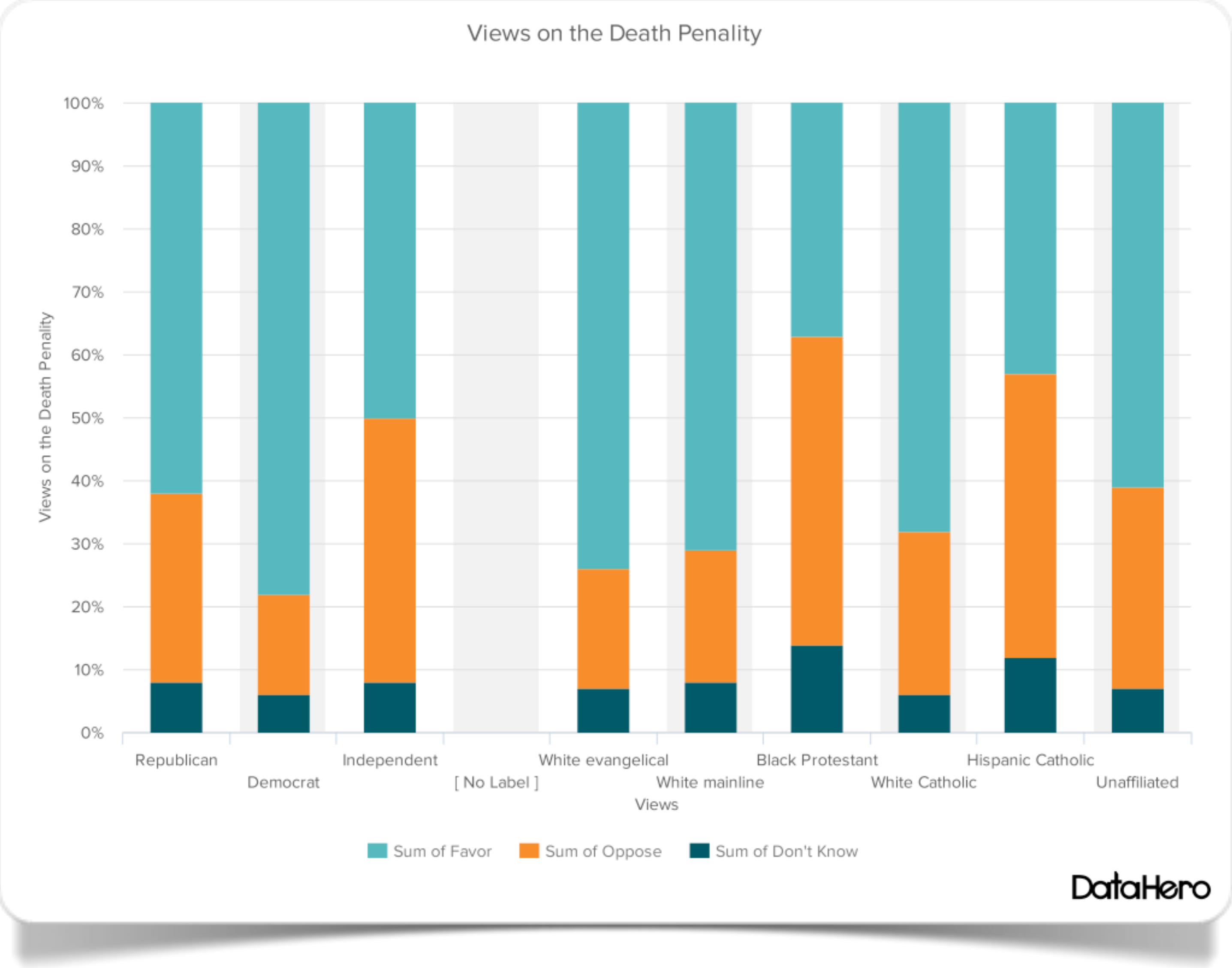


Fig. 4.10 Opinions on the death penalty in the United States for political orientation and religious affiliation (val.% - made with DataHero)

DataHero allows to create stacked bar charts using the Stacked Column type of visualization. If we choose the option % we can use the same size option to carry over on the chart the size of the different categories on a proportionate basis. Figure 4.10 shows the opinions of a sample of US citizens about the death penalty (source: Pew Research Center Survey 2010). The percentage values of the possible answers (favor, oppose, don’t know) are represented on the veritcal axis. The categories are stacked on the horizontal axis and divided into two main groups: political orientation and religious creed, with a sub-category for Caucasians, African-Americans and Hispanics. As we can see there are still many in favor of the death penalty. We can find the most opposed to it among African-American Protestants and Hispanic Catholics. The ethnic-cultural identity is stronger than the religious identity.

The **stacked area** chart can be used for percentages as an improved version of the stacked bar chart. The stack graphs are among the most aesthetically effective charts available in Slemma. Thanks to the Percentage option we can represent the percentage distribution of the frequencies of the different categories (Fig. 4.11).

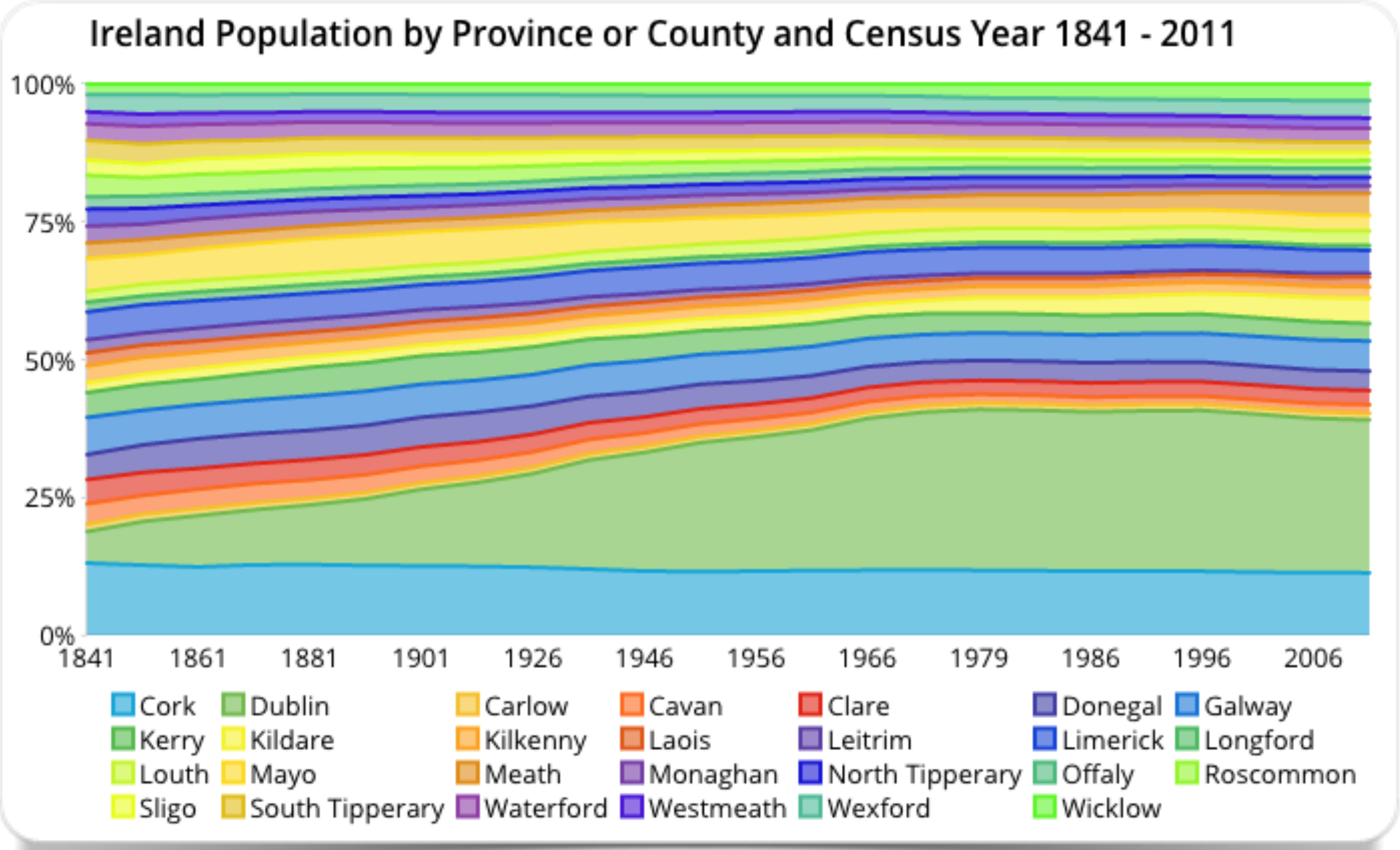


Fig. 4.11 Ireland resident population by county from 1841 to 2011 (made with Slemma)

Chapter 5

Relationship and distribution

Block histogram

5

Relationship and Distribution: categorical variable vs a single numeric variable

The **block histogram** is a unique chart whose main objective is to represent the distribution of the classes of a categorical variable in comparison to the values of a numeric variable. Usually, its use calls for the representation of the numeric variable on the horizontal axis and the determination of its subintervals, on which a series of blocks, each corresponding to a specific class of the categorical variable, has to be “stacked”. The vertical arrangement of the blocks (classes) is such that the classes to which a higher value of X is given - in the corresponding sub-range of values - will be higher on the vertical axis with respect to those with a lower value of X.

