

COURSEWARE

EDF

Data Literacy Professional Courseware

Auteur: Michel Dekker

EDF Data Literacy
Professional Courseware

Colophon

Title: EDF Data Literacy Professional Courseware

Authors: Michel Dekker

Publisher: Van Haren Publishing, 's-Hertogenbosch

ISBN Hard Copy: 978 94 018 0985 6

Edition: First edition, first print, March 2023

Design: Van Haren Publishing, 's-Hertogenbosch

Copyright: © Van Haren Publishing 2023

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.

Publisher about the Courseware

The Courseware was created by experts from the industry who served as the author(s) for this publication. The input for the material is based on existing publications and the experience and expertise of the author(s). The material has been revised by trainers who also have experience working with the material. Close attention was also paid to the key learning points to ensure what needs to be mastered.

The objective of the courseware is to provide maximum support to the trainer and to the student, during his or her training. The material has a modular structure and according to the author(s) has the highest success rate should the student opt for examination. The Courseware is also accredited for this reason, wherever applicable.

In order to satisfy the requirements for accreditation the material must meet certain quality standards. The structure, the use of certain terms, diagrams and references are all part of this accreditation. Additionally, the material must be made available to each student in order to obtain full accreditation. To optimally support the trainer and the participant of the training assignments, practice exams and results are provided with the material.

Direct reference to advised literature is also regularly covered in the sheets so that students can find additional information concerning a particular topic. The decision to leave out notes pages from the Courseware was to encourage students to take notes throughout the material.

Although the courseware is complete, the possibility that the trainer deviates from the structure of the sheets or chooses to not refer to all the sheets or commands does exist. The student always has the possibility to cover these topics and go through them on their own time. It is recommended to follow the structure of the courseware and publications for maximum exam preparation.

The courseware and the recommended literature are the perfect combination to learn and understand the theory.

-- Van Haren Publishing

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.

Table of content

	<i>--- Slide number</i>	<i>--- Page number</i>
Reflection		6
Agenda		8
Effective Data Foundation	(2)	11
Introduction	(7)	14
Read Data	(13)	17
What is Data?	(14)	17
Summarize Data	(21)	21
Consume Data	(28)	24
Check your Data	(48)	34
Work with Data	(58)	39
Creating Data	(59)	40
Data Quality	(66)	43
Acquiring & Cleaning	(72)	46
Managing	(94)	57
Analyze Data	(108)	64
Expectations	(111)	66
Thinking shortcuts	(115)	68
Types of Analysis	(135)	78
Analytical skills	(151)	86
Argue with Data	(175)	98
Explore to Explain	(177)	99
CHRTTS	(182)	101
Storytelling	(190)	104
Practice exam info		120
Syllabus		121

Self-Reflection of understanding Diagram

‘What you do not measure, you cannot control.’ – Tom Peters

Fill in this diagram to self-evaluate your understanding of the material. This is an evaluation of how well you know the material and how well you understand it. In order to pass the exam successfully you should be aiming to reach the higher end of Level 3. If you really want to become a pro, then you should be aiming for Level 4. Your overall level of understanding will naturally follow the learning curve. So, it’s important to keep track of where you are at each point of the training and address any areas of difficulty.

Based on where you are within the Self-Reflection of Understanding diagram you can evaluate the progress of your own training.

<i>Level of Understanding</i>	<i>Before Training (Pre-knowledge)</i>	<i>Training Part 1 (1st Half)</i>	<i>Training Part 2 (2nd Half)</i>	<i>After studying / reading the book</i>	<i>After exercises and the Practice exam</i>
<i>Level 4 I can explain the content and apply it .</i>					
<i>Level 3 I get it! I am right where I am supposed to be.</i>					Ready for the exam!
<i>Level 2 I almost have it but could use more practice.</i>					
<i>Level 1 I am learning but don't quite get it yet.</i>					

(Self-Reflection of Understanding Diagram)

Write down the problem areas that you are still having difficulty with so that you can consolidate them yourself, or with your trainer. After you have had a look at these, then you should evaluate to see if you now have a better understanding of where you actually are on the learning curve.

Troubleshooting

Problem areas:

Topic:

Part 1

Part 2

You have gone through the book and studied.

You have answered the questions and done the practice exam.

Timetable

Start	Duration	Subject	Start slide
09:00	00:15	Intro & Agenda	1
09:15	00:30	What is data?	14
09:45	00:15	Summarize data	21
10:00	00:25	Consume Data	28
10:25	00:20	Chart Types	38
10:45	00:10	BREAK	
10:55	00:25	Check your Data - The Fork	48
11:20	00:40	Check your Data - exercise	52
12:00	00:10	Wrap up	56
12:10	00:50	LUNCH	
13:00	00:10	Intro & Agenda	58
13:10	00:15	Creating	59
13:25	00:20	Data Quality	66
13:45	00:10	BREAK	
13:55	00:55	Aquire & Clean	72
14:50	00:50	Join	84
15:40	00:10	Managing	94
15:50	00:30	KPI Mantras	97
16:20	00:10	Wrap up	106

Start	Duration	Subject	Start slide
09:00	00:15	Intro & Agenda	108
09:15	00:20	Expectations	111
09:35	00:45	Shortcuts	115
10:20	00:40	Data Analysis	135
11:00	00:10	BREAK	
11:10	00:40	Analytical skills	151
11:50	00:10	Wrap up	173
12:00	01:00	LUNCH	
13:00	00:15	Intro & Agenda	175
13:15	00:40	Explore to Explain	177
13:55	00:55	CHRTTS	182
14:50	00:20	Storytelling ARC	190
15:10	00:10	BREAK	
15:20	00:35	Storytelling principles	193
15:55	00:25	Storytelling exercise	212
16:20	00:10	Wrap up	214

Data Literacy

how to read, work with, analyze & argue with data



COURSEWARE



©2023 - All training materials are sole property of Van Haren Publishing BV and are not to be reproduced in any form or shape without written permission.

Effective Data Foundation

not-for-profit collective,
who enables **professionals** to
leverage data to make
sustainable business
decisions



2

©2023 Van Haren Publishing BV.

DATA

Analysis / Literacy / Management / Visualization


3

©2023 Van Haren Publishing BV.



Certification

Effective Data Foundation



Congratulations to
Marly Buitenhuis
for achieving the
Data Visualization
certificate from the Effective Data Foundation.

4

©2023 Van Haren Publishing BV.



Certification



60 multiple choice questions

within **60** minutes

at least **65%** correct to pass

5

©2023 Van Haren Publishing BV.



Training structure



**Read
data**

25%



**Work with
data**

25%



**Analyze
data**

25%



**Argue with
data**

25%

6

©2023 Van Haren Publishing BV.



Data driven...



Data literacy

Employees recognizing data in the workplace as an **asset**



Employees **trust** their decisions more when based on data



Employees are **confident** in their data literacy skills



Introduction

*Data in the hands of a few data experts can be powerful but data at the fingertips of **many is truly transformational***

Brent Dykes - author of Effective Data Storytelling

Data literacy

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

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



Read
data



Work
with data



Analyze
data



Argue
with data

The Data Journey

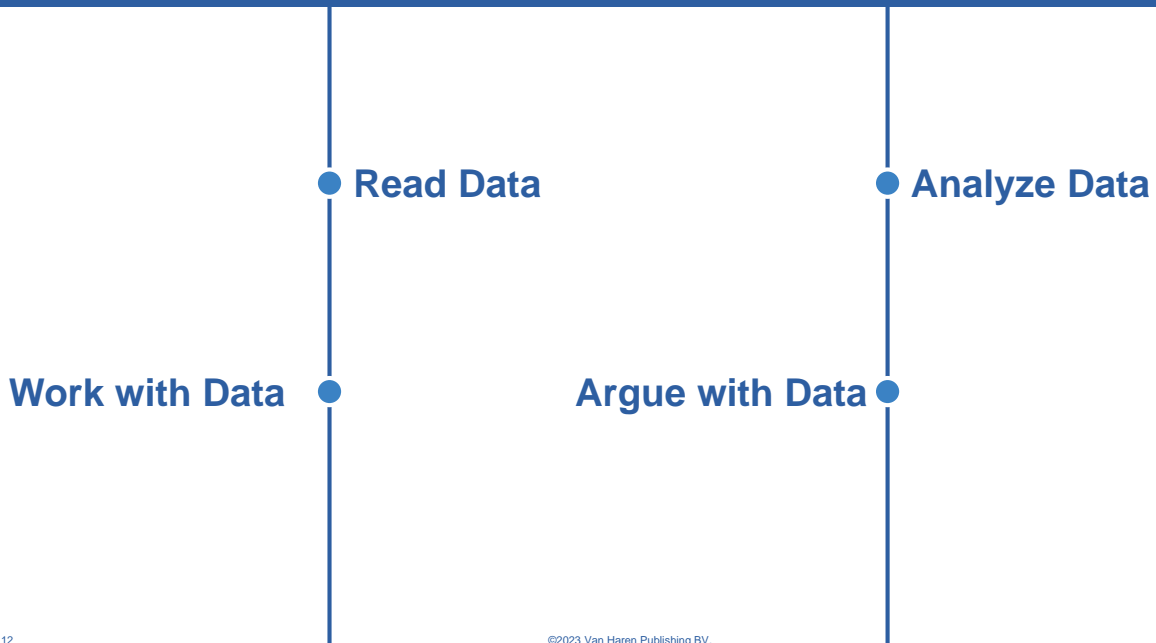


11

©2023 Van Haren Publishing BV.



Training Agenda



12

©2023 Van Haren Publishing BV.



Training Agenda



Read
data

What is Data?

Summarize Data

Consume Data

Check your Data

13

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

What is Data?



Data has a better idea

14

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Data is not Reality



15

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

What is Data?

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

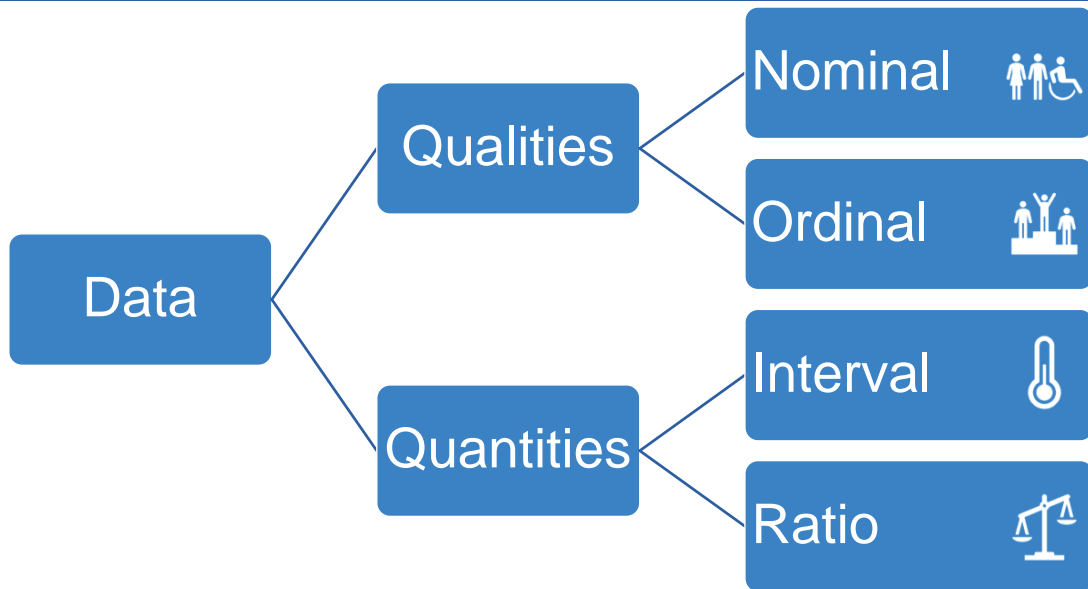
Source: unknown

16

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Quantity or Quality



17

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Types of scales

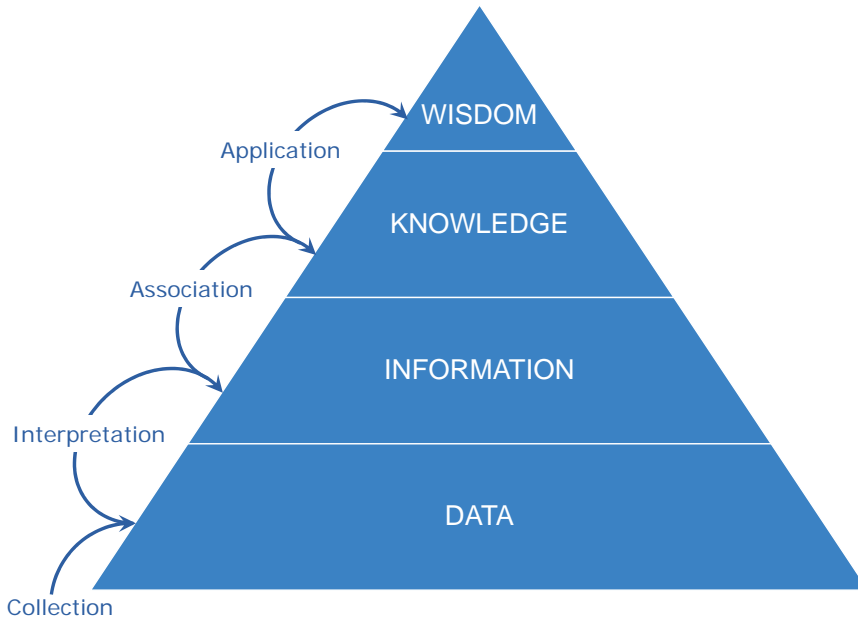


18

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Why do we use Data?

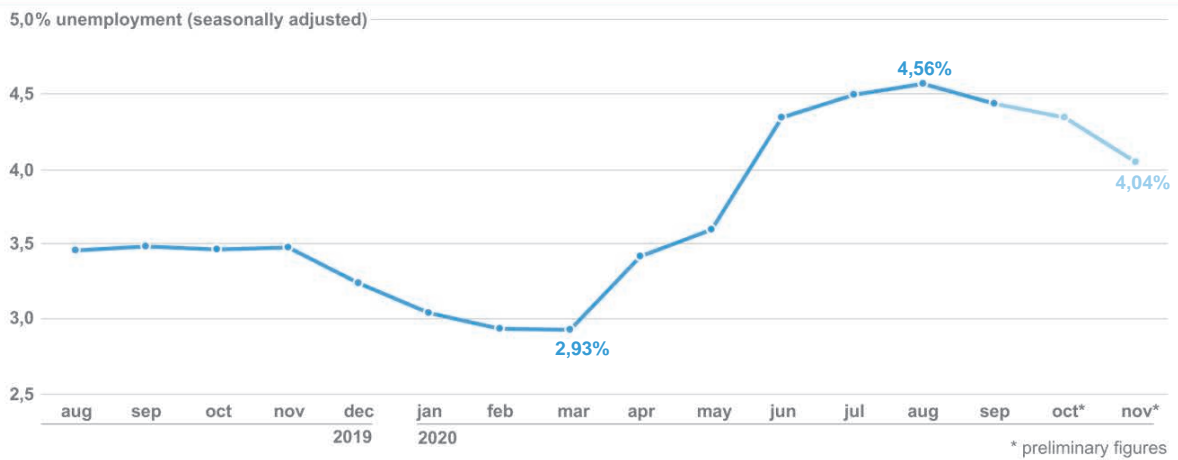


19 Source: Jones, Ben, Data Literacy Fundamentals: Understanding the Power & Value of Data

©2023 Van Haren Publishing BV.



Unemployment



$$\text{Unemployment rate} = \frac{\text{Unemployed labor force}}{\text{Labor force}}$$

20 Source: CBS – Statistics Netherlands

©2023 Van Haren Publishing BV.



Summarize data

*I didn't come here
to be average*

Source: Michael Jordan



21

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Summarize data



LeBron James



Lionel Messi



Tom Brady

 **6,690 + 3**
inhabitants

 **16.9 km²**
surface

 **18.5K\$**
average income

 **37.4K\$**
average income



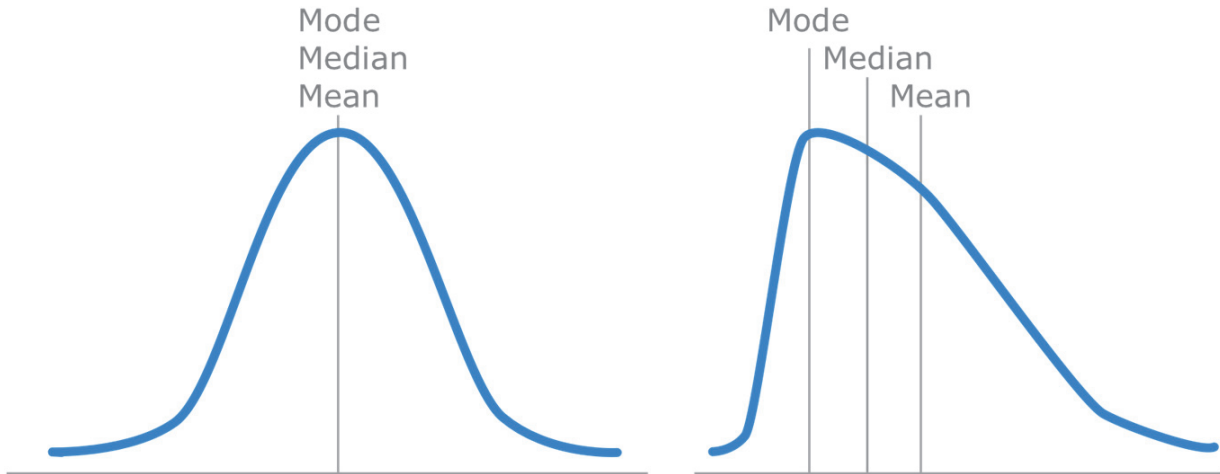
22

Source: https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income & <https://www.sportsc.com/personalities/athletes/2022/100-highest-paid-athletes-in-the-world-2022-full-list-1234674901/>

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Summarize data



23

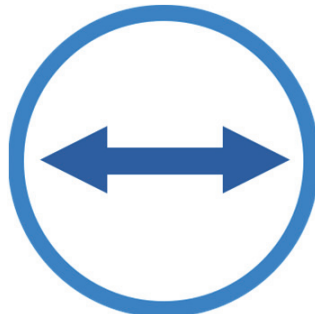
©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Summarize data



**Central
Tendency**



Spread



Shape

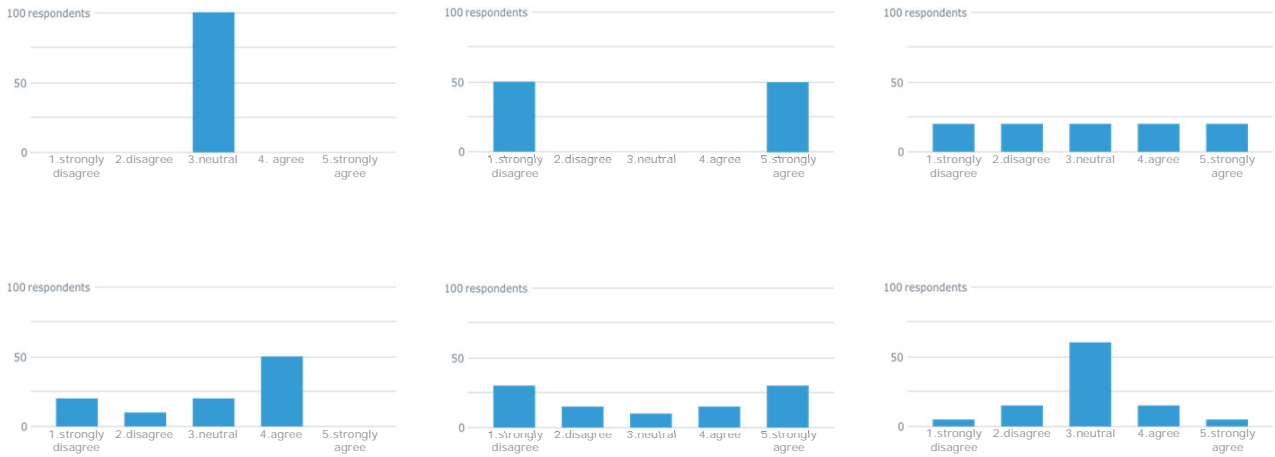
24 Source: Stephen S. Few, Now you see it, 2009

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Summarize data – Likert scale

Mean = 3.0

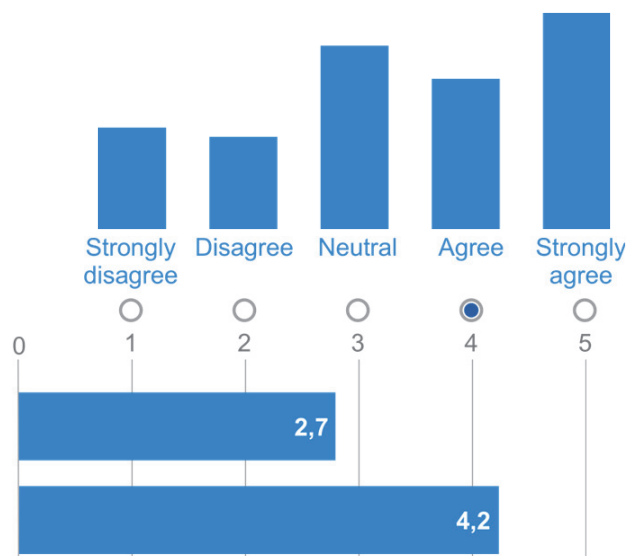


25

©2023 Van Haren Publishing BV.



Summarize data – Likert scale

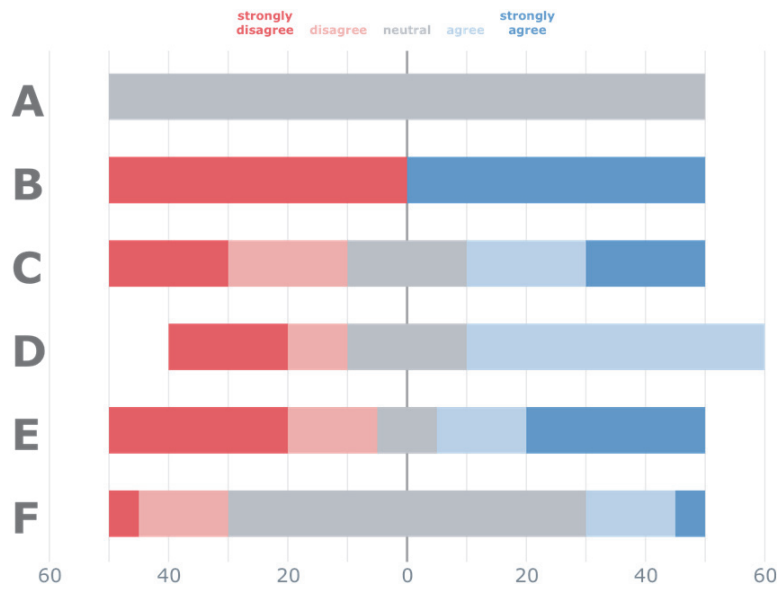


26

©2023 Van Haren Publishing BV.



Summarize data – Likert scale



27

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Consume data



28

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Consume data



"We used to use intuition; now we use analytics."

2014 IBM ad*

* and almost everyone else in BI/AI, at some point or another

Effective
DATA
Foundation

29

©2023 Van Haren Publishing BV.

Consume data



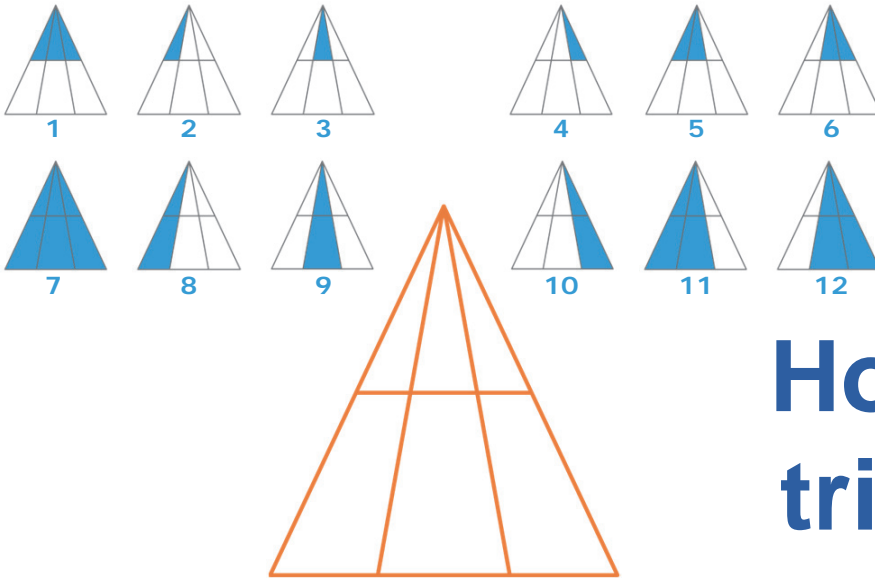
How does she feel?

Effective
DATA
Foundation

30 Photo by nappy on Pexels.com

©2023 Van Haren Publishing BV.

Consume data



How many triangles?

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

©2023 Van Haren Publishing BV.



Thinking, Fast and Slow

System 1

Fast

Unconscious

Automated

Everyday decisions

Error prone

System 2

Slow

Conscious

Effortful

Complex decisions

Reliable

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

©2023 Van Haren Publishing BV.



Consume Data

5'2"

Height Napoleon

~~157cm~~
Height Napoleon

168cm

Height Napoleon

165cm

Avg height



33 Source: James Gillray - 1805

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Consume data

**"We used to use
intuition;
then we used
analytics.**

Now we use both"

Ben Jones

34

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

BIG numbers



Every hour: **55 million** PET bottles sold in the world

35 Source: <https://graphics.reuters.com/ENVIRONMENT-PLASTIC/0100B275155/index.html>

©2023 Van Haren Publishing BV.



BIG numbers



1.3 billion
bottles every
day

36 Source: <https://graphics.reuters.com/ENVIRONMENT-PLASTIC/0100B275155/index.html>

©2023 Van Haren Publishing BV.



How solid is your opinion?

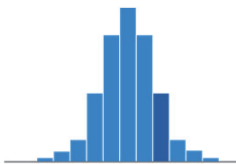
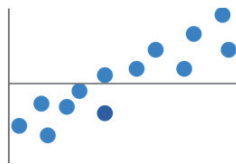
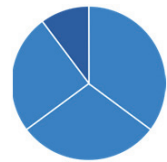


37 Source: Ben Jones

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Read carefully

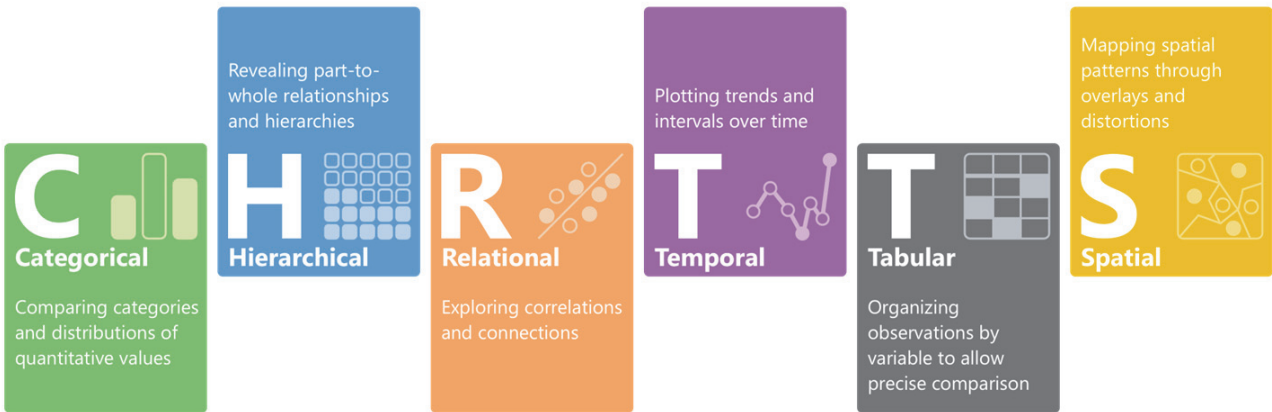


38 Source: Stephen S. Few, Show me the Numbers, 2004

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Read carefully

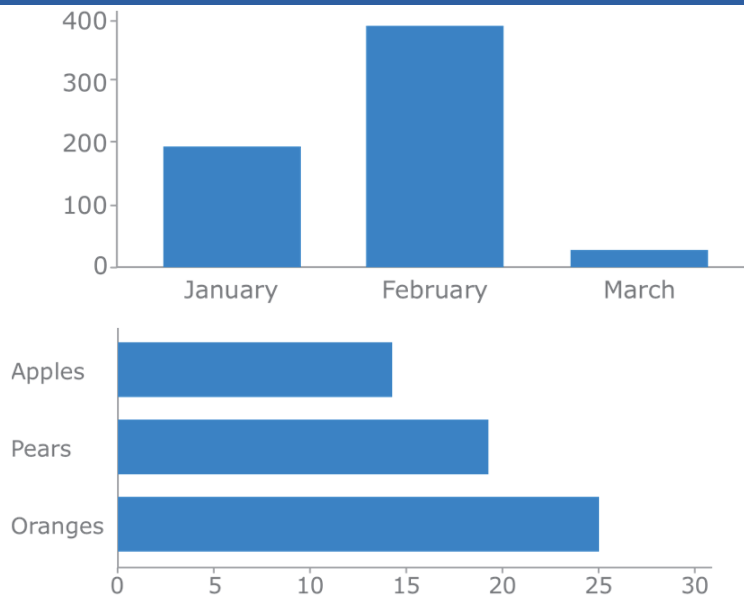


39 Based on: Data Visualization, Andy Kirk

©2023 Van Haren Publishing BV.



Read carefully



40

©2023 Van Haren Publishing BV.

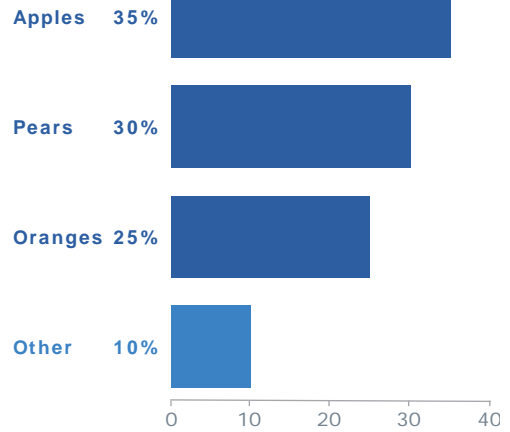
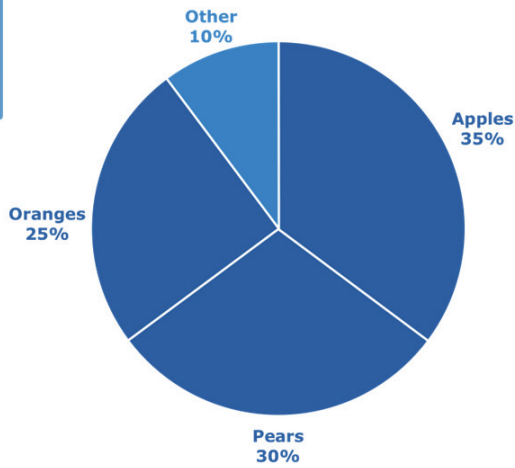


Read carefully

Revealing part-to-whole relationships and hierarchies



Hierarchical

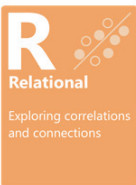


41

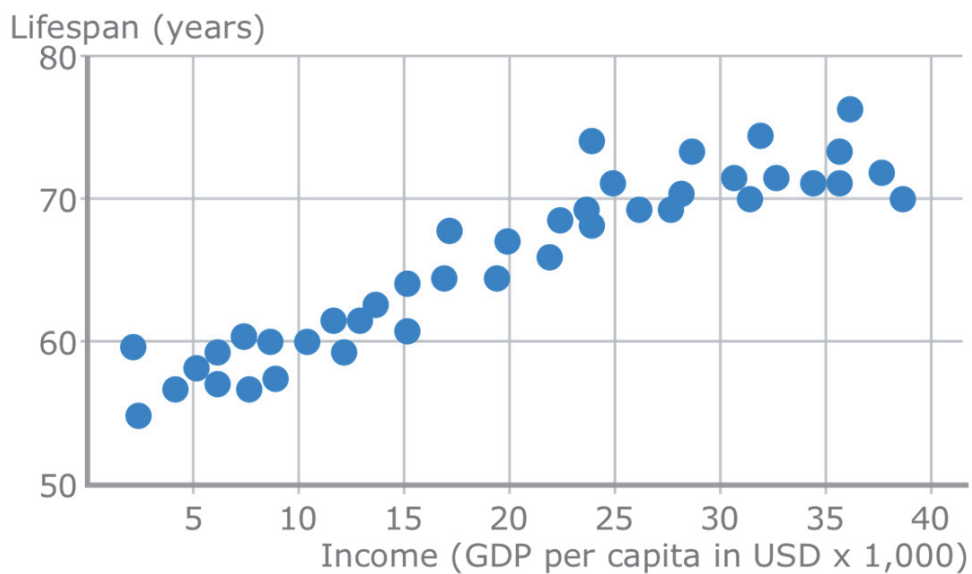
©2023 Van Haren Publishing BV.



Read carefully



Relational
Exploring correlations and connections



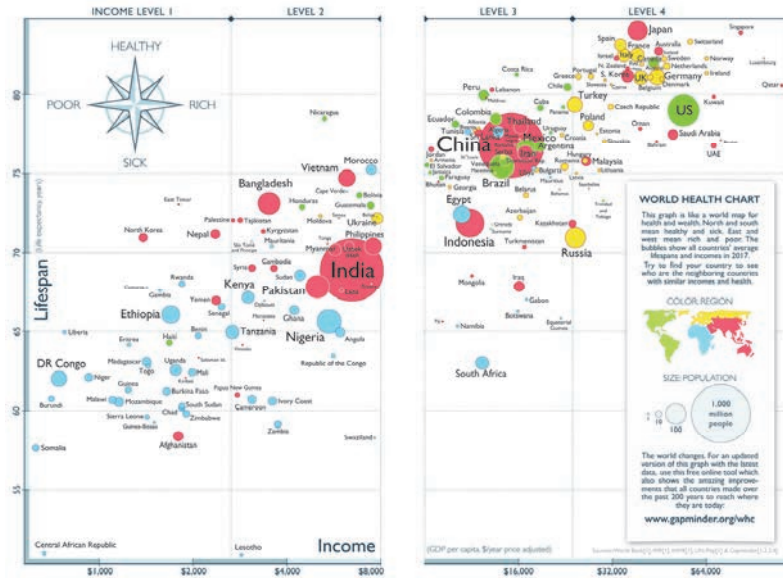
42

©2023 Van Haren Publishing BV.



Read carefully

R
Relational
Exploring correlations and connections



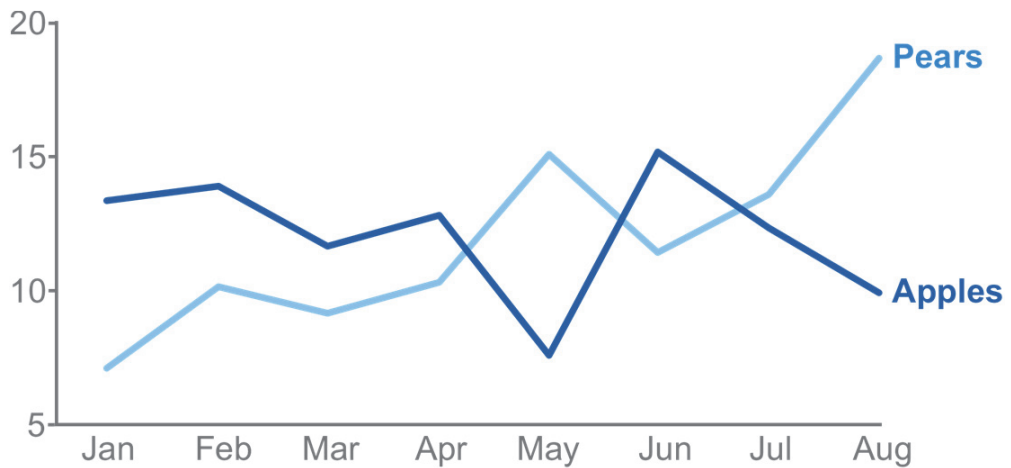
43 Source: Factfulness, Hans. Ola & Anna Rosling, 2018

©2023 Van Haren Publishing BV.



Read Carefully

Plotting trends and intervals over time
T
Temporal



44

©2023 Van Haren Publishing BV.



Read carefully

T
Tabular
 Organizing observations by variable to allow precise comparison

	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

45

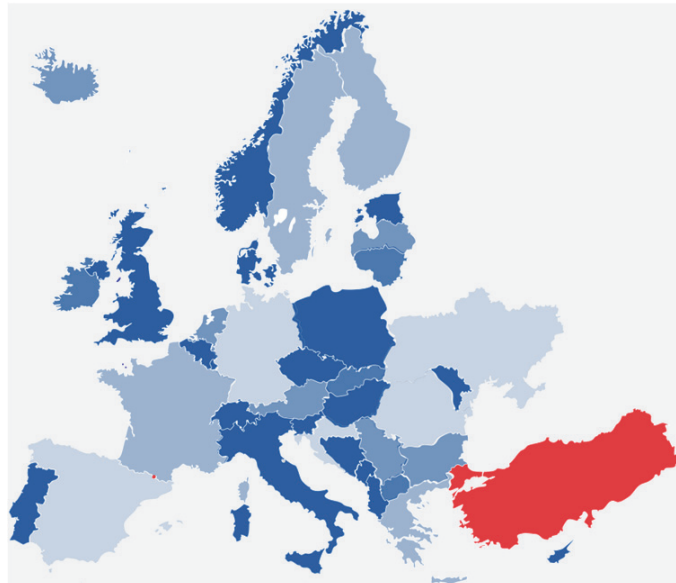
©2023 Van Haren Publishing BV.



Read carefully

Mapping spatial patterns through overlays and distortions

S
Spatial



46

©2023 Van Haren Publishing BV.

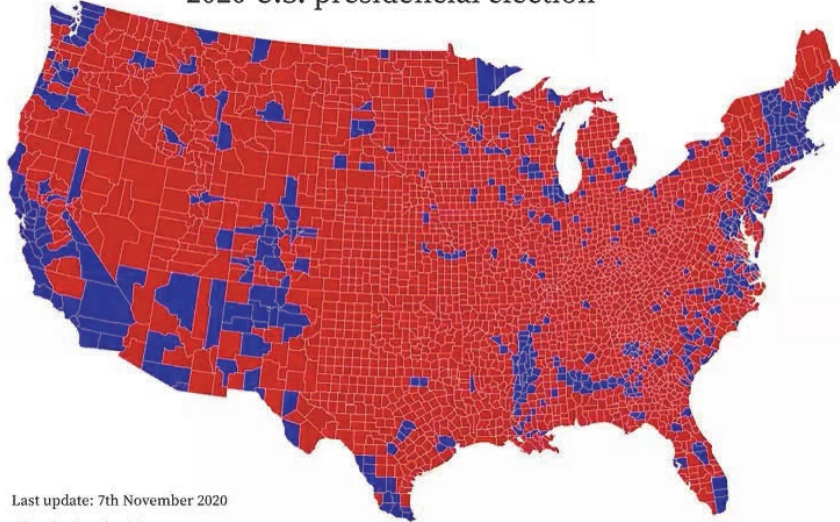


Read carefully

Mapping spatial patterns through overlays and distortions



2020 U.S. presidential election



Last update: 7th November 2020

@karim_douieb



47

©2023 Van Haren Publishing BV.

Check your Data: Data Claim Checklist





- 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?

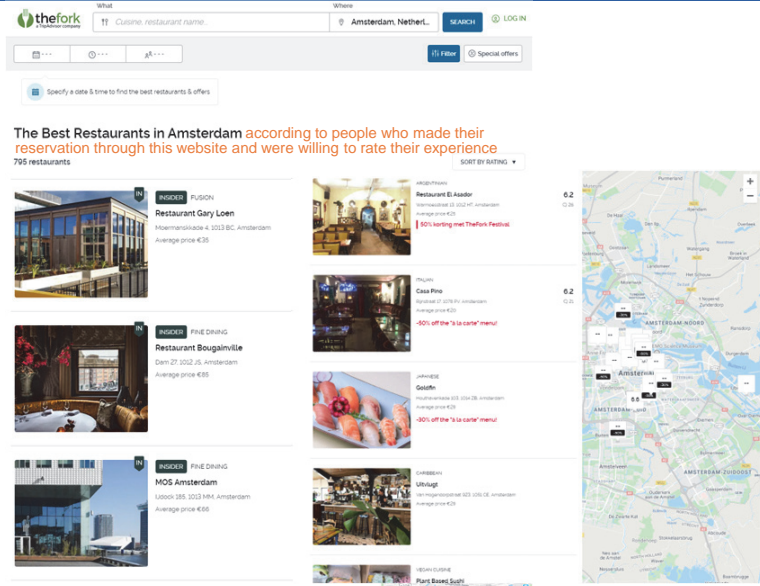


48 Photo by Glenn Carstens-Peters on Unsplash

©2023 Van Haren Publishing BV.

Check your Data: Data Claim Checklist





-  Why does the data matter to you?
-  Does the claim match the data?
-  Does the claim seem plausible?
-  What comparison needs to be made?




49

©2023 Van Haren Publishing BV.

Check your Data: Data Claim Checklist

-  Why does the data matter to you?
-  Does the claim match the data?
-  Does the claim seem plausible?
-  What comparison needs to be made?

☰ How is my rate calculated?

When you leave a review, you evaluate 3 criterias:

- The quality of the dishes which corresponds to 50% of the rate.
- The treatment and the service, which corresponde to 25% of the rate.
- The ambience of the restaurant, which corresponds to 25% of the rate.

Trending Art

How can I cancel

...TheFork uses the Bayesian method for calculating the average score...

Your opinion is important to us! Concentrate on your gastronomic experience and think what would you like to read about the restaurant when you book with **TheFork**.

How can I modify

Can I use my Yu

Information About The Reviews









About The Reviews

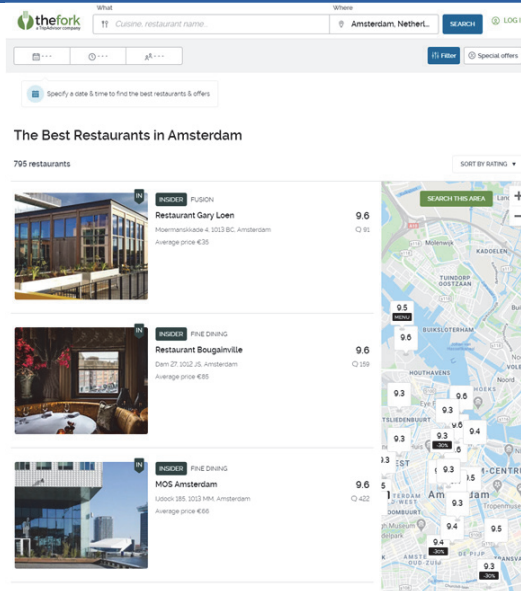


50

©2023 Van Haren Publishing BV.

Check your Data: Data Claim Checklist

-  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?



What Where

thefork
Cuisine: restaurant name: Amsterdam, Netherl. SEARCH LOGIN

Specify a date & time to find the best restaurants & offers

The Best Restaurants in Amsterdam

795 restaurants SORT BY RATING

Restaurant Name	Rating	Average Price
Restaurant Gary Loen	9.6	€35
Restaurant Bougainville	9.6	€35
MOS Amsterdam	9.6	€35

51

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Exercise 1

let's **PRACTICE**

52

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation


Let's check some data

 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?

EXPAT INFO CAREER HOUSING EDUCATION LIFESTYLE EXPAT SERVI

Dutch employment crisis: More job vacancies than unemployed

17 August 2021, by Victoria Séveno

The latest figures from **Statistics Netherlands (CBS)** expose the extent of the Dutch employment crisis, revealing that in the second quarter of 2021, **job vacancies in the Netherlands** outnumbered the number of **unemployed**.

Job vacancies and unemployment levels in the Netherlands

While many were left concerned about the impact of the **coronavirus pandemic** on the Dutch economy and labour market, figures from CBS confirm that the financial aid from the **Dutch government** not only prevented thousands of redundancies, but has actually led to another crisis altogether.

Unemployment has been falling and the number of vacancies has been steadily rising since mid-**2020**, but peaked this year between April and June, when there were 106 job vacancies for every 100 unemployed people. In this period, there were 327,000 jobs available - the highest number ever recorded in **the Netherlands**. At the end of the second quarter, there were 39 vacancies for every 1,000 jobs in the Dutch labour market - the highest figure ever recorded by CBS.

The jobs available in trade, business services, and the **Dutch healthcare sector** account for half of all vacancies in the Netherlands. The catering industry has also been severely affected by the pandemic, with the total number of vacancies in this sector doubling to 27,000 in the second quarter; at the end of June, there were 82 vacancies for every 1,000 jobs in this sector.

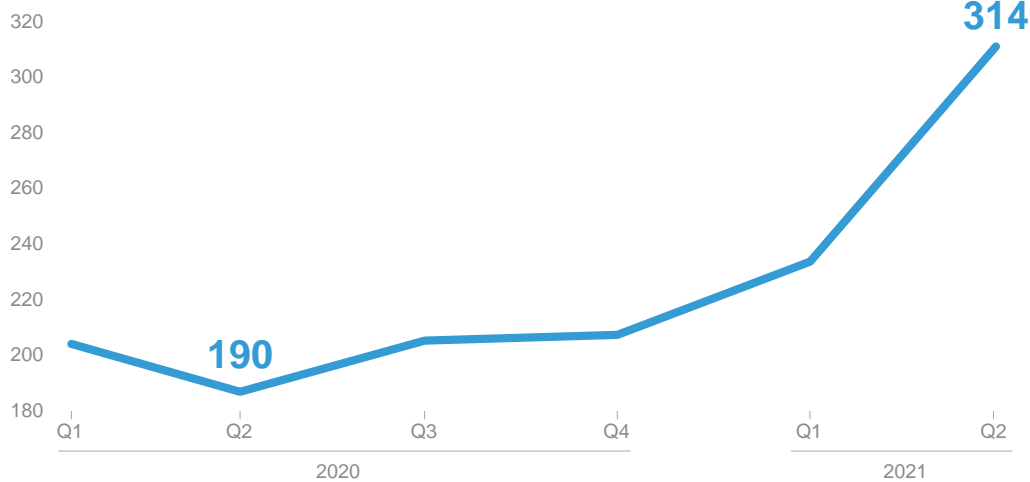
53 Source: I am Expat

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Let's check some data

340 vacancies (seasonally adjusted x 1.000)



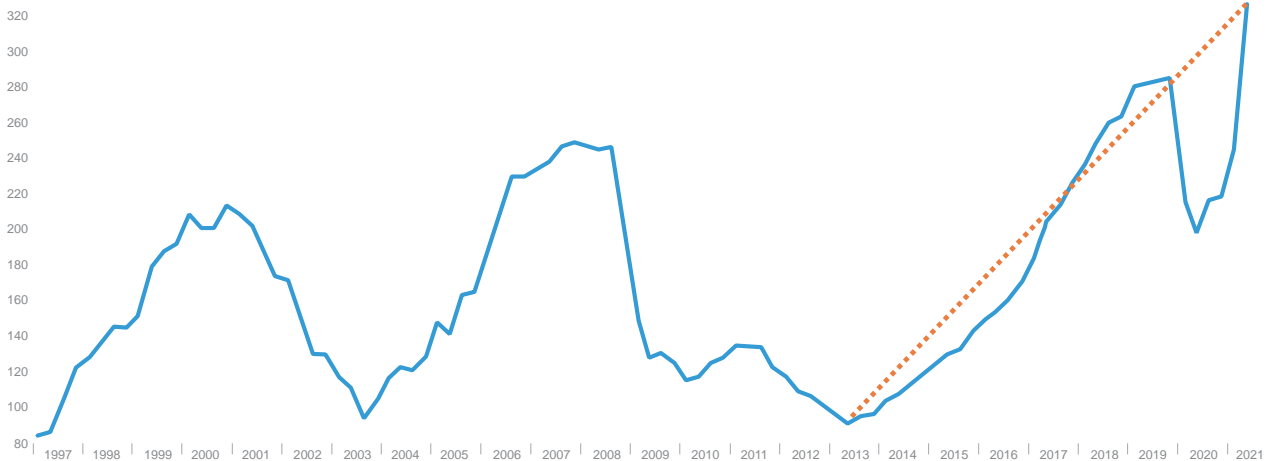
54 Source: CBS Statistics Netherlands

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Let's check some data

340 vacancies (seasonally adjusted x 1.000)



55 Source: CBS Statistics Netherlands

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Wrap up Read Data



Data is not reality



Aggregation comes at a price



No Pain no Gain



Read your data with care



Check data arguments

56

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

3 C's of Data Literacy



Curiosity



Creativity



Critical Thinking

57 Source: Jordan Morrow

©2023 Van Haren Publishing BV.



Training Agenda



**Work with
data**

Creating Data

Data Quality

Acquiring & Cleaning

Managing

58

©2023 Van Haren Publishing BV.



Creating data



59 Photo by Simon Kadula & Johan Mouchet on Unsplash

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Creating: Automated or Manual

60

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Creating data - manual



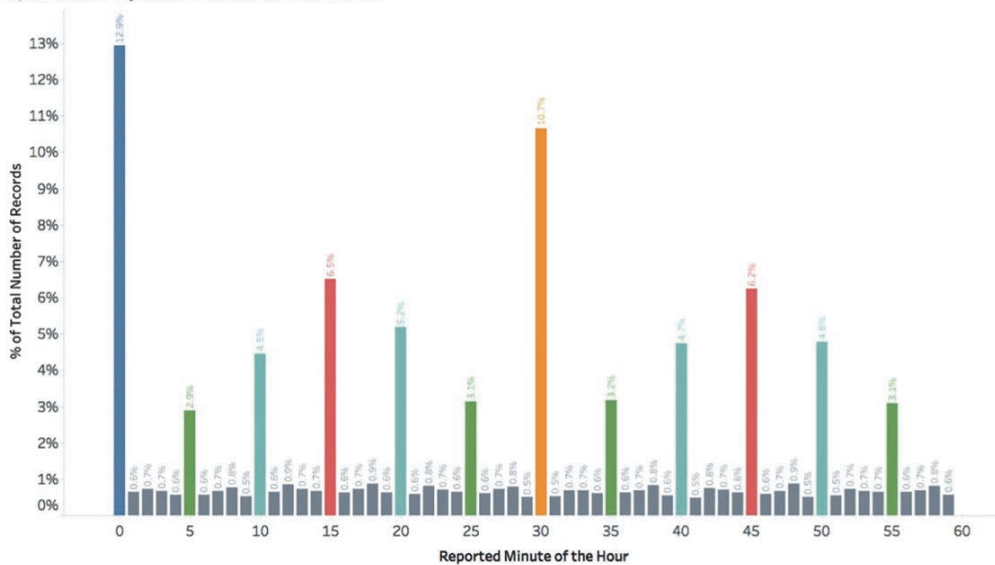
61 Photo: Mike Focus

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Creating data - manual

Reported strikes by minute of the hour, non-null values

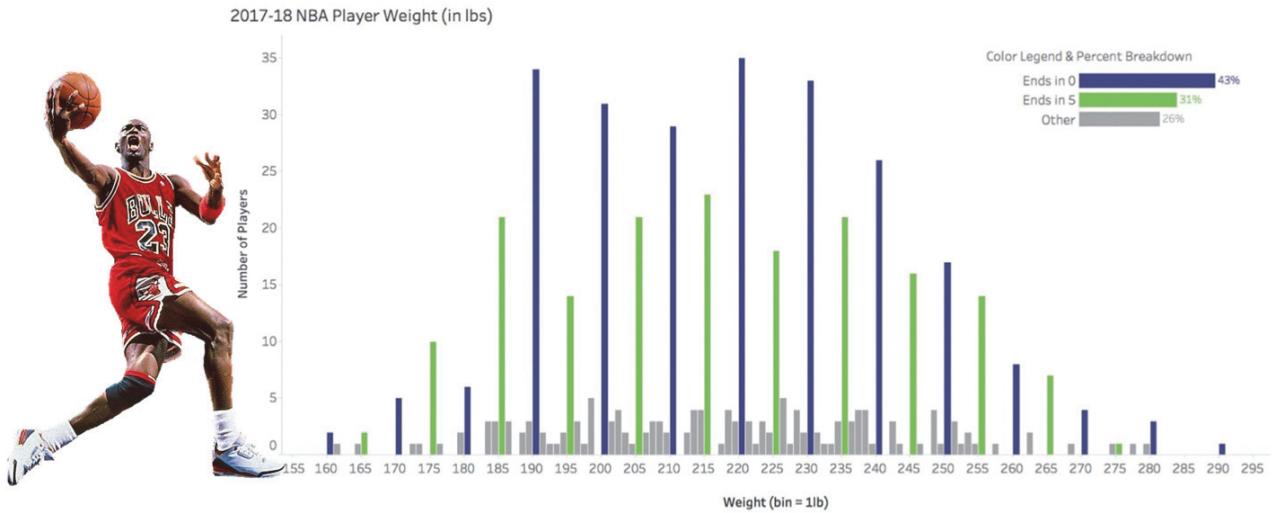


62 Source: Jones, Ben. Avoiding Data Pitfalls

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Creating data - manual

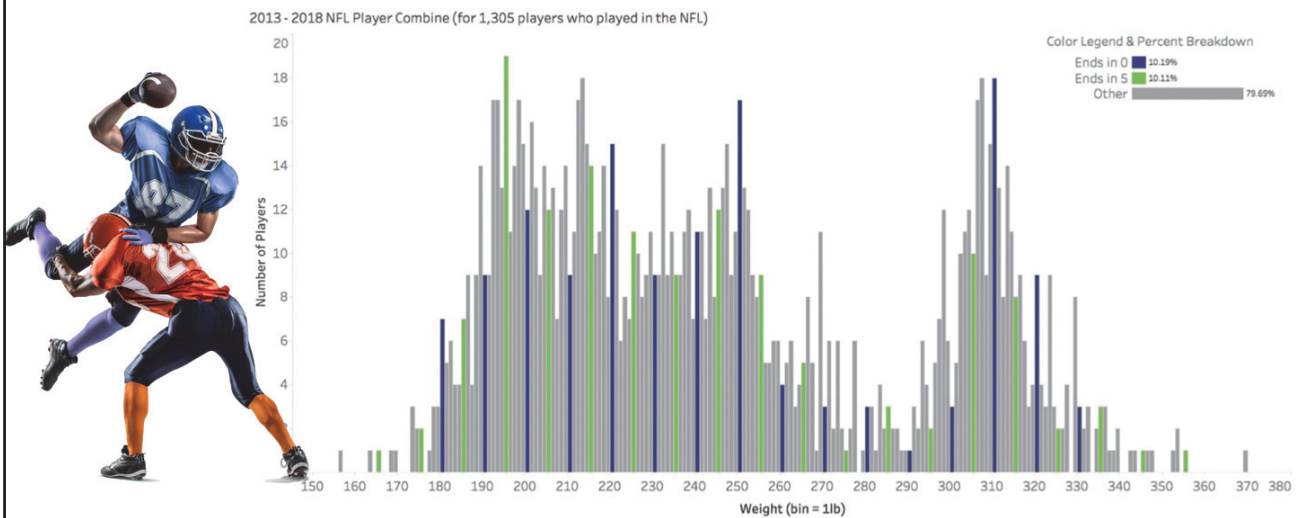


63 Source: Jones, Ben. Avoiding Data Pitfalls

©2023 Van Haren Publishing BV.



Creating data - manual



64 Source: Jones, Ben. Avoiding Data Pitfalls

©2023 Van Haren Publishing BV.



Creating data - manual



65 Photo: <https://hero-health.org/wp-content/uploads/2019/02/optional-mandatory-crop.jpg>

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Data Quality



66 Source: Data Driven, Thomas C. Redman

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Data Quality

Data is of high **quality**, if the data correctly **represents the real-world** construct that the data describes

67 Source: https://en.wikipedia.org/wiki/Data_quality

©2023 Van Haren Publishing BV.



Data Quality

-  **Completeness**
-  **Timeliness**
-  **Uniqueness**
-  **Consistency**
-  **Accuracy**

68

©2023 Van Haren Publishing BV.



Data Quality: avoid confusing data with reality



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



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



Understand the **limitations** and inaccuracies of each step in the process.



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

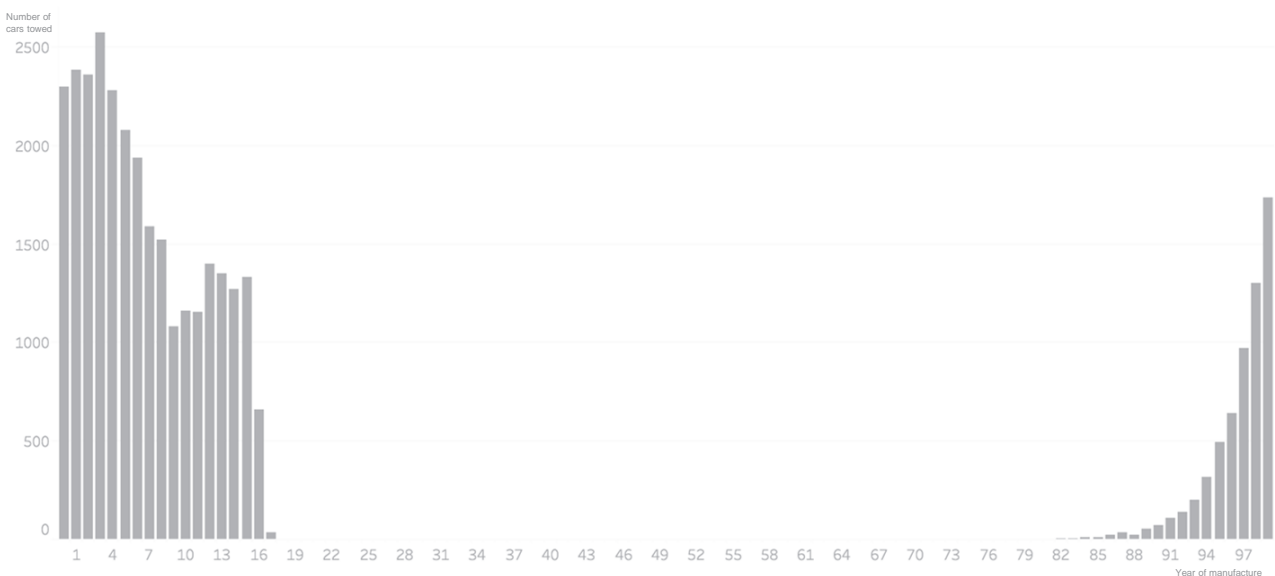


Seek to **understand** the **motives** of the people collecting and reporting. (biases or incentives)



Visualize the data and investigate any shifts, outliers, and trends for possible **discrepancies**.

Data Quality



Data Quality



71 Source: Jones, Ben. Avoiding Data Pitfalls

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Acquire & Clean

Data is like garbage. You'd better know what you are going to do with it before you collect it.

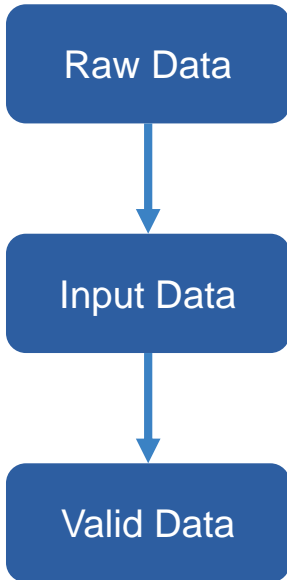
Source: Mark Twain

72

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Acquire & Clean



Make the data "technical" readable

- File Conversion
- String normalization
- Numeric normalization
- Date normalization

Make the data valid

- Missing values
- Implausible values
- Implausible combinations
- Restructure

73 Source: Mark van der Loo & Edwin de Jonge CBS – Statistics Netherlands

©2023 Van Haren Publishing BV.



Acquire & Clean

*Tidy datasets
are all alike, but
every messy
dataset is messy
in its own way.*



74 Source: Tidy Data – Hadley Wickham 2014

©2023 Van Haren Publishing BV.



Tidy data

	Canada	Germany	Australia	South Africa
Marketing	151	493	317	
Operations	959		326	826
Purchasing	528	805	699	425
Sales	381	470	931	90
Finance		459	228	

	Marketing	Operations	Purchasing	Sales	Finance
Canada	151	959	528	381	
Germany	493		805	470	459
Australia	317	326	699	931	228
South Africa		826	425	90	

Dataset is a collection of **values**
Values are either **numbers** or **strings**

75

©2023 Van Haren Publishing BV.



Tidy data

Every **value** belongs to a **variable** and an **observation**

Variable: contains all **values** that measure the **same attribute** across units

Observation: contains all **values** measures on the **same unit** across attributes

	Marketing	Operations	Purchasing	Sales	Finance
Canada	151	959	528	381	
Germany	493		805	470	459
Australia	317	326	699	931	228
South Africa		826	425	90	

76

©2023 Van Haren Publishing BV.



Tidy data

Observations

Variables

Country	Business Line	Items sold
Canada	Marketing	151
Germany	Marketing	493
Australia	Marketing	317
South Africa	Marketing	
Canada	Operations	959
Germany	Operations	
Australia	Operations	326
South Africa	Operations	826
Canada	Purchasing	528
Germany	Purchasing	805
Australia	Purchasing	699
South Africa	Purchasing	425
Canada	Sales	381
Germany	Sales	470
Australia	Sales	931
South Africa	Sales	90
Canada	Finance	
Germany	Finance	459
Australia	Finance	228
South Africa	Finance	

77

©2023 Van Haren Publishing BV.



Tidy data

Five most common problems with messy datasets



Column **headers are values**, not variable names



Multiple variables are stored in **one column**



Variables are stored in both **rows and columns**



Multiple types of observations are stored in the **same table**



A **single** observation is stored in **multiple tables**

78 Source: Tidy Data – Hadley Wickham 2014

©2023 Van Haren Publishing BV.



Exercise 2

let's PRACTICE

Tidy data

Which common data structure problems apply to this set?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
1	StartDate	EndDate	Duration	Finished	ResponseId	Latitude	Longitude	Lang	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	
2	12-2-2020 12:15	12-2-2020 12:16	55	True	R_PLEB7VLC8E5C1	52.3182068	4.951004028	NL	Ripe	Ripe	Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Unripe	Ripe	
3	12-2-2020 12:15	12-2-2020 12:17	95	True	R_2zNCZjpr8qA11W	52.3426056	4.863098145	NL	Very Ripe	Very Ripe	Ripe	Override	Override	Unripe	Ripe	Ripe	Almost Ripe	Very Ripe	Very Ripe
4	12-2-2020 12:16	12-2-2020 12:17	79	True	R_2E4zthzYRR8RWju	52.3426056	4.863098145	NL	Very Ripe	Ripe	Ripe	Override	Override	Unripe	Ripe	Ripe	Unripe	Very Ripe	
5	12-2-2020 12:16	12-2-2020 12:17	60	True	R_3eEccEzLiLyNkkj	52.0218048	4.700393677	NL	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Almost Ripe	Ripe	
6	12-2-2020 12:16	12-2-2020 12:17	105	True	R_1hGph1vC8LbM8mZ	52.3294067	4.873901367	EN-GB	Ripe	Almost Ripe	Unripe	Very Ripe	Override	Unripe	Unripe	Unripe	Unripe	Almost Ripe	
7	12-2-2020 12:16	12-2-2020 12:17	67	True	R_2701CelJiNMygmO	52.3426056	4.863098145	NL	Very Ripe	Very Ripe	Ripe	Override	Override	Unripe	Ripe	Ripe	Almost Ripe	Very Ripe	
8	12-2-2020 12:17	12-2-2020 12:18	57	True	R_Y7D4ngotIB8tA1	52.3182068	4.951004028	NL	Very Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Unripe	Ripe	
9	12-2-2020 12:16	12-2-2020 12:18	88	True	R_10rimsSuyQjyGba	51.5189056	-0.0921936	EN-GB	Ripe	Ripe	Almost Ripe	Override	Override	Almost Ripe	Ripe	Ripe	Almost Ripe	Very Ripe	
10	12-2-2020 12:17	12-2-2020 12:18	66	True	R_3qdMfSquQlczWak	52.3294067	4.873901367	EN-GB	Very Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Unripe	Ripe	
11	12-2-2020 12:16	12-2-2020 12:18	112	True	R_KpnHicKBRQ8AY9	52.3426056	4.863098145	NL	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Almost Ripe	Very Ripe	
12	12-2-2020 12:17	12-2-2020 12:18	76	True	R_22D1pDVUJaiH29	52.3426056	4.863098145	NL	Very Ripe	Almost Ripe	Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Ripe	Unripe	
13	12-2-2020 12:17	12-2-2020 12:18	76	True	R_brtVZ4ubxL0Gcx	52.3294067	4.873901367	NL	Very Ripe	Very Ripe	Ripe	Override	Override	Unripe	Ripe	Ripe	Unripe	Very Ripe	
14	12-2-2020 12:16	12-2-2020 12:18	104	True	R_29GKcpOTYmV4HYO	52.3294067	4.873901367	EN-GB	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Almost Ripe	Ripe	Unripe	Ripe	
15	12-2-2020 12:17	12-2-2020 12:20	138	True	R_bjCVCepG24n4FH	52.3182068	4.951004028	NL	Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Almost Ripe	Unripe	Ripe	
16	12-2-2020 12:17	12-2-2020 12:20	111	True	R_3hGyyNnh9crVj	52.3811951	5.248092651	NL	Override	Very Ripe	Ripe	Override	Override	Unripe	Ripe	Ripe	Almost Ripe	Very Ripe	
17	12-2-2020 12:19	12-2-2020 12:20	104	True	R_11gVtRFsW3YbQc	52.3426056	4.863098145	NL	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Almost Ripe	Unripe	Ripe	
18	12-2-2020 12:19	12-2-2020 12:20	83	True	R_2WOhYh07FWV1mf	52.366394	4.849594116	EN-GB	Very Ripe	Unripe	Unripe	Very Ripe	Override	Very Ripe	Unripe	Unripe	Unripe	Unripe	
19	12-2-2020 12:20	12-2-2020 12:22	102	True	R_3f428uasaalQ4Mq	52.3824005	4.899500615	NL	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Almost Ripe	Ripe	
20	12-2-2020 12:19	12-2-2020 12:22	188	True	R_ujFG5V5w1Xaqjwl	52.7207031	4.734298706	NL	Very Ripe	Ripe	Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Almost Ripe	Ripe	
21	12-2-2020 12:19	12-2-2020 12:23	274	True	R_25AhjRejz4d3Ghe	52.3426056	4.863098145	EN-GB	Ripe	Almost Ripe	Unripe	Very Ripe	Override	Unripe	Almost Ripe	Ripe	Unripe	Ripe	
22	12-2-2020 12:22	12-2-2020 12:24	70	True	R_2VkeMRKkTX7dGgh	52.3182068	4.951004028	NL	Ripe	Ripe	Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Unripe	Ripe	
23	12-2-2020 12:23	12-2-2020 12:24	60	True	R_p0lyfIcc5f8pYf	52.3426056	4.863098145	NL	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Almost Ripe	Ripe	
24	12-2-2020 12:23	12-2-2020 12:25	82	True	R_20NjQ4cLlOuP0eoh	52.3182068	4.951004028	EN-GB	Ripe	Very Ripe	Ripe	Override	Override	Unripe	Almost Ripe	Ripe	Almost Ripe	Ripe	
25	12-2-2020 12:23	12-2-2020 12:25	139	True	R_3qfSeVmcIqgRE3f	52.366394	4.849594116	EN-GB	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Almost Ripe	Unripe	Ripe	
26	12-2-2020 12:25	12-2-2020 12:26	62	True	R_2mtoHLBz4HAB	52.3426056	4.863098145	NL	Ripe	Almost Ripe	Unripe	Override	Override	Unripe	Ripe	Ripe	Unripe	Ripe	
27	12-2-2020 12:23	12-2-2020 12:26	176	True	R_110QZ7L1KjyruZO	52.3182068	4.951004028	EN-GB	Almost Ripe	Almost Ripe	Unripe	Very Ripe	Override	Unripe	Ripe	Ripe	Unripe	Ripe	
28	12-2-2020 12:25	12-2-2020 12:26	87	True	R_3et8sU2Z2DokZK	52.3426056	4.863098145	NL	Very Ripe	Ripe	Ripe	Very Ripe	Override	Unripe	Almost Ripe	Ripe	Ripe	Almost Ripe	
29	12-2-2020 12:25	12-2-2020 12:26	79	True	R_WqFcuLsaiLg1z1	52.3426056	4.863098145	NL	Very Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Almost Ripe	Unripe	Very Ripe	
30	12-2-2020 12:24	12-2-2020 12:27	125	True	R_12KQcX8BwREqLz	52.3426056	4.863098145	NL	Ripe	Almost Ripe	Almost Ripe	Very Ripe	Very Ripe	Unripe	Almost Ripe	Ripe	Almost Ripe	Ripe	
31	12-2-2020 12:26	12-2-2020 12:27	82	True	R_3DqeVmh2QsJUHINY	52.3426056	4.863098145	NL	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Almost Ripe	Ripe	
32	12-2-2020 12:25	12-2-2020 12:27	120	True	R_1rv7okYuhMvmSc	52.3462982	4.820800781	NL	Ripe	Almost Ripe	Unripe	Very Ripe	Override	Unripe	Almost Ripe	Ripe	Unripe	Ripe	
33	12-2-2020 12:26	12-2-2020 12:28	78	True	R_3kMp52pn8J2lcnN	52.3182068	4.951004028	NL	Very Ripe	Ripe	Ripe	Very Ripe	Override	Almost Ripe	Ripe	Ripe	Almost Ripe	Ripe	

Tidy data

How would this set look like in a Tidy format?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
1	StartDate	EndDate	Duration	Finished	Responded	Latitude	Longitude	Lang	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	
2	12-2-2020 12:15	12-2-2020 12:16	55	True	R_PLEB7Vlc8E5C1	52,3182068	4,951004028	NL	Ripe	Ripe	Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Ripe	Unripe	Ripe
3	12-2-2020 12:15	12-2-2020 12:17	95	True	R_2zNCZjpr8Aii1W	52,3426056	4,863098145	NL	Very Ripe	Very Ripe	Ripe	Override	Override	Unripe	Ripe	Ripe	Ripe	Almost Ripe	Very Ripe
4	12-2-2020 12:16	12-2-2020 12:17	79	True	R_2E4zhzYRDBRwju	52,3426056	4,863098145	NL	Very Ripe	Ripe	Ripe	Override	Override	Unripe	Ripe	Ripe	Ripe	Unripe	Very Ripe
5	12-2-2020 12:16	12-2-2020 12:17	60	True	R_3eIFcEzLilyNkxj	52,0218048	4,700393677	NL	Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Ripe	Almost Ripe	Ripe
6	12-2-2020 12:16	12-2-2020 12:17	105	True	R_1hGph1vC8LbMBmZ	52,3294067	4,873901367	EN-GB	Ripe	Almost Ripe	Unripe	Very Ripe	Override	Unripe	Unripe	Unripe	Unripe	Unripe	Almost Ripe
7	12-2-2020 12:17	12-2-2020 12:18	67	True	R_27O1CeJjJNypmO	52,3426056	4,863098145	NL	Very Ripe	Very Ripe	Ripe	Override	Override	Unripe	Ripe	Ripe	Ripe	Almost Ripe	Very Ripe
8	12-2-2020 12:17	12-2-2020 12:18	57	True	R_yL7D4ngoTt8IBAt	52,3182068	4,951004028	NL	Very Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Ripe	Unripe	Ripe
9	12-2-2020 12:17	12-2-2020 12:18	88	True	R_10rimSouyQjGba	51,5189056	-0,0921936	EN-GB	Ripe	Ripe	Almost Ripe	Override	Override	Almost Ripe	Ripe	Ripe	Ripe	Almost Ripe	Very Ripe
10	12-2-2020 12:17	12-2-2020 12:18	66	True	R_3qdMfaqQlozWak	52,3294067	4,873901367	EN-GB	Very Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Ripe	Unripe	Ripe
11	12-2-2020 12:16	12-2-2020 12:18	112	True	R_XpnHzkBRQR8AY9	52,3426056	4,863098145	NL	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Ripe	Almost Ripe	Very Ripe
12	12-2-2020 12:17	12-2-2020 12:18	76	True	R_2ZD1pDvVUaiH29	52,3426056	4,863098145	NL	Very Ripe	Almost Ripe	Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Ripe	Unripe	Ripe
13	12-2-2020 12:17	12-2-2020 12:18	76	True	R_brtY24ubxblOGcx	52,3294067	4,873901367	NL	Very Ripe	Very Ripe	Ripe	Override	Override	Unripe	Ripe	Ripe	Ripe	Unripe	Very Ripe
14	12-2-2020 12:16	12-2-2020 12:18	104	True	R_296Kpc07YmYAHYD	52,3294067	4,873901367	EN-GB	Very Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Almost Ripe	Unripe	Ripe
15	12-2-2020 12:17	12-2-2020 12:20	138	True	R_bICVCCpG24n4FH	52,3182068	4,951004028	NL	Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Almost Ripe	Unripe	Ripe
16	12-2-2020 12:18	12-2-2020 12:20	111	True	R_3hGyynNhh9crVj	52,3811951	5,248092651	NL	Override	Very Ripe	Ripe	Override	Override	Unripe	Ripe	Ripe	Almost Ripe	Unripe	Very Ripe
17	12-2-2020 12:19	12-2-2020 12:20	104	True	R_11gvRF5vW3YbQc	52,3426056	4,863098145	NL	Very Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Almost Ripe	Unripe	Ripe
18	12-2-2020 12:19	12-2-2020 12:20	83	True	R_2W0rYm07FWVImf	52,366394	4,849594116	EN-GB	Very Ripe	Unripe	Unripe	Very Ripe	Very Ripe	Unripe	Unripe	Unripe	Unripe	Unripe	Unripe
19	12-2-2020 12:20	12-2-2020 12:22	102	True	R_23fPvVlOgQAN4	52,3426056	4,863098145	EN-GB	Very Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Ripe	Almost Ripe	Ripe
20	12-2-2020 12:20	12-2-2020 12:22	102	True	R_3F428uasaQ4MQ	52,3824005	4,899505615	NL	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Ripe	Unripe	Very Ripe
21	12-2-2020 12:19	12-2-2020 12:22	188	True	R_ujFG5V5w1Xaqjwl	52,7207031	4,734298706	NL	Very Ripe	Ripe	Ripe	Very Ripe	Very Ripe	Unripe	Ripe	Ripe	Ripe	Almost Ripe	Ripe
22	12-2-2020 12:19	12-2-2020 12:23	274	True	R_25AhjReir2d3GhE	52,3426056	4,863098145	EN-GB	Ripe	Almost Ripe	Unripe	Very Ripe	Override	Unripe	Almost Ripe	Ripe	Ripe	Unripe	Ripe
23	12-2-2020 12:22	12-2-2020 12:24	70	True	R_2VkeMRXkTX7dGgh	52,3182068	4,951004028	NL	Ripe	Ripe	Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Ripe	Unripe	Ripe
24	12-2-2020 12:23	12-2-2020 12:24	60	True	R_p0lyfKz5f8jXT	52,3426056	4,863098145	NL	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Ripe	Almost Ripe	Ripe
25	12-2-2020 12:23	12-2-2020 12:25	82	True	R_3qJmQ4cLUuPoeh	52,3182068	4,951004028	EN-GB	Very Ripe	Ripe	Ripe	Very Ripe	Override	Unripe	Almost Ripe	Ripe	Ripe	Almost Ripe	Ripe
26	12-2-2020 12:23	12-2-2020 12:25	139	True	R_3qrF5eYmchPE3F	52,366394	4,849594116	EN-GB	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Almost Ripe	Unripe	Ripe
27	12-2-2020 12:25	12-2-2020 12:26	62	True	R_2mntoHtLBrzIHAB	52,3426056	4,863098145	NL	Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Ripe	Unripe	Ripe
28	12-2-2020 12:23	12-2-2020 12:26	176	True	R_110Q71KUYryeUO	52,3182068	4,951004028	EN-GB	Almost Ripe	Almost Ripe	Unripe	Very Ripe	Override	Unripe	Ripe	Ripe	Ripe	Unripe	Ripe
29	12-2-2020 12:25	12-2-2020 12:26	87	True	R_3et8sU2X2DoKZX	52,3426056	4,863098145	NL	Very Ripe	Ripe	Ripe	Very Ripe	Override	Unripe	Almost Ripe	Ripe	Ripe	Almost Ripe	Ripe
30	12-2-2020 12:25	12-2-2020 12:26	79	True	R_WqfFcUcjalig1	52,3426056	4,863098145	NL	Very Ripe	Ripe	Almost Ripe	Very Ripe	Override	Unripe	Ripe	Ripe	Almost Ripe	Unripe	Very Ripe
31	12-2-2020 12:24	12-2-2020 12:27	125	True	R_12KQXb0eREGqLz	52,3426056	4,863098145	NL	Ripe	Almost Ripe	Unripe	Very Ripe	Very Ripe	Unripe	Ripe	Almost Ripe	Ripe	Almost Ripe	Ripe
32	12-2-2020 12:26	12-2-2020 12:27	82	True	R_3Dqvm2ZcUJmHNY	52,3426056	4,863098145	NL	Very Ripe	Ripe	Almost Ripe	Override	Override	Unripe	Ripe	Ripe	Ripe	Almost Ripe	Ripe
33	12-2-2020 12:25	12-2-2020 12:27	120	True	R_1rv7oKvUyMvmSc	52,3462982	4,820800781	NL	Ripe	Almost Ripe	Unripe	Very Ripe	Override	Unripe	Almost Ripe	Ripe	Ripe	Unripe	Ripe
34	12-2-2020 12:26	12-2-2020 12:28	78	True	R_3kMpsZpr8i2GcN	52,3182068	4,951004028	NL	Very Ripe	Ripe	Ripe	Very Ripe	Override	Almost Ripe	Ripe	Ripe	Ripe	Almost Ripe	Ripe



Tidy data

How would this set look like in a Tidy format?

Surveys

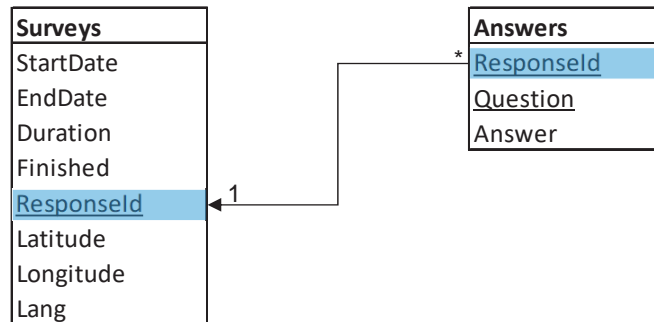
StartDate	EndDate	Duration	Finished	Responded	Latitude	Longitude	Lang
12-2-2020 12:15	12-2-2020 12:16	55	True	R_PLEB7Vlc8E5C1	52,3182068	4,951004028	NL
12-2-2020 12:15	12-2-2020 12:17	95	True	R_2zNCZjpr8Aii1W	52,3426056	4,863098145	NL
12-2-2020 12:16	12-2-2020 12:17	79	True	R_2E4zhzYRDBRwju	52,3426056	4,863098145	NL
12-2-2020 12:16	12-2-2020 12:17	60	True	R_3eIFcEzLilyNkxj	52,0218048	4,700393677	NL
12-2-2020 12:16	12-2-2020 12:17	105	True	R_1hGph1vC8LbMBmZ	52,3294067	4,873901367	EN-GB
12-2-2020 12:16	12-2-2020 12:17	67	True	R_27O1CeJjJNypmO	52,3426056	4,863098145	NL
12-2-2020 12:17	12-2-2020 12:18	57	True	R_yL7D4ngoTt8IBAt	52,3182068	4,951004028	NL
12-2-2020 12:16	12-2-2020 12:18	88	True	R_10rimSouyQjGba	51,5189056	-0,0921936	EN-GB
12-2-2020 12:17	12-2-2020 12:18	66	True	R_3qdMfaqQlozWak	52,3294067	4,873901367	EN-GB
12-2-2020 12:16	12-2-2020 12:18	112	True	R_XpnHzkBRQR8AY9	52,3426056	4,863098145	NL
12-2-2020 12:17	12-2-2020 12:18	76	True	R_2ZD1pDvVUaiH29	52,3426056	4,863098145	NL
12-2-2020 12:17	12-2-2020 12:18	76	True	R_brtY24ubxblOGcx	52,3294067	4,873901367	NL
12-2-2020 12:16	12-2-2020 12:18	104	True	R_296Kpc07YmYAHYD	52,3294067	4,873901367	EN-GB
12-2-2020 12:17	12-2-2020 12:20	138	True	R_bICVCCpG24n4FH	52,3182068	4,951004028	NL
12-2-2020 12:18	12-2-2020 12:20	111	True	R_3hGyynNhh9crVj	52,3811951	5,248092651	NL
12-2-2020 12:19	12-2-2020 12:20	104	True	R_11gvRF5vW3YbQc	52,3426056	4,863098145	NL
12-2-2020 12:19	12-2-2020 12:20	83	True	R_2W0rYm07FWVImf	52,366394	4,849594116	EN-GB
12-2-2020 12:20	12-2-2020 12:22	102	True	R_23fPvVlOgQAN4	52,3426056	4,863098145	EN-GB
12-2-2020 12:20	12-2-2020 12:22	102	True	R_3F428uasaQ4MQ	52,3824005	4,899505615	NL
12-2-2020 12:19	12-2-2020 12:22	188	True	R_ujFG5V5w1Xaqjwl	52,7207031	4,734298706	NL
12-2-2020 12:19	12-2-2020 12:23	274	True	R_25AhjReir2d3GhE	52,3426056	4,863098145	EN-GB
12-2-2020 12:22	12-2-2020 12:24	70	True	R_2VkeMRXkTX7dGgh	52,3182068	4,951004028	NL
12-2-2020 12:23	12-2-2020 12:25	82	True	R_p0lyfKz5f8jXT	52,3426056	4,863098145	NL
12-2-2020 12:23	12-2-2020 12:25	82	True	R_3qJmQ4cLUuPoeh	52,3182068	4,951004028	EN-GB
12-2-2020 12:23	12-2-2020 12:25	139	True	R_3qrF5eYmchPE3F	52,366394	4,849594116	EN-GB
12-2-2020 12:25	12-2-2020 12:26	62	True	R_2mntoHtLBrzIHAB	52,3426056	4,863098145	NL
12-2-2020 12:23	12-2-2020 12:26	176	True	R_110Q71KUYryeUO	52,3182068	4,951004028	EN-GB
12-2-2020 12:25	12-2-2020 12:26	87	True	R_3et8sU2X2DoKZX	52,3426056	4,863098145	NL
12-2-2020 12:25	12-2-2020 12:26	79	True	R_WqfFcUcjalig1	52,3426056	4,863098145	NL
12-2-2020 12:24	12-2-2020 12:27	125	True	R_12KQXb0eREGqLz	52,3426056	4,863098145	NL
12-2-2020 12:26	12-2-2020 12:27	82	True	R_3Dqvm2ZcUJmHNY	52,3426056	4,863098145	NL
12-2-2020 12:25	12-2-2020 12:27	120	True	R_1rv7oKvUyMvmSc	52,3462982	4,820800781	NL
12-2-2020 12:26	12-2-2020 12:28	78	True	R_3kMpsZpr8i2GcN	52,3182068	4,951004028	NL
12-2-2020 12:27	12-2-2020 12:29	141	True	R_3kBLDu7GCKGNFZ	52,3426056	4,863098145	NL

Answers

Responded	Question	Answer
R_PLEB7Vlc8E5C1	Q1	Ripe
R_PLEB7Vlc8E5C1	Q2	Ripe
R_PLEB7Vlc8E5C1	Q3	Ripe
R_PLEB7Vlc8E5C1	Q4	Very Ripe

Tidy data

How would this set look like in a Tidy format?

