**EBA 5005 Practice Module in Specialised Predictive Modelling & Forecasting**

**Final Report for Optimising Supply Chain Efficiency: An Analysis of Sales, Pricing, and Customer Behaviour Using the DataCo Dataset**

**EBA5005\_SmartDataCo.doc**

**Date of Report**

17 November 2024

**Team Name or Group No**

Smart DataCo

**Team Members**

Lam Chui Mei (A0269736L)

Lee Mei Teng (A0229878E)

Muhammad Hyqel Bin Razali (A0266745R)

Priya Deshpande (A0285996Y

### Table of Contents

1. **Introduction and Background**
2. **Industry Overview**
3. **Problem Statement & Objectives**
4. **Project Design**
5. **Scope of Work**
   * 5.1 Dataset Used and Data Description**,**Data Collection and Preparation
   * 5.3 Analytical Models and Techniques
   * 5.4 Evaluation Methodologies
6. **Data Understanding and Data Processing**
7. **Modelling and Evaluation**
   * 7.1 Time-Series Forecasting
     + 7.1.1 Objectives
     + 7.1.2 Methodology
     + 7.1.3 Model Evaluation
     + 7.1.4 Results
   * 7.2 Survival Analysis
     + 7.2.1 Objectives
     + 7.2.2 Methodology
     + 7.2.3 Model Evaluation
     + 7.2.4 Results
   * 7.3 Geographical Analysis
     + 7.3.1 Objectives
     + 7.3.2 Methodology
     + 7.3.3 Model Evaluation
     + 7.3.4 Results
   * 7.4 Fraud Detection
     + 7.4.1 Objectives
     + 7.4.2 Methodology
     + 7.4.3 Model Evaluation
     + 7.4.4 Results
   * 7.5 Repricing Strategies
     + 7.5.1 Objectives
     + 7.5.2 Methodology
     + 7.5.3 Model Evaluation
     + 7.5.4 Results
8. **Outcome Discussions and Analysis**
9. **Recommendations and Prescriptive Measures**
10. **Conclusions**
11. **Acknowledgements**
12. **References**

# Executive Summary

This report presents a data-driven approach to optimise supply chain performance for DataCo, focusing on sales, pricing, and customer behaviour insights. Our analysis utilised advanced analytics techniques to address key objectives, including delivery performance, sales and profitability trends, shipping and order status, geographic risk, and product category performance. By targeting these areas, the project aims to uncover actionable insights for enhancing efficiency, reducing delays, and improving customer satisfaction.

**Key Findings and Analyses**

1. **Delivery Performance and Risk Analysis**: Delivery delays pose a significant risk, particularly in regions with consistent shipping challenges, such as Puerto Rico. Analysis of ‘Days for shipping (real)’ versus scheduled times revealed patterns in delays. This insight led to the development of a Region-Specific Inventory Optimisation plan for high-risk areas, recommending buffer stock levels to mitigate delays.
2. **Sales and Profitability Insights**: Sales-centric segmentation highlighted high-sales and high-elasticity items, especially during volatile periods. These insights will guide targeted promotional strategies and discount optimisation. Additionally, we examined 'Sales After Discount' as a key metric, identifying top-performing states and flagging those with declining trends to optimise resource allocation across regions.
3. **Customer Retention** : Survival Analysis conducted to investigate the statistically significant factors impacting potential churn risks. This approach supports personalised marketing strategies and loyalty program initiatives to enhance customer retention and engagement.
4. **Product-Specific and Geographic Segmentation**: Our analysis uncovered product-specific demand trends and geographic variance in demand volatility, assisting in tailored inventory strategies. This segmentation aids in forecasting demand more accurately by product and region, further supporting inventory optimisation.
5. ~~Temporal Analysis for Strategic Period Identification: A focused analysis on key sales periods allowed us to identify high and low-demand phases. By aligning promotions, staffing, and stock levels with these periods, DataCo can ensure peak efficiency during high-demand times and reduce overhead in slower periods.~~

**Recommendations**

Our findings emphasise the need for region-specific inventory adjustments and a data-driven approach to marketing. The use of PyCaret’s time series AutoML for demand forecasting enables precision in managing volatile periods, while our analysis of discount impact and profitability offers a basis for pricing strategies that balance competitiveness and margin protection. This project ultimately recommends leveraging advanced analytics to build an agile, responsive supply chain framework that supports DataCo’s operational and customer satisfaction goals.

# Introduction

The supply chain industry is a critical component of the global economy, playing a central role in the movement of goods, information, and services from producers to consumers. As global markets become increasingly interconnected and competitive, the demand for supply chain efficiency, flexibility, and responsiveness has surged. With the rise of digital technologies, big data analytics, and automation, businesses are increasingly adopting advanced tools and methods to optimise their supply chain operations. This transformation is not only aimed at reducing costs but also enhancing customer satisfaction, improving product availability, and mitigating risks associated with logistics and distribution.

This project leverages advanced analytics on the Dataco’s, a supply chain organisation, dataset to improve supply chain efficiency, optimise sales performance, enhance pricing strategies, and increase customer retention. The goal is to foster data-driven decision-making within the organisation through the implementation of predictive models, fraud detection systems, and dynamic repricing strategies. The project design incorporates both business and technical objectives that align with enhancing operational efficiency and profitability.

#### 1.1 Business Objectives Summary

The project aims to achieve the following business objectives:

* **Enhance Sales Performance**: By developing dynamic pricing strategies, including repricing models that adapt to market demands and competitive factors.
* **Optimise Customer Retention and Delivery Performance**: Improve customer loyalty and reduce churn by analyzing behavior patterns, allowing for targeted retention efforts that enhance customer satisfaction.
* **Fraud Detection**: Minimize financial losses and protect the company's reputation by identifying and preventing fraudulent transactions through automated detection models.
* **Geographical Optimisation**: Identifying regions with varying demand patterns and sales volatility through geographical analysis, which will support inventory optimisation and strategic marketing.

#### 1.2 Technical Objectives Summary

Technically, the project aims to:

* **Time-Series Forecasting:** Develop predictive models to accurately forecast sales and demand patterns, supporting data-driven pricing strategies and efficient inventory management.
* **Survival Analysis:** Implement survival analysis methods to predict customer retention trends and assess the longevity of customer relationships, enabling targeted retention strategies.
* **Geographical Analysis:** Conduct geographical analysis to identify regional sales patterns and assess shipping delays, optimizing inventory distribution and logistics for enhanced efficiency.
* **Fraud Detection:** Build predictive models to detect fraudulent activities within the supply chain, minimizing financial risks and preventing potential losses.
* **Dynamic Repricing Strategies:** Leverage demand forecasting and competitive pricing analysis to implement adaptive pricing strategies, ensuring optimal revenue generation through market-responsive pricing.

**1.3 Key Assumptions**

This analytics project is founded on several critical assumptions. These assumptions, as tabulated below, will guide our analysis and help identify actionable insights.

| **Key Assumptions** | **Description** |
| --- | --- |
| **Customer** **Relationship Understanding** | Assume that the "**customers**" in this context are resellers, and the "**order**" related information pertains to their end-customers. This distinction is vital for accurately analysing sales dynamics and customer interactions. |
| **Market Demand Stability** | Assume that there is a baseline level of demand for products that can be analysed over time, allowing for reliable forecasting and trend analysis. |
| **Data Quality and Availability** | Assume that the historical sales data is sufficiently accurate and comprehensive to identify trends and inform future predictions. Reliable data is foundational for any analytics effort. |

# Project Design

This project is designed to optimise various facets of the supply chain using advanced analytics. The design focuses on achieving key business outcomes, such as enhancing sales performance, optimising pricing, improving customer retention, detecting fraud, and managing inventory levels. The approach integrates predictive models, systematic data preparation, and an iterative model development process to ensure alignment with business goals.

#### Data Collection and Preparation

The dataset used in this project consists of historical supply chain records, including information on sales, shipping, geographical locations, and order statuses. The data is carefully prepared through cleaning, transformation, and feature engineering to ensure it is ready for model implementation. This preparation process addresses missing values, eliminates inconsistencies, and enhances the data's suitability for predictive modelling.

#### Analytical Techniques and Iterative Improvement

The project design includes several analytical models to address the business objectives:

* **Time-Series Forecasting:** Predicts future demand and manages inventory levels to optimise stock availability.
* **Survival Analysis:** Assesses customer retention by analysing the frequency and duration of interactions, helping identify factors influencing customer loyalty and churn.
* **Geographical Analysis:** Identifies regions with varying demand patterns, high sales potential, or frequent delays, enabling better inventory allocation and targeted marketing.
* **Fraud Detection:** Identifies potentially fraudulent transactions in the supply chain to minimise financial risks.
* **Repricing Strategies:** Utilises demand forecasting and competitive pricing analysis to dynamically adjust prices, maximising revenue by responding to market shifts.

Each model will be evaluated against key performance metrics, with continuous improvement through testing and user feedback to ensure maximum effectiveness.

#### Success Criteria

Success criteria define the measurable goals that determine the effectiveness of each analytical solution in this project. These criteria act as benchmarks for assessing performance, guiding the refinement of models and strategies to ensure they deliver value. By setting clear, measurable targets, success criteria provide a focused approach to each solution, highlighting areas of improvement and ensuring alignment with the project’s overall objectives.

* **Time-Series Forecasting**: xxxx
* **Survival Analysis**: Identify statistically significant factors at confidence intervals of 95% or p-value below 0.05
* **Geographical Analysis**: Achieve a precision rate of X% xxxx
* **Fraud Detection**: Set a recall target of 85% to enhance the detection of fraudulent transactions, safeguarding against unnecessary operational cost and improving supply chain security.
* **Repricing Strategies**: success of the repricing strategy will be measured by achieving a target revenue uplift of 10% through elasticity-based price adjustments.

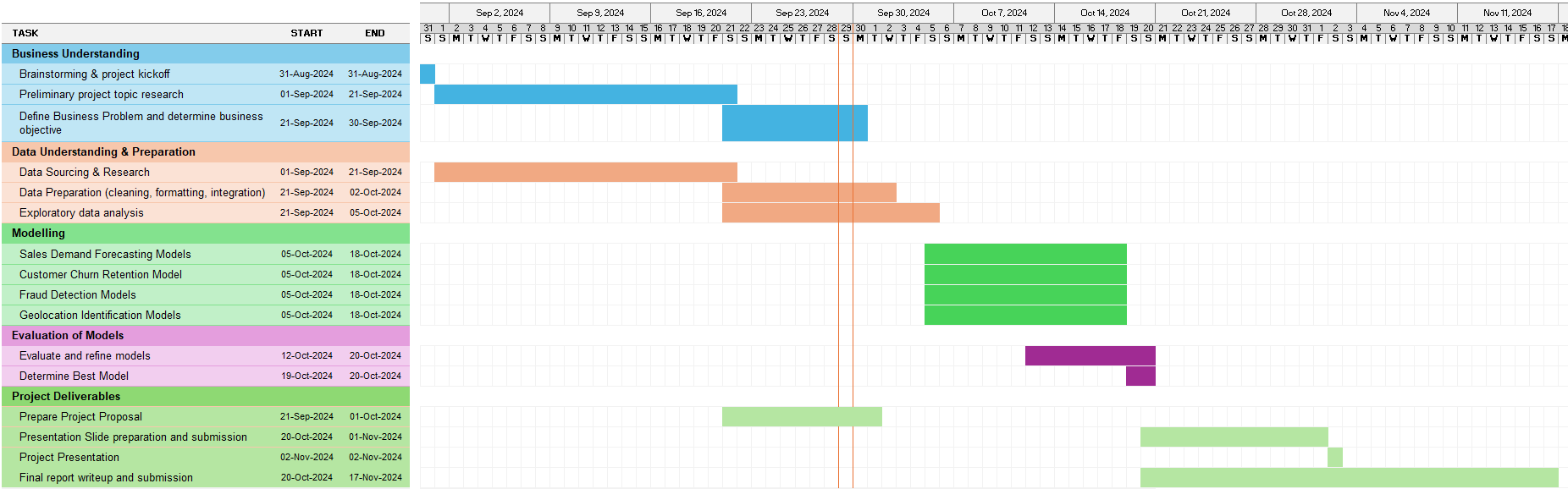
These criteria ensure that each solution contributes to the project’s goals by delivering actionable insights and measurable results.

#### Integration for Business Decision-Making

Upon refinement, the models will be integrated into the business's decision-making framework. This integration will allow the company to leverage the insights generated from the models to optimise pricing, marketing, and operational processes, driving the achievement of the project’s overarching business objectives.

#### Project Efforts and Timeline

The Gantt chart below provides a visual representation of the project timeline, outlining key phases, activities, and milestones. It serves as a roadmap for tracking progress and ensuring timely completion of the project.



# Scope of work

The scope of this project is to apply advanced analytics on the "DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS" dataset to improve various aspects of supply chain management, with a specific focus on enhancing sales performance, optimising pricing strategies, improving customer retention, detecting and mitigating fraud, and optimising facility locations. The project spans data preparation, model development, evaluation, and integration into decision-making processes to support the business’s data-driven objectives.

#### CRISP-DM Methodology

To guide the project, we follow the CRISP-DM framework, which ensures that the analysis is structured, iterative, and aligned with both the business and technical objectives. Below is the CRISP-DM table for this project:

| **CRISP-DM Phase** | **Relevant Actions** |
| --- | --- |
| **Business Understanding** | Defined objectives to increase sales, improve stock availability, optimise pricing (including repricing strategies), and detect fraud in the supply chain. |
| **Data Understanding** | Analysed supply chain data with variables such as 'Sales,' 'Order Status,' 'Product Price,' 'Region,' and competitor data for pricing analysis. |
| **Data Collection & Preparation** | Performed data cleaning, transformation, and feature engineering to make the dataset suitable for modeling. |
| **Modeling** | Implemented time-series forecasting, survival analysis, geographical segmentation, fraud detection models, and dynamic repricing models. |
| **Evaluation** | Validated models using performance metrics, fine-tuning based on business requirements and predictive accuracy. |
| **Deployment** | Integrated model insights into business processes for inventory optimisation, demand forecasting, dynamic pricing, and strategic planning. |

#### 

**5.1 Dataset**

The dataset used in this project is the "DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS" from Kaggle, which contains comprehensive supply chain data from DataCo Global, a company involved in activities such as provisioning, production, sales, and commercial distribution. The dataset comprises three files: “DataCoSupplyChainDataset.csv,” “DescriptionDataCoSupplyChain.csv,” and “tokenized\_access\_logs.csv.” Of these, two CSV files are relevant to our analyses, as detailed in the table below. The data dictionary for the dataset can be found in Annex

| **S/N** | **Dataset Name** | **Data Type** | **Description** |
| --- | --- | --- | --- |
| 1 | DataCoSupplyChainDataset.csv | Small, Non-streaming, Structured | This contains key supply chain information, including customer details, orders, shipping, and products. This dataset serves as the primary basis for the analyses in this project. The descriptions of each data field can be found in Annex. |
| 2 | DescriptionDataCoSupplyChain.csv | Small, Non-streaming, Structured | This contains the descriptions of the fields in the DataCoSupplyChainDataset.csv. |

**Data Selected for Analyses**

[Add table with selected data fields and name corresponding models]

The following table outlines the selected data fields from the dataset, along with the corresponding analyses in which each field is utilised as an input feature or output variable. This overview highlights the relationships between the data fields and various analytical approaches, providing clarity on how each field contributes to understanding sales performance and customer behaviour.

| **S/N** | **Data Field** | **Analyses** |
| --- | --- | --- |
| 1 | Order date (DateOrders) | Sales Trend Analysis, Time-Series Forecasting, Survival Analysis |
| 2 | Sales After Discount | Sales Trend Analysis, Root Cause Analysis, Product Performance Analysis, Time-Series Forecasting, Repricing Strategies |
| 3 | Customer Id | Root Cause Analysis, Survival Analysis |
| 4 | Delivery Status | Delivery Performance Analysis, Root Cause Analysis, Survival Analysis |
| 5 | Late\_delivery\_risk | Delivery Performance Analysis, Root Cause Analysis, Classification Models |
| 6 | Days for shipment (scheduled) | Delivery Performance Analysis, Time-Series Forecasting |
| 7 | Days for shipping (real) | Delivery Performance Analysis, Root Cause Analysis |
| 8 | Category Id | Product Performance Analysis |
| 9 | Category Name | Product Performance Analysis |
| 10 | Order Item Quantity | Product Performance Analysis, Inventory Optimisation |
| 11 | Product Price | Repricing Strategies |
| 12 | Order Item Discount Rate | Repricing Strategies |
| 13 | Shipping date (DateOrders) | Inventory Optimisation |
| 14 | Order Item Id | (For identification, where applicable) |

In Section 7, under modelling, we will explore feature engineering techniques that will enhance the dataset's predictive power. ~~This will include the creation of interaction features among key variables such as delivery performance metrics and sales figures. By strategically engineering these features, we aim to capture complex relationships that can improve model accuracy and provide deeper insights into sales performance and customer behavior.~~

**Data Preparation**

In the data preparation phase, data cleaning, transformation, and feature engineering will be conducted to enhance data quality and ensure readiness for modelling. Key steps in this phase include:

* **Data Cleaning**: Removal of missing or inconsistent values, identification and elimination of outliers, and standardisation of data formats.
* **Feature Engineering**: Development of new features relevant to the objectives, such as demand volatility, days since last purchase, churn event and regional demand metrics.
* **Scaling and Transformation**: Data normalisation and transformation to improve model performance and accuracy.
* **Class Imbalance Handling**: Addressing class imbalances in variables such as fraud detection to improve model reliability.

