

Data Analytics: BrightThreads Case Study

Course: Introduction to Data Analytics by Meta

Framework: OSEMN (Obtain, Scrub, Explore, Model, iNterpret) Process

Case Study: BrightThreads Boutique - Sales Growth Strategy

Executive summary:

This document demonstrates the complete data analytics lifecycle applied to BrightThreads Boutique, showcasing skills in data cleaning, exploratory analysis, predictive modeling, and business strategy development.

Key Deliverables:

- **Strategic Analysis:** SMART goals and KPI framework
- **Data Quality Assessment:** Comprehensive data cleaning documentation
- **Exploratory Insights:** Sales patterns, customer behavior analysis
- **Predictive Models:** Sales forecasting and customer segmentation
- **Business Recommendations:** Inventory optimization and marketing strategy

Obtaining and Scrubbing Data

Anna owns a clothing boutique in New York, called BrightThreads. She sells a mix of clothing brands and chooses items for her store that she believes her clients will like. She also sells online.

Anna is working on long-term planning for the upcoming year at BrightThreads. Business has been going well, but she would really like to increase sales and potentially open up a second location in a different neighborhood. Next year, Anna would like to increase her total sales by 10%. This would be a very good year for Anna and BrightThreads, but it seems doable based on the last few quarters and with some hard work.

Using this information, answer the questions below regarding the obtain and scrub stages of the OSEMN process. Add your answers to the template below.

In this scenario, what is a SMART goal that would benefit from data analysis?

S – Specific: Increase total sales and open a second store location in a new neighborhood.

M – Measurable: Achieve a 10% increase in total sales compared to the previous year.

A – Achievable: Based on recent quarterly performance and growth trends, this goal is realistic and attainable with focused marketing and operations planning.

R – Relevant: Opening a new store aligns with the company's long-term growth strategy and directly supports the goal of increasing sales.

T – Time-bound: Achieve the 10% sales increase and open the new store within the next year

What is a Primary KPI that would be useful to analyze for this goal?

A Primary KPI that would be useful to analyze for this goal is:

Total Sales Growth (%)

Formula:

$\text{Sales Growth (\%)} = (\text{Current Year Sales} - \text{Previous Year Sales}) / \text{Previous Year Sales} \times 100$

This KPI directly measures progress toward Anna's goal of increasing total sales by **10%** next year.

What relevant data would you gather in this scenario?

Historical sales, customer demographics, marketing performance, and inventory data.

How do you imagine you could obtain this data? What sources would you gather data from? Specifically, note what kind of data (first-party, third-party) and what methods you might use (survey, web analytics).

Type of Data	Source	Method	Type
Sales transactions	POS system, accounting software	Internal reports, database exports	First Party
Customer demographics & feedback	Loyalty programs, customer surveys	Online forms, in-store surveys	First Party
Website and social media analytics	Google Analytics, Meta Insights	Web analytics tools	First Party

Anna at BrightThreads has begun the process of gathering data to help analyze current sales.

She has collected data on recent online sales directly from the online storefront.

Anna has isolated 4 different segments that each have issues that need to be fixed. You can access each segment in the four sheets in this one spreadsheet. Click on each sheet for a different segment of the dataset. You can click on the tabs at the bottom of the spreadsheet to move between sheets. Review the image below for a preview:



The four sheets are accessible by clicking the tabs at the bottom of the spreadsheet.

[https://docs.google.com/spreadsheets/d/1Fa0xXTUDworwJ2_B5L1nmQUix8ROushLIBt_o64MdMk/edit
?usp=sharing](https://docs.google.com/spreadsheets/d/1Fa0xXTUDworwJ2_B5L1nmQUix8ROushLIBt_o64MdMk/edit?usp=sharing)

Using what you know about data validity, do you think the data Anna has gathered is valid? Why or why not?

Yes, the data Anna has gathered is valid because it was collected directly from the company's online storefront and tracked through Google Analytics. These systems automatically capture purchase activity for every transaction, ensuring that the data is objective, reliable, and free from human bias or manual entry errors.

What issue did you identify in segment 1 of the data?

Two duplicate records were found and removed to maintain data accuracy and prevent double-counting of sales.

What issue did you identify in segment 2 of the data?

There was inconsistent formatting in the **customer_zip** and **item_cost** columns. These inconsistencies were corrected to ensure uniform data structure and proper numerical formatting.

What issue did you identify in segment 3 of the data?

Two missing records were found and removed. After cleaning, only unique and complete records remain in the dataset.

What issue did you identify in segment 4 of the data?

Incorrect values were identified in the **item_cost** column — some prices were unrealistically high (e.g., 5999) or unusually low (e.g., 0.069). The low price was corrected based on similar product entries, and the high-priced record was removed because it did not match any valid item.

Exploring and Modeling Data

Anna from BrightThreads is exploring some data from last quarter's online sales.

The data was gathered from the BrightThreads online store.

Access: https://docs.google.com/spreadsheets/d/1K-ROPHxQfF2XV0b7hTvzqE2KIMwkhEDIG1ULLH_4Z2U/edit?usp=sharing

and click on Use Template in the upper right corner to access the dataset. Please note you will need to be logged into a Google account.

Review the following data and charts, then share what you can learn in the exploration stage of the OSEMN process.

Using this information, answer the questions below regarding the explore and model stages of the OSEMN process. Add your answers to the template below.

What are some things you can talk about this dataset? For instance, what does the size of the dataset tell you?

The dataset shows individual sales transactions from BrightThreads' online store.
The size of the dataset indicates how many sales occurred during the first quarter of 2023.
This helps assess overall business activity and sales frequency during that period.

What kind of data is in this dataset? (Numerical, categorical, etc.)

- **Numerical:** quant_items_per_order, order_total, total_sales
- **Categorical:** customer_id, zip_code, item_category, sales_level, day_of_week
- **Date:** sale_date

Reviewing this data, what is the minimum value in the order_total column? What is the maximum value in order_total column?

- **Minimum value:** \$39.99
- **Maximum value:** \$149.99

What kind of chart would you use to help visualize this data?

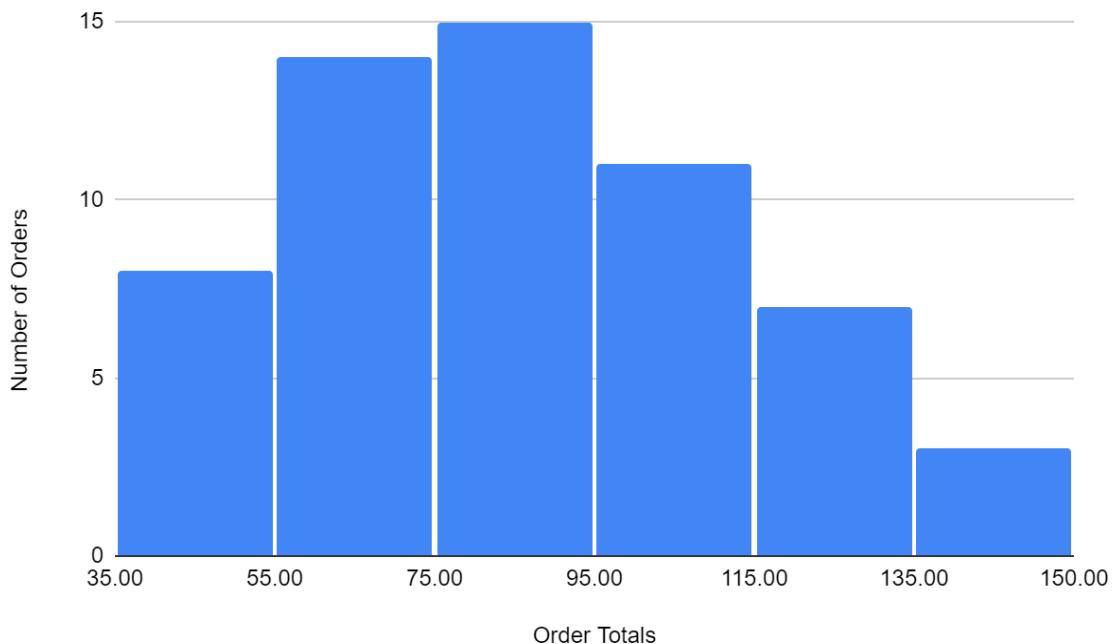
- **Bar Chart:** Compare total sales by day
- **Line Chart:** Show sales trends over time
- **Scatter Plot:** Examine relationships between quantity and order total
- **Histogram:** Visualize distribution of order totals

Based on what you have learned, would you add an additional column to this dataset using feature engineering? For instance, using the sales dates, would it be helpful to add in the day of the week data?

Yes, additional columns would provide more analytical value:

- **day_of_week** – Extracted from sale_date to identify which days have the highest sales.
- **total_sales** – Calculated as quant_items_per_order \times order_total to measure total transaction value per order.
- **sales_level** – Categorized using this formula to classify transaction value levels.

Anna has created the following chart to explore the relationship between order totals and the number of orders.



Based on the data in this chart, what would be a good title for this chart?

Distribution of Order Totals

What does this chart tell you about the number of orders in relation to the amount someone spends per order?

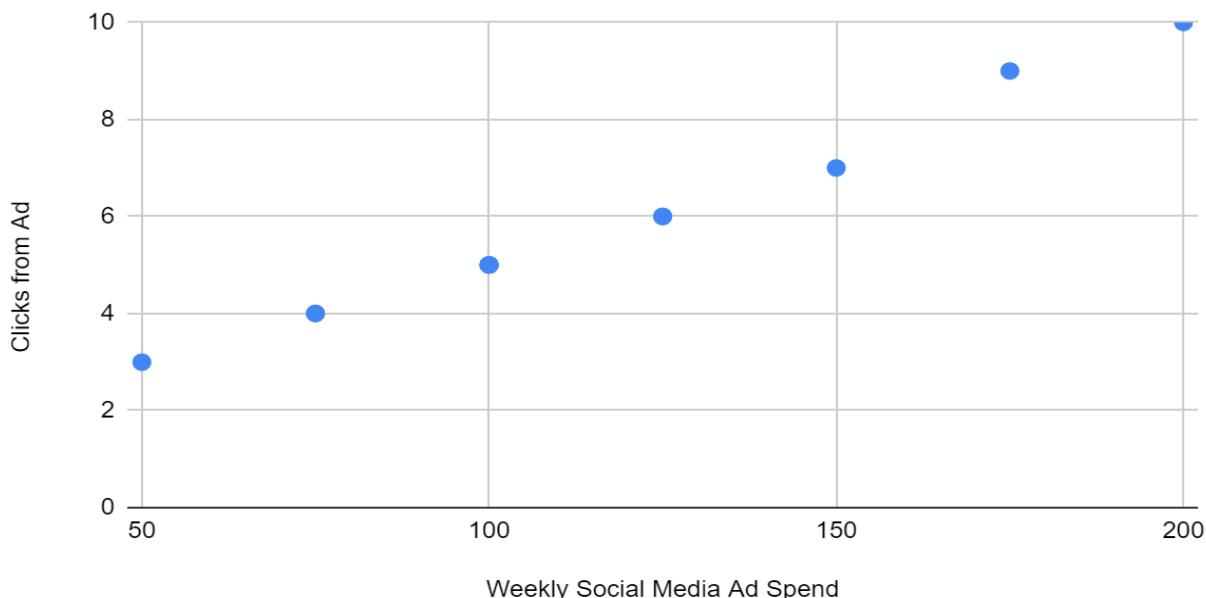
The majority of items purchased fall within a specific mid-range price band. There are fewer sales of the very cheapest and the most expensive items.

What range do most of the orders tend to be in?

Most orders tend to be in the range of: \$75.00 to \$95.00

Anna has also been analyzing data on the amount of money she spends on social media ads and how many clicks to the BrightThreads website they are generating.

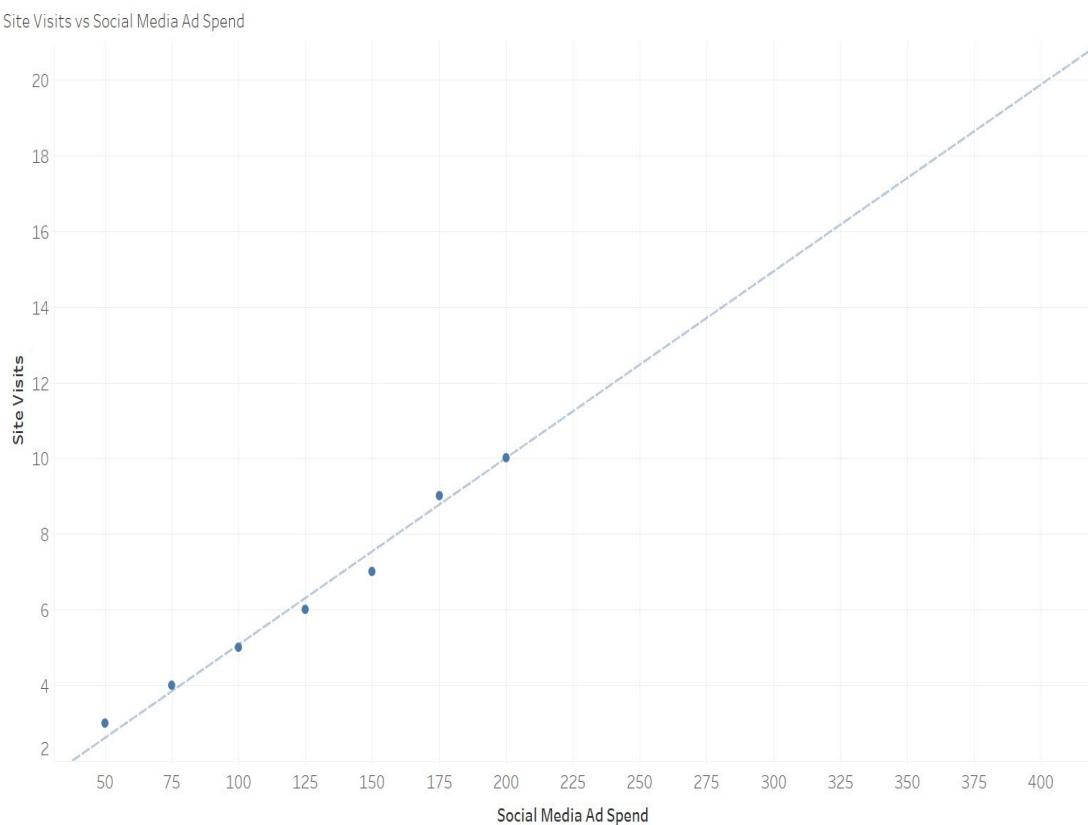
Site Visits vs. Social Media Ad Spend



Do you notice any correlations between the variables in this chart? If so, how would you describe them?

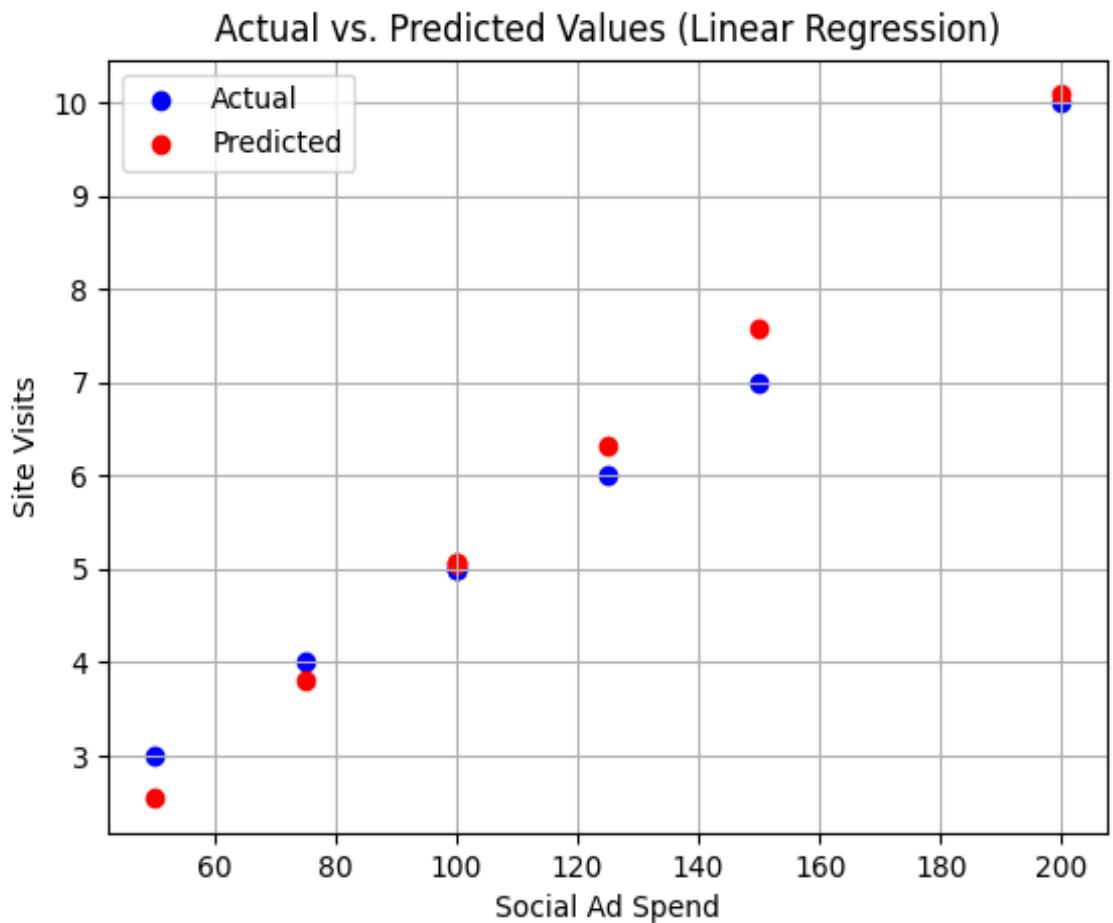
It shows positive correlation as amount spend on ad increase, the clicks on ad also increase.

Anna has learned a lot while exploring the data she has gathered. Now, it's time to model some of this data.



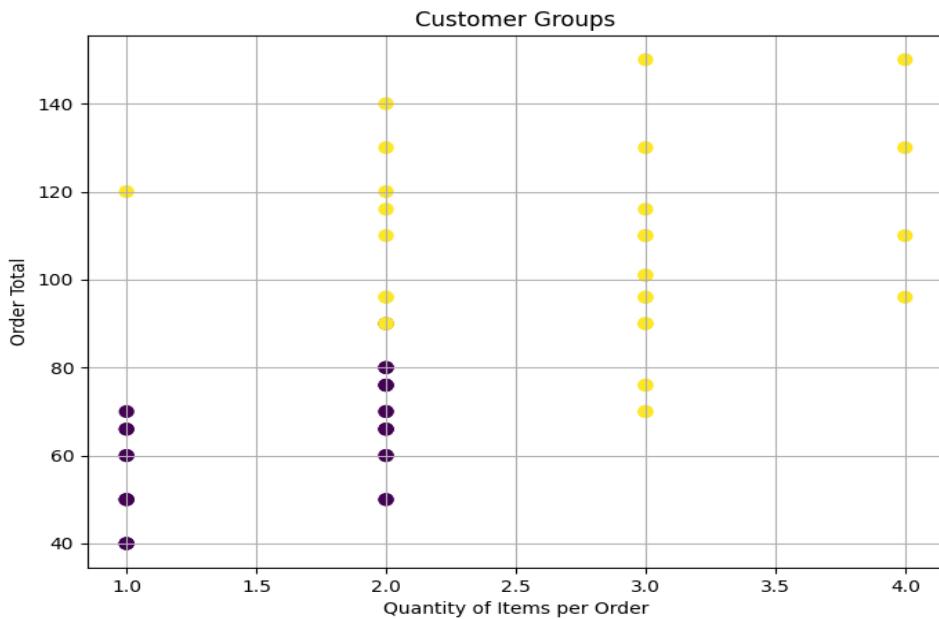
Reviewing this linear regression model, roughly how many site visits can be expected if the marketing budget is increased to \$250?

The model predicts that with a \$250 weekly ad spend, BrightThreads can expect to get about 12 to 13 site visits.



Review this linear regression model which shows the actual data values and the values predicted by the model when given a test set. Do you think that this model is sufficient for general use for this data? Why or why not?

This model is inaccurate and unreliable. Using it to make business decisions or predictions would likely lead to poor outcomes because its predictions are frequently wrong. A different, more complex model should be explored to better capture the underlying trend in the data.



Review this clustering model. A clustering algorithm has been used and identified two groups. How would you describe the two different customer groups? Why?

- **Group 1 (The "Value" or "Occasional" Shoppers):**
 - Spending: Lower order totals, roughly \$40 to \$80.
 - Behavior: They purchase smaller quantities, typically 1 to 2 items per order.

- **Group 2 (The "High-Value" or "Bulk" Shoppers):**
 - Spending: Higher order totals, ranging from \$70 to \$150.
 - Behavior: They purchase larger quantities, typically 2 to 4 items per order.

You are trying to forecast BrightThreads sales in the coming quarter- what model might you use? Why did you choose this?

I would use a Time Series Forecasting Model (like ARIMA) to predict BrightThreads' future sales. This model analyzes historical sales trends and seasonality, making it more accurate for time-based predictions than simple linear regression. It helps identify patterns and forecast upcoming sales more reliably.

iNterpreting Data

Anna has learned many things using data analysis. She has prepared a presentation to show to BrightThreads stakeholders. As a reminder, her goal is to grow sales by 10% in the upcoming year, and this presentation will cover what she's learned and how she plans to accomplish this goal.

Access [Anna's presentation](#).

Review the presentation, then share your thoughts on Anna's interpretation of the data at the end of OSEMN process.

Using this information, answer the questions below regarding the interpret stage of the OSEMN process. Add your answers to the template below.

What was the objective for this analysis?

The objective was to determine if BrightThreads could achieve a 10% increase in sales in the upcoming year by analyzing current sales data, identifying top-performing products, and forecasting future performance based on advertising and inventory strategies.

How does the date answer Anna's questions?

The data confirms that a 10% sales increase is achievable. It identifies key levers for growth: focusing inventory on top-selling items and reallocating social media advertising budgets to higher-performing channels, both of which are shown to directly increase sales and site visits.

How can Anna apply this in a business context?

Anna can immediately reallocate social media ad spending to the best-performing channels. She can also work with her buyers to shift inventory towards a more tailored selection of the most popular items. She should plan to reevaluate these strategies in six months to measure their impact.

What slides in the presentation shared the recap of the project?

The slides titled "What The Data Shows" and "Moving Forward" serve as the project recap, summarizing the key findings and the actionable plan.

What slides in the presentation covered the methods used in the project?

The slide titled "Methods For Analysis" covers the methods, mentioning the use of historical sales/advertising data, categorical and time-series analysis, and predictive models like linear regression.

What slides in the presentation included visualization of the project?

Slides containing charts and models visually supported the analysis, for example:

- "This chart shows our sales numbers for the last two years"
- "This model shows how much we'll need to sell each quarter..."
- "This chart shows our current top-selling items"
- "This model shows our potential sales increases..."
- "This chart shows our current spending on social media advertising..."
- "This model shows our potential increased site visits..."

What slides in the presentation provided an explanation of the project?

The slides "Project Objectives" and "Methods For Analysis" provide a clear explanation of what the project set out to do and how it was accomplished.

What slides in the presentation offered recommendations after the project?

The "Moving Forward" slide offers specific, time-bound recommendations, including immediate ad reallocation, a future inventory shift, and a plan to reevaluate in six months.

In your opinion, what parts of the presentation were meant to explain, engage, and enlighten the audience? Why?

- **Explain:** The "Project Objectives" and "Methods For Analysis" slides explain the *what* and *how* of the project in a clear, structured way for the stakeholders.
- **Engage:** The initial question, "Can we increase sales by 10%...?" and the use of visuals (charts/models) engage the audience by posing a central mystery and then using data to tell the story.
- **Enlighten:** The "What The Data Shows" slide enlightens the audience by revealing the conclusive

answer and the core insights, transforming raw data into understandable and actionable business intelligence.

In your opinion, what parts of the presentation were the setup, buildup, climax, and conclusion? Why?

- **Setup:** The "BrightThreads Goals" and "Project Objectives" slides set the stage by introducing the central goal and the plan to achieve it.
- **Buildup:** The series of slides showing charts and models constitutes the buildup. Each piece of evidence (sales trends, top items, ad performance) builds a stronger and stronger case for the final conclusion.
- **Climax:** The slide with the answer "Absolutely!" is the climax. It's the pivotal moment where the central question is definitively answered, delivering the key takeaway.
- **Conclusion:** The "What The Data Shows" and "Moving Forward" slides form the conclusion, summarizing the proof and outlining the concrete, actionable next steps for the business.

Conclusion

By systematically obtaining, scrubbing, exploring, modeling, and interpreting data, we confidently determined that a 10% increase in sales is an achievable goal for BrightThreads. More importantly, the analysis provided a clear, evidence-based roadmap to get there, turning a business ambition into a data-informed strategy. This project underscores that in the modern retail landscape; strategic growth is not just about intuition—it's about intelligence derived from data.