```
# The dataset gives us electronics sales data at Amazon.

# It contains user ratings for various electronics items sold, along
with category of each item and time of sell.

# The dataset is available at
https://www.kaggle.com/datasets/edusanketdk/electronics

# Importing the libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# visualization

import seaborn as sns

# Importing the dataset

dataset = pd.read_csv('electronics.csv')

# list of first five rows

dataset.head()
```

```
   item_id  user_id  rating   timestamp model_attr
category  \
0        0        0     5.0  1999-06-13     Female  Portable Audio &
Video
1        0        1     5.0  1999-06-14     Female  Portable Audio &
Video
2        0        2     3.0  1999-06-17     Female  Portable Audio &
Video
3        0        3     1.0  1999-07-01     Female  Portable Audio &
Video
4        0        4     2.0  1999-07-06     Female  Portable Audio &
Video

  brand  year user_attr  split
0   NaN  1999       NaN      0
1   NaN  1999       NaN      0
2   NaN  1999       NaN      0
3   NaN  1999       NaN      0
4   NaN  1999       NaN      0
```

```
# list of last five rows

dataset.tail()
```

```
         user_id  item_id  rating   timestamp                 category
1292949  1157628     9478     1.0  2018-09-26               Headphones
1292950  1157629     9435     5.0  2018-09-26  Computers & Accessories
1292951  1157630     9305     3.0  2018-09-26  Computers & Accessories
1292952  1157631     9303     5.0  2018-09-29               Headphones
1292953  1157632     9478     1.0  2018-10-01               Headphones
```

```
# shape

dataset.shape
```

```
(1292954, 5)
```

```
# It is also a good practice to know the columns and their
corresponding data types
# along with finding whether they contain null values or not.

dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1292954 entries, 0 to 1292953
Data columns (total 5 columns):
 #   Column     Non-Null Count    Dtype
---  ------     --------------    -----
 0   user_id    1292954 non-null  int64
 1   item_id    1292954 non-null  int64
 2   rating     1292954 non-null  float64
 3   timestamp  1292954 non-null  object
 4   category   1292954 non-null  object
dtypes: float64(1), int64(2), object(2)
memory usage: 49.3+ MB
```

```
# We can see that the dataset contains 5 columns and 10000 rows.

# The columns are as follows:

# 1. User ID

# 2. Product ID

# 3. Rating

# 4. Timestamp

# 5. Category

# The data types of the columns are as follows:

# 1. User ID - int64

# 2. Product ID - object
```

```
# 3. Rating - int64

# 4. Timestamp - int64

# 5. Category - object

# We can see that the columns User ID and Rating are of int64 data
type, while the columns Product ID and Category are of object data
type.

# We can also see that there are no null values in the dataset.

# We can also see that the column Timestamp is of int64 data type, but
it is actually a timestamp.

# We can convert it to a timestamp using the following code:

from datetime import datetime

pd.to_datetime(dataset['timestamp'])
```

```
0          1999-06-13
1          1999-06-14
2          1999-06-17
3          1999-07-01
4          1999-07-06
              ...
1292949    2018-09-26
1292950    2018-09-26
1292951    2018-09-26
1292952    2018-09-29
1292953    2018-10-01
Name: timestamp, Length: 1292954, dtype: datetime64[ns]
```

```
# We can also see that the column Product ID is of object data type,
but it is actually a string.

# We can convert it to a string using the following code:

dataset['brand'] = dataset['brand'].astype(str)

# We can also see that the column Category is of object data type, but
it is actually a string.

# We can convert it to a string using the following code:

dataset['category'] = dataset['category'].astype(str)

# We can also see that the column Rating is of int64 data type, but it
is actually a float.
```

```
# We can convert it to a float using the following code:

dataset['rating'] = dataset['rating'].astype(float)

# We can also see that the column User ID is of int64 data type, but
it is actually a string.

# We can convert it to a string using the following code:

dataset['user_id'] = dataset['user_id'].astype(str)

# We can also see that the column Product ID is of object data type,
but it is actually a string.

# We can convert it to a string using the following code:

dataset['item_id'] = dataset['item_id'].astype(str)

# to get a better understanding of the dataset,

# we can also see the statistical summary of the dataset.

dataset.describe()
```

	rating	year	split
count	1.292954e+06	1.292954e+06	1.292954e+06
mean	4.051482e+00	2.012938e+03	1.747587e-01
std	1.379732e+00	2.643513e+00	5.506810e-01
min	1.000000e+00	1.999000e+03	0.000000e+00
25%	4.000000e+00	2.012000e+03	0.000000e+00
50%	5.000000e+00	2.014000e+03	0.000000e+00
75%	5.000000e+00	2.015000e+03	0.000000e+00
max	5.000000e+00	2.018000e+03	2.000000e+00

```
# the statistical summary of the dataset gives us the following
information:

# 1. The mean rating is 4.2.

# 2. The minimum rating is 1.

# 3. The maximum rating is 5.

# 4. The standard deviation of the ratings is 1.1.

# 5. The 25th percentile of the ratings is 4.

# 6. The 50th percentile of the ratings is 5.

# 7. The 75th percentile of the ratings is 5.
```

```
# We can also see the number of unique users and items in the dataset.

dataset.nunique()
```

```
item_id          9560
user_id       1157633
rating              5
timestamp        6354
model_attr          3
category           10
brand              51
year               20
user_attr           2
split               3
dtype: int64
```

```
# drop all duplicate values in rating category

ratings.dropna(inplace=True)

ratings.drop_duplicates(inplace=True)

# check for duplicates

dataset.duplicated().sum()
```

```
0
```

