

# Retail

## Context:

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

## Data Description:

**InvoiceNo:** Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

**StockCode:** Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

**Description:** Product (item) name. Nominal.

**Quantity:** The quantities of each product (item) per transaction. Numeric.

**InvoiceDate:** Invoice Date and time. Numeric, the day and time when each transaction was generated.

**UnitPrice:** Unit price. Numeric, Product price per unit in sterling.

**CustomerID:** Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

**Country:** Country name. Nominal, the name of the country where each customer resides.

## Problem statement

It is a business critical requirement to understand the value derived from a customer. RFM is a method used for analyzing customer value.

Perform customer segmentation using RFM analysis. The resulting segments can be ordered from most valuable (highest recency, frequency, and value) to least valuable (lowest recency, frequency, and value). Identifying the most valuable RFM segments can capitalize on chance relationships in the data used for this analysis.

## Approach:

Following pointers will be helpful to structure your findings.

1. Perform a preliminary data inspection and Data cleaning
  - a. Check for missing data and formulate apt strategy to treat them.
  - b. Are there any duplicate data records? Remove them if present.
  - c. Perform Descriptive analytics on the given data.
2. Cohort Analysis: A cohort is a group of subjects who share a defining characteristic. We can observe how a cohort behaves across time and compare it to other cohorts.
  - a. Create month cohorts and analyse active customers for each cohort.
  - b. Also Analyse the retention rate of customers. Comment.

3. Build a RFM model – Recency Frequency and Monetary based on their behaviour. Recency is about when was the last order of a customer. It means the number of days since a customer made the last purchase. If it's a case for a website or an app, this could be interpreted as the last visit day or the last login time.

Frequency is about the number of purchase in a given period. It could be 3 months, 6 months or 1 year. So we can understand this value as for how often or how many a customer used the product of a company. The bigger the value is, the more engaged the customers are. Could we say them as our VIP? Not necessary. Cause we also have to think about how much they actually paid for each purchase, which means monetary value.

Monetary is the total amount of money a customer spent in that given period. Therefore big spenders will be differentiated with other customers such as MVP or VIP.

- a. Calculate RFM metrics.
  - i. Recency as the time in no. of days since last transaction
  - ii. Frequency as count of purchases done
  - iii. Monetary value as total amount spend
- b. Build RFM Segments.
  - i. Give Recency Frequency and Monetary scores individually by dividing them in to quartiles.

Note: Rate "Recency" for customer who have been active more recently better than the less recent customer, because each company wants its customers to be recent

Rate "Frequency" and "Monetary Value" higher label because we want Customer to spend more money and visit more often.

- ii. Combine three ratings to get a RFM segment (as strings)
    - iii. Get the RFM score by adding up the three ratings.
  - c. Analyse the RFM Segments by summarizing them and comment on the findings.
4. Create clusters using k means clustering algorithm.
  - a. Prepare the data for the algorithm.
    - i. If the data is Un Symmetrically distributed, manage the skewness with appropriate transformation.
    - ii. Standardize / scale the data.
  - b. Decide the optimum number of clusters to be formed
  - c. Analyse these clusters and comment on the results.
5. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:
  - a) Country-wise analysis to demonstrate Average spend. Use a bar chart show monthly figures.
  - b) Bar graph of top 15 products which are mostly ordered by the users to show the number of products sold.
  - c) Bar graph to show the count of orders Vs. hours throughout the day. What are the peak hours per your chart?
  - d) Plot the distribution of RFM values using histogram and frequency-charts.
  - e) Plot error(cost) vs no of clusters selected
  - f) Visualize to compare the RFM values of the clusters using heatmap

## Data Cleaning:

1. Perform a preliminary data inspection and data cleaning.
  - a. Check for missing data and formulate an apt strategy to treat them.
  - b. Remove duplicate data records.
  - c. Perform descriptive analytics on the given data.

- Calculating the missing value % Contribution in Dataset
- Dropping the missing values in the dataset
- Changing the datatype as per business understanding

## Step 2: Data Preparation

*We are going to analysis the Customers based on below 3 factors:*

- R (Recency): Number of days since last purchase
  - F (Frequency): Number of transactions
  - M (Monetary): Total amount of transactions (revenue contributed)
- 
- Create New Attribute : Monetary
  - Create New Attribute: Frequency
  - Merge merge dataset
  - Create New Attribute : Recency
  - Convert datetime to proper datatype
  - Compute the last date to know the last transaction date
  - Compute the difference between max date and transaction date
  - Compute the last transaction date to get the recency of customers
  - Extract the exact days
  - Merge the datasets to get the final RFM model

**There are 2 types of outliers and we will treat outliers as it can skew our dataset**

1. Statistical
  2. Domain specific
- 
- Outlier Analysis of amount, frequency and recency
  - Removing (Statistical) outliers for Amount
  - Removing (Statistical) outliers for Frequency
  - Removing (Statistical) outliers for Recency

## Rescaling the attribute

It is extremely important to rescale the variable so that they have comparable scale. There are two common ways of rescaling:

1. Min-Max Scaler
2. Standardization (mean-0, sigma-1)

Here, we will use Standardisation

- Rescaling the attributes, Instantiate Standard scaler and Fit-Transform

## Step 4 : Building the Model

### K-means Clustering

K-means is one of the simplest and popular unsupervised machine learning algorithms.

The algorithm works as follows:

- first we initialise K points, called means, randomly.
  - We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorised in that mean so far.
  - We repeat the process for given number of iterations and at the end, we have our clusters
- K-means with some arbitrary k value
  - Elbow Curve / SSD

### Silhouette Analysis

$\text{silhouette score} = (p-q)/(\max(p,q))$

p is the mean distance to the points in the nearest cluster that the data point is not a part of q is the mean intra-cluster distance to all the points in its own cluster.

- The value of the silhouette score range lies between -1 to 1
- A score closer to 1 indicates that the data point is very similar to other data points in the clusters
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster

Ans: For n\_clusters = 2, the silhouette score is 0.541842117113117  
For n\_clusters = 3, the silhouette score is 0.5084896296141937  
For n\_clusters = 4, the silhouette score is 0.48175475872338963  
For n\_clusters = 5, the silhouette score is 0.4639646901931184  
For n\_clusters = 6, the silhouette score is 0.4173990086284566  
For n\_clusters = 7, the silhouette score is 0.416300170967371  
For n\_clusters = 8, the silhouette score is 0.4025418765314729

- Final model with K=3
- Create Boxplot to visualize ClusterId Vs Frequency
- Create Boxplot to visualise ClusterId Vs Frequency
- Create Boxplot to visualise ClusterId Vs Recency

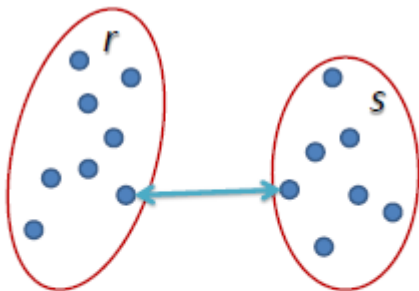
## **Hierarchical Clustering**

Hierarchical Clustering involves creating have a predetermined ordering from top to bottom. For example, all files and folders on the harddisk are organized in a hierarchy. There are two types of hierarchical clustering:

1. Divisive
2. Agglomerative

### **Single linkage:**

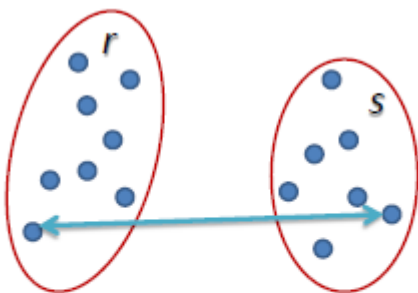
In Single linkage hierarchical clustering, the distance between the 2 clusters is defined as the shortest distance between two points in each cluster. For example, distance between clusters  $r$  and  $s$  to the left is equal to the length of the arrow between their two closest point



$$L(r, s) = \min(D(x_{ri}, x_{sj}))$$

### **Complete Linkage**

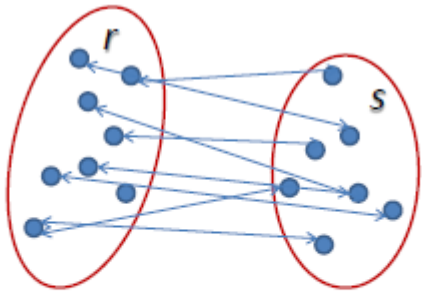
In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. for example, the distance between clusters  $r$  and  $s$  to the left is equal to the length of the arrow between their two furthest points.



$$L(r, s) = \max(D(x_{ri}, x_{sj}))$$

### Average Linkage:

In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters  $r$  and  $s$  to the left is equal to the average length each arrow between connecting the points of one cluster to the other.


$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

### Step 5: Final Analysis

*Inference:*

K-means clustering with 3 cluster ids:

- Customers with cluster id 2 are the customers with high amount of transactions as compared to other customers
- Customers with cluster id 2 are the frequent buyers
- Customers with cluster id 0 are not recent buyers and hence of least of importance from business point of view

Hierarchical Clustering with 3 Cluster Ids:

- Customers with Cluster\_labels 2 are the customers with high amount of transactions as compared to other customers
- Customers with cluster\_labels 2 are frequent buyers
- Customers with cluster\_labels 0 are not recent buyers and hence least of importance from business point of view