Data Visualization Practitioner's Guide

EDF certification handout

Auteur: Michel Dekker

Van Haren PUBLISHING

Effective DATA Foundation

Data Visualization
Practitioner's Guide

Effective DATA Foundation

Colophon

Title:	Data Visualization Practitioner's Guide
Authors:	Michel Dekker
Publisher:	Van Haren Publishing, ‘s-Hertogenbosch
ISBN Hard Copy:	978 94 018 1170 5
Edition:	First edition, first print, May 2024
Design:	Van Haren Publishing, ‘s-Hertogenbosch
Copyright:	© Van Haren Publishing 2024
	For further information about Van Haren Publishing please e-mail us at: info@vanharen.net or visit our website: www.vanharen.net

Other publications by Van Haren Publishing

Van Haren Publishing (VHP) specializes in titles on Best Practices, methods and standards within four domains:
- IT and IT Management
- Architecture (Enterprise and IT)
- Business Management and
- Project Management

Van Haren Publishing also publishes on behalf of leading organizations and companies: ASL BiSL Foundation, BRMI, CA, Centre Henri Tudor, Gaming Works, IACCM, IAOP, IFDC, Innovation Value Institute, IPMA-NL, ITSqc, NAF, KNVI, PMI-NL, PON, The Open Group, The SOX Institute.

Topics are (per domain):

IT and IT Management
ABC of ICT
ASL®
CATS CM®
CMMI®
COBIT®
e-CF
ISO/IEC 20000
ISO/IEC 27001/27002
ISPL
IT4IT®
IT-CMFtm
IT Service CMM
ITIL®
MOF
MSF
SABSA
SAF
SIAMtm
TRIM
VeriSMtm

Enterprise Architecture
ArchiMate®
GEA®
Novius Architecture Method
TOGAF®

Business Management
BABOK ® Guide
BiSL® and BiSL®-Next
BRMBOK™
BTF
EFQM
eSCM
IACCM
ISA-95
ISO 9000/9001
OPBOK
Six Sigma
SOX
SqEME®

Project Management
A4-Project Management
DSDM/Atern
ICB/NCB
ISO 21500
MINCE®
M_o_R®
MSP®
P3O®
PMBOK ® Guide
Praxis®
PRINCE2®

For the latest information on VHP publications, visit our website: www.vanharen.net.

1 Course overview

To harness the potential of data, it is crucial for data professionals to present it in a visual format that is both interactive and comprehensible to management and decision makers. The human brain has a strong inclination towards visual information, making data visualization an incredibly powerful tool for the efficient analysis, interpretation, and communication of data.

This certification demonstrates that qualified professionals have mastered the required skills to visualize data effectively, ensuring that important results will not escape their notice.

The training consists of six modules, each with its own weight towards the certification exam:

	Weight	Topic
Introduction	-	Introduction to the data visualization field and basic quantitative thinking.
Human perception	20%	About the principal components of visual perception, to optimize our visualizations for human consumption.
Visualizing data	30%	About applying the knowledge of our visual perception to data visualization and the introduction of the CHRTTS model to select chart types.
Data viz design	25%	About the detailed choices we need to consider when designing our visuals. We also look at the management dashboard and the important role of color.
Storytelling	15%	How to make use of our capacity to tell and consume stories to process data insights.
Workflow	10%	How to implement the effective data visualization practices in our daily work.

2 Introduction

“It’s less a technology problem, more a people problem.” (Aron Pilhofer – Executive Director of Digital for the Guardian).

Most visualization tools are very capable of creating visualizations, but this doesn’t make you a data visualization expert.

2.1 Why do we visualize data?

Figure 1 Why do we visualiza data?

Our visual system, composed of our eyes and brains, excels at recognizing objects, comparing patterns and shapes, and swiftly detecting trends and anomalies. This ability plays a crucial role in the visualization of data, enabling us to comprehend information more rapidly and effectively. By gaining a deeper understanding of the data, we can extract valuable insights that inform data-driven decision-making. Ultimately, the central element in data visualization is people, as they are instrumental in this process.

Why do we visualize data?

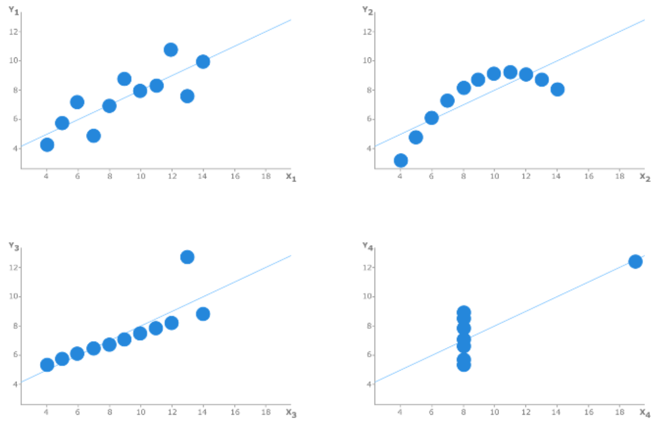
Dataset I		Dataset II		Dataset III		Dataset IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

15 Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Figure 2 Anscombe's quartet

But is it really necessary to visualize the data using charts? Can't we simply use tables instead? Have a look at Anscombe's quartet. Here you have four datasets. Each with two variables and eleven observations. If you take your time, it should not be hard to detect the differences. But look what happens when you visualize this data using some basic charts.

Why do we visualize data?



16 Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Figure 3 Anscombe's quartet visualized

It took you probably less than a second to see how different these four datasets are.

Does this mean we should stop using tables and always switch to charts? NO. We need both.

The key strengths of a table are to look up and compare individual values. It also has particular benefits in terms of precision.

The key strengths of a chart are revealing relationships among multiple values and when the message is contained in the shape of the values.

Data visualization is all about communicating a (data) message and is very similar to an informative conversation between two persons. Paul Grice created several principles (called ’maxims’) to describe an informative conversation. These maxims of conversation can also be applied to data visualization:

Be informative (maxims of quantity):

- Make your contribution as informative as is required. Provide all the information which is necessary for the purpose of the current exchange; don’t leave out anything important. You can only do this if you clearly understand the purpose. Without a purpose you can’t have a good conversation nor an effective data visualization.
- Do not make your contribution more informative than is required.
  Leave out any unnecessary details that aren’t important to the current exchange. Concentrate on the bare minimum.

Be truthful (maxims of quality):

- Do not say what you believe to be false.
  Avoid stating information that you believe might be wrong unless there is some compelling reason to do so.

If you do choose to include it, then provide a disclaimer that points out your doubts regarding this information.

- Do not say anything for which you lack evidence.
  Avoid including information that you can't back up with supporting evidence. If you do choose to include such information for some reason, provide a disclaimer that points out your doubts.

Be clear (maxims of manner):

- Avoid obscurity of expression.
  Avoid forms which are difficult to understand: e.g. chart types that your audience does not know how to read.
- Avoid ambiguity.
  Test to see if others read and interpret your intentions correctly. Make sure there is as little ambiguity as possible.
- Be brief.
  Provide information in a concise manner, that allows your recipient to focus on the key message.
- Be orderly.
  Provide information in an order that makes sense and makes it easy for your recipient to process it.

Be relevant (maxims of relation):

- Make sure that all the information is relevant to the current exchange.

Data visualization is clearly all about communication. According to Jock Mackinlay, there are two fundamental requirements for effective data visualization:

Expressiveness

- If the relevant information of a dataset is expressed by the visualization, and only this.
- Tell the truth and nothing but the truth. (don't lie, not even by omission).

Effectiveness

- A visualization addresses the capabilities of the human visual system: it's easier to understand.

- Use code that humans are best at decoding. (best = faster and/or more accurate).

2.2 Quantitative thinking

Figure 4 Definition of statistics

Here is the definition of statistics: "A discipline that concerns the collection, organization, analysis, interpretation, and presentation of data." For data visualization we don't have to be experts in statistics, but familiarity with some basic concepts will improve our ability to create effective data visualizations.

2.2.1 Correlation vs causation

Correlation is about how two variables are related: when one variable changes, the other tends to change too. If the variables are positively correlated, they both increase or decrease together. If they are negatively correlated, as one variable goes up, the other goes down, and vice versa. It's important to remember that correlation doesn't mean there is a cause-and-effect relationship between the variables. It only shows that there is an association between them.

Causality, on the other hand, is about a cause-and-effect relationship between two variables. In a causal relationship, changes in one variable directly lead to changes in the other. Proving causality goes beyond just finding a correlation. It requires strong evidence that shows how one variable influences the other, while also ruling out the possibility of other factors or coincidences causing the observed relationship.

By default, we need to assume there is NO causation.

2.2.2 The power of context

Figure 5 Power of context

Without appropriate context a number is just a number. We need to include enough context in our data visualizations to avoid possible misinterpretations.

An expert in data analysis, Donald Wheeler, even goes one step further: “Don't trust anyone who doesn't provide context about their data”.

In summary, context is essential for framing data, making it relevant and understandable to the intended audience. Successful data visualizations not only present data but also provide the necessary context to facilitate accurate interpretation and meaningful insights.

2.2.3 Summarizing data

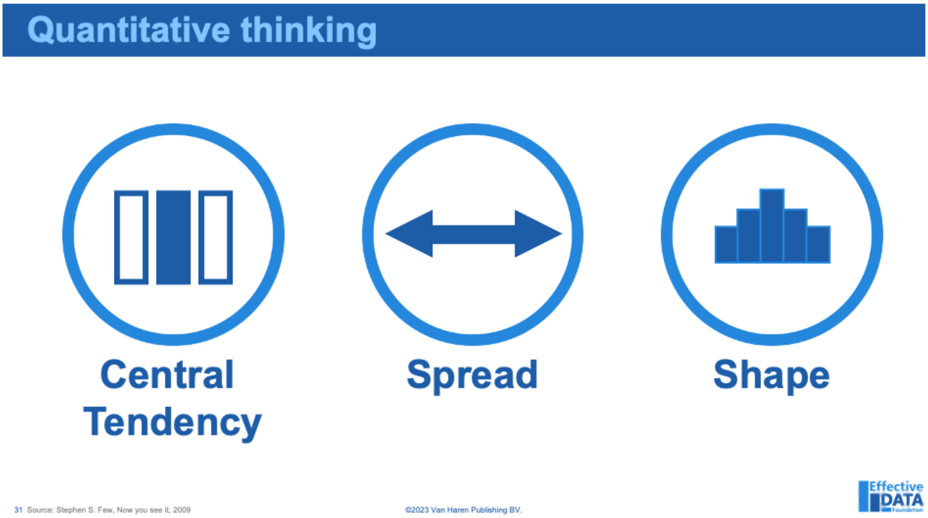


Figure 6 Three characteristics to summarize data

These three characteristics summarize the distribution of any set of values when displayed visually:

- Measure of central tendency:
  - Mean: The mathematical average: the sum of a set of values divided by the number of values in that set.
  - Median: The value which is precisely in the middle of a sorted dataset.
  - Mode: The value that appears most frequently in a dataset.
- Spread
  This provides information about the variability or dispersion of the data points. It allows us to understand variability and possibly identify outliers. Common measures of spread include range (difference between minimum and maximum values), variance and standard deviation.
- Shape
  The two summary measures above (central tendency and spread) can hide important patterns in your data. Therefore, it is important to always investigate the shape

of your data set by visualizing it. Popular ways to visualize the data set are histogram, frequency polygon and strip plots.

It is important to add that if you leave any of these three characteristics out you increase the risk of fooling yourself and your audience.

2.2.4 Charts

Charts represent relationships in quantitative data by showing those relationships visually. With this, you can easily match the proper quantitative message you want to share with the structural design that can most effectively do the job. Charts vary primarily in terms of the types of relationships they use to share, so it's helpful to understand the specific types of relationships that charts can represent.

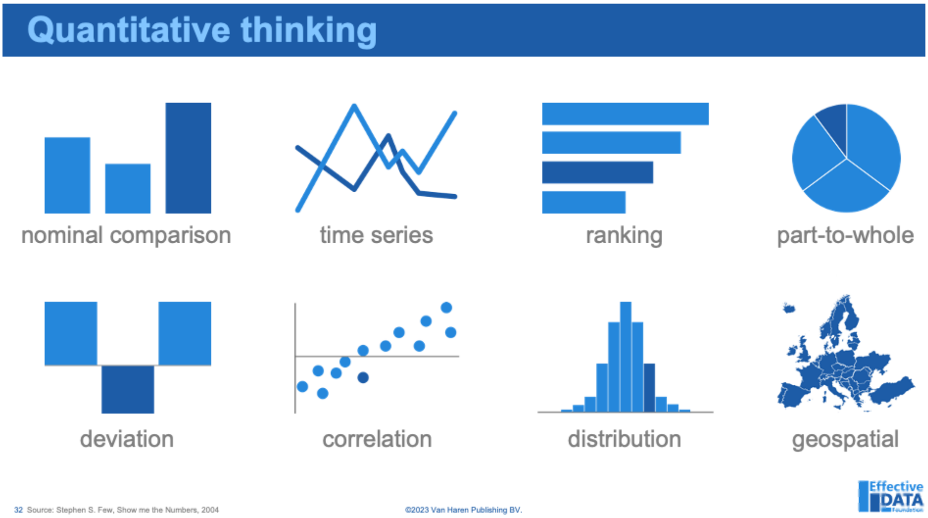


Figure 7 8 Main relationships in data

There are eight main relationships in graphs:

1. **Nominal comparison**: A simple comparison of values for a set of unordered items.
2. **Time series**: Values display how something has changed through time.

3. **Ranking**: Values are ordered by size (ascending or descending).
4. **Part-to-whole**: Values represent parts (proportions) of a whole.
5. **Deviation**: The difference between two sets of values.
6. **Correlation**: Comparison of two paired sets of values to determine if there is a relationship between them.
7. **Distribution**: Counts of values per interval from lowest to highest.
8. **Geospatial**: Values are displayed on a map to show their location.

These eight main relationships will help us choose the most appropriate charts.

3 Human Perception

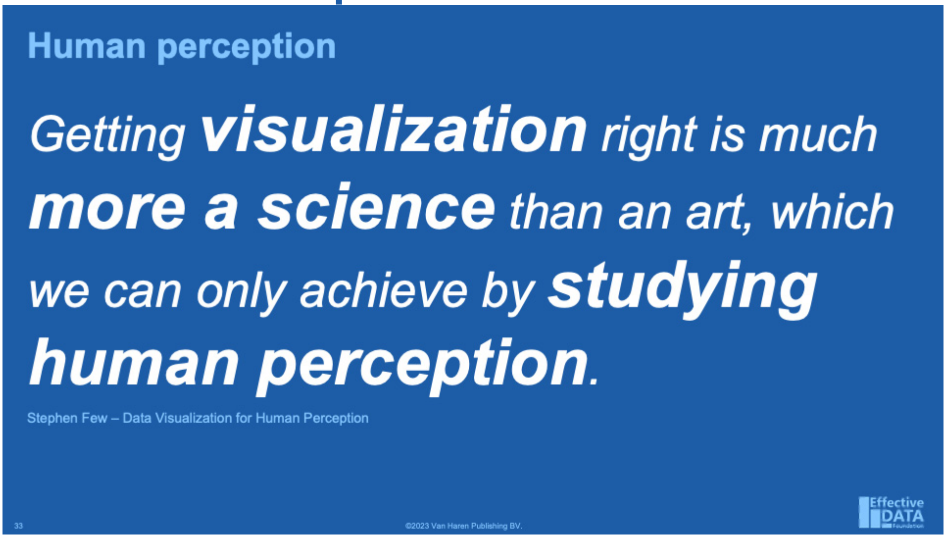


Figure 8 Human perception

Knowing how people perceive things is essential when making a data visualization. A message in visualization can come across as entirely different than intended by making wrong design decisions.

Our vision is the most powerful and efficient way we receive information from the world around us. It accounts for about 70% of our sensory input. The brain plays a crucial role in how we interpret what we see. The objects we perceive are composed of different visual elements, such as length, width, shape, size, and color. These individual attributes, also known as "pre-attentive attributes", allow us to quickly distinguish and recognize them without conscious effort. Our eyes transmit this information to the "visual cortex", a part of the brain located at the back of our head, through a fast connection, enabling rapid processing of visual data.

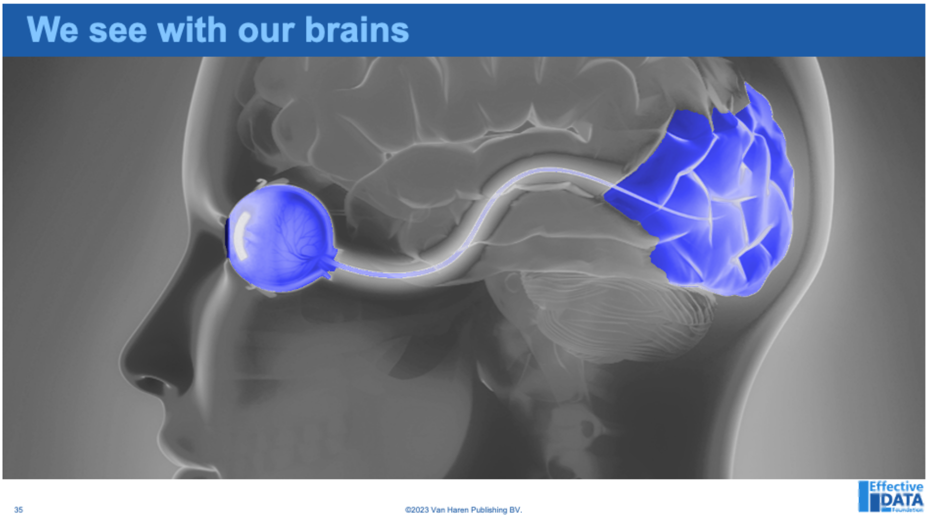


Figure 9 Our visual system

Some basic signals can be processed by the visual cortex itself. Anything that the visual cortex can process itself is what we call fast.

When other parts of the brain need to get involved, it is slow.

To give you an idea about what is fast or slow: most of us blink our eyes around 20 thousand times each day. But we do not “perceive” our surroundings as a blinking film. This is thanks to our visual cortex, who filters out the blinking.

3.1 Pre-attentive attributes

Visual analytics leverages pre-attentive attributes to guide you down the most useful paths. Pre-attentive attributes are information we can process visually almost immediately, before sending the information to the attention processing parts of our brain. The pre-attentive attributes are shown below.

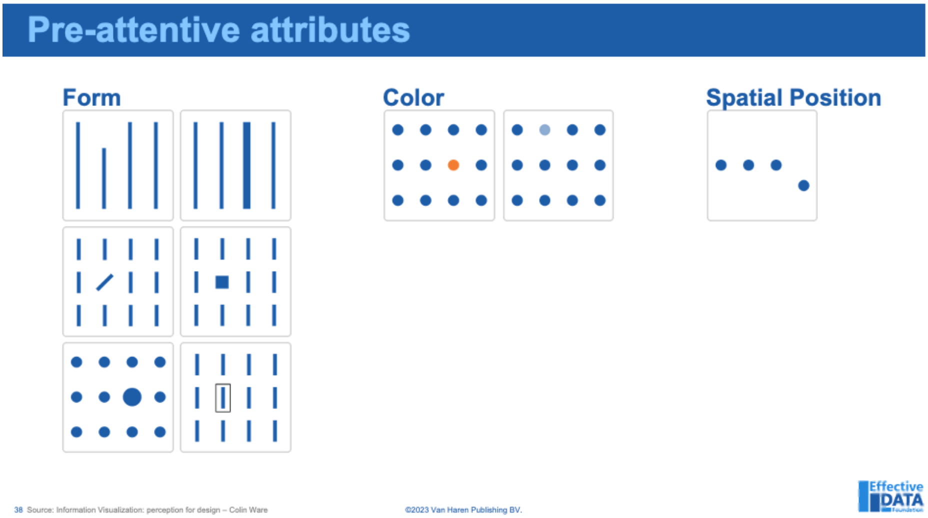


Figure 10 Pre-attentive attributes

Form:

- Line length: We can clearly distinguish the different lengths. The longer the line, the higher the value it represents. A typical chart that uses this attribute is the bar/column chart.
- Line width: Here we see the difference in width (Sankey diagram, Napoleon's march to Moscow) where wider means a greater value.
- Line orientation: We use this on a clock to read the time, but also in a line chart to show any increase or decrease between two points. Another chart that uses line orientation is the pie chart: it shows the size of the pie-piece.
- Shape: Although this is a pre-attentive attribute, it does not allow us to encode numeric values (large, too small or vice versa). In general, we use this attribute to identify markers in a visualization, like in a dot plot, strip plot or scatter plot.
- Size: The area size is another clear attribute to use in encoding values. The bigger the size, the bigger the value. Often applied in proportional symbol maps and bubble charts.

- Enclosure: Like shape this attribute only works for identification purposes, and not for expressing an amount or size.

Color:

- Hue: Popular attribute to identify an object as being a member of a group (same color). Does not work well to express a quantity, because most people are not familiar with the natural order of colors.
- Intensity: Color intensity is a popular attribute to express quantities. Dark color represents a high value and light color a low value.

Spatial position

- 2D position: One of the most popular attributes to express quantities. Used, amongst others, in line charts and scatter plots.

Movement

- This training only addresses the static attributes.

Line length and 2D position achieve the highest precision. However, precision is not always needed, so visuals that use less precise attributes also have their purpose.

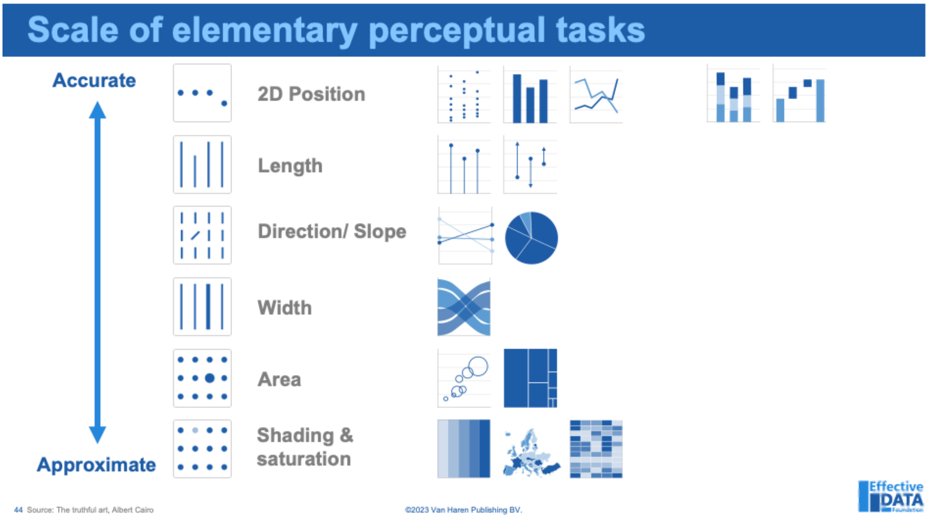


Figure 11 Order of pre-attentive attributes on accuracy

The elementary perceptual tasks (pre-attentive attributes) can be grouped into six categories:

1. **2D position**: Bar & column chart, line chart, stacked bar chart, waterfall chart.
2. **Length**: Lollipop chart, dumbbell chart.
3. **Direction & slope**: Slope chart, pie/donut chart.
4. **Width**: Sankey chart.
5. **Area**: Bubble chart, tree map.
6. **Shading & color saturation**: Choropleth map, heatmap.

The higher the attribute in this list the more accurate it can be read. The lower you get on this list, the more approximate are the values received.

3.2 Memory

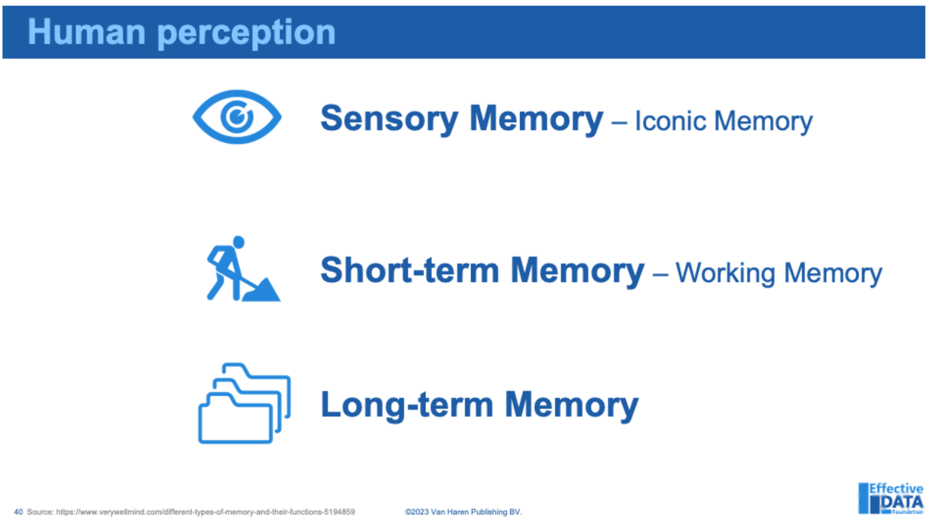


Figure 12 Memory types

Roughly speaking, we have three types of memory:

- **Sensory memory:**
  - Extension of our senses.
  - Flight or fight.
  - Visual aspect is called iconic memory.
  - Large capacity, very short duration (fractions of seconds).
- **Short-term memory** (working memory):

    - Small capacity - maximum four pieces of information.
    - Very dynamic - we can apply operations to the information (hence working memory).
    - Not really stored for the long run.
- **Long term memory:**
    - Huge capacity for permanent storage.
    - Knowledge and experiences.
    - Three processes -save, hold, retrieve.
    - Slow and hard to retrieve.

As designers, we especially need to consider sensory and short-term memory. An essential part of this is pre-attentive processing.

3.3 Gestalt principles

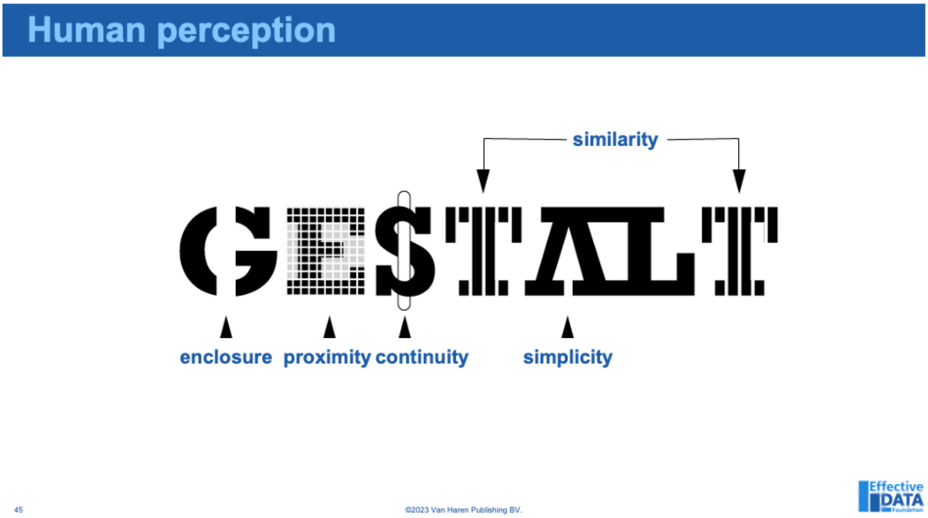


Figure 13 Gestalt

Gestalt is a psychology term which means "unified whole". It refers to theories of visual perception developed by German psychologists in the 1920s. These theories attempt to describe how people tend to organize visual elements into groups or unified wholes when certain principles are applied. These principles are:

- **Enclosure:** Objects that appear to have a boundary around them, for instance formed by a border or area of common color, are perceived as a group.
- **Proximity:** Objects that are close together are perceived as a group.
- **Continuity:** Objects that are aligned together or appear to be a continuation of one another are perceived as a group.
- **Simplicity:** We perceive the simplest shape possible, even when it is made up of several different shapes.
- **Similarity:** Objects that share similar attributes, like color or shape, are perceived as a group.

These principles help to get around the limitations of our short-term memory. We can remember and process more by putting small parts together into a whole.

3.4 Thinking, fast and slow

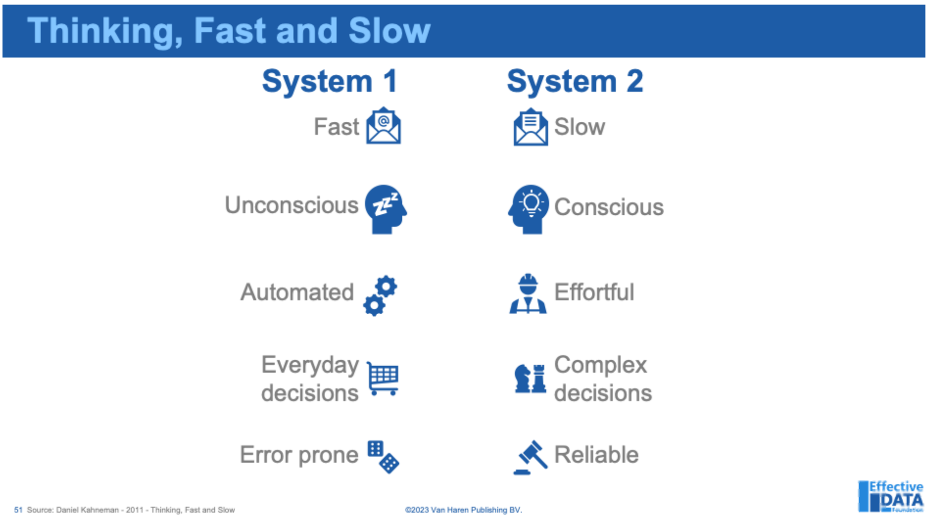


Figure 14 System 1 and System 2 thinking

In his popular book Daniel Kahneman describes the two main systems of thinking, as described below.

System 1: Works automatically and fast, with little or no effort and no sense of control. Think of making simple everyday decisions like solving the sum 2+2, identifying the source of a particular sound, or driving a car on an empty road.

System 2: Involves conscious attention to the mental effort expended, making complicated calculations. We often link this system's functioning to subjective experience, choice, and concentration. Like deciding to buy a new house, calculating 13 * 28, or comparing the price-quality ratio of a dishwasher.

Data is abstract in nature, and therefore always considered as complex. For any data analysis or decision-making we need to make use of System 2 (analytical brain), because System 1 (our instinct) tends to make us believe we know (but we don’t). Data (visualization) is never intuitive. It should always be read carefully.

4 Visualizing Data

Figure 15 It's not the ink...

4.1 Start with WHY?!

There is no "perfect chart", it all depends on the question you want to answer.

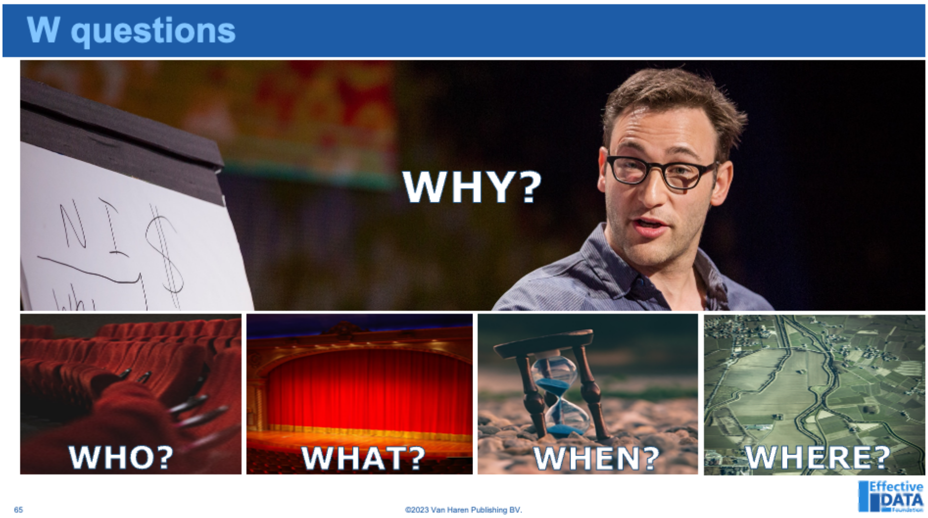


Figure 16 W questions

To determine what the question is, always start by asking the five W questions:

- **Why**: To get the real purpose of the question/request.
- **Who**: For whom are you doing this, who is the audience: your colleague, manager, department, company, or outside clients? They all have different needs, and they all have different knowledge levels.
- **What**: Our visuals need to focus on signals, not noise. Noise is random, and explaining noise does not make any sense. Also make sure reading your visual is as easy as possible for the audience. Avoid indirect measures. Strengthen your visualization by including direct measures.
- **When**: Time is a great way to show more context; what has happened before, is there a trend, how is the flow, are there ups and/or downs or not?
- **Where**: What type of medium are you presenting the result on? Is it on a smart watch, a mobile, tablet, laptop, a screen, a poster, or even a billboard? And does physical location play a role in answering the questions?

4.2 Table or chart?

When do you choose a table, and when do you select a chart?

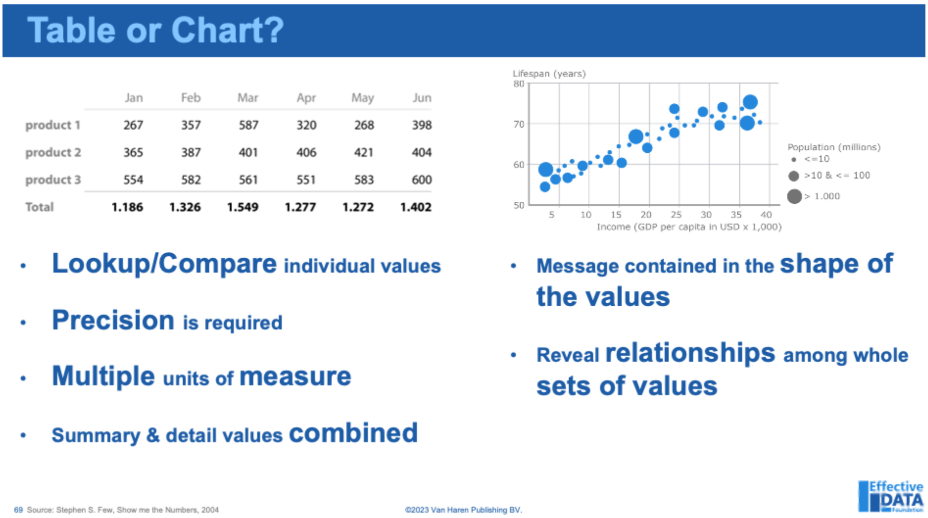


Figure 17 Table or chart?

Use a TABLE when:

- The display will be used to look up and/or compare individual values.
- Precise values are required.
- The quantitative values include more than one unit of measure.
- Both detail and summary values are included.

Use a CHART when:
- The message is contained in the shape of the values (patterns, trends, and exceptions).
- The display will be used to reveal relationships between whole sets of values.

4.3 CHRTTS

CHRTTS is a mnemonic to help organize your memory and enable you to think about which chart(s) to use for your visualization needs.

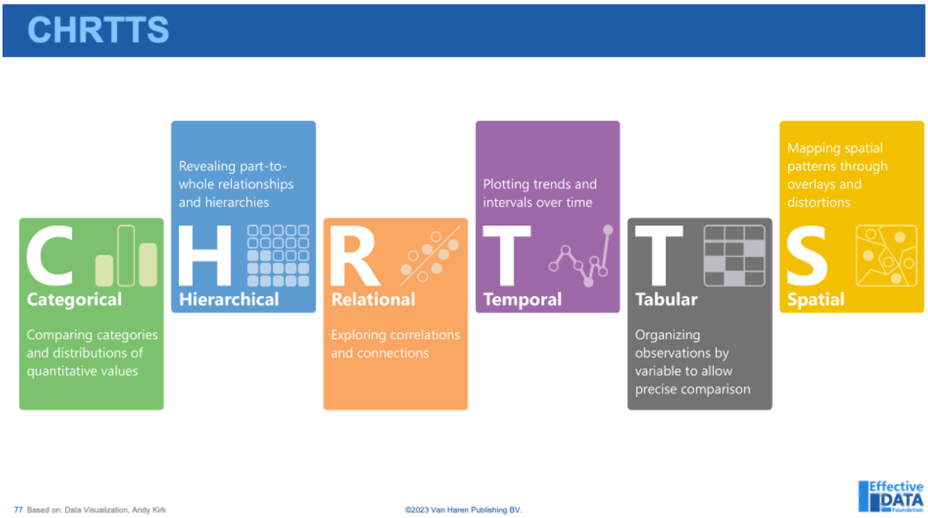


Figure 18 CHRTTS

The factors to consider are:
- Categorical: Comparing categories and distributions of quantitative values.
- Hierarchical: Revealing part-to-whole relationships and hierarchies.

- Relational: Exploring correlations and connections.
- Temporal: Plotting trends and intervals over time.
- Tabular: Organizing observations by variables to allow precise comparison.
- Spatial: Mapping spatial patterns through overlays and distortions.

4.3.1 Categorical

Comparing categories is about quickly and easily seeing the differences between the same values of different categories.

The most effective form for comparing values in a graphical form is with a bar or column chart. The categories' values are easy to compare because all bars have the same starting point and the same width. The reader can see and compare the relative length of the bars. In other words, the length represents a particular value of the category.

There are other options for comparing categories than the bar chart (by the way, the bar chart was introduced in 1785 by William Playfair, a clear example of proven technology).

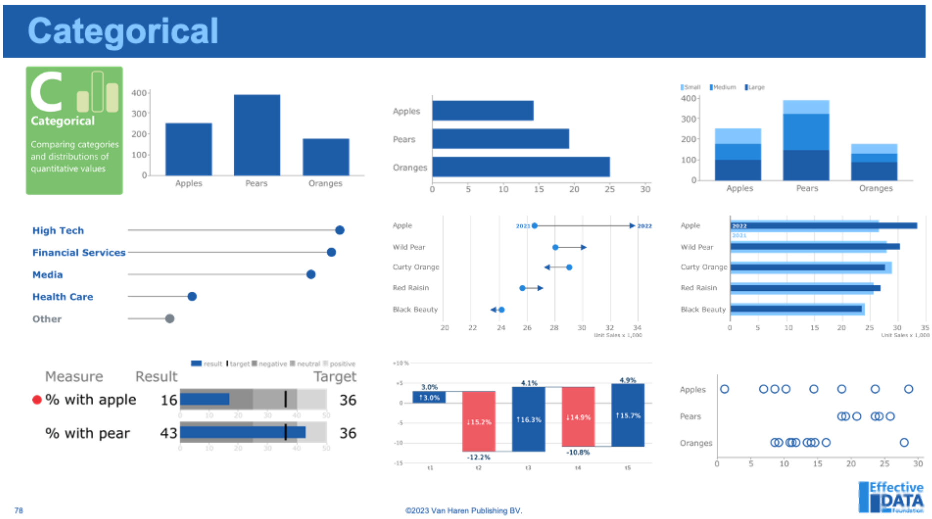


Figure 19 Categorical examples

Typical charts used for comparisons are (as shown above):

- Column chart.
- Bar chart.
- Stacked bar chart.
- Lollipop chart.
- Dumbbell chart.
- Lipstick chart.
- Bullet graph.
- Waterfall chart.
- Strip plot.

For showing distributions we often use:

- Histogram.
- Strip plot.
- Frequency polygon.
- Box plot.
- Violin plot.

And for deviation and ranking we often use the bar or column chart.

4.3.2 Hierarchical

In this section we look at the charts that can reveal part-to-whole relationships and hierarchies.

A part-to-whole relationship simply means understanding how individual pieces fit into the bigger picture. It's like seeing how different sections build up to a whole story. It's about recognizing that every part contributes to the overall structure or meaning of something larger.

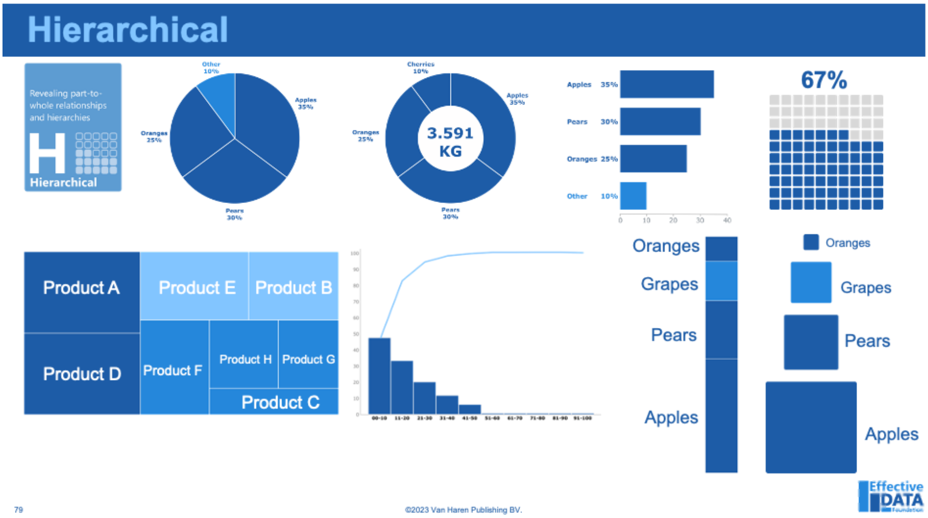


Figure 20 Hierarchical examples

The most commonly used chart for showing a part-to-whole relationship is the pie chart, or the derived form: the donut chart. There is a lot of debate surrounding the use of these visuals. We will discuss this later in the next chapter, “Designing data visualizations”.

Alternatives to the pie chart are:

- Bar chart.
- Waffle chart.
- Tree map.
- 100% stacked bar chart.
- Proportional symbol.

The Pareto chart visualizes the individual contributions (the columns) and the total (line) to show when you reach a certain threshold such as 80%.

4.3.3 Relational

The visuals in this section can help you to understand how two or more variables are related or correlated. Just keep in mind that two measures might be correlated, but it does not mean there is a causal relationship.

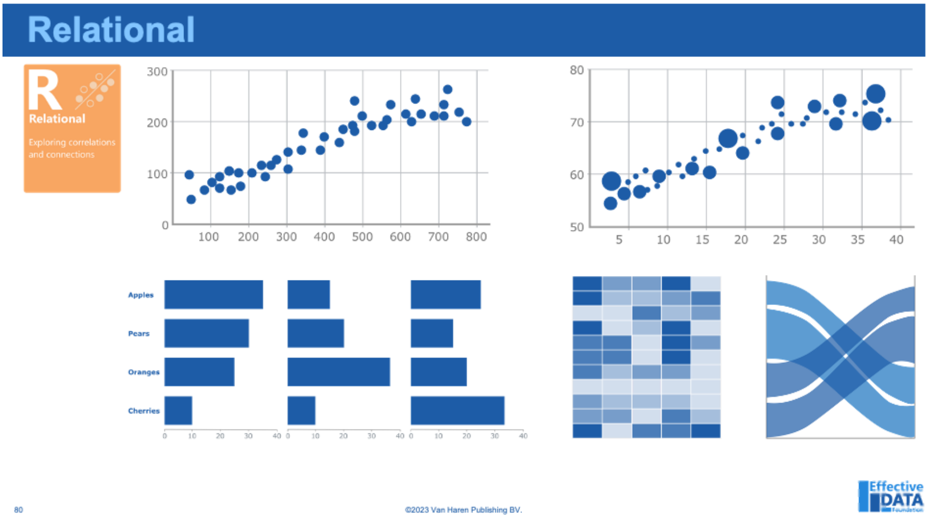


Figure 21 Relational examples

The charts shown here are:

- Scatter plot.
- Bubble chart.
- Merged bar chart.
- Heatmap.
- Sankey chart.

4.3.4 Temporal

Temporal is for situations where you want to show the importance of time in relation to your data. In contrast to bars, which emphasize values, lines emphasize continuity and flow from one value to the next. Lines are best used to display a continuous data series over a period of time. Lines are used for showing trends, acceleration or deceleration and volatility, including sudden peaks or troughs.

Figure 22 Temporal examples

Typical chart types used for this are:

- Line chart.
- Radar chart (also known as web chart or spider chart).
- Area chart (stacked).
- Column chart.
- Slope chart.
- Gantt chart.

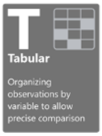
The line chart and column chart are by far the most popular types in Temporal. But when do you select the line and when do you select the column? (like you see in the example above).

When your objective is to show the form and direction of the data over time, the line chart is the better choice. When you want to compare the results of a specific period with the results of another period (close to each other), the column chart does a better job. Also, when you want to compare multiple series within one period (in the example the light blue and dark blue series), the column chart is more effective, though it also has a tendency to clutter the visual.

4.3.5 Tabular

It may not be the first thing that comes to mind when you think of data visualization, but the table is an essential basic shape. It is powerful in comparing (with great precision) individual values. The table also makes it possible to compare values of different units of measure (e.g. percentage of the total, average, amounts of money, numbers, etc.).

Tabular



		Jan	Feb	Mar	Apr	May	Jun
market A	product 1	267	357	587	320	268	398
	product 2	365	387	401	406	421	404
	product 3	554	582	561	551	583	600
market B	product 4	267	357	587	320	268	398
	product 5	365	387	401	406	421	404
	product 6	554	582	561	551	583	600

Market	Product	Month	Measure
market A	product 1	Jan	267
market A	product 2	Jan	365
market A	product 3	Jan	554
market B	product 1	Jan	267
market B	product 2	Jan	365
market B	product 3	Jan	554
…	…	…	…

82

Figure 23 Tabular examples

A matrix (aka cross-table) and a table visualization are both types of data representation used in data analysis. However, they differ in their purpose and layout:

- A matrix (on the left) is a two-dimensional representation of data that is used to visualize the relationships between categories or dimensions (with a hierarchy). In a matrix visualization, the cells of the matrix represent the intersections between the rows and columns, and the values in the cells can be used to represent a variety of measures, such as counts, sums, or averages. Matrix visualizations are typically used to highlight patterns and trends in large data sets.
- A table (on the right) is a simple arrangement of data in rows and columns, where each row represents an observation, and each column represents a field or

attribute of that observation. Tables are used to display detailed information about individual observations and to compare different attributes side by side.

4.3.6 Spatial

When location is relevant to your message, you can decide to use a spatial visualization. The most common type is a geographical map, which shows a city, a country, a region or even the world. Have you ever thought of visualizing a building, a floor, or a room? These objects also contain locations and can also be used in this category.

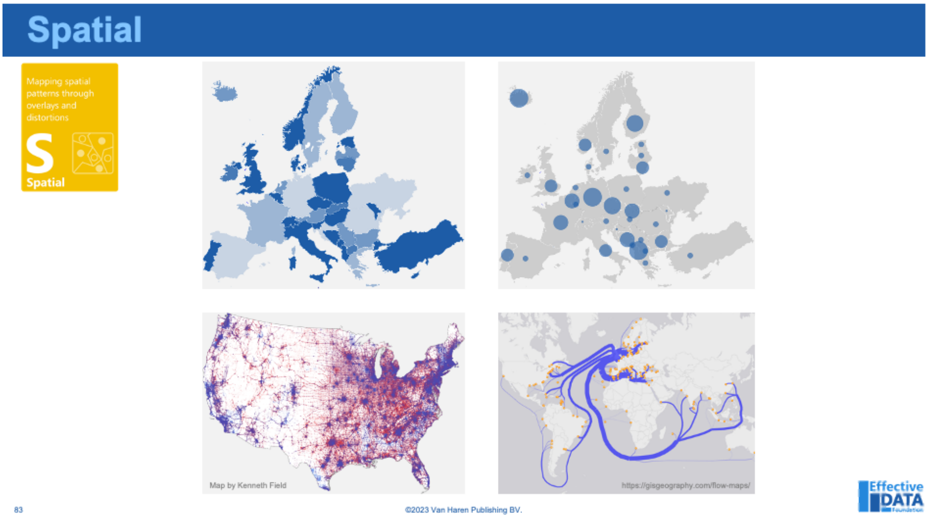


Figure 24 Spatial examples

Often, the maps are used to color a specific area. This map type is called a choropleth map. You can also position an object (often a circle) on each location and increase/decrease the size based on a quantity. This is called a proportional symbol map. The dot map is where you add a marker to a map for each observation. In the example shown you see about 126 million dots (markers), one for each voter during the presidential elections in 2016. Another popular use of maps is to include flows (flow map). In the example above you see English coal exports in 1864. The wider the line, the more coal is exported to a destination.

5 Dataviz Design

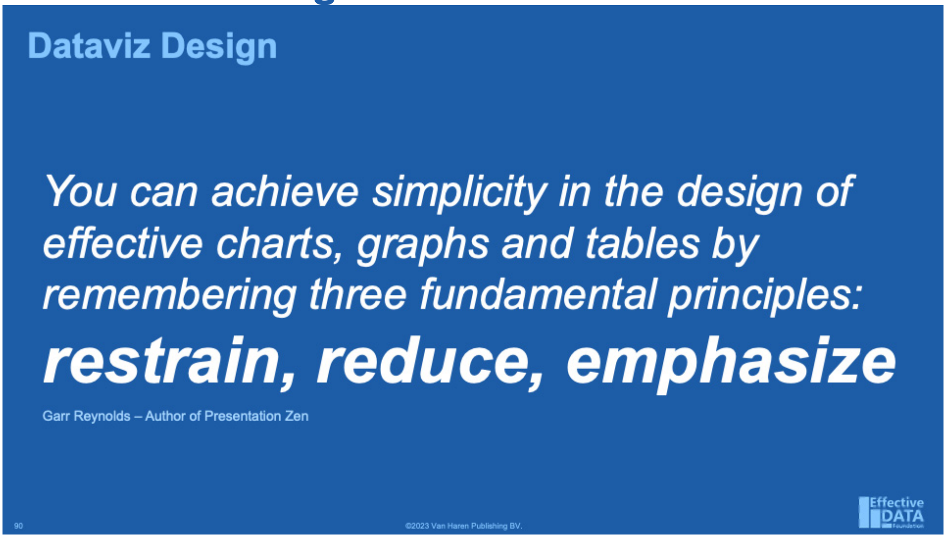


Figure 25 Data visualization design

5.1 Categorical

Comparing categories is about quickly and easily seeing the differences between the same values of different categories.

5.1.1 Bar/column chart

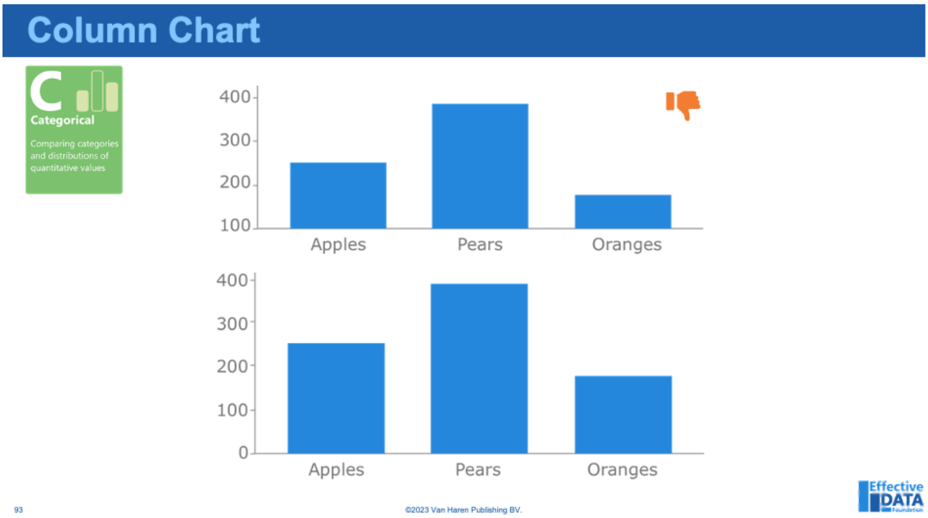


Figure 26 Column chart