**5.2 Proposed Models and Techniques**

The following analytical models and techniques will be used to address the project’s business objectives, with evaluation methodologies tailored to each objective:

| **SN** | **Models and Techniques** | **Objectives** | **Evaluation Methodologies** |
| --- | --- | --- | --- |
| 1 | Time-Series Predictive Models | Enhance Sales Performance and Optimise Pricing Strategies | Mean Squared Error (MSE), Mean Absolute Error (MAE) |
| 2 | Survival Analysis | Improve Customer Retention | Kaplan Meier Plot, Log-Rank Test |
| 3 | Anomaly Detection Models | Optimise Fraud Detection | Confusion Matrix, Precision, Recall |
| 4 | Geospatial Analysis Tools | Optimise Facility Location and Improve Logistics Efficiency | Return on Investment (ROI) |

#### Evaluation and Improvement

Each model will undergo rigorous testing against relevant performance metrics, with iterative refinement based on evaluation results. Feedback loops will be incorporated to enhance model accuracy and adaptability in response to changing business needs.

#### Integration for Business Decision-Making

The final outputs of this project, including time-series forecasts, survival analysis results, geographical insights, fraud detection models, and repricing strategies, will be integrated into business processes. This integration will empower DataCo to make data-driven decisions, optimise operations, and achieve long-term business sustainability.

The Scope of Work is designed to comprehensively address each of the project objectives, ensuring that each phase is executed with precision and aligned with DataCo’s strategic goals.

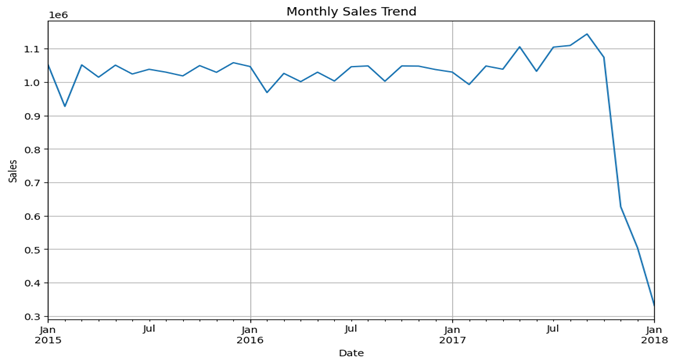
1. **Data Understanding**

6.2.1 **Monthly and Weekly Sales Trends Analysis**

The monthly and weekly sales trends provide insights into the sales performance from January 2015 to January 2018. This analysis reveals patterns of stability, seasonal peaks, and a notable sales decline in late 2017. These trends offer valuable information for refining revenue strategies, adjusting operational plans, and preparing for high-demand periods.

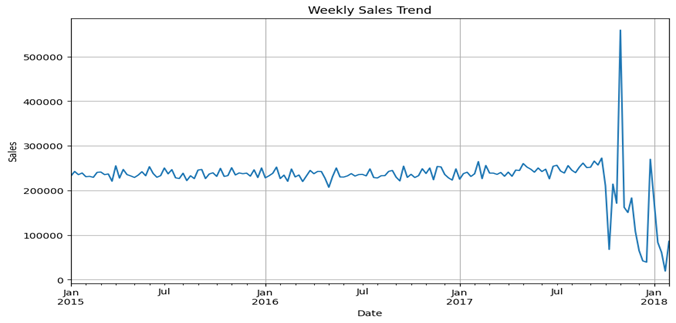
#### Monthly Sales Trend

Monthly sales remained relatively stable from 2015 through mid-2017, with only minor fluctuations. However, a sharp decline in late 2017 signals potential underlying issues that may have impacted revenue generation. This decline could be attributed to changes in consumer demand, market conditions, or internal operational challenges. Identifying the causes behind this decrease is crucial to developing strategies that mitigate similar risks in the future, thereby ensuring a more consistent revenue stream.



#### Weekly Sales Trend

The weekly sales trend shows a stable pattern with prominent seasonal peaks towards the end of each year, particularly in late 2017. These spikes are likely influenced by holiday shopping periods and promotional activities, which traditionally drive higher sales volumes. Recognizing these high-demand periods provides opportunities to optimize inventory, enhance marketing efforts, and ensure that stock levels align with increased customer demand. By strategically planning around these seasonal spikes, the business can maximize sales and improve customer satisfaction.

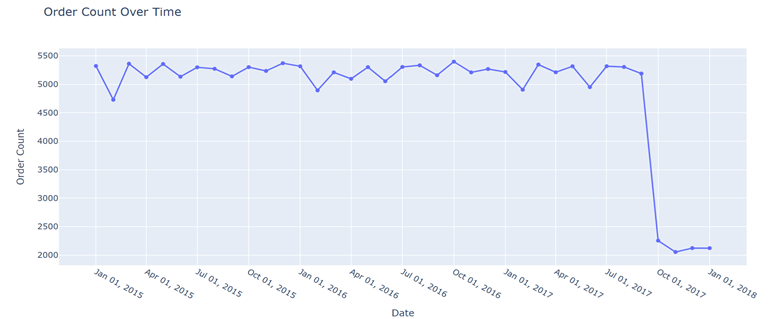


6.2.2 **Order Count and Status Analysis**

An examination of order counts and order statuses over time offers insights into the company's order fulfillment efficiency and operational health. Trends in order statuses highlight both strengths in fulfillment processes and potential areas for improvement.

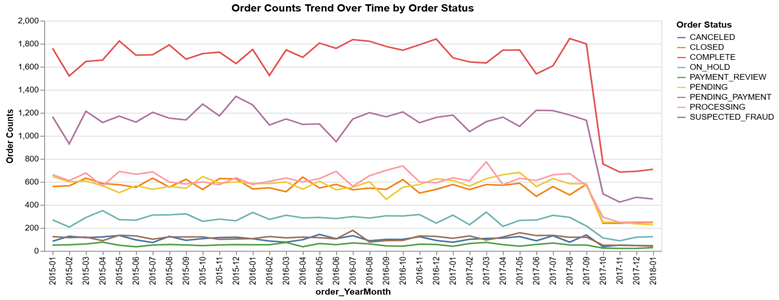
#### Order Count Over Time

Order counts remained stable for most of the analysis period, closely mirroring the monthly sales trend. However, a significant decline in late 2017 aligns with the observed drop in monthly sales, suggesting that the reduction in order volume contributed to the overall sales downturn. This trend may be the result of external market influences or internal operational challenges that affected the company's ability to process and fulfill orders.



#### Order Counts by Status Over Time

An analysis of order statuses over time reveals that 'Complete' orders were dominant until late 2017, when a marked decrease in completion rates was observed. Concurrently, there was an increase in orders categorized as 'Canceled,' 'On Hold,' and other non-complete statuses. This shift indicates potential bottlenecks or inefficiencies in the order processing and fulfillment workflow. By identifying and addressing these operational issues, the company can enhance order completion rates, reduce cancellations, and improve overall customer satisfaction.

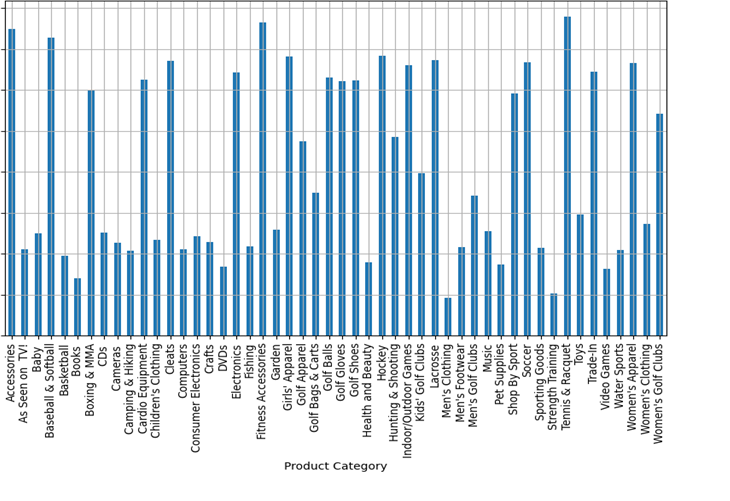


6.2.3 **Demand by Category, Customer State, and Region**

The distribution of demand across product categories, states, countries, and regions provides essential insights for optimizing inventory management, tailoring marketing strategies, and ensuring efficient resource allocation to meet customer demand in key areas.

#### Top Product Categories by Demand

Analysis of product categories reveals that 'Cleats,' 'Men’s Footwear,' and 'Women’s Apparel' are among the most in-demand items. These high-demand categories should be prioritized in inventory planning to ensure that stock levels can meet customer needs without delays. By focusing on these popular categories, the business can prevent stockouts, maintain customer satisfaction, and capitalize on revenue opportunities.



#### Top Customer States by Demand

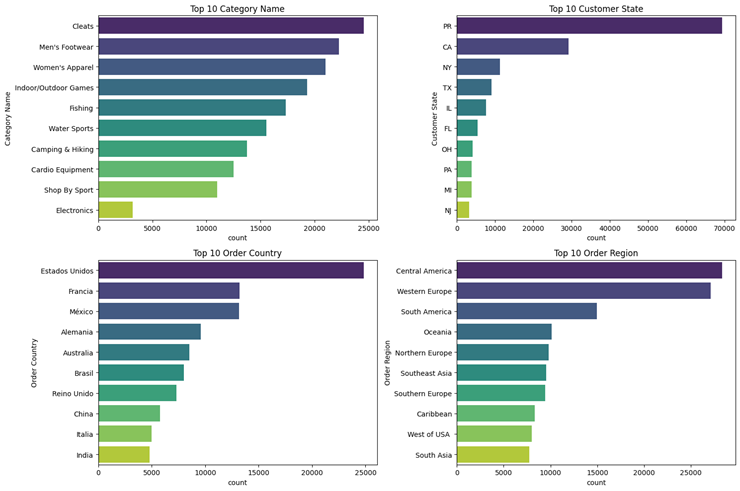
Demand by state shows that 'PR' (Puerto Rico) and 'CA' (California) lead in order volume. Knowing which states have the highest demand allows for the development of localized marketing and inventory strategies, ensuring that resources are allocated efficiently to meet regional demand.

#### Top Order Countries by Demand

Internationally, 'Estados Unidos' (United States) and 'Francia' (France) show the highest order counts. This information supports international logistics planning, ensuring that supply chain resources align with demand levels across key countries.

#### Top Order Regions by Demand

At the regional level, 'Central America' and 'Western Europe' are among the highest-demand areas. Recognizing demand at a regional level allows the business to optimize logistics, reduce delivery times, and improve service quality in geographically diverse markets.

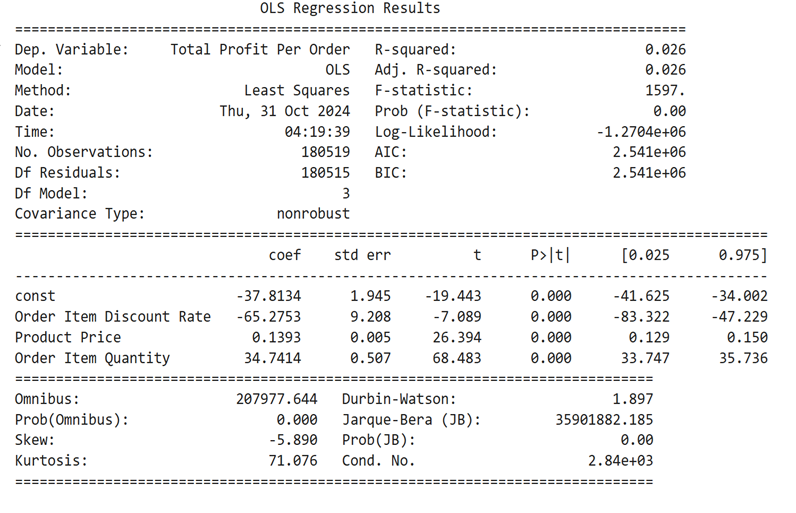


6.2.4 **Profitability Analysis Using OLS Regression**

To better understand the factors affecting profitability, an Ordinary Least Squares (OLS) regression analysis was conducted using variables such as discount rate, product price, and order quantity. The results offer actionable insights for refining pricing and discounting strategies.

#### OLS Regression Results

The regression analysis indicates that *Order Item Discount Rate* negatively impacts profit, suggesting that higher discounts reduce overall profitability. Conversely, *Product Price* and *Order Item Quantity* have positive impacts on profit, though product price shows a relatively smaller effect. The low R-squared value suggests that additional variables may need to be included in the model to provide a more comprehensive view of the factors influencing profitability.



1. **Modelling and Evaluation**

**7.1 Time-Series Forecasting**

**7.1.1 Objectives**

The objective of the time-series forecasting component is to develop predictive models that can accurately forecast sales and demand patterns within DataCo’s supply chain. By gaining insights into future sales trends, DataCo aims to make inventory management decisions, ultimately enhancing revenue generation and customer satisfaction. The forecasted demand data will enable **optimised inventory management**.

**7.1.2 Key Assumptions**

Taking Puerto Rico as an example, our key assumption is that if a warehouse is located in the middle of Puerto Rico, efficient inventory management will enable stocks to reach resellers’ locations within one day. According to geographical information from Britannica and other reputable sources, Puerto Rico measures approximately 179 km (111 miles) in length and 63 km (39 miles) in width, covering a total area of about 9,104 square kilometers (3,515 square miles).

With a warehouse situated near Caguas, most locations on the island can typically be reached within half a day for supply chain logistics; however, some remote areas may require more time for delivery. For instance, Cabo Rojo is about 185 km (115 miles) from San Juan, with a driving time of approximately 2 hours and 11 minutes under normal traffic conditions. In contrast, Fajardo, located on the northeastern coast, is roughly 45 minutes to 1 hour from San Juan.

By anticipating high-demand periods, DataCo can ensure adequate stock levels to facilitate deliveries within one day and avoid stockouts, especially in volatile regions or for high-demand products.

### 7.1.3 Methodology

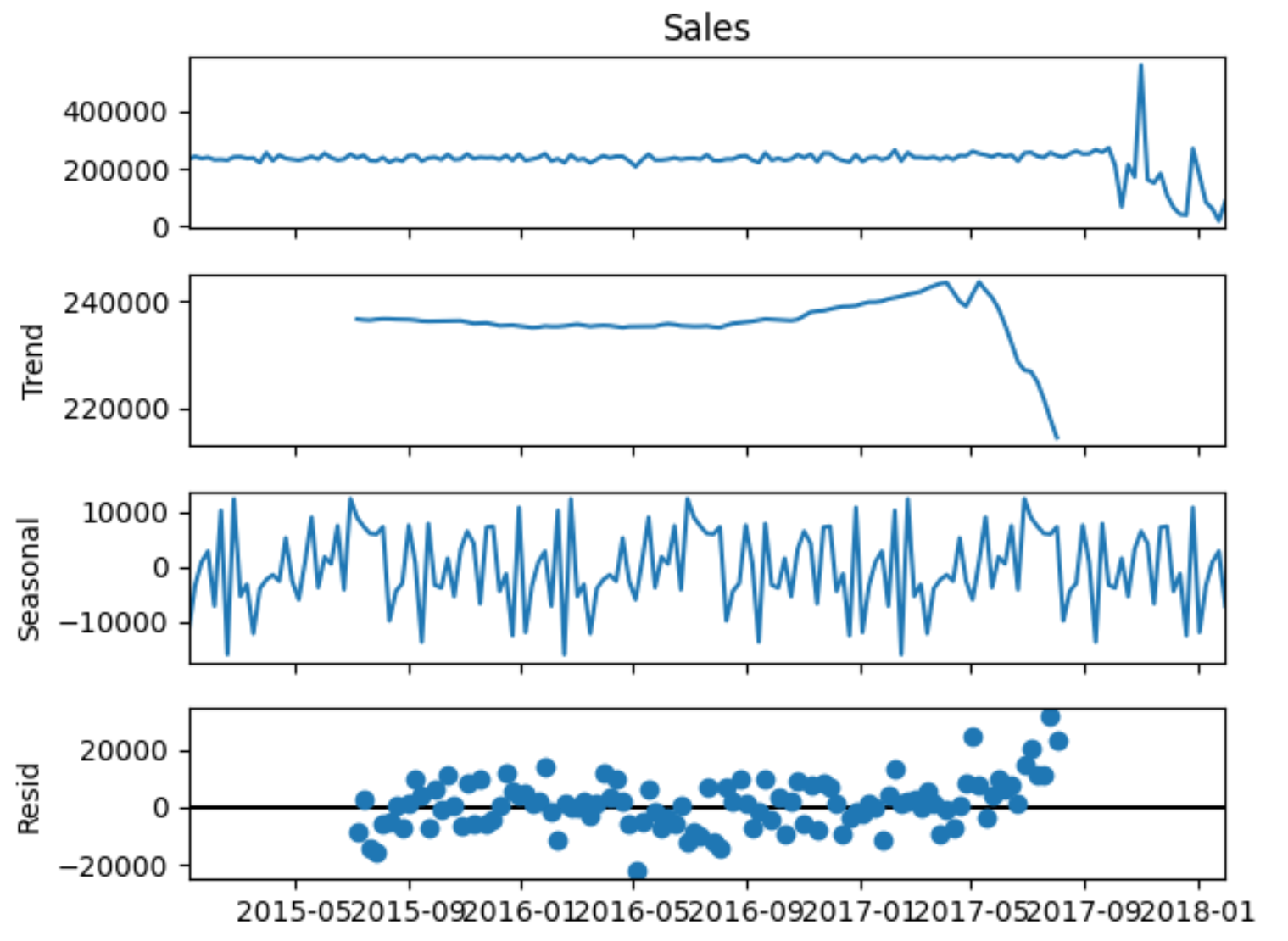
This section outlines the structured approach we followed to conduct our analysis and develop our forecasting models.

Our analysis followed a structured approach consisting of several key steps:

1. **Identification of Stable and Volatile Periods**

We analysed historical sales data to detect periods of stability and volatility.

Given the steep decline in monthly sales and the high volatility observed in the weekly sales trends during our exploratory data analysis, we will now decompose the weekly sales time series. This decomposition will provide a clearer understanding of its various components.



In addition to confirming the trends identified earlier, the plots reveal potential issues with the residuals, which could adversely affect the accuracy of our forecasting models. Furthermore, the plots do not provide clear evidence of seasonality.

Augmented Dickey-Fuller (ADF) test was run on the sales data separately, and a p-value of 0.073491 was obtained, suggesting that we may not have strong evidence against the null hypothesis of a unit root, indicating that while the series is not strongly stationary, there is some indication of stationarity after differencing (as suggested by the order of differencing used).

Next, we used statistical metrics like standard deviation to classify these periods.

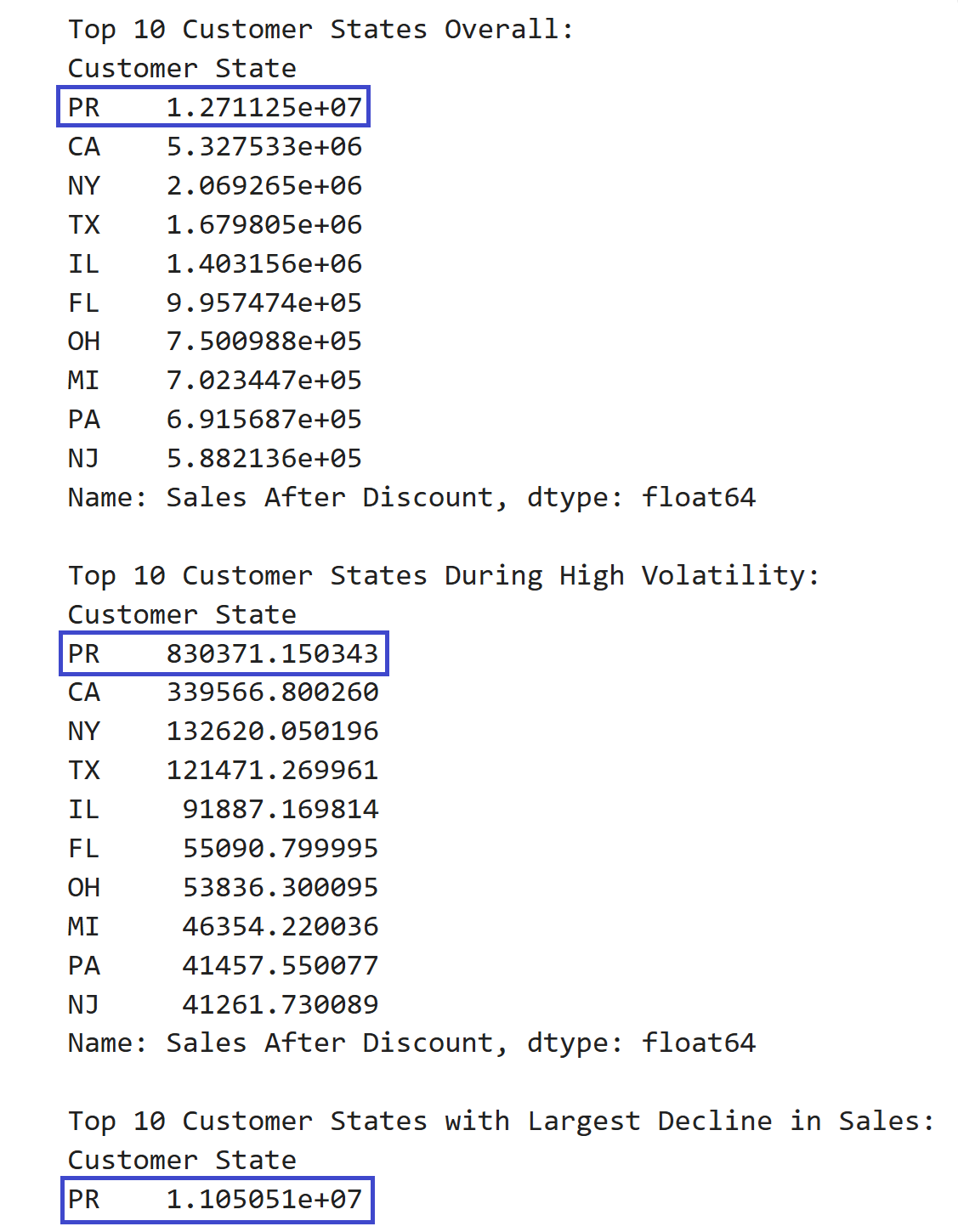


The variance is plotted to illustrate the identified cut-off date determined by this method, which is found to be the week of October 11, 2017.



1. State Selection

Next, we identified a state with the highest sales during stable periods and the largest declines afterward.

****

This analysis led us to select the PR state for this study.

1. **Technique Comparison**

We compared neural network models with non-neural network models, utilizing PyCaret, which includes the Auto ARIMA functionality, to identify the best-performing techniques.

To provide clearer insights into the output of the SARIMAX model, we modeled Auto ARIMA separately. Consequently, this section is organized into three subsections, all of which were performed on monthly data for the top-selling product, “Perfect Fitness Perfect Rip Deck”.

The performance of these three methodologies will be evaluated based on their forecasting accuracy, measured in Root Mean Squared Error (RMSE), for the top-selling product.

Following this analysis, one of these techniques will be selected as the primary model to forecast sales across all 118 products in our inventory. This systematic approach ensures that we adopt the most effective forecasting strategy tailored to our specific sales data characteristics.

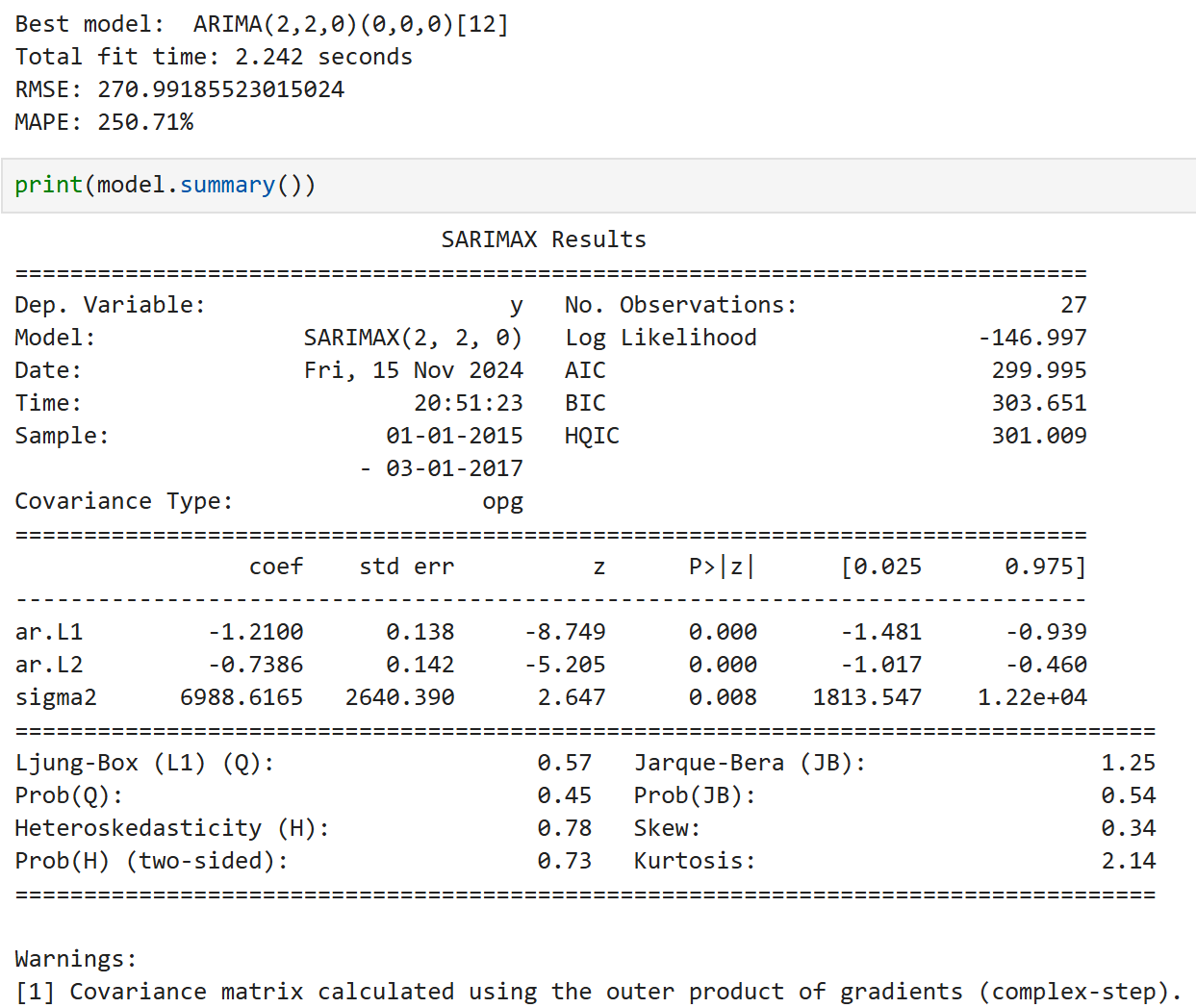
**(i) SARIMAX**

We begin with the implementation of the Seasonal Autoregressive Integrated Moving Average (SARIMAX) analysis conducted on the monthly sales data from January 2015 to January 2018. This statistical modeling approach aims to uncover underlying patterns and relationships within the sales data, enabling us to gain insights into past performance and inform future predictions. By examining various model parameters, fit statistics, and diagnostic tests, we evaluate the effectiveness of the SARIMAX model in capturing the dynamics of the sales data and assess its predictive capabilities.

The figure below illustrates the configuration used to execute the Auto ARIMA analysis on the dataset with the PMDARIMA package.



The following analysis details the key findings and implications derived from this modeling exercise.



The output of the SARIMAX model indicates that a Seasonal Autoregressive Integrated Moving Average (SARIMA) approach was utilized to analyze your monthly sales data from January 2015 to January 2018. The selected model is SARIMAX(2, 2, 0), which implies that there are two autoregressive terms, two levels of differencing to achieve stationarity, and no moving average terms.

