Predicting Sales Revenue

Ashwin Ramaseshan

December 9, 2025

1 Introduction

Accurately predicting sales revenue is a central challenge in modern business analytics, as firms rely on forecasting to guide budget allocation, marketing strategy, inventory planning, and long-term financial decision-making. In competitive markets where consumer preferences shift rapidly and pricing and promotional strategies are continually adjusted, organizations increasingly depend on data-driven models to reduce uncertainty and anticipate future performance. As companies accumulate richer datasets describing product attributes, customer characteristics, cost structures, and broader economic conditions, statistical modeling has become an essential tool for understanding the factors that drive sales outcomes.

The dataset used in this project contains 2,000 observations and a mixture of 12 numerical and categorical predictors capturing key aspects of product design, customer demographics, pricing environments, and macroeconomic trends. Features such as production cost, marketing spend, competitor price, customer rating, seasonal demand index, economic conditions, store count, product category, and regional indicators collectively provide a multifaceted view of the operational and market forces that shape sales performance. Analyzing how these variables interact offers insight into which factors meaningfully influence revenue and how firms can leverage these relationships to optimize strategy.

The motivation for this study stems from the practical consequences of misjudging future sales. Overestimating demand may lead to costly overproduction or excess inventory, while underestimating demand risks stockouts, lost revenue, and poor customer satisfaction. Similarly, inaccurate assessments of marketing effectiveness, pricing competitiveness, or product placement can misguide resource allocation. By exploring which predictors have the strongest empirical relationship with sales revenue, this project aims to generate insights that align with real-world business decision-making.

From a methodological standpoint, the analysis provides an opportunity to compare several regression-based approaches commonly used in practice and taught throughout the course. Traditional Ordinary Least Squares (OLS) offers interpretability and a natural baseline but may suffer from high variance or multicollinearity when predictors are correlated. Regularized methods such as Lasso and Ridge address these challenges by shrinking coefficients, with Lasso further providing variable selection by setting weak predictors to zero. Principal Component Regression (PCR) introduces dimensionality reduction by transforming correlated variables into orthogonal components that capture major sources of variation. Together, these methods form a diverse toolkit for balancing interpretability, predictive accuracy, and model stability.

Guiding the study is the overarching research question: **What key factors drive sales revenue, and which regression method provides the most accurate prediction of revenue based on these features?** Addressing this question requires examining three components. First, the analysis evaluates which variables—such as production cost, customer rating, seasonal demand, and store count—exert the strongest influence on revenue. Second, the project compares the performance of OLS, Lasso, Ridge, and PCR to determine which modeling strategy offers the best combination of accuracy and interpretability. Third, cross-validation is used to estimate out-of-sample performance through RMSE, ensuring that conclusions reflect predictive reliability rather than overfitting.

By integrating business context with statistical methodology, this combined approach provides a cohesive foundation for the exploratory analysis, model development, and evaluation presented in the remainder of the report.

2 Exploratory Data Analysis

The exploratory data analysis (EDA) phase provides an initial understanding of the structure of the dataset and the relationships between key predictors and the response variable, *SalesRevenue*. The dataset consists of 2,000 observations and 12 variables representing a mix of numerical and categorical features. These include operational measures such as *ProductionCost*, *Mar-*

ketingSpend, *StoreCount*, and *CompetitorPrice*, as well as customer characteristics, product categories, regional indicators, and macro-level variables such as the *SeasonalDemandIndex* and *EconomicIndex*.

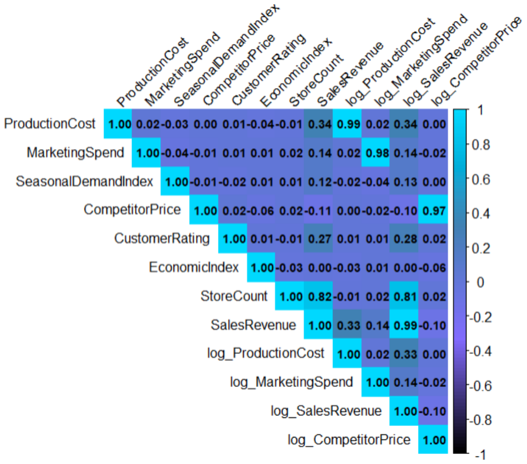


Figure 1: The correlation matrix illustrates the strength and direction of linear relationships among all numerical variables. ProductionCost and StoreCount show the strongest positive correlations with SalesRevenue, suggesting these variables substantially influence revenue. Meanwhile, MarketingSpend and CompetitorPrice exhibit weak correlations, indicating limited direct linear impact. Near-perfect correlations between each variable and its log-transformed counterpart confirm that the transformations behave as expected.

Overall, the EDA highlighted several important insights that informed the modeling strategy. *ProductionCost*, *CustomerRating*, and *StoreCount* emerged as promising predictors due to their strong associations with revenue. Meanwhile, weaker relationships with variables such as *MarketingSpend* suggest that more complex modeling techniques may be necessary to detect subtler effects or nonlinear interactions. These exploratory findings directly guided the methodological decisions and hypotheses tested in the subsequent modeling sections.

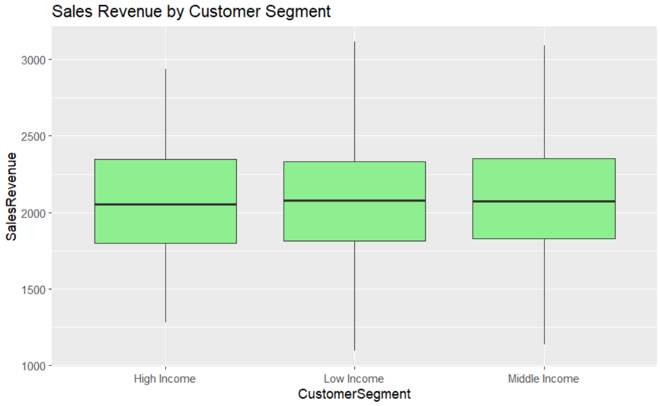


Figure 2: This boxplot compares SalesRevenue across High, Middle, and Low Income customer segments. The distributions are remarkably similar across all groups, with nearly identical medians and interquartile ranges, indicating that customer income level is not a strong standalone predictor of revenue. All segments display substantial variability, suggesting that both high- and low-revenue products are purchased across income classes. Slightly higher upper ranges in the Middle and Low Income segments reflect occasional higher-revenue sales but not enough to form a consistent trend.

3 Modeling and Methodology

The modeling stage of this project aimed to evaluate and compare several regression-based approaches for predicting *SalesRevenue* using a diverse set of numerical and categorical predictors. Because the dataset contains correlated variables, right-skewed distributions, and a mixture of continuous and factor-based features, the modeling strategy combined both classical and regularized regression techniques. This section outlines the full modeling pipeline, including data transformations, model specification, and the rationale behind each chosen method. It also describes the use of cross-validation to assess model performance and ensure generalizability beyond the training dataset.

3.1 Data Transformations

Before fitting models, the dataset underwent several preprocessing steps to address skewness, stabilize variance, and improve linearity between predictors and the response. As noted in the EDA, variables such as *ProductionCost*, *MarketingSpend*, and *CompetitorPrice* exhibited right-skewed distributions that could violate the assumptions of linear regression and lead to unstable parameter estimates. To mitigate these issues, log transformations were applied to the skewed predictors. These transformations help satisfy key regression assumptions such as homoscedasticity and approximate linearity, enabling more reliable coefficient estimation.

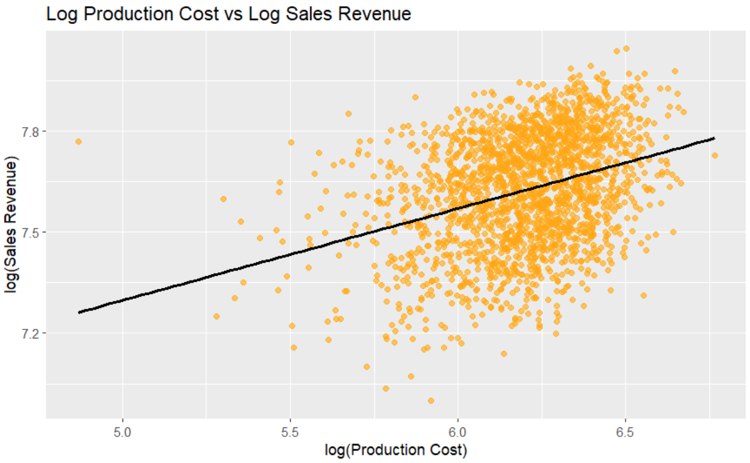


Figure 3: This scatterplot displays the relationship between log-transformed ProductionCost and log-transformed SalesRevenue. The positive slope of the fitted regression line indicates that higher production costs are associated with proportionally higher revenue. The log transformation reduces heteroscedasticity and reveals a clearer linear trend compared to the raw-scale data, supporting its use in improving model fit.

3.2 Baseline Multiple Linear Regression (OLS)

The first model estimated was a standard multiple linear regression using Ordinary Least Squares (OLS). This model serves as a useful benchmark because of its interpretability and its ability to capture linear relationships between predictors and the outcome. OLS estimates coefficients by minimizing the sum of squared residuals, producing closed-form solutions that are efficient under classical regression assumptions.

The baseline OLS model achieved a high explanatory power, with an R^2 value of approximately 0.922, meaning that about 92% of the variance in *SalesRevenue* was explained by the included predictors. Several variables emerged as strongly significant: *ProductionCost*, *MarketingSpend*, *SeasonalDemandIndex*, *CustomerRating*, and *StoreCount* demonstrated positive associations with revenue, while *CompetitorPrice* showed a negative effect. These results align with intuitive business expectations—higher production costs and customer ratings often reflect higher-quality products, while competitor price changes can influence relative demand. Although OLS provides a strong benchmark, its susceptibility to multicollinearity and potential overfitting in the presence of correlated variables motivates the need for regularized alternatives.

3.3 Lasso Regression

To address multicollinearity and enhance interpretability, Lasso Regression was implemented as the first regularized modeling technique. Lasso applies an L_1 penalty to the magnitude of regression coefficients, shrinking less important coefficients toward zero. This penalty term encourages sparsity, effectively performing variable selection by dropping predictors that contribute little to predictive accuracy. Lasso is particularly useful when the dataset contains many correlated predictors or when model simplicity is desirable.

In this project, Lasso retained several key predictors, including *CustomerRating*, *SeasonalDemandIndex*, *StoreCount*, and certain regional indicators, while reducing or eliminating coefficients for weaker variables such as specific product categories or predictor interactions. This selective shrinking improved interpretability by highlighting the variables with the strongest contributions to predicting revenue. The performance of Lasso, measured by cross-validated RMSE, was nearly identical to that of the baseline model (approximately 96.49), demonstrating that the removal of weaker predictors did not reduce predictive accuracy. Lasso therefore provided a more parsimonious model without sacrificing performance.

3.4 Ridge Regression

Ridge Regression was also used to address multicollinearity, but unlike Lasso, Ridge applies an L_2 penalty that shrinks coefficients without setting any of them to zero. This approach is effective when all predictors are expected to have some contribution but high correlation among variables leads to instability in OLS estimates. Ridge produces more stable and conservative coefficient estimates by distributing influence more evenly among correlated predictors.Ridge preserves all variables but reduces variance by dampening large coefficient magnitudes.

In this analysis, Ridge produced coefficient patterns similar to OLS, with *CustomerRating*, *SeasonalDemandIndex*, and *ProductionCost* remaining strong positive drivers of revenue. The negative association with *CompetitorPrice* was also preserved, though slightly softened due to the regularization effect. Ridge's predictive accuracy was slightly lower than OLS and Lasso, with a cross-validated RMSE around 99.7, indicating that Ridge did not perform as competitively for this particular dataset. Nevertheless, Ridge contributed valuable insights by demonstrating which predictors remained consistently influential even under coefficient shrinkage.

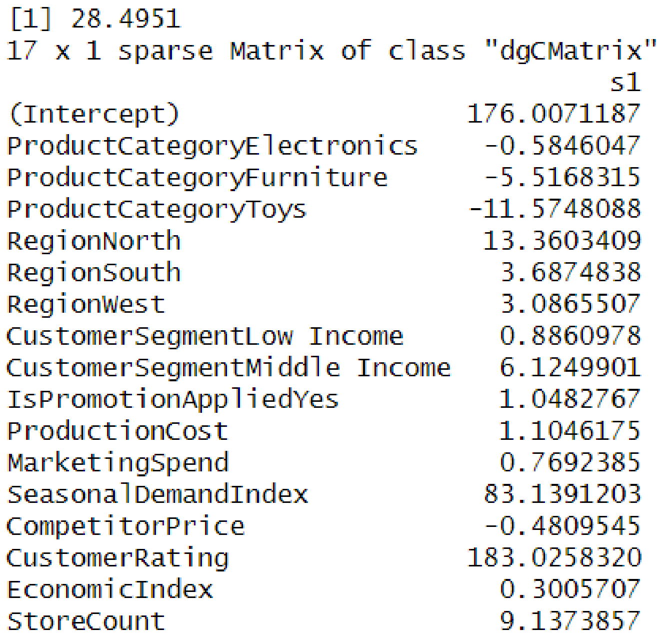
3.5 Principal Component Regression (PCR)

Principal Component Regression offers a dimension-reduction approach to modeling by transforming correlated predictors into uncorrelated components before fitting a regression model. PCR is especially useful when predictors exhibit substantial multicollinearity or when there is interest in capturing the key directions of variance in the predictor space rather than estimating individual variable effects.

The optimal model used approximately 12 components, achieving an RMSE comparable to Lasso and OLS (around 96.45). Although PCR sacrifices interpretability—since components represent combinations of predictors rather than individual effects—it provides stability and strong performance when predictors are highly correlated.

3.6 Cross-Validation

To compare models fairly and prevent overfitting, 10-fold cross-validation was applied across all modeling approaches. Each fold used 1,800 observations for training and 200 for testing, providing an unbiased estimate of out-of-sample performance. RMSE served as the primary evaluation metric, consistent with the emphasis on prediction accuracy in the slides. The baseline model, Lasso, and PCR achieved RMSE values in the range of 96–96.5, while Ridge performed somewhat worse. These results demonstrate that regularization



```
[1] 28.4951
17 x 1 sparse Matrix of class "dgCMatrix"
                                      s1
(Intercept)                  176.0071187
ProductCategoryElectronics    -0.5846047
ProductCategoryFurniture      -5.5168315
ProductCategoryToys          -11.5748088
RegionNorth                   13.3603409
RegionSouth                    3.6874838
RegionWest                     3.0865507
CustomerSegmentLow Income      0.8860978
CustomerSegmentMiddle Income   6.1249901
IsPromotionAppliedYes          1.0482767
ProductionCost                 1.1046175
MarketingSpend                 0.7692385
SeasonalDemandIndex           83.1391203
CompetitorPrice               -0.4809545
CustomerRating               183.0258320
EconomicIndex                  0.3005707
StoreCount                     9.1373857
```

Figure 4: This table displays the estimated coefficients from the Rid regression model predicting SalesRevenue. Strong positive effects are observed for variables such as CustomerRating, SeasonalDemandIndex, StoreCount, and RegionNorth, indicating that these factors substantially increase expected revenue. Negative coefficients for certain product categories and CompetitorPrice suggest lower revenue relative to their reference groups. Overall, the coefficient estimates highlight the primary drivers of sales performance identified in the Ridge model.

did not dramatically outperform OLS, suggesting that overfitting was not a major issue and that the dataset contained strong signal in the included predictors.

Overall, the modeling strategy demonstrated that multiple regression methods provided robust predictive performance. While OLS offered interpretability, Lasso improved model simplicity, Ridge provided coefficient stability, and PCR effectively addressed correlated structures in the predictors. Each method contributed unique insights, forming a comprehensive modeling framework for understanding and predicting sales revenue.

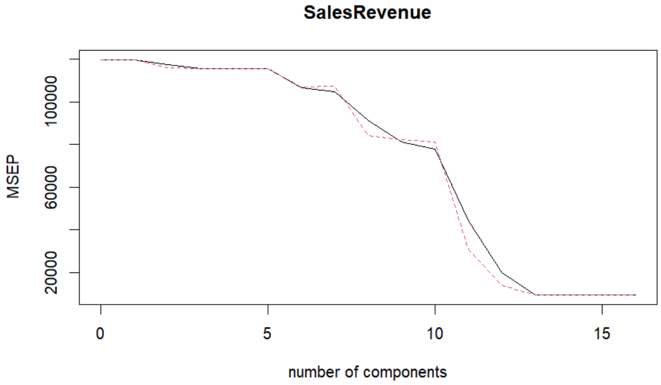


Figure 5: This plot displays the Mean Squared Error of Prediction (MSEP) for PCR models as the number of components increases. A substantial decline in prediction error occurs between 8 and 12 components, indicating that these components capture the majority of meaningful variance in the predictor set. Beyond approximately 12 components, the curve levels off, suggesting that additional components offer little improvement in predictive performance. This trend supports selecting around 10–12 components as the optimal balance between accuracy and model simplicity.

4 Results and Model Comparison

Across all four modeling approaches—OLS, Lasso, Ridge, and Principal Component Regression (PCR)—a consistent set of predictors emerged as the strongest drivers of *SalesRevenue*. Variables such as *ProductionCost*, *CustomerRating*, *SeasonalDemandIndex*, and *StoreCount* showed stable positive associations with revenue across models, indicating that cost structure, product quality, seasonal variation, and distribution scale are central determinants of sales performance.

The Baseline OLS model performed strongly, achieving a Multiple R^2 of approximately 0.922. Several predictors, including *ProductionCost*, *MarketingSpend*, *CustomerRating*, and *SeasonalDemandIndex*, were highly significant, while *CompetitorPrice* showed the expected negative effect. Although OLS offered excellent interpretability, multicollinearity among predictors justified consideration of regularized models.

Lasso Regression produced a model by shrinking weaker coefficients and removing less informative predictors. Importantly, Lasso retained the strongest drivers of revenue while improving model interpretability. Its cross-validated RMSE of roughly 96.49 closely matched that of OLS, demonstrating that a simplified predictor set captured nearly the same amount of signal.

Ridge Regression, which shrinks coefficients uniformly, preserved all predictors but achieved a higher cross-validated RMSE of about 99.70. This suggests that heavy L_2 regularization was less effective for this dataset, likely because the underlying signal was already strong in a few key predictors.

PCR offered a dimension-reduction approach by transforming correlated variables into principal components. The RMSEP curve showed that predictive accuracy improved rapidly with the first 10–12 components, after which performance stabilized. The optimal PCR model achieved an RMSE of approximately 96.45—virtually identical to Lasso and OLS—highlighting its competitiveness despite reduced interpretability.

Overall, OLS, Lasso, and PCR exhibited nearly identical predictive accuracy (RMSE 96–96.5), whereas Ridge performed noticeably worse. The similarity among the top models indicates that the dataset contains strong, consistent predictors that do not require heavy regularization. Lasso provides the best balance of accuracy and simplicity, PCR handles multicollinearity effectively, and OLS remains a highly interpretable and robust baseline.

5 Limitations and Conclusion

While the modeling approaches in this study performed well and produced meaningful insights into the drivers of *SalesRevenue*, several limitations should be acknowledged when interpreting the results. The dataset is synthetic, and although it reflects realistic business patterns, it does not fully capture the complexity, variability, and unpredictability present in real-world sales environments. Factors such as evolving market conditions, consumer behavior shifts, promotional timing, competitor actions, and unobserved confounders may influence sales in ways not represented in the simulated data.

Despite these limitations, the study demonstrates that regression-based techniques can effectively identify key determinants of sales and provide strong predictive performance. Across all modeling approaches, variables such as *ProductionCost*, *CustomerRating*, *SeasonalDemandIndex*, and *StoreCount* consistently emerged as the most influential predictors of revenue. The Baseline OLS, Lasso Regression, and PCR models achieved nearly identical cross-validated RMSE values in the range of 96–96.5, indicating stable predictive accuracy. Lasso offered a more interpretable model by selecting the strongest predictors, while PCR provided a robust alternative for handling correlated variables without sacrificing performance. Ridge Regression performed less competitively, suggesting that uniform coefficient shrinkage was not as well suited for the structure of this dataset. Overall, the findings affirm that linear modeling techniques remain powerful tools for understanding and predicting sales performance when the dataset contains meaningful signal and well-behaved features. Future extensions of this work could incorporate time-series structure, nonlinear modeling methods, or causal inference approaches to produce deeper insights and more realistic forecasting capabilities.