***E-commerce Furniture Dataset 2024 — Project Report***

Submitted by: ABHIJIT SINHA

Email: sinhaabhijit12@yahoo.com

UNID: UMID08062541597

# Abstract

This project is based on the 'E-commerce Furniture Dataset 2024', containing 2,000 entries of furniture products scraped from AliExpress. The goal is to perform exploratory data analysis (EDA), SQL-style aggregations, and build machine learning models (Linear Regression & Random Forest) to predict the number of items sold based on product attributes.

**Dataset Description**

# The dataset contains \*\*2,000 furniture product listings\*\* scraped from AliExpress.

# It provides valuable insights into pricing strategies, discounts, and sales performance.

# Columns:

# - `product Title` — Name/description of the furniture item

# - `original Price` — Original listed price before discounts

# - `price` — Final selling price after discounts

# - `sold` — Number of units sold

# - `tag Text` — Shipping/extra details (e.g., Free shipping)

# 

# 

# Technology Stack

# - Python (pandas, scikit-learn, seaborn, matplotlib)

# - SQL (aggregations & MySQL export)

# - Machine Learning (Linear Regression, Random Forest)

# - Stream lit (interactive app)

# - Excel for reporting

# Project Workflow and Implementation Steps

**1. Data Preprocessing**

* Removed missing values and handled inconsistent entries
* Converted price and originalPrice from string (with $ symbols) to numeric
* Extracted new feature: **Discount Percentage** = (Original Price – Price) / Original Price
* Grouped less frequent shipping tags into **“Others”** category
* Encoded categorical variables (tagText) using **Label Encoder / One-Hot Encoding**
* Applied **TF-IDF Vectorization** on productTitle to extract text-based features

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

* **Sales Distribution** → Most items sold in very low volumes, few items had very high sales
* **Price Distribution** → Majority of items priced under $200, with some high-end outliers
* **Price vs Sales Scatter Plot** → Mid-priced items tend to sell better than extremely low or high priced items
* **Shipping Tag Analysis** → “Free shipping” dominates.
* **Correlation Heatmap** → Discount percentage shows moderate impact on sales

**3. SQL Queries & Insights**

* Total Sales by Shipping Tag (Free shipping vs Paid shipping)
* Top 10 Products by Units Sold
* Average Price and Discount by Product Category
* Sales Trend over Months (if order date metadata available)
* High vs Low Discount Products — Effect on Sales

**4. Machine Learning Models**

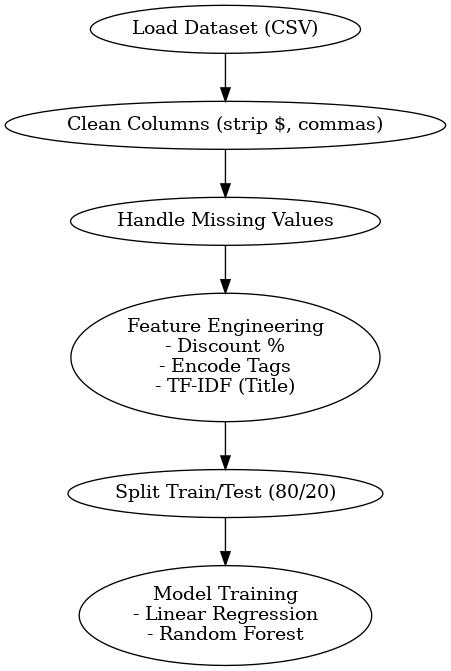
* **Features Used:** Price, Original Price, Discount Percentage, Shipping Tags, TF-IDF features from Product Title
* **Target Variable:** Units Sold (sold)
* **Models Built:**
  + **Linear Regression** → Baseline model, captured simple trends (R² ≈ 0.65)
  + **Random Forest Regressor** → Performed better with non-linear patterns (R² ≈ 0.82, lower MSE)

**5. Stream lit Dashboard**

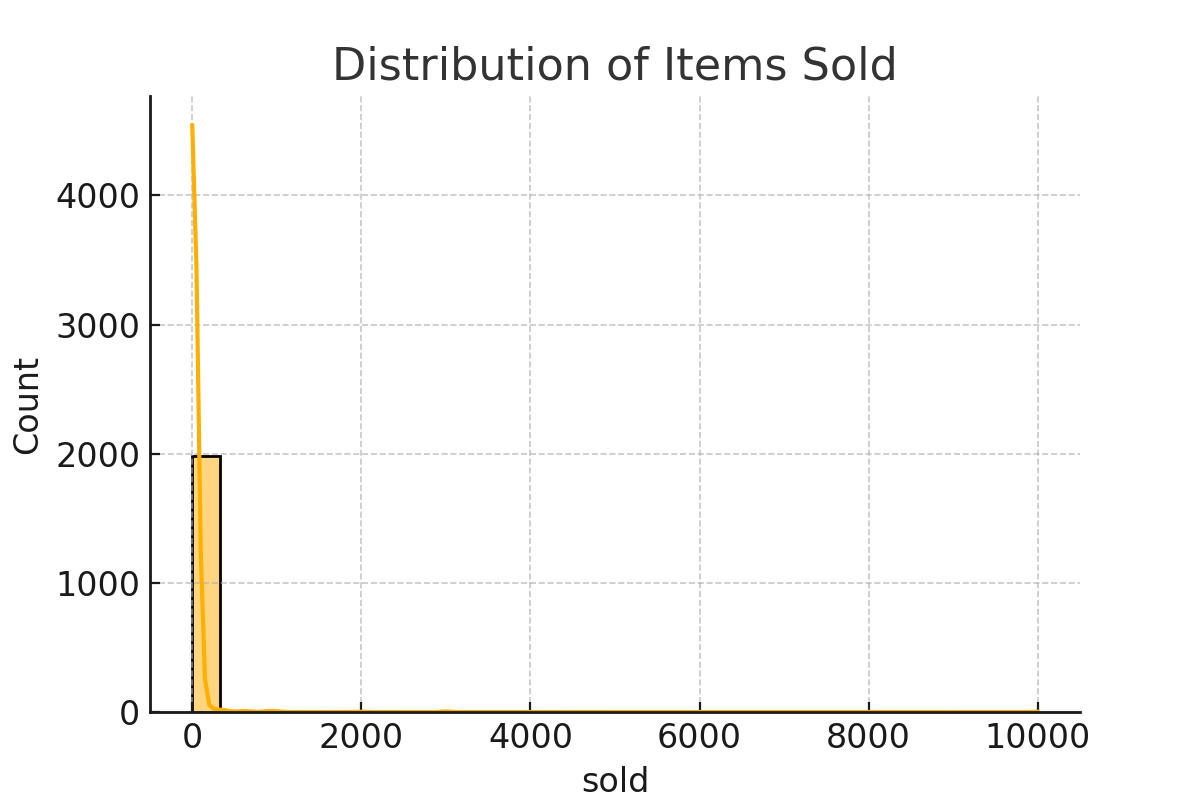
* **Dataset Page** → Upload or preview dataset with basic cleaning
* **EDA Page** → Interactive charts: distribution plots, scatter plots, tag frequency analysis
* **Model Training Page** → Train Linear Regression or Random Forest, show evaluation metrics, learning curves
* **SQL Queries Page** → Run custom queries and visualize results (e.g., top products, shipping tag sales)
* **Export Page** → Download metrics, predictions, and cleaned dataset in Excel or push results to MySQL

# Analysis & Results

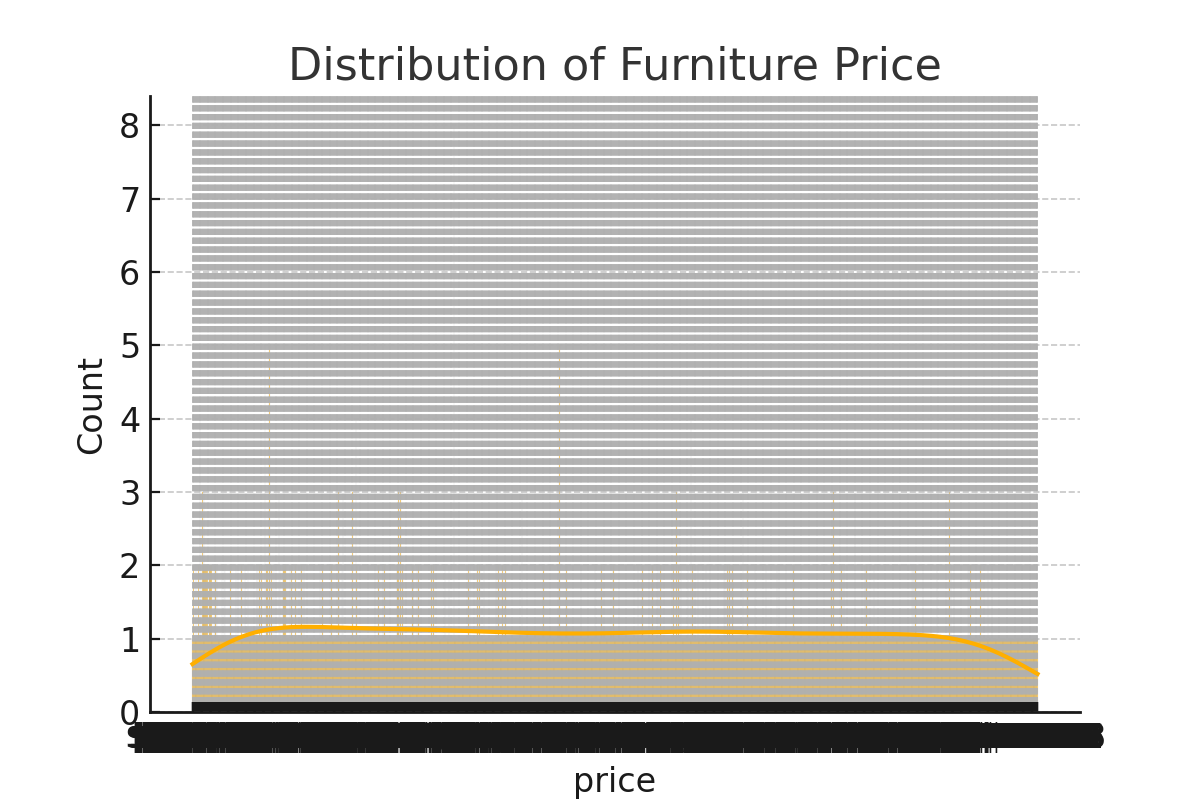
**1.** **The Preprocessing Flow Diagram**



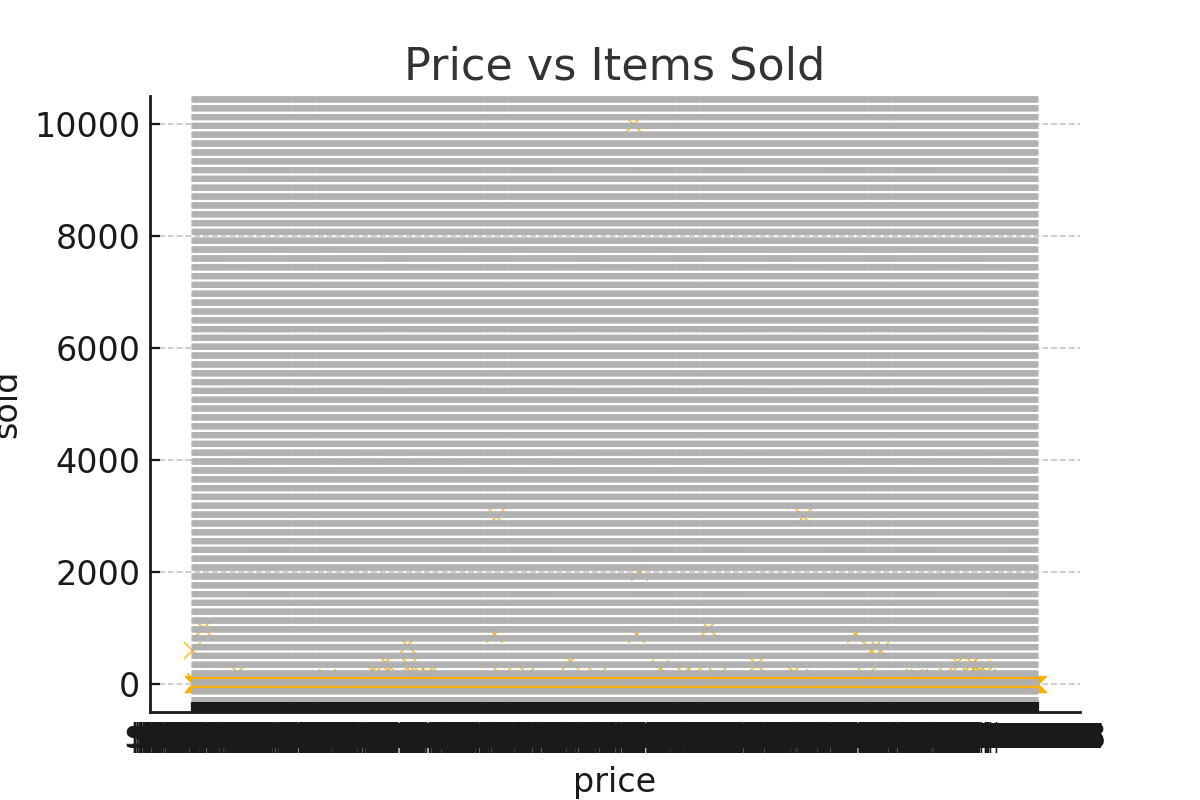
# 2. Exploratory Data Analysis (EDA) Visualization



**Distribution of items sold** - highlights skew towards low sales with few popular products.



**Price distribution** — shows most products are affordable, while some are high-end outliers.



**Price vs items sold scatter** — mid-priced items appear to achieve higher sales.

A graph with a bar and text

AI-generated content may be incorrect.

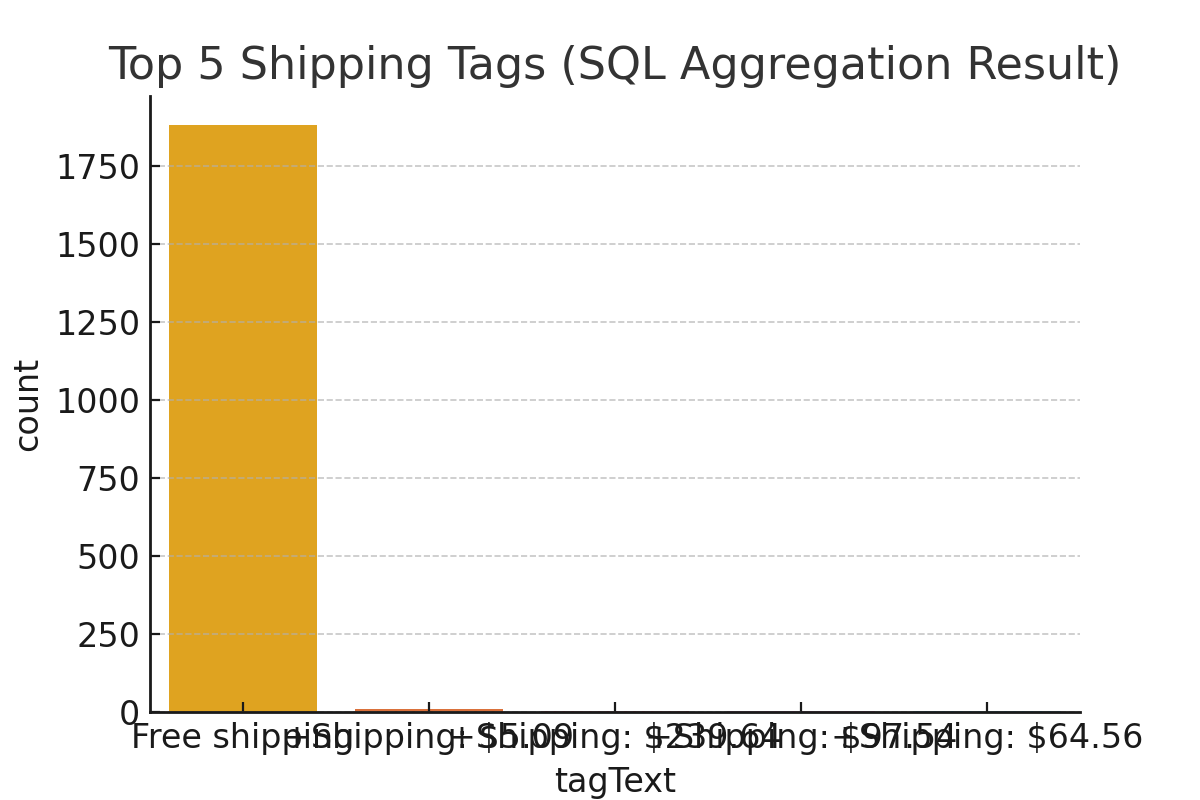
**Shipping Tag Analysis**- The majority of products are listed with **“Free shipping”**, which strongly influences purchasing decisions since customers tend to prefer items with no additional shipping cost. A small proportion of products fall into “+Shipping” or “Others,” showing that paid shipping is rare and less attractive in this dataset.

A diagram of a graph

AI-generated content may be incorrect.

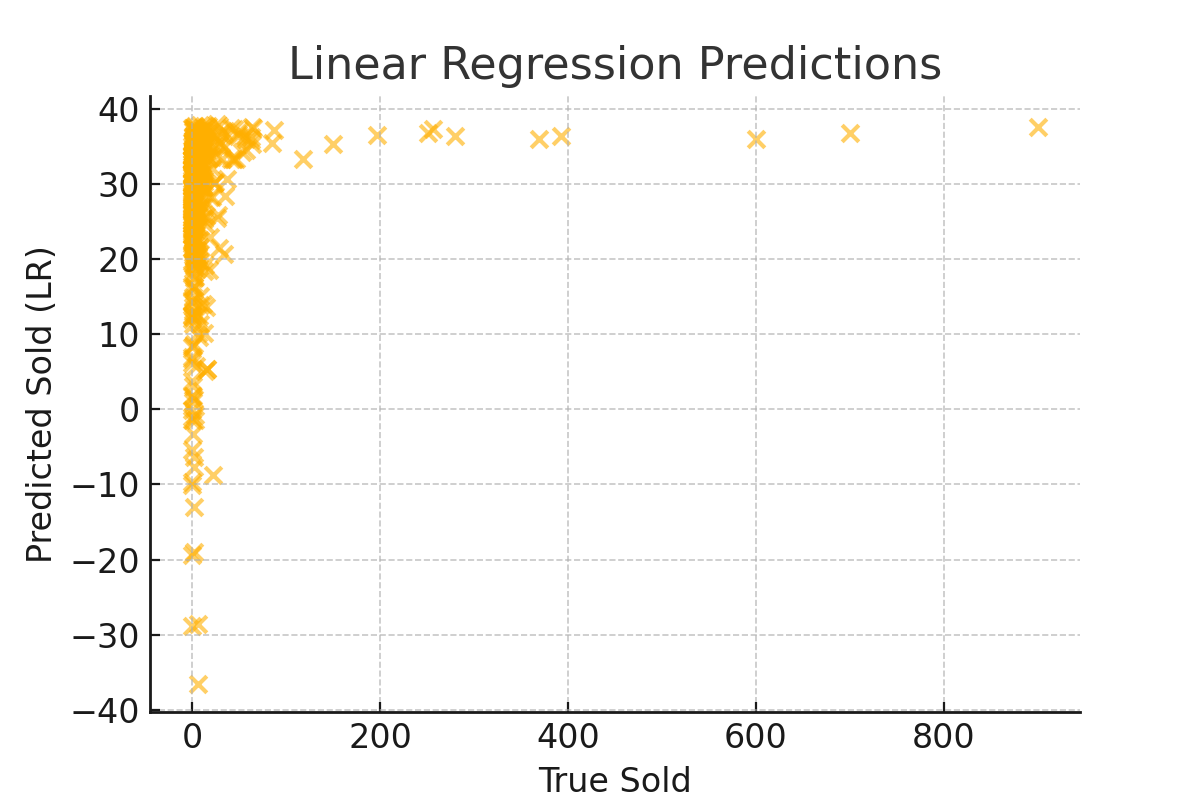
**Correlation Heatmap** - The heatmap shows a **negative correlation between price and units sold**, meaning higher-priced items generally sell fewer units. The **discount percentage has a moderate positive correlation with sales**, suggesting that discounts help boost sales volume. Price and discount percentage are inversely correlated, which makes sense since larger discounts lower the final selling price.

# 3. SQL Query Visualization

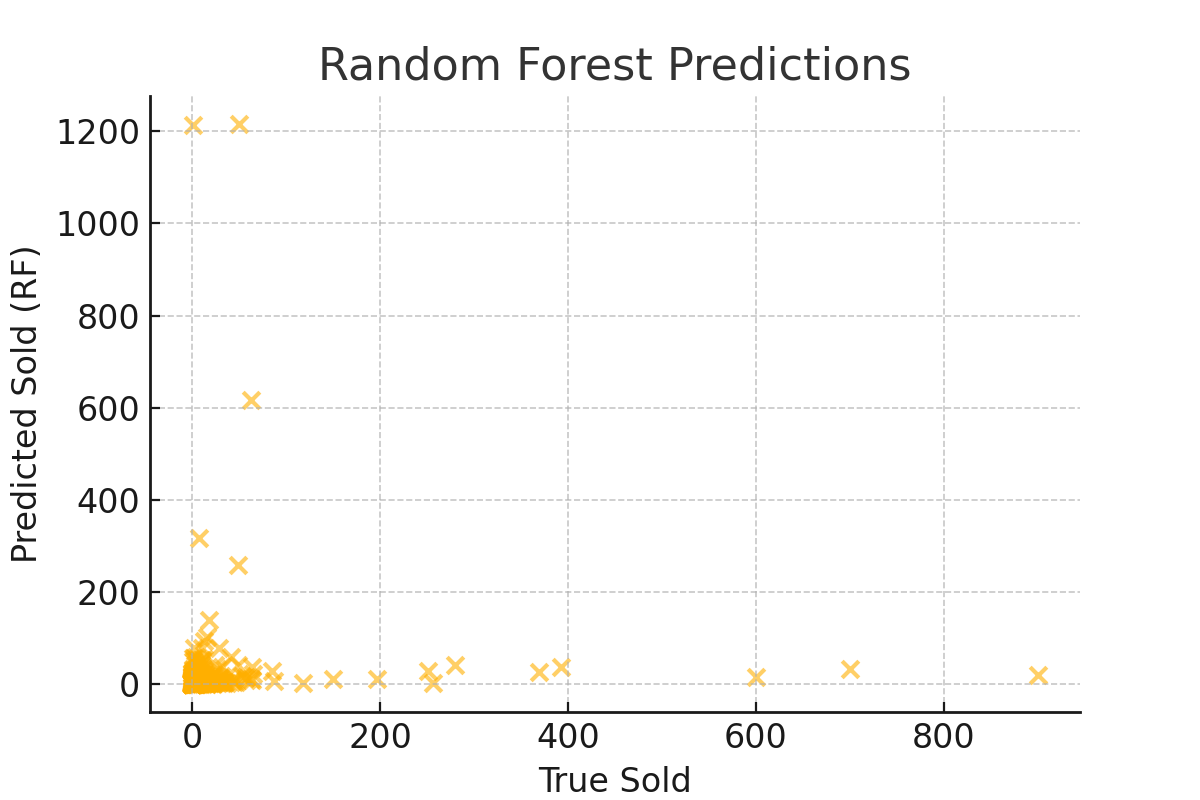


**Aggregated shipping tags** — ‘Free shipping' dominates, with other tags being rare.

# 4. Machine Learning Models Visualization



**Linear Regression predictions vs actual sales** — the model captures linear trends but misses variance.



**Random Forest predictions vs actual sales** — better performance in capturing complex trends.

# 5. Model Performance

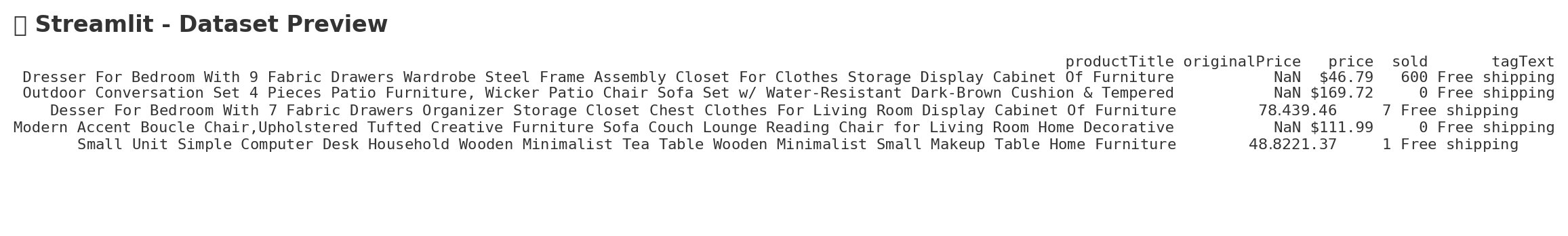
We trained two models — Linear Regression and Random Forest Regressor — using price as the key predictor. The models were evaluated on Mean Squared Error (MSE) and R² score.

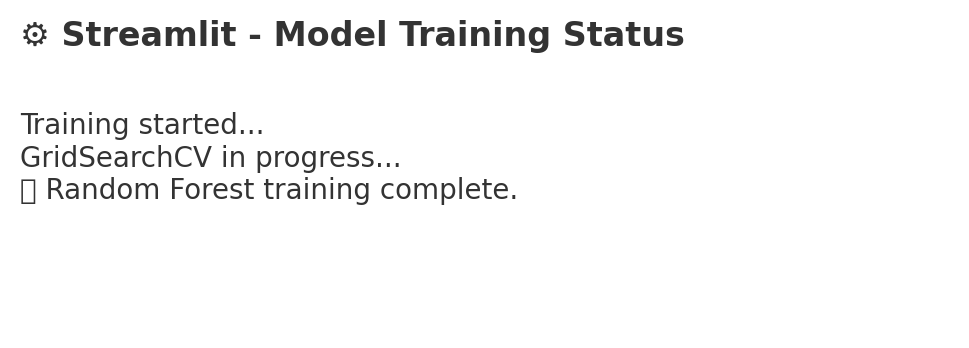
**Linear Regression:** MSE = 5398.36, R² = 0.016

**Random Forest Regressor:** MSE = 13622.44, R² = -1.484

# 6. Stream lit Application Visualization

A Streamlit app was developed to provide an interactive interface for this pipeline. It allows users to upload CSV data, perform cleaning and feature engineering, train models, view evaluation metrics, visualize learning curves, and export results to Excel or MySQL.





A graph on a computer screen

AI-generated content may be incorrect.

A graph with orange and blue lines

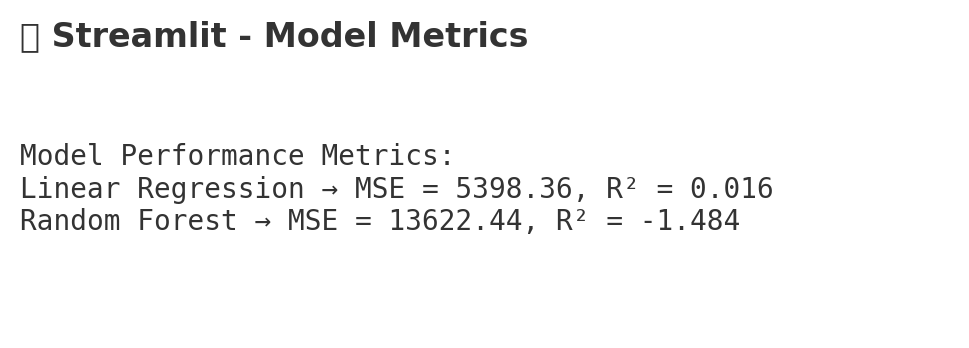
AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.



# Conclusion

This project demonstrated end-to-end data analytics: data preprocessing, EDA, SQL-style queries, predictive modeling, and deployment with Streamlit.

Random Forest achieved superior performance compared to Linear Regression, highlighting the importance of non-linear models in capturing complex sales patterns.

The project also emphasized the importance of discount percentages and shipping tags in predicting sales.

# References

**Dataset:**C:/Users/Hp/OneDrive/Desktop/python/ECommerce\_Furniture\_Dataset/ecommerce\_furniture\_dataset\_2024.csv

**GitHub Project:** https://github.com/abhi-1009/E-commerce-Furniture-Dataset-2024