Coursework: Protecting data

The purpose of this coursework is to apply data de-identification techniques. The goal is to find the right balance between privacy and utility, while producing **one dataset** that would be used for three use cases.

We provide you with a dataset, and you should apply a de-identification strategy. As deliverables, you should provide us with the de-identified dataset and a short report explaining clearly the methods applied, an assessment of both the privacy level of your strategy and the reduction of the utility entailed by your solution. You are also required to submit a notebook with the code you used to de-identify the dataset.

You should design de-identification strategies as seen in class, to prevent threats to the privacy of users contained in the dataset assuming the dataset will be **publicly released**. What theoretical guarantees are you providing if any, what you would qualify as sensitive information in this dataset, how are you choosing your privacy parameters if any, and what algorithm you're using to reach the guarantees should be clearly explained in the report among other relevant aspects.

There are no correct solutions. What we are interested in is your reasoning and your ability to describe in a short report the trade-offs you made and quantitative reassurances you can provide (both in terms of privacy and utility). We also would like to see a description of the remaining privacy risks and how this might put users at risk. Multiple strategies can be used to evaluate the utility of your anonymised dataset, ranging from a machine learning perspective to a raw evaluation of marginal distributions. We however expect you to explain your reasoning when evaluating utility and emphasise the limitations of your approach.

At the end of the report, we would also like you to briefly (~0.5 page) discuss potential changes you would make to your de-identification strategies and risk estimation if the dataset was **to only be shared with trusted researchers**, instead of publicly released.

The dataset (available on Scientia) contains 100 000 records. The attributes of this dataset are: area, postcode, dob (date of birth), gender, ethnic_group, phone_number, marital_status, qualifications, occupation, income, home_ownership, distance_to_work_km (0 means working from home).

The 3 use cases for the dataset are:

- Use case 1: To study if there is any pay inequity between racial groups.
- Use case 2: To analyse how income ownership rates differ by region.
- Use case 3: To study the relationship between distance travelled to work and age.

<u>Submission</u>

- Max. 4 pages and max. 1500 words - no appendices. The report needs to be self-contained (contain all the information one needs to evaluate your solution).
- Don't write any code in your report. Only pseudo-code and only if absolutely necessary.

- The de-identified dataset, in a CSV format.
- A clean Jupyter notebook used to de-identify the dataset.
- We **strongly recommend** working together on this and not “splitting the work”.
- Submit your report via Scientia.