CUSTOMER SEGMENTATION FOR AN ONLINE RETAIL BUSINESS

Overview

In this project, we will use unsupervised machine learning techniques to segment customers of a UK-based and registered, non-store online retail store based on their purchase behavior. The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers. We will use the Online Retail dataset which contains all the transactions for the online store between 01/12/2010 and 09/12/2011.

Problem Statement

Customer segmentation is the practice of dividing a company’s customers into groups that reflect similarities among customers in each group. The goal of segmenting customers is to make a decision on the way to relate to customers in each segment so as to maximize the worth of every customer to the business.

Our project aims to assist the online retail business understand it's customers' segments based on purchase behavior and assist the company in predicting the right group of new customers for their advertising campaigns.

1. Defining the Question

Specifying the Question

Which customer segments do we have based on purchase behavior?

Defining the Metric for Success

This research will be considered successful when we will be able to come up with clusters that are easily understandable and interpretable for the business based on customers' purchase behavior.

Research Objectives

Main Objective

We aim to segement the Customers based on RFM using Kmeans clusters so that the company can target its customers efficiently.

Specific Objectives

1. To find the relationship between country and item mostly purchased
2. To identify the country with the highest sales in a particular year
3. To perform Kmeans clustering on RFM metrics of the data
4. To identify the most popular products
5. To identify customers who frequently make purchases on the online retail store

Relevance of the data

This dataset is relevant in our research of identifying and segmenting customers based on their purchase behaviour. It includes variables such as total cost of each transaction, the quantity of items purchased, the date and time of the transaction, the customer ID and the country of customer's residence.

2. Reading and checking the data

```
(541909, 8)
```

Observations

> The dataset has 541909 rows and 8 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype         
---  ------       --------------   -----         
 0   InvoiceNo    541909 non-null  object        
 1   StockCode    541909 non-null  object        
 2   Description  540455 non-null  object        
 3   Quantity     541909 non-null  int64         
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64       
 6   CustomerID   406829 non-null  float64       
 7   Country      541909 non-null  object        
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

Observations

> From the output, the dataset has some null values on the description and customerID columns. All other columns do not have any null values. The CustomerID column does not have the appropriate datatype while all other columns have the appropriate datatype.

Explaining the Columns

1. InvoiceNo: Invoice number - A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
2. StockCode: Product (item) code -A 5-digit integral number uniquely assigned to each distinct product.
3. Description: Product (item) name.
4. Quantity: The quantities of each product (item) per transaction.
5. InvoiceDate: Invoice date and time - The day and time when a transaction was generated.
6. UnitPrice: Unit price - Product price per unit in sterling (Â£).
7. CustomerID: Customer number - A 5-digit integral number uniquely assigned to each customer.
8. Country: Country name - The name of the country where a customer resides

3. External Data Source Validation

Dr. Daqing Chen, Course Director: MSc Data Science. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK. link

4. Data Cleaning

Checking for duplicates

Observations

> There are a total of 5268 duplicated values in the dataset.

Dropping the duplicates

Observations

> After dropping the duplicated values, the dataset has 536641 rows and 8 columns.

Checking missing values

```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

Observations

> Our dataset has a total of 1454 missing values in the description column and 135080 in the Customerid column.

Removing missing values

Observations

> The new dataframe has 406829 rows and 8 columns

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

Observations

> All the missing values have been dropped.

Removing canceled invoice numbers

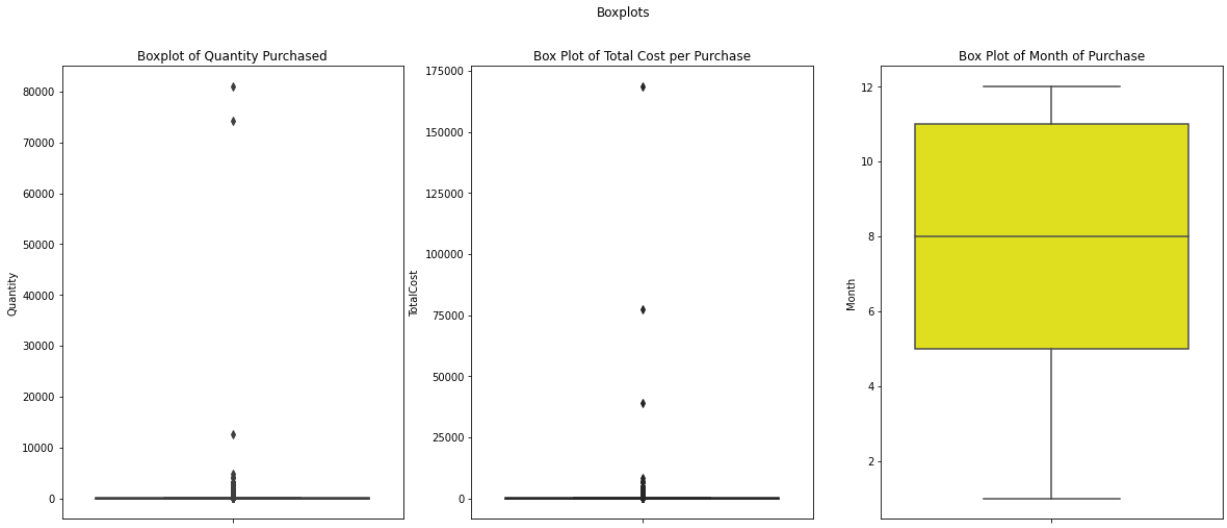
Observations

> After removing the canceled invoice numbers, the new dataframe has 397924 rows and 8 columns

Checking for Outliers

```
<Figure size 1440x576 with 3 Axes>
```

Download



Observations

> The output of the boxplots show that there are outliers. We will go ahead and check how many records of outliers we have and decide whether to remove them or use the data as is. We will deal with the outliers later

```
10.0
15.119999999999997
6.0
```

Observations

> The output shows that there are 10, 15.11999 and 6 records that are outside the interquatile ranges in the variables, Quantity, TotalCost and Month respectively. This data is useful in our reasearch and will therefore not be removed.

Checking distinctive values

```
InvoiceNo      18536
StockCode       3665
```

```
Description     3877
Quantity         302
InvoiceDate    17286
UnitPrice        441
CustomerID      4339
Country           37
TotalCost       2940
Month             12
dtype: int64
```

5. Exploratory Data Analysis

We grouped the data by the "Country" and "Description" columns and summed the quantity of each item. We then sorted the data by the quantity of each item and got the top item for each country. Finally, we used the seaborn library to create a bar plot that shows the relationship between country and the item mostly purchased.

```
<Figure size 432x288 with 1 Axes>
```

Download

We grouped the dataframe by country and sum the total sales for each country in both 2010 and 2011. We then find the country with the highest sales in 2010 and 2011 using the idxmax() function and print out the results.

```
Country with highest sales in 2010: United Kingdom
Country with highest sales in 2011: United Kingdom
```

Country with highest sales in 2010 and 2011 is the United Kingdom.

We grouped the data by product description and summed the quantity of each product, resulting in a new dataframe that shows the total quantity of each product sold. We then sorted this dataframe by the quantity in descending order and printed the top 10 most popular products by quantity sold.

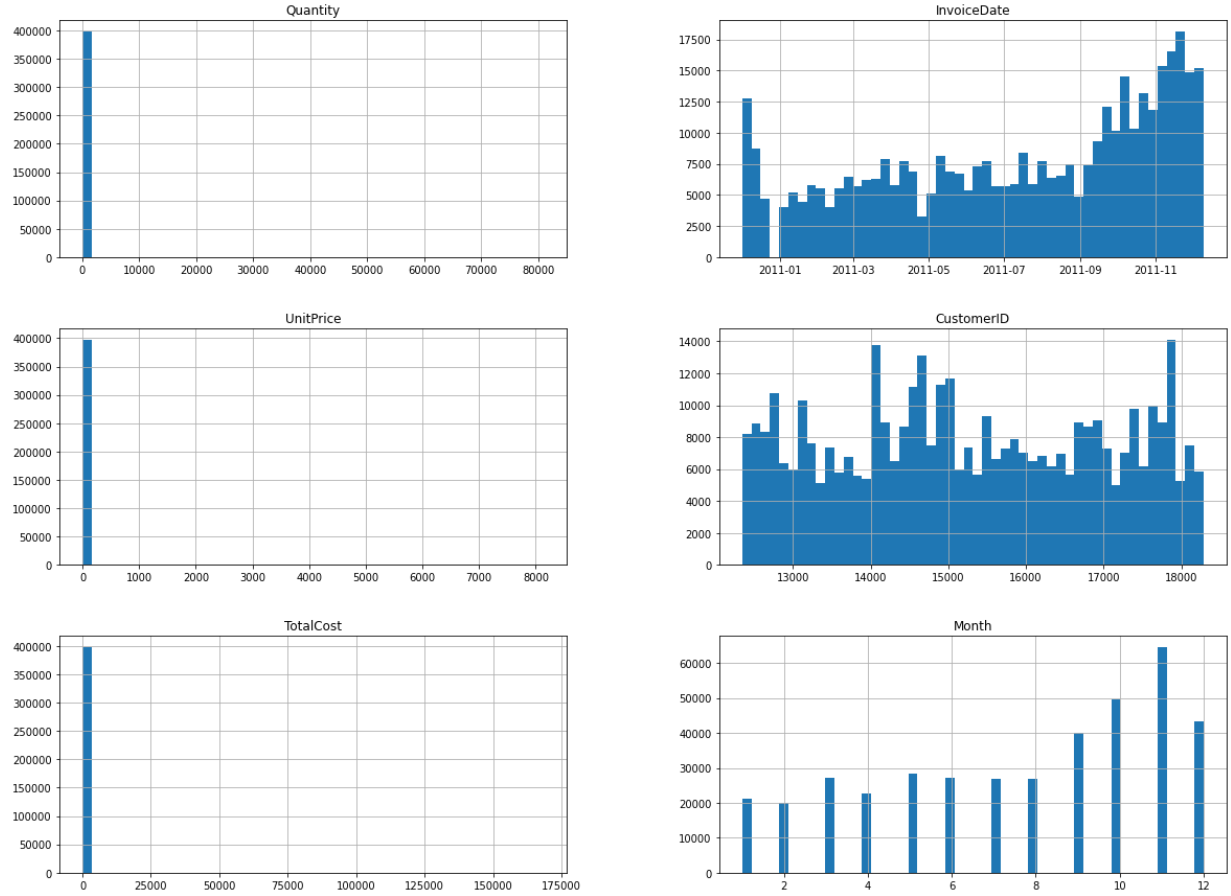
```
Description
PAPER CRAFT , LITTLE BIRDIE           80995
MEDIUM CERAMIC TOP STORAGE JAR        77916
WORLD WAR 2 GLIDERS ASSTD DESIGNS     54415
JUMBO BAG RED RETROSPOT               46181
WHITE HANGING HEART T-LIGHT HOLDER    36725
ASSORTED COLOUR BIRD ORNAMENT         35362
PACK OF 72 RETROSPOT CAKE CASES       33693
POPCORN HOLDER                        30931
RABBIT NIGHT LIGHT                    27202
MINI PAINT SET VINTAGE                26076
Name: Quantity, dtype: int64
```

PAPER CRAFT , LITTLE BIRDIE is the most popular product.

We plotted histograms to show the distribution of each feature in the data set

```
<Figure size 1440x1080 with 6 Axes>
```

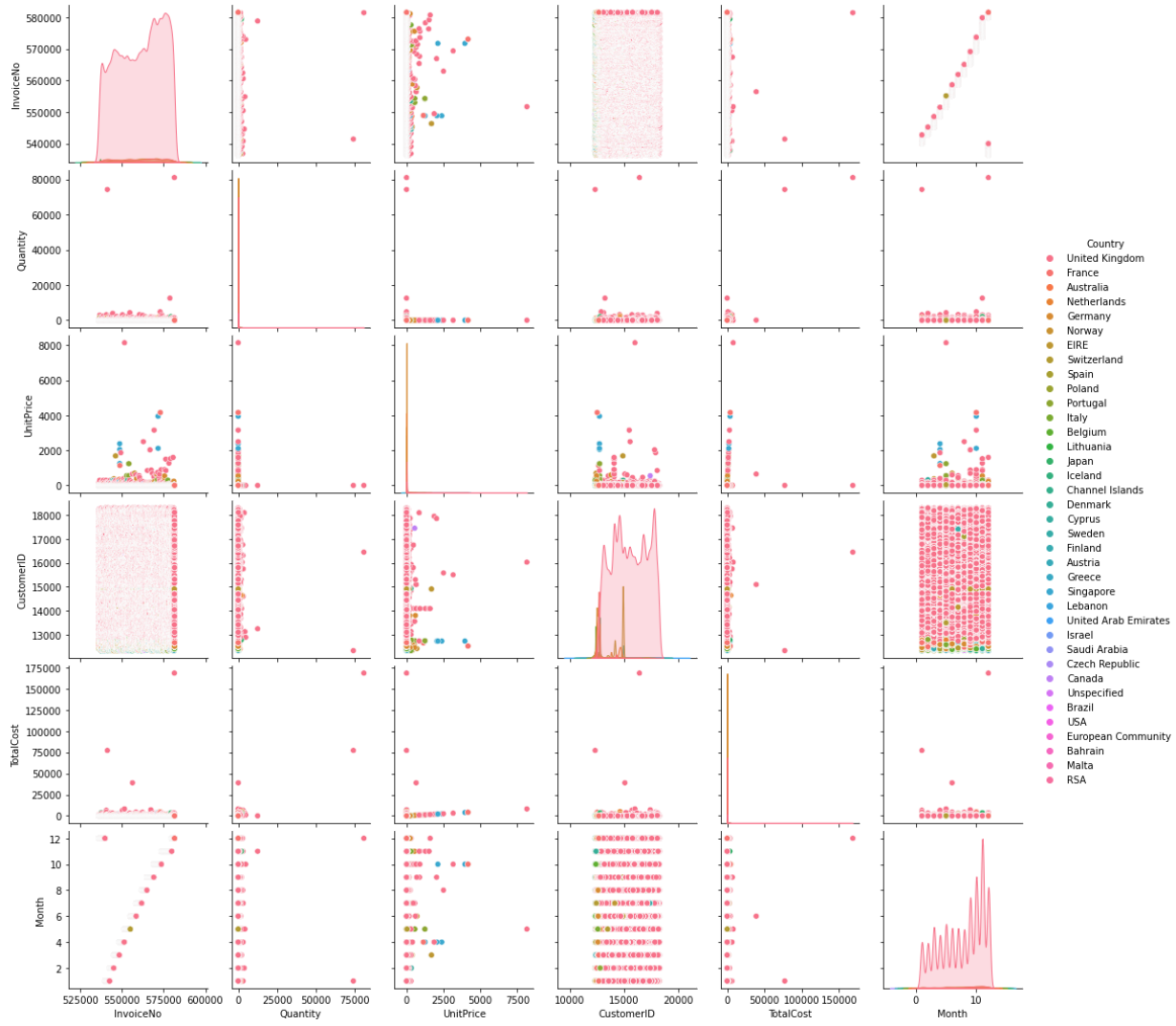
Download



We plotted a pairplot to visualize information about the distribution of the data and relationships between variables

```
<Figure size 1224x1080 with 42 Axes>
```

Download

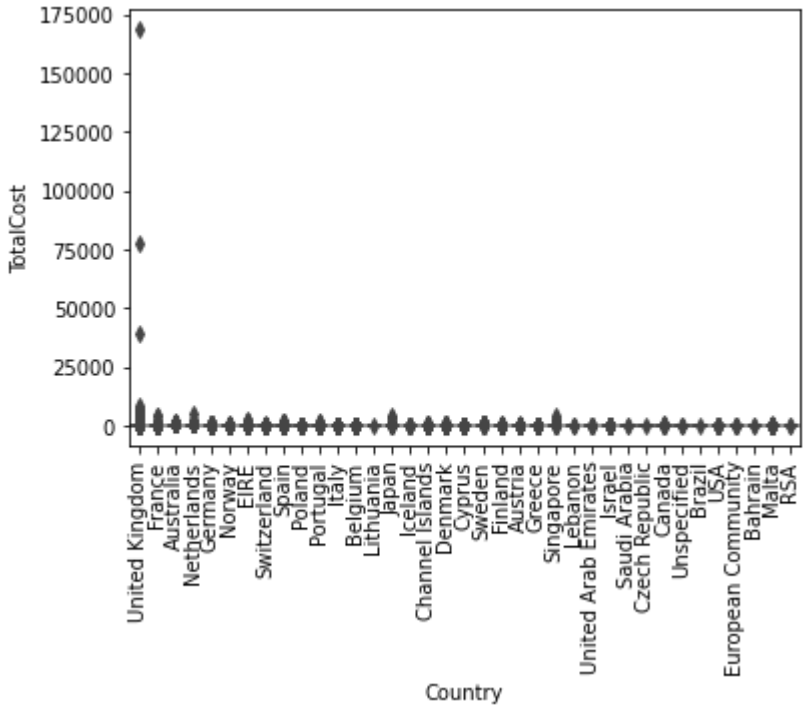


we observed that the distribution of total cost for customers from the United Kingdom appears to be more spread out than for customers from France, as seen in the scatter plots on the diagonal. Additionally, one could observe a positive correlation between quantity and total cost for customers from Germany, as seen in the scatter plot between those two variables

We then plotted boxplots for the various features

```
<Figure size 432x288 with 1 Axes>
```

Download

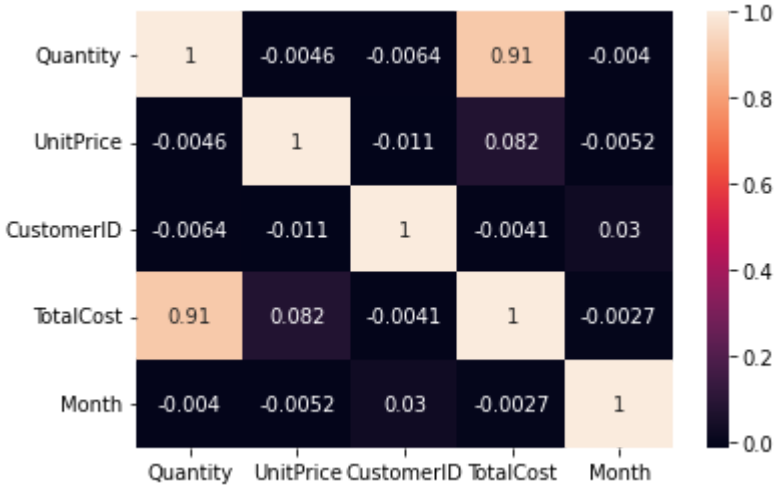


The box plot shows that countries such as France, Germany, and Spain have relatively lower total costs compared to the United Kingdom. The United Kingdom has the highest total costs, which could indicate that it is the country with the most sales and/or the highest average purchase value. The box plot also shows that the spread of total costs is greater for the United Kingdom compared to other countries. The presence of outliers in the box plot indicates that there are some very high total cost transactions for certain countries.

The plotted heatmap below was for visualizing the strength and direction of the relationship between different variables in the data.

```
<Figure size 432x288 with 2 Axes>
```

Download



The strong positive correlation (0.91) between the two variables, Quantity and TotalCost, would indicate that as the quantity of items purchased increases, so does the total cost of the purchase.

#**6. Data Preparation**

##RFM Metrics **Creating a new dataframe "rfm" for kmeans cluster and segment analysis.** We are going to analyse the Customers based on below 3 factors:

- R (Recency): Number of days since last purchase
- F (Frequency): Number of tracsactions
- M (Monetary): Total amount of transactions (TotalCost)

Creating a new dataframe rfm1

```
rfm1=rfm
```

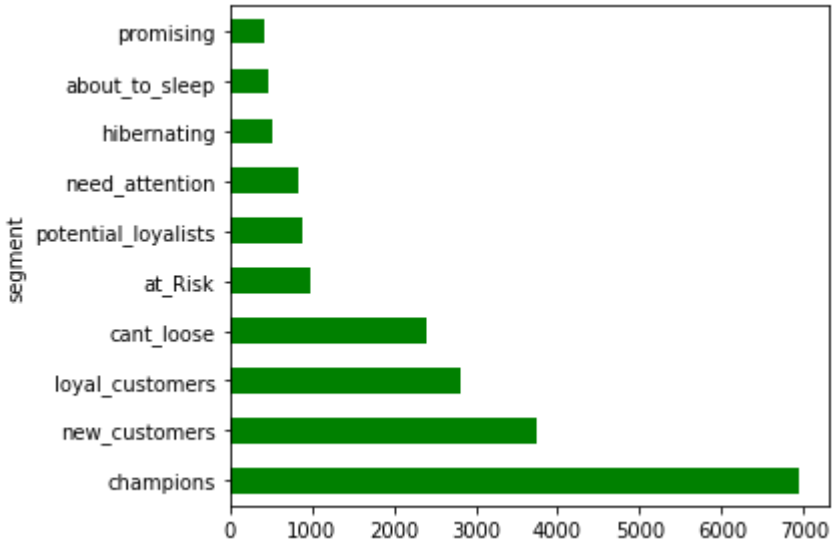
Now our RFM1 dataframe is ready to be used.

Segment variable is representing each customer's segmentation status. So we can get the descriptive statistics based on Segment variable.

This statistics are better understood when graphically showed.

```
<Figure size 432x288 with 1 Axes>
```

Download



As we can see, champions segment is bringing us the largest amount of revenue. Loyal_customers and cant_loose are following champions segment.

We then calculated the ratio of the customer segments compared to the total sum of revenue and the number of customers within segments and ratio of them in total customer numbers.

segment	TotalCost	monetary_ratio	number	number_ratio
champions	4218731.990	47.340802	607	13.989398
loyal_customers	2337784.570	26.233616	828	19.082738
at_Risk	569172.930	6.387015	578	13.321042
hibernating	556892.142	6.249205	1049	24.176077
potential_loyalists	454701.401	5.102464	505	11.638626
new_customers	205817.320	2.309594	55	1.267573

```
cant_loose            193875.051        2.175583      81      1.866790
need_attention        175565.740        1.970123     207      4.770684
about_to_sleep        152570.290        1.712078     321      7.398018
promising              46296.470        0.519519     108      2.489053
```

	TotalCost	monetary_ratio	number	number_ratio
segment				
champions	4218731.990	47.340802	607	13.989398
loyal_customers	2337784.570	26.233616	828	19.082738
at_Risk	569172.930	6.387015	578	13.321042
hibernating	556892.142	6.249205	1049	24.176077
potential_loyalists	454701.401	5.102464	505	11.638626
new_customers	205817.320	2.309594	55	1.267573
cant_loose	193875.051	2.175583	81	1.866790
need_attention	175565.740	1.970123	207	4.770684
about_to_sleep	152570.290	1.712078	321	7.398018
promising	46296.470	0.519519	108	2.489053

Above is the table of segments, monetary sum(TotalCost) of the segments, monetary ration in total revenue (TotalCost), number of customers per segment and the ratio of it.

Champions segment is representing only %14.00 of total customers while generating some %47.34 of the total revenue. So this is why they are called CHAMPIONS!

Loyal_customers segment is following as the second largest segment about revenue. Representing %19.08 of the total customers, loyal_customers segment is generating %26.23 of the total revenue.

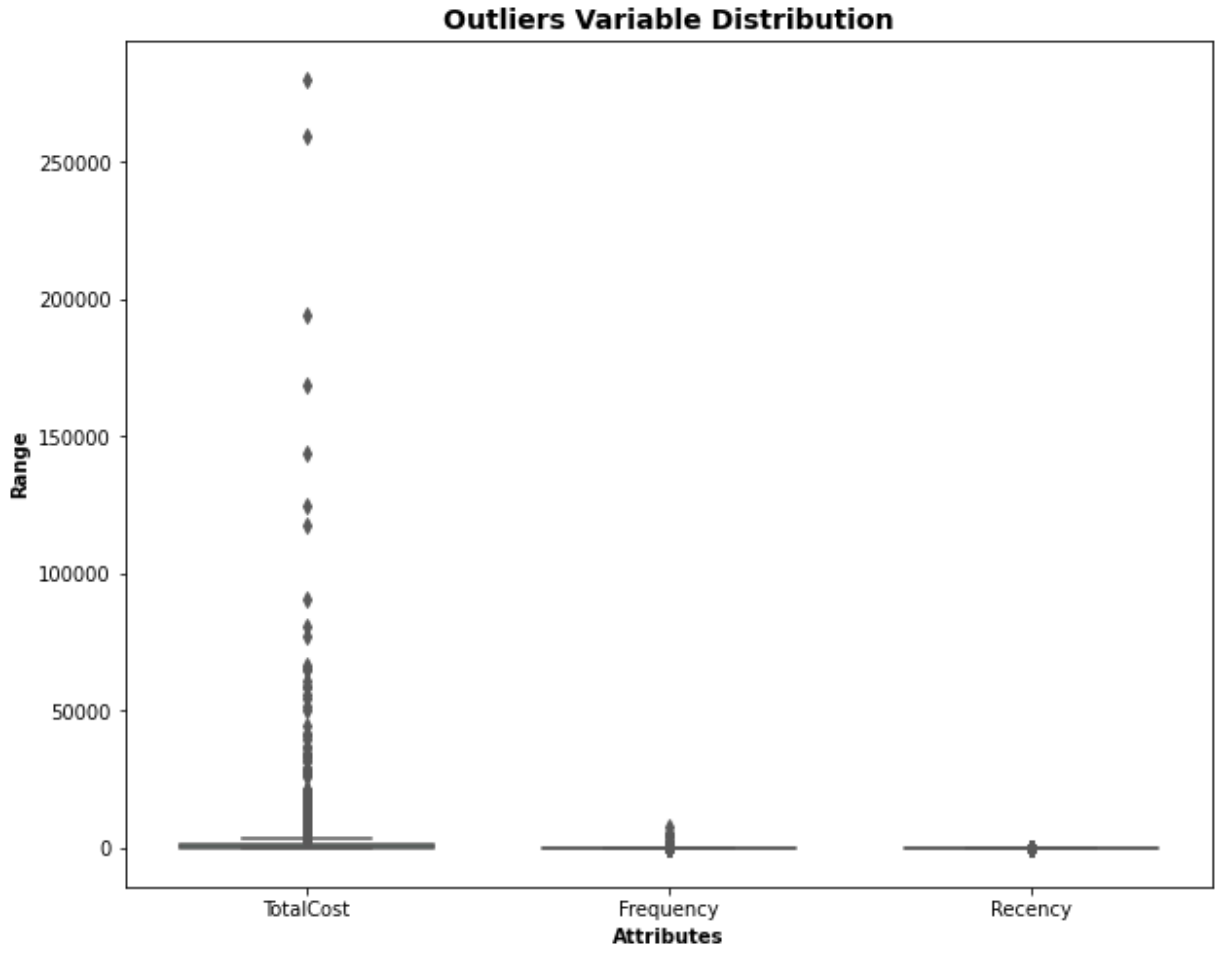
So basically we can assume that %32.72 of the total customers are generating %65.61 of the total revenue.

Outliers

```
Text(0.5, 0, 'Attributes')
```

```
<Figure size 720x576 with 1 Axes>
```

Download



Standardization

Standardization is needed when we want to protect the weights of each value. In this case it's a must!

```
   TotalCost  Frequency   Recency
0   1.645726   1.067832 -0.917872
1   0.262124  -0.460601 -0.189272
2   0.240269  -0.035474 -0.748198
3  -0.543361  -0.602310  2.156221
4   0.652411   0.085991 -0.578524
```

	TotalCost	Frequency	Recency
0	1.645726	1.067832	-0.917872
1	0.262124	-0.460601	-0.189272
2	0.240269	-0.035474	-0.748198
3	-0.543361	-0.602310	2.156221
4	0.652411	0.085991	-0.578524

7. Building the Model

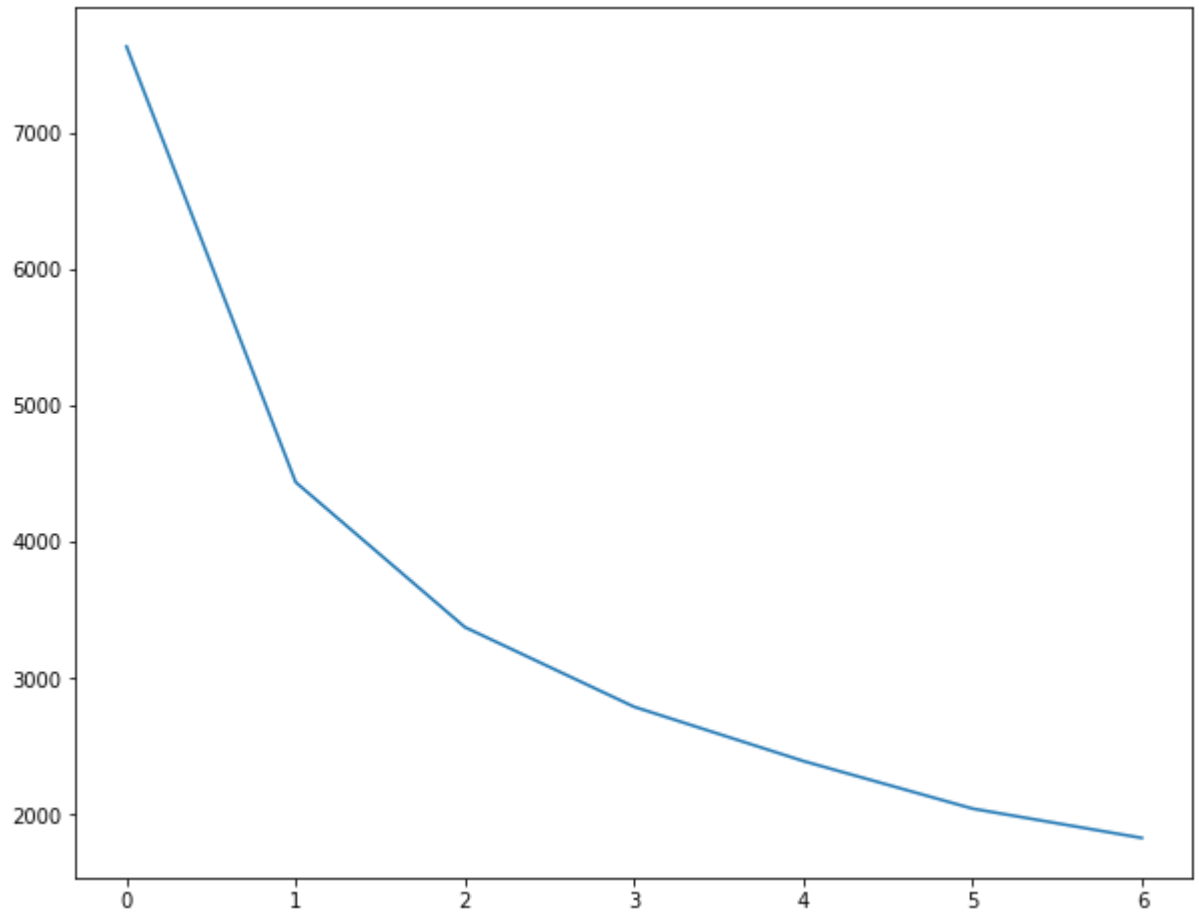
K Means clustering

Finding the Optimal Number of Clusters using The Elbow Method

```
[<matplotlib.lines.Line2D at 0x7ff8bc500460>]
```

```
<Figure size 720x576 with 1 Axes>
```

Download



The drastic change at some point shows the best number of clusters. That's why curve at the 3 clusters shows the elbow like shape hence Optimal number of clusters is 3.

###Cluster validation using Silhouette analysis

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 cluster,

A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

```
For n_clusters=2, the silhouette score is 0.540625413768531
For n_clusters=3, the silhouette score is 0.5087756865796796
For n_clusters=4, the silhouette score is 0.4851654323063987
```

```
For n_clusters=5, the silhouette score is 0.46619818582629496
For n_clusters=6, the silhouette score is 0.41629422231681457
For n_clusters=7, the silhouette score is 0.413435373101441
For n_clusters=8, the silhouette score is 0.40863087267455217
```

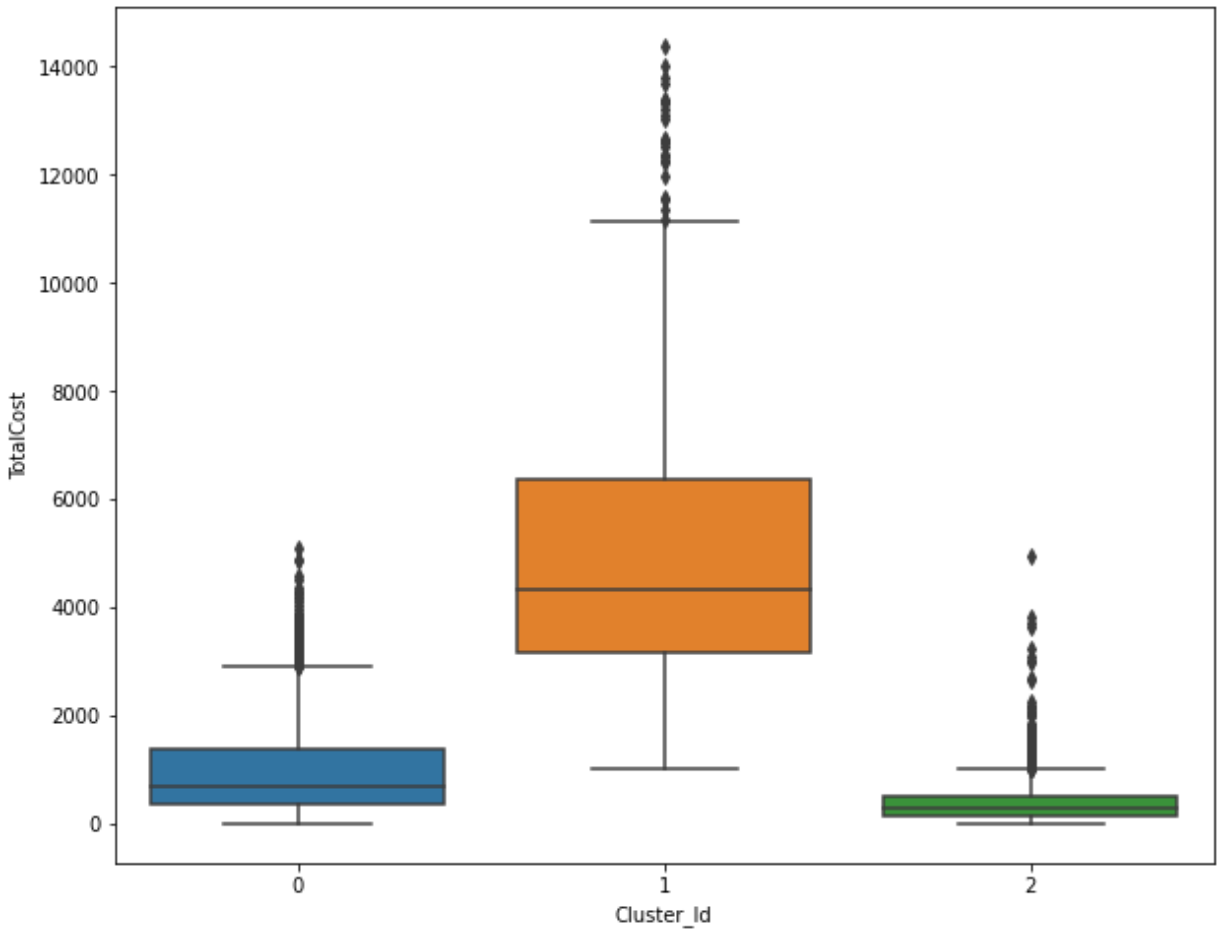
At cluster 3 we have a the silhouette score as 0.5087756865796796 which is statistically sound hence we confirm the validity of using the 3 clusters chosen

Modeling

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff8b6be3970>
```

```
<Figure size 720x576 with 1 Axes>
```

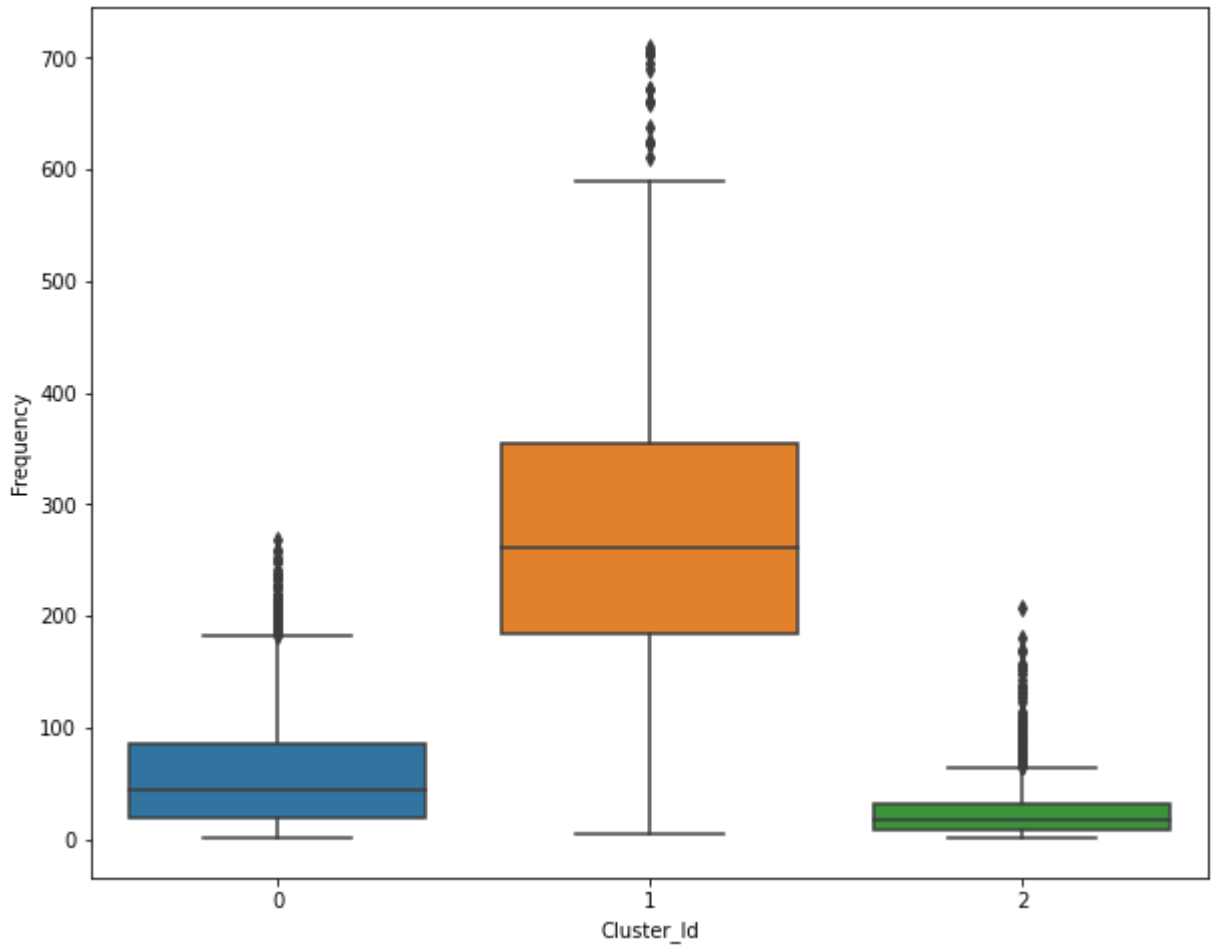
Download




```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff8b4eb9700>
```

```
<Figure size 720x576 with 1 Axes>
```

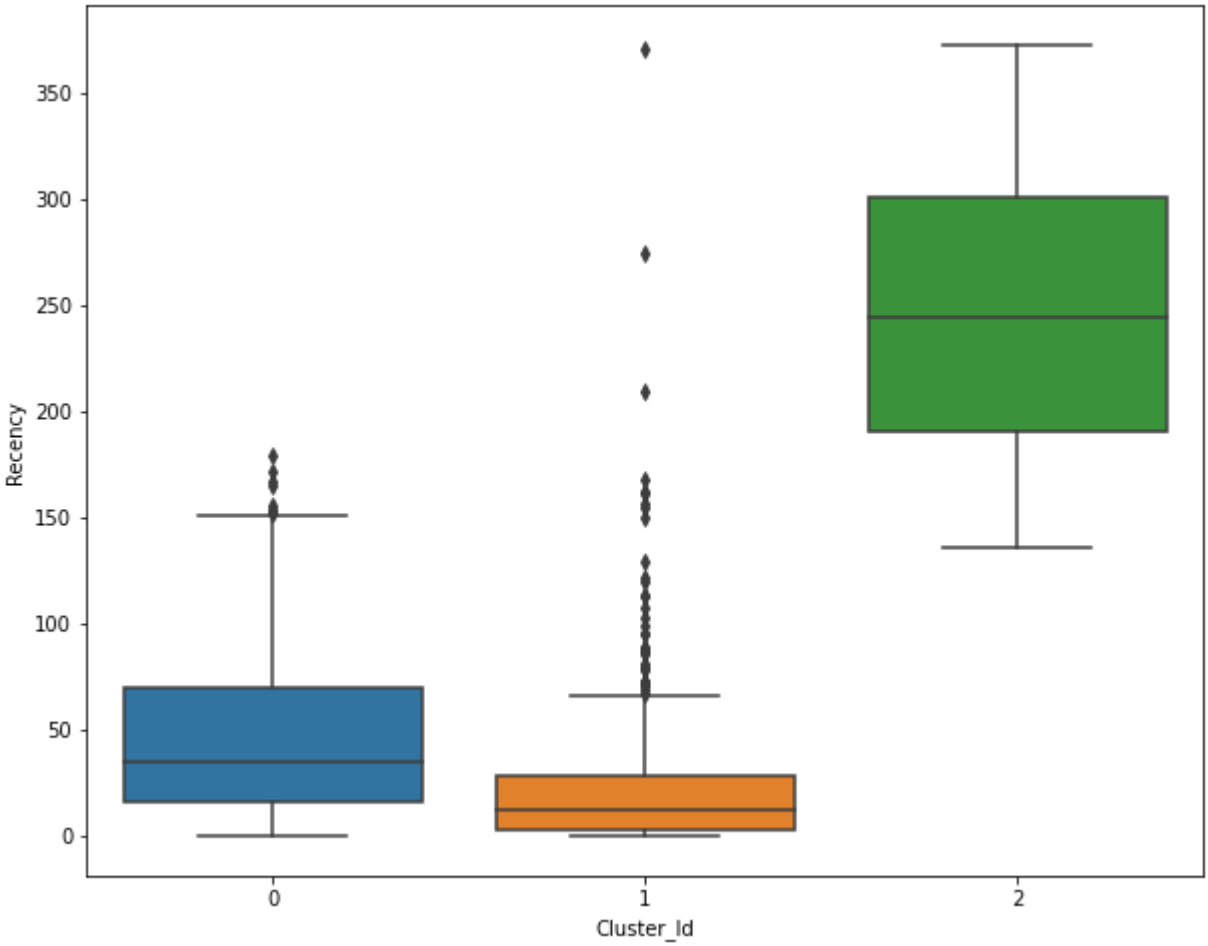
Download




```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff8b0c14d90>
```

```
<Figure size 720x576 with 1 Axes>
```

Download



###Evaluating model performance

We use the Calinski-Harabasz Score & Davies-Bouldin index to validate the silhouette score results from earlier and confirm statistical integrity of the model.

```
Calinski-Harabasz Score: 3631.9014311240626
```

```
Davies-Bouldin Score: 0.8502790798770803
```

From the high Calinski-Harabasz Score: 3631.9014311240626 and the low Davies-Bouldin Score: 0.8502790798770803 we find the kmeans clustering model application is accurate.

8. Conclusions

K-Means Clustering with 3 Clusters:

Customers with Cluster Id 1 are the customers with high amount of transactions as compared to other customers.

Customers with Cluster Id 1 are frequent buyers.

Customers with Cluster Id 2 are not recent buyers and hence least of importance from business point of view.

The majority of the customers are from the United Kingdom, and the country also has the highest sales in both 2010 and 2011. This suggests that the company should focus on expanding their market in the United Kingdom to increase their revenue.

The most popular products are those that are priced at a lower cost, which suggests that the company should focus on producing and promoting more affordable products to attract more customers.

The data also shows that there are some customers who frequently make purchases on the online retail store. These customers should be identified and targeted with personalized marketing campaigns and special offers to encourage them to continue making purchases.

From the box plot, it is clear that countries like France, Germany and Spain have more outliers in terms of total cost. This indicates that there are some customers in these countries who make high-value purchases. The company should focus on targeting these customers with high-end products and services.

From the heatmap, it is clear that Quantity and UnitPrice are highly correlated with each other. This suggests that the company should focus on increasing the quantity of products sold as this will increase the revenue.

The histograms show that most of the data is skewed to the left, indicating that the majority of the purchases are low-value transactions. The company should focus on increasing the average purchase value by promoting high-value products and services.

From the pairplot, we can observe that the distribution of TotalCost for different countries is different. This suggests that the company should focus on implementing country-specific strategies to increase sales.

The table of most purchased items for each country provides a clear picture of which items are popular in which countries. The company should focus on producing more of these popular items and promoting them in the relevant countries.