Capstone Project

Customer Segmentation

Introduction

- Businesses all over the world are growing every day. With the help of technology, they have access to a wider market and hence, a large customer base.

- Customer segmentation refers to categorizing customers into different groups with similar characteristics.

- Customer segmentation can help businesses focus on each customer group in a different way, in order to maximize benefits for customers as well as the business.

- This project mainly deals in segmenting customers of an online business store in the UK.

Data

- The data being used here is a transnational data of an online store based in the UK, which mainly sells unique all-occasion gifts.
- The data has 5,41,908 rows and 8 columns.
- The columns are as follows :-
- **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.
- **StockCode** : Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

- **Description** : Product (item) name. Nominal.
- **Quantity** : The quantities of each product (item) per transaction. Numeric.
- **InvoiceDate** : Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **UnitPrice** : Unit price. Numeric, Product price per unit in sterling.
- **CustomerID** : Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- **Country** : Country name. Nominal, the name of the country where each customer resides.
- Each row represents a different item purchased by the customer.

Steps in the project

- The project has been completed in 5 steps :-

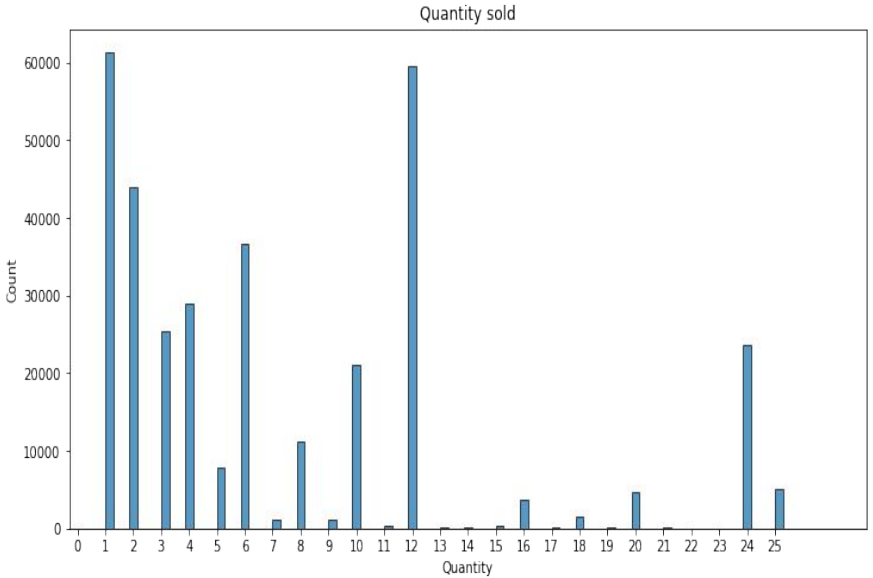
1. Data Cleaning
2. Exploratory Data Analysis (EDA)
3. Data Transformation
4. Clustering
5. Cluster Profiling

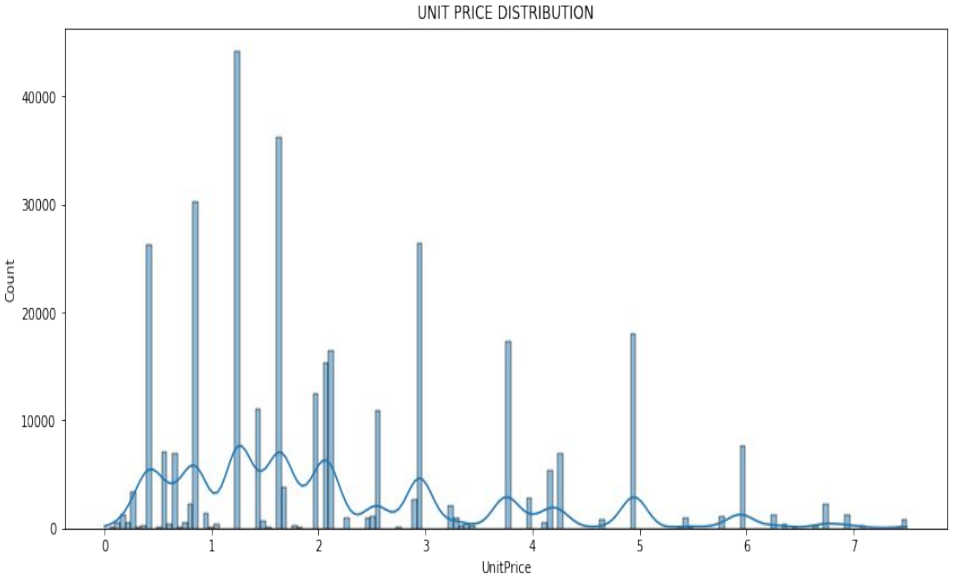
1. Data Cleaning

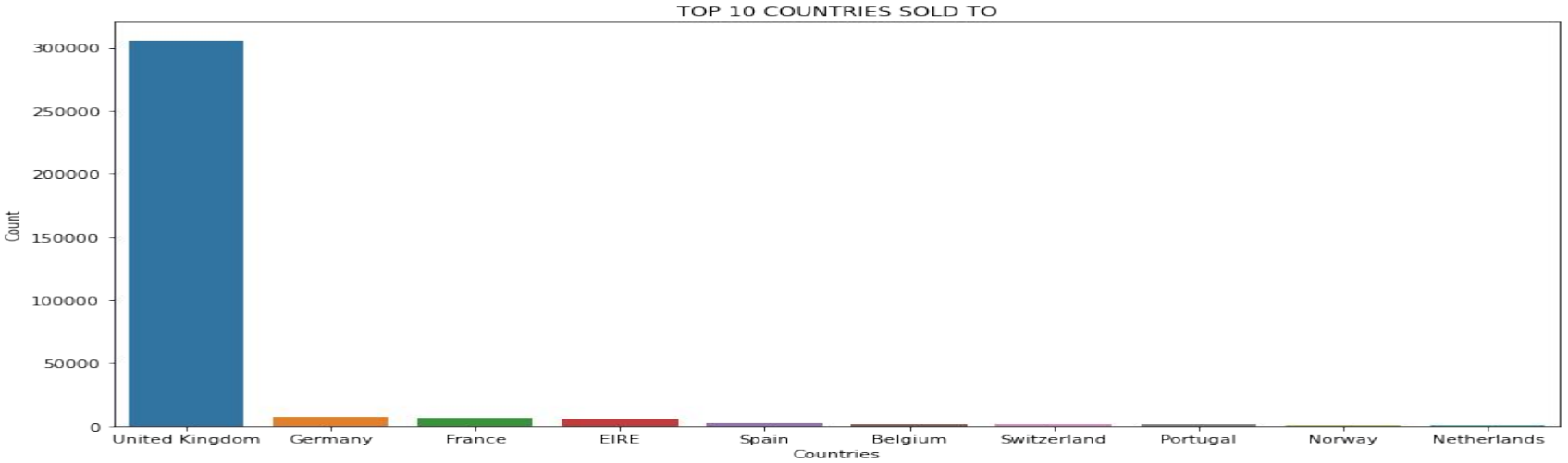
- After importing the data, the data must be cleaned.
- In this case, there are null values present in the ‘CustomerID’ and ‘Description’ column. These have to be dropped as there is no way of filling them strategically.
- Cancelled orders exist in the data, these too have been removed.
- Date, month and year were extracted from the ‘InvoiceDate’ column.
- Outliers in the ‘Quantity’ and ‘UnitPrice’ columns have been removed.
- ‘StockCode’ and ‘InvoiceDate’ column have been removed.

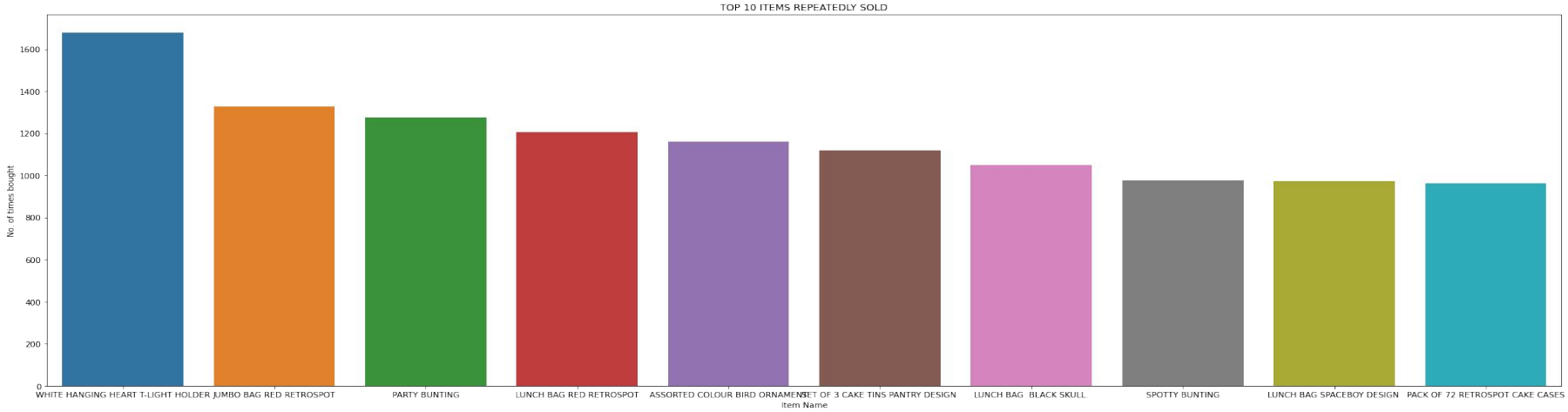
2. Exploratory Data Analysis (EDA)

