Data Management

Working with data

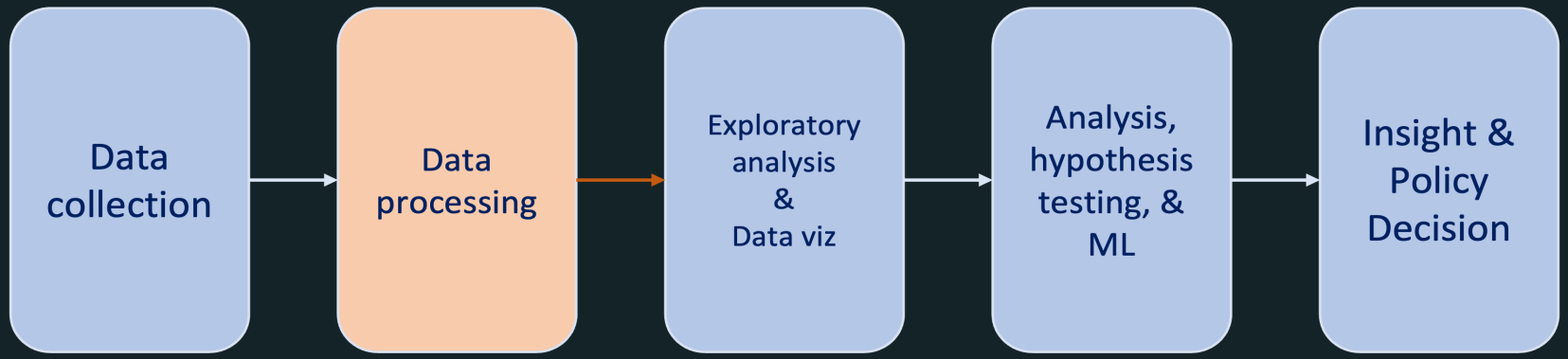
Malka Guillot

HEC Liège | ECON2306

Table of contents

What is data?

Data lifecycle



Objective: get structured & tidy data

Data wrangling (data munging)

= The process of transforming raw data to a set of data tables that can be used for a variety of downstream purposes such as analytics

Tidy data

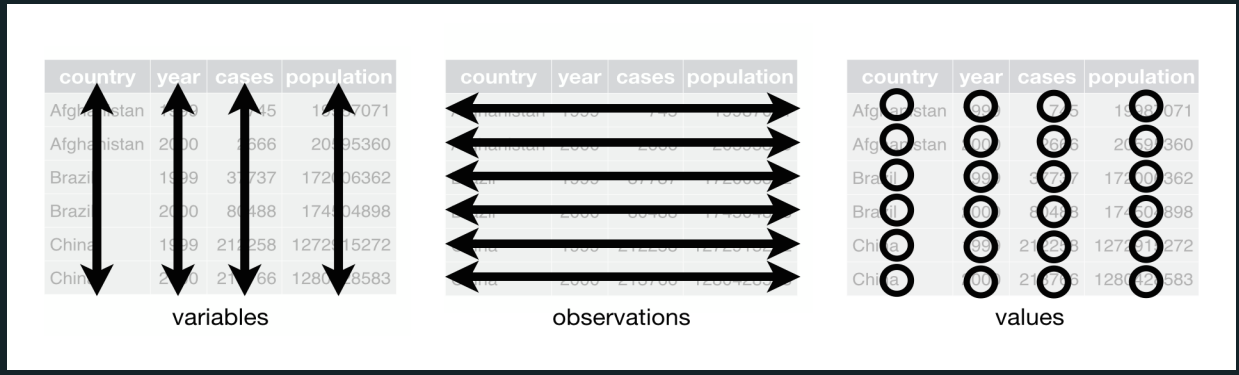
Principles

1. Each observation forms a row
2. Each variable forms a column
3. Each type of observational unit forms a table.
4. Each observation has a unique identifier (ID)

Advantages:

- **easier** to work with.
- finding errors and issues with data are usually easier with tidy data tables
- more transparent → helps other users to understand
- easy to extend:
  - new observations added as new rows;
  - new variables as new columns.

Structured data



Example of messy data

	Treatment A	Treatment B
John Smith	-	2
Jane Doe	16	11
Mary Johnson	3	1

Example of tidy data

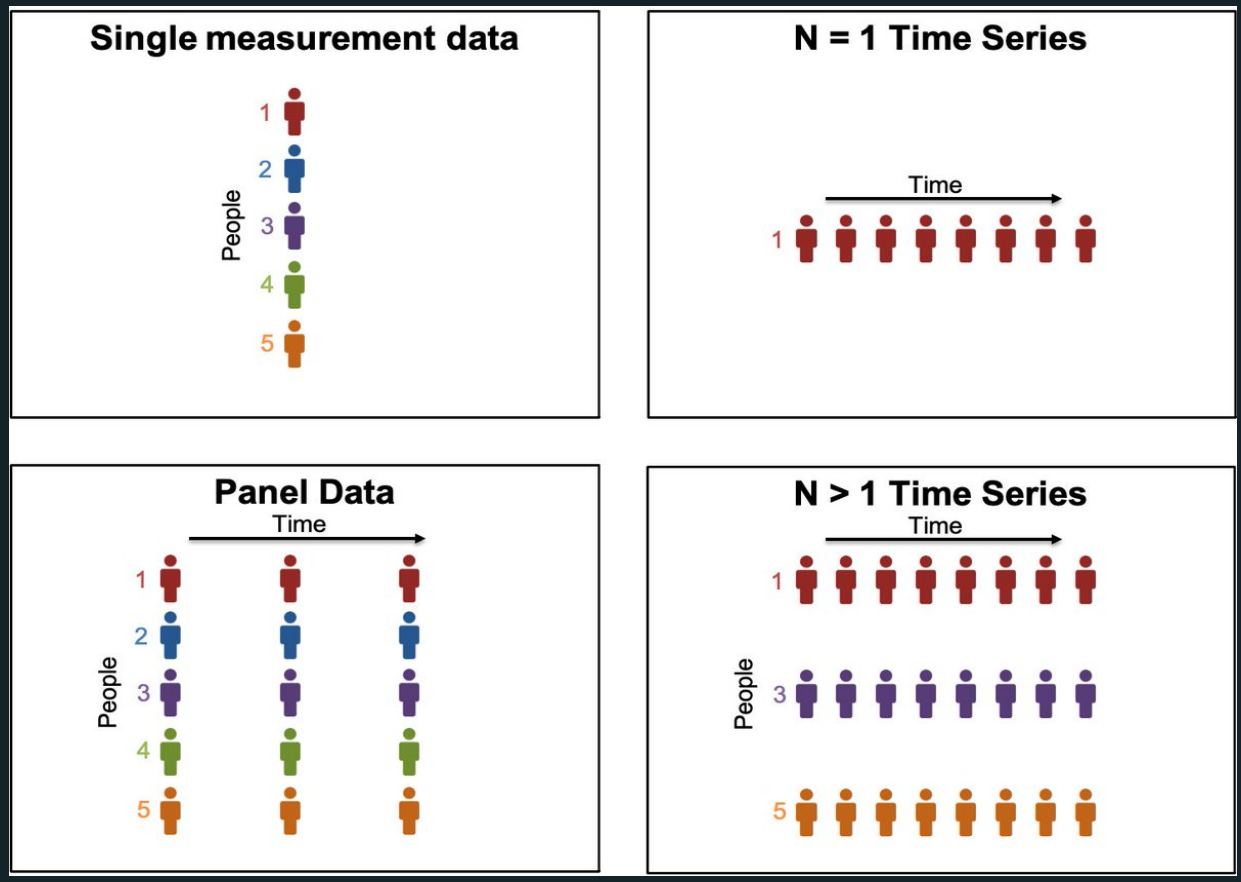
Name	Treatment	Result
John Smith	a	-
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Unstructured data?

⇒ we will be mainly working with structured data + learning how to go from unstructured to structured

Data structures

- cross-sectional data
- time series data
- panel or longitudinal data
  - balanced or not



Variables types

What type of variable do you know?

Variables

- Quantitative vs. qualitative variables
- Stock vs. flow variables

Data cleaning

Filter out duplicates

- duplicates: some observations appearing more than once in the data.
- May be the result of human error or the features of data source

Missing values

- Need to be identified & should be counted
- Potential selection bias: is data missing at random?
- Solutions:
  - Restrict the analysis to observations with non-missing values for **all** variables
  - Imputation : Fill in some value for the missing values, such as the mean or median value.

Extreme values

Substantially larger or smaller values for one or a handful of observations.

- Need conscious decision.
- Is this an error? (drop or replace)
- Is this not an error but not part of what we want to talk about? (drop)
- Is this an integral feature of the data? (keep)

Data wrangling: common steps

1. Write a code - it can be repeated and improved later
2. Understand the structure of the dataset:
    - create data tables, recognize links. Draw a schema.
3. Start by looking into the data table(s) to spot issues
4. Store data in tidy data tables.
5. Get each variable in an appropriate format
6. Have a description of variables
7. Make sure values are in meaningful ranges; correct non-admissible values or set them as missing
8. Identify missing values and store them in an appropriate format. Make edits if needed.
9. Document every step of data cleaning

Data operations

Tidying messy datasets

- **Objective** :
  - Prepare data in a standardized way prior to the analysis.
- **Tool** :
  - pandas package of handling data
  - os package of path

```
import pandas as pd
```

0. The Pandas DataFrame

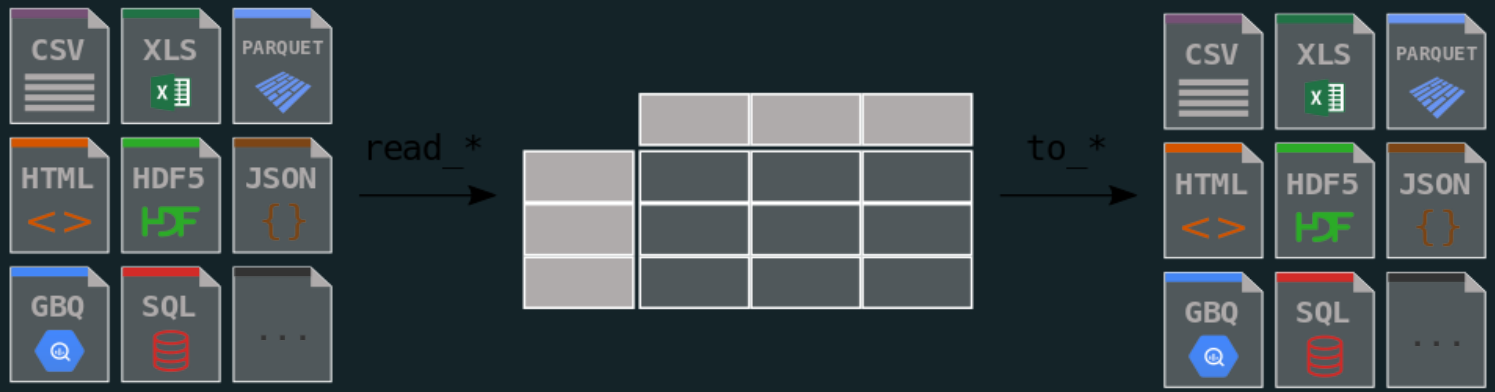
Let's create a table by hand

```
df = pd.DataFrame(
    {
        "Name": [
            "Braund, Mr. Owen Harris",
            "Allen, Mr. William Henry",
            "Bonnell, Miss. Elizabeth",
        ],
        "Age": [22, 35, 58],
        "Sex": ["male", "male", "female"],
    }
)
```

	Name	Age	Sex
0	Braund, Mr. Owen Harris	22	male
1	Allen, Mr. William Henry	35	male
2	Bonnell, Miss. Elizabeth	58	female

Each column in a DataFrame is a Series: cf. exercise

0. The Pandas DataFrame



Importing data

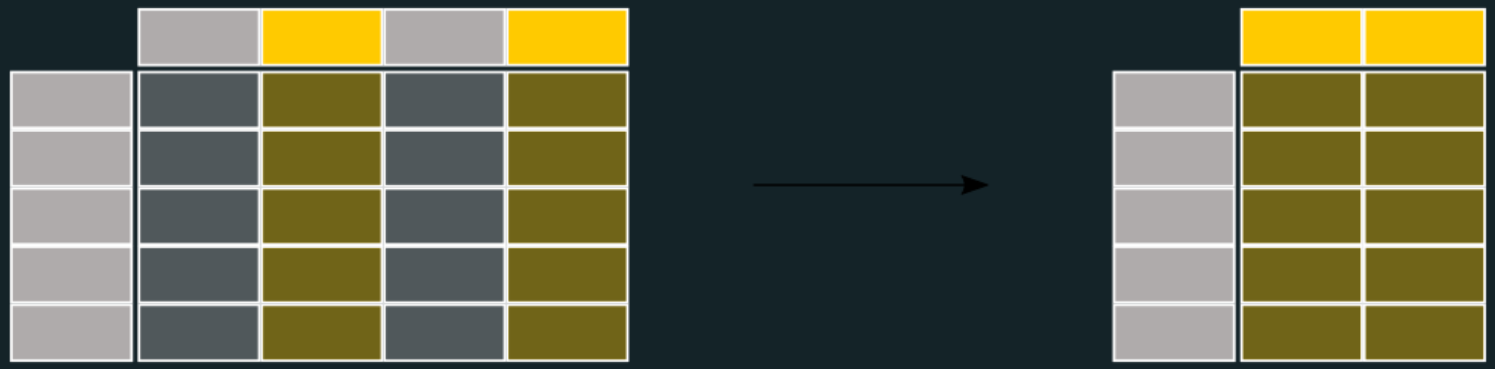
```
titanic = pd.read_csv("data/titanic.csv")
```

Exporting data

```
titanic.to_excel("titanic.xlsx", sheet_name="passengers", index=False)
```

1. Select / slicing

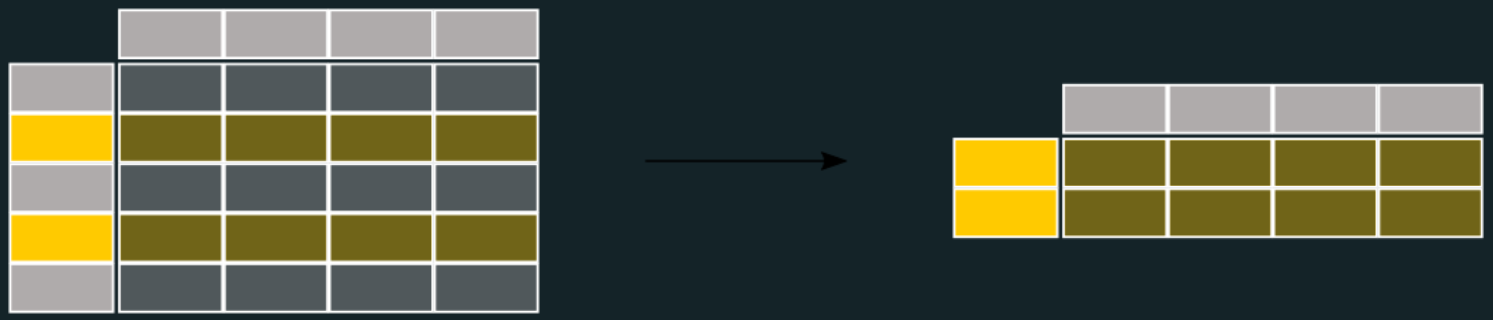
Select only some of the columns:



```
df["Name"]
df[["Name", "Age"]]
```

1. Select / slicing

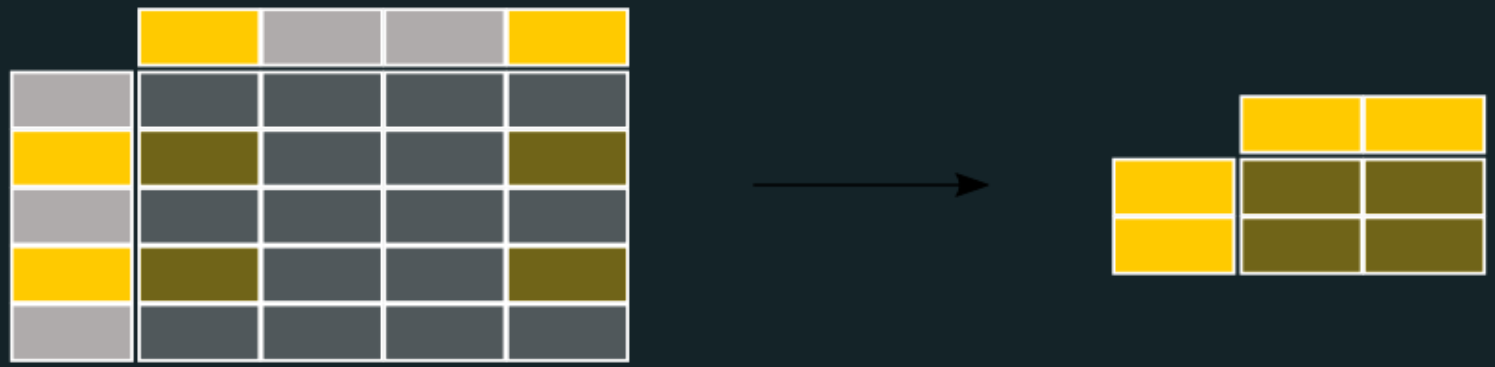
Select only some of the rows:



```
above_35 = titanic[titanic["Age"] > 35]
above_35.head()
```

1. Select / slicing

Select a combinations of rows and columns:



```
adult_names = titanic.loc[titanic["Age"] > 35, "Name"]
```

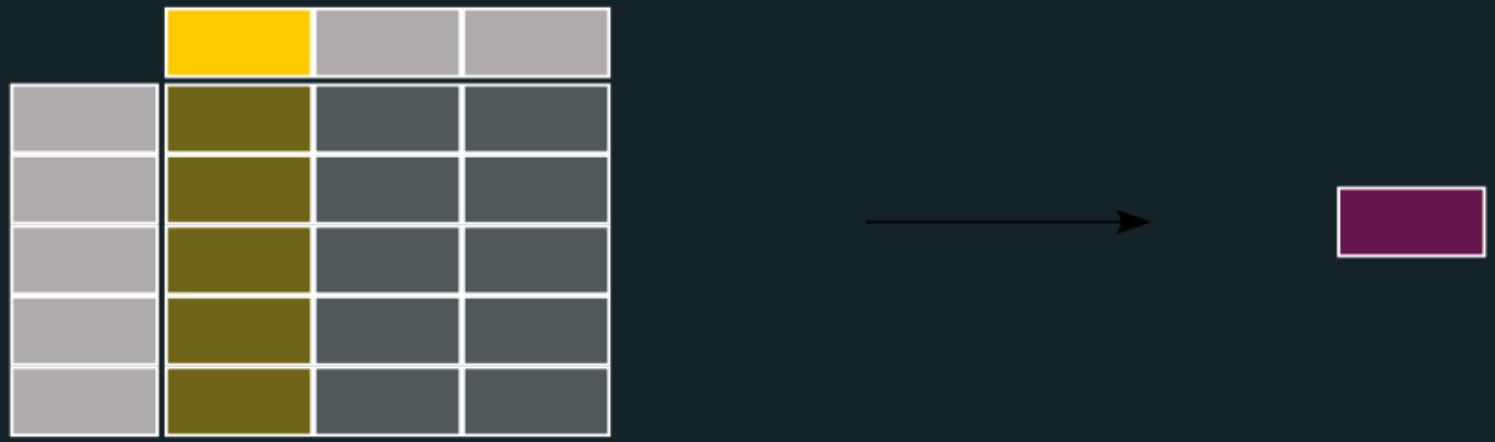
Challenge

Take a few minutes to :

1. Download this dataset and save it into a folder
2. Open a jupyter notebook in the same folder
3. Import the dataset in your notebook
4. Visualise the data (you can use df.head())
5. Create a new dataframe with only the country and points columns
6. Select all observations that are from Cyprus

2. Aggregate / reduce

Combine values across a column into a single value



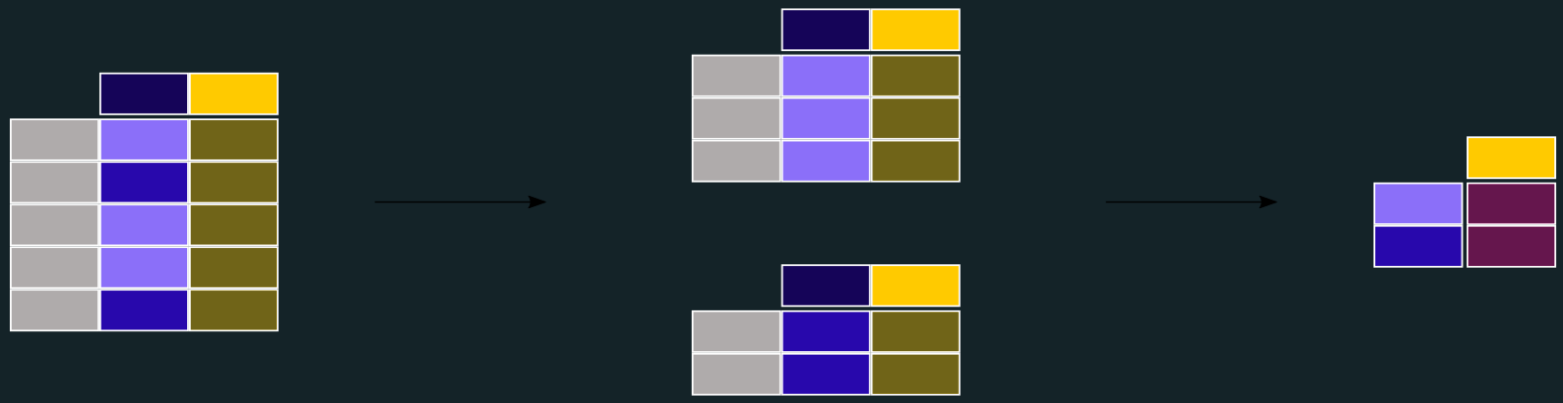
```
titanic["Age"].mean()
titanic[["Age", "Fare"]].median()
```

3. Map

Apply a function to every row, possibly creating more or fewer columns

4. Group by

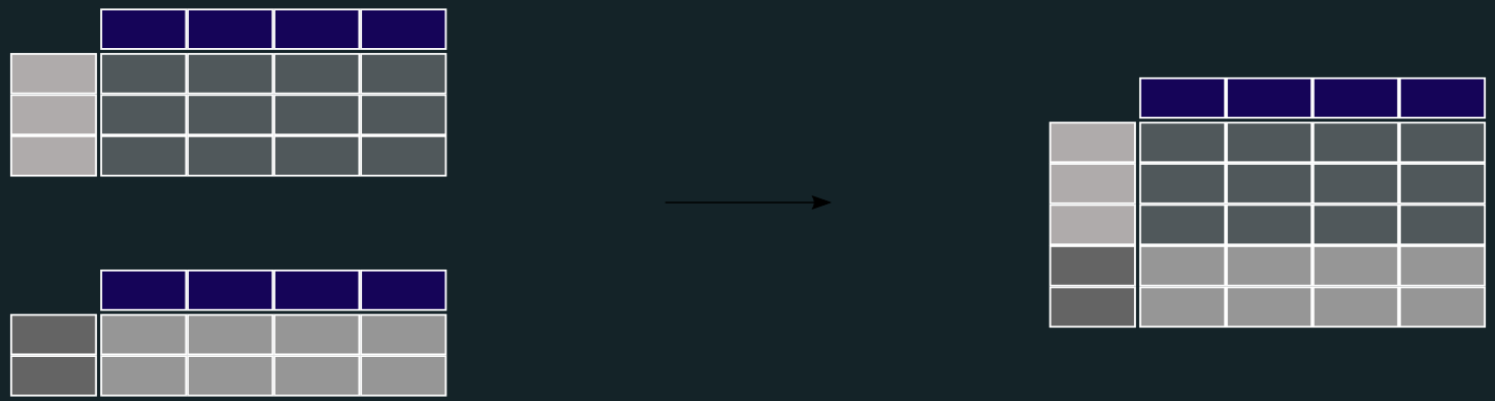
Apply a function to every row, possibly creating more or fewer columns



```
titanic[["Sex", "Age"]].groupby("Sex").mean()
```

5. Combine data from multiple tables

Concatenation



Merge

Exploratory data analysis

Preliminaries: cleaning data

1. Look at the data
2. Transform the variables into a known type
3. The type matters for what we do with them
    - For **aggregation**
        - flows ⇒ summmed
        - stocks ⇒ averaged
    - For **plotting the distribution**:
        - Qualitative ⇒ bar chart with frequencies
        - Quantitative ⇒ histogram

5 reasons to do EDA!

1. To check data cleaning (part of iterative process)
2. To guide subsequent analysis (for further analysis)
3. To give context of the results of subsequent analysis (for interpretation)
4. To ask additional questions (for specifying the (research) question)
5. Offer simple, but possibly important answers to questions.

Summary statistics + graphics

Summary statistics

- For any given variable, a *statistic* is a meaningful number that we can compute from a dataset.
- Basic *summary* statistics describe the most important features of distributions of variables.
- Example?

Distribution of a variable

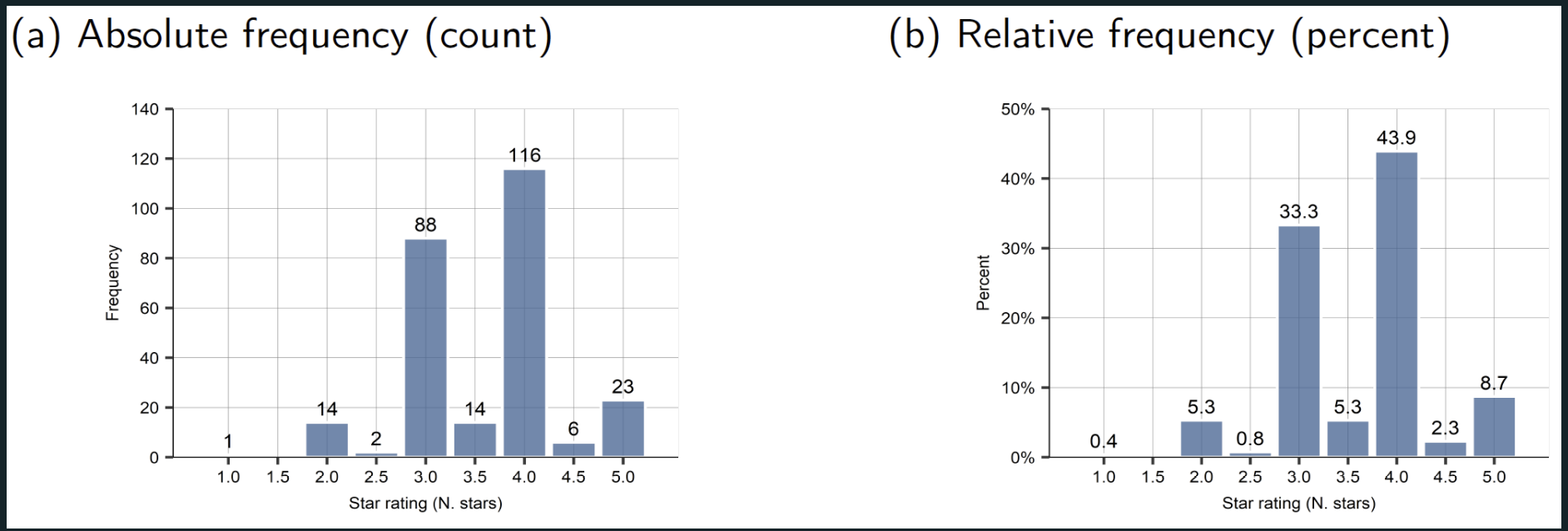
- All variables have a **distribution**
- The distribution of a variable tells the **frequency of each value** of the variable in the data
  - Absolute frequencies (number of observations)
  - Relative frequencies (percent of observations)
- Beware of **missing values**: proportion can be relative to all observations OR only observations with non-missing values (usual choice).

Histograms

Histogram reveals important properties of a distribution.

- Number and location of **modes**:
  - peaks in the distribution that stand out from their immediate neighborhood.
- Approximate regions for center and **tails**
- Symmetric or not
  - Asymmetric distributions have a long (left or right) tail
- **Extreme values**: values that are very different from the rest.

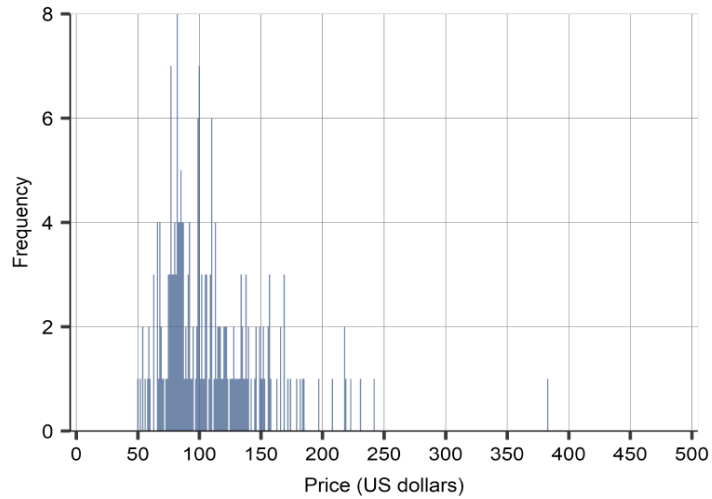
Hotel rating histograms



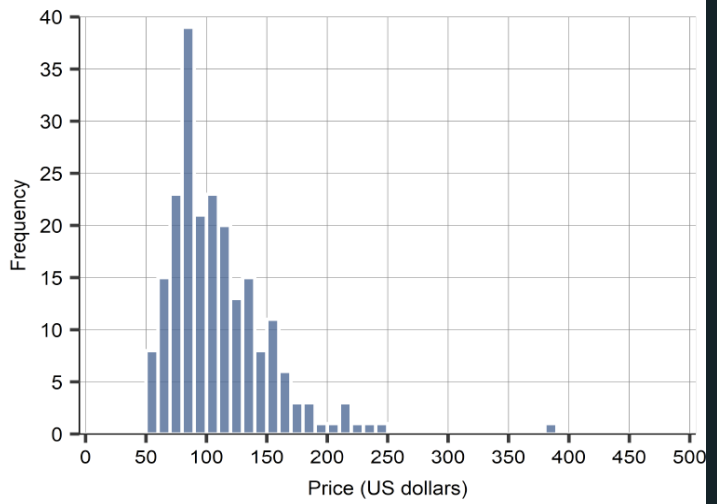
Source: hotels-vienna dataset. Vienna, Hotels only, for a 2017 November weekday

Hotel price histograms

(a) Histogram: individual values



(b) Histogram: 20$ bins



Source: hotels-vienna dataset. Vienna, Hotels only, for a 2017 November weekday

Theoretical distributions

Theoretical distributions can be helpful

- Have well-known properties!
- If variable in our data well approximated by a theoretical distribution –> attribute properties to the variable
- Real life, many variables surprisingly close to theoretical distributions.
- Will be useful when generalizing from data

Normal distribution

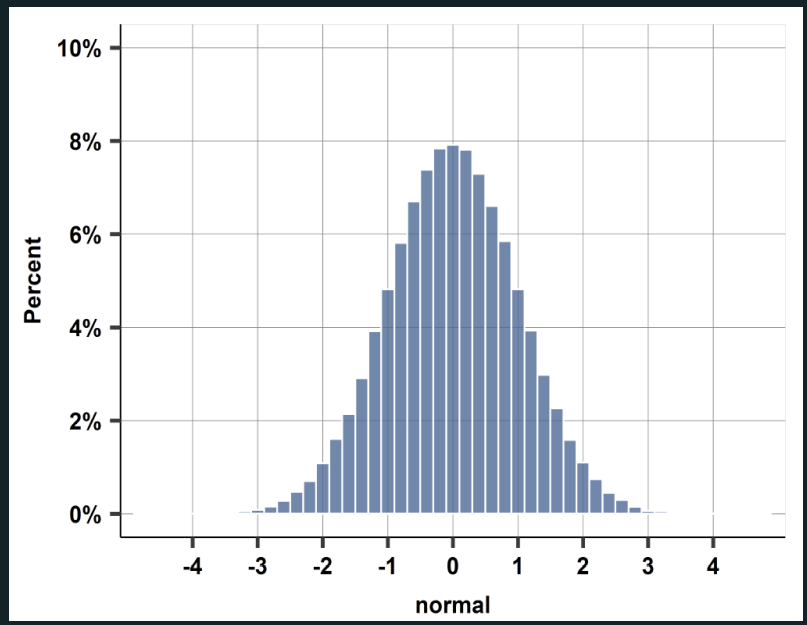
Histogram is bell-shaped

Distribution is captured by two parameters:

- μ is the mean
- σ is the variance

Symmetric = median, mean (and mode) are the same.

Example: height of people, IQs, ect.



The log-normal distribution

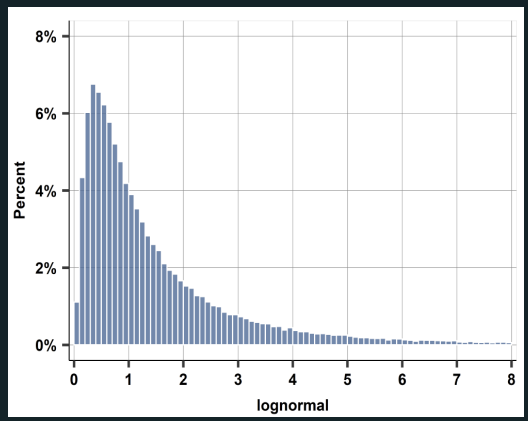
Asymmetrically distributed with long right tails.

Steps:

- start from a normally distributed r.v. (x),
- transform it: (e^x) and
- the resulting variable is distributed log-normal.

Always non-negative

Example distributions of income, or firm size.



Income and log-income

Figure: income

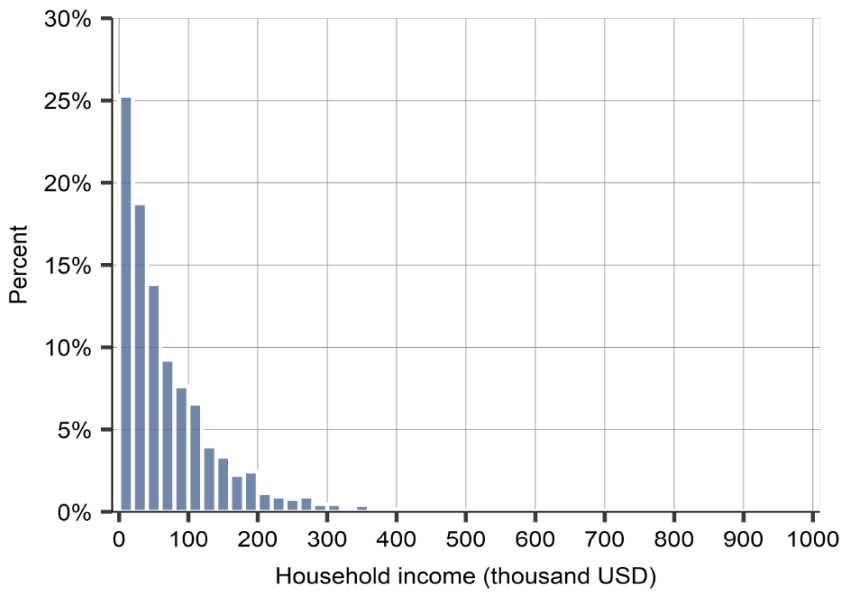
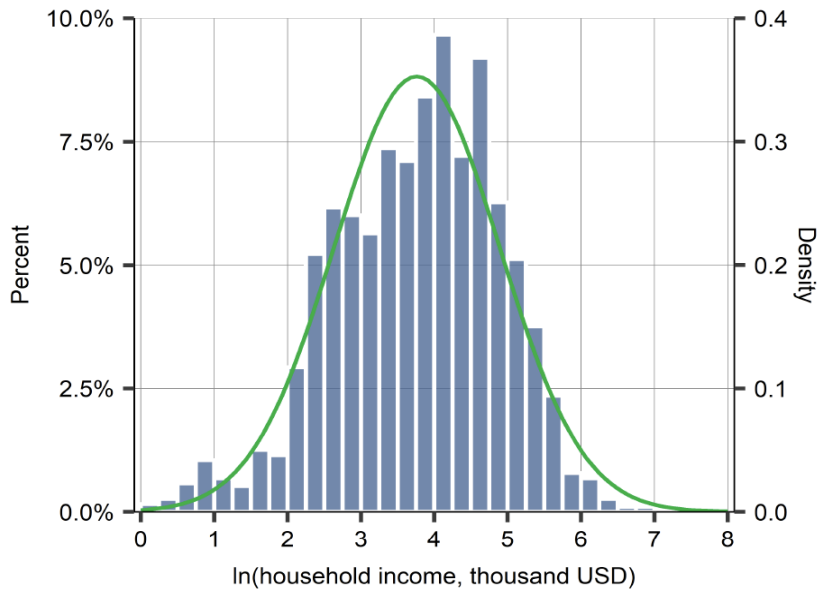


Figure: log income



The power law distribution

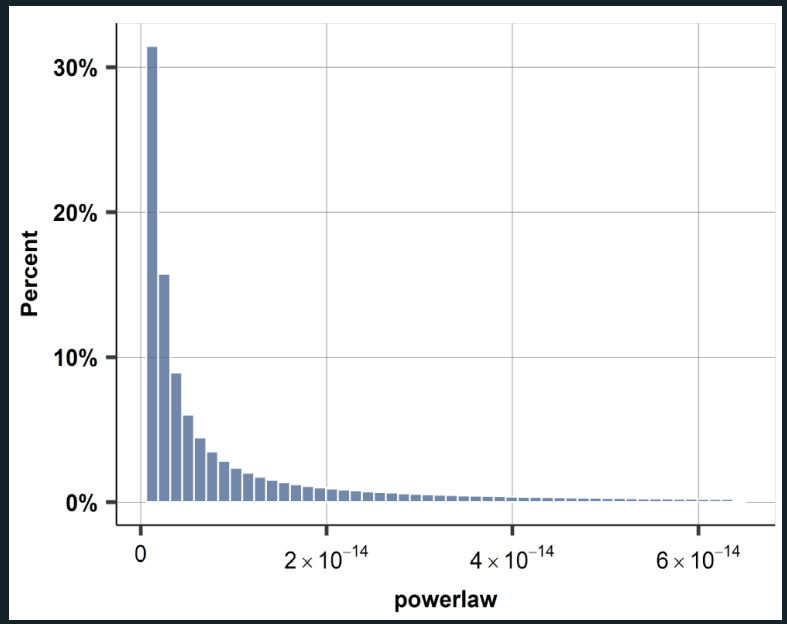
Also called as Pareto distribution

Very large extreme values - well approximated

Relative frequency of close-by values are the same along large and small values

Real world: many examples, but often not the whole distribution

Example: frequency of words, city population, wealth



Next

- Implementation using the introduction to pandas notebook
- More on summary statistics in the comparison & correlation lecture slides
- More on data visualisation principles and practice later.