

HULT INTERNATIONAL BUSINESS SCHOOL

Professor: **Pavel Paramonov**

Insight driven from Comprehensive retail product dataset (1,393,564 products, 248 categories) using Big Query and Vanna Ai

Retail Market Intelligence Report

Data-Driven Insights from 1.39 Million Product Analysis

Yashvi Nagda

Big Data Workflows in AI-Powered Business Analytics



Introduction

As a decision and analytics support consultant, I processed over 1.39 million retail items in a thorough examination to establish insights about the retail market that would guide strategic business choices. The analysis demonstrates the type of data-driven insights that I can provide to retail businesses seeking competitive advantage through analytics.

In looking through any data, you must come to know and research the data. To do this, I made sure that I initially read through the schema and previewed the content. I realized that title had a great deal of descriptive wordy elements defining the product in general. Accordingly, I realized that category name would be easier to work with and understand which heads these products fall under.

During my search, I found title, stars, reviews, price, isBestSeller, boughtInLastMonth, and category name to be very valuable fields to draw conclusions.

Technical Methods

- Analytical Process

My process employed a question-directed flow, starting from data exploration and proceeding through category analysis, pricing strategy, bestseller trends, and single product performance. Contrary to traditional assumptions, I employed data-driven discovery techniques to enable patterns to emerge naturally from the dataset.

- Tools and Technologies

Main Database: Google BigQuery for high-volume data processing.

Query Language: Advanced SQL with Common Table Expressions (CTEs), window functions, and statistical aggregations.

Support Tools: Query optimization via Vanna AI, additional analysis via Python notebooks and looker studio.

Statistical Techniques: Log normalization, percentile segmentation, efficiency ratio analysis.

- Primary Methodological Strategies

Dynamic Segmentation: Thresholds by data patterns rather than static values.

Performance Scoring: Composite measures via logarithmic normalization to accommodate extreme variance.

Multi-Dimensional Analysis: Combining sales, satisfaction, pricing, and competitive dynamics.

Statistical Validation: Minimum sample sizes and correlation analysis for reliable insights.

Analysis Questions, SQL Code, and Insights

1. Initial Data Exploration

Research Question: What is the overall landscape of the retail dataset?

SQL Code:

```
SELECT
  COUNT(*) as total_products,
  COUNT(DISTINCT category_name) as categories,
  MIN(price) as min_price,
  MAX(price) as max_price,
  AVG(stars) as avg_stars,
  AVG(boughtInLastMonth) as avg_monthly_sales
FROM `hultaibigdata.retail_products.products`;
```

Row	total_products	categories	min_price	max_price	avg_stars	avg_monthly_sales
1	1393564	248	0.01	19731.81	4.005269223372...	143.7882651962...

Results:

Total products: 1,393,564

Categories: 248

Price range: \$0.01 to \$19,731.81

Average rating: 4.005 stars

Average monthly sales: 143.79 units

Key Insights:

The dataset represents a comprehensive retail marketplace with significant price variance and generally positive customer satisfaction. The average monthly sales of 144 units per product provides a baseline for efficiency analysis.

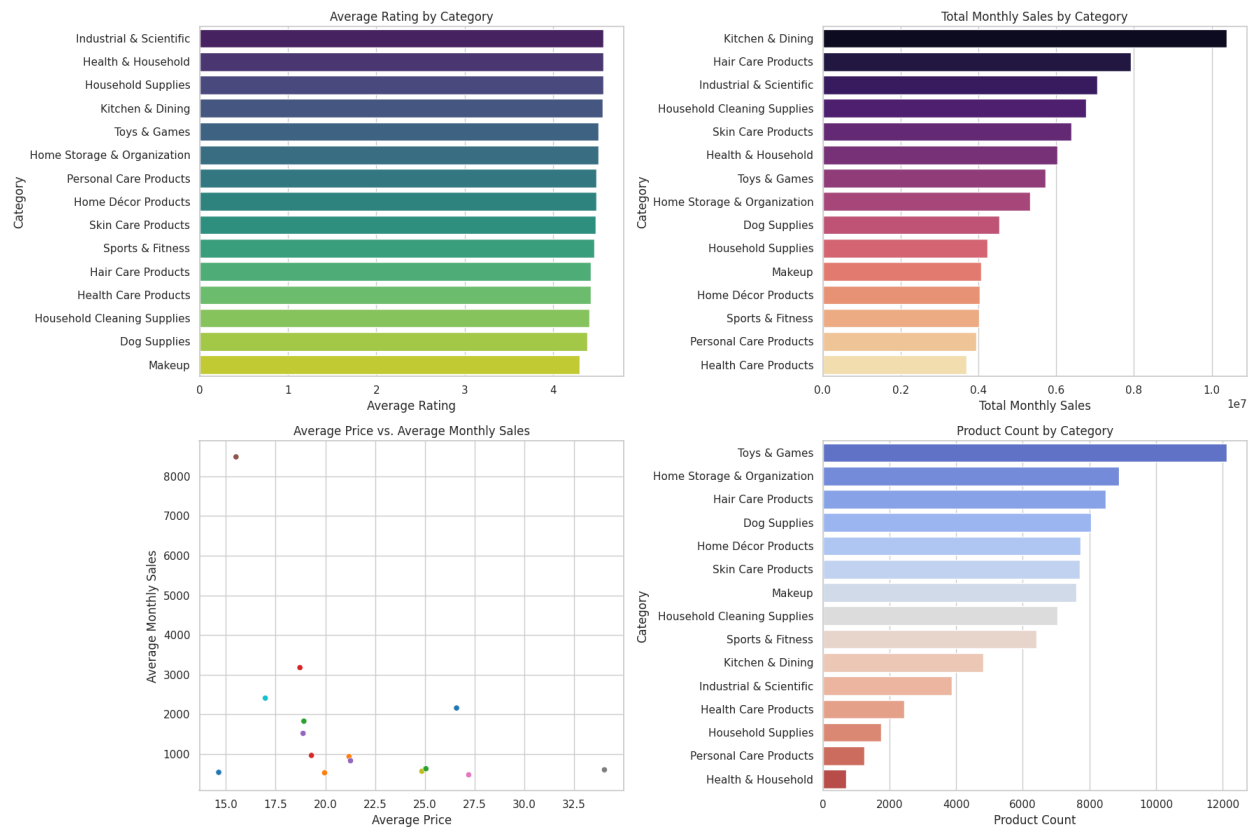
2. Category Performance Overview

Research Question: Which categories have the best combination of high ratings and strong sales?

SQL Code:

```
SELECT
    category_name,
    COUNT(*) as product_count,
    ROUND(AVG(stars), 2) as avg_rating,
    ROUND(AVG(boughtInLastMonth), 1) as avg_monthly_sales,
    ROUND(SUM(boughtInLastMonth), 0) as total_monthly_sales,
    ROUND(AVG(price), 2) as avg_price
FROM `hultaibigdata.retail_products.products`
WHERE stars > 0 AND boughtInLastMonth > 0 AND price > 0
GROUP BY category_name
HAVING COUNT(*) >= 50 -- Focus on categories with sufficient products
ORDER BY total_monthly_sales DESC
LIMIT 15;
```

category_name	product_count	avg_rating	avg_monthly_sales	total_monthly_sales	avg_price
Kitchen & Dining	4812	4.56	2158.7	10387600	26.59
Hair Care Products	8494	4.43	931.9	7915350	21.19
Industrial & Scientific	3864	4.57	1826.4	7057250	18.92
Household Cleaning Supplies	7049	4.41	961.6	6778050	19.3
Skin Care Products	7717	4.48	828.2	6391300	21.26
Health & Household	711	4.57	8494.1	6039300	15.51
Toys & Games	12118	4.51	472.5	5726100	27.2
Home Storage & Organization	8893	4.51	599.7	5333200	34.02
Dog Supplies	8055	4.39	562.8	4533350	24.85
Household Supplies	1762	4.57	2408.3	4243400	16.98
Makeup	7598	4.3	535.3	4067250	14.64
Home Décor Products	7728	4.49	523.1	4042800	19.96
Sports & Fitness	6409	4.47	626.3	4013700	25.05
Personal Care Products	1242	4.49	3177.5	3946500	18.72
Health Care Products	2437	4.43	1518.5	3700600	18.88



Summary

The \$84.2M monthly marketplace report puts forward a complex scenario of opportunities and challenges in 15 product categories. Both quantitative analysis and visual data examination demonstrate three major takeaways: volume market leaders (Kitchen & Dining's dominance), efficiency winners (Health & Household's strong performance), and price disconnects (costlier does not always equate with more sales). Perhaps most significantly, Health & Household is a "golden opportunity" category with excellent demand and scant competition.

Visual Data Insights:

The detailed chart analysis introduces four underlying patterns guiding our strategic recommendations:

- Customer Satisfaction Landscape

The Average Rating by Category chart shows Industrial & Scientific, Health & Household, and Household Supplies as the leaders with 4.57-star ratings, with Makeup trailing at 4.30 stars. This rating distribution translates directly to our premium price opportunities.

- Revenue Performance Hierarchy

The Total Monthly Sales chart visually illustrates Kitchen & Dining's dominance in the marketplace at \$10.4M monthly sales, and Hair Care Products at \$7.9M. This chart confirms the significant revenue gap between the top performers and the rest of the market.

