In [2]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import datetime as dt
```

In [3]:

```
df=pd.read_csv(r"C:\Users\kruna\OneDrive\Desktop\mega project data\Amazon Sales data.csv")
df.head()
```

Out[3]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	9925	255.28	159.42	2533654.00	1582243.50	951410.50
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2804	205.70	117.11	576782.80	328376.44	248406.36
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	1779	651.21	524.96	1158502.59	933903.84	224598.75
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2014	8102	9.33	6.92	75591.66	56065.84	19525.82
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2013	5062	651.21	524.96	3296425.02	2657347.52	639077.50

In [14]:

```
df.columns
```

Out[14]:

```
Index(['Region', 'Country', 'Item Type', 'Sales Channel', 'Order Priority',
       'Order Date', 'Order ID', 'Ship Date', 'Units Sold', 'Unit Price',
       'Unit Cost', 'Total Revenue', 'Total Cost', 'Total Profit'],
      dtype='object')
```

In [15]:

```
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 [16]:

```
df.isnull().sum()
```

Out[16]:

```
Region            0
Country           0
Item Type         0
Sales Channel     0
Order Priority    0
Order Date        0
Order ID          0
Ship Date         0
Units Sold        0
Unit Price        0
Unit Cost         0
Total Revenue     0
Total Cost        0
Total Profit      0
dtype: int64
```

In [17]:

```
df['Order_Date']=pd.to_datetime(df['Order Date'])
df['Ship_Date'] = pd.to_datetime(df['Ship Date'])
```

In [18]:

```
df["os_lead_time"]=df['Ship_Date']-df['Order_Date']
df.head()
```

Out[18]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	9925	255.28	159.42	2533654.00	1582243.50	951410.50
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2804	205.70	117.11	576782.80	328376.44	248406.36
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	1779	651.21	524.96	1158502.59	933903.84	224598.75
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2014	8102	9.33	6.92	75591.66	56065.84	19525.82
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2013	5062	651.21	524.96	3296425.02	2657347.52	639077.50

In [19]:

```
df.groupby('Region')['os_lead_time'].mean()
```

Out[19]:

```
Region
Asia                                    28 days 17:27:16.363636363
Australia and Oceania                   24 days 06:32:43.636363636
Central America and the Caribbean       26 days 17:08:34.285714285
Europe                                  24 days 03:16:21.818181818
Middle East and North Africa                      24 days 04:48:00
North America                                     25 days 16:00:00
Sub-Saharan Africa                                19 days 21:20:00
Name: os_lead_time, dtype: timedelta64[ns]
```

In [20]:

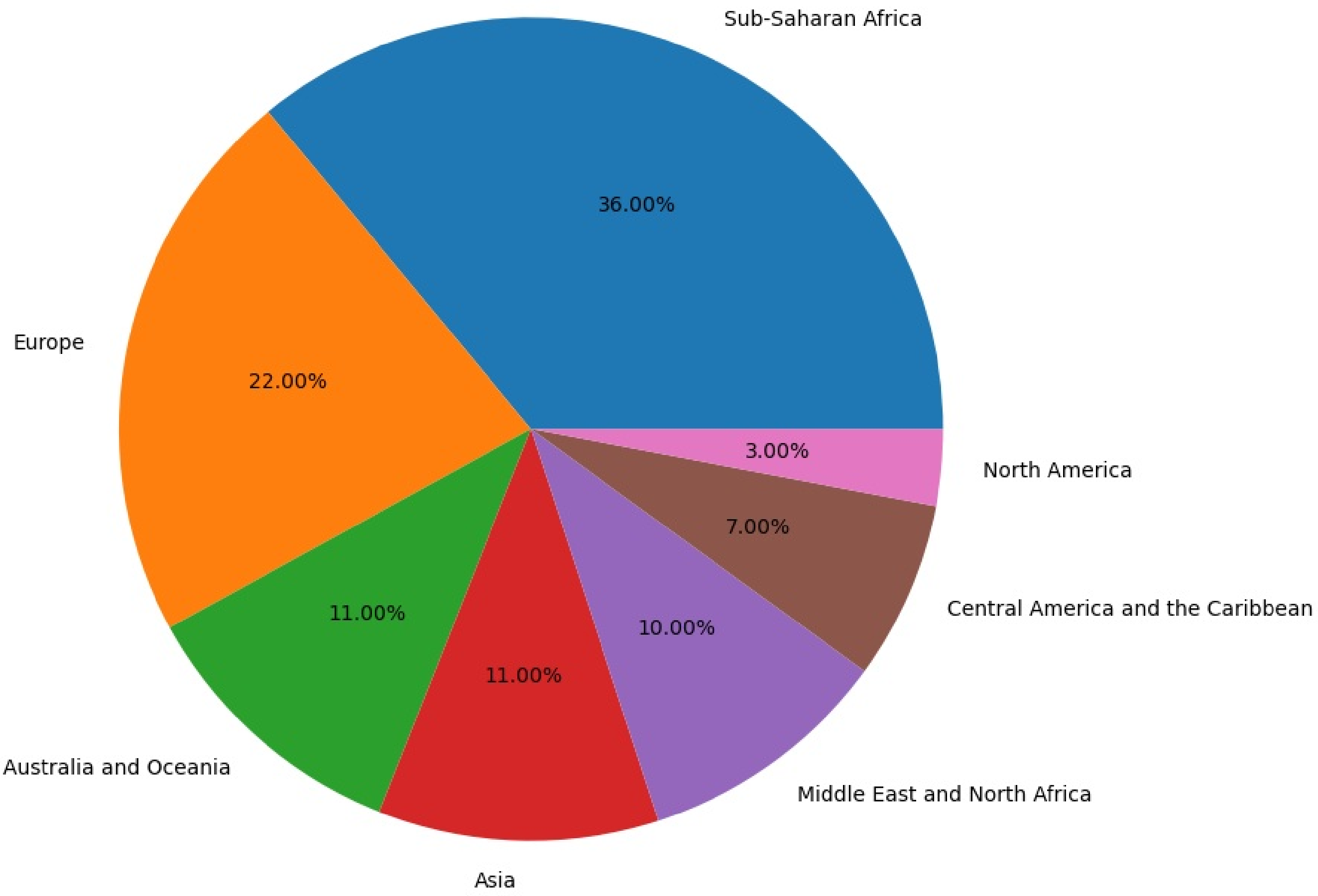
```
df.groupby('Country')['os_lead_time'].mean()
```

Out[20]:

```
Country
Albania           44 days 00:00:00
Angola             4 days 00:00:00
Australia         18 days 16:00:00
Austria            7 days 00:00:00
Azerbaijan        30 days 00:00:00
                        ...
The Gambia        17 days 06:00:00
Turkmenistan      24 days 00:00:00
Tuvalu            30 days 00:00:00
United Kingdom    40 days 00:00:00
Zambia             1 days 00:00:00
Name: os_lead_time, Length: 76, dtype: timedelta64[ns]
```

In [21]:

```
region_names=df.Region.value_counts().index
x=df.Region.value_counts().values
# Pie Chart for Regions
fig,ax = plt.subplots(figsize=(9,9))
plt.pie(x,labels=region_names,autopct='%1.2f%%')
plt.show()
```



In [22]:

```
country_val=df["Country"].value_counts()
country_val
```

Out[22]:

```
Country
The Gambia                4
Sierra Leone              3
Sao Tome and Principe     3
Mexico                    3
Australia                 3
                         ..
Comoros                   1
Iceland                   1
Macedonia                 1
Mauritania                1
Mozambique                1
Name: count, Length: 76, dtype: int64
```

In [23]:

```
df['Order_Date']=pd.to_datetime(df['Order Date'])
df['Ship_Date'] = pd.to_datetime(df['Ship Date'])
df=df.drop(columns=['Order Date'])
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 16 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 ID        100 non-null    int64
 6   Ship Date       100 non-null    object
 7   Units Sold      100 non-null    int64
 8   Unit Price      100 non-null    float64
 9   Unit Cost       100 non-null    float64
 10  Total Revenue   100 non-null    float64
 11  Total Cost      100 non-null    float64
 12  Total Profit    100 non-null    float64
 13  Order_Date      100 non-null    datetime64[ns]
 14  Ship_Date       100 non-null    datetime64[ns]
 15  os_lead_time    100 non-null    timedelta64[ns]
dtypes: datetime64[ns](2), float64(5), int64(2), object(6), timedelta64[ns](1)
memory usage: 12.6+ KB
```

In [24]:

```python
df['Order Month'] = df['Order_Date'].dt.month
df['Order Year'] = df['Order_Date'].dt.year
df.drop(columns=['Order_Date'])
```

Out[24]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit	S
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	669165933	6/27/2010	9925	255.28	159.42	2533654.00	1582243.50	951410.50	
1	Central America and the Caribbean	Grenada	Cereal	Online	C	963881480	9/15/2012	2804	205.70	117.11	576782.80	328376.44	248406.36	
2	Europe	Russia	Office Supplies	Offline	L	341417157	5/8/2014	1779	651.21	524.96	1158502.59	933903.84	224598.75	
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	514321792	7/5/2014	8102	9.33	6.92	75591.66	56065.84	19525.82	
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	115456712	2/6/2013	5062	651.21	524.96	3296425.02	2657347.52	639077.50	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
95	Sub-Saharan Africa	Mali	Clothes	Online	M	512878119	9/3/2011	888	109.28	35.84	97040.64	31825.92	65214.72	
96	Asia	Malaysia	Fruits	Offline	L	810711038	12/28/2011	6267	9.33	6.92	58471.11	43367.64	15103.47	
97	Sub-Saharan Africa	Sierra Leone	Vegetables	Offline	C	728815257	6/29/2016	1485	154.06	90.93	228779.10	135031.05	93748.05	
98	North America	Mexico	Personal Care	Offline	M	559427106	8/8/2015	5767	81.73	56.67	471336.91	326815.89	144521.02	
99	Sub-Saharan Africa	Mozambique	Household	Offline	L	665095412	2/15/2012	5367	668.27	502.54	3586605.09	2697132.18	889472.91	

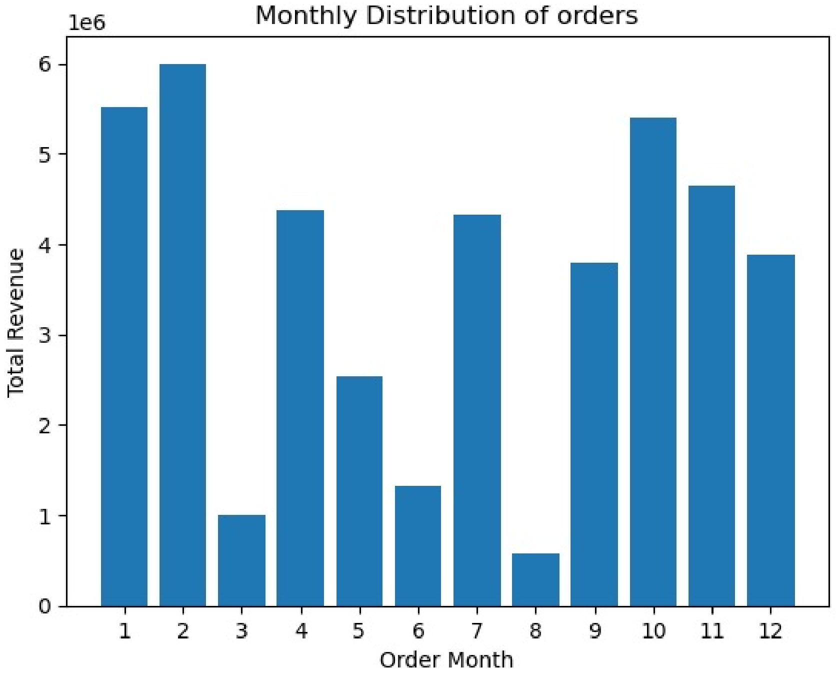
100 rows × 17 columns

In [26]:

```python
plt.bar(df['Order Month'], df['Total Revenue'])
plt.title('Monthly Distribution of orders')
plt.xticks([1,2,3,4,5,6,7,8,9,10,11,12])
plt.xlabel('Order Month')
plt.ylabel('Total Revenue')
```

Out[26]:

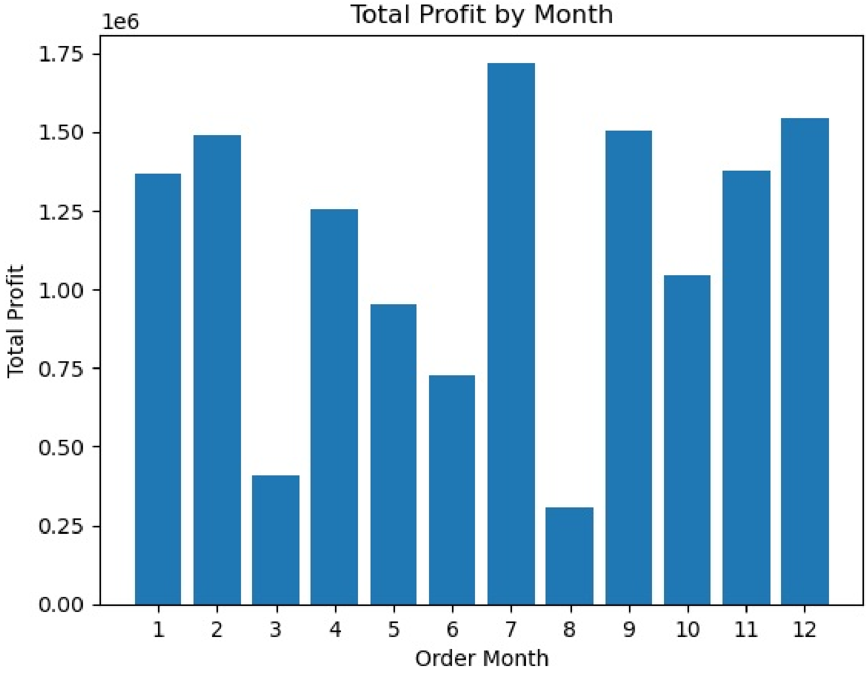
```
Text(0, 0.5, 'Total Revenue')
```



In [35]:

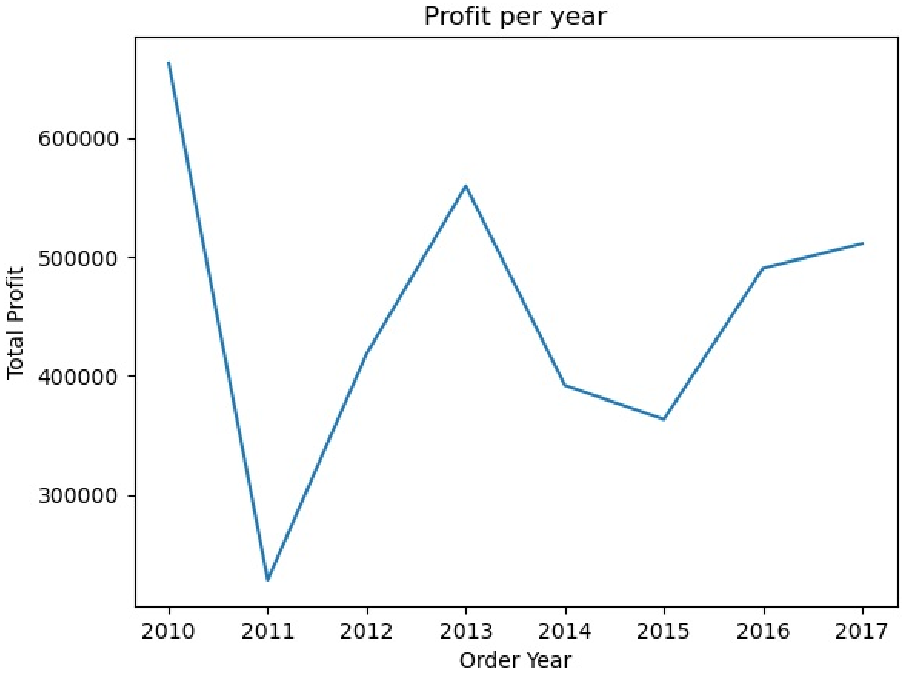
```python
plt.bar(df['Order Month'], df['Total Profit'])
plt.title('Total Profit by Month')
plt.xticks([1,2,3,4,5,6,7,8,9,10,11,12])
plt.xlabel('Order Month')
plt.ylabel('Total Profit')
```

Out[35]:
```
Text(0, 0.5, 'Total Profit')
```



In [37]:
```
df.groupby('Order Year')['Total Profit'].mean().plot()
plt.xlabel('Order Year')
plt.ylabel('Total Profit')
plt.title('Profit per year')
```

Out[37]:
```
Text(0.5, 1.0, 'Profit per year')
```



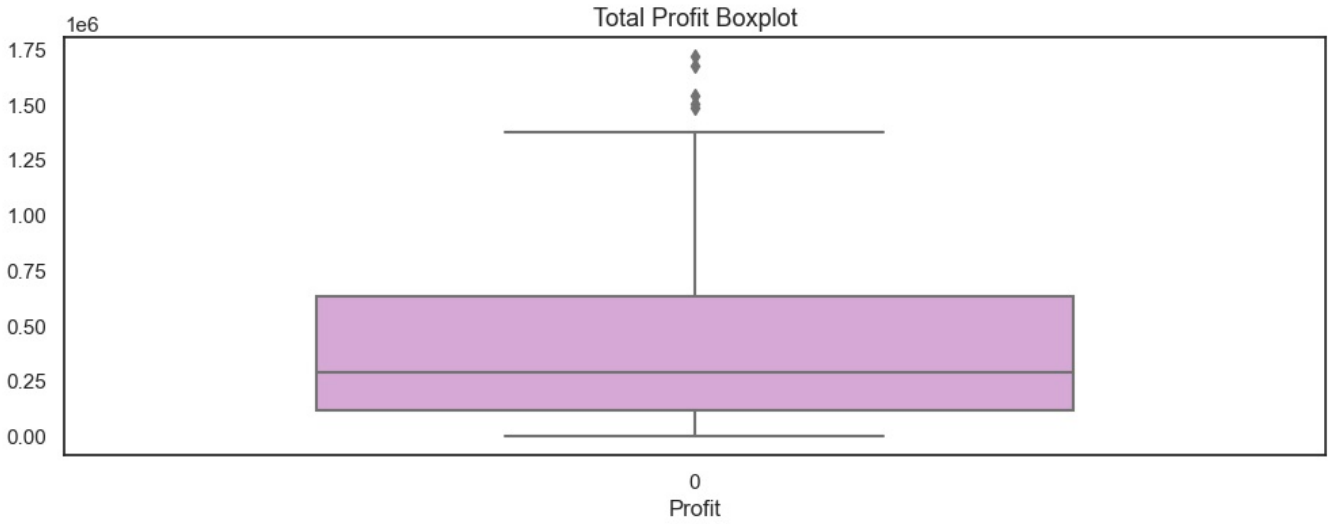
In [40]:
```
revenue_by_category = df.groupby('Item Type')['Total Revenue'].sum().sort_values(ascending=False)
revenue_by_category
```

Out[40]:
```
Item Type
Cosmetics          36601509.60
Office Supplies    30585380.07
Household          29889712.29
Baby Food          10350327.60
Clothes             7787292.80
Cereal              5322898.90
Meat                4503675.75
Personal Care       3980904.84
Vegetables          3089057.06
Beverages           2690794.60
Snacks              2080733.46
Fruits               466481.34
Name: Total Revenue, dtype: float64
```

In [39]:
```
profit_by_category = df.groupby('Item Type')['Total Profit'].sum().sort_values(ascending=False)
profit_by_category
```

Out[39]:
```
Item Type
Cosmetics          14556048.66
Household           7412605.71
Office Supplies     5929583.75
Clothes             5233334.40
Baby Food           3886643.70
Cereal              2292443.43
Vegetables          1265819.63
Personal Care       1220622.48
Beverages            888047.28
Snacks               751944.18
Meat                 610610.00
Fruits               120495.18
Name: Total Profit, dtype: float64
```

In [41]:
```
sns.set(style='white')
fig, ax = plt.subplots(figsize=(12, 4))
sns.boxplot(df['Total Profit'], color="plum", width=.6)
plt.title('Total Profit Boxplot', fontsize=13)
plt.xlabel('Profit')
plt.show()
```



In [50]:
```
def detect_outliers(df, column):
    threshold = 2
    mean = np.mean(column)
    std = np.std(column)
    outliers = []

    for i, value in enumerate(column):
        z_score = (value - mean) / std
        if np.abs(z_score) > threshold:
            outliers.append(i)
            print(df.loc[i])

    return outliers
```

In [51]:
```
outliers = detect_outliers(df, df["Total Profit"])
```

```
Region            Central America and the Caribbean
Country                                    Honduras
Item Type                                 Household
Sales Channel                               Offline
Order Priority                                    H
Order ID                                  522840487
Ship Date                                 2/13/2017
Units Sold                                     8974
Unit Price                                   668.27
Unit Cost                                    502.54
Total Revenue                            5997054.98
Total Cost                               4509793.96
Total Profit                             1487261.02
Order_Date                      2017-02-08 00:00:00
Ship_Date                       2017-02-13 00:00:00
os_lead_time                        5 days 00:00:00
Order Month                                       2
Order Year                                     2017
Name: 13, dtype: object
Region                         Europe
Country                   Switzerland
Item Type                   Cosmetics
Sales Channel                 Offline
Order Priority                      M
Order ID                    249693334
Ship Date                  10/20/2012
Units Sold                       8661
Unit Price                      437.2
Unit Cost                      263.33
Total Revenue               3786589.2
Total Cost                 2280701.13
Total Profit               1505888.07
Order_Date        2012-09-17 00:00:00
Ship_Date         2012-10-20 00:00:00
os_lead_time         33 days 00:00:00
Order Month                         9
Order Year                       2012
Name: 30, dtype: object
Region                           Asia
Country                       Myanmar
Item Type                   Household
Sales Channel                 Offline
Order Priority                      H
Order ID                    177713572
Ship Date                    3/1/2015
Units Sold                       8250
Unit Price                     668.27
Unit Cost                      502.54
Total Revenue               5513227.5
Total Cost                  4145955.0
Total Profit                1367272.5
Order_Date        2015-01-16 00:00:00
Ship_Date         2015-03-01 00:00:00
os_lead_time         44 days 00:00:00
Order Month                         1
Order Year                       2015
Name: 33, dtype: object
Region                         Europe
Country                       Iceland
Item Type                   Cosmetics
Sales Channel                  Online
Order Priority                      C
Order ID                    331438481
Ship Date                  12/31/2016
Units Sold                       8867
Unit Price                      437.2
Unit Cost                      263.33
Total Revenue               3876652.4
Total Cost                 2334947.11
Total Profit               1541705.29
Order_Date        2016-12-31 00:00:00
Ship_Date         2016-12-31 00:00:00
os_lead_time          0 days 00:00:00
Order Month                        12
Order Year                       2016
Name: 46, dtype: object
Region            Middle East and North Africa
Country                               Pakistan
Item Type                            Cosmetics
Sales Channel                          Offline
Order Priority                               L
Order ID                             231145322
Ship Date                            8/16/2013
Units Sold                                9892
Unit Price                               437.2
Unit Cost                               263.33
Total Revenue                        4324782.4
Total Cost                          2604860.36
Total Profit                        1719922.04
Order_Date                 2013-07-05 00:00:00
Ship_Date                  2013-08-16 00:00:00
os_lead_time                  42 days 00:00:00
Order Month                                  7
Order Year                                2013
Name: 74, dtype: object
Region            Australia and Oceania
Country                          Samoa 
Item Type                     Cosmetics
Sales Channel                    Online
Order Priority                        H
Order ID                      670854651
Ship Date                      8/7/2013
Units Sold                         9654
Unit Price                        437.2
Unit Cost                        263.33
Total Revenue                 4220728.8
Total Cost                   2542187.82
Total Profit                 1678540.98
Order_Date          2013-07-20 00:00:00
Ship_Date           2013-08-07 00:00:00
os_lead_time           18 days 00:00:00
Order Month                           7
Order Year                         2013
Name: 79, dtype: object
Region                         Europe
Country                       Romania
Item Type                   Cosmetics
```

```
Sales Channel                      Online
Order Priority                          H
Order ID                        660643374
Ship Date                      12/25/2010
Units Sold                           7910
Unit Price                          437.2
Unit Cost                          263.33
Total Revenue                   3458252.0
Total Cost                      2082940.3
Total Profit                    1375311.7
Order_Date            2010-11-26 00:00:00
Ship_Date             2010-12-25 00:00:00
os_lead_time             29 days 00:00:00
Order Month                            11
Order Year                           2010
Name: 93, dtype: object
```

In [52]:

```
print(outliers)
```

```
[13, 30, 33, 46, 74, 79, 93]
```

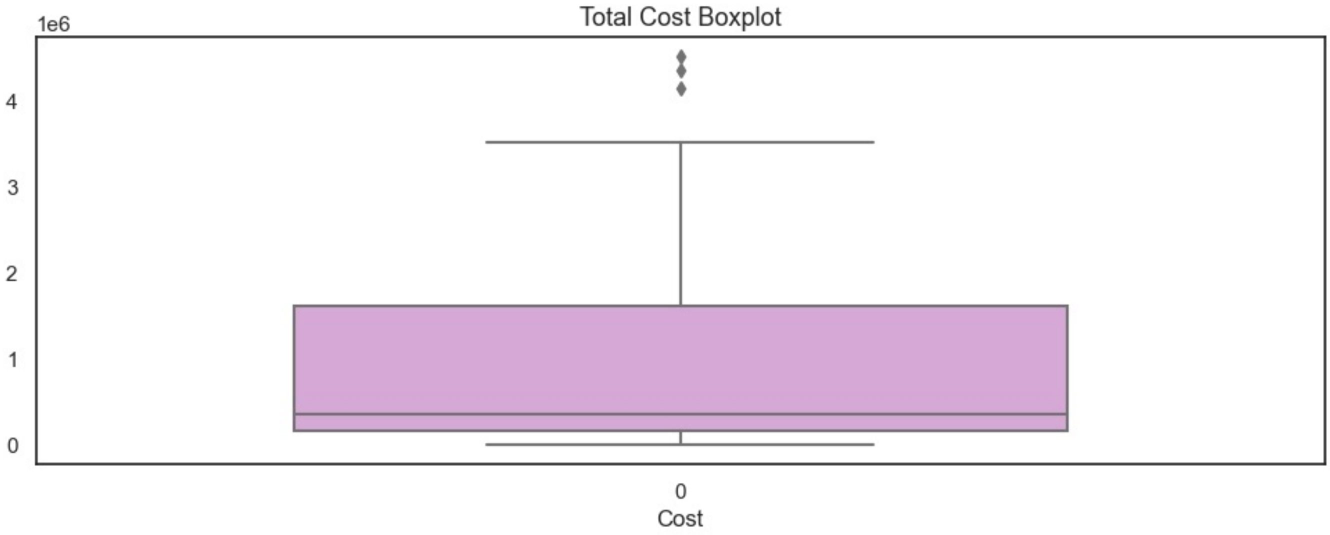
In [53]:

```
total_outlier= len(outliers)
print("The list has", total_outlier , "outliers in Total Profit column of dataset ")
```

```
The list has 7 outliers in Total Profit column of dataset
```

In [54]:

```
sns.set(style='white')
fig, ax = plt.subplots(figsize=(12, 4))
sns.boxplot(df['Total Cost'], color="plum", width=.6)
plt.title('Total Cost Boxplot', fontsize=13)
plt.xlabel('Cost')
plt.show()
```



In [55]:

```
def detect_outliers(df, column):
    threshold = 2
    mean = np.mean(column)
    std = np.std(column)
    outliers = []

    for i, value in enumerate(column):
        z_score = (value - mean) / std
        if np.abs(z_score) > threshold:
            outliers.append(i)
            print(df.loc[i])

    return outliers
```

In [56]:

```
outliers = detect_outliers(df, df["Total Cost"])
```

```
Region            Central America and the Caribbean
Country                                    Honduras
Item Type                                 Household
Sales Channel                               Offline
Order Priority                                    H
Order ID                                  522840487
Ship Date                                 2/13/2017
Units Sold                                     8974
Unit Price                                   668.27
Unit Cost                                    502.54
Total Revenue                            5997054.98
Total Cost                               4509793.96
Total Profit                             1487261.02
Order_Date                      2017-02-08 00:00:00
Ship_Date                       2017-02-13 00:00:00
os_lead_time                      5 days 00:00:00
Order Month                                       2
Order Year                                     2017
Name: 13, dtype: object
Region                           Asia
Country                       Myanmar
Item Type                   Household
Sales Channel                 Offline
Order Priority                      H
Order ID                    177713572
Ship Date                    3/1/2015
Units Sold                       8250
Unit Price                     668.27
Unit Cost                      502.54
Total Revenue               5513227.5
Total Cost                  4145955.0
Total Profit                1367272.5
Order_Date        2015-01-16 00:00:00
Ship_Date         2015-03-01 00:00:00
os_lead_time         44 days 00:00:00
Order Month                         1
Order Year                       2015
Name: 33, dtype: object
Region                           Asia
Country                        Brunei
Item Type             Office Supplies
Sales Channel                  Online
Order Priority                      L
Order ID                    320009267
Ship Date                    5/8/2012
Units Sold                       6708
Unit Price                     651.21
Unit Cost                      524.96
Total Revenue              4368316.68
Total Cost                 3521431.68
Total Profit                 846885.0
Order_Date        2012-04-01 00:00:00
Ship_Date         2012-05-08 00:00:00
os_lead_time         37 days 00:00:00
Order Month                         4
Order Year                       2012
Name: 38, dtype: object
Region                         Europe
Country                     Lithuania
Item Type             Office Supplies
Sales Channel                 Offline
Order Priority                      H
Order ID                    166460740
Ship Date                  11/17/2010
Units Sold                       8287
Unit Price                     651.21
Unit Cost                      524.96
Total Revenue              5396577.27
Total Cost                 4350343.52
Total Profit               1046233.75
Order_Date        2010-10-24 00:00:00
Ship_Date         2010-11-17 00:00:00
os_lead_time         24 days 00:00:00
Order Month                        10
Order Year                       2010
Name: 68, dtype: object
Region                  North America
Country                        Mexico
Item Type                   Household
Sales Channel                 Offline
Order Priority                      C
Order ID                    986435210
Ship Date                  12/12/2014
Units Sold                       6954
Unit Price                     668.27
Unit Cost                      502.54
Total Revenue              4647149.58
Total Cost                 3494663.16
Total Profit               1152486.42
Order_Date        2014-11-06 00:00:00
Ship_Date         2014-12-12 00:00:00
os_lead_time         36 days 00:00:00
Order Month                        11
Order Year                       2014
Name: 75, dtype: object
```

In [57]:

```
print(outliers)
```

```
[13, 33, 38, 68, 75]
```

In [58]:

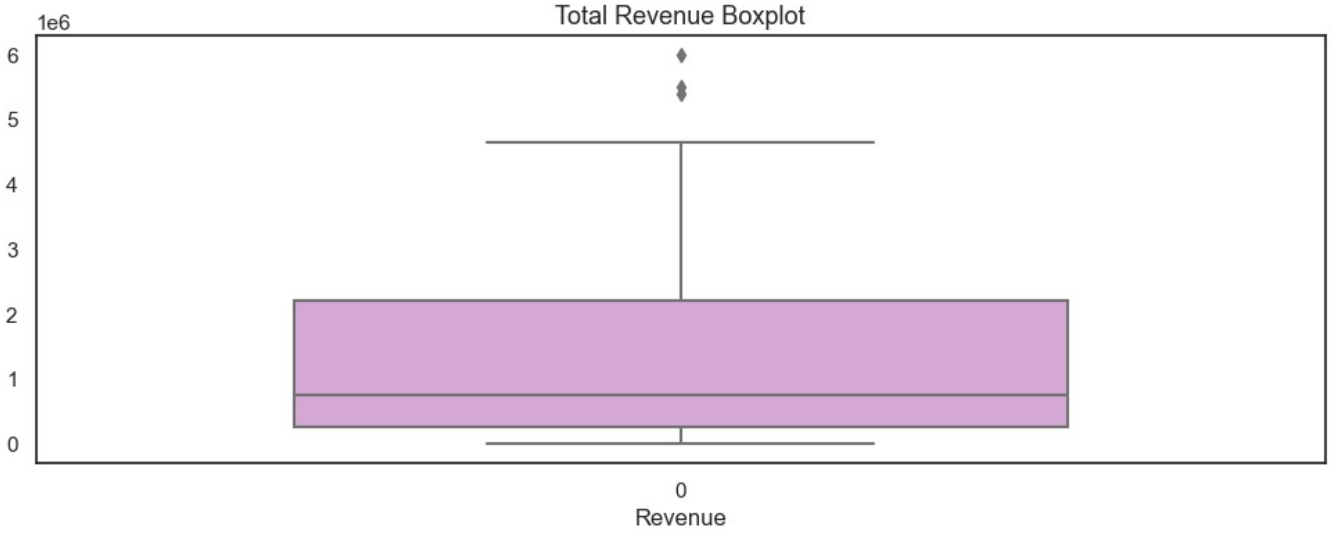
```
total_outlier= len(outliers)
print("The list has", total_outlier , "outliers in Total Cost column of dataset ")
```

```
The list has 5 outliers in Total Cost column of dataset
```

In [59]:

```
sns.set(style='white')
```

```
fig, ax = plt.subplots(figsize=(12, 4))
sns.boxplot(df['Total Revenue'], color="plum", width=.6)
plt.title('Total Revenue Boxplot', fontsize=13)
plt.xlabel('Revenue')
plt.show()
```



In [60]:

```
def detect_outliers(df, column):
    threshold = 2 ## 3rd standard deviation
    mean = np.mean(column)
    std = np.std(column)
    outliers = []
    for i, value in enumerate(column):
        z_score = (value - mean) / std
        if np.abs(z_score) > threshold:
            outliers.append(i)
            print(df.loc[i])
    return outliers
```

In [61]:

```
outliers = detect_outliers(df, df["Total Revenue"])
```

```
Region            Central America and the Caribbean
Country                                    Honduras
Item Type                                 Household
Sales Channel                               Offline
Order Priority                                    H
Order ID                                  522840487
Ship Date                                 2/13/2017
Units Sold                                     8974
Unit Price                                   668.27
Unit Cost                                    502.54
Total Revenue                            5997054.98
Total Cost                               4509793.96
Total Profit                             1487261.02
Order_Date                      2017-02-08 00:00:00
Ship_Date                       2017-02-13 00:00:00
os_lead_time                        5 days 00:00:00
Order Month                                       2
Order Year                                     2017
Name: 13, dtype: object
Region                           Asia
Country                       Myanmar
Item Type                   Household
Sales Channel                 Offline
Order Priority                      H
Order ID                    177713572
Ship Date                    3/1/2015
Units Sold                       8250
Unit Price                     668.27
Unit Cost                      502.54
Total Revenue               5513227.5
Total Cost                  4145955.0
Total Profit                1367272.5
Order_Date        2015-01-16 00:00:00
Ship_Date         2015-03-01 00:00:00
os_lead_time         44 days 00:00:00
Order Month                         1
Order Year                       2015
Name: 33, dtype: object
Region                           Asia
Country                        Brunei
Item Type             Office Supplies
Sales Channel                  Online
Order Priority                      L
Order ID                    320009267
Ship Date                    5/8/2012
```

```
Units Sold                       6708
Unit Price                     651.21
Unit Cost                      524.96
Total Revenue              4368316.68
Total Cost                 3521431.68
Total Profit                 846885.0
Order_Date        2012-04-01 00:00:00
Ship_Date         2012-05-08 00:00:00
os_lead_time         37 days 00:00:00
Order Month                         4
Order Year                       2012
Name: 38, dtype: object
Region                         Europe
Country                     Lithuania
Item Type             Office Supplies
Sales Channel                 Offline
Order Priority                      H
Order ID                    166460740
Ship Date                  11/17/2010
Units Sold                       8287
Unit Price                     651.21
Unit Cost                      524.96
Total Revenue              5396577.27
Total Cost                 4350343.52
Total Profit               1046233.75
Order_Date        2010-10-24 00:00:00
Ship_Date         2010-11-17 00:00:00
os_lead_time         24 days 00:00:00
Order Month                        10
Order Year                       2010
Name: 68, dtype: object
Region            Middle East and North Africa
Country                               Pakistan
Item Type                            Cosmetics
Sales Channel                          Offline
Order Priority                               L
Order ID                             231145322
Ship Date                            8/16/2013
Units Sold                                9892
Unit Price                               437.2
Unit Cost                               263.33
Total Revenue                        4324782.4
Total Cost                          2604860.36
Total Profit                        1719922.04
Order_Date                 2013-07-05 00:00:00
Ship_Date                  2013-08-16 00:00:00
os_lead_time                  42 days 00:00:00
Order Month                                  7
Order Year                                2013
Name: 74, dtype: object
Region                  North America
Country                        Mexico
Item Type                   Household
Sales Channel                 Offline
Order Priority                      C
Order ID                    986435210
Ship Date                  12/12/2014
Units Sold                       6954
Unit Price                     668.27
Unit Cost                      502.54
Total Revenue              4647149.58
Total Cost                 3494663.16
Total Profit               1152486.42
Order_Date        2014-11-06 00:00:00
Ship_Date         2014-12-12 00:00:00
os_lead_time         36 days 00:00:00
Order Month                        11
Order Year                       2014
Name: 75, dtype: object
```

In [62]:

```
print(outliers)
```

```
[13, 33, 38, 68, 74, 75]
```

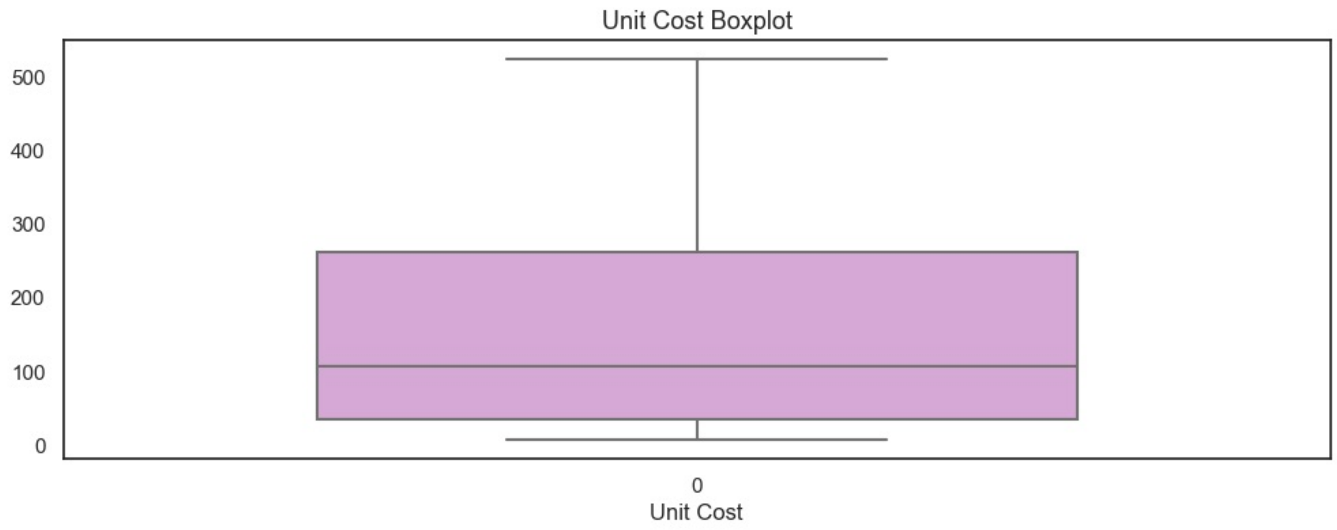
In [67]:

```
total_outlier= len(outliers)
print("The list has", total_outlier , "outliers in Total Revenue column of dataset ")
```

```
The list has 6 outliers in Total Revenue column of dataset
```

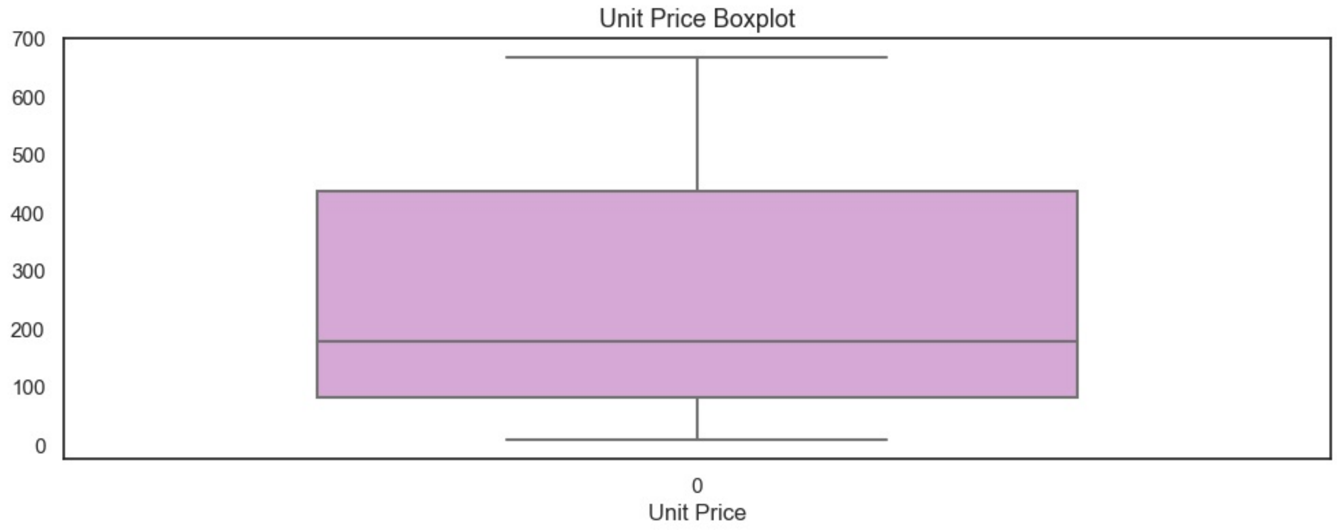
In [68]:

```
sns.set(style='white')
fig, ax = plt.subplots(figsize=(12, 4))
sns.boxplot(df['Unit Cost'], color="plum", width=.6)
plt.title('Unit Cost Boxplot', fontsize=13)
plt.xlabel('Unit Cost')
plt.show()
```



In [70]:

```
sns.set(style='white')
fig, ax = plt.subplots(figsize=(12, 4))
sns.boxplot(df['Unit Price'], color="plum", width=.6)
plt.title('Unit Price Boxplot', fontsize=13)
plt.xlabel('Unit Price')
plt.show()
```



In [71]:

```
revenue_by_category = df.groupby('Item Type')['Total Revenue'].sum().sort_values(ascending=False)
revenue_by_category
```

Out[71]:

```
Item Type
Cosmetics          36601509.60
Office Supplies    30585380.07
Household          29889712.29
Baby Food          10350327.60
Clothes             7787292.80
Cereal              5322898.90
Meat                4503675.75
Personal Care       3980904.84
Vegetables          3089057.06
Beverages           2690794.60
Snacks              2080733.46
Fruits               466481.34
Name: Total Revenue, dtype: float64
```

In [72]:

```
profit_by_category = df.groupby('Item Type')['Total Profit'].sum().sort_values(ascending=False)
profit_by_category
```

Out[72]:

```
Item Type
Cosmetics          14556048.66
Household           7412605.71
Office Supplies     5929583.75
Clothes             5233334.40
Baby Food           3886643.70
Cereal              2292443.43
Vegetables          1265819.63
Personal Care       1220622.48
Beverages            888047.28
Snacks               751944.18
Meat                 610610.00
Fruits               120495.18
Name: Total Profit, dtype: float64
```

In [73]:

```
print(df[['Total Revenue', 'Total Cost', 'Total Profit']].corr())
```

In [75]:

```
print(df[['Total Revenue', 'Total Cost', 'Total Profit']].corr())
```

```
               Total Revenue  Total Cost  Total Profit
Total Revenue       1.000000    0.983928      0.897327
Total Cost          0.983928    1.000000      0.804091
Total Profit        0.897327    0.804091      1.000000
```

In [86]:

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df["Item Type"] = le.fit_transform(df["Item Type"])
df["Sales Channel"] = le.fit_transform(df["Sales Channel"])
df["Order Priority"] = le.fit_transform(df["Order Priority"])
```

In [87]:

```
df.head()
```

Out[87]:

	Item Type	Sales Channel	Order Priority	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit	Order_Date	Ship_Date	os_lead_time	Order Month	Order Year
0	0	0	1	9925	255.28	159.42	2533654.00	1582243.50	951410.50	2010-05-28	2010-06-27	30 days	5	2010
1	2	1	0	2804	205.70	117.11	576782.80	328376.44	248406.36	2012-08-22	2012-09-15	24 days	8	2012
2	8	0	2	1779	651.21	524.96	1158502.59	933903.84	224598.75	2014-05-02	2014-05-08	6 days	5	2014
3	5	1	0	8102	9.33	6.92	75591.66	56065.84	19525.82	2014-06-20	2014-07-05	15 days	6	2014
4	8	0	2	5062	651.21	524.96	3296425.02	2657347.52	639077.50	2013-02-01	2013-02-06	5 days	2	2013

In [88]:

```
X = df[['Item Type', 'Sales Channel', 'Order Priority', 'Units Sold', 'Unit Price', 'Unit Cost', 'Total Revenue
y = df['Total Profit']
```

In [89]:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

In [90]:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
```

In [91]:

```
X_train = scaler.fit_transform(X_train)
```

In [92]:

```
X_test = scaler.transform(X_test)
```

In [93]:

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score
model = LinearRegression()
model.fit(X_train,y_train)
```

Out[93]:

```
▾ LinearRegression
LinearRegression()
```

In [94]:

```
model_pred = model.predict(X_test)
```

In [95]:

```
model_pred
```

Out[95]:

```
array([2.25246900e+05, 4.36446250e+05, 6.32512500e+05, 8.52235800e+04,
       4.35499200e+05, 5.03358750e+05, 1.22686500e+05, 7.47939490e+05,
       7.82812000e+03, 9.51410500e+05, 6.34745900e+05, 1.50588807e+06,
       7.66835040e+05, 1.36727250e+06, 1.19685000e+05, 6.39077500e+05,
       1.46875140e+05, 2.35601160e+05, 6.06834720e+05, 5.32525000e+04,
       2.55718080e+05, 1.25802000e+03, 1.30091800e+04, 1.87545030e+05,
       5.39196480e+05, 1.22865120e+05, 7.55559000e+04, 1.51034700e+04,
       5.84073870e+05, 8.99040600e+04])
```

In [96]:

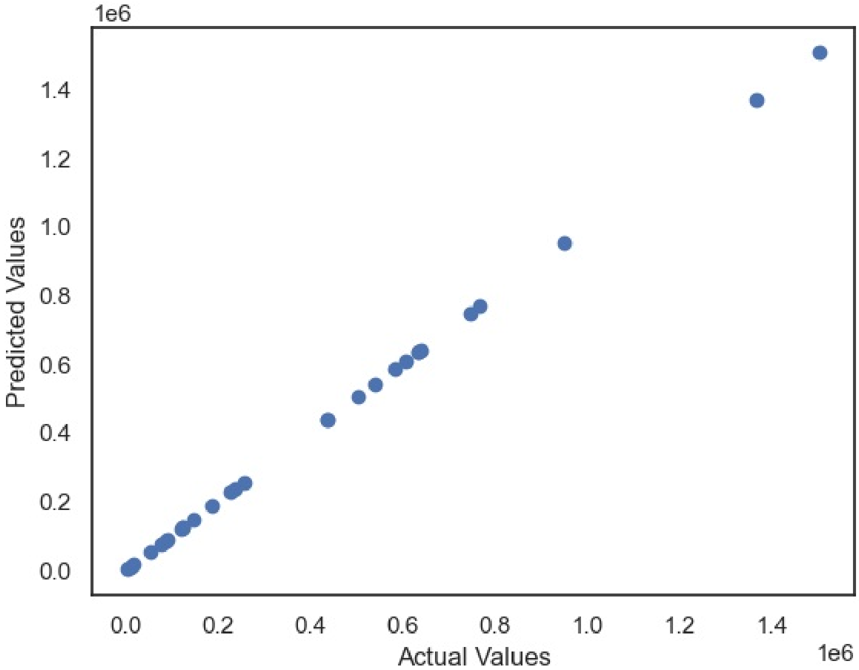
```
from sklearn.metrics import r2_score
score = r2_score(model_pred,y_test)

accuracy_pct = score * 100
print("Accuracy: {:.2f}%".format(accuracy_pct))
```

```
Accuracy: 100.00%
```

In [97]:

```
import matplotlib.pyplot as plt
plt.scatter(y_test, model_pred)
plt.xlabel('Actual Values')
plt.ylabel('Predicted Values')
plt.show()
```



```
In [ ]:
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js