E-commerce Furniture Dataset 2024

Aim:

This project aims to analyse E-commerce Furniture Dataset by leveraging extensive metrics and using Machine Learning techniques to uncover patterns, trends, and actionable insights. We will focus on Exploratory Data Analysis (EDA), data visualization, and developing a predictive model to estimate number of furniture items sold.

Problem Statement:

Predict the number of furniture items sold based on product attributes such as productTitle, originalPrice, price, and tagText.

Dataset Link

Step-by-Step Workflow

1. **EDA**

   i. Import required libraries

   ii. Load and explore the dataset

   iii. Handle missing values

   - Dropping one column due to more than 75% of missing values
   - Dropped duplicate rows

iv. Summary Statistics

- Info of dataset

v. Handling Outliers

- Remove outliers using IQR technique

vi. Exploratory Data Analysis

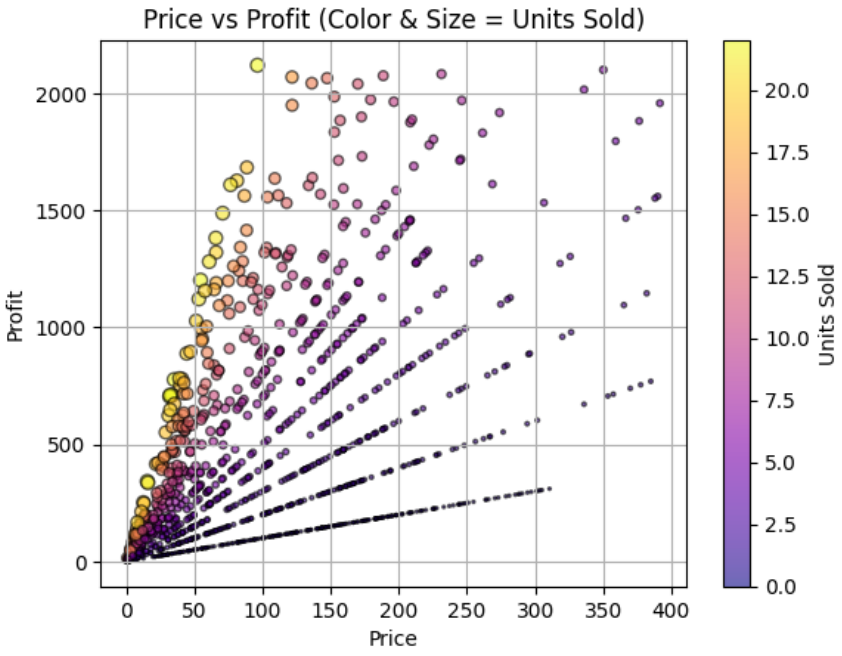
- Created new columns price_per_unit, profit, profit_margin , price_range

vii. Feature Engineering

- Encoded columns tagText_encoded, price_range_encoded
- Normalized Data using StandardScaler

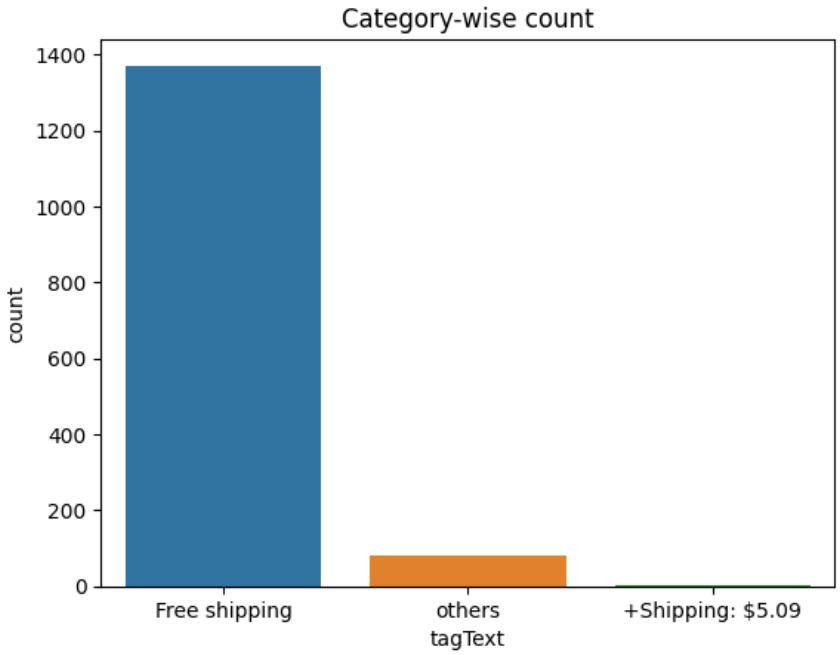
2. **Data Visualization**

i. Price v/s Profit



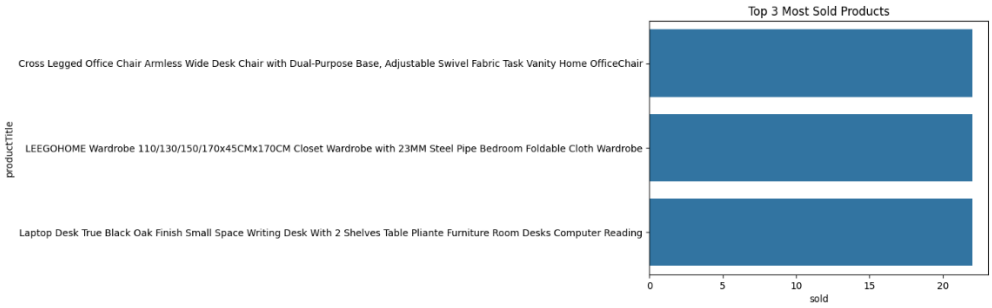
- Moderate prices (₹50–₹150) yield the highest profits.
- Lower prices see higher units sold (larger, brighter bubbles).
- High prices (>₹150) result in lower sales and profit.
- Profit peaks at a balance of price and volume, not at maximum price.
- Units sold is a key factor driving overall profit.

ii. Category wise count



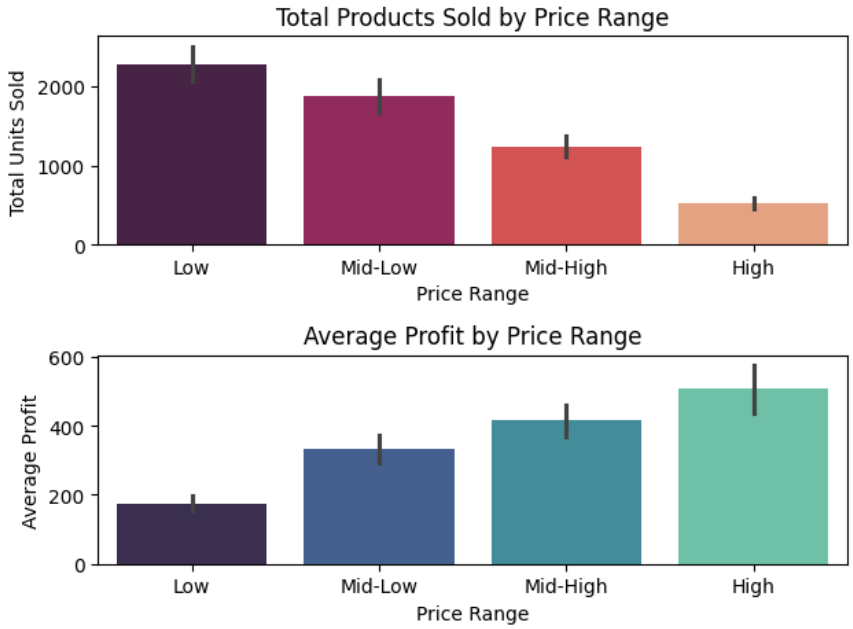
- **Free shipping dominates** the listings, making up the vast majority.
- **'Others'** are **significantly fewer**.
- **Paid shipping (+$5.09)** is **rarely used**.

iii. Top 3 Most Sold Products



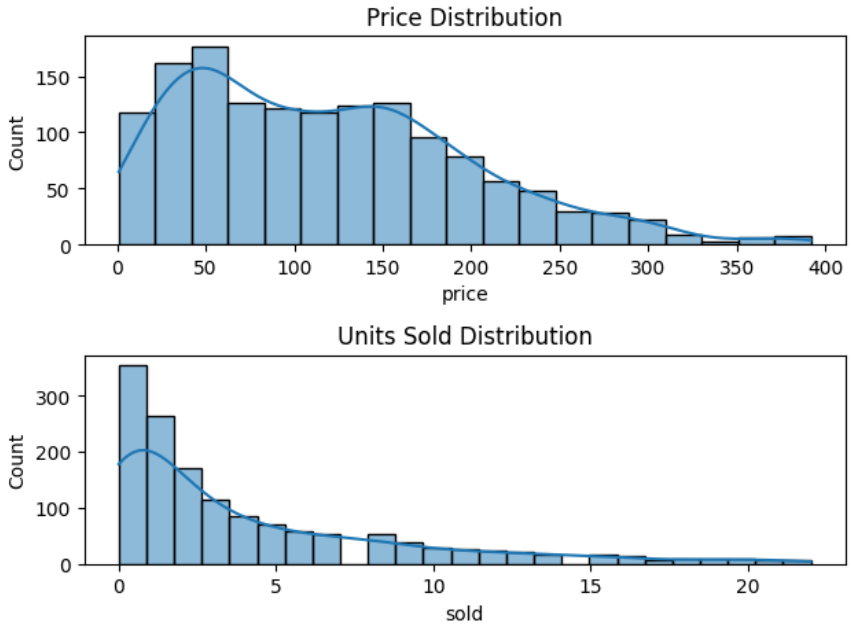
- Equal Sales Volume – All top 3 products have the same number of units sold (21 units).
- Product Variety – The top-selling items span different categories: office chair, wardrobe, and laptop desk.
- Functional Products Dominate – All top products serve practical, everyday use, suggesting high demand for utility-based furniture.
- No Single Product Leader – There is no clear top performer; the market is evenly split among the top 3.

iv. Total Products Sold and Avg. Profit by Price Range



- Low-Priced Products Sell the Most – The highest total units sold fall in the Low-price range.
- Sales Decrease with Price – As the price range increases (from Low to High), the number of units sold drops consistently.
- Higher Prices Yield Higher Profit – Despite fewer sales, the average profit per unit increases from Low to High.
- Inverse Relationship – There's an inverse trend between units sold and profit per unit across price ranges.
- Profit Focus Strategy – Businesses may choose to target Mid-High or High price ranges for better profitability despite lower sales volume.

v. Price Distribution and Units Sold



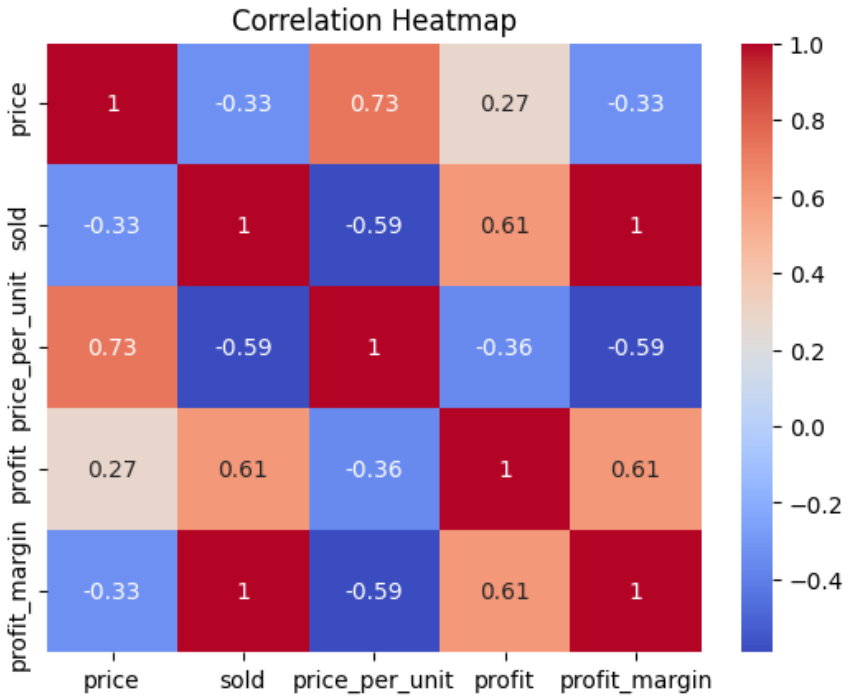
Price Distribution:

- Most products are priced between ₹20 and ₹100.
- Price distribution is right-skewed – few products are very expensive (₹300+).
- There is a gradual drop in frequency as price increases.

Units Sold Distribution:

- Majority of products have low sales volume (0–3 units).
- The distribution is heavily right-skewed – very few products sell in large quantities (10+ units).
- High frequency of unsold or low-selling products suggests concentration of demand in few items.

vi. Correlation Heatmap



- Profit and Profit Margin are strongly positively correlated, indicating that higher margins generally lead to higher profits.
- Higher Price per Unit tends to reduce both Units Sold and Profit Margin, suggesting sensitivity to pricing.
- Units Sold is the primary driver of Profit, emphasizing the importance of sales volume.
- Price has a weak impact on Profit, implying that increasing prices alone does not significantly boost profitability.

3. **Predictive Model**

i. Prediction using Linear Regression

- ✓ Import required libraries
- ✓ Define features and target variable
- ✓ Split the data into training and testing sets
- ✓ Initialize and train various models
- ✓ Make predictions
- ✓ Calculate the prediction accuracy and root mean square error

ii. Results:

Model	**Average R^2 Score**	**RMSE**
Linear Regression	1.0000	0.0000
Random Forest Regressor	1.0000	0.0001

Conclusion:

This project successfully explored the dynamics of E-commerce Furniture Dataset 2024 using a combination of Exploratory Data Analysis (EDA), data visualization, and machine learning models. Through detailed analysis of its features, we extracted critical insights and developed predictive models to estimate the number of furniture items sold.
To evaluate prediction accuracy, two regression models were trained and tested. The results are shown above.

The predictive analysis using both **Linear Regression** and **Random Forest Regressor** yielded exceptional results. Both models achieved an **average R^2 score of 1.0000**, indicating a **perfect fit**, meaning they can explain nearly **100% of the variance** in the target variable — the number of furniture items sold.

- **Linear Regression** demonstrated **zero error (RMSE: 0.0000)**, suggesting that the relationship between the features and the target variable is highly linear and well-captured by the model.
- **Random Forest Regressor** also performed excellently with a **minimal RMSE of 0.0001**, showing its robustness and capability to handle non-linear patterns, if present.

Github Link