Sales Forecasting Project Documentation

Objective

The objective of this project is to analyze historical sales data to forecast future trends and identify key drivers of sales performance. The end goal is to provide actionable insights for better pricing, inventory, and product strategy decisions.

Dataset Overview

- **Columns include:**
  `Invoice Date`, `Product`, `Units Sold`, `Price per Unit`, `Total Sales`, `Operating Profit`, `State`, etc.

- **Time Period:**
  Spans several months of retail sales data.

- **Format:**
  The original dataset contained string-formatted currency values and some missing or inconsistent entries.

Data Cleaning and Preparation

Cleaning Steps:

- Converted currency-formatted strings (e.g., "₹120.00") into numeric format for key columns.

- Handled missing values using median imputation or row removal, depending on criticality.

- Removed duplicate rows and standardized text fields.

- Converted `Invoice Date` into `datetime` format for time series analysis.

- Resampled data to monthly frequency for smoother trends.

Result:

A clean, analysis-ready DataFrame (`df`) with consistent data types, no null values, and well-prepared for modeling and visualization.

Feature Engineering

- **Date-Based Features**: Extracted Year, Month, and Day from `Invoice Date`.

- **Lag Features**: Created lag variables for past 1, 2, and 3 months of sales.

- **Rolling Averages**: Used 3-month rolling means to smooth volatility.

- **Categorical Handling**: Label-encoded categorical fields for use in models if applicable.

Exploratory Data Analysis (EDA)

Visualizations:

- **Boxplots**: For `Units Sold`, `Price per Unit`, and `Operating Profit` (to detect outliers).

- **Line Plot**: Showed monthly sales trends over time.

- **Correlation Matrix**: Revealed relationships between numeric variables.

- **Scatter Plot**: Compared `Units Sold` against `Total Sales` for pattern recognition.

Key Insights:

- Seasonal sales spikes observed from July to November.

- Highest sales volume occurs in the ₹100–₹150 price band.

- Apparel and footwear are top-performing product categories.

- South and Southeast regions consistently outperform others.

Forecasting & Modeling

Model Evaluation:

- RMSE (Root Mean Square Error)

- MAPE (Mean Absolute Percentage Error)

- R^2 Score

Strategic Recommendations

1. **Stock Up for Peak Months (Jul–Nov)**
   Historical and predicted trends suggest heavy demand during these months. Early stocking will reduce missed sales.

2. **Focus Pricing Around ₹100–₹150**
   This price range consistently shows higher sales volume and conversion, making it ideal for bundling and promotions.

3. **Double Down on Top Categories**
   Invest more in products like apparel and footwear which are stable and predictably profitable.

4. **Expand in Strong Regions (e.g., South)**
   High-performing regions should receive targeted campaigns and distribution focus to amplify returns.

Summary

This project used robust cleaning, visual exploration, and forecasting foundations to derive business insights and guide decision-making. The deliverables support pricing, inventory, and regional strategy optimization using data-backed methods.