Online Retail Customer Segmentation

T. Sai Harish Sarma

Contents

Problem Statement

- In this Project, Our task is to identify major segments on a transnational data set which contains the transactions occurring between 01/12/2010 and 09/12/2011 for a UK based and registered non-store online retail. The company mainly sells unique all- occasion gifts. Many customers of the company are wholesalers.

Business Context

- Businesses are growing rapidly and serving many customers. So, It is very important to categorize their customers to understand the customer and Business behavior. It also helps in marketing and Business development.

Data Description

- Invoice No: Invoice Number (Some Invoice No's are with letter 'C', means cancelled Transaction)(Numeric)
- Stock Code: Stock Name Code
- Description: Description of the product (Numeric)
- Quantity: Quantity bought (Numeric)
- Invoice Date: Invoice Date (Date Time)
- Unit Price: Price per Unit (Numeric)
- Customer ID: Unique Customer ID (Numeric)
- Country: Location

Data Description

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
sns.set()

import warnings
warnings.filterwarnings('ignore')
import datetime as dt

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from yellowbrick.cluster import SilhouetteVisualizer
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import DBSCAN

from sklearn import metrics
```

```
# Loading Data Set
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
filepath= ('/content/drive/MyDrive/Colab Notebooks/Clustering Unsupervised ML Project/Online Retail.xlsx')
data=pd.read_excel(filepath)
```

```
#Size of the dataet
data.shape
```

```
(541909, 8)
```

Data Description

```
# Details of the Dataset
data.info()
```

```
<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
```

```
# Checking for Null Values
data.isnull().sum()
```

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

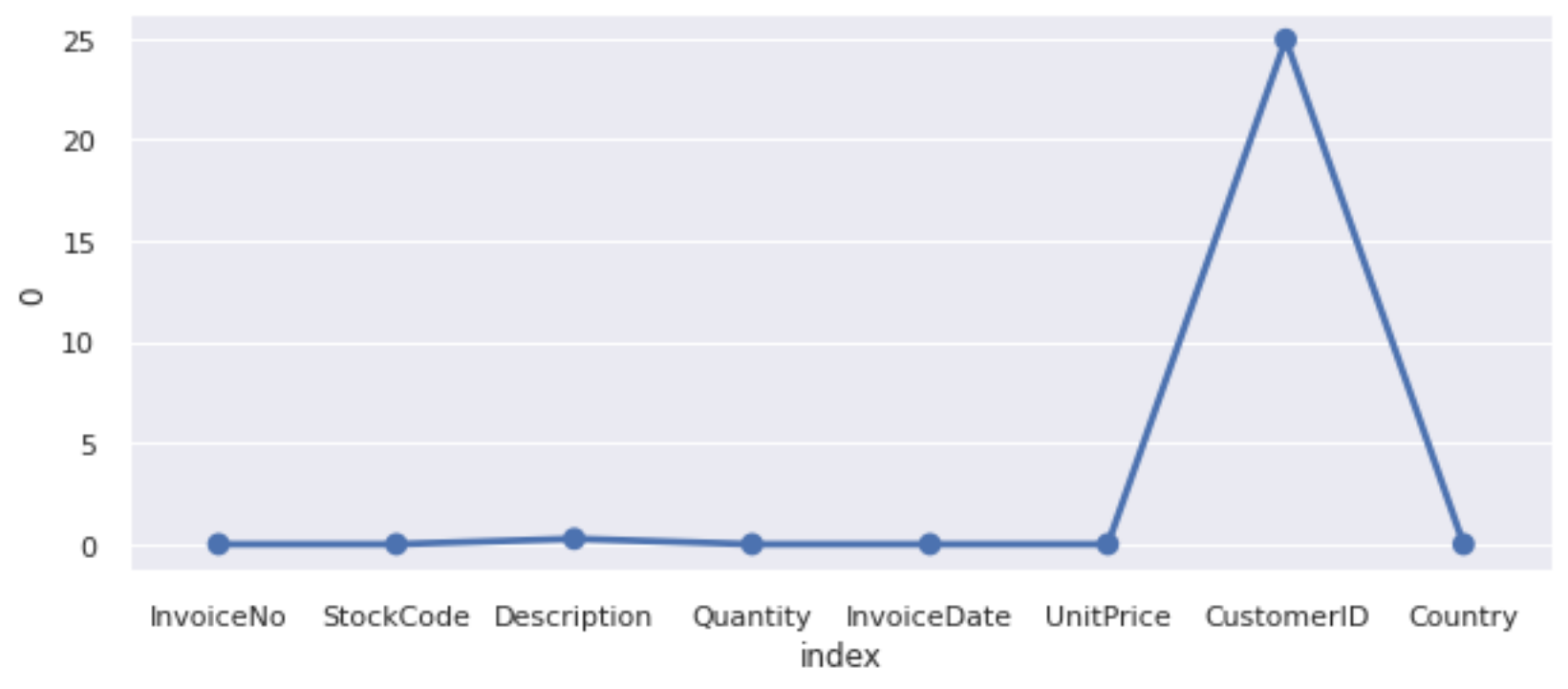
Data Cleaning

Handling Missing Values:

Customer ID is having 24.92% of missing values.
Description is having 0.26% of missing values.

```
# Checking for Null Values after removal of Nulls
data.isnull().sum()
```

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



Data Cleaning

- We have 5225 duplicate/ repeated entries.
- We can drop them from the dataset.
- We have few cancelled Invoice Nos indicated with C. We can remove them.
- There are 8872 cancelled Invoices.

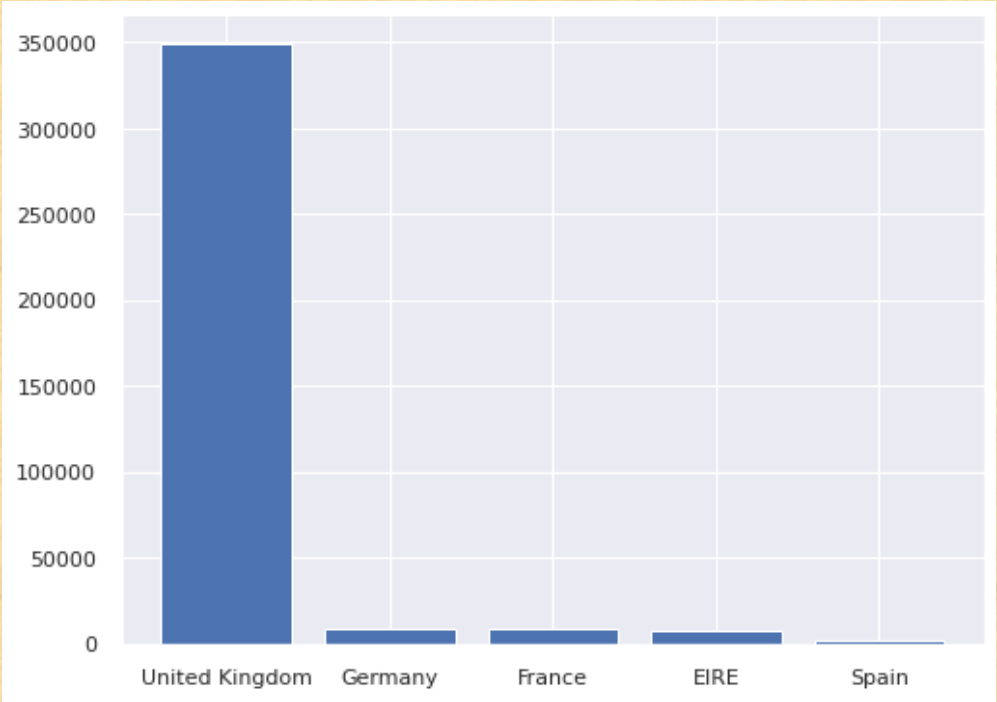
Duplication Check:

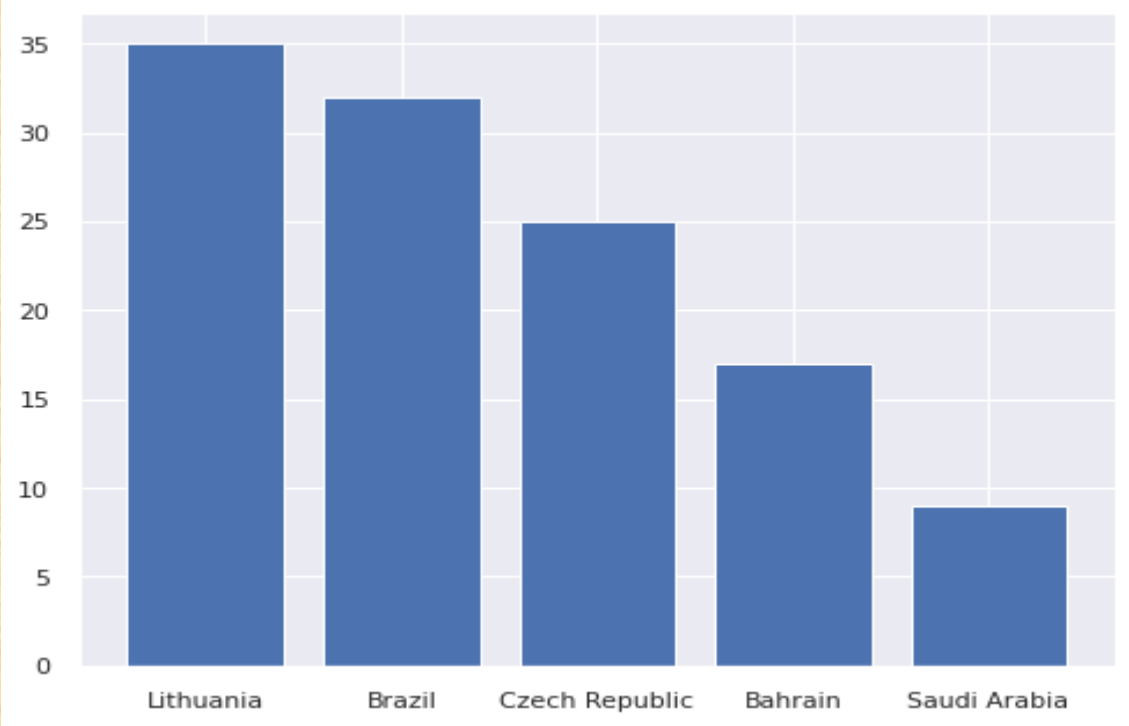
```
[1314] #Using Duplicated() Method
       len(data[data.duplicated()])

       5225
```

Data Analysis

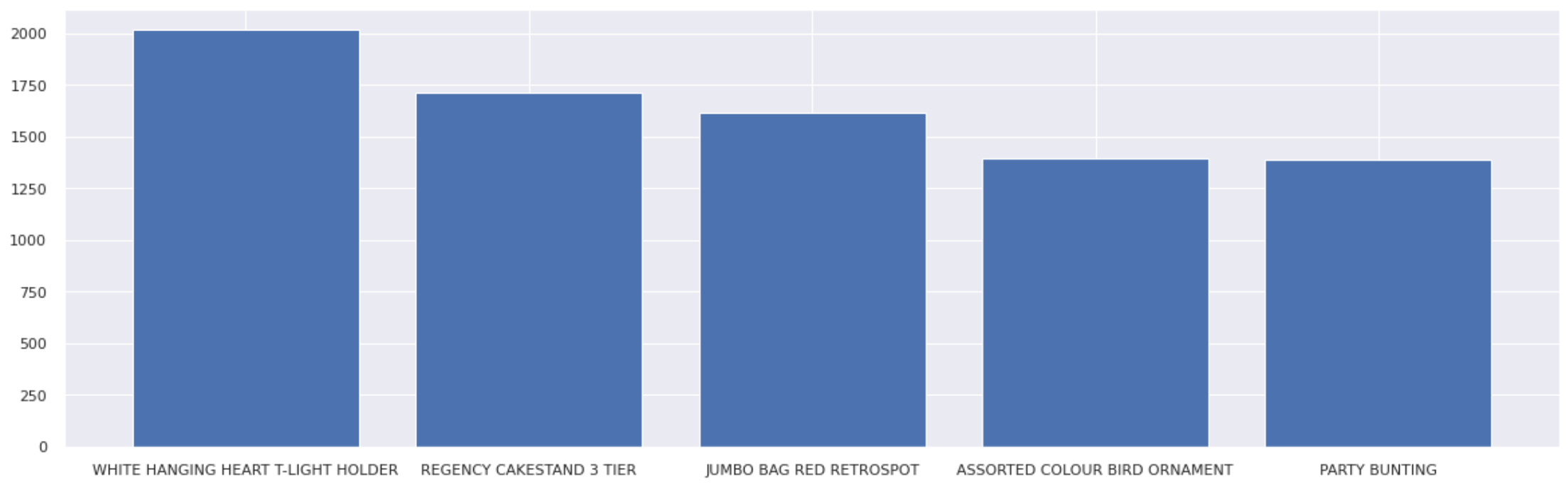
Country:





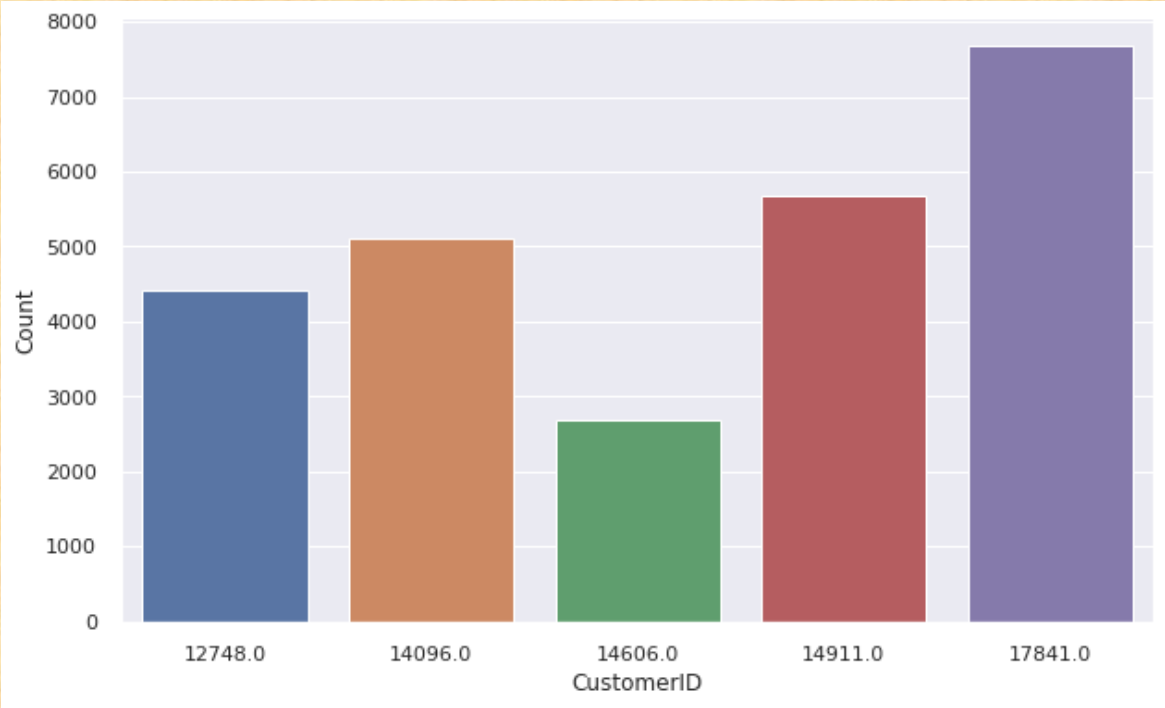
Data Analysis

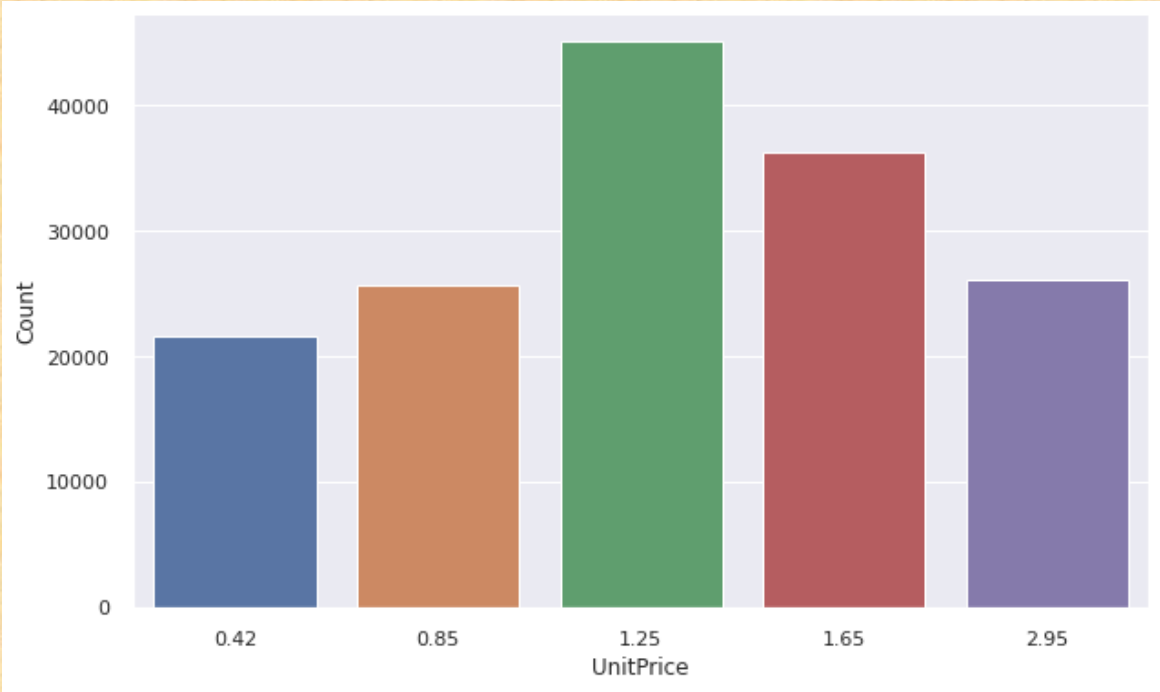
Description:



There are 3877 Unique Descriptions available in Dataset.

Data Analysis

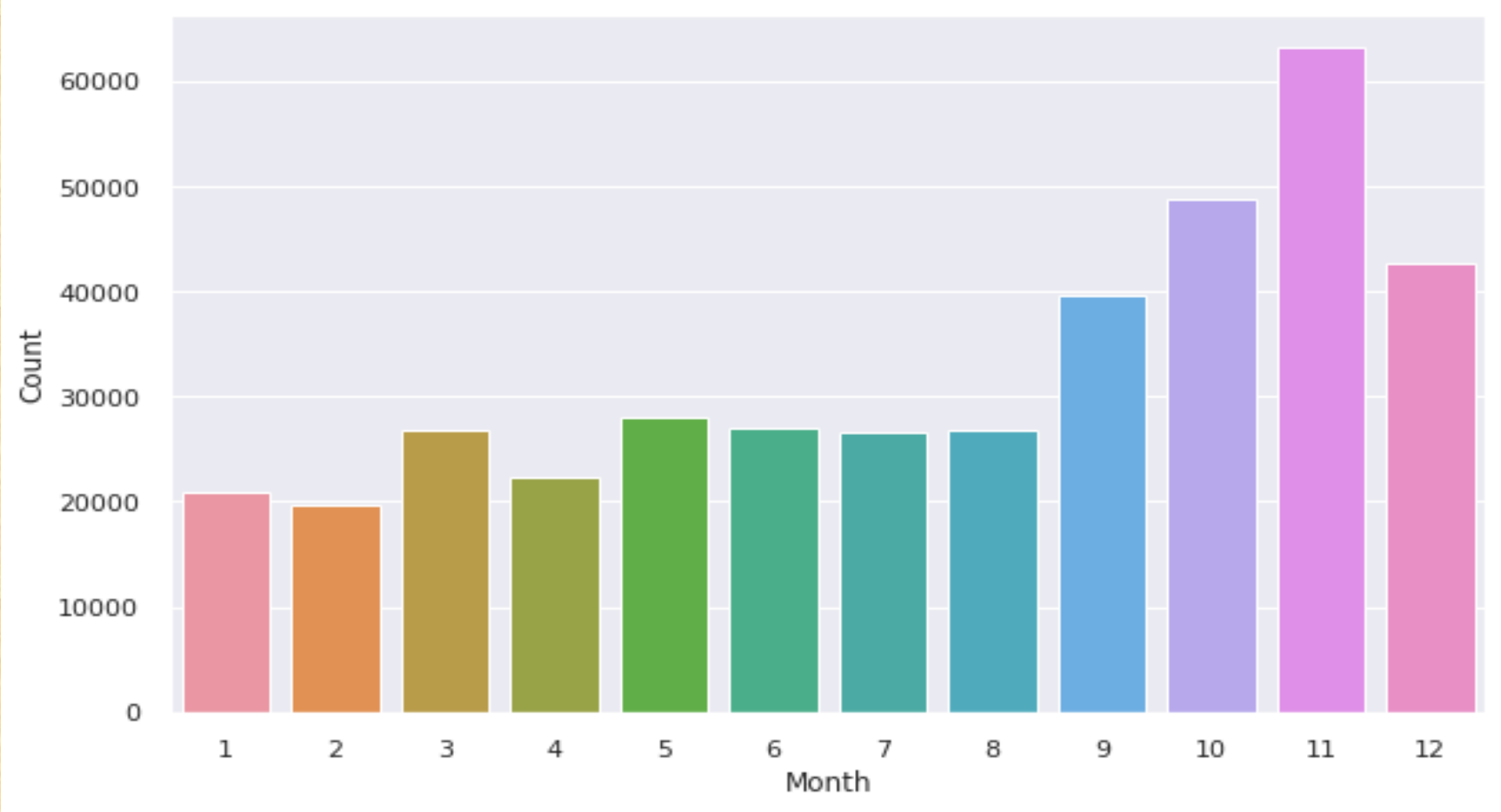




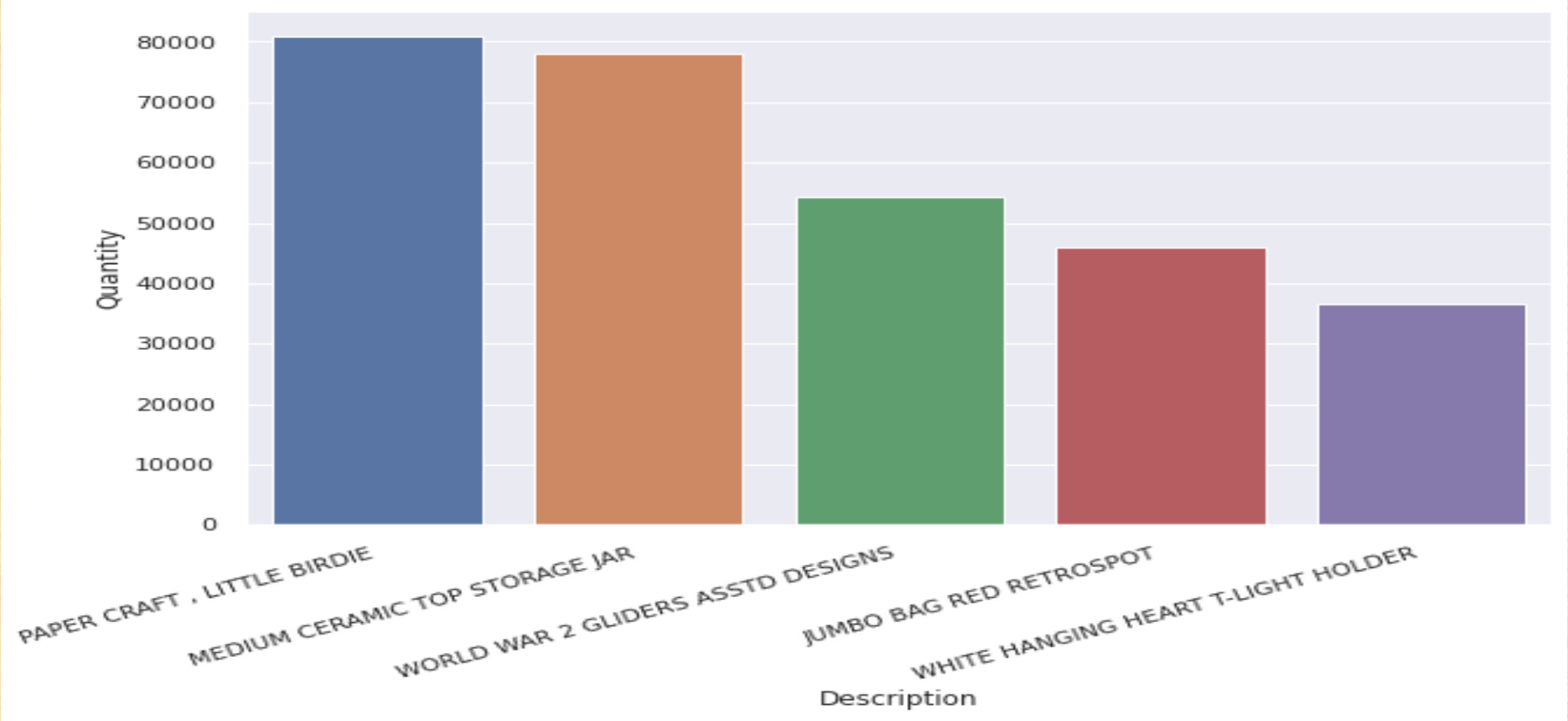
Customer ID : 17841 is top most customer by having large count in No. of Purchase.

The Products having price of 1.25 dollars are the selling products.

Data Analysis



Data Analysis



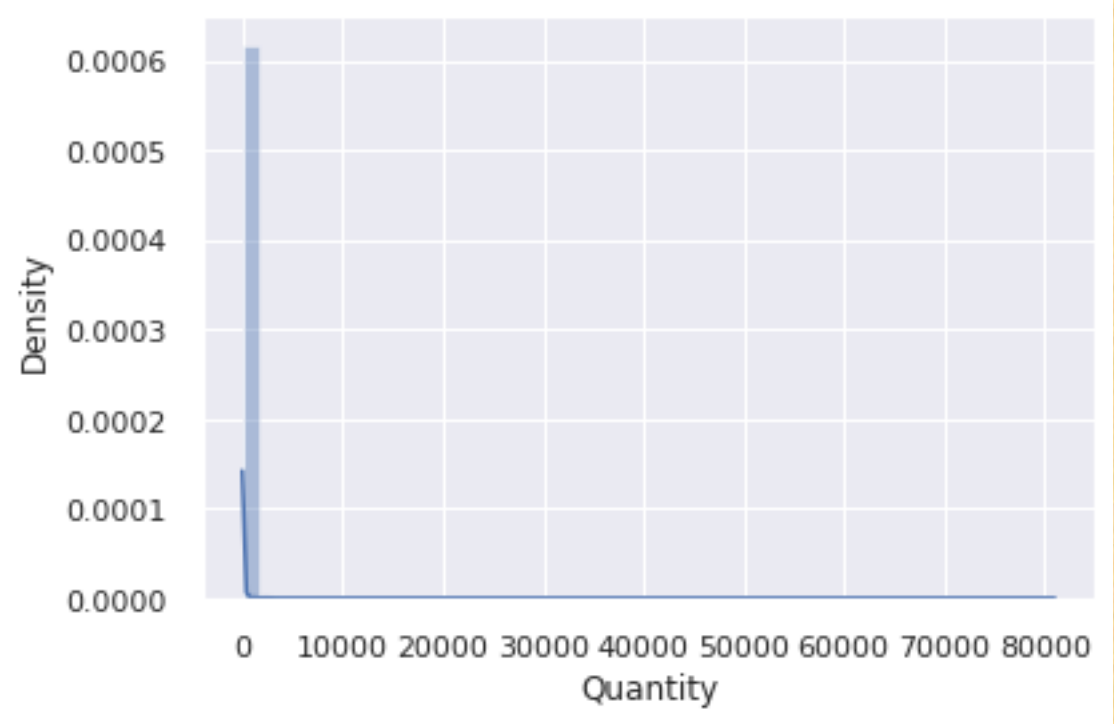
The product having description as "PAPER CRAFT , LITTLE BIRDIE" is the most selling product in store.

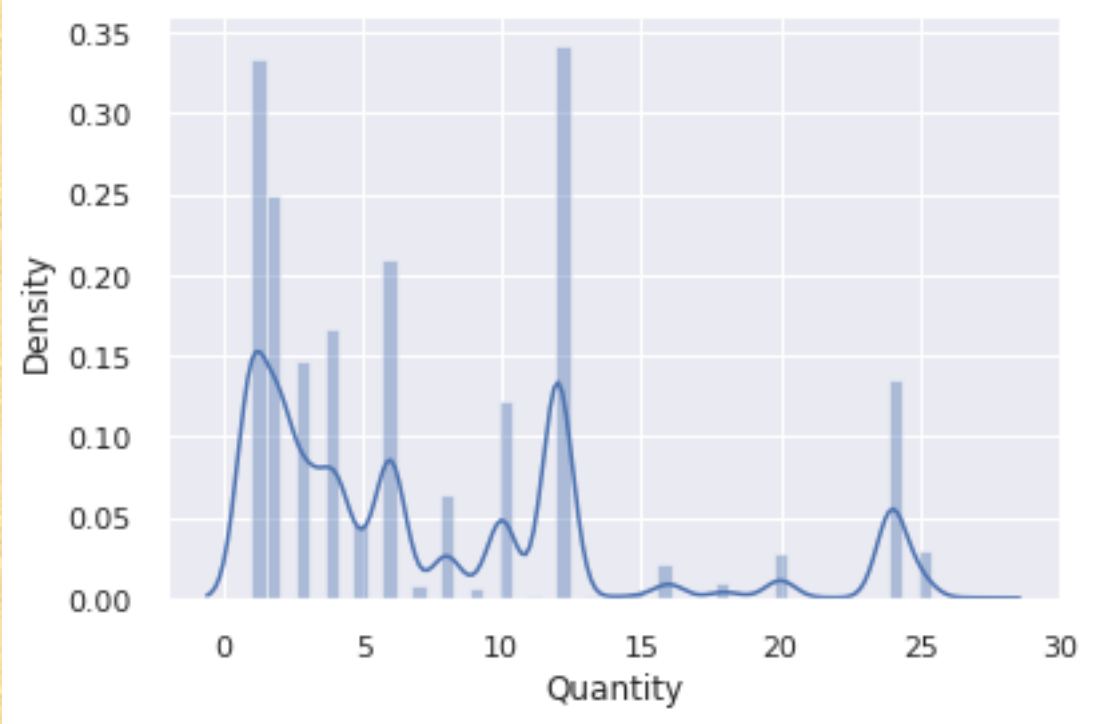
Feature Engineering

```
# Creating a function for outlier removal
def outliers_removal(data, column):
  Q1=data[column].quantile(0.25)
  Q3=data[column].quantile(0.75)
  IQR=Q3-Q1
  Ulimit= Q3+(IQR*1.5)
  Llimit= Q1-(IQR*1.5)
  if Llimit <0:
    data=data[data[column]<=Ulimit]
  else:
    data=data[(data[column]>=Llimit) & (data[column]<=Ulimit)]
  return data
```

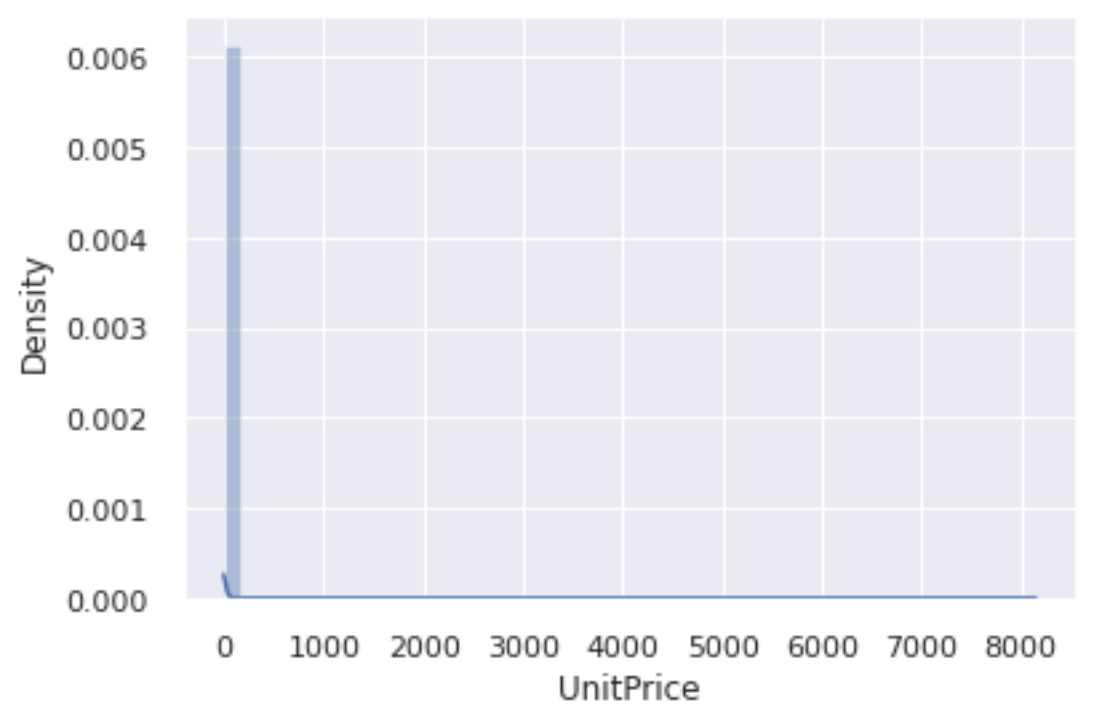
```
#Applying Outlier Function on Columns
data=outliers_removal(data=data, column='Quantity')
data=outliers_removal(data=data, column='UnitPrice')
```

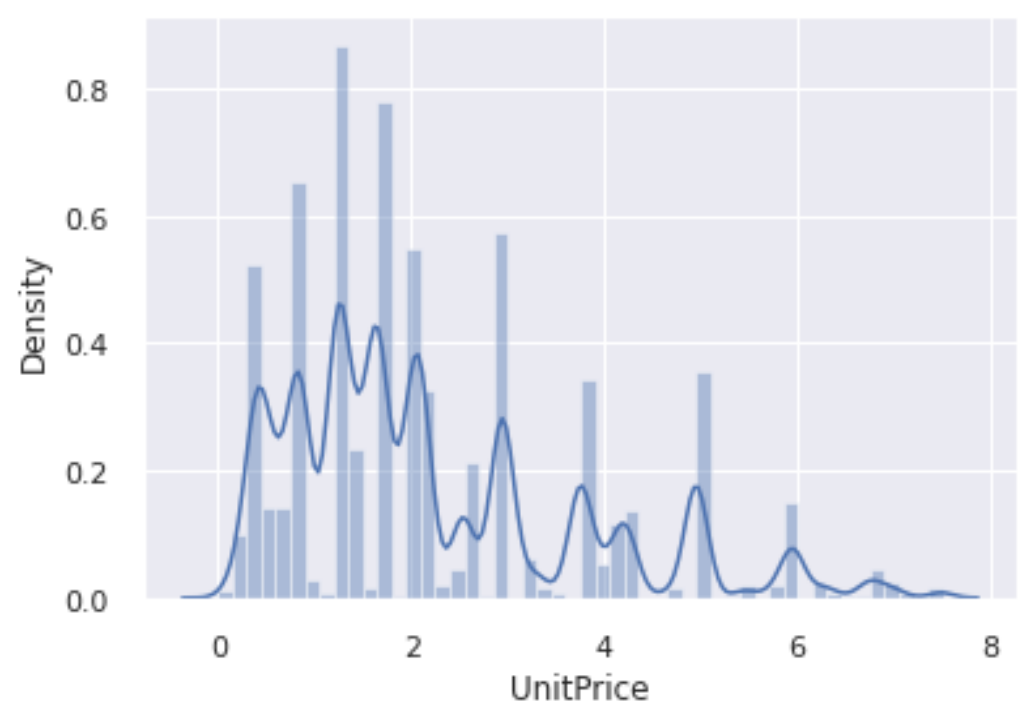
Feature Engineering



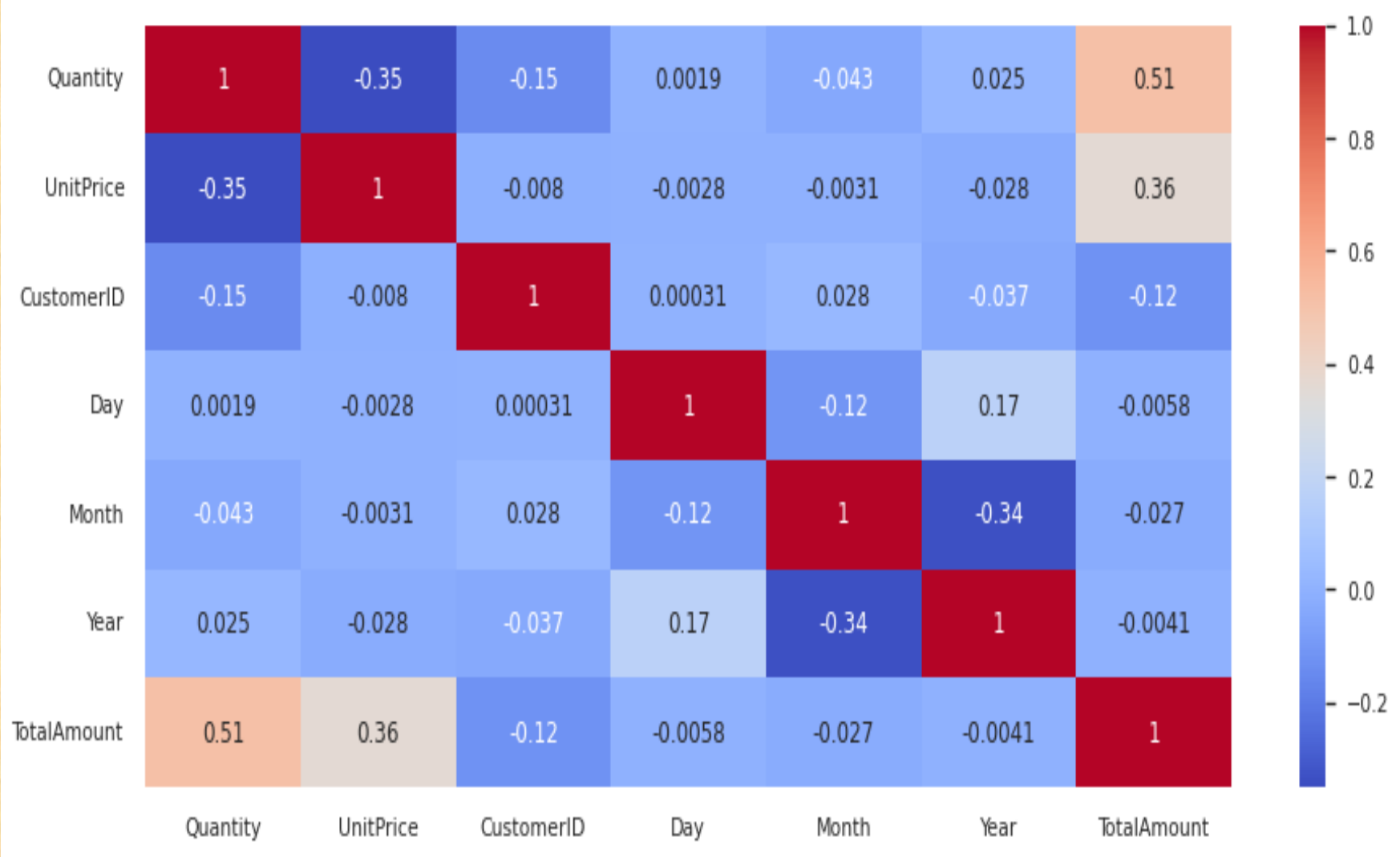


Feature Engineering





Feature Engineering



RFM Analysis

- RFM - Recency, Frequency and Monetory is a Marketing Analysis tool used for customer segmentation.
- Recency: How recently user bought/ visited.
- Frequency: How regularly user purchase/ visits.
- Monetory: How much revenue generated by that user.
- A the part of our project, we make a Data frame by extracting above features and use them for Clustering.

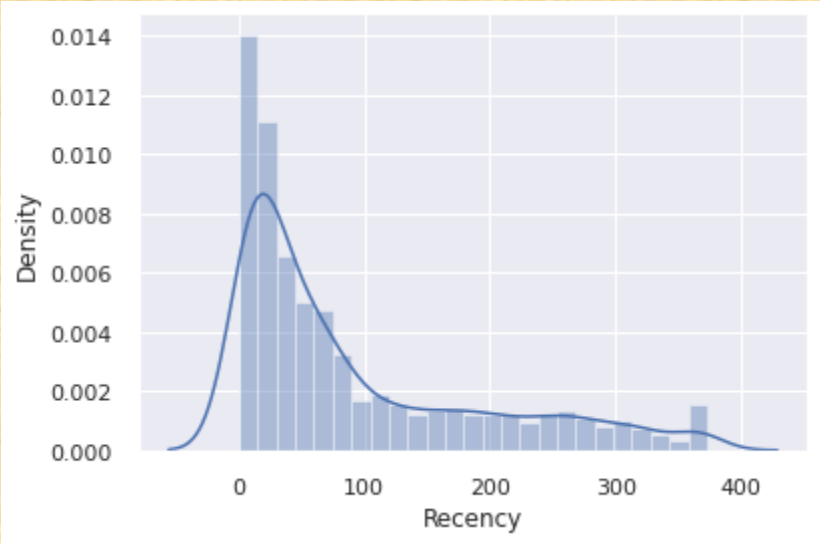
RFM Analysis

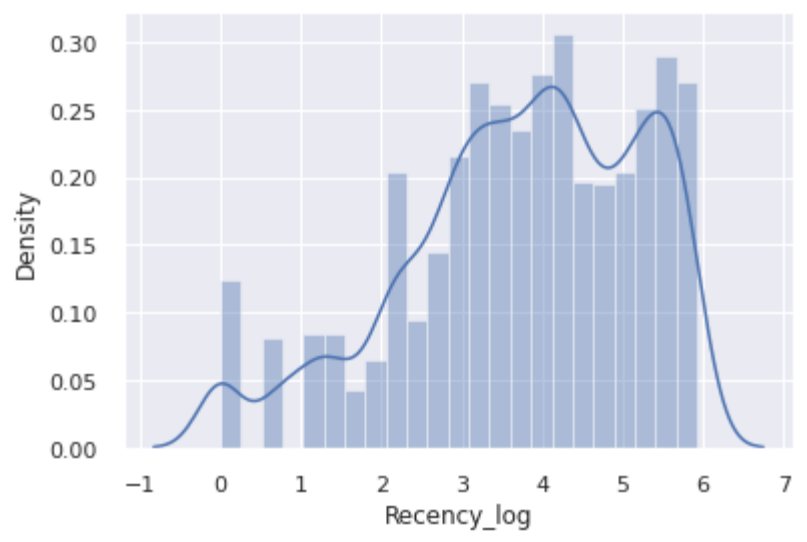
```
# Information of Dataframe
RFMDF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4339 entries, 0 to 4338
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   CustomerID  4339 non-null   float64
 1   Recency     4339 non-null   int64
 2   Frequency   4339 non-null   int64
 3   Monetory    4339 non-null   float64
dtypes: float64(2), int64(2)
memory usage: 135.7 KB
```

RFM Analysis

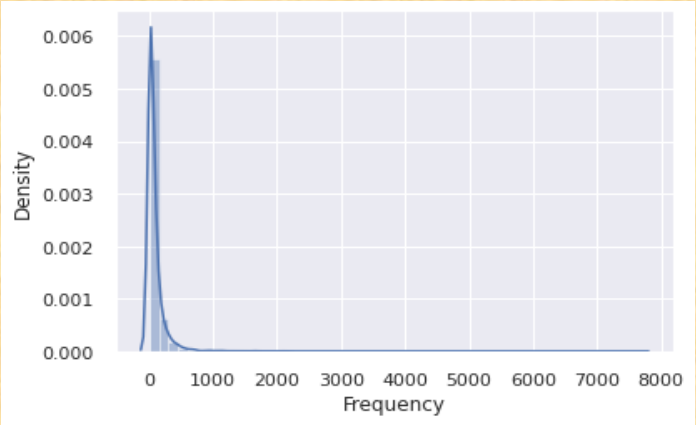
Log transformation:

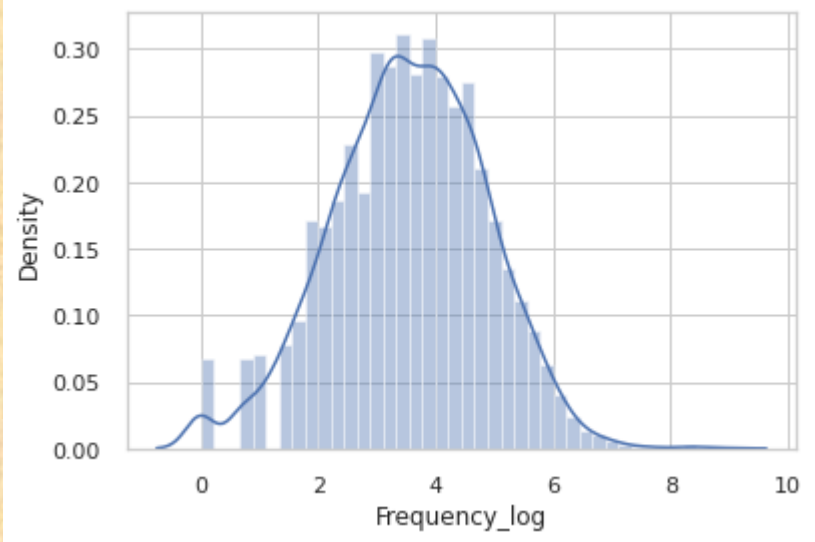


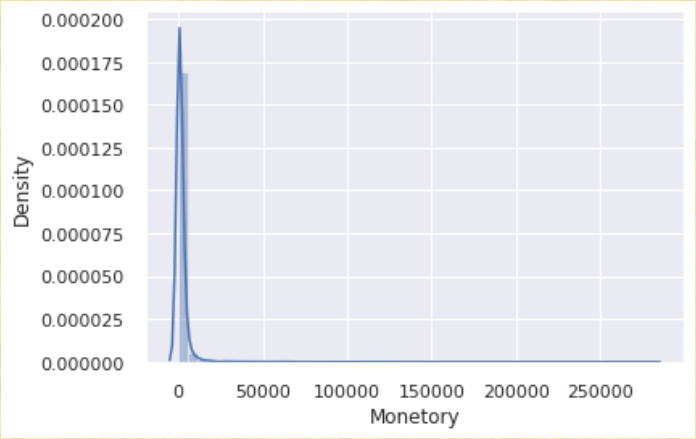


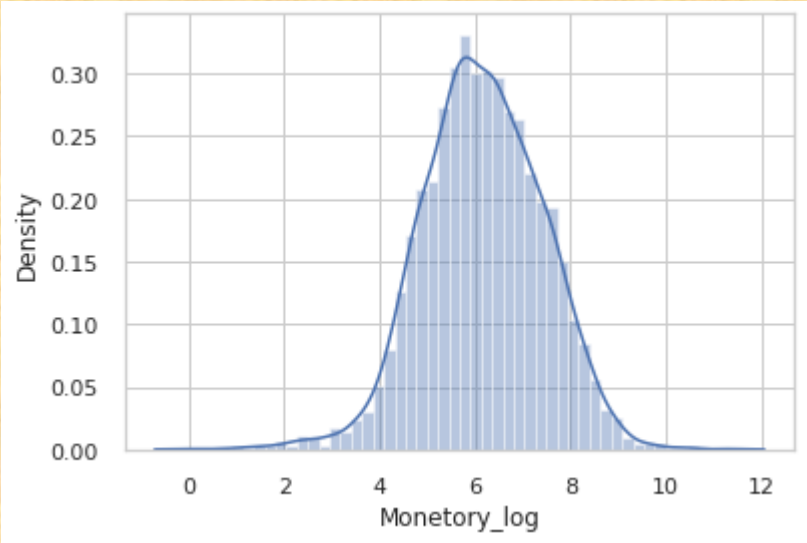
RFM Analysis

- Log transformation:





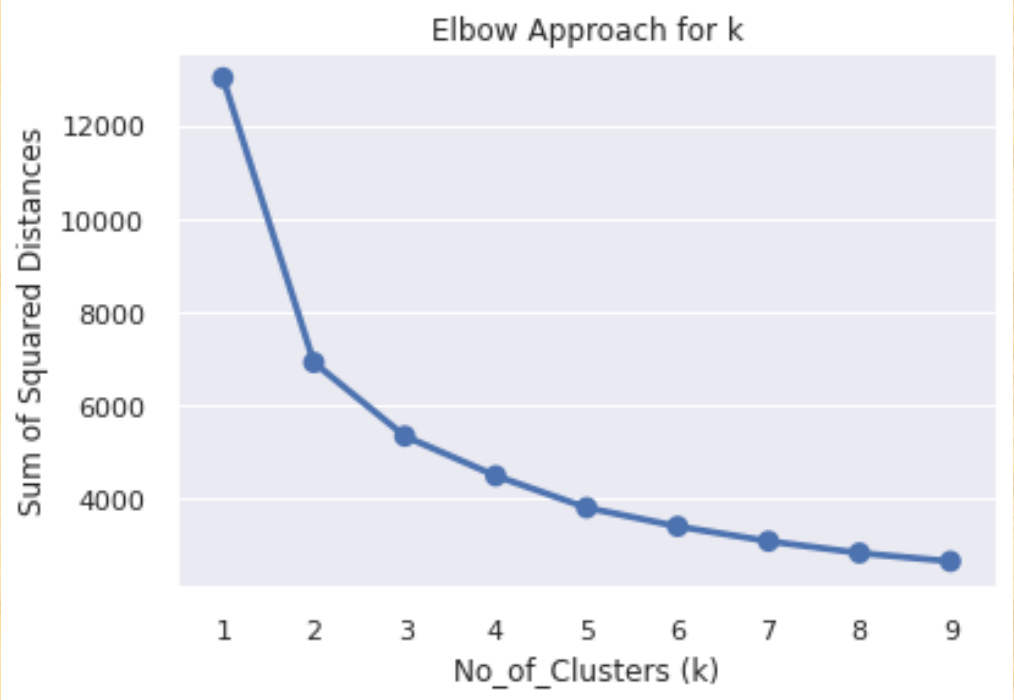




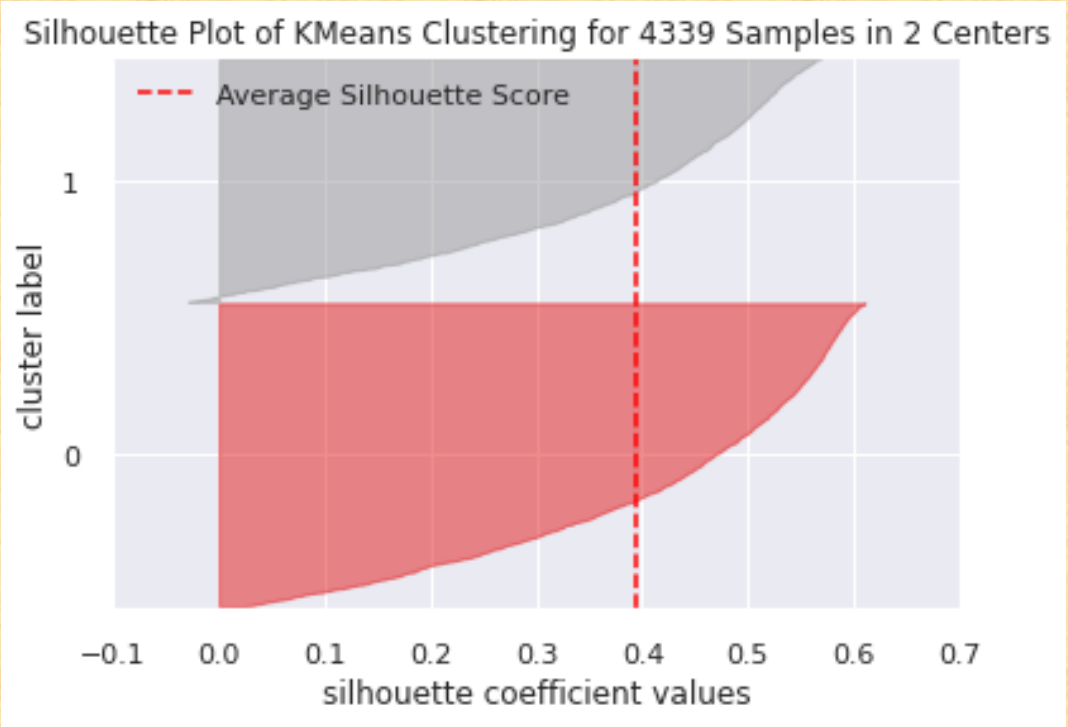
Preprocessing

```
# Using Standard Scaler () Method to stabdardize the data.
scaler=StandardScaler()
```

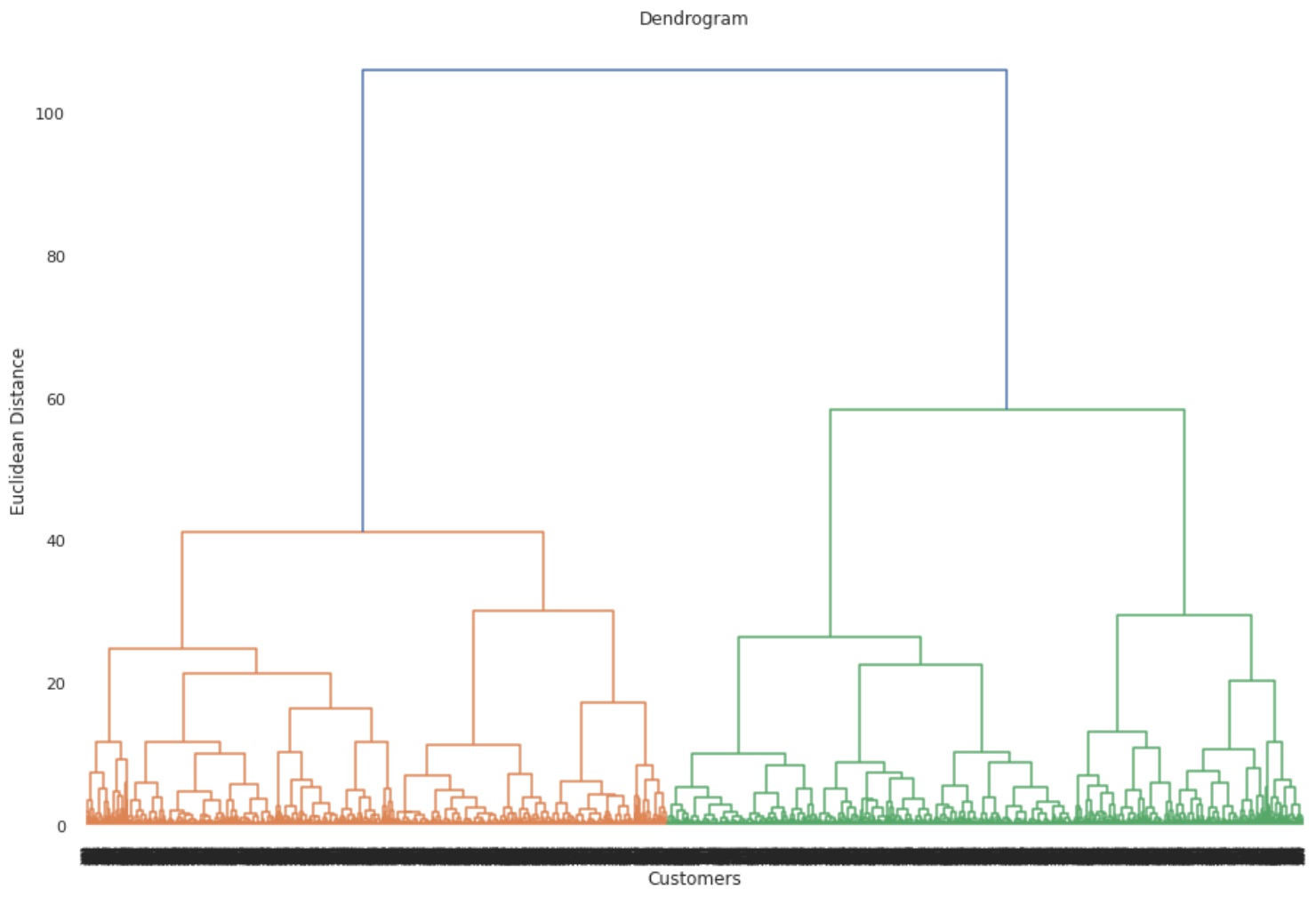
Modelling – KMeans & Elbow Curve



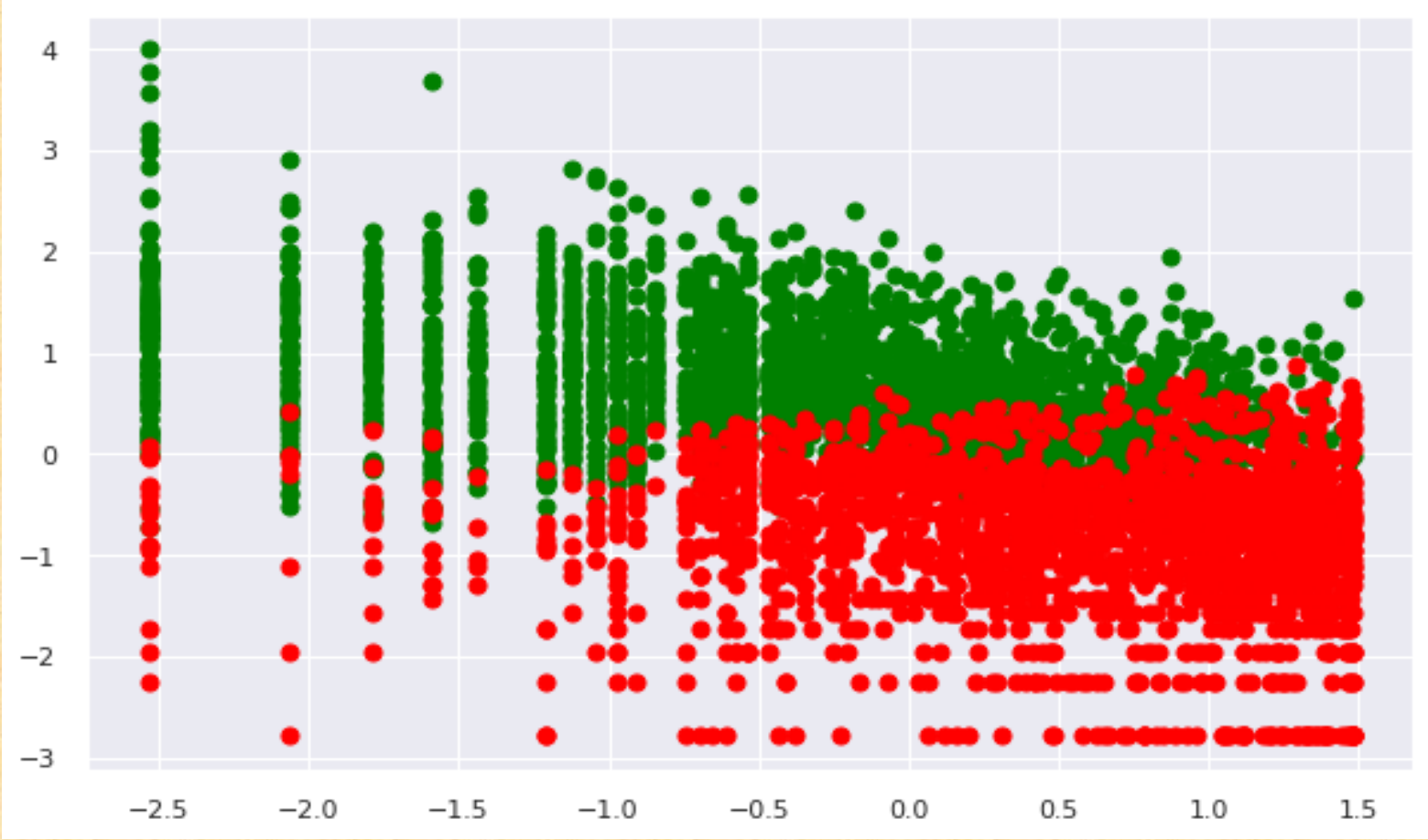
Modelling – KMeans & Silhouette Analysis



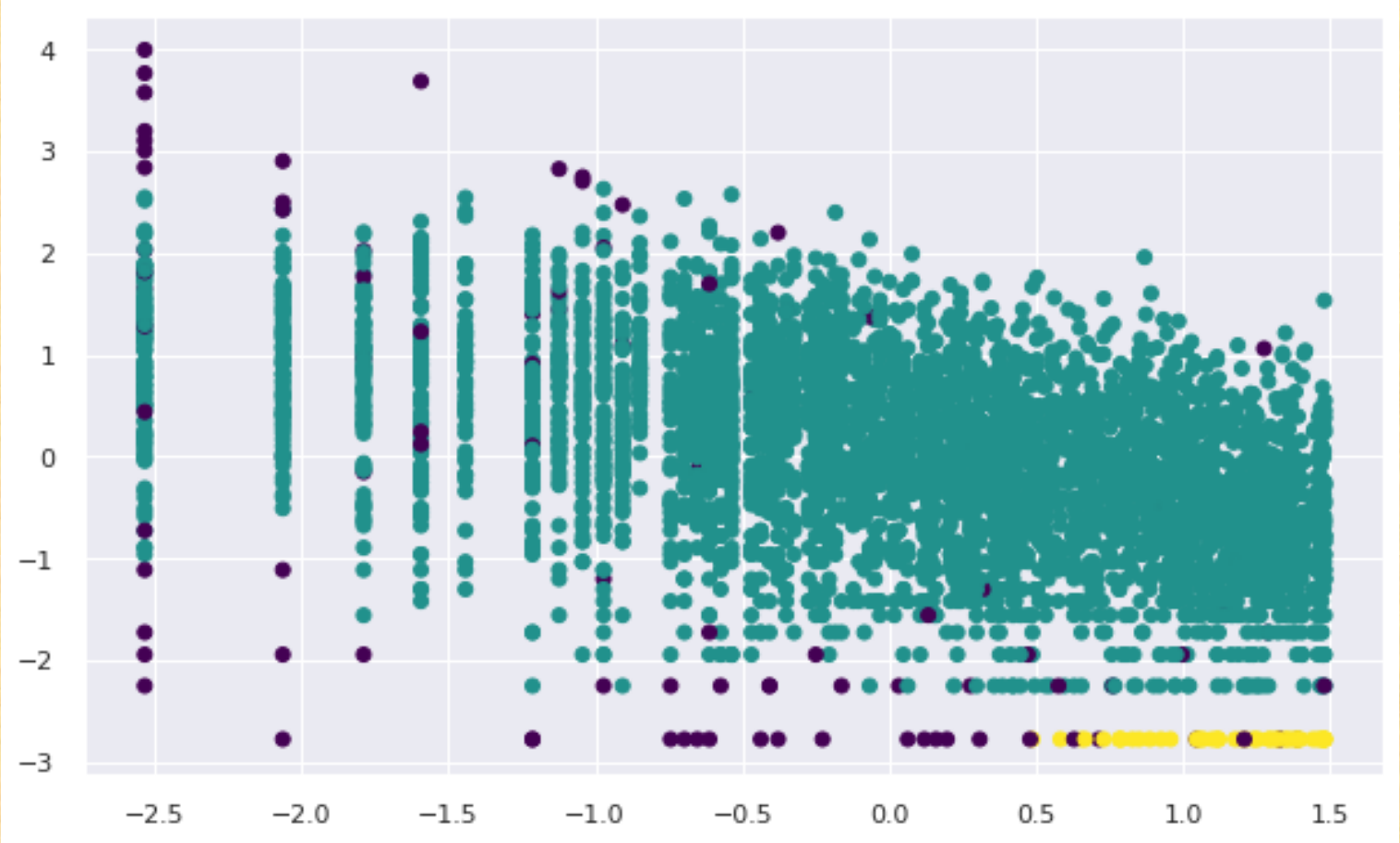
Modelling - Dendrogram



Modeling – Hierarchal Clustering



Modeling – DB SCAN



Clusters

	Recency			Frequency			Monetory			
	mean	min	max	mean	min	max	mean	min	max	count
Model1Clusters										
0	140.389140	1	373	24.998355	1	174	471.304797	1.00	77183.60	2431
1	30.459644	1	372	173.983753	1	7676	4057.373130	150.61	280206.02	1908

Conclusion

A. EDA Outcomes:

The retail store has a large share in local region i.e., UK.

The store has least market share in Saudi Arabia.

There are 3877 Unique Descriptions available in Dataset.

Customer ID : 17841 is top most customer by having large count in No. of Purchaces.

Customer with ID 14646 is the buyer of large quantity of the store.

The product having description as "PAPER CRAFT , LITTLE BIRDIE" is the most selling preoduct in store.

B. Challenges:

1. Missing Values - Description & CustomerID are having 0.26% & 24.92% of missing values respectively.

2. Duplicated Data - 5225 in count.

3. Outliers - Quantity & Unit Price Columns.

Conclusion:

C. Modelling Summary:

1. We can observe that the TWO clusters are clearly formed using KMeans- Elbow Approach.

2. We can observe that the TWO clusters are clearly defined using KMeans- Silhouette Analysis Approach.

3. The TWO clusters are clearly separated using Hierarchal clustering using Dendrogram Approach.

4. The clusters are clearly separated using DB SCAN clustering. DB Scan also creates few noise data points on clustering, which can exempted. Hence, the no. of optimal clusters can be 2.

5. Model1 i.e., KMeans with Elbow Approach having highest Score, Hence, we can conclude that it works better for clustering on this data.

6. Finally, We formed TWO Clusters:

Cluster 1 - Low Recency, High Frequency and High Monetory Values

Cluster 2 - High Recency, Low Frequency and Low Monetory Values

Thankyou