AS4: Exploratory Data Analysis with Clustering

Imagine that you are a data scientist working at a marketing department of a UK-based non-store online retailer. The company mainly sells unique all-occasion gifts and many of its customers are wholesalers. Your team is planning on carrying out a targeted marketing promotion. For this, you are asked to conduct an exploratory data analysis (EDA) and to identify customer segments. After some careful consideration, you decided to utilize the recency, frequency, monetary value (RFM) segmentation framework which, based on the practice / foundation, summarizes customers' purchase history using the following three variables:

- **Recency (R)** is a measure of how recently a customer has made a transaction,
- **Frequency (F)** is a measure of how often a customer makes a transaction, and
- **Monetary value (M)** is a measure of how much money a customer spends.

The retailer's transaction data contains all the purchases occurring between December 01, 2010 ~ December 09, 2011, and is stored in the file, ***onlineRetail.csv***, which is accessible in the assignment section on Canvas.

The variables that are included in the data set are as follows:

- `InvoiceNo`: The invoice number, which is a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
- `StockCode:` The product (item) code, which is a 5-digit integral number uniquely assigned to each distinct product. It can be accompanied by a trailing uppercase letter.
- `Description:` The Product (item) name.
- `Quantity:` The quantities of each product (item) per transaction.
- `InvoiceDate:` The invoice day and time when a transaction was generated. (follows a MM/DD/YY format)
- `UnitPrice:` The product price per unit in sterling (£).
- `CustomerID:` The customer number, which is a 5-digit integral number uniquely assigned to each customer.
- `Country:` The name of the country where a customer resides.

Please make sure to submit a document created using R Markdown with your answers to each question in the assignment.

INSTRUCTIONS:

1) Import and Examine the Data

a) Import the CSV file into R using fread() and take a look at the data (e.g., dim, head, summary, etc.)

b) Examine the data by printing out the unique number of customers, the unique number of products purchased, as well as the unique number of transactions.

2) Compute the RFM Variables

c) Convert the InvoiceDate into a date obj. then create a variable called Recency by computing the number of days until the last day of purchase in the dataset (i.e., Dec. 09, 2011) since last purchase for each customer.

d) Create a variable called Frequency and Monetary for each customer in the data.

3) Removing Outliers (i.e., Winsorizing)

e) Visualize the RFM variables with box plots.

f) It seems that there are extreme values in the RFM variables. Remove these extreme values/outliers by keeping only the values that are within the 99^{th} percentile.

4) Scaling the Variables

g) To prep the data for clustering, we will need to scale the features/variables. Create another data.table obj. called RFM_Scaled which contains the CustomerID and the standardized RFM variables.

5) Running K-Means Clustering

h) Convert RFM_Scaled to a matrix. (p.s., do not forget to remove the CustomerID from the matrix!)

i) Set seed at 2021 and run k-means clustering (set k = 4).

j) Attach the cluster numbers (i.e., km.out$cluster) onto RFM_Scaled.

6) Examining the Clusters

k) Compute the average of RFM for each cluster. Do we observe any difference between the clusters? Can we label them? Which of the clusters do you think are the most suitable for us to run target marketing campaigns and how?

l) Based on the list of top selling products, you could further develop your target marketing strategies. Print out the top 5 most selling products in terms of sales revenue (i.e., sum of sales amount = quantity x unit price) for each cluster.

Extra Credit:

EC1) You are interested in finding out if there is any seasonality (variation by month) in ***purchase frequency*** of the 5 top/best sellers. Compute purchase frequency of the top 5 selling products by month and visualize it using ggplot2.

EC2) Do we observe any seasonality? Please explain verbally.

EC3) When using k-means clustering, the number of clusters should be predetermined, and this should be firmly backed by domain knowledge or a proven theory. However, we could also take a data-driven approach by using methods such as the Elbow method or the Silhouette method which can easily be done using the packages like `factoextra` and `NbClust`. Explain whether k = 4 is a reasonable decision using the Elbow/Silhouette method.