- Univariate Analysis

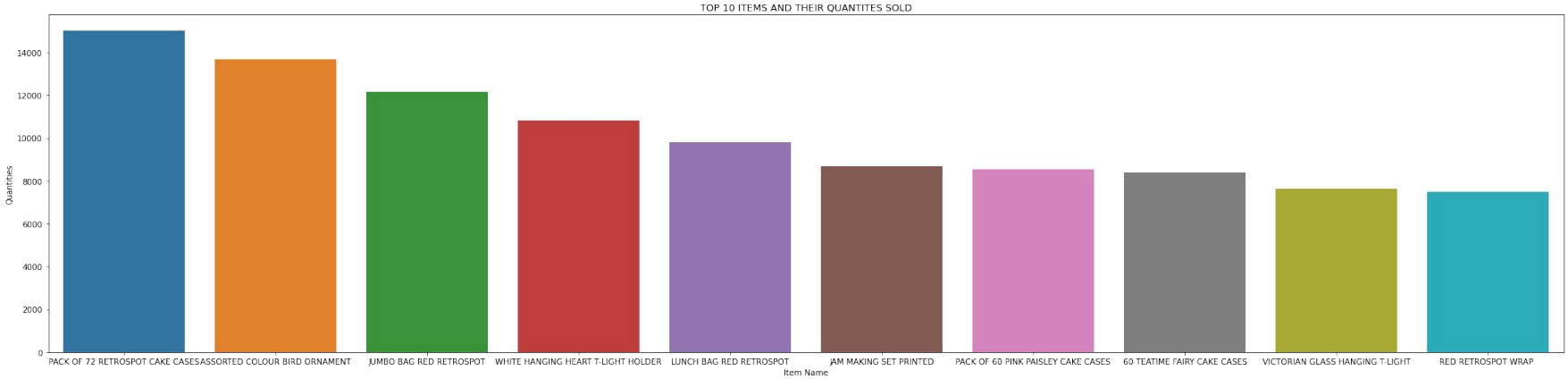


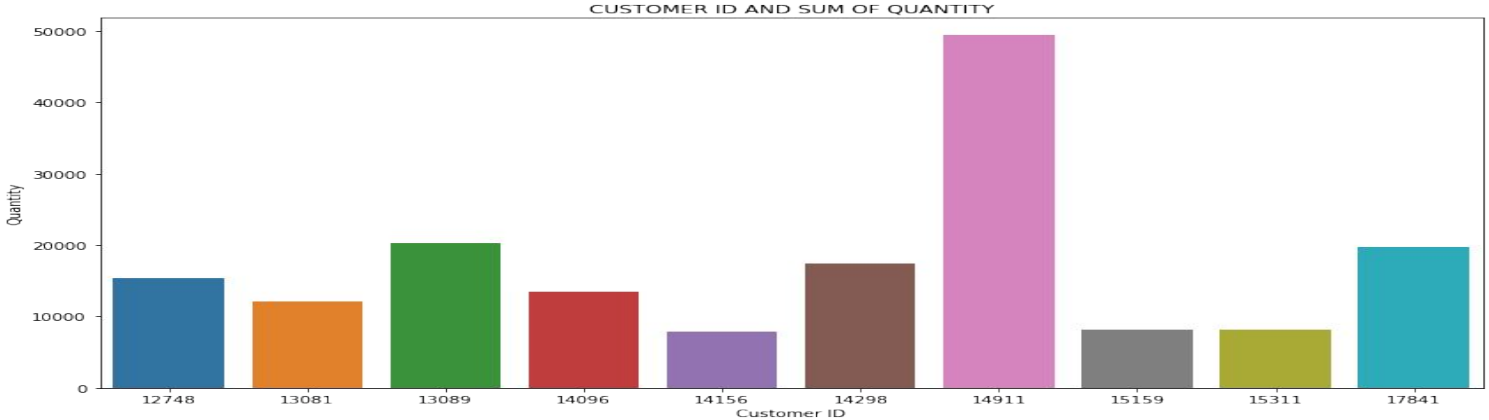



TOP 10 COUNTRIES SOLD TO
Count
300000
250000
200000
150000
100000
50000
0
United Kingdom
Germany
France
EIRE
Spain
Belgium
Switzerland
Portugal
Norway
Netherlands
Countries

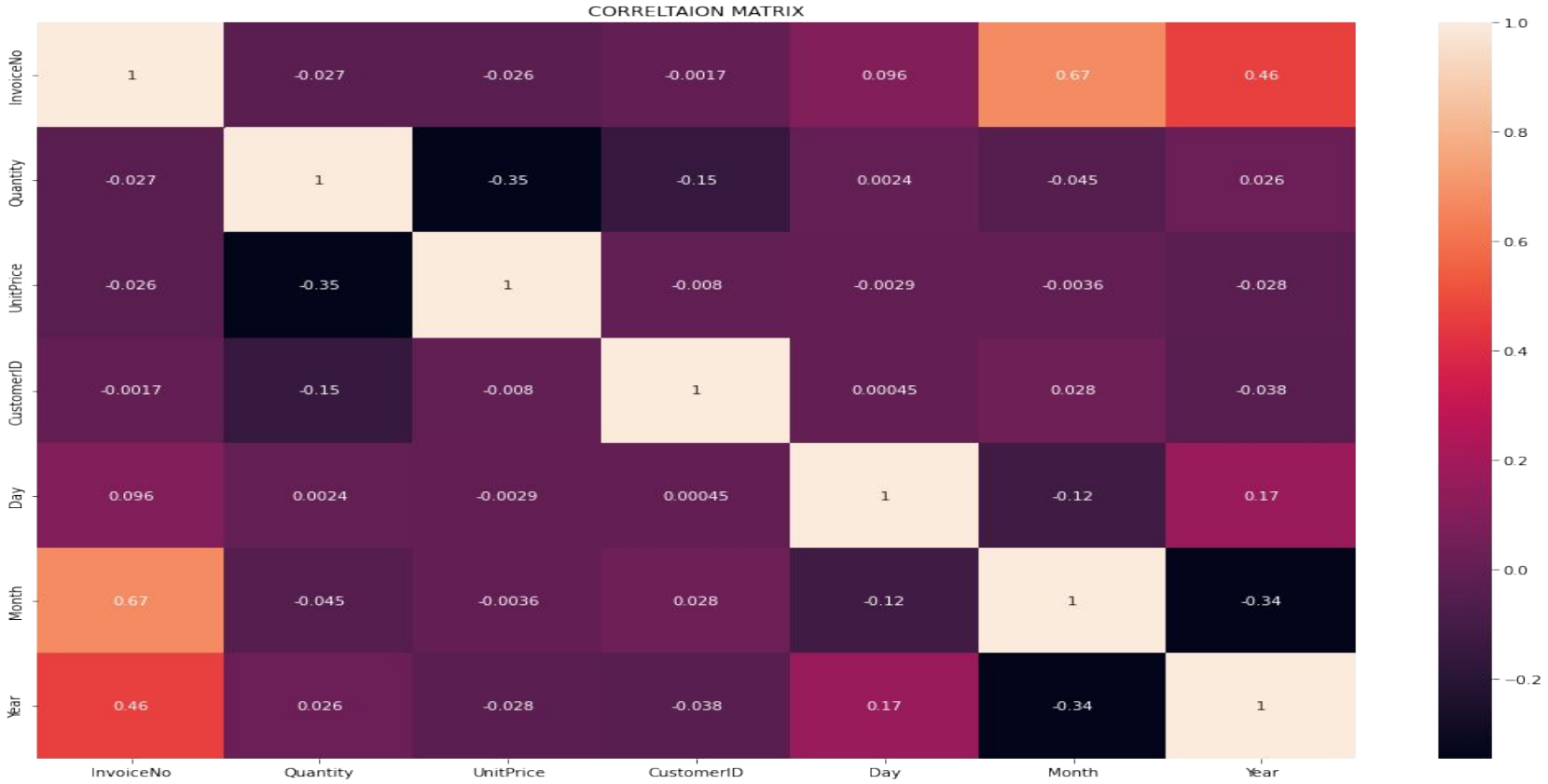

TOP 10 ITEMS REPEATEDLY SOLD
No. of times bought
1600
1400
1200
1000
800
600
400
200
0
WHITE HANGING HEART T-LIGHT HOLDER
JUMBO BAG RED RETROSPOT
PARTY BUNTING
LUNCH BAG RED RETROSPOT
ASSORTED COLOUR BIRD ORNAMENT
SET OF 3 CAKE TINS PANTRY DESIGN
LUNCH BAG BLACK SKULL
SPOTTY BUNTING
LUNCH BAG SPACEBOY DESIGN
PACK OF 72 RETROSPOT CAKE CASES
Item Name

- Bivariate Analysis





- Bivariate Analysis



Insights from EDA

- Quantities of products sold, mostly range from 1-12 units, also many purchases are in 24 units as well. This signifies units sold are in a dozen or two.
- Unit prices of products are mainly less than 3 pounds, there are some products with higher values as well.
- ‘White hanging heart T-light holder’ is most repeated order.
- The UK has the highest sales, this is logical as the company is UK based.
- Months of October, November and December repeat most frequently.
- Customer with the ‘ID 14911’ has purchased the most quantities.
- 'Pack of 72 retrospot cake cases' were sold the most in terms of quantity, around 15000 units.

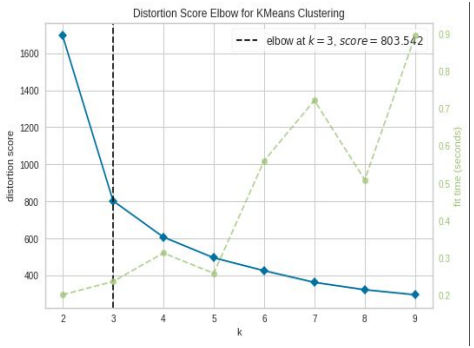
3. Data Transformation

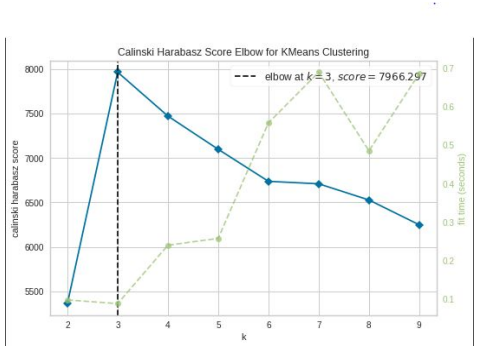
- In this section, a Recency, Frequency and Monetary (RFM) analysis about the data is done.
- Recency signifies the days since order, frequency signifies the number of times the customer is been billed and monetary signifies the sales each customer has provided.
- It can be seen that, frequency and monetary variables have a linear trend and frequency of orders have been high recently.
- The RFM dataframe is grouped on the basis of customer ID. The data now contains 4192 rows or customers.
- The ranges in the data differ, hence, the data is also scaled using a Standard Scaler.

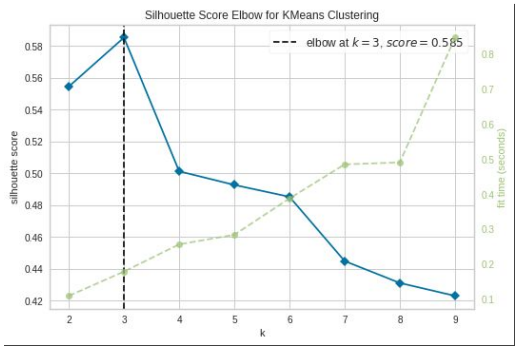
4. Clustering

- In this section, we use KMeans algorithm to cluster the customers into different segments.
- To identify the optimum number of clusters, we use the elbow method and silhouette analysis.
- With both the methods, 3 clusters is optimum in this case.
- A KMeans model with 3 clusters is developed and customers are segmented into different clusters.

Elbow Method & Silhouette Score







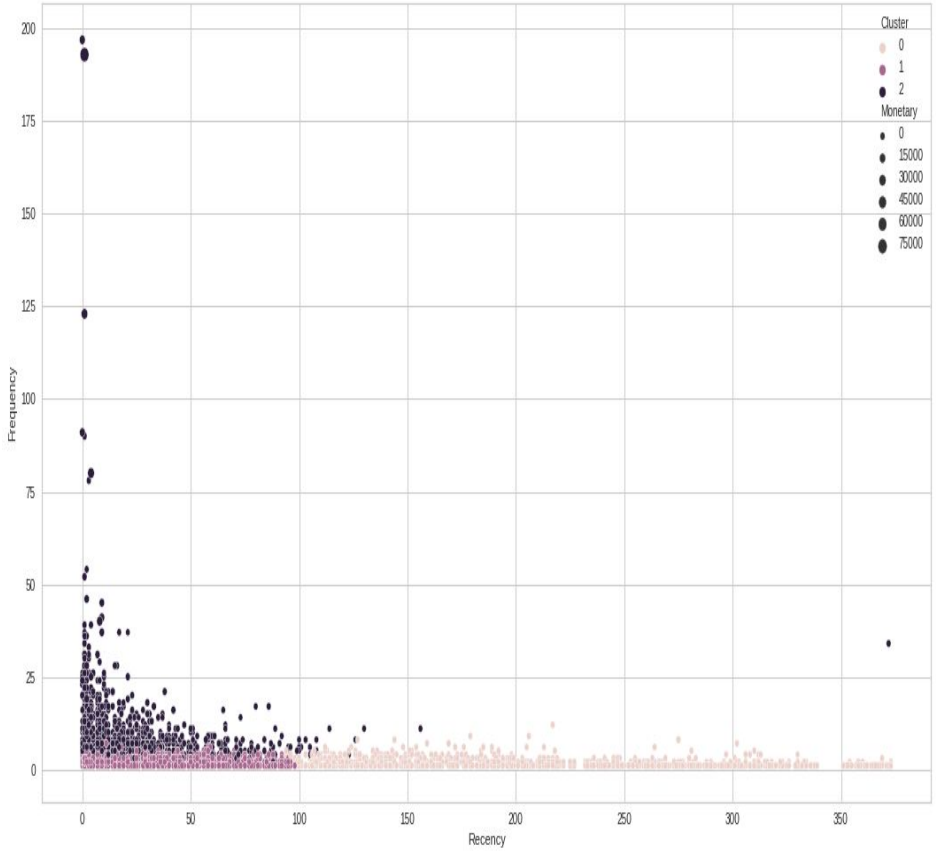
	Clusters	Silhouette Score
0	2.0	0.554560
1	3.0	0.585352
2	4.0	0.501268
3	5.0	0.492773
4	6.0	0.481417
5	7.0	0.442762
6	8.0	0.431277
7	9.0	0.416096
8	10.0	0.392698

