1 Introduction

Our goal in this coursework was to balance the utility of the de-identified dataset based on the three use cases while also ensuring the users' privacy. We managed to achieve 6-anonymity in the dataset without losing too many records. Specifically, we deleted around 10% of the original dataset and kept the most relevant columns for all the use cases. In the dataset, columns such as on_benefits, income, and credit_score are identified as sensitive information due to their direct impact on individuals' privacy and financial security. Other columns like area, age, gender, marital_status, qualifications, occupation, num_children, home_ownership are considered quasi-identifiers, as they do not directly identify a person; however, multiple could be combined to uniquely identify a person.

2 Methodology

2.1 Initial Steps

To begin de-identifying the dataset, we initially removed the 'name' column as it links records directly to individuals and is not essential for any of the intended purposes. Following that, we excluded the 'dob' column since the dataset includes an 'age' column, which is more general yet still serves the necessary purpose for the third use case. Additionally, we eliminated the 'postcode' column because of its specificity, noting that the 'area' column provides comparable information.

2.2 Utilising histograms and bar charts

Following our initial de-identification steps, we analysed the numerical columns using histograms, along with bar charts for categorical columns, to discern the empirical distributions of these columns. This analysis guided our decision-making process to protect individual privacy while preserving data utility.

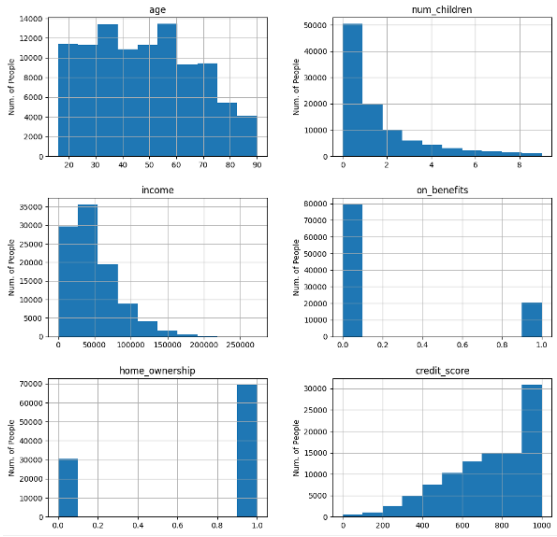


Figure 1: Histograms of numerical columns

2.2.1 Grouping the age column

- **Reasoning:** As depicted in Figure 1, upon reviewing the 'age' column, a notable disparity emerged for individuals aged between 80-90 and 70-80 within the dataset. Around 4000 records pertained to these age brackets, significantly fewer than the average count of 9000 observed across other age ranges. Consequently, we decided to segment the data into six distinct age groups (17-27, 28-38, 39-49, 50-60, 61-72, 73-90). This approach safeguards against revealing specific ages while establishing meaningful divisions, effectively combating the challenge posed by the scarcity of individuals beyond 73 years old.

- **Effect on Utility:** Age is utilised in the third use case, which investigates the fairness and bias of the credit score algorithm. Through this approach, we were able to minimise the impact on utility by not deleting any records from the data.

2.2.2 Grouping the number of children

- **Reasoning:** We adjusted the 'num_children' column due to its right-skewed distribution. We devised four categories ('0', '1', '2-3', '4-9'). Given the dataset's skewness, we aimed to retain the predominant groups ('0' and '1'), while reducing the uniqueness of individuals with more than 2 children.
- **Effect on Utility:** Regarding the impact on utility, this categorisation aimed to maintain the majority representations ('0' and '1') intact while providing a more balanced and less unique representation for individuals with larger families. By creating broader categories, we aimed to reduce the identification risk of individuals with larger family sizes. This restructuring maintains the dataset's utility by providing meaningful groupings without compromising significant variations in family size representations.

2.2.3 Grouping 'income'

- **Reasoning:** To protect individual identities, the sensitive data 'Income' was divided into £5000 segments, a choice influenced by its significance in the second use case. Our investigation unveiled a 7.7% gender pay gap, where men typically earn more than women on average [1]. Choosing larger income ranges might have obscured these differences, especially given the dataset's average income of £47,656. By selecting £5000 intervals, our goal was to ensure that the 7.7% gender gap, on average, would remain discernible within these segments.
- **Effect on Utility:** This approach helps avoid merging unfairly compensated genders into the same income group, allowing for clearer identification of wage inequalities and decreasing individual uniqueness.

2.2.4 Grouping 'credit_score'

- **Reasoning:** Upon examining the histogram, it became evident that the 'credit_score' data exhibited a left-skewed pattern, indicating that most individuals possessed credit scores falling between 900 and 999. Our analysis revealed that the 'credit_score' column aligned with the Experian credit score scale. Consequently, we segmented the column into the Experian credit score ranges (0-560, 561-720, 721-880, 881-960, 961-999) to enhance individual privacy within the dataset.
- **Effect on Utility:** The alignment with the Experian scale ensures continued relevance and compatibility with industry standards, thereby maintaining the dataset's utility for any credit-related investigations or assessments [2].

2.3 Removing sparsely populated areas

- **Reasoning:** Upon examining the distributions of 'area', we found that 'Isle of Scilly' and 'City of London' had the fewest people and could be re-identified using the 'area' attribute.
- **Effect on Utility:** Since both of these areas only held 15 records combined, removing them would have little impact on the utility of the dataset.

2.4 Merging 'Other' and 'Apprenticeship' together

- **Reasoning:** Considering that 'Other' and 'Apprenticeship' categories diverge from conventional education levels (School, High-School, University), we merged them into a single category within the 'qualifications' column. This consolidation aimed to balance the distribution, especially since these categories had fewer records compared to others.
- **Impact on Utility:** Instead of deleting records, we combined these attributes due to their deviation from the standard education levels utilised in the first use-case analysis.

2.5 Clustering Areas together

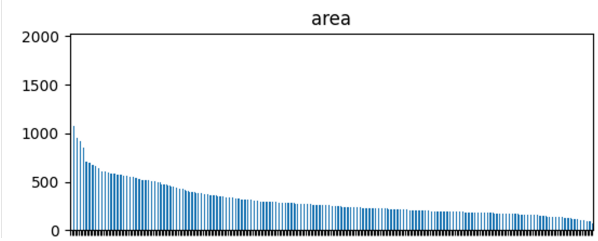


Figure 2: The distribution of areas

- **Reasoning:** Due to the 'area' column having 319 distinct categories, each with limited records, we implemented the K-means algorithm to cluster these areas. Prioritizing the first use case, we grouped areas sharing similar qualification level proportions into clusters. For example, if area X showed a distribution of 20% records at level 1, 20% at level 2, 30% at level 3, 10% at level 4, and 20% at 'other' and 'apprenticeship', we aimed to find analogous distributions in other areas and group them. Our exploration involved testing multiple cluster counts, evaluating k-anonymity, and intra-cluster variability. Ultimately, settling on 10 area groups with a maximum cluster difference of 0.18 and an average intra-cluster difference of 0.03.
- **Effect on Utility:** We preserve utility for the first use-case due to the theoretical guarantees that the K-means algorithm provides (similar areas clustered together). Additionally, the low average intra-cluster difference signifies minimal variation in qualifications among the areas within each cluster.

2.6 Removing 'marital_status' and combining school levels

- **Reasoning:** Upon assessing the impact of each column on achieving k-anonymity, 'marital_status' emerged as the primary influencer, demanding the deletion of the maximum number of records to achieve k-anonymity. Intended for use in Case 3, we conducted chi-squared analyses for each 'marital_status' category for all the features against credit scores. The consistent associations across all different marital_status categories suggested that marital status when combined with other features, does not influence the credit scores. Further research, including insights from Experian, corroborated this, leading to our decision to remove the 'marital_status' column [3]. Furthermore, we combined 'Level 1' and 'Level 2' qualifications because both were categorized as 'School' and had limited records, as noted from the histograms.
- **Effect on Utility:** The chi-squared test and insights from Experian validated that marital status does not significantly affect credit scores.

