Privacy Preserving Outlier Detection Exercise 1

Jonas Schulze

November 13, 2023

1 Dataset and personally identifiable information

I have chosen the CrossFit dataset already provided in the Moodle forum. The dataset consists of 28 attributes. In the following I will discuss which classify as personally identifiable information (PII). First off we have the **athlete_id** which works as a primary key for the table. The **name** of the athlete is an obvious explicit PII as it allows a one-to-one mapping to the corresponding athlete. Next, the **region, team, affiliate, gender, age, height and weight** serve as quasi-identifiers as in combination they can uniquely identify individuals.

2 Pseudonymisation

As the name of the athlete is an explicit personally identifiable information we would like to pseudonymize these. I use *anonymizedf* to substitute the name with a fake name which will replace the original entries. To reproduce these and any of the following modifications, the athletes dataset needs to be located in the already existing data folder. Within said folder the python file, after completion, will also have stored the final modified dataset and the lookup tables of the upcoming randomization process.

3 Randomization

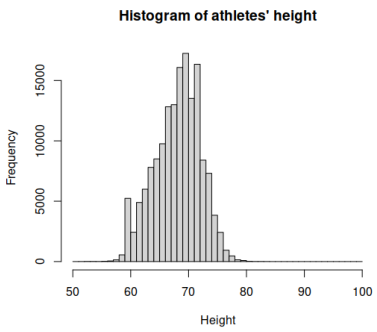
For randomizing I choose friendlywords which can generate human readable randomized strings. I use it to replace the team, region and affiliate. The package also allows themed strings e.g. teams which I will fittingly use to substitute the original CrossFit teams. The substituted words have been written to a csv file functioning as a look up table per attribute to match the original value with the altered, generated one. Examples for such have been uploaded in the data folder.

4 Aggregation

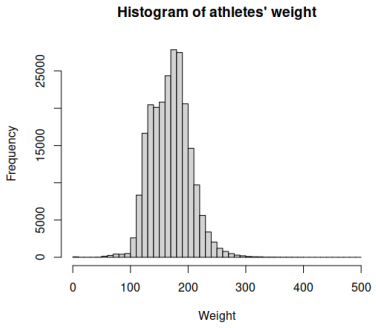
I use the pandas *cut* method to aid me in aggregating the age of the athletes. I choose a bin size of 10 years. Hence, athletes aged 21 and 29 would all together land in the (20-29] bin, hiding the exact age and helping in complicating using age as a quasi identifier. Only for ages 90 and above I put them all into the same bin, as the amount of athletes that old are slim and the oldest person in the dataset is apparently 125 years old (data analysis as such were done with the attached R script).

5 Perturbation

For perturbing I take both the height and the weight of the athletes as they are a prime candidate due to their Gaussian distribution. Thus I add Gaussian noise to both distributions given their respective standard deviations. After filtering out inhumane height and weight entries (e.g. the max height in the dataset is 8388607), we are left with a standard deviation of 34.22 for the weight and 3.99 for the height of the athletes which have been discovered using the enclosed R script. Furthermore our athletes have a mean weight of 170.67 and mean weight of 68.57. A visualization of the original distributions can be seen in figure 1.

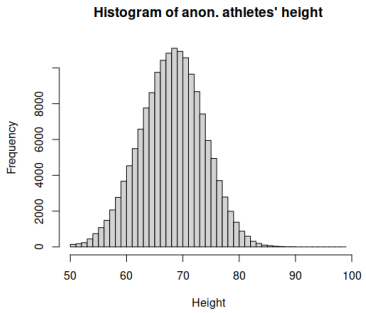


(a) Athletes' height

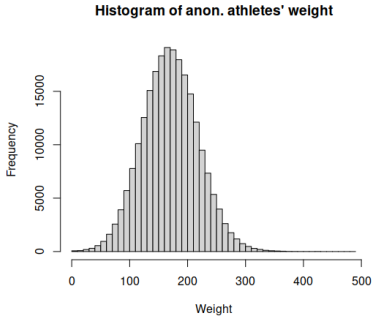


(b) Athletes' weight

Figure 1: Original distributions



(a) Athletes' height



(b) Athletes' weight

Figure 2: Altered distributions

I am now adding Gaussian noise with mean 0 and the respective standard deviations to the two distributions. After perturbation, our means are nearly identical with 170.72 and 68.57 respectively and as desired. Our standard deviation has increased, as expected after adding Gaussian noise, to 48.24 for the weight and 5.63 for the height. The resulting distributions can be examined in figure 2.

6 Data Analysis

6.1 Numerical Data

I use the information loss formula (IL1s information loss measure) provided on slide 11 in the deck of lecture 4 to calculate the information loss of the numerical data I have altered, namely the weight and height of the athletes. The computation can be found in the provided R script. We end up with an information loss of 0.69. The IL1 is rather small and thus the values are somewhat close to the original and retain utility. In combination with the more generalized distribution as seen in figure 2 we see that we have a trade off solely a small amount of information lost for a more obscured distribution of individuals and less disclosure risks. The noise could still be increased if the reduction of risk of re-identification is more important than the utility. However as we are working with a dataset covering athletic accomplishments of CrossFit performers, the sensibility of the data is not too excessive and the performed perturbations seem fitting and reasonable.

6.2 Categorical Data

For categorical data I will look at the change in unique values for the altered attributes. Common measurement are also the number of records changed and number of missing values. Though these are straightforward as we have changed every record of the chosen attributes and furthermore not

forcefully removed values where there used to be entries. I have intentionally chosen a smaller pool of randomized values to be swapped in ($\frac{2}{3}$ of the original number of unique values) to achieve a certain level of overlap to ensure there is no clear one-to-one mapping of original and the substituted entries. Thus, names, regions, affiliates and teams only have roughly two thirds amount of unique values as in the original dataset. Furthermore the ages have reduced from 53 distinct values to 7 bins as described above. With this we tend more to a generalized dataset but only to a degree at which we have gained a fair amount of anonymity. We only exchange little information and should be still able to obtain valuable insights in the dataset.