Exploratory Data Analysis of Superstore

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

In [3]:
```
dataset= pd.read_excel("Copy of Superstore_USA.xlsx")
```

In [4]:
```
dataset.head(2)
```

Out[4]:

	Row ID	Order Priority	Discount	Unit Price	Shipping Cost	Customer ID	Customer Name	Ship Mode	Customer Segment
0	18606	Not Specified	0.01	2.88	0.50	2	Janice Fletcher	Regular Air	Corporate
1	20847	High	0.01	2.84	0.93	3	Bonnie Potter	Express Air	Corporate

2 rows × 24 columns

In [5]:
```
dataset.shape
```

Out[5]:
```
(9426, 24)
```

In [6]:
```
dataset.isnull().sum()
```

Out[6]:
```
Row ID                    0
Order Priority            0
Discount                  0
Unit Price                0
Shipping Cost             0
Customer ID               0
Customer Name             0
Ship Mode                 0
Customer Segment          0
Product Category          0
Product Sub-Category      0
Product Container         0
Product Name              0
Product Base Margin      72
Region                    0
State or Province         0
City                      0
Postal Code               0
Order Date                0
Ship Date                 0
Profit                    0
Quantity ordered new      0
Sales                     0
Order ID                  0
dtype: int64
```

In [7]:
```
dataset['Product Base Margin'].fillna(dataset['Product Base Margin'].mean(), inp
```

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_12612\4071489688.py:1: FutureWarning:
A value is trying to be set on a copy of a DataFrame or Series through chained as
signment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work becau
se the intermediate object on which we are setting values always behaves as a cop
y.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.meth
od({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to pe
rform the operation inplace on the original object.


  dataset['Product Base Margin'].fillna(dataset['Product Base Margin'].mean(), in
place = True)
```

Order Date

In [8]:
```
dataset.info()
```

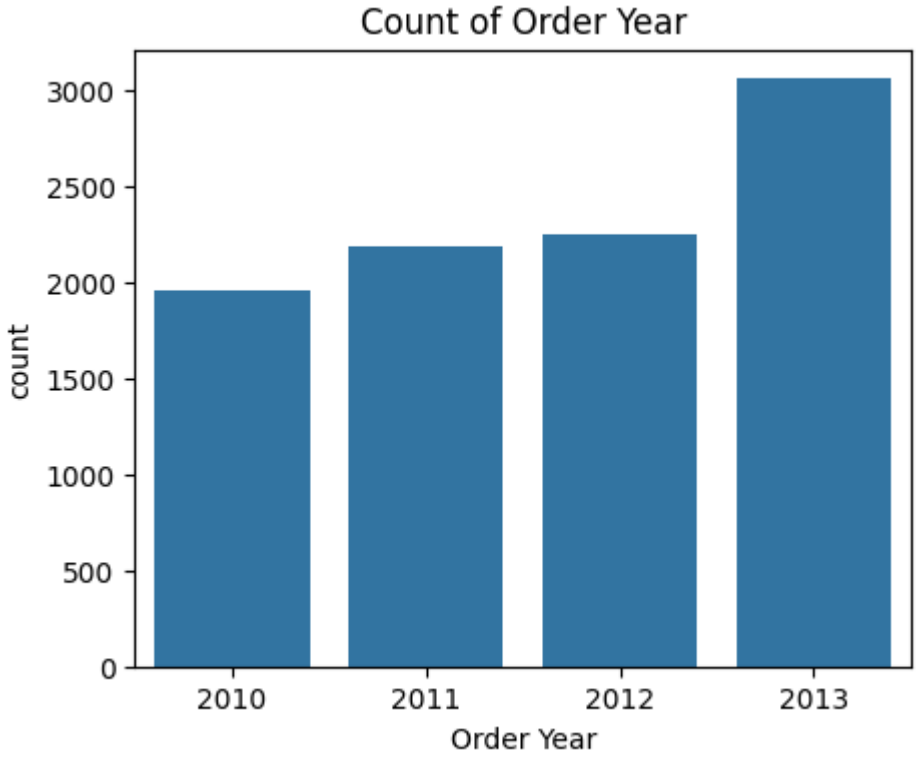
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9426 entries, 0 to 9425
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Row ID                9426 non-null   int64
 1   Order Priority        9426 non-null   object
 2   Discount              9426 non-null   float64
 3   Unit Price            9426 non-null   float64
 4   Shipping Cost         9426 non-null   float64
 5   Customer ID           9426 non-null   int64
 6   Customer Name         9426 non-null   object
 7   Ship Mode             9426 non-null   object
 8   Customer Segment      9426 non-null   object
 9   Product Category      9426 non-null   object
 10  Product Sub-Category  9426 non-null   object
 11  Product Container     9426 non-null   object
 12  Product Name          9426 non-null   object
 13  Product Base Margin   9426 non-null   float64
 14  Region                9426 non-null   object
 15  State or Province     9426 non-null   object
 16  City                  9426 non-null   object
 17  Postal Code           9426 non-null   int64
 18  Order Date            9426 non-null   datetime64[ns]
 19  Ship Date             9426 non-null   datetime64[ns]
 20  Profit                9426 non-null   float64
 21  Quantity ordered new  9426 non-null   int64
 22  Sales                 9426 non-null   float64
 23  Order ID              9426 non-null   int64
dtypes: datetime64[ns](2), float64(6), int64(5), object(11)
memory usage: 1.7+ MB
```

In [9]:
```
dataset["Order Year"] = dataset['Order Date'].dt.year
```

In [10]:
```
dataset["Order Year"].value_counts()
```

Out[10]:
```
Order Year
2013    3054
2012    2241
2011    2179
2010    1952
Name: count, dtype: int64
```

In [11]:
```
plt.figure(figsize= (5,4))
sns.countplot(x="Order Year", data=dataset)
plt.title("Count of Order Year")
plt.show()
```



'''Insights:- The sales data from 2010 to 2013 shows a clear upward trend. In 2010, the total sales were 1,952 units, which increased to 2,179 units in 2011. The growth continued in 2012 with 2,241 units sold, culminating in a significant rise to 3,054 units in 2013. This consistent increase indicates successful business strategies, potential market expansion, or improved product demand. Detailed analysis of the factors contributing to this growth could provide valuable insights for further strategic planning.'''

Order Priority

In [20]:
```
dataset['Order Priority'].value_counts()
```

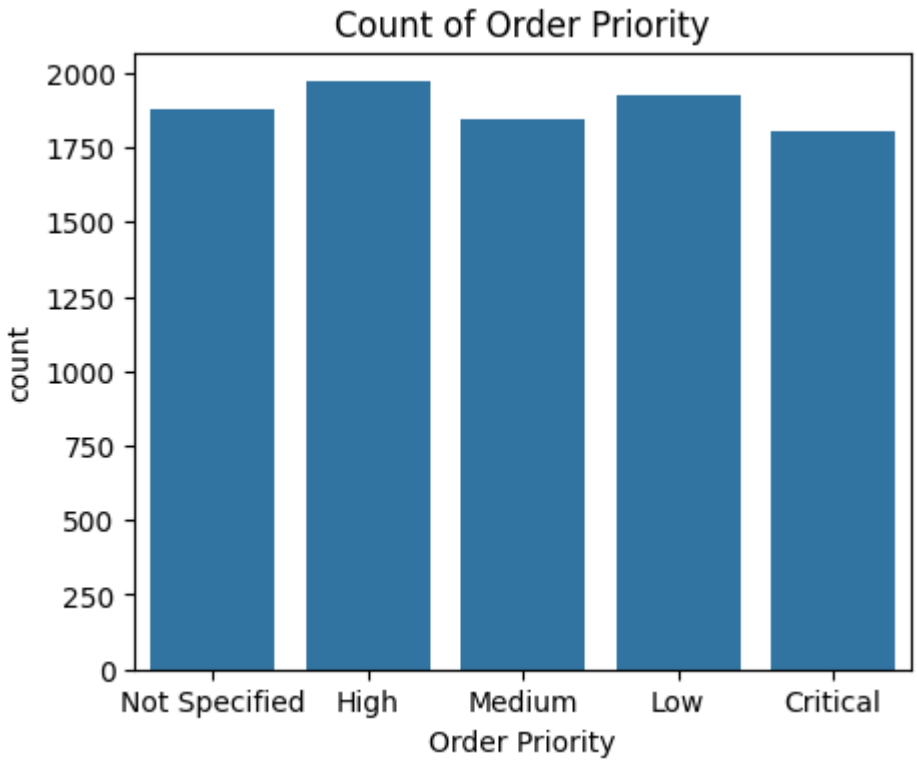
Out[20]:
```
Order Priority
High             1970
Low              1926
Not Specified    1881
Medium           1844
Critical         1805
Name: count, dtype: int64
```

In [15]:
```
dataset['Order Priority'].unique()
```

Out[15]:
```
array(['Not Specified', 'High', 'Medium', 'Low', 'Critical', 'Critical '],
      dtype=object)
```

In [19]:
```
dataset["Order Priority"]=  dataset["Order Priority"].replace("Critical ", "Crit
```

In [27]:
```
plt.figure(figsize= (5,4))
sns.countplot(x="Order Priority", data=dataset)
plt.title("Count of Order Priority")
plt.savefig("Count of Order Priority.jpg")
plt.show()
```



'''Insights:- The analysis of order priorities reveals a balanced distribution among different priority levels. 'High' priority orders lead with 1,970 units, followed closely by 'Low' priority orders at 1,926 units. 'Not Specified' and 'Medium' priorities are nearly equal, with 1,881 and 1,844 units respectively. 'Critical' priority orders, while still significant, account for 1,805 units. This distribution suggests a diverse range of customer urgency and the need for a flexible approach to handling orders across various priority levels. Further investigation could help optimize resource allocation and improve service efficiency.'''

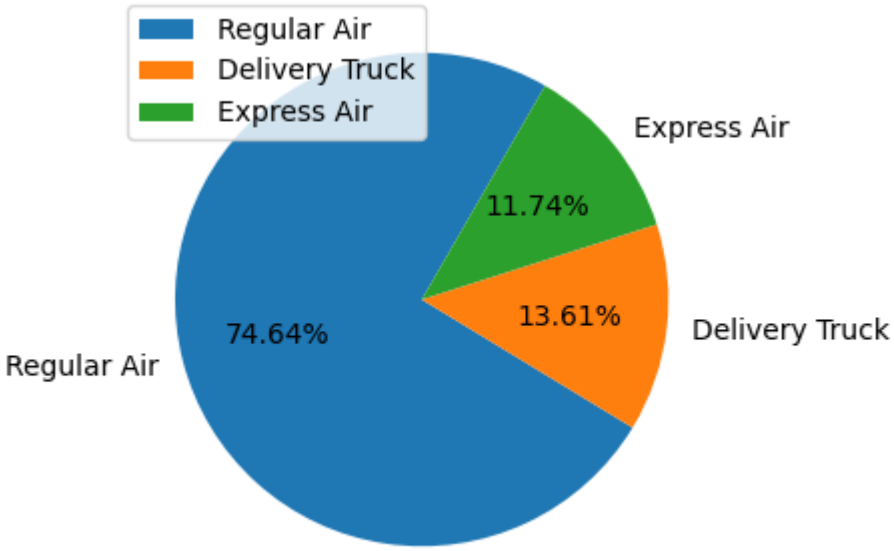
Ship Mode

In [29]:
```
dataset['Ship Mode'].value_counts()
```

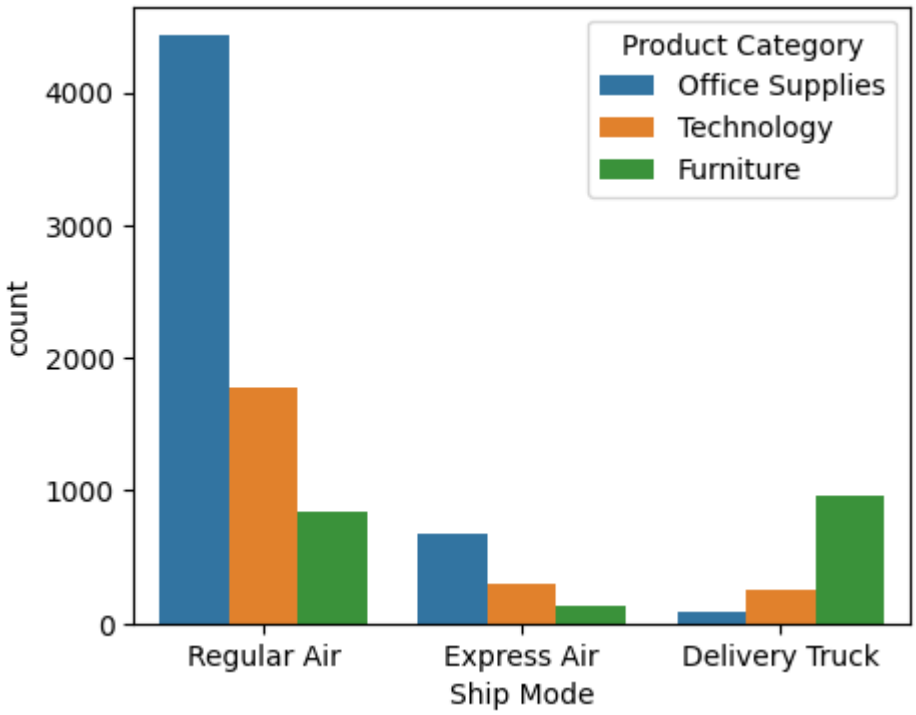
Out[29]:
```
Ship Mode
Regular Air       7036
Delivery Truck    1283
Express Air       1107
Name: count, dtype: int64
```

In [34]:
```
x= dataset['Ship Mode'].value_counts().index
y= dataset['Ship Mode'].value_counts().values
```

In [44]:
```
plt.figure(figsize= (5,4))
plt.pie(y, labels= x, startangle= 60, autopct = "%0.2f%%")
plt.legend(loc= 2)
plt.show()
```



In [47]:
```
#Bivariate Analysis
plt.figure(figsize= (5,4))
sns.countplot(x= "Ship Mode", data= dataset, hue= "Product Category")
plt.show()
```



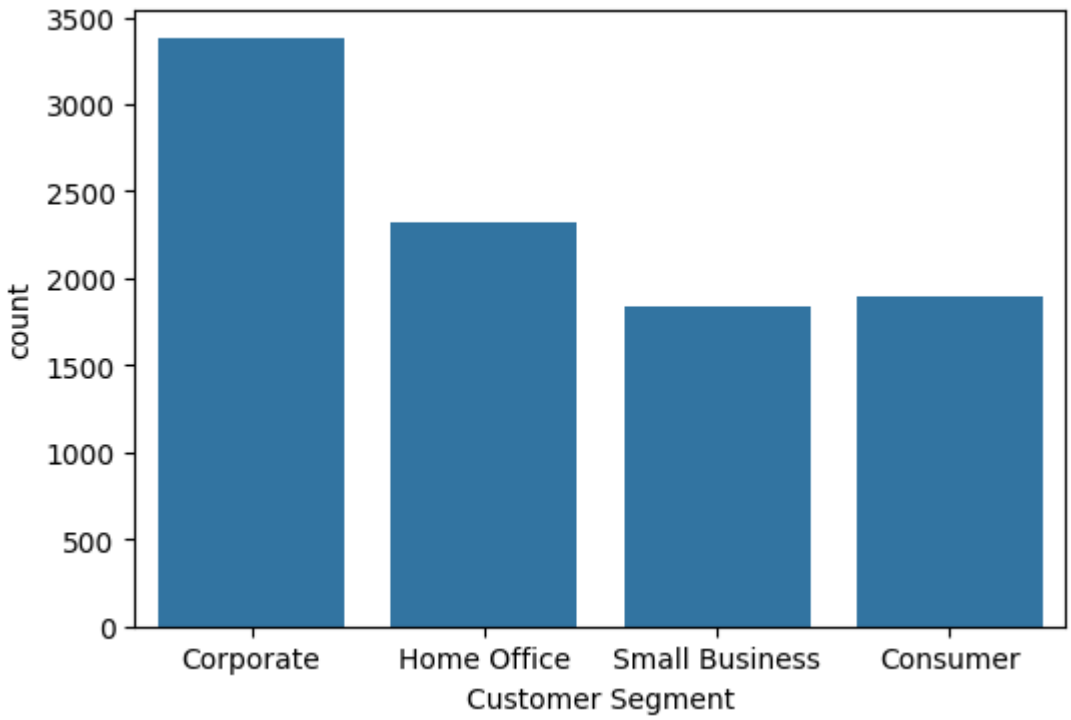
'''Insights:- The shipping mode data indicates a predominant preference for 'Regular Air' shipping, accounting for 7,036 shipments. 'Delivery Truck' follows with 1,283 shipments, and 'Express Air' is used for 1,107 shipments. The overwhelming use of 'Regular Air' suggests it is the most cost-effective and efficient option for most customers. The relatively lower usage of 'Express Air' indicates it is reserved for urgent or high-value orders, while 'Delivery Truck' serves as a reliable alternative for certain deliveries. Understanding the factors driving these choices can help optimize logistics and improve overall shipping strategies.'''

Customer Segment

In [12]:
```
dataset['Customer Segment'].value_counts()
```

Out[12]:
```
Customer Segment
Corporate          3375
Home Office        2316
Consumer           1894
Small Business     1841
Name: count, dtype: int64
```

In [51]:
```
plt.figure(figsize= (6,4))
sns.countplot(x="Customer Segment", data= dataset)
plt.show()
```



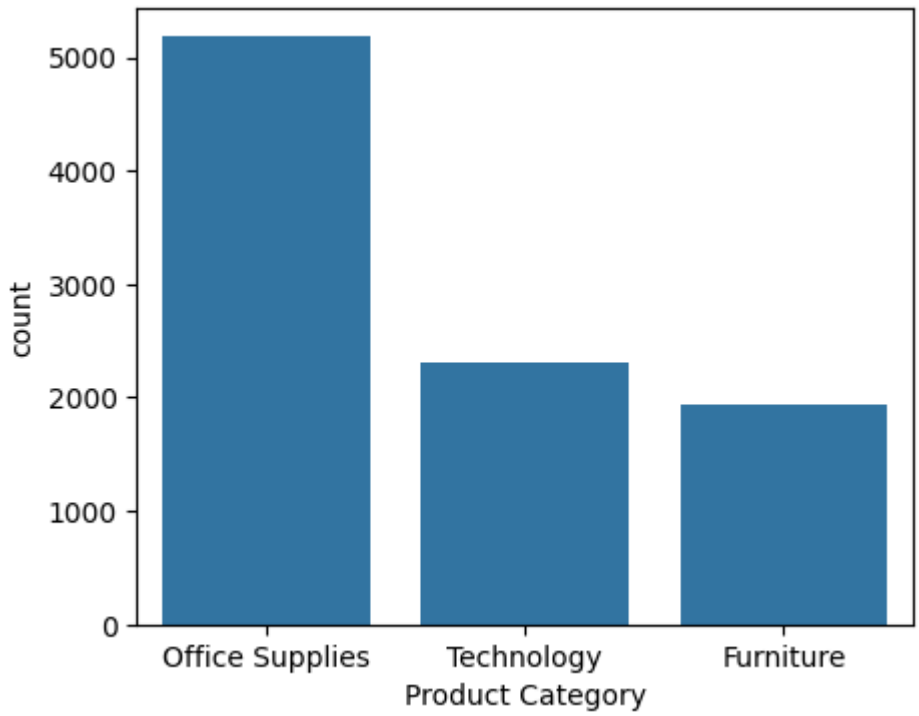
'''Insights:- The customer segment analysis reveals that the 'Corporate' segment is the largest, with 3,375 units, indicating a strong business-to-business market presence. 'Home Office' follows with 2,316 units, showing significant engagement from individual professionals and small teams. The 'Consumer' segment accounts for 1,894 units, while 'Small Business' contributes 1,841 units. This distribution highlights the importance of tailored strategies for each segment, with potential growth opportunities in the 'Consumer' and 'Small Business' areas. Focusing on the needs of these diverse segments can help drive further sales and customer satisfaction.'''

Product Category

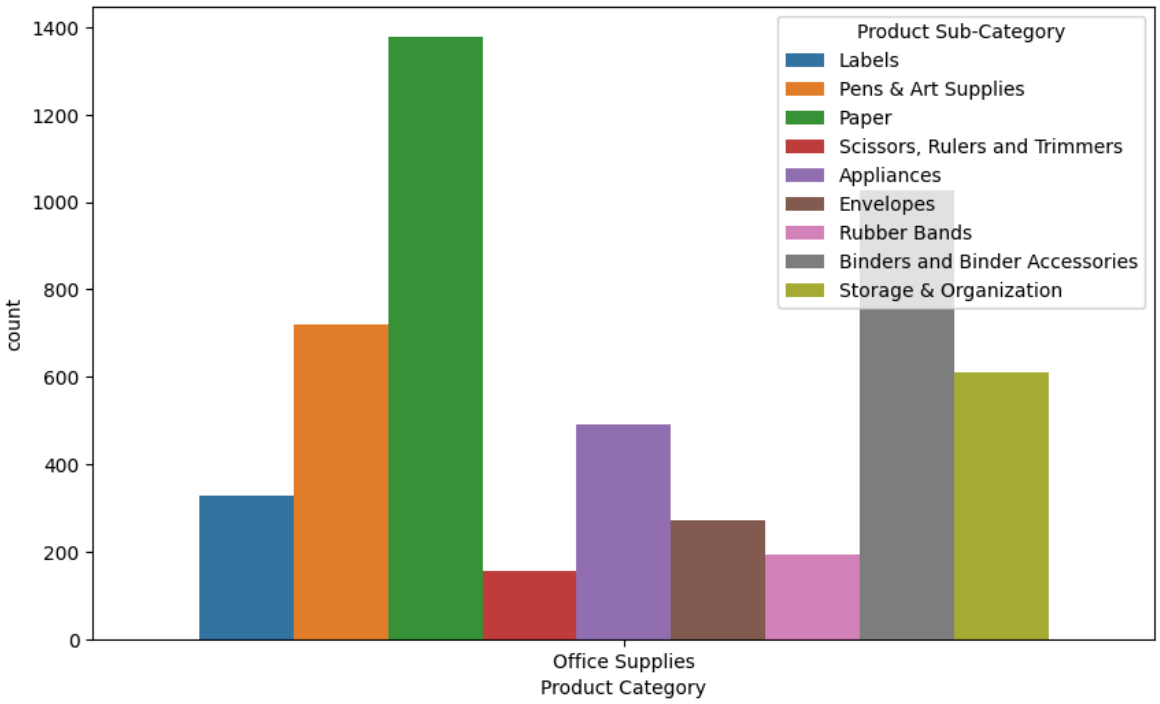
In [13]:
```
dataset['Product Category'].value_counts()
```

Out[13]:
```
Product Category
Office Supplies     5181
Technology          2312
Furniture           1933
Name: count, dtype: int64
```

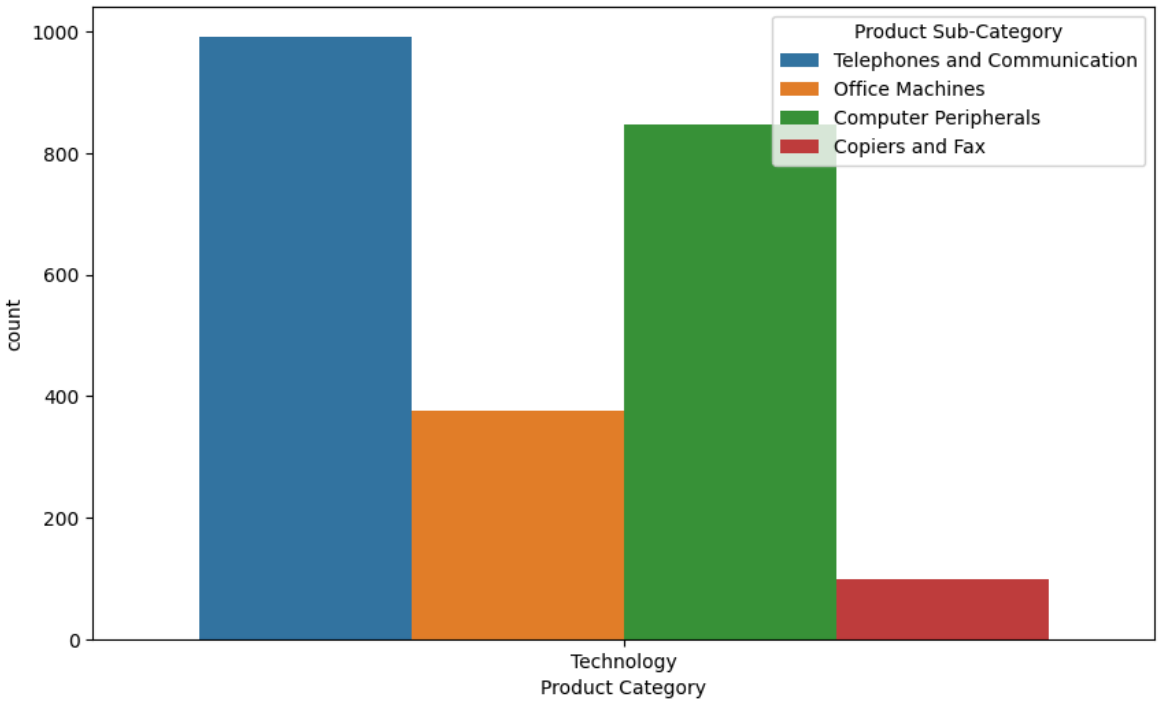
In [52]:
```
plt.figure(figsize= (5,4))
sns.countplot(x= "Product Category", data= dataset)
plt.show()
```



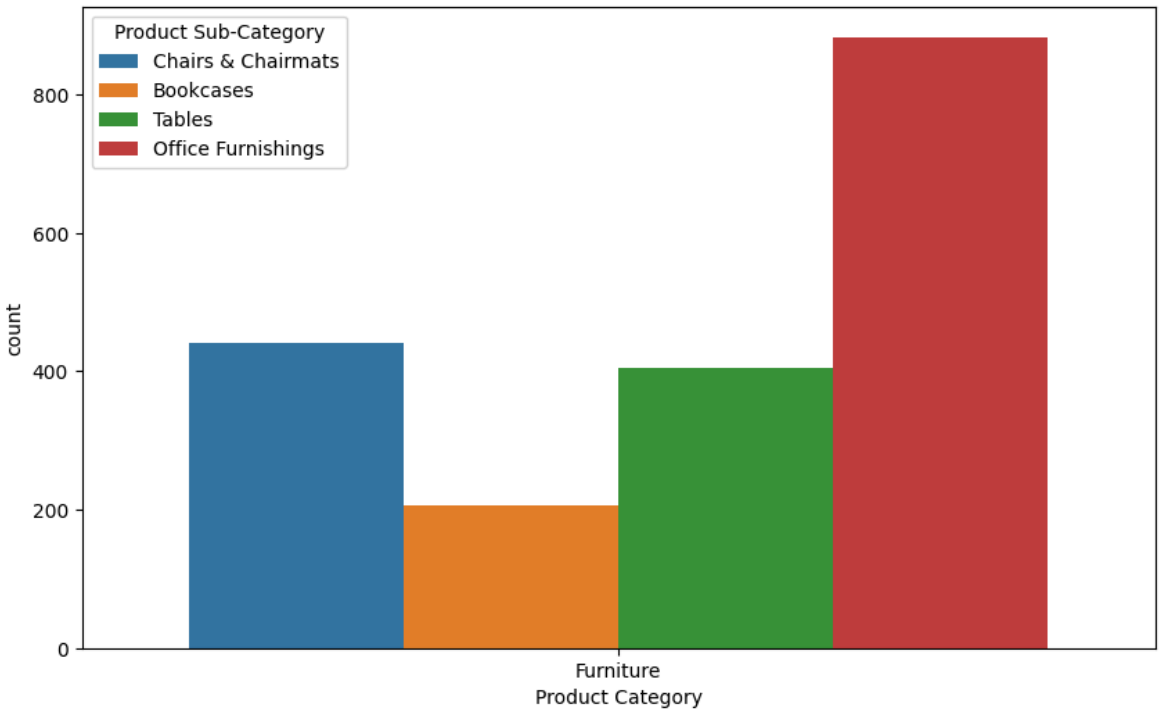
In [61]:
```
plt.figure(figsize= (10,6))
sns.countplot(x= "Product Category", data= dataset[dataset["Product Category"]==
plt.show()
```



In [66]:
```
plt.figure(figsize= (10,6))
sns.countplot(x= "Product Category", data= dataset[dataset["Product Category"]==
plt.show()
```



In [63]:
```
plt.figure(figsize= (10,6))
sns.countplot(x= "Product Category", data= dataset[dataset["Product Category"]==
plt.show()
```

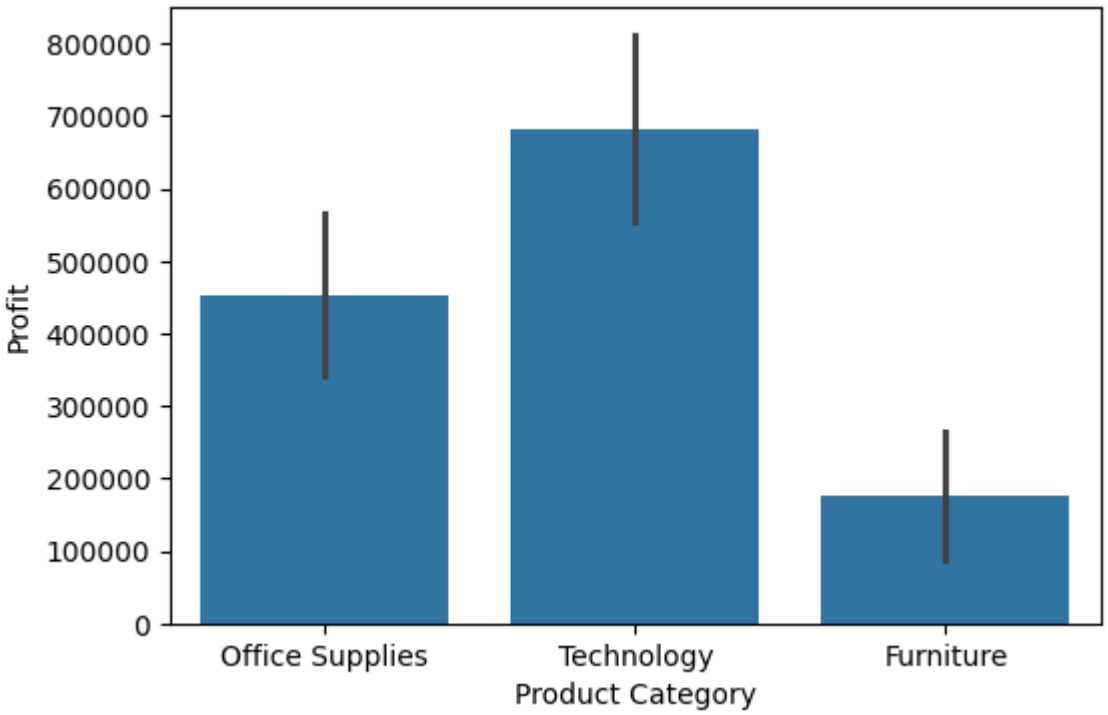


'''Insights:- The analysis of product categories shows that 'Office Supplies' dominate sales with 5,181 units, indicating strong demand in this category. 'Technology' products are the next significant category, with 2,312 units sold, reflecting the ongoing need for tech-related items in both personal and professional contexts. 'Furniture' accounts for 1,933 units, showing steady demand in this category as well. The data suggests that focusing on office supplies and technology products can yield substantial returns, while

opportunities for growth in the furniture category also exist. Tailored marketing and inventory strategies for each category could enhance overall sales performance.'''

Profit

In [85]:
```
plt.figure(figsize= (6,4))
sns.barplot(x="Product Category", y= "Profit", data= dataset, estimator= 'sum')
plt.show()
```



'''Insights:- This bar chart indicates that Technology products are the most lucrative, suggesting a potential focus area for maximizing profit. While Office Supplies also contribute substantially, the lower profit from Furniture highlights an opportunity for strategic improvement to enhance profitability in this category.'''

State or Province

In [89]:
```
dataset["State or Province"].value_counts()[:5]
```

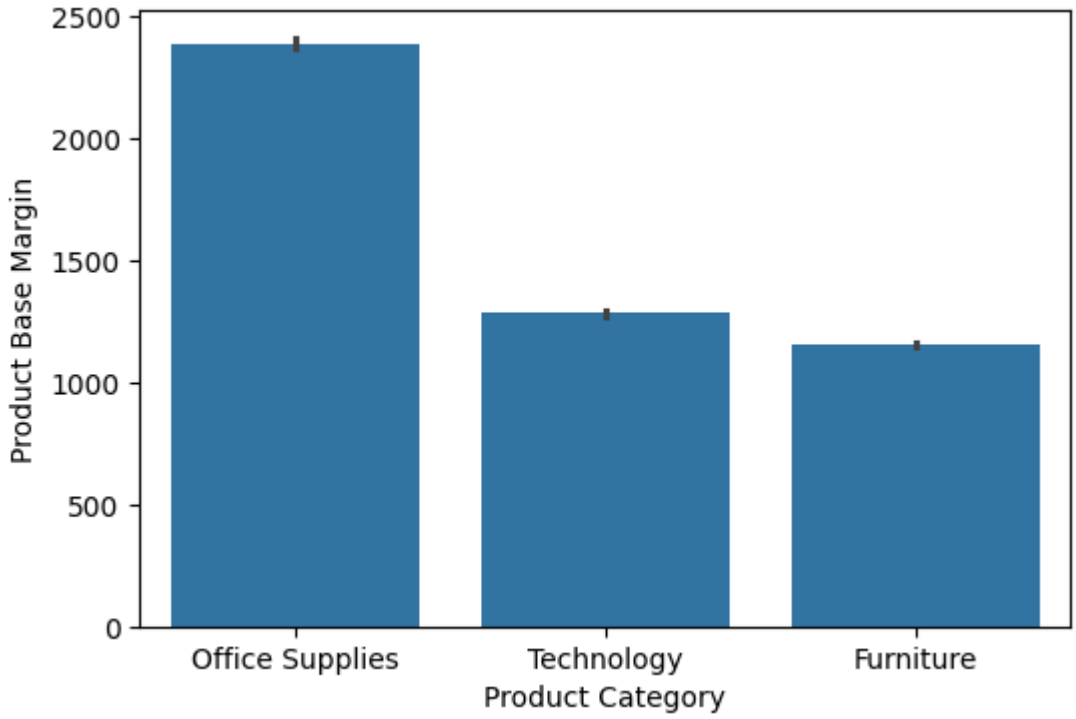
Out[89]:
```
State or Province
California    1021
Texas          646
Illinois       584
New York       574
Florida        522
Name: count, dtype: int64
```

'''The sales data by state indicates that California leads with 1,021 units, highlighting its significant market presence. Texas follows with 646 units, Illinois with 584 units, New York with 574 units, and Florida with 522 units. This distribution suggests that these states are key markets, potentially due to their large populations and economic activity. Targeted

marketing and sales strategies in these regions could further capitalize on existing demand and drive growth. Analyzing the specific factors contributing to high sales in these areas could provide insights for expanding market reach in other states.'''

Product Base Margin

In [90]:
```
plt.figure(figsize= (6,4))
sns.barplot(x="Product Category", y= "Product Base Margin", data= dataset, estim
plt.show()
```



'''Insights:- This data suggests that while Office Supplies offer the greatest margin, the margins for Technology and Furniture are comparatively lower. To enhance overall profitability, strategies might be considered to increase the margins of Technology and Furniture products. Identifying cost efficiencies or premium pricing opportunities in these categories could be beneficial.'''

'''Coclusion:- The analysis of sales data from 2010 to 2013 reveals significant growth, particularly in California and other key states, with a balanced distribution of order priorities and a strong preference for 'Regular Air' shipping. Corporate customers dominate sales, but there are opportunities for growth in the Consumer and Small Business segments. While Technology products lead in overall profitability, Office Supplies achieve the highest product base margin, indicating potential cost efficiencies. Furniture, with lower profits and margins, presents an area for strategic improvement. These insights suggest a focus on enhancing margins in Technology and Furniture categories and leveraging the high margins of Office Supplies to optimize profitability and drive further growth.'''