

Sales Dataset Analysis: Final Report and Presentation

Final Report

1. Introduction to Dataset

The sales_data_sample.csv dataset contains **2,823 rows** and **25 columns**, capturing global retail transactions from **2003 to 2005**. Key variables include order date, product line, customer location, quantity ordered, unit price, and sales revenue. The dataset allows comprehensive analysis of revenue drivers across time, geography, and product segments.

2. Data Cleaning Summary

Step	Action	Notes
Missing Values	STATE: Filled with "Unknown" ADDRESSLINE2: Left as-is	~1,500 and ~2,500 missing respectively
Duplicate Rows	Checked and removed	0 duplicates found
Date Format	Converted ORDERDATE to datetime format	Enabled time-series analysis
Text Standardization	Capitalized STATUS, DEALSIZE, and PRODUCTLINE	Unified categories
Numeric Conversion	Ensured QUANTITYORDERED, PRICEEACH, SALES are float	Prepared for analysis

Cleaning Log Summary:

- Left ADDRESSLINE2 untouched due to non-essential role.
 - Replaced missing STATE with "Unknown" for clarity.
 - No duplicate records found.
 - ORDERDATE transformed to datetime to support trend analysis.
 - Standardized categorical fields to avoid inconsistent labels.
-

3. Exploratory Data Analysis (EDA) Report

Descriptive Statistics:

- **Quantity Ordered:** Mean = 35.1, Range = 6 to 97
- **Price Each:** Mean = \$83.7, Range = \$26.88 to \$100
- **Sales:** Mean = \$3,553.9, Max = \$14,082.8

Frequency Analysis:

- **Top Product Lines:** Classic Cars (967), Vintage Cars (607), Motorcycles (331)
- **Top Countries by Orders:** USA (1,004), Spain (342), France (314)

Outlier Detection:

- Using IQR, 81 outlier sales records > \$4,508 were found—suggestive of bulk deals or anomalies.

Visualizations:

- **Bar Chart:** Sales by Product Line

Classic Cars dominate, followed by Vintage and Motorcycles.

- **Bar Chart:** Sales by Country

USA leads, followed by Spain and France.

- **Histogram:** Quantity Ordered

Typical order size falls between 25–45 units.

- **Histogram:** Price Each

Prices cluster between \$70–\$100, indicating premium product range.

- **Box Plot:** Sales Distribution

Highlights outliers and spread of sales values.

4. Advanced Analysis Summary

Total Sales by Product Line:

- Classic Cars: \$3.9M

- Vintage Cars: \$1.9M
- Motorcycles: \$1.17M

Sales by Country:

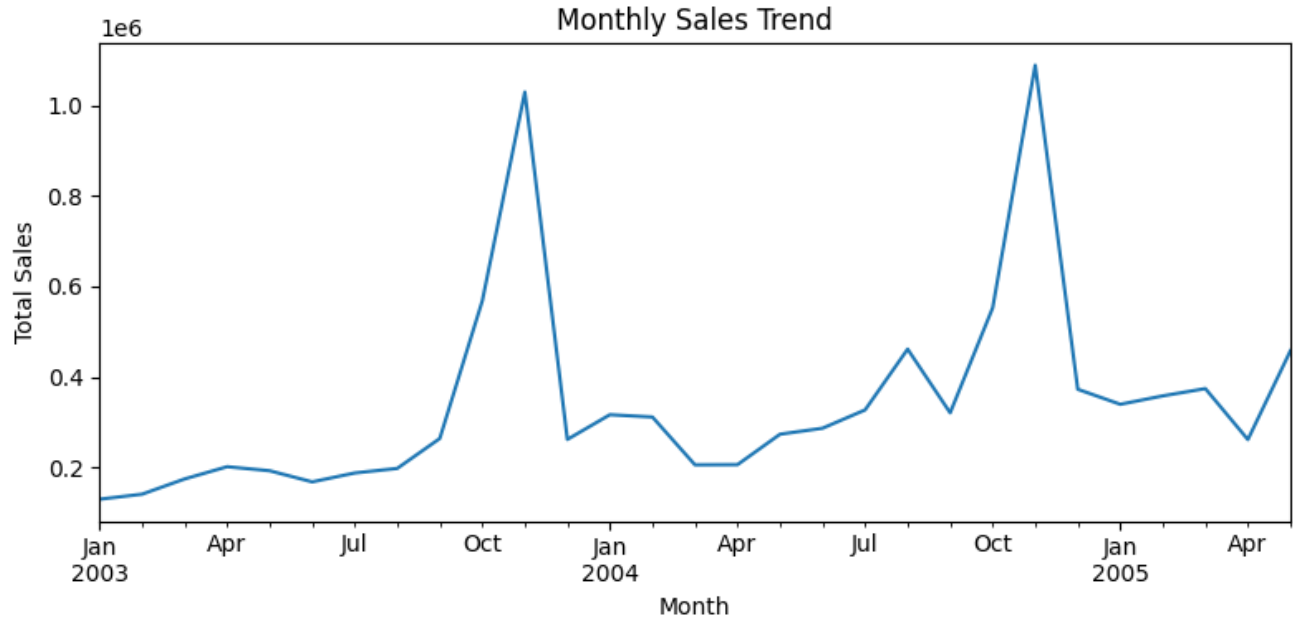
- USA: \$3.63M
- Spain: \$1.22M
- France: \$1.11M

