Online Retail Dataset Analysis

Dataset link:https://www.kaggle.com/datasets/abhishekrp1517/online-retail-transactions-dataset

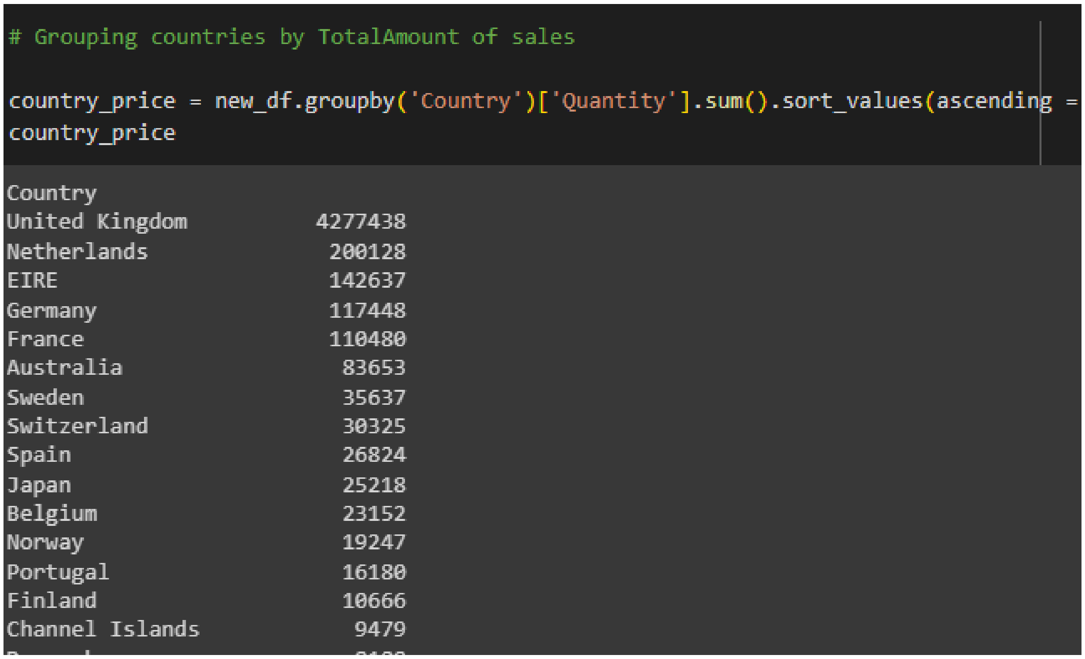
Dataset features

1. InvoiceNo
2. StockCode
3. Description
4. Quantity
5. Invoice Date
6. UnitPrice
7. CustomerID
8. Country

- Exploratory Data Analysis (EDA) is a crucial step in the data analysis process. Its main goal is to understand the data, gain insights, and identify patterns, relationships, and potential issues.
- Clustering algorithms are used in various fields and applications to group similar data points together based on certain features or characteristics.

I did an exploratory data analysis (EDA), and used clustering algorithms to segment customers effectively with the given dataset.