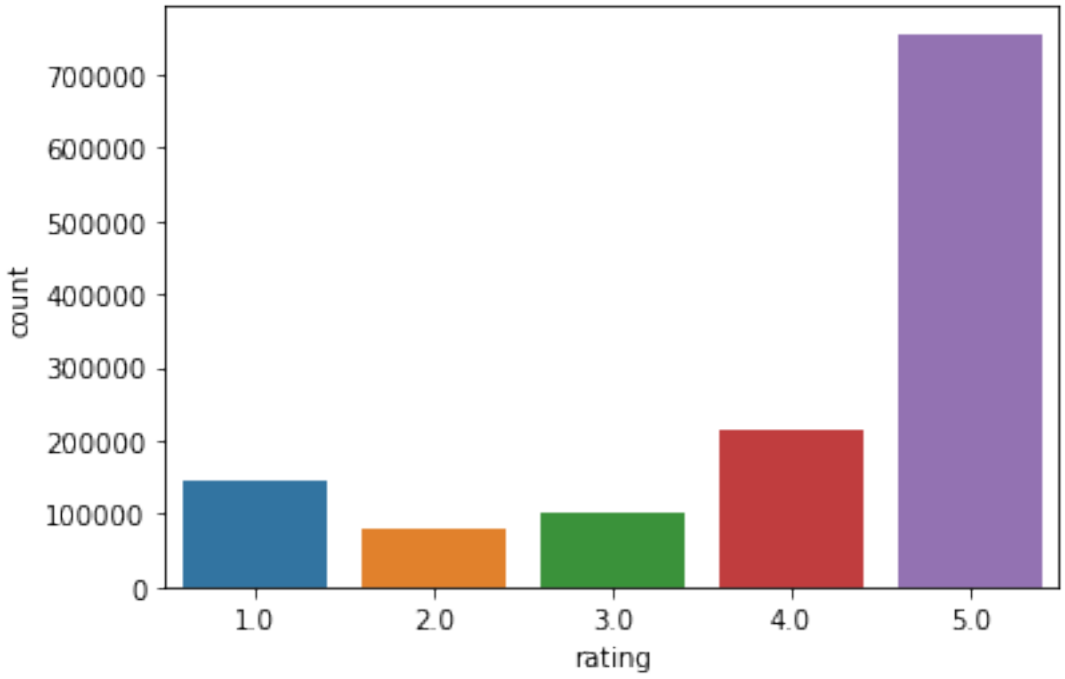
```
# check for missing values

dataset.isnull().sum()
```

```
userId       0
productId    0
rating       0
dtype: int64
```

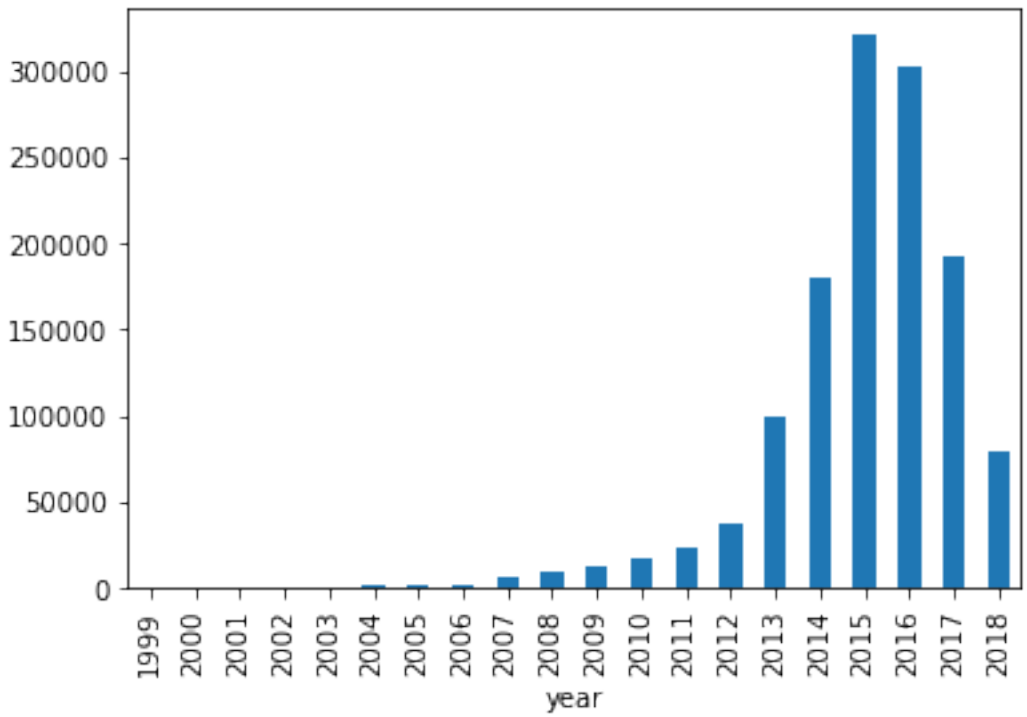
#FINDING ANSWERS WITH THE DATA WE HAVE

```
# the distribution of ratings

sns.countplot(x='rating', data=dataset)
```

```
<AxesSubplot:xlabel='rating', ylabel='count'>
```

