Data visualization techniques:
An analysis of the project “from Data to Viz”

Introduction

This document illustrates the results of the study of the project “*from Data to Viz*”[1]. “*From Data to Viz*” provides a comprehensive glossary of data visualization techniques and suggests when to use a specific visualization technique based on the data type.

The rest of the document reports a summary of the main data formats supported by “*from Data to Viz*” along with suggested data visualization techniques, and examples for each main data format.

The choice of the appropriate visualization

“*From Data to Viz*” provides a decision tree based on input data format. This tree leads to a set of formats representing the most common dataset types.

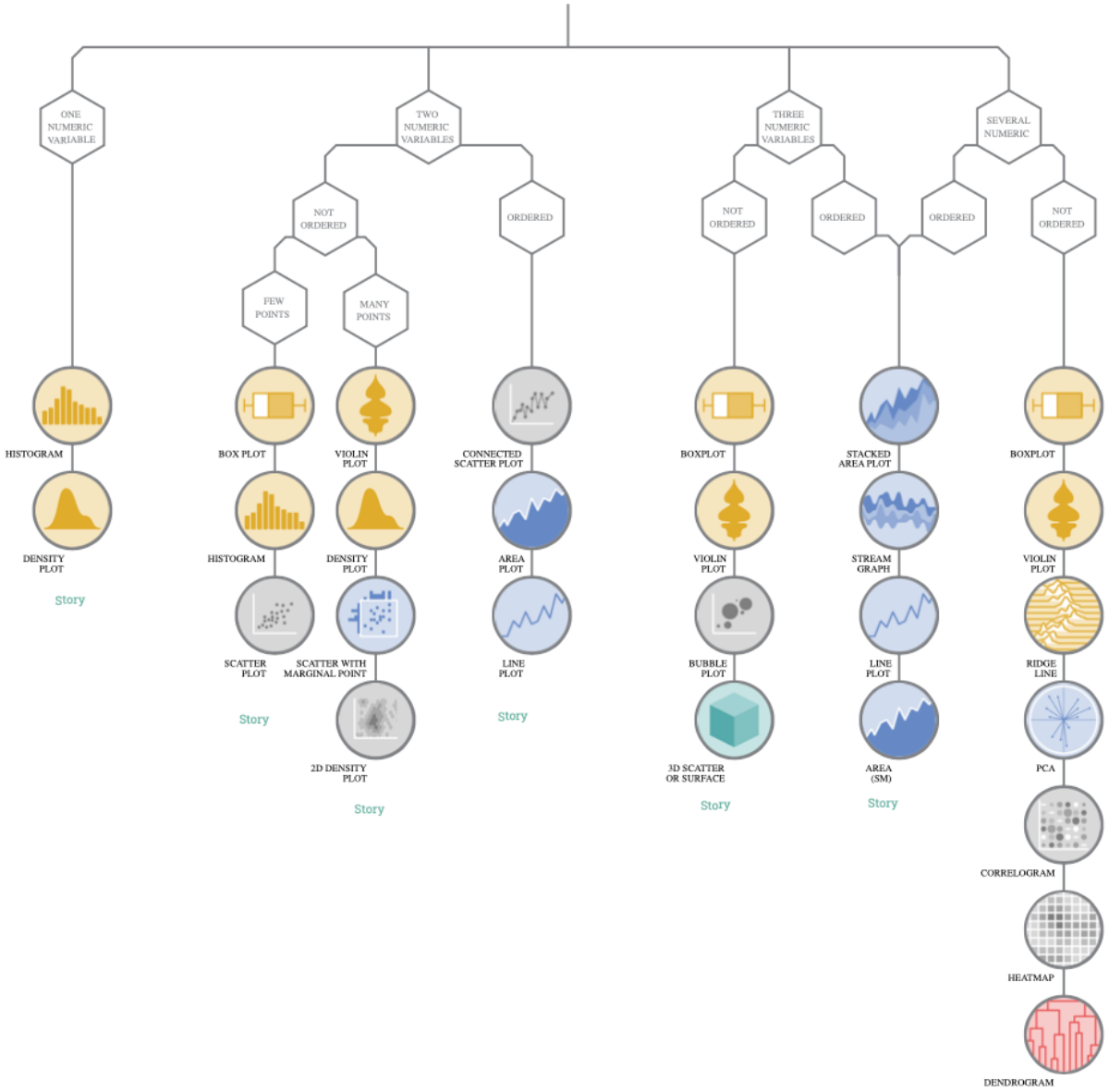
“*From Data to Viz*” supports six different main input data formats:

- Numeric
- Categoric
- Numeric & Categoric
- Maps
- Network
- Timeseries

The remainder of this section reports, for each input data format, the corresponding decision tree along with a brief description.

[1] https://www.data-to-viz.com

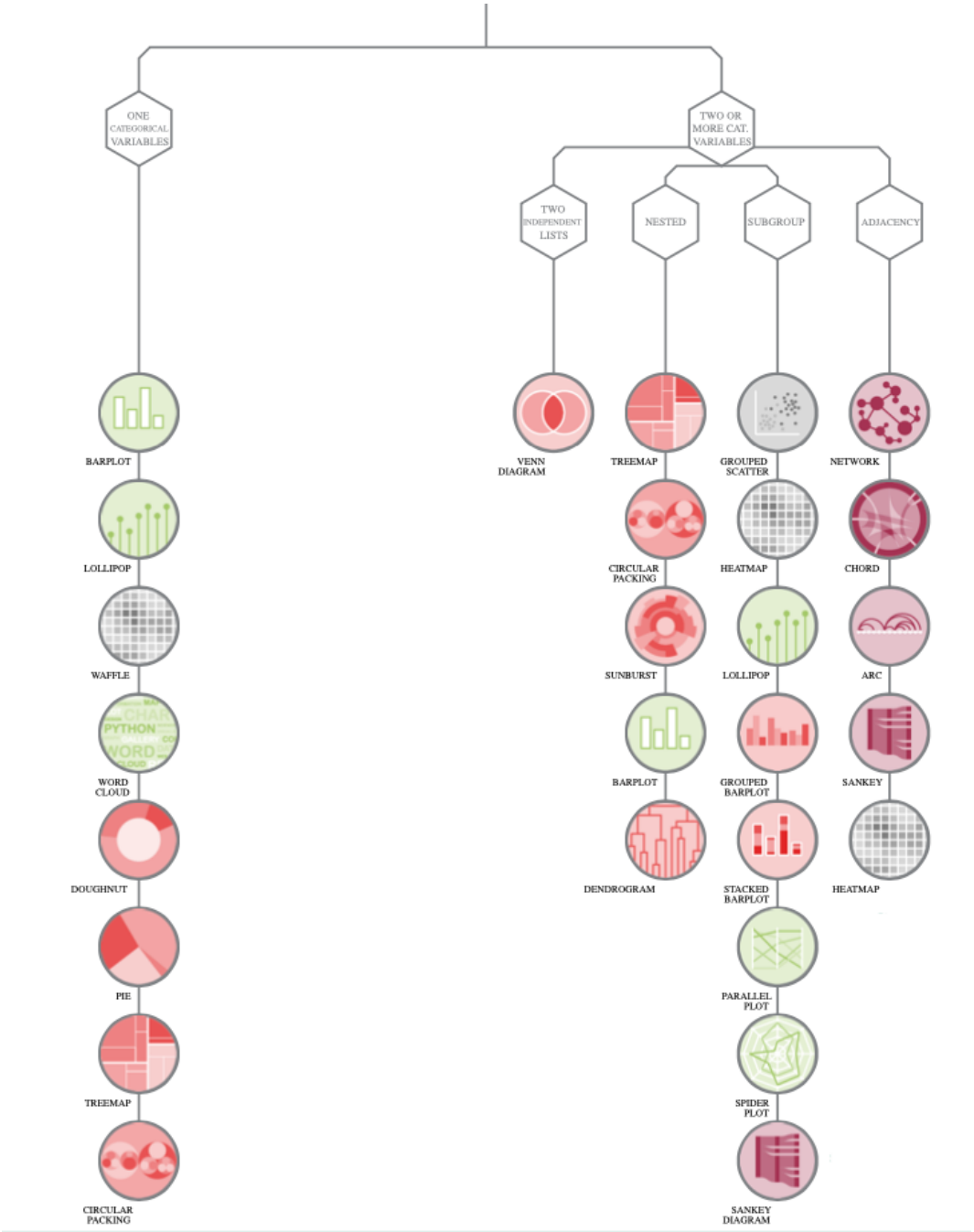
Numeric



As showed by the figure, "*from Data to Viz*" suggests several approaches to visualize numeric data, including *distribution graphs* (in yellow), *correlation graphs* (in grey), *part of a whole graphs* (in red) and *evolution graphs* (in blue).

The choice of the appropriate graph is driven by the characteristics of the data, such as (i) the number of variables, (ii) the number of data points and (iii) the relevance of the order. For example, *histogram* and *density plots* are suggested to visualize (without a specific order) less than three variables, while *box plots* and *violin plots* are suggested to visualize more than one unordered variable. Other data visualization techniques such as *scatter plots*, *2d density plots*, *correlogram* and *heat maps* are also suggested to visualize correlation among unordered numeric variables. In case of ordered variables, "*from Data to Viz*" mostly suggests the use of *evolution* graphs, such as *line plots*, *area plots*, *stream graphs*, etc.

Categoric



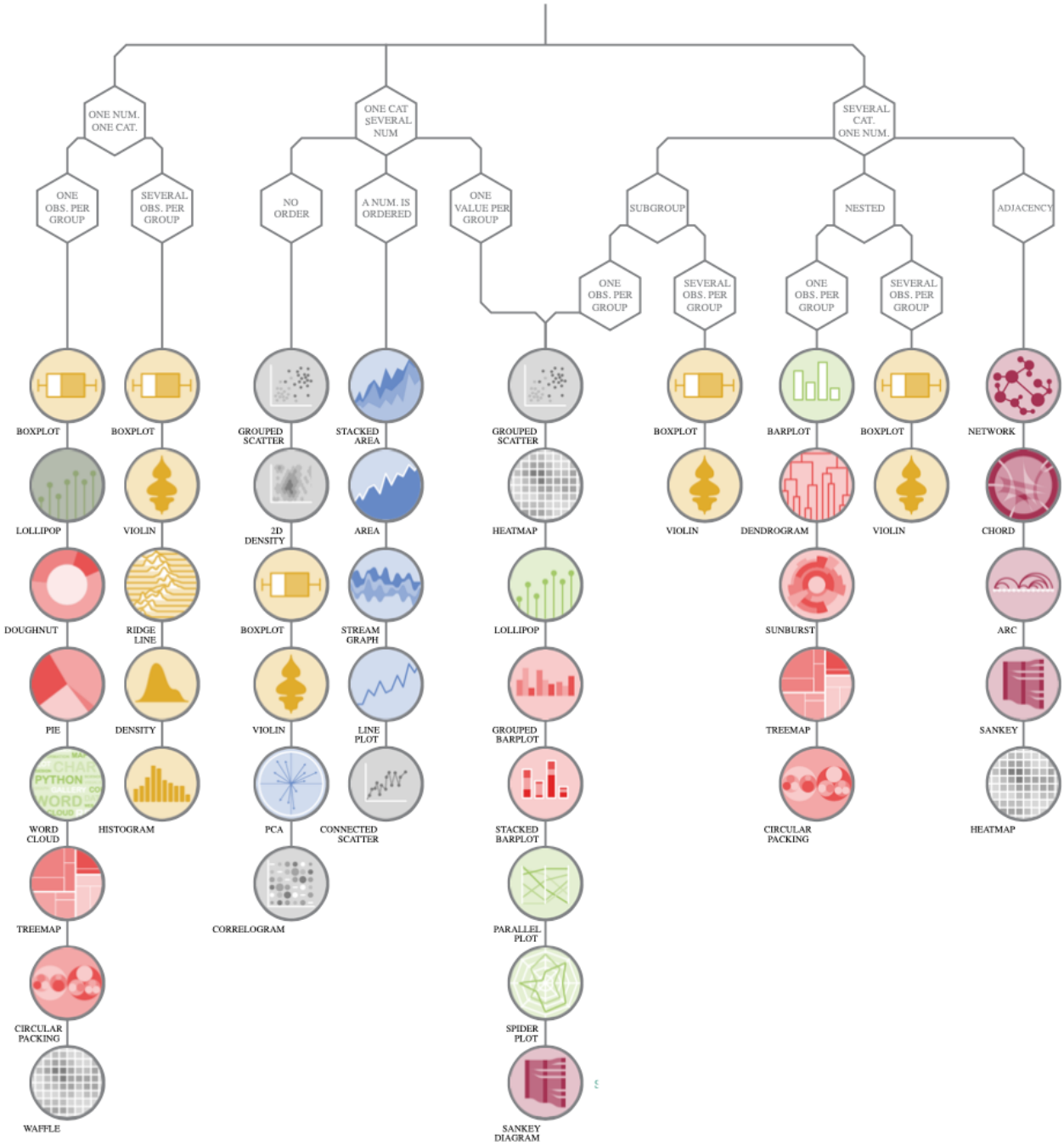
Similarly to numeric variables, the choice of the appropriate visualization for categoric data is mainly driven by the number of variables. In case of an individual categoric variable, "*from Data to Viz*" mostly suggests the use of *ranking* graphs, such as *bar plots*, *word cloud* and *lollipop*, or *part of a whole graphs*, such as *doughnuts*, *pies*, *tree maps*, *circular packing* or *waffles*. On the other hand, when multiple categoric variables are involved, the decision tree provides four different cases:

- Two independent lists
- Nested (each entity is separately identifiable but also part of larger data organizations)
- Subgroup (every combination among variables is possible)
- Adjacency

In case of two independent lists, "*from Data to Viz*" suggests the use of *Venn diagrams* to show the size of the overlap between them. *Part of a whole graphs*, such as *treemaps*, *sunbursts* and *dendrograms*, are instead suggested for nested categoric variables. For subgroups "from Data to Viz" suggests a wide range graphs, including *correlation graphs* (e.g., *heatmaps*), *ranking graphs* (e.g., *spider plots*), *part of a whole graphs* (e.g., *grouped barplots*) and *flow graphs* (e.g., *Sankey diagrams*). In order to visualize adjacencies among categoric variables, *flow graphs*, such as *networks*, *chords*, *Sankey diagrams* and *archs,* are suggested.

Numeric & Categoric

The following figure shows the decision tree for data involving both numeric and categoric variables:



The decision tree involves three different cases:

- One numeric variable and one categoric variable
- One numeric variable and several categoric variables
- Several numeric variables and one categoric variable

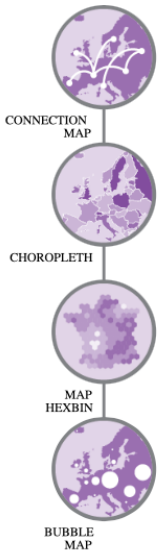
In the first case (one num. and one cat.), if the data involves an individual numeric observation per categoric value (i.e., group), "*from Data to Viz*" suggests the use of *part of a whole graphs* (e.g., *pie*, *doughnuts*, *treemaps*) or *ranking graphs* (e.g, *word clouds*, *lollipops*). When there are multiple observations per group, instead, it suggests *distribution graphs* (e.g., *boxplots, violin plots, ridge lines*).

In the second case (one cat. and sev. num.), *correlation graphs* and some *distribution graphs* (*box plots* and *violin plots*) are suggested for the unordered visualization of numeric variables. *Evolution graphs, such as area plots* and *line plots,* are instead suggested when the order must be visualized. If there is an individual observation per categorical value (i.e., group), "*from Data to Viz*" suggests either *part of a whole graphs* (*grouped* and *stacked barplots*)*, correlation graphs* (*heatmap* and *grouped scatterplots*), *ranking graphs* (*lollipop*, *parallel* and *spyder plots*) or *Sankey diagrams*.

In the third case (one num. and sev. cat.), the decision tree suggests *box plots* and *violin plots* to visualize both subgroups and nested subgroups with several observation per subgroup, while it suggests *part of a whole graphs* for nested subgroups with an individual observation per subgroup.

Adjacencies are instead provided through *flow graphs* or heatmaps*.*

Maps



In order to visualize data related to maps, "*From Data to Viz*" suggests four different graphs: *connection maps, choropleths, maps hexbin* and *bubble maps*.

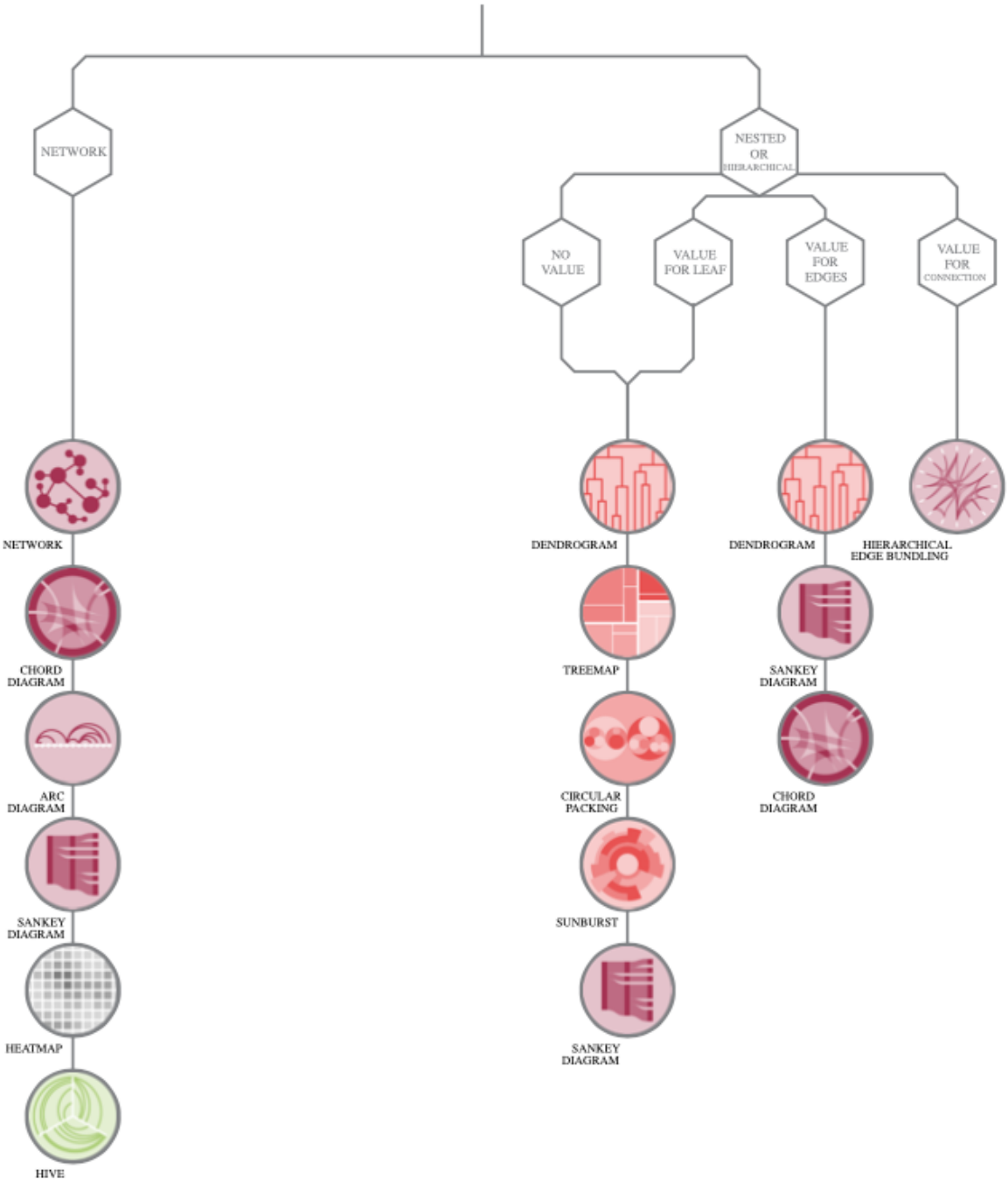
Connection maps allow to show the connection between several positions on a map. The link between 2 places can be drawn with a straight line or, more commonly, the shortest route between them.

Choropleth maps display divided geographical areas or regions that are colored in relation to a numeric variable. It allows to study how a variable evolutes along a territory.

Hexbin map is a kind of *choropleth map*, which splits a geographic area in a multitude of hexagons. A numeric value is attributed to each hexagone to remove the bias introduced by the different region size.

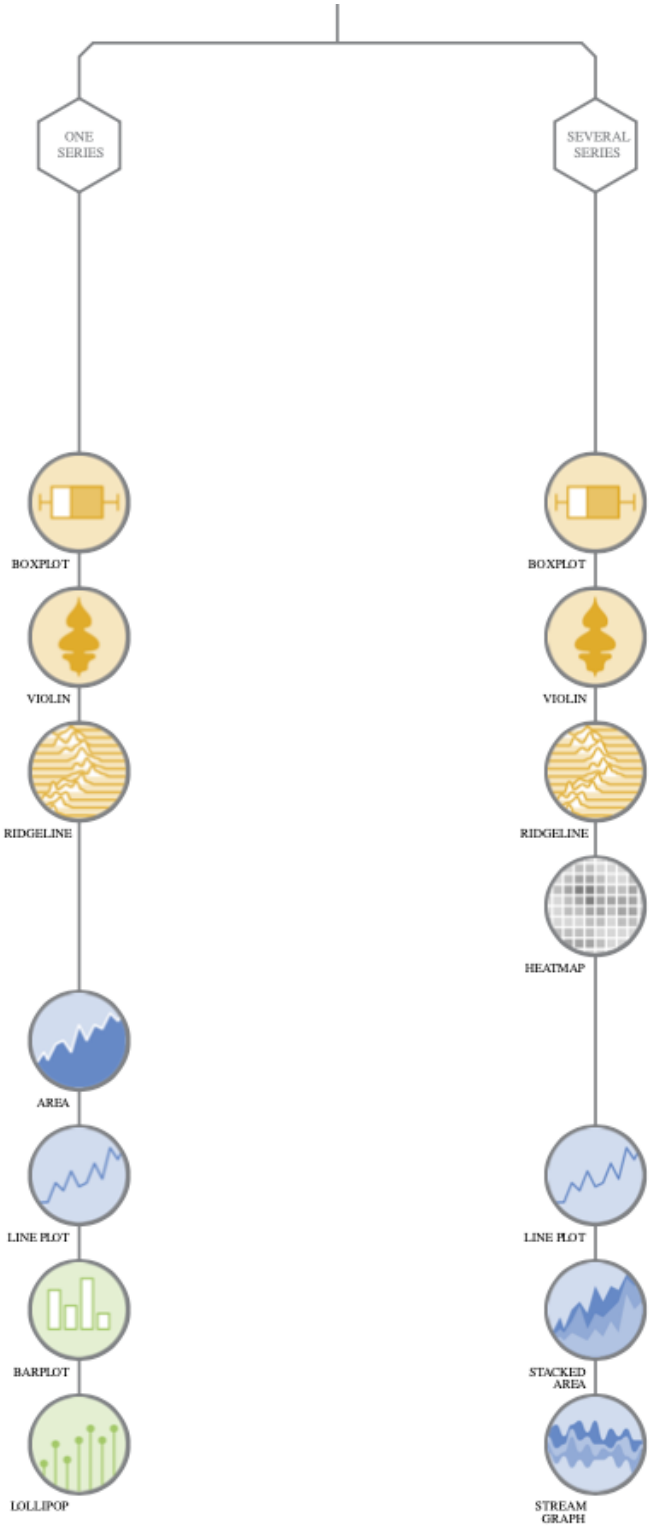
A *bubble map* uses circles of different size to represent a numeric value on a territory. It is possible to display a bubble per geographic coordinate, or a bubble per region.

Network



"*From Data to Viz*" mostly suggests *flow graphs* to visualize networks, including *networks graphs, chord diagrams, arc diagrams and Sankey diagrams.* In case of nested/hierarchical networks, *part of a whole graphs,* such as *treemaps*, *circular packings*, *sunbursts* and *dendrograms*, are also suggested.

Time series



Box plots, violin plots, ridgelines and *line plots* are suggested for both individual and multiple time series. In case of individual time series *bar plots, lollipops* and *area plots* are also suggested. *Heatmaps, stacked area plots* and *stream graphs* are instead suggested for multiple time series.

Examples

This section reports examples of data visualization,s for each main input data format.

Numeric

Apartment price vs ground living area - Scatter plot

This example considers the price of 1460 apartments (SalePrice) and their ground living area (GrLivArea). The dataset is composed by two numeric variables, which looks like the table below.

GrLivArea	SalePrice
1710	208500
1262	181500
1786	223500
1717	140000
2198	250000
1362	143000

A *scatterplot* helps to explore the correlation between sale price and living area.

The chart shows a quite obvious relation between price and ground living area.

Categoric

Brassens lyrics – Word cloud

This example considers the lyrics of a famous French singer (Georges Brassens). The data set is composed by a list of words which appear in Brassens' lyrics. The data set is structured as the table showed below.

Brassens
d'avoir
d'hélène
l'pornographe
georges
sète

The *wordcloud* graph provides a common way to visualize the frequencies of words.

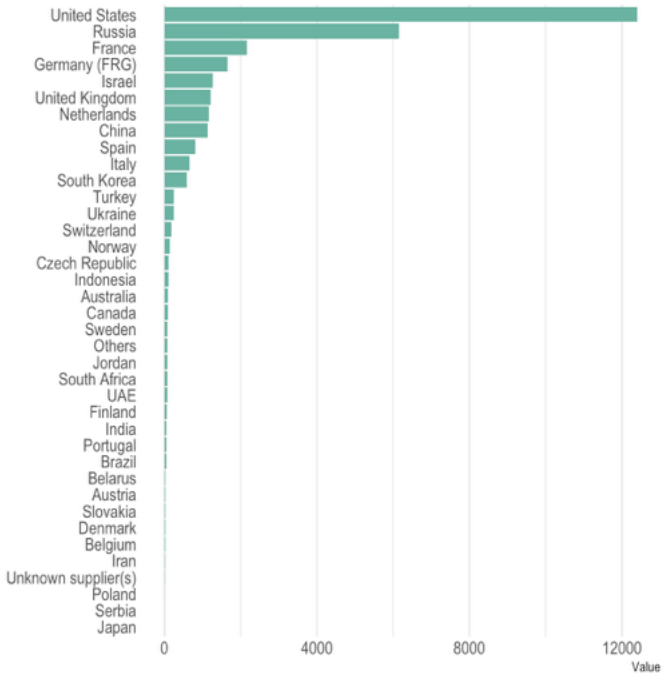
Numeric & Categoric

Who sells more weapons – Bar plots

This example considers the quantity of weapons exported by the top 50 largest exporters in 2017. The dataset involves one categoric and one numeric variable, and it structured as the table showed below.

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131

The following *bar plot* provides a sense of the differences in weapon exportations among the top 50 countries.

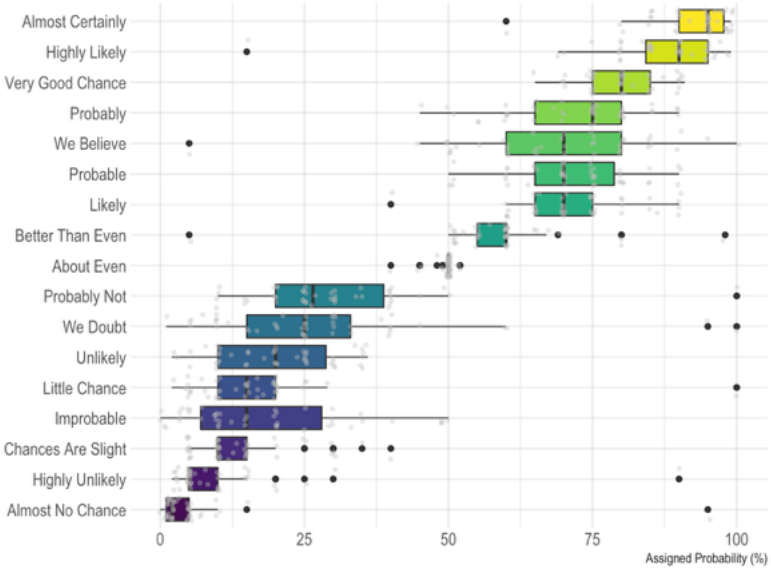


Perception of phrase probability – Box plots

This example considers how people perceive probability vocabulary. The dataset is composed by a set of probabilities assigned to phrases. Each item of the data set represents the probability (in terms of percentage) of a particular phrase as perceived by an individual. The data set looks like the table showed below.

text	value
Improbable	33
Almost Certainly	98
Likely	60
Almost Certainly	98
Unlikely	10
Probably Not	25
About Even	50
Probably	75

Boxplots allow to summarize the distributions of probabilities for each phrase.



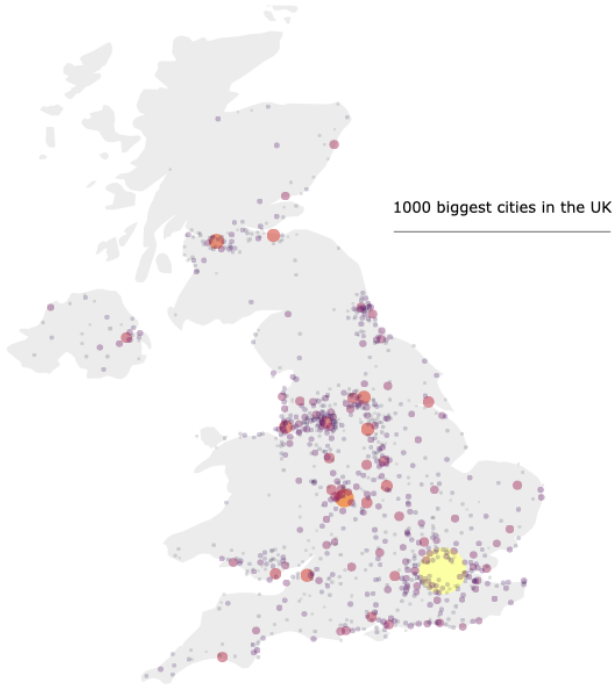
Maps

The Biggest UK Cities - Bubble map

This example considers the population of 925 cities in the UK. The data set looks like the table below.

lat	long	pop	name
51.65	-3.14	10146	Abercarn-Newbridge
51.72	-3.46	33048	Aberdare
57.15	-2.10	184031	Aberdeen
51.83	-3.02	14251	Abergavenny
53.28	-3.58	17819	Abergele

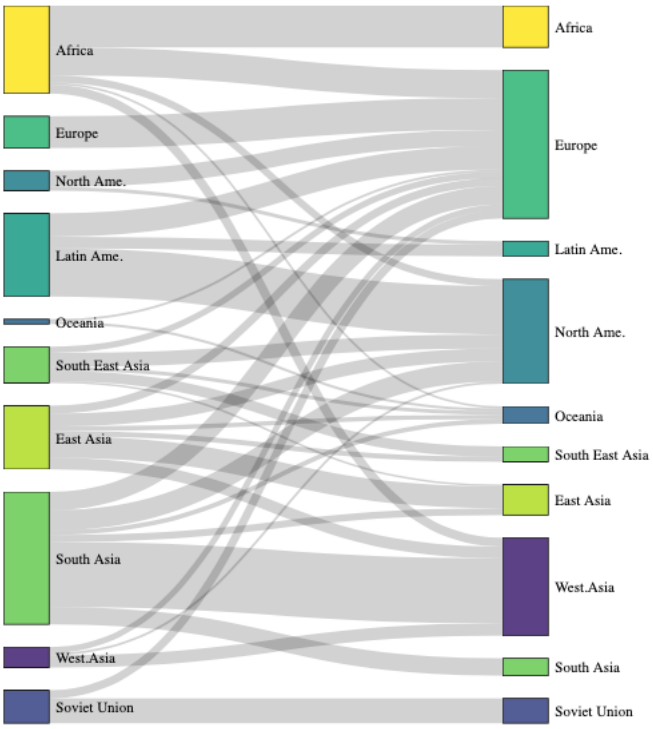
Bubble maps are probably the most common way to visualize this kind of data sets. In *bubble maps,* one circle is drawn per provided geographic position, and the size of the bubble is proportional to its corresponding numeric value.



Network

Migration flows – Sankey diagram

Sankey diagram is a good way to represent the migration flows, since it is appropriate to visualize directed and weighted networks.



Time series

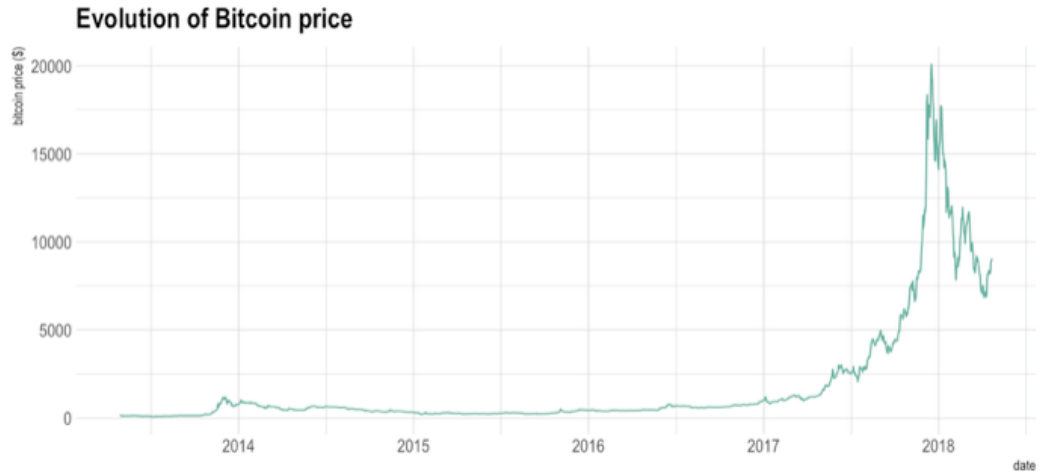
Bitcoin price – Line plot

This example considers the evolution of the bitcoin price between April 2013 and April 2018. The dataset is composed by two columns: the first column, date, represents an ordered numeric variable. The second, value gives the bitcoin price.

date	value
2013-04-28	135.98
2013-04-29	147.49
2013-04-30	146.93

date	value
2013-05-01	139.89
2013-05-02	125.60

Line plots represent the most common way to represent this kind of dataset. As showed by the plot below, it allows to give a good overview of the bitcoin price on the period.



Conclusion

"*From Data to Viz*" provides a well-defined approach to select the appropriate data visualization according to the data input format. The choice of the appropriate graph is driven by a set of decision trees, which consider types, numbers and characteristics of variables involved in the analysis. This document provided an overview of the data visualization selection process for each main data input format (i.e., numeric, categoric, numeric & categoric, maps, network, and time series). Additionally, the document reported, for each format, examples of effective data visualization techniques.