# MINI PROJECT

## 1. PROBLEM STATEMENT

Designing a mini project focused on a superstore sales dataset, where aim to perform data preprocessing and visualization tasks. The project will involve cleaning and transforming the data, handling missing values and outliers, and then creating insightful visualizations to uncover patterns, trends, and anomalies in the sales data. The goal is to provide actionable insights for improving the store's performance and optimizing sales strategies. The project will require Python programming skills, data analysis libraries (e.g., pandas, matplotlib, and seaborn ) and will culminate in a comprehensive report and presentation summarizing the findings and recommendations.

## 2. DATA PREPROCESSING

Superstore sale dataset preprocessing involves the following steps:

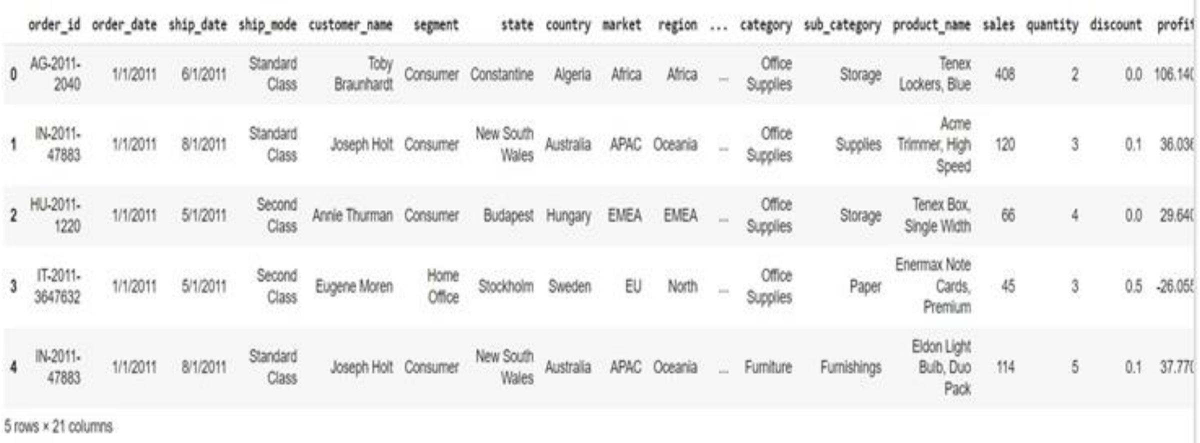
1. Data Cleaning: Removing duplicates, handling missing values, and correcting any erroneous entries to ensure data integrity.

2. Data Transformation: Standardizing data types, converting categorical variables into numerical representations (encoding), and scaling features if needed.
3. Feature Selection: Identifying and selecting relevant features to reduce dimensionality and improve model performance.
4. Outlier Detection: Identifying and handling outliers that may skew the analysis or modeling results.
5. Data Splitting: Dividing the dataset into training, validation, and test sets to evaluate and validate models effectively.

## PROGRAM:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df=pd.read_csv('/content/SuperStoreOrders.csv')
df.head()
```



| | order_id | order_date | ship_date | ship_mode | customer_name | segment | state | country | market | region | ... | category | sub_category | product_name | sales | quantity | discount | profit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | AG-2011-2040 | 1/1/2011 | 6/1/2011 | Standard Class | Toby Braunhardt | Consumer | Constantine | Algeria | Africa | Africa | ... | Office Supplies | Storage | Tenex Lockers, Blue | 408 | 2 | 0.0 | 106.140 |
| 1 | IN-2011-47883 | 1/1/2011 | 8/1/2011 | Standard Class | Joseph Holt | Consumer | New South Wales | Australia | APAC | Oceania | ... | Office Supplies | Supplies | Acme Trimmer, High Speed | 120 | 3 | 0.1 | 36.036 |
| 2 | HU-2011-1220 | 1/1/2011 | 5/1/2011 | Second Class | Annie Thurman | Consumer | Budapest | Hungary | EMEA | EMEA | ... | Office Supplies | Storage | Tenex Box, Single Width | 66 | 4 | 0.0 | 29.640 |
| 3 | IT-2011-3647632 | 1/1/2011 | 5/1/2011 | Second Class | Eugene Moren | Home Office | Stockholm | Sweden | EU | North | ... | Office Supplies | Paper | Enermax Note Cards, Premium | 45 | 3 | 0.5 | -26.055 |
| 4 | IN-2011-47883 | 1/1/2011 | 8/1/2011 | Standard Class | Joseph Holt | Consumer | New South Wales | Australia | APAC | Oceania | ... | Furniture | Furnishings | Eldon Light Bulb, Duo Pack | 114 | 5 | 0.1 | 37.770 |

5 rows × 21 columns

```
df.shape
```

```
(19911, 21)
```

```
df.columns
```

```
Index(['order_id', 'order_date', 'ship_date',
'ship_mode', 'customer_name', 'segment',
'state', 'country', 'market', 'region',
'product_id', 'category', 'sub_category',
'product_name', 'sales', 'quantity',
'discount', 'profit', 'shipping_cost',
'order_priority', 'year'], dtype='object')
```

```
df.dtypes
```

```
vorder_id          object
order_date        object
ship_date         object
ship_mode         object
customer_name     object
segment           object
state             object
country           object
market            object
region            object
product_id        object
category          object
sub_category      object
```

```
product_name        object
sales               object
quantity           float64
discount           float64
profit             float64
shipping_cost      float64
order_priority      object
year               float64
dtype: object
```

```
df.isnull().sum()
```

```
order_id           0
order_date         0
ship_date          0
ship_mode          0
customer_name      0
segment            0
state              0
country            0
market             0
region             0
product_id         0
category           1
sub_category       1
product_name       1
sales              1
quantity           1
```

```
discount          1
profit            1
shipping_cost     1
order_priority    1
year              1
dtype: int64
```

```
df['country'].value_counts()
```

```
United States    4084
Mexico           1057
Australia        1050
France           1032
Germany           868
                 ...
Tajikistan          1
Macedonia           1
Mauritania          1
South Sudan         1
Sri Lanka           1
Name: country, Length: 141, dtype: int64
```

```
df['category'].unique()
```

```
array(['Office Supplies', 'Furniture', 'Technology',
nan], dtype=object)
```

```
df['category'].value_counts()
```

```
Office Supplies    12115
Technology          3994
Furniture           3801
Name: category, dtype: int64
```

```
df['sub_category'].nunique()
```

```
17
```

```
df['sub_category'].value_counts()
```

```
Binders        2337
Storage        1989
Art            1907
Paper          1364
Phones         1340
Chairs         1303
Furnishings    1236
Accessories    1201
Labels         1018
Supplies        933
Envelopes       931
Fasteners       929
Bookcases       927
Copiers         855
Appliances      707
Machines        598
```

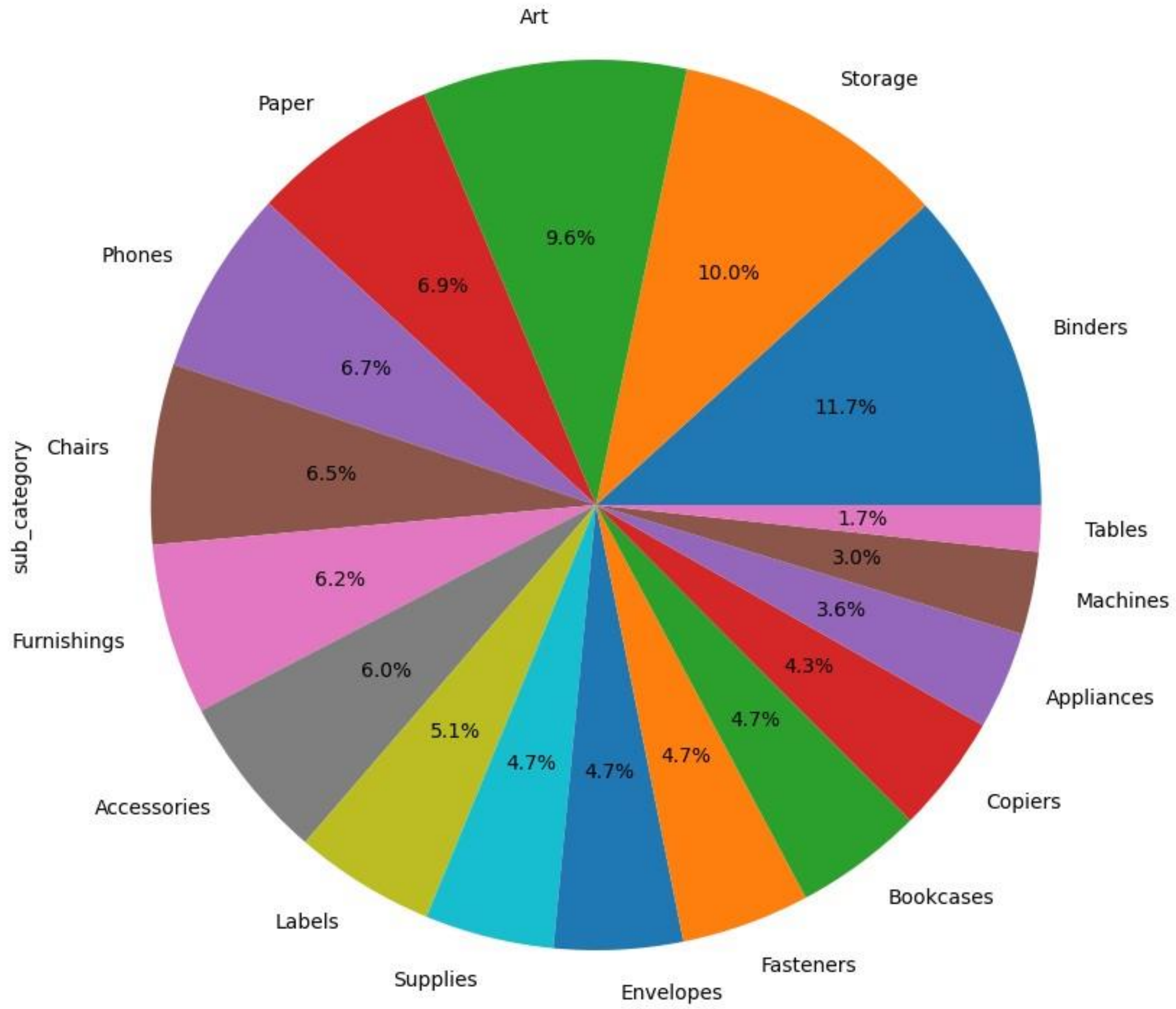
```
Tables           335
Name: sub_category, dtype: int64
```

## 3. DATA VISUALIZATION

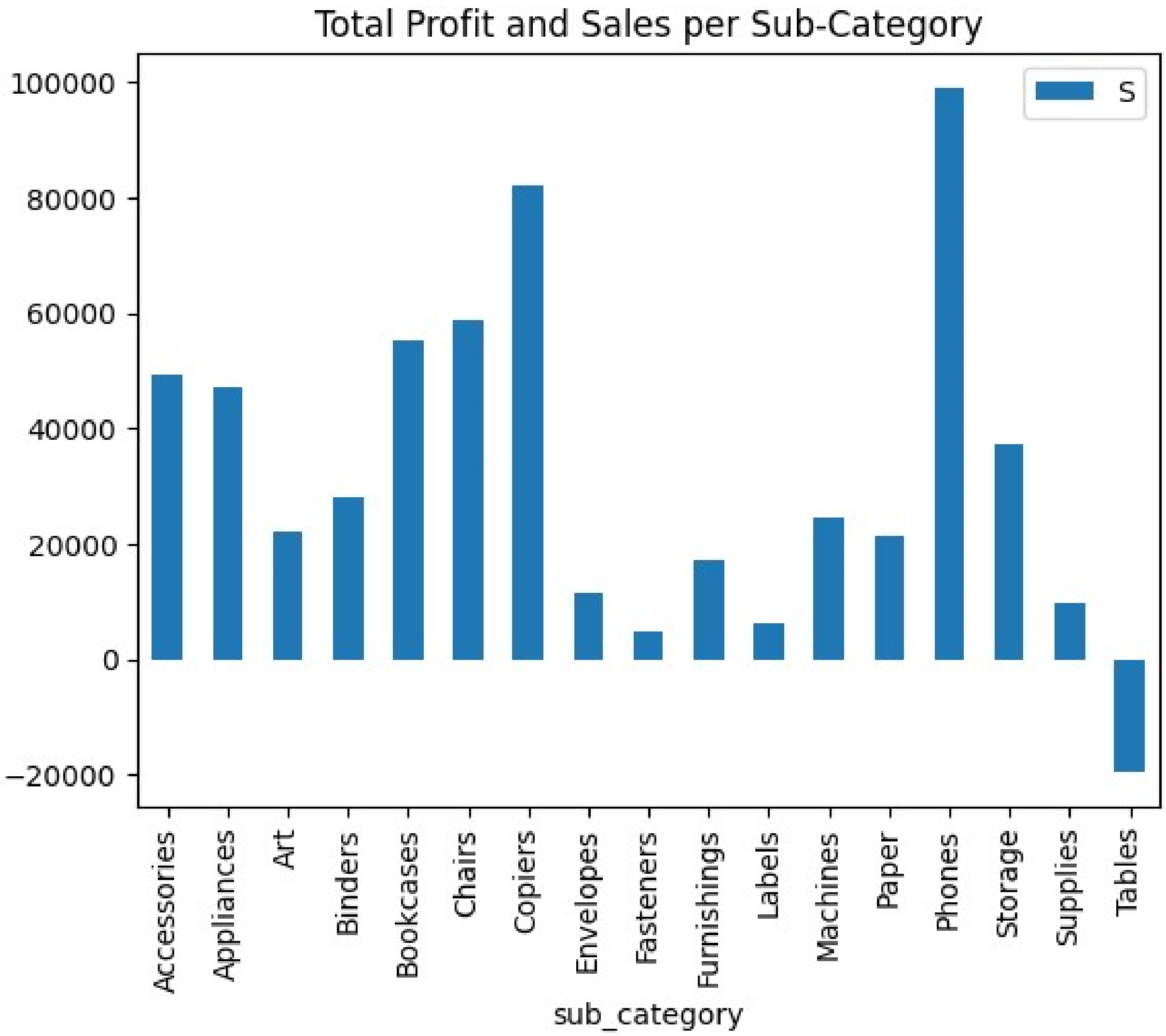
A superstore sale dataset visualization is a graphical representation of sales data from a large retail store. It typically uses charts, graphs, or other visual elements to illustrate key sales metrics, such as revenue, product categories, and trends over time. These visualizations help businesses and analysts gain insights into sales performance, identify patterns, and make informed decisions for inventory management, marketing strategies, and more. They provide a clear and concise way to communicate complex sales information, aiding in data-driven decision-making and strategic planning.

### PROGRAM:

```
plt.figure(figsize=(12,10))
df['sub_category'].value_counts().plot.pie(autopct="%1.1f%%")
plt.show()
```



```
df.groupby('sub_category')['profit','sales'].agg(['sum']).plot.bar()
plt.title('Total Profit and Sales per Sub-Category')
plt.legend('Profit')
plt.legend('Sales')
plt.show()
```

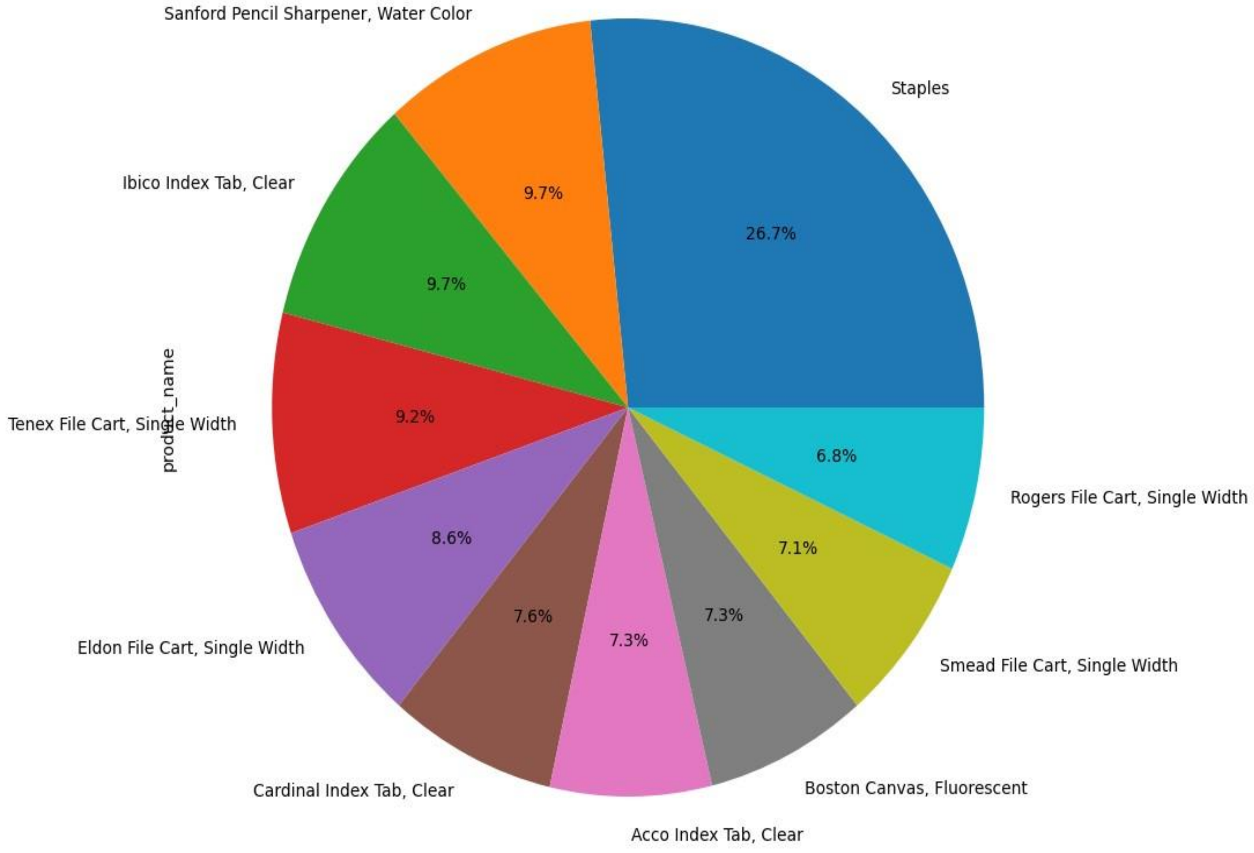


```
df['product_name'].nunique()
```

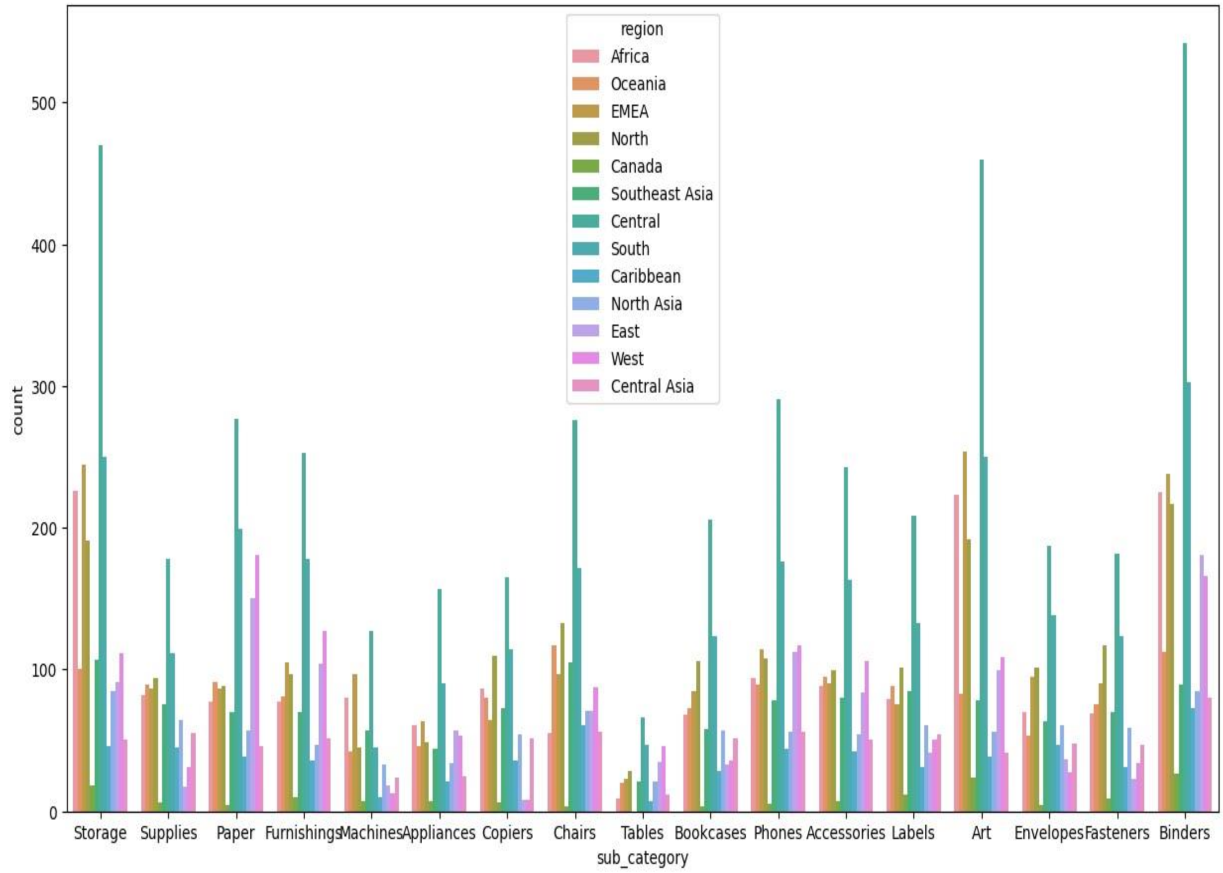
```
3532
```

plt.figure(figsize=(12,10))

df['product_name'].value_counts().head(10).plot.pie(autopct="% 1.1f%%")

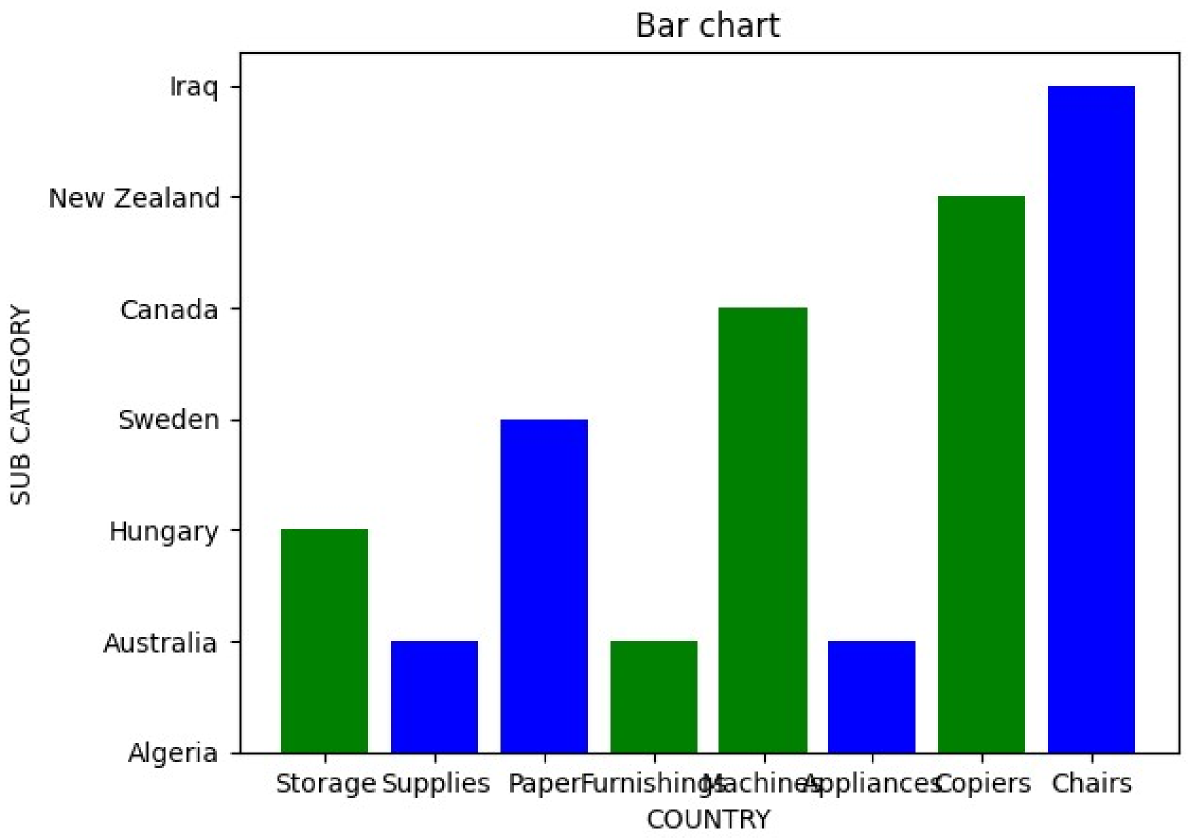


```
plt.figure(figsize=(15,8))
sns.countplot(x="sub_category", hue="region",
data=df)
plt.show()
```

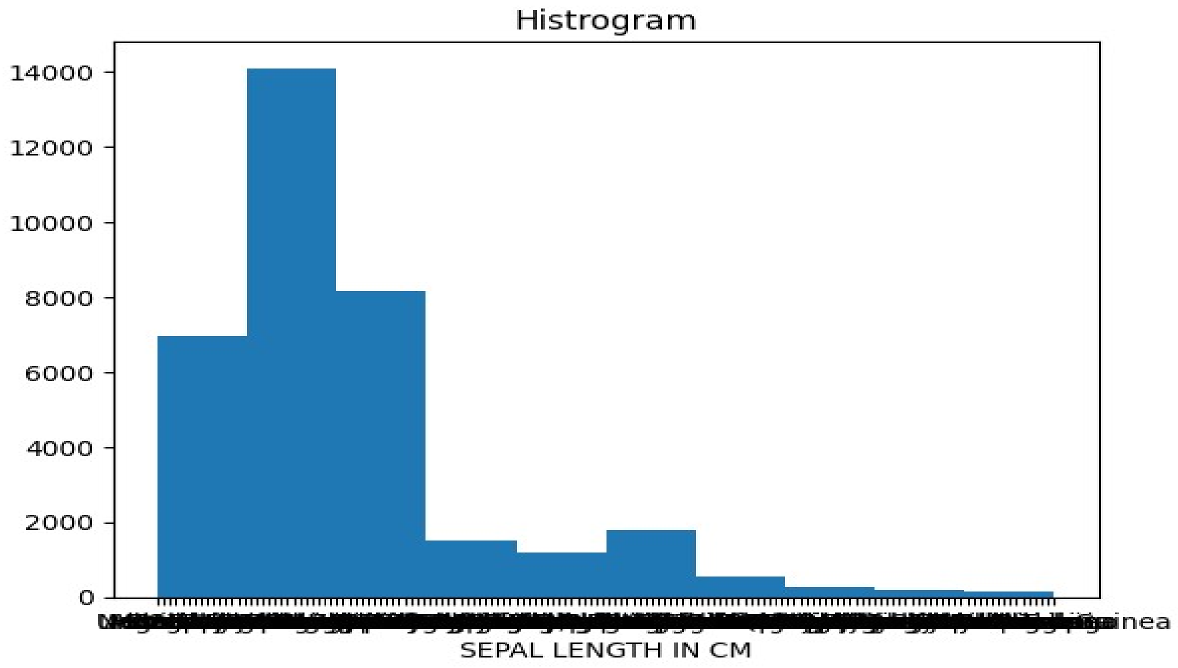


```
x=df['sub_category'].head(10)
```

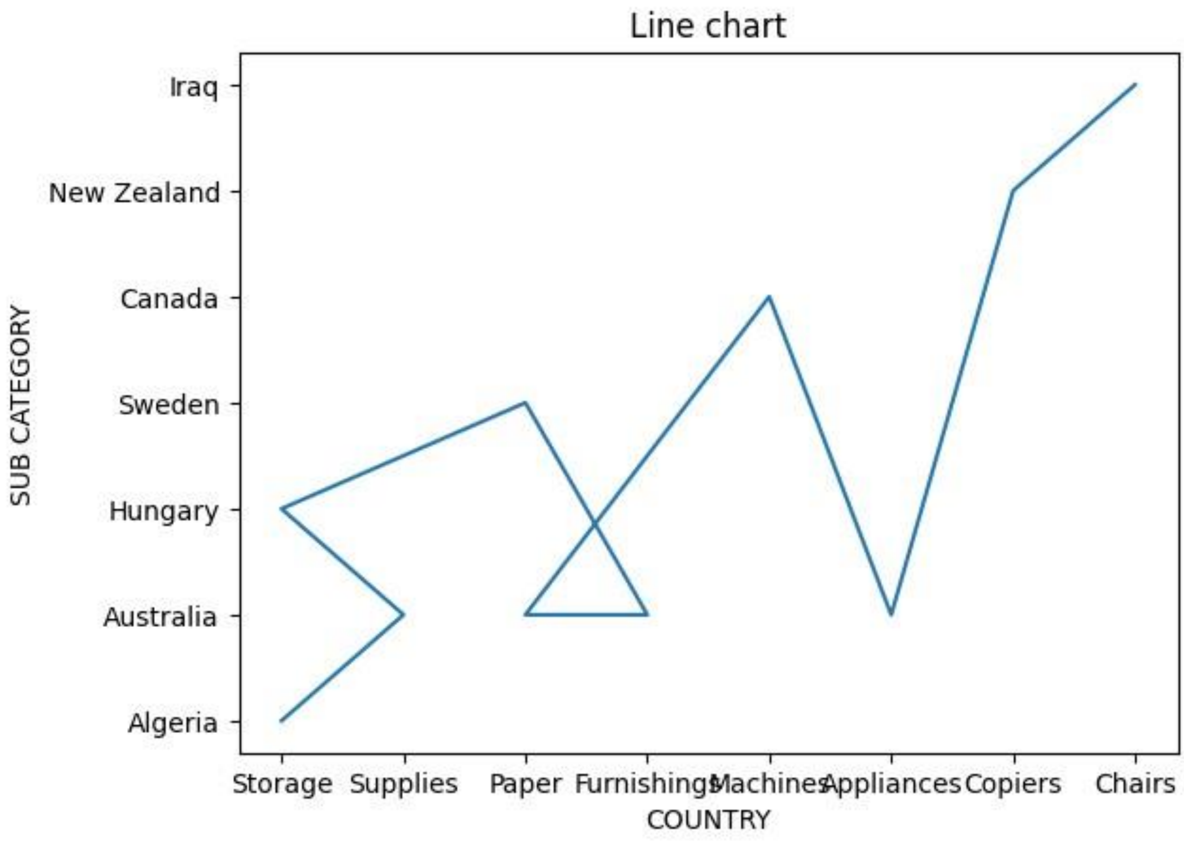
```
y=df['country'].head(10)


plt.bar(df['country'],df['sub_category'],color=
['green','blue'])
plt.title("Bar chart")
plt.xlabel('COUNTRY')
plt.ylabel('SUB CATEGORY')
plt.show()
```



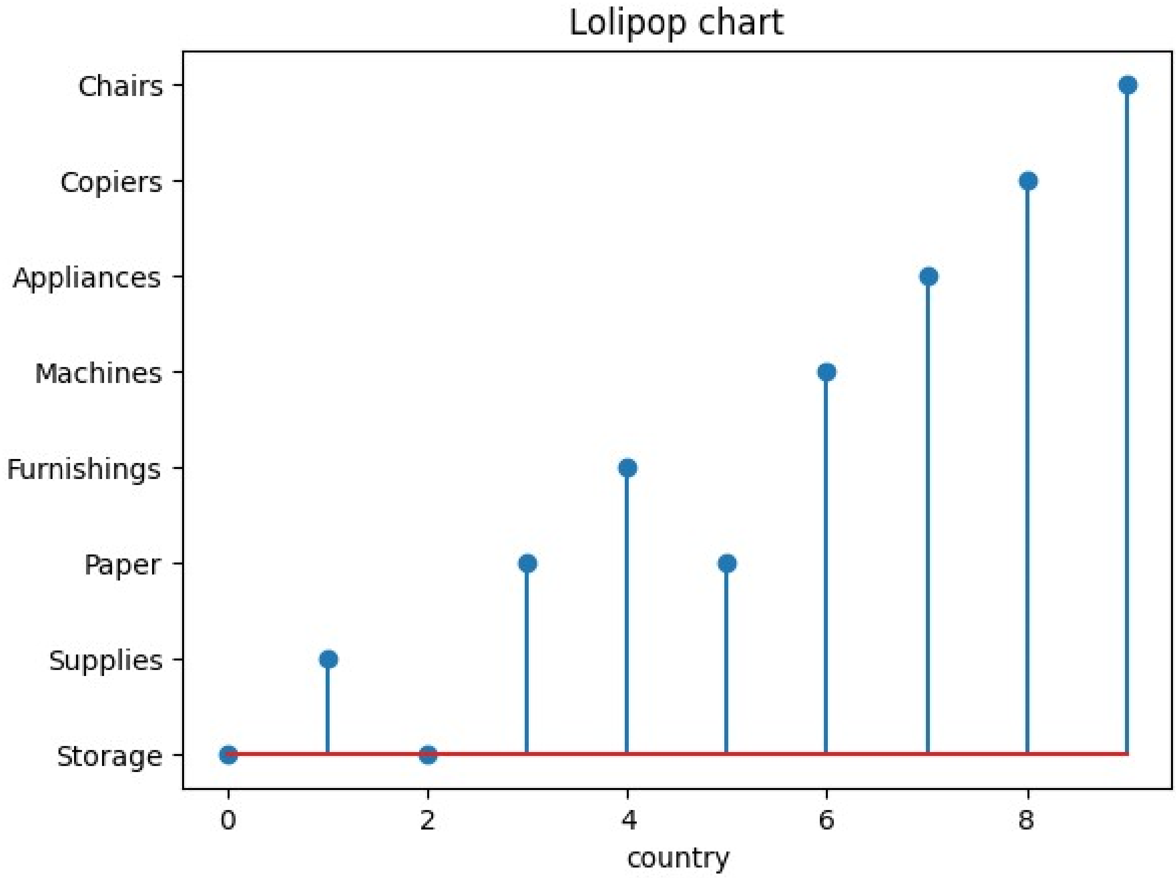
```
plt.hist(df['country'])
plt.title("Histrogram")
plt.xlabel('country')
plt.show()
```



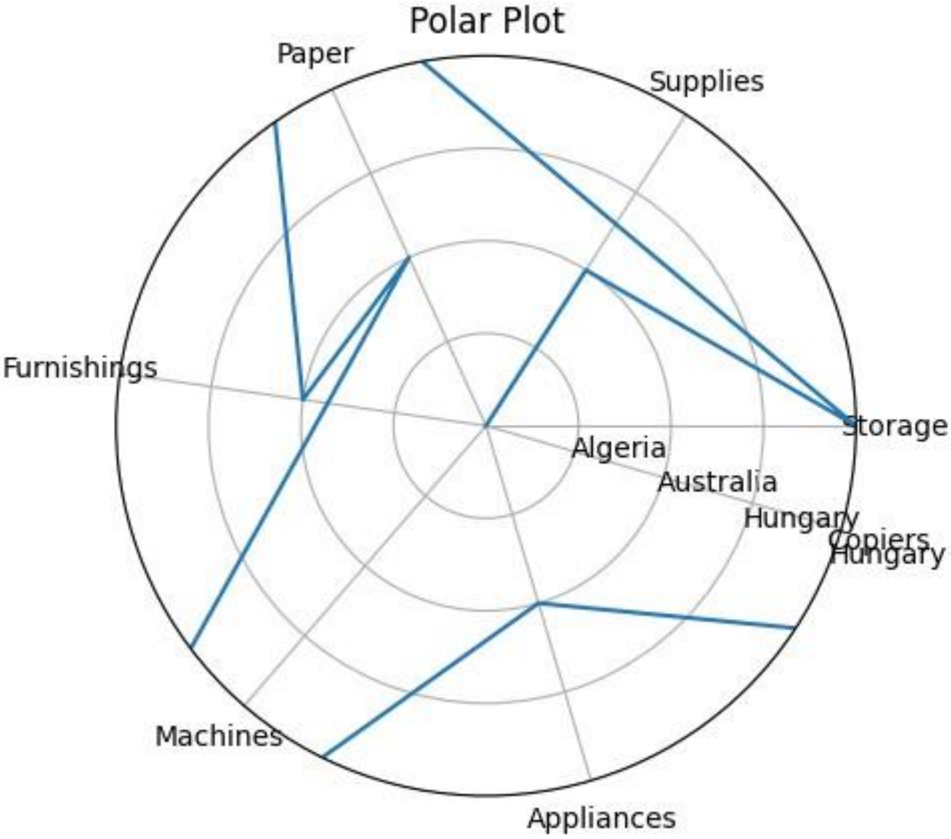
```
plt.plot(x,y)
plt.title("Line chart")
plt.xlabel('COUNTRY')
plt.ylabel('SUB CATEGORY')
plt.show()
```



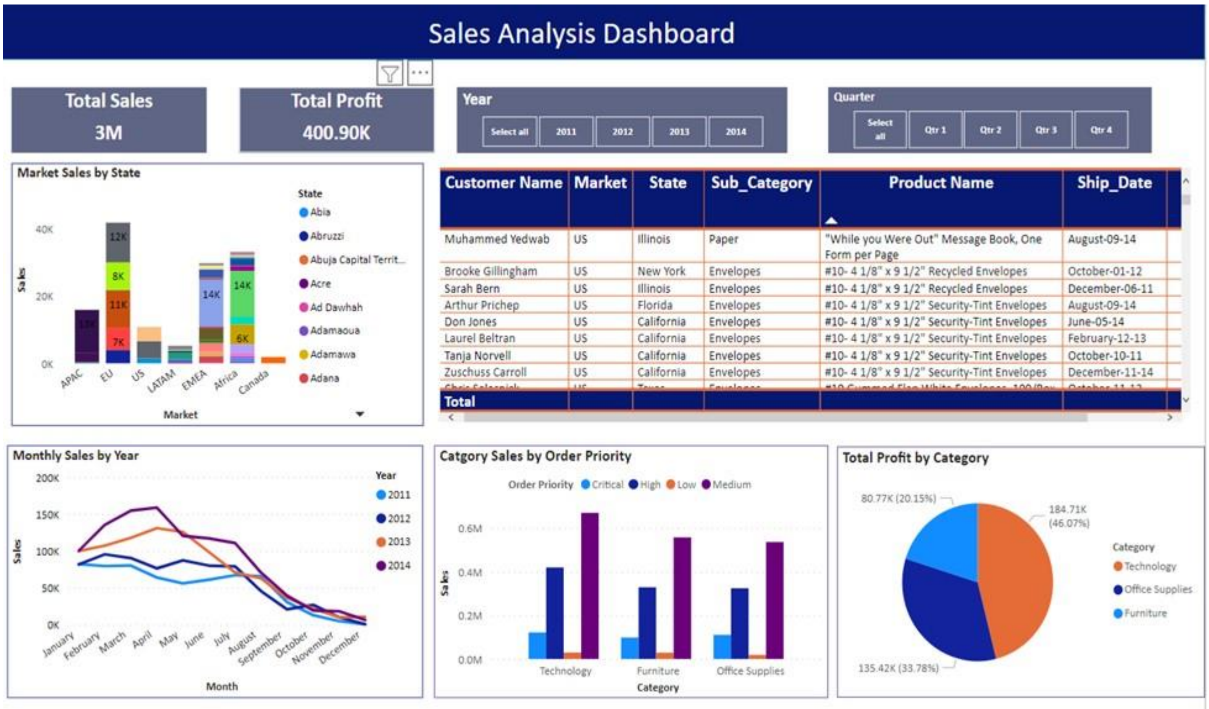
```
plt.stem(x)
plt.title("Lolipop chart")
plt.xlabel('country')
plt.show()
```



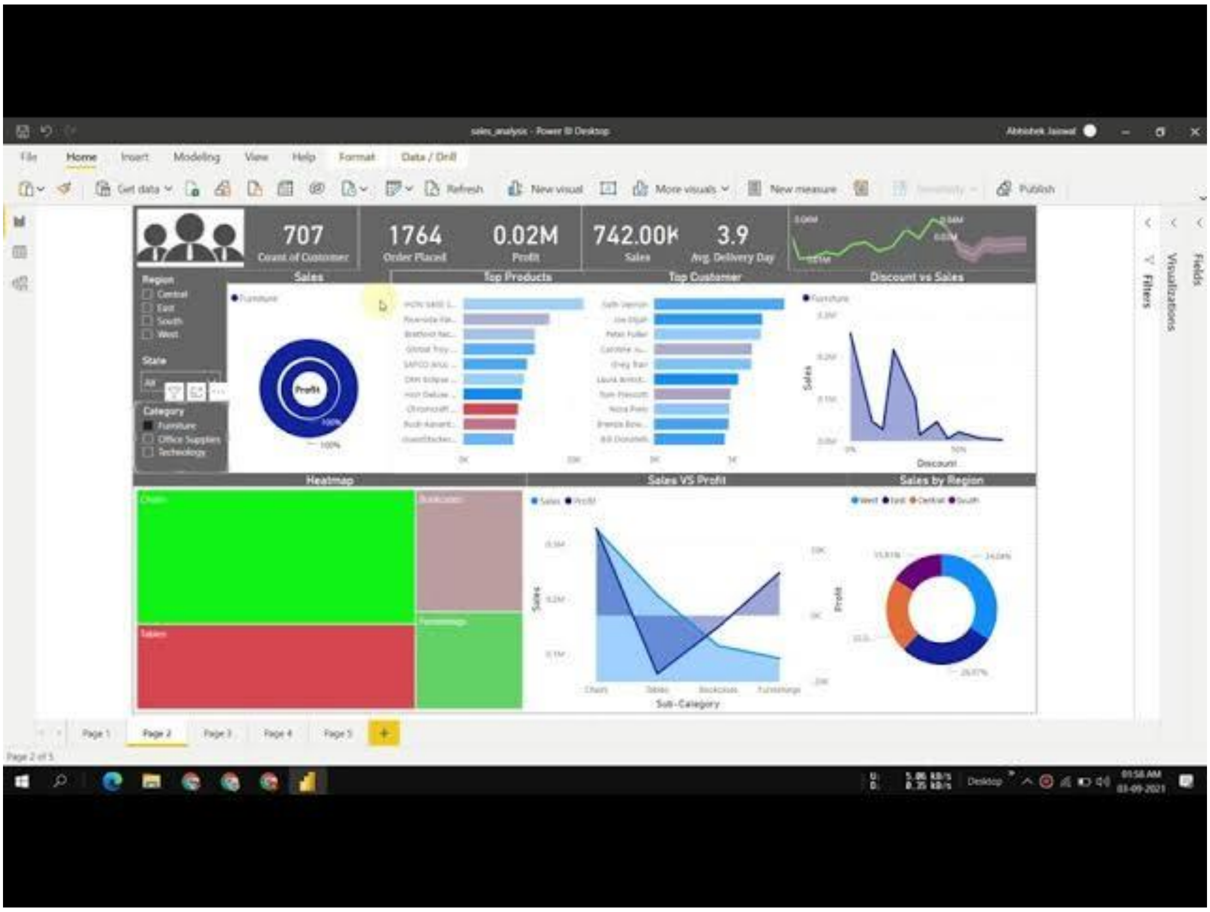
```
fig, ax = plt.subplots(subplot_kw={'projection':
'polar'})
ax.plot(x, y)
ax.set_rmax(2)
ax.set_rticks([0.5, 1, 1.5, 2])
ax.set_rlabel_position(-22.5)
ax.grid(True)
ax.set_title("Polar Plot", va='bottom')
plt.show()
```

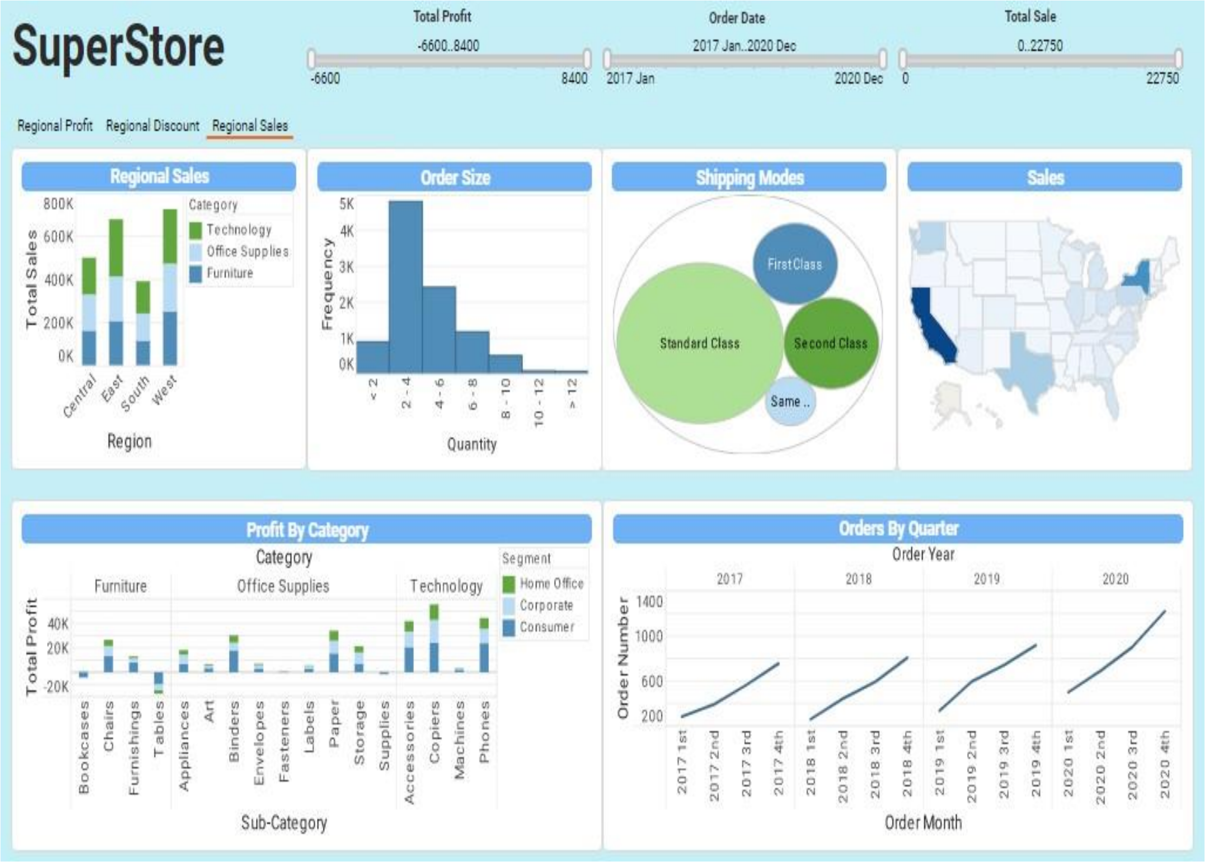

Polar Plot
Paper
Supplies
Furnishings
Storage
Algeria
Australia
Hungary
Copiers
Hungary
Machines
Appliances

# DASHBOARD



| Customer Name | Market | State | Sub_Category | Product Name | Ship_Date |
|---|---|---|---|---|---|
| Muhammed Yedwab | US | Illinois | Paper | "While you Were Out" Message Book, One Form per Page | August-09-14 |
| Brooke Gillingham | US | New York | Envelopes | #10- 4 1/8" x 9 1/2" Recycled Envelopes | October-01-12 |
| Sarah Bern | US | Illinois | Envelopes | #10- 4 1/8" x 9 1/2" Recycled Envelopes | December-06-11 |
| Arthur Prichep | US | Florida | Envelopes | #10- 4 1/8" x 9 1/2" Security-Tint Envelopes | August-09-14 |
| Don Jones | US | California | Envelopes | #10- 4 1/8" x 9 1/2" Security-Tint Envelopes | June-05-14 |
| Laurel Beltran | US | California | Envelopes | #10- 4 1/8" x 9 1/2" Security-Tint Envelopes | February-12-13 |
| Tanja Norvell | US | California | Envelopes | #10- 4 1/8" x 9 1/2" Security-Tint Envelopes | October-10-11 |
| Zuschuss Carroll | US | California | Envelopes | #10- 4 1/8" x 9 1/2" Security-Tint Envelopes | December-11-14 |
| Total | | | | | |




SuperStore
Total Profit
-6600..8400
-6600
8400
Order Date
2017 Jan..2020 Dec
2017 Jan
2020 Dec
Total Sale
0..22750
0
22750
Regional Profit
Regional Discount
Regional Sales
Regional Sales
800K
600K
400K
200K
0K
Total Sales
Central
East
South
West
Region
Category
Technology
Office Supplies
Furniture
Order Size
5K
4K
3K
2K
1K
0K
Frequency
< 2
2 - 4
4 - 6
6 - 8
8 - 10
10 - 12
> 12
Quantity
Shipping Modes
First Class
Standard Class
Second Class
Same ..
Sales
Profit By Category
Category
Furniture
Office Supplies
Technology
Segment
Home Office
Corporate
Consumer
40K
20K
-20K
Total Profit
Bookcases
Chairs
Furnishings
Tables
Appliances
Art
Binders
Envelopes
Fasteners
Labels
Paper
Storage
Supplies
Accessories
Copiers
Machines
Phones
Sub-Category
Orders By Quarter
Order Year
2017
2018
2019
2020
1400
1000
600
200
Order Number
2017 1st
2017 2nd
2017 3rd
2017 4th
2018 1st
2018 2nd
2018 3rd
2018 4th
2019 1st
2019 2nd
2019 3rd
2019 4th
2020 1st
2020 2nd
2020 3rd
2020 4th
Order Month

Superstore Profitability Overview
Profit Ratio
12.5%
Total Profit
$286,397
Total Sales
$2,297,201

# REPORT