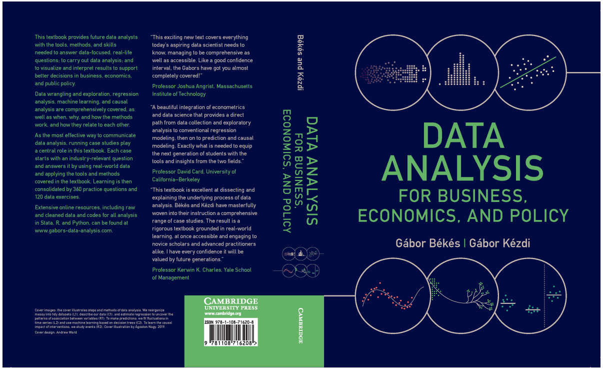
Data Analysis is a Process

Gábor Békés (Central European University and CEPR)

University College London – *SSDS Seminar*

7 December 2021

This talk is based on my Data Analysis textbook



- Cambridge University Press, 2021
- cambridge.org/bekeskezdi
- gabors-data-analysis.com
- github.com/gabors-data-analysis/da_case_studies

- Follow us
  - twitter.com/Gabors_Data
  - facebook.com/gaborsdata

Data Analysis is a process

1. First comes a research topic and a specific question
2. Data collection is the foundation for all empirical work
3. Cleaning and organizing the data is a necessary and time-consuming part
4. Exploratory data analysis helps both data preparation and analysis
5. Analytical work tests hypotheses and estimates model(s)
6. Results shall be presented in a user friendly way
7. Finally, we answer the original question and discuss generality

1 It starts with a question: From a topic to x, y and z

- Look for a topic that you care about / genuinely curious about the result
- Find a specific question - often about some relationship

- Translate to a causal question - think about an intervention
  - Data comes from an RCT experiment - easiest. Random assignment
  - Analysis is based on a natural experiment - hard to find, easy to do
  - Observational data - easiest to find, hardest to analyze

1 It starts with a question: From a topic to x, y and z

- Look for a topic that you care about / genuinely curious about the result
- Find a specific question - often about some relationship

- Translate to a causal question - think about an intervention
  - Data comes from an RCT experiment - easiest. Random assignment
  - Analysis is based on a natural experiment - hard to find, easy to do
  - Observational data - easiest to find, hardest to analyze

- Find an y (outcome) and x (treatment, causal variable)
  - Must start about measurement at the start, too

- With observational data, we must isolate the causal effect in the hard way
  - think about z variables that may prevent a causal analysis (such as confounders)

1 Case study: From a topic to x, y and z

- What makes some firm have better management?
- Founder / family ownership and management quality

1 Case study: From a topic to x, y and z

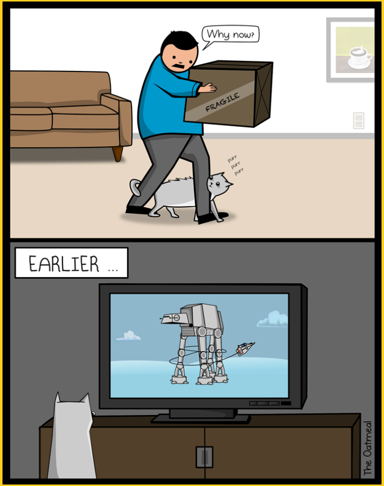
- What makes some firm have better management?
- Founder / family ownership and management quality

- Does having founders as owners make firm have a better management?
  - Thought experiment: take founder owned firms, and randomly sell stakes and see what happens later

- y (outcome) is management quality, and x (treatment) is ownership
- Confounders z: Institutions...
- Using data collected by a survey that measures management quality

1 Key point: Have an interesting question and measure it

- Having an interesting question is great
- Until you know what is y and what is x,and know how they may be measured, you don't have a project

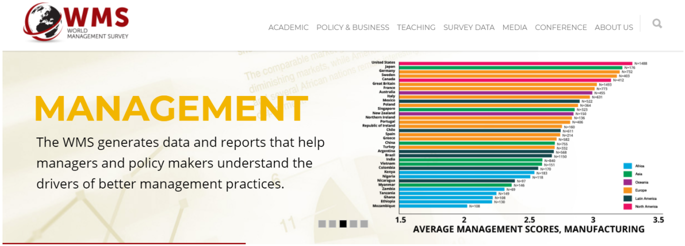


2 Data collection is the foundation for empirical work

- Two ways to think about the research question and data collection
- A: Formulating a question and collecting appropriate data to answer it
- B: Assessing whether the available data can help answer the question.
- Many forms of data collection
  - Administrative data - large, but hard to get access
  - Online data 1: Download/API - great, some cases, not always available
    - Many great source: World Bank, FRED, EBRD, US Census, Kaggle, etc
  - Online data 2: Web scraping - great, cleaning is exhaustive, some coding skills
  - Survey - focused, time consuming, hard to know if will work in advance

2 Case study : Management quality data collection

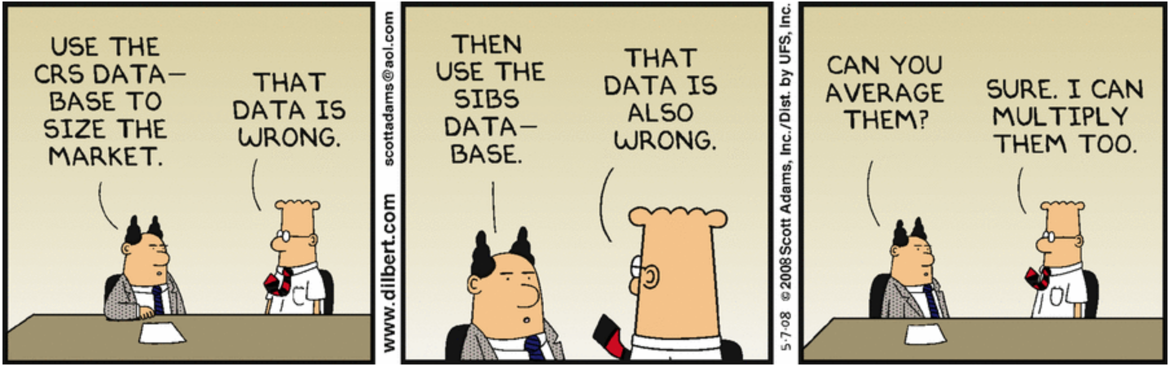
- World Management Survey (WMS) - centralized questions, global www.worldmanagementsurvey.org - Survey on firms and management.
- Scorecard for 18 monitoring, targets and incentives practices such as lean management
- Management quality = score (average)
- Standardized. Piloted. Public.



2 Key point: Unique dataset is massive advantage

- A unique dataset, if decent quality, is a massive advantage
- Web scraping, survey, joining data from various sources

Detour: working with bad data could be an upset



3 Cleaning and organizing the data is a necessary and time-consuming part

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. Filled with decisions.

Understanding and storing

- start from raw data
- understand the structure and content
- understand links between tables
- big data - engineering

Data cleaning

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

3 Case study : Prepping WMS data

- Check errors and weird values
  - Years of schooling, numerical variable, 999 means missing
- Drop, impute when for missing values.
  - Dropped observations when key variables are missing (14%)
- Filter for purpose.
  - we dropped the few firms with less than 50 employees or with more than 5000 employees (3%).

- Some decisions are necessary for analysis
- Some decisions are arbitrary

Detour: 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

3 Key point: Reproducible data wrangling is essential, time-consuming

- About 80% of analytical project time is wrangling, cleaning and managing data
- "Data and code or it did not happen".

4 Exploratory data analysis helps preparation and analysis

- Exploratory Data Analysis (EDA)
- Linked to data preparation
  - Give context to the eventual results
  - Help deciding the details of the analytical method to be applied.
- Creates first core (descriptive) results
- Guides deeper research
- Compare conditional means, distributions.
  - Tables, graphs.

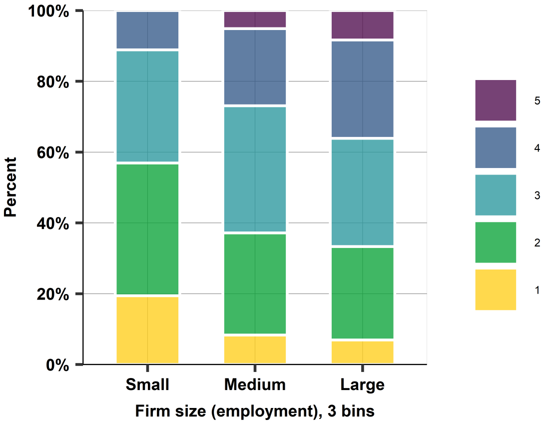
4 Case study: exploratory data analysis

- Pre-study: sample design
  - Understand distributions, understand measure of quality
  - Tabulate subgroups: industry, country
  - Tabulate ownership types - decide what to keep and not
    - Process: maybe go back to causal thinking and cleaning
- Describe patterns,
  - show correlations between management quality and ownership
  - Show correlations with some z variables
    - Process: depending on results: go to analysis, or back causal thinking

4 Case study - Management quality and firm size

- Lean management score 1–5
- Firm size: small, medium, large
- Conditional probability:
  - share of score=1 conditional on being a small firm is about 20%.
  - share of score=5 conditional on being a large firm is about 10%.
- Shows a pattern of association

Note: Source: Management quality is an average score of 18 variables. Firm size is number of emoployees. `wms-management-survey` data. Mexican sample, n=300.



4 Key point: a good descriptive table or a graph is great

- Often a good descriptive table, or a scatterplot with a regression line will be enough to convince readers that there is something going on.
- Even if not, it's informative

5 Analytical work tests and estimates model(s)

- Aim is always to get closer to causality

- Cross section OLS – think hard about causality
- Difference in differences – could a change be driven by something else?
- Panel fixed effects and event studies
  - when intervention varies over time, or happens frequently or continuous
  - often the closest we can come with observational data
- Matching – great was to ensure common support
- Regression discontinuity – nice if you can find one
- Instrumental variables - hardly ever works convincingly.
  - Unless randomization in background

5 Case study: OLS and matching

- Cross sectional data – OLS, matching
  - Propensity score matching on the nearest neighbor: for a group of treated observations finds untreated ones with similar characteristics
  - Here: group by industry, country, firm age, technology type.
    - Algorithm
- Very similar results - matching suggests dropping some types of firms with only family or only public
- Key benefit of matching was to realize there are some type of firms that have no similar counterpart

5 Key point: Getting closer to causality is hard work

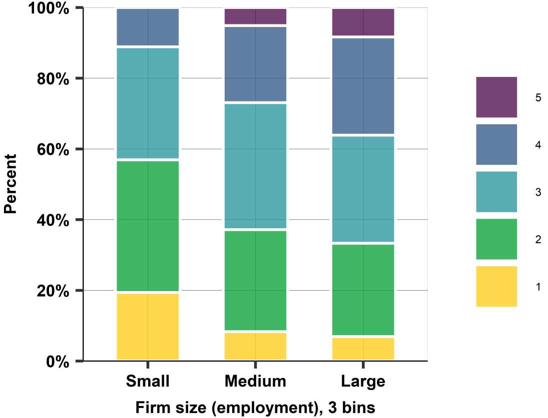
- Sometimes you can find a smart trick like RDD
- Often it's painful discussion of how far you are from a causal interpretation

6 Communicating results in a user friendly way

- Interpretation and effective presentation of the results
- Data visualization summarize findings / convey messages.
- There are rules and help to make good tables and graphs
  - Helping the user understand, tailor to audience
  - Make sure scaffolding is there, too.

6 Case study - Designing a graph

- Craft setup: to shows a pattern of association, create three groups of firm size
- Decide graph type: Stacked bar to show relative frequency
- Pick a color scheme (viridis)
- Add a note with with key info, such subset, N, variable definition



6 Key point: Develop graphical skills

- Creating good graphs may be practiced and done better
- Massively useful skill in real life

7 Answer the original question and discuss generality

- Answer the question
  - Precisely from your favorite model
  - More generally
- Must make a stand and discuss how you take the results. Reliable? Causal?
- Generalizing to the dataset you care about
  - Statistical Inference: SE, CI, p-values in the population
  - External validity: Beyond the dataset and population
- Statistical inference and external validity are both important
  - Sometimes trade-offs. Both important

7 Case study: result and interpretation

- The quality of management is lower, on average, by about 30% of a standard deviation, in founder/family-owned firms than other firms
  - of the same country, industry, size, age, with the same proportion of college-educated workers, and with a similar number of competitors.
- Public ownership is closely linked to management quality, there is likely a causal link
- Many uncontrolled variation - can't be sure.

7 Key point: Show the result and discuss problems

- Talk causality and internal validity: Be honest about the result.
- Talk external validity: what to expect when your model is used outside
- You have a paper if you can summarize findings in a few tweets.

7 Key point: Show the result and discuss problems

- Talk causality and internal validity: Be honest about the result.
- Talk external validity: what to expect when your model is used outside
- You have a paper if you can summarize findings in a few tweets.

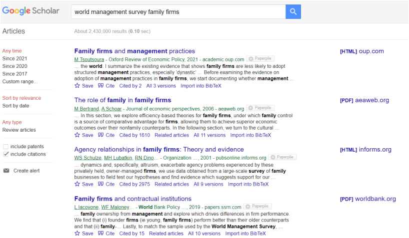
Tools to help the process

- Great deal of technology and tools to help the data analysis process
- Review a few for each of the seven steps

1 Read up on your topic, and manage references

- Reading up on your topic, and research question
- Research - Google Scholar, Repec, great repositories of papers in social sciences

- Several tools to manage bibliography and references, like: Paperpile, or Zotero



2. Doing surveys online

- Collecting data with a survey is oldest data collection method
- Several online platforms to help
- Data collection is hard, because
  - Writing and testing questions hard
  - Low response rate

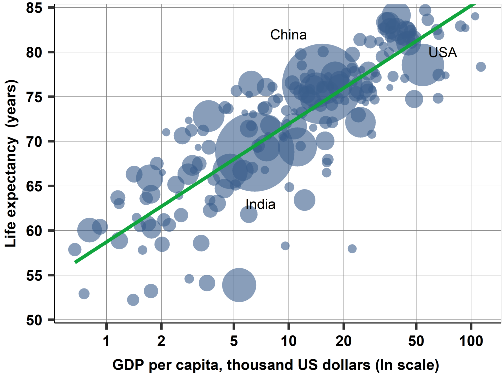
- surveymonkey.com
- docs.google.com/forms

3. Coding environments for data wrangling and analysis

- Coding for data wrangling and analysis - reproducible research
- Stata, R, Python (+Matlab, Gretl, SPSS, SAS, Julia)
  - Stata: academia, NGO, government in rich countries
  - R: academia, government, statistics, consulting, journalism
  - Python: computer science, finance, academia
- Coding environments help a great deal
  - Rstudio is designed for R, but works with many languages
  - Jupyter notebook is designed for Python but works with many languages

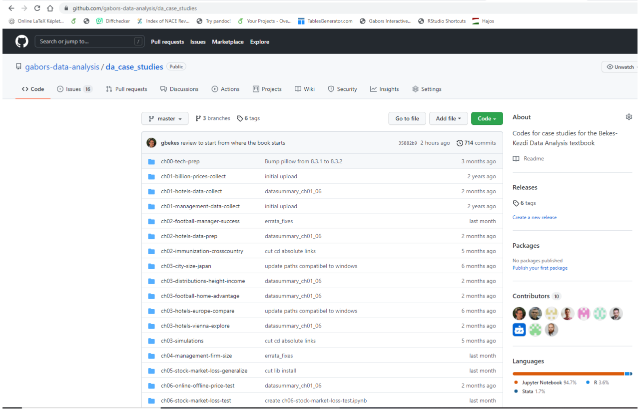
4. Data exploration and visualizaton with GGplot (R) and plotnine (Python)

- Everybody can learn basics of good graph making...
  - Lot of online help
  - r-graph-gallery.com
- R: Ggplot, Python: plotnine (same syntax)
  - versatile, must invest in learning a way of design thinking
- Graph here (ggplot): few lines + reproducible.



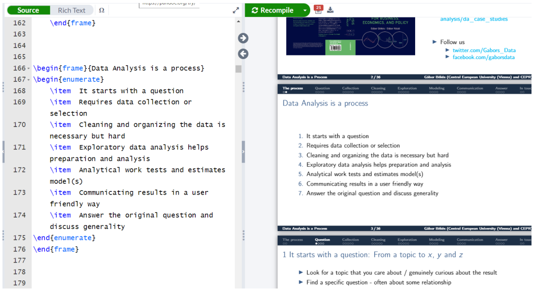
5. Doing reproducible research with Git and Github

- Reproducible research
- Git is version control system
- Github is a cloud based code repository system based on git
- All code for my textbook is hosted on Github: github.com/gabors-data-analysis/da_case_studies



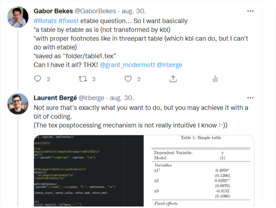
6. Writing up a thesis and presentation

- Tex/latex is a document preparation system (like MS Word).
  - User has full control.
- Overleaf is a cloud solution
  - For easy use of latex and collaboration
- This presentation is written in latex, edited in Overleaf
  - The textbook, too

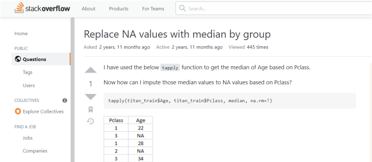


7 Benefit from online communities

- hrefhttps://twitter.comTwitter is the social media platform to learn
  - Rstats, EconTwitter and many more
  - Regular discussion of methods, coding tricks, new packages.



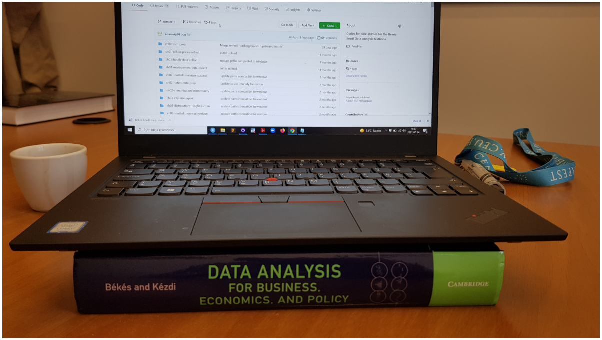
- Stackoverflow is a community of coders
  - Find answers to questions in R, Python and more
  - Pose questions / answer them



Review of tools

- Read up on your topic with Google Scholar, manage references with Paperpile
- Doing surveys online with Google Forms and SurveyMonkey
- Coding environment for reproducible research: R/Rstudio and Python/Jupyter
- Data exploration and visualization with ggplot (R) and Plotnine (Python).
- Doing reproducible research with Git and Github
- Writing up a thesis and presentation in Latex and Overleaf
- You are not alone - benefit from online community with Twitter and Stackoverflow

Thanks and keep in touch



- twitter.com/GaborBekes
- LinkedIn/bekesgabor
- facebook.com/gaborsdata
- twitter.com/Gabors_Data