Exercise 2

Research question: Did the new curriculum of the bachelor studies of computer science at the university of innsbruck affect the number of women starting a computer science study?

Data Source:

- (optional) University of Innsbruck Database of immatriculated Students
- LinkedIn: finished bachelor degree in computer science & ongoing studies in bachelor computer science

a) General biases and issues

We most likely have a **population bias** in our dataset: There probably is an underlying gender distribution to the population interested in studying Computer Science.

We would maybe encounter **redundancy** in our dataset, since the same person can have multiple accounts. This could cause the same person to be in the dataset multiple times.

Also **content production biases** are possible because a user made a mistake when putting their field of studies (e.g. they study architecture but set computer science).

Mitigate such problems:

- population: the population bias can be tackled by determining the population distribution of genders during the old curriculum and comparing it to the population distribution during the new curriculum time window
- redundancy: watch for multiple entries of the same person in the dataset and clean them
- content production biases: content production bias can be eliminated with checking their current occupation (if they have any) and check if their job's branch is in IT

b) Data source/origin level

External Bias: There might be people, which set their education falsely to present themself better for certain jobs.

Non-individual agents: We could encounter bot generated or fake accounts.

Mitigate such problems:

- external bias: A similar approach as in the 'content production bias': try to check how long they did not find a new job or search for loose periods in their CV
- non-individual agents: check how thorough the account information is and set a certain threshold (how much data was given in order to consider the account in our survey). Also, if there is a way to track activity (follows, likes, posts), that could be useful too.

c) Data collection/processing

Data collection bias: LinkedIn might not provide us with enough information.

Solution:

- Use more precise data to count student and finished students from a more regulated system, e.g. database from universities

Data cleaning bias: We might run into problems when cleaning our dataset because some outliers are very hard to detect - e.g. multiple accounts per person, fake or bot generated accounts.

Solution:

- Either use a more regulated system to gather the data
- Use algorithms to check different attributes of an account, if it is present in the dataset, (e.g. activity, past education, current occupation etc.) to predict a behavior in order to properly filter not wanted values

d) Analysis

Descriptive Analysis: Because we want to find the correlation among variables of interest, in our example the correlation between number of female students and the change of the curriculum. However, given that the goal of such work is to summarize complex datasets in a manageable way, they may also conceal important details, potentially misleading the reader into the wrong conclusions.

To avoid wrong conclusions maybe a combination of descriptive analysis and **Observational Studies** can lead to a more unbiased outcome. Observational studies goals are to answer the question ‘why’ and to determine the causation of the analysis.

e) Evaluate and interpret

We try to evaluate the outcome with the correlation of the number of female students after the curriculum change with the number of female students before.

Issues: No correlation, which will result in adapting our research question and gathering data.

Mitigate: detect the population bias