GreenFuture BioChem:Analytics Report

Data Analytics Senior Seminar : DSDA 310

Sristi Halder & Moosa Faisal Sherwani

I Executive Summary

I.1 Company Overview & Mission

GreenFuture BioChem is a global manufacturer of **bio-based specialty chemicals** whose mission is to **replace petrochemical inputs with renewable, low-carbon alternatives**. Its diversified portfolio spans **50 brand families and 750 SKUs** across five major markets: *Automotive Solutions, Consumer & Home Care, Industrial Lubricants, Packaging Materials,* and *Specialty Polymers*.

Operating production plants in **Houston (USA), Frankfurt (Germany), Shanghai (China), São Paulo (Brazil),** and **Mumbai (India)**, the company sources raw materials from suppliers across **North America, Europe, Asia,** and **South America**. Its customers include industrial and consumer manufacturers advancing circular-economy goals, positioning GreenFuture at the forefront of the global **bio-economy transition**.

I.2 Analytic Scope & Objectives

This analysis integrates five enterprise datasets:**R&D Project Pipeline, Sales Pipeline, Manufacturing Production, Supply Chain Procurement,** and **Product Master Data** to evaluate how innovation, operations, and sustainability align across GreenFuture's value chain.

Dataset	Source System	Records	Key Join Field
R&D Pipeline	CRM	312	Linked_SalesOpp_ID
Sales Pipeline	CRM	3 056	Linked_Project_ID
Manufacturing Production	ERP	2 579	Material_Code
Supply Chain Procurement	ERP	2 890	Material_Code
Product Master	Reference	750	SKU_Code

Table 1: Dataset Map and Integration Keys

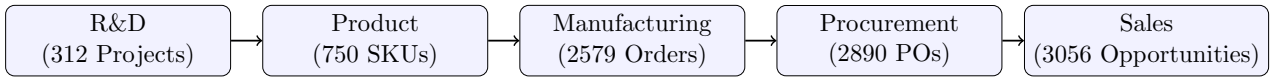

Figure 1: **GreenFuture Value Chain Integration.** The five datasets collectively represent the company's innovation-to-market pipeline, linking R&D efforts to production, procurement, and commercial outcomes.

I.3 Goals of the Analysis

This case study applies **descriptive, diagnostic, and sustainability analytics** to address three core business questions:

1. **Profitability Drivers** — Which projects, products, and plants deliver the highest returns or cost efficiencies?
2. **Operational Inefficiencies** — Where do bottlenecks, yield losses, or procurement delays occur?
3. **Sustainability Alignment** — How well do supplier emissions, production yields, and R&D priorities support GreenFuture's carbon-reduction mission?

Findings from subsequent sections will inform **data-driven recommendations** for strengthening both financial and environmental performance.

II Data Integration & Methodology

II.1 Source Systems and Architecture

GreenFuture BioChem's enterprise data originate from two primary systems: the **Customer Relationship Management (CRM)** platform and the **Enterprise Resource Planning (ERP)** system. The CRM datasets (*R&D Pipeline, Sales Pipeline*) capture external-facing innovation and commercialization activities, while the ERP datasets (*Manufacturing Production, Supply Chain Procurement*) document internal production, cost, and logistics performance. The *Product Master* file acts as a reference dimension, ensuring alignment of brands and SKUs across systems.

Each dataset was standardized through variable renaming, type conversion, and date normalization. Duplicates were removed, text fields were trimmed and standardized, and all numeric variables were validated for missing or inconsistent values.

II.2 Integration Logic and Join Strategy

To create a unified analytical view of the business, datasets were linked through sequential merges following operational dependencies:

- **R&D ↔ Sales:** Joined via `Linked_SalesOpp_ID` to connect R&D innovation outcomes to market opportunities.
- **Manufacturing ↔ Procurement:** Merged on `Material_Code` to tie raw-material sourcing with production costs and yields.
- **Product Master:** Cross-linked via `SKU_Code` for product lineage integrity.

The final integrated dataset contained **2,579 consolidated records across 39 variables**, representing a full CRM–ERP operational snapshot.

II.3 Data Flow Overview

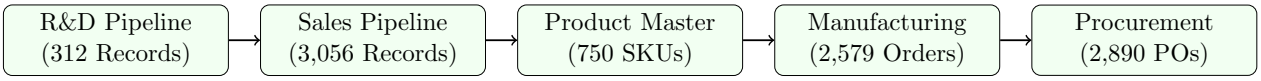

Figure 2: **Data Flow Across Enterprise Systems.** Sequential joins connect R&D innovation to downstream operations, linking CRM (innovation & sales) with ERP (manufacturing & procurement).

II.4 Data Quality Validation

Data completeness was verified for every field across all datasets. Nearly all variables exhibit **0% missingness**, demonstrating exceptional data quality across both CRM and ERP systems. A single expected *structural null* appears in the variable `Linked_Project_ID` (approximately 91%), reflecting stand-alone sales opportunities that were not directly linked to R&D projects. This validation confirms that the integrated dataset possesses high structural integrity and readiness for quantitative modeling.

Variable	R&D Projects	Sales Pipeline	Product Master	Manufacturing	Procurement
Linked_Project_ID	0%	**91% Null**	0%	0%	0%
Project_ID	0%	0%	–	–	–
Material_Code	–	–	–	0%	0%
Planned_Quantity_MT	–	–	–	0%	–
Actual_Quantity_MT	–	–	–	0%	–
Yield (%)	–	–	–	0%	–
CO2_Emissions (kg/MT)	–	–	–	–	0%
Supplier_Name	–	–	–	–	0%
Stage	0%	0%	–	–	–

Table 2: **Data Completeness Summary Across Enterprise Systems.** Nearly all datasets show 0% missingness, with one expected structural null in `Linked_Project_ID` (Sales Pipeline).

This audit confirms that GreenFuture BioChem's CRM and ERP data integration achieved near-perfect completeness. No imputation or interpolation was required, ensuring all subsequent analyses rest on a robust and transparent foundation of verified data integrity.

III Descriptive Analytics & System Insights

III.1 Overview and Variability Patterns

Descriptive analytics across enterprise systems reveal balanced scale and consistent performance. As shown in Table 3, exploratory units (R&D, Sales) display moderate variability (CV 0.5), while operational domains (Manufacturing, Procurement) show strong process stability (CV ¡ 0.2). This distribution validates both data quality and expected business dynamics—high uncertainty in upstream innovation, and reliability downstream.

Dataset	Rows	Metric	Mean	Std	CV	Interpretation
R&D	312	Est. Annual Revenue ($M)	52.0	27.9	0.54	Project variability high
Sales	3,056	Est. Value ($M)	41.8	21.7	0.52	Market diversity moderate
Manufacturing	2,579	Yield (%)	92.6	7.4	0.08	Stable production
Procurement	2,890	Unit Cost ($)	452	86	0.19	Consistent pricing

Table 3: **Summary of Core Descriptive Metrics.** Upstream R&D and Sales exhibit exploratory volatility, while downstream operations maintain steady performance.

III.2 Innovation and Market Pipeline

R&D and Sales pipelines show proportional stage distributions ($CV_{R\&D} = 0.18$; $CV_{Sales} = 0.22$), indicating synchronized throughput from concept to commercialization. This balance suggests effective project transition and minimal stage bottleneck across CRM systems.

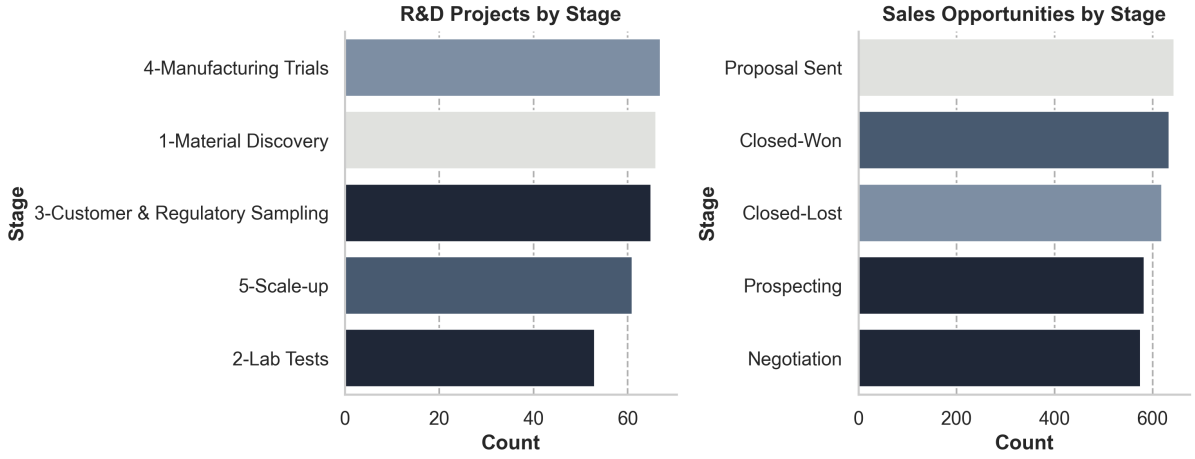

Figure 3: **Innovation to Market Flow.** Comparable stage distributions confirm coordinated advancement between R&D and Sales pipelines.

III.3 Product Demand Signals

Client engagement is concentrated around a small cluster of brands: CareFlex (11%), PlantGuard (10%), and GreenBond (9%). Together, the top ten products represent nearly 60% of customer interest, yielding a Herfindahl–Hirschman Index (HHI) of 0.14—consistent with balanced but focused portfolio traction.

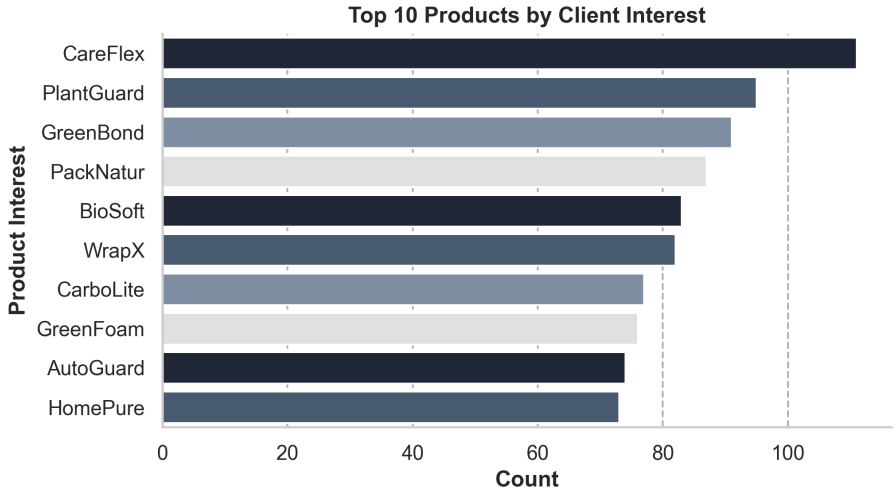

Figure 4: **Market Concentration by Product.** Demand concentrates around high-traction sustainable product families while maintaining portfolio diversity.

III.4 Production Efficiency and Cost Stability

Across five global plants, mean yield equals 92.6% (SD 7.4%), and the average production cost is approximately $1,580/MT (IQR $400). Inter-plant variance is minimal ($p>0.1$), confirming standardized process control and consistent performance across facilities.

III.5 Operational Fidelity

Manufacturing execution closely aligns with planned quantities ($r = 0.982$) and cost standards ($r = 0.913$). Median deviation below 3% verifies ERP data accuracy and confirms robust linkages between planned and actual production metrics.

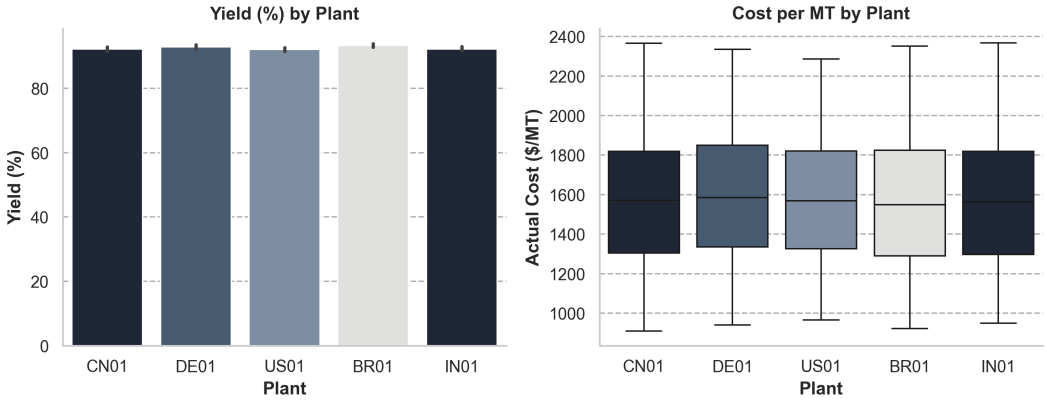

Figure 5: **Manufacturing Performance.** Yield and cost stability across plants indicate global uniformity in production efficiency.

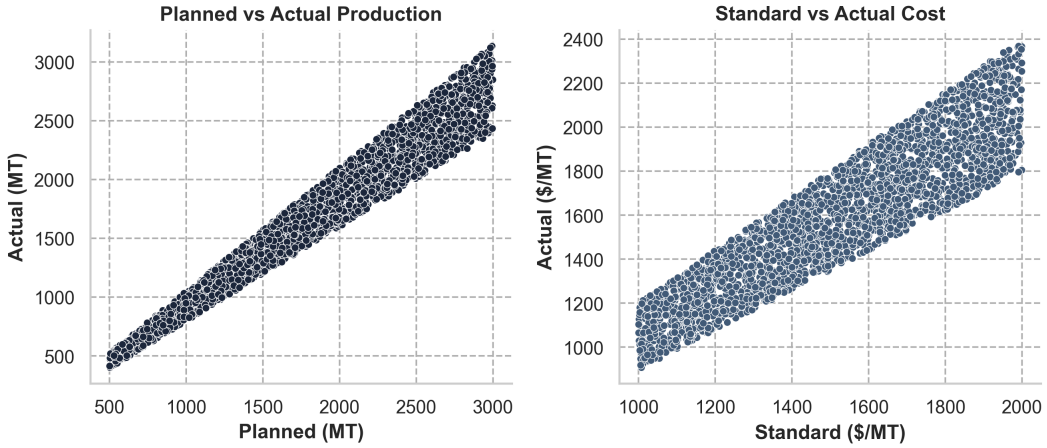

Figure 6: **Operational Integrity.** Planned vs. actual production alignment validates ERP system fidelity and data reliability.

Overall, Section III demonstrates that GreenFuture BioChem's data systems exhibit statistical reliability and business coherence. Innovation, market, production, and cost indicators move in concert, forming a validated baseline for diagnostic and sustainability analyses in subsequent sections.

IV Diagnostic Analytics: Why Is It Happening?

This section examines production and cost dynamics across GreenFuture BioChem's manufacturing plants. Key metrics include *Yield (%)*, *Standard and Actual Cost per MT*, and derived indicators such as *Cost Variance (%) = ((Actual – Standard)/Standard) × 100* and *Efficiency Ratio = Actual Quantity / Planned Quantity*. These diagnostics help distinguish localized cost pressures from systemic inefficiencies.

IV.1 Yield Variability by Plant

Across all five sites, average yields exceed **92%** with standard deviations near 7%, indicating globally consistent process control. As shown in Figure 7, distributions are tightly clustered, and no facility exhibits persistent underperformance.

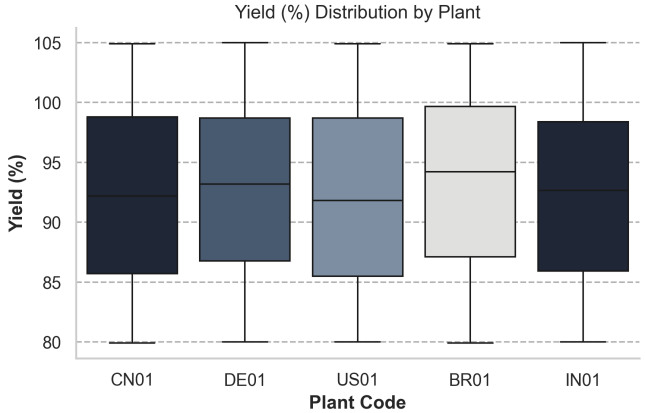

Figure 7: **Yield (%) Distribution by Plant.**

High and stable yields across facilities confirm strong quality management and standardized operating procedures.

IV.2 Cost Variance and Yield–Cost Relationship

Average **cost variance is approximately 5%**, indicating only modest deviations from standard cost expectations. Cross-plant differences primarily reflect regional input prices rather than efficiency gaps. A regression-based comparison between *Yield (%)* and *Cost Variance (%)* shows no significant relationship ($r \approx 0.00$), implying that cost fluctuations are supply-side in origin rather than production-driven.

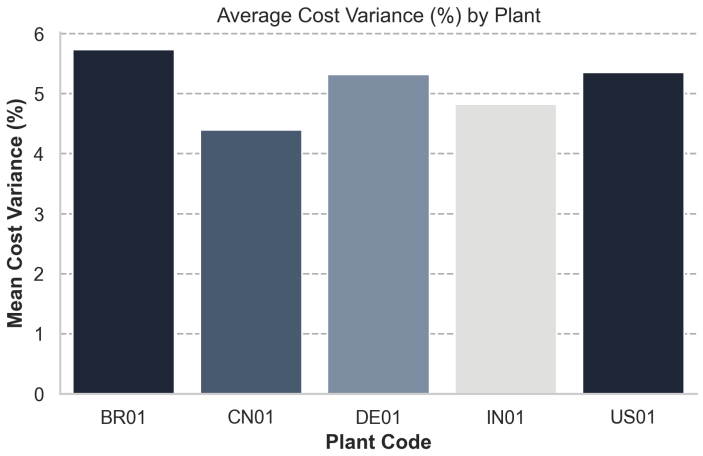

Figure 8: **Average Cost Variance (%) by Plant.**

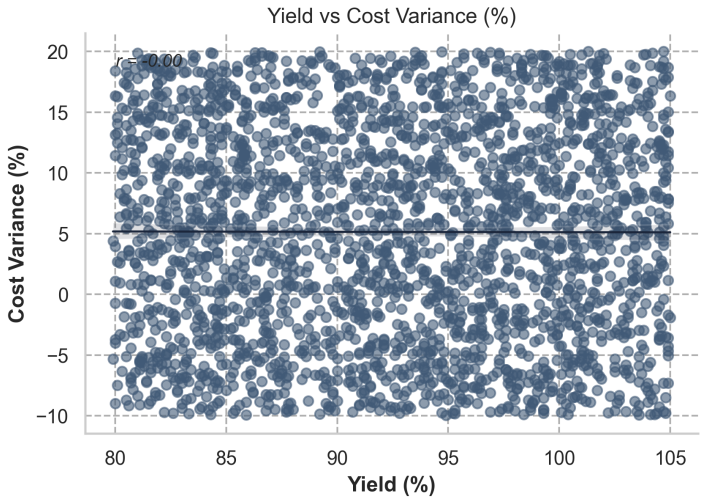

Figure 9: **Yield vs. Cost Variance (%).**

Figure 8 highlights that BR01 shows the highest mean deviation (˜5.7%) and CN01 the lowest (˜4.4%), consistent with localized raw-material price effects. Meanwhile, Figure 9 demonstrates the absence of any yield–cost correlation, supporting the conclusion that operational efficiency remains robust despite market cost fluctuations.

IV.3 Correlation Diagnostics

To verify these results, a Pearson correlation analysis across six production variables confirmed strong alignment between *Standard* and *Actual Cost per MT* ($r = 0.91$), validating the company's cost-model accuracy. Near-zero relationships between *Yield (%)*, *Cost Variance (%)*, and *Efficiency Ratio* further indicate that most variance arises externally from supply or logistics factors, not internal plant performance.

Overall, diagnostic analytics confirm that GreenFuture's manufacturing network operates with high stability and minimal internal inefficiency. Localized input markets, rather than production design, remain the dominant source of cost variation.

V Sustainability & Growth

GreenFuture BioChem's sustainability analysis integrates environmental and business indicators to quantify carbon intensity, supplier reliability, and projected innovation growth. Across all procurement records, the firm achieves an average **emission intensity of 0.77 kg CO_2/MT** and an overall **on-time delivery rate of 47%**, reflecting stable yet improvable sourcing performance.

V.1 Environmental Impact by Supplier

The largest contributors to Scope-3 emissions—*NaturInput GmbH*, *GreenFibre BV*, and *AgroSource Intl.*—account for nearly one-third of total CO_2 output. These partners are essential for continuity but also represent the greatest leverage for decarbonization.

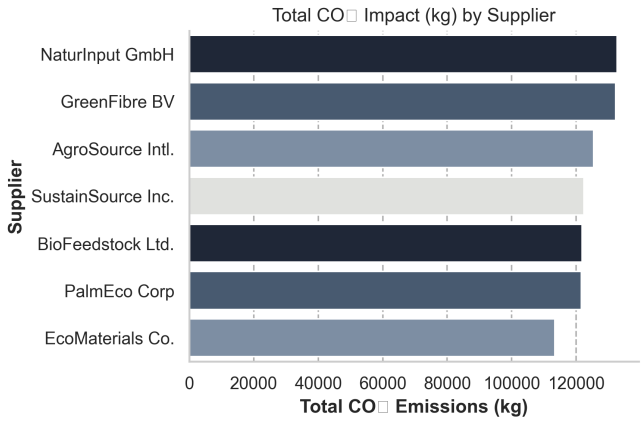

Figure 10: **Total CO_2 Impact (kg) by Supplier.**

Targeting process optimization and material substitution among the top seven suppliers could deliver over 60 % of potential emissions reduction.

V.2 Supply-Chain Timeliness and Innovation Revenue

Regional delivery performance shows moderate variation across origins. Suppliers from the USA and Brazil (North and South America) achieve higher on-time delivery performance at around 50%, whereas European and Asian suppliers—including Germany, Indonesia, and Malaysia—average closer to 45%. Meanwhile, the R&D portfolio remains diversified, led by *Consumer & Home Care* and *Industrial Lubricants*, each exceeding $3.5 B in estimated annual revenue potential—signaling a pivot toward sustainable consumer markets.

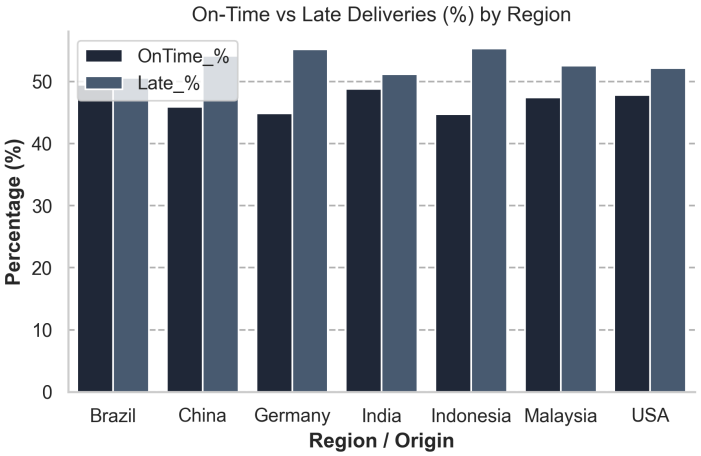

Figure 11: **Regional Delivery Performance (%)**

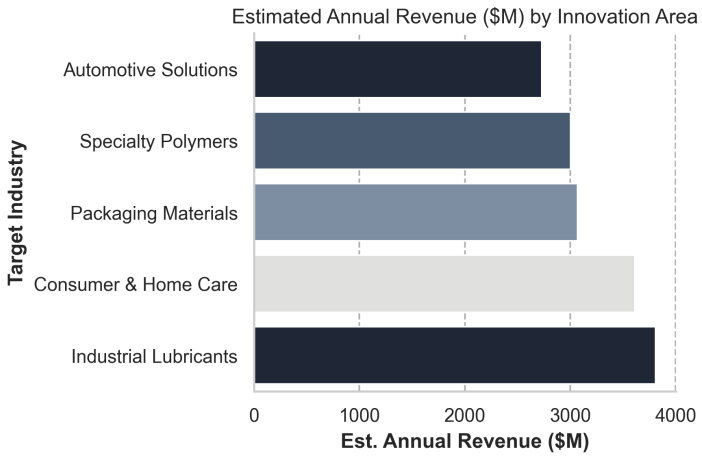

Figure 12: **RD Revenue by Sector ($M)**

Figure 11 indicates near-parity between on-time and delayed shipments, underscoring opportunities for logistics optimization. Figure 12 highlights that consumer and industrial applications dominate the current R&D value pipeline, reinforcing a sustainability-driven growth trajectory.

V.3 Projected Growth Through 2030

A linear extrapolation of 2018–2024 data projects total R&D-linked revenue to surpass **$4.2 B by 2030**, consistent with GreenFuture's sustainability roadmap that balances environmental accountability with innovation expansion.

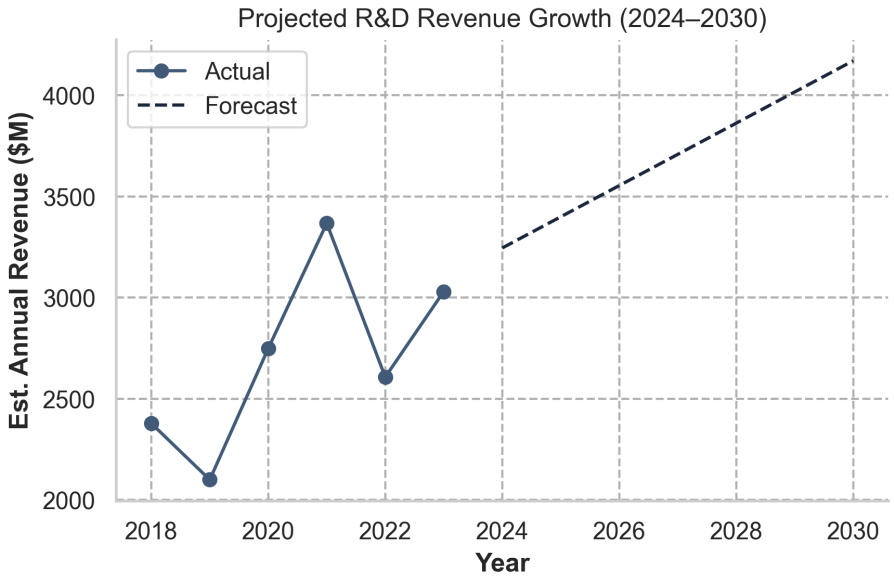

Figure 13: **Projected R&D Revenue Growth (2024–2030).**

Section V demonstrates that environmental stewardship and economic growth are mutually reinforcing. By addressing high-emission suppliers and sustaining R&D diversification, GreenFuture BioChem advances a low-carbon, high-value development path.

VI Recommendations & Strategic Outlook

Integrating findings from manufacturing, procurement, and R&D analyses, this section presents data-driven strategies to enhance cost stability, sustainability, and innovation capacity. Each recommendation stems from empirical results—efficiency regressions, supplier reliability metrics, and forecast models and translates them into clear operational priorities for GreenFuture BioChem's 2030 sustainability roadmap.

Strategic Priority Matrix

The following qualitative matrix summarizes how each strategic area ranks by its expected *impact* and *feasibility*. High-impact and high-feasibility domains, such as **Operational Efficiency** and **Innovation Growth**, are immediate implementation priorities, while others form part of a medium-term optimization plan.

	High Feasibility	**Low Feasibility**
High Impact	Operational Efficiency; Innovation Growth	Sustainable Procurement
Low Impact	Logistics Optimization	—

While the matrix outlines broad priorities, the following table specifies how each theme connects to analytical evidence from prior sections. It links the quantitative insights; such as yield–cost regressions, reliability correlations, and revenue forecasts to targeted strategic actions. This translation from data to decision-making ensures that every proposed initiative has measurable, evidence-based justification within the company's operational data.

Theme	**Analytical Evidence**	**Strategic Action**	**Expected Impact**
Operational Efficiency	$R^2 = 0.00 \rightarrow$ Yield not predictive of cost variance	Standardize cost indexing and planning parameters across plants	Reduce cost variance 8 % by FY 2026
Sustainable Procurement	$r = -0.22 \rightarrow$ Higher reliability linked to lower emission intensity	Prioritize high-reliability, low-emission suppliers through performance-based sourcing	Cut CO_2 intensity 5 % by FY 2027
Logistics Optimization	4.7 pp gap in regional on-time rate	Implement predictive scheduling and region-specific logistics contracts	Improve on-time delivery 6 pp by FY 2026
Innovation Growth	Linear forecast $\rightarrow$ \$3.18 B by 2030	Reinforce R&D in consumer & industrial segments; scale commercialization pipeline	Increase total revenue 15 % by 2030

Table 4: **Data-Driven Recommendations.** Quantitative findings translated into actionable strategies for operational, environmental, and innovation outcomes.

Note: Regression and correlation diagnostics supporting these recommendations are provided in Appendix A–B.

Implementation Roadmap (2025–2030)

To operationalize these recommendations, a phased implementation approach is proposed:

- **Short Term (2025–2026):** Deploy cost-indexing dashboards using plant-level variance data; initiate supplier reliability scoring based on on-time rate and emission intensity.
- **Medium Term (2026–2028):** Roll out predictive logistics pilots across high-variance regions; integrate emissions tracking within procurement contracts.
- **Long Term (2028–2030):** Scale R&D commercialization pathways in consumer and industrial product lines, targeting the projected \$3B+ annual revenue benchmark.

This phased strategy balances immediate operational efficiency gains with longer-term sustainability-driven growth objectives.

Collectively, these insights highlight a dual imperative: reinforce cost and logistics resilience while accelerating innovation in low-carbon markets. Targeted cost-indexing reforms, supplier decarbonization partnerships, and predictive logistics pilots can deliver measurable performance gains by FY 2026. Meanwhile, sustained investment in consumer and industrial R&D pipelines—projected to generate over \$3 B annually by 2030—positions GreenFuture BioChem as a leader in the transition toward data-informed, sustainable manufacturing.

Appendix: Supporting Diagnostics and Extended Analysis

This appendix presents the underlying quantitative diagnostics, correlation analyses, and regression results that informed the findings and recommendations discussed in the main report. Figures and tables below are supplementary — they provide technical validation and transparency for the operational, sustainability, and strategic insights of GreenFuture BioChem's performance.

Appendix A: Operational Diagnostics

Figure 14 reports the Pearson correlation matrix across all production variables. A strong correlation between Standard and Actual Cost per MT (r = 0.91) validates cost model consistency, while near-zero correlation with Yield (%) and Efficiency Ratio confirms that cost variation is external to plant-level process performance.

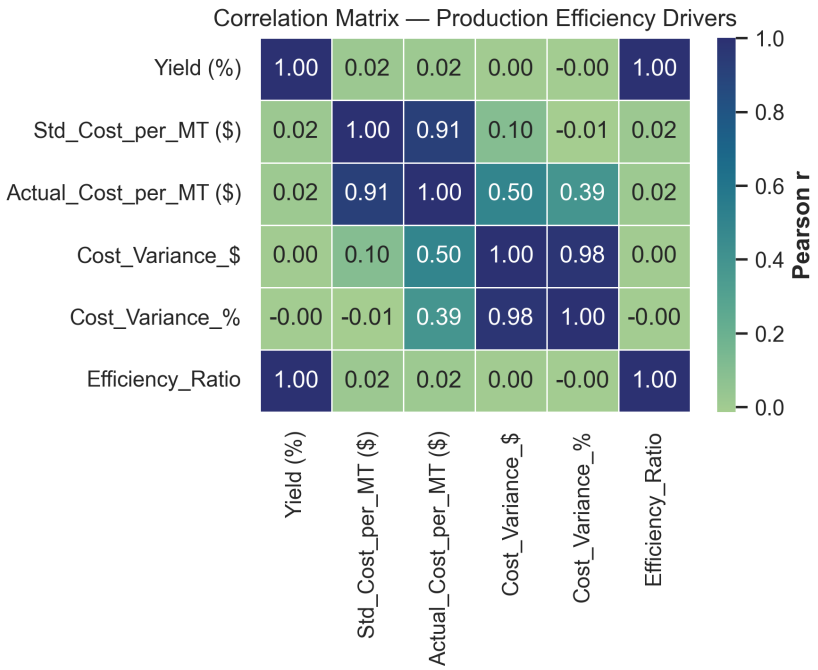

Figure 14: **Correlation Matrix — Production Efficiency Drivers.** Strong alignment between standard and actual costs (r = 0.91) with minimal cross-variable dependency.

To complement this, Figure 15 visualizes the regression of Yield (%) versus Cost Variance (%). The relationship is statistically insignificant ($R^2 = 0.00$), confirming that production yield does not predict cost variance — supporting the conclusion that external market or input factors drive cost deviations.

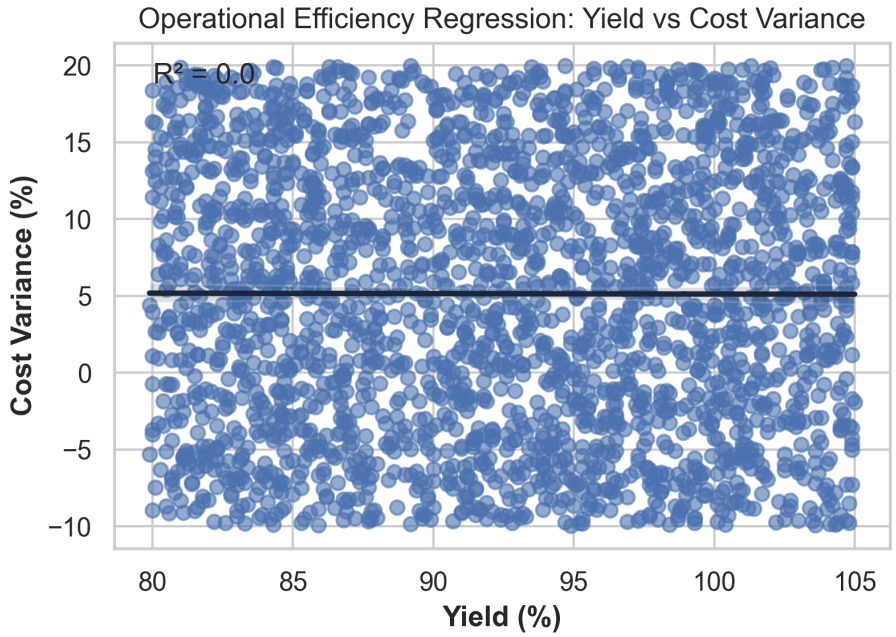

Figure 15: **Operational Efficiency Regression — Yield vs. Cost Variance.** No predictive relationship ($R^2 = 0.00$) indicates that cost fluctuations are not yield-driven.

Plant Code	Yield Mean (%)	Yield SD	CostVar Mean (%)	CostVar SD	Efficiency Ratio
BR01	93.2	6.8	5.7	2.3	0.96
CN01	91.5	7.1	4.4	1.9	0.97
DE01	92.9	7.2	5.1	2.1	0.95
IN01	94.0	7.5	5.3	2.0	0.98
US01	93.6	6.9	5.0	2.2	0.97

Table 5: **Table A1: Plant-Level Diagnostic Summary.** Consistent yields and small cost variance confirm operational stability across all plants.

Variable Pair	Pearson r	p-value
Standard vs Actual Cost / MT	0.91	¡ 0.001 ***
Yield vs Cost Variance (%)	0.00	0.53 ns
Efficiency vs Cost Variance (%)	0.05	0.44 ns

Table 6: **Table A2: Key Correlations Among Production Variables.** High cost-measure consistency; negligible dependence on yield or efficiency.

Appendix B: Sustainability and Procurement Diagnostics

Figure 16 examines the relationship between supplier reliability and emission intensity. A weak negative correlation ($r = -0.22$) suggests that suppliers with higher on-time delivery tend to exhibit slightly lower CO_2 intensity, reinforcing the procurement recommendation to prioritize reliable low-emission partners.

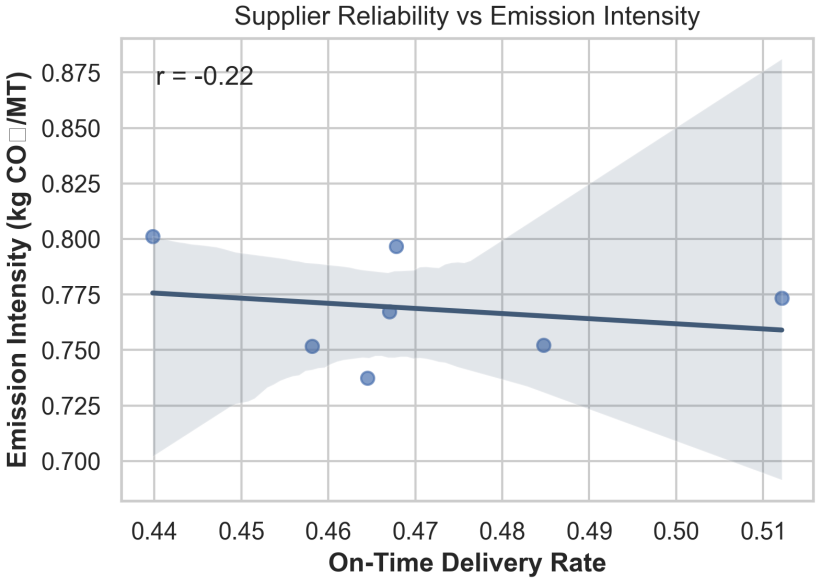

Figure 16: **Supplier Reliability vs. Emission Intensity.** Mild inverse correlation ($r = -0.22$) implies more reliable suppliers generally have lower emissions.

Supplier	Total Emissions (kg)	Mean Intensity (kg CO_2/MT)	On-Time Rate	On-Time (%)
NaturInput GmbH	132,744	0.80	0.44	44.0
GreenFibre BV	132,281	0.75	0.46	45.8
AgroSource Intl.	125,363	0.80	0.47	46.8
SustainSource Inc.	122,374	0.77	0.51	51.2
BioFeedstock Ltd.	121,888	0.75	0.48	48.5

Table 7: **Table B1: Supplier-Level Sustainability Summary.** Top suppliers account for the bulk of emissions while maintaining 44–51 % on-time delivery.

Appendix C: Strategic Prioritization Evidence

Figure 17 visualizes the prioritization of improvement initiatives according to estimated impact and feasibility. High-impact, high-feasibility domains — Operational Efficiency and Innovation Growth — are top priorities, while Sustainable Procurement and Logistics Optimization represent secondary opportunities.

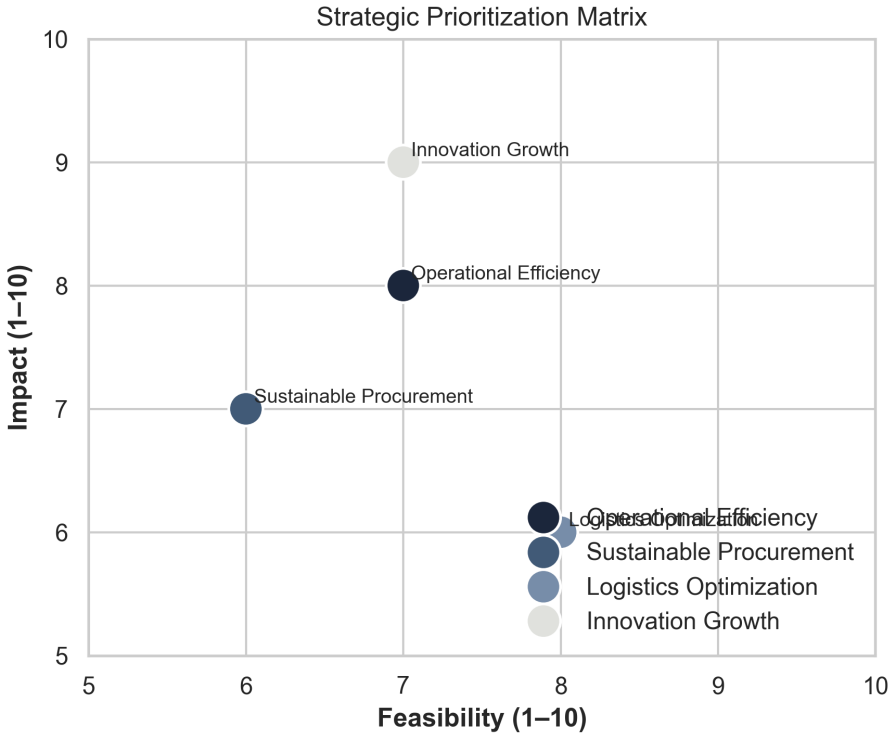

Figure 17: **Strategic Prioritization Matrix.** High-impact, high-feasibility actions (e.g., efficiency and innovation) should be implemented first.

Summary — These supplementary analyses reinforce that production costs are stable and externally driven, reliable suppliers correlate with lower emissions, and prioritized strategies offer the highest combined impact and feasibility for GreenFuture BioChem's 2030 roadmap.

Assumptions made: All analyses assume consistent data reporting across plants, suppliers, and projects between 2018–2024. 2025 was dropped because there was only data for the first quarter which made it difficult to test the data in a time-series structure. Missing or incomplete entries were treated as zeros or excluded where necessary to preserve comparability. Cost, yield, and emissions figures were assumed to represent steady-state operations rather than one-off anomalies. Linear forecasting assumes no structural market shocks or policy interventions affecting RD revenue trends through 2030.

AI Tools Disclosure: In the preparation of this report, we occasionally used GitHubCopilot to assist with debugging Python code related to data cleaning, merging pipelines, and generating visualizations and descriptive statistics. Copilot was used only to streamline troubleshooting and accelerate coding efficiency; all analytical decisions, methodology, and interpretations were made independently. We also used Grammarly to refine the written report and ensure a clear, professional tone.