Key aspects of the output are as follows:

1. **Target/Dependent Variable**: The model targets monthly sales totals (y) from January 2015 to January 2018, with a total of 37 observations available for analysis.
2. **Model Fit and Performance Metrics**:
   * The Log Likelihood value is -146.997, which helps in assessing the fit of the model; higher values suggest a better fit.
   * The Akaike Information Criterion (AIC) is 299.995, the Bayesian Information Criterion (BIC) is 303.651, and the Hannan-Quinn Information Criterion (HQIC) is 301.009; lower values indicating a better model fit.
3. **Coefficient Estimates**:
   * The coefficients for the autoregressive terms (ar.L1 and ar.L2) are -1.2100 and -0.7386, respectively. Both coefficients are statistically significant with p-values of 0.000, indicating a strong negative correlation of the current value with its past values. The confidence intervals for these coefficients range from approximately -1.481 to -0.939 for ar.L1 and -1.017 to -0.460 for ar.L2.
   * The sigma² coefficient, which represents the variance of the residual errors, is estimated at 6988.6165, with a p-value of 0.008 indicating statistical significance.
4. **Statistical Tests**:
   * The Ljung-Box test results indicate no significant autocorrelation in the residuals (Prob(Q) = 0.45), which is a good sign since it suggests that the model has captured the underlying structure of the data well.
5. **Error Metrics**:
   * The RMSE is approximately 270.99, which provides a measure of how well the model's predictions match the actual data, with lower values indicating better predictive power.
   * The Mean Absolute Percentage Error (MAPE) is extremely high at 250.71%, indicating that on average, the model’s predictions could differ by a significant percentage from actual values.

In summary, the SARIMAX model has identified strong autoregressive characteristics in the sales data, and while it fits the data relatively well as indicated by the statistical tests, the very high MAPE suggests that there may still be considerable room for improvement in terms of predictive accuracy. Further steps could include refining the model by exploring additional seasonal components or external factors.

**(ii) Comparison between Neural-Network and Non-Neural Network Techniques**

In the quest for accurate demand forecasting for Smart DataCo's product names,, we conducted a comprehensive comparison of various modelling approaches, including Long-Short Term Memory (LSTM) networks, non-neural network models using PyCaret (including Auto ARIMA).

Figure XYZ. Code Snippets for LSTM



Figure XYZ. Results for LSTM

Figure XYZ. Code Snippets for PyCaret Time Series Experiments



Figure . Results for PyCaret Time Series Experiments

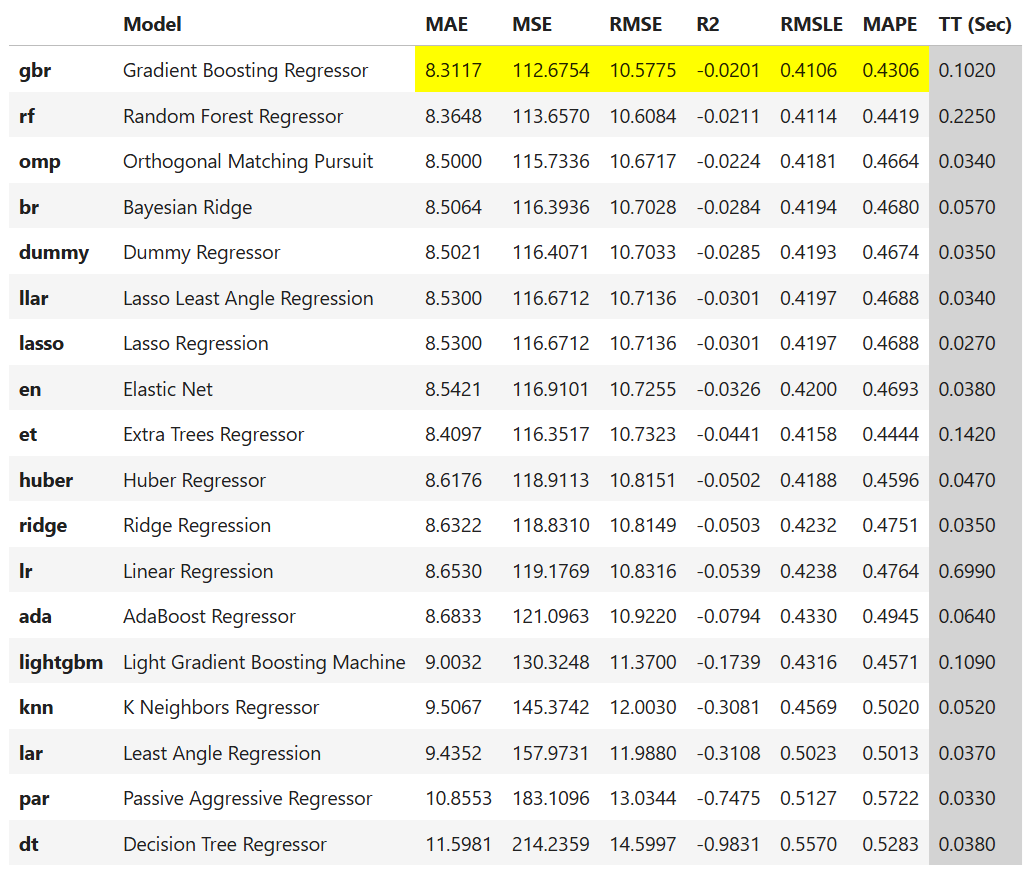


Table XYZ. Performance of Neural Network and Non-Neural Network Models

| **Approach** | **Best RMSE** |
| --- | --- |
| **LSTM** | 16.0575 |
| **Non-Neural Network Modelling using PyCaret Time Series Experiments**  (Best Model: Gradient Boosting Regressor) | **10.5775** |
| **Auto ARIMA using PMDARIMA**  (Best Model: SARIMAX) | 270.99 |

Among the three approaches, the Croston model implemented using PyCaret achieved the lowest MAPE of **10.7721**, indicating superior performance.

In this view, our team selected the PyCaret approach for menu item-level modelling.

1. Modeling with Best Techniques: We utilized PyCaret's AutoML capabilities to forecast time series across 118 products efficiently. This tool allowed for rapid experimentation and improved model performance.
2. Feature Selection: We started with all available features and then removed irrelevant features that did not contribute to state-level data. We also eliminated highly correlated features. We attempted LASSO regression, but it did not lead to the removal of any additional features.
3. Feature Engineering: We created rolling statistics and temporal features to capture trends and seasonality that may affect demand.
4. Volatility Analysis: We identified that sales from January 2015 to October 2017 were relatively stable. However, we observed high volatility from October 2017 to January 2018. This significant change prompted us to concentrate on the PR state due to its impact on sales.
5. Data Preparation: The final features we retained for modeling include days for shipping, order item discount rate, product price, product name, and order date. These variables are crucial for understanding demand patterns.
6. **Model Selection**:

**4.2**

**4.3**

1. **[Add cross validation] Model Training and Testing**: The dataset was split into training (January 2015 to September 2017) and testing (October 2017 to January 2018) sets. This split allowed for testing model performance during both stable and volatile periods.

[Add details and screenshots]

[Add tables]

[Add figures]

[Add code snippets]

1. **Hyperparameter Tuning**: Grid search and automated tuning were applied to optimise model parameters for ARIMA, SARIMA, and Prophet, improving forecast accuracy by minimising errors on the test set.

[Add details and screenshots]

[Add tables]

[Add figures]

[Add code snippets]

### 

### 7.1.3 Model Evaluation

Model evaluation was performed using standard performance metrics to assess forecast accuracy and identify the best-fitting model for sales prediction:

* **Mean Absolute Percentage Error (MAPE)**: MAPE enabled an assessment of forecast accuracy as a percentage, allowing for easy comparison across periods and different demand levels.
* **Root Mean Squared Error (RMSE)**: RMSE provided a comprehensive metric that captured both the magnitude and direction of errors, further validating the robustness of the selected model.

[Add details and screenshots]

[Add tables]

[Add figures]

[Add code snippets]

Based on these metrics, [...] and [...] were selected as the top-performing models, as they offered the most consistent accuracy, particularly in the volatile periods leading up to January 2018.

### 7.1.4 Results

The time-series forecasting models provided several actionable insights:

* **Projected Demand Patterns**: The final model forecasted a continued period of stabilised sales post-January 2018, with intermittent peaks that could guide seasonal inventory adjustments.
* **Inventory Recommendations**: Predicted high-demand periods informed inventory management strategies, suggesting stockpiling before peak sales periods to mitigate stockouts and optimise warehousing costs.
* **Dynamic Pricing Opportunities**: By identifying future demand fluctuations, the model supports repricing strategies to adjust prices during anticipated high-demand periods, maximising sales revenue and competitiveness.
* **Trend Analysis**: Analysis of the steep decline in sales from October 2017 until stabilisation in January 2018 highlighted areas for further investigation, such as supply chain constraints or external factors that may have contributed to the volatility. This period-specific insight underscores the importance of continual monitoring to adapt strategies proactively.

[Add details and screenshots]

[Add tables]

[Add figures]

[Add code snippets]

The time-series forecasting component has established a foundational tool for DataCo to enhance decision-making in inventory, pricing, and operational efficiency, ensuring alignment with the company’s business objectives for sustained growth and competitiveness.

**7.2 Survival Analysis [Clerys]**

The objectives of the survival analysis in this supply chain study are to enhance customer retention and satisfaction by identifying and understanding the factors impacting customer churn and late delivery risks.

Through these analyses, the project intends to provide actionable insights that will enable the formulation of effective strategies for improving overall service delivery and customer engagement. Enhanced understanding of the underlying causes of customer churn and delivery delays will not only improve operational efficiencies but also bolster customer loyalty and trust in the brand.

**7.2.1 Customer Churn Analysis**

7.2.1.1 Objective

The objective for customer churn analysis is to Identify the statistically significant factors that contribute to customer attrition. This understanding will facilitate the development of targeted strategies aimed at mitigating churn, thereby improving customer retention rates. The analysis seeks to uncover patterns and predictors of churn, which could range from demographic characteristics to interaction histories and service usage patterns.

7.2.1.2 Methodology

7.2.1.2.1 Data Preparation

The supply chain dataset does not come with features describing customer status, but it contains the purchase history of each customer including the order date, order quantity and products purchased. Feature engineering done for this customer churn analysis.

| New Columns | Description |
| --- | --- |
| Last Purchase date | Retrieve the most recent order date for each customer ID |
| Days since last purchase | Calculated the differences in number of days between Last Purchase date and the last order date in the dataset which is 31 Jan 2018 |
| Churn Event | Binary variable, with 1 = Churn, 0 = Not Churn  Assumption: Customers with Days since last purchase greater than 365 days are considered as Churned |

**7.2.1.2.2 Implementation in JMP**

Survival Analysis are conducted using JMP by fitting the Kaplan Meier survival plot and Proportional Hazard Model. The variables identified for the analysis:

**Time to event**: Days since last purchase

**Event**: Churn Event (1 = Churn, 0 = Not churn)

**Censor code**: 0

**Variables**: Customer segment, Market, Order Region, Delivery Status, Days for shipping (real)



Figure x.x Fitting Kaplan Meier Analysis on Customer Churn by Customer Segment

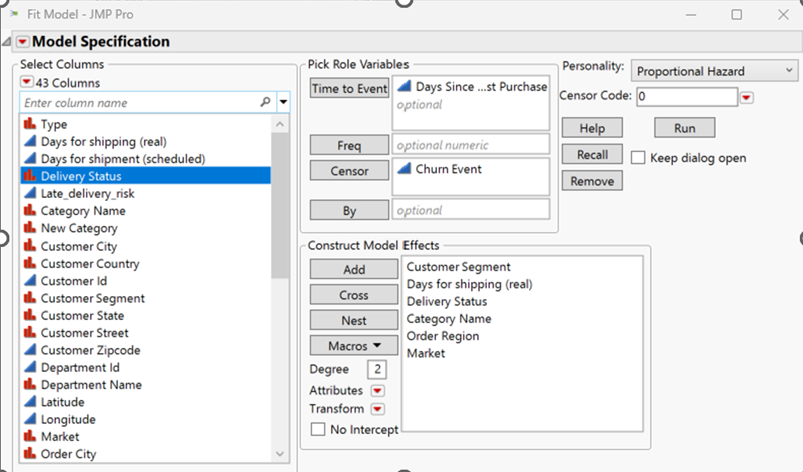


Figure x.x Fitting Proportional Hazard Model on Customer Churn

**7.2.1.3 Model Evaluation**

Statistically significant factors are identified based on the p-values derived from the Log-Rank Test in the Kaplan-Meier analysis and the Effect Summary and Whole Model sections of the Proportional Hazards Model. P-values less than 0.05 indicate statistical significance.

**Kaplan Meier:** The statistical significance of differences between survival curves can be assessed using the Log-Rank Test, where a p-value less than 0.05 typically indicates significant differences.

**Proportional Hazard Model:** Statistical significance of the covariates is evaluated through the Effect Summary and Whole Model sections, where p-values are reported. Specifically, the p-values in the Effect Summary indicate the significance of individual covariates, while the Whole Model p-value (from the likelihood ratio test comparing the model with and without the covariates) assesses the overall fit and utility of the model. P-values less than 0.05 are considered significant, confirming that the included factors meaningfully influence the time to the event in question.

7.2.1.4 Result

**7.2.1.4.1 Kaplan Meier**

Table below show the summary of the Kaplan Meier result for each variable. Based on the p-value under Log-Rank Test, only days for shipping (real) is not significant.

|  | Kaplan Meier | |  |
| --- | --- | --- | --- |
| Variables | Log-Rank | Wilcoxon | Result |
| Customer Segment | <.0001 | <.0001 | Significant |
| Market | <.0001 | <.0001 | Significant |
| Order Region | <.0001 | <.0001 | Significant |
| Delivery Status | 0.0304 | 0.0564 | Significant |
| Days for shipping (real) | 0.6059 | 0.1236 | Not significant |

The corresponding result with the survival plot displayed in JMP are shown in the Figure X.X. and Figure X.X. However, the survival plots do not reveal a clear pattern indicating which level significantly impacts the likelihood of customer churn.

|  |
| --- |

Figure X.X Kaplan Meier result in JMP for Customer Segment, Delivery Status and Market

|  |
| --- |

Figure X.X Kaplan Meier result in JMP for Days for shipping (real) and Order Region

**7.2.1.4.2 Proportional Hazard Model**

When fitting all the variables mentioned earlier into Proportional Hazard Model, Days for shipping (real) remains insignificant while the rest of the variables are significant. Another model excluding Days for shipping (real) from analysis shows similar results. Market, Delivery Status, Customer Segment and Order Region are statistically significant with p-value less than 0.05 based on the Effect Wald Test.

|  |
| --- |

Figure X.X Proportional Hazard Model result for multiple variables

To further investigate each variable and gain insight into which levels increase or decrease the likelihood of customer churn, proportional hazard models are run separately for each variable. This approach allows for a detailed assessment of the impact of individual variables on churn. The results under parameter estimates indicate the following:

**Market**: Africa has higher likelihood to churn, while Latin America has lowest likelihood to churn.

**Order Region**:

· Top 3 with high likelihood to churn: East Africa, Central Africa, West Africa.

· Bottom 3 with low likelihood to churn: Central America, South America, Caribbean.

**Customer Segment**: Corporate customers are more likely to churn while consumer are less likely to churn.

**Delivery Status**: Late delivery increases the likelihood of churn, while advance shipping reduces the likelihood to churn.

|  |
| --- |

Figure X.X Proportional Hazard Model result for Delivery Status and Customer Segment respectively

|  |
| --- |

Figure X.X Proportional Hazard Model result for Market, Order Region and Days for shipping (real) respectively

**7.2.1.4.3 Overall Insight**

The finding from both Kaplan Meier and Proportional Hazard Model shows consistent results.

| Variables | Kaplan Meier | | Proportional Hazard | |
| --- | --- | --- | --- | --- |
| Log-Rank | Wilcoxon | Effect Wald Test | Whole Model |
| Customer Segment | <.0001 | <.0001 | <.0001 | <.0001 |
| Market | <.0001 | <.0001 | <.0001 |
| Order Region | <.0001 | <.0001 | <.0001 |
| Delivery Status | 0.0304 | 0.0564 | 0.0313 |
| Days for shipping (real) | 0.6059 | 0.1236 | 0.2471 |

The findings regarding customer segment, market, and region suggest that efforts should be directed towards targeting Corporate customers and the African market when developing marketing strategies for customer retention.

On the other hand, Days for shipping (real) is not a statistically significant factor impacting customer churn; however, the delivery status is found to be statistically significant. Delivery status is determined by comparing the "Days for shipping (scheduled)" with the "Days for shipping (real)," where a shipment is considered late if the actual shipping duration exceeds the expected timeframe. This implies that managing customer expectations regarding delivery performance is crucial to improve customer retention. Delivery performance should be taken into account when working on customer retention.

**7.2.2. Late delivery risk analysis**

7.2.2.1 Objectives

Investigate the determinants of late deliveries within the supply chain to pinpoint factors that contribute to late delivery. By identifying these critical factors, the study will guide improvements on delivery performance, ultimately enhancing customer satisfaction through more reliable and timely deliveries.

7.2.2.2 Methodology

7.2.2.2.1 Data preparation

The supply chain dataset consists of late delivery risk which is a binary variable related to the delivery status. There are 4 values under delivery status, which include shipping on time, shipping cancelled, advance shipping and late delivery. The late delivery risk is 1 when the delivery status is late delivery, while 0 could be shipping on time, shipping cancelled or advance shipping. Data with shipping cancelled status are excluded from late delivery risk analysis since it is not relevant to the event and could be bias to the analysis.

7.2.2.2.2 Implementation

Similarly, Survival Analysis are conducted using JMP by fitting the Kaplan Meier survival plot and Proportional Hazard Model using the variables identified below:

**Time to event**: Days for shipping (real)

**Event**: Late Delivery risk (1= late delivery, 0 = advance shipping, on time delivery)

**Censor code**: 0

**Variables**: Shipping Mode, Market, Order Region, Order City, Department Name, Category Name

In this dataset, all the delivery are completed with confirmed delivery status, be it late or not late. There is no on-going delivery with unknown delivery status beyond the observation period of this dataset and hence literally no censoring for this late delivery risk analysis. However, JMP does not has dedicated “event” roles when feeding in parameter for analysis. The event, late delivery risk is input under “Censor” roles in JMP with censor code 0 for the software to learn that late delivery (1) is considered failed event.

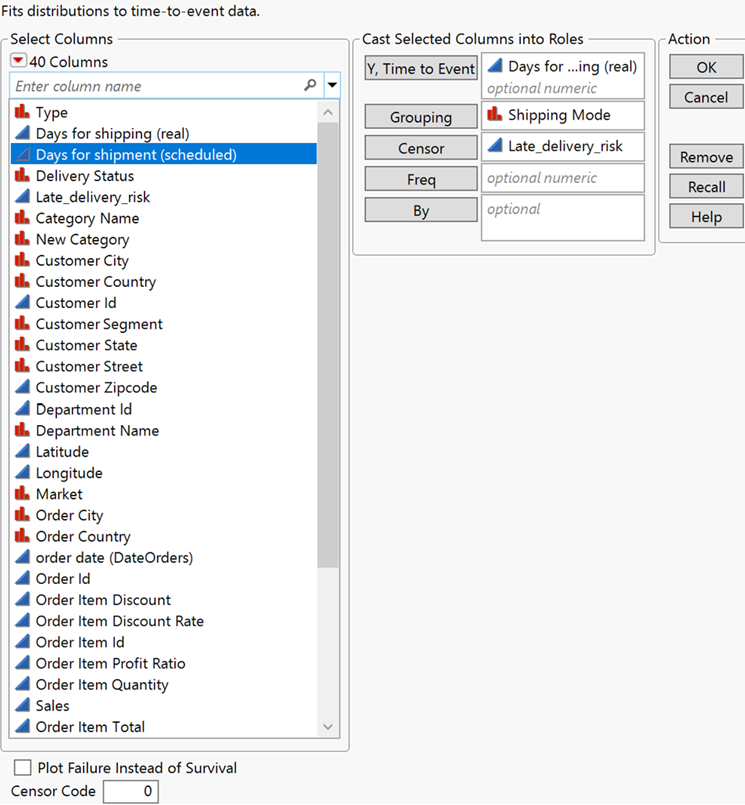


Figure X.X Fitting Kaplan Meier Analysis on Late Delivery risk by Shipping mode

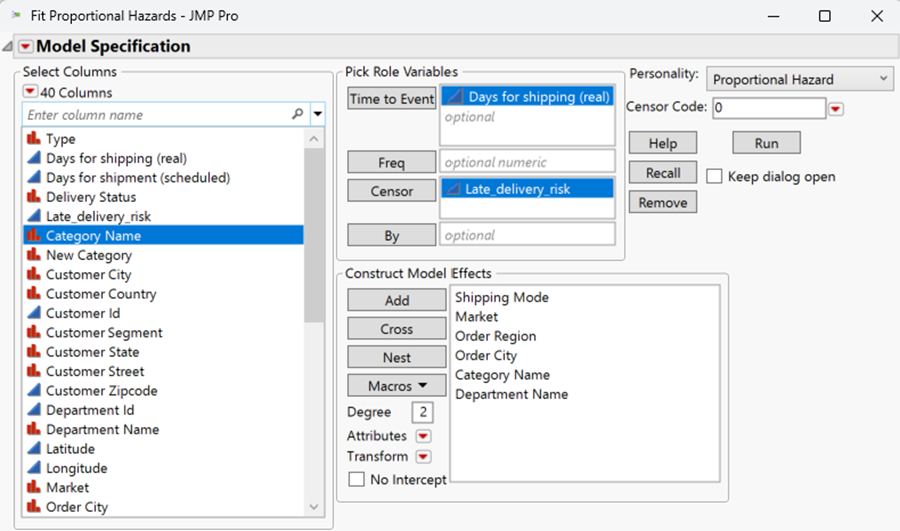


Figure x,x Fitting Proportional Hazard Model on Late Delivery risk

7.2.3 Model Evaluation

Similar to the 7.2.1.3

7.2.2.4 Results

7.2.2.4.1 **Kaplan Meier**

For late delivery risk analysis, the Kaplan Meier result shows that Shipping Mode, Market, Order Region, Department Name, and Category Name are all significant with p-value less than 0.05 for both Log-Rank Test and Wilcoxon.

|  | Kaplan Meier | |  |
| --- | --- | --- | --- |
| Variables | Log-Rank | Wilcoxon | Result |
| Shipping Mode | <.0001 | <.0001 | Significant |
| Market | 0.0297 | 0.0055 | Significant |
| Order Region | <.0001 | <.0001 | Significant |
| Department Name | 0.0178 | 0.0016 | Significant |
| Category Name | 0.0006 | <.0001 | Significant |

Looking at the survival plot shown in Figure X.X , Figure X.X and Figure X.X, generally it is not obvious to determine which level increase or decrease the late delivery risk, except for Shipping Mode. The result for Shipping Mode shows that First Class is always getting late delivery.

|  |
| --- |

Figure X.X Kaplan Meier result in JMP for Shipping Mode and Market

|  |
| --- |

Figure X.X Kaplan Meier result in JMP for Order Region and Department Name

|  |
| --- |

Figure X.X Kaplan Meier result in JMP for Category Name

**7.2.2.4.2 Proportional Hazard Model**

To further investigate the factors impacting late delivery risk, several Proportional Hazard Models are run using different combinations of variables. The initial model, which includes all five variables, reveals that certain variables, such as Department Name, Market, and Order Region, are not significant according to the Effect Wald Test. These insignificant variables are subsequently removed, and the Proportional Hazard Model is rerun. This process is repeated with different combinations of variables, ultimately finding that only Shipping Mode is statistically significant.

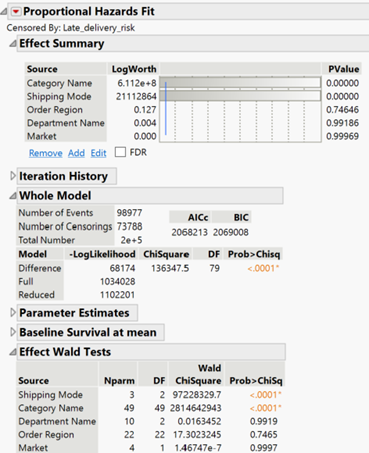


Figure X.X Proportional Hazard Model result for multiple variables

|  |
| --- |

Figure X.X Proportional Hazard Model result for Shipping Mode and Category Name respectively

**7.2.2.4.3 Overall Insight**

Kaplan Meier Graph suggested that Shipping Mode, Market, Order Region, Department Name, Category Name are significantly impacting the late delivery risk. However, Proportional Hazards model suggests that only shipping mode is significant.

| Variables | Kaplan Meier | | Proportional Hazard Model | |
| --- | --- | --- | --- | --- |
| Log-Rank | Wilcoxon | Effect Wald Test | Whole Model |
| Shipping Mode | <.0001 | <.0001 | <.0001 | <.0001 |
| Market | 0.0297 | 0.0055 | 0.2217 |
| Order Region | <.0001 | <.0001 | 0.6012 |
| Department Name | 0.0178 | 0.0016 | 0.0313 |
| Category Name | 0.0006 | <.0001 | 0.5589 |

Each shipping mode is assigned different delivery expectations. For "First Class" shipping mode, deliveries are expected within one day; however, they consistently take two days, resulting in delays. To ensure satisfactory on-time delivery service, it is advisable to either improve the delivery performance for First Class shipping or adjust expectations to more realistic timelines.

**7.3 Geographical Analysis [Priya]**

**a-** **Geographical Analysis using LSTM-** Capture and predict sales trends for each region

**7.3.1 a Objectives**

The Long Short-Term Memory (LSTM) model was employed to forecast demand across regions on a weekly basis, leveraging historical sales data to capture long-term demand patterns and short-term fluctuations. This model enables more accurate demand planning, allowing the company to anticipate customer needs and optimize inventory distribution.

**7.3.2 a Methodology**

In implementing the LSTM model for geographical demand forecasting, we followed a structured approach that included data normalization, sequence generation, and feature selection to capture key temporal and regional patterns in sales data. The model architecture was designed to leverage two LSTM layers with dropout and batch normalization, optimized with Mean Squared Error (MSE) and the Adam optimizer, ensuring the model could generalize effectively. Training was conducted with an 80/20 train-test split to assess the model's predictive accuracy. This approach enabled the model to effectively capture regional demand trends, supporting more accurate inventory planning and resource allocation.

**7.3.3 a Model Evaluation**

#### 1. Exploratory Data Analysis (EDA)

**Key Observations from EDA**:

* **Demand Distribution**: Geographic demand was uneven, with high sales concentrations in certain states, cities, and regions, highlighting the need for regional segmentation in the forecasting model.
* **Temporal Patterns**: Seasonal fluctuations and recurring peaks in sales data suggest that demand forecasting should account for seasonality.
* **Shipping Delays**: Variability in shipping times was observed, potentially impacting customer satisfaction and sales. This indicates the need for efficient logistics planning.

#### 2. Feature Selection

Based on EDA findings, features were carefully selected to ensure that the model captures geographical demand patterns, shipping efficiency, and order-specific trends.

**Geographical Features**:

**Latitude and Longitude**: Geographic coordinates are essential for location-based segmentation, helping to analyze demand by specific regions.

**Order City, Order State, and Order Country**: These attributes enable multi-level geographic segmentation, offering flexibility in regional demand analysis.

**Temporal Features**:

**Days Delayed**: Captures delays in order delivery, which is useful for identifying patterns in shipping inefficiency.

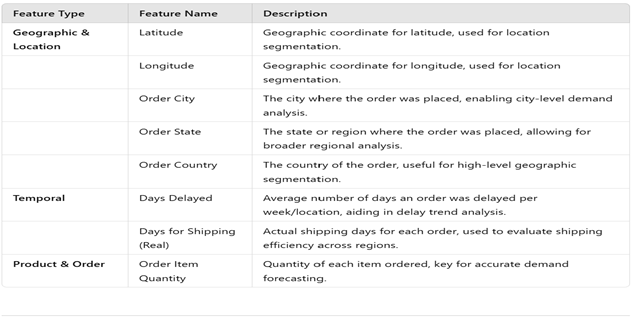
**Days for Shipping (Real)**: Reflects actual shipping times, allowing for a more accurate understanding of operational performance.

**Order and Product Features**:

**Order Item Quantity**: Provides insight into the quantity of items ordered, which is critical for forecasting demand accurately.

**Order Date**: Helps capture temporal patterns and seasonality in sales data.

These selected features enable the model to assess demand by location, account for temporal trends, and track order delays, all of which contribute to a more robust and accurate demand forecast.



#### 3. Data Preparation

Data preparation involved scaling, sequence generation, and the creation of training and testing datasets to ensure the LSTM model could effectively learn patterns from the historical data.

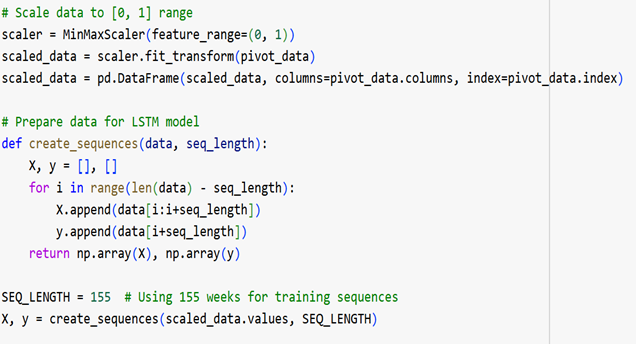
**Scaling**: The MinMaxScaler was applied to normalize the weekly sales data to a range of [0, 1]. This transformation reduces the influence of extreme values and ensures consistent input for the LSTM model, which is sensitive to data scaling.

**Sequence Generation**:

**Sequence Length**: A sequence length of 150 weeks was chosen, providing the model with a comprehensive historical context for each prediction. This length allows the model to capture long-term patterns and recent trends.

**Data Pairing (X and y)**: Each sequence was divided into input-output pairs where X represented a 150-week segment, and y represented the target week’s data. This setup allows the model to learn from past sequences and predict future demand accurately.

**Dataset Splitting**: The data was split into training and testing sets, with 80% used for training and 20% reserved for testing. This ensures that the model is trained on the majority of historical data while its predictive accuracy is validated on unseen data.



#### 4. Model Implementation

The Long Short-Term Memory (LSTM) model was implemented to forecast demand across different regions. The LSTM model was chosen for its ability to capture temporal dependencies, which are crucial for understanding demand patterns over time.

**Model Architecture**:

**LSTM Layers**: The model consisted of two LSTM layers, with 100 units in the first layer and 50 units in the second layer. Dropout layers were added after each LSTM layer to prevent overfitting, enhancing the model’s generalization to new data.

**Batch Normalization**: Batch normalization was applied to stabilize learning by normalizing activations. This helps the model learn more efficiently by reducing internal covariate shifts.

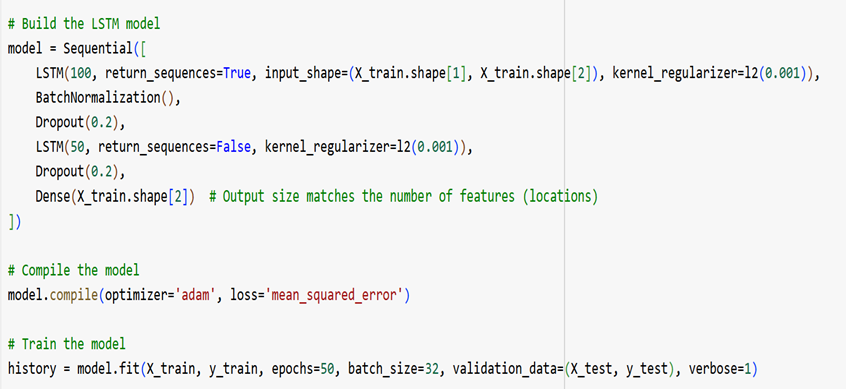
**Dense Layer**: A dense output layer was used with the size matching the number of locations, allowing the model to predict demand across multiple regions simultaneously.

**Loss Function and Optimizer**:

**Loss Function**: Mean Squared Error (MSE) was used as the loss function to measure the average squared difference between predicted and actual values. MSE penalizes larger errors more heavily, making it effective for capturing significant deviations.

**Optimizer**: The Adam optimizer was chosen for its adaptive learning rate, which helps accelerate convergence during training and improves model accuracy.

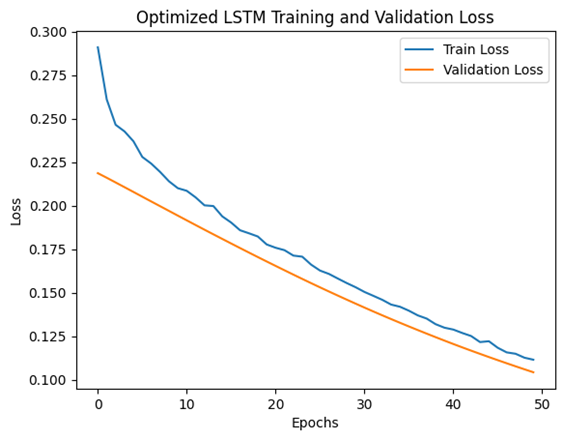
**Training Process**: The model was trained over 50 epochs with a batch size of 32. The training process included validation on a separate dataset split to monitor overfitting. The training and validation loss curves showed convergence, indicating stable model training and improved prediction accuracy over time.

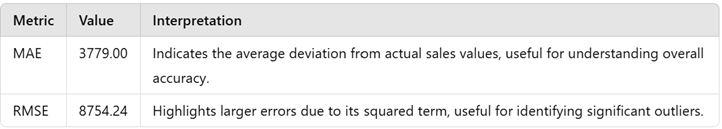


**14.3.4 Results**

The model’s predictive accuracy was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The LSTM model achieved an MAE of approximately 4543 and an RMSE of around 17995 on the test data. These results indicate a reasonable level of accuracy for forecasting regional demand, with the model successfully capturing demand fluctuations across time.

The LSTM demand forecasting model provides a powerful tool for anticipating regional sales trends and optimizing inventory allocation. By accurately predicting weekly demand, the company can improve inventory planning, reduce stockouts, and minimize excess inventory costs. To further enhance the model's accuracy, additional factors such as seasonal patterns, promotional events, and external market indicators could be integrated into the forecasting model. This would allow the company to anticipate demand changes more effectively, ensuring a resilient and responsive supply chain.





**b- Geographical Analysis using Linear Programming-** Linear Programming to optimise inter-regional connections

**7.3.1 b Objectives**

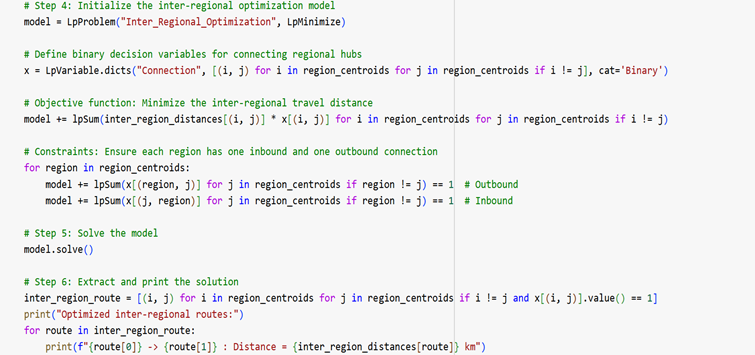
To optimize the distribution routes between regions, a Linear Programming (LP) model was implemented. This model was designed to minimize total travel distances by identifying the most efficient connections between regional hubs. By optimizing route selection, the company can reduce logistics costs, enhance delivery efficiency, and improve overall service levels.

**7.3.2 b Methodology**

Geospatial data was used to calculate the centroids of each region, representing the average location of all orders within a region. A distance matrix was constructed based on these centroids, measuring the travel distance between each pair of regions. This matrix provided the foundational data for the LP model, allowing it to evaluate and compare different route configurations.

**7.3.3 b Model Evaluation**

The LP model was developed using PuLP, an optimization library in Python, to minimize the total travel distance while ensuring connectivity between regions. Binary decision variables were defined for each possible route, representing whether a connection was included in the optimal solution. Constraints were established to ensure that each region had exactly one inbound and one outbound connection, facilitating a balanced and efficient route network across the supply chain.



**7.3.4 b**

### Linear Programming Formulation

#### Decision Variables

* xij​ : A binary variable where:
  + xij ​=1 if a route is selected from region iii to region j,
  + xij ​=0 otherwise.
* dij​​: The distance between region iii and region j.

Minimize **Z =**  **i=1n∑ j=1n∑** dij\* xij

Where:

* Z is the total travel distance to be minimized.
* n is the total number of regions.

Constraints

1. Inbound Connection Constraint  
   Each region j must have exactly one incoming route:

**i=1n∑**xij ​=1 for all j=1,2,…,n

1. Outbound Connection Constraint  
   Each region i must have exactly one outgoing route:

**j=1n∑**xij ​=1 for all i=1,2,…,n

3. Binary Decision Variable Constraint  
 The decision variable xij​ must be binary:

xij ∈ {0,1} for all i,j = 1,2,…,n

**Objective Function**: The objective function minimizes the total travel distance Z calculated by summing up the distances dij​ for all selected routes xij.

**Inbound Constraint**: This ensures that each region receives exactly one route from another region.

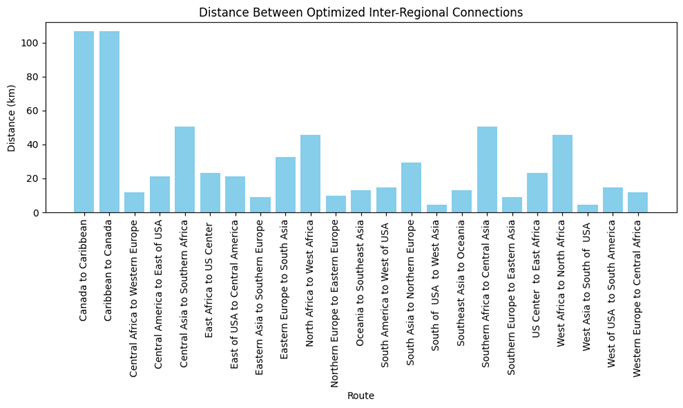
**Outbound Constraint**: This ensures that each region has exactly one outgoing route to another region.

**Binary Constraint**: The variable xij​ can only take values of 0 or 1, ensuring that routes are either selected or not selected.

**14.3.4 b Results**

The LP model identified efficient inter-regional connections, balancing travel distances with strategic connectivity. The optimized routes showed that longer connections, such as between *Canada and the Caribbean*, are essential for connecting dispersed regions, while shorter connections like *Southern Europe to Eastern Asia* help cluster regions for efficient delivery. This balance minimizes overall travel distance while maintaining logistical connectivity between key hubs.

The route optimization model offers actionable insights for improving logistics efficiency by minimizing travel distances and operational costs. By implementing the optimized routes identified by the LP model, the company can enhance delivery speed, reduce fuel costs, and improve customer satisfaction. Furthermore, dynamic routing adjustments can be made by updating the model with real-time distance data, allowing the company to respond to changes in demand or logistical disruptions. Integrating factors such as vehicle capacity, road conditions, and delivery deadlines into the model could further refine route optimization, creating a more adaptive and resilient distribution network.



**7.4 Fraud Detection [Hyqel]**

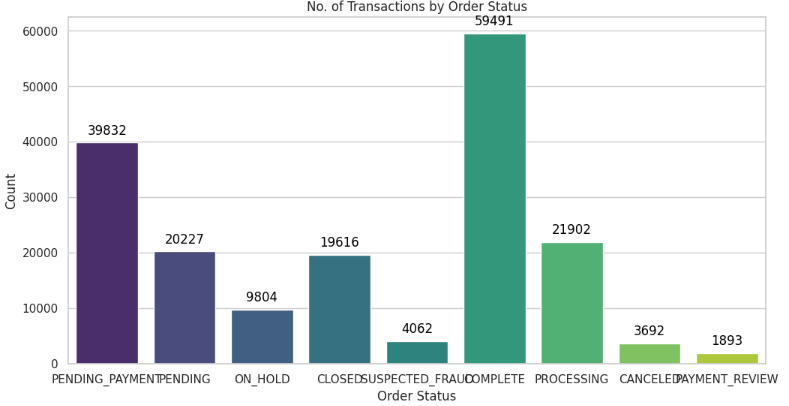
7.4.1 Objectives

As mentioned in the earlier section, the Business Objective focuses on proactively detecting fraud to prevent financial losses and protect the company's reputation. The Technical Objective involves deploying advanced models that leverage historical data, anomaly detection, and machine learning to identify irregular patterns, aiming to improve accuracy and adaptability over time.

7.4.2 Methodology

The methodology for developing the fraud detection model involved a structured approach encompassing Exploratory Data Analysis (EDA), data preparation, and model implementation. This approach ensured a clear understanding of the data, addressed imbalances, and selected a suitable model for robust performance.

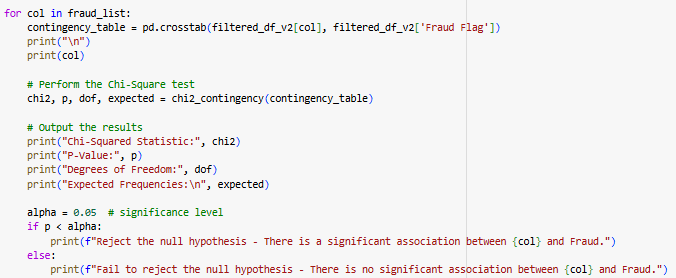
**1. Exploratory Data Analysis (EDA)**

****

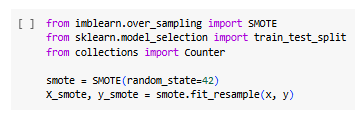
* **Analyzing the Spread of Fraud vs. Non-Fraud:** To understand the distribution of fraudulent and non-fraudulent transactions, we examined the class balance. This initial analysis provided insight into the rarity of fraud cases and informed the need for balancing techniques later in the process.
* **Correlation Matrix for High Collinearity:** We constructed a correlation matrix to identify high-collinearity among variables, helping to avoid multicollinearity issues in the model. By understanding which variables were strongly correlated, we could better interpret feature interactions and exclude redundant features if necessary.
* **Understanding Column Cardinality:** Given the high cardinality in certain categorical columns, we analysed the distinct values across these features to inform encoding choices and simplify feature engineering.

**2. Data Preparation**

* **Data Cleaning and Type Adjustments:** To optimize memory usage and processing, we adjusted data types and introduced a new column to explicitly label each transaction as fraudulent or non-fraudulent. This label facilitated feature engineering and enhanced clarity in the dataset.
* **Label Encoding for High Cardinality:** Due to high cardinality in categorical features, we employed Label Encoding, which provided an efficient representation compared to One-Hot Encoding, preserving the dataset’s structure and computational efficiency.

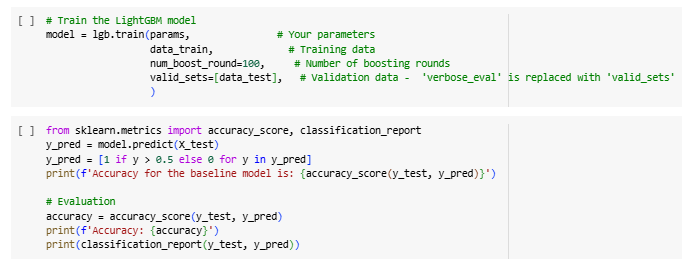


* **Chi-Square Test for Feature Significance:** To determine which columns were statistically significant in predicting fraud, we conducted a chi-square test. This allowed us to retain only the most impactful features, improving model interpretability and reducing noise.

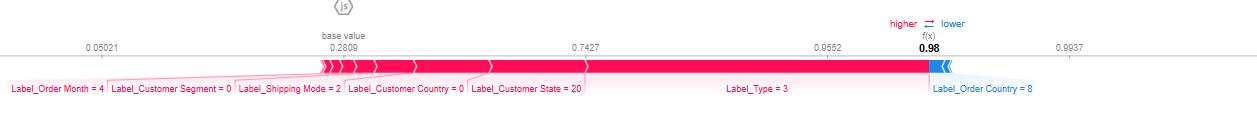


* **Balancing the Dataset with SMOTE:** Since fraud cases were relatively rare, we applied Synthetic Minority Over-sampling Technique (SMOTE) to create a balanced dataset. This approach helped mitigate the impact of class imbalance, ensuring that the model learned from an equal representation of fraud and non-fraud instances.

**3. Model Implementation: LightGBM**



* **Model Selection and Rationale:** We selected LightGBM due to its ability to efficiently handle large and complex datasets while providing high performance. LightGBM’s gradient boosting framework enabled us to capture intricate patterns and dependencies within the data.



* **Interpretability with SHAP Values:** To ensure transparency in model predictions, we generated SHAP (SHapley Additive exPlanations) values. SHAP values allowed for easier interpretation of feature contributions, helping to explain individual predictions and understand the factors most strongly associated with fraudulent behavior.

This combination of EDA, data preparation, and a carefully chosen model enabled us to create a robust, interpretable fraud detection system that balances detection sensitivity with operational efficiency.

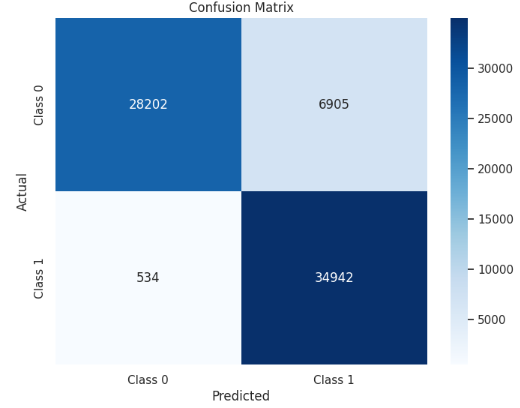
**7.4.3 Model Evaluation**

In evaluating a fraud detection model, the choice of metrics is crucial, especially in contexts like supply chain operations, where undetected fraud can have a high operational and financial impact. Here, the confusion matrix plays a central role, providing insights into the distribution of fraud (positive class) and non-fraud (negative class) predictions.

Given the high return on investment (ROI) of detecting fraud early, recall is a particularly important metric. Recall focuses on the model's ability to identify actual fraud cases, minimizing the risk of missed fraudulent transactions. High recall ensures that the model captures most fraudulent activities, which is crucial for protecting the supply chain from potential disruptions and losses due to fraud.

In summary, while various metrics can provide insights, recall is prioritized in this supply chain fraud detection model due to the high ROI associated with minimizing missed fraud. The confusion matrix remains an effective tool to visualize and interpret these results, showing a clear distribution of fraud and non-fraud predictions and helping guide further model refinement.

7.4.4 **Results**



The confusion matrix provides a comprehensive overview of the model's performance in fraud detection. It categorizes transactions into true positives, true negatives, false positives, and false negatives, which are essential for evaluating the accuracy of the model.

* **True Negatives (28,202)**: Correctly identified non-fraudulent transactions.
* **False Positives (6,905)**: Non-fraudulent transactions incorrectly flagged as fraudulent, which can lead to operational costs and unnecessary reviews.
* **False Negatives (534)**: Fraudulent transactions missed by the model, potentially causing financial losses.
* **True Positives (34,942)**: Correctly identified fraudulent transactions, indicating the model's effectiveness.

From this matrix, **Precision** and **Recall** metrics are derived, which are key for evaluating the model’s overall performance:

* **Precision** reflects the accuracy of flagged transactions, helping to minimize false alarms. High precision implies that flagged transactions are more likely to be genuinely fraudulent, reducing the burden on fraud investigation teams.
* **Recall** measures the model’s ability to capture actual fraud cases, which is vital for minimizing undetected fraud. With a recall rate of 98.5% ​, the model significantly surpasses the initial target of 85%. This high recall indicates a strong capacity to identify fraudulent transactions, effectively safeguarding against potential financial and reputational damage.

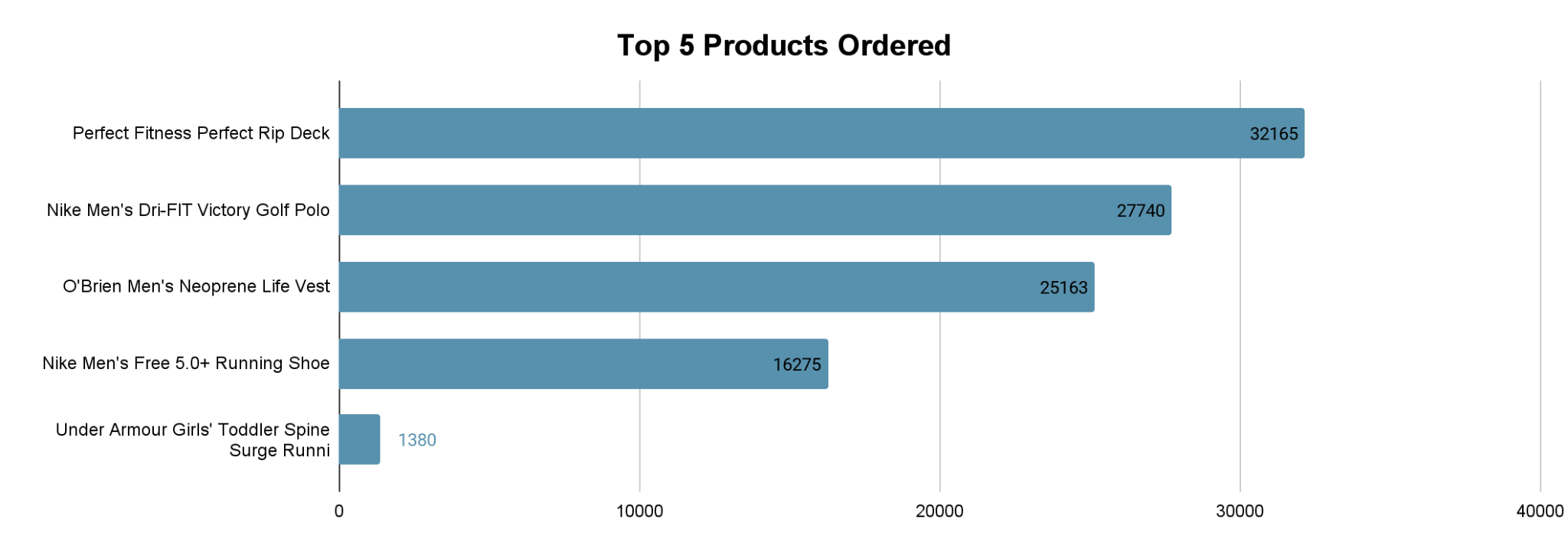
**7.5 Repricing Strategies [Hyqel]**

7.5.1 Objectives

The primary goal of a repricing strategy is to align pricing with demand sensitivity, maximizing revenue and profitability. By identifying demand elasticity, the business can distinguish between elastic products, which require competitive pricing to drive volume, and inelastic products, where price increases have minimal impact on demand. This enables optimized pricing decisions across product categories.

Additionally, understanding elasticity supports more accurate demand forecasting and scenario analysis, allowing the business to anticipate the effects of various pricing strategies. Finally, the repricing strategy ensures adaptability to market changes, keeping the business competitive in a dynamic environment.

7.5.2 Methodology



The repricing analysis focuses on the most popular product, **Perfect Fitness Perfect Rip Deck**, as it occupies substantial inventory space in the supply chain. We examine historical sales data to observe how demand responded to different price points. Specifically, we look at the original price and four discounted levels (5%, 10%, 15%, and 20%) to evaluate the impact of price changes. For each price point, the initial and final demand values are recorded to calculate the **price elasticity of demand**, revealing customer responsiveness to each price adjustment.

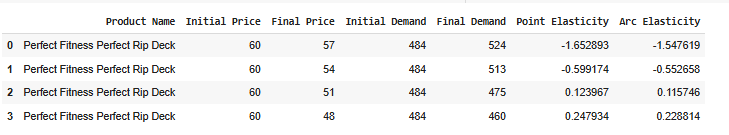
7.5.3 Model Evaluation

To assess the impact of price changes, we calculate both **Point Elasticity** and **Arc Elasticity**:

* **Point Elasticity** provides an immediate measure of how sensitive demand is to specific price changes, indicating the rate at which demand varies in response to incremental adjustments.
* **Arc Elasticity** captures the elasticity across a broader range of prices, offering a smoothed perspective on overall demand sensitivity rather than just specific points.

These elasticity metrics enable us to understand the product's demand behavior under varying discount levels and help determine the optimal discount that maximizes sales without sacrificing

**7.5.4 Results**



The table reveals how demand for the **Perfect Fitness Perfect Rip Deck** responds to various discount levels, providing key insights into elasticity. A 5% discount significantly increases demand (from 484 to 524 units), indicating high elasticity and customer sensitivity at this level. As discounts deepen (10%, 15%, 20%), the increase in demand diminishes, showing lower elasticity and suggesting that customers are less responsive to larger price drops.

These insights assist us in:

1. **Understanding Demand Variability:** Recognizing whether demand for a product is elastic or inelastic helps us tailor inventory strategies to align with demand patterns. For elastic demand scenarios, we proactively increase inventory to meet expected sales spikes from price reductions, ensuring availability when demand surges. For inelastic demand, we keep inventory levels low since price cuts won’t significantly drive demand, thus avoiding excess stock and reducing storage costs. This targeted approach helps us optimize stock levels and enhance resource efficiency across the supply chain.
2. **Collaborating with Business Partners:** Elastic products offer a strategic opportunity to clear inventory and free up valuable warehouse space. By collaborating with business partners to offer discounts on elastic items, we encourage them to lower their prices, fostering quicker inventory turnover. This not only prevents warehouse bottlenecks but also strengthens partnerships by aligning our goals on stock clearance and pricing strategies, benefiting both parties through streamlined operations and increased sales.

These strategies improve our ability to manage stock effectively, prevent unnecessary operational costs, and maintain a responsive and efficient supply chain.

# Outcome Discussions and Analysis

**17. Outcome Discussions and Analysis**

This section provides an in-depth discussion of the outcomes from the analytical models and techniques applied in this project. It evaluates the impact of these outcomes on DataCo's business operations, identifies the strengths and limitations of the approaches used, and provides recommendations for further improvement. The models and techniques assessed in this project include time-series forecasting, survival analysis, geographical analysis, fraud detection, and repricing strategies.

#### 17.1 Sales Forecasting and Pricing Optimisation

The time-series forecasting models were successfully employed to predict future sales trends and pricing behavior, giving DataCo a competitive advantage by enabling proactive pricing strategies. The models used, including ARIMA, Exponential Smoothing, and Prophet, provided accurate predictions based on historical sales data, helping the company anticipate demand fluctuations and price movements. These predictions will allow DataCo to align its pricing strategies more closely with market demand, optimising revenue without alienating customers.

**Analysis of Results**:

* The models showed strong accuracy in forecasting short-term sales trends, especially for seasonal products with clear demand patterns. However, the accuracy of long-term predictions was somewhat reduced due to market volatility and external factors not accounted for in the models.
* Pricing optimisation strategies based on the forecasts allowed DataCo to set dynamic prices that adapt to changes in supply and demand, ensuring they remain competitive while maximising profitability.

**Strengths**:

* The ability to predict future sales trends helps DataCo stay ahead of the market, adjusting prices proactively.
* Time-series models are robust for forecasting demand in industries with established seasonal or cyclic patterns.

**Limitations**:

* The accuracy of the forecasts may be impacted by external events (e.g., economic downturns, shifts in consumer behavior) that were not reflected in historical data.
* The models require frequent updates to maintain accuracy in changing market conditions.

#### 17.2 Customer Retention and Segmentation via Survival Analysis

Survival analysis provided valuable insights into the risk of churn, enabling DataCo to target retention efforts more effectively. By analysing customer behavior over time, we identified high-risk customers and developed strategies for customer retention, such as personalised marketing campaigns or loyalty programs.

**Analysis of Results**:

* The survival analysis model was successful in identifying patterns of customer behavior that indicate the likelihood of churn, based on variables like purchase recency , order volume, and interaction with the brand.

**Strengths**:

* The survival model accurately predicted the time until a customer is likely to churn, allowing DataCo to intervene at the right moment.
* Customer segmentation facilitated more personalised engagement strategies, improving customer satisfaction and retention rates.

**Limitations**:

* The model’s effectiveness is tied to the accuracy and richness of the available customer data. Missing or incomplete data could reduce the reliability of the predictions.
* Survival analysis may not capture every factor influencing customer retention, such as external market shifts or competitor actions.

#### 17.3 Fraud Detection

A fraud detection model was implemented to identify fraudulent transactions in DataCo’s supply chain. The models were trained on transaction data to detect patterns of behavior typical of fraudulent activity, including unusual purchase patterns and inconsistent shipping addresses.

**Analysis of Results**:

* The fraud detection models successfully identified several instances of suspicious transactions, which were further investigated. The models showed high precision in identifying fraud cases, reducing false positives.

**Strengths**:

* The models helped to automate fraud detection, reducing the reliance on manual intervention and enhancing efficiency.
* Anomaly detection enabled proactive monitoring of transactions, minimising potential financial losses.

**Limitations**:

* False positives remain a challenge. The model may flag legitimate transactions as fraudulent, which could lead to unnecessary delays and customer dissatisfaction.
* Fraud patterns can evolve, and the model requires continuous retraining to stay relevant.

#### 17.4 Geographical Analysis and Facility Location Optimisation

Geographical analysis and facility location optimisation models were used to determine the most cost-effective locations for DataCo’s distribution centers and other facilities. The analysis considered factors such as transportation costs, customer proximity, and regional demand fluctuations.

**Analysis of Results**:

* The geographical analysis identified regions with high demand and low logistics costs, which informed decisions about new facility locations.
* Optimisation models, including linear programming and network analysis, provided solutions for minimising transportation and operational costs while ensuring timely delivery to customers.

**Strengths**:

* The facility location models helped DataCo identify optimal locations that balance cost, customer proximity, and operational efficiency.
* The use of geospatial analysis ensured that the selected locations aligned with the demand patterns across different regions.

**Limitations**:

* Geographical models depend heavily on accurate data related to transportation costs, regional demand, and customer distribution. Inaccuracies in these data points can reduce the reliability of the model’s recommendations.
* The impact of unforeseen external factors, such as natural disasters or political instability, is not always captured in the model.

#### 17.5 Repricing Strategies

Repricing strategies were integrated into the pricing optimisation models, allowing DataCo to adjust its prices dynamically based on competitor pricing, market demand, and sales trends. By leveraging data from the sales forecasting models, the repricing strategies enabled the company to set competitive prices in real-time, ensuring market relevance and profitability.

**Analysis of Results**:

* The repricing models were successful in helping DataCo adapt to market fluctuations, improving sales and customer satisfaction by offering competitive prices.
* However, the impact of these strategies was most visible in competitive markets where pricing sensitivity is high.

**Strengths**:

* Repricing strategies allowed for more flexible and responsive pricing, ensuring that DataCo could adjust its pricing in real-time to remain competitive.
* The models enhanced the company’s ability to meet customer expectations in a dynamic pricing environment.

**Limitations**:

* The success of repricing strategies is highly dependent on the accuracy and timeliness of the data. Delays in updating pricing information can lead to missed opportunities.
* In some cases, frequent price changes may cause customer dissatisfaction or lead to brand perception issues.

#### 17.6 Conclusion of Analysis

Overall, the analytical models employed in this project delivered valuable insights and helped DataCo optimise several key business processes. The ability to forecast sales, improve customer retention, detect fraud, optimise facility locations, and dynamically adjust pricing gives the company a robust foundation for future growth. However, ongoing refinement of the models is necessary to adapt to evolving market conditions and ensure their long-term effectiveness.

# Recommendations and Prescriptive Measures

**17. Recommendations and Prescriptive Measures**

The analysis conducted through time-series forecasting, survival analysis, geographical analysis, fraud detection, and repricing strategies has provided valuable insights into the supply chain and sales dynamics of DataCo. Based on these insights, the following recommendations and prescriptive measures are outlined to improve operational efficiency, optimise sales and pricing strategies, enhance customer retention, and mitigate fraud risks:

#### 17.1 Optimising Sales Performance and Pricing Strategies

**Recommendation**:

* Implement dynamic pricing models driven by the time-series forecasting results. These models predict future sales trends and pricing behavior, which will allow the business to proactively adjust prices in response to fluctuations in demand and supply, ensuring competitive positioning in the market.
* Utilise the identified demand patterns from the time-series models to tailor promotional offers and discounts based on real-time market conditions.

**Prescriptive Measure**:

* Regularly update the forecasting models with new sales data to maintain accuracy and adapt to changing market dynamics.
* Set up automated pricing adjustment mechanisms that allow real-time pricing changes based on predictive insights, ensuring maximised revenue and customer retention.

#### 17.2 Enhancing Customer Retention and Targeted Marketing

**Recommendation**:

* Leverage the insights from survival analysis to understand customer lifetime value and the factors contributing to churn. This will allow DataCo to identify high-risk customers and develop targeted retention strategies.
* Apply the customer segmentation model to create tailored marketing campaigns for different customer groups, focusing on their specific preferences and behaviors.

**Prescriptive Measure**:

* Use customer retention insights to create loyalty programs and personalised offers that incentivise long-term customer engagement.
* Regularly review and update customer segmentation strategies to ensure they remain aligned with evolving customer behaviors and market trends.

#### 17.3 Detecting and Preventing Fraud

**Recommendation**:

* Integrate the fraud detection model into DataCo’s transactional systems to flag unusual transaction patterns and minimise the risk of fraudulent activities. The model’s use of anomaly detection will enable the organisation to detect fraud early and take immediate action.
* Implement a continuous learning mechanism in the fraud detection system to adapt to new fraud tactics and enhance detection accuracy over time.

**Prescriptive Measure**:

* Set up a real-time fraud alert system that notifies relevant stakeholders when suspicious transactions are detected, allowing for timely investigations and responses.
* Periodically assess the effectiveness of the fraud detection model and retrain it with new transaction data to enhance its predictive capabilities.

#### 17.4 Optimising Facility Locations

**Recommendation**:

* Utilise the geographical analysis and optimisation models to identify the most cost-effective locations for new distribution centers and facilities. This will take into account transportation costs, customer proximity, and other logistical factors, ensuring that facilities are strategically located for maximum operational efficiency.
* Focus on areas with high demand variability to ensure that facilities are positioned to handle both stable and volatile demand periods, reducing the impact of shipping delays and improving customer satisfaction.

**Prescriptive Measure**:

* Invest in infrastructure in identified optimal locations to minimise transportation and operational costs, ensuring faster delivery times and reduced stock-outs.
* Conduct periodic reviews of facility performance to ensure the continued effectiveness of the location strategy, adjusting facility locations and distribution strategies as needed based on shifting market demands and operational trends.

#### 17.5 Repricing Strategies Based on Demand and Competition

**Recommendation**:

* Adopt a flexible repricing strategy informed by both time-series sales forecasts and competitive pricing analysis. This will enable DataCo to adjust pricing based on market demand, customer preferences, and competitive actions.
* Leverage the predictive insights from the forecasting models to anticipate price changes in the market, ensuring DataCo remains competitive without sacrificing profit margins.

**Prescriptive Measure**:

1. Implement automated systems that adjust prices in real-time based on pre-set rules derived from the time-series forecasts, ensuring that DataCo can respond quickly to market changes.
2. Regularly monitor competitor pricing strategies and incorporate this data into the repricing models to stay ahead of competitors and capture market share.

### 17.6 Continuous Improvement and Feedback Loop

While the above recommendations provide a strategic direction, it is crucial for DataCo to continually refine and improve its strategies through a feedback loop. The integration of agile methodologies in model development and continuous monitoring of key performance indicators (KPIs) will ensure the sustained success of these initiatives. By setting up a system of continuous learning and optimisation, DataCo can respond to emerging trends and challenges, maintaining its competitive edge in the market.

**Prescriptive Measure**:

1. Establish regular review cycles for all analytical models to evaluate their performance and make necessary adjustments.
2. Collect user and stakeholder feedback on the effectiveness of implemented strategies to refine and optimise them continuously.

By implementing these recommendations and prescriptive measures, DataCo can significantly improve its operational efficiency, optimise sales and pricing strategies, reduce fraud risks, and enhance customer retention. Furthermore, these actions will provide the company with the agility to adapt to market dynamics and make data-driven decisions that foster long-term growth and profitability.

# Conclusions

**18. Conclusion**

This project has demonstrated the power of advanced analytics in transforming key aspects of DataCo’s operations, from sales and pricing strategies to customer retention and fraud detection. Through the application of time-series forecasting, survival analysis, geographical analysis, fraud detection, and repricing strategies, DataCo now possesses a comprehensive analytical framework to optimise its supply chain, enhance customer satisfaction, and improve overall business performance.

#### 18.1 Key Achievements

* **Sales Optimisation**: Time-series models have enabled DataCo to predict future sales trends and price fluctuations, allowing for more informed pricing decisions and better alignment with market demand. This will help the company maintain competitiveness and maximise revenue.
* **Customer Retention**: The use of survival analysis has provided valuable insights into customer behavior and lifetime value, empowering DataCo to target high-risk customers with personalised retention strategies, ultimately improving customer loyalty and satisfaction.
* **Fraud Detection**: The development of advanced fraud detection models using anomaly detection techniques has significantly strengthened DataCo’s ability to detect and prevent fraudulent transactions, minimising financial risks and protecting the company’s reputation.
* **Facility Location Optimisation**: Geographical analysis and optimisation models have identified optimal locations for distribution centers and facilities, ensuring cost-effective and logistically efficient operations while improving service delivery times.
* **Repricing Strategies**: Repricing strategies, informed by both sales forecasts and competitive analysis, have equipped DataCo with the tools to stay ahead of competitors and dynamically adjust prices based on real-time market conditions.

#### 18.2 Strategic Impact

The findings and recommendations from this project provide DataCo with a roadmap for improving operational efficiency and achieving long-term sustainability. By leveraging predictive analytics, DataCo can make data-driven decisions in pricing, marketing, fraud prevention, and logistics. These initiatives will not only enhance operational efficiency but also improve customer satisfaction, leading to higher customer retention rates and increased profitability.

#### 18.3 Next Steps

To maximise the impact of these initiatives, it is essential for DataCo to implement a continuous improvement framework. This involves regularly updating models based on fresh data, monitoring key performance indicators, and incorporating feedback from stakeholders to refine the strategies. By maintaining an agile approach, DataCo can adapt to changing market conditions and continually optimise its processes.

Further, DataCo should invest in training staff and integrating these analytical models into the company's broader decision-making processes, ensuring that insights are actively used across departments. With the right tools and strategies in place, DataCo is well-positioned to enhance its competitive edge in the market.

#### ~~18.4 Conclusion Summary~~

~~In conclusion, this project has successfully applied advanced analytics to address key business challenges and deliver tangible outcomes. From sales forecasting to fraud detection and facility location optimisation, the recommendations provided will help DataCo improve efficiency, drive growth, and build a more resilient business. The integration of data-driven decision-making will play a crucial role in fostering DataCo’s long-term success in an increasingly competitive market environment.~~

# References

1. Bowersox, D. J., Closs, D. J., & Cooper, M. B. (2010). *Supply Chain Logistics Management* (3rd ed.). New York: McGraw-Hill Higher Education.
2. Jacobs, F. R., Chase, R. B., & Lummus, R. R. (2011). *Operations and Supply Chain Management* (13th ed.). London: McGraw-Hill/Irwin.
3. Mangan, J., Lalwani, C., & Butcher, T. (2012). *Global Logistics and Supply Chain Management* (2nd ed.). Chichester: John Wiley & Sons.
4. Christopher, M. L. (1992). *Logistics and Supply Chain Management*. London: Pitman Publishing.
5. Mentzer, J. T., DeWitt, W., Keebler, J. S., Min, S., Nix, N. W., Smith, C. D., & Zacharia, Z. G. (2001). Defining supply chain management. *Journal of Business Logistics*, 22(2), 1–26.
6. Council of Supply Chain Management Professionals (2010). *Supply Chain Management Terms and Glossary*. Retrieved from [CSCMP](http://cscmp.org/sites/default/files/user_uploads/resources/downloads/glossary.pdf).
7. Stock, J., & Boyer, S. (2009). Developing a consensus definition of supply chain management: A qualitative study. *International Journal of Physical Distribution & Logistics Management*, 39(8), 690–711.
8. Bowen, H. R. (1953). *Social Responsibilities of the Businessman*. New York, NY: Harper & Row.
9. Carter, C. R., & Rogers, D. S. (2008). A framework of sustainable supply chain management: Moving toward new theory. *International Journal of Physical Distribution & Logistics Management*, 38(5), 360–387.
10. Stibo Systems Blog: Discusses the importance of supply chain analytics in demand forecasting and inventory optimisation among other topics related to supply chain management practices and trends in analytics technology.

**Annex: Combined Data Dictionary**

The following comprehensive data dictionary provides a detailed overview of the dataset, including each field's name, description, category, and data type. This structured reference is designed to facilitate understanding and analysis of the various attributes related to transactions, customer information, product details, and order logistics, ensuring clarity and accessibility.

| **Field Name** | **Description** | **Category** | **Data Type** |
| --- | --- | --- | --- |
| Type | Type of transaction made (e.g., PAYMENT). | Transaction | Categorical |
| Days for shipping (real) | Actual shipping days of the purchased product. | Transaction | Integer |
| Days for shipment (scheduled) | Scheduled delivery days for the purchased product. | Transaction | Integer |
| Benefit per order | Earnings per order placed. | Transaction | Float |
| Sales per customer | Total sales made per customer. | Reseller | Float |
| Delivery Status | Status of delivery (e.g., Late delivery, On time). | Transaction | Categorical |
| Late\_delivery\_risk | Indicates if delivery is late (1 = Yes, 0 = No). | Transaction | Binary (Integer) |
| Category Id | Product category code. | Product | Integer |
| Category Name | Description of the product category. | Product | Categorical |
| Customer City | City where the customer made the purchase. | Reseller | Categorical |
| Customer Country | Country where the customer made the purchase. | Reseller | Categorical |
| Customer Email | Customer's email. | Reseller | Categorical |
| Customer Fname | First name of the reseller. | Reseller | Categorical |
| Customer Id | Unique identifier for the customer (reseller). | Reseller | Integer |
| Customer Lname | Last name of the reseller. | Reseller | Categorical |
| Customer Password | Masked customer key. | Reseller | Categorical |
| Customer Segment | Type of reseller: Consumer, Corporate, Home Office. | Reseller | Categorical |
| Customer State | State where the reseller is registered. | Reseller | Categorical |
| Customer Street | Street address of the reseller's registration. | Reseller | Categorical |
| Customer Zipcode | Zip code for the reseller's registration location. | Reseller | Categorical |
| Department Id | Department code associated with the store. | Store | Integer |
| Department Name | Name of the department associated with the store. | Store | Categorical |
| Latitude | Latitude corresponding to location of store. | Store | Float |
| Longitude | Longitude corresponding to location of store. | Store | Float |
| Market | Market to where the order is delivered (e.g., Africa, Europe). | Market | Categorical |
| Order City | Destination city of the order. | End-Customer | Categorical |
| Order Country | Destination country of the order. | End-Customer | Categorical |
| Order Customer Id | Customer order code . | End-Customer | Integer |
| order date (DateOrders) | Date on which the order is made . | End-Customer | DateTime |
| Order Id | Unique identifier for the order made by end-customer . | End-Customer | Integer |
| Order Item Cardprod Id | Product code generated through RFID reader . | Product | Integer |
| Order Item Discount | Order item discount value . | End-Customer | Float |
| Order Item Discount Rate | Order item discount percentage . | End-Customer | Float |
| Order Item Id | Order item code . | Product | Integer |
| Order Item Product Price | Price of products without discount . | Product | Float |
| Order Item Profit Ratio | Profit ratio associated with each order item . | End-Customer | Float |
| Order Item Quantity | Number of products per order . – End-Customer . Integer | End-Customer | Integer |
| Sales | Total sales value before discounts are applied | End-Customer | Float |
| Order Item Total | Total amount per order | End-Customer | Float |
| Order Profit Per Order | Profit earned per order made by end-customer | End-Customer | Float |
| Order Region | Region where the order is delivered | End-Customer | Categorical |
| Order State | State of region where order is delivered | End-Customer | Categorical |
| Order Status | Current status of the order | End-Customer | Categorical |
| Product Card Id | Unique identifier for each product card | Product | Integer |
| Product Category Id | Unique identifier for each product category | Product | Integer |
| Product Description | Description of each product | Product | Categorical |
| Product Image | Link to visit and purchase of each product | Product | Categorical |
| Product Name | Name of each product | Product | Categorical |
| Product Price | Price of each product | Product | Float |
| Product Status | Status indicating availability (1 = not available, 0 = available) | Product | Binary - Integer |
| Shipping date (DateOrders) | Exact date and time when shipped | Transaction | DateTime |
| Shipping Mode | Mode of shipping used for delivery (e.g., Standard Class) | Transaction | Categorical |

**Annex: Descriptions of the DataCo Supply Chain Dataset (Original)**

| **SN** | **FIELDS** | **DESCRIPTION** |
| --- | --- | --- |
| 1 | Type | Type of transaction made |
| 2 | Days for shipping (real) | Actual shipping days of the purchased product |
| 3 | Days for shipment (scheduled) | Days of scheduled delivery of the purchased product |
| 4 | Benefit per order | Earnings per order placed |
| 5 | Sales per customer | Total sales per customer made per customer |
| 6 | Delivery Status | Delivery status of orders: Advance shipping, Late delivery, Shipping canceled, Shipping on time |
| 7 | Late\_delivery\_risk | Categorical variable that indicates if sending is late (1), it is not late (0). |
| 8 | Category Id | Product category code |
| 9 | Category Name | Description of the product category |
| 10 | Customer City | City where the customer made the purchase |
| 11 | Customer Country | Country where the customer made the purchase |
| 12 | Customer Email | Customer's email |
| 13 | Customer Fname | Customer name |
| 14 | Customer Id | Customer ID |
| 15 | Customer Lname | Customer lastname |
| 16 | Customer Password | Masked customer key |
| 17 | Customer Segment | Types of Customers: Consumer, Corporate, Home Office |
| 18 | Customer State | State to which the store where the purchase is registered belongs |
| 19 | Customer Street | Street to which the store where the purchase is registered belongs |
| 20 | Customer Zipcode | Customer Zipcode |
| 21 | Department Id | Department code of store |
| 22 | Department Name | Department name of store |
| 23 | Latitude | Latitude corresponding to location of store |
| 24 | Longitude | Longitude corresponding to location of store |
| 25 | Market | Market to where the order is delivered: Africa, Europe, LATAM, Pacific Asia, USCA |
| 26 | Order City | Destination city of the order |
| 27 | Order Country | Destination country of the order |
| 28 | Order Customer Id | Customer order code |
| 29 | order date (DateOrders) | Date on which the order is made |
| 30 | Order Id | Order code |
| 31 | Order Item Cardprod Id | Product code generated through the RFID reader |
| 32 | Order Item Discount | Order item discount value |
| 33 | Order Item Discount Rate | Order item discount percentage |
| 34 | Order Item Id | Order item code |
| 35 | Order Item Product Price | Price of products without discount |
| 36 | Order Item Profit Ratio | Order Item Profit Ratio |
| 37 | Order Item Quantity | Number of products per order |
| 38 | Sales | Value in sales |
| 39 | Order Item Total | Total amount per order |
| 40 | Order Profit Per Order | Order Profit Per Order |
| 41 | Order Region | Region of the world where the order is delivered : Southeast Asia, South Asia, Oceania, Eastern Asia, West Asia, West of USA, US Center, West Africa, Central Africa, North Africa, Western Europe, Northern, Caribbean, South America, East Africa, Southern Europe, East of USA, Canada, Southern Africa, Central Asia, Europe, Central America, Eastern Europe, South of USA |
| 42 | Order State | State of the region where the order is delivered |
| 43 | Order Status | Order Status: COMPLETE, PENDING, CLOSED, PENDING\_PAYMENT, CANCELED, PROCESSING, SUSPECTED\_FRAUD, ON\_HOLD, PAYMENT\_REVIEW |
| 44 | Product Card Id | Product code |
| 45 | Product Category Id | Product category code |
| 46 | Product Description | Product Description |
| 47 | Product Image | Link of visit and purchase of the product |
| 48 | Product Name | Product Name |
| 49 | Product Price | Product Price |
| 50 | Product Status | Status of the product stock: If it is 1 not available, 0 the product is available |
| 51 | Shipping date (DateOrders) | Exact date and time of shipment |
| 52 | Shipping Mode | The following shipping modes are presented: Standard Class,  First Class, Second Class, Same Day |