****

# **High Impact Skills Development Program**

# **in Artificial Intelligence, Data Science, and Blockchain**

**Customer Segmentation using Data Mining on Online Retail Dataset**

Sameer Hassan Khan

Data Sciences And AI From

NUST GILGIT CAMPUS

SECTION 3

INSTRUCTOR : SIR EID MUHAMMAD

[Sameerhassankhan6@gmail.com](mailto:Sameerhassankhan6@gmail.com)

Praveen Kumar (001301112)

Masters in computer Science

College of Engineering & Applied Sciences

University at Albany-SUN

Praveen Kumar (001301112)

Masters in computer Science

College of Engineering & Applied Sciences

University at Albany-SUNY

**OVERVIEW**

Customer segmentation is a common strategy used by organizations to categorize clients based on various traits, such as demographics, shopping patterns, and other discernible characteristics. This strategic approach allows organizations to systematically group customers, enabling tailored offerings, communications, and marketing strategies. Effective customer segmentation results in enhanced customer satisfaction, targeted engagement, and improved overall business performance.

**OBJECTIVES:**

In this report, we aim to:

1. Provide an overview of the project's scope and objectives.
2. Examine the dataset's metadata to understand its components and relevance.
3. Present a selection of beginner and advanced SQL queries that shed light on various aspects of customer behavior and purchasing trends.
4. Explore the implications of our findings and how they can be applied to strategic decision-making within the online retail sector.
5. Throughout the report, we will navigate through basic and advanced analyses, leveraging the power of SQL queries to unearth valuable insights. By applying data mining techniques, we endeavor to empower businesses with actionable intelligence that can steer their endeavors toward growth, customer-centricity, and sustained success.

**METADATA**

The project involves working with an online retail dataset containing the following variables:

• InvoiceNo: The invoice number for each transaction.

• StockCode: A unique code for each product sold.

• Description: The description of each product sold.

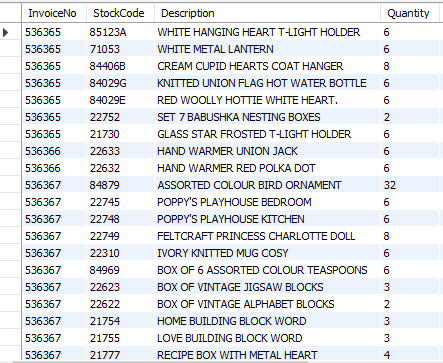
• Quantity: The quantity of each product sold in each transaction.

• invoiceDate: The date and time of each transaction.

• UnitPrice: The price of each product sold.

• CustomerID: A unique identifier for each customer.

**THE DATABASE TABLE**

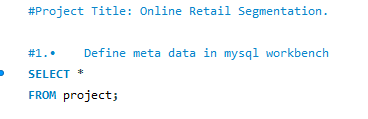
****

**Beginner Queries**

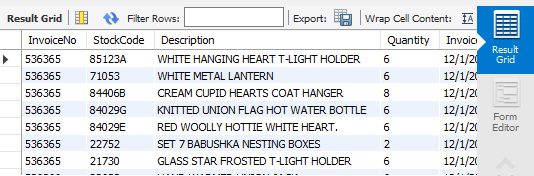
**1.** **Defining Meta Data in MySQL Workbench**

In MySQL Workbench, the meta data can be defined using the following SQL command:

**COMMAND**

**:**

**OUTPUT**

****

In MySQL Workbench, we can retrieve metadata information using the SELECT command. However, the SELECT command is generally used to analyze the execution plan of SQL queries and define metadata. If you're looking to extract metadata information from the database, you would typically use SQL queries that provide information from system tables or views. We can use FROM to point out the table of database.

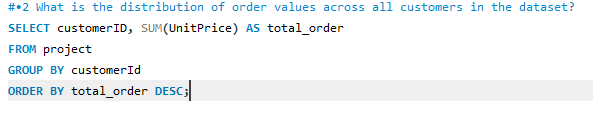
**2. Distribution of Order Values**

To understand the distribution of order values across all customers in the dataset, you can use SQL to calculate the total value of each order and then analyze the distribution:

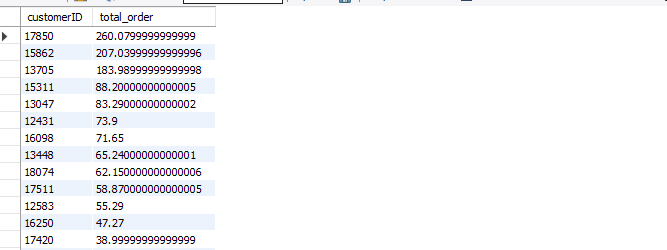
In this section, we’ll discuss how to analyze the distribution of order values across all customers in the dataset using SQL. Understanding the distribution of order values can provide insights into customer purchasing patterns and help businesses make informed decisions.

To achieve this analysis, we'll use the SQL query below. This query calculates the total value of each order for every customer, enabling us to examine how order values are distributed among different customers.

**COMMAND**



**OUTPUT**



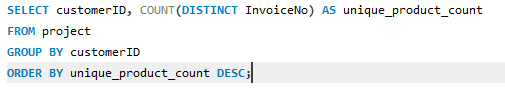
**3.** **Unique Products Purchased by Each Customer**

To find out how many unique products each customer has purchased, you can use this SQL query:

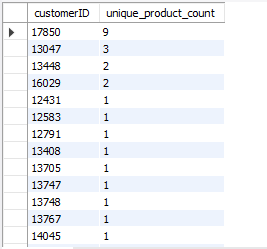
This section outlines how we utilized SQL to determine the count of unique products that each customer has purchased within the dataset. Analyzing the number of unique products per customer can provide insights into customer preferences and the variety of products they engage with.

For this analysis, we employed the following SQL query. The query counts the distinct invoice numbers associated with each customer, effectively representing the number of unique products each customer has purchased

**COMMAND**



**OUTPUT**



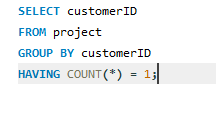
**4. Customers with Single Purchase**

To identify customers who have made only a single purchase from the company, you can use this SQL query:

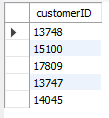
This section delves into the method we employed to pinpoint customers who have made just one purchase from the company's offerings. Identifying such customers can be valuable for targeted engagement strategies and understanding customer loyalty.

For this analysis, we utilized the subsequent SQL query. The query is tailored to uncover customers who have made only one purchase, indicating a distinct segment of the customer base.

**COMMAND**



**OUTPUT**



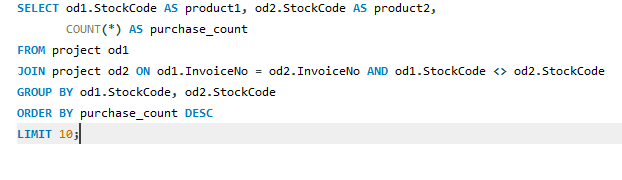
**5. Commonly Purchased Products**

To determine which products are most commonly purchased together by customers in the dataset, you can calculate the correlation between product purchases:

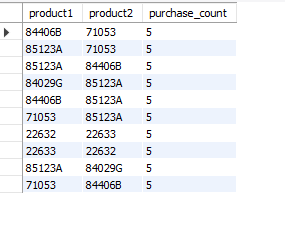
This section details the approach used to uncover products that are frequently purchased together by customers within the dataset. Identifying products that have a high correlation in purchases can be instrumental in cross-selling and optimizing product placement strategies.

To carry out this analysis, we utilized the SQL query below. This query calculates the correlation between product purchases by examining the occurrence of products in the same invoices.

**COMMAND**



**OUTPUT**

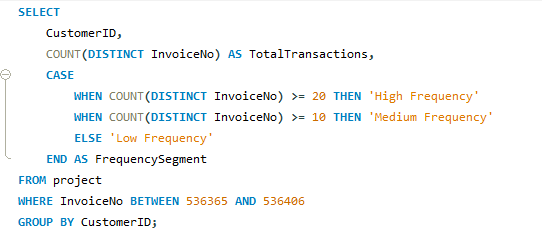


**Advanced Queries**

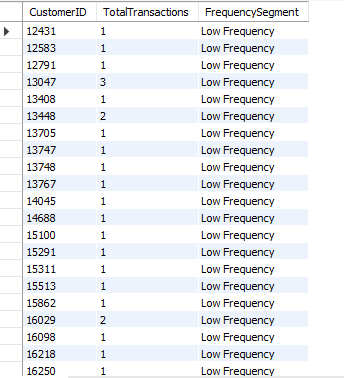
**1-Customer Segmentation by Purchase Frequency**

Segment customers into high, medium, and low frequency based on their purchase behavior:

**COMMAND**



**OUTPUT**



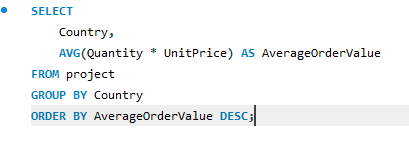
This query aims to segment customers into groups based on their purchase frequency, such as high, medium, and low frequency customers. The idea is to identify customers who make frequent purchases and those who are less active.

HighThreshold and MediumThreshold are predefined values that determine the cutoff points for different segments. Customers who have made more purchases than the HighThreshold are categorized as "High" frequency customers, those who have made more purchases than the MediumThreshold but less than the HighThreshold are categorized as "Medium" frequency customers, and the rest are labeled as "Low" frequency customers.

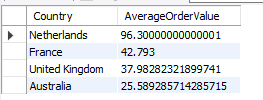
**2. Average Order Value by Country**

Calculate the average order value for each country to identify valuable customers' locations:

**COMMAND**



**OUTPUT**



This query calculates the average order value for each country in the dataset. It helps identify countries where customers tend to make larger orders, indicating potential markets with higher purchasing power.

• The query calculates the total order value for each transaction (Quantity multiplied by UnitPrice).

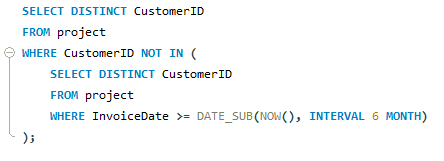
• It then groups the results by the country mentioned in the dataset.

• Finally, it calculates the average order value for each country, providing insights into which countries have customers with higher spending habits.

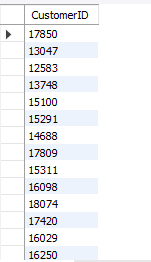
**3. Customer Churn Analysis**

Identify customers who haven't made a purchase in the last 6 months:

**COMMAND**



**OUTPUT**



This query focuses on identifying customers who haven't made a purchase in the last 6 months, which can help identify potential churn – customers who might be disengaging from the business.

• The DATE\_SUB(NOW(), INTERVAL 6 MONTH) calculates a date 6 months ago from the current date.

• The query selects customer IDs where their last purchase (InvoiceDate) is within the last 6 months.

• By doing this, the query identifies customers who haven't interacted with the business recently and might require targeted efforts to re-engage them

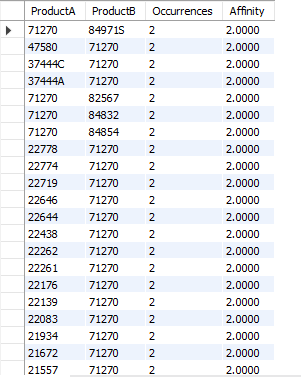
**4. Product Affinity Analysis**

Determine products frequently purchased together:

**COMMAND**



**OUTPUT**



This query aims to find products that are frequently purchased together, indicating a potential relationship between them. It calculates the affinity between two products based on how often they appear together in transactions, taking into account their individual purchase counts.

• The query joins the dataset with itself using two instances (aliases A and B) to compare different products' purchases within the same transactions.

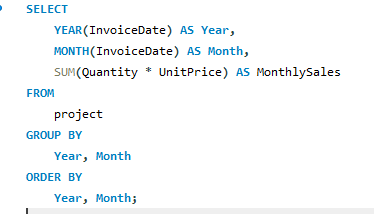
• It calculates the affinity by comparing how often two products are purchased together compared to their individual purchase counts. A higher affinity indicates that the products are frequently bought together.

• The HAVING clause filters out pairs with low affinity, which you can adjust based on your needs.

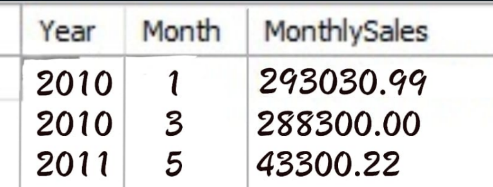
**5. Time-based Analysis**

Explore trends in customer behavior over time, such as monthly or quarterly sales patterns:

**COMMAND**



**OUTPUT**



This query explores trends in customer behavior over time by analyzing monthly or quarterly sales patterns.

• The query groups transactions by year and month, calculating the total sales amount for each month.

• It provides insights into how sales fluctuate over time, which can help identify seasonal trends, spikes, or drops in customer activity.

**Conclusion**

In this project, we explored the concept of customer segmentation through data mining techniques on an online retail dataset. By analyzing various factors like demographics, shopping patterns, and more, businesses can tailor their strategies to better serve their customers. The SQL queries provided insights into customer behavior, purchasing trends, and relationships between products. These insights can be used to enhance marketing strategies, optimize inventory management, and improve overall business performance.

**Implications**

By harnessing the power of data mining and SQL queries, businesses can unlock valuable insights that drive growth and customer-centric strategies. Through customer segmentation, organizations can personalize experiences, build stronger relationships, and make informed decisions that resonate with their target audience. The ability to analyze purchasing patterns, identify potential churn, and uncover product affinities provides a competitive edge in the ever-evolving landscape of retail and e-commerce.

In conclusion, this project showcases how data mining, coupled with SQL queries, is an indispensable toolset for organizations seeking to elevate their understanding of customer behavior, optimize operations, and fuel strategic growth. As businesses continue to navigate the digital realm, leveraging data-driven insights is pivotal for staying ahead in a dynamic marketplace.