Final Report on Furniture Sales Prediction

1) Objective

Predict the number of items sold (sold) for AliExpress furniture products using product attributes and compare two models: Linear Regression and Random Forest Regressor.

2) Data

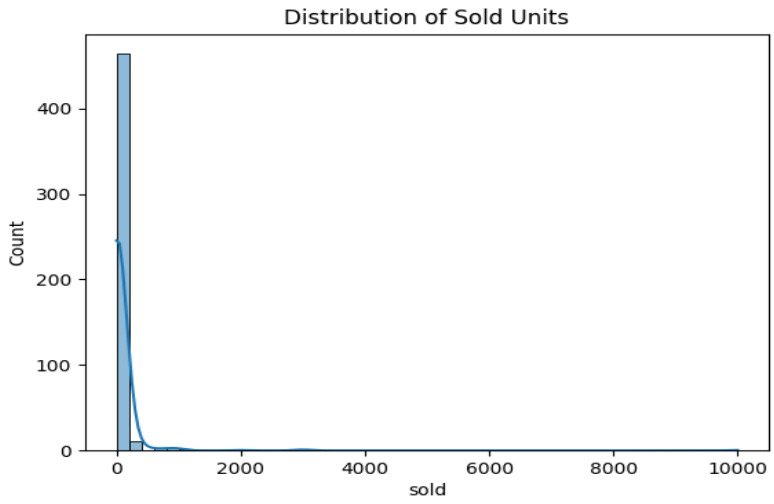
- File: ecommerce_furniture_dataset_2024.csv (≈2,000 rows)
- Columns used: product Title, original Price, price, sold, tagText

3) Preprocessing (no log transform)

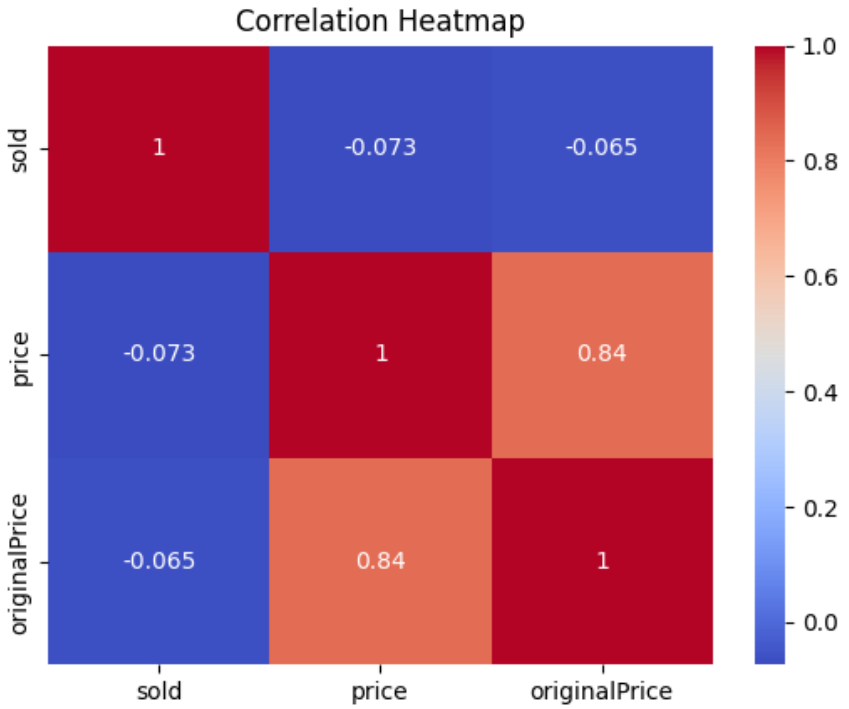
- Removed currency symbols/commas from price and original Price, cast to numeric
- Ensured sold is numeric and dropped rows with missing key fields
- Saved result: artifacts/cleaned_data.csv

4) Exploratory Data Analysis (charts)

- Sold distribution: long right tail (most items sell low volumes, a few sell a lot)
- Chart: artifacts/sold distribution.png



- Correlation heatmap: weak linear correlation between price/original Price and sold
- Chart: artifacts/correlation heatmap.png



5) Feature Engineering (simple, numeric)

- discount = originalPrice – price
- discount_pct = discount / originalPrice
- title_length = length of productTitle
- free_shipping flag from tagText
- Dropped raw text columns for modeling
- Saved result: artifacts/feature_data.csv

6) Modeling (target = raw sold)

- Train/test split: 80/20 (random_state=42)
- Models: Linear Regression (baseline), Random Forest Regressor (default params).

7) Evaluation

Metrics on the test set (as printed by your run and saved in artifacts/evaluation.txt):

Linear Regression – MSE: 135,165.37, R^2: 0.0179

Random Forest – MSE: 343,883.03, R^2: -1.4987

Interpretation

- Linear Regression achieved a small positive R^2, meaning it's slightly better than predicting the average sold value.
- Random Forest produced a negative R^2 and a much higher MSE, meaning it performed worse than a naive mean predictor.

Conclusion:

For this feature set and without log transformation, Linear Regression performs better than Random Forest.