College Dublin
Computing • IT • Business

BUSINESS REPORT

Customer Transactions – Online Retail

Insight Report

2023

PREPARED BY:

Bobbi McDermott

CONTENTS

INTRODUCTION

The dataset is very large comprising of 1067371 invoice purchases detailing products, quantity, price, customer and country of purchase. This data spans a 2-year period, 2009 – 2011. This information will help establish buying trends. As there is no information on the customers themselves, profiling them is not an option. A more prudent approach would be to perform an RFM analysis, to establish patterns in Recency, Frequency and Monetary which benefits from increased customer retention, response rate, conversion rate and revenue (Correia, 2016) with the correct marketing strategy

SECTION 1: BUSINESS UNDERSTANDING

This is an online retail business supplying an assortment of household and garden products. The sales are sporadic and concentrated primarily in the UK. Objectives would be to either boost sales across the geographic portfolio or limit their catchment area. The reason for this is inventory and operational costs cannot be justified if sales are low in these areas. Secondly, for the machine learning algorithm can produce customer segmentation which is an effective way of preventing customer attrition by spotting and seizing up- and cross-selling opportunities (Dessi, 2020). Lastly, Market Basket Analysis will create more efficient cross-selling, up-selling, and product placement tactics(quantzig, 2021)

SECTION 2: DATA UNDERSTANDING

The data presented represents a collection of sales to 43 different countries worldwide. The data shows that the sales are disproportionate in terms of region and customer engagement. By a large majority the United Kingdom dominate the online transactions (figure 1) large datasets tend to yield more accurate models ("IBM Documentation," 2021a), however there is a lot of misleading, missing data that needs to be refined, relevant attributes prioritized, for a more accurate result. The Price data is inaccurate and there are lot of missing Customer ID's which are removed as it lessens the accuracy of the model and cannot determine data touch points for business process (Srinivasan, 2019)

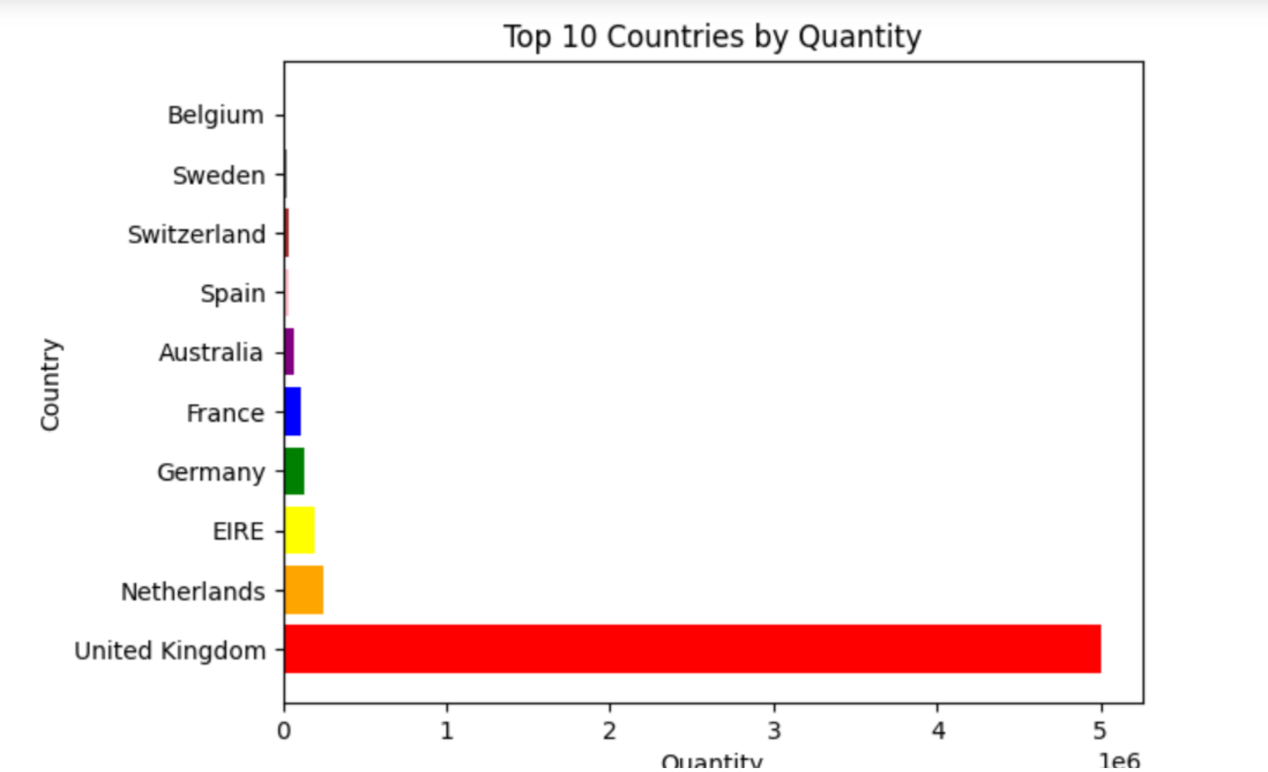


Figure 1

SECTION 3: DATA PREPARATION

1. Data Cleaning

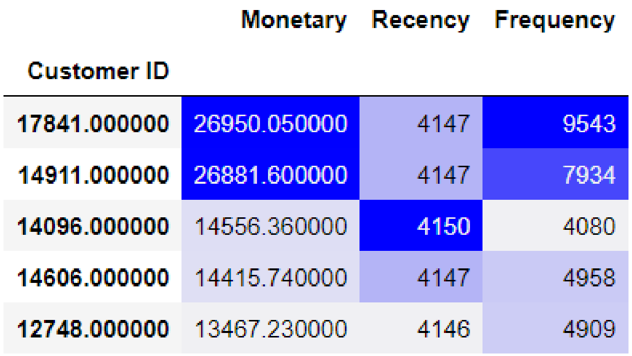
- The two sheets were merged, and year column added
- Values under 1 were removed in the Quantity and Price features
- Null values in Customer ID were dropped
- This equalized the Stock Code and Description
- Removed time portion in InvoiceDate

- Used statistical method (IQR) to reduce outliers in the Price column
- Encoded ‘Description’, ‘Quantity’ and ‘Year’ for a numeric dataframe for clustering
- Label Encoder for Customer ID

2. Recency / Frequency / Monetary

- Derived Recency from Invoice Date to current date
- Derived Frequency from Customer ID
- Derived Monetary from Price
- Merged these into a new dataframe, Figure 2

RFM analysis is pivotal in determining marketing strategies in Campaign improvement, segmentation, and more in-depth target group research(Correia, 2016)



	Monetary	Recency	Frequency
Customer ID			
17841.000000	26950.050000	4147	9543
14911.000000	26881.600000	4147	7934
14096.000000	14556.360000	4150	4080
14606.000000	14415.740000	4147	4958
12748.000000	13467.230000	4146	4909

Figure 2

SECTION 4: MODELING

Clustering

I used 3 forms of clustering for prediction to ensure I got the most accurate results.

1. Kmeans

This method indicated that purchases were being made were in the middle of the spend category but were increasing in terms of recency. The same pattern was noticed in KMenoids.

2. KMedoids

KMenoid is more robust to outliers therefore I proceeded as the 'Price' column did propose problems. It validated that spending was higher with more recent spenders.

3. Hierarchical

Hierarchical clustering established a more robust result in that it showed the separation between the clusters clearly. It is also a good algorithm to use for businesses as it is continuous in monitoring performance. It affirmed that the patterns of transactions seen in the previous algorithms were correct. This method of clustering is also more interpretable for data that has skewed values such as the 'Price' column ("Hierarchical Cluster Analysis · UC Business Analytics R Programming Guide," n.d.)

Market Basket Analysis

I performed 2 Market Basket Analyses to establish buying patterns to help with marketing for desirable items. Items sold the most as per figure 3.

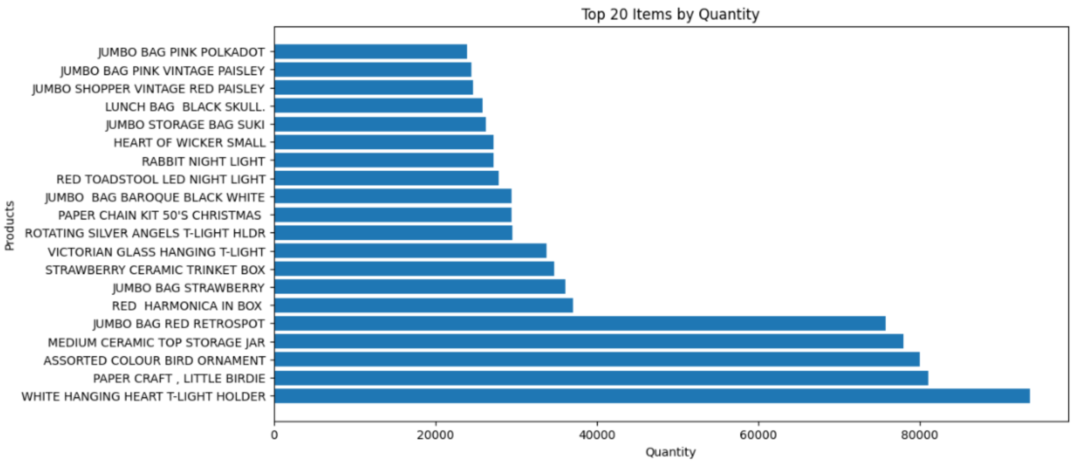


Figure 3

4. Apriori

I segmented the dataset into high, medium and low spend from the RFM dataframe and fed it into the cleaned dataframe. This was in part to reduce the data to run through MBA as it won’t perform, but it’s helpful in segmented marketing strategies. I also used a sample of the data as it was again too big.

5. FP Growth

FP Growth performs the same type of analysis on the data, but compacts items set, like a tree format. So that it breaks item sets down to more manageable patterns (Verma, 2021)

6. Conclusion

Both algorithms performed well and produced similar items in each deployment. There were different levels of confidence, for example SET 3 RED SPOT TIN TEA, COFFEE, SUGAR and RECYCLING BAG RETROSPOT were frequent items in low spend in both algorithms

SECTION 5: EVALUATION

There was very minimal correlation between spending and recency and frequency and recency. Therefore, it would suggest a marketing strategy toward preventing churn would be the most advantageous.

Ideally it would be logical to find other retailers that supply similar products to those produced by Market Basket Analysis and analysis their costing and marketing strategy. (Shubhanshugupta, 2021)

SECTION 6: DEPLOYMENT

7. Planning

We need to take the results of the Models to help plan infrastructure changes and marketing strategies. (“Crisp DM methodology,” n.d.)The Clustering is indicative that sales are consistent, however vary between recency and frequency. It centers around the middle spenders. MBA has been broken down so the rules here will assist with planning (IBM, 2021)

8. Monitoring

The Accuracy of the models performed well so data updates would need to be done regularly to update the rest of the CRISP DM strategy (“IBM Documentation,” 2021b)

SECTION 7: CONCLUSION

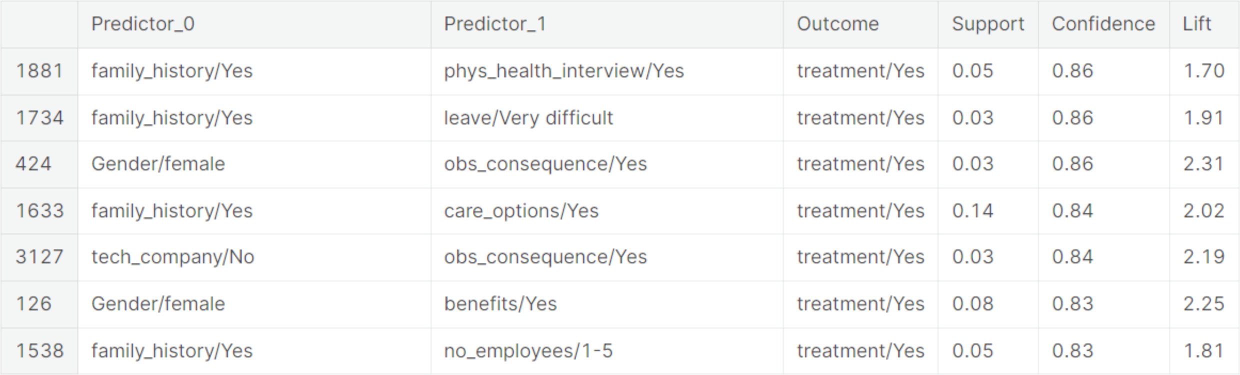
The data was difficult to work with as the price column provided very little assurance in the validity of the data overall. However, the trends that emerged from the machine learning algorithms were uniformed and therefore would provide a reasonable assumption to a marketing strategy going forward.

SECTION 7: MENTAL HEALTH SURVEY- CASE STUDY

Mental Health Survey

This study is a good example of an Association Rules analysis. It was a survey taken that comprises of information such an age, gender, location, history, opinions and work. It was taken to assess frequency of mental health by location and predictors and attitudes in the workplace

It cleans the dataset by gender and performs association rules on the basis that the data prediction and outcome. It takes four phases, count, support, confidence and lift to create tuples for this prediction. The predictions are categorized in a binary format as per Association Rules, it runs through the data and indicators and formulates the associations between the indicators. It then supplies a table with the Support / Confidence and Lift of these associations. The Author then has an 'Outcome' column where by parameters are outlined to decide whether treatement is required or not, Figure 4.



	Predictor_0	Predictor_1	Outcome	Support	Confidence	Lift
1881	family_history/Yes	phys_health_interview/Yes	treatment/Yes	0.05	0.86	1.70
1734	family_history/Yes	leave/Very difficult	treatment/Yes	0.03	0.86	1.91
424	Gender/female	obs_consequence/Yes	treatment/Yes	0.03	0.86	2.31
1633	family_history/Yes	care_options/Yes	treatment/Yes	0.14	0.84	2.02
3127	tech_company/No	obs_consequence/Yes	treatment/Yes	0.03	0.84	2.19
126	Gender/female	benefits/Yes	treatment/Yes	0.08	0.83	2.25
1538	family_history/Yes	no_employees/1-5	treatment/Yes	0.05	0.83	1.81

Figure 4

The summary of this data is enormously beneficial to caregivers as a 'guidance'. Guidelines are often formulated using data such as this and lead to improved patients outcomes (Setkowski et al., 2021) This is an excellent example of how Association rules can be used in predictions of various indicators in the medical field and can show how it could be used in other scenarios whereby indicators are present, such as criminology for offender detection (Sevri et al., 2017)

BIBLIOGRAPHY

Correia, J., 2016. How RFM Analysis Boosts Sales | Blast Analytics & Marketing. Blast Analytics. URL https://www.blastanalytics.com/blog/rfm-analysis-boosts-sales (accessed 4.16.23).

Crisp DM methodology, n.d. . Smart Vision Europe. URL https://www.sv-europe.com/crisp-dm-methodology/ (accessed 4.17.23).

Dessi, G., 2020. RFM Customer Segmentation with K Means [WWW Document]. URL https://rstudio-pubs-static.s3.amazonaws.com/671942_d7c20cc5f25d4fc2ac33c3556fc13e88.html (accessed 4.16.23).

Hierarchical Cluster Analysis · UC Business Analytics R Programming Guide [WWW Document], n.d. URL https://uc-r.github.io/hc_clustering (accessed 4.17.23).

IBM, 2021. IBM Documentation [WWW Document]. URL https://www.ibm.com/docs/en/spss-modeler/18.2.2?topic=deployment-planning (accessed 4.16.23).

IBM Documentation [WWW Document], 2021a. URL https://www.ibm.com/docs/en/spss-modeler/saas?topic=understanding-describing-data (accessed 4.16.23).

IBM Documentation [WWW Document], 2021b. URL https://www.ibm.com/docs/en/spss-modeler/18.2.2?topic=deployment-planning-monitoring-maintenance (accessed 4.16.23).

quantzig, 2021. Benefits of Market Basket Analysis for Business to Gain a Winning Edge. Quantzig. URL https://www.quantzig.com/blog/benefits-market-basket-analysis/ (accessed 4.16.23).

Setkowski, K., Boogert, K., Hoogendoorn, A.W., Gilissen, R., van Balkom, A.J.L.M., 2021. Guidelines improve patient outcomes in specialised mental health care: A systematic review and meta-analysis. Acta Psychiatr Scand 144, 246–258. https://doi.org/10.1111/acps.13332

Sevri, M., Karacan, H., Akcayol, M., 2017. Crime Analysis Based on Association Rules Using Apriori Algorithm. International Journal of Information and Electronics Engineering 7, 99–102. https://doi.org/10.18178/IJIEE.2017.7.3.669

Shubhanshugupta, 2021. Become a Story Telling Ninja: Present Data Science Models to Stakeholders - Shubhanshu Gupta. URL https://shubhanshugupta.com/art-of-story-telling/ (accessed 4.16.23).

Srinivasan, S., 2019. Business and Data Understanding in Data Science Lifecycle. Medium. URL https://medium.com/@srivatsan88/business-and-data-understanding-in-data-science-lifecycle-58f8e0588c66 (accessed 4.16.23).

Verma, Y., 2021. Apriori vs FP-Growth in Market Basket Analysis - A Comparative Guide [WWW Document]. Analytics India Magazine. URL https://analyticsindiamag.com/apriori-vs-fp-growth-in-market-basket-analysis-a-comparative-guide/ (accessed 4.16.23).