Data Transformations

with dplyr

Chuvakin Sergey

«School of Advanced Studies»

November 2, 2020

Outline

- Where data can be obtained?
- LTE process
- Relational database
- What is ID, what is row
- Tidy data
- Cheatsheat - what should be kept in mind

Data Sources

Where I can get data?

- Self gathering
  - Interview
  - Survey
  - Obsevation
- Public data
  - http://crimestat.ru/
  - https://rosstat.gov.ru/ (former gks)
  - WHO data
  - World Values Surevey
  - European Social Survey
- Data warehouse, data mart, data lakes - internal statistics of some company

Data Sources

ETL processes

Extract, Transform, Load - common process in business companies. All the company data are stored in spacial place named *Database*. Usually enterprises use Structured Query Language (SQL) to store and view the data. The raw data (recently obtained, non transformed) named Data Lake, while Data Warehouse and Data Mart are terms for transformed data (Note: terms maybe different in deferent companies).

ETL - process of moving and transforming the data from one source to another. It's essential part data science in enterprise. See OLAP and OLTP for better understanding.

Data sctructure

Relational data

name	age	country
Natalia	11	Iceland
Ned	6	New York
Zenas	14	Ireland
Laura	8	Kenya

Data sctructure

Relational data

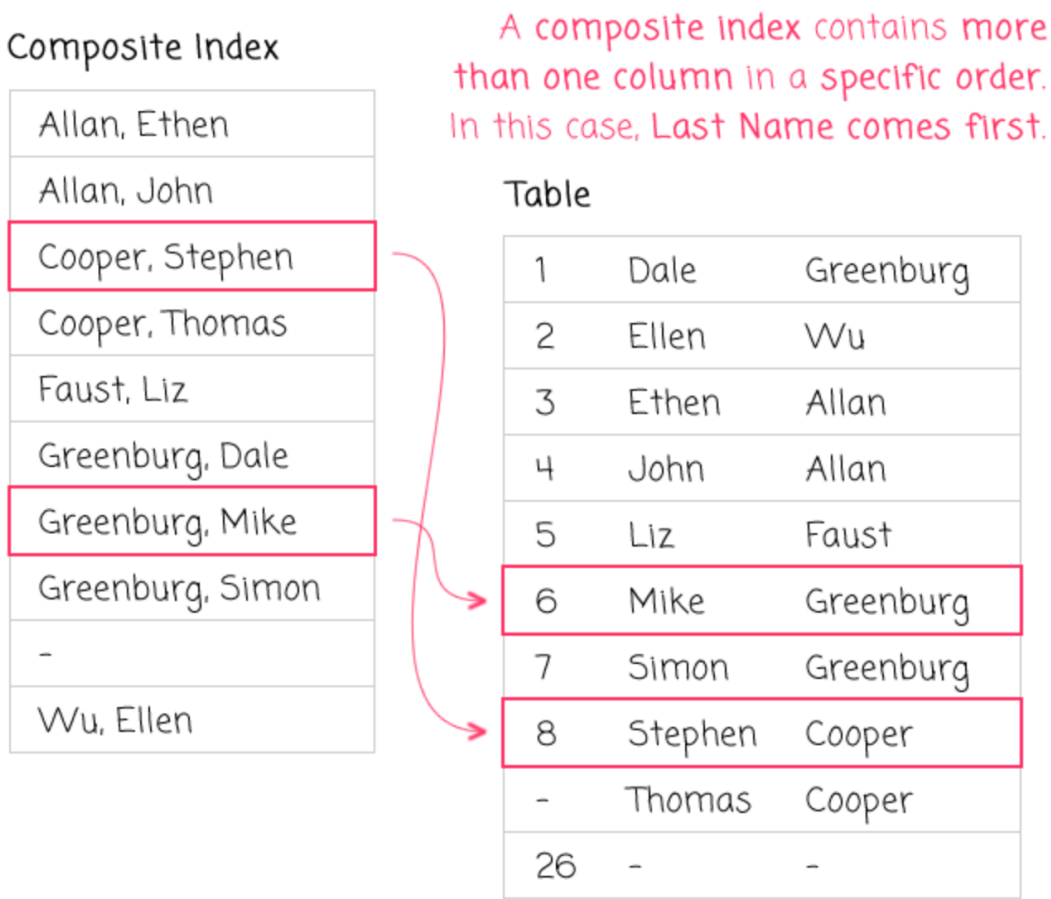
Each row - observation.
Each column - feature.
Otherwise - it's not relational database.

What it gives to social science?
Each row - it's separate vector aka separate person with set of features like sex, age, political preferences et etc.

Data sctructure

Relational data

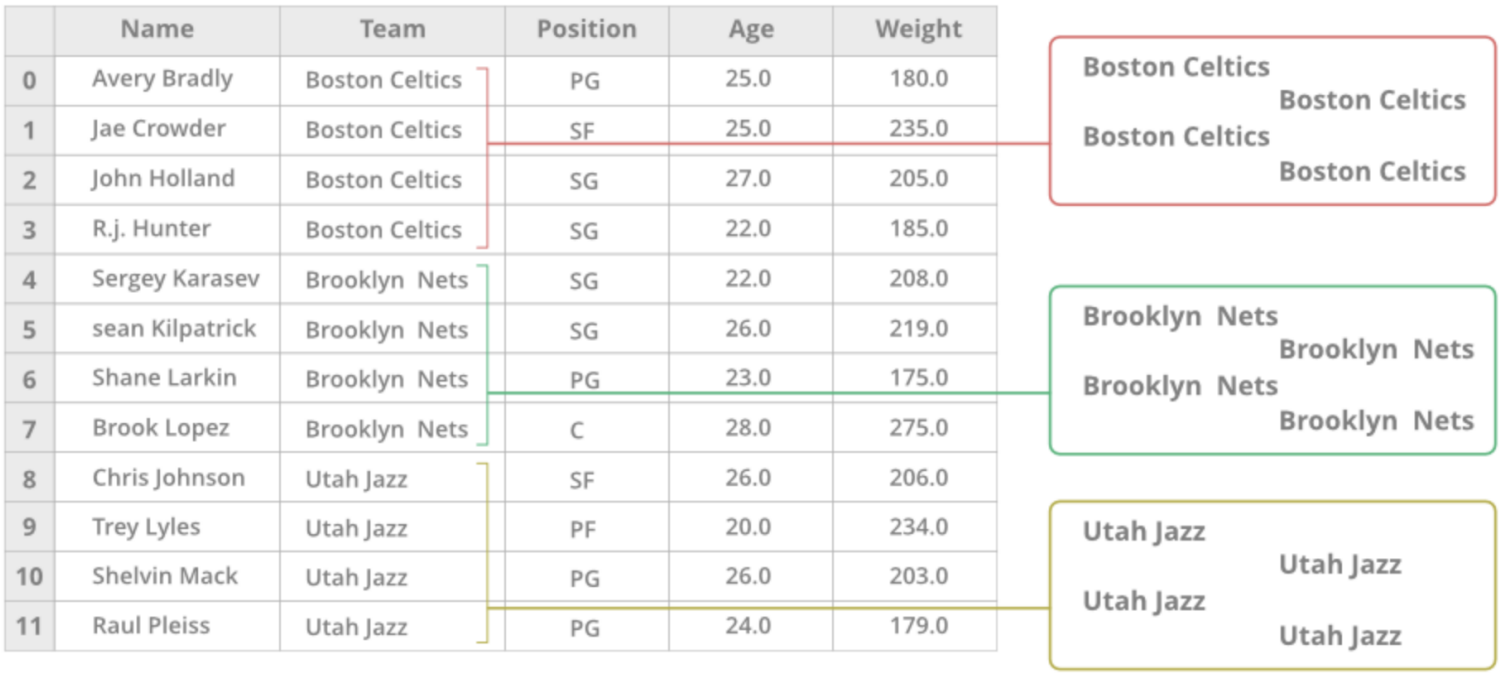
Each row by default - is unique value, but it's not always true.



Data sctructure

Relational data

Sometimes you'll need to group your data to get one observation per row.



	Name	Team	Position	Age	Weight
0	Avery Bradly	Boston Celtics	PG	25.0	180.0
1	Jae Crowder	Boston Celtics	SF	25.0	235.0
2	John Holland	Boston Celtics	SG	27.0	205.0
3	R.j. Hunter	Boston Celtics	SG	22.0	185.0
4	Sergey Karasev	Brooklyn Nets	SG	22.0	208.0
5	sean Kilpatrick	Brooklyn Nets	SG	26.0	219.0
6	Shane Larkin	Brooklyn Nets	PG	23.0	175.0
7	Brook Lopez	Brooklyn Nets	C	28.0	275.0
8	Chris Johnson	Utah Jazz	SF	26.0	206.0
9	Trey Lyles	Utah Jazz	PF	20.0	234.0
10	Shelvin Mack	Utah Jazz	PG	26.0	203.0
11	Raul Pleiss	Utah Jazz	PG	24.0	179.0

Data sctructure

Tidy Data

«Data tidying» - making data clean and tidy, ready for analysis and modeling.
Some rules to keep in mind:

- One row should be one obeservaton of your research question
- Data should avoid redundancy (data duplicated in several places)
- Variables (features) should be only on columns, otherwise to much missings are in data.

Don't worry if smth is not clear - it becomes transparent with exprience. Obligatory reading in English and in Russian.

Data sctructure

Cheatsheet

Some things to check before analysis

- Encoding! Does every symbols were read.
- Try to avoid spacial symbols like: $%* , any slashes, white spaces in columns naming.
- Cyrillic symbols it's mauvais ton because of problems with encodings in diferents OS.
- Use sommon formats like csv, tsv or at least xlsx
- Try to figure out what means every column in your data.
- Explore what is unique ID in your data!
- Explore how your data could be grouped
- Explore basic statistics of your data
- Plot all the variables, discover the data types in your data.