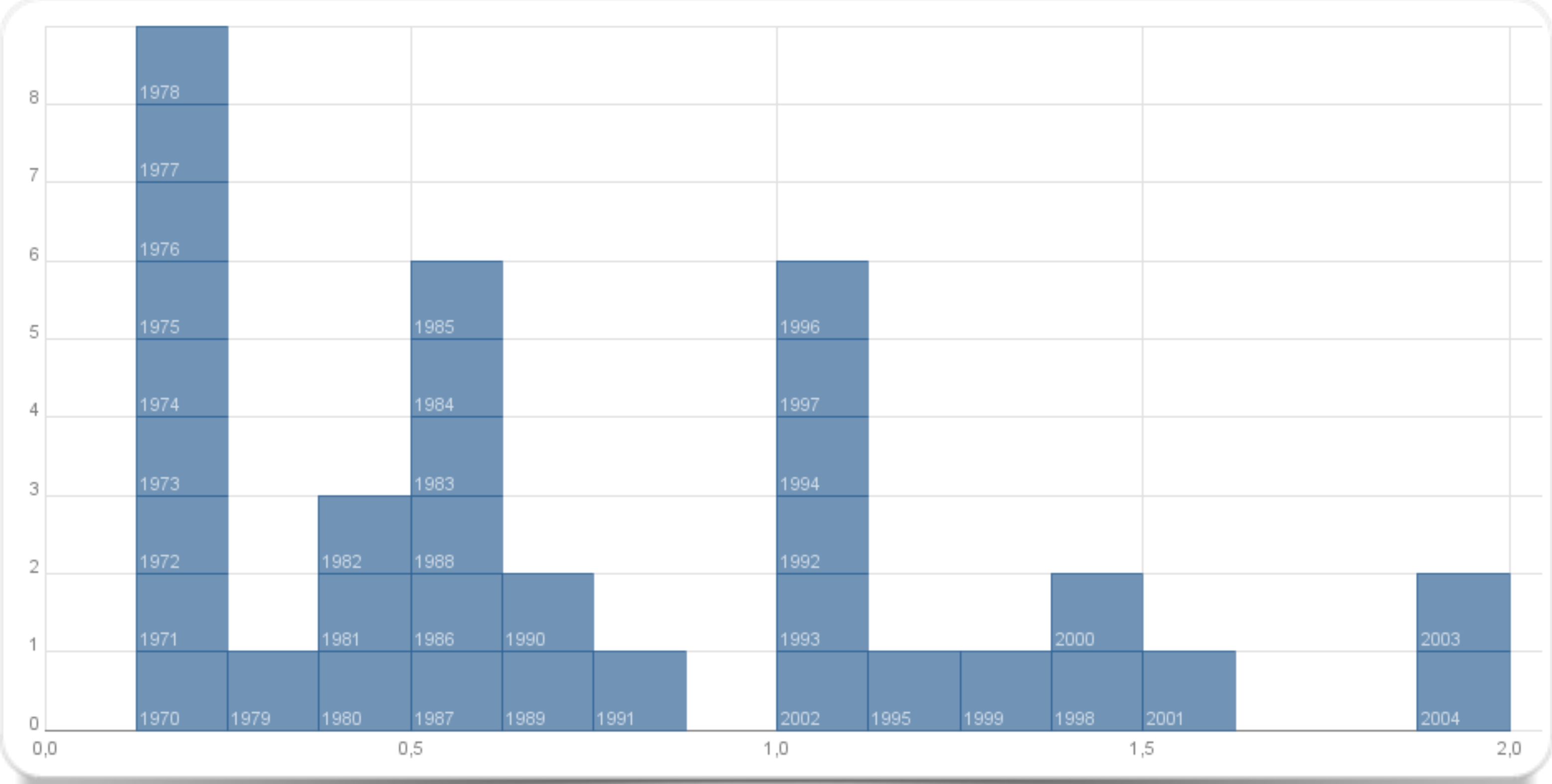
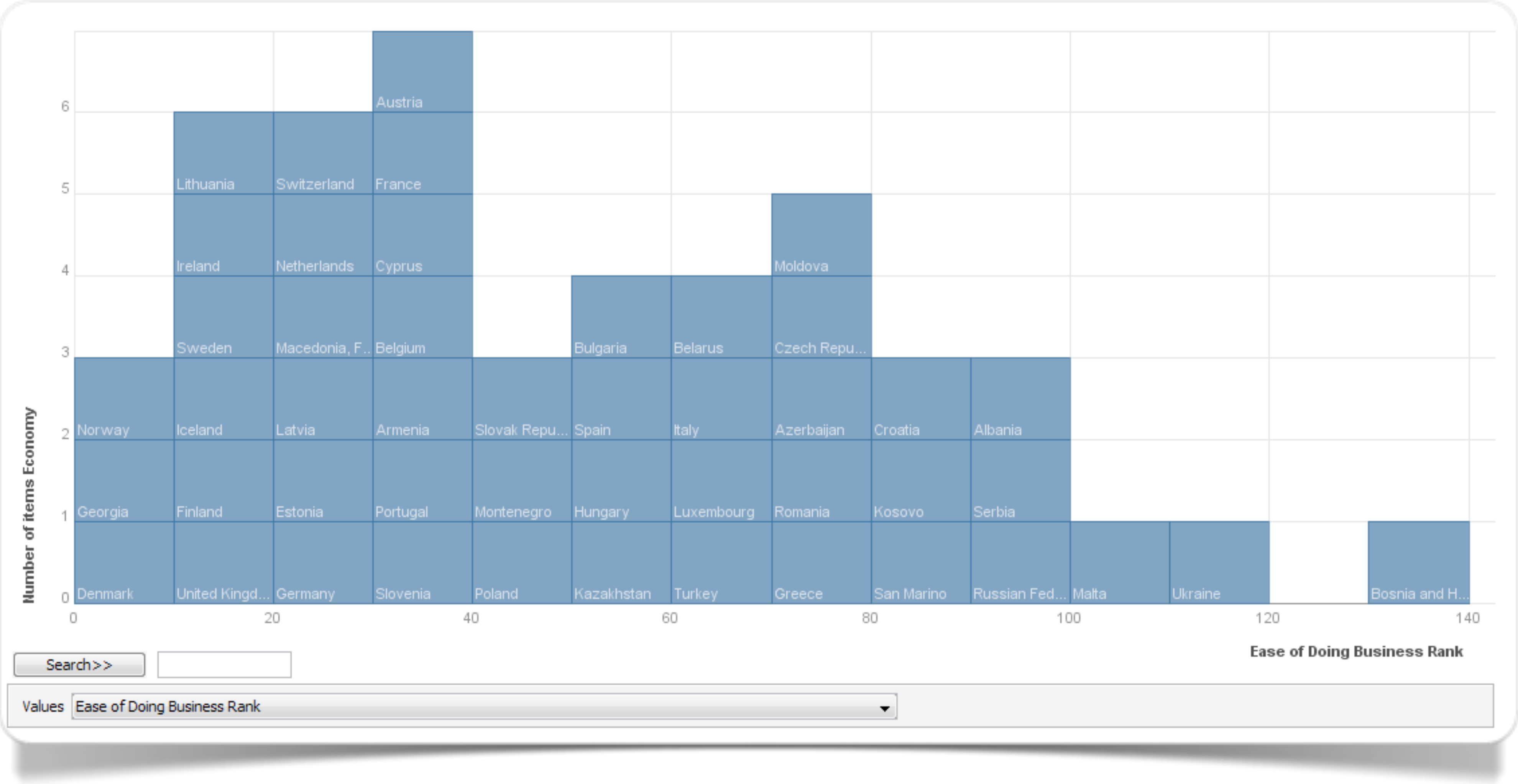


Fig. 5.1 Block histogram (stacked area chart) made with Many Eyes

In the example of figure 5.2 we can compare the European countries according to some economic indicators (this tool of Many Eyes currently is not available).



Values Ease of Doing Business Rank

Fig. 5.2 Block histogram showing the ranking of the different European economies (elab. Many Eyes)

Chapter 6

Relationship and composition

Parallel coordinates

6

Relationship and composition of many quantitative variables

The **parallel coordinates** (d'Ocagne, 1885; Inselberg, 1985) are a tool used to visualize and analyze multivariate data. In the graphic area there are as many vertical lines (axis or dimension), parallel and equidistant from each other, as the numeric variables to be represented.

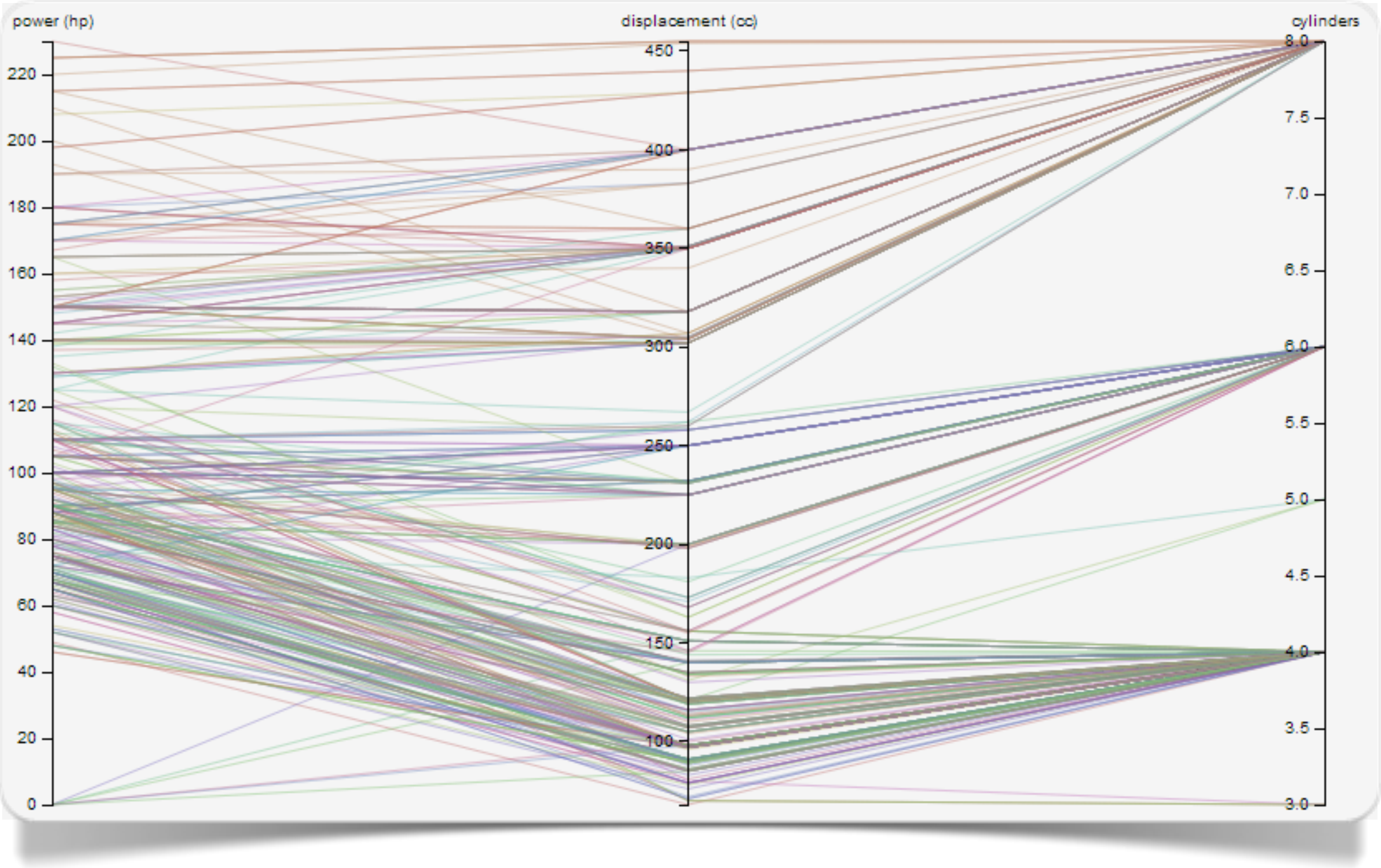


Fig. 6.1 Parallel coordinates made with Raw

The statistical units are represented by a broken line whose apex lies exactly in correspondence of each vertical line. The position of the apex on the i-th axis corresponds to the i-th coordinate of the point (see Wikipedia).

Raw allows to select a number n of numeric variables to be used as vertical lines (axis or dimensions). Moreover, it is possible to select a categorical **variable** as to identify, for example, units belonging to the same class with a single colour.

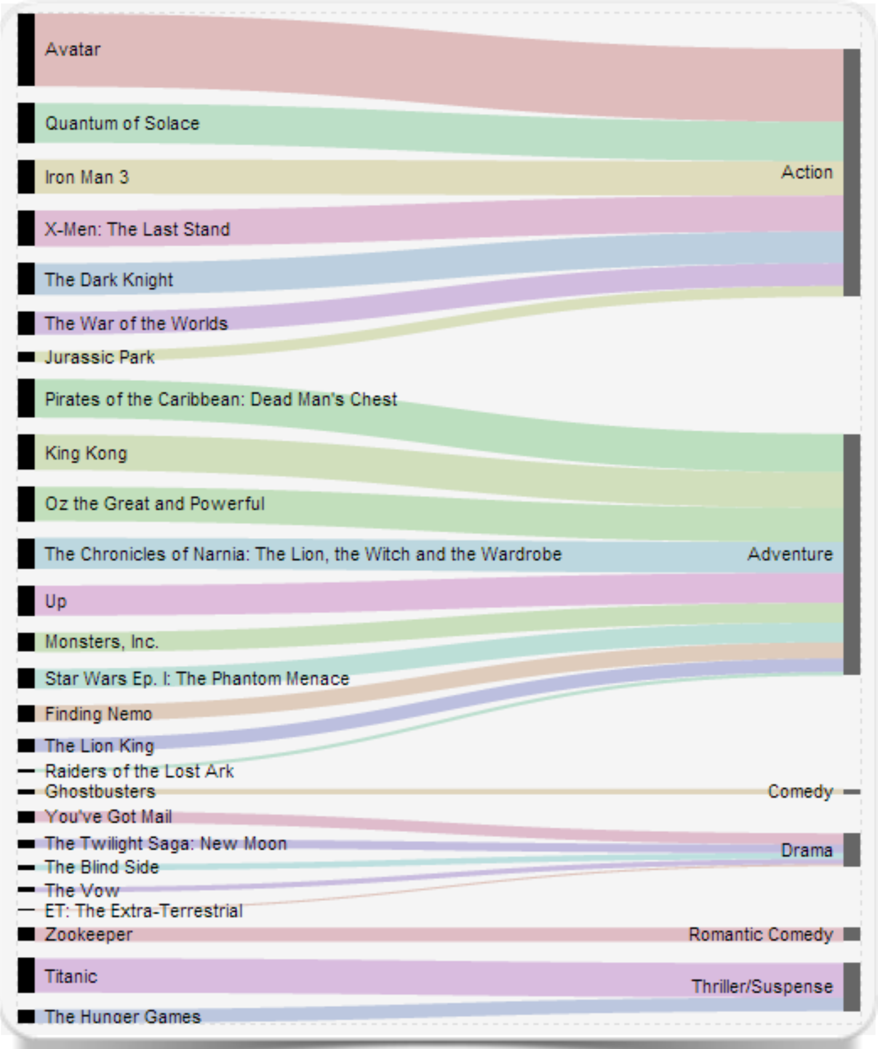


Fig. 6.2 Alluvial diagram made with Raw

The **alluvial diagram** is a unique type of **flow diagram**, in which the qualitative variables are organized in parallel lines and blocks (clusters of nodes). In the transition from a qualitative variable to another one, we can see branchings or unifications of different streams based on the contemporary affiliation of each individual statistical unit to the different classes of the qualitative variables involved. Through one or more quantitative variables of size, a “thickness” can be assigned to the flow of each cluster.

Raw allows to select a number of qualitative variables to be used as steps of the alluvial diagram. Optionally, we can also use numeric variables to make new steps.

Chapter 7

Comparison and distribution

Box and whiskers plot

Bagplot

7

Comparison and distribution between measurements of position and measures of dispersion

Known as **Box-plot**, or as **box and whiskers plot** (Tukey, 1977), this type of chart is mainly used in statistics to compare positions (mean, median, etc.) and measures of dispersion (standard deviation, interquartile range, etc.) across different groups of units belonging to the same variable. A further advantage of this representation lies in the possibility of interpreting the distributive nature of the data within each group. The whiskers, in particular, help to highlight a greater or a smaller dispersion below or above the respective measurement of position.

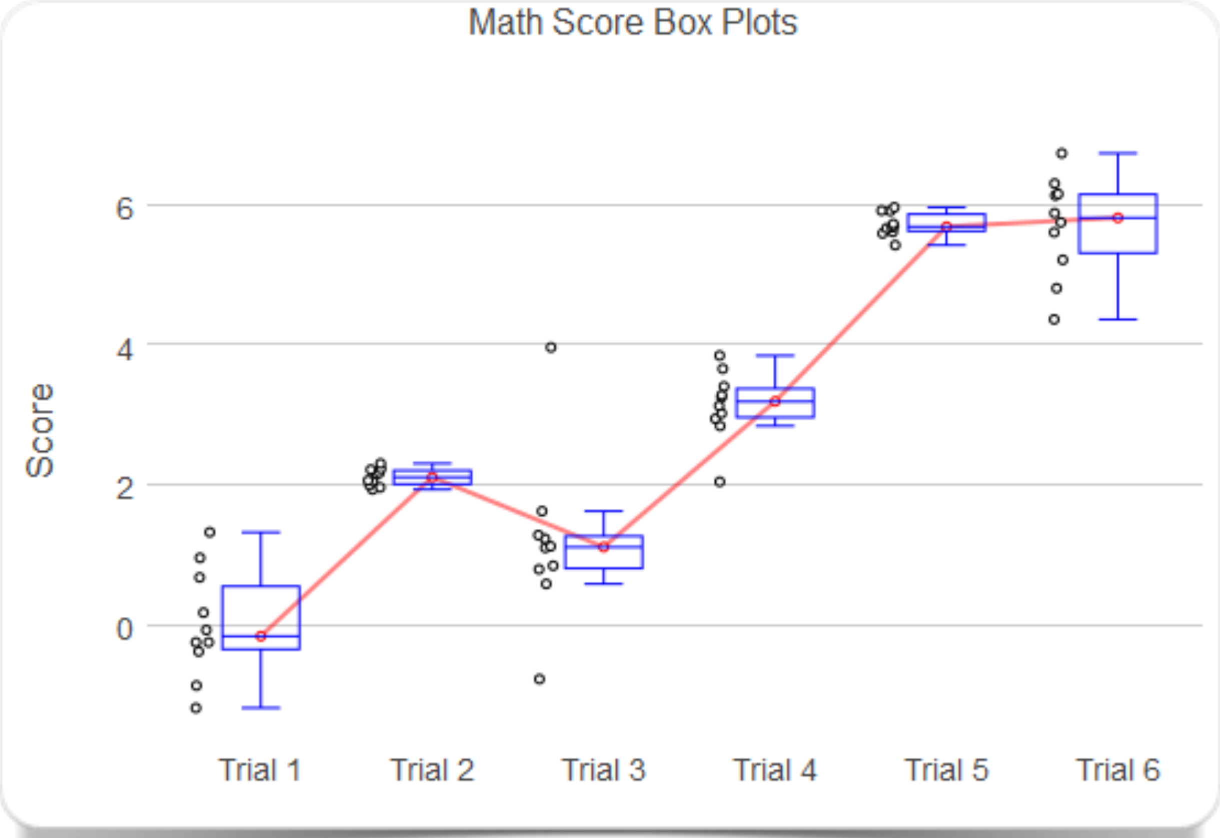


Fig. 7.1 Box and Whiskers Plot realizzato con plotly

Box and whiskers plot can be generated with plotly (as in figure 7.1). The program requires the user to specify only one column of numeric values for each box and whiskers plot, then automatically calculate their respective measures of central tendency and variation.

Comparison and distribution between two quantitative variables

The **bagplot** (Rousseeuw e al., 1999) is the two-dimensional representation of the box-plot. The bagplot shows bivariate measures of central tendency (mean, median, etc.) as well as a darker region and a lighter one around them. In the case of central tendency represented by a median, the dark region might represent values within the range of values closer to the median (eg., the range defined defined by the 25-th and 75-th percentile), while the "fence" which delimits the lighter region could represent the area delimited, for example, by the 15-th and 85-th percentile. Observation outside the “fence” will be considered outliers.

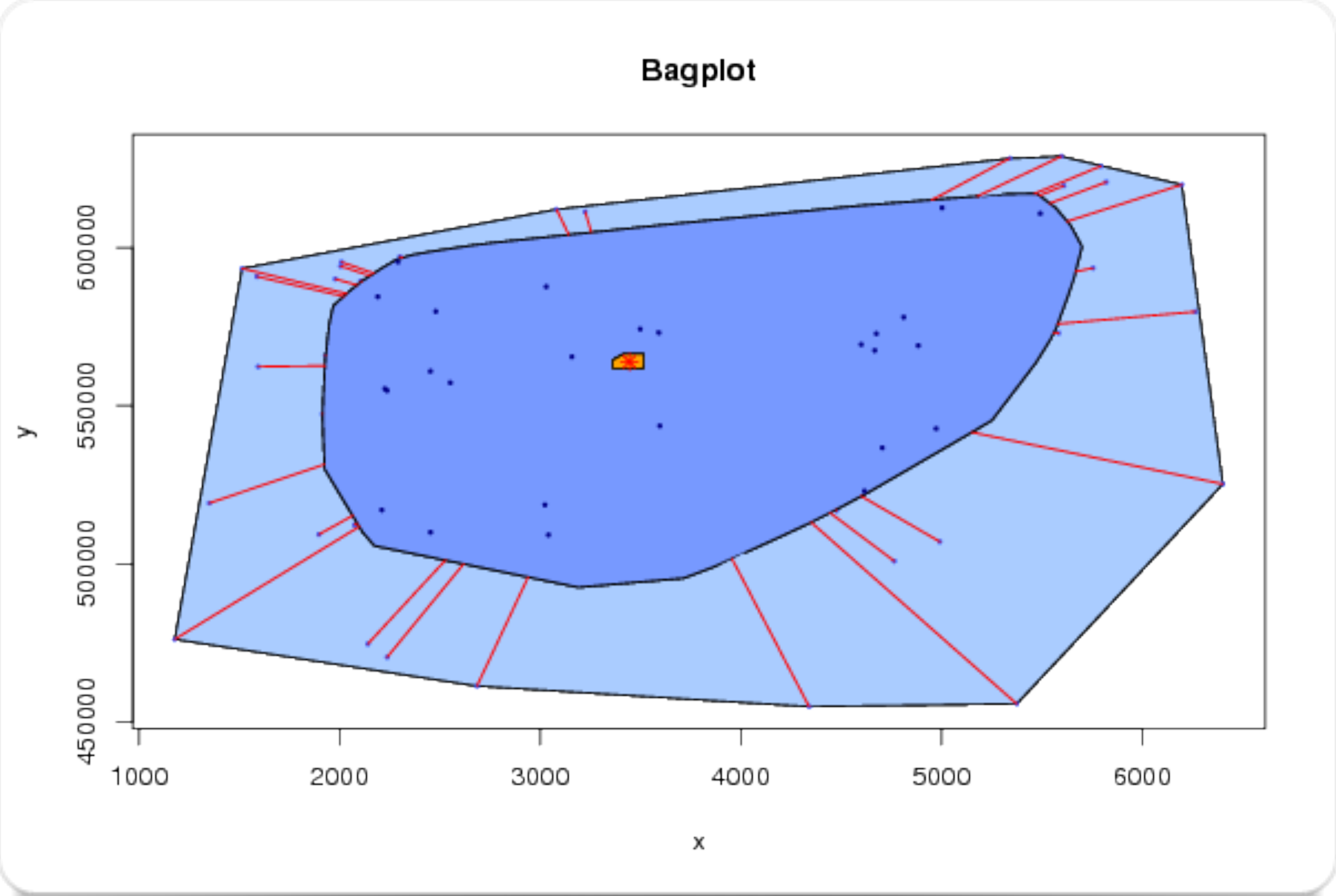


Fig. 7.2 Bagplot made with Wessa

With Wessa we can generate and manage every visual and content features of a bagplot (fig. 7.2; click *Compute*). For the special function implemented in Wessa we use the R: *rpart package*.

Chapter 8

Over-time comparison

Line graph

Sparkline

8

Over-time comparison of quantitative variables

The **line graph** (Harary & Norman, 1960) for several categories is really clear when there are few time steps (intervals), usually shown on the horizontal axis from the oldest one to newest. On such occasions it is possible to draw a series of lines passing through the different data-points of each category. This way we can easily compare the evolution of data for these categories over time.

Line graphs are among the available visualizations in Datawrapper. To generate them it is possible to use the typical 4 step interface. This type of visualization is the ideal tool for creating line graphs (Fig. 8.1).

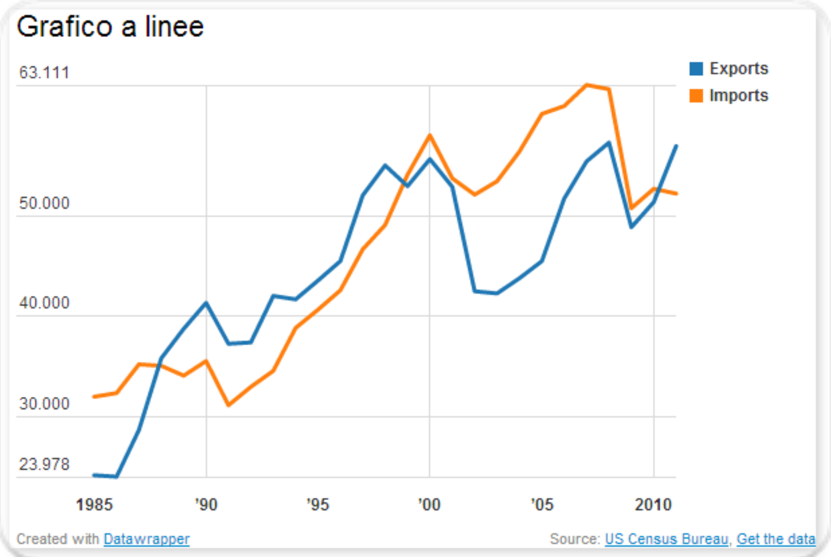


Grafico a linee

Created with Datawrapper

Source: US Census Bureau, Get the data

Fig. 8.1 Line graph made with Datawrapper

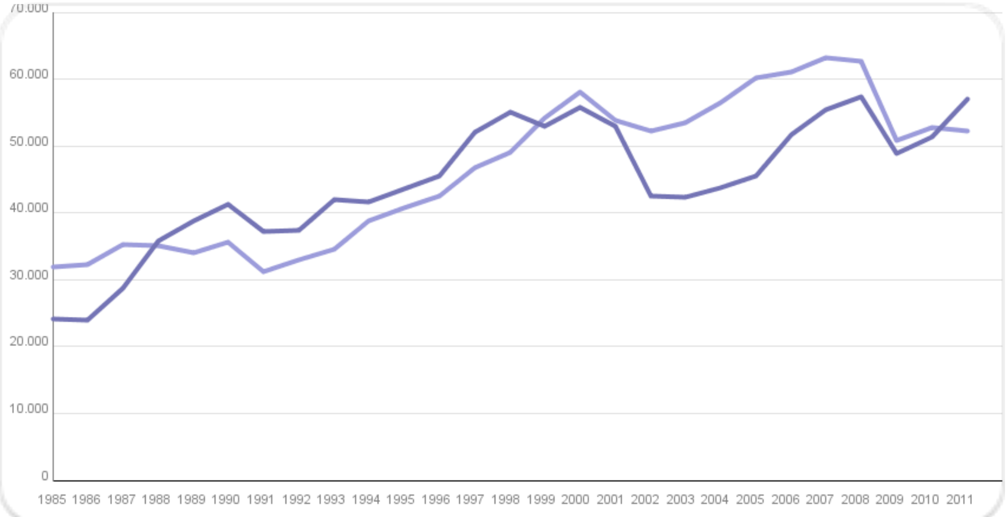


Fig. 8.2 Line graph made with IBM Watson Analytics

IBM Watson Analytics allows to generate line graphs using the Line Graph mode of visualization. In Figure 8.3 we can see a graphical representation of the evolution of the per capita income in the US (source DECD-CT).

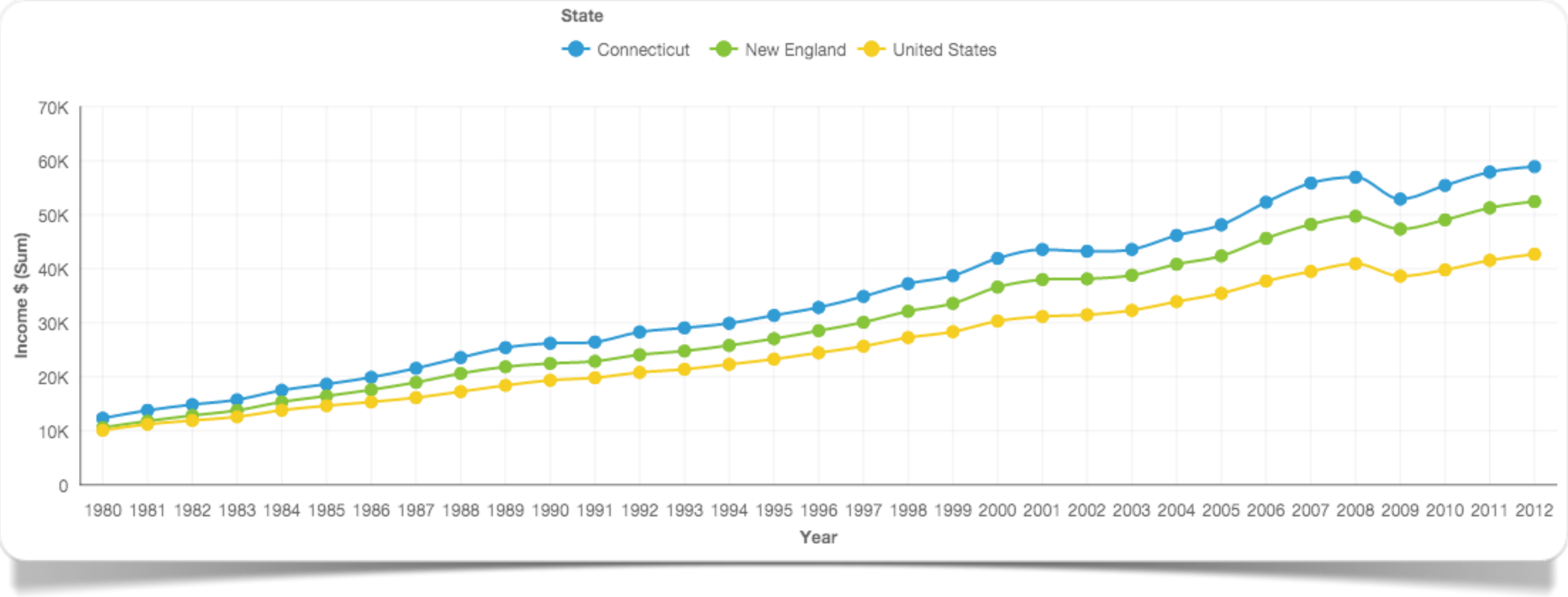


Fig. 8.3 Per capita income in the US from 1980 to 2012 (made with IBM Watson Analytics)

Over-time comparison: high-density

A **sparkline** (Tufte, 2004) is generally characterized by two main features: small size and high density data. The sparkline shows trends and variations associated with a particular measurement (temperature, financial trends) in the simplest possible way. In general, the tool used to depict a sparkline can be a line graph, a scatterplot or a bar chart.

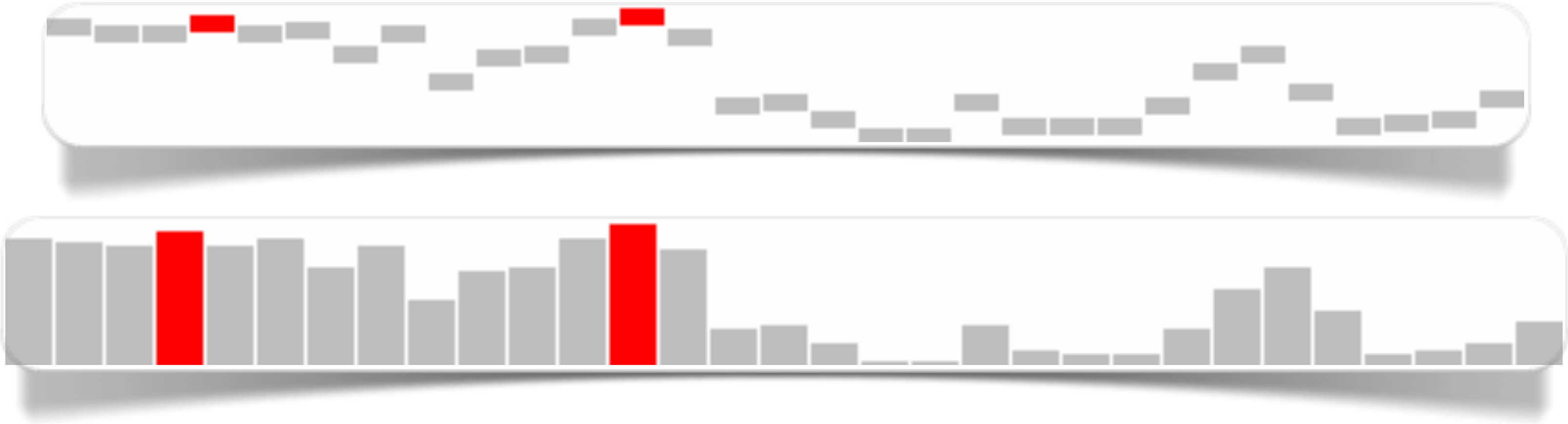


Fig. 8.4 Sparkline creati con Sparkline bitworking

Designed by Joe Gregorio, **Sparklines bitworking Project** is an online tool that allows to easily generate line or bar sparklines, managing all the graphical aspects of them (Fig. 8.4)

Chapter 9

Concentration

Lorenz curve

9

Concentration of a quantitative variable

The **Lorenz curve** (Lorenz, 1905) is the main tool of representation of concentration indexes. The curve is drawn in a plane on which the horizontal axis shows the relative cumulative frequencies, while the vertical axis shows the respective aggregate quantities. The area between the curve and the line of equality (the 45-degree line) is called the area of concentration and can be used as the basis for the definition of specific concentration ratios. The greater the concentration, the greater this area will be.

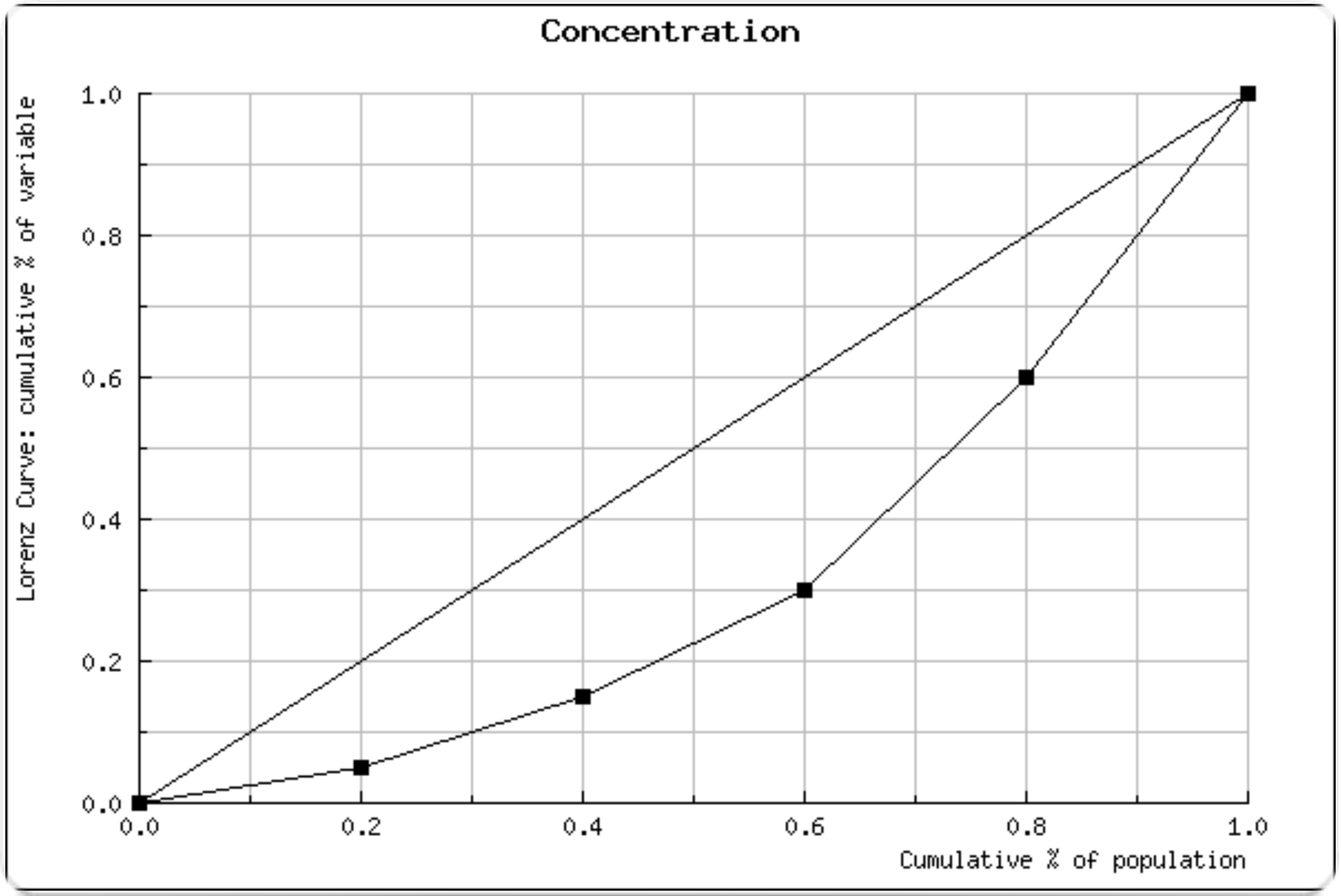
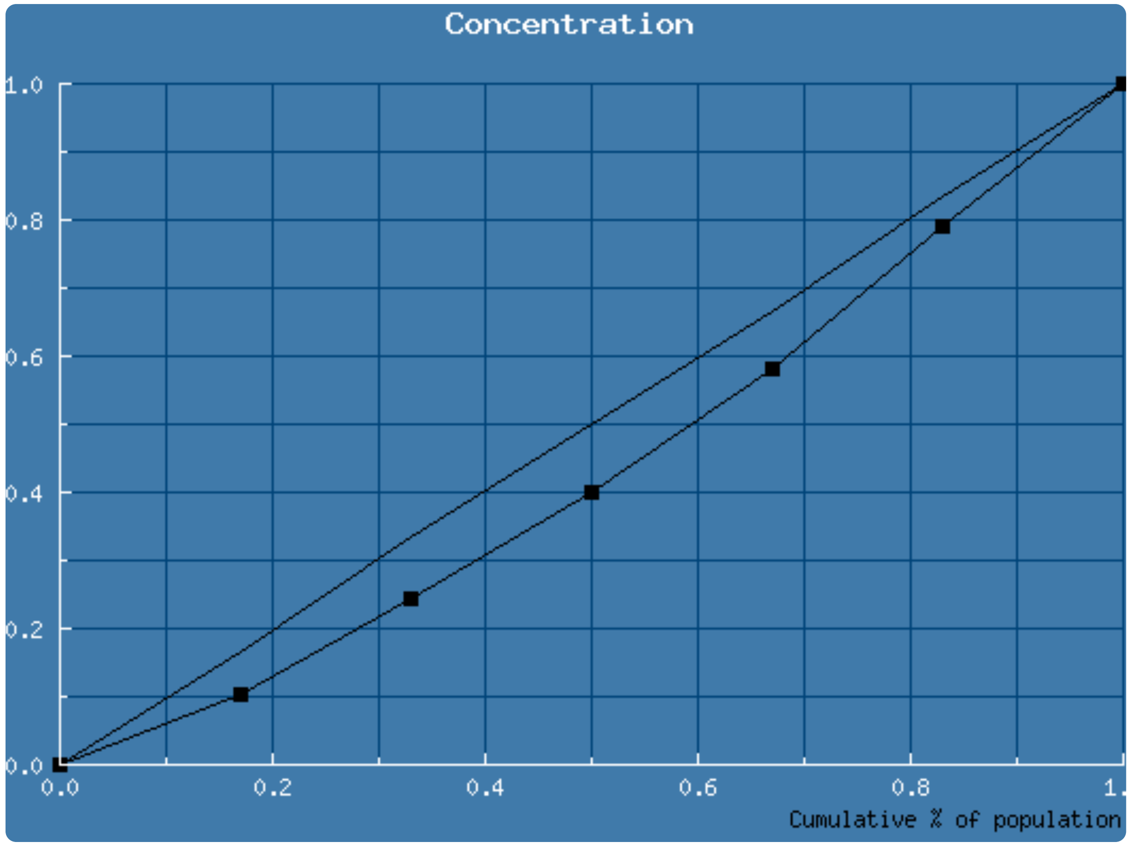


