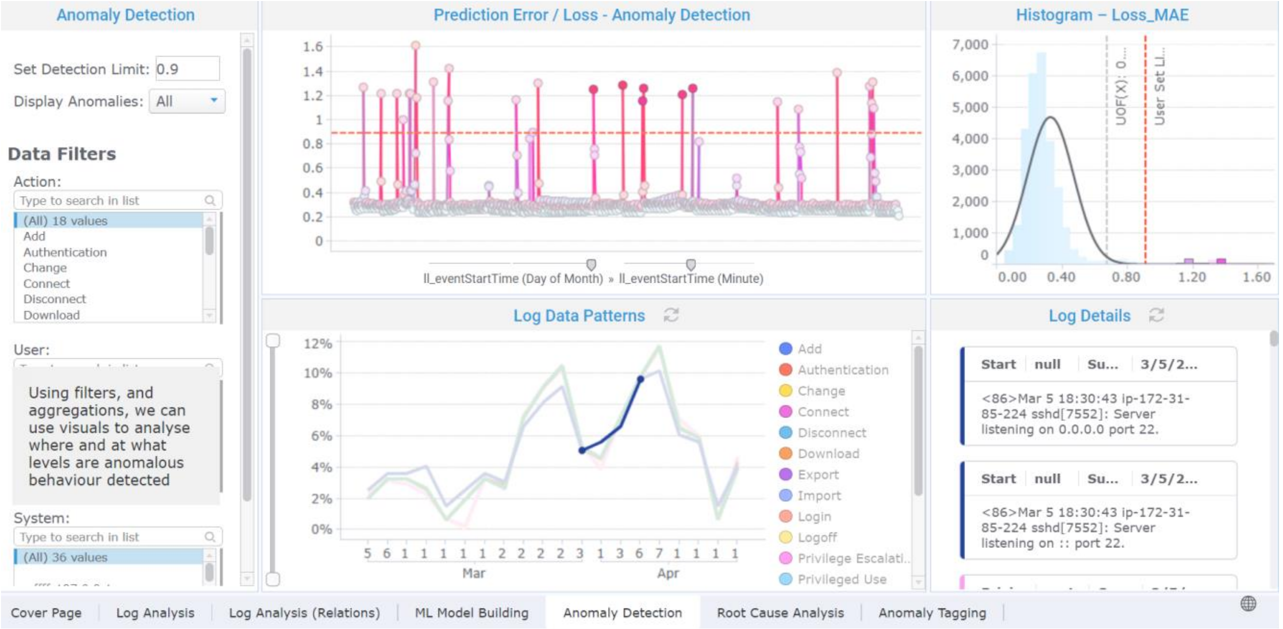
Universidade de Aveiro
Departamento de Electrónica,
Telecomunicações e Informática

Data



Set Detection Limit: 0.9

Display Anomalies: All

Using filters, and aggregations, we can use visuals to analyse where and at what levels are anomalous behaviour detected

Start | null | Su... | 3/5/2...

<86>Mar 5 18:30:43 ip-172-31-85-224 sshd[7552]: Server listening on 0.0.0.0 port 22.

Start | null | Su... | 3/5/2...

<86>Mar 5 18:30:43 ip-172-31-85-224 sshd[7552]: Server listening on :: port 22.

Anomaly Detection

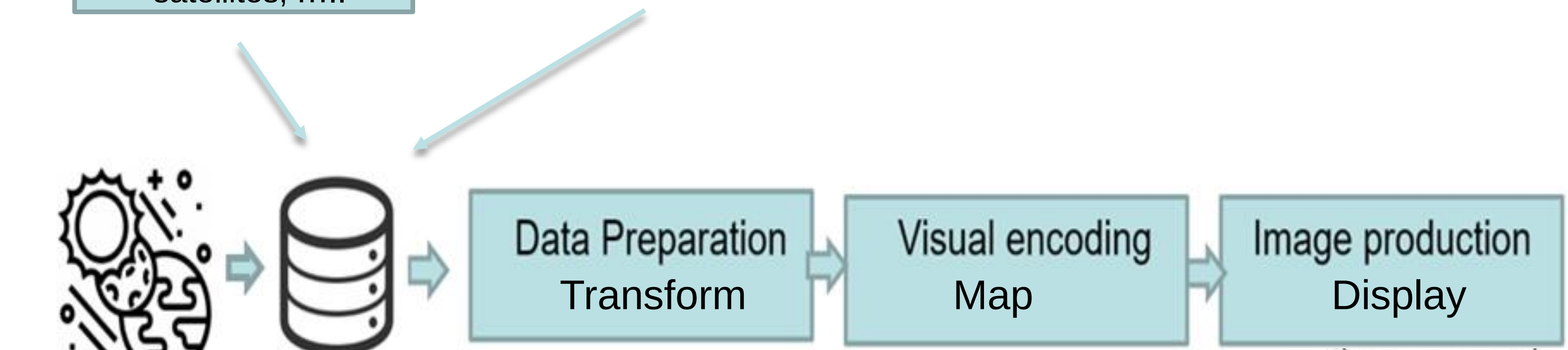
Beatriz Sousa Santos, University of Aveiro, 2025

Data is a proxy to the phenomena to analyse and understand

Measured data:

CAT, MR, sensors,
ultra sound,
laser digitizers,
satellites, …..

Simulated data:
Finite Element
Analysis, Numeric
methods, ……



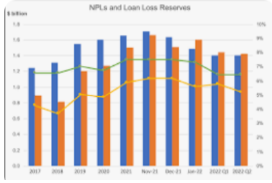
Data Preparation
Transform

Visual encoding
Map

Image production
Display

Phenomenon
World/simulation

Data



Adequate data pre-processing is vital!

- Data may have a lot of different forms and there are many techniques and systems to visualize them

- A data classification is important to:

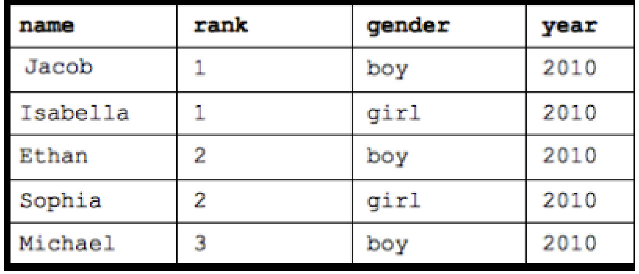
  - predict what visualization techniques are adequate

  - make easier the communication about the data

  - allow a more systematic approach to Visualization

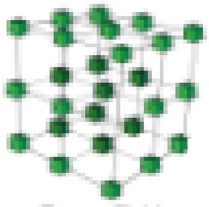
  ….

Data Abstraction

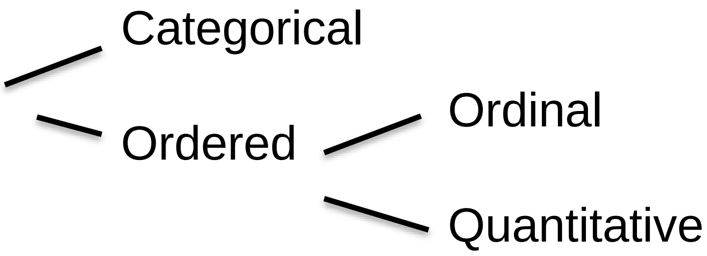


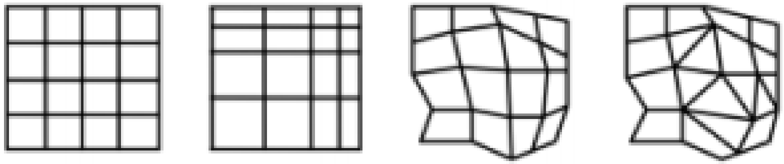
name	rank	gender	year
Jacob	1	boy	2010
Isabella	1	girl	2010
Ethan	2	boy	2010
Sophia	2	girl	2010
Michael	3	boy	2010

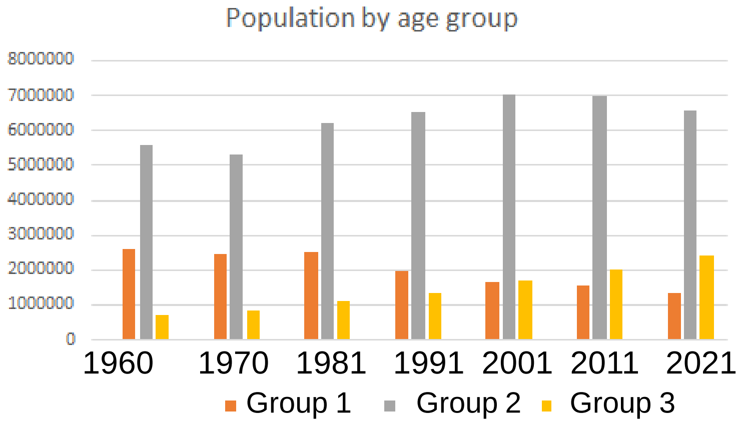
- Four basic dataset types:
  - Tables
  - Networks
  - Fields
  - Geometry



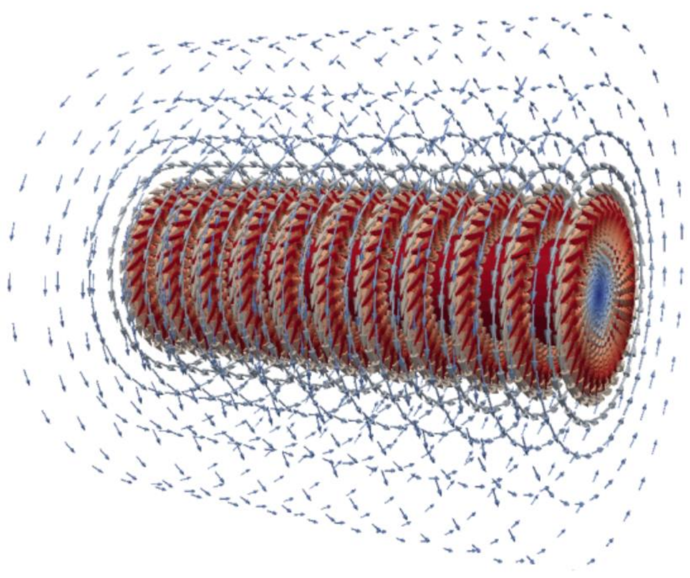
- Five basic datatypes
  - Items
  - Attributes
  - Links
  - Positions
  - Grids







3D/4D Spatial Data -> SciVis



https://www.paraview.org/

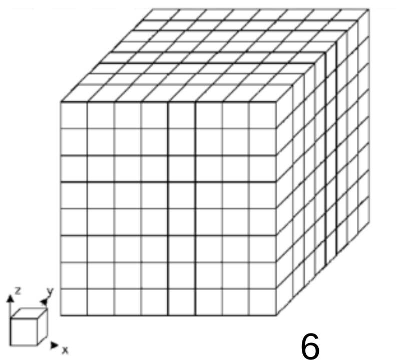
Attribute/variable

Census Year	Population by age group		
	Group 1: 0-14 years	Groups 2: 15-64 years	Group 3: 65 + years
1960	2591955	5588868	708569
1970	2451850	5326515	832760
1981	2508673	6198883	1125458
1991	1972403	6552000	1342744
2001	1656602	7006022	1693493
2011	1572329	6979785	2010064
2021	1331188	6588239	2423639

cell

Item/ object

Tabular Data -> InfoVis



- Data representation level:
  - Qualitative (or categorical)
  - Quantitative (or numeric)

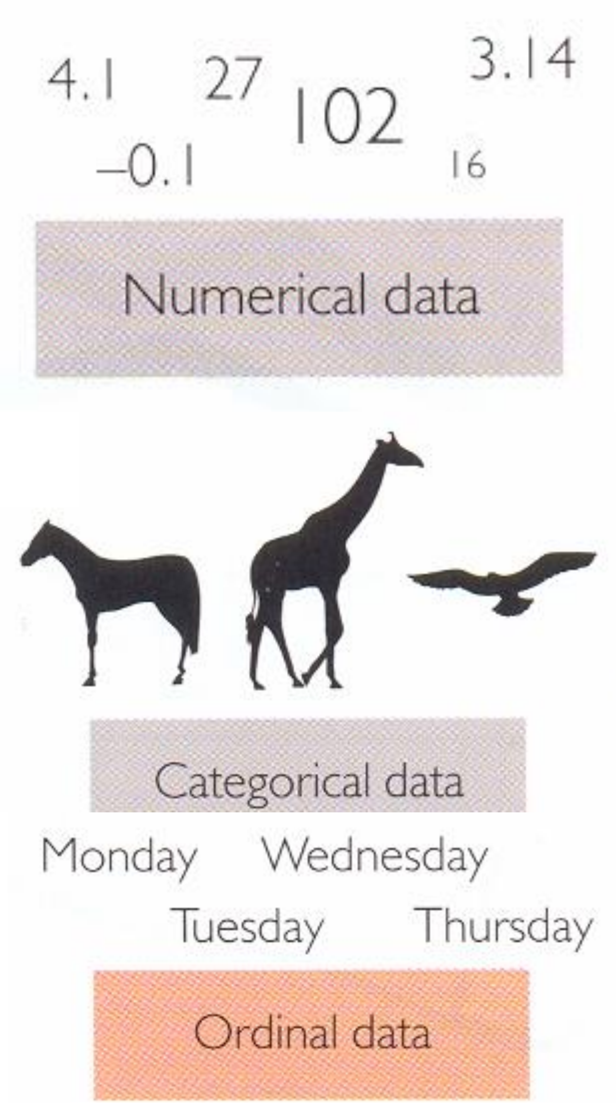
- Data nature:
  - Continuous
  - Discrete

Computer data are discrete
but the phenomena may be continuous

- Measuring scale:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Nominal, Ordinal > categorical

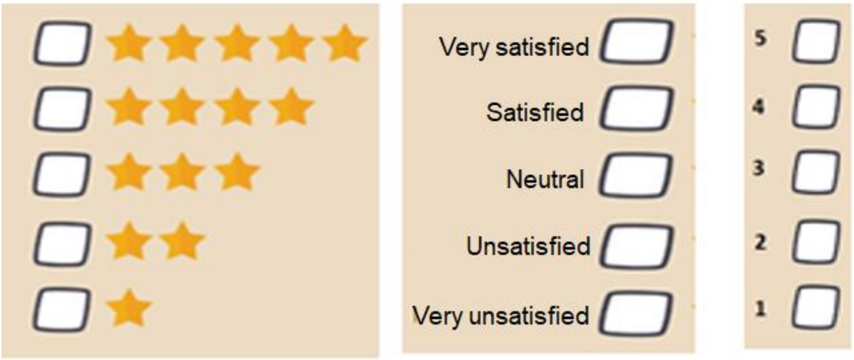
Interval, Ratio > quantitative



(Spence, 2007)

- What data are obtained from a survey using a Likert-type scale?

“How do you rate this product?”
“How satisfied are you with our service?”
“How do you grade this presentation?”



- Qualitative?
- Ordinal?
- Quantitative (or numeric)?

- Examples of measuring scales and types of data:
  - nominal --> car brands, gender, animal species…
  - ordinal --> week days, preferences, levels measured in a Likert-type scale
  - Interval --> date, IQ, temperatures in ºC
  - Ratio --> temperatures in ºK, weight, height

- The ratio scale represents the highest level of representation, has a non-arbitrary zero (unlike the interval scale)

- This is a general classification and might be used to select the statistical methods to use with the data

- In what scale are the following variables measured?

Length (m)

Weight (kg)

Latitude (º W/E)

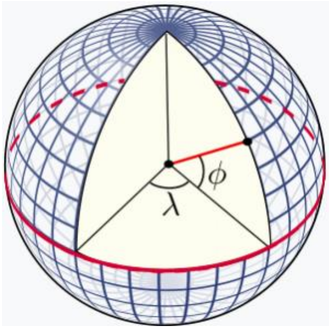
Longitude (º S/N)

Altitude (m)

Height (m)

Temperature (º F)

Pressure (Pa)



Example: beyond the structure of the data to Visualize: look at the phenomenon

- Consider a data set with three columns:

latitude *longitude* *d*

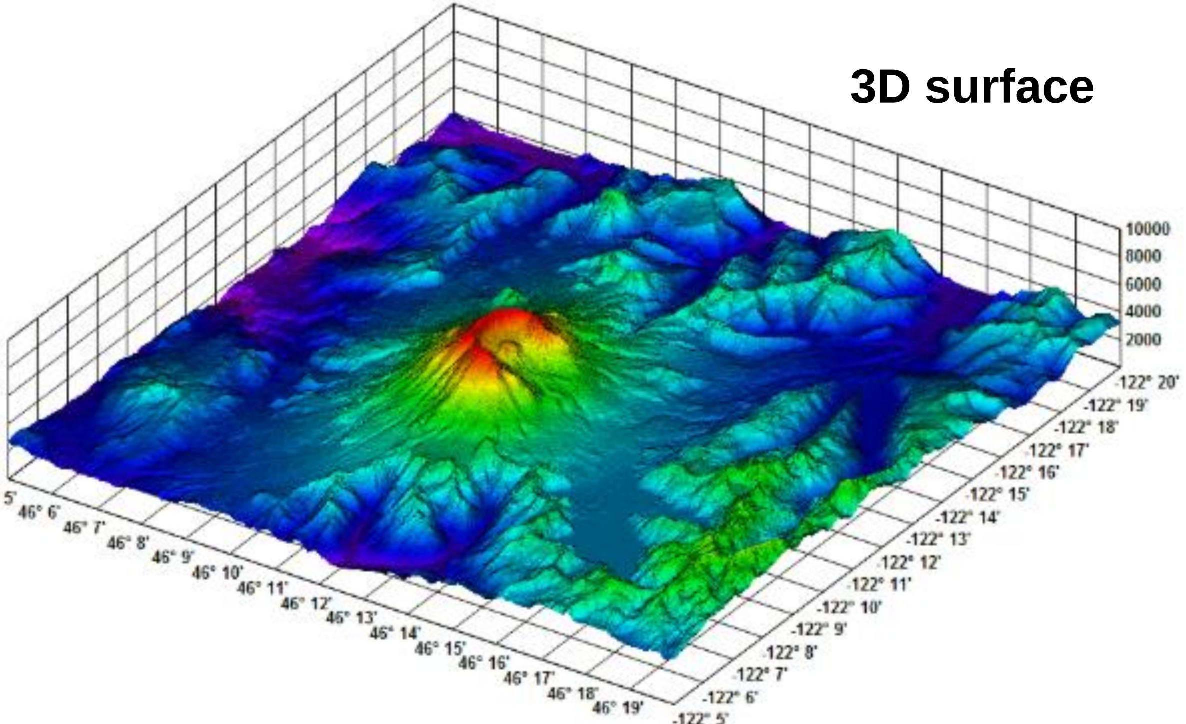
Iso-contours

- Which is the most adequate way to visualize these data?

- If *d* is depth or altitude?

the selected visualization technique may involve interpolation

(e.g. isocontours, isosurfaces, 3D surface)



3D surface

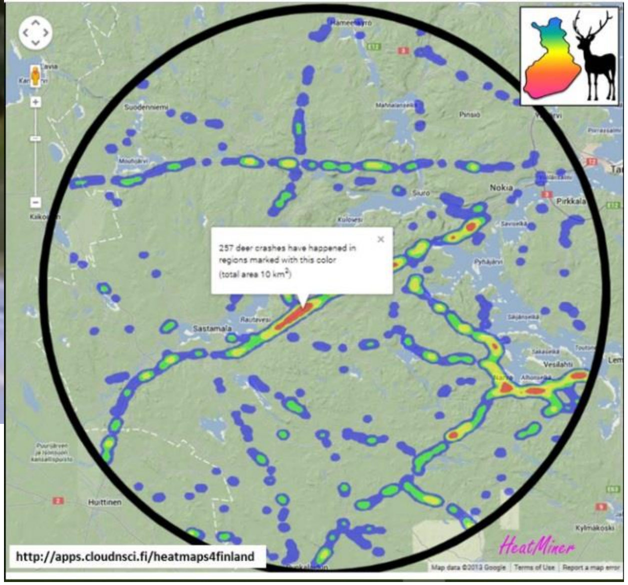
latitude *longitude* *d*

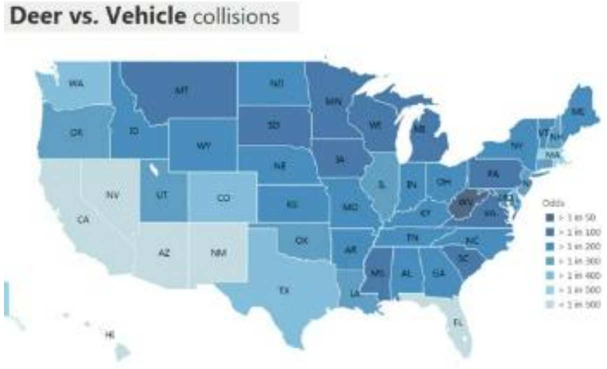
- What if data represent location and the number of “deer crash” accidents?

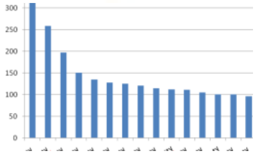
Heat map

Choropleth

http://cloudnsci.fi/wiki/index.php?n=Applications.Heatmaps4Finland


257 deer crashes have happened in regions marked with this color (total area 10 km²)
Sastamala
Rautavesi
Kulovesi
Siuro
Nokia
Pirkkala
Huittinen
HeatMiner
http://apps.cloudnsci.fi/heatmaps4finland





- Interpolation and contours don’t make sense!

Know the data structure is not enough

It is necessary to know the phenomenon behind the data as well as knowing the tasks (questions) of the users!

A partir de 26 de Fevereiro

- Segunda-feira -> DETI – 4.1.19
- Quinta-feira -> Anf. 12.2.1

Data preparation

- Data preparation is **very important and very time consuming**

- Several phases and terms:

  Data pre-processing

  Data wrangling

  Data cleaning, Data tiding …

  Data transformation

Data integrity becomes more essential when the volume of data increases

“Brilliant visualizations cannot redeem bad data!”

