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

In [3]:
```
df=pd.read_csv("Amazon Sales data.csv")
```

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Region          100 non-null    object 
 1   Country         100 non-null    object 
 2   Item Type       100 non-null    object 
 3   Sales Channel   100 non-null    object 
 4   Order Priority  100 non-null    object 
 5   Order Date      100 non-null    object 
 6   Order ID        100 non-null    int64  
 7   Ship Date       100 non-null    object 
 8   Units Sold      100 non-null    int64  
 9   Unit Price      100 non-null    float64
 10  Unit Cost       100 non-null    float64
 11  Total Revenue   100 non-null    float64
 12  Total Cost      100 non-null    float64
 13  Total Profit    100 non-null    float64
dtypes: float64(5), int64(2), object(7)
memory usage: 11.1+ KB
```

In [5]:
```
df.head()
```

Out[5]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Unit Sol
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	992
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	280
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	177
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2014	810
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2013	506

In [7]:
```
df['Order ID'].duplicated()
```

Out[7]:
```
0     False
1     False
2     False
3     False
4     False
      ...
95    False
96    False
97    False
98    False
99    False
Name: Order ID, Length: 100, dtype: bool
```

In [9]:
```
df.drop_duplicates('Order ID')
```

Out[9]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship D
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2(
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2(
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2(
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2(
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2(
...	...	...	...	...	...	...	...	
95	Sub-Saharan Africa	Mali	Clothes	Online	M	7/26/2011	512878119	9/3/2(
96	Asia	Malaysia	Fruits	Offline	L	11/11/2011	810711038	12/28/2(
97	Sub-Saharan Africa	Sierra Leone	Vegetables	Offline	C	6/1/2016	728815257	6/29/2(
98	North America	Mexico	Personal Care	Offline	M	7/30/2015	559427106	8/8/2(
99	Sub-Saharan Africa	Mozambique	Household	Offline	L	2/10/2012	665095412	2/15/2(

100 rows × 14 columns

