**Superstore Sales Data Analysis — Project Documentation**

**Objective**

Analyze Superstore sales data to uncover trends, seasonal patterns, and product performance insights, ultimately providing actionable recommendations for improving sales forecasting and inventory management.

### ****Dataset****

* **Source**: Superstore dataset (commonly used in analytics projects).
* **Structure**:
  + **Transactions**: Includes fields like Order Date, Ship Date, Category, Sub-Category, Sales, Profit, Quantity, Discount, and Region.
  + **Granularity**: Transaction-level data for multiple years across different geographic regions.
* **Size**: [Insert total rows and columns after inspection].
* **Format**: CSV file.

**Project Goals**

1. **Identify** sales trends and performance variations across regions and product categories.
2. **Detect** seasonality patterns for better demand forecasting.
3. **Evaluate** product profitability and customer purchase behavior.
4. **Recommend** strategies to optimize pricing, discounting, and inventory.

**1. Data Collection**

* Imported sales data from the Superstore CSV file into Python using **Pandas**:

import pandas as pd

df = pd.read\_csv("Superstore.csv")

* Checked dataset shape and schema:

df.shape

df.info()

df.describe()

* Verified data integrity by:
  + Checking for **missing values** with df.isnull().sum().
  + Identifying **duplicates** using df.duplicated().sum().

**2. Data Exploration**

* Reviewed column meanings and data types.
* Used **value counts** and grouping:

df['Category'].value\_counts()

df.groupby('Region')['Sales'].sum()

* Initial findings:
  + Certain regions contribute disproportionately to sales.
  + Office Supplies and Technology categories have different profitability margins.
* Identified **top-performing sub-categories** and **low performers**.

**3. Data Preprocessing**

* **Handled missing values**:
  + For categorical columns: Filled with mode.
  + For numerical columns: Filled with mean/median as appropriate.
* **Removed duplicate records** using:

df.drop\_duplicates(inplace=True)

* **Date formatting**:
  + Converted Order Date and Ship Date to datetime:

df['Order Date'] = pd.to\_datetime(df['Order Date'])

* + Extracted **year**, **month**, and **day** for time-series analysis.

**4. Exploratory Data Analysis (EDA)**

* **Sales Trends Over Time**:
  + Line plots of monthly sales and profit using **Matplotlib** and **Seaborn**.
* **Seasonality Detection**:
  + Grouped sales by month to identify peak sales periods.
* **Correlation Analysis**:
  + Used df.corr() to find relationships between Discount, Profit, and Sales.
  + Observed negative correlation between **Discount** and **Profit**.
* **Category & Region Analysis**:
  + Bar charts for category-level sales.
  + Heatmaps for region-category performance.

**5. Feature Selection**

* Selected impactful features for deeper analysis:
  + **Discount** (high influence on profitability).
  + **Category / Sub-Category**.
  + **Region**.
  + **Order Date** (for trend & seasonality analysis).
* Dropped irrelevant or redundant fields.

**6. Conclusion**

**Key Insights**:

* Sales peak during certain months (holiday seasons).
* Excessive discounts harm profitability despite increasing sales volume.
* Specific regions and product sub-categories dominate total sales.

**Recommendations**:

1. Optimize discount strategies to balance sales and profitability.
2. Focus marketing on high-performing regions during peak months.
3. Review underperforming sub-categories for potential removal or re-pricing.
4. Use seasonal demand patterns for inventory planning.

Pairplot of different perspectives and corelation of columns