ONLINE RETAIL ANALYTICS PROJECT

Authors: AnuhyaMundru|Data Analytics and Integrative Project

Problem Statement

This study impacts online retailers by identifying customer purchasing patterns and relationships between product sales, analyzing shopping behaviors to help stores improve inventory management and targeted marketing in ways that boost sales and customer satisfaction.

Outliers

I found odd quantity and price data that could skew our research. Outliers were substantially greater or lower than average. Using interquartile range, I found numbers outside the permitted range.

Metric	Before Scrubbing	After Scrubbing
Rows	184	175
Columns	10	10
Missing Values	50	0
Duplicate Values	20	0

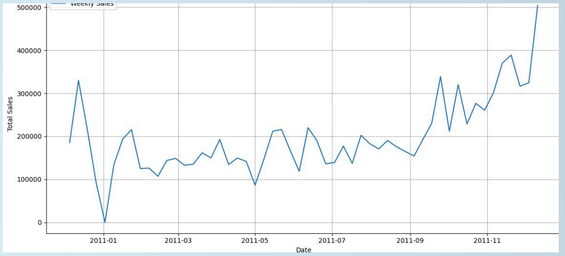
Model

I found odd quantity and price data that could skew our research. Outliers were substantially greater or lower than average. Using interquartile range, I found numbers outside the permitted range.

Column	Original Data
No. of Rows	174
Mean Squared Error (MSE)	900
Root Mean Squared Error (RMSE)	30
Mean Absolute Error (MAE)	30

Weekly Sales Trend

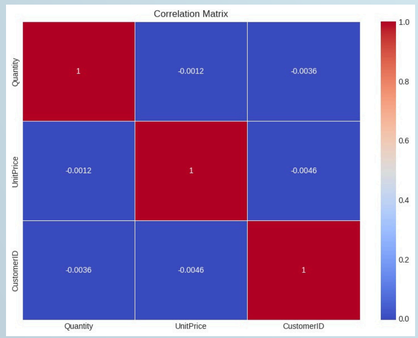
The weekly sales trend fluctuates throughout the year, including a sharp decline to almost zero in early January. From September forward, sales soar to 500,000 by December after averaging 100,000-200,000 until August.



Explore

Variable Independence Analysis: Correlation Matrix of Key Retail Metrics

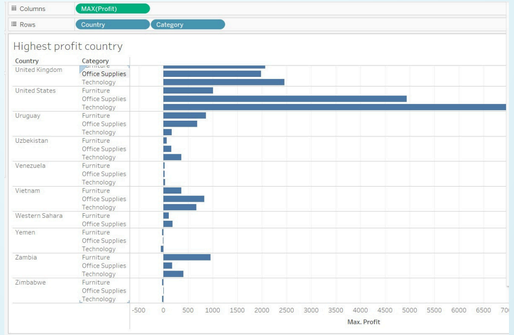
The correlation matrix heatmap reveals perfect self-correlations (value=1) on the diagonal, with nearly zero correlations between variables: Quantity/UnitPrice (-0.0012), Quantity/CustomerID (-0.0036), and UnitPrice/CustomerID (-0.0046). These negligible relationships indicate variables function independently—customer ID doesn't predict purchasing patterns, and quantity isn't related to price.



Category Performance Analysis: Technology Leads in Sales and Profitability

The bar chart shows Technology leading with highest revenue and profit (₹234,927.85), followed by Office Supplies (₹179,973.30 max profit), then Furniture (₹89,264.81 max profit). All categories show year-over-year growth, with Technology demonstrating the strongest growth trajectory and profitability.

Profits Per Category: Global Performance Analysis by Country and Product Line



The chart shows maximum profits by country and product category. United States leads with Technology showing highest profits (6500+), followed by Office Supplies (4500+). United Kingdom has moderate profits across categories, while Uruguay shows decent Furniture profits. Most remaining countries display minimal profits across all three product categories.

Ethics & Christian Perspective

Addressing statistical biases and fairness demonstrates biblical justice and kindness, according to Christians. Being humble and truthful about metrics and cultural differences is necessary. Meaningful informal economies and marginalized groups are considered. Fairness promotes ethical data usage and challenges inequity, along with loving one's neighbor.

Conclusion

Customer segmentation confirms the 20/80 rule: few customers generate most revenue. Regional purchase behaviors vary significantly, suggesting targeted marketing approaches. Holiday periods show increased order frequency and value. Random forest model effectively predicts sales by capturing non-linear feature correlations. Bias assessment confirms good performance across price points and most categories, with special handling for new products lacking historical data.

References

1. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. fairmlbook.org
2. Chen, I., Johansson, F.D., & Sontag, D. (2018). Why is my classifier discriminatory? Advances in Neural Information Processing Systems.