Security and Privacy BSSEPRI1KU
Autumn Semester 2025

Final Project

Group meow

IT UNIVERSITY OF COPENHAGEN

IT University of Copenhagen
November 8th, 2025

1 Introduction

The purpose of this project is to analyse and anonymise an election survey dataset to protect individual privacy while preserving the data's analytical utility.

2 Anonymisation

Most anonymisations were performed using string based generalisation and category grouping, where detailed values were transformed into broader or partially hidden representations to reduce reidentification risk while preserving analytical meaning.

2.1 Anonymisation methods

2.1.1 Age

To anonymise the age attribute, the date of birth values were first converted into a numerical value for age. Then, we decided to add noise of ±5 years to each individual's age to reduce the risk of reidentification of specific people. This noise was controlled by a random seed so that the results remain reproducible. Adding noise before grouping prevents boundary cases and rare combinations from revealing true ages, giving an extra layer of protection beyond simple categorisation.

Finally, the noisy ages were grouped into broader age intervals (<30, 30–49, 50–64, and 65+). This process keeps useful information for analysis while preventing anyone from accurately tracing back to a person.

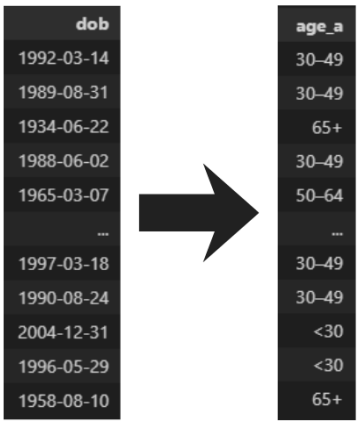

Figure 1: Anonymised age

2.1.2 Education

Education levels were simplified into three broad categories ("Lower", "Higher", "Other") to reduce unecessary details of the data and reduce the chance of reidentifying specific individuals. More specifically, anything below university level education was classified

as "Lower", and anything above that as "Higher". Cases such as "Not stated" were classified as "Other".

The original dataset included many specific education types that made individuals more distinguishable, especially when combined with other demographic attributes. By grouping them into broader categories, we protect privacy while keeping the data meaningful for analysing overall relationships between education level and voting behavior.

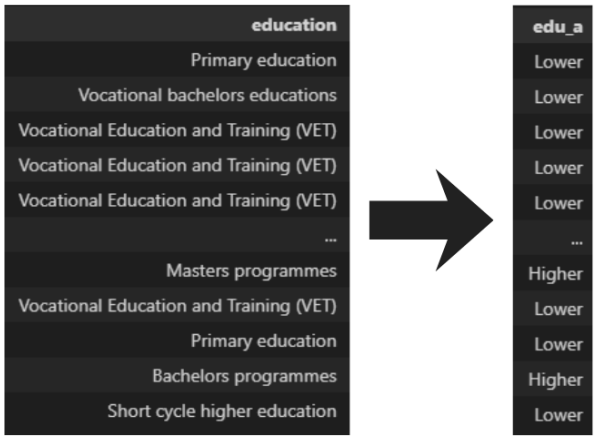

Figure 2: Anonymised education

2.1.3 Citizenship

Citizenship was simplified into two categories - "EU" and "non EU" to prevent reidentification through rare nationalities while keeping the information relevant for regional analysis. The original data included many specific citizenship values, and some of them occurred only a few times, which increases the risk that someone could be indirectly recognized when combined with other demographic variables.

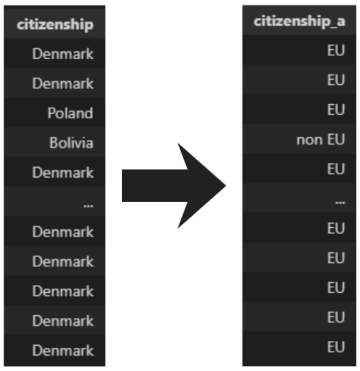

Figure 3: Anonymised citizenship

2.1.4 Marital status

Marital status was simplified into two categories, "Married" and "Single", to remove unnecessary detail and protect individuals with less common status such as "Divorced" or "Widowed". This approach reduces the risk of identifying people based on unique combinations of personal attributes while still preserving useful information about relationship patterns relevant for analysis. Although alternative grouping methods were considered, this binary classification provided the clearest balance between privacy protection and analytical usefulness in our opinion.

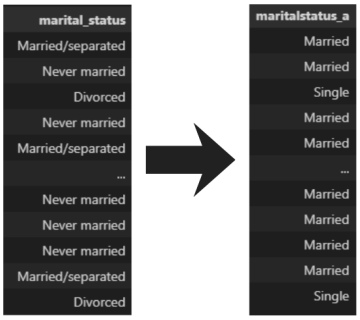

Figure 4: Anonymised marital status

2.1.5 ZIP code

ZIP codes were partially anonymised by keeping only the first two digits and replacing the rest with "xx", which maintains a regional reference while hiding the exact locations. To further protect privacy, as we had the most sample uniques due to ZIP code, codes belonging to individuals who were unique in their demographic combination were replaced entirely with a star (*) using a method called suppression. This two step process prevents potential reidentification of people living in small or sparsely populated areas while still preserving enough geographic information to support meaningful analysis.

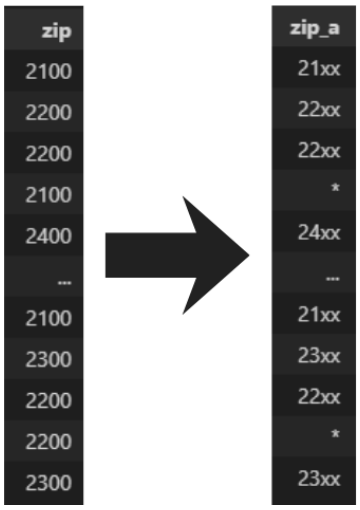

Figure 5: Anonymised ZIP codes

2.2 Anonymisation results

2.3 Disclosure risk

Disclosure risk was calculated using the individual risk measure approach. For each record, the number of individuals sharing the same quasi-identifiers ($x.f$) was counted, and the individual risk was estimated as $1/x.f$. The average and maximum risks were then derived across all records to evaluate dataset safety. This method directly reflects the reidentification probability for each equivalence class and aligns with the goal of minimising both average and maximum individual risk.

Table 1: Disclosure risk metrics for anonymised dataset

Metric	**Value**	**Interpretation**
Average risk ($\bar{r}$)	0.34	Average individual has 34% re-identification chance
Proportion of uniques	0.085	8.5% of records are unique
Proportion with $r(x) > 0.2$	0.515	51.5% of records belong to small, higher-risk groups

Average risk It indicates that on average, each individual has a 34% chance of being reidentified within their equivalence class. While this value is not extremely high, it suggests moderate exposure. Lowering this value further would require quite heavy generalisation or suppression of the key quasi identifiers (as we've already done quite a lot), which would significantly distort the dataset and make the statistical analysis less useful.

Proportion of uniques It means that 8.5% of records are completely unique based on their quasi-identifiers. These cases are the most critical for disclosure risk because a unique combination can directly reveal an individual's identity.

k-anonymity Sadly, we were not able to achieve k-anonymity > 1.The use of controlled random noise (in age) reduces the precision of quasi identifiers, protecting unique data records. This is to make up for the fact we could not achieve k-anonymity > 1 without distorting the dataset further, which would have significantly reduced its analytical value.

Proportion with $r(x) > 2$ It shows that around half of the dataset belongs to relatively small equivalence classes, meaning many individuals share only a few quasi-identifier combinations. This highlights that while broad anonymity was achieved, further generalisation could still improve protection.

3 Trade-off

3.1 Polling stations vs. e-votes

In this section we were investigating possibility of election votes manipulation in favor of red or green political party.
To address this possibility, we set a hypothesis:

H_0: There is **no significant difference** between the political preferences expressed in the survey and the election results.

H_1: There is **a significant difference** between the political preferences expressed in the survey and the election results.

3.1.1 Vote distribution overview

The table compares the distribution of votes between the survey and the official election results, divided by voting method (E-votes and Offline votes).

The Green party received the largest share in both survey and election data, followed by the Red party, while invalid votes remained minimal across all categories.

3.1.2 Shares

Since both datasets are not-balanced, we calculated shares to understand better insights and differences.

The table below shows the relative shares of votes for each party across both voting methods.

Table 2: Survey and Election Results by Voting Method and Party (Vote counts)

Party	**Red**	**Green**	**Invalid vote**	**Total**
Survey results				
E-votes	22	41	0	63
Offline votes	48	86	3	137
Total	**70**	**127**	**3**	**200**
Election results				
E-votes	130	206	1	337
Offline votes	278	406	18	702
Total	**408**	**612**	**19**	**1039**

Compared to the survey, the election results show a slight increase in support for the Red party and a corresponding decrease for the Green party in both E-votes and Offline votes. The share of invalid votes remains low.

Table 3: Survey and Election Results by Voting Method and Party (Shares and Change in Percentage Points)

Voting method	**Red (%)**	**Green (%)**	**Invalid vote (%)**
Survey results			
E-votes	34.9	65.1	0.0
Offline votes	35.0	62.8	2.2
Election results			
E-votes	38.6	61.1	0.3
Offline votes	39.6	57.8	2.6
Δ Change (pp) [Election − Survey]			
E-votes	+3.7	−4.0	+0.3
Offline votes	+4.6	−5.0	+0.4

3.1.3 Statistical tests

In order to see if these changes are statistically significant, we decided to perform a chi-squared test per each channel - online and offline and both together

Results In all cases, the p-value was greater than 0.05. This indicates that the differences in vote shares between the survey and the official results are not statistically significant, meaning that the observed variations in party support and vote counts. Whether for offline, online, or combined voting following the same distribution.

Table 4: Chi-squared Test Results for E-votes, Offline Votes, and Combined Data

Voting method	χ^2	**p-value**
E-votes	0.508	0.77578
Offline votes	1.156	0.56111
Combined (E-votes + Offline votes)	1.488	0.47510

Therefore, we **fail to reject the null hypothesis**, suggesting that statistically the elections were not manipulated in favor of either the Red or the Green party. Furthermore, the proportion of invalid votes also follows the same distribution.

3.2 Demographics vs. party

After anonymisation, p-values in most tests (apart from citizenship) increased slightly, meaning the relationships between demographics and voting behavior became weaker. This suggests that the anonymisation successfully reduced the identifiability of the individuals and protected sensitive correlations. This is because there is now less predictability, which makes it harder to reidentify someone based on their traits.

At the same time, the main statistical patterns (associations between education, age, and party preference) were still observable, which means the data retained a reasonable level of analytical utility.

Overall, the trade-off seems acceptable. Disclosure risk was lowered without completely eliminating meaningful analytical results. Stronger anonymisation could further reduce risk, but it would likely remove the key relationships necessary for social science interpretation.

p-values	Party	
Demographics	Non-anonymised	Anonymised
Gender	0,1312	0,1312
Education	0,0000	0,0001
Age	0,0000	0,0009
Citizenship	0,9821	0,4592
Marital status	0,0001	0,0016
ZIP	0,1527	0,3556

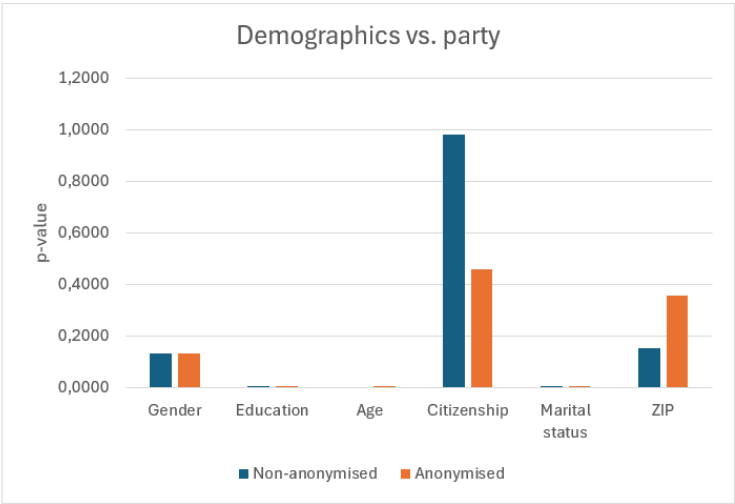

Figure 6: p-values for demographics vs. party

3.3 Demographics vs. e-vote

However, after anonymisation, associations between demographics and e-vote mostly became stronger. This likely happened because merging small categories and removing rare cases decreased noise and made patterns clearer.

Even though the associations became stronger, the overall analytical utility of the dataset remains good. The anonymised data stil allows for meaningful interpretation of how certain demographic factors may relate to e-voting, without losing detail.

All in all, the balance between privacy and utility is still reasonable, but a bit more grouping of variables could help reduce the risk of unwanted identification at a trade-off for utility.

p-values	Evote	
Demographics	Non-anonymised	Anonymised
Gender	0,3206	0,3206
Education	0,3807	0,3045
Age	0,6007	0,0283
Citizenship	0,6774	0,5872
Marital status	0,3093	0,2161
ZIP	0,3551	0,0818

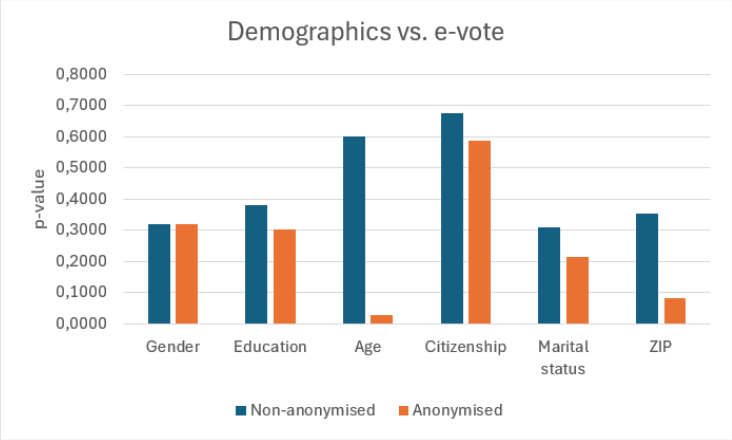

Figure 7: p-values for demographics vs. e-vote

4 Conclusions

This project showed the challenges of balancing data privacy and analytical utility. Through application of generalisation, suppression and controlled noise, the dataset achieved a reasonable level of anonymisation while still allowing meaningful statistical analysis.

Overall, we believe that the trade-off between privacy and utility we achieved is acceptable. The anonymisation successfully lowered the risk of disclosure without destroying the structure of the data or interpretability of the results.

5 Appendix