Project on Product Sales Analysis by Gaurav and Bhawna.

INTRODUCTION

This Product Sales Analysis is about to determine the different data available in out dataset.

It shows the categororial division of electric appliances.

Also shows increase or decrease of sales or a change in price according to the year.

with that rating of products as per the consumers.

IMPORTING LIBRARIES

```
#importing required libraries.

import warnings
warnings.filterwarnings('ignore')
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline


# 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	brand	year	user_attr
0	0	0	5.0	1999-06-13	Female	Portable Audio & Video	NaN	1999	NaN
1	0	1	5.0	1999-06-14	Female	Portable Audio & Video	NaN	1999	NaN

```
# list of last five rows

dataset.tail()
```

	item_id	user_id	rating	timestamp	model_attr	category	brand	year	user_attr	split
1292949	9478	1157628	1.0	2018-09-26	Female	Headphones	Etre Jeune	2017	NaN	0
1292950	9435	1157629	5.0	2018-09-26	Female	Computers & Accessories	NaN	2017	NaN	0
1292951	9305	1157630	3.0	2018-09-26	Female	Computers & Accessories	NaN	2016	NaN	0
1292952	9303	1157631	5.0	2018-09-29	Male	Headphones	NaN	2018	NaN	0
1292953	9478	1157632	1.0	2018-10-01	Female	Headphones	Etre Jeune	2017	Female	0

```
# shape

dataset.shape
```

```
(1292954, 10)
```

```
# It help to know the columns and their corresponding data types

dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1292954 entries, 0 to 1292953
Data columns (total 10 columns):
 #   Column      Non-Null Count    Dtype
```

```
 ---  ------      --------------    -----
  0   item_id     1292954 non-null  int64
  1   user_id     1292954 non-null  int64
  2   rating      1292954 non-null  float64
  3   timestamp   1292954 non-null  object
  4   model_attr  1292954 non-null  object
  5   category    1292954 non-null  object
  6   brand       331120 non-null   object
  7   year        1292954 non-null  int64
  8   user_attr   174124 non-null   object
  9   split       1292954 non-null  int64
dtypes: float64(1), int64(4), object(5)
memory usage: 98.6+ MB
```

```
# 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

```
# 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
```

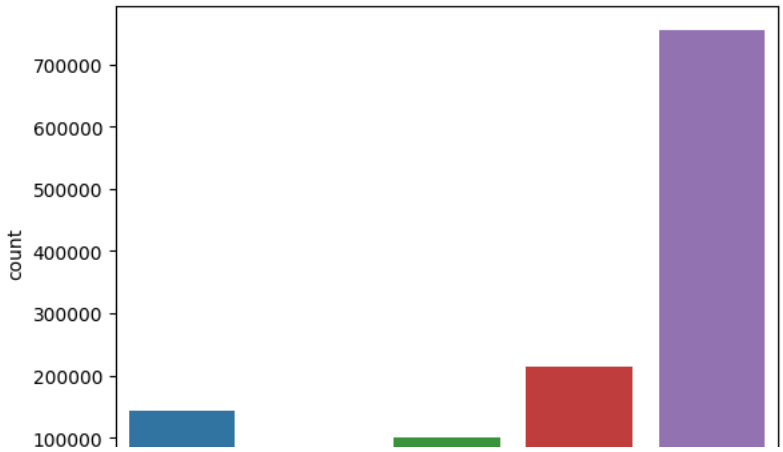
```
# check for missing values

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

```
item_id             0
user_id             0
rating              0
timestamp           0
model_attr          0
category            0
brand               0
year                0
user_attr     1118830
split               0
dtype: int64
```

FINDING ANSWERS WITH THE DATA WE HAVE

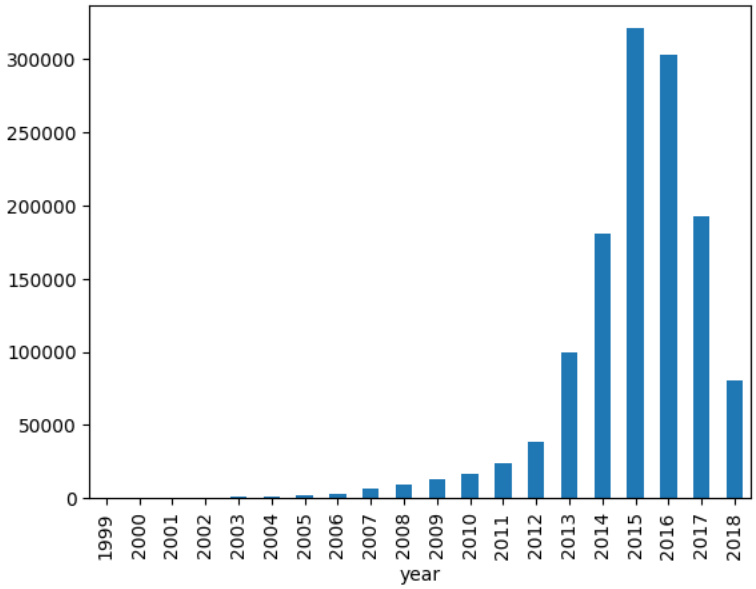
```
# the distribution of ratings

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

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



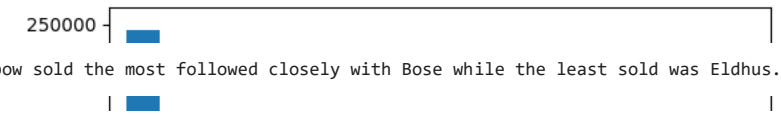

```
# what was the best year of sales

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

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

```
<Axes: 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='bar')
```

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

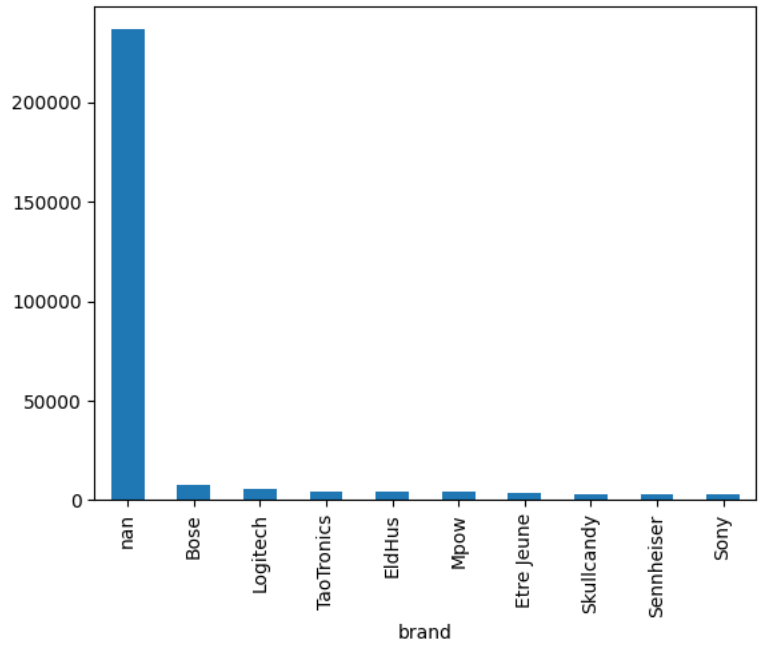



```
# Mpow sold the most followed closely with Bose while the least sold was Eldhus.
```



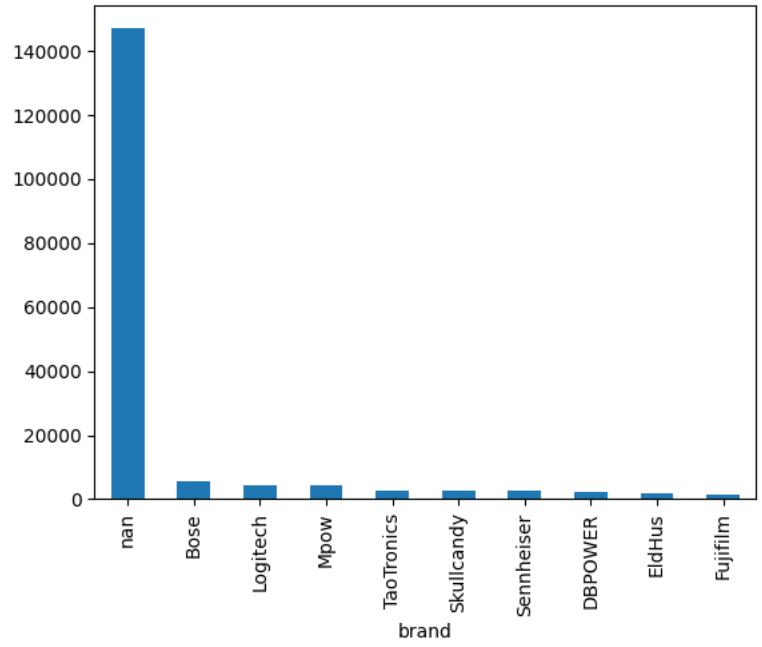
```
# what product sold the most in 2016

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

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



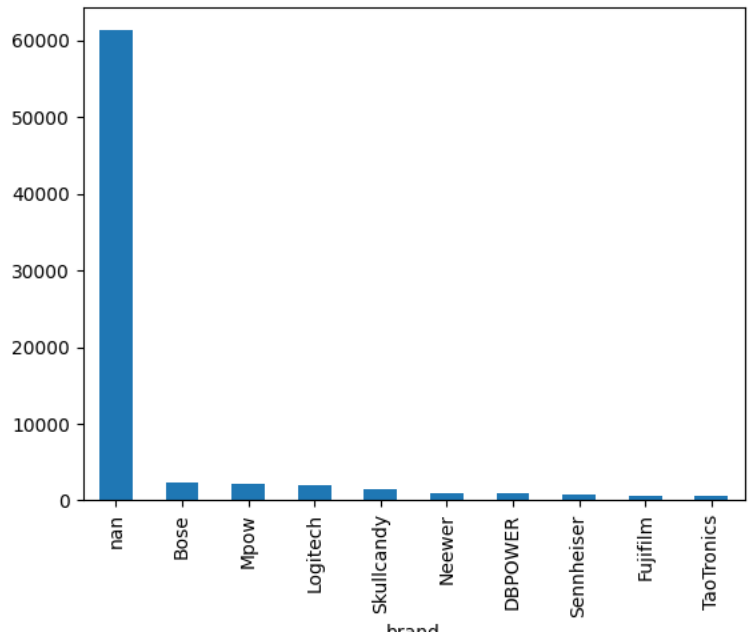

```
# the top 3 products sold in 2016 were Bose, Logitech & TaoTronics

# what product sold the most in 2017

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

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



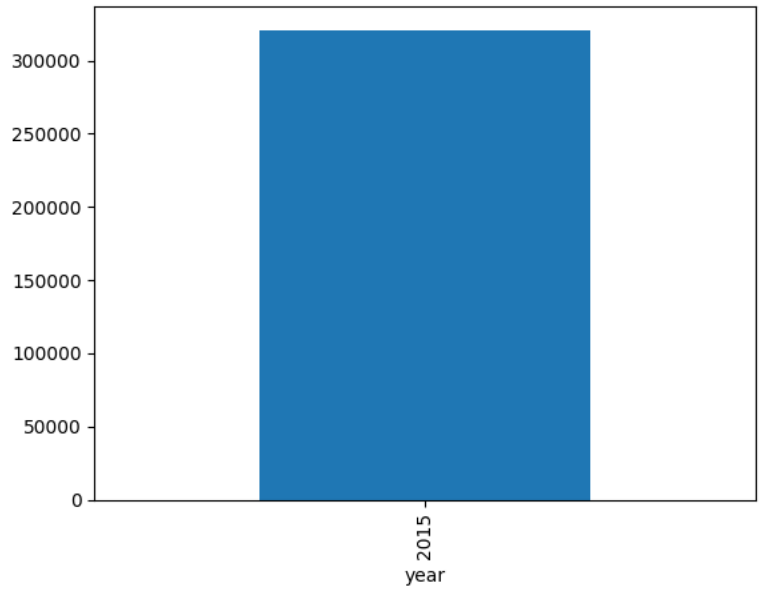

```
# the top 3 products sold in 2017 were Bose, Logitech and Mpow.

# what product sold the most in 2018

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

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



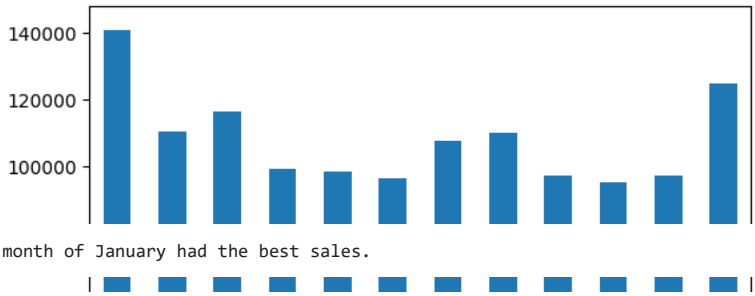

```
# the top 3 products sold in 2018 were Bose, Mpow and Logitech.

# How much was made in sales in the year 2015

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

```
<Axes: 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')
```

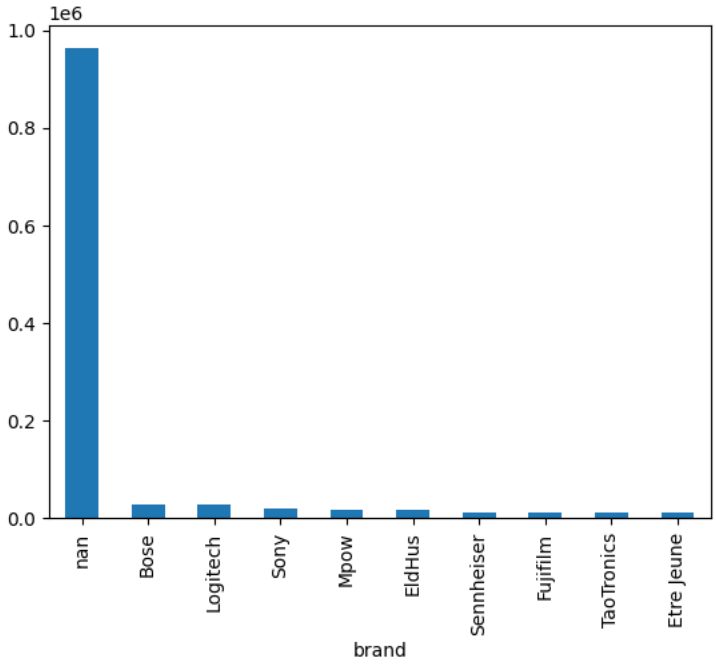
```
<Axes: xlabel='month'>
```




```
# 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='bar')
```

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



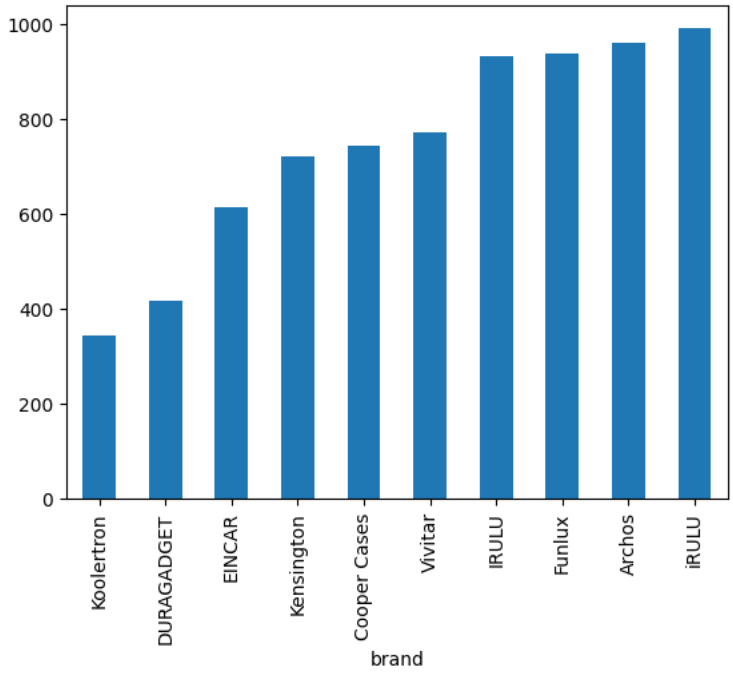

```
# We can see that the brand name of Bose sold the most followed closely with Logitech.

# What product by category sold the most?

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

```
<Axes: 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
```

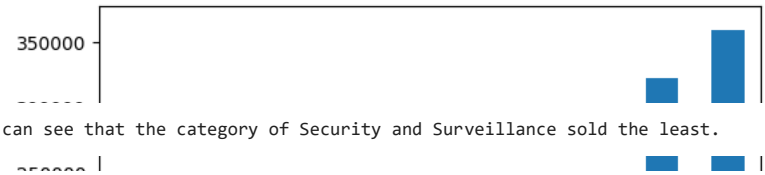

```
# What product by brand name sold the least?

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

```
<Axes: 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')
```

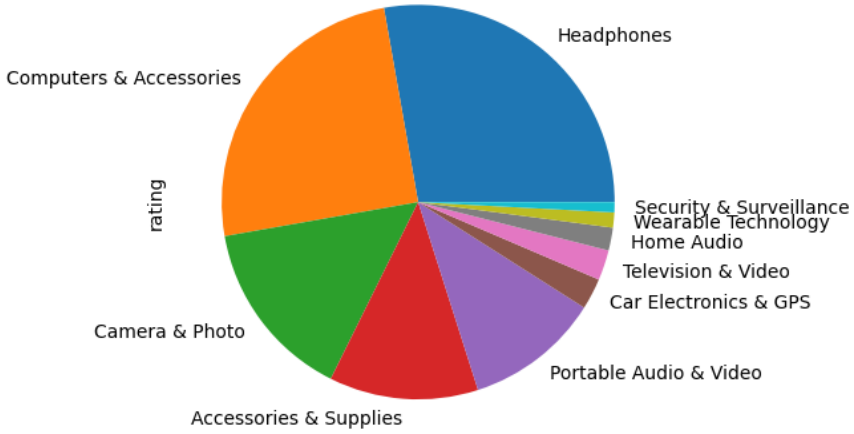
```
<Axes: xlabel='category'>
```




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

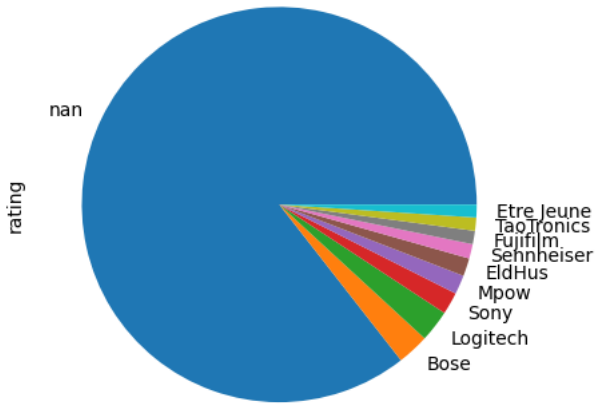
```
# category percentage sales

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

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




```
# brand percentage sales

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

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




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

conclusion of our analysis

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

2.The month of January had the best sales.

3.We can see that the brands Bose and Logitech sold the most

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

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

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