Data-Driven E-commerce Intelligence: Customer Analytics and Revenue Forecasting System

Applied Data Science and Business Intelligence Project

Project Repository: `https://github.com/soufianboukir/ecom-analytics-platform`

Dashboard URL: `https://ecom-analytics-forecasting-platform.streamlit.app`

Soufian Boukir

Abstract

This project develops an end-to-end e-commerce analytics and forecasting platform using a synthetic Amazon-style dataset of 100,000 transactions (2020-2024). The system integrates sales analytics, customer and product analysis, and revenue forecasting within an interactive Streamlit dashboard. Among four evaluated forecasting models, XGBoost achieved the best performance (RMSE: 61,570, MAPE: 3.71%). Forecasts indicate stable but stagnant revenue through 2026, highlighting the need for strategies such as market expansion, discount optimization, and customer loyalty programs.

Keywords: E-commerce Analytics, Time Series Forecasting, Business Intelligence

1 Introduction

E-commerce platforms generates a lot of transactional data across orders, customers, products, etc. but this data is often not be used for analyzing or decision making due to it's unstructured form. Key questions such as identifying high-value customers, understanding the impact of pricing and discounts, analyzing geographic performance, and predicting future revenue trends remain unanswered without a structured analytical system. This project solves this problem by designing an integrated data analytics and forecasting system that transforms raw sales data into meaningful insights and predictive models. The objective is to enable data-driven decision-making by combining descriptive analytics, and time-series forecasting into a unified framework for e-commerce business intelligence.

2 Dataset Understanding and Preparation

2.1 Dataset Description

Amazon sales dataset, This dataset contains 100,000 synthetic Amazon-style e-commerce sales transactions, designed to closely resemble real-world online retail behavior. With 20 clean and well-structured columns, it captures detailed information about customers, products, pricing, payments, logistics, and order outcomes.

Order Details
OrderID, OrderDate, OrderStatus, SellerID

Customer Information
CustomerID, CustomerName, City, State, Country

Product & Revenue
ProductID, ProductName, Category, Brand, Quantity, UnitPrice, Discount, Tax, ShippingCost, TotalAmount, PaymentMethod

	Quantity	UnitPrice	Discount	Tax	ShippingCost	TotalAmount
count	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	3.001400	302.905748	0.074226	68.468902	7.406660	918.256479
std	1.413548	171.840797	0.082583	74.131180	4.324057	724.508332
min	1.000000	5.000000	0.000000	0.000000	0.000000	4.270000
25%	2.000000	154.190000	0.000000	15.920000	3.680000	340.890000
50%	3.000000	303.070000	0.050000	45.250000	7.300000	714.315000
75%	4.000000	451.500000	0.100000	96.060000	11.150000	1349.765000
max	5.000000	599.990000	0.300000	538.460000	15.000000	3534.980000

Figure 1: Descriptive statistics of numerical columns.

2.2 Data Cleaning

Data cleaning is the process of preparing data for analysis by removing or fixing data that is incorrect, incomplete, or duplicated within a dataset. This dataset is already completly cleaned and it's perefect for next stages of analytics.

2.3 Exploratory Data Analysis

Exploratory Data Analysis used to explore, summarize, and visualize data to understand its structure, detect patterns, identify anomalies, and check relationships between variables before applying any machine learning or statistical models.

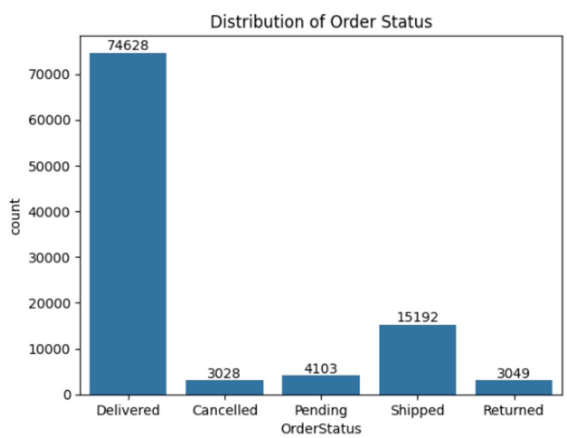

Figure 2: Order status distribution.

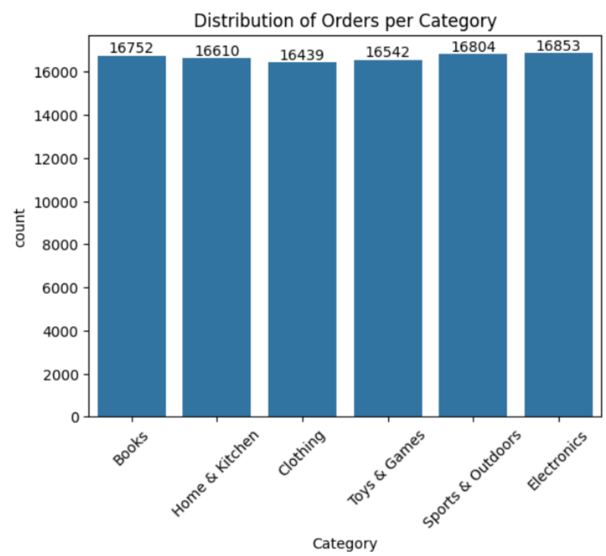

Figure 3: Product category distribution.

Figure 2: The Delivered status is dominant, accounting for more than 70% of all orders, followed by the Shipped status.

Figure 3: All categories exhibit a nearly uniform order distribution, with approximately 16k orders per category.

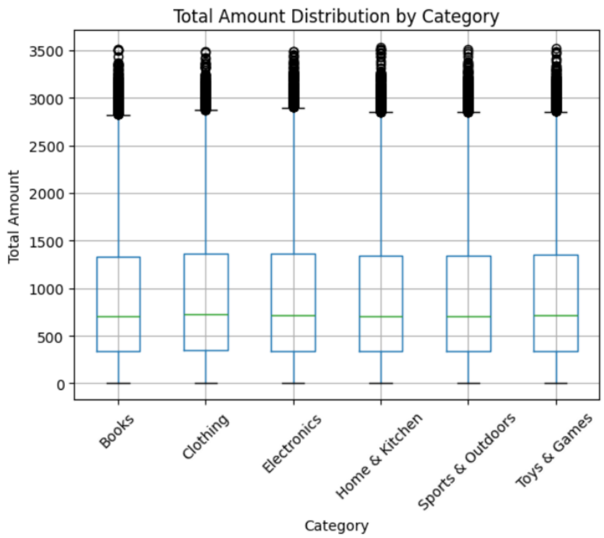

Figure 4: Total amount by category.

