Overload Ware Labs AI

Data Analysis

Kaggle dataset - E Commerce shipping Data

PDF – Report

Supply Chain - Shipping Data Analysis

Name	Srinivas Senthil Kumar
Degree	MSc Supply Chain Management
Position	Data Analyst Intern
City	Cologne. Germany
Tools	Python, Pandas, Seaborn
Contact	srinivas982902@gmail.com

Submitted on 08.08.2025

Abstract

This project analyses the shipping data of an international e-commerce shipping company to improve operational in efficiencies and customer experience challenge with the help of supply chain analytics. With the help of Python with its Pandas and Seaborn libraries, important relationships among shipping class, customer satisfaction, and product importance handling as well as shipping and delivery delays were identified. This research concluded that more than 40% percent of deliveries were late and this was particularly true for the “Ship” mode and certain warehouses such as Warehouse F. Suggested strategies focus on improving the accuracy of customer delivery expectations, bettering product importance handling, auditing warehouse operations, and incorporating automated systems to improve customer satisfaction and logistics and supply chain efficiency.

Table of Contents

1. Introduction

In the modern e-commerce environment, the retention of a customer hinges on their experience with deliveries. This analysis seeks to determine the delivery performance of an international e-commerce shipping operation using a real-world dataset. The aim is to establish strategies to mitigate the reasons for shipping delays and analyze the effect of shipping methods, warehouses, discounting, and product importance on shipping performance. This project seeks to enhance the logistics team by providing decision support using data analytics and visualization tools, this project aims to support decision-making that drives efficiency and improves customer service outcomes.

1.1 Problem Statement

The issue which affects many companies today, even with rising demand and increased spending in logistics, a major part of e-commerce orders still face delays, resulting in customer dissatisfaction and internal operational inefficiencies. Limited visibility into the root cause underlying in shipping modes, warehouse activities, product prioritization, and evaluation of underlying shipping modes further complicate the problem. Such systems which usually depend on input of data by personnel, and which are updated maybe monthly, do not offer real-time information. One major drawback of this nature of data is that real-time information cannot be produced, which limits the decision-making process by presenting managers as well as stakeholders with data that may be out of date, and therefore business decisions may not reflect the current needs of the market (Mohamed, 2024). In today's world where the expectation of customers is high, and business decisions are made based on time and accuracy of information, delay or inaccurate information on inventory may lead to complaints from clients and inefficient business operations because wrong decisions would have been made due to lack of information (Mohamed, 2024). These inefficiencies do not only intermit the actual flow of the supply chain but also increases operation cost, and hence the company's profitability and competitiveness.

Companies that lack a holistic understanding of warehouse operations may end up making baseless assumptions instead of informed decisions. Consequently, there is a need to assess shipping data to find the gaps and inefficiencies that undermine delivery performance and customer satisfaction. These capabilities are especially important in industries characterised by high demand variability and in which timely decisions are needed to sustain competitive advantage. In this way, the study will provide important findings to companies to

uncover patterns and find the inefficiencies that impact the supply chain process and to improve the area that affected and helps in faster decision making.

1.2 Research Objectives

This study aims to:

- To identify the extent and frequency of delivery delays across the e-commerce supply chain.
- To assess the effect of the mode of transportation on the shipping and delivery schedules.
- To evaluate the impact of discount strategies on customer satisfaction and delivery outcomes.
- To assess whether high-priority products are given preferential treatment in logistics operations.
- To detect underperforming warehouses contributing to delays.
- To explore the relationship between delivery reliability and customer satisfaction metrics (ratings, queries).
- To provide actionable recommendations for optimizing delivery strategies, improving logistics operations, and enhancing customer experience.

1.3 Research Questions:

- What percentage of shipments is delayed, and which factors affect the most to these delays?
- How do different shipping modes (Road, Ship, and Flight) impact in the delivery?
- Is there a relationship between discount levels and delivery status?
- Do high-priority or high-importance products receive faster and more reliable deliveries?
- Which warehouses are associated with higher delivery delays, and why?
- How does delivery status correlate with customer ratings and customer service interactions?
- Are operational inefficiencies (like warehouse delays or poor routing) linked to lower customer satisfaction?

2. Overview

Data Source: Kaggle: https://www.kaggle.com/datasets/prachi13/customer-analytics

Dataset name: Train.csv

Context:

An international e-commerce company based wants to discover key insights from their customer database; it seems they want to use some of the most advanced machine learning techniques to study their customers and to find and improve the inefficient areas. Data described as follows:

2.1 Data Description

Data Description

ID: ID Number of Customers.

Warehouse block: The Company have big Warehouse which is divided in to block such as A, B, C, D, E.

Mode of shipment: The Company Ships the products in multiple way such as Ship, Flight and Road.

Customer care calls: The number of calls made from enquiry for enquiry of the shipment.

Customer rating: The Company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best).

Cost of the product: Cost of the Product in US Dollars.

Prior purchases: The Number of Prior Purchase.

Product importance: The Company has categorized the product in the various parameters such as low, medium, high.

Gender: Male and Female.

Discount offered: Discount offered on that specific product.

Weight in Gms: It is the weight in grams.

Reached on time: It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.

2.2 Project Goal

An International E commerce shipping company wants to discover key business insights and to use the advanced Machine learning techniques to study their customer database.

The goal of this project is to analyse shipping and delivery performance data from an international e-commerce company in order to uncover operational inefficiencies and customer service gaps across the supply chain.

By exploring key factors such as shipping mode, warehouse performance, product prioritization, and customer feedback, this analysis aims to identify the root causes of delivery delays and their impact on customer satisfaction.

The insights derived will support data-driven decision-making to optimize logistics operations, improve delivery reliability, and enhance the overall customer experience — critical outcomes for any supply chain-focused business.

In short our main objective is on identifying patterns in shipping delays and customer behaviour to help logistics teams optimize delivery strategies and improve customer satisfaction.

3 Data Preparation

In any real-world analytics project, data is messy, so it's crucial to clean it before analysis or visualization to avoid flawed conclusions and also which affects the reliability of insights. This involves handling missing values, converting formats, removing duplicates or outliers, and creating new variables if necessary.

3.1 Data Cleaning

Data cleaning is necessary in finding and fixing errors, inconsistencies, or missing values in the dataset to ensure accuracy and reliability. In this project, I have performed these following steps:

- Looked into the shape of data how big it is

- Getting the info of the data

- Checked for rows with missing or invalid delivery status and customer ratings.

- Standardized categorical fields like shipping mode, delivery status, and warehouse labels.

- Checked for and handled outliers in numerical columns like Discount Offered and Customer Care Calls.

- Ensured the correct data types for each column, such as converting ratings to integers.

Clean data is crucial to avoid misleading insights and to ensure that visualizations and statistical summaries accurately reflect reality.

3.2 Data Mining

Data mining involves finding patterns and relationships within large datasets using statistical, mathematical, or machine learning techniques. For this analysis:

- Grouped data by shipping mode, warehouse, and product importance to uncover hidden trends in delays.

- Identified relationships between delivery status and other factors like customer ratings, discount offered, and number of customer care calls.

- To detect underperforming warehouses and modes through aggregations and comparisons.

Data mining goes beyond surface-level patterns and reveals deeper operational inefficiencies or trends in customer behavior.

3.3 Data Visualization

Data visualization changes raw data into visual formats like charts and graphs to make patterns, trends, and outliers easier to interpret. In this project:

- Bar charts were used to compare delay rates across shipping modes and warehouses.

- Box plots showed how discounts and delivery status are related.

- Heatmaps and correlation plots highlighted relationships between customer feedback and delivery timelines.

Visualizations help both analysts and stakeholders quickly understand complex patterns and make informed decisions.

3.2.1 Heatmap

Heatmap of the data for checking the correlation between the features and target column to check which numerical features influence whether a shipment is on time or delayed, and to understand the relationships between features.

This supports:

Feature selection for predictive modelling

Operational decision-making (e.g., which causes delays)

Simplifying analysis by identifying unimportant variables

Fig: 1 Correlation heatmap

Range:

- +1.0 = strong positive correlation
- = no correlation
- –1.0 = strong negative correlation

Key Insights from Heatmap

- Discount Offered have high positive correlation with Reached on Time or Not of 40%.
- Weights in gram have negative correlation with Reached on Time or Not -27%.
- Discount Offered and weights in grams have negative correlation -38%.
- Customer care calls and weights in grams have negative correlation -28%.
- Customer care calls and cost of the product have positive correlation of 32%.
- Prior Purchases and Customer care calls have slightly positive correlation

4. Exploratory Data Analysis (EDA)

EDA is a process of analysing data visually and statistically to identify patterns, outliers, and relationships and it aids in identifying inefficiencies in supply chains, such as delays across shipping modes, product importance and discounts, and warehouse performance.

5. Visual Insights & Key Findings

Analysing the given data

5.1 On-Time vs. Delayed Deliveries

Question1: What % of shipments is delayed?

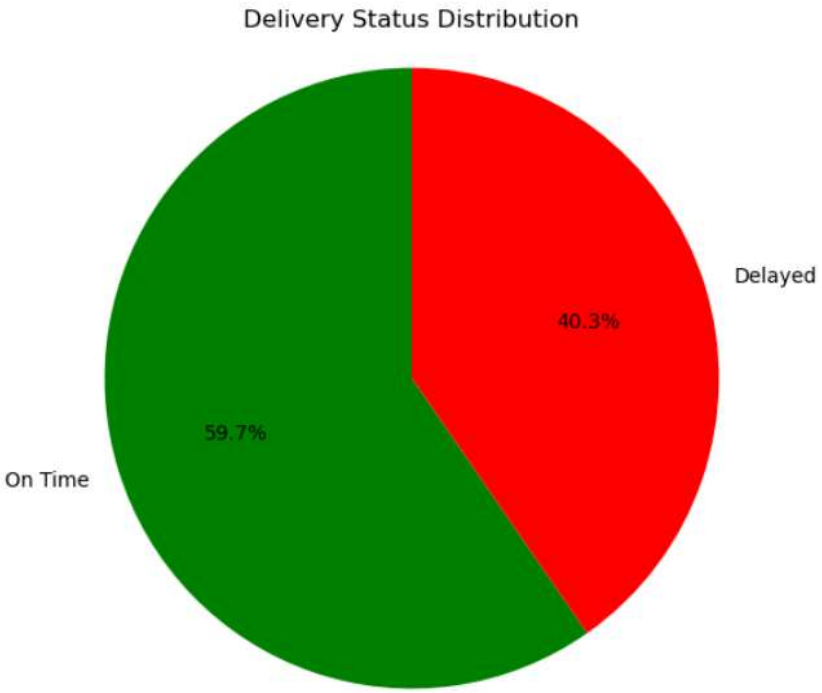

Fig 2: Delivery status distribution

Business Insights:

As we can see in the above pie chart for delivery status distribution over 40% of orders are delayed and only 59.7% products which were on time.

Approximately 4 out of 10 deliveries are late which is a high impacted customers and resulted in customer dissatisfaction.

Recommendation:

Improve delivery processes with better route planning or better mode of transport and suggested to use real-time tracking systems if applicable.

5.2 Shipping Mode Performance

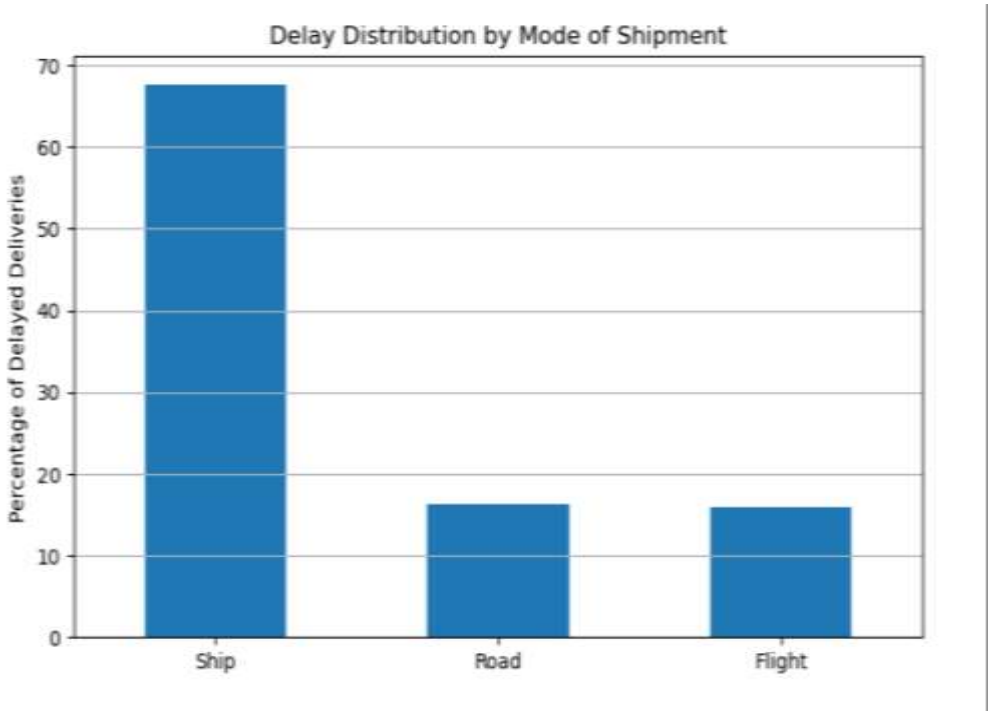

Fig 3: Mode of Shipment distribution

Business Insights:

As we can see in the above bar chart that ship mode accounts for 67% of all shipments which has the highest delay rate compared to road and flight mode.

Since ship is the least predictable and slowest mode of transport is likely responsible for most delays which causes bottlenecks to the company.

Recommendation:

Inspect partnerships with faster couriers (Air freight) for critical regions or products.

5.3 Discounts vs. Delivery Performance

Fig 4: Discount vs. Delivery performance

Business Insights:

As we can see in the above bar graph that discounts didn't cause delays in fact on-time orders had bigger discounts.

On more thing that higher discounts are often offered when products are delayed.

Discounting maybe used as reactive compensation - This suggests that the business might be giving discounts to recover from poor delivery performance, not to improve or prevent it.

Recommendation:

Instead of relying on discounts, I recommend to fix the root causes of delays.

5.4 Product Importance vs. Delivery Timeliness

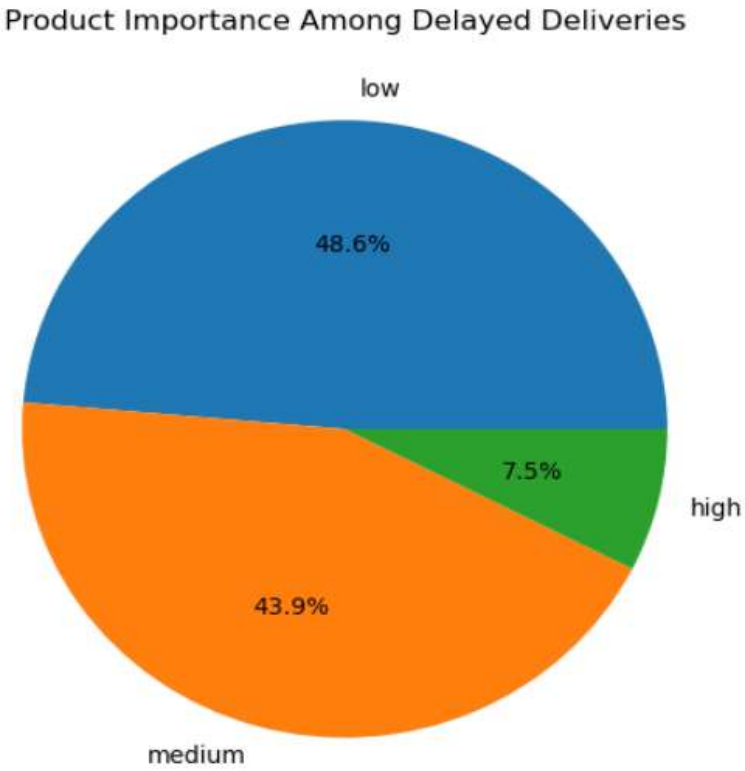

Fig: 5 Product importance vs. Delayed deliveries

Business Insights:

As we can see in the above pie chart that delivery rates are similar across all product importance levels.

Even among delayed orders, nearly 49% are low-importance, and only 7.5% are high-importance.

Recommendation:

Shows some priority handling which is already happening in the supply chain, but still 7.5% delays for high-importance items exist

Implement priority queues or Service Level Agreement tiers — a formalized commitment

Focus on service quality and delivery experience across all tiers

5.5 Warehouse Performance

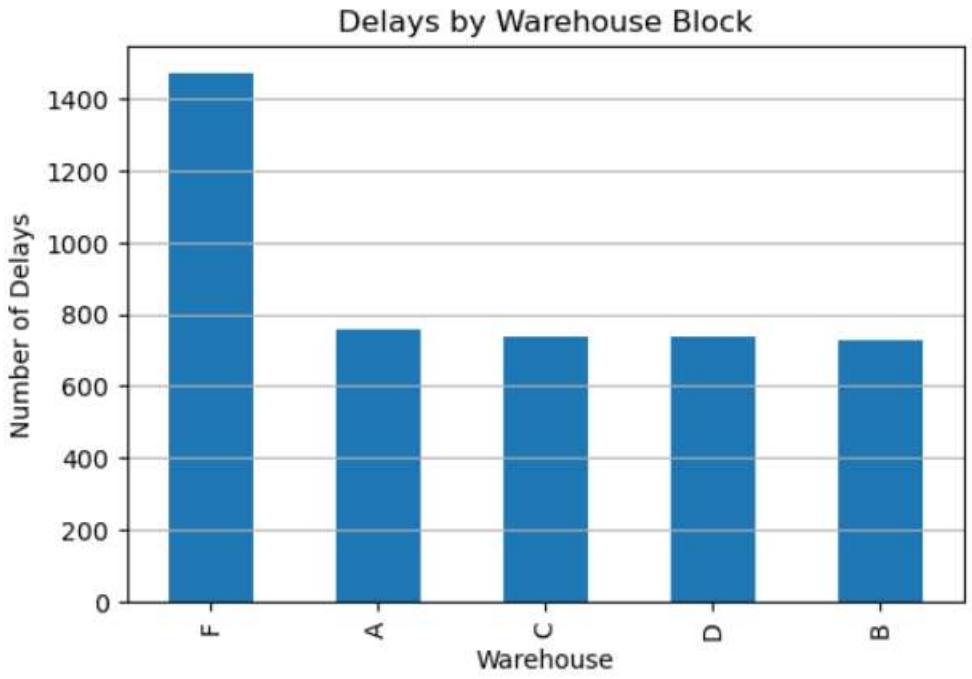

Fig: 6 Warehouse Block

Business Insights:

The above bar graph shows that warehouse F shows the highest number of late deliveries, followed by A and C.

Operational issues or capacity overload in specific warehouses may lead to delays or potential staffing or layout issues.

Recommendation:

Immediate audit performance at Warehouse F (staffing, layout, dispatch speed).

Balance order volume across warehouses or reallocate resources.

Assess staffing levels and equipment functionality.

Review order processing workflows and pick paths.

Evaluate carrier pickup schedules and work with the order loads.

5.6 Customer Rating vs. Delivery Status

Fig: 7 Customers Rating vs. Delivery Status

Business insights:

Above graph shows that higher customer ratings (4-5 stars) correlate with on-time deliveries, while lower ratings (1-2 stars) show slightly fewer on-time deliveries.

While delivery timeliness influences satisfaction, the correlation is weaker than expected, suggested other factors also significantly impact customer ratings.

Recommendation:

To improve overall satisfaction, delivery performance must be addressed alongside product quality, customer support, and pricing considerations.

5.6.1 Product Importance vs. On-Time Delivery

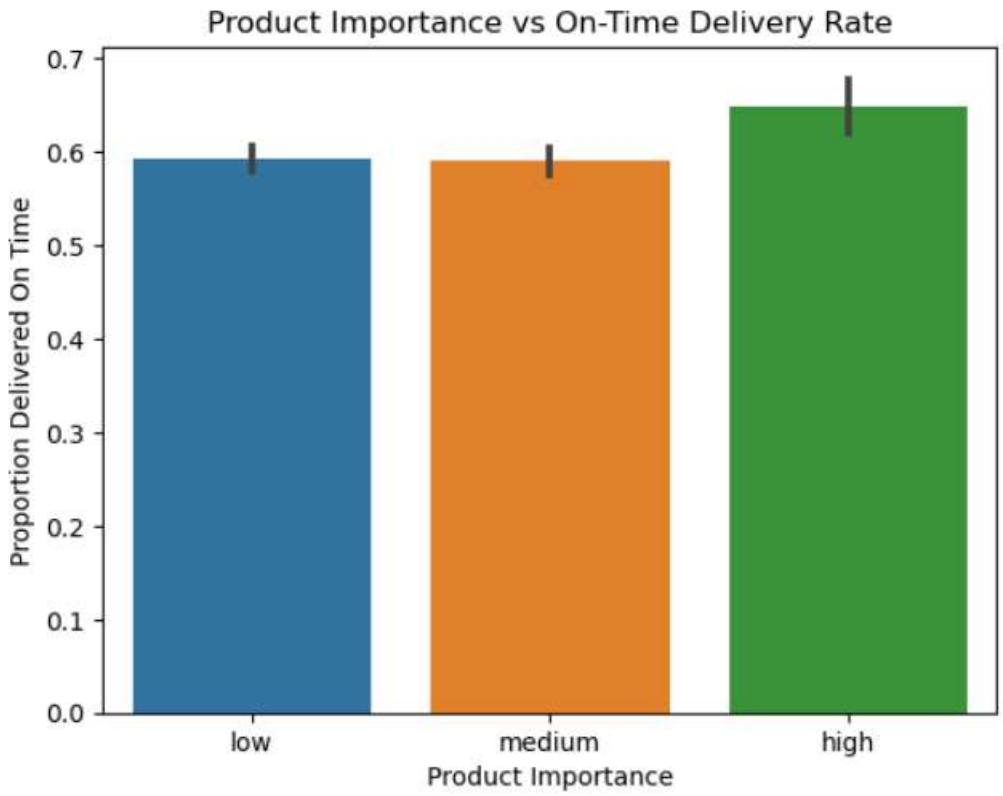

Fig: 8 Product importance vs. On-Time Delivery

Business Insight:

High-priority products do not consistently receive higher ratings since we can see that low importance products as well as receive good ratings

Companies shouldn't assume high-priority products yield higher customer satisfaction — they must still ensure quality and service for all items since customers expect top service for all products, not just important ones.

Recommendation:

I recommend ensuring quality across the board, regardless of product priority and Implement true prioritization in the supply chain to ensure high-importance products receive preferential handling and expedited shipping.

5.6.2 Product Importance vs. Customer Rating

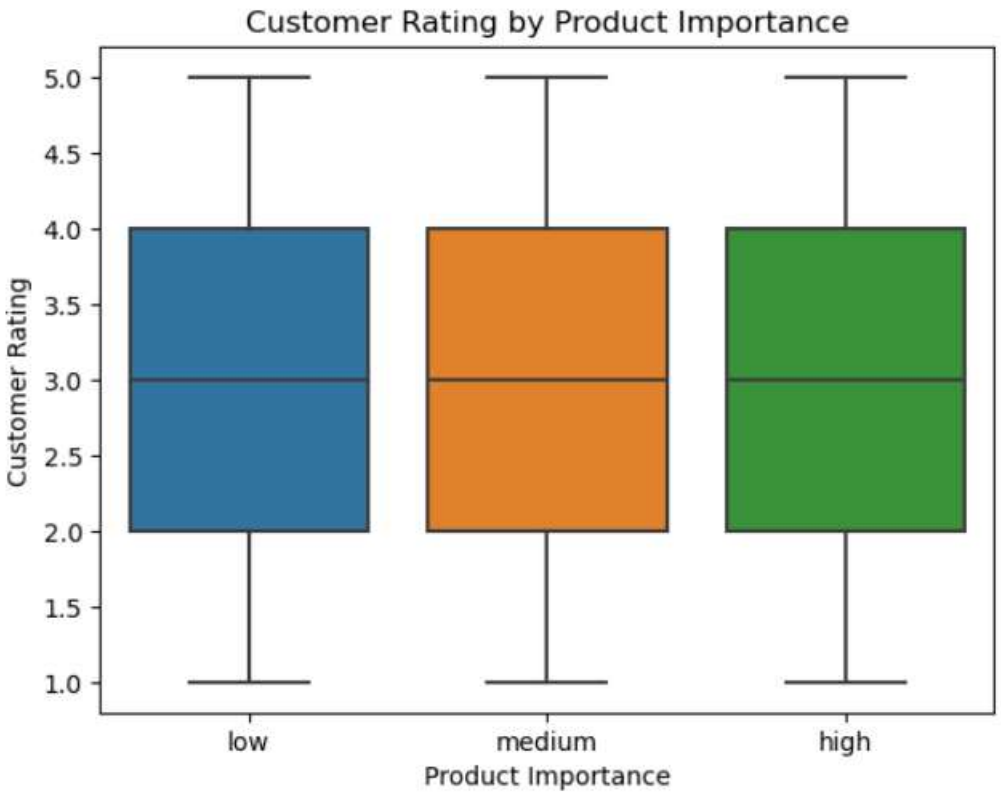

Fig: 9 Product importance vs. Customer Rating

Business Insights:

In the above graph product Importance does not strongly influence customer rating as all products are handled equally regardless of high and low importance.

Customers expect consistent service, regardless of how the product is labelled internally by the company. If a high-importance product is delayed or handled poorly, they may actually rate it lower, because expectations are higher.

Recommendation:

One recommendation here is improve service quality and reliability across all product importance levels, not just high-priority items. Implement consistent standards in packaging, handling, and communication between the stakeholders.

5.7 Customer Service Calls vs. Delivery

Fig: 10 Customer Service Calls vs. Delivery

Business Insights:

Delayed shipments generate a little wider range and higher average number of customer service calls compared to on-time deliveries which has increased support team workload

Higher customer service costs

Potential for escalations and complaints

Resource diversion from other service areas

Hidden cost of delays

Recommendation:

Each percentage point improvement in on-time delivery can reduce customer service load and associated costs.

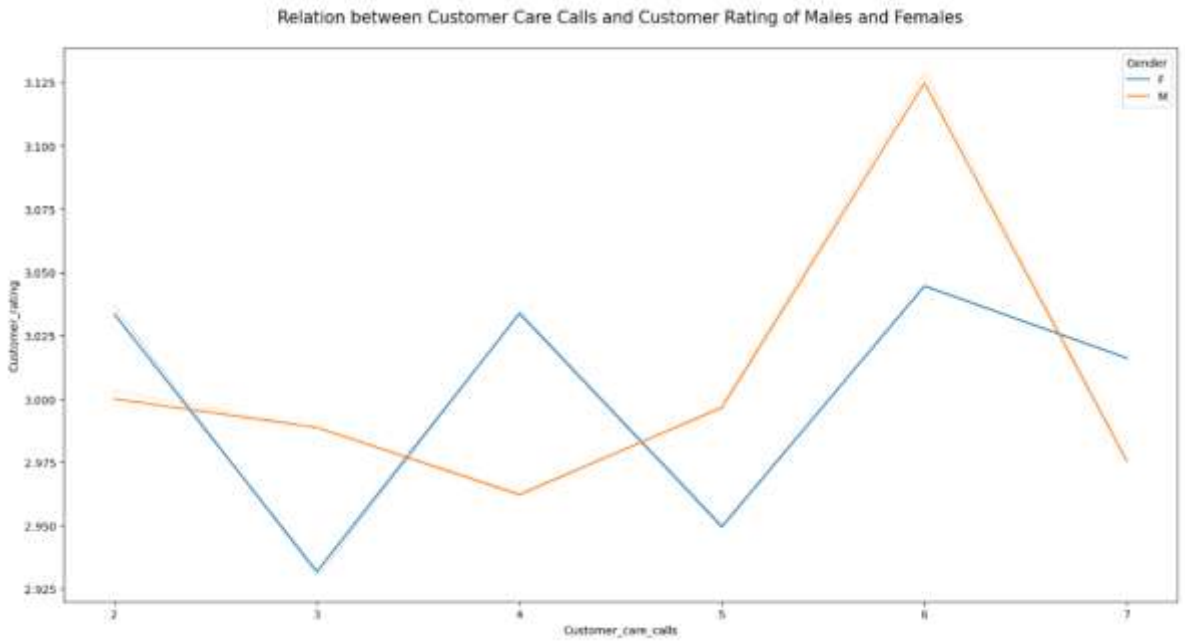

Fig: 11 Customer care calls vs. Customer rating

Business insights:

As we can see in the above graph that lower customer ratings received more customer calls maybe the product has an issue in terms of delivery or other.

And male customers may be more likely to call support even at similar satisfaction levels

High call volume can strain support teams.

Gender segmentation can help in targeted communication and training support staff.

Recommendation:

Reduce the delays to reduce calls and workload

Consider personalized responses based on user behavior — males may want more frequent updates or reassurance.

6. Strategic Recommendations:

- **Switch mode of transport from Ship to Air/Road for Critical Routes:**
  Evaluate air freight partnerships for high-value or time-sensitive products.

- **Audit and Upgrade Warehouse F:** Investigate staffing, layout, and process bottlenecks in high-delay warehouses.

- **Implement Real-Time Tracking and Route Optimization:** Use predictive analytics and logistics tools to proactively address delays.

- **Introduce Tiered Service Levels:** Formalize Service Level Agreements (SLAs) to ensure better handling of high-priority shipments.

- **Rebalance Customer Support Resources:**
  Link support demand forecasts to delivery performance data to improve resource planning.

- **Improve Cross-Department Coordination:**
  Align operations, logistics, customer care, and discounting strategies for a cohesive delivery experience.

7. Limitations & Assumptions

- Lack of Time-Series Data:
  Shipment trends over time were not explored in depth due to lack of timestamps or daily logs.

- Limited Context on External Factors:
  Traffic, weather, and regional disruptions were not included but may significantly influence delivery times.

- Assumption of Causation from Correlation:
  Observations are correlational —predictive ML models could offer deeper insights.

- Single Dataset Scope:
  Data from only one company limits generalization; insights may not fully translate across industries or geographies

8. Conclusion

- Through this analysis highlights key operational and customer experience challenges in the shipping lifecycle of an e-commerce business.

- Despite offering high-priority labels or discounts, much systemic inefficiency still lead to delivery delays — notably from shipping mode choices and underperforming warehouses.

- Addressing these issues with targeted operational changes, customer-centric strategies, and improved data-driven logistics can significantly improve delivery reliability and customer satisfaction.

- By publishing this analysis, I aim to showcase how data analysis can uncover impactful business insights from real-world operations and guide strategic decisions across supply chain networks.

Bibliography/ List of References

https://www.researchgate.net/publication/366020486_E-commerce_Data_Analysis_and_Visualization_using_Random_Forest_Regression_and_Prophet_model

https://www.researchgate.net/publication/394167965_Shipping_Route_Optimization_and_Risk_Management_based_on_Big_Data_Analysis

https://www.researchgate.net/publication/390169930_E-commerce_and_Shipping_Preferences_-_The_Impact_of_E-Commerce_on_Student_Preferences_for_Shipping_and_Delivery_Services

List of Figures

Declaration of Own Work

I hereby declare, that the work for this project was solely undertaken by me and that no help was provided from other sources than those allowed.

Cologne, Germany 08.08.2025 Signature