**CUSTOMER GROUP ANALYSIS REPORT AT AN ONLINE SOUVENIR STORE**

**Topic: Customer Relationship Management**

**Owner : DANG Trung Kien**

**Hà Nội – 2023**

**TABLE OF CONTENTS**

[**1. Introduction** 1](#_Toc161591158)

[**2. Data preparation** 2](#_Toc161591159)

[**2.1 Data set description** 2](#_Toc161591160)

[**2.2 Data cleaning** 4](#_Toc161591161)

[**3. Building a data warehouse.** 9](#_Toc161591162)

[**3.1 Identifying data units.** 9](#_Toc161591163)

[**3.2 Dimensions** 9](#_Toc161591164)

[**3.3 Design Fact table** 12](#_Toc161591165)

[**3.4 Relationships between Dimension tables and Fact table** 13](#_Toc161591166)

[**3.5 Import data into the data warehouse.** 13](#_Toc161591167)

[**4. Visualize and explore the dataset.** 14](#_Toc161591168)

[**4.1 Connect the data warehouse to PowerBI** 14](#_Toc161591169)

[**4.2 Data Analysis** 15](#_Toc161591170)

[**4.2.1 Overall Analysis** 15](#_Toc161591171)

[**4.2.2 Detailed analysis focusing on customers** 16](#_Toc161591172)

[**5. Conclusion** 24](#_Toc161591173)

[**Data set used in the report** 25](#_Toc161591174)

[**References** 25](#_Toc161591175)

# **1.** **Introduction**

Customer Relationship Management (CRM) is a technology-based business strategy that focuses on customer relationships. It helps businesses increase customer interaction, improve service quality, and increase sales.

CRM includes a system for customer information management, customer relationship management, and customer interaction management. The customer information system (CIS) stores customer information such as name, address, phone number, email, purchase history, and interactive activities, helping businesses understand more about their customers. Customer Relationship Management (CRM) includes strategies for retaining customers, building long-term relationships, and meeting customer needs. Customer Interaction Management (CIM) helps businesses interact with customers through many different channels such as email, phone, online chat, and social networks, helping to increase interaction and enhance customer experience.

CRM applications are very diverse. Some outstanding applications include Sales Management, Customer Service Management, Marketing, Campaign Management, and Quality Management.

**CRM needs:** CRM needs are often related to optimizing the customer relationship management process and improving service quality to meet customer needs. Specifically, some CRM needs include:

* **Enhance customer interaction:** Businesses must improve their ability to interact with customers across many channels to increase customer engagement and satisfaction.
* **Develop effective marketing strategies:** Businesses must develop marketing strategies that suit customer needs to increase interaction and attract customers.
* **Improve the ability to manage customer information:** Businesses need to professionally manage customer information to optimize the customer relationship management process.
* **Optimize sales process:** Businesses must optimize sales processes to enhance business efficiency and increase sales.
* **Marketing campaign management:** Businesses must manage marketing campaigns to achieve the highest business efficiency.
* **Enhance the ability to monitor and evaluate the effectiveness of business activities**: Businesses must monitor and evaluate the effectiveness of business activities to optimize business strategies and increase business efficiency.

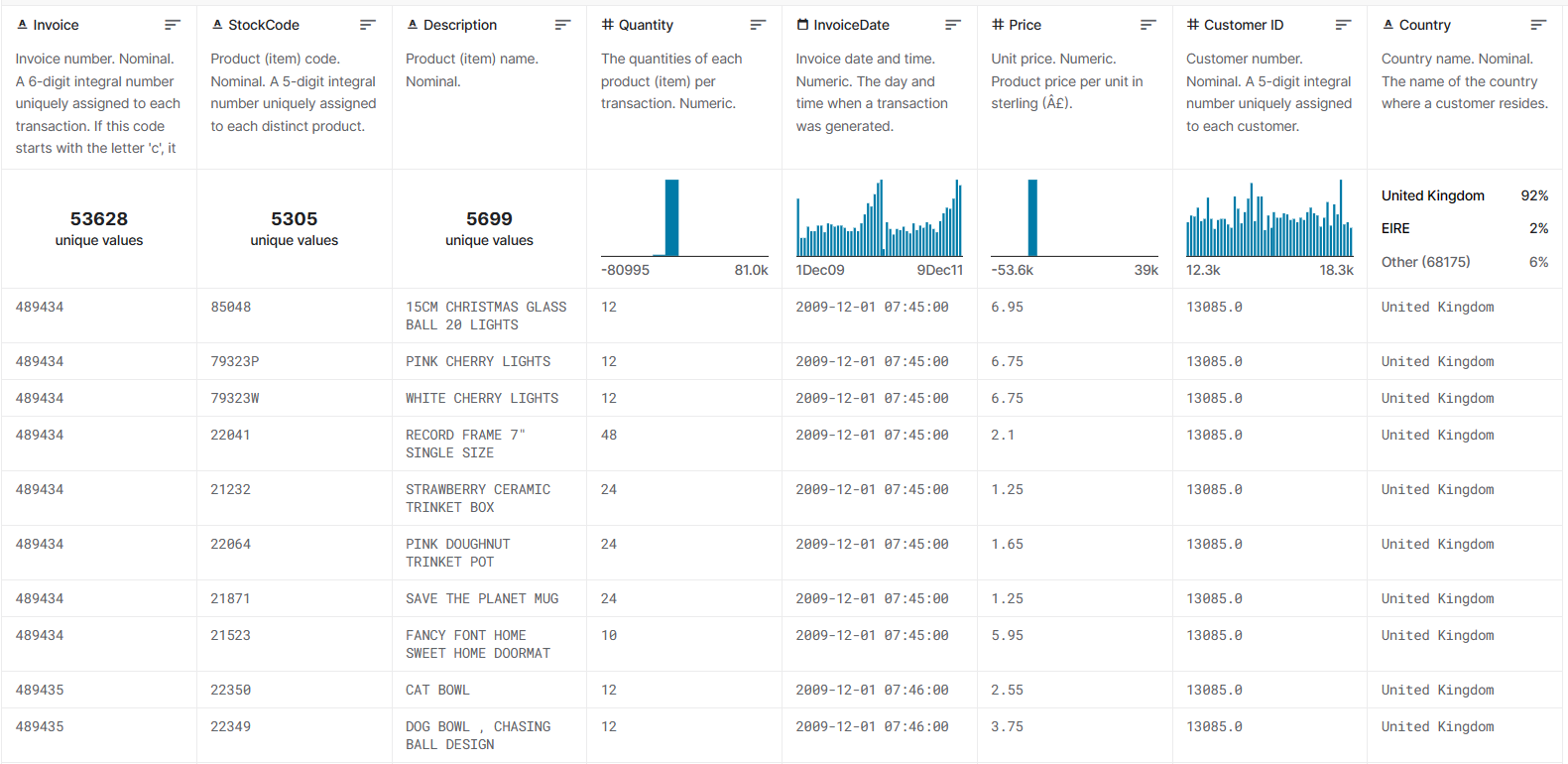
With customer grouping combined with available information in the data set, questions are asked to understand better customer segments and their habits in the data set, such as:

* Who are the store's main customer groups? Loyal customers, new customers, or other customers?
* Where do those customers come from? Where do new and loyal customers often come from?
* What are those customers looking for in the store? What customer groups usually buy from the store? Are there any differences between groups?
* What are their shopping habits like? Do they tend to buy on particular days, hours, or months?

# **2. Data preparation**

## **2.1 Data set description**

The data in the project is taken from the Online Retail II UCI dataset provided on Kaggle by author MIYABON. Retail II UCI is a data set containing customer transaction information at an online retail store in the UK from December 1, 2009 to December 9, 2011, with products mainly being gifts and customers that may include wholesalers.



***Figure 1****: Overview of the dataset*

**Invoice**:  A 6-digit integral number uniquely assigned to each transaction.

**StockCode**: A 5-digit integral number uniquely assigned to each distinct product.

**Description**: Product (item) name

**InvoiceDate**: The day and time when a transaction was generated.

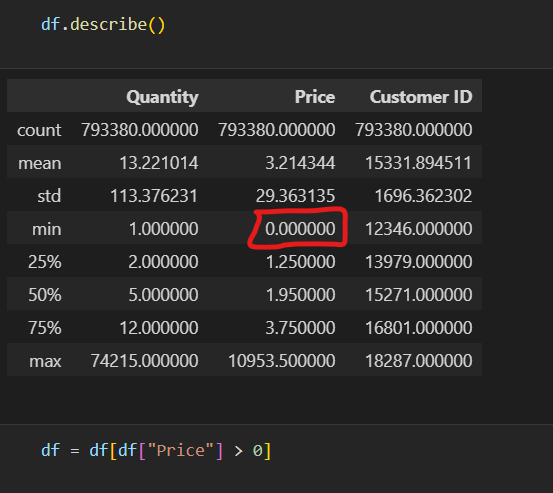
**Price**: price of a respective product

**CustomerID**:  A 5-digit integral number uniquely assigned to each customer.

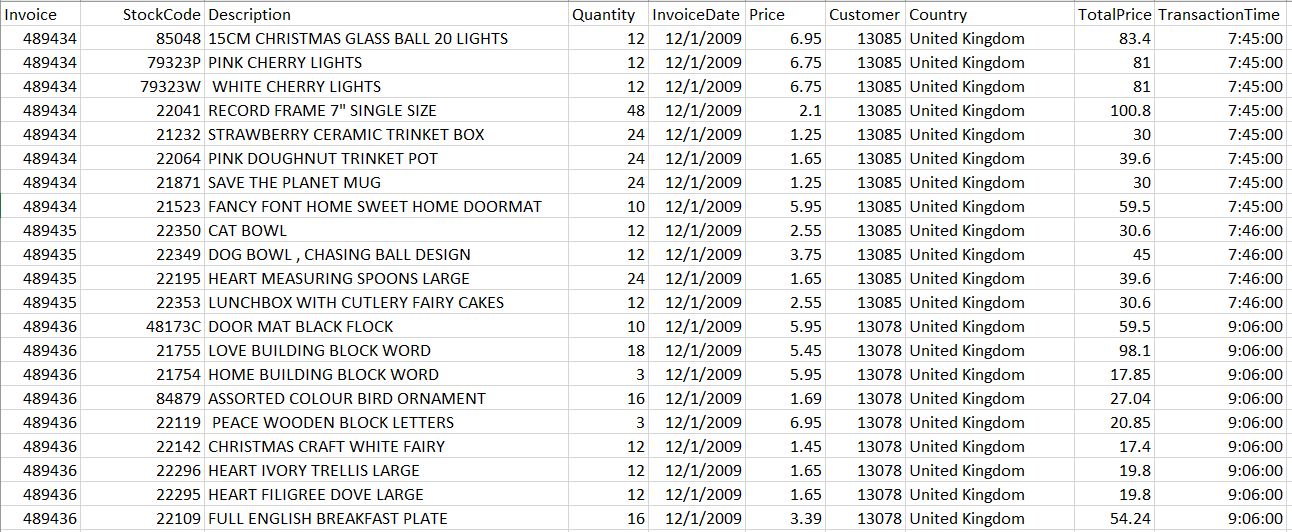
**Country**: The name of the country where a customer resides.

## **2.2 Data cleaning**

Python was used to eliminate NaN data and noise from the dataset. In the figure below, products with a price of 0 can be observed, which may affect subsequent calculations. Therefore, these products will be excluded.



***Figure 2****: Removing noisy data.*

Afterward, add a column TotalPrice (Quantity \* Price), and split the InvoiceDate column from datetime format into separate date and time formats. Once the data processing with Python is completed, export the data to a flat-file format for further processing.

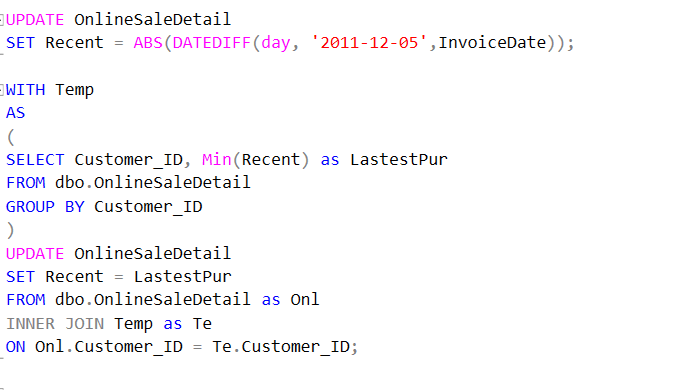
***Figure 3****: Overview of the data after removing outliers with Python.*

Create a database to store raw data, and proceed to add data as a table within the database. Add columns Frequency, Recent, FrequencyRating, RecentRating, MonetaryRating, and RFM\_Score to serve customer segmentation and RFM analysis purposes. Specifically:

**Frequency**: The number of times each user makes purchases at the store.

**Recent**: The number of days since the most recent purchase made by each user at the store.

**MonetaryRating**: Ranking based on the total amount of money spent by a customer.

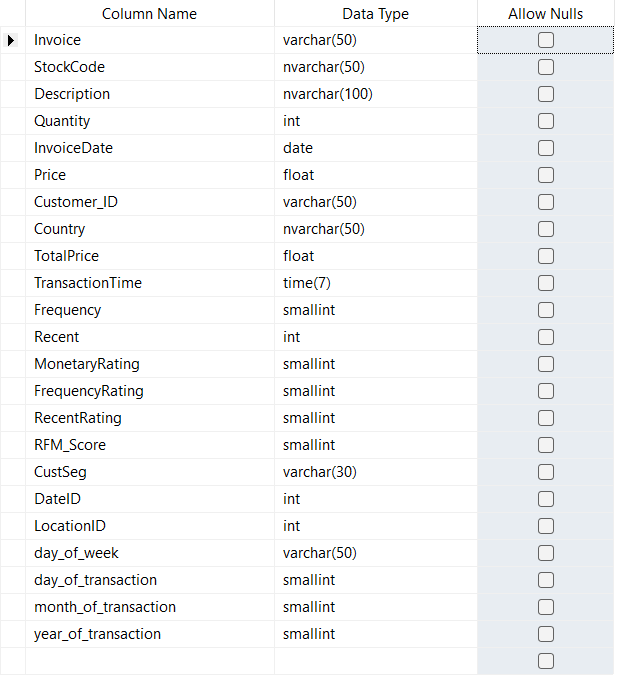


***Figure 4****: Update the data for the Recent column.*

After obtaining the aforementioned columns, proceed to calculate the RFM scores and classify customers into different customer segments.

|  |  |
| --- | --- |
| VIP | The group of customers who have made recent purchases, purchase frequently, and spend the most. |
| Loyal | The group of customers who spend a considerable amount of money and make purchases frequently. |
| Potential | The group of customers who have made recent purchases, spending a significant amount of money, and making more than one purchase. |
| Recent | Recent purchasers but not frequent buyers. |
| Promising | Recent purchasers but not high spenders. |
| Need Attention | Customers with an average RFM score. |
| Sleep | Customers below average RFM score. |
| At Risk | Customers who have made multiple purchases with a significant amount spent in the past |
| Can’t lose | Customers who spend a large amount of money frequently but have yet to make purchases recently. |
| Hibernating | Customers who have not made purchases for a long time, spend little, and do not make frequent purchases. |
| Lost | Customers with lowest RFM score. |

***Table 1****: Customer segmentation*

After segmenting customers, proceed to break down the order creation date into day, month, year, and day of the week, and add an ID for the date and location to facilitate the design of the data warehouse. However, with the same product ID, there are multiple different descriptions, so afterward, the product IDs will be reassigned based on each unique description. Finally, we will compile the cleaned dataset ready for the design of the data warehouse with a structure, as shown in Figure 5. 

***Figure 5****: The structure of the dataset after adjustments.*

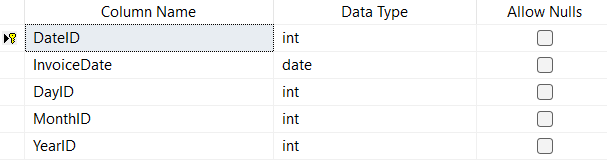
# **3. Building a data warehouse.**

## **3.1 Identifying data units.**

The Fact table stores information about customer transactions with specific products. Each row in the Fact table corresponds to a customer's transaction for a specific product at the store.

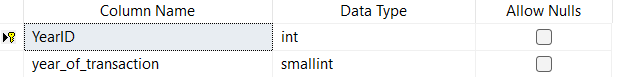
## **3.2 Dimensions**

**Time dimension(dim\_date)**: provides detailed information about the time the transaction was performed such as: day of the week, day, month, year.



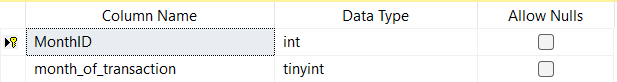
***Figure 6****: Time dimension structure*

**Dim year(dim\_year)**: Contains information about the year of transaction



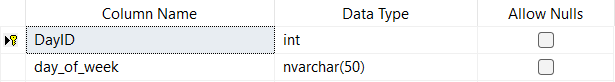
***Figure 7****: Structure of transaction year dimension*

**Dim month(dim\_month)**: contains information about the month in which the transaction was made



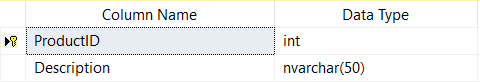
***Figure 8****: Structure of the transaction month*

**Dim\_day(dim\_day)**: contains information about the transaction day of the week

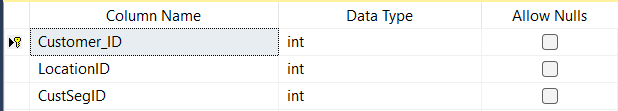


***Figure 9****: Structure of the transaction month*

**Product dimension(dim\_product)**: Contains information related to a certain product such as productID, description.

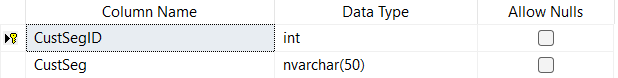
***Figure 10****: Product dimensional structure*

**Customer dimension(dim\_customer)**: Contains information related to a customer.

******

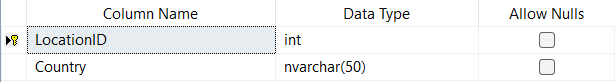
***Figure 11****: Customer dimension structure*

**Customer classification dimension (dim\_customer\_segmentation)**: Contains information related to a customer's classification calculated using the RFM index.

******

***Figure 12****: Dimensional structure of customer classification*

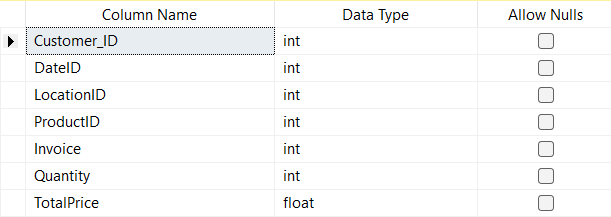
**Dimension of customer location(dim\_location)**: Contains information related to a customer's classification calculated using the RFM index.

******

***Figure 13****: Customer location dimensional structure*

## **3.3 Design Fact table**

The Fact table will store information about each customer's transaction with each specific product, including full details about the time, quantity, and total value of the order.



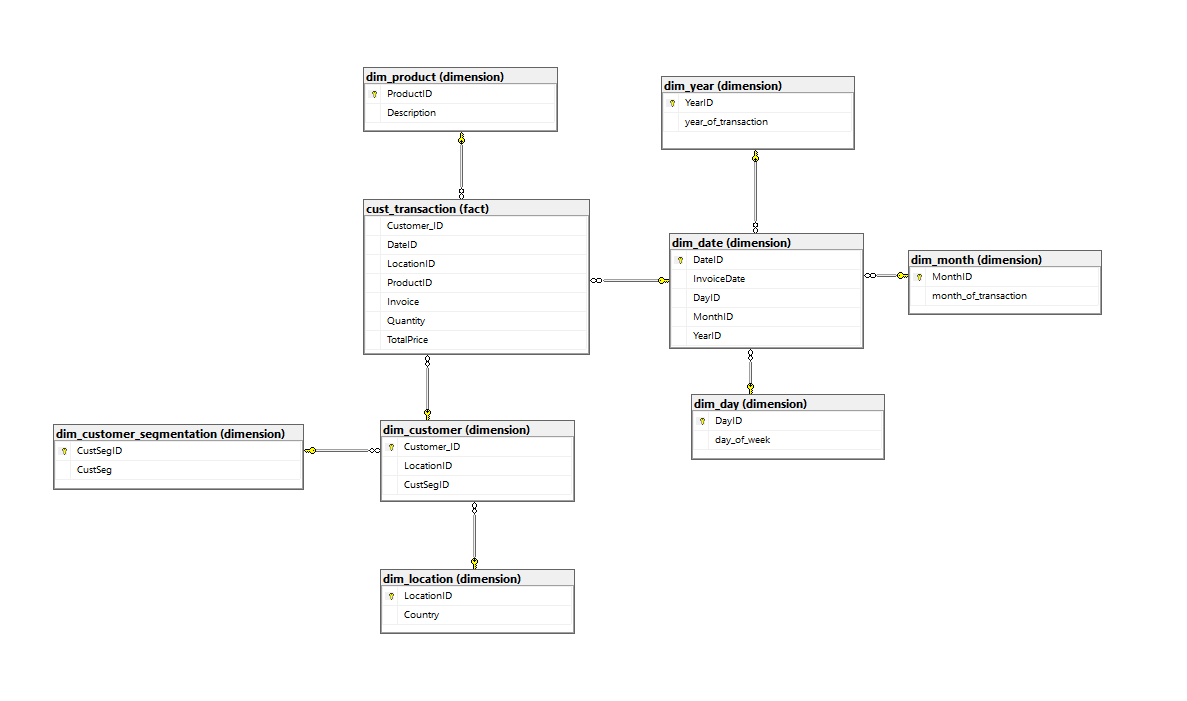
***Figure 14****: Fact table structure*

**Invoice**: Invoice number of the transaction

**Quantity**: Number of products purchased

**TotalPrice**: Price of the corresponding product quantity

## **3.4 Relationships between Dimension tables and Fact table**



***Figure 15****: Relationship between Fact table and Dimension tables*

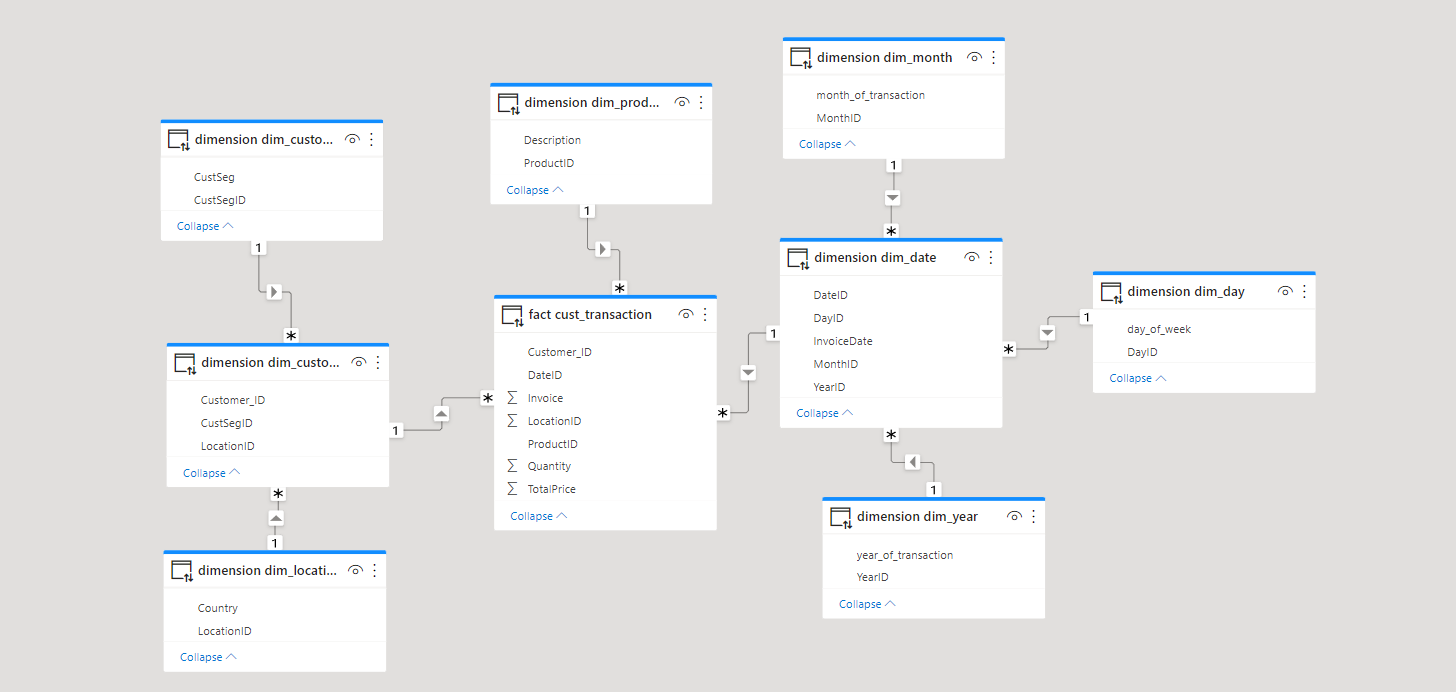
## **3.5 Import data into the data warehouse.**

After building the data warehouse, query the necessary information in the database containing the pre-processed data, export it to flat files, and load it into corresponding tables and fields in the data warehouse. For convenience in management, it is essential to reformat some data types in certain columns. For example, converting customer\_ID from nvarchar to int may be necessary.

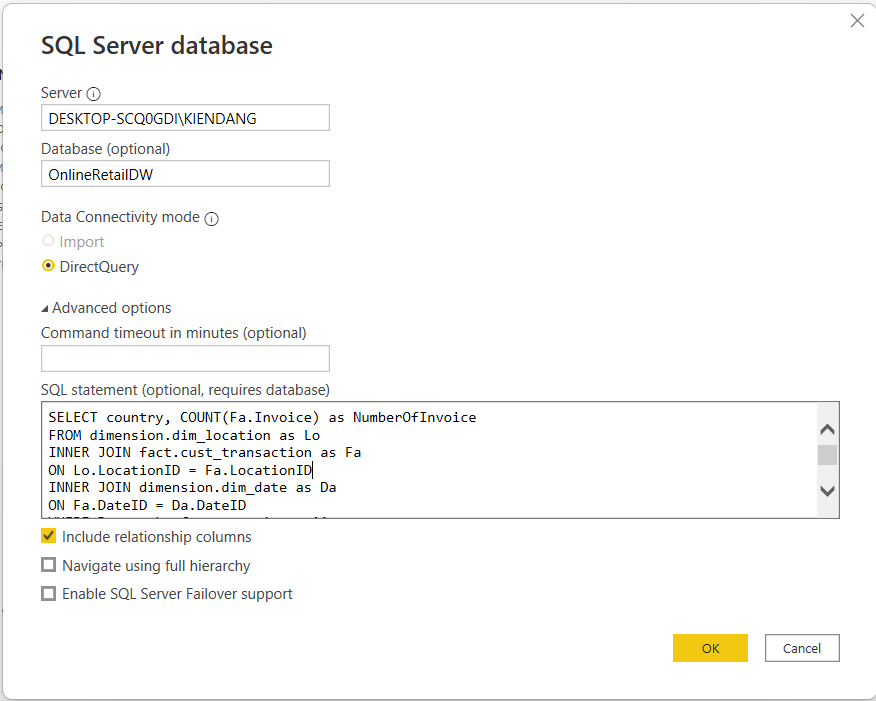
# **4. Visualize and explore the dataset.**

## **4.1 Connect the data warehouse to PowerBI**

Select SQL Server from the toolbar of PowerBI. Proceed to enter the server name and data warehouse name, then select DirectQuery. Choosing DirectQuery will execute queries to retrieve data directly from the data warehouse instead of storing it in memory on PowerBI as with import. Upon successfully connecting the data source and PowerBI, the model section will display the selected tables and their relationships.



***Figure 16****: Relationship between tables after connecting to data source*

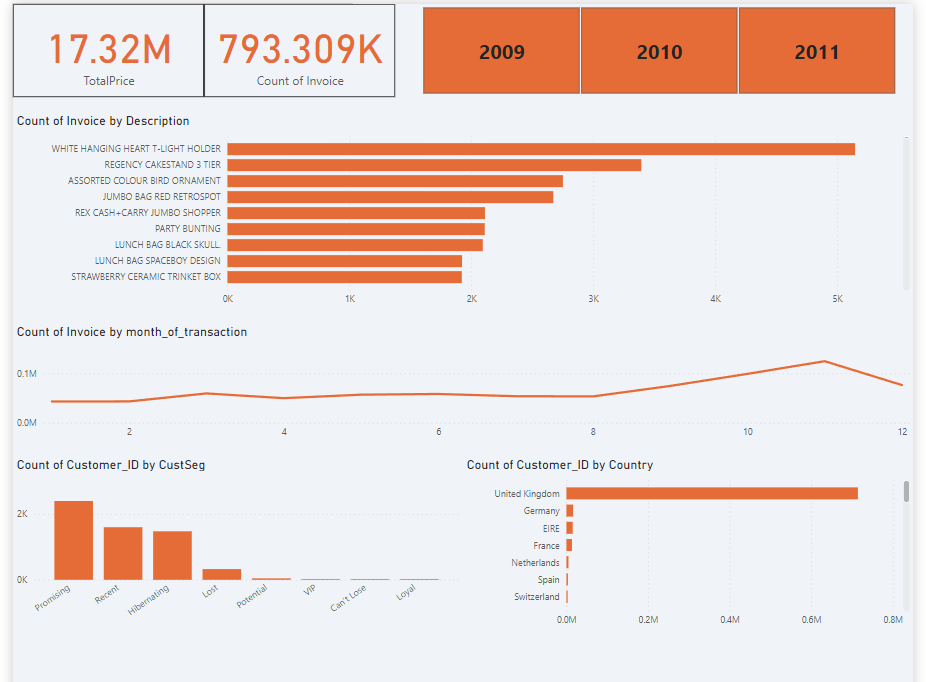
In addition to default querying, PowerBI provides the ability to query on-demand to interact with the data warehouse according to user requests. 

***Figure 17****: Execute queries on PowerBI*

## **4.2 Data Analysis**

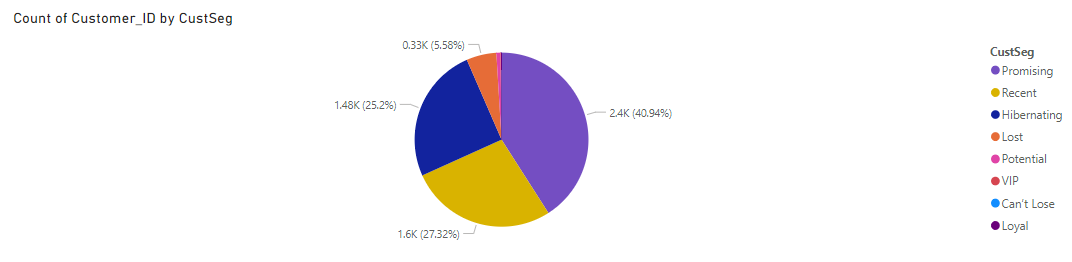
### **4.2.1 Overall Analysis**

Before diving into detailed analysis, let's take an overview of the dataset by combining the information provided in the dataset with additional insights derived from customer segmentation.

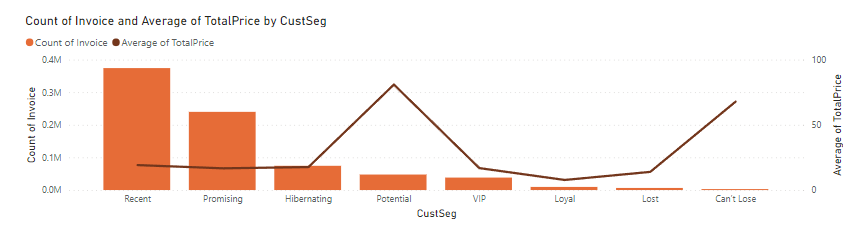
***Figure 18****: Overview analysis of the dataset*

The report indicates that over two yearsfrom December 2009 to December 2011, the store processed 793,309 orders, generating a total revenue of £17.32 million. The best-selling product was the 'White Hanging Heart T-Light Holder' with 5,147 orders placed. New customer segment accounts for a significant portion, predominantly from the United Kingdom. Based on the monthly order chart, it is observed that towards the end of the year, there is an increasing trend in customer purchases, with two sudden spikes in March and a significant surge from September to November, reaching its peak in November.

### **4.2.2 Detailed analysis focusing on customers**

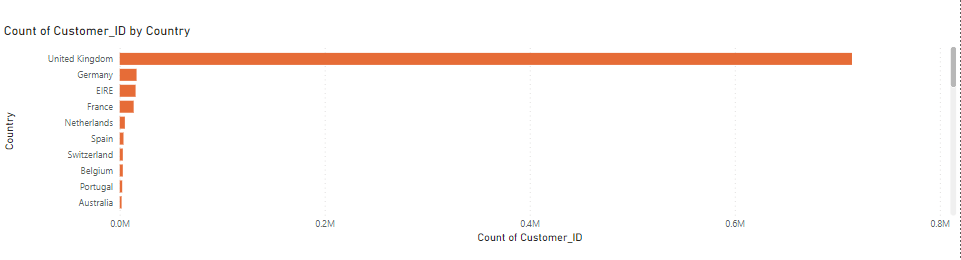
******

***Figure 19****: Customer group distribution*

With the above chart, the 5 major customer segments are Promising, Recent, Hibernating, Lost, and Potential, with most of them being customers who have either churned or have not made purchases for a long time. Only Promising, Potential, and Recent segments have recently interacted with the store. Therefore, it can be concluded that the primary customer base at the store consists mainly of new customers.

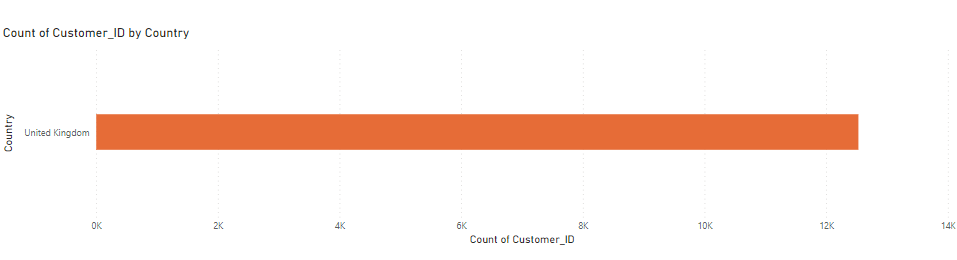
***Figure 20****: Number of purchases and average order value of each customer group*

With Figure 20, for new customer segments like Recent and Promising, they make a significant number of purchases; however, their average order value is relatively low. On the other hand, segments like Potential and Can't Lose make fewer purchases, especially the Can't Lose segment, which has the lowest purchase frequency among all segments, but the average order value of these two segments is much higher compared to other segments.

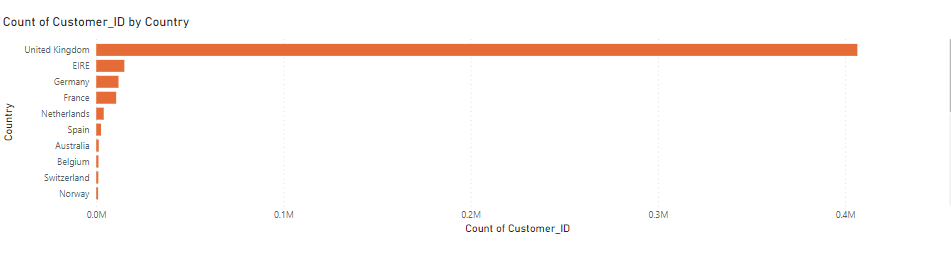
**

***Figure 21****: Distribution of customers across countries*

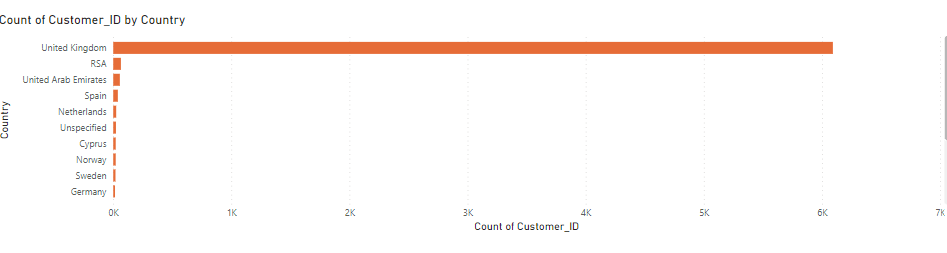
In Figure 21, the majority of customers come from the United Kingdom, with some customers from various other countries, although the number is not significant. However, this distribution will have certain differences for each specific customer segment.



***Figure 22****: Country distribution of Can't lose and Loyal customer groups*

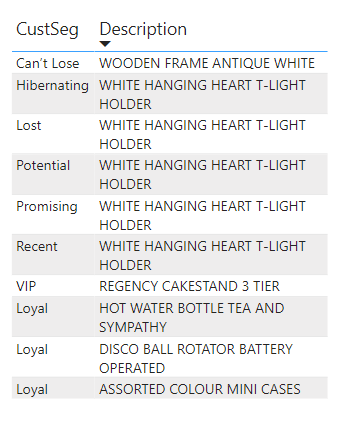
**

***Figure 23****: Country distribution of VIP, Potential and Recent customer groups*

**

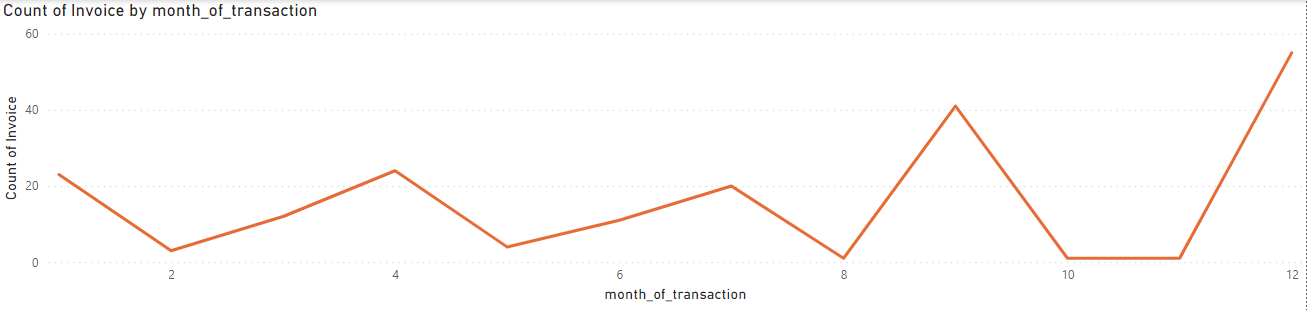
***Figure 24****: Country distribution of Lost customer group*

It can be observed that the Loyal and Can’t lose customer segments mainly originate from the United Kingdom, which can be readily explained and supported by the fact that the store operates online and is based in the UK, thus attracting a loyal customer base who frequently purchase products. Additionally, VIP, Potential, and Recent customers, apart from the UK, predominantly come from European countries such as Ireland (EIRE), Germany, France, the Netherlands, etc. This indicates that the store has attracted local customers and those from other European countries. Examining the distribution chart of the Lost customer segment, it is evident that the store's policies or products may not align well with countries in Africa or the Arab region. This is clearly evidenced by the high churn rates in these regions, second only to the United Kingdom.

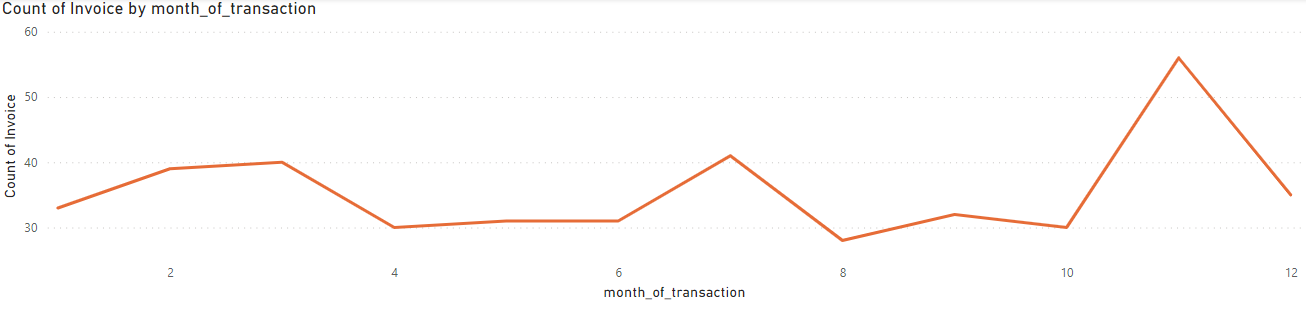
***Figure 25****: Most purchased products by each customer group* 

The data retrieved from Figure 25 shows that customers belonging to the Can’t Lose group may not make frequent purchases, but their order values are high due to their preference for high-value items such as wooden products. Both new customer groups and those who have not made purchases for a long time tend to buy similar low-cost consumer products. These products may initially attract customers to the store, but they fail to retain them for some reason. On the other hand, VIP and Loyal customer groups prefer utility items (storage boxes, baking trays) or electronic devices. Particularly, the Loyal customer group exhibits consistent preferences, as they purchase three different types of products in equal quantities. This indicates that utility items or electronic devices may not have the same initial appeal as consumer products. However, they have the potential to retain customers and bring them back to the store.

In general, customers' shopping habits tend to increase in March and November. However, this trend may vary slightly for each customer group.



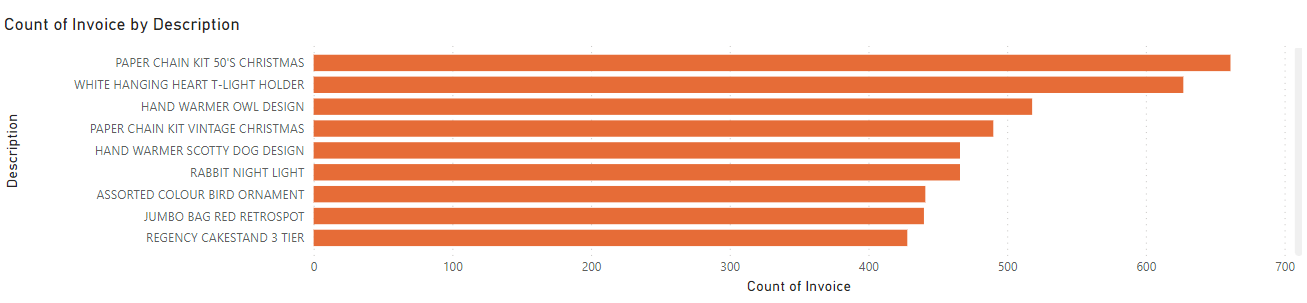
***Figure 26****: Monthly shopping habits of Can't lose customer group*



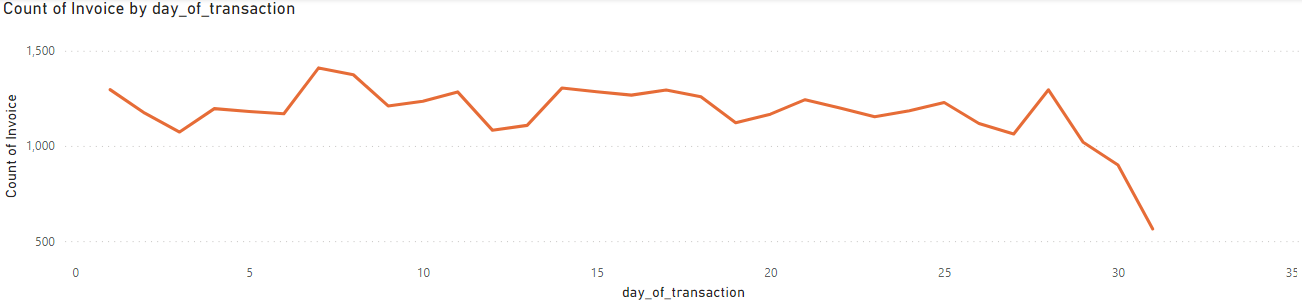
***Figure 27****: Monthly shopping habits of Loyal customer group*

For the two customer groups, Can’t Lose and Loyal, although they are all from the same specific geographical area, the United Kingdom, their shopping habits show significant differences. The Can’t Lose customer group does not make frequent purchases, leading to a shopping pattern characterized by multiple peaks and troughs. They tend to shop at the beginning and end of each quarter, except for the fourth quarter, and reduce spending between quarters. Particularly, towards the end of the year, compared to other groups, they tend to make the most purchases, focusing heavily on December, and seem to spend very little in October and November. As for the Loyal customer group, they make more frequent purchases, so their chart shows fewer peaks and troughs. This group tends to buy more at the beginning and end of the year, especially in November.

For other groups such as Recent, VIP, or Potential, their purchasing trends do not differ much from the overall trend. They tend to buy less at the beginning of the year and as the year progresses, their shopping tendencies increase, reaching a peak around November. The fact that these customer groups tend to buy more in November maybe because this is the time leading up to major Western holidays such as Christmas or New Year. This is reflected in the product data that customers typically purchase during this month.

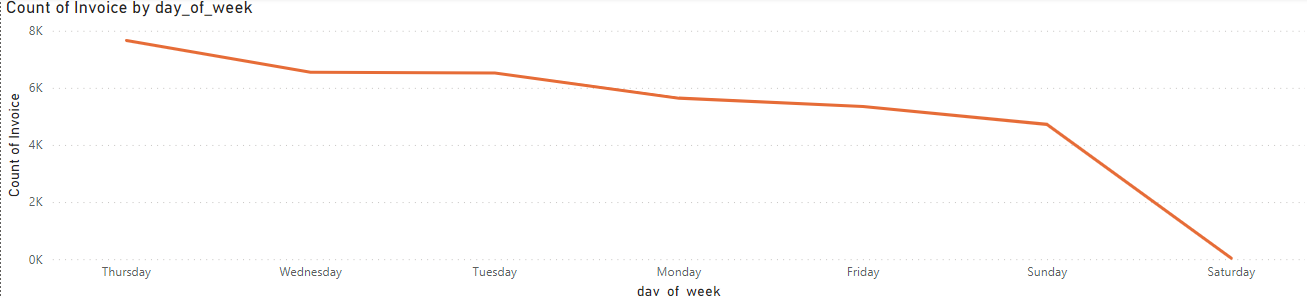


***Figure 28****: The product is purchased a lot by customers in November*



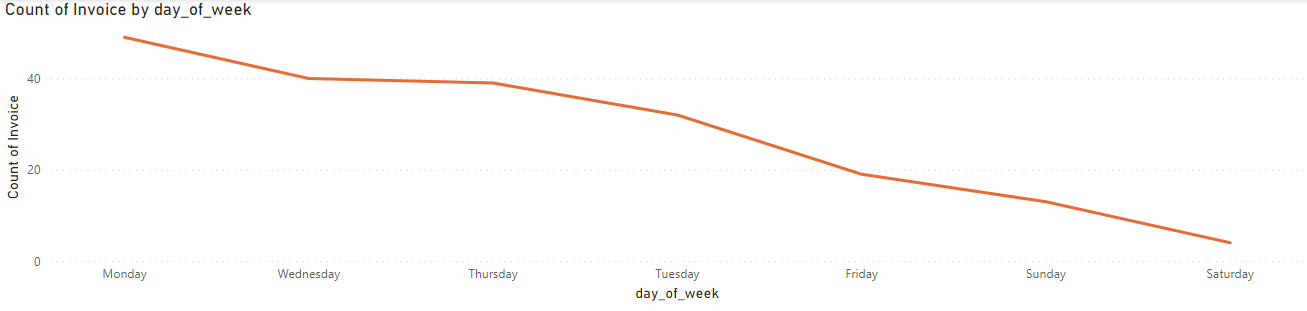
***Figure 29****: Purchasing trends by day of customer groups*

Regarding daily purchasing trends, there is little difference among the groups. Most follow the general trend of buying more at the beginning of the month and decreasing towards the end.



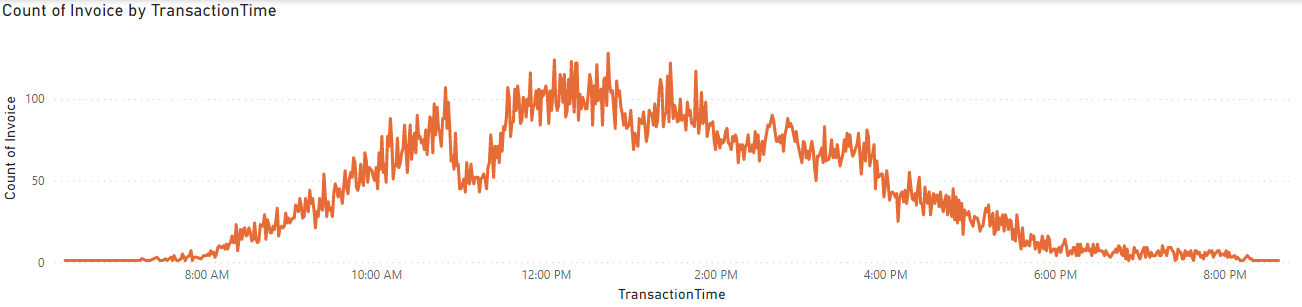
***Figure 30****: purchasing trends by day of the week of customer groups*

Regarding purchasing trends by day of the week, all groups spend more on weekdays and less on weekends. There's a common trend among the groups, particularly not spending on Saturdays. However, this pattern holds true for most groups, except for the Can't Lose group.



***Figure 31****: Daily purchasing trends of the Can't lose group*

The Can't Lose group also doesn't frequently make purchases on weekends, especially on Saturdays. However, unlike other groups, the Can't Lose group tends to do most of their shopping on weekdays, particularly on Mondays.



***Figure 32****: Purchasing trends by hour of the day*

Looking at Figure 32, we can observe that customers tend to shop more during business hours, especially from midday to mid-afternoon, where the shopping trend peaks compared to other hours. Customers also don't frequently make purchases in the evening, starting after regular business hours. This pattern is consistent across all groups.

# **5. Conclusion**

We constructed a data warehouse based on customer transactions from the Online Retail II UCI dataset. Subsequently, we analyzed the dataset to identify patterns and anomalies, aiming to improve customer relationships and increase revenue for the store. After analysis, it can be concluded that there are differences in geography and shopping habits among different customer groups. However, to draw more conclusions and facilitate more accurate analysis, the dataset must be expanded and provide more information about customers and the store.

# **Data set used in the report**

<[Online Retail II UCI | Kaggle](https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci)>

# **References**

Ralph Kimball, Margy Ross. The Data Warehouse Toolkit.

<[RFM Analysis For Successful Customer Segmentation - Putler](https://www.putler.com/rfm-analysis/#Visualizing_RFM_data)>.