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

```
# supress warnings
from warnings import filterwarnings
filterwarnings('ignore')
```

```
pwd= os.getcwd()
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.m
```

```
!pip install openpyxl xlrd
```

```
Requirement already satisfied: openpyxl in /usr/local/lib/python3.10/dist-packages (3
Requirement already satisfied: xlrd in /usr/local/lib/python3.10/dist-packages (2.0.1
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.10/dist-packages
```

```
file_path = '/content/drive/MyDrive/ML/Amazon Sales data.csv'
df = pd.read_csv(file_path)

# Display the first few rows of the dataframe
df.head()
```

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Unit Sol
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	992
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	280
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	177
3	Sub-Saharan Africa	Sao Tome and	Fruits	Online	C	6/20/2014	514321792	7/5/2014	810

```
dataset=df.copy()
dataset.head()
```

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Unit Sol
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	992
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	280
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	177
3	Sub-Saharan Africa	Sao Tome and	Fruits	Online	C	6/20/2014	514321792	7/5/2014	810

```
dataset.shape
```

```
(100, 14)
```

```
dataset.dtypes
```

```
Region             object
Country            object
Item Type          object
Sales Channel      object
Order Priority     object
Order Date         object
Order ID            int64
Ship Date          object
Units Sold          int64
Unit Price        float64
Unit Cost         float64
Total Revenue     float64
Total Cost        float64
Total Profit      float64
dtype: object
```

```
dataset["Order Date"]=pd.to_datetime(dataset["Order Date"])
dataset["Ship Date"]=pd.to_datetime(dataset["Ship Date"])
dataset["Order ID"]=dataset["Order ID"].astype(str)
dataset.dtypes
```

```
Region                    object
Country                   object
Item Type                 object
Sales Channel             object
Order Priority            object
Order Date        datetime64[ns]
Order ID                  object
Ship Date         datetime64[ns]
Units Sold                 int64
Unit Price               float64
Unit Cost                float64
Total Revenue            float64
Total Cost               float64
Total Profit             float64
dtype: object
```

```
for i in dataset.columns:

    print("\nno.of unique values present in column %s are %i\n"%(i,dataset[i].nunique()))
    print(dataset[i].unique())
```

```

no.of unique values present in column Region are 7

['Australia and Oceania' 'Central America and the Caribbean' 'Europe'
 'Sub-Saharan Africa' 'Asia' 'Middle East and North Africa'
 'North America']

no.of unique values present in column Country are 76

['Tuvalu' 'Grenada' 'Russia' 'Sao Tome and Principe' 'Rwanda'
 'Solomon Islands' 'Angola' 'Burkina Faso' 'Republic of the Congo'
 'Senegal' 'Kyrgyzstan' 'Cape Verde' 'Bangladesh' 'Honduras' 'Mongolia'
 'Bulgaria' 'Sri Lanka' 'Cameroon' 'Turkmenistan' 'East Timor' 'Norway'
 'Portugal' 'New Zealand' 'Moldova ' 'France' 'Kiribati' 'Mali'
 'The Gambia' 'Switzerland' 'South Sudan' 'Australia' 'Myanmar' 'Djibouti'
 'Costa Rica' 'Syria' 'Brunei' 'Niger' 'Azerbaijan' 'Slovakia' 'Comoros'
 'Iceland' 'Macedonia' 'Mauritania' 'Albania' 'Lesotho' 'Saudi Arabia'
 'Sierra Leone' "Cote d'Ivoire" 'Fiji' 'Austria' 'United Kingdom'
 'San Marino' 'Libya' 'Haiti' 'Gabon' 'Belize' 'Lithuania' 'Madagascar'
 'Democratic Republic of the Congo' 'Pakistan' 'Mexico'
 'Federated States of Micronesia' 'Laos' 'Monaco' 'Samoa ' 'Spain'
 'Lebanon' 'Iran' 'Zambia' 'Kenya' 'Kuwait' 'Slovenia' 'Romania'
 'Nicaragua' 'Malaysia' 'Mozambique']

no.of unique values present in column Item Type are 12

['Baby Food' 'Cereal' 'Office Supplies' 'Fruits' 'Household' 'Vegetables'
 'Personal Care' 'Clothes' 'Cosmetics' 'Beverages' 'Meat' 'Snacks']

no.of unique values present in column Sales Channel are 2

['Offline' 'Online']
```

```
no.of unique values present in column Order Priority are 4

['H' 'C' 'L' 'M']

no.of unique values present in column Order Date are 100

<DatetimeArray>
['2010-05-28 00:00:00', '2012-08-22 00:00:00', '2014-05-02 00:00:00',
 '2014-06-20 00:00:00', '2013-02-01 00:00:00', '2015-02-04 00:00:00',
 '2011-04-23 00:00:00', '2012-07-17 00:00:00', '2015-07-14 00:00:00',
 '2014-04-18 00:00:00', '2011-06-24 00:00:00', '2014-08-02 00:00:00',
 '2017-01-13 00:00:00', '2017-02-08 00:00:00', '2014-02-19 00:00:00',
 '2012-04-23 00:00:00', '2016-11-19 00:00:00', '2015-04-01 00:00:00',
 '2010-12-30 00:00:00', '2012-07-31 00:00:00', '2014-05-14 00:00:00',
 '2015-07-31 00:00:00', '2016-06-30 00:00:00', '2014-09-08 00:00:00',
 '2016-05-07 00:00:00', '2017-05-22 00:00:00', '2014-10-13 00:00:00',
 '2010-05-07 00:00:00', '2014-07-18 00:00:00', '2012-05-26 00:00:00',
 '2012-09-17 00:00:00', '2013-12-29 00:00:00', '2015-10-27 00:00:00',
 '2015-01-16 00:00:00', '2017-02-25 00:00:00', '2017-05-08 00:00:00',
 '2011-11-22 00:00:00', '2017-01-14 00:00:00', '2012-04-01 00:00:00',
 '2012-02-16 00:00:00', '2017-03-11 00:00:00', '2010-02-06 00:00:00',
 '2012-06-07 00:00:00', '2012-10-06 00:00:00', '2015-11-14 00:00:00',
 '2016-03-29 00:00:00', '2016-12-31 00:00:00', '2010-12-23 00:00:00',
 '2014-10-14 00:00:00', '2012-01-11 00:00:00', '2010-02-02 00:00:00',
 '2013-08-18 00:00:00', '2013-03-25 00:00:00', '2011-11-26 00:00:00'
```

```
round(dataset.isna().mean()*100,2)
```

```
Region            0.0
Country           0.0
Item Type         0.0
Sales Channel     0.0
Order Priority    0.0
Order Date        0.0
Order ID          0.0
Ship Date         0.0
Units Sold        0.0
Unit Price        0.0
Unit Cost         0.0
Total Revenue     0.0
Total Cost        0.0
Total Profit      0.0
dtype: float64
```

```
df_num = dataset.select_dtypes(include="number")
df_num.head()
```

	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
0	9925	255.28	159.42	2533654.00	1582243.50	951410.50
1	2804	205.70	117.11	576782.80	328376.44	248406.36
2	1779	651.21	524.96	1158502.59	933903.84	224598.75
3	8102	9.33	6.92	75591.66	56065.84	19525.82
4	5062	651.21	524.96	3296425.02	2657347.52	639077.50

```
df_col = dataset.drop(df_num.columns,axis=1)
df_col.head()
```

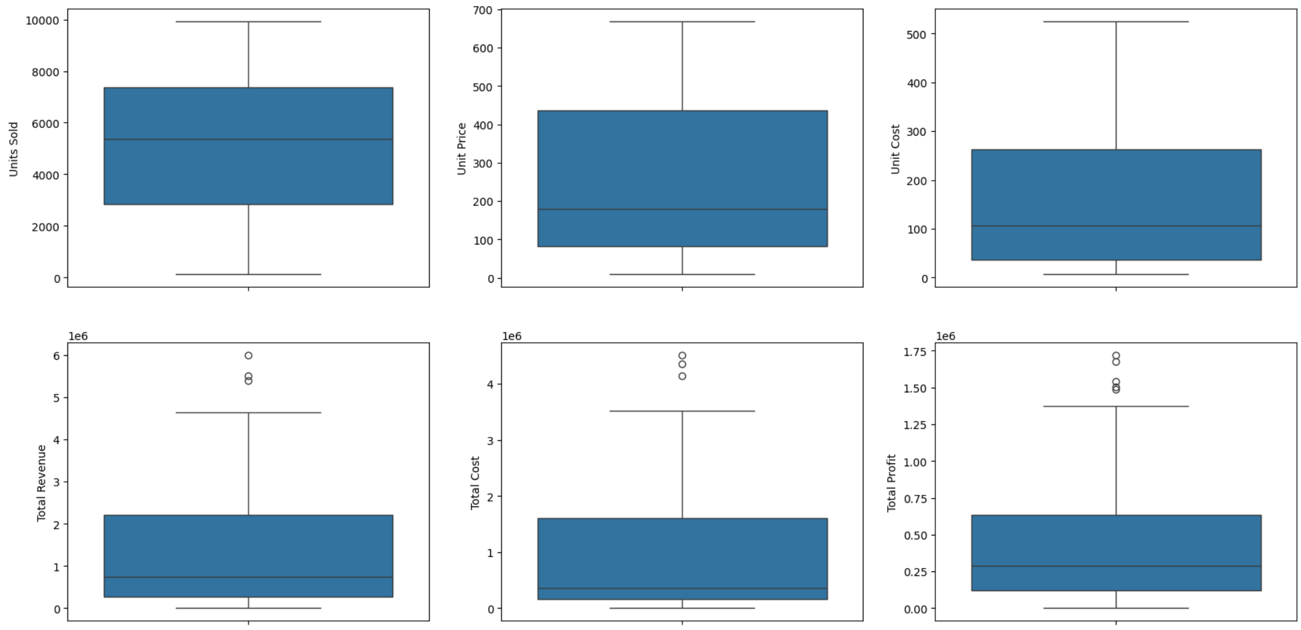
	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	2010-05-28	669165933	2010-06-27
1	Central America and the Caribbean	Grenada	Cereal	Online	C	2012-08-22	963881480	2012-09-15
2	Europe	Russia	Office Supplies	Offline	L	2014-05-02	341417157	2014-05-08

```
df_num.describe()
```

	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
count	100.000000	100.000000	100.000000	1.000000e+02	1.000000e+02	1.000000e+02
mean	5128.710000	276.761300	191.048000	1.373488e+06	9.318057e+05	4.416820e+05
std	2794.484562	235.592241	188.208181	1.460029e+06	1.083938e+06	4.385379e+05
min	124.000000	9.330000	6.920000	4.870260e+03	3.612240e+03	1.258020e+03
25%	2836.250000	81.730000	35.840000	2.687212e+05	1.688680e+05	1.214436e+05
50%	5382.500000	179.880000	107.275000	7.523144e+05	3.635664e+05	2.907680e+05
75%	7369.000000	437.200000	263.330000	2.212045e+06	1.613870e+06	6.358288e+05
max	9925.000000	668.270000	524.960000	5.997055e+06	4.509794e+06	1.719922e+06

```
t=1
plt.figure(figsize=[20,10])
for i in df_num.columns:
    plt.subplot(2,3,t)
    sns.boxplot(dataset[i])

    t=t+1
plt.show()
```



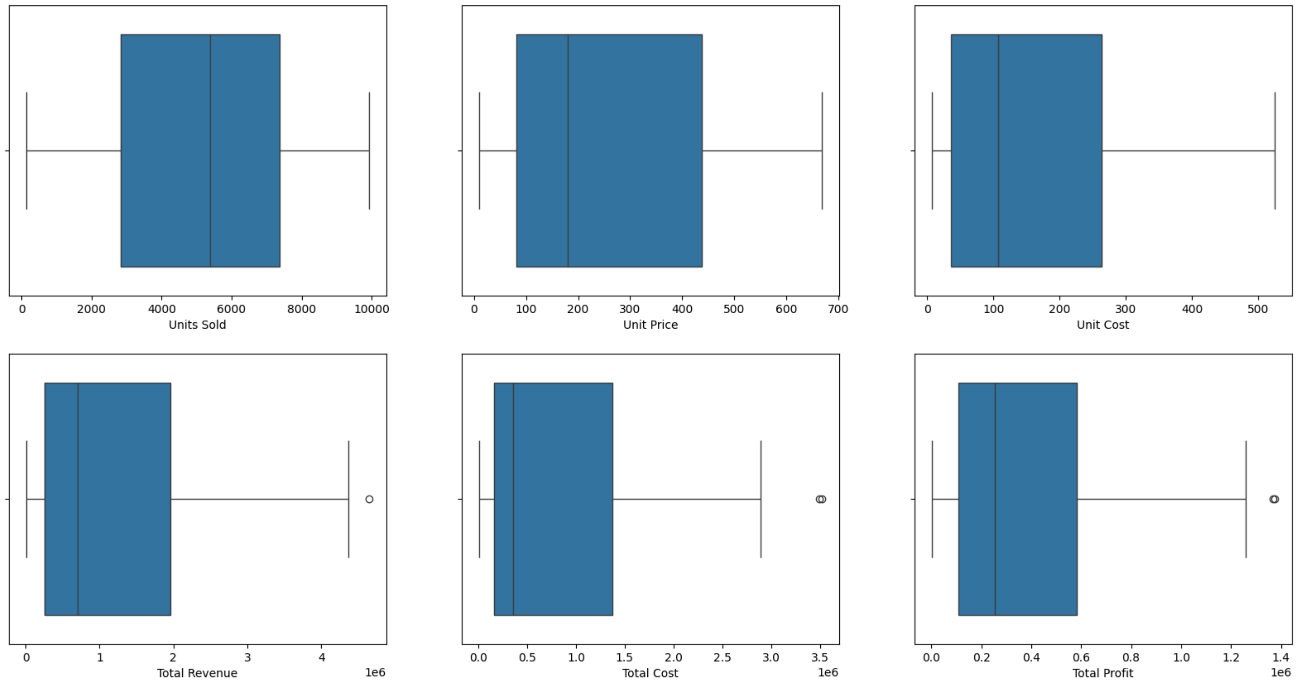
```
for i in df_num.columns:

    q1=df_num[i].quantile(0.25)
    q3=df_num[i].quantile(0.75)
    iqr=q3-q1
    dataset[i]=dataset[i][((dataset[i]>q1-iqr*1.5)&(dataset[i]<q3+iqr*1.5))]




t=1
plt.figure(figsize=[20,10])
for i in df_num.columns:
    plt.subplot(2,3,t)
    sns.boxplot(x=dataset[i])

    t=t+1
plt.show()
```



```
dataset.dropna(inplace= True)
```

```
dataset.shape
```

```
(93, 14)
```

```
dataset["delivery lead time"]=abs(dataset["Order Date"]-dataset["Ship Date"])
```

```
dataset.head()
```

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Un Pri
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	2010-05-28	669165933	2010-06-27	9925	255.
1	Central America and the Caribbean	Grenada	Cereal	Online	C	2012-08-22	963881480	2012-09-15	2804	205.
2	Europe	Russia	Office Supplies	Offline	L	2014-05-02	341417157	2014-05-08	1779	651.
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	2014-06-20	514321792	2014-07-05	8102	9.

```
dataset["sales_year"]= pd.DatetimeIndex(dataset["Order Date"]).year
dataset["sales_month"]= pd.DatetimeIndex(dataset["Order Date"]).month
dataset["sales_month_year"]= dataset["Order Date"].dt.to_period("M")
dataset.head()
```

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Un Pri
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	2010-05-28	669165933	2010-06-27	9925	255.
1	Central America and the Caribbean	Grenada	Cereal	Online	C	2012-08-22	963881480	2012-09-15	2804	205.
2	Europe	Russia	Office Supplies	Offline	L	2014-05-02	341417157	2014-05-08	1779	651.
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	2014-06-20	514321792	2014-07-05	8102	9.
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2013-02-01	115456712	2013-02-06	5062	651.

