Problem statement:

The goal of the company is to know their customer spending habits so they can cater to their customer base by devising a marketing strategy accordingly. By analyzing the purchasing behaviour of the customers the company wants to divide them into different categories by this they aim to market their product accordingly to different customer bases.

At the end of the analysis we would have different categories of customers with different attributes for instance customers shelling out a very less amount of money but consistently over a period of time and customers whose spending ticket is large but their frequency might be less etc.

Overview of dataset

This is a transnational data collection that comprises all transactions performed

between December 1, 2010, and December 9, 2011 by a UK-based and

registered non-store online retailer. The company sells unique gift items for all occasions. However, wholesalers make up a large portion of the firm's clientele. The dataset contains information about Invoice

No, Stock code, Description, Quantity, Invoice date, Unit price, Customer id

and the Country.

Dataset link

<https://archive.ics.uci.edu/ml/machine-learning-databases/00352/>

The rationale behind the model:

Breaking down the working of the clustering in simple words - it divides the data points (in our case the customers of the online retailer) into several different clusters that have similar attributes to those of other groups. The basic goal is to segregate different groups with similar purchasing habits and assign them into clusters.

K-means is one of the most common clustering methods used widely which is a centroid-based model. K is a set of groups or clusters into which the data can be divided based on the similarities and differences in their attribute values. Each cluster has a centre where observations from nearby clusters are most comparable. The amount of similarities to be employed in K-means clustering can be expressed as the separation between two observations. The calculation is then performed to ascertain which cluster each observational member belongs to using this distance. New cluster centres are identified with each additional observation, and new observations are reassigned to the relevant cluster.

The K-means clustering method is used to find groups that haven't been explicitly specified in the data. This can be used to both find unknown groups in huge data sets and to confirm business hypotheses about the kinds of groups that already exist. As soon as the algorithm has been run and the groups have been created, any extra data can be easily assigned to the correct group.

I was inclined towards using the following variables and used the same for clustering the customers.

* 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?

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: This cluster has a very large number of customers who placed the orders but their order value was minor. Their order quantities are tiny, their spending ticket size is also low, however, they place a lot of orders.
* Cluster 2: This cluster consists of the largest group of our customers. These are the customers who place a few orders with an average of 3.2 orders per customer and also have the smallest basket size.
* Cluster 3: This cluster consists of only 15 customers, however, their average ticket spending per customer is $3,507 per order and they also order very frequently. The average is 98 orders per customer, their purchase items is also very diverse with an average of 728 distinct items purchased. It would be safe to assume that these 15 customers would be large corporations who rely on our clients for their supplies.
* Cluster 4: This cluster also consists of only 15 clients but, these are the ones who buy a substantial amount of items per order with an average of 3,539 items per order. Large corporations to whom our client company supplies would fall into cluster 4.
* Cluster 5: Cluster 5 consists of a varied group of customers who place many orders for different categories of items. These customers can be labelled as economical ones.

Summing it up on the basis of our findings we could say that segmentation is a very productive tool to understand the customers spending habits which can help the company to approach different marketing strategies.