Analyzing Amazon Sales data

December 31, 2023

Importing the required libraries

[1]:
```
import numpy as np
import pandas as pd
import seaborn as sn
import matplotlib.pyplot as mp
import warnings
warnings.filterwarnings('ignore')
```

Reading the data

[2]:
```
data = pd.read_csv(r'C:\Users\DELL\Desktop\Internship\Amazon Sales data.csv')
```

Trying to view the top 10 rows of the data

[3]:
```
data.head(10)
```

[3]:

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	9925
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2804
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	1779
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2014	8102
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2013	5062
5	Australia and Oceania	Solomon Islands	Baby Food	Online	C	2/4/2015	547995746	2/21/2015	2974
6	Sub-Saharan Africa	Angola	Household	Offline	M	4/23/2011	135425221	4/27/2011	4187
7	Sub-Saharan Africa	Burkina Faso	Vegetables	Online	H	7/17/2012	871543967	7/27/2012	8082
8	Sub-Saharan Africa	Republic of the Congo	Personal Care	Offline	M	7/14/2015	770463311	8/25/2015	6070
9	Sub-Saharan Africa	Senegal	Cereal						

```
9        Online              H  4/18/2014  616607081  5/30/2014        6593

   Unit Price  Unit Cost  Total Revenue  Total Cost  Total Profit
0      255.28     159.42     2533654.00  1582243.50     951410.50
1      205.70     117.11      576782.80   328376.44     248406.36
2      651.21     524.96     1158502.59   933903.84     224598.75
3        9.33       6.92       75591.66    56065.84      19525.82
4      651.21     524.96     3296425.02  2657347.52     639077.50
5      255.28     159.42      759202.72   474115.08     285087.64
6      668.27     502.54     2798046.49  2104134.98     693911.51
7      154.06      90.93     1245112.92   734896.26     510216.66
8       81.73      56.67      496101.10   343986.90     152114.20
9      205.70     117.11     1356180.10   772106.23     584073.87
```

Checking the data set number of rows and columns

[4]:
```
data.shape
```

[4]:
```
(100, 14)
```

Rearanging the columns

[5]:
```
data = data[["Order Date","Order ID","Order Priority","Ship Date","Item␣
↪Type","Region","Country","Sales Channel","Units Sold","Unit Price","Unit␣
↪Cost","Total Revenue","Total Cost","Total Profit"]]
```

Checking whether columns re rearranged or not by by viewing the top 10 rows of the data

[6]:
```
data.head(10)
```

[6]:
```
  Order Date   Order ID Order Priority  Ship Date        Item Type  \
0  5/28/2010  669165933              H  6/27/2010        Baby Food
1  8/22/2012  963881480              C  9/15/2012           Cereal
2   5/2/2014  341417157              L   5/8/2014  Office Supplies
3  6/20/2014  514321792              C   7/5/2014           Fruits
4   2/1/2013  115456712              L   2/6/2013  Office Supplies
5   2/4/2015  547995746              C  2/21/2015        Baby Food
6  4/23/2011  135425221              M  4/27/2011        Household
7  7/17/2012  871543967              H  7/27/2012       Vegetables
8  7/14/2015  770463311              M  8/25/2015    Personal Care
9  4/18/2014  616607081              H  5/30/2014           Cereal

                           Region                Country Sales Channel  \
0              Australia and Oceania                 Tuvalu       Offline
1  Central America and the Caribbean                Grenada        Online
2                             Europe                 Russia       Offline
3                Sub-Saharan Africa  Sao Tome and Principe        Online
4                Sub-Saharan Africa                 Rwanda       Offline
5              Australia and Oceania        Solomon Islands        Online
```

```
6                 Sub-Saharan Africa                 Angola       Offline
7                 Sub-Saharan Africa           Burkina Faso        Online
8                 Sub-Saharan Africa  Republic of the Congo       Offline
9                 Sub-Saharan Africa                Senegal        Online

   Units Sold  Unit Price  Unit Cost  Total Revenue  Total Cost  Total Profit
0        9925      255.28     159.42     2533654.00  1582243.50     951410.50
1        2804      205.70     117.11      576782.80   328376.44     248406.36
2        1779      651.21     524.96     1158502.59   933903.84     224598.75
3        8102        9.33       6.92       75591.66    56065.84      19525.82
4        5062      651.21     524.96     3296425.02  2657347.52     639077.50
5        2974      255.28     159.42      759202.72   474115.08     285087.64
6        4187      668.27     502.54     2798046.49  2104134.98     693911.51
7        8082      154.06      90.93     1245112.92   734896.26     510216.66
8        6070       81.73      56.67      496101.10   343986.90     152114.20
9        6593      205.70     117.11     1356180.10   772106.23     584073.87
```

[7]:
```
#Checking the number of horizontal rows in the data.
data.axes[0]
```

[7]:
```
RangeIndex(start=0, stop=100, step=1)
```

[8]:
```
#Checking the data types of the each columns as each row contains different␣
 ↪data types
data.dtypes
```

[8]:
```
Order Date         object
Order ID            int64
Order Priority     object
Ship Date          object
Item Type          object
Region             object
Country            object
Sales Channel      object
Units Sold          int64
Unit Price        float64
Unit Cost         float64
Total Revenue     float64
Total Cost        float64
Total Profit      float64
dtype: object
```

[9]:
```
#Checking whether any columns contain null values
data.columns.isnull()
```

[9]:
```
array([False, False, False, False, False, False, False, False, False,
       False, False, False, False, False])
```

[10]:
```
np.corrcoef(data.loc[:,'Total Revenue'].iloc[:],data.loc[:,'Total Profit'].
↪iloc[:])
```

[10]:
```
array([[1.        , 0.89732687],
       [0.89732687, 1.        ]])
```

[11]:
```
data.set_index('Order ID',inplace=True)
```

[12]:
```
data.head(10)
```

[12]:

Order ID	Order Date	Order Priority	Ship Date	Item Type	Region	Country	Sales Channel	Units Sold	Unit Price	Unit Cost	Total Revenue
669165933	5/28/2010	H	6/27/2010	Baby Food	Australia and Oceania	Tuvalu	Offline	9925	255.28	159.42	2533654.00
963881480	8/22/2012	C	9/15/2012	Cereal	Central America and the Caribbean	Grenada	Online	2804	205.70	117.11	576782.80
341417157	5/2/2014	L	5/8/2014	Office Supplies	Europe	Russia	Offline	1779	651.21	524.96	1158502.59
514321792	6/20/2014	C	7/5/2014	Fruits	Sub-Saharan Africa	Sao Tome and Principe	Online	8102	9.33	6.92	75591.66
115456712	2/1/2013	L	2/6/2013	Office Supplies	Sub-Saharan Africa	Rwanda	Offline	5062	651.21	524.96	3296425.02
547995746	2/4/2015	C	2/21/2015	Baby Food	Australia and Oceania	Solomon Islands	Online	2974	255.28	159.42	759202.72
135425221	4/23/2011	M	4/27/2011	Household	Sub-Saharan Africa	Angola	Offline	4187	668.27	502.54	2798046.49
871543967	7/17/2012	H	7/27/2012	Vegetables	Sub-Saharan Africa	Burkina Faso	Online	8082	154.06	90.93	1245112.92
770463311	7/14/2015	M	8/25/2015	Personal Care	Sub-Saharan Africa	Republic of the Congo	Offline	6070	81.73	56.67	496101.10
616607081	4/18/2014	H	5/30/2014	Cereal	Sub-Saharan Africa	Senegal					

```
616607081        Online        6593      205.70     117.11     1356180.10
```

```
           Total Cost  Total Profit
Order ID
669165933  1582243.50     951410.50
963881480   328376.44     248406.36
341417157   933903.84     224598.75
514321792    56065.84      19525.82
115456712  2657347.52     639077.50
547995746   474115.08     285087.64
135425221  2104134.98     693911.51
871543967   734896.26     510216.66
770463311   343986.90     152114.20
616607081   772106.23     584073.87
```

[13]:
```
data1 = data[["Units Sold","Unit Price","Unit Cost","Total Revenue","Total␣
↪Cost","Total Profit"]]
```

[14]:
```
#Checking the covariance between different factors
data1.cov()
```

[14]:
```
                 Units Sold    Unit Price     Unit Cost  Total Revenue  \
Units Sold     7.809144e+06 -4.640481e+04 -4.850918e+04   1.826973e+09
Unit Price    -4.640481e+04  5.550370e+04  4.377593e+04   2.587902e+08
Unit Cost     -4.850918e+04  4.377593e+04  3.542232e+04   1.966455e+08
Total Revenue  1.826973e+09  2.587902e+08  1.966455e+08   2.131684e+12
Total Cost     1.135124e+09  2.012054e+08  1.580833e+08   1.557145e+12
Total Profit   6.918495e+08  5.758482e+07  3.856216e+07   5.745386e+11

                 Total Cost  Total Profit
Units Sold     1.135124e+09  6.918495e+08
Unit Price     2.012054e+08  5.758482e+07
Unit Cost      1.580833e+08  3.856216e+07
Total Revenue  1.557145e+12  5.745386e+11
Total Cost     1.174922e+12  3.822231e+11
Total Profit   3.822231e+11  1.923155e+11
```

[15]:
```
#Checking correlation coefficient between different factors using the␣
↪"pearson"method
data1.corr(method='pearson')
```

[15]:
```
               Units Sold  Unit Price  Unit Cost  Total Revenue  Total Cost  \
Units Sold       1.000000   -0.070486  -0.092232       0.447784    0.374746
Unit Price      -0.070486    1.000000   0.987270       0.752360    0.787905
Unit Cost       -0.092232    0.987270   1.000000       0.715623    0.774895
Total Revenue    0.447784    0.752360   0.715623       1.000000    0.983928
Total Cost       0.374746    0.787905   0.774895       0.983928    1.000000
```

```
Total Profit     0.564550    0.557365   0.467214       0.897327    0.804091
```

```
               Total Profit
Units Sold         0.564550
Unit Price         0.557365
Unit Cost          0.467214
Total Revenue      0.897327
Total Cost         0.804091
Total Profit       1.000000
```

The high value of Pearson correlation coefficient between Total Revenue and Total Profit indicates that these two variables are closely related to each other. If revenue generated is high, then more profit will be generated and vice versa. The negative value of correlation coefficient among Units Sold and Unit Cost implies that quantity of products is inversely proportional to their cost. Same is the scenario with Units Sold and Units Price. Lesser the number of units of a product available, more will be it's price.

[16]:
```
#Calculating the average profit generated for a product
np.average(data1['Total Profit'])
```

[16]:
```
441681.98399999994
```

[17]:
```
#Calculating the maximum profit earned
np.max(data1['Total Profit'])
```

[17]:
```
1719922.04
```

[18]:
```
#Calculating the minimum profit earned
np.min(data1['Total Profit'])
```

[18]:
```
1258.02
```

[19]:
```
#Calculating the variance between the total Profit
np.var(data1['Total Profit'])
```

[19]:
```
190392340968.9648
```

[20]:
```
#Maximum and minimum profit generated are  1719922.04 and  1258.02
```

[21]:
```
np.max(data1['Total Revenue'])
```

[21]:
```
5997054.98
```

[22]:
```
np.min(data1['Total Revenue'])
```

[22]:
```
4870.26
```

[23]:
```
np.var(data1['Total Revenue'])
```

[23]: 2110366986501.2166

[24]:
```
np.percentile(data['Total Revenue'],50,axis=0,overwrite_input=True)
```

[24]: 752314.36

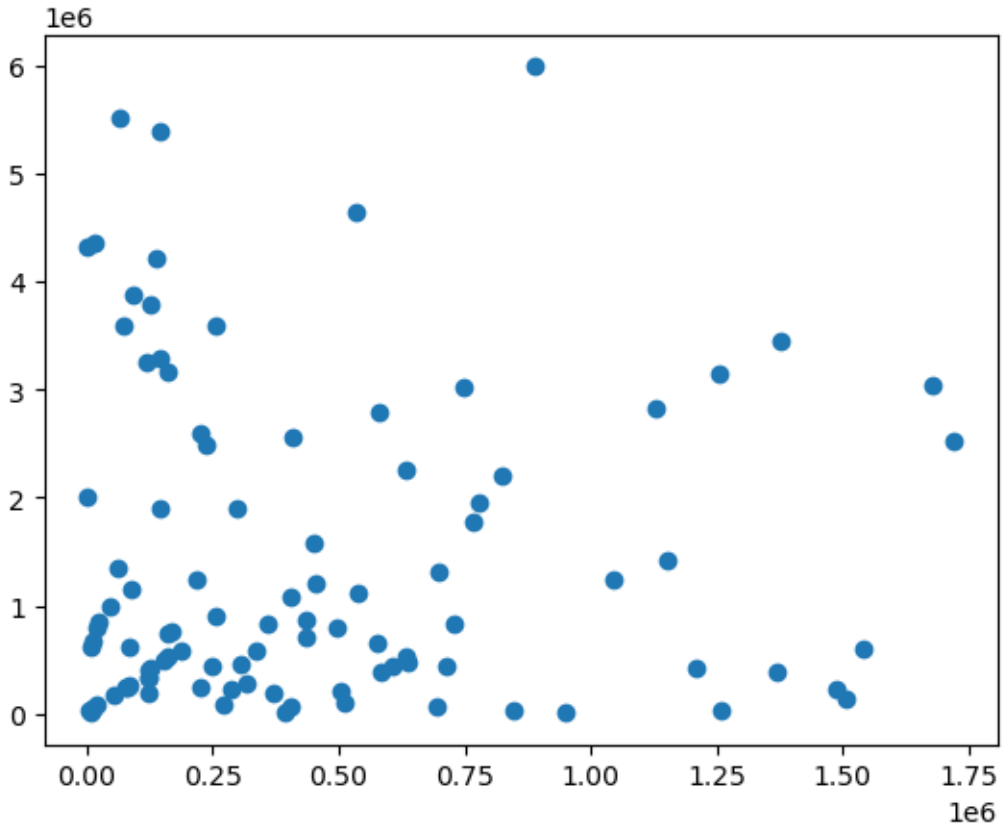
[25]:
```
np.median(data1['Total Revenue'])
```

[25]: 752314.36

Maximum and minimum revenue generated by the product are 5997054.98 and 4870.26. Revenue has very high variability in it's distribution. The median revenue generated is 752314.36.

[26]:
```
#Scatter plot between total profit and total revenue
mp.scatter(data1['Total Profit'],data['Total Revenue'])
```

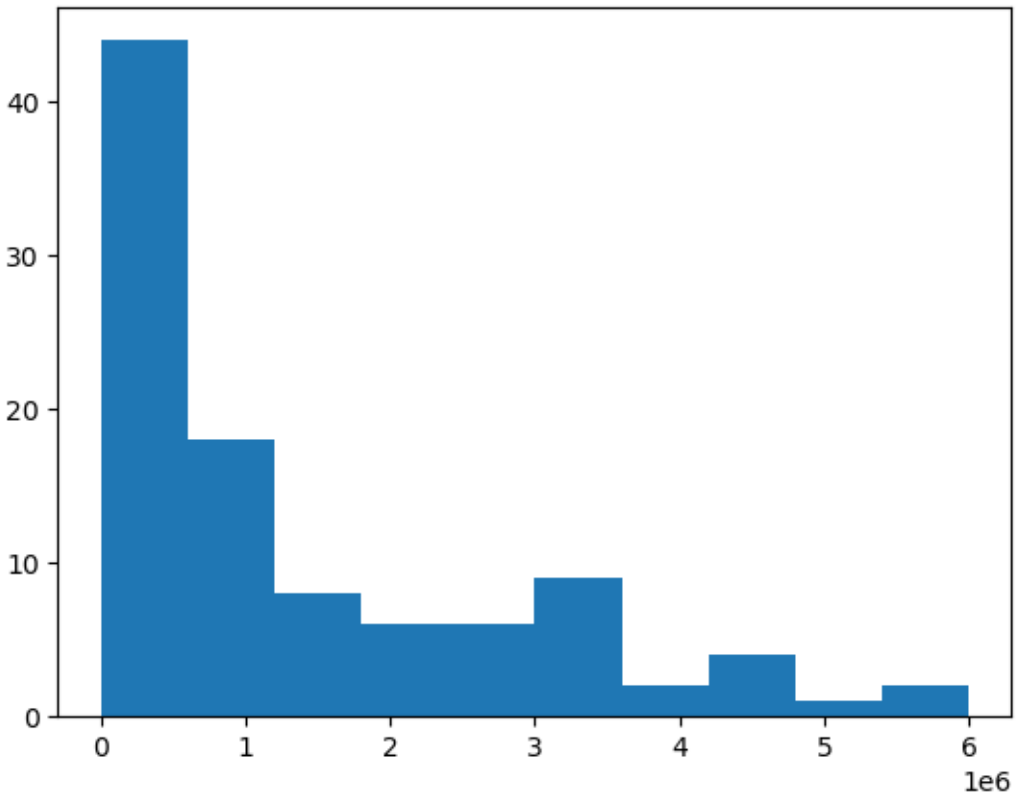
[26]: <matplotlib.collections.PathCollection at 0x218c45726d0>



The Above scatter plot also suggests that total profit and total revenue are directly proportional to each other.

