In [1]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```

In [5]:
```
df = pd.read_csv(r"C:\Users\pl\Downloads\sales_data_sample.csv",encoding='latin1'
```

In [6]:
```
df.head()
```

Out[6]:

| | ORDERNUMBER | QUANTITYORDERED | PRICEEACH | ORDERLINENUMBER | SALES | ORDERDATE |
|---|---|---|---|---|---|---|
| **0** | 10107 | 30 | 95.70 | 2 | 2871.00 | 2/24/2003 0:00 |
| **1** | 10121 | 34 | 81.35 | 5 | 2765.90 | 5/7/2003 0:00 |
| **2** | 10134 | 41 | 94.74 | 2 | 3884.34 | 7/1/2003 0:00 |
| **3** | 10145 | 45 | 83.26 | 6 | 3746.70 | 8/25/2003 0:00 |
| **4** | 10159 | 49 | 100.00 | 14 | 5205.27 | 10/10/2003 0:00 |

5 rows × 25 columns

In [7]:
```
df.describe()
```

Out[7]:

| | ORDERNUMBER | QUANTITYORDERED | PRICEEACH | ORDERLINENUMBER | SALES | |
|---|---|---|---|---|---|---|
| **count** | 2823.000000 | 2823.000000 | 2823.000000 | 2823.000000 | 2823.000000 | 282 |
| **mean** | 10258.725115 | 35.092809 | 83.658544 | 6.466171 | 3553.889072 | |
| **std** | 92.085478 | 9.741443 | 20.174277 | 4.225841 | 1841.865106 | |
| **min** | 10100.000000 | 6.000000 | 26.880000 | 1.000000 | 482.130000 | |
| **25%** | 10180.000000 | 27.000000 | 68.860000 | 3.000000 | 2203.430000 | |
| **50%** | 10262.000000 | 35.000000 | 95.700000 | 6.000000 | 3184.800000 | |
| **75%** | 10333.500000 | 43.000000 | 100.000000 | 9.000000 | 4508.000000 | |
| **max** | 10425.000000 | 97.000000 | 100.000000 | 18.000000 | 14082.800000 | |

In [8]:
```
df.shape
```

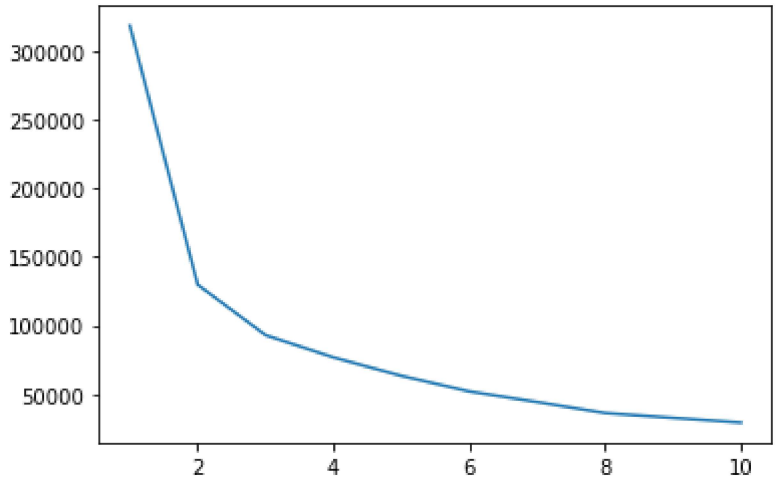
Out[8]: (2823, 25)

In [9]:
```
df = df[['QUANTITYORDERED', 'ORDERLINENUMBER']]
df = df.dropna(axis = 0)
```

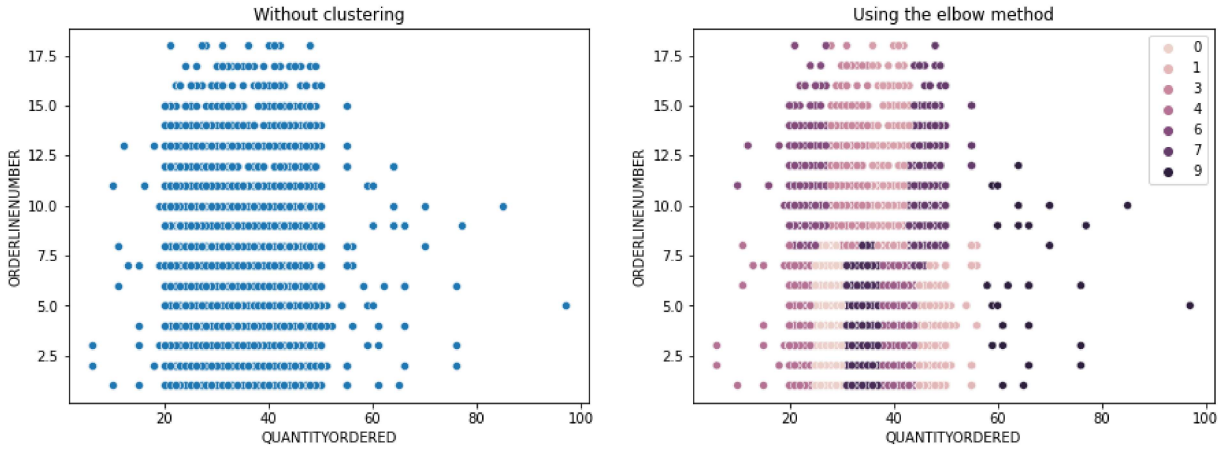
In [10]:
```
wcss = []

for i in range(1, 11):
    clustering = KMeans(n_clusters=i, init='k-means++', random_state=42)
    clustering.fit(df)
    wcss.append(clustering.inertia_)

ks = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
sns.lineplot(x = ks, y = wcss);
```




In [11]:
```
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(15,5))
sns.scatterplot(ax=axes[0], data=df, x='QUANTITYORDERED', y='ORDERLINENUMBER').se
sns.scatterplot(ax=axes[1], data=df, x='QUANTITYORDERED', y='ORDERLINENUMBER', hu
```



In [12]:
```
df.describe().T
```

Out[12]:

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **QUANTITYORDERED** | 2823.0 | 35.092809 | 9.741443 | 6.0 | 27.0 | 35.0 | 43.0 | 97.0 |
| **ORDERLINENUMBER** | 2823.0 | 6.466171 | 4.225841 | 1.0 | 3.0 | 6.0 | 9.0 | 18.0 |

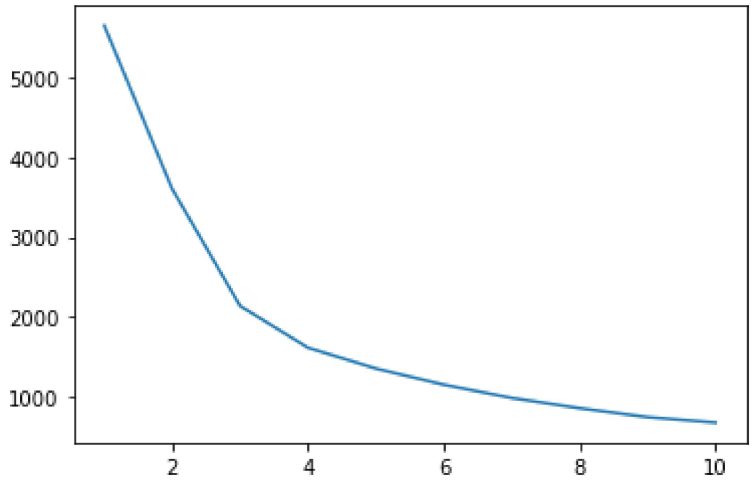
In [13]:
```
from sklearn.preprocessing import StandardScaler

ss = StandardScaler()
scaled = ss.fit_transform(df)
```

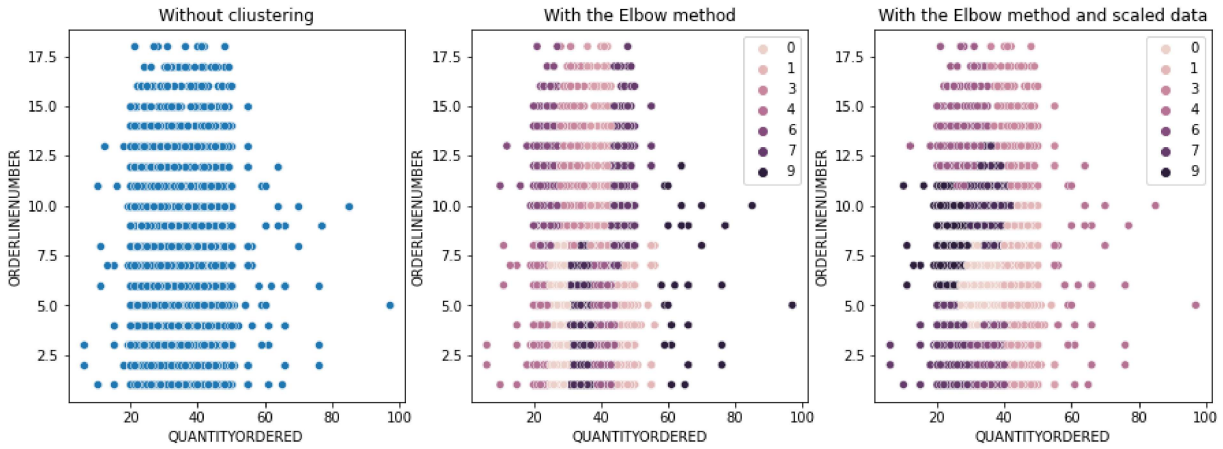
In [14]:
```
wcss_sc = []

for i in range(1, 11):
    clustering_sc = KMeans(n_clusters=i, init='k-means++', random_state=42)
    clustering_sc.fit(scaled)
    wcss_sc.append(clustering_sc.inertia_)

ks = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
sns.lineplot(x = ks, y = wcss_sc);
```



In [15]:
```
fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(15,5))
sns.scatterplot(ax=axes[0], data=df, x='QUANTITYORDERED', y='ORDERLINENUMBER').se
sns.scatterplot(ax=axes[1], data=df, x='QUANTITYORDERED', y='ORDERLINENUMBER', hu
sns.scatterplot(ax=axes[2], data=df, x='QUANTITYORDERED', y='ORDERLINENUMBER', hu
```




In [ ]: