



Durham  
University

# Data Cleaning and Analytics

## *Topic 1 Introduction* *--Week 1&2*

jingyun.wang@durham.ac.uk

References:

<https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4>

<https://www.scribbr.com/methodology/data-collection/>

# When and How?

- Term 1 Michaelmas
- Lectures, reading materials, F2F interaction
- Office Hours: Wednesday 4-5pm in zoom ([zoom meeting link](#))
- Assessment: evaluated 100% via coursework :
  - **Release date: 27/10/2023**
  - **Submission date: 12/12/2023 14:00 GMT**
  - **Feedback to students: 19/01/2024**

# Reading List



## Reading List

Visible to students ▼

The Library reading list for this module.

Data Science

Edit ▼

View & Export ▼

+ My Lists

ACADEMIC YEAR 2023-24 By George Koulieris Created 4 months ago | Updated 4 months ago Linked to COMP2271

COMP2271

Table of Contents ▼

Type: All ▼

Filter: All ▼

Citation Style: None ▼

Search

Probability  
Recommended reading



Probability and statistics

Book - by Morris H. DeGroot; Mark J. Schervish - 2014 - Fourth edition - Importance not set ▼

VIEW ONLINE



Data Collection and Cleaning



Python data science handbook: essential tools for working with data

Book - by Jacob T. Vanderplas - 2016 - First edition - Essential ▼

VIEW ONLINE



Python for data analysis: data wrangling with pandas, NumPy, and IPython

Book - by Wes McKinney - 2018 - Second edition - Essential ▼

VIEW ONLINE



Data mining: practical machine learning tools and techniques

Book - by I. H. Witten; Eibe Frank; Mark A. Hall - c2011 - 3rd ed - Importance not set ▼

VIEW ONLINE



Data science from scratch: first principles with Python

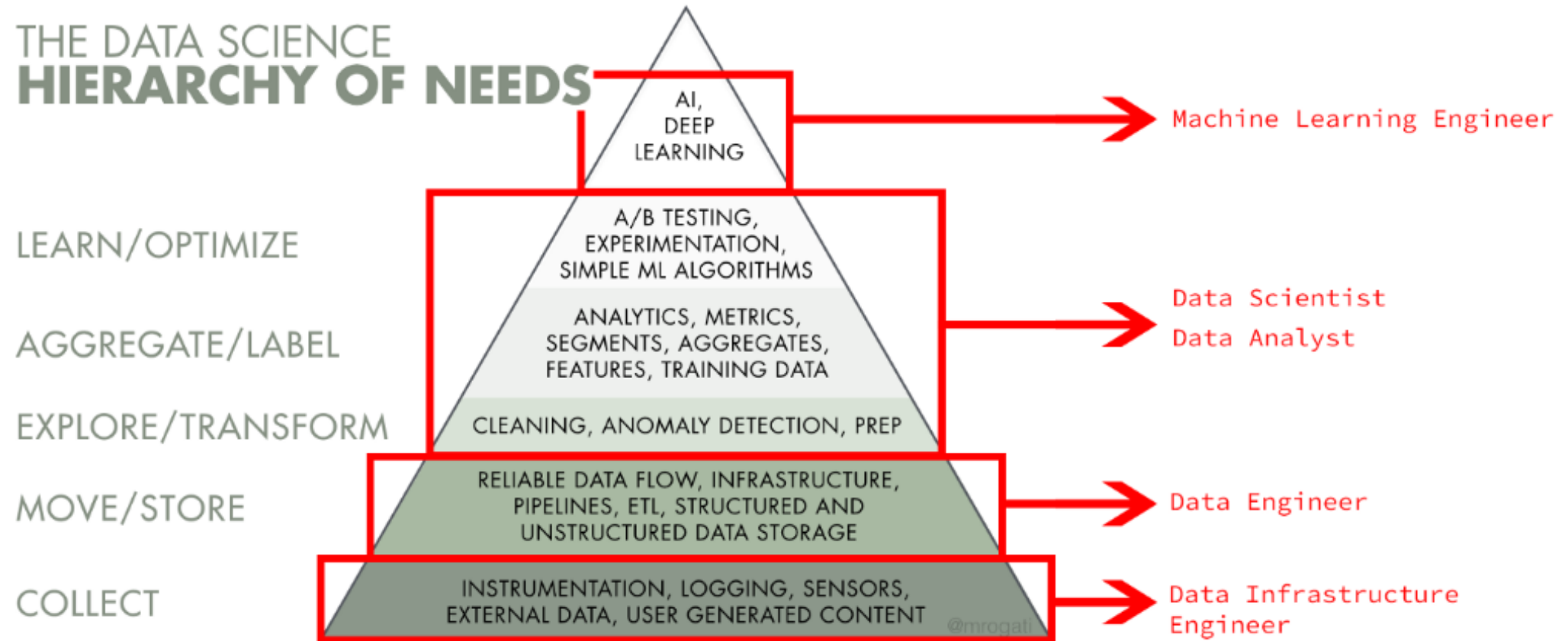
Book - by Joel Grus - 2019 - Second edition - Essential ▼

VIEW ONLINE



# THE DATA SCIENCE HIERARCHY OF NEEDS

(Monica Rogati)



**Data collection:** a systematic process of gathering observations or measurements.

Data collection is the process of **gathering** and **measuring** information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes.

- a research component in all study fields, including physical and social sciences, humanities, and business.
- While methods vary by discipline, the emphasis on ensuring **accurate** and **honest** collection remains the same.
- The goal for all data collection is to capture **quality evidence** that allows analysis to lead to **the formulation of convincing and credible answers** to the questions that have been posed.

[https://en.wikipedia.org/wiki/Data\\_collection](https://en.wikipedia.org/wiki/Data_collection)

# Daily data collection-Internet cookies



What Are Cookies? And How They Work

<https://www.youtube.com/watch?v=rdVPfIECed8>

# To collect high-quality data that is relevant to your purposes, follow these steps.

## Step 1: Define the aim of your research

- what is the practical or scientific issue that you want to address and why does it matter?
- Next, formulate one or more research questions that precisely define what you want to find out.

Depending on your research questions, you might need to collect quantitative or qualitative data:

- Quantitative data is expressed in ***numbers and graphs*** and is analyzed through ***statistical methods***.
- Qualitative data is expressed in ***words*** and analyzed through interpretations and categorizations.

## Step 2: Choose your data collection method

Method	When to use	How to collect data
Experiment	To test a causal relationship.	Manipulate variables and measure their effects on others.
Survey	To understand the general characteristics or opinions of a group of people.	Distribute a list of questions to a sample online, in person or over-the-phone.
Interview/focus group	To gain an in-depth understanding of perceptions or opinions on a topic.	Verbally ask participants open-ended questions in individual interviews or focus group discussions.
Observation	To understand something in its natural setting.	Measure or survey a sample without trying to affect them.
Ethnography	To study the culture of a community or organization first-hand.	Join and participate in a community and record your observations and reflections.
Archival research	To understand current or historical events, conditions or practices.	Access manuscripts, documents or records from libraries, depositories or the internet.
Secondary data collection	To analyze data from populations that you can't access first-hand.	Find existing datasets that have already been collected, from sources such as government agencies or research organizations.
User Log	To evaluate a design new system or to analyze the user behaviors	Record the user daily data automatically



**Step 3: Plan your data collection procedures:** to make accurate observations or measurements of the variables you are interested in

- **Operationalization:** turning abstract conceptual ideas into measurable observations.
- **Sampling:** will determine how you recruit participants or obtain measurements for your study
- **Standardizing procedures:** write a detailed manual( to collect data in a consistent way)→reliability of the data
- **Creating a data management plan**
  - anonymize and safeguard the data to prevent leaks of sensitive information
  - perform transcriptions or data entry in systematic ways to minimize distortion.
  - backed up

# Data Cleaning and Analytics

1. Introduction
2. Working with various types of files
3. Pandas
4. Data cleaning and preparation
5. Data visualization and relevant Python libraries
6. Data Analytics and Reporting
7. Decision making (Dr Nelly Bencomo will be the guest lecturer)

# Data cleansing

**Data cleansing** is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.<sup>[1]</sup>

[https://en.wikipedia.org/wiki/Data\\_cleansing#Data\\_quality](https://en.wikipedia.org/wiki/Data_cleansing#Data_quality)

# High-quality data criteria

- **Validity**
- **Accuracy**
- **Completeness**
- **Consistency**
- **Uniformity**

# Data quality—(1)Validity

The degree to which the measures conform to defined business rules or constraints.

- Data-Type Constraints:** values in a particular column must be of **a particular datatype**, e.g., boolean, numeric, date, etc.
- Range Constraints:** typically, numbers or dates should fall within **a certain range**.
- Mandatory Constraints:** certain columns **cannot be empty**.
- Unique Constraints:** a field, or a combination of fields, must be **unique across a dataset**.
- Set-Membership constraints:** values of a column come from a set of discrete values, e.g. enum values. For example, a person's gender may be male or female.
- Foreign-key constraints:** as in relational databases, a foreign key column can't have a value that does not exist in the referenced primary key.
- Regular expression patterns:** text fields that have to be in **a certain pattern**. For example, phone numbers may be required to have the pattern (999) 999–9999.
- Cross-field validation:** **certain conditions** that span across multiple fields must hold. For example, a patient's date of discharge from the hospital cannot be earlier than the date of admission.

# Data quality—(2)Accuracy

The degree of conformity of a measure to **a standard or a true value**.

- While defining all possible valid values allows invalid values to be easily spotted, it does not mean that they are accurate. A valid street address mightn't actually exist.
- Accuracy is very hard to achieve through data-cleansing in the general case because it requires accessing **an external source of data** that contains the true value. Accuracy has been achieved in some cleansing contexts, notably customer contact data, by using external databases that match up zip codes to geographical locations (city and state) and also help verify that street addresses within these zip codes actually exist.
- Another thing to note is the difference between **accuracy** and **precision**. Saying that you live on the earth is, actually true. But, not precise. Where on the earth?. Saying that you live at a particular street address is more precise.

# Data quality—(3)Completeness

The degree to which all required measures are known.

- Incompleteness is almost impossible to fix with data cleansing methodology: one cannot infer facts that were not captured when the data in question was initially recorded.
  - questioning the original source if possible, say **re-interviewing** the subject. But the subject may give a different answer or maybe hard to reach again.
- In the case of systems that insist certain columns should not be empty, one may work around the problem by designating a value that indicates "**unknown**" or "**missing**", but the supplying of default values does not imply that the data has been made complete.

# Data quality—(4) Consistency

The degree to which the data is consistent, within the same data set or across multiple data sets.

- Inconsistency occurs when **two data items in the data set contradict** each other. e.g., a customer is recorded in two different systems as having two different current addresses, and only one of them can be correct.
- Fixing inconsistency is not always possible: it requires a variety of strategies - e.g., deciding which data were recorded **more recently**, which data source is likely to be **most reliable** (the latter knowledge may be specific to a given organization), or simply trying to find the truth by **testing** both data items (e.g., calling up the customer).



# Data quality—(5)Uniformity

The degree to which the data is specified using **the same unit of measure**.

- In datasets pooled from different locales, weight may be recorded either in pounds or kilos.
- must be converted to a single measure using an arithmetic transformation.

# How to produce high-quality data?

The workflow--an iterative process

- 1. Inspection:** Detect unexpected, incorrect, and inconsistent data.
- 2. Cleaning:** Fix or remove the anomalies discovered.
- 3. Verifying:** After cleaning, the results are inspected to verify correctness.
- 4. Reporting:** A report about the changes made and the quality of the currently stored data is recorded.

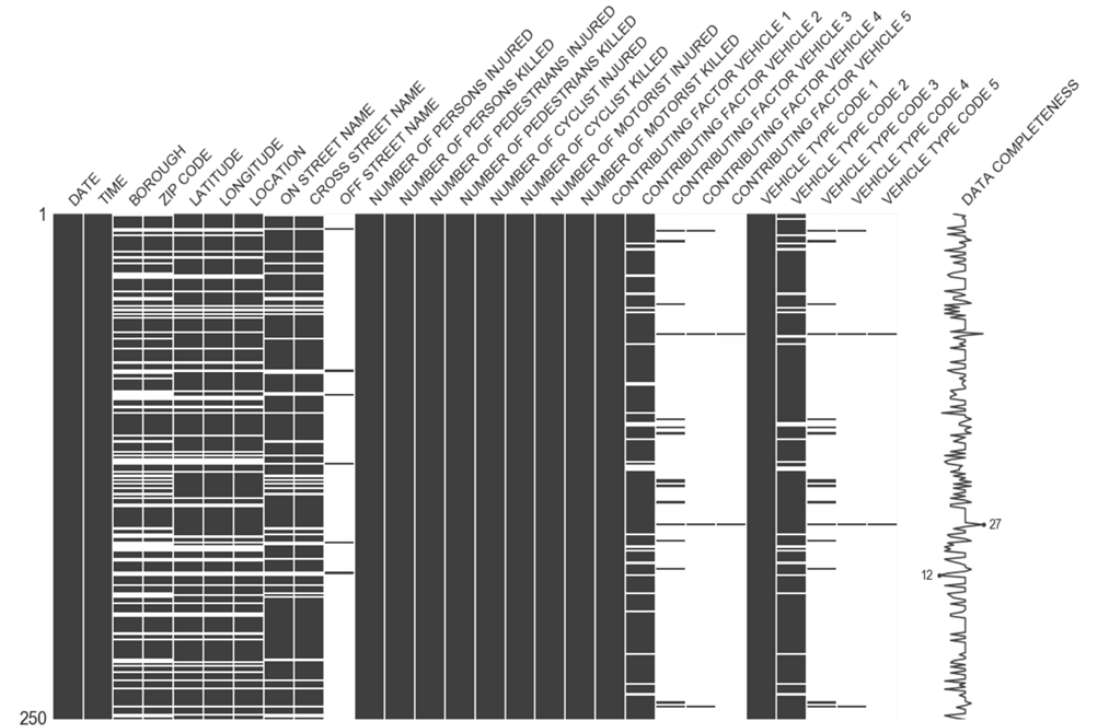
# Inspection--exploring the underlying data for error detection

- **Data profiling**: A **summary statistics** about the data, is really helpful to give a general idea about the quality of the data.
  - For example, check whether a particular column conforms to a particular standard or **pattern**. Is the data column recorded as a string or number?
  - How many values are **missing**?. How many unique values in a column, and their **distribution**?. Is this data set is linked to or have a **relationship** with another?

# Inspection--exploring the underlying data for error detection

## ➤ Visualizations

By analyzing and visualizing the data using statistical methods such as mean, standard deviation, range, or quantiles, one can find values that are **unexpected** and thus **erroneous**.



Nullity matrix (Aleksey Bilogur)

<https://github.com/ResidentMario/missingno>

# Cleaning: remove, correct, or impute incorrect data

**Irrelevant data:** those that are not actually needed, and don't fit under the context of the problem we're trying to solve.

- Only if you are sure that a piece of data is unimportant, you may drop it. Otherwise, explore the **correlation matrix** between feature variables.
- And even though you noticed no correlation, you should ask someone who is **domain expert**. You never know, a feature that seems irrelevant, could be very relevant from a domain perspective such as a clinical perspective

telephone, weight, blood test → health status

Learning Styles, cognitive load → learning performance

# Cleaning--Duplicates

data points that are repeated in your dataset

often happens when for example

- Data are combined from different sources
- The user may hit submit button twice thinking the form wasn't actually submitted.
- A request to online booking was submitted twice correcting wrong information that was entered accidentally in the first time.
- A common symptom is when two users have the same identity number. Or, the same article was scrapped twice. And therefore, they simply should be removed.

News about Ransomware

# Cleaning--Type conversion

numbers → numerical data types.

dates → date objects/Unix timestamp (number of seconds)

Categorical values ↔ numbers

Note: The values that can't be converted to the specified type should be converted to NA value (or any), with a warning being displayed. This indicates the value is incorrect and must be fixed.

# Cleaning--Syntax errors

- **Remove white spaces:** Extra white spaces at the beginning or the end of a string should be removed.

```
" hello Durham " => "hello Durham"
```

- **Pad strings:** Strings can be padded with spaces or other characters to a certain width. For example, some numerical codes are often represented with **prepending zeros** to ensure they always have the same number of digits.

```
929 => "000929"
```

