



Garbage in, garbage out: how to assure data quality for visualisation

Direction Access to and Reuse of Public Information

Unit EU Open Data and CORDIS

Sector EU Open Data

The context



This training course is organised in the scope of OP project within the ISA2 programme

ISA2 supports the development of **digital solutions** enabling public administrations, businesses and citizens in Europe to benefit from **interoperable cross-border** and **cross-sector public services**.

How OP is involved in ISA2?

OP is aiming at developing open data related activities in the areas of:

- Data visualisation
- Linked open data
- Persistent identification



Upcoming training & workshop sessions

Topic	Type of session	Lux + webex	Bxl
Garbage in, garbage out: how to assure data quality for visualisation	Training	30/04	13/05
Exploratory data analysis through data visualisation (R-ggplot2)	Training	24/05	04/06
Telling your story through data visualisation	Training	25/06	28/06
Making great online data visualisations without coding	Workshop	26/06	-
Going beyond bars and lines: practising non-standard data visualisation	Training	24/09	Sep-Oct
Making data visualisations like a pro: D3.js	Workshop	25/09	-
Applying data visualisation best practices in real use cases	workshop	24/10	-

and [webinars](#) (topic like for the trainings) ... stay tuned!

Materials will be published on <https://data.europa.eu/euodp/en/knowledge-center>



Data visualisation events in 2019



EU Datathon 2019

13 June 2019

Residence Palace - Brussels

<https://publications.europa.eu/eudatathon>
op-datathon@publications.europa.eu



EU DataViz 2019

12 November 2019

European Convention Center - Luxembourg

<https://publications.europa.eu/eudataviz>
op-eu-dataviz@publications.europa.eu



Where to find the information about our training workshop and webinar sessions?

Visit our Data Visualisation Community to find all the information about previous and upcoming sessions at:

<https://webgate.ec.europa.eu/fpfis/wikis/display/EUODDVC/Data+Visualisation+-+Training+Package+2019>



Agenda

- 09:00 Introduction
 - Pitfalls in data quality
- 10:30 Coffee break
 - Metadata
- 12:00 Lunch
- 13:00 Assessing and measuring data quality
 - Exercise: improving data quality
- 14:30 Coffee break
 - Transforming data for visualisation
 - ETL and data management
- 16:30 Q&A



1. INTRODUCTION





Participants

Institution/DG and role?
What data do you work with?
Expectations for today?






2.

PITFALLS IN DATA QUALITY FOR VISUALISATION





"Garbage in, garbage out" describes the concept that flawed, or nonsense input data produces nonsense output or "garbage".

[Wikipedia](#)

Garbage in, garbage out. Or rather more felicitously: the tree of nonsense is watered with error, and from its branches swing the pumpkins of disaster.

Nick Harkaway, *The Gone-Away World*



1 - 10 - 100 rule for data quality



Data quality pitfalls

Incorrect interpretation

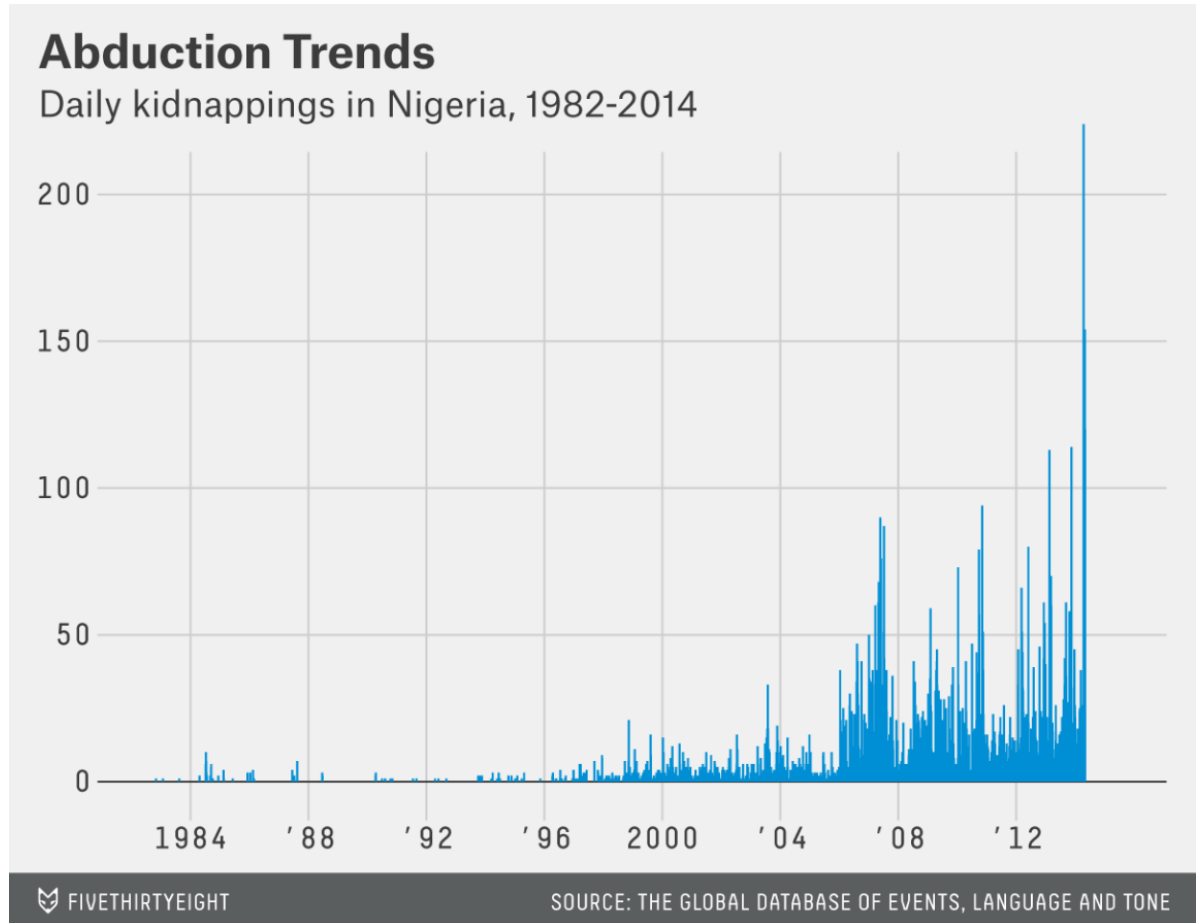
Why does Popeye eats spinach to get stronger?



Data quality pitfalls

Incorrect interpretation

Case: Fivethirtyeight and Nigerian kidnappings



Data quality pitfalls

Wrong precision

Case: Why did people keep trying to recover their lost or stolen phones in this house in Atlanta?

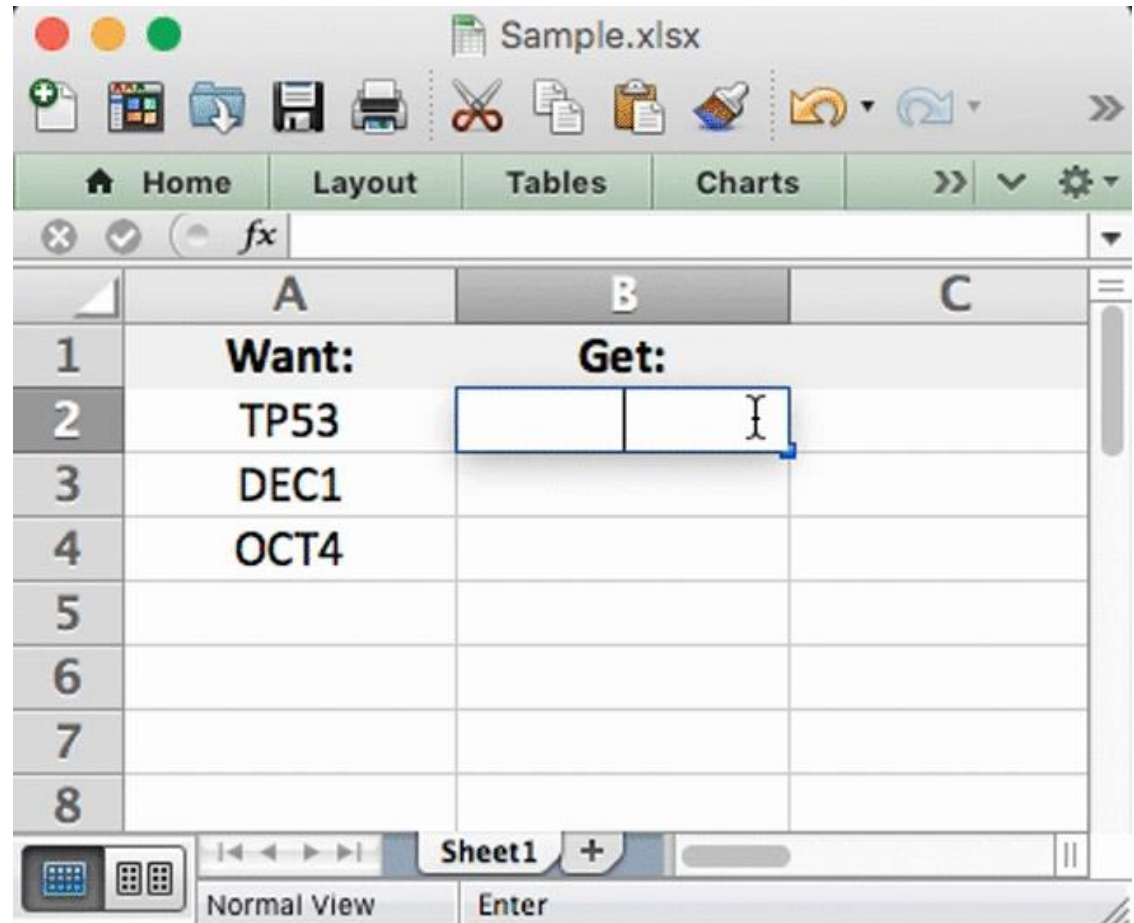


Data quality pitfalls

Transformation errors

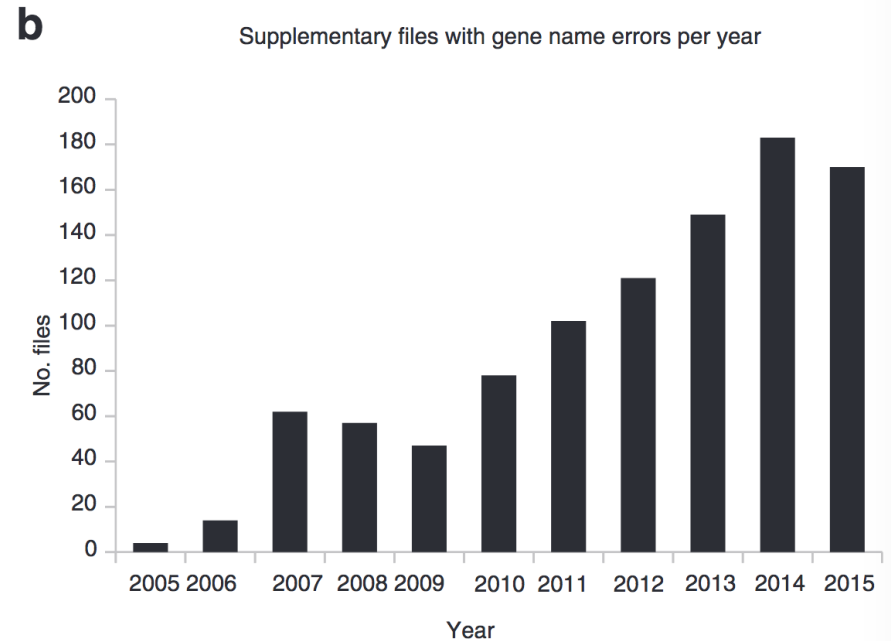
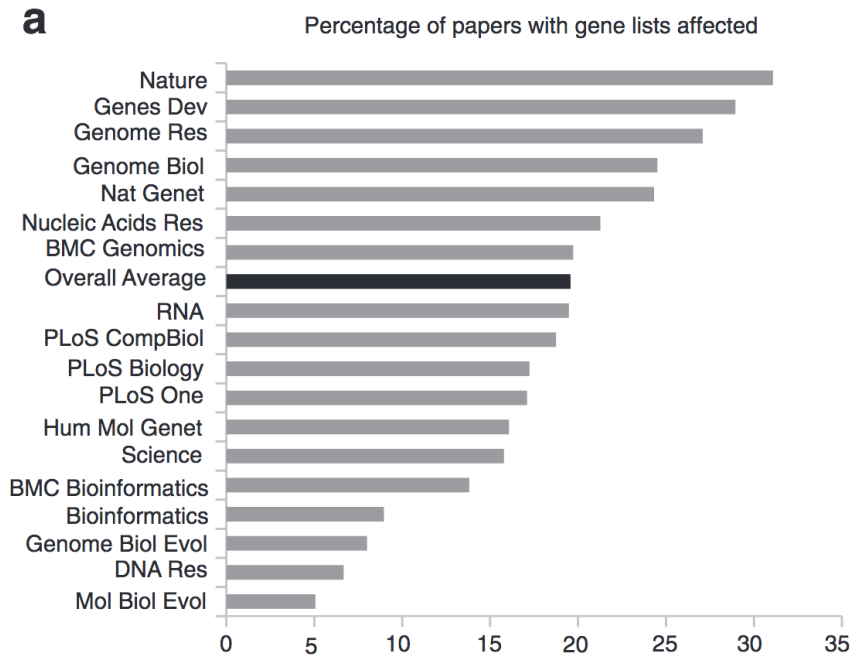
What will happen with the names of genes like “SEPT2” and “MARCH2” in Microsoft Excel?

What will happen to an identifier like “2310009E13”?



Data quality pitfalls

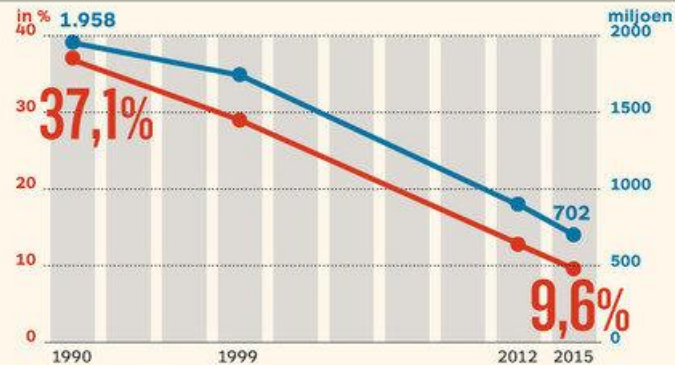
Transformation errors



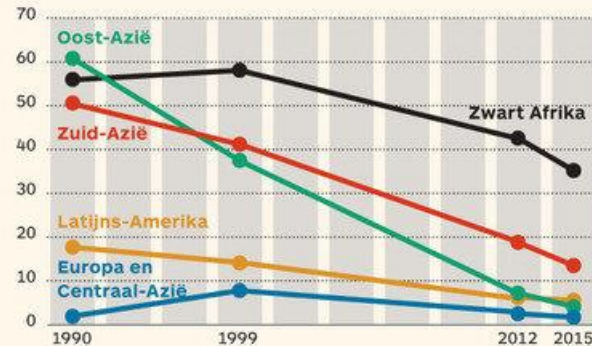
Data quality pitfalls Incorrect data types

What went wrong with
this chart?

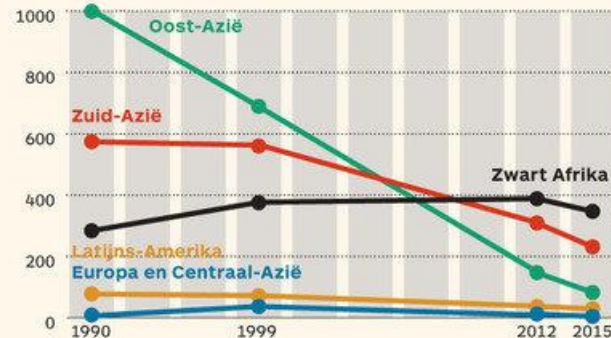
EXTREME ARMOEDE BEDRAAGT NOG SLECHTS
EEN VIERDE VAN KWARTEEUW GELEDEN



IN ZWART AFRIKA NEEMT EXTREME ARMOEDE ECHTER
VEEL MINDER SNEL AF (in procent)



IN ABSOLUTE CIJFERS STEEG HET AANTAL AFRIKANEN
DAT IN EXTREME ARMOEDE LEEFT (in miljoen)



Bron: Wereldbank

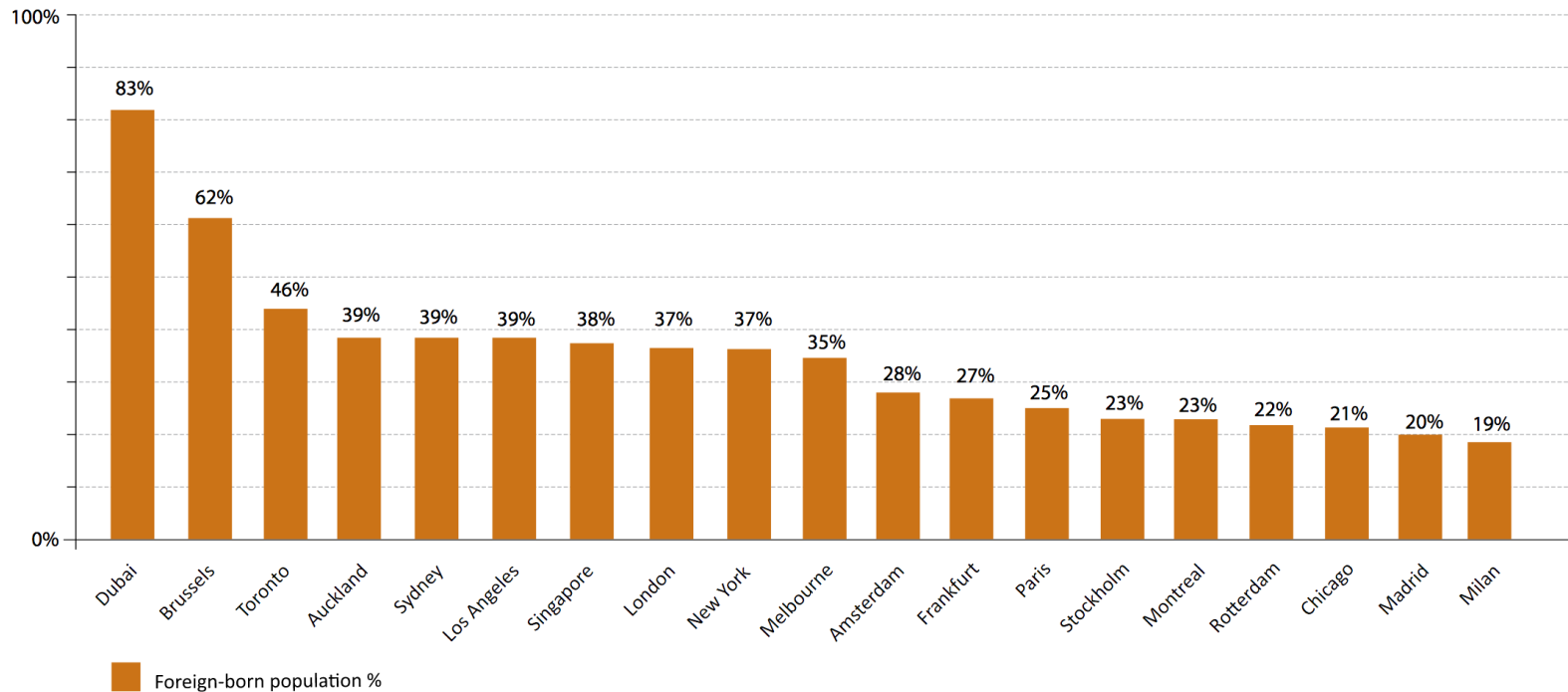


Data quality pitfalls

Definition mismatch



Figure 1 Foreign-born population in major cities



Source: Compiled by IOM from various sources – see list at the end of the References section.

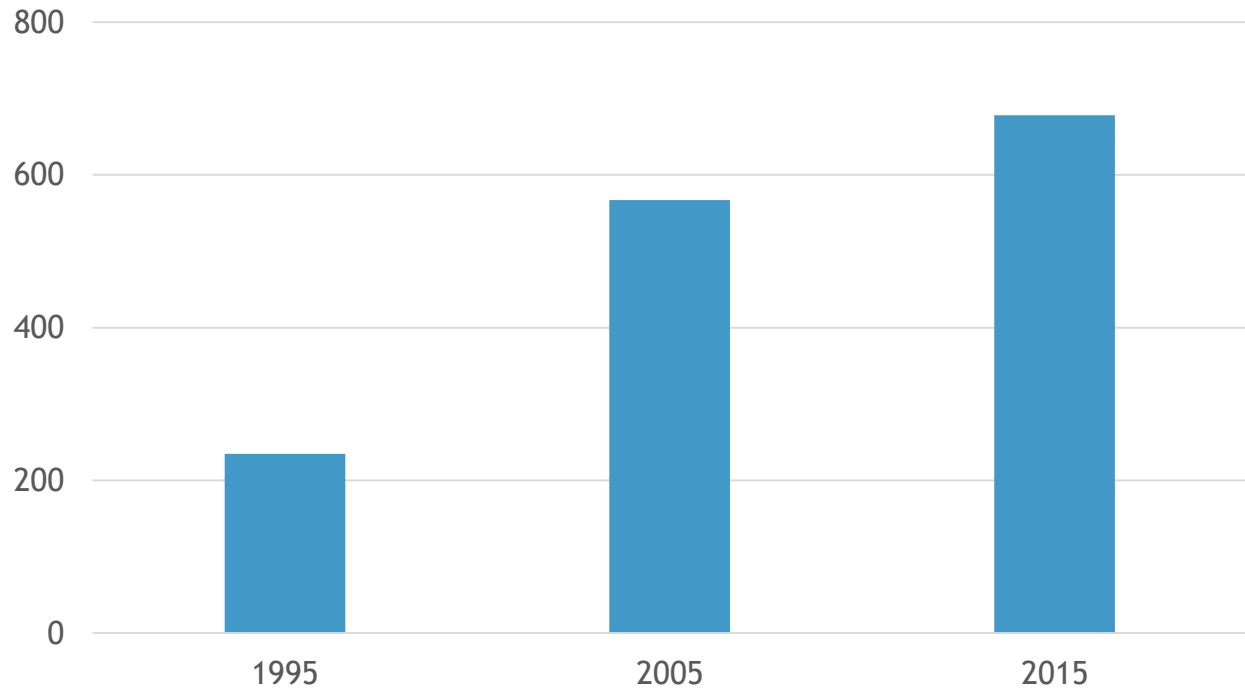


Data quality pitfalls

Definition mismatch

What is wrong with this data?

Year	Pollution	Region
1995	235	EU
2005	567	EU
2015	678	EU



Data quality pitfalls

Definition mismatch

Unemployed person: Eurostat

An **unemployed** person is defined by Eurostat, according to the guidelines of the International Labour Organization, as:

- someone aged 15 to 74 (in Italy, Spain, the United Kingdom, Iceland, Norway: 16 to 74 years)
- without work during the reference week
- available to start work within the next two weeks (or has already found a job to start within the next three months)
- actively having sought employment at some time during the last four weeks

The **unemployment rate** is the number of people unemployed as a percentage of the labour force.



Data quality pitfalls

Definition mismatch

Unemployed person: INSEE

Unemployed persons for the purposes of the population census are persons:

- aged 15 or over
- who declared themselves to be unemployed (either registered or not with Pôle Emploi) unless they have also explicitly declared that they were not looking for work
- and on the other hand the persons (of 15 or more years old) who declared themselves spontaneously neither in employment, nor unemployed, but who nevertheless declared to look for an employment.

Note

An unemployed person (population census) is not necessarily an unemployed person in the sense of ILO.



3.

METADATA



Metadata

What is it?



Metadata
=
Data about the data

How was data collected?

Who collected the data?

For what purpose was the data collected?

What definitions are used?

What units are the data expressed in?

What constraints does the data have?



Metadata

What is it?

Operational metadata

How was data accessed?

How was data transformed?

Particularly important for time series and financial data:

- Corrected for seasonality?
- Adjusted for inflation?
- Adjusted for purchasing power?



Metadata

What is it?

Technical metadata

File formats

Software used

Hardware used



Metadata

Why does it matter?

It makes data manageable

Process data

Maintain data

Integrate data

Share data

Audit data



Metadata

Why does it matter?

Without metadata, an organisation...

- ...is like a library without a card catalogue
- ...cannot manage its data as an asset
- ...cannot manage its data at all



Metadata

Why does it matter?

It makes data interpretable

What does the data represent?

What are the limitations?

Definitions establish a common vocabulary



Exercise

Reading metadata

1. Pick a data set from ec.europa.eu/eurostat/data/database
2. Go to “Explanatory text (metadata)”
3. Answer following questions, and think about the consequences for the use of the numbers in visualisation and communication

How was data collected?

For what purpose was the data collected?

What definitions are used?

Operational metadata?

Technical metadata?

What are the limitations (conclusions, integration with other sources, etc.) of the data?



4.

ASSESSING AND MEASURING DATA QUALITY



Data quality

Encoding

◆ ◆ ◆ ◆
æ-řå-åŒ-ã•
◆ ◆ ◆ ◆ ◆ ◆ ◆ ◆
U+00B6

Encodings map bits (1's and 0's) to characters

When the wrong encoding is used, weird characters are introduced, or text may be completely nonsensical

UTF-8 is the most used encoding

Sometimes the cause of weird characters is fonts missing some characters



Data quality

Missing values

“”
NA
-
X
-9999
0

Are missing values present in the data?
How are missing values encoded?
What does a missing value represent?
How to handle missing data?





Data quality

Duplicate records

If duplicate records are present: are they expected? How should they be handled?

If something seems like it should be unique (like a column named “ID”), verify that it is



Data quality

Unstandardised data

GB

Gr. Britain

Great Britain

Great Brittain

great britain

GREAT BRITAIN

“There is no worse way to screw up data than to let a single human type it in, without validation.”

Typing errors

Synonyms

Difference in capitalisation

Difference in units (MW vs GW)

Difference in formatting (1 March, 2019 vs 2019-03-01)





Data quality

Ambiguity in data

Column names may not unambiguously indicate what data they contain and in what units data is expressed

Don't assume what data represents and what the units are: consult the metadata



Data quality

Suspicious values

65535

255

9999

0000

1970-01-01T00:00:00Z

January 1st, 1900

0°N 0°E

All of these are indications of errors made by humans or computers.





Data quality

Outliers

Very high or very low numbers

Strings/categories occurring only once in the data

Sort or visualize the data to check for outliers

Do the outliers make sense?

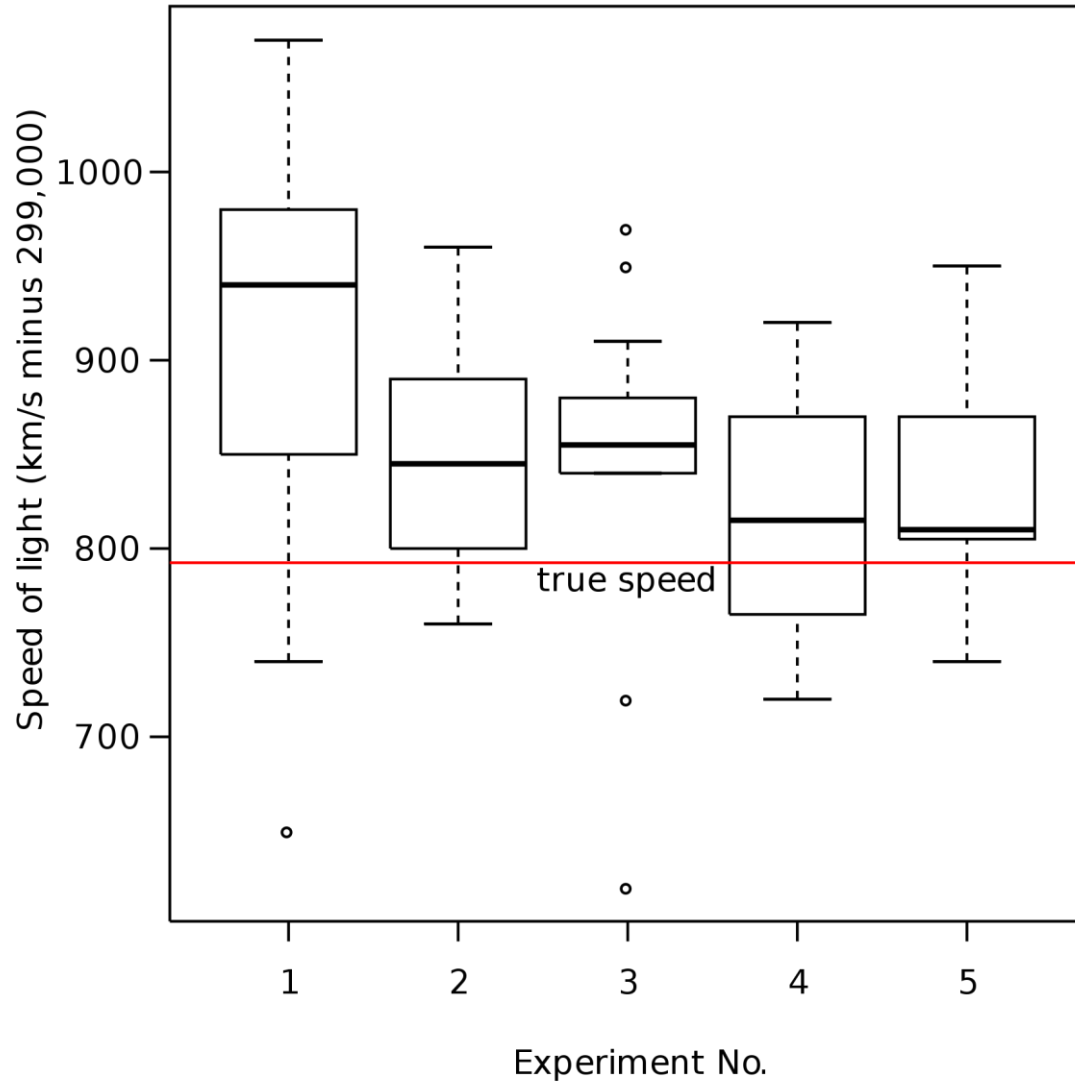
Are the outliers representing errors in the data?

Would it make sense to remove them from the data for analysis and visualisation?



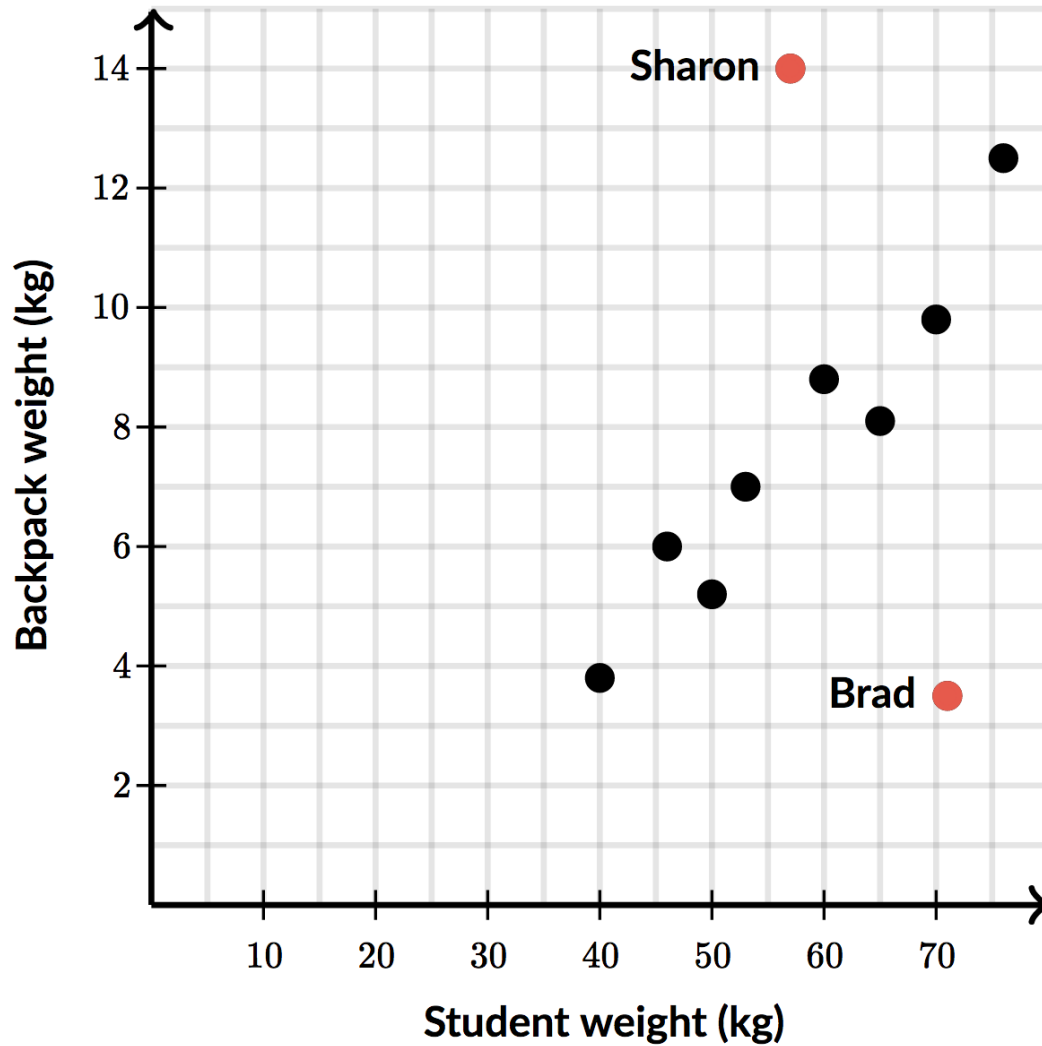
Data quality

Outliers

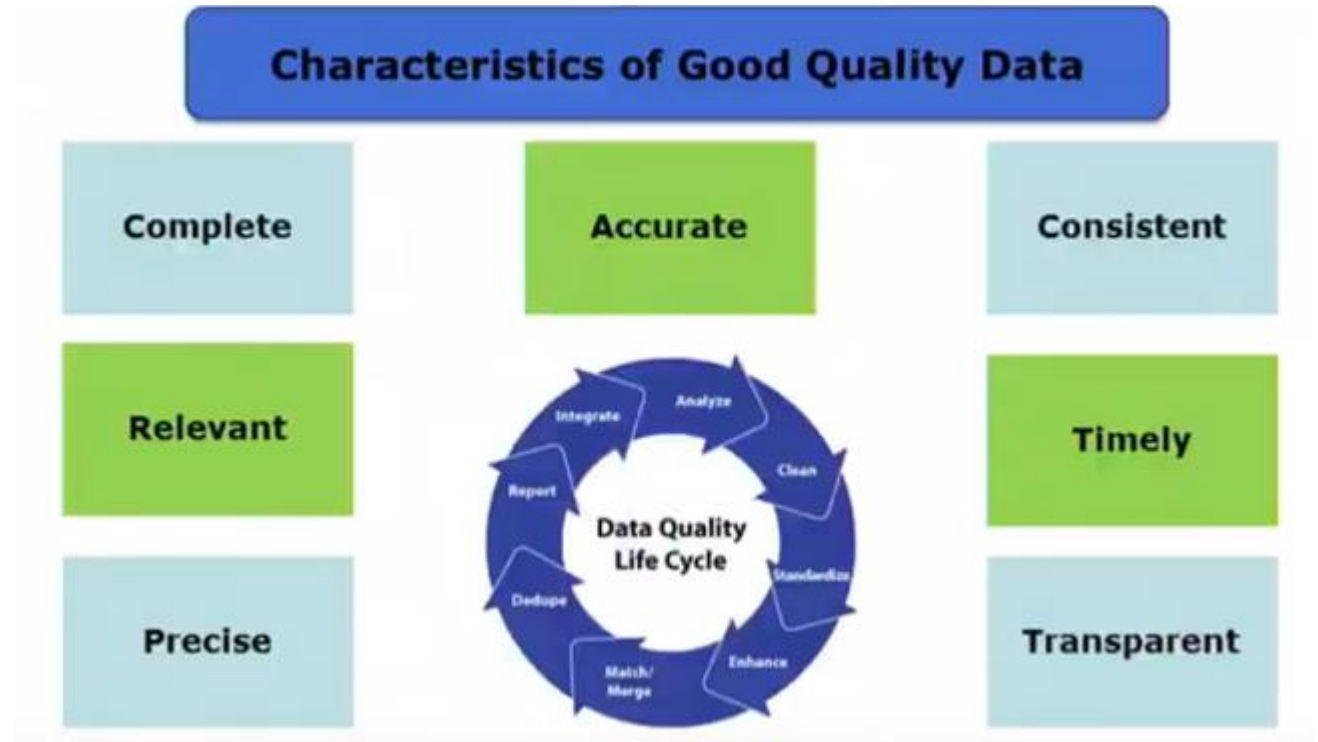


Data quality

Outliers



Data quality Characteristics



Data quality Data bullet proofing

After doing calculations or analysis:

Perform spot checks: pick some records, recalculate new values manually

Alternative calculation method: can you arrive at the same number (average, total, etc.) by taking a different calculation path?



Data quality

Data profiling

Collect information about the data and assess its quality

Detect anomalies and inconsistencies

Understand data structure and relations

Assess usage of the data

Number of records, number of columns?

Duplicates?

Missing values: encoding, number?

Data type of each column

Distributions and frequencies

Ranges: min and max values, outliers

ID (or “key”) column?

Read metadata



Data quality

Data profiling

Overview Output Destinations **Profile** Dependencies Data sources

All data



Results profile by column

#	FMID	ABC	MarketName	Primary_Website_or_URL	ABC	Address	ABC	Location_Description	Seasons		
Valid	38	Valid	38	Valid	22	Valid	38	Valid	38		
Mismatched	0	Mismatched	0	Mismatched	4	Mismatched	0	Mismatched	0		
Empty	0	Empty	0	Empty	12	Empty	0	Empty	0		
<p>Minimum 1,000,059 Lower quartile 1,001,683 Median 1,004,332 Upper quartile 1,005,310 Maximum 1,008,391</p>		<p>Top 20 values</p> <ul style="list-style-type: none"> 58th and Chester Farmer... 1 57th Street Greenmarket 1 52nd and Haverford Farm... 1 3rd Street N (hwy 11) b... 1 3rd & Curry St. Farmers... 1 38th & Meridian Farmers... 1 33rd and Diamond Farmer... 1 32nd Street/Waverly Far... 1 3 French Hens French Co... 1 29th and Wharton Farmer... 1 25th Avenue Farmers' Ma... 1 22nd and Tasker Farmers... 1 2012 Wood County Farmer... 1 17th Ave Market 1 175th Street Greenmarket 1 		<p>Top 14 values</p> <ul style="list-style-type: none"> http://www.grow NYC.org 6 http://www.foodtrustmar... 4 http://www.sandlercente... 1 http://www.peoplesfoodc... 1 http://www.pcfma.com/ma... 1 http://www.iatp.org/min... 1 http://www.foodtrustmar... 1 http://www.experimental... 1 http://www.carsonfarmer... 1 http://www.abquptowngro... 1 http://www.abingtonsage... 1 http://www.abingdonfarm... 1 http://www.6701burnetro... 1 http://www.3frenchhensm... 1 		<p>Top 20 values</p> <ul style="list-style-type: none"> 507 Harrison Street, Kal... 1 3rd St N (Hwy 11) besid... 1 3rd & Curry Street, Cars... 1 3808 North Meridian Str... 1 362 Plymouth Street, Abi... 1 29th and Wharton Street... 1 27 W Potomac Street, Bru... 1 2349 S Hwy 127, Russell ... 1 22nd and Tasker Streets... 1 201 Market Street, Virgi... 1 1st Ave - E 92nd & 93 S... 1 194 W 25th Avenue, San M... 1 1622 6th St NE, Minneapo... 1 1400 U Street NW, Washin... 1 12th & Brandywine Stree... 1 		<p>Top 7 values</p> <ul style="list-style-type: none"> Other 16 Private business parki... 8 Local government build... 7 Faith-based instituti... 3 Closed-off public stre... 2 On a farm from: a bar... 1 Co-located with wholes... 1 		<p>Top 20 values</p> <ul style="list-style-type: none"> {"Date":"","Time":""} 115 {"Date":"","Time":"Tu... 3 {"Date":"06/17/2013 t... 1 {"Date":"06/06/2013 t... 1 {"Date":"06/06/2013 t... 1 {"Date":"06/06/2013 t... 1 {"Date":"06/06/2013 t... 1 {"Date":"05/11/2013 t... 1 {"Date":"05/11/2013 t... 1 {"Date":"05/04/2013 t... 1 {"Date":"05/01/2013 t... 1 {"Date":"04/20/2013 t... 1 {"Date":"04/20/2013 t... 1 {"Date":"01/01/2013 t... 1 {"Date":"01/01/2013 t... 1 	



Exercise

Improving data quality in Excel

Some common data quality issues and what to do about them

Download

learno.net/uploads/downloads/reactors.csv

Follow the steps to clean the data and improve
data quality

Exercise

Improving data quality in Excel

1. Metadata

www.world-nuclear.org/information-library/facts-and-figures/reactor-database-data/reactor-database-search.aspx

Exercise

Improving data quality in Excel

2. Encoding

Data => Get External Data
=> From Text

Set encoding with “File origin”

Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the Data Type that best describes your data.

Delimited - Characters such as commas or tabs separate each field.
 Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: File origin:

Preview of selected data:

Preview of file /Users/maarten/Documents/clients/trasys/modules/T2-.../reactors.csv.

1	"url", "Name", "Name_link", "Country", "Capacity", "Status", "Type", "Owner", "Operator", "Connection"
2	"http://www.world-nuclear.org/information-library/facts-and-figures/reactor-database.aspx?source=%7B%22qu
3	"http://www.world-nuclear.org/information-library/facts-and-figures/reactor-database.aspx?source=%7B%22qu
4	"http://www.world-nuclear.org/information-library/facts-and-figures/reactor-database.aspx?source=%7B%22qu
5	"http://www.world-nuclear.org/information-library/facts-and-figures/reactor-database.aspx?source=%7B%22qu
6	"http://www.world-nuclear.org/information-library/facts-and-figures/reactor-database.aspx?source=%7B%22qu
7	"http://www.world-nuclear.org/information-library/facts-and-figures/reactor-database.aspx?source=%7B%22qu
8	"http://www.world-nuclear.org/information-library/facts-and-figures/reactor-database.aspx?source=%7B%22qu
9	"http://www.world-nuclear.org/information-library/facts-and-figures/reactor-database.aspx?source=%7B%22qu

Cancel < Back Next > Finish

Exercise

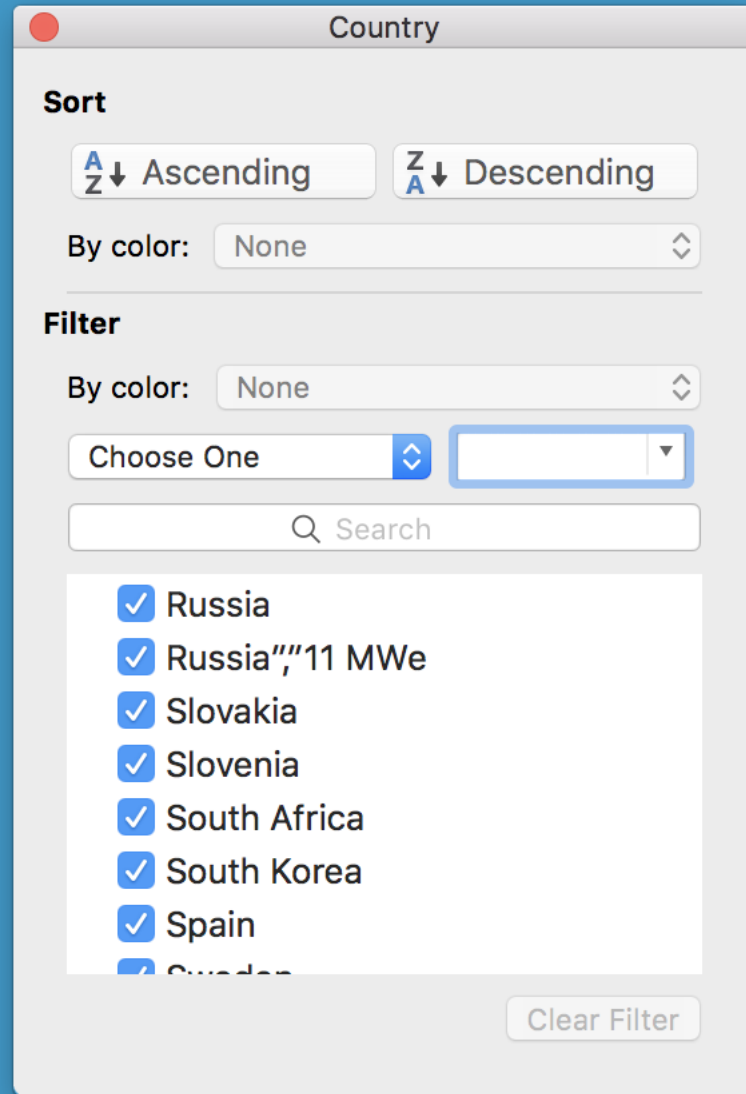
Improving data quality in Excel

3. Inspection

Click column header => number of rows is displayed

Visual inspection: formats, alignment (numbers are right aligned, text left)

Filter => see unique values in each column (check Blanks)



Exercise

Improving data quality in Excel

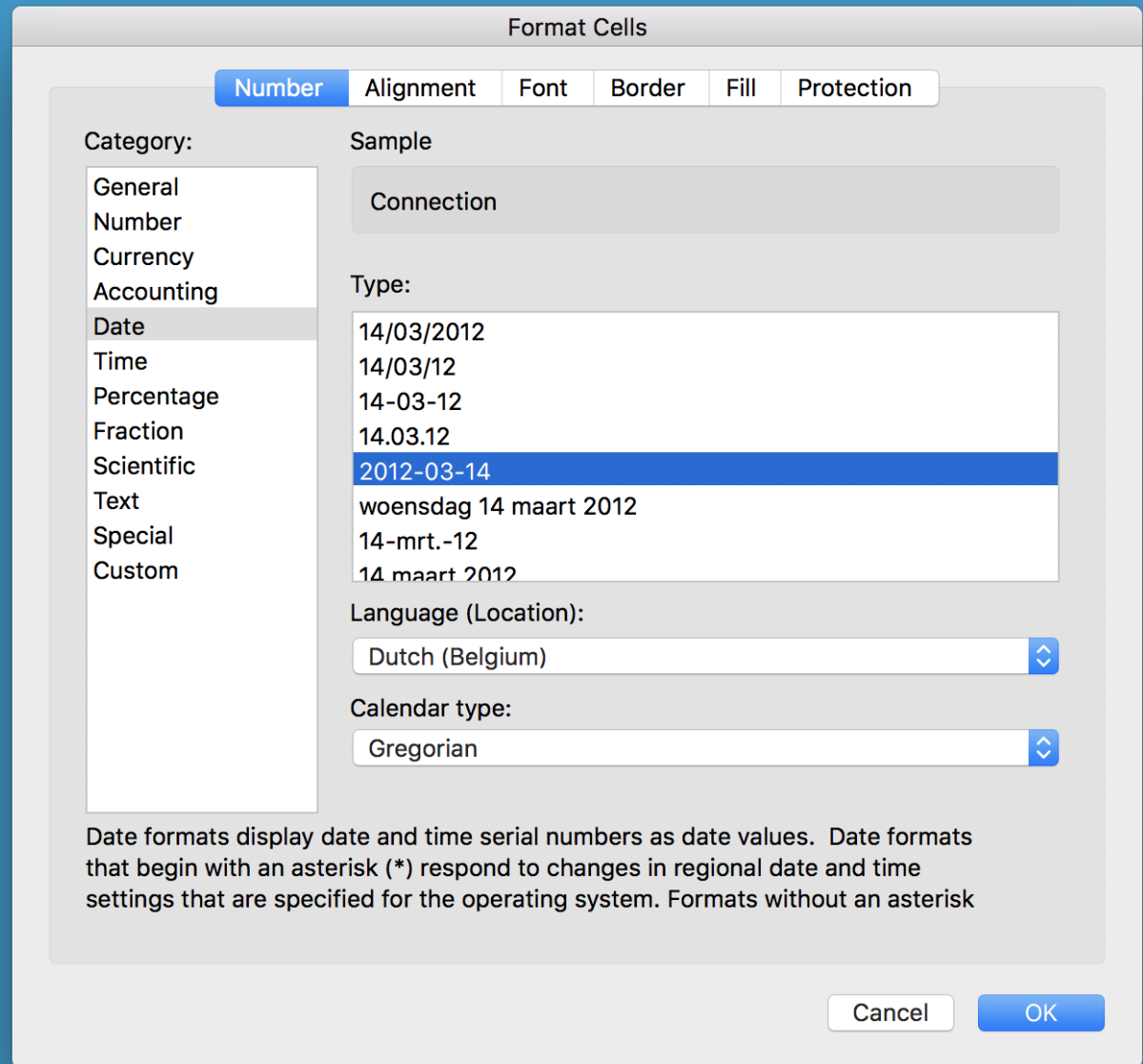
4. Formatting

Useful dataformat is the ISO format (yyyy-mm-dd)

Select cells => right click
=> format cells

Formatting numbers can also be helpful (thousand separator, decimal places ,...)

Formatting does not alter the stored data!



Exercise

Improving data quality in Excel

5. Formulas

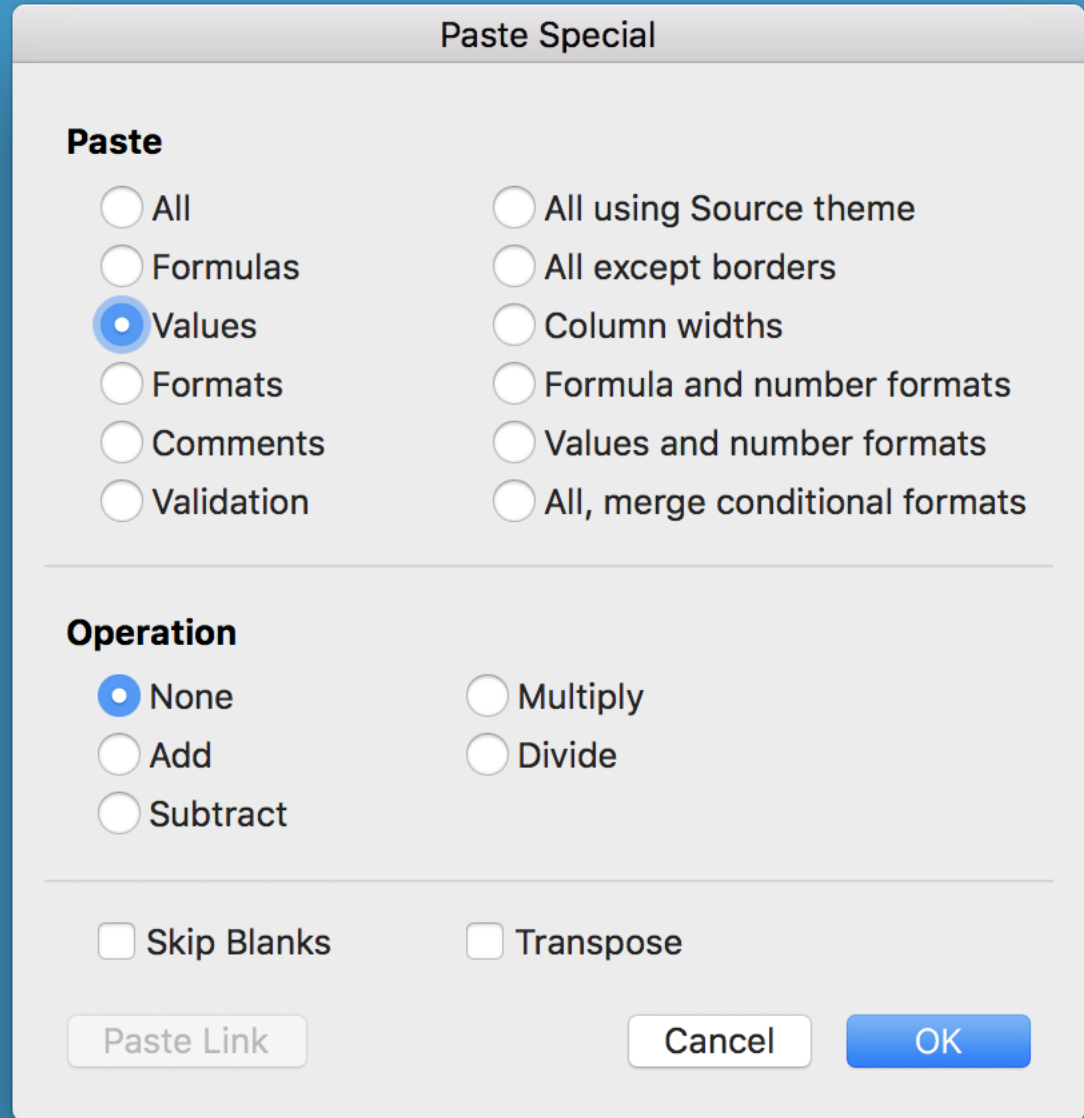
Start with “=”

=year() extracts the year from a date

=proper() capitalises first letter of every word in text

=left(), =right() gets the first and last characters of text

To convert formulas to actual values: Copy => Paste special => Values



6. TRANSFORMING DATA FOR VISUALISATION



Data preparation

Deriving measures

Calculate averages or medians

=average(), =median()

Calculate indices

=value/startingvalue*100

Normalise data: per capita, etc.

=\$/population



Data preparation

Data normalisation

Data normalisation often involves joining data (eg. Country population joined to electricity consumption allows to calculate the per capita electricity consumption)

Glue datasets together based on a common key

Left, right, inner and outer joins

Excel: vlookup()



Data preparation

Data aggregation

Geographically

In time

By category

Sum, count, average,
etc.

Aggregation = grouping records
and calculating measures at the
group level

Excel: pivot tables



Exercise

Preparing data for
visualisation

Pivot tables and aggregation

Top 10 of the biggest nuclear power plants

Top 10 of countries with the highest nuclear
power capacity

Timeline of installed capacity

Exercise

Improving data quality in Excel

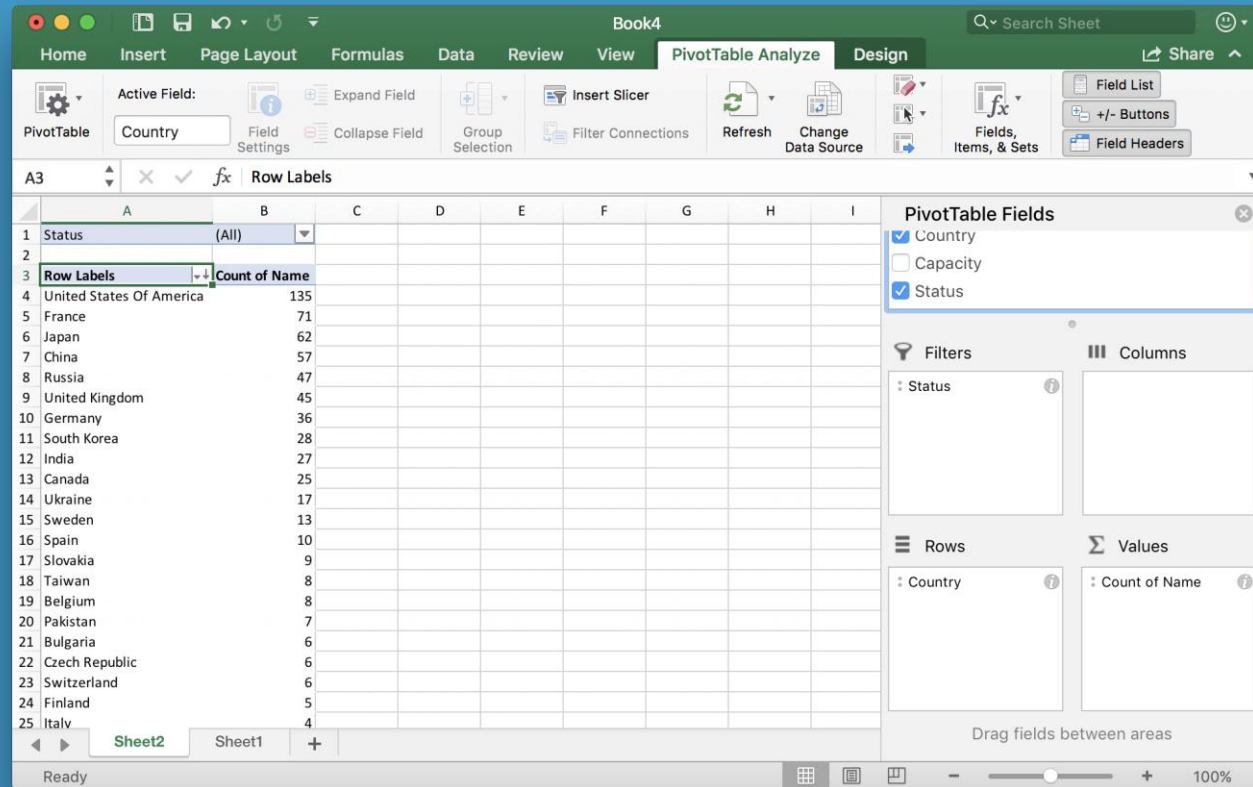
6. Aggregating

Grouping records, summarise data for each group

Insert => Pivot Table

Count, average, sum

Filter data





7.

ETL AND DATA MANAGEMENT



ETL

What is it?

Extract
Transform
Load

Consolidate multiple data sources into 1

Transform includes cleansing data

Automate the data processing



ETL Solutions on the market

Open source-generic vs custom-commercial

On premise vs cloud

Coding vs graphical UI

Batch processing vs streaming

Figure 1. Magic Quadrant for Data Integration Tools



Source: Gartner (July 2018)



Q&A



Resources

Readings about data quality

[The Quartz guide to bad data](#)

[Avoiding Garbage In, Garbage Out: The Importance of Data Quality](#)

[Applying the 1-10-100 rule to your data quality](#)

[What is metadata and why is it as important as the data itself?](#)

Excel

[Excel functions](#)

[Excel VLOOKUP function](#)

[Excel PivotTable](#)

[Cleaning data in Excel video course](#)



Upcoming training & workshop sessions

Topic	Type of session	Lux + webex	Bxl
Garbage in, garbage out: how to assure data quality for visualisation	Training	30/04	13/05
Exploratory data analysis through data visualisation (R-ggplot2)	Training	24/05	04/06
Telling your story through data visualisation	Training	25/06	28/06
Making great online data visualisations without coding	Workshop	26/06	-
Going beyond bars and lines: practising non-standard data visualisation	Training	24/09	Sep-Oct
Making data visualisations like a pro: D3.js	Workshop	25/09	-
Applying data visualisation best practices in real use cases	workshop	24/10	-

and [webinars](#) (topic like for the trainings) ... stay tuned!

Materials will be published on <https://data.europa.eu/euodp/en/knowledge-center>

