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

In [2]:
```
!pip install pyarrow
```

```
Requirement already satisfied: pyarrow in ./anaconda3/lib/python3.11/site-pack
ages (11.0.0)
Requirement already satisfied: numpy>=1.16.6 in ./anaconda3/lib/python3.11/sit
e-packages (from pyarrow) (1.24.3)
```

In [3]:
```
all_data = pd.read_feather(r'/Users/riyalachuriya/Desktop/Python Project/Sales_
```

In [4]:
```
all_data.head(6)
```

Out[4]:

	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	None	None	None	None	None	None
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 [5]:
```
all_data.shape
```

Out[5]:
```
(186850, 6)
```

In [6]:
```
all_data.isnull()
```

Out[6]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	False	False	False	False	False	False
1	True	True	True	True	True	True
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	False	False	False	False
...	...	...	...	...	...	...
186845	False	False	False	False	False	False
186846	False	False	False	False	False	False
186847	False	False	False	False	False	False
186848	False	False	False	False	False	False
186849	False	False	False	False	False	False

186850 rows × 6 columns

In [7]:

```
all_data.isnull().sum()
```

Out[7]:

```
Order ID            545
Product             545
Quantity Ordered    545
Price Each          545
Order Date          545
Purchase Address    545
dtype: int64
```

In [8]:

```
all_data.dropna(how="all")
```

Out[8]:

	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
...	...	...	...	...	...	...
186845	259353	AAA Batteries (4-pack)	3	2.99	09/17/19 20:56	840 Highland St, Los Angeles, CA 90001
186846	259354	iPhone	1	700	09/01/19 16:00	216 Dogwood St, San Francisco, CA 94016
186847	259355	iPhone	1	700	09/23/19 07:39	220 12th St, San Francisco, CA 94016
186848	259356	34in Ultrawide Monitor	1	379.99	09/19/19 17:30	511 Forest St, San Francisco, CA 94016
186849	259357	USB-C Charging Cable	1	11.95	09/30/19 00:18	250 Meadow St, San Francisco, CA 94016

186305 rows × 6 columns

In [9]:

```
all_data = all_data.dropna(how="all")
```

In [10]:

```
all_data.shape
```

Out[10]:

```
(186305, 6)
```

In [11]:

```
all_data.isnull().sum()
```

Out[11]:

```
Order ID            0
Product             0
Quantity Ordered    0
Price Each          0
Order Date          0
Purchase Address    0
dtype: int64
```

In [12]:

```
all_data.duplicated()
```

Out[12]:

```
0         False
2         False
3         False
4         False
5         False
          ...  
186845    False
186846    False
186847    False
186848    False
186849    False
Length: 186305, dtype: bool
```

In [13]:

```
all_data.duplicated().sum()
```

Out[13]:

```
618
```

In [14]:

```
all_data[all_data.duplicated()]
```

Out[14]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
31	176585	Bose SoundSport Headphones	1	99.99	04/07/19 11:31	823 Highland St, Boston, MA 02215
1149	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
1155	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
1302	177795	Apple Airpods Headphones	1	150	04/27/19 19:45	740 14th St, Seattle, WA 98101
1684	178158	USB-C Charging Cable	1	11.95	04/28/19 21:13	197 Center St, San Francisco, CA 94016
...	...	...	...	...	...	...
186563	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
186632	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
186738	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
186782	259296	Apple Airpods Headphones	1	150	09/28/19 16:48	894 6th St, Dallas, TX 75001
186785	259297	Lightning Charging Cable	1	14.95	09/15/19 18:54	138 Main St, Boston, MA 02215

618 rows × 6 columns

In [15]:

```
all_data.drop_duplicates()
```

Out[15]:

	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
...	...	...	...	...	...	...
186845	259353	AAA Batteries (4-pack)	3	2.99	09/17/19 20:56	840 Highland St, Los Angeles, CA 90001
186846	259354	iPhone	1	700	09/01/19 16:00	216 Dogwood St, San Francisco, CA 94016
186847	259355	iPhone	1	700	09/23/19 07:39	220 12th St, San Francisco, CA 94016
186848	259356	34in Ultrawide Monitor	1	379.99	09/19/19 17:30	511 Forest St, San Francisco, CA 94016
186849	259357	USB-C Charging Cable	1	11.95	09/30/19 00:18	250 Meadow St, San Francisco, CA 94016

185687 rows × 6 columns

In [16]:
```
all_data = all_data.drop_duplicates()
```

In [17]:
```
all_data.shape
```

Out[17]:
```
(185687, 6)
```

In [18]:
```
#which is the best month for sale
```

In [19]:
```
all_data ['Order Date']
```

Out[19]:
```
0         04/19/19 08:46
2         04/07/19 22:30
3         04/12/19 14:38
4         04/12/19 14:38
5         04/30/19 09:27
               ...
186845    09/17/19 20:56
186846    09/01/19 16:00
186847    09/23/19 07:39
186848    09/19/19 17:30
186849    09/30/19 00:18
Name: Order Date, Length: 185687, dtype: object
```

In [20]:
```
all_data ['Order Date'][0].split('/')[0]
```

Out[20]:

```
'04'
```

In [21]:

```python
def return_month(x):
    return x.split('/')[0]
```

In [22]:

```python
all_data ['Month'] = all_data ['Order Date'].apply(return_month)
```

In [23]:

```python
all_data.dtypes
```

Out[23]:

```
Order ID            object
Product             object
Quantity Ordered    object
Price Each          object
Order Date          object
Purchase Address    object
Month               object
dtype: object
```

In [24]:

```python
all_data ['Month'].astype(int)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[24], line 1
----> 1 all_data ['Month'].astype(int)

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/generic.py:6324, in 
NDFrame.astype(self, dtype, copy, errors)
   6317     results = [
   6318         self.iloc[:, i].astype(dtype, copy=copy)
   6319         for i in range(len(self.columns))
   6320     ]
   6322 else:
   6323     # else, only a single dtype is given
-> 6324     new_data = self._mgr.astype(dtype=dtype, copy=copy, errors=errors)
   6325     return self._constructor(new_data).__finalize__(self, method="asty
pe")
   6327 # GH 33113: handle empty frame or series

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/internals/managers.p
y:451, in BaseBlockManager.astype(self, dtype, copy, errors)
    448 elif using_copy_on_write():
    449     copy = False
--> 451 return self.apply(
    452     "astype",
    453     dtype=dtype,
    454     copy=copy,
    455     errors=errors,
    456     using_cow=using_copy_on_write(),
    457 )

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/internals/managers.p
y:352, in BaseBlockManager.apply(self, f, align_keys, **kwargs)
    350         applied = b.apply(f, **kwargs)
    351     else:
--> 352         applied = getattr(b, f)(**kwargs)
    353     result_blocks = extend_blocks(applied, result_blocks)
    355 out = type(self).from_blocks(result_blocks, self.axes)

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/internals/blocks.py:
511, in Block.astype(self, dtype, copy, errors, using_cow)
    491 """
    492 Coerce to the new dtype.
    493 
   (...)
    507 Block
    508 """
    509 values = self.values
--> 511 new_values = astype_array_safe(values, dtype, copy=copy, errors=error
s)
    513 new_values = maybe_coerce_values(new_values)
    515 refs = None

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/dtypes/astype.py:24
2, in astype_array_safe(values, dtype, copy, errors)
    239     dtype = dtype.numpy_dtype
    241 try:
--> 242     new_values = astype_array(values, dtype, copy=copy)
    243 except (ValueError, TypeError):
    244     # e.g. _astype_nansafe can fail on object-dtype of strings
    245     #  trying to convert to float
    246     if errors == "ignore":

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/dtypes/astype.py:18
7, in astype_array(values, dtype, copy)
    184     values = values.astype(dtype, copy=copy)
    186 else:
--> 187     values = _astype_nansafe(values, dtype, copy=copy)
    189 # in pandas we don't store numpy str dtypes, so convert to object
    190 if isinstance(dtype, np.dtype) and issubclass(values.dtype.type, str):

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/dtypes/astype.py:13
8, in _astype_nansafe(arr, dtype, copy, skipna)
    134     raise ValueError(msg)
    136 if copy or is_object_dtype(arr.dtype) or is_object_dtype(dtype):
    137     # Explicit copy, or required since NumPy can't view from / to obje
ct.
--> 138     return arr.astype(dtype, copy=True)
    140 return arr.astype(dtype, copy=copy)

ValueError: invalid literal for int() with base 10: 'Order Date'
```

In [25]:
```
all_data['Month'].unique()
```

Out[25]:
```
array(['04', '05', 'Order Date', '08', '09', '12', '01', '02', '03', '07',
       '06', '11', '10'], dtype=object)
```

In [26]:
```
all_data['Month']== 'Order Date'
```

Out[26]:
```
0         False
2         False
3         False
4         False
5         False
          ...
186845    False
186846    False
186847    False
186848    False
186849    False
Name: Month, Length: 185687, dtype: bool
```

In [27]:
```
filter1 = all_data['Month']== 'Order Date'
```

In [28]:
```
all_data[~filter1]
```

Out[28]:

	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
...	...	...	...	...	...	...	...
186845	259353	AAA Batteries (4-pack)	3	2.99	09/17/19 20:56	840 Highland St, Los Angeles, CA 90001	09
186846	259354	iPhone	1	700	09/01/19 16:00	216 Dogwood St, San Francisco, CA 94016	09
186847	259355	iPhone	1	700	09/23/19 07:39	220 12th St, San Francisco, CA 94016	09
186848	259356	34in Ultrawide Monitor	1	379.99	09/19/19 17:30	511 Forest St, San Francisco, CA 94016	09
186849	259357	USB-C Charging Cable	1	11.95	09/30/19 00:18	250 Meadow St, San Francisco, CA 94016	09

185686 rows × 7 columns

In [29]:
```
all_data = all_data[~filter1]
```

In [30]:
```
import warnings
from warnings import filterwarnings
filterwarnings('ignore')
```

In [31]:
```
all_data ['Month'] = all_data ['Month'].astype(int)
```

In [32]:
```
all_data ['Quantity Ordered'] = all_data ['Quantity Ordered'].astype(int)
all_data ['Price Each'] = all_data ['Price Each'].astype(float)
```

In [33]:
```
all_data.dtypes
```

```
Out[33]: Order ID             object
         Product              object
         Quantity Ordered      int64
         Price Each          float64
         Order Date           object
         Purchase Address     object
         Month                 int64
         dtype: object
```

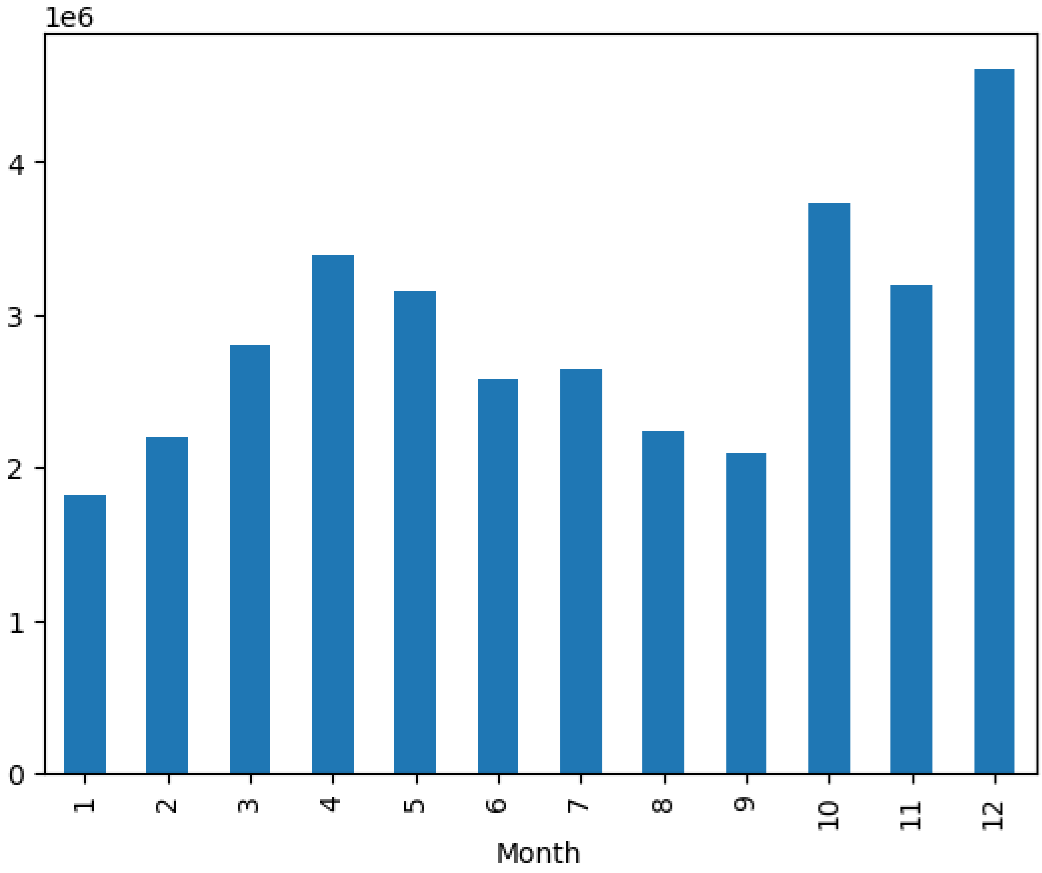
```
In [34]: all_data ['Sale'] = all_data ['Quantity Ordered'] * all_data ['Price Each']
```

```
In [35]:  all_data.groupby(['Month']) ['Sale'].sum()
```

```
Out[35]: Month
         1     1821413.16
         2     2200078.08
         3     2804973.35
         4     3389217.98
         5     3150616.23
         6     2576280.15
         7     2646461.32
         8     2241083.37
         9     2094465.69
         10    3734777.86
         11    3197875.05
         12    4608295.70
         Name: Sale, dtype: float64
```

```
In [36]:  all_data.groupby(['Month']) ['Sale'].sum().plot(kind= 'bar')
```

```
Out[36]: <Axes: xlabel='Month'>
```



In [37]:
```
#last month of year is highesh might be because of new year and christmas
```

In [38]:
```
all_data['Purchase Address'][0].split(',')[1]
```

Out[38]:
```
' Dallas'
```

def city(x):

```
return x.split(',')[1]
```

In [39]:
```
#another method
```

In [40]:
```
all_data['city'] = all_data['Purchase Address'].str.split(',').str[1]
```

all_data['city'] = all_data['Purchase Address'].apply(city)

In [41]:
```
all_data['city']
```

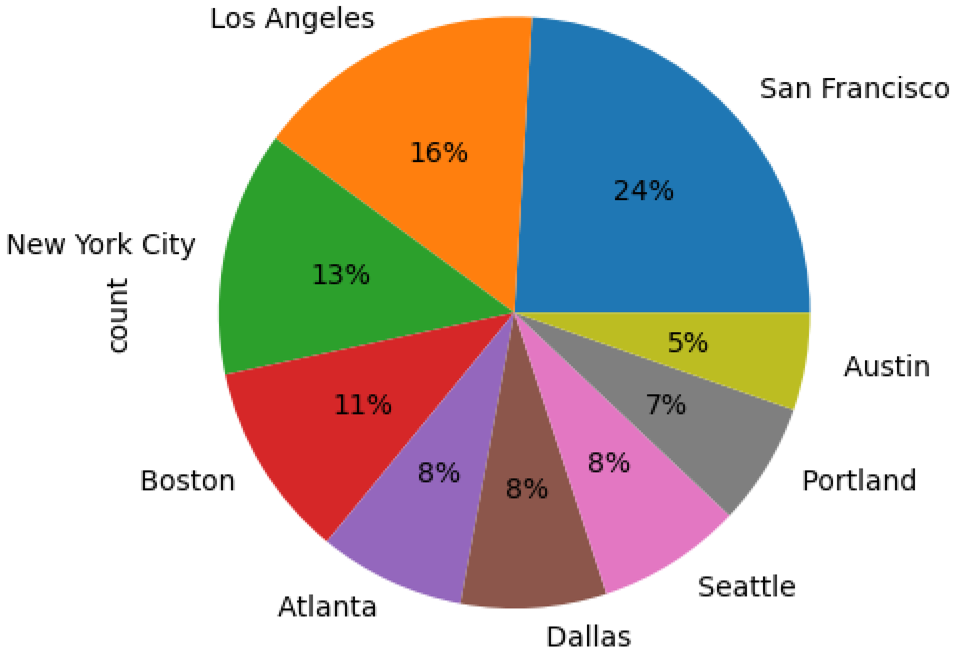
Out[41]:
```
0                  Dallas
2                  Boston
3             Los Angeles
4             Los Angeles
5             Los Angeles
                 ...
186845        Los Angeles
186846      San Francisco
186847      San Francisco
186848      San Francisco
186849      San Francisco
Name: city, Length: 185686, dtype: object
```

In [42]:
```
pd.value_counts(all_data['city'])
```

Out[42]:
```
city
 San Francisco    44662
 Los Angeles      29564
 New York City    24847
 Boston           19901
 Atlanta          14863
 Dallas           14797
 Seattle          14713
 Portland         12449
 Austin            9890
Name: count, dtype: int64
```

In [43]:
```
pd.value_counts(all_data['city']).plot(kind='pie' , autopct = '%1.0f%%')
```

Out[43]:
```
<Axes: ylabel='count'>
```



```
In [44]: #san francisco, los angeles, new york city, boston have highest purchases made
```

```
In [45]: all_data.columns
```

```
Out[45]: Index(['Order ID', 'Product', 'Quantity Ordered', 'Price Each', 'Order Date',
                'Purchase Address', 'Month', 'Sale', 'city'],
               dtype='object')
```

```
In [46]: count_df = all_data.groupby(['Product']).agg({'Quantity Ordered':'sum' , 'Pric
```

```
In [47]: product = count_df['Product'].values
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
File ~/anaconda3/lib/python3.11/site-packages/pandas/core/indexes/base.py:365
3, in Index.get_loc(self, key)
   3652 try:
-> 3653     return self._engine.get_loc(casted_key)
   3654 except KeyError as err:

File ~/anaconda3/lib/python3.11/site-packages/pandas/_libs/index.pyx:147, in p
andas._libs.index.IndexEngine.get_loc()

File ~/anaconda3/lib/python3.11/site-packages/pandas/_libs/index.pyx:176, in p
andas._libs.index.IndexEngine.get_loc()

File pandas/_libs/hashtable_class_helper.pxi:7080, in pandas._libs.hashtable.P
yObjectHashTable.get_item()

File pandas/_libs/hashtable_class_helper.pxi:7088, in pandas._libs.hashtable.P
yObjectHashTable.get_item()

KeyError: 'Product'

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
Cell In[47], line 1
----> 1 product = count_df['Product'].values

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/frame.py:3761, in Da
taFrame.__getitem__(self, key)
   3759 if self.columns.nlevels > 1:
   3760     return self._getitem_multilevel(key)
-> 3761 indexer = self.columns.get_loc(key)
   3762 if is_integer(indexer):
   3763     indexer = [indexer]

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/indexes/base.py:365
5, in Index.get_loc(self, key)
   3653     return self._engine.get_loc(casted_key)
   3654 except KeyError as err:
-> 3655     raise KeyError(key) from err
   3656 except TypeError:
   3657     # If we have a listlike key, _check_indexing_error will raise
   3658     #  InvalidIndexError. Otherwise we fall through and re-raise
   3659     #  the TypeError.
   3660     self._check_indexing_error(key)

KeyError: 'Product'
```

In [ ]:

```
count_df = count_df.reset_index()
```

In [ ]:

```
fig , ax1 = plt.subplots()

ax2 = ax1.twinx()
ax1.bar(count_df['Product'],count_df['Quantity Ordered'],color = 'pink')
ax2.plot(count_df['Product'],count_df['Price Each'],color = 'blue')
ax1.set_xticklabels(product,rotation='vertical',size ='12')
```

```
ax1.set_ylabel('odered count')
ax2.set_ylabel('avg price of product')
```

In [ ]:

```
#AAA batteries which have lowest price is sold most
```

In [48]:

```
all_data['Product'].value_counts()[0:5]
```

Out[48]:

```
Product
USB-C Charging Cable         21859
Lightning Charging Cable     21610
AAA Batteries (4-pack)       20612
AA Batteries (4-pack)        20558
Wired Headphones             18849
Name: count, dtype: int64
```

In [49]:

```
most_sold_product = all_data['Product'].value_counts()[0:5].index
```

In [50]:

```
most_sold_product
```

Out[50]:

```
Index(['USB-C Charging Cable', 'Lightning Charging Cable',
       'AAA Batteries (4-pack)', 'AA Batteries (4-pack)', 'Wired Headphones'],
      dtype='object', name='Product')
```

In [51]:

```
all_data['Product'].isin(most_sold_product)
```

Out[51]:

```
0          True
2         False
3         False
4          True
5          True
          ...  
186845     True
186846    False
186847    False
186848    False
186849     True
Name: Product, Length: 185686, dtype: bool
```

In [52]:

```
most_sold_product_df = all_data[all_data['Product'].isin(most_sold_product)]
```

In [53]:

```
most_sold_product_df
```

Out[53]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sale	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
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
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
8	176564	USB-C Charging Cable	1	11.95	04/12/19 10:58	790 Ridge St, Atlanta, GA 30301	4	11.95	Atlanta
...	...	...	...	...	...	...	...	...	...
186840	259349	AAA Batteries (4-pack)	1	2.99	09/01/19 22:14	911 River St, Dallas, TX 75001	9	2.99	Dallas
186842	259350	USB-C Charging Cable	1	11.95	09/30/19 13:49	519 Maple St, San Francisco, CA 94016	9	11.95	San Francisco
186844	259352	USB-C Charging Cable	1	11.95	09/07/19 15:49	976 Forest St, San Francisco, CA 94016	9	11.95	San Francisco
186845	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
186849	259357	USB-C Charging Cable	1	11.95	09/30/19 00:18	250 Meadow St, San Francisco, CA 94016	9	11.95	San Francisco

103488 rows × 9 columns

In [54]:

```
most_sold_product_df.groupby(['Month' , 'Product']).size()
```

Out[54]:

```
Month  Product
1      AA Batteries (4-pack)        1037
       AAA Batteries (4-pack)       1084
       Lightning Charging Cable     1069
       USB-C Charging Cable         1171
       Wired Headphones             1004
2      AA Batteries (4-pack)        1274
       AAA Batteries (4-pack)       1320
       Lightning Charging Cable     1393
       USB-C Charging Cable         1511
       Wired Headphones             1179
3      AA Batteries (4-pack)        1672
       AAA Batteries (4-pack)       1645
       Lightning Charging Cable     1749
       USB-C Charging Cable         1766
       Wired Headphones             1512
4      AA Batteries (4-pack)        2062
       AAA Batteries (4-pack)       1988
       Lightning Charging Cable     2197
       USB-C Charging Cable         2074
       Wired Headphones             1888
5      AA Batteries (4-pack)        1821
       AAA Batteries (4-pack)       1888
       Lightning Charging Cable     1929
       USB-C Charging Cable         1879
       Wired Headphones             1729
6      AA Batteries (4-pack)        1540
       AAA Batteries (4-pack)       1451
       Lightning Charging Cable     1560
       USB-C Charging Cable         1531
       Wired Headphones             1334
7      AA Batteries (4-pack)        1555
       AAA Batteries (4-pack)       1554
       Lightning Charging Cable     1690
       USB-C Charging Cable         1667
       Wired Headphones             1434
8      AA Batteries (4-pack)        1357
       AAA Batteries (4-pack)       1340
       Lightning Charging Cable     1354
       USB-C Charging Cable         1339
       Wired Headphones             1191
9      AA Batteries (4-pack)        1314
       AAA Batteries (4-pack)       1281
       Lightning Charging Cable     1324
       USB-C Charging Cable         1451
       Wired Headphones             1173
10     AA Batteries (4-pack)        2240
       AAA Batteries (4-pack)       2234
       Lightning Charging Cable     2414
       USB-C Charging Cable         2437
       Wired Headphones             2091
11     AA Batteries (4-pack)        1970
       AAA Batteries (4-pack)       1999
       Lightning Charging Cable     2044
       USB-C Charging Cable         2054
       Wired Headphones             1777
12     AA Batteries (4-pack)        2716
       AAA Batteries (4-pack)       2828
       Lightning Charging Cable     2887
       USB-C Charging Cable         2979
```

```
        Wired Headphones            2537
dtype: int64
```