Exploitary Analysis

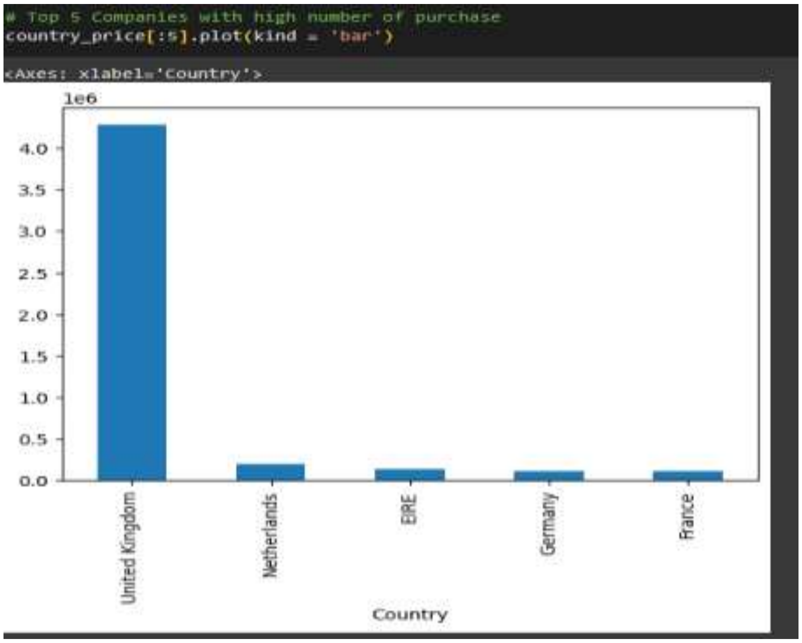


```
# Grouping countries by TotalAmount of sales

country_price = new_df.groupby('Country')['Quantity'].sum().sort_values(ascending =
country_price
```

```
Country
United Kingdom          4277438
Netherlands              200128
EIRE                     142637
Germany                  117448
France                   110480
Australia                 83653
Sweden                    35637
Switzerland               30325
Spain                     26824
Japan                     25218
Belgium                   23152
Norway                    19247
Portugal                  16180
Finland                   10666
Channel Islands            9479
```

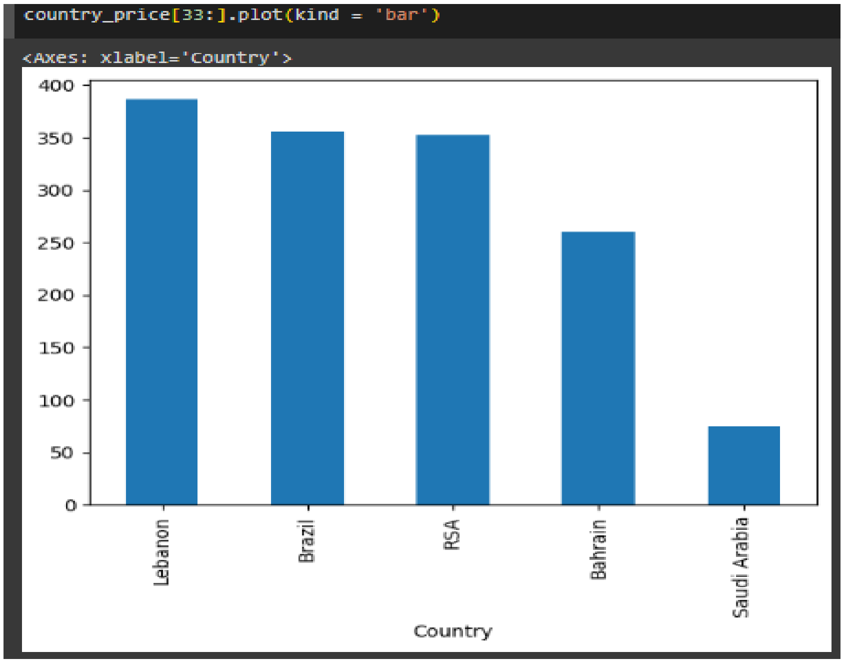
1 .Visualizing of top 5 companies by total amount of sales



```
# Top 5 Companies with high number of purchase
country_price[:5].plot(kind = 'bar')
```

```
<Axes: xlabel='Country'>
```

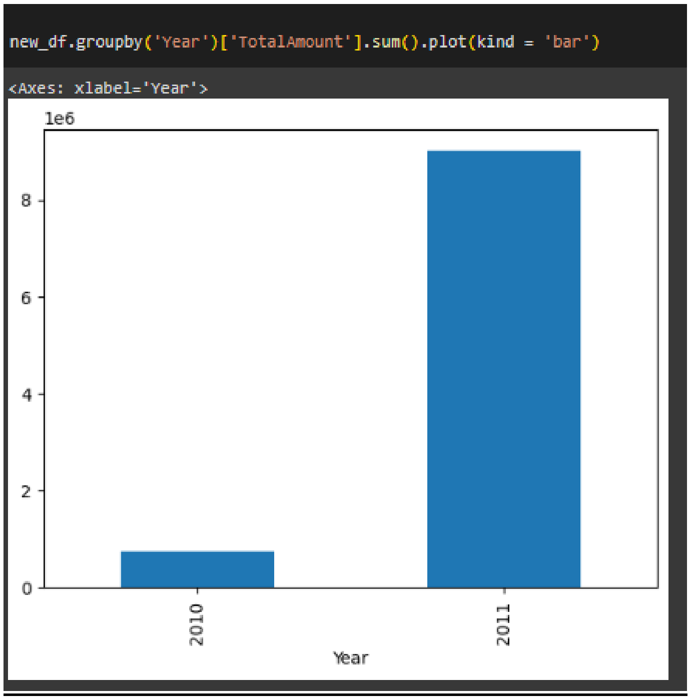
2. Visualizing of least 5 companies by total amount of sale



```
country_price[33:].plot(kind = 'bar')
```

```
<Axes: xlabel='Country'>
```

3. Total sales for different years



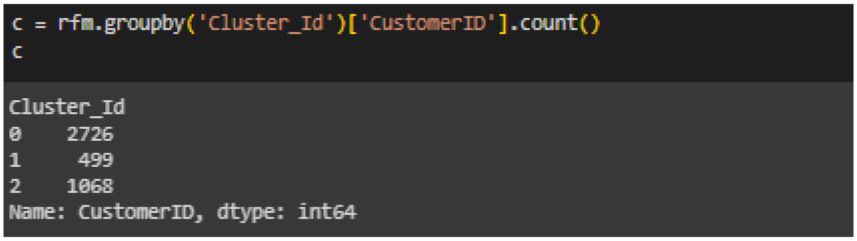
```
new_df.groupby('Year')['TotalAmount'].sum().plot(kind = 'bar')
```

```
<Axes: xlabel='Year'>
```

- By the exploratory data analysis we can see,
  - ➢ The year 2011 has the highest no of sales.
  - ➢ The United Kingdom has the highest no of sales and Saudi Arabia has the least no of sales.

Use of clustering algorithms

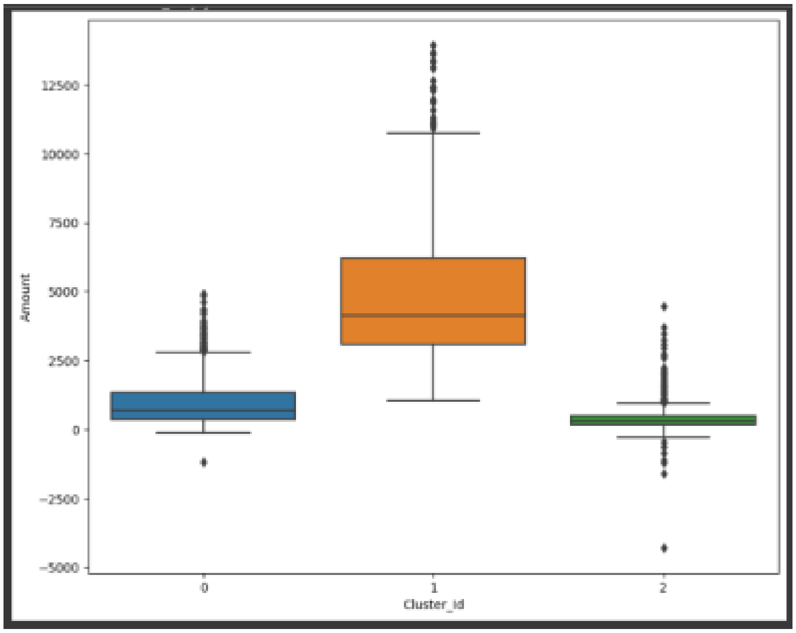
- I divided the dataset into three clusters and did a cluster analysis using K-means clustering.

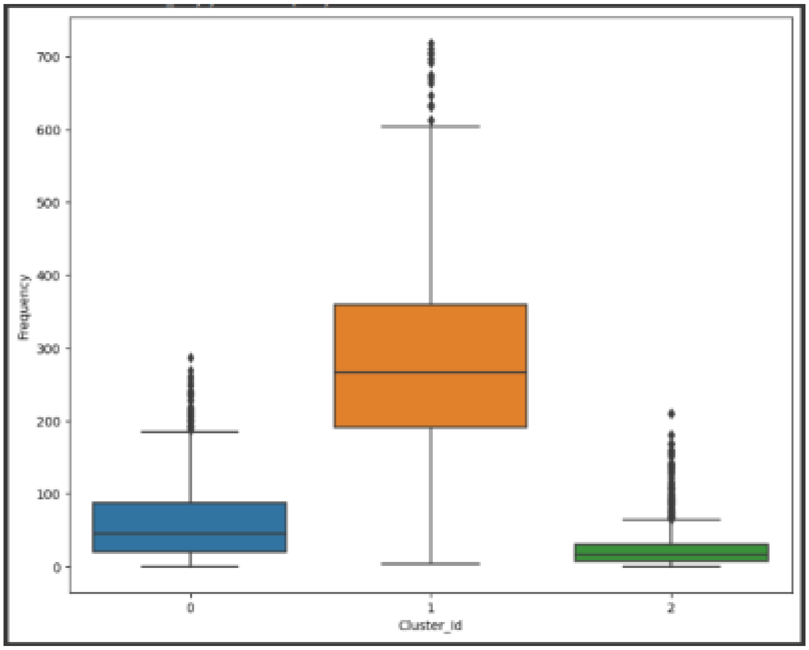


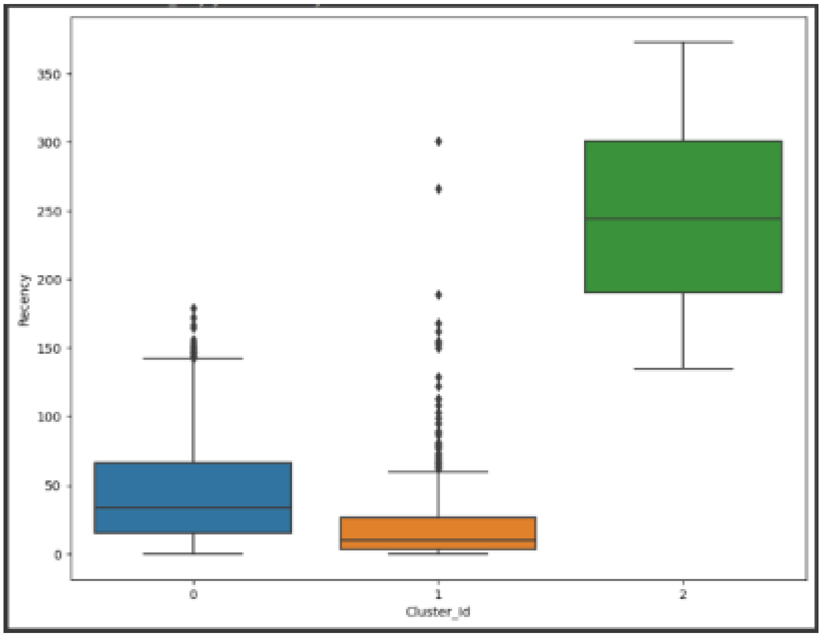
```
c = rfm.groupby('Cluster_Id')['CustomerID'].count()
c
```

```
Cluster_Id
0    2726
1     499
2    1068
Name: CustomerID, dtype: int64
```

Some of the plots I created used to display the results including Amount , Frequency and recency of sales according to several clusters are shown below.







- By examining the above analysis we can conclude that,
  - Customers with Cluster Id 1 are the customers with a 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 a business point of view.