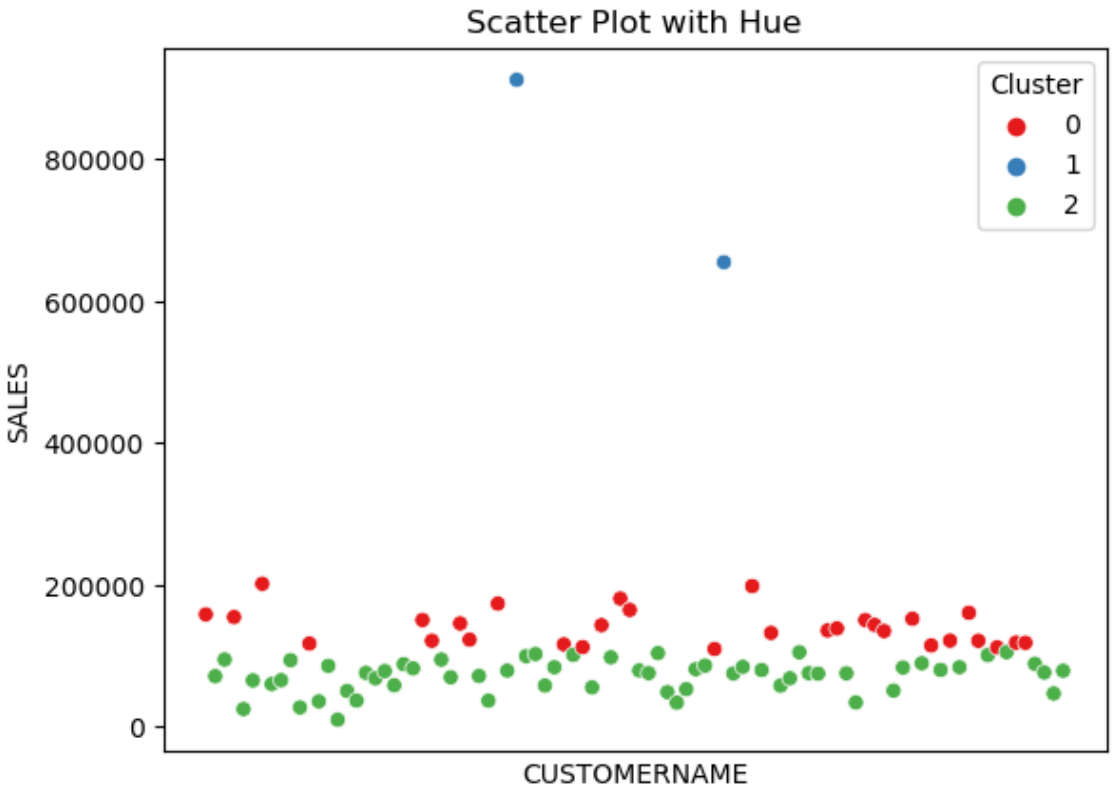
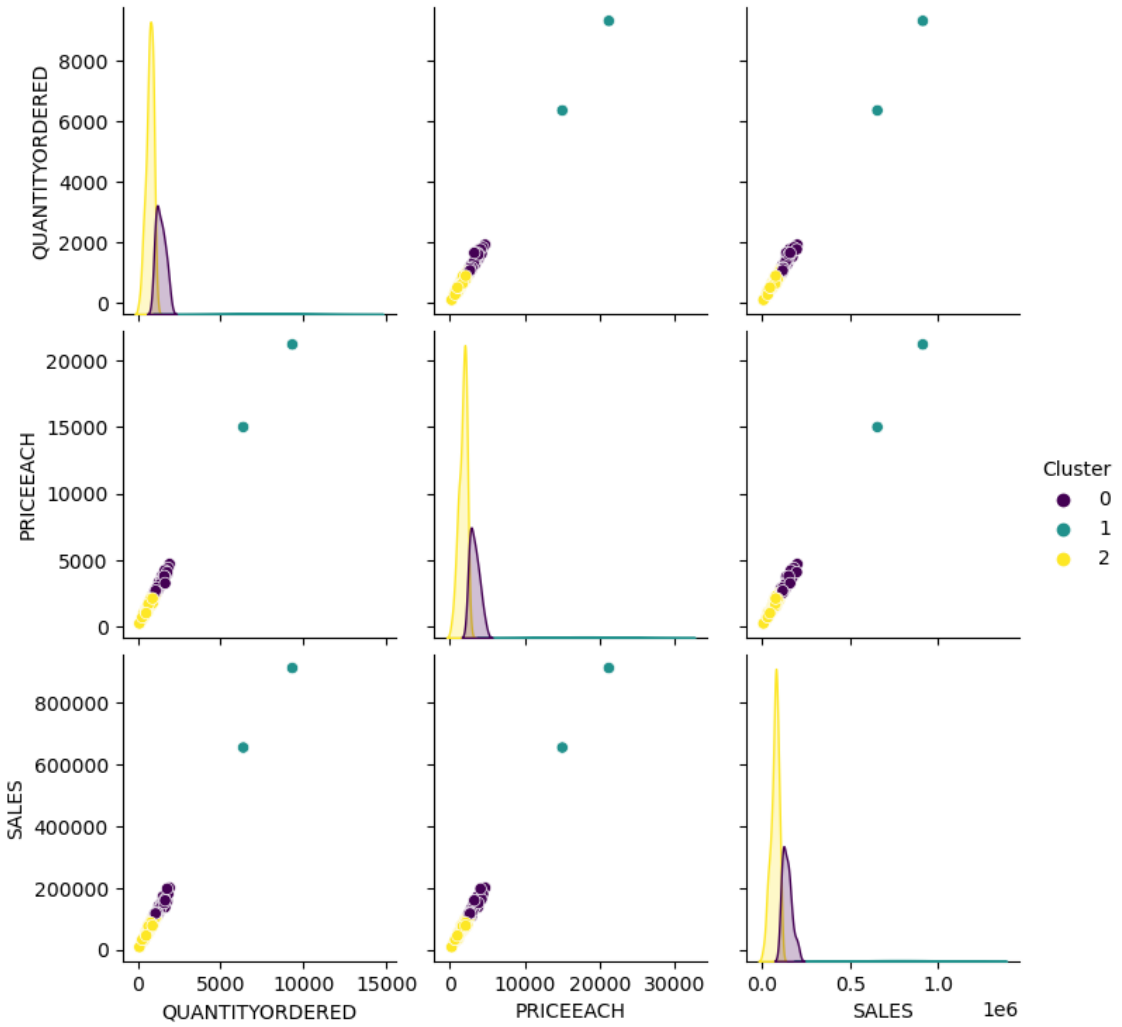
In [20]:

```
import seaborn as sns

# Scatter plot with hue
sns.scatterplot(x='CUSTOMERNAME', y='SALES', hue='Cluster',palette='Set1',
# Remove X-axis labels
plt.xticks([])
plt.title('Scatter Plot with Hue')
plt.show()
```



In [21]:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Assuming 'hue_column' is the column you want to use as the hue
sns.pairplot(df, hue='Cluster', palette='viridis')
plt.show()
```



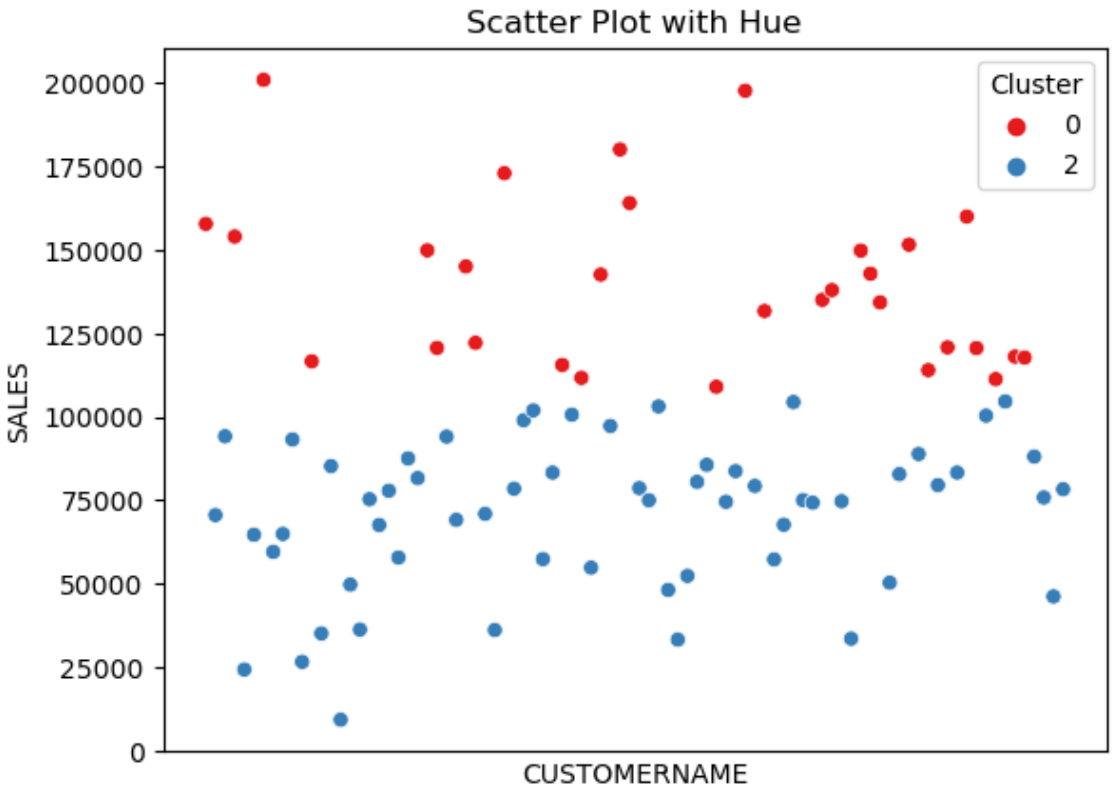
FILTER THE OUTLIERS

In [22]:

```
# Remove rows with 'Cluster' column value equal to -1
df = df[df['Cluster'] != 1]
```

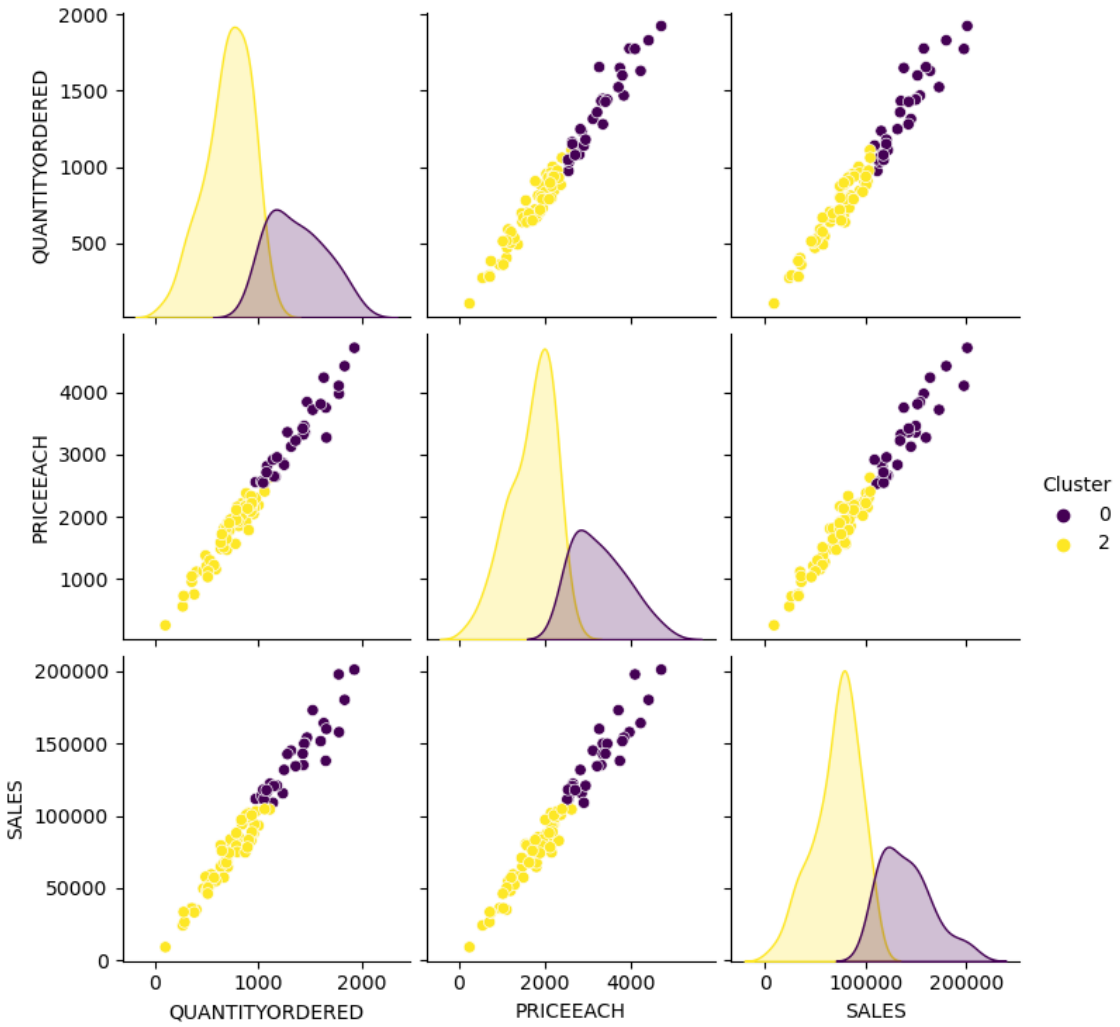
In [23]:

```
import seaborn as sns

# Scatter plot with hue
sns.scatterplot(x='CUSTOMERNAME', y='SALES', hue='Cluster',palette='Set1',
# Remove X-axis labels
plt.xticks([])
plt.title('Scatter Plot with Hue')
plt.show()
```



In [24]:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Assuming 'hue_column' is the column you want to use as the hue
sns.pairplot(df, hue='Cluster', palette='viridis')
plt.show()
```



ONCE THE OUTLIERS ARE OUT RE DO THE EXPERIMENT

In [25]:
```
df
```

Out[25]:

	QUANTITYORDERED	PRICEEACH	SALES	Cluster	CUSTOMERNAME
0	1778	3975.33	157807.81	0	AV Stores, Co.
1	687	1701.95	70488.44	2	Alpha Cognac
2	843	2218.41	94117.26	2	Amica Models & Co.
3	1469	3843.67	153996.13	0	Anna's Decorations, Ltd
4	270	558.43	24179.96	2	Atelier graphique
...	...	...	...	...	...
87	1078	2713.09	117713.56	0	Vida Sport, Ltd
88	787	2108.11	88041.26	2	Vitachrome Inc.
89	647	1720.14	75754.88	2	Volvo Model Replicas, Co
90	511	1030.99	46084.64	2	West Coast Collectables Co.
91	895	2131.78	78240.84	2	giftsbymail.co.uk

90 rows × 5 columns

In [26]:
```
# Drop columns PRODUCTCODE bc it has the info as PRODUCTLINE

columns_to_drop = ['Cluster']
df = df.drop(columns=columns_to_drop)
df
```

Out[26]:

	QUANTITYORDERED	PRICEEACH	SALES	CUSTOMERNAME
0	1778	3975.33	157807.81	AV Stores, Co.
1	687	1701.95	70488.44	Alpha Cognac
2	843	2218.41	94117.26	Amica Models & Co.
3	1469	3843.67	153996.13	Anna's Decorations, Ltd
4	270	558.43	24179.96	Atelier graphique
...	...	...	...	...
87	1078	2713.09	117713.56	Vida Sport, Ltd
88	787	2108.11	88041.26	Vitachrome Inc.
89	647	1720.14	75754.88	Volvo Model Replicas, Co
90	511	1030.99	46084.64	West Coast Collectables Co.
91	895	2131.78	78240.84	giftsbymail.co.uk

90 rows × 4 columns

In [27]:
```
df_result = df.copy()
```

In [28]:
```
columns_to_drop = ['CUSTOMERNAME']
df = df.drop(columns=columns_to_drop)
```

In [29]:
```
df
```

Out[29]:

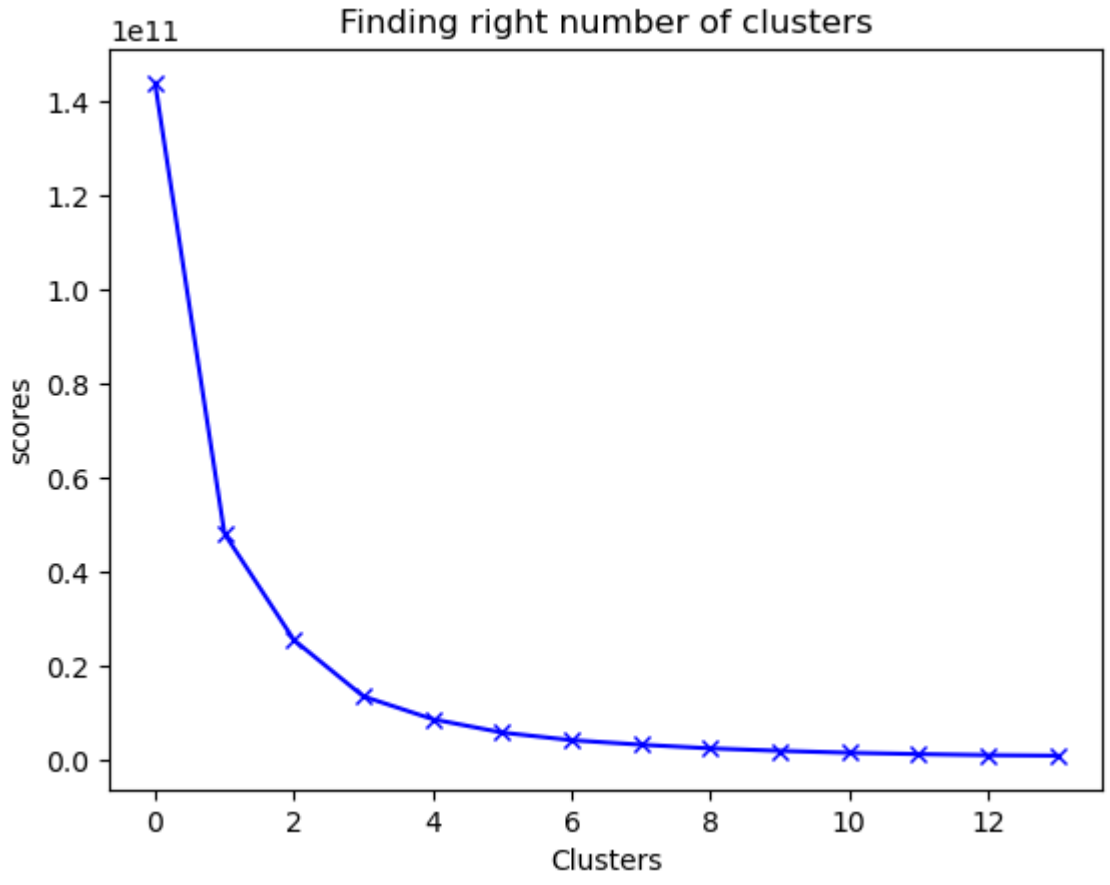
	QUANTITYORDERED	PRICEEACH	SALES
0	1778	3975.33	157807.81
1	687	1701.95	70488.44
2	843	2218.41	94117.26
3	1469	3843.67	153996.13
4	270	558.43	24179.96
...	...	...	...
87	1078	2713.09	117713.56
88	787	2108.11	88041.26
89	647	1720.14	75754.88
90	511	1030.99	46084.64
91	895	2131.78	78240.84

90 rows × 3 columns

FIND THE OPTIMAL NUMBER OF CLUSTERS USING ELBOW METHOD

In [30]:

```
import warnings
warnings.filterwarnings("ignore")

scores = []

range_values = range(1, 15)

for i in range_values:
  kmeans = KMeans(n_clusters = i)
  kmeans.fit(df)
  scores.append(kmeans.inertia_) # intertia is the Sum of squared distances

plt.plot(scores, 'bx-')
plt.title('Finding right number of clusters')
plt.xlabel('Clusters')
plt.ylabel('scores')
plt.show()

# Kmeans details in Sklearn: https://scikit-learn.org/stable/modules/genera
# From this we can observe that, 5th cluster seems to be forming the elbow
# Note that curve will change everytime we run the cell
```



APPLY K-MEANS METHOD

In [31]:
```
# Specify the number of clusters (you need to choose this based on your prc
n_clusters = 3

# Extract numerical columns from the DataFrame
X = df.values

# Apply K-Means clustering
kmeans = KMeans(n_clusters=n_clusters, random_state=42)
df['Cluster'] = kmeans.fit_predict(X)
```

In [32]:
```
# Add 'Value_df2' column from df2 to df1
df['CUSTOMERNAME'] = df_result['CUSTOMERNAME']
```

In [33]:
```
df
```

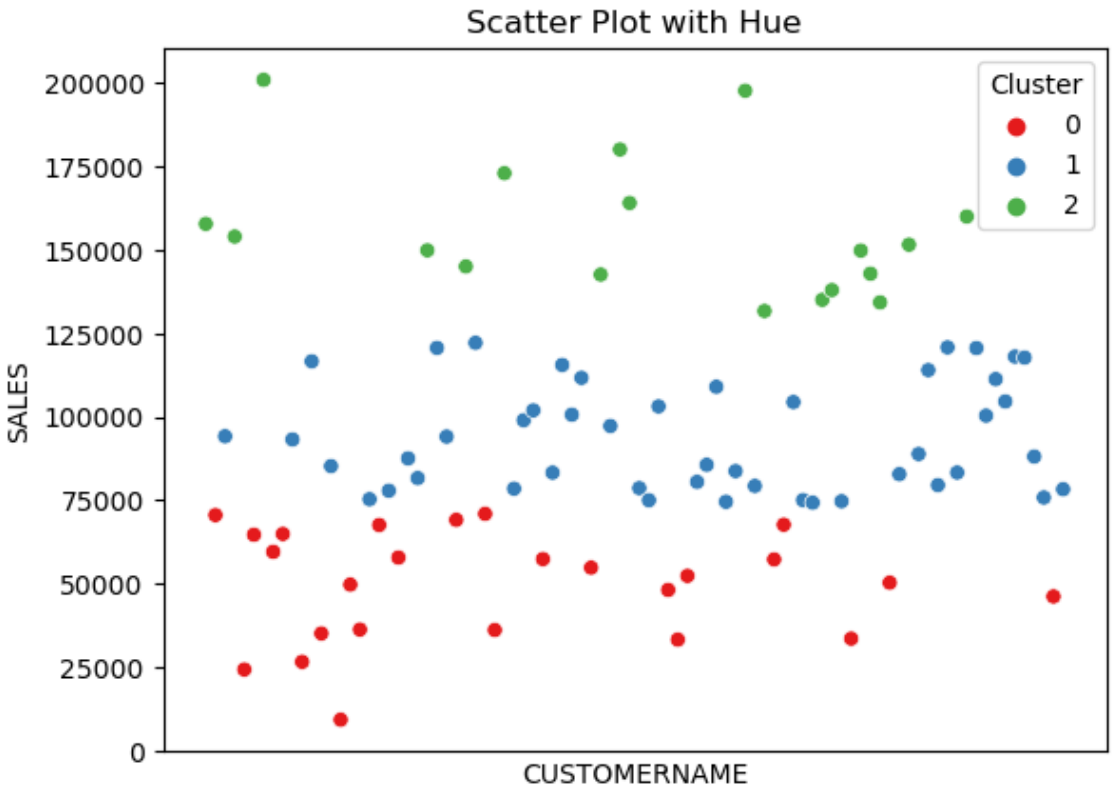
Out[33]:

	QUANTITYORDERED	PRICEEACH	SALES	Cluster	CUSTOMERNAME
0	1778	3975.33	157807.81	2	AV Stores, Co.
1	687	1701.95	70488.44	0	Alpha Cognac
2	843	2218.41	94117.26	1	Amica Models & Co.
3	1469	3843.67	153996.13	2	Anna's Decorations, Ltd
4	270	558.43	24179.96	0	Atelier graphique
...	...	...	...	...	...
87	1078	2713.09	117713.56	1	Vida Sport, Ltd
88	787	2108.11	88041.26	1	Vitachrome Inc.
89	647	1720.14	75754.88	1	Volvo Model Replicas, Co
90	511	1030.99	46084.64	0	West Coast Collectables Co.
91	895	2131.78	78240.84	1	giftsbymail.co.uk

90 rows × 5 columns

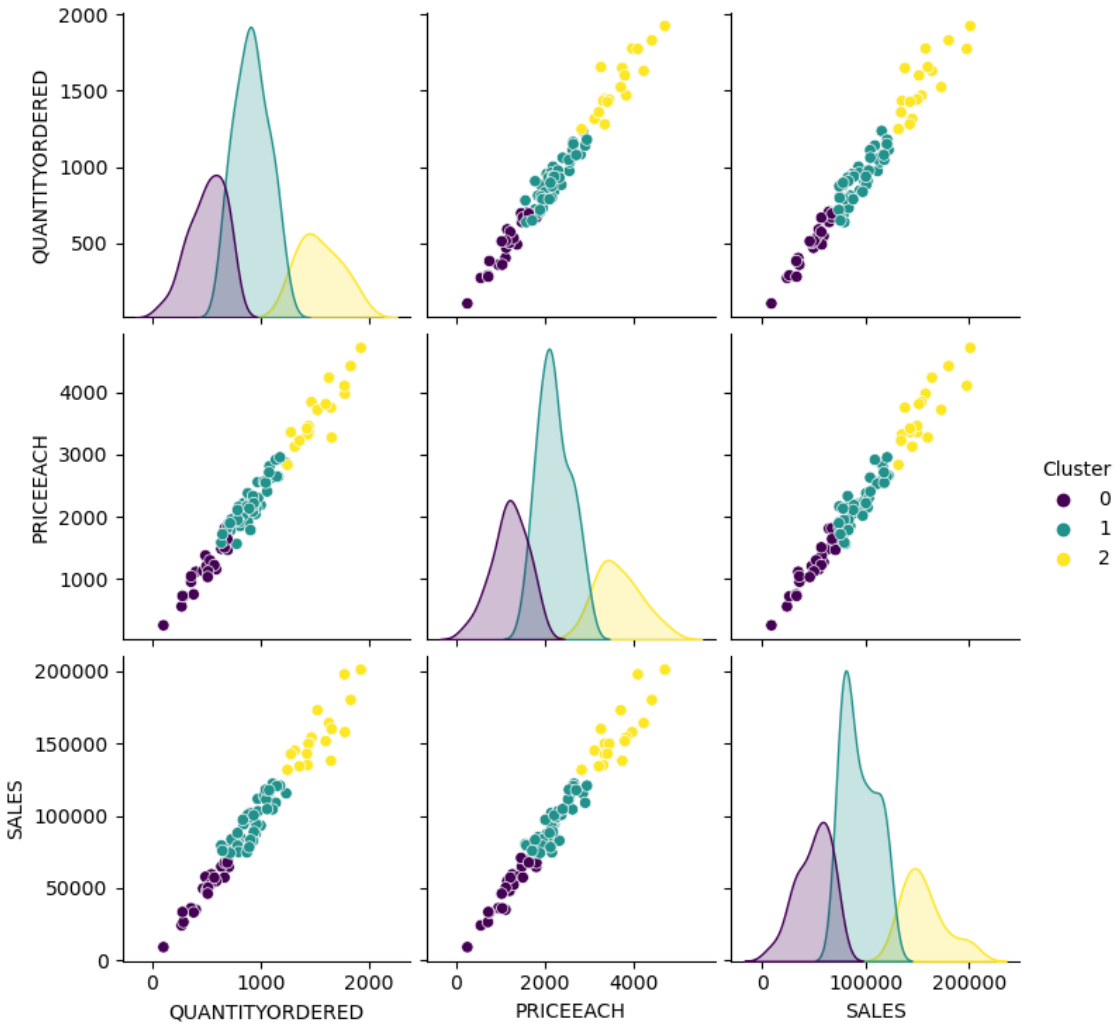
In [34]:

```
import seaborn as sns

# Scatter plot with hue
sns.scatterplot(x='CUSTOMERNAME', y='SALES', hue='Cluster',palette='Set1',
# Remove X-axis labels
plt.xticks([])
plt.title('Scatter Plot with Hue')
plt.show()
```



In [35]:

```
import seaborn as sns
import matplotlib.pyplot as plt

# Assuming 'hue_column' is the column you want to use as the hue
sns.pairplot(df, hue='Cluster', palette='viridis')
plt.show()
```



In [ ]: