Capstone Project2

Table of Contents

Problem Objective

An online retail store is trying to understand the various customer purchase patterns for their

firm, this project is to give enough evidence-based insights to provide the same

Problem Statement 2: An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence-based insights to provide the same.

Dataset Information: The online_retail.csv contains 387961 rows and 8 columns.

Feature Name	Description
Invoice	Invoice number
StockCode	Product ID
Description	Product Description
Quantity	Quantity of the product
InvoiceDate	Date of the invoice
Price	Price of the product per unit
CustomerID	CustomerID
Country	Region of Purchase

Data description, various insights from the data.

From the dataset of the retail store has 387961 rows and 8 columns.

Features are as follows

1. Invoice number is unique identification number
2. Product ID of the purchase
3. Description of Product
4. Quantity of the product
5. Date of the invoice
6. Price of the product per unit
7. Customer ID
8. Region of Purchase

Data preprocessing Steps and Inspiration

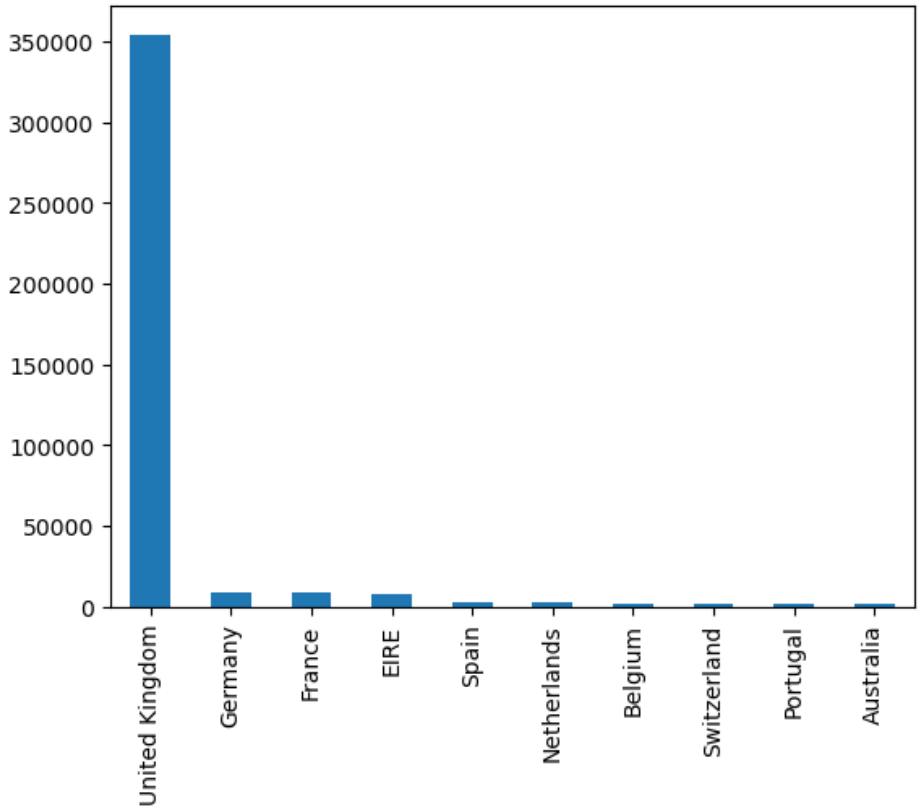
The preprocessing of the data included the following steps:

1. Step.1: Load data
2. Step.2: Perform Exploratory Data Analysis
    a. Confirm number of records in the data how they are distributed
    b. Check data types
    c. Check for missing data, invalid entries, duplicates
    d. Examine the correlation of the independent features.
    e. RFS analysis
3. Step.3: Model Predictions, two approaches:
    a. RFS model
    b. K-means clustering, confusion matrix

Model Evaluation and Technique

Top 5 countries sales count wise in the cleaned up data :

UK, Germany, France, Eire, Spain



RFM Analysis:

RFM (Recency, Frequency, Monetary) analysis is a customer segmentation technique that uses past purchase behaviour to divide customers into groups. RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services.

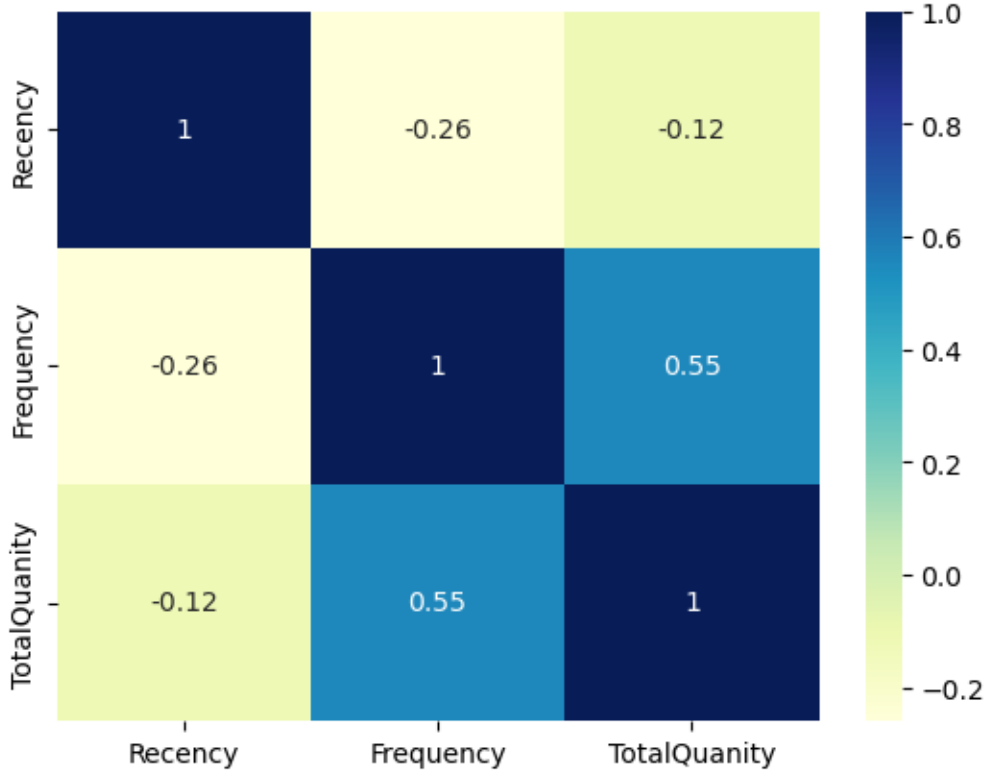
RECENCY (R): Days since last purchase

FREQUENCY (F): Total number of purchases

MONETARY VALUE (M): Total money this customer spent. We will create those 3 customer attributes for each customer.
Recency

RFM Table Visualisation

Now we will look at the correlation between the Recency, Frequency and Monetary part of the RFM table which will be an integral part of customer segmentation



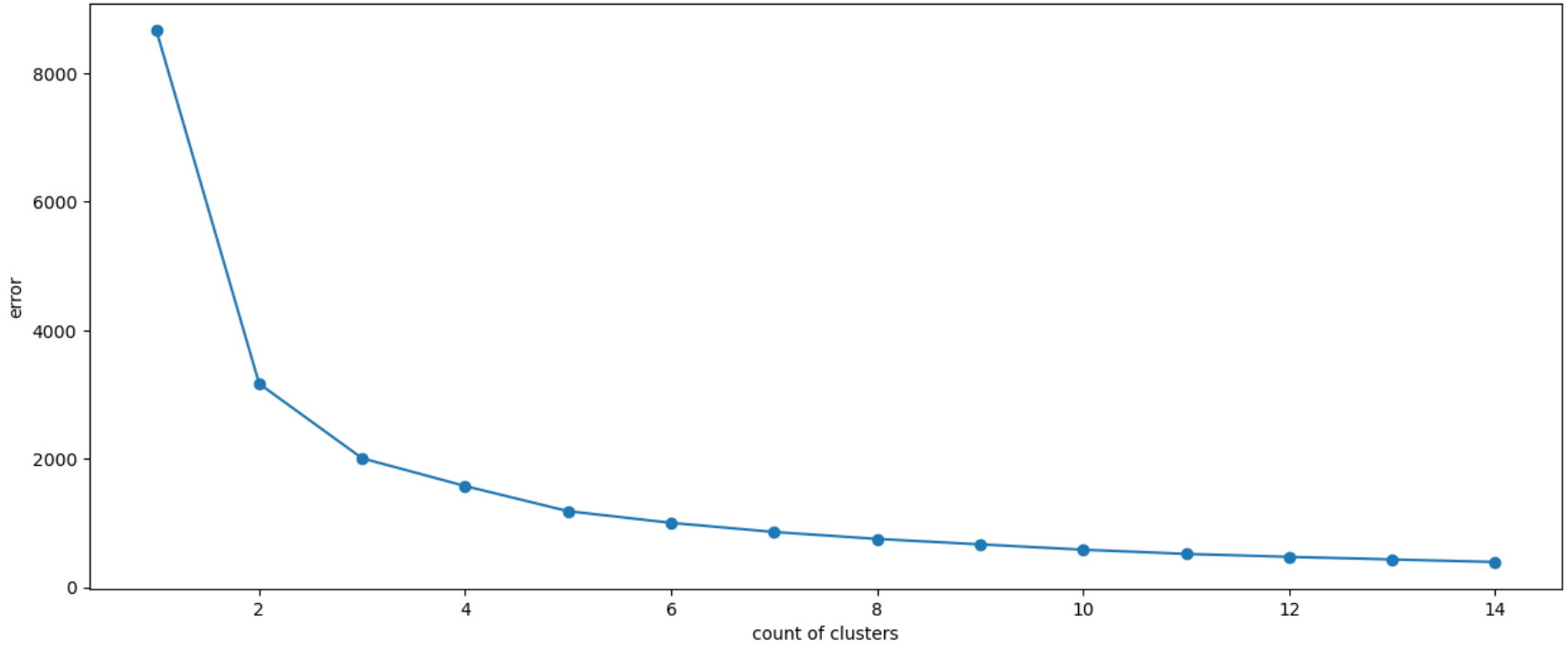
Conclusion To gain even further insight into customer behaviour, we can dig deeper in the relationship between RFM variables.

RFM model can be used in conjunction with certain predictive models like K-means clustering, Logistic Regression and Recommendation Engines to produce better informative results on customer behaviour.

We will go for K-means since it has been widely used for Market Segmentation and it offers the advantage of being simple to implement.
Applying PCA to reduce the dimensions and the correlation between Frequency and Monetary features.

Model Training

K-Means Clustering



Inferences:

We observe from the elbow plot a sharp bend after the number of clusters increase by 2. Silhouette Score is also the highest for 2 clusters.

But there is also a significant reduce in cluster error as number of clusters increase from 2 to 4 and after 4, the reduction is not much.

So, we will choose n_clusters = 4 to properly segment our customers.

Confusion matrix: KNN

```
Confusion matrix

 [[244   0   0   0   0   0]
 [  1 121   0   2   0   0]
 [  0   0 160   0   0   0]
 [  0   4   0 271   0   2]
 [  0   0   0   0 208   0]
```

```
 [  1   0   1   0   0 287]]

True Positives(TP) =  244

True Negatives(TN) =  121

False Positives(FP) =  0

False Negatives(FN) =  1
```

Confusion matrix: decisiontreeclassifier

```
confusion matrix
[[244   0   0   0   0   0]
 [  0 121   0   3   0   0]
 [  0   0 159   0   0   1]
 [  0   1   0 276   0   0]
 [  0   0   0   0 208   0]
 [  1   0   0   2   0 286]]

Accuracy of DecisionTreeClassifier: 99.3855606758832
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	244
1	0.99	0.98	0.98	124
2	1.00	0.99	1.00	160
3	0.98	1.00	0.99	277
4	1.00	1.00	1.00	208
5	1.00	0.99	0.99	289

```
    accuracy                           0.99      1302
   macro avg       0.99      0.99      0.99      1302
weighted avg       0.99      0.99      0.99      1302
```

Naive bayes

```
confussion matrix
[[242   0   0   0   0   2]
 [  0 123   0   1   0   0]
 [  0   0 141   0   4  15]
 [  0   1   0 274   0   2]
 [  0   0   0   0 208   0]
 [  0   0   0   4   0 285]]
```

Accuracy of Naive Bayes model: 97.7726574500768

	precision	recall	f1-score	support
0	1.00	0.99	1.00	244

1	0.99	0.99	0.99	124
2	1.00	0.88	0.94	160
3	0.98	0.99	0.99	277
4	0.98	1.00	0.99	208
5	0.94	0.99	0.96	289
accuracy			0.98	1302
macro avg	0.98	0.97	0.98	1302
weighted avg	0.98	0.98	0.98	1302

Conclusion

We saw that using classification models like Logisitc Regression, KNeighborsClassifier, DecisionTree we predicted the clusters for customers using RFM dataset as independent variables and Cluster as the target variable. The clusters predicted by the classification models perfectly aligns with K-Means clustering. So, we can conclude that our clusters are correct.

Summary:
The work described in this notebook is based on a database providing details on purchases made on an E-commerce platform over a period of one year. Each entry in the dataset describes the purchase of a product, by a particular customer and at a given date. In total, approximately ~ 4000 clients appear in the database. Given the available information, I decided to develop a classifier that allows to anticipate the type of purchase that a customer will make, as well as the number of visits that he will make during a year, and this from its first visit to the E-commerce site.

The next part of the analysis consisted of some basic data visualization. This was done in order to get insights regarding the country which was using the E-commerce website the most. I used basic plots in order to show the results of my analysis. I also tried to analyse other important factors such as the Gross Purchase by a country as well as which following description was

used the most. The final part of the analysis was the customer segmentation part.

The main way to go around with this process is to use the RFM (Recency, Frequency, Monetory) table to sort the customer in the groups. After creating the RFM table I used K-Means clustering (Elbow curve and Silhoutte scores) in order to create 4 clusters in which the customers should be Segmented. After each of the customers were segmented into their respective groups. I used models such as Logisitc Regression, KNeighborsClassifier ,DecisionTree in order the cross the accuracy of the clustering which resulted in an accuracy score 0.98. Hence, I conclude the customer segmentation was done which effective methods and high accuracy.