In [8]:
```
import pandas as pd
import os

all_files_data = pd.DataFrame()

files = [file for file in os.listdir("./Sales_Data")]

for file in files:

    df = pd.read_csv("./Sales_Data/"+file)

    all_files_data = pd.concat([all_files_data, df])

all_files_data.to_csv("./Sales_Data/all_months_data.csv", index=False)
```

Out[8]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
373690	259349	AAA Batteries (4-pack)	1	2.99	09/01/19 22:14	911 River St, Dallas, TX 75001
373691	259350	Google Phone	1	600	09/30/19 13:49	519 Maple St, San Francisco, CA 94016
373692	259350	USB-C Charging Cable	1	11.95	09/30/19 13:49	519 Maple St, San Francisco, CA 94016
373693	259351	Apple Airpods Headphones	1	150	09/01/19 19:43	981 4th St, New York City, NY 10001
373694	259352	USB-C Charging Cable	1	11.95	09/07/19 15:49	976 Forest St, San Francisco, CA 94016
373695	259353	AAA Batteries (4-pack)	3	2.99	09/17/19 20:56	840 Highland St, Los Angeles, CA 90001
373696	259354	iPhone	1	700	09/01/19 16:00	216 Dogwood St, San Francisco, CA 94016
373697	259355	iPhone	1	700	09/23/19 07:39	220 12th St, San Francisco, CA 94016
373698	259356	34in Ultrawide Monitor	1	379.99	09/19/19 17:30	511 Forest St, San Francisco, CA 94016
373699	259357	USB-C Charging Cable	1	11.95	09/30/19 00:18	250 Meadow St, San Francisco, CA 94016

In [9]:
```
all_data = pd.read_csv("./Sales_Data/all_months_data.csv")

all_data.tail(10)
```

Out[9]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
373690	259349	AAA Batteries (4-pack)	1	2.99	09/01/19 22:14	911 River St, Dallas, TX 75001
373691	259350	Google Phone	1	600	09/30/19 13:49	519 Maple St, San Francisco, CA 94016
373692	259350	USB-C Charging Cable	1	11.95	09/30/19 13:49	519 Maple St, San Francisco, CA 94016
373693	259351	Apple Airpods Headphones	1	150	09/01/19 19:43	981 4th St, New York City, NY 10001
373694	259352	USB-C Charging Cable	1	11.95	09/07/19 15:49	976 Forest St, San Francisco, CA 94016
373695	259353	AAA Batteries (4-pack)	3	2.99	09/17/19 20:56	840 Highland St, Los Angeles, CA 90001
373696	259354	iPhone	1	700	09/01/19 16:00	216 Dogwood St, San Francisco, CA 94016
373697	259355	iPhone	1	700	09/23/19 07:39	220 12th St, San Francisco, CA 94016
373698	259356	34in Ultrawide Monitor	1	379.99	09/19/19 17:30	511 Forest St, San Francisco, CA 94016
373699	259357	USB-C Charging Cable	1	11.95	09/30/19 00:18	250 Meadow St, San Francisco, CA 94016

In [10]:
```
all_data['Month'] = all_data['Order Date'].str[0:2]
all_data.head()
```

Out[10]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	04
1	NaN	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	04
3	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	04
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	04

In [13]:
```
all_data.info()
all_data.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 373700 entries, 0 to 373699
Data columns (total 7 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   Order ID          372610 non-null  object
 1   Product           372610 non-null  object
 2   Quantity Ordered  372610 non-null  object
 3   Price Each        372610 non-null  object
 4   Order Date        372610 non-null  object
 5   Purchase Address  372610 non-null  object
 6   Month             372610 non-null  object
dtypes: object(7)
memory usage: 20.0+ MB
```

Out[13]:
```
Order ID            1090
Product             1090
Quantity Ordered    1090
Price Each          1090
Order Date          1090
Purchase Address    1090
Month               1090
dtype: int64
```

In [17]:
```
nan_df = all_data[all_data.isna().any(axis=1)]
nan_df.head()
all_data = all_data.dropna()
all_data.isnull().sum()
```

Out[17]:
```
Order ID            0
Product             0
Quantity Ordered    0
Price Each          0
Order Date          0
Purchase Address    0
Month               0
dtype: int64
```

In [24]:
```
all_data = all_data[all_data['Order Date'].str[0:2] != 'Or']
```

In [25]:
```
all_data['Month'] = all_data['Month'].astype("int8")
```

In [29]:
```
all_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 371900 entries, 0 to 373699
Data columns (total 7 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   Order ID          371900 non-null  object
 1   Product           371900 non-null  object
 2   Quantity Ordered  371900 non-null  object
 3   Price Each        371900 non-null  object
 4   Order Date        371900 non-null  object
 5   Purchase Address  371900 non-null  object
 6   Month             371900 non-null  int8  
dtypes: int8(1), object(6)
memory usage: 20.2+ MB
```

In [30]:
```
all_data.head()
```

Out[30]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	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	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 09:27	333 8th St, Los Angeles, CA 90001	4

In [31]:
```
all_data['Quantity Ordered'] = pd.to_numeric(all_data["Quantity Ordered"])
all_data['Price Each'] = pd.to_numeric(all_data['Price Each'])
all_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 371900 entries, 0 to 373699
Data columns (total 7 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   Order ID          371900 non-null  object 
 1   Product           371900 non-null  object 
 2   Quantity Ordered  371900 non-null  int64  
 3   Price Each        371900 non-null  float64
 4   Order Date        371900 non-null  object 
 5   Purchase Address  371900 non-null  object 
 6   Month             371900 non-null  int8   
dtypes: float64(1), int64(1), int8(1), object(4)
memory usage: 20.2+ MB
```

In [32]:
```
all_data['Sales'] = all_data['Quantity Ordered'] * all_data['Price Each']
all_data.head()
```

Out[32]:

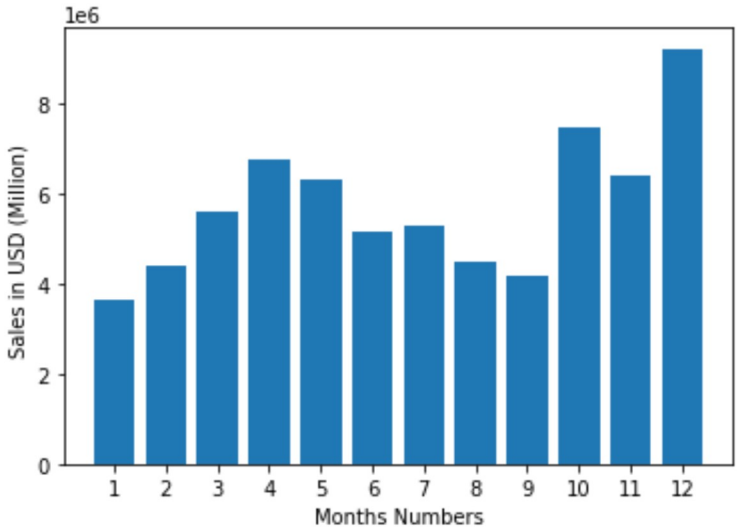
	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	4	23.90
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4	99.99
3	176560	Google Phone	1	600.00	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 09:27	333 8th St, Los Angeles, CA 90001	4	11.99

In [60]:
```
results = all_data.groupby('Month').sum()
print(results)
```

```
       Quantity Ordered    Price Each         Sales
Month
1                 21806  3.623537e+06  3.644513e+06
2                 26898  4.377769e+06  4.404045e+06
3                 34010  5.582416e+06  5.614201e+06
4                 41116  6.735342e+06  6.781340e+06
5                 37334  6.270250e+06  6.305214e+06
6                 30506  5.124051e+06  5.155605e+06
7                 32144  5.265079e+06  5.295552e+06
8                 26896  4.460691e+06  4.488936e+06
9                 26218  4.169984e+06  4.195120e+06
10                45406  7.431110e+06  7.473454e+06
11                39596  6.361201e+06  6.399206e+06
12                56228  9.176831e+06  9.226887e+06
```

In [71]:
```
import matplotlib.pyplot as plt
import numpy as np
months = range(1,13)

plt.bar(months, results['Sales'])
plt.xticks(months)
plt.xlabel('Months Numbers')
plt.ylabel("Sales in USD (Million) ")
plt.show()
```



In [76]:
```
def get_city(address):

    return address.split(',')[1]

def get_state(address):

    return address.split(',')[2].split(' ')[1]

all_data['City'] = all_data['Purchase Address'].apply(lambda x: get_cit
y(x) + ', ' + get_state(x))

all_data.head(10)
```

