```python
import pandas as pd
df=pd.read_csv("C:\\Users\\shaik\\Downloads\\
sales_data_sample.csv",encoding='latin1')
print(df.head())
```

```
   ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER
SALES  \
0        10107               30      95.70                2  2871.00

1        10121               34      81.35                5  2765.90

2        10134               41      94.74                2  3884.34

3        10145               45      83.26                6  3746.70

4        10159               49     100.00               14  5205.27


         ORDERDATE   STATUS  QTR_ID  MONTH_ID  YEAR_ID  ...  \
0   2/24/2003 0:00  Shipped       1         2     2003  ...
1    5/7/2003 0:00  Shipped       2         5     2003  ...
2    7/1/2003 0:00  Shipped       3         7     2003  ...
3   8/25/2003 0:00  Shipped       3         8     2003  ...
4  10/10/2003 0:00  Shipped       4        10     2003  ...

                    ADDRESSLINE1  ADDRESSLINE2           CITY STATE  \
0        897 Long Airport Avenue           NaN            NYC    NY
1             59 rue de l'Abbaye           NaN          Reims   NaN
2  27 rue du Colonel Pierre Avia           NaN          Paris   NaN
3             78934 Hillside Dr.           NaN       Pasadena    CA
4                7734 Strong St.           NaN  San Francisco    CA

  POSTALCODE COUNTRY TERRITORY CONTACTLASTNAME CONTACTFIRSTNAME
DEALSIZE
0      10022     USA       NaN              Yu             Kwai
Small
1      51100  France      EMEA         Henriot             Paul
Small
2      75508  France      EMEA        Da Cunha           Daniel
Medium
3      90003     USA       NaN           Young            Julie
Medium
4        NaN     USA       NaN           Brown            Julie
Medium

[5 rows x 25 columns]
```

```python
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2823 entries, 0 to 2822
```

```
Data columns (total 25 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   ORDERNUMBER       2823 non-null   int64  
 1   QUANTITYORDERED   2823 non-null   int64  
 2   PRICEEACH         2823 non-null   float64
 3   ORDERLINENUMBER   2823 non-null   int64  
 4   SALES             2823 non-null   float64
 5   ORDERDATE         2823 non-null   object 
 6   STATUS            2823 non-null   object 
 7   QTR_ID            2823 non-null   int64  
 8   MONTH_ID          2823 non-null   int64  
 9   YEAR_ID           2823 non-null   int64  
 10  PRODUCTLINE       2823 non-null   object 
 11  MSRP              2823 non-null   int64  
 12  PRODUCTCODE       2823 non-null   object 
 13  CUSTOMERNAME      2823 non-null   object 
 14  PHONE             2823 non-null   object 
 15  ADDRESSLINE1      2823 non-null   object 
 16  ADDRESSLINE2      302 non-null    object 
 17  CITY              2823 non-null   object 
 18  STATE             1337 non-null   object 
 19  POSTALCODE        2747 non-null   object 
 20  COUNTRY           2823 non-null   object 
 21  TERRITORY         1749 non-null   object 
 22  CONTACTLASTNAME   2823 non-null   object 
 23  CONTACTFIRSTNAME  2823 non-null   object 
 24  DEALSIZE          2823 non-null   object 
dtypes: float64(2), int64(7), object(16)
memory usage: 551.5+ KB
None
```

```
print(df.describe())
```

```
        ORDERNUMBER  QUANTITYORDERED    PRICEEACH  ORDERLINENUMBER  \
count   2823.000000      2823.000000  2823.000000      2823.000000   
mean   10258.725115        35.092809    83.658544         6.466171   
std       92.085478         9.741443    20.174277         4.225841   
min    10100.000000         6.000000    26.880000         1.000000   
25%    10180.000000        27.000000    68.860000         3.000000   
50%    10262.000000        35.000000    95.700000         6.000000   
75%    10333.500000        43.000000   100.000000         9.000000   
max    10425.000000        97.000000   100.000000        18.000000   

              SALES       QTR_ID     MONTH_ID     YEAR_ID         MSRP

count   2823.000000  2823.000000  2823.000000  2823.00000  2823.000000

mean    3553.889072     2.717676     7.092455  2003.81509   100.715551
```

```
std     1841.865106     1.203878     3.656633     0.69967    40.187912

min      482.130000     1.000000     1.000000  2003.00000    33.000000

25%     2203.430000     2.000000     4.000000  2003.00000    68.000000

50%     3184.800000     3.000000     8.000000  2004.00000    99.000000

75%     4508.000000     4.000000    11.000000  2004.00000   124.000000

max    14082.800000     4.000000    12.000000  2005.00000   214.000000
```

```
#Handling missing values
#1 Checking missing value
print(df.isnull().sum())
```

```
ORDERNUMBER            0
QUANTITYORDERED        0
PRICEEACH              0
ORDERLINENUMBER        0
SALES                  0
ORDERDATE              0
STATUS                 0
QTR_ID                 0
MONTH_ID               0
YEAR_ID                0
PRODUCTLINE            0
MSRP                   0
PRODUCTCODE            0
CUSTOMERNAME           0
PHONE                  0
ADDRESSLINE1           0
ADDRESSLINE2        2521
CITY                   0
STATE               1486
POSTALCODE            76
COUNTRY                0
TERRITORY           1074
CONTACTLASTNAME        0
CONTACTFIRSTNAME       0
DEALSIZE               0
dtype: int64
```

```
#2 filling missing values
df.fillna("NA",inplace=True)
print(df.isnull().sum())
```

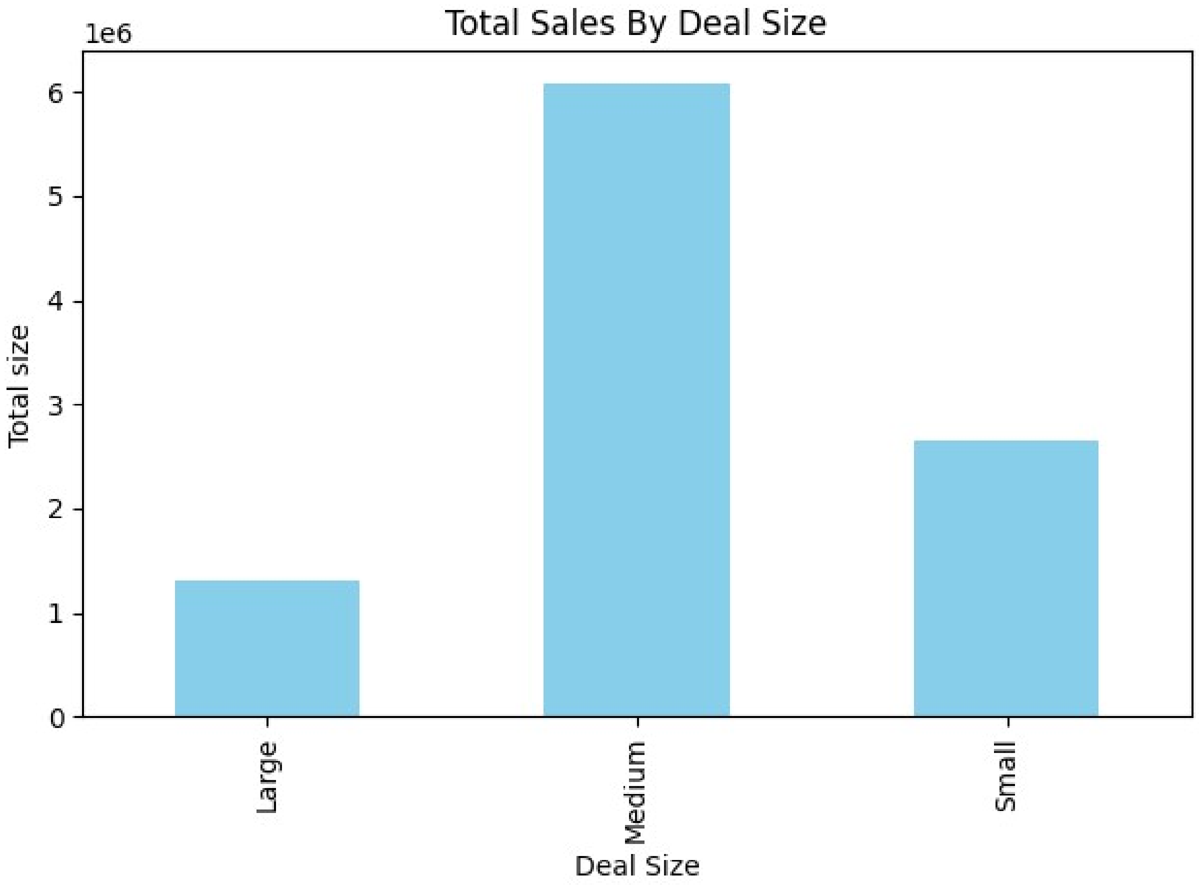
```
ORDERNUMBER         0
QUANTITYORDERED     0
PRICEEACH           0
```

```
ORDERLINENUMBER     0
SALES               0
ORDERDATE           0
STATUS              0
QTR_ID              0
MONTH_ID            0
YEAR_ID             0
PRODUCTLINE         0
MSRP                0
PRODUCTCODE         0
CUSTOMERNAME        0
PHONE               0
ADDRESSLINE1        0
ADDRESSLINE2        0
CITY                0
STATE               0
POSTALCODE          0
COUNTRY             0
TERRITORY           0
CONTACTLASTNAME     0
CONTACTFIRSTNAME    0
DEALSIZE            0
dtype: int64
```

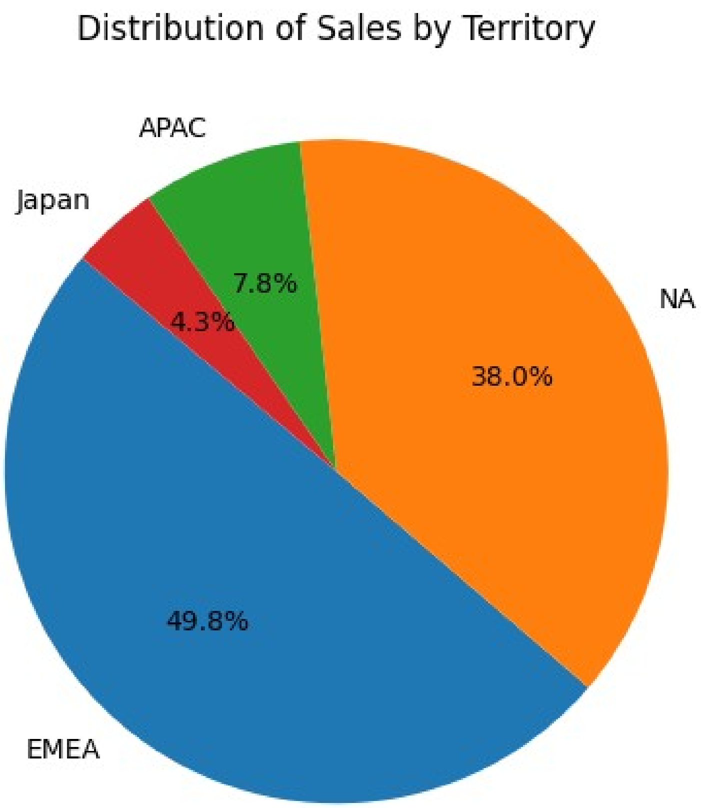
```
#descriptive statistics
df['DEALSIZE'].value_counts()
```

```
DEALSIZE
Medium    1384
Small     1282
Large      157
Name: count, dtype: int64
```

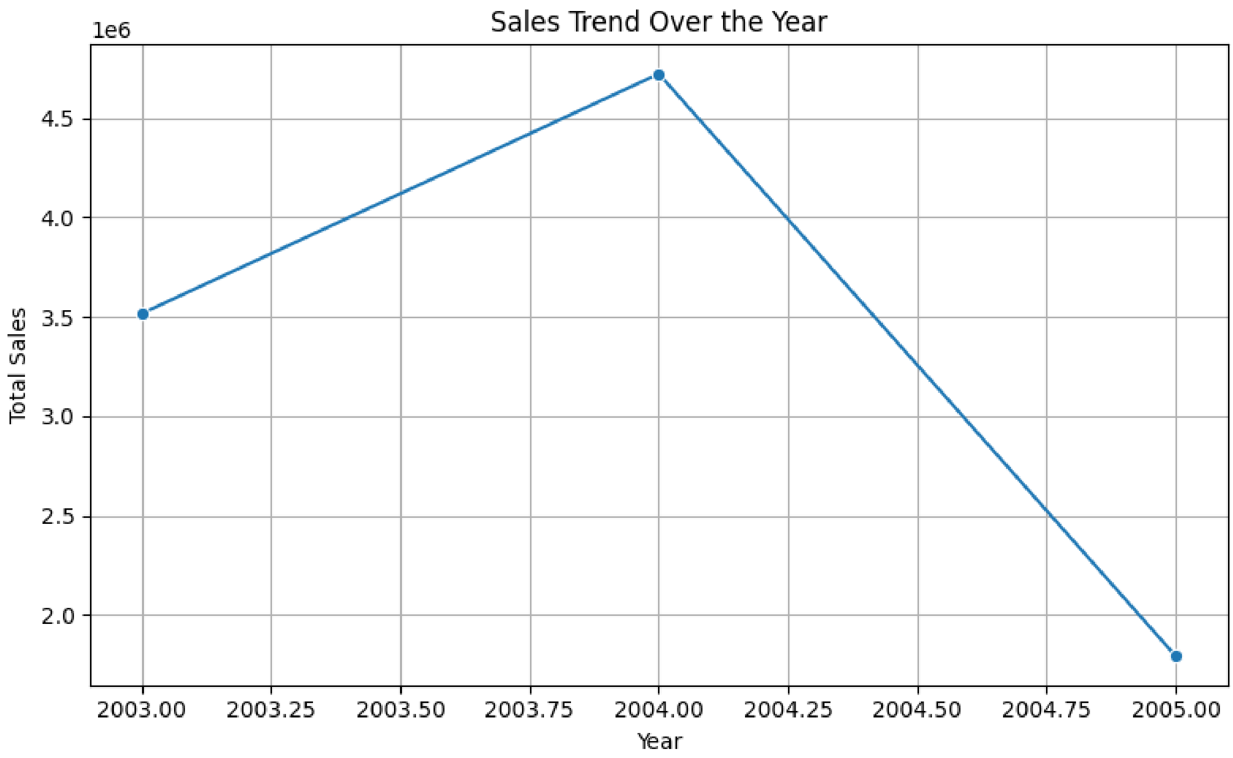
```
#DATA VISUALIZATION
#TOTAL SALES BY DEAL SIZE

import matplotlib.pyplot as plt

sales_by_dealsize=df.groupby('DEALSIZE')['SALES'].sum()
sales_by_dealsize.plot(kind='bar',color='skyblue')
plt.title('Total Sales By Deal Size')
plt.xlabel('Deal Size')
plt.ylabel('Total size')
plt.tight_layout()
plt.show()
```



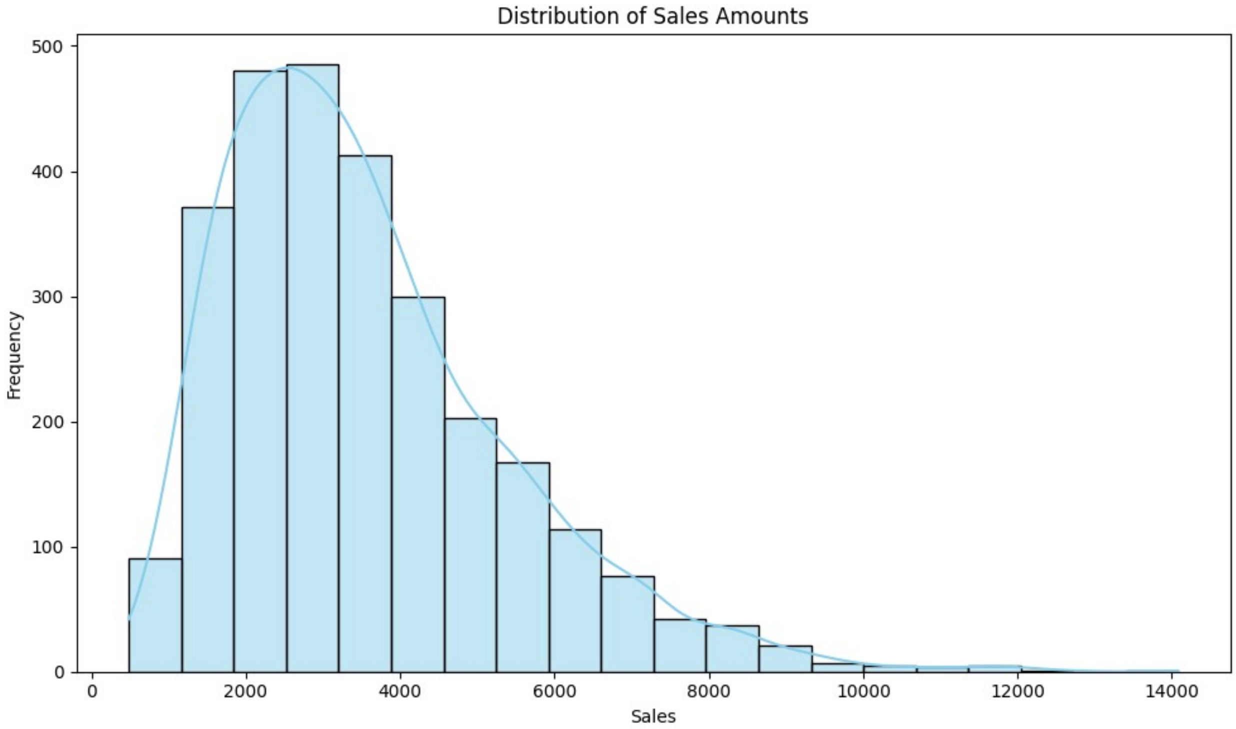
```
#Pie Chart of Territory Distribution
import matplotlib.pyplot as plt

territory_counts=df['TERRITORY'].value_counts()
territory_counts.plot(kind='pie',autopct='%1.1f%%',startangle=140)
plt.title('Distribution of Sales by Territory')
plt.ylabel('')
plt.tight_layout()
plt.show()
```



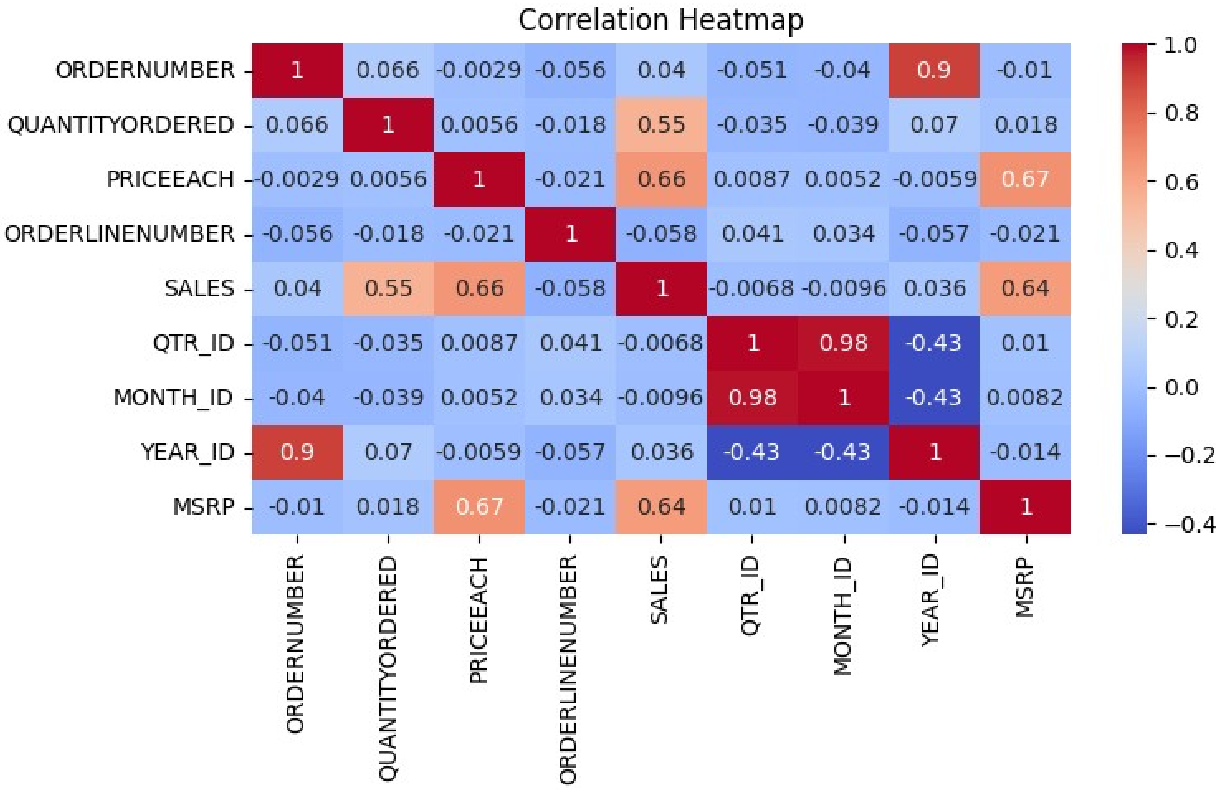
```python
# Sales by Year
import matplotlib.pyplot as plt
import seaborn as sns

sales_by_year=df.groupby('YEAR_ID')['SALES'].sum().reset_index()
plt.figure(figsize=(8,5))
sns.lineplot(x='YEAR_ID',y='SALES', data=sales_by_year,marker='o')
plt.title("Sales Trend Over the Year")
plt.xlabel("Year")
plt.ylabel("Total Sales")
plt.grid(True)
plt.tight_layout()
plt.show()
```



```python
#Sales Distribution
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10,6))
sns.histplot(df['SALES'],bins=20,kde=True,color='skyblue',edgecolor='b
lack')
plt.title("Distribution of Sales Amounts")
plt.xlabel("Sales")
plt.ylabel("Frequency")
plt.tight_layout()
plt.show()
```



```
#Correlation Heatmap
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(8,5))
sns.heatmap(df.corr(numeric_only=True),annot=True,cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.tight_layout()
plt.show()
```


Correlation Heatmap
ORDERNUMBER 1 0.066 -0.0029 -0.056 0.04 -0.051 -0.04 0.9 -0.01
QUANTITYORDERED 0.066 1 0.0056 -0.018 0.55 -0.035 -0.039 0.07 0.018
PRICEEACH -0.0029 0.0056 1 -0.021 0.66 0.0087 0.0052 -0.0059 0.67
ORDERLINENUMBER -0.056 -0.018 -0.021 1 -0.058 0.041 0.034 -0.057 -0.021
SALES 0.04 0.55 0.66 -0.058 1 -0.0068 -0.0096 0.036 0.64
QTR_ID -0.051 -0.035 0.0087 0.041 -0.0068 1 0.98 -0.43 0.01
MONTH_ID -0.04 -0.039 0.0052 0.034 -0.0096 0.98 1 -0.43 0.0082
YEAR_ID 0.9 0.07 -0.0059 -0.057 0.036 -0.43 -0.43 1 -0.014
MSRP -0.01 0.018 0.67 -0.021 0.64 0.01 0.0082 -0.014 1
ORDERNUMBER
QUANTITYORDERED
PRICEEACH
ORDERLINENUMBER
SALES
QTR_ID
MONTH_ID
YEAR_ID
MSRP
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4