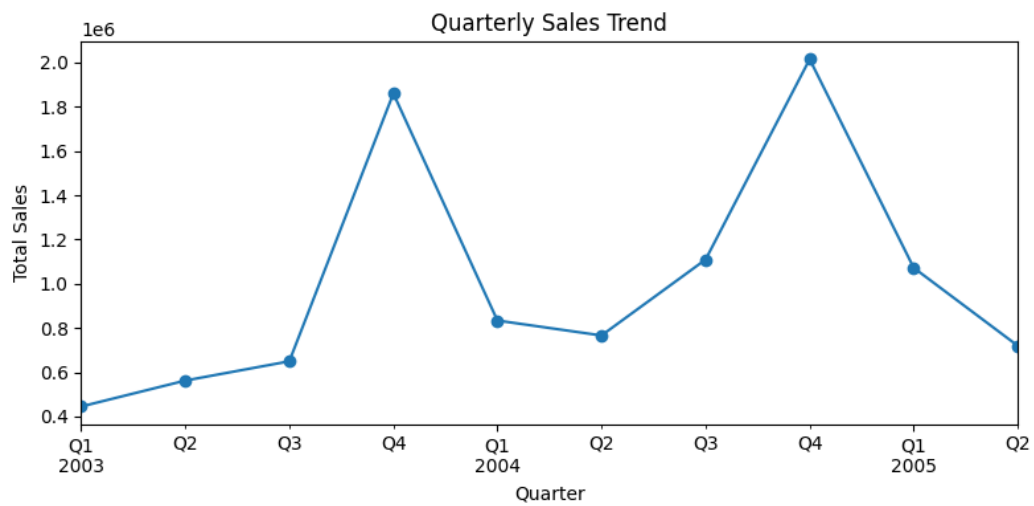
Top 5 Products by Revenue:

- S18_3232: \$288K
- Followed by S24_1937, S18_4409, etc.

Sales Trends:

- **Monthly & Quarterly Sales Line Charts** show clear peaks mid-year and Q4, indicating seasonal demand.

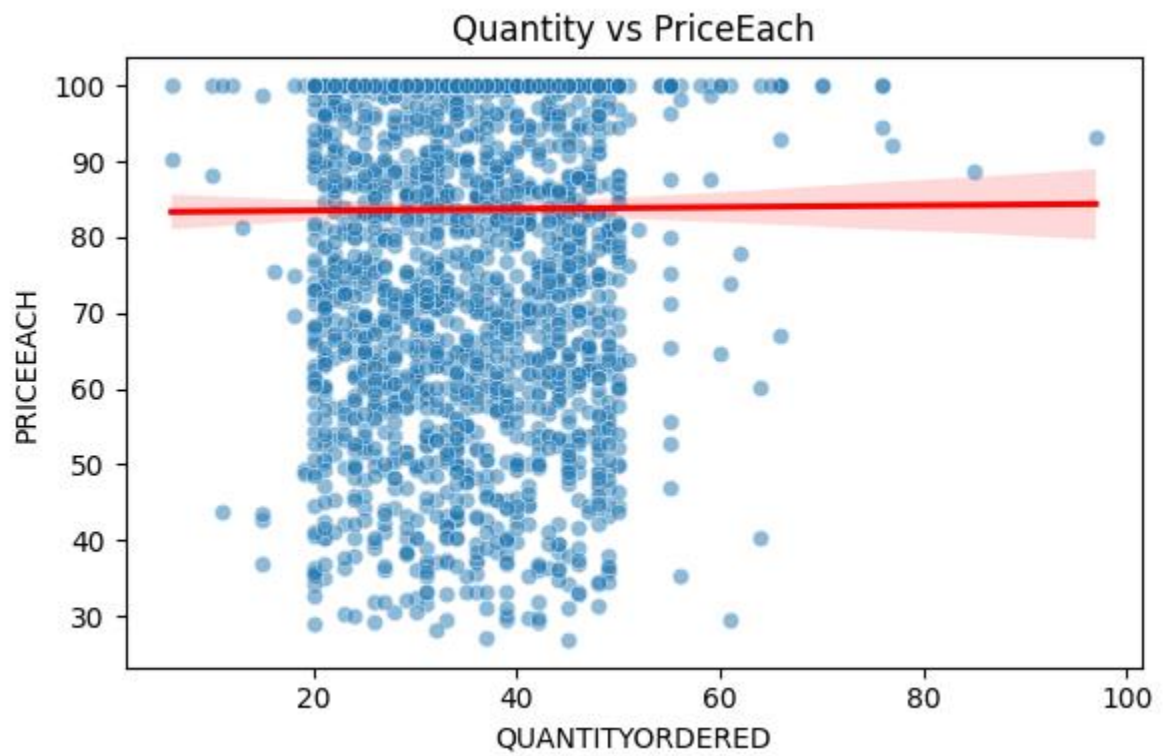




Quantity
vs. Price

Correlation:

- $r \approx 0.0056 \rightarrow$ No meaningful relationship; higher quantities do not lead to discounts.



Key Takeaways:

- Classic Cars are the strongest revenue driver.
- USA is the primary market with room to expand in EMEA.
- Significant seasonality observed.
- High-value orders need further business review.