EBA5001: Practice Module in Analytics Project Management

Proposal for Harnessing and Implementing Data Analytics at Sole Fitness SG

Date of Proposal: 24th February 2025
Team Name: FitForward Analytics (Group 1)

Team Members:

Daniel Loh Guo Juin	A0149952R
Kuo Yu Chen	A0314462L
Lim Boon Yee	A0315645A
Lim Sze Zhong	A0312363N
Torana	A0199212A

1.0 Executive Summary

This report presents the outcomes of a data analytics project undertaken in partnership with Sole Fitness Singapore, a local SME specializing in fitness equipment. The engagement was conducted as part of the EBA5001 module on Analytics Project Management at the National University of Singapore. The project's objective was to help Sole Fitness leverage data analytics to improve product performance, optimize advertising efficiency, and establish a robust data infrastructure for long-term business intelligence.

Three core business problems were identified through initial discussions with Sole Fitness:

1. The lack of a unified data infrastructure.
2. Overdependence on treadmills as the primary revenue source.
3. Advertisement expenditure is spread across platforms without clear ROI tracking.

To address these challenges, the team constructed a complete end-to-end analytics solution, comprising:

- A PostgreSQL data warehouse integrating sales and advertising data across the website, Lazada, and Shopee.
- A suite of derived metrics such as Revenue Driver Score (RDS) and Return on Advertising Spend (ROAS).
- Clustering analysis to identify high-potential products and underperformers.
- ETL pipelines and Power BI dashboards to visualize business insights.

This project not only delivers tangible business value to Sole Fitness but also serves as a practical application of data management, modelling, and visualizations techniques for the project team.

2.0 Table of Contents

3.0 List of Figures

4.0 List of Tables

5.0 Introduction & Background

For this project our team engaged by local SME, Sole Fitness Singapore, an e-commerce retailer specializes in sporting goods. As Sole Fitness started its business operations in 2013, initially operating from a physical storefront in Telok Blangah and over time increased its online presence via a website, Shopee, and Lazada stores. The company faced low online sales during its early years. However, the surge in demand for home-based fitness equipment during the COVID-19 pandemic in 2020 marked a turning point significantly increasing online order volumes. Since then, Sole Fitness has experienced steady growth through its digital channels. Sole fitness recognized this opportunity to enhance business performance through their digital channels and our team led the implementation of data analytics processes.

This project aims to apply data analytic techniques to help Sole Fitness Singapore optimize product performance, enhance advertising efficiency, and streamline data management. By leveraging trend analysis, segmentation, and correlation studies, the project will identify key revenue drivers and recommend strategies such as discounting, bundling, or product phase-outs to maximize sales and diversify revenue streams.

E-commerce has reshaped retail, enabling businesses to scale beyond brick-and-mortar storefronts. Online stores and third-party marketplaces like Shopee and Lazada have made a multitude of products from different merchants more accessible to consumers, intensifying competition1. As the industry matures, digital advertising dominates marketing spend 1,2, helping e-commerce brands drive traffic and conversions. Rising ad costs, shifting consumer behavior, and seasonal demand fluctuations are key factors businesses must consider when optimizing marketing effectiveness. Online promotional campaigns also influence purchasing trends, contributing to pricing and revenue variations.

The ease of entry and unlimited digital shelf space in e-commerce allows for vast product assortments, yet many online retailers discover that a single best-selling "hero product" often generates most of their sales, accounting for a significant portion of their overall revenue2. However, over-reliance on a few top products carries risks as trends shift and markets saturate3. To maintain long-term stability, product mix diversification would ensure a more balanced revenue distribution.

As e-commerce continues to evolve, shifting market dynamics, competition, and data-driven decision making will shape the industry's long-term growth and sustainability.

6.0 Project Management

6.1 Project Vision

This project aims to deliver business value to Sole Fitness Singapore by establishing a scalable data infrastructure that will enable data-driven decision-making to diversify product revenue, optimise ad spending, and enhance business visibility.

6.2 Stake Holder Analysis

The following stakeholder analysis was done to gain an understanding of the concerns of the stakeholders and to address them.

Stakeholder	Key Concerns	Response to Key Concerns
CEO of Sole Fitness	1. Concerns over data privacy especially regarding revenue and profits. 2. Doubts over feasibility of project in terms of useability of company data for data analytics. 3. Current business practices are storing segmented excel files in shared drives.	1. Data will be kept confidential and only data that is necessary for understanding will be presented. Sensitive information will not be shared or exposed. 2. A review will be done on available data and data will be cleaned to make sure it can be used effectively for the project. 3. Project will propose a centralized and structured data storage approach that can scale with future analytics needs.
Marketing Director	1. Difficulty in tracking advertising spend across multiple fitness equipment categories. 2. A need to develop process to optimize ad-expenditure. 3. Concerns over effectiveness of current ad-expenditure.	1. Consolidation of ad spending information into a single, easy-to-read dashboard to improve visibility across the categories. 2. Analytical work to include metric to access returns on ad-expenditure which will also provide insights on which products can do with less ad-expenditure cutting overall budget.

Sales and Operations Head	1. Inconsistent product information (categories/ SKU) across platforms . 2. Treadmills make a majority of Sole Fitness revenue, has aspirations to explore other products that business can focus on pushing.	1. To standardise data to ensure consistency across platforms. 2. Analysis done for project will include efforts to identify products that have potential.

Table 1:Sole Fitness Stakeholder Analysis

6.2 Project Requirements Envisioning

Phase	Sub-Phase	Key Deliverables
Business Understanding	Preliminary discussions with Sole Fitness team	Documentation of Sole Fitness's business operations, context, and stakeholder expectations.
		Identification of current challenges.
		Data request and confirmation of data availability from Sole Fitness.
	Defining Business Problem	Defining business problems.
		Formulating business and technical objectives to align to business problems.
Data Understanding	Initial Data Review	Identifying origins of data sources.
		Summary of data structures and variable definitions.
		Listing of data quality or completeness issues identified.
Data Preparation	Data Cleaning and Integration	Cleaning datasets with errors corrected and missing values addressed.
		Unifying of data sources.
	Data Validation and Cleaning	Ensuring consistency between fields.
		Validate data with consistent logic and referential integrity.
Transformation and Aggregation	Feature Engineering	Formulation of derived metrics.
		Calendar dimension table for time-based analysis.
	Data Aggregation	Consolidation of sales and ad spend data at weekly and monthly levels.
		Reduce product categorisation into consistent hierarchical levels.
Data Modelling and Loading	Data Modelling	Building infrastructure data models for products, sales, categories, and advertising.

		Establishing relationships between entities for analytics.
	ETL Pipeline Implementation	Designing automated extraction, transformation, and loading of cleaned data.
Analytics Modelling	EDA	Conduct exploratory data analysis.
	Analytical Modelling	Analytical modelling and clustering to address business problems and achieve technical objective.
	Construct ETL Pipeline	Design data extraction process from Sole Fitness platforms (Website, Lazada and Shopee).
		Automate data integration process.
		Establish data loading process into dashboards.
Development	Development of Dashboard	Design dashboards tailored to monitoring needs.
		Implement automated dashboard refresh linked to ETL outputs.
	Scheduling periodic data quality checks	Develop periodic data validation schedule.
	Monitoring of metrics developed	Track key performance indicators (KPIs) in dashboards according to business needs.

Table 2:Project Requirements Phases with Key Deliverables

6.3 Initial Release Plan

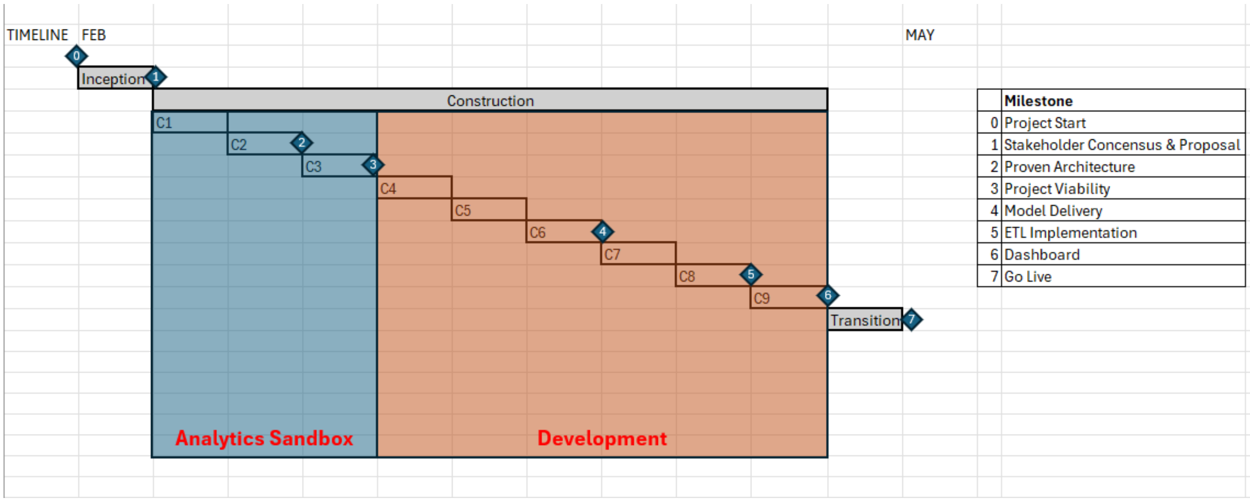
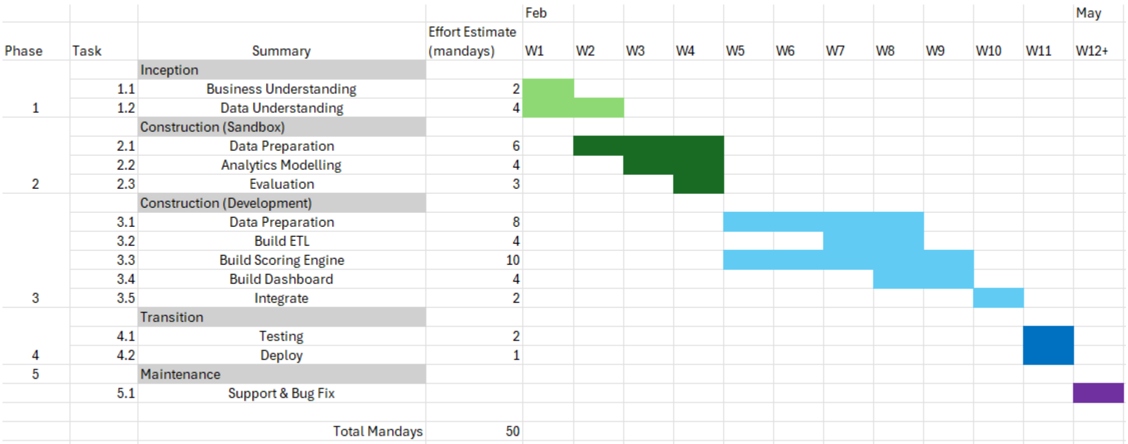


Figure 1 Project Timeline illustrating the phased progression from Inception through Construction to Transition

The project timeline is structured across three primary phases: Inception, Construction, and Transition, spanning from February to May. The Inception phase initiates the project, focuses on achieving Stakeholder Consensus & Proposal (1), establishing alignment and agreement on project objectives and scope. The Construction phase is subdivided into two key streams: the Analytics Sandbox (blue section) and Development (red section). Progress through Construction is marked by the achievement of critical milestones, including Proven Architecture (2), Project Viability (3), Model Delivery (4), ETL Implementation (5), and Dashboard Completion (6). Activities are split across sprints (C1 to C9), ensuring systematic progression towards each deliverable. The project reaches its completion in the Transition phase, where the final milestone, Go Live (7), signifies the operational deployment. This structured timeline provides clear visibility into project phases, dependencies, and key decision points essential for successful execution.

6.3.1 Resource Planning



				Feb											May
Phase	Task	Summary	Effort Estimate (mandays)	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12+
		Inception													
	1.1	Business Understanding	2												
1	1.2	Data Understanding	4												
		Construction (Sandbox)													
	2.1	Data Preparation	6												
	2.2	Analytics Modelling	4												
2	2.3	Evaluation	3												
		Construction (Development)													
	3.1	Data Preparation	8												
	3.2	Build ETL	4												
	3.3	Build Scoring Engine	10												
	3.4	Build Dashboard	4												
3	3.5	Integrate	2												
		Transition													
	4.1	Testing	2												
4	4.2	Deploy	1												
5		Maintenance													
	5.1	Support & Bug Fix													
		Total Mandays	50												

Figure 2 Project work breakdown by phase, task, effort estimates, and timeline

The project team consists of five dedicated roles: Project Manager, Business Analyst, Data Engineer, Data Scientist, and Developer. Resource allocation is aligned with task specialization to ensure efficient execution:

- The Project Manager (PM) oversees the overall scope, resource coordination, and timeline adherence across all phases.
- The Business Analyst (BA) plays a key role during the Inception phase, particularly in Business Understanding (1.1) and Data Understanding (1.2), by facilitating communication with stakeholders.
- The Data Engineer (DE) is primarily responsible for Data Preparation (2.1, 3.1) and Infrastructure Setup, supporting the construction of robust ETL processes (3.2).
- The Data Scientist (DS) is tasked with Analytics Modelling (2.2) and contributes to Data Cleaning efforts, supporting the Evaluation (2.3) of model performance.
- The Developer (DEV) is focused on the Build Dashboard (3.4) task, translating analytical outputs into a user-friendly interface.

Following Construction, the Transition phase addresses Testing (4.1) and Deployment (4.2), with cross-functional collaboration between the Data Engineer, Developer, and Project Manager. The final Maintenance phase ensures ongoing Support and Bug Fixes (5.1), safeguarding solution stability post-deployment.

6.4 Initial Architecture

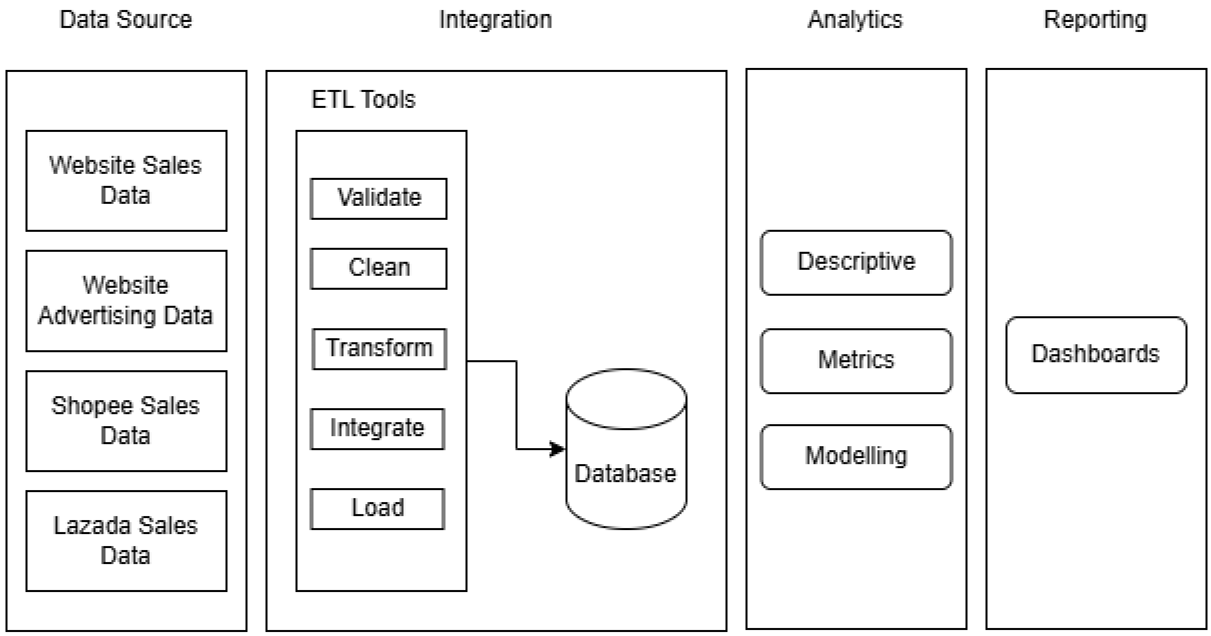


Figure 3 Initial system architecture diagram

The initial system architecture illustrates a streamlined end-to-end data pipeline, designed to support data-driven decision-making through comprehensive analytics and reporting capabilities. The architecture is composed of four main layers: Data Source, Integration, Analytics, and Reporting.

In the Data Source layer, raw data is ingested from multiple platforms, namely Website Sales Data, Website Advertising Data, Shopee Sales Data, and Lazada Sales Data. These diverse datasets feed into the Integration layer, where Python-based ETL tools perform essential preprocessing tasks, including Validation, Cleaning, Transformation, Integration, and Loading into a centralized PostgreSQL database.

The Analytics layer is responsible for deriving actionable insights from the consolidated data. Two key activities are undertaken in this layer:

- Metrics Development: Tracking key performance indicators related to sales trends and advertising effectiveness.
- Clustering Modelling: Applying clustering algorithms to group products based on performance characteristics, supporting strategic decisions on which products to promote or scale down.

In the final Reporting layer, insights are visualized through Power BI dashboards, which are updated on a weekly basis to reflect the latest available data. These dashboards are designed to support the information needs of key stakeholders, primarily the CEO of Sole Fitness, Marketing Director and the Sales and Operations Head, enabling timely and informed decision-making across marketing, sales, and operational functions.

This modular architecture ensures high data quality through systematic ETL processing, enables rich analytics for both descriptive and predictive use cases, and supports scalable, user-centric reporting.

6.5 Project Risk Management

Risk ID	Risk - Short Name	Risk Description	Impact	Probability	Risk Severity	Mitigation
R1a	Diversification – Market Entry Failure	Diversifying into the product may fail due to existing low demand, or misalignment with customer needs, leading to financial loss.	8	8	64	To ensure that the demand is valid, we have conducted preliminary market study based on the market demand for the various fitness equipment categories together with Sole Fitness. Some of the details are as below: • For treadmills, the Asia Pacific market (inclusive of Singapore) was valued at USD900M in 2024 with an expected CAGR of 7.3% up till 2031 (Cognitive Market Research, 2024). • For exercise bikes, the Asia Pacific market was valued at USD131M in 2024 with an expected CAGR of 7.0% till 2029 (Cognitive Market Research, 2022). • For functional trainer and multi-gym equipment, the Asia Pacific market is valued at USD1,414M in 2025 with an expected CAGR of 5.3% (Cognitive Market Research, 2022). With this market data, we further discussed with Sole Fitness on the feasibility of the diversification requirement and the addressable market size as well as future potential growth of the other fitness categories other than treadmills. The alignment was that the

Risk ID	Risk - Short Name	Risk Description	Impact	Probability	Risk Severity	Mitigation
						diversification requirement is a valid one and will be beneficial for the longevity of Sole Fitness.
R1b	Diversification – Core Product Neglect	Since Sole Fitness company's resources are limited, shifting resources (financial, human, operational) to the new venture risks underfunding the core product, reducing its revenue and market position.	6	8	48	To ensure that before diversifying, rigorous analysis is done to identify the respective product category: 1) Core Products 2) Products with high potential of success 3) Low performing products Ensure sufficient resources remain dedicated to core products to safeguard business continuity, even as efforts to expand into new revenue streams alongside Core product offerings. Resources should be reallocated from low performing products (such as products identified to be phased out), and not from the core products towards the products with high potential of success.

Risk ID	Risk - Short Name	Risk Description	Impact	Probability	Risk Severity	Mitigation
R2	Sole Fitness Data Input Quality & Lack of Data Driven Culture (Negative Risk)	Initial EDA into the dataset that Sole Fitness has provided us with has surfaced many inconsistencies and differing data format used in the input file. This may lead to an unexpected data input into the dashboard, with no exceptions being coded for it, causing the data pipeline to fail in handling these data issues.	8	8	64	To ensure that Sole Fitness can utilize the analytics solution and dashboard provided: • Data dictionary will be provided to Sole Fitness for each of the respective attribute in the input file. • Input validation with sound data validation rules (e.g. format checks, range checks, consistency checks across related fields, and mandatory field checks) will be coded out according to the data dictionary so that inconsistent / unexpected input will be called out as an error requiring the data user's attention to rectify. • In the longer-term view, to establish Data Governance Processes and a Data Driven Culture at Sole Fitness, proper workflows will need to be fleshed out. These include defining roles and responsibilities for data ownership and stewardship, implementation of data quality monitoring and reporting and establishing feedback loops and issue resolution processes.

Risk ID	Risk - Short Name	Risk Description	Impact	Probability	Risk Severity	Mitigation
R3	Lack of existing Power BI Automated Pipeline in Sole Fitness (Negative Risk)	As a Small-Medium Enterprise, Sole Fitness does not utilize dashboards for their analytics processes. The team at Sole Fitness usually does manual data processing in Excel, which is their current medium of analysis, causing increased errors and inefficiencies. In addition, not having any Professional Tableau License or Power BI Pro License	8	6	48	To establish a tactical dashboard with our dashboarding platform of choice (Power BI) with automated data pipeline & scheduled refreshes: • We will be providing Sole Fitness with the Python and PostgreSQL code needed to setup the data pipeline from the ground up, with data validation, data cleaning and ETL steps in place for loading into the database periodically. • Localhost server will be initially used on the Sole Fitness's employee's workstation such that running costs to be kept to a minimum. • We have liaised with Sole Fitness and the company will purchase a monthly subscription of the Power BI Pro License, and we will be helping to set up a Power BI Gateway and schedule weekly automated refreshes for the dashboard that will be published to their Sole Fitness Power BI Account.

Risk ID	Risk - Short Name	Risk Description	Impact	Probability	Risk Severity	Mitigation
		subscription will limit their access to the advanced capabilities of the analytics platforms.				
R4	Lack of Technical Expertise in Sole Fitness (Negative Risk)	Sole Fitness only possesses one Marketing Director who is currently responsible for the data collection and processing for further analytics. There is no dedicated analytics team in Sole Fitness. As such, users may not have the required technical skills to maintain and	8	7	56	To ensure that the benefits of the project can be realized post system deployment, several action items will be pursued as below: • We will be providing a simple training slides and detailed documentation to assist Sole Fitness employee in understanding how the system works and how to operate & maintain the pipeline and dashboard for proper handover. • Post roll out, there will be an extra 1-month maintenance phase (on top of the existing maintenance phase specified) for this project. With one of our teammates having the connection with the owner of the company, there will be an extra Extended Lifecycle Support (ELS) phase beyond this particular project timeline where a communication channel will be

Risk ID	Risk - Short Name	Risk Description	Impact	Probability	Risk Severity	Mitigation
		troubleshoot the system. This may result in a prolonged downtime and unresolved issues down the road, which meant that user satisfaction and long-term viability of the project may not be fulfilled, due to one of the key deliverables being at risk of being not accessible to Sole Fitness users.				established between Sole Fitness and our team and continued assistance will be provided so that Sole Fitness will have access to the expertise when needed until the end of the extra ELS phase when they would have picked up the relevant skills needed for proper maintenance.

Table 3: Project Risk Management

6.6 Project User Stories

6.6.1 Roles

Project Manager (PM): Oversees all phases, coordinates cross-functional tasks, manages risks, and ensures timeline adherence.

Business Analyst (BA): Leads Inception (Business/Data Understanding), defines KPIs.

Data Engineer (DE): Owns Data Prep, ETL, and Infrastructure tasks.

Data Scientist (DS): Drives Analytics Modelling, Evaluation, and metric calculations.

Developer (DV): Builds Dashboards and supports Testing/Deployment.

6.6.2 User Story Table

P1 is Inception, P2 is Construction (Sandbox), P3 is Construction (Development), P4 is Transition, and P5 is Maintenance

ID	Phase	Role	User Story	Depend-encies	Story Points
US-01	P1	BA	Define business objectives and KPIs for all 3 business problems, so that stakeholders have clear success metrics aligned with organizational goals.	None	3
US-02	P1	BA	Document data sources and validate alignment with business goals, so that the stakeholders can ensure data relevance and prevent scope creep.	US-01	2
US-03	P1	BA	Conduct feasibility study to evaluate the viability of the project, so that the stakeholders can make informed decisions and ensure the success of the project	US-01 US-02	1
US-04	P1	DE	Proposal initial architecture and infrastructure requirements for technical development	US-02	1
US-05	P2	DE	Clean and preprocess raw data for product and ad-spend analysis so that the Data Scientist receives high-quality data for accurate modelling.	US-04	4
US-06	P2	DS	Perform clustering analysis (k-means) to segment products (Problem 2) so that underperforming products can be identified for optimization.	US-02 US-05	5
US-07	P2	DS	Perform category, product analysis to obtained details product recommendation so Sole Fitness can focus on specific products.	US-06	3

ID	Phase	Role	User Story	Depend-encies	Story Points
US-08	P2	DS	Perform Revenue Stability analysis over different months so that non-stable products will not be recommended.	US-06	2
US-09	P2	DS	Analyse ads spend vs. sales correlation (Problem 3) so that the Marketing team can validate whether advertisement spending are effective.	US-02 US-05	5
US-10	P2	DS	Formulate new metric based on ad spending (Problem 3) so that best performing product category with ad spending can be identified and budget allocated further.	US-02 US-05	5
US-11	P2	DS	Finalize an evaluation report for product recommendations and ad spend insights so that the Project Manager can present actionable insights to stakeholders.	US-06 US-09 US-10	1
US-12	P3	DE	Design data models and define relationships between entities so that the database supports scalable, efficient, and accurate data storage and retrieval	US-04	4
US-13	P3	DE	Convert data cleaning, preprocessing, and feature engineering steps into reusable Python scripts so that the process is automated, consistent, and easy to maintain	US-05 US-06 US-09 US-10	3
US-14	P3	DE	Build production-grade ETL pipelines that ingest and process data in near real-time so that dashboards reflects up-to-date information	US-13	4
US-15	P3	DEV	Develop a Power BI dashboard with role-level security access (Data Owner, Marketing, Sales) so that users access only relevant data, enhancing security and usability.	US-11 US-14	4
US-16	P3	DEV	Integrate analytics outputs (clusters, metrics) into the dashboard So that decision-makers have a single source of truth for performance metrics.	US-11 US-14 US-15	3
US-17	P4	DEV	Conduct end-to-end testing of ETL pipelines and dashboard functionality, so that all components work seamlessly before deployment.	US-16	1
US-18	P4	DE	Deploy pipelines and dashboard to production so that stakeholders can use the solution.	US-16 US-17	1
US-19	P4	PM	Finalize user documentation and conduct stakeholder training so that users can understand the concepts behind the dashboard post-deployment.	US-18	1

ID	Phase	Role	User Story	Depend-encies	Story Points
US-20	P5	DEV	Address dashboard bugs and user feedback so that the solution remains reliable and user-friendly.	US-18	1
US-21	P5	DE	Optimize ETL pipeline performance so that data latency issues do not disrupt decision-making.	US-18	1
US-22	P5	PM	Conduct review meetings so that the solution evolves with changing business needs.	US-11 US-18	1

Table 4: Project User Stories

6.6.3 Burn Down Chart

The burn down chart is as shown below:

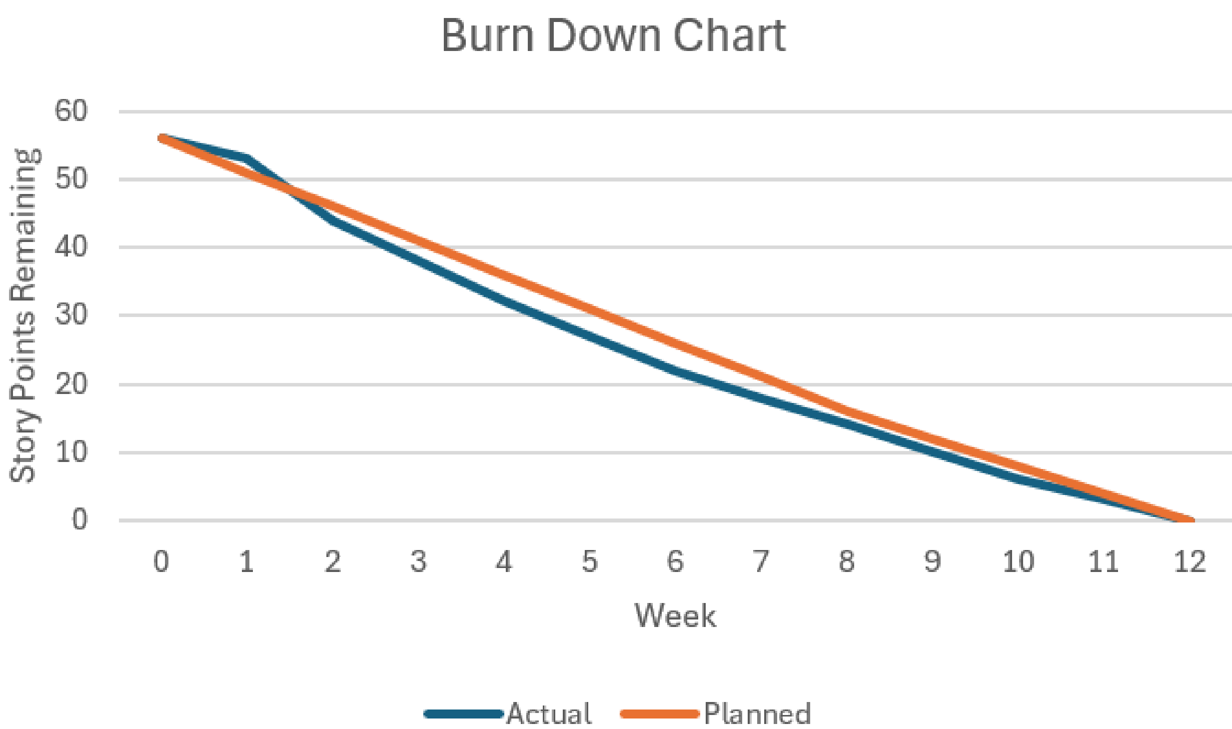


Figure 4 Project Burn Down Chart

7.0 Business Problem & Business/ Technical Objectives

Based on preliminary exploratory data analysis and discussions with the business, the business problem and objectives have been determined to be as below.

7.1 Business Problem 1

Business Problem 1: The business relies on a manual, Excel-based data entry and dashboarding process, which is time-consuming, error-prone, and inefficient. This limits the company's ability to gain timely and accurate insights on sales revenue and advertising impact.

Business Objective 1: To establish a data-driven culture and infrastructure that empowers continuous improvement in sales strategies, optimizes advertising investments for maximum long-term return, and enables timely, informed decisions across the organization, contributing to the business's sustainable growth and resilience.

Technical Objective 1: Automated Dashboarding and Reporting
Develop a data pipeline that ensures consistent, clean, and structured data for reporting and visualization. This includes designing standardized data models to eliminate duplicates, handle missing values, and enforce data integrity. The pipeline will extract, transform, and load (ETL) data into these models, ensuring uniformity across sales, marketing, and product listing datasets. Additionally, dashboard updates will be fully automated, providing real-time insights with accurate and reliable data. This solution enhances reporting efficiency, improves data quality, and ensures stakeholders have timely, actionable insights for decision-making.

7.2 Business Problem 2

Business Problem 2:
Despite offering over 50 product categories, Sole Fitness currently derives nearly half of its total revenue from a single product category – treadmills. While it is positive that treadmills have consistently performed well as the business' core revenue driver, this level of dependence is concerning as it introduces significant concentration risk. If treadmill sales were disrupted due to reasons including supply chain breakdowns, market saturation, or shifting consumer demand, the company could face substantial revenue loss and operational instability. Meanwhile, the broad range of other product categories may either hold untapped growth potential or have been consistently underperforming. Without a strategic approach to reduce this revenue concentration risk and diversify income streams, Sole Fitness remains highly exposed to future disruptions and market volatility.

Business Objective 2: The aim is to achieve a more balanced and resilient revenue distribution across Sole Fitness's product portfolio by increasing the combined revenue contribution from non-treadmill categories. High-potential categories that have been overlooked or under-leveraged will be identified and prioritized for targeted promotional or inventory efforts to drive revenue growth and product turnover. This approach seeks to diversify the business's revenue streams and reduce overreliance on a single product category, thereby mitigating future concentration risk. At the same time, categories that have been persistently underperforming will be flagged for evaluation, enabling data-driven resource allocation decisions on whether to optimize, scale down, or phase them out. These will ultimately support the business's long-term continuity and agility.

Technical Objective 2: Product Performance Analytics
To support this objective, a structured product performance analytics framework would be implemented. In addition to conventional sales metrics such as pageviews, add-to-cart counts, purchases, and revenue, the framework would incorporate newly derived metrics designed to provide deeper insight into product performance across stages of the customer funnel. Products and categories would be segmented based on monetization efficiency and conversion behavior to help surface current top performers, underperformers, and overlooked opportunities. K-Means clustering analysis would be applied to further substantiate these insights, grouping products by shared behavioral patterns across multiple performance dimensions. A seasonality and revenue stability check would also be conducted to refine product prioritization. This approach is expected to guide focused, data-driven decisions on which product categories to scale, optimize, or phase out.

7.3 Business Problem 3

Business Problem 3:
The business currently allocates its advertising budget across multiple advertising platforms (Google Ads, Facebook Ads, and Bing Ads) without a clear understanding of the impact on revenue growth. The reliance on a pay-per-click model across these advertisement platforms generally results in a noticeable correlation between advertisement spending and product views, making it challenging to assess the effectiveness of advertisement spending over time. Furthermore, the absence of a measurable ad spending metric prevents the business from optimizing advertisement spending based on item performance. With upcoming business plans to increase advertising expenditures, there is a concern about wasting valuable resources and missing opportunities to maximize viewership, customer engagement, and sales due to the lack of a data-driven approach to advertisement spending.

Business Objective 3:
To strategically realign advertising spend based on data-validated item performance, boosting marketing efficiency, enhancing customer acquisition, and driving revenue growth by optimizing advertising investment across different product categories.

Technical Objective 3: Advertising Optimization
Analyse the relationship between advertising spend and user engagement by evaluating product sales data across the various categories. This includes measuring ad performance and identifying correlations between pay-per-click (PPC) spending and product views. Develop efficiency metrics, such as Return on Advertisement Spending (ROAS), to compare performance across categories and optimize budget allocation based on data-driven insights.

8.0 Scope of Work

Works done will leverage available data to provide actionable insights and recommendations for Sole Fitness, specifically addressing product assortment optimization, advertising effectiveness and strategic promotions. The project will also establish a foundation for a data-driven culture through data pipeline automation and interactive dashboarding.

8.1 Data Foundation (Phase 1 and Phase 2 of Project Design)

- Inventory and access multiple data sources (Excel files) which contains a combination of sales data, product data and advertising data.
- Clarification with Sole Fitness on field's meaning and potential limitations.
- Aggregate and clean across data sources to ensure there are no duplicates and removing rows with missing data.
- Recategorize inconsistent products, in sales data that do not align with product data.
- Retrieve product category and SKU from product name if unavailable.
- Create new categories for products of special nature, i.e. (rental, display units, maintenance and servicing, rehab ...etc).
- Create a calendar dimension table that will map each sales row to its corresponding week number and month number.
- Feature engineering to derive relevant metrics on advertising efficiency and sales velocity
- Develop data models to represent key business entities (e.g., products, customers, sales, advertising spend) and create relationships between data models.
- Establish ETL processes that ensure data accuracy, consistency and accessibility.

8.2 Analytical Modelling (Phase 3 of Project Design)

- Leverage traditional domain-driven analytics, focusing on derived metrics for core analysis. Support insights with clustering algorithms to segment products based on key performance indicators like sales and customer engagement, to identify high-potential and underperforming products, including assortment decisions. Silhouette Score / Elbow method to be used to evaluate the models.
- Construct linear regression models to quantify the relationship between advertising spend and product views. Goodness of Model Fit (R^2) scores can be used to evaluate the models.
- Utilize correlation (evaluated using correlation score) to uncover product associations and patterns in customer purchasing behavior.
- Generate specialized metric Return on Advertisement Spending, ROAS for Sole Fitness advertisement spending and item sales data.

8.3 Insight Delivery (Phase 4 and Phase 5 of Project Design)

- Business Problem 2: Product Prioritisation Recommendations
    - Based on product performance analytics, products within selected non-treadmill categories will be recommended to Sole Fitness' marketing team to guide resource prioritization and planning of downstream marketing strategies.
    - It should be noted that specific marketing initiatives (e.g., promotions, bundling, or view-generation tactics) will not be prescribed as part of this analysis. These decisions will remain under the purview of Sole Fitness's internal marketing team. However, general strategic considerations will be highlighted to guide further planning.
- Overall Product Recommendation. Recommend products to focus on based on the consolidation of insights from solutions for Business Problem 2 and Business Problem 3.
- Develop 3 dashboards based on key business interests:
    - Sales Trends – Highlight top-performing products with metrics like revenue, units sold, and performance index to guide advertising priorities.
    - Category Performance – Identify well-performing product categories with sales distribution, customer engagement, and potential growth opportunities.
    - Advertising Efficacy – Track ad spending vs. product views & ROAS, to monitor item performance and optimize marketing efforts.
    - Integrate developed dashboards with the existing ETL pipeline to enable real-time or automated data refreshes for up-to-date performance monitoring.

9.0 Data Management Process

9.1 Data Sources

1. Sole Fitness Consolidated Website Report

Excel file containing information collected from the e-commerce shop hosted by Sole Fitness. The file has 3 sheets, extracted once for the year 2024. (600kb)

Sheet Summary

a) Sole Fitness Website's Sales Analytics Data aggregated weekly
b) Sole Fitness Advertising Spending Data aggregated weekly
c) Sole Fitness Website's Product Category Data

Known Issues

- Some product names in analytics data and product category data are manually entered (hybrid of free text and product name from website)
- Product name is used as a shared column for sales analytics and product category data
- The hierarchical structure of categories (main and subcategories) in this system is flexible and not strictly enforced, allowing a single subcategory to belong to multiple main categories.
- Items on discount are given a product category
- Category names in product category data are manually entered (free text)
- Category names in advertising data and product category data do not align fully

2. Sole Fitness Consolidated Lazada & Shopee Report

Excel file containing information collected from the e-commerce shop hosted by Lazada. The file has 2 sheets, extracted once for the year 2024. (100kb)

Sheet Summary

a) Lazada Sales Analytics Data aggregated monthly
b) Lazada Product Category Data
c) Shopee Sales Analytics Data aggregated monthly
d) Shopee Product Category Data

Known Issues

- Lazada & Shopee product categories are not aligned with website's product categories
- Lazada & Shopee sales analytics columns are different from website's analytics data

- Product URL is used as a shared column for Lazada's sales analytics and product category.
- Product SKU is used as a shared column for Lazada's sales analytics and product category.
- Lazada & Shopee's hierarchical structure of categories are different from websites' (main, sub categories, sub types)
- Category names in product category data are manually entered (free text)
- Missing categories found in product category

9.2 Data Set Used & Description

Data Dictionary

Variable	Description	Data Type	Sample Values(s)
Month	This field indicates the month of ads spending	Datetime (DD/MM/YYYY)	1/2/2024
Week	This field indicates the week of ads spending	Datetime (DD/MM/YYYY)	5/2/2024
Platform	This field indicates the sales platform which the ads direct viewing traffic to	String	WEBSITE
Medium	This field indicates the platform or source where the advertisement is being displayed	String	Google Ads
Category	This field indicates the type of fitness equipment the ads spending is used for	String	TREADMILL
Campaign	This field indicates the name of the advertising campaign the ads spending is categorized under	String	DSA – Treadmill
Cost	This field indicates the cost of the ad spending	Float	$80.85

Table 5:Sole Fitness Paid Ads Raw Data Table Data Dictionary

Variable	Description	Data Type	Sample Values(s)
Category	This field indicates the type of fitness equipment the ads spending is used for	String	TREADMILL
Campaign	This field indicates the name of the advertising campaign the ads spending is categorized under	String	DSA – Treadmill

Table 6:Sole Fitness Paid Ads Category Table Data Dictionary

Variable	Description	Data Type	Sample Values(s)
MONTH	This field indicates the month of the sales revenue	Datetime (DD/MM/YYYY)	1/10/2024
WEEK	This field indicates the week of the sales revenue	Datetime (DD/MM/YYYY)	1/10/2024
PLATFORM	This field indicates the platform through which the item sales has occurred	String	Shopee
CATEGORY 1	This field indicates the type of fitness equipment category	String	FREE WEIGHTS
CATEGORY 2	This field indicates the specific type of fitness equipment within a broader category	String	Dumbbells
ITEM NAME	This field indicates the specific name of the fitness equipment product on the sales platforms	String	Rubber Hexagonal Dumbbells Set with 3-Tier Rack
USERS	This field indicates the number of users that have interacted with the item	Integer	4
ITEMS VIEWED	This field indicates the number of non-unique views for the item	Integer	4
ITEMS ADDED TO CART	This field indicates the number of times the item has been added to cart for the sales platform	Integer	1
ITEMS CHECKED OUT	This field indicates the number of times the item has been checked out on the sales platform	Integer	1

ITEMS PURCHASED	This field indicates the number of purchases for the item on the sales platform	Integer	1
ITEM REVENUE	This field indicates the total sales revenue for the item	Float	3603.09

Table 7:Sole Fitness Revenue Raw Data Table Data Dictionary

Variable	**Description**	**Data Type**	**Sample Values(s)**
Date	This field indicates the week that the item is added onto the platforms	Datetime (DD/MM/YYYY)	11/3/2024
PLATFORM	This field indicates the platform to which the item has been added	String	Lazada
ITEM NAME	This field indicates the specific name of the fitness equipment product on the sales platforms	String	Livepro Tactical Weight Vests - Beige
CATEGORY 1	This field indicates the type of fitness equipment category	String	FUNCTIONAL FITNESS
CATEGORY 2	This field indicates the specific type of fitness equipment within a broader category	String	WEIGHTED VESTS

Table 8:Sole Fitness Revenue Category Table Data Dictionary

9.3 Data Integration

Different Keys Across Reports

Different reports used SKU, URL, and Product Name inconsistently as identifiers.
SKU was selected as the common key across all platforms to ensure consistent integration and a reliable join across datasets.

Product Name Matching Using Semantic Search

Product names were matched using Sentence-BERT text embeddings to find the nearest product name associated with SKU data.
A similarity threshold (e.g., 0.85) was applied; matches falling below this threshold were flagged for manual review to ensure accurate matching.

Inconsistent Product Categories Across Platforms

Products were labelled differently across sources (e.g., "3. Functional & Accessories" in Lazada vs. "FUNCTIONAL FITNESS" on the website).
Categories were standardized through a mapping process and manual review to ensure consistency across all platforms.

Source Filename Retention for Traceability

The original source filename was retained alongside each record to preserve data lineage and enable traceability back to the source file if any issues or discrepancies arise.

Enforced Source Priority Rules for Conflicting Data

In cases where conflicting information was found across platforms, the main website was explicitly designated as the authoritative source of truth.
Data from other sources was overridden to align with the main website's information to maintain data consistency and integrity.

9.4 Data Validation & Cleaning

Key Integrity Checks

- Duplicate Keys
  Detected and handled duplicate entries based on the integration keys (primarily URL and SKU, depending on data source).
  Duplicates were found both within individual source files and after integration; appropriate deduplication rules were applied.
- Enforced Correct Data Types
  Verified that all numerical fields in the analytics dataset (e.g., views, purchases, revenue) contained valid numeric types (integers or floats).
  Empty strings ("") and invalid formats were flagged and corrected.
- Date Format Standardization
  Enforced the yyyy-mm-dd format consistently across all date fields to ensure compatibility for downstream processing.

Analytics Data Consistency Checks

- Purchase-Revenue Consistency
  Flagged rows where purchases were recorded without corresponding revenue, and vice versa, for further manual investigation.
- Purchase-Views Consistency
  Flagged rows where purchases occurred without corresponding view counts, indicating potential data quality issues.

Product and Category Data Cleaning

- Spelling Corrections
  Corrected manual spelling errors identified in product names and category names (e.g., "WIEGHTS" corrected to "WEIGHTS").
- Category Formatting and Standardization
  - Converted all category names to uppercase.
  - Removed leading and trailing whitespaces.
- Category Consistency Across Sources
  Aligned category definitions between website analytics and website advertising datasets to maintain consistency.
- Category Reduction and Recategorization
  Rationalized Lazada and Shopee’s three-level category structure (Main Category →

Sub Category → Sub Type) into the website's two-level structure (Main Category → Sub Category).
Where mappings were ambiguous, business logic guided the consolidation, based on product characteristics and hierarchy.

Product SKU and URL Cleaning

- Removal of Product Variant SKUs
  Removed SKU variants when multiple SKUs pointed to the same URL.
  Since URL was treated as the primary key, retaining multiple SKUs would cause duplication and inconsistencies.
- Correction of SKU-URL Mismatches
  Manually identified and corrected mismatches where SKUs were linked to incorrect URLs.
- Manual Completion of Missing Data
  Supplemented missing SKU, URL, and category information where feasible, based on domain knowledge and cross-referencing datasets.

Range and Null Value Checks (Additional Validation)

- Range Validation for Numeric Columns
  Verified that numerical values such as views, purchases, and revenue were non-negative, unless justified (e.g., refund scenarios).
- Null and Empty Field Checks
  Conducted checks for nulls and empty strings in critical fields (e.g., SKU, URL, Category) after integration and cleaning.

9.5 Data Transformation

Sales Analytics Aggregation

Sales data was aggregated by SKU at both weekly and monthly levels to support different reporting granularity. Weekly aggregation applied to website-based data sources (sales and advertising), while monthly aggregation was conducted across all datasets, ensuring alignment and comparability for later analysis.

Feature Engineering of Performance Metrics

Derived key performance indicators (KPIs) such as revenue_driver_score (revenue/views) and conversion_rate (add-to-cart/views) to evaluate product effectiveness. Basic safeguards were applied to avoid divide-by-zero errors and ensure numeric integrity in derived columns.

Post-Transformation Edge Case Handling

Following feature engineering, checks were applied to detect and flag anomalies in derived metrics. This included spotting implausibly high ratios, negative values, or mismatches (e.g., products with extremely high revenue despite very low views). These safeguards improve downstream interpretability and prevent skew in dashboards or models.

Creation of Time Dimension Table

A separate time dimension table was created to enhance time-based reporting. Each unique date was enriched with additional attributes including year, quarter, month, and ISO week number, standardizing time semantics across datasets.

9.6 Data Modelling

9.6.1 Data Modelling Approach

In designing the data model, a relational schema normalized to the Third Normal Form (3NF) was adopted to ensure consistency, eliminate redundancy, and maintain data integrity across multiple sources. The primary goal was to create a robust and scalable structure capable of supporting long-term analytical and reporting needs across different e-commerce platforms.

The starting point was a denormalized dataset that combined cleaned and integrated data into a single table for sales and another for advertising. Each record contained detailed fields such as date, SKU, URL, product name, categories, platform name, and performance metrics (views, purchases, revenue, etc.).

While denormalization provided convenience for exploratory analysis, it introduced several structural issues:

- Data redundancy: Repeated platform names, product names, and category values across thousands of rows.
- Update anomalies: Changes to a product name or category would require updates across multiple rows, increasing the risk of inconsistency.
- Scalability limitations: Adding new data sources, metrics, or product attributes would have required schema changes and reprocessing large datasets.
- Lack of referential structure: Relationships between platforms, products, and categories were implicit and not formally defined.

To address these challenges, a normalized schema was introduced, organizing the data into distinct entities with clear keys and relationships. This provided a foundation for an extensible architecture that supported efficient querying, minimized storage duplication, and ensured reliable joins across tables.

9.6.2 Overview of Denormalized Datasets

Post transformation, there are two denormalized datasets, one for sales and one for advertising. These datasets consolidated raw inputs into flat, analysis-ready tables, but carried redundancy and lacked formal structure.

Sales Dataset Fields

- report_date: Aggregated start date (weekly for website data, monthly for all platforms)

- platform_name: Name of the platform (e.g., website, Shopee, Lazada)
- url: Product URL
- sku_code: Unique SKU for the product
- product_name: Text name of the product
- main_category, sub_category: Category labels
- user_count: Number of unique users visiting the product URL
- view_count: Total product views
- addtocart_count: Add-to-cart events for the product
- purchase_count: Number of purchases
- revenue: Total revenue generated
- revenue_driver_score: Ratio of revenue to views
- cart_to_view_ratio: Ratio of add-to-cart to views
- sourcefile: Name of the original file used for traceability

Advertising Dataset Fields

- report_date: Aggregated start date (always monthly)
- ad_medium: Advertising channel (e.g., Google Ads, Facebook Ads, Bing Ads)
- platform_name: Platform associated with the campaign (e.g., website, Shopee, Lazada)
- sub_category: Product sub-category targeted in the campaign
- campaign_name: Campaign identifier
- adspend_total: Total advertising spend for that category-platform combination
- sourcefile: File of origin, retained for lineage and audit purposes

While both datasets were useful for unified views of business performance, their structure was not ideal for scaling or relational integrity. Shared values such as platform names, sub-categories, and campaign names were repeated across many rows, motivating the shift toward normalization.

9.6.3 Entity Identification and Table Design

Platform

Represents each e-commerce or advertising platform where sales or campaigns occurred.

Key fields: platform_code, platform_name

Product

Captures information specific to individual SKUs and associated metadata.

Key fields: sku_code, product_name

Category

Categories are used for classification and filtering but vary across platforms with a hierarchical nature

Key fields: category_id, main_category, sub_category

Product-Platform-Url

Shows the url of each product per platform

Key fields: sku_code, product_name, platform_code, platform_name, url

Product-Category

Identifies the category for each product

Key fields: pcat_id, main_category, sub_category, sku_code, product_name

Sales Record

Represents time-based product performance.

Key fields: report_date, sku_code, product_name, platform_code, platform_name, metrics (e.g., views, purchases, revenue)

Advertising Campaign

Tracks aggregated ad spend by platform, sub-category, and medium.

Key Fields: report_date, platform_code, platform_name, sub_category, ad_medium, campaign_name, adspend_total

Calander Dimension

Table with year, month, week number for aggregation and comparison of data across different sources and time periods.

Key Fields: calendar_date, year, quarter, month, week, day

Here is the logical diagram from the above

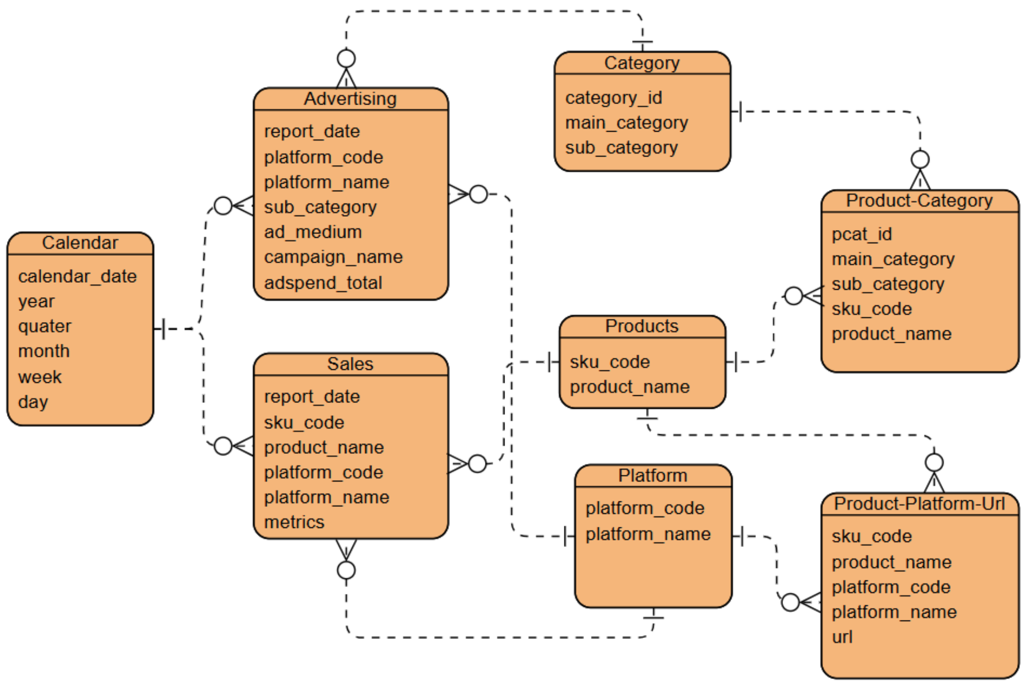


Figure 5 Initial Logical Model

9.6.4 Normalization to 3rd Normal Form

Platform

Primary Key: platform_code

Description: Unique code assigned for each sales or advertising platform (e.g., website, Shopee, Lazada).

Product

Primary Key: sku_code

Description: Each product is uniquely identified by its SKU across platforms.

Category

Split into 2 tables:

Main Category

- Primary Key: maincat_id

Sub Category

- Primary Key: subcat_id

Description: Categories are uniquely defined by a combination of main and subcategories to reflect product classification hierarchy.

Product-Platform-Url

Primary Key: ppuid

Foreign Key: platform code, sku_code

Description: Increment integer created sequentially by database for each product url per platform. Optional relationship to platform and product table because not every product has to be on a platform and not every platform may be available for products to be sold on.

Product-Category

Primary Key: category_id

Foreign Key: sku_code, subcat_id

Description: Increment integer created sequentially by database for each product's category. Optional relationship to product and subcategory because not every subcategory may have products in it and not every product may have been assigned a subcategory.

Sales Record

Split into 2 tables, one for the weekly aggregation of sales data and another for monthly aggregation of sales data

Primary Key: sales_id

Foreign Key: report_date, platform_code, sku_code

Description: Increment integer created sequentially by database for each sales record. Optional relationship to platform and product because sales records are only registered if a purchase occurs on a platform of a particular product

Advertising Campaign

Split into 2 tables, one for the weekly aggregation of advertising data and another for monthly aggregation of advertising data

Primary Key: ad_id

Foreign Key: report_date, platform_code, subcat_id

Description: Increment integer created sequentially by database for each advertisement record.

Calendar Dimension

Primary Key: calendar_date

Description: date value across a span of 10 years from 2020 to 2030

9.7 Data Dictionary (Post-Normalization)

Platform Table

Column Name	Key	Unique	Required	Data Type	Max Length	Sample Values	Default	Description
platform_code	PK	Y	Y	char	3	LAZ	-	Code for platform
platform_name		Y	Y	varchar	10	Lazada	-	Name of platform

Table 9:Platform Table Data Dictionary

Products Table

Column Name	Key	Unique	Required	Data Type	Max Length	Sample Values	Default	Description
sku_code	PK	Y	Y	varchar	10	F60	-	Unique product identifier
product name		N	Y	varchar	128	Sole F60 Folding Treadmill	-	Product name
is_active		N	N	boolean		TRUE	-	Is SKU code in use or deprecated

Table 10:Products Table Data Dictionary

Platform-Product-Url Table

Column Name	Key	Unique	Required	Data Type	Max Length	Sample Values	Default	Description
ppuid	PK	Y	Y	integer		14	-	Sequential number generated for unique identifier
platform_name	FK	Y	Y	varchar	10	Lazada	-	Name of platform
sku_code	FK	Y	Y	varchar	10	F60	-	Unique product identifier
url		Y	Y	varchar	256	https://www.lazada.sg/products/sole-f60-treadmill-i1622748217.html	-	Url to ecommerce shop listing

Table 11:Platform-Product-Url Table Data Dictionary

Category-Main Table

Column Name	Key	Unique	Required	Data Type	Max Length	Sample Values	Default	Description
maincat_id	PK	Y	Y	integer		2	-	Sequential number generated for unique identifier
main_category		Y	Y	varchar	50	CARDIO	-	First level hierarchical for categories

Table 12:Category-MainTable Data Dictionary

Category-Sub Table

Column Name	Key	Unique	Required	Data Type	Max Length	Sample Values	Default	Description
subcat_id	PK	Y	Y	integer		16	-	Sequential number generated for unique identifier
sub_category		Y	Y	varchar	50	TREADMILL	-	Second level hierarchical for categories
main_category	FK	N	Y	varchar	50	CARDIO	-	First level hierarchical for categories

Table 13:Category-Sub Table Data Dictionary

Product-Category Table

Column Name	Key	Unique	Required	Data Type	Max Length	Sample Values	Default	Description
category_id	PK	Y	Y	integer		3	-	Sequential number generated for unique identifier
sku_code	FK	Y	Y	varchar	10	F60	-	Unique product identifier
subcat_id	FK	Y	Y	integer		16	-	Sequential number generated for unique identifier

Table 14:Product-Category Table Data Dictionary

Advertising-Monthly-Rpt Table

Column Name	Key	Unique	Required	Data Type	Max Length	Sample Values	Default	Description
adm_id	PK	Y	Y	integer		42	-	sequential number generated for unique identifier
report_date		N	Y	date		10/2/2022	-	date of report that ads data was aggregated on
ad_medium		N	Y	varchar	20	Google Ads	-	location of ads placement (platform that ads were placed)
platform_code	FK	N	N	char	3	Website	-	ecommerce platform of which ads directed traffic to
subcat_id	FK	N	Y	integer		16	-	generic category for which ads were targetting
campaign_name		N	N	varchar	50	Treadmill campaign ads	-	ads campaign name
adspend_total		N	Y	float		982.12	-	total spent on ads campaign
source_file		N	N	varchar	128	budgetreport2024.xlsx	-	original tracking file used for data lineage
creadtedon		N	N	datetime		21/1/2025 10:12	now	datetime of when report was created

Table 15:Advertising-Monthly-Rpt Table Data Dictionary

Sales-Monthly-Rpt Table

Column Name	Key	Unique	Required	Data Type	Max Length	Sample Values	Default	Description
salesm_id	PK	Y	Y	integer		584	-	Sequential number generated for unique identifier
report_date		N	Y	date		10/2/2022	-	Date of report that sales data was aggregated on
platform_code	FK	N	N	char	3	Website	-	E-commerce platform of which ads directed traffic to
sku_code	FK	Y	Y	varchar	10	F60	-	Unique product identifier
user_count		N	N	integer		56	-	Number of users accessing the product listing
view_count		N	N	integer		2910	-	Number of views for the product listing
addtocart_count		N	N	integer		15	-	Number of add to cart for the product listing
purchase_count		N	N	integer		5	-	Number of purchases for the product listing
revenue		N	N	float		5419	-	Revenue from the product listing
revenue_driver_score		N	N	float		1.862199313	-	Engineered feature of total revenue divided by total views
cart_to_vew_ratio		N	N	float		0.005154639	-	Engineered feature of total add to card divided by total views
source_file		N	N	varchar	128	salesreport2024.xlsx	-	Original tracking file used for data lineage

creadtedon	N	N	datetime	21/1/2025 10:12	now	Datetime of when report was created

Table 16:Sales-Monthly-Rpt Table Data Dictionary

9.8 ERD

This ERD diagram illustrates a relational database schema for Sole Fitness, capturing product, category, platform, advertising, and sales data with supporting calendar dimensions to enable time-based reporting.

Weekly sales and advertising tables have been removed for clarity. The tables sales weekly and advertising weekly are identical to sales monthly and advertising monthly respectively.

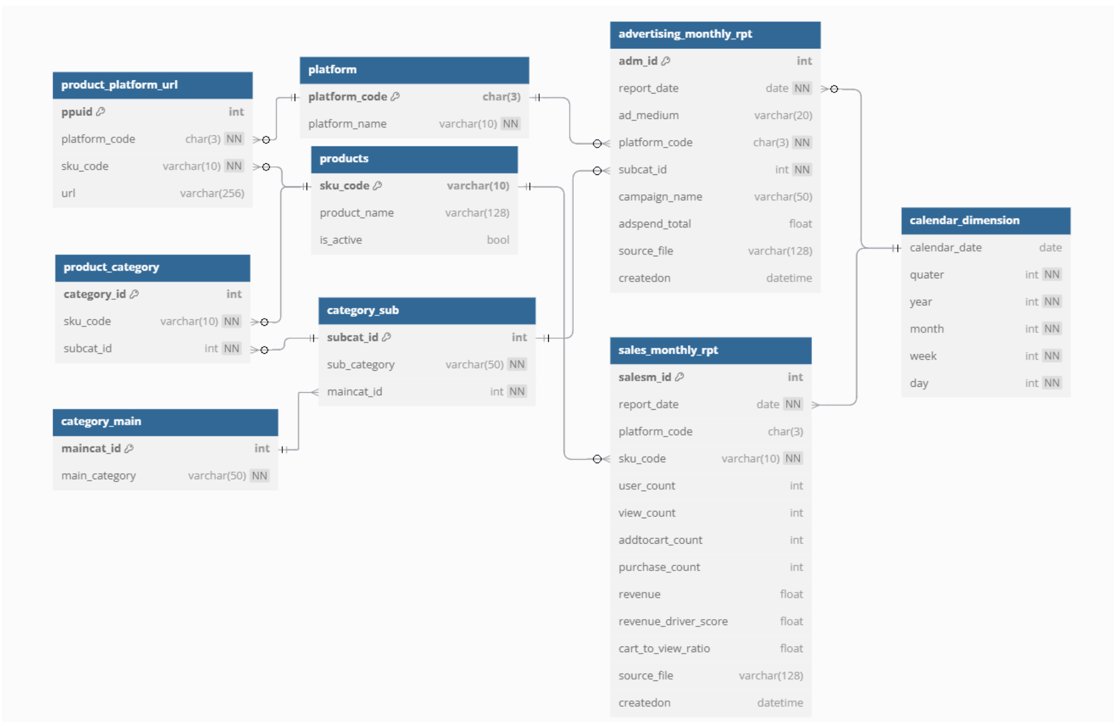


Figure 6: Logical Entity Relationship Diagram (ERD)

The DB tables will be loaded into the Power BI dashboard for further dashboard development. Please see below for the physical ERD from Power BI.

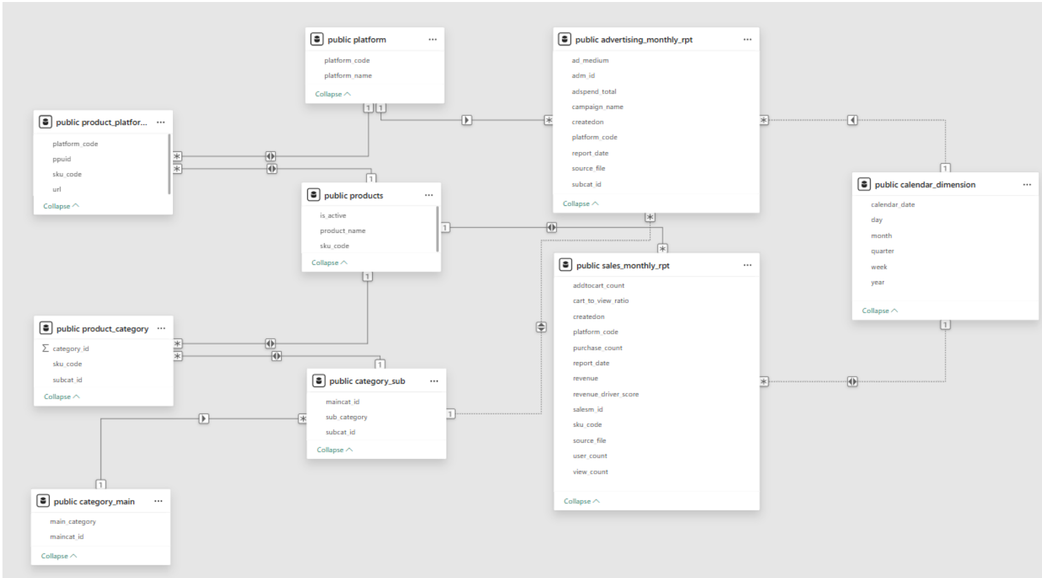


Figure 7: Physical Entity Relationship Diagram (ERD)

9.9 Data Governance & Security

9.91 Power BI Desktop Application Security Role Setup

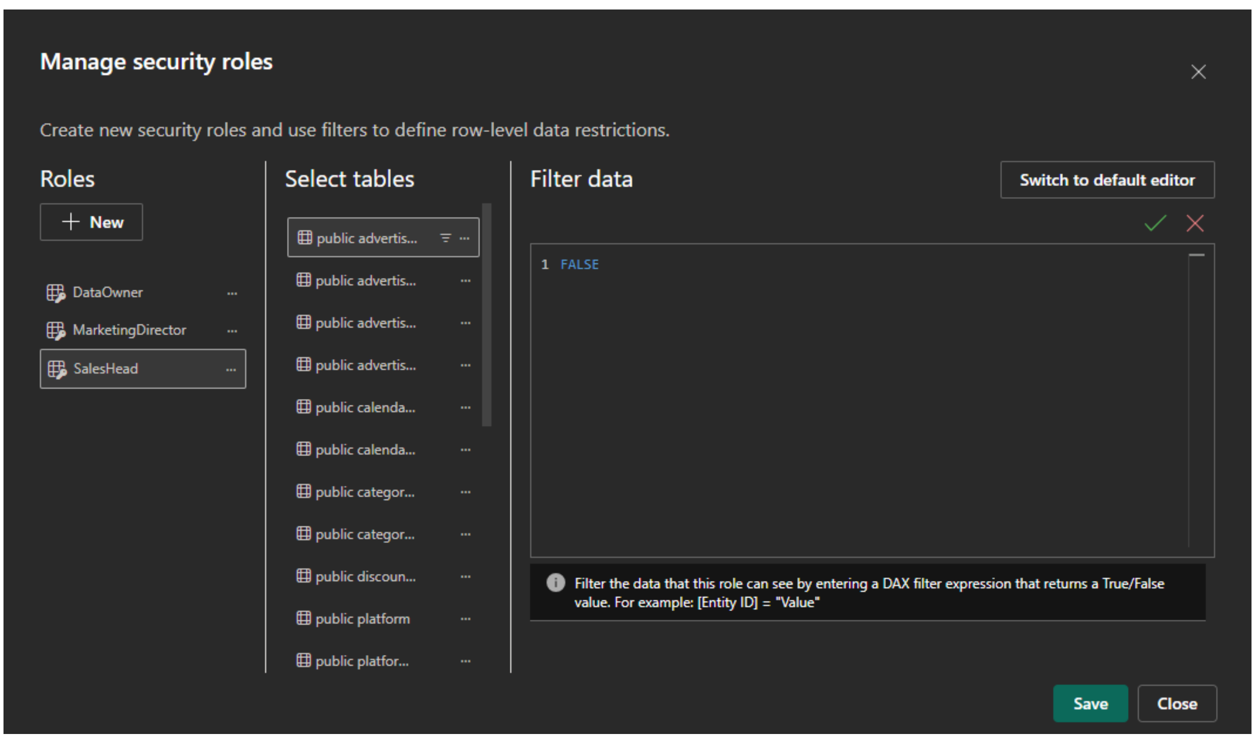


Figure 8: Image of Power BI Desktop Application Security Roles Setup for Sole Fitness Dashboards

To ensure data governance and security is adhered to in Sole Fitness, Row-Level Security will be implemented on the dashboard in the Power BI Desktop application as seen in Figure 8. This is done by creating three separate roles: 1) DataOwner; 2) MarketingDirector and 3) SalesHead based on the existing Sole Fitness business functions.

For the role of DataOwner, there will be full access to all of the data presented on the dashboard. This role will be assigned to CEO of the company. For the role of MarketingDirector, advertisement spending and metrics will be accessible, while access to the sales trend and categorical performance will be restricted to website only through the "platform" table. For the role of SalesHead, the sales trend and performance by category (including Revenue Driver Score) will be accessible, but advertisement metrics will be restricted through the "FALSE" setting of the "advertising_monthly_rpt" table.

9.9.2 Power BI Service Data Security

After the roles have been setup in the Power BI Desktop application, the dashboard is published onto the Power BI Service Workspace. To fully set up the Row-Level Security for the dashboard, the relevant Sole Fitness employees' email will be added as Members to each of the Roles as per the Table 17 and Figure 9 below.

Access to future employees will be provided on an as-needed basis by Sole Fitness, based on their job responsibilities and requirements.

