Mini Project

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

1.Merge the files in one dataframe

In [2]:
```
files = [file for file in os.listdir(r'C:\Users\keerti chouhan\Desktop\sales')

sales_df = pd.DataFrame()

for file in files:
    df = pd.read_csv('C:/Users/keerti chouhan/Desktop/sales/'+file)
    sales_df = pd.concat([sales_df,df])
```

In [3]:
```
sales_df.head()
```

Out[3]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	176558	USB-C Charging Cable	2	11.95	04/19/19 8:46	917 1st St, Dallas, TX 75001
1	NaN	NaN	NaN	NaN	NaN	NaN
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215
3	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001

In [4]:
```
sales_df.tail()
```

Out[4]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
11681	259353	AAA Batteries (4-pack)	3	2.99	09/17/19 20:56	840 Highland St, Los Angeles, CA 90001
11682	259354	iPhone	1	700	09/01/19 16:00	216 Dogwood St, San Francisco, CA 94016
11683	259355	iPhone	1	700	09/23/19 7:39	220 12th St, San Francisco, CA 94016
11684	259356	34in Ultrawide Monitor	1	379.99	09/19/19 17:30	511 Forest St, San Francisco, CA 94016
11685	259357	USB-C Charging Cable	1	11.95	09/30/19 0:18	250 Meadow St, San Francisco, CA 94016

In [5]:
```
sales_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 186850 entries, 0 to 11685
Data columns (total 6 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   Order ID          186305 non-null  object
 1   Product           186305 non-null  object
 2   Quantity Ordered  186305 non-null  object
 3   Price Each        186305 non-null  object
 4   Order Date        186305 non-null  object
 5   Purchase Address  186305 non-null  object
dtypes: object(6)
memory usage: 10.0+ MB
```

2.Clean the data.

In [6]:
```
sales_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 186850 entries, 0 to 11685
Data columns (total 6 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   Order ID          186305 non-null  object
 1   Product           186305 non-null  object
 2   Quantity Ordered  186305 non-null  object
 3   Price Each        186305 non-null  object
 4   Order Date        186305 non-null  object
 5   Purchase Address  186305 non-null  object
dtypes: object(6)
memory usage: 10.0+ MB
```

In [7]:
```
sales_df.isna().sum()
```

Out[7]:
```
Order ID            545
Product             545
Quantity Ordered    545
Price Each          545
Order Date          545
Purchase Address    545
dtype: int64
```

In [8]:
```
sales_df=sales_df.dropna()
```

In [9]:
```
sales_df
```

Out[9]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	176558	USB-C Charging Cable	2	11.95	04/19/19 8:46	917 1st St, Dallas, TX 75001
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215
3	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
5	176561	Wired Headphones	1	11.99	04/30/19 9:27	333 8th St, Los Angeles, CA 90001
...	...	...	...	...	...	...
11681	259353	AAA Batteries (4-pack)	3	2.99	09/17/19 20:56	840 Highland St, Los Angeles, CA 90001
11682	259354	iPhone	1	700	09/01/19 16:00	216 Dogwood St, San Francisco, CA 94016
11683	259355	iPhone	1	700	09/23/19 7:39	220 12th St, San Francisco, CA 94016
11684	259356	34in Ultrawide Monitor	1	379.99	09/19/19 17:30	511 Forest St, San Francisco, CA 94016
11685	259357	USB-C Charging Cable	1	11.95	09/30/19 0:18	250 Meadow St, San Francisco, CA 94016

186305 rows × 6 columns

In [10]:
```
sales_df.isna().sum()
```

Out[10]:
```
Order ID            0
Product             0
Quantity Ordered    0
Price Each          0
Order Date          0
Purchase Address    0
dtype: int64
```

In [11]:
```
sales_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 186305 entries, 0 to 11685
Data columns (total 6 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   Order ID          186305 non-null  object
 1   Product           186305 non-null  object
 2   Quantity Ordered  186305 non-null  object
 3   Price Each        186305 non-null  object
 4   Order Date        186305 non-null  object
 5   Purchase Address  186305 non-null  object
dtypes: object(6)
memory usage: 9.9+ MB
```

3.Change the object type column into integer type or float type.

In [12]:
```
sales_df[sales_df["Price Each"]=="Price Each"].index
```

Out[12]:
```
Int64Index([  519,  1149,  1155,  2878,  2893,  3036,  3209,  3618,  4138,
             4645,
            ...
             8644,  9325,  9502,  9615,  9954, 10000, 10387, 11399, 11468,
            11574],
           dtype='int64', length=355)
```

In [13]:
```
sales_df.drop(sales_df[sales_df["Price Each"]=="Price Each"].index,inplace=Tru
```

In [14]:
```
sales_df[sales_df["Price Each"]=="Price Each"].index
```

Out[14]:
```
Int64Index([], dtype='int64')
```

In [15]:
```
sales_df["Price Each"]=sales_df["Price Each"].astype('float64')
```

In [16]:
```
sales_df["Order ID"]=sales_df["Order ID"].astype('int64')
```

In [17]:
```
sales_df["Quantity Ordered"]=sales_df["Quantity Ordered"].astype('int32')
```

In [18]:
```
sales_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 182735 entries, 0 to 11685
Data columns (total 6 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   Order ID          182735 non-null  int64  
 1   Product           182735 non-null  object 
 2   Quantity Ordered  182735 non-null  int32  
 3   Price Each        182735 non-null  float64
 4   Order Date        182735 non-null  object 
 5   Purchase Address  182735 non-null  object 
dtypes: float64(1), int32(1), int64(1), object(3)
memory usage: 9.1+ MB
```

4.Get the month value from the order date?

In [19]:
```
import datetime
month=pd.to_datetime(sales_df['Order Date']).dt.month
month
```

Out[19]:
```
0        4
2        4
3        4
4        4
5        4
        ..
11681    9
11682    9
11683    9
11684    9
11685    9
Name: Order Date, Length: 182735, dtype: int64
```

In [20]:
```
sales_df['month']=month
sales_df
```

Out[20]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	month
0	176558	USB-C Charging Cable	2	11.95	04/19/19 8:46	917 1st St, Dallas, TX 75001	4
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4
5	176561	Wired Headphones	1	11.99	04/30/19 9:27	333 8th St, Los Angeles, CA 90001	4
...	...	...	...	...	...	...	...
11681	259353	AAA Batteries (4-pack)	3	2.99	09/17/19 20:56	840 Highland St, Los Angeles, CA 90001	9
11682	259354	iPhone	1	700.00	09/01/19 16:00	216 Dogwood St, San Francisco, CA 94016	9
11683	259355	iPhone	1	700.00	09/23/19 7:39	220 12th St, San Francisco, CA 94016	9
11684	259356	34in Ultrawide Monitor	1	379.99	09/19/19 17:30	511 Forest St, San Francisco, CA 94016	9
11685	259357	USB-C Charging Cable	1	11.95	09/30/19 0:18	250 Meadow St, San Francisco, CA 94016	9

182735 rows × 7 columns

5.Which was the most productive month in terms of sales?

In [21]:
```
sales_df['Sales']=sales_df['Quantity Ordered']*sales_df['Price Each']
```

In [22]:
```
sales_df.head()
```

Out[22]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	month	Sales
0	176558	USB-C Charging Cable	2	11.95	04/19/19 8:46	917 1st St, Dallas, TX 75001	4	23.90
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4	99.99
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	11.99
5	176561	Wired Headphones	1	11.99	04/30/19 9:27	333 8th St, Los Angeles, CA 90001	4	11.99

In [23]:
```
a=sales_df.groupby('month')["Sales"].sum().sort_values(ascending=False)
a
```

Out[23]:
```
month
12    4557905.42
10    3679254.16
4     3336376.42
11    3149785.09
5     3101881.04
3     2755969.40
7     2587444.91
6     2524464.99
8     2191698.31
2     2158127.48
9     2050361.26
1     1786511.29
Name: Sales, dtype: float64
```