[27]:
```
#hecking the total revenue in histogram graph
mp.hist(data1['Total Revenue'])
```

[27]:
```
(array([44., 18.,  8.,  6.,  6.,  9.,  2.,  4.,  1.,  2.]),
 array([4.87026000e+03, 6.04088732e+05, 1.20330720e+06, 1.80252568e+06,
        2.40174415e+06, 3.00096262e+06, 3.60018109e+06, 4.19939956e+06,
        4.79861804e+06, 5.39783651e+06, 5.99705498e+06]),
 <BarContainer object of 10 artists>)
```



[28]:
```
#Calculating the correlation coefficient between total profit and total revenue
np.correlate(data1['Total Revenue'],data1['Total Profit'])
```

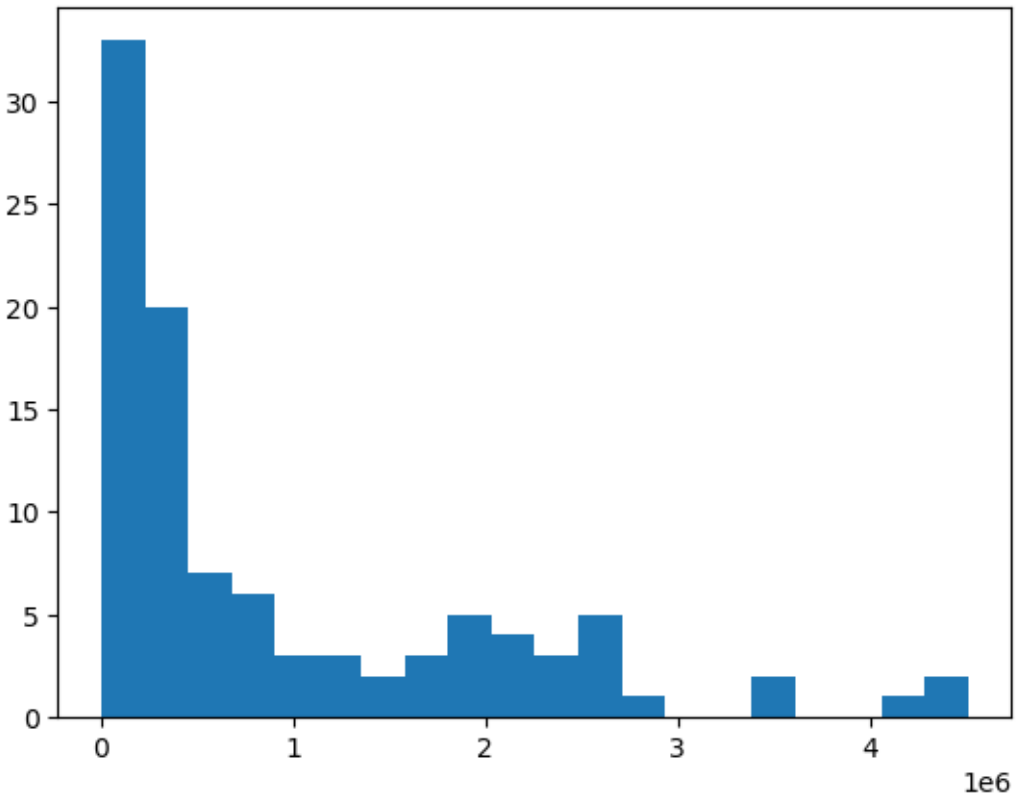
[28]:
```
array([1.17543797e+14])
```

[29]:
```
np.histogram(data1['Total Cost'],bins=10)
```

[29]:
```
(array([53, 13,  6,  5,  9,  8,  1,  2,  0,  3], dtype=int64),
 array([3.61224000e+03, 4.54230412e+05, 9.04848584e+05, 1.35546676e+06,
        1.80608493e+06, 2.25670310e+06, 2.70732127e+06, 3.15793944e+06,
        3.60855762e+06, 4.05917579e+06, 4.50979396e+06]))
```

[30]:
```
mp.hist(data1['Total Cost'],bins=20)
```

[30]:
```
(array([33., 20.,  7.,  6.,  3.,  3.,  2.,  3.,  5.,  4.,  3.,  5.,  1.,
         0.,  0.,  2.,  0.,  0.,  1.,  2.]),
 array([3.61224000e+03, 2.28921326e+05, 4.54230412e+05, 6.79539498e+05,
        9.04848584e+05, 1.13015767e+06, 1.35546676e+06, 1.58077584e+06,
        1.80608493e+06, 2.03139401e+06, 2.25670310e+06, 2.48201219e+06,
        2.70732127e+06, 2.93263036e+06, 3.15793944e+06, 3.38324853e+06,
        3.60855762e+06, 3.83386670e+06, 4.05917579e+06, 4.28448487e+06,
        4.50979396e+06]),
 <BarContainer object of 20 artists>)
```



[31]:
```
data1.corr(method='pearson')
```

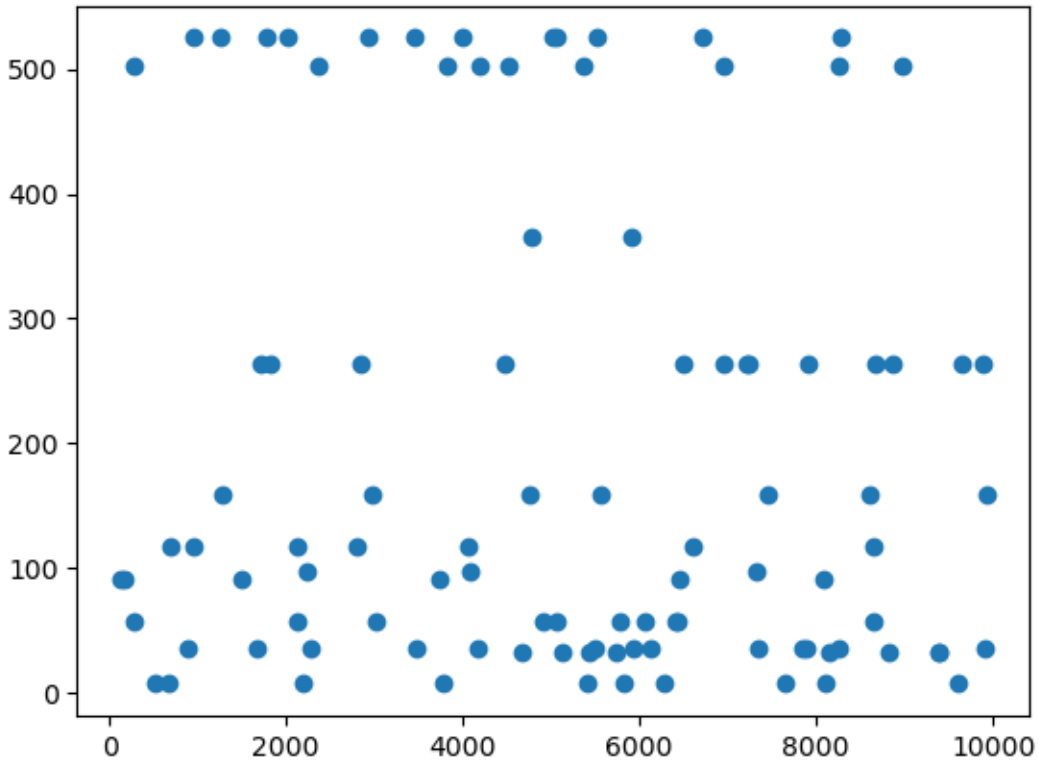
[31]:

	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	\
Units Sold	1.000000	-0.070486	-0.092232	0.447784	0.374746	
Unit Price	-0.070486	1.000000	0.987270	0.752360	0.787905	
Unit Cost	-0.092232	0.987270	1.000000	0.715623	0.774895	
Total Revenue	0.447784	0.752360	0.715623	1.000000	0.983928	
Total Cost	0.374746	0.787905	0.774895	0.983928	1.000000	
Total Profit	0.564550	0.557365	0.467214	0.897327	0.804091	

```
                Total Profit
Units Sold          0.564550
Unit Price          0.557365
Unit Cost           0.467214
Total Revenue       0.897327
Total Cost          0.804091
Total Profit        1.000000
```

[32]:
```
mp.scatter(data1['Units Sold'],data1['Unit Cost'])
```

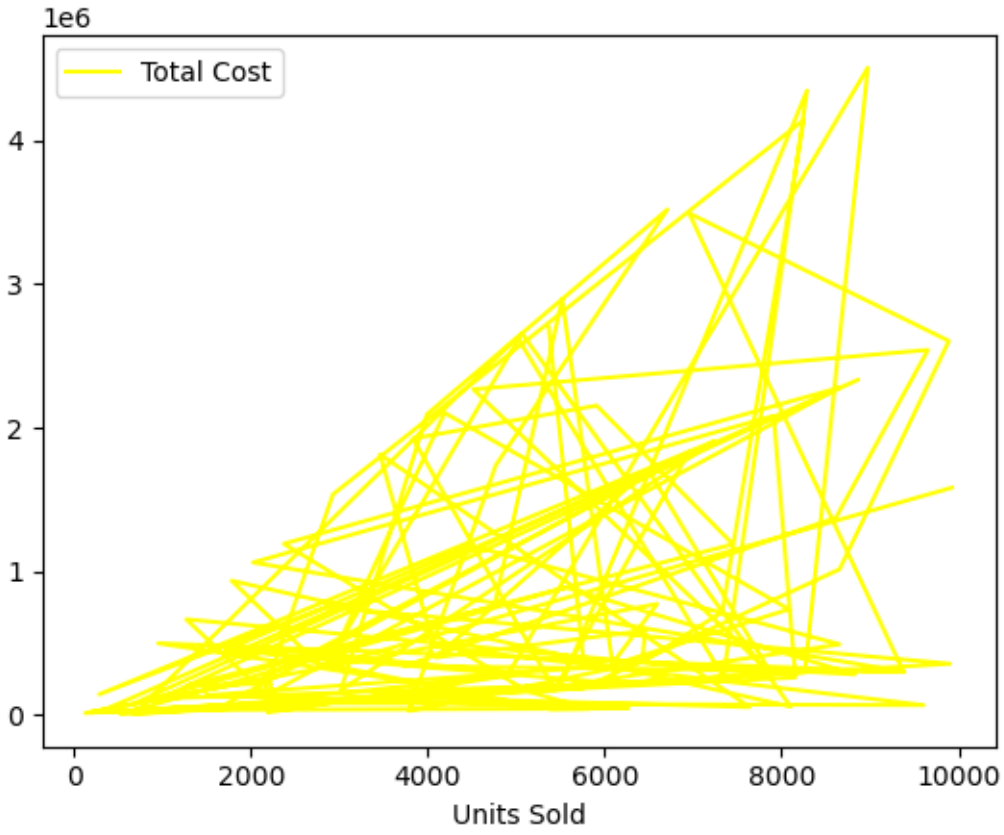
[32]:
```
<matplotlib.collections.PathCollection at 0x218c50fc690>
```



The above scatter plot implies that the two variables 'Units Sold' and 'Unit Cost' are inversely proportional to each other to some extent. When more units of a product are sold, the unit cost of that product becomes lesser and vice versa.

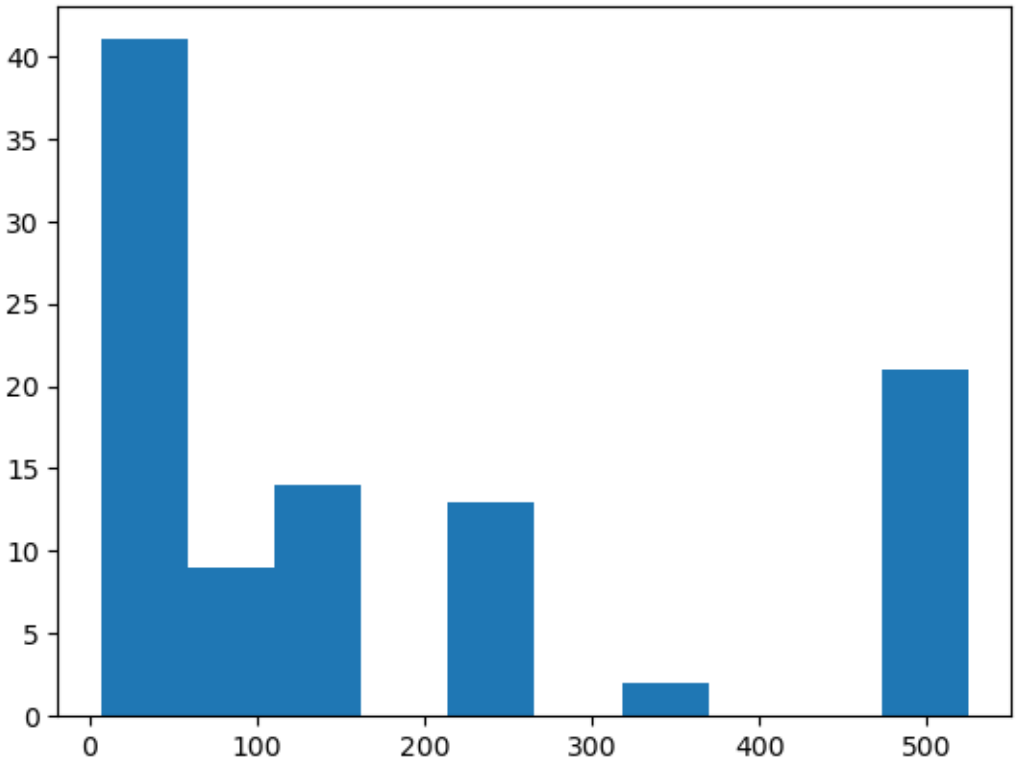
[33]:
```
data1.plot.line(x='Units Sold',y='Total Cost',subplots=True,color={'Total Cost':
↪ 'yellow'})
```

[33]:
```
array([<Axes: xlabel='Units Sold'>], dtype=object)
```



[34]:
```
mp.hist(data1['Unit Cost'])
```

[34]:
```
(array([41.,  9., 14.,  0., 13.,  0.,  2.,  0.,  0., 21.]),
 array([  6.92 ,  58.724, 110.528, 162.332, 214.136, 265.94 , 317.744,
        369.548, 421.352, 473.156, 524.96 ]),
 <BarContainer object of 10 artists>)
```



[35]:
```
np.min(data1['Unit Cost'])
```

[35]: 6.92

[36]:
```
np.max(data1['Unit Cost'])
```

[36]: 524.96

[37]:
```
np.mean(data1['Unit Cost'])
```

[37]: 191.048

[38]:
```
np.std(data1['Unit Cost'])
```

[38]: 187.2647759029979

[39]:
```
np.var(data1['Unit Cost'])
```

[39]: 35068.096294000024

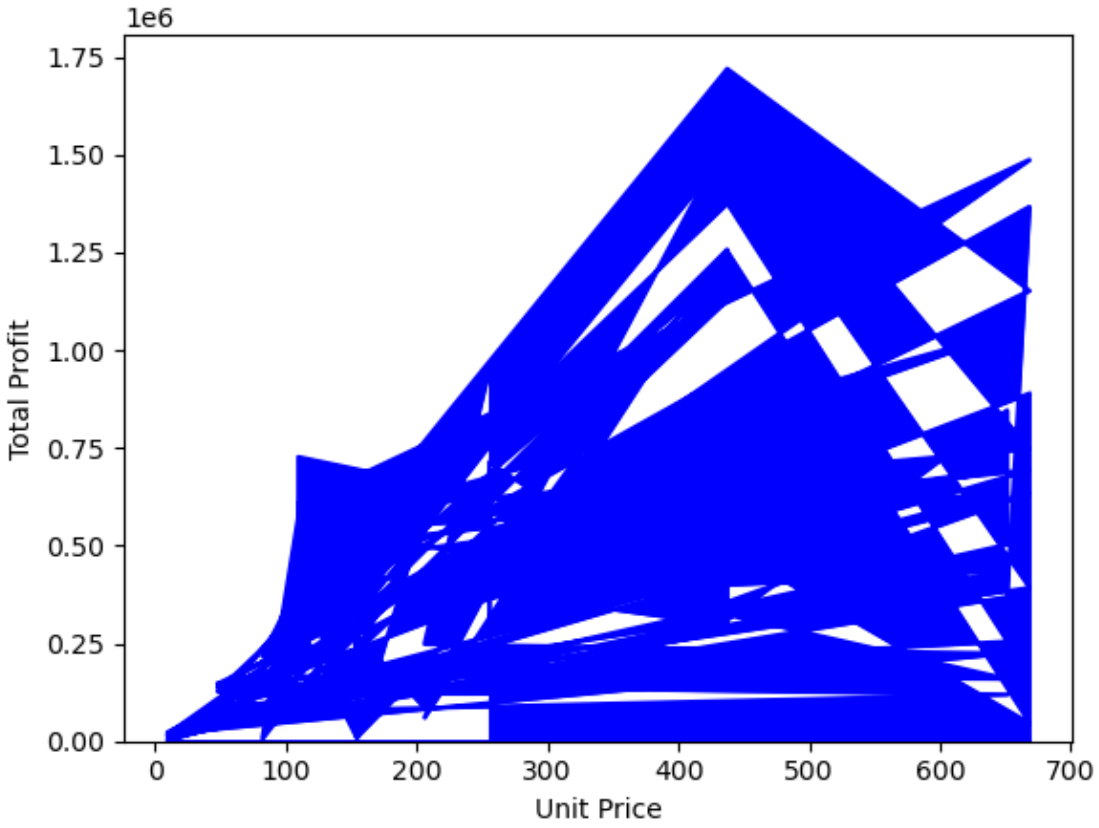
[40]:
```
np.median(data1['Unit Cost'])
```

[40]: 107.275

Maximum and minimum unit costs are 6.92 and 524.96 respectively. Average unit cost of a product is 191.05. The Unit Cost varies considerably throughout it's distribution. The median cost of a unit stands at 107.28.

[41]:
```
area_plot = data.plot.area(x='Unit Price',y='Total␣
 ↪Profit',color='blue',stacked=True,legend=None)
mp.ylabel('Total Profit')
```

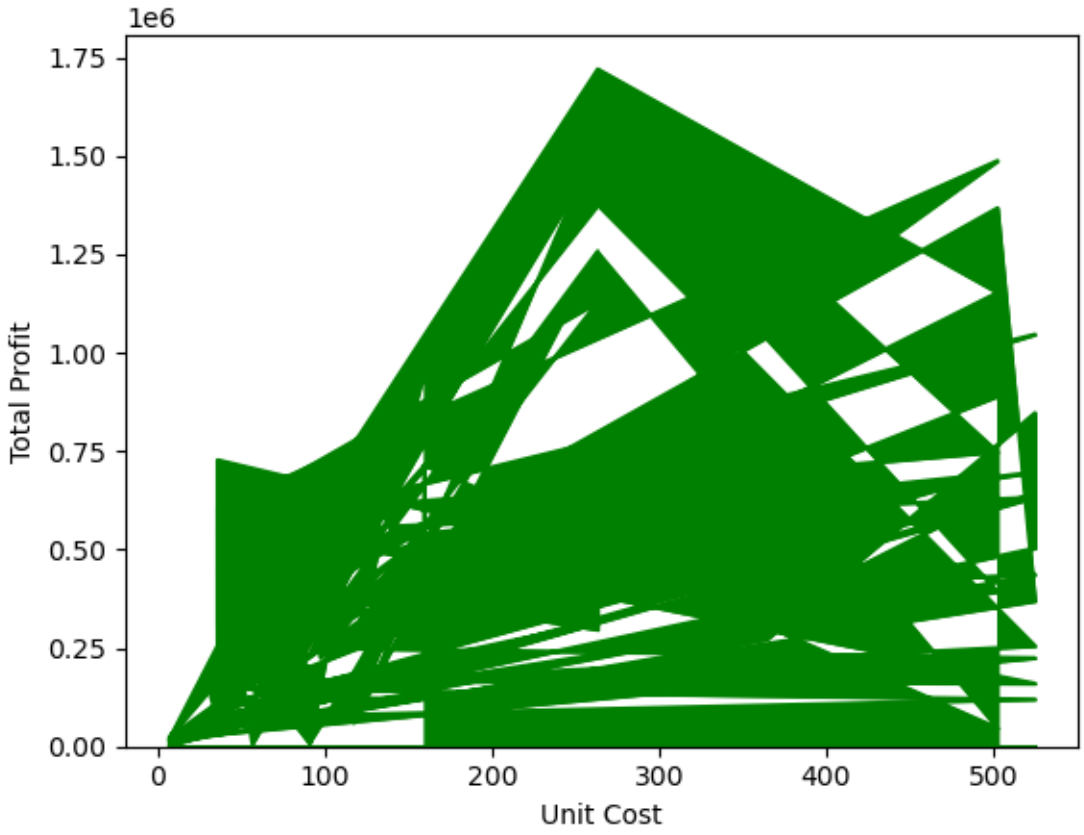
[41]:
```
Text(0, 0.5, 'Total Profit')
```



From the above plot we can conclude that the maximum profit has been generated in the unit price range of 400- 500.

[42]:
```
area_plot = data.plot.area(x='Unit Cost',y='Total␣
 ↪Profit',color='g',stacked=True,legend=None)
mp.ylabel('Total Profit')
```

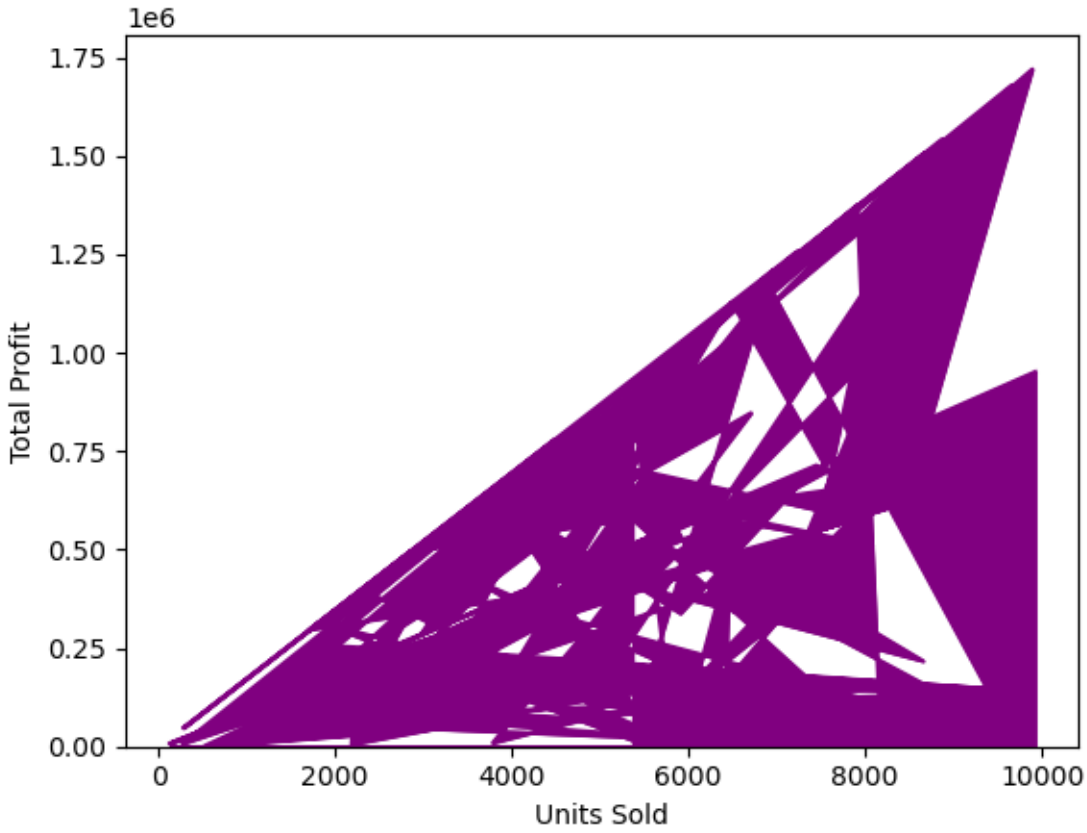
[42]:
```
Text(0, 0.5, 'Total Profit')
```



From the above plot we can conclude that the maximum profit has been generated in the unit cost range of 200- 300.

[43]:
```
data1.plot.area(x='Units Sold',y='Total Profit',color='purple',legend=None)
mp.ylabel('Total Profit')
```

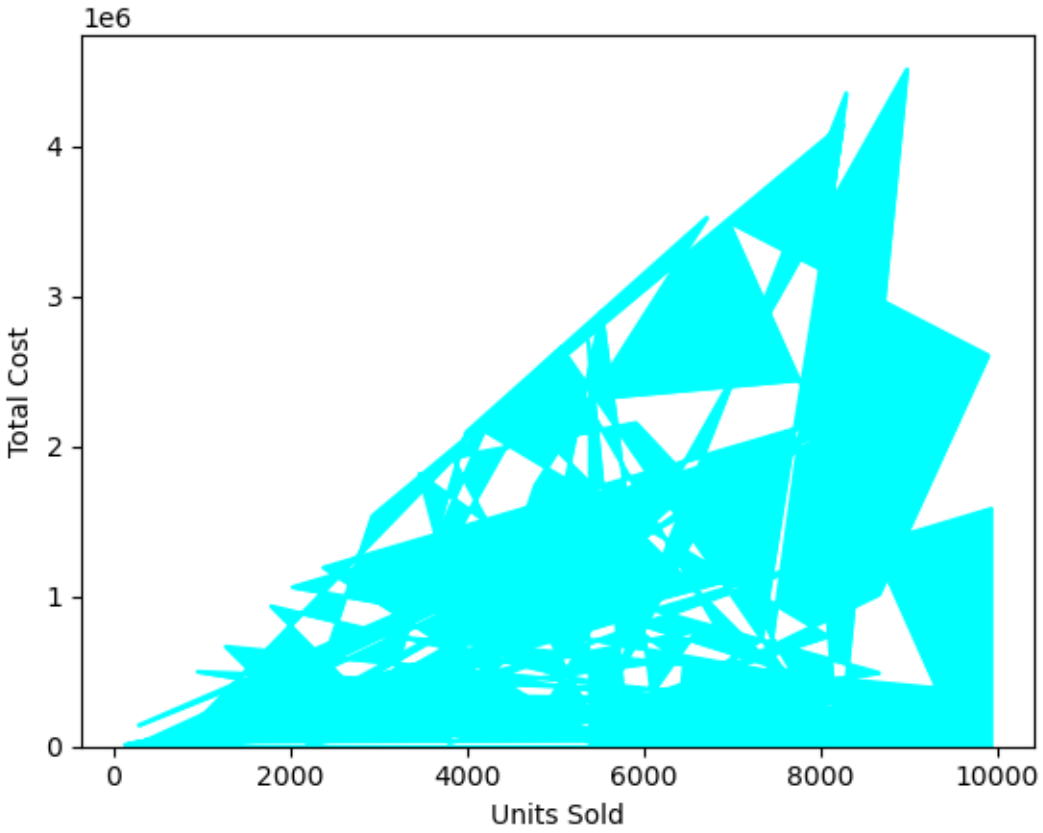
[43]:
```
Text(0, 0.5, 'Total Profit')
```



From the above plot we can conclude that maximum profit has been generated when the number of units sold were between 8000 and 10000. More the number of units sold, more will be the profit generated.

[44]:
```
data1.plot.area(x='Units Sold',y='Total Cost',color='aqua',legend=None)
mp.ylabel('Total Cost')
```

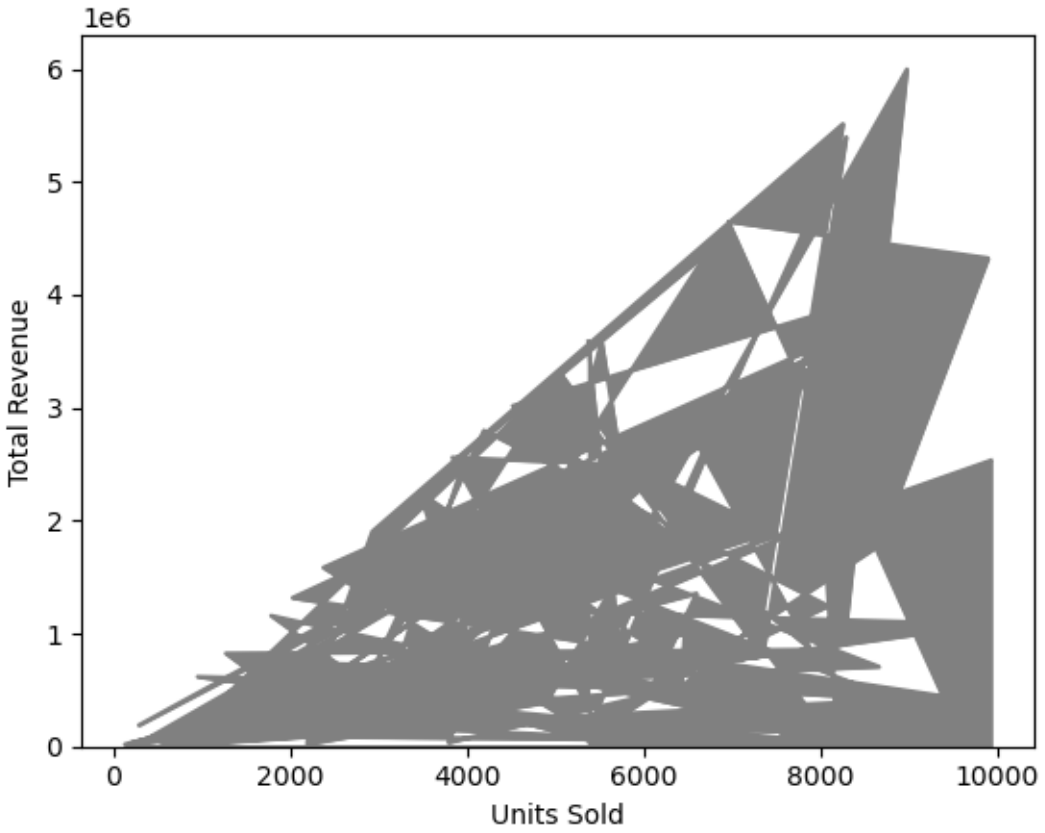
[44]:
```
Text(0, 0.5, 'Total Cost')
```



From the above plot we can conclude that maximum cost has been generated when the units sold were above 8000 and below 10000.

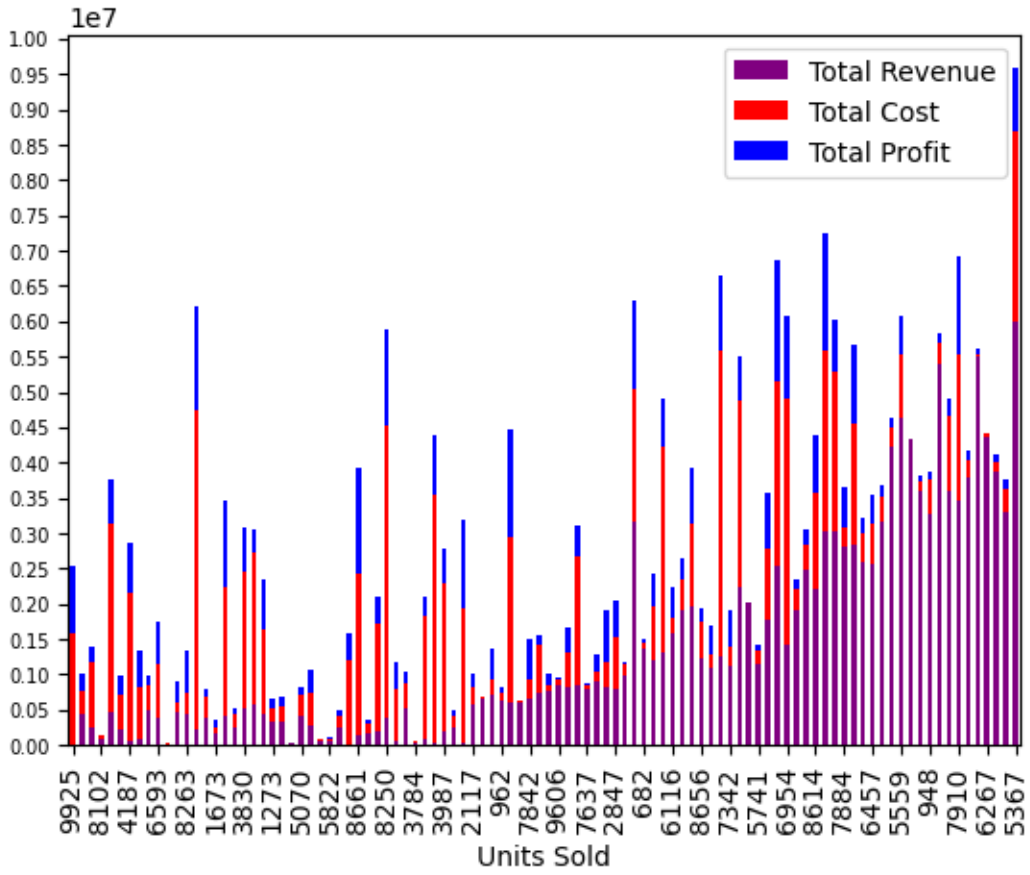
[45]:
```
data1.plot.area(x='Units Sold',y='Total Revenue',color='grey',legend=None)
mp.ylabel('Total Revenue')
```

[45]:
```
Text(0, 0.5, 'Total Revenue')
```



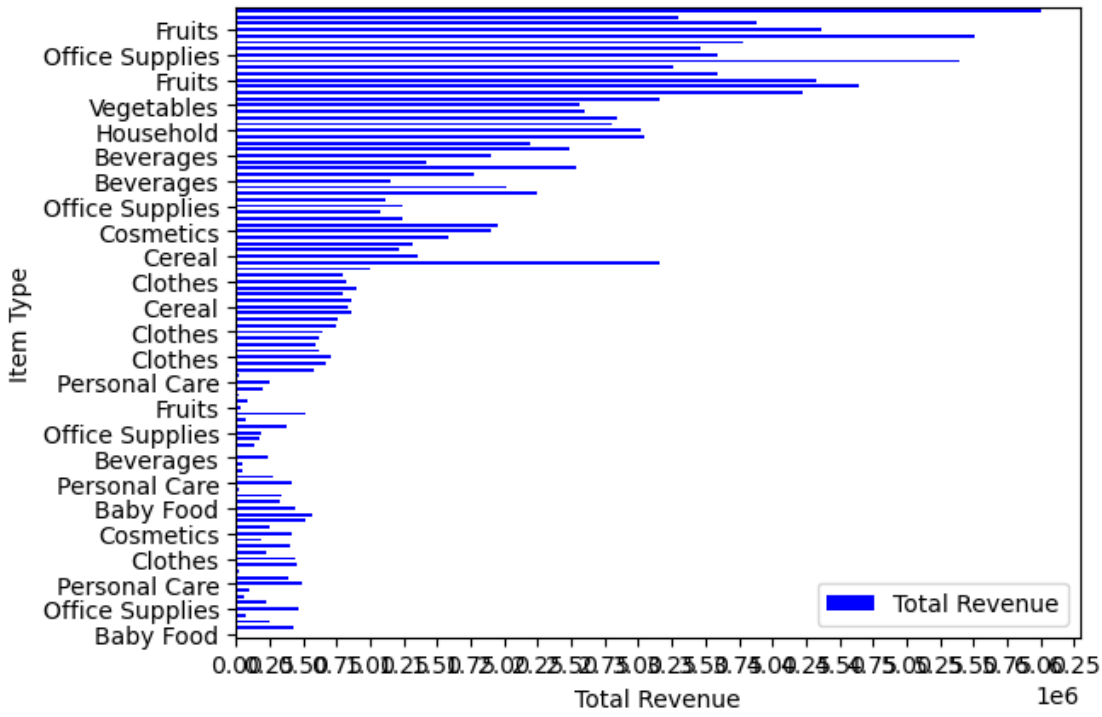
From the above plot we can conclude that maximum revenue has been generated when 8000-10000 units of products were sold.

[46]:
```
bar_plot = data.plot.bar(x='Units Sold',y=['Total Revenue','Total Cost','Total␣
↪Profit'],color=['purple','red','blue'],stacked=True,rot=True)
mp.xticks(rotation=90)
mp.locator_params(nbins=38)
mp.tick_params(axis='y', which='major', labelsize=7)
```



[47]:
```
data.plot.barh(x='Item Type',y='Total Revenue',color='blue')
mp.locator_params(nbins=28)
mp.xlabel('Total Revenue')
```

[47]: Text(0.5, 0, 'Total Revenue')



[48]:
```
#Finding the unique values of all the categories of item type accordinding to␣
 ↪the hash table.
data['Item Type'].unique()
```

[48]:
```
array(['Baby Food', 'Cereal', 'Office Supplies', 'Fruits', 'Household',
       'Vegetables', 'Personal Care', 'Clothes', 'Cosmetics', 'Beverages',
       'Meat', 'Snacks'], dtype=object)
```

[49]:
```
#Rearranging the items columnaccording to their Unique values.
items = ['Baby Food', 'Cereal', 'Office Supplies', 'Fruits', 'Household',
       'Vegetables', 'Personal Care', 'Clothes', 'Cosmetics', 'Beverages',
       'Meat', 'Snacks']
```

[50]:
```
data['Item Type'] = pd.Categorical(data['Item␣
 ↪Type'],categories=items,ordered=True)
```

[51]:
```
#Checking the items are rearranged or not
data
```

[51]:

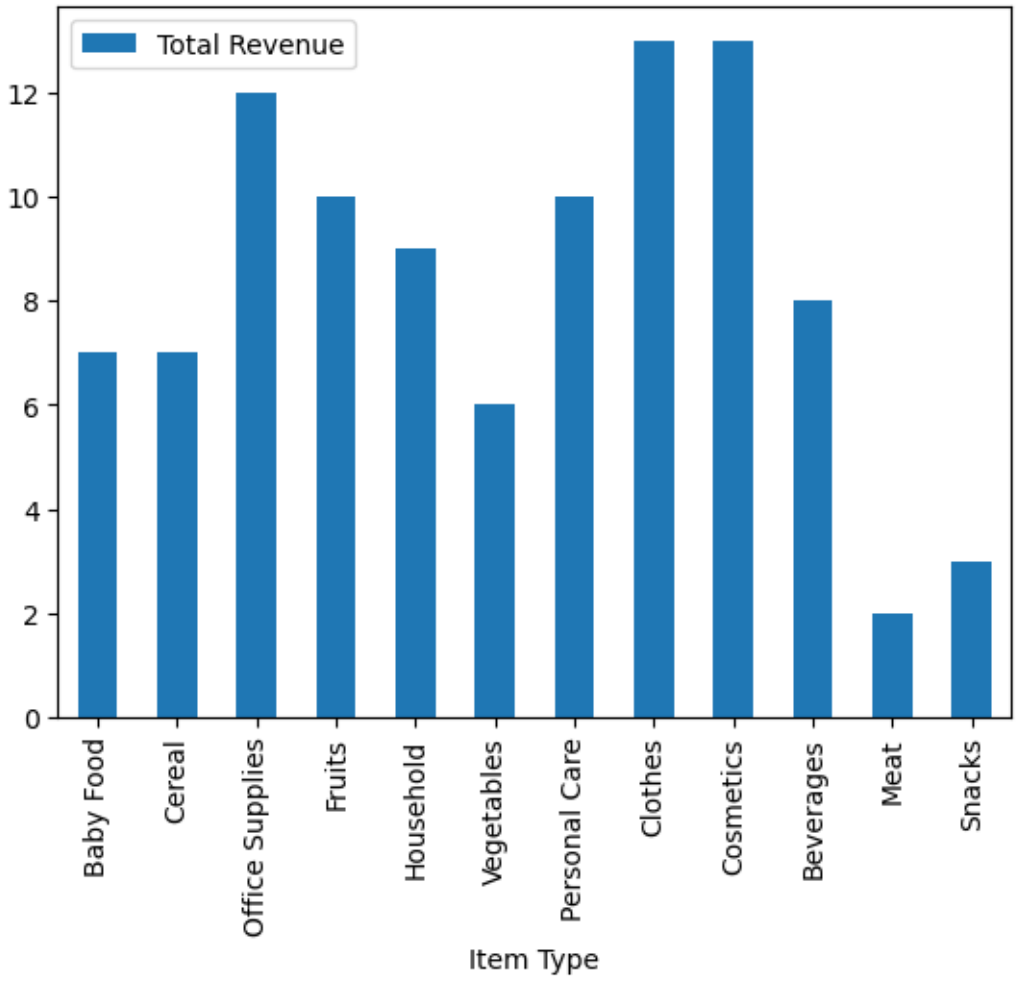
Order ID	Order Date	Order Priority	Ship Date	Item Type
669165933	5/28/2010	H	6/27/2010	Baby Food
963881480	8/22/2012	C	9/15/2012	Cereal
341417157	5/2/2014	L	5/8/2014	Office Supplies

```
514321792   6/20/2014             C    7/5/2014          Fruits
115456712    2/1/2013             L    2/6/2013  Office Supplies
…                …             …         …               …
512878119   7/26/2011             M    9/3/2011         Clothes
810711038  11/11/2011             L   12/28/2011          Fruits
728815257    6/1/2016             C   6/29/2016      Vegetables
559427106   7/30/2015             M    8/8/2015   Personal Care
665095412   2/10/2012             L   2/15/2012       Household

                                    Region                Country  \
Order ID
669165933              Australia and Oceania                 Tuvalu
963881480  Central America and the Caribbean                Grenada
341417157                             Europe                 Russia
514321792                 Sub-Saharan Africa  Sao Tome and Principe
115456712                 Sub-Saharan Africa                 Rwanda
…                                        …                     …
512878119                 Sub-Saharan Africa                   Mali
810711038                               Asia               Malaysia
728815257                 Sub-Saharan Africa           Sierra Leone
559427106                      North America                 Mexico
665095412                 Sub-Saharan Africa             Mozambique

          Sales Channel  Units Sold  Unit Price  Unit Cost  Total Revenue  \
Order ID
669165933       Offline        9925      255.28     159.42        4870.26
963881480        Online        2804      205.70     117.11      435466.90
341417157       Offline        1779      651.21     524.96      247956.32
514321792        Online        8102        9.33       6.92       75591.66
115456712       Offline        5062      651.21     524.96      471336.91
…                   …           …           …          …             …
512878119        Online         888      109.28      35.84     5513227.50
810711038       Offline        6267        9.33       6.92     4368316.68
728815257       Offline        1485      154.06      90.93     3876652.40
559427106       Offline        5767       81.73      56.67     3296425.02
665095412       Offline        5367      668.27     502.54     5997054.98

           Total Cost  Total Profit
Order ID
669165933  1582243.50     951410.50
963881480   328376.44     248406.36
341417157   933903.84     224598.75
514321792    56065.84      19525.82
115456712  2657347.52     639077.50
…                …             …
512878119    31825.92      65214.72
810711038    43367.64      15103.47
```

```
728815257    135031.05      93748.05
559427106    326815.89     144521.02
665095412   2697132.18     889472.91

[100 rows x 13 columns]
```

[52]:
```
pd.pivot_table(data,values='Total Revenue',index='Item Type',aggfunc='count').
↪plot(kind='bar')
```

[52]:
```
<Axes: xlabel='Item Type'>
```

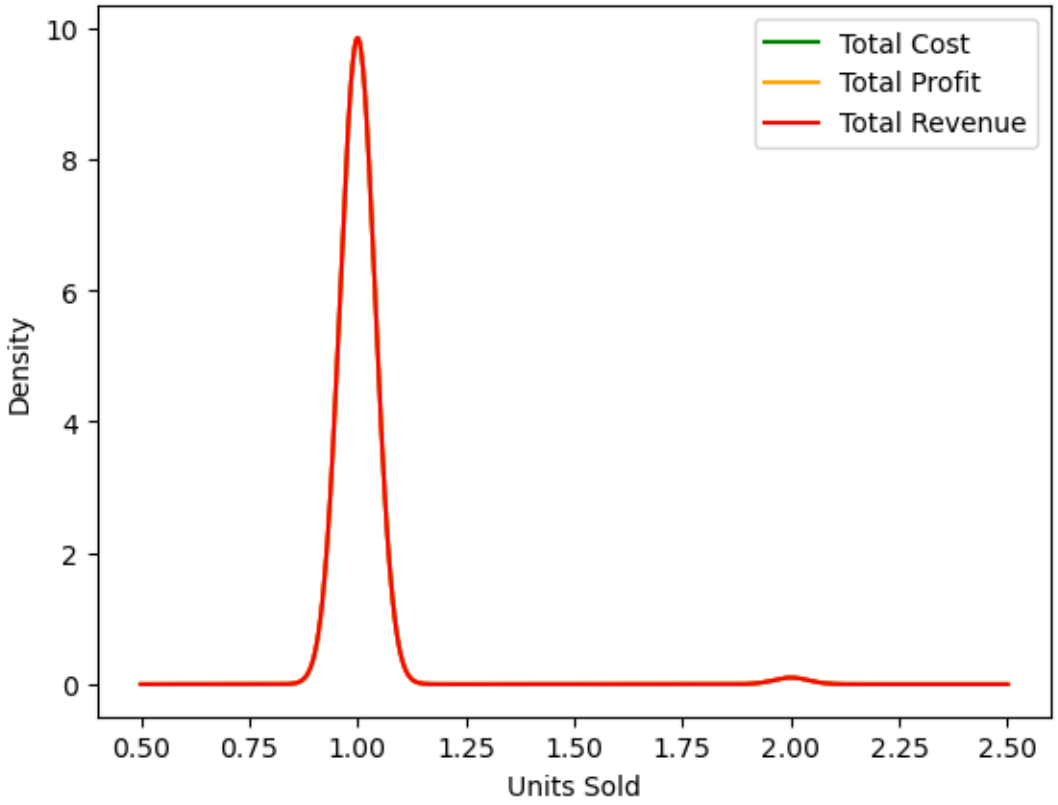


From the above graph we can conclude that maximum revenue has been generated from the items 'Clothes' and 'Cosmetics' closely followed by 'Office Supplies'.

[53]:
```
pd.pivot_table(data,values=['Total Revenue','Total Cost','Total␣
↪Profit'],index='Units Sold',aggfunc='count').
↪plot(kind='kde',color=['green','orange','red'],stacked=True)
```

```
mp.xlabel('Units Sold')
```

[53]: Text(0.5, 0, 'Units Sold')



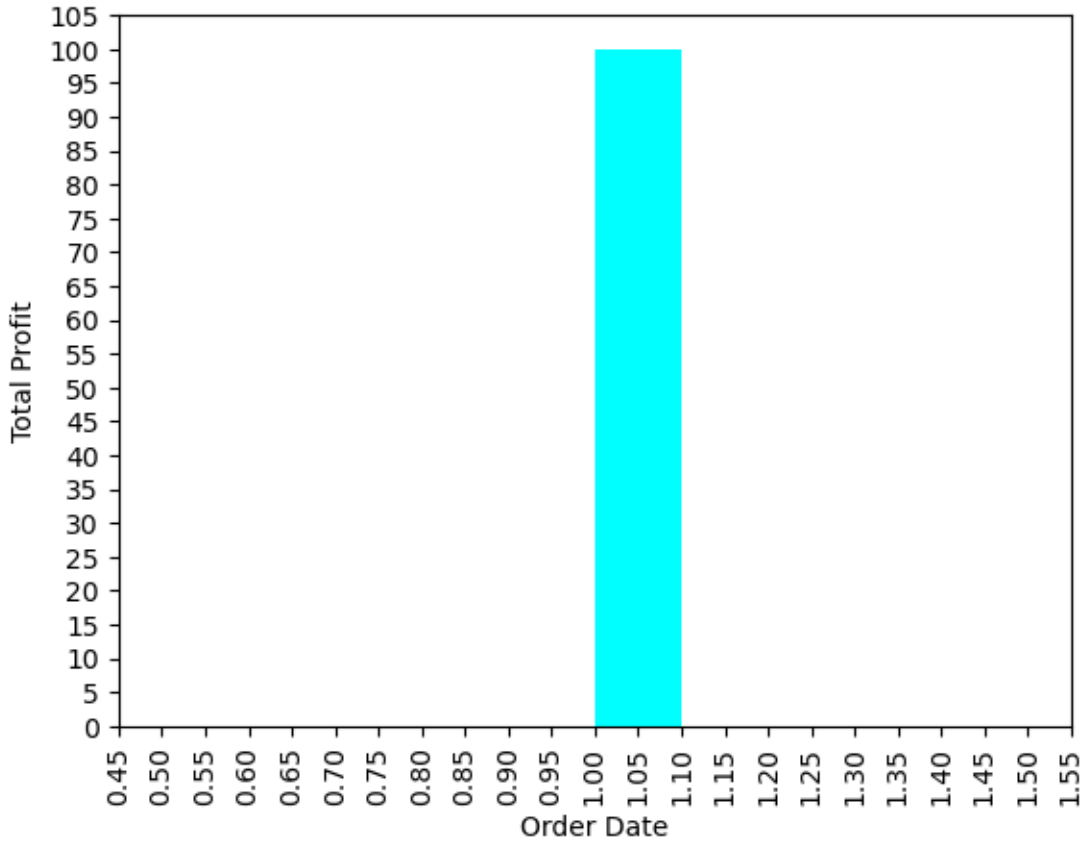
[54]:
```
data['Order Date'].unique()
```

[54]:
```
array(['5/28/2010', '8/22/2012', '5/2/2014', '6/20/2014', '2/1/2013',
       '2/4/2015', '4/23/2011', '7/17/2012', '7/14/2015', '4/18/2014',
       '6/24/2011', '8/2/2014', '1/13/2017', '2/8/2017', '2/19/2014',
       '4/23/2012', '11/19/2016', '4/1/2015', '12/30/2010', '7/31/2012',
       '5/14/2014', '7/31/2015', '6/30/2016', '9/8/2014', '5/7/2016',
       '5/22/2017', '10/13/2014', '5/7/2010', '7/18/2014', '5/26/2012',
       '9/17/2012', '12/29/2013', '10/27/2015', '1/16/2015', '2/25/2017',
       '5/8/2017', '11/22/2011', '1/14/2017', '4/1/2012', '2/16/2012',
       '3/11/2017', '2/6/2010', '6/7/2012', '10/6/2012', '11/14/2015',
       '3/29/2016', '12/31/2016', '12/23/2010', '10/14/2014', '1/11/2012',
       '2/2/2010', '8/18/2013', '3/25/2013', '11/26/2011', '9/17/2013',
       '6/8/2012', '6/30/2010', '2/23/2015', '1/5/2012', '4/7/2014',
       '6/9/2013', '6/26/2013', '11/7/2011', '10/30/2010', '10/13/2013',
       '10/11/2013', '7/8/2012', '7/25/2016', '10/24/2010', '4/25/2015',
       '4/23/2013', '8/14/2015', '5/26/2011', '5/20/2017', '7/5/2013',
```

```
       '11/6/2014', '10/28/2014', '9/15/2011', '5/29/2012', '7/20/2013',
       '10/21/2012', '9/18/2012', '11/15/2016', '1/4/2011', '3/18/2012',
       '2/17/2012', '1/16/2011', '2/3/2014', '4/30/2012', '10/23/2016',
       '12/6/2016', '7/7/2014', '6/13/2012', '11/26/2010', '2/8/2011',
       '7/26/2011', '11/11/2011', '6/1/2016', '7/30/2015', '2/10/2012'],
      dtype=object)
```

[55]:
```
pd.pivot_table(data,values='Total Profit',index='Order Date',aggfunc='count').
  ↪plot(kind='hist',color='aqua',stacked=False,legend=None)
mp.xticks(rotation=90)
mp.ylabel('Total Profit')
mp.locator_params(nbins=32)
mp.xlabel('Order Date')
```

[55]:
```
Text(0.5, 0, 'Order Date')
```



[56]:
```
data.describe()
```

[56]:
```
         Units Sold  Unit Price   Unit Cost  Total Revenue    Total Cost  \
count   100.000000  100.000000  100.000000   1.000000e+02  1.000000e+02
mean   5128.710000  276.761300  191.048000   1.373488e+06  9.318057e+05
std    2794.484562  235.592241  188.208181   1.460029e+06  1.083938e+06
min     124.000000    9.330000    6.920000   4.870260e+03  3.612240e+03
25%    2836.250000   81.730000   35.840000   2.687212e+05  1.688680e+05
50%    5382.500000  179.880000  107.275000   7.523144e+05  3.635664e+05
75%    7369.000000  437.200000  263.330000   2.212045e+06  1.613870e+06
max    9925.000000  668.270000  524.960000   5.997055e+06  4.509794e+06

       Total Profit
count  1.000000e+02
mean   4.416820e+05
std    4.385379e+05
min    1.258020e+03
25%    1.214436e+05
50%    2.907680e+05
75%    6.358288e+05
max    1.719922e+06
```

[57]:
```
data
```

[57]:
```
           Order Date Order Priority   Ship Date        Item Type  \
Order ID
669165933   5/28/2010              H   6/27/2010        Baby Food
963881480   8/22/2012              C   9/15/2012           Cereal
341417157    5/2/2014              L    5/8/2014  Office Supplies
514321792   6/20/2014              C    7/5/2014           Fruits
115456712    2/1/2013              L    2/6/2013  Office Supplies
…                …              …          …                …
512878119   7/26/2011              M    9/3/2011          Clothes
810711038  11/11/2011              L  12/28/2011           Fruits
728815257    6/1/2016              C   6/29/2016       Vegetables
559427106   7/30/2015              M    8/8/2015    Personal Care
665095412   2/10/2012              L   2/15/2012        Household

                                      Region                Country  \
Order ID
669165933                 Australia and Oceania                Tuvalu
963881480  Central America and the Caribbean                Grenada
341417157                             Europe                 Russia
514321792                 Sub-Saharan Africa  Sao Tome and Principe
115456712                 Sub-Saharan Africa                 Rwanda
…                                        …                    …
512878119                 Sub-Saharan Africa                   Mali
810711038                               Asia               Malaysia
728815257                 Sub-Saharan Africa           Sierra Leone
```

```
559427106                    North America                 Mexico
665095412                Sub-Saharan Africa             Mozambique

          Sales Channel  Units Sold  Unit Price  Unit Cost  Total Revenue  \
Order ID
669165933       Offline        9925      255.28     159.42        4870.26
963881480        Online        2804      205.70     117.11      435466.90
341417157       Offline        1779      651.21     524.96      247956.32
514321792        Online        8102        9.33       6.92       75591.66
115456712       Offline        5062      651.21     524.96      471336.91
…                 …          …          …          …           …
512878119        Online         888      109.28      35.84     5513227.50
810711038       Offline        6267        9.33       6.92     4368316.68
728815257       Offline        1485      154.06      90.93     3876652.40
559427106       Offline        5767       81.73      56.67     3296425.02
665095412       Offline        5367      668.27     502.54     5997054.98

           Total Cost  Total Profit
Order ID
669165933  1582243.50     951410.50
963881480   328376.44     248406.36
341417157   933903.84     224598.75
514321792    56065.84      19525.82
115456712  2657347.52     639077.50
…                …           …
512878119    31825.92      65214.72
810711038    43367.64      15103.47
728815257   135031.05      93748.05
559427106   326815.89     144521.02
665095412  2697132.18     889472.91

[100 rows x 13 columns]
```

[58]:
```
data['Region'].unique()
```

[58]:
```
array(['Australia and Oceania', 'Central America and the Caribbean',
       'Europe', 'Sub-Saharan Africa', 'Asia',
       'Middle East and North Africa', 'North America'], dtype=object)
```

[59]:
```
regions = ['Australia and Oceania', 'Central America and the Caribbean',
       'Europe', 'Sub-Saharan Africa', 'Asia',
       'Middle East and North Africa', 'North America']
```

[60]:
```
data['Region'] = pd.Categorical(data['Region'],categories =␣
 ↪regions,ordered=True)
```

[61]:
```
data
```

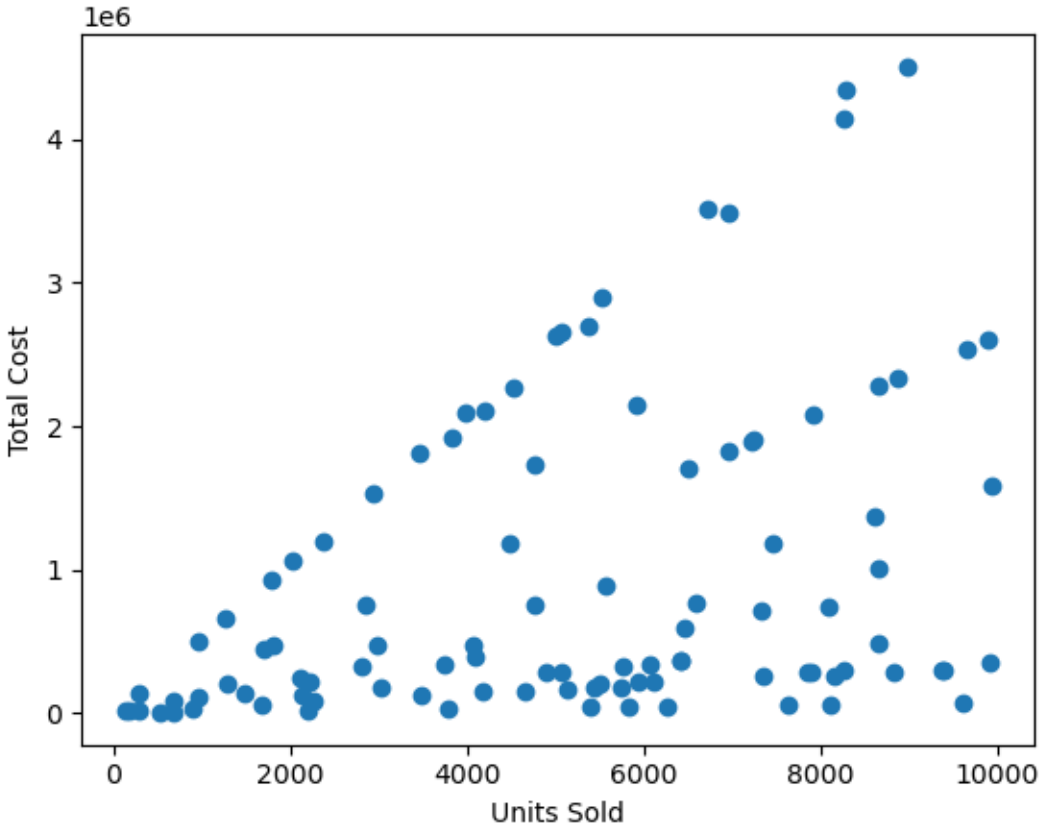
[61]:

Order ID	Order Date	Order Priority	Ship Date	Item Type	Region	Country	Sales Channel	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
669165933	5/28/2010	H	6/27/2010	Baby Food	Australia and Oceania	Tuvalu	Offline	9925	255.28	159.42	4870.26	1582243.50	951410.50
963881480	8/22/2012	C	9/15/2012	Cereal	Central America and the Caribbean	Grenada	Online	2804	205.70	117.11	435466.90	328376.44	248406.36
341417157	5/2/2014	L	5/8/2014	Office Supplies	Europe	Russia	Offline	1779	651.21	524.96	247956.32	933903.84	224598.75
514321792	6/20/2014	C	7/5/2014	Fruits	Sub-Saharan Africa	Sao Tome and Principe	Online	8102	9.33	6.92	75591.66		
115456712	2/1/2013	L	2/6/2013	Office Supplies	Sub-Saharan Africa	Rwanda	Offline	5062	651.21	524.96	471336.91		
…	…	…	…	…	…	…	…	…	…	…	…		
512878119	7/26/2011	M	9/3/2011	Clothes	Sub-Saharan Africa	Mali	Online	888	109.28	35.84	5513227.50		
810711038	11/11/2011	L	12/28/2011	Fruits	Asia	Malaysia	Offline	6267	9.33	6.92	4368316.68		
728815257	6/1/2016	C	6/29/2016	Vegetables	Sub-Saharan Africa	Sierra Leone	Offline	1485	154.06	90.93	3876652.40		
559427106	7/30/2015	M	8/8/2015	Personal Care	North America	Mexico	Offline	5767	81.73	56.67	3296425.02		
665095412	2/10/2012	L	2/15/2012	Household	Sub-Saharan Africa	Mozambique	Offline	5367	668.27	502.54	5997054.98		

```
514321792    56065.84      19525.82
115456712  2657347.52     639077.50
…                …             …
512878119    31825.92      65214.72
810711038    43367.64      15103.47
728815257   135031.05      93748.05
559427106   326815.89     144521.02
665095412  2697132.18     889472.91

[100 rows x 13 columns]
```

[62]:
```
mp.scatter(data['Units Sold'],data['Total Cost'])
mp.xlabel('Units Sold')
mp.ylabel('Total Cost')
```

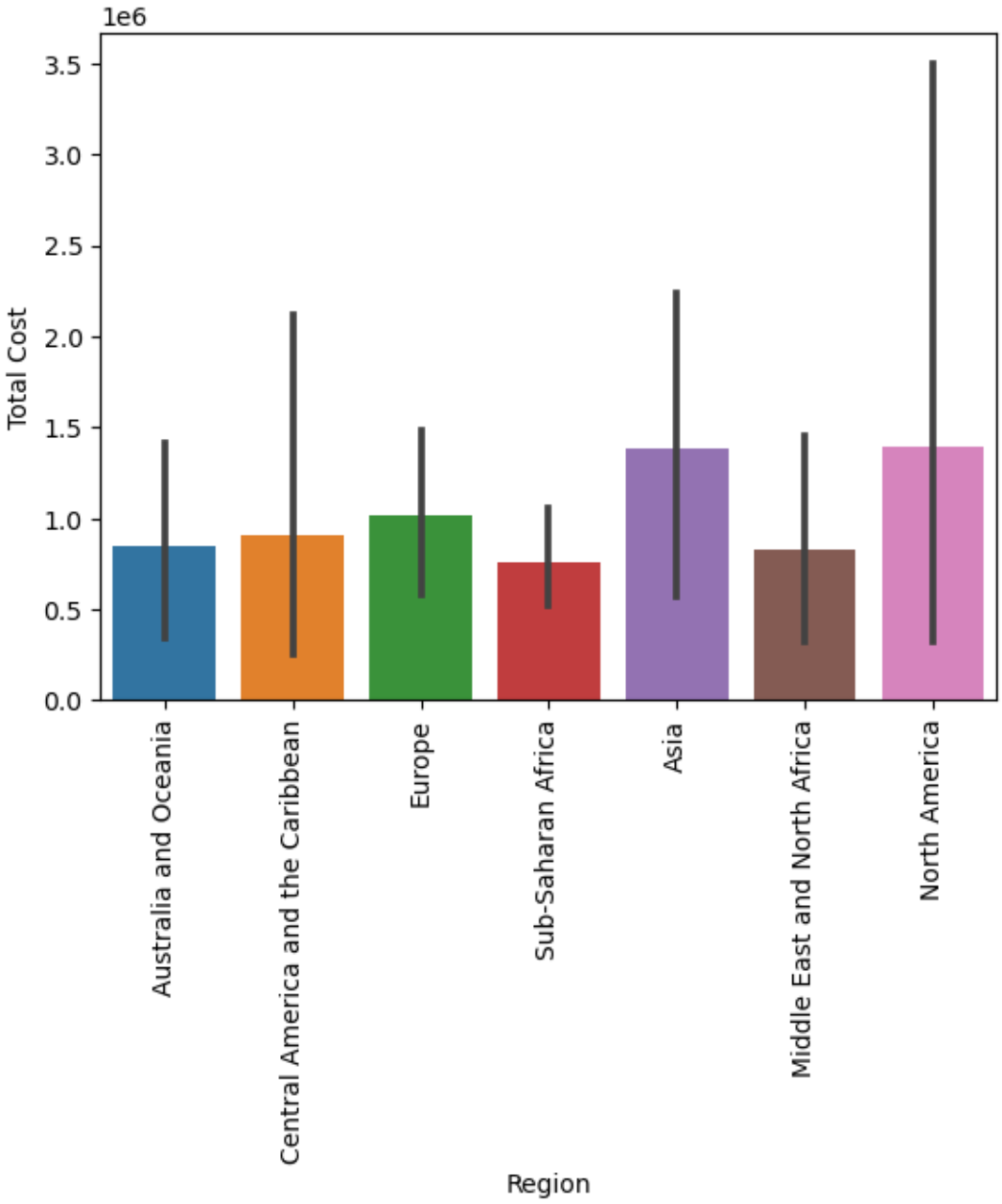
[62]:
```
Text(0, 0.5, 'Total Cost')
```



From the above scatter plot we can conclude that more the number of units sold of a product, more will be the total cost associated with it.

[63]:
```
sn.barplot(x='Region',y='Total Cost',data=data)
mp.xticks(rotation=90)
```

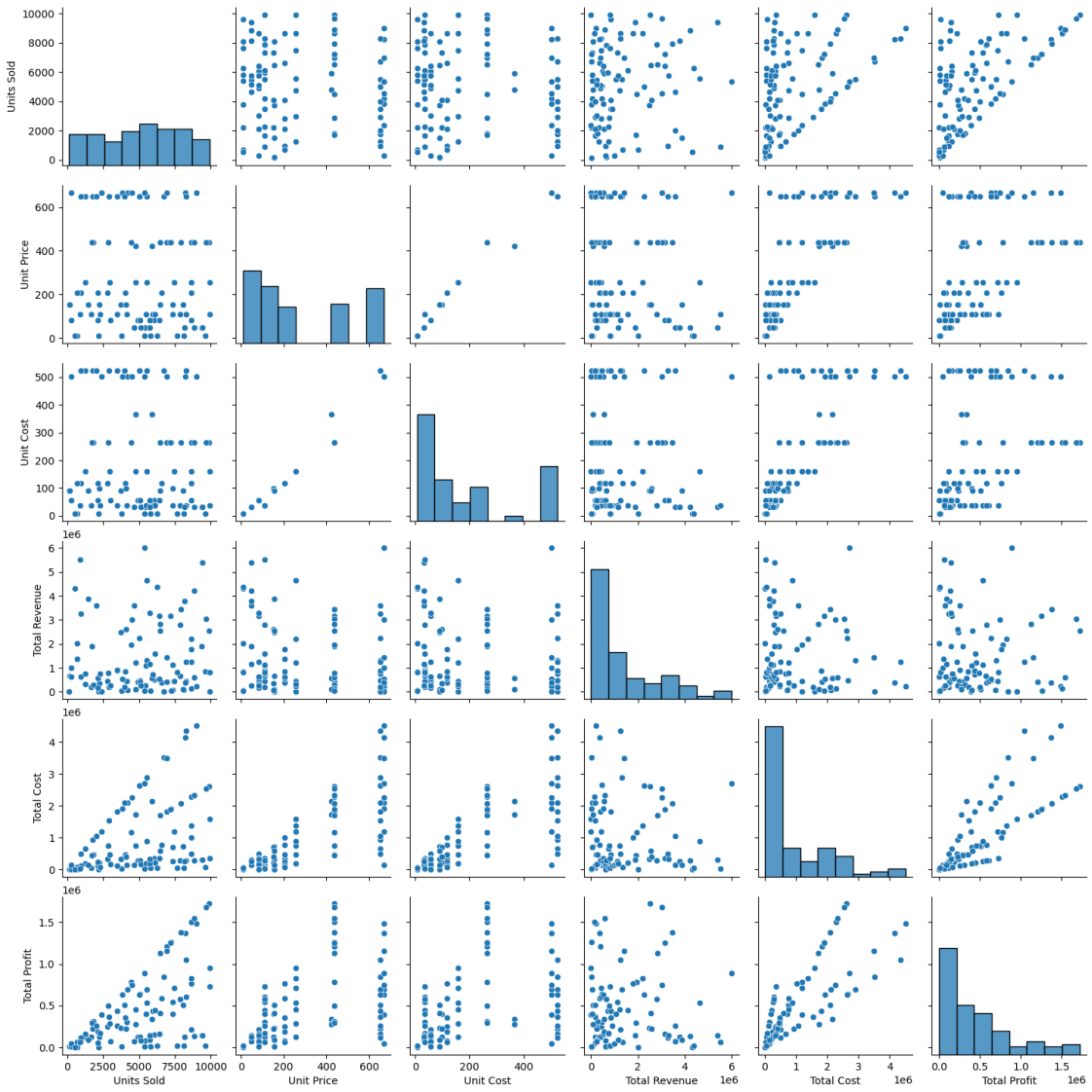
[63]:
```
(array([0, 1, 2, 3, 4, 5, 6]),
 [Text(0, 0, 'Australia and Oceania'),
  Text(1, 0, 'Central America and the Caribbean'),
  Text(2, 0, 'Europe'),
  Text(3, 0, 'Sub-Saharan Africa'),
  Text(4, 0, 'Asia'),
  Text(5, 0, 'Middle East and North Africa'),
  Text(6, 0, 'North America')])
```



Cost of items is maximum in Asia and North America, and minimum in Sub-Saharan Africa.

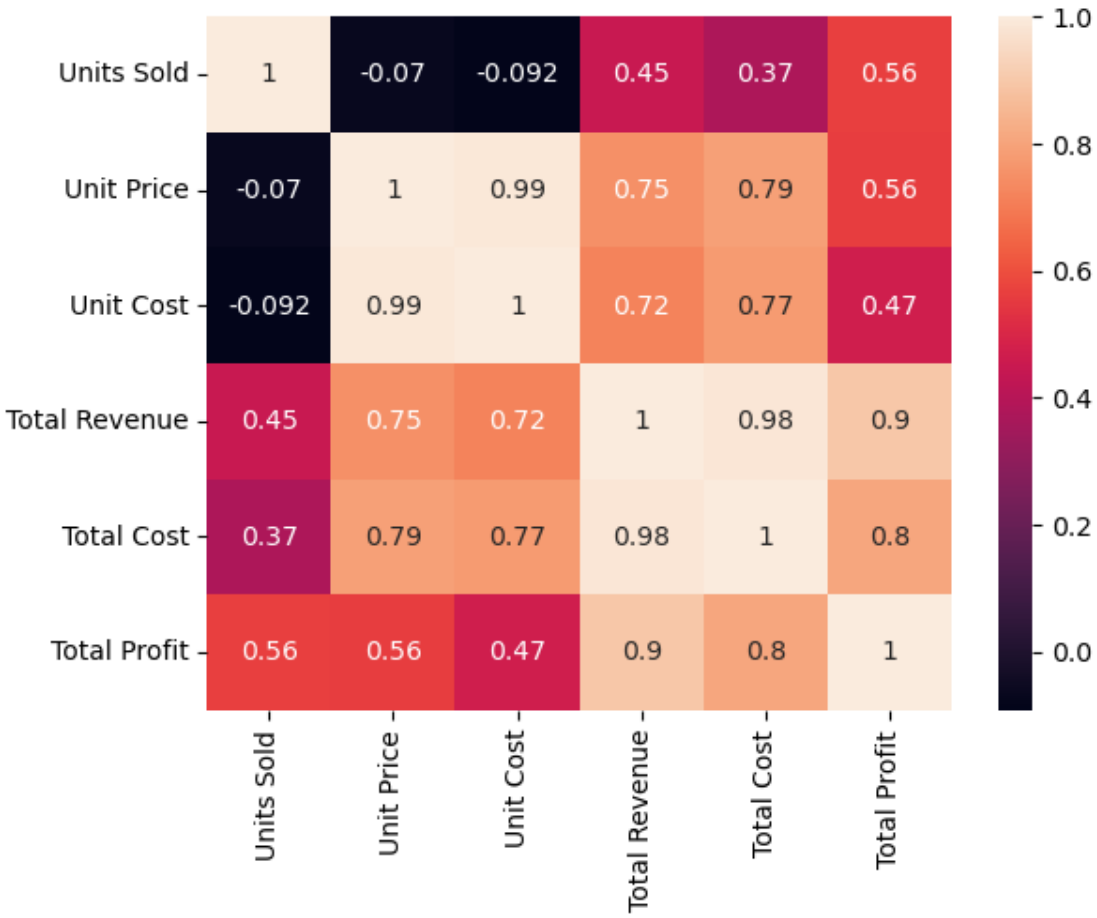
[64]:
```
sn.pairplot(data)
```

[64]:
```
<seaborn.axisgrid.PairGrid at 0x218c7c20550>
```



[65]:
```
sn.heatmap(data1.corr(),annot=True)
```

[65]: <Axes: >



From the above heatmap, we can infer that Total Cost is strongly related to Unit Price,Unit Cost and Total Profit. Units Sold and {Unit Price and Unit Cost} are completely independent. Unit Cost, Unit Price and Total Cost are almost completely independent of Total Revenue.

[66]:
```
data['Country'].unique()
```

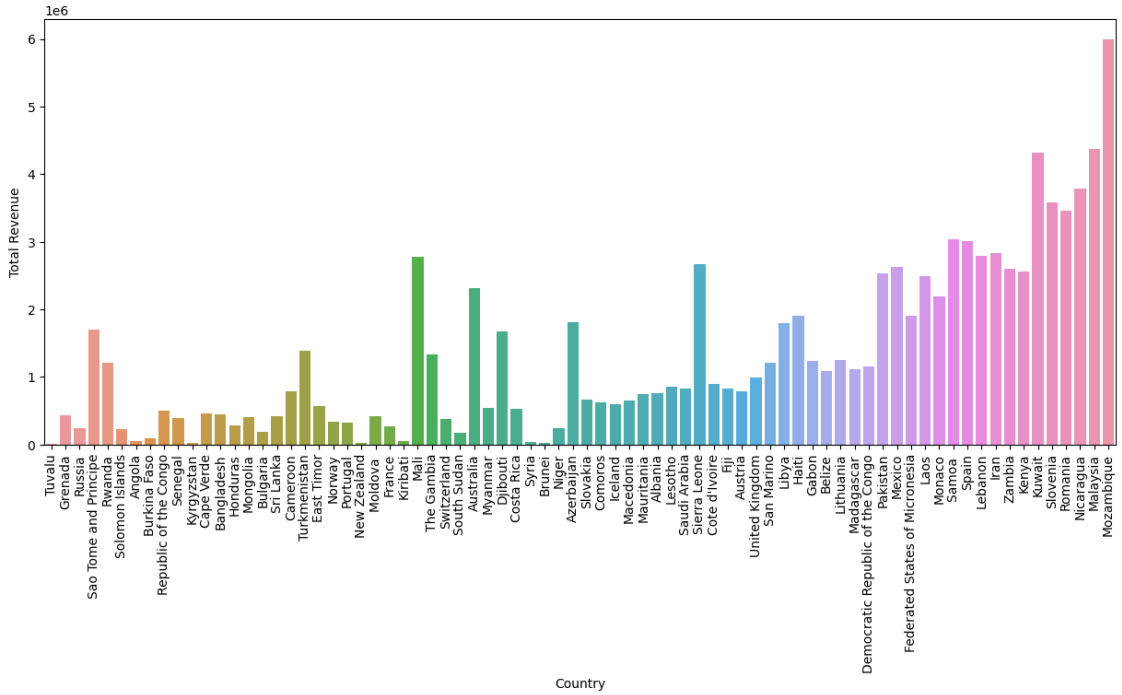
[66]:
```
array(['Tuvalu', 'Grenada', 'Russia', 'Sao Tome and Principe', 'Rwanda',
       'Solomon Islands', 'Angola', 'Burkina Faso',
       'Republic of the Congo', 'Senegal', 'Kyrgyzstan', 'Cape Verde',
       'Bangladesh', 'Honduras', 'Mongolia', 'Bulgaria', 'Sri Lanka',
       'Cameroon', 'Turkmenistan', 'East Timor', 'Norway', 'Portugal',
       'New Zealand', 'Moldova ', 'France', 'Kiribati', 'Mali',
       'The Gambia', 'Switzerland', 'South Sudan', 'Australia', 'Myanmar',
       'Djibouti', 'Costa Rica', 'Syria', 'Brunei', 'Niger', 'Azerbaijan',
       'Slovakia', 'Comoros', 'Iceland', 'Macedonia', 'Mauritania',
       'Albania', 'Lesotho', 'Saudi Arabia', 'Sierra Leone',
       "Cote d'Ivoire", 'Fiji', 'Austria', 'United Kingdom', 'San Marino',
       'Libya', 'Haiti', 'Gabon', 'Belize', 'Lithuania', 'Madagascar',
```

```
       'Democratic Republic of the Congo', 'Pakistan', 'Mexico',
       'Federated States of Micronesia', 'Laos', 'Monaco', 'Samoa ',
       'Spain', 'Lebanon', 'Iran', 'Zambia', 'Kenya', 'Kuwait',
       'Slovenia', 'Romania', 'Nicaragua', 'Malaysia', 'Mozambique'],
      dtype=object)
```

[67]:
```
countries = ['Tuvalu', 'Grenada', 'Russia', 'Sao Tome and Principe', 'Rwanda',
       'Solomon Islands', 'Angola', 'Burkina Faso',
       'Republic of the Congo', 'Senegal', 'Kyrgyzstan', 'Cape Verde',
       'Bangladesh', 'Honduras', 'Mongolia', 'Bulgaria', 'Sri Lanka',
       'Cameroon', 'Turkmenistan', 'East Timor', 'Norway', 'Portugal',
       'New Zealand', 'Moldova ', 'France', 'Kiribati', 'Mali',
       'The Gambia', 'Switzerland', 'South Sudan', 'Australia', 'Myanmar',
       'Djibouti', 'Costa Rica', 'Syria', 'Brunei', 'Niger', 'Azerbaijan',
       'Slovakia', 'Comoros', 'Iceland', 'Macedonia', 'Mauritania',
       'Albania', 'Lesotho', 'Saudi Arabia', 'Sierra Leone',
       "Cote d'Ivoire", 'Fiji', 'Austria', 'United Kingdom', 'San Marino',
       'Libya', 'Haiti', 'Gabon', 'Belize', 'Lithuania', 'Madagascar',
       'Democratic Republic of the Congo', 'Pakistan', 'Mexico',
       'Federated States of Micronesia', 'Laos', 'Monaco', 'Samoa ',
       'Spain', 'Lebanon', 'Iran', 'Zambia', 'Kenya', 'Kuwait',
       'Slovenia', 'Romania', 'Nicaragua', 'Malaysia', 'Mozambique']
```

[68]:
```
data['Country'] = pd.
 ↪Categorical(data['Country'],categories=countries,ordered=True)
```

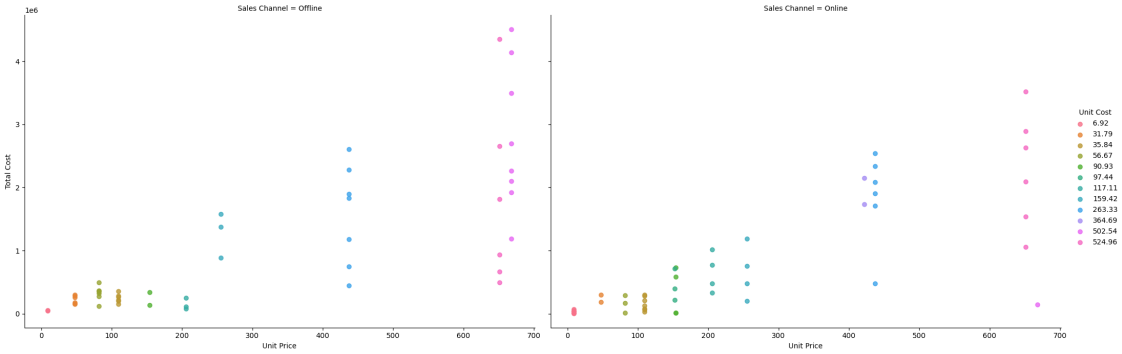
[69]:
```
mp.figure(figsize=(15,6))
sn.barplot(x='Country', y='Total Revenue', data=data, ci=None)
mp.xticks(rotation=90)
mp.tick_params(axis='x', which='major', labelsize=10)
```



From the above we can conclude Mozambique is the country where maximum revenue has been generated followed by Kenya.

[70]:
```
sn.lmplot(x='Unit Price',y='Total Cost',data=data,col='Sales Channel',hue='Unit␣
↪Cost',aspect=1.5,height=7)
```

[70]:
```
<seaborn.axisgrid.FacetGrid at 0x218cb254190>
```



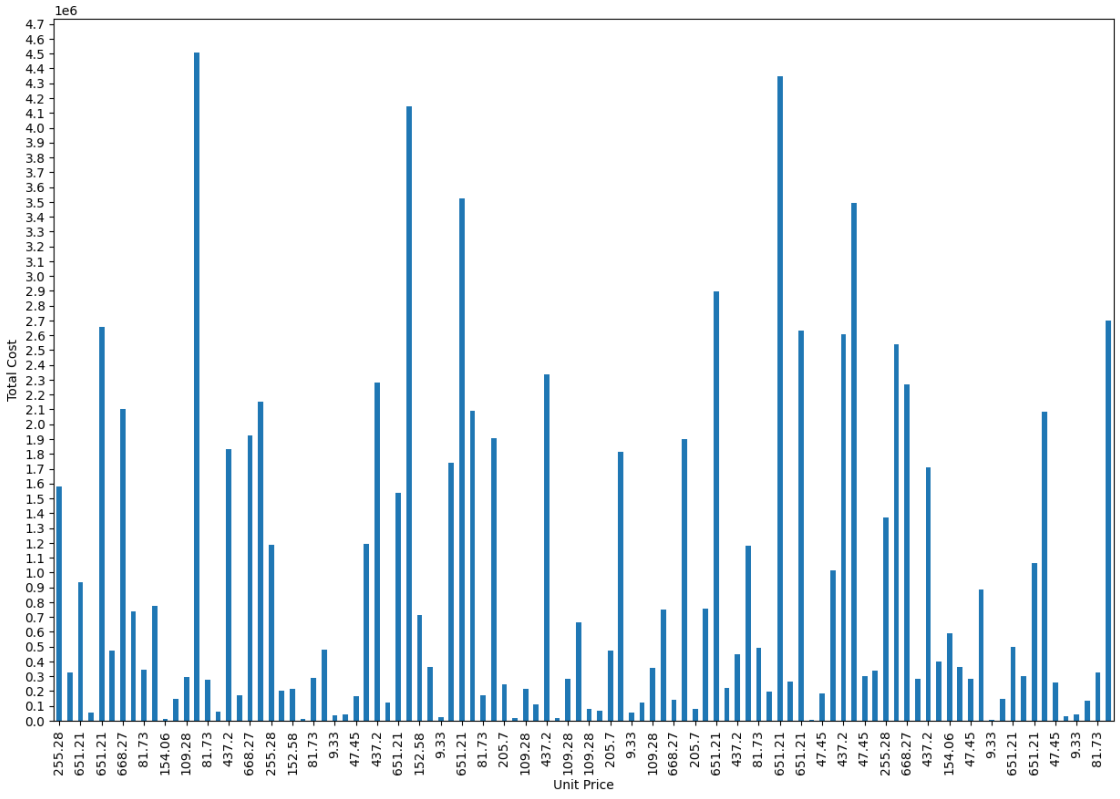
[71]:
```
data.sort_values(by='Unit Price')
```

[71]:
```
            Order Date Order Priority   Ship Date  Item Type  \
Order ID
```

```
142278373    9/8/2014              H   10/4/2014     Fruits
508980977   9/17/2013              H  10/24/2013     Fruits
162052476  11/22/2011              L   12/3/2011     Fruits
514321792   6/20/2014              C    7/5/2014     Fruits
810711038  11/11/2011              L  12/28/2011     Fruits
…               …             …          …        …
886494815   5/26/2012              L    6/9/2012  Household
213487374  10/21/2012              L  11/30/2012  Household
955357205    1/5/2012              L   2/14/2012  Household
441619336  12/30/2010              L   1/20/2011  Household
665095412   2/10/2012              L   2/15/2012  Household

                              Region                Country Sales Channel  \
Order ID
142278373         Australia and Oceania            New Zealand        Online
508980977            Sub-Saharan Africa  Sao Tome and Principe       Offline
162052476  Middle East and North Africa                 Syria        Online
514321792            Sub-Saharan Africa  Sao Tome and Principe        Online
810711038                          Asia               Malaysia       Offline
…                                 …                      …           …
886494815            Sub-Saharan Africa             The Gambia       Offline
213487374                        Europe                  Spain       Offline
955357205                        Europe         United Kingdom        Online
441619336                          Asia           Turkmenistan       Offline
665095412            Sub-Saharan Africa              Mozambique       Offline

           Units Sold  Unit Price  Unit Cost  Total Revenue  Total Cost  \
Order ID
142278373        2187        9.33       6.92       20404.71    15134.04
508980977        7637        9.33       6.92      802333.76    52848.04
162052476        3784        9.33       6.92       35304.72    26185.28
514321792        8102        9.33       6.92       75591.66    56065.84
810711038        6267        9.33       6.92     4368316.68    43367.64
…                …          …          …              …           …
886494815        2370      668.27     502.54        6279.09  1191019.80
213487374        4513      668.27     502.54     3015902.51  2267963.02
955357205         282      668.27     502.54      994765.42   141716.28
441619336        3830      668.27     502.54      524870.06  1924728.20
665095412        5367      668.27     502.54     5997054.98  2697132.18

           Total Profit
Order ID
142278373      5270.67
508980977     18405.17
162052476      9119.44
514321792     19525.82
810711038     15103.47
```

```
…                …
886494815     392780.10
213487374     747939.49
955357205      46735.86
441619336     634745.90
665095412     889472.91

[100 rows x 13 columns]
```

[72]:
```
data.plot.bar(x='Unit Price',y='Total Cost',legend=None,figsize=(15,10),rot=0)
mp.ylabel('Total Cost')
mp.xticks(rotation=90)
mp.locator_params(nbins=90)
```



From the above bar graph we can conclude that higher the value of unit price of a product, more will be the total cost of it.

[73]:
```
np.cov(data['Unit Price'],data['Total Cost'])
```

[73]:
```
array([[5.55037038e+04, 2.01205393e+08],
       [2.01205393e+08, 1.17492213e+12]])
```

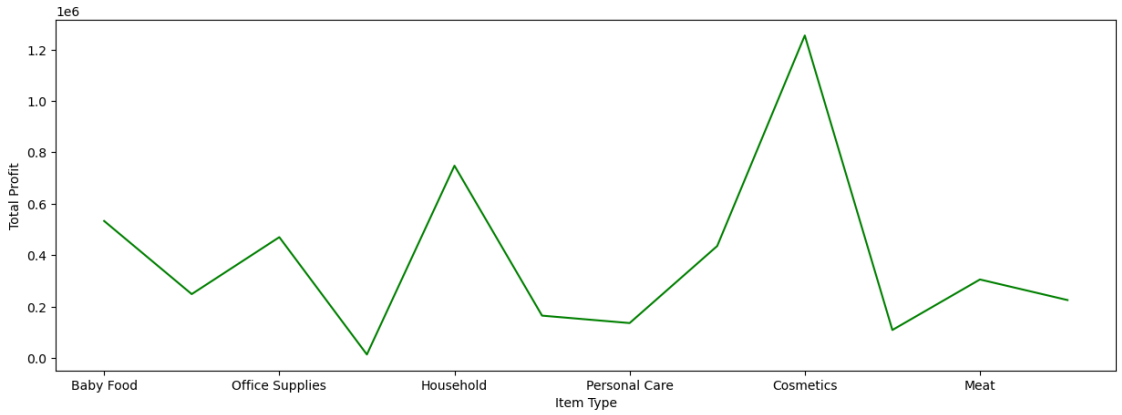
[74]:
```
np.corrcoef(data['Unit Price'],data['Total Cost'])
```

[74]:
```
array([[1.        , 0.78790543],
       [0.78790543, 1.        ]])
```

The high value of degree of correlation between 'Unit Price' and 'Total Cost' variables indicates that they are almost directly proportional to each other and highly dependent on each other.

[75]:
```
pd.pivot_table(data,index='Item Type',values='Total Profit',aggfunc=np.median).
  ↪plot(kind='line',color='green',figsize=(15,5),legend=None)
mp.ylabel('Total Profit')
```

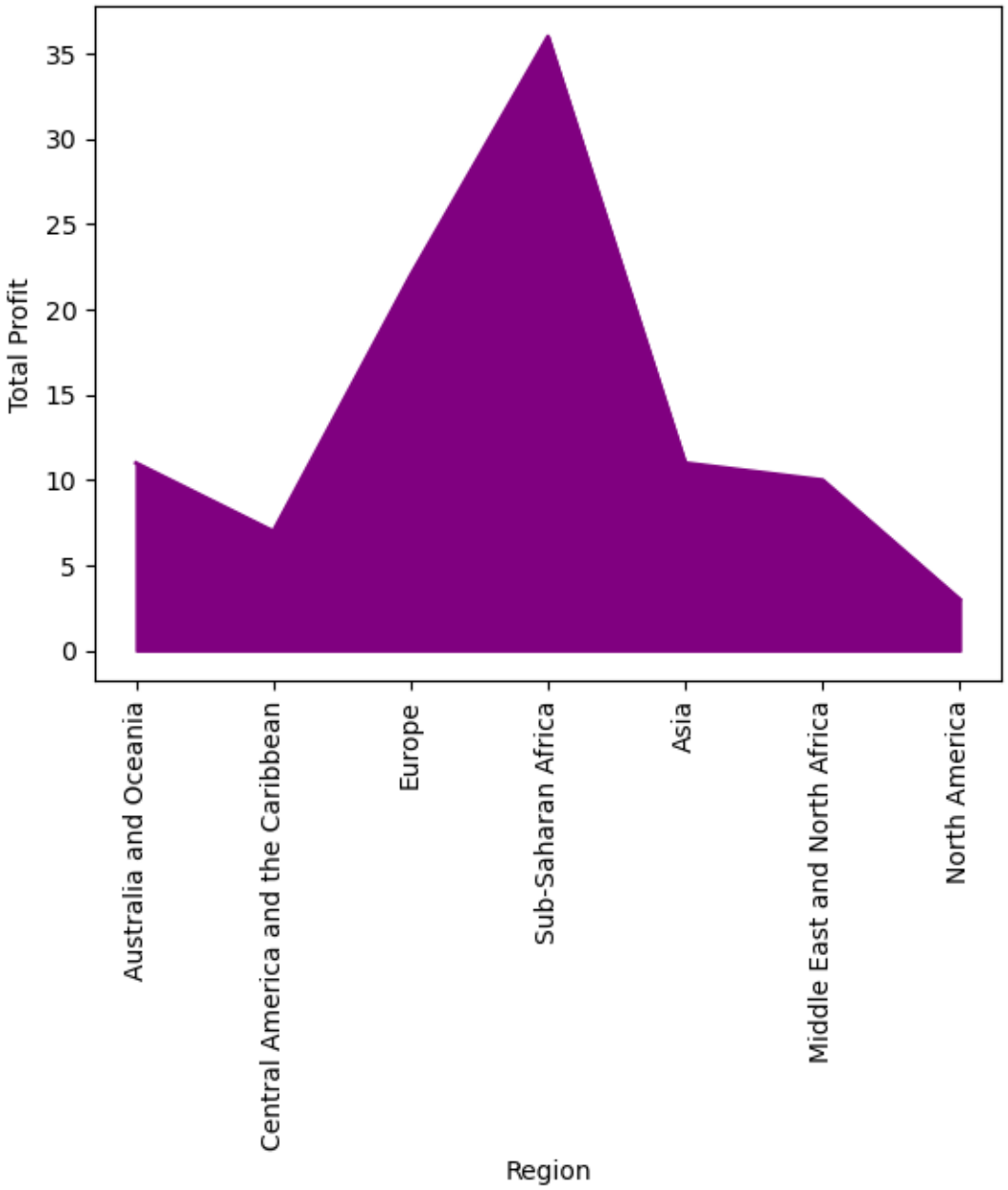
[75]:
```
Text(0, 0.5, 'Total Profit')
```



From the above plot we can conclude that maximum of the total profit is received by cosmetics item type.

[76]:
```
data.groupby('Region')['Total Profit'].count().
  ↪plot(kind='area',color=['purple','brown','blue','green'])
mp.xticks(rotation=90)
mp.ylabel('Total Profit')
```

[76]:
```
Text(0, 0.5, 'Total Profit')
```



from the above plot we can conclude that the Maximum profit has been generated in the Sub-Saharan African region while minimum profit has been generated in the North American region.

[77]:
```
data['Order Priority'].unique()
```

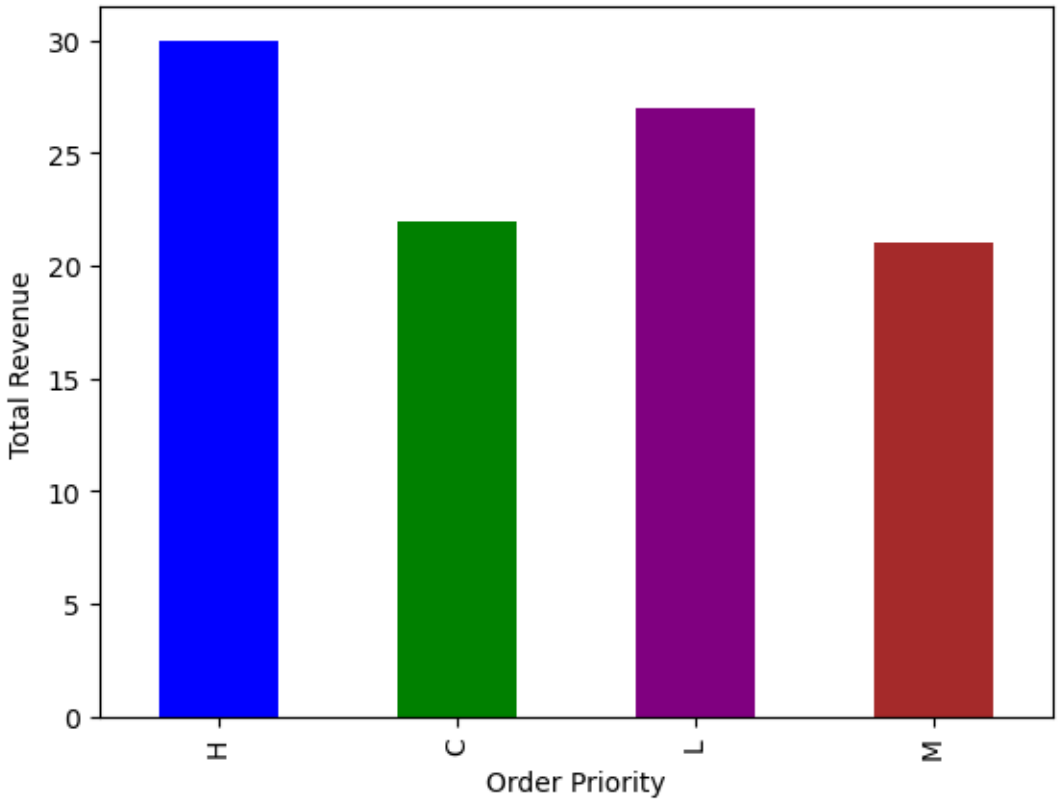
[77]:
```
array(['H', 'C', 'L', 'M'], dtype=object)
```

[78]:
```
order_priorities = ['H', 'C', 'L', 'M']
```

[79]:
```
data['Order Priority'] = pd.Categorical(data['Order␣
↪Priority'],categories=order_priorities,ordered=True)
```

[80]:
```
data.groupby('Order Priority')['Total Revenue'].count().
↪plot(kind='bar',color=['blue','green','purple','brown'])
mp.ylabel('Total Revenue')
```

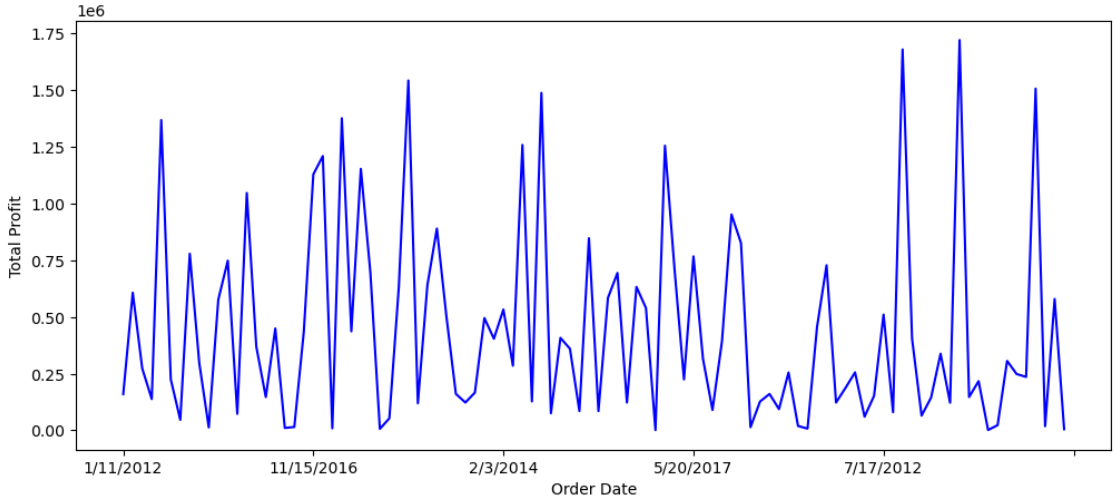
[80]: Text(0, 0.5, 'Total Revenue')



From the above bar graph we can conclude that maximum profit has been generated by products having order priority 'H' while minimum profit has been obtained in case of 'C' priority product orders.

[81]:
```
mp.figure(figsize=(12,5))
data.groupby('Order Date')['Total Profit'].sum().plot(kind='line',color='blue')
mp.ylabel('Total Profit')
```

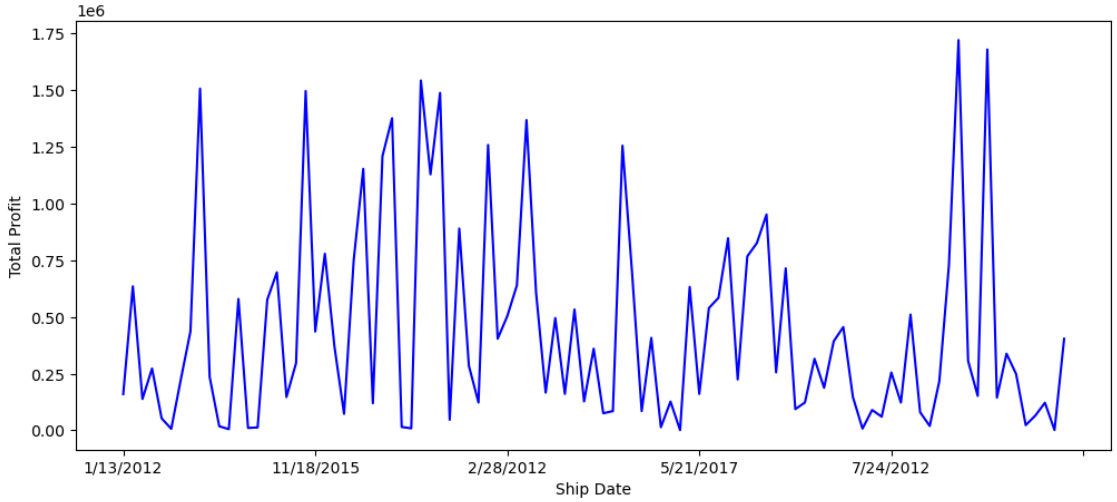
[81]: Text(0, 0.5, 'Total Profit')



From the above graph we can conclude that maximum profit has been achieved during the year 2012.

[82]:
```
mp.figure(figsize=(12,5))
data.groupby('Ship Date')['Total Profit'].sum().plot(kind='line',color='blue')
mp.ylabel('Total Profit')
```

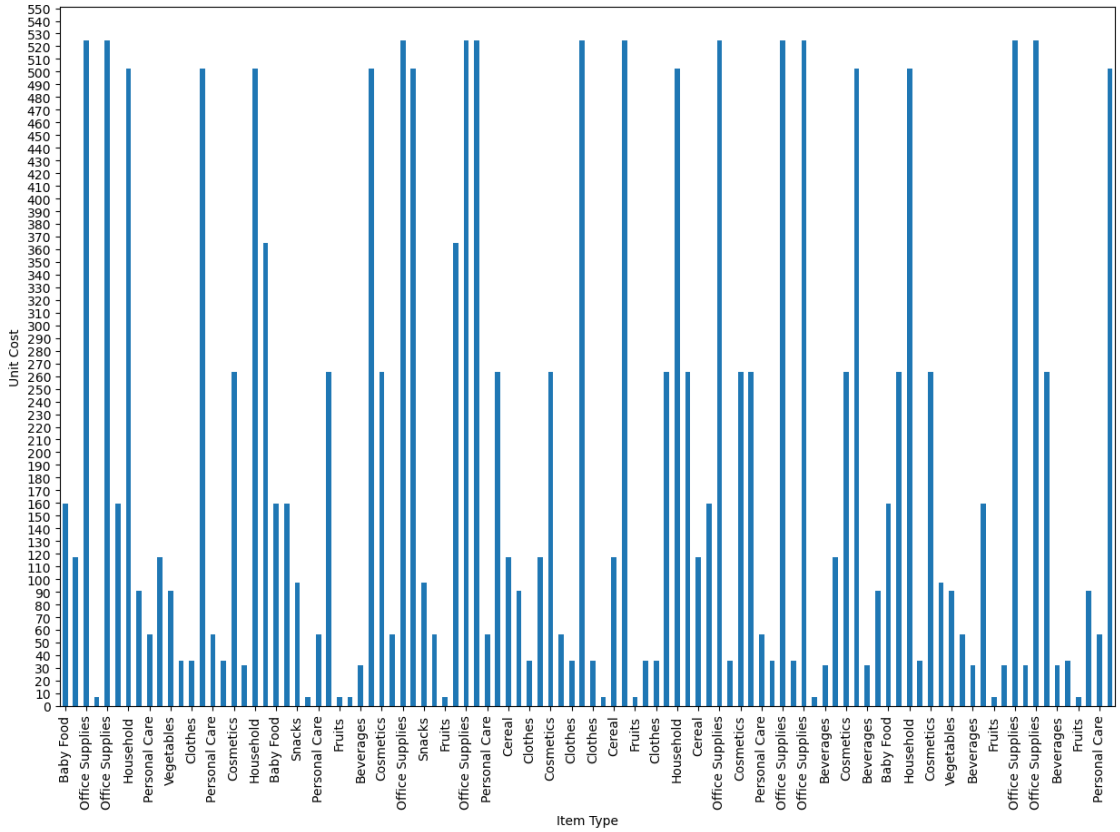
[82]: Text(0, 0.5, 'Total Profit')



As mentioned above maximum profit has been generated during the year 2012.

[83]:
```
data.plot.bar(x='Item Type',y='Unit Cost',legend=None,figsize=(15,10),rot=0)
mp.ylabel('Unit Cost')
```

```
mp.xticks(rotation=90)
mp.locator_params(nbins=90)
```



From the above bar plot we can conclude that office supplies and some items has the maximum unit cost and fruits has minimum unit cost.

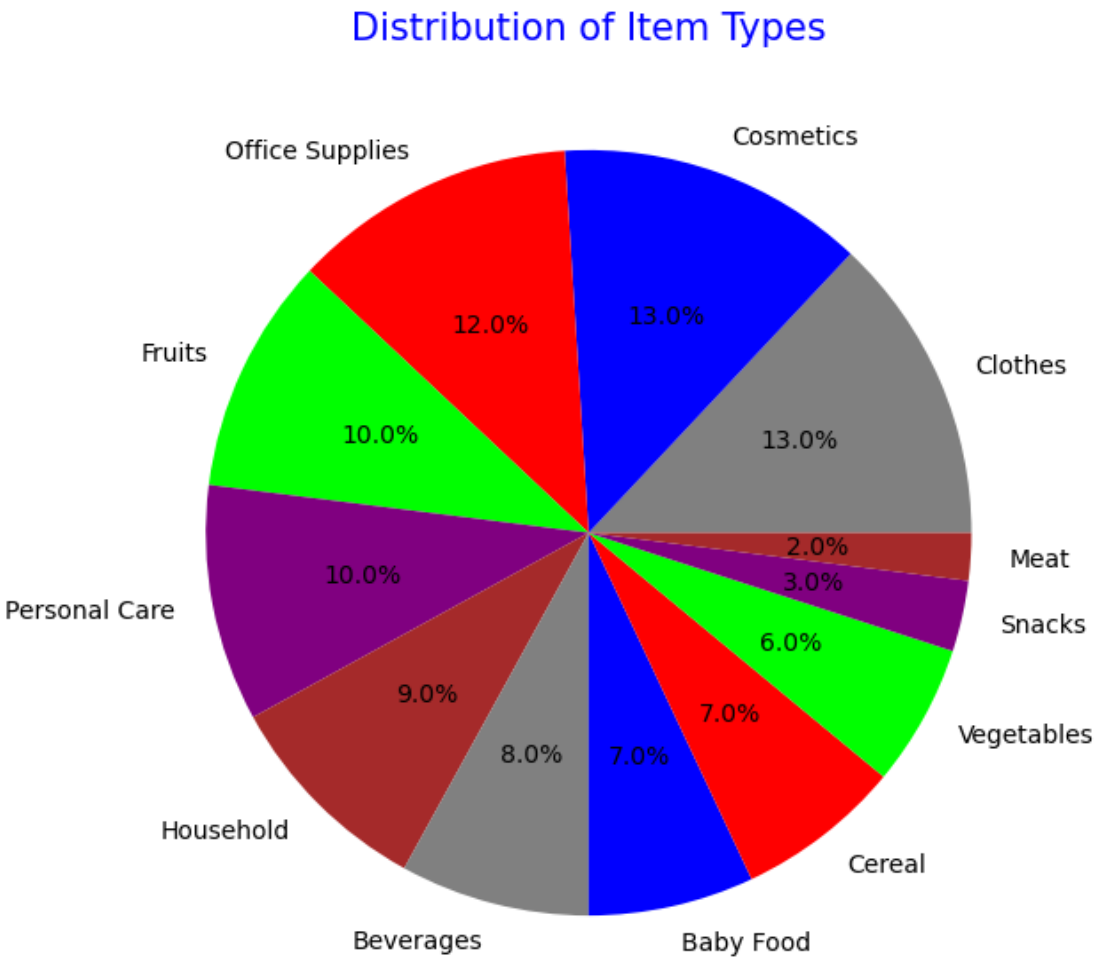
[84]:
```
data['Item Type'].dropna(inplace=True)
```

[85]:
```
labels = data['Item Type'].value_counts().index
```

[86]:
```
sizes = data['Item Type'].value_counts().values
colors = ['grey','blue','red','lime','purple','brown']
```

[87]:
```
mp.figure(figsize=(7,7))
mp.pie(sizes,labels=labels,colors=colors,autopct='%1.1f%%')
mp.title('Distribution of Item Types',fontsize=15,color='blue')
```

[87]:
```
Text(0.5, 1.0, 'Distribution of Item Types')
```



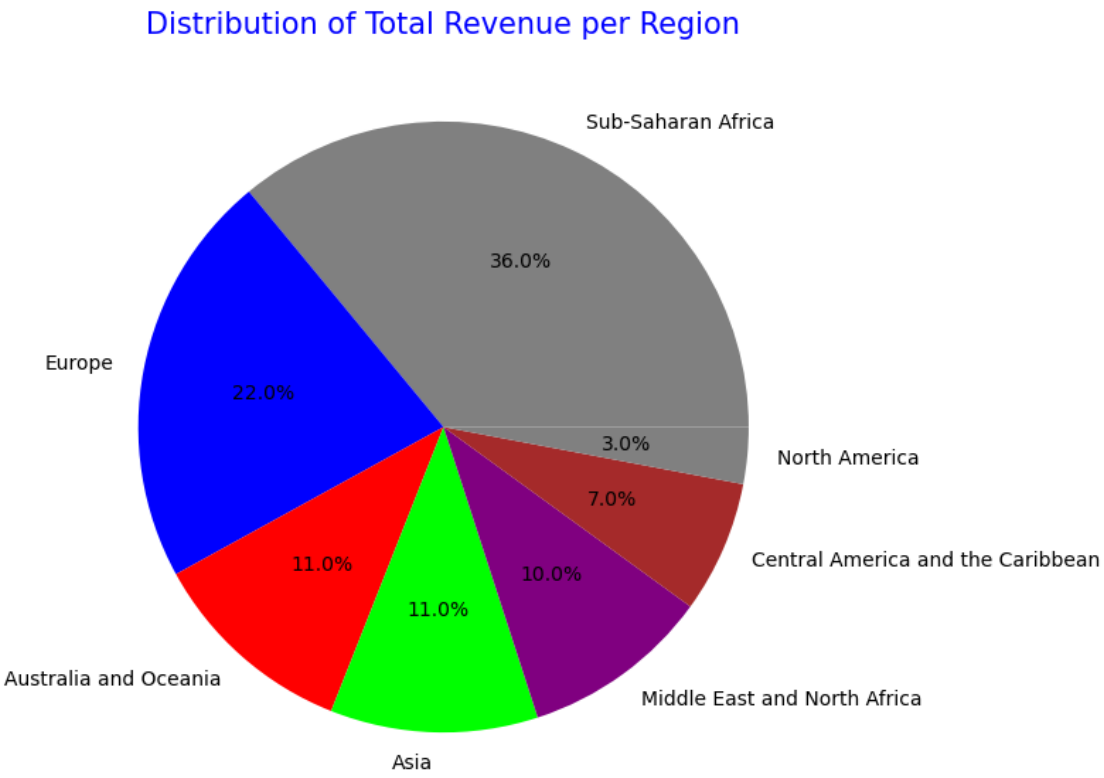
From the above pie chart we can conclude that clothes and cosmetics are the most purchased items while meat and snacks are the least purchased ones.

[88]:
```
labels = data['Region'].value_counts().index
```

[89]:
```
sizes = data['Region'].value_counts().values
colors = ['grey','blue','red','lime','purple','brown']
```

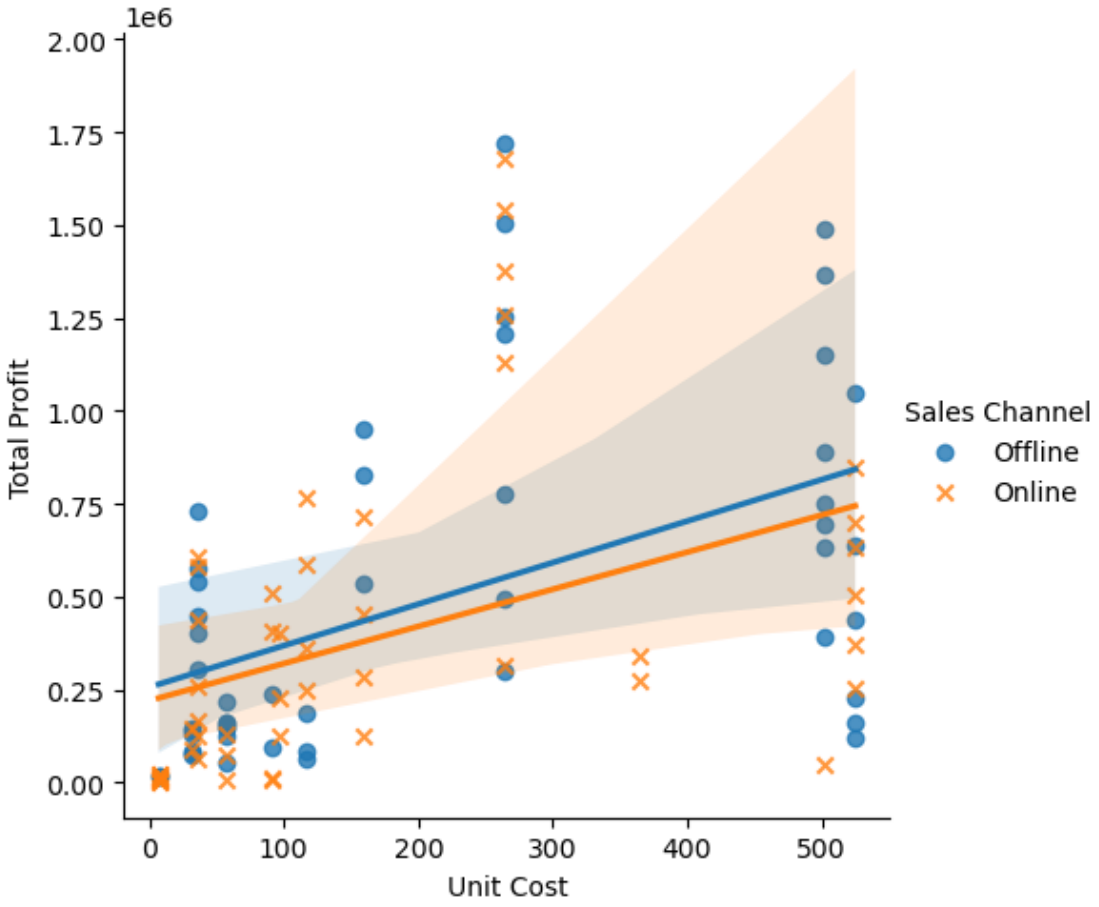
[90]:
```
mp.figure(figsize=(7,7))
mp.pie(sizes,labels=labels,colors=colors,autopct='%1.1f%%')
mp.title('Distribution of Total Revenue per Region',fontsize=15,color='blue')
```

[90]:
```
Text(0.5, 1.0, 'Distribution of Total Revenue per Region')
```



[91]:
```
sn.lmplot(x='Unit Cost',y='Total Profit',data=data,height=5,aspect=1,hue='Sales␣
↪Channel',logx=False,truncate=True,ci=100,y_jitter=2.
↪2,scatter=True,fit_reg=True,markers=['o','x'])
```

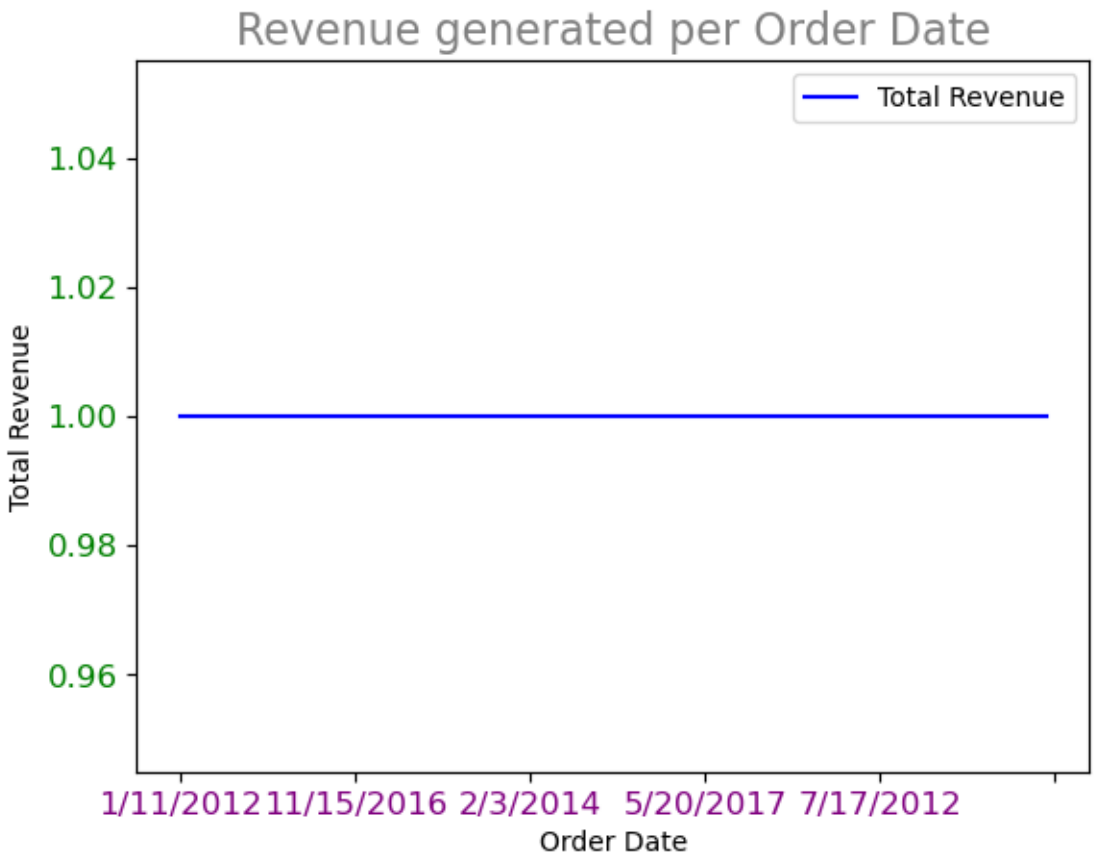
[91]:
```
<seaborn.axisgrid.FacetGrid at 0x218cec2c950>
```



From the above LM plot we can conclude that total profit keeps on increasing with increase in unit cost.

[92]:
```
pd.pivot_table(index='Order Date',values='Total␣
 ↪Revenue',data=data,aggfunc='count').
 ↪plot(kind='line',color='blue',legend=True)
mp.ylabel('Total Revenue')
mp.yticks(fontsize=12,color='green')
mp.xticks(fontsize=12,color='purple')
mp.title('Revenue generated per Order Date',fontsize=16,color='grey')
```

[92]:
```
Text(0.5, 1.0, 'Revenue generated per Order Date')
```



From the above plot we can conclude that the total revenue remains constant for every year.