Exploratory data analysis

From project setup to data visualisation

Applied Data Science using R

Prof. Dr. Claudius Gräbner-Radkowitsch
Europa-University Flensburg, Department of Pluralist Economics
www.claudius-graebner.com | @ClaudiusGraebner | claudius@claudius-graebner.com

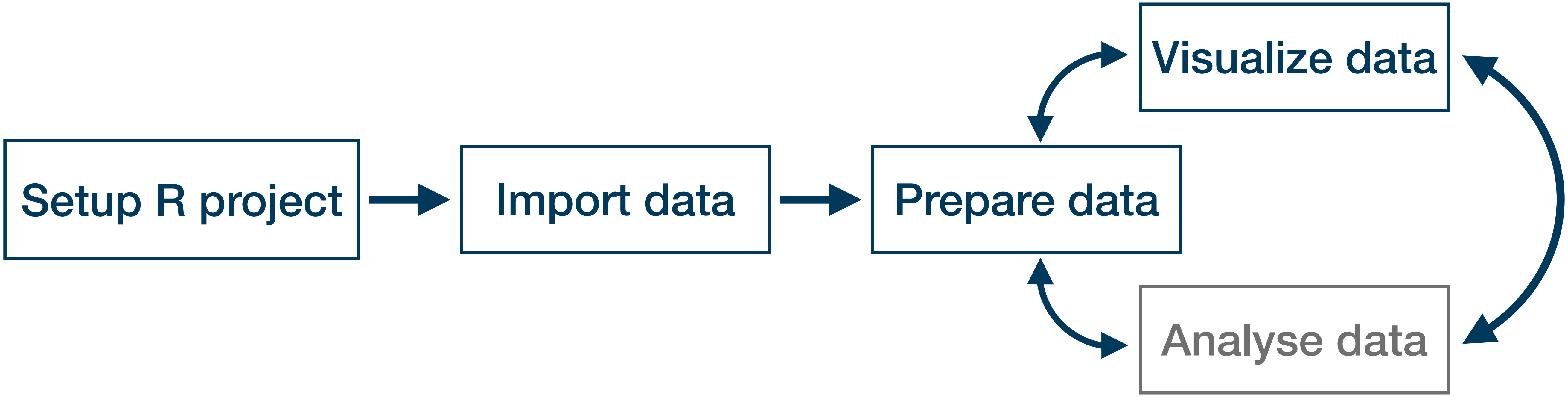
Europa-Universität Flensburg

Europa-Universität Flensburg
International Institute of Management and Economic Education
Department of Pluralist Economics

Goals for today

I. Practice and recap of concepts from the following sessions:

1) Project setup
2) Data import
3) Data preparation
4) Visualisation



General overview

- We are interested the relationship between `inequality`, `female labor force participation` and `child mortality`.
- Goal: get an overview of whether a relationship between the variables exist.

Step 1: Set up an R project

- Create an R project with the standard directory structure!
- Time: 3 minutes

Step 2: gather and import data

- The data we need is available via the following sources:
- World Bank: https://www.worldbank.org/en/home
  - GDP per capita in PPP (constant 2017 international $)
  - Mortality rate, under-5 (per 1,000 live births)
- Standardized World Income Inequality Database: https://fsolt.org/swiid/
  - Gini index of disposable income

Task

- Collect the data and save it in the `raw` directory
- Read the data sets into R
- Time: 20 minutes

Step 2: gather and import data

Hint for the future

- The package `WDI` can be used to collect data from the World Bank.
  1. Search the the data code on the World Bank Data Website
  2. Use `WDI::WDI()` to download the data
  3. Save the data in the raw folder
- See the code example in the exercise solutions

Most larger data repositories have
corresponding R packages to collect data easily.
Check the internet 👩🏼‍💻

Step 3: Prepare data

Task

- Transform the World Bank Data into a tidy format
- Merge the World Bank and Gini data using `dplyr::inner_join()`
- Save the data into the `tidy` directory
- Time: 20 minutes

Step 4: visualise the data

Task

- Compute country averages for all variables for the time period 2010 - 2020
- Create a scatter plot to illustrate the relationship between…
  - Income inequality and child mortality
  - Income inequality and income
  - Income and child mortality
- Save your plot into the `output` directory
- Interpret your results
- Time: 20 minutes

Further exercises

- Use your tidy data from before but consider only data after 1990 and until 2019
- Add a variable indicating whether a country is rich or poor
    - First, remove all countries without a GDP observation from the data
    - To define rich countries, compute the 80% quantile of GDP per capita in 1990; all countries with GDP per capita above this threshold are rich, the rest is poor;
    - Hint: its easiest to create a second data set, compute the classification, and then merge it with the original data
- Add data on CO_2 emissions (`EN.ATM.CO2E.KT`) and population size (`SP.POP.TOTL`) from the World Bank database to your data set; make sure you do not create new missing values during the merge
- Study the relation of CO_2 emissions per capita and GDP per capita using scatter plots
- Compute the share of rich and poor countries, as defined above, of total CO_2 emissions per capita as well as total population over time

Possible solutions are provided via the course website