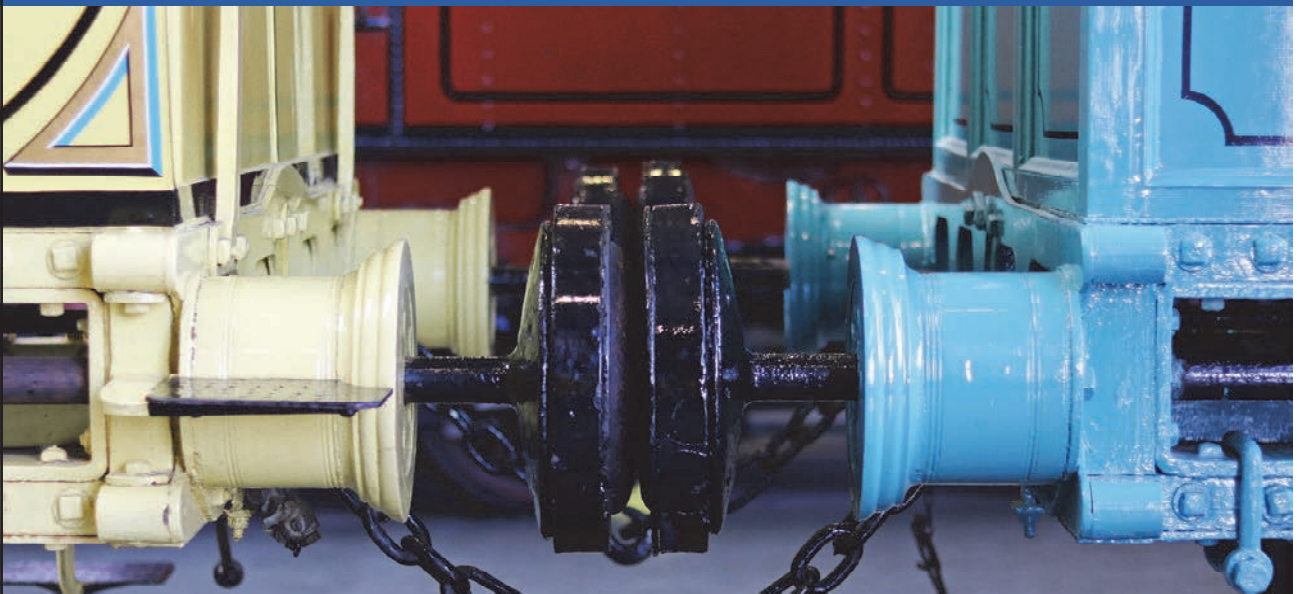


83

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Combine data



84

©2023 Van Haren Publishing BV.

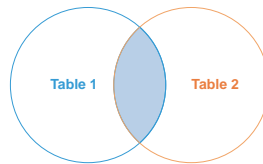
Effective
DATA
Foundation

Combine Data

Emp_ID	Emp_FirstName	Emp_LastName
1	Andy	Khan
2	Josh	Olsson
3	Suki	Noris
4	Nadine	Smith
5	Felicity	Torrance

HW_ID	Emp_ID	Highway	Towed
70	4	Interstate 70	2
95	3	Interstate 95	8
97	5	Interstate 97	4
170	6	Interstate 170	7
195	3	Interstate 195	6

27

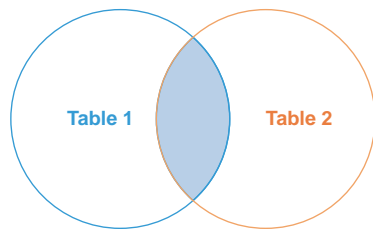


Emp_ID	Emp_FirstName	Emp_LastName	HW_ID	Highway	Towed
3	Suki	Noris	95	Interstate 95	8
3	Suki	Noris	195	Interstate 195	6
4	Nadine	Smith	70	Interstate 70	2
5	Felicity	Torrance	97	Interstate 97	4

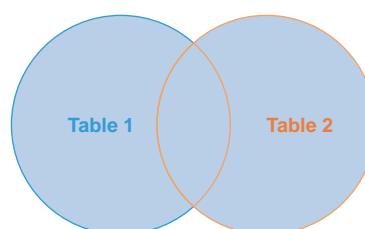
20

Combine data

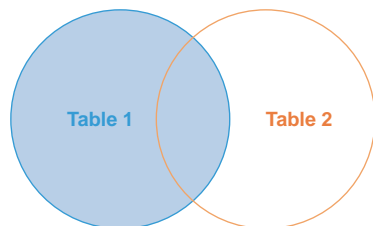
INNER JOIN



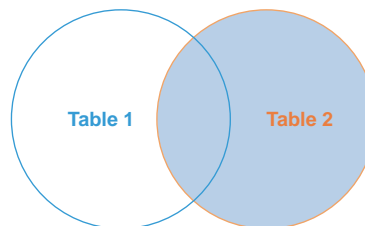
OUTER JOIN



LEFT JOIN



RIGHT JOIN



Combine data

table 1

1		
2		

inner join

1				
---	--	--	--	--

outer join

1				
2				
3				
4				

union

1		
2		
1		
3		
4		

table 2

1		
3		
4		

left join

1				
2				

right join

1				
3				
4				

87

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Exercise 3

let's PRACTICE

88

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Combine data

We have a dataset with all suppliers from The Netherlands.
Another dataset contains all supplied products from suppliers world wide.

How can we combine the two datasets to address the following questions:

- Which suppliers from The Netherlands have supplied us products?
- What is the total value of all products supplied by Dutch suppliers?
- How many suppliers from The Netherlands have actually supplied us?

Suppliers from NL

Suppl_ID	Name	City
S1	Janssen	Amersfoort
S2	Johnsson	Diemen
S3	Petersen	Assen
S4	Karelsen	Maastricht
S5	Severijns	Leiden

Supplied products

Prod_ID	Suppl_ID	Value
P1	S2	200.000
P2	S3	360.000
P3	S125	180.000
P4	S274	450.000
P5	S5	230.000

89

©2023 Van Haren Publishing BV.



Combine data

We have a dataset with all suppliers from The Netherlands.
Another dataset contains all supplied products from suppliers world wide.

How can we combine the two datasets to address the following questions:

- How much is the average supplied value delivered by suppliers from NL?
- How many suppliers have not delivered to us yet? Expressed as a percentage of total of suppliers from NL.

Suppliers from NL

Suppl_ID	Name	City
S1	Janssen	Amersfoort
S2	Johnsson	Diemen
S3	Petersen	Assen
S4	Karelsen	Maastricht
S5	Severijns	Leiden

Supplied products

Prod_ID	Suppl_ID	Value
P1	S2	200.000
P2	S3	360.000
P3	S125	180.000
P4	S274	450.000
P5	S5	230.000

90

©2023 Van Haren Publishing BV.



Combine data

We have a dataset with all suppliers from The Netherlands.
Another dataset contains all suppliers we prefer to avoid (Deny list).

How can we combine the two datasets to address the following question:

- Which suppliers are NOT on the Deny list?

Suppliers from NL

Suppl_ID	Name	City
S1	Janssen	Amersfoort
S2	Johnsson	Diemen
S3	Petersen	Assen
S4	Karelsen	Maastricht
S5	Severijns	Leiden

Deny list

BL_ID	Suppl_ID
B1	S2
B2	S32
B3	S4
B4	S48
B5	S55

91

©2023 Van Haren Publishing BV.



Combine data

We have a dataset with all suppliers as registered by department A.
Another dataset contains all suppliers as registered by department B.

How can we combine the two datasets to address the following questions:

- Which suppliers are registered by both departments?
- How many suppliers (part of whole) are registered by both, only A and only B?

Suppliers dep. A

Suppl_ID	Name
S1	Janssen
S2	Johnsson
S3	Petersen
S4	Karelsen
S5	Severijns

Suppliers dep. B

Suppl_ID	Name
S1	Janssen
S2	Johnsson
S5	Severijns
S6	Pietersen
S7	Koninks

92

©2023 Van Haren Publishing BV.



Combine data

Our organisation has two departments: A and B.
We have a dataset with all suppliers of department A.
Another dataset contains all suppliers of department B.

How can we combine the two datasets to address the following question:
- Who are all our suppliers on a company level?

Suppliers dep. A

Suppl_ID	Name
S1	Janssen
S2	Johnsson
S3	Petersen
S4	Karelsen
S5	Severijns

Suppliers dep. B

Suppl_ID	Name
S1	Janssen
S2	Johnsson
S5	Severijns
S6	Pietersen
S7	Koninks

93

©2023 Van Haren Publishing BV.



Managing data

*Data management is about
**people, processes,
and technology, in that
order.***

Source: unknown

94

©2023 Van Haren Publishing BV.



Managing data

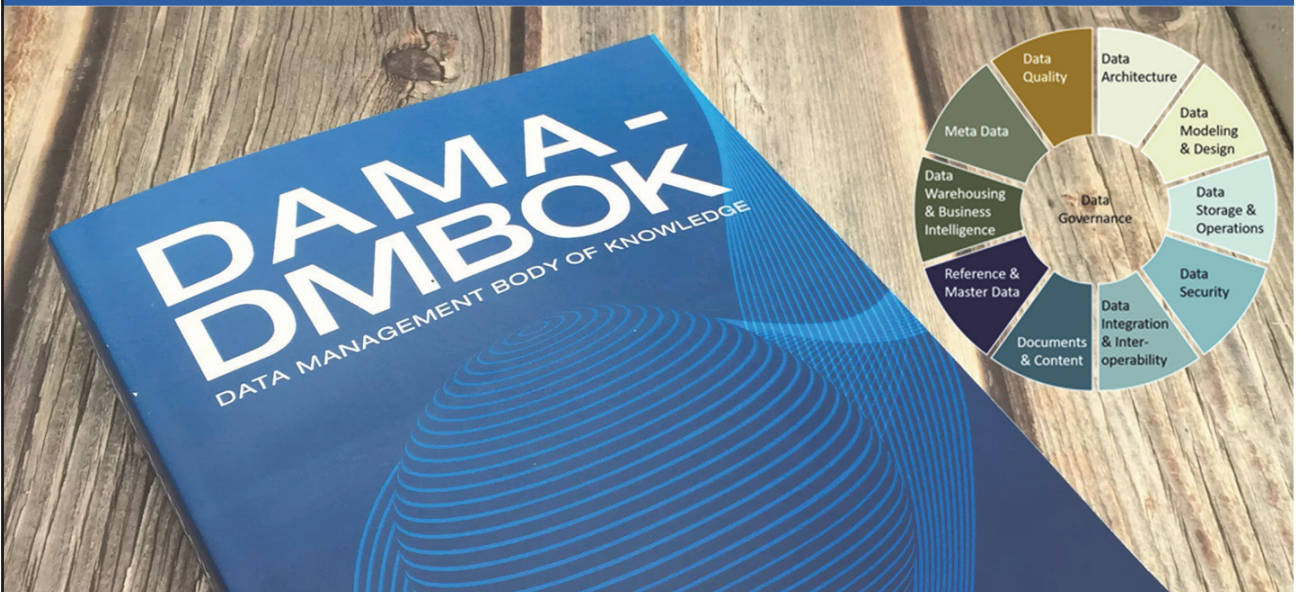


95

©2023 Van Haren Publishing BV.



Managing data



96

©2023 Van Haren Publishing BV.



KPI Mantras

KPI

A way to measure
how something or
someone is doing

97 Source: Coen de Bruijn, Key Performance Illusions

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

KPI Mantras



 Success loves **speed**

 **Patterns**, not points

98 Source: Stacey Barr, PuMP

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

KPI Mantras: PATTERNS, not points

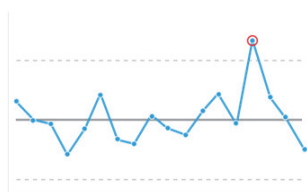


99 Source: Stacey Barr, PuMP

©2023 Van Haren Publishing BV.

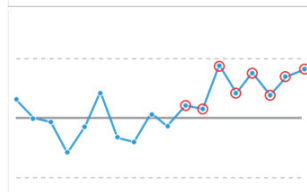


KPI Mantras: PATTERNS, not points



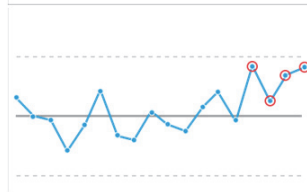
Special case

Measure falls outside of the “control limits”



Long term

Eight points in a row on the same side of the central line



Short term

Three points (or 3 out of 4) in a row closer to the control limit than to the central line

100 Source: Stacey Barr, PuMP

©2023 Van Haren Publishing BV.



KPI Mantras



 Success loves **speed**

 **Patterns**, not points

 It's a **process**, not an event



101 Source: Stacey Barr, PuMP

©2023 Van Haren Publishing BV.

KPI Mantras: It's a process, not an event

The PuMP® Performance Measure Blueprint



102 Source: Stacey Barr, PuMP

©2023 Van Haren Publishing BV.

KPI Mantras



 Success loves **speed**

 **Patterns**, not points

 It's a **process**, not an event

 **Buy-in**, not sign-off

 **Results** before measures

KPI Mantras: Results before measures

Begin with the result in mind (write it down):

Hotel employees are responsive to customer needs

List sensory evidence of the result:








- Customers comment on how friendly employees are
- When a customer wants something, employees get it for them quickly
- Whenever a customer walks past an employee, the employee greets them
- Employees anticipate customer needs, ready to act before the customer asks

Create potential measures:






Potential measure	Strength	Feasibility	Select?
1. Customer satisfaction rating for employee friendliness	L	M	
2. Cycle time to respond to customer needs	H	H	X
3. Number of customers greeted per employee per day	M	L	
4. Customer satisfaction rating for employee responsiveness	H	M	X
5. Customer compliments	L	L	

KPI Mantras



-  Success loves **speed**
-  **Patterns**, not points
-  It's a **process**, not an event
-  **Buy-in**, not sign-off
-  **Results** before measures
-  **Should**, **can** and **will**
-  **Practical**, not perfect

Wrap up Work with Data

-  Avoid the data-reality **gap**
-  **Tidy** your data
-  Be careful with **linking** data
-  **Manage** your data
-  Apply a **KPI** definition process

3 C's of Data Literacy



Curiosity



Creativity



Critical Thinking

Data literacy

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

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



Read
data



Work
with data



Analyze
data



Argue
with data

What is analysis?

a **detailed examination** of anything **complex** in order to **understand** its nature or to determine its **essential features**:

a thorough study

Merriam-Webster Dictionary

Training Agenda



Analyze
data

Expectations

Thinking shortcuts

Types of Analysis

Analytical skills

Expectations

2 4 8

111

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Expectations



What do I **expect** to see/find?



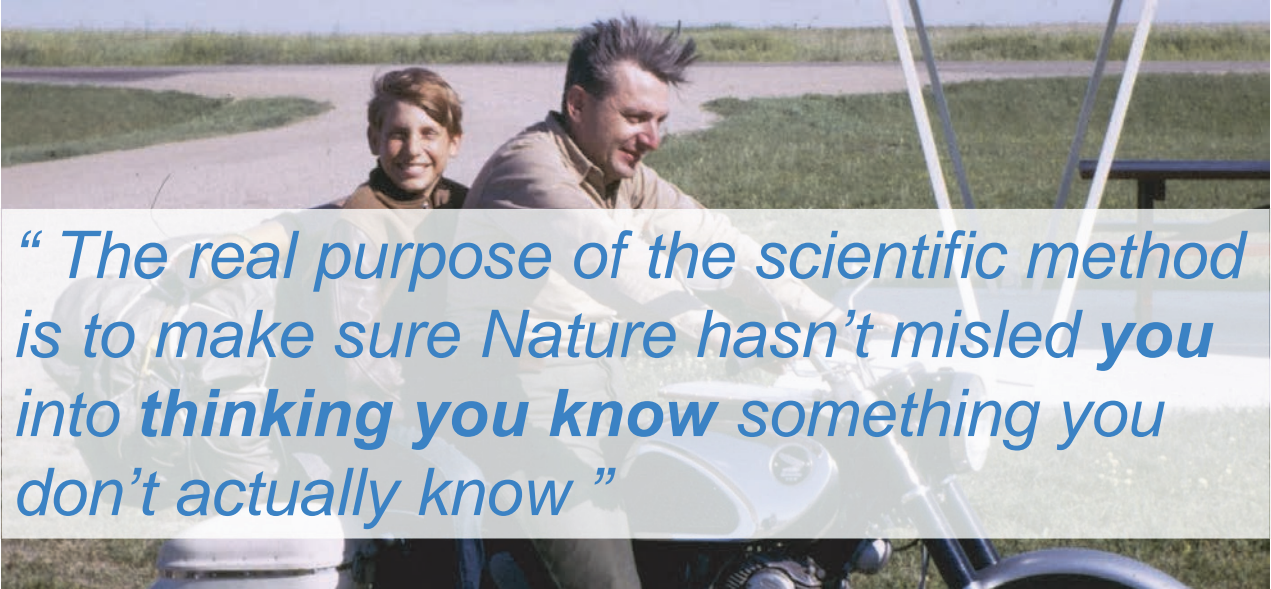
What would I **like** to see/find?

112

©2023 Van Haren Publishing BV.

Effective
DATA
Foundation

Expectations: Scientific Method

A photograph of a man and a young boy sitting on a motorcycle. The man is in the driver's seat, and the boy is sitting behind him. They are outdoors on a paved area with grass in the background.

“ The real purpose of the scientific method is to make sure Nature hasn’t misled **you** into **thinking you know** something you don’t actually know ”

Expectations: Scientific Method



What do I **expect** to see/find?

Translate your question into one or multiple *hypotheses*;

Hypothesis =
an educated guess at an answer to a question

Focus on: **dis**confirm the *hypothesis*

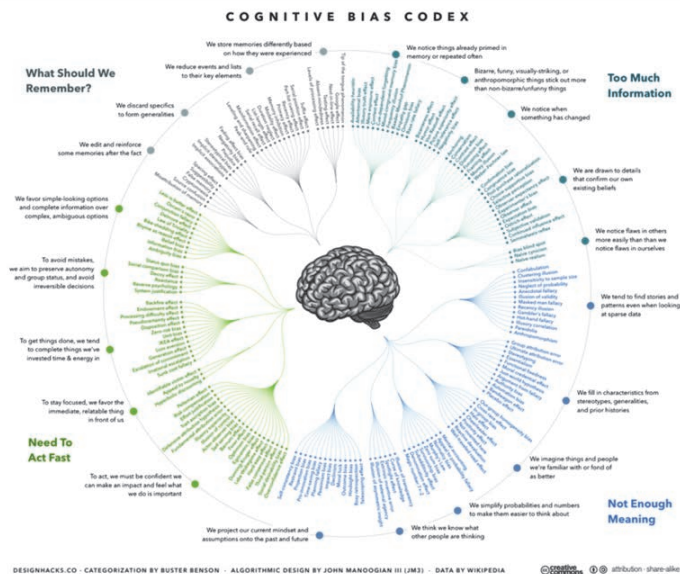
Shortcuts in thinking



115

©2023 Van Haren Publishing BV.

Shortcuts in thinking



116

©2023 Van Haren Publishing BV.

Thinking, Fast and Slow

System 1

Fast 

Unconscious 

Automated 

Everyday decisions 

Error prone 

System 2

Slow 

Conscious 

Effortful 

Complex decisions 

Reliable 

Shortcuts in thinking



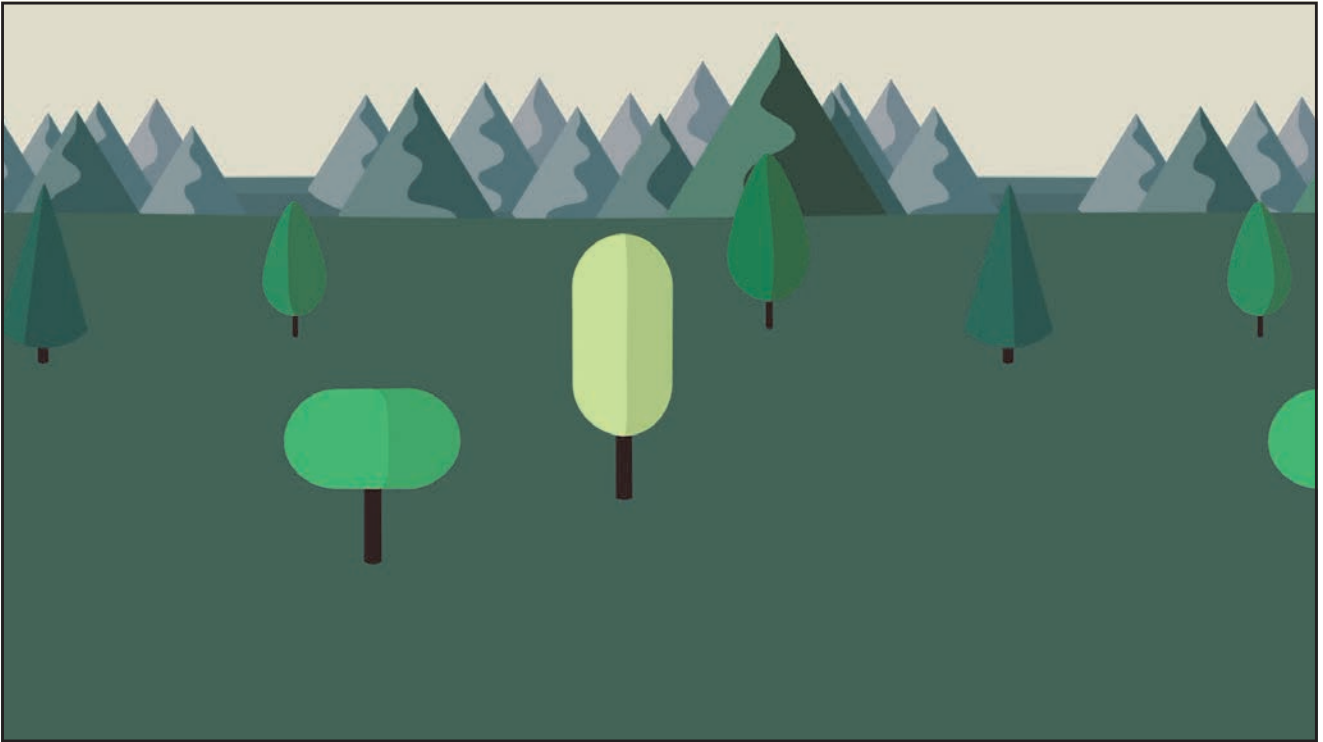
Cognitive Biases:

our systematic inclinations towards certain patterns of erroneous thinking (irrationality)



Logical Fallacies:

flaws in an argument that weakens the argument or makes the conclusion invalid



Confirmation Bias

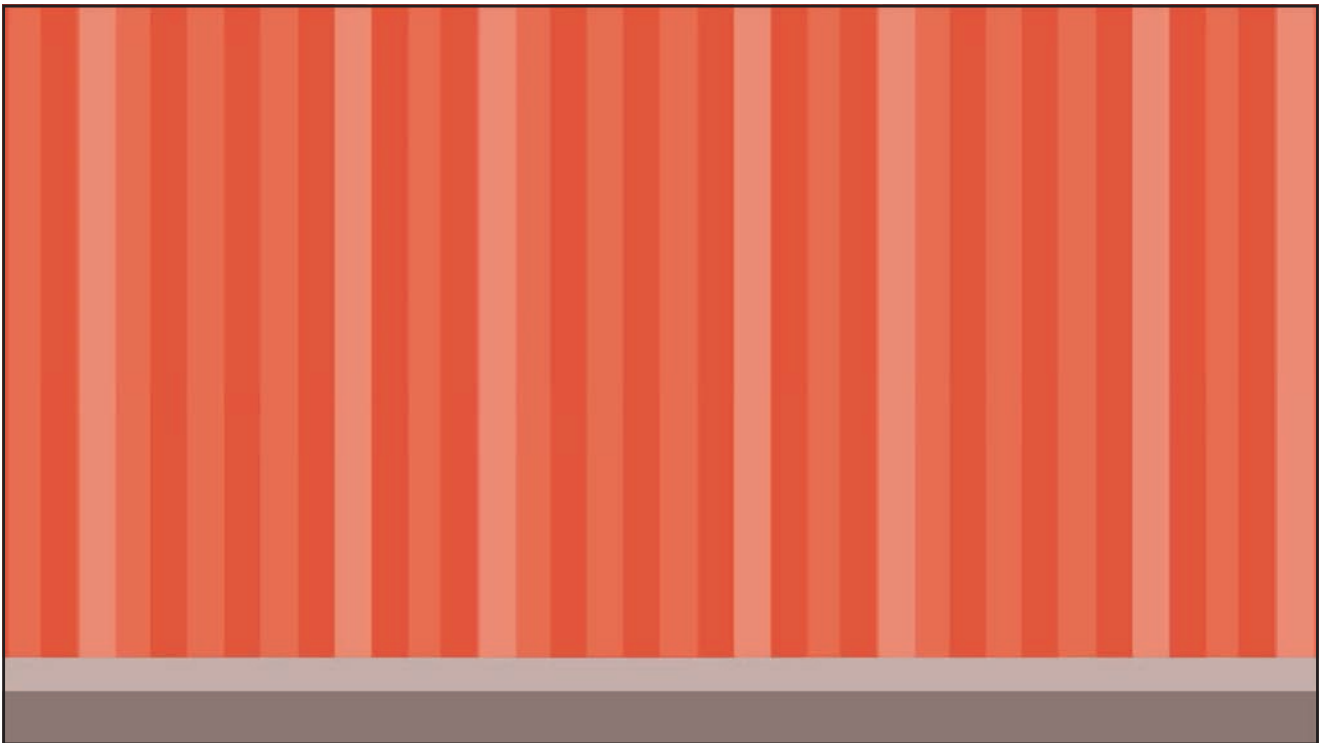
What the human being is best at doing is interpreting all new information so that
prior conclusions remain intact.

Warren Buffet



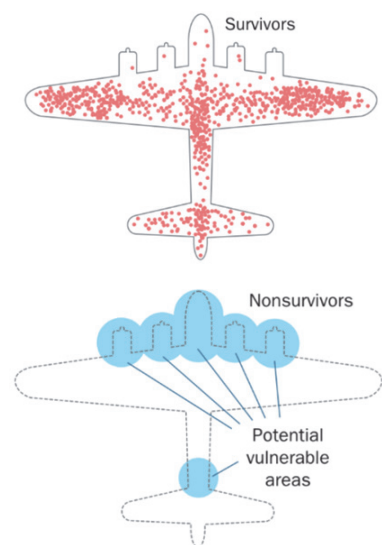
Photo: Thomson Reuters, LLC





Survivorship Bias

Tendency to **focus** only on what **succeeded** or survived while **ignoring** what **failed** or didn't survive.



Shortcuts



Who the hell wants to hear actors talk?

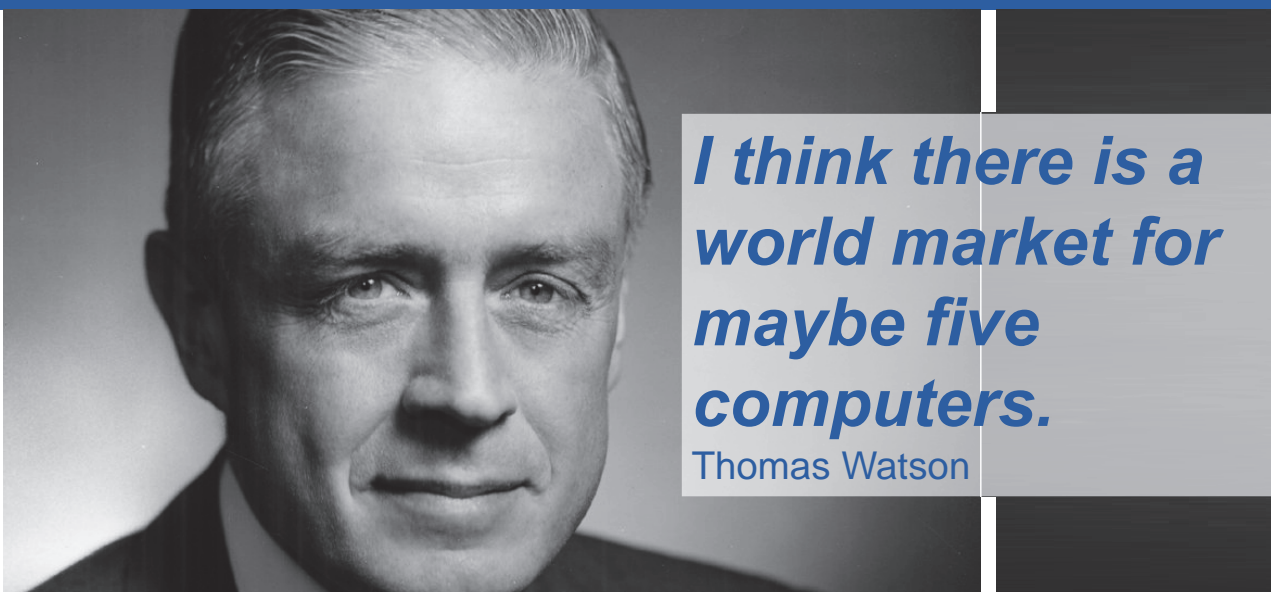
Harry M. Warner



123

©2023 Van Haren Publishing BV.

Shortcuts



I think there is a world market for maybe five computers.

Thomas Watson



124

©2023 Van Haren Publishing BV.

Curse of Knowledge



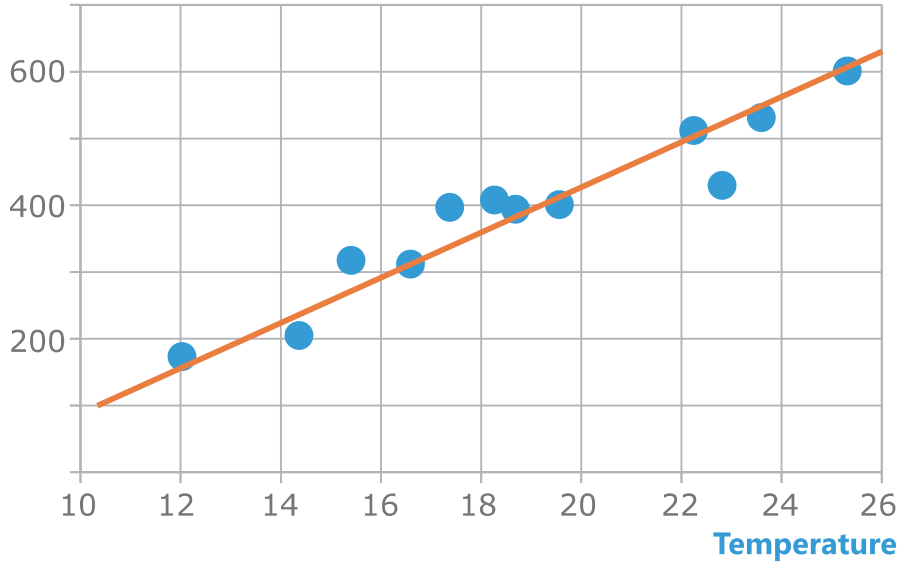
Tendency to assume **other people** have the necessary **context or knowledge** to follow what you're communicating

Causation



Correlation

Ice Cream revenue

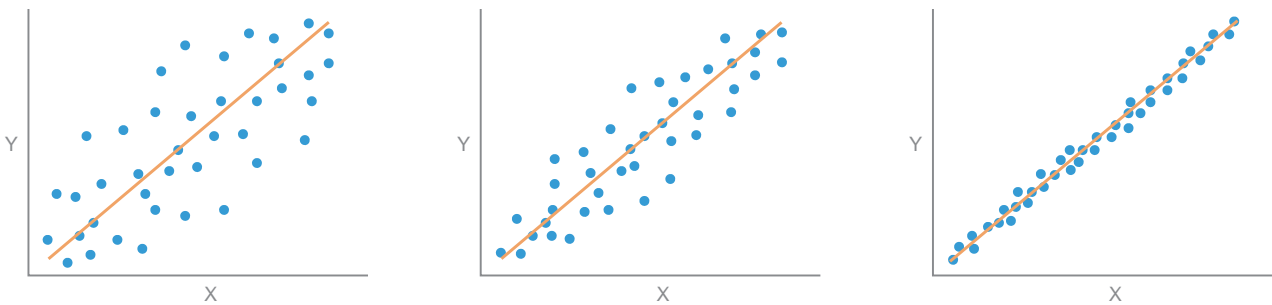


127

©2023 Van Haren Publishing BV.



Correlation



128

©2023 Van Haren Publishing BV.



Shortcuts

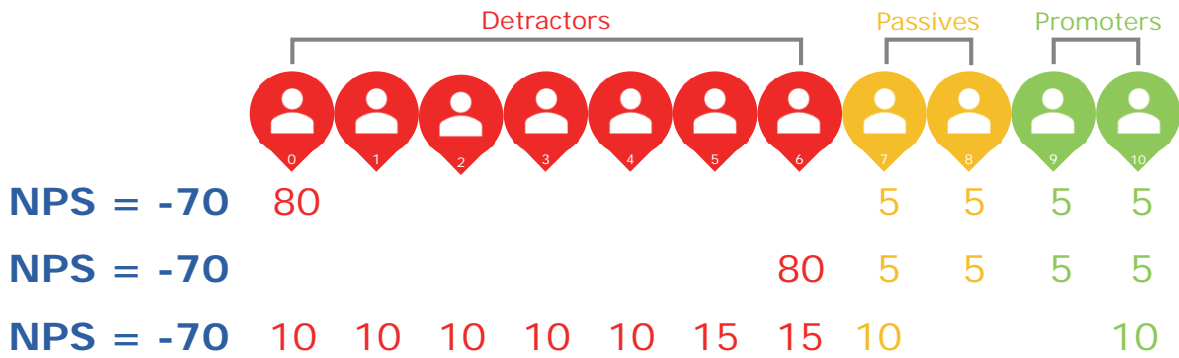
On a scale of 0 to 10,
how likely are you to
recommend our company?

Shortcuts



$$\text{NPS} = \text{Promoters\%} - \text{Detractors\%}$$

Shortcuts

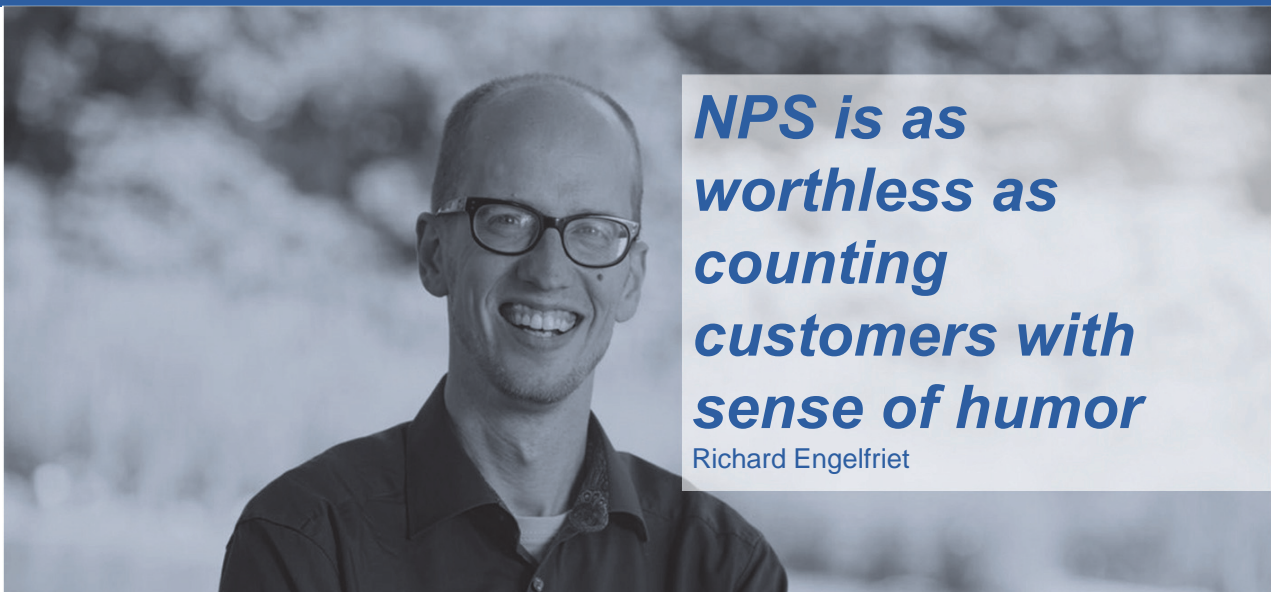


131

©2023 Van Haren Publishing BV.



Shortcuts



132

©2023 Van Haren Publishing BV.



Shortcuts

How likely are you to recommend Windows 10 to a friend or colleague?

1 2 3 4 5

Not at all likely Extremely likely

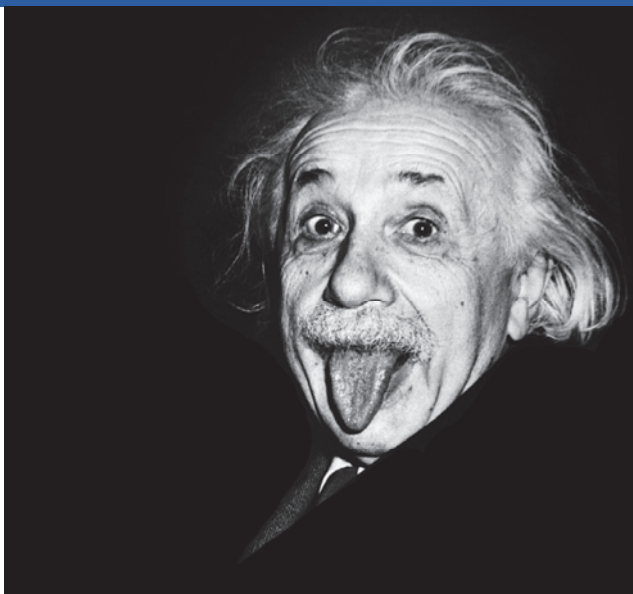
1 2 3 4 5

Not at all likely Extremely likely

Please explain why you gave this score.

I need you to understand that people don't have conversations where they randomly recommend operating systems to one another

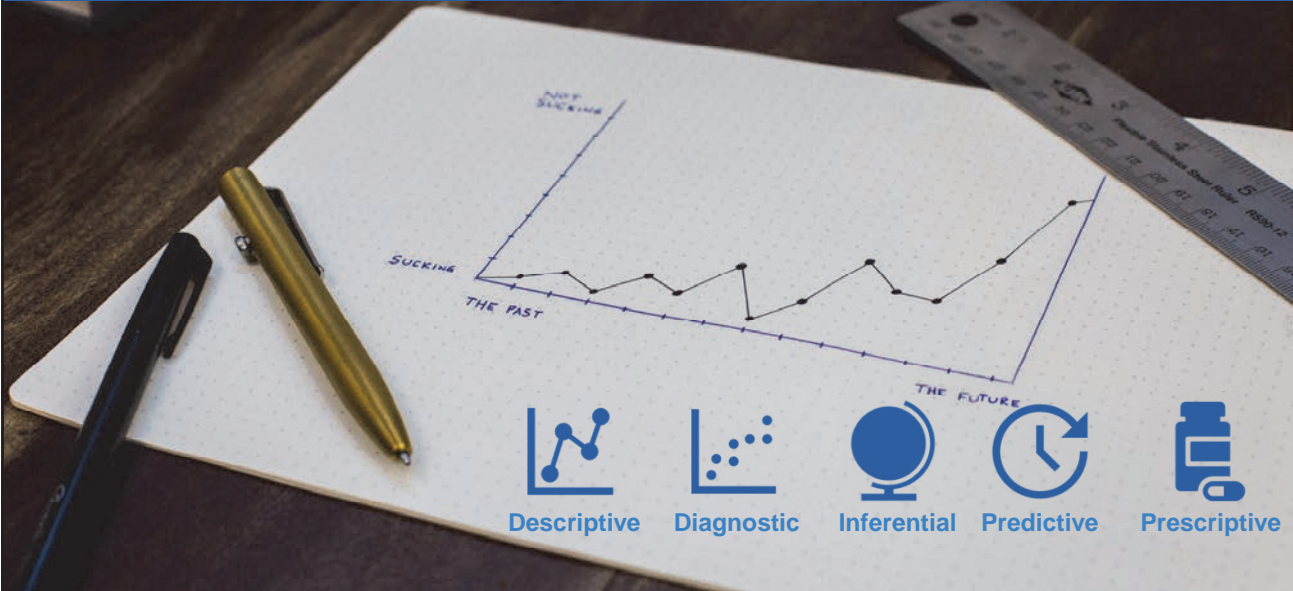
Shortcuts



Everything should be made as simple as possible, but no simpler.

Albert Einstein

Data Analysis



Descriptive

Diagnostic

Inferential

Predictive

Prescriptive



135

©2023 Van Haren Publishing BV.

Descriptive Analysis



Counts



Min/Max



Sums (totals)



Ratios/
Proportions/
Percent



Measure of
central tendency



Measure of
dispersion



136

©2023 Van Haren Publishing BV.

Descriptive Analysis



9.999

11 / 17.242

179

421

Diagnostic Analysis



Drilling down



Find correlations



Spot outliers



Diagnostic Analysis



Drilling down



139

©2023 Van Haren Publishing BV.

Diagnostic Analysis



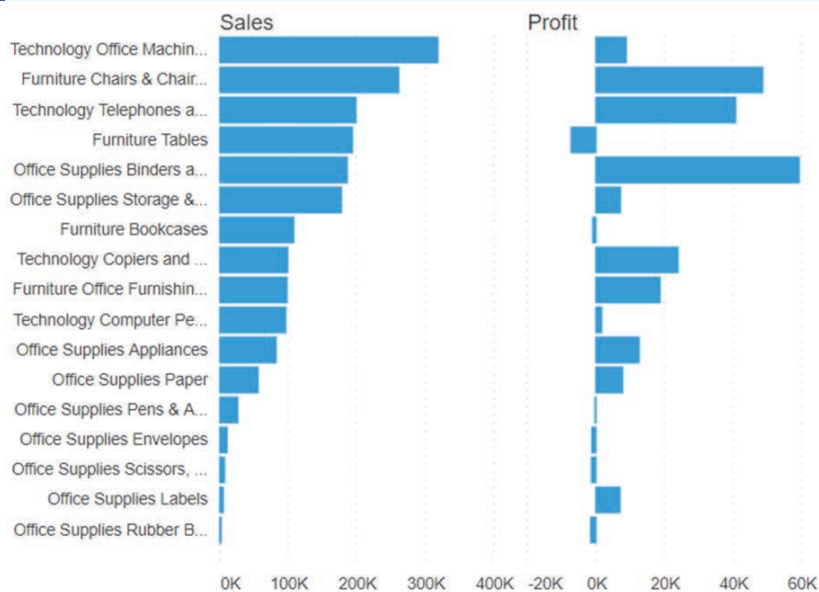
Drilling down



Find correlations



Spot outliers



140

©2023 Van Haren Publishing BV.

Inferential Analysis



Population



Sample



Sample size (N)



Representative sample

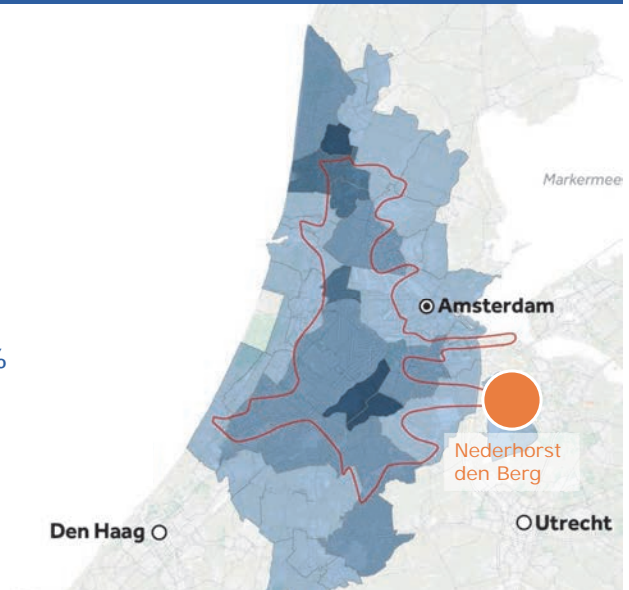
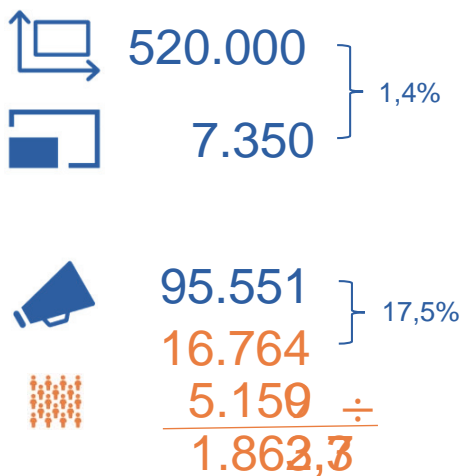


Under-coverage bias

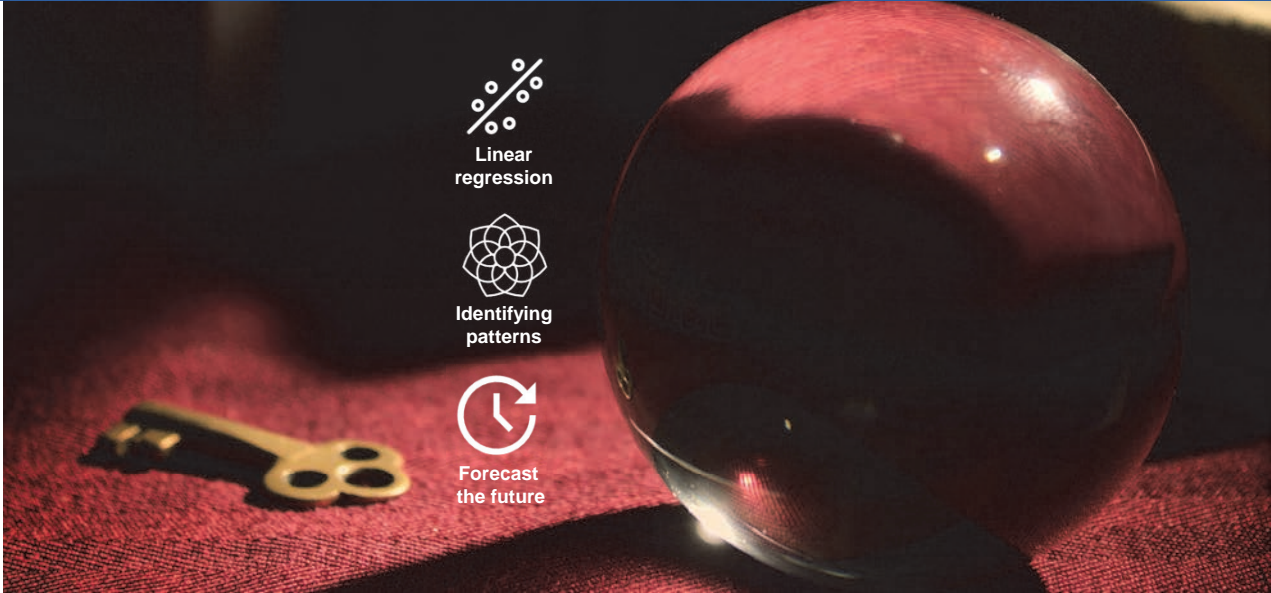


Self-selection bias

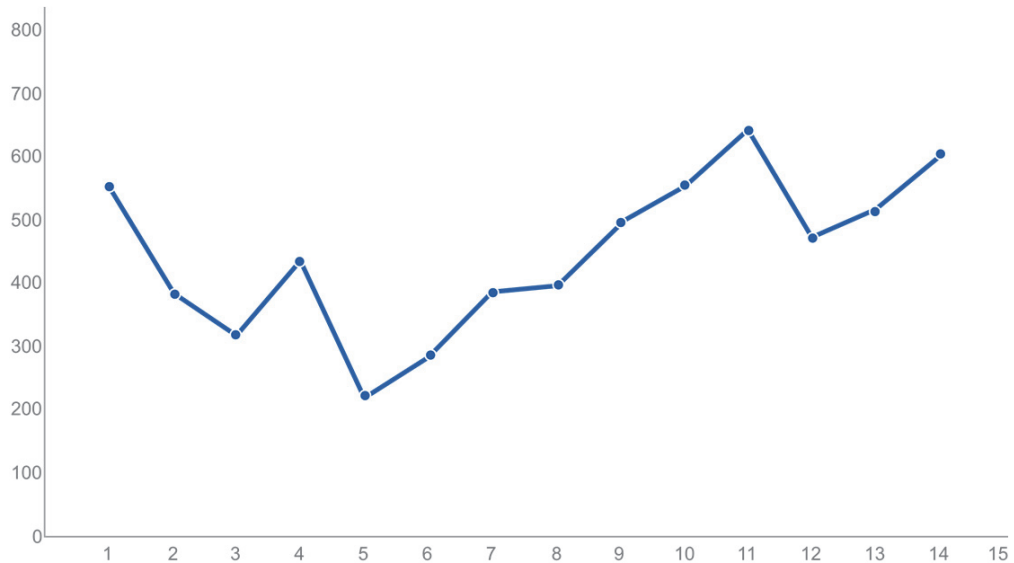
Inferential Analysis



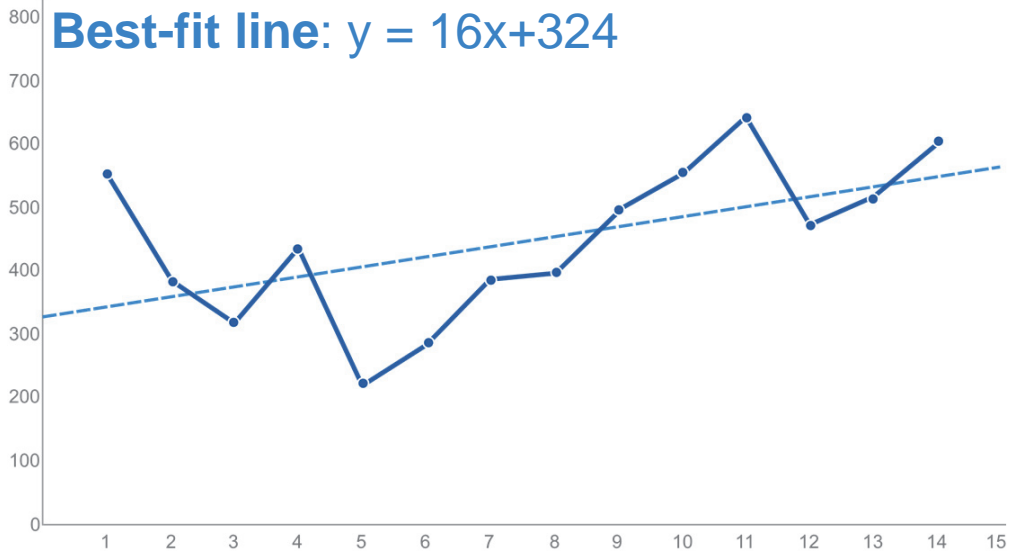
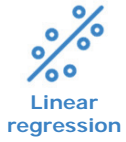
Predictive Analysis



Predictive Analysis



Predictive Analysis



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

©2023 Van Haren Publishing BV.

Predictive Analysis



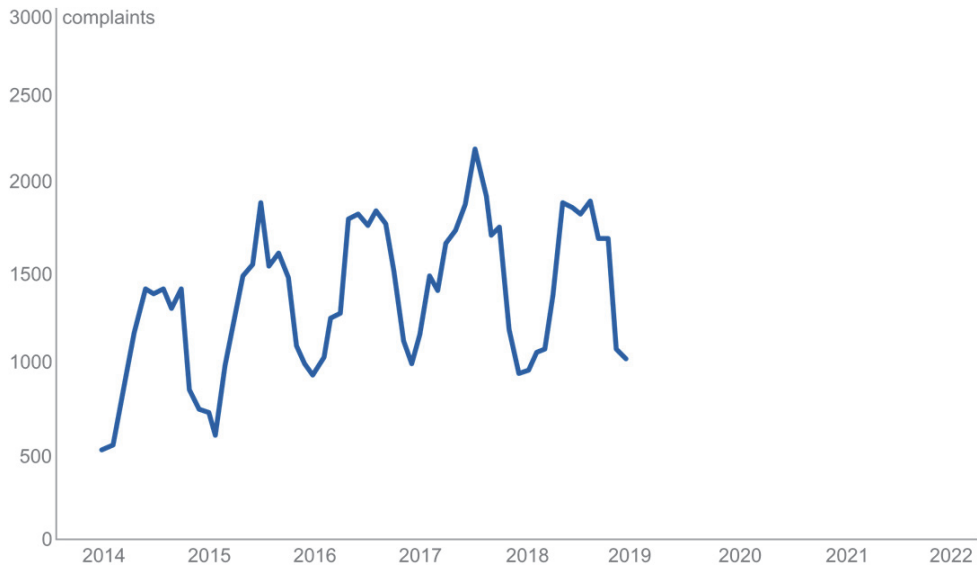
146 Source: Jones, Ben. Data Literacy Fundamentals: Understanding the Power & Value of Data

©2023 Van Haren Publishing BV.

Predictive Analysis


Linear regression


Identify patterns



Effective DATA Foundation

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

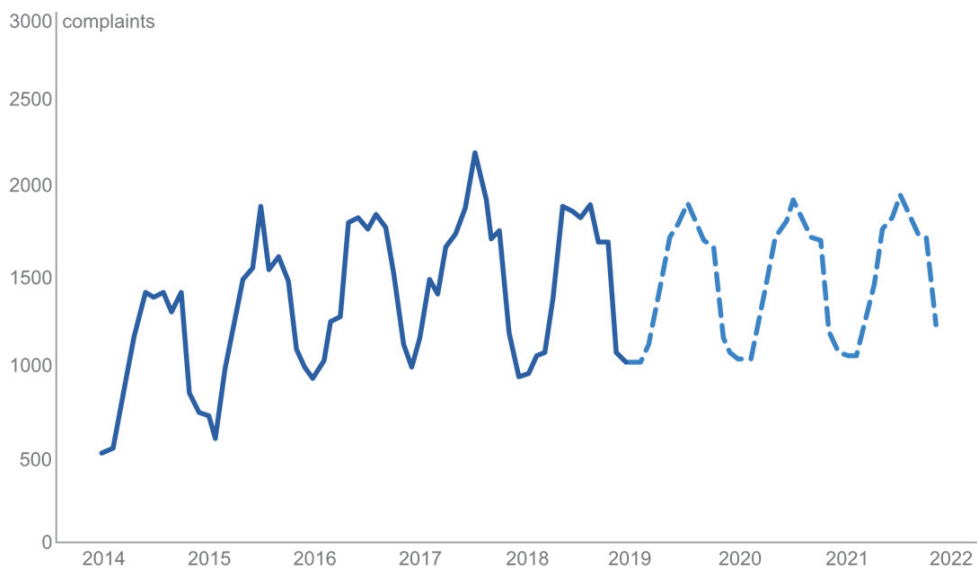
©2023 Van Haren Publishing BV.

Predictive Analysis


Linear regression


Identify patterns


Forecast the future



Effective DATA Foundation

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

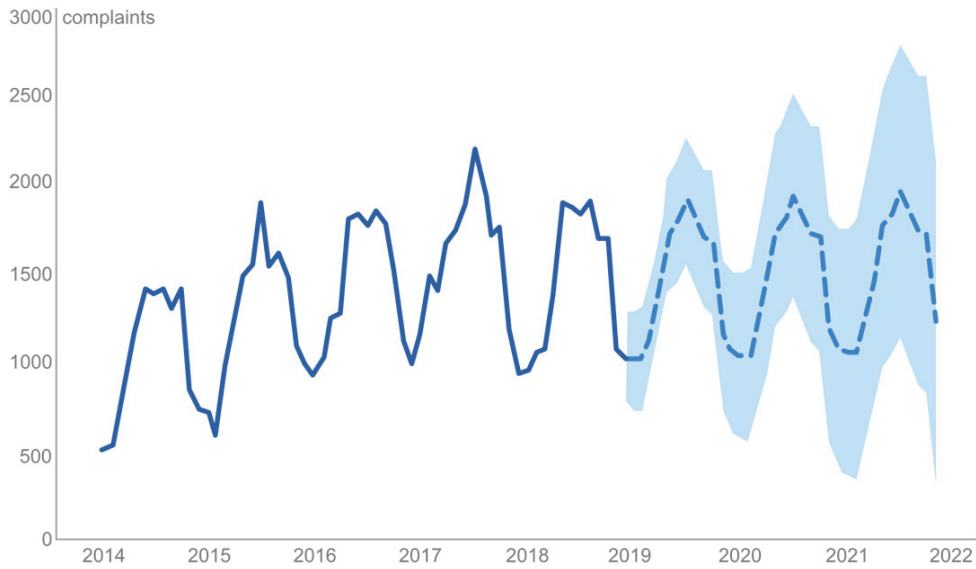
©2023 Van Haren Publishing BV.

Predictive Analysis

Linear regression

Identify patterns

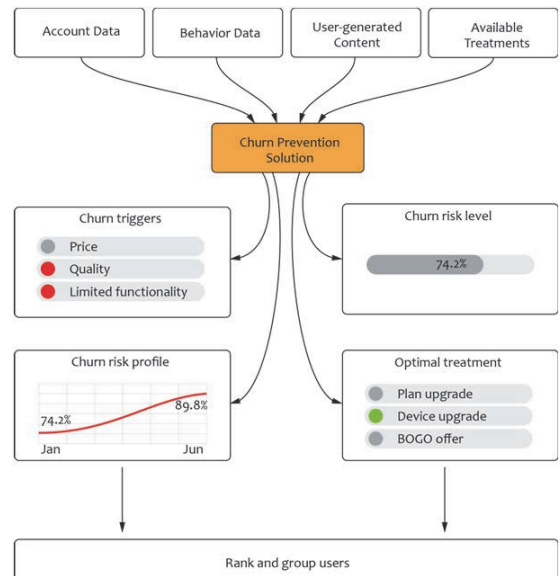
Forecast the future



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

©2023 Van Haren Publishing BV.

Prescriptive Analysis



Source: <https://blog.griddynamics.com/customer-churn-prevention-prescriptive-solution-using-deep-learning/> & <https://kpi-max.com/churn-rate/>

©2023 Van Haren Publishing BV.

Analytical Skills



First principle for understanding data

No **data** have **meaning** apart from their **context**

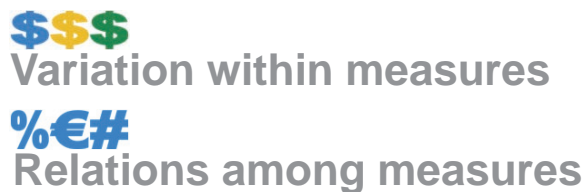
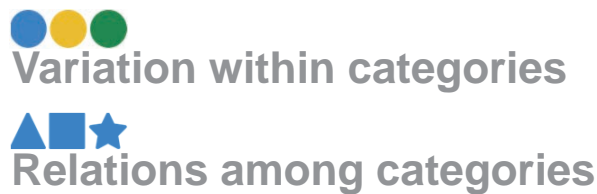
- *Trust no one who cannot, or will not, provide the context for their figures.*
- *Stop reporting comparisons between pairs of values except as part of a broader comparison.*
- *Start using graphs to present current values in context.*

What is Data?

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

Source: unknown

Qualities & Quantities



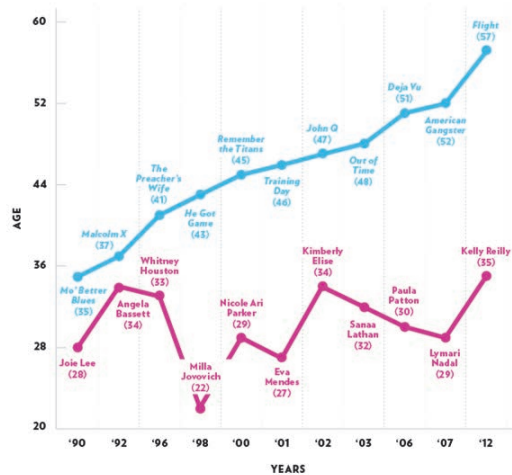
Exercise 4

let's PRACTICE

Exercise: Leading men age, but...

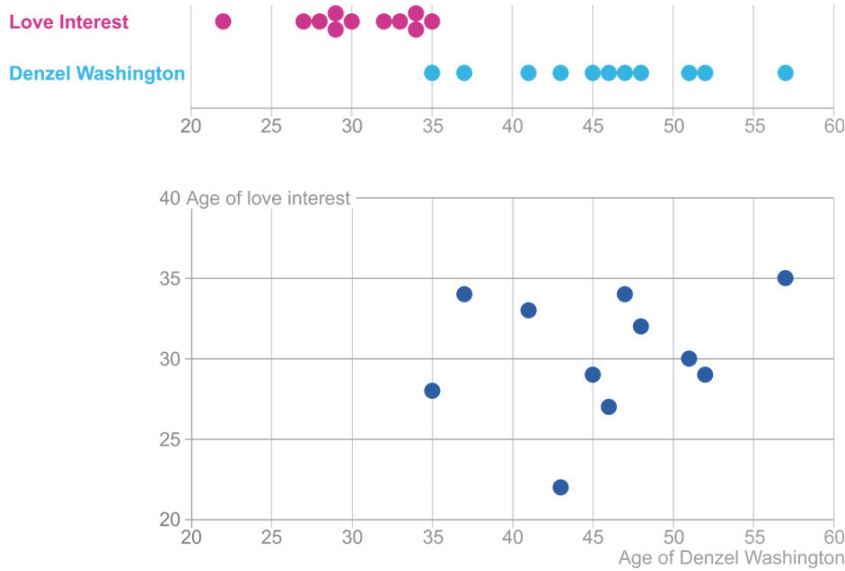
Leading Men Age, But Their Love Interests Don't

DENZEL WASHINGTON



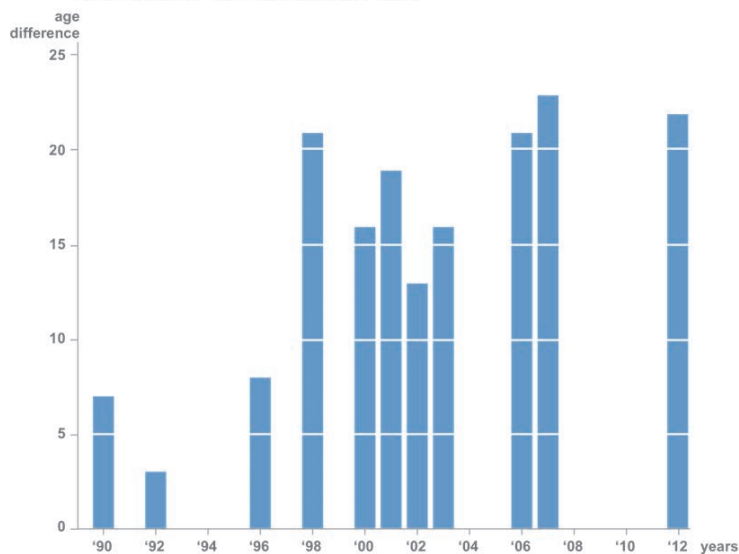
WASHINGTON'S AGE WHEN MAKING MOVIE LOVE INTEREST'S AGE

Exercise: Leading men age, but...



Exercise: Leading men age, but...

DENZEL WASHINGTON



Variations within categories



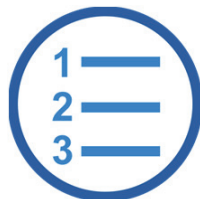
tally



sum



average



ranking



part-to-whole



159

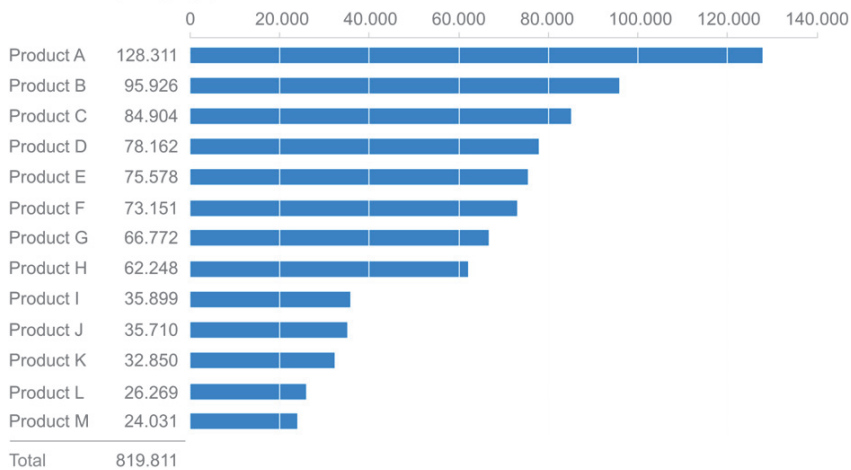
©2023 Van Haren Publishing BV.

Variations within categories



ranking

Revenue (in €) by product



160

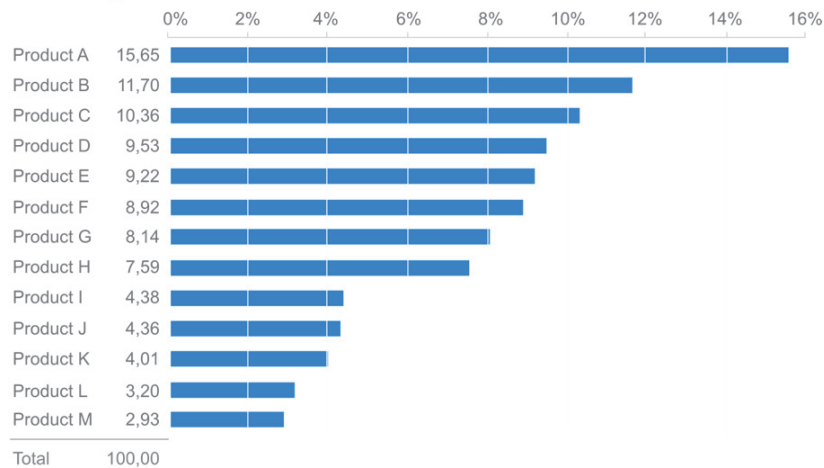
©2023 Van Haren Publishing BV.

Variations within categories



part-to-whole

Percentage of total revenue



161

©2023 Van Haren Publishing BV.

Relations among categories



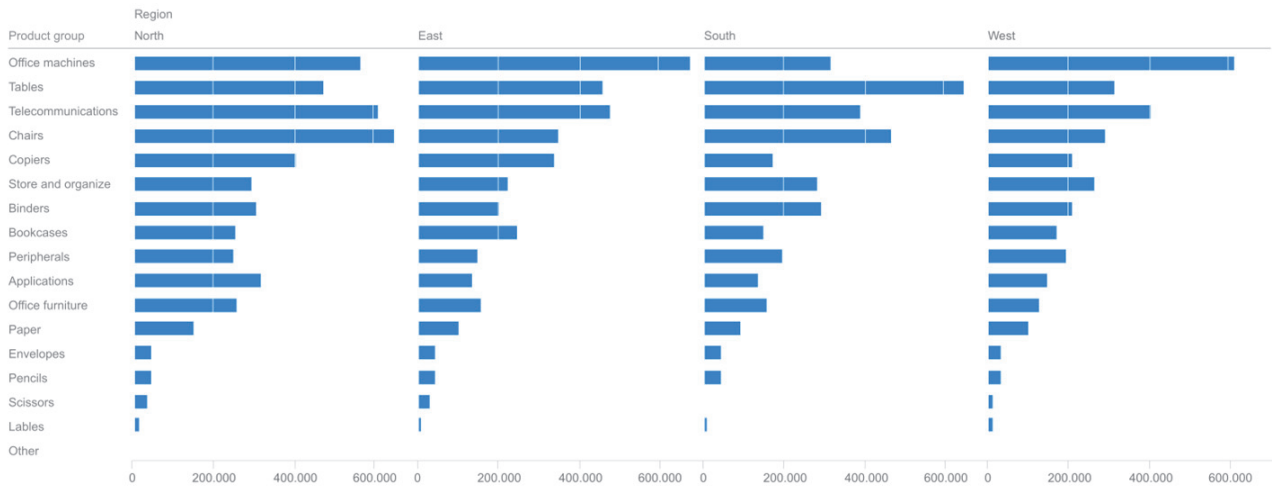
Product group	Region			
	North	East	South	West
Office machines	563.395	673.390	321.105	610.807
Tables	471.751	454.887	652.965	316.405
Telecommunications	613.410	475.653	394.726	405.524
Chairs	651.564	348.052	469.652	292.478
Copiers	404.175	343.117	173.833	209.237
Store and organize	299.116	227.534	280.367	263.166
Binders	309.262	203.847	294.907	214.942
Bookcase	258.919	246.411	145.818	171.504
Peripherals	250.718	150.974	198.649	195.535
Applications	317.079	133.946	136.944	149.023
Office Furniture	259.389	159.443	149.828	129.434
Paper	150.710	98.576	96.958	100.210
Envelopes	47.531	49.608	43.691	33.256
Pencils	45.807	42.625	42.908	35.768
Scissors	36.376	30.577	4.729	9.315
Lables	14.062	7.692	6.298	10.930
Other	5.815	3.416	3.089	2.687



162

©2023 Van Haren Publishing BV.

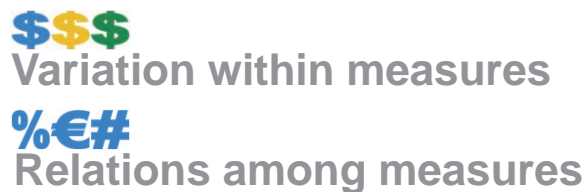
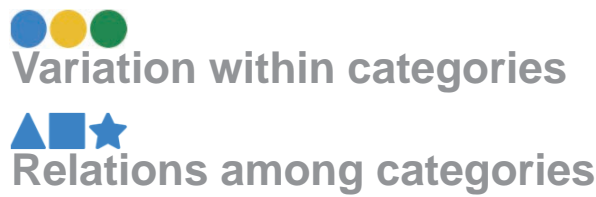
Relations among categories



163

©2023 Van Haren Publishing BV.

Qualities & Quantities



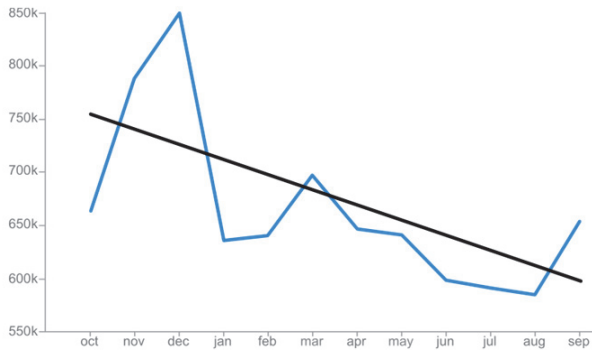
164 Source: S. Few – Signal

©2023 Van Haren Publishing BV.

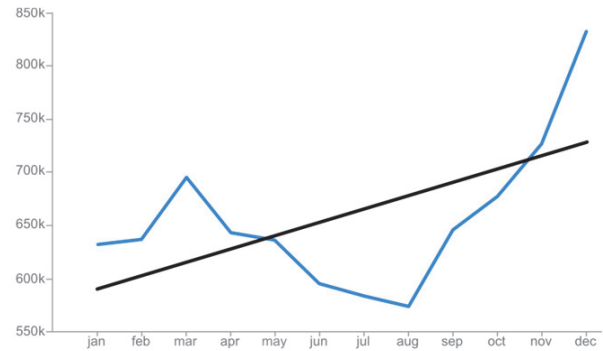
Variation within measures



Expenses from October 2021 until September 2022



Expenses from January 2022 until December 2022



165

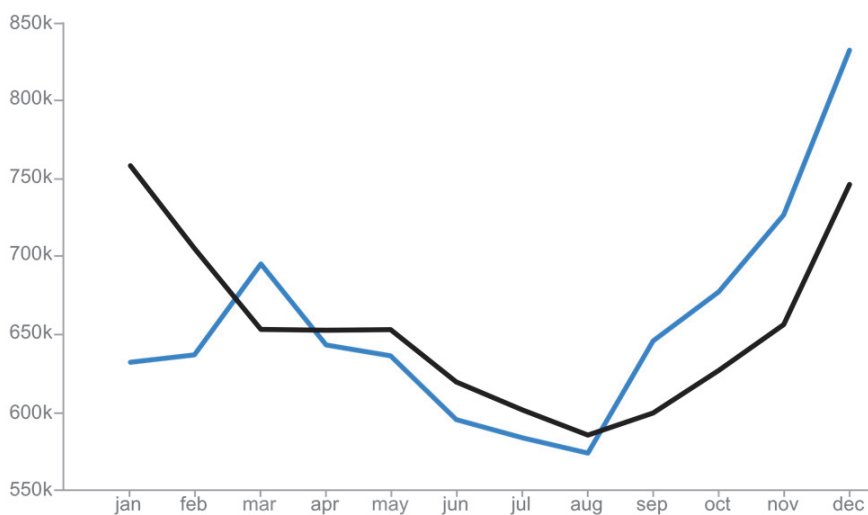
©2023 Van Haren Publishing BV.



Variation within measures



Expenses from January 2022 until December 2022

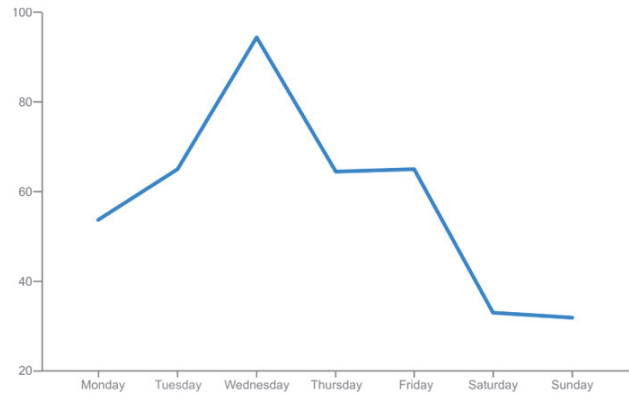
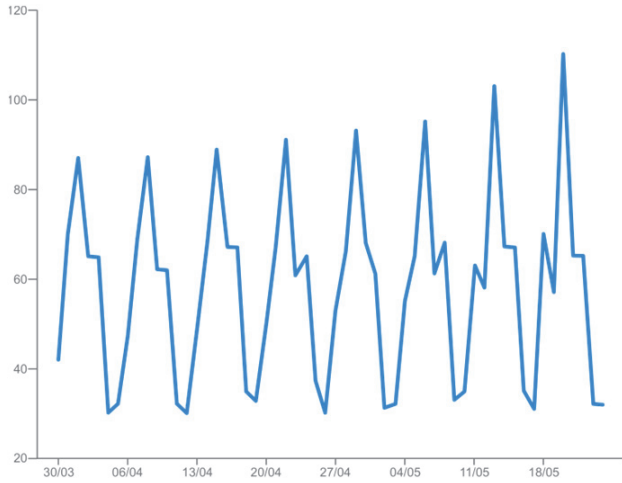


166

©2023 Van Haren Publishing BV.



Variation within measures

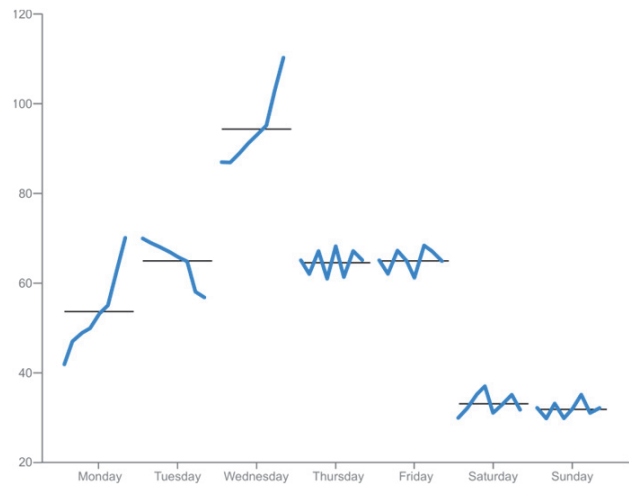
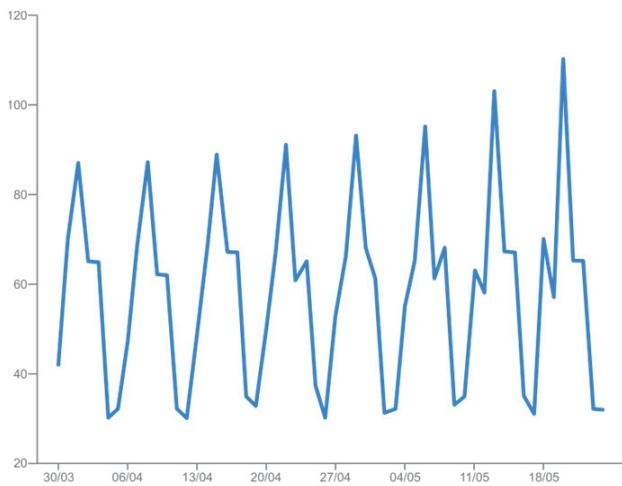


167

©2023 Van Haren Publishing BV.



Variation within measures



168

©2023 Van Haren Publishing BV.



Qualities & Quantities



 Variation within categories

 Relations among categories

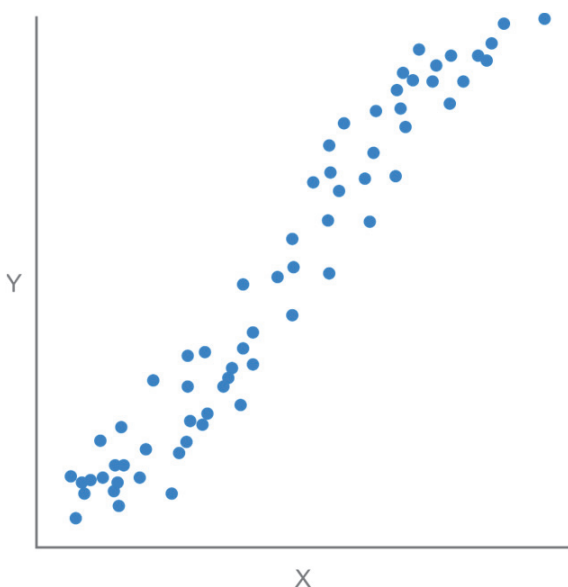


 Variation within measures

 Relations among measures

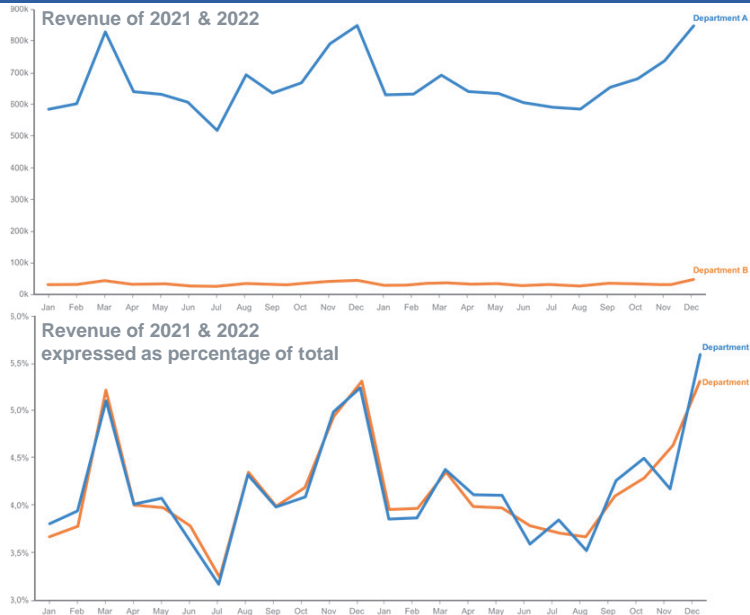
Relations among measures





Relations among measures

%€#



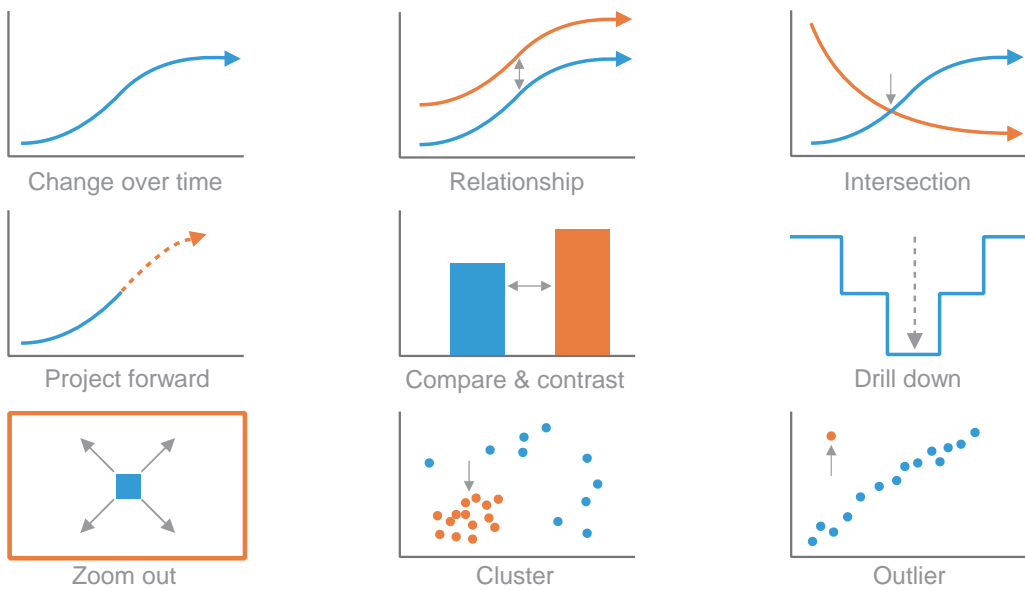
Effective DATA Foundation

171

©2023 Van Haren Publishing BV.

Look for patterns

%€#







Effective DATA Foundation

172 Source: Brent Dykes – Effective Data Storytelling

©2023 Van Haren Publishing BV.

Wrap up Analyze Data

-  What do you **expect** to see/find?
-  Be aware of your **biases**
-  Select the appropriate **analysis**
-  Train your analytical **skills**

3 C's of Data Literacy



Curiosity



Creativity



Critical Thinking

Argue with data

*No one has ever made a decision because of a number.
They need a story.*

Daniel Kahneman

Training Agenda



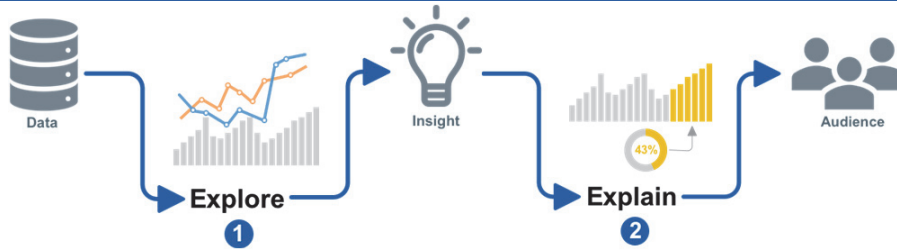
Argue with
data

Explore to Explain

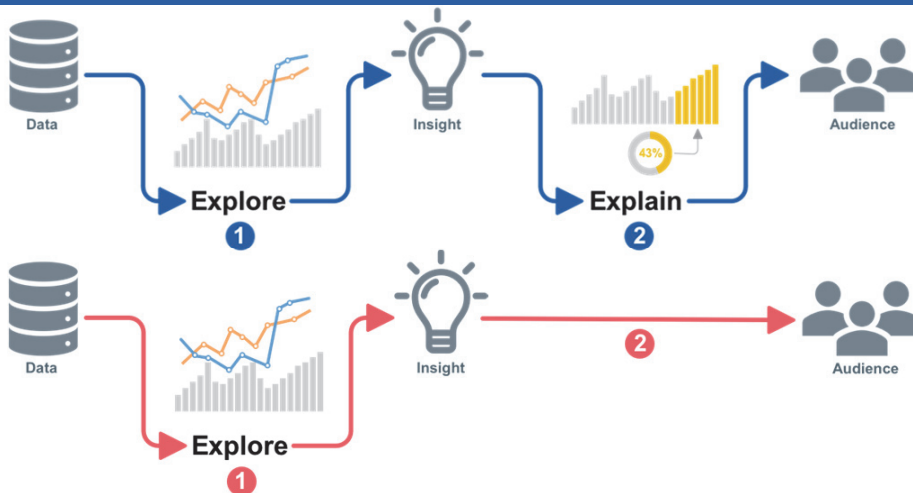
CHRTTS

Storytelling

Explore to Explain

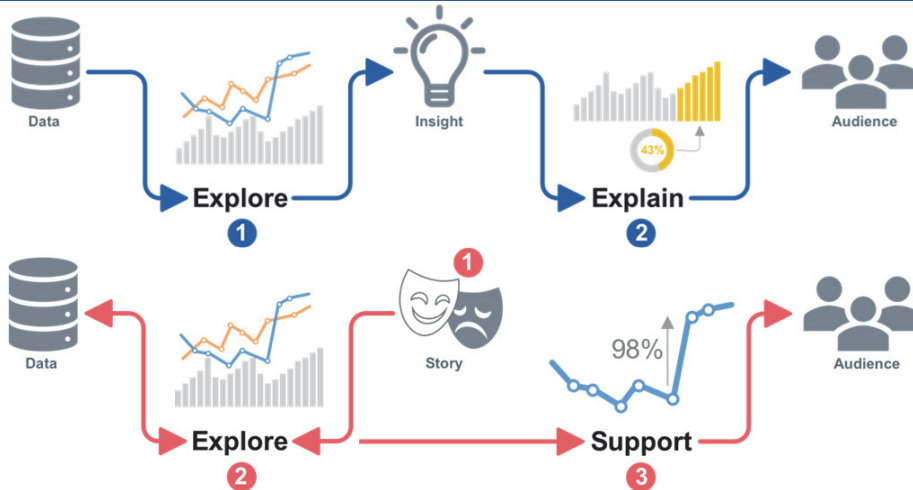


Explore to Explain – the data cut



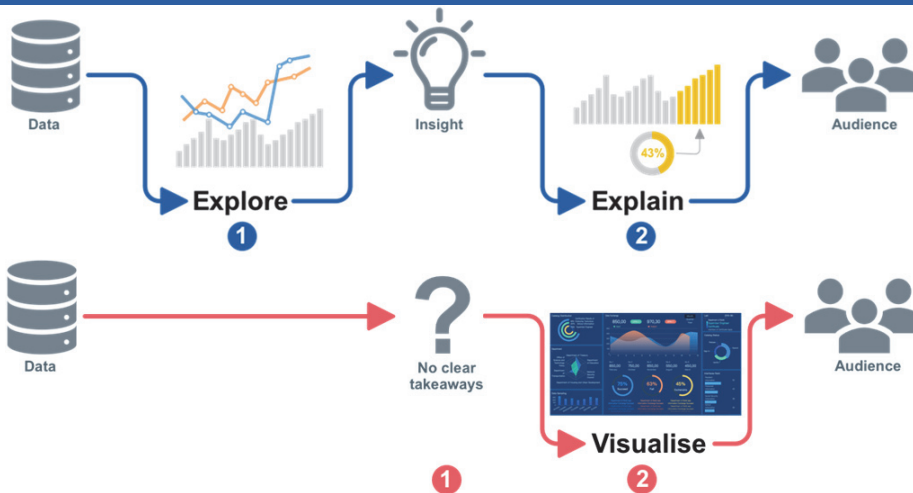
- You feel the data speaks for itself because the evidence is so strong;
- You aren't sure how the audience will receive or interpret the results;
- You haven't spent much time tailoring your charts to your audience.

Explore to Explain – the data cameo



- You already know the narrative you want to tell before examining any data;
- You are selecting data that supports a particular viewpoint;
- You aren't looking to disprove your preferred viewpoint.

Explore to Explain – the data decoration



- You don't have a clear focus or emphasis for the visuals you're creating;
- You are more focused on the data visualization tool than the actual data;
- You want to visualize the data so other people with more domain expertise can make better sense of the numbers.

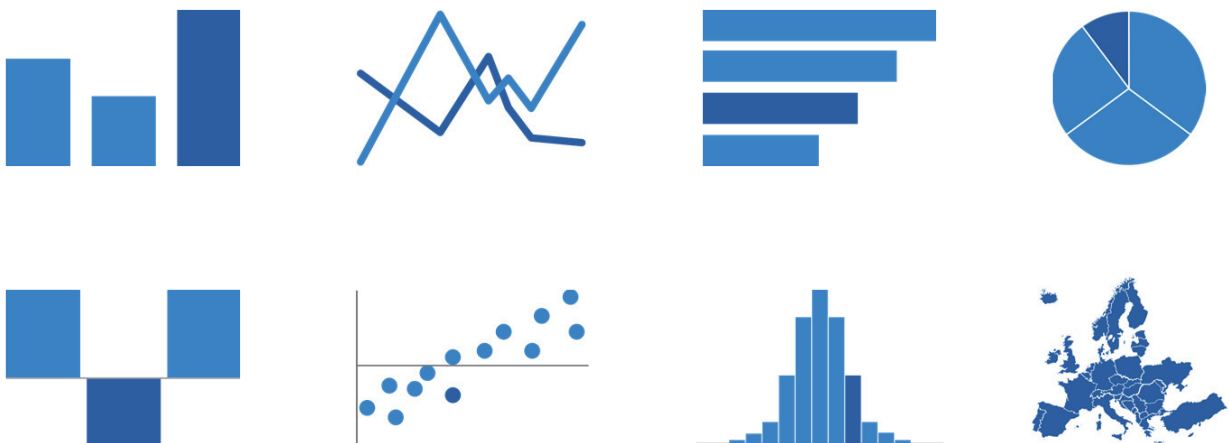
Explore to Explain



181 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

Design carefully



182 Source: Stephen S. Few, Show me the Numbers, 2004

©2023 Van Haren Publishing BV.

Design carefully

C
Categorical
Comparing categories and distributions of quantitative values




Revealing part-to-whole relationships and hierarchies

H
Hierarchical



R
Relational
Exploring correlations and connections



Plotting trends and intervals over time

T
Temporal




T
Tabular
Organizing observations by variable to allow precise comparison



Mapping spatial patterns through overlays and distortions

S
Spatial



183 Based on: Data Visualization, Andy Kirk

©2023 Van Haren Publishing BV.



Exercise 5

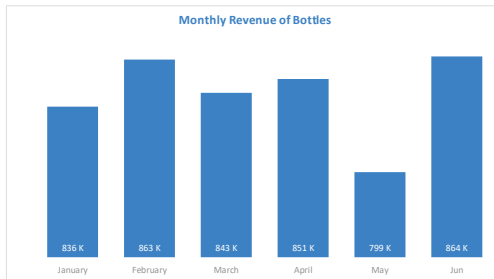
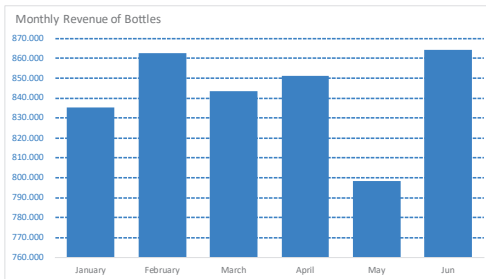
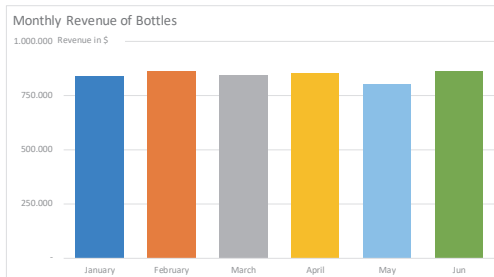
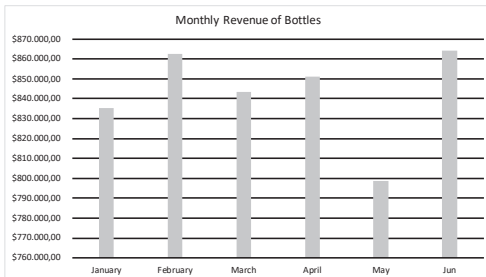
let's **PRACTICE**

184

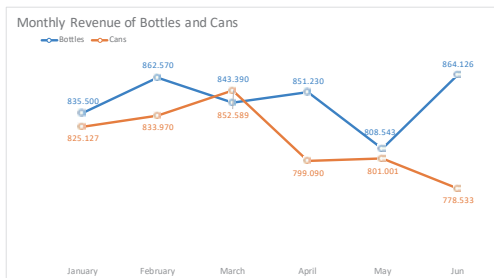
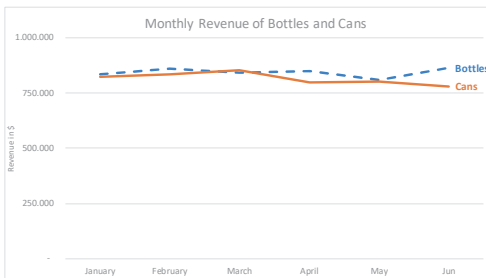
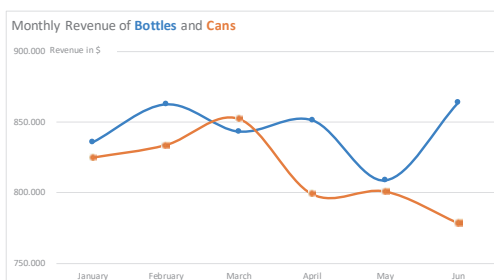
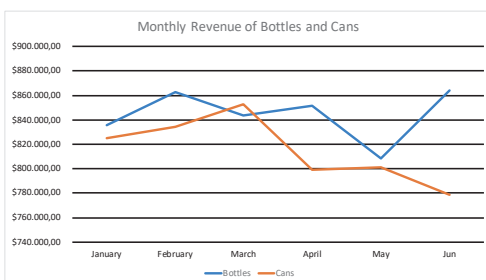
©2023 Van Haren Publishing BV.



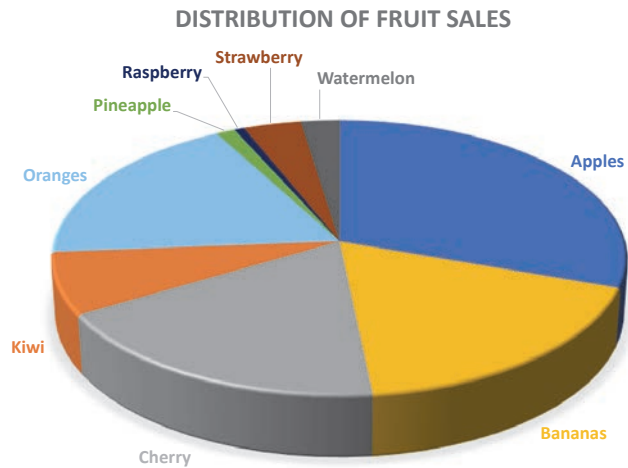
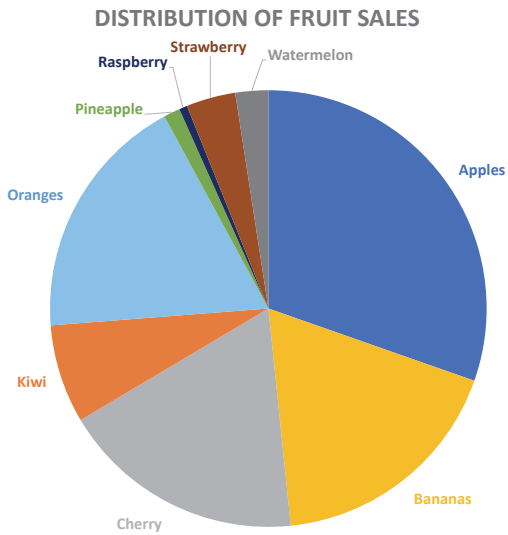
Exercise: Four Column Charts



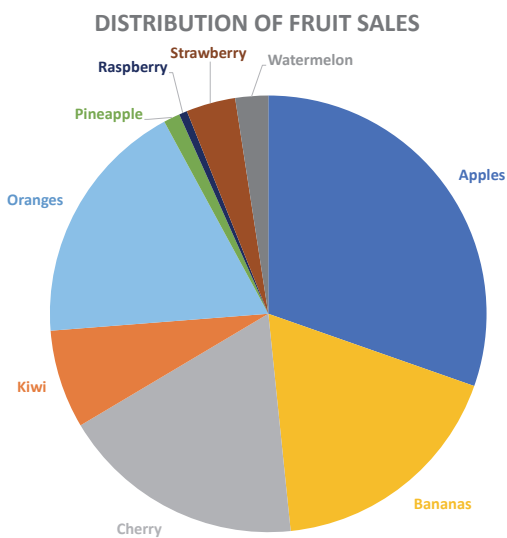
Exercise: Four Line Charts



Distribution of Fruit sales

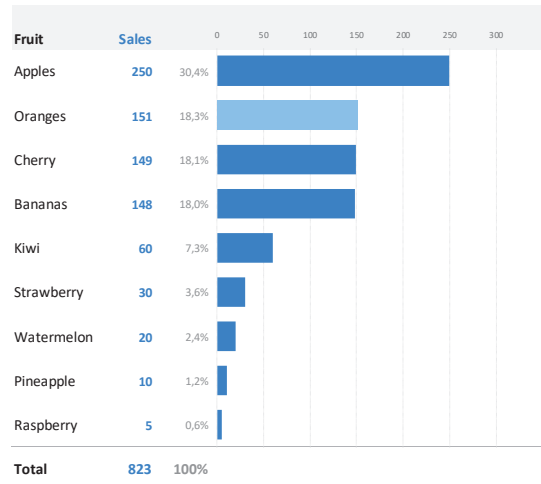
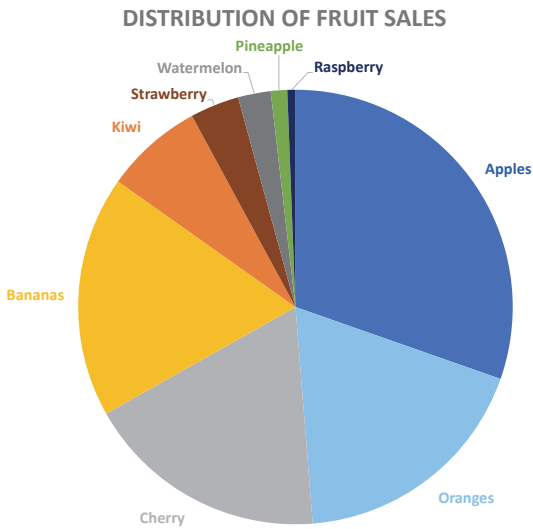


Distribution of Fruit sales – option 1

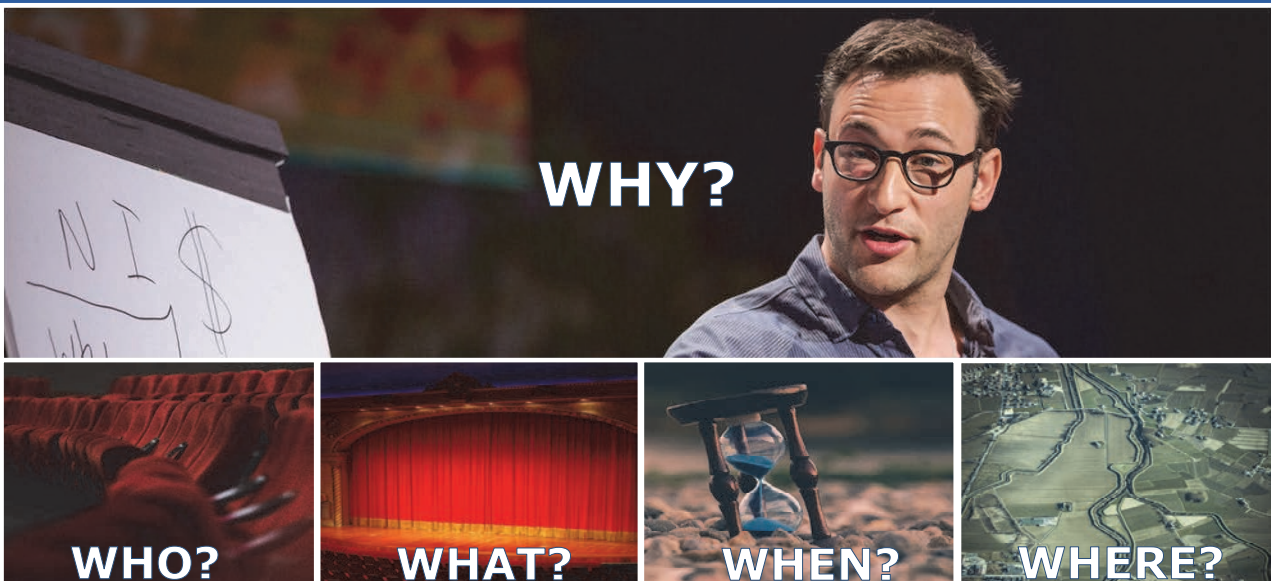


Fruit	Sales	0	50	100	150	200	250	300
Apples	250	30,4%						
Bananas	148	18,0%						
Cherry	149	18,1%						
Kiwi	60	7,3%						
Oranges	151	18,3%						
Pineapple	10	1,2%						
Raspberry	5	0,6%						
Strawberry	30	3,6%						
Watermelon	20	2,4%						
Total	823	100%						

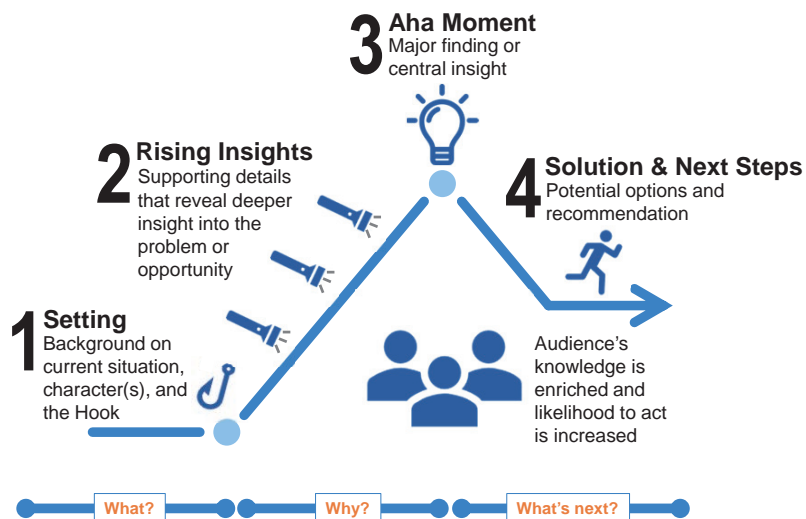
Distribution of Fruit sales – option 2



W questions



Storytelling Arc



191 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.



Storytelling principles



The ability to pivot from exploratory to explanatory in the analysis process is what separates effective data storytellers from everyone else who is attempting to share data.

192 Brent Dykes - author of Effective Data Storytelling

©2023 Van Haren Publishing BV.



Visual storytelling principles

PART 1 – THE SETUP

 RIGHT DATA

 RIGHT VISUALIZATIONS

 RIGHT CONFIGURATIONS

PART 2 – THE POLISH

 REMOVE NOISE

 FOCUS ATTENTION

 MAKE APPROACHABLE

 INSTILL TRUST

Visual storytelling principles

 RIGHT DATA

If you don't have **sound data**, it is difficult to find **meaningful insights.**

Brent Dykes

Visual storytelling principles

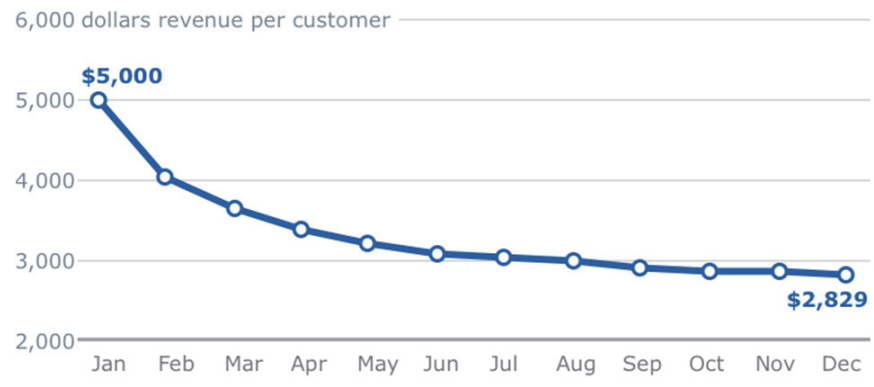
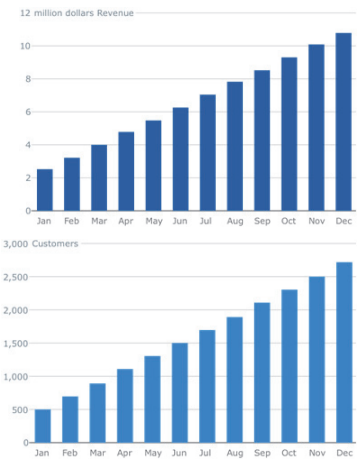
RIGHT DATA



195 Source: Effective data storytelling – Brent Dykes

Visual storytelling principles

RIGHT DATA

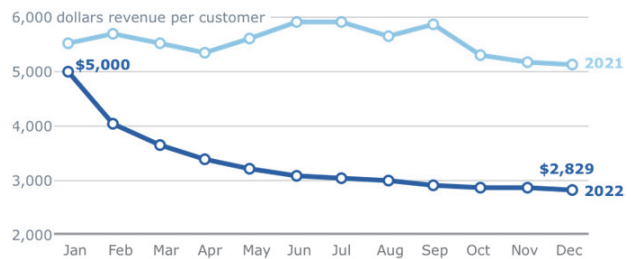
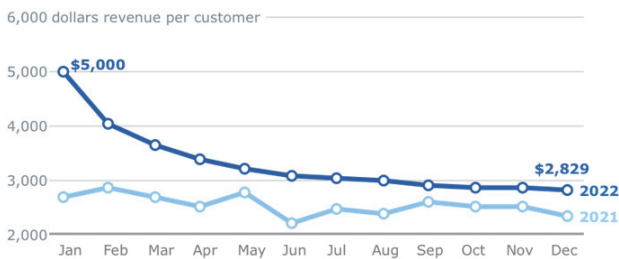
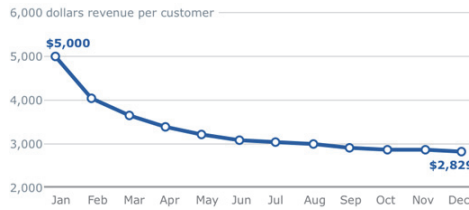


196 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

Visual storytelling principles

RIGHT DATA

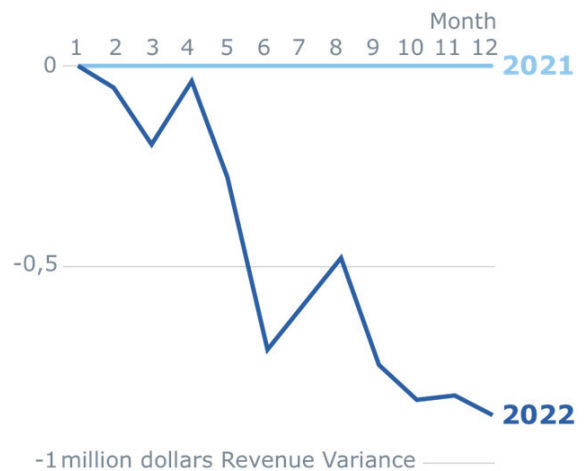
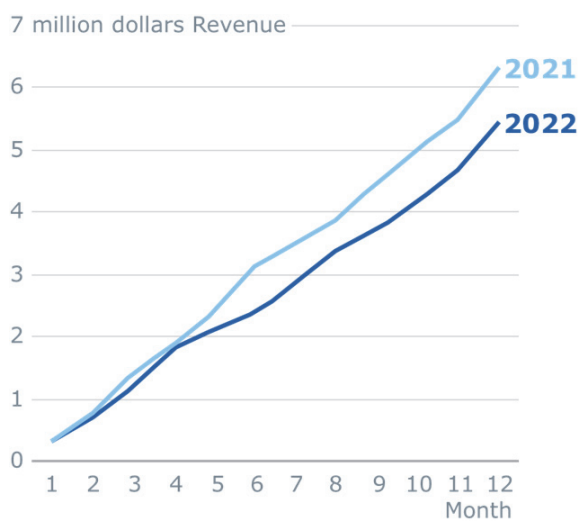


197 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

Visual storytelling principles

RIGHT DATA

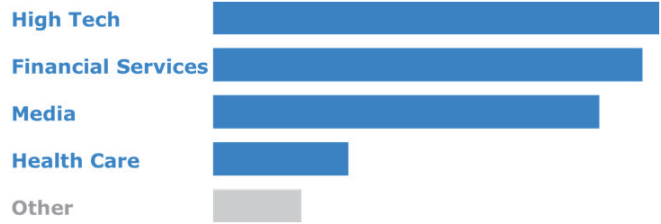
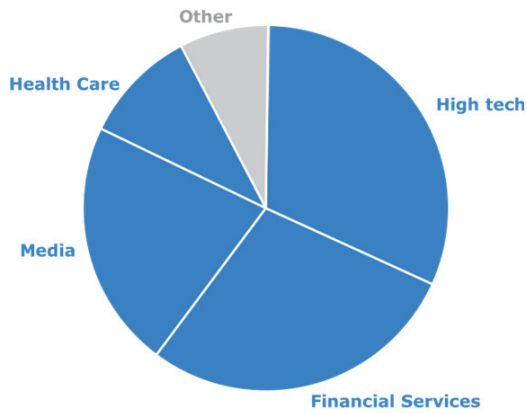


198 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

Visual storytelling principles

RIGHT VISUALIZATIONS



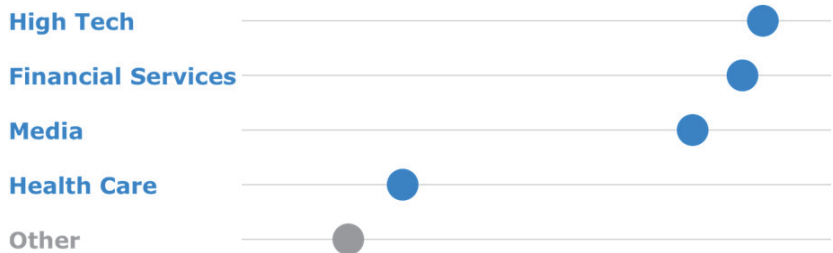
199 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.



Visual storytelling principles

RIGHT VISUALIZATIONS



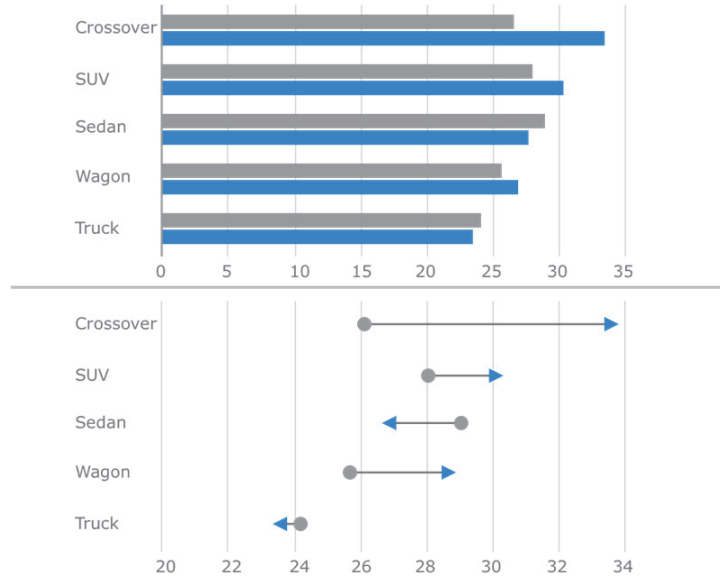
200 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.