In [55]:

```
most_sold_product_df.groupby(['Month' , 'Product']).size().unstack()
```

Out[55]:

Product	AA Batteries (4-pack)	AAA Batteries (4-pack)	Lightning Charging Cable	USB-C Charging Cable	Wired Headphones
Month					
1	1037	1084	1069	1171	1004
2	1274	1320	1393	1511	1179
3	1672	1645	1749	1766	1512
4	2062	1988	2197	2074	1888
5	1821	1888	1929	1879	1729
6	1540	1451	1560	1531	1334
7	1555	1554	1690	1667	1434
8	1357	1340	1354	1339	1191
9	1314	1281	1324	1451	1173
10	2240	2234	2414	2437	2091
11	1970	1999	2044	2054	1777
12	2716	2828	2887	2979	2537

In [56]:

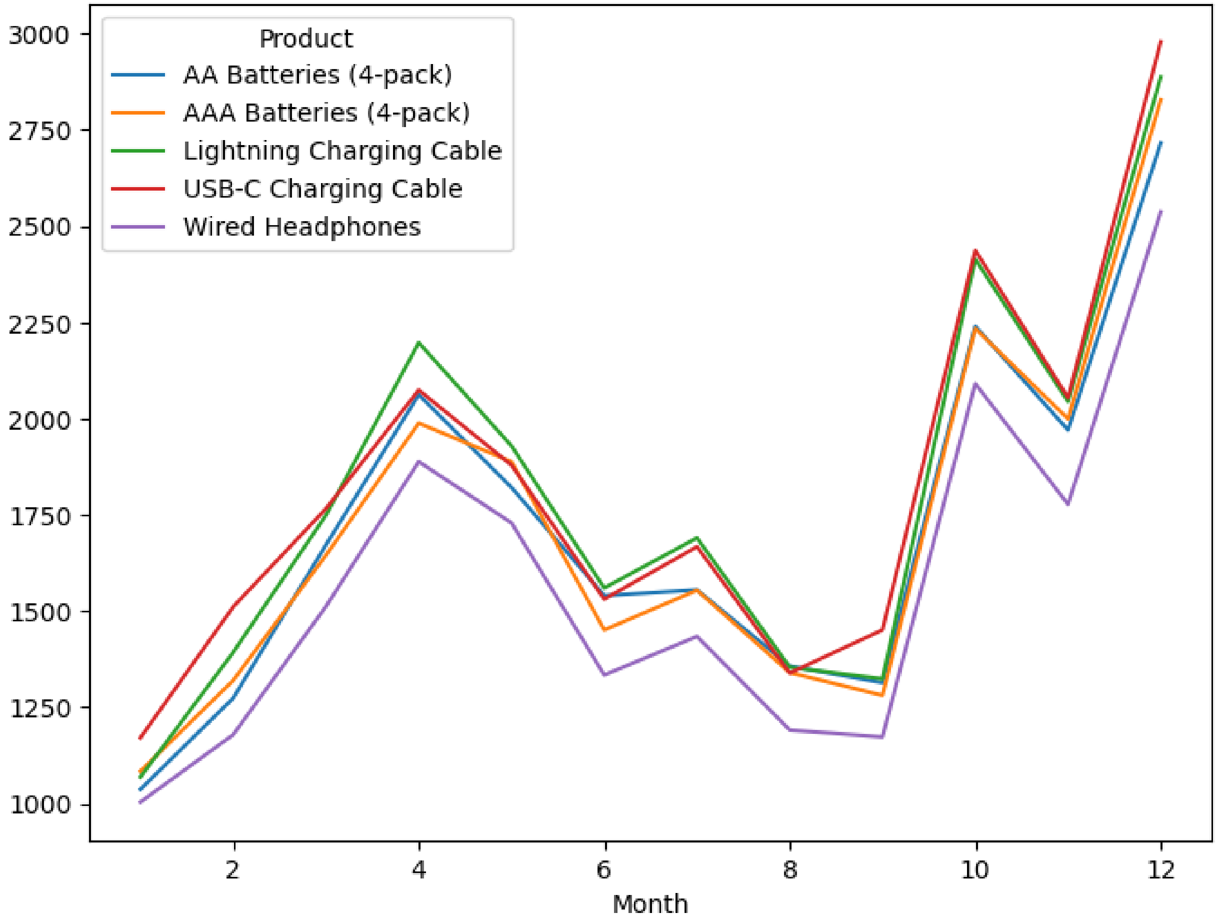
```
pivot = most_sold_product_df.groupby(['Month' , 'Product']).size().unstack()
```

In [57]:

```
pivot.plot(figsize=(8,6))
```

Out[57]:

```
<Axes: xlabel='Month'>
```



In [58]:
```
all_data['Order ID']
```

Out[58]:
```
0         176558
2         176559
3         176560
4         176560
5         176561
           ...
186845    259353
186846    259354
186847    259355
186848    259356
186849    259357
Name: Order ID, Length: 185686, dtype: object
```

In [59]:
```
all_data['Order ID'].duplicated(keep = False)
```

Out[59]:
```
0         False
2         False
3          True
4          True
5         False
          ...
186845    False
186846    False
186847    False
186848    False
186849    False
Name: Order ID, Length: 185686, dtype: bool
```

In [60]:

```
all_data[all_data['Order ID'].duplicated(keep = False)]
```

Out[60]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sale	city
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
18	176574	Google Phone	1	600.00	04/03/19 19:42	20 Hill St, Los Angeles, CA 90001	4	600.00	Los Angeles
19	176574	USB-C Charging Cable	1	11.95	04/03/19 19:42	20 Hill St, Los Angeles, CA 90001	4	11.95	Los Angeles
32	176586	AAA Batteries (4-pack)	2	2.99	04/10/19 17:00	365 Center St, San Francisco, CA 94016	4	5.98	San Francisco
...	...	...	...	...	...	...	...	...	...
186792	259303	AA Batteries (4-pack)	1	3.84	09/20/19 20:18	106 7th St, Atlanta, GA 30301	9	3.84	Atlanta
186803	259314	Wired Headphones	1	11.99	09/16/19 00:25	241 Highland St, Atlanta, GA 30301	9	11.99	Atlanta
186804	259314	AAA Batteries (4-pack)	2	2.99	09/16/19 00:25	241 Highland St, Atlanta, GA 30301	9	5.98	Atlanta
186841	259350	Google Phone	1	600.00	09/30/19 13:49	519 Maple St, San Francisco, CA 94016	9	600.00	San Francisco
186842	259350	USB-C Charging Cable	1	11.95	09/30/19 13:49	519 Maple St, San Francisco, CA 94016	9	11.95	San Francisco

14128 rows × 9 columns

In [61]:

```
df_duplicated = all_data[all_data['Order ID'].duplicated(keep = False)]
```

In [62]:

```
df_duplicated.groupby(['Order ID'])['Product'].apply (lambda x : ','. join(x))
```

Out[62]:

```
Order ID
141275              USB-C Charging Cable,Wired Headphones
141290     Apple Airpods Headphones,AA Batteries (4-pack)
141365                 Vareebadd Phone,Wired Headphones
141384                 Google Phone,USB-C Charging Cable
141450           Google Phone,Bose SoundSport Headphones
                               ...
319536               Macbook Pro Laptop,Wired Headphones
319556                    Google Phone,Wired Headphones
319584                          iPhone,Wired Headphones
319596                  iPhone,Lightning Charging Cable
319631    34in Ultrawide Monitor,Lightning Charging Cable
Name: Product, Length: 6879, dtype: object
```

In [63]:

```
df_duplicated.groupby(['Order ID'])['Product'].apply (lambda x : ','. join(x))
```

Out[63]:

	Order ID	grouped_products
0	141275	USB-C Charging Cable,Wired Headphones
1	141290	Apple Airpods Headphones,AA Batteries (4-pack)
2	141365	Vareebadd Phone,Wired Headphones
3	141384	Google Phone,USB-C Charging Cable
4	141450	Google Phone,Bose SoundSport Headphones
...	...	...
6874	319536	Macbook Pro Laptop,Wired Headphones
6875	319556	Google Phone,Wired Headphones
6876	319584	iPhone,Wired Headphones
6877	319596	iPhone,Lightning Charging Cable
6878	319631	34in Ultrawide Monitor,Lightning Charging Cable

6879 rows × 2 columns

In [64]:

```
dup_products = df_duplicated.groupby(['Order ID'])['Product'].apply (lambda x
```

In [65]:

```
df_duplicated .merge(dup_products, how = 'left' , on = 'Order ID')
```

