Data Management

Working with data

Malka Guillot

HEC Liège | ECON2306

Table of contents

What is data?

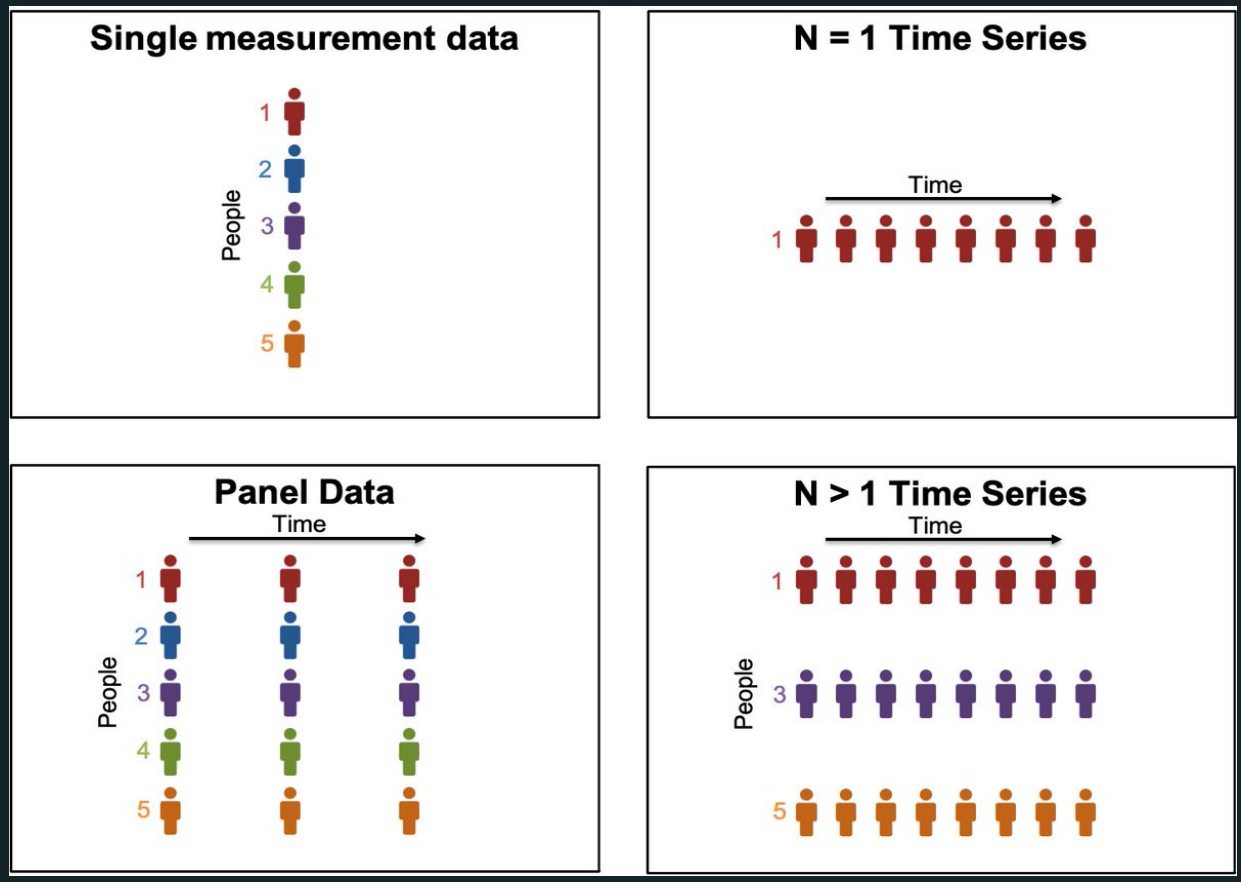
Structured data

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



Looking for data?

- Search engines:
  - google datasetsearch
  - https://fr.statista.com/
- Institutional reposities
  - Open data Liège
  - The official portal for European data
  - OECD data
  - Statbel and Open data Belgium
  - Open data France (also here)
- List of resources:
  - Common-built archive
  - Personal lists here, here or here
  - Registry of Open Data on AWS

You do not find what you are looking for?

Some solutions:

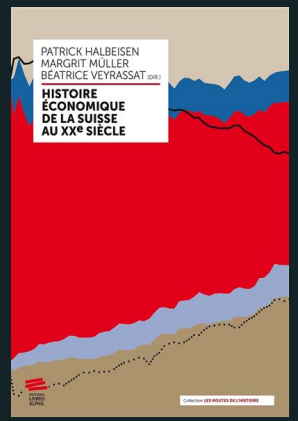
- Data collection
  - webscraping
  - asking for the data
  - administrative sources
- Implement a survey

Example

How did the under-developed and agricultural Switzerland from the 1850s become such an economically successful country?

A research question in history... where data are central

- Timing of the development of certain economic sectors?
- Geographical distribution of the firms?
- Interaction with public policies ?



Example: REFLEX project

- *Approach*: The Swiss Commercial Registry (1883-)

№ 1 — II. Theil 1883 IIe Partie — № 1

Schweizerisches Handelsamtsblatt

Feuille officielle suisse du commerce — Foglio ufficiale svizzero di commercio

Bern, 4. Januar — Berne, le 4 Janvier — Berna, li 4 Gennaio

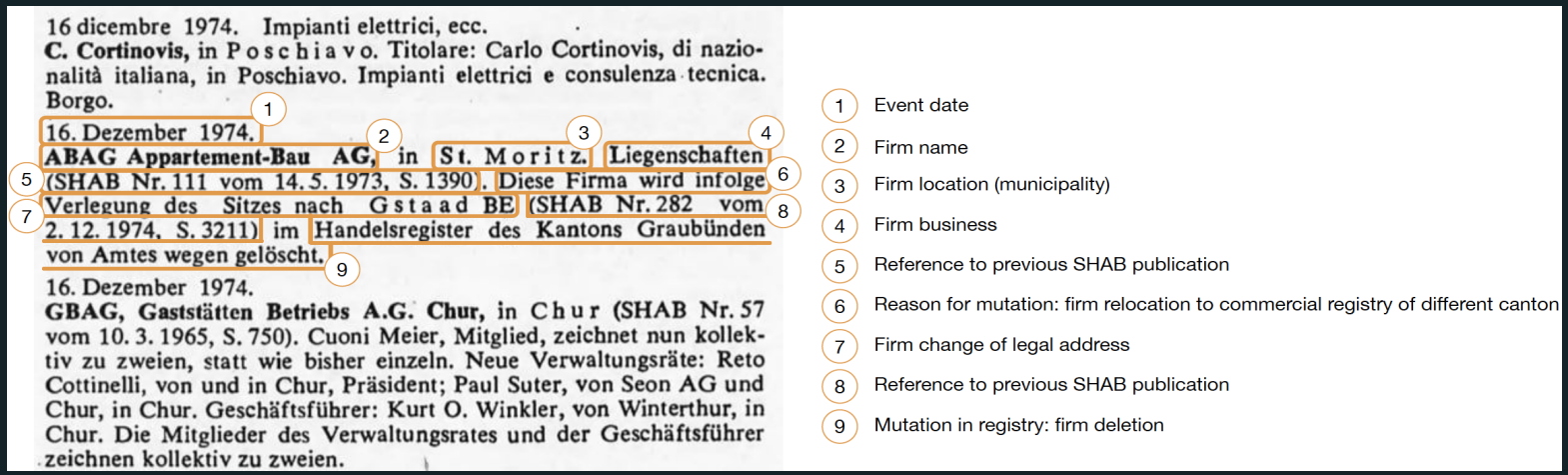
Publikationsorgan der eidgenössischen Departemente für Finanzen, Zoll und Handel
Organe de publicité des Départements fédéraux des Finances, des Péages et du Commerce
Organo di Pubblicità dei Dipartimenti federali per le Finanze, i Dazi ed il Commercio

Jährlicher Abonnementspreis Fr. 5. — Abonnemente nehmen alle Postämter sowie die Expedition des *Schweiz. Handelsamtsblattes* in Bern entgegen.
Abonnement annuel Fr. 5. — On s'abonne auprès des bureaux de poste et à l'expédition de la *Feuille officielle suisse du commerce* à Berne.
Prezzo delle associazioni Fr. 5. — Associazioni presso gli uffizi postali ed alla spedizione del *Foglio ufficiale svizzero di commercio* a Berna.

- *Methods*:
  - Transforming the pdf into structured data using:
    - google vision
    - Open AI
    - deep learning

Example: REFLEX project

- *Objectives*:



16 dicembre 1974. Impianti elettrici, ecc.
C. Cortinovis, in Poschiavo. Titolare: Carlo Cortinovis, di nazionalità italiana, in Poschiavo. Impianti elettrici e consulenza tecnica. Borgo.

(1) 16. Dezember 1974.
(2) **ABAG Appartement-Bau AG,** in (3) St. Moritz. (4) Liegenschaften (5) (SHAB Nr. 111 vom 14. 5. 1973, S. 1390). Diese Firma wird infolge (6) (7) Verlegung des Sitzes nach Gstaad BE (SHAB Nr. 282 vom 2. 12. 1974, S. 3211) im (8) Handelsregister des Kantons Graubünden von Amtes wegen gelöscht. (9)

16. Dezember 1974.
GBAG, Gaststätten Betriebs A.G. Chur, in Chur (SHAB Nr. 57 vom 10. 3. 1965, S. 750). Cuoni Meier, Mitglied, zeichnet nun kollektiv zu zweien, statt wie bisher einzeln. Neue Verwaltungsräte: Reto Cottinelli, von und in Chur, Präsident; Paul Suter, von Seon AG und Chur, in Chur. Geschäftsführer: Kurt O. Winkler, von Winterthur, in Chur. Die Mitglieder des Verwaltungsrates und der Geschäftsführer zeichnen kollektiv zu zweien.

1. Event date
2. Firm name
3. Firm location (municipality)
4. Firm business
5. Reference to previous SHAB publication
6. Reason for mutation: firm relocation to commercial registry of different canton
7. Firm change of legal address
8. Reference to previous SHAB publication
9. Mutation in registry: firm deletion

- *Challenges*:
  - Correcting for OCR mistakes
  - Building a firm identifyer
  - 3 languages
  - ...

Sampling

Exploratory data analysis

Variables types

What type of variable do you know?

Variables

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

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.

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)

Next

- Implementation using the introduction to pandas notebook
- More on summary statistics in the comparison & correlation lecture slides