# Importing essential libraries

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

# Reading xlsx file and converting into dataframe

In [8]:
```
df = pd.read_excel('Superstore_USA.xlsx')
df.head(3)
```

Out[8]:

| | Row ID | Order Priority | Discount | Unit Price | Shipping Cost | Customer ID | Customer Name | Ship Mode | Cus Se |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 18606 | Not Specified | 0.01 | 2.88 | 0.50 | 2 | Janice Fletcher | Regular Air | Cor |
| **1** | 20847 | High | 0.01 | 2.84 | 0.93 | 3 | Bonnie Potter | Express Air | Cor |
| **2** | 23086 | Not Specified | 0.03 | 6.68 | 6.15 | 3 | Bonnie Potter | Express Air | Cor |

3 rows × 24 columns

# Understanding Data

In [30]:
```
df.shape
```

Out[30]:
```
(9426, 24)
```

In [10]:
```
df.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  9354 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 [12]:
```
df.describe()
```

Out[12]:

| | Row ID | Discount | Unit Price | Shipping Cost | Customer ID | Bas |
|---|---|---|---|---|---|---|
| **count** | 9426.000000 | 9426.000000 | 9426.000000 | 9426.000000 | 9426.000000 | 935 |
| **mean** | 20241.015277 | 0.049628 | 88.303686 | 12.795142 | 1738.422236 | |
| **min** | 2.000000 | 0.000000 | 0.990000 | 0.490000 | 2.000000 | |
| **25%** | 19330.250000 | 0.020000 | 6.480000 | 3.192500 | 898.000000 | |
| **50%** | 21686.500000 | 0.050000 | 20.990000 | 6.050000 | 1750.000000 | |
| **75%** | 24042.750000 | 0.080000 | 85.990000 | 13.990000 | 2578.750000 | |
| **max** | 26399.000000 | 0.250000 | 6783.020000 | 164.730000 | 3403.000000 | |
| **std** | 6101.890965 | 0.031798 | 281.540982 | 17.181203 | 979.167197 | |

# Missing data analysis

In [16]:
```
df.isnull().sum()
```

Out[16]:
```
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 [24]:
```
df['Product Base Margin'].fillna(df['Product Base
Margin'].mean(),inplace=True)
```

In [26]:
```
df.isnull().sum()
```

Out[26]:
```
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     0
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
```

# Univariate Data Analysis

## Order Priority

In [35]:
```
df['Order Priority'].value_counts()
```

Out[35]:
```
Order Priority
High            1970
Low             1926
Not Specified    1881
Medium          1844
Critical        1804
Critical           1
Name: count, dtype: int64
```

In [37]:
```
df['Order Priority'].unique()
```

Out[37]:
```
array(['Not Specified', 'High', 'Medium', 'Low', 'Critical', 'Critical
'],
      dtype=object)
```

In [57]:
```
df['Order Priority']=df['Order Priority'].replace('Critical
','Critical')
```

In [59]:
```
df['Order Priority'].unique()
```

Out[59]:
```
array(['Not Specified', 'High', 'Medium', 'Low', 'Critical'], dtype=obje
ct)
```

In [87]:
```
plt.figure(figsize=(10,3))
plt.title("Order Priority analysis",fontsize=24)
plt.xlabel("Order Priority",fontsize=14)
plt.ylabel("Counts",fontsize=14)
sns.countplot(data=df,x='Order Priority')
plt.show()
```


## Ship Mode

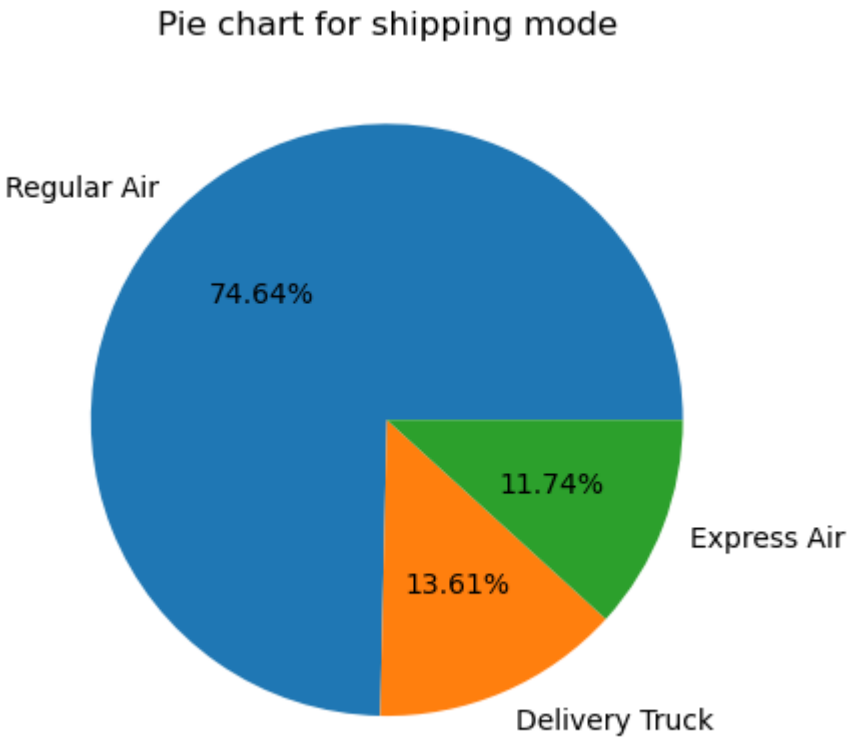
In [91]:
```
df['Ship Mode'].value_counts()
```

Out[91]:
```
Ship Mode
Regular Air       7036
Delivery Truck    1283
Express Air       1107
Name: count, dtype: int64
```

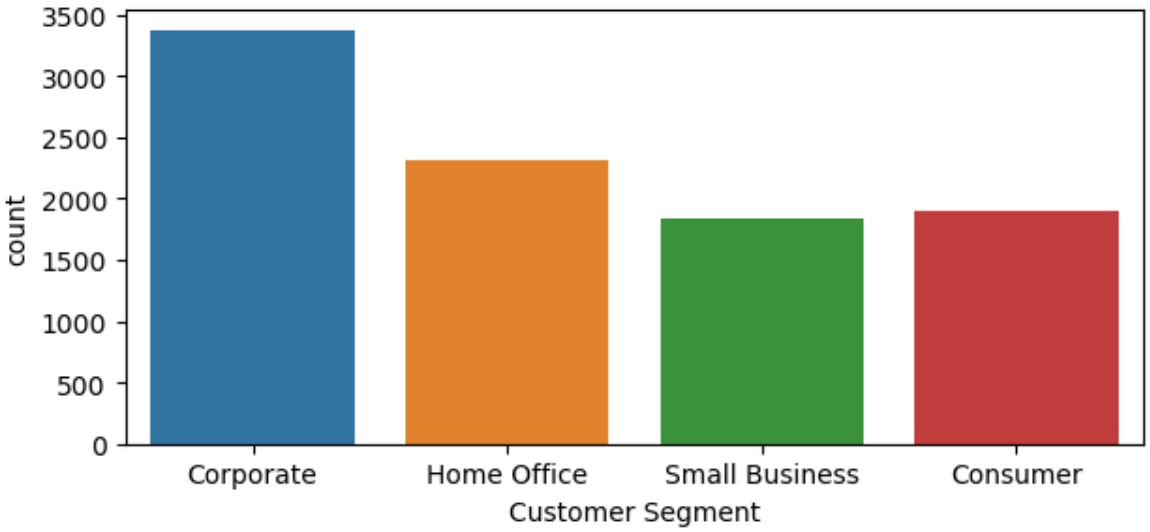
In [99]:
```
x = df['Ship Mode'].value_counts().index
y=df['Ship Mode'].value_counts().values
```

Out[99]:
```
array([7036, 1283, 1107])
```

In [111…
```
plt.title("Pie chart for shipping mode")
plt.pie(y,labels=x,autopct="%0.2f%%")
plt.show()
```

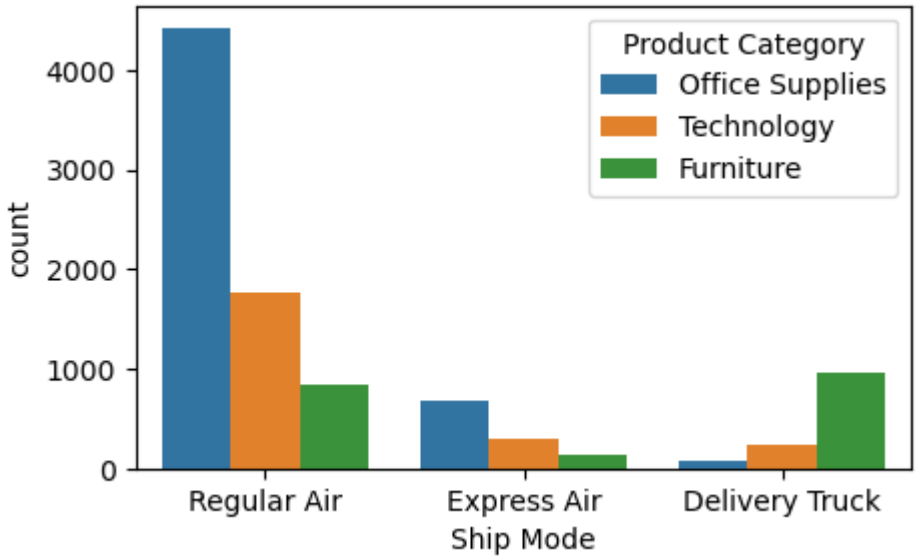



In [134…
```
plt.figure(figsize=(7,3))
sns.countplot(data=df,x='Customer Segment')
plt.show()
```

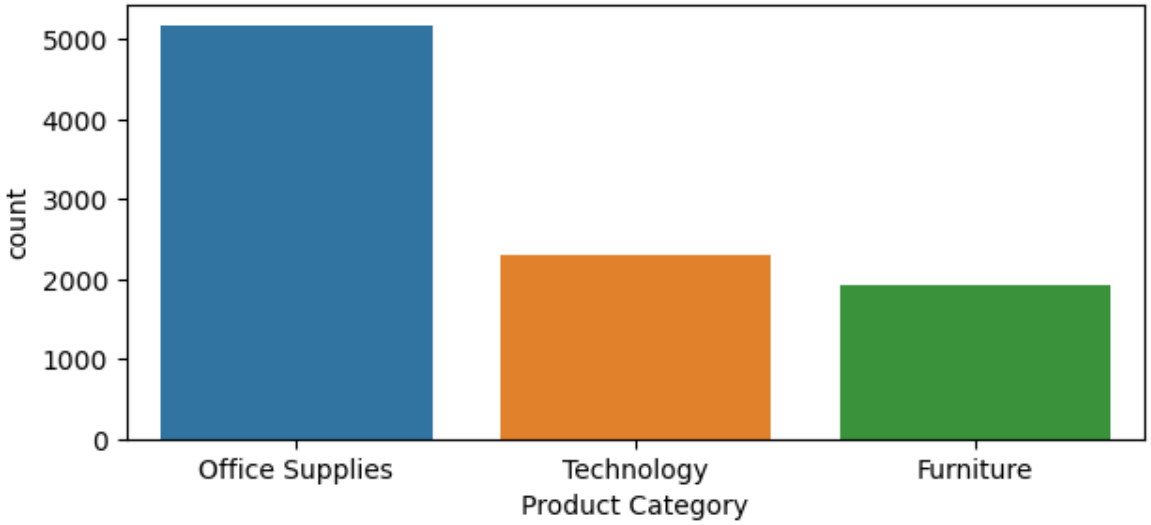



# Bivariate Data Analysis

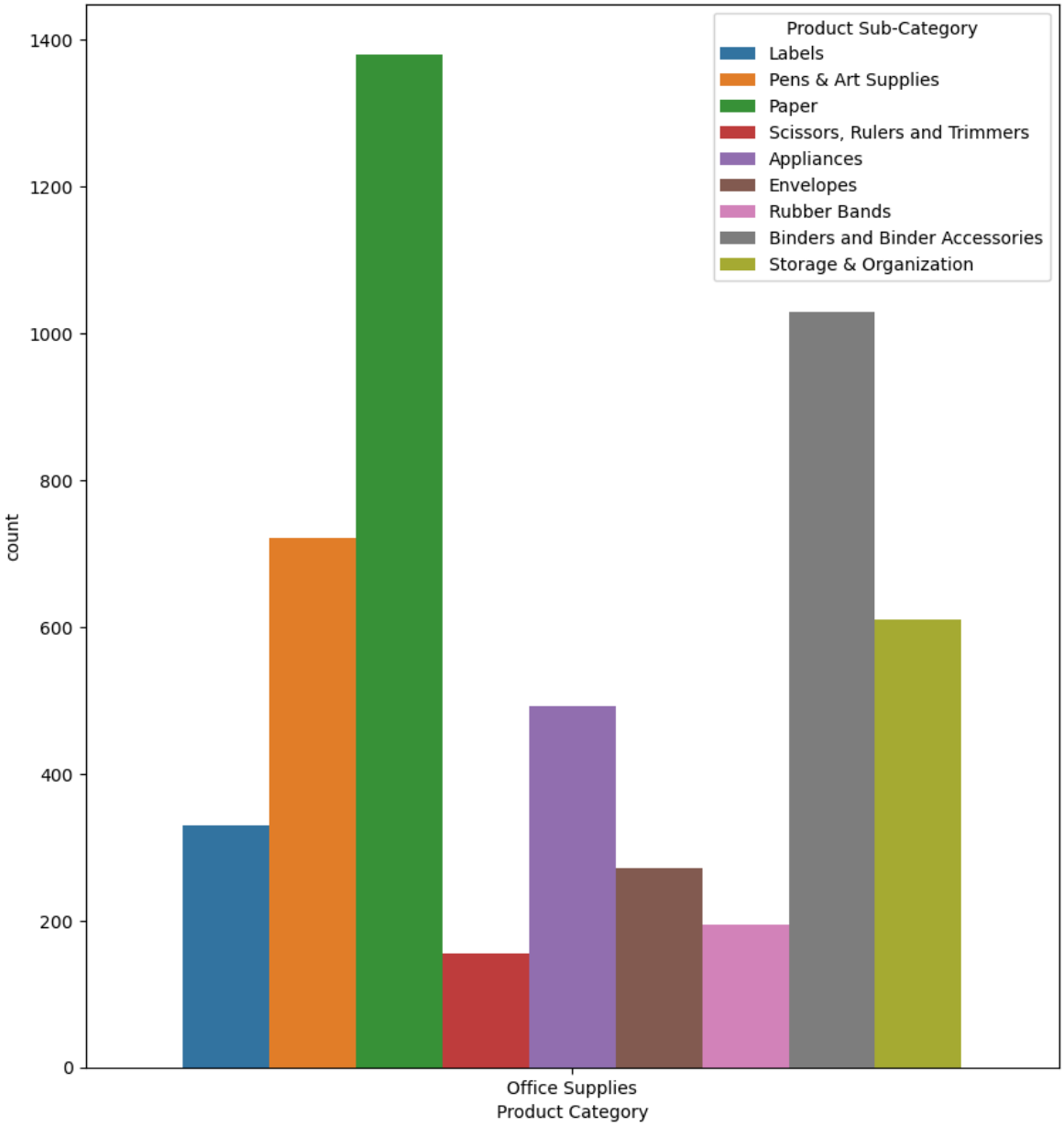
In [126…
```
plt.figure(figsize=(5,3))
sns.countplot(data=df,x='Ship Mode',hue='Product Category')
plt.show()
```




In [160…
```
plt.figure(figsize=(7,3))
sns.countplot(data=df,x='Product Category')
plt.show()
```

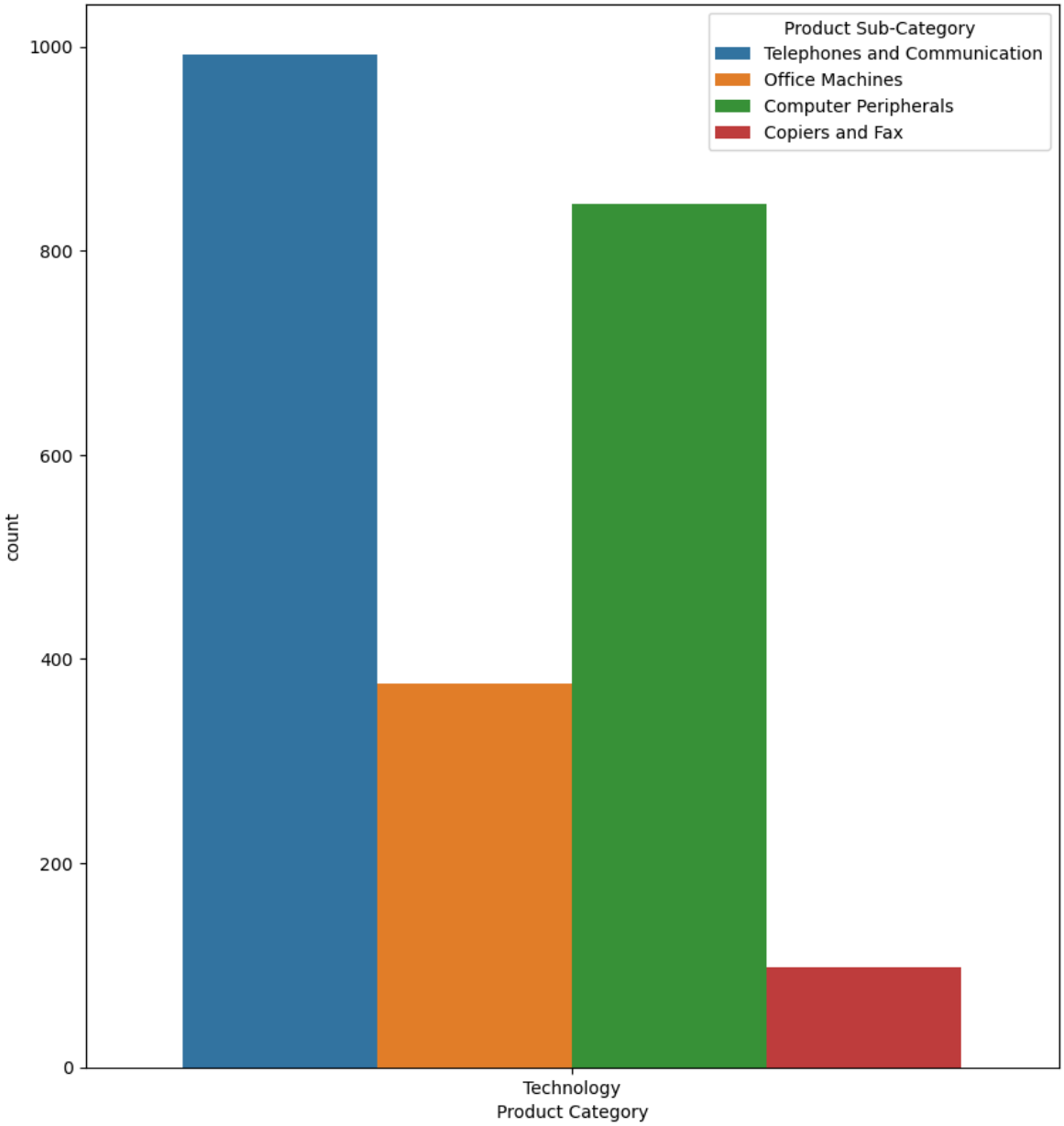



In [168…
```
plt.figure(figsize=(10,11))
sns.countplot(data=df[df['Product Category']=='Office
Supplies'],x='Product Category', hue='Product Sub-Category')
plt.show()
```



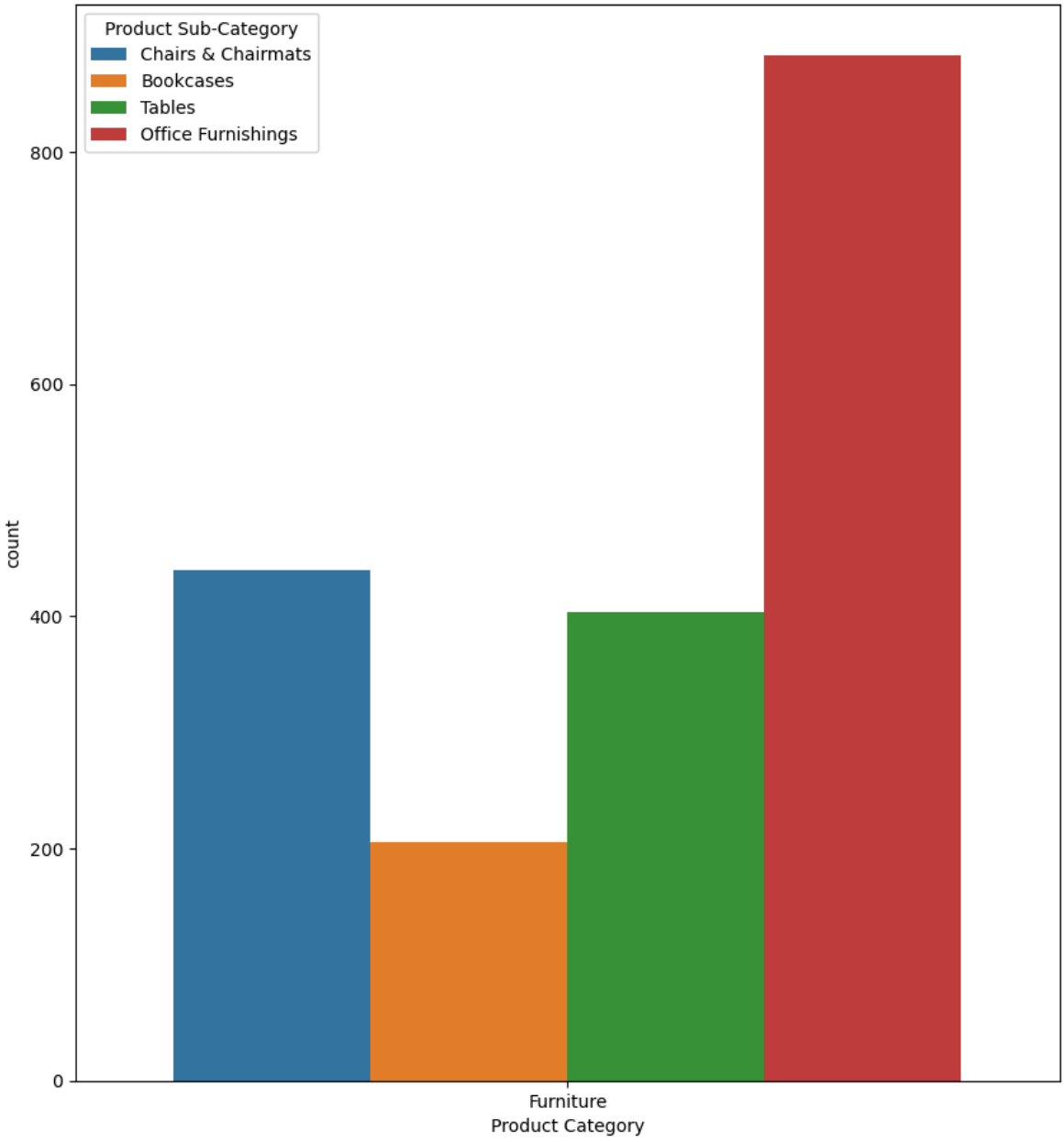

In [162…

```
plt.figure(figsize=(10,11))
sns.countplot(data=df[df['Product Category']=='Technology'],x='Product
Category', hue='Product Sub-Category')
plt.show()
```




In [164…

```
plt.figure(figsize=(10,11))
sns.countplot(data=df[df['Product Category']=='Furniture'],x='Product
Category', hue='Product Sub-Category')
plt.show()
```



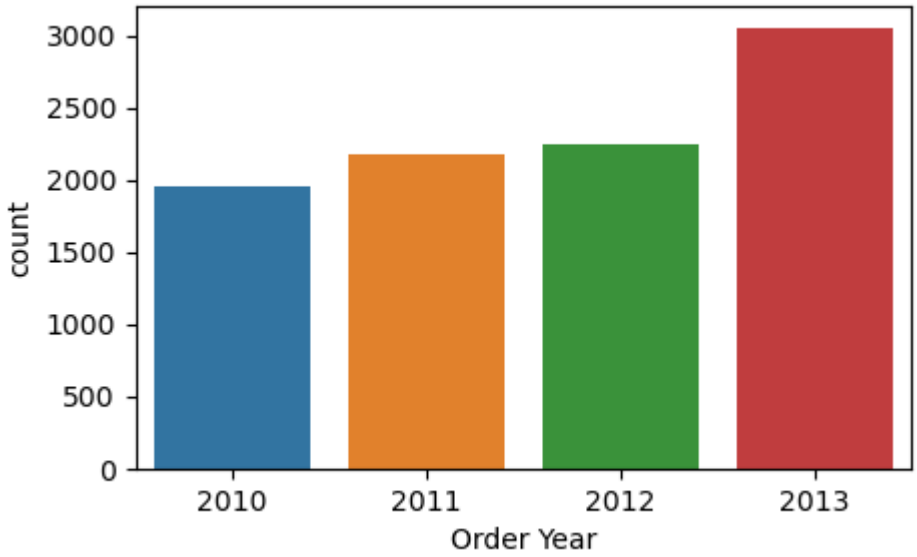

# Time Series Analysis

In [172…
```
df['Order Year']=df['Order Date'].dt.year
```

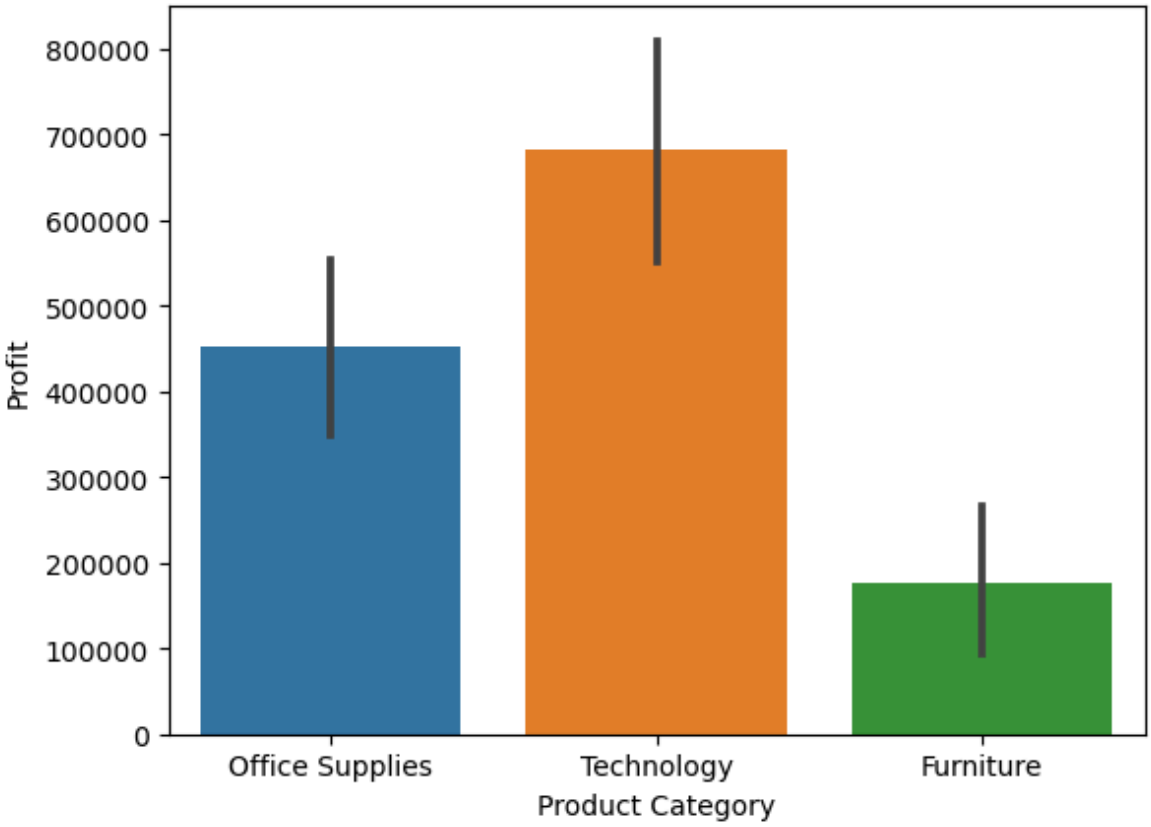
In [180…
```
df['Order Year'].value_counts()
```

Out[180…
```
Order Year
2013    3054
2012    2241
2011    2179
2010    1952
Name: count, dtype: int64
```

In [188…
```
plt.figure(figsize=(5,3))
sns.countplot(data=df,x='Order Year')
plt.show()
```




In [196…
```
sns.barplot(x='Product Category',y='Profit',data=df,estimator='sum')
plt.show()
```




In [198…
```
df['State or Province'].value_counts()
```

Out[198…

```
State or Province
California              1021
Texas                    646
Illinois                 584
New York                 574
Florida                  522
Ohio                     396
Washington               327
Michigan                 327
Pennsylvania             271
North Carolina           251
Indiana                  241
Minnesota                239
Massachusetts            222
Georgia                  214
Virginia                 198
Maryland                 178
Colorado                 177
New Jersey               177
Wisconsin                169
Oregon                   168
Tennessee                166
Missouri                 161
Iowa                     156
Utah                     146
Arizona                  134
Kansas                   133
Maine                    128
Alabama                  125
Arkansas                 123
Idaho                    114
South Carolina           105
Oklahoma                 104
Louisiana                 89
New Mexico                84
Kentucky                  83
Connecticut               82
Mississippi               78
Nebraska                  77
District of Columbia      68
Vermont                   61
New Hampshire             54
Montana                   49
West Virginia             43
Nevada                    43
North Dakota              34
South Dakota              28
Wyoming                   21
Rhode Island              20
Delaware                  15
Name: count, dtype: int64
```

In [ ]: