ATENEO DE MANILA UNIVERSITY

LOYOLA SCHOOLS

School of Science and Engineering

Department of Information Systems and Computer Science

Second Semester

2024-2025

ISCS 30.19

Big Data Analysis

Group Project 2

Written by:

Damalerio, Adrian Lance

Go, Jared

Tan, Tristan

A.Y. 2024-2025

Executive Summary

This paper applies big data analytics to the Looker E-Commerce dataset to identify strategies for maximizing profit and revenue by optimizing product performance and targeting high-value customer segments and geographic markets. The analysis reveals that profitability is concentrated in a few key regions, particularly China and the United States, and among middle-aged users. Premium products and brands such as outerwear and Calvin Klein drive the bulk of revenue and profit, while a large portion of the catalog suffers from low sell-through or underperformance. Inventory insights uncovered over 5,000 poorly converting products and 129 items with near-zero sales, signaling clear opportunities for deprioritization. By implementing K-Means clustering, the paper also segments products into strategic profiles, enabling data-driven inventory and pricing decisions. Overall, the findings highlight how big data can guide e-commerce companies in reallocating resources, improving targeting, and sustaining long-term growth in a competitive digital market.

Table of Contents

Introduction

The rapid expansion of e-commerce has transformed the way businesses interact with consumers, creating new opportunities and challenges for revenue growth and operational efficiency. In this dynamic environment, success is increasingly becoming dependent on the company's ability to harness and interpret large volumes of data. Big data analysis empowers businesses to go beyond traditional decision-making, uncovering deep insights into customer behavior, product performance, and market trends.

By systematically analyzing patterns and trends in sales, inventory, and user engagement, e-commerce companies can optimize their operations, tailor marketing strategies, and make informed product decisions that directly impact profitability. In today's world, the ability to identify high-value customer segments, forecast demand accurately, and manage inventory intelligently is no longer a competitive advantage. It has become a necessity for survival in the crowded ecommerce space.

This paper emphasizes the critical role of big data analytics in driving ecommerce success. Through the application of data science techniques to real-world e-commerce datasets, it demonstrates how businesses can move from reactive decision-making to proactive, strategic management. In today's world where customer preferences shift rapidly and competition is fierce, leveraging data is key to maximizing profit, sustaining growth, and securing long-term market relevance.

About the Dataset

This analysis is based on the Looker Ecommerce Dataset, which offers a comprehensive view of online retail operations. The dataset integrates multiple sources, providing detailed information across inventory management, product attributes, order transactions, and user demographics.

Specifically, it combines data from five key files: inventory.csv, order_items.csv, products.csv, orders.csv, and users.csv. Together, these sources create a unified picture of how products move through the sales cycle, how customers engage with the platform, and how revenue and profit are generated.

The dataset covers:

- **Inventory Behavior:** 490,705 inventory entries tracking stock availability and sales timing.
- **Sales Activity:** 125,226 orders and 181,759 individual order items, capturing purchase trends and product performance.

- **Product Metadata:** 29,120 products, with details on pricing, cost, brand, category, and department.
- **Customer Profiles:** 100,000 user records including age, gender, and country information.

In addition to basic transaction data, the dataset includes profitability metrics, order statuses, and temporal features such as stocked and sold timestamps. This breadth of coverage enables a multi-dimensional analysis of both supply-side (inventory and products) and demand-side (orders and users) factors that influence ecommerce performance.

The size and scale of the Looker Ecommerce Dataset make it an ideal foundation for investigating strategies to optimize product performance, target high-value customer segments, and ultimately maximize business profitability.

Research Objective and Guiding Questions

To guide the analysis of the dataset, this paper is structured around the following key business problem:

"How can the business maximize profit and revenue by optimizing product performance and targeting high-value customer segments and geographic opportunities?"

This central business problem will be explored through thematic areas with each having their own sub questions on the matter. The following thematic areas are shown below:

Company Portfolio: Profitability Insights

1. How has Looker E-Commerce performed over time in terms of profitability? (2019-2024)
2. Which country contributes the most and least profit to Looker E-Commerce?
3. Do user attributes such as gender and age correlate with higher spending or profitability?

Product Performance

4. Which products generate the highest and lowest total sales volume, revenue and profit? (All-time)
5. Which product categories generate the highest and lowest total sales volume, revenue and profit?
6. Which product brands generate the highest and lowest total sales volume, revenue and profit?

Inventory and Sell-Through Effectiveness

7. What is the sold-stocked ratio (sold versus stocked) per product?

8. Which products are overstocked or have low conversion despite high inventory?
9. Are there products with zero or near-zero sales that should be deprioritized?

Modeling

Findings

The findings to each subproblem are listed below including the methodology used to come up with the results. These findings will then be used to answer the main question of "How can the business maximize profit and revenue by optimizing product performance and targeting high-value customer segments and geographic opportunities?"

Subquestion 1

"How has Looker E-Commerce performed over time in terms of profitability? (2019-2024)"

Methodology:

To evaluate how the platform's profitability performed from 2019 to 2024 and identify any trends, the analysis first focused on cleaning and preparing the transaction data. The "sold_at" timestamps were parsed into a consistent datetime format, with invalid entries safely coerced and dropped to ensure data integrity. Afterward, each transaction was assigned to a specific month using a standardized "YYYY-MM" format. Aggregation techniques were then applied to group the data by month, calculating the total profit earned within each period. The resulting monthly profits were then organized chronologically to showcase any clear trend within the time period. Finally, a line was plotted connecting all the data points to visually represent the company's monthly profitability patterns, allowing for the identification of growth periods, declines, or seasonal behaviors across the six-year range.

Code:

```
# Safely parse 'sold_at' using mixed format handling
df_analysis['sold_at'] = pd.to_datetime(df_analysis['sold_at'],
format='mixed', errors='coerce')

# Drop any rows where sold_at couldn't be parsed
df_analysis = df_analysis.dropna(subset=['sold_at'])

# Extract the month in 'YYYY-MM' format
df_analysis['month'] =
df_analysis['sold_at'].dt.to_period('M').astype(str)
```

```
# Group by month to calculate total profit
monthly_profit = df_analysis.groupby('month').agg(
    total_profit=('profit', 'sum')
).reset_index()

# Sort by chronological month
monthly_profit['month'] = pd.to_datetime(monthly_profit['month'])
monthly_profit = monthly_profit.sort_values('month')

# Plot: Monthly profit trend
plt.figure(figsize=(14, 6))
sns.lineplot(data=monthly_profit, x='month', y='total_profit', marker='o')
plt.title('Total Monthly Profit Over Time')
plt.xlabel('Month')
plt.ylabel('Total Profit')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Results and Interpretation:

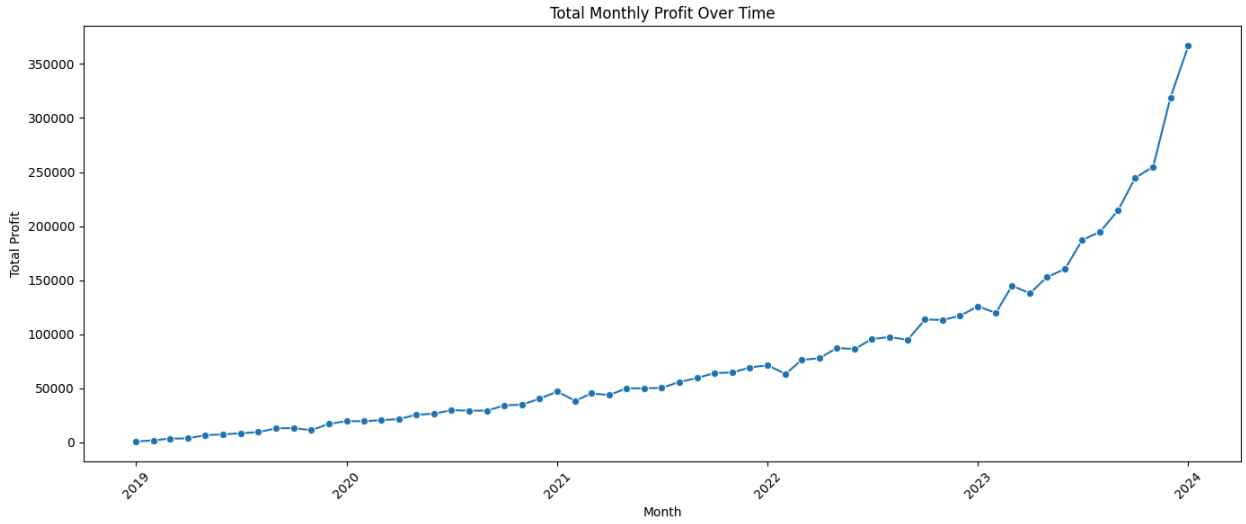

Figure 1. Profitability Analysis of Looker E-Commerce

The results show a clear upward trajectory in total monthly profit from 2019-2024. Early in 2019, profits were modest but consistently positive, indicating a stable start. From 2019

through mid-2022, profit growth was steady and increased gradually. It was also marked by minor fluctuations that suggest periodic variations in sales or operational efficiency.

Starting in late 2022 to 2023, the results showed an observable and sustained increase in profitability. From 2023 to 2024, monthly profits began rising at an exponential rate, with total profits reaching their highest levels by early 2024. This inflection point suggests that the company may have implemented successful growth strategies, like expanding product offerings, improving customer acquisition, or optimizing operations, which led to accelerated revenue generation and improved margins.

Overall, the company's performance over time reflects strong and robust profitability along with rapid growth observed over the past two years. The trend suggests that Looker E-Commerce has transitioned from a phase of stable but moderate profitability to a phase of rapid and potentially scalable growth, positioning the company strongly for continued success if current trends persist.

Subquestion 2

“Which country contributes the most and least profit to Looker E-Commerce?”

Methodology

To determine which countries contribute the most to the company’s profitability, inconsistencies in country names within the dataset were first addressed by standardizing entries, such as replacing "España" with "Spain" and "Deutschland" with "Germany." After ensuring consistency, the transactional data per country were grouped. For each country, three key metrics were calculated: the total profit, the total revenue, and the total number of sales (measured by counting individual order items). Aggregating the data in this manner provided a holistic view of each country's overall contribution to the business. To enhance interpretability, the results were visualized through bar plots, with three pairs of charts generated. Each pair displayed the top and bottom eight countries based on total profit, total revenue, and total sales, respectively. These visualizations supplemented the summary table, making it easier to compare country-level performance across different dimensions. The resulting summary table was then sorted in descending order by total profit, allowing for straightforward identification of the highest and lowest performing countries in terms of profitability.

Code:

```
# Group by country and sum profit
df_country = df.copy()

# Create a mapping of country names that need to be unified
country_mapping = {
    'España': 'Spain',
```

```
    'Deutschland': 'Germany'
}

# Replace country names based on the mapping
df_country['country'] = df_country['country'].replace(country_mapping)

# Group by country and sum profit
country_profit = df_country.groupby('country').agg(
    total_profit=('profit', 'sum'),
    total_revenue=('revenue', 'sum'),
    total_sales=('order_items_id', 'count')
).reset_index().sort_values(by='total_profit', ascending=False)

# Display the result
print(country_profit)

# Bar plot for top 8 countries
plt.figure(figsize=(12, 6))
sns.barplot(x='total_profit', y='country', data=country_profit.head(8),
palette='crest')
plt.title('Top 8 Countries by Total Profit')
plt.xlabel('Total Profit (in millions)')
plt.ylabel('Country')
plt.tight_layout()
plt.show()

# Bottom 8 countries
plt.figure(figsize=(12, 6))
sns.barplot(x='total_profit', y='country', data=country_profit.tail(8),
palette='flare')
plt.title('Bottom 8 Countries by Total Profit')
plt.xlabel('Total Profit (in millions)')
plt.ylabel('Country')
plt.tight_layout()
plt.show()

# Bar plot for top 8 countries
plt.figure(figsize=(12, 6))
sns.barplot(x='total_revenue', y='country', data=country_profit.head(8),
palette='crest')
```

```
plt.title('Top 8 Countries by Total Revenue')
plt.xlabel('Total Revenue (in millions)')
plt.ylabel('Country')
plt.tight_layout()
plt.show()

# Bar plot for bottom 8 countries
plt.figure(figsize=(12, 6))
sns.barplot(x='total_revenue', y='country', data=country_profit.tail(8),
palette='flare')
plt.title('Bottom 8 Countries by Total Revenue')
plt.xlabel('Total Revenue (in millions)')
plt.ylabel('Country')
plt.tight_layout()
plt.show()

# Bar plot for top 8 countries
plt.figure(figsize=(12, 6))
sns.barplot(x='total_sales', y='country', data=country_profit.head(8),
palette='crest')
plt.title('Top 8 Countries by Total Sales')
plt.xlabel('Total Sales')
plt.ylabel('Country')
plt.tight_layout()
plt.show()

# Bar plot for bottom 8 countries
plt.figure(figsize=(12, 6))
sns.barplot(x='total_sales', y='country', data=country_profit.tail(8),
palette='flare')
plt.title('Bottom 8 Countries by Total Sales')
plt.xlabel('Total Sales')
plt.ylabel('Country')
plt.tight_layout()
plt.show()
```

Results and Interpretation:

Table 1. Country Contribution to Profit, Revenue, and Number of Sales

Country	**Total Profit**	**Total Revenue**	**Total Sales**
China	1,647,702.00	3,177,001.00	53,203
United States	1,071,953.00	2,067,572.00	34,707
Brasil	685,105.70	1,319,042.00	22,359
South Korea	258,205.50	497,781.00	8,235
United Kingdom	225,269.50	432,328.10	7,050
France	224,338.70	430,836.30	7,282
Germany	196,819.90	380,087.20	6,574
Spain	194,210.20	372,972.60	6,144
Japan	115,406.30	221,489.60	3,686
Australia	97,385.12	188,351.90	3,222
Belgium	55,994.78	107,811.00	1,784
Poland	12,542.19	23,749.94	402
Colombia	480.67	949.98	19
Austria	42.98	84.49	2

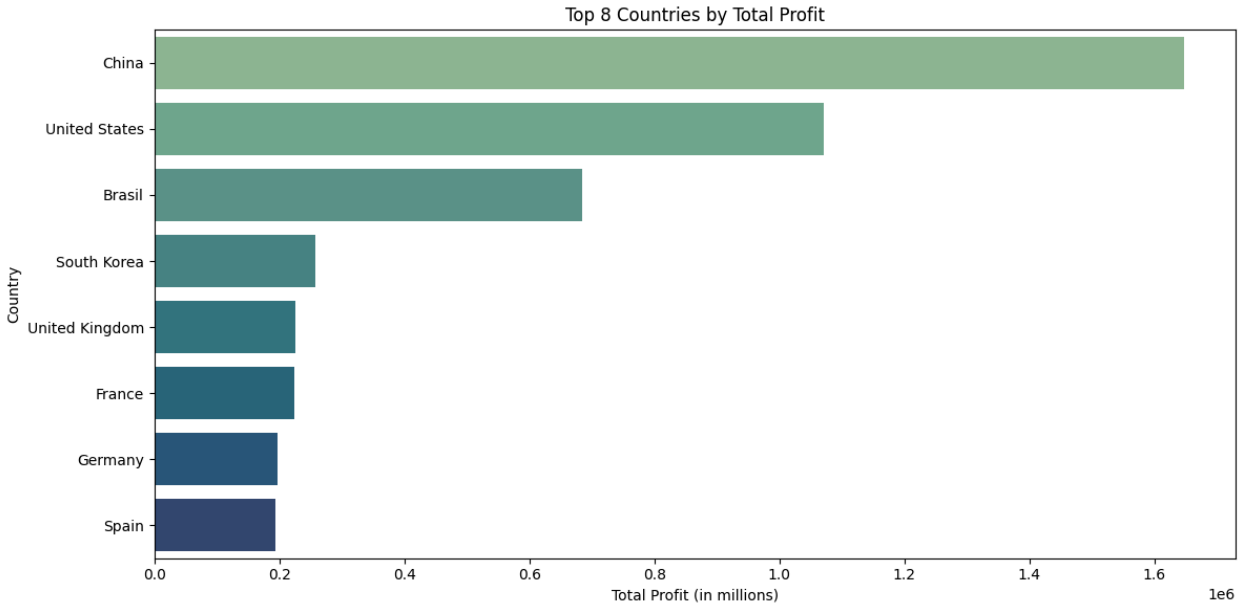

Figure 2. Top 8 Countries by Total Profit

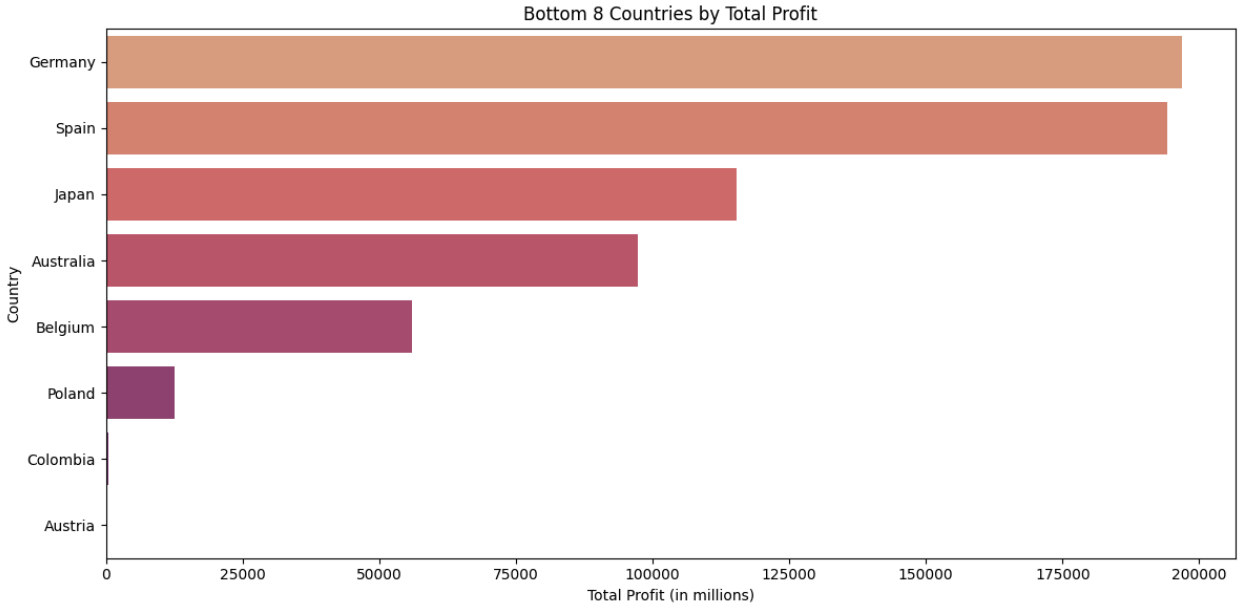

Figure 3. Bottom 8 Countries by Total Profit

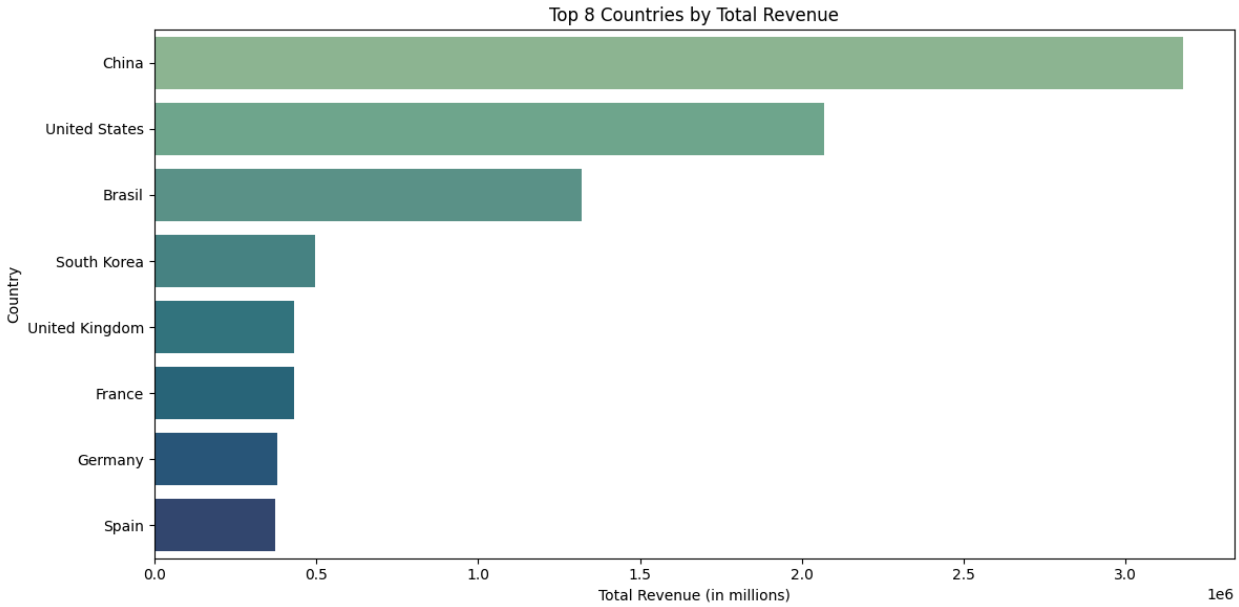

Figure 4. Top 8 Countries by Total Revenue

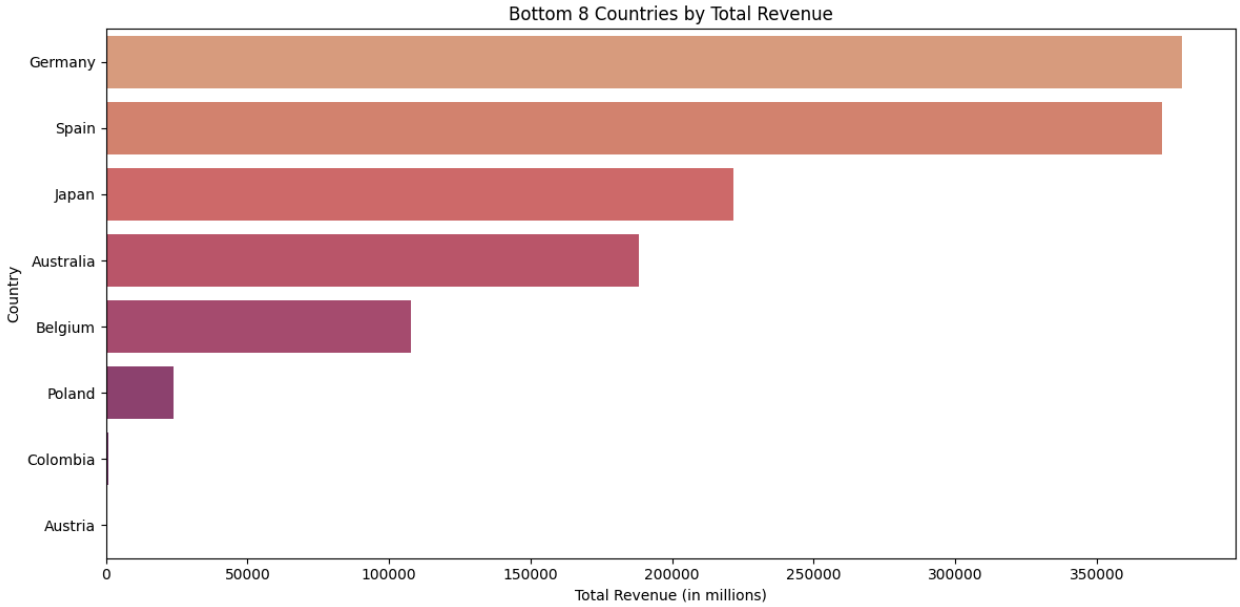

Figure 5. Bottom 8 Countries by Total Revenue

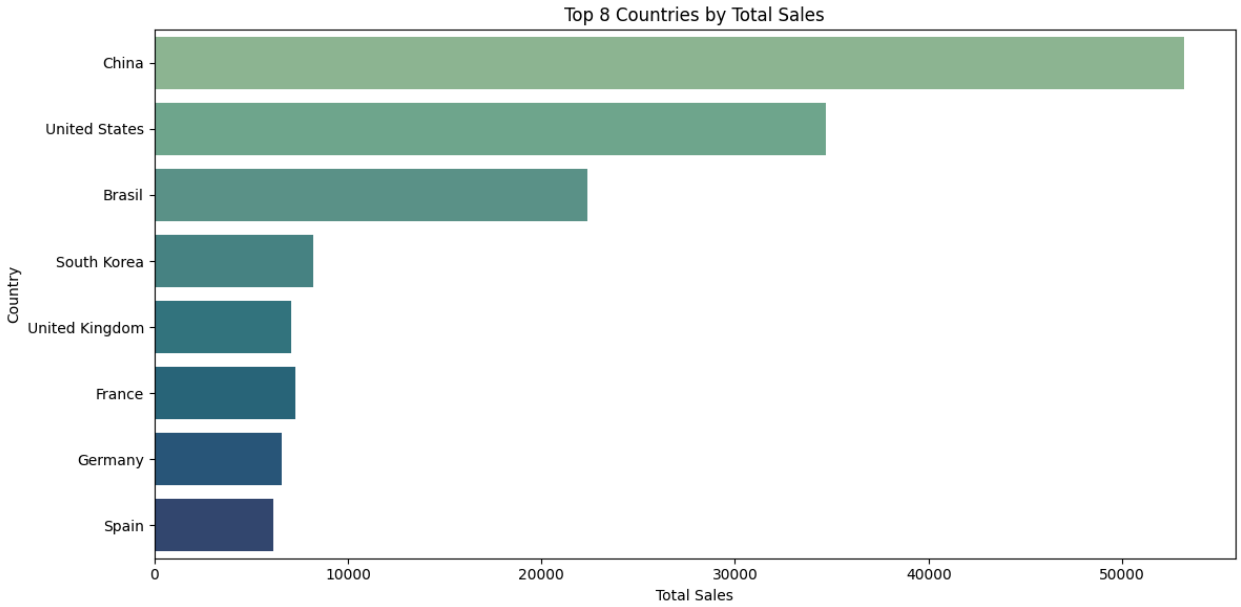

Figure 6. Top 8 Countries by Total Sales

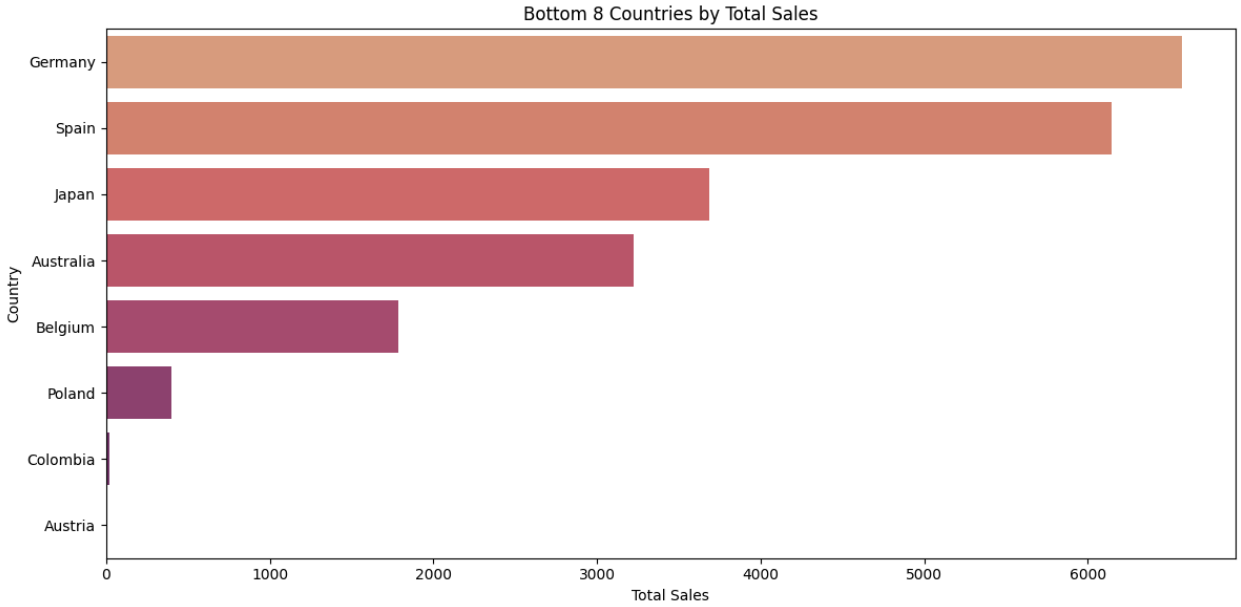

Figure 7. Bottom 8 Countries by Total Sales

The analysis of country-level contributions to the company's profitability reveals notable differences across markets. As shown in Table 1 and illustrated in Figures 2–7, China is the largest contributor by a significant margin, generating the highest total profit ($1,647,702), total revenue ($3,177,001), and total number of sales (53,203 orders). Following China, the United States ranks second across all three metrics, with a total profit of $1,071,953 and 34,707 sales.

Brasil consistently ranks third, showing strong sales performance and profitability relative to other countries.

Conversely, countries such as Austria, Colombia, and Poland represent the least profitable markets. Austria posted the lowest total profit ($42.98), with only 2 recorded sales, while Colombia and Poland also showed very low figures both in profit and volume. These findings are further supported by the bottom eight country bar plots, where Austria, Colombia, and Poland consistently appear at the lower end across total profit, revenue, and sales.

Overall, the results suggest that Looker E-Commerce's profitability is heavily concentrated in a few key countries, particularly China and the United States, while smaller markets contribute minimally. This highlights a potential opportunity for the company to either invest more strategically in high-performing regions or reevaluate its presence in underperforming markets.

Subquestion 3

"Do user attributes such as gender and age correlate with higher spending or profitability?"

Methodology:

To determine whether user attributes such as gender and age correlate with higher spending or profitability, the transactional data was first grouped by gender. For each gender group, three metrics were calculated: total revenue, total profit, and the number of unique users. This aggregation provided insight into which gender segments contributed most significantly to the company's financial performance.

For age analysis, users were categorized into defined age brackets: <18, 18–24, 25–34, 35–44, 45–54, 55–64, and >65. Age groups were created using binning techniques to ensure consistent classification. Similar to the gender grouping, total revenue, total profit, and unique user counts were calculated for each age group.

To visualize the results, bar plots were generated showing total profit by gender and total profit by age group. These visualizations helped highlight spending and profitability patterns across different demographic profiles, supporting deeper interpretation of customer behavior trends.

Code:

```
# Group by gender
gender_profile = df.groupby('gender').agg(
    total_revenue=('revenue', 'sum'),
    total_profit=('profit', 'sum'),
    user_count=('user_id', 'nunique')
```

```
).reset_index()

# Define age brackets
bins = [0, 17, 24, 34, 44, 54, 64, 120]
labels = ['<18', '18-24', '25-34', '35-44', '45-54', '55-64', '65+']
df['age_group'] = pd.cut(df['age'], bins=bins, labels=labels)

# Group by age group
age_profile = df.groupby('age_group').agg(
    total_revenue=('revenue', 'sum'),
    total_profit=('profit', 'sum'),
    user_count=('user_id', 'nunique')
).reset_index()

# PLOT 1: Gender Profile (Barplot)
plt.figure(figsize=(10, 5))
sns.barplot(data=gender_profile, x='gender', y='total_profit',
palette='pastel')
plt.title('Total Profit by Gender')
plt.ylabel('Total Profit')
plt.xlabel('Gender')
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 6))
sns.barplot(data=age_profile, x='age_group', y='total_profit',
palette='Set2')
plt.title('Total Profit by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Total Profit')
plt.tight_layout()
plt.show()
```

Results and Interpretation:

Figure 8. Total Profit by Gender

Figure 9. Total Profit by Age Group

The results as shown in Figures 8 and 9 illustrate how profitability varies based on user attributes such as gender and age group. The analysis by gender (Figure 8) shows that both male and female users contribute significantly to total profit, with male users generating slightly

higher total profit than female users. Although the difference is relatively modest, it suggests that male customers tend to spend slightly more or purchase more profitable products compared to female customers.

When examining age groups (Figure 9), a more pronounced pattern emerges. Users aged 55–64 generated the highest total profit. This is closely followed by the 25–34, 35–44, and 45–54 brackets. These middle-aged to pre-retirement groups appear to be the most profitable customer segments for the company. In contrast, the <18 and >65 groups contributed the least to overall profit, with both segments showing similar profit levels. This suggests that younger and older users are less significant drivers of profitability compared to the prime working-age demographic (25–64 years old).

Overall, the results indicate that user attributes such as age have a stronger correlation with profitability than gender. Specifically, middle-aged customers represent the highest-value segment, while gender differences, though present, are less substantial.

Subquestion 4

"Which products generate the highest and lowest total sales volume, revenue and profit? (All-time)"

Methodology:

To determine which products generated the highest and lowest sales volume, revenue, and profit over all time, the transactional dataset was grouped at the product level, combining both product_id and name to ensure clarity between products with similar names or identifiers. For each product, three key metrics were calculated:

- Total Revenue: The sum of all revenues attributed to that product.
- Total Profit: The sum of all profits attributed to that product.
- Sales Volume: The total number of times the product was sold (measured by the count of order items).

Once these metrics were aggregated, the products were then sorted in both descending and ascending order for each metric. The top 10 and bottom 10 products were identified separately for total revenue, total profit, and sales volume to provide a comprehensive view of product performance across multiple dimensions.

Code:

```
# Group by product name
product_performance = df.groupby(['product_id', 'name']).agg(
    total_revenue=('revenue', 'sum'),
    total_profit=('profit', 'sum'),
```

```
    times_sold=('order_items_id', 'count'),  # frequency of sales'
).reset_index()

# Top 10 products by profit
top_10_profit = product_performance.sort_values(by='total_profit',
ascending=False).head(10)

# Top 10 products by revenue
top_10_revenue = product_performance.sort_values(by='total_revenue',
ascending=False).head(10)

# Top 10 products by sales volume
top_10_sales = product_performance.sort_values(by='times_sold',
ascending=False).head(10)

# Plot: Top 10 Products by Revenue
plt.figure(figsize=(12, 6))
sns.barplot(x='total_revenue', y='name', data=top_10_revenue,
palette='viridis')
plt.title('Top 10 Products by Total Revenue')
plt.xlabel('Total Revenue')
plt.ylabel('Product Name')
plt.tight_layout()
plt.show()

# Plot: Top 10 Products by Profit
plt.figure(figsize=(12, 6))
sns.barplot(x='total_profit', y='name', data=top_10_profit,
palette='viridis')
plt.title('Top 10 Products by Total Profit')
plt.xlabel('Total Profit')
plt.ylabel('Product Name')
plt.tight_layout()
plt.show()

# Plot: Top 10 Products by Sales

plt.figure(figsize=(12, 6))
sns.barplot(x='times_sold', y='name', data=top_10_sales,
palette='coolwarm')
```

```
plt.title('Top 10 Products by Total Sales')
plt.xlabel('Total Sales')
plt.ylabel('Product Name')
plt.tight_layout()
plt.show()
```

Results and Interpretation:

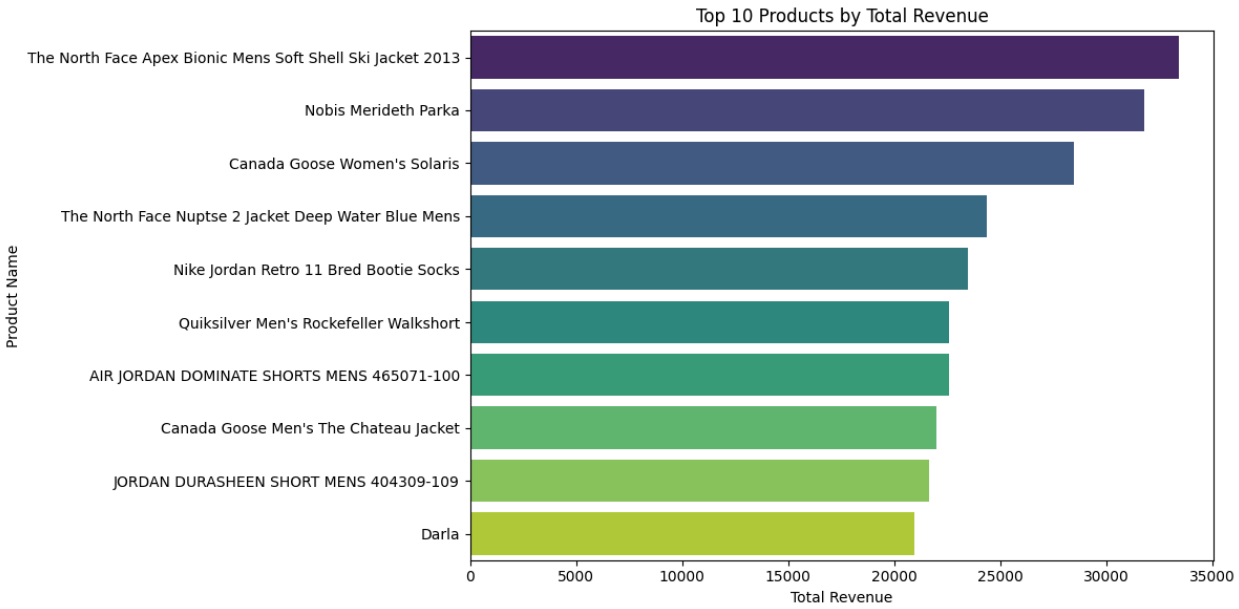

Figure 10. Top 10 Products by Total Revenue

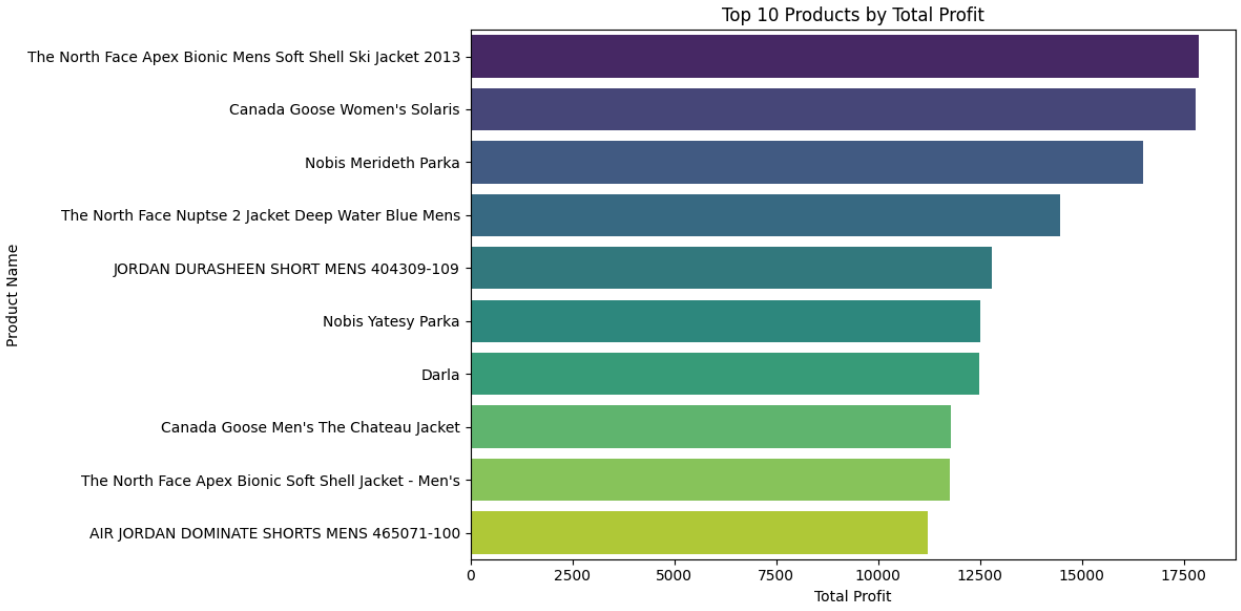

Figure 11. Top 10 Products by Total Profit

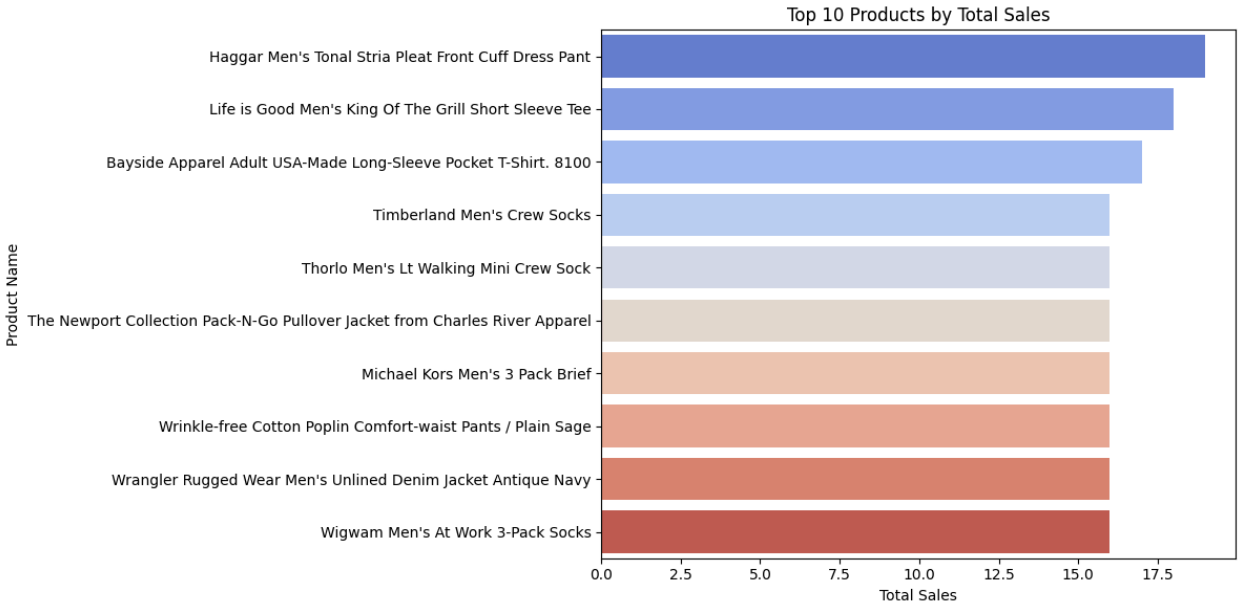

Figure 12. Top 10 Products by Total Sales

The results, as shown in Figures 10 to 12, illustrate how product performance differs across three key business metrics: total revenue, total profit, and total sales volume. Each chart highlights the top 10 products, offering insights into the most impactful items in the Looker E-Commerce catalog.

Figure 10 presents the top 10 products by total revenue. Leading the list is *The North Face Apex Bionic Mens Soft Shell Ski Jacket 2013*, generating the highest all-time revenue, followed closely by *Nobis Merideth Parka* and *Canada Goose Women's Solaris*. These products are characterized by high price points and consistent customer demand. The presence of multiple outerwear items and premium brands such as Canada Goose and The North Face suggests that winter apparel, particularly high-end jackets, contribute significantly to the company's revenue stream.

Figure 11 displays the top 10 products by total profit, and the pattern closely mirrors the revenue chart, with several overlaps in product names. *The North Face Apex Bionic Mens Soft Shell Ski Jacket 2013* again leads, indicating that its profitability is not solely based on sales volume, but also on a healthy profit margin. Notably, *Canada Goose Women's Solaris* and *Nobis Merideth Parka* also appear at the top, reaffirming their strong contribution to both revenue and profit. However, a few entries such as *Nobis Yatesy Parka* and *The North Face Apex Bionic Soft Shell Jacket - Men's* are more prominent here than in the revenue chart, suggesting they offer better margins relative to their sale volume.

In contrast, Figure 12 reveals the top 10 products by total sales volume, and the results diverge significantly from the previous two charts. Products like *Haggar Men's Tonal Stria Pleat*

Front Cuff Dress Pant, *Life is Good Men's King Of The Grill Short Sleeve Tee*, and *Bayside Apparel Adult USA-Made Long-Sleeve Pocket T-Shirt* dominate in terms of units sold, but these are generally lower-priced items. The complete absence of high-revenue and high-profit items from this chart underscores the importance of distinguishing between volume-driven popularity and value-driven profitability. In this case, while these items are widely purchased, they likely offer low margins and contribute modestly to overall revenue and profit.

Overall, the results suggest that while certain products achieve success through high transaction volume, others succeed through premium pricing and margin efficiency. Products that appear across all three charts—such as *The North Face Apex Bionic Mens Soft Shell Ski Jacket 2013* and *Canada Goose Women's Solaris*—should be considered strategic assets. These high-value items combine strong demand with high profitability and are critical to sustaining business growth. On the other hand, high-volume, low-margin products may still play an important role in attracting customers and generating traffic but should be managed with tighter cost controls to ensure operational efficiency.

Subquestion 5

"Which product categories generate the highest and lowest total sales volume, revenue and profit?"

Methodology:

To evaluate which product categories generate the highest and lowest total sales volume, revenue, and profit, the dataset was first grouped by the category field. This allowed the analysis to aggregate all product-level transactions under their respective category labels.
Three key performance metrics were calculated for each category:

- Total Revenue: The cumulative sales income generated by all products within the category.
- Total Profit: The total profit accumulated from all product sales under the category.
- Sales Volume: The total number of order items sold, indicating how frequently products from that category were purchased.

After computing these metrics, the dataset was sorted to identify the top 10 and bottom 10 categories for each metric. This enabled the detection of both high-performing and underperforming categories based on revenue, profitability, and purchase frequency. To enhance interpretability, the top 10 categories by total revenue were visualized using a horizontal bar plot. The x-axis represents the total revenue, while the y-axis lists the corresponding category names. The plot provides a clear visual comparison of revenue contributions across different product categories, supporting deeper insights into the commercial performance of each category within the business portfolio.

Code:

```
# Group by product name
category_performance = df.groupby('category').agg(
    total_revenue=('revenue', 'sum'),
    total_profit=('profit', 'sum'),
    times_sold=('order_items_id', 'count')  # frequency of sales
).reset_index()

# Top 10 categories by profit
top_10_profit = category_performance.sort_values(by='total_profit',
ascending=False).head(10)

# Bottom 10 categories by profit
bottom_10_profit = category_performance.sort_values(by='total_profit',
ascending=True).head(10)

# Top 10 categories by revenue
top_10_revenue = category_performance.sort_values(by='total_revenue',
ascending=False).head(10)

# Bottom 10 categories by revenue
bottom_10_revenue = category_performance.sort_values(by='total_revenue',
ascending=True).head(10)

# Top 10 categories by sales volume
top_10_sales = category_performance.sort_values(by='times_sold',
ascending=False).head(10)

# Bottom 10 categories by sales volume
bottom_10_sales = category_performance.sort_values(by='times_sold',
ascending=True).head(10)

# Plot: Top 10 Categories by Revenue
plt.figure(figsize=(12, 6))
sns.barplot(x='total_revenue', y='category', data=top_10_revenue,
palette='viridis')
plt.title('Top 10 Categories by Total Revenue')
plt.xlabel('Total Revenue (by millions)')
plt.ylabel('Category Name')
plt.tight_layout()
```

```
plt.show()

# Plot: Bottom 10 Categories by Revenue

plt.figure(figsize=(12, 6))
sns.barplot(x='total_revenue', y='category', data=bottom_10_revenue,
palette='viridis')
plt.title('Bottom 10 Categories by Total Revenue')
plt.xlabel('Total Revenue')
plt.ylabel('Category Name')
plt.tight_layout()
plt.show()

# Plot: Top 10 Categories by Profit

plt.figure(figsize=(12, 6))
sns.barplot(x='total_profit', y='category', data=top_10_profit,
palette='viridis')
plt.title('Top 10 Categories by Total Profit')
plt.xlabel('Total Profit')
plt.ylabel('Category Name')
plt.tight_layout()
plt.show()

# Plot: Bottom 10 Categories by Profit

plt.figure(figsize=(12, 6))
sns.barplot(x='total_profit', y='category', data=bottom_10_profit,
palette='viridis')
plt.title('Bottom 10 Categories by Total Profit')
plt.xlabel('Total Profit')
plt.ylabel('Category Name')
plt.tight_layout()
plt.show()
# Plot: Top 10 Categories by Sales

plt.figure(figsize=(12, 6))
sns.barplot(x='times_sold', y='category', data=top_10_sales,
palette='viridis')
plt.title('Top 10 Categories by Total Sales')
```

```
plt.xlabel('Total Sales')
plt.ylabel('Category Name')
plt.tight_layout()
plt.show()

# Plot: Bottom 10 Categories by Sales

plt.figure(figsize=(12, 6))
sns.barplot(x='times_sold', y='category', data=bottom_10_sales,
palette='viridis')
plt.title('Bottom 10 Categories by Total Sales')
plt.xlabel('Total Sales')
plt.ylabel('Category Name')
plt.tight_layout()
plt.show()
```

Results and Interpretation:

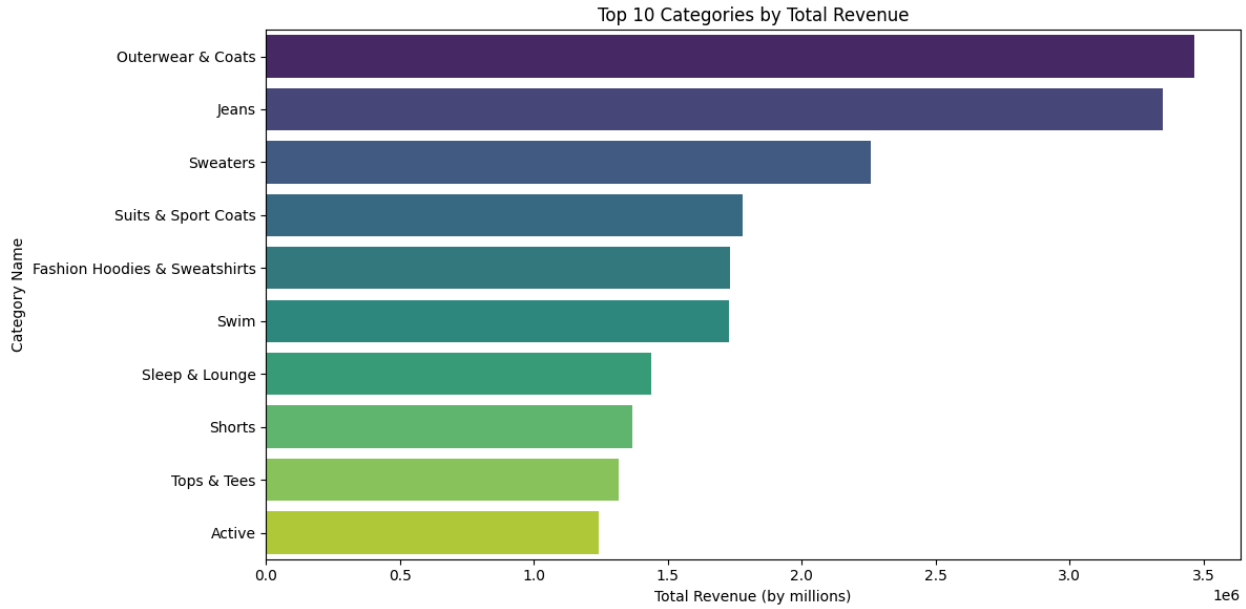

Figure 13. Top 10 Categories by Total Revenue

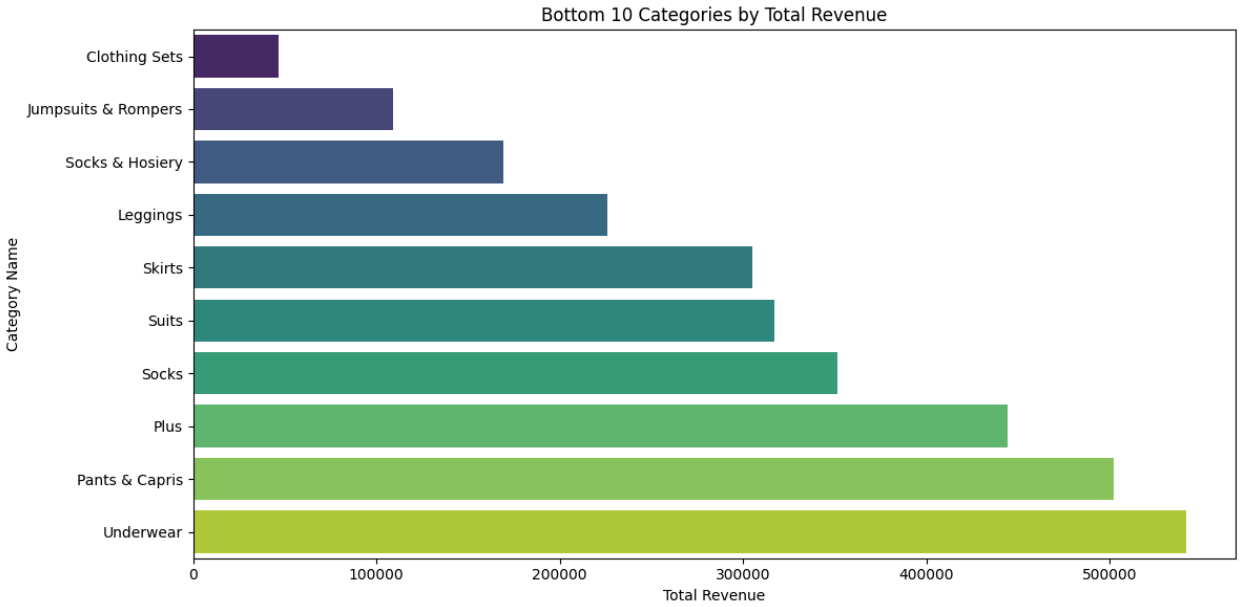

Figure 14. Bottom 10 Categories by Total Revenue

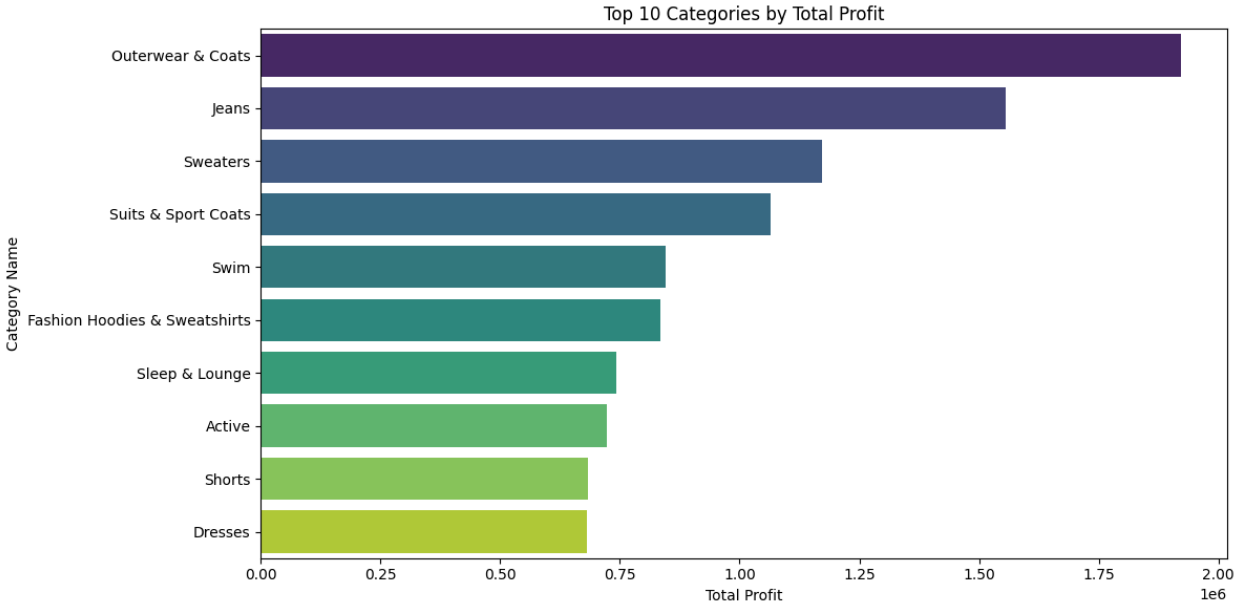

Figure 15. Top 10 Categories by Total Profit

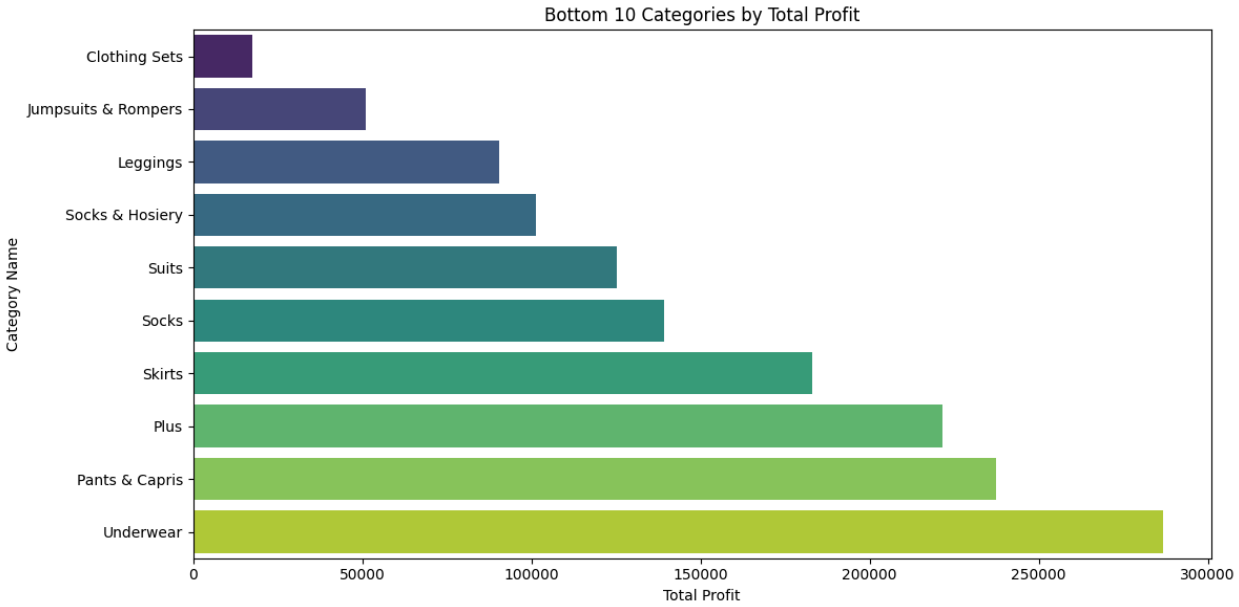

Figure 16. Bottom 10 Categories by Total Profit

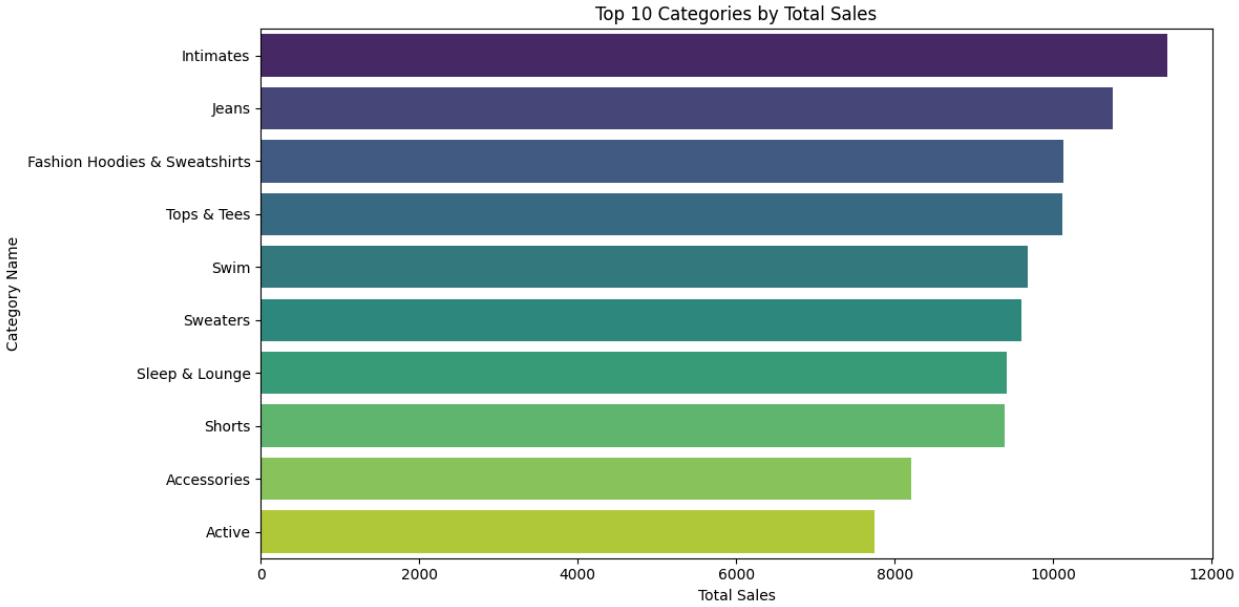

Figure 17. Top 10 Categories by Total Sales

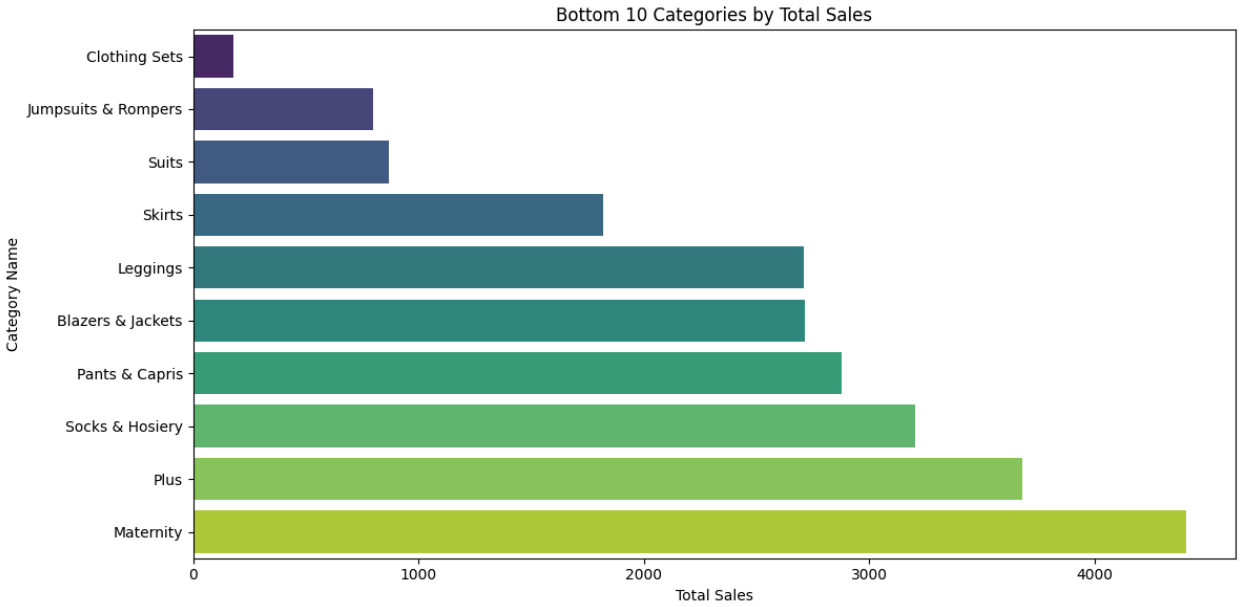

Figure 18. Bottom 10 Categories by Total Sales

The results, as shown in Figures 13 to 18, illustrate how product category performance varies significantly across total revenue, total profit, and total sales volume. These dimensions provide a comprehensive view of which categories are driving business value for Looker E-Commerce.

Figures 13 and 14 highlight the top and bottom 10 product categories by total revenue. The results show that Outerwear & Coats leads by a considerable margin, followed by Jeans, Sweaters, and Suits & Sport Coats. These categories represent high-ticket items that, while not always the most frequently sold, contribute significantly to total earnings due to their premium pricing. On the other end, categories such as Clothing Sets, Jumpsuits & Rompers, and Socks & Hosiery generate the least revenue, indicating limited market demand or lower average selling prices. This distribution suggests that revenue is concentrated in fewer, higher-value categories, while a long tail of lower-revenue categories may be underperforming.

Figures 15 and 16 display the top and bottom 10 categories by total profit, revealing a pattern that closely resembles the revenue breakdown. Outerwear & Coats, Jeans, and Sweaters again dominate the top-performing list, reinforcing that their high revenue is also matched by strong margins. Interestingly, categories like Dresses and Shorts appear in the top 10 profit chart despite not being in the top 10 for revenue, suggesting they offer better profit margins relative to their selling price. Meanwhile, the least profitable categories, including Clothing Sets, Jumpsuits & Rompers, and Leggings, which align with their poor revenue performance, making them prime candidates for reassessment or optimization.

Figures 17 and 18 focus on sales volume across categories. Here, a different trend emerges. The highest-selling categories include Intimates, Jeans, Fashion Hoodies & Sweatshirts, and Tops & Tees. These categories are characterized by high-frequency purchases, often driven by affordability and broad appeal. In contrast, the lowest-selling categories, such as Clothing Sets, Jumpsuits & Rompers, Suits, and Maternity, show limited demand. Notably, some high-volume categories like Intimates and Tops & Tees do not appear in the top revenue or profit rankings, suggesting that while they drive customer transactions, their lower pricing or margins limit their financial impact.

The results demonstrate that high-performing categories vary depending on the metric used. Outerwear & Coats and Jeans are consistently strong across all three measures—revenue, profit, and volume—making them strategic pillars of the product portfolio. On the other hand, categories like Clothing Sets, Jumpsuits & Rompers, and Suits consistently underperform, both in terms of financial returns and customer demand. These insights can guide targeted business strategies: investing in top-performing categories for growth, reevaluating the pricing and positioning of mid-tier categories, and potentially phasing out the persistently low-performing ones to optimize inventory and operational efficiency.

Subquestion 6

"Which product brands generate the highest and lowest total sales volume, revenue and profit?"

Methodology:

To evaluate which product brands generate the highest and lowest total sales volume, revenue, and profit, the dataset was aggregated by the brand attribute. This grouping enabled the analysis to capture the performance of each brand as a whole, regardless of the number of products under it.

Three key business metrics were computed for each brand:

- Total Revenue: The sum of all revenue generated by the brand's products.
- Total Profit: The cumulative profit derived from all sales attributed to that brand.
- Sales Volume: The number of individual items sold under the brand, measured by the total count of order items.

After calculating these metrics, the results were sorted in both descending and ascending order for each performance measure. This allowed for the identification of the top 10 and bottom 10 brands based on revenue, profit, and sales volume. This brand-level analysis helps to identify which brands are consistently driving profitability and sales for the business, and which may be underperforming despite product availability. It also lays the foundation for strategic decisions around brand promotion, inventory allocation, partnership renewal, or possible deactivation of underperforming suppliers.

Code:

```
# Group by product name
brand_performance = df.groupby('brand').agg(
    total_revenue=('revenue', 'sum'),
    total_profit=('profit', 'sum'),
    times_sold=('order_items_id', 'count')  # frequency of sales
).reset_index()

# Top 10 categories by profit
top_10_profit = brand_performance.sort_values(by='total_profit',
ascending=False).head(10)

# Bottom 10 categories by profit
bottom_10_profit = brand_performance.sort_values(by='total_profit',
ascending=True).head(10)

# Top 10 categories by revenue
top_10_revenue = brand_performance.sort_values(by='total_revenue',
ascending=False).head(10)

# Bottom 10 categories by revenue
bottom_10_revenue = brand_performance.sort_values(by='total_revenue',
ascending=True).head(10)

# Top 10 categories by sales volume
top_10_sales = brand_performance.sort_values(by='times_sold',
ascending=False).head(10)

# Bottom 10 categories by sales volume
bottom_10_sales = brand_performance.sort_values(by='times_sold',
ascending=True).head(10)

# Plot: Top 10 Brands by Revenue

plt.figure(figsize=(12, 6))
sns.barplot(x='total_revenue', y='brand', data=top_10_revenue,
palette='viridis')
plt.title('Top 10 Brands by Total Revenue')
plt.xlabel('Total Revenue')
plt.ylabel('Brand Name')
```

```
plt.tight_layout()
plt.show()

# Plot: Bottom 10 Brands by Revenue

plt.figure(figsize=(12, 6))
sns.barplot(x='total_revenue', y='brand', data=bottom_10_revenue,
palette='viridis')
plt.title('Bottom 10 Brands by Total Revenue')
plt.xlabel('Total Revenue')
plt.ylabel('Brand Name')
plt.tight_layout()
plt.show()

# Plot: Top 10 Brands by Profit

plt.figure(figsize=(12, 6))
sns.barplot(x='total_profit', y='brand', data=top_10_profit,
palette='viridis')
plt.title('Top 10 Brands by Total Profit')
plt.xlabel('Total Profit')
plt.ylabel('Brand Name')
plt.tight_layout()
plt.show()

# Plot: Bottom 10 Brands by Profit

plt.figure(figsize=(12, 6))
sns.barplot(x='total_profit', y='brand', data=bottom_10_profit,
palette='viridis')
plt.title('Bottom 10 Brands by Total Profit')
plt.xlabel('Total Profit')
plt.ylabel('Brands Name')
plt.tight_layout()
plt.show()

# Plot: Top 10 Brands by Sales

plt.figure(figsize=(12, 6))
```

```
sns.barplot(x='times_sold', y='brand', data=top_10_sales,
palette='viridis')
plt.title('Top 10 Brands by Total Sales')
plt.xlabel('Total Sales')
plt.ylabel('Brand Name')
plt.tight_layout()
plt.show()

# Plot: Bottom 10 Brands by Sales

plt.figure(figsize=(12, 6))
sns.barplot(x='times_sold', y='brand', data=bottom_10_sales,
palette='viridis')
plt.title('Bottom 10 Brands by Total Sales')
plt.xlabel('Total Sales')
plt.ylabel('Brand Name')
plt.tight_layout()
plt.show()
```

Results and Interpretation:

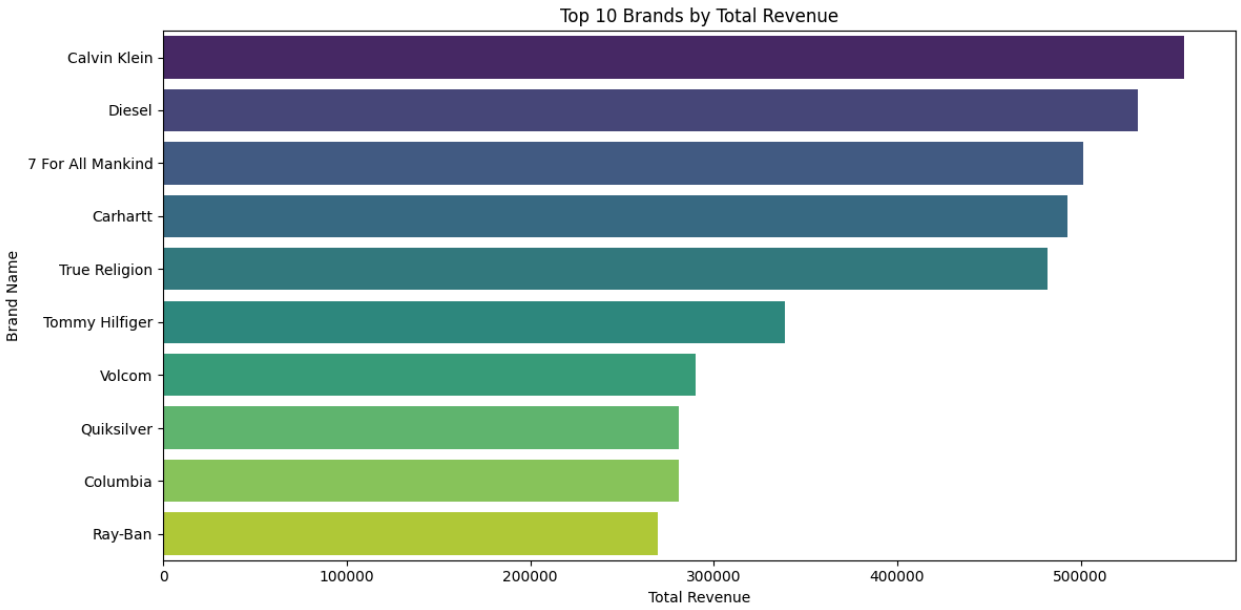

Figure 19. Top 10 Brands by Total Revenue

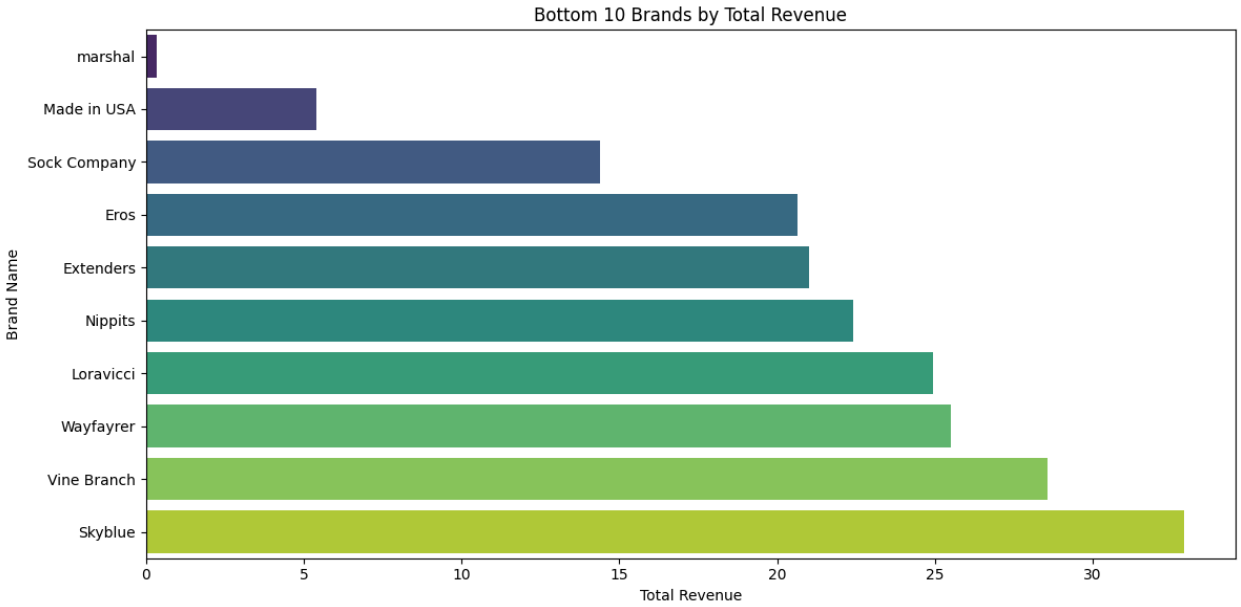

Figure 20. Bottom 10 Brands by Total Revenue

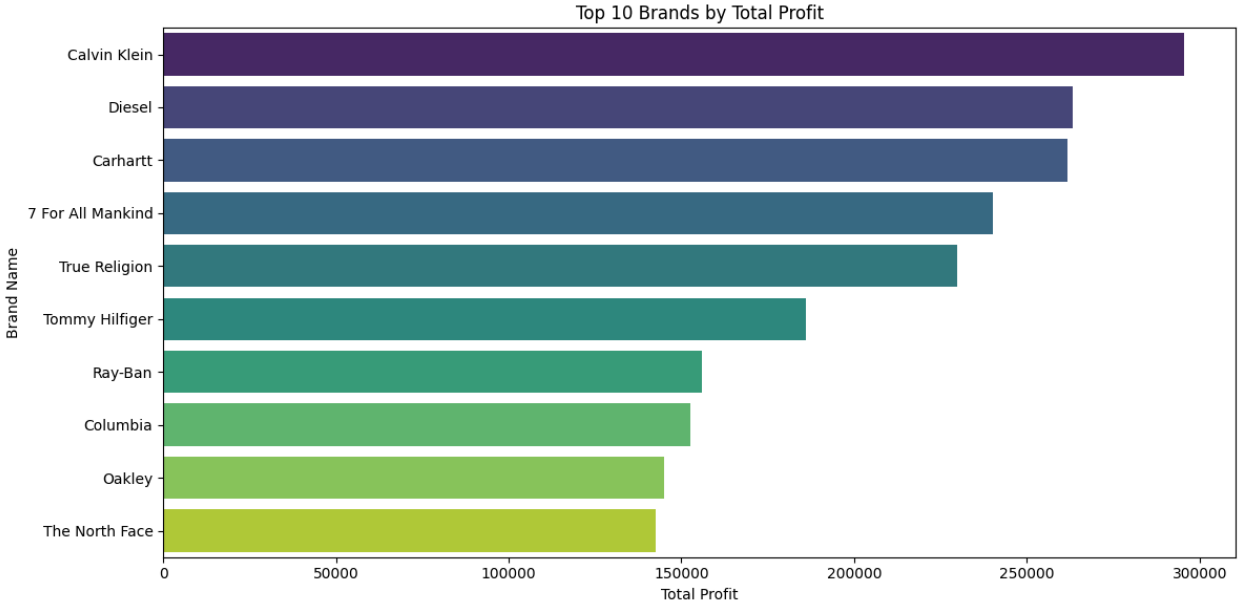

Figure 21. Top 10 Brands by Total Profit

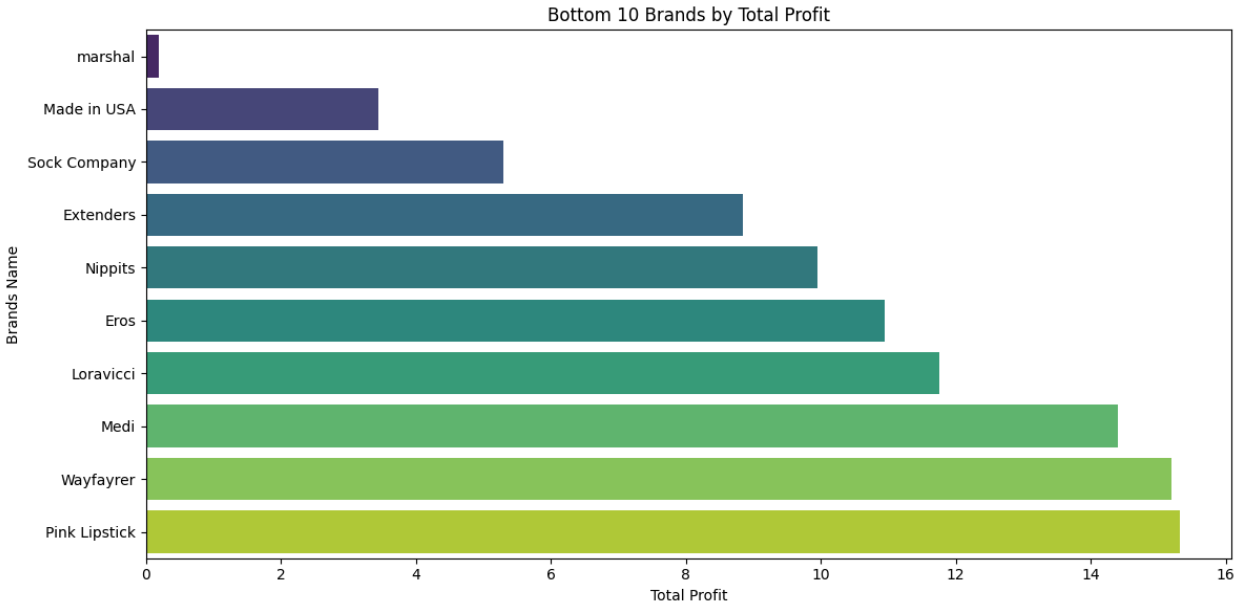

Figure 22. Bottom 10 Brands by Total Profit

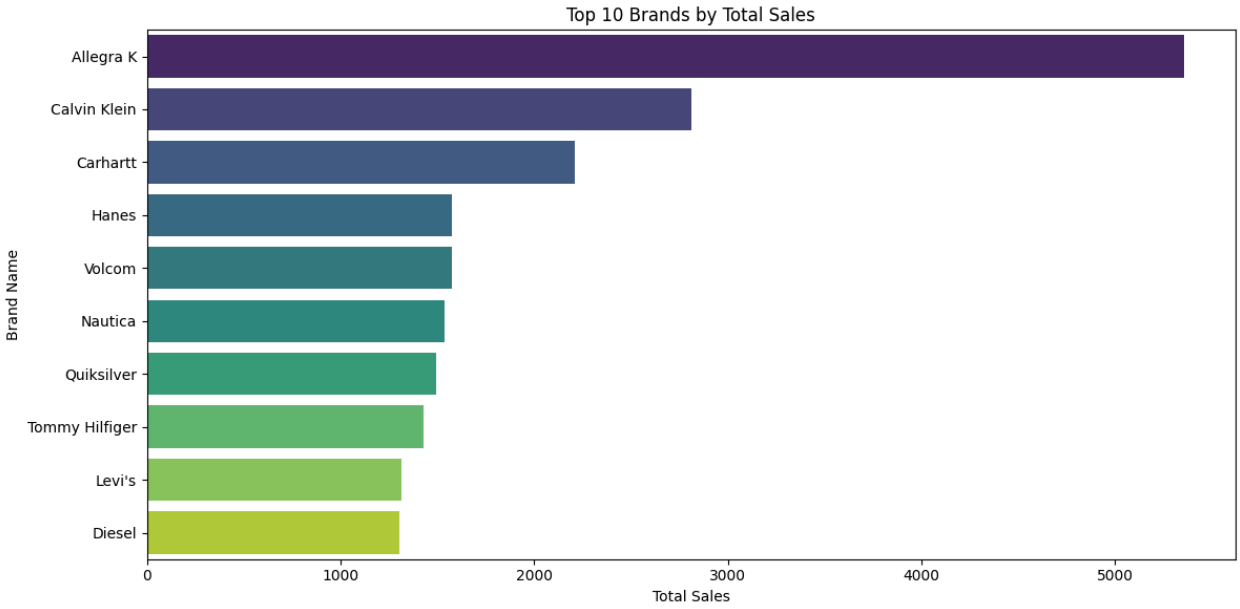

Figure 23. Top 10 Brands by Total Sales

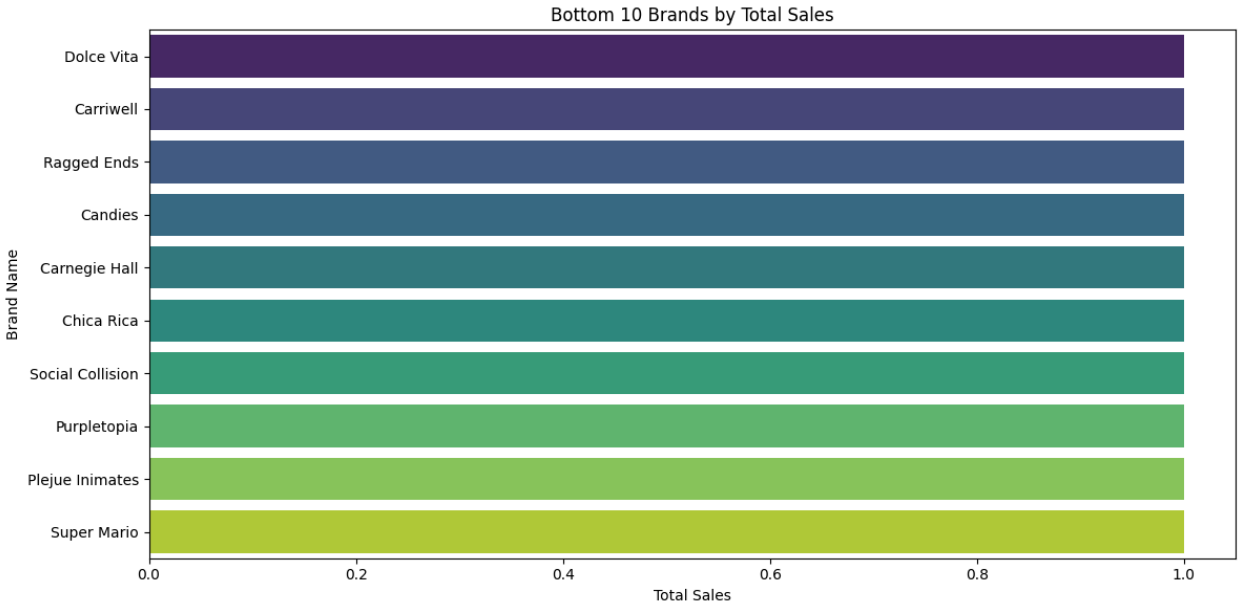

Figure 24. Bottom 10 Brands by Total Sales

The results, as shown in Figures 19 to 24, highlight significant variations in brand performance when measured by total revenue, profit, and sales volume.

Figures 19 and 20 present the top and bottom 10 brands by total revenue. The brand Calvin Klein leads with the highest revenue, followed closely by Diesel, 7 For All Mankind, and Carhartt. These brands are associated with premium product lines and strong market visibility, which likely contribute to their revenue dominance. On the other end of the spectrum, brands such as marshal, Made in USA, and Sock Company posted the lowest revenue figures, indicating limited sales activity or a smaller presence in the product catalog.

Figures 21 and 22 reveal a similar pattern in terms of total profit. Once again, Calvin Klein ranks first, followed by Diesel, Carhartt, and 7 For All Mankind, reflecting both high unit prices and strong profit margins. Interestingly, while Ray-Ban and The North Face also appear in the top 10 profit chart, they are not present in the top 10 for revenue, suggesting highly profitable pricing strategies. Meanwhile, the brands with the lowest profit—marshal, Made in USA, and Sock Company—mirror those in the bottom revenue chart, emphasizing consistently poor financial performance across both metrics.

Figures 23 and 24 illustrate the top and bottom 10 brands by total sales volume. Here, a notable shift occurs. Allegra K dominates by a wide margin in terms of units sold, far exceeding other brands like Calvin Klein, Carhartt, Hanes, and Volcom. This indicates that Allegra K offers products with broad appeal or lower price points, driving high volume but potentially lower margins. In contrast, brands such as Dolce Vita, Carriwell, and Social Collision recorded the lowest number of total sales, suggesting weak demand or niche offerings.

These findings underscore the importance of analyzing brand performance across multiple dimensions. High-revenue and high-profit brands like Calvin Klein, Diesel, and Carhartt are strategic to the company's financial success and should be prioritized for marketing, inventory investment, and long-term partnership development. Conversely, brands that consistently appear in the bottom 10 across all metrics—such as marshal and Sock Company—may need to be reevaluated for delisting, renegotiation, or repositioning. Moreover, the contrast between high-volume brands like Allegra K and high-margin brands like Ray-Ban highlights differing business models—volume-driven vs. margin-driven success. A balanced portfolio that considers both strategies can help Looker E-Commerce optimize profitability while maintaining strong market coverage.

Subquestion 7

"What is the sold-stocked ratio (sold versus stocked) per product?"

Methodology:

The sell-through rate is a metric that measures how efficiently inventory is converted into sales. It is calculated as the percentage of units sold relative to the number of units stocked for a given product. To determine the sell-through rate for each product, the analysis began by identifying how many unique inventory items were initially stocked per product. This was calculated by counting distinct inventory item IDs associated with each product. To determine the number of items sold, the dataset was filtered to include only records where inventory items were linked to an actual order, and the number of sales per product was then aggregated. These two datasets (stocked quantities and sold quantities) were merged by product ID to enable direct comparison. Any product with no recorded sales was assigned a sold quantity of zero. The sell-through rate was then computed as the percentage of sold items relative to total stocked items for each product. To visualize the distribution of sell-through rates across the product catalog, a histogram was generated, offering a clear overview of which products perform well in terms of turnover and which may be underperforming.

Code:

```
# Quantity Stocked = unique inventory items per product
stocked = df[['product_id', 'inventory_item_id']].drop_duplicates()
stocked_count = stocked.groupby('product_id').agg(
    quantity_stocked=('inventory_item_id', 'count')
).reset_index()

# Quantity Sold = only rows where inventory_item_id is linked to an order
sold = df[~df['order_items_id'].isna()]
```

```
sold_count = sold.groupby('product_id').agg(
    quantity_sold=('order_items_id', 'count')
).reset_index()

# Merge
sell_through = pd.merge(stocked_count, sold_count, on='product_id',
how='left')

# Fill unsold with 0
sell_through['quantity_sold'] = sell_through['quantity_sold'].fillna(0)

# Compute sell-through rate
sell_through['sell_through_rate'] = (sell_through['quantity_sold'] /
sell_through['quantity_stocked']) * 100
# Step 3: Histogram
plt.figure(figsize=(10, 5))
sns.histplot(sell_through['sell_through_rate'], bins=40, kde=True,
color='skyblue')
plt.title('Histogram of Sell-Through Rates per Product')
plt.xlabel('Sell-Through Rate (%)')
plt.ylabel('Number of Products')
plt.tight_layout()
plt.show()
```

Results and Interpretation:

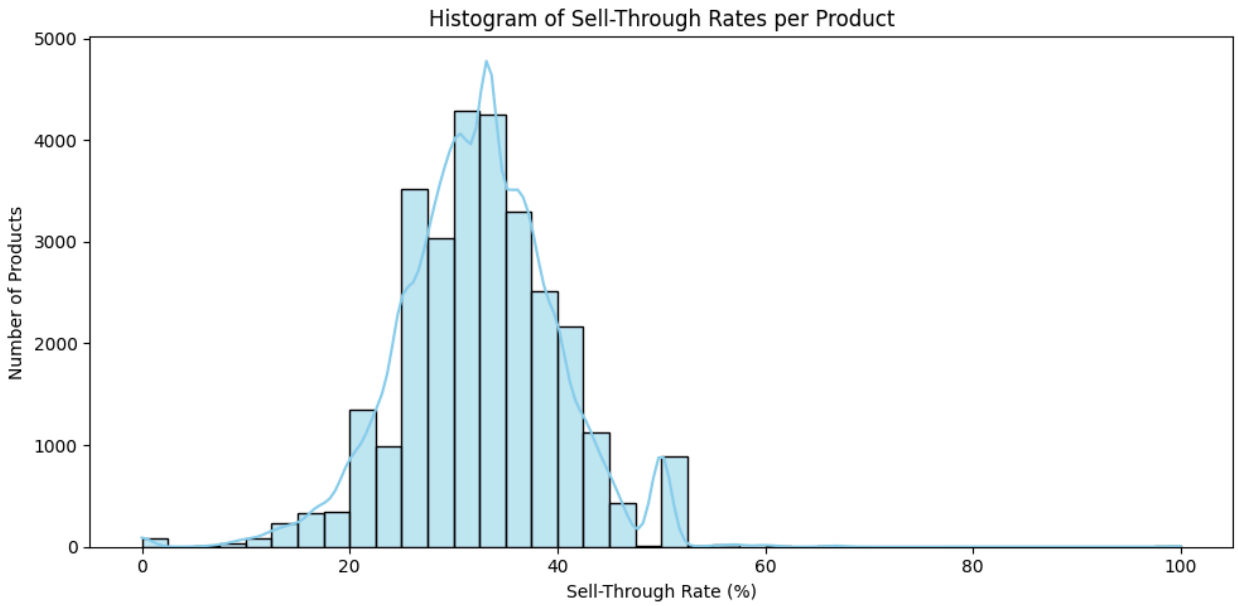

Figure 25. Histogram of Sell-Through Rates per Product

The histogram in Figure 25 illustrates the distribution of sell-through rates across all products in the dataset. The distribution is unimodal and slightly skewed to the right, with the majority of products exhibiting sell-through rates between 20% and 40%. The peak occurs around the 30–35% mark, indicating that this is the most common sell-through range. Very few products achieved sell-through rates above 60%, and near-zero sales performance is rare but present, suggesting minimal instances of products that remained almost entirely unsold.

These results suggest that while most products do convert a portion of their stock into sales, full inventory turnover is uncommon. The relatively moderate sell-through rates imply that Looker E-Commerce may be overstocking certain items or lacking effective demand forecasting for a large portion of its catalog. Identifying and focusing on products with consistently high sell-through rates could help the business optimize inventory allocation, reduce holding costs, and improve overall profitability.

Subquestion 8

"Which products are overstocked or have low conversion despite high inventory?"

Methodology:

To identify products that are overstocked or have low conversion rates despite high inventory levels, the analysis segmented products into sell-through rate (STR) buckets. The STR was divided into four categories: 0–25%, 25–50%, 50–75%, and 75–100%. Using binning techniques, each product was assigned to a bucket based on its STR value. This allowed for a structured categorization of product performance in terms of inventory turnover.

After classification, the number of products in each bucket was tallied to assess how many products fell within each conversion range. These counts were displayed in a summary table and visualized using a bar chart to clearly highlight which ranges were most common. The focus was on identifying products in the lowest bucket (0–25%), which represent items that may be overstocked or underperforming in sales relative to their inventory levels. This approach helps isolate potential inefficiencies in inventory management and supports targeted intervention strategies for under-converting products.

Code:

```
# Define STR buckets
bins = [0, 25, 50, 75, 100]
labels = ['0-25%', '25-50%', '50-75%', '75-100%']

sell_through['conversion_bucket'] =
pd.cut(sell_through['sell_through_rate'], bins=bins, labels=labels,
include_lowest=True)

# Count products in each bucket
bucket_counts =
sell_through['conversion_bucket'].value_counts().sort_index().reset_index(
)
bucket_counts.columns = ['Sell-Through Rate Bucket', 'Number of Products']


# Step 3: Print the table
print(bucket_counts.to_string(index=False))

# Bar Chart
plt.figure(figsize=(8, 5))
sns.barplot(x='Sell-Through Rate Bucket', y='Number of Products',
data=bucket_counts, palette='Spectral')
plt.title('Distribution of Products by Sell-Through Rate Buckets')
plt.xlabel('Sell-Through Rate Bucket')
plt.ylabel('Number of Products')
plt.tight_layout()
plt.show()
```

Results and Interpretation:

Table 2. Number of Products per STR Bin

STR Bucket	Number of Products
0-25%	5,039
25-50%	23,947
50-75%	55
75-100%	5

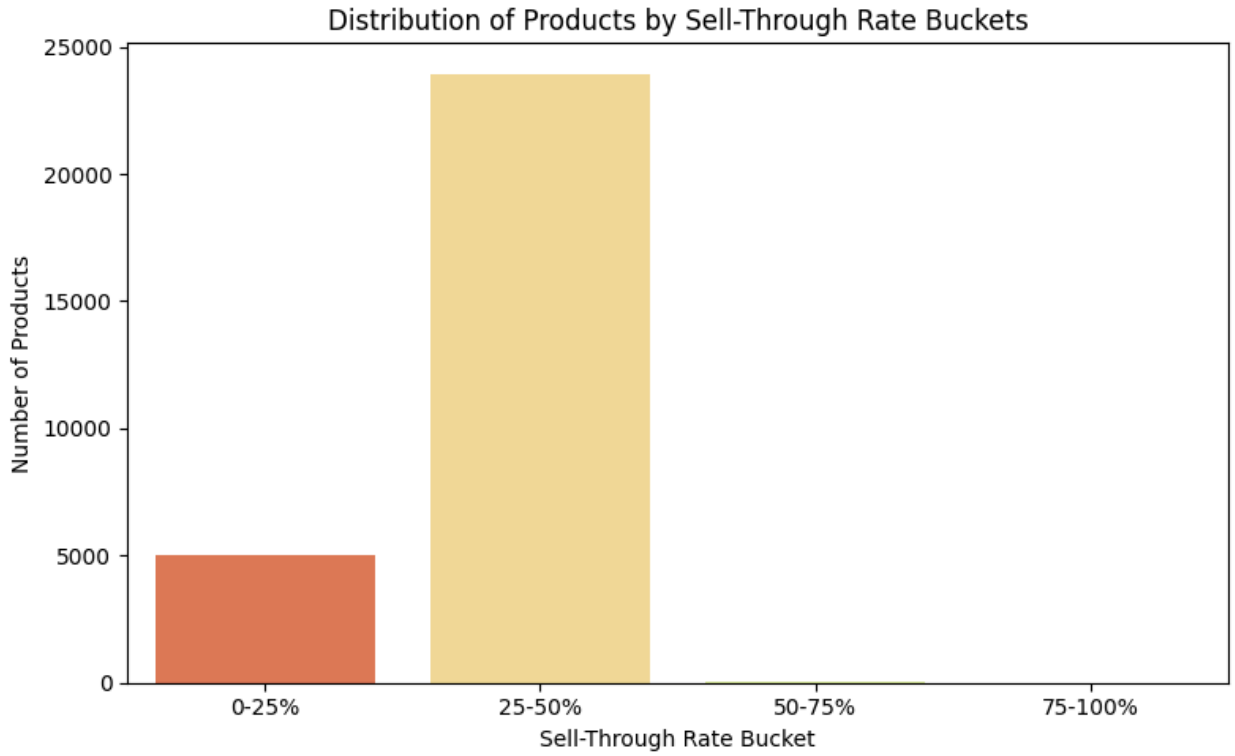

Figure 26. Distribution of Products by STR

Figure 26 and Table 2 present the distribution of products based on their STR, categorized into four buckets: 0–25%, 25–50%, 50–75%, and 75–100%. The majority of products (23,947) fall within the 25–50% range, indicating moderate conversion but not optimal inventory turnover. More critically, a substantial 5,039 products fall within the 0–25% bucket, signaling low-conversion or overstocked items that are underperforming despite being available in inventory.

On the other hand, only 55 products reached a sell-through rate of 50–75%, and an extremely small subset of just 5 products achieved 75–100%. This highlights how rare high inventory efficiency is across the catalog. This skewed distribution implies that most products are not selling at a rate that justifies current stock levels.

The results point to a clear opportunity for operational improvement. Products in the lowest STR bucket should be closely evaluated for potential markdowns, bundling, or discontinuation, while top-performing items could guide future stocking and marketing strategies. Overall, the data underscores the need for more responsive inventory planning to minimize holding costs and improve sales efficiency.

Subquestion 9

"Are there products with zero or near-zero sales that should be deprioritized?"

Methodology:

To identify products that may warrant deprioritization due to poor sales performance, the analysis focused on detecting two specific groups: products with zero sales and those with close to zero sell-through rates. First, products with zero sales were identified by counting all entries where the quantity sold was equal to zero. These represent items that were stocked but never purchased. After identifying products with zero sales, the analysis filtered for products with a STR of less than 10%, a threshold for poor conversion. This was done while ensuring that these products had at least 10 units stocked to eliminate noise from low-inventory edge cases. The counts from both segments were then summed to determine the total number of underperforming products that may be strong candidates for deprioritization. This approach provides a focused and data-driven method for flagging items that consume inventory space and capital without generating meaningful revenue.

Code:

```
# Count of zero-sale products
zero_sales_count = (sell_through['quantity_sold'] == 0).sum()

# Count of near-zero sell-through products (<10%) with at least 10 stocked
near_zero_sales_count = sell_through[
    (sell_through['sell_through_rate'] < 10) &
    (sell_through['quantity_sold'] > 0) &
    (sell_through['quantity_stocked'] >= 10)
].shape[0]

# Total unique deprioritized products
total_deprioritized = zero_sales_count + near_zero_sales_count
```

```
# Print results
print(f"Products with zero sales: {zero_sales_count}")
print(f"Products with near-zero sales (<10% STR):
{near_zero_sales_count}")
print(f"Total products to consider for deprioritization:
{total_deprioritized}")
```

Results and Interpretation:

```
Products with zero sales: 87
Products with near-zero sales (<10% STR): 42
Total products to consider for deprioritization: 129
```

Figure 27. Results Showing Number of Products Warranting Deprioritization

The results shown in Figure 27 identified a total of 129 products that may be strong candidates for deprioritization based on poor sales performance. Specifically, 87 products recorded zero sales, indicating that they were stocked but never purchased by any customer. Additionally, 42 products had near-zero STRs, which was defined as less than 10% of their inventory sold. These products represent inventory that is taking up resources without generating any returns.

The presence of both unsold and severely underperforming products highlights potential inefficiencies in inventory planning, product-market fit, or visibility. Continuing to allocate shelf space, marketing effort, or capital to these items may not be justifiable. These findings suggest a clear opportunity to deprioritize or phase out such products in favor of better-performing items, allowing the company to reduce holding costs and improve inventory turnover.

K-Means Clustering Model

“Can products be segmented into distinct clusters based on sales performance, profitability, and inventory levels to reveal meaningful product profiles?”

Methodology:

To explore whether products can be segmented into meaningful groups based on performance and inventory characteristics, K-Means clustering was applied to create product profiles. The analysis began by aggregating key metrics for each product. The metrics were total_stocked, total_sold, and total_profit. These features represent inventory availability, sales volume, and financial performance.

Before clustering, the selected features were standardized using StandardScaler to ensure that variables on different scales did not disproportionately influence the clustering algorithm.

K-Means clustering was then performed to group products based on their similarity across these attributes. To determine the optimal number of clusters, the elbow method was used by plotting the within-cluster sum of squares (inertia) across a range of cluster values, helping to identify the point where additional clusters no longer significantly improve the model.

The final model assigned each product to a specific cluster, allowing for post-clustering interpretation. Cluster characteristics were examined by comparing average stocked quantity, sold quantity, and profitability within each group. This enabled the identification of product profiles such as high-profit/high-sales items, overstocked low-sellers, or low-volume but high-margin products. The resulting clusters offer actionable insights for tailoring inventory, pricing, and promotional strategies.

Code:

```
# Aggregate by product_id and name (if available)
product_summary = df.groupby(['product_id', 'name']).agg(
    total_stocked=('inventory_item_id', 'count'),
    total_sold=('order_items_id', 'count'),
    total_profit=('profit', 'sum'),
    total_revenue=('revenue', 'sum')
).reset_index()
product_summary.columns
# Compute correlation matrix
corr_matrix = product_summary[['total_stocked', 'total_sold',
'total_profit', 'total_revenue']].corr()

# Plot heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f',
linewidths=0.5)
plt.title("Feature Correlation Matrix")
plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.tight_layout()
plt.show()
# drop total revenue (highly correlated with total profit)
product_summary = product_summary.drop(columns=['total_revenue'])
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.preprocessing import StandardScaler
from itertools import combinations
```

```
# Step 2: Prepare features
features = ['total_sold', 'total_profit', 'total_stocked']
X = product_summary[features]
X_scaled = StandardScaler().fit_transform(X)

# Step 3: Try KMeans for k = 2 to 7
results = []

for k in range(2, 9):
    try:
        kmeans = KMeans(n_clusters=k, random_state=42)
        labels = kmeans.fit_predict(X_scaled)
        score = silhouette_score(X_scaled, labels)
        results.append({
            'k': k,
            'score': score
        })
    except Exception as e:
        print(f"Skipping k={k} due to error: {e}")

# Step 4: Compile and sort results
results_df = pd.DataFrame(results)
results_df                =                results_df.sort_values(by='score',
ascending=False).reset_index(drop=True)

# Step 5: Print sorted k results
print("🔝 Top k values based on silhouette score:\n")
for i, row in results_df.iterrows():
    print(f"Rank {i+1} | k = {row['k']} | Score = {row['score']:.4f}")

# Fit KMeans with k=3
kmeans = KMeans(n_clusters=3, random_state=42)
product_summary['cluster'] = kmeans.fit_predict(X_scaled)

# View cluster centers
# Define scaler
scaler = StandardScaler()
scaler.fit(X) # fit the scaler to your original data
```

```
cluster_centers =
pd.DataFrame(scaler.inverse_transform(kmeans.cluster_centers_),
columns=features)
print("Cluster Centers:\n", cluster_centers)
```

Results and Interpretation:

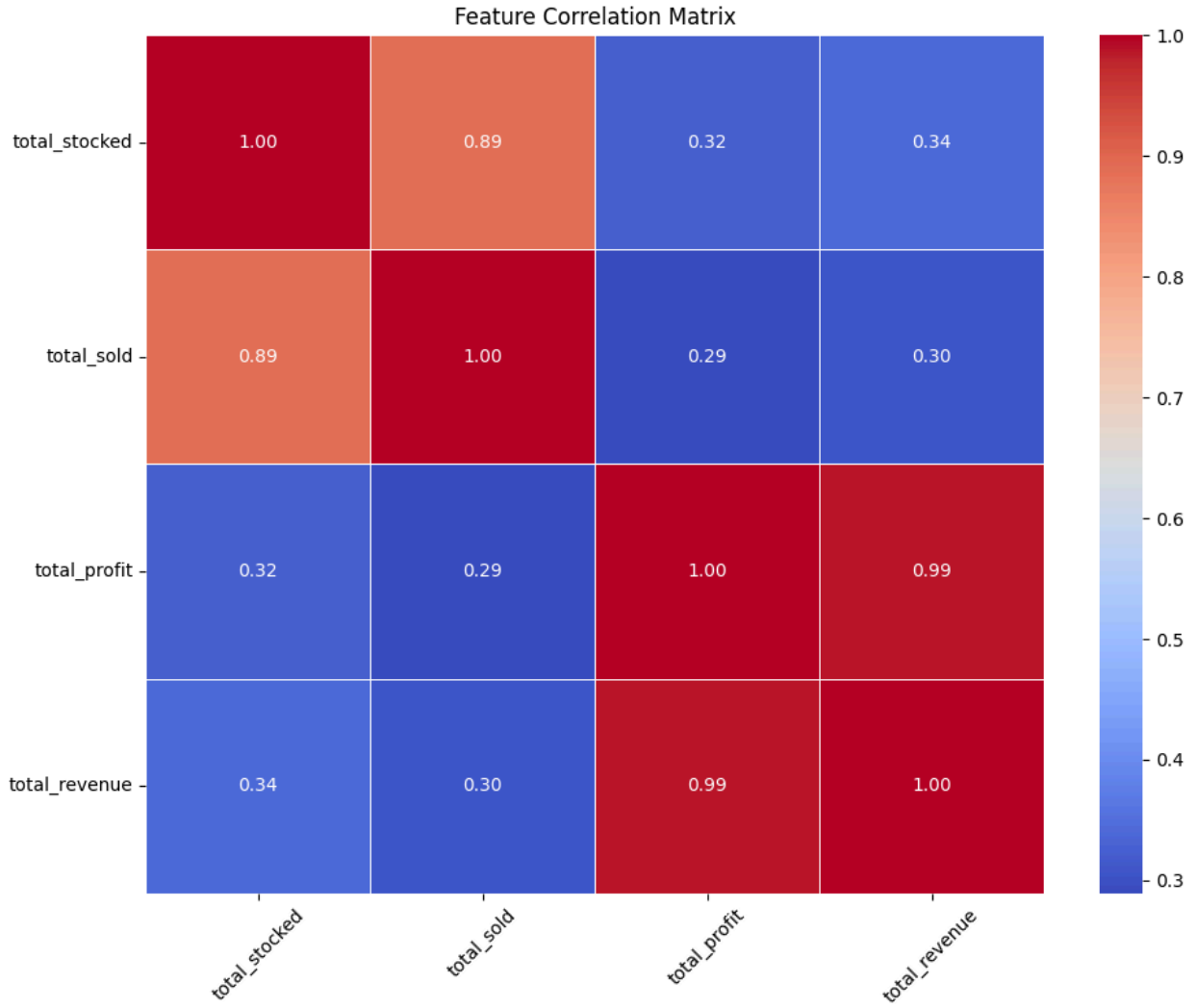

Figure 27. Correlation Matrix of Aggregate Features

The correlation matrix seen in Figure 27 shows that while total_sold and total_stocked are strongly correlated (r = 0.89), both are only weakly correlated with total_profit and total_revenue (ranging from 0.29 to 0.34). This suggests that sales volume and inventory levels are closely linked, but do not always translate proportionally into profitability. This gives us an insight that supports the need for clustering to uncover hidden product patterns.

Table 3. Top k Values Based on Silhouette Scores

Rank	k	Score
1	3	0.4417
2	2	0.4386
3	4	0.4324
4	5	0.3878
5	6	0.3384
6	7	0.3084
7	8	0.2741

K-Means clustering was applied to standardized values of total_sold, total_stocked, and total_profit. Based on the silhouette scores (Table 3), the optimal number of clusters was found to be k = 3 (score = 0.4417), indicating a well-defined segmentation structure. The three resulting clusters each exhibit distinct product profiles:

- Cluster 0
    - Moderate sales and stock levels with moderate profitability (≈ $551 profit, 7.4 units sold)
- Cluster 1
    - Low stock and sales levels with the lowest profitability (≈ $319 profit, 3.7 units sold)
- Cluster 2
    - High sales and stock levels with significantly higher profitability (≈ $3,082 profit, 7.4 units sold)

This segmentation confirms that products can be meaningfully grouped based on performance metrics. In particular, Cluster 2 stands out as a high-performing segment, while Cluster 1 includes underperforming products that may require intervention such as pricing adjustments, inventory cuts, or marketing support. These insights enable more informed inventory and product life cycle strategies tailored to each cluster's behavior.

Summary of Findings

The analysis of the Looker E-Commerce dataset reveals key insights that collectively address how the business can maximize profit and revenue through improved product performance and strategic targeting. First, profitability has shown a strong upward trend from

2019 to 2024, particularly accelerating in the last two years. This indicates that recent strategic shifts may be yielding scalable growth. Country-level analysis revealed that profitability is heavily concentrated in a few markets, with China and the United States dominating across profit, revenue, and sales, while countries like Austria and Colombia contribute minimally. User demographic analysis showed that middle-aged users (ages 25–64) are the most profitable segment, while gender has a more modest influence on spending behavior. Product-level analysis identified key high-revenue and high-profit items, mostly premium outerwear, as well as high-volume, low-margin products that serve different strategic purposes. Similarly, certain product categories and brands, like Outerwear & Coats and Calvin Klein, consistently outperformed others, while categories like Jumpsuits & Rompers and brands such as Marshal underperformed across all metrics. Inventory analysis highlighted a moderate sell-through rate across the catalog, with most products turning over only 25–50% of their stock, and over 5,000 products falling into the lowest performance bucket. Additionally, 129 products were identified as having zero or near-zero sales, signaling opportunities for deprioritization. Finally, K-Means clustering revealed three distinct product profiles based on stock, sales, and profitability, enabling targeted inventory and pricing strategies. These findings provide a comprehensive understanding of where the business currently generates value and where inefficiencies lie. This offers clear guidance for maximizing profitability through focused investments in high-performing markets, segments, and products, while reducing exposure to underperforming areas.

Conclusion

This study set out to answer how Looker E-Commerce can maximize profit and revenue by optimizing product performance and targeting high-value customer segments and geographic opportunities. Through detailed big data analysis, the findings highlight that profitability is concentrated in key regions such as China and the United States, and among middle-aged customer segments. High-performing products, particularly premium outerwear and brands like Calvin Klein, play a significant role in driving both revenue and profit. However, the business also carries a substantial number of underperforming or overstocked products, with over 5,000 items showing poor sell-through and 129 products flagged for potential deprioritization. These inefficiencies indicate a need for more precise inventory management and demand alignment.

To support continued growth and profitability, Looker E-Commerce should prioritize top-performing markets, customer segments, and product categories while streamlining or eliminating consistently underperforming items. Additionally, the use of K-Means clustering demonstrated how unsupervised learning can uncover meaningful product groupings, which can inform differentiated pricing, stocking, and promotional strategies. Moving forward, the company would benefit from adopting real-time analytics and machine learning models for demand forecasting, dynamic pricing, and personalized recommendations. These tools would

enable Looker E-Commerce to shift from reactive decisions to proactive, data-driven strategies, ensuring sustained competitiveness in the rapidly evolving e-commerce landscape.

References

Looker Ecommerce BigQuery Dataset (2025)
http://kaggle.com/datasets/mustafakeser4/looker-ecommerce-bigquery-dataset/data