2.7 Clustering Occupations together

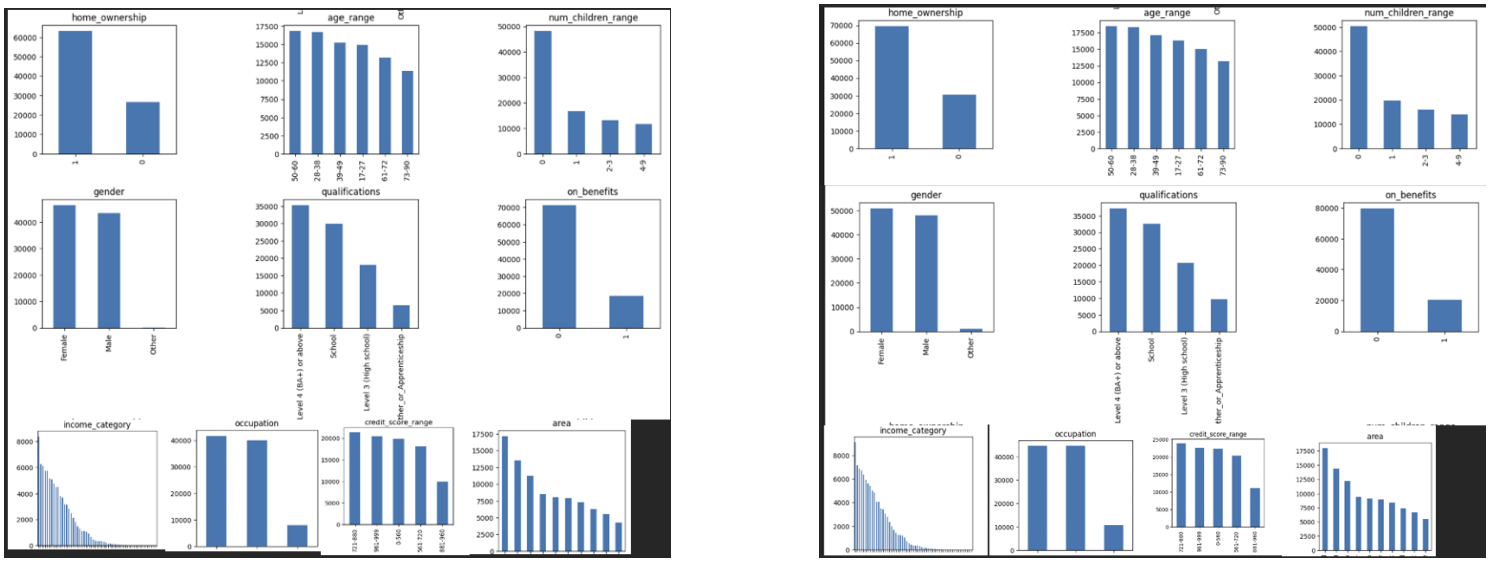
- **Reasoning:** Upon conducting a chi-squared analysis between occupations and income, we identified certain occupation categories exhibiting analogous patterns to other groups. This observation prompted us to employ the k-means algorithm on occupations, clustering them based on income group ratios. Experimenting with various k-values and employing similar evaluation techniques as before, we settled on generating 3 clusters. This choice aimed to avoid isolating 'No occupation' into its own cluster due to its strong association with the sensitive attribute 'on_benefits' (97% of 'No occupation' fall under 'on_benefits').
- **Effect on Utility:** As mentioned before, k-means guarantees that occupations with similar incomes are grouped together. This was done due to potential usage in the second use-case.

2.8 Pick the best K for k-anonymity

We tested k-anonymity for different k values ranging from 1-10, and we decided to pick k = 6 as it would allow us to keep 90% of the original dataset. Our analysis revealed that the maximum size of an equivalence class is 438,

with an average size of 20. The current de-identified dataset lacks l-diversity due to potential implications for utility. Implementing l-diversity would require widening income ranges, possibly obfuscating gender-based income disparities if they exist, which would impact the second use case. Moreover, adjustments to the other two sensitive columns, 'on_benefits' and 'credit_score', are constrained. The 'on_benefits' column, being binary, cannot be further simplified. Additionally, altering the 'credit_score', which adheres to the official Experian range, would compromise accuracy in the third use case.

3 Use-cases Guarantees



(a) Before Removing Records

(b) After Removing Records

Figure 3: Comparison of Record Removal

As previously mentioned, we deleted around 10% of the original dataset. To ensure that the deletion did not impact the data distribution, we created histograms before and after the deletion. As can be seen in 3 there has not been any significant shift in the distribution, thus not altering the utility and statistical properties of the dataset.

4 Privacy Guarantees – Risks

We ensure a 6-anonymous dataset with an average equivalence class size of 20, and the largest class comprises 438 records. Thus, on average, an attacker equipped with all quasi-identifiers can not distinguish an individual from 19 others. However, our de-identified dataset remains susceptible to homogeneity attacks due to a minimum l-diversity of 1. This vulnerability extends across all sensitive attribute combinations. Additionally, semantic attacks are viable as financial status inference is possible without adequate l-diversity. The 'Income' categorization, for instance, labels someone earning £120,000 - £124,999 similarly to another individual earning £125,000 - £129,999, both possibly considered rich.

References

[1] B. Francis-Devine and D. Pyper, "The gender pay gap," *Parliamentary paper, House of Commons library briefing*, vol. 7068, pp. 1–38, 2020.

[2] [Online]. Available: https://www.experian.co.uk/consumer/experian-credit-score.html

[3] J. Akin, "What happens to your credit when you get married?" Jan 2021. [Online]. Available: https://www.experian.com/blogs/ask-experian/credit-education/life-events/marriage-and-credit