Homework Assignment 1

ST2DMG - Data Management and Ethics

Data Quality Analysis

Analyze and report on the quality of three datasets for an e-commerce company: Customer Dataset, Sales Dataset, and Product Dataset. Evaluate them across six data quality dimensions and provide a comprehensive report.

Datasets Provided:

Customer Dataset: Customer ID, title, name, email, phone number, address, birthday, age, subscription date.

Sales Dataset: Product ID, customer ID, transaction date, currency, amount.

Product Dataset: Product name, category, price, availability.

Data Quality Analysis

Objective :

Assess and report on the following data quality dimensions for each dataset:

- **Completeness**: Identify missing or incomplete data.
- **Uniqueness:** Detect duplicate records and ensure unique identifiers.
- **Timeliness:** Confirm data aligns with expected timeframes.
- **Validity:** Verify conformity to formats and standards.
- **Accuracy:** Ensure data reflects real-world values.
- **Consistency:** Check for alignment within and across datasets.

Data Quality Analysis

Tasks

1. Define data quality rules.
2. Use AWS or another platform to analyze the data.
3. Create a report including:
   - Data quality metrics per dimension for each dataset, along with an overall average score.
   - Extracts of rows with issues.
   - Recommendations for corrections.

Data Protection

On the same provided datasets

Instructions:

- Classify each column (then table) into one of the following categories : [Public, Internal, Confidential, Restricted]
- Identify Privacy Risks
- Suggest Mitigation Measures
- For the set of informations that you think should be encrypted (with justifications), Show the raw values then the encrypted values using AES-256

Deliverable format : A mail attaching PDF report including, required elements and screenshots of code and results

Group work allowed (4 max) : list to be provided before   19/12/2024

Submission Deadline: 21/01/2025

Good Luck !