Knowledge Discovery and Data Mining Project Report

Introduction

Knowledge discovery is a Computer Science subject that helps to extract valuable knowledge from acquired data. In that spirit, this project aims to use gathered data about retail and derive some so-called association rules, which will be used to the benefit of the online store.
This report aims to not only explain the theoretical notions at hand, but also motivate them in the current context by offering some applicable suggestions.

Dataset overview

The dataset used in this project is the online retail data containing a collection of transaction records that provide detailed information about online purchases for a UK-based online retail.The dates of the transactions are between 1/12/20210 and 1/12/2011. In total it has 541909 records and 8 different attributes.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-I	6	12/1/10 8:26	2,55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/1/10 8:26	3,39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COA	8	12/1/10 8:26	2,75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT	6	12/1/10 8:26	3,39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHIT	6	12/1/10 8:26	3,39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING	2	12/1/10 8:26	7,65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIG	6	12/1/10 8:26	4,25	17850	United Kingdom
536366	22633	HAND WARMER UNION JAC	6	12/1/10 8:28	1,85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA	6	12/1/10 8:28	1,85	17850	United Kingdom
536367	84879	ASSORTED COLOUR BIRD O	32	12/1/10 8:34	1,69	13047	United Kingdom
536367	22745	POPPY'S PLAYHOUSE BEDR	6	12/1/10 8:34	2,1	13047	United Kingdom
536367	22748	POPPY'S PLAYHOUSE KITCH	6	12/1/10 8:34	2,1	13047	United Kingdom
536367	22749	FELTCRAFT PRINCESS CHAR	8	12/1/10 8:34	3,75	13047	United Kingdom

Here is a short description of each of the different attributes

InvoiceNo : This integer value represents the invoice number, each transaction has a unique invoice number when multiple rows share the same it means all of the represented products were part of one order.

StockCode : A code consisting of numbers and letters that is unique to each product.

Description : The name of the product, always in capital letters and unique to the product.

Quantity : An integer representing the number of designated products ordered on that specific transaction.

InvoiceDate : A timestamp in format dd/mm/yy h:m that records when the order was sent.

UnitPrice : A tuple representing the price per unit of product in euro.

CustomerID : An integer uniquely representing the customer that made the order.

Country : A string representing the country where the customer is.

Data preprocessing

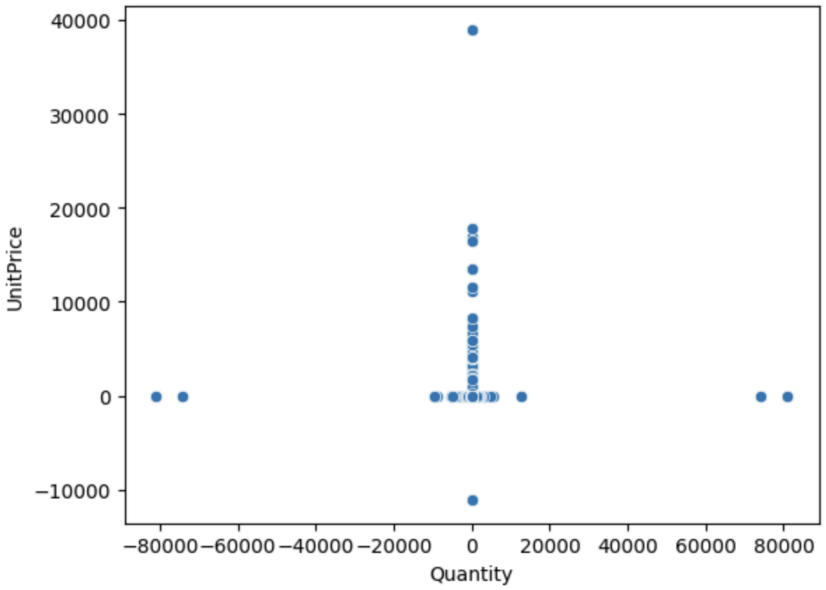
The first task we have done was data preprocessing to improve the data quality before analysing it. In this process start by identifying the wrong data, then we either clean it or repair it if it's possible.

1) **Duplicate values:**

We didn't find any duplicate records in the dataset.

2) **Negative values:**

The dataset has two attributes "UnitPrice" and "Quantity" which should be positive or greater than zero to make sense in any transaction. We plotted these two attributes in a scatterplot to see if they contain any negative values which was the case as shown in the graph below.



Scatter plot of Quantity and UnitPrice attributes

We decided to delete the records containing a negative UnitPrice because it can affect our analysis, for instance if we calculate the value of the revenue or the most products bought in which negative values of the attribute UnitPrice will make a difference in the results.

We also decided to drop the records containing the negative quantities, because the negative quantities are always related to cancelled transactions as shown in the table below whenever we have a negative quantity the InvoiceNo starts with "C" to indicate cancellation of an order.

***Table1:** Negative quantities*

InvoiceNo	StockCode	Description	Quantity
C537251	21891	TRADITIONAL WOOD	-3
C537251	22747	POPPY'S PLAYHOUSE	-6
C537251	22454	MEASURING TAPE BA	-8
C537251	22327	ROUND SNACK BOXE	-4
C537251	21915	RED HARMONICA IN	-4
C537251	84347	ROTATING SILVER AN	-9

This could be a separate dataset that could be analysed on its own but in our case we focused on analysing transitions that contribute to the profit of the online store.

3) **Null values:**

We found records with null CustomerID and Description.The missing description could be repaired by looking for the associated StockCode in other records and get its description, but we found that there is a link between the the missing descriptions and the missing customer IDs, and we decided to drop records with null CustomerID as there was no way to repair it, since the data didn't contain any other attributes to uniquely identify the customer except the CustomerID.

Here is a summary data preprocessing results:

Original number of rows : **541909**

No. of rows in DF after removing duplicates : **541909**

No. of rows in DF after removing negative prices : **541907**

No. of rows in DF after removing negative quantities : **531283**

No. of rows in DF after removing empty customers ID : **397924**

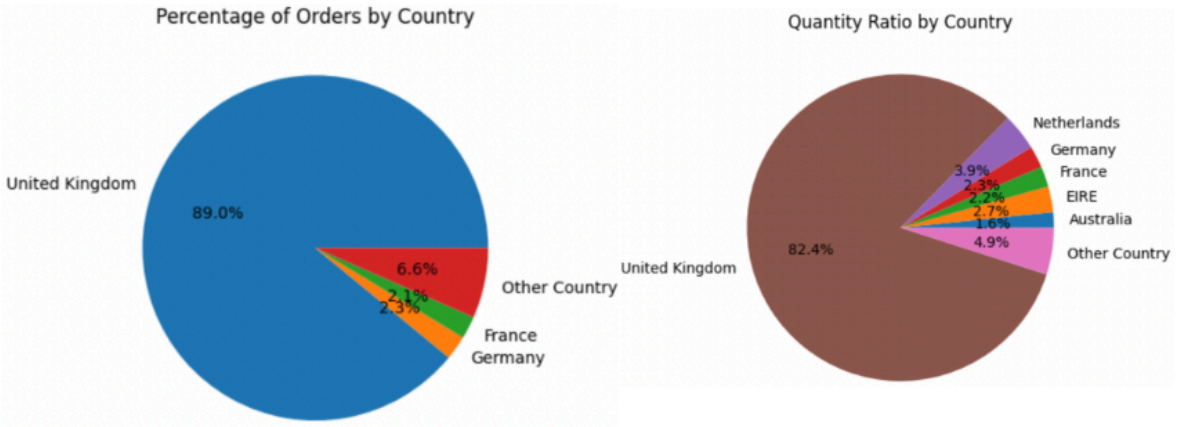
No. of rows in DF after removing empty description : **397924**

In total, we dropped **143985 rows** and kept **397924 rows** for analysis.

Data visualisation and analysis

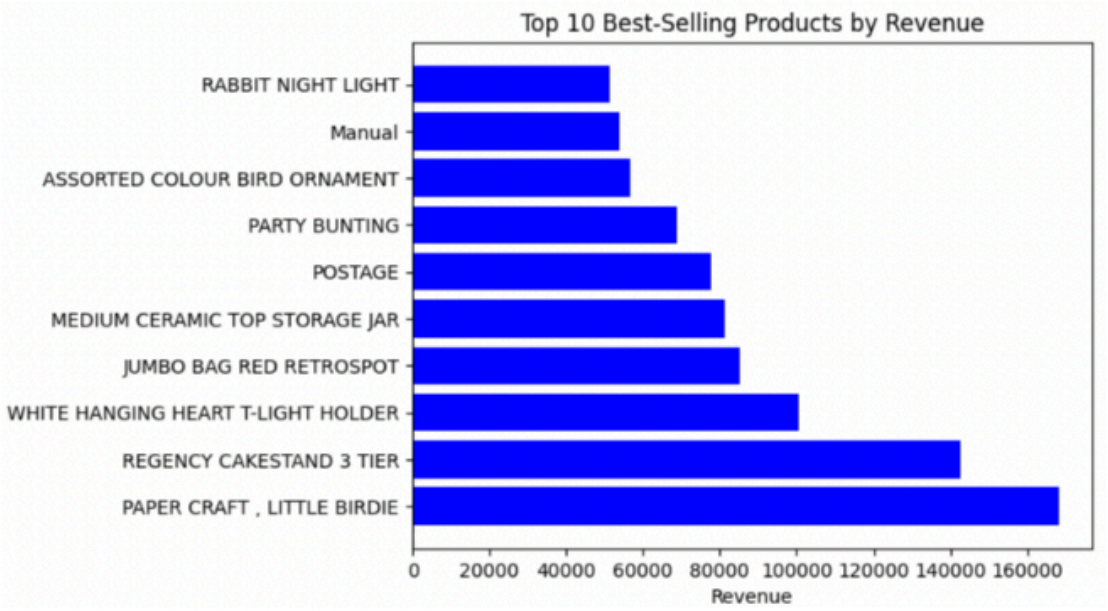
Some valuable insights were to be found in the cleaned up data, two aspects we focused on were

Orders and Quantity by Country

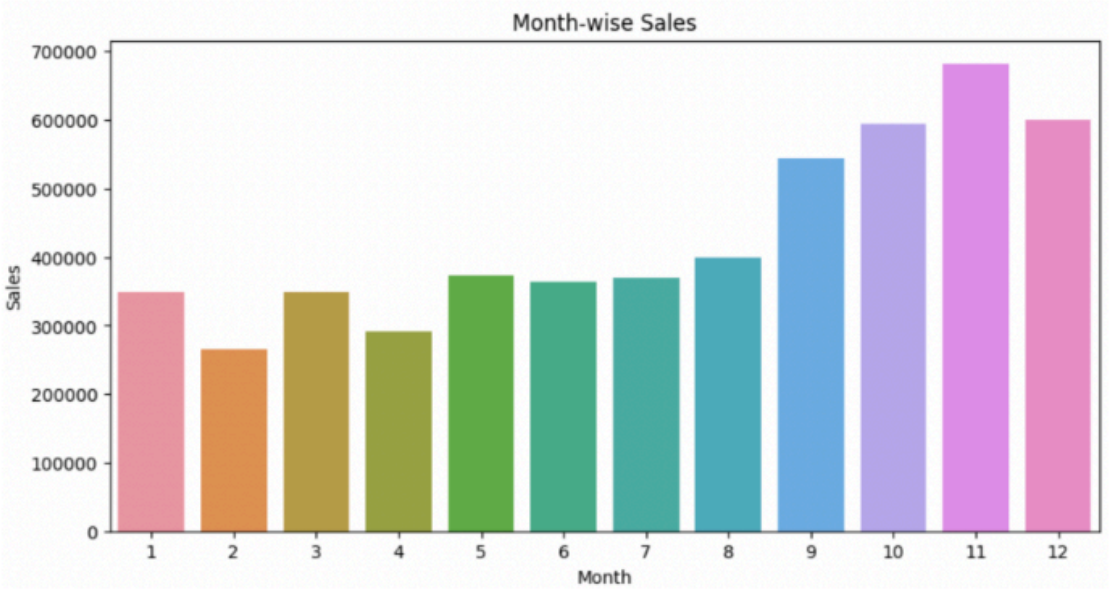


The geospatial nature of the data, by looking at where the orders originated from and where most of the products were sent. This is helpful so that the retail team knows their customers better.

Top Best-Selling Products



The top 10 best selling products is an interesting data to gather to know what products are in high demand and produce most of the revenue. (see in improvement part that this metric is not the best one to gather)



The temporal nature of the data gives valuable information about the customer's habits and allows a better preparation for the logistics thanks to the month-by-month sales breakdown. The retail team knows when to expect most of its orders and can plan accordingly.

Association discovery

We have chosen association discovery from the proposed data analysis techniques because it's widely used in the domain of retail and customer behaviour analysis.

Association discovery, also known as market basket analysis, is a data mining technique used to discover relationships among items in a large dataset, in other words it helps to identify patterns and dependencies between the items, it basically looks for items that co-occur together. In the case of retail data, association discovery is used to find products that often are purchased together, these rules are represented in a conditional form like "if product A is bought, then product E is bought", we can have more than one item in an association rule for example "if product A and product E are bought, then product C is bought".

In the context of association discovery, there are three important metrics we used: support, confidence and lift.
Support: it measures the popularity of an itemset in the dataset. Higher support means the itemset appears more frequently in the dataset, in other words it appears often in the transactions.

$$Support(A) = No.\ of\ transactions\ containing\ A\ /\ Total\ no.\ of\ transactions$$

Confidence: it measures the conditional probability of an association rule, for instance the probability that product E is bought given that product R is bought, so the values range from 0 to 1. It basically indicates the strength of the relationship between the items in the association rules, a strong association has a confidence close to 1.

$$Confidence(A \rightarrow B) = Support(A \cup B)\ /\ Support(A)$$

Lift: it gives us more information about the strength of associations, A lift value greater than 1 indicates a positive association, simply it means that the presence of product A increases the chance that product E is present. Aloft value less than 1 indicates a negative association and a lift value equal to 1 means that there is no association between the products.

$$Lift(A \rightarrow B) = Support(A \cup B)\ /\ (Support(A) \times Support(B))$$

To analyse the retail dataset using association discovery , we used the **Apriori algorithm,** which is a common algorithm used in this technique. It simply generates the frequent itemsets, then it derives the rules based on the defined thresholds like minimum confidence...etc.

Here are the resulting association rules :

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(60 TEATIME FAIRY CAKE CASES)	(PACK OF 72 RETROSPOT CAKE CASES)	0.034198	0.054569	0.018735	0.547847	10.039479
1	(ALARM CLOCK BAKELIKE CHOCOLATE)	(ALARM CLOCK BAKELIKE GREEN)	0.018040	0.040947	0.011372	0.630385	15.395009
2	(ALARM CLOCK BAKELIKE CHOCOLATE)	(ALARM CLOCK BAKELIKE RED)	0.018040	0.044220	0.012108	0.671202	15.178723
3	(ALARM CLOCK BAKELIKE IVORY)	(ALARM CLOCK BAKELIKE GREEN)	0.023848	0.040947	0.013745	0.576329	14.074872
4	(ALARM CLOCK BAKELIKE ORANGE)	(ALARM CLOCK BAKELIKE GREEN)	0.018899	0.040947	0.011331	0.599567	14.642375
5	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.032684	0.040947	0.017140	0.524406	12.806810
6	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.040947	0.044220	0.026426	0.645355	14.594209
7	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.044220	0.040947	0.026426	0.597595	14.594209
8	(ALARM CLOCK BAKELIKE IVORY)	(ALARM CLOCK BAKELIKE RED)	0.023848	0.044220	0.015422	0.646655	14.623621
9	(ALARM CLOCK BAKELIKE ORANGE)	(ALARM CLOCK BAKELIKE RED)	0.018899	0.044220	0.012681	0.670996	15.174061

To go deeper : Threeway association

In order to delve deeper into association rules and still find valuable insights, we found threeway association rules depicted in the following table :

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
901	(POPPY'S PLAYHOUSE BEDROOM , POPPY'S PLAYHOUSE...	(POPPY'S PLAYHOUSE KITCHEN)	0.011060	0.018686	0.010035	0.907317	48.607021
619	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENC...	(GREEN REGENCY TEACUP AND SAUCER)	0.023522	0.037279	0.021040	0.894495	23.994742
613	(REGENCY CAKESTAND 3 TIER, PINK REGENCY TEACUP...	(GREEN REGENCY TEACUP AND SAUCER)	0.016670	0.037279	0.014620	0.877023	23.526037
900	(POPPY'S PLAYHOUSE KITCHEN, POPPY'S PLAYHOUSE ...	(POPPY'S PLAYHOUSE BEDROOM)	0.011599	0.017048	0.010035	0.865116	50.746188
896	(REGENCY CAKESTAND 3 TIER, PINK REGENCY TEACUP...	(ROSES REGENCY TEACUP AND SAUCER)	0.016670	0.042242	0.014297	0.857605	20.302132
620	(GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY...	(ROSES REGENCY TEACUP AND SAUCER)	0.024817	0.042242	0.021040	0.847826	20.070631
626	(REGENCY CAKESTAND 3 TIER, GREEN REGENCY TEACU...	(ROSES REGENCY TEACUP AND SAUCER)	0.020177	0.042242	0.016832	0.834225	19.748643
632	(JUMBO BAG PINK POLKADOT, JUMBO BAG STRAWBERRY)	(JUMBO BAG RED RETROSPOT)	0.015807	0.086319	0.012516	0.791809	9.173106
644	(JUMBO BAG PINK POLKADOT, JUMBO STORAGE BAG SUKI)	(JUMBO BAG RED RETROSPOT)	0.015160	0.086319	0.011869	0.782918	9.070107
600	(ALARM CLOCK BAKELIKE GREEN, ALARM CLOCK BAKEL...	(ALARM CLOCK BAKELIKE RED)	0.014836	0.047313	0.011599	0.781818	16.524267

The usage of two antecedents to yield one consequent allows a more important granularity and to have some more intricate associations.

Importance of association rules

Association rules yield some valuable metrics to measure the link between product buying.

It will help a retailing team to understand how a product A (called antecedent) and a product B (called consequent) are often associated through the orders done by the customers.
This means that if when buying the antecedent, the consequent is often bought, the lift will reflect this and give the retailing team a good idea of the strength of the link between the two items.

This metric can be leveraged in multiple ways :

1.
    a. Product placement (physical store) : Putting strongly associated products next to each other in a physical store will easily display those products to clients likely to buy them during the same purchase.
    b. Suggestion (online store) : Suggesting strongly associated products to a customer browsing or buying the antecedent product will display the consequent to clients likely to buy it.
2. Target advertisement : Advertising consequent products to consumers that bought antecedents, via any way (leaflet advertisement, TV advertisement, email advertisement) will be an effective way to promote the mentioned consequent product to the right audience.
3. Conditional promotion : Creating a sale where the products benefit from a lower price when bought together is a good idea as it incentivizes customers to buy a product they were already likely to buy. It also encourages the customer to buy those at the same place as opposed to obtaining the two products from independent retail stores.
4. Bundle : Creating a bundle where the two products are bought as one is likely to be effective as a non negligible amount of customers are looking to buy both antecedents and consequents.

Conclusion

In conclusion, this study about retailing data focused on association rules yielded some usable and practical insights as demonstrated in this report.

Some key points to improve :

- The data did not allow to calculate benefits so revenue was calculated instead, it would be preferable to remove the costs associated with the products in order to know which ones yield more for the company.
- Some other knowledge discovery techniques could be applied in order to gather more informations, for instance clustering could be used to predict data concerning a new order (country of origin for instance could be a predicted output)

Group Members:
Amani KEDDAM
Eliott BONTE