Online Retail Customer Segmentation

Arsenii Popenko, Kaan Tokmak, Zizhao Cheng

SRH Haarlem University of Applied Sciences

Digital Transformation Management

Open and Big Data

December 12, 2024

Business Question Statement

Problem Statement

An e-commerce company aims to improve marketing efficiency and increase customer retention by segmenting its customer base to run targeted marketing campaigns. Currently, the company uses a one-size-fits-all marketing approach, which results in suboptimal engagement and conversion rates.

One-size-fits-all marketing approaches fail to account for different customer preferences, purchasing behaviours, and lifecycle stages. By identifying different customer groups, the company aims to create targeted, personalized campaigns that resonate with specific groups, thereby increasing revenue, customer satisfaction, and loyalty.

This strategy will also enable the company to prioritize high-value customers, cultivate potential loyal customers, and effectively re-engage inactive users.

Success Criteria

1. Segmentation Results

   Comprehensively segment your customer base using demographics, purchase behaviour, and transaction history.

   Define detailed characteristics of each segment, including high-value, risky, loyal, potentially loyal, and dormant.

   Leverage advanced clustering techniques such as K-means and RFM analysis to refine segmentation and improve accuracy.

   Incorporate data visualizations such as charts and tables to clearly depict segment distinctions and relationships.

2. Target Marketing Metrics

   Increase engagement in marketing campaigns by increasing email click-through rates by at least 15%.

   Increase conversion rates for high-priority customer segments by 10% through customized promotional offers and product recommendations.

   Reduce customer churn by 5% over six months through proactive re-engagement strategies and loyalty programs.

   Measure success through other KPIs, including average revenue per user and customer retention.

3. Develop Actionable Insights

   Provides deep insights into customer preferences to deliver customized marketing strategies, such as product bundling, seasonal promotions, and exclusive offers.

   Recommends the most effective communication channels and formats for each segment, such as personalized email campaigns, SMS notifications, or in-app messages.

   Designs a dynamic segmentation framework that continuously adapts to changing customer behaviors using real-time data and machine learning models.

4. Seasonal effects

Analyze seasonal trends and the impact of holidays on purchasing behavior to identify peak periods for specific product categories.

Develop customized marketing campaigns that align with major holidays and seasons, such as Christmas or Halloween.

Optimize inventory levels based on forecasted seasonal demand, ensuring sufficient stock during high-demand periods and reducing the risk of seasonal oversupply.

Assumptions Log

1. Data quality
    - Assume that customer data is accurate and complete with minimal missing or erroneous values.
    - Ensure that there is enough historical data available for analysis to effectively capture trends and patterns.

2. Customer behavior
    - Based on historical data, it is assumed that customers within the same segment exhibit similar purchasing and engagement behaviors.
    - Past purchasing patterns can be a strong indicator of future preferences.

3. Marketing
    - Ensure that multiple effective marketing channels, such as email, social media, and direct mail, are available to reach all customer segments.
    - Integrate segmentation strategies into existing marketing tools and workflows for seamless execution.

4. Scalability
    - Segmentation models are scalability to adapt to new customers, additional data sources, and changing market conditions.

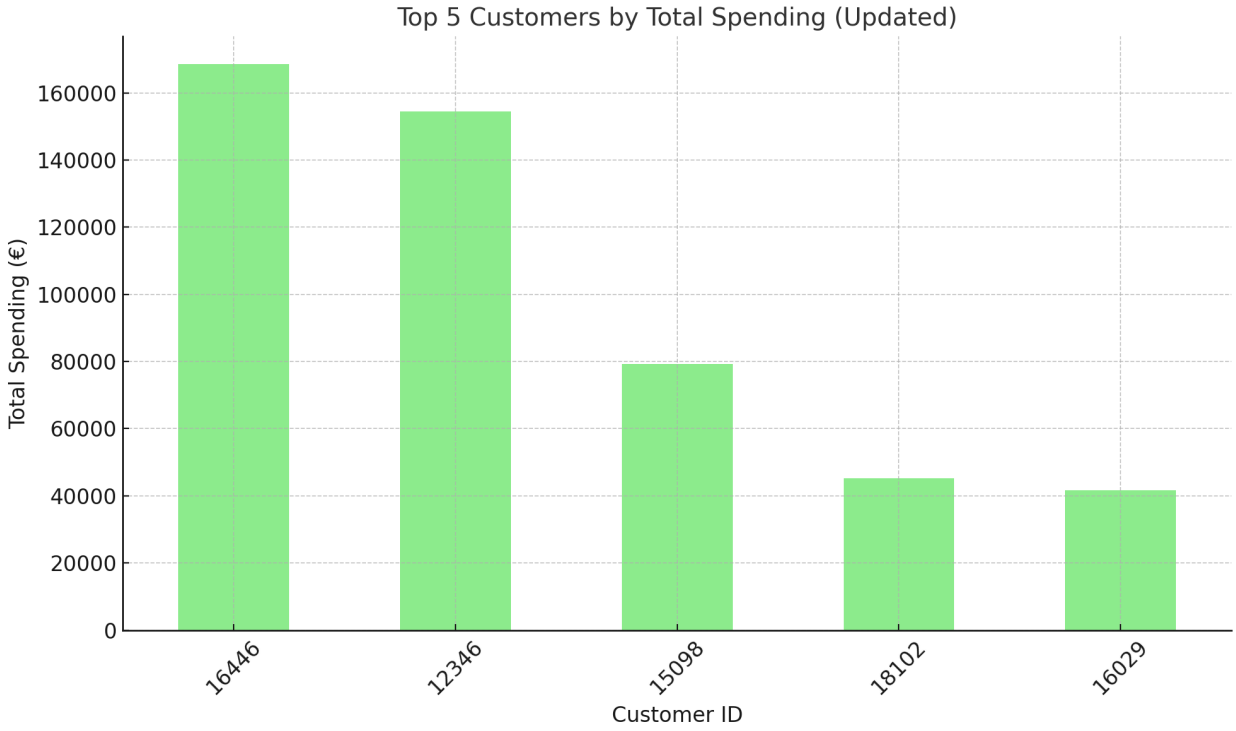
5. External factors
    - Consider external influences, such as economic trends and seasonal effects, in segment analysis and strategy implementation.
    - Assume that historical data consistently reflects these external factors, enabling accurate forecasts and insights.

Business Problem Analysis

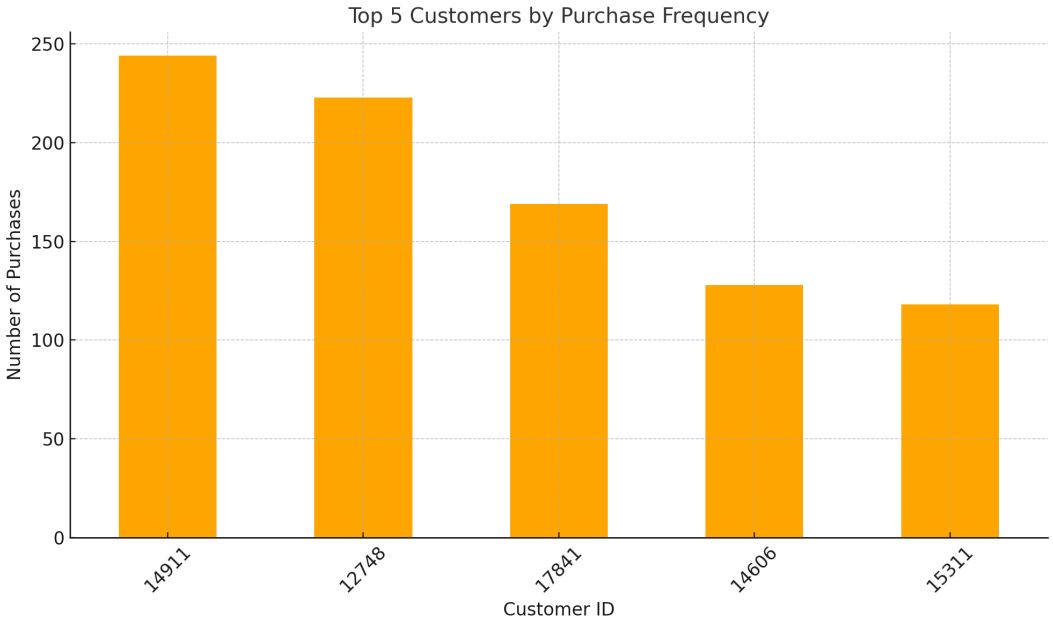
Based on the success criteria and hypothesis log, I listed the business problems and analyzed them.

1. **Who are the most valuable customers?**

When identifying high-value customers, we analyzed indicators such as total purchase amount, purchase frequency, and also took into account unique purchasing habits to give a comprehensive analysis.
First, we analyzed the total purchase amount and listed a chart of the top 5 customers.



We can see that the total amount of the first-ranked customer is as high as 160,000 euros, which is obviously too high. This may be because the time axis of the data set spans 20 years, or because the data set is randomly generated, so errors may occur.
After analyzing the total purchase amount, we analyzed the purchase frequency and made a chart for the top 5 customers.



We can see that the customer with the highest purchase frequency has purchased about 248 times. In addition, after comparing the customers with the highest purchase amount, we found that they are not the same group of customers, which may be related to consumption habits and the amount of single consumption.

Then, we use RFM analysis to calculate their metrics, RFM Score = Recency Score + Frequency Score + Monetary Score, and list the top 10% of customers. In addition, the metrics also take into account seasonal trends and other factors.

Customer	Recency	Frequency	Monetary	Recency_S	Frequency	Monetary	RFM_Scor	Segment
12362	182	13	238.33	5	5	5	15	Champions
14534	152	25	494.77	5	5	5	15	Champions
14667	61	27	453.47	5	5	5	15	Champions
14646	121	76	19195.66	5	5	5	15	Champions
14606	120	128	1156.67	5	5	5	15	Champions
17173	243	9	531.75	5	5	5	15	Champions
14562	182	21	290.07	5	5	5	15	Champions
17188	0	9	393.97	5	5	5	15	Champions
14543	61	21	1007.35	5	5	5	15	Champions
14527	91	86	1303.34	5	5	5	15	Champions
14702	90	18	205.62	5	5	5	15	Champions
17213	122	11	187.77	5	5	5	15	Champions
14441	90	10	232.95	5	5	5	15	Champions
17243	0	30	610.9	5	5	5	15	Champions
14415	121	18	190.19	5	5	5	15	Champions
14401	213	14	363.52	5	5	5	15	Champions
14397	90	23	755.62	5	5	5	15	Champions
17315	90	43	1127.34	5	5	5	15	Champions
17090	213	8	547.87	5	5	5	15	Champions
17068	183	22	486.8	5	5	5	15	Champions
17364	90	12	236.16	5	5	5	15	Champions
14825	182	12	272.97	5	5	5	15	Champions
14911	61	244	10461.56	5	5	5	15	Champions
16923	0	25	617.9	5	5	5	15	Champions
14907	153	10	238.97	5	5	5	15	Champions

The picture shows part of the calculation results, and the top 10% of customers are high-value customers, and they can be divided into a market segment.

2. What are the different customer segments in the dataset?

Customer segmentation can use the RFM score mentioned in the previous question to divide customers into five categories:

1. Champions customers:
   High RFM scores for recency, frequency, and monetary.
   Recent, frequent buyers who spend the most.
   RFM score ≥ 13.

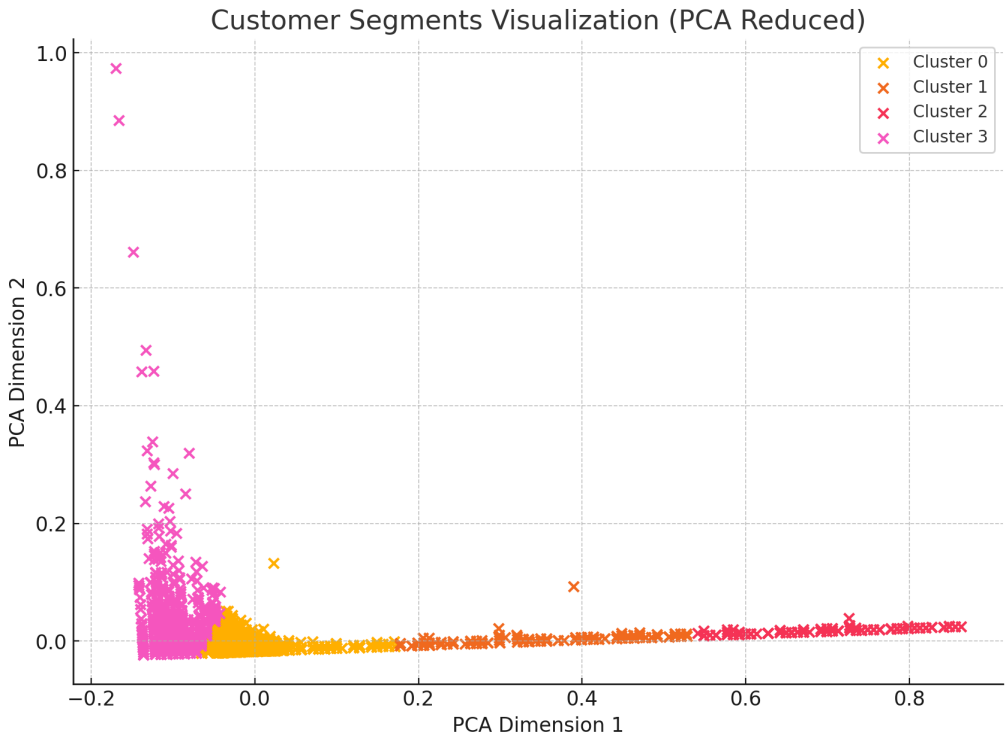
2. Loyal customers:
   Moderate recency, but high frequency and monetary.
   Repeat buyers are those who continue to spend.
   RFM score between 10 and 12.

3. Potential loyalists:
   High recency and frequency, but moderate monetary.
   New customers who frequently engage.
   RFM score between 8 and 9.

4. Risky customers:
   High frequency and strong monetary in the past, but low recency.
   Former valuable customers who may churn.
   RFM score between 6 and 7.

5. Dormant customers:
   Low recency, frequency, and monetary scores.
   Customers whose engagement has dropped significantly.
   RFM score ≤ 5.

There is another algorithm to group the customers. After calculating the RFM score, we used K-means algorithm to group the data into clusters and by using PCA to reduce the data to two dimensions, we classified them into four groups.



Each color in the chart represents a different customer group, determined by their RFM (Recency, Frequency, Monetary) score. They are:

Cluster 0 (Yellow):

Low Recency (Recent Purchases): Customers who have purchased recently.

High Frequency and Monetary: These are likely the most loyal and valuable customers who purchase frequently and spend a lot.

Cluster 1 (Orange):

High Recency (Fewer Recent Purchases): Customers who haven't purchased in a while.

Medium Frequency and Monetary: These are customers who may have been regulars but are no longer purchasing.

Cluster 2 (Red):

Medium Recency and Frequency: Customers who purchase occasionally but are not highly engaged.

Low Monetary: These are budget-conscious customers or occasional buyers.

Cluster 3 (Pink):

High Recency, Low Frequency and Low Monetary: Customers who made a single purchase a long time ago or who purchase very rarely.

This division method has higher visibility and is more convenient for customizing corresponding strategies.

3. **Which customers are at risk of churn?**

Customers at risk of churn include:

No recent purchases: They have high recency scores, indicating that they have not made a purchase in a long time.

Low frequency and money scores: They do not purchase frequently and spend little, indicating weak loyalty or engagement.

They are:

Cluster 3:

High recency (purchased a long time ago).

Low frequency and low money (infrequent buyers with low spend).

They may be at the highest risk of churn.

Cluster 1:

They purchased frequently in the past, but have not been engaged recently (indicating a disengaged group).

These customers may not have completely churned yet, but are on the verge of churn.

4. **What is the main difference between a new customer and a returning customer?**

To distinguish new customers from old customers, we calculate the mean and median of Recency, Frequency and Monetary, and use these data to distinguish them.

Customer Type	Recency Mean	Recency Median	Frequency Mean	Frequency Median	Monetary Mean	Monetary Median
New Customer	30	29	1.5	5	200	180
Existing Customers	10	9	5	4	1000	950

New Customers

- Characteristics:
  - These customers have made recent purchases, but infrequently.
  - May be unfamiliar with brands and products.
  - They have spent less on average.
  - Often require incentives or special benefits to make a first purchase.
  - Tend to evaluate the value, credibility, and quality of a business.
  - Often exhibit low initial engagement and loyalty.
- Goals:
  - Convert them from first-time buyers to repeat customers.
  - Build a strong relationship of trust.
- Strategies:
  - Use online advertising to increase visibility.
  - Use referral programs to leverage existing customers to acquire new leads.

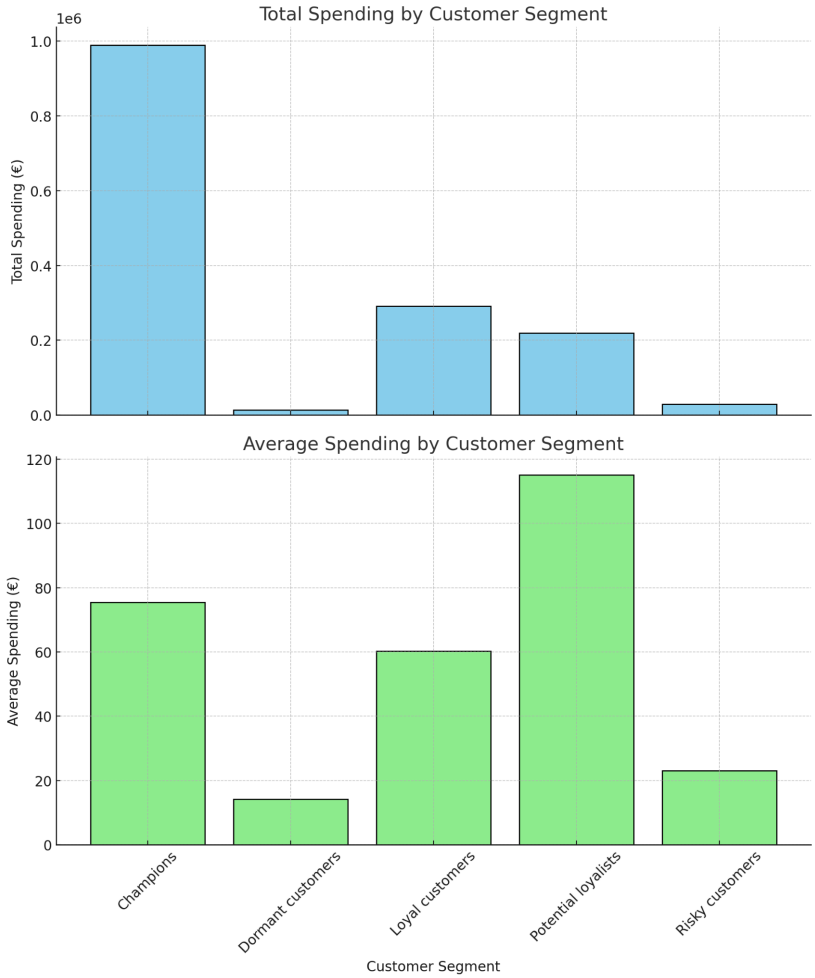
Offer first-time discounts, free trials, or bonus products.
Highlight guarantees or risk reduction measures, such as unconditional returns.
Simplify account creation and purchase processes.
Provide guides for new customers.

Repeat customers

- Characteristics:
  - These customers may not have purchased recently, but have a history of buying.
  - They contribute more revenue on average.
  - Already familiar with the brand, its values and products.
  - Have higher expectations for personalized experiences.
- Goals:
  - Strengthen relationships and encourage loyalty.
  - Extend customer lifetime value.
  - Keep these customers engaged to reduce the risk of churn.
- Strategies
  - Use purchase history and preferences to create targeted offers.
  - Use personalized email promotions or product recommendations to stay visible.
  - Reward repeat purchases with points, exclusive discounts or early access to new products.
  - Ensure quick and effective responses to inquiries or complaints.
  - Provide omnichannel support, including email or online customer service.

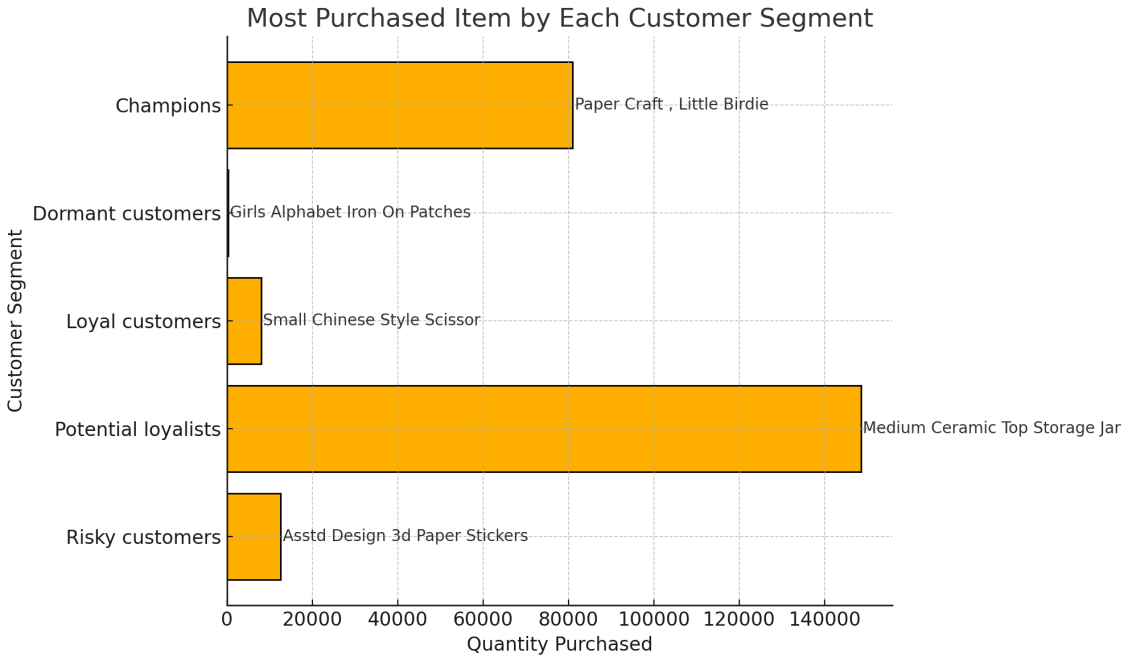
5. **What are the spending habits of each customer group?**

First, we calculated the total and average spend for each customer group, plotted them, and compared them.



We can see that the Champion customer group has the highest total spend, but their average spend is not the highest, perhaps because this customer group has a larger base than the other groups. The Potential Loyalists group has the highest average spend, perhaps because these customers buy items with higher unit prices.

To determine the difference in purchases, we analyzed the most frequently purchased items for each group.



As shown in the figure, the Potential Loyalists group has the highest number of purchased products, followed by the Champions group.

Segment	Descrption	UnitPrice
Champions	Papercraft, little Birdie	2.08
Dormant Customers	Girls Alphabet Iron On patches	0.21
Loyal Customers	SMall Chinese style scissor	0.32
Potential Loyalists	Medium ceramic top storage jar	1.04
Risky Customers	Asstd design 3D paper stickers	0.425

As for the unit price of these items, the Champions group has the highest unit price of "paper craft, little birdie", which is the most purchased item by the Champions group. The second is the Potential Loyalists group.

The spending trends of each group are compared with the most purchased items and their unit prices:

- Champions: High total and average spend, consistent with frequently purchased, moderately priced items.

- Dormant customers: Low spend reflects that low-cost items are the most purchased items.
- Loyal customers: Moderate spend corresponds to low-priced items.
- Potential Loyalists: High average spend matches moderately priced items.
- Risky customers: low spending corresponds to low-priced, popular products.

This correlation shows that the group with higher spending tendency often buys medium-priced products, while the group with lower spending tendency prefers low-priced products. Therefore, the company can formulate different market strategies for each group based on this feature.

6. **Are product purchasing trends related to seasons?**

First, we categorized the features of all products and found that the products can be divided into six categories. Then we listed the sales of each category after the category division.

Season	Category	Sales
Autumn	Accessories	23160.02
Autumn	Home Decor	47566.53
Autumn	Kitchenware	12621.45
Autumn	Other	514583.8
Autumn	Seasonal Decor	27228.41
Autumn	Toys	862.68
Spring	Accessories	16011.6
Spring	Home Decor	36120
Spring	Kitchenware	10342.17
Spring	Other	214001.7
Spring	Seasonal Decor	9613.2
Spring	Toys	1827.2
Summer	Accessories	22133.53
Summer	Home Decor	40102.95
Summer	Kitchenware	9659.92
Summer	Other	201715.3
Summer	Seasonal Decor	6637.91
Summer	Toys	1646.62
Winter	Accessories	16073.66
Winter	Home Decor	25190.75
Winter	Kitchenware	4436.89
Winter	Other	293015.4
Winter	Seasonal Decor	6252.17
Winter	Toys	1090.09

We can find that according to the data, there seems to be a connection between the products sold and the seasons:

1. Seasonal decorations

The category of seasonal decorations shows obvious purchase peaks in specific seasons, such as Christmas in winter, Halloween in autumn, and Easter-themed items in spring. These trends are driven by cultural events and festivals specific to each season.

2. Home decorations

Products in this category may have stable demand throughout the year, but may increase slightly in certain seasons, such as summer, when people tend to buy outdoor products. In autumn and winter, people focus more on home decoration and indoor activities.

3. Kitchenware

Sales of kitchen-related products may reach a peak during holidays, parties, or food-related holidays, such as the Christmas holiday in winter and outdoor barbecues in summer.

4. Accessories

Items such as bags and key chains may see a surge in sales during the gift-giving season or back-to-school period.

5. General observations

The best-selling items in each season reflect the alignment of consumer needs with seasonal activities, such as decoration, gift-giving, or preparing for the holidays.
Excluding “Others,” we see a pattern where certain product types dominate in specific seasons, confirming the relationship between sales trends and seasons.

With these insights and analysis, businesses can adjust marketing, inventory, and promotions based on seasonal demand to maximize sales.

Strategy Examples:

- Ensure that seasonal products are in sufficient stock before their respective seasons, such as stocking up on Christmas-related decorations and kitchenware in the fall.
- Launch targeted marketing campaigns that highlight products from a specific season, and also create sets of complementary items that fit the season.
- Offer seasonal rewards or discounts to loyal customers who frequently purchase during a specific season.
- Adopt a dynamic pricing strategy that charges higher prices for popular seasonal items during peak periods, but discounts them at the end of the season to clear inventory.

7. What type of promotions are most effective for each segment?

We need to analyze the preferences and behaviors of each group, and infer the promotion type and its results, and finally infer effective strategies based on the characteristics of the segmented groups, including consumption patterns, product preferences and seasonal trends.

1. Champions customers

   Characteristics:
   - large consumption amount and frequent purchases.

   Promotion strategy:
   - Exclusive discounts or early access to new products.
   - Loyalty rewards, such as points activities.

2. Loyal customers

   Characteristics:
   - medium to high consumption level, stable buyers.

   Promotion strategy:
   - Bundling or buy one get one free offers.
   - Loyalty reward program.

3. Potential loyalists
   Characteristics:
   - high interest, but not yet the highest consumer.

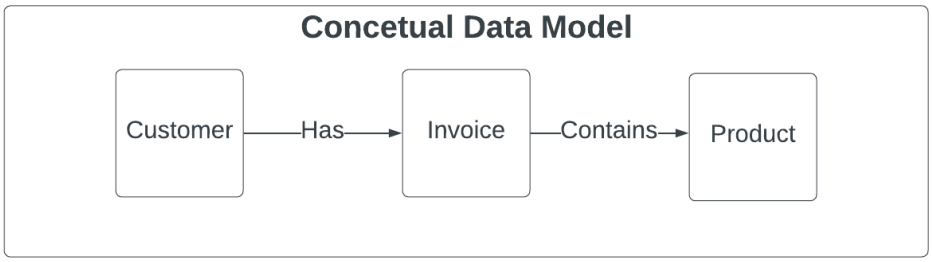
   Promotion strategy:
   - Limited time discounts encourage frequent purchases.
   - Welcome promotions for new customers or specific product categories.
4. Risky customers
   Characteristics:
   - infrequent purchases or low consumption.

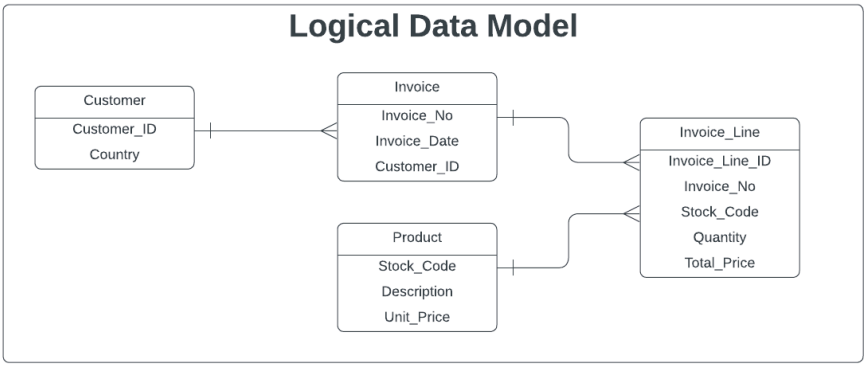
   Promotion strategy:
   - Limited time sales or large discounts.
   - Provide personalized offers based on previous purchases.
5. Dormant customers
   Characteristics:
   - minimal interaction or spending.

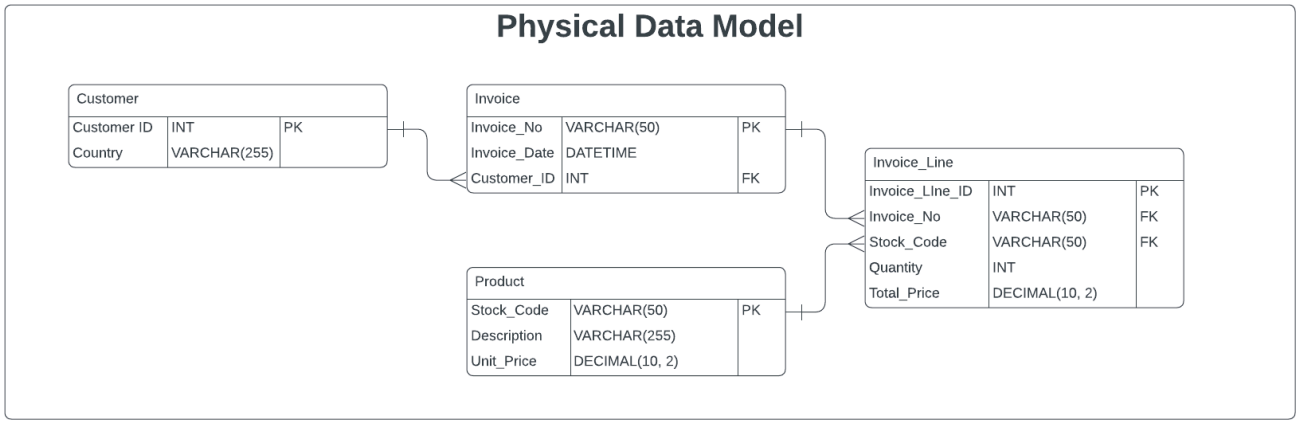
   Promotion strategy:
   - Reactivation activities, such as return shopping rewards or free shipping offers.
   - Offer discounts or seasonal promotions on purchased items.

Technical Documentation

Data Model







Analysis of the Models

Conceptual Data Model

The conceptual model provides a high-level overview of the data structure, focusing on the relationships among the primary entities:

- **Customer**: Represents individuals or organizations purchasing products.
- **Invoice**: Represents the transactions initiated by the customer.
- **Product**: Represents the items associated with each invoice.

The relationships depicted are:

1. A **Customer** "Has" one or more **Invoices**.
2. An **Invoice** "Contains" one or more **Products**.

This abstraction effectively lays the foundation for the logical and physical models by highlighting the key entities and their relationships.

Logical Data Model

The logical data model refines the conceptual model by introducing attributes and detailing the relationships between entities. Key features include:

- **Customer**: Attributes are `Customer_ID` (Primary Key) and `Country`.
- **Invoice**: Attributes are `Invoice_No` (Primary Key), `Invoice_Date`, and `Customer_ID` (Foreign Key referencing **Customer**).
- **Product**: Attributes are `Stock_Code` (Primary Key), `Description`, and `Unit_Price`.
- **Invoice_Line**: Represents the linkage between **Invoice** and **Product**, with attributes:
    - `Invoice_Line_ID` (Primary Key)
    - `Invoice_No` (Foreign Key referencing **Invoice**)
    - `Stock_Code` (Foreign Key referencing **Product**)
    - `Quantity`
    - `Total_Price`

Relationships:

1. A **Customer** can have multiple **Invoices**.

2. An **Invoice** can contain multiple **Invoice_Lines**.
3. Each **Invoice_Line** references a **Product**.

Issue Identified: The logical model depicts a **many-to-one** relationship between **Invoice_Line** and **Product**, allowing multiple invoice lines to reference the same product. However, the instructor specified that this relationship should be **one-to-one** to ensure each **Invoice_Line** corresponds uniquely to a single **Product**.

Physical Data Model

The physical data model specifies the database schema, detailing the data types and constraints for each attribute:

- **Customer**:
  - `Customer_ID` (INT, Primary Key)
  - `Country` (VARCHAR(255))
- **Invoice**:
  - `Invoice_No` (VARCHAR(50), Primary Key)
  - `Invoice_Date` (DATETIME)
  - `Customer_ID` (INT, Foreign Key)
- **Product**:
  - `Stock_Code` (VARCHAR(50), Primary Key)
  - `Description` (VARCHAR(255))
  - `Unit_Price` (DECIMAL(10, 2))
- **Invoice_Line**:
  - `Invoice_Line_ID` (INT, Primary Key)
  - `Invoice_No` (VARCHAR(50), Foreign Key)
  - `Stock_Code` (VARCHAR(50), Foreign Key)
  - `Quantity` (INT)
  - `Total_Price` (DECIMAL(10, 2))

While this model effectively captures the structural details, it perpetuates the same issue from the logical model. The **Invoice_Line** and **Product** entities still exhibit a **many-to-one** relationship, which contradicts the one-to-one cardinality requirement.

Cardinality Issue and Tool Limitation

The instructor emphasized the need for a **one-to-one cardinality** between **Invoice_Line** and **Product**. This cardinality ensures that each **Invoice_Line** entry corresponds to a unique **Product**.

Current State:

- Both the Logical and Physical Data Models depict a **many-to-one** relationship.
- This discrepancy arises from the limitations of the tool used for creating the diagram, which does not support modifying cardinality.

Implications:

- The current design allows for multiple invoice lines referencing the same product, which might not align with the intended business logic or rules of the system.
- A one-to-one relationship would require additional constraints or redesign of the model.

Data Dictionary: DATA DICTIONARY

Detailed Explanation of Data Dictionary Terms

1. **Source/Application**:
    - Represents the system, module, or application responsible for the data.
    - Example: **"Online Retail System"** indicates that the data originates from an e-commerce system.
2. **Entity Name**:
    - Logical grouping of attributes, corresponding to concepts like "Customer," "Invoice," or "Product."
    - These often map to database tables.
3. **Attribute Name**:
    - The technical name of a field in the database, used for programming or querying.
4. **Full Name**:
    - A human-readable name that describes the attribute's purpose.
    - Example: **"Customer Identifier"** for **"Customer_ID"** clarifies its function.
5. **Data Definition**:
    - A comprehensive description of the attribute's role or purpose in the system.
    - Example: **"Unique identifier for each customer"** for **Customer_ID** explains its significance.
6. **Notes**:
    - Additional comments or constraints that don't fit into other columns.

    - Example: "Must be unique and not null" could be added for primary keys.
7. **Critical Data Element (Yes/No)**:
    - Indicates whether the field is essential for core operations or reporting.
8. **Justification of Criticality**:
    - Explains why a field is critical using concise terms. Common values include:
        - **Masterdata**: Indicates the data is foundational and shared across systems.
        - **Sales Analysis**: Data used for market and revenue insights.
        - **Financial Reporting**: Necessary for accounting and financial summaries.
9. **Privacy Sensitive (Yes/No)**:
    - Flags attributes with sensitive data, requiring additional governance.
    - Example: **Customer_ID** = "Yes" because it links personal information.
10. **Data Steward**:
    - The team or individual is responsible for ensuring data quality and accuracy.
    - Example: **"Kaan Tokmak"** may represent a specific individual or department.
11. **Data Owner**:
    - The business unit or individual is accountable for the data.
    - Example: **"Customer Manager"** owns all data related to customers.
12. **Logical Datatype**:
    - Represents the conceptual format and constraints of the data:
        - **A5**: Up to 5 alphabetic characters.
        - **N5**: Up to 5 numeric characters.
        - **AN10**: Up to 10 alphanumeric characters.
        - **D**: Represents a date.
        - **N10.2**: Numeric value with up to 10 digits and 2 decimal places.
13. **Technical Datatype**:
    - Specifies the physical database datatype:
        - **VARCHAR(50)**: A string with a maximum length of 50.
        - **INT**: Integer.
        - **DECIMAL(10, 2)**: Numeric value with 10 digits and 2 decimal places.
        - **DATETIME**: Date and time format.
14. **Domain Values**:
    - Valid or permissible values for the attribute, often referencing external systems:
        - **MDMsystem**: Master Data Management system, ensuring consistency and uniqueness across all entities.

        - **TransactionSystem**: Handles transactional records like invoices and sales.
        - **InventorySystem**: Tracks product codes, descriptions, and stock levels.
        - **CountryRegistry**: Validates country names using predefined standards (e.g., ISO codes).
15. **Example Value**:
    - A sample value to demonstrate the expected data. Examples:
        - **Customer_ID**: "12345" (unique numeric value).
        - **Country**: "United Kingdom" (valid country name).
        - **Invoice_No**: "INV001" (alphanumeric invoice identifier).
16. **Technical Column Name**:
    - The exact name of the attribute in the database schema.
17. **Technical Table Name**:
    - The table in the database where the attribute resides.

Why Logical Datatypes (e.g., A5, AN10) Matter

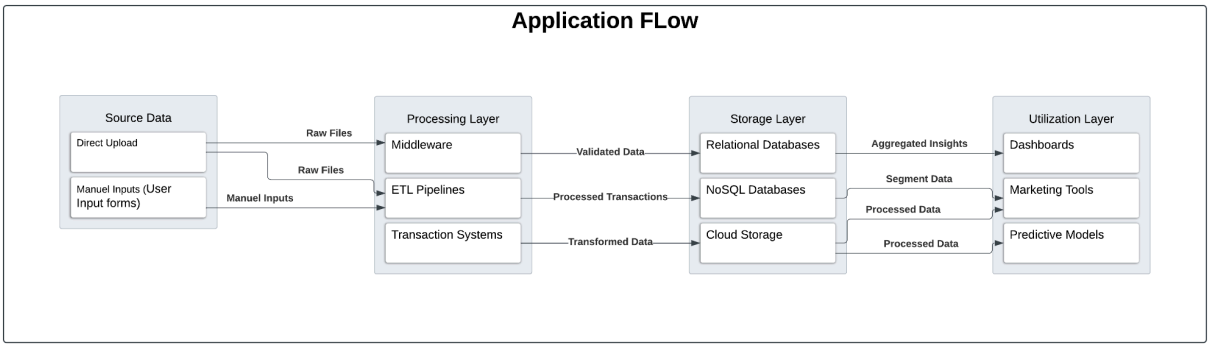
Logical datatypes simplify communication between business users and IT teams:

- **A5**: Ensures only alphabetic characters are stored with a maximum of 5.
- **AN10**: Validates a mix of letters and numbers within a 10-character limit.
- **D**: Enforces date formats without worrying about underlying database implementation.

Why Domain Values (e.g., MDMsystem) Are Important

Domain values ensure that data adheres to predefined rules or external standards:

- **MDMsystem**: Guarantees unique identifiers for entities like customers or products.
- **TransactionSystem**: Maintains the integrity of transactional data (e.g., invoices).
- **InventorySystem**: Aligns product data across inventory management and sales systems.



Overview of Layers

The diagram consists of four key layers, each responsible for distinct stages of the data flow:

1. **Source Data Layer** - Where the raw data originates.
2. **Processing Layer** - Where the data is transformed, validated, and processed.
3. **Storage Layer** - Where processed data is stored and managed.
4. **Utilization Layer** - Where the data is used to generate insights, drive decisions, or power tools.

Layer-wise Explanation

Source Data Layer

This is the entry point for all data into the system. It has two sources:

- **Direct Upload**: Files are uploaded directly into the system, providing raw data for further processing.
- **Manual Inputs (User Input Forms)**: Users manually enter data through forms, which is fed into the processing layer.

This layer ensures that data collection is flexible and accommodates both automated and manual inputs.

Processing Layer

This layer transforms raw data into structured, validated, and actionable formats:

- **Middleware**: Serves as an intermediary to validate and route the raw data to the appropriate processing pipelines or systems. It ensures data quality and conformity.
- **ETL Pipelines** (Extract, Transform, Load):
    - Extracts data from the source.
    - Transforms it into a usable format.
    - Loads it into the subsequent systems or storage.
- **Transaction Systems**: Manage and process transactional data, converting it into meaningful formats for downstream consumption.

The processing layer guarantees that only clean, structured data moves forward.

Storage Layer

This layer acts as the repository for all processed data, providing structured storage options:

- **Relational Databases**: Store structured, validated data, supporting queries and aggregations.
- **NoSQL Databases**: Handle unstructured or semi-structured data, enabling fast access to dynamic or hierarchical data.
- **Cloud Storage**: Stores large volumes of transformed data for scalable, long-term retention.

This layer ensures data is stored in the right format and structure for its intended use.

Utilization Layer

This layer enables end-users to derive insights and take actions using the processed data:

- **Dashboards**: Visualize aggregated insights for quick decision-making.
- **Marketing Tools**: Use segment data to personalize campaigns and enhance customer engagement.
- **Predictive Models**: Leverage processed data to forecast trends and guide strategic decisions.

The utilization layer closes the loop by turning raw data into actionable business insights.

Flow of Data

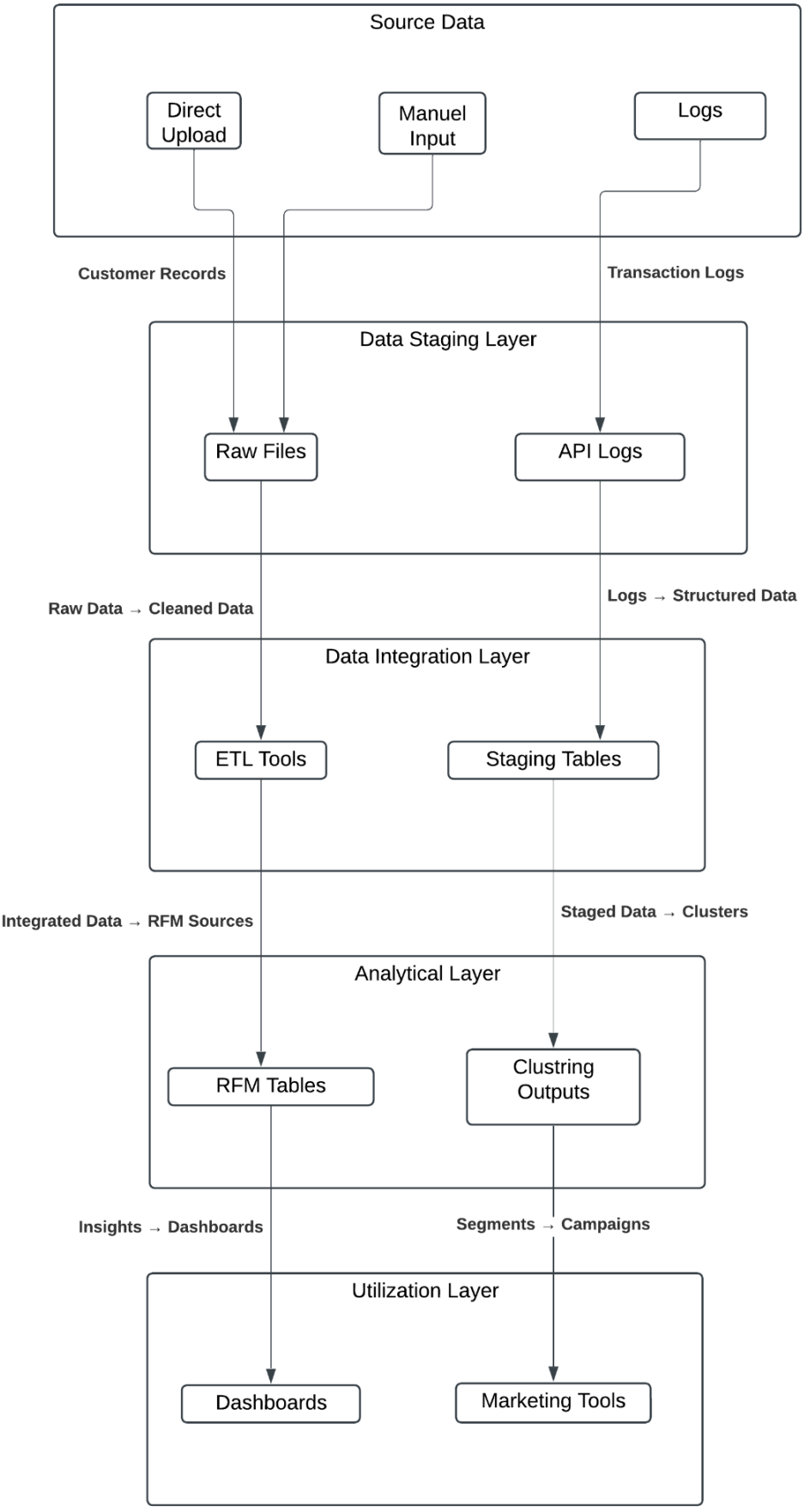
- Raw files or manual inputs are funneled from the **Source Data Layer** into the **Processing Layer**.

- The **Processing Layer** transforms and validates data, which is then forwarded to the **Storage Layer**.
- The **Storage Layer** organizes and retains data for use in the **Utilization Layer**, where the data powers tools and insights for business purposes.

Key Takeaways

- **Scalability**: The architecture supports diverse data sources and storage types.
- **Flexibility**: Allows both structured and unstructured data to coexist and flow seamlessly.
- **Actionability**: Ensures the end-users can derive real-world value from data through tools and insights.

This flow represents a modern, modular approach to data management, ensuring data integrity and utility throughout the process.


Source Data
Direct Upload
Manuel Input
Logs
Customer Records
Transaction Logs
Data Staging Layer
Raw Files
API Logs
Raw Data → Cleaned Data
Logs → Structured Data
Data Integration Layer
ETL Tools
Staging Tables
Integrated Data → RFM Sources
Staged Data → Clusters
Analytical Layer
RFM Tables
Clustring Outputs
Insights → Dashboards
Segments → Campaigns
Utilization Layer
Dashboards
Marketing Tools

Source Data Layer

The process begins with the collection of raw data from various sources. These include:

- Direct Upload: Automated submission of data files or system-generated inputs.
- Manual Input: Human-entered data, often through user interfaces or forms.
- Logs: Automatically generated transaction or activity logs that track system or user events.

At this stage, the data is categorized as:

- Customer Records: Data related to customer profiles, interactions, or behaviors.
- Transaction Logs: Detailed records of system or user activities, such as purchases or logins.

This raw data forms the starting point for the data flow.

Data Staging Layer

The staging layer is where raw data is temporarily stored and organized to prepare for further transformation. It includes:

- Raw Files: Unprocessed data files from direct uploads.
- API Logs: Structured log data generated by API interactions.

This stage ensures that data is collected and stored systematically, setting the stage for cleaning and integration in the next step.

Data Integration Layer

In this critical stage, the raw data is cleaned, transformed, and structured to ensure it is usable for analysis. The key components include:

- ETL Tools: Extract, transform, and load data by applying rules and logic to clean and standardize it.
- Staging Tables: Temporary storage for transformed data before it moves to the analytical layer.

This stage ensures data integrity and prepares it for deeper analysis and modeling.

Analytical Layer

Once integrated, the data is processed to extract actionable insights. This layer focuses on:

- RFM Tables: Segment customers based on recency, frequency, and monetary value metrics, helping to identify key behaviors.
- Clustering Outputs: Apply advanced analytics to group data into clusters, such as customer segments or behavioral patterns.

These analyses provide a foundation for targeted strategies and decision-making.

Utilization Layer

The final stage involves applying insights to achieve business goals and improve decision-making. The data is used for:

- Dashboards: Visualize insights, metrics, and trends for stakeholders in an accessible format.
- Marketing Tools: Use clustered data to design personalized campaigns, optimize customer engagement, and drive sales.

This layer closes the loop by ensuring that the refined data delivers actionable outcomes.

Data Management

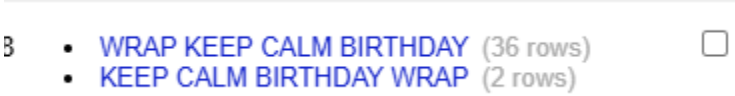
Data Inventory

Field	**Description**
Dataset Name	Online Retail Dataset
Originator	Kaggle
Data Entities	Customer, Invoice, Description, InvoiceLine
Format	CSV
Physical Structure	Tabular
Location	File: `/mnt/data/Online Retail (Complete).csv`
Number of Records	22,072

Changes	From 541,000 Rows to 22,072 Rows
Changes	Added 2 columns

Dataset Cleaning Process

- Changed all capital letters to first capital letter for easy reading.
- Deleted rows with blank content. A total of 1454 rows.
- Deleted rows with wrong content such as "?", "Lost" and "Missing", a total of 200 rows.
- Merge items with the same items but different names, such as Wrap Keep Calm Birthday and Keep Calm Birthday Wrap are obviously the same content, but with different names. All similar items are combined into a unified name, a total of 80.



Quality Assessment Report

Critical Data Elements

- **CustomerID**: Links customers to their transactions. Critical: Yes, personal information.
- **Invoice_No**: Uniquely identifies a transaction. Critical: Yes, uniquely identifies each transaction.
- **Stock_Code**: Represents unique products. Critical: Yes, identification of specific item purchased.
- **Quantity**: Indicates units sold. Critical: Yes, number of units sold per transaction.
- **Unit_Price**: Represents the price of products. Critical: Yes, represents the price per unit for each product.
- **Invoice_Date**: Timestamp for transactions. Critical: Yes, serves as the timestamp for transactions.
- **TotalPrice:** Total price of units sold

Criteria for “good” data

Completeness: No missing critical fields. Before data cleaning the dataset had 541,041 rows and had missing values, duplicate values and blanks. After cleaning the dataset had 22,072 rows. Deleted rows with blank content. A total of 1454 rows.

Consistency: The dataset is consistent with format for each field with valid relationships and no contradictory data .

Accuracy: Valid numerical values formatted consistently (e.g., Quantity > 0, Unit_Price > 0).

Timeliness: Transactions are correctly timestamped.

Uniqueness: No duplicate records.

Selection Criteria Documentation

Criteria

1. **Relevance**: Data must include customer purchase history, invoice details, and product information.
2. **Accuracy**: Data must have valid and accurate numerical and categorical fields (e.g., Quantity > 0).
3. **Completeness**: Critical fields such as CustomerID, Invoice_No, Stock_Code, Quantity, Unit_Price, and Invoice_Date must be populated.
4. **Timeliness**: Data should cover a meaningful period to capture customer lifecycle behaviors.
5. **Consistency**: No conflicting or duplicate entries within the dataset.
6. **Volume**: Data should contain enough records to ensure statistical validity of segmentation models.
7. **Compatibility**: Data format (CSV) must be processable with standard tools.

Ethical considerations

Based on our analysis and goals, there are some ethical considerations that we should be aware of during project implementation:

1. Privacy and Data Security

   The dataset contains personal information, such as CustomerID, which may involve personal identification.

   - Ensure compliance with GDPR data privacy laws.

- Implement anonymization or pseudonymization techniques to protect customer identities.
- Restrict access to sensitive data to internal personnel only.

2. Consent and Data Ownership
   Verify that customer data is collected with explicit consent for analysis and segmentation.
   - Inform customers how their data will be used and obtain consent for secondary purposes such as targeted marketing.
3. Fairness and Bias
   Using clustering or RFM analysis may inadvertently favor certain customer groups.
   - Regularly evaluate segmentation results to ensure that no group is unfairly excluded.
   - Use unbiased criteria for segmentation and evaluate the impact of recommendations on different demographics.
4. Accuracy and Representativeness
   Since the dataset is cleaned and only a small part of the large dataset is used for analysis, there is a risk that the analysis data will be inaccurate if invalid data points are omitted.
   - Ensure that the selected dataset is representative and does not have extreme values.
   - Validate data cleansing processes to ensure they do not introduce errors or unfairly remove valid records.
   - Clearly disclose any limitations or assumptions about the dataset in reports.
5. Marketing practices
   Personalized marketing based on segmentation may lead to intrusive advertising.
   - Ensure all advertising is ethical and avoids using sensitive words.
   - Ensure transparency to customers about why they receive certain promotions.
6. Reporting transparency
   Reports should accurately reflect findings and avoid overgeneralization or exaggeration.
   - Provide context for all insights, including potential sources of error.
7. Long-term impact
   Segmentation frameworks should maintain ongoing ethical monitoring as they adapt to changes in customer behavior.
   - Review and update segmentation criteria regularly to reflect long-term ethical and fair practices.
   - Address potential unintended consequences of dynamic marketing strategies, such as oversaturation of targeted promotions.

Lessons learned Reflection

Jacky Cheng

1. Understand the importance of data cleaning

I recognize the importance of data cleaning for data analysis. The original data set contains many problems, such as missing values, incorrect entries, and inconsistent naming conventions. Without data cleaning, analysis is impossible. In addition, too much data may lead to inconvenience during analysis. Selecting the most relevant and high-quality data is easier to manage and ensures faster and more efficient analysis. This is also my first time using OpenRefine, which is very helpful for data cleaning.

2. The role of customer segmentation

This is the first time I have conducted RMF analysis myself, using it to divide customers into different categories based on their purchasing behavior and preferences. I learned that it can effectively help companies allocate resources, make strategic decisions, and improve return on investment.

3. Challenges in dealing with large data sets

This is also the first time we have dealt with large data sets, and we need to carefully plan and develop a robust strategy. We must also ensure the validity and clarity of the data, while considering the impact of data volume on performance and efficiency. We must also make ethical considerations and learn lessons from them.

Kaan Tokmak

Through this project, I gained a deeper understanding of the critical processes involved in data management. I learned how data flows seamlessly from its source to the utilization layer, highlighting the significance of each stage, including data staging, integration, and analytical layers. The creation and analysis of diagrams, such as the application flow and data models, improved my ability to visualize complex systems and their interdependencies. Additionally, working on the data dictionary emphasized the importance of data governance and the need to define attributes, ensure consistency, and maintain data quality. Overall, this experience enhanced my knowledge of designing efficient data management frameworks and reinforced the value of structured approaches in handling organizational data.

Arsenii Popenko

Understood the importance of data inventory in identifying and organising key datasets which was the foundation for effective analysis. Data completeness and consistency significantly impact the quality of insights, reinforcing the need for cleaning and validation process. By focusing on critical data, I developed a more targeted approach to ensuring completeness, consistency, and accuracy in these fields.