BUDGET OPTIMIZATION

Data Transformation & Analysis

December 2025

Executive Summary

Goal: Combine R&D + Marketing into Total_Spend to reduce multicollinearity; deliver a stable schema for modeling.

Recommendation: Use Total_Spend + Administration as primary model inputs; use scaled versions for scale-sensitive models.

The Challenge

Our budget data contained multiple spending components that told largely the same story: R&D Spend, Marketing Spend, and Administration were highly correlated. Models using these separate inputs became unstable and unreliable. Predictions varied wildly depending on minor data fluctuations.

Problem: When predictors are highly correlated (a condition called "multicollinearity"), the model cannot reliably determine how much credit each input deserves. This creates unstable coefficients and unpredictable forecasts.

Our Solution

We engineered a single, powerful metric: Total_Spend = R&D + Marketing. This simple combination captures the overall investment magnitude without redundancy.

Why this works: Rather than asking "which component matters," we ask "how much are we investing overall, and what regional adjustments apply?" This gives us a clean, interpretable signal.

Result: Budget Core Schema (BCS)

Feature	Purpose
Total_Spend	Core signal: overall investment magnitude
Administration	Operational control & overhead effects
Profit	Business outcome (target)
State_California / State_Florida / State_New York	Regional effects (encoded 0/1)
Total_Spend_scaled / Administration_scaled	Standardized inputs for modeling

Data Quality & Transformation Pipeline

- Step 1: Load & Normalize — Standardized column names and data types.
- Step 2: Feature Engineering — Created Total_Spend = R_D_Spend + Marketing_Spend.
- Step 3: One-Hot Encoding — State_California, State_Florida, State_New York.
- Step 4: Scaling — z-score standardization for Total_Spend and Administration.
- Step 5: VIF Analysis — Validate multicollinearity before and after fix.
- Step 6: Final Schema — Model-ready BCS dataset.

Key Metrics

Rows: 993 | Columns: 11 | States present: California, Florida, New York

Scaling stats (mean, std):

- Total_Spend: mean=307572.21, std=137488.72

- Administration: mean=122967.22, std=12647.65

Fixing Multicollinearity — VIF Before/After

Original schema (R_D_Spend, Administration, Marketing_Spend):

Feature	VIF
R_D_Spend	25.16
Administration	1.63
Marketing_Spend	22.80

Reduced schema (Total_Spend, Administration):

Feature	VIF
Total_Spend	1.42
Administration	1.42

Output Artifacts Generated

- ./outputs/processed_data.csv — Model-ready BCS
- ./outputs/scaled_features.csv — Standardized inputs
- ./outputs/vif_scores.csv — Multicollinearity diagnostics
- ./outputs/Data_Transformation_Report_BCS.pdf — Full transformation report
- ./outputs/processed_outputs_bundle.zip — All artifacts bundled

Business Impact & Next Steps

- Model Stability: Lower VIF → more reliable coefficients and forecasts.
- Immediate: Fit baseline linear regression (Total_Spend + Administration), report coeffs & RMSE.

- Short Term: Cross-validated Ridge/Lasso if expanding features.
- Medium Term: Panel/time-series if temporal data is available.

Key Takeaways

- Problem Solved: Multicollinearity eliminated through smart feature engineering.
- Quality Assured: Rigorous transformation pipeline ensures data integrity.
- Ready for Action: BCS is production-ready for forecasting and planning.
- Future-Proof: Repeatable process supports advanced analytics initiatives.