

CA1: CLUSTERING AND MARKET BASKET ANALYSIS

Applying clustering techniques and
performing market basket analysis
on a commercial dataset

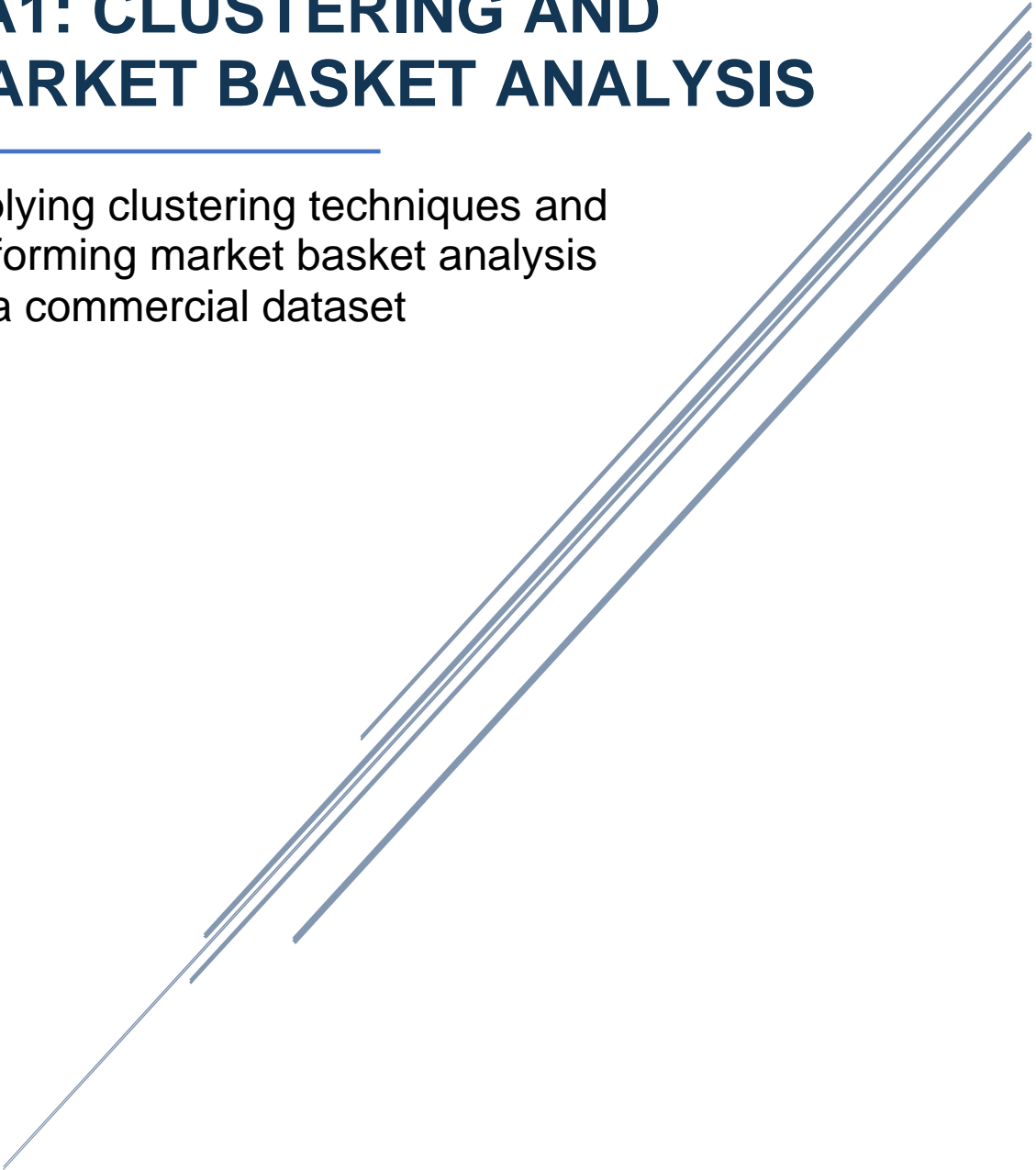


Table of Contents

0. INTRODUCTION.....	2
1. BUSINESS UNDERSTANDING.....	2
2. DATA DESCRIPTION.....	2
3. DATA PREPARATION	2
4. MODELING- UNSUPERVISED LEARNING	3
4.1. CLUSTERING TECHNIQUES.....	3
4.1.1. RFM ANALYSIS	3
4.1.2. K-MEANS.....	3
4.1.2. K-MEDOIDS	5
4.1.3. FUZZY C-MEANS.....	6
4.2. MARKET BASKET ANALYSIS TECHNIQUES	6
4.2.1. APRIORI PRINCIPLE ALGORITHM	7
4.2.2. FP-GROWTH.....	8
5. EVALUATION OF MODELS.....	9
5.1. EVALUATING AND CONCLUSIONS FOR CLUSTERING ALGORITHMS	9
5.2. EVALUATING AND CONCLUSIONS FOR MBA ALGORITHMS.....	9
6. RESEARCHING- ADDITIONAL MBA APPLICATION	10
APPENDIX.....	10
a) Appendix 1	11
b) Appendix 2.....	11
c) Appendix 3	11
REFERENCE LIST	12

0. INTRODUCTION

In the following assignment, I am going to analyse a transnational data set that contains customer transactions for an online retailer occurred between 2009 and 2011. Together with this PDF document, a Jupyter Notebook is attached where you can consult all the codes used for the analysis.

1. BUSINESS UNDERSTANDING

Data analysis has become an essential tool for large companies today. By using data analysis techniques, companies can obtain valuable information about their customers, products and internal processes, which allows them to make more informed and strategic decisions.

2. DATA DESCRIPTION

The dataset contains customer transactions for an online retailer occurred between 01/12/2009 and 09/12/2011. The size of the data set before cleaning is 106371 rows or observations and 8 columns or attributes (106371 x 8). Our original columns are: "Invoice", "StockCode", "Description", "Quantity", "InvoiceDate", "Price", "Customer ID" and "Country". As we can see in the Table 1A we have 4 categorical variables, 3 categorical and 1 datetime.

3. DATA PREPARATION

In order to apply ML algorithms, it is important to clean and manipulate the data in a way that makes it useful for analysis. In this case:

- The "StockCode" attribute has been removed because it is just a label that provides the same information as the "Description" attribute.
- Removed 34531 duplicate rows.
- Removed 4275 missing values from the "Item_name" attribute and 235063 from the "Customer_ID" attribute.
- Negative values or values equal to 0 have been eliminated from the variables "Price" and "Quantity" because it does not make sense to have negative prices and quantities for the focus of my analysis.
- The outliers of the variables "Price" and "Quantity" have been eliminated using the Interquartile range technique (see Table 2A).

After all these procedures, we get a new clean dataset named as "df_cleaned" (663329 rows x 7 columns) which will be used for the ML (see Table 1).

	Invoice	Item_name	Quantity	Invoice_Date	Price	Customer_ID	Country
0	489434	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
4	489434	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom
5	489434	PINK DOUGHNUT TRINKET POT	24	2009-12-01 07:45:00	1.65	13085.0	United Kingdom

Table 1. Display of “df_cleaned”.

4. MODELING- UNSUPERVISED LEARNING

4.1. CLUSTERING TECHNIQUES

4.1.1. RFM ANALYSIS

RFM Analysis is a way to use data based on existing customer behavior to predict how a new customer is likely to act in the future (Delval, 2021). RFM analysis ranks each customer based on three key metrics: **Recency (R)**, **Frequency (F)** and **Monetary (M)**.

After applying this analysis I got a new dataset called “rfm_df”(5679 rows x 3 columns) (see Table 2), which I scaled in order to have all the values in the same scale “rfm_scaled” and I used for clustering (see Table 3).

	Recency	Frequency	Monetary
Customer_ID			
12346.0	551	11	372.86
12347.0	24	8	3888.01
12348.0	270	4	312.36
12349.0	40	3	2635.04
12350.0	332	1	294.40

	Recency	Frequency	Monetary
Customer_ID			
0	1.572438	0.445491	-0.306621
1	-0.950196	0.184717	0.633554
2	0.227352	-0.162982	-0.322803
3	-0.873607	-0.249907	0.298430
4	0.524133	-0.423756	-0.327606

Table 2. Display of “rfm_df (before scaling). Table 3. Display of “rfm_scaled” (after scaling).

4.1.2. K-MEANS

One of the requirements to run the K-Means algorithm is to define a number of clusters. For this, “The Elbow Method” has been used (see Table 4).

According to the graph (see Table 4), the optimal number of groups is 3, since it is from that moment on when the dispersion of the data does not experience sudden changes.

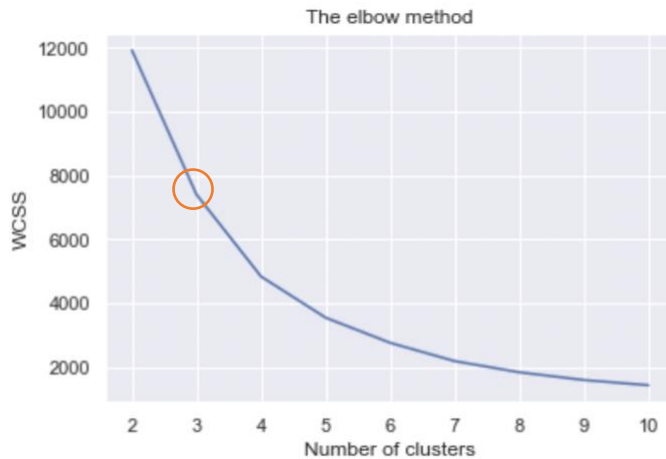


Table 4. Cluster Sum of Squares- "The elbow method".

After running the K-means algorithm on our data and analysing the mean values obtained (see Table 5) for the attributes "Recency", "Frequency" and "Monetary", we got:

1. **Cluster 0** → Formed by 2005 low-level customers, who have not bought here for a long time, and whose purchase frequency is low, as well as the money spent.
2. **Cluster 1** → Formed by 3665 middle-level customers. Mean values for recency, frequency, monetary.
3. **Cluster 2** → Formed only by 9 premium customers. Customers who buy frequently and spend a lot of money.

	Recency	Frequency	Monetary
Cluster_Kmeans			
0	475.167082	2.110723	452.731762
1	84.664939	7.467667	1960.972367
2	66.000000	195.888889	59245.203333

Table 5. RFM mean values per cluster (K-Means).

Let's visualize them with scatterplot (see Table 6). For visualisation I took "Frequency" and "Monetary" variables because are highly correlated (see Table 3A).



Table 6. K-Means Clustering Results.

4.1.2. K-MEDOIDS

For the K-Medoids algorithm, 3 has also been taken as the number of clusters to generate. The results obtained are the following (see Table 7):

1. **Cluster 0** → Formed by 741 premium customers.
2. **Cluster 1** → Formed by 2006 low-level customers.
3. **Cluster 2** → Formed by 2932 middle level customers.

	Recency	Frequency	Monetary
Cluster_Kmedoids			
0	55.394062	22.731444	6652.648092
1	475.297607	2.156032	468.348551
2	91.782742	4.159277	940.920296

Table 7. RFM mean values per cluster (K-Medoids).

Let's visualize them with scatterplot (see Table 8).

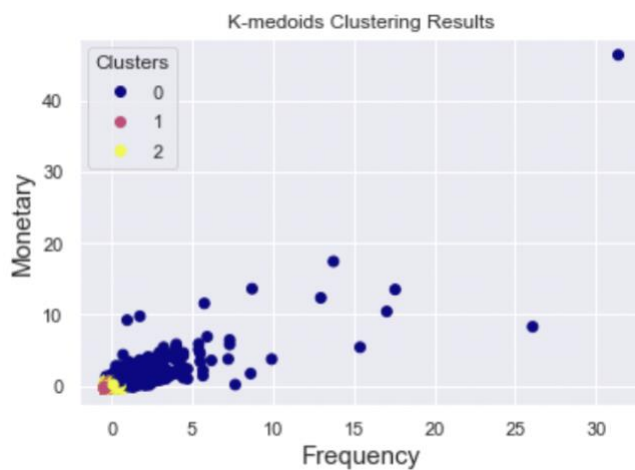


Table 8: K-Medoids Clustering Results.

4.1.3. FUZZY C-MEANS

The results obtained with FCM are the following (see Table 9):

1. **Cluster 0** → Formed by 3341 premium customers.
2. **Cluster 1** → Formed by 448 premium customers.
3. **Cluster 2** → Formed by 1890 middle-level customers.

		Recency	Frequency	Monetary
Cluster_Fuzzy				
	0	95.595331	4.829392	1154.625193
	1	50.176339	29.508929	8713.083607
	2	487.692063	2.121164	458.642070

Table 9. RFM mean values per cluster (Fuzzy C-Means).

Let's visualize them with scatterplot (see Table 10).

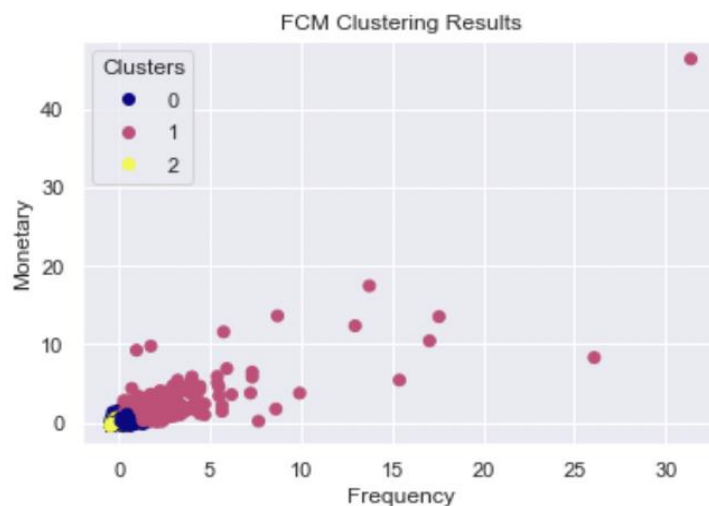


Table 10: Fuzzy C-Means Clustering Results.

4.2. MARKET BASKET ANALYSIS TECHNIQUES

In order to apply these algorithms, the original dataset "df_clean" has been modified. How? Applying one hot encoding to the variable "Item_name" in order to get the number of sales per product for each invoice. Thus, we get the dataset "df_mba" (5 rows x 4840 columns) that looks like this (see Table 11):

	DOORMAT UNION JACK GUNS AND ROSES	3 STRIPEY MICE FELTCRAFT	4 PURPLE FLOCK DINNER CANDLES	50'S CHRISTMAS GIFT BAG LARGE	ANIMAL STICKERS	BLACK PIRATE TREASURE CHEST	BROWN PIRATE TREASURE CHEST	CAMPHOR WOOD PORTOBELLO MUSHROOM	CHERRY BLOSSOM DECORATIVE FLASK	DOLLY GIRL BEAKER	...	ZINC STAR T- LIGHT HOLDER	ZINC SWEET SOAP D
Invoice													
489434	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
489435	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
489436	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
489437	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
489439	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

Table 11. Display of “df_mba”.

These two algorithms are based on association rules. In order to understand how these algorithms work, it is necessary to remember these 3 ratios:

- **Support:** is the probability that an item set appears in the data.
- **Confidence:** the probability that product B (consequent) is purchased given the purchase of product A (antecedent).
- **Lift:** measures how much more likely an item is to be purchased when another item is purchased, compared to its likelihood of being purchased in general.

4.2.1. APRIORI PRINCIPLE ALGORITHM

The item that appears the most in the shopping cart is “white hanging heart t-light holder” with a probability of 1.17% (see Table 12).

	support	itemsets
160	0.117312	(WHITE HANGING HEART T-LIGHT HOLDER)
8	0.064021	(ASSORTED COLOUR BIRD ORNAMENT)
62	0.063661	(JUMBO BAG RED RETROSPOT)
102	0.056678	(PARTY BUNTING)
74	0.056138	(LUNCH BAG BLACK SKULL.)

Table 12. Items support (Apriori Principle).

The association rules obtained can be seen in Table 13.

	antecedents	consequents	confidence	support	lift
2	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.786108	0.021370	25.865591
3	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.703156	0.021370	25.865591
16	(SWEETHEART CERAMIC TRINKET BOX)	(STRAWBERRY CERAMIC TRINKET BOX)	0.685904	0.021730	14.302803
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.670911	0.020531	20.093600
15	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.659574	0.028804	5.622392
0	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.614901	0.020531	20.093600
19	(WOODEN PICTURE FRAME WHITE FINISH)	(WOODEN FRAME ANTIQUE WHITE)	0.597005	0.027485	11.540256
18	(WOODEN FRAME ANTIQUE WHITE)	(WOODEN PICTURE FRAME WHITE FINISH)	0.531286	0.027485	11.540256
6	(LOVE BUILDING BLOCK WORD)	(HOME BUILDING BLOCK WORD)	0.528736	0.023438	9.666157
4	(HEART OF WICKER SMALL)	(HEART OF WICKER LARGE)	0.496010	0.024218	9.945235

Table 13. Association rules (Apriori Principle).

To interpret the table, I take the first line as an example. It means that the probability that a customer will buy "roses regency teacup and saucer" after buying "green regency teacup and saucer" is 78%, that is, a very high probability. The value of the lift ratio is very high, which indicates a positive association. Obviously, the company has to review each of these rules to make strategic sales decisions. In general, the items with the highest % confidence are those related to decoration and ceramics.

4.2.2. FP-GROWTH

Using FP-Growth algorithm, we obtained that the item that appears the most in the shopping cart is "white hanging heart t-light holder" with a probability of 1.17%, the same result as Apriori Principle algorithm (see Table 14).

	support	itemsets
26	0.117312	(WHITE HANGING HEART T-LIGHT HOLDER)
3	0.064021	(ASSORTED COLOUR BIRD ORNAMENT)
371	0.063661	(JUMBO BAG RED RETROSPOT)
178	0.056678	(PARTY BUNTING)
142	0.056138	(LUNCH BAG BLACK SKULL.)

Table 14. Items support (FP-Growth).

Let's see if happen the same with the association rules obtained (see Table 15).

	antecedents	consequents	confidence	support	lift
350	(POPPY'S PLAYHOUSE LIVINGROOM , POPPY'S PLAYHO...	(POPPY'S PLAYHOUSE KITCHEN)	0.922460	0.010340	50.703380
274	(PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	(GREEN REGENCY TEACUP AND SAUCER)	0.893170	0.015286	32.855259
345	(POPPY'S PLAYHOUSE LIVINGROOM)	(POPPY'S PLAYHOUSE KITCHEN)	0.882743	0.011959	48.520345
348	(POPPY'S PLAYHOUSE KITCHEN, POPPY'S PLAYHOUSE ...	(POPPY'S PLAYHOUSE BEDROOM)	0.864662	0.010340	52.836211
343	(POPPY'S PLAYHOUSE BEDROOM)	(POPPY'S PLAYHOUSE KITCHEN)	0.847985	0.013877	46.609857
273	(PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	(ROSES REGENCY TEACUP AND SAUCER)	0.845771	0.015286	27.828707
269	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.835180	0.018073	30.722103
346	(POPPY'S PLAYHOUSE LIVINGROOM)	(POPPY'S PLAYHOUSE BEDROOM)	0.827434	0.011210	50.561347
171	(KITCHEN METAL SIGN)	(BATHROOM METAL SIGN)	0.791353	0.012618	29.632676
271	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.790859	0.017114	26.021904

Table 15. Association rules (FP-Growth).

It means that the probability that a customer will buy " poppy's playhouse kitchen" after buying " poppy's playhouse livingroom" is 92%, that is, a very high probability, probably because there are pieces of the same toy. The items with the highest % confidence are those related to toys and ceramic plates.

5. EVALUATION OF MODELS

5.1. EVALUATING AND CONCLUSIONS FOR CLUSTERING ALGORITHMS

For evaluating the performance of K-Means, K-Medoids and Fuzzy C-Means algorithms I used Silhouette coefficient which returns the following scores:

- Silhouette score K-means: 0.5308913216480491
- Silhouette score K-medoids: 0.539288244348331
- Silhouette score FCM: 0.5698915034888655

Based on these scores, the algorithm which performs better is Fuzzy C-Means.

Thanks to the clustering algorithms, it has been possible to segment customers into 3 clusters. It is important to adopt strategies that encourage the frequency of purchase and the money spent to increase, as well as taking care of premium customers.

5.2. EVALUATING AND CONCLUSIONS FOR MBA ALGORITHMS

For evaluating the performance of Apriori Principle and FP-Growth algorithms I used time function in order to know the execution time of each algorithm. I got the next results:

- Apriori algorithm execution time: 7.606765985488892
- FP-Growth algorithm execution time: 5.1822898387908936

The algorithm which performs better is FP-Growth because took less time and also because founded more frequent itemsets (688) that Apriori Principle (180).

Thanks to the application of MBA algorithms, sufficient product association rules have been obtained for strategic decision making regarding the product.

6. RESEARCHING- ADDITIONAL MBA APPLICATION

The MBA technique can be applied to the health sector. How? There are different ways to apply it. For example, the MBA can analyse patient diagnostic data and find patterns in the diseases that are most frequently diagnosed together. Also analyse patient treatment data and find patterns in the most frequently prescribed treatments.

APPENDIX

a) Appendix 1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1067371 entries, 0 to 1067370
Data columns (total 8 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Invoice          1067371 non-null object
1   StockCode       1067371 non-null object
2   Description     1062989 non-null object
3   Quantity        1067371 non-null int64
4   InvoiceDate      1067371 non-null datetime64[ns]
5   Price           1067371 non-null float64
6   Customer ID     824364 non-null float64
7   Country         1067371 non-null object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 65.1+ MB
```

Table 1A. Data types of data columns

b) Appendix 2

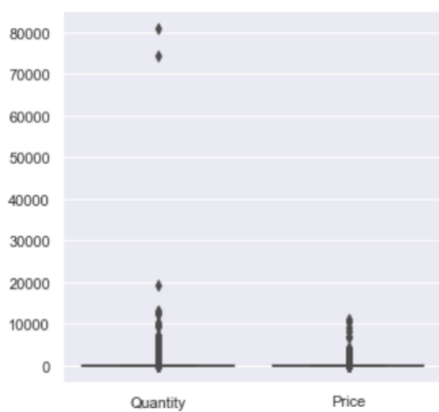


Table 2A. Outliers in “Quantity” and “Price” variables

c) Appendix 3

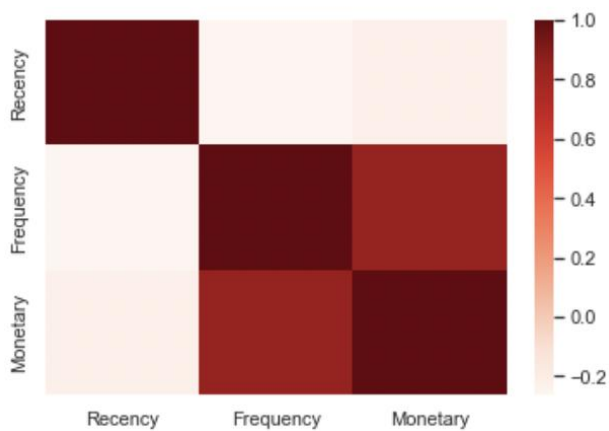


Table 3 A. Heatmap of RFM features

REFERENCE LIST

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

Hotz, N. (2022). *CRISP-DM*. [online] Data Science Project Management. Available at: <https://www.datascience-pm.com/crisp-dm-2/>.

Rao, A.B., Kiran, J.S. and G, P. (2021). Application of market–basket analysis on healthcare. *International Journal of System Assurance Engineering and Management*. doi:<https://doi.org/10.1007/s13198-021-01298-2>.

SERT, B. (n.d.). *ASSOCIATION & RFM ANALYSIS*. [online] kaggle.com. Available at: <https://www.kaggle.com/code/baturalpsert/association-rfm-analysis/notebook> [Accessed 16 Apr. 2023].