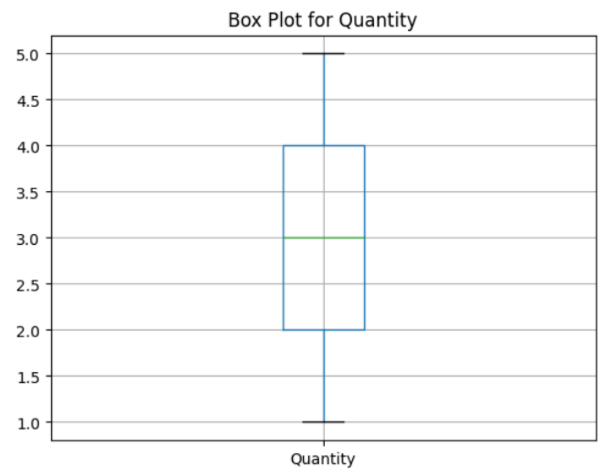

Figure 5: Box plot for quantity.

Figure 4: All categories show similar total amounts and comparable average order values, with a few outliers present.

Figure 5: Approximately 50% of orders have a quantity between 2 and 4 units, with an average order quantity of 3 units.

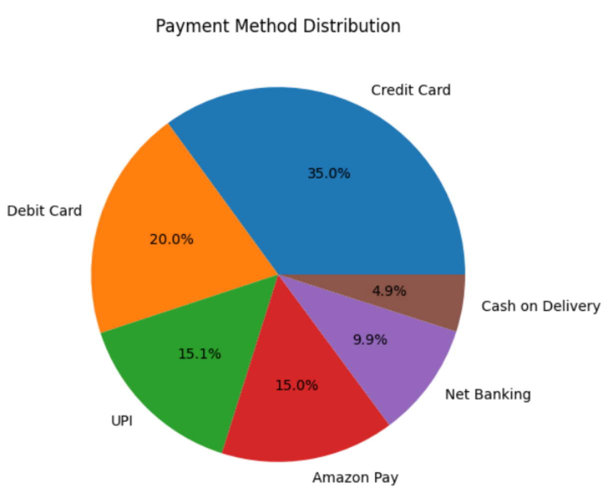

Figure 6: Payment method distribution.

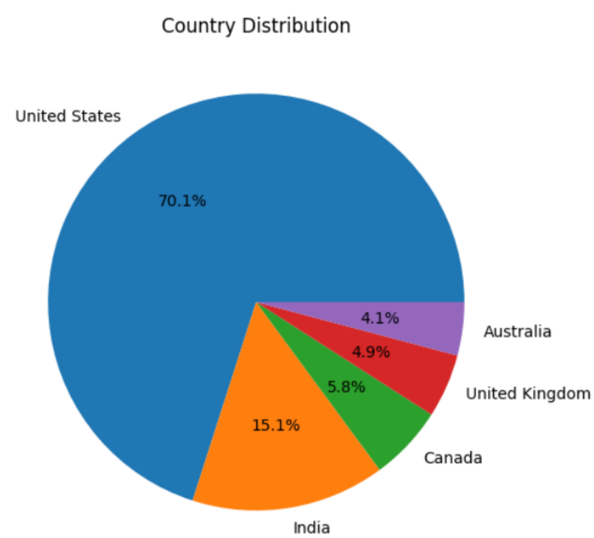

Figure 7: Country disribution.

Figure 6: The Credit card payment method is dominant, accounting for more than 35%.

Figure 7: The US accounts for the majority of orders, representing more than 70% of the total, followed by India.

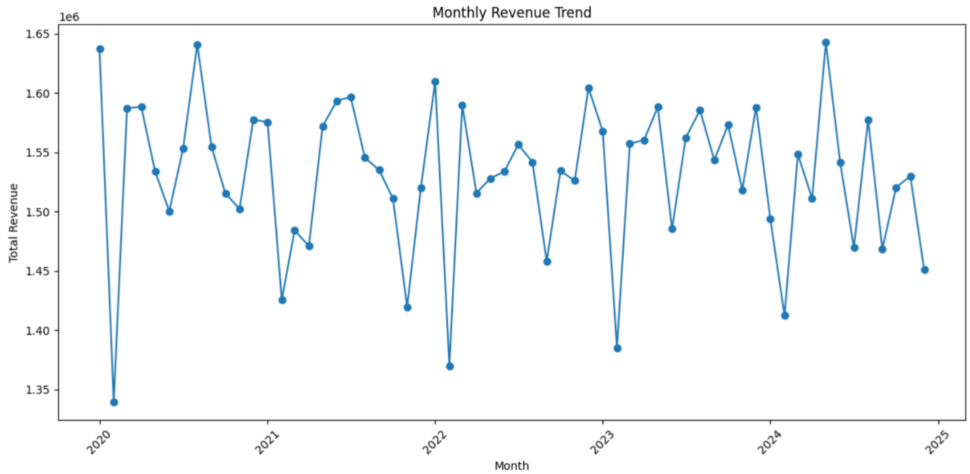

Figure 8: Monthly revenue trend.

Figure 8: Revenue fluctuates consistently between approximately 1.34M$ and 1.65M$ over the five-year period, with no clear upward or downward trend, indicating relatively stable business performance. Additionally, Month 2 records the lowest revenue in almost every year.

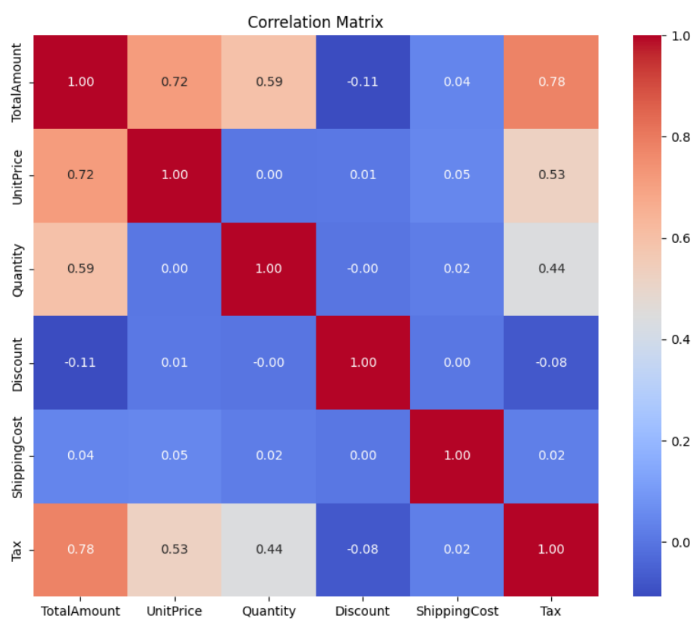

Figure 9: Correlation matrix for numerical columns.

Figure 9: TotalAmount is primarily influenced by Tax (0.78), UnitPrice (0.72), and Quantity (0.59), while Discount shows a slight negative relationship. Shipping-Cost has minimal impact on revenue, and UnitPrice and Quantity appear largely independent.

3 Modeling

We applied multiple predictive models to forecast monthly e-commerce revenue and evaluate their performance under a time-series framework. The modeling phase is designed to compare both simple statistical approaches and more advanced machine learning techniques in order to identify the most effective method for capturing revenue patterns over time.

Four models are implemented: a Naive baseline model, Linear Regression, XGBoost, and Prophet. The **Naive model** assumes that future revenue will follow the most recent observed value (tommorow's revenue is toddday's revenue), serving as a reference benchmark. **Linear Regression** and **XGBoost** use engineered lag and temporal features to learn relationships between past values and future revenue. **Prophet** is used as a statistical forecasting model that captures trend and seasonality components directly from the time series.

Model	MAE	RMSE	R^2	MAPE (%)
Naive	71,581.17	79,320.13	-2.2640	4.78%
Linear Regression	54,909.46	65,713.65	-1.2402	3.71%
XGBoost	55,246.91	61,570.98	-0.9667	3.71%
Prophet	70,473.94	83,586.02	-2.6245	4.76%

Figure 10: Model performance comparison.

Figure 10: It is important to note that all models report negative R^2 values, including XGBoost (R^2 = -0.967), which despite achieving the best RMSE and MAPE, still falls below the performance of a simple mean predictor. This is a direct consequence of the dataset's flat and noisy revenue structure — monthly revenue oscillates within a narrow 1.34M$-1.65M$ band with no discernible trend or seasonality, making the historical mean an inherently strong baseline. In such conditions, negative R^2 does not indicate model failure, but rather reflects the difficulty of the forecasting task itself. The selection of XGBoost as the primary forecasting model is therefore based on its relative superiority in RMSE and MAPE, which remain the most operationally meaningful metrics in this context.

4 Dashboard Overview

The dashboard provides an interactive interface that consolidates the results of the analysis and predictive modeling into a set of visual insights. It is designed to support decision-making by presenting key business metrics, customer behavior patterns, product and geographic performance, and revenue forecasting results in a clear and accessible format. Through dynamic visualizations and filters, the dashboard enables users to explore the data from multiple perspectives and understand both historical trends and future revenue projections.

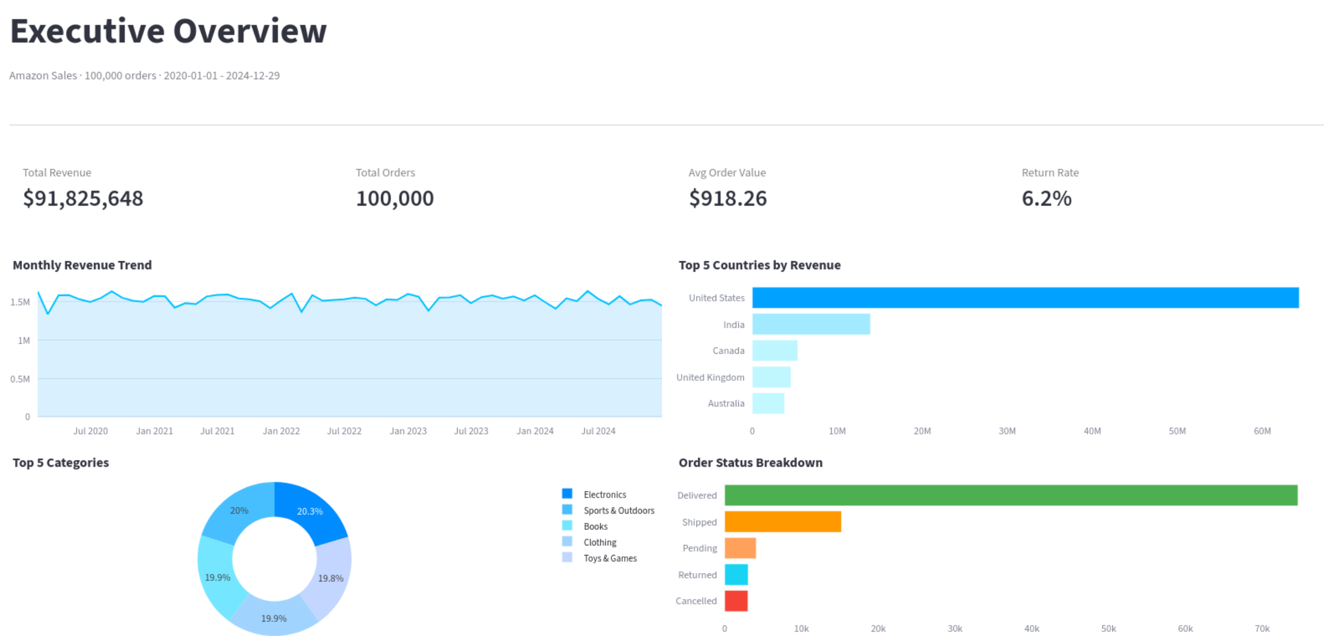

Figure 11: Page-1 Executive Overview.

The Amazon Sales dataset covers 100,000 orders between 2020 and 2024, generating a total revenue of 91.8M$ with an average order value of 918$ and a low return rate of 6.2%. Revenue remained relatively stable throughout the period, averaging around 1.5M$ per month with no major growth or decline trend. The United States represents the dominant market, contributing the majority of total revenue. Product revenue distribution is highly balanced across all categories, with each contributing approximately 20% of total sales. Operationally, most orders are successfully delivered, while returned and cancelled orders remain minimal, indicating a stable fulfillment process.

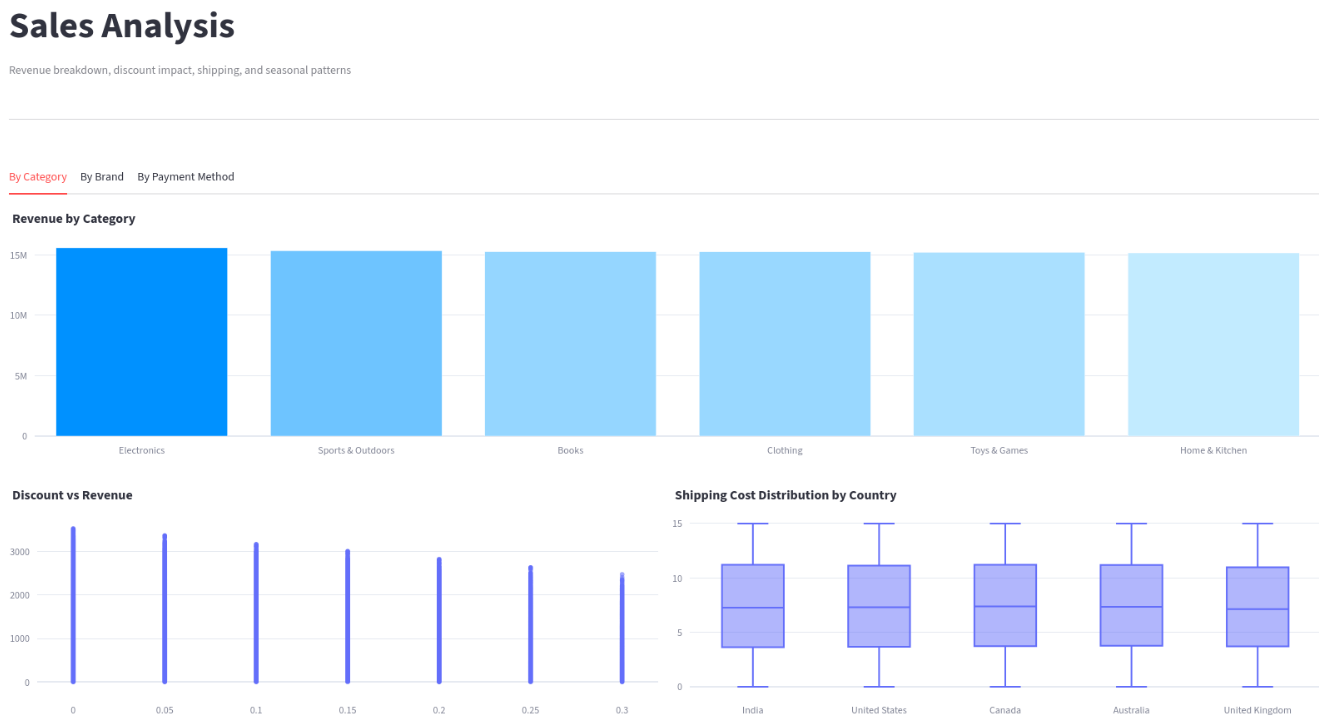

Figure 12: Page-2 Sales Analysis

Revenue by Category shows a highly balanced product mix, with all categories generating nearly identical revenue (15M$ each), indicating no dominant category.

The Discount vs Revenue analysis suggests that discounts have little impact on revenue generation, consistent with the weak negative correlation observed previously.

Additionally, shipping cost distributions remain similar across all countries, reflecting a standardized pricing structure for logistics operations.

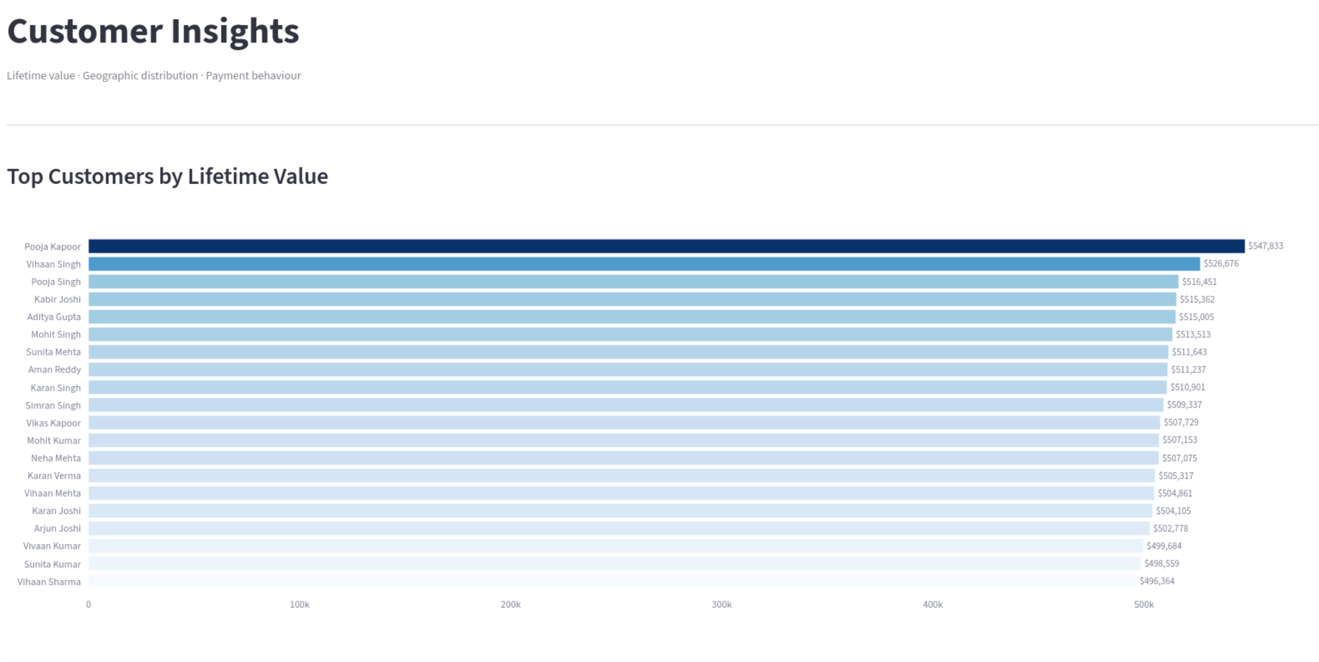

Figure 13: Page-3 Customer Insights

The top 20 customers show a remarkably compressed LTV range of just 496K$-548K$, with Pooja Kapoor leading at 547,833$ and Vihaan Sharma at the bottom with 496,364$ — a spread of less than 52K$ across all 20 customers.

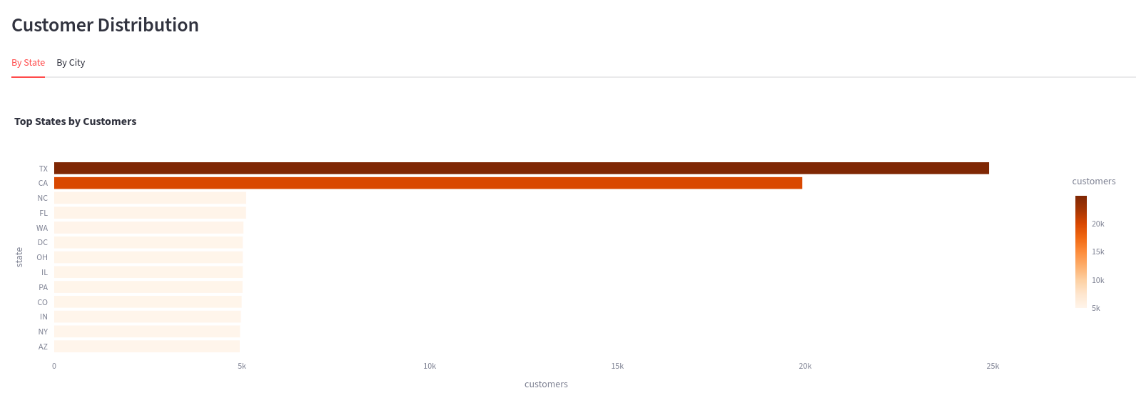

Figure 14: Page-3 Customer Insights

Texas (TX) and California (CA) are overwhelmingly the dominant markets, with 25K and 20K customers respectively — together accounting for a disproportionately large share compared to every other state.

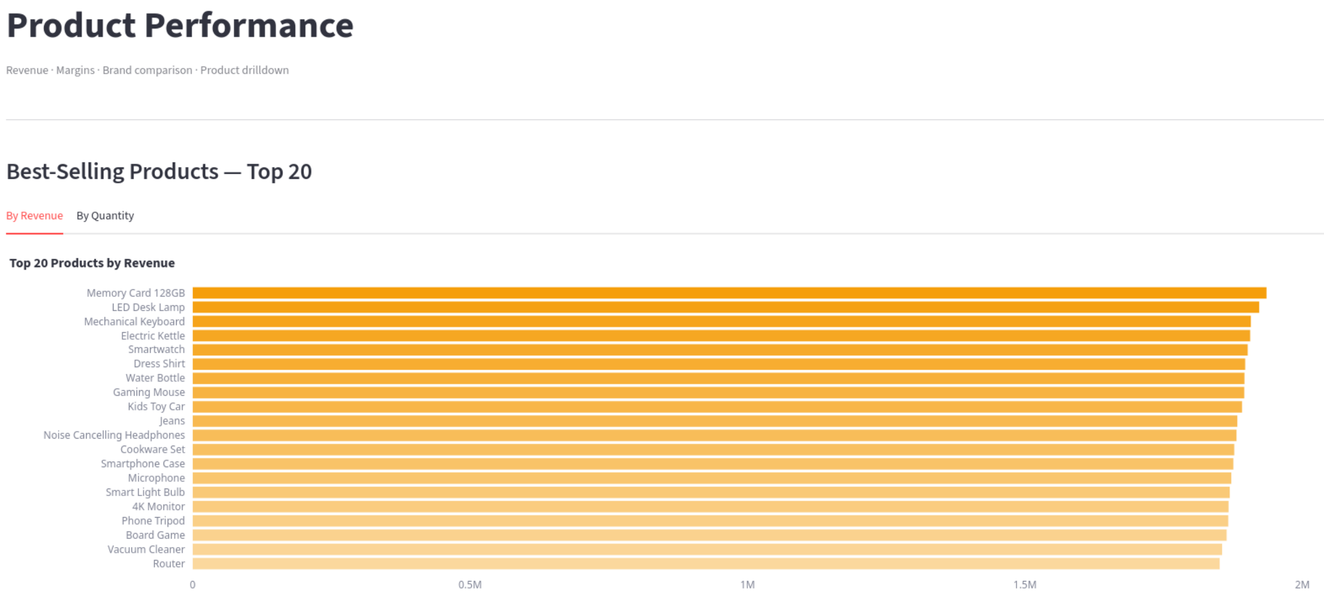

Figure 15: Page-4 Product Performance

The top 20 products by revenue are tightly clustered between 1.5M$ and 2M$, with Memory Card 128GB leading narrowly ahead of LED Desk Lamp and Mechanical Keyboard. The negligible gap between first and last place — spanning diverse categories from Electronics to Clothing to Toys — again confirms the dataset's characteristic pattern of near-perfect revenue uniformity across all dimensions.

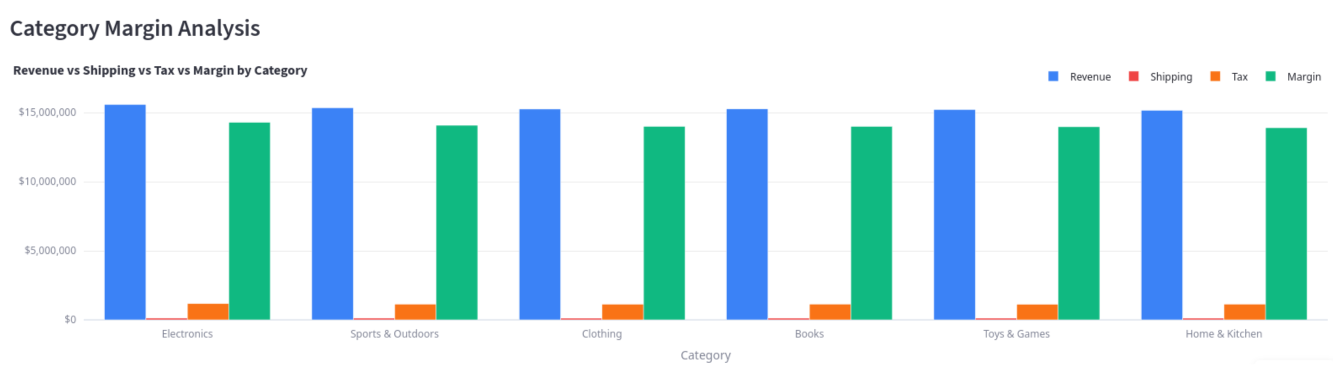

Figure 16: Page-4 Category Margin Analysis

All six categories generate nearly identical revenue of 15M$, with margins consistently landing around 14M$ — indicating a remarkably high and uniform margin rate of roughly 93% across the board. Tax (1M$) is the only meaningful deduction, while Shipping Cost is negligible (nearly invisible in the chart), suggesting either subsidized or free shipping across all categories

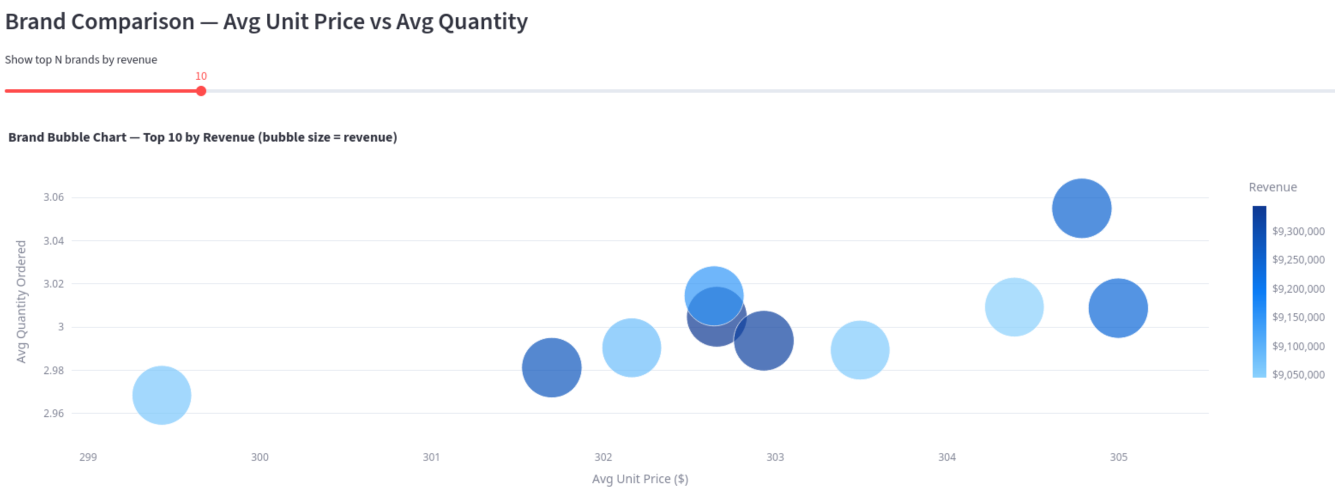

Figure 17: Page-4 Brand Comparison

All top 10 brands are compressed into an extremely narrow price range of 299-305 and quantity range of 2.96-3.06 units confirming that no brand stands out as a premium or budget player. Revenue across brands similarly clusters tightly between 9.05M$ and 9.3M$, with the UrbanStyle brand (highest price 3.05, highest quantity 3.06) achieving the best combined performance.

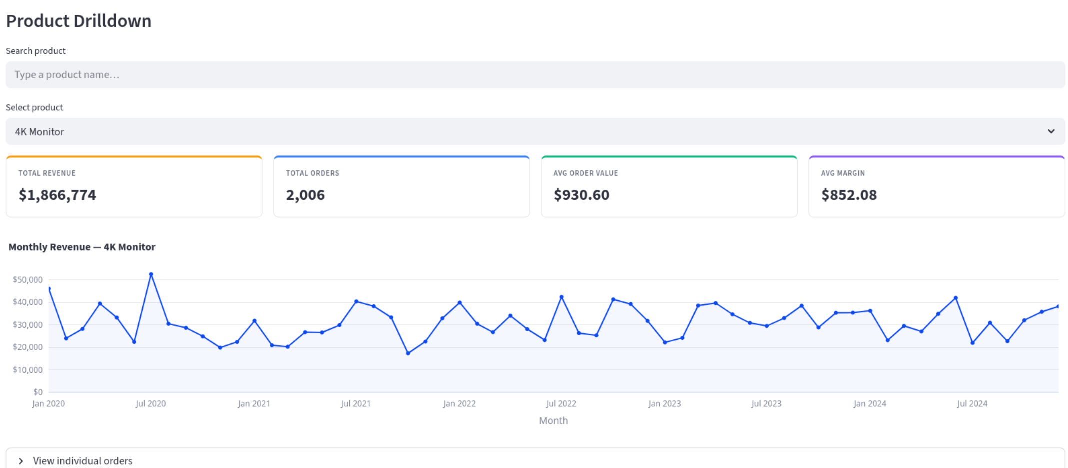

Figure 18: Page-4 Product Drill Down

The Product Drilldown acts as a product-level investigation tool, letting users search and select any product to instantly view its isolated KPIs — total revenue, orders, average order value, and margin — without being diluted by category or brand aggregations.

