NAP QUEENS – ASSIGNMENT ROUND

Name : Abhilash Nagisetty

Registration Number: 20MID0201

ROLE APPLIED : Data Analyst

Data Analytics | Python

Data Source:

Use the provided sales data in spreadsheet format. The data contains information about sales transactions, including date, product, quantity, and revenue.

Data source file:

https://docs.google.com/spreadsheets/d/1KagwoQLy1quKvT_82amuS-x3UnsoIX4J6p02ewbjQNA/edit?usp=sharing

Perform basic data exploration and visualization using the provided dataset, "Global-Superstore." Develop your own data storytelling narrative based on the insights you uncover.

Context About the Dataset:

The dataset appears to be a sales transaction dataset from a global superstore. It contains detailed information about individual sales orders, including order dates, shipping dates, customer details, product details, sales figures, and other related metrics.

Key Columns in My Dataset

1. Order ID: This is the unique identifier for each order.
2. Order Date: This column tells me when the order was placed.
3. Ship Date: Here, I can see when the order was shipped.
4. Ship Mode: This indicates the mode of shipping (e.g., Same Day, Second Class, First Class, Standard Class).
5. Customer ID: Each customer has a unique identifier.
6. Customer Name: This is the name of the customer.
7. Segment: It shows the customer segment (e.g., Consumer, Corporate, Home Office).
8. City: The city where the customer is located.

9. State: The state where the customer is located.

10. Country: The country where the customer is located.

11. Postal Code: The postal code of the customer's location.

12. Market: This represents the market region (e.g., US, APAC, EU, Africa, LATAM).

13. Region: This indicates the geographical region (e.g., East, Central, Oceania, Africa, North Asia, etc.).

14. Product ID: Each product has a unique identifier.

15. Category: The category of the product (e.g., Technology, Furniture, Office Supplies).

16. Sub-Category: The sub-category of the product (e.g., Accessories, Phones, Chairs, Tables).

17. Product Name: The name of the product.

18. Sales: The total sales amount for the order.

19. Quantity: The quantity of the product ordered.

20. Discount: Any discount applied to the order.

21. Profit: The profit made from the order.

22. Shipping Cost: The cost of shipping the order.

23. Order Priority: The priority of the order (e.g., Critical, High, Medium, Low).

Potential Insights from My Dataset

1. Sales Performance:
    - I can analyze total sales and profit by different dimensions such as product category, customer segment, and region.
    - It will be interesting to identify the best and worst-performing products in terms of sales and profit.
2. Customer Analysis:
    - I can segment customers based on their purchase behaviour.
    - Identifying top customers and their purchasing patterns could provide valuable insights.
    - Analyzing customer distribution across different geographical locations will be insightful.
3. Shipping Analysis:

   - I can evaluate shipping performance based on different shipping modes.
   - Analyzing the relationship between shipping cost and profit will be valuable.

4. Order Priority:
   - Counting and analyzing the distribution of order priorities will give a good overview.
   - I can determine if there's any correlation between order priority and other factors such as profit, sales, and shipping time.

5. Discount Analysis:
   - Analyzing the impact of discounts on sales and profit will be crucial.
   - Identifying if higher discounts lead to increased sales volumes but decreased profit margins will be insightful.

Data Exploration and Visualization

1. Order Priority Distribution:
   - I can create a count plot of orders by order priority to understand the distribution of order priorities.

2. Sales by Category:
   - A bar chart will help me visualize total sales for each product category.

3. Profit by Region:
   - I can use a map or a bar chart to visualize profit distribution across different regions.

4. Sales and Profit over Time:
   - A line chart will help me analyze trends in sales and profit over time.

5. Customer Segmentation:
   - Pie charts or bar charts can be used to analyze the distribution of different customer segments.

By performing these analyses, I can uncover valuable insights that can help me make informed business decisions, improve sales strategies, optimize shipping methods, and enhance customer satisfaction.

DATA EXPLORATION USING PYTHON

Notebook – https://drive.google.com/file/d/16BVLXA-0mPtm_odFGm-TXRoXcWP8-xZz/view?usp=sharing

```
In [11]: import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns

         # Load the data
         file_path = 'C:/Users/abhis/Downloads/Global-Superstore(1).csv'
         data = pd.read_csv(file_path)
```

```
In [12]: # Basic data exploration
         print("First few rows of the dataset:")
         print(data.head())
```

```
First few rows of the dataset:
   Row ID         Order ID  Order Date   Ship Date       Ship Mode Customer
ID  \
0   32298   CA-2012-124891   7/31/2012   7/31/2012        Same Day    RH-194
95
1   26341    IN-2013-77878    2/5/2013    2/7/2013    Second Class    JR-162
10
2   25330    IN-2013-71249  10/17/2013  10/18/2013     First Class    CR-127
30
3   13524  ES-2013-1579342   1/28/2013   1/30/2013     First Class    KM-163
75
4   47221     SG-2013-4320   11/5/2013   11/6/2013        Same Day     RH-94
95

      Customer Name      Segment           City            State  ...   \
0       Rick Hansen     Consumer  New York City         New York  ...
1     Justin Ritter    Corporate     Wollongong  New South Wales  ...
2      Craig Reiter     Consumer       Brisbane       Queensland  ...
3  Katherine Murray  Home Office         Berlin           Berlin  ...
4       Rick Hansen     Consumer          Dakar            Dakar  ...

         Product ID    Category Sub-Category  \
0   TEC-AC-10003033  Technology  Accessories
1   FUR-CH-10003950   Furniture       Chairs
2   TEC-PH-10004664  Technology       Phones
3   TEC-PH-10004583  Technology       Phones
4  TEC-SHA-10000501  Technology      Copiers

                                        Product Name     Sales Quantity  \
0  Plantronics CS510 - Over-the-Head monaural Wir...  2309.650        7
1          Novimex Executive Leather Armchair, Black  3709.395        9
2                  Nokia Smart Phone, with Caller ID  5175.171        9
3                     Motorola Smart Phone, Cordless  2892.510        5
4                     Sharp Wireless Fax, High-Speed  2832.960        8

  Discount    Profit  Shipping Cost  Order Priority
0      0.0  762.1845         933.57        Critical
1      0.1 -288.7650         923.63        Critical
2      0.1  919.9710         915.49          Medium
3      0.1  -96.5400         910.16          Medium
4      0.0  311.5200         903.04        Critical

[5 rows x 24 columns]
```

In [14]:

```
print("\nSummary statistics:")
print(data.describe())
```

```

Summary statistics:
             Row ID   Postal Code         Sales      Quantity      Discount  \
count  51290.00000   9994.000000  51290.000000  51290.000000  51290.000000   
mean   25645.50000  55190.379428    246.490581      3.476545      0.142908   
std    14806.29199  32063.693350    487.565361      2.278766      0.212280   
min        1.00000   1040.000000      0.444000      1.000000      0.000000   
25%    12823.25000  23223.000000     30.758625      2.000000      0.000000   
50%    25645.50000  56430.500000     85.053000      3.000000      0.000000   
75%    38467.75000  90008.000000    251.053200      5.000000      0.200000   
max    51290.00000  99301.000000  22638.480000     14.000000      0.850000   

             Profit  Shipping Cost  
count  51290.000000   51290.000000  
mean      28.610982      26.375915  
std      174.340972      57.296804  
min    -6599.978000       0.000000  
25%        0.000000       2.610000  
50%        9.240000       7.790000  
75%       36.810000      24.450000  
max     8399.976000     933.570000  
```

In [16]:

```
print("\nData types and missing values:")
print(data.info())
```

```
Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51290 entries, 0 to 51289
Data columns (total 24 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Row ID          51290 non-null  int64  
 1   Order ID        51290 non-null  object 
 2   Order Date      51290 non-null  object 
 3   Ship Date       51290 non-null  object 
 4   Ship Mode       51290 non-null  object 
 5   Customer ID     51290 non-null  object 
 6   Customer Name   51290 non-null  object 
 7   Segment         51290 non-null  object 
 8   City            51290 non-null  object 
 9   State           51290 non-null  object 
 10  Country         51290 non-null  object 
 11  Postal Code     9994 non-null   float64
 12  Market          51290 non-null  object 
 13  Region          51290 non-null  object 
 14  Product ID      51290 non-null  object 
 15  Category        51290 non-null  object 
 16  Sub-Category    51290 non-null  object 
 17  Product Name    51290 non-null  object 
 18  Sales           51290 non-null  float64
 19  Quantity        51290 non-null  int64  
 20  Discount        51290 non-null  float64
 21  Profit          51290 non-null  float64
 22  Shipping Cost   51290 non-null  float64
 23  Order Priority  51290 non-null  object 
dtypes: float64(5), int64(2), object(17)
memory usage: 9.4+ MB
None
```

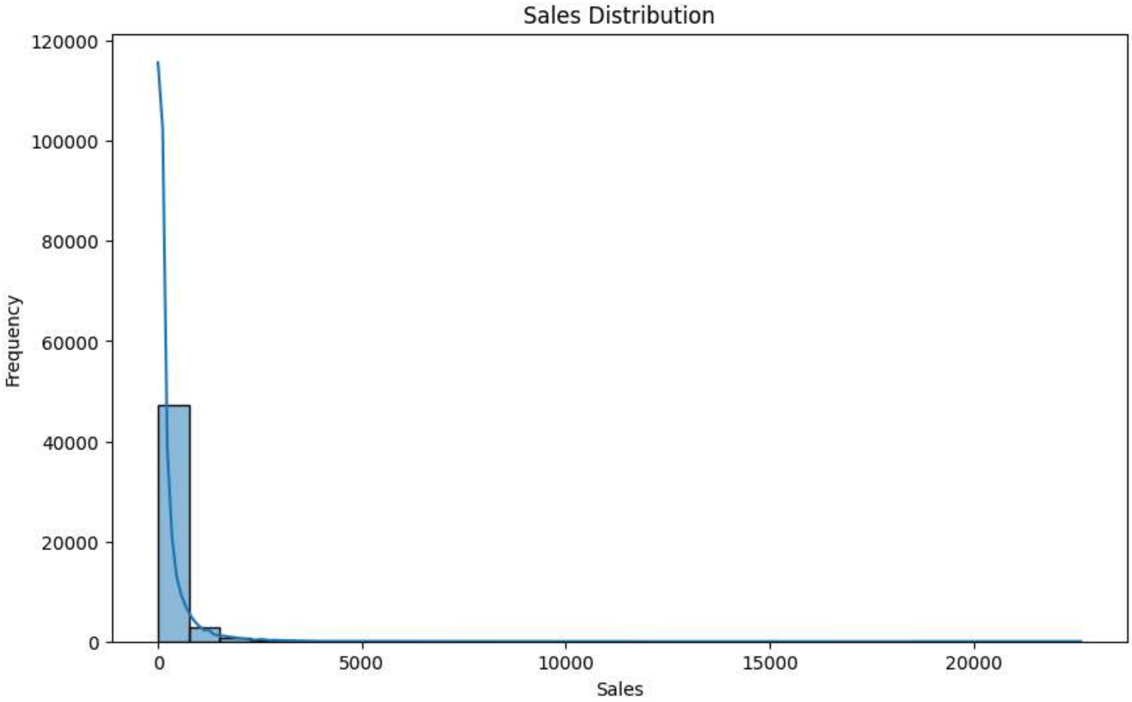
In [17]:

```
# Check for missing values
print("\nMissing values in each column:")
print(data.isnull().sum())
```

```
Missing values in each column:
Row ID                0
Order ID              0
Order Date            0
Ship Date             0
Ship Mode             0
Customer ID           0
Customer Name         0
Segment               0
City                  0
State                 0
Country               0
Postal Code       41296
Market                0
Region                0
Product ID            0
Category              0
Sub-Category          0
Product Name          0
Sales                 0
Quantity              0
Discount              0
Profit                0
Shipping Cost         0
Order Priority        0
dtype: int64
```

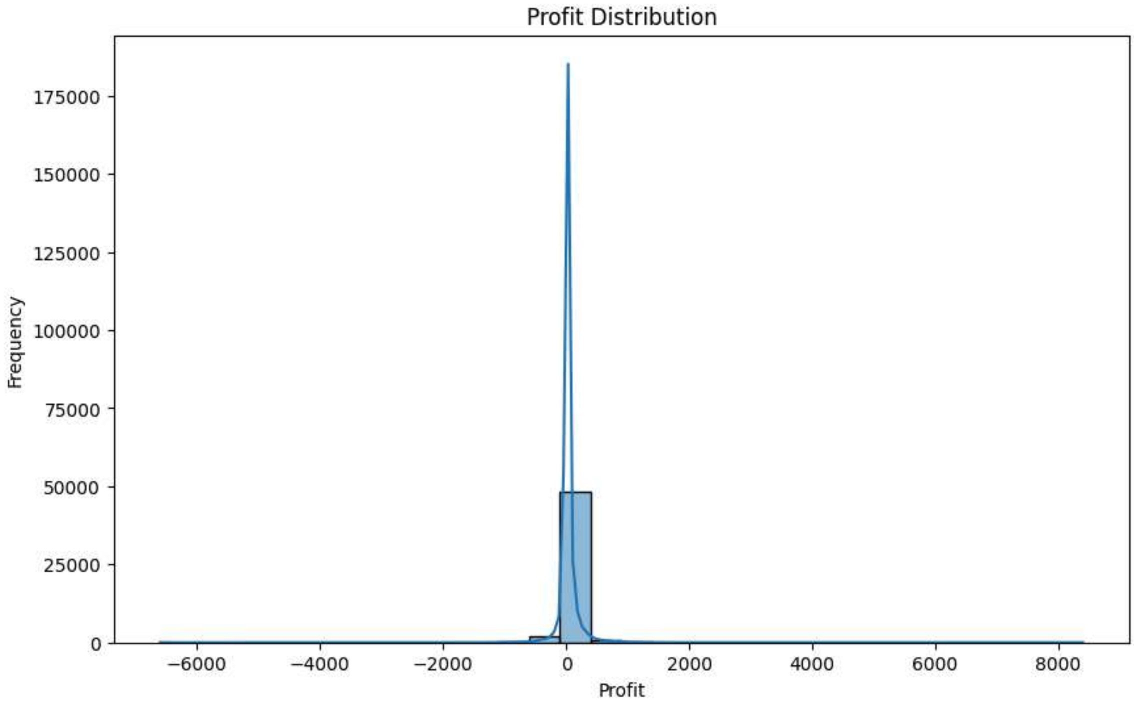
In [18]:

```
'''
1. Sales Distribution
This plot shows the distribution of sales values in the dataset.
It helps to understand the overall range and frequency of sales.
We use a histogram with a Kernel Density Estimate (KDE) to see the distribu
'''
plt.figure(figsize=(10, 6))
sns.histplot(data['Sales'], bins=30, kde=True)
plt.title('Sales Distribution')
plt.xlabel('Sales')
plt.ylabel('Frequency')
plt.show()
```



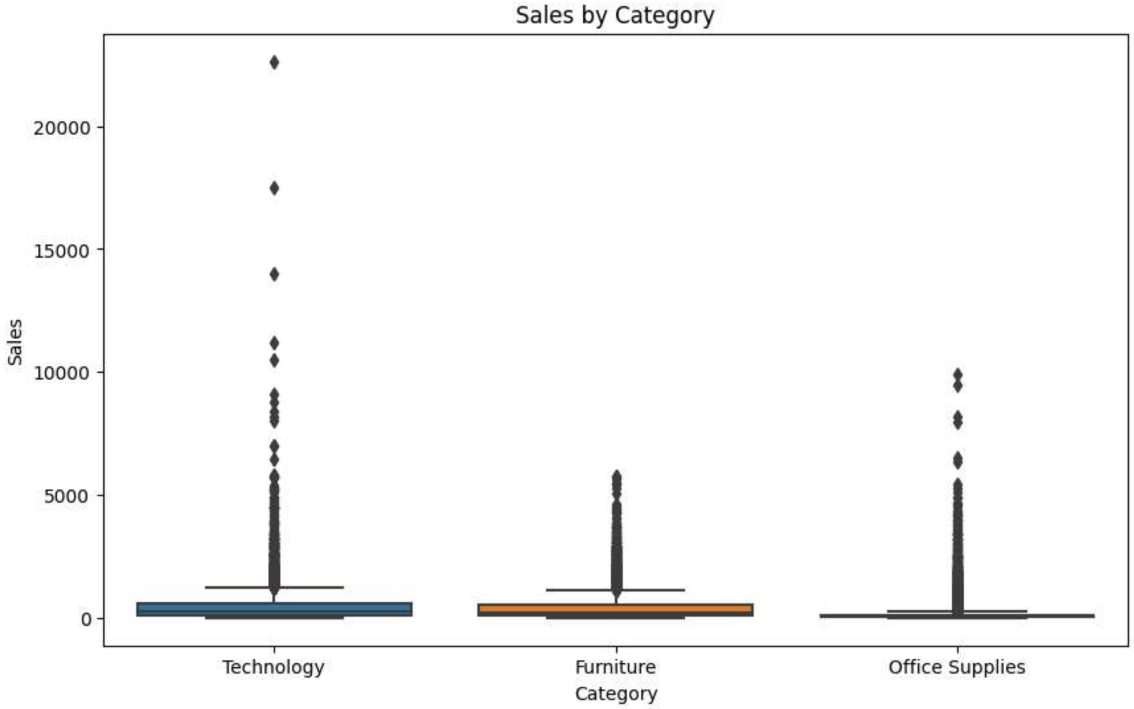
In [19]:

```
'''
2. Profit Distribution
This plot shows the distribution of profit values in the dataset.
It helps to understand the overall range and frequency of profits.
We use a histogram with a Kernel Density Estimate (KDE) to see the distribu
'''
plt.figure(figsize=(10, 6))
sns.histplot(data['Profit'], bins=30, kde=True)
plt.title('Profit Distribution')
plt.xlabel('Profit')
plt.ylabel('Frequency')
plt.show()
```

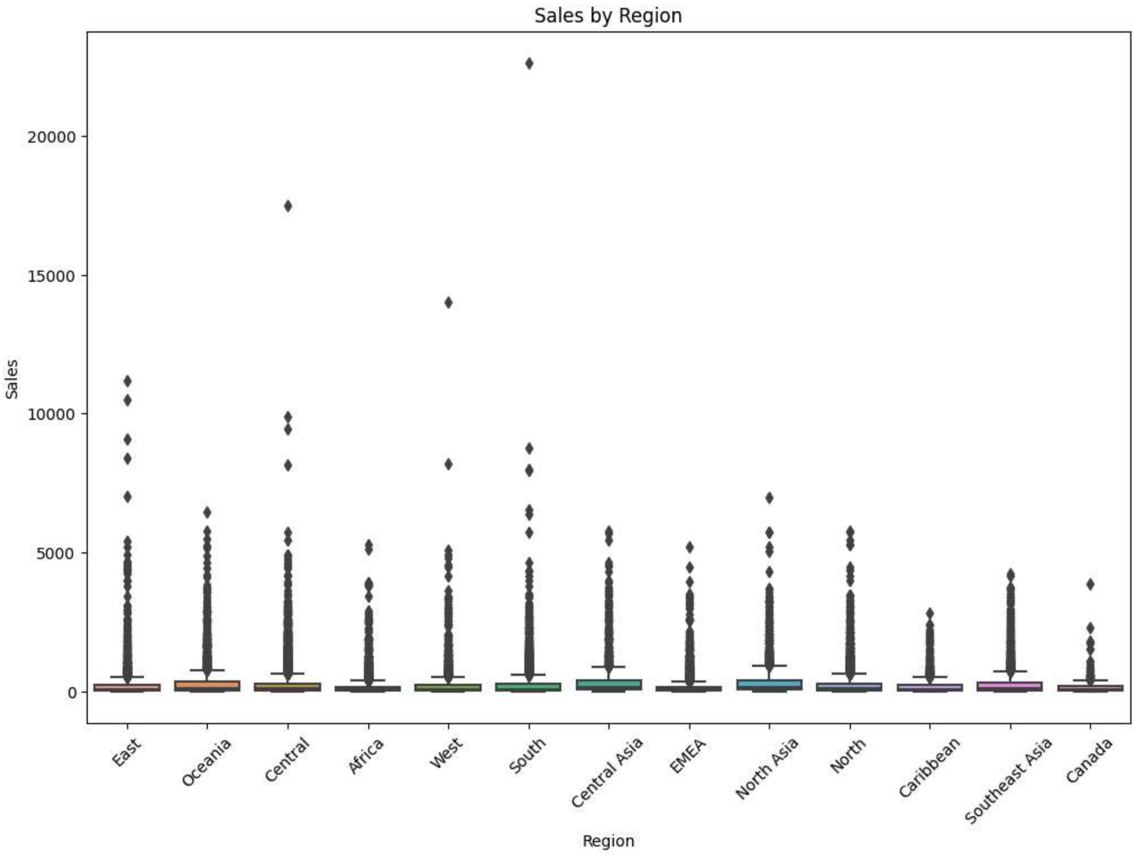


In [20]:

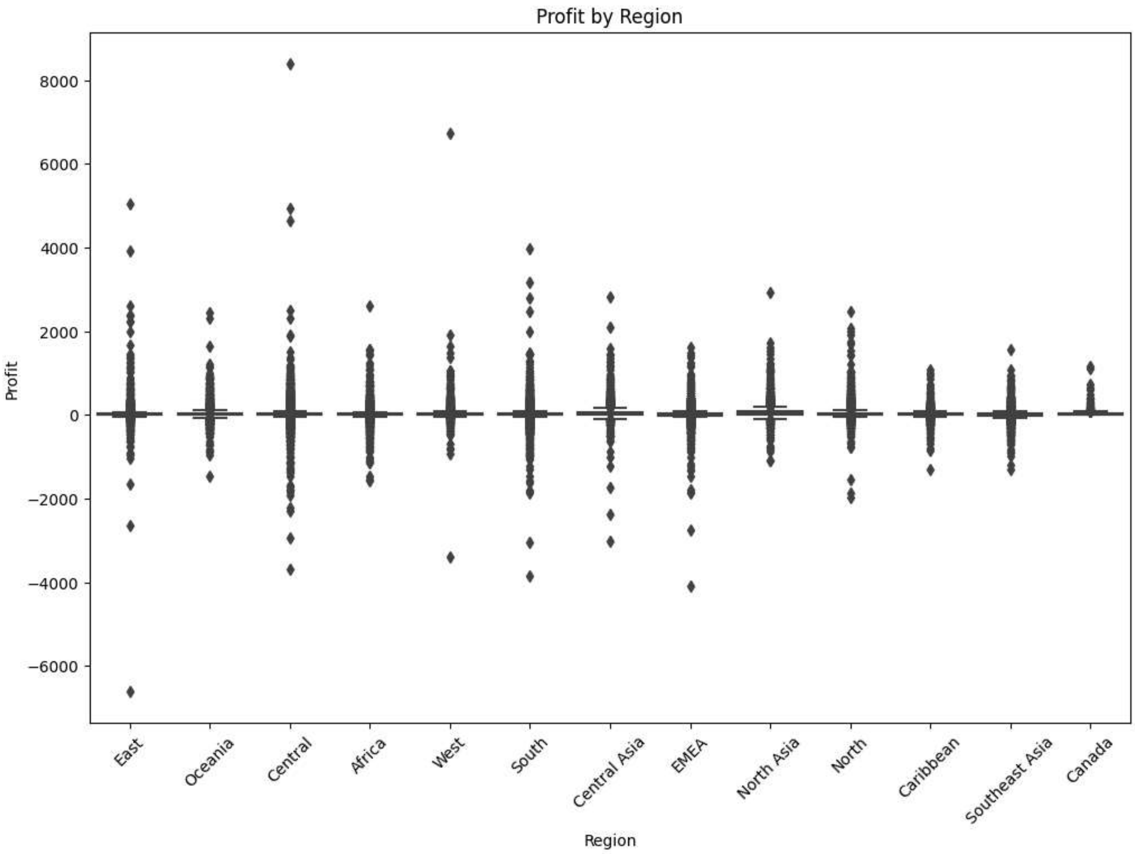
```
'''
3. Sales by Category
This box plot shows the distribution of sales for different product categor
It helps to identify which categories generate higher or lower sales.
Box plots are useful for displaying the median, quartiles, and potential ou
'''
plt.figure(figsize=(10, 6))
sns.boxplot(x='Category', y='Sales', data=data)
plt.title('Sales by Category')
plt.xlabel('Category')
plt.ylabel('Sales')
plt.show()
```



```
In [21]: '''
         4. Sales by Region
         This box plot shows the distribution of sales for different regions.
         It helps to identify regional differences in sales performance.
         Box plots allow comparison across different regions, showing central tenden
         '''
         plt.figure(figsize=(12, 8))
         sns.boxplot(x='Region', y='Sales', data=data)
         plt.title('Sales by Region')
         plt.xlabel('Region')
         plt.ylabel('Sales')
         plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
         plt.show()
```

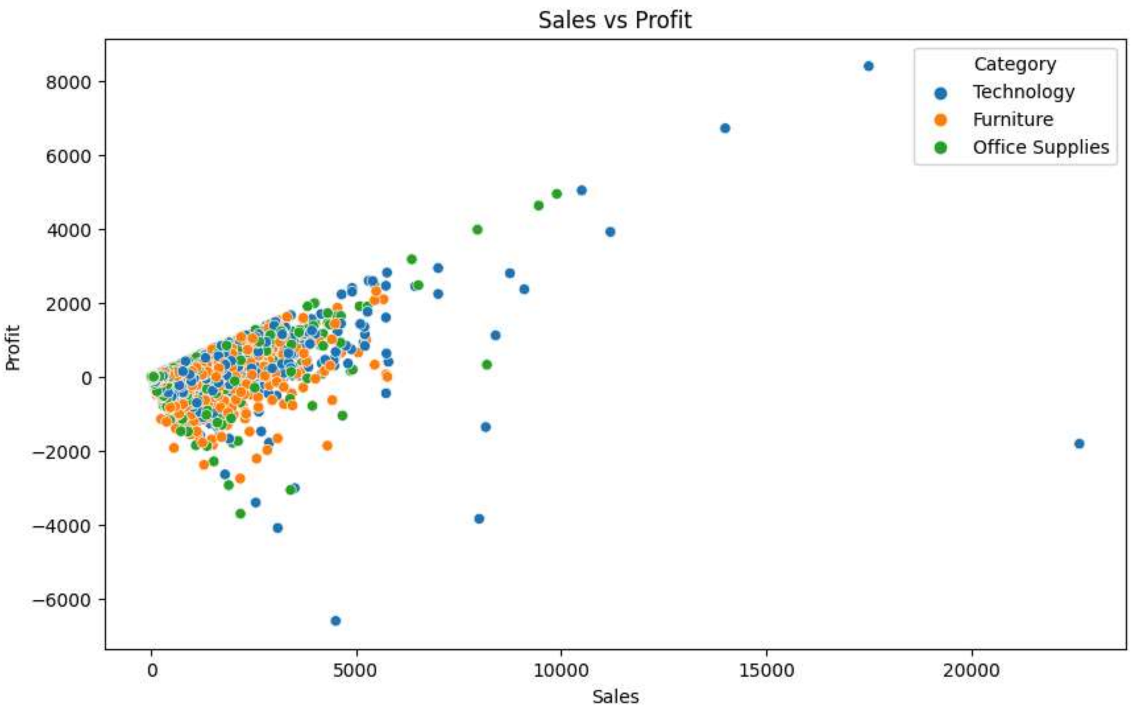


In [22]:
```
'''
5. Profit by Region
This box plot shows the distribution of profit for different regions.
It helps to identify regional differences in profitability.
Box plots allow comparison across different regions, showing central tenden
'''
plt.figure(figsize=(12, 8))
sns.boxplot(x='Region', y='Profit', data=data)
plt.title('Profit by Region')
plt.xlabel('Region')
plt.ylabel('Profit')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.show()
```



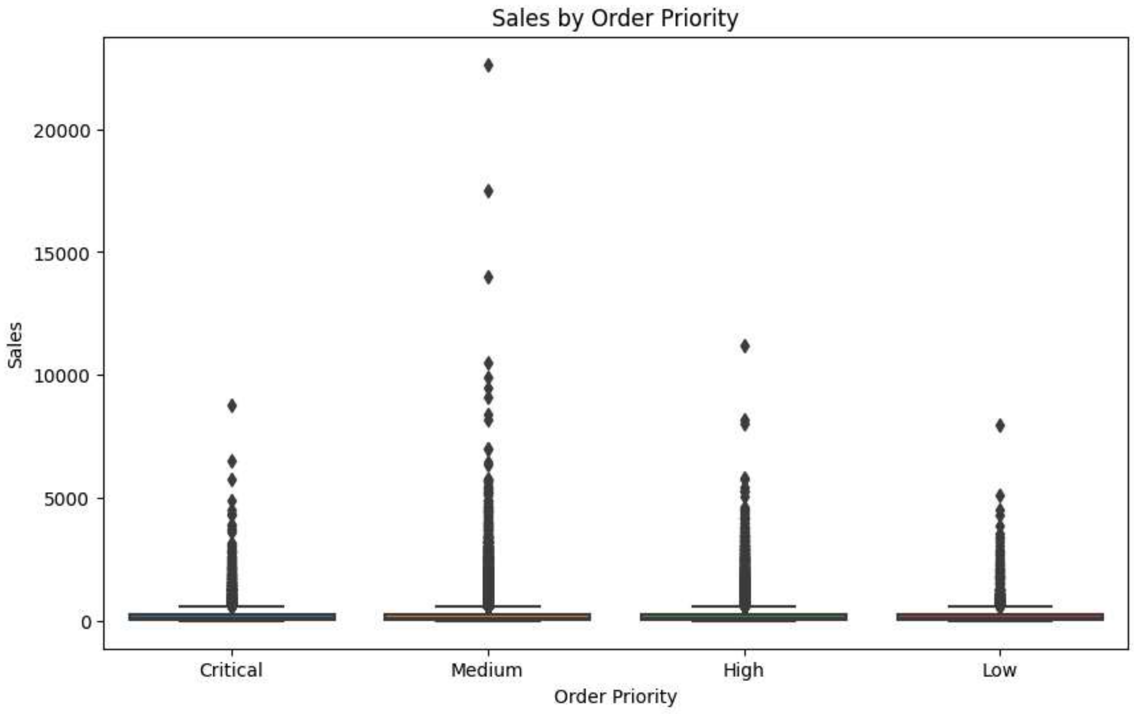
In [23]:

```
'''
6. Sales vs Profit
This scatter plot shows the relationship between sales and profit.
Each point represents a transaction, and the color indicates the product ca
It helps to visualize how sales and profit are related and identify trends
'''
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Sales', y='Profit', hue='Category', data=data)
plt.title('Sales vs Profit')
plt.xlabel('Sales')
plt.ylabel('Profit')
plt.legend(title='Category')
plt.show()
```



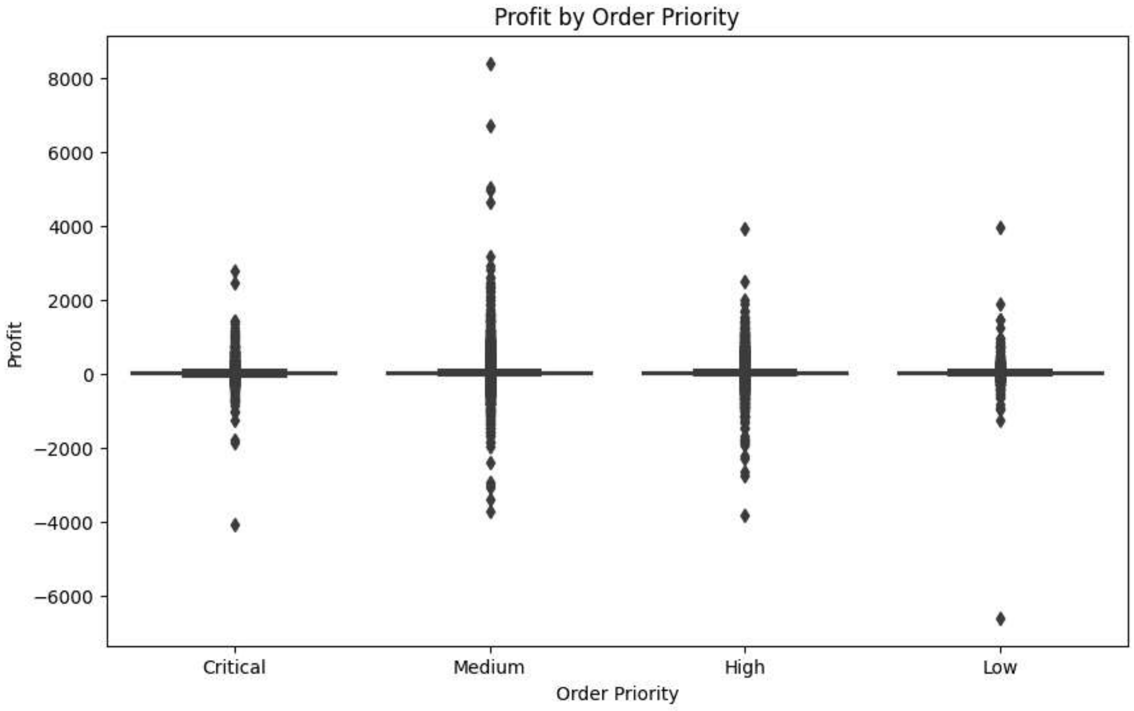
In [24]:

```
'''
7. Sales by Order Priority
This box plot shows the distribution of sales for different order prioritie
It helps to identify how order priority impacts sales.
Box plots display the median, quartiles, and potential outliers for sales a
'''
plt.figure(figsize=(10, 6))
sns.boxplot(x='Order Priority', y='Sales', data=data)
plt.title('Sales by Order Priority')
plt.xlabel('Order Priority')
plt.ylabel('Sales')
plt.show()
```



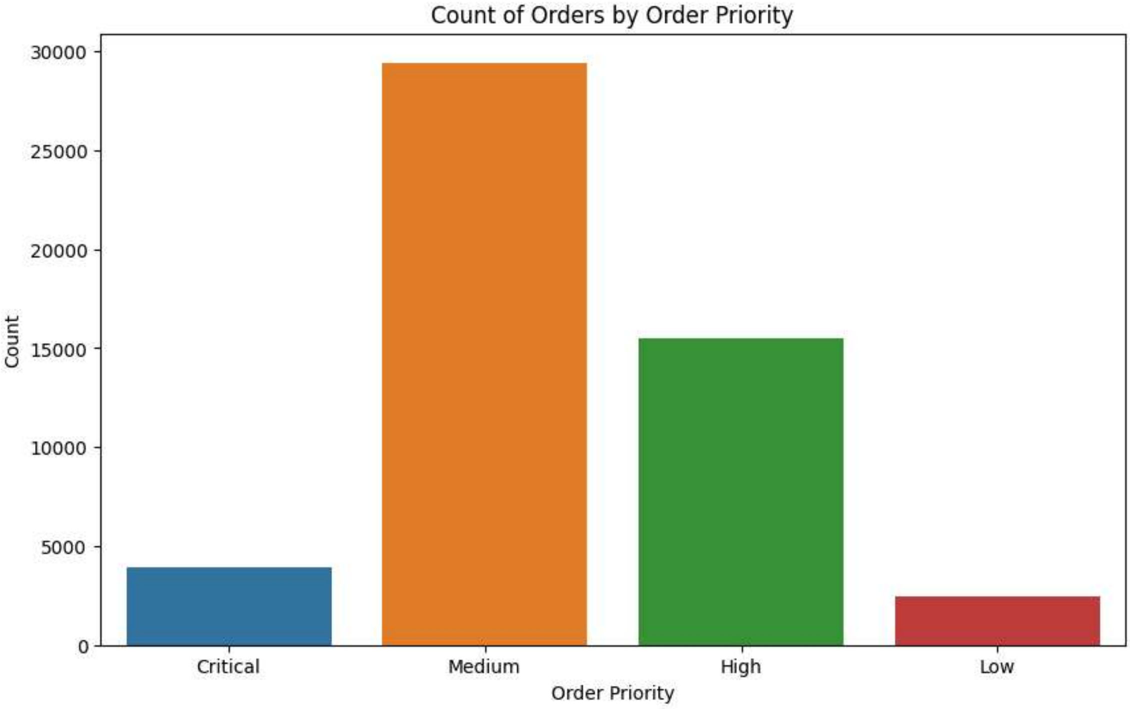
In [25]:

```
'''
8. Profit by Order Priority
This box plot shows the distribution of profit for different order prioriti
It helps to identify how order priority impacts profitability.
Box plots display the median, quartiles, and potential outliers for profit
'''
plt.figure(figsize=(10, 6))
sns.boxplot(x='Order Priority', y='Profit', data=data)
plt.title('Profit by Order Priority')
plt.xlabel('Order Priority')
plt.ylabel('Profit')
plt.show()
```



In [26]:

```
'''
9. Count of Orders by Order Priority
This count plot shows the number of orders for each order priority.
It helps to understand the frequency of different order priorities.
Count plots display the total number of occurrences for each category.
'''
plt.figure(figsize=(10, 6))
sns.countplot(x='Order Priority', data=data)
plt.title('Count of Orders by Order Priority')
plt.xlabel('Order Priority')
plt.ylabel('Count')
plt.show()
```



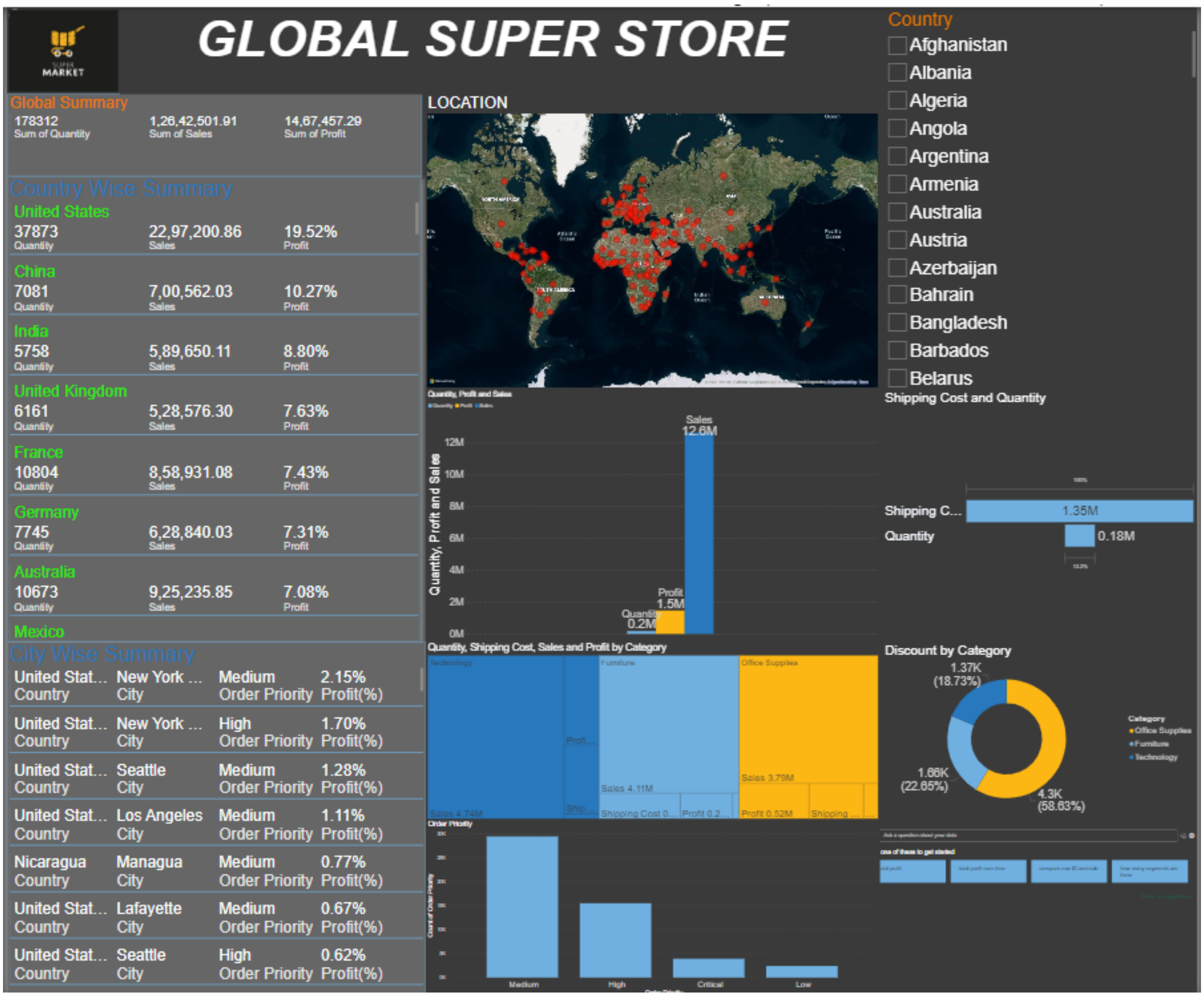
In [ ]:

DATA EXPLORATION USING POWERBI

DASHBOARD – https://drive.google.com/file/d/13AsQ0JZJmfpo0WBfL34dIN2rWo4nRbcL/view?usp=sharing

Here I utilized PowerBI to further explore and visualize the sales data from a global superstore. PowerBI's intuitive interface and robust visualization tools allowed me to easily uncover insights and trends within the dataset. By leveraging PowerBI's capabilities, I could create clear and effective visualizations that highlight key aspects of the data, such as order priority distribution, sales performance, profit margins, and customer segmentation. This approach enabled me to efficiently analyze and interpret the data, making the entire process of data exploration both straightforward and insightful.

Here is a rough understanding about the dashboard



GLOBAL SUPER STORE

Global Summary

178312 Sum of Quantity	1,26,42,501.91 Sum of Sales	14,67,457.29 Sum of Profit

Country Wise Summary

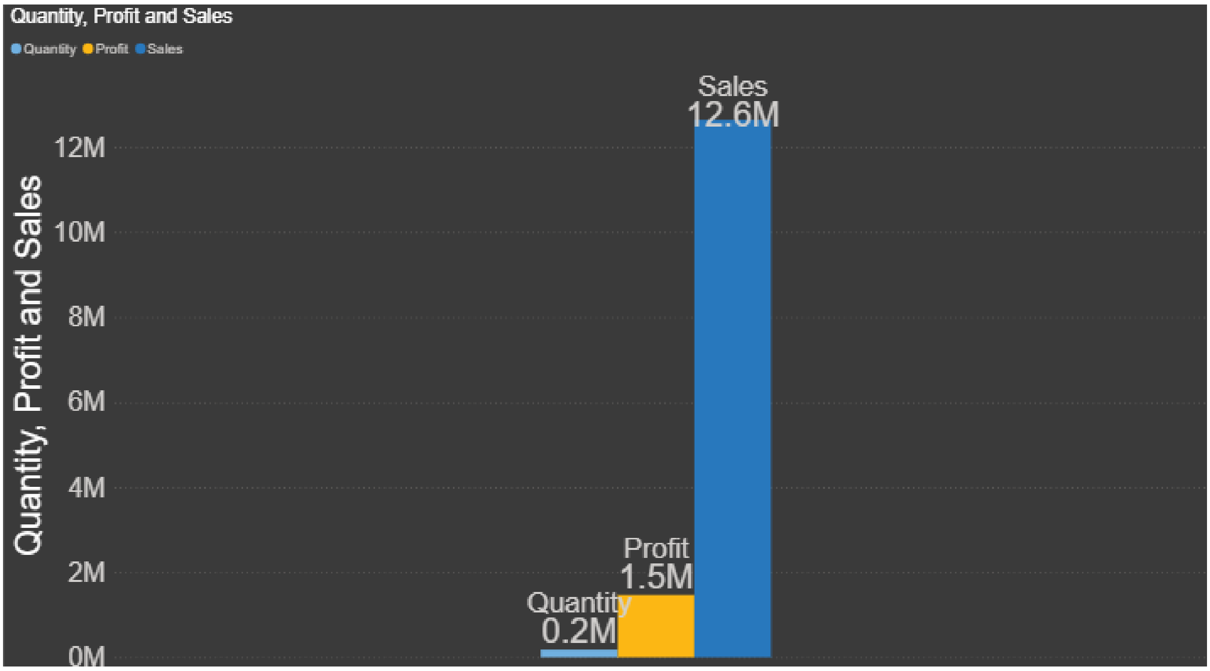
United States		
37873 Quantity	22,97,200.86 Sales	19.52% Profit
China		
7081 Quantity	7,00,562.03 Sales	10.27% Profit
India		
5758 Quantity	5,89,650.11 Sales	8.80% Profit
United Kingdom		
6161 Quantity	5,28,576.30 Sales	7.63% Profit
France		
10804 Quantity	8,58,931.08 Sales	7.43% Profit
Germany		
7745 Quantity	6,28,840.03 Sales	7.31% Profit
Australia		
10673 Quantity	9,25,235.85 Sales	7.08% Profit
Mexico		

City Wise Summary

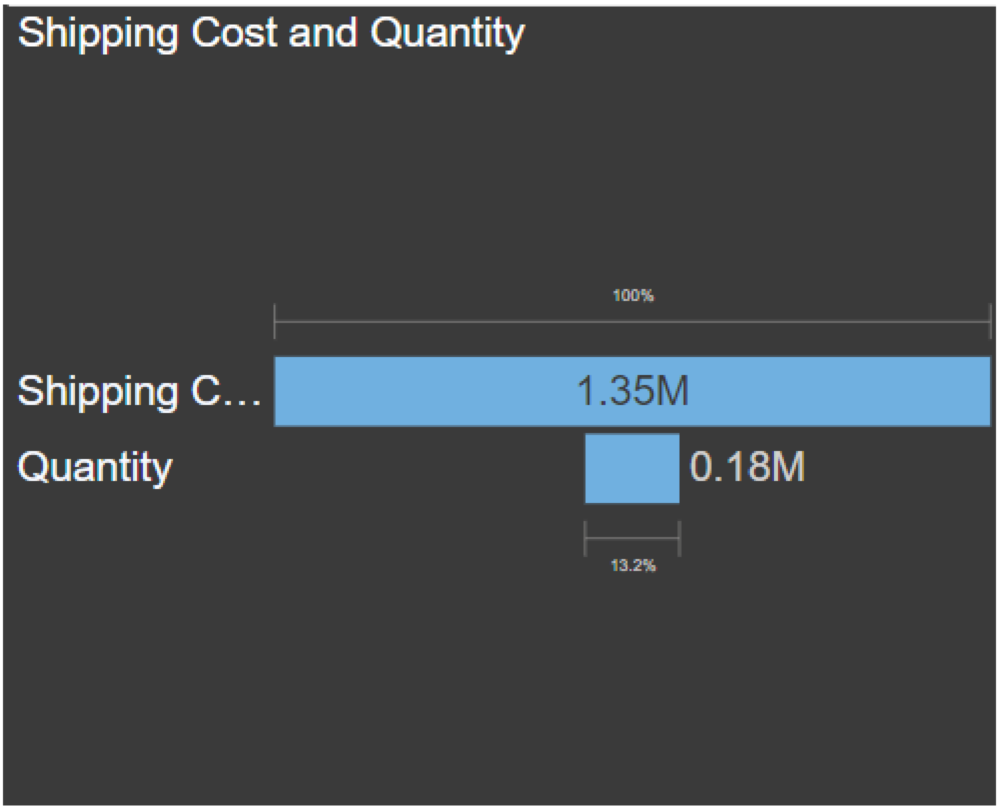
United Stat... Country	New York ... City	Medium Order Priority	2.15% Profit(%)
United Stat... Country	New York ... City	High Order Priority	1.70% Profit(%)
United Stat... Country	Seattle City	Medium Order Priority	1.28% Profit(%)
United Stat... Country	Los Angeles City	Medium Order Priority	1.11% Profit(%)
Nicaragua Country	Managua City	Medium Order Priority	0.77% Profit(%)
United Stat... Country	Lafayette City	Medium Order Priority	0.67% Profit(%)
United Stat... Country	Seattle City	High Order Priority	0.62% Profit(%)

Country
- □ Afghanistan
- □ Albania
- □ Algeria
- □ Angola
- □ Argentina
- □ Armenia
- □ Australia
- □ Austria
- □ Azerbaijan
- □ Bahrain
- □ Bangladesh
- □ Barbados
- □ Belarus

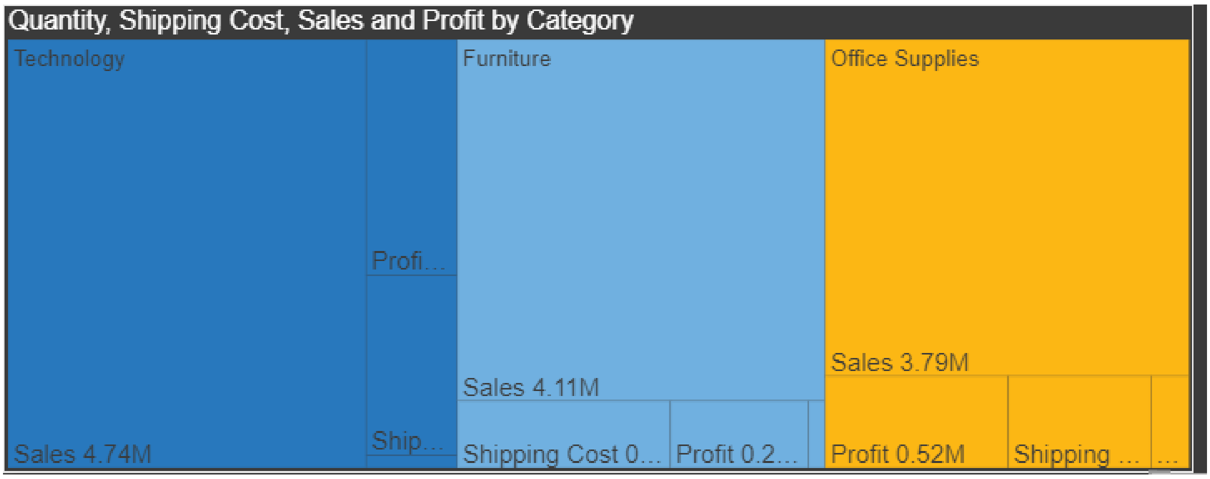
General context of the visualizations:



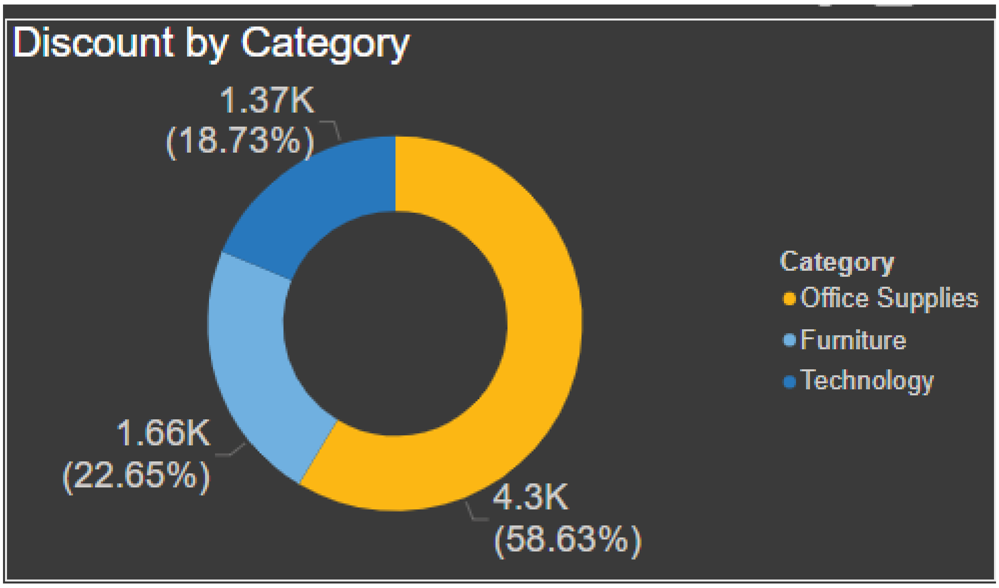
The stacked column chart I created in PowerBI includes quantity, sales, and profit. This chart allows me to understand the relationship and distribution of these metrics over a chosen country.



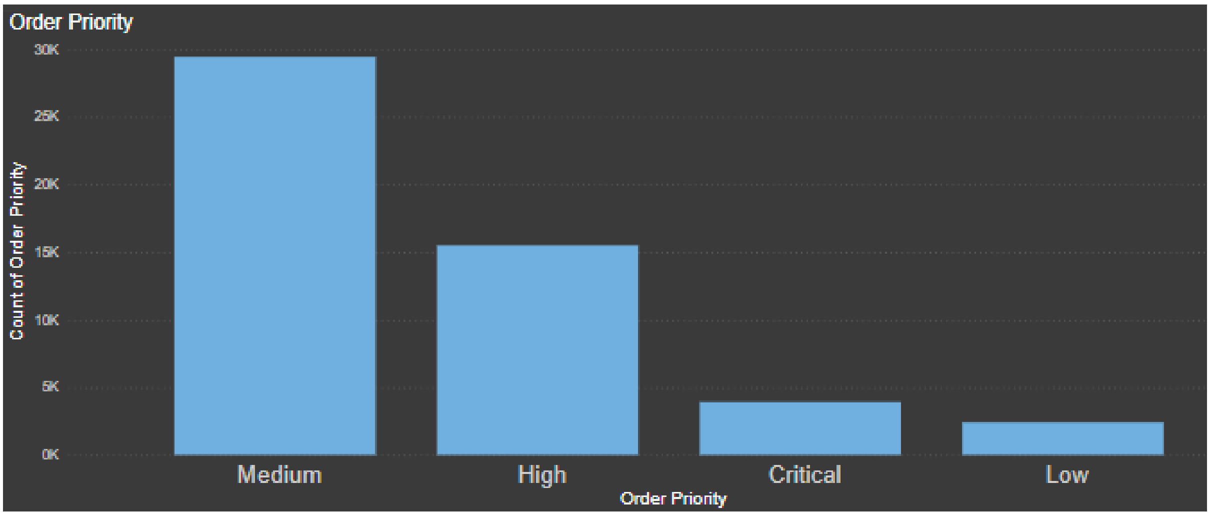
I created a multi-row card in PowerBI to display shipping costs and quantities for different segments of the data. This visualization allows me to quickly understand the distribution and impact of shipping costs and quantities across various dimensions like country,city,region



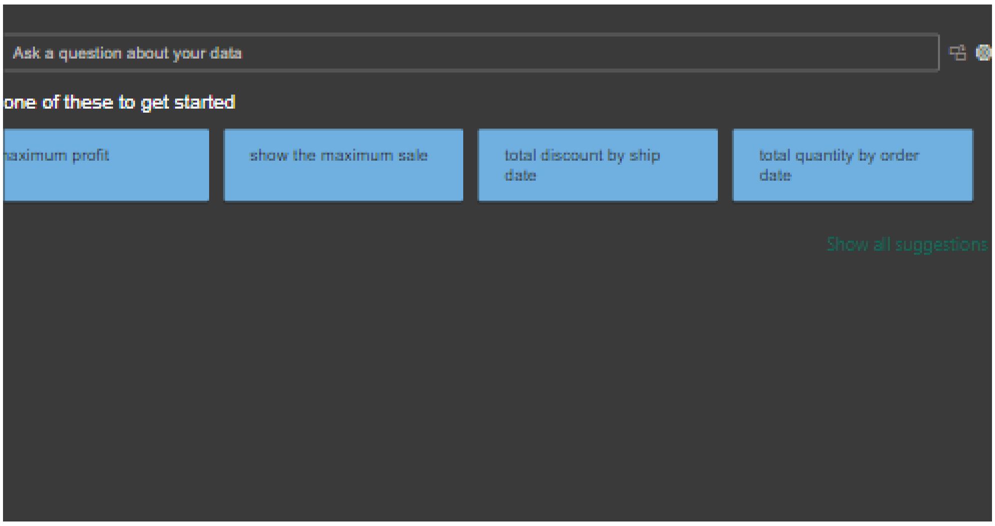
I created a tree map in PowerBI to visualize the quantity, shipping cost, and profit by product category. The tree map provides a hierarchical view that helps me understand the relative size and contribution of each category in terms of these metrics.



I created a donut chart in PowerBI to visualize the distribution of discounts across different product categories. This chart helps me understand which categories are receiving the most discounts and can indicate areas where promotional activities are concentrated.



I created a stacked column chart in PowerBI to visualize the count of orders by order priority. This chart helps me understand the distribution of orders across different priority levels, providing insights into customer behavior and operational efficiency.


Ask a question about your data
one of these to get started
aximum profit
show the maximum sale
total discount by ship date
total quantity by order date
Show all suggestions

I incorporated a Q&A visualization into my PowerBI dashboard, allowing users to ask natural language questions about the data and receive immediate, interactive answers. This feature enhances user engagement and provides instant insights without needing predefined reports or visualizations.