High Value Customer Identification

Team:
Anuradha Tidke
Priya Johny
Rakshith Reddy Eleti
Venkata Gangadhar Naveen Palaka

Introduction

❑ UK-based online retail store contains sales data for different products for the period of one year (Nov 2016 to Dec 2017) and have updated the dataset (in 2020).

❑ The project objective is to find significant customers for the business who make high purchases of their favorite products.

❑ The packages used to implement the network are NumPy, Pandas, Seaborn, Matplotlib, Sklearn and Scipy.

❑ The data mining algorithm used is **K-means Clustering** and it is in standard form.

❑ **Elbow method** approach is used for metric evaluation.

Data Cleaning & Preprocessing

Data Description

- Trans-national dataset that contains all the transactions occurring between Nov-2016 to Dec-2017 for a UK-based online retail store.

- Source: Kaggle

- The data contains over 541909 entries and 9 columns

  “InvoiceNo” -[object]

  “Stock Code” -[object]

  “Description” -[object]

  “Quantity” -[int64]

  “CustomerID” -[float64]

  “Country” -[object]

  “InvoiceDate” -[datetime64]

  “Unit Price” -[float64]

Data Cleaning

❑ Checking unique data points, datatype, null values, duplicates.

❑ Dropped 1 empty column

❑ Drop NaN’s/null values

❑ Replace whitespaces or symbols

❑ Extracting year, month and date from ‘InvoiceDate’

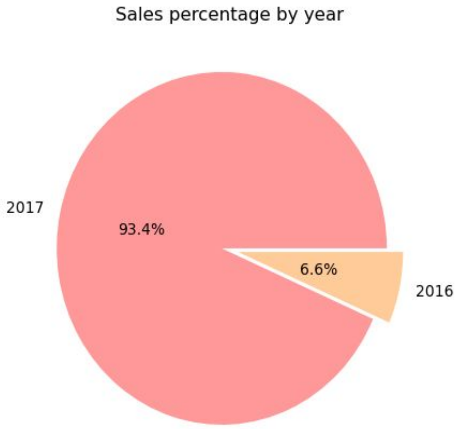
❑ Adding new variable ‘TotalExpense’ -> [‘Quantity’]*[‘UnitPrice’]

❑ Change datatype of ‘CustomerId’ from float to integer

Exploratory Data Analysis

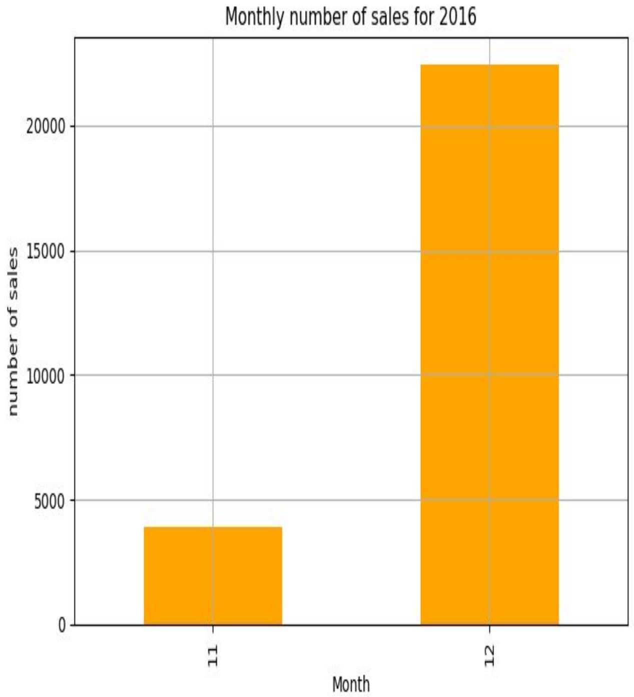
Plots

- **Plotting Year Wise Number of Sales**

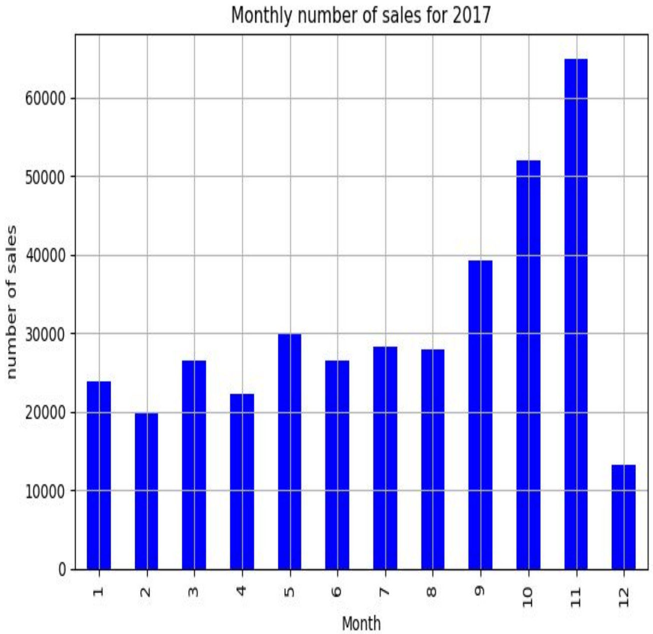


We can observe in the pie chart, in the year 2016 it records 6.6% which is very low as compared to the sales in the year 2017 which is 93.4%

- # Plotting Monthly Sales by Year

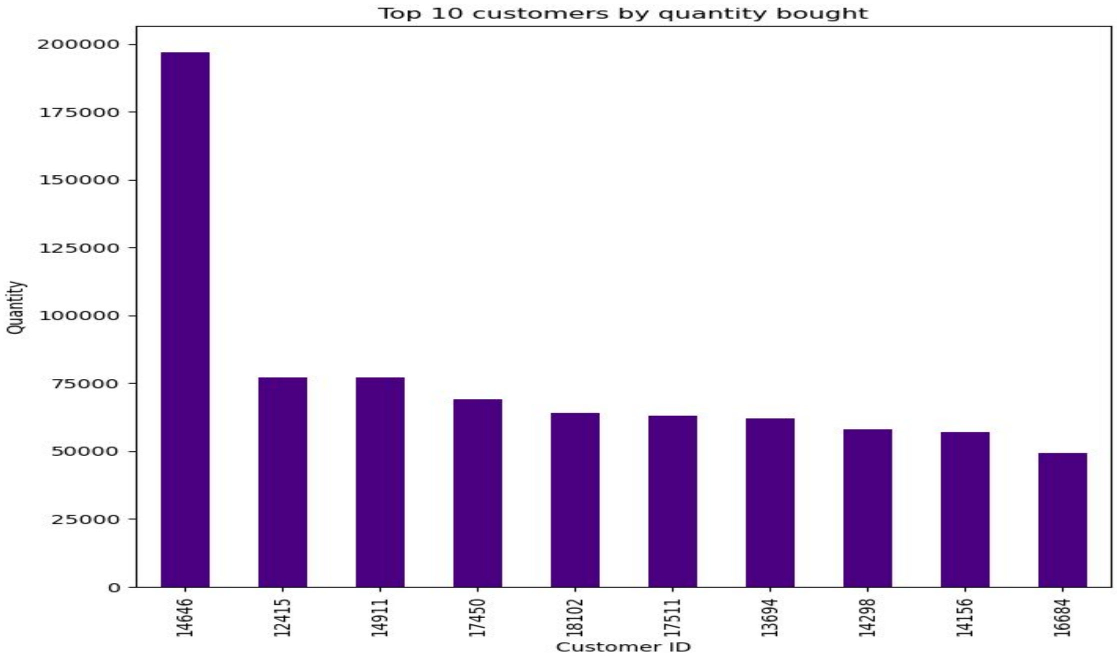


We can see in the year 2016, November has less than 5000 sales whereas in December it has more 20000 sales.



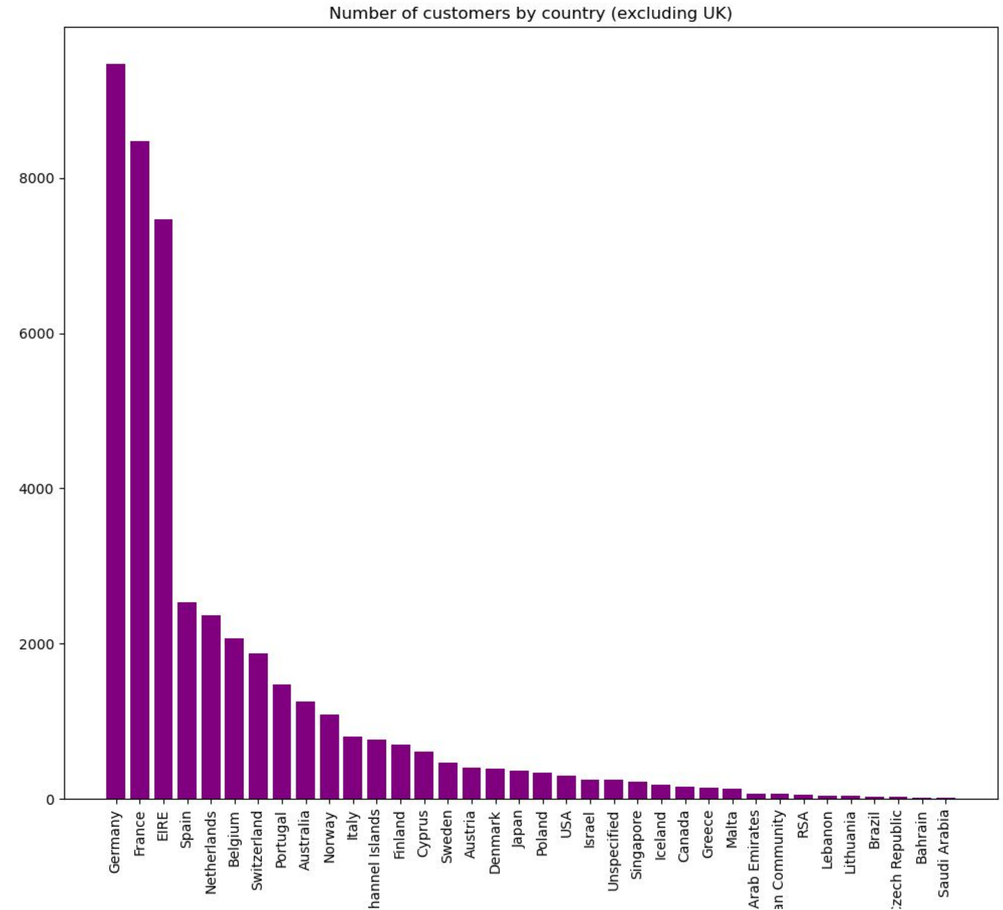
For the year 2017, The first six months sales fluctuate up and down not crossing more than 30k sales, but from July we can observe it starts increasing drastically till November which records highest number of sales of more than 60k, then suddenly sales drop in December to lowest of all.

- **Plotting Top 10 Customers by Quantity Bought**



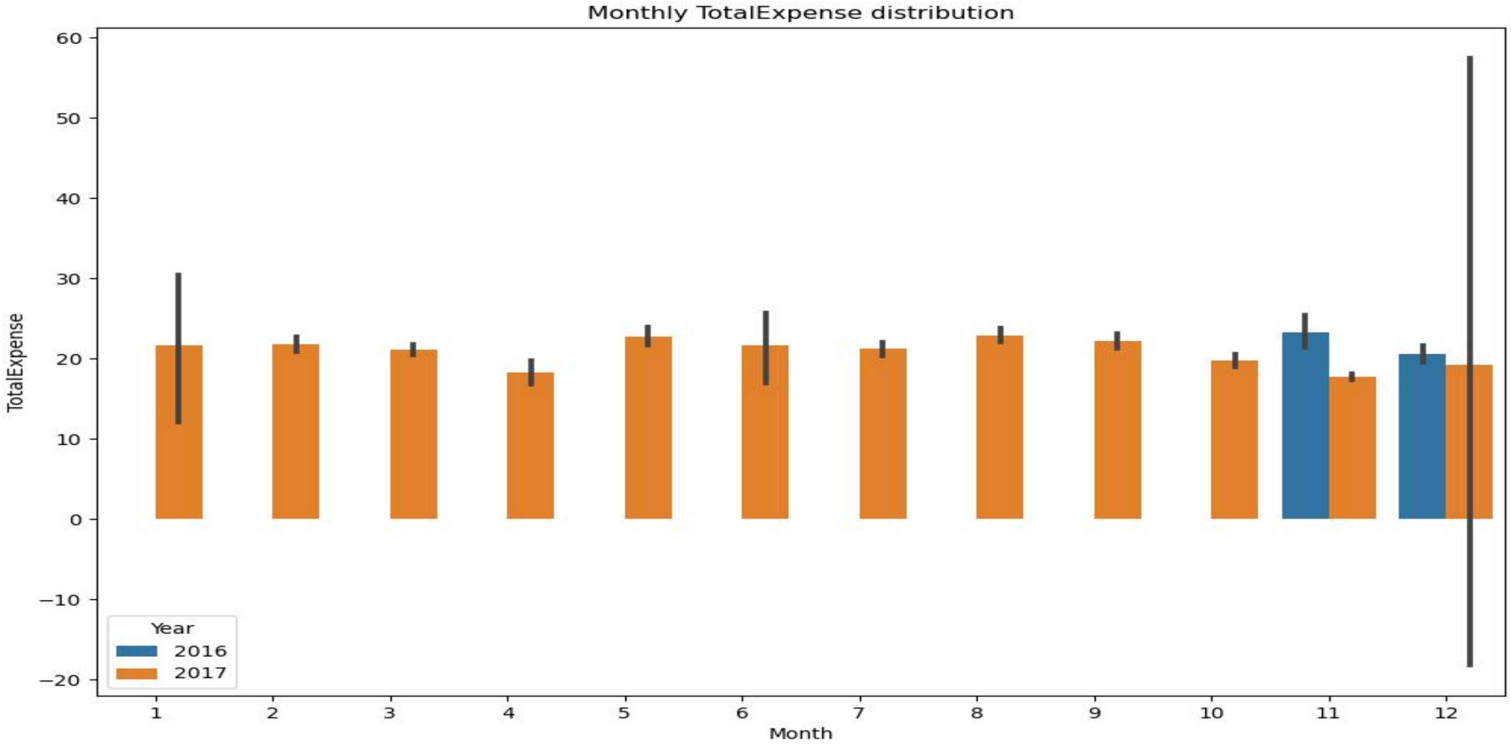
Top 10 customers who bought more, the customerID with 14646 has bought highest number of quantities, who can be considered as the highest valued customer.

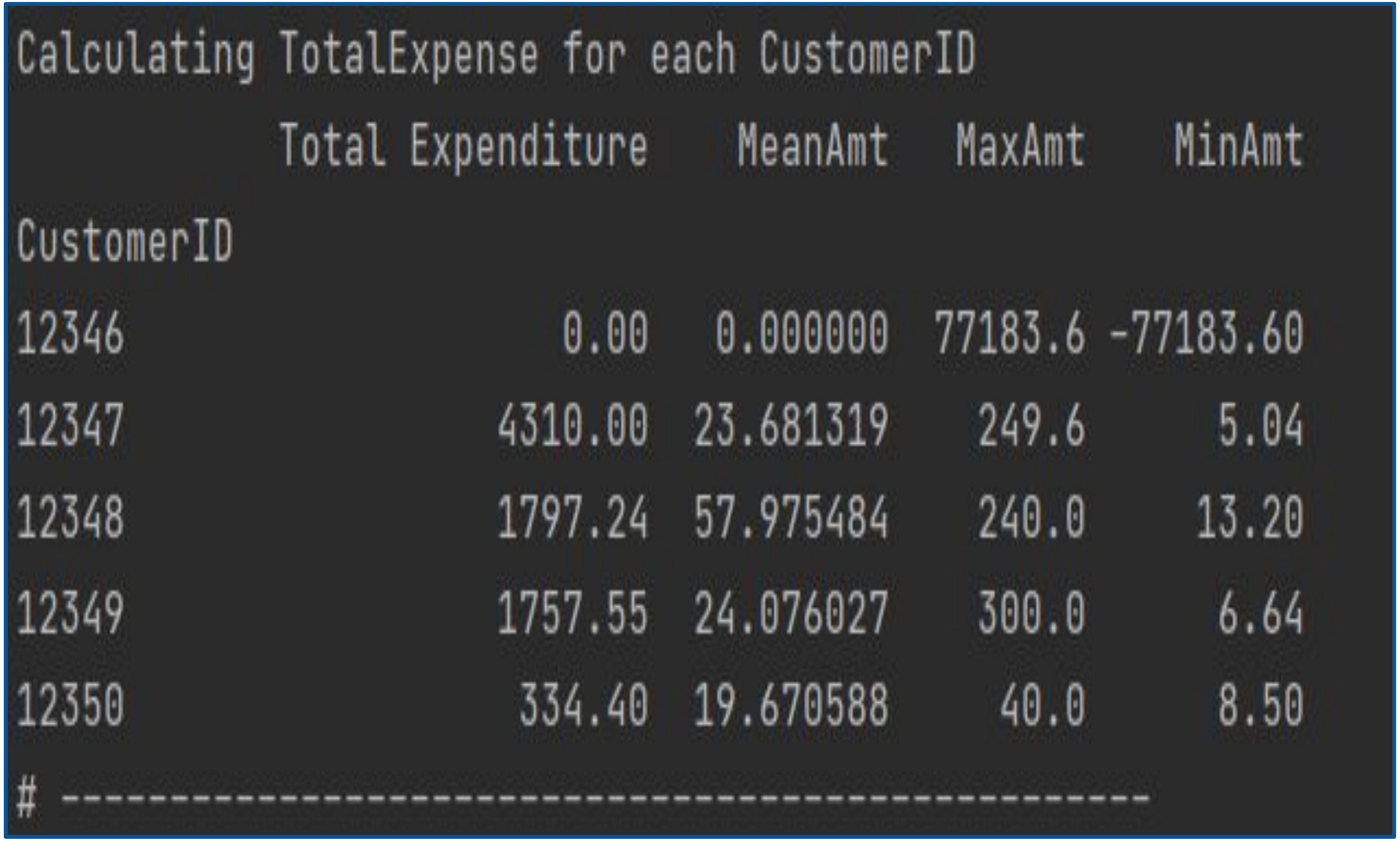
- **Country Wise Sales (except UK)**



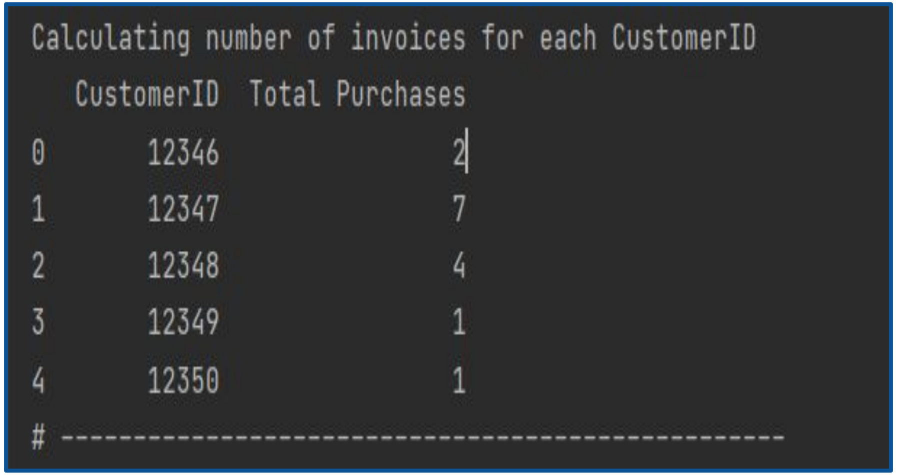
We can see in the above bar chart, Germany records second highest sales (9,477) (after UK) followed by France. We already know UK has the first highest sales (3,56,616) with major differences.

- **Plotting Monthly Total Expense Distribution**

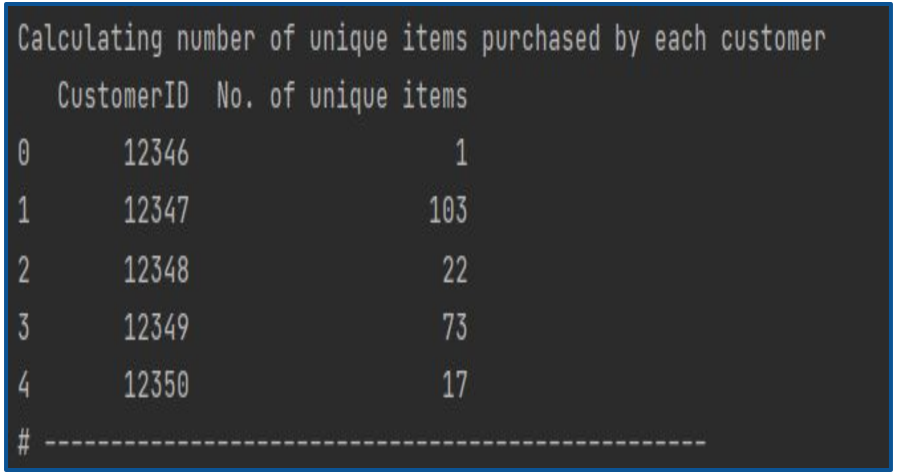


- 

```
Calculating TotalExpense for each CustomerID
            Total Expenditure   MeanAmt   MaxAmt    MinAmt
CustomerID
12346                    0.00  0.000000  77183.6 -77183.60
12347                 4310.00 23.681319    249.6      5.04
12348                 1797.24 57.975484    240.0     13.20
12349                 1757.55 24.076027    300.0      6.64
12350                  334.40 19.670588     40.0      8.50
# ---------------------------------------------------
```

- 

```
Calculating number of invoices for each CustomerID
   CustomerID  Total Purchases
0       12346                2
1       12347                7
2       12348                4
3       12349                1
4       12350                1
# --------------------------------------------------
```

- 

```
Calculating number of unique items purchased by each customer
   CustomerID  No. of unique items
0       12346                    1
1       12347                  103
2       12348                   22
3       12349                   73
4       12350                   17
# ----------------------------------------------
```

Algorithm

Algorithm Description

Clustering Algorithms Used

❑ K-Means

Tasks to be Performed

❑ Use the clustering methodology to segment customers into groups

❑ Clustering algorithm used: K-means

❑ Identify the right number of customer segments (value of k) using elbow method

❑ Identify which cluster groups the highly valued customers

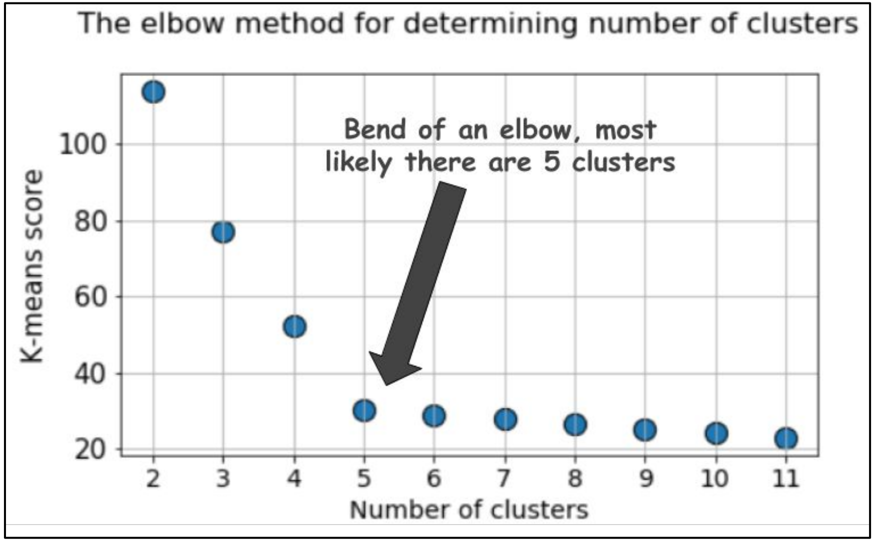
Performance Metrics

Clustering is often done for such analytics with the goal of market segmentation.

It is, therefore, easily conceivable that, depending on the number of clusters, appropriate marketing personnel will be allocated to the problem.

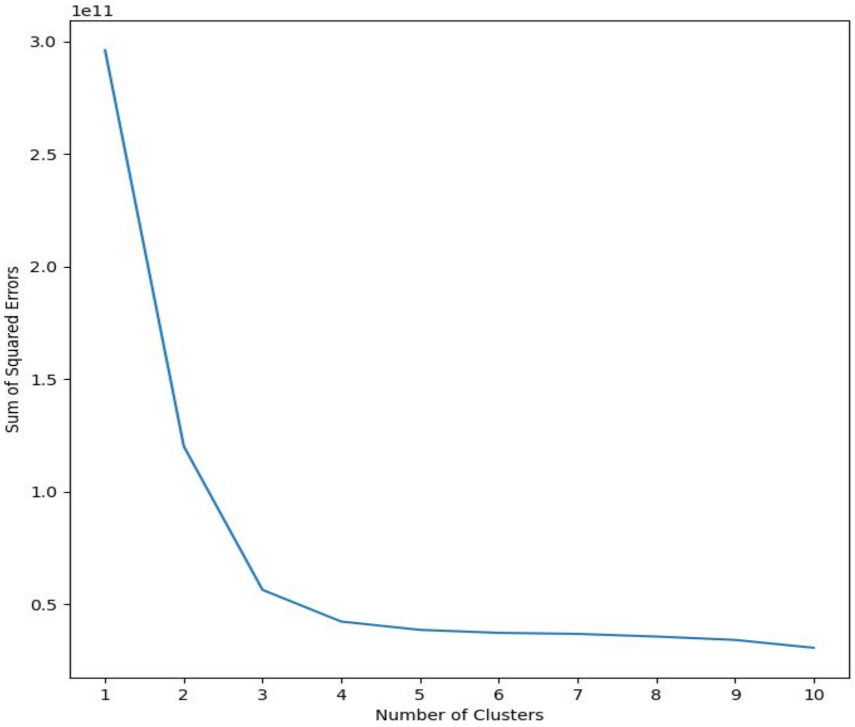
Consequently, a wrong assessment of the number of clusters can lead to sub-optimum allocation of precious resources.

The Elbow Method



It involves running the algorithm multiple times over a loop, with an increasing number of cluster choices and then plotting a clustering score as a function of the number of clusters.

Plotting SSE Against Various Values of k

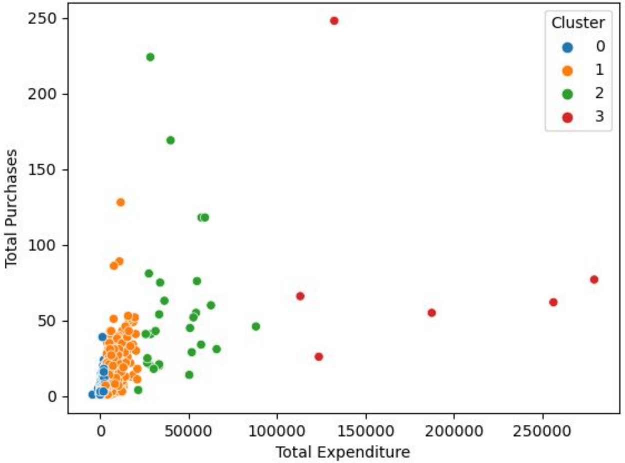


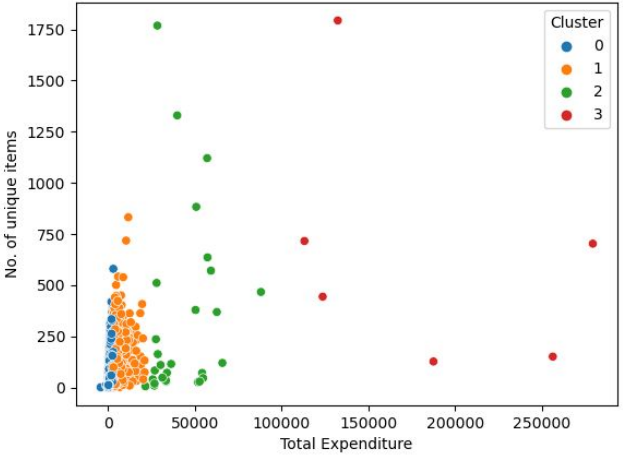
For each customer ID, we calculated:

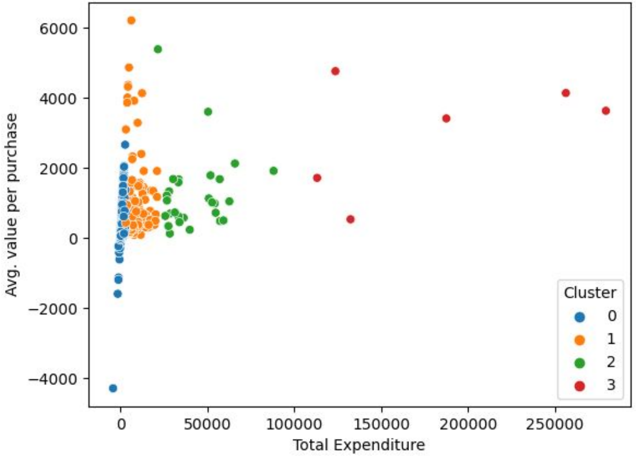
- TotalExpense
- Number of purchases
- Average amount per purchase
- Number of unique items bought

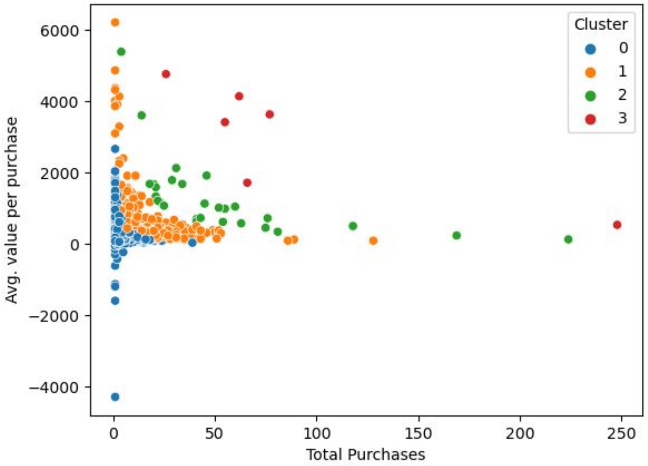
Using Elbow method, we can decide k = 4 as an appropriate number of clusters.

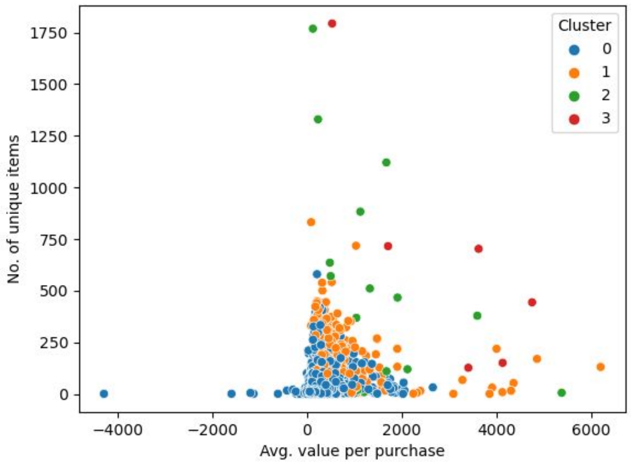
Scatter Plots Showing Different Clusters

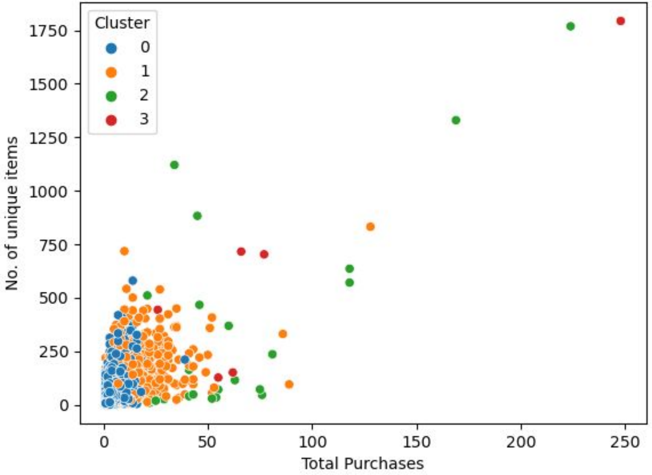


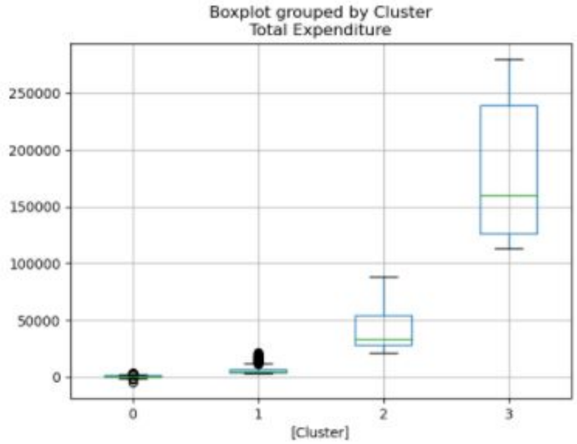


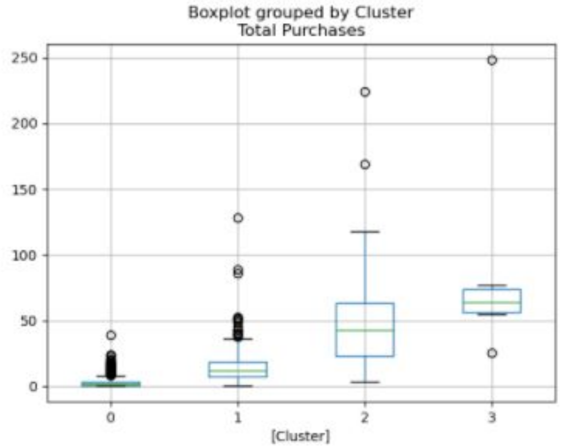


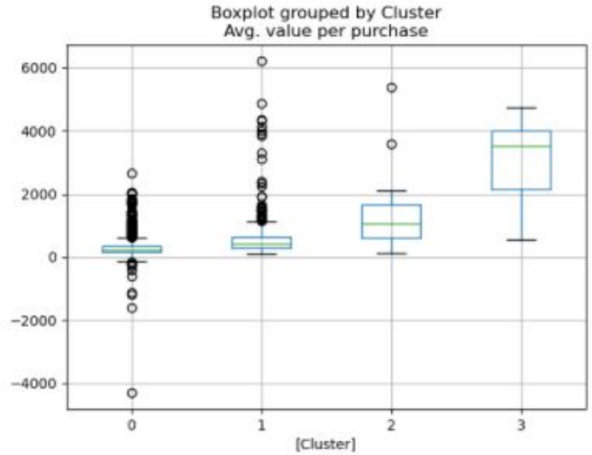

Cluster
0
1
2
3
Avg. value per purchase
6000
4000
2000
0
−2000
−4000
0
50
100
150
200
250
Total Purchases

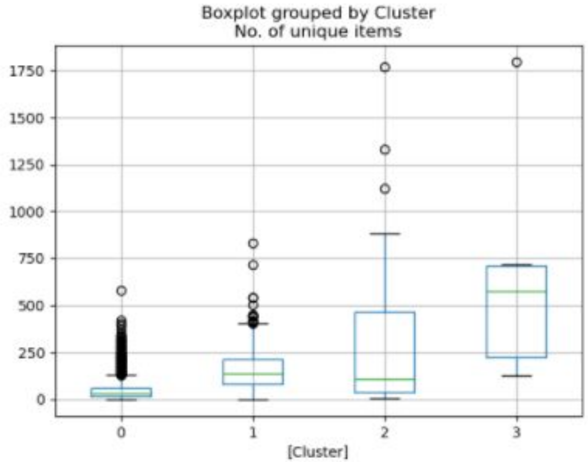

Cluster
0
1
2
3
No. of unique items
1750
1500
1250
1000
750
500
250
0
−4000
−2000
0
2000
4000
6000
Avg. value per purchase


Cluster
0
1
2
3
No. of unique items
1750
1500
1250
1000
750
500
250
0
0
50
100
150
200
250
Total Purchases


Boxplot grouped by Cluster
Total Expenditure
250000
200000
150000
100000
50000
0
0
1
2
3
[Cluster]


Boxplot grouped by Cluster
Total Purchases
250
200
150
100
50
0
0
1
2
3
[Cluster]


Boxplot grouped by Cluster
Avg. value per purchase
6000
4000
2000
0
−2000
−4000
0
1
2
3
[Cluster]

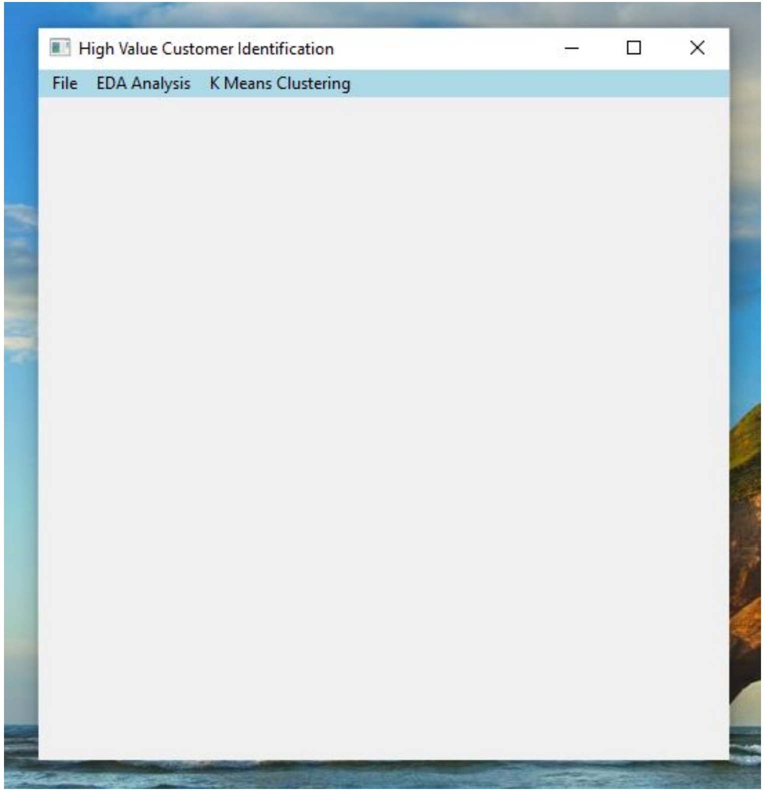

Boxplot grouped by Cluster
No. of unique items
1750
1500
1250
1000
750
500
250
0
0
1
2
3
[Cluster]

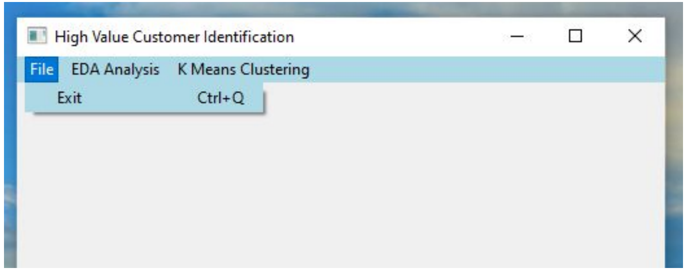
Conclusions

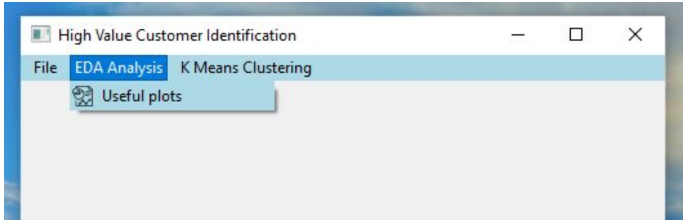
Based on these results from the cluster feature graphs and the box plots, we can categorize the clusters as follows:

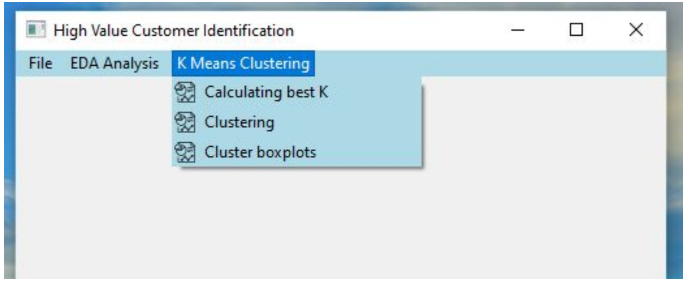
- **Cluster 0** and **1 need** to be focussed **more** on, in terms of **discounts** and other **offers** in order to increase their purchase numbers.
- **Cluster 2** consists of **regular** value but **loyal customers** who visit the store pretty often.
- **Cluster 3** can be considered as **high-valued customers** for whom a loyalty program should be rolled out. These customers are loyal as well seeing the number of purchases.

Graphical User Interface

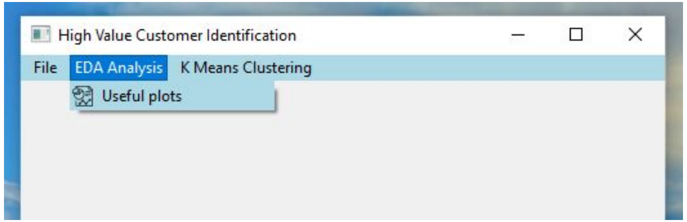


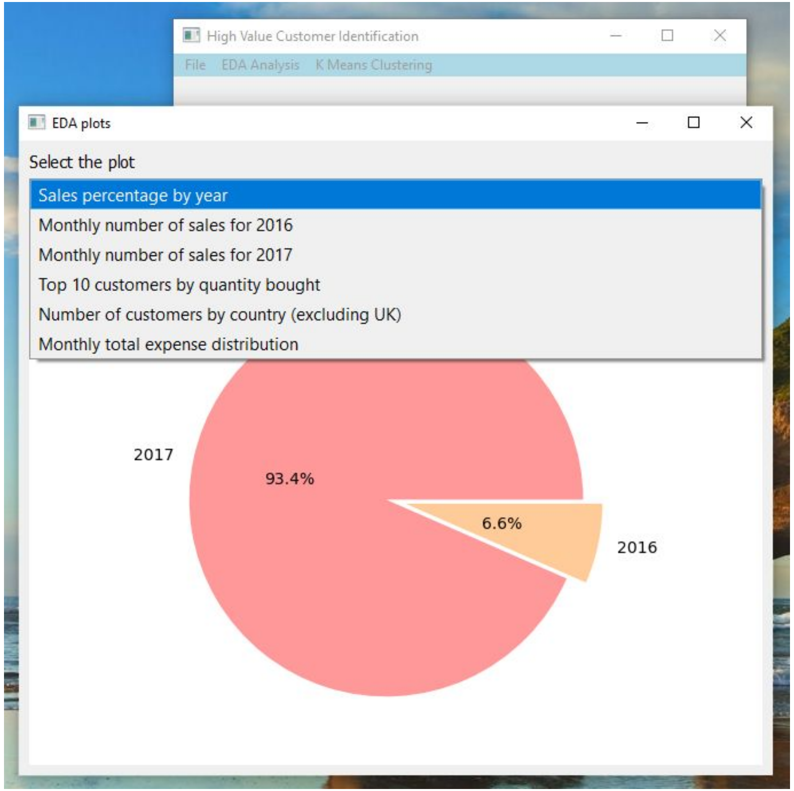




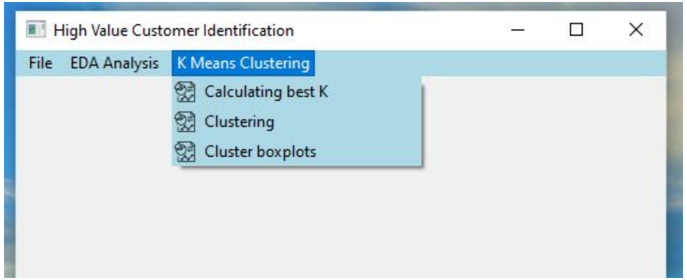


Graphical User Interface

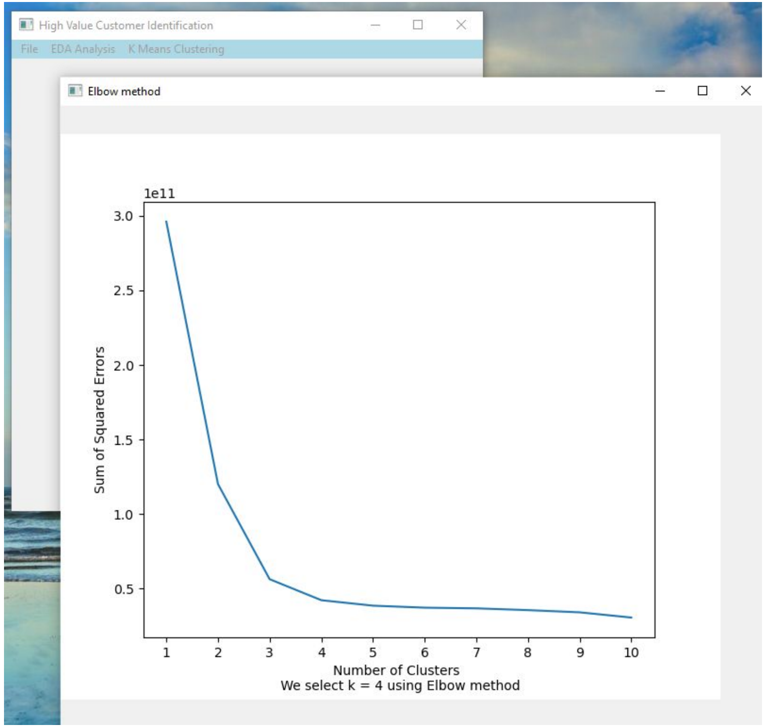



High Value Customer Identification
File
EDA Analysis
K Means Clustering
EDA plots
Select the plot
Sales percentage by year
Monthly number of sales for 2016
Monthly number of sales for 2017
Top 10 customers by quantity bought
Number of customers by country (excluding UK)
Monthly total expense distribution
2017
93.4%
6.6%
2016

Graphical User Interface

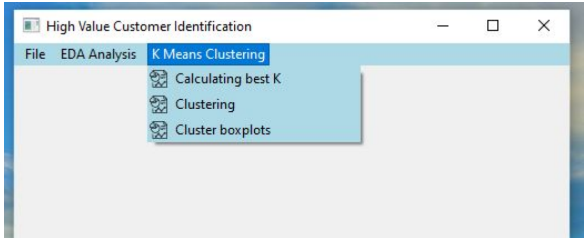


K Means Clustering -> Calculating best K

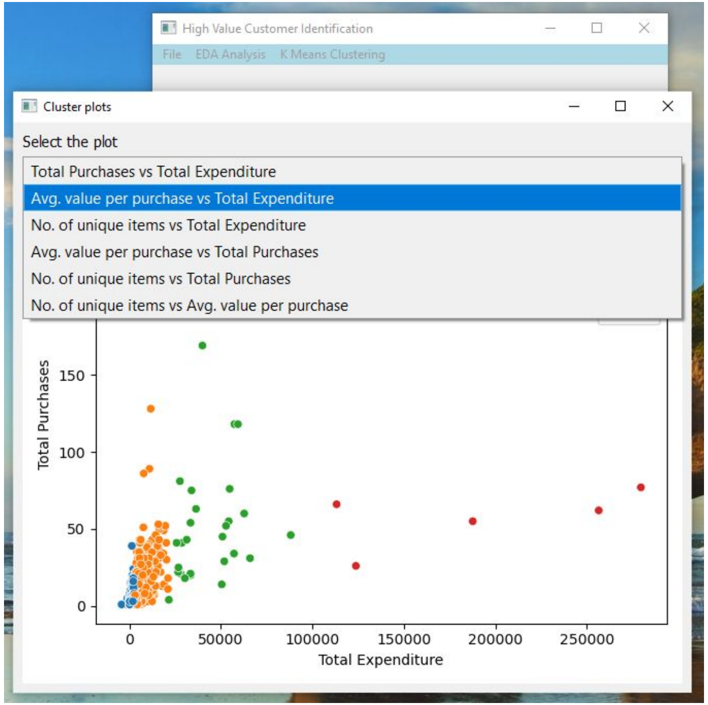


We select k = 4 using Elbow method

Graphical User Interface



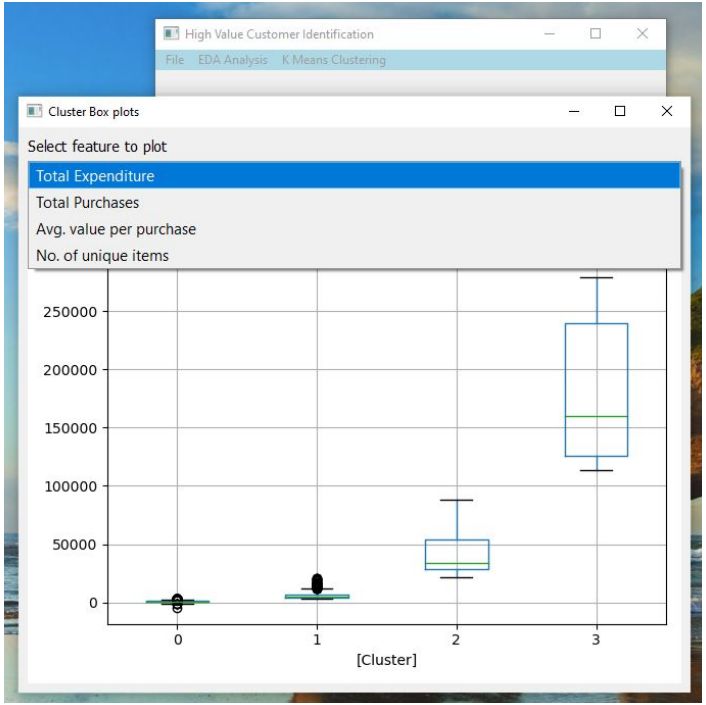
K Means Clustering -> Clustering



Select the plot

- Total Purchases vs Total Expenditure
- Avg. value per purchase vs Total Expenditure
- No. of unique items vs Total Expenditure
- Avg. value per purchase vs Total Purchases
- No. of unique items vs Total Purchases
- No. of unique items vs Avg. value per purchase

K Means Clustering -> Cluster Boxplots



Select feature to plot

- Total Expenditure
- Total Purchases
- Avg. value per purchase
- No. of unique items

Questions?