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

Out[10]:

	Order ID	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost
count	1.000000e+02	100.000000	100.000000	100.000000	1.000000e+02	1.000000e+02
mean	5.550204e+08	5128.710000	276.761300	191.048000	1.373488e+06	9.318057e+05
std	2.606153e+08	2794.484562	235.592241	188.208181	1.460029e+06	1.083938e+06
min	1.146066e+08	124.000000	9.330000	6.920000	4.870260e+03	3.612240e+03
25%	3.389225e+08	2836.250000	81.730000	35.840000	2.687212e+05	1.688680e+05
50%	5.577086e+08	5382.500000	179.880000	107.275000	7.523144e+05	3.635664e+05
75%	7.907551e+08	7369.000000	437.200000	263.330000	2.212045e+06	1.613870e+06
max	9.940222e+08	9925.000000	668.270000	524.960000	5.997055e+06	4.509794e+06

In [11]:
```
df.head()
```

Out[11]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Unit Sol
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	992
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	280
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	177
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2014	810
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2013	506

Exploratory Data Analysis

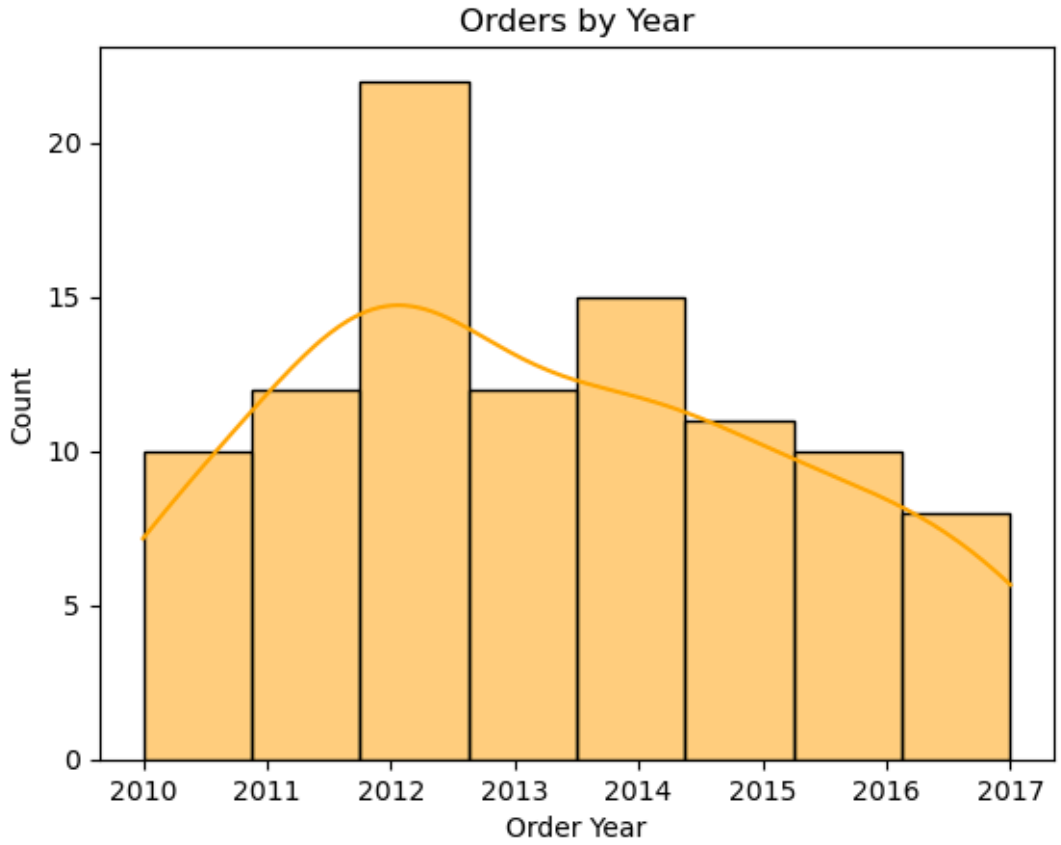
In [13]:
```
numeric_df=df.select_dtypes(include=['int64','float64'])
```

In [14]:
```
plt.figure(figsize=(10,8))
sns.heatmap(numeric_df.corr(),annot=True)
```

Out[14]: <AxesSubplot:>

In [16]:
```
df['Order Year']= pd.to_datetime(df['Order Date']).dt.year
sns.histplot(x=df['Order Year'],kde=True, color='orange')
plt.title('Orders by Year')
```

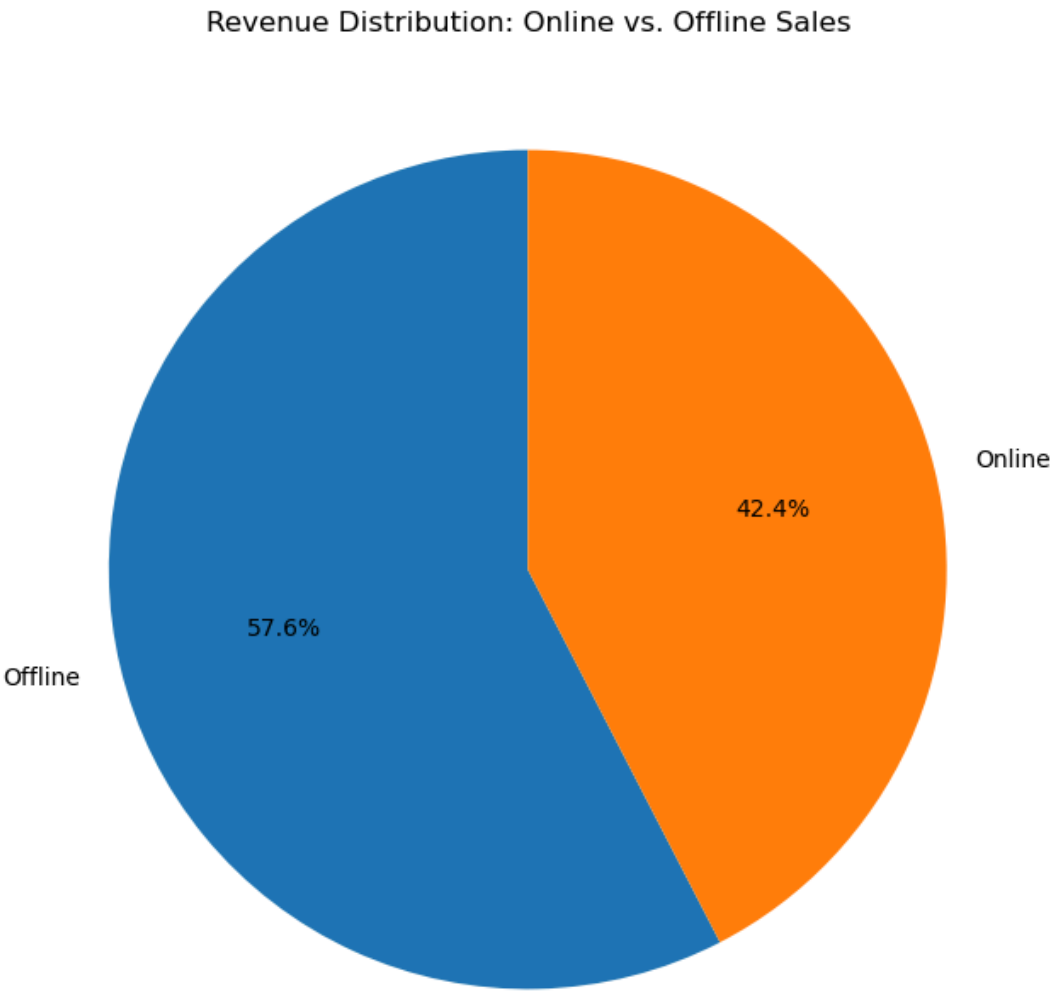
Out[16]:
```
Text(0.5, 1.0, 'Orders by Year')
```



In [17]:
```
revenue_type=df.groupby('Sales Channel')['Total Revenue'].sum()
```

In [19]:
```
plt.figure(figsize=[10,8])
plt.pie(revenue_type, labels=['Offline','Online'], autopct='%1.1f%%', s
plt.title('Revenue Distribution: Online vs. Offline Sales')
```

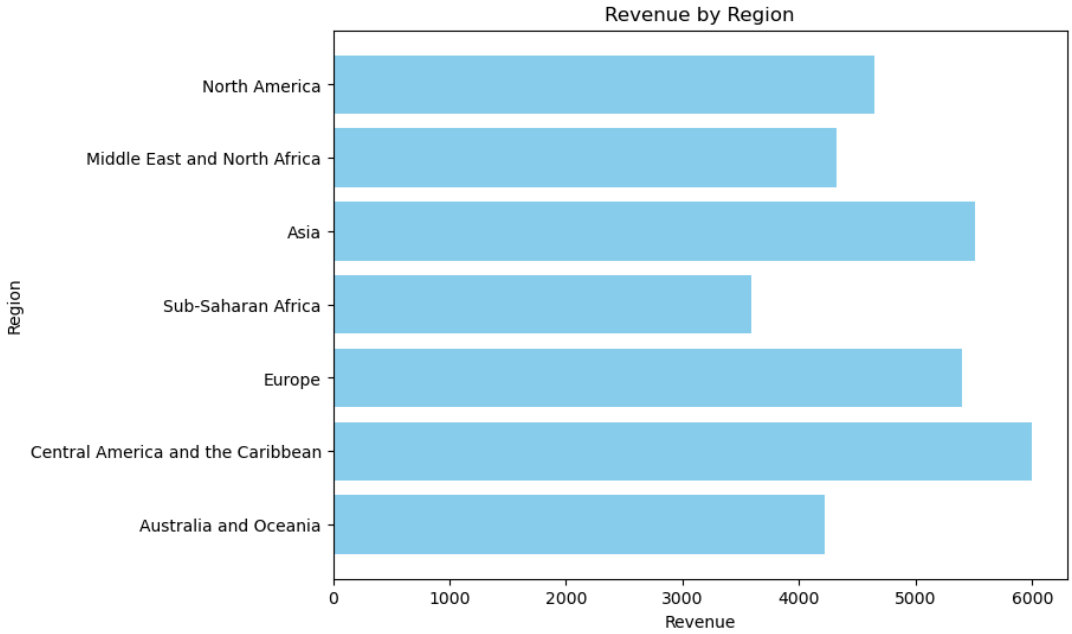
Out[19]: Text(0.5, 1.0, 'Revenue Distribution: Online vs. Offline Sales')



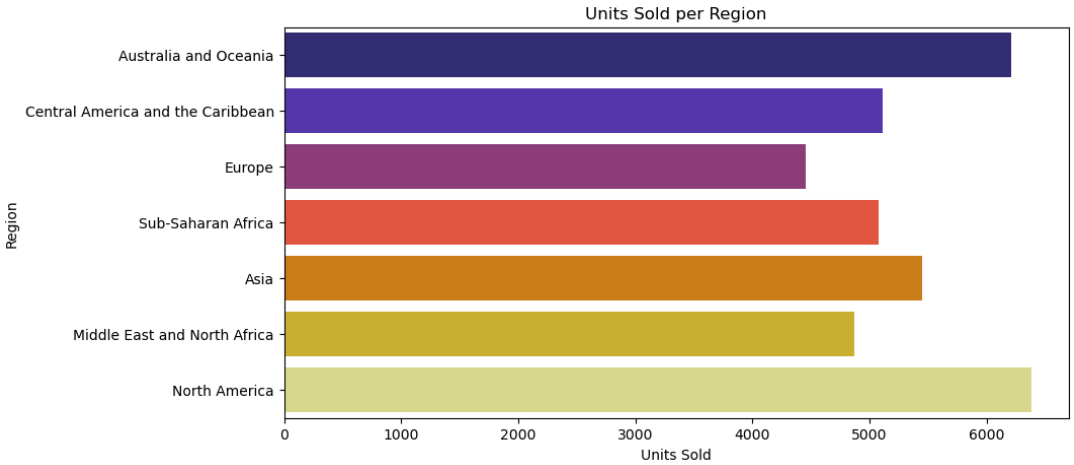
In [21]:
```
df['Revenue_in_thousands']=df['Total Revenue']/1000
plt.figure(figsize=[8,6])
plt.barh(df['Region'],df['Revenue_in_thousands'], color='skyblue')

#Add labels and title
plt.xlabel('Revenue')
plt.ylabel('Region')
plt.title('Revenue by Region')
```

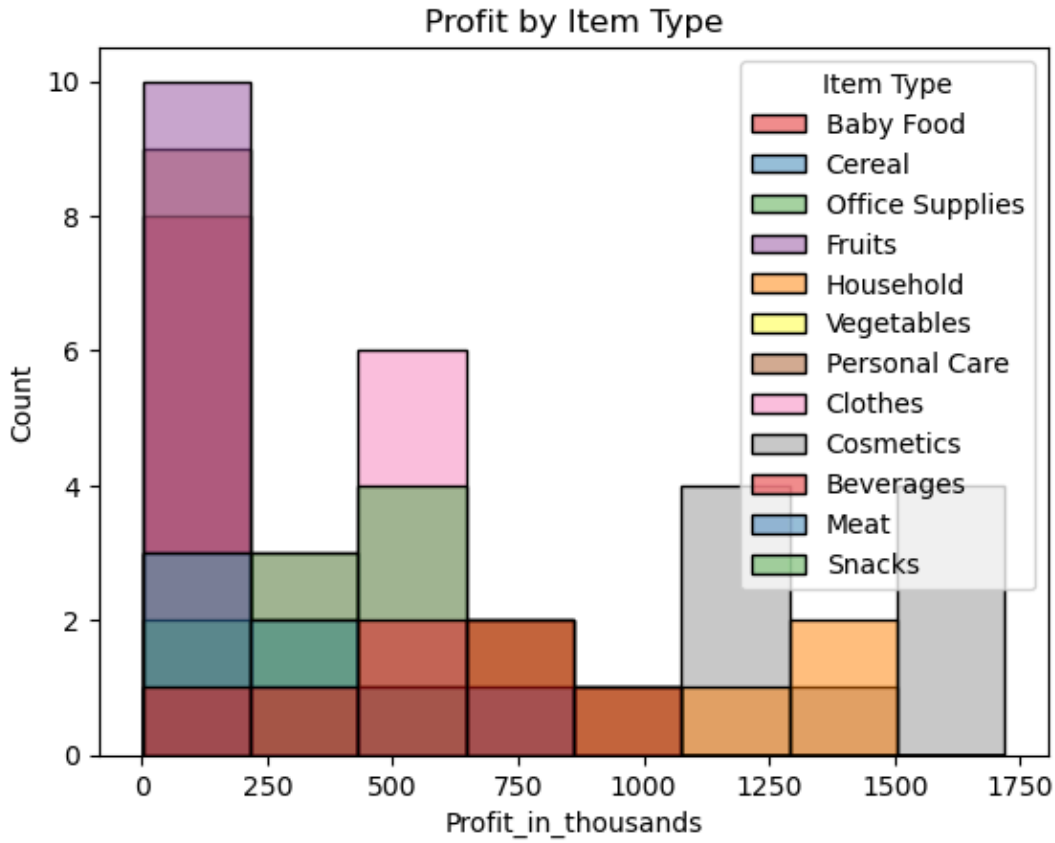
Out[21]: Text(0.5, 1.0, 'Revenue by Region')



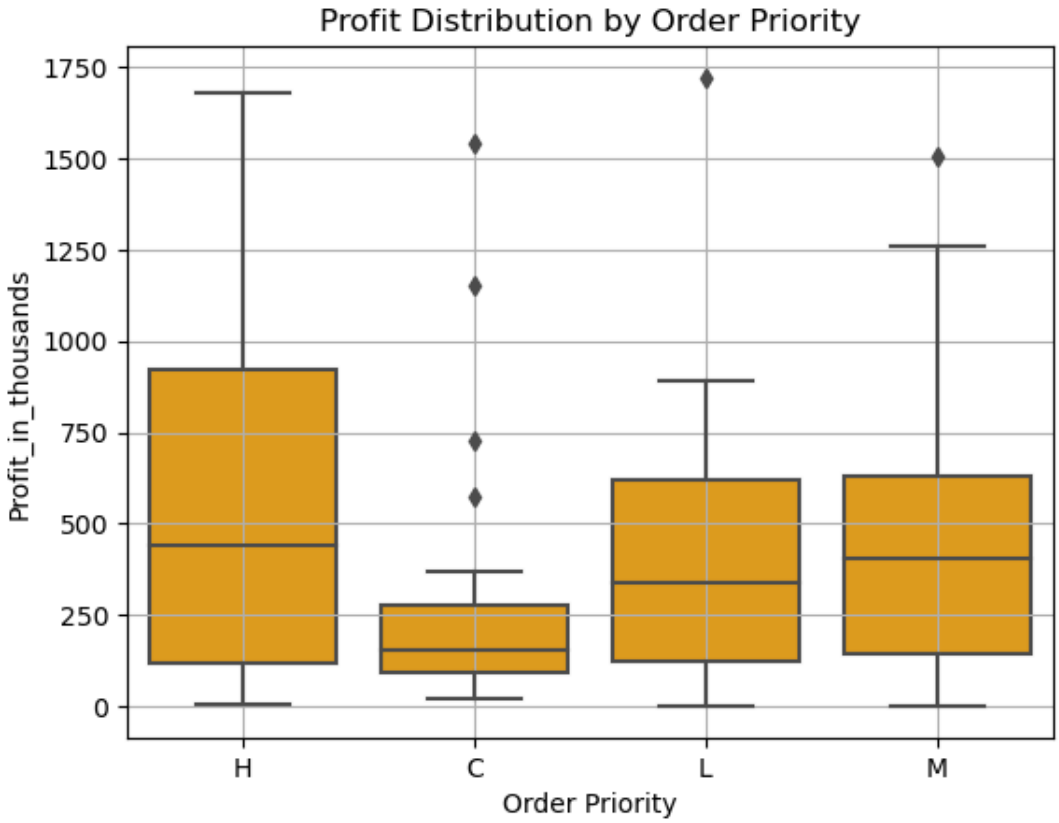
In [25]:
```
plt.figure(figsize=(10,5))
sns.barplot(y='Region',x='Units Sold',data=df, palette='CMRmap', ci=Non
plt.title('Units Sold per Region')
plt.show()
```



In [27]:
```
df['Profit_in_thousands']=df['Total Profit']/1000
sns.histplot(x=df['Profit_in_thousands'],data=df,hue='Item Type', palet
plt.title('Profit by Item Type')
plt.show()
```

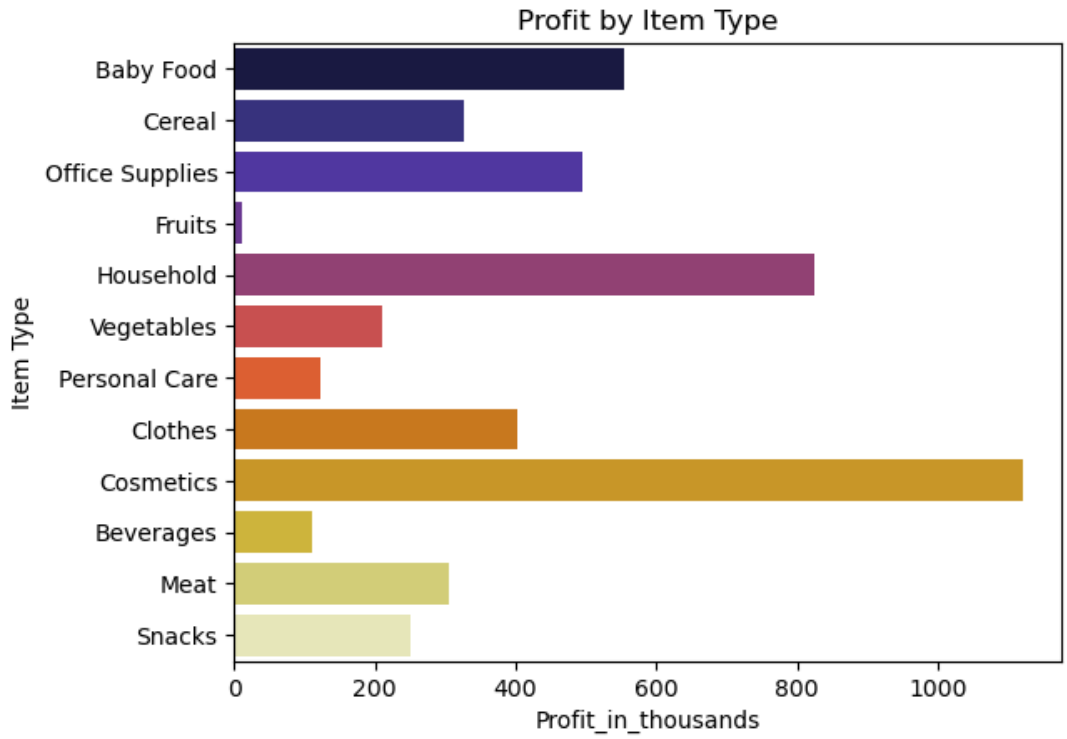


In [30]:
```
sns.boxplot(y='Profit_in_thousands',data=df,x='Order Priority',color='o
plt.grid()
plt.title('Profit Distribution by Order Priority')
plt.show()
```



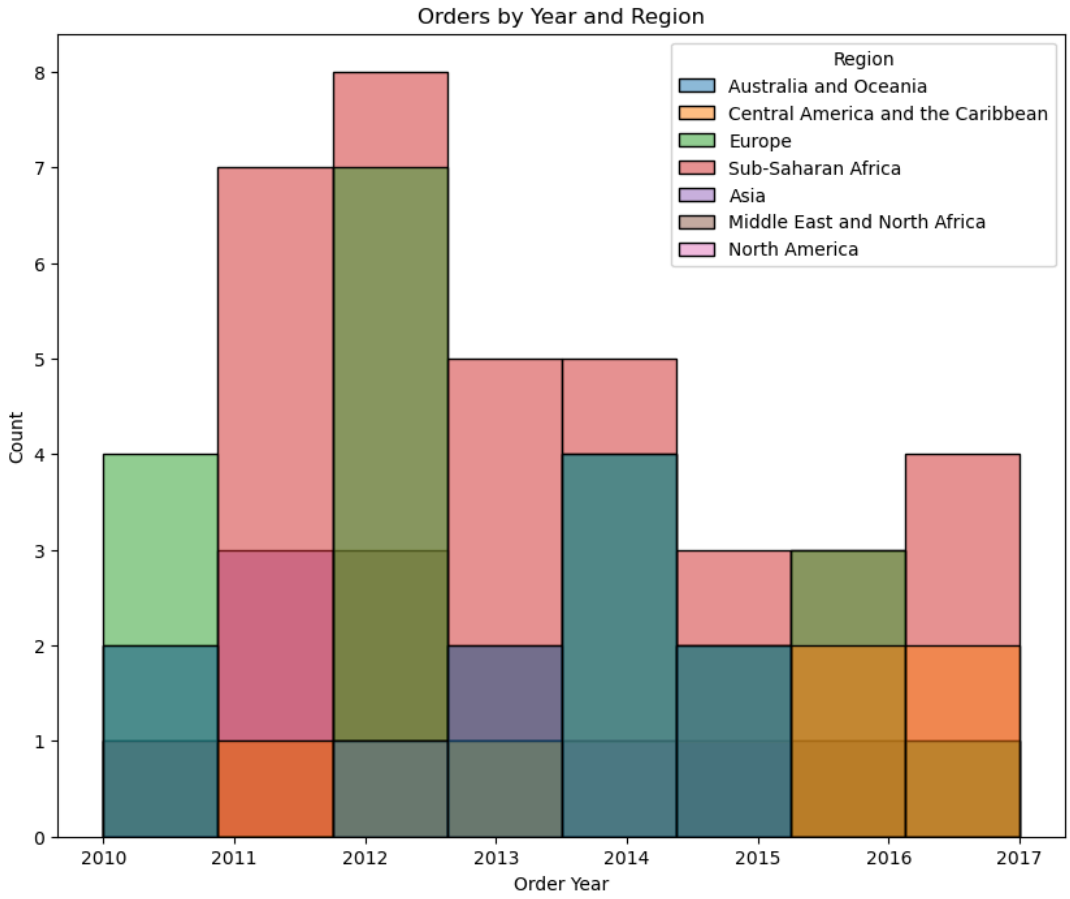
In [31]:
```
sns.barplot(x=df['Profit_in_thousands'],data=df,y='Item Type',palette='
plt.title('Profit by Item Type')
```

Out[31]:
```
Text(0.5, 1.0, 'Profit by Item Type')
```

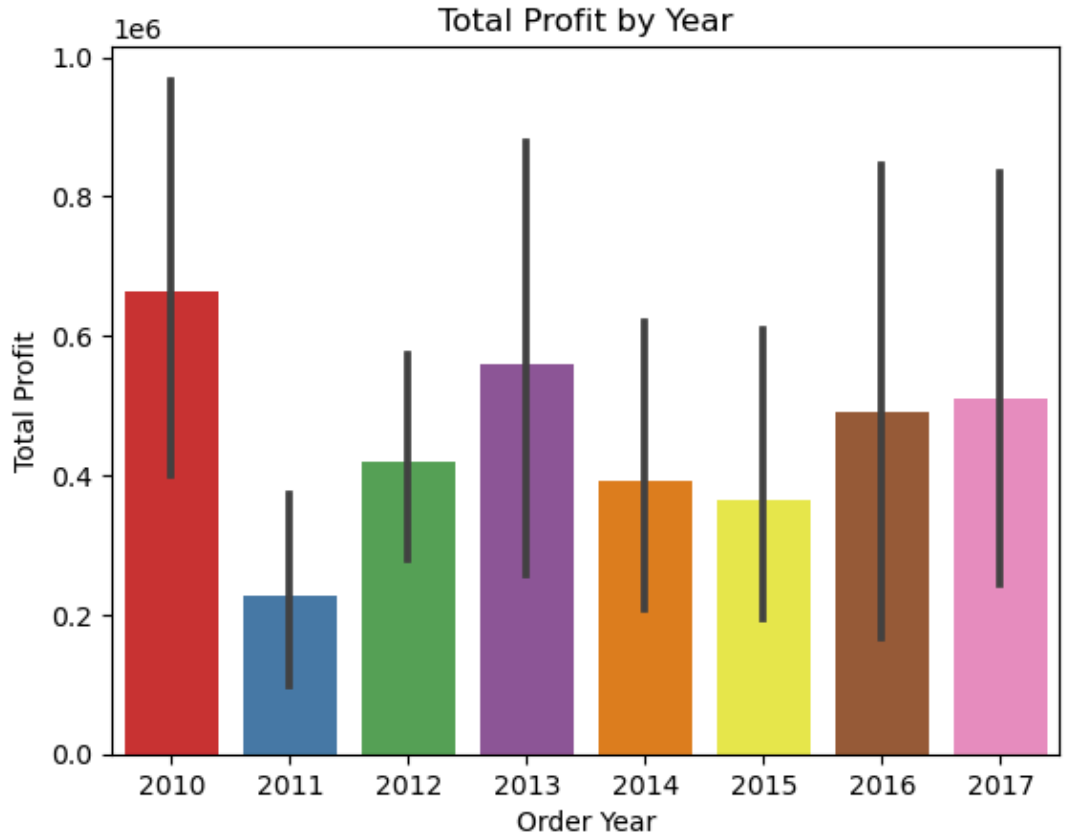


In [32]:
```
plt.figure(figsize=(10,8))
sns.histplot(x='Order Year', data=df, hue='Region')
plt.title('Orders by Year and Region')
```

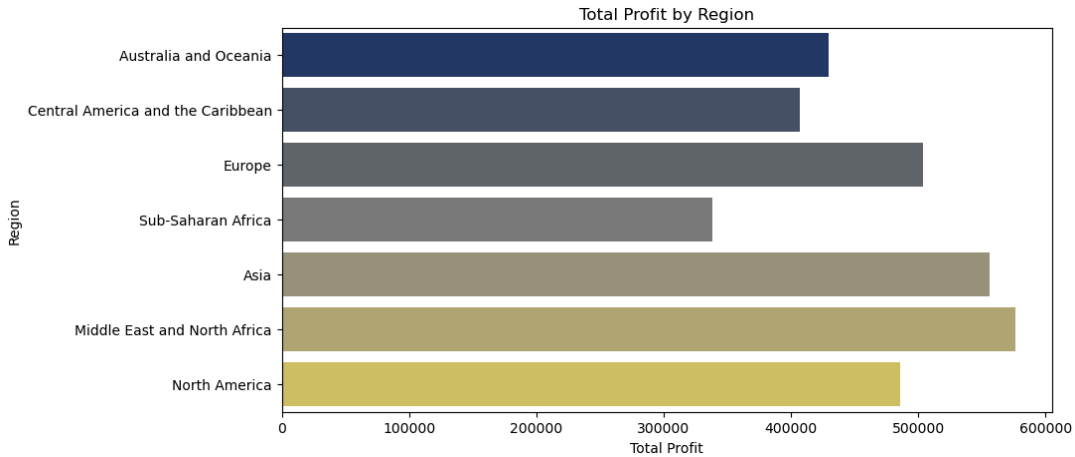
Out[32]:
```
Text(0.5, 1.0, 'Orders by Year and Region')
```



In [35]:
```
sns.barplot(x='Order Year', data=df, y='Total Profit',palette='Set1')
plt.title('Total Profit by Year')
plt.show()
```



In [43]:
```
.figure(figsize=(10, 5))
.barplot(x='Total Profit', data=df, y='Region', palette='cividis',ci=No
.title('Total Profit by Region')
.show()
```



In [ ]:

In [ ]: