[52]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

[2]:
```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
```

[10]:
```
df_cs = pd.read_csv("Customers.csv")
df_ts = pd.read_csv("Transactions.csv")
```

[14]:
```
df_cs
```

[14]:

	CustomerID	CustomerName	Region	SignupDate
0	C0001	Lawrence Carroll	South America	2022-07-10
1	C0002	Elizabeth Lutz	Asia	2022-02-13
2	C0003	Michael Rivera	South America	2024-03-07
3	C0004	Kathleen Rodriguez	South America	2022-10-09
4	C0005	Laura Weber	Asia	2022-08-15
...	...	...	...	...
195	C0196	Laura Watts	Europe	2022-06-07
196	C0197	Christina Harvey	Europe	2023-03-21
197	C0198	Rebecca Ray	Europe	2022-02-27
198	C0199	Andrea Jenkins	Europe	2022-12-03
199	C0200	Kelly Cross	Asia	2023-06-11

200 rows × 4 columns

[13]:
```
df_ts
```

[13]:

	TransactionID	CustomerID	ProductID	TransactionDate	Quantity	TotalValue	Price
0	T00001	C0199	P067	2024-08-25 12:38:23	1	300.68	300.68
1	T00112	C0146	P067	2024-05-27 22:23:54	1	300.68	300.68
2	T00166	C0127	P067	2024-04-25 07:38:55	1	300.68	300.68
3	T00272	C0087	P067	2024-03-26 22:55:37	2	601.36	300.68
4	T00363	C0070	P067	2024-03-21 15:10:10	3	902.04	300.68
...	...	...	...	...	...	...	...
995	T00496	C0118	P037	2024-10-24 08:30:27	1	459.86	459.86
996	T00759	C0059	P037	2024-06-04 02:15:24	3	1379.58	459.86
997	T00922	C0018	P037	2024-04-05 13:05:32	4	1839.44	459.86
998	T00959	C0115	P037	2024-09-29 10:16:02	2	919.72	459.86
999	T00992	C0024	P037	2024-04-21 10:52:24	1	459.86	459.86

1000 rows × 7 columns

[42]:
```
df = pd.merge(df_ts, df_cs, on='CustomerID')
```

[43]:
```
df_agg = df.groupby('CustomerID').agg({
    'TotalValue': 'sum',
    'Quantity': 'sum',
    'TransactionID': 'count',
    'Region': 'first'
}).reset_index()
```

[30]:
```
df_ts
```

[30]:

	TransactionID	CustomerID	ProductID	TransactionDate	Quantity	TotalValue	Price	Total_Spend	Average_Spend	Recency	Frequency
0	T00001	C0199	P067	2024-08-25 12:38:23	1	300.68	300.68	1979.28	494.820000	155 days 12:02:09.016597	4
1	T00112	C0146	P067	2024-05-27 22:23:54	1	300.68	300.68	2570.80	642.700000	245 days 02:16:38.016597	4
2	T00166	C0127	P067	2024-04-25 07:38:55	1	300.68	300.68	3232.88	538.813333	277 days 17:01:37.016597	6
3	T00272	C0087	P067	2024-03-26 22:55:37	2	601.36	300.68	6604.23	943.461429	307 days 01:44:55.016597	7
4	T00363	C0070	P067	2024-03-21 15:10:10	3	902.04	300.68	3125.49	781.372500	312 days 09:30:22.016597	4
...	...	...	...	...	...	...	...	...	...	...	...
995	T00496	C0118	P037	2024-10-24 08:30:27	1	459.86	459.86	3434.77	572.461667	95 days 16:10:05.016597	6
996	T00759	C0059	P037	2024-06-04 02:15:24	3	1379.58	459.86	7073.28	884.160000	237 days 22:25:08.016597	8
997	T00922	C0018	P037	2024-04-05 13:05:32	4	1839.44	459.86	4781.85	956.370000	297 days 11:35:00.016597	5
998	T00959	C0115	P037	2024-09-29 10:16:02	2	919.72	459.86	3137.18	1045.726667	120 days 14:24:30.016597	3
999	T00992	C0024	P037	2024-04-21 10:52:24	1	459.86	459.86	3627.02	518.145714	281 days 13:48:08.016597	7

1000 rows × 11 columns

1000 rows × 11 columns

[31]:
```
df = df_cs.merge(df_ts[['CustomerID', 'Total_Spend', 'Average_Spend', 'Recency', 'Frequency']], on='CustomerID', how='left')
df
```

[31]:

	CustomerID	CustomerName	Region	SignupDate	Total_Spend	Average_Spend	Recency	Frequency
0	C0001	Lawrence Carroll	South America	2022-07-10	3354.52	670.904	374 days 21:27:37.016597	5.0
1	C0001	Lawrence Carroll	South America	2022-07-10	3354.52	670.904	132 days 15:39:14.016597	5.0
2	C0001	Lawrence Carroll	South America	2022-07-10	3354.52	670.904	295 days 00:39:32.016597	5.0
3	C0001	Lawrence Carroll	South America	2022-07-10	3354.52	670.904	265 days 21:28:48.016597	5.0
4	C0001	Lawrence Carroll	South America	2022-07-10	3354.52	670.904	86 days 07:36:16.016597	5.0
...	...	...	...	...	...	...	...	...
996	C0200	Kelly Cross	Asia	2023-06-11	4758.60	951.720	47 days 21:34:42.016597	5.0
997	C0200	Kelly Cross	Asia	2023-06-11	4758.60	951.720	275 days 05:34:12.016597	5.0
998	C0200	Kelly Cross	Asia	2023-06-11	4758.60	951.720	196 days 04:04:04.016597	5.0
999	C0200	Kelly Cross	Asia	2023-06-11	4758.60	951.720	139 days 14:49:44.016597	5.0
1000	C0200	Kelly Cross	Asia	2023-06-11	4758.60	951.720	116 days 20:07:16.016597	5.0

1001 rows × 8 columns

[44]:
```
df_agg = df.groupby('CustomerID').agg({
    'TotalValue': 'sum',
    'Quantity': 'sum',
    'TransactionID': 'count',
    'Region': 'first'
}).reset_index()
```

[45]:
```
df_agg = pd.get_dummies(df_agg, columns=['Region'], drop_first=True)
```

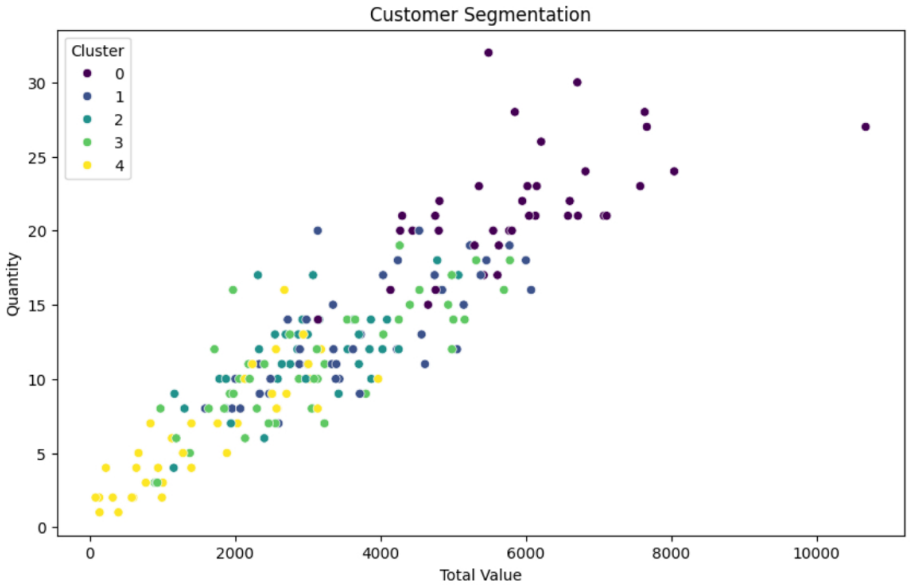
[46]:
```
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df_agg.drop(columns=['CustomerID']))
```

[47]:
```
k = 5
kmeans = KMeans(n_clusters=k, random_state=42)
clusters = kmeans.fit_predict(df_scaled)
```

[48]:
```
df_agg['Cluster'] = clusters
```

[50]:
```
from sklearn.metrics import davies_bouldin_score, silhouette_score
db_index = davies_bouldin_score(df_scaled, clusters)
silhouette_avg = silhouette_score(df_scaled, clusters)
```

[53]:
```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='TotalValue', y='Quantity', hue='Cluster', data=df_agg, palette='viridis')
plt.title('Customer Segmentation')
plt.xlabel('Total Value')
plt.ylabel('Quantity')
plt.legend(title='Cluster')
plt.show()
```



[54]:
```
print(f"Number of clusters formed: {kmeans.n_clusters}")
print(f"Davies-Bouldin Index: {db_index:.4f}")
print(f"Silhouette Score: {silhouette_avg:.4f}")
```

```
Number of clusters formed: 5
Davies-Bouldin Index: 0.9385
Silhouette Score: 0.4008
```

[ ]: