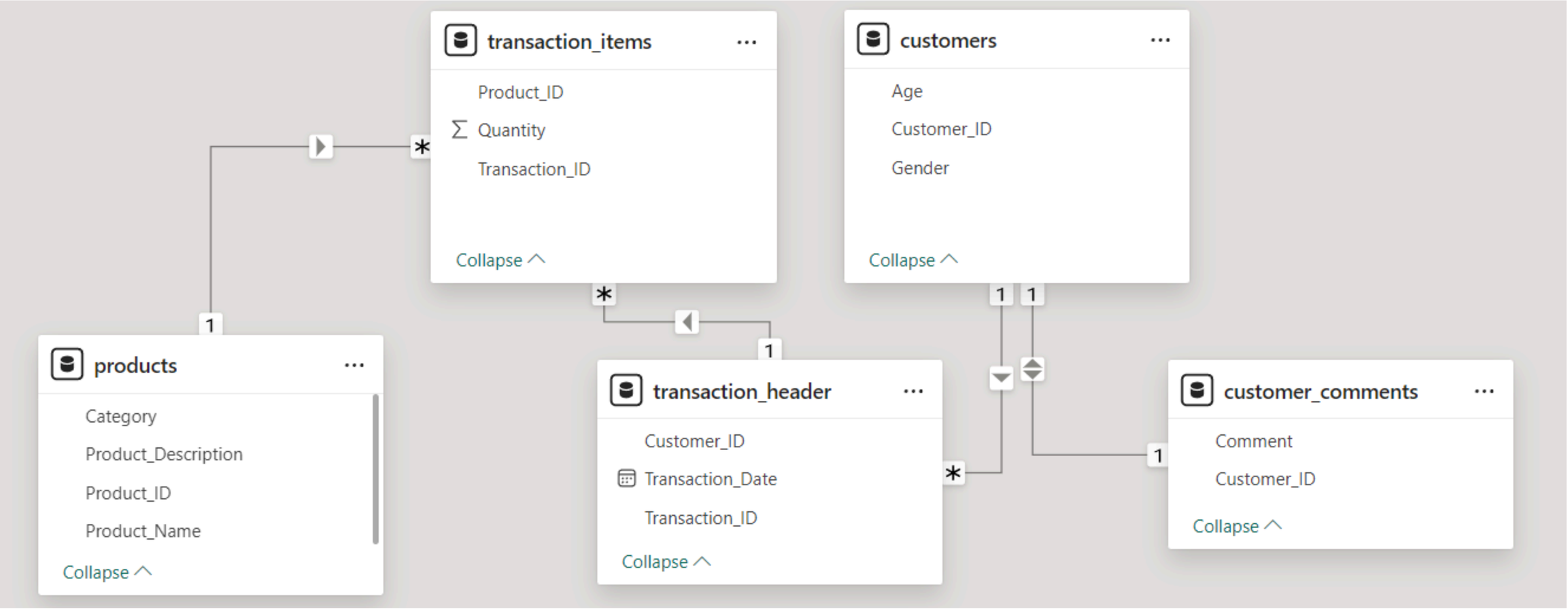
MADT 8102

Lab2 - Feature Store Design

Exercise 1: Understanding common data structure

ER Diagram



Data Dictionary

1. products

- Category: Category of the product (e.g., electronics, groceries).
- Product_Description: Detailed description of the product.
- Product_ID: Unique identifier for each product (Primary Key).
- Product_Name: Name of the product.

2. transaction_items

- Product_ID: Foreign Key linking to products. Refers to the product purchased in the transaction.
- Quantity: The number of units of the product purchased in the transaction.
- Transaction_ID: Foreign Key linking to transaction_header. Refers to the specific transaction.

3. transaction_header

- Customer_ID: Foreign Key linking to customers. Refers to the customer who made the transaction.
- Transaction_Date: Date when the transaction occurred.
- Transaction_ID: Unique identifier for each transaction (Primary Key).

4. customers

- Customer_ID: Unique identifier for each customer (Primary Key).
- Age: Age of the customer.
- Gender: Gender of the customer.

5. customer_comments

- Customer_ID: Foreign Key linking to customers. Refers to the customer who provided the comment.
- Comment: Text comment or feedback provided by the customer.

https://colab.research.google.com/drive/1gG8m7DG0QNhJT2j7T4gd6IPcn4_hmPrd#scrollTo=UuT48U0HceG6

Exercise 2: Feature Engineering in Python

Feature	Description	Origin
Customer_ID	Unique identifier for each customer.	From customers.csv dataset.
Total_Transactions	Total number of unique transactions per customer.	Derived from transaction_header.csv by counting unique Transaction_ID for each Customer_ID.
Total_Quantity	Total quantity of items purchased by the customer across all transactions.	Derived from transaction_items.csv by summing Quantity for each Customer_ID.
Total_Spending	Total spending by the customer (quantity bought).	Calculated by summing Quantity for each customer (could be refined if price data is available).
Gender_Female	Binary indicator of whether the customer is female (1 if female, 0 otherwise).	One-hot encoded from the Gender field in customers.csv.
Gender_Male	Binary indicator of whether the customer is male (1 if male, 0 otherwise).	One-hot encoded from the Gender field in customers.csv.
Recency	Number of days since the customer's most recent transaction.	Derived from transaction_header.csv, calculated as the difference between the current date and last transaction.
Avg_Transactions_Per_Month	Average number of transactions per month.	Derived from transaction_header.csv by calculating monthly frequency of transactions for each customer.
Avg_Quantity_Per_Transaction	Average quantity of items bought per transaction.	Calculated by dividing total quantity by the number of transactions for each customer.
Most_Frequent_Category	The most frequently bought product category for each customer.	Extracted from products.csv and aggregated to find the mode for each Customer_ID.
Distinct_Products_Bought	The number of distinct products purchased by each customer.	Derived from transaction_items.csv, counting unique Product_ID for each customer.
Avg_Spending_Per_Transaction	Average spending per transaction.	Total quantity divided by the number of transactions for each customer.
Weekend_Transactions	Number of transactions made on weekends.	Boolean feature derived from transaction_header.csv based on Transaction_Date falling on a weekend.
Total_Comments	Total number of comments left by the customer.	Counted from customer_comments.csv by Customer_ID.
Avg_Sentiment_Score	Average sentiment score derived from customer comments (positive or negative).	Basic sentiment analysis applied to customer_comments.csv using keywords for positive and negative comments.

Exercise 2: Feature Engineering in Python

Feature	Description	Origin
Total_Electronics_Bought	Total quantity of items bought in the "Electronics" category.	Summed from products.csv where Category is "Electronics".
Distinct_Transaction_Dates	Number of distinct transaction dates for each customer.	Count of unique Transaction_Date for each customer from transaction_header.csv.
Max_Quantity_In_Single_Transaction	Maximum quantity of items bought in a single transaction.	Maximum value of Quantity from transaction_items.csv for each customer.
Product_Diversity	Ratio of distinct products bought to the total quantity of products.	Calculated by dividing distinct product count by total quantity for each customer.
Avg_Time_Between_Transactions	Average time (in days) between each transaction.	Calculated from transaction_header.csv by finding the average time gap between transactions.
Days_Since_First_Transaction	Days since the customer's first transaction.	Calculated as the difference between the first transaction date and the current date.
Favorite_Brand	The brand the customer buys most frequently.	Extracted from product descriptions in products.csv and aggregated by mode for each Customer_ID.
Total_High_End_Products_Bought	Total quantity of high-end products bought (e.g., products with "Pro" or "Ultra" in their name).	Filtered from Product_Name in products.csv and summed for each customer.
Product_Feedback_Comments	Number of comments where the customer mentioned a product.	Counted from customer_comments.csv where the comment contains the word "product".
Most_Bought_Product	The product bought most frequently by the customer.	Mode of Product_Name from transaction_items.csv for each Customer_ID.
Spending_On_Most_Bought_Product	Total spending on the customer's most frequently bought product.	Summed for the most frequent product for each customer.
Avg_Sentiment_Per_Product	Average sentiment score for products the customer purchased.	Calculated as the average sentiment for product-related comments.
Total_Words_In_Comments	Total number of words in customer comments.	Summed from the word counts in customer_comments.csv.

Exercise 2: Feature Engineering in Python

https://colab.research.google.com/drive/1s8Rq9Vc39WcwdPJtrhhFf01NgdzOMoe5#scrollTo=S9C1yRw0xHOI

Comment Extracting (RegEx)

Customer_ID	Comment	Price_Mention	Delivery_Mention	Quality_Mention	Sentiment_Score	Sentiment
1	I love this product! The performance is am	No Price Mention	No Delivery Issue	No Product Quality Mention	10	Positive
2	Not satisfied with the quality, the screen b	No Price Mention	No Delivery Issue	Product Quality Mention	7	Positive
3	Fast shipping, excellent service, and the p	No Price Mention	Delivery Issue	No Product Quality Mention	9	Positive
4	The product is okay but not as described.	Price Mention	No Delivery Issue	No Product Quality Mention	5	Neutral
5	Highly recommend this product! It exceed	No Price Mention	No Delivery Issue	No Product Quality Mention	7	Positive
6	The worst purchase I've ever made. Woul	No Price Mention	No Delivery Issue	No Product Quality Mention	2	Negative

Product Description Extracting (RegEx)

Product_ID	Product_Name	Category	Product_Description	Extracted_Nouns	Brands	Features	Product_Types	Technical_Specs	Brand	Product_Type
1	Product_1	Electronics	BrandX SuperWidget 2000 - High Performance, 32GB RAM, 256GB SSD	['BrandX', 'SuperWidget', 'High', 'Performance', 'RAM', 'SSD']	['BrandX']	['Performance', 'RAM']	['SuperWidget']	['RAM']	['BrandX']	['SuperWidget']
2	Product_2	Electronics	BrandY MegaPhone Pro - Lightweight, 128GB Storage, 6GB RAM	['BrandY', 'MegaPhone', 'Pro', 'Lightweight', 'Storage', 'RAM']	['BrandY']	['Storage', 'RAM']	['MegaPhone']	['RAM']	['BrandY']	['MegaPhone']
3	Product_3	Electronics	BrandZ PowerTablet - 16GB Storage, 2GB RAM, 10-inch display	['BrandZ', 'PowerTablet', 'Storage', 'RAM']	['BrandZ']	['Storage', 'RAM']	['PowerTablet']	['RAM']	['BrandZ']	['PowerTablet']
4	Product_4	Electronics	BrandA UltraCamera - Professional, 64MP, 4K Video, 128GB Storage	['BrandA', 'UltraCamera', 'Professional', 'Video', 'Storage']	['BrandA']	['Video', 'Storage']	['UltraCamera']	[]	['BrandA']	['UltraCamera']
5	Product_5	Electronics	BrandB SmartWatch 3 - GPS, Heart Rate Monitor, 16GB Storage	['BrandB', 'SmartWatch', 'GPS', 'Heart', 'Rate', 'Monitor', 'Storage']	['BrandB']	['GPS', 'Heart', 'Rate', 'Monitor', 'Storage']	['SmartWatch']	[]	['BrandB']	['SmartWatch']

Exercise 3: Designing Feature Store for Predictive model

Customer Features Table

Column	Description
Customer_ID	Unique identifier for each customer
Date	Date when the features were updated
Total_Transactions	Total number of transactions
Total_Quantity	Total number of products bought
Total_Spending	Total spending (quantity of products bought)
Recency	Days since the customer's last transaction
Avg_Transactions_Per_Month	Average number of transactions per month
Avg_Quantity_Per_Transaction	Average number of products bought per transaction
Most_Frequent_Category	Product category bought most frequently
Distinct_Products_Bought	The number of distinct products bought
Product_Diversity	Ratio of distinct products to total quantity bought
Avg_Time_Between_Transactions	Average time between transactions (in days)
Days_Since_First_Transaction	Days since the customer's first transaction
Favorite_Brand	The customer's most frequently bought brand
Total_High_End_Products_Bought	Number of high-end products bought

Exercise 3: Designing Feature Store for Predictive model

Product Features Table

Column	Description
Customer_ID	Unique identifier for each customer
Date	Date when the features were updated
Most_Bought_Product	The most frequently bought product
Spending_On_Most_Bought_Product	Total spending on the customer's most frequently bought product
Avg_Sentiment_Per_Product	Average sentiment for product-related comments
Product_Feedback_Comments	Number of product-related comments

Our Team

SURAPAT V.

6610424015

THITIPORN T.

6610424017

WARIT Y.

6610424025

CHANYANUCH B

6610424031

Thank you