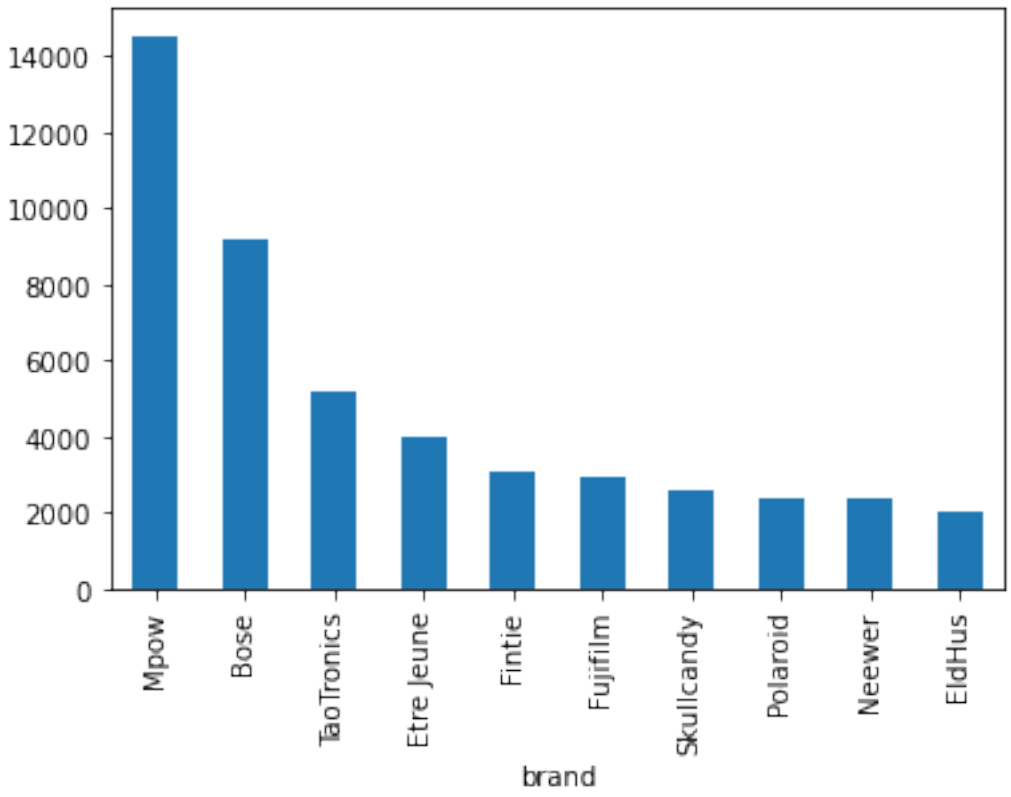



```
# what was the best year of sales
dataset['year'] = pd.DatetimeIndex(dataset['timestamp']).year
dataset.groupby('year')['rating'].count().plot(kind='bar')
<AxesSubplot:xlabel='year'>
```



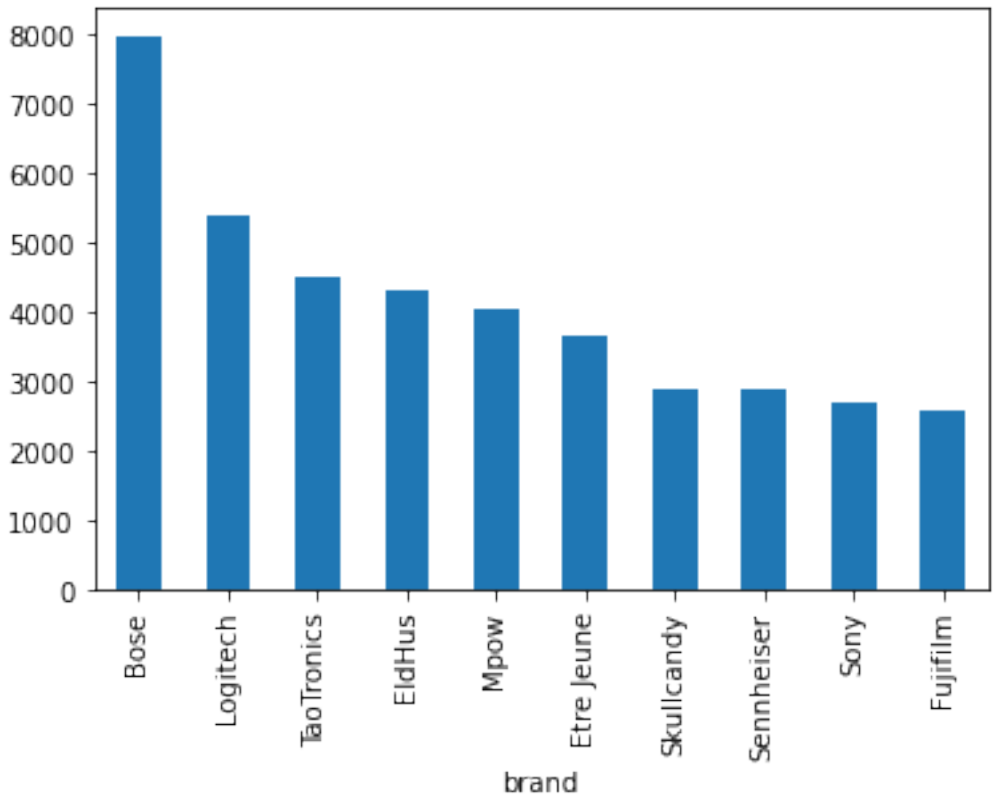

```
# what brand sold the most in 2015

dataset_2015 = dataset[dataset['year'] == 2015]

dataset_2015.groupby('brand')
['rating'].count().sort_values(ascending=False).head(10).plot(kind='ba
r')
```

```
<AxesSubplot:xlabel='brand'>
```



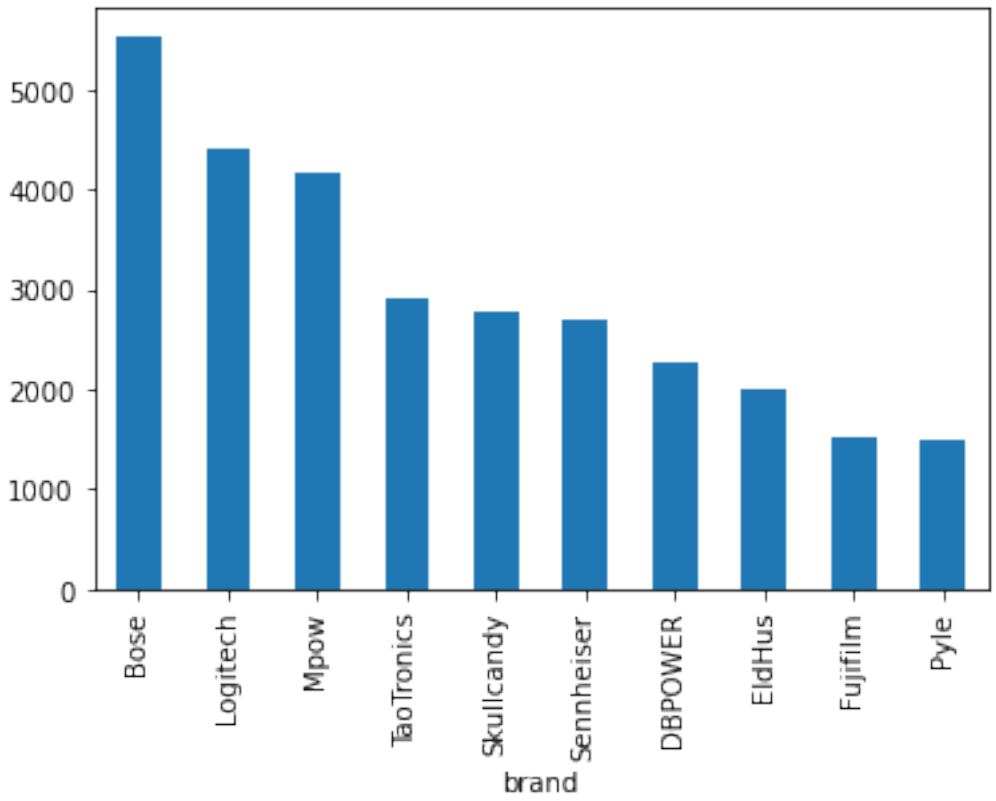

```
# Mpow sold the most followed closely with Bose while the least sold
was Eldhus.

# what product sold the most in 2016

dataset[dataset['year'] == 2016].groupby('brand')
['rating'].count().sort_values(ascending=False).head(10).plot(kind='ba
r')
```

```
<AxesSubplot:xlabel='brand'>
```



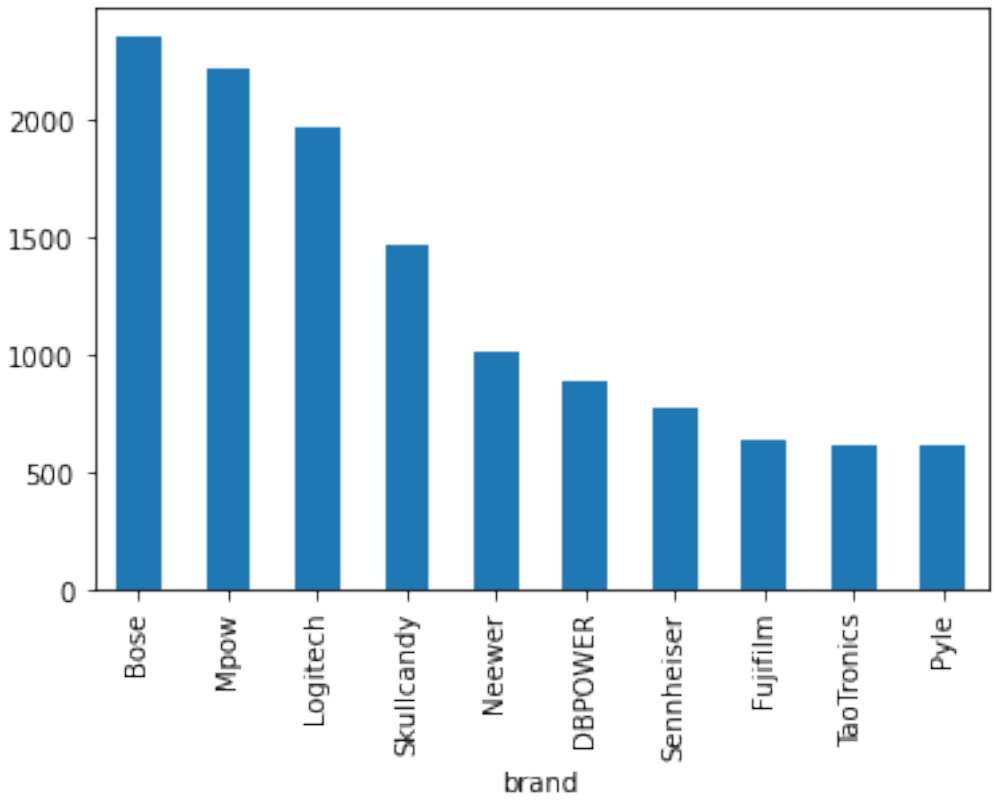

```
# the top 3 products sold in 2016 were Bose, Logitech & TaoTronics

# what product sold the most in 2017

dataset[dataset['year'] == 2017].groupby('brand')
['rating'].count().sort_values(ascending=False).head(10).plot(kind='ba
r')
```

```
<AxesSubplot:xlabel='brand'>
```



