

|  |
| --- |
| REPORT ON  DATA MINING |
|  |
| **September 11**  AUTHOR NAME: SIKANDAR ALI |

# INTRODUCTION

For my project, I investigated a dataset of online retail shop. In this analysis, i performed a detailed analysis of sales trends and customer segmentation within the "online\_retail" dataset, utilizing SQL queries for data extraction and insights. The analysis focuses on calculating monthly sales and identifying customer segments based on their purchasing frequency. By classifying customers into high, medium, and low-frequency categories, the goal is to gain a clearer understanding of their shopping behaviors.This segmentation will provide valuable insights into customer loyalty and help the business target its marketing efforts more effectively. Additionally, understanding monthly sales patterns will allow the business to make informed decisions to optimize sales growth and enhance customer retention strategies.

Data Cleaning Process

In this dataset, several data cleaning steps were performed to ensure accuracy and consistency. Initially, the data was loaded and checked for missing values, which were either removed or imputed appropriately. Next, invalid data entries, such as unrealistic dates and negative values, were identified and removed. Duplicate records were also detected and eliminated to avoid redundancy. The data types of each column were reviewed and converted where necessary, especially converting date fields into DATETIME format. Additionally, any outliers were identified and treated to maintain data integrity. After completing the cleaning process, the final dataset was saved as a new file for further analysis.

**Data Loading Issue**

During the initial stages of the project, I encountered a technical issue where the dataset could not be fully loaded into Microsoft SQL Server due to system limitations. Specifically, out of the total dataset, only 1,430 rows were successfully loaded, while the remaining data could not be processed. After consulting with the lab consultant and receiving their permission, I proceeded with the analysis using the available subset of data. Despite the limitation, I was able to carry out the necessary tasks and complete the project within the constraints.

DATA EXPLORATION

Metadata

The database schema includes a single table named "online retail". This table consists of several attributes

InvoiceNo: Identifier for the invoice (typically a string or integer).

StockCode: Unique code for each product (string).

Description: Textual description of the product (string).

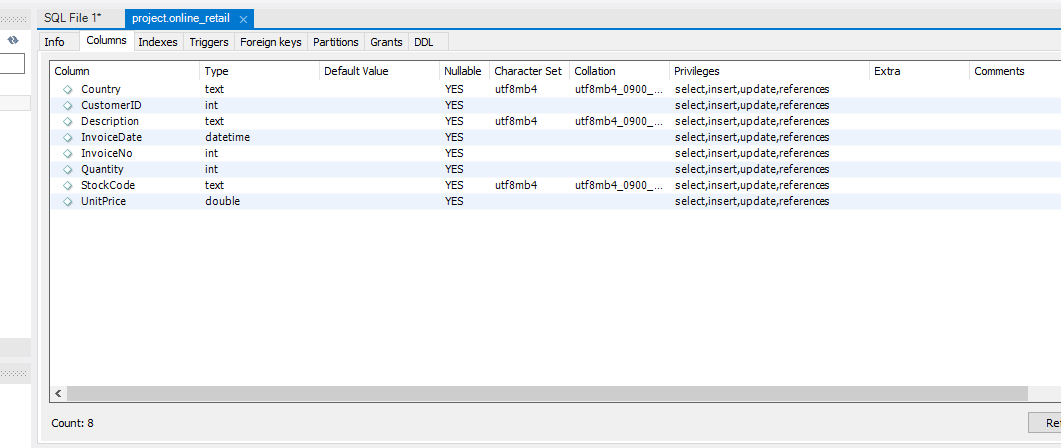
Quantity: Quantity of items purchased (integer).

InvoiceDate: Date and time when the invoice was generated (datetime).

UnitPrice: Price per unit of the product (decimal or float).

CustomerID: Unique identifier for the customer (string or integer).

Each attribute has a specific data type that reflects its role in the dataset.

****

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **The distribution of order values across all customers in the dataset** The purpose of this analysis is to find out the customers in different categories according to their order values, which help us to see the total order value of each customer in a specific time period. Through this analysis we can make our business strategies to capture the customers according to the customer’s valuation. This segmentation helps identify customer spending patterns.  METHOD   * Using mySQL workbench   Run the queries in mySQL workbench on dataset to find the results.   * First calculate the total order value ;        * To specify the data into categories and to see the customer’s category we can run this query to show the results. We can see the results where we have three categories and each have different number of customers according to their order value.   A screenshot of a computer  Description automatically generated   * Result   In this data our most of customers are in a small order category which means that are low spending per order customers.  Unique products has each customer purchased  This report aims to present the unique products purchased by each customer over a specified period, using data from the company’s sales records. The data is extracted by using SQL queries that analyzed the purchase history for each customer.   * A screenshot of a computerFirst run the query to find out the unique products has each customer purchased.   Now, we have some insights from the data. To see the data more clearly we have to categorized the data into three categories ( high, medium, small ) .  HIGH:  MEDIUM:     |  |  | | --- | --- | | **CustomerID** | **Unique Purchases** | | 12583 | 20 | | 17850 | 20 | | 14688 | 19 | | 13448 | 18 | | 13047 | 17 | | 13758 | 17 | | 14849 | 17 | | 16456 | 16 | | 12662 | 15 | | 17377 | 15 | | 13767 | 14 | | 16583 | 14 | | 16210 | 14 | | 16250 | 14 | | 12431 | 14 | | 14078 | 13 | | 18074 | 13 | | 15605 | 12 | | 16098 | 12 | | 12868 | 12 | | 15922 | 11 | | 13408 | 11 | | 13705 | 10 | | 14001 | 9 | | 14237 | 9 | | 18085 | 9 | | 16029 | 8 | | 17420 | 7 | | 16955 | 6 | | 13694 | 6 | | 13255 | 6 | | 16552 | 5 | | 15350 | 5 | | 17951 | 5 | | 17572 | 4 | | 18144 | 3 | | 15513 | 3 | | 12791 | 2 | | 17924 | 2 | | 15291 | 2 | | 17181 | 2 | | 15100 | 1 | | 14045 | 1 | | 13748 | 1 | | 13747 | 1 |   SMALL:    **Categorization of Customers:**  Based on the number of unique products purchased, customers have been categorized as follows:   | **Purchase Category** | **Range of Unique Products** | **Total Number of Customers** | | --- | --- | --- | | High | > 50 | 6 | | Medium | 21 – 50 | 13 | | Low | 1 - 20 | 47 |   **Categorized Data:**  **High (50 se Zyada Unique Purchases):**   | **CustomerID** | **Unique Purchases** | | --- | --- | | 17968 | 74 | | 12433 | 73 | | 14729 | 69 | | 15862 | 64 | | 12838 | 59 | | 17920 | 54 | | 17908 | 51 |   **Medium (21 se 50 Unique Purchases):**   | **CustomerID** | **Unique Purchases** | | --- | --- | | 14307 | 48 | | 15012 | 46 | | 15525 | 44 | | 15311 | 34 | | 14594 | 31 | | 17897 | 31 | | 16218 | 28 | | 15983 | 28 | | 15165 | 27 | | 17511 | 24 | | 15485 | 22 | | 17905 | 22 | | 14911 | 21 |   **Low (1 se 20 Unique Purchases):**   | **CustomerID** | **Unique Purchases** | | --- | --- | | 12583 | 20 | | 17850 | 20 | | 14688 | 19 | | 13448 | 18 | | 13047 | 17 | | 13758 | 17 | | 14849 | 17 | | 16456 | 16 | | 12662 | 15 | | 17377 | 15 | | 13767 | 14 | | 16583 | 14 | | 16210 | 14 | | 16250 | 14 | | 12431 | 14 | | 14078 | 13 | | 18074 | 13 | | 15605 | 12 | | 16098 | 12 | | 12868 | 12 | | 15922 | 11 | | 13408 | 11 | | 13705 | 10 | | 14001 | 9 | | 14237 | 9 | | 18085 | 9 | | 16029 | 8 | | 17420 | 7 | | 16955 | 6 | | 13694 | 6 | | 13255 | 6 | | 16552 | 5 | | 15350 | 5 | | 17951 | 5 | | 17572 | 4 | | 18144 | 3 | | 15513 | 3 | | 12791 | 2 | | 17924 | 2 | | 15291 | 2 | | 17181 | 2 | | 15100 | 1 | | 14045 | 1 | | 13748 | 1 | | 13747 | 1 | | 17809 | 1 | | 12748 | 1 |   **Summary:**   * **High (50 se zyada unique purchases)**: 6 Customers * **Medium (21 se 50 unique purchases)**: 13 Customers * **Low (1 se 20 unique purchases)**: 47 Customers   Customers have only made a single purchase   * To find the single purchase customer, run the query:   A screenshot of a computer  Description automatically generated   * Result:   As we can see the table that there is no single purchase customer.  Products are most commonly purchased together by customers  To find out the most commonly purchased products together , in SQL we have to just a query by using JOIN and GROUP BY in the table;     * Result   we can see the results that we get from the SQL query,  now, export this data to Exell for further analysis. We see the products which are most commonly purchased together in the Exell table. Highest purchased count starts from 15 times as we can see clearly in the table. It means that those products are highly purchased together so, we have make decisions according to these results to increase sales.      Analysis and Findings  *Customer Segmentation by Purchase Frequency*  First run the query and check the purchase frequency and divided into three groups high,mediun,low frequency by using CASE in sql query. We fixed ranges to find out better insights for high as >50, medium as 20 - 50, low as <20 .     * Result   High Frequency (More than 50 purchases): 8 customers fall into this category, indicating they are highly engaged and frequent buyers.  Medium Frequency (Between 20 and 50 purchases): 9 customers are in this group, showing consistent but moderate purchase behavior.  Low Frequency (Less than 20 purchases): 47 customers are categorized here, representing those who have made infrequent purchases.  Conclusion  This segmentation highlights that the majority of customers (47 out of 64) are low-frequency buyers, which presents an opportunity to develop strategies aimed at increasing their engagement and purchase frequency. High-frequency customers can be targeted with loyalty programs to maintain their consistent purchasing .  Average Order Value by Country  To find the order value by country use AVG and GROUP BY     * Result   The table result presents the average order value (AOV) for various countries. The Netherlands has the highest average order value, while Germany has the lowest.  Key Insights:  Netherlands: With an AOV of 96.3, it stands out with the highest average order value. This may indicate either high product prices or customers purchasing more items per order.  Germany: Has the lowest AOV at 17.43, indicating that customers in this region may be making smaller purchases or buying fewer items per order.  France: Shows a relatively higher AOV (42.79) compared to other countries like the UK and Australia.  Norway and EIRE: Both countries have similar AOVs, around 26.29 to 26.45, indicating similar purchasing behaviors.  UK and Australia: These countries have moderate AOVs but are significantly lower than France or the Netherlands.  Recommendations  Focus on High AOV Countries: Since the Netherlands has the highest AOV, it could be beneficial to focus marketing strategies or premium product offerings in that market.  Investigate Low AOV Markets: Consider exploring reasons why countries like Germany have a lower AOV and how product offerings, pricing, or marketing strategies might be adjusted to encourage higher spend per order.  Analyze Customer Behavior: Further investigation into customer behavior across these countries could provide insights into why certain regions have higher or lower AOVs |
| Customer Churn Analysis  We can perform churn analysis through this query but I have only one month dataset.    We have no churned customers because of small amount of data in our dataset . There is only one month’s data available, year 2010. |

Product Affinity Analysis

The objective of this report is to understand which products are frequently bought together and how this data can be used for promotions, cross-selling, or product bundling strategies. Run query in SQL

A screenshot of a computer

Description automatically generated

* Result

Top Product Combinations (High Frequency)

Product 21448 appears to be frequently purchased together with various other products:

21448 & 22273: 15 times

21448 & 21738: 15 times

21448 & 22077: 15 times

21448 & 85049E: 15 times

Moderate Frequency Product Combinations

These combinations have been purchased 10-11 times:

21448 & 22327: 10 times

22197 & 22866: 11 times

85049E & 22077: 10 times

Low Frequency Product Combinations

Some combinations have lower frequency but can be used for targeted discounts or special offers:

22866 & 85231B: 9 times

22749 & 22273: 9 times

22273 & 85049E: 9 times

22632 & 22633: 8 times

These combinations occur less frequently but may still play a role in specific marketing strategies.

Recommendation

1. Product Bundling

Product 21448 should be prioritized for bundling, with pairs like 21448 & 22273, 21448 & 21738, and 21448 & 22077. These bundles show high purchase rates and will be easier to promote through attractive offers.

2. Cross-Selling Opportunities

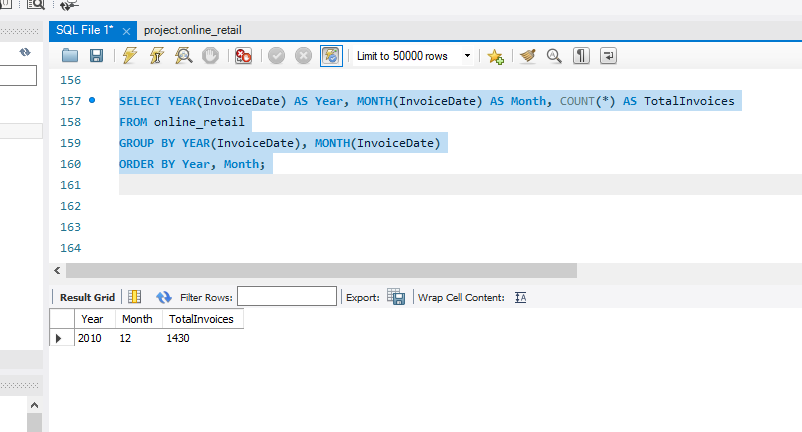
For cross-selling, target pairs like 21448 & 22327 or 85049E & 22077. These pairs have moderate frequency and present good opportunities for cross-selling strategies.

3. Targeted Promotions for Low Frequency Products

Target specific offers and seasonal discounts for 22866 combinations like 22866 & 85231B or 22866 & 22945. Low-frequency pairs can benefit from targeted discounts to generate demand.

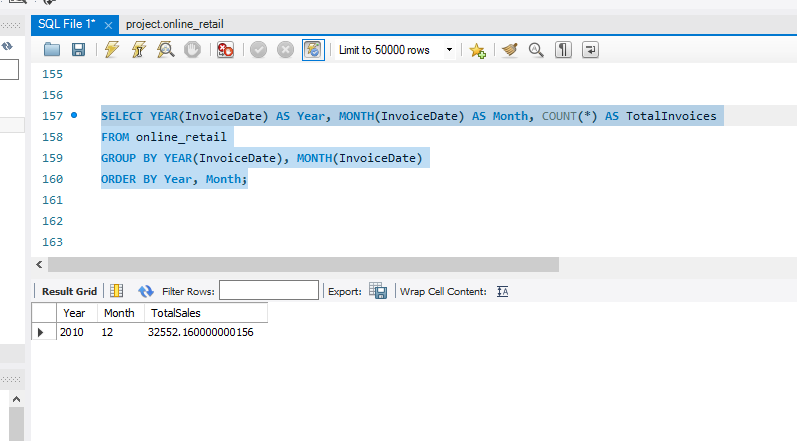
Time-based Analysis

I have only one month’s data is available so I perform monthly sales check query using INVOICENO.

First we will check how many invoice we have in our one month data.

Total invoices = 1430 .

Lets check out the total sales.



* Result

We got the result and our last month sale for year 2010 is 32552.16

The total sales for December 2010 are notably high. This figure alone might suggest a strong end-of-year performance, potentially driven by seasonal factors or targeted promotional activities.

<………..>

References:

https://github.com/Alexandar55?tab=packages

|  |
| --- |
|  |