Furniture Sales Analysis & Prediction Report

1. Introduction

This project focuses on analyzing furniture sales data, identifying key trends, and building predictive models to forecast product sales. The goal is to help the store optimize pricing, shipping policies, and inventory planning by leveraging data insights.

---

2. Data Preprocessing

The raw dataset contained missing values and inconsistencies. The following cleaning steps were applied:

- **Filled missing originalPrice values** with the corresponding price (assuming no discount).
- **Filled missing shipping_cost values** with 0.0 (assuming free shipping).
- Verified that all missing values were handled and data types were correct.

---

3. Exploratory Data Analysis (EDA) Insights

3.1 Sales Distribution

- The sales distribution is **heavily skewed**, with most furniture products selling in very low quantities.
- A few items achieve **exceptionally high sales** (up to ~10,000 units), significantly impacting total revenue.
- **Recommendation:** Prioritize marketing and inventory for these high-performing items while finding ways to boost sales for low-volume products.

3.2 Price vs Sales

- Most products are priced **below $500** and sell in relatively low quantities.
- A few **low-priced** items achieve exceptionally high sales.
- Higher-priced products generally have **much lower sales volumes**.
- **Insight:** Affordability plays a key role in driving demand.

3.3 Price Distribution

- Most products are priced between **$50 and $200**.
- A small number of premium products are priced **above $1,000**, creating a long-tail distribution.

- **Conclusion:** The store primarily targets the budget to mid-range market, with limited high-end offerings.

3.4 Original Price vs Sales

- Most products have an **original price below $500** and sell in low quantities.
- High-priced products (**>$1,000**) have very limited sales.
- **Insight:** Lower pricing is strongly associated with higher demand.

3.5 Shipping Cost Impact

- Products with **free shipping** tend to sell in much higher volumes than those with paid shipping.
- Paid shipping items generally have lower and more consistent sales.
- **Recommendation:** Offering free shipping could significantly boost sales.

3.6 Correlation Observations

- Strong **price–original price correlation** → Discounts are generally proportional to the original price.
- Sales are concentrated at **low price ranges**.
- A few extreme outliers exist (very high sales or prices), which could skew averages.
- Sparse sales in the **mid–high price range**.

---

4. Machine Learning Models

Two regression models were tested to predict sold (log-transformed):

Model	MSE	R^2
Linear Regression	1.3465	0.2728
Tuned Random Forest	1.1691	0.3686

Observation:
The **Tuned Random Forest** outperformed Linear Regression in both MSE and R^2, showing better predictive power and lower error.

---

5. Model Performance Analysis

5.1 Actual vs Predicted – Random Forest

The scatter plot compares predicted and actual sales.

- The **red dashed line** represents perfect predictions.
- Many points lie close to the diagonal, showing good accuracy for most products.
- Predictions for **high-selling items** show greater deviation, suggesting the model struggles with extreme values.
- The spread is consistent with the moderate **R^2 score (0.3686)**.

Business Takeaway:
Improving predictions for high-selling products could significantly enhance inventory and marketing efficiency.

---

6. Key Business Recommendations

1. **Boost marketing & inventory** for top-selling products while finding ways to promote low-selling items.
2. **Maintain affordable pricing** (below $500) to capture the largest demand segment.
3. **Implement free shipping** where possible to increase sales volumes.
4. **Explore targeted discounts** for mid-price items to stimulate demand.
5. **Enhance predictive models** by incorporating more features and handling outliers to improve high-sales forecasting.

---

7. Conclusion

This analysis highlights that **pricing and shipping policies** are crucial drivers of sales.
The predictive model (Random Forest) provides a reasonable starting point for demand forecasting, but further optimization is needed for high-value outliers.
By applying these insights, the store can make **data-driven decisions** to increase revenue and improve customer satisfaction.