Overview

We are going to consider a dataset of electronics sales data at Amazon. It contains user ratings for various electronics items sold, along with the category of each item and time of sale.

We will use Python libraries (Pandas, Numpy, Matplotlib & Seaborn) to analyze and answer business questions for sales data. The data contains hundreds of thousands of electronics store purchases broken down by month, product type, cost, purchase address, etc.
The dataset can be downloaded here.
https://github.com/AnudipAE/DANLC/blob/master/cleaned.csv

In this analysis, we will be using Jupyter Notebook.

STEP 1:

Exploratory Data Analysis [EDA]

This is the process by which we shall critically perform initial investigations of the data we have to discover patterns, to spot anomalies, test hypotheses and to check assumptions with the help of summary statistics and graphical representations.

It is how we get to understand the data we have and gather many insights from it. It is more of making sense of the data we have before working with it.

```python
# 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('https://raw.githubusercontent.com/AnudipAE/DANLC/master/cleaned.csv')

# list of first five rows

dataset.head()
```

Output:

index	item_id	user_id	rating	timestamp	gender	category	brand	year	month	quantity	unitprice	amount
0	7	131	4	36692	Female	Home Audio	Philips	2000	6	5	6360	31800
1	19	231	5	36891	Female	Camera	Canon	2000	12	10	9955	99550
2	14	233	5	36893	Female	Camera	Kodak	2001	1	9	7639	68751
3	14	257	5	36926	Female	Camera	Kodak	2001	2	7	5097	35679
4	14	269	5	36952	Female	Camera	Kodak	2001	3	10	6472	64720

To take a look at the first five rows we use the pandas function “ .head()”. Similarly “.tail()” returns the last five observations of the data set.

```python
# list of last five rows

dataset.tail()
```

Output:

```
# list of last five rows

dataset.tail()
```

	item_id	user_id	rating	timestamp	gender	category	brand	year	month	quantity	unitprice	amount
45161	7828	1157458	5	43341	Female	Headphones	Bose	2018	8	7	5925	41475
45162	8624	1157504	5	43342	Female	Headphones	Pyle	2018	8	7	9717	68019
45163	9513	1157527	5	43344	Male	Headphones	Mpow	2018	9	8	9197	73576
45164	9125	1157555	3	43348	Female	Headphones	EldHus	2018	9	10	8848	88480
45165	9478	1157632	1	43374	Female	Headphones	Etre Jeune	2018	10	7	7717	54019

To know the total number of rows and columns in the data set we use ".shape" as shown below.

```
# shape

dataset.shape
```

Output:

```
# shape

dataset.shape
```

```
(45166, 12)
```

Inference:
Dataset comprises 45166 Rows and 12 columns.

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.

```
# 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()
```

Output:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45166 entries, 0 to 45165
Data columns (total 12 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   item_id    45166 non-null  int64 
 1   user_id    45166 non-null  int64 
 2   rating     45166 non-null  int64 
 3   timestamp  45166 non-null  int64 
 4   gender     45166 non-null  object
 5   category   45166 non-null  object
 6   brand      45166 non-null  object
 7   year       45166 non-null  int64 
 8   month      45166 non-null  int64 
 9   quantity   45166 non-null  int64 
 10  unitprice  45166 non-null  int64 
 11  amount     45166 non-null  int64 
```

Inference:

No Variable column has null/missing values

We can see that the dataset contains 12 columns and 45166 rows.

The columns are as follows:

1. item_id
2. user_id
3. rating
4. timestamp
5. gender
6. category
7. brand
8. year
9. month
10. quantity
11. unitprice
12. amount

The data types of the columns are as follows:

1. item_id int64
2. user_id int64
3. rating int64
4. timestamp int64
5. gender object
6. category object
7. brand object
8. year int64
9. month int64
10. quantity int64
11. unitprice int64

12. amount	int64

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 there are no null values in the dataset. The column Timestamp is of int64 data type.

The column Product ID is of object data type, but it is actually a string, the column Category is of object data type, but it is actually a string.

To get a better understanding of the dataset, we can also see the statistical summary of the dataset using the function ".describe()".

This includes count, mean, median (or 50th percentile) standard variation, min-max, and percentile values of columns as shown below.

```
# to get a better understanding of the dataset,

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

dataset['rating'].describe()
```

Output:

```
count    45166.000000
mean         4.218594
std          1.221118
min          1.000000
25%          4.000000
50%          5.000000
75%          5.000000
max          5.000000
Name: rating, dtype: float64
```

Inference:
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 1.22
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.

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

dataset.nunique()
```

Output:

```
item_id       1892
user_id      40401
rating           5
timestamp     4179
gender           2
category        10
brand           50
year            19
month           12
quantity         6
unitprice     5001
amount       19611
dtype: int64
```

<u>Dealing With Missing Values</u>

There can be multiple reasons why certain values are missing from the data. Reasons for the missing data from the dataset affect the approach of handling missing data. So it’s necessary to understand why the data could be missing.

Some of the reasons are listed below:
Past data might get corrupted due to improper maintenance.
Observations are not recorded for certain fields due to some reasons.
There might be a failure in recording the values due to human error.
The user has not provided the values intentionally.

```
# check for missing values

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

Output:

```
item_id       0
user_id       0
rating        0
timestamp     0
gender        0
category      0
brand         0
year          0
month         0
quantity      0
unitprice     0
amount        0
```

Image: Checking sum of Null Values

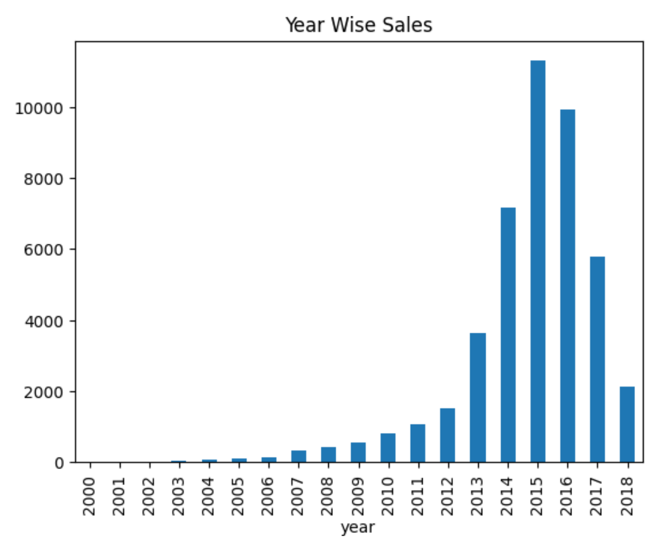
<u>Finding Answers with the Data Using Visualizations</u>

To make it easier to understand, we are going to use matplotlib and seaborn that we earlier imported to visualize our results with simple bar charts. This will make it easier to answer questions that might arise from the data set.

i.) What was the best year of sales?

```
# what was the best year of sales

dataset.groupby('year')['amount'].count().plot(kind='bar',title='Year
Wise Sales')
```

Output:

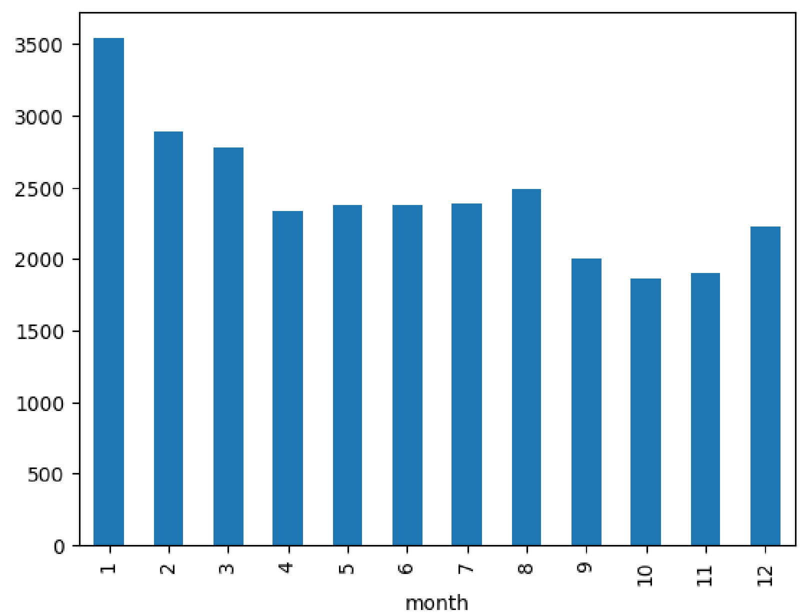


Inference:
From the graph we just plotted we can see that year 2015 had the best sales out of all years.
There was a steady increase of sales from the year 2007 to 2015 then a slight decline in 2016. That decline in sales was big in the following years of 2017 and 2018.

ii.) Which was the best month for sales between 2015 t0 2018

```
# We can see that the year 2015  to 2018 had the best sales.

# what was the best month of sales
dataset_2015_2018 = dataset[(dataset['year'] >= 2015) & (dataset['year']
<= 2018)]

dataset_2015_2018.groupby('month')['rating'].count().plot(kind='bar')
```

Output:



Inference:
January was the month when most sales were made across the product categories and over the years.

iii.) What brand sold the most in the highest selling year(2015 to 2018)

```
# what brand sold the most in 2015 to 2018

dataset_2015_2018 = dataset[(dataset['year'] >= 2015) & (dataset['year']
<= 2018)]
dataset_2015_2018.groupby('brand')['amount'].sum().sort_values(ascending
=False).head(10)\
.plot(kind='bar',title='Brand Wise Top 10 Sales 2015 to
2018',y='amount')
```

Output:

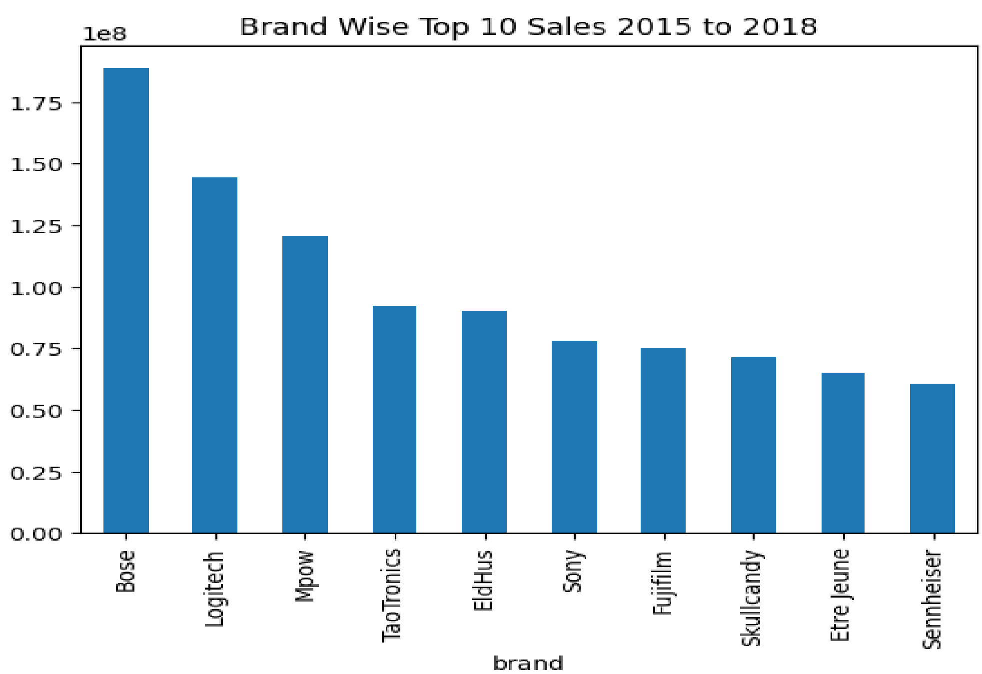


Image: Best selling Brand

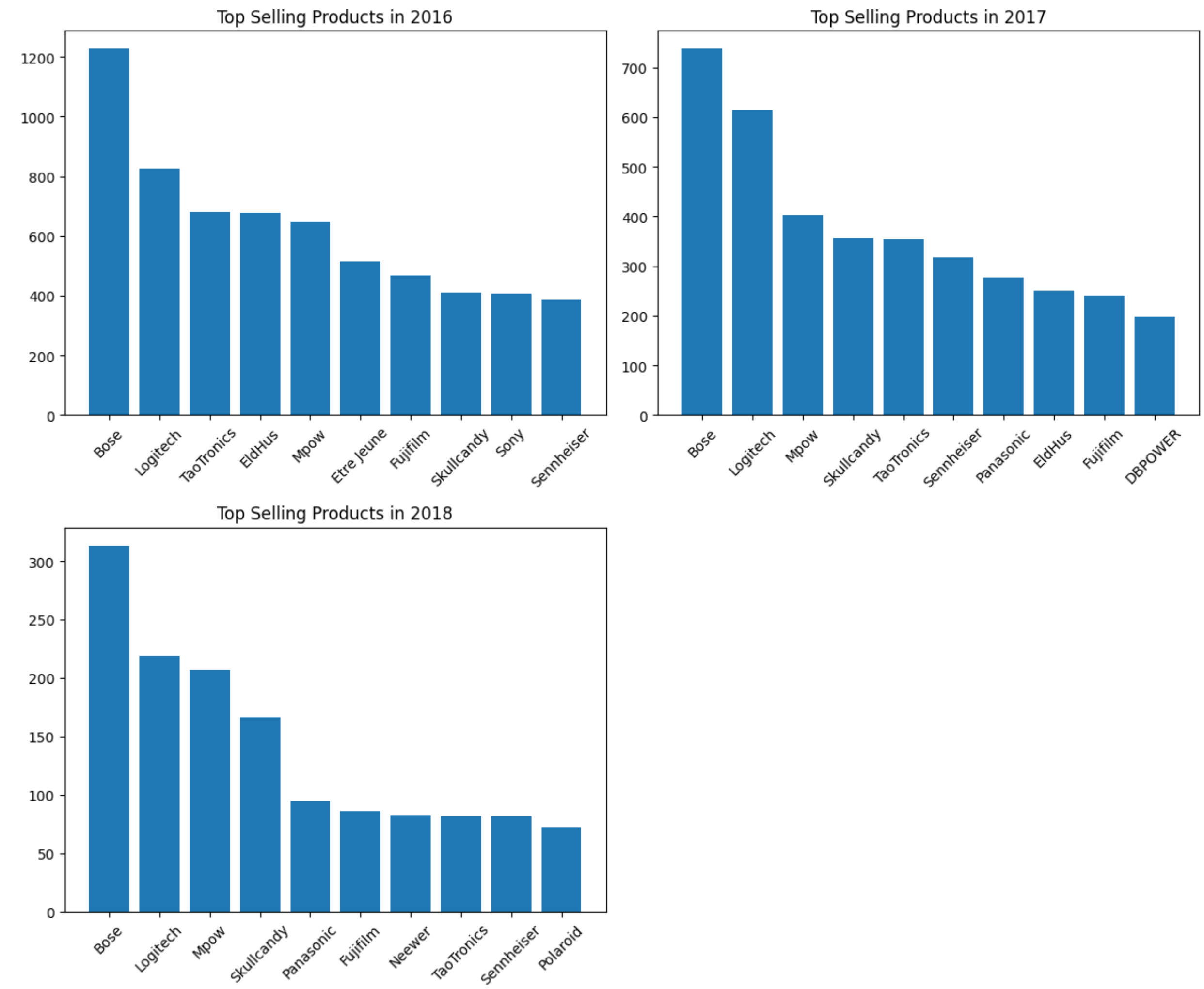
Inference:
Bose was the brand with the most sales in 2015 to 2018 followed by Logitech.

iv.) What products sold the most in the three years 2016, 2017 & 2018

```
# Create subplots with 2 rows and 2 columns
fig, axs = plt.subplots(2, 2, figsize=(12, 10))
```

```
# Plot for 2016
top_selling_2016 = dataset[dataset['year'] ==
2016].groupby('brand')['rating'].count().sort_values(ascending=False).he
ad(10)
axs[0, 0].bar(top_selling_2016.index, top_selling_2016)
axs[0, 0].set_title('Top Selling Products in 2016')
axs[0, 0].tick_params(axis='x', rotation=45)  # Rotate x-axis labels

# Plot for 2017
top_selling_2017 = dataset[dataset['year'] ==
2017].groupby('brand')['rating'].count().sort_values(ascending=False).he
ad(10)
axs[0, 1].bar(top_selling_2017.index, top_selling_2017)
axs[0, 1].set_title('Top Selling Products in 2017')
axs[0, 1].tick_params(axis='x', rotation=45)  # Rotate x-axis labels

# Plot for 2018
top_selling_2018 = dataset[dataset['year'] ==
2018].groupby('brand')['rating'].count().sort_values(ascending=False).he
ad(10)
axs[1, 0].bar(top_selling_2018.index, top_selling_2018)
axs[1, 0].set_title('Top Selling Products in 2018')
axs[1, 0].tick_params(axis='x', rotation=45)  # Rotate x-axis labels

# Hide the empty subplot
axs[1, 1].axis('off')

# Adjust layout for better appearance
plt.tight_layout()

# Show the plots
plt.show()
```

Output:



Inference:
There has been one consistent Brand product with the most sales in the 3 years and it is Bose.
The second most sold brand's products have been Logitech.

- 2016 (Bose and Logitech)
- 2017 (Bose and Logitech)
- 2018 (Bose and Logitech)

v.) What product by category sold the most between 2015 to 2018?

```
# # What product by category sold the most between 2015 to 2018?
dataset2015_2018 = dataset[(dataset['year'] >= 2015) & (dataset['year']
<= 2018)]
```

```
dataset2015_2018.groupby('category')['amount'].sum().sort_values(ascending=False).head(10).plot(kind='bar',title='Top 10 Most Sold Product Category 2015 to 2018')
```

Output:

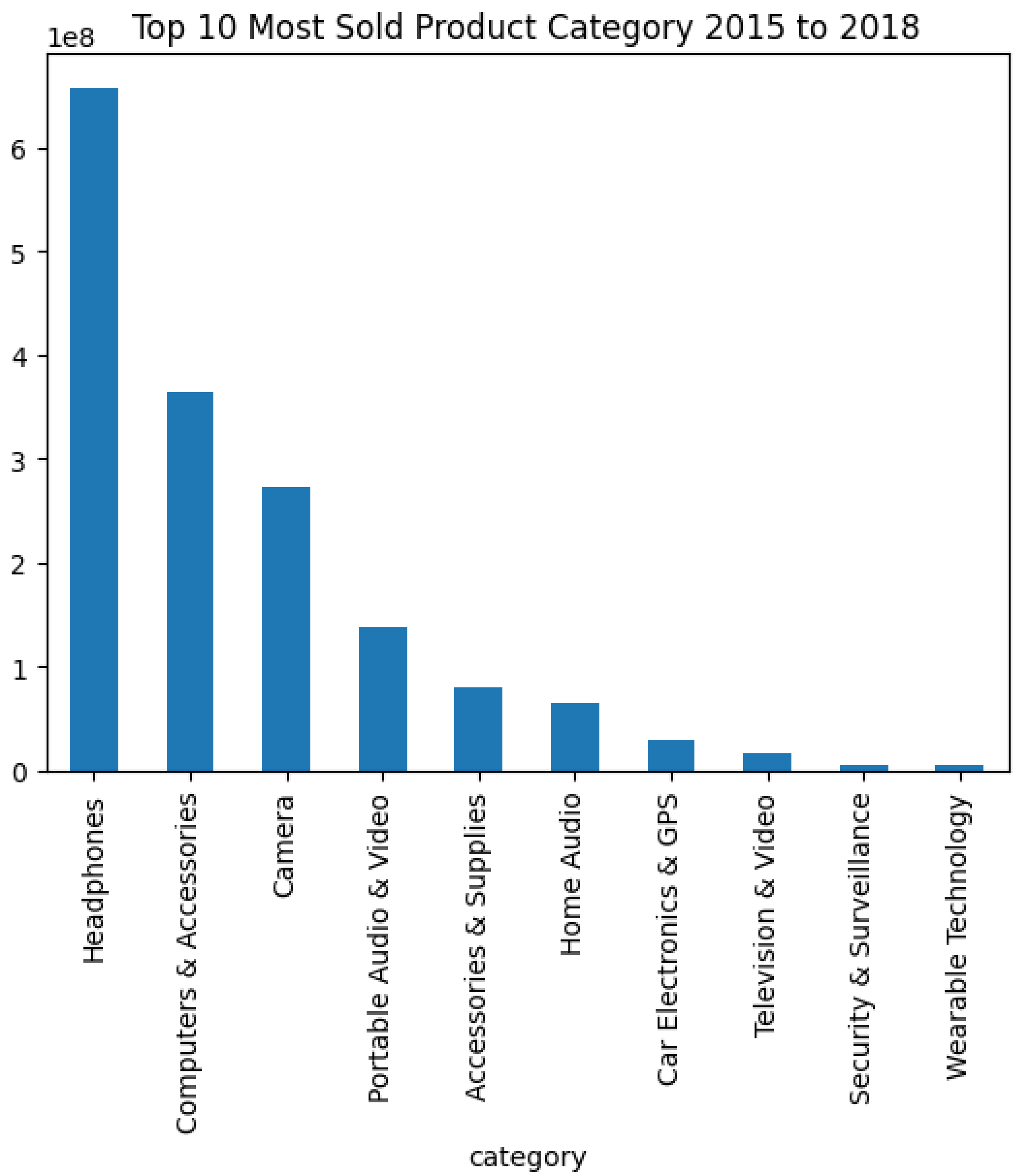


Image: Product by Category that sold the most

Inference:
We can see that the category of Headphones sold the most, computers and accessories were sold the second most while cameras sold the third most .

vi.)What product by category sold the least between 2015 to 2018?

```
# What product by brand name sold the least between 2015 to 2018?
dataset2015_2018 = dataset[(dataset['year'] >= 2015) & (dataset['year']
<= 2018)]
dataset2015_2018.groupby('category')['amount'].sum().sort_values(ascendi
ng=True).head(10).plot(kind='bar',title='10 Least  Sold Product Brand
2015 to 2018')
```

Output:

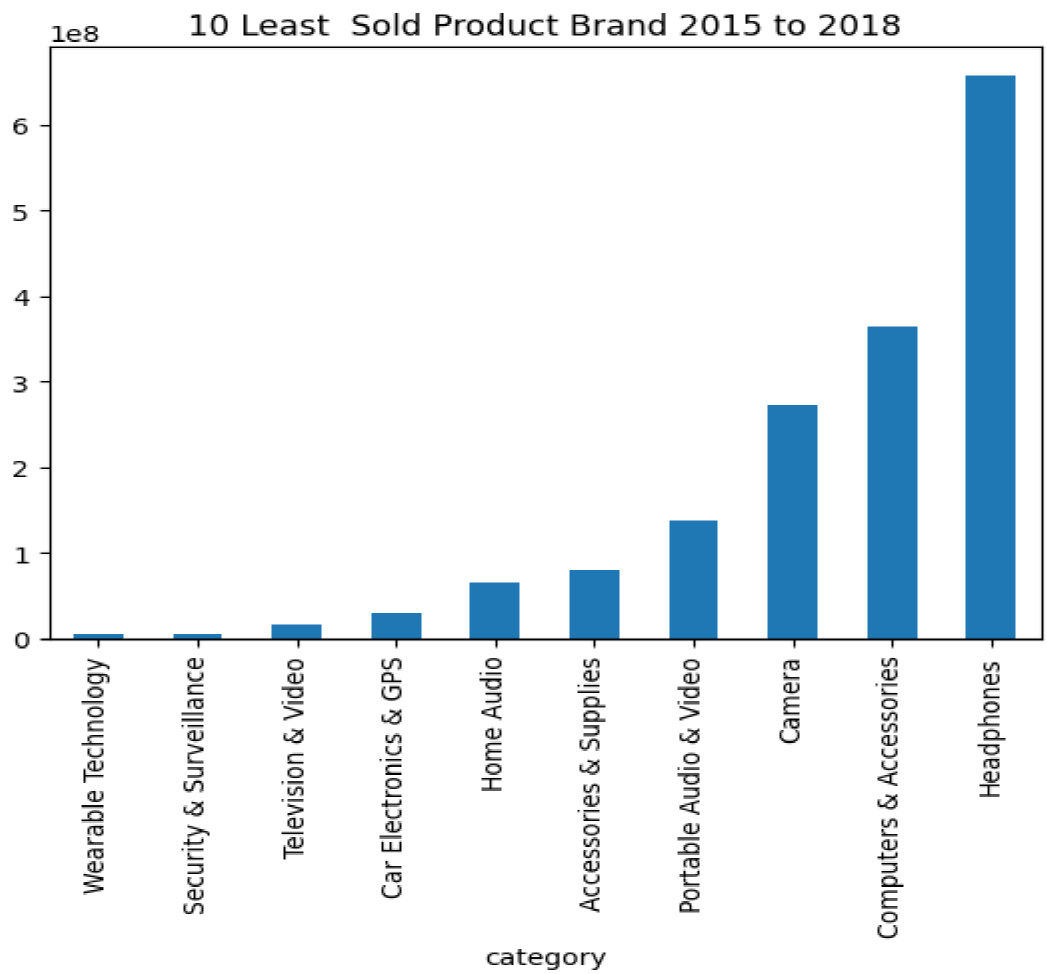


Image: Product by Category that sold the least

Inference:
We can see that the category of Wearable Technology sold the least followed closely by Security and Surveillance.

vii.) What product by brand name sold the least between 2015 to 2018?

```
# What product by brand name sold the least between 2015 to 2018?
dataset2015_2018 = dataset[(dataset['year'] >= 2015) & (dataset['year']
<= 2018)]
dataset2015_2018.groupby('brand')['amount'].sum().sort_values(ascending=
True).head(10).plot(kind='bar',title='10 Least  Sold Product Brand 2015
to 2018')
```

Output:

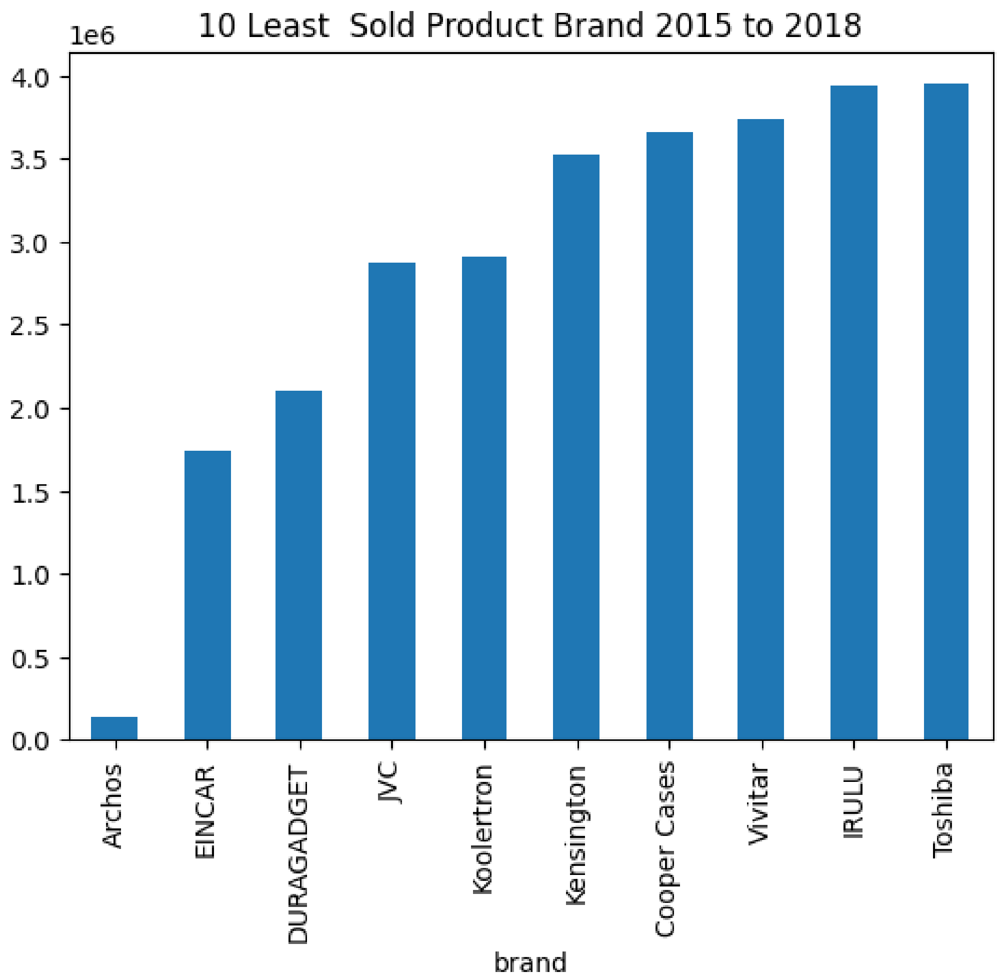


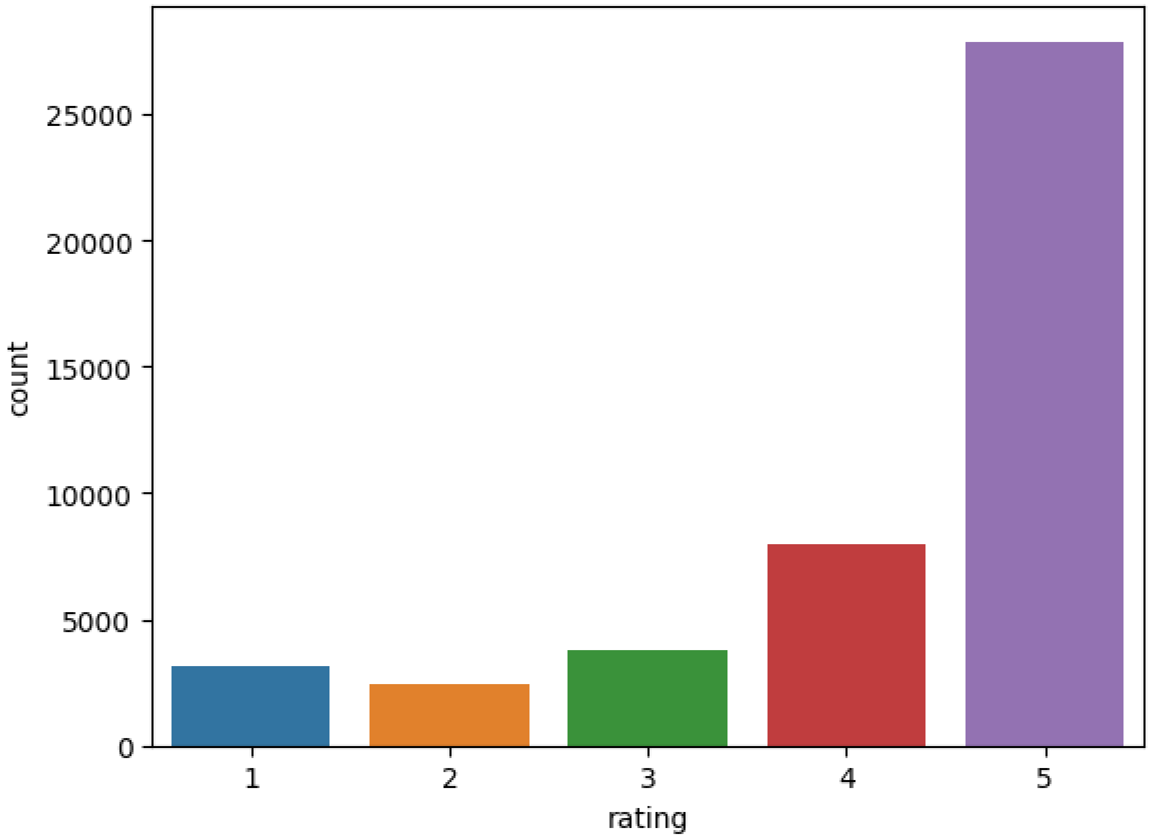
Image: Product by brand name sold the least

Inference:
Archos sold the least followed closely with EINCAR.

viii.) Ratings Distribution

```
# # the distribution of ratings

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

Output:



Inference:
Most Products were rated 5

ix.) Best rated brands

```
# What is the most rated brand name  between 2015 to 2018?
dataset2015_2018 = dataset[(dataset['year'] >= 2015) & (dataset['year']
<= 2018)]
dataset2015_2018.groupby('brand')['rating'].mean().sort_values(ascending
=False).head(10).plot(kind='bar',title='10 most rating Brand 2015 to
2018')
```

Output:

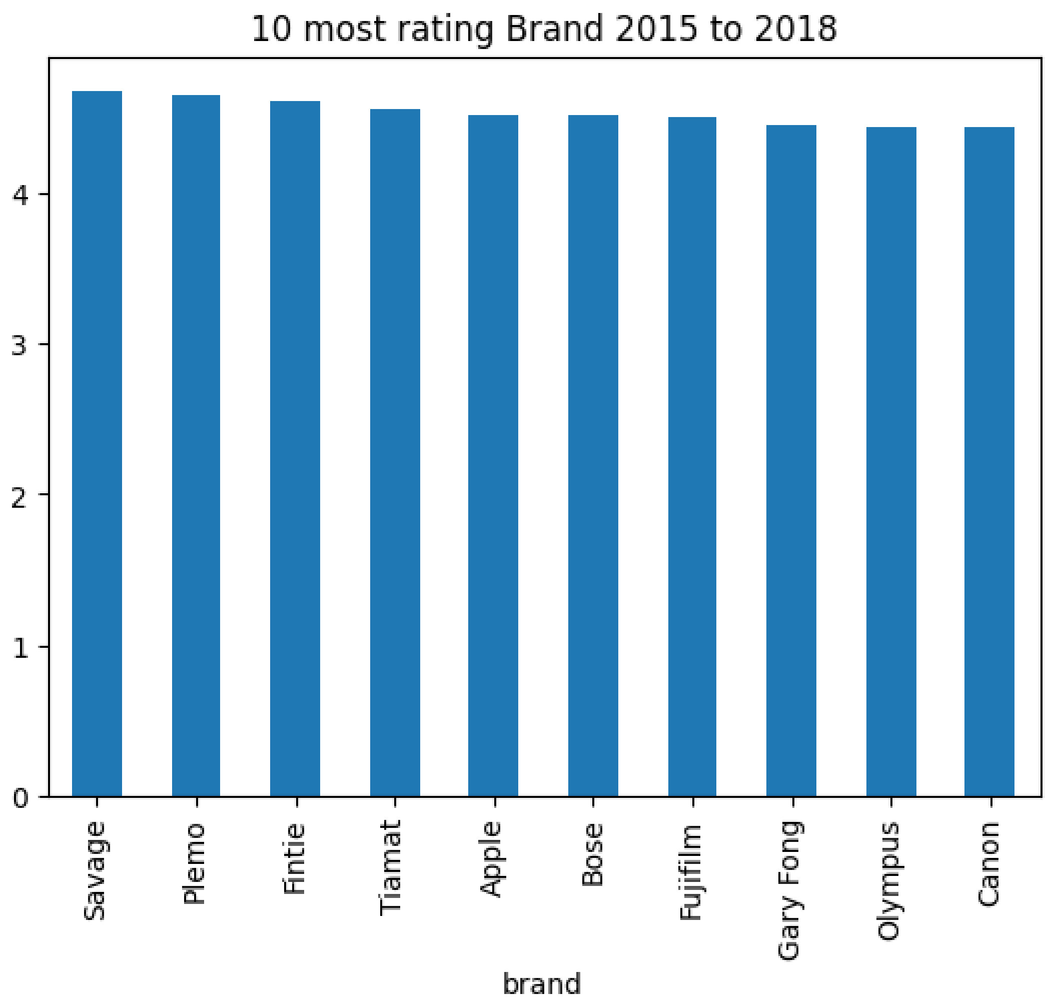


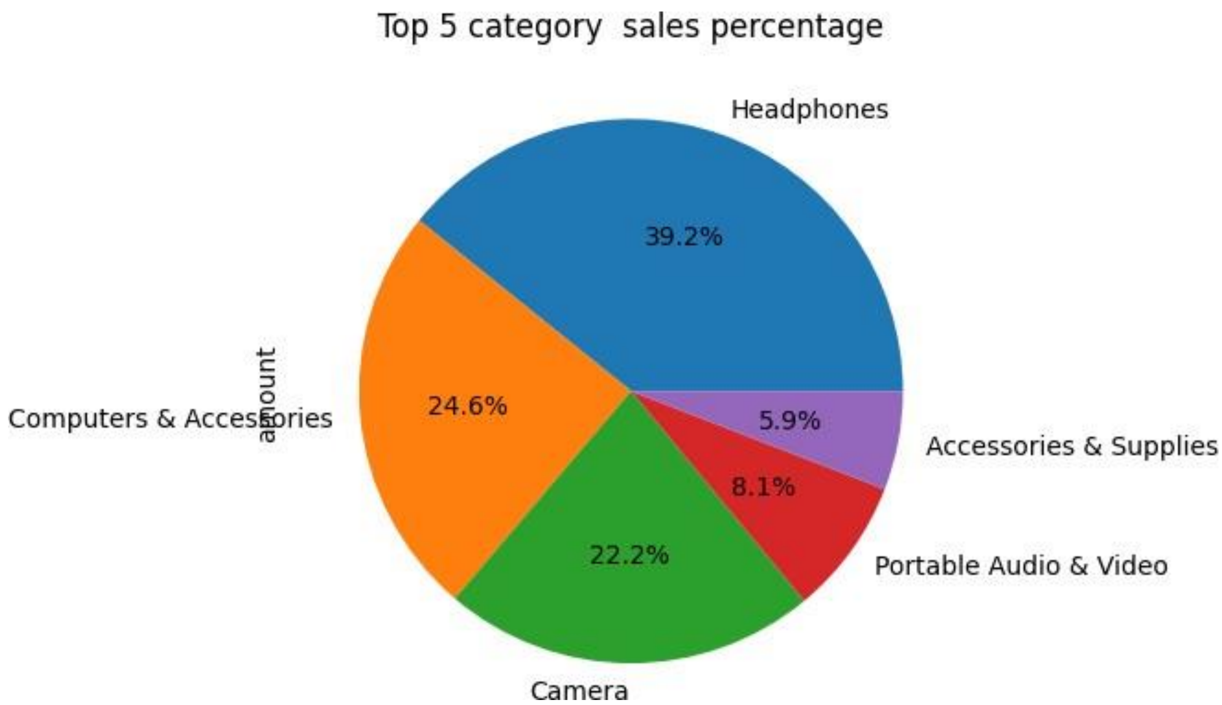
Image: Best brands by rating

Inference:
Savage and Plemo were the brands with the highest ratings.

x) Top 5 category sales percentage

```
# category percentage sales

dataset.groupby('category')['amount'].sum().sort_values(ascending=False)
.head(5).plot(kind='pie', autopct='%1.1f%%',title='Top 5 category  sales
percentage')
```

Output:

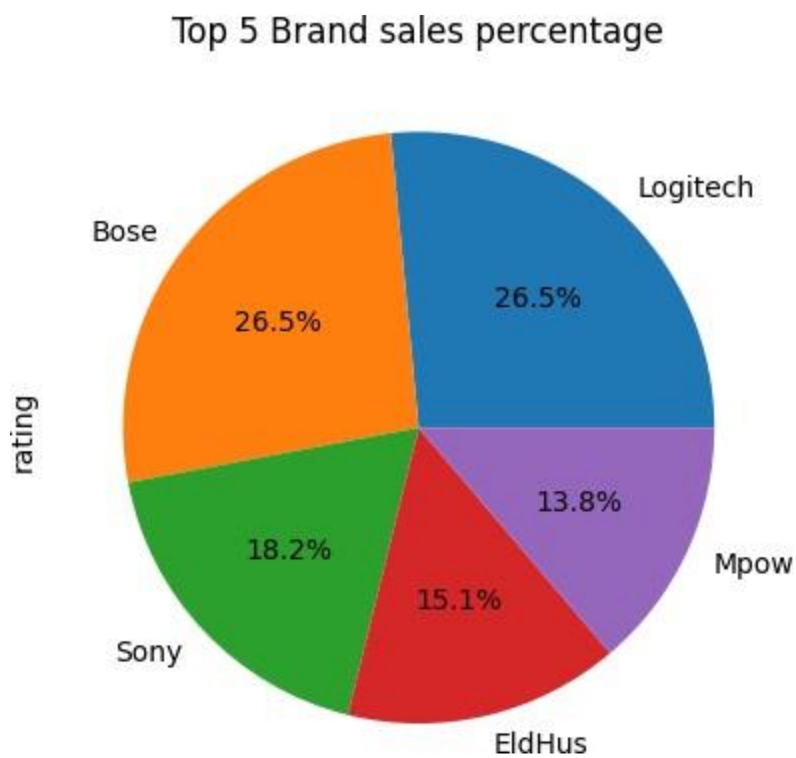


Inference:
Headphones sales % is the highest followed by Computers & Accessories.

xi) Brand wise sales percentage

```
# brand wise sales percentage

dataset.groupby('brand')['rating'].count().sort_values(ascending=False).
head(5).plot(kind='pie', autopct='%1.1f%%',title='Top 5 Brand wise sales
percentage')
```

Output:

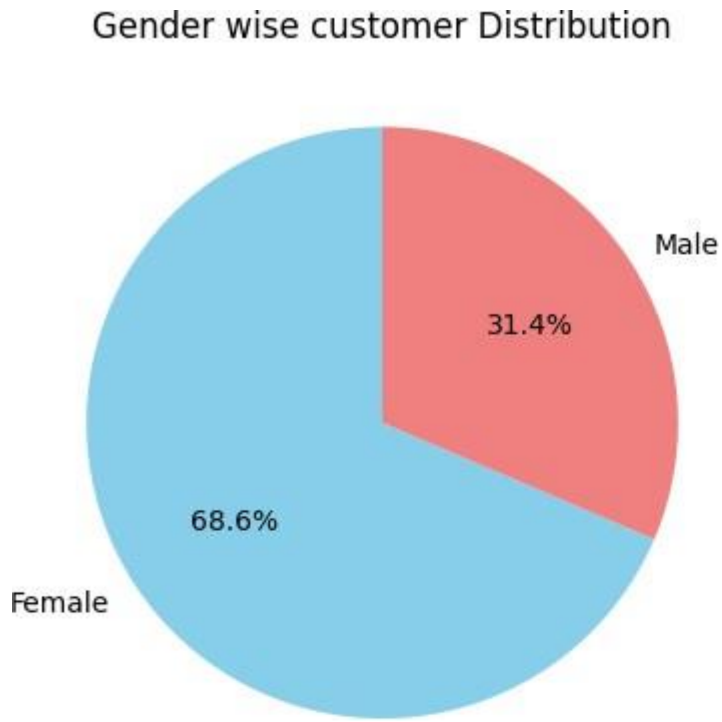


Inference:
Bose and Logitech sales % is the highest followed by Sony.

xii) Gender wise customer distribution

```
# Gender wise customer distribution
gender_distribution = dataset['gender'].value_counts()
plt.pie(gender_distribution, labels=gender_distribution.index,
autopct='%1.1f%%', startangle=90, colors=['skyblue', 'lightcoral'])
plt.title('Gender wise customer Distribution')
plt.show()
```

Output:



Inference:
Most of the customers are in Female categories.

Conclusion:

- 2015 was the best year in terms of sales and profit
- Headphones was the category with most sales followed closely with Computer and Accessories while the least sales were made in the Category Security & Surveillance.
- There has been a steady rise in sales from 2007 to 2015 and a sharp decline from 2016 to 2018.
- The brand name Bose sold the most followed by Logitech.
- The brand Archos sold the least followed closely with EINCAR..
- Most products were rated 5.
- Best rated brands were Savage and Plemo.

The above analysis should help you to understand and explore further on the reasons behind the popularity and/or poor sales of the products. With this foresight a company can make decisions whether to continue production/sales of a specific product for the future.