COMP0047 — Plan for 2nd half of term

- Week 6
  - Feb 22 — Session on coursework / tips to write good reports
  - Feb 24 — Case study 3
- Week 7
  - Mar 1 — Intro to case study 4 (correlations & PCA)
  - Mar 3 — Case study 4 (1st part)
- Week 8
  - Mar 8 — Interactive session on coursework (breakout rooms + Q&A)
  - Mar 10 — Case study 4 (2nd part)
- Week 9
  - Mar 15 — Intro to case study 5 (networks)
  - Mar 17 — Case study 5 (1st part)
- Week 10
  - Mar 22 — Interactive session on coursework (breakout rooms + Q&A)
  - Mar 24 — Case study 5 (2nd part)

Tips on how to write good reports

- Recap of assignment and deadline
- Marking criteria
- Writing tips

Coursework - Data Science report

- **Final mark of the module: 100% Coursework**
    - Due April 6 (2nd week after end of term)
    - **Individual report** — You will formulate a data-driven question on a dataset and tackle it with methods of your choice, detailing the results in a report

Data Science report - Learning outcomes

- **Learning outcome #1**: how to download, import and handle a large and complex dataset

- **Learning outcome #2**: how to formulate data-driven research questions

- **Learning outcome #3**: how to leverage data for business purposes

- **Learning outcome #4**: learn how to analyse and manipulate data that are different (and richer) from “typical” financial data (which will be the subject of most case studies)

Data Science report - due April 6

- Identify a **data-driven research question** to investigate with the Yelp dataset and tackle it with suitable methods

- **Examples**
  - What factors (e.g., Wi-Fi, parking, etc.) correlate strongly with high/low ratings?
  - Do certain types of cuisines systematically attract higher/lower ratings?
  - Are certain cuisines more/less successful in certain cities?
  - Do different locations of chain restaurants receive similar ratings regardless of the neighbourhood they’re in?
  - Are ratings influenced by seasons/weather?

Data Science report - due April 6

- Summarize your results in an **individual report** of less than **1500 words**, with up to **4 display items (plots or tables)** with captions of less than **150 words** (excluded from total word count)

- **2 dedicated sessions** with group discussions, breakout rooms and Q&A:
  - March 8 (week 8)
  - March 22 (week 10)

Marking criteria

1. Justification of approaches used
2. Clarity (both in text and plots / tables)
3. Consistency of language and mathematical notation
4. Soundness of results

Equally
important:
25% of final
mark each

Originality in is not a marking criterion! Prioritise coming up with a well posed question over coming up with a “fancy” one

Justification of approaches used

1. Walk the reader through every hypothesis and choice you make (assume the reader is a graduate student with basic knowledge of probability and statistics)

2. Explain why you chose to do something over the alternatives (e.g., why did you choose to consider a certain class of distributions?)

3. Using a package downloaded from the Internet does **NOT** qualify as a methodology: you have to explain what it does and demonstrate that you understand it

Justification of approaches used

1. Walk the reader through every hypothesis and choice you make (assume the reader is a graduate student with basic knowledge of probability and statistics)

2. Explain why you chose to do something over the alternatives (e.g., why did you choose to consider a certain class of distributions?)

3. Using a package downloaded from the Internet does **NOT** qualify as a methodology: you have to explain what it does and demonstrate that you understand it

Clarity (text)

1. Write using clear language and notation, **ask yourself whether you would understand your own report**

2. Avoid colloquial language, use formal language as in a scientific paper

3. Use the word limit to your advantage: write short sentences where each word matters!

Clarity (plots & tables)

1. Use 2-3 significant digits when presenting numerical results (e.g., 1.23 instead of 1.23456789)

2. Show good-quality figures (no cropping!)

3. All plots **must** have: axis labels, legends where needed, clearly discernible symbols and lines

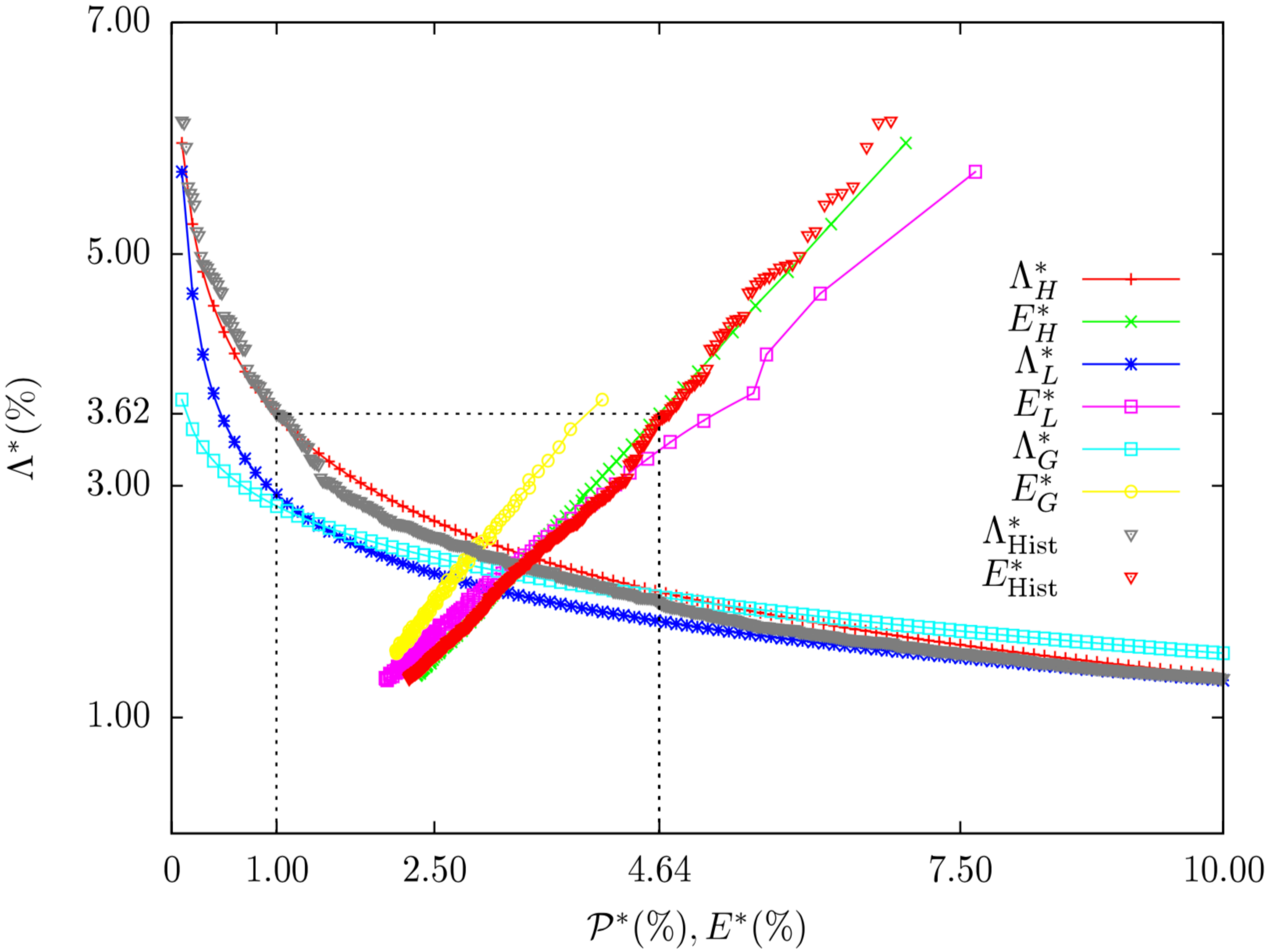
Clarity (plots & tables)

