In [2]:
```
import pandas as pd
import numpy as np
customers = pd.read_csv('./data/Customers.csv')
transactions = pd.read_csv('./data/Transactions.csv')
print(customers.head())
print(transactions.head())
```

```
  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
  TransactionID CustomerID ProductID      TransactionDate  Quantity  \
0        T00001      C0199      P067  2024-08-25 12:38:23         1
1        T00112      C0146      P067  2024-05-27 22:23:54         1
2        T00166      C0127      P067  2024-04-25 07:38:55         1
3        T00272      C0087      P067  2024-03-26 22:55:37         2
4        T00363      C0070      P067  2024-03-21 15:10:10         3

   TotalValue   Price
0      300.68  300.68
1      300.68  300.68
2      300.68  300.68
3      601.36  300.68
4      902.04  300.68
```

In [3]:
```
data = transactions.merge(customers, on='CustomerID')
customer_features = data.groupby('CustomerID').agg({
    'TotalValue': 'sum',
    'Quantity': 'sum',
    'TransactionID': 'count'
}).reset_index()
customer_features = customer_features.merge(customers[['CustomerID', 'Region']],
print(customer_features.head())
```

```
  CustomerID  TotalValue  Quantity  TransactionID         Region
0      C0001     3354.52        12              5  South America
1      C0002     1862.74        10              4           Asia
2      C0003     2725.38        14              4  South America
3      C0004     5354.88        23              8  South America
4      C0005     2034.24         7              3           Asia
```

The TotalValue column shows the total spending of each customer. Customers with higher totals indicate high-value clients who are crucial for the business.Focus on retaining high-spending customers, as they contribute more to overall revenue. Consider offers to boost retention.

Some customers, like C0004, have made more frequent purchases which can indicate higher engagement.Highly engaged customers tend to make more purchases. Offering time-limited promotions or regular updates could further encourage repeat transactions.

Customers buying large quantities might be bulk purchasers or need products frequently. Offering subscription-based models or bulk discounts could appeal to these customers.

The business could explore targeted marketing campaigns in South America to capitalize on high transaction volume.

High spenders may also be the most loyal customers, making them good candidates for personalized or VIP treatment. Their frequent purchases suggest they might appreciate exclusive offers.

In [4]:
```
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
customer_features['Region'] = encoder.fit_transform(customer_features['Region'])
print(customer_features.head())
```

```
  CustomerID  TotalValue  Quantity  TransactionID  Region
0      C0001     3354.52        12              5       3
1      C0002     1862.74        10              4       0
2      C0003     2725.38        14              4       3
3      C0004     5354.88        23              8       3
4      C0005     2034.24         7              3       0
```

South American customers seem to contribute more to revenue. A deeper focus on this region could help the business increase its revenue stream by offering region-specific promotions or deals.

Customers in South America seem to buy products in bulk, suggesting a preference for bulk deals or larger quantities. Offering bundled packages or subscription models could align with their purchasing behavior.

Targeted marketing campaigns for South American customers, offering incentives like discounts or exclusive promotions, could increase transaction frequency even further.

High-spending customers appear to be concentrated in South America. This suggests a strategic focus on this region could result in significant returns, particularly through loyalty programs or region-specific deals.

Customers from Region 0 (Asia) have lower total spend and fewer transactions on average. While Asia has potential, it appears under-exploited in comparison to South America.

In [5]:
```
from sklearn.preprocessing import StandardScaler
features = customer_features[['TotalValue', 'Quantity', 'TransactionID', 'Region
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)
print(scaled_features[:5])
```

```
[[-0.06170143 -0.12203296 -0.01145819  1.23740234]
 [-0.87774353 -0.44800021 -0.46749414 -1.41989693]
 [-0.40585722  0.20393428 -0.46749414  1.23740234]
 [ 1.03254704  1.67078689  1.35664965  1.23740234]
 [-0.78392861 -0.93695108 -0.92353008 -1.41989693]]
```

High-spending customers should be targeted with premium offers, while low-spending customers could benefit from special promotions to boost their purchases

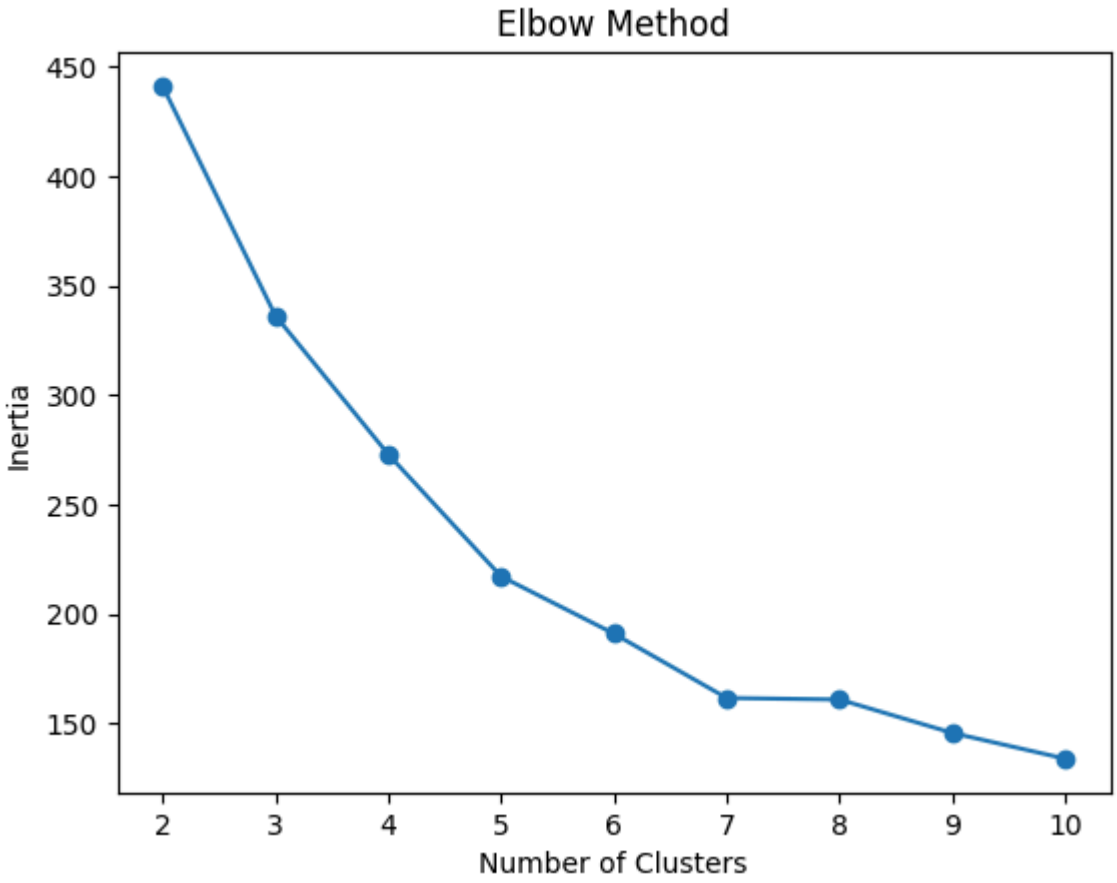
Customers with higher scaled values for Quantity tend to buy more products. These customers are purchasing in bulk.

Customers who make more transactions have a higher scaled value, showing they are frequent buyers.Regular buyers are important to retain.

South American customers are more engaged.

With all features scaled, it's easier to compare different customers and understand their purchasing patterns.

In [6]:
```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
inertia = []
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(scaled_features)
    inertia.append(kmeans.inertia_)
plt.plot(range(2, 11), inertia, marker='o')
plt.title('Elbow Method')
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia')
plt.show()
```



In [7]:
```
kmeans = KMeans(n_clusters=4, random_state=42)
customer_features['Cluster'] = kmeans.fit_predict(scaled_features)

print(customer_features.head())
```

	CustomerID	TotalValue	Quantity	TransactionID	Region	Cluster
0	C0001	3354.52	12	5	3	2
1	C0002	1862.74	10	4	0	1
2	C0003	2725.38	14	4	3	2
3	C0004	5354.88	23	8	3	0
4	C0005	2034.24	7	3	0	1

Customers have been grouped into 4 distinct clusters. Some clusters show high spending (e.g., Cluster 0 and Cluster 2), while others show lower spending patterns.

Clusters with higher Quantity purchased like Cluster 2 may indicate bulk buyers who prefer larger quantities.

Clusters with more frequent transactions like Cluster 0 represent highly engaged customers who consistently make purchases.

The Region variable may show variation between clusters. Customers in Cluster 0 or Cluster 2 might represent regions with stronger engagement, such as South America.

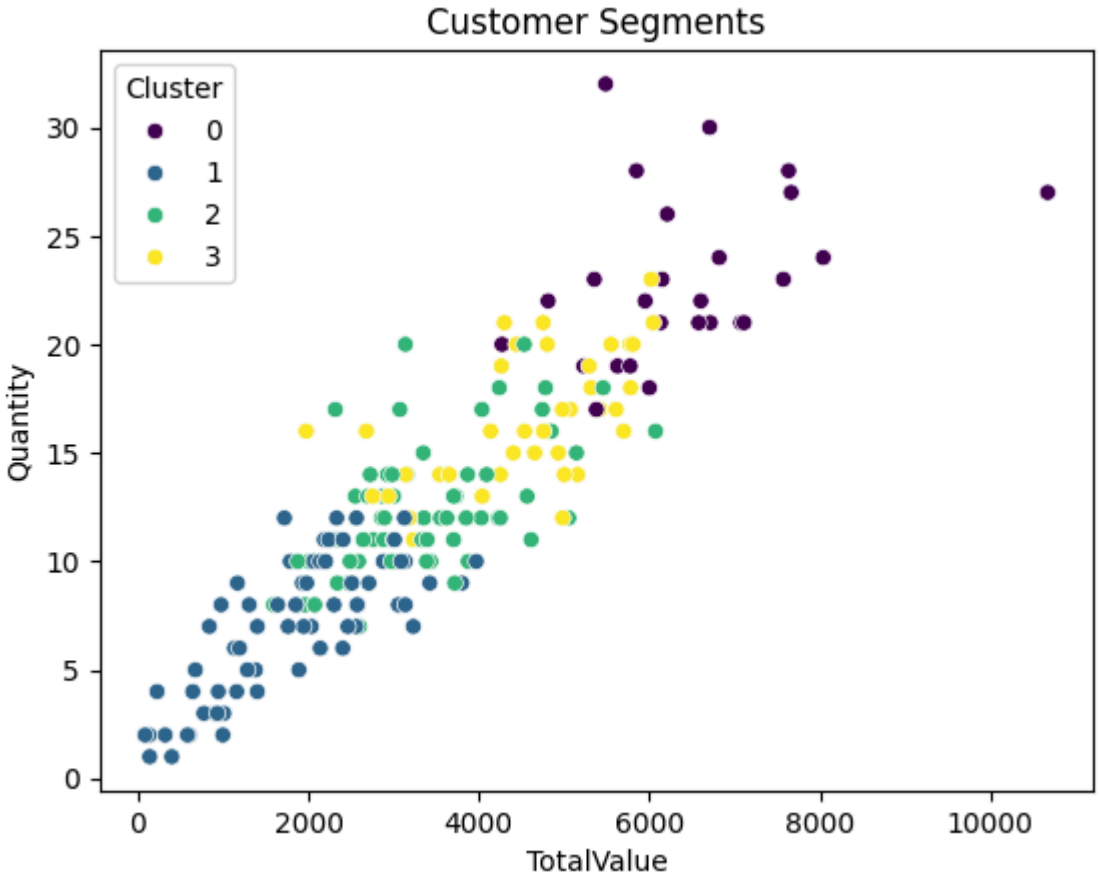
Different clusters likely have unique preferences regarding transaction size, frequency, and spending.

In [8]:
```
from sklearn.metrics import davies_bouldin_score
db_index = davies_bouldin_score(scaled_features, customer_features['Cluster'])
print(f"Davis-Bouldin Index: {db_index}")
```

```
Davis-Bouldin Index: 1.0259136043470107
```

The company should consider refining the clustering process like adjusting the number of clusters or using different clustering techniques to achieve even more distinct customer segments.

In [9]:
```
import seaborn as sns
sns.scatterplot(
    x=customer_features['TotalValue'],
    y=customer_features['Quantity'],
    hue=customer_features['Cluster'],
    palette='viridis'
)
plt.title('Customer Segments')
plt.show()
```



The analysis helps identify distinct customer groups for targeted marketing and better product recommendations. It also provides insights to improve customer engagement, loyalty, and sales strategies.