Superstore Sales Dataset EDA

August 16, 2023

1 Superstore Sales Dataset EDA

EDA of four years of global superstore sales data

- **Access dataset HERE**

[1]:
```
#importing packahes

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

[2]:
```
# importing dataset

df = pd.read_csv(r"/Users/scipio/Downloads/Sales_Dataset_Project.csv")

df.head()
```

[2]:

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category
0	1	CA-2017-152156	08/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-BO-10001798	Furniture	Bookcases
1	2	CA-2017-152156	08/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-CH-10000454	Furniture	Chairs
2	3	CA-2017-138688	12/06/2017	16/06/2017	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036.0	West	OFF-LA-10000240	Office Supplies	Labels
3	4	US-2016-108966	11/10/2016	18/10/2016	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311.0	South	FUR-TA-10000577	Furniture	Tables
4	5	US-2016-108966	11/10/2016	18/10/2016	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311.0	South	OFF-ST-10000760	Office Supplies	Storage

```
                                          Product Name     Sales
0                  Bush Somerset Collection Bookcase  261.9600
1  Hon Deluxe Fabric Upholstered Stacking Chairs,…  731.9400
2  Self-Adhesive Address Labels for Typewriters b…   14.6200
3      Bretford CR4500 Series Slim Rectangular Table  957.5775
4                      Eldon Fold 'N Roll Cart System   22.3680
```

[71]:
```
df.isnull().sum()
```

[71]:
```
Row ID            0
Order ID          0
Order Date        0
Ship Date         0
Ship Mode         0
Customer ID       0
Customer Name     0
Segment           0
Country           0
City              0
State             0
Postal Code      11
Region            0
Product ID        0
Category          0
Sub-Category      0
Product Name      0
Sales             0
dtype: int64
```

There are 11 null values in the *'Postal Code'* column.

[8]:
```
df.shape
```

[8]:
```
(9800, 18)
```

[63]:
```
# coverting 'Order Date' to datetime format
df['Order Date'] = pd.to_datetime(df['Order Date'])
```

[64]:
```
# Min and Max dates
df['Order Date'].agg(['min','max'])
```

[64]:
```
min   2015-01-02
max   2018-12-30
Name: Order Date, dtype: datetime64[ns]
```

[65]:
```
# Range of Data
df['Order Date'].max() - df['Order Date'].min()
```

[65]:
```
Timedelta('1458 days 00:00:00')
```

[15]:
```
# Number of unqiue Customers in the dataset
df['Customer Name'].nunique()
```

[15]:
```
793
```

[16]:
```
# Number of unqiue Cities in the dataset
df['City'].nunique()
```

[16]:
```
529
```

[17]:
```
# Number of unqiue States
df['State'].nunique()
```

[17]:
```
49
```

[20]:
```
# Number of regions
df['Region'].nunique()
```

[20]:
```
4
```

[18]:
```
# Number of Segments
df['Segment'].nunique()
```

[18]:
```
3
```

[21]:
```
# Number of Categories
df['Category'].nunique()
```

[21]:
```
3
```

[23]:
```
# Number of Sub Categories
df['Sub-Category'].nunique()
```

[23]:
```
17
```

[24]:
```
# Number of prodcuts
df['Product Name'].nunique()
```

[24]:
```
1849
```

[6]:
```
# sales statistical metrics

round(df['Sales'].describe(),2)
```

[6]:
```
count     9800.00
mean       230.77
std        626.65
```

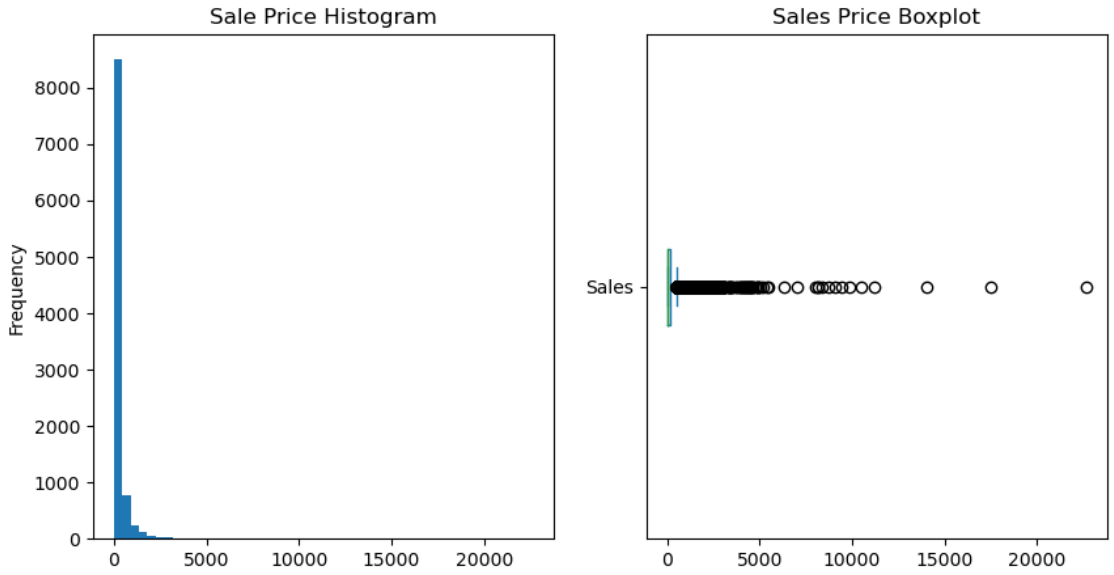
```
min          0.44
25%         17.25
50%         54.49
75%        210.60
max      22638.48
Name: Sales, dtype: float64
```

[70]:
```
# Creating a subplot

fig,axs = plt.subplots(nrows = 1, ncols = 2, figsize = (10,5))

df['Sales'].plot(kind = 'hist', ax = axs[0], title = 'Sale Price Histogram',␣
 ↪bins = 50)
df['Sales'].plot(kind = 'box', vert = False, title = 'Sales Price Boxplot')
```

[70]:
```
<Axes: title={'center': 'Sales Price Boxplot'}>
```



Based on the *Sales Price Histogram* there is a right skew in the sales price distribution. This is also indicated in the difference between the mean and median values. Based on the *Sales Price Boxplot* a large portion of the sales prices are outliers. Additionally, the *Sales Price Boxplot* confirms the right skew indicated in the *Sale Price Histogram*.