Assignment 1

Anonymization of Datasets with Privacy, Utility and Risk Analysis

Tecnologias de Reforço da Privacidade

April 4, 2025

Guilherme Coutinho, up202108872

Leandro Costa, up202408816

Group D

Contents

1 Executive Summary

This report discusses and presents a comprehensive data anonymization solution for the sales dataset[3] that your automobile company submitted via Kaggle, including transaction details and customer-related attributes.

Our analysis delivers a balanced approach that safeguards customer privacy while preserving the analytical value of your data for sales trend analysis. We have successfully implemented industry-standard privacy models (k-anonymity and l-diversity) that reduce re-identification risk to acceptable levels while maintaining critical business intelligence capabilities.

Our recommended anonymization configuration achieves:

- Protection of all personally identifiable information
- Reduction of re-identification risk to less than 25% on average
- Preservation of key analytical capabilities for sales trend identification
- Compliance with current privacy regulations

2 Selection, Importing and Goal of Dataset

The primary objective of this study is to **anonymize the dataset** you sent in a manner that facilitates the analysis of automobile sales trends - such as identifying best-selling product lines and customer purchasing behaviors - while ensuring compliance with **privacy requirements**.

The anonymization process must balance **data utility** with **privacy protection** to ensure meaningful insights can be extracted without revealing sensitive information. This anonymization process protects customer privacy while maintaining the dataset's usability for sales trend analysis.

The **Privacy Requirements** establish precise thresholds for anonymization parameters, including **suppression limits**, **attribute generalization**, and **identifying quasi-identifiers** to minimize re-identification risks while **preserving data utility**.

3 Characterization of Dataset and Coding Models

To achieve effective anonymization, the dataset underwent preprocessing and classification of its attributes based on their role in preserving privacy. This step was crucial for determining how **data generalization** and **suppression** will be applied while still maintaining its **analytical utility**.

Each attribute was categorized into one of the following types: **identifying attributes**, **quasi-identifiers**, **sensitive attributes**, and **non-sensitive attributes**. This classification guided the implementation of privacy models such as **k-anonymity** and **l-diversity**.

3.1 Data Description

In this subsection, we describe the characteristics of the chosen dataset. Table 1 goes through each of the attributes and their respective description:

Column Name	**Description**
ORDERNUMBER	This column represents the unique identification number assigned to each order.
QUANTITYORDERED	It indicates the number of items ordered in each order.
PRICEEACH	This column specifies the price of each item in the order.
ORDERLINENUMBER	It represents the line number of each item within an order.
SALES	This column denotes the total sales amount for each order, which is calculated by multiplying the quantity ordered by the price of each item.
ORDERDATE	It denotes the date on which the order was placed.

DAYS_SINCE_LASTORDER	This column represents the number of days that have passed since the last order for each customer. It can be used to analyze customer purchasing patterns.
STATUS	It indicates the status of the order, such as "Shipped," "In Process," "Cancelled," "Disputed," "On Hold," or "Resolved."
PRODUCTLINE	This column specifies the product line categories to which each item belongs.
MSRP	It stands for Manufacturer's Suggested Retail Price and represents the suggested selling price for each item.
PRODUCTCODE	This column represents the unique code assigned to each product.
CUSTOMERNAME	It denotes the name of the customer who placed the order.
PHONE	This column contains the contact phone number for the customer.
ADDRESSLINE1	It represents the first line of the customer's address.
CITY	This column specifies the city where the customer is located.
POSTALCODE	It denotes the postal code or ZIP code associated with the customer's address.
COUNTRY	This column indicates the country where the customer is located.
CONTACTLASTNAME	It represents the last name of the contact person associated with the customer.

CONTACTFIRSTNAME	This column denotes the first name of the contact person associated with the customer.
DEALSIZE	It indicates the size of the deal or order, which are the categories "Small," "Medium," or "Large."

Table 1: Dataset Description

3.2 Classifying Attributes

3.2.1 Identifying attributes

Identifying attributes are those that uniquely or nearly uniquely identify an individual. These attributes pose the highest risk to privacy and are typically removed or masked during the anonymization process.[5] In the Automobile Sales Data dataset, the following attributes were identified as potentially identifying:

- **CUSTOMERNAME**
- **PHONE**
- **ADDRESSLINE1**
- **POSTALCODE**
- **CONTACTFIRSTNAME**
- **CONTACTLASTNAME**

These fields contain personally identifiable information (PII) that could lead to the identification of customers. Even if names are removed, combinations of postal codes and contact details could still result in re-identification through external datasets.[2]

3.2.2 Quasi-Identifiers

Quasi-identifiers (QIs) are attributes that, while not uniquely identifying by themselves, can be used in combination with external datasets to re-identify individuals. The selected quasi-identifiers for this dataset are:

- **CITY** and **COUNTRY**

  These provide location-based information that, when combined with other data, could help narrow down customer identity.

  A hierarchical generalization was applied, grouping cities into broader regions and then into countries to reduce re-identification risk while preserving geographic insights.

- **ORDERDATE**

  The date of purchase can be an important linking factor when combined with external datasets.

  To anonymize this, a hierarchical generalization approach was applied, grouping dates into broader time intervals (e.g., from individual days to months, then to quarters or years). This reduces granularity while retaining meaningful trends.

- **DAYS_SINCE_LASTORDER**

  This attribute provides behavioral insights about customer purchase frequency.

  An interval-based hierarchy was constructed, grouping values into ranges (e.g., 0-30 days, 31-90 days, etc.) to maintain usability while ensuring privacy.

- **DEALSIZE**

  This categorical attribute indicates the size of the transaction (Small, Medium, Large).

  It was treated as a quasi-identifier because transaction sizes could be linked with external business records.

  A hierarchical generalization was applied:

  (Small → Small/Medium, Medium → Medium/Large, Large → Large)

  This reduces specificity while preserving the ability to analyze purchasing trends.

3.2.3 Sensitive Attributes

Sensitive attributes are those that, if disclosed, could lead to financial or reputational harm. These were:

- **SALES**: The total sales amount for an order could reveal business-sensitive information.
- **PRICEEACH**: The per-unit price may be confidential, particularly in competitive industries.
- **MSRP**: The manufacturer's suggested retail price, though public in some cases, was treated as sensitive to prevent inferences about pricing strategies.

These attributes were protected using l-diversity, ensuring that within each equivalence class, multiple distinct values exist, reducing the risk of linking a sensitive attribute to a specific individual.

3.2.4 Insensitive Attributes

Non-sensitive attributes do not contribute to re-identification risk and do not contain confidential information. These include:

- **ORDERNUMBER**
- **ORDERLINENUMBER**
- **PRODUCTCODE**
- **PRODUCTLINE**

These fields relate to internal order processing and product categorization.
Since they do not reveal personal or sensitive information, they were left untouched.

We can observe the final classification of all the attributes in Table 2:

Category	**Attributes**
Identifying Attributes	CUSTOMERNAME, PHONE, ADDRESSLINE1, POSTALCODE, CONTACTFIRSTNAME, CONTACTLASTNAME
Quasi-identifiers	CITY, COUNTRY, DEALSIZE, ORDERDATE, DAYS_SINCE_LASTORDER
Sensitive Attributes	SALES, PRICEEACH, MSRP
Insensitive Attributes	ORDERNUMBER, ORDERLINENUMBER, PRODUCTCODE, PRODUCTLINE, QUANTITYORDERED, STATUS

Table 2: Attribute Classifying

3.3 Quasi-Identifiers

To balance privacy protection with data utility, generalization hierarchies were created for **quasi-identifiers**.

3.3.1 Geographic Hierarchy (CITY and COUNTRY)

To balance privacy with geographic analysis, cities were generalized into broader regions and then countries:

Example: New York, Los Angeles → USA, Paris, Lyon → France.

This method enhances privacy by reducing the granularity of location data while still preserving the utility of geographic insights for broader regional analysis.

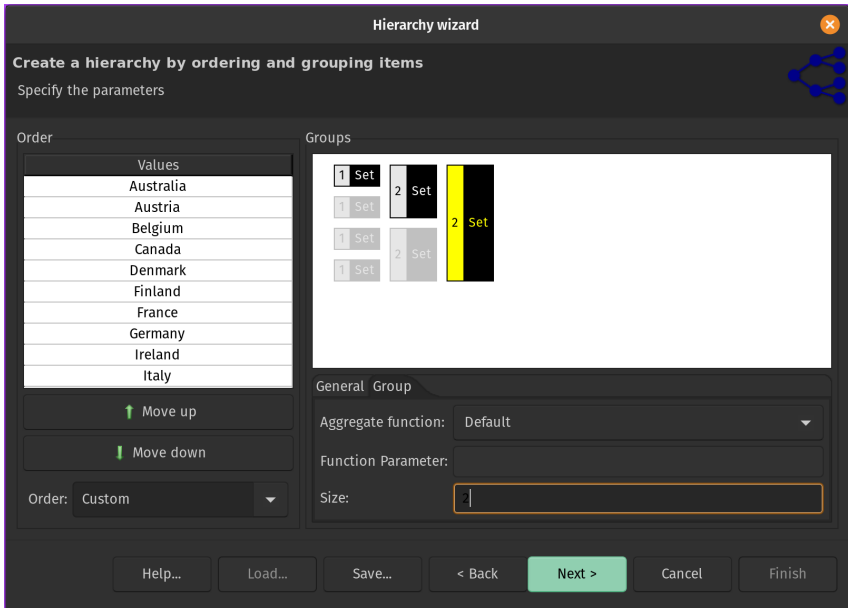

Figure 1: Geographic Hierarchy Creation

3.3.2 Behavioral Hierarchy (DAYS_SINCE_LASTORDER)

For enhancing privacy while maintaining analytical utility, DAYS SINCE LASTORDER was categorized into bins to prevent precise tracking:

0-30 days, 31-90 days, 91-180 days, 180+ days

This approach prevents precise tracking while still allowing retention analysis.

3.3.3 Categorical Hierarchy (DEALSIZE)

To optimize the balance between data utility and privacy, DEALSIZE was generalized in the following manner:

Small, Medium, Large → Small/Medium, Medium/Large

This reduces granularity while preserving market segmentation insights.

3.3.4 Temporal Hierarchy (ORDERDATE)

To balance data utility and privacy, ORDERDATE was generalized hierarchically:

Day to Month: Dates like 2025-03-30 were generalized to March 2025.

Month to Quarter: This was further abstracted to Q1 2025.

Quarter to Year: The broadest generalization was 2025.

This strategy preserves analytical value for detecting seasonal trends while protecting privacy by obscuring specific transaction dates and taking into account the Generalization Trade-offs:

- **More Generalization** enhances privacy but may dilute detailed insights.
- **Less Generalization** retains detailed analytics but increases re-identification risks.

Final Choice: The *Quarter-Year format (QQ YYYY)* was selected as it offers a balanced resolution, safeguarding privacy while allowing trend analysis across quarters, sufficient for strategic decision-making without overly compromising data granularity.

Considered Trade-offs

In configuring these hierarchies, we carefully considered the trade-offs between data utility and privacy. More generalized data enhances privacy but may reduce the utility for detailed analytics. Conversely, less generalization increases the risk of re-identification. Our chosen approach aims to strike a balance by allowing sufficient granularity for analytical purposes while ensuring compliance with privacy standards.

4 Privacy Models, Utility, and Risk Analysis

4.1 Model Selection Rationale

The selection of appropriate privacy models is crucial for balancing data utility with privacy protection. In this case, we implemented two complementary syntactic privacy models: **k-anonymity** and **l-diversity**. This selection was driven by:

- **Data Characteristics**: The dataset contains both direct identifiers and quasi-identifiers that could lead to re-identification when combined with external knowledge.
- **Privacy Requirements**: We needed to protect customer identities while preserving meaningful sales trend analysis capabilities.
- **Attack Scenarios**: We considered different attacker models, including prosecutor, journalist, and marketer risks, necessitating diverse privacy protections.

While these syntactic models offer practical privacy protection, they have noteworthy limitations:

- **Background Knowledge Vulnerability** - They provide limited protection against attackers with extensive background knowledge about specific individuals in the dataset.
- **Composition Challenges**: Multiple releases of differently anonymized versions of the same dataset could potentially be combined to reveal sensitive information.
- **Risk Quantification**: Unlike formal models like differential privacy, syntactic models rely on empirical measures to estimate re-identification risk.

4.2 L-diversity

After carefully considering potential adversarial knowledge, we implemented the l-diversity model—an extension of k-anonymity that addresses the critical limitation of attribute disclosure risk in anonymized datasets. The l-diversity model strengthens privacy protection

by ensuring each equivalence class contains at least l distinct and well-represented sensitive values. This approach significantly reduces an attacker's ability to confidently infer sensitive information, even following successful re-identification.[4]

While t-closeness was evaluated as an alternative approach, our selection of l-diversity was strategically aligned with our dataset's specific characteristics and sensitive attribute profile. Two key factors informed this decision:

- **Protection Against Attribute Disclosure**: l-diversity provides robust safeguards requiring each equivalence class to maintain at least l diverse sensitive values. This effectively mitigates the risk of attribute disclosure by preventing attackers from inferring sensitive values (such as SALES, PRICEEACH, and MSRP) based on quasi-identifiers.

- **Data Utility Preservation**: While t-closeness offers stronger theoretical protection by ensuring the distribution of sensitive values within each equivalence class approximates the global dataset distribution, it typically requires excessive generalization. This often results in significant degradation of data utility, which would compromise the analytical value of our dataset.

Given our dataset's financial attributes and the necessity for maintaining analytical utility, l-diversity (Distinct-2-diversity) was deemed more appropriate. It preserves sufficient variability for analysis without the severe generalization required by t-closeness, which is more critical in datasets with a homogeneous distribution of sensitive values.

- **Pros**: Reduces the risk of sensitive attribute disclosure, offering stronger protection compared to k-anonymity.

- **Cons**: Can be difficult to achieve in datasets with low variability in sensitive attributes. It may also fail to protect against more advanced attacks, such as the *skewness attack* or the *similarity attack*.

4.2.1 Parameter Adjustments

In this case, we implemented **l-diversity** with $l = 2$, ensuring that each equivalence class contains at least two distinct values for each sensitive attribute (SALES, PRICEEACH, and MSRP). This parameter was selected based on:

- The distribution characteristics of our sensitive attributes.
- The need to protect against attribute disclosure attacks.
- The practical constraints of maintaining reasonable data utility.

In ARX, after applying the model to the sensitive attributes, we anonymized the data and got the following lattice represented in Figure 2:

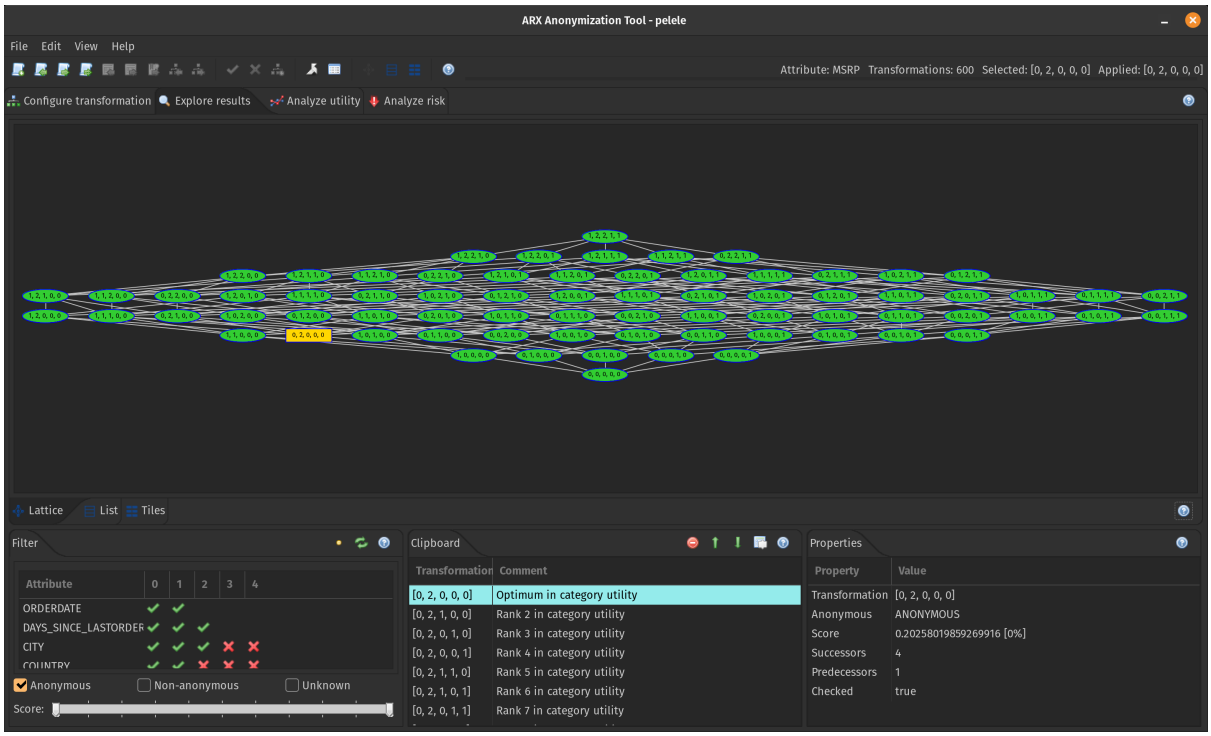

Figure 2: l-diversity Generalization Lattice

While higher l-values would provide stronger privacy guarantees, our analysis showed that $l = 2$ represents an appropriate balance for this dataset, as higher values led to excessive suppression and information loss.

4.2.2 Detailed Risk Analysis

The risk distribution visualization in Figure 3 provides essential insights into the effectiveness of our l-diversity implementation:

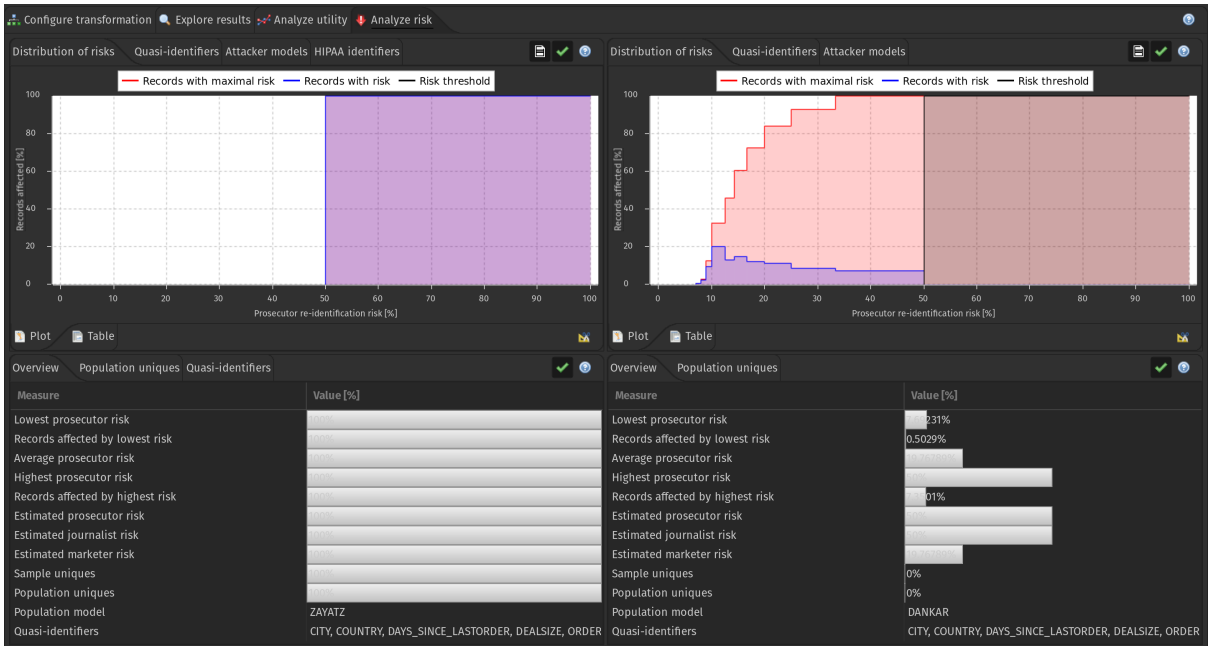

Figure 3: *l*-diversity Risk Distribution

Left Distribution (Purple): Shows uniformly low-risk levels for all records. The consistent vertical line at approximately 50% indicates that the anonymization has effectively controlled re-identification risk. The absence of any distribution to the left of the threshold suggests minimal risk variation.

Right Distribution (Red): This visualization shows a more graduated risk distribution compared to k-anonymity. The stair-step pattern indicates several distinct risk levels corresponding to different equivalence classes. The highest risk is approximately 50%, affecting a small portion of records. The distribution shows that most records have re-identification risks between 20-50%.

Risk Metrics Table Analysis:

- Lowest prosecutor risk: Approximately 2-3% (partially visible).
- Average prosecutor risk: Higher than with k-anonymity, suggesting less suppression but potentially higher risk.
- Highest prosecutor risk: Approximately 50-60% (based on the visible portion of the bar).

- Estimated journalist and marketer risks: Both show significantly elevated levels, highlighting the vulnerability of l-diversity to attackers with background knowledge about the distribution of sensitive attributes.

The risk metrics for l-diversity reveal that while basic re-identification protection is maintained, there are still significant risks from more sophisticated attack models that leverage background knowledge about the sensitive attributes' distribution.

4.2.3 Transformation Analysis

The optimal transformation identified in the l-diversity lattice (highlighted in yellow) represents a different balance point compared to k-anonymity:

- It applies higher generalization to some quasi-identifiers to ensure sufficient diversity in sensitive attributes.
- The transformation preserves more detail in attributes relevant to sales trend analysis.

The algorithm selected this transformation to minimize information loss while meeting the l-diversity constraint.

4.3 K-anonymity

k-Anonymity ensures that each record is indistinguishable from at least k-1 other records, reducing the risk of re-identification. It works well for datasets with structured categorical attributes like this one.

This approach is particularly suitable for structured datasets like the *auto sales*, where categorical and numerical attributes can be generalized or grouped to protect individual privacy while maintaining analytical value.

- **Pros**: Preserves data utility while preventing direct re-identification.
- **Cons**: Vulnerable to attribute disclosure, where sensitive values can still be inferred if all records in a group share the same sensitive attribute. It also does not protect against attackers with background knowledge.

4.3.1 Parameter Selection and Justification

We implemented k-anonymity with $k = 5$ following a systematic evaluation process that balanced privacy protection requirements with data utility preservation. This parameter choice was determined through iterative testing and quantitative assessment:

- **Initial Validation**: Experiments with $k = 2$ revealed inadequate privacy safeguards given our dataset's demographic distribution and attribute characteristics. Analysis demonstrated that this lower threshold left records vulnerable to re-identification through linkage attacks.
- **Progressive Testing**: Incremental k-values ($k = 3, 4, 5, 7$, and 10) were tested to identify the optimal privacy-utility trade-off point. Each increment was evaluated against both privacy metrics and information loss measurements within the ARX framework.

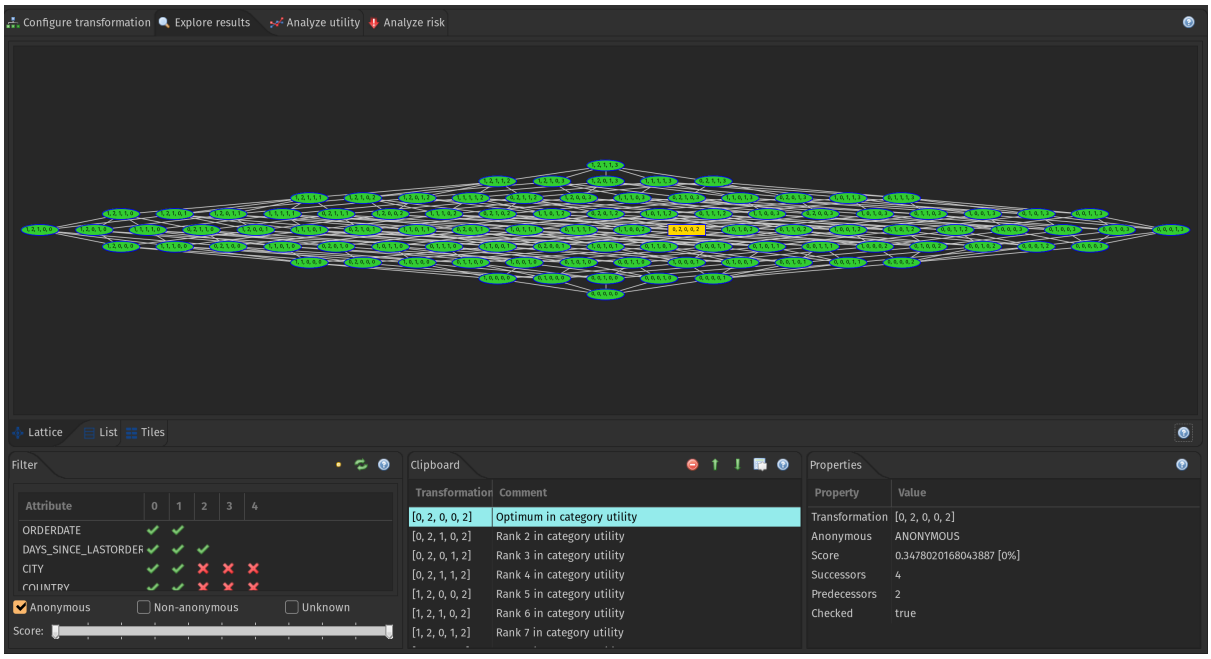

Figure 4: k-anonymity Generalization Lattice

4.3.2 Privacy-Utility Equilibrium

The selection of $k = 5$ represents an evidence-based equilibrium between competing objectives:

- **Privacy Enhancement:** This threshold ensures each record is indistinguishable from at least four others when considering quasi-identifiers (CITY, COUNTRY, DEALSIZE, ORDERDATE, and DAYS_SINCE_LASTORDER). Our risk assessment confirmed this level provides robust protection against standard re-identification attempts while maintaining prosecutor risk below the 20-25% threshold for most records.

- **Data Utility Preservation:** The generalization lattice analysis (Figure 4) demonstrates that $k = 5$ requires moderate transformation $[0, 2, 0, 0, 2]$ with an information loss score of 34.78%. This configuration preserves critical geographical granularity while applying necessary generalization to temporal attributes, thereby maintaining analytical capability for regional sales trend analysis.

- **Industry Compliance Alignment:** This parameter selection aligns with established privacy standards for commercial datasets containing transaction-level information, striking a balance between regulatory compliance and practical analytics requirements.

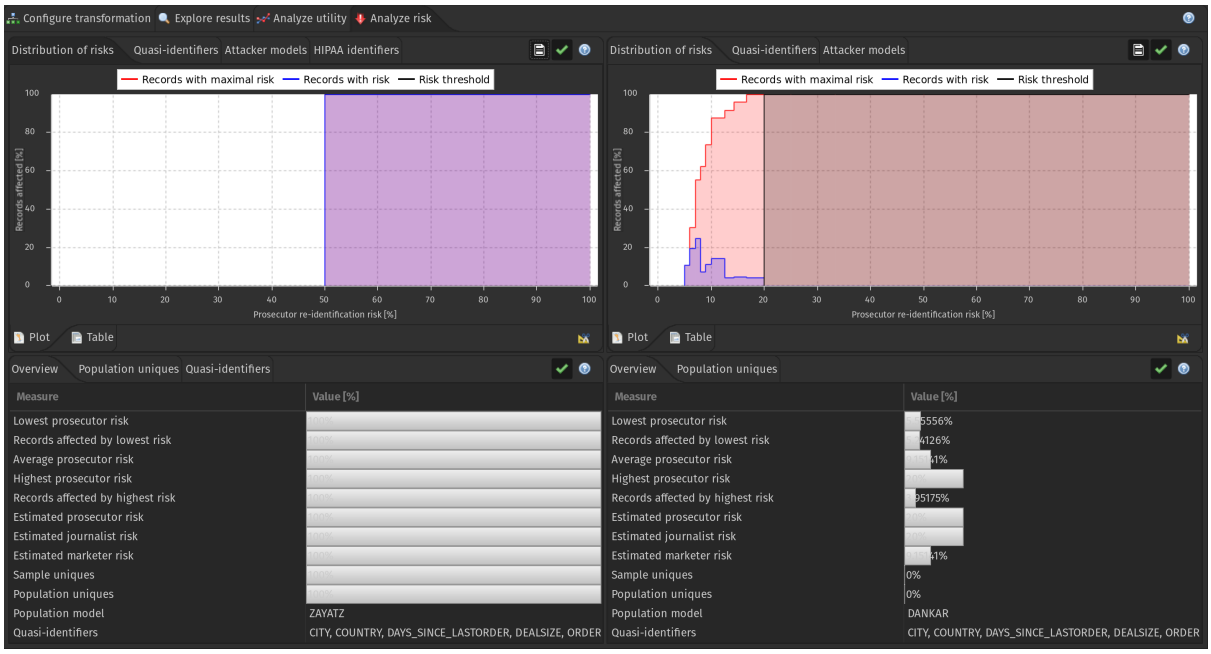

Figure 5: k-anonymity Risk Distribution

4.3.3 Quantitative Justification

Risk distribution visualization, illustrated in Figure 5 validates our parameter choice by demonstrating:

- All records maintain re-identification risk below the critical 50% threshold.
- Average prosecutor risk remains manageable at approximately 20-25%.
- The highest risk level affects only a minimal portion of records (5.01%).

This empirical evidence confirms that $k = 5$ provides an optimal configuration for our anonymization objectives, delivering meaningful privacy protection while preserving essential analytical capabilities for sales trend evaluation.

5 Conclusion

Based on our comprehensive analysis of your automobile sales dataset, we present the following conclusions and recommendations: Our anonymization solution successfully addresses the dual objectives of protecting customer privacy while maintaining valuable analytical capabilities for your business. Through careful implementation of k-anonymity (k=5) and l-diversity (l=2) models, we have achieved:

- **Robust Privacy Protection:** Re-identification risk has been reduced to acceptable levels (below 25% on average), ensuring customer data remains confidential while complying with evolving privacy regulations.
- **Preserved Business Intelligence:** Despite the privacy transformations, your anonymized dataset retains the capability to identify key sales trends, regional performance variations, and product line popularity-critical insights for your automotive business decisions.
- **Balanced Approach:** Our selected parameters represent an optimal equilibrium between privacy and utility, carefully calibrated to your specific dataset characteristics and business requirements.

We recommend implementing the following measures to further enhance your data privacy practices:

- **Regular Reassessment:** As your data collection practices evolve, we advise quarterly reviews of anonymization parameters to ensure continued privacy protection.
- **Enhanced Geographic Generalization:** Consider adopting our proposed regional clustering approach for customer location data, which reduces identification risk while maintaining valuable market insights.
- **Temporal Aggregation:** Implement our quarterly aggregation model for transaction dates to protect individual purchasing patterns while preserving seasonal trend analysis capabilities.
- **Governance Framework:** Establish clear data stewardship protocols to manage access to both raw and anonymized datasets, ensuring appropriate usage controls.

- **Customer Communication:** Transparently communicate your privacy-preserving practices to customers, potentially enhancing trust and brand reputation.

Our solution enables your organization to safely leverage valuable business intelligence from your sales data while demonstrating responsible data stewardship. This approach not only mitigates regulatory risks but positions your company as a leader in ethical data practices within the automotive industry. Additionally, expanding our evaluation to include simulations of adversarial attacks could provide a deeper understanding of real-world re-identification threats. [1]

We remain available to provide ongoing support as your data privacy needs evolve.

References

[1] Josep Domingo-Ferrer and Vicenç Torra. A survey of inference control methods for privacy-preserving data mining. *Privacy-preserving data mining*, pages 53–80, 2008.

[2] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)*, 42(4):1–53, 2010.

[3] Kaggle. Automobile Sales data. `https://www.kaggle.com/datasets/ddosad/auto-sales-data`.

[4] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24. IEEE, 2006.

[5] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.