Or

“Garbage in garbage out …”

Cleansing Data

- Data is dirty: it contains typos, inconsistencies, fails in some way to meet a standard…

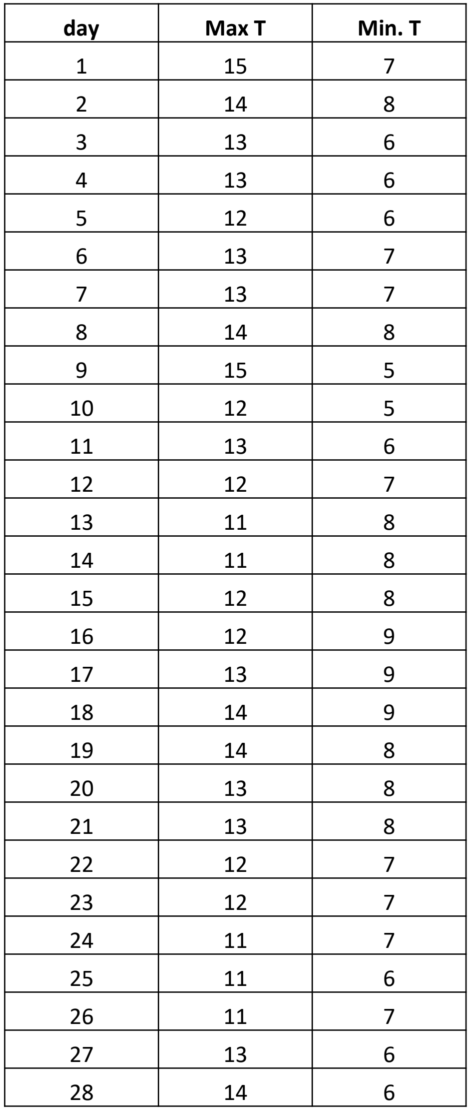
Transforming Data

(at the variable level)

  - Encoding
  - Aggregation
  - Derived data
  - Removal
  - Standardization

Revisiting previous examples:

Max and Min temperatures along the month of February (in ºC):



day	Max T	Min. T
1	15	7
2	14	8
3	13	6
4	13	6
5	12	6
6	13	7
7	13	7
8	14	8
9	15	5
10	12	5
11	13	6
12	12	7
13	11	8
14	11	8
15	12	8
16	12	9
17	13	9
18	14	9
19	14	8
20	13	8
21	13	8
22	12	7
23	12	7
24	11	7
25	11	6
26	11	7
27	13	6
28	14	6

Q4- How were the daily temperature ranges?

Q5 – What was the maximum temperature range?

- Should we use a derived variable to answer Q4 and Q5?

- What if we are addressing an audience in the USA?
  Should we use some other temperature unit?

Census Year	Population by age group		
	Group 1: 0-14 years	Groups 2: 15-64 years	Group 3: 65 + years
1960	2591955	5588868	708569
1970	2451850	5326515	832760
1981	2508673	6198883	1125458
1991	1972403	6552000	1342744
2001	1656602	7006022	1693493
2011	1572329	6979785	2010064
2021	1331188	6588239	2423639

- What if we are studying the senior population?
- Or only children?
- Or everyone?

More examples:

Cleaning Data

Birth date: Feb/30/2000

Temperature: -300 °K

City: Lixboa

Transforming Data

- Encoding – answers to an open question need to be parsed and coded
- Aggregation – detail may be excessive (age: <18; 18-40; 41-65; >65)
- Derived data – add new relevant variables ($T_{range}= T_{max}-T_{min}$)
- Removal – remove data that are not needed
- Standardization – M/F; °C or °F

Main bibliography

- Camões, J., *Data at Work : Best practices for creating effective charts and information graphics in Microsoft Excel*, Pearson Education, 2016

  https://learning.oreilly.com/library/view/data-at-work/9780134268798/title.html

- Spence, R., *Information Visualization, Design for Interaction*, 2nd ed., Prentice Hall, 2007

- Munzner, T., *Visualization Analysis and Design*, A K Peters, 2014

  https://learning.oreilly.com/library/view/visualization-analysis-and/9781466508910/cover.xhtml

- Kirk, A., *Data Visualization : a successful design process*. Packt Publishing, 2012

  https://learning.oreilly.com/library/view/data-visualization-a/9781849693462/cover.html