Machine Learning for Business – CA1
Summary Report

Report by: Leandro Marigo
Lecturer: Sam Weiss
Date: April 2023

About the dataset

Customer transactions data from a global online retailer from 2009 to 2011.

Atribute information:

Invoice: Unique identifier of a transaction which can include several different products.
StockCode: Unique identifier of a product.
Quantity: Quantity of product bought in a invoice/transaction.
InvoiceDate: Date which an invoice / transaction took place.
Price: Unit price of a product.
Customer: ID / Unique identifier for a customer.
Country: Country in which the invoice / transaction took place.

Licence

Introduction

I have been given access to a dataset containing customer transactions for an online retailer and tasked with using machine learning tools to gain and report on business insights. The audience for this report are non-specialists.

The main goals are:

Clustering

Apply and evaluate various clustering techniques with the aim of generating actionable insights from the data.

- Apply at least 3 clustering algorithms to the dataset.
- Evaluate the performance of the algorithms and make a recommendation as to which gives the "best" results.
- Include in my report my own interpretation of the results.

Market Basket Analysis

Perform a market basket analysis of the transaction data.

- Include in my report a comparison and evaluation of at least two algorithms.

Research and summarise an application of market basket analysis that is not transaction data.

The report is structured in different sections as outlined below:

1. Exploratory Data Analysis

The dataset contains 8 features and 1,067,371 observations/rows.
The columns are a mix of data types such as Object/String, numerical with one datetime which is the "InvoiceDate" column.

All variables are skewed and some of the main variables for the clustering purposes has a high cardinality that will need to be further investigated / transformed before applying the models.
Negative values are treated as errors and removed from the data as this could impact on the clustering and visualization of the customers segmentation.

Dataset Statistics		Dataset Insights	
Number of Variables	8	`Customer ID` has 243007 (22.77%) missing values	Missing
Number of Rows	1.0674×10^{06}	`Quantity` is skewed	Skewed
Missing Cells	247389	`Price` is skewed	Skewed
Missing Cells (%)	2.9%	Dataset has 34335 (3.22%) duplicate rows	Duplicates
Duplicate Rows	34335	`Invoice` has a high cardinality: 53628 distinct values	High Cardinality
Duplicate Rows (%)	3.2%	`StockCode` has a high cardinality: 5305 distinct values	High Cardinality
Total Size in Memory	324.7 MB	`Description` has a high cardinality: 5698 distinct values	High Cardinality
Average Row Size in Memory	319.0 B	`Quantity` has 22950 (2.15%) negatives	Negatives
Variable Types	Categorical: 3 Numerical: 3 DateTime: 1 GeoGraphy: 1		

Overall observations of EDA:

Dataprep.eda provides a broad overview for Exploratory Data Analysis

- We can see immediately the distribution of all the Features to give an insight of how to approach cleaning and Feature Selection
- All features are skewed, there are missing values and duplicate data that will be handled in the data cleaning phase.
- In the column "Quantity" there are several records with a negative number, and a few negative numbers in the column "Price". Both columns will be further investigated.

Describe, shape and info shows an overall structure and statistical metrics of the dataset

- Data type variables:
  - Object = 4 variables
  - int64 = 1 variable
  - datetime64[ns] = 1 variable
  - float64 = 2 variables

Missing values

- Missing values identified in columns Description and Customer ID and will be handled at later stage during data cleaning.

As the goal will be to form clusters containing valuable information about the transactions, customers and products bought ,some of the variables won't be useful for this purpose as they do not add information and are not corelated with the group of data that will form the different clusters. These variables will be dropped:

- **Either Description or StockCode:** As they represent the same information (Product) in different formats. I will decide later after more investigation which one to drop.

2. Data Cleaning / Feature Engineering

At this stage duplicate and missing values were removed, the column “Description” was dropped as it was overlapping with the column “StockCode”.

Based on the column “InvoiceDate” a new column was created with the “Year” of the transaction. I was also added a new column named “Total_Sales” which calculates the customers total spend using the columns “Price” multiplied by the “Quantity”. That new column “Total_Sales” will be one of the key variables for the clustering and customer segmentation purposes.

