# E-COMMERCE DATA CLUSTERING AND CUSTOMER SEGMENTATION

Prepared for:

*EAS 507 Final Project*

*Yao Ji #22*

*Siddharth Satyakam ##44*

*Yuan Hui #20*

University at Buffalo The State University of New York

# Introduction

## Data Introduction

- This UK-based transnational data set contains all the online transactions occurring between 12/01/2010 and 12/09/2011 for customers from different countries.
- It has 541909 observations and 8 variables

## Objectives

- This project aims at analyzing the customers' online purchase behaviors
- To divide the customers into groups based on the analysis of their similar shopping behaviors and also to anticipate the potential purchases made by new customers.

## Methods

- Exploratory Data Analysis (EAD)
- Clustering (K-means)
- RFM analysis (Recency , Frequency, Monetary Value)

## Data Source

Kaggle challenge: https://www.kaggle.com/carrie1/ecommerce-data

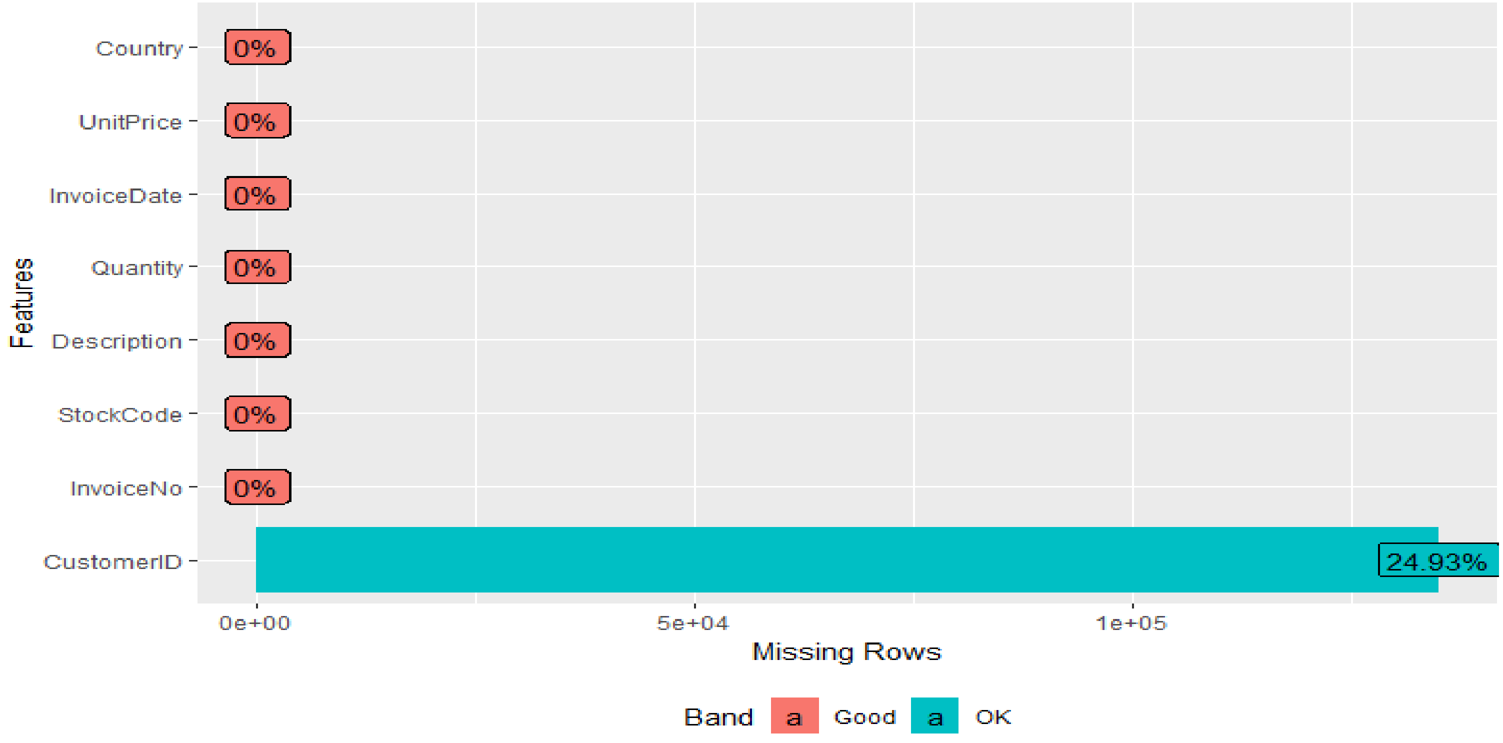
# Variable Description

- **InvoiceNo:** invoice number, unique for each transaction (code starts with letter ‘c’ indicates a cancellation).
- **StockCode:** uniquely assigned to each distinct product
- **Description:** product (item) name.
- **Quantity:** numerical variable. The quantities of each product per transaction.
- **InvoiceDate:** Invice Date and time.
- **UnitPrice:** numerical variable, Unit price for each product.
- **CustomerID:** uniquely assigned to each customer.
- **Country:** where each customer resides.

# Exploratory Data Analysis

## Missing Values

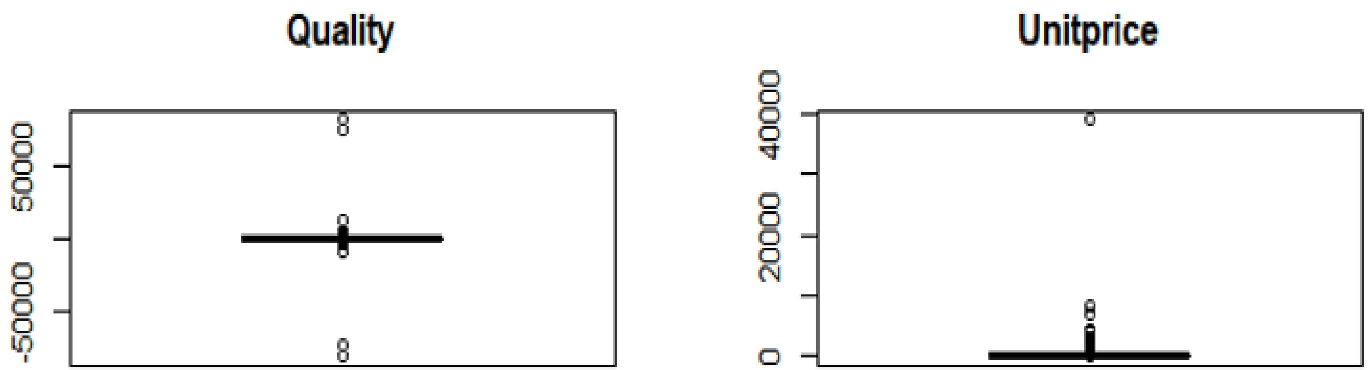
- Variable CustomerID has 135080 missing values, nearly 25% of total values
- We remove these NAs from CustomerID



# Exploratory Data Analysis

## Numerical Variables with Outliers

- Item sales = Unitprice * Quantity, because the sum of item sales (negative and positive sales are equal) is 0, they do not affect the result of sales for each item and total sales. Thus we do not remove the outliers from variables in this case.

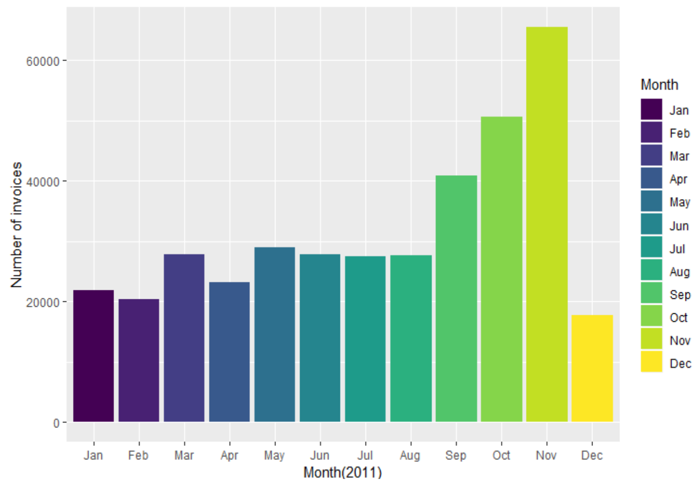


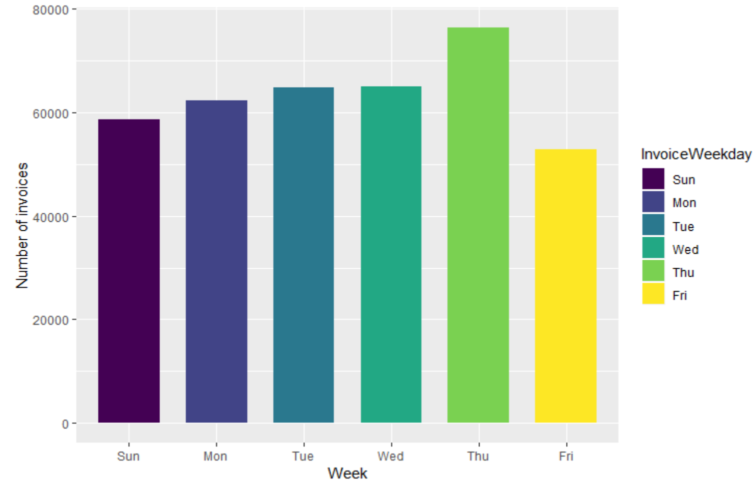

```
> #outliers of Quantity and unitprice
> customer[(customer$Quantity>50000 | customer$Quantity< -50000.00),]#(-11062.06, 38970)
       InvoiceNo StockCode                    Description Quantity      InvoiceDate UnitPrice CustomerID        Country
61620     541431     23166 MEDIUM CERAMIC TOP STORAGE JAR    74215 1/18/2011 10:01      1.04      12346 United Kingdom
61625    C541433     23166 MEDIUM CERAMIC TOP STORAGE JAR   -74215 1/18/2011 10:17      1.04      12346 United Kingdom
540422    581483     23843    PAPER CRAFT , LITTLE BIRDIE    80995  12/9/2011 9:15      2.08      16446 United Kingdom
540423   C581484     23843    PAPER CRAFT , LITTLE BIRDIE   -80995  12/9/2011 9:27      2.08      16446 United Kingdom
> customer[(customer$UnitPrice<0 | customer$UnitPrice>30000.00),]#(-11062.06, 38970)
       InvoiceNo StockCode Description Quantity     InvoiceDate UnitPrice CustomerID        Country
222682   C556445         M      Manual       -1 6/10/2011 15:31  38970.00      15098 United Kingdom
```

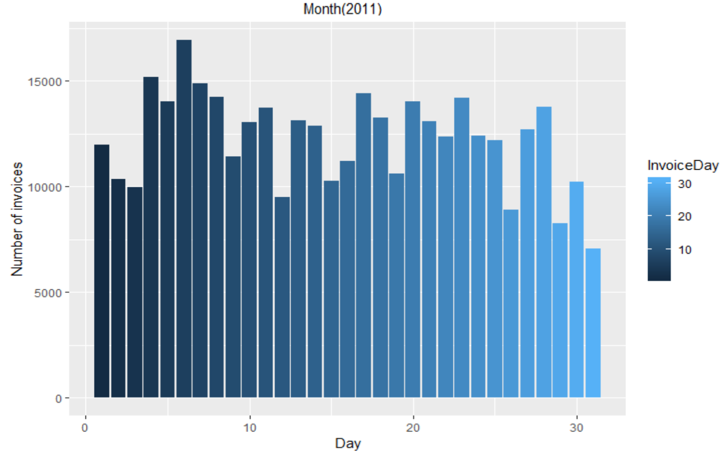
# Exploratory Data Analysis

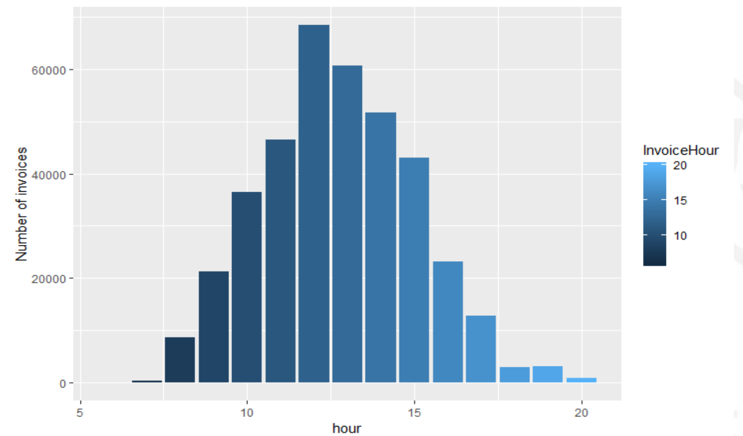
**Dealing with variable "InvoiceDate"**

- Transformed InvoiceDate into datetime variable and extract the Month, Week, Day, Hour and Date from it.