Power BI Roles	**Employee Job Designation**
DataOwner	Sole Fitness CEO
MarketingDirector	Sole Fitness Marketing Director
SalesHead	Sole Fitness Sales and Operations Head

Table 17:Power BI Dashboard Row-Level Security (RLS) Designation for Sole Fitness

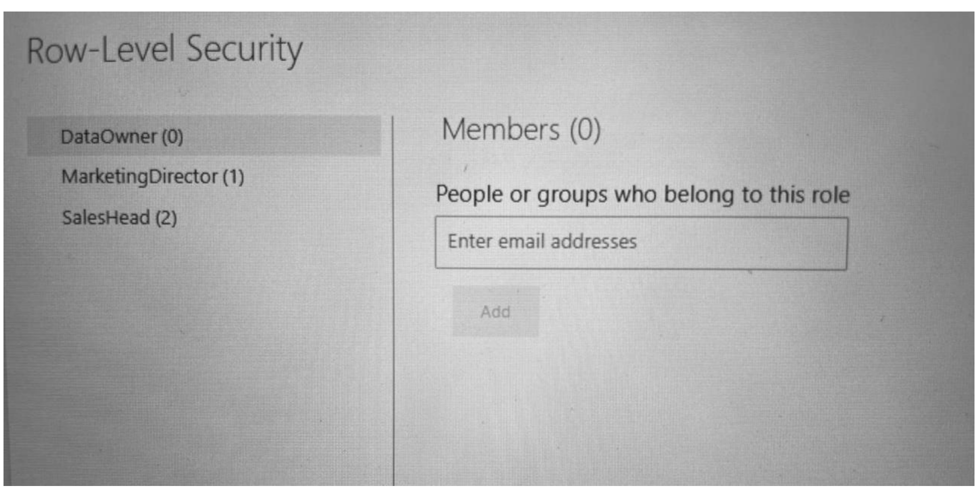


Figure 9: Image of Power BI Service Row-Level Security Setup for Sole Fitness Dashboards

10.0 Analysis for Business Problem 2

10.1 Background and Introduction

The business is burdened with an excess of underperforming products and an over-reliance on treadmill offerings, leading to an inefficient and high-risk model.

To streamline operations and optimize its product lineup, the company aims to identify items for discontinuation or phase-out.

Additionally, to enhance resilience and mitigate risk from dependency on treadmill products, the business seeks to diversify revenue streams by identifying high-potential non-treadmill products, fostering a more balanced and sustainable portfolio.

The Figure 10 below shows the two problems that the business is facing.

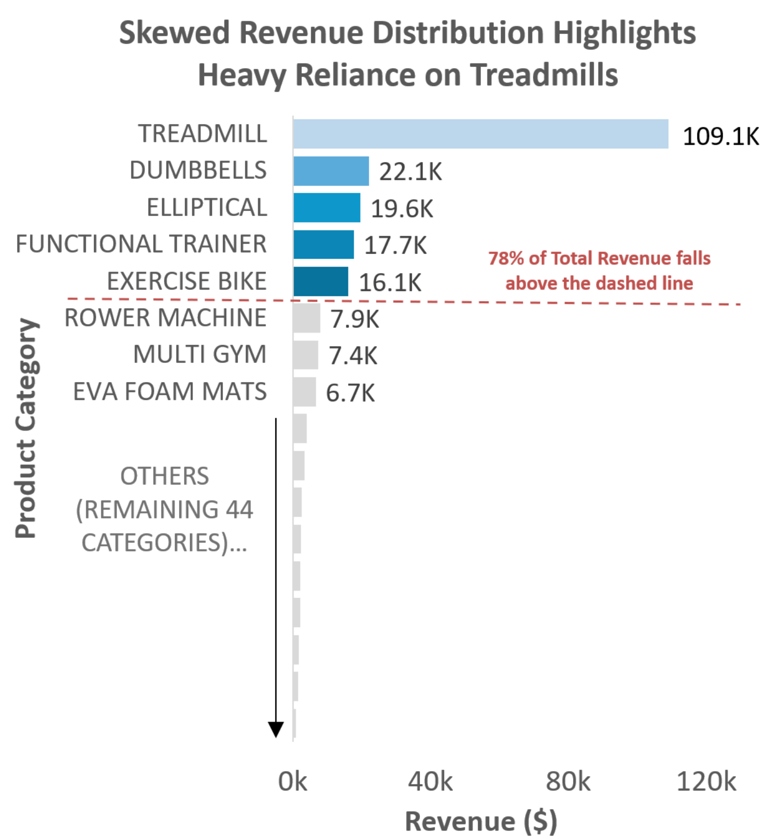


Figure 10 Excessive Dependence on Treadmill and Overload of Low Yield Products

The top 5 categories result in 78% of total revenue of the business, with treadmills taking up approximately 46% of total revenue of the business. There are also multiple other categories which have little to no revenue.

10.2 Analytics Methodology

Sole Fitness has limited money and time. It is not practical to optimize every product. Running tests on too many products at once can be expensive and lead to missed opportunities. As such, the approach below was used to strike a balance and provide the business with a short list of products which they can focus on for optimization.

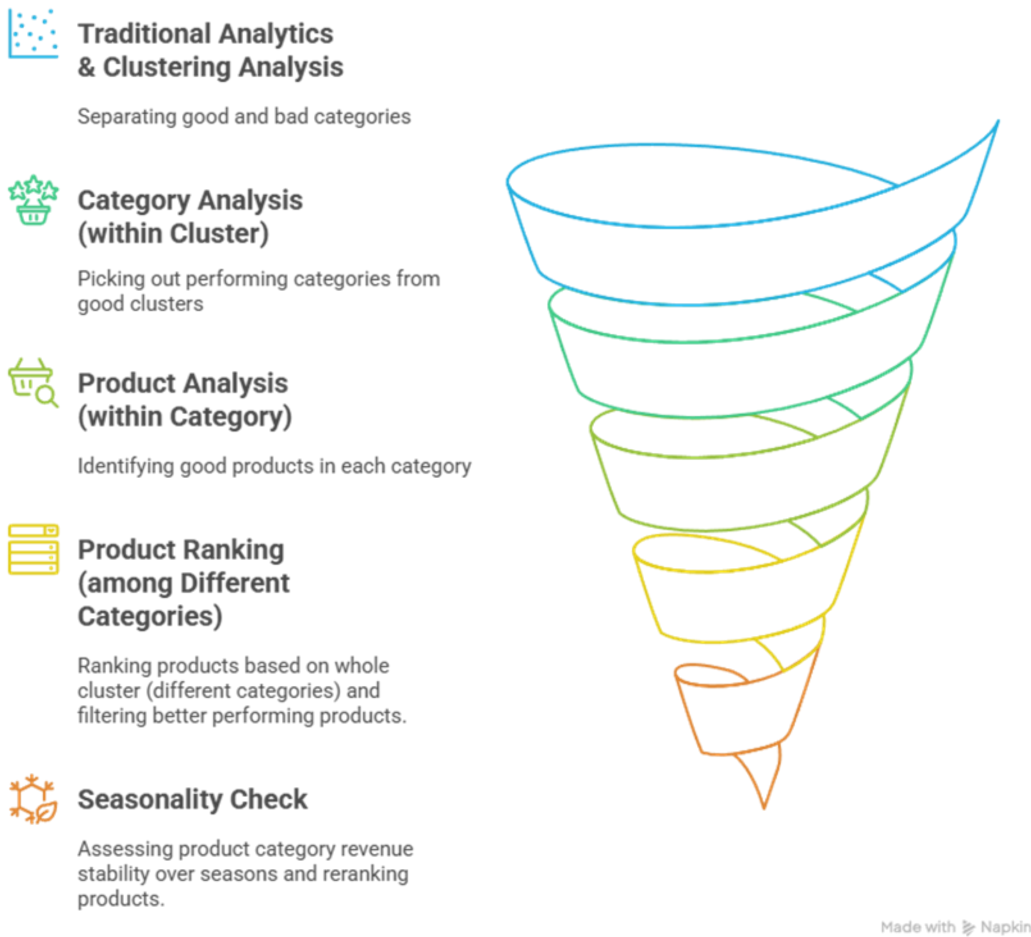


Figure 11 Analytics Methodology to Narrow Down on Product Recommendation

Conventional Analytics

The initial step involves leveraging traditional analytical methods to extract insights on key metrics. While conventional analytics provides valuable profiling data, it is inherently limited in its ability to capture complex interdependencies among metrics. Instead of analyzing each metric in isolation, the process focuses on gathering performance trends that inform the subsequent clustering stage.

Clustering Analysis

Once conventional insights have been obtained, clustering analysis is employed to segment products into distinct groups based on shared characteristics and behavioral patterns. This method ensures that all interdependent relationships between key metrics are considered, allowing businesses to better understand product dynamics beyond surface-level trends. Clustering eliminates the need for exhaustive comparisons across numerous categories by organizing products based on inherent similarities.

Category Analysis

Within each cluster, categories are analyzed to identify top-performing segments. This step ensures that subsequent optimizations focus on the most relevant and impactful product groups. By selecting categories with strong performance indicators, Sole Fitness can channel their resources effectively without spreading efforts too thin across lower-priority segments.

Product Analysis

Following category selection, products within these high-potential categories are examined in detail. Individual product performance is assessed to determine conversion efficiency, visibility, and revenue contribution. This stage further refines the focus by shortlisting top-performing products that show significant potential for optimization.

Product Ranking

Once key products have been identified, a broader comparison is conducted across different categories within various clusters. This ensures that products are not just optimized within their respective categories but also weighed against other high-performing items from different clusters. This step helps determine which products should be prioritized.

Seasonality Check

To ensure long-term sustainability, shortlisted products undergo seasonality analysis. This step evaluates whether a product maintains consistent demand throughout the year or exhibits fluctuations based on seasonal trends. Understanding seasonal dependencies is crucial in refining inventory decisions, marketing schedules, and promotional strategies.

Final Product Recommendations

The final recommendation is then made based on all the above considerations.

10.3 Key Metrics for Product Performance

The business has provided data on key performance metrics, including page views, add-to-cart actions, purchases, and revenue. These metrics align with the various stages of a customer's buying journey—ranging from initial attention to interest, desire, and ultimately, the decision to purchase. The primary objective is to facilitate a smoother progression of products through this funnel, ensuring more items transition from visibility to purchase, thereby driving overall revenue growth.

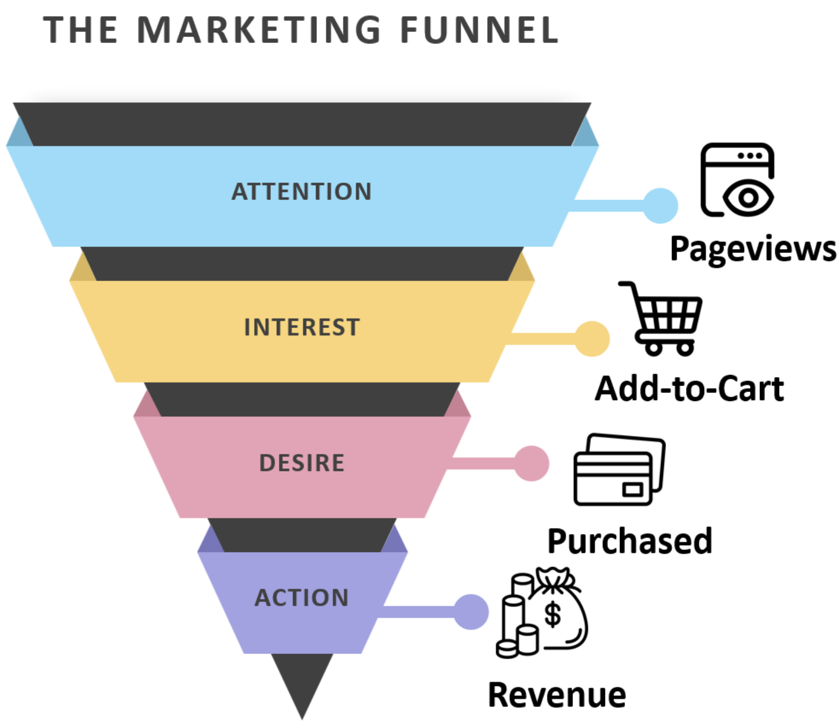


Figure 12 Marketing Funnel - From views to revenue

Beyond the fundamental metrics, additional insights can be derived by analyzing their interrelationships, such as ratio-based evaluations. Following discussions with the business, the analysis focuses on two key derived metrics: the Revenue Driver Score (RDS) and the Cart-to-View Ratio (CtV).

10.3.1 Revenue Driver Score (RDS)

The Revenue Driver Score quantifies the revenue generated per page view, providing a direct measure of a product's ability to convert visibility into sales. The Revenue Driver Score is essentially the amount of revenue made, from each page view.

$$Revenue\ Driver\ Score = \frac{Revenue}{Page\ Views}$$

A low Revenue Driver Score may indicate that despite high page views, revenue remains low. Initially, it was expected that lower-priced items would exhibit lower RDS values compared to higher-priced items. However, analysis revealed this assumption to be inaccurate. Observations suggest that customers experience lower purchasing resistance for affordable items than for higher-priced ones. As a result, RDS serves as a valuable indicator of product performance, independent of price considerations.

10.3.2 Cart-to-View Ratio (CtV)

The Cart-to-View Ratio can be calculated using the formula below:

$$Cart - to - View\ Ratio\ (CtV) = \frac{Add - to - Cart}{Page\ Views}$$

The Cart-to-View Ratio (CtV) serves as an indicator of customer interest in a product. While a high number of page views may suggest that targeted ads successfully attracted potential customers to the landing page, it does not necessarily reflect genuine interest or purchase intent. Many visitors may browse without any intention of making a purchase, whether out of curiosity or exposure to marketing efforts.

However, when a customer takes the additional step of adding a product to their cart, it signals a deeper level of interest and intent. This action requires a conscious decision and effort, suggesting that the customer views the product as a potential purchase rather than mere window shopping. Consequently, the CtV ratio provides a more reliable measure of engagement, helping businesses differentiate between casual visitors and prospective buyers.

A high CtV ratio indicates strong alignment between customer interest and the product's appeal, whereas a low ratio may highlight areas for improvement—such as pricing adjustments, enhanced product descriptions, or more persuasive calls-to-action. Analyzing this metric alongside other performance indicators can support strategic decisions aimed at increasing conversions and optimizing revenue potential.

10.4 Conventional Analytics on Product Metrics

10.4.1 RDS – Page Views Quadrant

With Revenue Driver Score (RDS) being a key and important metric, the information was plotted on a chart, with RDS plotted against Page Views to analyze the relationship between product visibility and revenue generation. This approach provides more actionable insights compared to plotting revenue directly against page views, as it accounts for efficiency rather than absolute revenue volume.

Revenue alone does not indicate how well a product converts visibility into sales, it only reflects total earnings, which can be influenced by factors like pricing rather than performance efficiency. A high-revenue product may simply have a high price point rather than strong customer engagement, while a low-revenue product may generate frequent sales but lack substantial earnings due to lower pricing. By contrast, RDS normalizes revenue by considering revenue generated per page view, allowing for a clearer comparison across different products.

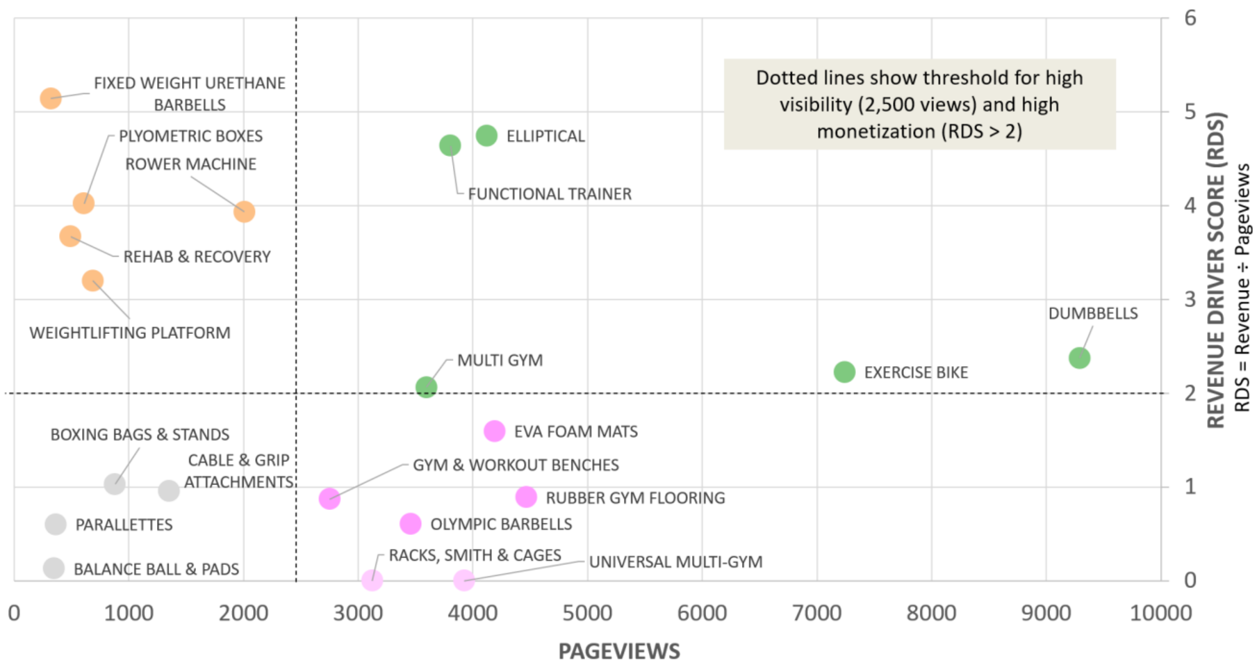


Figure 13 Product Category Distribution Across Page Views and Revenue Driver Score (RDS) excl. Treadmill

By plotting Revenue Driver Score (RDS) against Page Views, four distinct quadrants emerge, each offering valuable insights into product performance and efficiency in converting visibility into revenue.

High RDS – High Page Views

Products in this quadrant demonstrate strong sales performance and high visibility. Not only do these items attract significant page views, but they also efficiently generate revenue per view, making them highly valuable assets. These products are likely well-optimized in terms of pricing, marketing, and product presentation, effectively capturing customer interest and driving conversions. To further maximize their potential, strategies such as increased ad spend, cross-selling opportunities, and enhanced promotional efforts can be explored.

High RDS – Low Page Views

Products in this segment have strong revenue generation relative to their visibility but suffer from limited exposure. While these items perform exceptionally well when seen by customers, they may not be reaching a large enough audience. This presents an opportunity for the business to increase awareness and expand market reach through targeted advertising, improved search visibility, or better product placement within the platform. Investing in these products' promotion could significantly boost their overall revenue contribution.

Low RDS – High Page Views

Products in this quadrant attract a large number of visitors but struggle to convert them into revenue. This could stem from issues such as unappealing pricing, ineffective product descriptions, lack of compelling imagery, or unclear value propositions. Customers may browse the product but hesitate to make a purchase due to perceived risks or insufficient incentives. To address this, the business can refine their messaging, adjust pricing strategies, enhance product imagery, or introduce promotions like discounts or bundling to improve conversion rates.

Low RDS – Low Page Views

These products face the greatest challenges, as they neither receive substantial visibility nor generate sufficient revenue. They are often underperforming due to poor discoverability, weak customer interest, or fundamental product issues such as misalignment with market demand. Based on the current business problem, the recommendation would be to phase out or discontinue these products, while focusing the efforts on the other quadrants.

The four quadrants can be summarized in the Figure 14 below, illustrating the varying relationships between Revenue Driver Score (RDS) and Page Views. Ideally, businesses would aim to strategically transition products from the Low RDS – High Page Views and High RDS – Low Page Views segments into the optimal quadrant of High RDS – High Page Views, where products benefit from both strong visibility and efficient revenue generation.

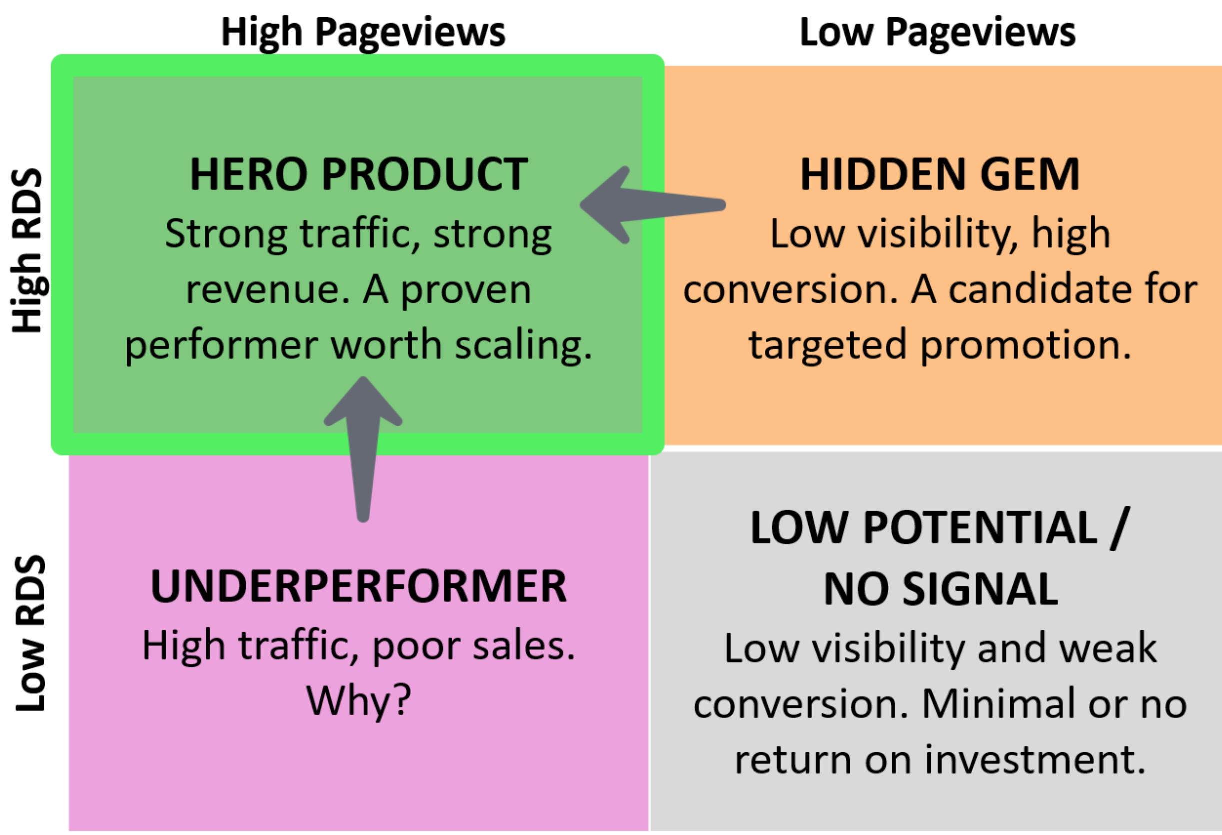


Figure 14 RDS - Page View Quadrants

10.4.2 Cart-to-View Ratio Category Comparison

A comparative analysis was conducted across the top 10 categories with the highest Cart-to-View (CtV) ratios, providing deeper insights into customer engagement and purchase intent. Notably, Rubber Gym Flooring and EVA Foam Mats, despite being categorized within the Underperformer quadrant, exhibited a Cart-to-View Ratio exceeding 10%.

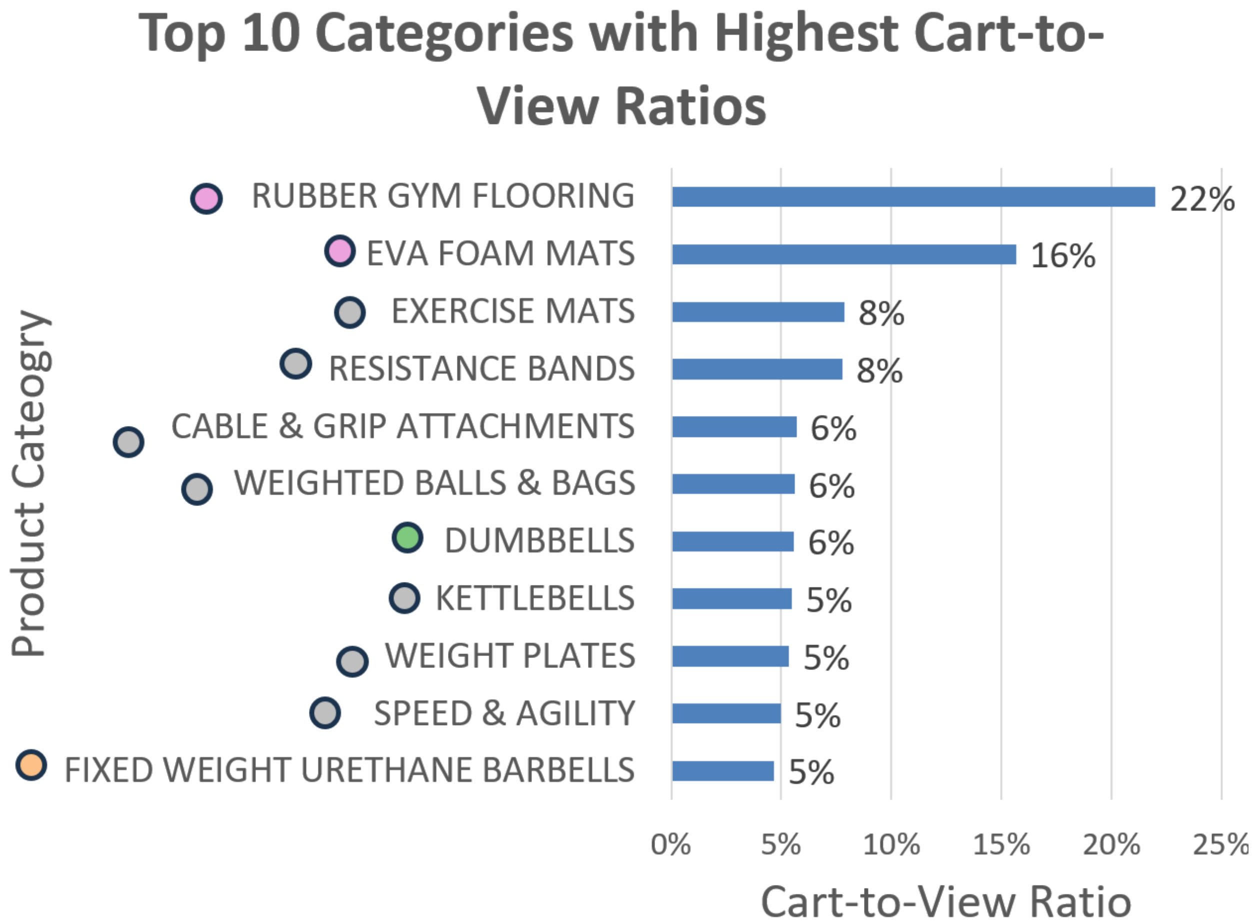


Figure 15: Top 10 Categories for Cart-to-View Ratio labelled with RDS Quadrant Profiling

This finding suggests that while these products may not have exceptionally high overall page views or revenue generation, they demonstrate strong customer interest, and a higher likelihood of conversion once viewed. A CtV ratio above 10% indicates that a significant portion of visitors actively add these items to their cart, signaling potential demand that may not be fully leveraged due to visibility or pricing constraints.

10.4.3 Limitations of Conventional Analytics

While conventional analytics provides valuable insights into individual metrics, it often falls short in capturing the complex relationships between multiple variables simultaneously. Traditional methods require analyzing each metric in isolation, which can lead to fragmented conclusions that overlook key dependencies and interactions. For instance, evaluating page views, conversion rates, and revenue separately may provide insights into product performance, but without understanding their interdependence, businesses may miss critical patterns that influence purchasing behavior.

Additionally, the conventional approach demands extensive manual effort, as each metric must be examined across multiple categories, compared, and prioritized. This process is time-consuming and exhausting, especially since the dataset contains more than 50 categories and more than 10x the number of products.

To overcome these challenges, clustering analysis offers a more comprehensive solution by grouping products based on shared characteristics and underlying behavioral trends. Unlike conventional analytics, clustering considers all relationships holistically, enabling businesses to uncover hidden patterns and categorize products more effectively. Once clustering has segmented the data, profiling and deep-diving analysis can be conducted with greater precision, ensuring that optimization efforts are targeted and resource-efficient.

It is important to recognize that conventional analytics serve as a foundational step, providing the necessary insights to inform clustering parameters and ensure meaningful segmentation.

10.5 Clustering Analysis

Before any analysis is conducted, it is important to know what features to include into the model, and conduct EDA (Exploratory Data Analysis) on the data to assess the suitability of each feature, identify potential transformations, and detect any redundant or highly correlated variables that may affect the integrity of the model.

10.5.1 Feature Engineering

Feature engineering is a crucial step in refining raw data into structured and interpretable metrics. By incorporating both base features and derived features, the model can better represent customer behavior and product performance.

The base features used include:

- Number of Views (items_viewed)
- Number of Carted (items_cart)
- Number of Check Outs (items_chk) – Not used due to high correlation with items_cart which will be explained later.
- Number of Purchases (items_pur)
- Amount of Revenue (items_rev)

Derived features were also created to capture the relationships between the different base features.

- Revenue Driver Score (rev_driver_score)
- Cart-to-View Ratio (add_cart_rate)
- Checkout to Cart Ratio (chkout_cart_rate)
- Purchases to Checkout Ratio (buy_chkout_rate)
- Purchases to Cart Ratio (buy_cart_rate)
- Purchases to View Ratio (buy_view_rate)

There were also other features developed but as they were highly correlated with other features, they shall not be mentioned here.

10.5.2 Transformation of Distribution of Data

Data transformation is essential to ensure that features are appropriately scaled and contribute effectively to the clustering process. This is necessary because k-means clustering relies on distance calculations, and features with different scales can disproportionately influence the results. Proper transformation helps in achieving more accurate and meaningful clusters.

To address this, several transformation methods were tested:

1) Standardization: This method standardizes features by removing the mean and scaling to unit variance. It ensures that all features are similarly spaced in terms of distance, providing a balanced contribution to the clustering algorithm. Standardization is particularly useful for algorithms that assume normally distributed data, as it transforms the data to have a mean of 0 and a standard deviation of 1. This helps in mitigating the effect of features with larger scales dominating the clustering process.

2) Log Transformation + Standardization: Applying a log transformation helps stabilize variance and make the data more normally distributed. This transformation is particularly useful for skewed data, as it can reduce the impact of extreme values and make the distribution more symmetric. After log transformation, standardization is applied to ensure features are appropriately scaled. This combination can enhance the clustering process by providing a more balanced representation of the data. The log transformation helps in compressing the range of values, making the data less skewed and more suitable for distance-based algorithms like k-means.

3) Box-Cox Transformation + Standardization: The Box-Cox transformation is a power transformation that aims to make the data more normally distributed. It is more flexible than log transformation as it can handle both positive and negative values. Following the Box-Cox transformation, standardization is applied to ensure features are appropriately scaled.

After thorough evaluation, it was determined that the combination of log transformation and standardization was the most reasonable and effective for our data. This approach provided a better balance between stabilizing variance and ensuring features were appropriately scaled for the clustering process. The log transformation helped in reducing skewness and making the data more normally distributed, while standardization ensured that all features contributed equally to the distance calculations in k-means clustering. This combination was particularly effective in handling features with varying scales and distributions, leading to more accurate and meaningful clusters.

Below is a typical feature of items_viewed which shows pre and post transformation. For this feature, there is not much difference between Box Cox Transformation and Log Transformation.

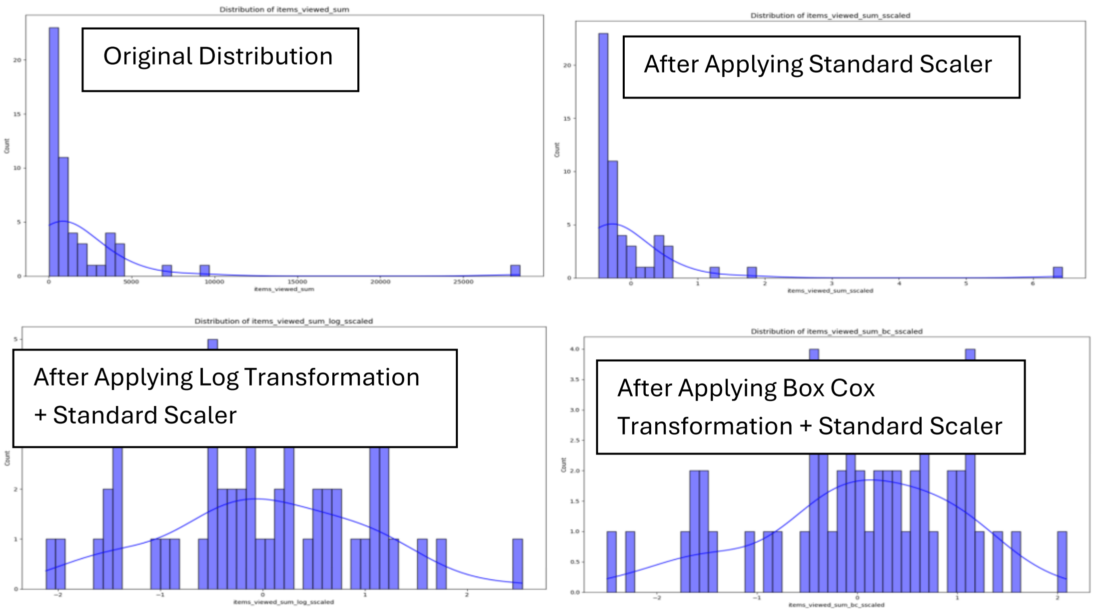


Figure 16 Distribution of Feature "items_viewed" Pre and Post Transformation

However, for the feature items_rev, where the numbers can be quite sparse, a Log Transformation might seem more logical than a Box Cox Transformation. This is one of the reasons we use Log Transformation instead of Box Cox Transformation.

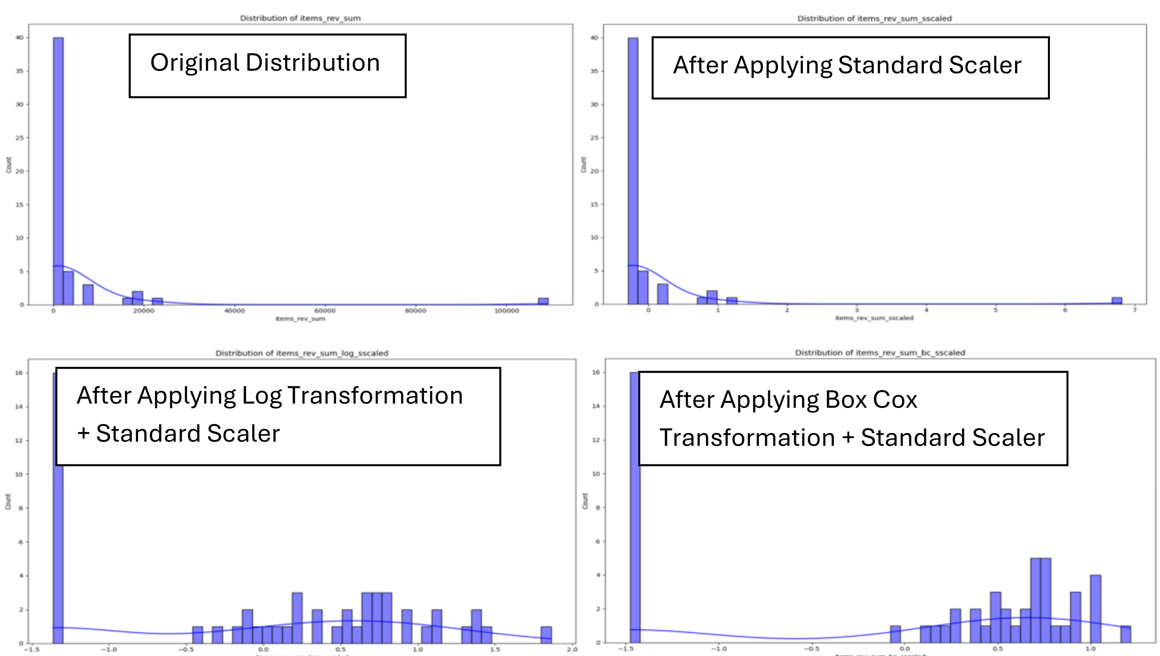


Figure 17 Distribution of Feature "items_rev" Pre and Post Transformation

The items_rev from the previous page had quite a number of rows where items_rev = 0. To improve the transformation, we also considered removing rows where items_rev = 0. The result is a much better distribution which we will keep in mind for now.

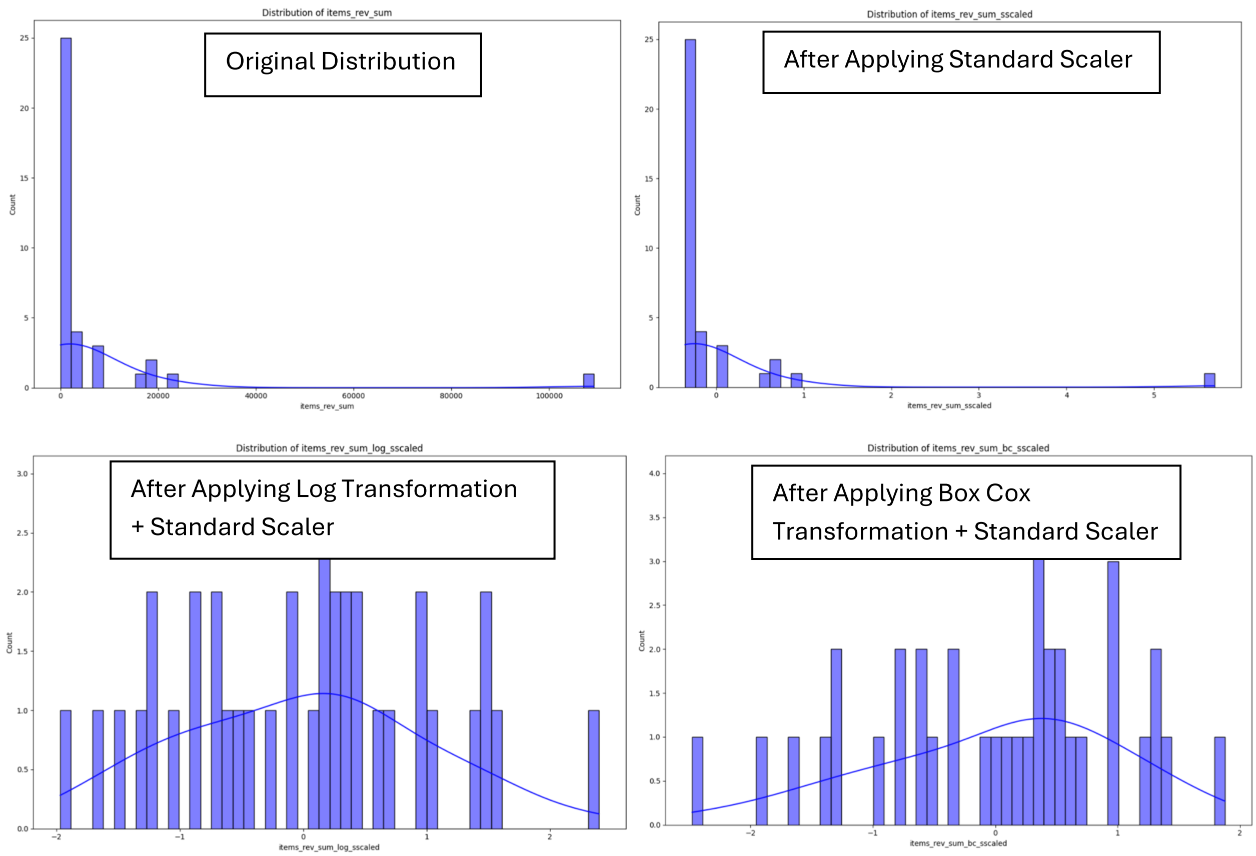


Figure 18 Distribution of Feature "items_rev" after removing zeroes, Pre and Post Transformation

Numerous combinations of clusterings were tested to better understand the data, resulting in the selection of two primary models. The first model includes all rows, including those where items_rev = 0, while the second model considers only rows where items_rev > 0. These models were chosen because they provide a clearer explanation of the data.

It should be noted that as the results of the second model are very similar in nature to the first model, it was decided to just showcase the first model in this report.

10.5.3 Correlation Check

The features available from the initial dataset are items_viewed, items_cart, items_chk, items_pur, and items_rev. Other columns are a derivation of the columns such that any underlying relationships can be detected by the model.

A correlation check was done on the different features that are to be put into the model.

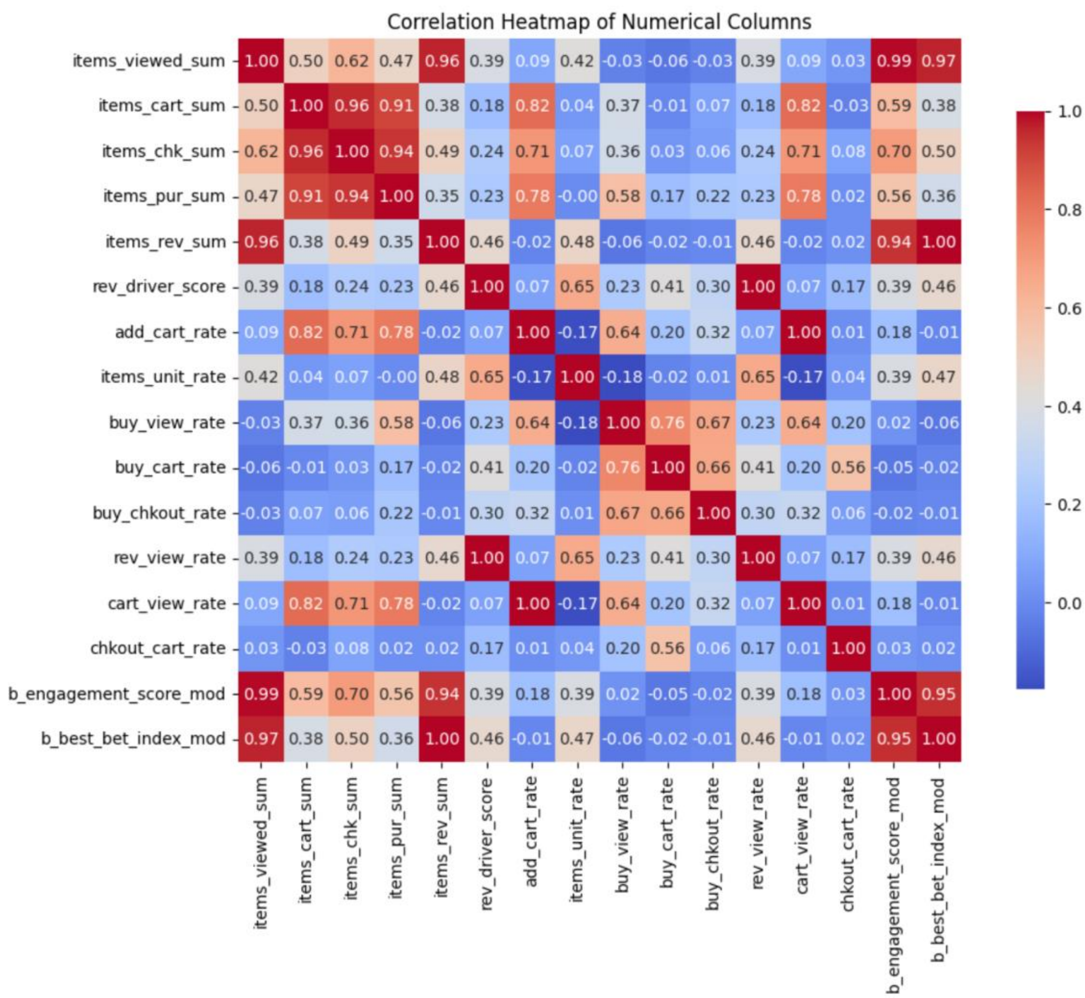


Figure 19 Correlation Heatmap for Features

A correlation check on the standardized form of the features was also conducted. The thought behind was that the model will be fed the standardized form of features and as such needs to be checked.

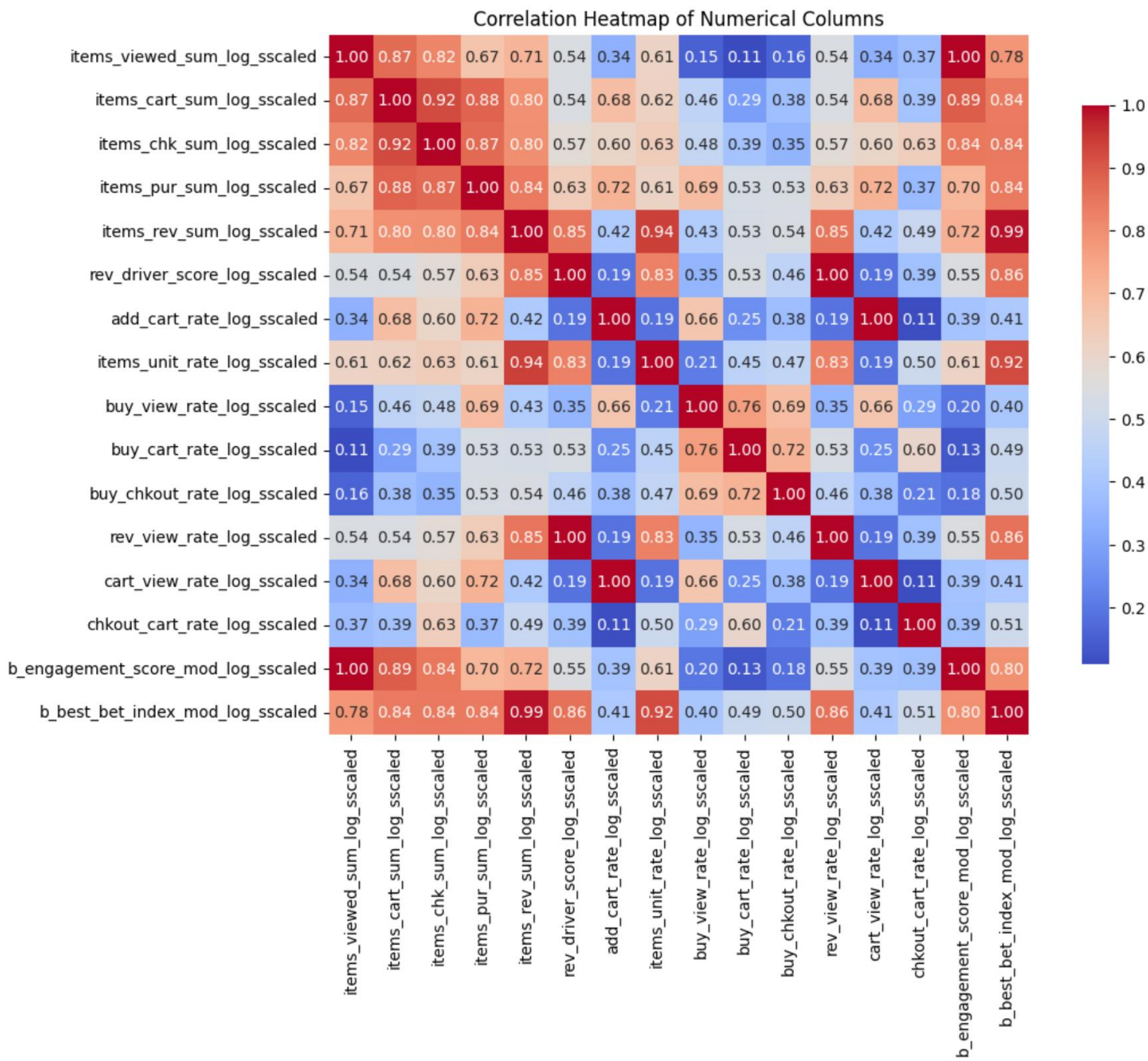


Figure 20 Correlation Heatmap for Features After Transformation (Log Scaled + Standard Scaled)

Since the standardized features have a higher correlation in general, it was decided to review the standardized features directly. Initially, we wanted to remove those where the correlation is more than 0.8, which include key features such as items_viewed, items_chk, items_pur, items_rev. However, considering that these key features hold significant importance, it was decided to keep those key features and remove the remaining features with a correlation of more than 0.9 instead.

After removing the other strongly correlated features, we are left with the below features.

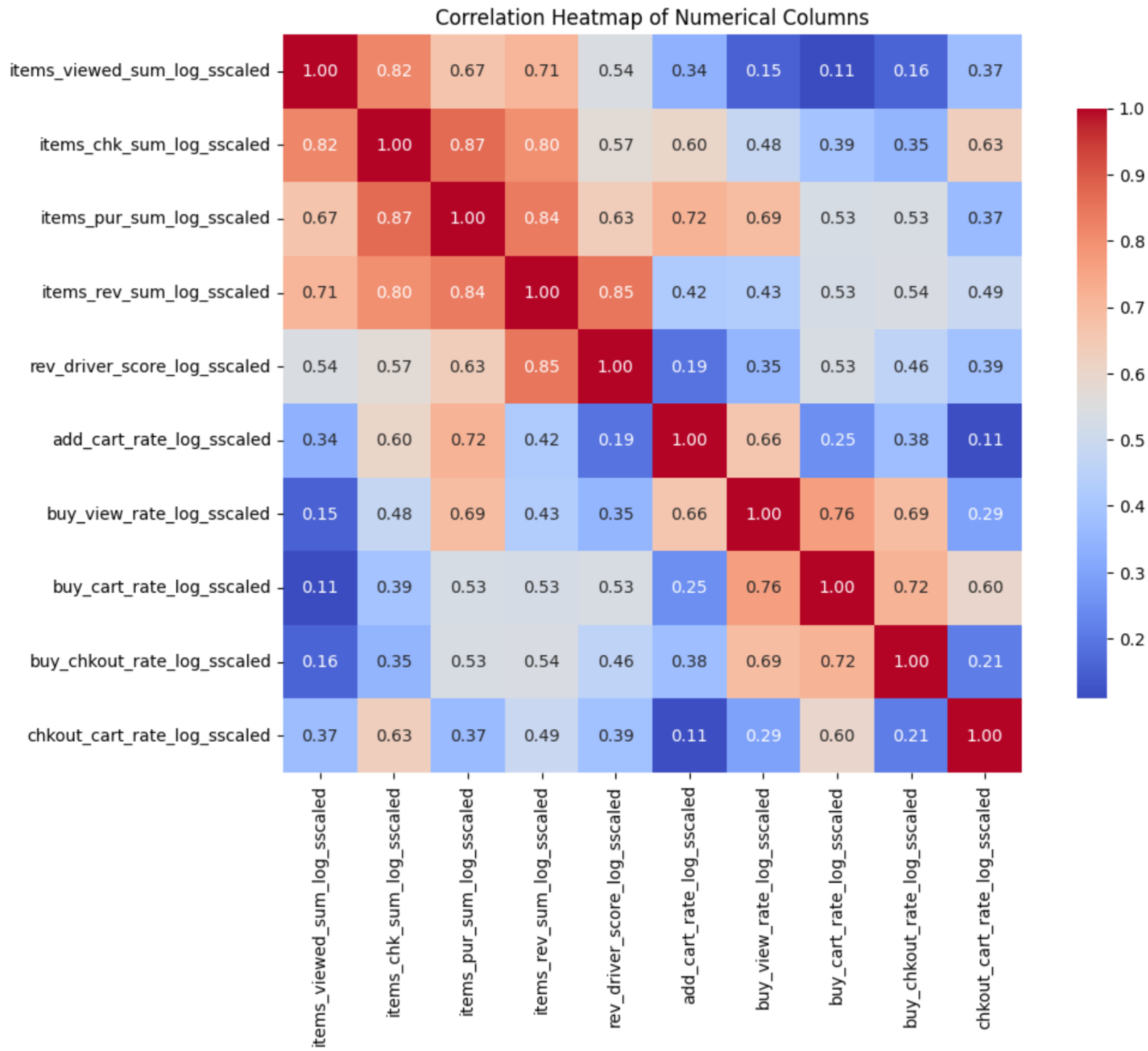


Figure 21 Correlation Heatmap For Features After Transformation (Log Scaled + Standard Scaled) After Selection

These features will constitute the foundational set for the experiments. Various combinations will be explored and incorporated into the model to achieve optimal clustering results and derive meaningful insights.

10.5.4 Modeling Parameters

Clustering Model 1 used the dataset where all items are included including where items_rev = 0. Columns used for Clustering Model 1 is as below:

1) items_viewed_sum_log_sscaled
2) items_chk_sum_log_sscaled
3) items_chk_sum_log_sscaled
4) items_chk_sum_log_sscaled
5) items_chk_sum_log_sscaled
6) add_cart_rate_log_sscaled
7) buy_view_log_sscaled
8) items_chk_sum_log_sscaled
9) items_chk_sum_log_sscaled
10) items_chk_sum_log_sscaled

For the Model Hyperparameters, it was set as below:

- No. of Clusters: Range between 2 to 7
- Max Iteration: Default of 300. (Tried 500, 750, 1000, with no significant differences in the result and centroids)
- Algorithm: Lloyd's
- Random State: 88

10.5.5 Modeling Technical Validation

For the technical evaluation, Silhouette and Inertia was used. There were a few options of different distance calculations to use. Euclidean distance was used due to:

- No Categorical features. All features used are continuous features.
- Low to mid-low dimensional data (10 features only)

Below is the chart for the Silhouette score and Inertia

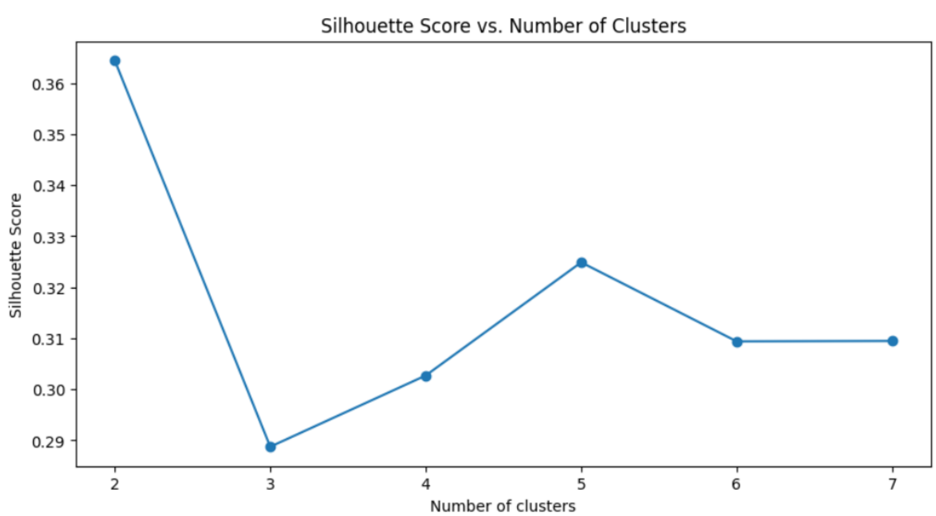


Figure 22 Cluster Model 1 Silhouette Score Chart

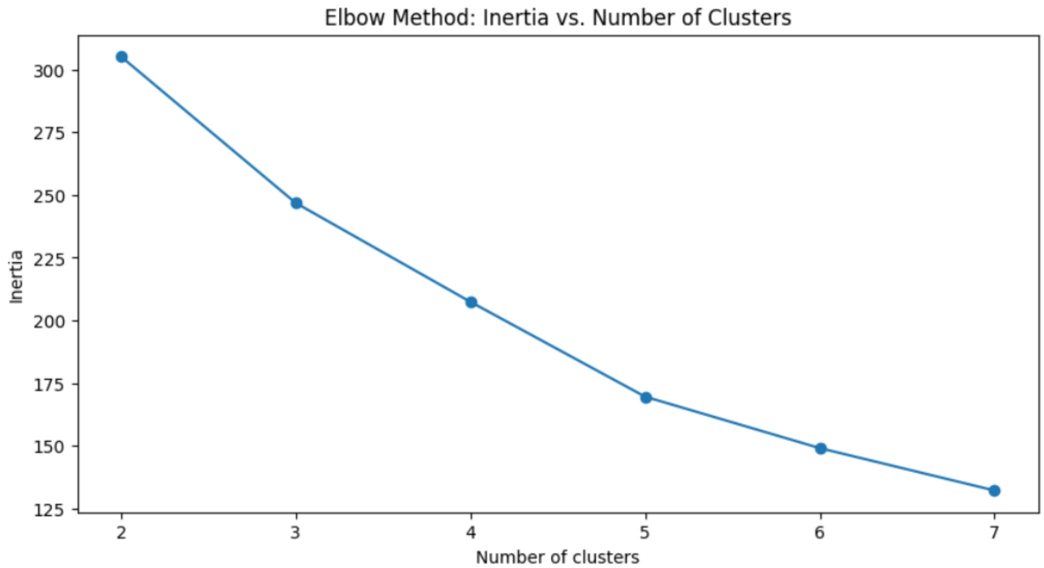


Figure 23 Cluster Model 1 Elbow Method Chart

The initial determination of the optimal number of clusters for the KMeans clustering was based on the Silhouette score and the Elbow method (inertia vs. number of clusters chart). The Silhouette score chart exhibits an unusual pattern, with the highest score at 2 clusters, followed by a sharp decline, and a minor peak at 5 clusters (which is still lower than the score at 2 clusters). The Elbow method also suggests an optimal number of 5 clusters, although the elbow is not very pronounced.

The unusual pattern in the Silhouette score chart can occur due to several reasons. The Silhouette score measures how similar each point is to its own cluster compared to other clusters. High scores indicate well-defined clusters, while low scores suggest overlapping or poorly defined clusters. In this case, the highest score at 2 clusters suggests that the data is most cohesively grouped into two clusters. The sharp decline and minor peak at 5 clusters may indicate that while 5 clusters provide some structure, they are not as well-defined as the 2-cluster solution. This behavior is expected given that the data is from an SME and does not have the volume typically seen in big data. Such datasets often exhibit more variability and irregular patterns, which can affect clustering results.

10.5.6 Modeling Business Validation (Profiling)

After evaluating the Silhouette score and the Elbow method, the profiles of the different clusters were examined to gain a deeper understanding of the data. Various clustering configurations were tested, specifically n = 4, 5, and 6 clusters, and their profiles were evaluated. Exploratory Data Analysis (EDA) was conducted, including plotting charts to understand and profile the data. Based on these analyses, it was concluded that the 5-cluster model is the most reasonable and aligns best with our understanding of the data.

On the left, is an example of evaluating the different clusters for different cluster models. There are a few interesting observations from the different clustering models that we can infer.

Clustering Model 1A (n=4)

	clus_0	clus_1	clus_2	clus_3
items_viewed_avg	961.81	871.00	318.71	4341.68
items_cart_avg	16.75	54.50	2.57	181.32
items_chk_avg	18.38	56.00	0.64	130.89
items_pur_avg	2.63	31.75	0.00	28.16
items_rev_avg	253.63	1174.75	0.00	12066.28
items_unit_rate_avg	120.39	51.03	0.00	838.22
rev_driver_score_avg	0.39	1.64	0.00	2.37

Table 18 Clustering Model 1A Profiling Results

Clustering Model 1B (n=5)

	clus_0	clus_1	clus_2	clus_3	clus_4
items_viewed_avg	1041.06	871.00	559.00	3489.00	5644.00
items_cart_avg	24.61	54.50	3.07	363.00	79.89
items_chk_avg	28.94	56.00	1.07	237.14	65.89
items_pur_avg	3.78	31.75	0.00	56.86	12.33
items_rev_avg	528.89	1174.75	0.00	5674.44	20452.92
items_unit_rate_avg	150.17	51.03	0.00	91.05	1612.44
rev_driver_score_avg	0.48	1.64	0.00	1.26	3.76

Table 19 Clustering Model 1B Profiling Results

Clustering Model 1C (n=6)

	clus_0	clus_1	clus_2	clus_3	clus_4	clus_5
items_viewed_avg	1001.08	871.00	318.71	4618.50	1925.33	4328.00
items_cart_avg	14.77	54.50	2.57	123.00	26.50	820.00
items_chk_avg	15.00	56.00	0.64	106.93	24.00	472.50
items_pur_avg	1.77	31.75	0.00	21.64	6.33	106.50
items_rev_avg	184.62	1174.75	0.00	12061.55	8565.72	5330.70
items_unit_rate_avg	122.46	51.03	0.00	524.68	1469.22	49.78
rev_driver_score_avg	0.32	1.64	0.00	1.35	4.28	1.24

Table 20 Clustering Model 1C Profiling Results

1) For clus_1 for the different models, all the average fields are the same.

2) For clus_2 for the different models, the average items_rev are 0. The items_viewed_avg is different though, which would mean that potentially some of the items where items_rev = 0 maybe in other clusters.

3) clus_3 from Clustering Model 1A, and clus_4 from Clustering Model 1B and 1C, have a much higher items_unit_rate_avg, then the other clusters. However, in Model 1A, the

items_pur_avg for clus_3 is just marginally lesser than clus_1; whereas in Model 1B and 1C, clus_4 has an items_pur_avg which is much lower than the remaining clusters. A discussion with the business teams have also mentioned that it is common for items with a higher unit rate to have less number of purchases.

4) clus_0 for the different clustering models, has the second lowest average rev_driver_score, with the lowest being the cluster with no revenue. For context, the rev_driver_score is a derivation of items_rev divide by items_viewed. A high rev_driver_score means that there is higher revenue per page view.

5) For all models, the items_cart_avg and the items_chk_avg seems to be very close to each other, except for Clustering Model 1C clus_5. This is expected as the correlation check earlier revealed that items_cart_avg and items_chk_avg had a strong correlation of over 0.9.

In Point 3, Model 1A has shown to be less accurate as the distinction between the items_pur_avg is not as obvious as in Model 1B and 1C. In Model 1C, there is some overlap between the clusters and the discernation of the averages are also not as clear as in Model 1B. For example, in Model 1C, clus_3 and clus_5 have similar items_viewed_avg; clus_0 and clus_4 have similar items_pur_avg; clus_1 and clus_5 have similar items_unit_rate_avg.

As such, based on the Silhouette score, the Elbow method, and the profiling done as above, it was decided that we will deep-dive using Model 1B.

Clustering Model 1B (n=5)

	clus_0	clus_1	clus_2	clus_3	clus_4	Overall Mean
items_viewed_avg	1041.06	871.00	559.00	3489.00	5644.00	1996.74
items_cart_avg	24.61	54.50	3.07	363.00	79.89	74.85
items_chk_avg	28.94	56.00	1.07	237.14	65.89	56.87
items_pur_avg	3.78	31.75	0.00	56.86	12.33	13.28
items_rev_avg	528.89	1174.75	0.00	5674.44	20452.92	4490.87
items_unit_rate_avg	150.17	51.03	0.00	91.05	1612.44	340.69
rev_driver_score_avg	0.48	1.64	0.00	1.26	3.76	1.09

Table 21 Clustering Model 1B with Overall Mean

As a side note, these profiling results are very similar to Clustering Model 2 (where Item revenue > 0), except that there was no clus_2. It is a major reason why the team decided to focus on Clustering Model 1B and not show Clustering Model 2 within this report.

The profiling was explored with relevance to:

1) the overall dataset mean (benchmarking) and
2) across the cluster values (relative differences).

For the comparison with the overall mean, values below the mean was labelled as Low and values above the mean was labelled as high. For the comparison across cluster values, clus_2 was ignored (as it is the lowest in all categories), and values below the median was labelled as Low, and values above the median was labelled as High.

Clustering Model 1B (n=5)	As Compared with Overall Mean				
	clus_0	clus_1	clus_2	clus_3	clus_4
items_viewed_avg	Low	Low	Low	High	High
items_cart_avg	Low	Low	Low	High	High
items_chk_avg	Low	Low	Low	High	High
items_pur_avg	Low	High	Low	High	High
items_rev_avg	Low	Low	Low	High	High
items_unit_rate_avg	Low	Low	Low	Low	High
rev_driver_score_avg	Low	High	Low	High	High

Table 22 Clustering Model 1B clusters as compared with overall mean

Clustering Model 1B (n=5)	Compared Among Clusters Excl. clus_2				
	clus_0	clus_1	clus_2	clus_3	clus_4
items_viewed_avg	Low	Low	NA	High	High
items_cart_avg	Low	Low	NA	High	High
items_chk_avg	Low	Low	NA	High	High
items_pur_avg	Low	High	NA	High	Low
items_rev_avg	Low	Low	NA	High	High
items_unit_rate_avg	High	Low	NA	Low	High
rev_driver_score_avg	Low	High	NA	Low	High

Table 23 Clustering Model 1B clusters as compared among individual clusters excluding clus_2

From the two different approaches, most had the same results except for:

1) clus_0 items_unit_rate_avg
2) clus_3 rev_driver_score_avg
3) clus_4 items_pur_avg

To simplify the explanation and categorization, the relative differences method will be used. The profiling will be done by comparing among the clusters.

For the profiling analysis, clus_2 will be ignored for now as the cluster has no purchase amount or revenue.

After consultation, the business agreed to prioritize the Revenue Driver Score, followed by other metrics. The observations below were obtained when comparing the different metrics to the revenue driver score.

		Page Views	
		High	Low
Revenue Driver	High	clus_4	clus_1
	Low	clus_3	clus_0

		Cart	
		High	Low
Revenue Driver	High	clus_4	clus_1
	Low	clus_3	clus_0

		Purchase Amount	
		High	Low
Revenue Driver	High	clus_0	clus_4
	Low	clus_3	clus_0

		Revenue	
		High	Low
Revenue Driver	High	clus_4	clus_1
	Low	clus_3	clus_0

		Unit Rate	
		High	Low
Revenue Driver	High	clus_4	clus_1
	Low	clus_0	clus_3

Figure 24: Clustering Model 1B Clustering Metrics Comparison - Revenue Driver

There are 3 observations regarding the metrics:

1) Revenue driver vs (Pageviews, Cart, Revenue) have the same profile.
2) Revenue driver vs Purchase Amount has a difference and swap for clus_1 and clus_4.
3) Unit Rate and Purchase Amount seem to be inverse. High unit rate clusters will be in the low purchase amount clusters.

There are also some observations on the clusters:

1) Clus_4 is a strong Revenue driver but has high unit rate and low purchase amount.
2) Clus_1 is a strong Revenue driver but has low unit rate and high purchase amount.
3) Clus_0 is a weak Revenue driver, has everything low, but a high unit rate.

Below are the scatter plots showing how the revenue driver score fares against the items viewed. The plots are separated out for better clarity. Clus_0 and clus_2 which is low performing is shown on one plot, and clus_1, clus_3, clus_4 is shown on another plot for further analysis. The labels are only shown for specific categories from clus_3 and clus_4 that are in the higher RDS score range or Items Viewed range.

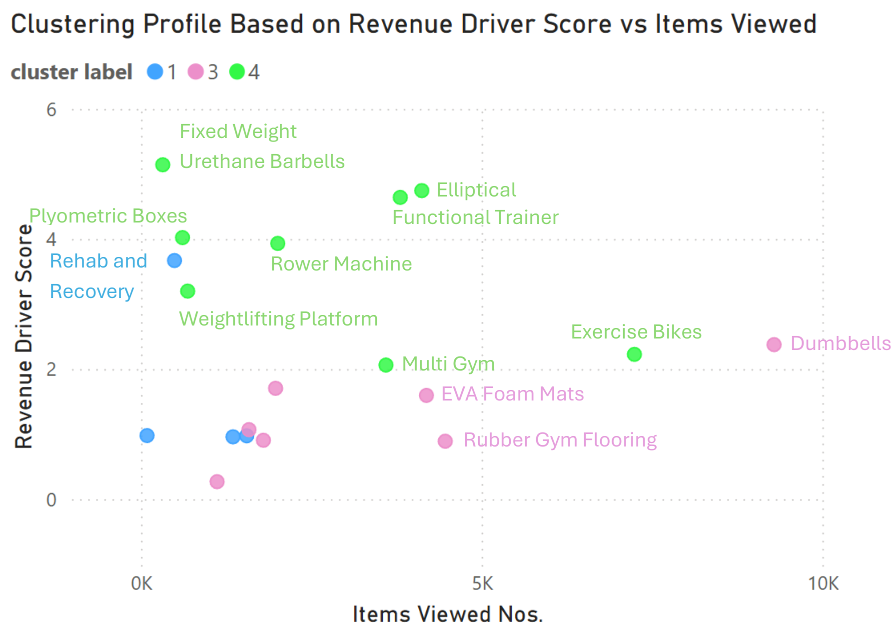


Figure 25 Clustering Model 1B Scatter Plot between RDS vs Item Viewed (clus_1, clus_3, clus_4)

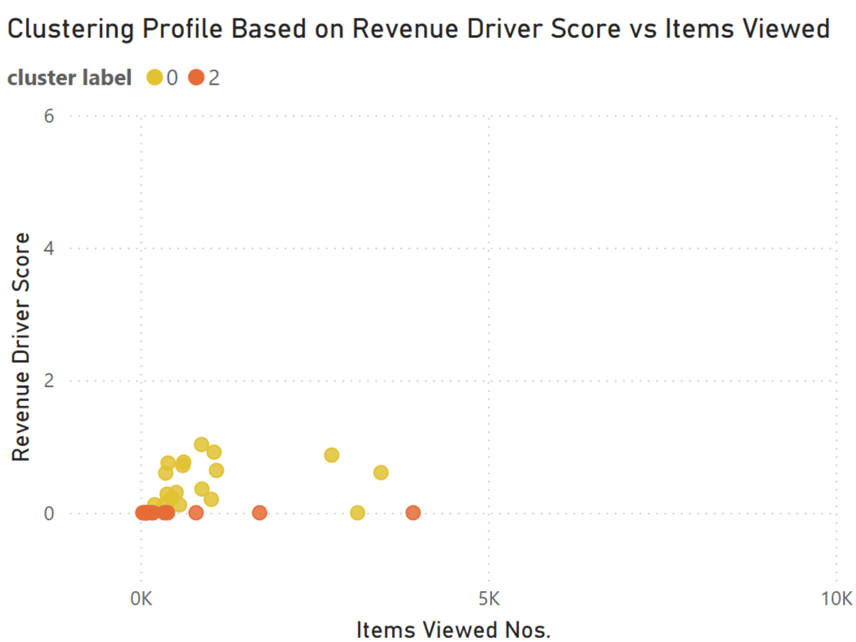


Figure 26: Clustering Model 1B Scatter Plot between RDS vs Item Viewed (clus_0, clus_2)

Based on the discussion with the business, clus_4 which is a high revenue driver with high page views is something the business should continue to thrive on, and clus_0 which has a low revenue driver with low page views should be phased out. The subsequent analysis will be on whether the business should focus on clus_1 or clus_3 to improve their sales performance.

To choose between clus_1 and clus_3, analysis was done on the purchase amount and the other metrics.

		Page Views	
		High	Low
Purchase Amount	High	clus_3	clus_1
	Low	clus_4	clus_0

		Revenue	
		High	Low
Purchase Amount	High	clus_3	clus_1
	Low	clus_4	clus_0

		Unit Rate	
		High	Low
Purchase Amount	High		clus_1 clus_3
	Low	clus_0 clus_4	

Figure 27: Clustering Model 1B Clustering Metrics Comparison - Purchase Amount

When looking at the purchase amount versus various metrics, focus should be put on clus_3 as it was observed that clus_3 was in the quadrant with high page views, high revenue but low unit rate.

		Cart	
		High	Low
Page Views	High	clus_4 clus_3	
	Low		clus_0 clus_1

		Purchase Amount	
		High	Low
Cart	High	clus_3	clus_4
	Low	clus_1	clus_0

Figure 28: Clustering Model 1B Clustering Metrics Comparison - Other Metrics

When looking at page views to cart, and cart to purchase amount, clus_3 was also observed to be in the high-high quadrant. As the metrics quadrant above doesn't really indicate the rate, but rather individual metrics, the below table was introduced that checks the average rate of the clusters based on the flow. (View -> Cart -> Checkout - > Purchases).

	Average Values			
Cluster	Cart to View Ratio (Cart / Views)	Checkout to Cart Ratio (Checkout / Cart)	Purchase to Checkout Ratio (Purchases / Checkout)	Revenue Driver Score (Revenue / Page Views)
0	0.027	1.301	0.207	0.479
1	0.053	1.159	0.652	1.645
2	0.006	0.144	0.000	0.000
3	0.097	0.691	0.307	1.258
4	0.018	0.904	0.264	3.756

Table 24: Clustering Model 1B Metrics Comparison

For this table, the revenue driver score was used instead of the revenue or unit rate (revenue / purchases) as the quantification is of a different nature. Revenue is in SGD, whereas the others are in terms of numbers.

From the table, focusing specifically on clus_1 and clus_3, the Cart to View ratio for clus_3 is much higher than clus_1. In fact, it has the highest cart to view ratio among the other clusters. This might imply that this category has the highest percentage of customers viewing these items and intending to buy it by carting it.

For the checkout to cart ratio and purchase to checkout ratio, clus_1 is much higher than clus_3, which might imply that customers are interested in the item but hesitate more to finalize the purchase. Strictly speaking, checkout to cart ratio should be less than 1 if customers are only allowed to checkout after carting the items. However, check out is allowed on most platforms before carting.

Below chart shows how clus_3 has better engagement than clus_1 thru the spread of the cart to view ratio. Clus_3 has either similar or higher cart to view ratio as clus_1, and also much higher item views.

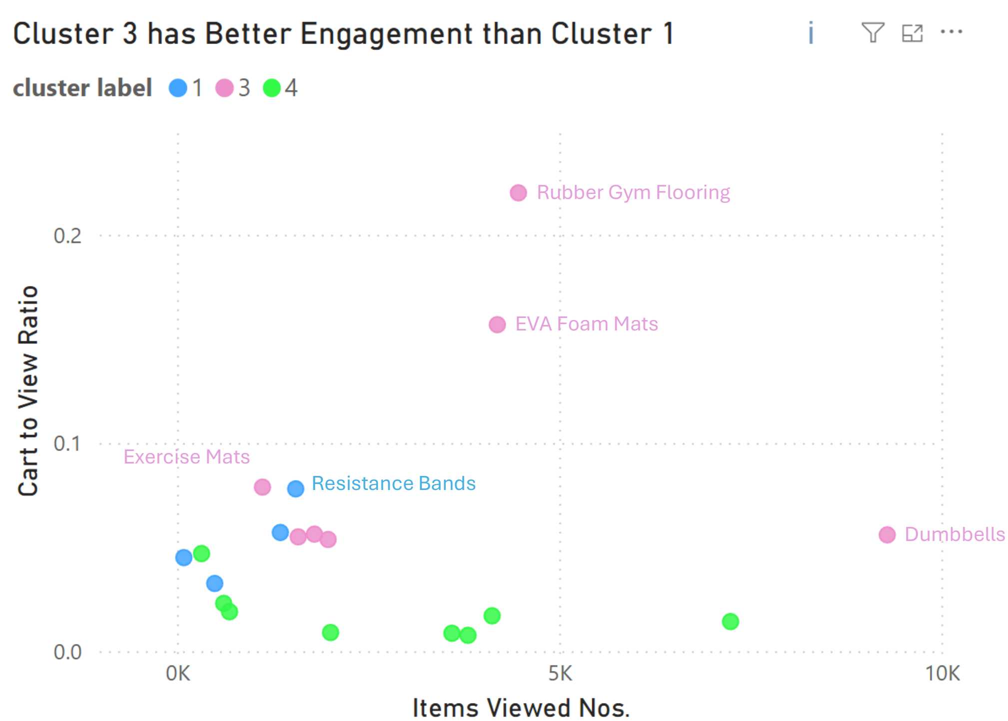


Figure 29: Clustering Model 1B Scatter Plot between Cart to View Ratio vs Item Viewed (clus_1, clus_3, clus_4)

The team recommends for the company to focus on clus_3 instead of clus_1 from Clustering Model 1B, due to the below reasons:

1) Higher number of items_viewed, items_cart, items_chk, items_pur, items_rev. This implies that clus_3 items are already performing better than clus_1, in terms of current exposure, and sales.
2) Clus_3 has a much higher cart to view ratio (>80%) as compared to clus_1, or any other product. This would imply that clus_3 items have a higher proportion of customers / viewers who are already engaged and interested in the products. The initial engagement of customers in clus_3 should be leveraged to improve the checkout and purchase conversion rates.
3) Even if clus_1 has a higher checkout to cart ratio or purchase to checkout ratio, the initial lack of engagement is a concern. The lower initial interest in clus_1 means that increasing views might not translate into a proportional increase in cart to view ratio, which are crucial for driving sales. Therefore, focusing on clus_3, where interest is already established, is a more strategic approach for optimizing the sales funnel and achieving higher overall sales.

The team was able to summarize and profile the clusters with the relevant action plan.

Cluster	Profile			Action
	Engagement	RDS	Item Price	
0	Low	Low	Expensive	To phase out.
1	Low	High	Affordable	To phase out.
2	Low	None	NA	To phase out.
3	High	Low	Affordable	To focus on increasing conversion for this cluster by means of discounts or bundling.
4	High	High	Expensive	Continue to focus sales on this sector.

Table 25 Clustering Model 1B Final Profiling

Based on this Clustering Profiles, the subsequent research will be done on the products or items from individual clusters and recommend accordingly.

10.6 Category and Product Analysis

10.6.1 Category and Product Analysis – Cluster 3

Focusing on clus_3, we can observe that when reviewing cart to view ratio and revenue driver score, 3 categories stand out. Dumbbells, EVA Foam Mats, and Rubber Gym Flooring. These 3 categories have significant views which suggest sufficient exposure, and good cart to view ratios which suggest customers are keen to get the products.

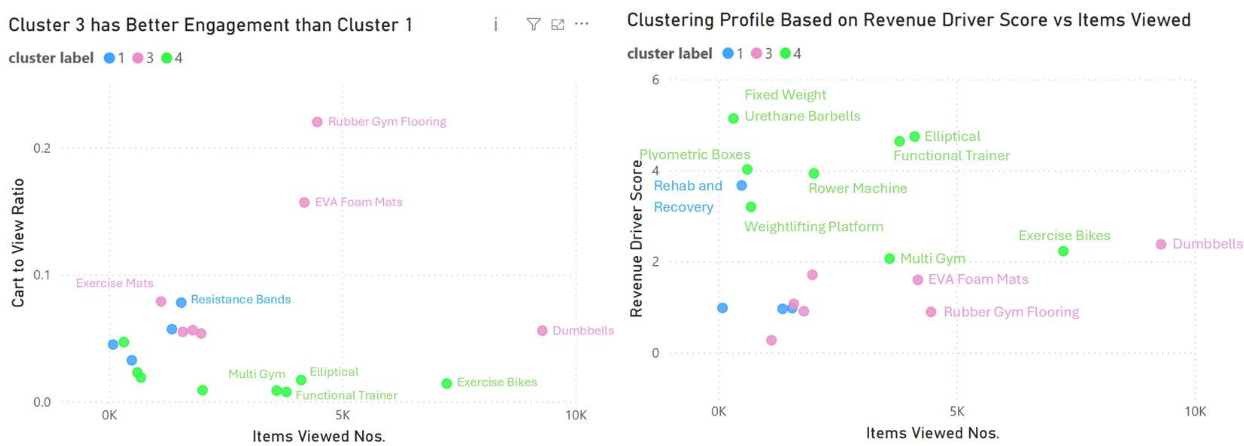


Figure 30: Cluster Profile shown previously

For Dumbbells, Urethane Round Dumbbells and 12-Sided Urethane Dumbbells have high cart to view ratios or revenue driver score. The rubber hexagonal dumbbell set with 3-tier rack seems to have a very high revenue driver score but very low cart to view ratio, which is not ideal.

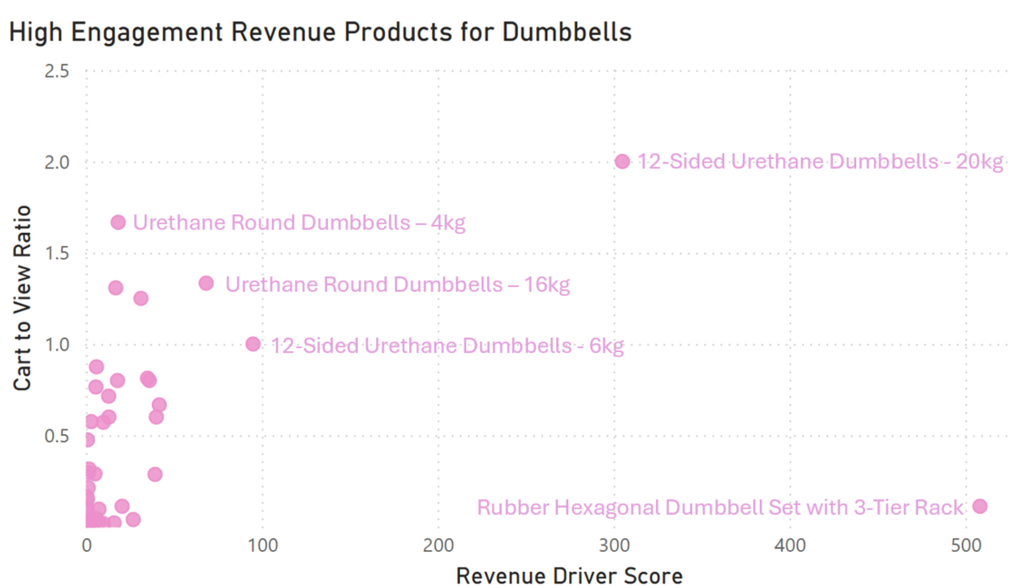


Figure 31: High Engagement Revenue Products for Dumbbells

For EVA Foam Mats, Martial Arts Eva Foam Tatami Mats – Red & Blue – 40mm has either a high cart to view ratio, or the Sensei Sports version has a high revenue driver score. The 50mm version seems to have a high revenue driver score but has a very low almost 0 cart to view ratio. The Grey EVA Foam Mat has a low cart to view ratio but is the third highest revenue driver score in the category.

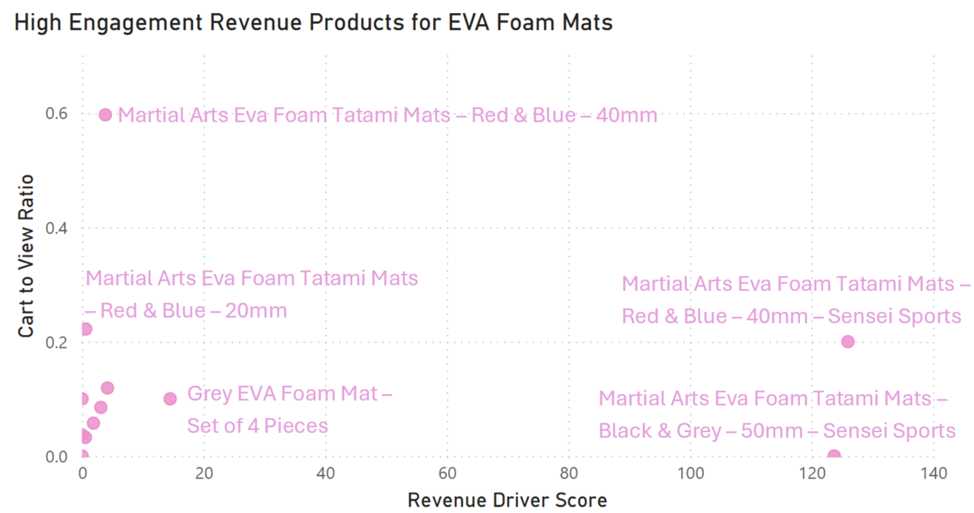


Figure 32: High Engagement Revenue Products for EVA Foam Mats

For Rubber Gym Flooring, most of the products have either a very low revenue driver score, or a very low cart to view ratio. Only the Commercial Gym Tiles – 20mm seems to have a good balance of Revenue Driver Score and Cart to View ratio.

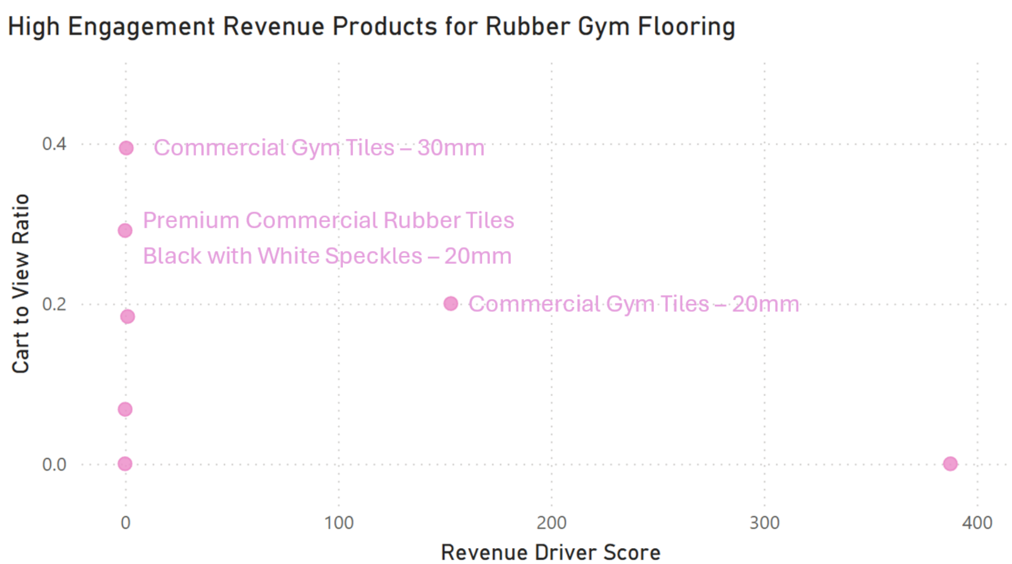


Figure 33: High Engagement Revenue Products for Rubber Gym Flooring

Based on the study of these 3 categories, the team would recommend these products for cluster 3:

Rank	Product	Reason
1	12-Sided Urethane Dumbbells – 20kg (RDS > 300, Cart to View > 2)	Highest RDS + Cart to View for whole cluster.
2	Commercial Gym Tiles – 20mm (RDS > 300, Cart to View > 0.2)	High RDS + Bare Minimum Cart to View
3	Martial Arts Eva Foam Tatami Mats – Red & Blue – 40mm – Sensei Sports (RDS > 120, Cart to View > 0.2)	High RDS + Bare Minimum Cart to View
4	12-Sided Urethane Dumbbells – 6kg (RDS > 90, Cart to View > 1)	Medium RDS + High Cart to View
5	Urethane Round Dumbbells – 16kg (RDS > 60, Cart to View > 1.3)	Medium RDS + High Cart to View
6	Urethane Round Dumbbells – 4kg (RDS > 15, Cart to View > 1.6)	Low RDS + High Cart to View

Table 26: Product Recommendation for Cluster 3

The recommendation focuses on RDS as a priority, followed by those with high Cart to View ratio as a secondary priority.

10.6.2 Category and Product Analysis – Cluster 4

Focusing on clus_4, we can observe that when reviewing cart to view ratio and revenue driver score, 4 categories stand out. Exercise Bikes, Elliptical, Functional Trainer, and Multi Gym. Despite Fixed Weight Urethane Barbells showing up as having the highest RDS, it was not considered due to the low views.

Figure 34: Cluster Profile shown previously

For Exercise Bikes, most of the products have relatively low cart to view ratio and low revenue driver score. Only 1 product stands out within the exercise bikes category, namely Sole LCR Exercise Recumbent Bike. The Sole SB1200 Spin Bike and Sole LCB Upright Exercise Bike have high cart to view ratio but very low revenue driver score, making them not an option.

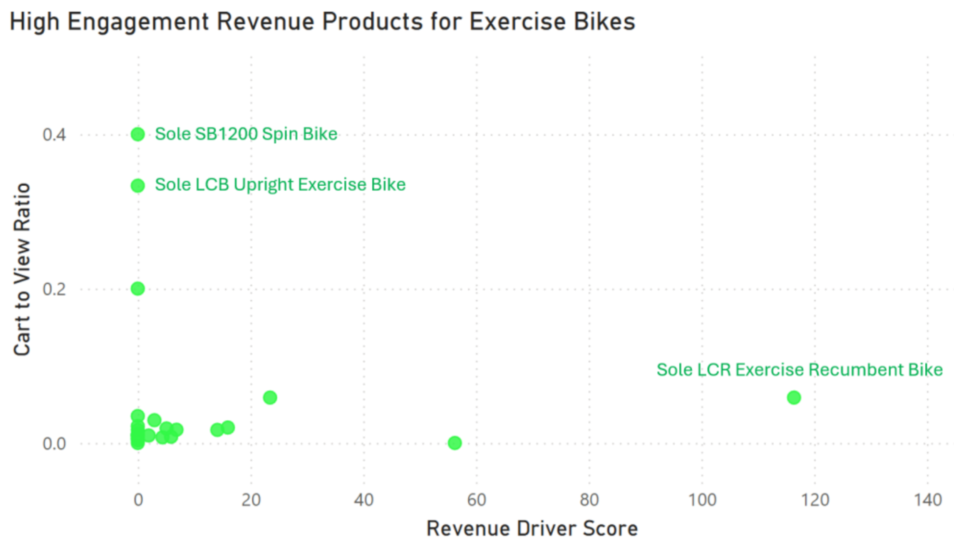


Figure 35: High Engagement Revenue Products for Exercise Bikes

For Ellipticals, similar to the exercise bikes, the Sole E95 Elliptical Cross Trainer has very low revenue driver score. As such, the only product in consideration are the Xterra FS3.0 Elliptical Cross Trainer, and the Sole E95 Elliptical Cross Trainer with Touch Screen.

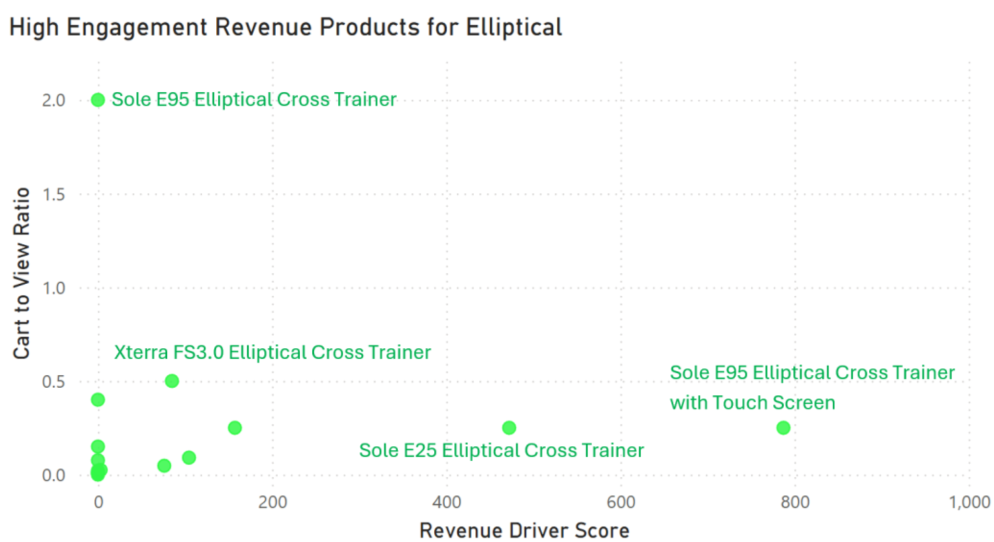


Figure 36: High Engagement Revenue Products for Ellipticals

For Functional Trainer, there are two more prominent products, Inspire FTX Functional Trainer, and Inspire FT2 Functional Trainer. Inspire FT1 Functional Trainer was not considered as it has a lower revenue driver score and cart to view ratio than the Inspire FT2 Functional Trainer.

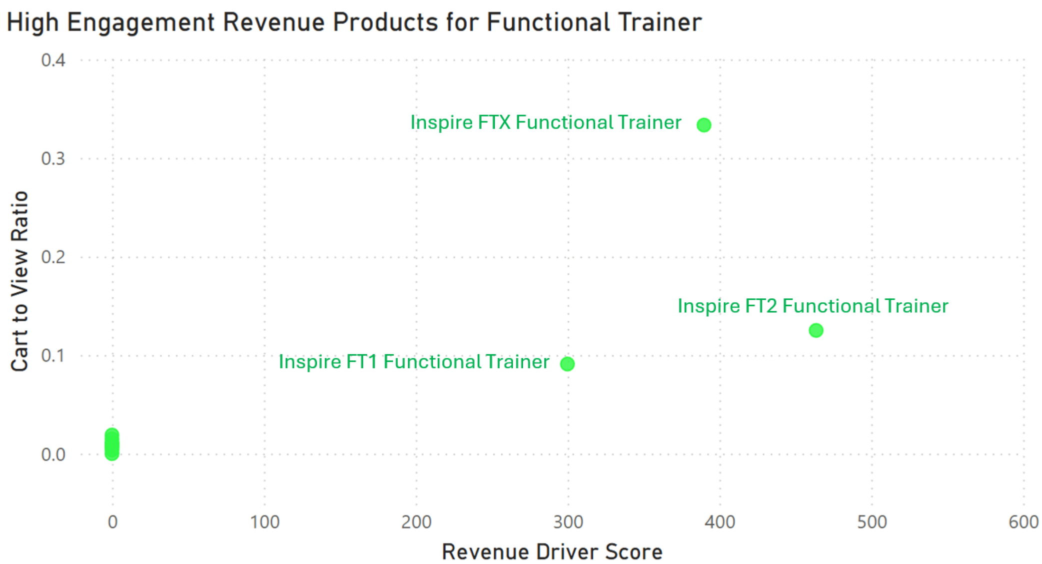


Figure 37: High Engagement Revenue Products for Functional Trainer

For Multi Gym, the two products that are more significant are the Inspire Multi Gym M2S and Inspire Multi Gym M2.

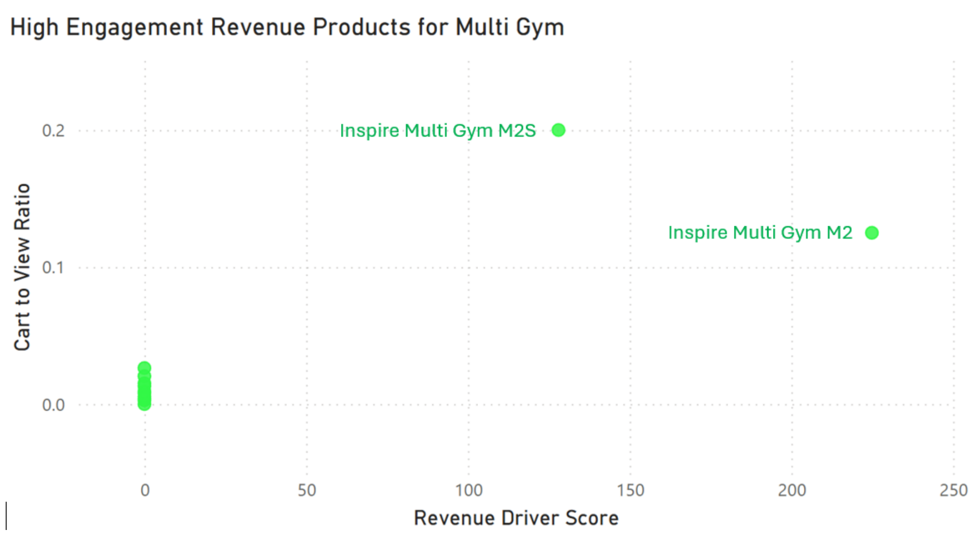


Figure 38: High Engagement Revenue Products for Multi Gym

Based on the study of these 4 categories, the team would recommend these products for cluster 4:

Rank	Product	Reason
1	Sole E95 Elliptical Cross Trainer with Touch Screen (RDS > 775, Cart to View > 0.25)	Highest RDS + Average Cart to View
2	Inspire FT2 Functional Trainer (RDS > 450, Cart to View > 0.1)	High RDS + Bare Minimum Cart to View
3	Inspire FTX Function Trainer (RDS > 375, Cart to View > 0.3)	High RDS + High Cart to View
4	Xterra FS3.0 Elliptical Cross Trainer (RDS > 85, Cart to View > 0.5)	Low RDS + Highest Cart to View for whole cluster

Table 27: Product Recommendation for Cluster 4

The recommendation focuses on RDS as a priority, followed by those with high Cart to View ratio as a secondary priority. The comparison happens within the cluster and not among the different clusters. As such, the benchmark of it being high or low, is dependent on the values from the cluster.

Inspire Multi Gym M2S, Inspire Multi Gym M2, and Sole LCR Exercise Recumbent Bike, was not included as part of the recommendation, due to not having an RDS higher than 250, and / or not having an above average Cart to View ratio.

10.7 Product Ranking (among different Categories / Clusters)

The consolidated product ranking according to RDS as the main priority is as below.

Rank	Product	Cluster	Category
1	Sole E95 Elliptical Cross Trainer with Touch Screen	Cluster 4	Elliptical
2	Inspire FT2 Functional Trainer	Cluster 4	Functional Trainer
3	Inspire FTX Functional Trainer	Cluster 4	Functional Trainer
4	12-Sided Urethane Dumbbells – 20kg	Cluster 3	Dumbbells
5	Commercial Gym Tiles – 20mm	Cluster 3	Rubber Gym Flooring
6	Martial Arts Eva Foam Tatami Mats – Red & Blue – 40mm – Sensei Sports	Cluster 3	EVA Foam Mat
7	12-Sided Urethane Dumbbells – 6kg	Cluster 3	Dumbbells
8	Xterra FS3.0 Ellipticial Cross Trainer	Cluster 4	Elliptical
9	Urethane Round Dumbbells – 16kg	Cluster 3	Dumbbells
10	Urethane Round Dumbbells – 4kg	Cluster 3	Dumbbells

Table 28: Product Ranking among Different Categories and Clusters

10.8 Seasonality Check

To check whether a category has stable revenue stability, the team decided to use a combination of the coefficient of variation and a simple check of how many months of revenue was obtained.

The coefficient of variation uses the formula as below:

$$Coefficient\ of\ Variation\ (CV) = \frac{Standard\ Deviation\ of\ Revenue\ within\ Category}{Mean\ of\ Revenue\ within\ Category}$$

For the mean and standard deviation, we included the values where revenue = 0 as that would be representative of the actual data and distribution.

As getting the CV itself isn't usually representative of the actual distribution, it was decided to use a simple check of how many months of revenue was positive over the period of one year.

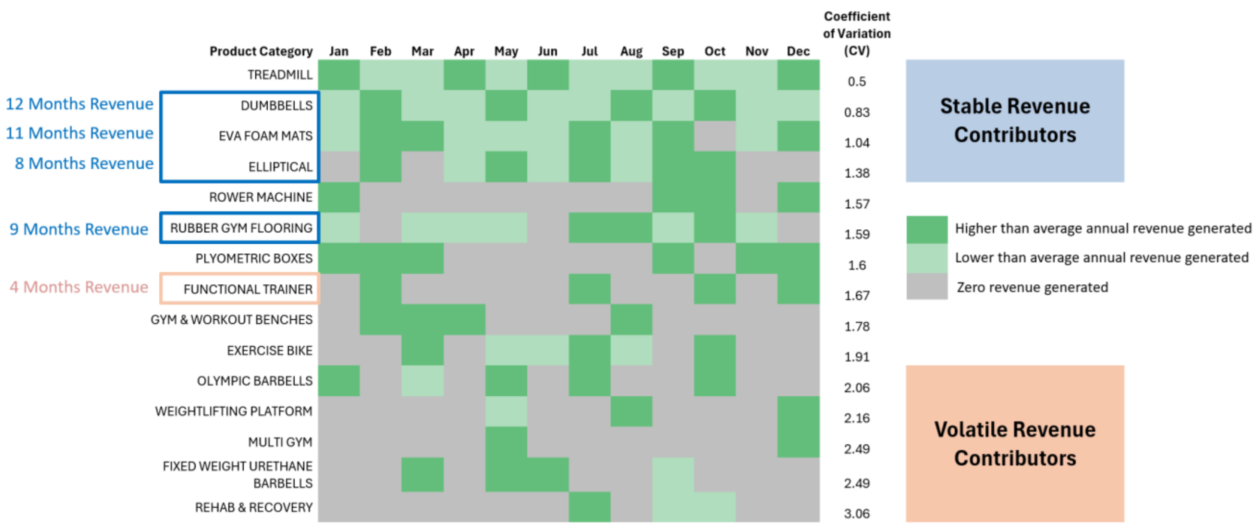


Figure 39: Revenue Seasonality Patterns Across Product Categories

Some observations from the Figure 39 above:

1) Treadmill has the best CV of 0.5 and 12 out of 12 months had positive revenue.
2) The category with the least months with revenue (2 out of 12 months) is Multi Gym.
3) The category with the worst CV is Rehab

For product recommendations, it would be advisable to focus on products which are stable revenue contributors or have some form of stability.

Based on the revenue stability, we can produce the below table.

Previous Rank	Product	Cluster	Category	Revenue Stability
1	Sole E95 Elliptical Cross Trainer with Touch Screen	Cluster 4	Elliptical	8 out of 12 months
2	Inspire FT2 Functional Trainer	Cluster 4	Functional Trainer	4 out of 12 months
3	Inspire FTX Functional Trainer	Cluster 4	Functional Trainer	4 out of 12 months
4	12-Sided Urethane Dumbbells – 20kg	Cluster 3	Dumbbells	12 out of 12 months
5	Commercial Gym Tiles – 20mm	Cluster 3	Rubber Gym Flooring	9 out of 12 months
6	Martial Arts Eva Foam Tatami Mats – Red & Blue – 40mm – Sensei Sports	Cluster 3	EVA Foam Mat	11 out of 12 months
7	12-Sided Urethane Dumbbells – 6kg	Cluster 3	Dumbbells	12 out of 12 months
8	Xterra FS3.0 Ellipticial Cross Trainer	Cluster 4	Elliptical	8 out of 12 months
9	Urethane Round Dumbbells – 16kg	Cluster 3	Dumbbells	12 out of 12 months
10	Urethane Round Dumbbells – 4kg	Cluster 3	Dumbbells	12 out of 12 months

Table 29: Product Ranking (old) with Revenue Stability Factor

Based on the table, it can be seen that the worst performing revenue stability in this list is from the functional trainer category with just 4 out of 12 months. There was a consideration to remove it completely from the recommendation, but the team thought it would be worth it to test the response and see whether the revenue stability would improve after the enhancements to the product marketing strategy was carried out.

10.9 Conclusion for Business Problem 2

The analysis and recommendations presented in this report directly address Sole Fitness's two core challenges: mitigating over-reliance on treadmills and streamlining an overloaded product portfolio. By prioritizing high-potential non-treadmill products in Clusters 3 and 4, the business can diversify revenue streams while phasing out underperforming categories (Clusters 0, 1, and 2).

10.9.1 Final Product Rankings

The re-ranking will be done based on the top product for each category by revenue stability, followed by the remaining products by revenue stability. The final ranking is as below.

Old Rank	New Rank	Product	Cluster	Category	Revenue Stability
4	1	12-Sided Urethane Dumbbells – 20kg	Cluster 3	Dumbbells	12 out of 12 months
6	2	Martial Arts Eva Foam Tatami Mats – Red & Blue – 40mm – Sensei Sports	Cluster 3	EVA Foam Mat	11 out of 12 months
5	3	Commercial Gym Tiles – 20mm	Cluster 3	Rubber Gym Flooring	9 out of 12 months
1	4	Sole E95 Elliptical Cross Trainer with Touch Screen	Cluster 4	Elliptical	8 out of 12 months
2	5	Inspire FT2 Functional Trainer	Cluster 4	Functional Trainer	4 out of 12 months
7	6	12-Sided Urethane Dumbbells – 6kg	Cluster 3	Dumbbells	12 out of 12 months
9	7	Urethane Round Dumbbells – 16kg	Cluster 3	Dumbbells	12 out of 12 months
10	8	Urethane Round Dumbbells – 4kg	Cluster 3	Dumbbells	12 out of 12 months
8	9	Xterra FS3.0 Ellipticial Cross Trainer	Cluster 4	Elliptical	8 out of 12 months
3	10	Inspire FTX Functional Trainer	Cluster 4	Functional Trainer	4 out of 12 months

Table 30: Product Ranking (new) with Revenue Stability Factor

10.9.2 General Marketing Recommendations for Clusters

While the following recommendations provide general marketing guidance, the actual strategy will be developed by the business's marketing team, leveraging their expertise in fitness equipment and consumer behavior to create tailored campaigns based on the identified high-potential products.

Cluster 4 Products

(Premium Products with High Engagement, High RDS, Low Cart-to-View Ratios)

Cluster 4 products, such as ellipticals and functional trainers, generate significant revenue per view but face lower cart-to-view ratios, likely due to their premium pricing. To address this, the following strategies are proposed:

- Value-Centric Marketing Campaigns
  Develop content that emphasizes long-term value and durability, such as video testimonials or comparative guides (e.g., “Commercial vs. Home Fitness Equipment: Why Premium Matters”). This approach aligns with the psychology of high-investment purchases, where customers seek reassurance about quality and longevity.

- Reducing Friction in High-Value Purchases
  Introduce flexible payment options (e.g., installment plans) to lower upfront cost barriers. Additionally, bundle these products with complementary Cluster 3 items (e.g., gym flooring) to enhance perceived value and incentivize cart additions.

- Leverage Trust Signals
  Highlight certifications, warranties, or endorsements from fitness professionals or associations to alleviate purchase hesitations. For example, featuring certifications like “TÜV SÜD” or “NSCA (National Strength and Conditioning Association)” on product pages can reinforce credibility.

Cluster 3 Products

(Everyday Gym Products with High Engagement, Low RDS, High Cart-to-View Ratios)

Cluster 3 products, such as dumbbells and gym flooring, attract significant customer interest (high cart-to-view ratios) but require optimization to improve conversion (cart to purchase conversion) efficiency. Recommendations include:

- Conversion-Focused Promotions
  Implement limited-time discounts or bundled offers (e.g., “Buy 3 Dumbbells, Get 1 Mat Free”) to capitalize on existing engagement. Such tactics aim to convert browsing intent into purchases without eroding long-term pricing integrity.

- Checkout Process Optimization
  Simplify checkout steps and prominently display free shipping thresholds (e.g., “Spend $200 for Free Delivery”) to reduce cart abandonment. A streamlined user experience can directly translate to higher conversion rates.

- Retargeting and User-Generated Content
  Deploy retargeting ads for users who abandoned carts, paired with customer-generated content (e.g., social media posts of customers using the products) to build social proof and reignite interest.

10.9.3 Implementation Strategies

To operationalize the recommendations in alignment with Sole Fitness’s risk-averse operational philosophy, a phased, iterative approach is proposed. This methodology balances innovation with prudence, allowing the business to validate strategies on a small scale before committing significant resources. Below is a sample structured implementation roadmap:

Phase 1: Pilot Testing (Month 1 to 3)

- Conduct A/B Tests on two products each from Cluster 3 promotions and Cluster 4 payment plans.
- Set success metrics to grade. Eg. Sustained 10% increase in RDS over two consecutive months. or 15% increase in cart-to-view ratio in bundled offers.

Phase 2: Incremental Execution (Month 4 to 6)

- Reallocate resources in small increments from underperforming clusters to top-ranked products.
- Track existing resource and revenue for treadmills are not impacted negatively.
- Use customer feedback loops to refine strategies.

The two phases can be done iteratively and in a safe, risk-averse manner that the business is comfortable with going forward with.

11.0 Analysis for Business Problem 3

11.1 Background and Introduction

To recap, Business Problem 3 tackles the issue of ineffective and unoptimized advertising spend across platforms (Google Ads, Facebook Ads, and Bing Ads) due to the absence of data-driven performance metrics.

The scope of Business Problem 3 is focused on the sales that occur on sole fitness website (https://www.solefitness.sg) as the advertisements direct customers to their website. Advertisements are not done to direct sales to Shopee and Lazada shops.

Through early engagement and consultations with the Operations Manager, our team developed the following understanding of Sole Fitness's business operations.

Sole Fitness marketing channels are through Google Ads, Facebook Ads, and Bing Ads, which are pay-per-click digital advertising platforms that allow businesses to place ads on search engines and social media networks (Facebook/Instagram) through keyword bidding (ensuring website comes out on top of searches) and audience targeting,

Sole Fitness allocates a fixed monthly budget of 3600$ for their advertisements and keywords bidding on the 3 platforms. Please refer to below to Table 31 for an overview of the budget allocated for advertising to drive sales on the website.

MONTH	PLATFORM	BUDGET	AD SPENT
Jan-24	WEBSITE	$3,600	$3,185
Feb-24	WEBSITE	$3,600	$3,823
Mar-24	WEBSITE	$3,600	$3,020
Apr-24	WEBSITE	$3,600	$3,534
May-24	WEBSITE	$3,600	$3,376
Jun-24	WEBSITE	$3,600	$2,348
Jul-24	WEBSITE	$3,600	$3,844
Aug-24	WEBSITE	$3,600	$3,357
Sep-24	WEBSITE	$3,600	$2,594
Oct-24	WEBSITE	$3,600	$2,293
Nov-24	WEBSITE	$3,600	$3,456
Dec-24	WEBSITE	$3,600	$3,210

Table 31: Ad-spend over Months

As Sole Fitness is a relatively new business, the business only has data of the ad-spent data collated for the year 2024. The BUDGET and AD SPENT are generated from aggregated values of the raw data in Table 5. From the data in Table 5, it is possible to create aggregated overviews of the budget allocated and % spent. Below is an example of the Ad-Spent in January 2024.

MONTH	WEEK	PLATFORM	BUDGET	SPENT
Jan-24	1-Jan	OVERALL	$840	$737
	8-Jan	OVERALL	$840	$670
	15-Jan	OVERALL	$840	$658
	22-Jan	OVERALL	$840	$766
	29-Jan	OVERALL	$240	$353
TOTAL			$3,600	$3,185

Table 32: January 2024 Weekly Ad Spending

Initial analysis of Table 31 revealed that, a total allocated budget of $43,200, of which $38,040 was actually spent. Meaning a budget utilization rate of approximately 88%.

In addition, the proportion of advertisement spending over website revenue stands at significant 16%. Without having an analytical method to allocate ad-spend budget, there are concerns that the advertisement spend is not maximizing viewership and might lead to a wastage of resources.

Hence, to solve the business problem of Sole Fitness not having a clear metric for advertisement spending effectiveness, our business objective is to assist Sole Fitness in strategically realign advertising spend based on item performance. This will be done through correlation analysis between advertising spend and product views as well as developing a new advertising metric, Return on Advertisement Spending (ROAS) which will be defined in the following section for use in Sole Fitness.

11.2 Key Metrics for Advertising Spend

These key metrics are used in the upcoming sections.

11.2.1 Correlation Score Between Ad Spend & Item Revenue

To calculate correlation score between Ad Spend and Item Revenue, we will be utilizing the below formula:

$$Correlation\ Between\ Ad\ Spend\ \&\ Item\ Revenue, r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

Where:

- x_i *and* y_i *are the individual sample points of ad spend & item revenue*
- $\overline{x}$ *and* $\overline{y}$ *are the means of the ad spend and item revenue data points*

11.2.1 Return on Advertisement Spending (ROAS)

To calculate ROAS, we will be utilizing the below formula:

$$\mathrm{ROAS} = \frac{Revenue\ from\ Ad\ Spending}{Ad\ Spending}$$

$$\mathrm{ROAS} = \frac{Revenue\ With\ Ad\ Spending - Revenue\ Without\ Ad\ Spending}{Ad\ Spending}$$

$$\mathrm{ROAS} = \frac{Mean\ Weekly\ Revenue\ With\ Ad\ Spending - Mean\ Weekly\ Revenue\ Without\ Ad\ Spending}{Mean\ Weekly\ Ad\ Spending}$$

A good ROAS for Sole Fitness should be above 1, as it meant that Sole Fitness is effective, gaining more dollars back in revenue than what they have invested in as advertising spending. Hence, any product category with ROAS below 1 is not generating sufficient returns to cover the advertising spending costs.

11.3 Exploratory Data Analysis on Business Problem 3

11.3.1 Summary Statistics for Ad Spending

Please see below (Table 5) for the Data Dictionary of Sole Fitness Paid Ads Raw Data.

Variable	Description	Data Type	Sample Values(s)
Month	This field indicates the month of ads spending	Datetime (DD/MM/YYYY)	1/2/2024
Week	This field indicates the week of ads spending	Datetime (DD/MM/YYYY)	5/2/2024
Platform	This field indicates the sales platform which the ads direct viewing traffic to	String	WEBSITE
Medium	This field indicates the platform or source where the advertisement is being displayed	String	Google Ads
Category	This field indicates the type of fitness equipment the ads spending is used for	String	TREADMILL
Campaign	This field indicates the name of the advertising campaign the ads spending is categorized under	String	DSA – Treadmill
Cost	This field indicates the cost of the ad spending	Float	$80.85

Data exploration for Business problem 3 begins with the summary statistics for ad-spent data in Table 5.

Summary Statistics for Ad Cost	
Mean	63.5
Median	29.5
Mode	0
Min	0
Max	571.98
Range	571.98
Variance	8073.57
Standard Deviation	89.9
Coefficient of Variation	1.41
Skewness	2.56
Kurtosis	7.94

Table 33: Summary Statistics for Advertising Cost

The advertising cost data, in the table above, exhibits significant variability and non-normality. The mean cost per campaign is $63.50, while the median is notably lower at $29.51, and the mode is $0.00, indicating a positively skewed distribution driven by a small number of high-cost outliers. This is supported by a skewness of 2.56 and a high kurtosis of 7.94, suggesting a sharp peak and heavy tails. The standard deviation ($89.85) and coefficient of variation (1.41) point to substantial dispersion relative to the mean. The wide cost range (from $0 to $571.98) and presence of zero-cost entries raise questions about campaign execution consistency.

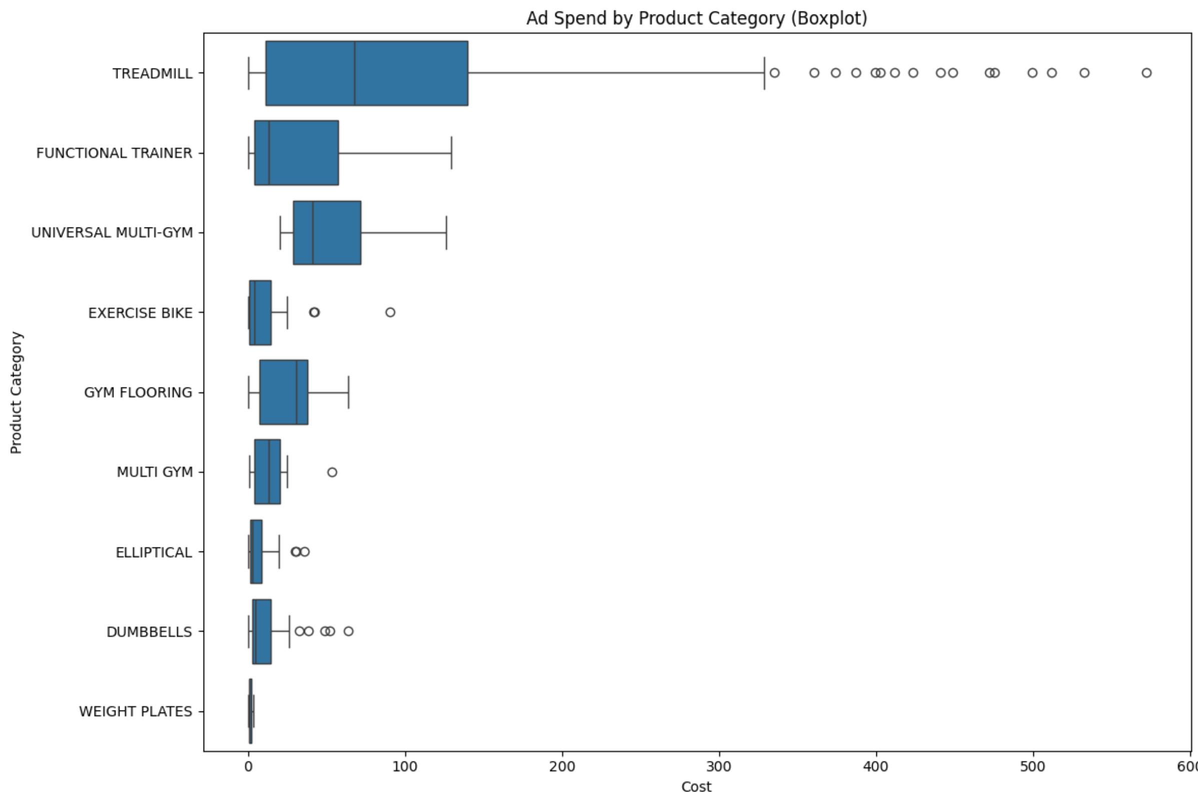


Figure 40: Box Plot of Distribution of Advertising Spend by Product Category

Sole Fitness performs advertisement spend on the entire Gym Flooring category and does not sub-divide the advertisement campaigns into the fitness equipment sub-categories. The category of Gym Flooring comprises several subcategories such as Artificial Turf, Eva Foam Mats, Gymnastic Mats, Rubber Gym Flooring. For advertising spend analysis purposes, these 4 sub-categories are lumped together as the main Gym Flooring category for insights and recommendations.

The boxplot in Figure 40 above, illustrates the distribution of advertising expenditure across different product categories.

Notably, Treadmills exhibit the widest interquartile range (IQR), indicating high variability in spending. As more campaigns were executed by sole fitness for treadmills, the presence of several outliers is not surprising.

In contrast, categories such as Functional Trainers, Exercise Bikes, and Gym Flooring demonstrate more compressed distributions, implying more consistent and lower levels of ad spending. The median advertising cost for these categories is also significantly lower compared to Treadmills, highlighting the imbalance in ad budget allocation.

Overall, the boxplot in Figure 40 suggests a concentrated advertising strategy focused heavily on Treadmills, with relatively limited investment in promoting other products in different categories. As per initial business studies and, this strategy decreases the ROAS and limits opportunities for revenue diversification.

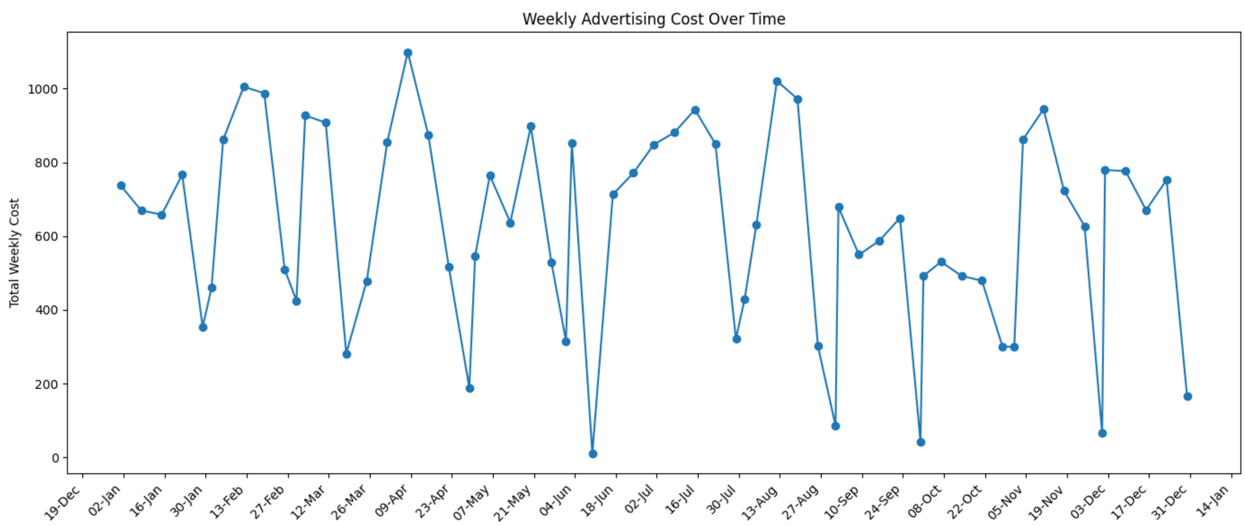


Figure 41: Weekly Advertising Cost Over Time

Figure 41 above represents a line chart depicting the total advertising expenditure incurred on a weekly basis throughout the year 2024. Each data point represents the aggregate cost of all advertising campaigns launched in that week across advertising mediums (Google Ads, Facebook Ads, and Bing Ads) and product categories.

The plot reveals considerable fluctuations in weekly spending. Several weeks recorded markedly lower spending levels, suggesting underutilization of the allocated weekly budget. For instance, spending dips are observed intermittently throughout the year, such as during the latter half of March and early October. Conversely, spikes in expenditure appear in early February, mid-July, and early December, indicating concentrated bursts of ad activity. In the coming sections these fluctuations will be compared to the revenue.

This inconsistent distribution of ad spends backed with knowledge of Sole Fitness static budget allocation of 3600$ every month is indicative that a data-informed approach is lacking.

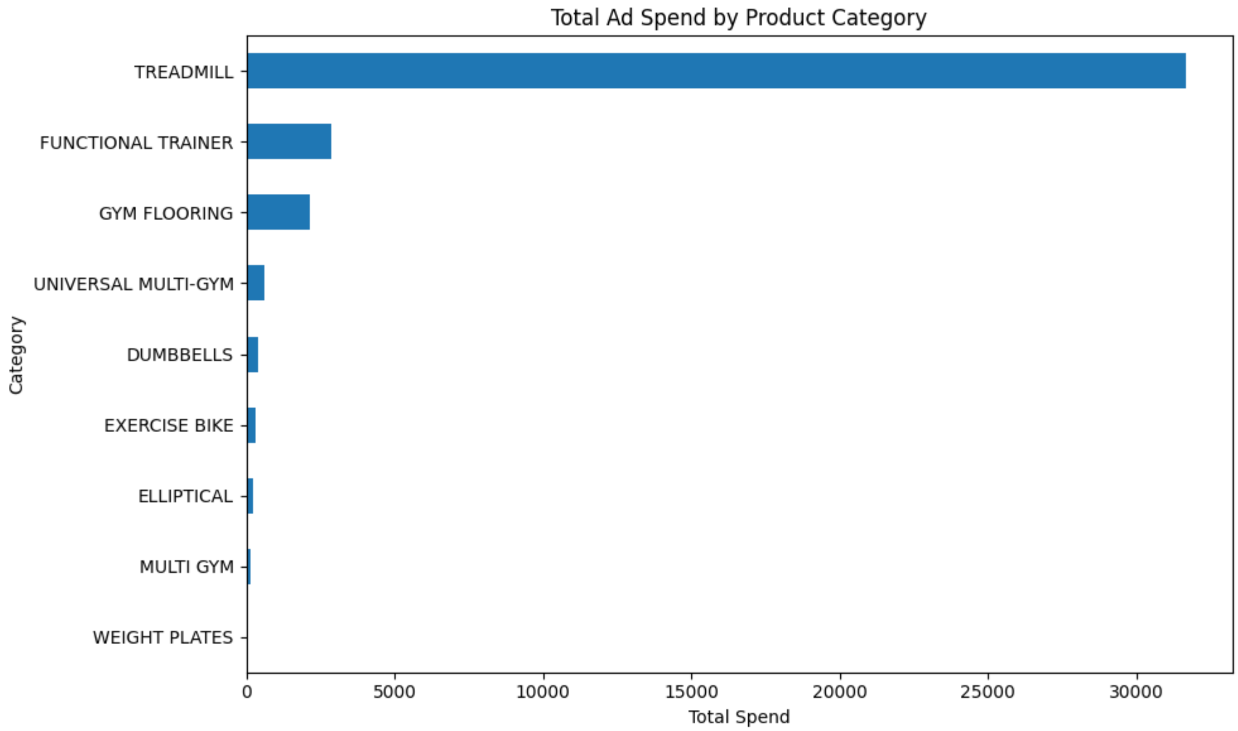


Figure 42: Total Advertising Spend by Product Category

Figure 42 above is a horizontal bar chart displaying the cumulative advertising expenditure allocated to each product category across all campaigns in 2024. As expected, the visualization reveals a pronounced imbalance in budget distribution, with the Treadmill category receiving a disproportionately high level of investment compared to all other product lines.

The next highest expenditure was for Functional Trainers and Gym Flooring, yet both were significantly lower, falling below $3,000 each. Categories such as Universal Multi-Gyms, Dumbbells, and Exercise Bikes saw even lower investments, while Ellipticals, Multi Gyms, and Weight Plates received minimal to negligible funding.

This solidifies the need for Sole Fitness to leverage data analytics to identify other high-potential categories that are currently underfunded, thereby improving both the efficiency of advertising investments and the robustness of the overall revenue generation.

11.3.2 Proportion of Ad Spending Over Revenue

Before analyzing the data based on the proposed technical objectives, we must delve deeper into the dataset. The combo charts below, provides a clearer view of the proportion of advertising spend to item revenue for each of the fitness equipment categories. The combo chart for weight plates and universal multi was omitted due to the absence of revenue for the particular category on the weeks with advertising spend and insignificant percentage of ad spend over annual revenue for each of the fitness categories (<1%).

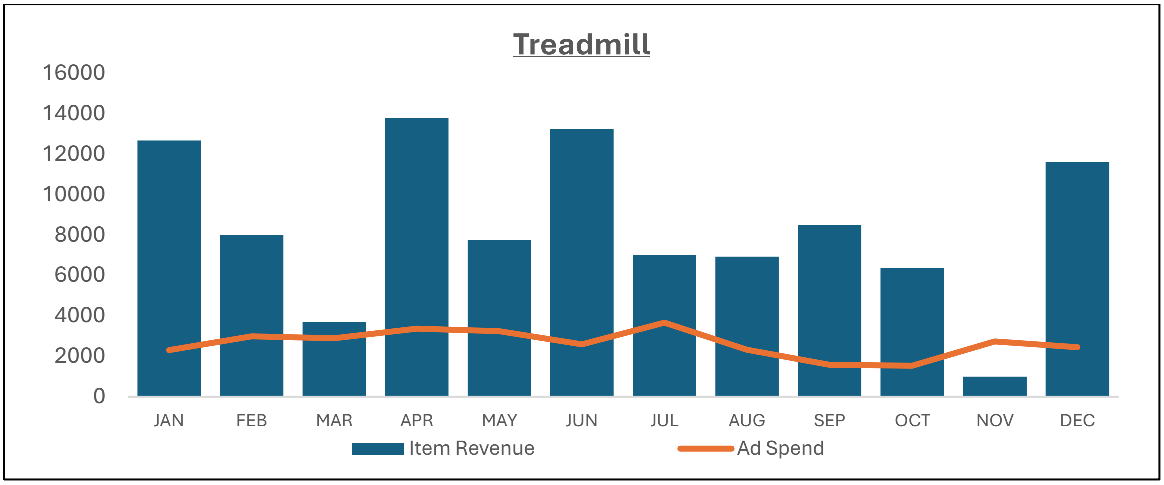


Figure 43: Combo Chart Showing Proportion of Ad Spend Over Revenue for Treadmill

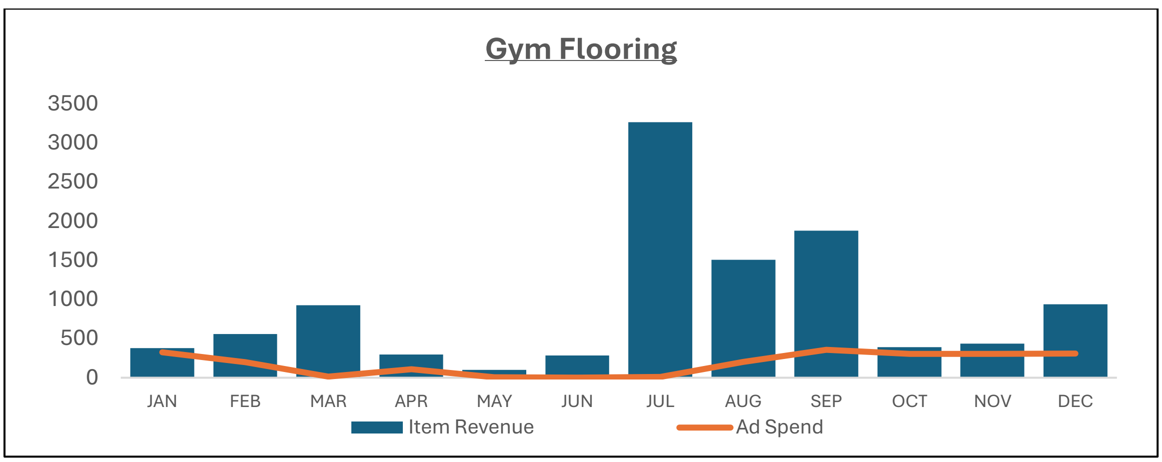


Figure 44: Combo Chart Showing Proportion of Ad Spend Over Revenue for Gym Flooring

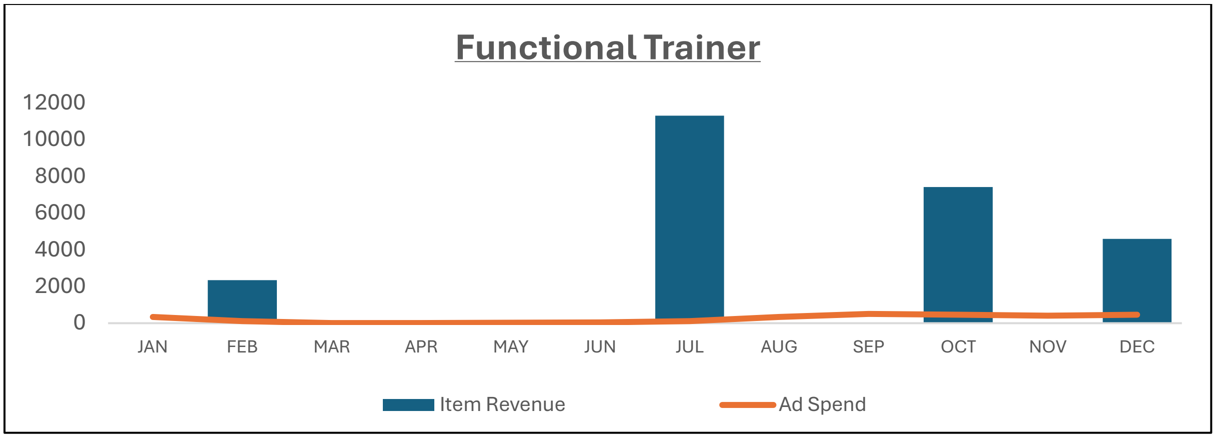


Figure 45: Combo Chart Showing Proportion of Ad Spend Over Revenue for Functional Trainer

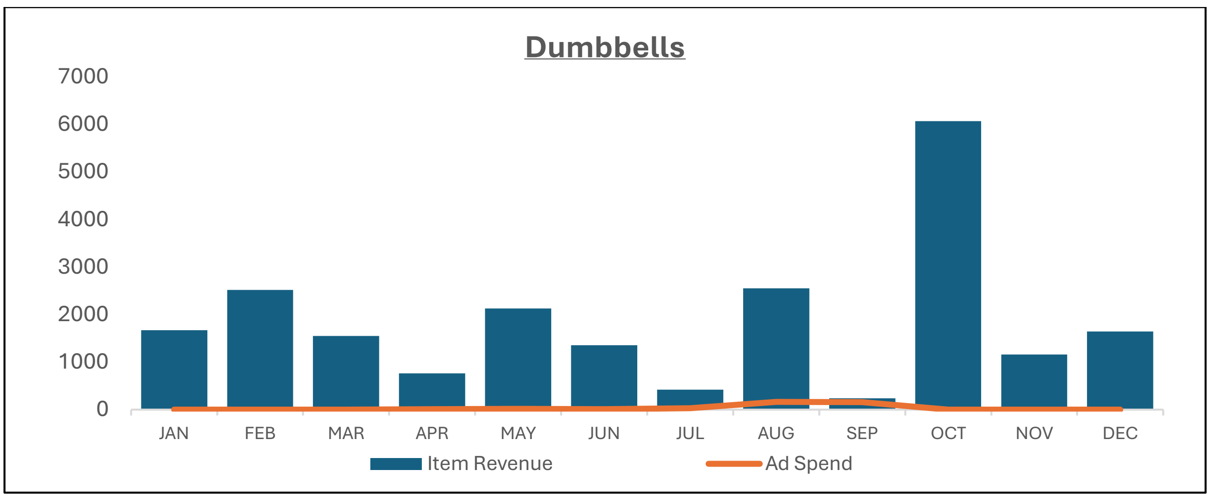


Figure 46: Combo Chart Showing Proportion of Ad Spend Over Revenue for Dumbbells

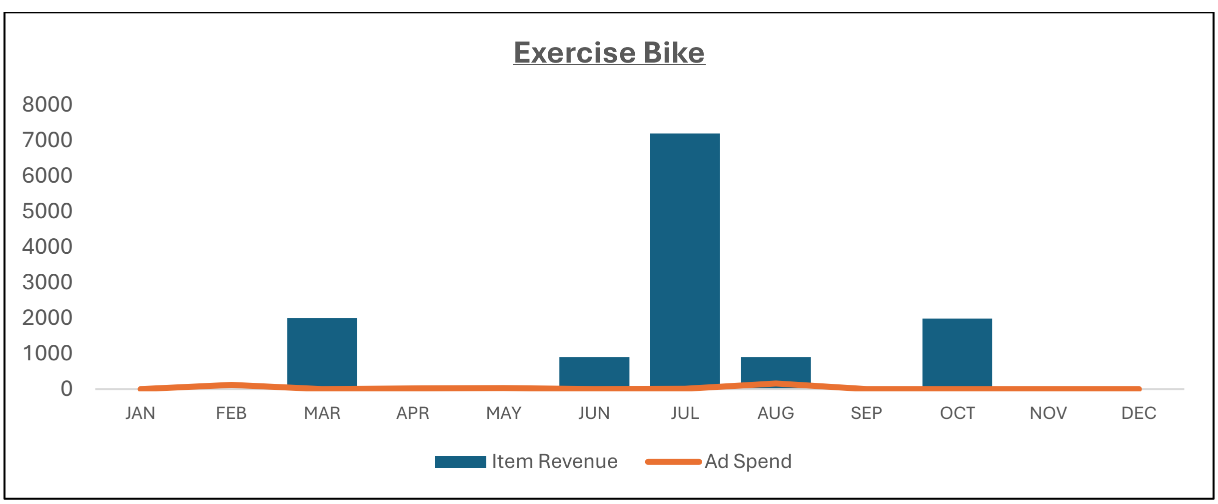


Figure 47: Combo Chart Showing Proportion of Ad Spend Over Revenue for Exercise Bike

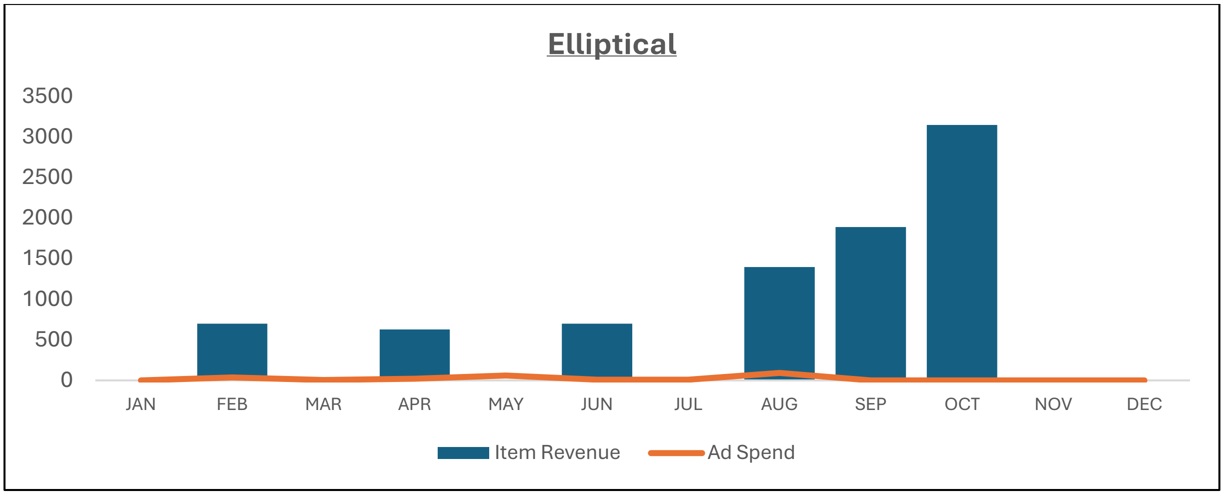


Figure 48: Combo Chart Showing Proportion of Ad Spend Over Revenue for Treadmill

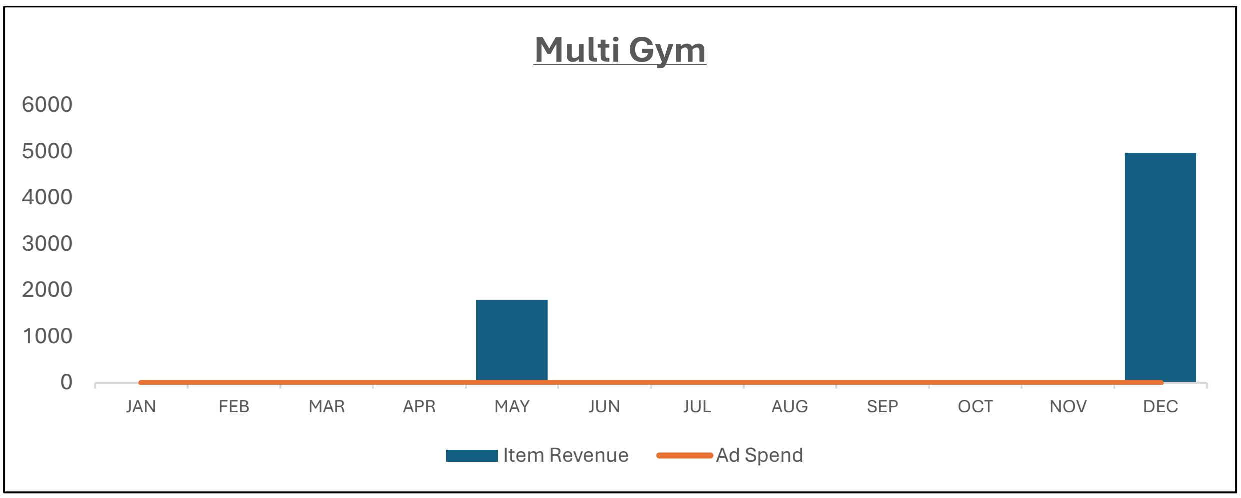


Figure 49: Combo Chart Showing Proportion of Ad Spend Over Revenue for Multi Gym

The table summary of the proportion of ad spend for each fitness equipment category is also included below in Table 34.

