

Data Literacy Practitioner's Guide

EDF Certification Handout

Michel Dekker



Data Literacy
Practitioner's Guide



Colophon

Title:	Data Literacy Practitioner's Guide
Subtitle:	EDF Certification Handout
Authors:	Michel Dekker
Publisher:	Van Haren Publishing, 's-Hertogenbosch
ISBN Hard Copy:	978 94 018 1131 6
Edition:	First edition, first print, March 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

No part of this publication may be reproduced in any form by print, photo print, microfilm or any other means without written permission by the publisher. Although this publication has been composed with much care, neither author, nor editor, nor publisher can accept any liability for damage caused by possible errors and/or incompleteness in this publication.

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 is also publishing on behalf of leading organizations and companies: ASLBiSL 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-CMF™
IT Service CMM
ITIL®
MOF
MSF
SABSA
SAF
SIAM™
TRIM
VeriSM™

Enterprise Architecture

ArchiMate®
GEA®
Novius Architectuur
Methode
TOGAF®

Business Management

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

Project Management

A4-
Projectmanagement
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

As kids, we learned our native language: reading, writing, and speaking. You learn a language by starting with grammar, trying it out, making mistakes and improving it.

We are flooded with facts and figures daily, but how well do we understand the meaning behind all those numbers? How can we turn this data into value in our day-to-day work and for the organization we work for? The ability to derive meaningful information from data is called Data Literacy. Making sense of data is no longer just a skill for data scientists and technology experts but an essential skill for all of us.

Data Literacy is the ability to read, work with, analyze and argue with data.

This course is the first step to make you aware of Data Literacy (as an essential skill) and how it impacts your work. You'll learn the fundamentals and how they relate to working in a data-informed organization. This will be enough for some people to raise the required questions when confronted with data. For others, this is the first step in becoming a fluent data speaker.

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

	Weight	Topic
Introduction		Introduction to Data Literacy
Read data	25%	The ability to read and interpret data correctly. Which questions we need to ask to avoid fooling ourselves.
Work with data	25%	What happens to data during its journey from the source to final consumption? How does this impact the understanding and possibilities of this data?
Analyze data	25%	Data needs to be analyzed, not only read. In this section we'll have a closer look at how we analyze data.
Argue with data	25%	Once we have found interesting insights in the data, we need to share this with our audience. In this part we'll look at what we need to do so our audience understands the data in the best possible way.

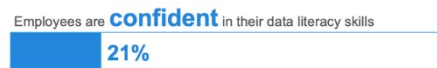
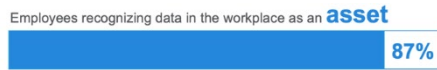
1	COURSE OVERVIEW	4
2	INTRODUCTION	8
3	READ DATA	10
3.1	What is data?	10
3.1.1	Data scales	11
3.1.2	Why do we use data?	13
3.2	Summarize data	14
3.2.1	Central tendency	15
3.2.2	How to summarize data	17
3.3	Consume data	18
3.3.1	CHRTTS	19
3.3.2	Categorical charts	22
3.3.3	Hierarchical charts	23
3.3.4	Relational charts	24
3.3.5	Temporal charts	25
3.3.6	Tabular cCharts	26
3.3.7	Spatial charts	27
3.4	Check your data	28
4	WORK WITH DATA	32
4.1	Creating data	32
4.1.1	Automated data	32
4.1.2	Manual data	32
4.2	Data quality	33
4.2.1	Data quality dimensions	33
4.2.2	Avoid confusing data	35
4.3	Acquiring & cleaning data	36
4.3.1	Tidy data	37
4.3.2	Combine data	38
4.4	Managing data	39
4.4.1	KPI mantras	40
5	ANALYZE DATA	43
5.1	Expectations	43
5.2	Thinking shortcuts	43

5.2.1	Confirmation bias	45
5.2.2	Survivorship bias	45
5.2.3	Curse of knowledge	45
5.2.4	Correlation vs causation	46
5.3	Types of analysis	47
5.3.1	Descriptive analysis	47
5.3.2	Diagnostic analysis	48
5.3.3	Inferential analysis	48
5.3.4	Predictive analysis	49
5.3.5	Prescriptive analysis	51
5.4	Analytical skills	51
5.4.1	Variations within categories	52
5.4.2	Relations among categories	53
5.4.3	Variations within measures	53
5.4.4	Relations among measures	54
5.4.5	Look for patterns	55
6	ARGUE WITH DATA	58
6.1	Explore to explain	58
6.1.1	The data cut	59
6.1.2	The data cameo	59
6.1.3	The data decoration	60
6.2	Effective data visualization	61
6.3	Storytelling with data	62
6.3.1	The Storytelling Arc	62
6.3.2	Visual storytelling principles	64
7	FURTHER READING	72

2 Introduction

When asked, most of us recognize data in the workplace as an asset. This was clearly demonstrated in “The Human Impact of Data Literacy” study in 2020 where 87% of employees recognized data as an asset.

But if we ask employees whether they trust their decisions more when based on data, only 37% says “yes”. And when asked about confidence in their own data literacy skills, only 21% confirmed this was the case.



Source: *The Human Impact of Data Literacy – 2020 – Qlik, Accenture, Data Literacy Project. N=9,000*

We clearly have a challenge when it comes to our abilities to effectively use the data around us.

The training consists of four modules (3.5 hours per module). Two modules per training day (7 hours per ILT classroom).

We follow the four main components from the Data Literacy definition by MIT: Read data, Work with data, Analyze data and Argue with data.

The ability to **read, work with, analyze, and argue with data**

Source: Raul Bhargava and Catherine D'Ignazio from MIT and Emerson College



We'll start with Read data, to learn and understand what data is and what aspects of the world it represents.

Next, we'll talk about

Work with data, this involves creating, acquiring, cleaning, and managing data.

The second day we start with Analyze data. This is about using data to make informed decisions. Filtering, sorting, aggregating, comparing, and performing other analytical operations.

Finally, we wrap up this training with the module Argue with data: Using data to support a larger narrative intended to communicate some message to a particular audience.

3 Read Data

3.1 What is data?

Data itself is not reality, it reflects reality, like the plane in the water. It can represent and provide insights into various aspects of reality. Data is a representation of information gathered from observations, measurements, or records. It reflects events, phenomena, or processes that occur in the real world.



It's important to note that data is always subject to limitations and biases. The way data is collected, the selection criteria, and the interpretation can influence the insights derived from it. Therefore, while data can provide valuable information, it should be analyzed critically and in conjunction with other sources of knowledge to form a comprehensive understanding of reality. Keep in mind: all data is wrong (incomplete), but some data is useful.

Data can be defined as a collection of raw facts or figures that is typically represented in a quantitative or qualitative form. It refers to an object or an event that is collected, stored, and processed by various systems, tools, and technologies.

Describes a **quality** or **quantity**
of some **object** or **event**.

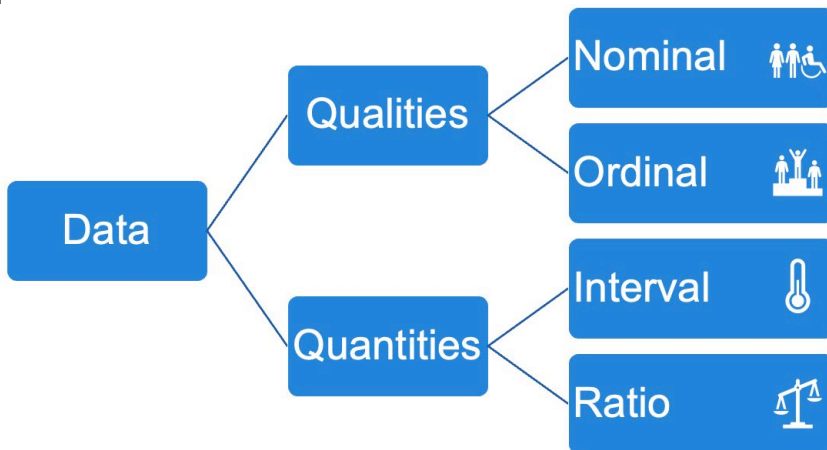
Source: unknown

The value and significance of data lie in its potential to be analyzed, interpreted, and transformed into meaningful insights and knowledge. Analyzing

data enables patterns, relationships, and trends to be identified, leading to informed decision-making and the discovery of new information. It allows us to capture and encode the properties of events and/or objects.

3.1.1 Data scales

Data can be defined as a collection of raw facts or figures that is typically represented in two types: a quantitative or a qualitative form.



Qualitative data represents a set of distinct categories or groups that an observation or data point can belong to and can have the types: nominal or ordinal.

If we use a supermarket as an example:

Nominal variables represent categories without any inherent order or ranking. Like the color¹ of packaging (red, green, pink, yellow, or blue), the product packaging type (a box, a bag, or a bottle), product type (food vs. non-food), or the product brand name.

Ordinal variables have distinct categories with a specific order or ranking. The order represents a relative ranking but not necessarily the magnitude between categories. For instance, the NUTRI-SCORE (A, B, C, D or E, a nutrition label), shelf position (eye level=high attention), or for spicy food (Mild, Medium, Hot, Very Hot, Extremely Hot).

You can't calculate with these categories despite the order because it is impossible to tell the difference between the two ordered values. Let's take the spicy food example: How big or small is the gap between Hot food and Very Hot?

Qualitative values do not have a natural numerical order or magnitude, so you can't calculate with a qualitative value.

A **quantitative** variable is a type of variable that represents numerical quantities or measurements. It is a characteristic or attribute that can be measured on a numerical scale. Unlike qualitative variables, which represent categories or qualities, quantitative variables provide quantitative information and allow for mathematical operations and statistical analysis.

Quantitative variables can be further classified into two subtypes: interval & ratio.

It's important to note that while **interval** variables have a continuous nature, they **do not possess a true zero point**. This means that statements like "twice as much" or "half as much" do not hold any absolute meaning when referring to interval variables.

¹ Officially color does have a natural order (color spectrum), but most people don't use colors in this way. In practice you can use color as a nominal variable, but officially it is an ordinal variable.

Therefore, caution should be exercised when interpreting and making comparisons based on interval data. So, it lacks a true zero point and requires careful interpretation and consideration when making comparisons or drawing conclusions. Zero does not mean there is nothing!

Examples for interval are: the store location (longitude, latitude), the maximum temperature (in Celsius) to keep the product, the year of product production.

A **ratio** variable is a type of quantitative variable that possesses all the characteristics of an interval variable, with the additional property of having a **true zero point**. Zero means there is nothing!

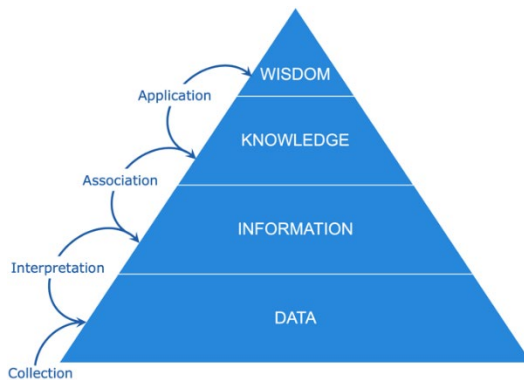
It allows for meaningful ratios, proportions, and arithmetic operations. Ratio variables can be analyzed using a wide range of statistical techniques and enable precise comparisons and inferences.

Examples: The number of items on shelf, number of items in a pack, weight, price.

3.1.2 Why do we use data?

The purpose of data is to shed light on us and on our environment, to help us distinguish between truth and falsehood, and to enable us to choose sensible courses of action to take.

In one word, the chief goal of data is wisdom.



Source: Jones, Ben. *Data Literacy Fundamentals: Understanding the Power & Value of Data*

Data is the raw material, commonly (but not exclusively) in digital form, that allows us to capture and encode facts about our world.

Information is data that has been organized and formatted so that it's useful to us in some way. In other words, it's the shape and meaning

of the data that turns it into information.

Information gets turned into **knowledge** when we incorporate it into our broader understanding of the world. We do this by linking the information we take in with other information, resulting in an accumulation of learning.

Learning is a matter of gathering knowledge; **Wisdom** is applying that knowledge.

We start with data collection, and then to turn that data into information we need to carry out accurate interpretation. And if we get from information to knowledge by making associations, then we get from knowledge to wisdom by correct application.

3.2 Summarize data

While averages provide valuable insights, it's important to note that they have limitations. They can be influenced by outliers, skewness in the data, or variations within subgroups.

Therefore, it's crucial to consider other statistical measures and explore the underlying data distribution for a comprehensive understanding.

Like this fictional example of Buena Vista City, Virginia:

In Buena Vista City live 6,690 inhabitants. The total surface area is 16,9 square kilometers. The average income of Buena Vista City is \$18,5K.



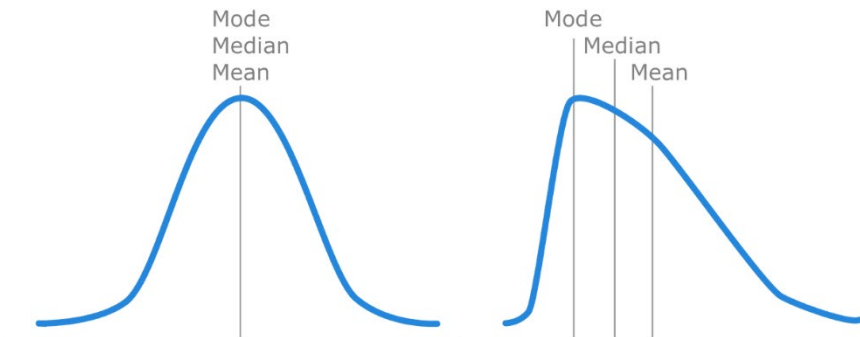
Suddenly, three men decide to buy new houses in Buena Vista City: LeBron James, Lionel Messi, and Tom Brady. The new inhabitant count of the city will be $6,690 + 3 = 6,693$.

The average income is going to change too with these new inhabitants. The new average income will rise to \$37K, meaning it doubles from the previous average!

This is a typical case of outliers and skewness in the data: the average income is not an accurate representation of the income of the city, because our new inhabitants highly skew the data.

3.2.1 Central tendency

The measures of central tendency, which include the mean, median, and mode, describe different aspects of the distribution of a dataset. Here's how each measure is defined and what it reveals about the data:



- **Mean:** The mean is calculated by adding up all the values in a dataset and dividing the sum by the total number of values. It represents the arithmetic average of the dataset. The mean is sensitive to extreme values or outliers because it considers every value in the dataset. It is commonly used when the data is normally distributed or symmetrically distributed.

Example:

Consider the dataset [2, 4, 6, 6, 8, 10].

- The mean is calculated as $(2 + 4 + 6 + 6 + 8 + 10 = 36) / 6 = 6$.

Consider the dataset [2, 4, 6, 6, 8, 100].

- The mean is calculated as $(2 + 4 + 6 + 6 + 8 + 100 = 126) / 6 = 21$.

- **Median:** The median is the middle value of an ordered dataset, separating the higher half from the lower half. To find the median, you arrange the values in ascending or descending order and identify the value in the middle. If the dataset has an even number of values, the median is the average of the two middle values. The median is not affected by extreme values and is useful when the data is skewed or contains outliers.

Example:

In the dataset [2, 4, 6, 6, 8, 10], the median is 6.

In the dataset [2, 4, 6, 6, 8, 100], the median is 6.

- **Mode:** The mode is the value or values that occur most frequently in a dataset. In other words, it represents the

most common value(s). A dataset can have no mode (when all values are unique), a single mode (when one value occurs more frequently than others), or multiple modes (when multiple values have the same highest frequency). The mode is useful for categorical or discrete data, but it can also be used with continuous data.

Example:

In the dataset [2, 4, 6, 6, 8, 10], the mode is 6 because it occurs twice, while other values occur only once.

In the dataset [2, 4, 6, 6, 8, 100], the mode is 6 because it occurs twice, while other values occur only once.

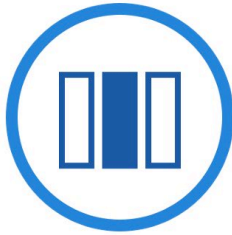
It's important to note that the choice of measure of central tendency depends on the nature of the data and the specific context of the analysis. Using multiple measures can provide a more comprehensive understanding of the data distribution and its characteristics.

3.2.2 How to summarize data

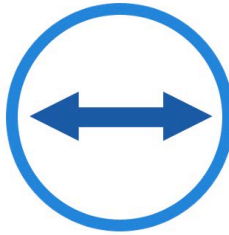
When we summarize data, we can't only look at a measure of central tendency (mean, median, mode). We also need consider two more characteristics of the data set.

To summarize a data set we need to consider all three of these data set characteristics:

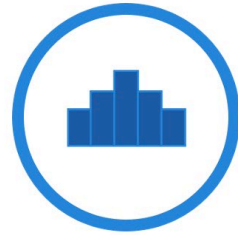
- Measure of central tendency (mean, median, mode).
- Spread (lowest to highest and their distance, standard deviation, sigma, etc.).
- Shape of the data.



**Central
Tendency**



Spread



Shape





Source: Stephen S. Few, Now you see it, 2009

3.3 Consume data

The cognitive reflection test (CRT) is a task designed to measure a person's tendency to override an incorrect "gut" response and engage in further reflection to find a correct answer; however, the validity of the assessment as a measure of "cognitive reflection" or "intuitive thinking" is under question.

According to Shane Frederick, there are two general types of cognitive activity called "system 1" and "system 2" (these terms were first used by Keith Stanovich and Richard West). System 1 is executed quickly without reflection, while system 2 requires conscious thought and effort².

² Source: https://en.wikipedia.org/wiki/Cognitive_reflection_test

System 1	System 2
Fast 	Slow 
Unconscious 	Conscious 
Automated 	Effortful 
Everyday decisions 	Complex decisions 
Error prone 	Reliable 

Source: Daniel Kahneman - 2011 - *Thinking, Fast and Slow*

The cognitive reflection test has a question that has an obvious but incorrect response given by system 1. The correct response requires the activation of system 2. For system 2 to be activated, a person must note that their first answer is incorrect, which requires reflection on their own cognition.

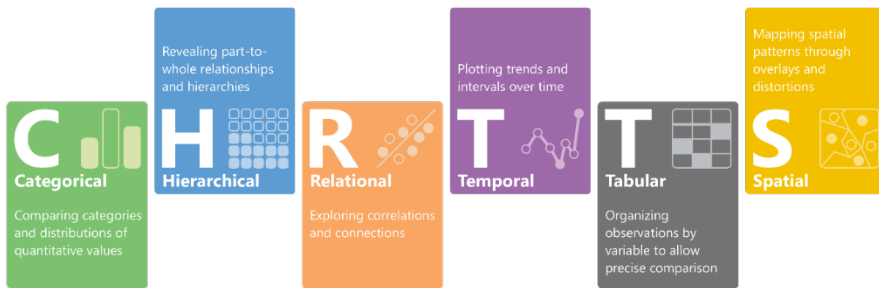
System 1: works automatically, unconscious, and fast, with little or no effort and no sense of control. Think of making simple everyday decisions like solving the sum $2+2$, identifying the source of a particular sound, or driving a car on an empty road. This is your instinct.

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 making a complex and reliable decision for buying a new house, or like solving the sum $13 * 28$, or comparing the price and quality ratio of a dishwasher. This is also called our analytical brain.

3.3.1 CHRTTS

Choosing the right chart for the right purpose is essential in data visualization because it directly impacts the effectiveness of conveying information, insights, and messages to your audience. Different types of charts are designed to represent specific types of data and patterns, and using the appropriate chart can make your data more accessible, understandable, and visually appealing. Here are some reasons why choosing the right chart is important:

1. **Clarity and communication**
Each chart type has its strengths in presenting certain types of data. Using the right chart ensures that the data is presented in a clear and concise manner, making it easier for the audience to grasp the main points without confusion.
2. **Relevance and context**
Different data visualization techniques provide different levels of detail and insight. Choosing the appropriate chart allows you to present the data in a way that is most relevant to your specific audience and context.
3. **Highlighting patterns and trends**
Some charts are better at highlighting patterns and trends in data, such as line charts for time series data or scatter plots for correlations. Choosing the right chart enables you to emphasize the key insights effectively.
4. **Comparison**
Certain charts, are excellent for comparing different categories or showing the composition of a whole. Selecting the right chart type helps you compare data points accurately.
5. **Avoiding misinterpretation**
Using an inappropriate chart can lead to misinterpretation of data. For example, using a pie chart to represent many categories can make it hard to read and understand.



Based on: Data Visualization, Andy Kirk

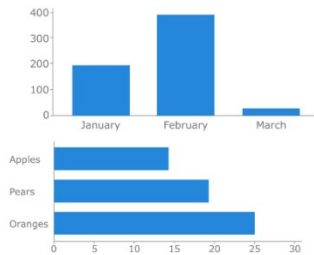
To support the process of choosing the right visual, you could use the acronym CHRTTS and hereby some chart types per category:

- **Categorical**
Bar chart, Column chart, Paired bar chart, Diverging bar chart, Dot plot, Marimekko, Isotype, Bullet chart, Bubble chart, Histogram, Pyramid chart, Strip plot, Box-and-whisker plot, Candlestick, Violin charts, Ridgeline plots, Raincloud, Stem and leaf plot, Bee swarm, Error-bars.
- **Hierarchy**
Pie chart, Donut chart, Tree map, Stacked bar chart, Waffle chart, Waterfall chart, Parallel coordinates, Chord diagram, Sunburst.
- **Relationships**
Scatter plot, Bubble chart, Merged bar chart, Heatmap, Sankey diagram.
- **Temporal**
Line chart, Bump chart, Cycle plot, Area charts, Gantt, Stream graph, Connected scatter plot, Slope graph.
- **Tabular**
Table, Matrix.
- **Spatial**
Choropleth, Dot density map, Tile grid map, Cartograms, Proportional symbol map, Flow maps.

Selecting the right chart for your data visualization depends on the nature of your data, the story you want to tell, and the audience you are addressing. Understanding the strengths and weaknesses of different chart types allows you to make informed decisions that enhance the impact of your data visualization.

3.3.2 Categorical charts

Comparing categories is all about quickly and easily seeing the



differences between the same values of different categories.

Bar charts are excellent for comparing data across different categories or groups.

The length of the bars

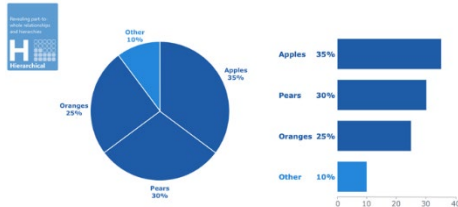
directly represents the values of the data points, making it easy to compare the magnitudes of different categories. Bar charts are ideal for representing discrete, categorical data, such as product categories, cities, months, or survey responses. Each category is represented by a separate bar, allowing you to display data for multiple categories side by side. They are simple and intuitive, making them easy for a wide range of audiences to understand. The clear visual representation of data values helps convey the main message quickly.

When you have data that can be ranked or ordered, bar charts allow you to present this information clearly. The order of the bars can represent the ranking or sorting of the categories, adding additional insights.

Bar charts are a valuable tool in the data visualization toolkit, particularly when dealing with categorical data and comparisons between different groups. By presenting data in a clear and concise manner, bar charts help audiences to understand the significance of the data and draw meaningful insights.

3.3.3 Hierarchical charts

Revealing part-to-whole relationships and hierarchies.



Part-to-whole: This kind of chart shows how the whole breaks up into constituent parts. A pie chart is the most common form for visualizing "part of a whole". There is much debate about the use

of this form. This mainly concerns the degree of precision with which the data is presented. This has to do with the corners of the pie slices in combination with the round shape of a pie.

The great power of a pie chart is that the circle (the entire pie) represents 100% in its entirety. In addition, round shapes are often found more beautiful/attractive than straight shapes.

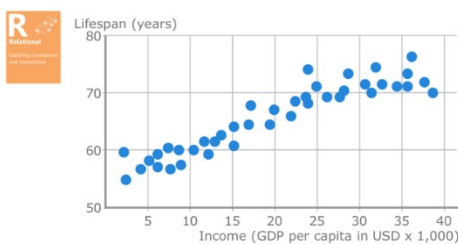
If we look at a pie chart, we translate its values by looking at the angle of a pie slice or the area of a pie slice.

If you are going to make or use pie charts, pay attention to these points:

1. Use a maximum of four pie slices. A pie chart remains legible if you limit the number of pie slices to a maximum of four pieces. If you have more categories, show the top three and add the remaining categories to the fourth (called Others).
2. Use one base color. Many different colors can be distracting so use one color and, if necessary, one other color to accentuate/emphasize a specific category. Then whiten the edges so that the pieces are clearly visible.
3. If possible (depending on the available space), put the labels and the values in the pie slices. This increases readability. With a legend next to the graph, the reader must read "back and forth". This is still possible with one chart, but if there are several, then it isn't very clear.

4. Refrain from using pie charts if you want to compare the pieces of a pie with the pieces in another pie. If you want to compare multiple values in a graph with other graphs, it is easier for the reader to compare by using a bar graph.
5. A reader compares the areas of the pieces (based on the angles). Consequently, it is intuitive to read a pie chart from top to bottom and then read it clockwise. Therefore, the ideal sorting of the pie pieces is to start at the top (at 12 o'clock) with the largest/most important point (clockwise). As a result, it is the first to receive the attention (it deserves). The other three pieces also start at 12 o'clock and then go counterclockwise in order of importance. The two most significant points are consequently at the top and are easy to compare because they have the same starting point, namely a vertical line. The two most minor important pieces will be placed where they will receive less attention. This ideal sorting is a bit more challenging in a software tool because you will have to sort the data (manually). If manual sorting is difficult or impossible, sorting from largest to smallest, starting at 12 o'clock, then going clockwise with the rest is the best method.

3.3.4 Relational charts



A scatter plot can be used to show a relationship between two (numerical) measured values. A point then represents each data point, and the horizontal (x-) axis

contains the first value, and the vertical (y-) axis contains the second value. For example, if we want to show the relationship between age and salary for a department, then age is on the horizontal axis, and salary is on the vertical axis.

This is a compelling way to show the relationship, find outliers, detect clusters, or determine the general shape. The only thing you must pay attention to is the number of points that are shown. If these become too many, then the visualization can become difficult to read (overplotting). A popular way to address overplotting is by making the marks partially transparent.

3.3.5 Temporal charts



The line diagram is great to show change over time, or movement or flow in data. Because it concerns evolution, it is a perfect way to visualize a measurement value

with intervals of (chronological) time. A straight line connects the different measurement values.

There are a few things to keep in mind when using a line chart:

1. **A line chart doesn't have to start at zero (like a bar chart)**

Unlike the bar chart, the line chart does not have to start the y-axis on 0. The values in a line chart are based on the 2D position, and not on any length.

2. **Order the horizontal axis values from lowest to highest**

When visualizing chronologically, it is easy for a reader to read if the time is ordered from lowest to highest. In the western world, we read from left to right, and if the axes are ordered similarly, it makes it easier to read.

3. **Don't use categories on the x-axis**

The line chart only works if the x-axis variable has a clear order and fixed distance between the variable values.

4. **Always show all intervals**

In addition to the previous point, do not omit gaps if there is no missing value! Always show all intervals in a

graph if there is a missing value or zero. A missing value is a strong signal that can’t be simply ignored.

5. Give the line some space

To show a clear comparison, it is good practice to keep the same amount of space both below the lowest value and above the highest value. To draw the line graph, use two-thirds of the area. Divide the remaining space above and below the graph.

6. Watch out for the spaghetti effect

A line chart is a great way of comparing multiple categories in one visualization. However, too many lines (more than 5/6) can create a kind of spaghetti which makes it harder to read. An alternative is to use “small multiples” or several small graphs in which one specific category is highlighted in each graph through the use of color.

7. Take the shortest route

Always draw the shortest line from point to point, so avoid using smoothed lines. Smoothing can obscure important details in the data and may lead to misinterpretations. Depending on the level of smoothing, it can either hide small-scale fluctuations or create false trends that do not exist in the actual data.

3.3.6 Tabular charts



	Jan	Feb	Mar	Apr	May	Jun
product 1	267	357	587	320	268	398
product 2	365	387	401	406	421	404
product 3	554	582	561	551	583	600
Total	1,186	1,326	1,549	1,277	1,272	1,402

	Jan	Feb	Mar	Apr	May	Jun
product 1	267	357	587	320	268	398
product 2	365	387	401	406	421	404
product 3	554	582	561	551	583	600
Total	1,186	1,326	1,549	1,277	1,272	1,402

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.).

A matrix and a table visualization are both types of data representation used in data analysis. However, they differ in their purpose and layout.

- A matrix 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. Great for human consumption.
- A table 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.

3.3.7 Spatial charts



If physical location is an essential factor (geographical), maps are a good tool. Using maps, the data is enriched with position and distance between places.

Here you have two European countries highlighted with the RED color: Andorra (between Spain and France) and Turkey. Which draws more attention?

Turkey, because it is bigger, so there are more red pixels.

Question is: does area play a role in our analysis?

The size of an area (here, a country) doesn't have to be related to the value shown with the color intensity (for example the number of patients). In other words, a large country seems to have a greater value because it uses a larger surface area of

the visualization. But the area says nothing about the measured value that is displayed.

Applying a map in this way only indicates the values’ geographical distribution. So usually, this shape is complemented by another visualization to add precision. It is also a good tool for identifying exceptions. Most people have no logical order in colors, so there are better options than using different colors. It is best to use only one color and use a dark variant for high values and a lighter variant for lower values

3.4 Check your data

Whenever you are exposed to data it is good to ask yourself the following eight questions:

-  Why does the data matter to you?
-  Does the claim match the data?
-  Does the claim seem plausible?
-  What comparison needs to be made?
-  Who’s saying it?
-  How was the data gathered?
-  What’s missing?
-  Is the data being distorted?



Based on: <https://www.geckoboard.com/best-practice/data-claim-checklist/>

1. **Why does the data matter to you?**

First you need to determine if the data makes you happy or sad. If you feel emotional about the data, it is best to find help from others. Ask others to explain what they conclude from the data. Whenever we feel emotional about data, we tend to fall into the traps of biases (confirmation bias).

2. **Does the claim match the data?**

Is the headline misleading? Has the data been over-simplified, over-inflated, or otherwise dramatized to become a sensational headline?

Has a generalization been made that doesn’t accurately

reflect the data?

What's the small print? Headlines often omit key details. Real life example: A 2013 Times article claimed, "More people have cell phones than toilets." However, by looking at the actual data, we find that more people have access to mobile phones than toilets. "Access" is a tricky word because it could mean that dozens of people share a single mobile phone, but the Times headline makes it sound like the number of cell phones exceeds the number of toilets.

3. **Does the claim seem plausible?**

Perform a sanity check of the claim. Do some quick back-of-the-envelope math or use your own prior knowledge. Are there other, more plausible explanations for the effect? Could they have made a mistake? Can you verify the claim in any other way? Perhaps you have access to other data or can pull a report from another source.

The less plausible the claim, the more heavily you'll want to scrutinize everything else.

Real life example: You're a customer support rep and your boss claims that "Our best customer support rep can resolve 800 tickets via phone a day." Let's do some quick math to see if this is plausible.

5 seconds (answer the phone and get the customer's name)

5 seconds (pull up customer's account and ask what the problem is)

10 seconds (customer explains the problem)

30 seconds (verify the problem or find the source)

40 seconds (fix the problem)

= 90 seconds per ticket or 40 tickets per hour

4. **What comparison needs to be made?**

Data is all about "compared to what?" Last week? Last year? Competitor(s)? Revenue?

Real life example: Several years ago, Colgate ran an advertising campaign claiming that "80% of dentists recommend Colgate." The implied comparison is that

dentists recommend Colgate over and above other brands. However, the Advertising Standards Authority discovered that in the survey, dentists could recommend more than one toothpaste. In fact, another competitor was recommended almost as often as Colgate was.

5. **Who's saying it?**

Are they an expert?

What is their agenda? Combined with the plausibility of the claim, this will affect how heavily you'll need to scrutinize the data.

Where did the data come from in the first place?

Real life example: In 1998, a research paper published in *The Lancet* claimed there was a link between certain vaccines and Autism. Several subsequent studies by independent organizations showed the author of the paper, Andrew Wakefield, manipulated the evidence to create the appearance of a link in his research.

Although he was a gastroenterologist and medical researcher, he wasn't an expert in toxicology, genetics, neurology, or other disciplines necessary to be an expert on autism. Additionally, he failed to disclose a conflict of interest as he received significant money to prove the vaccine was dangerous.

6. **How was the data gathered?**

How did they arrive at their conclusion/claim?

Often it's not that easy to gather the exact data you need/want. What was their methodology? Have any approximations been made? Were these done sensibly? Is there too much extrapolation? Were best practices followed (such as significance tests, sampling biases avoided, etc.)?

Example: Suppose you want to know how long it takes a cup of coffee (at 140 degrees Fahrenheit) to cool to room temperature. After observing for three minutes, you find the coffee cools by five degrees every minute.

If you then extrapolate that data (extending the trend of five degrees cooler per minute), you could end up with the ridiculous conclusion that after 30 minutes, the

coffee would freeze.

This extrapolation fails to consider physical limits (coffee cannot become colder than room temperature) and that the rate of cooling slows as it gets closer to room temperature.

7. **What's missing?**

Was their sample representative of the whole?

Has the data been "cherry picked" (i.e., only using the information that they want)?

Do you have other data that would help put the claim into context?

Real life example: Global warming is an often-debated topic where both "sides" have trends to back their claims. This is achieved by cherry picking only the data that supports their position, while omitting the rest.

8. **Is the data being distorted?**

In addition to cherry picking, other tactics might be employed. For example, the line chart axis might be cropped, or a misleading average might be shown.

Real life example: In 2012, Fox Business showed a chart visualizing the impact if Bush tax cuts were to expire. The top tax rate would change from 35% to 39.6%. However, the axis was cropped - beginning at 34% instead of 0% which made the tax increase appear larger than it actually was.

4 Work With Data

4.1 Creating data

Data can be created through a variety of methods, including automated and manual processes.

4.1.1 Automated data

Automated data creation typically involves the use of technology such as sensors, scanners, or software programs that collect and process data automatically without human intervention.



For example, a website may automatically track user behavior and generate data about which pages are visited and how long users stay on each page. Similarly, a manufacturing plant may use sensors to collect data about the temperature, pressure, and other variables in its production process.

A sensor is set up to detect certain activities or conditions. A lot of assumptions are often made about the conditions under which the data is collected, so this can go terribly wrong if those conditions change.

4.1.2 Manual data

On the other hand, manual data creation involves humans actively collecting and entering data into a system. This can be done through a variety of methods such as surveys, interviews, or manual data entry. For example, a marketing research firm may conduct a survey to collect data about customer preferences, or a government agency may collect census data through door-to-door surveys.

It's important to note that data creation methods can also be a combination of automated and manual processes. For instance, an online retailer may use automated software to track sales data but also employ human data analysts to

interpret the data and make recommendations for business strategy.

People might say that a lot can go wrong with this method of data entry; that is true, we humans are very good at approximations. On the other hand, the advantage is that people can deal flexibly with changing circumstances (emergence of mobile numbers instead of fax numbers).

4.2 Data quality

Often, we use the following (practical) definition of data quality: *“Data is of high quality if the data is fit for the intended purpose.”*

But what happens when the purpose changes? Or if you want to use the data more broadly? Like:

You have addresses collected to send letters to, and now you want to visit the address... But what if the collected address is a PO box?

What is a customer in your organization? A customer may be slightly different in the Finance department (payer) than in the Sales department (decision maker).

There is also another definition of data quality: *“Data is of high quality if the data correctly represents the real-world construct that the data describes.”*

This is an ideal picture, but it is virtually unattainable. It requires data to represent reality perfectly. And this is never the case.

4.2.1 Data quality dimensions

A data quality dimension is a measurable aspect or characteristic of data. The word dimension appeals to the idea that the quality of data can be assessed by looking at different characteristics of the data. There are endless lists of data quality dimensions and a full overview can't be given.

Therefore, we will only cover a short list with dimensions that are frequently used in practice:

- **Completeness**

Data is considered “complete” when it fulfills expectations of comprehensiveness. Let’s say that you ask the customer to supply his or her name. You might make a customer’s middle name optional, but if you have the first and last name, the data is complete.

There are things you can do to improve this data quality dimension. You’ll want to assess whether all the requisite information is available, and whether there are any missing elements.

- **Timeliness**

Is your information available at the point it’s needed? That data quality dimension is called “timeliness.” Let’s say that you need financial information every quarter; if the data is ready when it’s supposed to be, it’s timely.

The data quality dimension of timeliness is a user expectation. If your information isn’t ready exactly when you need it, it doesn’t fulfil that dimension.

- **Uniqueness**

“Unique” information means that there’s only one instance of it appearing in a database. As we know, data duplication is a frequent occurrence. “Daniel A. Robertson” and “Dan A. Robertson” may well be the same person.

Meeting this data quality dimension involves reviewing your information to ensure that none of it is duplicated.

- **Consistency**

At many companies, the same information may be stored in more than one place. If that information matches, it’s considered “consistent.” For example, if your human resources information systems say an employee doesn’t work there anymore, yet your payroll says he’s still receiving a check, that’s inconsistent.

To resolve issues with inconsistency, review your data sets to see if they’re the same in every instance. Are there any instances in which the information conflicts with itself?

- **Accuracy**

The term “accuracy” refers to the degree to which information accurately reflects an event or object described. For example, if a customer’s age is 32, but the system says she’s 34, that information is inaccurate.

What steps can you take to improve your accuracy? Ask yourself whether the information represents the reality of the situation. Is there incorrect data (that needs to be fixed)?

4.2.2 Avoid confusing data

As we have seen before, data can be confusing. Data is merely a representation of reality, so make sure not to confuse it with actual reality. Often, we receive data from others or share it with others. There are some pitfalls to consider in preventing confusing data.

- **Clearly understand the operational definitions of all metrics**

Make sure definitions of metrics are clear. For example, what does it mean when you count a “customer”? Is a customer somebody who visits your store, or is a customer someone who actually makes a purchase?

- **Draw the data collection steps as a process flow diagram**

- A process flow diagram greatly increases the understanding of how and when data is collected, helping you to place it in the correct context. **Understand the limitations and inaccuracies of each step in the process**

In every step of the data collection or processing process there may be limitations or inaccuracies. For example, if you are manually counting traffic at a certain road, you may miss some counts during lunch or at night.

- **Identify any changes in method or equipment over time**

Changing your method or equipment can introduce unwanted variances in your data. For instance, changing