Multivariate Analysis Final Report

r11250023 陳貞諺

April 20, 2023

Q1. RFM Analysis for OnlineRetail Data

Introduction

We are going to do RFM (Recency, Frequency, Monetary) analysis for a UK-based and registered non-store online retail. That is, to segment consumer based on their purchasing patterns showed by Recency, Frequency, and Monetary. After that, we suggest corresponding actions for the brand to improve sales or maintain customers.

The dataset "OnlineRetail.csv" from UCI repository contains all the transactions occurring between 01/12/2010 and 09/12/2011 for the brand. There are 541909 observations and 8 attributes as following:

1. 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.
2. StockCode: Product/item code (nominal). A 5-digit integral number uniquely assigned to each distinct product.
3. Description: Product/item name (nominal).
4. Quantity: The quantities of each product/item per transaction (numeric).
5. InvoiceDate: Invoice date and time (numeric). The day and time when each transaction was generated.
6. UnitPrice: Unit price (numeric). Product price per unit in sterling.
7. CustomerID: Customer number (nominal). A 5-digit integral number uniquely assigned to each customer.
8. Country: Country name (nominal). The name of the country where each customer resides.

Data cleaning

Before we construct the corresponding RFM attributes, we take a glimpse into the data. First, we need to delete the objects missing CustomerID and with non-positive UnitPrice. After that, we notice that the transaction is a cancellation when the quantity is negative. Therefore, it is not meaningless for the data with negative quantity and we have done data cleaning now.

Construction of Recency, Frequency, and Monetary

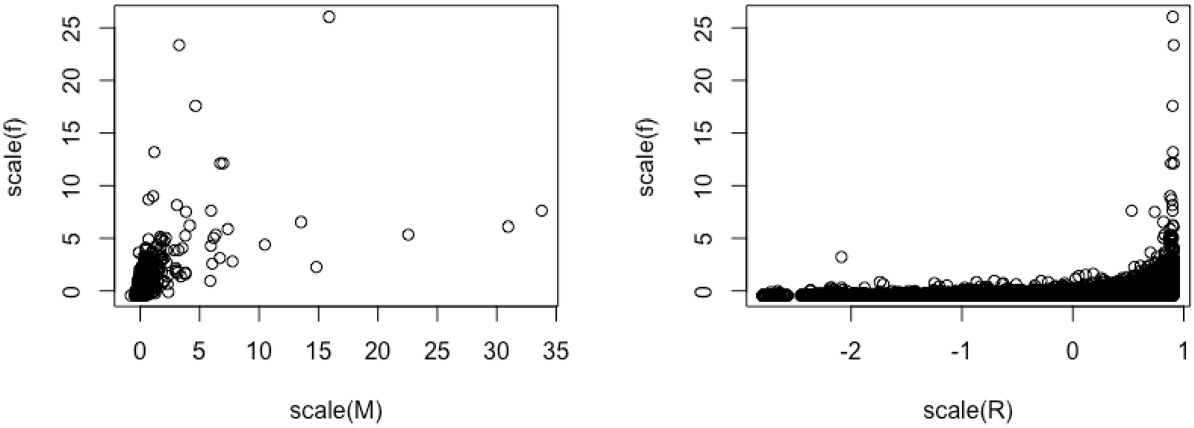
1. Recency: for each customer, we record their latest transaction time as recency and save it in a numerical way, so the customer who is higher in this variable means the latest transaction of him/her is nearer. For example, "2010-12-01 08:26:00"<"2010-12-01 08:28:00". This is not the traditional recency.
2. Frequency: the objects with the same InvoiceNo should be viewed as the same transaction, so we calculate the frequency of each customer by calculating the number of different InvoiceNo for each CustomerID.

3. Monetary: we simply calculate the monetary by the sum of Quantity*UnitPrice for each CustomerID.

Obviously, the scale of recency are much larger than monetary and frequency, which would effects the results of classification depending on distance. Therefore, I standardize these three variables.

