Project Overview

This project analyzes sales data from Dmart, focusing on key business metrics like sales, profit, discounts, and customer behavior. The dataset contains 8,047 records with 17 attributes, including customer details, product categories, order information, and financials. The analysis covers data cleaning, outlier detection, exploratory data analysis (EDA), and visualizations to uncover trends and patterns.

Objective

To analyze sales performance, profitability, and discount impact while identifying key trends, top customers, and optimization opportunities.

```python
#importing necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
```

Data Loading

```python
df = pd.read_excel("C:/Users/KIIT/Desktop/Projects/DMart Sales/DMart 
Data Store.xlsx")

# to preview the data
df
```

```
             Order ID Order Date     Customer Name         Country  \
0     BN-2011-7407039 2011-01-01        Ruby Patel          Sweden
1     AZ-2011-9050313 2011-01-03    Summer Hayward  United Kingdom
2     AZ-2011-6674300 2011-01-04  Devin Huddleston          France
3     BN-2011-2819714 2011-01-04       Mary Parker  United Kingdom
4     BN-2011-2819714 2011-01-04       Mary Parker  United Kingdom
...               ...        ...               ...             ...
8042   AZ-2014-766953 2014-12-31      Jose Gambino  United Kingdom
8043  BN-2014-4140795 2014-12-31   Daniel Hamilton     Netherlands
8044  BN-2014-4140795 2014-12-31   Daniel Hamilton     Netherlands
8045  BN-2014-4140795 2014-12-31   Daniel Hamilton     Netherlands
8046   AZ-2014-766953 2014-12-31      Jose Gambino  United Kingdom

                     State        City   Region      Segment     Ship 
Mode  \
0                Stockholm   Stockholm    North  Home Office  Economy 
Plus
1                  England   Southport    North     Consumer
Economy
2     Auvergne-Rhône-Alpes     Valence  Central     Consumer
Economy
3                  England  Birmingham    North    Corporate
```

```
Economy
4                  England  Birmingham    North    Corporate
Economy
...                    ...         ...      ...          ...
...
8042               England  Maidenhead    North    Corporate
Economy
8043         North Brabant   Eindhoven  Central  Home Office  Economy
Plus
8044         North Brabant   Eindhoven  Central  Home Office  Economy
Plus
8045         North Brabant   Eindhoven  Central  Home Office  Economy
Plus
8046               England  Maidenhead    North    Corporate
Economy

             Category Sub-Category                           Product
Name  \
0     Office Supplies        Paper            Enermax Note Cards,
Premium
1           Furniture    Bookcases     Dania Corner Shelving,
Traditional
2     Office Supplies          Art  Binney & Smith Sketch Pad, Easy-
Erase
3     Office Supplies          Art             Boston Markers, Easy-
Erase
4     Office Supplies      Storage            Eldon Folders, Single
Width
...               ...          ...
...
8042        Furniture    Bookcases   Ikea Stackable Bookrack,
Traditional
8043  Office Supplies          Art      BIC Pencil Sharpener,
Fluorescent
8044  Office Supplies      Binders          Avery Binder Covers,
Recycled
8045       Technology     Machines                    StarTech Phone,
Red
8046       Technology       Phones              Motorola Audio Dock,
VoIP

      Discount  Sales  Profit  Quantity  Feedback?
0          0.5     45     -26         3      False
1          0.0    854     290         7       True
2          0.0    140      21         3       True
3          0.5     27     -22         2       True
4          0.5     17      -1         2       True
...        ...    ...     ...       ...        ...
8042       0.0    245      91         2       True
```

```
8043       0.5     30     -10         2      False
8044       0.5     23      -6         4       True
8045       0.5    108     -19         3      False
8046       0.0    867     251         5      False

[8047 rows x 17 columns]

#to check for the first 10 rows
df.head(10)

          Order ID Order Date        Customer Name         Country  \
0  BN-2011-7407039 2011-01-01           Ruby Patel          Sweden
1  AZ-2011-9050313 2011-01-03       Summer Hayward  United Kingdom
2  AZ-2011-6674300 2011-01-04     Devin Huddleston          France
3  BN-2011-2819714 2011-01-04          Mary Parker  United Kingdom
4  BN-2011-2819714 2011-01-04          Mary Parker  United Kingdom
5   AZ-2011-617423 2011-01-05         Daniel Burke          France
6   AZ-2011-617423 2011-01-05         Daniel Burke          France
7  AZ-2011-2918397 2011-01-07   Fredrick Beveridge          France
8  AZ-2011-2918397 2011-01-07   Fredrick Beveridge          France
9  AZ-2011-2918397 2011-01-07   Fredrick Beveridge          France

                         State              City   Region      Segment
\
0                    Stockholm         Stockholm    North  Home Office

1                      England         Southport    North     Consumer

2         Auvergne-Rhône-Alpes           Valence  Central     Consumer

3                      England        Birmingham    North    Corporate

4                      England        Birmingham    North    Corporate

5         Auvergne-Rhône-Alpes        Echirolles  Central  Home Office

6         Auvergne-Rhône-Alpes        Echirolles  Central  Home Office

7  Provence-Alpes-Côte d'Azur  La Seyne-sur-Mer  Central    Corporate

8  Provence-Alpes-Côte d'Azur  La Seyne-sur-Mer  Central    Corporate

9  Provence-Alpes-Côte d'Azur  La Seyne-sur-Mer  Central    Corporate


       Ship Mode         Category Sub-Category  \
0  Economy Plus  Office Supplies        Paper
1       Economy        Furniture    Bookcases
2       Economy  Office Supplies          Art
3       Economy  Office Supplies          Art
4       Economy  Office Supplies      Storage
```

```
5      Priority  Office Supplies          Art
6      Priority  Office Supplies          Art
7      Priority        Furniture    Bookcases
8      Priority  Office Supplies    Fasteners
9      Priority  Office Supplies      Storage

                                   Product Name  Discount  Sales
Profit  \
0                   Enermax Note Cards, Premium       0.5     45     -
26
1            Dania Corner Shelving, Traditional       0.0    854
290
2         Binney & Smith Sketch Pad, Easy-Erase       0.0    140
21
3                    Boston Markers, Easy-Erase       0.5     27     -
22
4                   Eldon Folders, Single Width       0.5     17
-1
5  Binney & Smith Pencil Sharpener, Water Color       0.0     90
21
6                   Sanford Canvas, Fluorescent       0.0    207
77
7                 Bush Floating Shelf Set, Pine       0.1    155
36
8             Accos Thumb Tacks, Assorted Sizes       0.0     33
2
9                     Smead Lockers, Industrial       0.1    716
143

   Quantity  Feedback?
0         3      False
1         7       True
2         3       True
3         2       True
4         2       True
5         3      False
6         4      False
7         1       True
8         3       True
9         4       True
```

```
#to check for the last 10 rows
df.tail(10)
```

```
             Order ID Order Date        Customer Name         Country
\
8037   AZ-2014-436448 2014-12-30     Georgia Arundale           Italy

8038  AZ-2014-1412225 2014-12-31          Leon Barnes  United Kingdom
```

```
8039  AZ-2014-4217323 2014-12-31          Evie Morton          France

8040  AZ-2014-8174835 2014-12-31         Eloise Sykes         Germany

8041  AZ-2014-7604524 2014-12-31  Rebecca Chamberlain         Germany

8042   AZ-2014-766953 2014-12-31         Jose Gambino  United Kingdom

8043  BN-2014-4140795 2014-12-31      Daniel Hamilton     Netherlands

8044  BN-2014-4140795 2014-12-31      Daniel Hamilton     Netherlands

8045  BN-2014-4140795 2014-12-31      Daniel Hamilton     Netherlands

8046   AZ-2014-766953 2014-12-31         Jose Gambino  United Kingdom


                       State        City   Region      Segment
Ship Mode  \
8037                Campania      Naples    South    Corporate
Economy
8038                 England   Worcester    North     Consumer
Priority
8039                Normandy        Caen  Central     Consumer
Economy Plus
8040  North Rhine-Westphalia   Bielefeld  Central     Consumer
Economy
8041                 Hamburg     Hamburg  Central  Home Office
Economy
8042                 England  Maidenhead    North    Corporate
Economy
8043           North Brabant   Eindhoven  Central  Home Office
Economy Plus
8044           North Brabant   Eindhoven  Central  Home Office
Economy Plus
8045           North Brabant   Eindhoven  Central  Home Office
Economy Plus
8046                 England  Maidenhead    North    Corporate
Economy

             Category Sub-Category                          Product
Name  \
8037  Office Supplies      Binders                  Acco Binder,
Economy
8038  Office Supplies      Storage       Fellowes Shelving, Single
Width
8039  Office Supplies      Storage          Fellowes Lockers, Wire
Frame
8040  Office Supplies     Supplies             Kleencut Shears,
Serrated
```

```
8041  Office Supplies      Binders       Wilson Jones Index Tab,
Economy
8042        Furniture    Bookcases  Ikea Stackable Bookrack,
Traditional
8043  Office Supplies          Art      BIC Pencil Sharpener,
Fluorescent
8044  Office Supplies      Binders          Avery Binder Covers,
Recycled
8045       Technology     Machines                   StarTech Phone,
Red
8046       Technology       Phones             Motorola Audio Dock,
VoIP

      Discount  Sales  Profit  Quantity  Feedback?
8037       0.0     45       6         3       True
8038       0.0    289      75         5       True
8039       0.1    557     217         3      False
8040       0.0    261      13         6       True
8041       0.0     32       8         5       True
8042       0.0    245      91         2       True
8043       0.5     30     -10         2      False
8044       0.5     23      -6         4       True
8045       0.5    108     -19         3      False
8046       0.0    867     251         5      False
```

Understanding the dataset

```
#the total column and rows
df.shape

(8047, 17)

#to check column names
df.columns

Index(['Order ID', 'Order Date', 'Customer Name', 'Country', 'State',
'City',
       'Region', 'Segment', 'Ship Mode', 'Category', 'Sub-Category',
       'Product Name', 'Discount', 'Sales', 'Profit', 'Quantity',
'Feedback?'],
      dtype='object')

#data types and null values
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8047 entries, 0 to 8046
Data columns (total 17 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
```

```
 0   Order ID       8047 non-null   object
 1   Order Date     8047 non-null   datetime64[ns]
 2   Customer Name  8047 non-null   object
 3   Country        8047 non-null   object
 4   State          8047 non-null   object
 5   City           8047 non-null   object
 6   Region         8047 non-null   object
 7   Segment        8047 non-null   object
 8   Ship Mode      8047 non-null   object
 9   Category       8047 non-null   object
 10  Sub-Category   8047 non-null   object
 11  Product Name   8047 non-null   object
 12  Discount       8047 non-null   float64
 13  Sales          8047 non-null   int64
 14  Profit         8047 non-null   int64
 15  Quantity       8047 non-null   int64
 16  Feedback?      8047 non-null   bool
dtypes: bool(1), datetime64[ns](1), float64(1), int64(3), object(11)
memory usage: 1013.9+ KB
```

```python
#to check for the null values
df.isnull().sum()
```

```
Order ID         0
Order Date       0
Customer Name    0
Country          0
State            0
City             0
Region           0
Segment          0
Ship Mode        0
Category         0
Sub-Category     0
Product Name     0
Discount         0
Sales            0
Profit           0
Quantity         0
Feedback?        0
dtype: int64
```

```python
#to check for duplicates
duplicate = df.duplicated().sum()
duplicate #2 duplicate rows
```

```
2
```

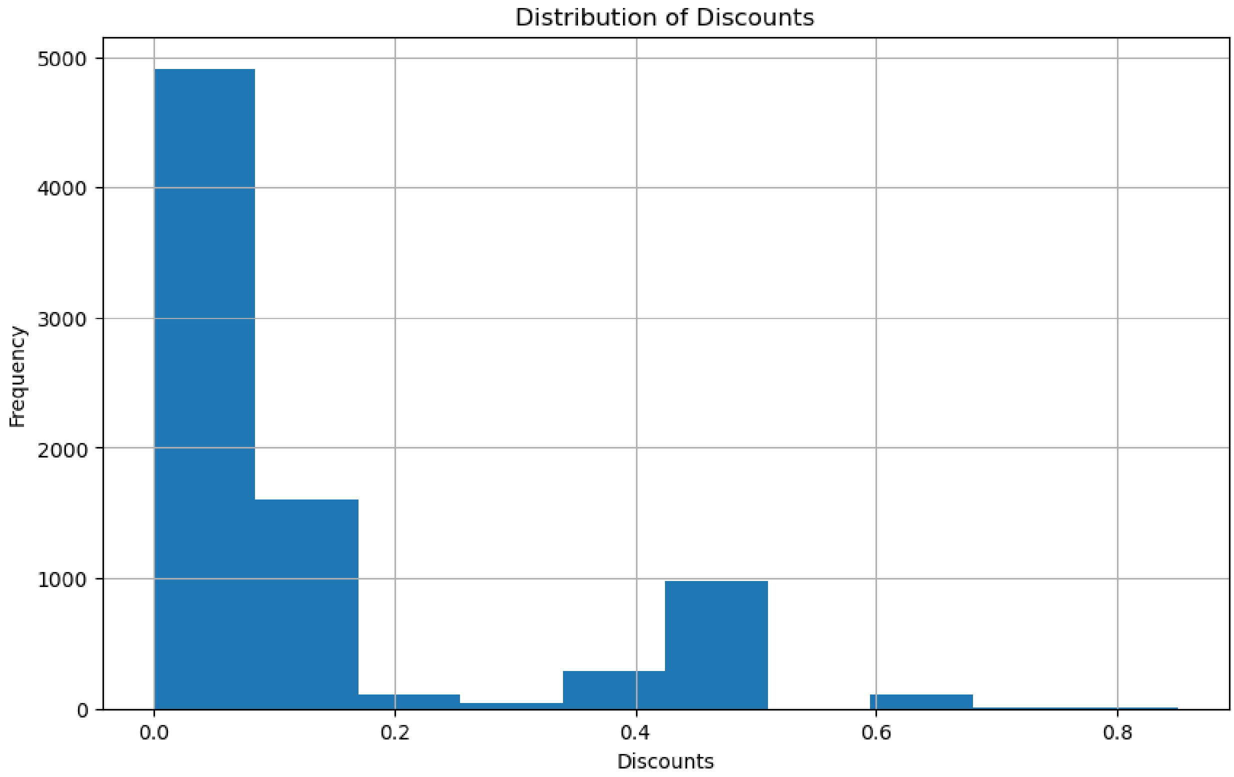
```python
#to check for the unique values
df.nunique()
```

```
Order ID         4117
Order Date       1214
Customer Name     792
Country            15
State             127
City              999
Region              3
Segment             3
Ship Mode           4
Category            3
Sub-Category       17
Product Name     1810
Discount           14
Sales            1248
Profit            845
Quantity           14
Feedback?           2
dtype: int64
```

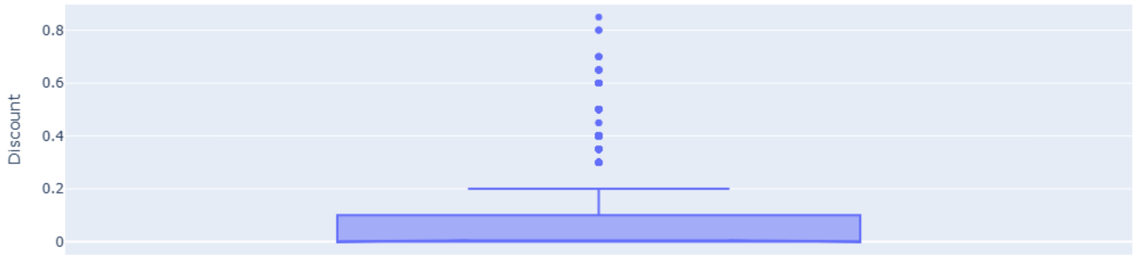
Outliers Detection and Treatment

For Discounts

```
df['Discount'].hist(bins = 10, figsize= (10,6))
plt.title('Distribution of Discounts')
plt.xlabel('Discounts')
plt.ylabel('Frequency')
plt.show()
```



```
#identifying the outliers
fig = px.box(df, y = 'Discount', notched = True)
fig.show()
```



```
q1 = df['Discount'].quantile(0.25)
q2 = df['Discount'].quantile(0.5)
q3 = df['Discount'].quantile(0.75)
iqr = q3 - q1

lb_dis = q1 - 1.5*iqr
lb_dis
```

```
-0.15000000000000002
```

```
ub_dis = q3 + 1.5*iqr
ub_dis
```

```
0.25
```

```
df[(df['Discount'] < lb_dis)  | (df['Discount'] > ub_dis)]
```

```
             Order ID Order Date    Customer Name         Country  \
0     BN-2011-7407039 2011-01-01       Ruby Patel          Sweden
3     BN-2011-2819714 2011-01-04      Mary Parker  United Kingdom
4     BN-2011-2819714 2011-01-04      Mary Parker  United Kingdom
10    BN-2011-3248724 2011-01-08      Archer Hort          France
11    BN-2011-3248724 2011-01-08      Archer Hort          France
...               ...        ...              ...             ...
8023  BN-2014-3913645 2014-12-29  Mark Washington         Ireland
8029  BN-2014-8679573 2014-12-30   Dennis Conaway     Netherlands
8043  BN-2014-4140795 2014-12-31  Daniel Hamilton     Netherlands
8044  BN-2014-4140795 2014-12-31  Daniel Hamilton     Netherlands
8045  BN-2014-4140795 2014-12-31  Daniel Hamilton     Netherlands

                                   State        City   Region
Segment  \
0                              Stockholm   Stockholm    North  Home
Office
3                                England  Birmingham    North
Corporate
4                                England  Birmingham    North
Corporate
10    Languedoc-Roussillon-Midi-Pyrénées    Toulouse  Central
Consumer
11    Languedoc-Roussillon-Midi-Pyrénées    Toulouse  Central
Consumer
...                                  ...         ...      ...
...
8023                              Dublin      Dublin    North  Home
Office
8029                       South Holland   The Hague  Central
Consumer
8043                       North Brabant   Eindhoven  Central  Home
Office
8044                       North Brabant   Eindhoven  Central  Home
Office
8045                       North Brabant   Eindhoven  Central  Home
Office

         Ship Mode         Category Sub-Category  \
0     Economy Plus  Office Supplies        Paper
3          Economy  Office Supplies          Art
4          Economy  Office Supplies      Storage
10         Economy        Furniture    Bookcases
```

```
11         Economy  Office Supplies          Art
...            ...              ...          ...
8023       Economy       Technology      Copiers
8029      Priority  Office Supplies   Appliances
8043  Economy Plus  Office Supplies          Art
8044  Economy Plus  Office Supplies      Binders
8045  Economy Plus       Technology     Machines

                           Product Name  Discount  Sales  Profit
Quantity  \
0          Enermax Note Cards, Premium       0.5     45     -26
3
3           Boston Markers, Easy-Erase       0.5     27     -22
2
4          Eldon Folders, Single Width       0.5     17      -1
2
10        Ikea Classic Bookcase, Metal       0.6    987   -1012
6
11      Binney & Smith Sketch Pad, Blue       0.5    116     -56
5
...                                 ...       ...    ...     ...
...
8023                  Sharp Ink, Laser       0.5    373    -254
6
8029            Cuisinart Blender, Silver       0.5     68     -62
2
8043  BIC Pencil Sharpener, Fluorescent       0.5     30     -10
2
8044       Avery Binder Covers, Recycled       0.5     23      -6
4
8045                StarTech Phone, Red       0.5    108     -19
3

      Feedback?
0         False
3          True
4          True
10         True
11        False
...         ...
8023      False
8029      False
8043      False
8044       True
8045      False

[1426 rows x 17 columns]
```

```
#Outlier Treatment
df['Discount'] = df['Discount'].apply(lambda x: ub_dis if x > ub_dis
else (lb_dis if x < lb_dis else x))
```

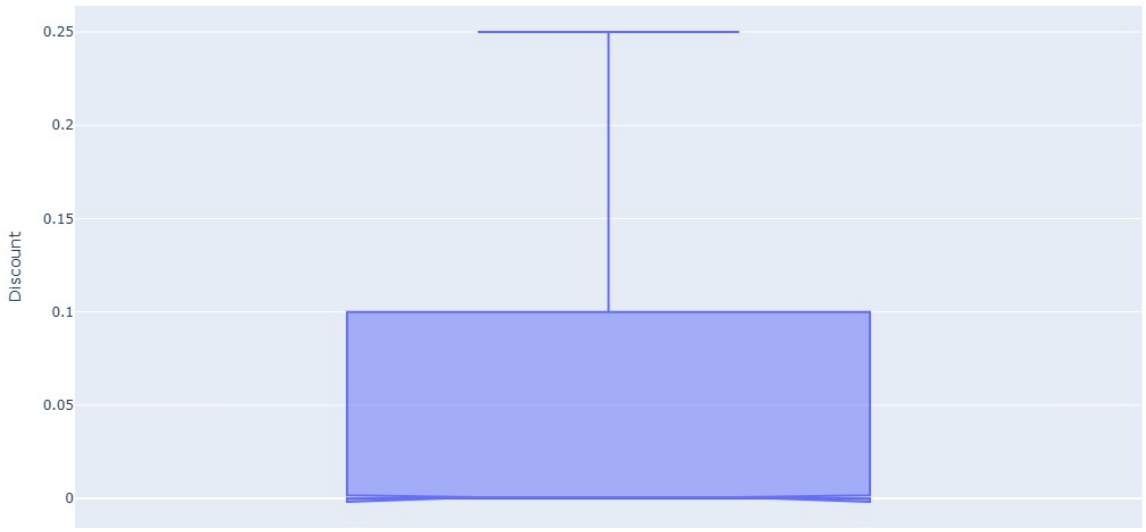
```
min_dis = df['Discount'].min()
min_dis
```

```
0.0
```

```
max_dis = df['Discount'].max()
max_dis
```

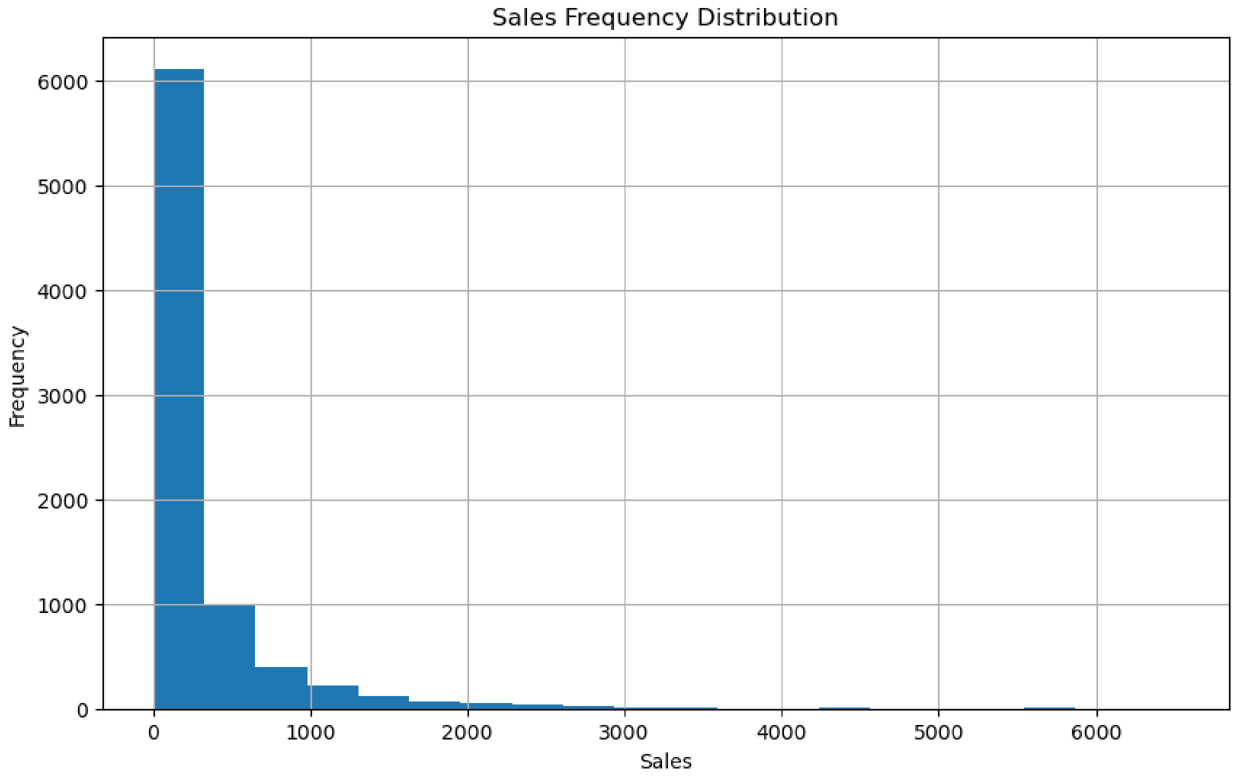
```
0.25
```

```
fig = px.box(df, y = 'Discount', notched = True, width=800,
height=600)
fig.show()
```

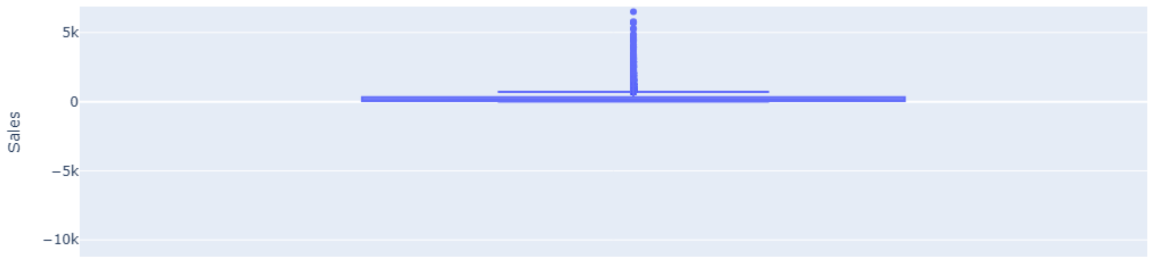


For Sales

```
df['Sales'].hist(bins = 20, figsize= (10,6))
plt.title('Sales Frequency Distribution')
plt.xlabel('Sales')
plt.ylabel('Frequency')
plt.show()
```



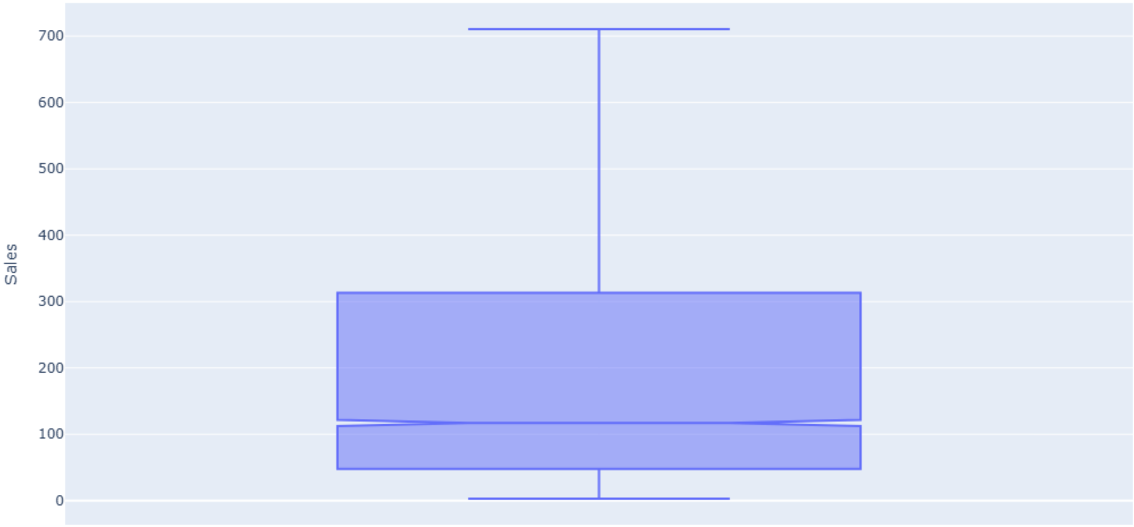
```
#identifying the outliers
fig = px.box(df, y = 'Sales', notched = True)
fig.show()
```



```
q1_sales = df['Sales'].quantile(0.25)
q2_sales = df['Sales'].quantile(0.5)
q3_sales = df['Sales'].quantile(0.75)
iqr_sales = q3_sales - q1_sales

lb_sales = (q1_sales - 1.5 * iqr_sales)
lb_sales
```

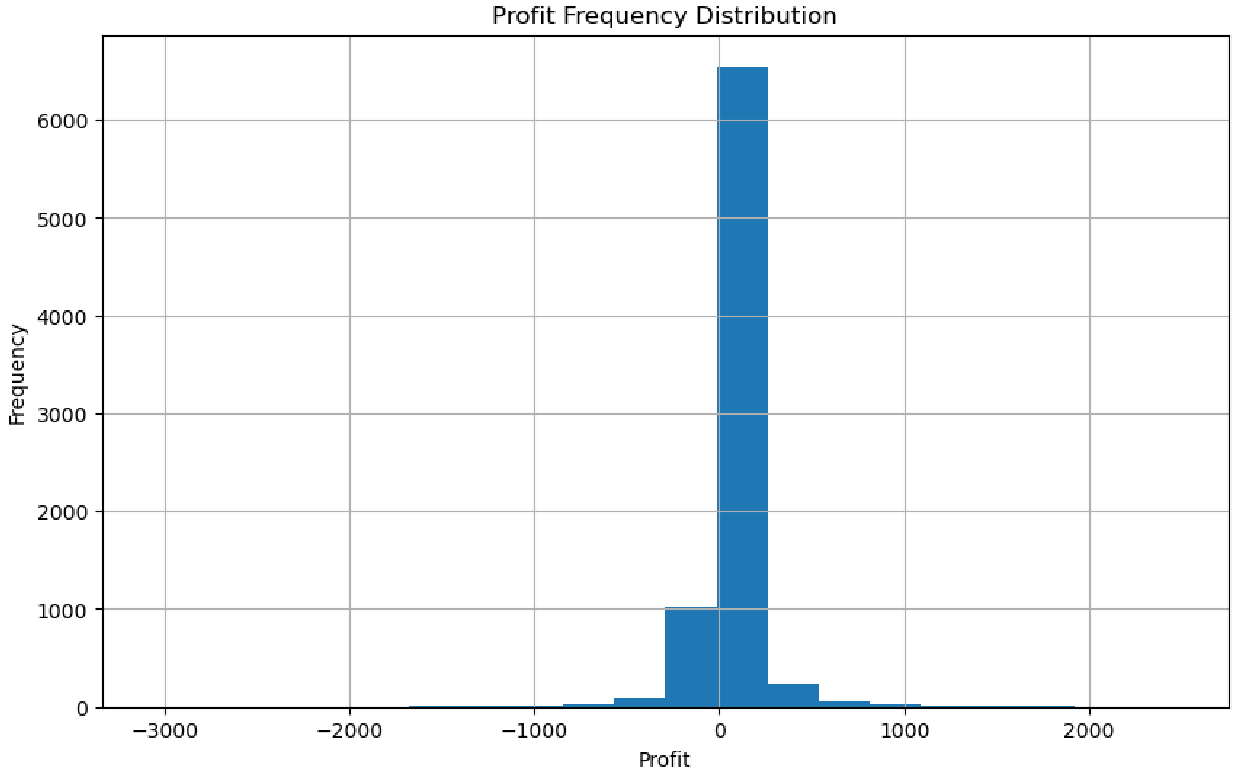
```
-349.5
```

```
lb_sales = max(0, q1_sales - 1.5 * iqr_sales)
lb_sales

0

ub_sales = q3_sales + 1.5*iqr_sales
ub_sales

710.5

#outlier treatment
df['Sales'] = df['Sales'].apply(lambda x: ub_sales if x > ub_sales
else (lb_sales if x < lb_sales else x))

min_sales = df['Sales'].min()
min_sales

3.0

max_sales = df['Sales'].max()
max_sales

710.5

fig = px.box(df, y = 'Sales', notched = True, width=800, height=600)
fig.show()
```

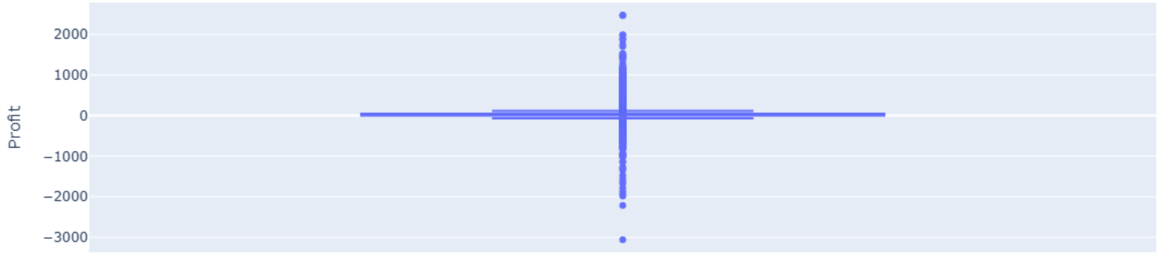


For Profit

```
df['Profit'].hist(bins = 20, figsize= (10,6))
plt.title('Profit Frequency Distribution')
plt.xlabel('Profit')
plt.ylabel('Frequency')
plt.show()
```



```
#identifying the outliers
fig = px.box(df, y = 'Profit', notched = True)
fig.show()
```



```
q1_profit = df['Profit'].quantile(0.25)
q2_profit = df['Profit'].quantile(0.5)
```

```
q3_profit = df['Profit'].quantile(0.75)
iqr_profit = q3_profit - q1_profit
```

```
lb_profit = q1_profit - 1.5*iqr_profit
lb_profit
```

```
-68.0
```

```
ub_profit = q3_profit + 1.5*iqr_profit
ub_profit
```

```
116.0
```

```
#outlier treatment
df['Profit'] = df['Profit'].apply(lambda x: ub_profit if x > ub_profit
else (lb_profit if x < lb_profit else x))
```

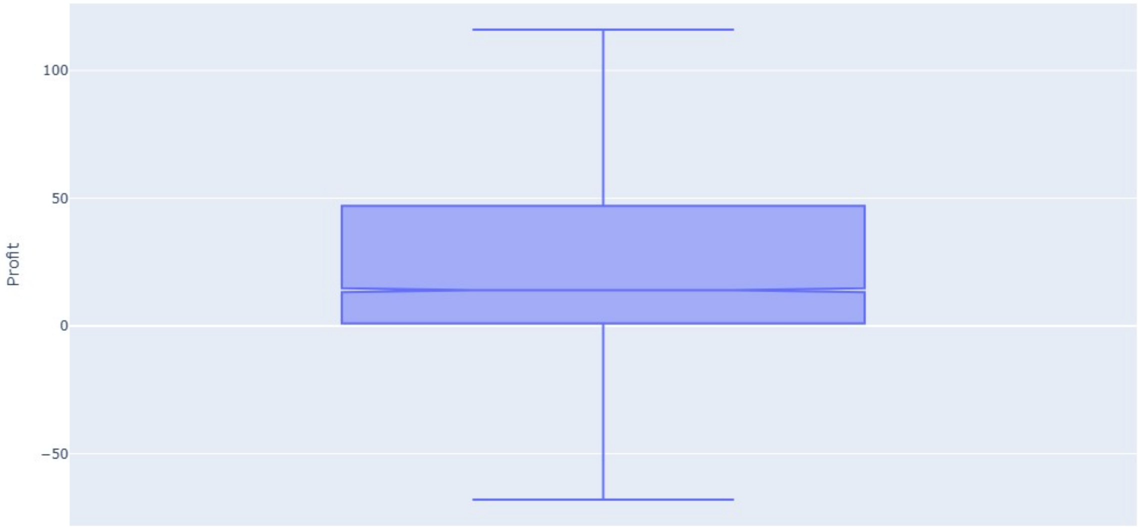
```
min_profit = df['Profit'].min()
min_profit
```

```
-68.0
```

```
max_profit = df['Profit'].max()
max_profit
```