1. Use 2-3 significant digits when presenting numerical results (e.g., 1.23 instead of 1.23456789)

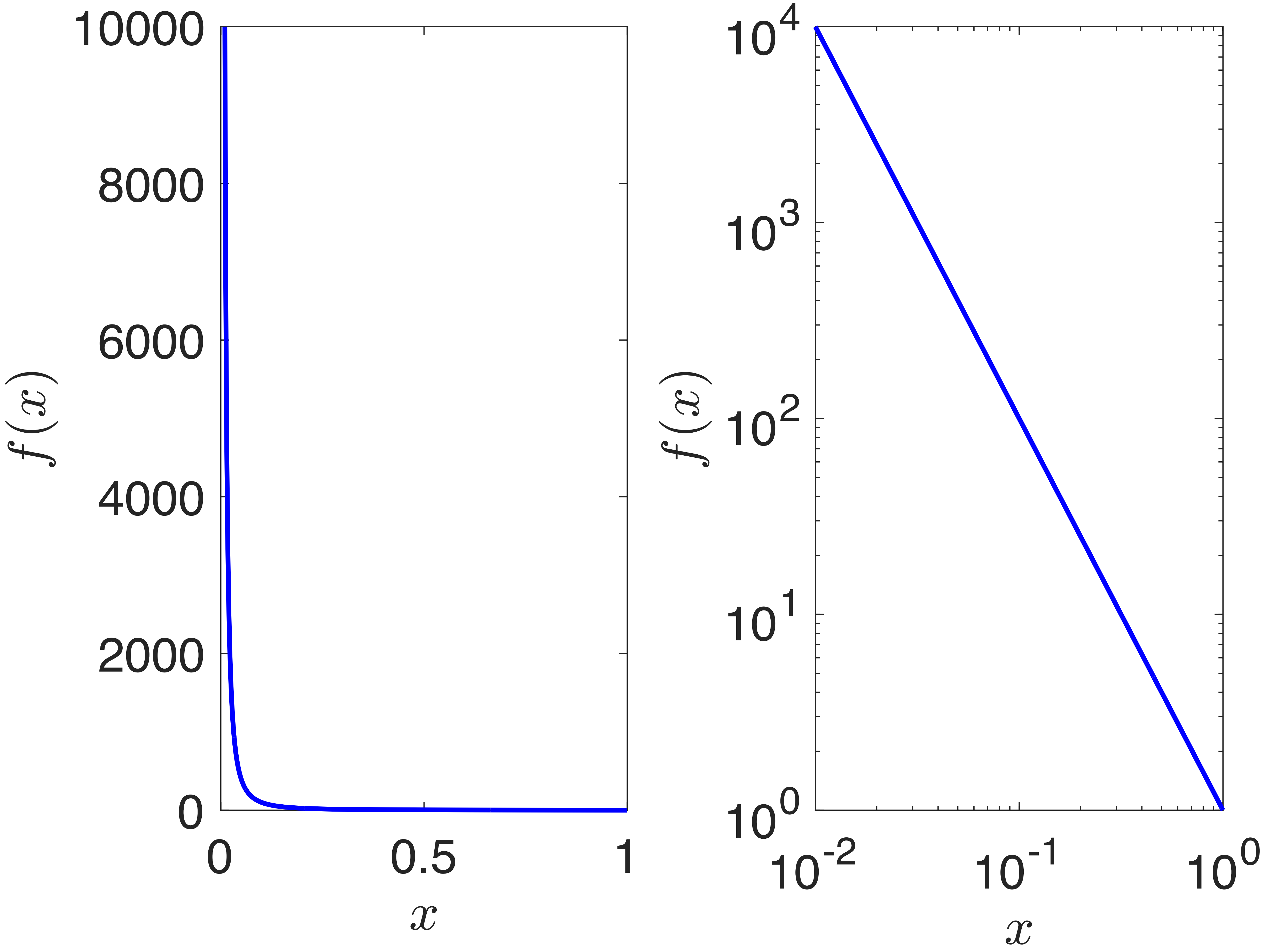
2. Show good-quality figures (no cropping!)

! 3. All plots **must** have: axis labels, legends where needed, clearly discernible symbols and lines !

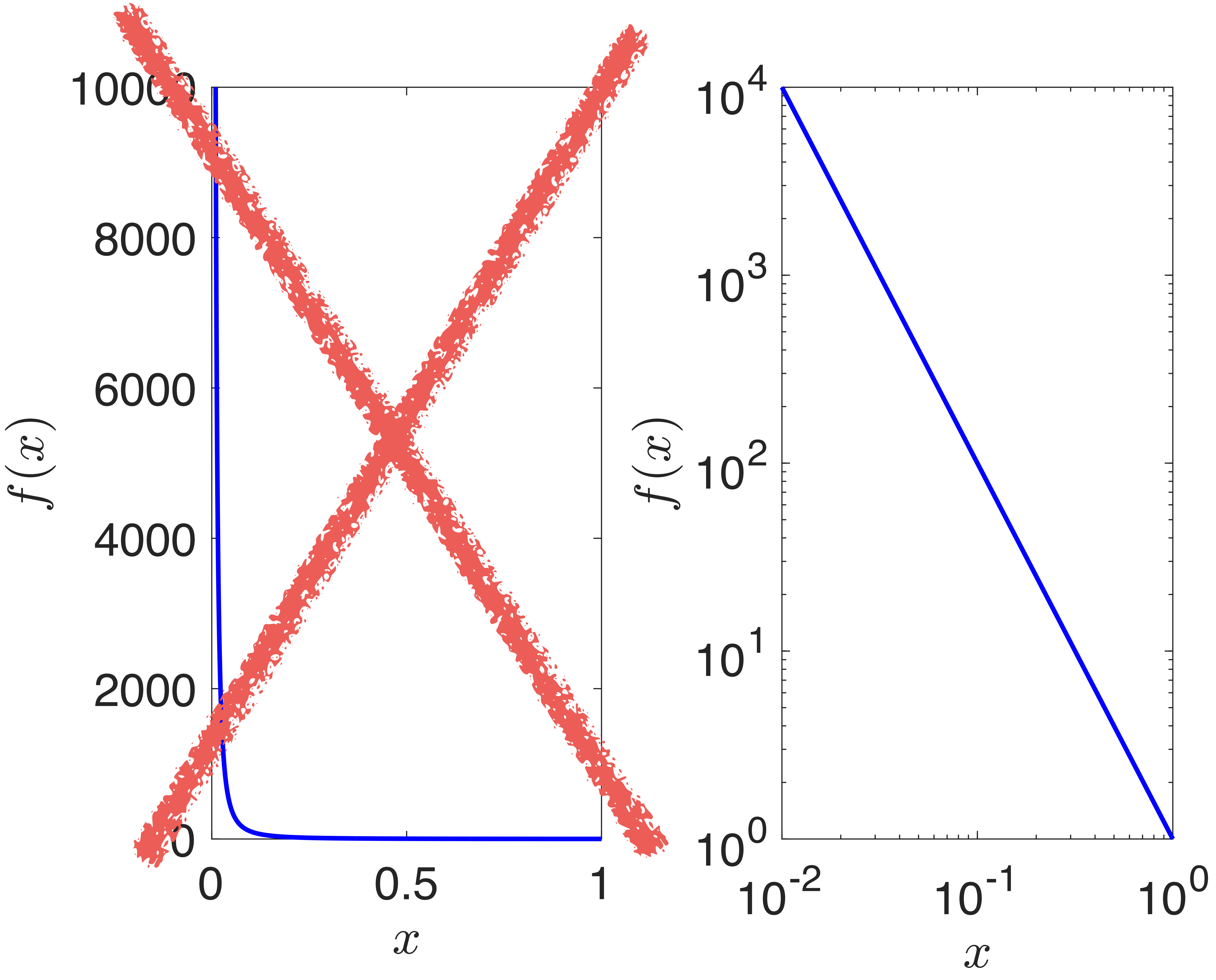
Example of a good plot



Use log scale when necessary!



Use log scale when necessary!



Consistency of language and notation

1. Use the same style throughout the report (e.g., do not switch present / past tense, etc)

2. Use consistent mathematical notation (e.g., do not rename variables, label plot axes accordingly, etc.)

Soundness of results

1. Be rigorous and discuss your results professionally

2. This means discussing statistical significance when appropriate, and discussing what worked and what did not work in your approach in relationship to your initial hypotheses

3. **Negative results are OK!** There's nothing wrong in disproving your own hypotheses. No need to "stretch" your results in order to make them work

Soundness of results

1. Be rigorous and discuss your results professionally

2. This means discussing statistical significance when appropriate, and discussing what worked and what did not work in your approach in relationship to your initial hypotheses

!

3. **Negative results are OK!** There’s nothing wrong in disproving your own hypotheses. No need to “stretch” your results in order to make them work

!

Suggested structure

1. **Introduction** - A brief overview of the data you used (e.g., what cities, what fields, how you cleaned the data, etc.), a presentation of the research questions and of the hypotheses you're making to address them

2. **Methodology** - Explain how you're going to tackle the research questions

3. **Results** - A summary of your main results, without much comment

4. **Discussion & conclusions** - Comment on your results: what did you find? Was it in agreement with your hypotheses? What worked or did not work? How could you improve the analyses?