Task – 3

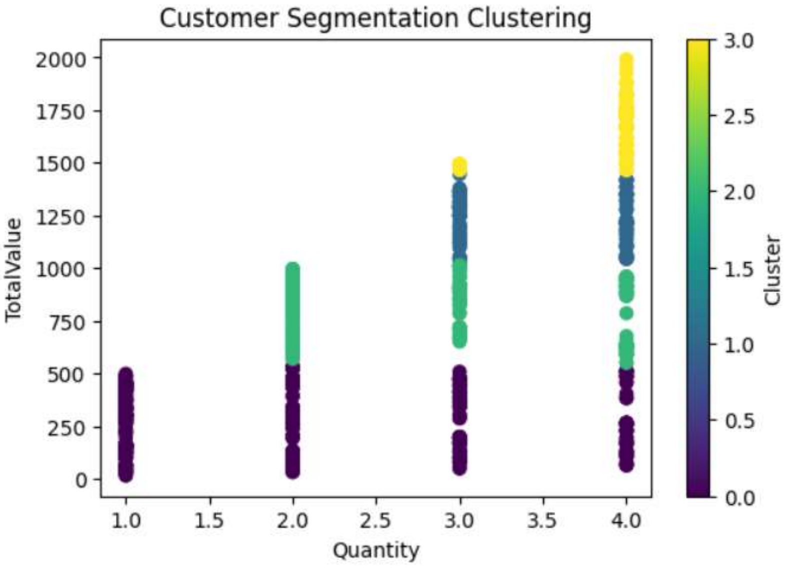
Customer Segmentation / Clustering

Perform customer segmentation using clustering techniques. Use both profile information (from Customers.csv) and transaction information (from Transactions.csv).

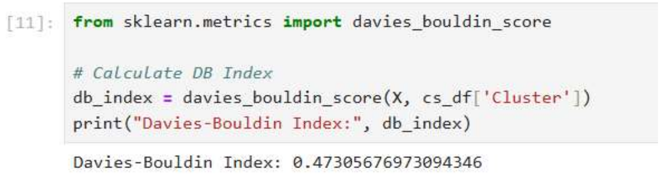
1. You have the flexibility to choose any clustering algorithm and any number of clusters in between(2 and 10).
2. Calculate clustering metrics, including the **DB Index**(Evaluation will be done on this).
3. Visualise your clusters using relevant plots.

1. The main goal of clustering is to find out **High-Value Customers** so, we need relevant features to identify **Customer Spending Behaviour**. That's why Selected specific columns such as "Region (Numeric)", "Quantity" and "TotalValue". After clustering, we will analyze which clusters have the highest average Total Value and customers in the **Highest-Value Cluster** will be considered as **High-Value Customers.**

2. Used K-Means Clustering, Initially used K = 4 because, we have Four Regions (South America, North America, Europe and Asia) it might form a cluster according to that.

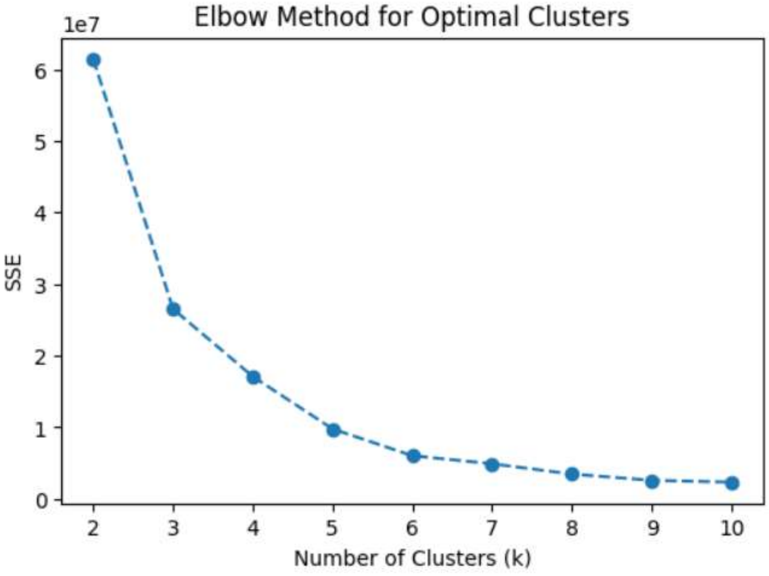


As main Evaluation Matrix is Davies-Bouldin Index for K = 4 we got DB Index = 0.4730. This is Low but we will use another Metric called as "Elbow Method" to identify closest number of K.



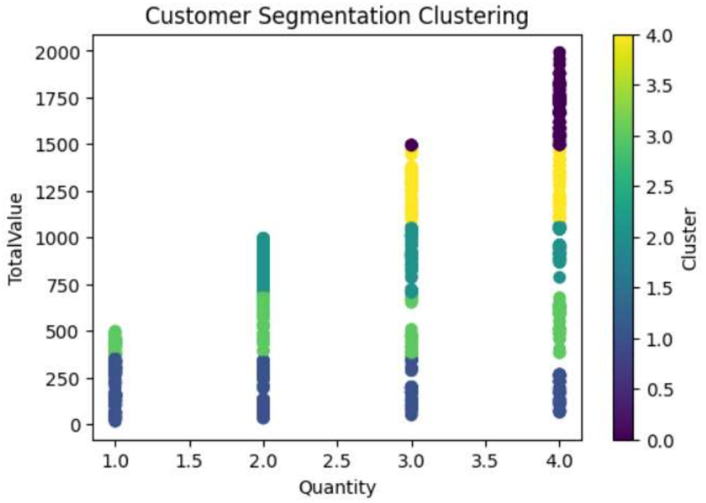
```
[11]: from sklearn.metrics import davies_bouldin_score

      # Calculate DB Index
      db_index = davies_bouldin_score(X, cs_df['Cluster'])
      print("Davies-Bouldin Index:", db_index)

      Davies-Bouldin Index: 0.47305676973094346
```

3. **Elbow Plot** :

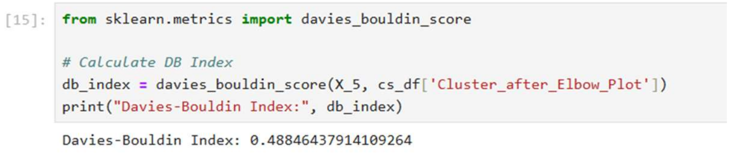


As mentioned in the task I have flexibility to choose any number of clusters between 2 to 10 range. Based, on Elbow Graph elbow appears to be around “K = 5” or “K = 6” So, we can perform next actions as per that.

4. Now, below clustering graph is of K = 5, we can see a High value cluster is has formed the “The Purple Colour one”. But, also needs to validate Quality of Cluster with DB Index.

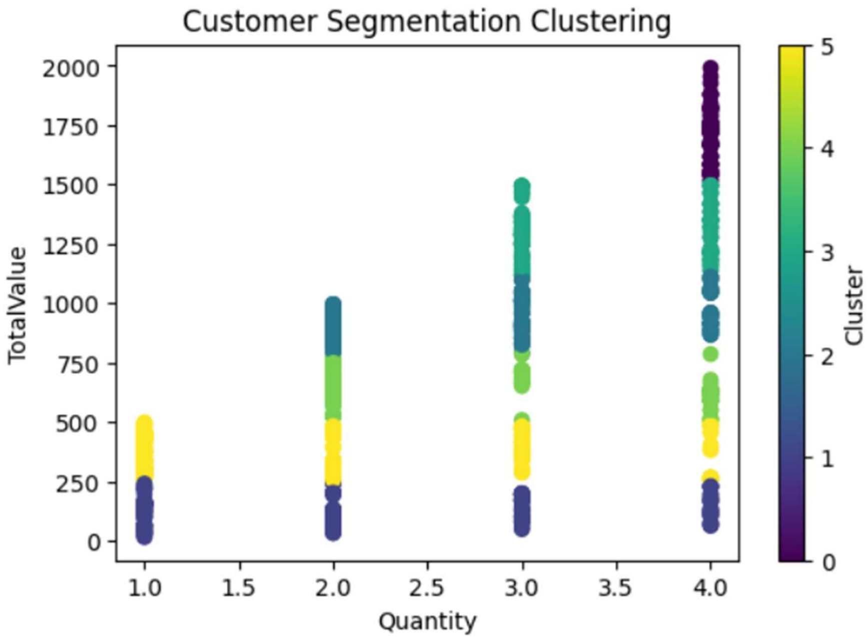


Davies-Bouldin Index for K = 5 is 0.4884 slightly higher than last one this might indicating a some problem.

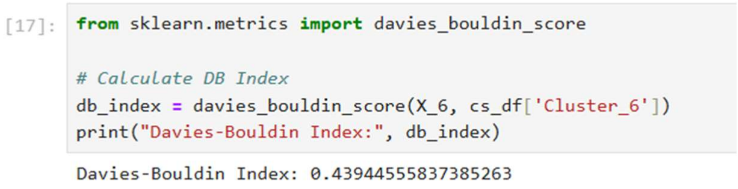


```
[15]: from sklearn.metrics import davies_bouldin_score

      # Calculate DB Index
      db_index = davies_bouldin_score(X_5, cs_df['Cluster_after_Elbow_Plot'])
      print("Davies-Bouldin Index:", db_index)

      Davies-Bouldin Index: 0.48846437914109264
```

5. Now, I am taken K = 6 which was also suggested by Elbow graph and you can see it has formed clear Six distinct clusters and we have likely got High-Value Cluster (Purple one). This cluster likely consists list of the Highest Spending Customers.

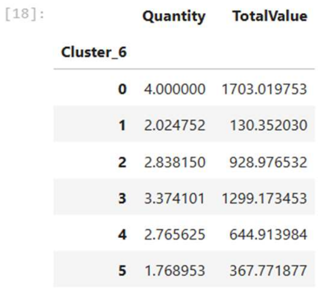


Davies-Bouldin Index for K = 6 is 0.4394 best among all indicating most distinct clusters.



```
[17]: from sklearn.metrics import davies_bouldin_score

      # Calculate DB Index
      db_index = davies_bouldin_score(X_6, cs_df['Cluster_6'])
      print("Davies-Bouldin Index:", db_index)

      Davies-Bouldin Index: 0.43944555837385263
```

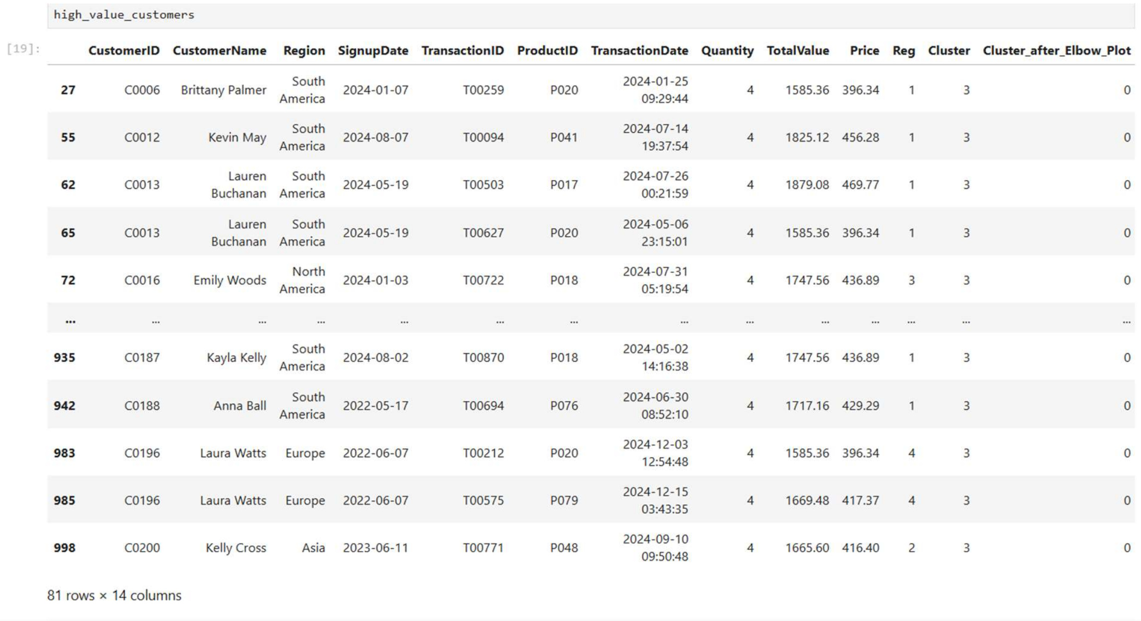
6. Perfect. We have successfully identified the **High-Value Cluster** i.e **Cluster – 0 (Purple).** This cluster consists of customers who have spent a minimum of $1703, making them highest value segment. Given their Purchasing Behaviour makes them key target group for business strategies.



[18]:

	Quantity	TotalValue
Cluster_6		
0	4.000000	1703.019753
1	2.024752	130.352030
2	2.838150	928.976532
3	3.374101	1299.173453
4	2.765625	644.913984
5	1.768953	367.771877

7. **High-Value Customers List** :



```
high_value_customers
```

[19]:

	CustomerID	CustomerName	Region	SignupDate	TransactionID	ProductID	TransactionDate	Quantity	TotalValue	Price	Reg	Cluster	Cluster_after_Elbow_Plot
27	C0006	Brittany Palmer	South America	2024-01-07	T00259	P020	2024-01-25 09:29:44	4	1585.36	396.34	1	3	0
55	C0012	Kevin May	South America	2024-08-07	T00094	P041	2024-07-14 19:37:54	4	1825.12	456.28	1	3	0
62	C0013	Lauren Buchanan	South America	2024-05-19	T00503	P017	2024-07-26 00:21:59	4	1879.08	469.77	1	3	0
65	C0013	Lauren Buchanan	South America	2024-05-19	T00627	P020	2024-05-06 23:15:01	4	1585.36	396.34	1	3	0
72	C0016	Emily Woods	North America	2024-01-03	T00722	P018	2024-07-31 05:19:54	4	1747.56	436.89	3	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
935	C0187	Kayla Kelly	South America	2024-08-02	T00870	P018	2024-05-02 14:16:38	4	1747.56	436.89	1	3	0
942	C0188	Anna Ball	South America	2022-05-17	T00694	P076	2024-06-30 08:52:10	4	1717.16	429.29	1	3	0
983	C0196	Laura Watts	Europe	2022-06-07	T00212	P020	2024-12-03 12:54:48	4	1585.36	396.34	4	3	0
985	C0196	Laura Watts	Europe	2022-06-07	T00575	P079	2024-12-15 03:43:35	4	1669.48	417.37	4	3	0
998	C0200	Kelly Cross	Asia	2023-06-11	T00771	P048	2024-09-10 09:50:48	4	1665.60	416.40	2	3	0

81 rows × 14 columns