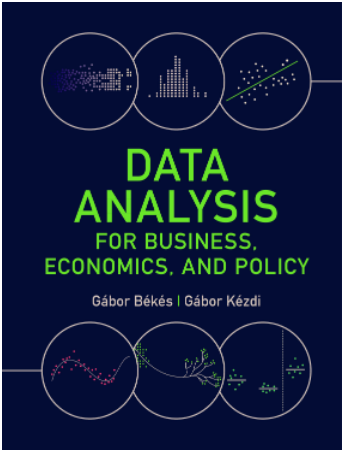
02 Preparing data for analysis

Gábor Békés

Data Analysis 1: Exploration

2023

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021
- gabors-data-analysis.com
  - Download all data and code: gabors-data-analysis.com/data-and-code/
- This slideshow is for Chapter 02

Motivation

- Does immunization of infants against measles save lives in poor countries? Use data on immunization rates in various countries in various years from the World Bank. How should you store, organize and use the data to have all relevant information in an accessible format that lends itself to meaningful analysis?

- You want to know, who has been the best manager in the top English football league. Have downloaded data on football games and on managers. To answer your question you need to combine this data. How should you do that? And are there issues with the data that you need to address?

Variable types: Qualitative vs quantitative

- Data can be born (collected, generated) in different form, and our variables may capture the quality or the quantity of a phenomenon.
- Quantitative variables are born as numbers. Typically take many values.
  - also called numeric variables
  - special case is time (date)
- Qualitative variables, also called categorical variables, take on a few values, with each value having a specific interpretation (belonging a category).
  - Another name used is categorical or factor variable.
  - binary variable (YES/NO) is special case.

Variable types - binary

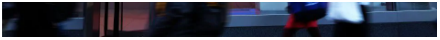
- A special case is a binary variable, which can take on two values
- ...yes/no answer to whether the observation belongs to some group. Best to represent these as 0 or 1 variables: 0 for no, 1 for yes.
- Sometimes binary variables with 0-1 values are called dummy variables or an indicator

- Flag - binary showing existence of some issue (such as missing value for another variable, presence in another dataset)

Variable types - formal definition

1. Nominal qualitative variables take on values that *cannot* be unambiguously ordered. Color, brands
2. Ordinal, or ordered variables take on values that are *unambiguously ordered*. All quantitative variables can be ordered; some qualitative variables can be ordered, too. Grades
3. "Interval" variables are ordered variables, with a *difference between values that can be compared*. Degree Celsius. Price in dollar.
4. "Ratio" (or "scale") variables are interval variables with the additional property: their ratios mean the same regardless of the magnitudes. This additional property also implies a meaningful zero in the scale. Distance in miles. Price in dollar.

Storing variables: Example the Washington Post (2016)



(Jewel Samad/AFP/Getty Images)

By **Christopher Ingraham**
August 26, 2016

A surprisingly high number of scientific papers in the field of genetics contain errors introduced by Microsoft Excel, according to an analysis recently published in the journal Genome Biology.

A team of Australian researchers analyzed nearly 3,600 genetics papers published in a number of leading scientific journals — like Nature, Science and PLoS One. As is common practice in the field, these papers all came with supplementary files containing lists of genes used in the research.

The Australian researchers found that roughly 1 in 5 of these papers included errors in their gene lists that were due to Excel automatically converting gene names to things like calendar dates or random numbers.

[This new model for training scientists could create a conflict of interest]

You see, genes are often referred to in scientific literature by symbols — essentially shortened versions of full gene names. The gene "Septin 2" is typically shortened as SEPT2. "Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase" gets mercifully shortened to MARCH1.

What you type	What you see	How Excel stores it
MARCH1	1-MAR	42430
SEPT2	2-SEP	42615

https://www.washingtonpost.com/news/wonk/wp/2016/08/26/an-alarming-number-of-scientific-pa

Data wrangling (data munging)

Data wrangling is 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.

[1] Understanding and storing

- start from raw data
- understand the structure and content
- create tidy data tables
- understand links between tables

[2] Data cleaning

- understand features, variable types
- filter duplicates
- look for and manage missing observations
- understand limitations

The tidy data approach

A useful concept of organizing and cleaning data is called the *tidy data* approach:

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:

- standard data tables that turn out to be easy to work with.
- finding errors and issues with data are usually easier with tidy data tables
- transparent, which helps other users to understand
- easy to extend. New observations added as new rows; new variables as new columns.

Simple tidy data table

Table: A simple tidy table

	Variables/columns		
	hotel_id	price	distance
Observations/rows	21897	81	1.7
	21901	85	1.4
	21902	83	1.7

Source: hotels-vienna data. Vienna, 2017 November weekend.

Tidy data table of multi-dimensional data

- The *tidy approach* - store xt data in data tables with each row referring to one cross-sectional unit observed in one time period.
- One row is one observation '*it*'.
- This is sometimes called the *long format* for xt data.
- The next row then may be the same cross-sectional unit observed in the next time period.
- Important and difficult task for analysts is to figure out the structure of multi-dimensional data and create tidy data tables.

- Also used: *wide format* - one row would refer to one cross-sectional unit, and different time periods are represented in different columns. Good for presenting and some analysis. Not to keep data.

Displaying immunization rates across countries

- xt panel of countries with yearly observations,
- Downloaded from the World Development Indicators data website maintained by the World Bank.
- Illustrate the data structure focusing on the two ID variables (country and year) and two other variables, immunization rate and GDP per capita.

Displaying immunization rates across countries – WIDE

Country	imm2015	imm2016	imm2017	gdppc2015	gdppc2016	gdppc2017
India	87	88	88	5743	6145	6516
Pakistan	75	75	76	4459	4608	4771

Wide format of country-year panel data, each row is one country, different years are different variables. imm: rate of immunization against measles among 12–13-month-old infants. gdppc: GDP per capital, PPP, constant 2011 USD. Source: `world-bank-vaccination` data

Displaying immunization rates across countries – LONG

Country	Year	imm	gdppc
India	2015	87	5743
India	2016	88	6145
India	2017	88	6516
Pakistan	2015	75	4459
Pakistan	2016	75	4608
Pakistan	2017	76	4771

Note: Tidy (long) format of country-year panel data, each row is one country in one year. imm: rate of immunization against measles among 12–13-month-old infants. gdppc: GDP per capital, PPP, constant 2011 USD. Source: `world-bank-vaccination` data.

Relational database

- The relational database is a concept of organizing information.
- It is a data structure that allows you map a concept set of information into a set of tables
- Each table is a made up of rows and columns
- Each row is a record (observation) identified with a unique identifier *ID* (also called *key*).
- Rows (observations) in a table can be linked to rows in other tables with a column for the unique ID of the linked row (*foreign ID*)

- Define these tables, understand structure
- Merge tables when needed

Identifying successful football managers

- Who have been the best football managers in England?
- We combine data from two sources for this analysis, one on teams and games, and one on managers.
- Data covers 11 seasons of English Premier League (EPL) games – 2008/2009 to 2018/2019
- The data comes from the website `www.football-data.co.uk`.
- Each observation is a single game. Key variables are
  - the date of the game
  - name of the home team, the name of the away team,
  - goals scored by the home team, goals scored by the away team

Identifying successful football managers

Table: Games data

Date	HomeTeam	AwayTeam	Home goals	Away goals
2018-08-19	Brighton	Man United	3	2
2018-08-19	Burnley	Watford	1	3
2018-08-19	Man City	Huddersfield	6	1
2018-08-20	Crystal Palace	Liverpool	0	2
2018-08-25	Arsenal	West Ham	3	1
2018-08-25	Bournemouth	Everton	2	2
2018-08-25	Huddersfield	Cardiff	0	0

Source: `football` data.

Identifying successful football managers

- Is this a tidy data table?

Identifying successful football managers

- Is this a tidy data table?
- It is.
- Each observation is a game, and each game is a separate row in the data table.
- Three ID variables identify each observation: date, home team, away team. The other variables describe the result of the game.
- From the two scores we know who won, by what margin, how many goals they scored, and how many goals they conceded.

Identifying successful football managers

- Could we have an alternative tidy table?

Identifying successful football managers

- Could we have an alternative tidy table?
- There is an alternative way to structure the same data table, which will serve our analysis better
- In this data table, each row is a game played by a team.
- It includes variables from the perspective of that team: when played, who the opponent was, and what the score was.

Identifying successful football managers

Table: Games data - long table version

Date	Team	Opponent team	Goals	Opponent goals	Home/away	Points
2018-08-19	Brighton	Man United	3	2	home	3
2018-08-19	Burnley	Watford	1	3	home	0
2018-08-19	Man City	Huddersfield	6	1	home	3
2018-08-19	Man United	Brighton	2	3	away	0
2018-08-19	Watford	Burnley	3	1	away	3
2018-08-19	Huddersfield	Man City	1	6	away	0

Identifying successful football managers

- Also a tidy data table, albeit a different one.
- It has twice as many rows as the original data table: Each game appears twice in this data table, once for each of the playing team's perspectives.
- New variable to denote whether the team at that game was the home team or the away team.
- Now we have two ID variables, one denoting the team, and one denoting the date of the game. The identity of the opponent team is a qualitative variable.

- Tidy data has some key features. But a given multi-dimensional data may be stored as tidy in multiple ways.

Identifying successful football managers

- Our second data table is on managers.
- One row is one manager-team relationship.
- Each manager may feature more than once in this data table if they worked for multiple teams.
- For each observation, we have the name of the manager, their nationality, the name of the team (club), the start time of the manager's work at the team, and the end time.

Identifying successful football managers

Table: Managers data

Name	Nat.	Club	From	Until
Arsene Wenger	France	Arsenal	1 Oct 1996	13 May 2018
Unai Emery	Spain	Arsenal	23 May 2018	Present*
Ron Atkinson	England	Aston Villa	7 June 1991	10 Nov 1994
Brian Little	England	Aston Villa	25 Nov 1995	24 Feb 1998
John Gregory	England	Aston Villa	25 Feb 1998	24 Jan 2002
Dean Smith	England	Aston Villa	10 Oct 2018	Present*
Alan Pardew	England	Crystal Palace	2 Jan 2015	22 Dec 2016
Alan Pardew	England	Newcastle	9 Dec 2010	2 Jan 2015

Source: `football` data. Present = 01 July 2019

Identifying successful football managers

- This is a relational dataset.
- One data table with team-game observations, and one data table with manager-team observations.
- To work with the data, we need to create a workfile, which is a single data table that is at the team-game level with the additional variable of who the manager was at the time of that game.
- but before we do, need need to merge them...

Relational data and linking data tables across observations

- Organize and store data in tidy data tables with appropriate ID variables,
- how to combine such table into a workfile to run our analysis?
- The process of pulling different variables from different data tables for well-identified entities to create a new data table is called linking, joining, merging, or matching.

Relational data and link data tables across observations

Matching (joining) depends on data structure

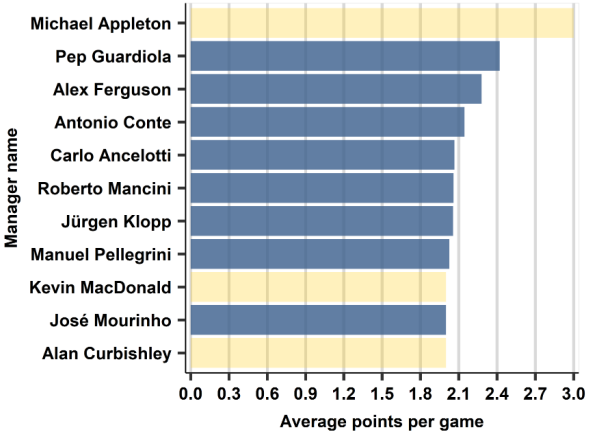
- one-to-one (1:1) matching: merging tables with the same type of observations.
  - Football teams and stadium now.
- many-to-one (m:1) or one-to-many (1:m) matching when in one of the data tables, a value may be matched to more than one values in the other table.
  - Football teams and their players now - many players in a team.
- many-to-many (m:m) matching when values in both tables could be matched to many others.
  - Football teams and their manager ever - some managers worked for multiple teams.

Identifying successful football managers

- Started with a relational dataset.
- Merged two data tables and created a work file
- With the workfile at hand, we can describe it.
- The workfile has 8360 team-game observations: in each of the 11 seasons, 20 teams playing 38 games (19 opponent teams twice; $11 \times 20 \times 38 = 8360$).
- There are 137 managers in the data.

Identifying successful football managers

- Remember: data is 11 seasons, EPL.
- spells at teams: if a manager worked for two teams, we consider it two cases.
- Success: average points per game
- Above 2.0



Complex data - tidy data- summary

- Creating a tidy data - generating a set of data tables that are easy to understand, combine and extend in the future.
- If relational data, IDs are essential
- Often raw data will not come in a tidy format, and you will need to work understanding the structure, relationships and find the individual ingredients.
- For analysis work, need to combine tidy data tables

AI and data structure

Generative AI (LLM) is good and helpful

- understands your data structure
- helps you combine datasets (suggests code / does it)
- helps even summarize the data

AI and data structure

Generative AI (LLM) is good and helpful

- understands your data structure
- helps you combine datasets (suggests code / does it)
- helps even summarize the data

Pay attention

- Check and debug a lot
- It does not know what you want. May need keep control.
- Best is to be very specific and focus on help with coding

Data wrangling: cleaning

- Entity resolution:
  - Dealing with duplicates
  - ambiguous identification
  - non-entity rows
- Missing values

Wrangling: Filter out Duplicates

- *duplicates*: some observations appearing more than once in the data.
- Duplicates may be the result of human error (when data is entered by hand), or the features of data source (e.g., data scraped from classified ads with some items posted more than once).
- Often, easy process. Just check and get rid of repeated observations
- Sometimes, same observation is featured number of times.
  - Need to investigate. Makes sense / an error?
  - Example: Daily stock quotes, some stock features twice, with different price.
  - Decision- what to keep. Sometimes no clear-cut way, but usually no big deal.

Entity identification and resolution

- More generally, you would need to have unique IDs
- It could be that two observations belong to two entities although ID is the same.
  - example: John Smith – there may be many
  - need to figure out, maybe assign unique IDs in raw data
- It could be that two observations have different ID but belong to same entity
  - need to figure out and have a single ID
- Unique IDs crucial. Numerical IDs are better

Entity resolution example

Team ID	Unified name	Original name
19	Man City	Manchester City
19	Man City	Man City
19	Man City	Man. City
19	Man City	Manchester City F.C.
20	Man United	Manchester United
20	Man United	Manchester United F.C.
20	Man United	Manchester United Football Club
20	Man United	Man United

Source: various sources

Getting rid of non-entity observations

- Rows that do not belong to an entity we want in the data table.
- Find them and drop them
- Such as: a summary row in a table that adds up, or averages, variables across all, or some, entities.
- Case study: a data table downloaded from the World Bank on countries often includes observations on larger regions, such as Sub-Saharan Africa

Missing values

- A frequent and important issue with variables is *missing values*.
- Missing values mean that the value of a variable is not available for some, but not all, observations.

- Scope: How much missing?
- Reason: Why missing?

Missing values: scope and reason

Key issues

1. Look at content of data - related to data quality (esp. coverage)
2. Missing values need to be identified.
   - Easy: "NA" (for "not available"), a dot ".", an empty space "".
   - Hard - binary 0 for no, 1 for yes, 9 for missing
   - Hard - percent 0-100, 9999 for missing
   - Hard numeric, range is 1-100000, 9999999999 for missing
3. Missing value is encoded. But when aggregate must pay attention.
4. Missing values should be counted. Missing values mean fewer observations with valid information. May actually have a lot fewer observations to work with than the size of the original dataset.
5. The third issue is potential selection bias. Is data missing at random?

Missing values - Understanding the selection process

- Random: When missing data really means no information, it may be the result of errors in the data collection process. Rare.
- In some other cases, missing just means "zero" or "no". In these instances, we should simply recode (replace) the missing values as "zero" or as "no".
- Often, values are missing systematically. Some survey respondents may not know the answer to a question or refuse to answer it, and such respondents are likely to be different from those who provide valid answers.

Missing values: what can we do?

Two basic options:

1. Restrict the analysis to observations with non-missing values for all variables used in the analysis.
2. *Imputation* - Fill in some value for the missing values, such as the mean or median.

Missing values: Some practical advice

- Focus on more fully filled variables. Often, simpler.
- Sometimes, informative if missing - create a new variable (called flag) to capture missing value and use this variable instead of the original.
  - For example, the original variable is a text of the Twitter handler.
  - Here, the binary variable that is 1 if the person has an account and 0 otherwise could be more interesting.
- Automatic missing variable filling packages. Advanced only.
- Always be conservative, impute if absolutely necessary!

Missing values: Some practical advice

- For binary variables: zero if yes/no.
- For qualitative nominal variables, you may add missing as a new value: white, blue red and missing.
  - What if binary qualitative?
- For ordinal variables, you may add missing as new value or recode missing to a neutral variable: high, average, low, with missing recoded as average.
- For quantitative variables - you may recode with mean or median
- if impute, create a flag and use it analysis

- Always be conservative, impute if absolutely necessary!

Practical data management

- Structure of files
- Naming files

Naming files

- Use file names that simple, machine readable. No spaces, punctuation(.,;), accented characters or capitalized letters
- Human readable. Deliberate use of "_" (part of info) and "-" (help read)
- Computer ordering friendly. Use numbers early on. Left pad (01 not 1)
- dates used in ISO: YYYY-MM-DD.
- Good examples
  - "bekes-kezdi_textbook-presentation_ceu2018.pdf
  - "ch02_organizing-data_world-bank-download_2017-06-01.tex"
- Never use:
  - thesis.pdf, mytext.doc
  - calculations1112018.xls, Gábor's-01.nov.19_work.pdf.

Structure of files

It is good practice to structure the data files at three levels.

- Raw data files
- Clean and tidy data files
- Workfile(s) for analysis

You may also organize your folders - maybe even before we start the analysis

- Wrangling (code)
- Analysis (code)
- Output (graphs, tables)
- Report (pdf, markdown)

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. Make sure one row in the data is one observation and manage duplicates
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

AI and data wrangling

Generative AI (LLM) is good and helpful

- understands your variables
- finds potential problems

https://chat.openai.com/share/c1f54966-d1ce-4412-acd6-d2f16f9abb53

AI and data wrangling

Generative AI (LLM) is good and helpful

- understands your variables
- finds potential problems

https://chat.openai.com/share/c1f54966-d1ce-4412-acd6-d2f16f9abb53

Pay attention

- You shall know more and may want different steps
- It does not know what you want. May need keep control.
- Eventually copy code – keep reproducible