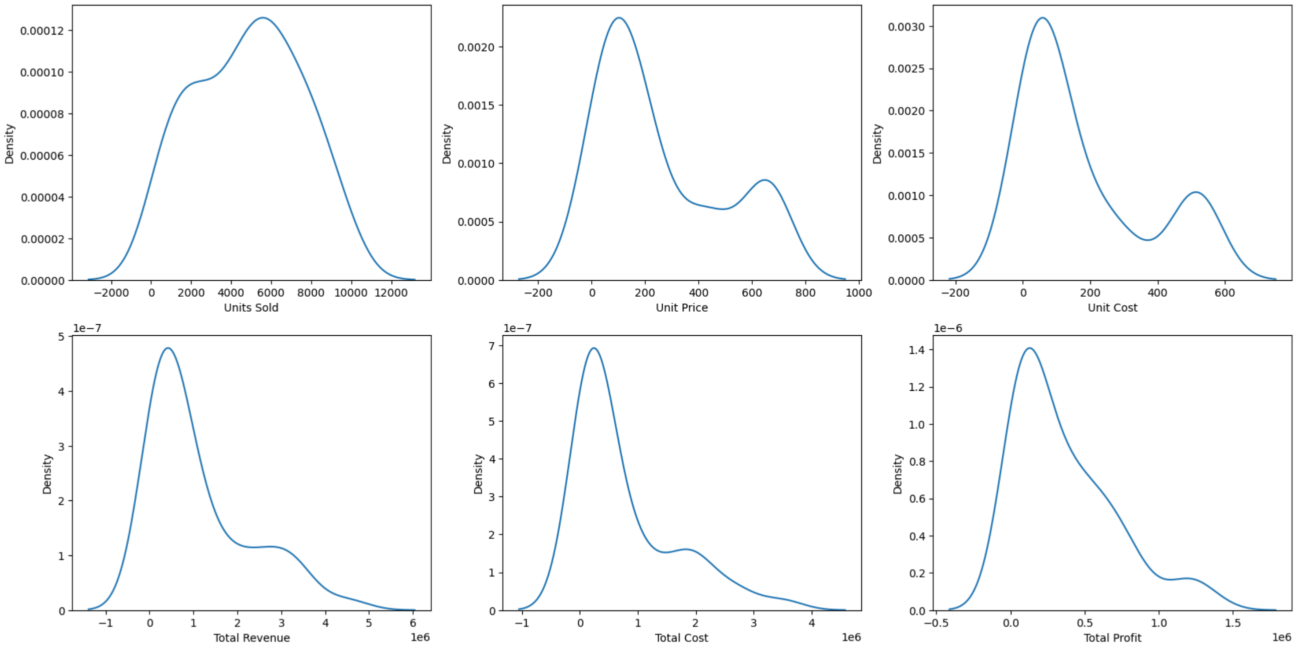
```
dataset["Sales Channel"]=dataset["Sales Channel"].replace({"Offline" : 1 ,"Online":0})

dataset.head()
```

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Un Pri
0	Australia and Oceania	Tuvalu	Baby Food	1	H	2010-05-28	669165933	2010-06-27	9925	255.
1	Central America and the Caribbean	Grenada	Cereal	0	C	2012-08-22	963881480	2012-09-15	2804	205.
2	Europe	Russia	Office Supplies	1	L	2014-05-02	341417157	2014-05-08	1779	651.
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	0	C	2014-06-20	514321792	2014-07-05	8102	9.
4	Sub-Saharan Africa	Rwanda	Office Supplies	1	L	2013-02-01	115456712	2013-02-06	5062	651.

```
t=1
plt.figure(figsize=[20,10])
for i in df_num.columns:
    plt.subplot(2,3,t)
    sns.kdeplot(x=dataset[i])

    t=t+1
plt.show()
```



```
dataset.to_excel(pwd+"\\amazon_sale_clean_data.xlsx",index= False)
```

```
!pip install ydata-profiling
import ydata_profiling as pp

pp.ProfileReport(dataset)
```