# **SUPPLY CHAIN ANALYSIS**

**RESEARCH QUESTIONS**

**RESEARCH QUESTION 1: How can we segment customers based on their purchasing behavior, shipping preferences, and order history?**

***PHASE 1***

#### The first phase of the project focused on identifying key variables that influence customer behavior and could be used for segmentation. These variables included Customer ID to track behavior at an individual level, sales per customer to capture monetary value, product category to understand buying preferences, shipping mode to reflect logistical preferences, order frequency to differentiate regular vs. one-time buyers, late delivery risk to measure customer experience, and geographic data (City, State, Country, Latitude, Longitude) to explore regional differences. While no specific algorithmic approach (such as RFM or clustering) was defined at this stage, the emphasis was on leveraging business-driven variables to support targeted marketing, logistics optimization, and customer retention strategies.

#### **PHASE 2**

#### In the implementation phase, segmentation was carried out using a structured analytical approach. First, RFM (Recency, Frequency, Monetary) analysis was performed, where customers were scored using quantile-based binning across the RFM dimensions to capture recent purchase activity, transaction frequency, and total spend. These RFM scores were then used as inputs for K-Means clustering, with the elbow method identifying four optimal clusters labeled as Loyal High-Value, New/Low-Value, Inactive Mid-Spenders, and Engaged Mid-Value. To enhance interpretability and reporting, Principal Component Analysis (PCA) was applied to reduce the 3D RFM data into a 2D space for visualization.

#### **REASON**

The initial goal in Phase 1 was broad segmentation based on multiple dimensions of behavior and preference. However, during implementation, the team focused specifically on transactional behavior, using RFM analysis as a proven and interpretable metric. This allowed for:

* Clear grouping of customers based on value and loyalty
* Practical cluster creation using unsupervised learning (K-means)
* Simplified visuals through PCA

The shift toward RFM and clustering offered a more structured, scalable, and business-relevant segmentation method while maintaining the spirit of the original research question.

**RESEARCH QUESTION 2: Can we predict future sales based on past transaction data, seasonality, and regional sales trends?**

**PHASE 1**

In the initial phase of the project, the focus was on predicting future sales using past transaction data, seasonality, and regional sales trends. The proposed approach included using features such as order date, sales per customer, product category, shipping mode, order quantity, and regional details like city, state, and country. Two forecasting models were planned: ARIMA, for short-term interpretable forecasting, and LSTM, for capturing long-term patterns in sequential sales data. The reasoning for these choices was that ARIMA works well for smaller, stable datasets, while LSTM is more effective when long-term dependencies exist. Managerial applications of this forecasting model were clearly outlined, such as helping with inventory management, financial planning, regional expansion, and seasonal discount planning.

**PHASE 2**

In the implementation phase, several models were tested, including baseline models (naïve, moving average), ARIMA (1,1,2), and various deep learning models such as univariate LSTM, multivariate LSTM, stacked LSTM, and a weighted ensemble of LSTM and naïve predictions. Data preparation focused mainly on sales and time-based features, such as day, month, year, lag values, and moving averages. Unlike Phase 1, regional data like city or state, as well as product category and sales per customer, were not included in the final modeling. The best-performing model was the weighted ensemble of LSTM and naïve predictions, which achieved the lowest RMSE and MAE among all approaches.

**REASON**

While Phase 1 aimed to incorporate cross-sectional diversity (customer behavior, product trends, regional variation), Phase 2 focused entirely on engineered time series features. This shift was likely driven by two key reasons. First, deep learning models like LSTM require consistent, numeric input, and incorporating categorical fields like region or shipping mode would have required additional preprocessing and complexity. Second, the most noticeable pattern in the data was time-driven, particularly a sharp sales spike in the test set, which was better handled by focusing on time lags and moving averages rather than cross-sectional differences.

In summary, the research question remained the same across both phases, but its execution shifted toward a data-driven solution that prioritized simplicity, model stability, and time-based predictability. The focus on stacking and ensemble methods in Phase 2 also reflected a practical adjustment to improve model robustness, especially given the sales volatility and data limitations observed in the test set.

**RESEARCH QUESTION 3: Can we identify fraudulent transactions based on order patterns, payment types, and late delivery risks?**

**PHASE 1**

In Phase 1, our objective was to explore the possibility of detecting fraudulent transactions by analyzing order patterns, payment types, and late delivery risks. The initial focus was on understanding the factors contributing to fraudulent activities within the supply chain. We began with basic data preparation, including cleaning the dataset to remove null values and standardizing columns. The analysis focused on key variables such as Payment Type, Order Amount, Late Delivery Risk, and Order Frequency, which were identified as potential indicators of fraudulent behavior. Our approach primarily relied on descriptive analysis to understand transaction patterns and identify any anomalies that could signal fraud. Although no sophisticated modeling techniques were applied, this initial phase laid the groundwork for fraud detection by revealing patterns such as high-risk payment methods, unusual order frequencies, and the correlation between delayed deliveries and potential fraud.

**PHASE 2**

In Phase 2, we advanced fraud detection by implementing a multi-model framework, including Logistic Regression, Random Forest, Gradient Boosting (LightGBM, XGBoost), and Artificial Neural Networks (ANN). Data preprocessing involved encoding variables, handling imbalance with SMOTE, and feature engineering. Model performance was evaluated using metrics like Accuracy, Precision, Recall, with XGBoost achieving the best results. This approach significantly improved fraud detection accuracy, providing a scalable, data-driven solution to mitigate financial risks.

**REASON**

The transition from Phase 1 to Phase 2 was driven by the need for a more sophisticated and reliable approach to fraud detection. In Phase 1, the analysis was limited to descriptive statistics, which, while useful for identifying basic patterns, lacked the ability to detect complex fraudulent behaviors. Phase 2 addressed this limitation by leveraging advanced machine learning models capable of automatically identifying fraudulent transactions based on complex patterns in the data. This shift from a simple, manual analysis to an automated, data-driven approach significantly improved fraud detection accuracy, providing a scalable solution that can be continuously updated to counter emerging fraud patterns.