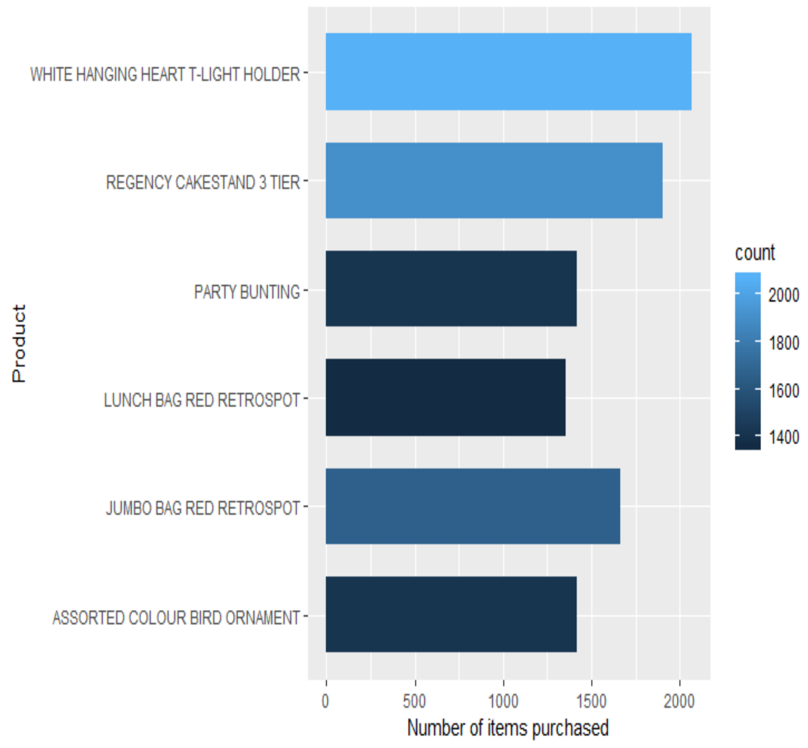



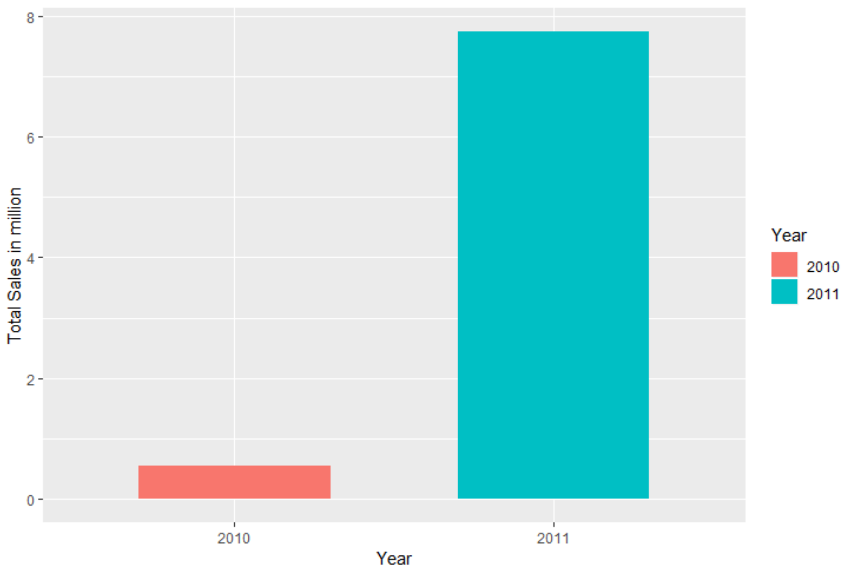



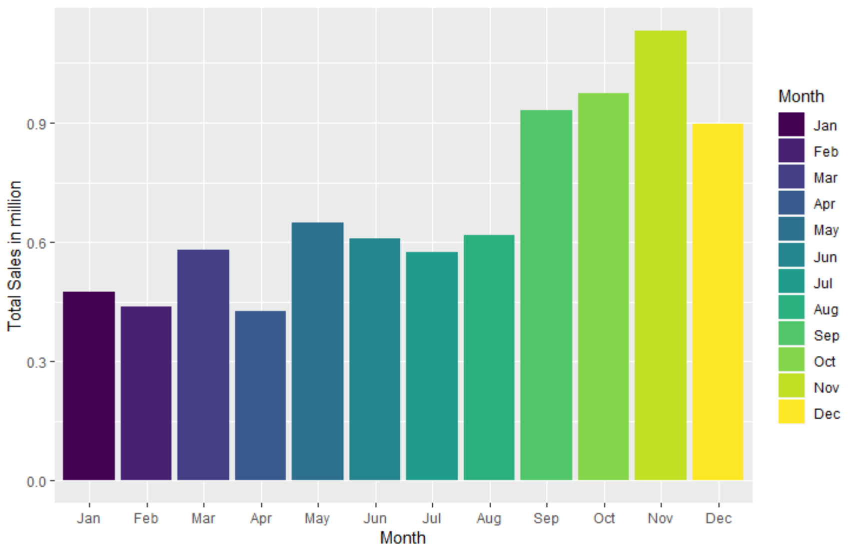



# Exploratory Data Analysis

## Total Sales for Years and Months

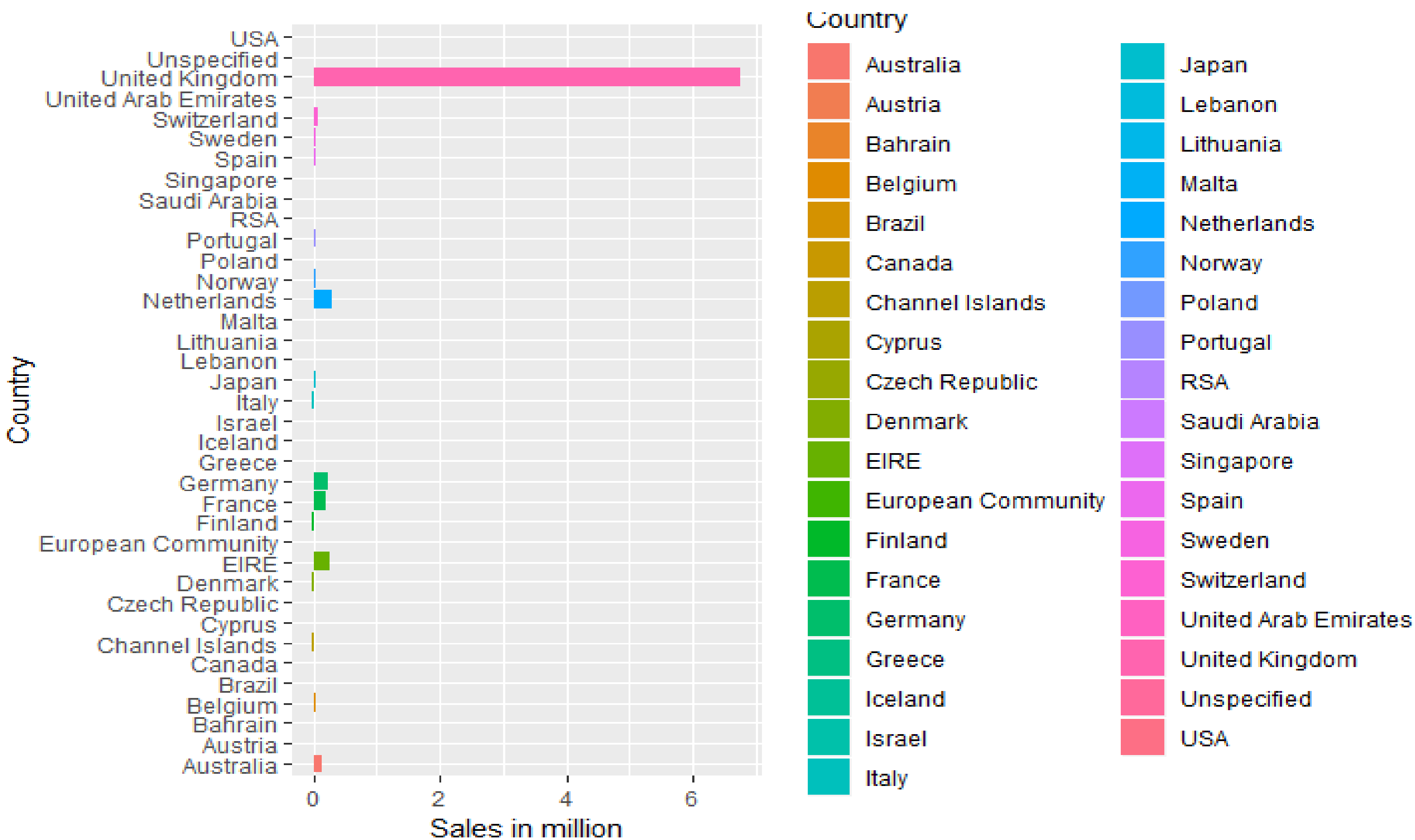


# Exploratory Data Analysis

**Total Sales for Each Country**



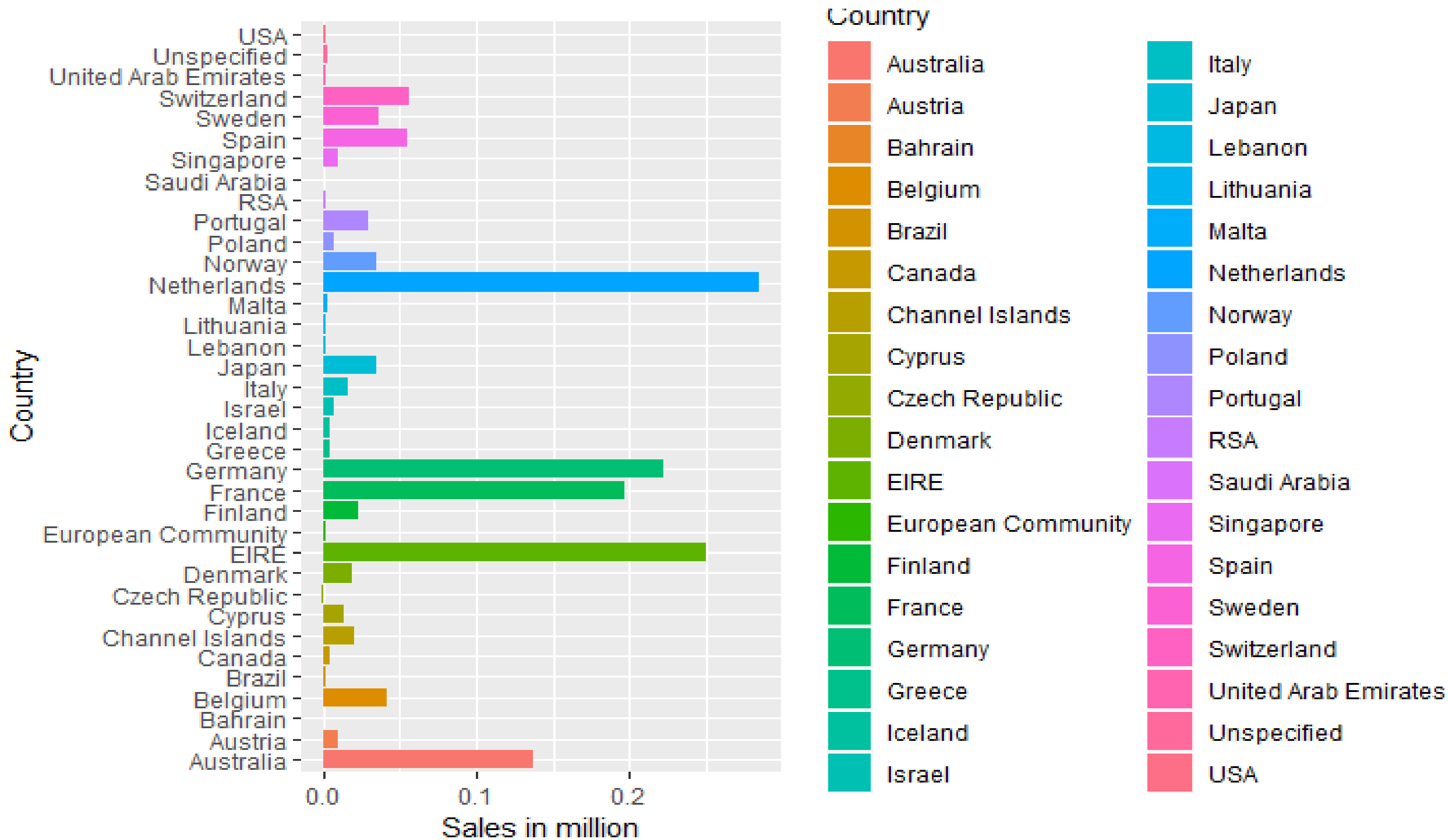
# Exploratory Data Analysis

## Top 10 Countries with Highest Sales

| Country | total_sales | customers | ave_comsuption |
|---|---|---|---|
| United Kingdom | 6767873 | 3950 | 1713 |
| Netherlands | 284661 | 9 | 31629 |
| EIRE | 250285 | 3 | 83428 |
| Germany | 221698 | 95 | 2334 |
| France | 196713 | 87 | 2261 |
| Australia | 137077 | 9 | 15231 |
| Switzerland | 55739 | 21 | 2654 |
| Spain | 54775 | 31 | 1767 |
| Belgium | 4091 | 25 | 1636 |
| Sweden | 36596 | 8 | 4574 |

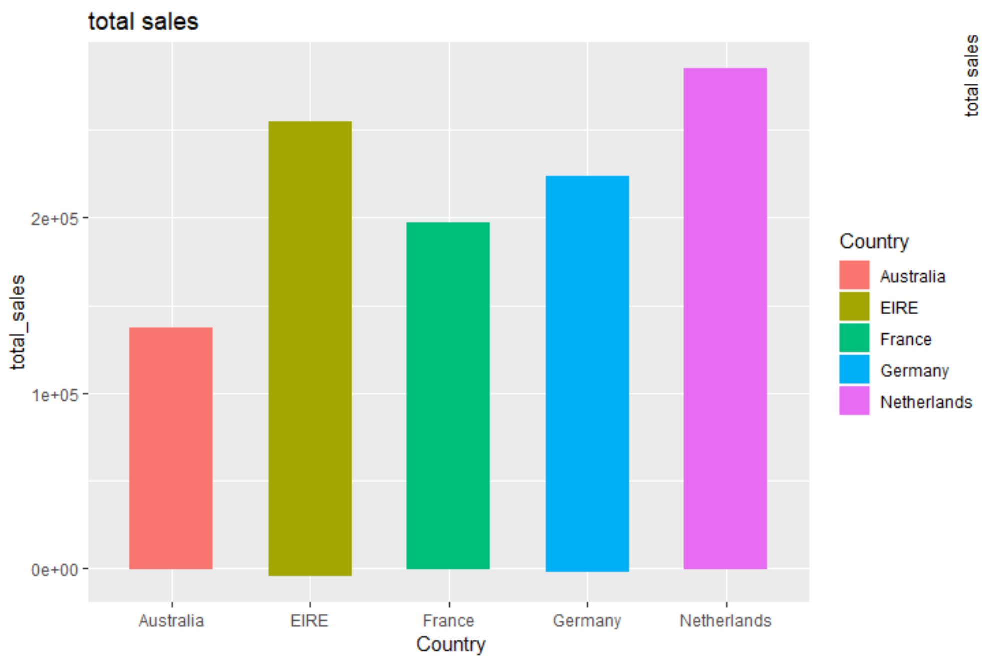
# Exploratory Data Analysis

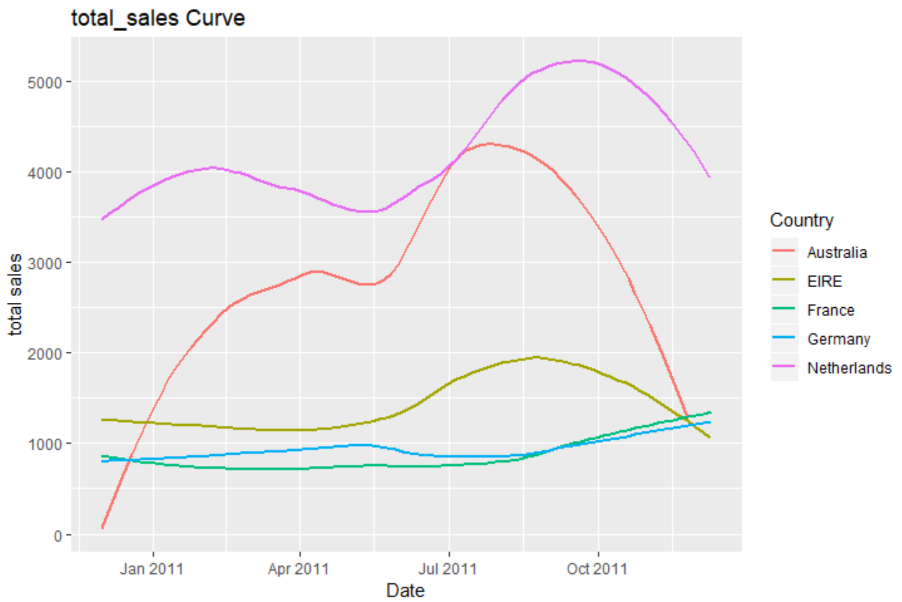
## Total Sales for Other Countries Besides UK



# Exploratory Data Analysis

## Top Five Countries with Highest Sales



# Clustering

- The customers are from 37 countries but the majority of them are from UK. The clusters will be built with
  - transactions from all countries
  - transactions from other countries except UK

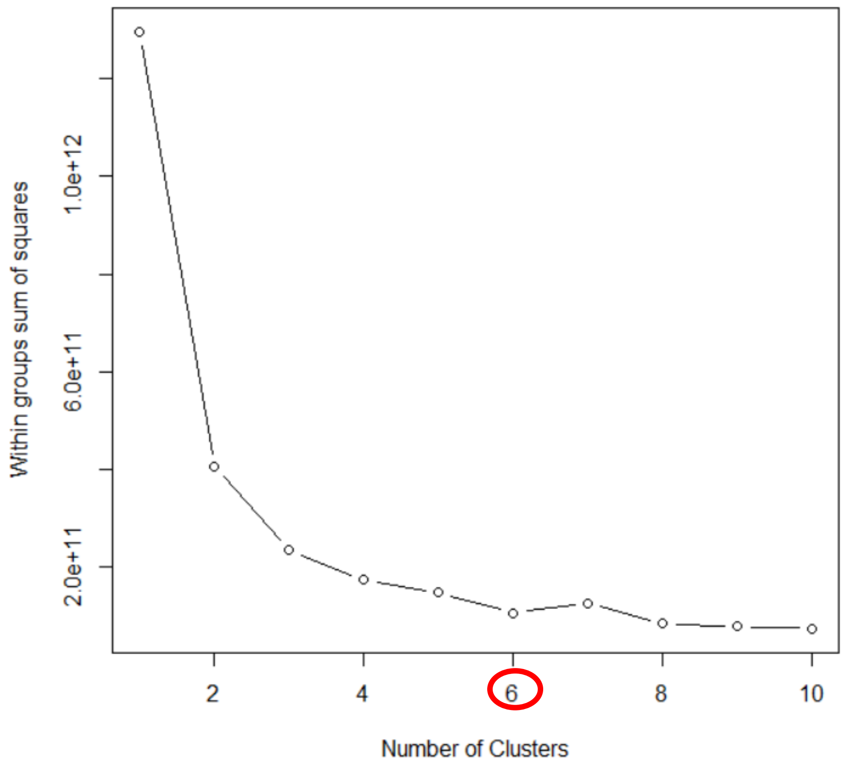
- Based on feature engineering, we know that *Quantity*, *UnitPrice*, *CustomerID*, *Item_Sales*, *Month*, *WeekDay*, *Day*, *Hour* are important features

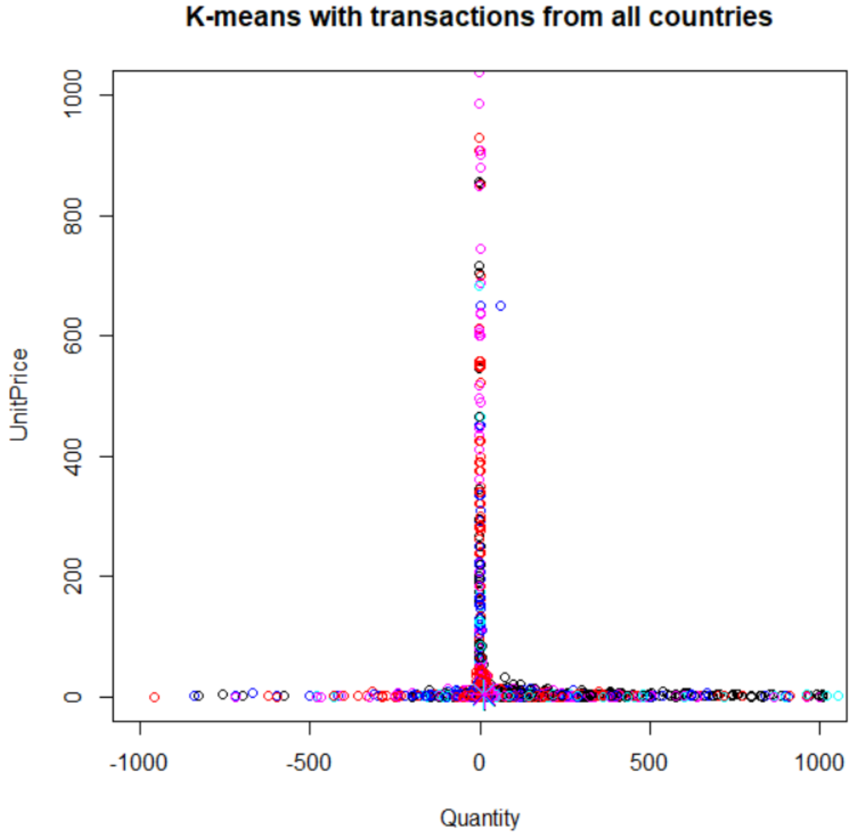
- Clustering models
  - K-Means (with/without PCA)
  - K-Medoids
  - Hierarchical clustering

