Pricing Analysis and Optimization in E-Commerce Products with Machine Learning

UCSB PSTAT 135: Big Data Analytics (S26) Final Project

Jayden Udall

June 11, 2026

I. Introduction

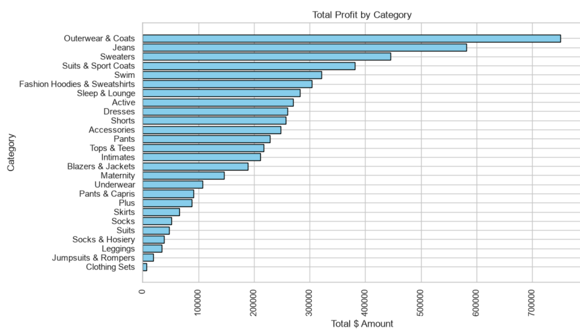
A common business problem is finding the optimal pricing strategies for company products, ranging from hundreds to millions across different platforms. The goal of this project is to apply statistical methods (consistent with common big data analytics practice) to an E-Commerce dataset to determine if this optimization can be achieved through machine learning algorithms. The ultimate goal is to predict profit margins based on features such as category, sale price, product cost, etc. I originally wanted to optimize pricing strategy; however, prices in this dataset are fixed and I am not able to analyze trends over different prices for the same product.

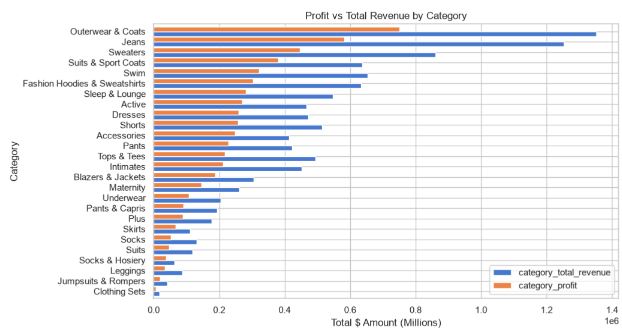
I chose this project because I am heavily interested in applying complex machine learning concepts to business intelligence and strategy. There is a clear goal in this context: to increase revenue through predictions based on past trends and simulations. This project will hopefully allow me to explore supervised and unsupervised learning algorithms and combine them with Monte Carlo optimization techniques. E-commerce datasets, especially large, can be very rich with information waiting to be discovered, and that is exactly the type of challenge I’m looking for with this project. I hope to build skills that would help to maximize profit margins and bring valuable insights to corporations through statistically sound data analysis.

There have been extensive studies on this topic that have focused on several core methodologies and goals, such as demand forecasting with Random Forests or LSTMs (deep neural networks), dynamic pricing using Reinforcement Learning agents, and price elasticity analysis. Many academic papers highlight that there are clear benefits to this approach, and studies have shown upwards of 10-20% profit increase in some companies compared to traditional manual pricing. Through this project, I hope to gain a better understanding of how the algorithms demonstrated in class can achieve this optimization, which learning approaches work, and which do not.

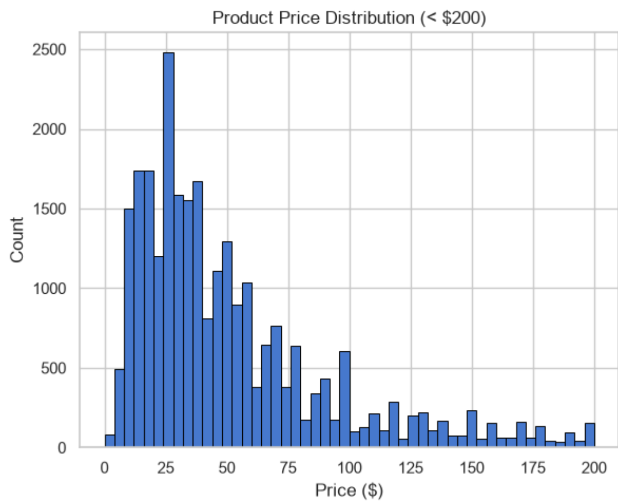
II. Data

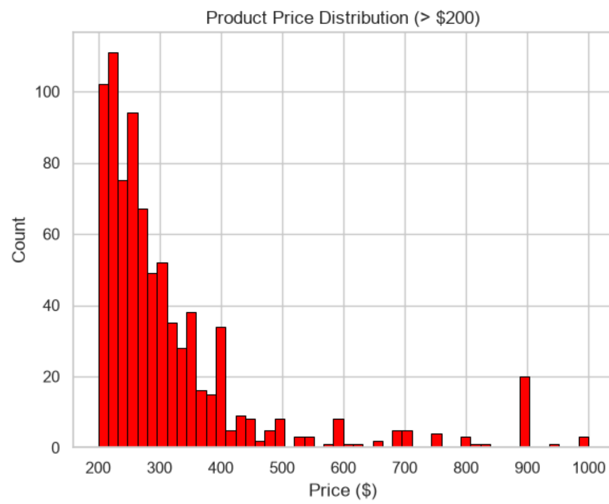
The dataset used in this project is the The Look eCommerce Dataset, a publicly available retail dataset hosted through Google Cloud BigQuery. The dataset was created by Google Cloud as part of its public data program and is designed to simulate the operations of a real online fashion retailer. It contains information about customers, products, orders, order items, inventory, distribution centers, and website events, allowing users to analyze customer behavior, sales performance, and business operations. The dataset consists of hundreds of thousands of records spread across multiple interconnected tables, making it a well-suited introduction for large-scale analytics and cloud-based data processing.



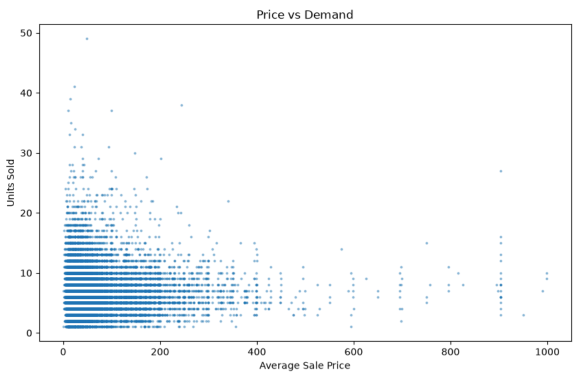


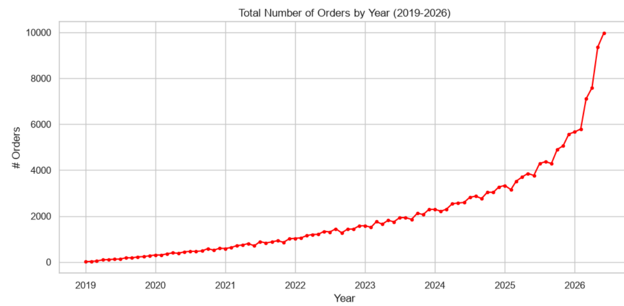
I chose this dataset because it provides a realistic representation of an eCommerce business and offers opportunities to answer meaningful business intelligence questions. Since the dataset includes transaction-level sales data, product information, and customer activity, it allows analysts to explore topics such as revenue generation, pricing strategies, customer purchasing patterns, inventory management, and product performance. In addition, the dataset is hosted directly in BigQuery, making it an excellent resource for practicing SQL, cloud analytics, and big data techniques.





The Look eCommerce Dataset was generated and curated by Google Cloud rather than collected from an actual retailer. The data is synthetically generated using realistic business rules and customer behavior patterns to mimic the structure and complexity of real-world retail data while avoiding privacy concerns. This approach creates a safe and accessible environment for learning data analytics, machine learning, and business intelligence skills.





Some example of product statistics are shown below:

Product name	**Avg cost**	**Avg sale price**	**Units sold**	**Total revenue**	**Profit**
The North Face Apex Bionic Soft Shell Jacket - Men's	425.88	903.0	27	24,381.0	12,882.20
Quiksilver Men's Rockefeller Walkshort	472.27	903.0	16	14,448.0	6,891.70
NIKE WOMEN'S PRO COMPRESSION SPORTS BRA *Outstanding Support and Comfort*	445.48	903.0	15	13,545.0	6,862.80
Canada Goose Women's Mystique	329.00	750.0	15	11,250.0	6,315.00
The North Face Denali Down Womens Jacket 2013	417.88	903.0	13	11,739.0	6,306.55
Darla	404.60	999.0	10	9,990.0	5,944.05
Mens Nike AirJordan Varsity Hoodie Jacket Grey / Black 451582-066	409.06	903.0	12	10,836.0	5,927.29

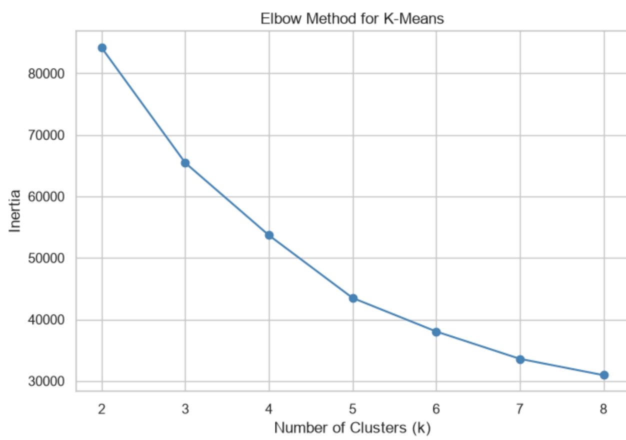
III. Methodology

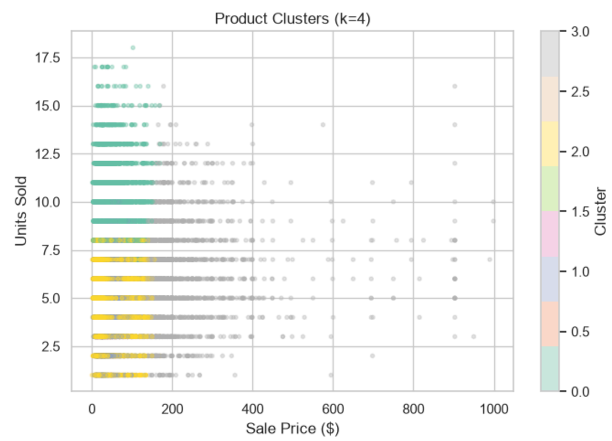
Building on the exploratory analysis, I applied unsupervised and supervised learning to the ecommerce transaction data to support pricing and demand insights. The analysis uses two levels of granularity: a SKU-level product table (29,049 unique product_id records) for segmentation and pricing prioritization, and the full order-line dataset (181,873 transactions) for PCA-based exploration of transaction-level variation. Unsupervised methods include K-Means clustering, which groups products into four segments based on standardized price, cost, margin, and units sold, and Principal Component Analysis (PCA), applied at both the SKU and order-line levels to reduce dimensionality and visualize how products and transactions separate by pricing, geography, and category. Supervised methods focus primarily on forecasting monthly category demand: Gradient Boosting (with 1-, 2-, and 3-month lag features) serves as a benchmark, while linear regression using lag-1 demand, a time trend, and their interaction provides the preferred forecast under a time-based train/test split. Rather than predicting profit from product attributes, which is largely mechanical once price and cost are known, the pricing analysis builds a margin-demand priority matrix that ranks categories for protection, repricing, or cost review, and flags SKUs below their category median margin. Models are evaluated using R^2, RMSE, MAE, and visual diagnostics including cluster profiles, PCA loadings, actual-vs-predicted plots, and the priority matrix. Together, these approaches identify natural product tiers, reveal the main drivers of transaction variation, and connect demand forecasts to actionable pricing priorities.

IV. Results

Part 1: Unsupervised Learning

To identify natural groupings in the product catalog, I applied K-Means clustering at the ID level (product_id) using four standardized features: sale price, product cost, margin percentage, and units sold. Standardization ensures no single variable dominates the distance metric. The goal is to discover distinct product components, such as premium high-profit items versus high-volume mid-range products, that can inform pricing strategies. I used the elbow method to select k clusters and then profile each segment by average price, demand, profit, and dominant product categories.

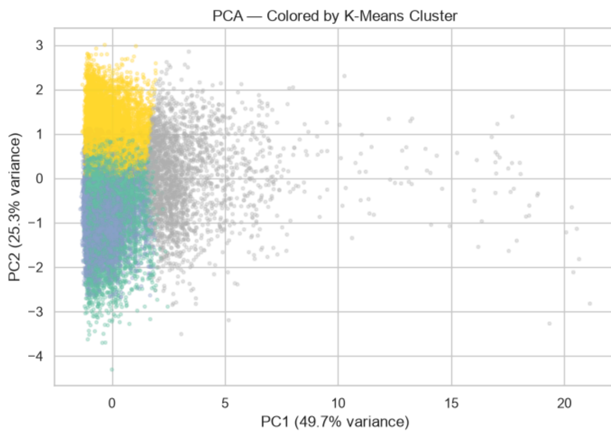


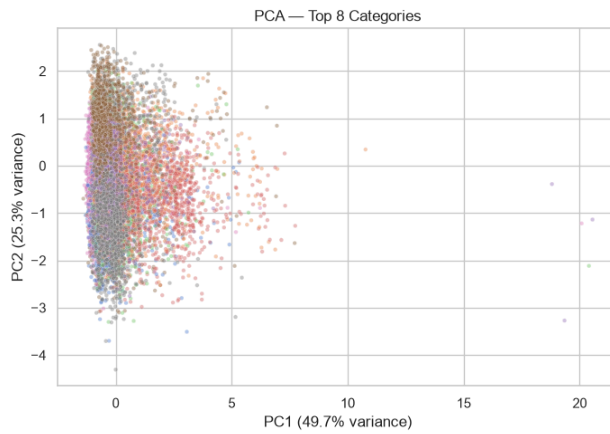


The k vs. inertia graph did not produce a trivial elbow, so I went with k = 4. This produced the graph on the left that separates the data into three distinct groups. The fourth was drowned out by noise. Below are the clusters identified along with some summary statistics:

Cluster	# Products	Avg price	Avg cost	Avg margin	Avg units sold	Avg total profit
0	7,204	47.21	22.70	0.52	9.37	229.27
1	9,447	40.88	22.84	0.44	5.29	94.52
2	9,950	46.11	19.79	0.57	4.93	129.79
3	2,448	218.70	102.66	0.52	6.27	739.90

Cluster 3 is the premium tier (~$219 avg price, ~$740 avg total profit, 2,448 products), while Clusters 0–2 cover mid-range products ($41–$47 avg price) with Cluster 0 showing the highest average demand (~9.4 units sold). Each cluster is dominated by different categories (e.g., Shorts and Sweaters in Cluster 0, Intimates and Tops & Tees in Cluster 1), confirming that products naturally group into distinct pricing and demand profiles useful for tiered pricing strategies. After clustering, I used Principal Component Analysis (PCA) to project the same standardized product features (sale price, product cost, margin percentage, and units sold) into two principal components. This reduces four dimensions to a 2D view while preserving as much variance as possible, making it easier to see how the K-Means clusters separate and whether product categories align with those groupings. I visualized the results with scatter plots colored by cluster and by top product categories to interpret which features drive the main axes of variation in the product catalog.

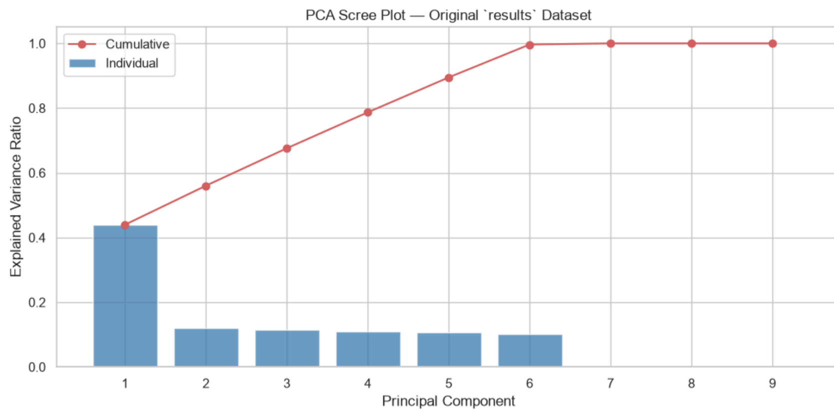


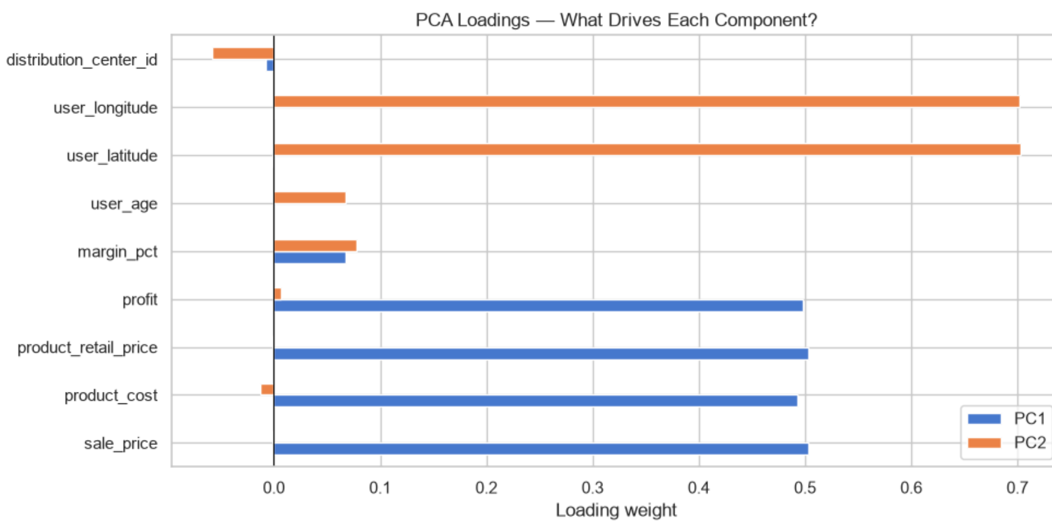


The total variance explained by the top two principal components is 49.7% and 25.3%, respectively, giving us a total percentage of 75%. PC1 primarily reflects the pricing and profitability axis, while PC2 adds a second dimension of variation based on margin and demand. Because this reduces the number of features, I can accept this 25% loss of information. The PCA data colored by cluster show clear separation; however, the categories overlap substantially, indicating that category alone is not as effective in explaining variation in product pricing as the combined predictors used in clustering.

Next, I ran PCA on all 181,873 order-line rows in the raw results table to capture transaction- and customer-level variation. I used numeric features including pricing (sale_price, product_cost), derived profit and margin, customer age and location, and distribution center. Features are standardized before fitting PCA, and I examined a scree plot, component loadings, and 2D scatter plots to identify the main axes of variation across orders and how they relate to product category, department, and customer demographics.

The scree plot shows a large drop of after PC1 followed by a uniform distribution of variance explained by the next five. PC1 explains 44% of variance while PC2 explains 12%. Together, the first two give us 56% of the total variation in the entire dataset.

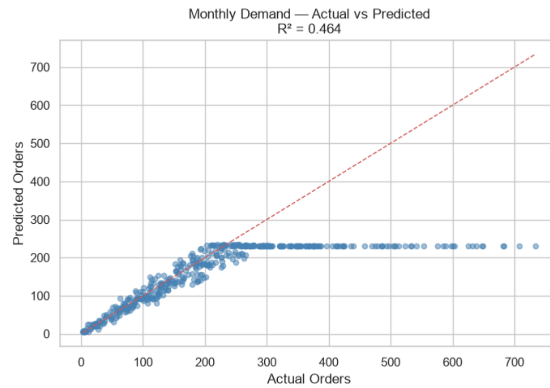




The PCA loadings chart shows how much each feature contributes to PC1 and PC2. On PC1, sale price, product cost, retail price, and profit all have strong positive loadings (~0.50), so PC1 is mainly a pricing/profit axis that identifies cheap vs expensive orders. On PC2, latitude and longitude dominate (~0.70), so PC2 reflects customer location, not product type. Margin, age, and distribution center load weakly on both axes, meaning the first two components are driven almost entirely by price level and geography.

Part 2: Supervised Learning

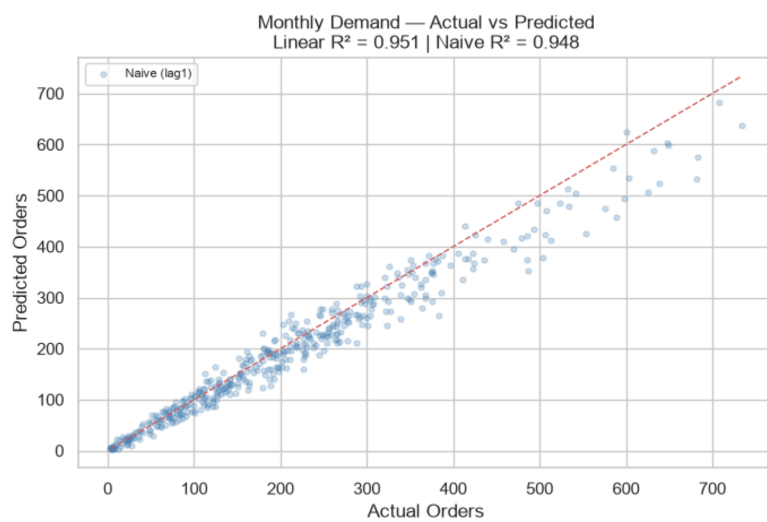
To forecast demand at the category level, I began by fitting a Gradient Boosting Regressor on monthly order counts for each product category. The model uses lag features (orders from the prior 1 and 2 months plus a 3-month rolling average) to predict the next month's volume. This is because monthly demand tends to be persistent. For example, if Jane ordered 600 items last month, she is likely to order around the same number of products this month. Training uses earlier months and testing uses later months on a time-based split so evaluation reflects real forecasting.



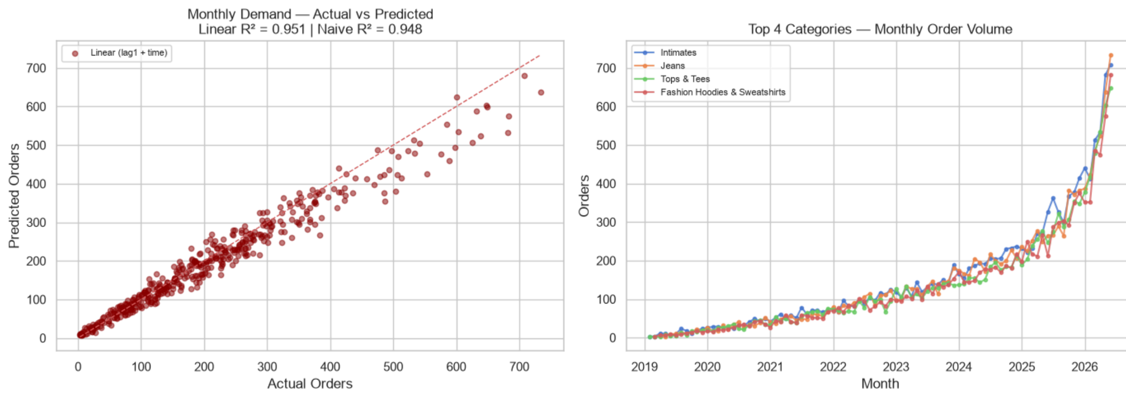
Actual vs predicted for Gradient Boosting Tree model

This model achieved a rather disappointing $R^2 \approx 0.46$ (RMSE ≈ 108 orders). In retrospect, I determined that this was because the Gradient Boosting model, which is a tree mode, does not extrapolate and only combines patterns seen in the training set. The model was unable to account for the surge in demand in 2026, which explains the poor performance.

This was a flawed approach, so I tried again using a linear regression model with features lag1 (previous month orders), t (month), and lag1 * t to account for surge, achieving far superior results.



Lag(1) baseline for comparison



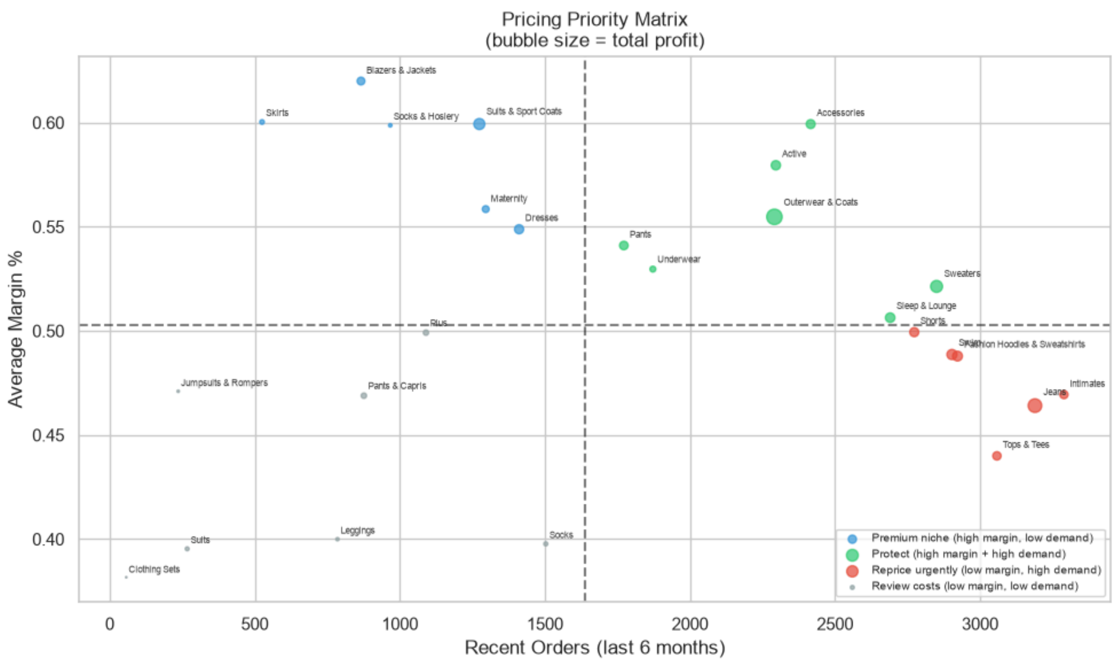
Linear (lag1 + time + lag1*time) and monthly order volumes

This is a surprising discovery. The results show that the previous month alone is already a great predictor of future demand, achieving a 0.94 R^2 value and highlighting the persistent nature of monthly demand. The linear regression model took into account the lag, time, and lag * time to provide a slight weight correction and take demand surge into account. This gave us a slight increase in accuracy and an R^2 value of 0.96, meaning that 96% of the variance was explained by the model. Using this model, we can accurately forecast product demand on a month-to-month basis. This is relevant because it allows us to forecast demand and adjust prices accordingly before the next month starts.

Now, I'll pivot into the pricing optimization strategy. The key idea is that profit and demand alone don't tell you what to do, as each product is unique. A category can have high profit because it sells a lot (high demand), or high profit because the margins are high. My approach is to categorize items into four distinct quadrants: no repricing necessary (high margin, high demand), reprice urgently (low margin, high demand), premium and niche (high margin, low demand), and reevaluate product (low demand, low margin). The four quadrants are built by comparing each of the 26 categories to the medians of both margin and demand. We can plot the categories based on this table:

Product category	**Avg margin pct**	**Avg price**	**Avg cost**	**Total units**	**Total profit**	**# SKUs**	**# Below median margin**	**Profit per unit**
Outerwear & Coats	0.5548	146.1071	65.0210	91,289	7,512,308.870	1,418	708	82.2906
Jeans	0.4641	97.8044	52.3926	12,764	582,191.2834	1,996	996	45.6120
Sweaters	0.5214	75.3770	36.1207	11,291	445,830.1338	1,732	859	39.4854
Suits & Sport Coats	0.5993	126.5556	50.7996	5,012	381,226.0365	739	365	76.0627
Swim	0.4886	57.8175	28.8850	11,449	321,469.3216	1,797	897	28.0784

⋮



This plot lets us visualize which categories are profit drivers, which are premiums, and which need to be reevaluated to increase profit.

Next, I took the optimization strategy even further down to the unique product level based on individual SKUs. The idea is to compare each product to its category medians, and only keep SKUs with greater than 5 units sold (enough volume to actually matter). I then rank by total profit and repricing impact, calculated as (margin_gap * price * units sold), giving us this table:

Product name	Product category	Sale price	Product cost	Margin pct	Category median margin	Units sold	Total profit	Repricing impact
The North Face Apex Bionic Soft Shell Jacket - Men's	Active	903.0	391.902	0.566	0.579	14	7155.372	164.346
Quiksilver Men's Rockefeller Walkshort	Shorts	903.0	472.269	0.477	0.499	16	6891.696	317.856
Mens Nike AirJordan Varsity Hoodie Jacket Grey / Black 451582-066	Outerwear & Coats	903.0	409.059	0.547	0.555	12	5927.292	86.688

⋮

With this information, we can analyze the classified repricing strategy of every individual product and determine whether repricing is necessary to maximize total profit or the product must be reevaluated as a whole.

V. Conclusion

This project analyzed 181,873 order lines across 29,049 SKUs and 26 product categories to answer a practical pricing question: where should an e-commerce retailer focus to protect margin, grow revenue, and prioritize repricing? Exploratory analysis showed that profitability is highly concentrated. A small set of premium outerwear and denim products drives outsized profit at the SKU level, while category-level totals reveal the same pattern at scale: Outerwear & Coats ($751,231), Jeans ($582,191), and Sweaters ($445,830) are the three largest profit contributors. At the same time, average SKU margin sits near 51%, but margins vary meaningfully within categories (average within-category standard deviation of 3.5%), which means category averages alone are not enough for pricing decisions.

Unsupervised learning helped structure that heterogeneity. K-Means clustering on price, cost, margin, and units produced four product segments, separating mid-market volume SKUs from a premium cluster (Cluster 3: ~$219 average price, ~$740 average total profit per SKU). PCA on pricing-related features captured roughly 75% of variance in the first two components (PC1 ~49.7%, PC2 ~25.3%), confirming that price, cost, and margin make up most of the product landscape. Together, these results show that the catalog is not one homogeneous market; it contains distinct economic segments that should be managed differently.

Supervised demand modeling connected those segments to forward-looking planning. An initial gradient boosting model for monthly category demand achieved only $R^2 = 0.464$ and systematically under-predicted 2025–2026 growth, effectively hitting a ceiling near prior training maxima rather than extrapolating rising demand. Replacing it with a simpler linear model using lag-1 demand plus time dramatically improved performance to $R^2 = 0.960$ and MAE = 20.4 orders per category-month, only modestly better than a naive lag-1 baseline ($R^2 = 0.948$, MAE = 23.0). The takeaway: for this dataset's accelerating demand period, trend-aware regression outperformed tree models that cannot extrapolate beyond historical ranges.

The pricing optimization section reframed the problem from "predict profit" to "prioritize action." Attempts to predict SKU profit from category alone explain little variation, while models using price, cost, and units are largely mechanical because profit is defined by those inputs. The actionable output is a priority matrix that joins recent six-month category demand with average margin. Categories above both medians, such as Outerwear & Coats, Sweaters, Sleep & Lounge, Active, and Accessories, should be protected from unnecessary discounting. High-margin but lower-demand categories like Suits & Sport Coats and Dresses are premium niches. The highest-leverage intervention targets are "Reprice urgently" categories: high demand but below-median margin, including Jeans (3,189 recent orders, 46.4% margin), Swim, Fashion Hoodies & Sweatshirts, Shorts, Tops & Tees, and Intimates. At the SKU level, 10,821 products sold at least five units while sitting below their category median margin, including many high-dollar items where even small margin gaps represent large dollar impact.

The trivial fact that demand is characterized by the market and cannot be predicted by price, custom margin, and category alone reinforces the main conclusion: the strongest decision framework in this project is not a single predictive score, but an integrated workflow: EDA to identify profit concentration, clustering/PCA to segment the catalog, demand forecasting to anticipate category volume, and a margin/demand priority matrix to direct repricing, cost review, and margin protection.

Overall, the project supports a clear business strategy. Protect profitable, high-demand categories; selectively invest in premium niches; urgently reprice high-volume, low-margin categories starting with Jeans and Intimates; and drill to SKU-level repricing for the 10,821 under-margin products with meaningful sales. The main empirical result is that combining demand forecasts ($R^2 \approx 0.96$) with margin economics yields more useful pricing guidance than trying to predict profit directly from product metadata. For a retailer operating on this data, the highest ROI next step is targeted price and cost review in the “Reprice urgently” quadrant, while preserving pricing power in top profit engines like Outerwear & Coats and Sweaters.

VI. References

Matplotlib: https://matplotlib.org/
Scikit-learn: https://scikit-learn.org/
Google Cloud Console Documentation: https://docs.cloud.google.com/docs

Pricing Strategy Optimization by Machine Learning in E-commerce

- Quan Liu and Yunkui Song. 2025. Pricing Strategy Optimization by Machine Learning in E-commerce. In Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence (DEAI '25). Association for Computing Machinery, New York, NY, USA, 760–765. https://doi.org/10.1145/3745238.3745358

Optimizing E-commerce Dynamic Pricing Using Aggregated Market Data and Cloud-Based Analytics

- Gupta, Medha & Jain, Jitender. (2024). Optimizing E-commerce Dynamic Pricing Using Aggregated Market Data and Cloud-Based Analytics. International Jmynal of Global Innovations and Solutions (IJGIS). 10.21428/e90189c8.ed197a70.
- https://www.researchgate.net/publication/387505222_Optimizing_E-commerce_Dynamic_Pricing_Using_Aggregated_Market_Data_and_Cloud-Based_Analytics