Outliers

After our data cleaning, all the data points are valid from the business point of view, since they are real transaction records. But when it comes to data analysis, we may need to remove outliers to get a decent result, which is depending on our data analysis methods.



We can see that there are some outliers on Monetary and Frequency. As a result, we need to deal with their outliers when we use some techniques that are sensitive to the outliers.

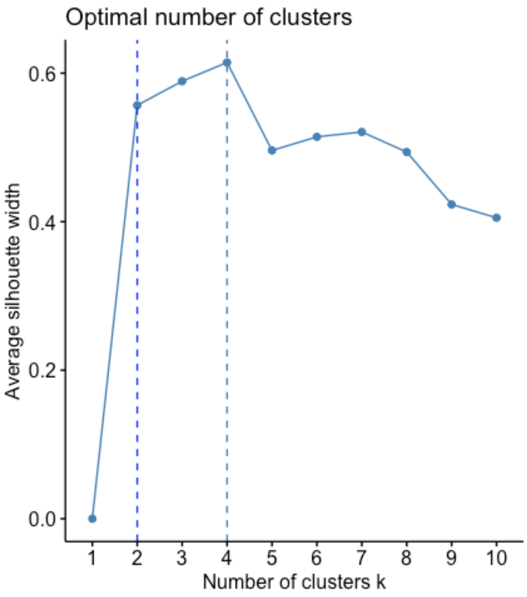
Customer Specification (Clustering)

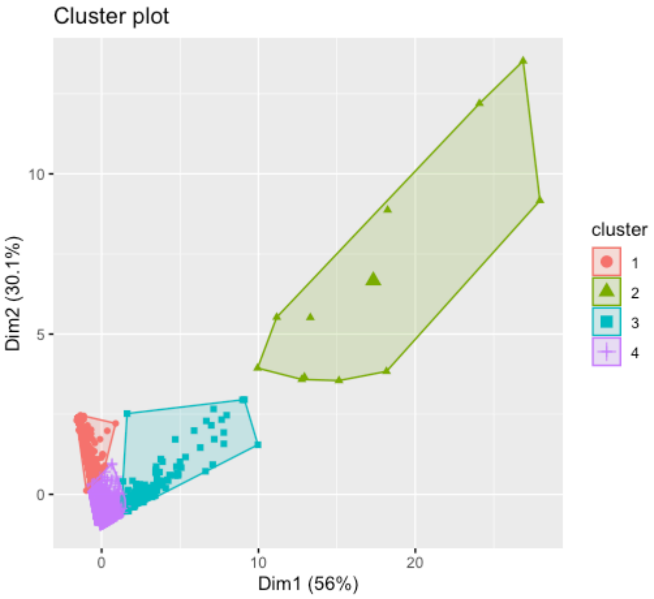
As RFM attributes are constructed, I cluster the customers in the following two ways. Both algorithms will segment similar data into same clusters. After grouping customers, we can then look into the characteristics of each group.

