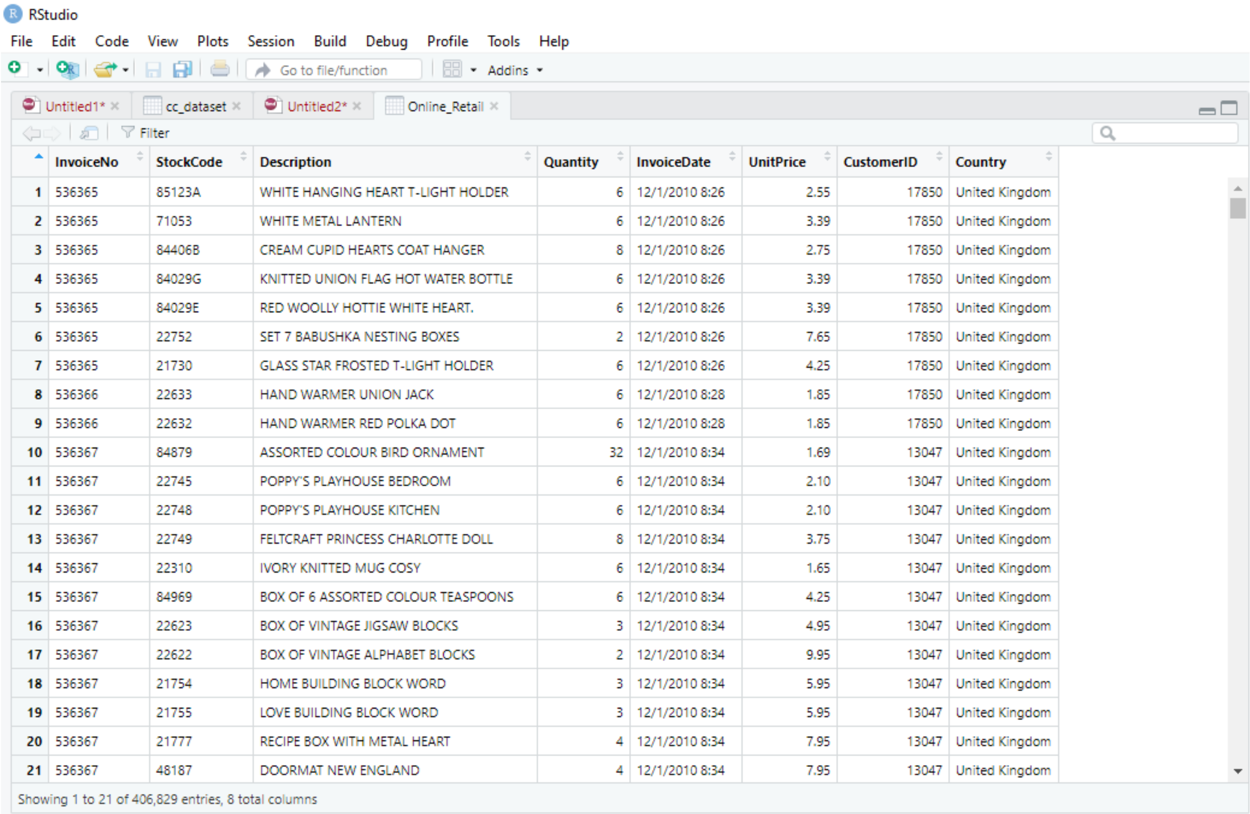
Online Retail Customer Segmentation

Problem Statement:

The company wants to know their customer spending habits so that they can cater better to the customers by structuring their marketing strategy accordingly. The company would like to divide the customers into different categories based on their purchase behavior, so that they can market their products appropriately to each category. Upon completion of analysis each category would have different attributes like customers spending less amount but consistently and customers spending more amount but not so frequently etc.

Dataset:

This is a transnational data collection that includes all transactions made by a UK-based and registered non-store online retailer between December 1, 2010, and December 9, 2011. The company specializes on selling one-of-a-kind presents for any occasion. Many customers of the company are wholesalers. The dataset contains information about Invoice No, Stock code, Description, Quantity, Invoice date, Unit price, Customer id and the Country.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom
9	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingdom
10	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047	United Kingdom
11	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	12/1/2010 8:34	2.10	13047	United Kingdom
12	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	12/1/2010 8:34	2.10	13047	United Kingdom
13	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	12/1/2010 8:34	3.75	13047	United Kingdom
14	536367	22310	IVORY KNITTED MUG COSY	6	12/1/2010 8:34	1.65	13047	United Kingdom
15	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	12/1/2010 8:34	4.25	13047	United Kingdom
16	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	12/1/2010 8:34	4.95	13047	United Kingdom
17	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	12/1/2010 8:34	9.95	13047	United Kingdom
18	536367	21754	HOME BUILDING BLOCK WORD	3	12/1/2010 8:34	5.95	13047	United Kingdom
19	536367	21755	LOVE BUILDING BLOCK WORD	3	12/1/2010 8:34	5.95	13047	United Kingdom
20	536367	21777	RECIPE BOX WITH METAL HEART	4	12/1/2010 8:34	7.95	13047	United Kingdom
21	536367	48187	DOORMAT NEW ENGLAND	4	12/1/2010 8:34	7.95	13047	United Kingdom

Showing 1 to 21 of 406,829 entries, 8 total columns

Here is the link for the dataset: https://www.kaggle.com/roshansharma/online-retail-transactions-in-uk/data

Approach for the model:

Clustering is a data mining process that uses customer data to divide consumers into groups in such a way that members of one group have a lot of attributes in common with each other while each group is different from other groups. The K-means approach is one of the most often used clustering methods. In simple terms, K is several groups/clusters that the data can be divided into based on their similarities and dissimilarities in attributes. Each cluster has a centre which has most similarities to its neighbouring observations. In K-means clustering the number of similarities to be used to cluster can be conveyed as a distance from one observation to the other. Then, using this distance, the calculation is done to determine which cluster each member of the observation belongs to. With each additional observation, new cluster centres are determined, and new observations are assigned to the appropriate cluster.

The variables we are interested in:

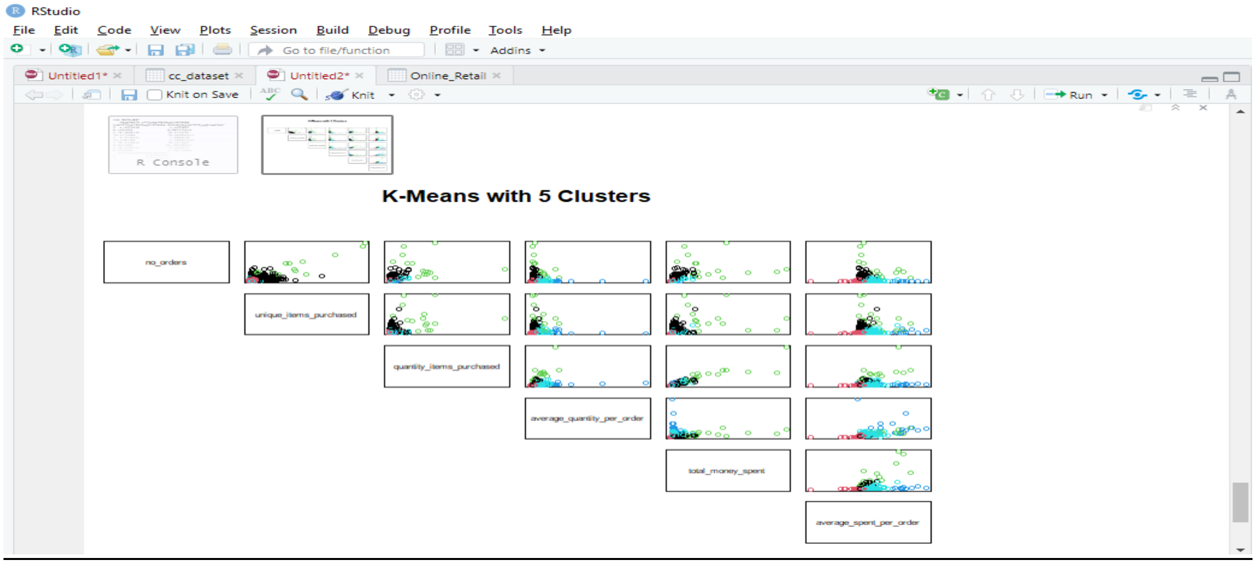
I decided to try clustering customers based on the following variables:

- What is the total number of orders a customer has placed?
- What is the total number of items a consumer has purchased?
- How many unique items has a customer purchased?
- How many products did a consumer buy on average per order?
- What is the total amount of money spent by a customer?
- How much does the average consumer spend on an order?

Why use K-means?

To locate groupings that haven't been explicitly identified in the data, the K-means clustering technique is utilized. This can be used to verify business assumptions about the types of groups that exist, as well as to identify unknown groups in large data sets. Any additional data may be readily allocated to the correct group once the algorithm has been run and the groups have been formed.

Based on my analysis these are the results I determined:



The analysis says that there are five types of customers, but the exact numbering of clusters will change every time as it Is randomly assigned in R:

- Cluster 1: consists of a big number of customers who place a significant number of minor orders. They have tiny order quantities and spend amounts, yet they place a lot of orders.
- Cluster 2: is the largest group of our customers. They only place a few orders (an average of 3.2 orders per customer) and have the smallest basket size - an average Cluster 1 order contains 138 items and costs $231.
- Cluster 3: 15 customers who buy a substantial quantity of items from us - they are most likely large corporations who rely on us as a supplier. These customers spend an average of $3,507 per order and order frequently, averaging 98 orders per customer. They buy the most diverse selection of things, with an average of 728 distinct items purchased.
- Cluster 4: 15 customers that don't place many orders but do so in substantial quantities (an average of 3,539 items per order). Large companies that we supply are also likely to fall into this category. They do, however, only buy a few goods from our catalogue (an average of 49 unique items).
- Cluster 5: A huge group of customers who place many orders for a variety of items. They are frugal consumers.

Conclusion:

Based on the above results we can conclude that segmentation is an essential tool used to understand customers based on their spending habits and purchase behaviour and based on the clusters we can say that the customers can be segmented into different categories , where most of the customers order a significant number of minor orders and a few customers buy less frequently (most likely they are large corporations) but when they do, they order in large quantities.

This can help the company market their products appropriately to different segments of customers and thus improve revenue and profits by leveraging the clustering results.

GitHub Link:

https://github.com/ndholaki/Nemin-Dholakia-ML-.git