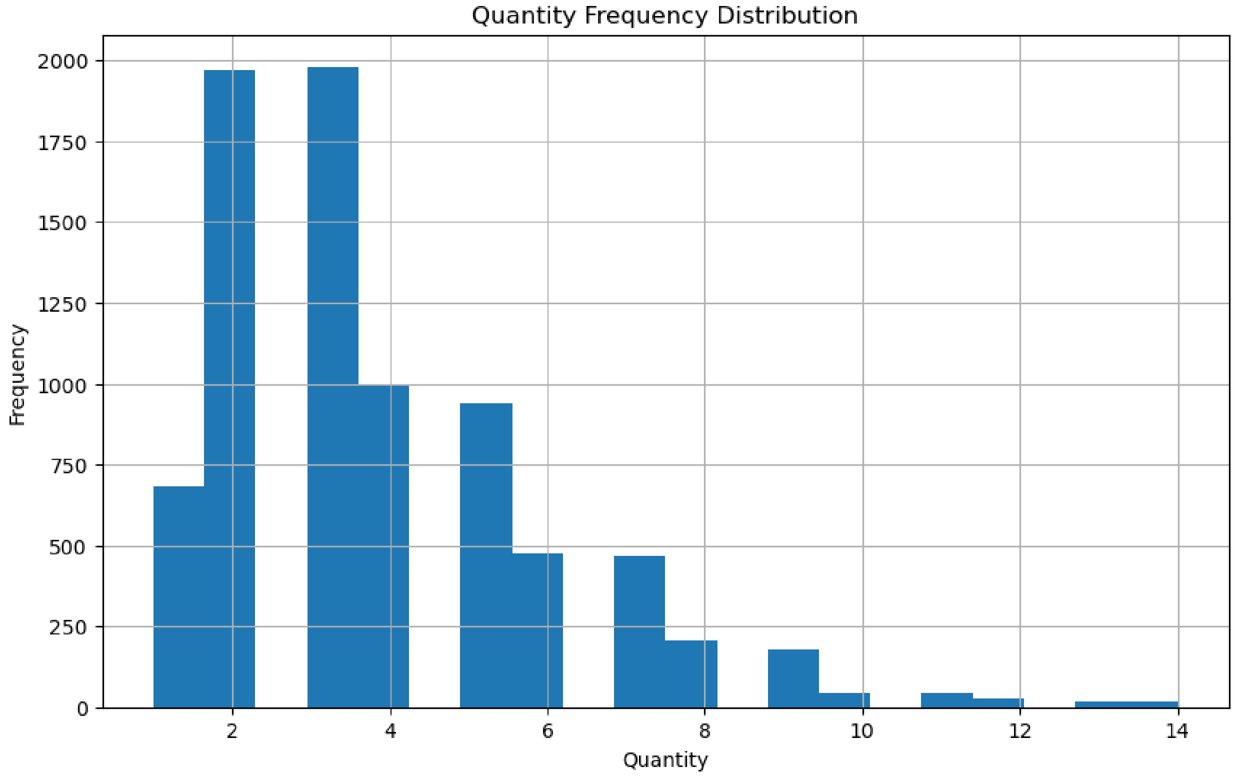
```
116.0
```

```
fig = px.box(df, y = 'Profit', notched = True, width=800, height=600)
fig.show()
```

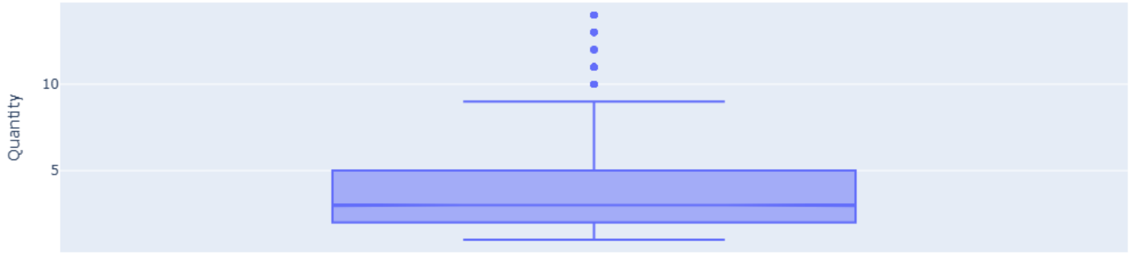


For Quantity

```
df['Quantity'].hist(bins = 20, figsize= (10,6))
plt.title('Quantity Frequency Distribution')
plt.xlabel('Quantity')
plt.ylabel('Frequency')
plt.show()
```



```
#identifying the outliers
fig = px.box(df, y = 'Quantity', notched = True)
fig.show()
```



```
q1_quan = df['Quantity'].quantile(0.25)
q2_quan = df['Quantity'].quantile(0.5)
```

```python
q3_quan = df['Quantity'].quantile(0.75)
iqr_quan = q3_quan - q1_quan
```

```python
lb_quan = q1_quan - 1.5*iqr_quan
lb_quan
```

```
-2.5
```

```python
lb_quan = max(0, q1_quan - 1.5 * iqr_quan)
lb_quan
```

```
0
```

```python
ub_quan = q3_quan + 1.5*iqr_quan
ub_quan
```

```
9.5
```

```python
#outlier treatment
df['Quantity'] = df['Quantity'].apply(lambda x: ub_quan if x > ub_quan
else (lb_quan if x < lb_quan else x))
```

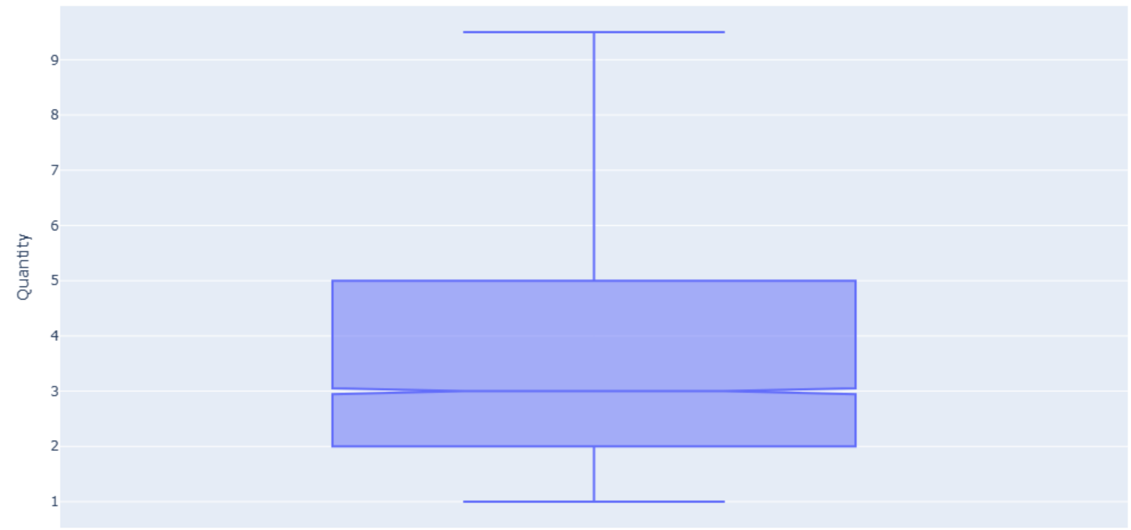
```python
min_quan = df['Quantity'].min()
min_quan
```

```
1.0
```

```python
max_quan = df['Quantity'].max()
max_quan
```

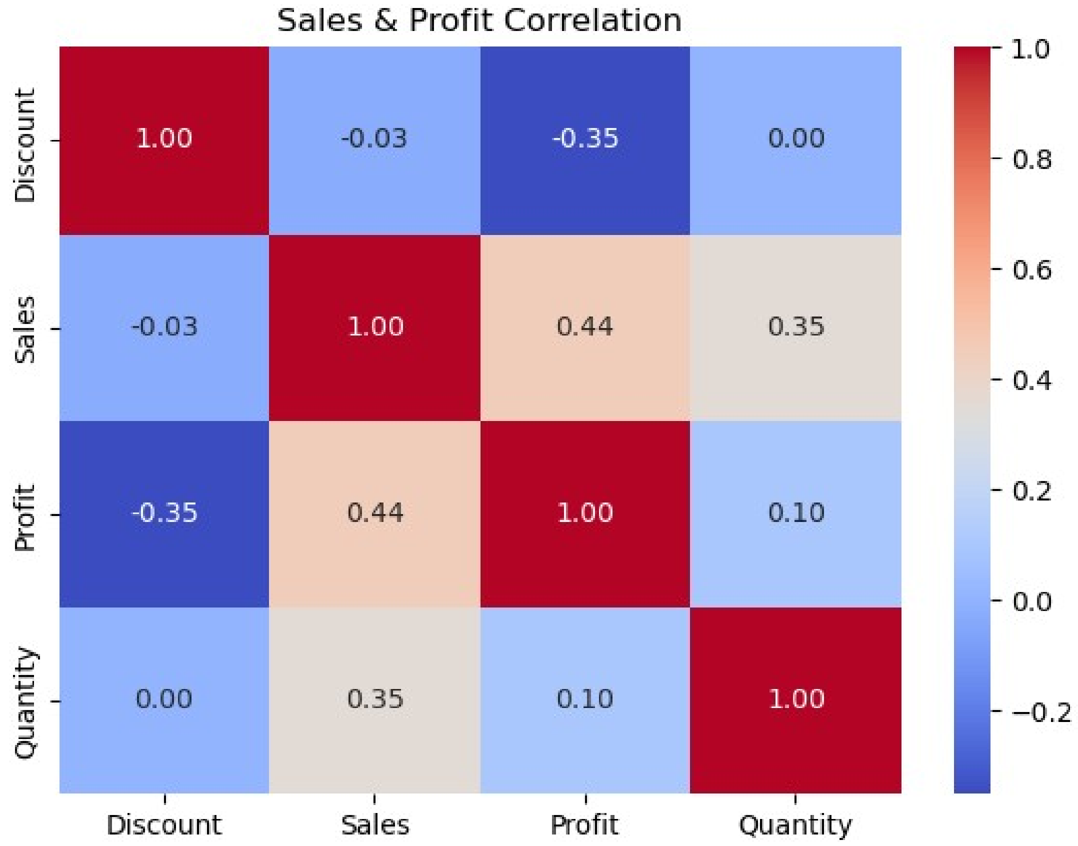
```
9.5
```

```python
fig = px.box(df, y = 'Quantity', notched = True, width=800,
height=600)
fig.show()
```



```
df.describe()

                                 Order Date     Discount        Sales
Profit  \
count                                  8047  8047.000000  8047.000000
8047.000000
mean   2013-04-19 12:25:40.748104704     0.068727   218.702498
24.168137
min              2011-01-01 00:00:00     0.000000     3.000000   -
68.000000
25%              2012-06-08 00:00:00     0.000000    48.000000
1.000000
50%              2013-06-11 00:00:00     0.000000   117.000000
14.000000
75%              2014-04-30 00:00:00     0.100000   313.000000
47.000000
max              2014-12-31 00:00:00     0.250000   710.500000
116.000000
std                                     NaN     0.096626   227.544531
48.971611

          Quantity
count  8047.000000
mean      3.735119
min       1.000000
25%       2.000000
50%       3.000000
75%       5.000000
```

```
max       9.500000
std       2.079027

#correlation between the columns having numerical values
df_numeric = df.select_dtypes(include=['number'])
correlation_matrix = df_numeric.corr()
correlation_matrix

          Discount     Sales    Profit  Quantity
Discount  1.000000  0.052092 -0.499028  0.000088
Sales     0.052092  1.000000  0.442246  0.363168
Profit   -0.499028  0.442246  1.000000  0.172306
Quantity  0.000088  0.363168  0.172306  1.000000

#Sales & Profit Correlation
plt.figure(figsize=(7, 5))
correlation_matrix = df.select_dtypes(include=['number']).corr()  # 
Ensuring only numeric columns are used
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', 
fmt=".2f")  # fmt=".2f" limits decimal places
plt.title("Sales & Profit Correlation")
plt.show()
```

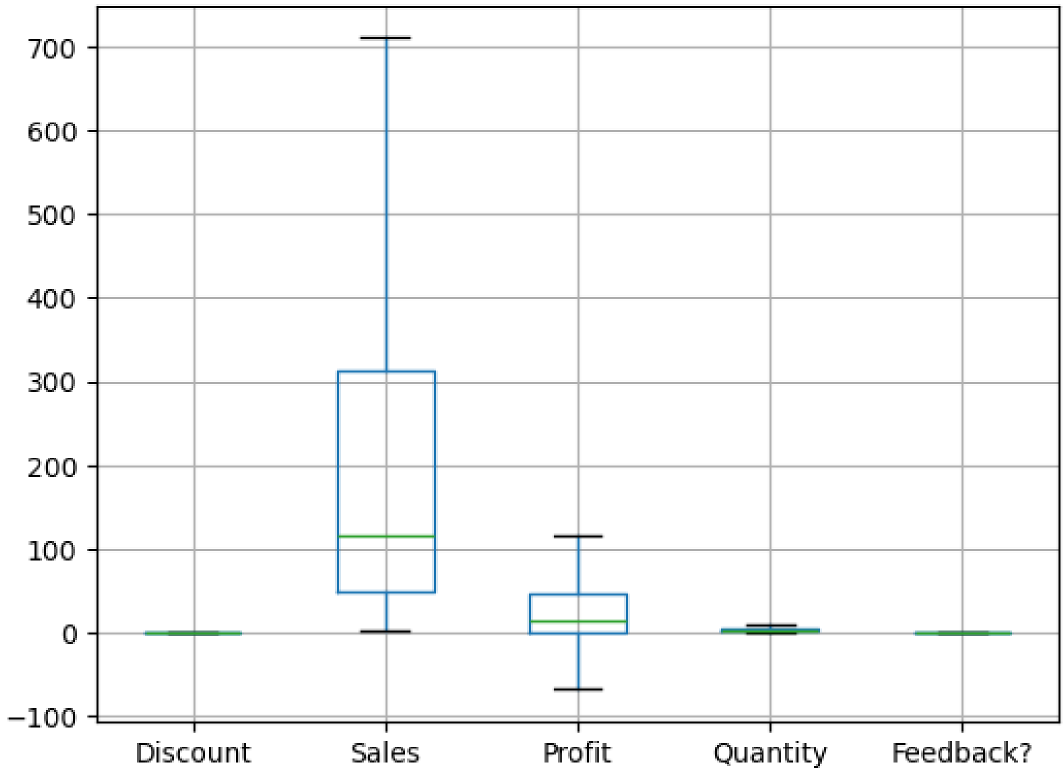


Sales and Profit are positively correlated (0.44), meaning higher sales generally lead to higher profit. Discount has a negative correlation (-0.49) with profit, meaning increasing discounts reduces profitability. Quantity and Profit have a weak correlation (0.17), indicating selling more units does not always mean higher profits.

Data Visualization

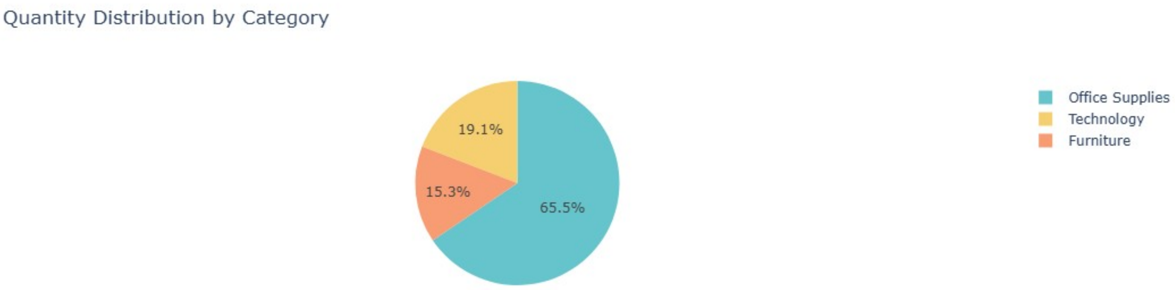
```
df.boxplot()
```

```
<Axes: >
```



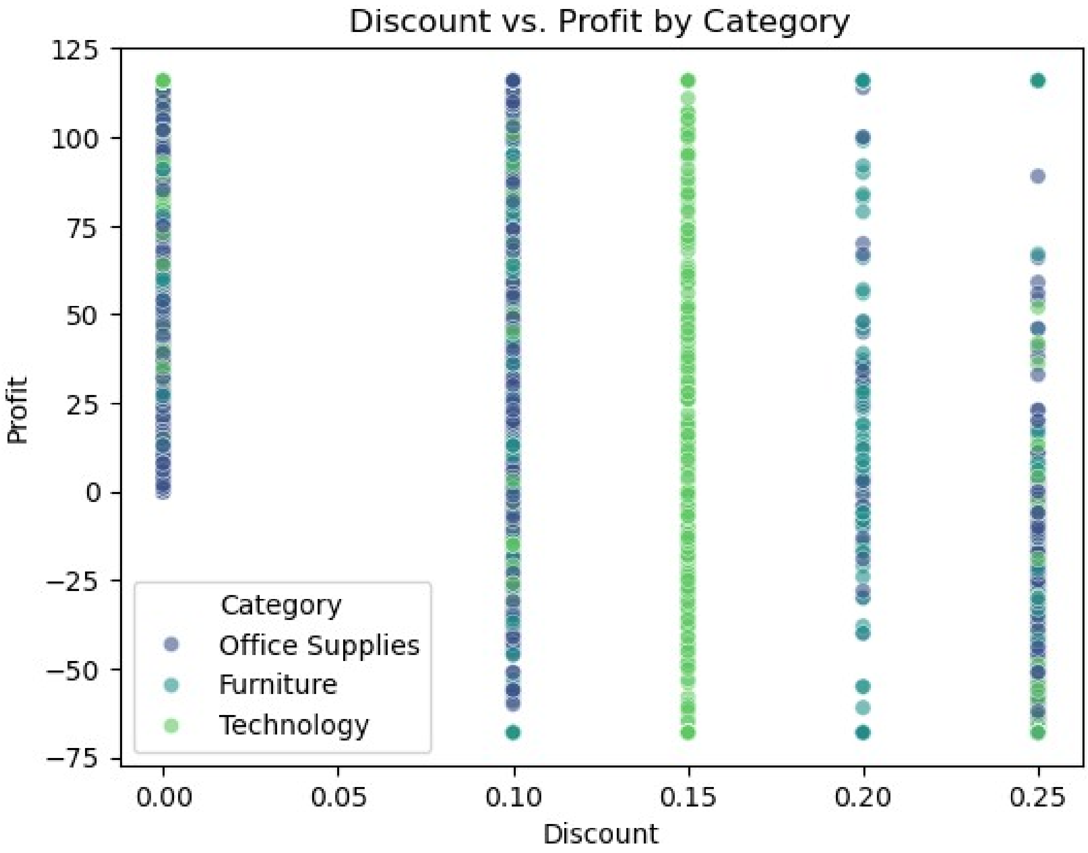
```
pip install seaborn --upgrade
```

```
Defaulting to user installation because normal site-packages is not writeableNote: you may need to restart the kernel to use updated packages.

Requirement already satisfied: seaborn in d:\anaconda\lib\site-packages (0.13.2)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in d:\anaconda\lib\site-packages (from seaborn) (1.26.4)
Requirement already satisfied: pandas>=1.2 in d:\anaconda\lib\site-packages (from seaborn) (2.2.2)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in d:\anaconda\lib\site-packages (from seaborn) (3.8.4)
Requirement already satisfied: contourpy>=1.0.1 in d:\anaconda\lib\
site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (1.2.0)
Requirement already satisfied: cycler>=0.10 in d:\anaconda\lib\site-
packages (from matplotlib!=3.6.1,>=3.4->seaborn) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in d:\anaconda\lib\
site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (4.51.0)
Requirement already satisfied: kiwisolver>=1.3.1 in d:\anaconda\lib\
site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (1.4.4)
Requirement already satisfied: packaging>=20.0 in d:\anaconda\lib\
site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (23.2)
Requirement already satisfied: pillow>=8 in d:\anaconda\lib\site-
packages (from matplotlib!=3.6.1,>=3.4->seaborn) (10.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in d:\anaconda\lib\
site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in d:\anaconda\
lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn)
(2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in d:\anaconda\lib\site-
packages (from pandas>=1.2->seaborn) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in d:\anaconda\lib\site-
packages (from pandas>=1.2->seaborn) (2023.3)
Requirement already satisfied: six>=1.5 in d:\anaconda\lib\site-
packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=3.4->seaborn)
(1.16.0)

#Quantity distribution by category
category_quantity = df.groupby('Category')
['Quantity'].sum().reset_index()
fig = px.pie(category_quantity, names='Category', values='Quantity',
title='Quantity Distribution by Category',
color_discrete_sequence=px.colors.qualitative.Pastel)
fig.show()
```



The pie chart shows the quantity percentage distribution amongst the categories.

```
#Discount Impact on Profit
sns.scatterplot(data=df, x='Discount', y='Profit', hue='Category',
alpha=0.6, palette='viridis')
plt.title("Discount vs. Profit by Category")
plt.xlabel("Discount")
plt.ylabel("Profit")
plt.show()
```

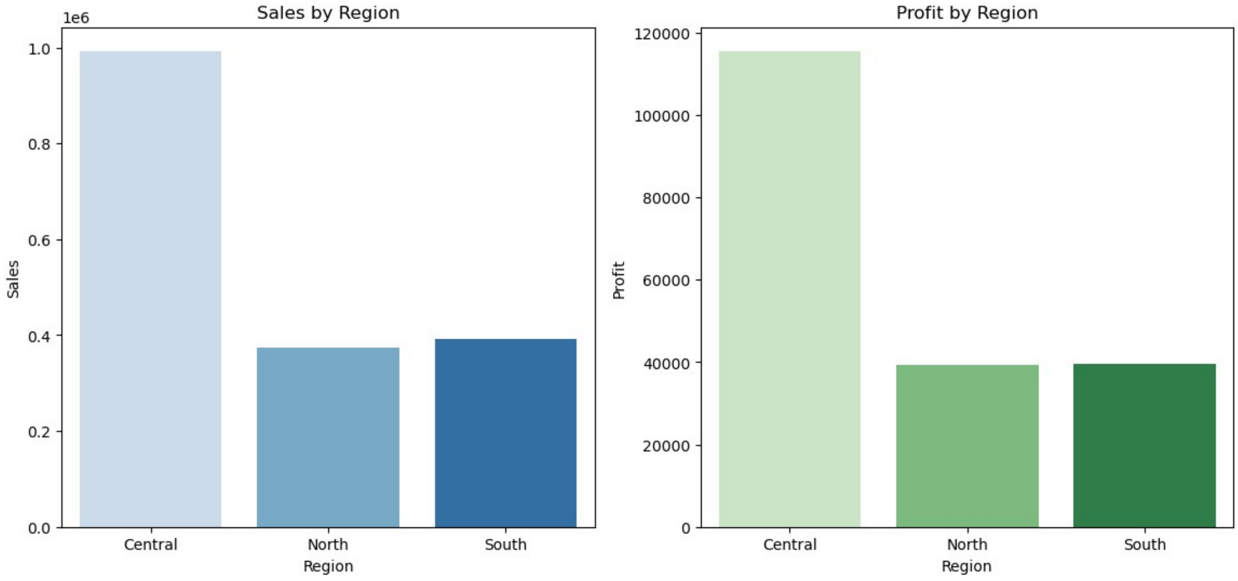


Products with high discounts (> 25%) tend to have negative profits. Some categories, especially Office Supplies and Technology, show major losses when discounts are high.

```
#Sales and Profit by Region
region_sales_profit = df.groupby('Region')[['Sales',
'Profit']].sum().reset_index()

fig, ax = plt.subplots(1, 2, figsize=(14, 6))
sns.barplot(data=region_sales_profit, x='Region', y='Sales', ax=ax[0],
palette='Blues')
ax[0].set_title("Sales by Region")
ax[0].set_xlabel("Region")
ax[0].set_ylabel("Sales")

sns.barplot(data=region_sales_profit, x='Region', y='Profit',
ax=ax[1], palette='Greens')
ax[1].set_title("Profit by Region")
```

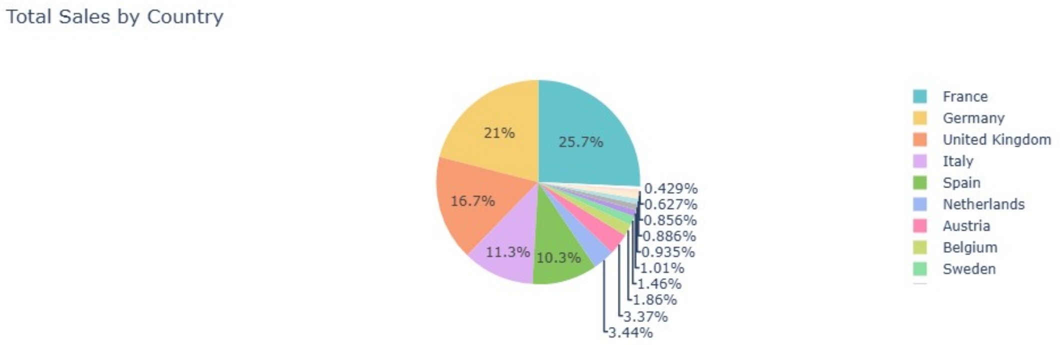
```
ax[1].set_xlabel("Region")
ax[1].set_ylabel("Profit")

plt.show()
```

```
C:\Users\KIIT\AppData\Local\Temp\ipykernel_19476\828807372.py:4:
FutureWarning:



Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.


C:\Users\KIIT\AppData\Local\Temp\ipykernel_19476\828807372.py:9:
FutureWarning:



Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.
```



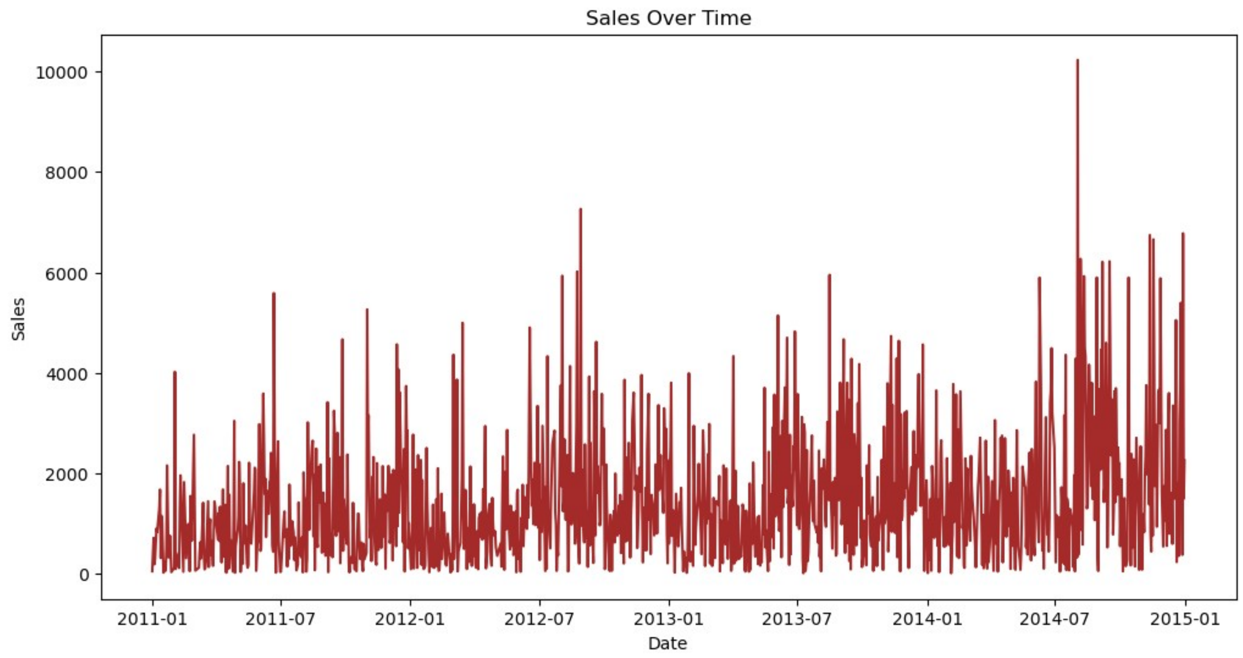
The North region has the highest sales and profit. The Central region shows decent sales but low profit, meaning operational costs may be higher.

```
#Sales by Country distribution
category_quantity = df.groupby('Country')['Sales'].sum().reset_index()
fig = px.pie(category_quantity, names='Country', values='Sales',
title='Total Sales by Country',
color_discrete_sequence=px.colors.qualitative.Pastel)
fig.show()
```



United Kingdom, Germany and France contribute the most to sales. Smaller markets like Netherlands and Ireland have minimal sales impact.

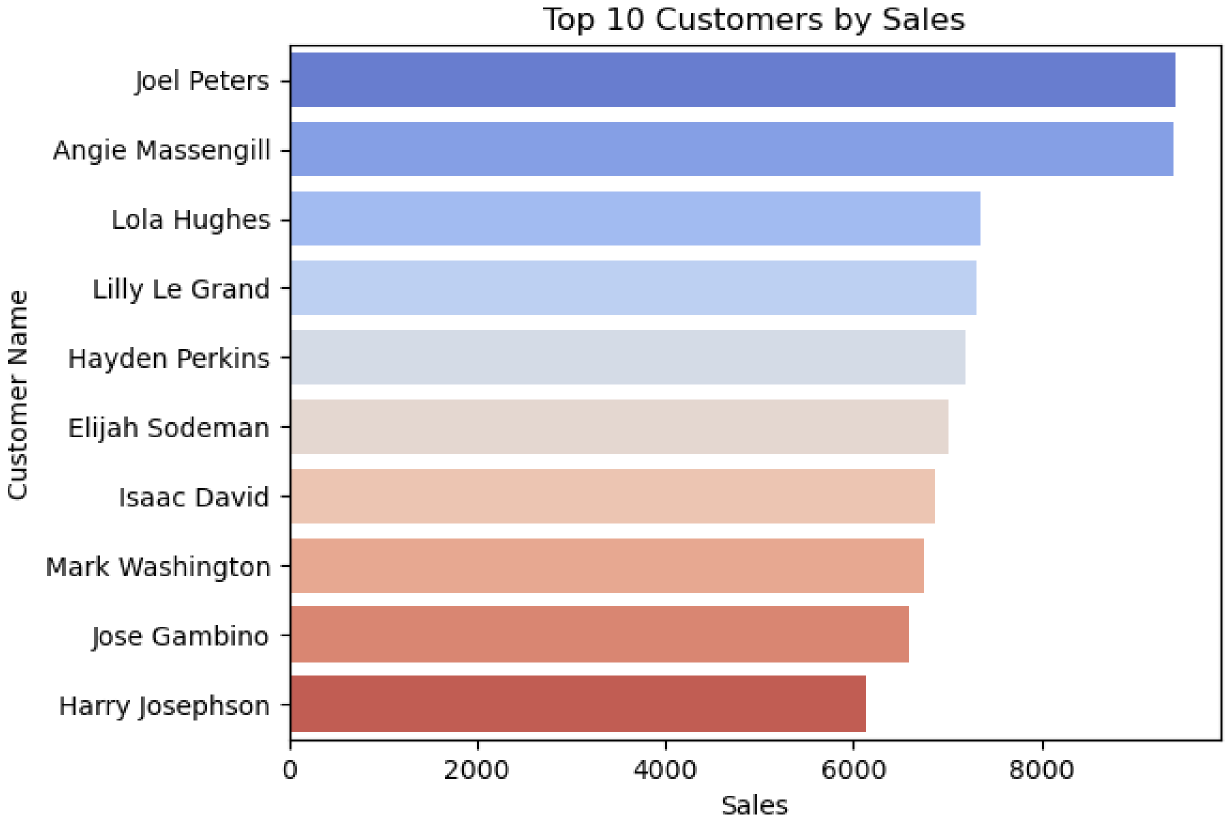
```
#Time Series Analysis
df['Order Date'] = pd.to_datetime(df['Order Date'])  # Ensure it's in
datetime format
sales_over_time = df.groupby('Order Date')
['Sales'].sum().reset_index()

plt.figure(figsize=(12, 6))
sns.lineplot(data=sales_over_time, x='Order Date', y='Sales',
color='brown')
plt.title("Sales Over Time")
plt.xlabel("Date")
plt.ylabel("Sales")
plt.show()
```



Sales show seasonal trends, with peaks in certain months.

```
#Top Customers by Sales
top_customers = df.groupby('Customer Name')
['Sales'].sum().nlargest(10).reset_index()
sns.barplot(data=top_customers, x='Sales', y='Customer Name',
palette='coolwarm')
plt.title("Top 10 Customers by Sales")
plt.xlabel("Sales")
plt.ylabel("Customer Name")
plt.show()
```

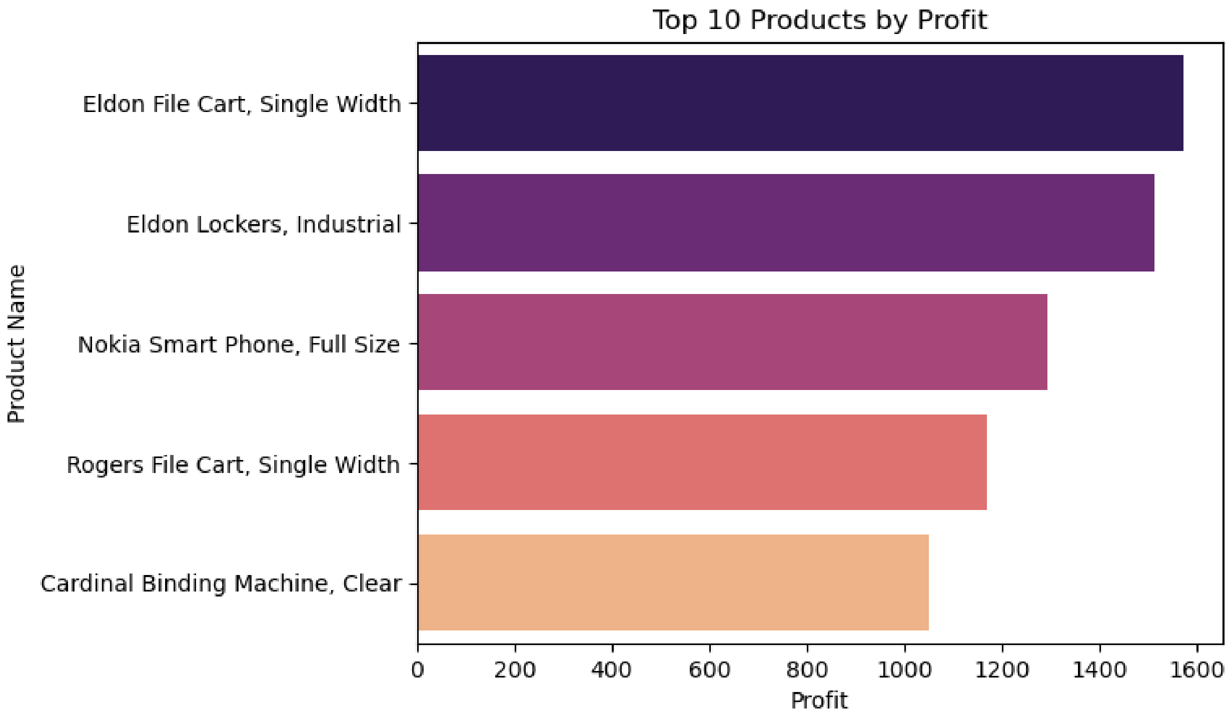
```
C:\Users\KIIT\AppData\Local\Temp\ipykernel_19476\1941503435.py:3:
FutureWarning:



Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `y` variable to `hue` and set
`legend=False` for the same effect.
```



A few key customers contribute a large portion of total sales. These top customers likely make repeated high-value purchases.

```python
#Top 5 products by Profit
top_products = df.groupby('Product Name')
['Profit'].sum().nlargest(5).reset_index()
sns.barplot(data=top_products, x='Profit', y='Product Name',
palette='magma')
plt.title("Top 10 Products by Profit")
plt.xlabel("Profit")
plt.ylabel("Product Name")
plt.show()
```

```
C:\Users\KIIT\AppData\Local\Temp\ipykernel_19476\315273585.py:3:
FutureWarning:



Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `y` variable to `hue` and set
`legend=False` for the same effect.
```



A few products generate the highest profits, indicating high-margin items. The most profitable products belong to Technology and Office Supplies categories.

Key Findings

Discounts negatively impact profit – A strong negative correlation (-0.49) indicates that excessive discounts reduce profitability. Optimizing discount strategies is crucial.

Some high-sales products generate minimal or negative profit, indicating a need for better pricing strategies.

The North region generates the highest sales and profit, while other regions need cost optimizations and better sales strategies.

There are clear end-of-year sales spikes, highlighting opportunities to align marketing efforts with peak demand.

A small group of customers accounts for a large portion of revenue, emphasizing the need for loyalty programs and personalized promotions.

These product categories have higher profitability, suggesting a focus on inventory management and targeted marketing for them.