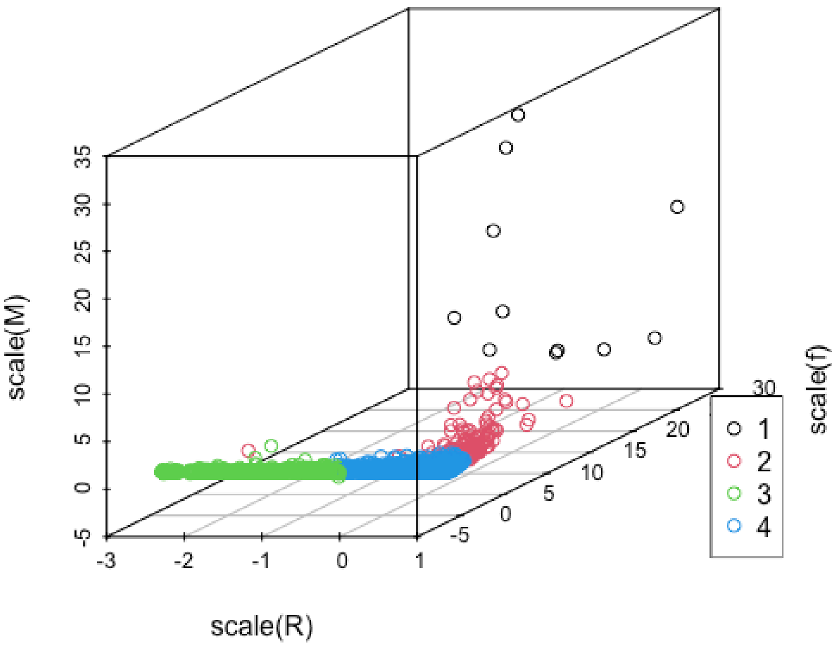
K-means

K-means is a method of clustering and will generate K clusters. This method randomly partitions the items into K initial clusters first and going through each element to reassign it to a cluster and mean within the cluster is nearest on Euclidean distance. Therefore, it would be sensitive to the outliers.

To determine the number of clusters K, I use Silhouette method to assess the fitness of clusters and it shows that when $K = 4$, the data is most appropriately clustered. The graphical results of clustering are showed below.







As we cluster the customers, we can observe that:

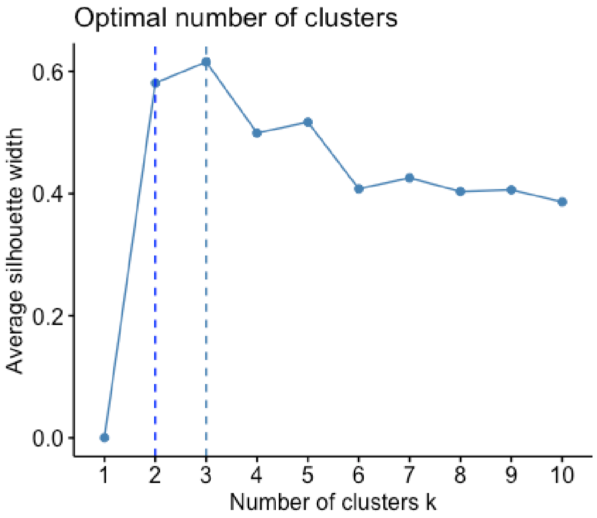
1. Group 1 is of the highest frequency and higher monetary. That is, those who have higher monetary are also regular customers.

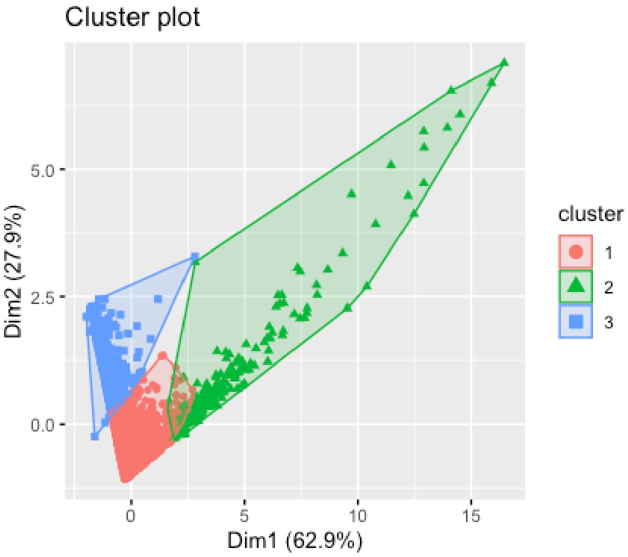
2. Group 2 has second high frequency and monetary and is of high recency.

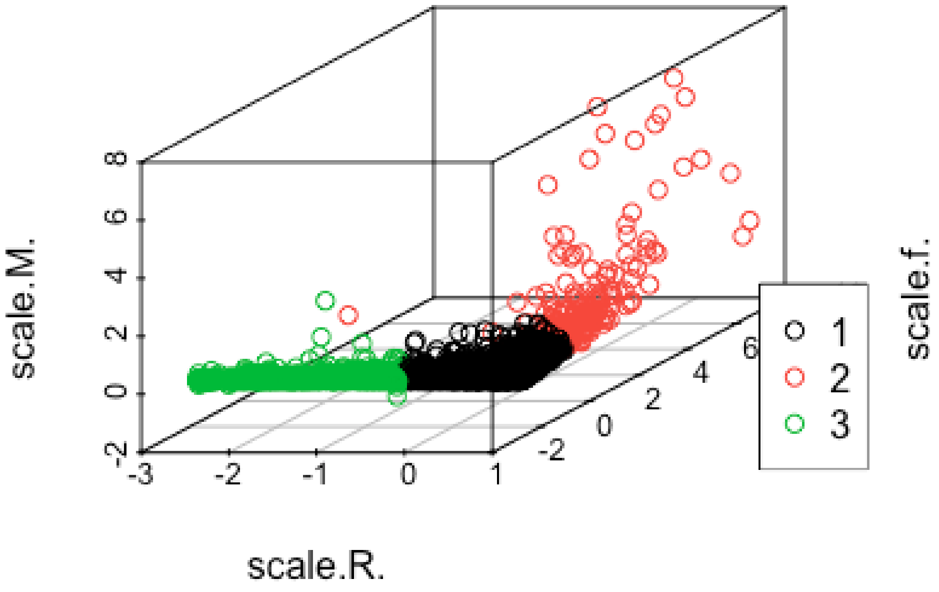
3. Group 3 and Group 4 have similar frequency and monetary, but Group 3 is of the lowest recency.

K-means (outliers removed)

We further examine the results with outliers removed.





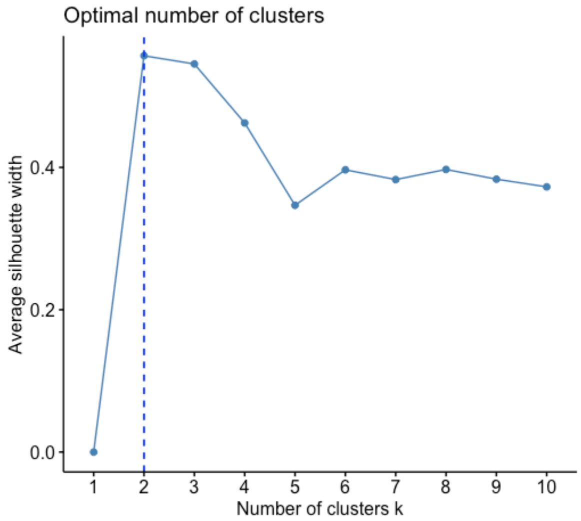


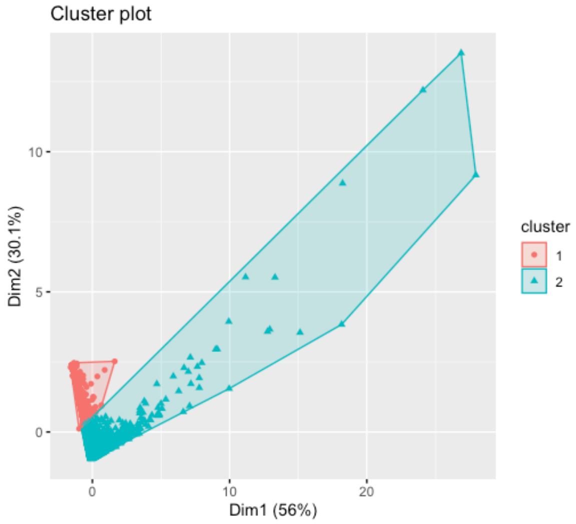
It seems removing outliers is to remove group 1 in the previous results.

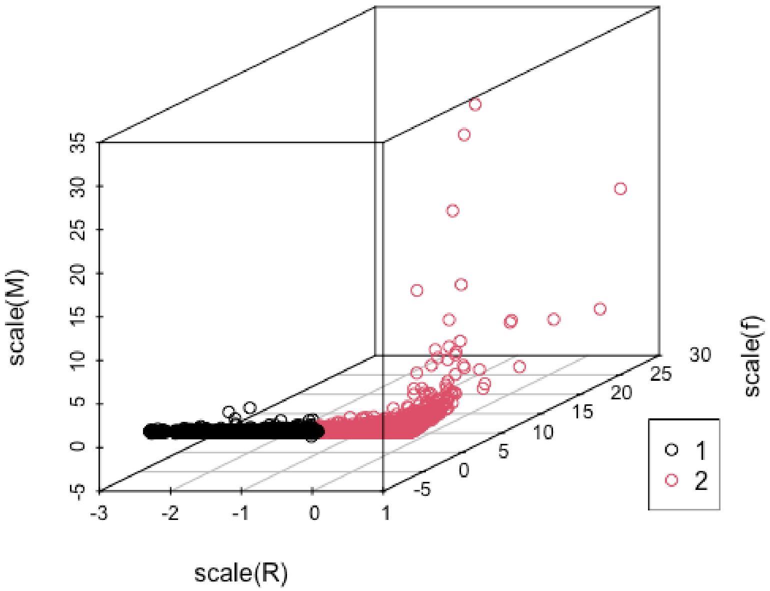
K-medoids

K-medoids is a method of clustering and will generate K clusters. This method randomly chooses K items to be the medoids (as a center) and assign each item to the nearest medoid to form K initial clusters. Then, for each cluster, swap the medoid to a better one with smallest mean and repeat to assign each item to the nearest medoid. K-medoids is more robust to the outliers.

To determine the number of clusters K, I use Silhouette method to assess the fitness of clusters and it shows that when $K = 2$, the data is most appropriately clustered. The graphical results of clustering are showed below.







We cluster the customers into two groups. Group 1 is of lower recency against Group 2 and Group 2 is variant in Frequency and Monetary.

Discussion between K-Means and K-Medoids

We have shown that the results of clustering are very different between K-means and K-medoids. This is reasonable since I do standardization in construction of RFM. K-means algorithm depends on the mean of Euclidean distance between data, so it would be more sensitive to our data than K-medoids. In my experiments with RFM attributes, if we do K-means without standardization of RFM, the result resembles K-medoids. Moreover, the results of K-medoids with or without standardization are similar.

There are similarities between the results of both two methods. Although K-medoids only suggests to form 2 clusters, Group 1 in K-medoids resembles Group 3 in K-means, and Group 2 in K-medoids resembles Group 1+Group 2+Group 4. This means both two methods suggested that the customers with the lowest recency should be specified from other customers.

Conclusion

When we collect the transaction data of each customer, we can do RFM analysis to know about customers' behavior. As we cluster the customers via some methods, we can do corresponding actions to improve our sales and maintain customers. Since we want to retrieve some features in our data, the outliers cannot be simply omitted and we need to compare the results of different clustering methods for interpretation. In this data, I will get use of our K-means clustering results above for precise customer segmentation and suggest that:

1. For those regular customers who have high monetary (Group 1), the brand should keep contact with them when they find someone's recency become smaller (his/her last transaction was long time before) by sending some coupons, and frequently send flyers to them.

2. For second regular customers who have second high monetary (Group 2), the brand should launch some exclusive promotions or limited time offer to improve their frequency and monetary.

3. For those who has low frequency and monetary (Group 3&4), the brand should place advertisements or launch some campaigns to increase its exposure and publicity and set up sales events to attract people to shop at their online retail more.