American University of Armenia

Business Analytics Project

UK Retailer

Team: Aram Barseghyan, Areg Hovakimyan, Sergo Poghosyan, Mehher Ghevandiani

Fall 2023

Abstract

The following paper is going to analyze transactions from a United Kingdom-based retailer company. Specifically, the paper is going to include the data description and the problems we encountered while cleaning the data. Throughout the analysis, we are going to perform the RFM customer segmentation technique and apply the KNN machine learning model to it. Additionally, we are going to focus on finding sales trends by seasonal patterns. Moreover, we are going to investigate sales performance across different countries and find sales trends and market potential. Also, we are going to get seasonal, revenue, and geographical analysis, as well as understand customer behavior.

Contents

Chapter 1

Introduction

In this project, we are going to consider a transactional dataset from a UK-based online retailer. In the dataset, we have transactions from 2010-2011. Below, you can see the data description of an Excel file.

Data Description

InvoiceNo: Code representing each unique transaction. If this code starts with the letter 'c', it indicates a cancellation.
StockCode: Code uniquely assigned to each distinct product.
Description: Description of each product.
Quantity: The number of units of a product in a transaction.
InvoiceDate: The date and time of the transaction.
UnitPrice: The unit price of the product in sterling.
CustomerID: Identifier uniquely assigned to each customer.
Country The: country of the customer.

Chapter 2.1

Data Manipulation

After introducing the data description, it is time to move on to the part where we will perform data manipulation. Specifically, we will discuss the problems we encountered during the data-cleaning process.

The first problem that we faced was when we tried to do exploratory data analysis. In this part, we have null values for customerID as we can see count shows 406829, although we have 541909 rows. Also, we have a similar problem with the Description. Therefore, we fixed these kinds of problems by dealing with null values. After removing the null values, we do not have any problem with missing data. In the same way, we removed the data junk, such as duplicated rows and canceled orders.

Chapter 2.2

Exploratory Data Analysis

In this chapter, we are going to focus on our data analysis and statistics. Let's start with basic statistics. Below, you can see the Exploratory Data Analysis.

	count	mean	std	min	25%	50%	75%	max
Quantity	541909.0	9.552250	218.081158	-80995.00	1.00	3.00	10.00	80995.0
UnitPrice	541909.0	4.611114	96.759853	-11062.06	1.25	2.08	4.13	38970.0
CustomerID	406829.0	15287.690570	1713.600303	12346.00	13953.00	15152.00	16791.00	18287.0

• The mean quantity of products is about 9.55
• The standard deviation of quantity is quite high, implying that our data is widely spread and has outliers
• The minimum value of quantity is negative, it might indicate that we have returned or canceled orders
• The mean of unitprice is 4.6 dollars
• The third quantile of unit price shows 4.13 dollars, although the maximum order price is 38970, which implies that we have outliers
• We have null values for customerID as we can see count shows 406829 although we have 541909 rows
• CustomerID ranges from 12346 to 18287

	count	unique	top	freq
InvoiceNo	541909	25900	573585	1114
StockCode	541909	4070	85123A	2313
Description	540455	4223	WHITE HANGING HEART T-LIGHT HOLDER	2369
InvoiceDate	541909	23260	10/31/2011 14:41	1114
Country	541909	38	United Kingdom	495478

• We have 25900 unique values for InvoiceNo, the most frequent one is number 573585, with 1114 frequency
• We have 4070 unique values for StockCode, the most frequent one is 85123A, with 2313 frequency
• We have 4223 unique values for Description, the most frequent one is White hanging Heart T-Light Holder, with 2369 frequency
• We have 23260 unique values for InvoiceDate, the most frequent date is 10/31/2021 at 14:41, with 1114 frequency
• We have 38 unique values for Country, the most frequent country is United Kingdom, with 495478 frequency (Almost all data we have comes from UK)

Apart from that, we looked into canceled orders. As mentioned before, the “InvoiceNo” Code represents each unique transaction. If this code starts with the letter 'c,' it indicates a cancellation. Therefore, we only looked at the codes that start with the letter ‘c’. We got a negative mean for quantity for canceled orders. We found out that the percentage of canceled transactions in the dataset is 2.26%.

Chapter 3.1

RFM Analysis

In this part, we performed RFM analysis. First of all, we found recency, frequency, and monetary components. Then, we performed the normalization of RFM components and found the RFM score. Afterward, in order to get more accurate results, we removed outliers from normalized components by using Isolation Forest unsupervised machine learning algorithm that is used for anomaly (outlier) detection.

Finally, after having accurate data and RFM scores, we were able to segment the customers into three groups and observe the number of customers in each group. As a result, we found that most of the customers belong to the “Risky Customers” segment. Then, in second place, we have the “Potential Loyalists” segment, and the smallest segment of customers is called the “Champions”. We concluded that this result is more than acceptable as it works the same way for all businesses.

Chapter 3.2

KNN Analysis

For this part, we used only our calculated Recency Frequency and Monetary components of RFM analysis. So, if someone gives us the values of these components, we can predict the customer segment immediately.

After performing all the necessary calculations, we got our final model and started the process of model validation. As a result, with a high confidence interval, we did not have underfitting or overfitting of the data, so we accepted the model.

Chapter 3.3

Seasonal Analysis

In this section, we analyzed sales trends over time using the InvoiceDate column and identified possible seasonal patterns, peak sales periods, and overall sales trends.

After performing all necessary steps and receiving plots, we concluded that there is a clear sales increase over this one-year. Although we do not have any seasonal patterns, we can see a slight increase in sales compared to the near weeks during February, June, and October. From the perspective of residuals, it is almost random. Therefore, we have done the seasonal analysis plot correctly, and we can trust the results.

Chapter 3.4

Revenue Analysis

For this part, we performed a revenue analysis. We explored distribution across different countries and customer segments.

First of all, we calculated proportional revenue generated by each country, and based on the results, although we had 90% of our sales generated from the United Kingdom, we have more extensive proportional sales for countries like the Netherlands or Australia. The United Kingdom sees frequent purchases, but they probably mostly consist of low-priced items. In contrast, top revenue-generating countries have fewer purchases. but each transaction involves more expensive products. This suggests that people from distant countries tend to buy higher-value items or a lot of them rather than everyday inexpensive ones.

In the next step, we got the proportional revenue generated by each segment. Not surprisingly, we have “Champions”, “Potential Loyalists”, and “Risky Customers” in the corresponding decreasing order.

Chapter 3.5

Customer Behavior and Geographical Analysis

In this section, in order to understand customer behavior from different countries, we derived the plot of the top 20 countries' customer segments by their mean monetary (M in RFM) value. Plot's idea was to have a chart to understand in what direction the company can work to maximize its sales. Japan is a prime example, with nearly 80% of its customers being champions who contribute significantly to the mean monetary value. It's important to note that the majority of our data comes from the United Kingdom, but for countries like Japan, Singapore, Switzerland, and the Channel Islands, the company should consider taking action because these countries show high mean revenues and a majority of customers are champions or potential loyalists.

Chapter 4

Conclusion

In conclusion, our analysis of UK retail transactions revealed key insights into seasonal sales trends, revenue distribution across countries and customer segments, and customer purchasing behavior. One of our main challenges was data cleaning, because we wanted a very precise customer segmentation with its predicting KNN model. We have done analysis for each column of our data and removed the unnecessary ones. One main issue of our data was the time period, it would be way better if we had a range of 5-10 years, in order to analyze the seasonality and maybe even spot cycles in it. Initially one of our main goals was also to make a recommendation system for the customers based on their previous purchases and preferences, however, it caused a lot of unsolvable problems for us. Initially, we assumed that the United Kingdom has the largest mean monetary value; however; after analysis, we found out that the United Kingdom is only in 14th place, and there are countries like Japan, Singapore, and Switzerland that have several times greater mean monetary value. These findings offer valuable guidance for retailers to optimize strategies in a dynamic market.