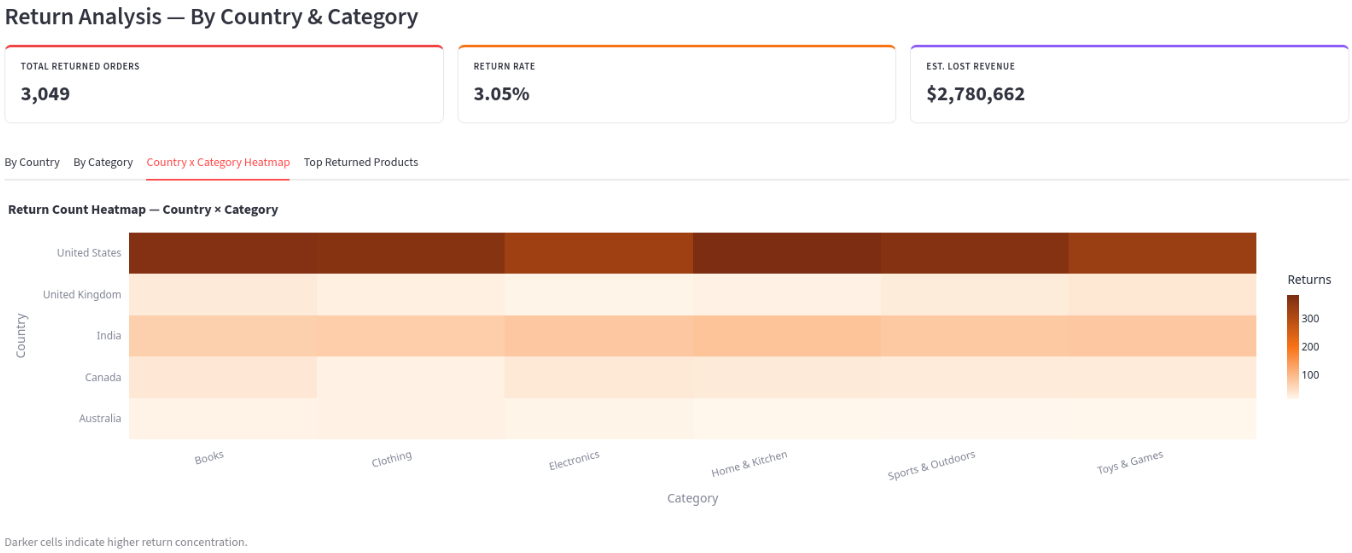

Figure 19: Page-4 Return Analysis-By Country x Category

Total returned products reached 3,049, representing a return rate of 3.05% and an estimated lost revenue of 2,780,662$. The United States recorded the highest number of returned orders, with Home and Kitchen being the most returned category, followed by Books. The United Kingdom ranked second in total returned orders.

Figure 20: Page-4 The most 10 returned products

Yoga Mat was the most returned product, with 81 returns and an estimated lost revenue of 72,690$. It was followed by Noise Cancelling Headphones, which recorded 77 returns and approximately 65,089$ in lost revenue. Overall, the top 10 most returned products each recorded between 60 and 85 returns.

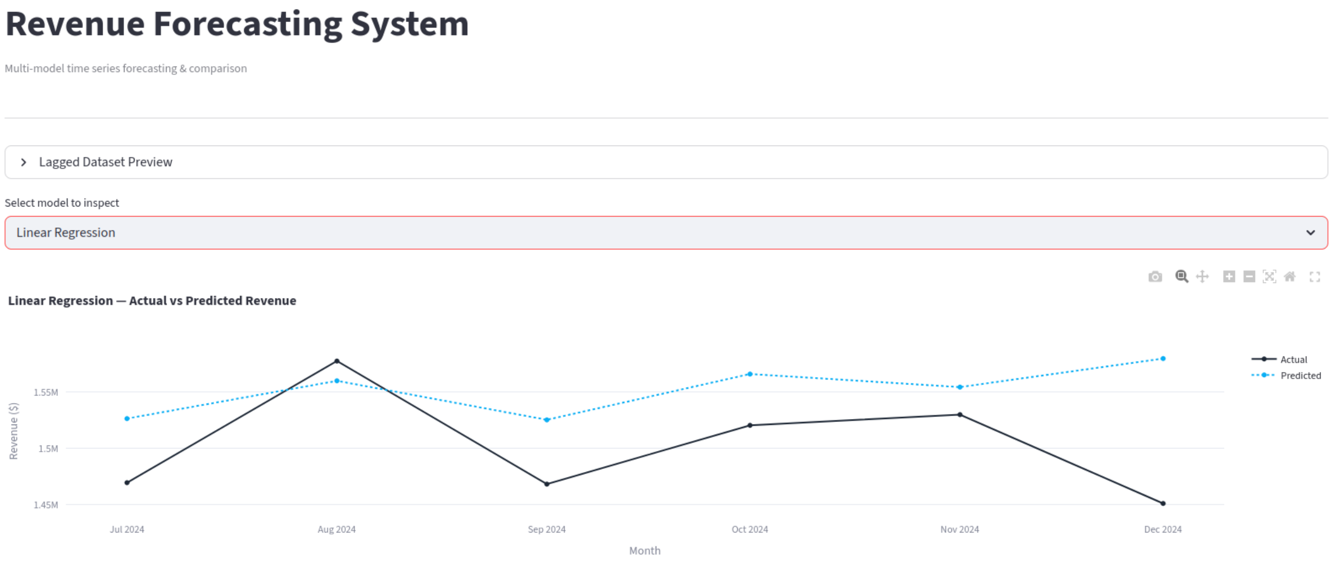

Figure 21: Page-5 Actual vs Predicted Revenue using Linear regression

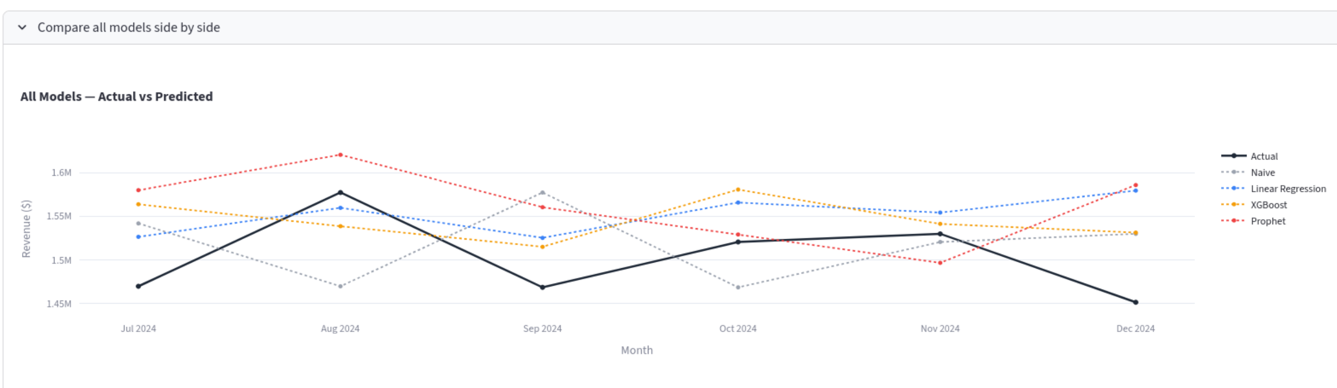

Figure 22: Page-5 Actual vs Predicted Revenue using Multiple Models

Figure 21: The Linear regression model captures the general level and direction reasonably well in the middle months but struggles with the sharp dips in July and December — predicting a smoother curve than reality delivers.

Figure 22: All four models consistently overestimate revenue across the 6-month test window, clustering between 1.53M$-1.63M$ while actual revenue dips as low as 1.45M$ in December — indicating that all models struggle to anticipate downward volatility. XGBoost and Linear Regression track closest to the actual line in the middle months (Aug-Nov), while Prophet is the furthest off.

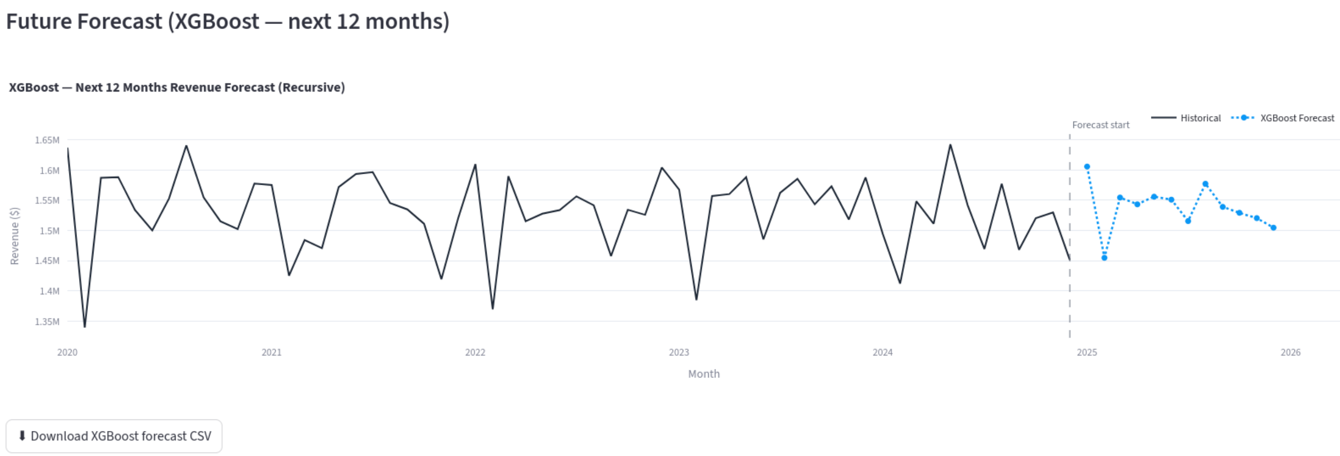

Figure 23: Page-5 Future Forecas (XGboost) Next 12 months

The forecast predicts revenue will remain within the familiar 1.45M$-1.62M$ band, consistent with the flat historical pattern, with an initial dip in february 2025 followed by a gradual stabilization around 1.50M$-1.55M$ toward end of 2025. This confirms the model has learned the business's core characteristic — stable, range-bound revenue with no structural growth trend — making it reliable for operational planning but also signaling that without strategic intervention, the business is forecasted to remain flat through 2026.

5 Business Recommendations

1. **Invest in United states especially Texas and California states:** Rather than spreading resources thin, prioritize deeper penetration — better logistics, localized promotions, and customer retention programs specifically in these two states before expanding elsewhere.

2. **Stop Relying on Discounts:** The data clearly shows discounts have virtually zero impact on revenue (correlation of -0.11). Money spent on discounting is essentially wasted — redirecting that budget toward product quality, faster shipping, or loyalty programs would likely yield better returns.

3. **No Star Product or Category better:** Everything is uniformly equal — categories, products, brands, customers. This means the business has no competitive anchor.

4. **Investigate the February Drop:** Understanding why February underperforms expectations could unlock a significant revenue opportunity.

5. **The 3.05% Return Rate Needs Root Cause Analysis:** While seemingly low, at 100,000 orders that's 3050 returned orders representing potentially 2.7M$ in lost revenue. Identifying which products, categories, or regions drive the most returns could recover a significant portion of that.

6. **Diversify into New Geographies:** Since revenue is capped domestically with TX and CA already saturated, the flat forecast suggests international expansion — particularly into India and the UK which already appear in the top 5 countries — could be the most impactful growth lever available.

7. **India is an Underutilized Market:** India ranks 2nd in customers but generates disproportionately low revenue, suggesting a localized pricing or product strategy could significantly unlock its untapped market potential.

8. **The Early 2025 Dip Needs Preparation:** The forecast shows a notable dip in early 2025 before stabilizing. Businesses should prepare for this by building cash reserves, reducing discretionary spending in Q1 2025, and front-loading inventory clearance before the dip hits.

6 Conclusion

This project successfully demonstrates the end-to-end development of an E-Commerce Analytics and Forecasting Platform, transforming raw transactional data into actionable business intelligence through interactive visualizations, statistical analysis, and machine learning-based forecasting.

The analysis revealed a business characterized by remarkable uniformity across every dimension - revenue, categories, products, brands, and customers all perform within narrow, near-identical ranges. While this consistency reflects operational stability, it also exposes a fundamental strategic gap: the absence of any standout product, loyal customer segment, or growth-driving market that could propel the business beyond its current flat trajectory.

The forecasting component benchmarked four models against a 6-month holdout period, with XGBoost emerging as the most reliable predictor for operational planning. Its 12-month recursive forecast confirms that without deliberate strategic intervention, revenue will continue to hover around 1.5M$/month through 2026 — a plateau rather than a growth curve.

From a technical standpoint, the platform demonstrates the practical value of combining descriptive analytics with predictive modeling in a unified, interactive environment — enabling decision-makers to move seamlessly from high-level KPIs down to individual product transactions, and from historical patterns to future projections.

Fianally, the insights generated throughout this analysis provide a clear and data-driven foundation for strategic decision-making — from geographic expansion and customer retention to product prioritization and forecasting-informed budgeting — positioning the business to move from stability toward sustainable growth.

7 References

Amazon Sales Dataset. Kaggle. Retrieved from `https://www.kaggle.com/datasets/rohiteng/amazon-sales-dataset`

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. The American Statistician, 72(1), 37–45.

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Streamlit. (2024). Streamlit — The fastest way to build data apps. Retrieved from `https://streamlit.io`

Plotly Technologies Inc. (2024). Plotly Python Graphing Library. Retrieved from `https://plotly.com/python`