





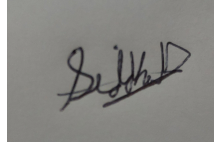
Analyzing Supply Chain Data with PySpark

- Team 8 Members: Harshit Kavathia, Shaunak Dhande, Siddharth Shankar

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	Signature
1	Harshit Kavathia	
2	Shaunak Dhande	

3	Siddharth Shankar	
---	-------------------	---

I. Executive Summary

Our analysis of supply chain data offers valuable insights into various aspects of supply chain operations, enabling informed decision-making for optimization and improvement. Through a comprehensive examination of the dataset, we identified top-selling product categories, average shipping times by different modes, percentages of late deliveries, and visualized delivery statuses for source-destination pairs.

By leveraging these insights, businesses can optimize logistics operations, mitigate delivery risks, and enhance customer satisfaction, ultimately driving competitiveness and success in the supply chain domain. Our classification model further aids in strategic decision-making to reduce late delivery risks, ensuring smoother and more efficient supply chain operations.

II. Data Description (1 page)

DataCo Global has provided a dataset of supply chain activities for analysis. This dataset is a valuable resource for applying machine learning algorithms and analysis. The dataset encompasses various areas of important registered activities within the supply chain, including provisioning, production, sales, and commercial distribution. Moreover, it enables the correlation of structured data with unstructured data for knowledge generation, providing a comprehensive understanding of supply chain operations. The dataset used in our supply chain optimization project contains essential information on transactions and orders within the supply chain network. It comprises 180,520 rows and 53 columns, providing a comprehensive view of various aspects involved in the supply chain process.

Data Field 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 indicating if shipping is late (1) or not (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's first name.
14. **Customer Id:** Customer ID.

15. **Customer Lname:** Customer's last name.
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's Zipcode.
21. **Department Id:** Department code of the store.
22. **Department Name:** Department name of the store.
23. **Latitude:** Latitude corresponding to the location of the store.
24. **Longitude:** Longitude corresponding to the location of the 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.
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: 1 if not available, 0 if the product is available.
51. **Shipping date (DateOrders):** Exact date and time of shipment.
52. **Shipping Mode: Shipping modes:** Standard Class, First Class, Second Class, Same Day.

The dataset provides a comprehensive view of various parameters associated with the supply chain process, including customer details, order specifics, product information, and shipping logistics. This rich dataset will serve as the foundation for our supply chain optimization project, enabling us to analyze and optimize various aspects of the supply chain network efficiently.

III. Research Questions

We aimed to first understand and analyse the given supply chain data to understand the problems or improvements the system needs which can be solved by using data.

1. Analysis and Problem Identification:
 - What are top selling categories in products
 - Average shipping times taken by different modes
 - Percentage of Late deliveries

- Identifying different delivery statuses for source-destination
 - Identify growing market regions for different products
2. Predicting future demand for product orders based on historical data
 3. How to identify and avoid late delivery for shipments?

IV. Methodology (1 page)

1. Analysis

a. Finding the top-selling categories using Pyspark Graphframes

The analysis of top-selling categories within the supply chain dataset of DataCo Global involves a structured methodology aimed at uncovering valuable insights for decision-making processes. Beginning with data preparation, relevant columns including 'Category Name' and 'Sales' are selected to form a new data frame, enabling a focused analysis on sales performance across different product categories. This initial step ensures that only essential data is retained for further processing, streamlining the analysis process.

Subsequently, a graph is constructed using the GraphFrames library, facilitating efficient graph-based computations. Vertices are generated based on distinct 'Category Name' values, representing individual categories within the supply chain. Edges are then created to establish relationships between categories and their corresponding sales values, forming a comprehensive graph structure. This graph serves as the foundation for aggregating and analyzing sales data coherently.

Aggregation and sorting operations are performed to compute the total sales for each category, with the aggregated data sorted in descending order of sales volume. This allows for the identification of the top-selling categories, with the top 10 categories selected for further analysis. The resulting insights provide stakeholders with valuable information regarding the most lucrative product categories within the supply chain, guiding strategic decisions related to inventory management, marketing strategies, and product development.

b. Average shipping times taken by different modes

Analyzing average shipping times across various shipping modes within the supply chain dataset of DataCo Global offers valuable insights into the efficiency and performance of delivery operations. This analysis employs a structured methodology designed to extract meaningful insights from the dataset. Initially, pertinent columns such as 'Shipping Mode', 'Delivery Status', and 'Days for shipping (real)' are selected to form a focused DataFrame, ensuring that only relevant data related to shipping modes and delivery timelines is included for analysis.

Subsequently, a graph is constructed using the GraphFrames library, facilitating the representation of relationships between shipping modes, delivery statuses, and corresponding shipping durations. This graph enables the visualization of shipping modes as nodes and delivery statuses as edges, with the duration of shipping serving as the weight of each edge. Through this graph-based approach, the average shipping time for each shipping mode is calculated by aggregating and averaging the 'Days for shipping (real)' values associated with each mode.

The visualization of average shipping times per shipping mode is presented using a bar plot, where the average shipping time (in days) is depicted on the y-axis, and different shipping modes are represented on the x-axis. This visualization provides stakeholders with a clear and

concise overview of the performance trends across different shipping modes. For instance, shipping modes with shorter average shipping times, such as 'Same Day', indicate efficient delivery processes, while longer average shipping times in modes like 'Second Class' and 'Standard Class' may signify areas for improvement.

c. Percentage of Late deliveries for every shipping mode

Analyzing the percentage of late deliveries by shipping mode within the supply chain dataset of DataCo Global involves a systematic approach to understand the timeliness of order fulfillment across different modes. This methodology unveils insights into the effectiveness of each shipping mode in meeting delivery deadlines and identifies potential areas for improvement. By focusing on shipping modes with higher percentages of late deliveries, businesses can implement targeted measures to enhance logistics efficiency, streamline operations, and mitigate delays, ultimately improving customer satisfaction and loyalty.

The analysis begins by calculating the overall percentage of late deliveries across all orders and shipping modes within the dataset. This calculation is performed by counting the total number of late deliveries and dividing it by the total number of orders, multiplied by 100 to obtain a percentage. This provides a baseline understanding of the prevalence of late deliveries in the supply chain network.

Subsequently, a graph is constructed using the GraphFrames library to represent relationships between shipping modes, delivery statuses, and the occurrence of late deliveries. Vertices are created based on distinct shipping modes, while edges represent relationships between shipping modes and delivery statuses, considering the late delivery risk as an attribute. This graph-based approach facilitates the aggregation of late delivery percentages for each shipping mode, providing insights into the timeliness of order fulfillment across different modes.

The visualization of late delivery percentages per shipping mode is presented using a bar plot, where each shipping mode is plotted on the x-axis, and the corresponding percentage of late deliveries is depicted on the y-axis.

d. Identifying different delivery statuses for source-destination

The methodology for identifying different delivery statuses for source-destination pairs, as depicted in the visualization provided, involves a systematic approach to extract and analyze pertinent data from the supply chain dataset. The aim is to gain insights into the relationships between source cities (customer cities) and destination cities (order cities), with the delivery status serving as the connecting relation.

Firstly, the supply chain dataset is loaded into a DataFrame, with relevant filters applied to narrow down the dataset to specific product categories. In this case, we focus on sporting goods, ensuring that the analysis is targeted and relevant to the context. Additionally, to manage computational complexity and facilitate visualization, a sample subset of the dataset is selected, limiting the analysis to a manageable size while still retaining representative data.

Next, distinct nodes representing customer cities and order cities are extracted from the filtered dataset. These nodes serve as the endpoints of the relationships that will be analyzed. By selecting distinct values for both customer cities and order cities, we ensure that each unique city is represented as a node in the graph.

Subsequently, edges are defined based on the relationships between source and destination cities, with the delivery status serving as the connecting relation. This information is extracted directly from the dataset and used to establish connections between source-destination pairs.

Each edge represents a specific delivery status associated with a particular source-destination pair, providing valuable insights into the distribution of delivery statuses across different city pairs.

Once the nodes and edges are defined, a graph is constructed using the GraphFrames library, which enables efficient graph-based computations. This graph represents the relationships between customer cities, order cities, and delivery statuses, forming a comprehensive network that visualizes the flow of products through the supply chain.

e. Sales vs Order Item quantity

The relationship between sales and order item quantity reveals potential insights into customer behavior and purchasing patterns. Understanding this correlation can aid businesses in optimizing inventory levels, identifying cross-selling opportunities, and evaluating the effectiveness of promotions. Higher order item quantities can impact production and operational costs. Businesses need to balance the benefits of increased sales with the costs associated with producing or acquiring larger quantities of items. A sudden increase in order quantity without a corresponding increase in sales could indicate a successful promotion or a new product launch.

In summary, the relationship between sales and order item quantity is a multifaceted aspect of business analysis which needs to be explored by both product and delivery companies as it is crucial to optimizing the supply chain process.

2. Product demand forecasting using Ensemble Machine learning

Demand forecasting enables businesses to foresee future demand patterns for their products or services based on a range of factors, such as historical data on sales, market trends, and customer behavior. Machine learning models can incorporate such fluctuations to forecast the demand for products, making it a valuable tool for strategic planning.

Our methodology searches through the large dataset for important features that could potentially contribute to Sports product sales and order rates over time. This could include product specific features such as sales, order-profit ratio, product category rankings but also people-centric factors such as regions with interest in certain products at specific times of the year (indicated by shipping dates)

Using ensemble modeling we account for all our model parameters when predicting the Product order trends for the future. Demand forecasting often deals with dynamic and evolving market conditions and hyperparameter tuning enables the model to adapt to changing trends and patterns in the data, ensuring that forecasts remain accurate and reliable over time.

The model predictions can then be used to identify expected spikes/downturns in order quantities for various products which can help both the delivery teams and parent companies come up with a robust shipping strategy ahead of time

3. Predicting the delivery status based on certain factors

As the global supply chain is getting more and more complex, there are multiple ways to deliver a single shipment. It is very much likely that a shipment is delivered wrongly or late. It becomes supremely important for businesses to deliver the shipments more efficiently. In our analysis, we want to make a provision to avoid late deliveries for businesses. We want to do it by using a classification machine learning model. We have trained three models namely, Logistic Regression, Decision Tree, and Random Forest Classifier. We have identified this as a classification problem because our target variable is categorical and we somehow wanted to predict the delivery status based on certain criteria.

We are using classification models to predict the delivery status of a given order. By this analysis, we can frame our shipping mode such that our shipment is not delivered late. To predict Delivery Status we are using the following parameters.

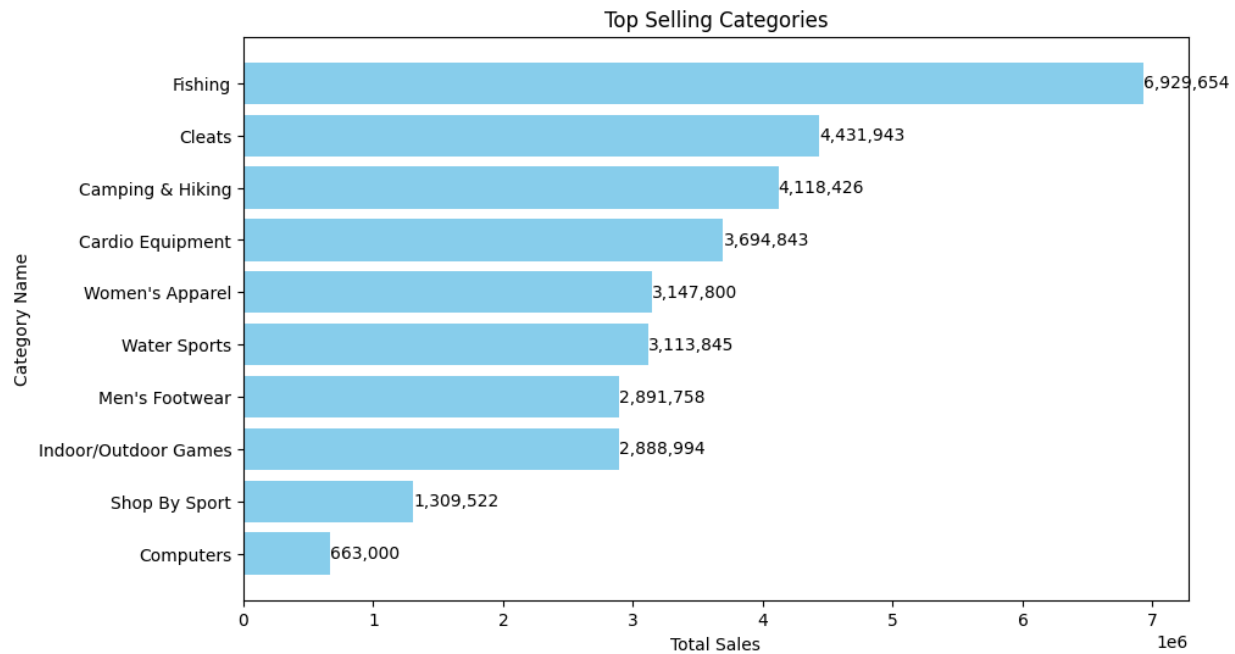
1. Market - Market to where the order is delivered: Africa, Europe, LATAM, Pacific Asia, USCA.
2. Order City - Destination city of the order
3. Order Country - Destination country of the order
4. Category Name - Product category
5. Shipping Mode - The mode by which the shipping is done: Standard Class, First Class, Second Class, Same Day
6. Latitude - Latitude corresponding to the location of the store
7. Longitude - Longitude corresponding to the location of the store
8. Order Item Quantity - Number of products per order

The idea behind this modeling is to facilitate the decision-making at the supplier's end to choose the perfect shipping mode and shop to supply the product on time. If somehow for a given source, destination, and shipping mode the delivery status comes out to be late then maybe we can change the source or shipping mode to make sure the shipment is delivered on time.

V. Results and Finding

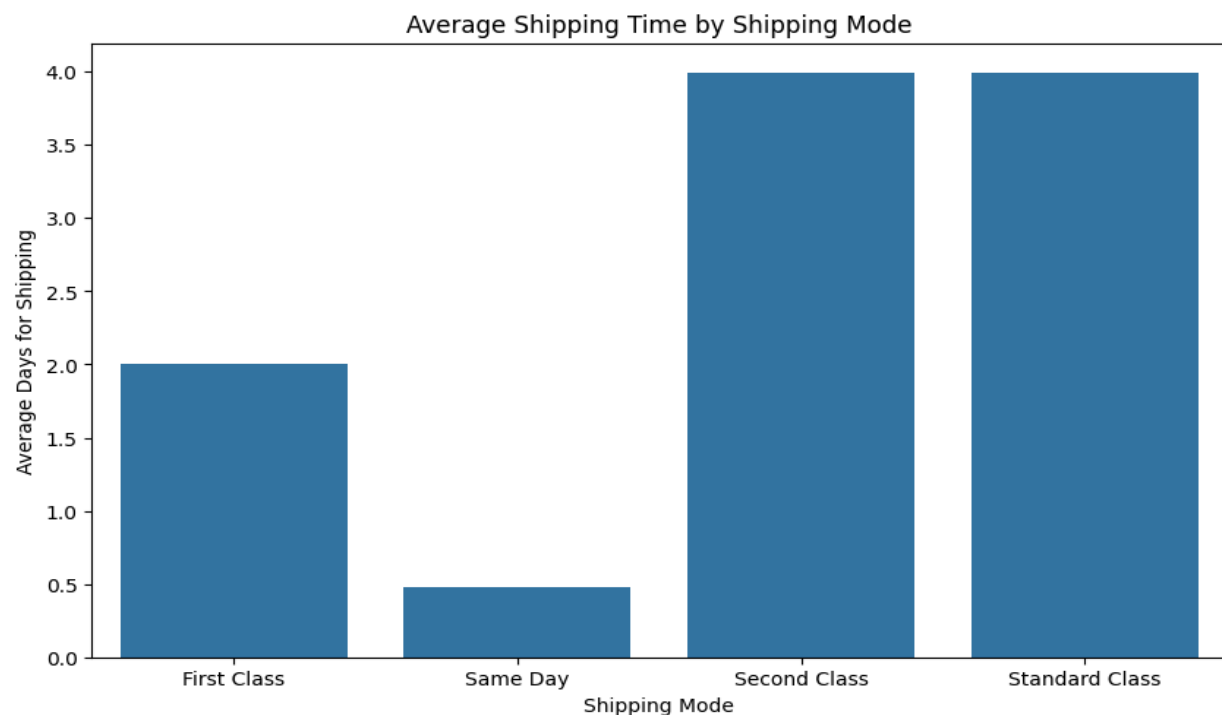
1. What are top selling categories in products

The analysis of top-selling categories within the supply chain dataset reveals compelling insights into product sales performance. From the visualization generated, it is evident that Fishing, Cleats, and Camping & Hiking emerge as the top three categories with the highest sales volume. This finding underscores the significance of these categories in driving revenue and underscores their importance in inventory management, marketing strategies, and product development initiatives. By leveraging this insight, businesses can optimize their inventory levels, tailor marketing campaigns to capitalize on popular categories, and refine product offerings to meet customer demand effectively, ultimately driving profitability and enhancing competitiveness in the market.



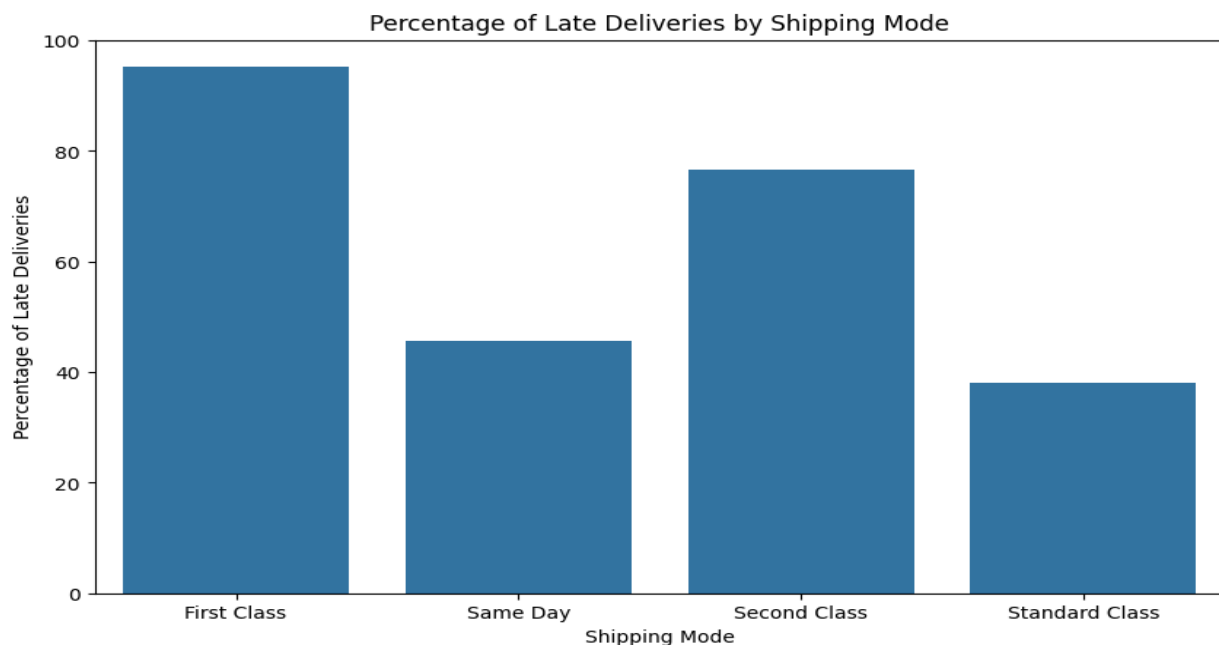
2. Average shipping times taken by different modes

The analysis of average shipping times by different modes within the supply chain dataset reveals notable performance variations across shipping modes. The visualization illustrates that the 'same day' mode exhibits the shortest average shipping time, indicating efficient delivery processes. Conversely, 'Second' and 'Standard' class modes demonstrate longer average shipping times, suggesting potential areas for improvement in delivery timelines. These findings underscore the importance of optimizing logistics strategies tailored to each shipping mode to enhance overall performance and customer satisfaction. By addressing delays and ensuring timely deliveries, businesses can optimize supply chain operations, improving efficiency, and enhancing the overall customer experience.



3. Percentage of Late deliveries for every shipment mode

The analysis of late deliveries by shipping mode provides valuable insights into the timeliness of order fulfilment across various modes within the supply chain. From the visualization generated, it is observed that certain shipping modes exhibit higher percentages of late deliveries compared to others. An interesting observation is that First class shipment mode has the least average days for shipping but at the same time has the highest percentage of late deliveries. This finding underscores the need for targeted interventions to enhance logistics efficiency and mitigate delays in specific shipping modes. By identifying areas for improvement, businesses can implement strategic measures to optimize delivery performance and ensure more reliable and timely order fulfilment processes. Ultimately, these efforts aim to enhance customer satisfaction and loyalty by providing a seamless and dependable delivery experience across all shipping modes.



4. Expected increased demand for product orders in certain regions:

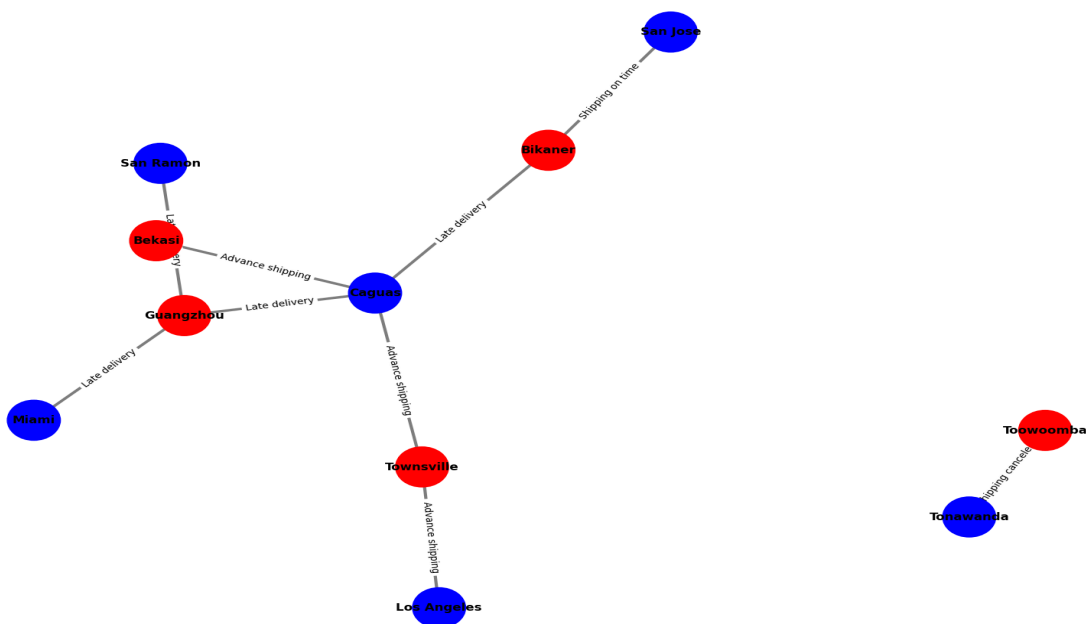
The predictive analysis of Product order trends based on the historical data gives us keen insight into the changing demand trends in various regions. The Sales - Order Item quantity is expected to increase for all Sports goods in particular for Soccer, Golf and Rugby. On average the Latin/South American and Australian market have a fairly high order count which is expected to increase in the coming months. This could potentially be due to the fact that Soccer and Rugby are very popular sports in these two regions.



5. Visualizing different delivery statuses for source-destination

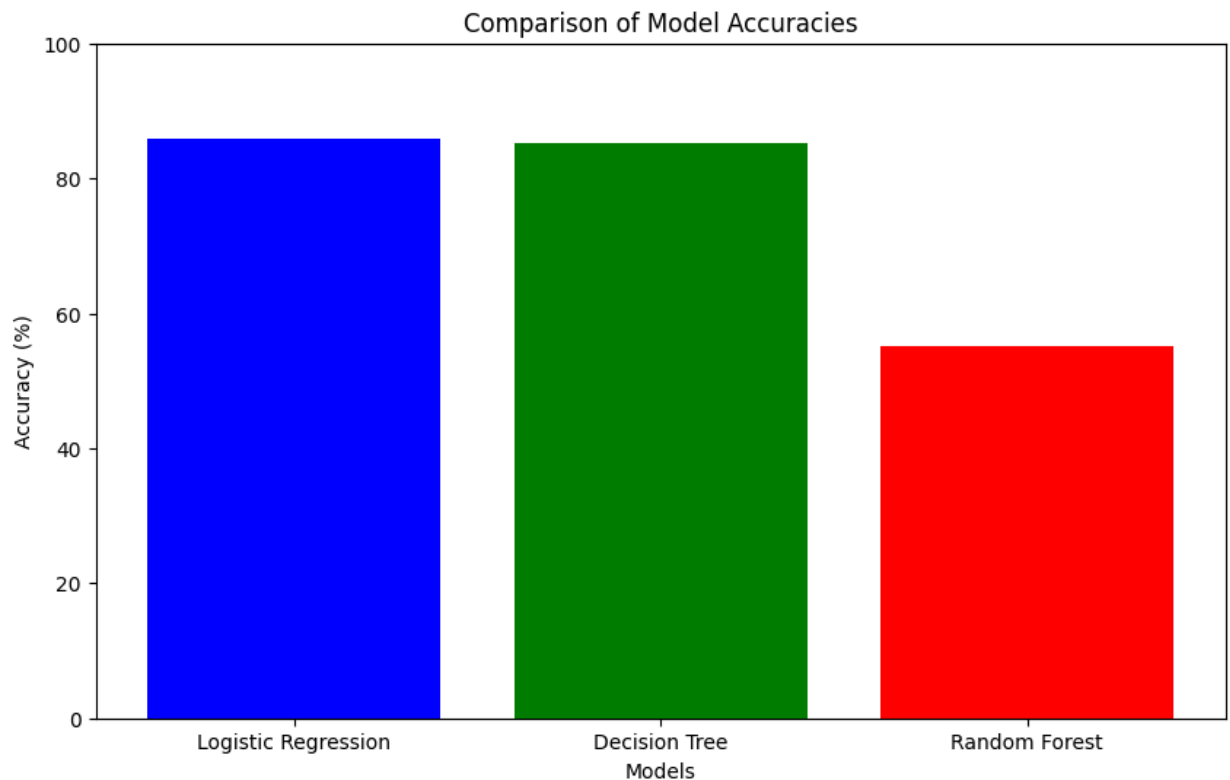
The visualization illustrates the source and destination cities of products within the supply chain, with the delivery status serving as the connecting relationship. Nodes representing customer cities are denoted in blue, while nodes representing order cities are depicted in red. Through this graph visualization, the connections between different cities and their associated delivery statuses are visualized, offering insights into the flow of products and their journey from origin to destination. Additionally, the edge labels provide information about the delivery status for each source-destination pair, enabling stakeholders to identify patterns, trends, and potential bottlenecks in the delivery process. This visualization facilitates a comprehensive understanding of the supply chain dynamics, allowing businesses to optimize logistics operations, address delivery challenges, and enhance overall efficiency in order fulfillment processes.

Graph Visualization



6. Classification model accuracies

We have trained the three classification models namely logistic regression, decision tree and random forest classifier. We found that logistic regression gives the maximum accuracy amongst all three models.



VI. Conclusion

In the initial phase of the project, we conducted an in-depth analysis to identify key insights and problem areas within the supply chain. Firstly, we determined the top-selling categories in products, revealing Fishing, Cleats, and Camping & Hiking as the top performers. This insight offers valuable guidance for inventory management, marketing strategies, and product development initiatives.

Next, Demand forecasting based on sales and region/market information provided us with insights for targeted marketing strategies. By accurately predicting future demand trends, businesses can adjust production levels, allocate resources efficiently, and tailor delivery efforts to specific regions. This strategic approach not only improves operational efficiency but also strengthens competitive advantage and drives sustainable growth.

Additionally, we examined the average shipping times taken by different modes, shedding light on delivery efficiency across various shipping methods. Furthermore, we analyzed the percentage of late deliveries, highlighting areas for improvement in meeting delivery deadlines. Lastly, we visualized different delivery statuses for source-destination pairs, providing insights into the flow of products and potential bottlenecks in the delivery process. These analyses collectively lay the groundwork for further investigation and strategic decision-making aimed at optimizing logistics operations, improving customer satisfaction, and driving overall business success in the supply chain domain. Our classification model aids strategic decision-making to reduce late delivery risk.