In [6]:

```
import numpy as np
import os
import matplotlib as mpl
import matplotlib.pyplot as plt
import pandas as pd
```

In [109…

```
#merging all the csv files into 1 file
```

In [227…

```
path = "/Users/muditkant/Desktop/Machine Learning/problem solving/Pandas-Data-Science-Task
files = [file for file in os.listdir(path) if not file.startswith('.')] # Ignore hidden fi

all_months_data = pd.DataFrame()

for file in files:
    current_data = pd.read_csv(path+"/"+file)
    all_months_data = pd.concat([all_months_data, current_data])

all_months_data.to_csv("all_data1.csv", index=False)
```

In [228…

```
df = pd.read_csv("/Users/muditkant/Downloads/Pandas-Data-Science-Tasks-master/SalesAnalysi
df.head()
```

Out[228…

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	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 [229…

```
#checking for null values
```

In [230…

```
df.isnull().values.any()
```

Out[230…

```
True
```

In [231…

```
df = df.dropna()
```

In [232…

```
#checking for null values
```

In [233…

```
df.isnull().values.any()
```

Out[233…

```
False
```

In [234…

```
temp = df[df["Order Date"].str[:2] == "Or"]
temp.head()
```

Out[234…

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
519	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
1149	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
1155	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
2878	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
2893	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address

In [235…

```
#order date consits of gibberish data
#removing gibberish values
```

In [236…

```
df = df[df["Order Date"].str[:2] != "Or"]
df.head()
```

Out[236…

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

In [237…

```
df["Month"] = df["Order Date"].str[:2]
df.head()
```

Out[237…

	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
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
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	04

In [238…

```
#converting month values into int
```

In [239…

```
df["Month"] = df["Month"].astype('int32')
df.head()
```

Out[239…

	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 [240…

```
## Q1: Sales associated with each order
```

In [241…

```
#converting Quantity Ordered and Price Each into numeric int
```

In [242…

```
df["Quantity Ordered"] = pd.to_numeric(df["Quantity Ordered"])
df["Price Each"] = pd.to_numeric(df["Price Each"])
```

In [243…

```
df["Sales"] = df["Quantity Ordered"] * df["Price Each"]
df.head()
```

Out[243…

	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 [246…

```
#Best month for sales and revenue generated
```

In [247…

```
graph = df.groupby("Month").sum()
graph
```

Out[247…

	Quantity Ordered	Price Each	Sales
Month			
1	10903	1811768.38	1822256.73

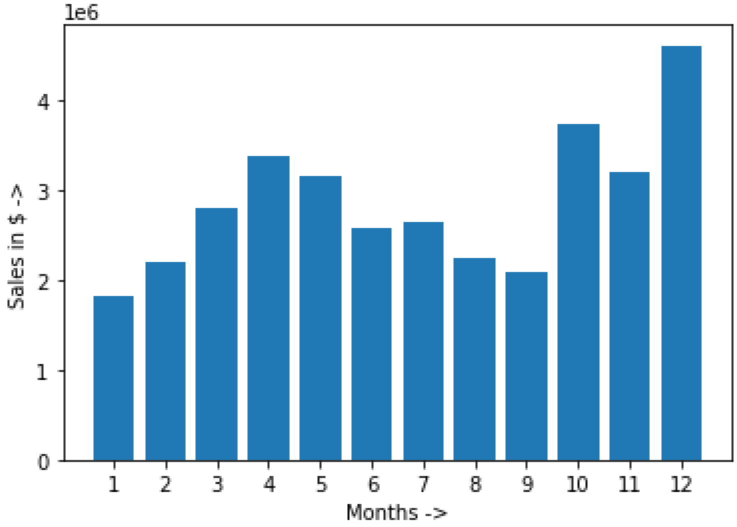
	Quantity Ordered	Price Each	Sales
Month			
2	13449	2188884.72	2202022.42
3	17005	2791207.83	2807100.38
4	20558	3367671.02	3390670.24
5	18667	3135125.13	3152606.75
6	15253	2562025.61	2577802.26
7	16072	2632539.56	2647775.76
8	13448	2230345.42	2244467.88
9	13109	2084992.09	2097560.13
10	22703	3715554.83	3736726.88
11	19798	3180600.68	3199603.20
12	28114	4588415.41	4613443.34

In [248…

```
#visualizing using matplotlib
```

In [267…

```
months = range(1,13)
plt.bar(months,graph["Sales"])
plt.xticks(months)
plt.ylabel('Sales in $ ->')
plt.xlabel("Months ->")
plt.show()
```



In [268…

```
#Which city has best sales
```

In [280…

```
df.head()
```

Out[280…

	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

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City
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 09:27	333 8th St, Los Angeles, CA 90001	4	11.99	Los Angeles

In [281…

```
#Making seperate column for city and extracting value
```

In [295…

```
def state(address):
    return address.split(",")[2].split(" ")[1]

df["City"] = df["Purchase Address"].apply(lambda x: x.split(',')[1] + " " + state(x))
```

In [318…

```
df.head()
```

Out[318…

	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

In [319…

```
#Q2: Best sales in which city
```

In [320…

```
result = df.groupby("City").sum()
result
```

Out[320…

	Quantity Ordered	Price Each	Month	Sales
City				
Atlanta GA	16602	2779908.20	104794	2795498.58
Austin TX	11153	1809873.61	69829	1819581.75
Boston MA	22528	3637409.77	141112	3661642.01
Dallas TX	16730	2752627.82	104620	2767975.40

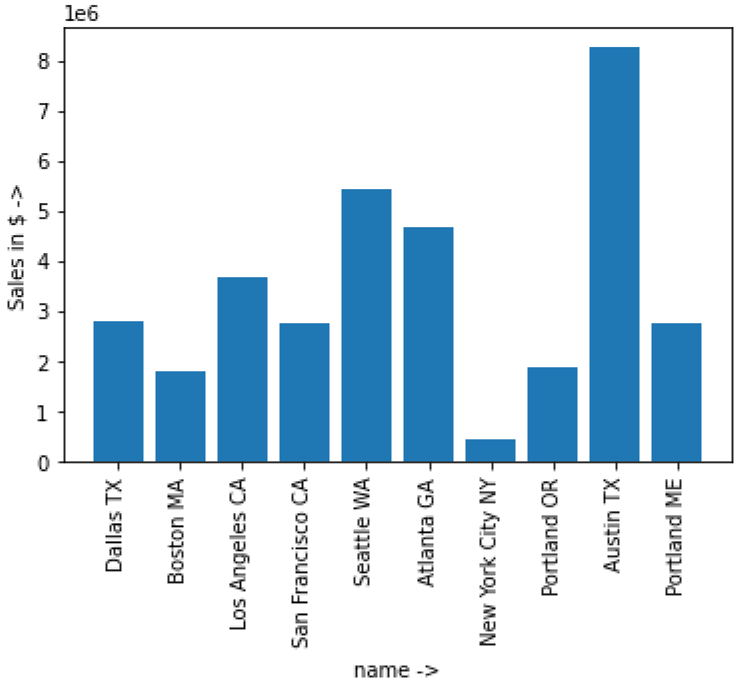
	Quantity Ordered	Price Each	Month	Sales
City				
Los Angeles CA	33289	5421435.23	208325	5452570.80
New York City NY	27932	4635370.83	175741	4664317.43
Portland ME	2750	447189.25	17144	449758.27
Portland OR	11303	1860558.22	70621	1870732.34
San Francisco CA	50239	8211461.74	315520	8262203.91
Seattle WA	16553	2733296.01	104941	2747755.48

In [312…

```
#visualizing using matplotlib
```

In [333…

```
cities = df['City'].unique()
plt.bar(cities,result["Sales"])
plt.xticks(cities,rotation = "vertical")
plt.ylabel('Sales in $ ->')
plt.xlabel("name ->")
plt.show()
```



In [334…

```
# In city colums it shows San Francisco CA bes sales
# While in visulalization it shows, austin TX
```

In [335…

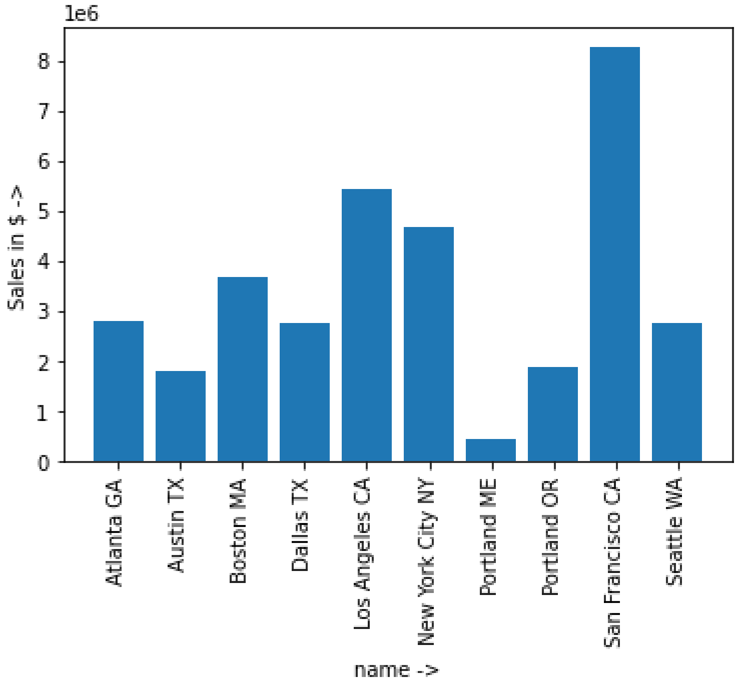
```
## Need to search why this happened.
```

In [336…

```
## X data and Y data needs to be in same order. That's why it's causing
```

In [344…

```
cities = [city for city, df in df.groupby(['City'])]
plt.bar(cities,result["Sales"])
plt.xticks(cities,rotation = "vertical")
plt.ylabel('Sales in $ ->')
plt.xlabel("name ->")
plt.show()
```



In [345…

```
#Q3: Best Time for advertisements.
```

In [346…

```
df.head()
```

Out[346…

	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

In [361…

```
df["Order Date"] = pd.to_datetime(df["Order Date"])
```

In [366…

```
df["Hour"] = df["Order Date"].dt.hour
df["Minutes"] = df["Order Date"].dt.minute
df.head()
```

Out[366…

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	Hour	Minutes

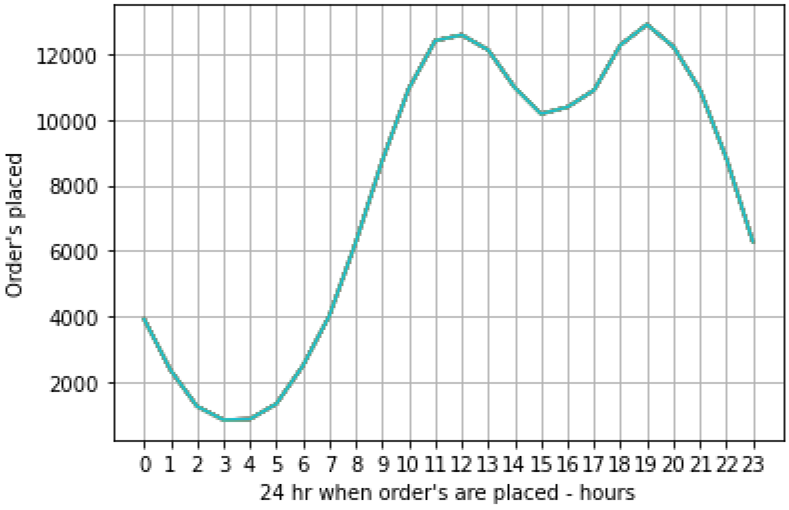
	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	Hour	Minutes
0	176558	USB-C Charging Cable	2	11.95	2019-04-19 08:46:00	917 1st St, Dallas, TX 75001	4	23.90	Dallas TX	8	46
2	176559	Bose SoundSport Headphones	1	99.99	2019-04-07 22:30:00	682 Chestnut St, Boston, MA 02215	4	99.99	Boston MA	22	30
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	38
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	38
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	27

In [364…

```
#visualizing using matplotlib
```

In [374…

```
Hour = [Hour for Hour, df in df.groupby(['Hour'])]
plt.plot(Hour, df.groupby(['Hour']).count())
plt.xticks(Hour)
plt.xlabel("24 hr when order's are placed - hours")
plt.ylabel("Order's placed")
plt.grid()
plt.show()
```



In [375…

```
#delivering advertisements just before Peak hours of sale: 11 - 12 & 18 - 19 hrs
```

In [376…

```
#Q4: What are the most often products sold in a groupof 2 or 3?
```

In [383…

```
df.head(20)
```

Out[383…

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	Hour	Minutes
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	46
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	30
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	38
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	38
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	27
6	176562	USB-C Charging Cable	1	11.95	2019-04-29 13:03:00	381 Wilson St, San Francisco, CA 94016	4	11.95	San Francisco CA	13	3
7	176563	Bose SoundSport Headphones	1	99.99	2019-04-02 07:46:00	668 Center St, Seattle, WA 98101	4	99.99	Seattle WA	7	46
8	176564	USB-C Charging Cable	1	11.95	2019-04-12 10:58:00	790 Ridge St, Atlanta, GA 30301	4	11.95	Atlanta GA	10	58
9	176565	Macbook Pro Laptop	1	1700.00	2019-04-24 10:38:00	915 Willow St, San Francisco, CA 94016	4	1700.00	San Francisco CA	10	38
10	176566	Wired Headphones	1	11.99	2019-04-08 14:05:00	83 7th St, Boston, MA 02215	4	11.99	Boston MA	14	5
11	176567	Google Phone	1	600.00	2019-04-18 17:18:00	444 7th St, Los Angeles, CA 90001	4	600.00	Los Angeles CA	17	18
12	176568	Lightning Charging Cable	1	14.95	2019-04-15 12:18:00	438 Elm St, Seattle, WA 98101	4	14.95	Seattle WA	12	18
13	176569	27in 4K Gaming Monitor	1	389.99	2019-04-16 19:23:00	657 Hill St, Dallas, TX 75001	4	389.99	Dallas TX	19	23

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	Hour	Minutes
14	176570	AA Batteries (4-pack)	1	3.84	2019-04-22 15:09:00	186 12th St, Dallas, TX 75001	4	3.84	Dallas TX	15	9
15	176571	Lightning Charging Cable	1	14.95	2019-04-19 14:29:00	253 Johnson St, Atlanta, GA 30301	4	14.95	Atlanta GA	14	29
16	176572	Apple Airpods Headphones	1	150.00	2019-04-04 20:30:00	149 Dogwood St, New York City, NY 10001	4	150.00	New York City NY	20	30
17	176573	USB-C Charging Cable	1	11.95	2019-04-27 18:41:00	214 Chestnut St, San Francisco, CA 94016	4	11.95	San Francisco CA	18	41
18	176574	Google Phone	1	600.00	2019-04-03 19:42:00	20 Hill St, Los Angeles, CA 90001	4	600.00	Los Angeles CA	19	42
19	176574	USB-C Charging Cable	1	11.95	2019-04-03 19:42:00	20 Hill St, Los Angeles, CA 90001	4	11.95	Los Angeles CA	19	42
20	176575	AAA Batteries (4-pack)	1	2.99	2019-04-27 00:30:00	433 Hill St, New York City, NY 10001	4	2.99	New York City NY	0	30

In [384…

```
#Order's have duplicate order ID:
```

In [387…

```
df = df[df['Order ID'].duplicated(keep=False)]
df['Grouped'] = df.groupby('Order ID')['Product'].transform(lambda x: ','.join(x))
df2 = df[['Order ID', 'Grouped']].drop_duplicates()
```

In [388…

```
from itertools import combinations
from collections import Counter
```

In [393…

```
count = Counter()

for row in df2['Grouped']:
    row_list = row.split(',')
    count.update(Counter(combinations(row_list, 1)))

for key,value in count.most_common(10):
    print(key, value)
```

```
('USB-C Charging Cable',) 2111
('iPhone',) 1867
('Lightning Charging Cable',) 1827
('Wired Headphones',) 1674
('Google Phone',) 1639
```

```
('Apple Airpods Headphones',) 974
('Bose SoundSport Headphones',) 820
('AAA Batteries (4-pack)',) 815
('AA Batteries (4-pack)',) 768
('Vareebadd Phone',) 601
```

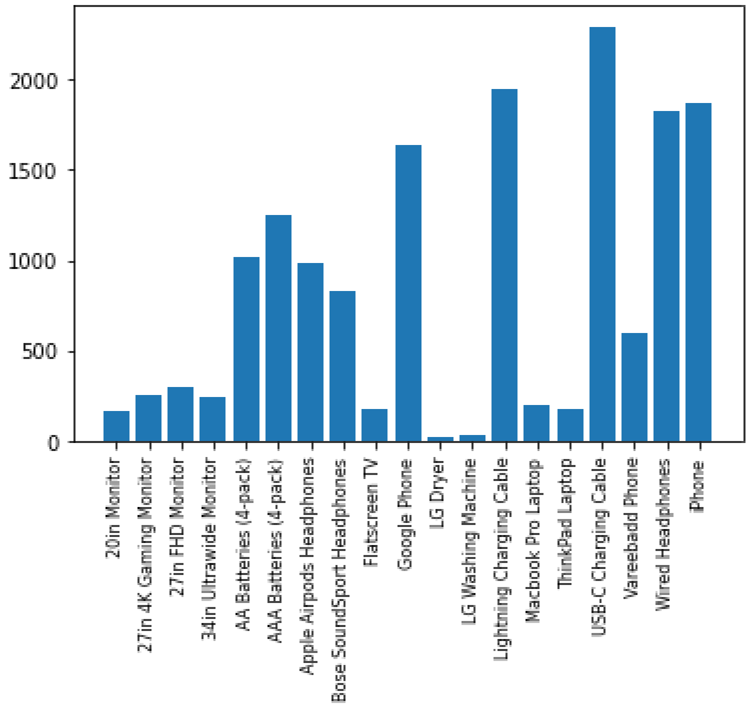
In [392…

```
# Q5: What product was sold the most and why
```

In [394…

```
product_group = df.groupby('Product')
quantity_ordered = product_group.sum()['Quantity Ordered']

keys = [pair for pair, df in product_group]
plt.bar(keys, quantity_ordered)
plt.xticks(keys, rotation='vertical', size=8)
plt.show()
```



In [399…

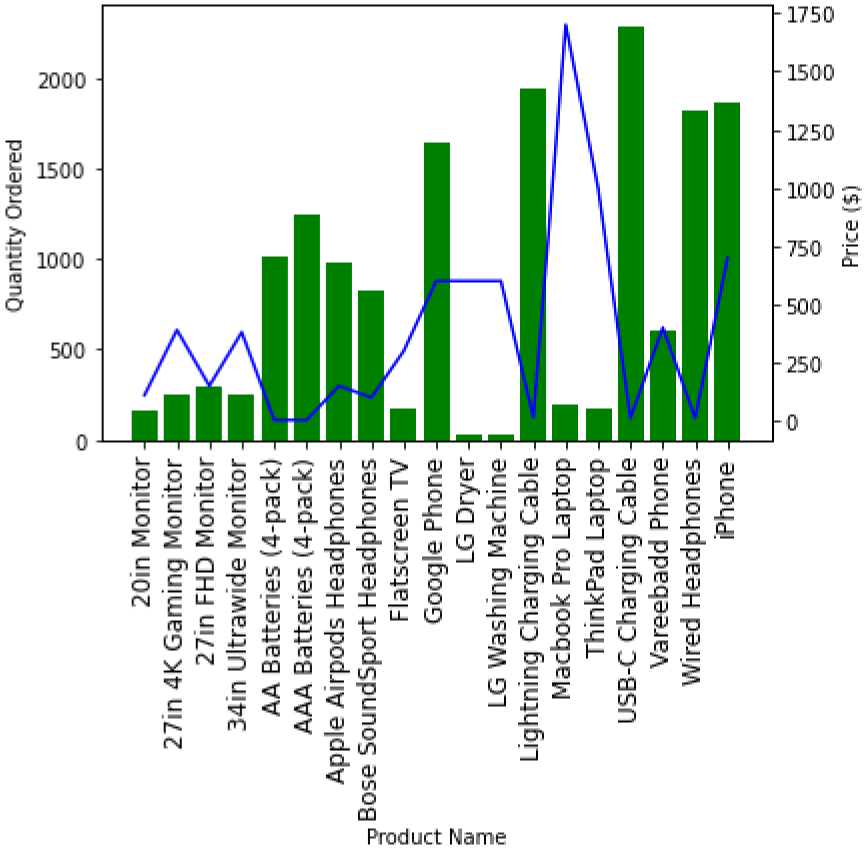
```
prices = df.groupby('Product').mean()['Price Each']
fig, ax1 = plt.subplots()

ax2 = ax1.twinx()
ax1.bar(keys, quantity_ordered, color='g')
ax2.plot(keys, prices, color='b')

ax1.set_xlabel('Product Name')
ax1.set_ylabel('Quantity Ordered')
ax2.set_ylabel('Price ($)')
ax1.set_xticklabels(keys, rotation='vertical',)

fig.show()
```

```
/var/folders/xy/4d96r82x4yx1r3rlsdld1cqm0000gn/T/ipykernel_2758/1460133011.py:11: UserWarn
ing: FixedFormatter should only be used together with FixedLocator
  ax1.set_xticklabels(keys, rotation='vertical', size=12)
/var/folders/xy/4d96r82x4yx1r3rlsdld1cqm0000gn/T/ipykernel_2758/1460133011.py:13: UserWarn
ing: Matplotlib is currently using module://matplotlib_inline.backend_inline, which is a n
on-GUI backend, so cannot show the figure.
  fig.show()
```


Quantity Ordered
Price ($)
Product Name
0
500
1000
1500
2000
0
250
500
750
1000
1250
1500
1750
20in Monitor
27in 4K Gaming Monitor
27in FHD Monitor
34in Ultrawide Monitor
AA Batteries (4-pack)
AAA Batteries (4-pack)
Apple Airpods Headphones
Bose SoundSport Headphones
Flatscreen TV
Google Phone
LG Dryer
LG Washing Machine
Lightning Charging Cable
Macbook Pro Laptop
ThinkPad Laptop
USB-C Charging Cable
Vareebadd Phone
Wired Headphones
iPhone

In [ ]: