Introduction to DA1

András Lajos Sári

Data Analysis 1: Exploration

2024

This course

1. This course introduces data collection and data wrangling (management), presentation and understanding of descriptive statistics and basics of visualization.

2. Cover classical statistics methods and their applications, such as data collection and sampling, generalization from the sample to the population and hypothesis testing.

Exploration

- Data Analysis 1 is about exploration
- Figuring out where it comes from, how it's structured, describing and understanding some key patterns
- Exploring data is a process

Exploration

- Data Analysis 1 is about exploration
- Figuring out where it comes from, how it's structured, describing and understanding some key patterns
- Exploring data is a process

START with idea

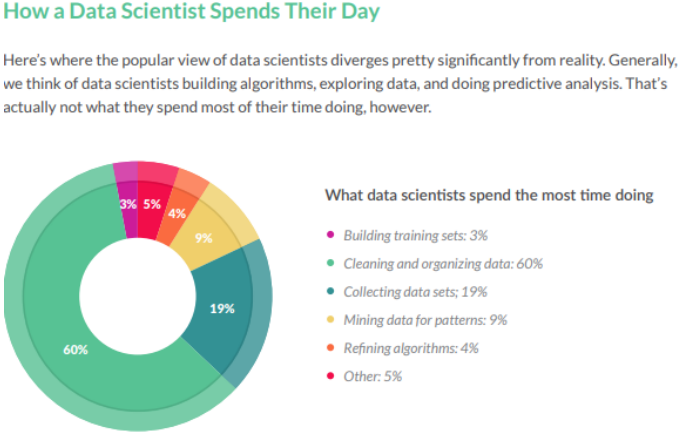
1. writing code –>
2. getting some result –>
3. interpreting that result –>
4. improved/altered idea –>
5. writing code –>
6. getting some result –>
7. interpreting that result –>
8. improved/altered idea –>

...

STOP if happy / run out of time

Data management is key task

- About 80% of data science tasks are composed of managing data, from understanding and altering features of the dataset and variables, to combining various datasets.



Data Analysis 1: Exploration - topics

1. Origins of data (data table, data quality, survey, sampling, ethics)
2. Preparing data for analysis (tidy data, source of variation, variable types, missing data, data cleaning)
3. Exploratory data analysis, Describing variables (probability, distributions, extreme values, summary stats), data visualization
4. Comparison and correlation (conditional probability, conditional distribution, conditional expectation, visual comparisons, correlation)
5. Generalizing from a dataset (repeated samples, confidence interval, standard error estimation via bootstrap and formula, external validity)
6. Testing hypotheses (null and alternative hypotheses, t-test, false positives / false negatives, p-value, testing multiple hypotheses)

Grading

- 20% start-of-the-class quizzes (10% coding - 10% theory)
  - on moodle
  - past lecture material
  - 2% - 4% - 4% for theory
  - be on time!
- 80% closed book exam (40% coding - 40% theory)
  - Textbook chapters 1-6
  - all parts (unless otherwise noted)
  - I will provide a mock exam for the theory part (please remind me, if I forget it!)