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

data = pd.read_csv("Global_Superstore.csv",encoding='latin1')
#import pandas as pd

#file_path = "Global_Superstore.csv"
#data = pd.read_csv(file_path, encoding='latin1')

data
```

```
       Row ID         Order ID  Order Date   Ship Date       Ship Mode
\
0       32298   CA-2012-124891  31-07-2012  31-07-2012        Same Day

1       26341    IN-2013-77878  05-02-2013  07-02-2013    Second Class

2       25330    IN-2013-71249  17-10-2013  18-10-2013     First Class

3       13524  ES-2013-1579342  28-01-2013  30-01-2013     First Class

4       47221     SG-2013-4320  05-11-2013  06-11-2013        Same Day

...       ...              ...         ...         ...             ...

51285   29002    IN-2014-62366  19-06-2014  19-06-2014        Same Day

51286   35398   US-2014-102288  20-06-2014  24-06-2014  Standard Class

51287   40470   US-2013-155768  02-12-2013  02-12-2013        Same Day

51288    9596   MX-2012-140767  18-02-2012  22-02-2012  Standard Class

51289    6147   MX-2012-134460  22-05-2012  26-05-2012    Second Class


      Customer ID     Customer Name      Segment           City  \
0        RH-19495       Rick Hansen     Consumer  New York City
1        JR-16210     Justin Ritter    Corporate     Wollongong
2        CR-12730      Craig Reiter     Consumer       Brisbane
3        KM-16375  Katherine Murray  Home Office         Berlin
4         RH-9495       Rick Hansen     Consumer          Dakar
...           ...               ...         ...            ...
51285    KE-16420   Katrina Edelman    Corporate           Kure
51286    ZC-21910  Zuschuss Carroll     Consumer        Houston
51287    LB-16795    Laurel Beltran  Home Office         Oxnard
51288    RB-19795        Ross Baird  Home Office       Valinhos
51289    MC-18100     Mick Crebagga     Consumer       Tipitapa
```

```
                  State  ...        Product ID         Category Sub-
Category  \
0             New York  ...   TEC-AC-10003033       Technology
Accessories
1      New South Wales  ...   FUR-CH-10003950        Furniture
Chairs
2           Queensland  ...   TEC-PH-10004664       Technology
Phones
3               Berlin  ...   TEC-PH-10004583       Technology
Phones
4                Dakar  ...  TEC-SHA-10000501       Technology
Copiers
...                ...  ...               ...              ...
...
51285        Hiroshima  ...   OFF-FA-10000746  Office Supplies
Fasteners
51286            Texas  ...   OFF-AP-10002906  Office Supplies
Appliances
51287       California  ...   OFF-EN-10001219  Office Supplies
Envelopes
51288        São Paulo  ...   OFF-BI-10000806  Office Supplies
Binders
51289          Managua  ...   OFF-PA-10004155  Office Supplies
Paper

                                              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
...                                                 ...       ...
...
51285                     Advantus Thumb Tacks, 12 Pack    65.100
5
51286  Hoover Replacement Belt for Commercial Guardsm...     0.444
1
51287       #10- 4 1/8" x 9 1/2" Security-Tint Envelopes    22.920
3
51288                            Acco Index Tab, Economy    13.440
2
51289            Eaton Computer Printout Paper, 8.5 x 11    61.380
3
```

```
      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
...        ...       ...            ...             ...
51285      0.0    4.5000           0.01          Medium
51286      0.8   -1.1100           0.01          Medium
51287      0.0   11.2308           0.01            High
51288      0.0    2.4000           0.00          Medium
51289      0.0    1.8000           0.00            High

[51290 rows x 24 columns]
```

```
data.describe()
```

```
             Row ID   Postal Code          Sales      Quantity
Discount  \
count  25882.000000  2.588200e+04  25882.000000  25882.000000
25882.000000
mean   22876.987057  5.519038e+04     74.783825      2.735337
0.112225
std    15605.923473  2.182829e-11     73.131964      1.588873
0.186080
min        1.000000  5.519038e+04      1.548000      1.000000
0.000000
25%     9579.250000  5.519038e+04     25.200000      2.000000
0.000000
50%    20329.500000  5.519038e+04     50.376600      2.000000
0.000000
75%    31111.750000  5.519038e+04     99.625875      4.000000
0.170000
max    51290.000000  5.519038e+04    593.730000      7.000000
0.602000

             Profit  Shipping Cost
count  25882.000000   25882.000000
mean      10.113063       6.746957
std       18.980380       6.182084
min      -41.880000       0.000000
25%        0.600000       2.050000
50%        6.840000       4.590000
75%       18.795000       9.580000
max       69.800000      26.960000
```

```
data.info()
```

```
<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
```

```
data.dtypes
```

```
Row ID              int64
Order ID           object
Order Date         object
Ship Date          object
Ship Mode          object
Customer ID        object
Customer Name      object
Segment            object
City               object
State              object
Country            object
Postal Code       float64
Market             object
Region             object
Product ID         object
Category           object
```

```
Sub-Category       object
Product Name       object
Sales             float64
Quantity            int64
Discount          float64
Profit            float64
Shipping Cost     float64
Order Priority     object
dtype: object
```

```
data.isnull().sum()
```

```
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
```

```
data["Postal Code"]
```

```
0        10024.0
1            NaN
2            NaN
3            NaN
4            NaN
           ...
51285        NaN
51286    77095.0
51287    93030.0
51288        NaN
```

```
51289        NaN
Name: Postal Code, Length: 51290, dtype: float64
```

```
data["Postal Code"].mean()
```

```
np.float64(55190.3794276566)
```

```
data["Postal Code"] = data["Postal Code"].fillna(data["Postal
Code"].mean())
```

```
data
```

```
       Row ID         Order ID  Order Date   Ship Date       Ship Mode
\
0       32298   CA-2012-124891  31-07-2012  31-07-2012        Same Day

1       26341    IN-2013-77878  05-02-2013  07-02-2013    Second Class

2       25330    IN-2013-71249  17-10-2013  18-10-2013     First Class

3       13524  ES-2013-1579342  28-01-2013  30-01-2013     First Class

4       47221     SG-2013-4320  05-11-2013  06-11-2013        Same Day

...       ...              ...         ...         ...             ...

51285   29002    IN-2014-62366  19-06-2014  19-06-2014        Same Day

51286   35398   US-2014-102288  20-06-2014  24-06-2014  Standard Class

51287   40470   US-2013-155768  02-12-2013  02-12-2013        Same Day

51288    9596   MX-2012-140767  18-02-2012  22-02-2012  Standard Class

51289    6147   MX-2012-134460  22-05-2012  26-05-2012    Second Class


      Customer ID     Customer Name      Segment           City  \
0        RH-19495       Rick Hansen     Consumer  New York City
1        JR-16210     Justin Ritter    Corporate     Wollongong
2        CR-12730      Craig Reiter     Consumer       Brisbane
3        KM-16375  Katherine Murray  Home Office         Berlin
4         RH-9495       Rick Hansen     Consumer          Dakar
...           ...               ...         ...            ...
51285    KE-16420   Katrina Edelman    Corporate           Kure
51286    ZC-21910  Zuschuss Carroll     Consumer        Houston
51287    LB-16795    Laurel Beltran  Home Office        Oxnard
51288    RB-19795        Ross Baird  Home Office       Valinhos
51289    MC-18100     Mick Crebagga     Consumer       Tipitapa

                 State  ...        Product ID         Category Sub-
Category  \
```

```
0             New York  ...   TEC-AC-10003033       Technology
Accessories
1      New South Wales  ...   FUR-CH-10003950        Furniture
Chairs
2           Queensland  ...   TEC-PH-10004664       Technology
Phones
3               Berlin  ...   TEC-PH-10004583       Technology
Phones
4                Dakar  ...  TEC-SHA-10000501       Technology
Copiers
...                ...  ...               ...              ...
...
51285        Hiroshima  ...   OFF-FA-10000746  Office Supplies
Fasteners
51286            Texas  ...   OFF-AP-10002906  Office Supplies
Appliances
51287       California  ...   OFF-EN-10001219  Office Supplies
Envelopes
51288        São Paulo  ...   OFF-BI-10000806  Office Supplies
Binders
51289          Managua  ...   OFF-PA-10004155  Office Supplies
Paper

                                              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
...                                                 ...       ...
...
51285                     Advantus Thumb Tacks, 12 Pack    65.100
5
51286  Hoover Replacement Belt for Commercial Guardsm...     0.444
1
51287       #10- 4 1/8" x 9 1/2" Security-Tint Envelopes    22.920
3
51288                             Acco Index Tab, Economy    13.440
2
51289            Eaton Computer Printout Paper, 8.5 x 11    61.380
3

      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
...        ...       ...             ...             ...
51285      0.0    4.5000            0.01          Medium
51286      0.8   -1.1100            0.01          Medium
51287      0.0   11.2308            0.01            High
51288      0.0    2.4000            0.00          Medium
51289      0.0    1.8000            0.00            High

[51290 rows x 24 columns]
```

```
data.isnull().sum()
```

```
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       0
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
```

```
data.duplicated().sum()
```

```
np.int64(0)
```

```
def detect_outliers_iqr(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
```

```
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df[(df[column] < lower_bound) | (df[column] > upper_bound)]

numerical_columns = data.select_dtypes(include=np.number).columns
for col in numerical_columns:
    outliers = detect_outliers_iqr(data, col)
    if not outliers.empty:
        print(f"Outliers detected in '{col}': {len(outliers)} rows")
        data = data[~data.index.isin(outliers.index)]
```

```
Outliers detected in 'Postal Code': 9994 rows
Outliers detected in 'Sales': 4548 rows
Outliers detected in 'Quantity': 1796 rows
Outliers detected in 'Discount': 1322 rows
Outliers detected in 'Profit': 5336 rows
Outliers detected in 'Shipping Cost': 2412 rows
```

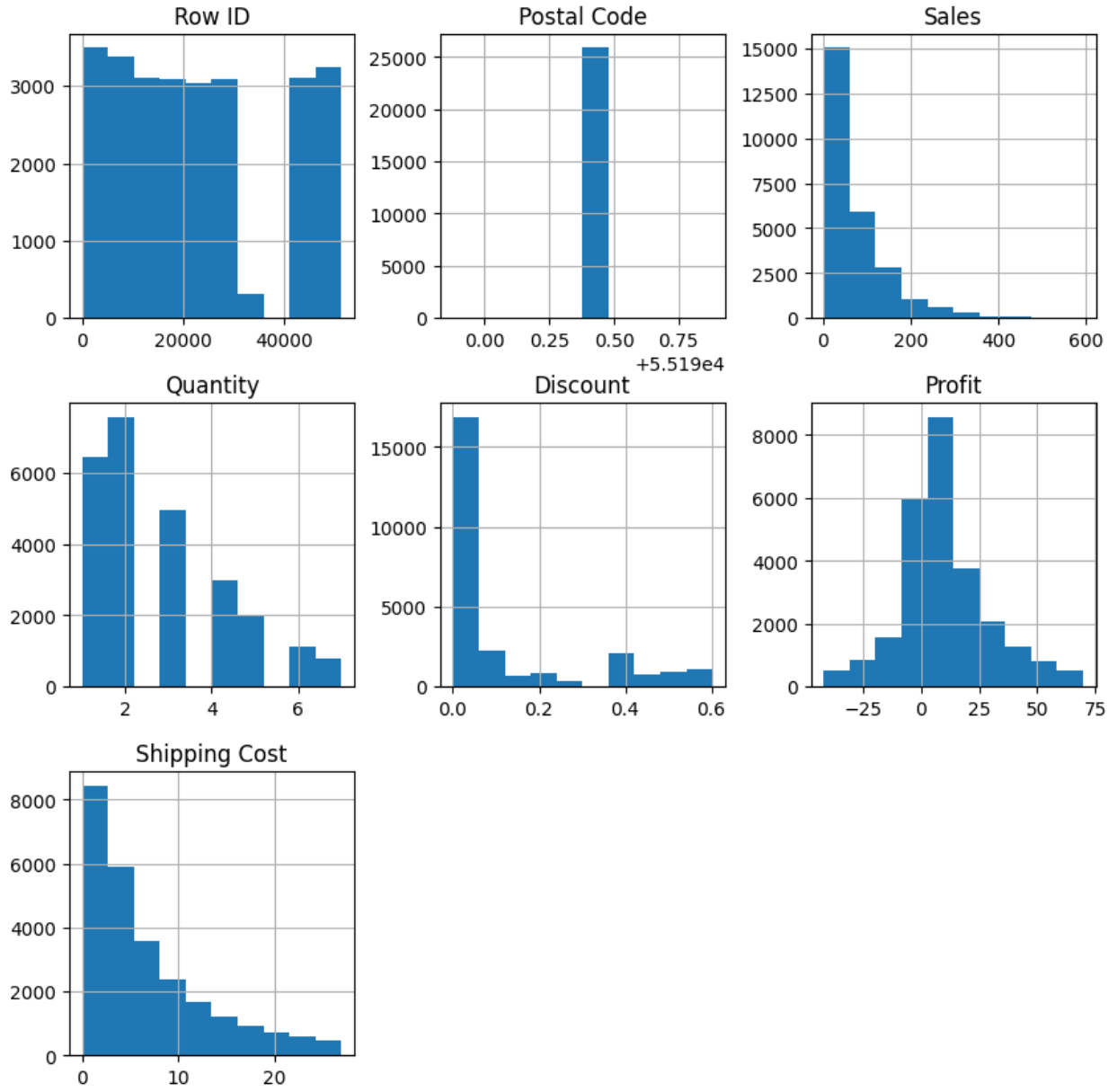
```
correlations=data.corr
```

```
correlations
```

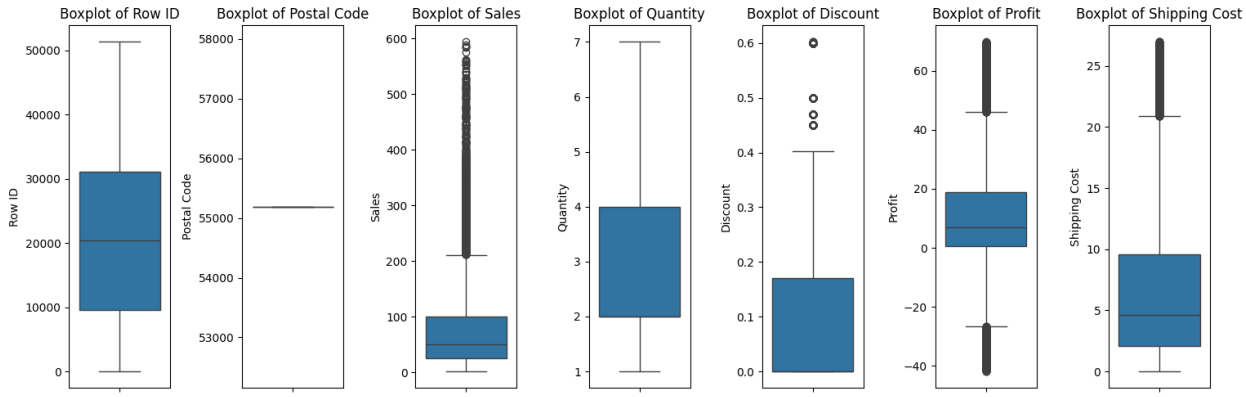
```
<bound method DataFrame.corr of        Row ID         Order ID  Order
Date   Ship Date       Ship Mode  \
11866    9427   MX-2013-111003  14-02-2013  19-02-2013    Second Class

11867   12307  ES-2014-2835815  02-09-2014  08-09-2014  Standard Class

11868   29218    IN-2014-55912  09-08-2014  13-08-2014  Standard Class

11870    1860   MX-2014-162180  21-10-2014  21-10-2014        Same Day

11871   20903    IN-2011-33253  07-09-2011  09-09-2011     First Class

...       ...              ...         ...         ...             ...

51283   24105    IN-2014-72327  30-05-2014  30-05-2014        Same Day

51284   24175    IN-2014-57662  05-08-2014  10-08-2014  Standard Class

51285   29002    IN-2014-62366  19-06-2014  19-06-2014        Same Day

51288    9596   MX-2012-140767  18-02-2012  22-02-2012  Standard Class

51289    6147   MX-2012-134460  22-05-2012  26-05-2012    Second Class


      Customer ID      Customer Name      Segment         City
State  \
11866    SV-20785   Stewart Visinsky     Consumer  Bucaramanga
Santander
```

```
11867    SW-20275   Scott Williamson     Consumer   Manchester
England
11868    RR-19315       Ralph Ritter     Consumer       Raipur
Chhattisgarh
11870    NG-18430      Nathan Gelder     Consumer    São Paulo
São Paulo
11871    PB-19210     Phillip Breyer    Corporate      Geelong
Victoria
...           ...                ...          ...          ...
...
51283    KH-16330    Katharine Harms    Corporate      Lucknow  Uttar
Pradesh
51284    DB-13270  Deborah Brumfield  Home Office   Townsville
Queensland
51285    KE-16420    Katrina Edelman    Corporate         Kure
Hiroshima
51288    RB-19795         Ross Baird  Home Office     Valinhos
São Paulo
51289    MC-18100      Mick Crebagga     Consumer     Tipitapa
Managua

       ...       Product ID         Category Sub-Category  \
11866  ...  TEC-MA-10002080       Technology     Machines
11867  ...  OFF-ST-10000648  Office Supplies      Storage
11868  ...  FUR-CH-10000026        Furniture       Chairs
11870  ...  TEC-CO-10000917       Technology      Copiers
11871  ...  OFF-BI-10004589  Office Supplies      Binders
...    ...              ...              ...          ...
51283  ...  OFF-PA-10000215  Office Supplies        Paper
51284  ...  OFF-BI-10002424  Office Supplies      Binders
51285  ...  OFF-FA-10000746  Office Supplies    Fasteners
51288  ...  OFF-BI-10000806  Office Supplies      Binders
51289  ...  OFF-PA-10004155  Office Supplies        Paper

                                  Product Name      Sales Quantity
Discount  \
11866                    Epson Printer, Durable  175.16000        1
0.000
11867               Eldon File Cart, Industrial  255.78000        2
0.000
11868                SAFCO Rocking Chair, Black  397.44000        3
0.000
11870               Hewlett Copy Machine, Color  176.38652        1
0.002
11871           Avery Binding Machine, Economy  131.86800        3
0.100
...                                        ...        ...      ...
...
51283           Eaton Parchment Paper, Premium   26.94000        2
```

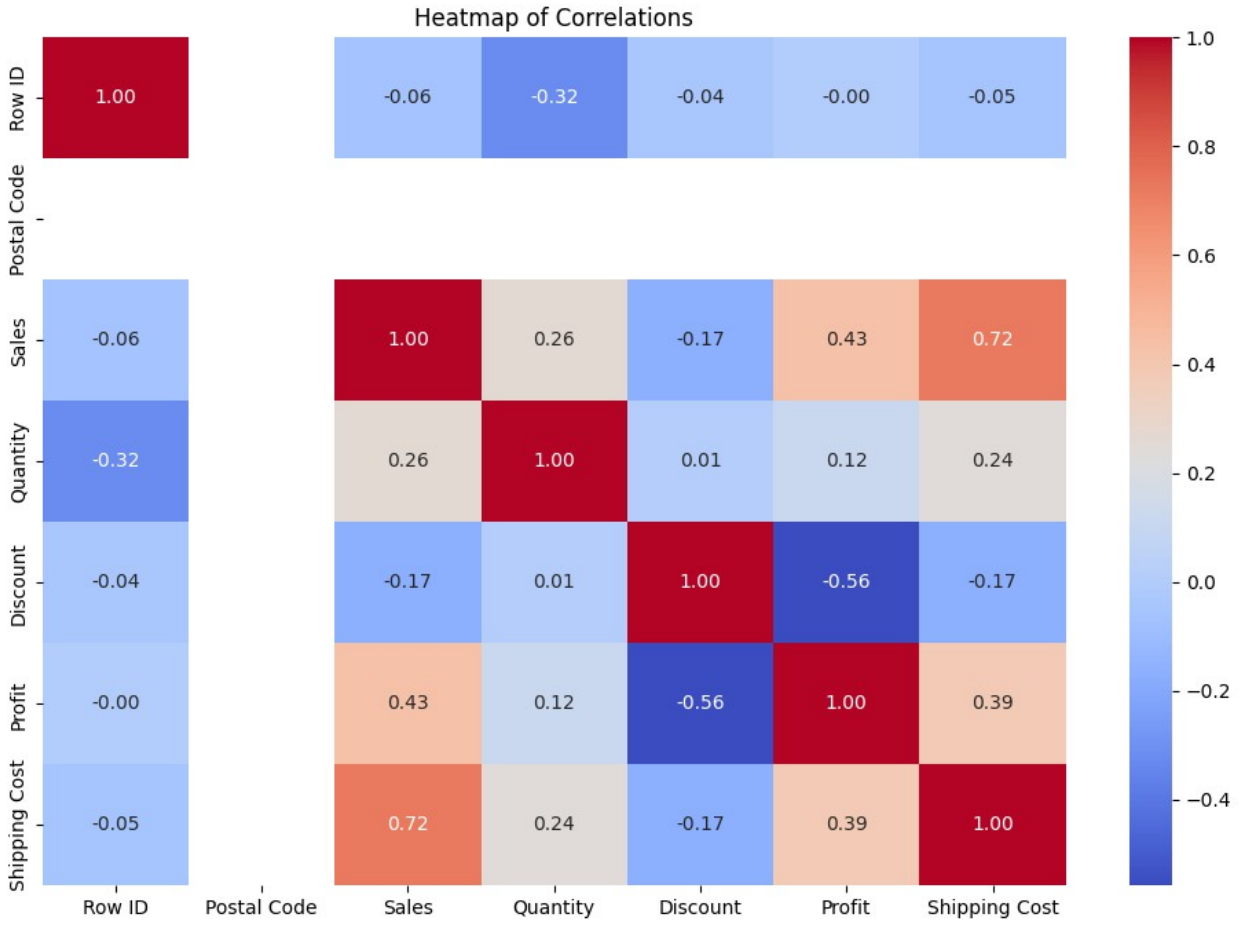
```
0.000
51284                     Avery Binder, Economy   58.05000         5
0.100
51285             Advantus Thumb Tacks, 12 Pack   65.10000         5
0.000
51288                   Acco Index Tab, Economy   13.44000         2
0.000
51289  Eaton Computer Printout Paper, 8.5 x 11   61.38000         3
0.000

          Profit  Shipping Cost  Order Priority
11866  40.28000          26.96            High
11867  30.66000          26.96          Medium
11868  15.84000          26.96          Medium
11870  26.14652          26.96        Critical
11871   5.77800          26.95            High
...         ...            ...             ...
51283   1.86000           0.01            High
51284  19.95000           0.01          Medium
51285   4.50000           0.01          Medium
51288   2.40000           0.00          Medium
51289   1.80000           0.00            High

[25882 rows x 24 columns]>


data.hist(figsize=(10,10))
plt.suptitle("Histograms of Numerical Columns", fontsize=16)
plt.show()
```




```
plt.figure(figsize=(15, 5))
for i, col in enumerate(numerical_columns, 1):
    plt.subplot(1, len(numerical_columns), i)
    sns.boxplot(data[col])
    plt.title(f"Boxplot of {col}")
plt.tight_layout()
plt.show()
```

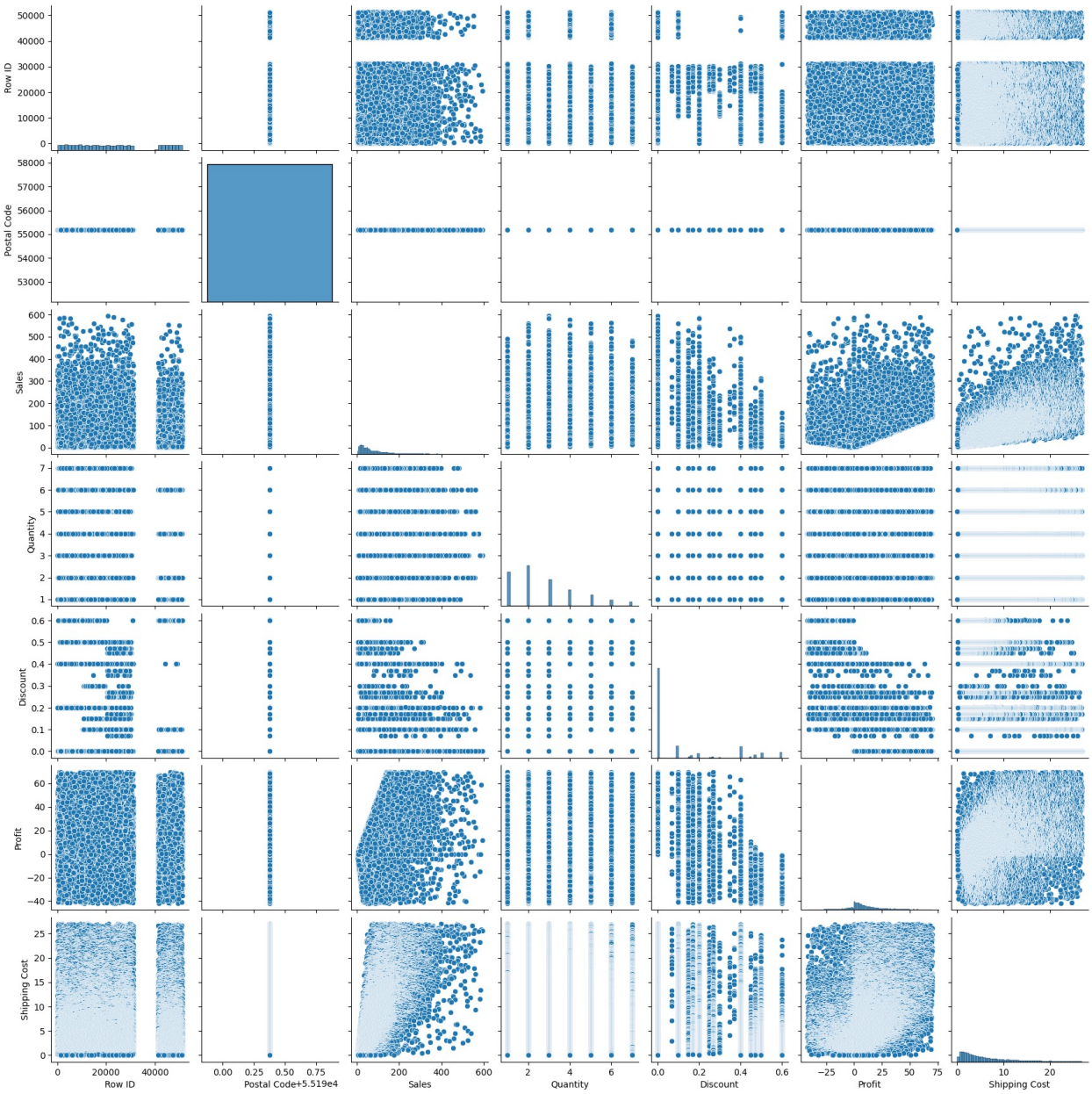



```
# Compute correlations for numeric columns only
correlations = data.select_dtypes(include=np.number).corr()

# Check if the correlation matrix is valid
if correlations.empty:
    print("No numeric columns to compute correlations.")
else:
    # Create heatmap
    plt.figure(figsize=(12, 8))
    sns.heatmap(correlations, annot=True, fmt=".2f", cmap="coolwarm",
cbar=True)
    plt.title("Heatmap of Correlations")
    plt.show()
```




```
sns.pairplot(data=data)
```

```
<seaborn.axisgrid.PairGrid at 0x274de7ab650>
```


Row ID
Postal Code
Sales
Quantity
Discount
Profit
Shipping Cost
Row ID
Postal Code+5.519e4
Sales
Quantity
Discount
Profit
Shipping Cost