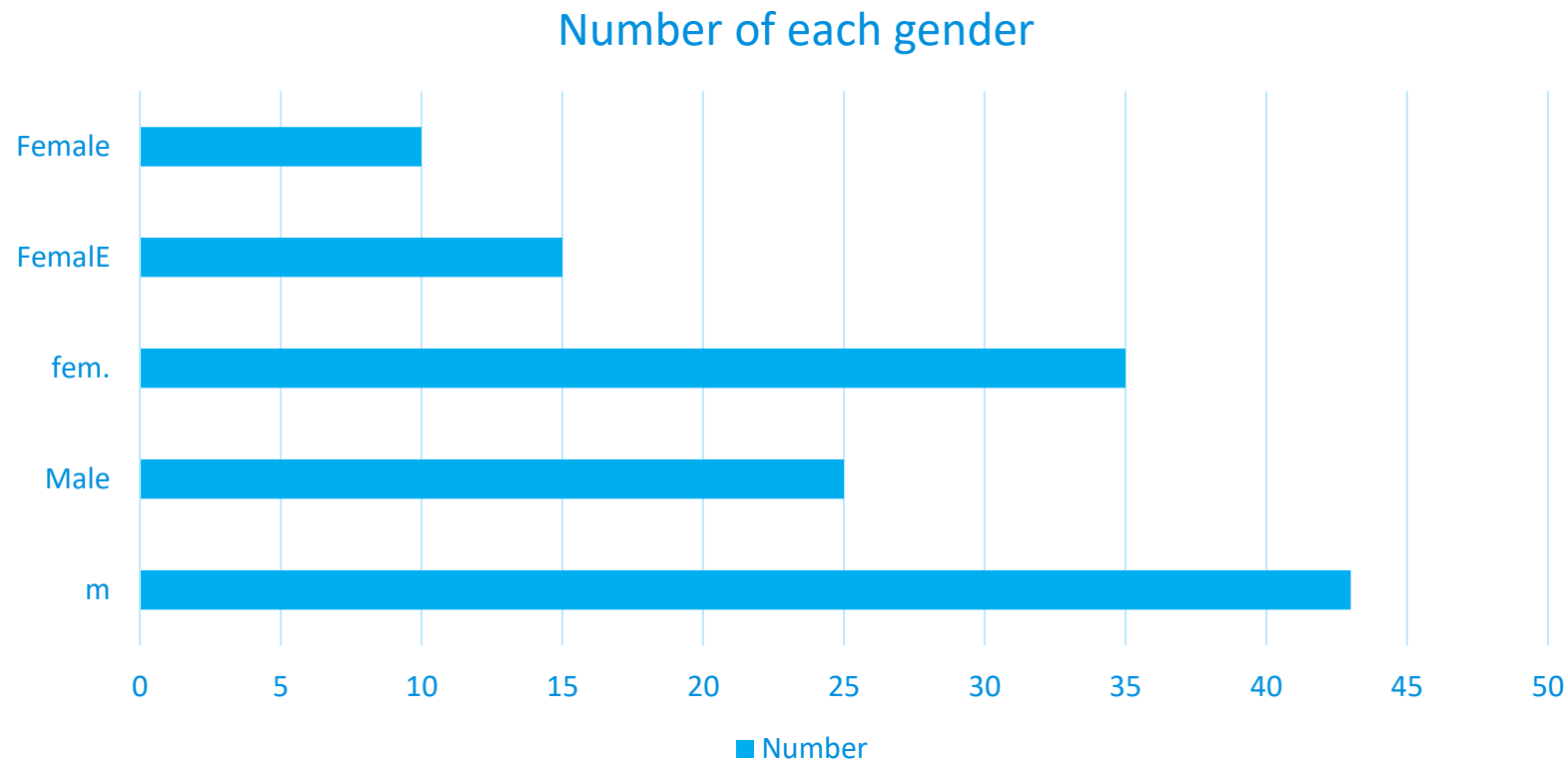
- **Fix typos:** Strings can be entered in many different ways, and no wonder, can have mistakes.

```
Gender: m/Male/fem./FemalE/Female
```



# Cleaning-- Syntax errors

A bar plot is useful to visualize all the unique values. One can notice some values are different but do mean the same thing



# Cleaning-- Syntax errors

The first solution is to manually **map each value** to either “male” or “female”.

```
dataframe['gender'].map({'m': 'male', 'f': 'female', ...})
```

The second solution is to use **pattern match**. For example, we can look for the occurrence of m or M in the gender at the beginning of the string. If you want to replace all the words starting with m end with whatever, to become "Male":

```
re.sub(r"^m[^\s]*", 'Male', 'malee', flags=re.IGNORECASE)
```



pattern



replacement



input string

# Cleaning-- Syntax errors

The third solution is to use **fuzzy matching**: An algorithm that identifies the distance between the expected string(s) and each of the given one. Its basic implementation counts how many operations are needed to turn one string into another.

Gender	male	female
m	3	5
Male	1	3
fem.	5	3
FemaleE	3	2
Femle	3	1

should replace all values that mean the same thing to one unique value

# Cleaning--Standardize

not only recognize the typos

but also put each value in the same standardized format.

- For **strings**, make sure all values are either in lower or upper case.
- For **numerical values**, make sure all values have a certain measurement unit.
- For **dates**, the USA version is not the same as the European version. Recording the date as a timestamp (a number of milliseconds) is not the same as recording the date as a date object.

# Cleaning--Scaling / Transformation

transform your data so that it fits within a specific scale, such as 0–100 or 0–1.

It can also help in making certain types of data easier to plot.

- For example, we might want to reduce skewness to assist in plotting (when having such many outliers). The most commonly used functions are **log**, **square root**, and **inverse**.

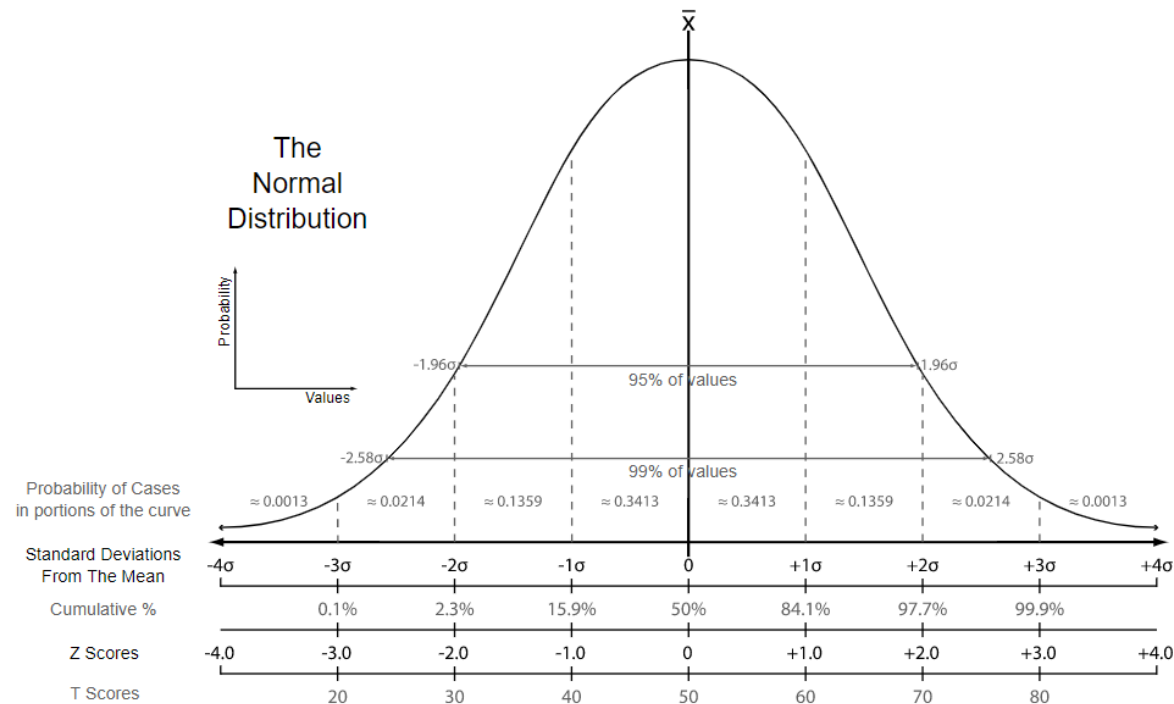
It can also take place on data that has **different measurement units** or **different scales**.

- Student scores on different exams can't be compared if these two exams are on a different scale. In this case, we need re-scale these two scores to take numbers, say, between 0–1. By scaling, we can plot and compare different scores.

# Cleaning-- Normalization

to change your observations so that they can be described as a normal distribution.

Normal distribution (Gaussian distribution), also known as the **bell curve**, is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean.



# Cleaning-- Normalization

While normalization also rescales the values into a range of 0–1, the intention here is to transform the data so that it is normally distributed



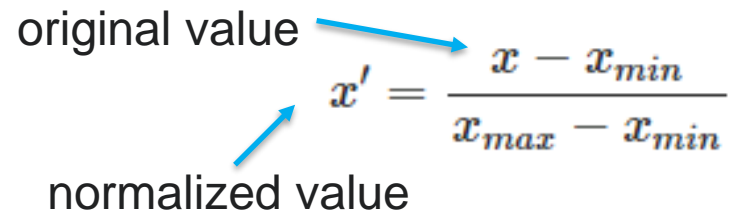
In most cases, we normalize the data if we're going to be using statistical methods that rely on normally distributed data.

