E-commerce Furniture Dataset Analysis Report

1. Project Objective

The aim of this project is to analyze sales data of furniture products listed on an e-commerce platform (AliExpress). By examining key product features such as **price, sales volume, and shipping conditions**, this project seeks to identify trends that influence customer purchasing behavior and lay the groundwork for predictive modeling.

2. Dataset Overview

- **Source:** Scraped from project pdf link
- **Entries:** 2,000 furniture listings
- **Key Features:**
  - `productTitle`: Name of the furniture item
  - `originalPrice`: Original listing price (many values missing)
  - `price`: Current selling price
  - `sold`: Number of units sold
  - `tagText`: Shipping info and promotional tags (e.g., "Free shipping")

3. Data Cleaning and Preparation

Actions Taken:

- Dropped the `originalPrice` column due to 75% missing values.
- Removed rows with missing shipping info (`tagText`).
- Converted price strings into numeric format by stripping `$` and commas.
- Normalized `tagText` values:
  - Grouped all values except `'Free shipping'` and `'+Shipping: $5.09'` into an `'others'` category.
- Encoded `tagText` using **LabelEncoder** for ML-readiness.

4. Exploratory Data Analysis (EDA)

a) Price Distribution

Most products are priced under $100. The price range is wide, but lower-cost items dominate.

b) Sales Distribution

Sales volume (`sold`) is heavily skewed — many products have low or zero sales, with a few high-selling outliers.

c) Shipping Impact

- '**Free shipping**' is by far the most common tag (~94% of products).
- Products with free shipping seem more likely to have higher sales, suggesting a potential influence.

d) Price vs. Sales Relationship

A scatter plot between price and units sold shows **no strong linear relationship**, but certain patterns suggest that mid-range priced products may sell better.

5. Feature Engineering

- **Price Normalization**: Converted price to numeric format for analysis.
- **Shipping Label Encoding**: Applied `LabelEncoder` to convert categorical shipping data into numeric form.
- Future possibilities include:
    - Using TF-IDF to extract keyword relevance from `productTitle`
    - Engineering a discount rate if `originalPrice` data is recovered

6. Model Readiness (Planned)

Although no models were trained in this version, the dataset is now cleaned and prepared for machine learning tasks like:

- **Regression modeling** to predict `sold` based on `price` and `tagText`
- **Clustering** to group similar products based on price/sales behavior
- **Classification** to identify high-performing products

7. Key Insights

- **Pricing Strategy:** The majority of high-selling items fall in the mid to low price range.
- **Shipping Incentives:** “Free shipping” is prevalent and could positively affect sales.
- **Low Sales Volume:** Most items sold very few units, pointing to either niche products or high competition.

8. Applications

Applications:

- **E-commerce Optimization:** Inform pricing and shipping strategies.
- **Sales Prediction:** Train ML models to forecast sales volume.
- **Product Tagging:** Automate tagging strategies for better visibility.