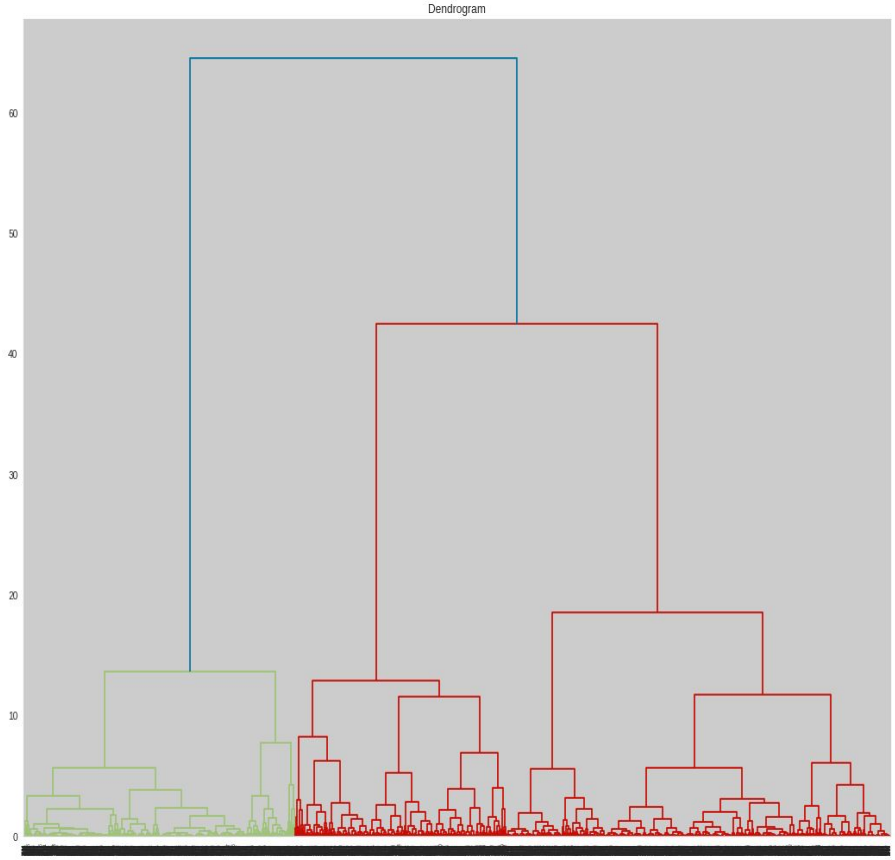
5. Cluster Profiling

Cluster	Recency	Frequency	Monetary
0	219.930303	1.675758	353.290455
1	38.098834	2.171016	455.651145
2	24.879552	10.002801	2830.897675

- After grouping the clusters, using the mean, each cluster can be named as follows :-

1. Cluster 2 - High value and loyal customers.
2. Cluster 1 - Average value and ordinary customers.
3. Cluster 0 - Low value and casual customers.


Cluster
0
1
2
Monetary
0
15000
30000
45000
60000
75000
Frequency
Recency
0
25
50
75
100
125
150
175
200
0
50
100
150
200
250
300
350


Dendrogram
0
10
20
30
40
50
60

Conclusion

- This project mainly focused on developing customer segments for a UK based online store, selling unique all occasion gifts.
- This project is worked through 5 major steps, starting from data cleaning till cluster profiling.
- Using a recency, frequency and monetary (RFM) analysis, the customers have been segmented into various clusters.
- The business can focus on these different clusters and provide to customers of each sector in a different way, which would not only benefit the customer but also the business at large.

THANK YOU!