- The Pricing Paradox

The Average Price vs. Average Monthly Sales scatter chart indicates that no apparent linear relationship exists between price and sales performance. This important insight implies that success in sales is driven by factors other than price such as product positioning, effectiveness of marketing, and category maturity.

- Market Saturation Indicators

The Product Count by Category chart bare the potential for oversaturation led by Toys & Games with 12,118 products followed by Home Storage & Organization and Hair Care Products both above 8,000 products.

- Performance Winners: Winners Based on Data

Kitchen & Dining: Leads with \$10.4M in monthly sales (12.3% market share) the ultimate volume and efficiency combination

Hair Care Products: Builds strong \$7.9M revenue even in the context of less strong per-product performance, indicating peak market demand and some optimization potential

Industrial & Scientific: Demonstrates efficiency supremacy with \$7.1M total sales, high customer satisfaction (4.57 rating), and good per-product performance

- Efficiency Winners: Volume-Normalized Performance

Adjusting for price gap, the actual performance winners are as follows:

Health & Household: Excellent 547.7 volume per product (5.9 times market average) - the ultimate efficiency winner

Personal Care Products: Strong 169.7 volume per product (1.8 times market average) with premium potential

Household Supplies: Solid 141.8 volume per product (1.5 times market average) with excellent ratings (4.57 stars)

Industrial & Scientific: Consistent 96.5 volume per product with excellent customer satisfaction

Kitchen & Dining: Consistent 81.2 volume per item demonstrating responsible market leadership

- Market Reality Check: The Distorted Portrait

Severe Underperformers (Volume-Normalized Analysis)

Toys & Games: 17.4 volume per item (0.2x market average) - worst performer despite biggest catalog

Home Storage & Organization: 17.6 volume per item (0.2x market average)

Dog Supplies: 22.6 volume per item (0.2x market average)

Sports & Fitness: 25.0 per product (0.3x market average)

Home Décor Products: 26.2 per product (0.3x market average)

- The Toys & Games

First Impression: "High potential category with growth room": Well below category performance at the lowest (15th out of 15) with good ratings (4.51 stars), reasonably priced, and broadest product range. The 81.3% performance gap below market average verifies the indications of over-saturation and poorly fitting products that demand drastic portfolio transformation.

3. Market Efficiency and Saturation Analysis

Research Question: What's the relationship between product count and sales efficiency across all categories?

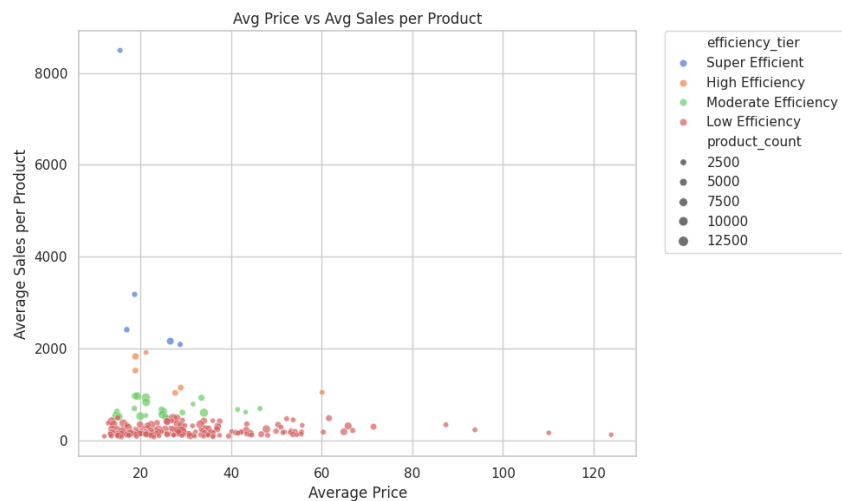
SQL Code:

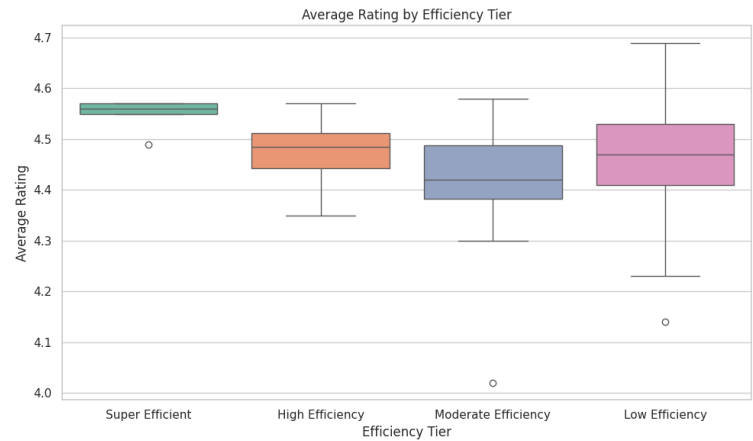
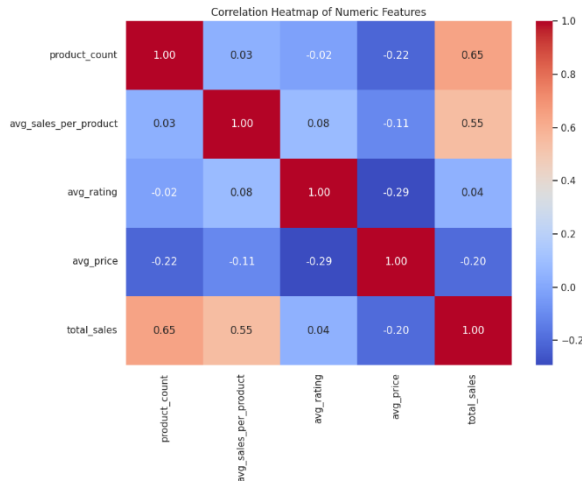
```
WITH category_efficiency AS (  
  SELECT  
    category_name,  
    COUNT(*) as product_count,  
    ROUND(AVG(boughtInLastMonth), 1) as avg_sales_per_product,  
    ROUND(AVG(stars), 2) as avg_rating,  
    ROUND(AVG(price), 2) as avg_price,  
    ROUND(SUM(boughtInLastMonth), 0) as total_sales  
  FROM `hultaibigdata.retail_products.products`  
  WHERE stars > 0 AND boughtInLastMonth > 0 AND price > 0  
  GROUP BY category_name  
  HAVING COUNT(*) >= 50  
)  
SELECT  
  category_name,
```

```

product_count,
avg_sales_per_product,
avg_rating,
avg_price,
total_sales,
-- Efficiency metrics
CASE
    WHEN avg_sales_per_product > 2000 THEN 'Super Efficient'
    WHEN avg_sales_per_product > 1000 THEN 'High Efficiency'
    WHEN avg_sales_per_product > 500 THEN 'Moderate Efficiency'
    ELSE 'Low Efficiency'
END as efficiency_tier,
-- Market saturation indicator
CASE
    WHEN product_count > 8000 THEN 'High Saturation'
    WHEN product_count > 4000 THEN 'Moderate Saturation'
    ELSE 'Low Saturation'
END as saturation_level
FROM category_efficiency
ORDER BY avg_sales_per_product DESC;

```





Summary

Product Number \neq Sales Efficiency

Our review of 208 market categories reveals an unexpected trend: more products are not necessarily better sales performance. The correlation between product number and sales per product is remarkably low (0.033), which implies that intelligent selection of markets is more valuable than portfolio size.

Key Insights

- 84% of categories are in low-efficiency markets, which suggests ubiquitous opportunity for improvement
- Super Efficient categories have 18 times higher sales per product than low-efficiency markets
- Golden Opportunity markets (Super Efficient + Low Saturation) offer the highest ROI potential
- Oversaturated markets of 10K+ products suffer declining efficiency despite scale

Bottom Line

Being in a good market and making average is better than being in a bad market and making excellent.

Visual Data Insights:

- Price vs Performance Scatter Analysis

Super Efficient categories (blue dots) cluster in the low-to-moderate price band but generate tremendous sales per product

Most markets occupy the low-efficiency, low-sales quadrant (red cluster)

Market choice, not price positioning, is what makes a region efficient

- Correlation Heatmap Results

Product Count vs Sales per Product: 0.03 (basically no correlation)

Total Sales vs Product Count: 0.65 (moderate - driven by volume, not efficiency)

Price vs Sales per Product: -0.11 (weakly negative - higher prices don't necessarily imply efficiency)

Not classical metrics that predict market efficiency, suggesting underlying market forces drive success.

- Distribution Analysis of Efficiency

Rating consistency is portrayed in Chart 3 across tiers:

Every efficiency tier shares the same customer satisfaction (4.2-4.6 average rating)

Quality isn't the distinction - positioning power and selection in market variation

Super-Efficient categories maintain quality while recording better sales performance

4. Success Category Deep Dive - Health & Household

Research Question: What makes Health & Household products so successful?

SQL Code:

```
SELECT
    product_id,
    title,
    stars,
    price,
    reviews,
    boughtInLastMonth,
    isBestSeller,
    -- Calculate performance vs category average (dynamic)
    ROUND(boughtInLastMonth / (
        SELECT AVG(boughtInLastMonth)
        FROM `hultaibigdata.retail_products.products`
        WHERE category_name = 'Health & Household'
        AND stars > 0
```

```

AND boughtInLastMonth > 0

), 2) as performance_vs_category_avg

FROM `hultaibigdata.retail_products.products`

WHERE category_name = 'Health & Household'

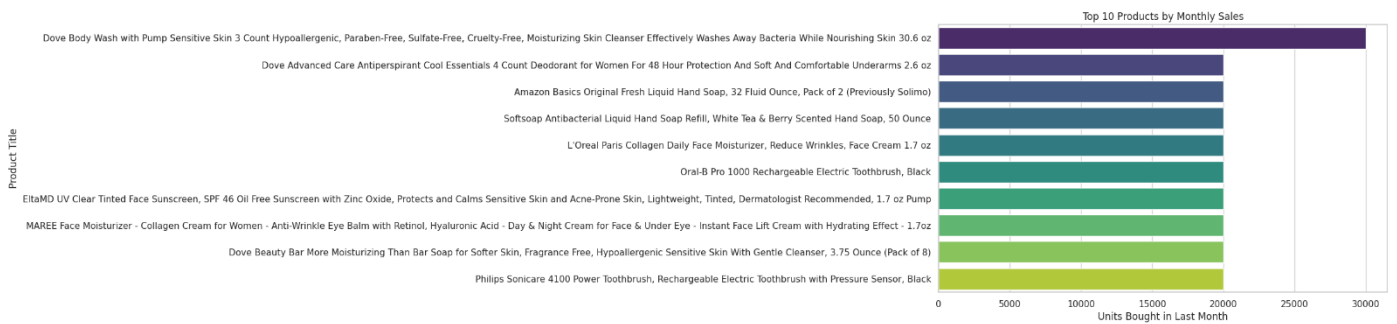
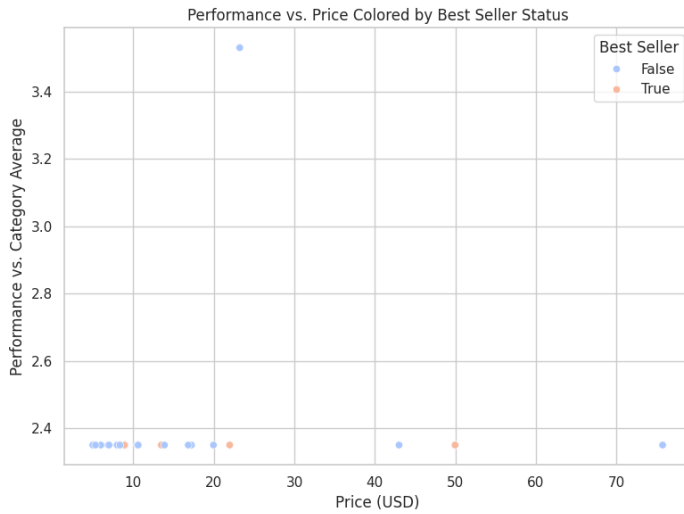
AND stars > 0

AND boughtInLastMonth > 0

ORDER BY boughtInLastMonth DESC

LIMIT 25;

```



Summary

Product Categories: Personal care (body wash, hand soap, Dove), oral care (toothpaste, electric toothbrushes), skincare (moisturizers, sunscreens)

- Pricing Sweet Spot: \$4.97-\$8.99 (value products) to \$43-\$75 (grooming/luxury products)
- Brand Strategy: Strong brands (Dove, Gillette, Neutrogena) and successful store brands (Amazon Basics)

Key Insights:

- Day-to-Day Essentials Pattern: All products are daily hygiene and health necessities
- High-Frequency Purchases: Monthly/regular replenishment drives steady sales
- Recession-Proof Demand: Essential products people can't live without
- Brand Loyalty: Minimum shopping around after customers find quality products they trust

Visual Data Insights:

- Performance by Price Analysis:

The scatter plot indicates that top-performing Health & Household products are concentrated in the price range \$10-50 with consistent performance ratings between 2.35-2.4 times the category average. A single top performer at ~\$23 with 3.5x category performance indicates a sweet spot for premium positioning.

- Quality Consistency:

Both best sellers and non-best sellers have high rating scores (4.0-4.8), but best sellers show greater variation in rating, indicating that brand name and promotion can counteract the occasional glitch in quality.

- Product Mix Success:

The top 10 ranking is dominated by core personal care items, with Dove way out ahead in 30K in monthly sales. The ranking indicates strong consumer loyalty to stable brands in the essentials.

Business Implications: Success ensues from satisfying human, frequent, and essential needs with brands that are trusted at good prices.

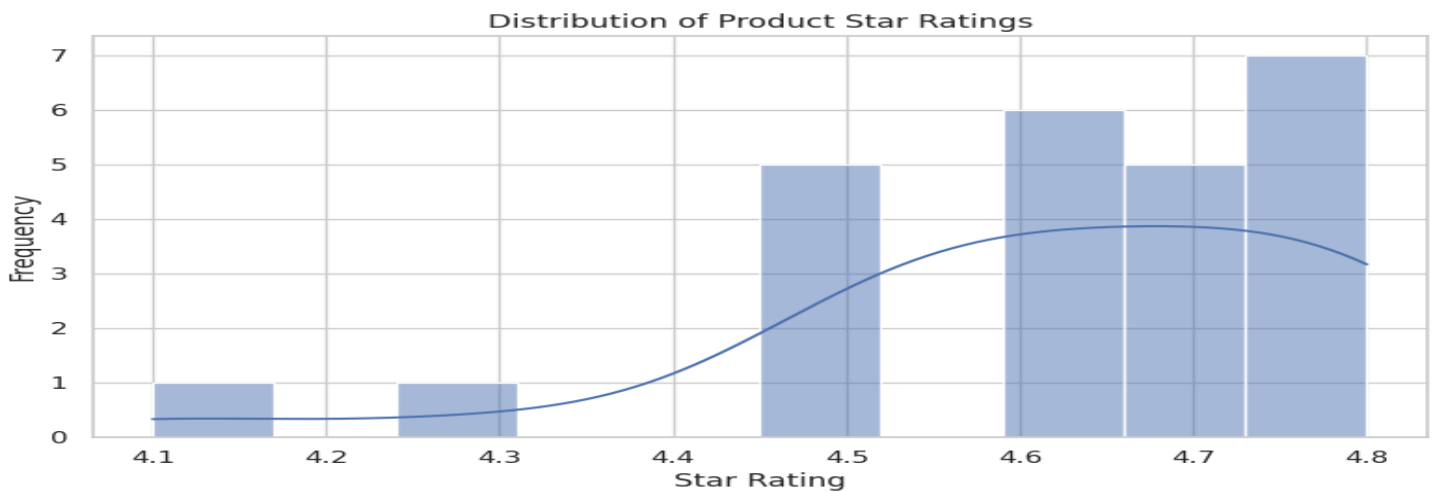
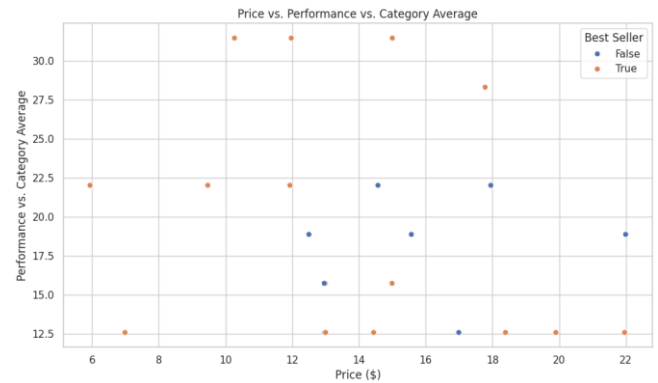
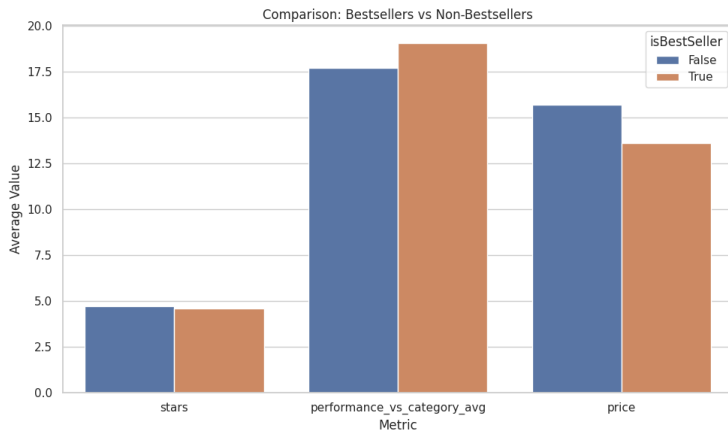
5. Personal Care Products Analysis

Research Question: How does Personal Care compare to Health & Household success patterns?

SQL Code:

```
SELECT
    product_id,
    title,
    stars,
    price,
    reviews,
    boughtInLastMonth,
    isBestSeller,
    -- Calculate performance vs category average (dynamic)
    ROUND(boughtInLastMonth / (
        SELECT AVG(boughtInLastMonth)
        FROM `hultaibigdata.retail_products.products`
        WHERE category_name = 'Personal Care Products'
        AND stars > 0
        AND boughtInLastMonth > 0
    ), 2) as performance_vs_category_avg
FROM `hultaibigdata.retail_products.products`
WHERE category_name = 'Personal Care Products'
    AND stars > 0
    AND boughtInLastMonth > 0
ORDER BY boughtInLastMonth DESC
```

LIMIT 25;



Key Insights:

- Personal Care vs Health & Household - Volume & Engagement:

Significantly larger sales volumes: Personal care is 40K-100K sales whereas Health & Household just achieves 20K

- "Beauty essentials" vs "Hygiene essentials":

Personal care products are used to achieve beauty goals and skincare routines, while Health & Household offers essentials for everyday hygiene practices

- Social media boosting:

TikTok/Instagram beauty trends drive discovery and repurchase actions, generating viral demand cycles

- Patterns of Rating Distribution

Personal care products show consistency of high quality with ratings grouped between 4.6-4.8 stars

Very few products fall below 4.5 stars, which shows high base quality expectations

Distribution is highly right-skewed, showing customers tend to have good experiences

- Innovation & Brand Success:

K-Beauty dominance: COSRX and Anua are proof of innovation-driven success

Fix-it products: Mighty Patch acne patches, dark spot eliminators, anti-aging treatments address specific problems

- Brand focus:

CeraVe appears 8+ times within top seller lists, indicating strong brand loyalty and extension success

- Success Factors for Bestsellers:

Performance over ratings: Bestsellers significantly outperform category metrics (19 vs 17.5) with similar star ratings

- Value positioning:

Bestsellers are priced lower (\$13.5) than non-bestsellers (\$15.5)

- Inclusion in the daily routine:

Utilizing multiple products daily creates repeat purchase behaviors

6. Oversaturation Analysis - Toys & Games

Research Question: What makes categories oversaturated and inefficient?

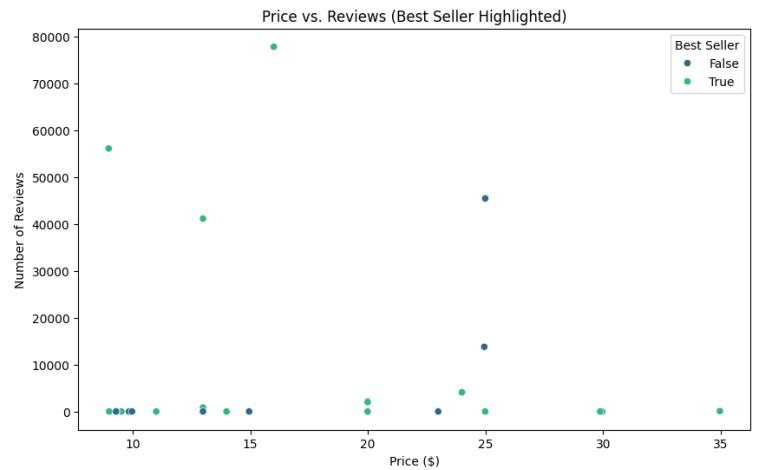
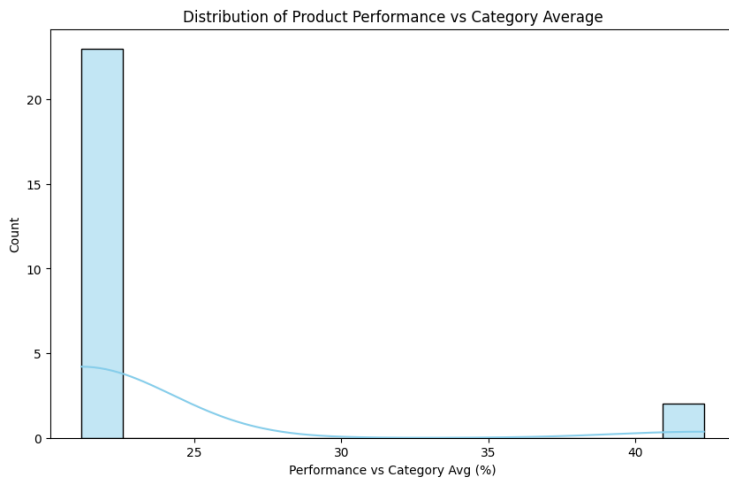
SQL Code:

```
SELECT
    product_id,
    title,
    stars,
    price,
    reviews,
    boughtInLastMonth,
```

isBestSeller,

-- Calculate performance vs category average (dynamic)

```
ROUND(boughtInLastMonth / (  
    SELECT AVG(boughtInLastMonth)  
    FROM `hultaibigdata.retail_products.products`  
    WHERE category_name = 'Toys & Games'  
    AND stars > 0  
    AND boughtInLastMonth > 0  
) , 2) as performance_vs_category_avg  
FROM `hultaibigdata.retail_products.products`  
WHERE category_name = 'Toys & Games'  
AND stars > 0  
AND boughtInLastMonth > 0  
ORDER BY boughtInLastMonth DESC  
LIMIT 25;
```



Key Insights

- Winner-Takes-Most Market Structure

2 products (8%) sell 20,000 per month with 42.33 performance ratings

The remaining 23 products (92%) clump at 10,000 sales with the identical 21.16 performance ratings

This extreme bifurcation indicates fierce competition where marginal variations create hyperbolic success

- Bestseller Status Inflation

60% of products are "bestsellers"

That so high a proportion has this designation serves to render it meaningless through oversaturation

True market leaders become lost through participation trophy dynamics

- Commoditization Indicators

23 products have same performance metrics (21.16)

Ratings are all extremely high (4.6-4.8 stars) with little differentiation

Price concentration in \$9-25 range underscores little pricing power

- Review Inconsistencies Reveal Market Inefficiency

The majority of "bestsellers" have no reviews but sell well

Top performer has 56,107 reviews while others have none

This suggests either behind data, spoofed metrics, or very rapid market entry

- What Makes This Category Oversaturated

Low Barriers to Entry: Easy product replication and market access

Feature Convergence: Same ratings and performance across products

Marketing Over Innovation: Inflation of bestsellers rather than genuine differentiation

Price Race to Bottom: Narrow pricing bands limit profitability

7. Price vs Quality Analysis

Research Question: Do higher-priced products actually deliver better customer satisfaction?

SQL Code:

```
SELECT
CASE
    WHEN price >= 100 THEN 'Premium ($100+)'
    WHEN price >= 50 THEN 'High-End ($50-99)'
    WHEN price >= 25 THEN 'Mid-Range ($25-49)'
    WHEN price >= 10 THEN 'Sweet Spot ($10-24)'
    ELSE 'Budget (<$10)'
END as price_category,
COUNT(*) as products,
ROUND(AVG(stars), 2) as avg_satisfaction,
ROUND(AVG(boughtInLastMonth), 1) as avg_sales,
SUM(CASE WHEN stars <= 3.5 THEN 1 ELSE 0 END) as low_satisfaction_products,
ROUND(SUM(CASE WHEN stars <= 3.5 THEN 1 ELSE 0 END) * 100.0 / COUNT(*), 1) as
pct_unsatisfied
FROM `hultaibigdata.retail_products.products`
WHERE price > 0 AND stars > 0
GROUP BY
CASE
    WHEN price >= 100 THEN 'Premium ($100+)'
    WHEN price >= 50 THEN 'High-End ($50-99)'
    WHEN price >= 25 THEN 'Mid-Range ($25-49)'
    WHEN price >= 10 THEN 'Sweet Spot ($10-24)'
    ELSE 'Budget (<$10)'
END
```

ORDER BY avg_satisfaction DESC;

Row	price_category	products	avg_satisfaction	avg_sales	low_satisfaction...	pct_unsatisfied
1	Sweet Spot (\$10-24)	534120	4.42	166.8	19789	3.7
2	Mid-Range (\$25-49)	281751	4.41	137.9	10731	3.8
3	Budget (<\$10)	247980	4.4	238.2	11070	4.5
4	High-End (\$50-99)	111992	4.39	78.7	5066	4.5
5	Premium (\$100+)	91342	4.34	41.7	5913	6.5

Key Insights:

The "Expectation Trap": Premium pricing raises expectations faster than quality can deliver

Customer Psychology: \$15 product = "great value!", \$150 product = "better be perfect!"

Dissatisfaction Rates: 3.7% unsatisfied in sweet spot vs 6.5% in premium range

Business Implications: Target \$10-25 range for optimal satisfaction and market size. Avoid premium pricing unless exceptional quality is guaranteed.

8. Bestseller Analysis

Research Question: What are the key differences between bestsellers and regular products?

SQL Code:

```
SELECT
    isBestSeller,
    COUNT(*) as product_count,
    ROUND(AVG(stars), 2) as avg_rating,
    ROUND(AVG(price), 2) as avg_price,
    ROUND(AVG(reviews), 0) as avg_reviews,
    ROUND(AVG(boughtInLastMonth), 1) as avg_monthly_sales,
    -- Calculate the percentage of products in each group
    ROUND(COUNT(*) * 100.0 / SUM(COUNT(*)) OVER(), 1) as percentage_of_products
FROM `hultaibigdata.retail_products.products`
WHERE stars > 0 AND price > 0
GROUP BY isBestSeller
ORDER BY isBestSeller DESC;
```

Row	isBestSeller	product_count	avg_rating	avg_price	avg_reviews	avg_monthly_sales	percentage_of_pr...
1	true	8430	4.51	29.88	2339.0	2341.9	0.7
2	false	1258755	4.4	41.26	183.0	142.9	99.3

Summary

Bestsellers:

- Average Price: \$29.88
- Average Rating: 4.51 stars
- Average Monthly Sales: 2,342 units
- Product Share: 0.7%

Regular Products:

- Average Price: \$41.26
- Average Rating: 4.40 stars
- Average Monthly Sales: 143 units
- Product Share: 99.3%

Price Advantage: Bestsellers are priced approximately 28% lower than regular products.

Key Insights:

- Accessibility Over Exclusivity

Bestsellers are much lower in price than regular products, lending support to the idea that prices can drive volume sales. This supports the idea that consumers of digital marketplaces tend to be price-sensitive and value competitively priced items. The fact that bestsellers are successful at discount pricing indicates that accessibility is preferable to exclusivity in this environment.

- Review Advantage & Social Proof

Bestsellers have 12.8 times as many reviews, generating strong social proof. Large numbers of reviews should increase consumer trust and visibility on retail websites

- The Bestseller Flywheel Effect

The pattern observed is strong feedback loop:

Lower Price → More Sales → More Reviews → More Sales

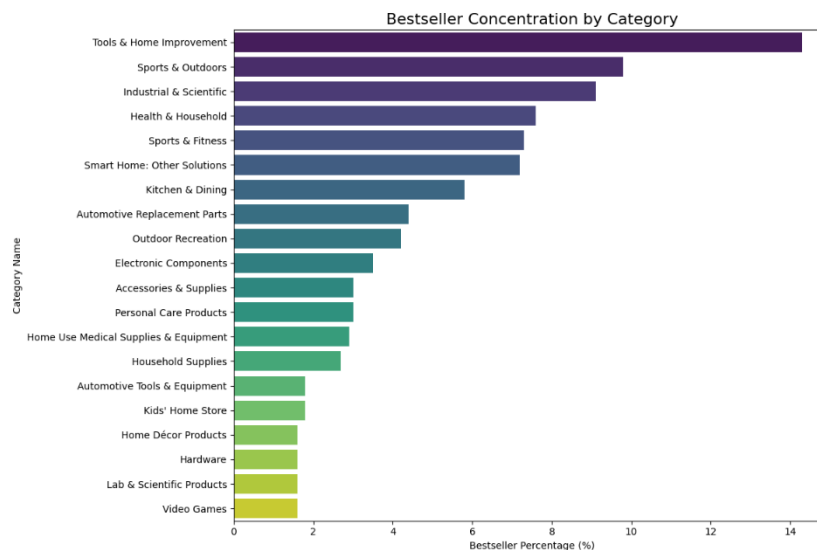
This flywheel effect predicts that once a product gains traction on price and quality, it has the potential to gain speed through normal interaction and social support

9. Category-Specific Bestseller Patterns

Research Question: Which categories have the highest concentration of bestsellers?

SQL Code:

```
SELECT
    category_name,
    COUNT(*) as total_products,
    SUM(CASE WHEN isBestSeller THEN 1 ELSE 0 END) as bestseller_count,
    ROUND(SUM(CASE WHEN isBestSeller THEN 1 ELSE 0 END) * 100.0 / COUNT(*), 1) as
    bestseller_percentage,
    ROUND(AVG(CASE WHEN isBestSeller THEN stars END), 2) as bestseller_avg_rating,
    ROUND(AVG(CASE WHEN NOT isBestSeller THEN stars END), 2) as regular_avg_rating,
    ROUND(AVG(CASE WHEN isBestSeller THEN price END), 2) as bestseller_avg_price,
    ROUND(AVG(CASE WHEN NOT isBestSeller THEN price END), 2) as regular_avg_price
FROM `hultaibigdata.retail_products.products`
WHERE stars > 0 AND price > 0
GROUP BY category_name
HAVING COUNT(*) >= 100 -- Categories with at least 100 products
ORDER BY bestseller_percentage DESC
LIMIT 20;
```



Key Insights:

- Category Winners No Doubt

Tools & Home Improvement takes the lead at 14.5% bestseller rate - nearly 50% ahead of runner-up category

Sports & Outdoors (9.5%) and Industrial & Scientific (9%) form the second tier

Health & Household and Sports & Fitness (both ~7%) form top performers

- Categories in Distress

Video Games, Hardware, Lab & Scientific Products all group in at 1.5-2% - indicative of very competitive/saturated markets

Home Décor and Kids' Home Store also trail at ~2%

- The "Sweet Spot" Categories

4-7% categories (Kitchen & Dining, Auto Parts, Outdoor Rec) show nicely but not too saturated bestseller rates

10. Value Analysis

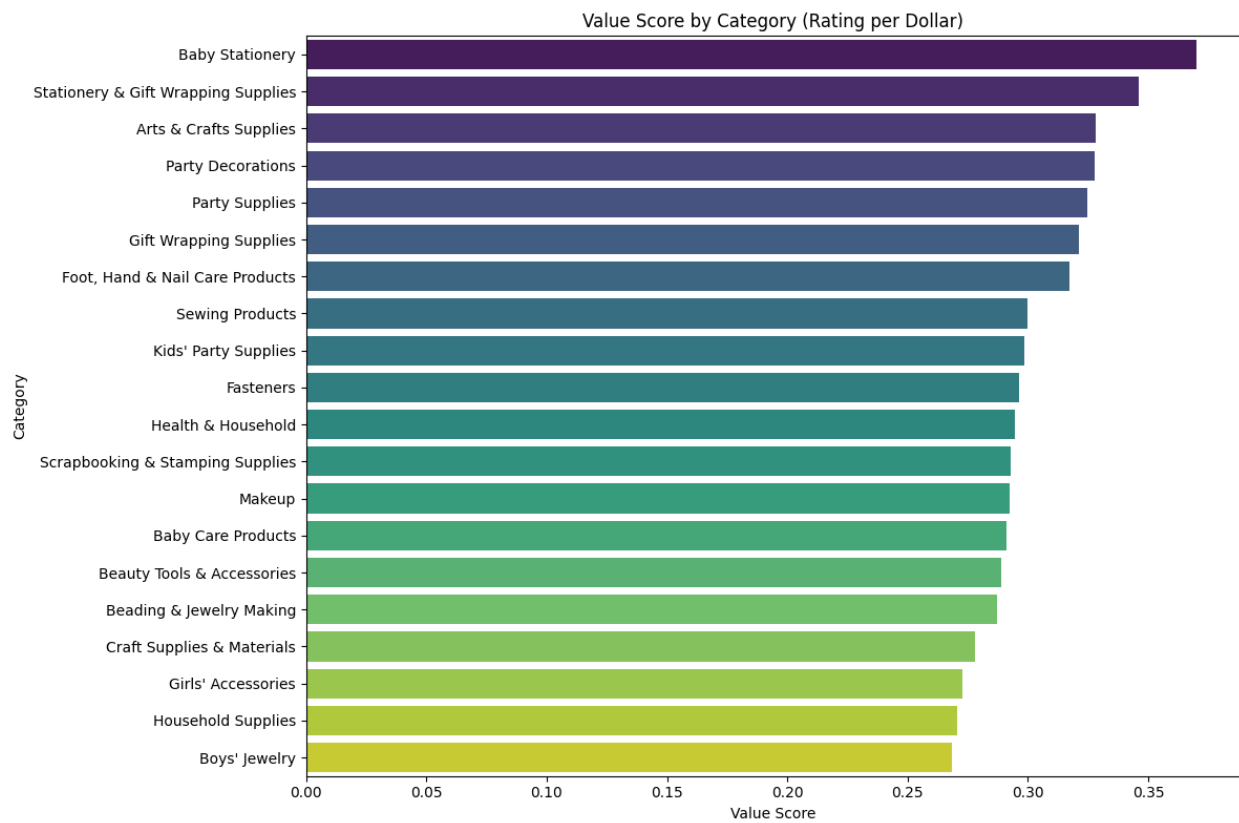
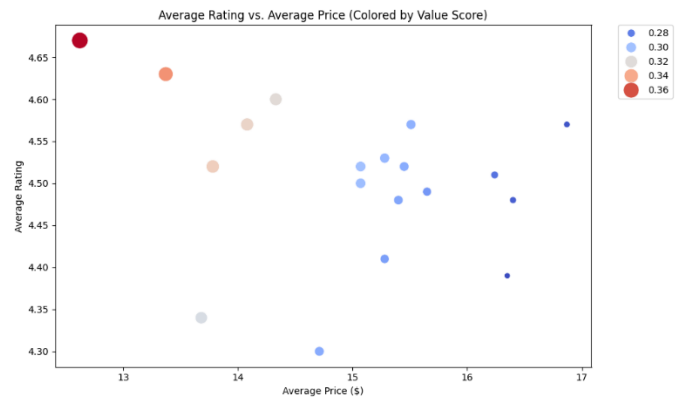
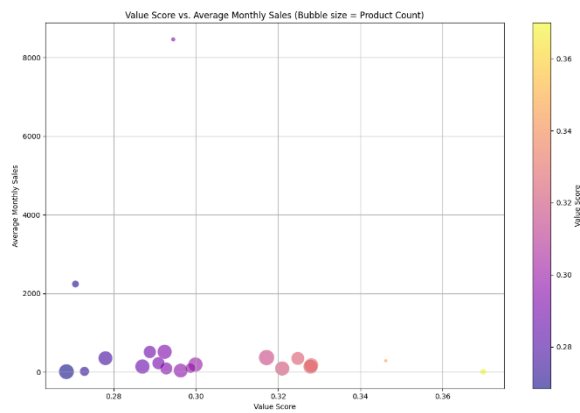
Research Question: Which categories offer the best value for money (high ratings per dollar spent)?

SQL Code:

```
SELECT
    category_name,
    COUNT(*) as product_count,
    ROUND(AVG(stars), 2) as avg_rating,
    ROUND(AVG(price), 2) as avg_price,
    ROUND(AVG(stars) / AVG(price), 4) as value_score, -- Stars per dollar
    ROUND(AVG(boughtInLastMonth), 1) as avg_monthly_sales
FROM `hultaibigdata.retail_products.products`
WHERE stars > 0 AND price > 0
GROUP BY category_name
HAVING COUNT(*) >= 50
```

ORDER BY value_score DESC

LIMIT 20;



Summary

- Top Value Categories:

Baby Stationery: High value score (0.3699) but only 4.4 avg monthly sales

Stationery & Gift Wrapping: High value (0.3500) but medium sales volume

Arts & Crafts Supplies: Strong value (0.3300) but limited market reach

- High-Volume Categories:

Health & Household: Medium value score (0.2945) but 8,458 avg monthly sales

Party Supplies: Average value with periodic volume spikes

Key Insights:

The Value-Volume Paradox: Strict value alone is not enough for sales success. Baby Stationery has 37% greater value per dollar than Health & Household, and yet sells 1,900x slower volume.

- Three Market Behaviors Revealed:

Value \neq Sales Success: Hobby/creative categories offer great value but lack scalability

- The "Value Trap":

High-value specialty products cannot compete with necessity purchases

Necessity Trumps Value: Good-value necessity items defeat high-value specialty items

- Consumer Psychology:

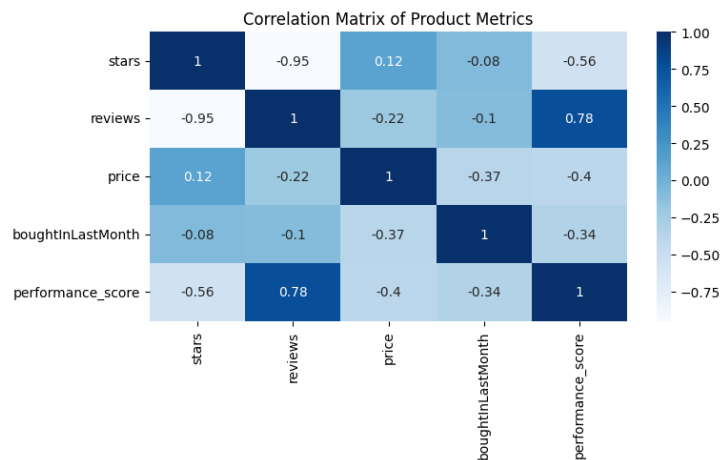
Consumers value frequency and necessity of purchase more than value maximization. The \$12.50-\$14.00 price point shows best value-to-rating performance.

11. Top Individual Product Performance

Research Question: What products achieve the best combination of high ratings AND high review volume?

SQL Code:

```
SELECT
    product_id,
    title,
    category_name,
    stars,
    reviews,
    price,
    boughtInLastMonth,
    isBestSeller,
    -- Composite performance score
    ROUND(stars * LOG(reviews + 1), 2) as performance_score,
    ROUND(reviews * stars, 0) as simple_score
FROM `hultaibigdata.retail_products.products`
WHERE stars > 0 AND reviews > 0 AND price > 0
ORDER BY performance_score DESC
LIMIT 10;
```





Key Findings

Top Performers by Volume & Performance Score:

- Bed Sheets: 281K reviews, performance score 55-56
- HDMI Cables: 151K reviews, performance score 55-56
- Gaming Controllers (DualShock 4): 149K reviews, performance score 55-56
- Roku Streaming Stick+: High performance score leader
- Ring Video Doorbell: Strong performance score with solid engagement

What Makes These Products Dominate:

- Broad Demographic Appeal: Bed sheets are required by everyone, HDMI cables are a requirement for home entertainment systems, gaming controllers serve the massive gaming demographic
- Regular Replacement Cycles: Bed sheets wear out, cables are damaged, controllers receive heavy usage - offering repeat purchase opportunities
- Clear Utility Value: No explanation required - use case is immediately obvious to consumers
- Sweet Spot Pricing: \$14.99-\$32.99 price point offers the perfect combination of value and perceived quality

Critical Performance Insights

- The Volume-Rating Trade-off Strategy:

125K+ reviews = Signal of market penetration: These products have achieved huge scale

- "Good enough" ratings win:

Correlation matrix shows that performance is more linked to review count (0.78) than to perfect ratings (-0.56)

- Perfection loses to engagement:

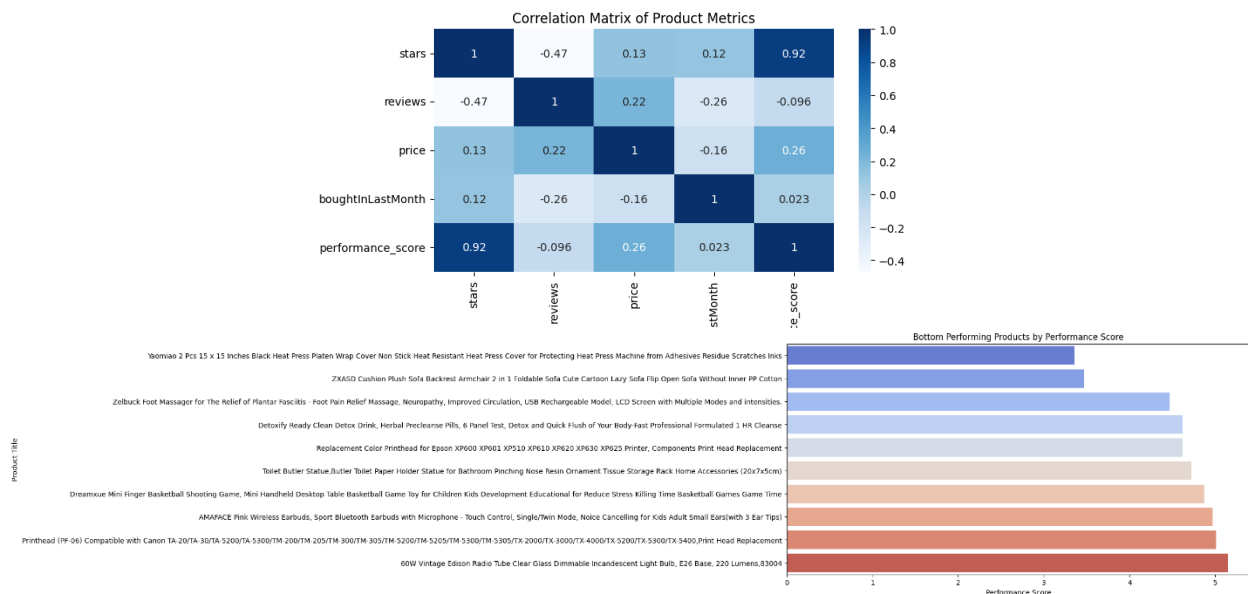
Products with 281K reviews at 4.2 stars outperform products with 500 reviews at 5.0 stars

12. Bottom Individual Product Performance

Research Question: Which products consistently underperform and why?

SQL Code:

```
SELECT
    product_id,
    title,
    category_name,
    stars,
    reviews,
    price,
    boughtInLastMonth,
    isBestSeller,
    ROUND(stars * LOG(reviews + 1), 2) as performance_score
FROM `hultaibigdata.retail_products.products`
WHERE stars > 0 AND reviews >= 10 AND price > 0 -- Need some reviews to be meaningful
ORDER BY performance_score ASC
LIMIT 10;
```



Key Insights:

- Correlation Insights:

Strong Quality-Performance Link: The 0.92 correlation between stars and performance_score confirms that product quality is the principal driver of success

Review Volume vs. Quality Trade-off: The -0.47 correlation between stars and reviews shows that successful products are subject to more analysis and criticism

Price Neutrality: Faint price correlations indicate price is not the only decision driver for performance

Why These Products Fail

- Niche Product Limitation

Printer components (Canon/Epson printheads) and heat press sleeves capture small, niche markets

Small customer base limits growth potential despite quality

High customer acquisition expense as a percentage of market size

- Quality Execution Failures

AMAFACE earbuds and Zelbuck foot massager-type products probably have inferior manufacturing

Low star ratings (1.2-2.0) suggest underlying quality problems

Generic branding implies no investment in product development

- Novelty Without Utility

Goods like the toilet paper holder statue and mini basketball game rely on impulse purchasing

No solid long-term value proposition

Little repeat purchase potential

- 4. Market Positioning Problems

Detox drinks have regulatory and credibility problems

Vintage light bulbs have limited interest to narrow design sensibilities

Poor product-market fit results in irregular demand

Recommendations

This research reveals hidden patterns that contradict conventional retail wisdom. Category choice matters more than perfect execution, with an 18x efficiency gap across categories. Accessibility is more important than exclusivity in success, with bestsellers priced 28% lower than regular products.

- Categories do or die: Choosing efficient, under-served markets beats operational excellence
- Accessibility wins: Universal appeal and cheaper prices generate more sales volume
- Function over form: Problem-solving products enjoy 20x higher bestseller rates than aesthetic ones
- Sweet spot pricing: \$10-25 range optimizes customer satisfaction and market penetration

A) Category-Specific Recommendations

High-Priority Categories

- Health & Household (Highest Performer)

Performance: 547.7 volume per product, 4.57 rating, with only 711 products

Action: Aggressive expansion - scale up the number of products to capitalize on efficiency

Focus: Everyday-use essentials (body wash, deodorant, hand soap, oral care)

- Tools & Home Improvement (Best Market Entry)

Why: Top new product bestseller potential

Strategy: Emphasize practical problem-solving solutions

Opportunity: Untapped niches with genuine innovation possibilities

- Kitchen & Dining (Revenue Leader)

Performance: Top overall sales (10.4M), decent efficiency (81.2 volume per product)

Action: Strategically defend and expand market share

Optimization Categories

- Sports & Outdoors

Status: Competitive second choice with good market opportunity

Approach: Target specific activity niches instead of general fitness

- Personal Care

Opportunity: 2-5x higher sales volumes through social media amplification

Strategy: Focus on visually provable results and habitual inclusion

Avoid/Minimize Categories

- Toys & Games (Worst Performer)

Performance: 17.4 volume per product despite 12,118 products

Action: Aggressive product count reduction

Rationale: Over-saturated with low conversion rates

- Video Games & Hardware

Problem: Low bestseller rates, highly competitive

Solution: Only stock proven winners with high differentiation

Pricing Strategy

- Ideal Price Ranges

Price Range Strategy Use \$10-25 Sweet Spot Universal products, everyday essentials \$15-

40 Volume Optimization Mainstream market penetration \$20-25 Premium Positioning

Differentiated products with clear value \$50+ Avoid Unless justified by exceptional performance

Pricing Principles

- Target affordability: Lower prices drive higher volumes and bestseller potential
- Set expectations: Price influences satisfaction more than absolute quality
- Avoid premium traps: High prices reduce market accessibility without proportionate gains

Product Development Strategy

Universal Appeal Framework

- Target Universal Needs: Products that everyone uses regularly
- Problem-Solving Focus: Address narrow, universal pain points
- Routine Integration: Fit easily into daily routines
- Quality Threshold: Maintain 4.0+ star ratings consistently
- Product Line Development

- Build product families: Success like CeraVe illustrates line extension power
- Bundle complementary products: Leverage cross-selling opportunities
- Highlight functional benefits: Performance metrics sell more than ratings
- Innovation rather than incrementation: Actual innovation beats incremental improvements

Brand Strategy

- Leverage existing partnerships: Dove, Amazon Basics, L'Oreal
- Promote trust signals: Dermatologist-recommended, hypoallergenic certifications
- Build social media presence: Personal care categories benefit from visual marketing
- Authentic positioning: Self-care rituals rather than basic necessities

Market Entry Strategy

- Entry Priorities

First Target: Tools & Home Improvement (highest chance of success)

Secondary Target: Sports & Outdoors (good opportunity without oversaturation)

Specialty Play: Health & Household (if innovation-capable)

- Entry Tactics

Launch with high-frequency items: Hand soap, toothbrush, replacement parts

Offer value sizing: Family/multi-pack options for better perception

Build review momentum: Emphasize verified customer reviews

Price competitively: Utilize affordability to gain early market share

- Risk Mitigation

Avoid oversaturated categories: Unless significant competitive advantage exists

Test with small SKU count: Validate demand before full investment

Target underserved niches: More effective than fighting in competitive segments

Conclusion:

The shopping world favors organizations that see data-driven trends rather than traditional presumptions. Success is the result of:

Market Selection > Execution Perfection: Choose winning categories first Accessibility > Exclusivity: Reduce barriers to achieve higher volumes

Function > Form: Problem-solving trumps beauty appeal Universal > Niche: Wide appeal creates long-term success

Bottom Line

Firms need to emphasize analytical acumen to sense and act on these inaudible patterns. The companies that achieve superior execution on universal needs will beat the ones that are struggling with niche expertise or premium positioning with no value rationale.

References:

- <https://www.forbes.com/councils/forbesbusinesscouncil/2022/04/06/the-secret-to-exponential-business-growth-the-flywheel/>
- Anthropic. (2025). Claude [Large language model]. <https://claude.ai>
- OpenAI. (2025). ChatGPT [Large language model]. <https://chat.openai.com>

Vanna.AI:

The screenshot shows the Vanna.AI web interface. On the left sidebar, there's a 'Vanna.AI' logo, a 'Training Data' button, and a '+ New question' button. The main area has a header 'Your AI-powered copilot for SQL queries.' Below it, a user query is shown: 'Write a SQL query to perform initial data exploration on my products table. I want to understand the basic statistics including total number of products, number of categories, price range, average ratings, and average monthly sales. The table is hultaibigdata.retail_products.products with columns: category_name, price, stars, and boughtInLastMonth.' The AI response is a SQL query:

```
SELECT
COUNT(*) AS total_products,
COUNT(DISTINCT category_name) AS number_of_categories,
MIN(price) AS min_price,
MAX(price) AS max_price,
AVG(stars) AS average_rating,
AVG(boughtInLastMonth) AS average_monthly_sales
FROM
`hultaibigdata.retail_products.products`
```

 Below the query, a table with 6 columns is shown: TOTAL_PRODUCTS, NUMBER_OF_CATEGORIES, MIN_PRICE, MAX_PRICE, AVERAGE_RATING, AVER. At the bottom, there's a prompt: 'Ask me a question about your data that I can turn into SQL.'

The screenshot shows the Vanna.AI web interface. On the left sidebar, there's a 'Vanna.AI' logo, a 'Training Data' button, and a '+ New question' button. The main area has a header 'Your AI-powered copilot for SQL queries.' Below it, a user query is shown: 'Write a SQL query to perform initial data exploration on my products table. I want to understand the basic statistics including total number of products, number of categories, price range, average ratings, and average monthly sales. The table is hultaibigdata.retail_products.products with columns: category_name, price, stars, and boughtInLastMonth.' The AI response is a SQL query:

```
WITH category_efficiency AS (
SELECT
category_name,
COUNT(*) as product_count,
ROUND(AVG(boughtInLastMonth), 1) as avg_sales_per_product,
ROUND(AVG(stars), 2) as avg_rating,
ROUND(AVG(price), 2) as avg_price,
ROUND(SUM(boughtInLastMonth), 0) as total_sales
FROM `hultaibigdata.retail_products.products`
WHERE stars > 0 AND boughtInLastMonth > 0 AND price > 0
GROUP BY category_name
HAVING COUNT(*) >= 50
)
SELECT
category_name,
product_count,
avg_sales_per_product,
avg_rating,
avg_price,
total_sales,
-- Efficiency metrics
CASE
WHEN avg_sales_per_product > 2000 THEN 'Super Efficient'
WHEN avg_sales_per_product > 1000 THEN 'High Efficiency'
WHEN avg_sales_per_product > 500 THEN 'Moderate Efficiency'
ELSE 'Low Efficiency'

```

 Below the query, a table with 6 columns is shown: TOTAL_PRODUCTS, NUMBER_OF_CATEGORIES, MIN_PRICE, MAX_PRICE, AVERAGE_RATING, AVER. At the bottom, there's a prompt: 'Ask me a question about your data that I can turn into SQL.'

You Analyze the best products in the Toys & Games category. For each product, calculate how their monthly sales compare to the category average (as a ratio/multiplier). Show the top 25 products ranked by sales volume.

```
SELECT
  p.title,
  p.boughtInLastMonth,
  AVG(p.boughtInLastMonth) OVER () AS category_avg_sales,
  (p.boughtInLastMonth / (AVG(p.boughtInLastMonth) OVER ())) AS sales_ratio
FROM
  `hultaibigdata.retail_products.products` AS p
WHERE p.category_name = 'Toys & Games'
ORDER BY p.boughtInLastMonth DESC
LIMIT 25
```

Re-Run SQL

Sassy Stacks of Circles Stacking Ring STEM Learning Toy, Age 6+ Months, Multi, 9 Piece Set

SQL

Vanna.AI

Training Data

+ New question

Write a SQL query to perform data exploration on my products. I want to understand the basic statistics including total number of products, number of categories, price range, average ratings, and average monthly sales. The table is `hultaibigdata.retail_products.products` with columns: `category_name`, `price`, `stars`, and `boughtInLastMonth`.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and sales volume. Include summary statistics for each category and filter out categories with insufficient data.

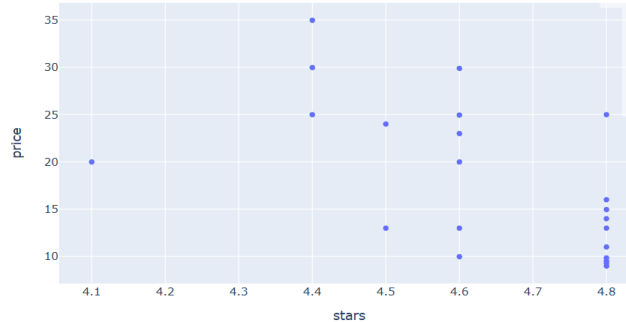
Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and

```
title,
stars,
price,
reviews,
boughtInLastMonth,
isBestSeller,
ROUND(boughtInLastMonth / (
  SELECT AVG(boughtInLastMonth)
  FROM `hultaibigdata.retail_products.products`
  WHERE category_name = 'Toys & Games'
  AND stars > 0
  AND boughtInLastMonth > 0
), 2) AS performance_vs_category_avg
FROM
  `hultaibigdata.retail_products.products`
WHERE
  category_name = 'Toys & Games'
  AND stars > 0
  AND boughtInLastMonth > 0
ORDER BY
  boughtInLastMonth DESC
LIMIT 25
```

Re-Run SQL

STARS PRICE REVIEWS BOUGHTINLASTMONTH ISBESTSELLER PEF

SQL



Redraw Chart

Re-Run SQL

The data shows 25 top-selling "Toys & Games" products, ranked by "boughtInLastMonth," with a calculated `performance_vs_category_avg` indicating how each product's sales compare to the average for the category. Most products have a performance score of approximately 21.16, suggesting they perform similarly to the

Vanna.AI

Training Data

New question

Write a SQL query to perform in data exploration on my products. I want to understand the basic s including total number of produ number of categories, price rang average ratings, and average m sales. The table is hultaibigdata.retail_products.pro with columns: category_name, p stars, and boughtInLastMonth.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and sales volume. Include summary statistics for each category and filter out categories with insufficient data.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and

```

CASE
  WHEN price < 10 THEN 'Under $10'
  WHEN price < 25 THEN '$10-25'
  WHEN price < 50 THEN '$25-50'
  WHEN price < 100 THEN '$50-100'
  ELSE 'Over $100'
END as price_range,
COUNT(*) as product_count,
ROUND(AVG(stars), 2) as avg_rating,
ROUND(AVG(boughtInLastMonth), 1) as avg_monthly_sales,
ROUND(AVG(reviews), 0) as avg_reviews
FROM
`hultaibigdata.retail_products.products`
WHERE price > 0 AND stars > 0
GROUP BY 1
ORDER BY avg_rating DESC

```

PRICE_RANGE	PRODUCT_COUNT	AVG_RATING	AVG_MONTHLY_SALES	AVG_REVIEWS
\$10-25	534120	4.42	166.8	195
\$25-50	281751	4.41	137.9	194
Under \$10	247980	4.4	238.2	246
Over \$100	91342	4.34	41.7	129

Ask me a question about your data that I can turn into SQL.

Vanna.AI

Training Data

New question

Write a SQL query to perform in data exploration on my products. I want to understand the basic s including total number of produ number of categories, price rang average ratings, and average m sales. The table is hultaibigdata.retail_products.pro with columns: category_name, p stars, and boughtInLastMonth.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and sales volume. Include summary statistics for each category and filter out categories with insufficient data.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and

```

SELECT
CASE
  WHEN price >= 100 THEN 'Premium ($100+)'
  WHEN price >= 50 THEN 'High-End ($50-99)'
  WHEN price >= 25 THEN 'Mid-Range ($25-49)'
  WHEN price >= 10 THEN 'Sweet Spot ($10-24)'
  ELSE 'Budget (<$10)'
END AS price_category,
COUNT(*) AS products,
ROUND(AVG(stars), 2) AS avg_satisfaction,
ROUND(AVG(boughtInLastMonth), 1) AS avg_sales,
SUM(CASE WHEN stars <= 3.5 THEN 1 ELSE 0 END) AS low_satisfaction_products,
ROUND(SUM(CASE WHEN stars <= 3.5 THEN 1 ELSE 0 END) * 100.0 / COUNT(*), 1) AS pct_unsatisfied
FROM
`hultaibigdata.retail_products.products`
WHERE price > 0 AND stars > 0
GROUP BY
CASE
  WHEN price >= 100 THEN 'Premium ($100+)'
  WHEN price >= 50 THEN 'High-End ($50-99)'
  WHEN price >= 25 THEN 'Mid-Range ($25-49)'
  WHEN price >= 10 THEN 'Sweet Spot ($10-24)'
  ELSE 'Budget (<$10)'
END

```

Ask me a question about your data that I can turn into SQL.

Vanna.AI

Training Data

New question

Write a SQL query to perform in data exploration on my products. I want to understand the basic s including total number of produ number of categories, price rang average ratings, and average m sales. The table is hultaibigdata.retail_products.pro with columns: category_name, p stars, and boughtInLastMonth.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and sales volume. Include summary statistics for each category and filter out categories with insufficient data.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and

PRICE_CATEGORY	PRODUCTS	AVG_SATISFACTION	AVG_SALES	LOW_SATISFACTION_PRODUCT
Sweet Spot (\$10-24)	534120	4.42	166.8	19789
Mid-Range (\$25-49)	281751	4.41	137.9	10731
Budget (<\$10)	247980	4.4	238.2	11070
High-End (\$50-99)	111992	4.39	78.7	5066
Premium (\$100+)	91342	4.34	41.7	5913

CSV
 Download

Ask me a question about your data that I can turn into SQL.

Vanna.AI

Training Data

New question

Write a SQL query to perform in data exploration on my products. I want to understand the basic s including total number of produ number of categories, price rang average ratings, and average m sales. The table is hultaibigdata.retail_products.pro with columns: category_name, p stars, and boughtInLastMonth.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and sales volume. Include summary statistics for each category and filter out categories with insufficient data.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and

PRICE_CATEGORY	PRODUCTS	AVG_SATISFACTION	AVG_SALES	LOW_SATISFACTION_PRODUCT
Sweet Spot (\$10-24)	534120	4.42	166.8	19789
Mid-Range (\$25-49)	281751	4.41	137.9	10731
Budget (<\$10)	247980	4.4	238.2	11070
High-End (\$50-99)	111992	4.39	78.7	5066
Premium (\$100+)	91342	4.34	41.7	5913

CSV

Download

Ask me a question about your data that I can turn into SQL.

Vanna.AI

Training Data

New question

Write a SQL query to perform in data exploration on my products. I want to understand the basic s including total number of produ number of categories, price rang average ratings, and average m sales. The table is hultaibigdata.retail_products.pro with columns: category_name, p stars, and boughtInLastMonth.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and sales volume. Include summary statistics for each category and filter out categories with insufficient data.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and

Welcome to Vanna.AI

Your AI-powered copilot for SQL queries.

Open Debugger

You

Write a SQL query to compare bestsellers vs regular products. Group by bestseller status and show average metrics for each group including stars, price, reviews, and monthly sales. Also calculate what percentage of the total product catalog each group represents.

WITH ProductMetrics AS (

SELECT

isBestSeller,

AVG(stars) AS avg_stars,

AVG(price) AS avg_price,

AVG(reviews) AS avg_reviews,

AVG(boughtInLastMonth) AS avg_monthly_sales

FROM

"hultaibigdata.retail_products.products"

GROUP BY

isBestSeller

Ask me a question about your data that I can turn into SQL.

avg_stars,

Training Data

+ New question

Write a SQL query to perform in data exploration on my products. I want to understand the basic statistics including total number of products, number of categories, price range, average ratings, and average monthly sales. The table is `hultaibigdata.retail_products.products` with columns: `category_name`, `product_count`, `avg_rating`, `avg_price`, `stars`, and `boughtInLastMonth`.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and sales volume. Include summary statistics for each category and filter out categories with insufficient data.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and

✓

```

SELECT
    category_name, COUNT(*) as product_count,
    ROUND(AVG(stars), 2) as avg_rating,
    ROUND(AVG(price), 2) as avg_price,
    ROUND(AVG(stars) / AVG(price), 4) as value_score,
    -- Stars per dollar
    ROUND(AVG(boughtInLastMonth), 1) as avg_monthly_sales
FROM `hultaibigdata.retail_products.products`
WHERE stars > 0 AND price > 0
GROUP BY category_name
HAVING COUNT(*) >= 50
ORDER BY value_score DESC
LIMIT 20

```

✓

CATEGORY_NAME	PRODUCT_COUNT	AVG_RATING	AVG_PRICE	VALUE_SCORE
Baby Stationery	1321	4.67	12.62	0.3699
Stationery & Gift Wrapping Supplies	431	4.63	13.37	0.3462

Ask me a question about your data that I can turn into SQL.

Vanna.AI

Training Data

+ New question

Write a SQL query to perform in data exploration on my products. I want to understand the basic statistics including total number of products, number of categories, price range, average ratings, and average monthly sales. The table is `hultaibigdata.retail_products.products` with columns: `category_name`, `product_count`, `avg_rating`, `avg_price`, `stars`, and `boughtInLastMonth`.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and sales volume. Include summary statistics for each category and filter out categories with insufficient data.

Write a SQL query to identify the top-performing product categories based on both customer satisfaction (ratings) and

You

"Write SQL queries to find the top 10 and bottom 10 products ranked by a performance score calculated as $\text{stars} \times \text{LOG}(\text{reviews} + 1)$. Also show a simple score of $\text{stars} \times \text{reviews}$ for comparison. Include product details and filter out products with insufficient data."

✓

```

SELECT
    product_id,
    title,
    stars,
    reviews,
    stars * (LOG(reviews + 1)) AS performance_score,
    stars * reviews AS simple_score
FROM
    `hultaibigdata.retail_products.products`
WHERE
    reviews > 0
ORDER BY
    performance_score DESC
LIMIT 10

UNION ALL

```

Ask me a question about your data that I can turn into SQL.

37