Fig. 9.1 Lorenz curve made with Wessa

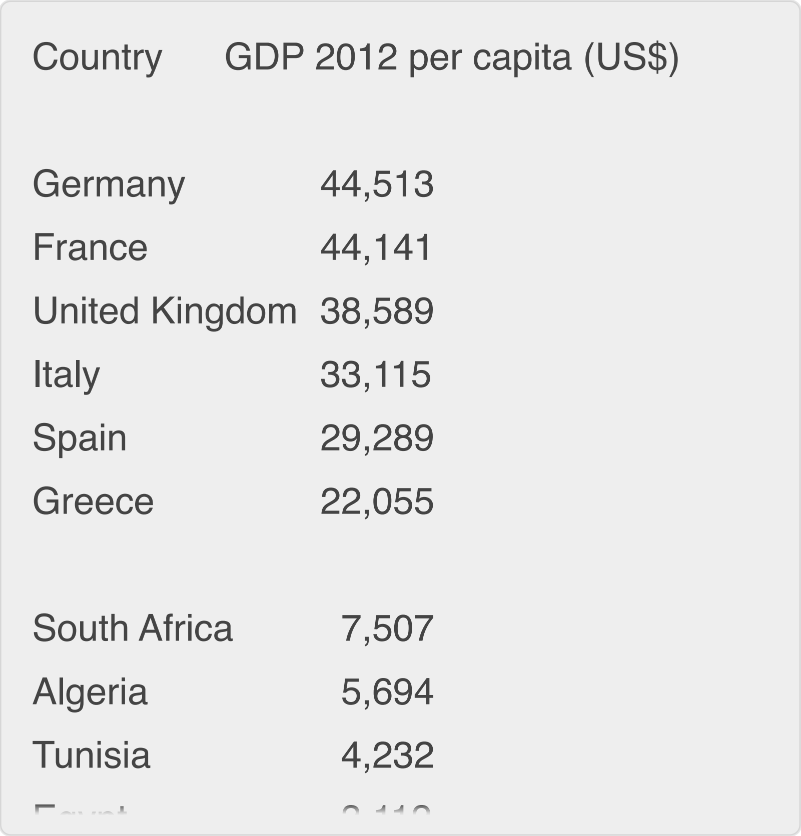
This Lorenz curve, or **graph of concentration** (Fig. 9.1), was made with Wessa with pre-entered data in the *Data* field.

In gallery 9.1 we can see the Lorenz curve of the data of scrollable figure 9.2 where we have the 2012 GDP per capita in US dollars for some European countries and some African countries (source: Internation Monetary Fund - Wikipedia).



Gallery 9.1 GDP per capita 2012: comparison between some European and African countries

GDP per capita in US dollars for Germany, France, UK, Italy, Spain, Greece



Country	GDP 2012 per capita (US$)
Germany	44,513
France	44,141
United Kingdom	38,589
Italy	33,115
Spain	29,289
Greece	22,055
South Africa	7,507
Algeria	5,694
Tunisia	4,232
Egypt	3,112

Fig. 9.2 GDP per capita data related to the graph of Gallery 9.1 - Comparison between the concentration curve of European and African countries

Chapter 10

Classification

ROC curve

Dendrogram

10

Classification of quantitative variable

The **ROC curve** is a tool widely used in biomedical statistics. It is a graphical plot that illustrates the performance of a binary classifier system, whose two axis generally represent the sensitivity and the value (1 - specificity) of a particular test. The data structure usually requires a numeric **variable** determined by a threshold value and a second, two-class variable (for example, “positive” or “negative”). The ROC curve allows to analyze the performance of the test across the range of variation of the values of the numeric variable. An area under curve (AUC) of 1 represents a perfect performance, while an area equal to 0.5 (ROC curve equivalent to a 45-degree line) represents a performance that has a probability equal to 0.5 of ranking a positive instance.

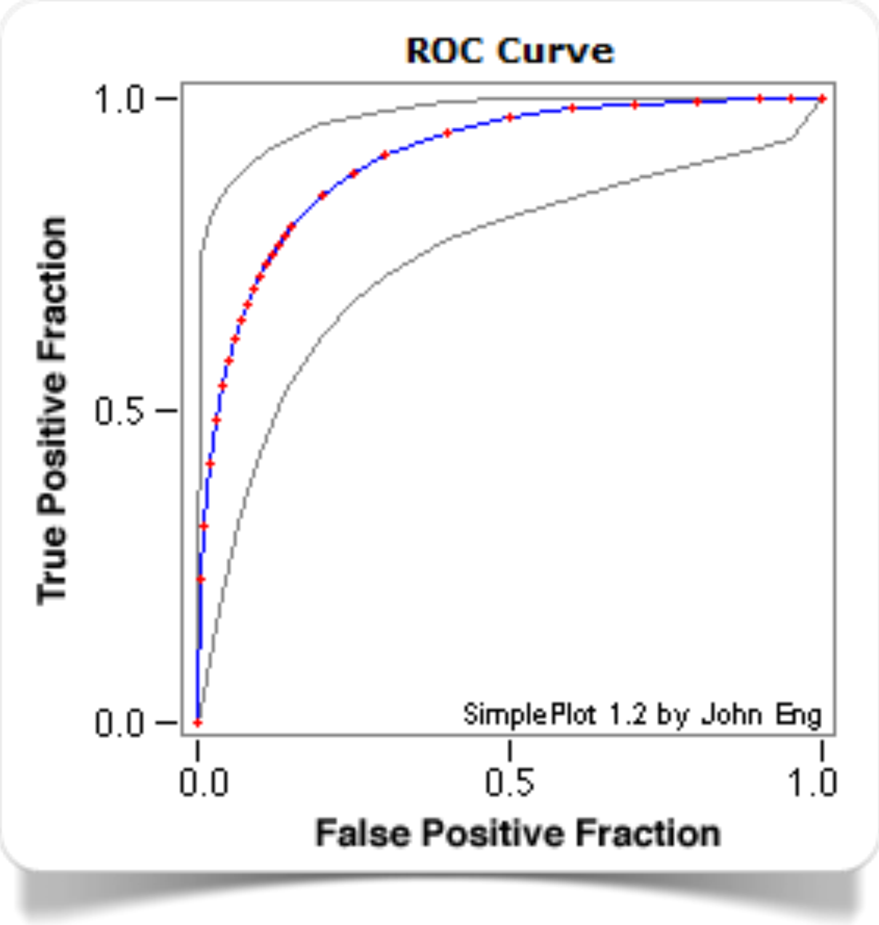


Fig. 10.1 ROC curve made with JROCFIT

JROCFIT is a web software made available by the Johns Hopkins University, Baltimore, Maryland, USA, to allow its students and other web users to generate ROC curves. The JROCFIT web page illustrates what kind of data format is accepted, as well as providing instructions on how to export the results. On the web we can get an analysis of a ROC curve through Wessa, using its open source interface R (see http://www.wessa.net/rwasp_logisticregression.wasp).

Classification: cluster analysis (hierarchical clustering)

The **dendrogram** is used to represent the results of an analysis of groups (cluster analysis) using the technique of hierarchical clustering. Each group is defined by at least a member (a group composed of a single cluster) to a maximum that is equal to the total number of clusters (a group containing all clusters).

The distance between the two ends of the graph defines the degree of homogeneity of the members belonging to the same group. The nearer to the base of the dendrogram (step 0) is the union between multiple clusters, the greater will be the degree of homogeneity of the clusters that form the resulting group.

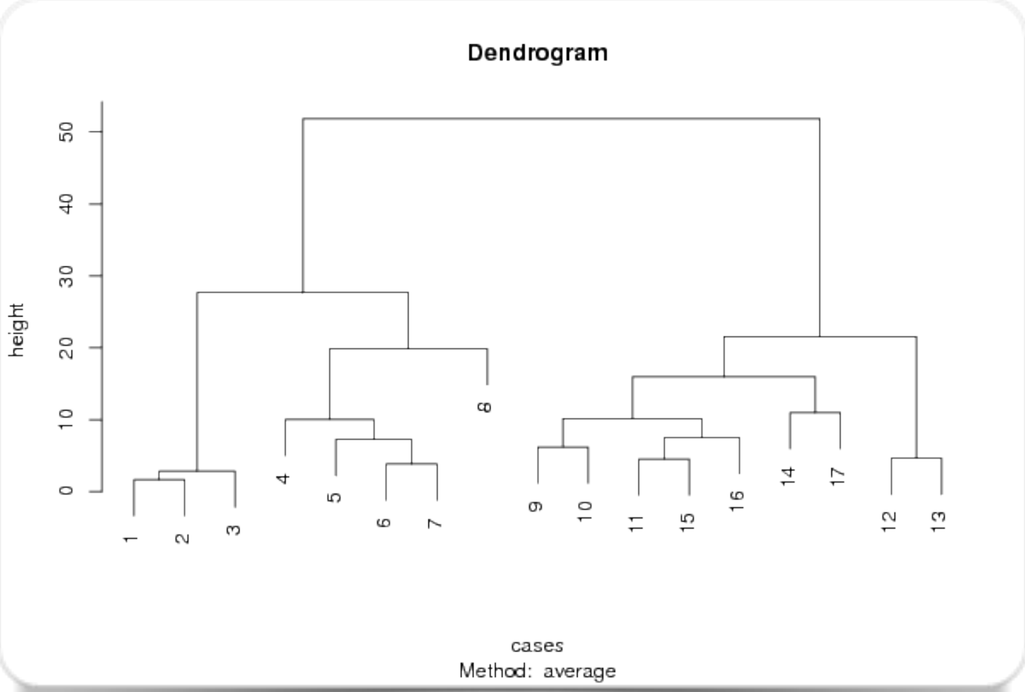


Fig. 10.2 Dendrogram made with Wessa

Thanks to Wessa we can generate dendrograms of all levels of complexity (Fig. 10.2; click on Compute). Since this is a graph traditionally made in the context of the analysis of groups, for the generation of this graph we resort to the use of the R: cluster package.

