Superstore Dataset Analysis Report

1. Data Loading and Initial Exploration

df = pd.read_csv('/content/Sample - Superstore.csv', encoding='latin-1')

display(df.head())

- **Dataset shape**: 9,994 rows × 21 columns.
- Dataset includes customer orders, shipment details, and sales metrics.

2. Descriptive Statistics

df.describe(include='all')

Key insights:

- **Unique Order IDs**: 5,009
- **Top Ship Mode**: "Standard Class" (59.7% of orders)
- **Top Segment**: "Consumer"
- **Average Sales**: $229.86
- **Average Profit**: $28.66
- **Max Sales**: $22,638.48
- **Max Profit**: $8,349.00

3. Dataset Info

df.info()

- No null values detected.
- Column types: object (15), int64 (3), float64 (3)
- Suitable for both categorical and numerical analysis.

4. Missing & Duplicate Values Check

df.isnull().sum()

df.duplicated().sum()

- **Missing values**: 0
- **Duplicate rows**: 0

5. Categorical Analysis

categorical_cols = df.select_dtypes(include='object').columns

```
for col in categorical_cols:

  print(df[col].value_counts())
```

- **Top Sub-Categories**: Binders, Paper, Furnishings
- **Most Sold Product**: "Staple envelope"

6. Univariate Analysis – Histograms

```
numeric_cols = df.select_dtypes(include=['float64', 'int64']).columns

for col in numeric_cols:

  sns.histplot(df[col], kde=True)
```

- **Sales**: Right-skewed (few high sales values)
- **Profit**: Some losses (negative profits)
- **Discount**: Peaks at 0.2
- **Quantity**: Mostly 2 to 5 items per order

7. Correlation Heatmap

```
sns.heatmap(df[numeric_cols].corr(), annot=True, cmap='coolwarm')
```

- High correlation between **Sales & Profit** (0.48)
- Negative correlation between **Discount & Profit** (-0.22)

8. Pairplot

```
sns.pairplot(df[numeric_cols])
```

- Visualizes pairwise relationships
- Reveals clusters and outliers, especially in Profit vs Sales

9. Bivariate Analysis

Profit vs Discount

```
sns.scatterplot(x='Discount', y='Profit', data=df)
```

- High discounts correlate with reduced profit (sometimes negative)

Total Profit per Category

```
sns.barplot(x='Category', y='Profit', data=df, estimator=sum)
```

- **Highest profit**: Technology
- **Lowest profit**: Furniture

Total Sales per Category

```
sns.barplot(x='Category', y='Sales', data=df, estimator=sum)
```

- **Highest sales**: Technology

- Sales don't always equate to high profits (e.g., Furniture)

Profit by Segment

sns.boxplot(x='Segment', y='Profit', data=df)

- Corporate and Home Office show more profit variation than Consumer

Here is a detailed, in-depth conclusion based on the analysis of the Superstore dataset from your Python notebook:

Conclusion from Superstore Dataset Analysis

(SHORT SUMMARY)

1. Business Performance Overview

The dataset captures extensive order-level details of a retail store, including products sold, customer segments, shipping details, and financial metrics (sales, discount, profit). A thorough analysis reveals that while the business appears to be generating high sales, profitability is not consistently aligned with revenue.

2. Customer Segmentation Insights

- The Consumer segment contributes to the majority of the sales volume (over 50%).

- However, Corporate and Home Office segments show more variability in profit, indicating higher-risk, higher-return behaviour.

- Profit distribution is skewed with numerous outliers, particularly in high-sale, low-profit transactions.

Implication: The company should segment marketing strategies. Target Consumer for volume, but optimize discounts in Corporate/Home Office to stabilize profits.

3. Product Category & Sub-Category Performance

- Technology products drive the highest profits and sales, making it the most lucrative category.

- Furniture, despite relatively high sales, shows poor profitability, likely due to high discounts or cost margins.

- Sub-categories like Chairs, Tables, and Bookcases from Furniture need deeper cost-benefit reviews.

Implication: Prioritize promotion of Technology products, reassess pricing and procurement of Furniture items.

4. Sales vs Discount vs Profit Dynamics

- Discounts show a clear negative correlation with Profit (≈ -0.22). In other words, as discounts increase, profit generally drops.

- There are many instances where large discounts led to net losses, especially on expensive items.

- While discounts may drive volume, they erode margins if not paired with cost control.

Implication: Introduce a discount ceiling or more strategic promotions only for profitable/high-margin items.

5. Sales Distribution and Outliers

- Sales is heavily right-skewed, meaning a few very large transactions drive a significant portion of revenue.

- Similarly, profit has a long tail, with both extremely high profits and losses.

- A boxplot and pair plot show several extreme outliers, indicating inconsistent performance at the order level.

Implication: Conduct order-level audits on extreme outliers to assess potential pricing, fulfilment or customer issues.

6. Regional Performance & Logistics

- Though not deeply visualized in this notebook, the data includes geographical details like Region, City, and Postal Code.

- This is a rich opportunity for geo-segmentation and supply chain optimization, particularly looking at:

    o Shipping costs vs profit by region

    o Inventory distribution effectiveness

Implication: Consider adding regional logistics KPIs to detect low-efficiency regions for cost-cutting or performance boosting.

7. Overall Data Health

- No missing values or nulls, indicating excellent data quality.

- No duplicates, making it reliable for modelling or reporting.

- The structure supports both exploratory data analysis and machine learning-based forecasting (e.g., predicting profit/loss or optimal discount).

Implication: Ready for advanced analytics, including regression modelling or customer lifetime value prediction.

Strategic Recommendations:

Focus Area	Recommendation
Discounts	Implement smart discounting algorithms to avoid profit erosion.
Product Strategy	Expand and push high-margin products, particularly in Technology.
Segments	Customize strategy by segment: Volume-driven for Consumer, Value-driven for Corporate/Home Office.
Cost Management	Reassess unprofitable sub-categories (e.g., Furniture) for cost leaks.
Regional	Leverage geographic data for micro-level decisions in inventory and logistics.