Data Science with R: Project

This document contains the problem statement with dataset information.

Dataset can be downloaded from the download section in the LMS.

Problem Statement:

A UK-based online retail store has captured the sales data for different products for the period of one year (Nov 2016 to Dec 2017). The organization sells gifts primarily on the online platform. The customers who make a purchase consume directly for themselves. There are small businesses that buy in bulk and sell to other customers through the retail outlet channel.

Analysis information:

Find the significant customers for the business who make high purchases of their favorite products. The organization wants to roll out an offer to the high-value customers after identification of segments. Use the clustering methodology to segment customers into groups:

- Use the following clustering algorithms:
  - K means
  - Hierarchical
- Identify the right number of customer segments
- Provide the number of customers who are highly valued
- Identify the clustering algorithm that gives maximum accuracy and explains robust clusters.
- If the number of observations is loaded in one of the clusters, break down that cluster further using clustering algorithm.

Variables in the Dataset:

- This is a transnational dataset that contains all the transactions occurring between Nov-2016 to Dec-2017 for a UK-based online retail store.
- Variable Information:
  - InvoiceNo: Invoice number (A 6-digit integral number uniquely assigned to each transaction)
  - StockCode: Product (item) code
  - Description: Product (item) name
  - Quantity: The quantities of each product (item) per transaction
  - InvoiceDate: The day when each transaction was generated
  - UnitPrice: Unit price (Product price per unit)
  - CustomerID: Customer number (Unique ID assigned to each customer)
  - Country: Country name (The name of the country where each customer resides)