# Clustering

## K-Means with transactions from all countries

- Six clusters is suggested from sum of square
- The adjusted rand index is 0.57

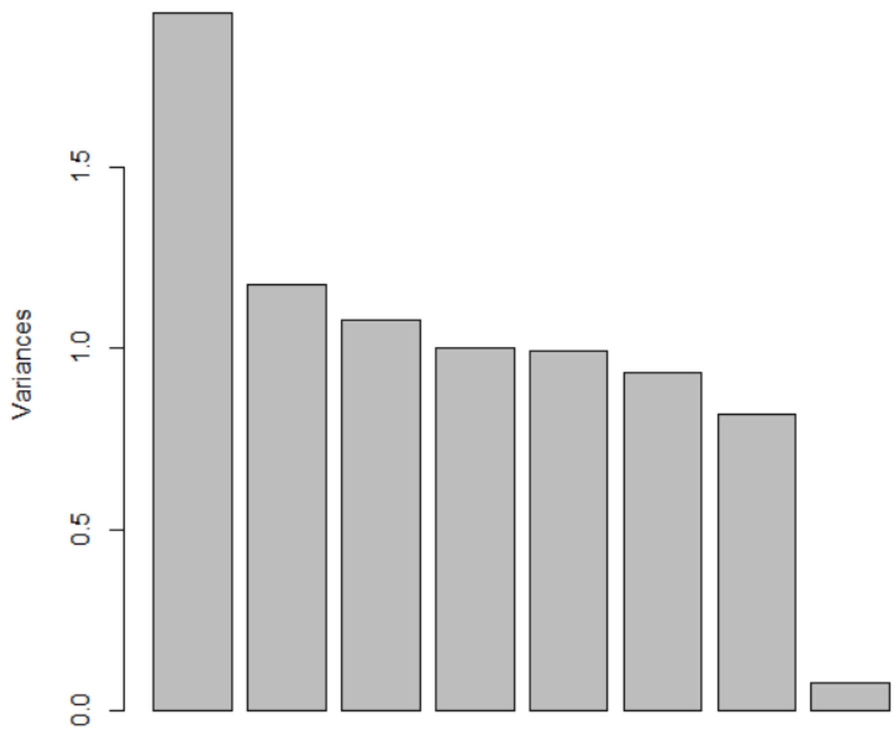


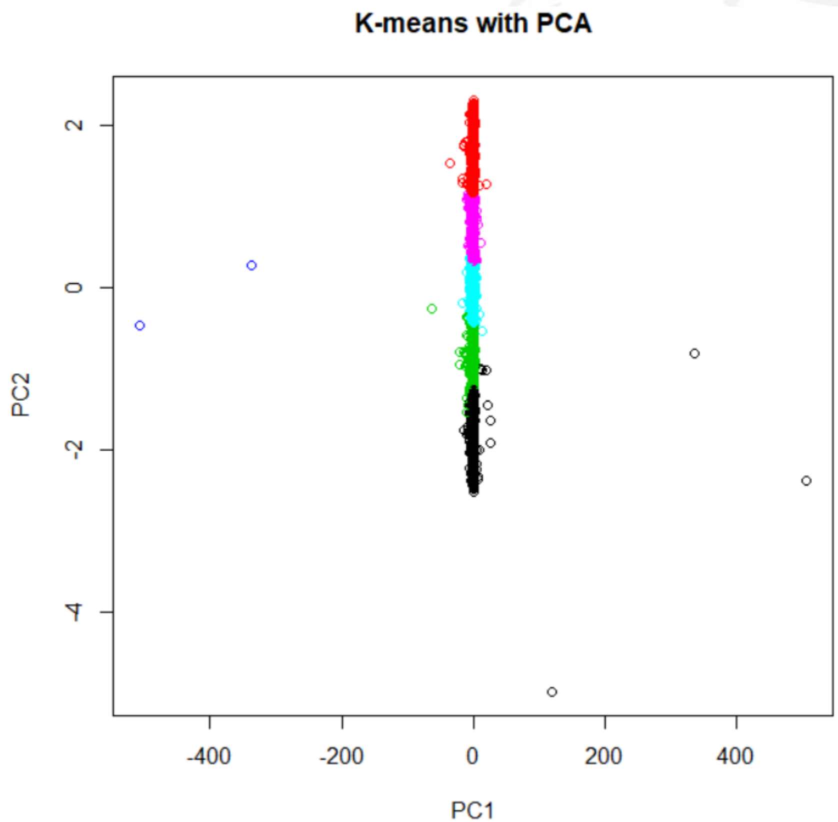



# Clustering

## K-Means with transactions from all countries (apply PCA)

- PCA is applied to original eight features and the first two components are selected
- The adjusted rand index is 0.5
- PCA does not help to improve the clusters

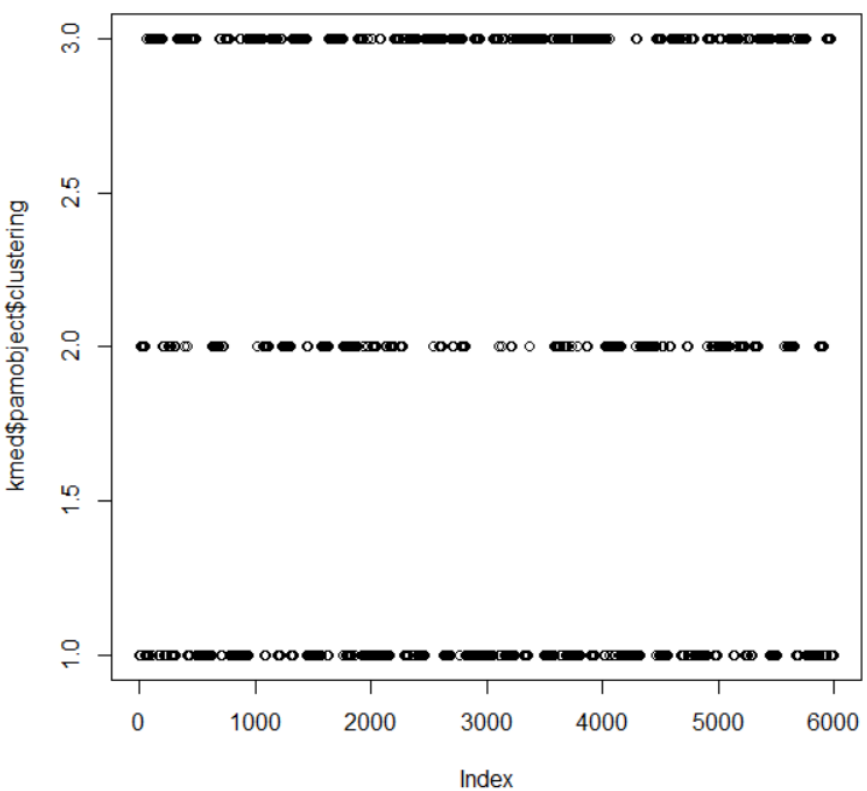


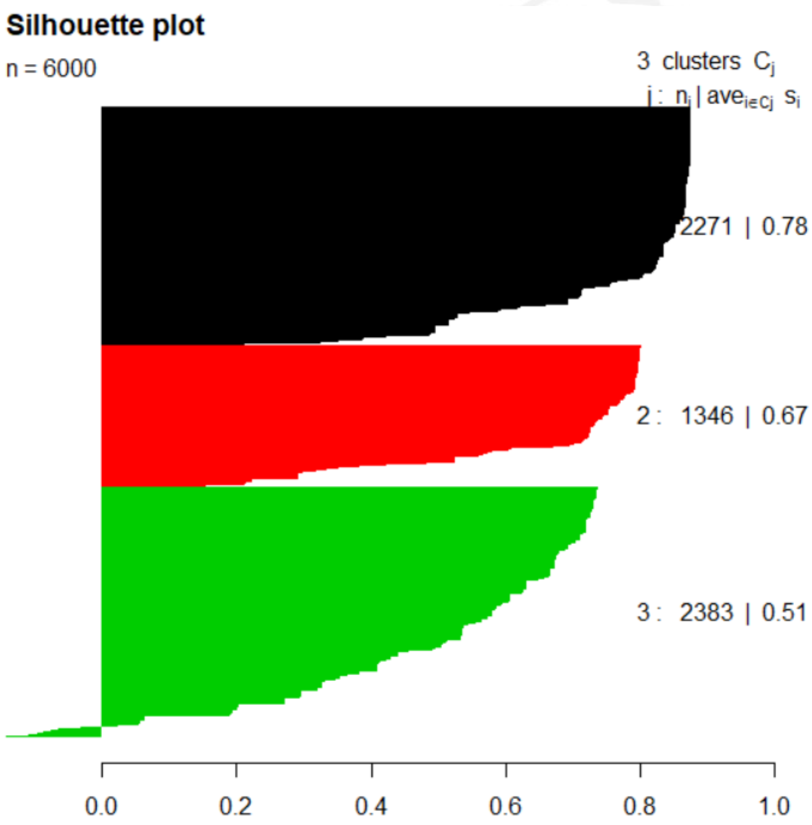



# Clustering

## K-Medoids with transactions from all countries

- 6000 observation is randomly selected due to the size limit
- The suggested k is 3
- Silhouettewidth is 0.65

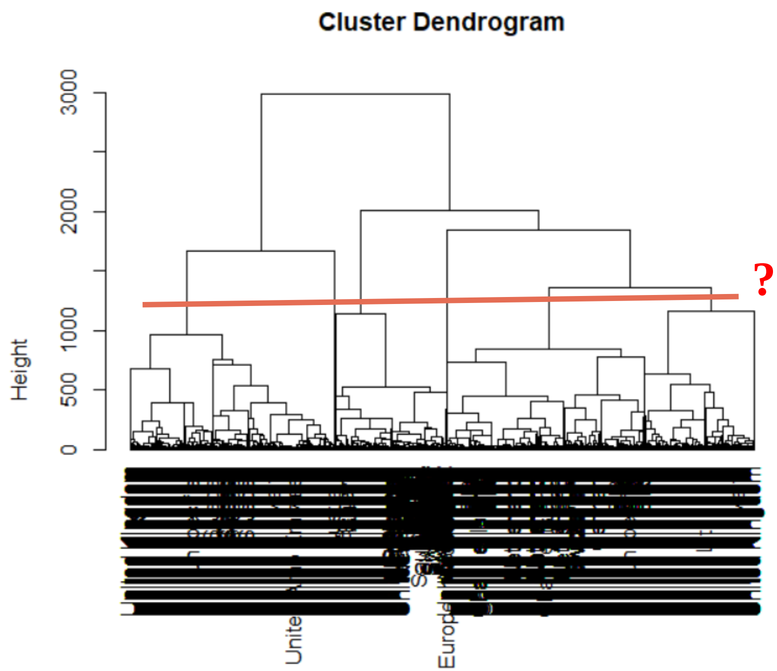


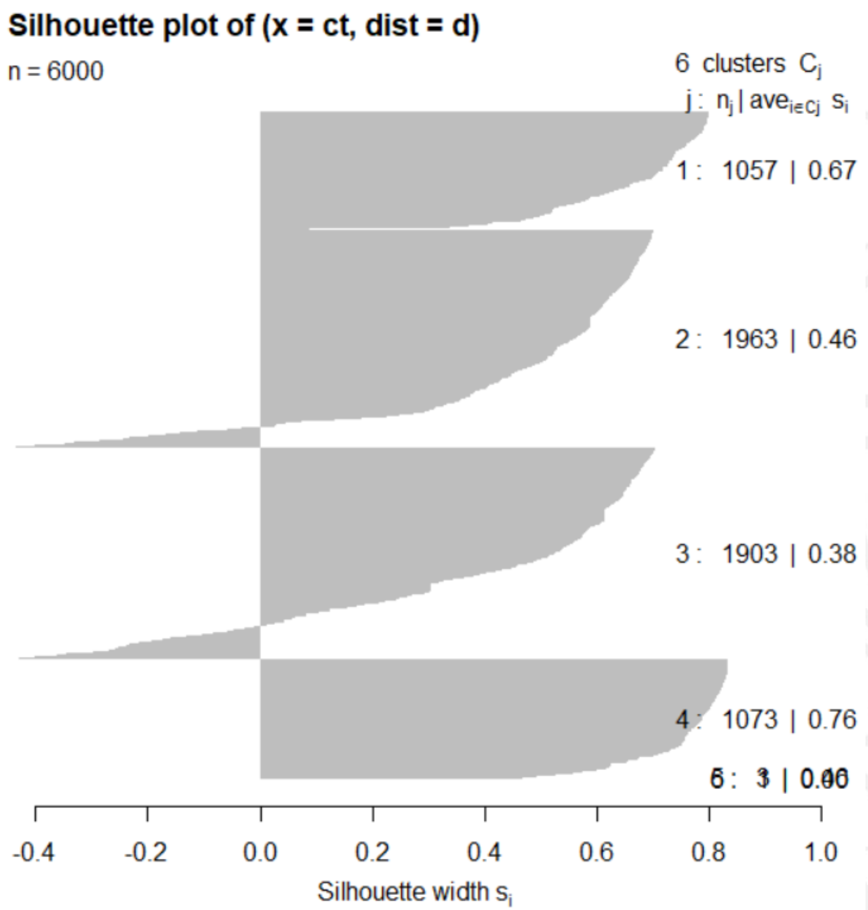



# Clustering

## Hierarchical clustering with transactions from all countries

- 6000 observation is randomly selected
- The average method is used
- The suggested k is 6 from dendrogram
- Silhouette width is 0.53

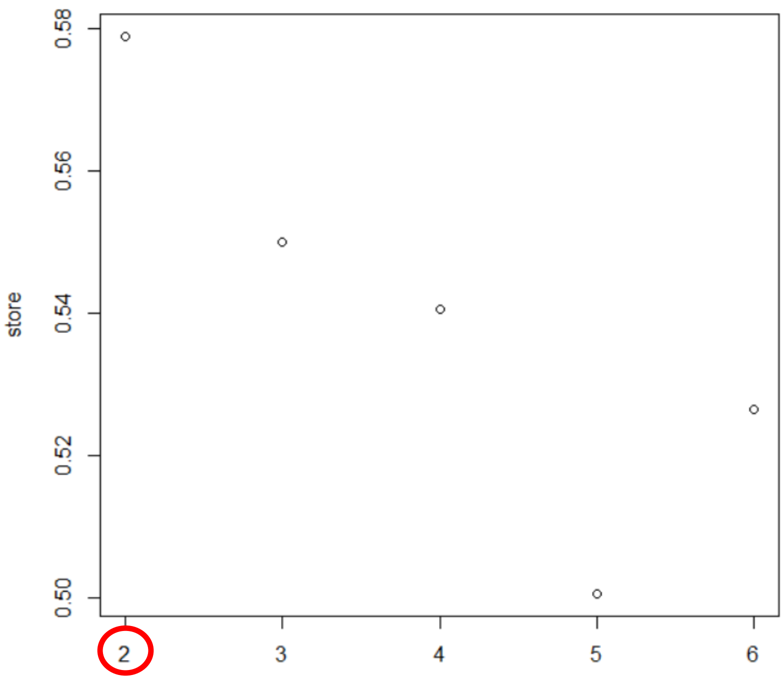


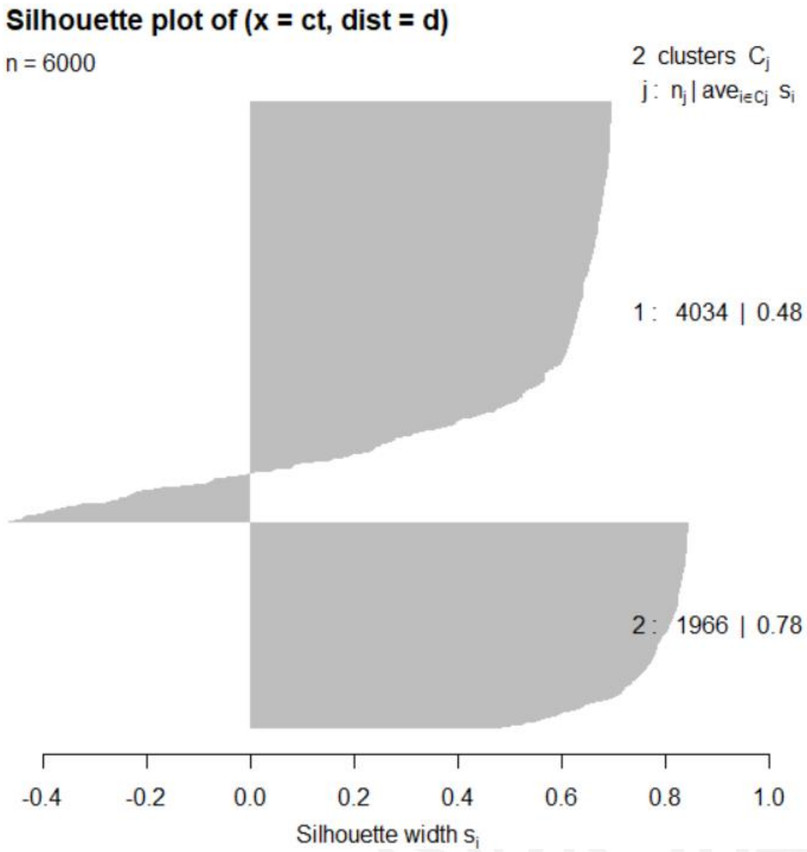



# Clustering

## Hierarchical clustering with transactions from all countries

- Based on Silhouette width, the optimal k is 2
- Silhouette width is 0.58

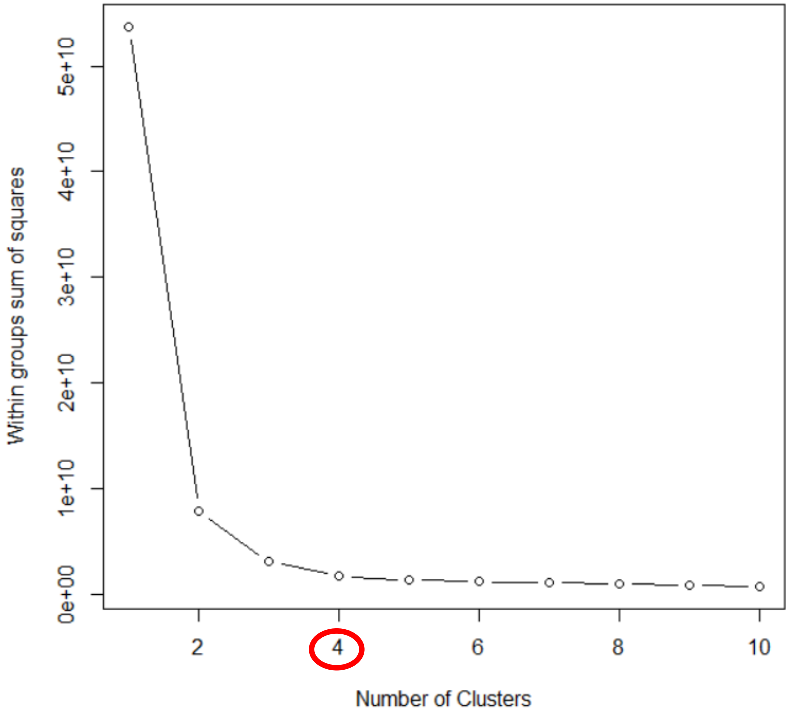


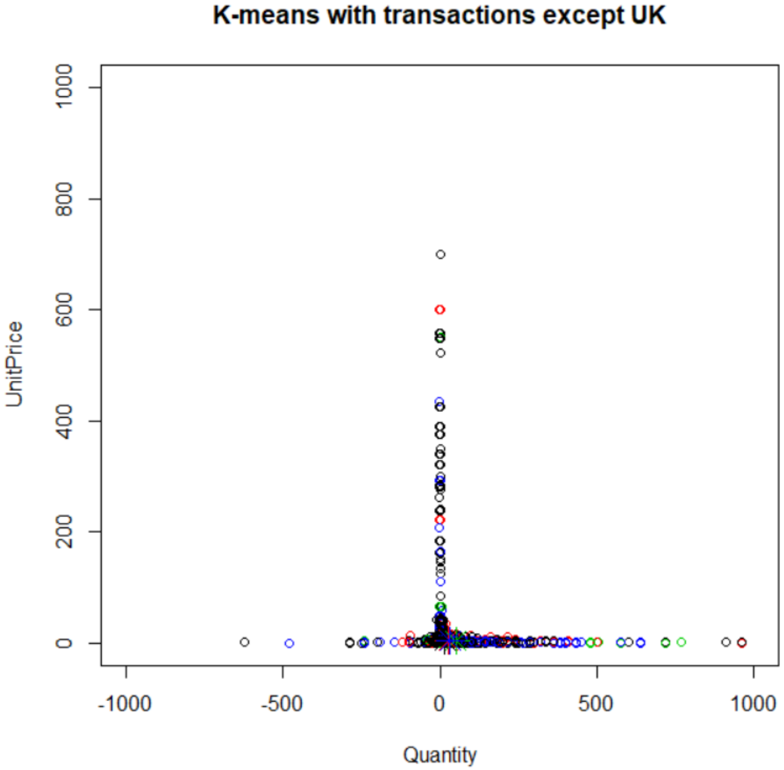



# Clustering

## K-Means with transactions except UK

- The total observation is decreased to 44951.
- There are 36 countries
- Four clusters is suggested from sum of square
- The adjusted rand index is 0.51

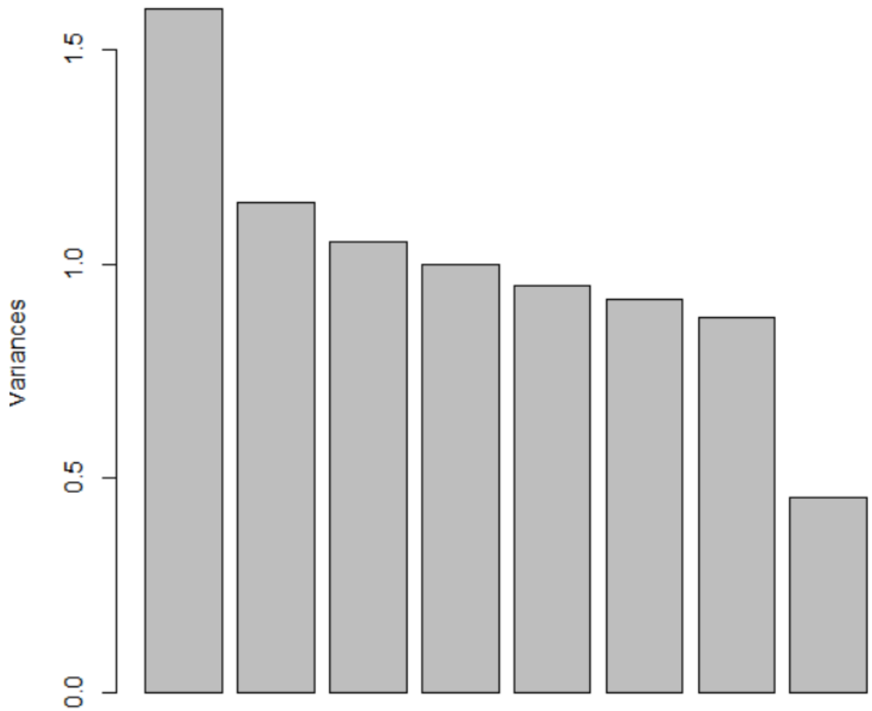


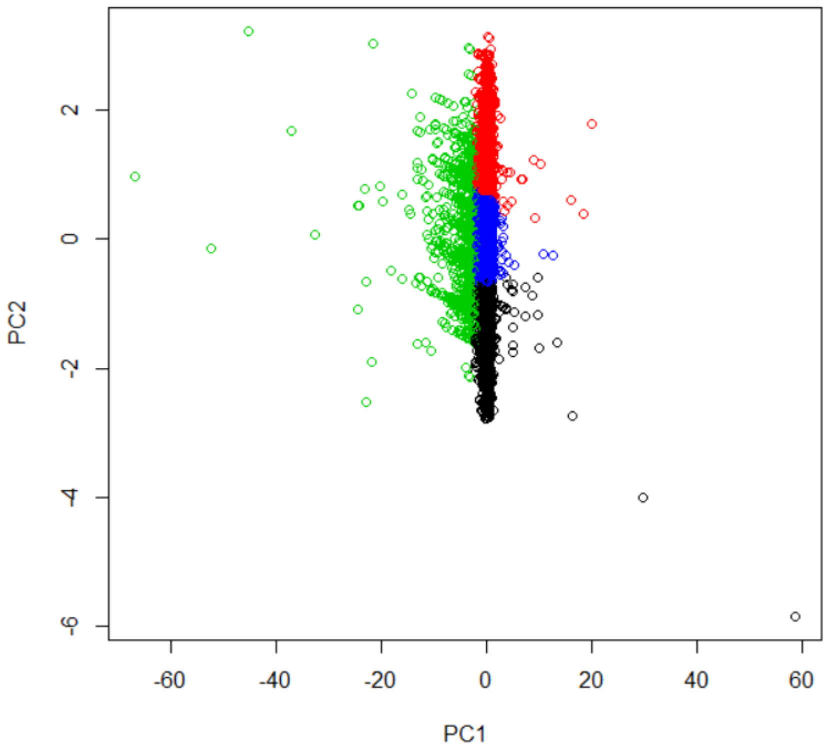



# Clustering

**K-Means with transactions except UK (apply PCA)**

- PCA is applied to original eight features and the first two components are selected
- The adjusted rand index is 0.47
- PCA does not help to improve the clusters

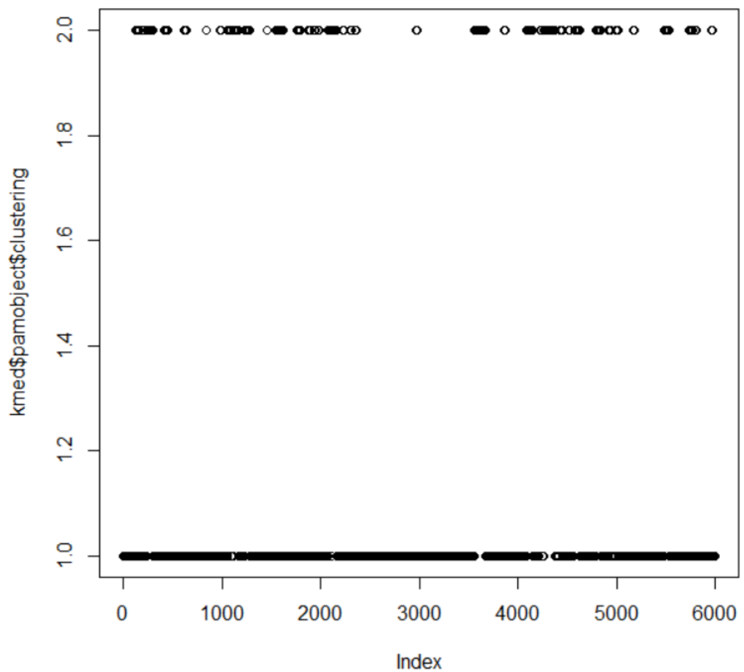


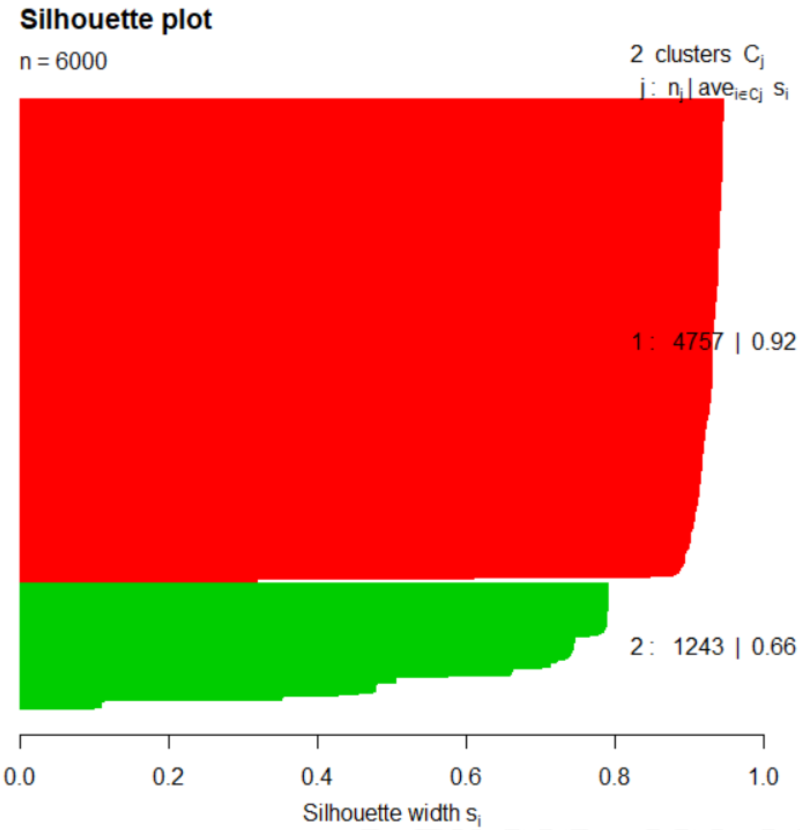



# Clustering

## K-Medoids with transactions except UK

- 6000 observation is randomly selected due to the size limit
- The suggested k is 2 and more observations are in cluster 1.
- Silhouettewidth is 0.87

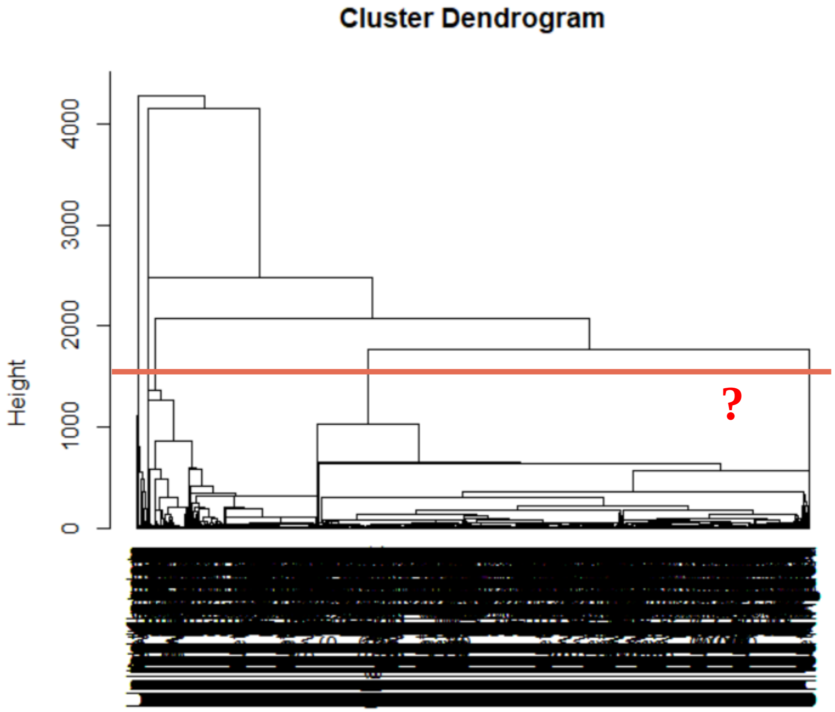


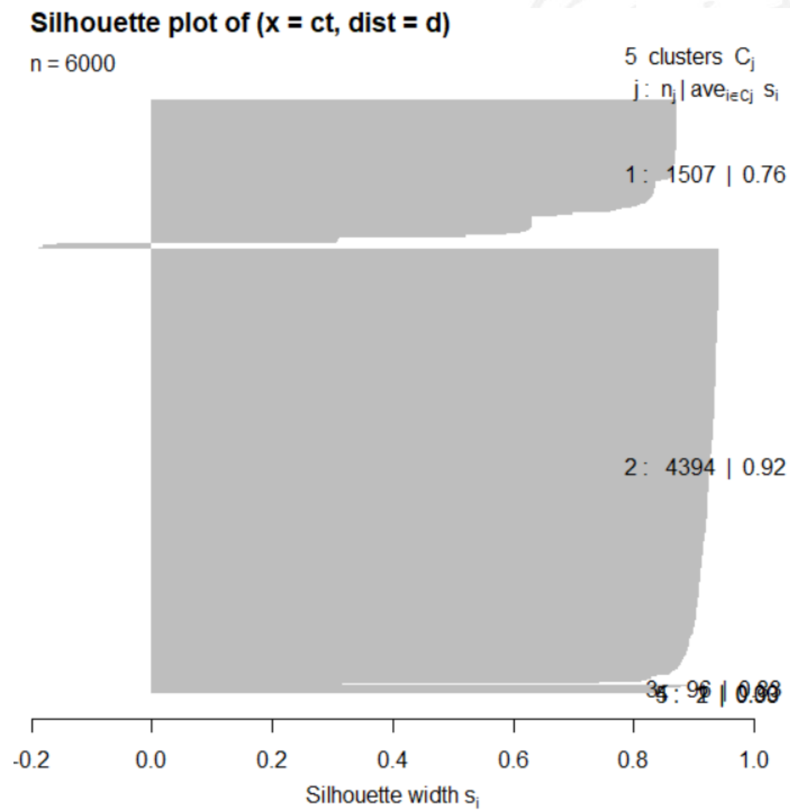



# Clustering

**Hierarchical clustering with transactions except UK**

- 6000 observation is randomly selected
- The average method is used
- The suggested k is 5 from dendrogram, which is the same from Silhouette width
- Silhouette width is 0.88



# Customer segmentation analysis

- Now having seen the working of the clustering and other method, we are now going for something more specialized.

- We would be using the concept of the RFM Analysis.

**RFM Analysis:**

- The RFM Analysis uses the past behavior of the customer or whatever behavior has been observed in the past over different values of the feature in consideration.

- It consists of 3 parts as obvious from the abbreviation:
    1. Recency: The time since the present date and the latest transaction date.
    2. Frequency: The number of transactions that have taken place.
    3. Monetary: The average amount that is being spent over each transaction.

# Customer segmentation analysis

- In our model the 3 features would represented as the:
    1. Recency: (Present Date – max(Invoice Date))
    2. Frequency: total (Invoice)
    3. Monetary: SUM(All bill Amounts) / Frequency
- Analysis Fundamental:

In RFM the customers are segmented into groups or clusters based on the 3 fundamental features of RFM that we just recently covered i.e.

1. Firstly we calculate the RFMs by grouping the data into subsets based on some feature value across which we want to draw the marketing model.
2. Then for each value of the feature across which we are drawing the marketing model we see the combination:

   (R,F,M)

   Higher the RFM better is the better is the feature value.
3. To accomplish this RFM segmentation across groups we draw the clusters using the Hierarchical clustering.

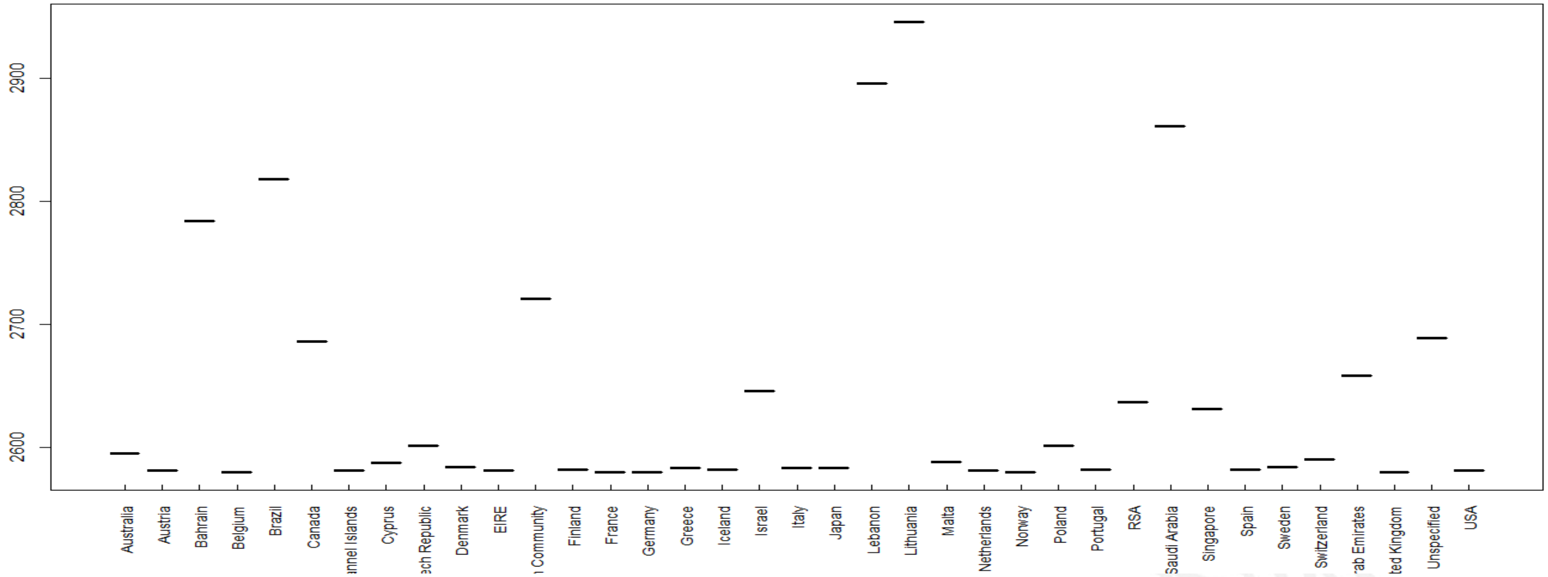
# Customer segmentation analysis

- **Application of RFM in our Modelling of Data sets:**

  **1.** Here firstly we created a new variable/predictor total_dollar ( total_dollar = total_Quantity_bought * total_Unit_Price ) followed by RFM Analysis. Then we observed the results:

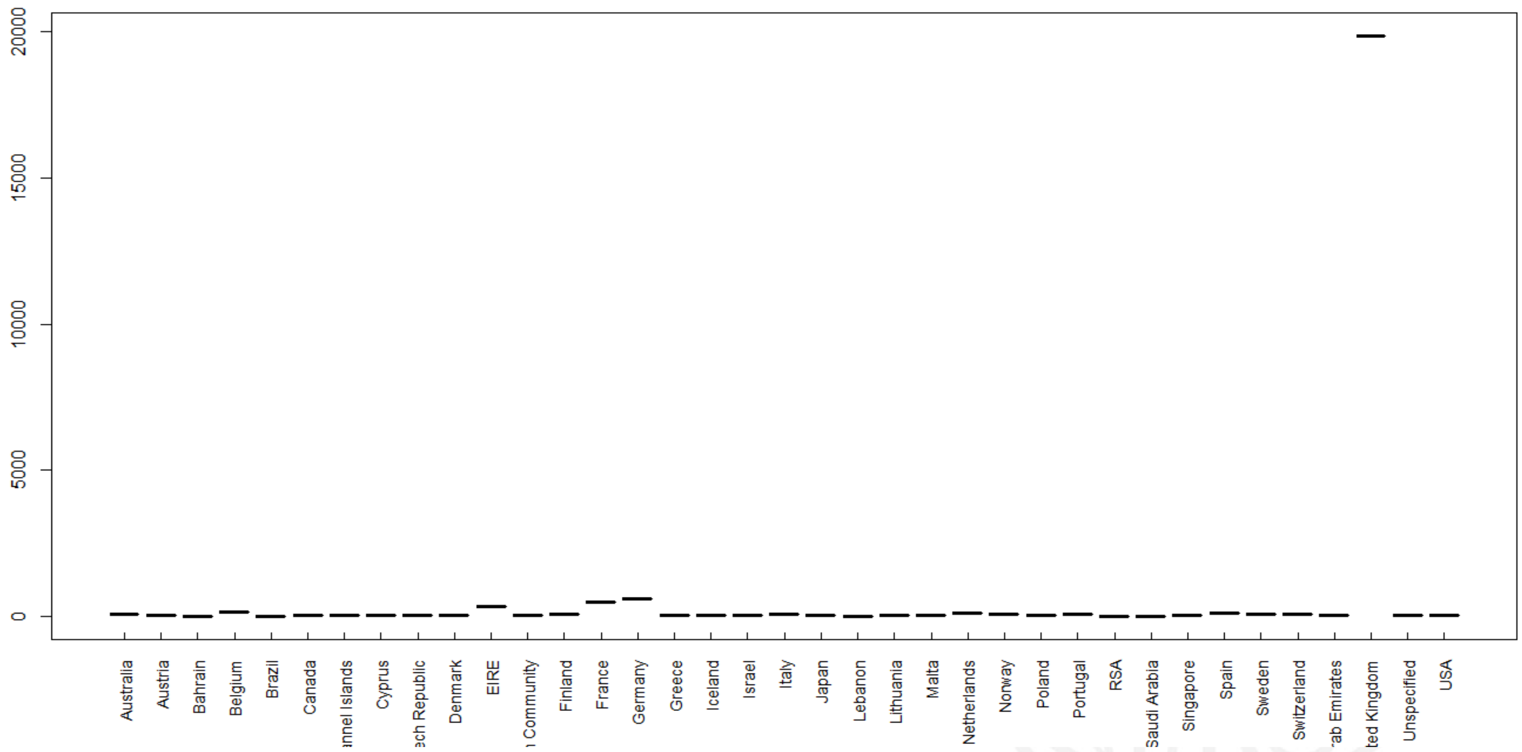
- **RFM Analysis over the Countries:**

  Recency:



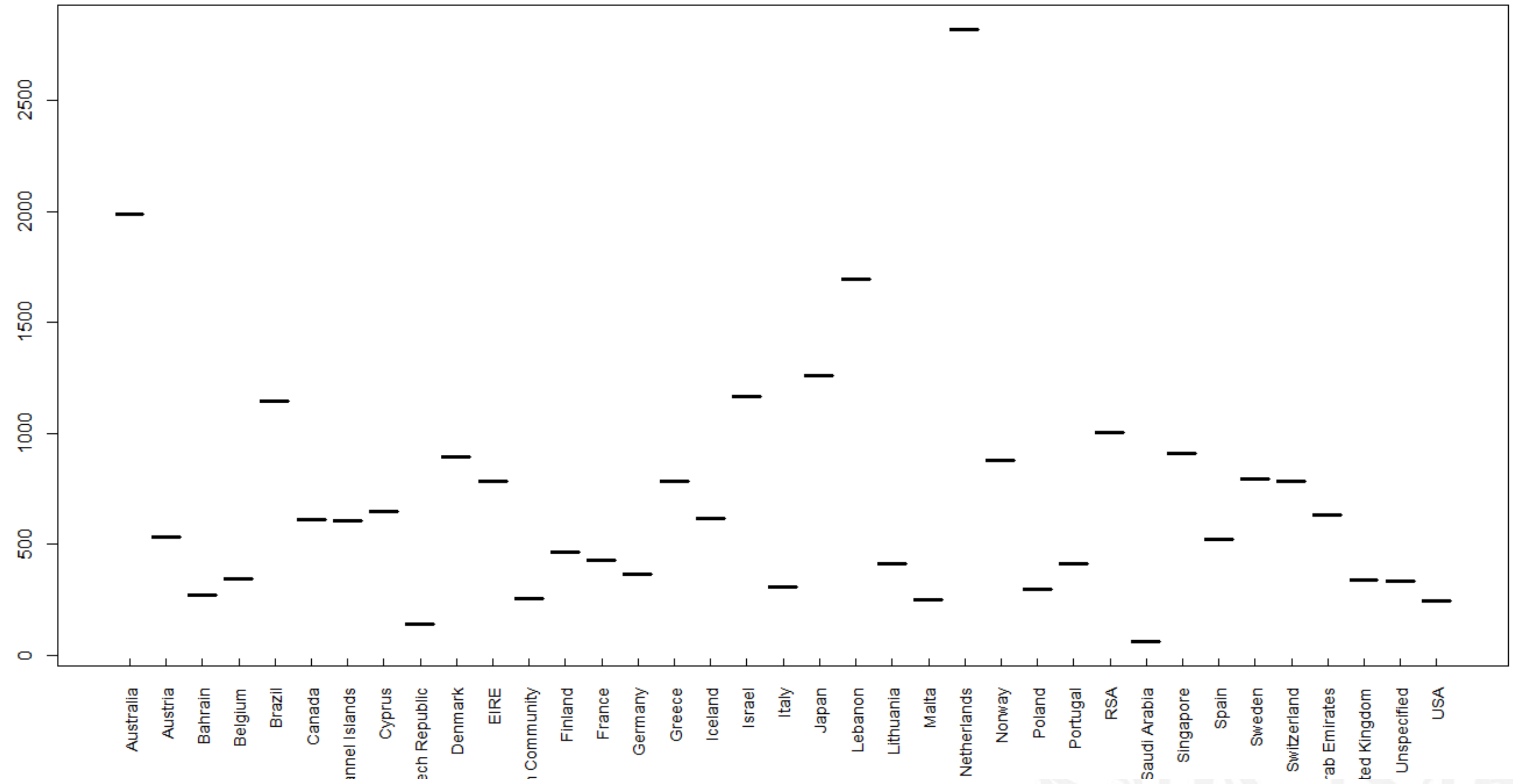
# Customer segmentation analysis

Frequency:



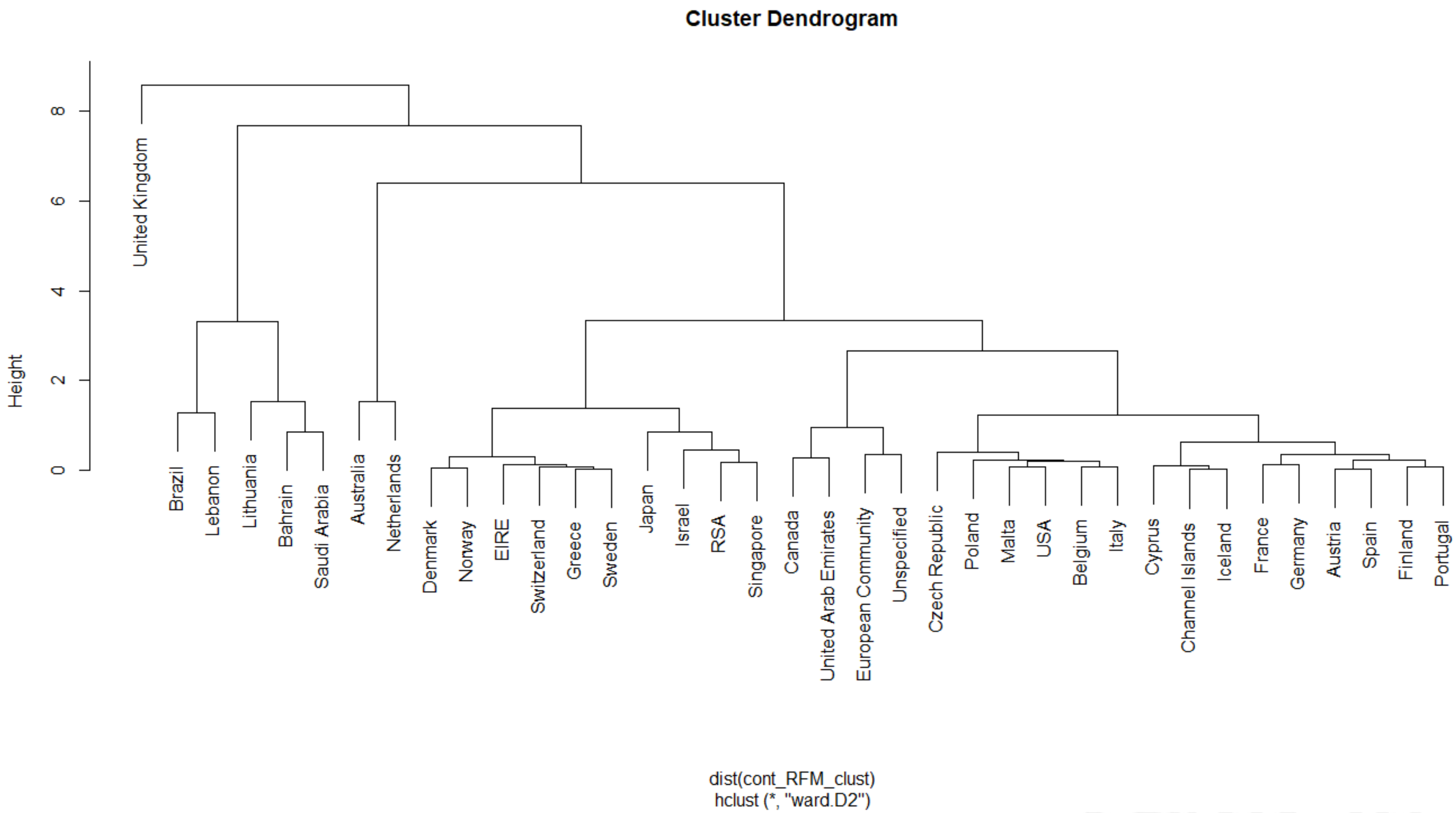
# Customer segmentation analysis

Monetary:



# Customer segmentation analysis

The Hierarchical Clustering based on the RFM:



# Customer segmentation analysis

The Hierarchical Clustering Table based on the RFM:

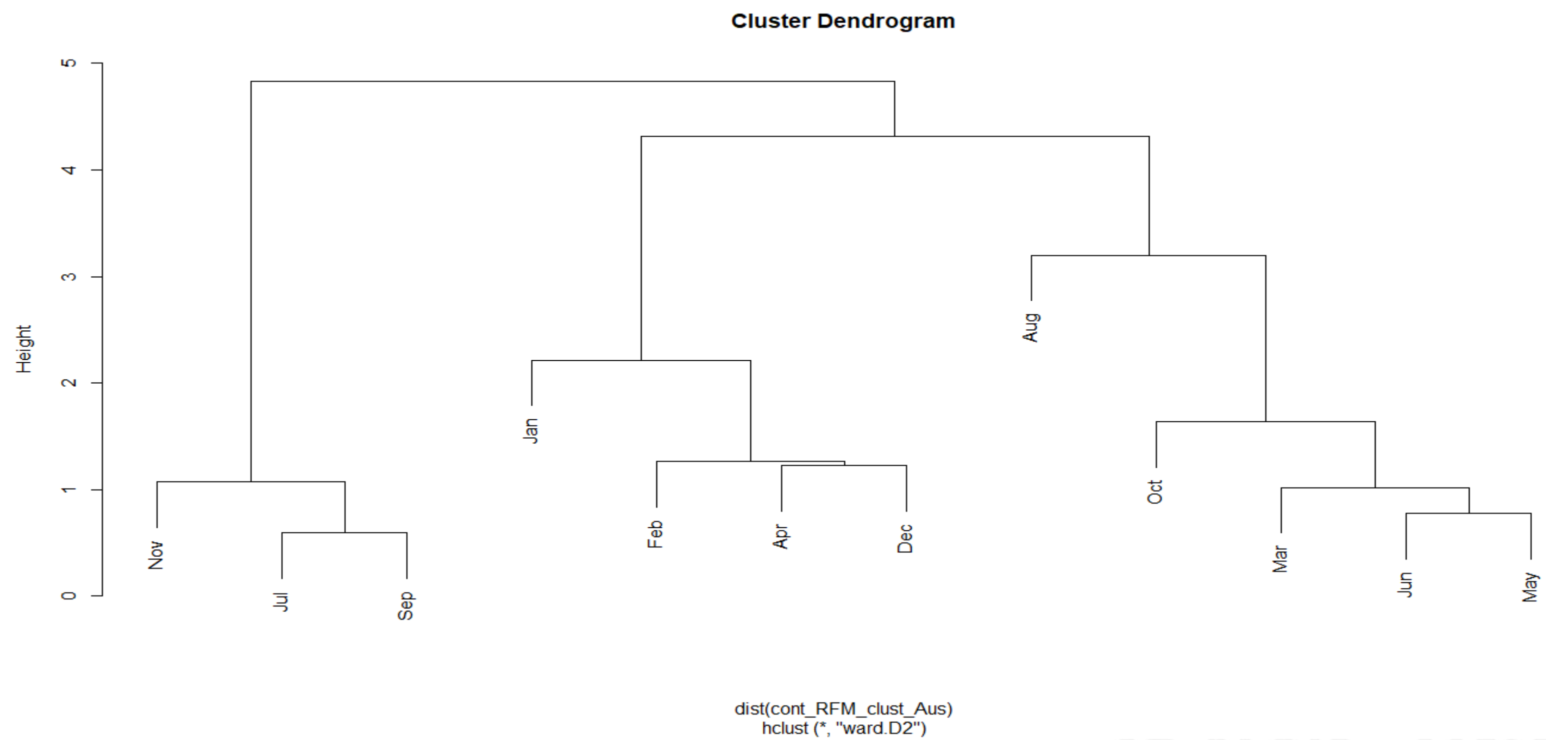
| | Group.1 | recency | freq | montery |
|---|---|---|---|---|
| 1 | 1 | -0.5015678 | -0.1580709 | 3.1031481 |
| 2 | 2 | -0.5347131 | -0.1517652 | -0.5308642 |
| 3 | 3 | 2.2954931 | -0.1833550 | -0.8256586 |
| 4 | 4 | 2.2278495 | -0.1838668 | 1.3061261 |
| 5 | 5 | 0.5181588 | -0.1824849 | -0.4465604 |
| 6 | 6 | -0.3808241 | -0.1673451 | 0.4070452 |
| 7 | 7 | -0.5827401 | 5.9138092 | -0.6628192 |

- Based on the data group with highest RFM is Group 3
- So we would be considering one of the countries falling in the group 3 for further customer analysis.
- We will consider Australia as it has a huge volume of data after UK.

# Customer segmentation analysis

In the Australia we further did RFM on the months to find which month had maximum RFMs :

Here we found the 3 cluster had maximum RFM so we will focus on one of the Months of group 3 i.e. MAY



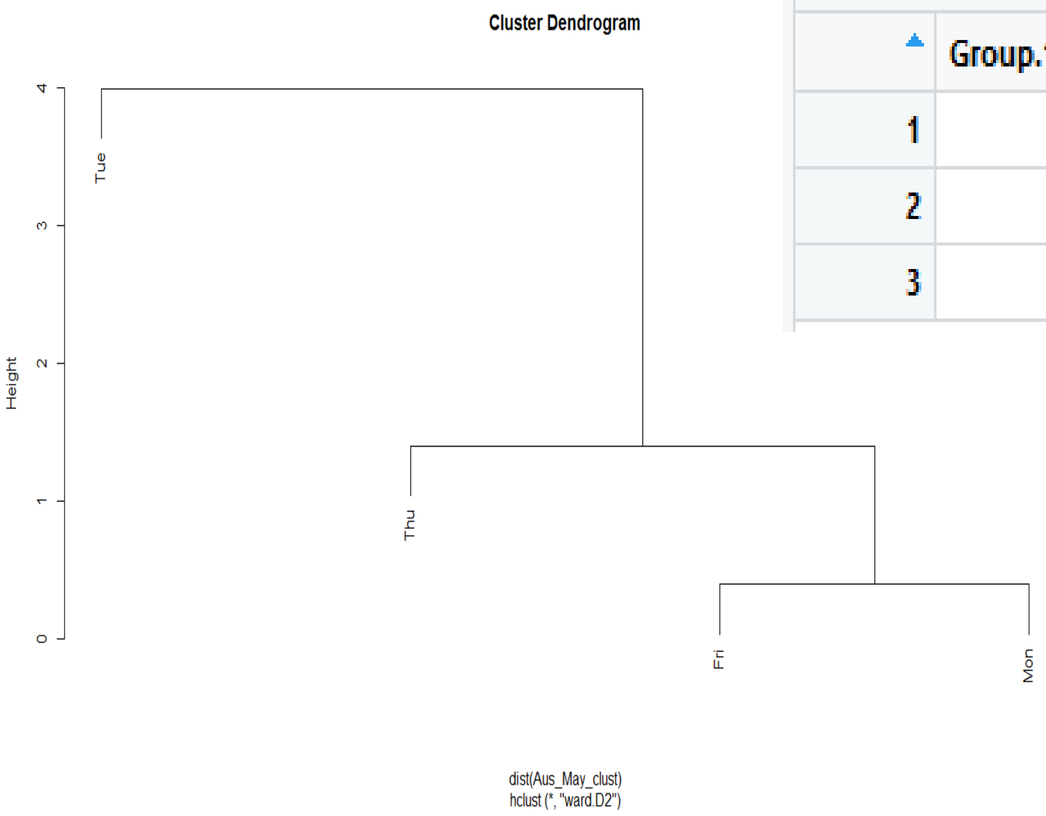
# Customer segmentation analysis

Now we did the customer analysis on the month of MAY for a better idea as which customers have been consistent in month of MAY over the years and hence can be looked upon for patterns in the advertisements and also for special offers

| | CustomerID | recency | freq | montery |
|---|---|---|---|---|
| 1 | 12415 | 2783 | 2 | 6345.58 |
| 2 | 12431 | 2772 | 3 | 312.45 |

# Customer segmentation analysis

We also did a further analysis on the data in general to find which are the days in the month of MAY we can amplify our sales by using the RFM:




| | Group.1 | WeekDay | recency | freq | montery |
|---|---|---|---|---|---|
| 1 | 1 | NA | 2781.5 | 1 | 619.225 |
| 2 | 2 | NA | 2791.0 | 1 | 475.160 |
| 3 | 3 | NA | 2772.0 | 2 | 5957.450 |

Here we found that in Australia in month of MAY the weekdays Monday and Fridays are the maximum RFM days so, should be focused on

# Conclusion

- In data preprocessing part, we analyzed each variable using EAD method. We have observed that 1) most transactions occurred in November; 2) UK had the most transactions followed by Australia, EIRE, France, Germany and Netherlands.
- When doing clustering to non-UK countries, there is slight improvement compared to using transactions from all countries, especially for using K-Medoids and Hierarchical clustering.
- Doing PCA for K-Means does not help in this dataset.
- K-Medoids and Hierarchical clustering perform better than K-Means. This may because there is weak ‘circle’ patterns in data
- The RFM method does not add much to the mix of clustering mechanisms but is just a more specialized method to find the possible opportunities for promotional events.
- By using RFM we are using the concept of the Recency ,Frequency and Monetary to able to gain a much more better perspective on the customer behavior over a range of features and hence find the exact customers /features values when we can boost our sales.

# Questions

- If there is a high possibility that the majority of data are belonging to one cluster and the rest small percent of data belong to multiple clusters, how could this affect clustering models? How to improve it?

- Can RFM be used as a sole method to gain insight into the behavior of customers ,more accurately could I use the method of RFM for complete modelling or only a section of modelling?

# THANK YOU FOR YOUR ATTENTION!

# Q & A!