Visual storytelling principles

RIGHT VISUALIZATIONS

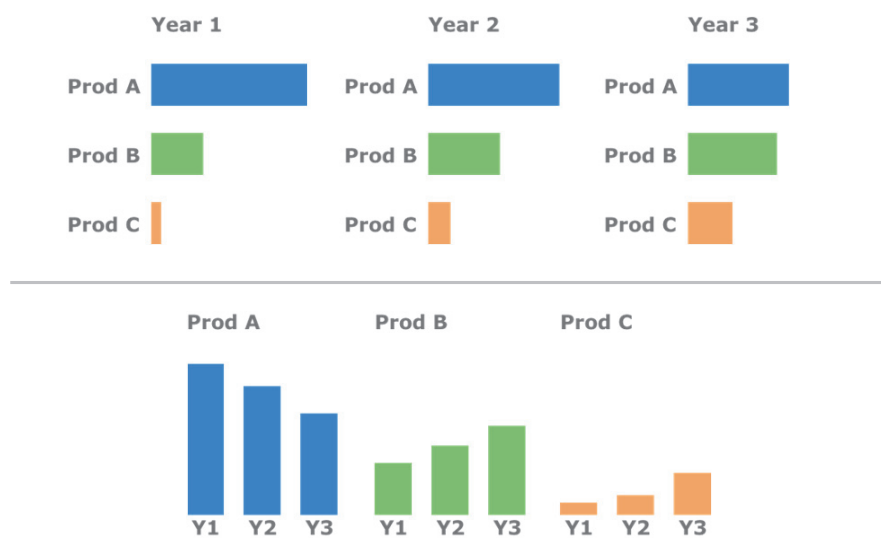


201 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

Visual storytelling principles

RIGHT CONFIGURATIONS

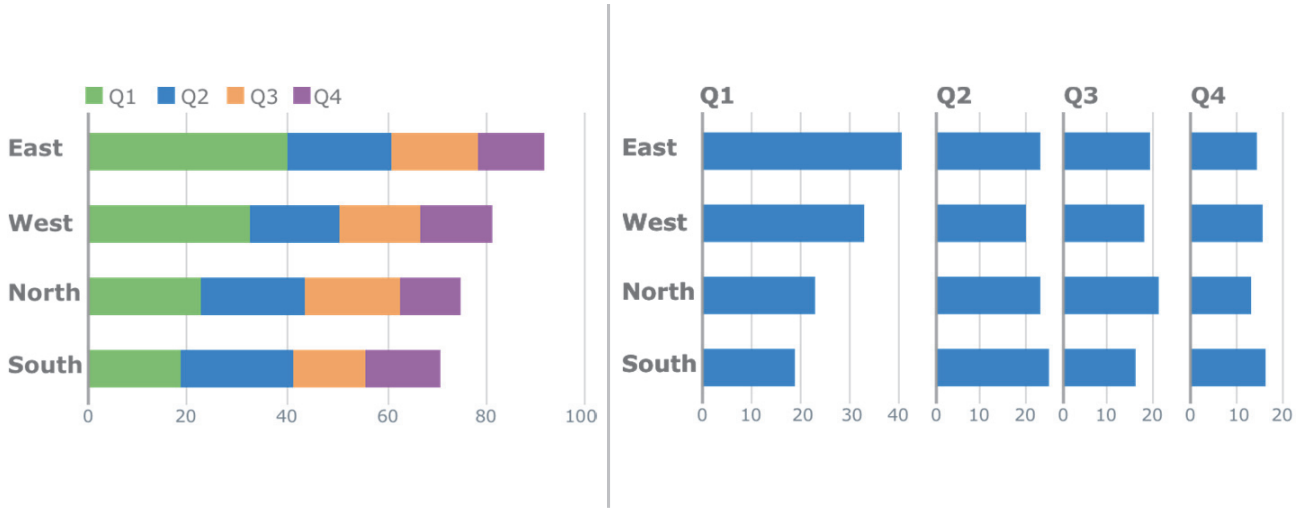


202 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

Visual storytelling principles

RIGHT CONFIGURATIONS



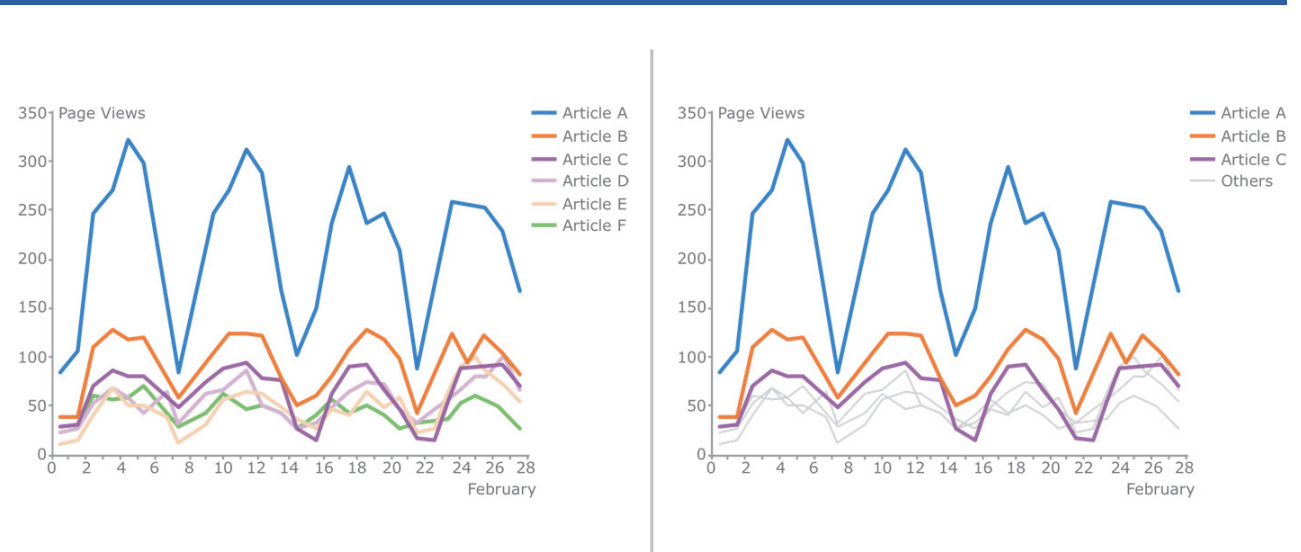
203 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.



Visual storytelling principles

REMOVE NOISE



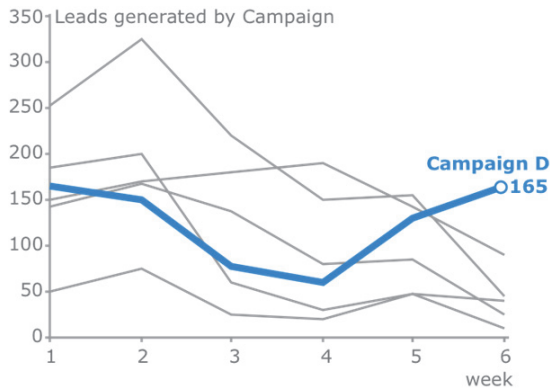
204 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

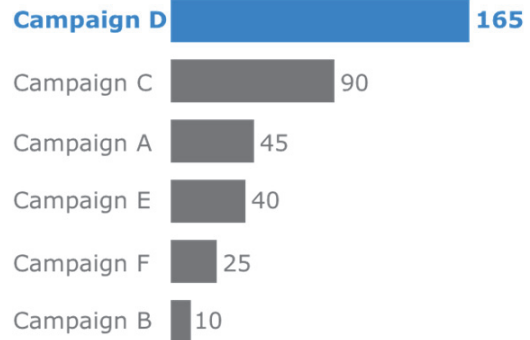


Visual storytelling principles

FOCUS ATTENTION



Leads generated by Campaign in Week 6



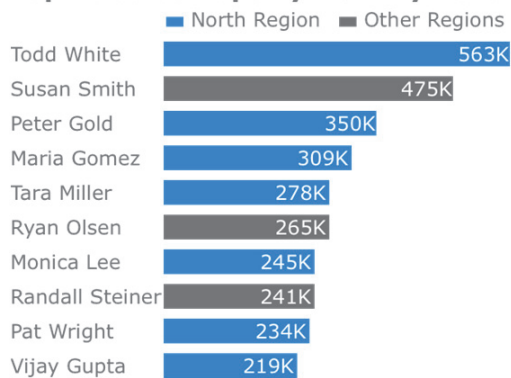
205 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

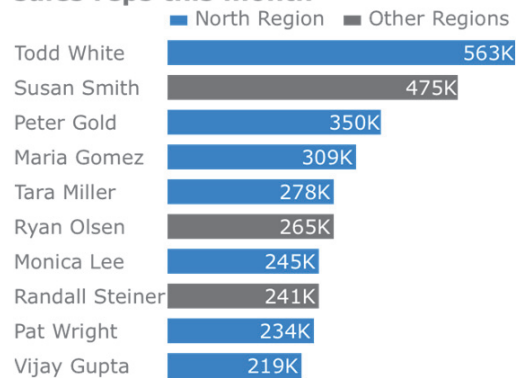
Visual storytelling principles

FOCUS ATTENTION

Top 10 Sales Reps by Monthly Revenue



North Region has 7 of the top 10 sales reps this month



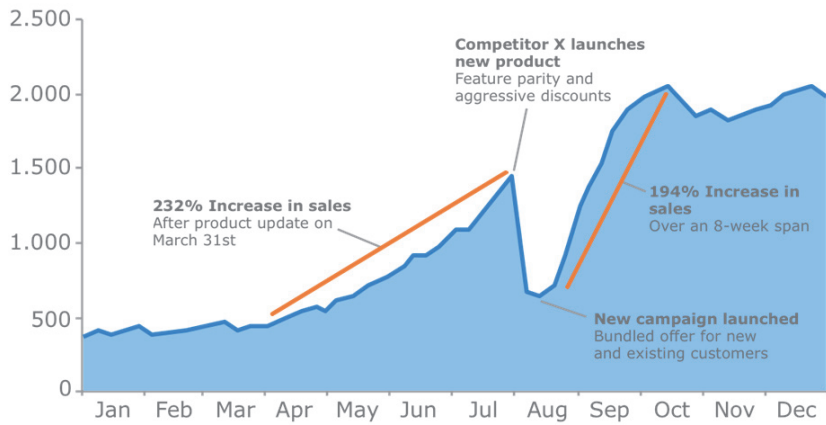
206 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

Visual storytelling principles

FOCUS ATTENTION

Bundled promotion drives 194% increase in sales



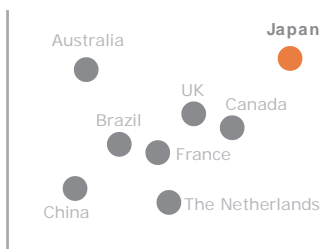
207 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

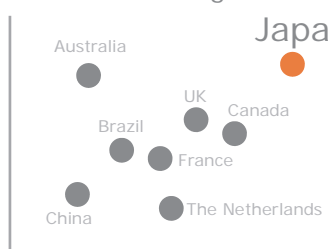
Visual storytelling principles

FOCUS ATTENTION

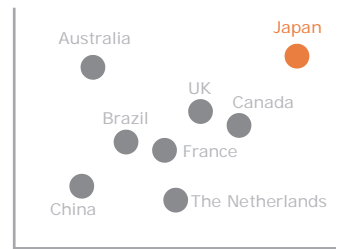
Japan is 35% higher than the next highest country



Japan is 35% higher than the next highest country



Japan is 35% higher than the next highest country

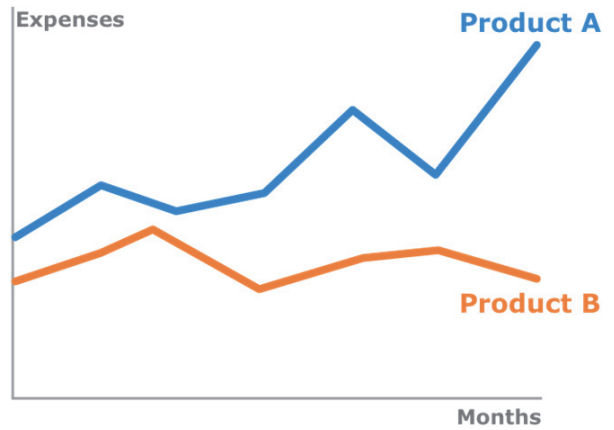
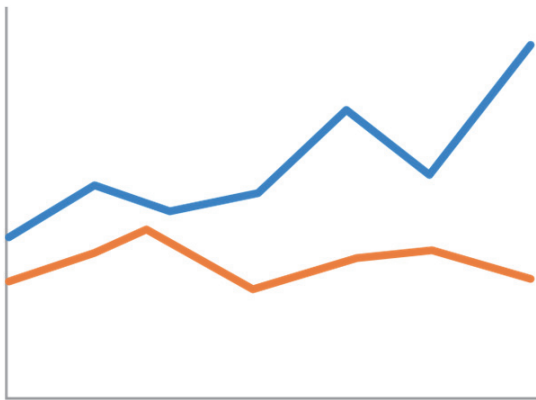


208 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

Visual storytelling principles

MAKE APPROACHABLE

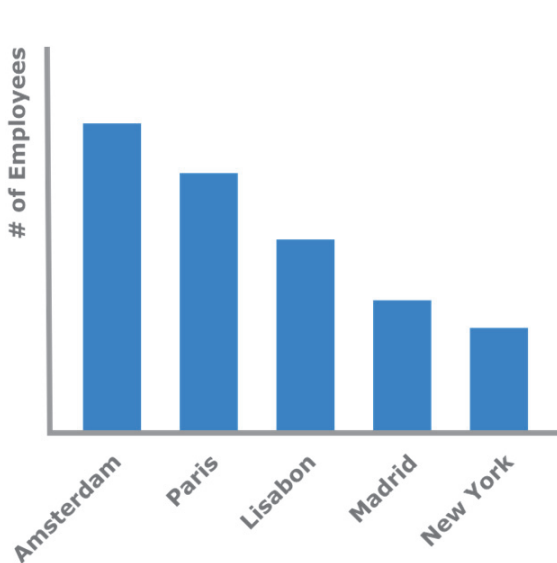


209 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

Visual storytelling principles

MAKE APPROACHABLE

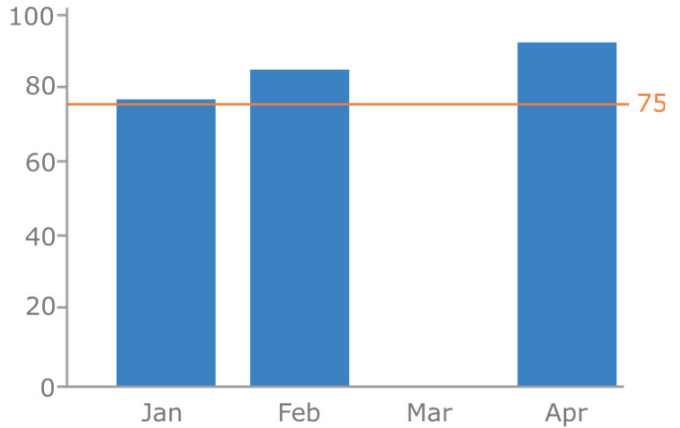
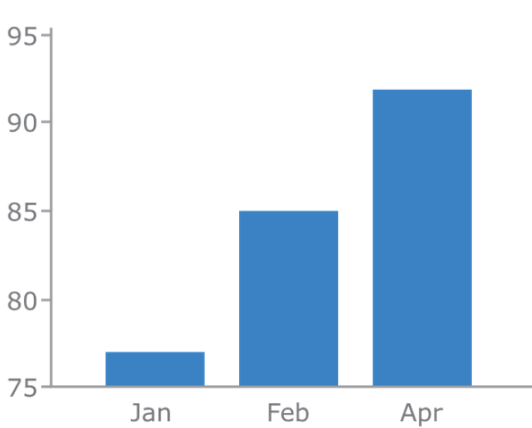


210 Source: Effective data storytelling – Brent Dykes

©2023 Van Haren Publishing BV.

Visual storytelling principles

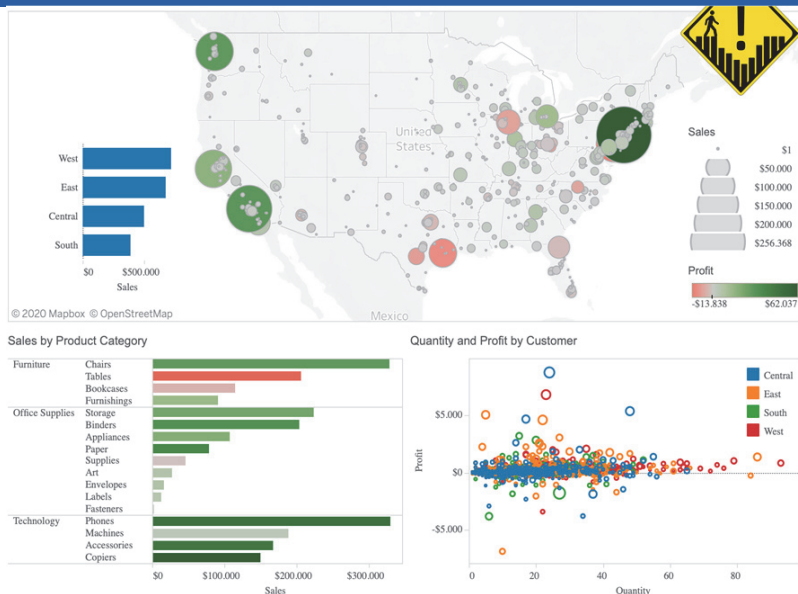
INSTILL TRUST



211 Source: Effective data storytelling – Brent Dykes

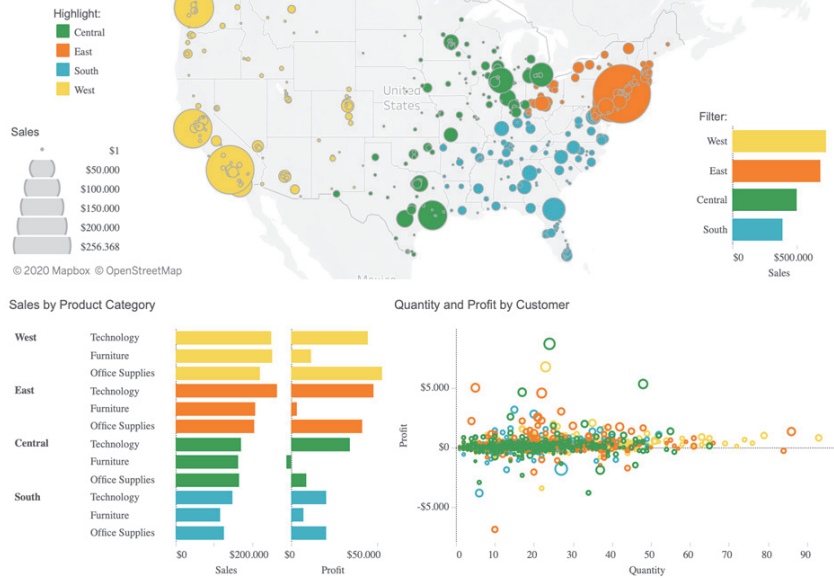
©2023 Van Haren Publishing BV.

Visual storytelling principles



212 Source: https://public.tableau.com/views/SalesConfusingColors/SalesConfusingColors?embed=y&:loadOrder=2&:display_count=y&:showTabory&:origin=viz_share_link/ ©2023 Van Haren Publishing BV.

Visual storytelling principles



Wrap up Argue with Data

 From Explore to Explain

 Design your charts

 Tell a Data Story

3 C's of Data Literacy



Curiosity



Creativity



Critical Thinking

Data literacy

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

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



Read
data



Work
with data



Analyze
data



Argue
with data

Data Literacy

how to read, work with, analyze & argue with data



COURSEWARE



©2023 - All training materials are sole property of Van Haren Publishing BV and are not to be reproduced in any form or shape without written permission.

Effective Data Foundation

not-for-profit collective,
who enables **professionals** to
leverage data to make
sustainable business
decisions



2

©2023 Van Haren Publishing BV.

PRACTICE EXAMS INFO

A practice exam is included within the online exam.

Once you enter the online exam environment, before doing the actual exam. You can do the practice exam test your performance before doing the exam.

The practice exam will give you an indication on how well you master the subject. This is not a guaranty, but an indication.

Be aware, if you solely do the practice exam questions several times you will master the answers rather that the subject. This can provide a wrong indication on how much you master the subject since the exam questions will probably vary from the practice exam.



Data Literacy Certification Syllabus

v. 1.0

TABLE OF CONTENTS

INTRODUCTION	3
DATA LITERACY - THE FOUNDATION OF DATA-DRIVEN DECISION MAKING	3
THE EFFECTIVE DATA FOUNDATION (EDF)	3
INTENDED AUDIENCE	3
THE EDF – DATA LITERACY CERTIFICATE	3
CERTIFYING ORGANIZATION	3
THE EDF – DATA LITERACY EXAM	4
PRACTICAL INFORMATION	4
LEVELS	4
YOUR INVESTMENT	4
PREPARATION AND RECOMMENDED LITERATURE	4
PREPARATION TRAINING	5
TOPICS OF THE EDF DATA LITERACY EXAM	6
EXAM STRUCTURE	6
EXAM TOPICS AND RECOMMENDED LITERATURE	7
TOPIC 1: READ DATA	7
TOPIC 2: WORK WITH DATA	7
TOPIC 3: ANALYZE DATA	7
TOPIC 4: ARGUE WITH DATA	7
EXAM REGULATIONS	8
GENERAL RULES	8
SHARING OF EXAM QUESTIONS IS ILLEGAL	8
FEEDBACK AND QUESTIONS	8

Introduction

Data literacy - the foundation of data-driven decision making

Numbers are central to our understanding of performance. They enable us to make informed decisions. The way we determine success or failure is almost always based on numbers. We derive great value from the stories that numbers tell, yet we rarely consider the significance of how we use them.

Data literacy is an umbrella term to cover all the skills required to understand, work with and share data effectively.

Understanding data requires a set of skills that are easy to learn, but in general far from intuitive. None of us are born with the capacity to understand data: it is a human abstract construct, so we all need to learn how to work with it.

We can learn a lot from our data in order to improve our processes and our lives. But before we get to the value of data, we need to have a better understanding of what data is and what it isn't.

The Effective Data Foundation (EDF)

The Effective Data Foundation (EDF) is a private collaboration that aims to promote the effective use of data.

Intended audience

The Effective Data Foundation – Data Literacy Certification is intended for anyone who uses data in their professional life to improve processes and performance.

Therefore, it is ideal for people for whom the concept of data is relatively new and who wish to become competent in using data. It is also for people already working in this field who wish to strengthen their knowledge and improve their skills.

The EDF – Data Literacy Certificate

This EDF Certification recognizes the awareness and understanding of the components of data literacy and how to foster its adoption and application for the benefit of everybody who needs to make decisions based on data.

The EDF Certification is achieved through an exam which demonstrates that a participant:

- Understands what data is and isn't;
- Is aware of the questions that need answered whenever you receive some data;
- Knows how to summarize data in the right way;
- Is aware of the main processes involved from data origins to usage;
- Understands what data quality is;
- Knows how to use data to measure performance;
- Is aware of the impact of our expectations on data analysis;
- Understands the main human biases involved in data analysis;
- Is aware of the principal types of data analysis;
- Understands the impact of context on analysis;
- Knows how to apply storytelling principles to data arguments

The syllabus outlines the knowledge that the candidate will be tested on during the exam. It also provides suggestions for preparation (background reading) and highlights the benefits of taking this exam.

Certifying Organisation

Van Haren certN

The EDF – Data Literacy Exam

You first need to have successfully completed the Effective Data Foundation Data Literacy Exam to obtain the EDF Data Literacy Certificate. The exam procedure is explained in this section.

Practical information

You must pass a multiple-choice exam in which your knowledge of effective data literacy will be tested to obtain an EDF Data Literacy Certificate. All exam candidates will access the online exam environment and need to answer 60 multiple-choice questions within 60 minutes. To pass, you must answer 65% of the questions correctly (or at least 39 of the 60 questions). Each question has precisely four possible answers, where one or multiple answers are correct. You will receive the result immediately after the exam. (Digital) Access to your certificate will be given once you have passed.

Number of questions:	60
Time (minutes) for the exam:	60 minutes
% Minimal passing grade:	65%
Open or closed book:	Closed
Language:	English
Exam format:	Online
Type of questions	Multiple choice
Are there also negative questions included in the exam (<i>for example: "which of the following is NOT a good data visualization?"</i>)	Yes. Candidates are advised to read the questions carefully

Levels

The EDF Data Literacy Certification tests candidates at levels 1 and 2 according to the Bloom Revised Taxonomy.

Bloom Level 1: Recall & Retention

We test candidates on their ability to memorize factual information, to retain information by collecting, remembering, and recognizing specific knowledge. Knowledge includes facts, terms, answers, or terminology.

Bloom Level 2: Understanding

We test candidates on their ability to construct meaning from oral, written, or graphical pieces of information. This is done by interpreting, summarizing, distracting, comparing, classifying, predicting, or explaining the message.

Your investment

The EDF Exam requires preparation, which means this is an investment in time for personal study and attention covering the subject of effective data literacy. You are completely free to do this in several ways and can consider self-study, reading the reference materials listed in the syllabus, or following a training programme which is designed in line with this syllabus.

Refer to the list of topics in this syllabus. Here you can see which subjects you will be tested on during the exam. The time it takes to prepare for the exam depends on your prior knowledge, experience, and training. Commercially offered training programmes that prepare for the Data Literacy Exam will typically last two to three days. You should allow sufficient time for self-study to address the subjects listed in this syllabus.

Preparation and recommended literature

During your exam preparation, you should familiarize yourself with the concepts of effective data literacy, for example by following a course and reading specified literature. There are ongoing publications about effective data literacy. So, it should be straightforward to find books, articles, blogs, vlogs, or videos about the different aspects.

We include a recommended reading list in this syllabus.

We also advise you to contact people who work with effective data literacy and observe what they do and the techniques they use - and also talk to them.

We have included the following in the syllabus to help you get started:

- Specifications of the examination material - divided into modules.
- The weighting of each individual module towards the overall exam.
- A list of key terms and concepts detailing what must be covered.
- Literature suggestions are available for newcomers in field. Note many of the data literacy concepts have been established for some time and are widely accepted with online and offline reference materials available.
- A practice exam is available online after purchasing an exam. The practice exam contains questions at the same level as the questions in the actual exam. The number of questions may differ from the actual exam. The actual exam includes 60 questions, and you will have 60 minutes to answer them.

Preparation training

We endorse the added value of thoroughly preparing for the Effective Data Literacy Exam and strongly recommend preparatory classroom training, webinars, and online eLearning journeys.

This can help you to understand the essence of effective data literacy and can give you practical examples. That said, it is not mandatory to follow specialized training.

The Effective Data Foundation does not accredit trainers, training institutions or training programmes. The composition and duration, organization, pricing, and execution of the training is the responsibility of the trainer.

Topics of the EDF Data Literacy Exam

In this section, you can read about how the Data Literacy Exam is structured and which subjects you will be tested on as a candidate. It is also a tool that you can use to prepare yourself for the test.

In this syllabus we indicate the topics that are covered in the exam and additional topics which are relevant for further study but are not covered in the exam. During the exam you will be tested on your general knowledge about:

Topic 1: Read data

- What is data?
- Summarize data
- Consume data
- Check your data

Topic 2: Work with data

- Creating data
- Data quality
- Acquiring & cleaning
- Managing data

Topic 3: Analyze data

- Expectations
- Thinking shortcuts
- Types of analysis
- Analytical skills

Topic 4: Argue with data

- Explore to explain
- Selecting the right visualizations
- Storytelling with data

Exam structure

The exam specifications describe the topics in the subject matter of the Data Literacy Exam, and their relative importance. Questions can be asked during the exam about the following subjects.

Topic	% Questions in the exam
1 Read data	25%
2 Work with data	25%
3 Analyze data	25%
4 Argue with data	25%

The following sections specify what knowledge is expected in each of these topics.

Exam topics and recommended literature

Topic 1: Read data

Goals:

- Describe the key properties of data (recall)
- Describe how to summarize data (recall)
- Recognize typical pitfalls when consuming data (comprehend)
- Describe the main questions to check any data expression (recall)

Recommended literature:

- Be Data Literate, Jordan Morrow
- Learning to See Data, Ben Jones
- Data Literacy Fundamentals, Ben Jones
- Thinking Fast and Slow, Daniel Kahneman
- Data Literacy, Peter Aiken & Todd Harbour

Topic 2: Work with data

Goals:

- Describe how data is created (recall)
- Describe the differences between machine and human generated data (recall)
- Describe the differences between mandatory and optional data (recall)
- Recognize the data quality dimensions (comprehend)
- Recognize the best data structure for analysis (comprehend)
- Describe the three basic data cleaning phases (recall)
- Describe how to define performance measures (recall)

Recommended literature:

- Read, Write, Think Data, Ben Jones
- Data Literacy Fundamentals, Ben Jones
- Becoming a Data Head, Alex J. Gutman & Jordan Goldmeier
- Avoiding Data Pitfalls, Ben Jones
- Innumeracy, John Allen Paulos
- Statistical Data Cleaning, Mark van der Loo & Edwin de Jonge
- Practical Performance Measurement, Stacey Barr
- Prove it, Stacey Barr

Topic 3: Analyze data

Goals:

- Recognize the impact of our expectations on our analysis results (comprehend)
- Describe the principal thinking shortcuts (recall)
- Describe the main types of analysis (recall)
- Describe the main analytical skills (recall)

Recommended literature:

- Read, Write, Think Data, Ben Jones
- Becoming a Data Head, Alex J. Gutman & Jordan Goldmeier
- The Signal and the Noise, Nate Silver
- Naked Statistics, Charles Wheelan
- Weapons of Math Destruction, Cathy O'Neal
- Statistical Data Cleaning, Mark van der Loo & Edwin de Jonge

Topic 4: Argue with data

Goals:

- Describe the key properties of the Explore phase (recall)
- Describe the key properties of the Explain phase (recall)
- Recognize the activities belonging to a specific phase (comprehend)
- Describe the typical data forgeries (recall)
- Describe the main steps in the data storytelling arc (recall)

Recommended literature:

- More Judgement than Data, Michael Jones
- Effective Data Storytelling, Brent Dykes

Exam regulations

General rules

An Data Literacy Certification via the Effective Data Foundation is a prestigious title, and fraud is not tolerated. Your exam will be immediately rejected if fraud is found to have been committed during or after completion of the exam. As a result, you will not be reimbursed for your examination fees.

If you fail to pass the exam, you will not receive a certificate. This also means that you must purchase and take a new exam for your certification. Every candidate only gets one attempt per exam to succeed.

Sharing of exam questions is illegal

It is not allowed to share exam questions with others or make them public. This is a violation of the copyright and IP of the Effective Data Foundation and the Certifying Body. Doing so can lead to legal action by the Certifying Body with potentially harmful consequences.

Feedback and questions

We have done our best to help you prepare for the Data Literacy Exam by publishing this syllabus.

We would like to know what you think of this syllabus and the exam. If you have any suggestions for us, we would love to hear from you.

Have fun, take your time preparing for the exam, and good luck. Naturally, we also wish you lots of fun in putting what you've learned into practice!

On behalf of the team – Effective Data Foundation.

Amsterdam, May 2022

