

Data Analysis 1

Gábor Békés

Data Analysis 1 – MS Business Analytics: Exploration

2023

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 is a process
- ▶ Using data analysis to answer a research question

START with research question

1. collect some data →
2. look at data, explore →
3. do some analysis →
4. look at data, explore →
5. do some analysis →
6. ...
7. Draw some conclusions →

STOP answering the question

Exploration

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

...

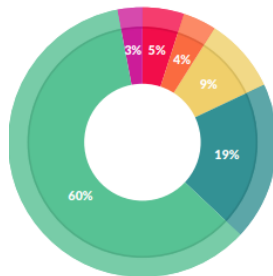
STOP if happy / run out of time

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.

How a Data Scientist Spends Their Day

Here's where the popular view of data scientists diverges pretty significantly from reality. Generally, we think of data scientists building algorithms, exploring data, and doing predictive analysis. That's actually not what they spend most of their time doing, however.



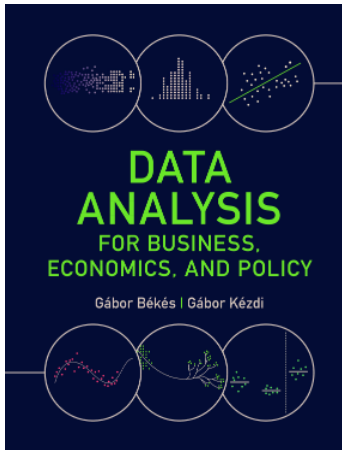
What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

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)

We'll use my textbook...



- ▶ Cambridge University Press, 2021
- ▶ gabors-data-analysis.com/
 - ▶ Download all data and code: gabors-data-analysis.com/data-and-code/
 - ▶ Errata page: check, report if find typos or errors: gabors-data-analysis.com/errata/
- ▶ Follow us
 - ▶ twitter.com/Gabors_Data
 - ▶ facebook.com/gaborsdata

AI and data analysis

- ▶ Throughout this course, will discuss how AI can be a co-pilot
- ▶ AI: LLM / ChatGPT, Bard, etc
- ▶ Co-pilot: You need to be in charge but AI will be used in every step to
 - ▶ Help you offer options, ideas
 - ▶ Help you weigh pros and cons, and make better decisions
 - ▶ Help check your output
 - ▶ Assist in technical steps, sometimes do it for you.
- ▶ Will discuss what is AI great / helpful / bad at.

Grading

Assessment

- ▶ Class participation (3%)
- ▶ Start-of-the-class Quizzes (7%)
 - ▶ Past lecture material
 - ▶ Simple question, close to practice question at the end of handout chapter
 - ▶ 3 quiz: 1+3+3 points (online on Moodle)
- ▶ Closed book exam (90%)
 - ▶ Textbook chapter 1-6
 - ▶ All sections including Under the Hood -unless otherwise noted
 - ▶ 90 minutes, short questions, like end of chapters and the quiz questions
 - ▶ Mock exam will provided