Exploratory Data Analysis (EDA) on Superstore Sales Data

Harshada Shirsale

December 13, 2025

Contents

1 Objective

Perform complete exploratory data analysis on a real-world Sample Superstore dataset The goal is to understand the data's structure and relationships before formal analysis.

2 Introduction

Steps to perform a complete **Exploratory Data Analysis (EDA)** on sample superstore dataset are following:

1. Data Loading
2. Data Cleaning
3. Visualization
4. Outlier Detection
5. Conclusion

Exploratory Data Analysis (EDA) is important to understand the dataset properly. By performing EDA, we can identify which variables are useful and which are not required for the analysis. In real-world scenarios, it is not practical to use all variables from a large dataset, as most of the time only two or three variables are needed for effective analysis. EDA helps in removing unnecessary data, leading to better results and improved accuracy. Only after performing EDA can we effectively answer questions based on the dataset with reliable accuracy.

Key Findings

Some of the important measures are as follows:

- Total Quantity = 37837
- Total Sales =2295526.002
- Total Profit = 286389.1208

Units of Measurement

- Sales and Profit are measured in US Dollars (USD).
- Quantity represents the number of units sold.

3 Dataset Overview

Name of the Dataset is **Sample - Superstore.xlsx** it's a excel file with proper information. This dataset contains different types of variables, including **Order Date** and **Ship Date**. Additionally, the country is fixed, with data distributed across different cities and regions. along with that the **Sales**, **Discount**, **Profit** are the most important variables are also their to understand the data well

Key Dataset Attributes

Number of records: 9994 rows $\times$ 21 columns

Key Variables

- Sales: Total revenue generated from each order.
- Profit: Net profit obtained after costs and discounts.
- Quantity: Number of units sold in each order.
- Discount: Discount rate applied to the order.
- Category: Broad classification of products.
- Sub-Category: Detailed classification within each product category.
- Segment: Type of customer segment.
- Region: Geographic region where the order was placed.
- State: State in which the order was placed.

Data Types

- Numerical:Postal Code,Sales, Quantity,Discount, Profit
- Categorical:order ID,Ship Mode,Customer ID,Customer Name , object,Segment , Country/Region , City , State
- datetime:Order Date , Ship Date
- ID:Row ID (Not used for modeling)

4 Missing Data Analysis

Missing values

In our Dataset we observed that their is only **Postal Code** has a missing values.

- Postal Code: 11 values missing ($\sim$11.01%)

Handle missing values (mean, median, mode, or drop) (optional)

1. Mean: Used when the values are numerical.
2. Median: Used when the values are numerical and ordinal. Ordinal means categories have a meaningful order.
3. Mode: Used when the values are categorical or discrete numerical data.

5 Visualization

Bar Chart

Use a bar chart to compare discrete categories, show data distribution, or highlight rankings between different groups. here we draw a bar chart for **States** variable to identify each count by it's States

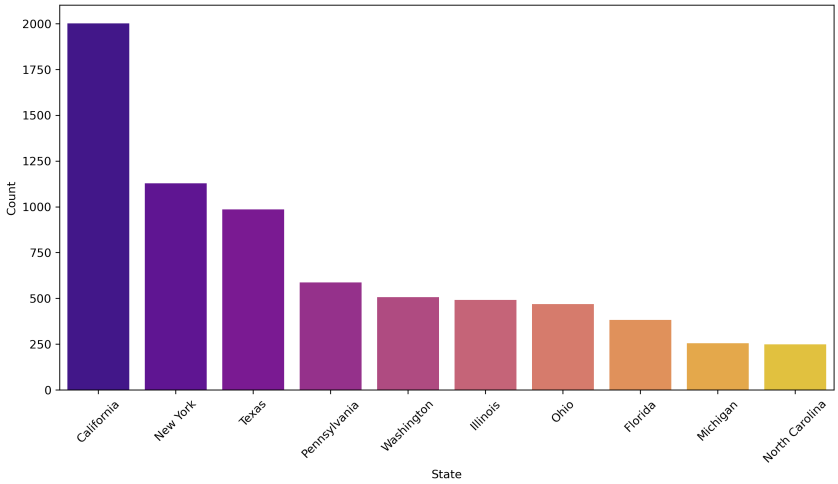

Figure 1: Top 10 States by Number of Records

Conclusion here we see **California** is the top 1 state by number of records and **North Carolina** is on the number 10 by the number of records. the count of **California** is approx. 2000 and the count of **North Carolina** is approx. 250.

Scatter plot

Scatter plot is nothing but a simple plot when we identify the relationship between the two variables like in our data we check what if there is there is any relation between **Sales** and **Profit** along with that **Discount** and **Profit** and most important is we draw a scatter plot only on numeric data.

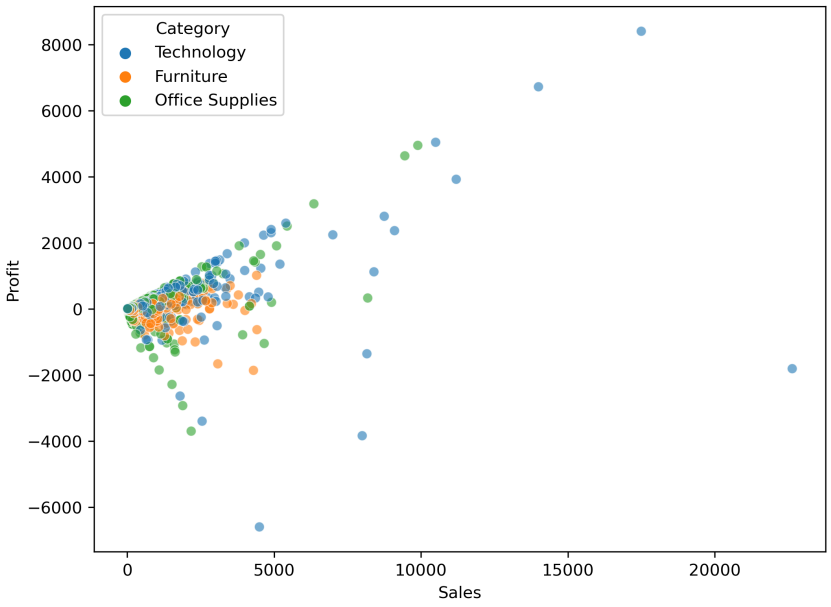

Figure 2: Scatter Plot: Sales vs Profit

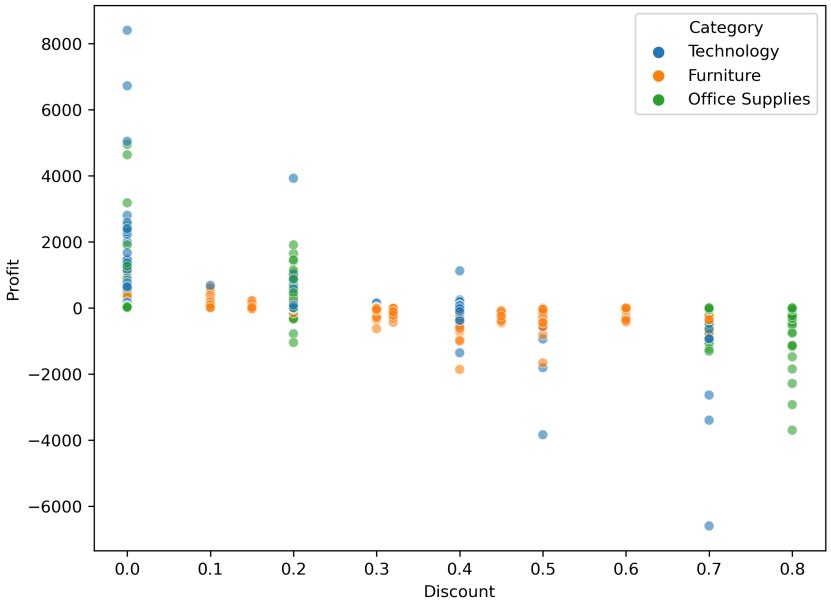

Figure 3: Scatter Plot: Discount vs Profit

Conclusion Figure 2 shows the Scatter plot between Sales vs Profit in this we clearly see the relation between them is positive as **Sales** increases the **Profit** also increases.

Figure 3 shows the Scatter plot between Discount vs Profit in this we clearly see the relation between them is negative as **Discount** increases the **Profit** decreases.

Histogram

Use for numeric distributions such as Sales, Profit, Quantity, Discount

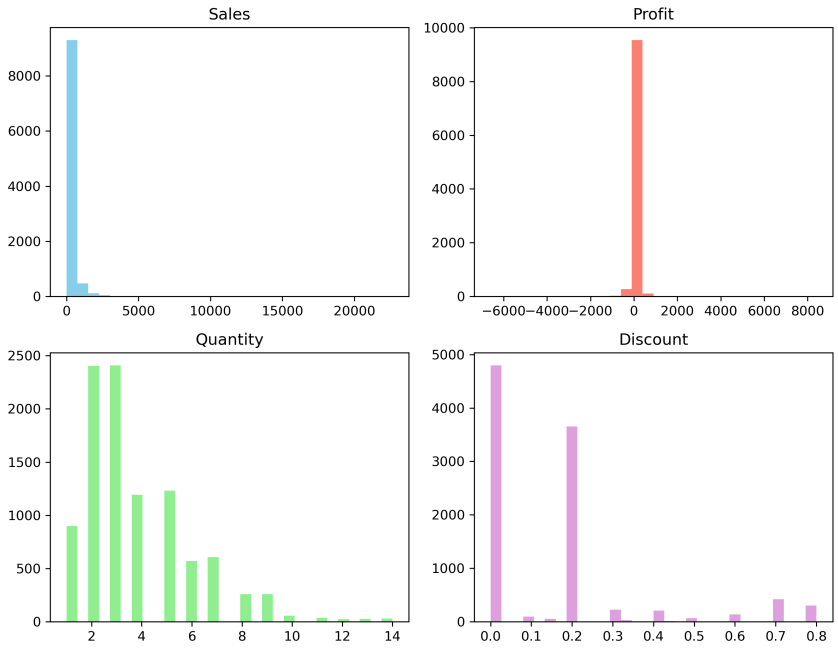

Figure 4: Distribution of Sales, profit, Discount, Quantity

Conclusion Figure 4 shows the Most orders have low sales and profit, with only a few orders having very high values. Customers usually buy small quantities of products. Discounts are generally small, and high discounts are given only in a few cases.

Heatmap

Used to visualize correlations between numeric variables such as Sales, Profit, Discount, and Quantity.A graphical representation of data that uses a system of color coding to represent different values.

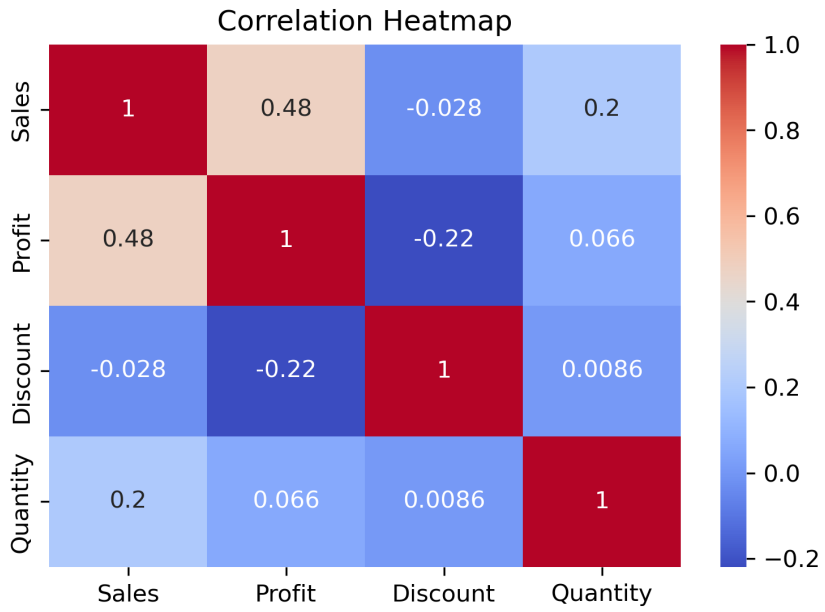

Figure 5: Correlation between Sales, profit, Discount, Quantity

Conclusion Figure 5 shows the Correlation between Sales, profit, Discount, Quantity where **Sales** and **Profit** are the most strongly related variables, while **Discount** shows weak or negative relationships with other numerical variables.

Boxplot

Outliers Extreme value that far away from rest of observations.

to identify the outliers which is present in our dataset the most appropriate tool is boxplot .

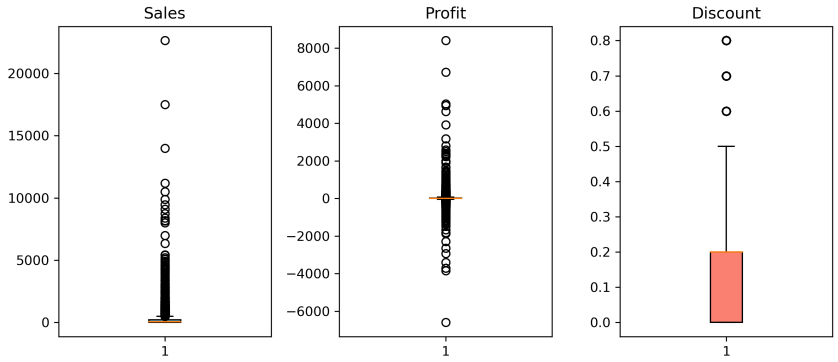

Figure 6: Boxplot of Sales, Profit, and Discount

Conclusion Figure 6 shows the Boxplot of Sales, Profit, and Discount. Use of boxplot is to identify outliers in **Sales**, **Profit**, and **Discount**. we observed that in **Sales** we see most outliers are present and in **Profit**, **Discount** as compare to Sales the outliers are few.

6 Conclusion

- In our Dataset **Sample - Superstore.xlsx** their are some missing values we have to identify. the column name **Postal Code** has the 11 missing values approx. to ($\sim$11.01%).
- Then next is to the how to handle such situation. how to handle missing values in our dataset we generally use simple approach like (**mean**, **mode**, **median**)
- Along with that we can observed all the data types of a variable and check if their is need to transfer the datatype of **int** to the **float** or vice-versa
- To more understand the data we perform visuals on it like **Bar Chart**, **Scatter plot**, **Histogram**, **Heatmap** and most important **Boxplot**
- In **Bar Chart** we see in **Category** in **(office supplies)** has the higesht count among all approx. 6000 Count. also for the **Sub-Category** in **(binders)** has highest among the rest of all approx. 1500 count.
- If go for **Region** wise then **West** has the highest count.
- Then we go for the Scatter plot to check their is any relationship between **Profit**, **Sales**, **Discount** and we identify their is relation between **Sales** and **Profit**.
- Then we go for the correlation to check this we are using **Heatmap**. using this graph we can identify their is positive relation between **Sales** and **Profit** that also we know using the scatter plot but it won't give a correlation value that why we used the heatmap.
- Then for check the outliers present in dataset we go for a **Boxplot** using this we can identify among **Sales**, **Profit**, **Discount** we observed **Sales** has more Outliers present.

Next Step

In the next step, we plan to use appropriate methods to handle the outliers present in the dataset. These techniques will help us decide whether to remove or retain the outliers in the dataset. This will be our next task.