Fitness Equipment Category	**Percentage of Ad Spend Over Revenue in 2024**
TREADMILL	31.48%
GYM FLOORING	19.49%
FUNCTIONAL TRAINER	11.11%
ELLIPTICAL	2.80%
EXERCISE BIKE	2.57%
MULTI GYM	2.17%
DUMBBELLS	1.77%
UNIVERSAL MULTI-GYM	*No Revenue*

Table 34: Summary of Proportion of Advertising Cost over Revenue

As seen from Figure 43 and Table 34, the treadmill category has the highest proportion of advertisement spending when compared to the monthly item revenue. This is expected given that treadmills are the anchor product for Sole Fitness, having grossed over 40% of the annual website revenue. The proportion of advertisement spending over revenue for treadmill is about 31% and is among the highest for the product categories with ad spend in Sole Fitness. This is followed by gym flooring and functional trainer with a proportion of 19% and 11% percentage respectively.

11.4 Ad Spend Analysis

11.4.1 Ad Spend Analysis Considerations

Next, we binned the weekly advertisement spending data points for each of the fitness equipment categories into the bins as below and found that a significant number of the ad spend is below $10. The <$10 data points take up about 31% of the dataset.

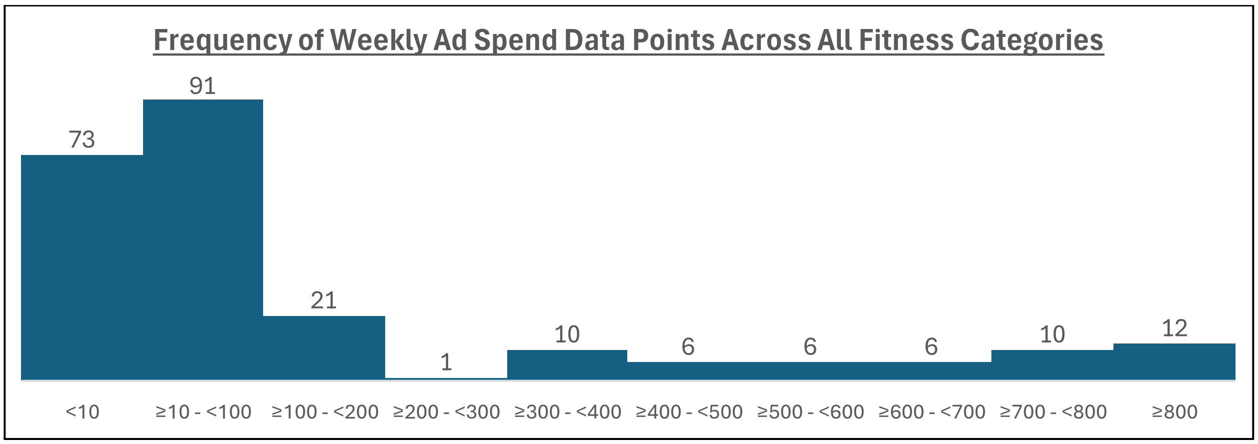


Figure 50: Bar Chart of Weekly Advertisement Spend For Respective Fitness Categories

To ensure a more informed analysis, we will be proceeding to remove the data points of weekly advertising spend with less than $10. There are two major reasons why such an exclusion is done.

From a domain perspective, <$10 ad spend is rather insignificant in the span of a week. When the advertising spending is below a certain threshold, the ad placement may not be at an ideal place due to the need for competitive bids for high visibility areas (Kovalenko, 2018). With insufficient spending, Sole Fitness will not be able to secure optimal placements for their ads. The small amount of spending also meant that there are issues with the ad targeting, and the audience may not fully understand what the ads are about. Therefore, the exclusion of <$10 weekly ad spend helps Sole Fitness to focus on campaigns with sufficient budget to test and refine key variables, thereby enhancing overall ad performance and return on advertising spend.

From a technical standpoint, excluding ad spending less than $10 is essential to ensure the accuracy and reliability of the analysis. Low-budget campaigns often lack sufficient data for meaningful insights due to limited reach and engagement. Moreover, there are other factors such as organic conversions and organic product views that would affect the relevant metrics more at this level of ad spending. With 31% of the data points being < $10, it will significantly affect the analysis result and recommendations. Therefore, ad spending less than $10 should be excluded from the analysis to prevent skewed results and ensure the robustness of the data. This threshold is set to eliminate campaigns that do not generate enough data points with sufficient range for reliable statistical analysis and optimization, which will improve the validity of the correlation study between views and ad spending.

11.4.2 Correlation Analysis & Linear Regression Between Ad Spend & Item Views

For correlation analysis between item views and ad spending, we will be analyzing the data on a weekly basis. After removing the data points with <$10 weekly ad spend, we proceeded with the plot out the scatter plot of weekly item views against weekly ad spend for fitness categories with ad spending.

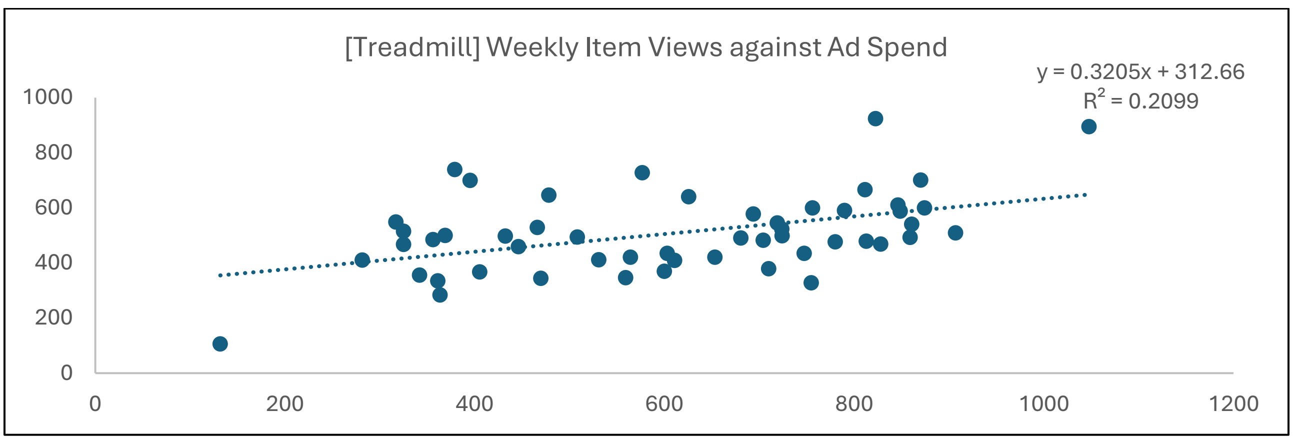


Figure 51: Scatterplot of Weekly Item Views Against Weekly Advertisement Spend For Treadmill

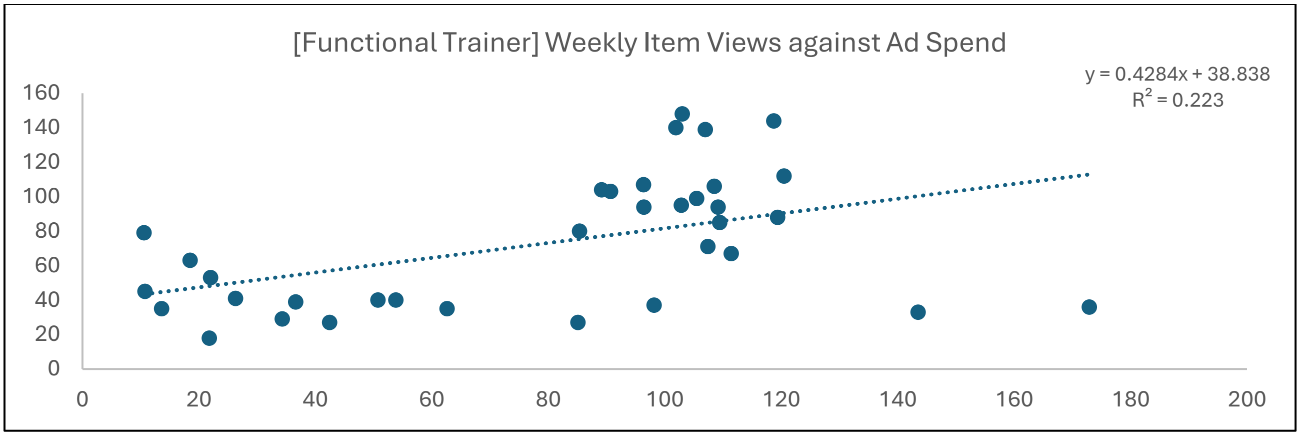


Figure 52: Scatterplot of Weekly Item Views Against Weekly Advertisement Spend For Functional Trainer

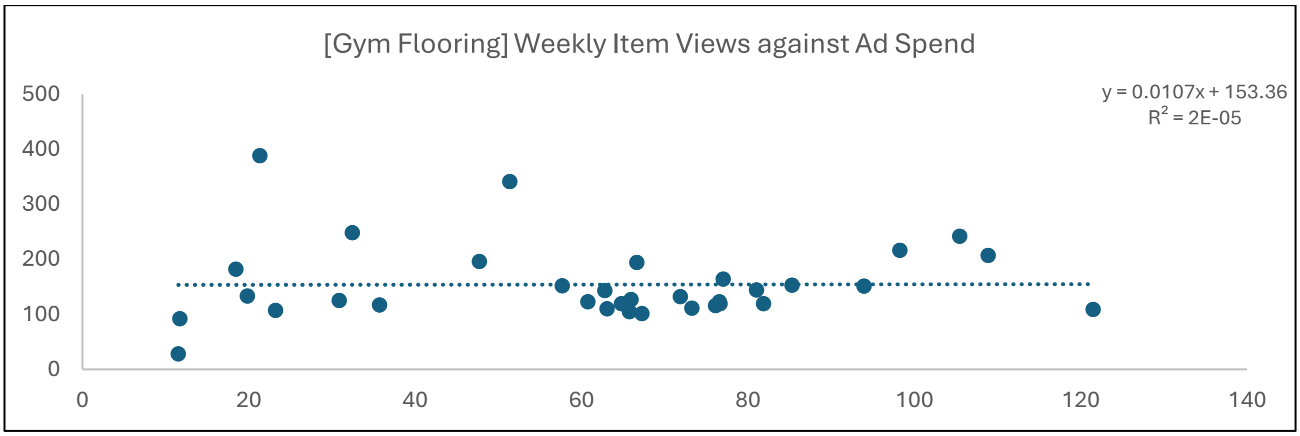


Figure 53: Scatterplot of Weekly Item Views Against Weekly Advertisement Spend For Gym Flooring

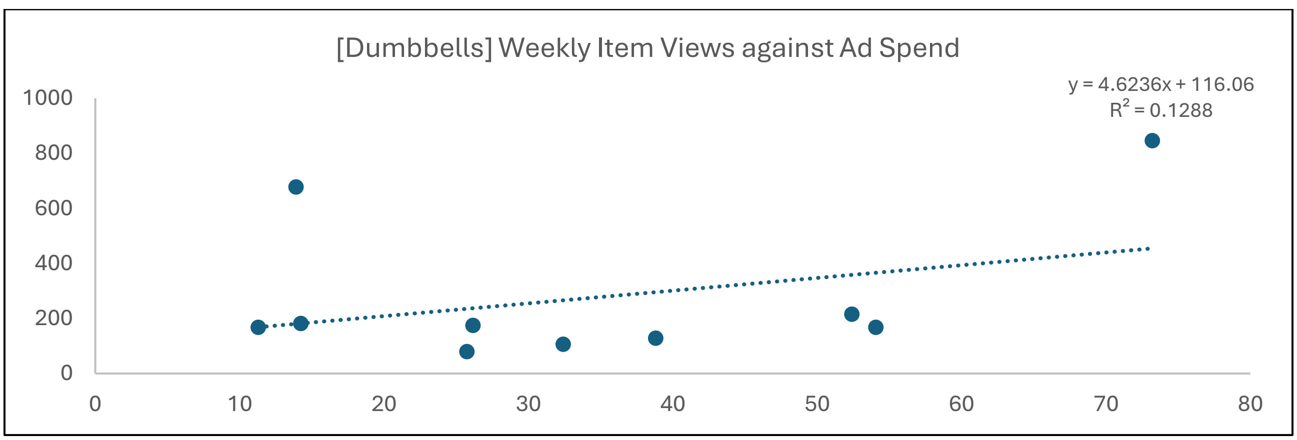


Figure 54: Scatterplot of Weekly Item Views Against Weekly Advertisement Spend For Dumbbells

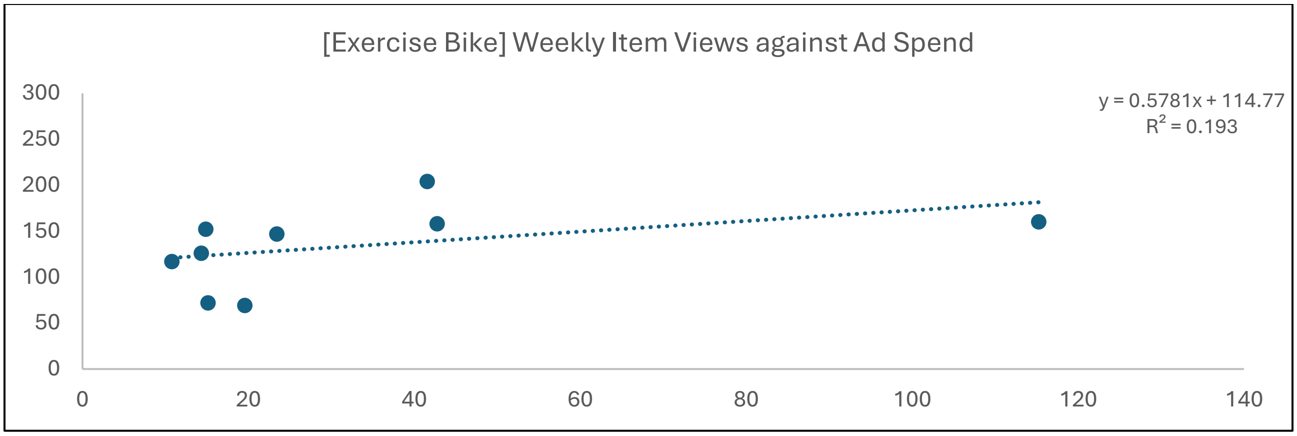


Figure 55: Scatterplot of Weekly Item Views Against Weekly Advertisement Spend For Exercise Bike

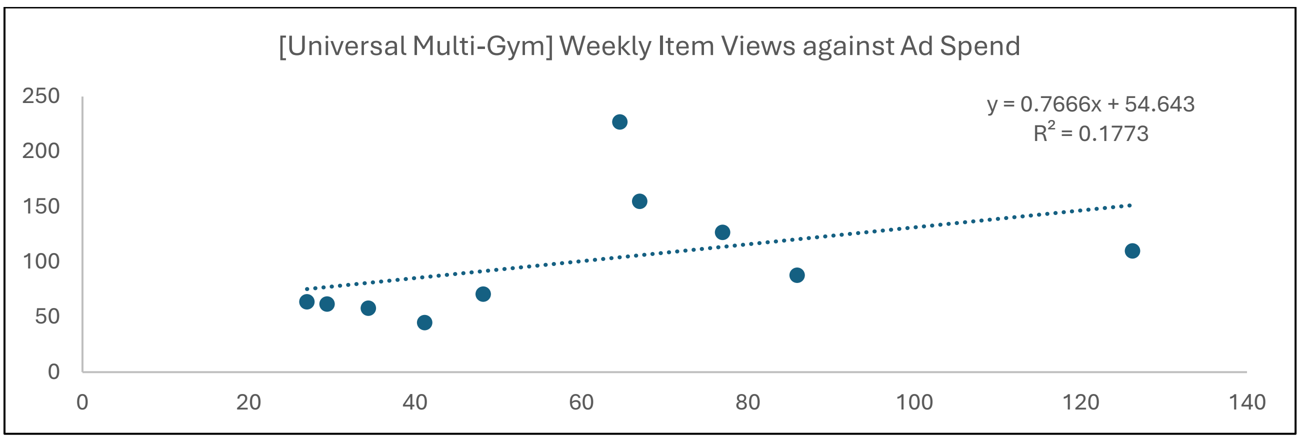


Figure 56: Scatterplot of Weekly Item Views Against Weekly Advertisement Spend For Universal Multi-Gym

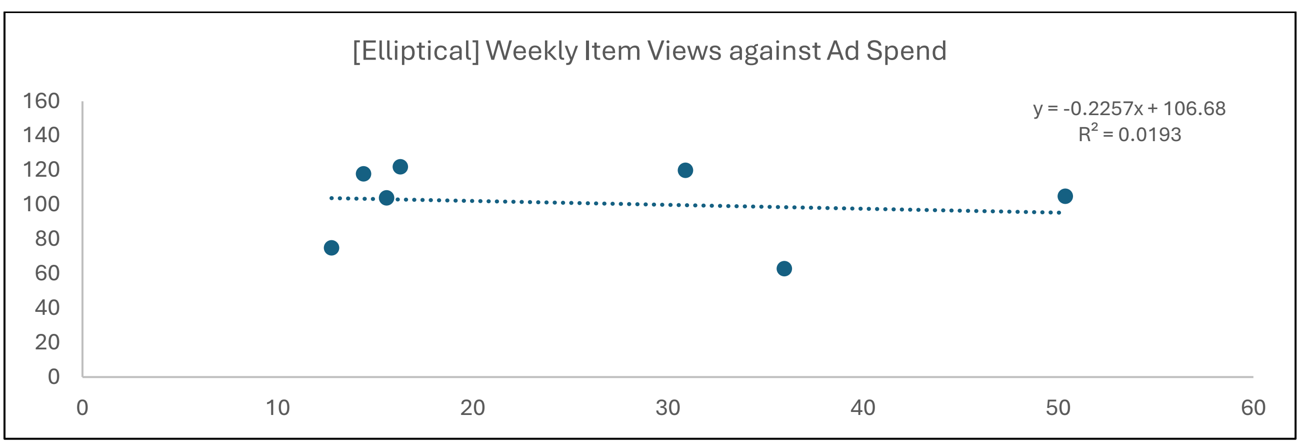


Figure 57: Scatterplot of Weekly Item Views Against Weekly Advertisement Spend For Elliptical

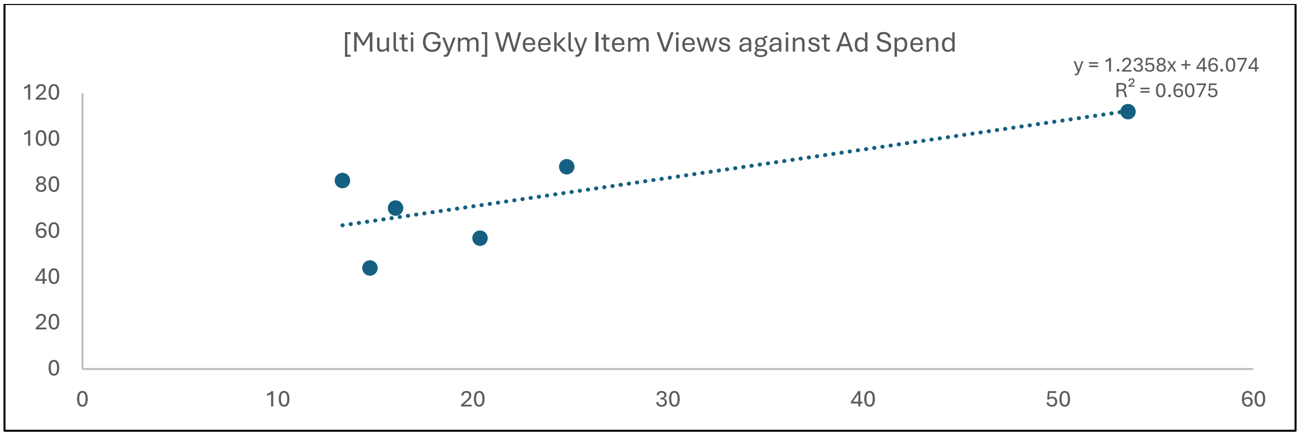


Figure 58: Scatterplot of Weekly Item Views Against Weekly Advertisement Spend For Multi Gym

In the individual scatterplot for the various fitness equipment categories, the simple linear regression line is also fitted onto the charts. Treadmill had the broadest range of weekly ad spending data points, ranging from $132 to $1,047. Remaining fitness equipment categories have weekly ad spending ranging from $10 to approximately $200.

Multi-gyms only have 6 data points of ad spending, which is not a sufficient number for proper linear regression and correlation analysis. Statistically, we will not be able to glean accurate data patterns from such a small number of weekly data points. Hence, the analysis results should not consider Multi-gym fitness equipment category.

Moving on, the goodness of fits measure of all the linear regression lines are not that high, except for multi gym (as seen in Table 35). The values of the fitness equipment category are all below 0.30, while the usual $R^2 > 0.7$ indicates that the model is a good fit. This would indicate that the simple linear regression lines are not capable of predicting the appropriate number of views with the amount of ad spending. As such, we will not be recommending Sole Fitness to use the linear regression models for their predictions.

Fitness Equipment Category	**R^2 Measure**
MULTI GYM	0.61
FUNCTIONAL TRAINER	0.22
TREADMILL	0.21
EXERCISE BIKE	0.19
UNIVERSAL MULTI-GYM	0.18
DUMBBELLS	0.13
ELLIPTICAL	0.02
GYM FLOORING	0.00

Table 35: Goodness of Fit Measure Between Item Views & Ad Spend For All Fitness Equipment

The bar chart below shows the summary of correlation between weekly item views and ad spend for all of the fitness equipment categories, excluding Multi-Gym which only had a small number of 6 data points.

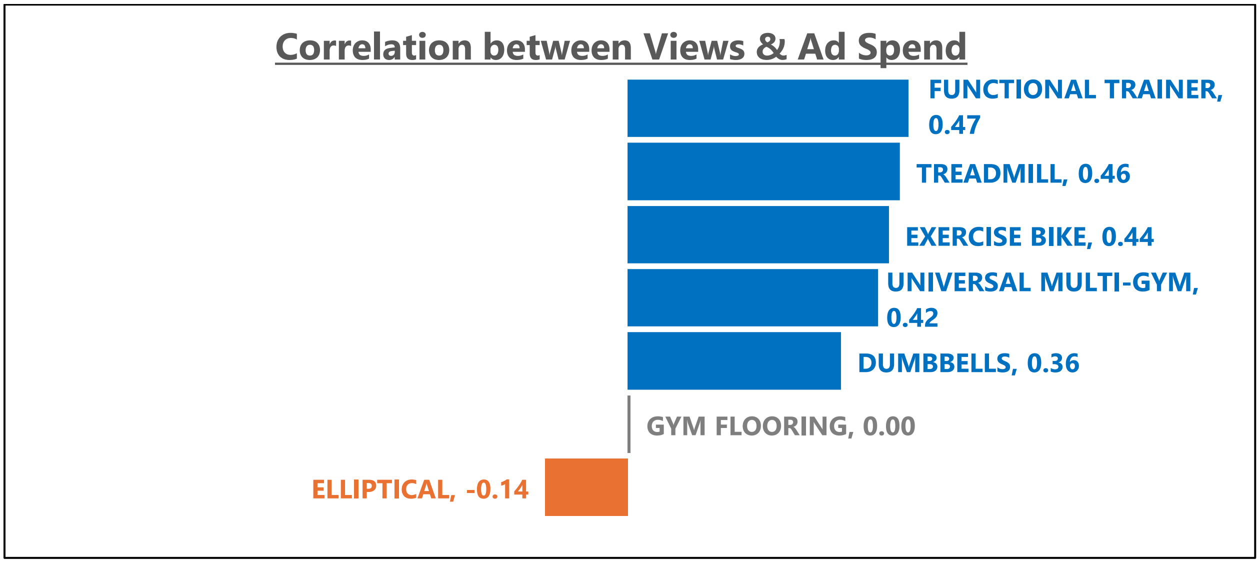


Figure 59: Summary Bar Chart of Correlation Between Weekly Item Views And Weekly Advertisement Spend

During our correlation analysis for views and ad spend, we found that the top 3 fitness equipment categories are functional trainer, treadmill and exercise bike. Majority of the fitness equipment categories have a weak positive correlation between item views and ad spend. Functional trainers have the best correlation, which means that Sole Fitness has done very well for the ad targeting accuracy, reaching the correct audience and with ads placement in the high visibility areas.

For elliptical, we noticed that it has a negative correlation. This is an outlier, and more data points are needed to make a definite conclusion as elliptical has only 7 data points. For Gym Flooring, the correlation is non-existent between views and ad spend. This could be due to many different factors such as high baseline organic views for this category, and further observation is needed for this category as well.

Our recommendation for the categories with positive correlation is for Sole Fitness to have incremental increases in small interval ($100) to ad spending and monitor for ad spend saturation over time using the correlation score. Ad spend saturation occurs when over-targeting of the same audience with the same ads happens repeatedly. Sole Fitness should be cautious of such situation where increases in ad spending led to a continuous decrease in the correlation score over a period of time, as it meant that ad spend saturation has occurred and Sole Fitness should relook at the ad targeting parameters to avoid re-targeting the same audience repeatedly.

11.4.3 ROAS Analysis

Moving on to the Return on Advertisement Spending (ROAS) analysis, we will be analyzing the ROAS for the respective fitness equipment categories using the formula that we have specified earlier.

From the bar chart below (Figure 60), functional trainer has the highest ROAS, while exercise bike has the most negative ROAS. Recalling from the earlier correlation analysis between views and ad spending, our top 3 categories are actually Functional Trainer, Treadmill and Exercise Bike, which coincidentally, is the top 2 and bottom 1 performer in terms of ROAS.

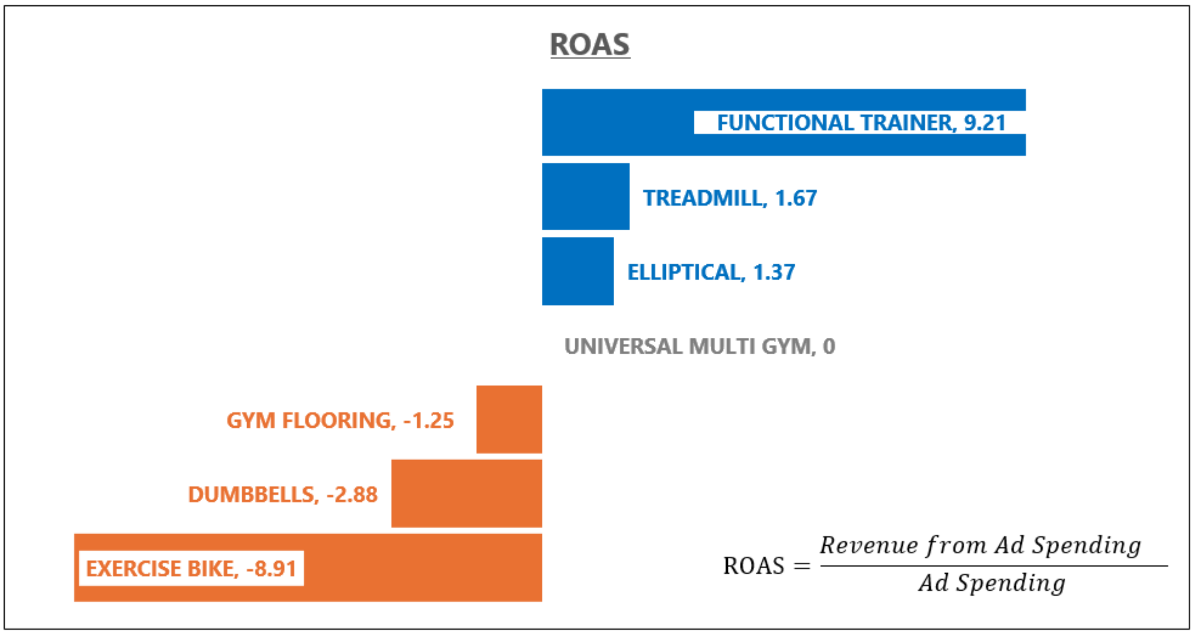


Figure 60: Bar Chart of ROAS for Various Fitness Equipment Categories

Our recommendation for Sole Fitness is to continue increase ad spending for Functional Trainer to get better returns and higher revenue, exploiting the high ROAS (9.21) of the fitness equipment category. The correlation score and ROAS over time should be monitored while the ad spending is increased, such that the ad saturation point is not reached.

As for treadmill, the current level of ad spending is sufficient, and we do not recommend further increases in ad spending due to the mature ROAS of 1.67. Simply increasing ad spending may not lead to higher revenue and may lead to a decrease in the ROAS metric for treadmill instead.

For exercise bike, which is the lowest performer of ROAS (-8.91), we would recommend Sole Fitness to stop the ad spending for this particular fitness equipment category. In addition, Sole Fitness should relook their advertisement campaigns due to the absence of returns, which meant that Sole Fitness is wasting precious resources on this non-performing category.

For gym flooring, the correlation analysis earlier indicated that there is no correlation between item views and ad spend, while the ROAS analysis result showed that gym flooring has a negative ROAS of -1.25. From the combo chart earlier (Figure 44), it appeared that gym flooring has seasonality in item revenue, and advertisement spending might be a smaller factor in influencing the product views and purchases count.

11.5 Conclusion for Business Problem 3

To enable optimized ad spending for the different fitness equipment categories, both Correlation Score between Item Views and Ad Spend and ROAS for the categories should be considered.

For fitness equipment categories without existing ad spend, if Sole Fitness is interested in starting advertisement spending for the categories, Sole Fitness should start with weekly advertisement spending of >$10 with incremental increases across the weeks while accumulating more data points if possible. At the same time, Sole Fitness should monitor the correlation score between ad spending and item views as well as the ROAS across time for those particular categories.

Based on the analysis of fitness equipment categories with existing and sufficient advertisement spending data points, Sole Fitness should focus advertisement spending on Functional Trainer category, which possess the best correlation score and ROAS metric, while maintaining the level of advertisement spend on treadmills, which has mature ROAS and is their anchor product. In addition, Sole Fitness should relook at advertisement spending and campaigns for fitness equipment categories with negative correlation and ROAS <1 (such as Exercise Bikes and Dumbbells), as the negative metrics may be due to insufficient data points or actual advertising underperformance of the fitness equipment category.

For all of the fitness equipment categories requiring further ad spending, Sole Fitness should continue to monitor the ROAS per campaign basis and perform additional AB testing, as ROAS is a constantly evolving metric and will vary across time, together with the correlation score of views and ad spend.

12.0 Sole Fitness Dashboards Overview

12.1 Sales Trend Dashboard

A Sales Trend dashboard was developed to provide a high-level overview of revenue and purchasing patterns over time. This enables the business to make swift decisions through the identification of top-performing products and revenue fluctuations.

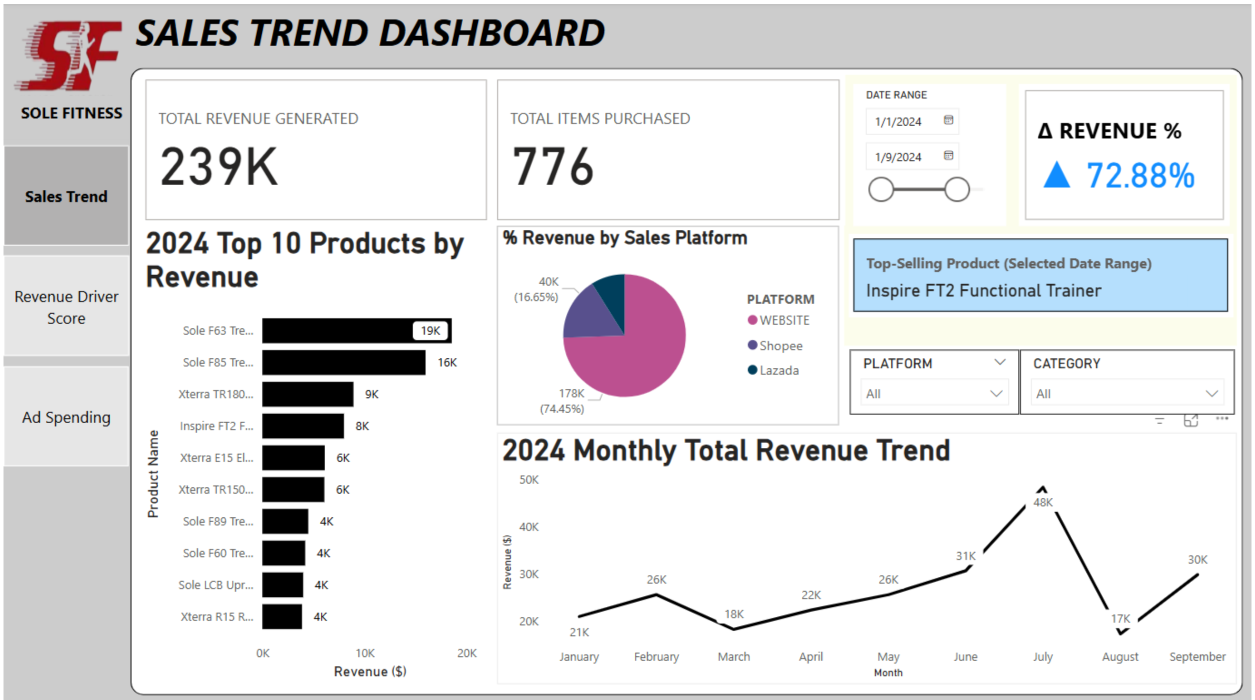


Figure 61: Sole Fitness Sales Trend dashboard

The dashboard layout prioritizes key performance indicators (KPIs) at the top, mainly the total revenue attained by the business and the number of items purchased. To provide more contextual KPIs, we wanted to further provide a means of comparison i.e. MoM change in revenue. To provide timeframe control, we included a date range slicer to allow users to filter by month or even a custom date range. This allows users to explore percentage revenue change across a custom timeframe. Additionally, a card element highlights the top-selling product within the selected time range, reinforcing actionable insights.

The bar chart titled *"2024 Top 10 Products by Revenue"* ranks individual products by total revenue contribution, offering immediate visibility into sales drivers. To provide cross-platform performance analysis, a pie chart showing platform-level revenue distribution (Website, Shopee, Lazada) was included as well, revealing the effectiveness of each channel. The line graph titled *"2024 Monthly Total Revenue Trend"* captures seasonality and revenue trajectory throughout the year.

The layout of the Sales Trend dashboard was intentionally designed to align with the *Gutenberg Diagram*, which models how users typically scan visual information from the top-left to the bottom-right in a diagonal flow.

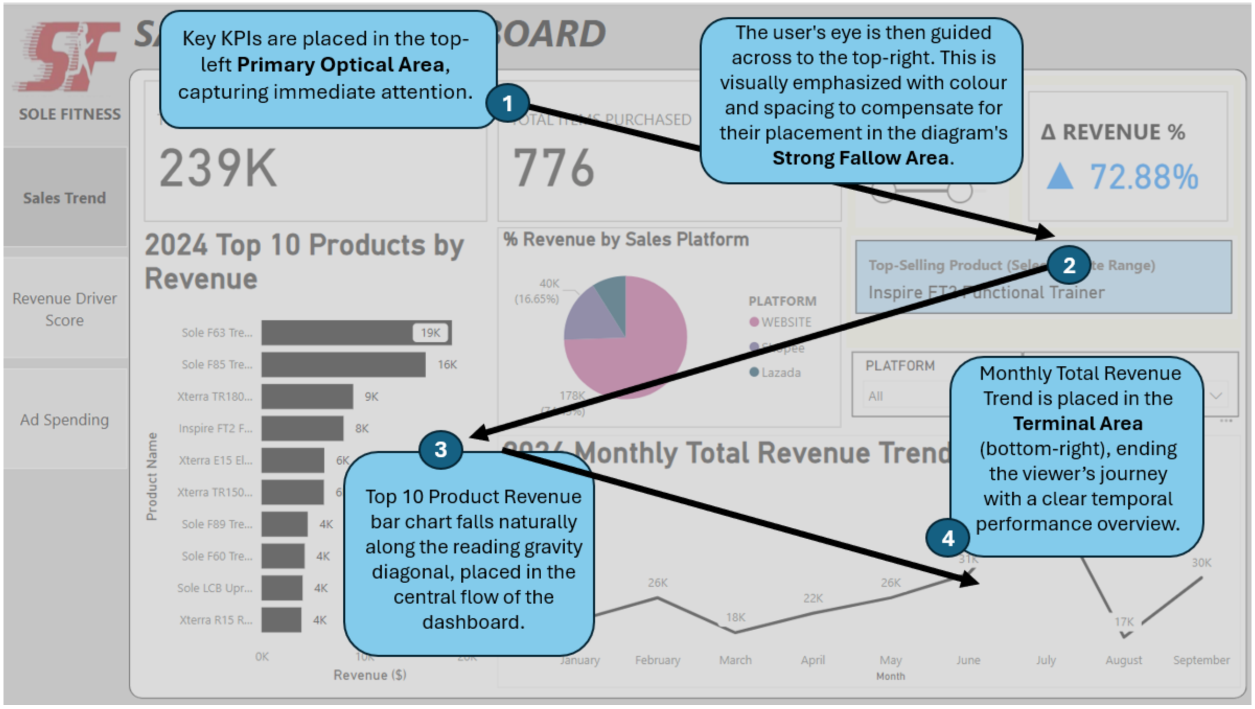


Figure 62: Representation of the applied principles from the Gutenberg Diagram on Sole Fitness Sales Trend dashboard

Key performance metrics are placed at the starting point to immediately anchor the viewer's attention. Important interactive elements i.e. adjusting the date range slicer to utilize the revenue change indicator, are positioned strategically to maintain engagement across less naturally emphasized areas. The central region and terminal area feature bar and line charts that reinforce the dashboard's main narrative: showcasing top products and overall revenue trends over time. This structured layout ensures that users absorb key insights efficiently.

The Gestalt design principles are also applied to enhance clarity and user navigation. The principle of proximity is observed in how related metrics (e.g., Total Revenue, Total Items Purchased, and Revenue % Change) are grouped in the top row, allowing users to immediately compare overall performance indicators. Similarity is applied through consistent visual formatting (e.g., font styles and card visuals), which helps users distinguish between metric types at a glance. The principle of common region is demonstrated through the use of boxed sections, such as the Revenue % panel and the top-selling product card, which visually bind related content together.

Lastly, continuity and figure-ground are reflected in the layout's clean grid and background contrast, which guides the viewer's eye from high-level KPIs at the top toward supporting charts below, reinforcing a logical flow of insight discovery.

12.2 Category Performance Dashboard

Given that the business manages over 50 product categories, a Category Performance dashboard was developed to provide a consolidated view of how each category is performing. This enables the business to identify well-performing categories, uncover untapped growth potential, and flag underperforming segments – supporting informed decision-making on future investment, promotion, and optimization strategies.

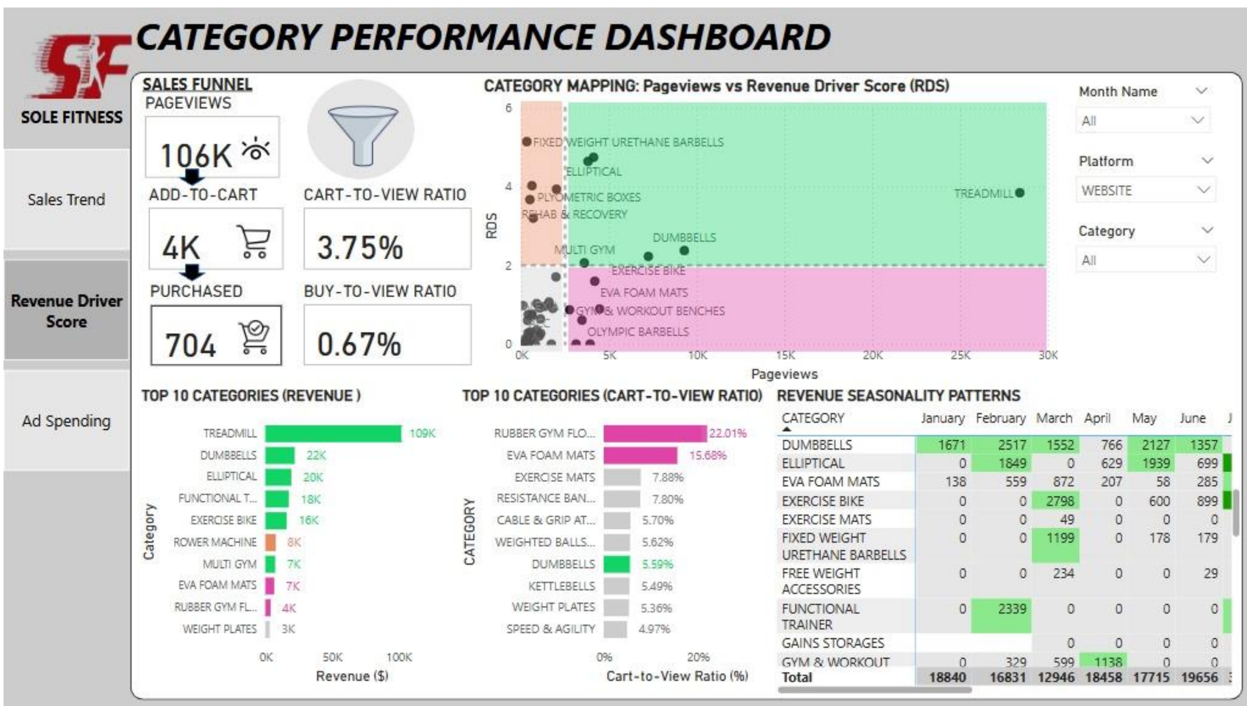


Figure 63: Sole Fitness Category Performance dashboard

The dashboard opens with an impactful sales funnel to reveal customer conversion rates across different stages – from pageviews to add-to-cart to purchases – pinpointing where major drop-offs occur. This funnel is supported by two key performance ratios: cart-to-view, which measures the percentage of product views that result in cart additions, and buy-to-view, which tracks the final conversion rate from views to purchases. These metrics provide clarity on where user interest is generated and where potential revenue may be slipping through.

To complement the 4-quadrant framework discussed in Business Problem 2, a centrally positioned scatter plot visualizes the segmentation of product categories based on pageviews and revenue contribution. This enables clear classification into our 4 segments: Hero Products, Hidden Gems, Underperformers, and Low Signal. This reinforces our suggested strategic approach in equipping the business with actionable insights for prioritization.

Following the diagonal reading gravity of the Gutenberg Diagram, the viewer's eye naturally continues to the bottom-left, where a bar chart of top revenue-generating categories is placed. Adjacent to this, a cart-to-view ratio bar chart highlights categories with strong purchase intent that may benefit from promotional support. Both bar charts implement bold bars and colors, tight spacing, and intuitive ranking, give them strong visual weight, effectively drawing user attention. Finally, a revenue seasonality matrix in the bottom-right offers a granular, month-by-month view of category volatility.

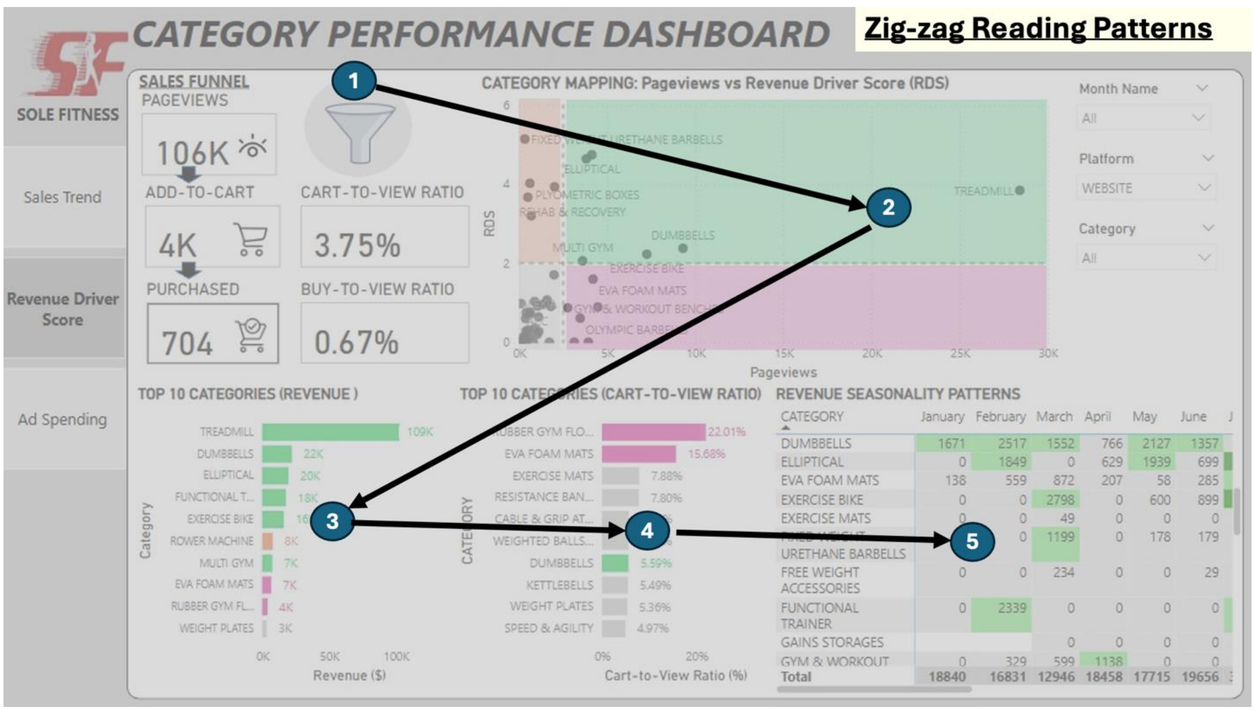


Figure 64: Zig-Zag Reading Flow applied to the Sole Fitness Category Performance Dashboard

Once again, the layout applies Gestalt principles such as proximity (grouping related KPIs), similarity (uniform chart styling), and continuity (logical visual flow), all of which contribute to a seamless and intuitive navigation experience.

Together, these visual elements support structured exploratory analysis, starting from "what sold," moving through "how it converted," and concluding with "when to act"—to guide strategic product and marketing decisions.

12.3 Ad Spend Dashboard

To assist Sole Fitness in further monitoring the ad spend key metrics for their ecommerce business, we have come up with a tactical dashboard Business Problem #3: Ad Spend Effectiveness). The intended audience for the dashboard would be the Data Owner, Marketing Team and Operation/Sales Team. Please see Figure 65 for the Ad Spending portion of the Dashboard.

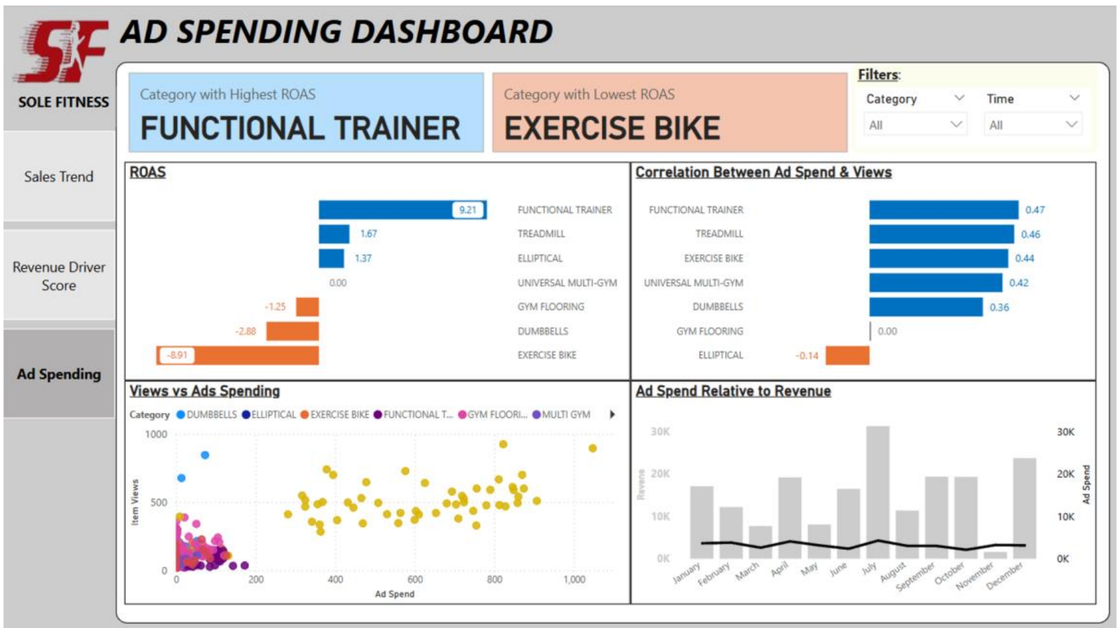


Figure 65: Sole Fitness Ad Spending Dashboard

For the Ad Spending Dashboard, key metrics that have been analyzed earlier, such as ROAS and Correlation Score between views and ad spend together with the Category with the Highest and Lowest ROAS, will be called out. The metrics are both automatically calculated with Power BI Measures, making those metrics accessible to Sole Fitness users as long as the data is supplied to the pipeline.

In this report, we will be showing the formula that the measures are being calculated with, with certain information (such as table names and column names) redacted and replaced to ensure the data security of the Sole Fitness dashboard.

The ROAS Measure is calculated as below:

```
ROAS =
VAR RevenueWithAdSpend = SUMX(
   VALUES('Data'[Category]),
   CALCULATE(AVERAGE('Data'[ItemRevenue]), 'Data'[AdSpend] >= 10))
VAR RevenueWithoutAdSpend = SUMX(
   VALUES('Data'[Category]),
   CALCULATE(AVERAGE('Data'[ItemRevenue]), 'Data'[AdSpend] < 10))
VAR AdSpend = SUMX(
   VALUES('Data'[Category]),
   AVERAGE('Data'[AdSpend]))
RETURN
DIVIDE(RevenueWithAdSpend - RevenueWithoutAdSpend, AdSpend)
```

Additionally, the Correlation Score Measure is calculated as below:

```
CorrelationBetweenViewsAndAdsSpend =
VAR MeanX = AVERAGE('Data'[AdSpend])
VAR MeanY = AVERAGE('Data'[ItemsViewed])
VAR SumProduct = SUMX('Data', ('Data'[AdSpend] - MeanX) * ('Data'[ItemsViewed] -
MeanY))
VAR SumSquareX = SUMX('Data', ('Data'[AdSpend] - MeanX) ^ 2)
VAR SumSquareY = SUMX('Data', ('Data'[ItemsViewed] - MeanY) ^ 2)
RETURN
DIVIDE(SumProduct, SQRT(SumSquareX * SumSquareY))
```

The highest ROAS category is also determined through the following Measure:

```
HighestROASCategory =
VAR Categories = VALUES('Data'[Category])
VAR ROASTable = ADDCOLUMNS(Categories, "ROAS",
   VAR RevenueWithAdSpend = CALCULATE(AVERAGE('Data'[ItemRevenue]),
'Data'[Category] = EARLIER('Data'[Category]), 'Data'[AdSpend] >= 10)
   VAR RevenueWithoutAdSpend = CALCULATE(AVERAGE('Data'[ItemRevenue]),
'Data'[Category] = EARLIER('Data'[Category]), 'Data'[AdSpend] < 10)
   VAR AdSpend = CALCULATE(AVERAGE('Data'[AdSpend]), 'Data'[Category] =
EARLIER('Data'[Category]), 'Data'[AdSpend] >= 10)
   RETURN DIVIDE(RevenueWithAdSpend - RevenueWithoutAdSpend, AdSpend)
)
VAR TopCategoryTable = TOPN(1, ROASTable, [ROAS], DESC)
RETURN
MAXX(TopCategoryTable, [Category])
```

While the lowest ROAS Category is determined through the following measure:

```
LowestROASCategory =
VAR Categories = VALUES('Data'[Category])
VAR ROASTable = ADDCOLUMNS(Categories, "ROAS",
    VAR RevenueWithAdSpend = CALCULATE(AVERAGE('Data'[ItemRevenue]),
'Data'[Category] = EARLIER('Data'[Category]), 'Data'[AdSpend] >= 10)
    VAR RevenueWithoutAdSpend = CALCULATE(AVERAGE('Data'[ItemRevenue]),
'Data'[Category] = EARLIER('Data'[Category]), 'Data'[AdSpend] < 10)
    VAR AdSpend = CALCULATE(AVERAGE('Data'[AdSpend]), 'Data'[Category] =
EARLIER('Data'[Category]), 'Data'[AdSpend] >= 10)
    RETURN DIVIDE(RevenueWithAdSpend - RevenueWithoutAdSpend, AdSpend)
)
VAR TopCategoryTable = TOPN(1, ROASTable, [ROAS], ASC)
RETURN
MAXX(TopCategoryTable, [Category])
```

In line with the best principles of Dashboarding, this dashboard is also created with the Zig-zag reading pattern in mind, as seen on Figure 66.

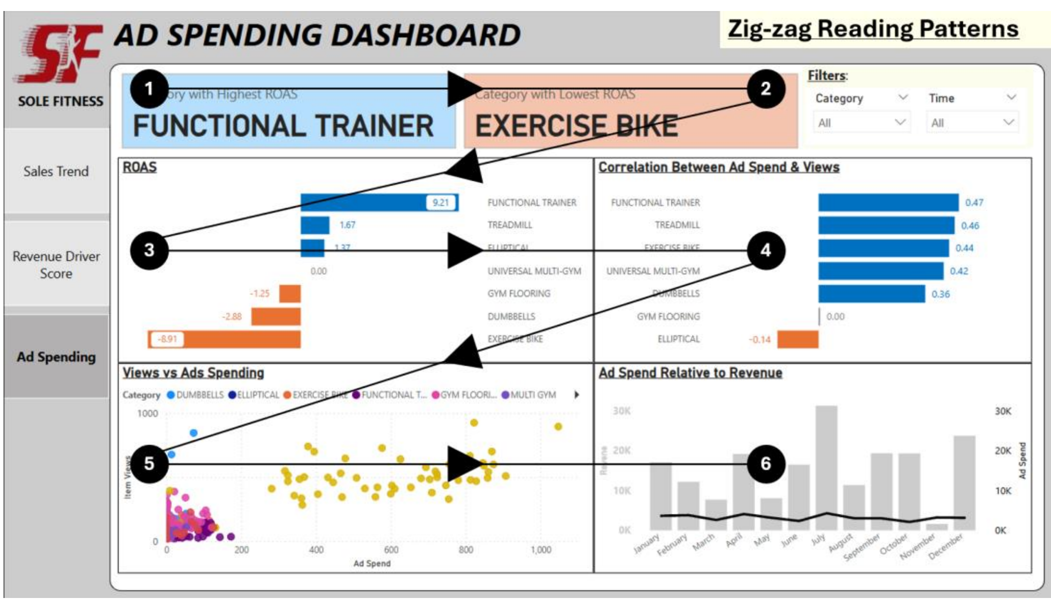


Figure 66: Sole Fitness Ad Spending Dashboard with Zig-Zag Reading Pattern Design Overlay

Readers will first have their attention directed to the most important metric: 1) Fitness Equipment Category with the Highest ROAS, which will be followed by 2) Fitness Equipment Category with the Lowest ROAS. They will then be able to glean more information from the 3) ROAS summary bar chart for all fitness equipment categories with ad spend and 4) Correlation Score Between Product Views and Ad Spend Bar Chart. This will be followed by supplementary information through the 5) Scatterplot Between Product Views and Ad Spend and 6) Combo Chart of Revenue and Ad Spend, both of which responds to the inherent Power BI Dynamic Filtering feature, making it easier to glean more information for each respective fitness equipment category.

In addition, referring to Figure 67, the Ad Spending Dashboard employs the Gestalt principle of Similarity in colour selection with colour blindness accommodation, where blue signifies positive metrics (Point 1a), orange signifies negative metrics (Point 1b) and grey signifies zero (Point 1c). The Gestalt principle of proximity is also in use with the position of the data labels being next to the individual bar on the charts (Point 3).

Moreover, to better accommodate for easier comparison between revenue and ad spend, the maximum amount on the vertical axis has been set to the same limit, while the axes themselves have their colour amended in line with the bar and line colour (Point 2).

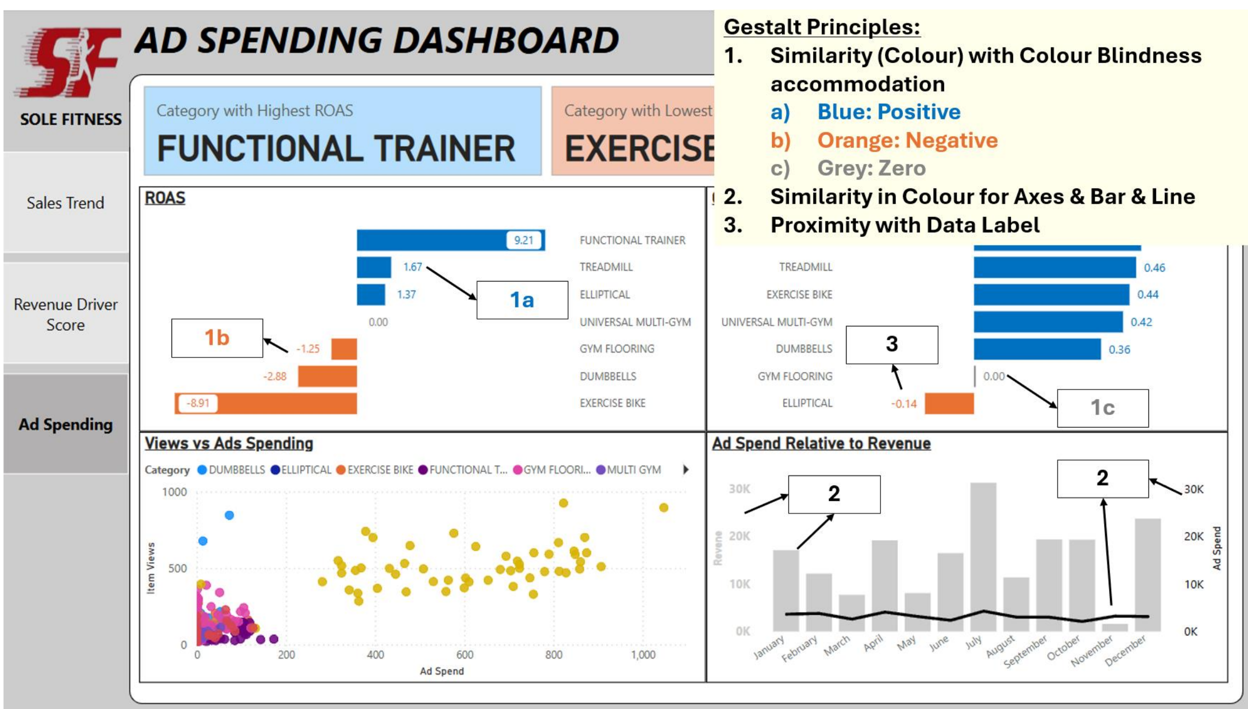


Figure 67: Sole Fitness Ad Spending Dashboard with Gestalt Principle of Similarity and Proximity Labeled

13.0 Outcome Discussion & Analysis (Insight Delivery)

The following recommendations synthesize findings from Business Problem 2 (product portfolio optimization) – Section 10.0, and Business Problem 3 (ad spend efficiency) – Section 11.0, to guide Sole Fitness in aligning resource allocation with revenue stability and advertising ROI. There is a minor change in the product prioritization.

Old Rank	New Rank	Product	Cluster	Category	Revenue Stability	Category ROAS
1	1	12-Sided Urethane Dumbbells – 20kg	Cluster 3	Dumbbells	12 out of 12 months	-2.88
2	2	Martial Arts Eva Foam Tatami Mats – Red & Blue – 40mm – Sensei Sports	Cluster 3	EVA Foam Mat	11 out of 12 months	-1.25
3	3	Commercial Gym Tiles – 20mm	Cluster 3	Rubber Gym Flooring	9 out of 12 months	-1.25
5	4	Inspire FT2 Functional Trainer	Cluster 4	Functional Trainer	4 out of 12 months	9.21
4	5	Sole E95 Elliptical Cross Trainer with Touch Screen	Cluster 4	Elliptical	8 out of 12 months	1.37
6	6	12-Sided Urethane Dumbbells – 6kg	Cluster 3	Dumbbells	12 out of 12 months	-2.88
7	7	Urethane Round Dumbbells – 16kg	Cluster 3	Dumbbells	12 out of 12 months	-2.88
8	8	Urethane Round Dumbbells – 4kg	Cluster 3	Dumbbells	12 out of 12 months	-2.88
10	9	Inspire FTX Functional Trainer	Cluster 4	Functional Trainer	4 out of 12 months	9.21
9	10	Xterra FS3.0 Ellipticial Cross Trainer	Cluster 4	Elliptical	8 out of 12 months	1.37

Table 36 Product Ranking to Work on Based on Consolidation of Analysis of Business Problem 2 and Business Problem 3

A recap on the ROAS as below:

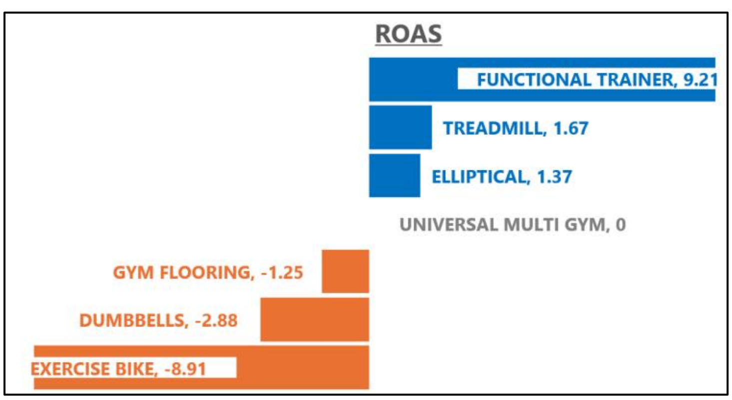


Figure 68 ROAS Results from Section 11.4.3

The justification for prioritization is split into Cluster 3 and Cluster 4.

<u>Cluster 3 – Conversion (Views to Revenue) Focus</u>

Dumbbells, despite their significantly negative ROAS (-2.88), are prioritized over Gym Flooring due to their perfect revenue stability (with revenue in 12 out of 12 months for the year) and the substantial upside potential inherent in their underperforming advertising efficiency. The severe ROAS deficit highlights critical inefficiencies in current strategies, suggesting that targeted optimizations—such as advertisement audience targeting optimization, adjustment of pricing models, or enhancing promotional messaging—could unlock disproportionately high returns.

While Gym Flooring exhibits a marginally better ROAS (-1.25), its lower revenue stability (with revenue in 9 out of 12 months for the year) limits the long-term impact of improvements. By focusing resources on Dumbbells, Sole Fitness can leverage their consistent revenue base to stabilize and amplify returns, whereas Gym Flooring's narrower upside and seasonal limitations offer less strategic value. Notably, the existing product ranking (Section 10.9.1) and ROAS metrics (Section 11.4.3) align in identifying Dumbbells as the priority, necessitating no further adjustments to the hierarchy.

<u>Cluster 4 – Scalability Focus</u>

Functional Trainers are prioritized over Ellipticals due to their exceptional ROAS (9.21), which signals strong responsiveness from audience and disproportionate revenue potential per ad dollar spent. While revenue stability is lower (with revenue in 4 out of 12 months for the year) for functional trainers, this metric underscores an opportunity to test scalable ad investments and assess whether heightened visibility can mitigate seasonal fluctuations. The high ROAS confirms that current campaigns resonate effectively, indicating that functional trainer advertisements are working very well for Sole Fitness, justifying strategic advertising budget increases to explore growth ceilings. Conversely, Ellipticals, though with more stable revenue (with revenue in 8 out of 12 months for the year), deliver diminishing returns (ROAS = 1.37) and are deprioritized to conserve resources for higher-impact opportunities.

14.0 Conclusion

In conclusion, this project delivered a scalable, data-driven solution to support Sole Fitness's growth ambitions and operational efficiency. Through the development of a unified data infrastructure, we addressed inefficiencies in manual excel data storing, mitigated product concentration risk, and established clearer links between advertising spend and product performance. The integration of ETL pipelines and dynamic dashboards enables visualized insights, empowering stakeholders to make informed, strategic decisions. Our analyses revealed actionable opportunities for product diversification and advertising optimization, laying the foundation for long-term business resilience. This project demonstrates the transformative potential of analytics in enhancing business performance in a competitive e-commerce landscape.

15.0 Future Recommendations

Currently, the available data limits the depth of analysis and modeling. To enhance feature engineering and unlock more advanced insights, the business should prioritize the collection of additional behavioral and external data. This includes metrics such as bounce rates, session duration, and customer navigation paths, as well as external signals like competitor pricing and customer sentiment from reviews. Establishing robust data pipelines for these sources is a necessary first step before implementing more sophisticated analytics and machine learning models.

To support scalable and consistent product decision-making, the development of a machine learning-based product recommendation engine is proposed. This system analyzes product performance metrics and cluster assignments to suggest specific actions such as promoting, bundling, discounting, or discontinuing products. Over time, the engine can be refined using feedback loops and business outcomes, enabling it to learn from past decisions and continuously improve its recommendations. This would serve as a valuable decision-support tool for product and marketing teams.

To further improve on the weekly aggregated Return on Advertisement Spending (ROAS) metric being used as a measure of advertisement spending effectiveness in Sole Fitness, a more specific campaign-based ROAS metric could be developed and utilized to optimize advertisement campaigns directly. This will also enable root-cause analysis to be conducted on the campaign level with such level of detail and will be able to assist Sole Fitness in increasing the effectiveness of their marketing efforts by singling out ineffective campaigns and prevent further wastage of resources.

16.0 Acknowledgements

Our team would like to express our deepest appreciation to Sole Fitness Singapore for their support and collaboration that enabled this project. We are especially grateful for their willingness to share their business data and provide candid insights into their operations.

The opportunity given to us, has been an exceptional learning experience for our team to apply classroom knowledge to real-world business challenges. The engagement with Sole Fitness has deepened our understanding of data integration, analytical modelling, and dashboard development in an e-commerce context.

We sincerely thank Sole Fitness for their time, openness and trust. Their trust has been instrumental to the completion of this project and the learning experience of our team.

And we are especially grateful to Mr. Revi Chandren, CEO of Sole Fitness, for granting us the opportunity to work with the company's business data and for his openness in sharing valuable insights that set the direction for the project.

Special thanks to the professors at NUS-ISS for their guidance and expertise throughout this project:

- Ang Ji Hua, Brian
- Brandon Ng Swee Tuan
- Charles Pang T-Howe
- Daniel Boey Swee Kee
- Dave Ho
- Marylyn Xiang

Your insights and encouragement have been instrumental in shaping the direction and outcome of our work. Thank you for your unwavering support.

17.0 References

Kovalenko, I. (2018, September 25). *Ad campaign testing. why low budgets don't work?* SmartyAds. https://smartyads.com/blog/ad-campaign-testing-why-lowest-budgets-never-yield-best-results

Cognitive Market Research. (2024, March 23). *Asia Pacific Treadmill Market is Growing at a CAGR of 7.3% from 2024 to 2031*. Cognitive Market Research. https://www.cognitivemarketresearch.com/regional-analysis/asia-pacific-treadmill-market-report

Cognitive Market Research. (2022, December 2). *The global Exercise Bikes Market size will be USD 569.6 million in 2024*. Cognitive Market Research. https://www.cognitivemarketresearch.com/exercise-bike-market-report

Cognitive Market Research. (2022, October 20). *Global Home Multi Gyms Market Report 2025 Edition, Market Size, Share, CAGR, Forecast, Revenue*. Cognitive Market Research. https://www.cognitivemarketresearch.com/home-multi-gyms-market-report