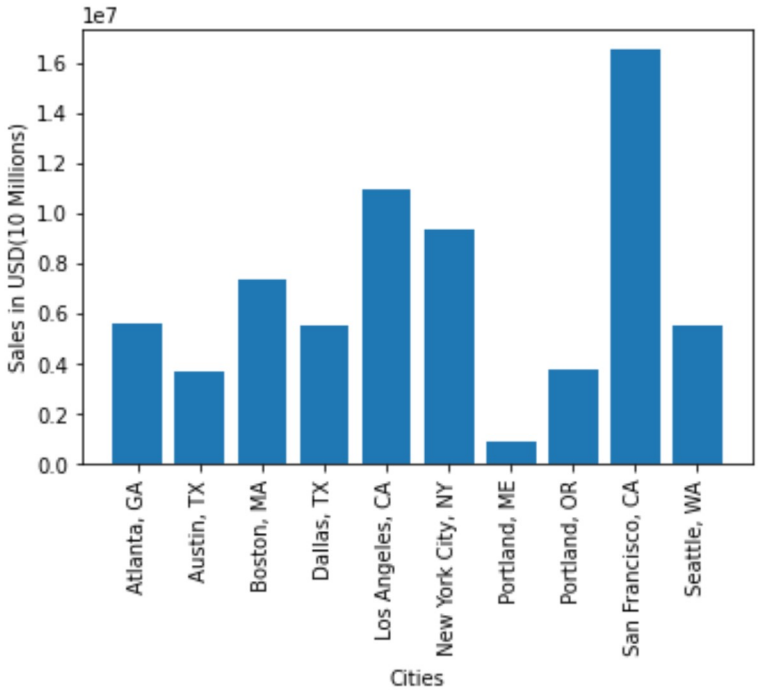
Out[76]:

	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 08:46	917 1st St, Dallas, TX 75001	4	23.90	Dallas, TX
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4	99.99	Boston, MA
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles, CA
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles, CA
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	4	11.99	Los Angeles, CA
6	176562	USB-C Charging Cable	1	11.95	04/29/19 13:03	381 Wilson St, San Francisco, CA 94016	4	11.95	San Francisco, CA
7	176563	Bose SoundSport Headphones	1	99.99	04/02/19 07:46	668 Center St, Seattle, WA 98101	4	99.99	Seattle, WA
8	176564	USB-C Charging Cable	1	11.95	04/12/19 10:58	790 Ridge St, Atlanta, GA 30301	4	11.95	Atlanta, GA
9	176565	Macbook Pro Laptop	1	1700.00	04/24/19 10:38	915 Willow St, San Francisco, CA 94016	4	1700.00	San Francisco, CA

In [78]:
```
city_result = all_data.groupby('City').sum()
print(city_result)
```

```
                  Quantity Ordered    Price Each     Month       Sales
City
 Atlanta, GA                 33204  5.559816e+06  209588.0  5.59099
7e+06
 Austin, TX                  22306  3.619747e+06  139658.0  3.63916
4e+06
 Boston, MA                  45056  7.274820e+06  282224.0  7.32328
4e+06
 Dallas, TX                  33460  5.505256e+06  209240.0  5.53595
1e+06
 Los Angeles, CA             66578  1.084287e+07  416650.0  1.09051
4e+07
 New York City, NY           55864  9.270742e+06  351482.0  9.32863
5e+06
 Portland, ME                 5500  8.943785e+05   34288.0  8.99516
5e+05
 Portland, OR                22606  3.721116e+06  141242.0  3.74146
5e+06
 San Francisco, CA          100478  1.642292e+07  631040.0  1.65244
1e+07
 Seattle, WA                 33106  5.466592e+06  209882.0  5.49551
1e+06
```

In [91]:
```
x = [city for city, df in all_data.groupby('City')]
y = city_result['Sales']
plt.bar(x,y)
plt.xticks(rotation=90)
plt.xlabel("Cities")
plt.ylabel("Sales in USD(10 Millions)")
plt.show()
```



In [93]:
```
all_data['Order Date'] = pd.to_datetime(all_data['Order Date'])
all_data['Hour'] = all_data['Order Date'].dt.hour
all_data['Minute'] = all_data['Order Date'].dt.minute
all_data.head()
```

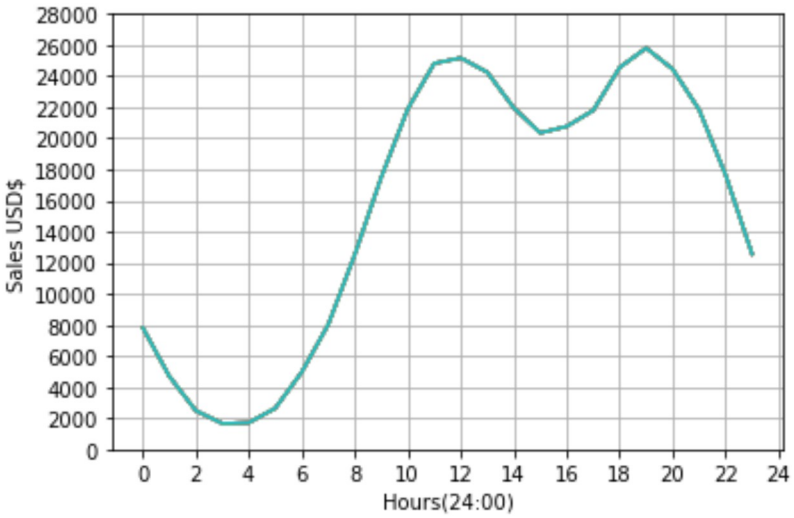
Out[93]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	Hou
0	176558	USB-C Charging Cable	2	11.95	2019-04-19 08:46:00	917 1st St, Dallas, TX 75001	4	23.90	Dallas, TX	8
2	176559	Bose SoundSport Headphones	1	99.99	2019-04-07 22:30:00	682 Chestnut St, Boston, MA 02215	4	99.99	Boston, MA	22
3	176560	Google Phone	1	600.00	2019-04-12 14:38:00	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles, CA	14
4	176560	Wired Headphones	1	11.99	2019-04-12 14:38:00	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles, CA	14
5	176561	Wired Headphones	1	11.99	2019-04-30 09:27:00	333 8th St, Los Angeles, CA 90001	4	11.99	Los Angeles, CA	9

In [95]:
```
hours = [hour for hour, df in all_data.groupby('Hour')]
print(hours)
```

```
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 1
9, 20, 21, 22, 23]
```

In [113]:
```
yy = all_data.groupby(['Hour']).count()
plt.plot(hours,yy)
plt.xticks(np.arange(0,25,2))
plt.yticks(np.arange(0,30000, 2000))
plt.xlabel("Hours(24:00)")
plt.ylabel("Sales USD$")
plt.grid()
plt.show()
```



In [145]:
```
all_data = all_data.drop_duplicates()
all_data.duplicated().sum()
df = all_data[all_data["Order ID"].duplicated(keep=False)]
df['Grouped'] = df.groupby('Order ID')['Product'].transform(lambda x:
','.join(x))
df = df[['Order ID', 'Grouped']].drop_duplicates()
df.head()
```

```
<ipython-input-145-335561c16f57>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/panda
s-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Grouped'] = df.groupby('Order ID')['Product'].transform(lambda
x: ','.join(x))
```

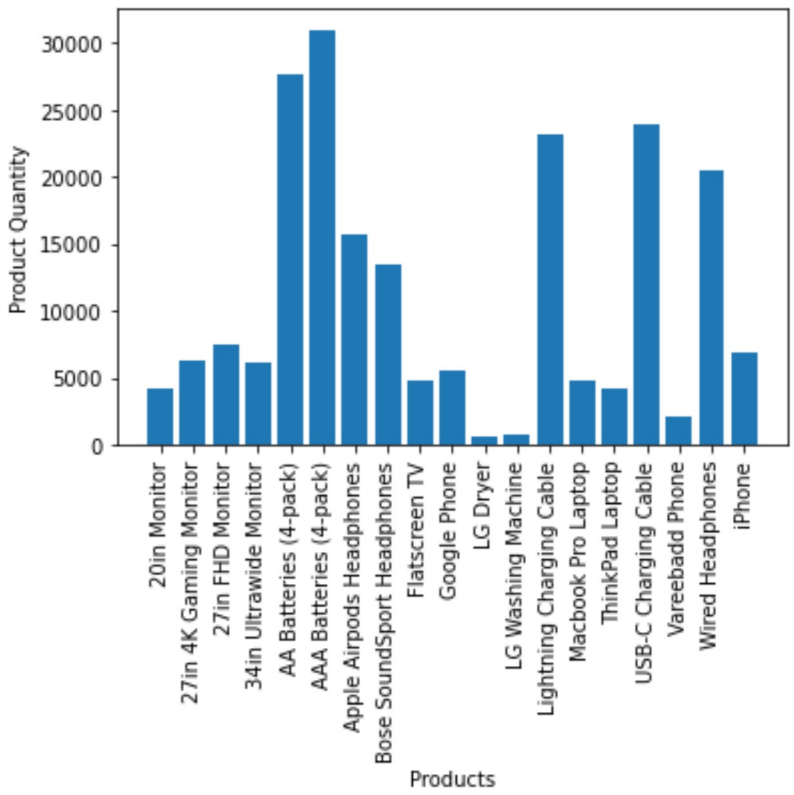
Out[145]:

	Order ID	Grouped
3	176560	Google Phone,Wired Headphones
18	176574	Google Phone,USB-C Charging Cable
32	176586	AAA Batteries (4-pack),Google Phone
119	176672	Lightning Charging Cable,USB-C Charging Cable
129	176681	Apple Airpods Headphones,ThinkPad Laptop

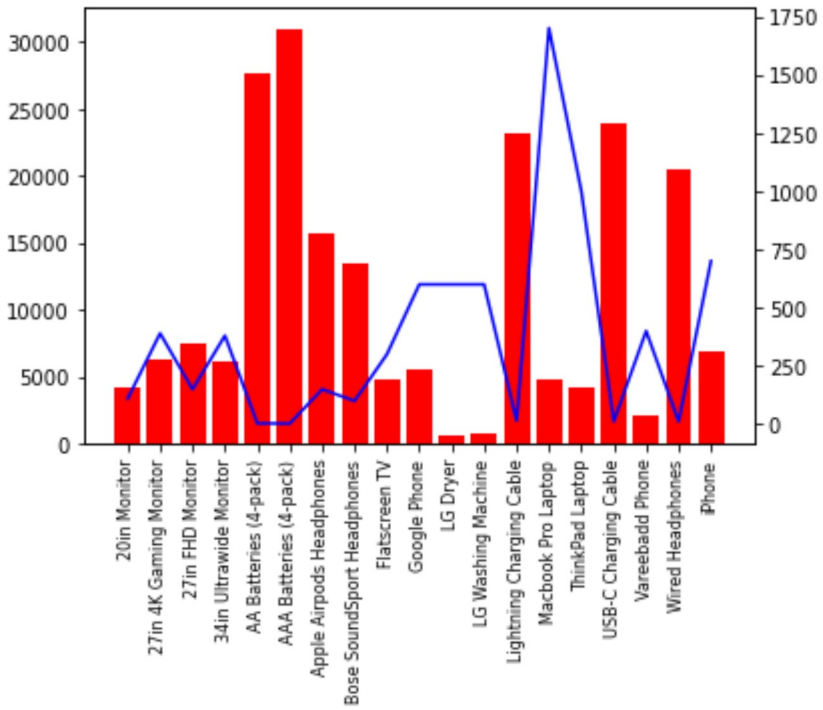
In [148]:
```
df['Grouped'].value_counts()
```

Out[148]:
```
iPhone,Lightning Charging Cable                                  886
Google Phone,USB-C Charging Cable                                857
iPhone,Wired Headphones                                          361
Vareebadd Phone,USB-C Charging Cable                             312
Google Phone,Wired Headphones                                    303
                                                                ...
20in Monitor,iPhone                                                1
27in FHD Monitor,Vareebadd Phone                                   1
Vareebadd Phone,27in FHD Monitor                                   1
Vareebadd Phone,Bose SoundSport Headphones,Flatscreen TV           1
20in Monitor,34in Ultrawide Monitor                                1
Name: Grouped, Length: 350, dtype: int64
```

In [152]:
```
product_group = all_data.groupby('Product')
product_qty = product_group['Quantity Ordered'].sum()
products = [product for product, df in product_group]
plt.bar(products,product_qty)
plt.xticks(rotation=90)
plt.xlabel("Products")
plt.ylabel("Product Quantity")
plt.show()
```



In [225]:
```
prices = all_data.groupby('Product').mean()['Price Each']
index = np.arange(len(products))
fig, ax1 = plt.subplots()
ax2 = ax1.twinx()
ax1.bar(products,product_qty,color='r')
ax2.plot(products, prices, 'b-')
ax1.set_xticklabels(products,rotation='vertical', size=8)
plt.show()
```



In [ ]: