LINMA2472 - Algorithms in data science

Project 3: Privacy

Authors : Heroufosse Gauthier
Degives Nicolas
Gandjeto Martine

2021 - 2022

1. Introduction

Nowadays, data are increasingly being collected to inform thinking and guide decision making. Some of these data may represent sensitive information about individuals. Therefore, it is important to respect the privacy of individuals as it is a basic human right. In this context, this 3rd project is an implementation of some methods to guarantee the anonymity of a dataset medical records, such that it can be shared with third-parties (or data users) without violating the privacy of patients. For this purpose, we consider two potential third-parties, namely sociologist and the US department of Health and Human services, which have respectively the following use case:

- Study the impact of stress and high-pressure environments on one's health ;
- Decide where to build new hospitals and which departments to have in these.

The ultimate goal is to reach a satisfying balance between privacy and utility that can be reached for these two use cases. Note that the two produce dataset will be simultaneously available. We need to keep this in mind to avoid link between the two which would decrease the privacy level.

2. Initial dataset

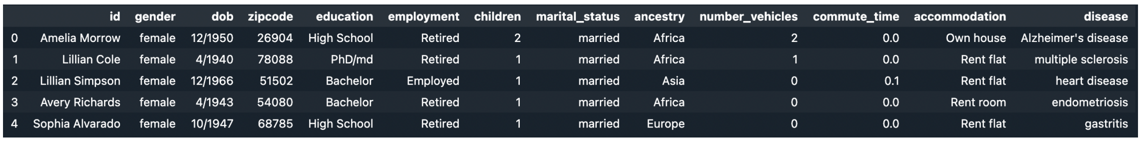
The initial dataset which contains 2000 records and 13 variables (detailed in Table 1) per record is construct as follow:

Table 1: Structure of the dataset

Variable Name	Description of the information
Id	Name of the patient
Gender	Male or female
Dob	Date of birth
Zipcode	Location
Education	Level reached
Employment statuts	Employee, Retired, . . .
Children	Number of children
Marital status	Married or not
Ancestry	The origins
Number vehicles	Number of vehicles for the household
Commute time	Average daily commuting time (in hours)
Accommodations	Type of housing
Disease	Illness of the patient

Table 2 below present a few row if the initial dataset :

Table 2: Five first rows of the initial dataset



	id	gender	dob	zipcode	education	employment	children	marital_status	ancestry	number_vehicles	commute_time	accommodation	disease
0	Amelia Morrow	female	12/1950	26904	High School	Retired	2	married	Africa	2	0.0	Own house	Alzheimer's disease
1	Lillian Cole	female	4/1940	78088	PhD/md	Retired	1	married	Africa	1	0.0	Rent flat	multiple sclerosis
2	Lillian Simpson	female	12/1966	51502	Bachelor	Employed	1	married	Asia	0	0.1	Rent flat	heart disease
3	Avery Richards	female	4/1943	54080	Bachelor	Retired	1	married	Africa	0	0.0	Rent room	endometriosis
4	Sophia Alvarado	female	10/1947	68785	High School	Retired	1	married	Europe	0	0.0	Rent flat	gastritis

As we can see, it contains sensitive information such as the disease which is relevant for the application. Only the name is a direct identifier but the other information represents a large set of quasi-identifiers.

3. Case 1: Health-based sociologists studies

In this first case the data is intended for sociologists to study the impact of stress and high-pressure environments on people's health. It is therefore interesting for them to have factors in people's lives that could be correlated with certain diseases or stress conditions.

In the mean time, we need to protect people's sensitives information. In order to reach a right level of privacy, we had as objectives to set both 2-anonymity and 2-diversity to our Data set (see Near & Abuah, 2021, for details on k-anonymity and k-diversity).

As we want through the data manipulation, we had to make a choice over the way we will proceed to achieve that level of anonymity. Our implementation could be based on the maximization of variables or on the maximization of individuals. This is why we chose to test two different implementations, and compare the two results to choose the best compromise. We will present both of them and then have a small discussion to explain how we chose to keep only one of them.

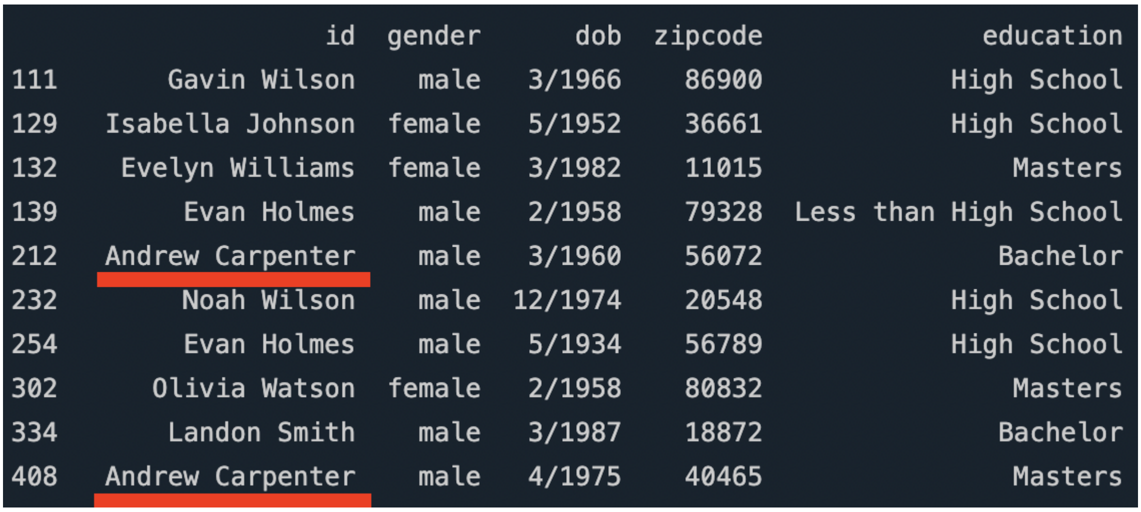
3.1 First Assay: variables maximization

In this first assay, our main objective was to preserve the initial structure of the dataset. But we still made the choice to get rid of both the *ancestry* and the *zipcode* column, because to us it provides no useful information about the health status or the stress of a person.

After those data suppression, our main process was to generalize the data to makes people less recognizable. We placed our variables (columns) into different classes to make the transformation fit the needs.

First, the *id* column was treated as a direct identifier. To pseudonymize this, we apply a cryptographic hash functions (SHA-224) to all the individuals. But as we thought our problem solved, we found that some people had the same name (as you can see in Table 3). So we apply a hash function based on name, but also birth date, zipcode and we added a random salt to make sure no one could make a brute-force attack.

Table 3: Identification of id redundancy



```
                   id  gender      dob  zipcode                education
111       Gavin Wilson    male   3/1966    86900              High School
129   Isabella Johnson  female   5/1952    36661              High School
132    Evelyn Williams  female   3/1982    11015                  Masters
139        Evan Holmes    male   2/1958    79328  Less than High School
212   Andrew Carpenter    male   3/1960    56072                 Bachelor
232        Noah Wilson    male  12/1974    20548              High School
254        Evan Holmes    male   5/1934    56789              High School
302      Olivia Watson  female   2/1958    80832                  Masters
334       Landon Smith    male   3/1987    18872                 Bachelor
408   Andrew Carpenter    male   4/1975    40465                  Masters
```

Secondly, we defined that the sensitive information in the dataset is the *disease*. But as it is needed for the research of scientists, we can not hide or modify in any way this information.

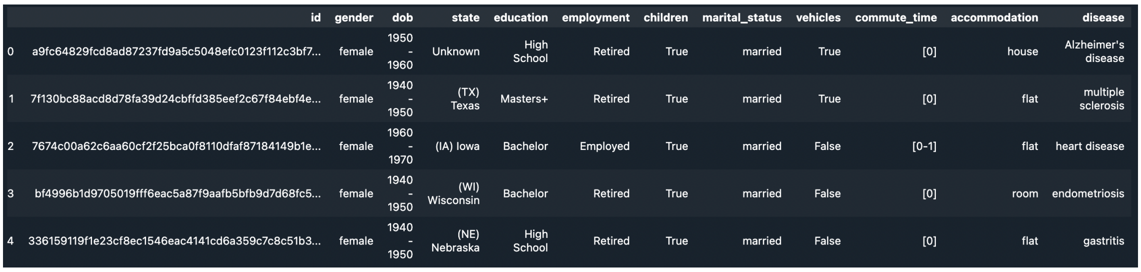
Finally, we defined the last set of variables as "quasi-identifier" columns. This regroup the last columns we haven't talk about yet. As all columns are formatted differently, we needed to apply a different transformation for every variable. We will explain those transformations and present the changes after the processing of the whole dataset.

The *dob* column gives the birth date of the individual. You probably already know that the age of the patient represents a first indication for certain diseases. Through the years, the body becomes tired and less resistant. Moreover, we can imagine that sociologists identify different stresses related to the age of the person. We well understand that this column is really important. However, the precise date is unnecessary. We make as assumption that knowing the yearly tens were people are born is precise enough. Like this, we stay with only 8 classes in our entire data set. That's far less than the 752 initially present.

The next columns we processed were the commute time, *accommodation*, *children*, *education* and *number of vehicles*. To all those columns, we proceeded to a generalization and reduce the number of unique values to 3 or less. In this way, we reduce both *number of vehicles* and *number of children* to True or False. For the *accommodation* column, we choose to keep only the type of housing (room, flat, house). For the *education* column, we create a new value named Masters+ regrouping both PhD/md and Masters. Concerning the *commute time*, we created three categories: [0], [0-1] and [1+]. The last two columns, *employment* and *marital status* were not modified, as we did not find any convenient way to do it.

In the end of all those manipulations, we obtained a data set with 11 columns and still our 2000 individuals. You can see below a sample of that dataset (Table 4).

Table 4: Sample of the final dataset for case 1.1



	id	gender	dob	state	education	employment	children	marital_status	vehicles	commute_time	accommodation	disease
0	a9fc64829fcd8ad87237fd9a5c5048efc0123f112c3bf7...	female	1950 - 1960	Unknown	High School	Retired	True	married	True	[0]	house	Alzheimer's disease
1	7f130bc88acd8d78fa39d24cbffd385eef2c67f84ebf4e...	female	1940 - 1950	(TX) Texas	Masters+	Retired	True	married	True	[0]	flat	multiple sclerosis
2	7674c00a62c6aa60cf2f25bca0f8110dfaf87184149b1e...	female	1960 - 1970	(IA) Iowa	Bachelor	Employed	True	married	False	[0-1]	flat	heart disease
3	bf4996b1d9705019fff6eac5a87f9aafb5bfb9d7d68fc5...	female	1940 - 1950	(WI) Wisconsin	Bachelor	Retired	True	married	False	[0]	room	endometriosis
4	336159119f1e23cf8ec1546eac4141cd6a359c7c8c51b3...	female	1940 - 1950	(NE) Nebraska	High School	Retired	True	married	False	[0]	flat	gastritis

But we were still far from achieve a 2-anonymity and even further from 2-diversity. We still had to remove unique individuals talking to quasi-identifier and then remove those who shares the same disease within an equivalence class. Hopefully, we implemented a tricky function on our dataframe that identify equivalence classes with unique sensitive information. Then, we just removed corresponding indexes and that's it !

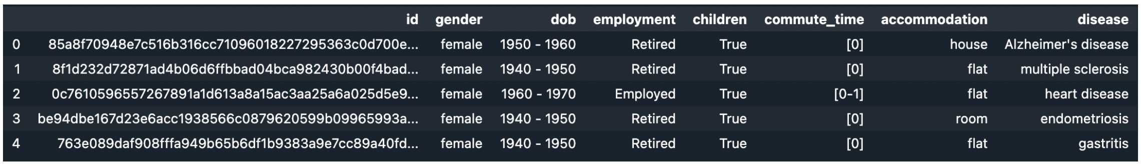
Finally, after this final data manipulation, we end up with 1318 rows × 11 columns. That means a loss of **34.1 %** in the rows, and only two columns. We achieved both 2-anonymity and 2-diversity, and we measure a t-closeness of **0.695**.

3.2 Second Assay: individuals maximization

In this second assay, we will make a slightly different assumption: we will maximize the number of individuals (rows) rather than the number of variables (columns). We will consider that these five columns: *ancestry*, *zipcode*, *education*, *marital status* and *number of vehicles* are not going to be useful. We can now work on the eight remaining columns.

The same transformation processes as presented above were applied. We obtain a data set with 8 columns and still our 2000 individuals. You can see below a sample of that data set (Table 4).

Table 5: Sample of the final data set for case 1.2



	id	gender	dob	employment	children	commute_time	accommodation	disease
0	85a8f70948e7c516b316cc71096018227295363c0d700e...	female	1950 - 1960	Retired	True	[0]	house	Alzheimer's disease
1	8f1d232d72871ad4b06d6ffbbad04bca982430b00f4bad...	female	1940 - 1950	Retired	True	[0]	flat	multiple sclerosis
2	0c7610596557267891a1d613a8a15ac3aa25a6a025d5e9...	female	1960 - 1970	Employed	True	[0-1]	flat	heart disease
3	be94dbe167d23e6acc1938566c0879620599b09965993a...	female	1940 - 1950	Retired	True	[0]	room	endometriosis
4	763e089daf908fffa949b65b6df1b9383a9e7cc89a40fd...	female	1940 - 1950	Retired	True	[0]	flat	gastritis

As before, we applied the function to identify equivalence classes with unique sensitive information. The results were as different as expected: we end up with 1914 rows × 8 columns. That means a loss of 5 columns but only **4.3 %** of the rows. We achieved both 2-anonymity and 2-diversity, and we measure a t-closeness of **0.597**.

3.3 Discussion about Case 1

Now that we have seen results for both assumptions, we will try to justify which simulation we prefer and why. We choose to go for the second assay with individuals maximization. Although the column number is kept, the loss of 1/3 of the rows is too important. We know that in scientific researches, the number of observations needs to be as high as possible. The links between variables we kept and diseases will be easier to evaluate with the second assay than with the first one. Moreover, the t-closeness is also smaller in the second case. That means we are less vulnerable to skewness attack as the data is more symmetrical than is the first assay.

To conclude, we will quickly review the manipulations we have performed to combat the most common types of attacks we may encounter on data sets:

- **Brute-force attack:** commonly used on direct identifier, iterate through the whole possibilities of the hash function to found the pattern. We prevented this by using id, zipcode and birth date all in one and adding a random salt with unknown size of course ;
- **Uniqueness attack:** used if multiple quasi-identifier known, regroup information from different data set to successfully identify individuals. We prevented this by removing the zipcode variable from this first data set, and using a random salt for the id's.
- **Homogeneity attacks:** used when individuals in the same equivalence class all have the same sensitive attribute value. To prevent this, we simply make sure that a 2-diversity is respected in all equivalence classes.
- **Skewness attacks:** used when the distribution of the sensitive attributes in a class is skewed. To prevent this, we make sure to reduce the t-closeness of our data set.

4. Case 2: US department of Health and Human services

For this second case, the data will be used by the US department of Health and Human services chairman to decide where to build new hospitals and which departments to have in these. Therefore, we imagine that the scientist will need the locations where there could be more patient and diseases that could be more represented in this area. As we did for the first case, we're trying to reach a right level of privacy, the objective is to set both at least 2-anonymity and 2-diversity to our Data set while maintaining the utility of the data.

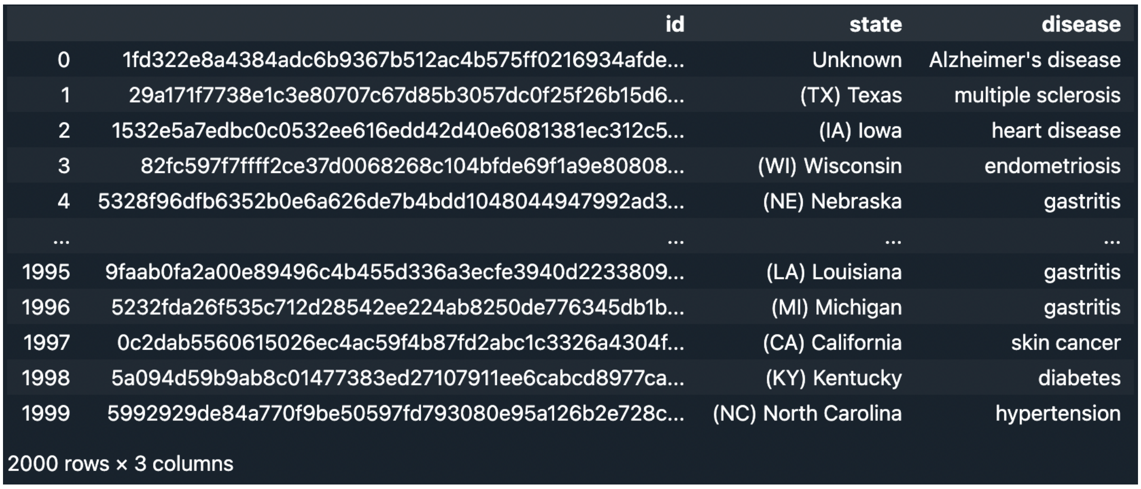
Again, the first step is the pseudonymization. As we did in the first case, we apply a cryptographic hash function with a random salt to avoid possible comparison between this new one and the previous one. This could allow to link the different quasi-identifiers and ruin the effort to anonymize the dataset.

Then, there are actually only few important information in this second case study. The localisation is only given by the zipcodes and the departments needed are suggested by the disease. All the other columns being useless to the case studied, we can simply get rid of them in order to reduce the number of quasi-identifiers.

The only quasi-identifier column left is the zipcode column. As we believe that statewide localization is sufficiently accurate, we choose to convert the zipcode to the corresponding US state. So we download a .csv file with zipcodes and the related US states on simplemaps, and we choose to use a 3-digits separation. Unfortunately, as the data set is not a real one, we found that some zipcode did not correspond to any US state so we set an Unknown value for those.

We end up with a dataset that has only three columns id, state, disease of which there is a sample below (Table 6):

Table 6: Sample of the final data set for case 2



	id	state	disease
0	1fd322e8a4384adc6b9367b512ac4b575ff0216934afde...	Unknown	Alzheimer's disease
1	29a171f7738e1c3e80707c67d85b3057dc0f25f26b15d6...	(TX) Texas	multiple sclerosis
2	1532e5a7edbc0c0532ee616edd42d40e6081381ec312c5...	(IA) Iowa	heart disease
3	82fc597f7ffff2ce37d0068268c104bfde69f1a9e80808...	(WI) Wisconsin	endometriosis
4	5328f96dfb6352b0e6a626de7b4bdd1048044947992ad3...	(NE) Nebraska	gastritis
...	...	...	...
1995	9faab0fa2a00e89496c4b455d336a3ecfe3940d2233809...	(LA) Louisiana	gastritis
1996	5232fda26f535c712d28542ee224ab8250de776345db1b...	(MI) Michigan	gastritis
1997	0c2dab5560615026ec4ac59f4b87fd2abc1c3326a4304f...	(CA) California	skin cancer
1998	5a094d59b9ab8c01477383ed27107911ee6cabcd8977ca...	(KY) Kentucky	diabetes
1999	5992929de84a770f9be50597fd793080e95a126b2e728c...	(NC) North Carolina	hypertension

2000 rows × 3 columns

After these few manipulation, we obtain a data set of 3 columns and 2000 records. That means that there is no loss in the rows thanks to the many removed columns removed. Moreover, it allowed to reach 6-anonymity and 4-diversity. We also measure a t-closeness of **0.303**.

5. Conclusion

Since we are unable to predict precisely the information required for the two studies, we deleted/generalized information that we did not consider useful in the case under consideration. However, in our opinion, the first case required more different data so that more columns were generalized and not deleted. The privacy of the first case is therefore more limited but the data is much more preserved.

As it was precised in the introduction, the two data set are simultaneously available. That's why it is necessary to check that the two files are not linked, which would allow to reconstruct a data set with more quasi-identifier in it. We have avoided this by using a random salt for the pseudonymization of the two cases, the id's are therefore really different between the 2 data sets. Moreover, the columns present in the two data sets (case 1.2 and case 2) are different except for the disease column which is the same in both. It remains however laborious to join these two data sets in one without knowing the salts and hash function used.

To go further, we identify a few points that could be improved:

- 2-anonymity and 2-diversity are maybe a bit too small considering the amount of freely available data on internet. It could be still quite easy to find people back using additional data ;
- Using different transformations for our class and compare the loss and the t-closeness we get for each simulation ;
- Work with additional indicator such as entropy to measure the loss of data.

Bibliography

Near, J. P., & Abuah, C. (2021). *Programming Differential Privacy*, vol. 1. URL `https://uvm-plaid.github.io/programming-dp/`

Other useful sources

- K-Anonymity (Github)
- K-Anonymity (Wikipedia)
- L-diversity (Wikipedia)
- Anonymizing Data with Pandas
- A Brief Overview of K-Anonymity using Clustering in Python