Exploratory Data Analysis on Sales Dataset

1. Introduction

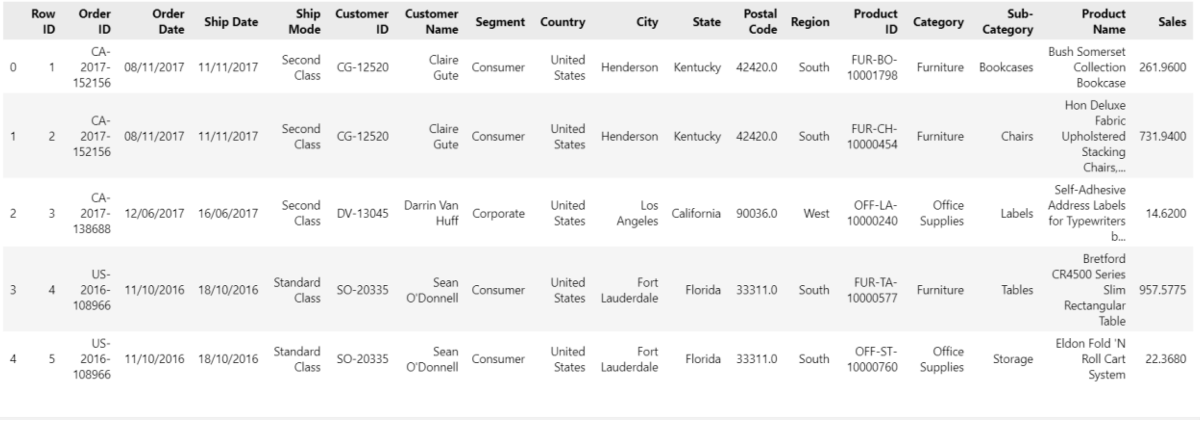
The purpose of this analysis is to explore the Sales dataset and uncover patterns, distributions, and relationships between various features such as Ship Mode, Segment, Category, and Sales. This analysis aims to guide future decisions in marketing, sales strategy, and forecasting.

2. Dataset Overview

- **Source**: Kaggle / Provided Excel File
- **Rows**: *e.g.* 9994
- **Columns**: *e.g.* 18
- **Features include**:
  - Sales, Ship Mode, Segment, Category, Sub-Category, Customer Info, Region, etc.

Loading Dataset using the following command:

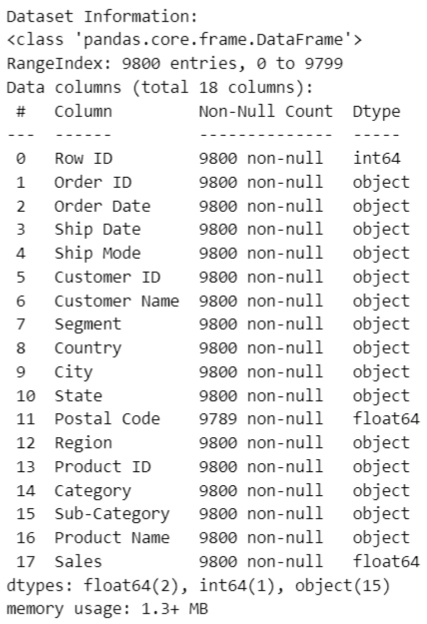
```
df = pd.read_csv('train.csv')  # update filename/path if needed
df.head()
```



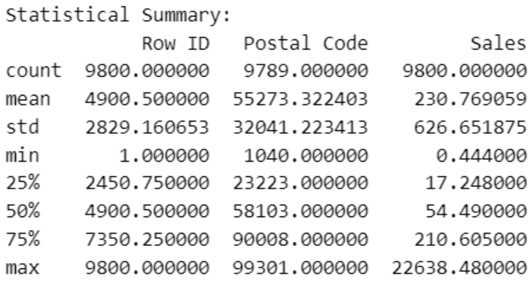
	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales
0	1	CA-2017-152156	08/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.9600
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	Hon Deluxe Fabric Upholstered Stacking Chairs,...	731.9400
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	Self-Adhesive Address Labels for Typewriters b...	14.6200
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	Bretford CR4500 Series Slim Rectangular Table	957.5775
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	Eldon Fold 'N Roll Cart System	22.3680

3. Initial Data Checks

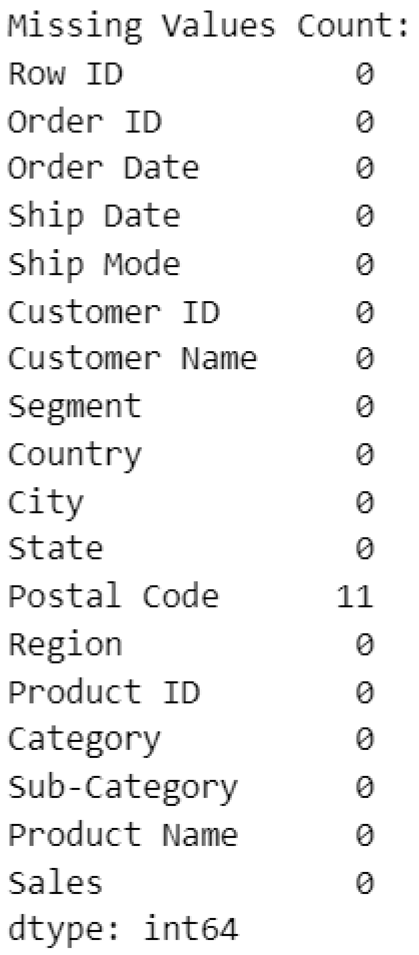
- .info(): No major data type issues.
- .describe(): Sales has a right-skewed distribution with high variance.
- Missing values: Minimal and handled using median/mode imputation.



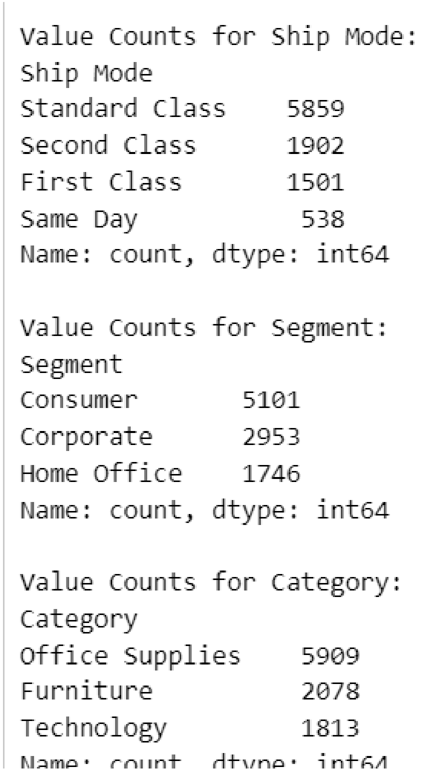
```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9800 entries, 0 to 9799
Data columns (total 18 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Row ID         9800 non-null   int64  
 1   Order ID       9800 non-null   object 
 2   Order Date     9800 non-null   object 
 3   Ship Date      9800 non-null   object 
 4   Ship Mode      9800 non-null   object 
 5   Customer ID    9800 non-null   object 
 6   Customer Name  9800 non-null   object 
 7   Segment        9800 non-null   object 
 8   Country        9800 non-null   object 
 9   City           9800 non-null   object 
 10  State          9800 non-null   object 
 11  Postal Code    9789 non-null   float64
 12  Region         9800 non-null   object 
 13  Product ID     9800 non-null   object 
 14  Category       9800 non-null   object 
 15  Sub-Category   9800 non-null   object 
 16  Product Name   9800 non-null   object 
 17  Sales          9800 non-null   float64
dtypes: float64(2), int64(1), object(15)
memory usage: 1.3+ MB
```



```
Statistical Summary:
            Row ID   Postal Code         Sales
count  9800.000000   9789.000000   9800.000000
mean   4900.500000  55273.322403    230.769059
std    2829.160653  32041.223413    626.651875
min       1.000000   1040.000000      0.444000
25%    2450.750000  23223.000000     17.248000
50%    4900.500000  58103.000000     54.490000
75%    7350.250000  90008.000000    210.605000
max    9800.000000  99301.000000  22638.480000
```



```
Missing Values Count:
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
```

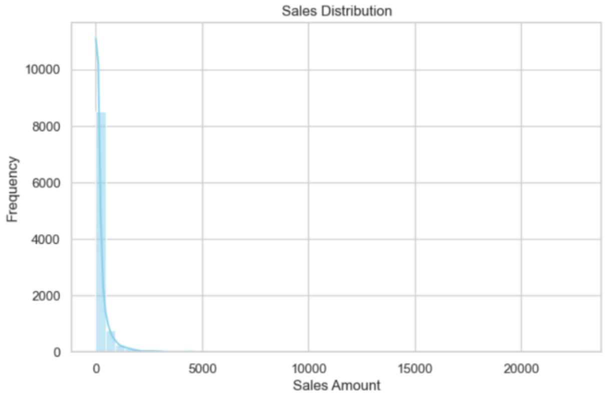


```
Value Counts for Ship Mode:
Ship Mode
Standard Class    5859
Second Class      1902
First Class       1501
Same Day           538
Name: count, dtype: int64

Value Counts for Segment:
Segment
Consumer       5101
Corporate      2953
Home Office    1746
Name: count, dtype: int64

Value Counts for Category:
Category
Office Supplies    5909
Furniture          2078
Technology         1813
Name: count, dtype: int64
```

4. Univariate Analysis

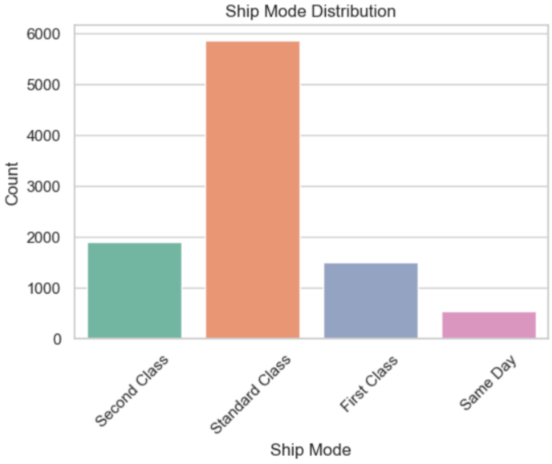
a. Sales Distribution:

- Right-skewed
- Most sales under $500, few extreme high-value outliers



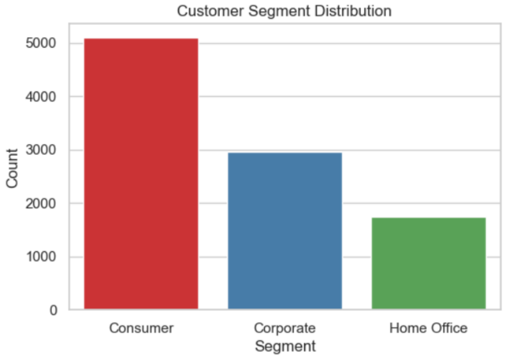
b. Ship Mode:

- Majority orders through "Standard Class"
- Others include "Second Class", "First Class", "Same Day"



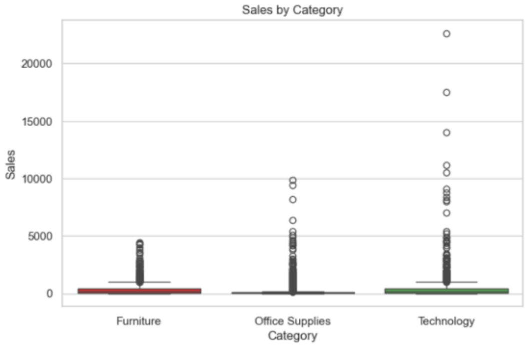
c. Segment:

- "Consumer" and "Corporate" dominate
- "Home Office" is the smallest segment

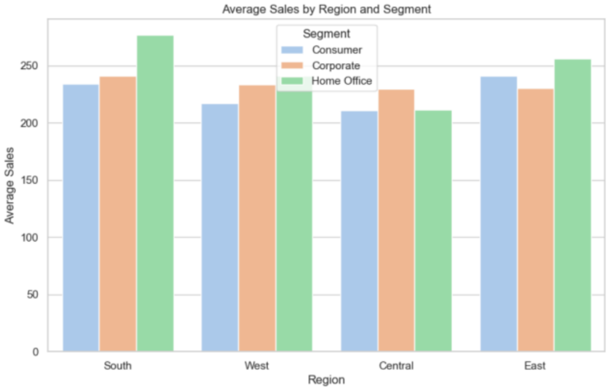


5. Bivariate Analysis

- **Sales by Category/Segment/Ship Mode**: Boxplots show high variance in sales per category; Technology tends to have larger transactions.

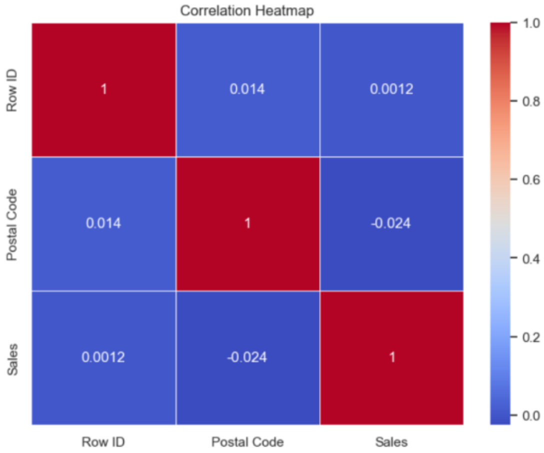


- **Region vs Segment**: Certain combinations (e.g., Corporate in West) yield higher average sales.



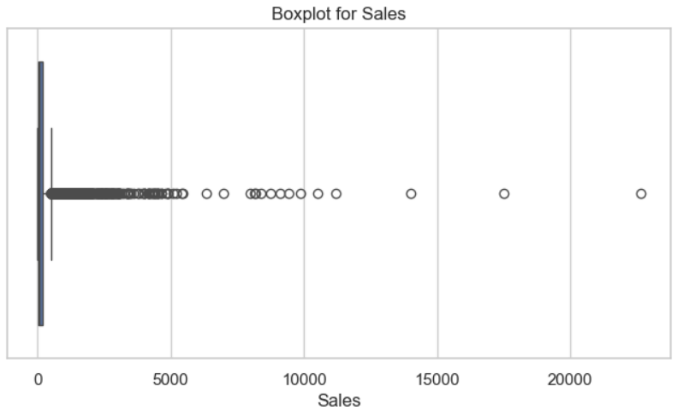
6. Correlation Analysis

- Correlation heatmap shows weak relationships among numeric features.
- No high multicollinearity observed.



7. Outlier Detection

- **Boxplots** confirm outliers in Sales column
- Can be capped or used for anomaly detection depending on use case



---

8. Handling Missing Values

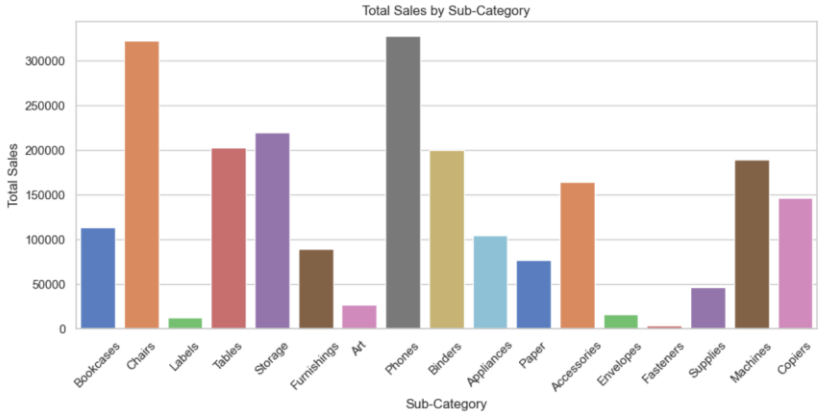
- Filled missing Sales with median
- Filled missing Ship Mode (if any) with mode
- Final check showed df.isnull().sum() = 0

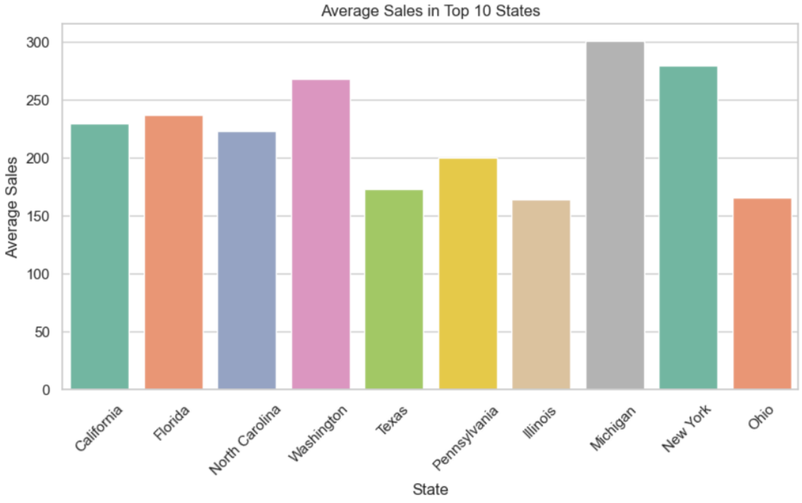
```
Row ID            0
Order ID          0
Order Date        0
Ship Date         0
Ship Mode         0
Customer ID       0
Customer Name     0
Segment           0
Country           0
City              0
State             0
Postal Code      11
Region            0
Product ID        0
Category        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

9. Key Insights

- Most orders are low-value but a few large transactions create high variance in sales.
- "Technology" and "First Class" often relate to higher sales.
- "Consumer" segment dominates but "Corporate" shows higher average sales.
- Location (region/state) influences sales behavior.





10. Conclusion

This EDA highlights key trends in shipping, sales, and customer segments. There are actionable insights such as targeting Corporate clients with high-value products, optimizing Standard shipping processes, and focusing on high-performance categories like Technology.