**DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS**

* Candidate Name : Manish
* Candidate E-Mail : [manish97cta@gmail.com](mailto:manish97cta@gmail.com)
* Project Assessment : Capstone

# PROJECT SUMMARY:

Business case and goal of the project:

In an era of escalating demand and intricate logistics, Dataco aims to revolutionize its global supply chain management by accurately predicting late deliveries. This project leverages cutting-edge machine learning techniques to analyse historical delivery data, ensuring timely insights and proactive measures to strengthen Dataco's delivery competence, customer satisfaction and strategic decision-making in supply chain management in the increasingly competitive world.

It also undermines the customer's trust in the company. As a result, the company loses credibility and profit over time through losing customers. By preventing late delivery, the company may help retain customers, increase customer lifetime value, and increase its ROI.

By harnessing the predictive power of machine learning, Dataco can optimize its supply chain operations, mitigate delivery risks, and foster long-term growth and success in the dynamic logistics landscape.

Dataco is a leading company in the field of global supply chain management, specializing in providing innovative solutions for logistics and delivery services.

The Dataco supply chain dataset is a comprehensive collection of historical data encompassing various aspects of the supply chain, including order processing, shipping, delivery, and customer interactions. This dataset contains a wealth of information, including order details, delivery statuses, customer demographics, product information, and more.

**Benefits**:

By proactively addressing late deliveries, Dataco can significantly enhance its operational efficiency and customer satisfaction levels. Timely deliveries not only bolster customer loyalty but also strengthen Dataco's position in the global supply chain market. Moreover, by leveraging machine learning insights, Dataco can streamline its logistics processes, reduce costs, and gain a competitive edge in the industry.

**The Goal and Insights of the project are as follows:**

1. Understanding the variables that are causing the late delivery.

Index:

# Analysis.

1. **Data Preprocessing**
2. **Data Visualization, Storytelling & Experimenting with charts.**
3. **Model**

Top of Form

1. **Analysis:**

The given dataset is from supply chain industry. This dataset consists of 1,80,519 rows and 53 columns.

**Summary of Categorical Attributes:**

We have total 24 categorical attributes in the dataset. Categories are summarized below in terms of description, number of levels and examples of levels.

**Categorical Attributes:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Description** | **No of Levels** | **Categories** |
| Type | Type of transaction made | 4 | DEBIT: 69295 TRANSFER: 49883 PAYMENT: 41725 CASH: 19616 |
| Delivery Status | Delivery status of orders | 4 | Late delivery: 98977 Advance shipping: 41592 Shipping on time: 32196 Shipping canceled: 7754 |
| Category Name | Description of the product category | 50 | Cleats: 24551 Men's Footwear: 22246 Women's Apparel: 21035 Indoor/Outdoor Games: 19298 Fishing: 17325 |
| Customer City | City where the customer made the purchase | 563 | Caguas: 66770 Chicago: 3885 Los Angeles: 3417 Brooklyn: 3412 New York: 1816 |
| Customer Country | Country where the customer made the purchase | 2 | EE. UU.: 111146 Puerto Rico: 69373 |
| Customer Email | Customer's email address | 1 | XXXXXXXXX |
| Customer Fname | Customer first name | 782 | Mary: 65150 James: 1835 Robert: 1759 Michael: 1680 David: 1625 |
| Customer Lname | Customer last name | 1109 | Smith: 64104 Johnson: 989 Brown: 909 Williams: 869 Jones: 859 |
| Customer Password | Masked customer key | 1 | XXXXXXXXX |
| Customer Segment | Types of Customers | 3 | Consumer: 93504 Corporate: 54789 Home Office: 32226 |
| Customer State | State to which the store where the purchase is registered belongs | 46 | PR: 69373 CA: 29223 NY: 11327 TX: 9103 IL: 7631 |
| Customer Street | Street to which the store where the purchase is registered belongs | 7458 | 9126 Wishing Expressway: 122 4388 Burning Goose Ridge: 117 4720 Noble Hills Wynd: 116 2878 Hazy Wagon Thicket: 113 398 Emerald Grove: 109 |
| Department Name | Department name of store | 11 | Fan Shop: 66861 Apparel: 48998 Golf: 33220 Footwear: 14525 Outdoors: 9686 |
| Market | Market to where the order is delivered | 5 | LATAM: 51594 Europe: 50252 Pacific Asia: 41260 USCA: 25799 Africa: 11614 |
| Order City | Destination city of the order | 3597 | Santo Domingo: 2211 New York City: 2202 Los Angeles: 1845 Tegucigalpa: 1783 Managua: 1682 |
| Order Country | Destination country of the order | 164 | Estados Unidos: 24840 Francia: 13222 Mexico: 13172 Alemania: 9564 Australia: 8497 |
| order date (DateOrders) | Date on which the order is made | 65752 | 4/29/2017 10:11 5 7/16/2015 14:47 5 9/13/2017 12:17 5 11/21/2016 23:12 5 6/5/2016 17:18 5 |
| Order Region | Region of the world where the order is delivered | 23 | Central America: 28341 Western Europe: 27109 South America: 14935 Oceania: 10148 Northern Europe: 9792 |
| Order State | State of the region where the order is delivered | 1089 | Inglaterra: 6722 California: 4966 Isla de Francia: 4580 Renania del Norte-Westfalia: 3303 San Salvador: 3055 |
| Order Status | Order Status | 9 | COMPLETE: 59491 PENDING\_PAYMENT: 39832 PROCESSING: 21902 PENDING: 20227 CLOSED: 19616 |
| Product Image | Link of visit and purchase of the product | 118 | http://images.acmesports.sports/Perfect+Fitness+Perfect+Rip+Deck 24515 |
| Product Name | Product Name | 118 | Perfect Fitness Perfect Rip Deck: 24515 Nike Men's CJ Elite 2 TD Football Cleat: 22246 Nike Men's Dri-FIT Victory Golf Polo: 21035 O'Brien Men's Neoprene Life Vest: 19298 Field & Stream Sportsman 16 Gun Fire Safe: 17325 |
| shipping date (DateOrders) | Exact date and time of shipment | 63701 | 10/25/2016 23:46 10 3/13/2017 3:26 10 2/1/2015 2:35 10 8/7/2016 20:33 10 4/24/2016 2:18 10 |
| Shipping Mode | The following shipping modes are presented | 4 | Standard Class: 107752 Second Class: 35216 First Class: 27814 Same Day: 9737 |

**Summary of Numerical Attributes:**

We have total 29 numerical attributes in the dataset. All attributes are summarized below in five-number summary as well as value count, mean, and standard deviation.

**Numerical Attributes:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute** | **Description** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| Days for shipping (real) | Actual shipping days of the purchased product | 180519 | 3.5 | 1.6 | 0 | 2 | 3 | 5 | 6 |
| Days for shipment (scheduled) | Days of scheduled delivery of the purchased product | 180519 | 2.9 | 1.4 | 0 | 2 | 4 | 4 | 4 |
| Benefit per order | Earnings per order placed | 180519 | 22 | 104.4 | -4275 | 7 | 31.5 | 64.8 | 911.8 |
| Sales per customer | Total sales per customer made per customer | 180519 | 183.1 | 120 | 7.5 | 104.4 | 164 | 247.4 | 1940 |
| Late\_delivery\_risk | Categorical variable that indicates if sending is late (1), it is not late (0). | 180519 | 0.5 | 0.5 | 0 | 0 | 1 | 1 | 1 |
| Category Id | Product category code | 180519 | 31.9 | 15.6 | 2 | 18 | 29 | 45 | 76 |
| Customer Id | Customer ID | 180519 | 6691.4 | 4162.9 | 1 | 3258.5 | 6457 | 9779 | 20757 |
| Customer Zipcode | Customer Zipcode | 180516 | 35921.1 | 37542.5 | 603 | 725 | 19380 | 78207 | 99205 |
| Department Id | Department code of store | 180519 | 5.4 | 1.6 | 2 | 4 | 5 | 7 | 12 |
| Latitude | Latitude corresponding to location of store | 180519 | 29.7 | 9.8 | -33.9 | 18.3 | 33.1 | 39.3 | 48.8 |
| Longitude | Longitude corresponding to location of store | 180519 | -84.9 | 21.4 | -158 | -98.4 | -76.8 | -66.4 | 115.3 |
| Order Customer Id | Customer order code | 180519 | 6691.4 | 4162.9 | 1 | 3258.5 | 6457 | 9779 | 20757 |
| Order Id | Order code | 180519 | 36221.9 | 21045.4 | 1 | 18057 | 36140 | 54144 | 77204 |
| Order Item Cardprod Id | Product code generated through the RFID reader | 180519 | 692.5 | 336.4 | 19 | 403 | 627 | 1004 | 1363 |
| Order Item Discount | Order item discount value | 180519 | 20.7 | 21.8 | 0 | 5.4 | 14 | 30 | 500 |
| Order Item Discount Rate | Order item discount percentage | 180519 | 0.1 | 0.1 | 0 | 0 | 0.1 | 0.2 | 0.3 |
| Order Item Id | Order item code | 180519 | 90260 | 52111.5 | 1 | 45130.5 | 90260 | 135389.5 | 180519 |
| Order Item Product Price | Price of products without discount | 180519 | 141.2 | 139.7 | 10 | 50 | 60 | 200 | 2000 |
| Order Item Profit Ratio | Order Item Profit Ratio | 180519 | 0.1 | 0.5 | -2.8 | 0.1 | 0.3 | 0.4 | 0.5 |
| Order Item Quantity | Number of products per order | 180519 | 2.1 | 1.5 | 1 | 1 | 1 | 3 | 5 |
| Sales | Value in sales | 180519 | 203.8 | 132.3 | 10 | 120 | 199.9 | 300 | 2000 |
| Order Item Total | Total amount per order | 180519 | 183.1 | 120 | 7.5 | 104.4 | 164 | 247.4 | 1940 |
| Order Profit Per Order | Order Profit Per Order | 180519 | 22 | 104.4 | -4275 | 7 | 31.5 | 64.8 | 911.8 |
| Order Zipcode | Order Zip Code | 24840 | 55426.1 | 31919.3 | 1040 | 23464 | 59405 | 90008 | 99301 |
| Product Card Id | Product code | 180519 | 692.5 | 336.4 | 19 | 403 | 627 | 1004 | 1363 |
| Product Category Id | Product category code | 180519 | 31.9 | 15.6 | 2 | 18 | 29 | 45 | 76 |
| Product Description | Product Description | 180519 |  |  |  |  |  |  |  |
| Product Price | Product Price | 180519 | 141.2 | 139.7 | 10 | 50 | 60 | 200 | 2000 |
| Product Status | Status of the product stock. If it is 1 not available, 0 the product is available | 180519 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

1. **Data Preprocessing:**

Cleaning the data for the modelling process has the foremost importance in order to achieve creditable conclusions. I checked all the attributes in my dataset for any attribute has less than 2 unique values.

**checked and handled null values:**

|  |  |
| --- | --- |
| **Attributes** | **Count Null Values** |
| Product Description | 180519 |
| Order Zipcode | 155679 |
| Customer Lname | 8 |
| Customer Zipcode | 3 |

**We checked for duplicates and found that there were no duplicates.**

**Unique Values of Attributes:**

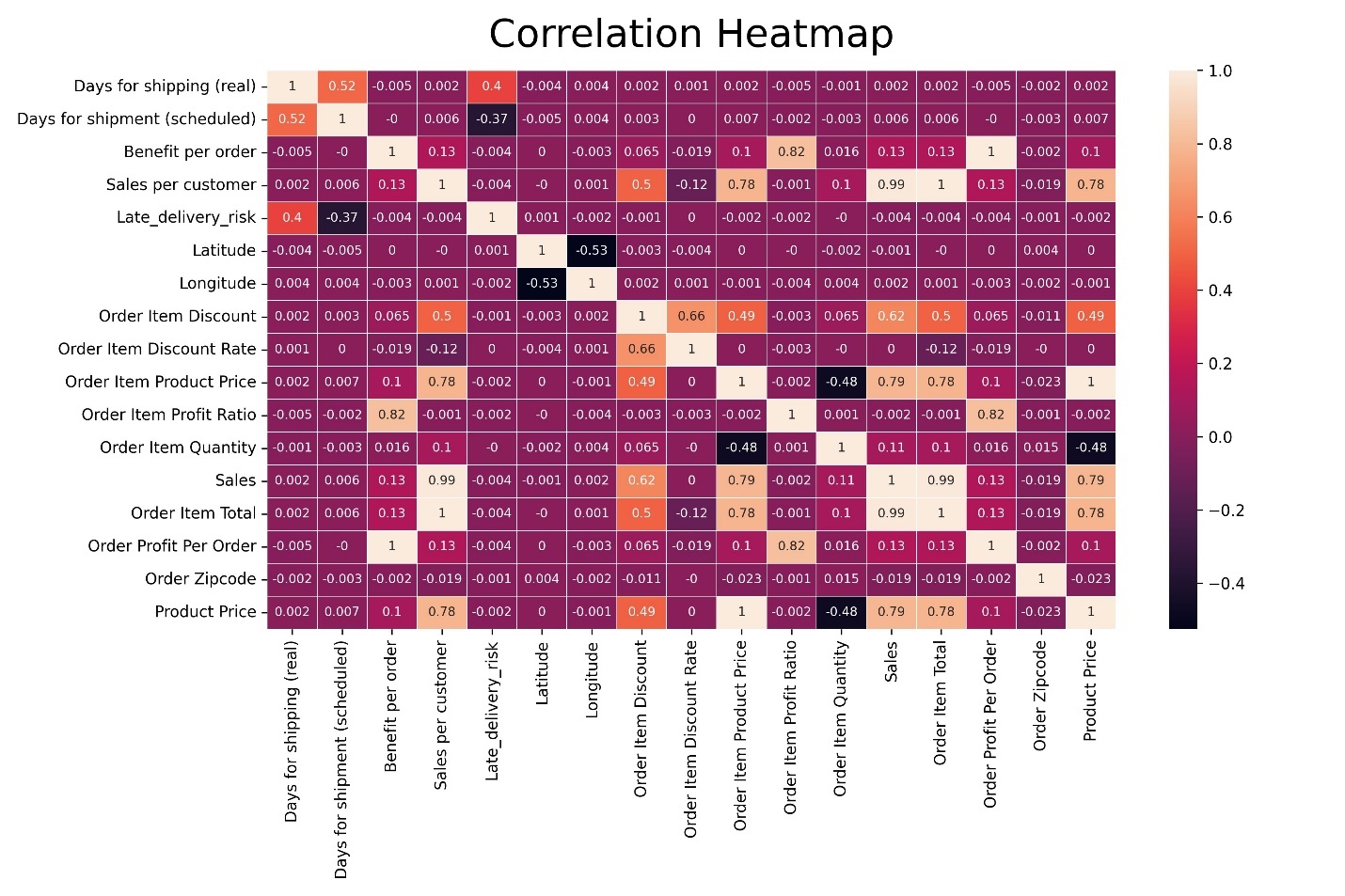
|  |  |
| --- | --- |
| **Attribute** | **Unique Values** |
| Customer Password | 1 |
| Customer Email | 1 |

All attributes that have less than 2 unique values can be dropped from the dataset as it offers no information for our models. So, we delete them.

**After reviewing attributes, we dropped irrelevant columns:**

'Category Id', 'Customer Fname', 'Customer Id', 'Customer Lname', 'Department Id', 'Order Customer Id', 'Order Id', 'Order Item Cardprod Id', 'Order Item Id', 'Product Card Id', 'Product Category Id', 'Product Image', ‘Customer Country’, ‘Customer State’, ‘Customer Street’

**Checking correlation heatmap between attributes to check whether different attributes give the same or similar information in the dataset:**



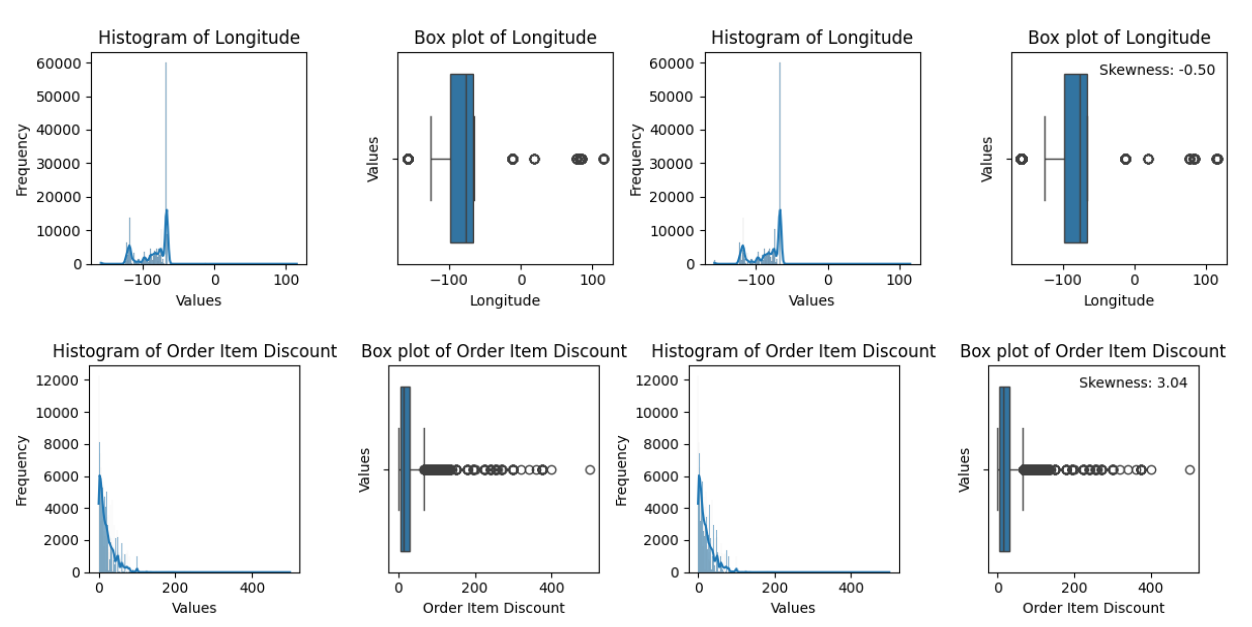
There are some highly correlated attributes in dataset so, I dropped attributes with correlation > 0.85.

|  |  |  |
| --- | --- | --- |
| **Attribute1** | **Attribute2** | **Correlation** |
| Sales per customer | Order Item Total | 1 |
| Order Item Product Price | Product Price | 1 |
| Benefit per order | Order Profit Per Order | 1 |
| Sales | Order Item Total | 0.989744 |
| Sales per customer | Sales | 0.989744 |

Handling categorical and numerical columns with large no of missing values and filing them.

**Checking for outliers and imputation:**

We used box plot and histogram to find how the data is distributed and outliers.

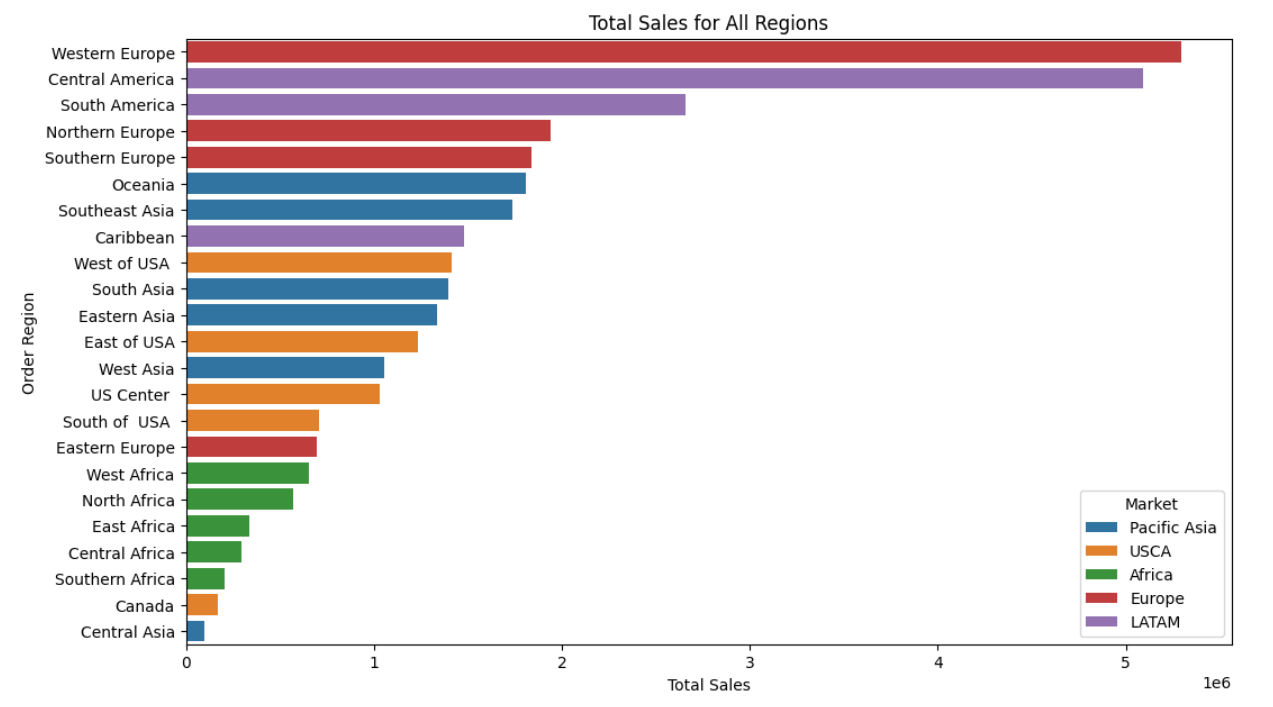


**Outliers were handled for the columns using Winsorize technique:**

For "Order Item Discount", "Order Item Profit Ratio", "Order Item Quantity", "Order Item Total",

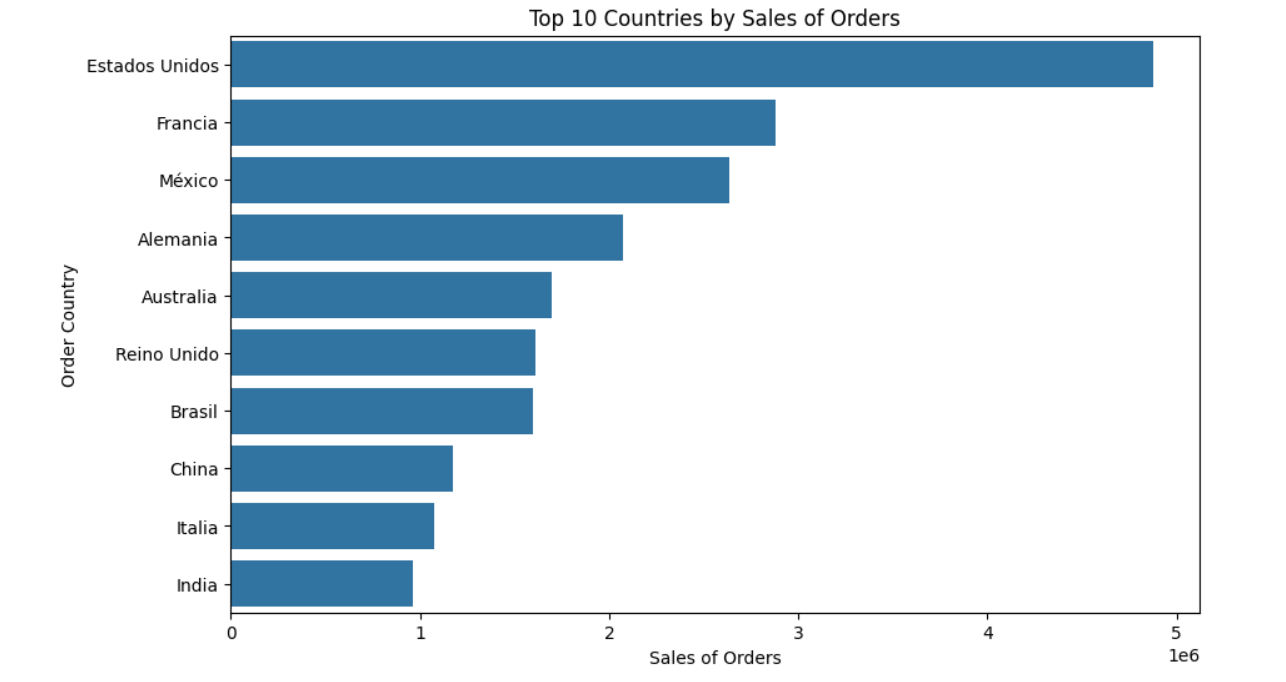
"Order Profit Per Order", "Order Zipcode", "Product Price" columns.

1. **Data Visualization, Storytelling & Experimenting with charts:**
2. **Sales Analysis for All Regions:**

****

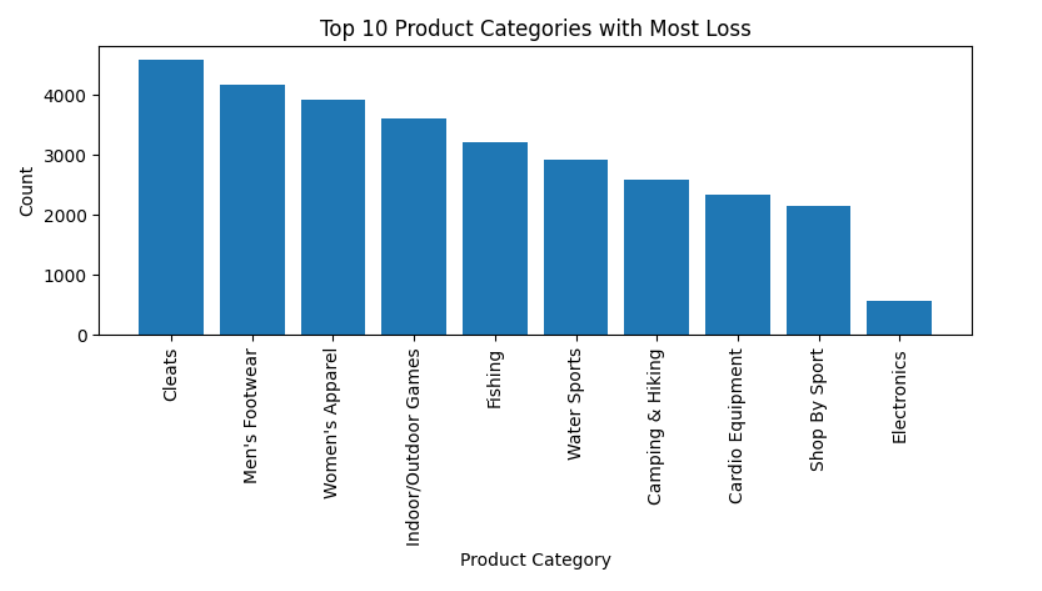
* Western Europe and Central America are the top regions in terms of total sales, indicating these regions are significant markets for the business.

1. **Top 10 Countries by Sales of Orders:**

****

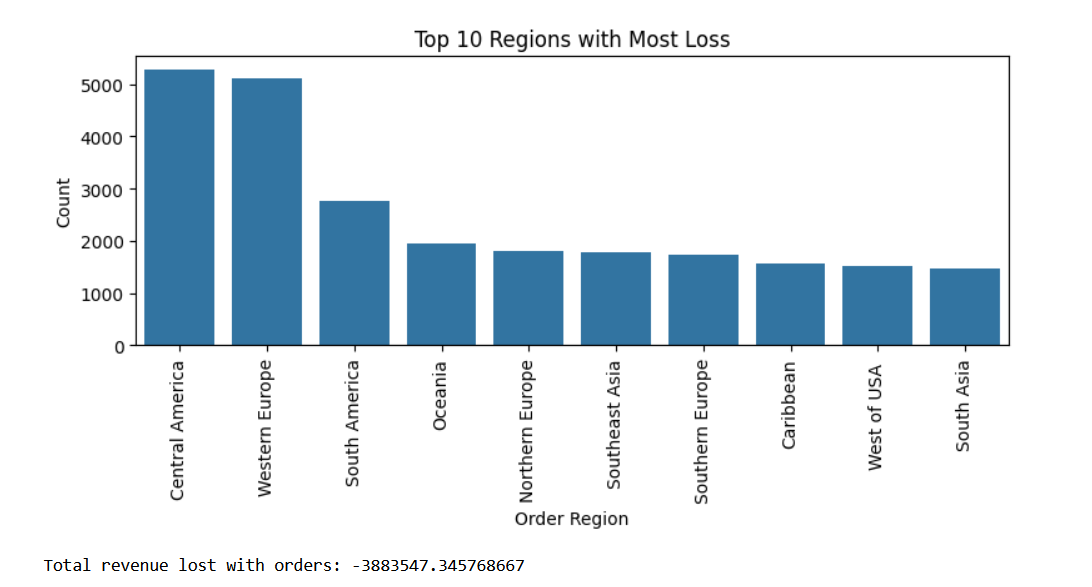
* The Estados Unidos, France, and Mexico are the top countries by sales of orders, suggesting they might have a high customer base or strong market demand.

1. **Top 10 Product Categories with Most Loss:**

****

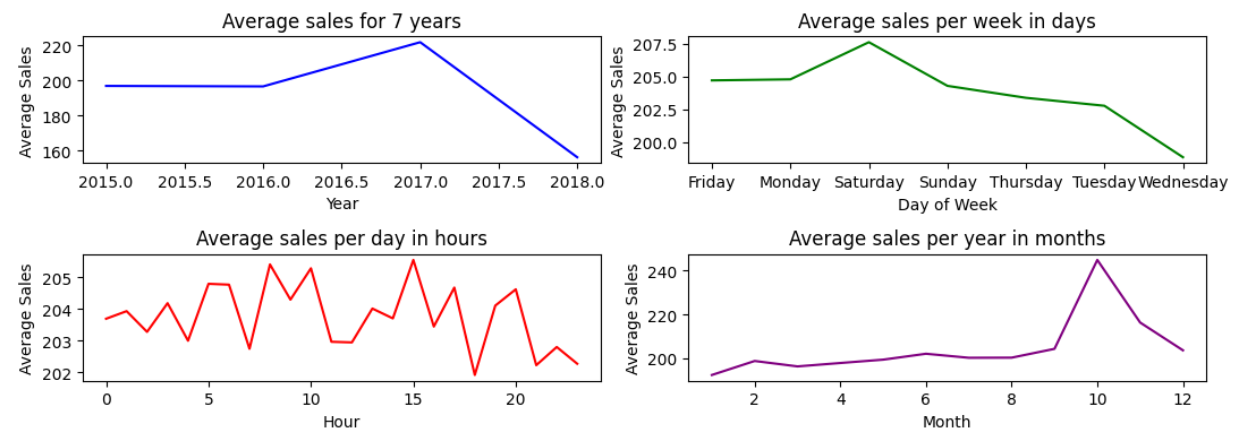
* Cleats are the top product category associated with the most losses.

1. **Top 10 Regions with Most Loss:**

****

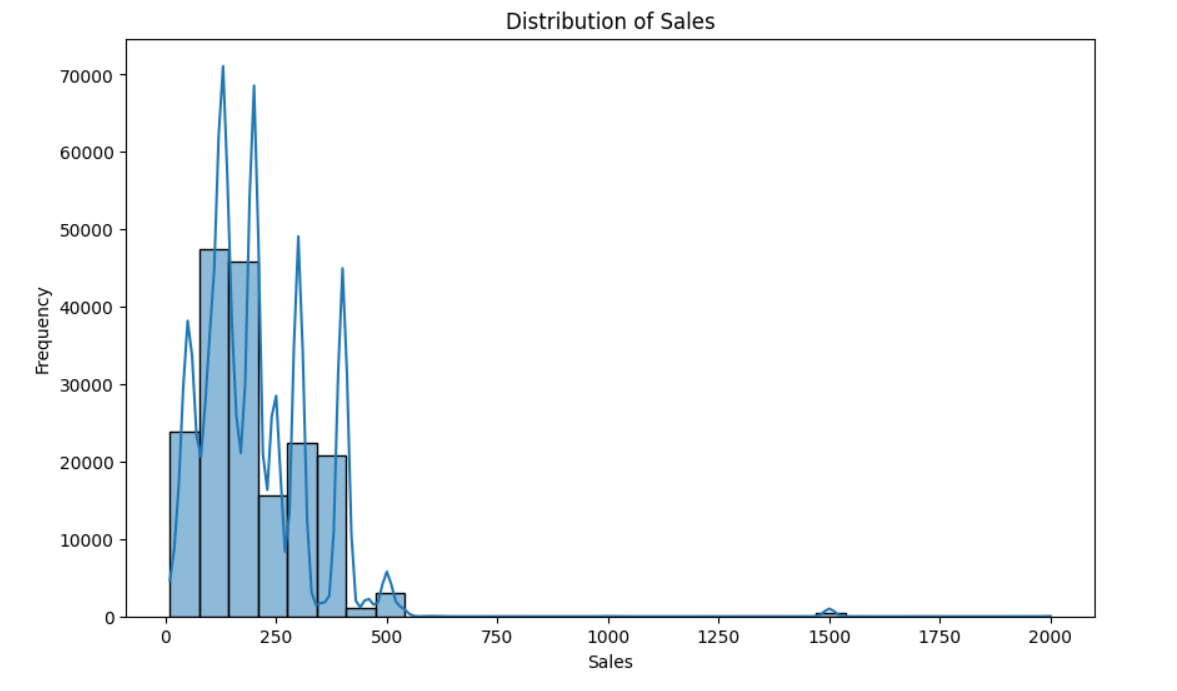
* Central America emerges as the region with the highest count of losses.

1. **What are the trends in average sales over year, month, week and hour:**

****

* Highest number of orders are placed by customers in 2017.
* Saturday recorded highest number of average sales and Wednesday with the least number of sales.
* Average sales per day in hours does not vary much.
* The most number of orders came in October.

1. **Distribution of Sales:**

****

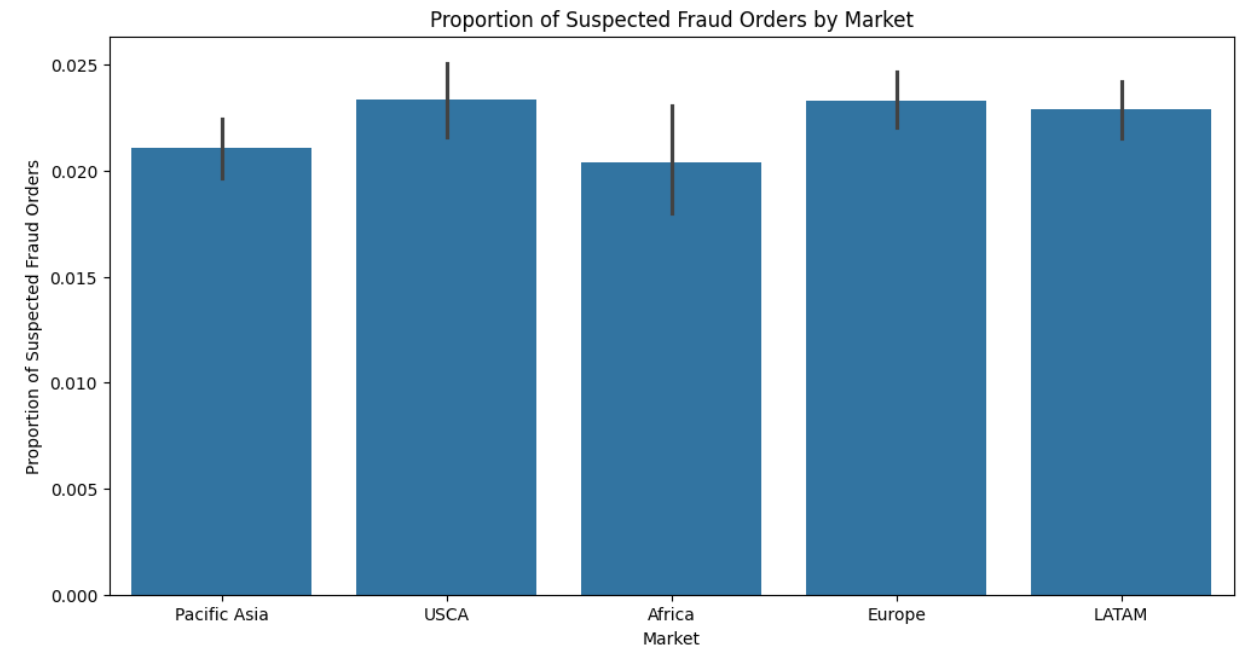
* Product with sales price range of 75-200 are frequently bought.

1. **Suspected Fraud orders per customer:**

****

* Consumers experience the highest number of frauds compared to Corporate and Home Office segments, highlighting potential vulnerabilities in the consumer-facing processes or channels.

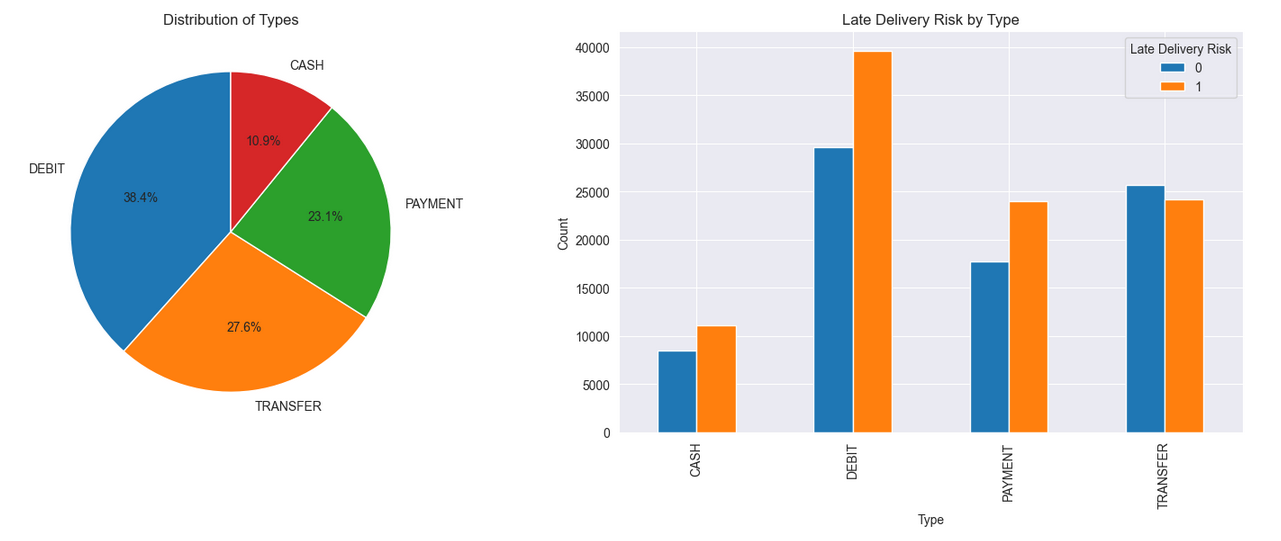
1. **Suspected Fraud Orders per Market:**

****

* Fraud proportions across different markets are relatively consistent, indicating that fraud detection measures might need improvement uniformly across markets.

**3a. Analysis of relevant features with Late Delivery:**

1. **Type v/s Late delivery:**



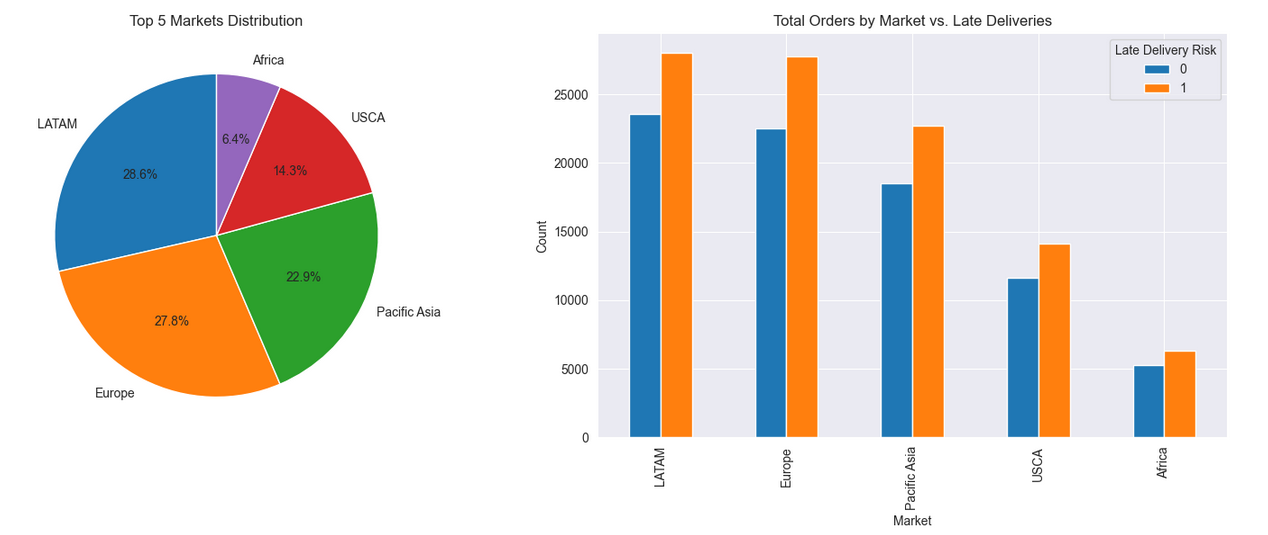
* DEBIT transaction is the most frequently used transaction.
* DEBIT transactions has high risk of late delivery and TRANSFER transactions has minimum risk.

1. **Customer Segment Distribution v/s Late delivery:**



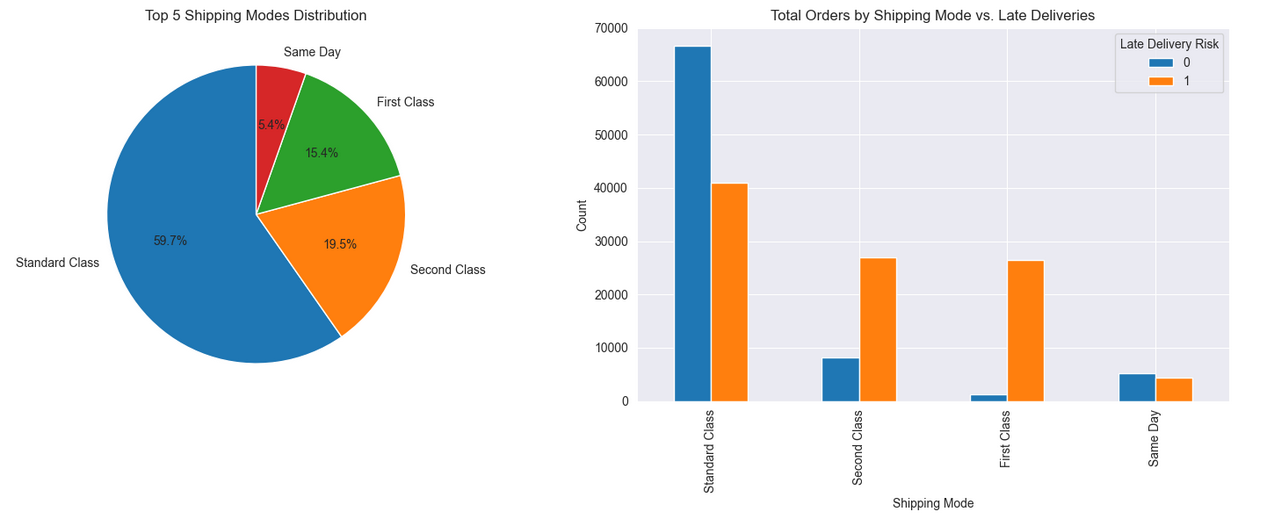
* CONSUMERS represent the largest segment, followed by CORPORATE and HOME OFFICE customers.
* Across all segments, there is a notable proportion of late deliveries, with CONSUMERS experiencing the highest count of late deliveries followed by CORPORATE and HOME OFFICE segments.

1. **Market Distribution v/s Late delivery:**



* LATAM is the largest market, followed by EUROPE, PACIFIC ASIA, USCA, and AFRICA.
* LATAM, EUROPE, and PACIFIC ASIA have the highest counts of late deliveries.

1. **Shipping mode distribution v/s Late delivery:**



* STANDARD CLASS is the most commonly chosen shipping mode, followed by SECOND CLASS, FIRST CLASS, and SAME DAY. Despite its popularity, STANDARD CLASS experiences a significant number of late deliveries.

1. **Category Name v/s Late delivery:**



* CLEATS is the most popular category, followed by MEN'S FOOTWEAR, WOMEN'S APPAREL, INDOOR/OUTDOOR GAMES, and FISHING.
* CLEATS, MEN'S FOOTWEAR, and WOMEN'S APPAREL, despite being popular, experience significant late deliveries.

1. **Order Item Quantities v/s Late delivery:**



* Orders with ORDER ITEM QUANTITY 1 has the most common number of total orders, but it also has a significant number of late deliveries.

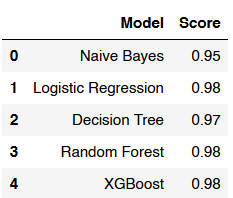
1. **Model:**

Late Delivery Risk being the target. Train-Test split was successfully done.

Scaling was done after split to the X\_train, to avoid data leak.

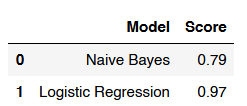
Build model’s and do test its accuracy.

**For Numerical Attributes:**



Logistic Regression, Random Forest, and XGBoost models are the best performers.

**For Categorical Attributes:**



Logistic Regression is the best performer.

**Some potential improvements and areas where the DataCo supply chain dataset may be facing challenges:**

1. **Delivery Time Optimization:**
   * Implementing predictive analytics models to forecast delivery times accurately based on historical data, seasonal trends, and other relevant factors.
   * Streamlining logistics operations and improving route planning to minimize delivery delays.
2. **Customer Segmentation and Targeted Strategies:**
   * Segmenting customers based on order history, preferences, and geographical location to tailor delivery services and communication channels.
   * Designing targeted marketing campaigns and promotions to incentivize timely orders and reduce late deliveries.
3. **Supplier Relationship Management:**
   * Strengthening partnerships with suppliers to ensure timely product deliveries and maintain adequate inventory levels.
   * Implementing performance monitoring systems to track supplier reliability and address any issues proactively.
4. **Inventory Management Optimization:**
   * Utilizing inventory management systems to optimize stock levels, reduce lead times, and prevent stockouts or overstock situations.
   * Implementing demand forecasting techniques to anticipate customer demand and adjust inventory levels accordingly.
5. **Quality Control and Product Inspection:**
   * Enhancing quality control measures and product inspection processes to minimize defective or damaged items, which can lead to delays in fulfillment.
   * Implementing quality assurance programs to ensure product consistency and customer satisfaction.
6. **Technology Integration and Automation:**
   * Leveraging advanced technologies such as IoT devices, RFID tracking, and automated order processing systems to streamline operations and improve efficiency.
   * Integrating supply chain management software to facilitate real-time monitoring, data analytics, and decision-making.
7. **Training and Development Programs:**
   * Providing training and development programs for warehouse staff, delivery personnel, and customer service representatives to enhance their skills and customer service abilities.
   * Fostering a culture of continuous improvement and innovation to adapt to changing market dynamics and customer expectations.
8. **Customer Feedback and Communication:**
   * Soliciting feedback from customers through surveys, reviews, and feedback forms to identify areas for improvement and address any issues promptly.
   * Improving communication channels with customers to provide updates on order status, delivery delays, and resolution of any issues or concerns.

Accurate forecasting of late deliveries is pivotal for effective supply chain management and maximizing profitability. By successfully predicting delays, a company can proactively take measures to prevent sales reductions, gaining a competitive edge over rivals. Furthermore, employing machine learning for late delivery prediction not only benefits the company but also has a positive impact on the environment. Implementing a late-delay forecasting model can significantly reduce the number of missed deliveries, thereby minimizing carbon dioxide emissions and fostering environmentally sustainable practices.

By prioritizing these improvement areas, DataCo can optimize its supply chain operations, shorten delivery lead times, mitigate late deliveries, and ultimately enhance customer satisfaction and loyalty. Continuous monitoring, analysis, and adaptation to evolving market conditions are vital for sustaining success and competitiveness in the industry.

## References[¶](https://www.kaggle.com/code/gelarerouzbahani/data-analysis-for-supply-chain#References)

1- Dataset link : <https://data.mendeley.com/datasets/8gx2fvg2k6/5>

2- Various Projects <https://github.com/>

3- Various Projects <https://kaggle.com/>

4- How Data Analytics Can Help Improve Your Supply Chain. Available at: <https://www.linkedin.com/pulse/how-data-analytics-can-help-improve-your-supply-chain-dfreight/?trk=public_post>