1 Deanonymization (5 Points)

Consider the following two tables. Table 1 shows fictional anonymized data from a social network, Table 2 is said to have been taken from a website for film and series reviews. Assume that all of the people in Table 2 are also in Table 1.

name	gender	age	city	favorite movie	favorite series	relationship status
*	female	20-23	Saarlouis	Thor: Love and Thunder	Hawkeye	not specified
*	diverse	24-27	Völklingen	El Camino	Lost	not specified
*	male	20-23	Merzig	Black Panther 2	Breaking Bad	in a relationship
*	male	20-23	Saarbrücken	El Camino	Rick and Morty	not specified
*	female	24-27	Zweibrücken	Spider-Man: No Way Home	The Last of Us	single
*	male	24-27	Zweibrücken	Minions 2	Rick and Morty	single
*	female	20-23	Saarlouis	Avatar 2	Solar Opposites	in a relationship

Table 1: Anonymized data from a social network.

name	movie/series	review date	star rating
Anna	The Last Of Us	09.04.2023	★★
Tim	Rick and Morty	27.11.2019	★★★★★
Lewis	Scream	20.03.2015	★★★
Lisa	Avatar 2	01.04.2023	★★★★★
Lewis	El Camino	01.03.2019	★★
Josh	Black Panther 2	01.01.2023	★★★★★
Sam	Game of Thrones	21.05.2019	★
Anna	Avatar 2	03.02.2023	★★★★★
Tim	The Big Bang Theory	15.03.2017	★★★
Sam	Madagascar	13.11.2021	★★★★
Josh	Rick and Morty	27.06.2020	★★
Sarah	Spider-Man: No Way Home	01.10.2022	★★★★★
Lisa	Spider-Man: No Way Home	06.07.2022	★
Josh	Lost	02.02.2023	★

Table 2: Data from a website for film and series reviews.

(a) What personal information do you get about the following people by linking the two data sources? Explain your approach.

1. Sarah (1 Point)
2. Anna (2 Points)

(b) Assign names to rows 2,3,4 and 6 from Table 1. Justify your choices. (2 Points)

2 Employers and Employees (5 Points)

In the following, we will look at an employer that stores the **final certificates** of all employees. Normally, only the following information is stored, i.e. basic information about the employee as well as their final grade.

[employees] : {[id: int, name: string, birth_year: int]}
[diploma] : {[employee: (employees), final_grade: float, supervisor: string]}

Due to a recent data leak, the employer now has access to various exam corrections and therefore also to the grades of all exams written. Note that for the sake of simplicity, we assume that **id** and **student_id** match. Furthermore, all employees studied at the same university, the university from which the data leak occurred.

[corrections] : {[exam_id: int, correction_date: date, student_id: int, grade: float]}

1. Explain how the employer could use the above information to detect falsified diplomas. Give at least one concrete approach. (1 Point)
2. Now specify an SQL query that provides the information from the previous exercise. Note that students can take exams more than once and only the best grade of an exam counts towards the final grade. (2 Points)
3. Assume that the employer has uncovered cases in which employees have submitted false diplomas. How should the employer react? (2 Points)

Hint: You may assume that all grades are weighted equally.

3 Separated NSA Entity-Relationship Model (5 Points)

In the lecture you have seen how the shopping behaviour of a person, combined with their living situation, can reveal certain information, e.g. whether an additional person might be hiding illegally in a household.

In this exercise, we will investigate whether even a weakened version of the entity-relationship model would have been sufficient to infer similar information. Assume the NSA entity-relationship model given in the lecture is split into two separate models. One model exclusively contains information about the citizens and in which household they live in. The other model contains information about customers and the groceries they buy. However, there is no obvious connection between the two models.

Given both ER models, answer the following questions and justify your answers.

1. What potentially harmful information can you infer if you only have access to the ER model and corresponding data given in Figure 1. (1 Point)
2. What potentially harmful information can you infer if you only have access to the ER model and corresponding data given in Figure 2. (2 Points)
3. What potentially harmful information can you infer if you have access to both ER models and their corresponding data? How exactly would you do that? Note that citizens and customers are different entity types. (2 Points)

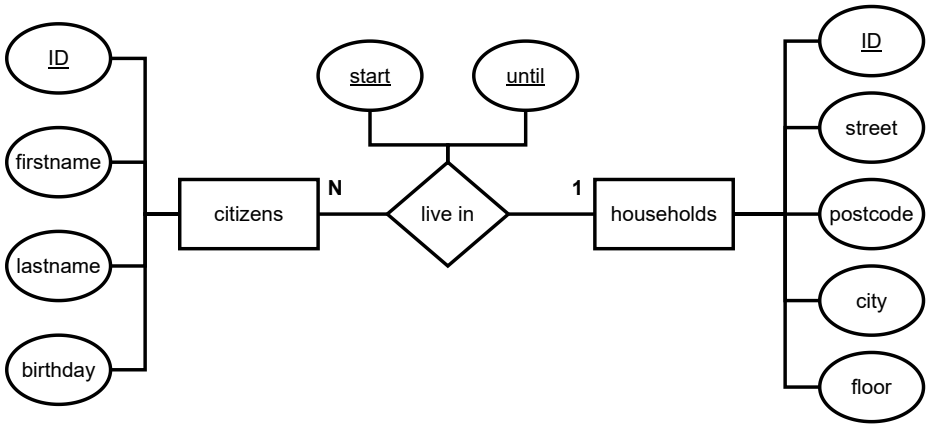

Figure 1: The living situation of citizens.

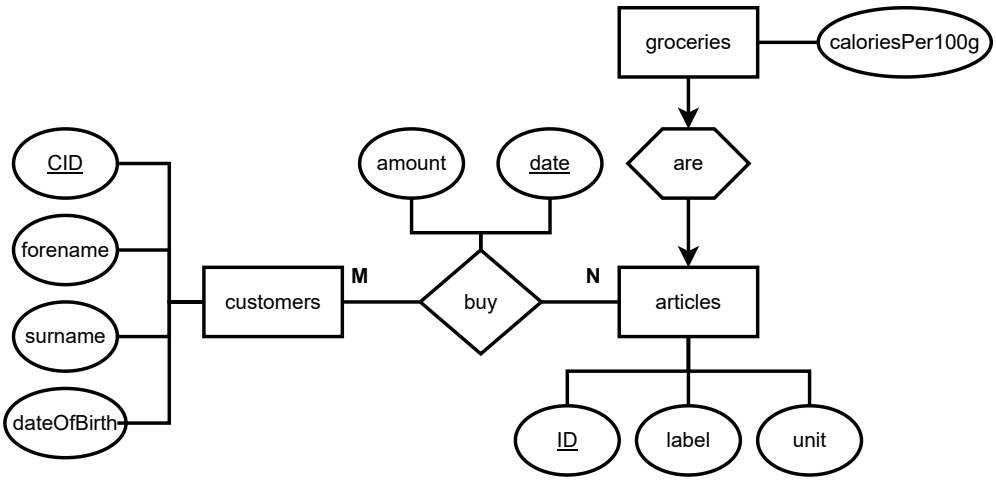

Figure 2: The shopping behaviour of customers.

4 Commissioner Equi-Join's Toughest Case (5 Points)

In this task, you slip into the role of Junior Inspector Group by. Use the enclosed notebook to solve the task. The supplied data set is an extension of the data in the NSA.ipynb notebook.

[households] : {[id: int, street: string, postcode: int, city: string, floor: int]}
[citizens] : {[id: int, firstname: string, lastname: string, birthday: timestamp]}
[live_in] : {[citizen_id:(citizens), start: timestamp, until: timestamp, household_id:(households)]}
[articles] : {[id: int, label: string, unit: string]}
[groceries] : {[id:(articles), caloriesPer100g: int]}
[purchases] : {[article_id:(articles), citizen_id:(citizens), date: timestamp, amount: real]}

Commissioner Equi-Join recently spoke to his superior and mentor, Chief Commissioner Theta-Join, about old cases, which drew Equi-Join's attention to one of his few unsolved cases. This is about a murder of the person John Doe, which took place on November 24th, 1943 at 3 pm. According to the autopsy report, the cause of death was found to be a very rare (and fictional!) poison, which however, according to the forensic pathologist Dr. Selection, can be made from a list of everyday foods. Concretely, the poison can be made with the following ingredients.

- Exactly 500 grams of carrots.
- At least two kilograms of apples.
- At least one kilogram of onions, but a maximum (inclusive) of three kilograms.

Note, that the `amount` in the purchases of these ingredients is given in 'kilogram', e.g. an amount of 0.2 equals 200 grams.

Dr. Selection also said that these foods can be used as poison for a maximum of 5 days (inclusive) after the ingredients are purchased, otherwise the effect would be too weak.

Based on this assumption, Commissioner Equi-Join then launched a review of documented purchases in local supermarkets. Unfortunately, he could not do much with this data. He then questioned witnesses about suspicious activities on the day of the murder, but this was also unsuccessful.

However, Commissioner Equi-Join is now convinced that with your help he can solve the case. To do this, he pulls out the old surveys of the witnesses. Unfortunately, due to the age of the documents, most of the statements have become illegible. All he can find is a page with information about how suspects return to their homes on the day of the murder. The following information about the addresses of the suspects is still legible:

- Address 1: ...13
- Address 2: ...bucht...
- Address 3: Kor...

In this regard, Commissioner Equi-Join gives you the register of residents of the time, which contains information about the local residents and their registered houses, as well as data on the purchases registered at the time.

With this data, can you help Commissioner Equi-Join solve his old case? Submit your solution as a SQL query that has the following output.

- The suspects' first names as 'First_Name'.
- The last names of the suspects as 'Last_Name'

You may use subqueries and views to solve this task. Note that all timestamps follow the format `YYYY-MM-DD HH:MI:SS`. Also explain in the `jupyter.txt` whether you can clearly identify a main suspect based on the data provided.

Submission

Solutions must be submitted in teams of 3 to 4 students by May, 30 2024, 10:00 a.m. via your personal status page in CMS using the Team Groupings functionality. Late submissions will not be graded!

Please note that there are two submissions, under **Theoratical** you submit your solutions to exercises 1, 2 and 3 as a PDF file.

Your solution to exercise 4 must be handed in under **Practical** as txt file.

Make sure that you only copy the complete Jupyter cells you want to add and that indentation and formatting are correct.