(Customer Segmentation)

An analysis based on a retail database

Project Report

Princi Sharma | Maitri Bhatt
ID: 001560408 | 001568126
psharma5@albany.edu | bbhatt@albany.edu

December 12, 2022

College of Arts Sciences
University at Albany (SUNY)
United States

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science

Professor Justin Curry

AMAT 502 - Modern Computing for Mathematicians

Contents

1 Abstract **i**

2 Introduction **i**
2.1 Business and Data Mining Goals i
2.2 Business Goals i
2.3 Data Mining Goals i

3 Data Description and Data Preparation **ii**

4 Exploratory Data Analysis and Data Visualization **ii**

5 Data Mining Solutions **vi**
5.1 K-Means vi
5.2 Association Rules x

6 Recommendation **xi**

7 Conclusion **xii**

1 Abstract

Retail businesses sell products or services to customers for their consumption. They sell products and services in-store, but some items may be sold online or over the phone and then shipped to the customer. In the retail business, customer satisfaction and profit making are the most important factors to consider and these will be backed up by strategic pillars such as marketing, management, finance, and operations.
Customer segmentation is defined as the process of grouping the customers into subsets with similar characteristics so that the businesses can market to those groups efficiently. This can be done based on transactions, demographic, geographic, and psychographic segmentation. Identifying customer needs, the businesses can get benefits such as market expansion, target production, marketing, market competitiveness and customer satisfaction.
Our goal is to review transaction history from an online business to analyze purchases. We utilized association algorithms and K-means clustering to identify items customers purchased frequently and segment the customers. Based on this analysis, we recommended advertising options to increase revenue and maintain customer satisfaction.

2 Introduction

2.1 Business and Data Mining Goals

It is of interest to retailers to increase their marketability and learn about their customers by viewing their purchase practices. In e-commerce, it is imperative to know what type of customers purchase which products. Classifying customers by knowing their purchasing practices would aid in marketing initiatives and will help to determine marketing strategies. Some of those marketing strategies are ways to improve the revenue, understand customer purchasing practices, target marketing and seek customer segments for company strategy management.
It is of interest to retailers to increase their marketability and learn about their customers by viewing their purchase practices. In e-commerce, it is imperative to know what type of customers purchase which products. Classifying customers by knowing their purchasing practices would aid in marketing initiatives and will help to determine marketing strategies. Some of those marketing strategies are ways to improve the revenue, understand customer purchasing practices, target marketing and seek customer segments for company strategy management.

2.2 Business Goals

1. To help improve the profits for the organization by identifying customer buying behaviors to predict the future buying patterns.

2. Recommend marketing strategies to improve the purchasing behaviors for categories.

2.3 Data Mining Goals

1. Segment customers based on several factors and correlate based on variables.

2. Identify the outliers and clean the dataset and develop models to identify future customer buying patterns.

3 Data Description and Data Preparation

The dataset was retrieved from the kaggle website to aid in segmenting customers for an e-commerce store, from the year 2010 to 2011. This dataset has categorical and numerical variables containing approximately 500,000 data rows with about 4000 different customers in 37 countries.
Additional columns were created such as Total Revenue, Canceled Quantity, Canceled Cost, and Return Customer for analysis and in building the data model.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Qty_cancelled	Cancelled_Cost	Total_Revenue	Return_Customer
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom	0	0	15.3	1
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom	0	0	20.34	1
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom	0	0	22	1
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom	0	0	20.34	1
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom	0	0	20.34	1
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom	0	0	15.3	1
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingdom	0	0	25.5	1
536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom	0	0	11.1	1
536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingdom	0	0	11.1	1

Figure 1: Dataset after preprocessing

Initially missing data were filled with the NA label while some were replaced with 0. We then created new columns that we believe will be good predictors of the algorithm in Excel using Vlookup and PivotTables.

4 Exploratory Data Analysis and Data Visualization

Exploratory Data Analysis was done to identify the buying patterns with the highest and the lowest demands, any outliers, trends, patterns, and relationships among the variables.

