Group Project: Statistics and Data Science

Description

It is time! You will apply the techniques and knowledge learned during your master in a Data Science project using real data. In a group of 3 with complementary skills and diverse background, you will conduct a statistical analysis, trying to identify a causal effect. However, a “fancy” data analysis will not be sufficient! In today’s world, it is crucial to be able to disseminate good science to fight misinformation. You will thus have to write an article, targeted to the general public, explaining the problem you are answering, the issues you are facing to conduct your analysis, the techniques you are using to solve these issues, the insights you gained, and the limitations behind your study.

Instructions

Suggested steps:

1. Define your “research question”. You should agree with your team to study a given topic.
2. Do some literature search to see what has been done in this area
3. Select the variables you will need for your analysis – and clean your dataset if needed
4. Conduct an Exploratory Data Analysis (EDA)
5. Identify the issues you are facing to identify a causal relation (e.g., endogeneity, sample bias)
6. Design and conduct a causal analysis
7. Discuss your results, limitations, and potential next steps

Deliverable:

- One article per group, written in a notebook, that is ready to publish on Medium.com
- Your notebook should be uploaded to Moodle
- You can structure your article as you wish. However, we do expect the following elements to be discussed at some point in the article:
  1. Context and and research question — why does the research question matters
  2. Discussion on causality and methodology
  3. Data sources and possible biases in the data
  4. Visual presentation of results
  5. Critical discussion of limitations and extensions
- For the things that you believe are too technical to go in the main body (data specifics, regression tables, etc), you can include them in an appendix
- Deadline: December 23, 23:59

Evaluation:

- 25% choice and justification of research question [domain expertise]
- 25% discussion of issues for causality, justification of techniques used [econometrics]
- 25% clarity, quality, and creativity of notebook [communication]
- 25% implementation in Python [programming]

Advice

- Respect each other, and respect each other point of view. No matter the background and skills, everybody has valuable input and should have their say in the project
- Your notebook should include text, images and graphs, lines of codes, and any creative content you might think of to help the dissemination of your study
- In the assignment, you discovered and cleaned some dataset. You are encouraged to reuse these datasets, but it is not mandatory. You can also complement these datasets with other ones if needed
- Use url link to import your data. You can upload your dataset to GitHub, Google drive, or similar to do so
- You can collaborate on notebooks using GitHub or similar. You can also use Google Colab (e.g., sharing your notebook in a Google drive folder) or Deepnote
- In the assignment, each of you picked a research question, selected some variables, and conducted an EDA. You are free to reuse what one or several member(s) of your team did. Discuss in your group what each of you did for the assignment, assess synergies, and find out if you can and want to build on that. You are also free to tackle a completely new issue
- Remember that a data science project is an iterative (and not linear) process. It is completely fine to engage in a way, and then realize you need more data/variables, a different statistical model, etc. The more you progress, the more information you get, which might affect your course of action.
- You will apply in this project what we have seen in this class (programming in Python, statistics, and causality models), but you should also rely on what you have learned in your other courses and previous experiences. A data science project is not only about computer science and statistics, but a key part of it is also domain expertise, i.e., what you know about a given topic
- You are free to organize your teamwork as you wish, but it’s never a bad idea to set up milestones and allocate tasks to each member
- Finally, remember to have fun! Our only goal in the project is to learn!