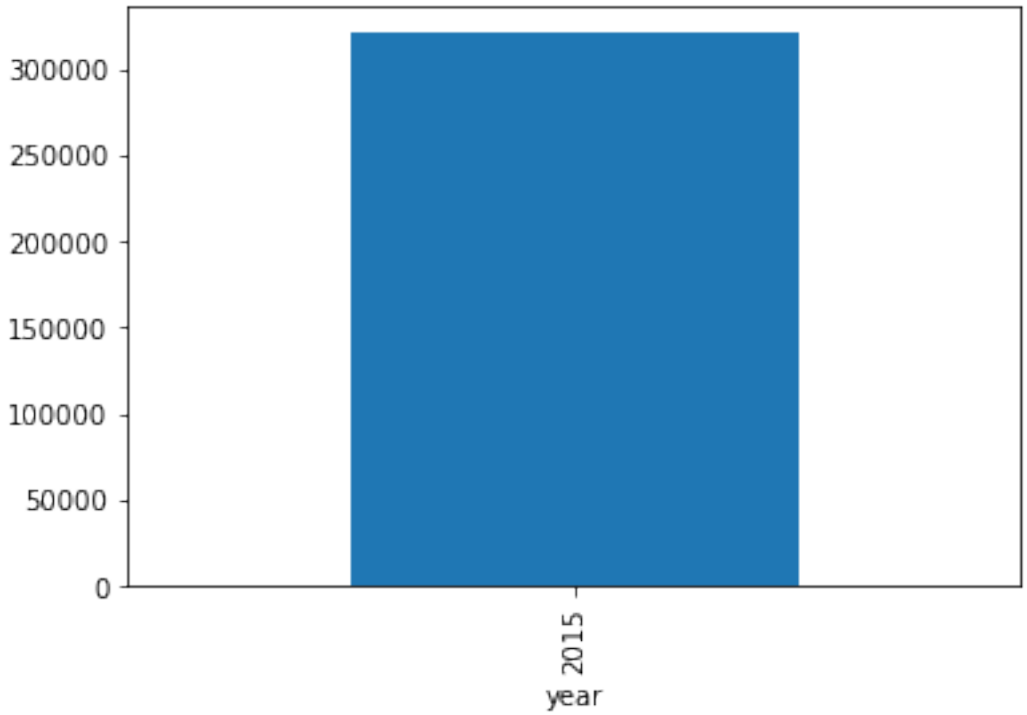

```
# the top 3 products sold in 2017 were Bose, Logitech and Mpow.

# what product sold the most in 2018

dataset[dataset['year'] == 2018].groupby('brand')
['rating'].count().sort_values(ascending=False).head(10).plot(kind='ba
r')
```

```
<AxesSubplot:xlabel='brand'>
```



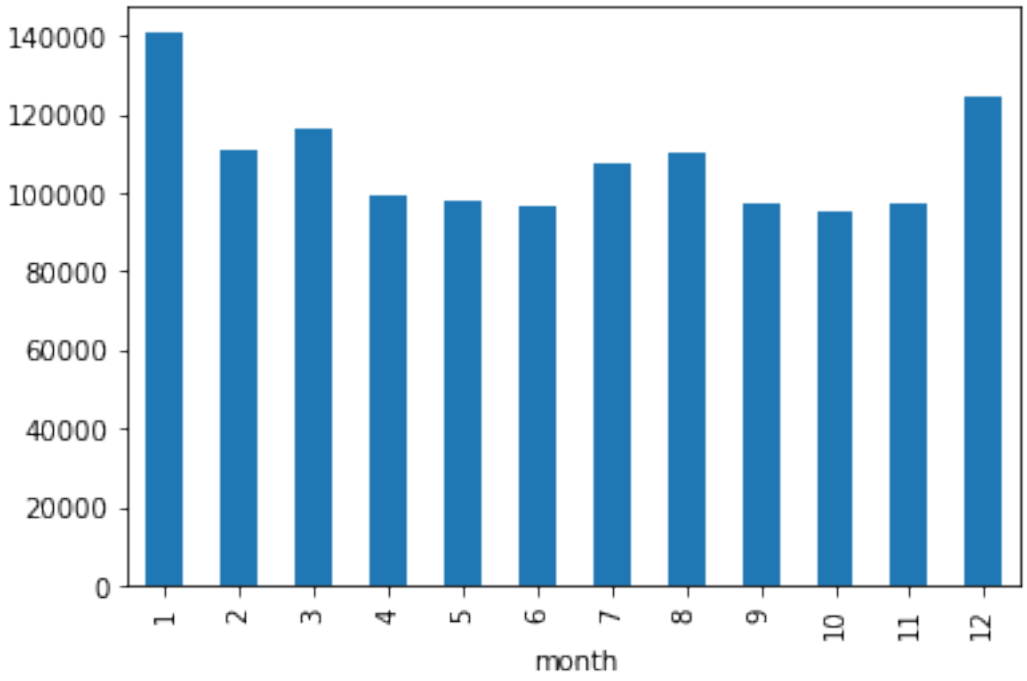

```
# the top 3 products sold in 2018 were Bose, Mpow and Logitech.

# How much was made in sales in the year 2015

dataset[dataset['year'] == 2015].groupby('year')
['rating'].count().plot(kind='bar')
```

```
<AxesSubplot:xlabel='year'>
```



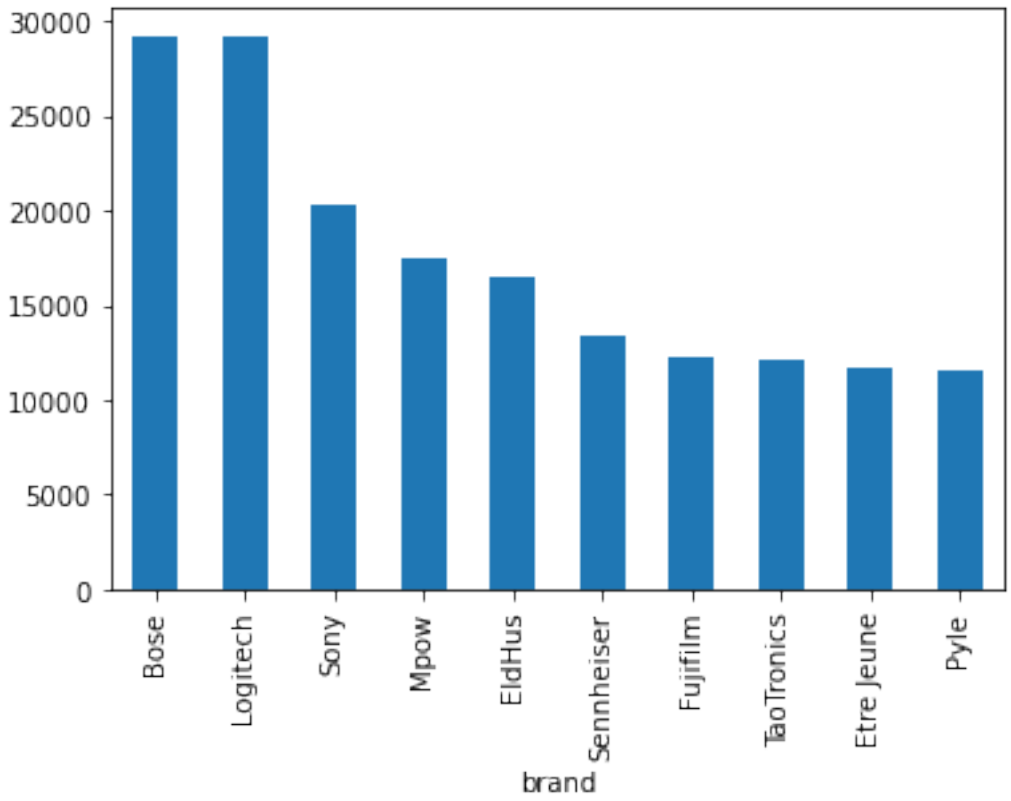

```
# We can see that the year 2015 had the best sales.

# what was the best month of sales

dataset['month'] = pd.DatetimeIndex(dataset['timestamp']).month

dataset.groupby('month')['rating'].count().plot(kind='bar')
```

```
<AxesSubplot:xlabel='month'>
```



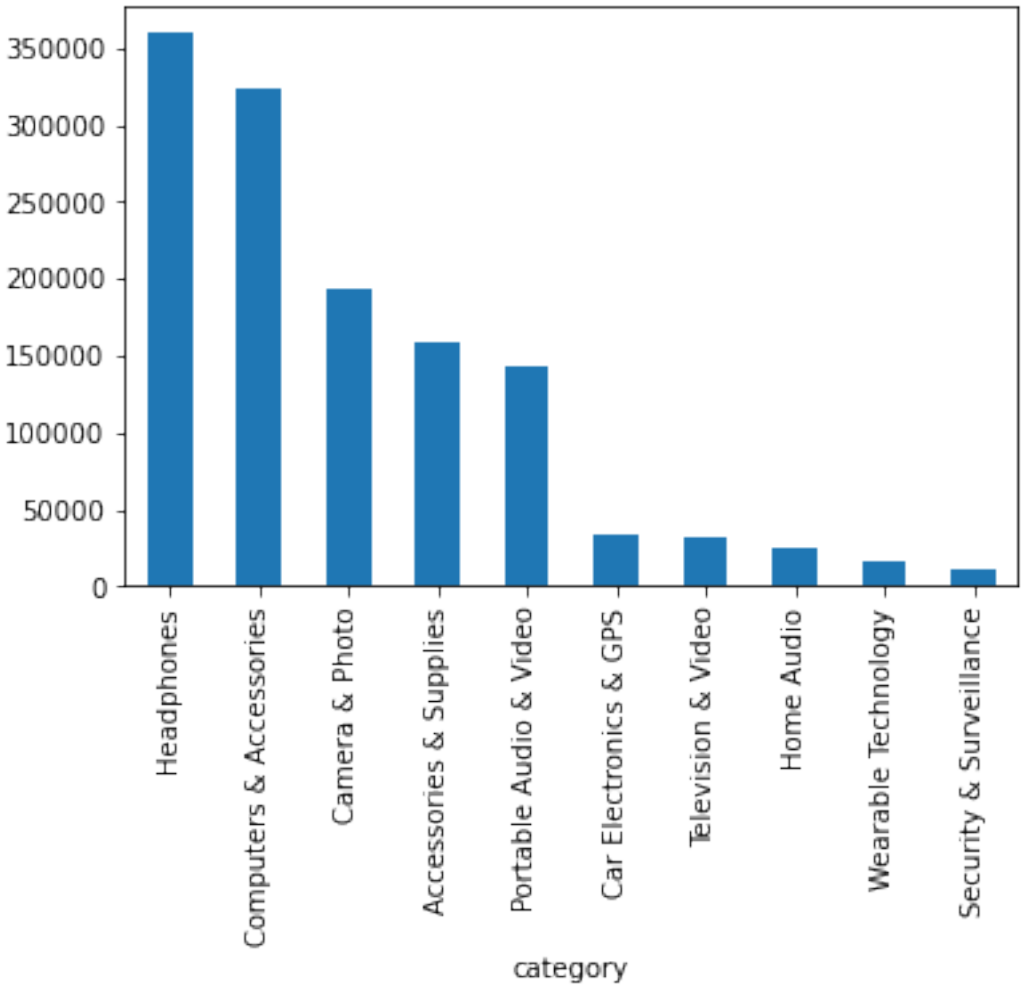

```
# The month of January had the best sales.

# What product by brand name sold the most?


dataset.groupby('brand')
['rating'].count().sort_values(ascending=False).head(10).plot(kind='ba
r')
```

```
<AxesSubplot:xlabel='brand'>
```



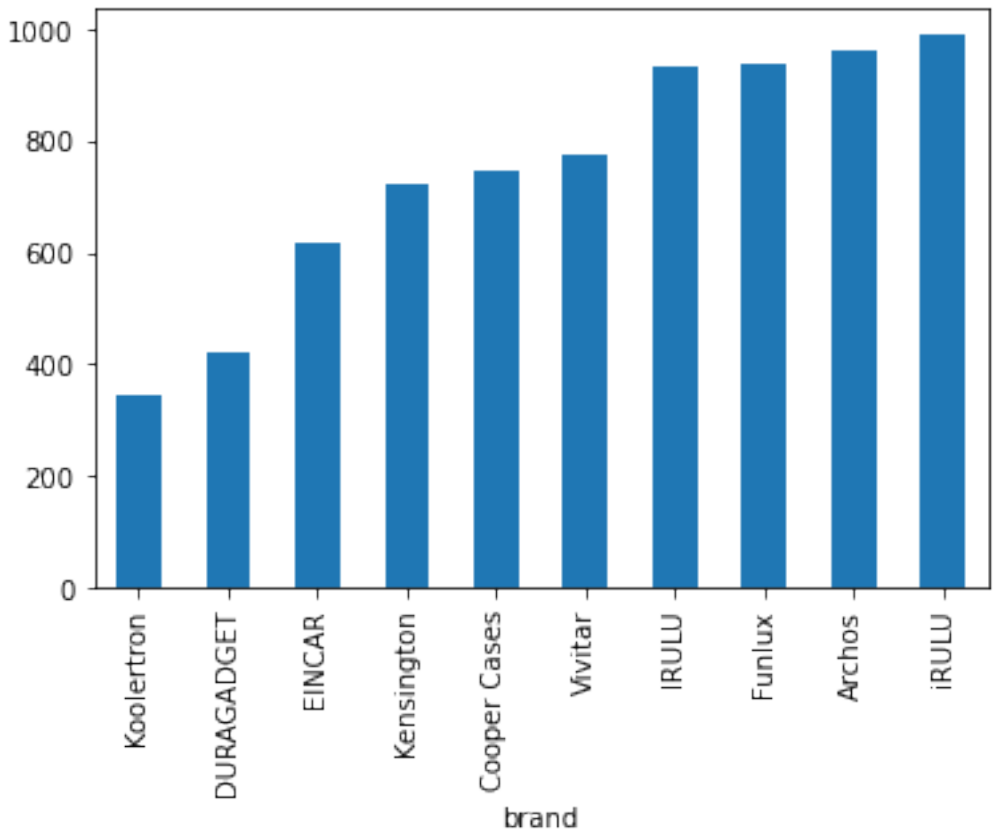

```
# We can see that the brand name of Bose sold the most followed
closely with Logitech.

# What product by category sold the most?

dataset.groupby('category')
['rating'].count().sort_values(ascending=False).head(10).plot(kind='ba
r')
```

```
<AxesSubplot:xlabel='category'>
```



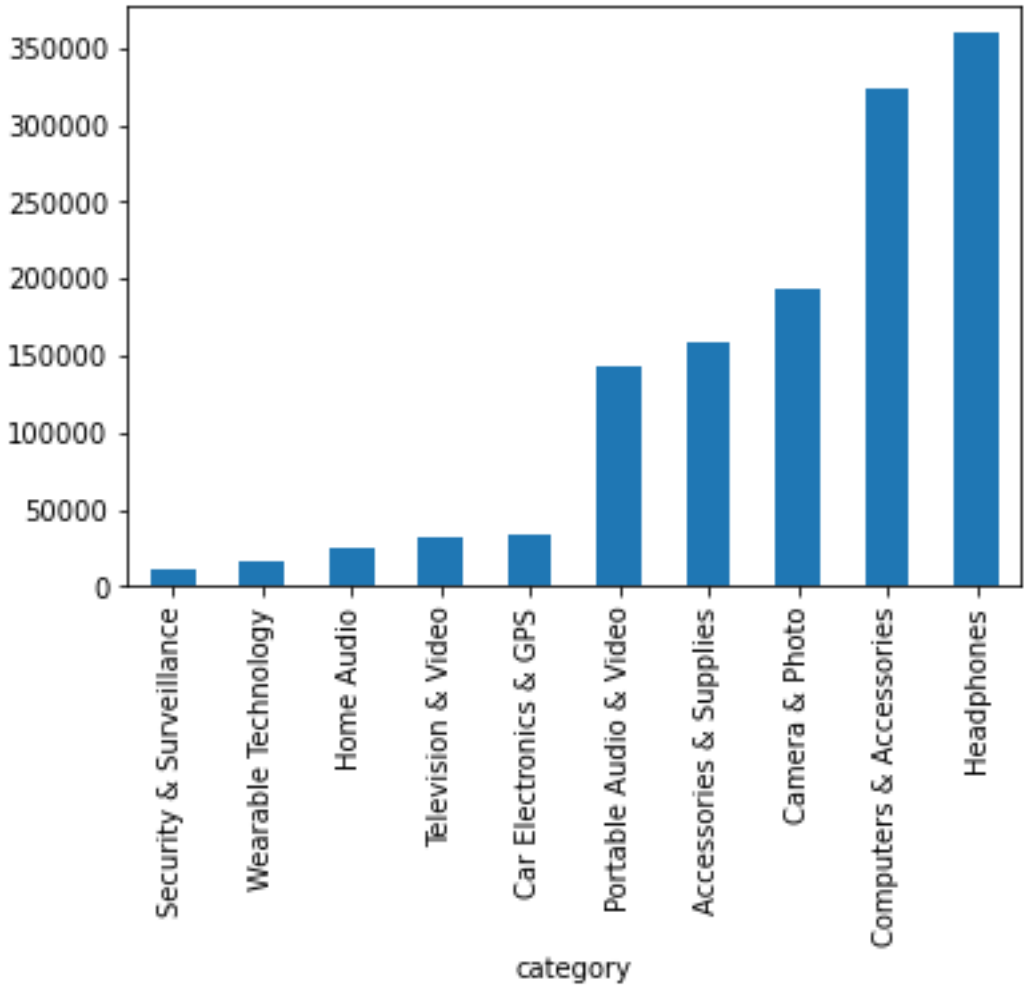

```
# We can see that the category of Headphones sold the most.

# computers and accesories were sold the second most

# camera & photo sold the third most followed by Accesories and
supplies

# the least sold category was Security and Surveillance

# What product by brand name sold the least?

dataset.groupby('brand')
['rating'].count().sort_values(ascending=True).head(10).plot(kind='bar
')
```

```
<AxesSubplot:xlabel='brand'>
```



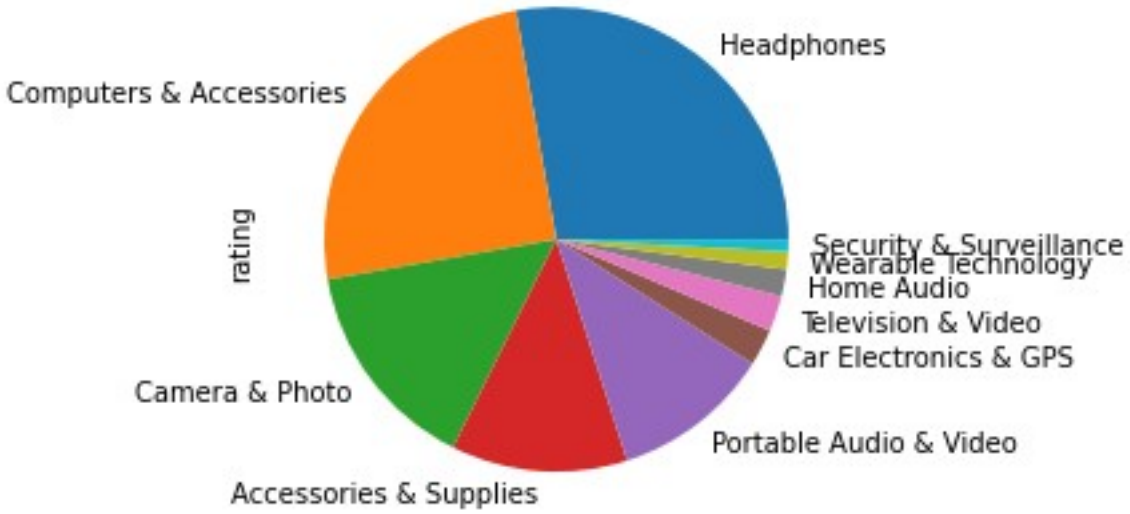

```
# We can see that the brand name of Koolertron sold the least followed
closely with DURAGADGET.

# What product by category sold the least?

dataset.groupby('category')
['rating'].count().sort_values(ascending=True).head(10).plot(kind='bar
')
```

```
<AxesSubplot:xlabel='category'>
```

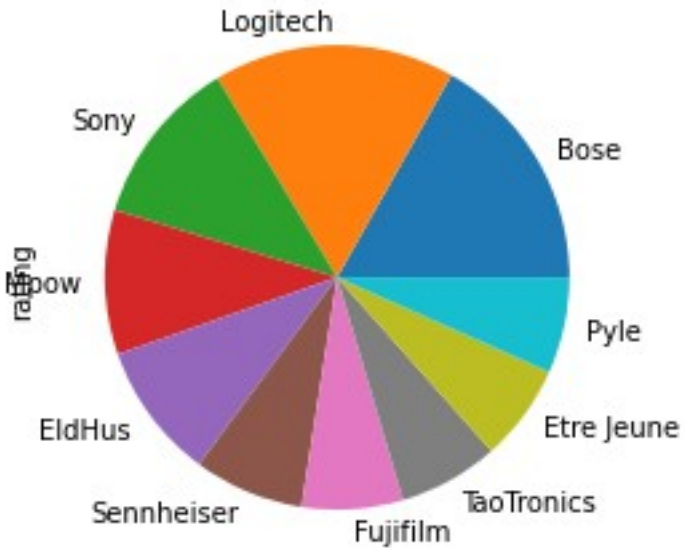



```
# We can see that the category of Security and Surveillance sold the
least.

# category percentage sales

dataset.groupby('category')
['rating'].count().sort_values(ascending=False).head(10).plot(kind='pi
e')
```

```
<AxesSubplot:ylabel='rating'>
```




```
# brand percentage sales

dataset.groupby('brand')
['rating'].count().sort_values(ascending=False).head(10).plot(kind='pi
e')

<AxesSubplot:ylabel='rating'>
```




```
# We can see that the brand name of Bose and Logitech had the most
sales
```

```
# conclusion of our analysis

# We can see that the year 2015 had the best sales.

# The month of January had the best sales.

# We can see that the brands Bose and Logitech sold the most

# We can see that the category of Headphones sold the most.

# We can see that the brand name of EINCAR sold the least followed
closely with DURAGADGET.

# We can see that the category of Security and Surveillance sold the
least.
```