United Kingdom	349201
Germany	9025
France	8326
EIRE	7226
Spain	2479
Netherlands	2359
Belgium	2031
Switzerland	1841
Portugal	1453
Australia	1181
Norway	1071
Italy	758
Channel Islands	747
Finland	685
Cyprus	603
Sweden	450
Austria	398
Denmark	380

Figure 2: Country-wise Quantity analysis

Our largest customer pool is from UK/Europe acquiring 90% of our customer base. The UK customers have placed bulk orders and taken up first place quantity-wise.

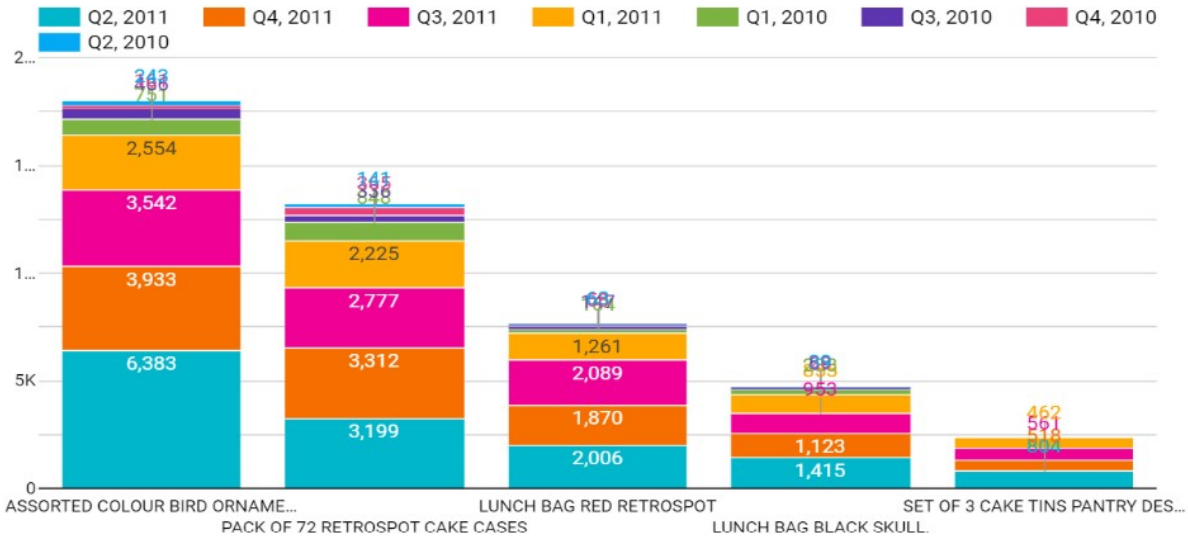

Figure 3: Customer-based Quantity analysis

With this chart we can derive that bulk orders are generally low with a lower frequency and do not contribute majorly to the revenue, implying that we need to target frequent consumers. As they generate revenue consistently and can be labeled as loyal buyers.

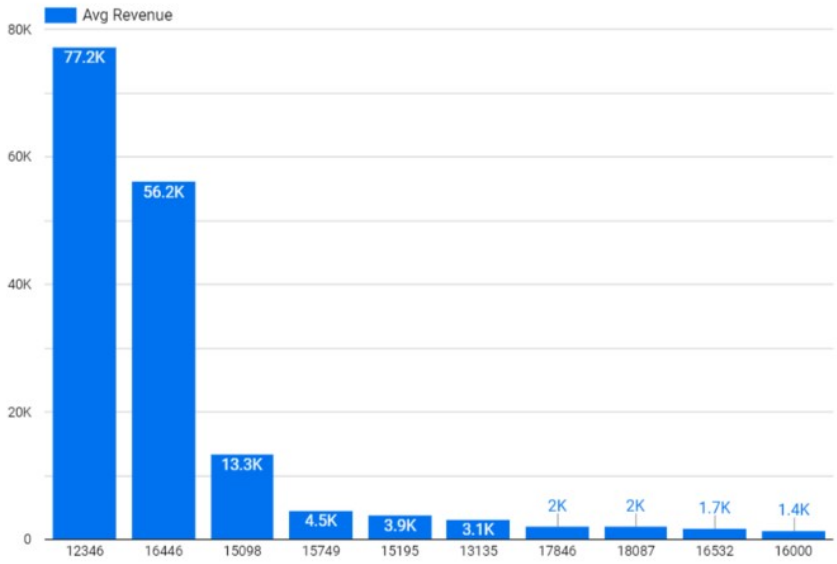

Figure 4: Customer-based Average Sales

When we do customer-base analysis, the above chart represents customers vs their average purchase amount. We can conclude that there is a significant gap between the purchasers' buying habits and their individual revenue patterns.

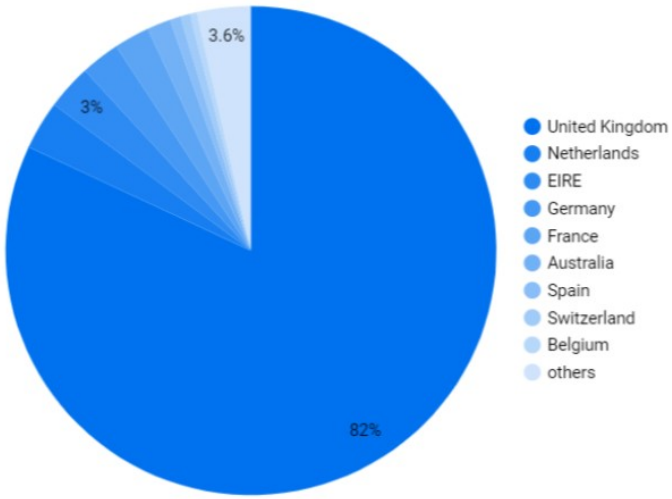

Figure 5: Country-wise Revenue collection

The top 10 customers are Europe based, placing bulk orders generating 82% of revenue, and the second highest contributors(quantity-wise) are from Germany at 2.6%, the Netherlands at 3.2% revenue-wise. Overall we have sold over 5 million products and generated 8 million in revenue so

by this we can calculate that an average sale per consumer is $22.

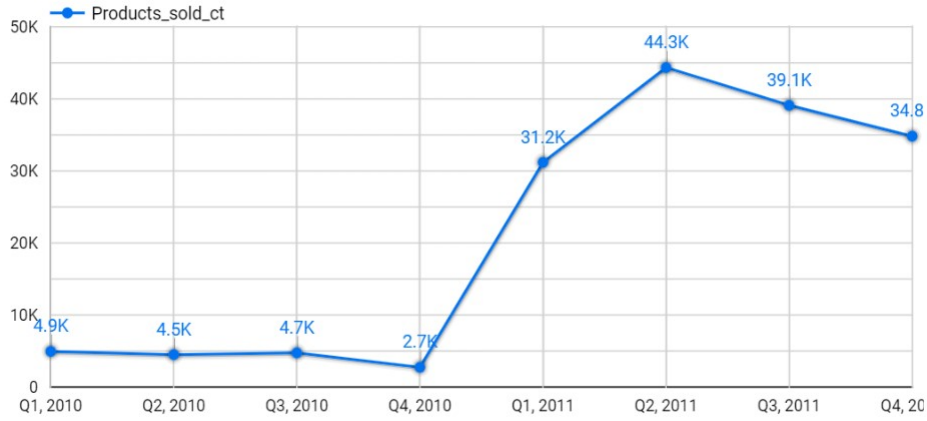

Figure 6: Quarterly Sales

The yearly comparison shows nearly 10X more sales in 2011 and the second quarter of 2011 month of June was the highest revenue-generating month.

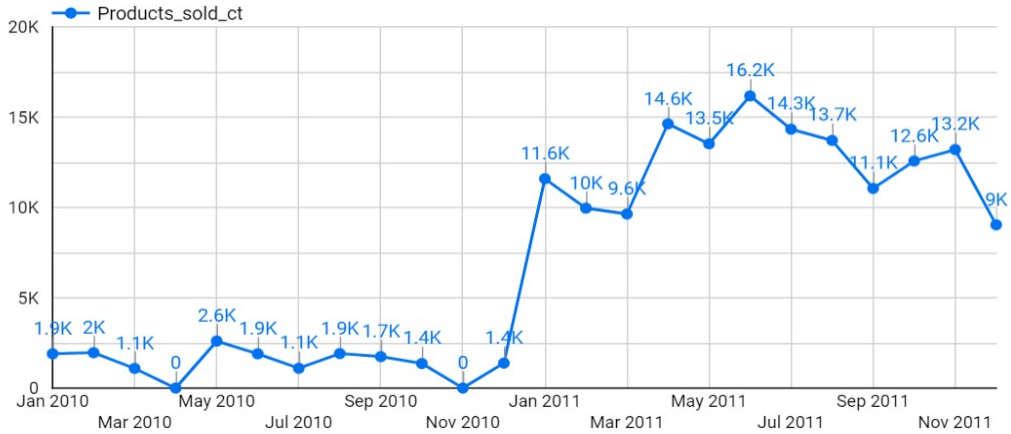

Figure 7: Monthly Sale

Description	TotalRevenue
JUMBO BAG RED RETROSPOT	1790
REGENCY CAKESTAND 3 TIER	1401.6
RECYCLING BAG RETROSPOT	1295
JUMBO BAG PINK POLKADOT	1253
JUMBO BAG STRAWBERRY	1074
SET OF 3 CAKE TINS PANTRY DESIGN	1020
SET OF 3 NOTEBOOKS IN PARCEL	974.4
TOY TIDY PINK POLKADOT	925

While accumulating the top 8 sellers of June which evidently shows the resemblance in the category and description. By this, we can rule out association analysis as one of the key algorithms for customer segmentation.
Thereafter the data cleansing was done by removing null values, and duplicated values, renaming columns, removing additional unwanted data such as canceled and discounted items, and normalizing the variables needed for the models; K-Means clustering and Association Rules were defined during our data analysis.

5 Data Mining Solutions

For this project, we mainly focused on developing K-Means clustering and Association Rules.

5.1 K-Means

The K-means clustering belongs to the partition based clustering family of algorithms where each sample in a dataset is assigned to exactly one cluster. Based on this Euclidean distance metric, we can describe the K-means algorithm as a simple optimization problem, an approach for minimizing the within-cluster sum of squared errors (SSE), which is called cluster inertia. An RFM analysis was conducted while doing the K-mean clustering. This is a strategy for consumer segmentation that divides customers into categories based on historical purchasing history. In order to identify clients who are more likely to respond to promotions and also for future personalized services, RFM helps group customers into different clusters.

1. RECENCY (R): Denotes how many days have passed since a consumer last made a transaction. This could be the last visit date/the last login time if it relates to a website.
2. FREQUENCY (F): The purchased quantity during a specific time. We can interpret how frequently and how many customers used a company's product.
3. MONETARY VALUE (M): Total amount of money a customer spent in that given period.

When performing the data pre-processing, we checked the K-means assumptions before implementing the K-means clustering model. ie. symmetric distribution of variables (not skewed), variables with same average values, and variables with same variance.

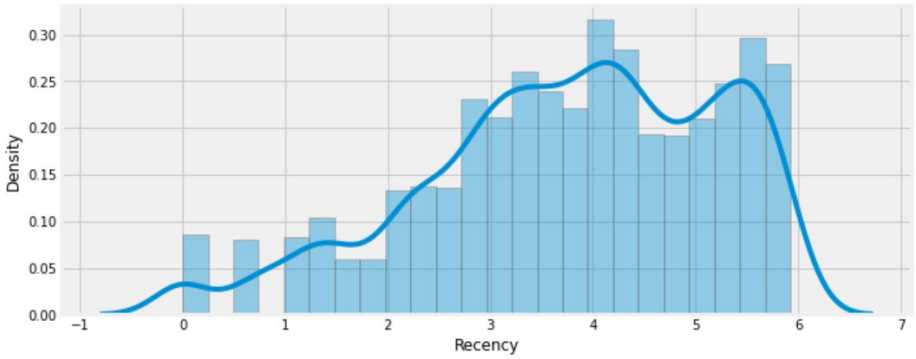

Figure 8: Skewness in Recency

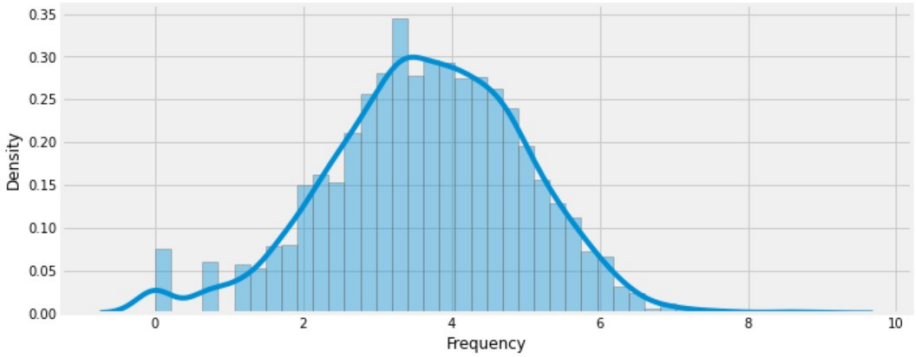

Figure 9: Skewness in Frequency

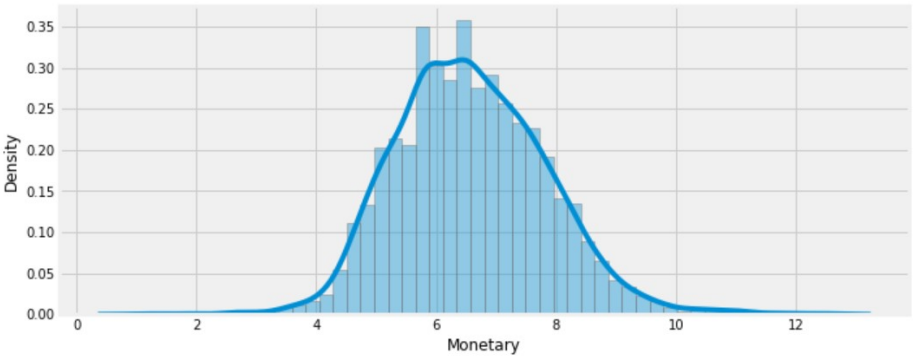

Figure 10: Skewness in Monetary

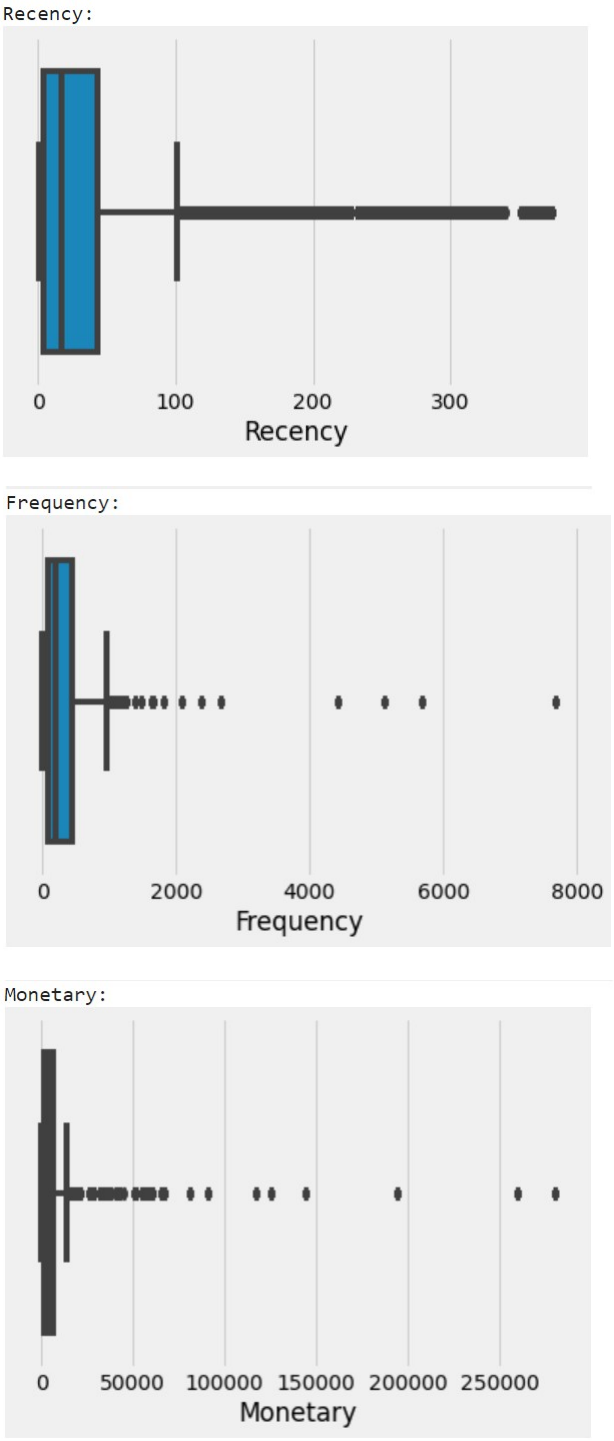

Figure 11: Outliers in RFM Profile

Thereafter, we performed K-means clustering that involved data pre-processing, choosing a number of clusters using the Elbow method, running K-means clustering on pre-processed data, and analyzing average RFM values.

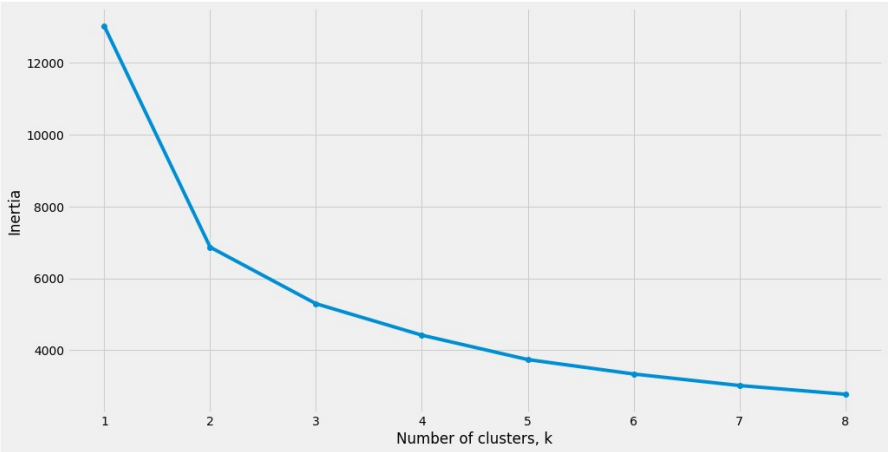

Figure 12: K-means Clustering

Three clusters with comparable buying patterns and customer attributes were largely found. Invoice date indicates recentness, invoice number indicates customer frequency, and total revenue indicates monetary worth. Data were normalized based on RFM segmentation and scores. Finally, clients are classified as Gold customers if their RFM score is greater than 10, Silver customers if it is between 5 and 10, and Bronze customers if it is less than 5.

	CustomerID	Customer_Segmenation	K_Cluster	Metric	Value
0	12347.0	Gold	1	Recency	-2.148776
1	12348.0	Sliver	2	Recency	0.383413
2	12349.0	Sliver	2	Recency	-0.575940
3	12350.0	Bronze	0	Recency	1.375606
4	12352.0	Sliver	2	Recency	-0.128755

Figure 13: Customer Segmentation using K-means

Cluster 0 - These are the customers with the low RFM scores. They will make occasional purchases and visit the platform when they have a specific product they'd like to buy. They are new users with the potential to become long-term consumers with high frequency and monetary value and can be targeted with special "new-user promotions" to instill brand loyalty.

Cluster 1 - We can say that these customers make purchases often and visit the platform frequently. Their monetary value is extremely high, indicating that they spend a lot when shopping

online. This could mean that users in this segment are likely to make multiple purchases frequently and are highly responsive to cross-selling and up-selling but their low recency suggests that they have not been extremely active on the platform recently, this could mean several things that they were disappointed with the service and switched to a competitor platform, they no longer have any interest in the products sold, or their customer ID changed as they re-registered onto the platform with different credentials.

Cluster 2 - Extremely low values in RFM which suggests the customer visited the platform in the past but did not like the products to purchase. Hence they are the old customers and the weakest segment that can be rolled out from the marketing segments.

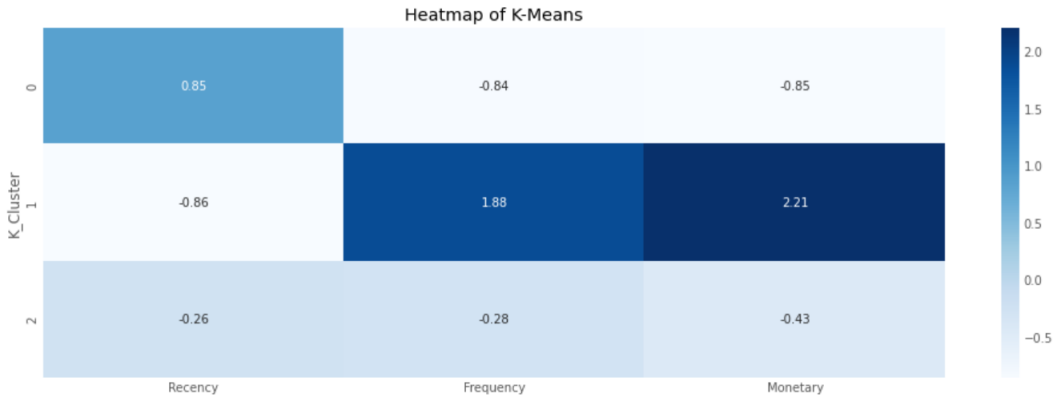

Figure 14: Heatmap of K-means

5.2 Association Rules

Association rules help to find the relationship between items. This primarily tries to find the rules that govern how or why such products/items are often bought together. Our aim is to identify which customers purchased items together and give an analysis on the frequent item set.

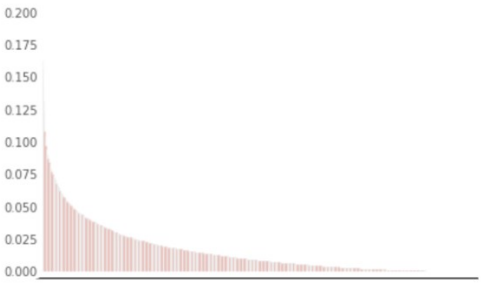

Figure 15: Items Vs Frequency

In this dataset the Apriori algorithm was used to generate frequent item sets according to the CustomerID and Description. Then the dataset was transposed to a binary incidence matrix to make

it easier to find the frequently purchased data. The rules indicate a strong dependence between the antecedent and consequent item sets and measure the strength of association implied by a rule. The rules were verified using confidence and lift ratio metrics. It helped to determine if there was a high value of confidence that could suggest strong association rules. Also, the lift ratio was used to compare confidence of the rule with a benchmark value. The larger the lift ratio, the greater the strength of the association.

While it would be ideal to utilize the description and lift ratio to verify the validity of the rules, we were unable to do so. The description included the quantity of the items as part of the description and further text analysis should be used to present clearer description of items, which is beyond the scope of this project. Therefore, to circumvent this error the antecedent support metric was used to illustrate that customers who bought one item, and also bought another item. The results showed that customers who bought party bunting also bought spotty bunting. In addition, 3 sets of cake tins were followed by 6 sets of spice tins, also the bulk cake cases were purchased. These items should be suggested to customers to purchase if they purchase one of those items in the item sets.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
249	((index, PARTY BUNTING))	((index, SPOTTY BUNTING))	0.163246	0.131658	0.068942	0.422316	3.207682	0.047449	1.503145
271	((index, SET OF 3 CAKE TINS PANTRY DESIGN))	((index, REGENCY CAKESTAND 3 TIER))	0.147567	0.203136	0.059258	0.401563	1.976818	0.029281	1.331575
265	((index, SET OF 3 CAKE TINS PANTRY DESIGN))	((index, RECIPE BOX PANTRY YELLOW DESIGN))	0.147567	0.108600	0.061102	0.414063	3.812716	0.045076	1.521322
289	((index, SET OF 3 CAKE TINS PANTRY DESIGN))	((index, SET OF 6 SPICE TINS PANTRY DESIGN))	0.147567	0.090616	0.062716	0.425000	4.690140	0.049344	1.581538
130	((index, JUMBO BAG RED RETROSPOT))	((index, LUNCH BAG RED RETROSPOT))	0.146415	0.122665	0.062716	0.428346	3.491990	0.044756	1.534731
...	...	...	...	...	...	...	...	...	...
535	((index, LUNCH BAG PINK POLKADOT), (index, LUN...	((index, LUNCH BAG CARS BLUE))	0.047729	0.110445	0.040120	0.840580	7.610844	0.034848	5.579936
488	((index, LUNCH BAG WOODLAND), (index, LUNCH BA...	((index, LUNCH BAG RED RETROSPOT))	0.047268	0.122665	0.040581	0.858537	6.999010	0.034783	6.201848
526	((index, LUNCH BAG PINK POLKADOT), (index, LUN...	((index, LUNCH BAG RED RETROSPOT))	0.046807	0.122665	0.040581	0.866995	7.067965	0.034840	6.596256
536	((index, LUNCH BAG CARS BLUE), (index, LUNCH B...	((index, LUNCH BAG RED RETROSPOT))	0.046345	0.122665	0.040120	0.865672	7.057177	0.034435	6.531268
303	((index, ALARM CLOCK BAKELIKE PINK), (index, A...	((index, ALARM CLOCK BAKELIKE RED))	0.044962	0.089463	0.041042	0.912821	10.203357	0.037020	10.444398

545 rows × 9 columns

Figure 16: Frequent Item Sets

6 Recommendation

Based on the data analysis, we found groupings, which we categorized as bronze, silver and gold. Marketing teams can utilize these groupings to promote their products to customers that fit that profile.
Bronze customers: The lowest profitable customers who rarely buy products. This can be due to their income level or the demand for the products. These customers will shop rarely and normally purchase in small quantities. We can recommend the buy one get one free method, to target these groups of customers which may encourage them to purchase more items based on the sales or promotions we've offered. These customers may also be willing to purchase free trial products which is another alternative we can try to convert them into a silver/gold customer.

Silver customers - Customers in this category, patronize more often and buy in small quantities or they purchase in large quantities on rare occasions. Visit websites more frequently but make modest purchases, or they visit infrequently but make major purchases once a month to acquire

the things all at once. This purchasing pattern creates an overall average profit for the business from this group of customers. The business can provide a reward system for these clients to buy more products with loyalty points. The business can also use a recommendation system for these customers which will pop up the associated products with the products based on the association rule models by figuring out the products they often buy which will allow them to buy more products.

Gold customers: The most profitable customers for the business. This group needs to be kept satisfied continuously as they purchase large quantities of products on a regular basis. The business can provide subscriptions at a lower price that will show them more products on the website for them to buy more frequently. And the business can become more customized for these customers by offering the best products at a discounted rate on special days. These are some recommendations for the customer categories to improve the business revenue.

7 Conclusion

We explored the data to find the avenues to assist the online retail business and provided solutions based on association rules and clustering. To maintain the data model, the dataset and the model could be run bi-weekly and should be analyzed on an on-going basis. When the accuracy of the model drops lower than 75%, the data and algorithm used should be updated to improve the model's accuracy and predictive power.
Though we provided some recommendations based on the results from our data mining algorithm, it is important to know that our data and analysis had some limitations. The missing values in the description and datasets, which could have contributed to the overall accuracy of the data. Moreover, the model was built based on data from 2010 and 2011, which is 10 years ago which could lead to the change in the model as well. Therefore, this model might change based on the latest version of data which were not taken into consideration in our model.
Lastly, we have limited data to help explain customer behavior such as customer feedback. Therefore, this model could not fully explain the extent of customer behavior for purchasing items. In future models, customer feedback and other fields can be used in conjunction with this dataset to explore the customer's behavior. Nevertheless, the model and results along with the recommendations presented are a good foundation to make suggestions on how the business can increase revenue. As the business problem and goals of this retail company change, they should update the parameters and measures used in the algorithm so that it properly reflects these new goals.