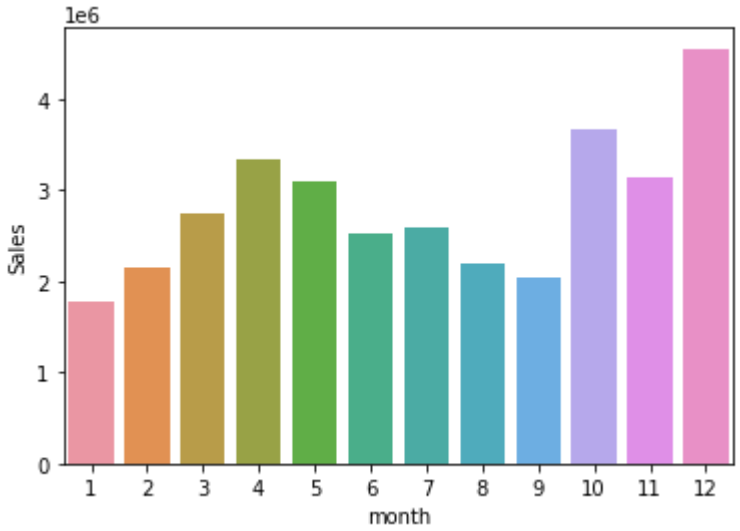
In [24]:
```
c_data = sales_df.groupby(by = 'month', as_index = False )['Sales'].sum()
c_data = c_data.sort_values(by = 'Sales', ascending = False)
c_data.head()
```

Out[24]:

	month	Sales
11	12	4557905.42
9	10	3679254.16
3	4	3336376.42
10	11	3149785.09
4	5	3101881.04

In [25]:
```
sns.barplot(x='month',y='Sales',data=c_data,ci=None)
```

Out[25]:
```
<AxesSubplot:xlabel='month', ylabel='Sales'>
```



The most productive month in terms of sales is December.

6.Which city had the highest number of sales?

In [26]:
```
sales_df['city']=sales_df['Purchase Address'].apply(lambda x: x.split(',')[1])
sales_df.head()
```

Out[26]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	month	Sales	city
0	176558	USB-C Charging Cable	2	11.95	04/19/19 8:46	917 1st St, Dallas, TX 75001	4	23.90	Dallas
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4	99.99	Boston
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles
5	176561	Wired Headphones	1	11.99	04/30/19 9:27	333 8th St, Los Angeles, CA 90001	4	11.99	Los Angeles

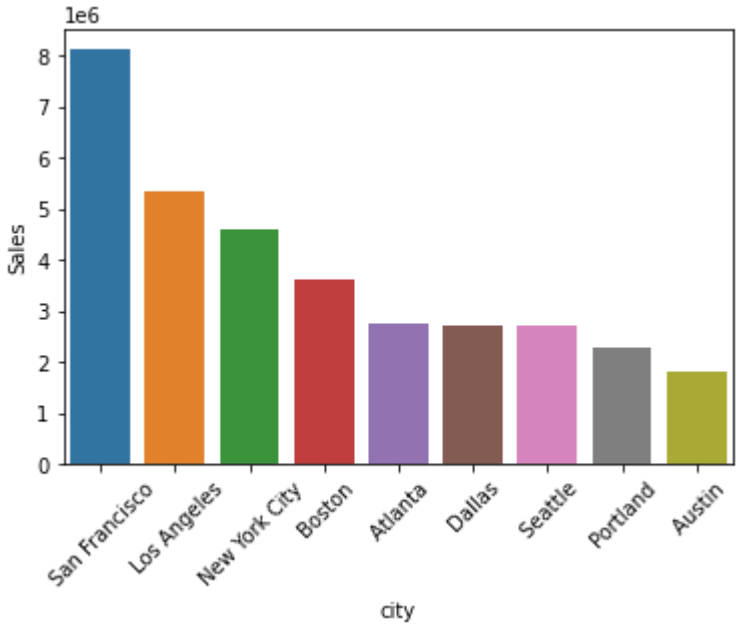
In [27]:
```
c_data = sales_df.groupby(by = 'city', as_index = False )['Sales'].sum()

c_data = c_data.sort_values(by = 'Sales', ascending = False)

c_data.head()
```

Out[27]:

	city	Sales
7	San Francisco	8124120.94
4	Los Angeles	5354039.93
5	New York City	4581658.91
2	Boston	3604080.86
0	Atlanta	2741642.05

In [28]:
```
sns.barplot(x='city',y='Sales',data=c_data,ci=None)
plt.xticks(rotation=45);
```

Out[28]:
```
(array([0, 1, 2, 3, 4, 5, 6, 7, 8]),
 [Text(0, 0, ' San Francisco'),
  Text(1, 0, ' Los Angeles'),
  Text(2, 0, ' New York City'),
  Text(3, 0, ' Boston'),
  Text(4, 0, ' Atlanta'),
  Text(5, 0, ' Dallas'),
  Text(6, 0, ' Seattle'),
  Text(7, 0, ' Portland'),
  Text(8, 0, ' Austin')])
```



The city had the highest number of sales is : San Francisco

7.At what time people mostly purchase the product?

In [29]:
```
hour= pd.to_datetime(sales_df['Order Date']).dt.hour
hour
```

Out[29]:
```
0         8
2        22
3        14
4        14
5         9
        ..
11681    20
11682    16
11683     7
11684    17
11685     0
Name: Order Date, Length: 182735, dtype: int64
```

In [30]:
```
sales_df['hour']=hour
sales_df.head()
```

Out[30]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	month	Sales	city	hour
0	176558	USB-C Charging Cable	2	11.95	04/19/19 8:46	917 1st St, Dallas, TX 75001	4	23.90	Dallas	8
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4	99.99	Boston	22
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles	14
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles	14
5	176561	Wired Headphones	1	11.99	04/30/19 9:27	333 8th St, Los Angeles, CA 90001	4	11.99	Los Angeles	9

In [37]:
```
k=sales_df.groupby(by='hour' ,as_index=False)['Quantity Ordered'].sum()
k.sort_values(by='Quantity Ordered' ,ascending=False)
```

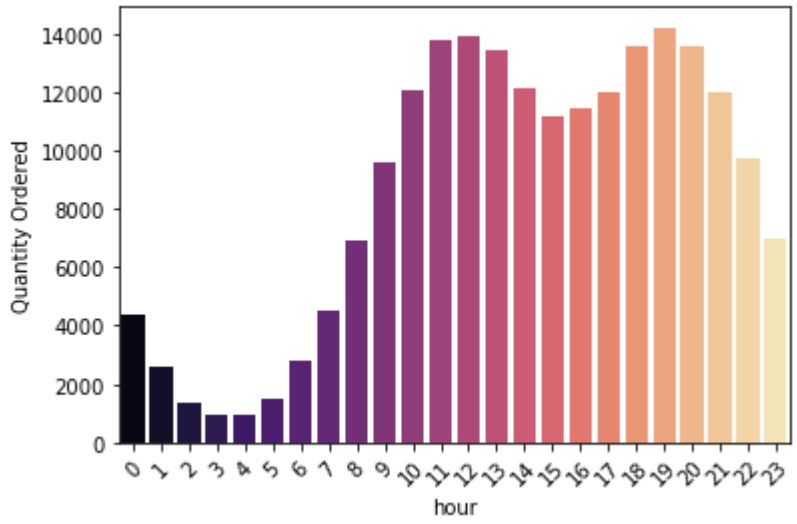
Out[37]:

	hour	Quantity Ordered
19	19	14228
12	12	13946
11	11	13763
18	18	13576
20	20	13565
13	13	13437
14	14	12163
10	10	12058
21	21	12030
17	17	12016
16	16	11420
15	15	11175
22	22	9715
9	9	9628
23	23	6955
8	8	6917
7	7	4483
0	0	4355
6	6	2767
1	1	2579
5	5	1463
2	2	1379
4	4	925
3	3	912

In [31]:
```
sales_df['hour'].value_counts().head()
```

Out[31]:
```
19    12685
12    12360
11    12202
18    12074
20    12040
Name: hour, dtype: int64
```

In [42]:
```
sns.barplot(x='hour',y='Quantity Ordered',data=k,palette='magma')
plt.xticks(rotation=45);
```



The Time in which people mostly purchase the product is between (18-21)

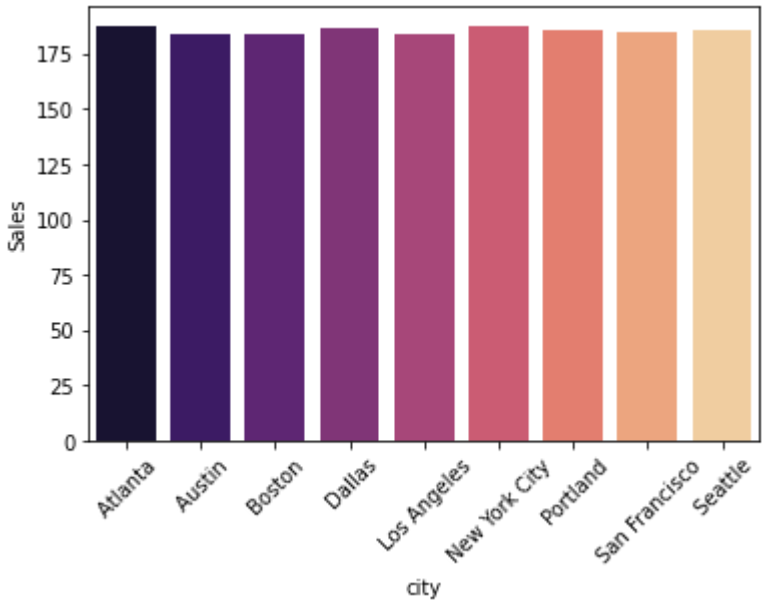
8What is the average purchase by city?

In [48]:
```
g=sales_df.groupby(by='city' ,as_index=False)['Sales'].mean()
g.sort_values(by='Sales' ,ascending=False)
```

Out[48]:

	city	Sales
0	Atlanta	187.578137
5	New York City	187.342939
3	Dallas	186.520741
8	Seattle	185.894153
6	Portland	185.682183
7	San Francisco	184.857580
2	Boston	184.116519
4	Los Angeles	183.943379
1	Austin	183.935096

In [50]:
```
sns.barplot(x='city',y='Sales',data=g,palette='magma')
plt.xticks(rotation=45);
```



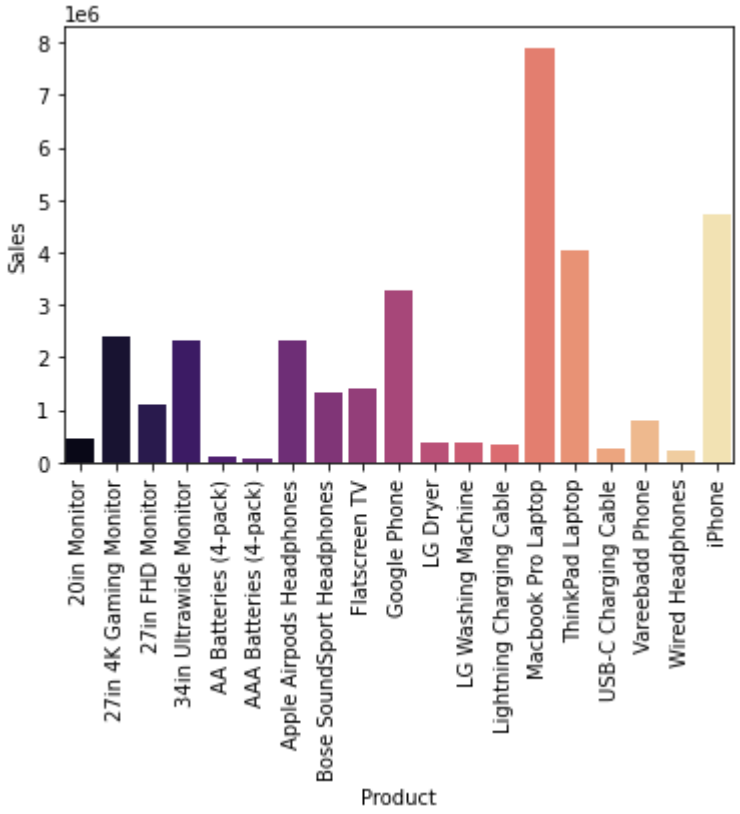
9.Which product has the highest sales?

In [53]:
```
g=sales_df.groupby(by='Product' ,as_index=False)['Sales'].sum()
g.sort_values(by='Sales' ,ascending=False)
```

Out[53]:

	Product	Sales
13	Macbook Pro Laptop	7896500.00
18	iPhone	4712400.00
14	ThinkPad Laptop	4053959.46
9	Google Phone	3264000.00
1	27in 4K Gaming Monitor	2392198.66
3	34in Ultrawide Monitor	2308819.24
6	Apple Airpods Headphones	2307450.00
8	Flatscreen TV	1417200.00
7	Bose SoundSport Headphones	1323467.64
2	27in FHD Monitor	1114275.71
16	Vareebadd Phone	809200.00
0	20in Monitor	446339.42
11	LG Washing Machine	389400.00
10	LG Dryer	384000.00
12	Lightning Charging Cable	341472.95
15	USB-C Charging Cable	281482.25
17	Wired Headphones	242209.99
4	AA Batteries (4-pack)	104248.32
5	AAA Batteries (4-pack)	91156.13

In [58]:
```
sns.barplot(x='Product',y='Sales',data=g,palette='magma');
plt.xticks(rotation=90);
```



Macbook Pro Laptop product has the highest sales

10. In Month of September, which product has the lowest sales?

In [63]:
```
sept=sales_df[sales_df['month']== 9]
sept
```

Out[63]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	month	Sales	city	hou
2251	238834	Apple Airpods Headphones	1	150.00	09/01/19 4:13	761 Forest St, San Francisco, CA 94016	9	150.00	San Francisco	
2731	239285	34in Ultrawide Monitor	1	379.99	09/01/19 1:09	373 1st St, San Francisco, CA 94016	9	379.99	San Francisco	
4124	240636	Lightning Charging Cable	1	14.95	09/01/19 2:07	63 1st St, Seattle, WA 98101	9	14.95	Seattle	
4569	241054	AAA Batteries (4-pack)	1	2.99	09/01/19 0:25	175 South St, San Francisco, CA 94016	9	2.99	San Francisco	
5914	242343	ThinkPad Laptop	1	999.99	09/01/19 2:44	510 Park St, Boston, MA 02215	9	999.99	Boston	
...	...	...	...	...	...	...	...	...	...	.
11681	259353	AAA Batteries (4-pack)	3	2.99	09/17/19 20:56	840 Highland St, Los Angeles, CA 90001	9	8.97	Los Angeles	2
11682	259354	iPhone	1	700.00	09/01/19 16:00	216 Dogwood St, San Francisco, CA 94016	9	700.00	San Francisco	1
11683	259355	iPhone	1	700.00	09/23/19 7:39	220 12th St, San Francisco, CA 94016	9	700.00	San Francisco	
11684	259356	34in Ultrawide Monitor	1	379.99	09/19/19 17:30	511 Forest St, San Francisco, CA 94016	9	379.99	San Francisco	1
11685	259357	USB-C Charging Cable	1	11.95	09/30/19 0:18	250 Meadow St, San Francisco, CA 94016	9	11.95	San Francisco	

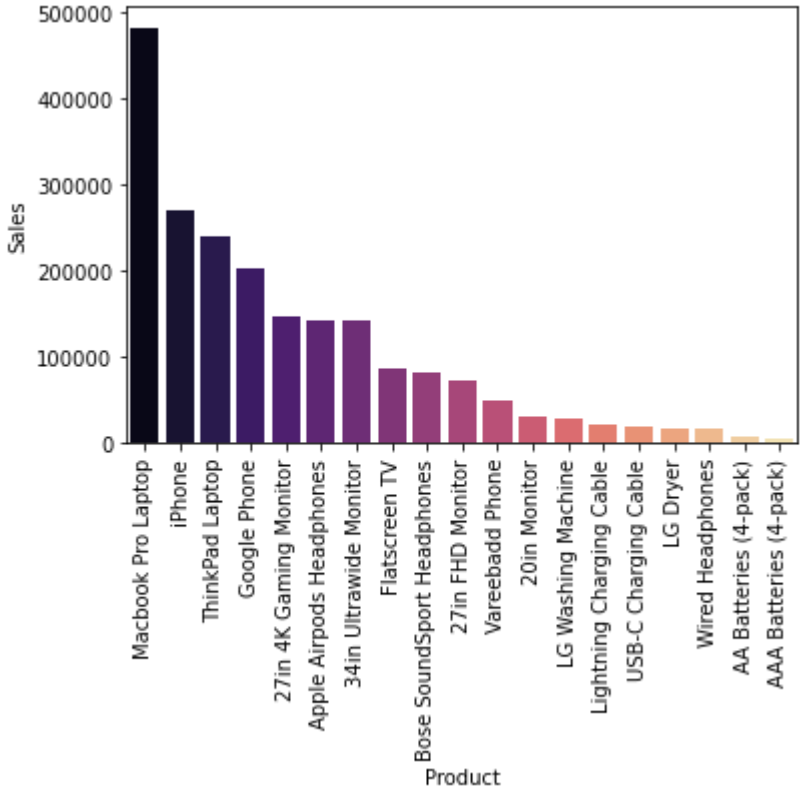
11375 rows × 10 columns

In [64]:
```
g=sept.groupby(by='Product' ,as_index=False)['Sales'].sum()
t=g.sort_values(by='Sales' ,ascending=False)
t
```

Out[64]:

	Product	Sales
13	Macbook Pro Laptop	482800.00
18	iPhone	269500.00
14	ThinkPad Laptop	239997.60
9	Google Phone	201600.00
1	27in 4K Gaming Monitor	146246.25
6	Apple Airpods Headphones	142500.00
3	34in Ultrawide Monitor	140976.29
8	Flatscreen TV	85500.00
7	Bose SoundSport Headphones	80891.91
2	27in FHD Monitor	71095.26
16	Vareebadd Phone	48000.00
0	20in Monitor	29587.31
11	LG Washing Machine	27600.00
12	Lightning Charging Cable	21049.60
15	USB-C Charging Cable	18713.70
10	LG Dryer	16800.00
17	Wired Headphones	15251.28
4	AA Batteries (4-pack)	6600.96
5	AAA Batteries (4-pack)	5651.10

In [66]:
```
sns.barplot(x='Product',y='Sales',data=t,palette='magma')
plt.xticks(rotation=90);
```



In month of sept,AAA Batteries (4-Pack) has lowest sale

In [ ]:
```
Question -Answers
```

5. Which was the most productive month in terms of sales?

December was the most productive month in terms of sales

6. Which city had the highest number of sales?

San Francisco has the highest number of sales

7. At what time people mostly purchase the product?

At 19:00 people mostly purchase the product

8. What is the average purchase by city?

Atlanta has the maximum average purchase by the cities.

9. Which product has the highest sales?

Macbook Pro Laptop has the highest sales.

10. In Month of September, which product has the lowest sales?

In month of september AAA Batteries (4-pack) has the lowest sale.

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: