Statistical Programming and Open Science Methods

Data wrangling and visualization fundamentals

Joachim Gassen
Humboldt-Universität zu Berlin

September 02, 2022

SFB/Transregio 266
ACCOUNTING FOR TRANSPARENCY

Time table October 11

When?	What?
09:00	Writing readable and reusable code
10:30	Coffee
11:00	Debugging tools
12:30	Lunch and coffee
13:30	Relational databases and the concept of normalized data
14:30	Data wrangling and visualization fundamentals
15:30	Assignments and wrap up
16:00	End of event

Disclaimer

A lot of what follows — including but not limited to the figures — are borrowed from Claus O. Wilke (2019): Fundamentals of Data Visualization, https://serialmentor.com/dataviz/index.html

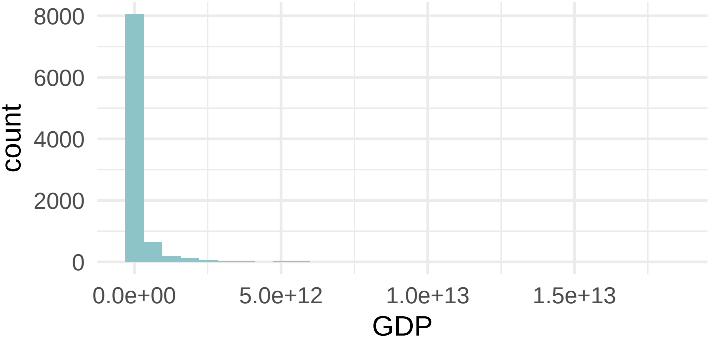
Data wrangling/cleaning

- We already talked a lot about data wrangling in the last session
- Special issues like data scraping will be covered in the 2nd course block in February
- Instead, I would like to focus on one issue that I believe to be paramount in applied econometrics: miss-coded data

Miss-coded data

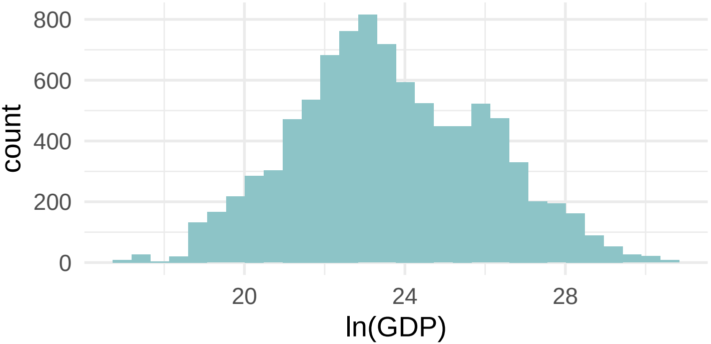
- Data in econometrics only rarely has unambiguous measurement properties
- Money is an opinion
- Magnitudes of measures are heavily affected by the size of the data generating unit (firm, countries, etc.)
- A common remedy is to size-adjust measures by deflating but this sometimes tends to introduce new problems

A macro example that we all know



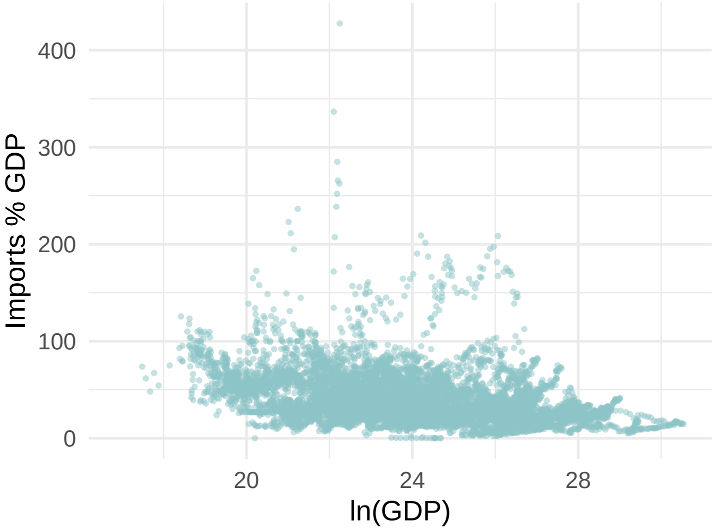
Definition: Gini index (World Bank estimate) Note: Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution. A Lorenz curve plots the cumulative percentages of total income received against the cumulative number of recipients, starting with the poorest individual or household. The Gini index measures the area between the Lorenz curve and a hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line. Thus a Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality. Source: World Bank, Development Research Group. Data are based on primary household survey data obtained from government statistical agencies and World Bank country departments. For more information and

A macro example that we all know



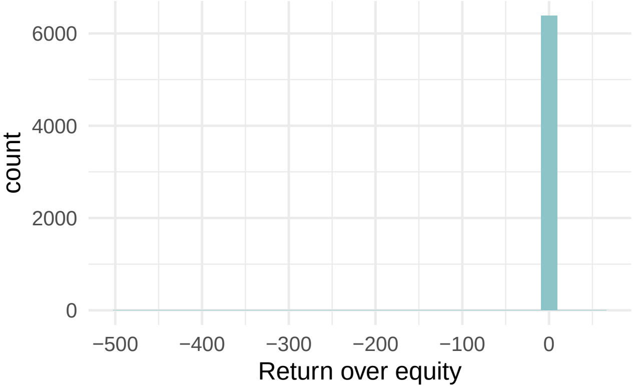
Definition: Gini index (World Bank estimate) Note: Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution. A Lorenz curve plots the cumulative percentages of total income received against the cumulative number of recipients, starting with the poorest individual or household. The Gini index measures the area between the Lorenz curve and a hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line. Thus a Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality. Source: World Bank, Development Research Group. Data are based on primary household survey data obtained from government statistical agencies and World Bank country departments. For more information and

The resulting deflator problem



Flawed variable definitions

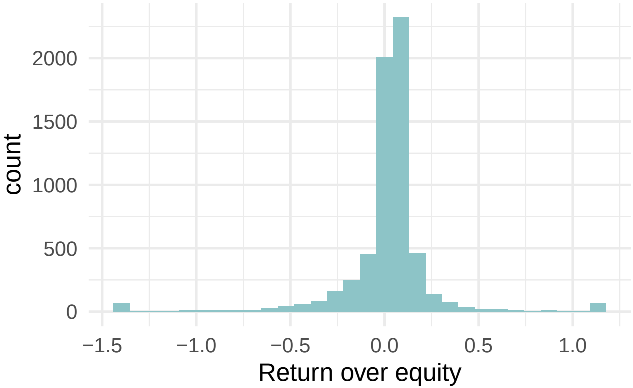
$$ROE_{i,t} = \frac{NI_{i,t}}{0.5(BVE_{i,t-1}+BVE_{i,t})}$$



Definition: Return on equity (net income divided by average equity). Data: Russel 3000 U.S. Index firms, fiscal years 2014-2016, as provided by the ExPanDaR R package, https://joachim-gassen.github.io/ExPanDaR/.

Flawed variable definitions: Winsorization

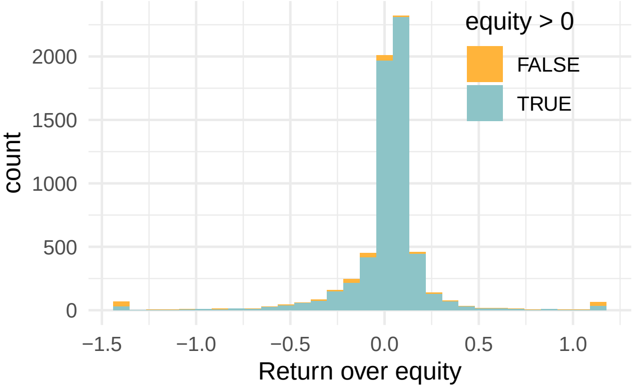
$$ROE_{i,t} = \frac{NI_{i,t}}{0.5(BVE_{i,t-1}+BVE_{i,t})}$$



Definition: Return on equity (net income divided by average equity). Data: Russel 3000 U.S. Index firms, fiscal years 2014-2016, as provided by the ExPanDaR R package, https://joachim-gassen.github.io/ExPanDaR/.

Flawed variable definitions: Unintended side-effects

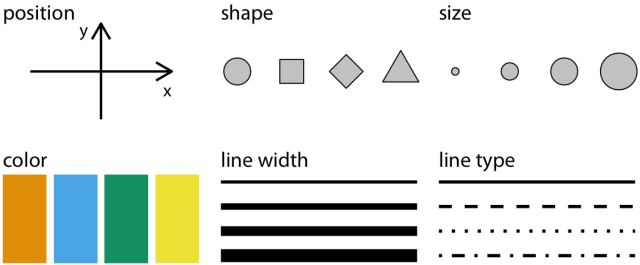
$$ROE_{i,t} = \frac{NI_{i,t}}{0.5(BVE_{i,t-1}+BVE_{i,t})}$$



Definition: Return on equity (net income divided by average equity). Data: Russel 3000 U.S. Index firms, fiscal years 2014-2016, as provided by the ExPanDaR R package, https://joachim-gassen.github.io/ExPanDaR/.

Data Visualization

Data visualization means mapping data on aesthetics



Data visualization for exploring data

- Adapt a structured workflow of analysis
- Visuals do not need to be nice but you should be able to produce them quick
- This is were interactive graphics really shine
- Shamelessly self-advertising example below

```r
library(ExPanDaR)
ExPanD(worldbank, cs_id = "country", ts_id = "year",
       var_def = worldbank_var_def)
```

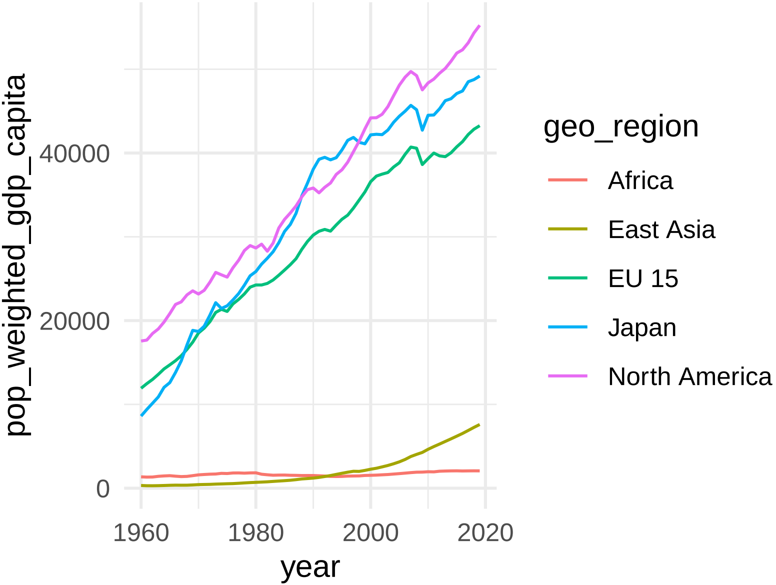
Data visualization for presenting data

- What is your story?
- Who is your target audience (academics in your field?, academics in other fields?, scientific media?, general audience?)
- What are the restrictions of your communication channel (paper?, slide presentation?, web site? social web?)
- Produce the best that you can do, factoring in the opportunities and limitations generated by the points above

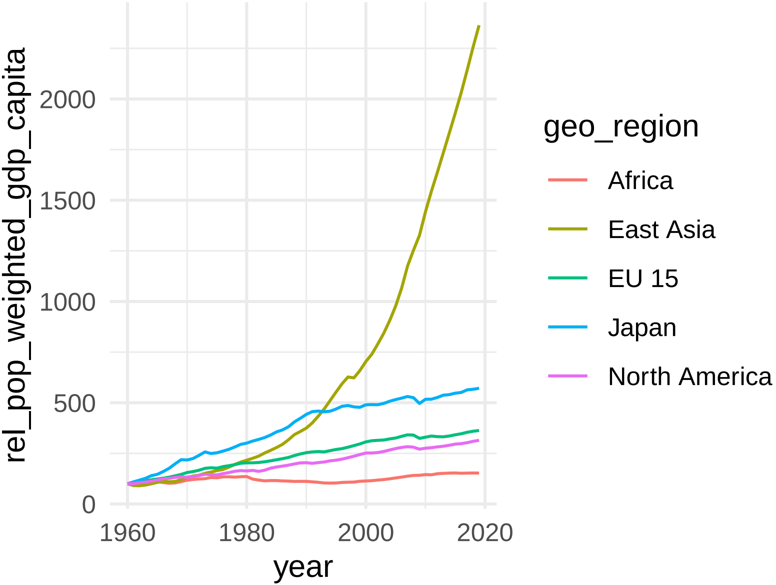
An example for data presentation

- We want to provide a tweet-able visual that communicates the sluggish development of African economies in the last 50 years
- The main point is to create awareness that the income gap between African economies and North-american/European economies is not closing over time while Asian economies have been catching up
- Of course we want our visual to be fully reproducible

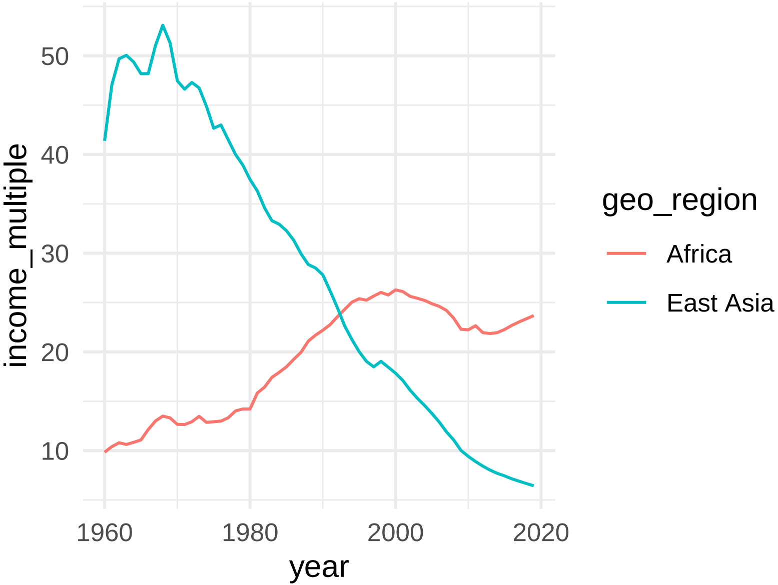
Zooming in on geographic regions



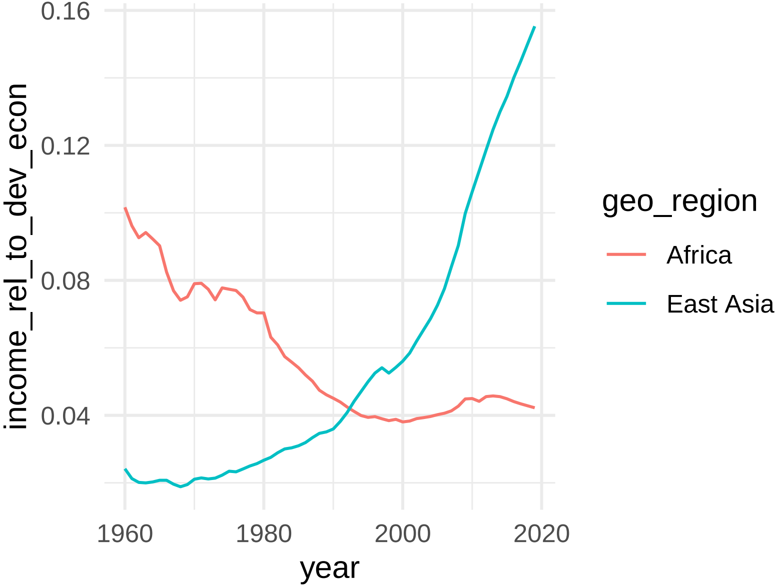
Base adjusting



Reducing information



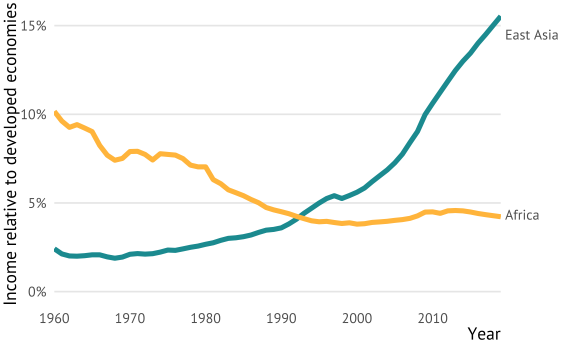
Catering to viewing habits (positive slopes — good)



Nicing up

Africa is not catching up

Average income per capita is deteriorating in Africa over time, relative to developed economies in Europe, Japan and North America but also relative to economies in east Asia.



Income relative to developed economies is the population weighted mean of country-level GDP measures from the respective region relative to countries from developed economies (Canada, EU 15, Japan, U.S.).
Data: World Bank. Code: https://github.com/joachim-gassen/sposm.