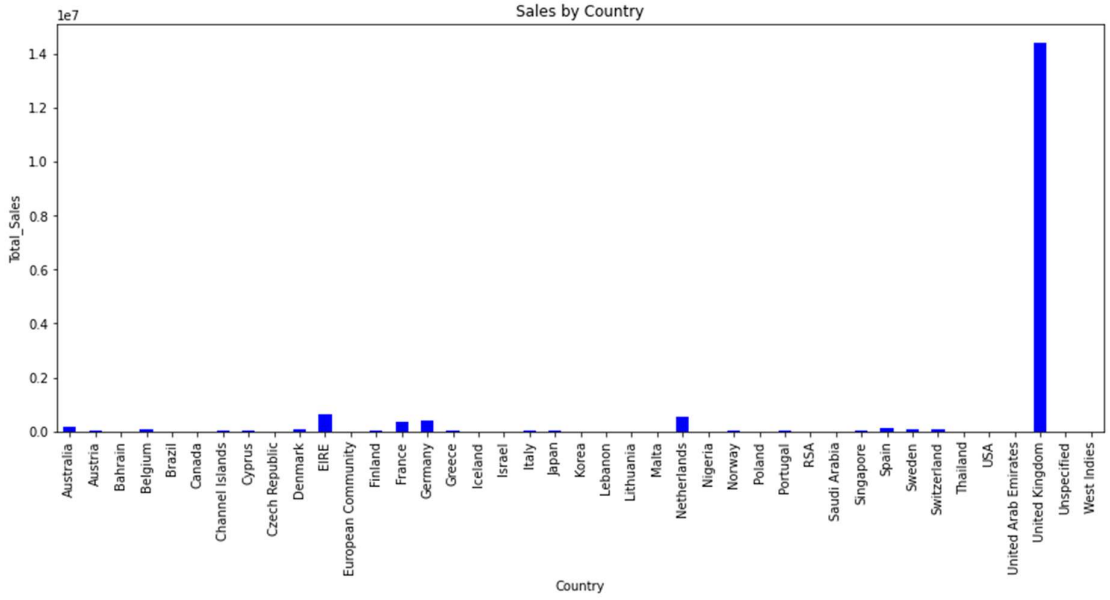
3. Data Visualization / Further EDA

At this stage I looked in more detail for outliers identified after visualizing the box plot. However I am treating those values as natural variation of the data as different customer types / behaviours and therefore will keep it for the clustering purposes.

"Natural variation can produce outliers which most of the time are not necessarily a problem. If the extreme value is a legitimate observation that is a natural part of the population you’re studying, you should leave it in the dataset." (Frost, 2019)

"Sometimes it’s best to keep outliers in your data. They can capture valuable information that is part of your study area. Retaining these points can be hard, particularly when it reduces statistical significance! However, excluding extreme values solely due to their extremeness can distort the results by removing information about the variability inherent in the study area. You’re forcing the subject area to appear less variable than it is in reality." (Frost, 2019)

It was also observed that most of the transactions / sales from the dataset took place in the UK. (Fig 1)



(Fig 1)

4. RFM Segmentation

In order to apply the clustering model and segment the customers I will use the RFM model to calculate new variables that will help understand how recently a customer has bought in the shop, how frequently they've engaged with the shop and how much money they've spent. RFM is a data modeling method is mainly used in the marketing fields to segment customers and better target marketing campaigns and strategies. The model is built using three key factors:

1. how recently a customer has transacted with a brand
2. how frequently they've engaged with a brand
3. how much money they've spent on a brand's products and services

Delval, F. (2021). What is RFM Analysis? [online] ActionIQ. Available at: https://www.actioniq.com/blog/what-is-rfm-analysis/.

This is what the head of the new data frame looks like: (Fig 2)

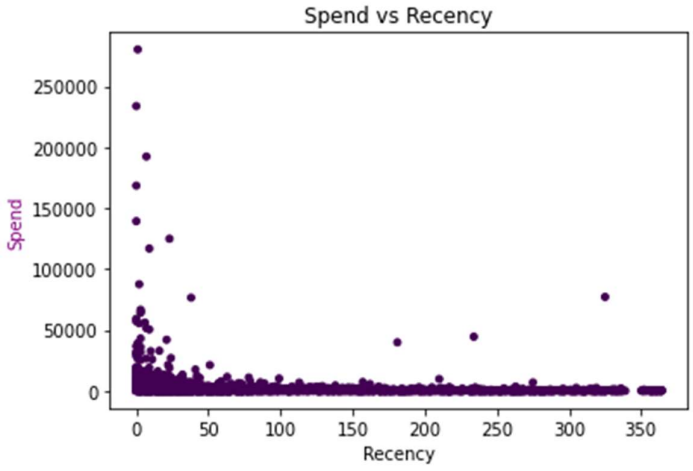
	Customer ID	Recency	Frequency	Spend
0	12346.00	325	1	77183.60
1	12347.00	1	151	3598.21
2	12348.00	74	31	1797.24
3	12349.00	18	73	1757.55
4	12350.00	309	17	334.40

(Fig 2)

5. DBSCAN before Pre-processing

Considering that the DBSCAN algorithm works well with high density and sparse data I tested the model before normalizing and pre-processing the data.

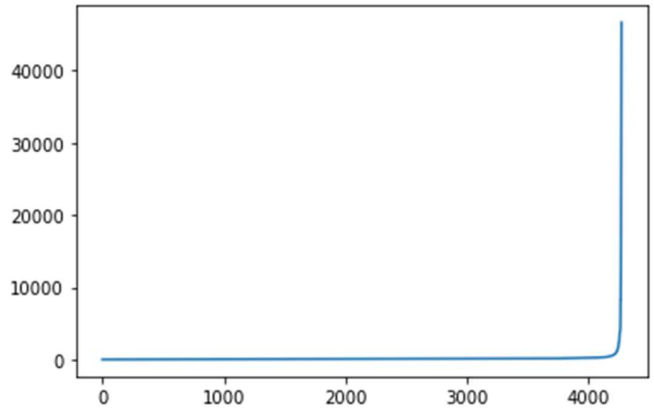
Using the new features "Spend" and "Recency" it was observed that the model was not performing well and was not able to create / segment the date into different clusters. (Fig 3)



(Fig 3)

It wasn't possible to fine tune the model using the k-nearest neighbours method as the data was still very sparse the distances graph was not able to capture and identify the optimal epsilon and therefore it was decided to proceed with normalizing the data before applying the models.

No good insights from the distances graph at this stage, as the slope is very steep in the corner of the graph: (Fig 4)



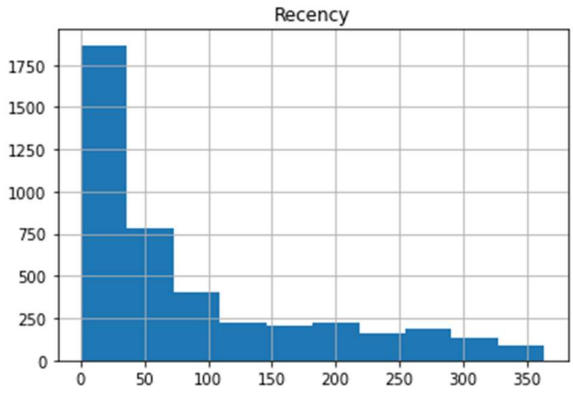
(Fig 4)

6. Log Transform

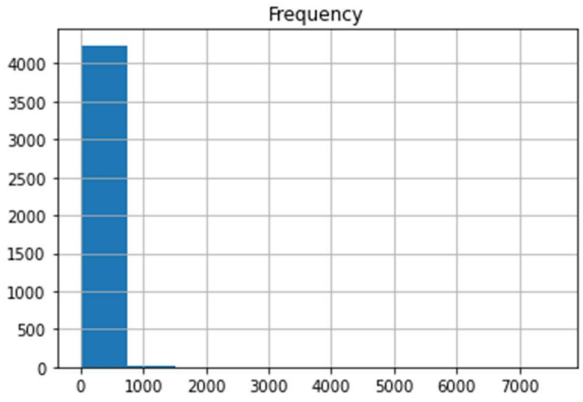
The features that will be used to apply the ML model are skewed and sparse and therefore it was applied the "log transformation" to normalize the data and reduce skewness so the model can work properly.

"This process is useful for compressing the *y*-axis when plotting histograms. For example, if we have a very large range of data, then smaller values can get overwhelmed by the larger values. Taking the log of each variable enables the visualization to be clearer." (Metcalf and Casey, 2016)

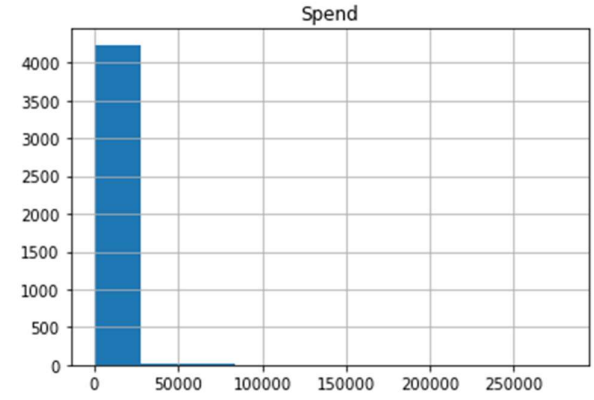
Distribution of the data for each variable before the log transformation: (Fig 5, 6, 7)



(Fig 5)

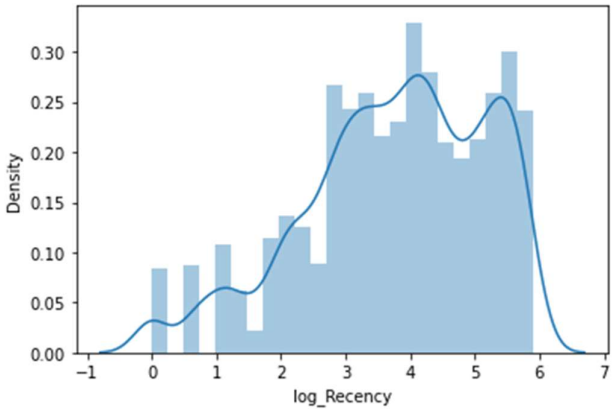


(Fig 6)

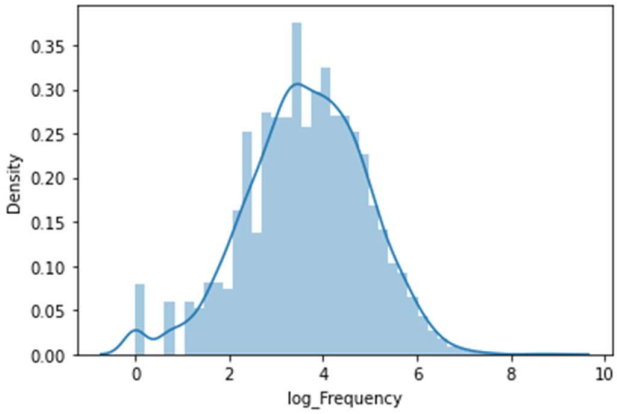


(Fig 7)

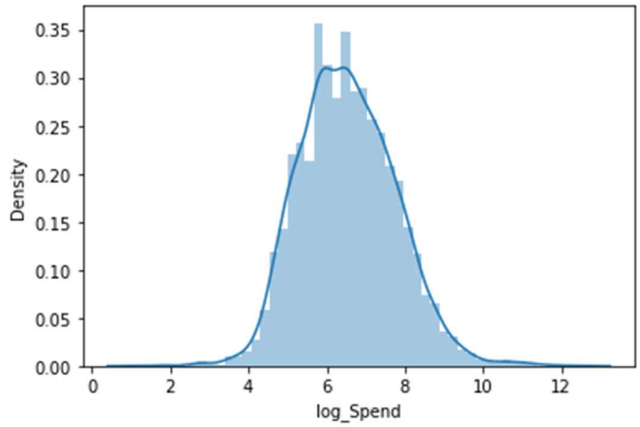
Distribution of the data for each variable after the log transformation: (Fig 8, 9, 10)



(Fig 8)



(Fig 9)



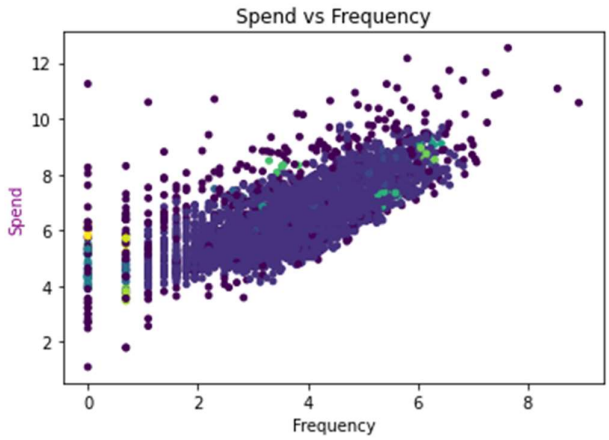
(Fig 10)

7. DBSCAN after Pre-processing

After the log transformation the model performed much better, however DBSCAN was still not able to identify concise clusters even after applying the optimal epsilon. I used the variables “Spend” and “Frequency” to plot the graph as those variables seem to have high and positive correlation.

Graph applying optimal epsilon. (Fig 11)

It is observed that no clusters were formed.



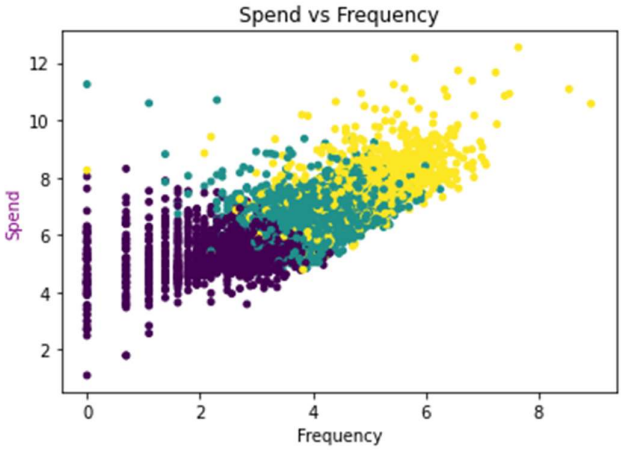
(Fig 11)

8. K-Means

The KMeans algorithm performed much better than DBSCAN, being able to form different clusters.

Using the Elbow method to identify the optimal number of clusters the elbow occurs between 3 and 5 clusters and when applying the Silhouette score to find the optimal number of cluster, 2 clusters got the highest score. I tested the model applying 2, 3 and 4 clusters and the best visual results for the customer segmentation from a marketing strategy standpoint was applying 3 clusters and visualizing them by the Total Spend per customer vs their Frequency in the shop.

We can immediately see that Cluster 0 (purple) represents customers with low spend and frequency, Cluster 1 (green) represents the average customer with average spend and frequency and Cluster 2 (yellow) are the Premium customers with high spend and frequency. (Fig 12)

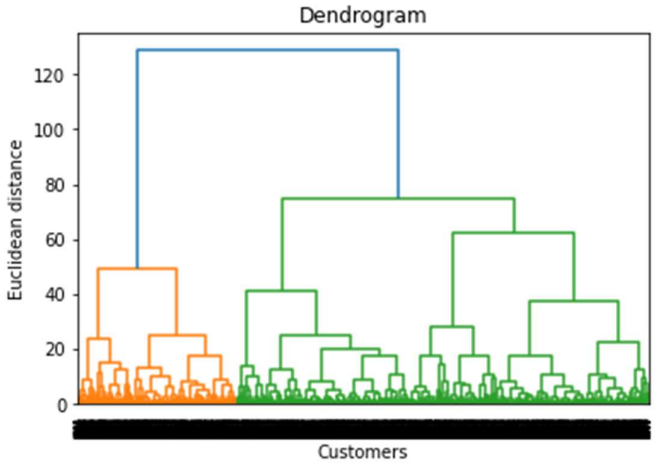


(Fig 12)

9. Hierarchical Clustering

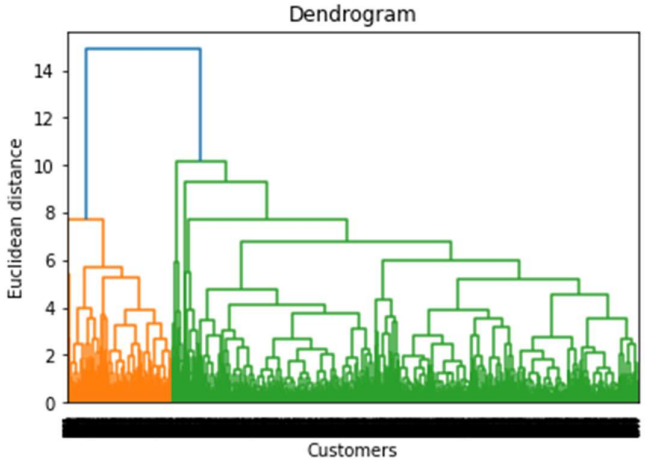
When applying the hierarchical clustering model and using dendrograms to visualize the formed clusters, the method ‘Ward’ performed better over the method ‘Complete’ and was able to form clearer defined clusters.

Dendrogram applying the method ‘Ward’. (Fig 13)



(Fig 13)

Dendrogram applying the method ‘Complete’. (Fig 14)



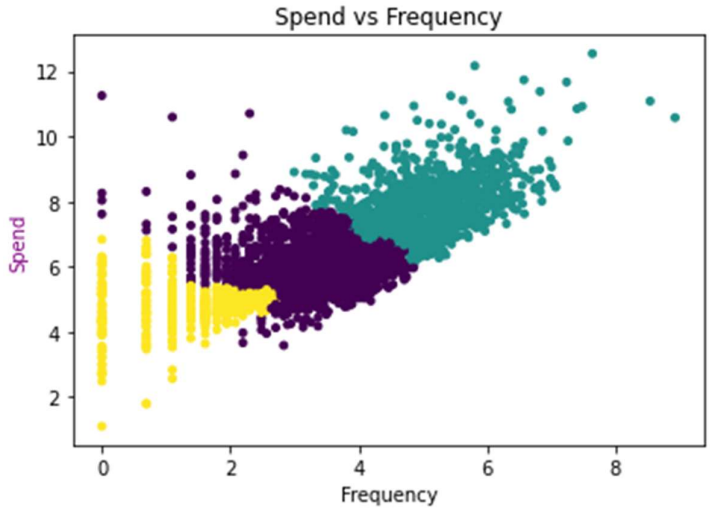
(Fig 14)

The method ‘Ward’ was chosen to visualize the clusters by the customers Spend vs Frequency and it is observed that the Hierarchical Clustering got similar results to Kmeans with a better distinction with no overlaps across the points.

Hierarchical Clustering results visualized by customers Spend vs Frequency. (Fig 15)

Interpretation of the results

We can see immediately that the Yellow cluster represents customers with low spend and low frequency, the purple represents the average customer with a few points captured in the high spend realm, and the green cluster are the premium customers with high spend and high frequency. Based on that we will be able develop different marketing strategies based on the different types of customers identified through hierarchical clustering.



(Fig 15)

10. MBA - Apriori

For applying the Market Basket Analysis models the original dataset was used and re-transformed by cleaning the data again, removing duplicates, missing values and negative values from columns “Price” and “Quantity”. As the UK region represents more than 80% of the data it was created a analysed a separate data frame for only the UK products.

Results of the MBA applying the Apriori algorithm. (Fig 16)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
12	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.05	0.14	0.03	0.71	5.04	0.03	2.92
15	(SWEETHEART CERAMIC TRINKET BOX)	(STRAWBERRY CERAMIC TRINKET BOX)	0.03	0.05	0.02	0.69	13.95	0.02	3.06
16	(WOODEN PICTURE FRAME WHITE FINISH)	(WOODEN FRAME ANTIQUE WHITE)	0.05	0.05	0.03	0.60	11.77	0.03	2.39
17	(WOODEN FRAME ANTIQUE WHITE)	(WOODEN PICTURE FRAME WHITE FINISH)	0.05	0.05	0.03	0.57	11.77	0.03	2.19
3	(LOVE BUILDING BLOCK WORD)	(HOME BUILDING BLOCK WORD)	0.04	0.05	0.02	0.53	10.02	0.02	2.01

(Fig 16)

11. MBA – FP-Growth

Results of the MBA applying the FP-Growth algorithm. (Fig 17)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
4	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.05	0.14	0.03	0.71	5.04	0.03	2.92
7	(SWEETHEART CERAMIC TRINKET BOX)	(STRAWBERRY CERAMIC TRINKET BOX)	0.03	0.05	0.02	0.69	13.95	0.02	3.06
10	(WOODEN PICTURE FRAME WHITE FINISH)	(WOODEN FRAME ANTIQUE WHITE)	0.05	0.05	0.03	0.60	11.77	0.03	2.39
11	(WOODEN FRAME ANTIQUE WHITE)	(WOODEN PICTURE FRAME WHITE FINISH)	0.05	0.05	0.03	0.57	11.77	0.03	2.19
1	(LOVE BUILDING BLOCK WORD)	(HOME BUILDING BLOCK WORD)	0.04	0.05	0.02	0.53	10.02	0.02	2.01

(Fig 17)

Both models (Apriori and FP-Growth) presented same results.

Interpretation of the MBA results

In the UK customers tend to buy more frequently items for their homes such as "Hanging Holders" in different collors, "Ceramic Trinket Boxes" in different formats and colors, "Wooden picture frames" in different colors and types, and many other home items.

By looking at the metrics, the probability of customers buying these products together varies between 71% and 53% from the highest to lowest, the highest being after customers buying the "RED HANGING HEART T-LIGHT HOLDER" there's a 71% probability they will also buy the "WHITE HANGING HEART T-LIGHT HOLDER", and the lowest being after customers buying the "LOVE BUILDING BLOCK WORD" there's a 53% probability they will also buy the "HOME BUILDING BLOCK WORD". Based on Lift and Leverage all items have a small depedency however looking at conviction and that metric being higher than 1 it means the products are indenpendet altough they have a high probability of being bought together.

12. Other types of MBA applications used in different Industries

The Market Basket Analysis algorithms are well known for data mining and association rules mainly in the retail sector in order to identify associated products that are normally bought together. However the MBA machine learning algorithms can also be used in other areas such as Finance, Manufacturing, Pharmaceutical, Customer Behaviour, Medicine and Telecom by basically implementing its association rules to help with identifying, predicting or detecting any kind of trend for each individual industry goals. (Chaudhary, n.d.)

Other examples of Market Basket Analysis outside of the retail industry is the recomendation engines used in the streaming aplications such as Netflix, Amazon, and Spotify, which uses MBA to make association rules and recommend what to watch or listen based on prior behaviour. MBA can also be used in the Pharmaceutical industry specifically for Adverse Drug Events (ADEs) where the "transactions" in this case will be the reports of ADEs, the items/products are the drugs taken, as well as other information like symptoms. MBA will simply use association rules to identify patterns of co-occurence in the reports and provide insights about what drugs and symptoms go together to prevent and make right drug recommendations for each case. --> Based on the research (Marketing, 2019)

13. Bibliography

Reference list

Bhardwaj, A. (2020). Silhouette Coefficient : Validating clustering techniques. [online] Medium. Available at: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c.

Çankaya, M.F. (2022). How To Normalize Your Unsupervised Data For Clustering Methods. [online] Medium. Available at: https://engineering.teknasyon.com/how-to-normalize-your-unsupervised-data-for-clustering-methods-9389298d20d5.

Chaudhary, S. (n.d.). Understanding Market Basket Analysis in Data Mining. [online] www.turing.com. Available at: https://www.turing.com/kb/market-basket-analysis.

Delval, F. (2021). What is RFM Analysis? [online] ActionIQ. Available at: https://www.actioniq.com/blog/what-is-rfm-analysis/.

Frost, J. (2019). Guidelines for Removing and Handling Outliers in Data - Statistics by Jim. [online] Statistics by Jim. Available at: https://statisticsbyjim.com/basics/remove-outliers/ [Accessed 28 Dec. 2022].

Marketing, S. (2019). Use Market Basket Analysis for More Than Market Baskets. [online] Syntelli Solutions Inc. Available at: https://www.syntelli.com/use-market-basket-analysis-for-more-than-market-baskets [Accessed 15 Apr. 2023].

Metcalf, L. and Casey, W. (2016). Cybersecurity and applied mathematics. Amsterdam: Elsevier.

Müller, A.C. and Guido, S. (2017). Introduction to machine learning with Python : a guide for data scientists. Beijing: O'reilly.

Rink, K. (2021). Four mistakes in Clustering you should avoid. [online] Medium. Available at: https://towardsdatascience.com/common-mistakes-in-cluster-analysis-and-how-to-avoid-them-eb960116d773.

Vanderplas, J.T. (2017). Python data science handbook : essential tools for working with data. Beijing Etc.: O'reilly, Cop.

Verma, Y. (2021). Why Data Scaling is important in Machine Learning & How to effectively do it. [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/why-data-scaling-is-important-in-machine-learning-how-to-effectively-do-it/.