Bharath Kumar Bolla

Apply data privacy measures (removal, secure storage, masking) and data anonymization techniques to the dataset provided.

Data ethics assignment

Part 2

# Table of Contents

[2.1 Data Classification: 2](#_Toc142250252)

[2.2 Attribute Masking: 3](#_Toc142250253)

[2.3 - Non-Sensitive Attribute Identification: 4](#_Toc142250254)

[2.4 - Non-Personal Attribute Identification: 5](#_Toc142250255)

[2.5 - Data Anonymization: 6](#_Toc142250256)

[2.6 - Updated Files Submission: 8](#_Toc142250257)

## 2.1 Data Classification:

To perform data classification and data minimization on the given dataset, we will first identify the sensitive attributes containing personally identifiable information (PII) and then determine which of these attributes can be removed or stored separately with limited access.

Sensitive Information (PII) in the dataset:

1. Member's Full Name

2. Member's Address

3. Member's Email

4. Member's Phone Number

5. Member's Date of Birth

6. Credit Card Number

Data Minimization:

Considering the sensitive nature of the attributes mentioned above, we should take the following steps for data minimization:

1. Remove or Masking Credit Card Number: Since the "Credit Card Number" column contains NaN values, indicating missing data, it is essential to remove this attribute entirely from the dataset or Credit card numbers are highly sensitive so should not be stored in the dataset without masking.

2. Store Sensitive Attributes Separately: Sensitive attributes like "Member's Full Name," "Member's Address," "Member's Email," "Member's Phone Number," and "Member's Date of Birth" should be stored separately with limited access. These attributes can be stored securely in a separate database or file, and access should be granted only to authorized personnel who need this information for specific purposes, such as customer support or legal compliance.

3. Anonymize Data: For the remaining sensitive attributes that are essential for analysis or modeling, such as "Member's Date of Birth," we can apply data anonymization techniques to protect individual identities. One approach is to use generalization, where we group individuals into age groups instead of specifying their exact birth dates. For example, instead of specifying individual dates of birth, we can use age groups like "18-25," "26-35," etc.

4. Non-Sensitive Attributes: Non-sensitive attributes like "Order," "SKU," "Created On," "Description," "Order Value," and "Delivery Address" can be retained in the dataset as they do not contain sensitive or personally identifiable information.

By performing data classification and data minimization, we ensure that sensitive information is protected, and access to this information is limited to authorized personnel, minimizing the risk of data breaches and unauthorized access.

## 2.2 Attribute Masking:

Attribute masking or anonymization involves transforming sensitive attributes in such a way that customer identities are protected while preserving the utility and integrity of the dataset for analysis or modeling. In the given dataset, the following attributes can be masked or anonymized:

1. Member's Full Name: Instead of using the actual names, we can replace them with pseudonyms or generic labels. For example, "John Doe" can be masked as "Customer A," "Jane Smith" as "Customer B," and so on.

2. Member's Address: To protect customer identities, we can generalize or truncate the addresses. For example, "19774 Kelly Square, South Dennis, DC 11507" can be masked as "Address A, City A, State A."

3. Member's Email: Email addresses can be anonymized by using generic labels or random strings. For instance, "stacyturner@example.org" can be masked as "customer123@example.com."

4. Member's Phone Number: Similar to email addresses, phone numbers can be masked using random strings or generic labels. For example, "+1-628-447-8451x2516" can be masked as "Phone A."

5. Member's Date of Birth: To protect individual ages, we can generalize the birth dates into age groups. For example, "1921-07-03" can be masked as "Age Group: 80-89."

6. Member's Cradit Card Number: Remove or mask credit card numbers, showing only few digits.

By applying attribute masking or anonymization to the above sensitive attributes, we ensure that customer identities are protected, and the risk of re-identification is minimized. This allows us to conduct analysis or modeling while preserving data privacy and complying with data protection regulations.

## 2.3 - Non-Sensitive Attribute Identification:

Non-sensitive attributes in the dataset are those that do not contain personally identifiable information (PII) or sensitive data. These attributes can still provide valuable insights for analysis or machine learning purposes and can be leveraged to improve various aspects of business operations. Some of the non-sensitive attributes in the given dataset and their potential uses are as follows:

1. Order: The order number can be used to track individual transactions, monitor purchase behavior, and analyze order patterns. It can help in identifying popular products or services, understanding customer preferences, and optimizing inventory management.

2. SKU (Stock Keeping Unit): SKUs can be used to categorize products and track inventory levels. Analyzing SKU data can help in identifying top-selling products, predicting demand, and optimizing supply chain operations.

3. Created On: The creation date of the order can be used to analyze order trends over time. It can help in identifying seasonal variations in demand, understanding peak order periods, and optimizing promotional strategies.

4. Description: The product descriptions can provide insights into the types of products being sold. Analyzing the description data can help in identifying product categories, understanding customer preferences, and optimizing product offerings.

5. Member's Purchase History: This attribute indicates the number of past purchases made by each member. Analyzing purchase history can help in identifying loyal customers, segmenting customers based on their buying behavior, and implementing targeted marketing strategies.

6. Order Value: The order value represents the total cost of each order. Analyzing order values can help in understanding customer spending patterns, identifying high-value customers, and optimizing pricing strategies.

7. Payment Method: This attribute indicates the method of payment used for each order. Analyzing payment methods can provide insights into customer payment preferences and help in optimizing payment processing systems.

8. Order Status: The order status indicates the current status of each order (e.g., pending, shipped, etc.). Analyzing order status data can help in tracking order fulfillment, identifying potential bottlenecks in the order processing pipeline, and improving customer service.

These non-sensitive attributes can be utilized to improve customer experience by personalizing product recommendations, enhancing operational efficiency by optimizing inventory and supply chain management, and generating business intelligence by understanding customer behavior and market trends. Proper analysis of these attributes can lead to data-driven decision-making, improved customer satisfaction, and overall business growth.

## 2.4 - Non-Personal Attribute Identification:

Non-personal attributes in the dataset are those that do not contain personally identifiable information (PII) and are not considered sensitive in nature. These attributes provide valuable information for data analysis, market research, and statistical modeling without posing privacy concerns. Some of the non-personal attributes in the given dataset and their potential benefits are as follows:

1. Order: The order number is a non-personal attribute that uniquely identifies each transaction. It can be leveraged in data analysis to track individual orders, analyze order patterns, and identify popular products or services. Market research can use order numbers to understand customer behavior, preferences, and purchase trends.

2. SKU (Stock Keeping Unit): SKUs are non-personal attributes used to categorize and identify products. They can be valuable in data analysis to track product sales, inventory levels, and demand patterns. Market research can utilize SKUs to segment products based on categories and study customer preferences for specific product types.

3. Created On: The creation date of the order is a non-personal attribute that can be used to analyze order trends over time. It is helpful in data analysis to understand seasonal variations in demand, identify peak order periods, and optimize promotional strategies. Market research can study order trends to develop marketing campaigns aligned with customer buying behavior.

4. Description: Product descriptions are non-personal attributes that provide insights into the types of products being sold. They can be used in data analysis to categorize products and identify product attributes driving customer preferences. Market research can utilize product descriptions to evaluate customer sentiment and identify potential gaps in product offerings.

5. Member's Purchase History: The number of past purchases made by each member is a non-personal attribute that can be used to segment customers based on buying behavior. It is valuable in data analysis to identify loyal customers, high-value customers, and patterns in repeat purchases. Market research can use purchase history to identify customer segments with different buying preferences.

6. Order Value: The order value represents the total cost of each order and is a non-personal attribute. It is useful in data analysis to understand customer spending patterns, identify high-value orders, and optimize pricing strategies. Market research can use order values to analyze customer segments with different spending capacities.

7. Payment Method: Payment methods used for each order are non-personal attributes that can be analyzed to understand customer payment preferences. Data analysis can reveal popular payment methods and identify potential opportunities for offering diverse payment options to customers.

8. Order Status: The current status of each order is a non-personal attribute that can be used to analyze order fulfillment and customer satisfaction. It helps in data analysis to identify bottlenecks in the order processing pipeline and improve overall customer service.

The potential benefits of leveraging non-personal attributes include data-driven decision-making, optimized business operations, improved customer experiences, and targeted marketing strategies. Non-personal attributes offer valuable insights into customer behavior, market trends, and product performance, enabling businesses to make informed and strategic decisions for growth and success.

## 2.5 - Data Anonymization:

Data anonymization is the process of modifying or transforming data in such a way that it no longer contains identifiable information about individuals. The main goal of data anonymization is to protect the privacy of individuals while allowing for data analysis, research, and sharing. Anonymization ensures that the data cannot be linked back to specific individuals, thereby reducing the risk of privacy breaches and unauthorized access.

Importance of Data Anonymization:

Data anonymization is crucial in preserving privacy and ensuring compliance with data protection regulations (e.g., GDPR, HIPAA). It allows organizations to share and analyze data without compromising the confidentiality of sensitive information. Some key reasons why data anonymization is important are:

1. Privacy Protection: Anonymization helps safeguard sensitive personal information, preventing potential misuse or identity theft.

2. Legal Compliance: Many data protection laws require organizations to protect the privacy of individuals by anonymizing data.

3. Data Sharing: Anonymization enables secure data sharing with external partners, researchers, or vendors, fostering collaboration without exposing sensitive details.

4. Risk Mitigation: Anonymization reduces the risk of data breaches or insider threats, as the data no longer contains personally identifiable information.

Anonymization Techniques:

Several anonymization techniques can be applied to the dataset to protect privacy. Some common techniques include:

1. Generalization: This involves replacing specific values with a broader category. For example, replacing exact dates of birth with age groups (e.g., 20-30, 30-40) helps protect individual identities while retaining the essential demographic information.

2. Suppression: This technique involves removing or suppressing certain attributes that are sensitive or highly identifiable. Columns containing personally identifiable information, such as names, addresses, and contact details, can be entirely removed or masked.

3. Randomization: Randomization involves adding noise or perturbation to numerical data, making it difficult to link specific records to individuals accurately. This can be useful in protecting sensitive numerical attributes.

4. Data Swapping: Data swapping involves exchanging data between individuals, effectively breaking the link between individual identities and their associated attributes.

5. Data Perturbation: Perturbation involves altering data values slightly to protect individual identities. This technique maintains data patterns but reduces the risk of re-identification.

6. K-Anonymity: K-anonymity ensures that each record in the dataset is indistinguishable from at least 'k' other records, thereby providing a higher level of privacy protection.

7. L-Diversity: L-diversity extends k-anonymity by ensuring that each group of 'k' indistinguishable records has at least 'l' distinct values for sensitive attributes.

Conclusion:

Data anonymization is a critical aspect of data privacy and security. By employing anonymization techniques such as generalization, suppression, and randomization, organizations can protect individual identities while retaining the utility of the data for analysis, research, and decision-making. Properly anonymized data enables responsible data sharing, adheres to data protection regulations, and safeguards individuals' privacy in an increasingly data-driven world.

## 2.6 - Updated Files Submission:

Data Privacy and Anonymization Report

1. Data Privacy Measures:

In order to protect sensitive information and ensure data privacy, the following measures have been applied to the dataset:

- Removal of Personally Identifiable Information (PII): Columns containing PII such as 'Member's Full Name', 'Member's Address', 'Member's Email', 'Member's Phone Number', and 'Credit Card Number' have been removed from the dataset. This ensures that sensitive personal details of customers are not included in the analysis.

- Secure Storage: The dataset is stored in a secure environment with limited access, ensuring that only authorized personnel can access and analyze the data. This helps to prevent unauthorized access to sensitive information.

- Attribute Masking: Non-sensitive attributes that can still be linked to individual identities have been masked to protect customer identities. For example, 'Member' column has been masked or anonymized to preserve the integrity of the dataset while ensuring customer privacy.

2. Data Anonymization Techniques:

Anonymization techniques have been applied to protect the privacy of individuals while allowing for data analysis. The following techniques have been used:

- Generalization: Age and date of birth information has been generalized into age groups to protect individual's exact age. For example, 'Member's Date of Birth' has been converted into 'Age Group' to provide age ranges instead of exact dates of birth.

- Suppression: Columns with sensitive information that cannot be anonymized or generalized have been suppressed or removed from the dataset to eliminate the risk of exposing personal data. For example, 'Member's Full Name', 'Member's Address', 'Member's Email', 'Member's Phone Number', and 'Credit Card Number' have been suppressed.

3. Anonymized Dataset:

The anonymized dataset is provided below:

```python

import pandas as pd

# Load the anonymized dataset

data\_anonymized = pd.read\_excel("DataEethicsAssignmentdataset\_anonymized.xlsx")

# Convert 'Created On' column to datetime type

data['OrderDate'] = pd.to\_datetime(data['Created On'], errors='coerce')

data['DOB'] = pd.to\_datetime(data['Member\'s Date of Birth'], errors='coerce')

data['DOB'] = pd.to\_datetime(data['DOB'], format='%Y-%m-%d')

# Calculate age using current date

current\_date = datetime.now()

data['Age'] = (current\_date - data['DOB']).astype('<m8[Y]')

print(data['Age'].sample(5))

# Only took non sensitive info for dataset

df\_model = data[["Member", "Member's Gender", "Payment Method", "Member's Membership Level",

"Order Status", "Age Group", "LastPurchaseDays", "Order Value Bin"]]

# Masking credit card numbers

num\_rows = len(data\_DDI)

random\_card\_numbers = [str(random.randint(10\*\*15, 10\*\*16 - 1)) for \_ in range(num\_rows)]

data\_DDI['Masked Credit Card Number'] = random\_card\_numbers

# Masking Email

def mask\_email(email):

parts = email.split('@')

username = parts[0]

masked\_username = re.sub(r'\S', '\*', username[:-3]) + username[-3:]

return masked\_username + '@' + parts[1]

data\_DDI["Masked Member's Email"] = data\_DDI["Member's Email"].apply(mask\_email)

# masking phone numbers

def mask\_phone(phone):

phone\_str = str(phone) # Convert phone number to string

return re.sub(r'\d', '\*', phone\_str[:-4]) + phone\_str[-4:]

data\_DDI["Masked Member's Phone Number"] = data\_DDI["Member's Phone Number"].apply(mask\_phone)

# Whole Code is for to mask Sensitive Data

# masking Member's Full Name

# Masking function

def mask\_name(name):

names = name.split(" ")

if len(names) > 1:

first\_name, last\_name = names[0], names[1]

masked\_last\_name = "\*" \* len(last\_name)

return f"{first\_name} {masked\_last\_name}"

else:

return "\*" \* len(names[0])

# Applying mask to the column

data\_DDI["Masked Member's Full Name"] = data\_DDI["Member's Full Name"].apply(mask\_name)

# Masking function for address

def mask\_address(address):

parts = address.split(", ")

masked\_parts = [part if idx == 0 else "\*" \* len(part) for idx, part in enumerate(parts)]

return ", ".join(masked\_parts)

# Applying mask to the columns

data\_DDI["Masked Member's Address"] = data\_DDI["Member's Address"].apply(mask\_address)

data\_DDI["Masked Delivery Address"] = data\_DDI["Delivery Address"].apply(mask\_address)

4. Conclusion:

By applying data privacy measures such as removal of PII, secure storage, and attribute masking, along with anonymization techniques like generalization and suppression, the dataset has been safeguarded to protect sensitive information and ensure the privacy of individuals. The anonymized dataset can now be used for data analysis, market research, and statistical modeling without compromising the privacy and confidentiality of the customers' personal details.