Name	Formula	Use
<a href="#">Standard score</a>	$\frac{X - \mu}{\sigma}$	Normalizing errors when population parameters are known. Works well for populations that are <a href="#">normally distributed</a> <sup>[2]</sup>
Student's t-statistic	$\frac{\hat{\beta} - \beta_0}{\text{s. e.}(\hat{\beta})}$	the departure of the estimated value of a parameter from its hypothesized value, normalized by its standard error.
Studentized residual	$\frac{\hat{e}_i}{\hat{\sigma}_i} = \frac{X_i - \hat{\mu}_i}{\hat{\sigma}_i}$	Normalizing residuals when parameters are estimated, particularly across different data points in <a href="#">regression analysis</a> .
Standardized moment	$\frac{\mu_k}{\sigma^k}$	Normalizing moments, using the standard deviation $\sigma$ as a measure of scale.
<a href="#">Coefficient of variation</a>	$\frac{\sigma}{\mu}$	Normalizing dispersion, using the mean $\mu$ as a measure of scale, particularly for positive distribution such as the <a href="#">exponential distribution</a> and <a href="#">Poisson distribution</a> .
<a href="#">Min-max feature scaling</a>	$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$	<a href="#">Feature scaling</a> is used to bring all values into the range [0,1]. This is also called unity-based normalization. This can be generalized to restrict the range of values in the dataset between any arbitrary points $a$ and $b$ , using for example $X' = a + \frac{(X - X_{\min})(b - a)}{X_{\max} - X_{\min}}$ .

# Cleaning-- Normalization

**Feature scaling**: is a method used to normalize the range of **independent variables** or **features of data**. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

- min-max scaling/ min-max normalization/Rescaling: the simplest method and consists in rescaling the range of features to scale the range in  $[0, 1]$  or  $[-1, 1]$ .



original value  $x$  is mapped to the numerator of the formula, and the normalized value  $x'$  is the result of the formula.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

normalized value

- is important in support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important.



# Cleaning-- Normalization

**Feature scaling**: is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as **data normalization** and is generally performed during the data preprocessing step.

- min-max scaling /min-max normalization/Rescaling
- Standard score(Standardization)/Z-score Normalization

original feature vector

$$x' = \frac{x - \bar{x}}{\sigma}$$

standard deviation

It's widely used in SVM, logistics regression and neural networks.

# Cleaning-- Normalization

**Feature scaling**: is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as **data normalization** and is generally performed during the data preprocessing step.

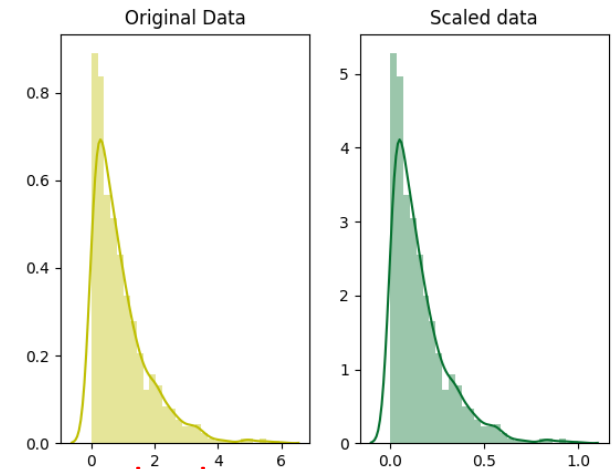
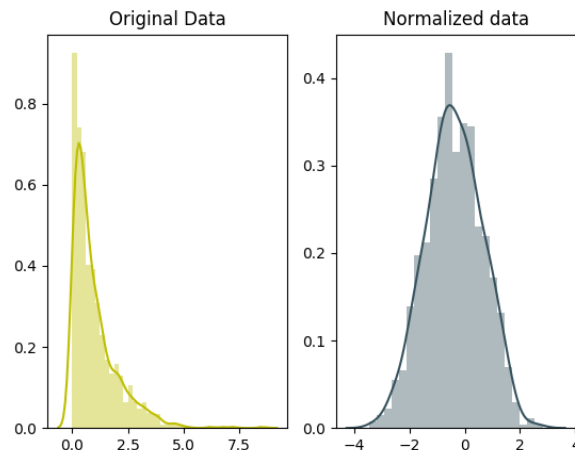
➤ **min-max scaling** /min-max normalization/Rescaling

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

➤ **Standard score**(Standardization)/Z-score Normalization

➤ **Mean normalization**

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)}$$



the use of a particular normalization method depends on the problem itself (or you may want to experiment which one better suits!).

# Cleaning--Missing values

- ❑ rarely clean and homogeneous.
  - ❑ the missing values are unavoidable
  - ❑ different data sources may indicate missing data in different ways.
- 
- Generally, two strategies:
    - using a mask that globally indicates missing values,
    - or choosing a sentinel value that indicates a missing entry.
  - In Pandas missing data is represented by two value:
    - None: a Python singleton object that is often used for missing data in Python code.
    - NaN (an acronym for Not a Number): a special floating-point value recognized by all systems that use the **standard IEEE floating-point representation**

# Cleaning--Missing values

## — One. Drop.

If the missing values in a column rarely happen and occur at random → drop observations (rows)

If most of the column's values are missing, and occur at random → drop the whole column

## — Two. Impute: calculate the missing value based on other observations

- 1) Using **statistical values** like mean, median. However, none of these guarantees unbiased data, especially if there are many missing values.
- 2) Using a **linear regression**. Based on the existing data, one can calculate the **best fit** line between two variables, say, house price vs. size  $m^2$ . It is worth mentioning that linear regression models are sensitive to outliers.
- 3) **Hot-deck**: Copying values from other similar records. This is only useful if you have enough available data. And, it can be applied to numerical and categorical data.

# Cleaning--Missing values $\neq$ default values $\neq$ “unknown”

## — One. Drop.

If the missing values in a column rarely happen and occur at random  $\rightarrow$  drop observations (rows)

If most of the column's values are missing, and occur at random  $\rightarrow$  drop the whole column

— **Two. Impute:** calculate the missing value based on other observations

— **Three. Flag:** reinforcing the pattern already exist by other features.

What if the data is missing is informative in itself?

when the missing data doesn't happen at random. Take for example a conducted survey where most people from a specific race refuse to answer a certain question.

# Cleaning--Outliers

## Interquartile range (IQR).

Given a list of values, sort, and then split them into 4 quarters. The range from the end of Q1 till start of Q3 (the middle 50% of the data) is the box in the box plot, where the line inside is the median. Let's take an example. Suppose the values of life expectancies column are:

1,2,2,2|2,3,3,3|5,8,10,15|26,30,45,50

## To calculate the quartiles and IQR

- \* Quartile 1 (Q1) = 25% =  $(2+2)/2 = 2 \Rightarrow$  25% of numbers  $\leq 2$
- \* Quartile 2 (Q2) = 50% =  $(3+5)/2 = 4 \Rightarrow$  50% of numbers  $\leq 4$  (median)
- \* Quartile 3 (Q3) = 75% =  $(15+26)/2 = 20.5 \Rightarrow$  75% of numbers  $\leq 20.5$
- \* IQR = Q3 — Q1 =  $20.5 - 2 = 18.5$  (not really helpful on its own, only when comparing different distributions i.e. data variability)

So, we can say that the “middle 50%” of countries, have life expectancies between 2 and 20.5 years old.

# Cleaning--Outliers

They are values that are significantly different from all other observations.

- Any data value that lies more than  $(1.5 * \text{IQR})$  away from the Q1 and Q3 quartiles is considered an outlier.
- Even weird, **suspicious values** that are unlikely to happen, should not be removed unless there is a good reason for that.
- It is also worth mentioning that **some models**, like linear regression, are very **sensitive to outliers**. In other words, outliers might throw the model off from where most of the data lie.

# Cleaning--In-record & cross-datasets errors

two or more values in the same row or across datasets that contradict with each other

- For example, if we have a dataset about the cost of living in cities. The total column must be equivalent to the sum of rent, transport, and food
- Similarly, a child can't be married. An employee's salary can't be less than the calculated taxes.



# Verifying

After cleaning the data, verify correctness by re-inspecting the data and making sure it rules and constraints do hold.

- For example, after filling out the missing data, they might violate any of the rules and constraints.

It might involve some manual correction if not possible otherwise.

# Reporting

Reporting **how healthy the data is**

software packages or libraries can generate reports of the changes made, which rules were violated, and how many times

In addition to logging the violations, the causes of these errors should be considered. Why did they happen in the first place?.

Distribution before and after data cleaning

# Summary

No matter how robust and strong the validation and cleaning process is, one will continue to suffer as new data come in. It is better to guard yourself against a disease instead of spending the time and effort to remedy it.

➤ **Research questions?**

➤ **How the data is collected, and under what conditions?**

- the location, timing, weather conditions, etc.

➤ **What does the data represent?--metadata**

➤ **What are the methods used to clean the data and why?**

Different methods can be better in different situations or with different data types.