Chapter 11

Bibliography and web resources

11

Bibliography

Aitchison J. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman & Hall, reprinted in 2003 with additional material by Caldwell, NJ: The Blackburn Press.

Chambers J. M., Cleveland W., Kleiner B., Tukey P. (1983) *Graphical Methods for Data Analysis.* Wadsworth International Group.

d'Ocagne M. (1885) *Coordonnées parallèles et axiales : Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles*. Paris: Gauthier-Villars.

Few S. (2006) *Information Dashboard Design: The Effective Visual Communication of Data*. Sebastopol, CA: O'Reilly Media.

Harary F., Norman R. Z. (1960) Some properties of line digraphs. *Rendiconti del Circolo Matematico di Palermo,* 9 (2): 161–169.

Lorenz M. O. (1905) Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, Vol. 9, No. 70: 209–219.

Pearson K. (1895) Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 186: 343–326.

Pearson K. (1904) On the Theory of Contingency and Its Relation to Association and Normal Correlation, in *Research Memoirs Biometric Series I*, Drapers' Company.

Playfair W. (1786) *The Commercial and Political Atlas: Representing, by Means of Stained Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure and Debts of England during the Whole of the Eighteenth Century*. London: Debrett; Robinson; and Sewell.

Playfair W. (1801) *Statistical Breviary; Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe*. London: Wallis.

Rousseeuw P. J., Ruts I., Tukey, J. W. (1999) The Bagplot: A Bivariate Boxplot. T*he American Statistician.* 53 (4): 382–387.

Shneiderman B., Plaisant C. (2009) Treemaps for space-constrained visualization of hierarchies. Dec. 26th, 1998, last updated November, 2013; *Retrieved,* June 11, 2014.

Sneath P. H. A. (1957) The application of computers to taxonomy. *Journal of General Microbiology*, 17 (1): 201–226.

Tufte E. (2004) Sparkline theory and practice. Edward Tufte forum, May 27; *Retrieved,* June 11, 2014.

Tukey J. W. (1977) *Exploratory Data Analysis.* Boston: Pearson, Addison-Wesley

Venn J. (July 1880). On the Diagrammatic and Mechanical Representation of Propositions and Reasonings. *Philosophical Magazine and Journal of Science*. 5 10 (59).

Web resources

D3js.org Data-Driven Documents (http://d3js.org/)

DataHero (https://datahero.com/)

Datawrapper (https://datawrapper.de/)

Google Search (https://www.google.com/)

IBM Watson Analytics (http://www.ibm.com/analytics/watson-analytics/)

JROCFIT (http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html)

Many Eyes (http://www-01.ibm.com/software/analytics/many-eyes/)

plotly (https://plot.ly/)

R - The R Project for Statistical Computing (http://www.r-project.org/)

Sparklines bitworking (http://sparklines.bitworking.info/)

Raw (http://raw.densitydesign.org/)

Slemma (https://slemma.com/)

Wessa (http://www.wessa.net/)

WolframAlpha (http://www.wolframalpha.com/)

SAPIENZA
UNIVERSITÀ DI ROMA

Department of Statistical Sciences

1. L.Giuliano. *Il valore delle parole. L'analisi automatica dei testi in Web 2.0*.
2. D. Schiavon, L. Giuliano. *Wizard grafico. Una guida alla visualizzazione dei dati numerici*.
3. L. Giuliano. *The Value of Words. Automatic Text Analysis Tools in Web 2.0.*
4. D. Frongia. *Introduzione alla Social Network Analysis*.
5. D. Schiavon, L. Giuliano. *Chart Wizard. A guide to graphical representation of numeric data*.
6. D. Bennato. *Social Media Analytics e scienze sociali* (in preparazione).

Data Science Series

Duccio Schiavon, Luca Giuliano
Chart Wizard. A guide to graphical
representation of numeric data
Rome : Department of Statistical Sciences
[2015] 70 p.

9 788890 875748

Absolute difference

The absolute difference of two real numbers x, y is given by |x – y| and it describes the distance on the real line between the points corresponding to x and y.

Termini del glossario correlati

Trascina termini correlati qui

Indice

Trova termine

Composition

With the term composition we mean the set of quantitative data, which are quantitative descriptions of the parts of some whole, conveying exclusively relative information (Aitchison, 1986). In statistics, the use of this type of data is frequent when each data point is a "fraction" of a set of non-negative numbers whose sum is 1. In general, each data point suggests the proportion (or "percentage ") of statistical units that correspond to a specific category within the set of total of categories that are in the data set.

Termini del glossario correlati

Trascina termini correlati qui

Indice

Trova termine

Chapter 1 - What would you like to show?

Contingency table

A contingency table (Pearson, 1904) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables involved in the analysis.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Chapter 2 - Relationship between two qualitative variables

Curve fitting

With interpolating curve we mean the function resulting from the process of curve fitting. Curve fitting is the process of constructing a curve, or a mathematical function, that has the best fit to a series of data points

Termini del glossario correlati

Trascina termini correlati qui

Indice

Trova termine

Chapter 2 - Relationship between two quantitative variables

Data journalism

Data journalism (or data-driven journalism) is a journalism specialty based on the analysis of large data sets that are, in most cases freely, available on the web (open data), and whose processing requires the use of open source tools.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Distribution

In statistics, the concept of distribution is mainly related to that of probability distribution and is used to visually suggest what may be the most appropriate statistical model to fit the data of the shape of the distribution. The distribution has the graphical role to highlight what could be the specific statistical properties of the population belonging to the analyzed dataset.

Termini del glossario correlati

Trascina termini correlati qui

Indice

Trova termine

Flow diagram

A flow chart is a type of diagram that represents an algorithm as a flow process: each step is marked by a specific graphic element (circles, squares, boxes, etc.) and each of them is connected to the other with links and arrows that show the sequence and direction of flow. Generally, this schematic representation has the function to illustrate the solution to a given problem.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Infographics

The infographic (information graphic) is a graphic visual representation of information in which numbers and text are arranged in a specific way and presented in an organized visual form. The techniques used to achieve this kinds of representation require specific computer and graphical skills, as weel as good presentation skills.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Chapter 1 - What would you like to show?

Open data

With the term “open data” we mean data that are freely available to everyone to use and republish as they wish. In order to meet the essential feature of "openness", data may not be restricted by patents or other mechanisms of control which may limit their reproduction. The only restrictions are those that refer to the obligation to cite any source or those that refer to any eventual modification made.

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Chapter 1 - Web based softwares

Relative difference

Realtive differences are usually used to compare quantities while taking into account the "sizes" of the things being compared. The comparison is expressed as a ratio and is a unitless number. By multiplying these ratios by 100 they can be expressed as percentages, so, in this case, the term percentage difference can also be used.

Termini del glossario correlati

Trascina termini correlati qui

Indice

Trova termine

Variability

In statistics, variability (also called statistical dispersion, scatter, or spread) measures the degree of dispersion of a distribution. In particular, an index of variability (variance, standard deviation, interquartile range, etc.) is used to describe how its values are far apart from the respective measures of central tendency (mean, median, mid-range, etc.).

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Chapter 2 - Relationship between two quantitative variables

Variable

In statistics, a variable is a characteristic of a statistical unit being observed that may assume more than one of a set of values to which a numerical measure or a category from a classification can be assigned (income, age, weight, etc. for numeric variables; "profession", "eye colour "," disease ", etc. for categorical variables).

Numeric variables can be classified into two categories:

• *continuous variables*, which may take on an infinite number of possible values between two distinct values (blood pressure, temperature, etc.).

• *discrete variables*, which may take on only a countable number of distinct values (number of children, number of legs of an animal, etc.).

Categorical variables are classified into two categories:

• *nominal variables*, that contains values indicating membership in one of several possible categories whose characteristics or qualities cannot be ranked (sex, race, transportation, etc.).

• *ordinal variables*, that contains values indicating membership in one of several possible categories which can be ordered by an order relation or a hierarchy (education, satisfaction, etc.).

Termini del glossario correlati

Trascina termini correlati qui

Indice Trova termine

Chapter 1 - What would you like to show?