In [1]:
```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
```

In [2]:
```
# --- Load the dataset with proper encoding ---
df = pd.read_csv("sales_data_sample.csv", encoding='latin1')

print("✅ Dataset loaded successfully!")
print("Shape:", df.shape)
print(df.head())
```

```
✅ Dataset loaded successfully!
Shape: (2823, 25)
   ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER    SALES  \
0        10107               30      95.70                2  2871.00
1        10121               34      81.35                5  2765.90
2        10134               41      94.74                2  3884.34
3        10145               45      83.26                6  3746.70
4        10159               49     100.00               14  5205.27

          ORDERDATE   STATUS  QTR_ID  MONTH_ID  YEAR_ID  ...  \
0   2/24/2003 0:00  Shipped       1         2     2003  ...
1    5/7/2003 0:00  Shipped       2         5     2003  ...
2    7/1/2003 0:00  Shipped       3         7     2003  ...
3   8/25/2003 0:00  Shipped       3         8     2003  ...
4  10/10/2003 0:00  Shipped       4        10     2003  ...

                    ADDRESSLINE1  ADDRESSLINE2           CITY STATE  \
0         897 Long Airport Avenue           NaN            NYC    NY
1              59 rue de l'Abbaye           NaN          Reims   NaN
2  27 rue du Colonel Pierre Avia           NaN          Paris   NaN
3              78934 Hillside Dr.           NaN       Pasadena    CA
4                 7734 Strong St.           NaN  San Francisco    CA

  POSTALCODE COUNTRY TERRITORY CONTACTLASTNAME CONTACTFIRSTNAME DEALSIZE
0      10022     USA       NaN              Yu              Kwai    Small
1      51100  France      EMEA         Henriot              Paul    Small
2      75508  France      EMEA        Da Cunha            Daniel   Medium
3      90003     USA       NaN           Young             Julie   Medium
4        NaN     USA       NaN           Brown             Julie   Medium

[5 rows x 25 columns]
```

In [3]:
```
# --- Automatically select all numerical columns for clustering ---
numerical_cols = df.select_dtypes(include=['int64', 'float64']).columns.tolist()
print("\nNumerical columns selected for clustering:", numerical_cols)
```

```
Numerical columns selected for clustering: ['ORDERNUMBER', 'QUANTITYORDERED', 'PR
ICEEACH', 'ORDERLINENUMBER', 'SALES', 'QTR_ID', 'MONTH_ID', 'YEAR_ID', 'MSRP']
```

In [4]:
```
# --- Preprocessing: Scale numerical features ---
X = df[numerical_cols].values
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

In [5]:
```
# --- Elbow Method to determine optimal K ---
wcss = []
K_range = range(1, 11)
```

```
for k in K_range:
    kmeans = KMeans(n_clusters=k, init='k-means++', random_state=42, n_init='aut
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_)
```

In [6]:
```
# --- K-Means Clustering (Choose optimal K, e.g., K=3) ---
optimal_k = 3
kmeans_model = KMeans(n_clusters=optimal_k, init='k-means++', random_state=42, n
df['Cluster'] = kmeans_model.fit_predict(X_scaled)
```

In [7]:
```
# --- Clean Output ---
print("\n--- K-Means Clustering Implementation ---")
print(f"Optimal Number of Clusters (Determined by Elbow Method): {optimal_k}\n")
print("First 5 rows with assigned cluster labels:")
print(df.head())
print(f"\nWCSS for K={optimal_k}: {wcss[optimal_k-1]:.2f}")
print("\nAnalysis: The elbow method suggests the optimal K is at the point where
print("Points of Improvement: Try feature selection, scaling methods, different
```

```
--- K-Means Clustering Implementation ---
Optimal Number of Clusters (Determined by Elbow Method): 3

First 5 rows with assigned cluster labels:
   ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER    SALES  \
0        10107               30      95.70               2  2871.00
1        10121               34      81.35               5  2765.90
2        10134               41      94.74               2  3884.34
3        10145               45      83.26               6  3746.70
4        10159               49     100.00              14  5205.27

         ORDERDATE   STATUS  QTR_ID  MONTH_ID  YEAR_ID  ... ADDRESSLINE2  \
0   2/24/2003 0:00  Shipped       1         2     2003  ...          NaN
1    5/7/2003 0:00  Shipped       2         5     2003  ...          NaN
2    7/1/2003 0:00  Shipped       3         7     2003  ...          NaN
3   8/25/2003 0:00  Shipped       3         8     2003  ...          NaN
4  10/10/2003 0:00  Shipped       4        10     2003  ...          NaN

            CITY STATE POSTALCODE COUNTRY TERRITORY CONTACTLASTNAME  \
0            NYC    NY      10022     USA       NaN             Yu
1          Reims   NaN      51100  France      EMEA        Henriot
2          Paris   NaN      75508  France      EMEA       Da Cunha
3       Pasadena    CA      90003     USA       NaN          Young
4  San Francisco    CA        NaN     USA       NaN          Brown

  CONTACTFIRSTNAME DEALSIZE Cluster
0             Kwai    Small       1
1             Paul    Small       2
2           Daniel   Medium       1
3            Julie   Medium       1
4            Julie   Medium       1

[5 rows x 26 columns]

WCSS for K=3: 16909.36

Analysis: The elbow method suggests the optimal K is at the point where the WCSS
curve bends sharply, indicating distinct clusters in the data.
Points of Improvement: Try feature selection, scaling methods, different K value
s, or clustering algorithms like DBSCAN or hierarchical clustering.
```

In [ ]: