Data Analysis 1

6 classes

1. origins data
2. preparing data
3. exploring data analysis, describing vatiables
4. comparison, correlation
5. generalizing from a dataset
6. testing hypotheses

DA1 is about exploration

60-80% of the time is collecting data, cleaning data

what is data:

* bunch of numbers used for argumentation

Data table (matrix):

* **observations** are the rows
* **Variables** are features or covariates (Columns)

storage : comma seperated values (can be anything: , ; or other)

pl:

**Pizza margareta prices Size of Margareta**

**Pizza forte** 1,200 Ft 32 cm  
 **online -> Data table**

**Don pepe** 1,300 Ft 30 cm

Data set

* containes several data tables (collection of data tables, typically related)

Data structures:

* cross- sectional (xsec): data have info on many units observed at the same time
* time series (tseries): on single unit observed many times
* multi-dimensional (panel): have multiple dimensions
  + - * many cross-sectional units observed many times
      * units observed in different space

Data quality: these big questions to be considered:

* Content: what is the substance a variable captures. Always check details
* Validity: is the content of variable close to intended content. "Durability" vs "Quality"
* Reliability: If we were to measure the same variable multiple times for the same observation it should give the same result.
* Comparability: in measurement across observations
* Coverage: Ideally complete coverage. In practice, they may not include all planned units (incomplete coverage).
* Unbiased selection: In incomplete coverage, observations that are included should be similar to all observations that were intended to be covered.

always understand the variables and observation -> do not need to memorize lists just understand know I’m doing

Data analysts should know their data:

* How data was born
* All details of measurement that may be relevant for their analysis
* To this end, consider having
  + - README.txt that describes where dataset comes from
    - VARIABLES.xls that provides basic information on your variables

Data Collection type: lot of info in ppt, check it!

* digital
* administrative

Representativeness:

* depends on the correct size
  + size: does it sufficient to answer your question?
* benchmarking and evaluation of data collection are the way to measure representativeness

Selection methods

1. Random sampling

Selection bias: The sample is different to the population