Out[65]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sale	city	g
0	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles	
1	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles	
2	176574	Google Phone	1	600.00	04/03/19 19:42	20 Hill St, Los Angeles, CA 90001	4	600.00	Los Angeles	G
3	176574	USB-C Charging Cable	1	11.95	04/03/19 19:42	20 Hill St, Los Angeles, CA 90001	4	11.95	Los Angeles	G
4	176586	AAA Batteries (4-pack)	2	2.99	04/10/19 17:00	365 Center St, San Francisco, CA 94016	4	5.98	San Francisco	
...	...	...	...	...	...	...	...	...	...	
14123	259303	AA Batteries (4-pack)	1	3.84	09/20/19 20:18	106 7th St, Atlanta, GA 30301	9	3.84	Atlanta	
14124	259314	Wired Headphones	1	11.99	09/16/19 00:25	241 Highland St, Atlanta, GA 30301	9	11.99	Atlanta	
14125	259314	AAA Batteries (4-pack)	2	2.99	09/16/19 00:25	241 Highland St, Atlanta, GA 30301	9	5.98	Atlanta	
14126	259350	Google Phone	1	600.00	09/30/19 13:49	519 Maple St, San Francisco, CA 94016	9	600.00	San Francisco	G
14127	259350	USB-C Charging Cable	1	11.95	09/30/19 13:49	519 Maple St, San Francisco, CA 94016	9	11.95	San Francisco	G

14128 rows × 10 columns

In [66]:
```
dup_products_df = df_duplicated .merge(dup_products, how = 'left' , on = 'Orde
```

In [67]:
```
dup_products_df.drop_duplicates(subset = ['Order ID'])
```

Out[67]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sale	city	
0	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles	
2	176574	Google Phone	1	600.00	04/03/19 19:42	20 Hill St, Los Angeles, CA 90001	4	600.00	Los Angeles	
4	176586	AAA Batteries (4-pack)	2	2.99	04/10/19 17:00	365 Center St, San Francisco, CA 94016	4	5.98	San Francisco	
6	176672	Lightning Charging Cable	1	14.95	04/12/19 11:07	778 Maple St, New York City, NY 10001	4	14.95	New York City	
8	176681	Apple Airpods Headphones	1	150.00	04/20/19 10:39	331 Cherry St, Seattle, WA 98101	4	150.00	Seattle	H
...	...	...	...	...	...	...	...	...	...	
14118	259277	iPhone	1	700.00	09/28/19 13:07	795 Willow St, New York City, NY 10001	9	700.00	New York City	
14120	259297	iPhone	1	700.00	09/15/19 18:54	138 Main St, Boston, MA 02215	9	700.00	Boston	
14122	259303	34in Ultrawide Monitor	1	379.99	09/20/19 20:18	106 7th St, Atlanta, GA 30301	9	379.99	Atlanta	
14124	259314	Wired Headphones	1	11.99	09/16/19 00:25	241 Highland St, Atlanta, GA 30301	9	11.99	Atlanta	
14126	259350	Google Phone	1	600.00	09/30/19 13:49	519 Maple St, San Francisco, CA 94016	9	600.00	San Francisco	

6879 rows × 10 columns

In [68]:

```
no_dup_df = dup_products_df.drop_duplicates(subset = ['Order ID'])
```

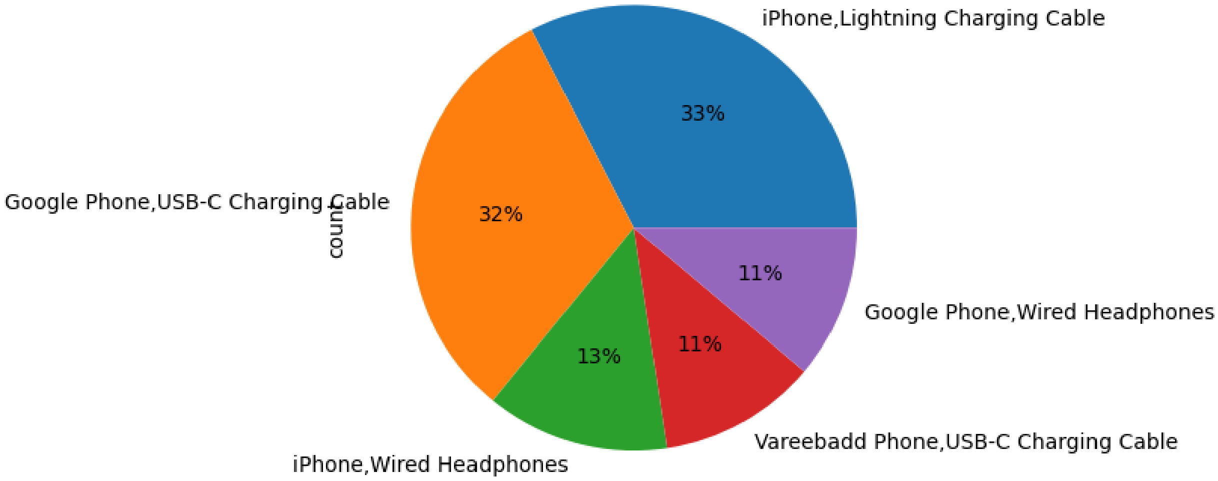
In [ ]:

In [69]:

```
no_dup_df['grouped_products'].value_counts()[0:5].plot(kind='pie' , autopct =
```

Out[69]:

```
<Axes: ylabel='count'>
```



In [ ]:

In [ ]: