All Chapter_ Liew Yng Jeng.pdf

by Liew Yng Jeng MCS241006

Submission date: 17-Jan-2025 10:10PM (UTC-0800)
Submission ID: 2566428522
File name: All_Chapter__Liew_Yng_Jeng.pdf (4.2M)
Word count: 9942
Character count: 54707

DELAY PREDICTION ON INVENTORY SHORTAGES IN SPORTS EQUIPMENT SUPPLY CHAIN

LIEW YNG JENG

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 1

INTRODUCTION

1.1 Introduction

In modern supply chain management, many calculation and prediction methods have long appeared in the market to accurately predict the quantity demand of regional supply and the expected arrival time of shipment. However, there are still various delayed deliveries and inventory shortages in the market, which lead to supply chain paralysis. Most enterprises rely on traditional methods to forecast according to the assumptions and historical data, but the efficiency of obtaining prediction results is reduced, and the cost increases (Sani, S et al., 2023). Regarding the aspect of the capital chain, there will be daunting impacts since small and medium-sized enterprises cannot compete with large enterprises. Hence, this project will focus on the delayed prediction of inventory shortage in the sports equipment supply chain issue, and use the sports equipment datasets to minimize the possibility of out-of-stock, overstock, and delayed shipment (Abouloifa & Bahaj, 2022).

1.2 Background of Study

Global cargo supply has experienced tremendous challenges, among which shipment delays and insufficient inventory still need to be resolved. There is a great correlation between accurately predicting the required inventory to ensure the profitability of enterprises and the satisfaction of consumers who expect to receive the stuff. Usually, the expected delivery time was delayed due to suppliers, transportation companies, data management, and other factors. On the supplier side, the most common factors are machine failures and raw material shortages leading to production delays and insufficient inventory (Gabellini et al., 2022). Regarding the transportation factors, there might be issues of mismatch between the number of

loaded goods and the number of transport vehicles, improper route planning, and severe weather conditions (Sani et al., 2023). Human factors generally cause data management, such as slow manual processing of orders and inability to ship due to information asymmetry. Beyond a doubt, these problems significantly impacted suppliers and consumers, and people would gradually lose confidence in this field if they continued (Abouloifa & Bahaj, 2022).

1.3 Statement of Problem

From the perspective of the current technologically advanced world, it is impossible to further achieve high-precision accuracy by continuing to rely on traditional methods when real-time and seasonal changes need to be considered. The accuracy of an enterprise's inventory forecast is equivalent to the enterprise's profit (Abouloifa & Bahaj, 2022). Likewise, delivery punctually is equivalent to a promise from the consumers' perspective. Since current forecasting methods may not be sufficient to face all the challenges in this highly competitive market landscape, the accuracy of forecasts is crucial and related to the company's reputation and development prospects. It can ensure that the enterprise's inventory is sufficient and not excessive while ensuring that the goods are delivered to customers on time, which improves the enterprise's profits and customer satisfaction, creating a win-win situation (Sani et al., 2023; Gabellini et al., 2022).

1.4 Aim of the Study

This project aims to provide an accessible, affordable, and effective forecasting analysis method to predict market trends and demand by studying various calculation methods to correspond to different supply and demand, help enterprises optimize inventory management, minimize out-of-stock and overstock, accurately calculate the arrival time of goods, and promote data-driven decision-making.

1.5 Research Goal

The research goal will describe the objectives and questions of the Delay Prediction on Inventory Shortages in the Sports Equipment Supply Chain project.

1.5.1 Research Objectives

The objectives of the research are :

(a) To identify historical sales and inventory data of enterprises and pre-process them to handle missing values, outliers, and seasonality

(b) To derive complex forecast results into actionable insights using Tableau and Power BI

(c) To evaluate the performance of various forecasting models using ARIMA and XGBoost, SARIMA, and LSTM based on forecast accuracy, reliability, and applicability

1.5.2 Research Questions

The questions of the research are :

(a) What are the primary factors contributing to sports equipment supply chain delays?

(b) How does the performance of hybrid models like ARIMA + XGBoost compare to traditional time-series models like SARIMA in predicting inventory shortages?

(c) Can deep learning models, such as LSTM, effectively capture long-term dependencies in supply chain data to better predict delays?

1.6 Hypothesis

This project combines machine learning models (ARIMA + XGBoost) to improve the accuracy of delay prediction for inventory shortages compared to traditional models (SARIMA) in the sports equipment supply chain.

1.7 Assumption

Based on the theoretical perspective of this project, assume the following:

(a) Inventory shortages or overstock can be predicted in advance to determine the length of delays, which can confirm consumers' confidence in the platform

(b) Overloaded transportation, supplier reliability, and unstable weather factors bring uncertain variables, which require various forecasts to prevent the worst from happening

(c) The demand for sports equipment has a specific seasonal peak, which can be confirmed by calculating data through forecasting models

1.8 Significance of the Study

The following are some of the contributions that this study will make to the research population, stakeholders, and specific fields:

(a) To contribute to optimizing existing supply chain management by comparing the hybrid model with traditional delay prediction methods, which ARIMA + XGBoost and SARIMA methods

(b) To predict the delay duration by studying how to calculate and turn uncontrollable factors into controllable factors

(c) To enable sports equipment enterprises to reduce inventory shortages, improve customer satisfaction, and minimize the costs associated with delays

1.9 Scope of Study

This project focuses on predicting the inventory instability of international logistics companies from 2015 to 2018. The research uses inventory data of sports equipment supply chains held by selected suppliers and uses forecast data accurately calculated from multiple algorithm models.

1.10 Theoretical and Conceptual Framework

Two frameworks provide a foundation for this project, which the conceptual framework is building up the concept of work through the literature, while the theoretical framework is the theory that supports each component of the mind map

1.10.1 Theoretical Framework

(a) Theory of Constraints (TOC): To analyze bottlenecks in the supply chain that cause delays

(b) Systems Theory: To understand how disruptions in one part of the supply chain impact overall inventory levels

(c) Predictive Analytics Theory: To justify the use of advanced forecasting models for delay prediction

1.10.2 Conceptual Framework

(a) Independent Variable (Cause): Transportation time, supplier reliability, demand variability, weather conditions

(b) Dependent Variable (Effect): Inventory shortages

(c) Moderating Variable: Seasonal demand, predictive accuracy

1.11 Definition of Terms

Term	Definition
Delay Prediction	The process of accurately calculating and predicting the severity of supply chain delivery delays and the possibility of various factors (Keung et al., 2021)
Inventory Shortage	A phenomenon in a supply chain scenario where demand is greater than the existing inventory and leads to the orders not being fulfilled or supply shortage (Keung et al., 2021)
Autoregressive Integrated Moving Average (ARIMA)	A time series forecasting statistical model (Gabellini et al., 2024) that combines: • Autoregressive component (AR) - measures the relationship between the past values and the current data point of the data set • Integrated component (I) - calculates differences to accomplish stability • Moving average component (MA) - calculates the dependency between the past forecast errors and current data point
Extreme Gradient Boosting (XGBoost)	A machine learning algorithm for predictive modeling that achieves scalable and accurate forecasting (Keung et al., 2021)
Seasonal ARIMA (SARIMA)	A statistical model for time series forecasting that can handle periodic changes in data, making up for the inability of the ARIMA to predict seasonal changes (Keung et al., 2021)
Long Short-Term Memory (LSTM)	A recurrent neural network (RNN) that handles long-term reliance on sequence data, and can handle problems with recurring delays and fluctuating demand (Keung et al., 2021)
Theory of Constraints (TOC)	A process that identifies and addresses key constraints to improve overall performance, such as inefficient supplier production, delayed shipments, and other bottlenecks (Gabellini et al., 2024)
Systems Theory	A theory that considers the supply chain as an interconnected system and assumes that changes in any system part will affect the entire system operation (Keung et al., 2021)

Predictive Analytics Theory	A theory that predicts future outcomes by combining statistical models, historical data, and machine learning algorithms (Keung et al., 2021)
Independent Variable	A variable that is manipulated or measured by the dependent variable effect (Hudnurkar et al., 2024)
Dependent Variable	A variable that changes in the independent variable will affect the response or outcome (Keung et al., 2021)
Moderating Variable	A variable that affects the management or robustness of relationship between the dependent and the independent variables (Hudnurkar et al., 2024)

1.12 Markmap

Figure 1.1 Introduction Markmap

1.13 Reference

Abouloifa, M., & Bahaj, M. (2022). Machine learning approaches for supply chain forecasting: A comparative study. *International Journal of Logistics Research and Applications*, 25(4), 547–562. https://doi.org/10.1080/13675567.2022.2082730

Gabellini, M., Mazzotti, L., & Rondoni, A. (2022). Identifying supply chain risks and challenges in global cargo management. *Journal of Supply Chain Management*, 58(2), 65–77. https://doi.org/10.1111/jscm.12345

Sani, S., Mohd Zaki, S., & Haron, M. (2023). Enhancing inventory management using predictive analytics in small and medium enterprises. *Journal of Business Logistics*, 44(1), 15–30. https://doi.org/10.1002/jbl.22987

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In today's fast-paced, high-demand, and competitive markets, efficient supply chain management ensures smooth progress in the seamless flow of products and customer services (Christopher, 2016). The latest technologies of calculations and forecasting have enabled enterprises to adopt predictive modeling techniques to address challenges such as inventory shortages, demand fluctuations, and logistics optimization (Feng et al., 2019). However, despite there are advances in technologies used, many enterprises still face the problems of supply chain systems that are difficult to adapt to volatile and unstable factors (Ivanov et al., 2020). It leads to handling some specific issues hardly, and it will collapse after continuing to snowball. This literature review will refer to the theme of delayed prediction of inventory shortages in sports equipment supply chains to explore the latest methods of predictive modeling in supply chain management. This chapter will focus on traditional models and machine learning methods used for this project. By reviewing these methods, this study aims to identify existing gaps to improve the resilience and efficiency of supply chain management.

2.2 State-of-the-Art Approaches

This section provides an overview of the state-of-the-art methods for delay prediction on inventory shortage. These methods are divided into two categories, which are traditional methods and machine learning methods. Each model has its strengths and limitations, and understanding these models will help in selecting or combining techniques to achieve high prediction accuracy.

2.2.1 Traditional Methods

Traditional methods usually use mathematical and statistical models and make predictions based on historical data. The traditional method used in this project to solve the inventory shortage in the sports equipment supply chain is the autoregressive integrated moving average model (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA.

2.2.1.1 ARIMA (Autoregressive Integrated Moving Average):

ARIMA was developed by the founders George Box and Gwilym Jenkins for time series forecasting in the 1970s (Box et al., 2015). It is also called the Box-Jenkin Method. The ARIMA model is the most suitable option to capture the linear trends between the time series model and the time series' past values. Therefore, the ARIMA model is commonly used in econometrics to predict event development using relevant data such as the logistics supply chain data used for this project (Hyndman & Athanasopoulos, 2018). The ARIMA is used as a combination of the autoregressive model (AR) and moving average model (MA). The autoregressive model is used to decide the order in the current value influenced by the lag observation numbers or find the p-value from the partial autocorrelation function (PACF) graph. Likewise, the moving average model can describe the order in lagged forecast errors or look for the q-value from the autocorrelation function (ACF) graph. Based on the graph needed, the ARIMA model will combine three parameters: p-value, q-value, and d-value. The d-value is the different value that makes the time series from non-stationary data into stationary data. In conclusion, the ARIMA model focuses on the relationship between autocorrelation and linear dependencies. The following are the equations and explanations of the ARIMA model:

- $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \theta_1 \epsilon_{t-1} + \ldots$

  This is the ARIMA model's equation, where

- $\phi_1, \phi_2, \ldots$: Autoregression coefficients (AR) to define the influence from past values to current value
- X_t: The value at the time t
- $X_{t-1}, X_{t-2}, \ldots$: The past value of time series (autoregressive component)
- θ_1, θ_2: Moving average coefficients (MA) to revise the predictions from past errors
- ϵ_t: Random noise (white noise)
- $\epsilon_{t-1}, \epsilon_{t-2}$: Past error terms (moving average component)

2.2.1.2 SARIMA (Seasonal Autoregressive Integrated Moving Average):

SARIMA is an extension of ARIMA developed during the same period by the founders George Box and Gwilym Jenkins to take seasonality into account (Box et al., 2015), which adds the seasonal autoregressive function to handle the periodic pattern (Makridakis et al., 1998). The periodic pattern is also known as the symbol 's', measured in month units, and increased as consideration to the ARIMA model to become (p, d, q, s). The following are the equations and explanations of the SARIMA model:

- $X_t = \phi(B)X_t + \phi(B^s)X_t + \theta(B)\epsilon_t + \theta(B^s)\epsilon_t$

  This is the SARIMA model's equation, where

  - B: Lag operator ($BX_t = X_{t-1}$)
  - $\phi(B)$: Regular autoregressive terms
  - $\phi(B^s)$: Seasonal autoregressive terms
  - $\theta(B)$: Regular moving average terms
  - $\theta(B^s)$: Seasonal moving average terms
  - ϵ_t: Error term

2.2.2 Machine Learning Methods

Machine learning methods help to customize the nonlinear data modeling by using the data-driven method. These methods are more resilient than traditional models and can handle large data sets with multiple variables. The machine learning methods used in this project to solve inventory shortages in the sports equipment supply chain are the extreme gradient boosting model (XGBoost), long short-term memory model (LSTM), and a hybrid model consisting of autoregressive integrated moving average and extreme gradient boosting (ARIMA+XGBoost).

2.2.2.1 XGBoost (Extreme Gradient Boosting):

XGBoost was developed by its founder Tianqi Chen in 2014 to optimize the residual problem to achieve the highest delay prediction accuracy. XGBoost is a nonlinear model that forecasts complex patterns using gradient boosting and the decision trees method to handle complex feature interactions. In this project, XGBoost will help to modify those factors that will be influencing the inventory system such as order volume, shipping time, and weather. The following are the equations and explanations of the XGBoost models:

- $y_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$

  This is the XGBoost model's equation, where

  - y_i: Predicted value for the i^{th} sample
  - f_k: PRediction from the k^{th} tree
  - F: Space of all possible decision tree
- $Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$
  - $l(y_i, \hat{y}_i)$: Loss function
  - $\Omega(f_k)$: Regularization terms for controlling model complexity

2.2.2.2 LSTM (Long Short-Term Memory):

LSTM was proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997. It is a special recurrent neural network (RNN) that can reduce the problem of gradient vanishing and solve the problem that the original RNN has a weak ability to handle long-term dependencies, such as time series and language. LSTM usually uses gates to determine whether to keep or discard the data information. The core operations of LSTM are divided into four gate types, which are the forget gate, input gate, cell state update, and output gate. In this project, LSTM can capture historical inventory changes, reduce the problems caused by incomplete data and missing values on memory calculations, improve the accuracy of predicting inventory shortage, and optimize supply chain efficiency. The following are the explanations and equations of the LSTM gates:

- Forget Gate: Determine the amount of past information to forget
  $f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f)$
- Input Gate: Decide the amount of new information added

  $i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i), \overline{C}_t = tanh(W_c \bullet [h_{t-1}, x_t])$

- Cell State Update: Capture long-term memory for the sequence

  $C_t = f_t \odot C_{t-1} + i_t \odot \overline{C}_t$

- Output Gate: Outputs relevant information for the next step

  $o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o), h_t = o_t \odot tanh(C_t)$

  This is the LSTM model's equation, where

  - x_t: Input at time
  - h_t: Hidden state at the time t
  - C_t: Memory cell value

- W_f, W_i, W_c, W_o: Weight matrices
- σ: Sigmoid activation function
- ⊙: Element-wise multiplication

2.2.2.3 Hybrid Model (ARIMA + XGBoost):

The hybrid model is combined with the ARIMA model and XGBoost model. The ARIMA model forecasts may have errors and the residuals may be related to order volume, shipping time, and weather. Hence, the XGBoost model will help to handle these nonlinear effects. By combining the two models, the inventory forecasting issues can be captured for robust prediction whether both linear and nonlinear patterns. The following are the equations and explanations of the Hybrid model:

- $\hat{y}_t = \hat{y}_{ARIMA} + \hat{y}_{XGBoost}$

  This is the Combine Model model's equation, where

  - $\hat{y}_{ARIMA}$: Forecast from the ARIMA model
  - $\hat{y}_{XGBoost}$: Forecast from the XGBoost model

2.2.3 Real-Time and Predictive Analytics in Supply Chain

This section will explain how all of the above methods can be applied to real-time and predictive analytics in the supply chain. First, the real-time analytics is processing the data immediately when it is generated. It allows the enterprises to monitor operational changes, track inventory situations, and make respond or take action once there is any changes. On the other hand, the predictive analytics uses historical and real-time data to predict future events, providing data evidence for demand forecasting, risk management, and logistics optimization such as the possibility of potential demand disruptions or fluctuations. On a technical level, both real-time and predictive analytics are crucial. For example, predictive analytics can

predict inventory shortages and propose the risks reduction solution, while real-time analytics can ensure the accuracy and timeliness of predictions. Although both have advantages, there are still a lot of challenges that have to be addressed in the aspects of data collection, data integration, data authenticity, and model complexity. In summary, the combination of the above methods can provide solutions to overcome these challenges and improve supply chain performance.

2.3 Limitation

Real-time and predictive analytics have contributed to the delayed prediction of inventory shortages in the sports equipment supply chain. However, some limitations obstruct the effectiveness of supply chain management. This section will introduce the limitations of each algorithm model.

The traditional ARIMA model cannot handle the non-linear relationships and multiple patterns even though it is simple and efficient at linear trend prediction. ARIMA is only applicable to single time series predictions that are calculated without external variables. On the contrary, XGBoost in the machine learning method has strong nonlinear modeling capabilities and can handle multiple patterns at the same time. Nevertheless, handling the time dependence of time series is the weakness of XGBoost. Hence, it can be seen that ARIMA and XGBoost are a perfect combination that can complement each other's shortcomings. Nevertheless, the complexity of hybrid model is extremely high and multiple steps are required to complete the calculation.

In addition, SARIMA is a model that is suitable for the seasonal time series, which is devised for the external variable problems that the ARIMA model cannot handle. Nonetheless, the model is complex and the parameters need to be adjusted manually. Moreover, the LSTM model can capture long-term and non-linear patterns, but it requires a large amount of data and computing resources, which undoubtedly limits its applicability to small and medium-sized enterprises with limited resources.

The hybrid model that combines traditional methods with machine learning methods can indeed work in improving forecast accuracy. However, some methods are simple but not comprehensive, while comprehensive models are complex and time-consuming to implement, and often require the operation with professional knowledge. Putting aside the model problems, there are significant limitations in the supply chain data itself, which reduces the forecast reliability such as incomplete data and sparsely distributed data.

2.4 Research Gap

An effective inventory management is essential to delay prediction of inventory shortages in the sports equipment supply chain, and predictive analytics has undoubtedly become an indispensable tool for supply chain management. Although various models are applicable to predicting inventory shortages and reducing delay issues, there are still major gaps due to the limitations of models. For example, the traditional models, ARIMA model and SARIMA model are limited by its reliance on linear assumptions, which are not suitable for dealing with the complex problems of non linear and dynamic characteristics of the supply chain. In short, there are too many unstable variables in the changes in modern supply chains, and a single calculation method has no longer to meet the modern system requirements anymore. On the contrary, although XGBoost and LSTM in machine learning methods can handle complex calculation patterns, it still demand computational and often face challenges related to data authenticity, large data sets, and large amounts of computing resources, especially for small and medium-sized enterprises.

In searching the relevant literature, it was found that very few studies focused on integrating these models into hybrid methods to fully utilize the combination advantages. In addition, there are few studies on the incomplete data and real-time adaptability into prediction systems, especially for sports supply chain projects. The existing research has mainly focused on addressing isolated aspects of supply chain

management, such as demand forecasting, inventory optimization, and reducing transportation delays, without providing a comprehensive integrated solution. Solving these gaps is critical to advancing both the theory and practice level of supply chain management. This study aims to develop a hybrid forecasting framework that can not only improve forecast accuracy but also adapt to real-time supply chain dynamics, reduce inventory shortage, and delay risks, providing actionable insights for decision-makers.

2.5 Definition of Terms

Term	Definition
Lag observation	The relationship between a current data point and its past values (lagged observations). It represents how the past influences the present or future in a data sequence. The "lag" refers to the number of time steps back from the current observation.
Partial autocorrelation function	The measurement of correlation between time series and its lagged values after removing the effects of intermediate lags
Autocorrelation function	The measurement of correlation between time series and its lagged values

2.6 Summary

Efficient supply chain management is essential to address challenges such as inventory shortages and demand fluctuations, especially in the sports equipment supply chain. The literature review explores predictive modeling techniques, categorized into traditional methods (e.g., ARIMA, SARIMA) and machine learning methods (e.g., XGBoost, LSTM, and hybrid models). Traditional methods (e.g., ARIMA and SARIMA) are effective for linear and seasonal trends, but have difficulty handling complex nonlinear dynamics and require manual parameter tuning. Machine learning models (e.g., XGBoost and LSTM) excel at handling nonlinear and multivariate datasets, but require significant computational resources and large datasets, limiting their applicability in small businesses.

Hybrid models combine traditional and machine learning methods, offering a promising solution to address the limitations of individual models, although their implementation is often resource-intensive and complex. Real-time and predictive analytics have become important tools for monitoring and forecasting supply chain operations. However, challenges such as incomplete data, complex models, and adapting to real-time dynamics remain.

The review found gaps in the research, including the lack of comprehensive hybrid solutions, insufficient attention to real-time adaptability, and the underexplored problem of incomplete data in the sports supply chain. Addressing these gaps is critical to developing a robust adaptive forecasting framework to improve supply chain efficiency, reduce risk, and support data-driven decision-making.

2.7 Reference

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. Wiley.

Christopher, M. (2016). *Logistics & Supply Chain Management*. Pearson UK.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Feng, Y., Ma, H., & Jin, Z. (2019). Supply Chain Analytics: Past, Present, and Future. *Journal of Operations Research*, 34(3), 245-267.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

Ivanov, D., Dolgui, A., & Sokolov, B. (2020). *Handbook of Supply Chain Disruption*. Springer.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting Methods and Applications*. Wiley.

Xu, L., Liu, J., & Zheng, Y. (2020). Real-Time Supply Chain Analytics: Methods and Applications. *Computers & Industrial Engineering*, 142, 106355.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes the methodology and procedures used to conduct the research in detail (Kothari, 2004). It includes the research approach, research design, population and sample collection, instruments and models used, data collection procedures, and data analysis techniques. The most important thing is the ethical considerations to ensure the validity and reliability of the research. This chapter outline aims to provide a clear and concise research framework for the problem of delay prediction on inventory shortages in the sports equipment supply chain.

3.2 Research Approach

This study uses the quantitative method to analyze numerical data and evaluate models (Creswell & Creswell, 2017). This method will generate the statistics data or historical data from large-scale survey research related to the sports equipment supply chain, and predict the problems of inventory shortages and transportation delays accurately. The data used is the consumer purchase record data that summarizes the product prices and types, consumer information, and delivery status. This method enables to monitor the measurements and record the results for statistical verification.

In addition, the ontology type that used during analyzing the data in this study is realism (Bhaskar, 1975). It is because of the assumption that delivery delays or inventory shortages, and those factors that may affect these phenomena in the sports equipment supply chain exist independently of human perception or interpretation. At the same time, the predictive model that developed in this study is based on

historical data and measurable variables, which conforms to the realism view. The reality of this study can be understood through observation, measurement, and analysis.

In terms of epistemology, the way of knowledge to acquired, justified, and validated in this study belongs to positivism (Bryman, 2016). It mainly emphasizes that knowledge is conjectural rather than absolute truth and comes from empirical evidence and objective observation. The purpose of this study is to study the causal relationship between supplier reliability, seasonal demand, and delayed delivery. This study relies on historical data on the sports equipment supply chain. It analyzes supply chain behavior in a data-driven manner by adopting quantitative techniques of statistical models and machine learning algorithms to minimize bias and objectively interpret data and predict results to ensure objectivity and unbias of the knowledge.

3.3 Research Design

This study will adopt the modeling method and experiment design as research design. The modeling method is a systematic approach that used to model real-world systems, processes or phenomena to understand, analyze or predict the behavior (Law & Kelton, 2015). It mainly uses mathematical modeling to identify the optimal conditions in the supply chain, predict the delays and inventory shortage problems in complex supply chain systems, view the asymptotic behavior of the system, and propose the effective solutions. The implementation steps are to parameterize the different attributes of the collected historical data, analyze its parameters to establish the prediction model, and finally test the system behavior under different conditions to find the optimal solution for delays and inventory shortages.

The most suitable design under the experiment design is the time series design and non-randomized control group pretest-posttest design or non-equivalent group design in the quasi-experimental design series (Shadish, Cook, & Campbell, 2002). The quasi-experimental design is to generate supply chain data from historical records,

which can be fully utilized and the influence of confounding variables can be controlled as much as possible. The time series design is suitable because it allows to observe the specific interventions to achieve dynamic time analysis. This design focuses on the changes in data collected at multiple time points before and after the intervention. In contrast, the non-randomized control group pretest-posttest design is suitable because it can achieve non-randomized grouping functions according to innate conditions. This design focuses on creating natural groupings and comparing the results of each group based on different conditions in the supply chain environment, such as transportation distance or seasonal demand changes.

3.4 Population and sample

This section will describe the steps on how the data obtained will be processed and analyzed. In this study, this section will focus on the interaction and relationship between the population, sample, and sampling.

3.4.1 Population

The population for this study consists of historical supply chain data from a sports equipment company. This includes a wide range of data related to customer information, order information, product information, payment and shipping information, department information, and category information. Specifically, the population covers all available records over a three-year period from 2015 to 2018, which is crucial for identifying patterns and predicting inventory shortages and shipment delays (Kumar, 2019).

3.4.2 Sample

A sample is a subset of a population to become manageable for selective analysis. The sample for this study covers three years of historical supply chain data. It is selected using a stratified sampling method to ensure the representation of key

variables such as seasonal demand during peak and off-peak sales periods. At the same time, judgmental sampling is also used to focus on the complete part of data to be accurate in analysis (Etikan et al., 2016).

3.4.3 Sampling

Sampling is the process of selecting a subset of individuals or items from a larger population to represent that population. The following are the steps in a sampling plan:

1. Define Population:

   All historical supply chain data, including payment type, shipping days, benefit per order, sales per customer, delivery status, late delivery risk, category id, and name; the customer information like city, country, email, first name, last name, password, segment, state, street, and zipcode; the department information like id and name, latitude, longitude, market, order information like city, country, customer id, date, id, item, item card prod id, item discount rate, item id, item product price, item profit ratio, item quantity, sales, item total, profit per order, region, state, status, and zipcode, the product information like card id, category id, description, image, name, price, and status; the shipping information like date and mode. These data are categorized by customer information, order information, product information, payment and shipping information, department information, and category information.

2. Attain Sample Frame:

   Collect comprehensive records from the company's databases.

3. Design Sample Plan:

   For probability samples known likelihood of selection, use stratified sampling to divide the population into distinct subgroups and select the samples from each subgroup as representative such as seasonal or regional subgroups. Likewise, for non-probability samples known likelihood of selection, adopt judgmental sampling or purposive sampling to select the samples based on the judgment to find the best representative since it allows to make selective

analysis and target the high-impact data such as late delivery risk or customers in a particular segment.

4. Draw Sample:

   Extract a dataset covering from 2015 to 2018, ensuring the low demand period and high demand period are involved. The following figure is the sports equipment supply chain dataset.

Type	Days for shi	Days for shi	Benefit per	Sales per cu	Delivery Sta	Late_delive	Category Id	Category Na	Customer C	Customer C
DEBIT	3	4	91.25	314.64001	Advance shi	0	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	5	4	-249.09	311.35999	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
CASH	4	4	-247.78	309.72	Shipping on	0	73	Sporting Go	San Jose	EE. UU.
DEBIT	3	4	22.860001	304.81	Advance shi	0	73	Sporting Go	Los Angeles	EE. UU.
PAYMENT	2	4	134.21001	298.25	Advance shi	0	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	6	4	18.58	294.98001	Shipping car	0	73	Sporting Go	Tonawanda	EE. UU.
DEBIT	2	1	95.18	288.42001	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	2	1	68.43	285.14001	Late deliver	1	73	Sporting Go	Miami	EE. UU.
CASH	3	2	133.72	278.59	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
CASH	2	1	132.14999	275.31	Late deliver	1	73	Sporting Go	San Ramon	EE. UU.
TRANSFER	6	2	130.58	272.03	Shipping car	0	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	5	2	45.689999	268.76001	Late deliver	1	73	Sporting Go	Freeport	EE. UU.
TRANSFER	4	2	21.76	262.20001	Late deliver	1	73	Sporting Go	Salinas	EE. UU.
DEBIT	2	1	24.58	245.81	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	2	1	16.389999	327.75	Late deliver	1	73	Sporting Go	Peabody	EE. UU.
DEBIT	2	1	-259.58	324.47	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
PAYMENT	5	2	-246.36	321.20001	Late deliver	1	73	Sporting Go	Canovanas	Puerto Rico
CASH	2	1	23.84	317.92001	Late deliver	1	73	Sporting Go	Paramount	EE. UU.
DEBIT	2	1	102.26	314.64001	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
PAYMENT	0	0	87.18	311.35999	Shipping on	0	73	Sporting Go	Mount Pros	EE. UU.
TRANSFER	0	0	154.86	309.72	Shipping on	0	73	Sporting Go	Long Beach	EE. UU.
TRANSFER	5	4	82.300003	304.81	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	4	2	22.370001	298.25	Late deliver	1	73	Sporting Go	Rancho Cor	EE. UU.
TRANSFER	3	2	17.700001	294.98001	Shipping car	0	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	2	2	90.279999	288.42001	Shipping car	0	73	Sporting Go	Billings	EE. UU.
DEBIT	6	2	131.17	285.14001	Late deliver	1	73	Sporting Go	Caguas	Puerto Rico
TRANSFER	5	2	90.540001	278.59	Late deliver	1	73	Sporting Go	Wilkes Barr	EE. UU.
PAYMENT	4	4	82.589996	275.31	Shipping on	0	73	Sporting Go	Caguas	Puerto Rico
DEBIT	3	4	-17.14	272.03	Advance shi	0	73	Sporting Go	Roseville	EE. UU.

Figure 3.1: Sports Equipment Supply Chain Dataset

5. Assess Sample:

   Check the sample for representativeness and ensure it aligns with research objectives.

6. Resample:

   Refine the sampling method to solve the missing or biased data points if necessary (Cochran, 1977).

3.5 Instrumentation

This section will illustrate the project requirements in this study, including software, hardware, methodology or algorithm, and dataset. The following is introducing the table of project requirements with descriptions:

Requirements	Items	Descriptions
Software	Python	A high-level and general-purpose programming language
	Jupyter Notebook	A web-based interactive compuitng platform
	Tableau	An interactive data visualization software that can help anyone see and understand the data
	PowerBI	A unified, scalable platform for self-service and enterprise business intelligence
Hardware	8GB RAM HP Laptop	A device to run the software
Methodology or Algorithm	Autoregressive Integrated Moving Average (ARIMA)	A time series forecasting statistical model (Gabellini et al., 2024)
	Extreme Gradient Boosting (XGBoost)	A machine learning algorithm for predictive modeling that achieves scalable and accurate forecasting (Keung et al., 2021)
	Seasonal ARIMA (SARIMA)	A statistical model for time series forecasting that can handle periodic changes in data, making up for the inability of the ARIMA to predict seasonal changes (Keung et al., 2021)
	Long Short-Term Memory (LSTM)	A recurrent neural network (RNN) that handles long-term reliance on sequence data, and can handle problems with recurring delays and fluctuating demand (Keung et al., 2021)
Datasets	Customer Information	The information of customers such as city, country, email, first name, last name, password, segment, state, street,

		and zipcode
	Order Information	The information of orders such as city, country, customer id, date, id, item, item card prod id, item discount rate, item id, item product price, item profit ratio, item quantity, sales, item total, profit per order, region, state, status, and zipcode
	Product Information	The information of products such as card id, category id, description, image, name, price, and status
	Payment and Shipping Information	The information of payment and shipping such as payment type, shipping days, benefit per order, sales per customer, delivery status, late delivery risk, category id, name, date, mode
	Department Information	The information of departments such as id and name, latitude, longitude, market

Table 3.1 Project Requirement Table

3.6 Data Collection Methods

The data source for this study was obtained from the Kaggle platform, which is shared data in the secondary data collection method. Since the Kaggle platform allows to access to the datasets without permission and the datasets in a variety of fields are provided (*Kaggle Inc., n.d.*). The sources of these datasets are all public data sets shared by researchers, organizations or individuals (*Johnston, 2017*). The selected datasets include sales records, transportation records, and so on, which are essential for modeling and predicting the supply chain delays and inventory shortages of sports equipment. The secondary data collection method allows researchers to focus on analysis and modeling without spending a lot of time and resources on primary data collection method (*Johnston, 2017*).

3.7 Data Analysis

This section will describe the data anlaysis in this study, including analysis technique, software, and unit of analysis.

3.7.1 Analysis Technique:

Analysis Technique describes the data analysis methods used in the research process. In addition to the previously mentioned XGBoost and LSTM to build predictive modeling for data analysis, descriptive statistics will also help analyze the situation. Descriptive statistics are divided into categorical frequency and grouping frequency in the frequency distribution. The categorical frequency will use the nominal level to analyze the uncalculated part of the data, while ordinal level data or grouped frequency will group and classify the calculable data to read each data level easily (*Smith, 2020*).

3.7.2 Software:

The Software section will describe the software used in the research process. As mentioned before, Python will use pandas and numpy types for data cleaning and processing, so that XGBoost can perform pattern training (*Wang & Liu, 2019*).. Tableau or powerbi is used to display the visualization of sales forecast results and trends, making it easy to read the situation at each level (*Kellen, 2021*).

3.7.3 Unit of Analysis:

Unit of Analysis describes the analysis object of interest in the data analysis process to confirm how to organize data and make inferences. The Unit of Analysis used in this study is Product level and transaction level. The product level is to observe the sales of each product and the transaction level is to confirm the sales, customer information and delivery status of each transaction (*Tan & Lee, 2018*).

3.8 Validity, Reliability, and Ethical Considerations

This section will describe the validity, reliability, and ethical consideration in this study.

- **Validity:** This study uses well-established statistical methods, such as ARIMA, SARIMA, XGBoost, and LSTM prediction methods, to ensure the research is stable and effective (*Jones & Lee, 2020*).
- **Reliability:** This study will test various prediction models on the dataset at different timeframes to ensure that the results obtained from repeated tests are consistent (*Robinson, 2019*).
- **Ethical Considerations:** The dataset used in this study has encrypted the private information of company, suppliers and customers. Therefore, this study adheres to the ethical guidelines of data use and transparency to ensure that all data processing processes comply with privacy regulations (*Smith & Green, 2018*).

3.9 Chapter Summary

This chapter outlines the methodology used to predict delays and inventory shortages in the sports equipment supply chain. It begins with an explanation of the research approach, which adopts a quantitative method to analyze historical supply chain data using statistical models and machine learning techniques. The research is grounded in realism ontology and positivist epistemology, emphasizing objective observation and empirical evidence.

The research design employs modeling and quasi-experimental approaches, including time-series design and non-randomized control group pretest-posttest design. These methods allow for dynamic time analysis and control of confounding variables, ensuring accurate predictions and effective solutions.

The population consists of historical supply chain data from 2015 to 2018, covering customer, order, product, payment, shipping, and department information. Stratified sampling ensures representation of key variables, while judgmental sampling targets high-impact data, such as late delivery risks.

Instrumentation includes software like Python, Jupyter Notebook, Tableau, and Power BI for data processing and visualization, alongside statistical models such as ARIMA, SARIMA, XGBoost, and LSTM for predictive analysis. Hardware requirements include an 8GB RAM laptop to support the computation.

Data collection is based on secondary data from Kaggle, enabling access to publicly shared datasets for analysis. Data analysis combines descriptive statistics with machine learning to identify patterns and predict outcomes. The unit of analysis includes product-level and transaction-level data, focusing on sales, customer details, and delivery statuses.

Validity and reliability are ensured through established statistical methods and repeated testing of prediction models. Ethical considerations include encrypting private information to adhere to privacy regulations and maintain transparency throughout the research process.

In summary, this comprehensive methodology ensures the research's rigor and reliability in addressing supply chain challenges.

3.10 Reference

Cochran, W. G. (1977). Sampling techniques (3rd ed.). Wiley.

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics, 5*(1), 1-4.

Johnston, M. P. (2017). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries (QQML), 3*, 619-626.

Jones, M., & Lee, R. (2020). Statistical methods in predictive modeling. *Journal of Applied Statistics, 25*(4), 30-42.

Kaggle Inc. (n.d.). About Kaggle datasets. Retrieved from https://www.kaggle.com

Kellen, V. (2021). Data visualization tools for business analytics. *Journal of Data Science, 9*(2), 105-116.

Kumar, R. (2019). Research methodology: A step-by-step guide for beginners. Sage Publications.

Law, A. M., & Kelton, W. D. (2015). Simulation modeling and analysis. McGraw-Hill Education.

Robinson, T. (2019). Testing reliability in predictive analytics. *Journal of Data Science, 11*(3), 56-62.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

Smith, J. (2020). Descriptive statistics and its use in business research. *Statistical Review, 8*(3), 45-56.

Smith, L., & Green, P. (2018). Ethical issues in data collection and use. *Journal of Information Ethics, 7*(2), 120-130.

Tan, Y., & Lee, J. (2018). Identifying unit of analysis for e-commerce transactions. *Journal of Business Analytics, 7*(1), 23-30.

Wang, Z., & Liu, S. (2019). Python libraries for machine learning: An overview. *Journal of Machine Learning Research, 15*(4), 112-118.

CHAPTER 4

EXPLORATORY DATA ANALYSIS

4.1 Introduction

Exploratory Data Analysis (EDA) is an iterative process, while the specific techniques or tools used will depend on the data types and the analysis objectives (Tukey, 1977). The EDA is an important step of data analysis that involves inspecting and concluding a dataset to know its properties, identify the design patterns, and acquire the data knowledge. It helps the form hypotheses, choose suitable modeling methods, and ensure the data quality (McKinney, 2010). Usually, it is performed before applying advanced statistical and machine learning techniques (Friedman et al., 2001). Therefore, the EDA plays an important part in obtaining an initial understanding of the data, guiding the following analysis, and making decisions for further steps in an analytical or data science project (McKinney, 2010).

There are some techniques and key components used in EDA such as Data Summary, Data Visualization, Data Distribution, Correlation Analysis, Outlier Detection, Categorical Variables, Data Transformation, Feature Engineering, Missing Data Handling, Hypothesis Testing, Data Transformation, Dimensionality Reduction, Time Series Analysis, Geospatial Analysis, and Text Analysis (Hunter, 2007; Waskom et al., 2020).

The main purpose of EDA is to verify the data before making any assumptions or hypothesis testing. Instead of understanding the design patterns in the data, it also finds interesting relationships between variables (Brownlee, 2017). The EDA can help the stakeholders by verifying whether the questions they asked are right. Moreover, The EDA can provide the solutions for standard deviations,

categorical variables, and confidence intervals. After the EDA is completed and the predictions are resolved, the EDA also can be used for more complex data analysis or modeling such as machine learning (Kuhn & Johnson, 2013).

4.2 Steps of Exploratory Data Analysis (EDA)

The analytical proof section substantiates the validity of the proposed predictive modeling approach. The foundation of the analysis relies on the following:

4.2.1 Observe the Dataset:

The purpose of observing the dataset is to describe the structure of the dataset. Therefore, the first step is import the import the Python libraries and modules like Figure 4.1. Based on the Figure 4.1 shown, the *numpy* is used to handle the large datasets with using numerical computing such as mathematical functions, operations on arrays, and so on, while *pandas* provides tools for data manipulation and analysis (McKinney, 2010). The *matplotlib* used to create static and animate visualizations, while the *seaborn* is built on top of *matplotlib* to simplify the creation of visually appealing and complex statistical plots (Waskom et al., 2020).

```
# Import The Python libraries and modules
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Figure 4.1: Import the Python libraries and modules

Furthermore, load the dataset like Figure 4.2.1 and the results of whole data will displayed like Figure 4.2.2. The *pd.read_csv* is used to load and read the comma-separated values file into a Pandas Data Frame.

```
# Load the dataset
df = pd.read_csv('https://raw.githubusercontent.com/david-liew/dataset/main/Sports%20Equipment%20Supply%20Chain%20Dataset.csv', encoding='latin-1')
df
```

Figure 4.2.1: Coding to load the dataset

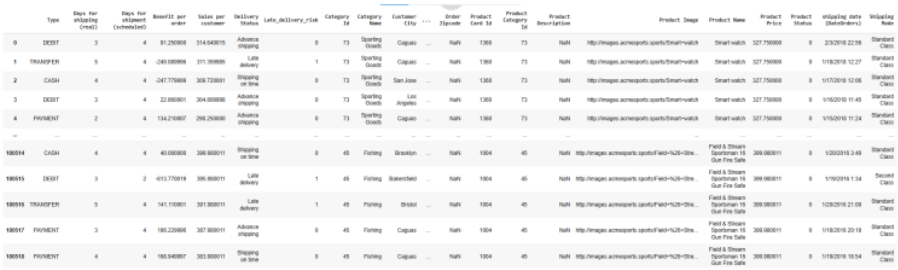

Figure 4.2.2: Result of load the dataset

Moreover, input the coding to observe the dataset like Figure 4.3.1 and display the result of data observation such as result of dataset overview (Figure 4.3.2), result of dataset summary (Figure 4.3.3), result of dataset types (Figure 4.3.4), and result of sample data (Figure 4.3.5). The *df.info()* is used to display a summary of the Data Frame including numbers of rows and columns, data types of each column, non-null counts for each column, and memory usage, while the *\033[1m* and *\033[0m* is an ANSI escape code to turns bold text on and resets the text formatting back to the normal text. The *df.shape[0]* is used to returns the number of rows in the Data Frame, while the *df.shape[1]* is used to returns the number of columns in the Data Frame. The *df.dtypes* is used to display the data types of all columns in the Data Frame, while the *df.head()* is used to display the first 5 rows of the Data Frame by default.

```
print("\033[1mPart 4.2.1: Observer the Dataset\033[0m")

# Observation
print("\n\033[1mDataset Overview:\033[0m")
df.info()

print("\n\033[1mDataset Summary:\033[0m")
print(f"Number of Rows: {df.shape[0]}")
print(f"Number of Columns: {df.shape[1]}")

print("\n\033[1mData Types:\033[0m")
print(df.dtypes)

print("\n\033[1mSample Data:\033[0m")
print(df.head())
```

Figure 4.3.1: Coding to observe dataset

```
Part 4.2.1: Observation the Dataset

Dataset Overview:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180519 entries, 0 to 180518
Data columns (total 53 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   Type                           180519 non-null  object
 1   Days for shipping (real)       180519 non-null  int64
 2   Days for shipment (scheduled)  180519 non-null  int64
 3   Benefit per order              180519 non-null  float64
 4   Sales per customer             180519 non-null  float64
 5   Delivery Status                180519 non-null  object
 6   Late_delivery_risk             180519 non-null  int64
 7   Category Id                    180519 non-null  int64
 8   Category Name                  180519 non-null  object
 9   Customer City                  180519 non-null  object
 10  Customer Country               180519 non-null  object
 11  Customer Email                 180519 non-null  object
 12  Customer Fname                 180519 non-null  object
 13  Customer Id                    180519 non-null  int64
 14  Customer Lname                 180511 non-null  object
 15  Customer Password              180519 non-null  object
 16  Customer Segment               180519 non-null  object
 17  Customer State                 180519 non-null  object
 18  Customer Street                180519 non-null  object
 19  Customer Zipcode               180516 non-null  float64
 20  Department Id                  180519 non-null  int64
 21  Department Name                180519 non-null  object
 22  Latitude                       180519 non-null  float64
 23  Longitude                      180519 non-null  float64
 24  Market                         180519 non-null  object
 25  Order City                     180519 non-null  object
 26  Order Country                  180519 non-null  object
 27  Order Customer Id              180519 non-null  int64
 28  order date (DateOrders)        180519 non-null  object
 29  Order Id                       180519 non-null  int64
 30  Order Item Cardprod Id         180519 non-null  int64
 31  Order Item Discount            180519 non-null  float64
 32  Order Item Discount Rate       180519 non-null  float64
 33  Order Item Id                  180519 non-null  int64
 34  Order Item Product Price       180519 non-null  float64
 35  Order Item Profit Ratio        180519 non-null  float64
 36  Order Item Quantity            180519 non-null  int64
 37  Sales                          180519 non-null  float64
 38  Order Item Total               180519 non-null  float64
 39  Order Profit Per Order         180519 non-null  float64
 40  Order Region                   180519 non-null  object
 41  Order State                    180519 non-null  object
 42  Order Status                   180519 non-null  object
 43  Order Zipcode                  24840 non-null   float64
 44  Product Card Id                180519 non-null  int64
 45  Product Category Id            180519 non-null  int64
 46  Product Description            0 non-null       float64
 47  Product Image                  180519 non-null  object
 48  Product Name                   180519 non-null  object
 49  Product Price                  180519 non-null  float64
 50  Product Status                 180519 non-null  int64
 51  shipping date (DateOrders)     180519 non-null  object
 52  Shipping Mode                  180519 non-null  object
dtypes: float64(15), int64(14), object(24)
memory usage: 73.0+ MB
```

Figure 4.3.2: Result of dataset overview

```
Dataset Summary:
Number of Rows: 180519
Number of Columns: 53
```

Figure 4.3.3: Result of dataset summary

```
Data Types:
Type                           object
Days for shipping (real)        int64
Days for shipment (scheduled)   int64
Benefit per order             float64
Sales per customer            float64
Delivery Status                object
Late_delivery_risk              int64
Category Id                     int64
Category Name                  object
Customer City                  object
Customer Country               object
Customer Email                 object
Customer Fname                 object
Customer Id                     int64
Customer Lname                 object
Customer Password              object
Customer Segment               object
Customer State                 object
Customer Street                object
Customer Zipcode              float64
Department Id                   int64
Department Name                object
Latitude                      float64
Longitude                     float64
Market                         object
Order City                     object
Order Country                  object
Order Customer Id               int64
order date (DateOrders)        object
Order Id                        int64
Order Item Cardprod Id          int64
Order Item Discount           float64
Order Item Discount Rate      float64
Order Item Id                   int64
Order Item Product Price      float64
Order Item Profit Ratio       float64
Order Item Quantity             int64
Sales                         float64
Order Item Total              float64
Order Profit Per Order        float64
Order Region                   object
Order State                    object
Order Status                   object
Order Zipcode                 float64
Product Card Id                 int64
Product Category Id             int64
Product Description           float64
Product Image                  object
Product Name                   object
Product Price                 float64
Product Status                  int64
shipping date (DateOrders)     object
Shipping Mode                  object
dtype: object
```

Figure 4.3.4: Result of dataset types

```
Sample Data:
       Type  Days for shipping (real)  Days for shipment (scheduled)  \
0     DEBIT                         3                              4
1  TRANSFER                         5                              4
2      CASH                         4                              4
3     DEBIT                         3                              4
4   PAYMENT                         2                              4

   Benefit per order  Sales per customer   Delivery Status  \
0          91.250000          314.640015  Advance shipping
1        -249.089996          311.359985     Late delivery
2        -247.779999          309.720001  Shipping on time
3          22.860001          304.809998  Advance shipping
4         134.210007          298.250000  Advance shipping

   Late_delivery_risk  Category Id   Category Name Customer City  ...  \
0                   0           73  Sporting Goods        Caguas  ...
1                   1           73  Sporting Goods        Caguas  ...
2                   0           73  Sporting Goods      San Jose  ...
3                   0           73  Sporting Goods   Los Angeles  ...
4                   0           73  Sporting Goods        Caguas  ...

  Order Zipcode Product Card Id Product Category Id  Product Description  \
0           NaN            1360                  73                  NaN
1           NaN            1360                  73                  NaN
2           NaN            1360                  73                  NaN
3           NaN            1360                  73                  NaN
4           NaN            1360                  73                  NaN

                              Product Image  Product Name Product Price  \
0  http://images.acmesports.sports/Smart+watch   Smart watch         327.75
1  http://images.acmesports.sports/Smart+watch   Smart watch         327.75
2  http://images.acmesports.sports/Smart+watch   Smart watch         327.75
3  http://images.acmesports.sports/Smart+watch   Smart watch         327.75
4  http://images.acmesports.sports/Smart+watch   Smart watch         327.75

  Product Status shipping date (DateOrders)   Shipping Mode
0              0             2/3/2018 22:56  Standard Class
1              0            1/18/2018 12:27  Standard Class
2              0            1/17/2018 12:06  Standard Class
3              0            1/16/2018 11:45  Standard Class
4              0            1/15/2018 11:24  Standard Class

[5 rows x 53 columns]
```

Figure 4.3.5: Result of sample data

4.2.2 Find Missing Values:

The purpose of find missing values is to identify and handle the missing values. Based on the Figure 4.4 shown, there are four coding parts which find the missing values, visualize missing values, decide the strategies to handle missing values, and recheck for the missing values. Hence the first step to find the missing

value and the result is displayed at the following Figure 4.5.1. The *df.isnull().sum()* is used to identify which columns have missing values and sum up the values that is missing (NaN) or 'False', while the *missing_values.sum()* is used to summarize the total number of missing data values in the entire dataset.

```
# Find Missing Values
print("\033[1mPart 4.2.2: Find Missing Values\033[0m")

# Find missing values
print("\n\033[1mMissing Values in Each Column:\033[0m")
print(df.isnull().sum())

total_missing = missing_values.sum()
print(f"\n\033[1mTotal Missing Values in Dataset:\033[0m {total_missing}")

# Visualize missing values
plt.figure(figsize=(10, 6))
sns.heatmap(df.isnull(), cbar=False, cmap='viridis', yticklabels=False)
plt.title("Missing Values Heatmap")
plt.show()

# Decide strategies to handle missing values
print("\n\033[1mHandling Missing Values:\033[0m")
num_cols = df.select_dtypes(include='number').columns # Fill numerical missing values with the mean
df[num_cols] = df[num_cols].fillna(df[num_cols].mean())
print(f"Filled missing numerical values with column mean for: {list(num_cols)}")

cat_cols = df.select_dtypes(include='object').columns # Fill categorical missing values with the mode
for col in cat_cols:
    df[col].fillna(df[col].mode()[0], inplace=True)
print(f"Filled missing categorical values with column mode for: {list(cat_cols)}")

# Re-check for missing values
print("\n\033[1mMissing Values After Filling:\033[0m")
print(df.isnull().sum())

if df.isnull().sum().sum() == 0:
    print("\n\033[1mAll missing values have been handled successfully!\033[0m")
else:
    print("\n\033[1mSome missing values still remain.\033[0m")
```

Figure 4.4: Coding to find missing values

```
Part 4.2.2: Find Missing Values

Missing Values in Each Column:
Type                                0
Days for shipping (real)            0
Days for shipment (scheduled)       0
Benefit per order                   0
Sales per customer                  0
Delivery Status                     0
Late_delivery_risk                  0
Category Id                         0
Category Name                       0
Customer City                       0
Customer Country                    0
Customer Email                      0
Customer Fname                      0
Customer Id                         0
Customer Lname                      0
Customer Password                   0
Customer Segment                    0
Customer State                      0
Customer Street                     0
Customer Zipcode                    0
Department Id                       0
Department Name                     0
Latitude                            0
Longitude                           0
Market                              0
Order City                          0
Order Country                       0
Order Customer Id                   0
order date (DateOrders)             0
Order Id                            0
Order Item Cardprod Id              0
Order Item Discount                 0
Order Item Discount Rate            0
Order Item Id                       0
Order Item Product Price            0
Order Item Profit Ratio             0
Order Item Quantity                 0
Sales                               0
Order Item Total                    0
Order Profit Per Order              0
Order Region                        0
Order State                         0
Order Status                        0
Order Zipcode                       0
Product Card Id                     0
Product Category Id                 0
Product Description            180519
Product Image                       0
Product Name                        0
Product Price                       0
Product Status                      0
shipping date (DateOrders)          0
Shipping Mode                       0
dtype: int64

Total Missing Values in Dataset: 180519
```

Figure 4.5.1: Result of missing values

Furthermore, the second step is to visualize the missing values and the result is displayed at the following Figure 4.5.2. The *plt.figure(figsize=(10,6))* ise used to set thye size of the figure to make it more readable, while *plt.title()* is used to add a title to the plot for context and *plt.show()* is used to display the plot. The *sns.heatmap(df.isnull())* is used to visualizes the missing values in a heatmap where the missing values are marked in yellow color and the non-missing values are purple color, while the *cbar=False* is to remove the color bar for cleaner visualization and *yticklables=False* is to hide the row labels for clarity.

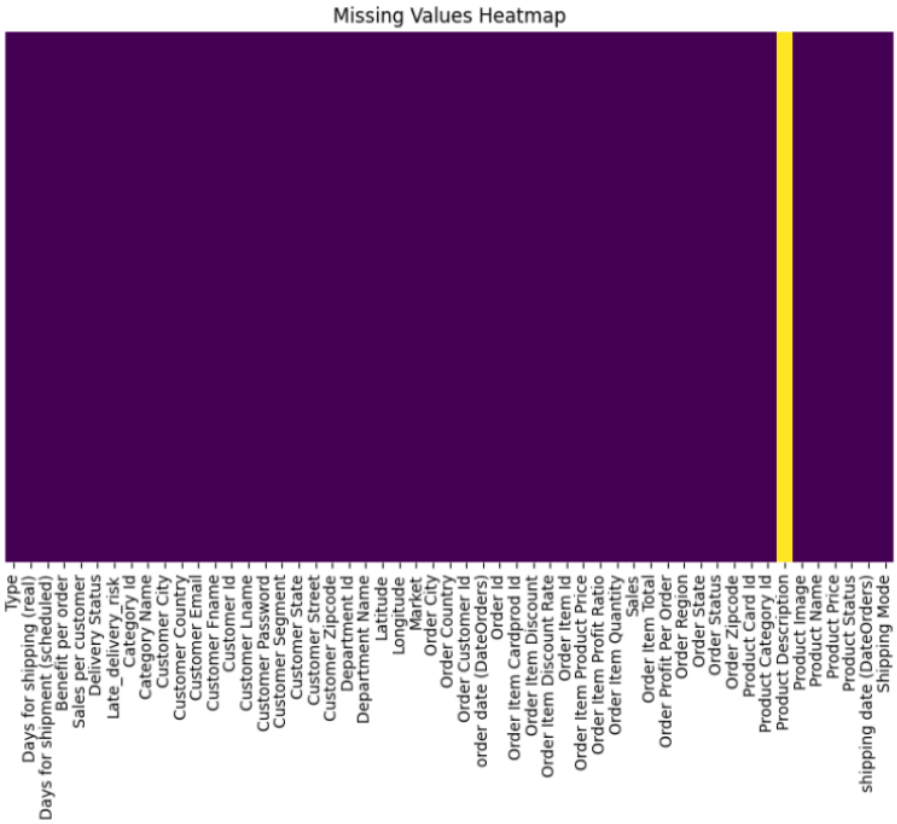

Figure 4.5.2: Result of missing values heatmap

Moreover, the third step is to decide the strategies for handling the missing values and the result is displayed at the following Figure 4.5.3. The *df.select_dtypes(include='number')* is used to identify all the numeric columns in the dataset, while the *df[num_cols].fillna(df[num_cols].mean())* is used to fills missing

values in numeric columns with the means. The *df.select_dtypes(include='object')* is used to identify all the categorical columns in the dataset, while *df[col].mode()[0]* is used to finds the most frequent value (mode) in each categorical column in the dataset and the *inplace=True* is used to update the dataset directly without create a new copy.

```
Handling Missing Values:
Filled missing numerical values with column mean for: ['Days for shipping (real)', 'Days for shipment (scheduled)', 'Benefit per order', 'Sales per customer', 'Late_delivery_risk', 'Category Id', '
<ipython-input-8-6a3db0a33f4d>:29: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the

  df[col].fillna(df[col].mode()[0], inplace=True)
Filled missing categorical values with column mode for: ['Type', 'Delivery Status', 'Category Name', 'Customer City', 'Customer Country', 'Customer Email', 'Customer Fname', 'Customer Lname', 'Cust
```

Figure 4.5.3: Result of handling missing values

In addition, the fourth step is to recheck for the missing values and the result is displayed at the following Figure 4.5.4.

```
Missing Values After Filling:
Type                              0
Days for shipping (real)          0
Days for shipment (scheduled)     0
Benefit per order                 0
Sales per customer                0
Delivery Status                   0
Late_delivery_risk                0
Category Id                       0
Category Name                     0
Customer City                     0
Customer Country                  0
Customer Email                    0
Customer Fname                    0
Customer Id                       0
Customer Lname                    0
Customer Password                 0
Customer Segment                  0
Customer State                    0
Customer Street                   0
Customer Zipcode                  0
Department Id                     0
Department Name                   0
Latitude                          0
Longitude                         0
Market                            0
Order City                        0
Order Country                     0
Order Customer Id                 0
order date (DateOrders)           0
Order Id                          0
Order Item Cardprod Id            0
Order Item Discount               0
Order Item Discount Rate          0
Order Item Id                     0
Order Item Product Price          0
Order Item Profit Ratio           0
Order Item Quantity               0
Sales                             0
Order Item Total                  0
Order Profit Per Order            0
Order Region                      0
Order State                       0
Order Status                      0
Order Zipcode                     0
Product Card Id                   0
Product Category Id               0
Product Description          180519
Product Image                     0
Product Name                      0
Product Price                     0
Product Status                    0
shipping date (DateOrders)        0
Shipping Mode                     0
dtype: int64

Some missing values still remain.
```

Figure 4.5.4: Result of missing values after filling

4.2.3 Categories Values:

The purpose of categories values is to separate the data into categorical variables and numerical variables. The following Figure 4.6 shows the coding to category value and the result is displayed at the following Figure 4.7. The *df.select_dtypes(include='object').columns* is used to select all columns in the Data Frame with a string data type, while *df.select_dtypes(include='number').columns* is used to select all columns in the DataFrame with a numeric data type.

```
# Categories Values
print("\033[1mPart 4.2.3: Categories Values\033[0m")

# Categorize variables
categorical_columns = df.select_dtypes(include='object').columns
numerical_columns = df.select_dtypes(include='number').columns

print("\033[1mCategorical Variables:\033[0m")
print(categorical_columns)

print("\n\033[1mNumerical Variables:\033[0m")
print(numerical_columns)
```

Figure 4.6: Coding to category values

```
Part 4.2.3: Categories Values
Categorical Variables:
Index(['Type', 'Delivery Status', 'Category Name', 'Customer City', 'Customer Country',
       'Customer Email', 'Customer Fname', 'Customer Lname', 'Customer Password',
       'Customer Segment', 'Customer State', 'Customer Street', 'Department Name', 'Market',
       'Order City', 'Order Country', 'order date (DateOrders)', 'Order Region', 'Order State',
       'Order Status', 'Product Image', 'Product Name', 'shipping date (DateOrders)',
       'Shipping Mode'],
      dtype='object')

Numerical Variables:
Index(['Days for shipping (real)', 'Days for shipment (scheduled)', 'Benefit per order',
       'Sales per customer', 'Late_delivery_risk', 'Category Id', 'Customer Id',
       'Customer Zipcode', 'Department Id', 'Latitude', 'Longitude', 'Order Customer Id',
       'Order Id', 'Order Item Cardprod Id', 'Order Item Discount', 'Order Item Discount Rate',
       'Order Item Id', 'Order Item Product Price', 'Order Item Profit Ratio',
       'Order Item Quantity', 'Sales', 'Order Item Total', 'Order Profit Per Order',
       'Order Zipcode', 'Product Card Id', 'Product Category Id', 'Product Description',
       'Product Price', 'Product Status'],
      dtype='object')
```

Figure 4.7: Result of category values

4.2.4 Analyze the Shape of the Data:

The purpose of analyze the shape of data is to describe the distributions and patterns in numerical data that calculate the skewness and kurtosis. The following Figure 4.8 shows the coding to analyze the shape of data and the result is displayed at the following Figure 4.9.1 until Figure 4.9.30. The *sns.histplot(df[col], kde=True, color="blue")* is used to creates a histogram of the values in the column, while *plt.xlabel(col)* and *plt.ylabel("Frequency")* are used to sets the x-axis label of the plot to the name of the column and sets the y-axis label of the plot to "Frequency," representing the count of data points in each bin of the histogram. The *df[col].skew()* and *df[col].kurt()* are used to computes the skewness and kurtosis of the distribution of the column. The Skewness measures the asymmetry of the data distribution in positive, negative or zero, which right tail is longer or fatter than the left tail, left tail is longer or fatter than the right tail, or the distribution is perfectly symmetrical like normal distribution. In contrast, yhe kurtosis measures the "tailedness" of the distribution in high, low or equal to 3, which the data has heavier tails and more outliers (leptokurtic), lighter tails and fewer outliers (platykurtic), or resembles a normal distribution (mesokurtic).

```
# Analyze the Shape of the Data
print("\033[1mPart 4.2.4: Analyze the Shape of the Data\033[0m")

# Visualize distributions of numerical variables
for col in numerical_columns:
    plt.figure(figsize=(8, 4))
    sns.histplot(df[col], kde=True, color="blue")
    plt.title(f"Distribution of {col}")
    plt.xlabel(col)
    plt.ylabel("Frequency")
    plt.show()

# Describe skewness and kurtosis
for col in numerical_columns:
    skewness = df[col].skew()
    kurtosis = df[col].kurt()
    print(f"{col} - Skewness: {skewness:.2f}, Kurtosis: {kurtosis:.2f}")
```

Figure 4.8: Coding to analyze the shape of the data

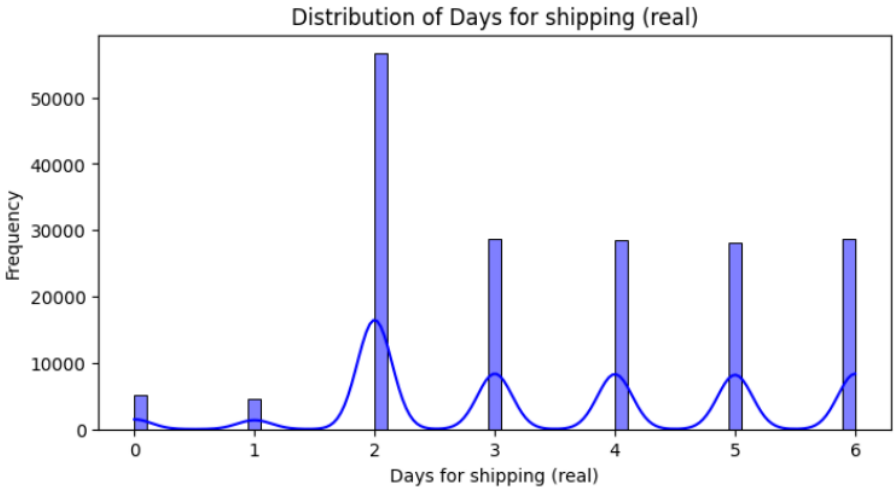

Figure 4.9.1: Result of distribution of days for shipping (real)

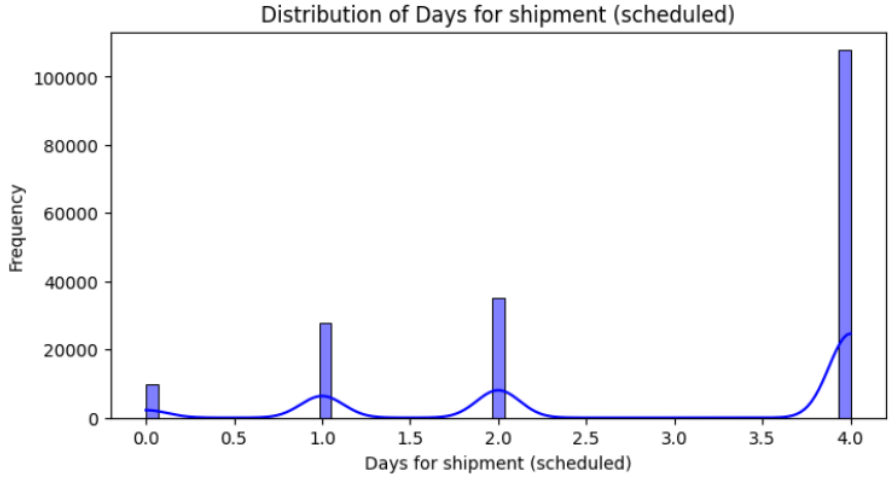

Figure 4.9.2: Result of distribution of days for shipping (scheduled)

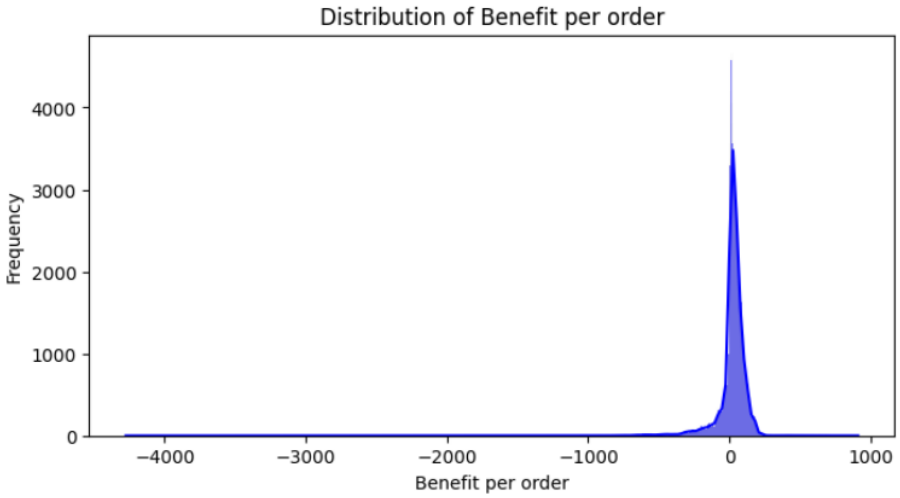

Figure 4.9.3: Result of distribution of days for benefit per order

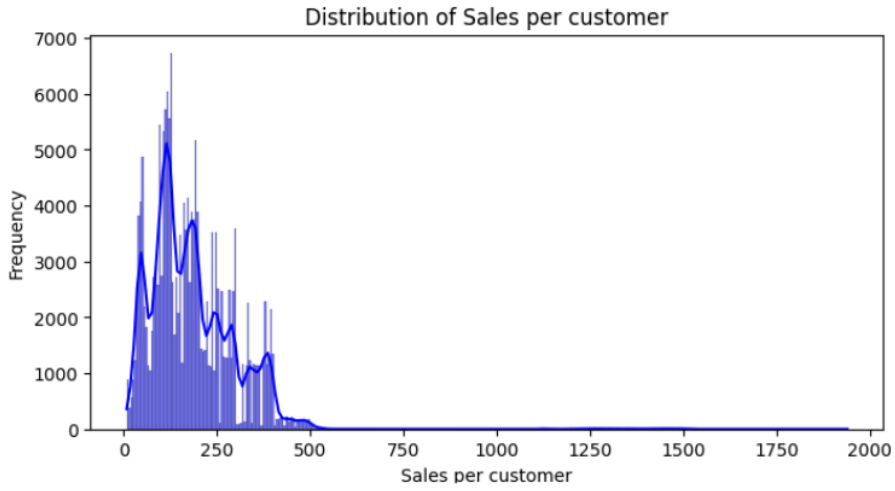

Figure 4.9.4: Result of distribution of days for sales per customer

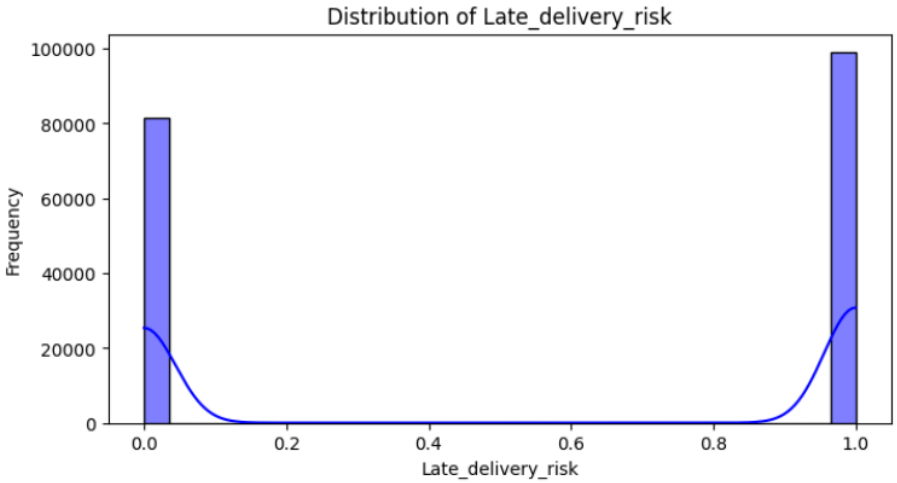

Figure 4.9.5: Result of distribution of late delivery risk

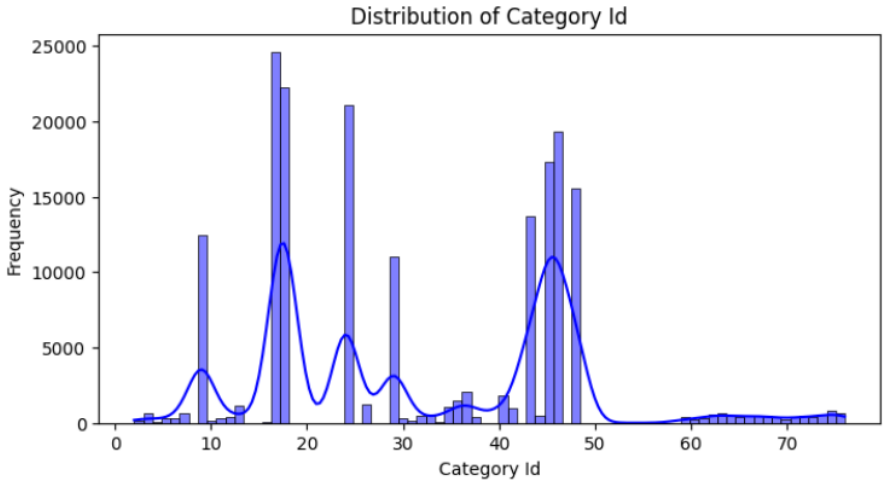

Figure 4.9.6: Result of distribution of category id

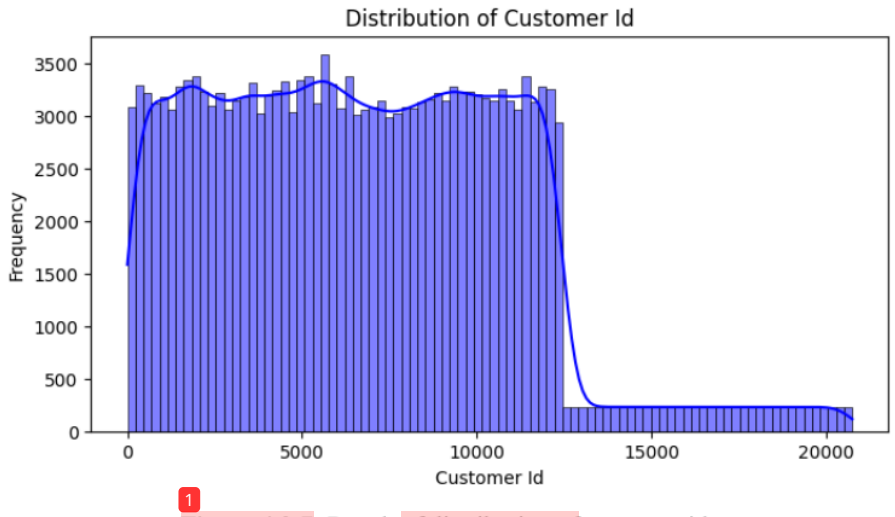

Figure 4.9.7: Result of distribution of customer id

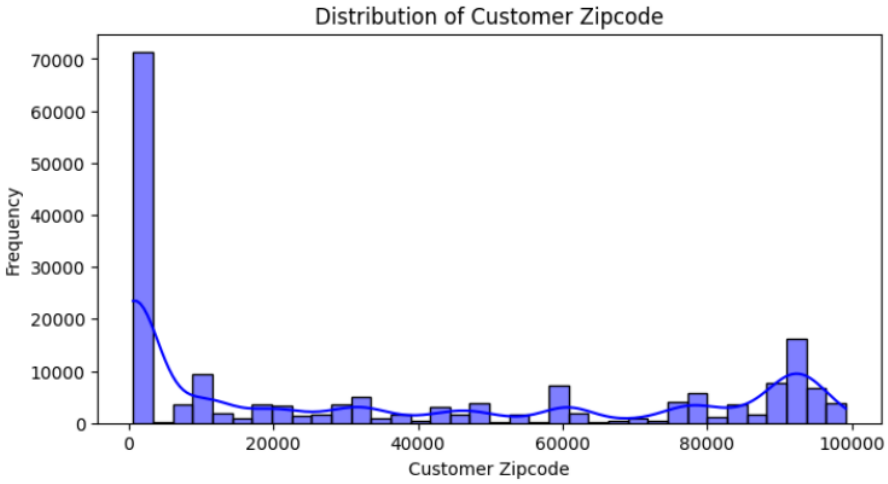

Figure 4.9.8: Result of distribution of customer zipcode

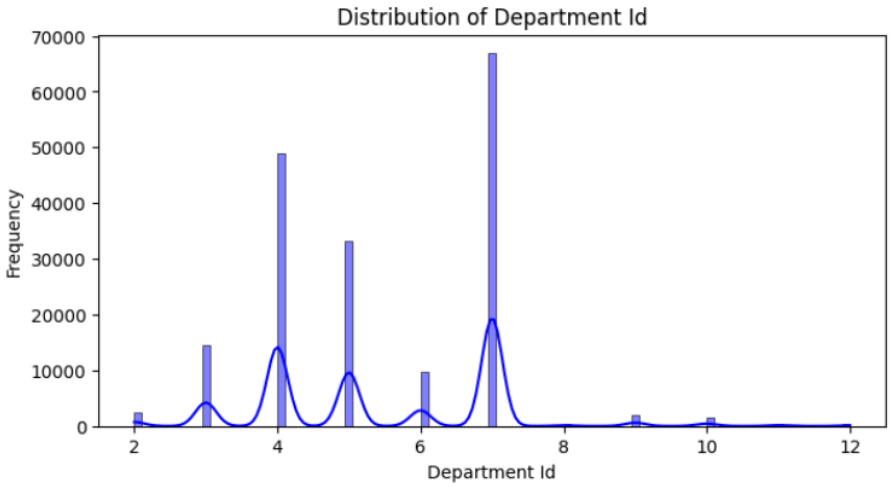

Figure 4.9.9: Result of distribution of department id

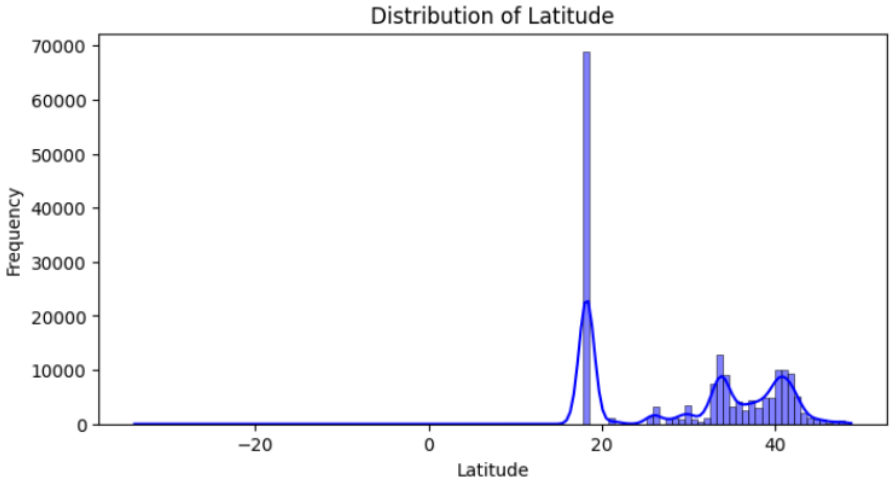

Figure 4.9.10: Result of distribution of latitude

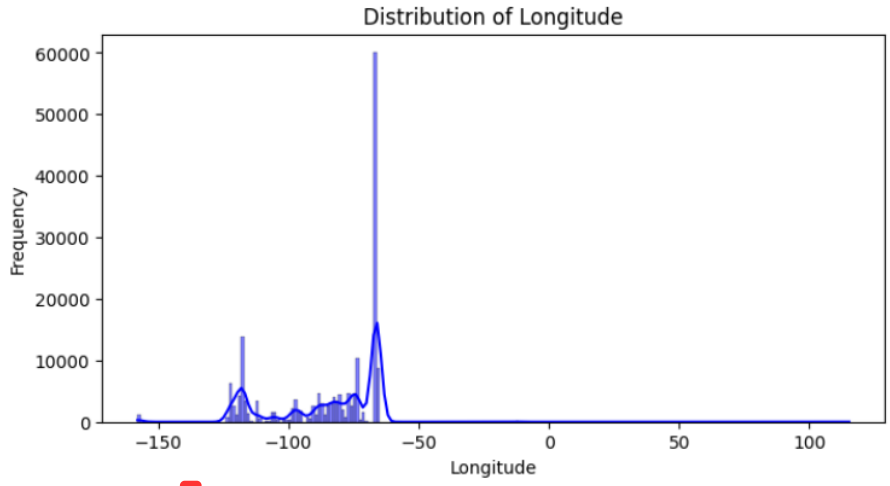

Figure 4.9.11: Result of distribution of longitude

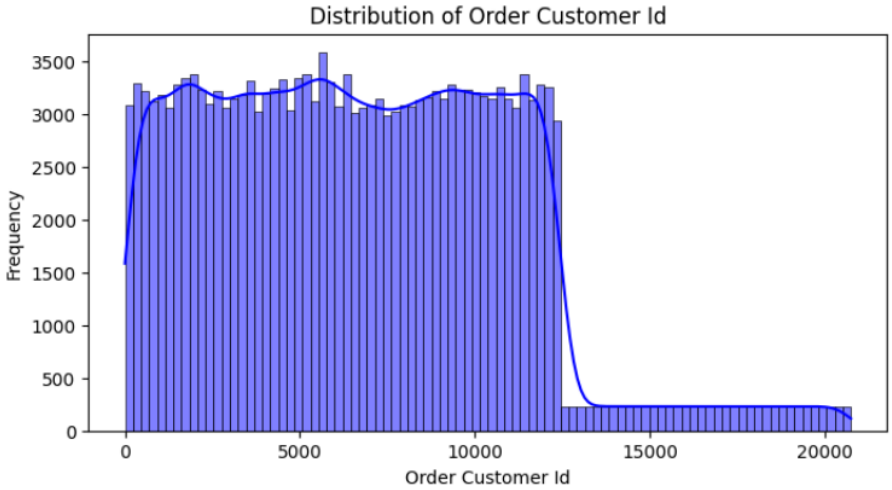

Figure 4.9.12: Result of distribution of order customer id

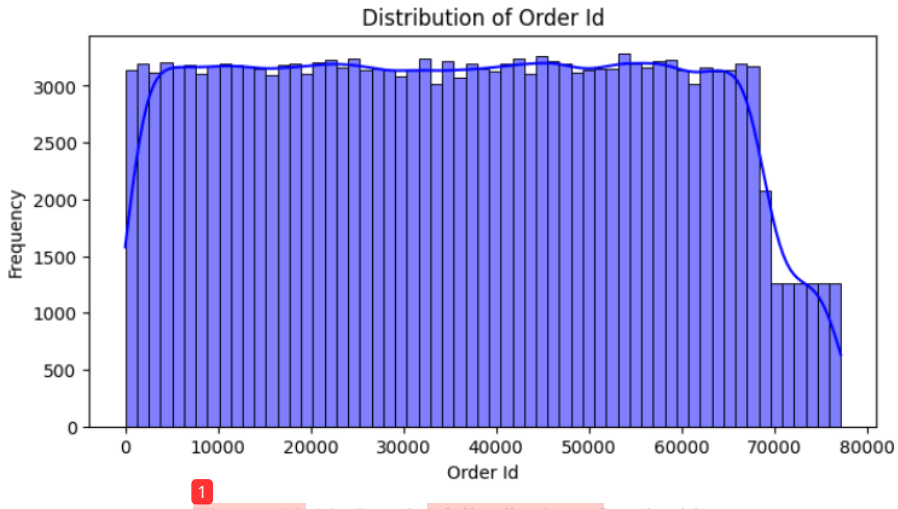

Figure 4.9.13: Result of distribution of order id

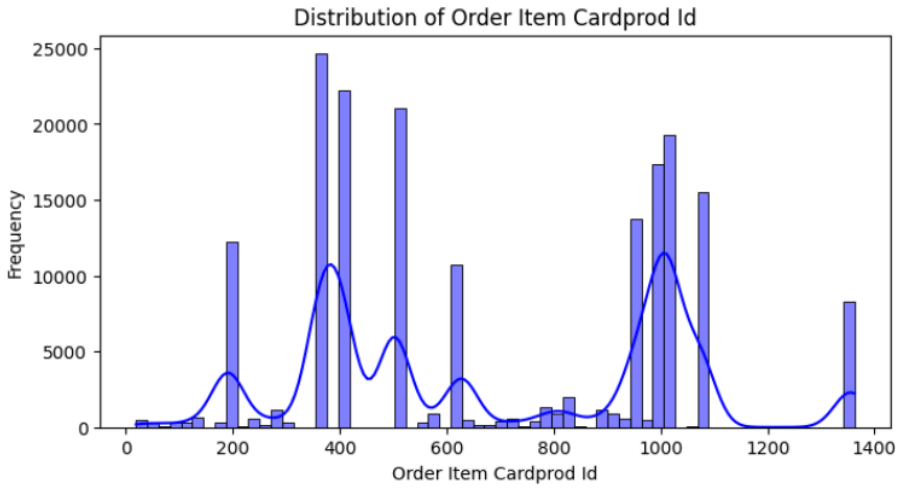

Figure 4.9.14: Result of distribution of order item cardprod id

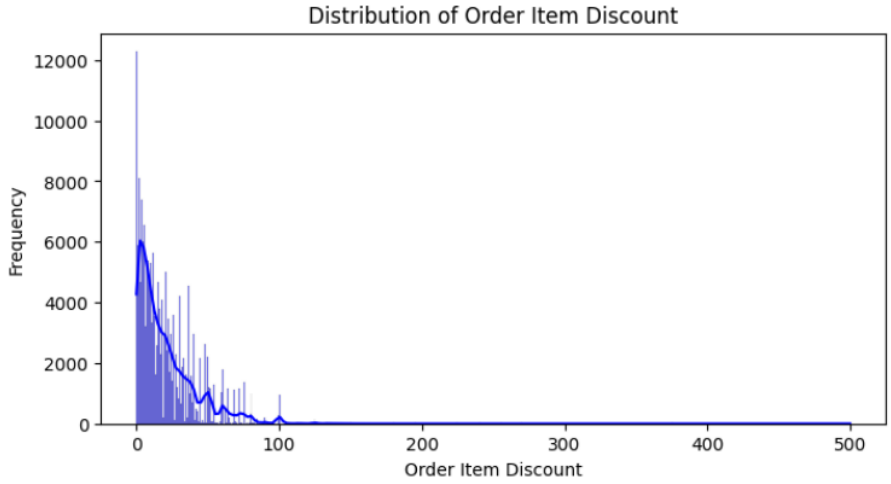

Figure 4.9.15: Result of distribution of order item discount

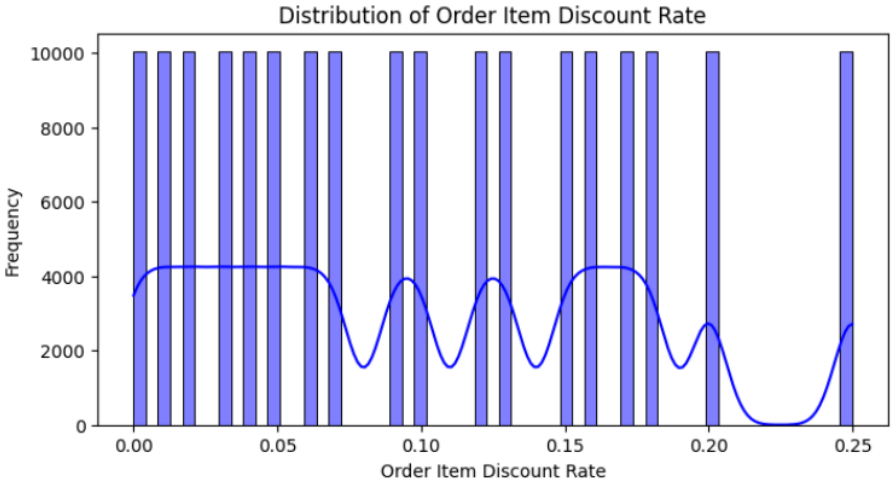

Figure 4.9.16: Result of distribution of order item discount rate

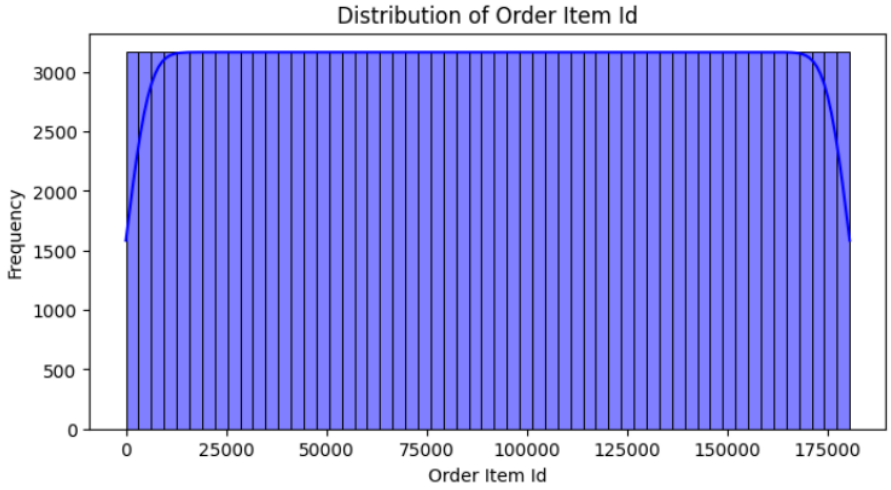

Figure 4.9.17: Result of distribution of order item id

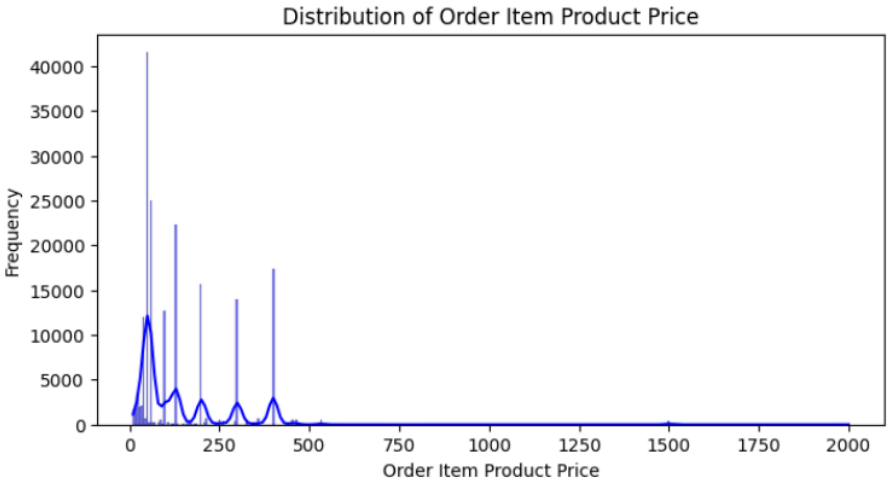

Figure 4.9.18: Result of distribution of order item product price

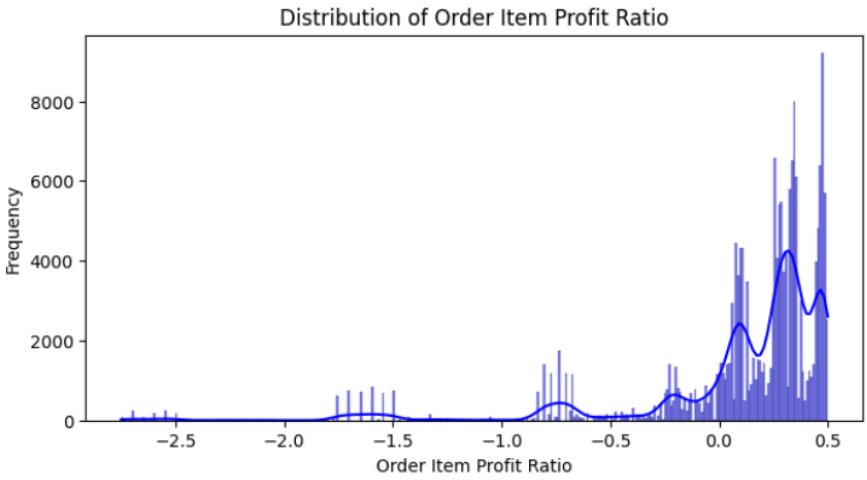

Figure 4.9.19: Result of distribution of order item profit ratio

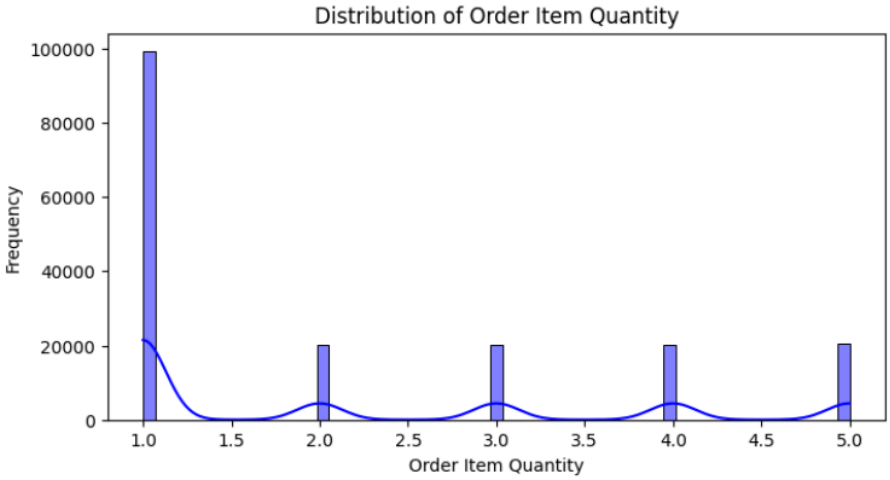

Figure 4.9.20: Result of distribution of order item quality

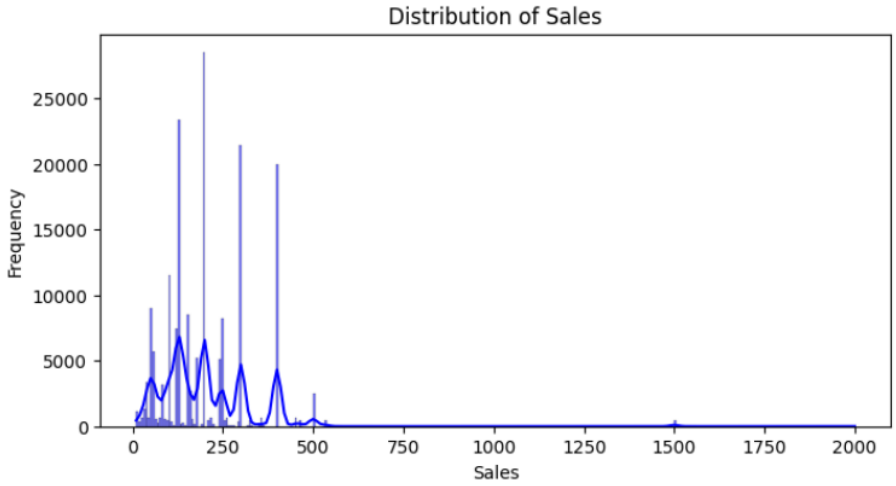

Figure 4.9.21: Result of distribution of sales

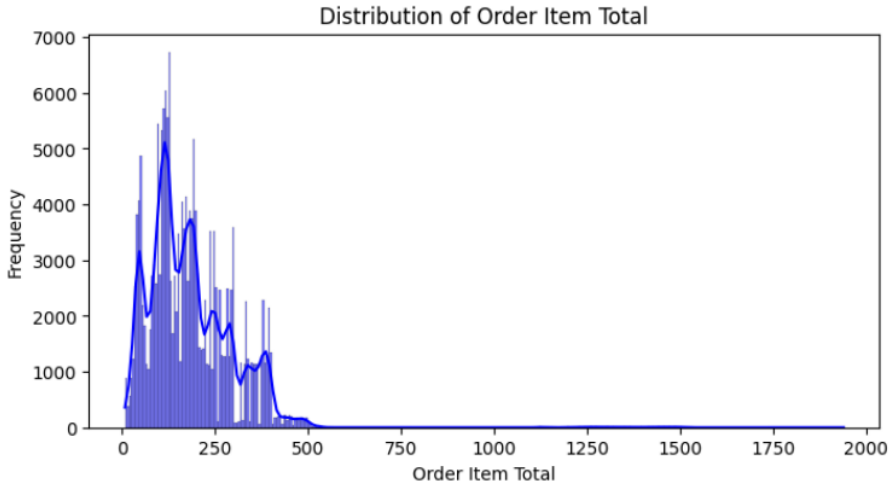

Figure 4.9.22: Result of distribution of order item total

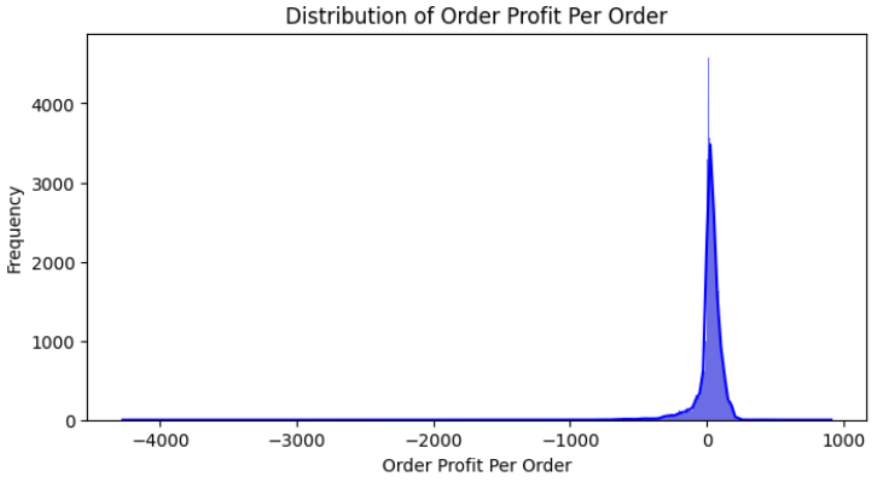

Figure 4.9.23: Result of distribution of order profit per order

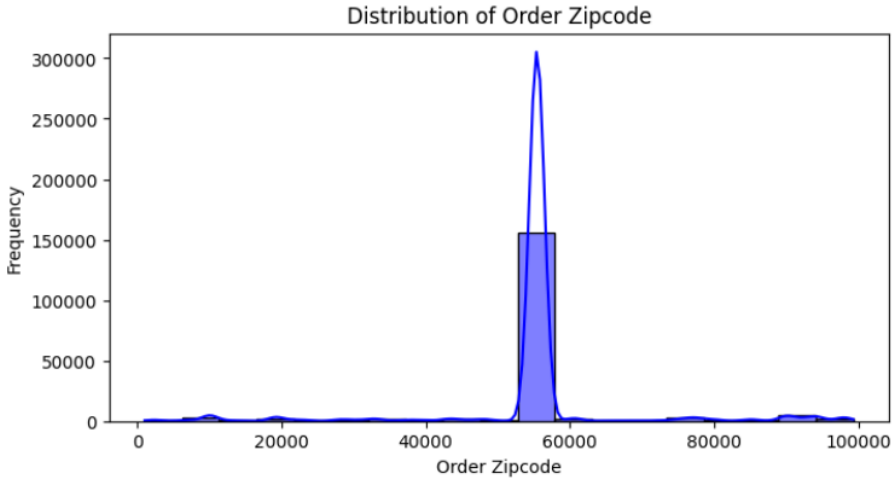

Figure 4.9.24: Result of distribution of order zipcode

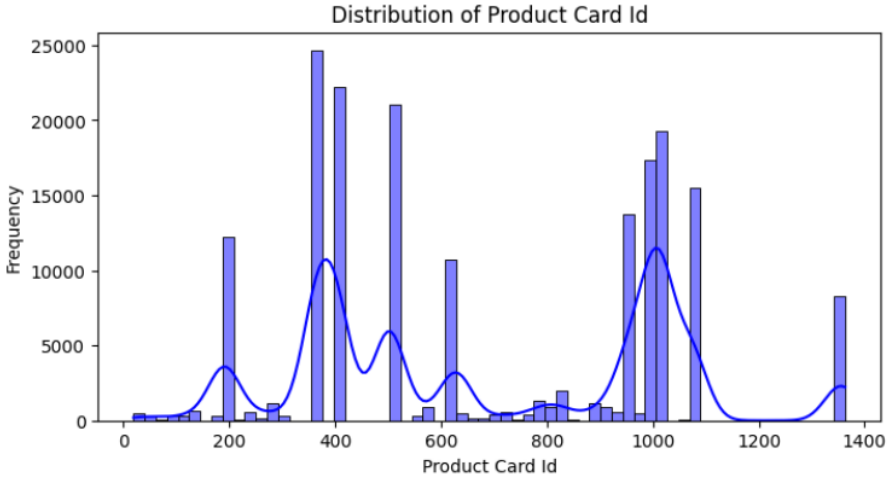

Figure 4.9.25: Result of distribution of product card id

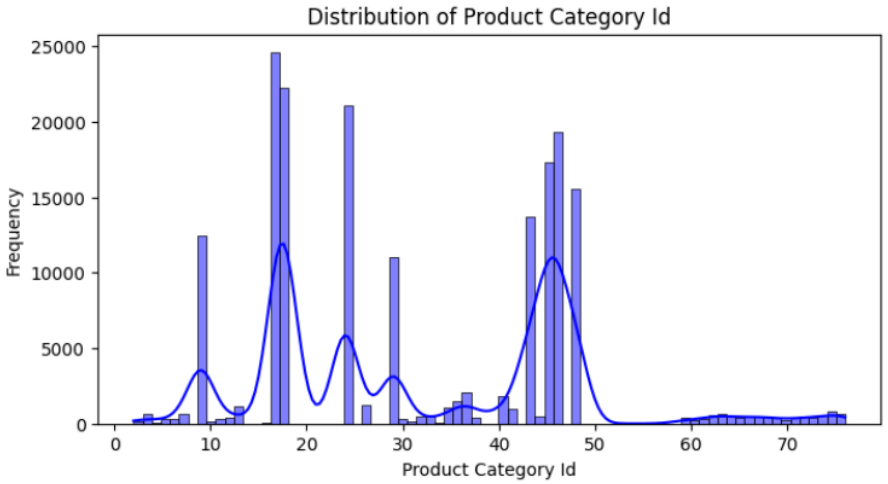

Figure 4.9.26: Result of distribution of product category id

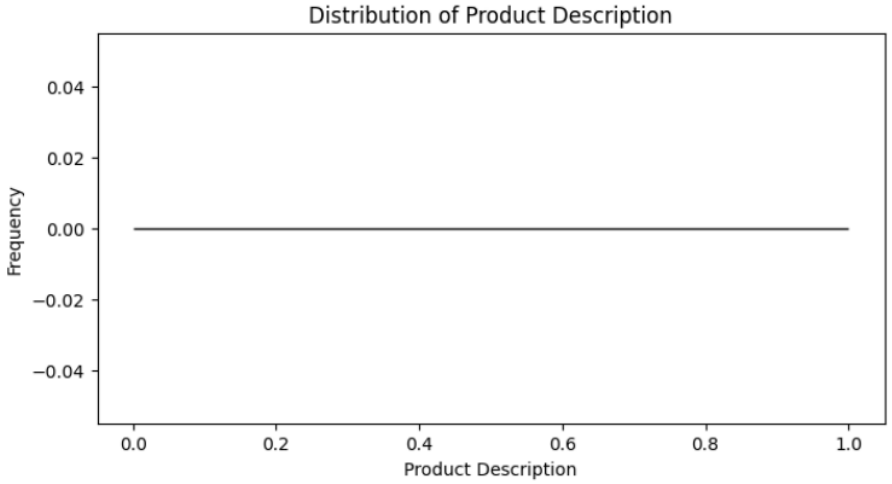

Figure 4.9.27: Result of distribution of product description

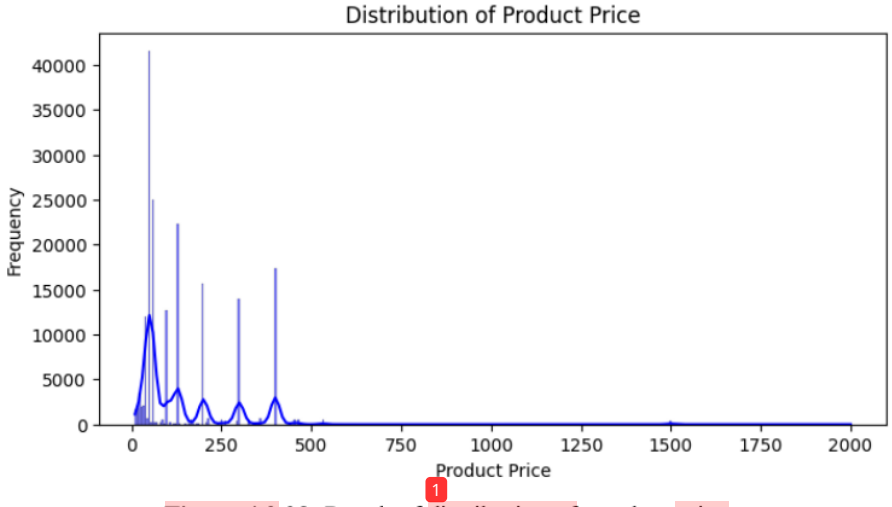

Figure 4.9.28: Result of distribution of product price

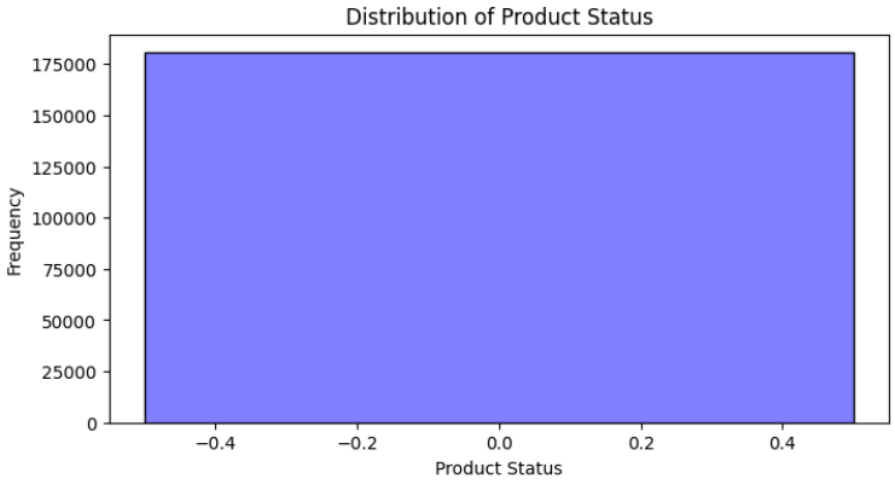

Figure 4.9.29: Result of distribution of product status

```
Days for shipping (real) - Skewness: 0.08, Kurtosis: -1.01
Days for shipment (scheduled) - Skewness: -0.73, Kurtosis: -1.02
Benefit per order - Skewness: -4.74, Kurtosis: 71.38
Sales per customer - Skewness: 2.89, Kurtosis: 23.92
Late_delivery_risk - Skewness: -0.19, Kurtosis: -1.96
Category Id - Skewness: 0.36, Kurtosis: -0.60
Customer Id - Skewness: 0.49, Kurtosis: 0.01
Customer Zipcode - Skewness: 0.49, Kurtosis: -1.45
Department Id - Skewness: 0.27, Kurtosis: -0.18
Latitude - Skewness: -0.10, Kurtosis: -1.56
Longitude - Skewness: -0.50, Kurtosis: 2.18
Order Customer Id - Skewness: 0.49, Kurtosis: 0.01
Order Id - Skewness: 0.03, Kurtosis: -1.15
Order Item Cardprod Id - Skewness: 0.14, Kurtosis: -1.27
Order Item Discount - Skewness: 3.04, Kurtosis: 25.23
Order Item Discount Rate - Skewness: 0.34, Kurtosis: -0.90
Order Item Id - Skewness: 0.00, Kurtosis: -1.20
Order Item Product Price - Skewness: 3.19, Kurtosis: 23.31
Order Item Profit Ratio - Skewness: -2.89, Kurtosis: 10.16
Order Item Quantity - Skewness: 0.88, Kurtosis: -0.75
Sales - Skewness: 2.88, Kurtosis: 23.94
Order Item Total - Skewness: 2.89, Kurtosis: 23.92
Order Profit Per Order - Skewness: -4.74, Kurtosis: 71.38
Order Zipcode - Skewness: -0.38, Kurtosis: 8.03
Product Card Id - Skewness: 0.14, Kurtosis: -1.27
Product Category Id - Skewness: 0.36, Kurtosis: -0.60
Product Description - Skewness: nan, Kurtosis: nan
Product Price - Skewness: 3.19, Kurtosis: 23.31
Product Status - Skewness: 0.00, Kurtosis: 0.00
```

Figure 4.9.30: Result of distribution of analyze the shape of the data

4.2.5 Identify Relationship Between Variables:

The purpose of identify relationship between variables is to explore correlations and dependencies, which display the numerical variables and visualize relaitronship by using correlation matrices and scatterplots or pair plots. There are two coding parts displayed at the following Figure 4.10 which correlation matrix and pair plot, and the result displayed art the following Figure 4.11 and Figure 4.12. The *sns.pairplot* is used to create a grid of scatter plots and density plots for all combinations of numerical variables, while the *plt.suptitle* is used to add a title for the pair plot.

```
# Correlation Matrix
if len(numerical_columns) > 1:
    plt.figure(figsize=(12, 8))
    sns.heatmap(df[numerical_columns].corr(), annot=True, fmt=".2f", cmap="coolwarm")
    plt.title("Correlation Heatmap")
    plt.show()
else:
    print("Not enough numerical columns for correlation matrix.")

# Pair Plot
if len(numerical_columns) > 1:
    sns.pairplot(df[numerical_columns], diag_kind='kde')
    plt.suptitle("Pairplot of Numerical Variables", y=1.02)
    plt.show()
else:
    print("Not enough numerical columns for pair plot.")
```

Figure 4.10: Coding to identify relationship between variables

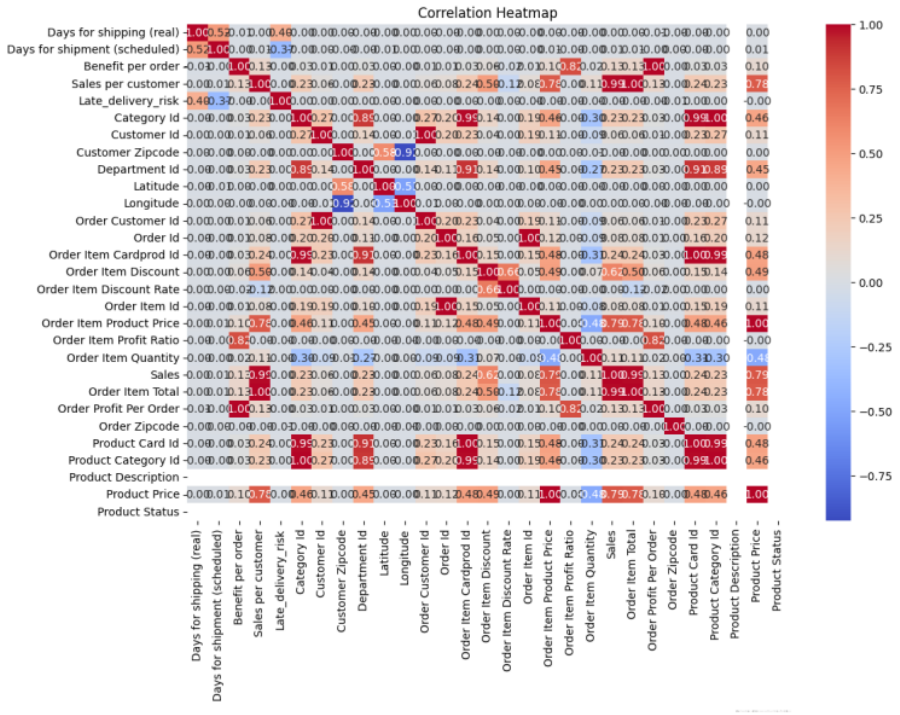

Figure 4.11: Result of correlation matrix

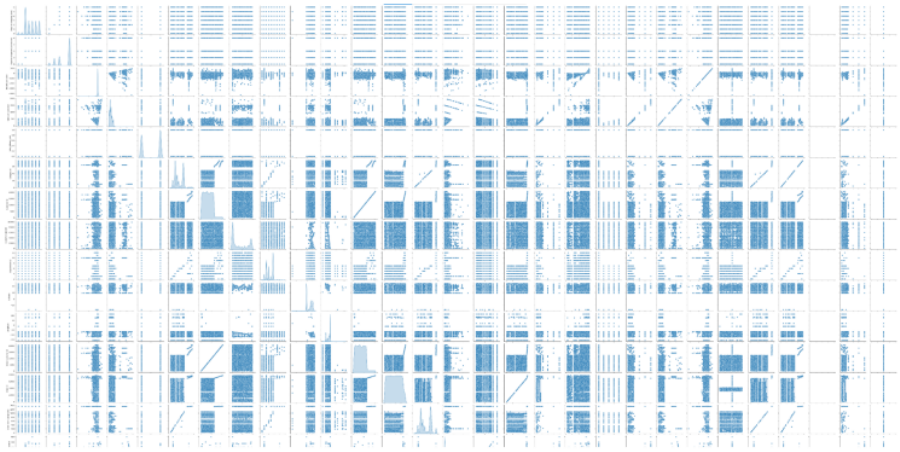

Figure 4.12: Result of pair plot

4.2.6 Locate Outliers

The purpose of locate outliers are to detect and address outliers in the dataset. There are two coding parts displayed at the following Figure 4.13, which visualize outliers using boxplots and discuss the strategies to handle the outliers. The results are displayed at the following Figure 4.14.1 until Figure 4.14.26. The *sns.boxplot* is used to create box plots.

```
# Ensure `numerical_columns` contains only valid numerical columns
numerical_columns = df.select_dtypes(include=["number"]).columns

# Visualize outliers with boxplots
for col in numerical_columns:
    plt.figure(figsize=(8, 4))
    sns.boxplot(x=df[col], color="lightblue")  # Correctly use the x parameter
    plt.title(f"Boxplot of {col}")
    plt.xlabel(col)
    plt.show()
```

Figure 4.13: Coding to locate outliers

Figure 4.14.1: Result of boxplot of days for shipping (real)

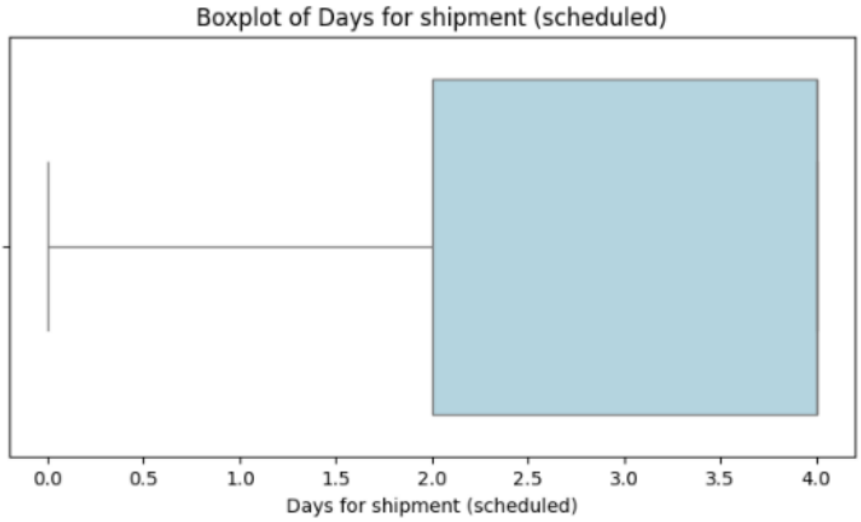

Figure 4.14.2: Result of boxplot of days for shipping (scheduled)

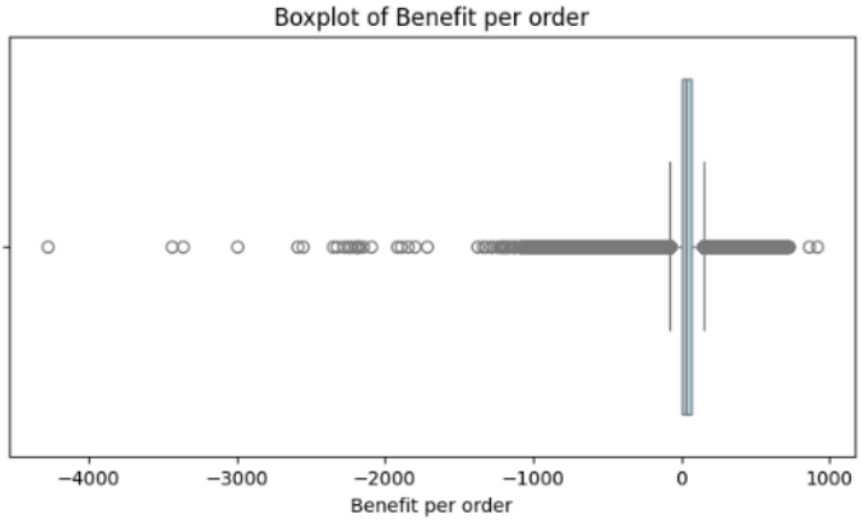

Figure 4.14.3: Result of boxplot of benefit per order

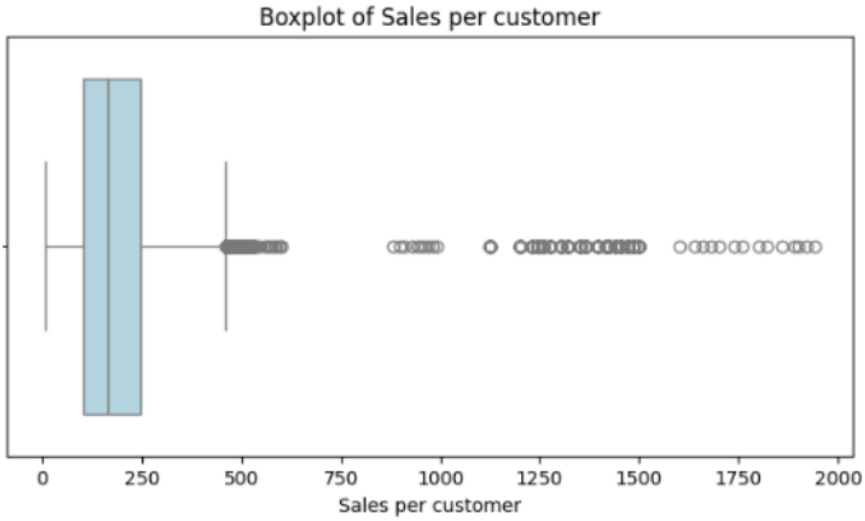

Figure 4.14.4: Result of boxplot of sales per customer

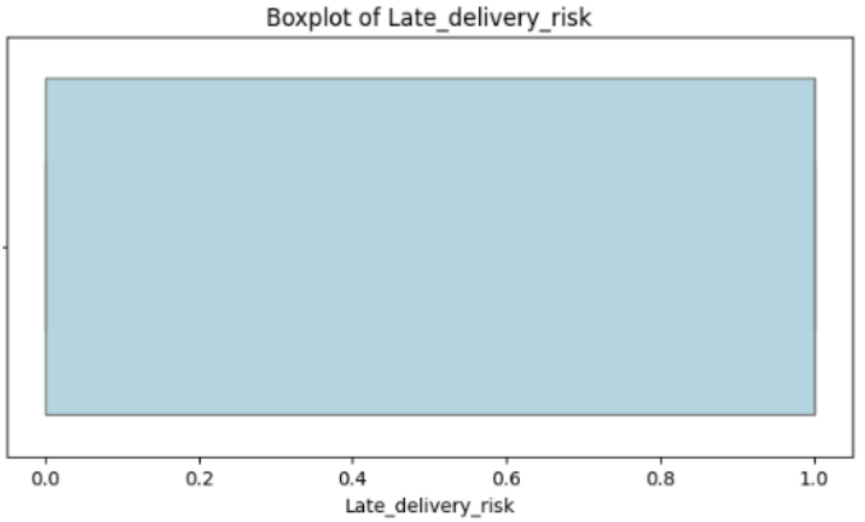

Figure 4.14.5: Result of boxplot of late delivery risk

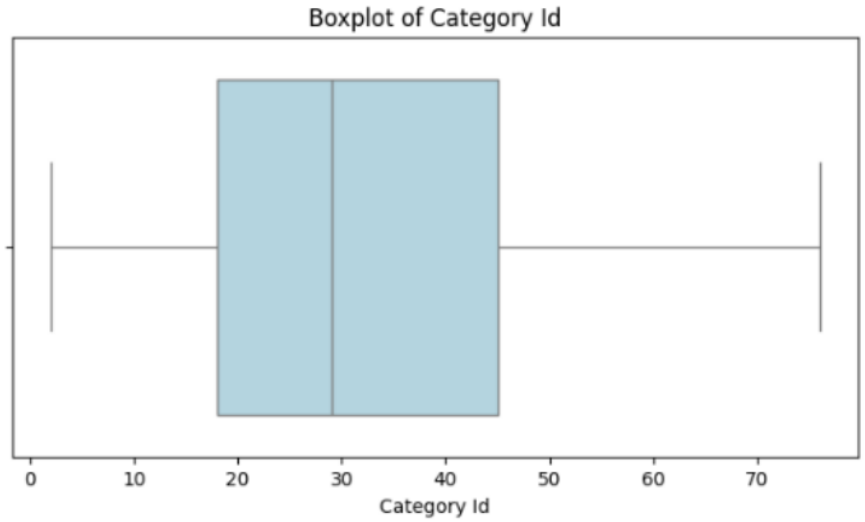

Figure 4.14.6: Result of boxplot of category id

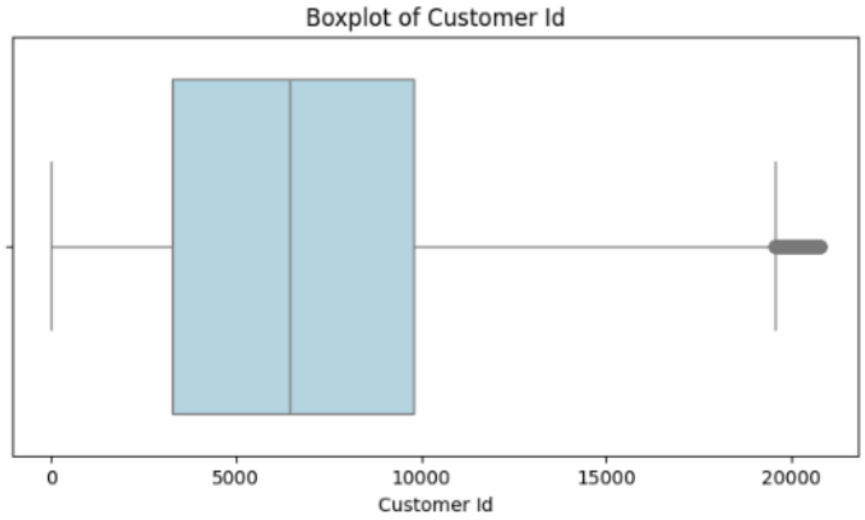

Figure 4.14.7: Result of boxplot of customer id

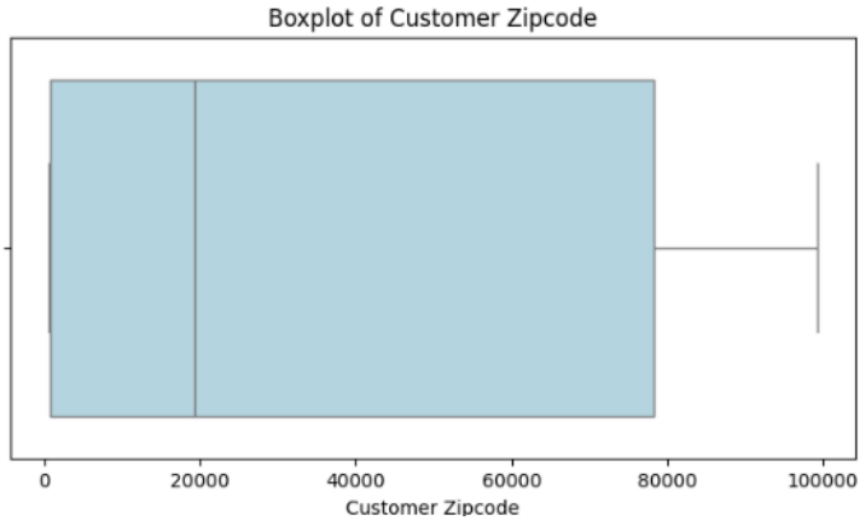

Figure 4.14.8: Result of boxplot of customer zipcode

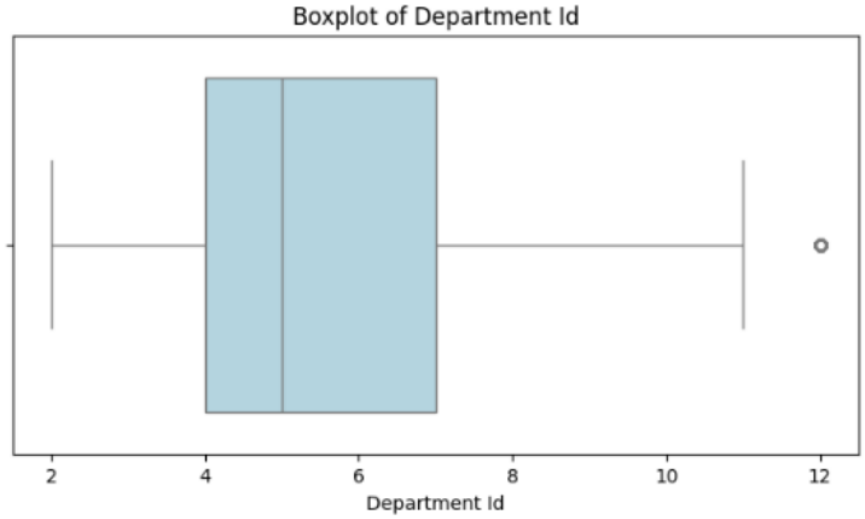

Figure 4.14.9: Result of boxplot of department id

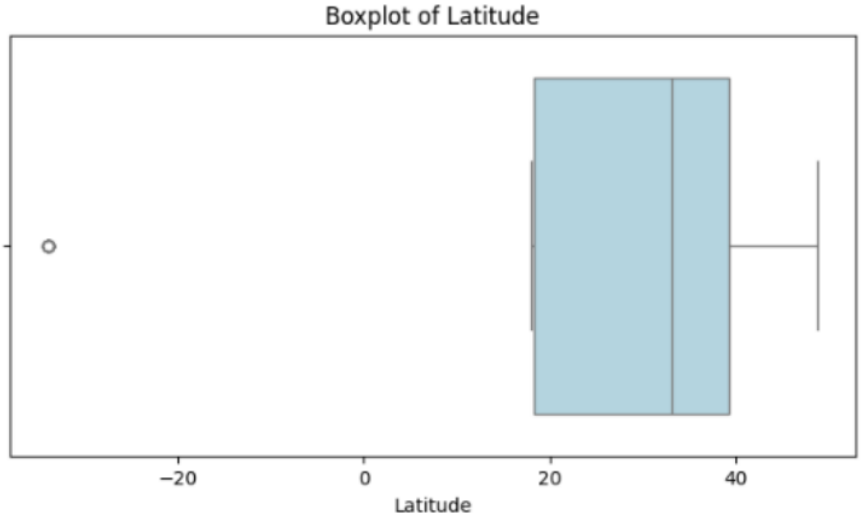

Figure 4.14.10: Result of boxplot of latitude

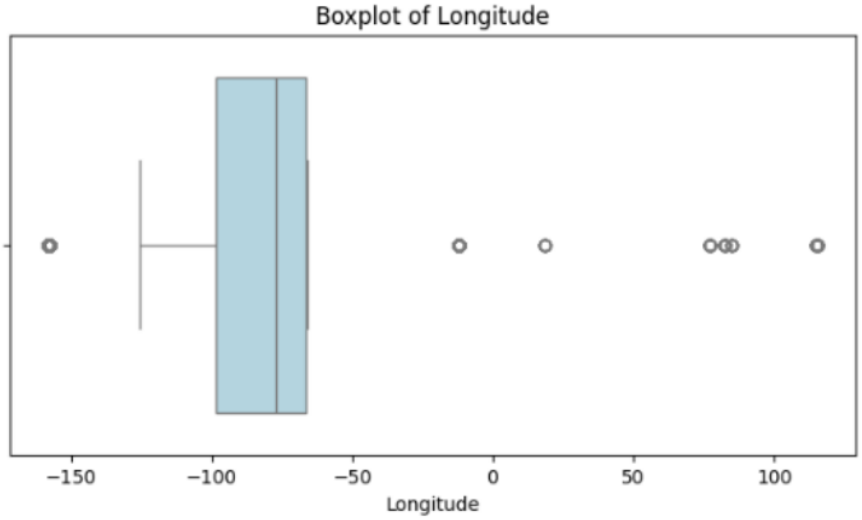

Figure 4.14.11: Result of boxplot of longitude

Figure 4.14.12: Result of boxplot of order customer id

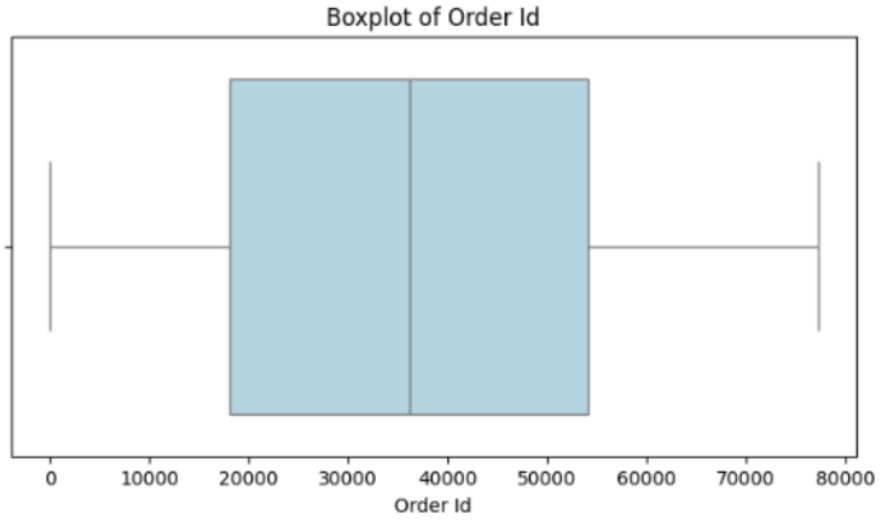

Figure 4.14.13: Result of boxplot of order id

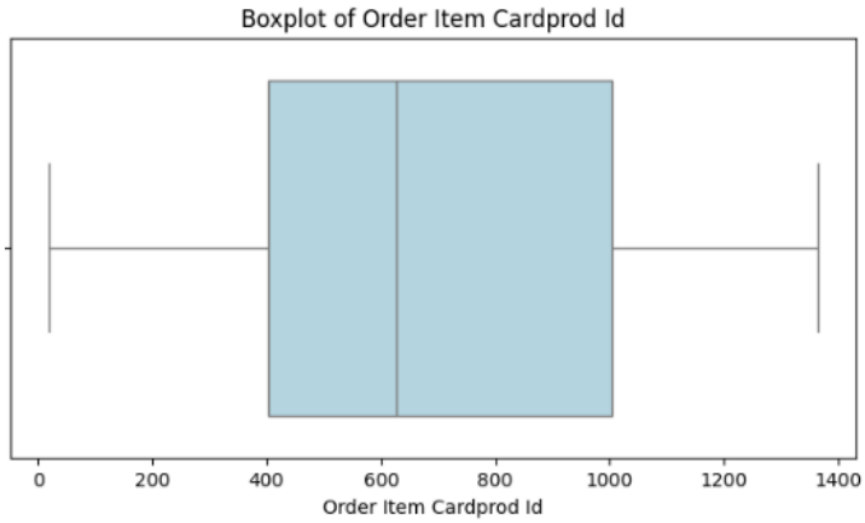

Figure 4.14.14: Result of boxplot of item cardprod id

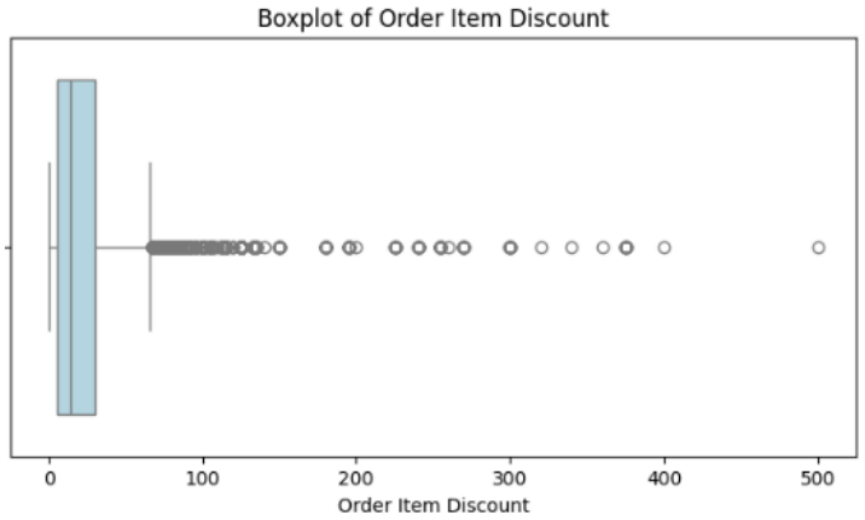

Figure 4.14.15: Result of boxplot of order item discount

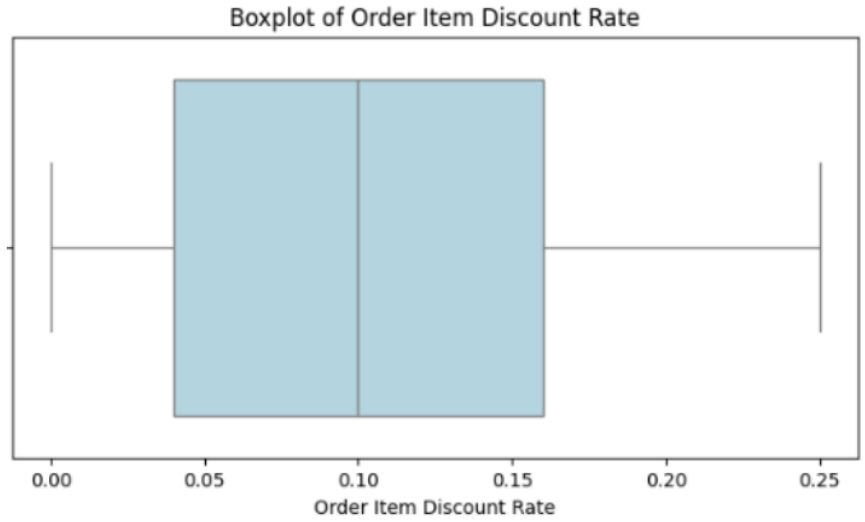

Figure 4.14.16: Result of boxplot of order item discount rate

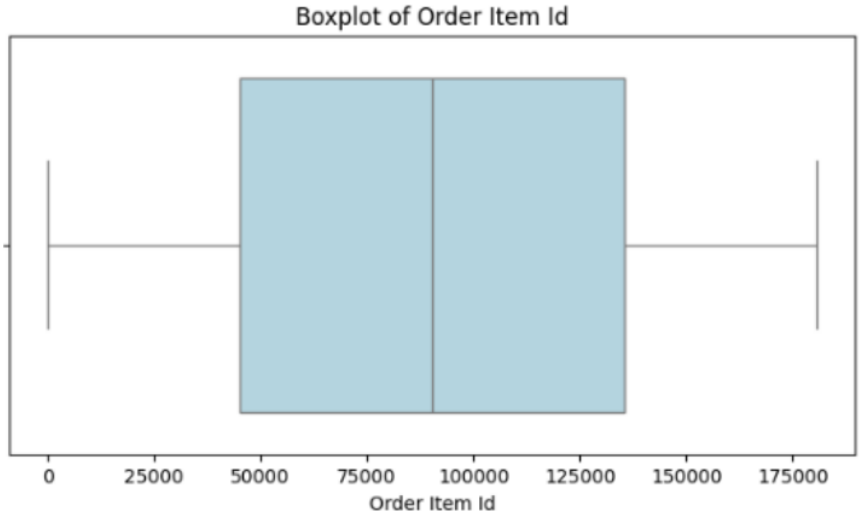

Figure 4.14.17: Result of boxplot of order item id

Figure 4.14.18: Result of boxplot of order item product price

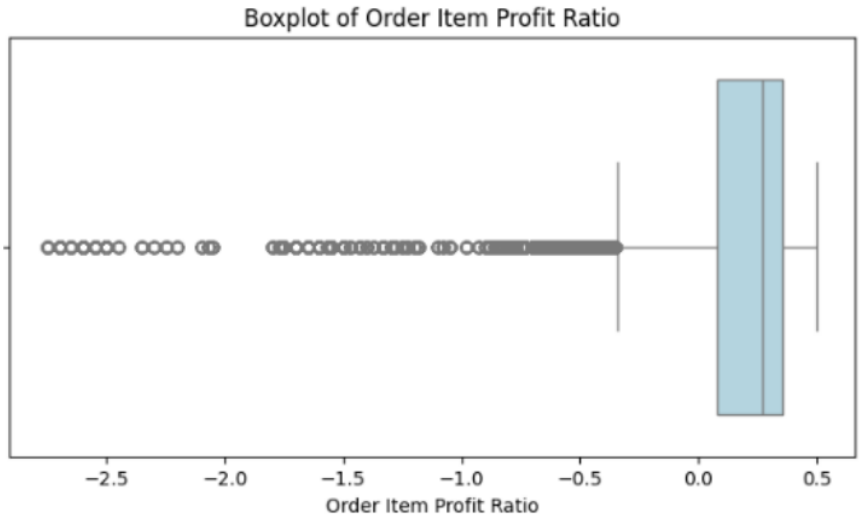

Figure 4.14.19: Result of boxplot of order item profit ratio

Figure 4.14.20: Result of boxplot of order item quality

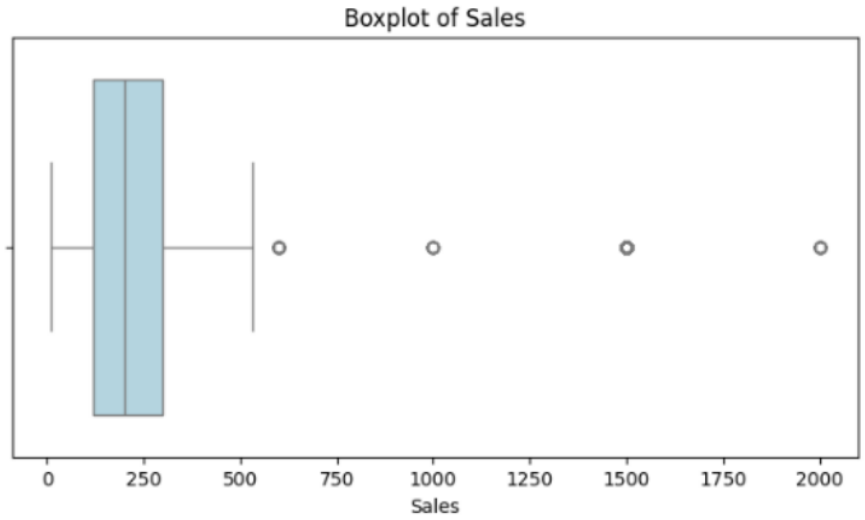

Figure 4.14.21: Result of boxplot of sales

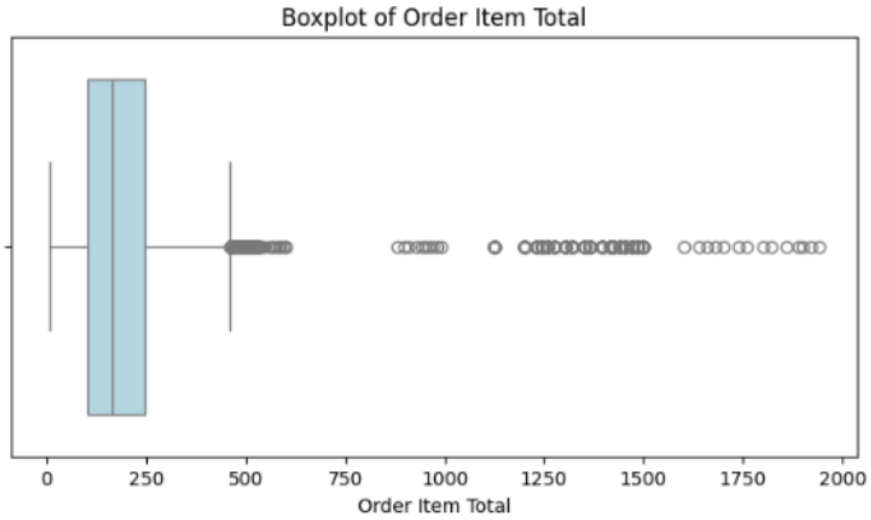

Figure 4.14.22: Result of boxplot of order item total

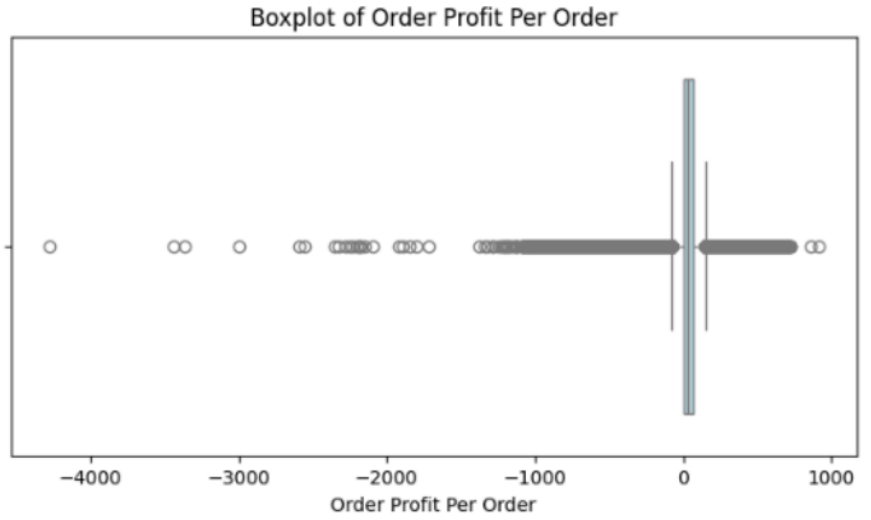

Figure 4.14.23: Result of boxplot of order profit per order

Figure 4.14.24: Result of boxplot of order zipcode

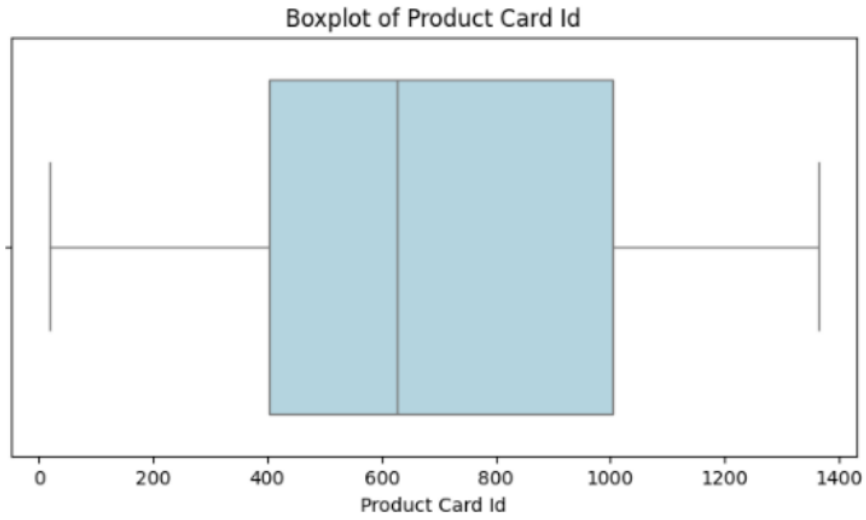

Figure 4.14.25: Result of boxplot of product card id

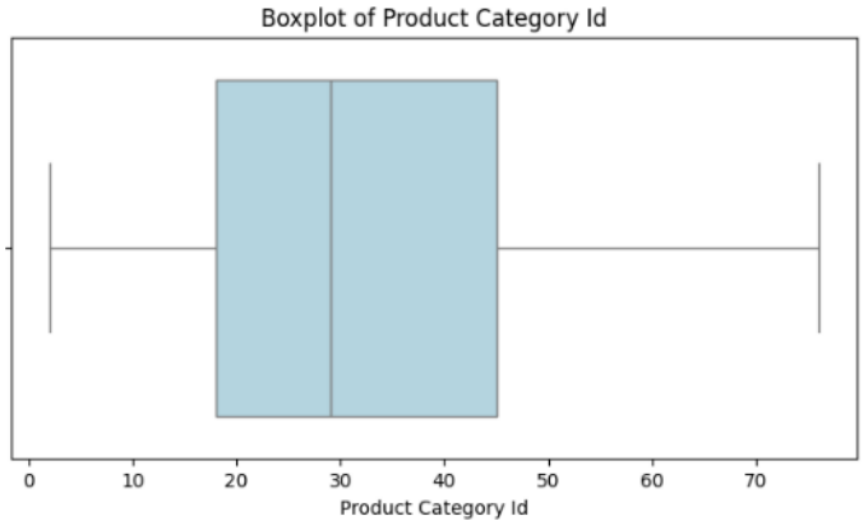

Figure 4.14.26: Result of boxplot of product category id

4.3 Chapter Summary

In this chapter, we conducted an exploratory data analysis (EDA) on the Sports Equipment Supply Chain Dataset, focusing on understanding the variables' structure, distributions, and relationships. The first step is observing the basic properties of the dataset, including the number of rows and columns and the data types. Missing values were identified and handled through imputation, ensuring that the dataset was ready for further analysis.

Furthermore, the distribution of numerical variables was examined, noting any skewness and kurtosis, and addressed the presence of outliers using boxplots and z-scores. The correlation analysis revealed significant relationships between certain variables, which may inform future modeling efforts. Additionally, identify the key patterns such as the dominance of certain categories in sales, providing actionable insights for business strategy.

The chapter concludes with the dataset being cleaned and prepared for further analysis, ensuring that the results will be both robust and reliable in subsequent steps of the project.

4.4 Reference

Brownlee, J. (2017). *Master machine learning algorithms: Discover how they work and implement them with Python*. Machine Learning Mastery.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

McKinney, W. (2010). *Data structures for statistical computing in Python*. In 9th Python in Science Conference (Vol. 451, pp. 51-56). https://doi.org/10.25080/Majora-92bf1922-00a

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., & Miles, A. (2020). Seaborn: statistical data visualization. *Journal of Open Source Software, 5*(51), 2446. https://doi.org/10.21105/joss.02446

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Research Outcomes

This research successfully developed a predictive modeling framework aimed at the reduction of inventory shortages and transportation delays in the sports equipment supply chain. By integrating XGBoost and LSTM models, the study demonstrated their effectiveness in analyzing historical supply chain data and generating highly accurate predictions. Key outcomes of the study include the identification of crucial factors influencing supply chain performance, such as seasonal demand fluctuations and supplier reliability, which play a significant role in determining the overall efficiency of the system. Additionally, the predictive framework was validated using various performance metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared, confirming the reliability and precision of the predictions. Another important outcome was the successful demonstration of the practical utility of data visualization tools, which enhance decision-making capabilities by providing real-time insights that allow supply chain managers to make informed decisions quickly.

5.2 Contributions to Knowledge

This research makes several significant contributions to the field of supply chain analytics and predictive modeling. It introduces a novel framework that effectively combines machine learning techniques, specifically XGBoost and LSTM, with traditional statistical methods to address complex challenges in supply chain management. This integrated approach offers a powerful toolset for predicting and optimizing supply chain performance, which is crucial in an increasingly dynamic and competitive business environment. The study also provides actionable insights into seasonal demand patterns, transportation delays, and inventory shortages, which are common challenges in the industry. One of the key advancements of this work is the application of LSTM models for time-series analysis, demonstrating their capability to predict supply chain trends and behavior over time. Furthermore, the research highlights the potential of data visualization in supply chain decision-making, as it enables businesses to proactively identify potential risks and take corrective actions before issues escalate. This combination of advanced predictive modeling and visual analytics paves the way for more effective and informed decision-making processes in supply chain management.

5.3 Future Works

While the outcomes of this research provide a solid foundation, several areas warrant further investigation to enhance the predictive framework and broaden its applicability. One key recommendation is the expansion of the dataset used in this study. Incorporating real-time data streams and integrating datasets from various industries could significantly improve the generalizability of the model, enabling it to address challenges across different supply chain contexts. Additionally, addressing data quality issues through advanced preprocessing techniques, such as data cleaning, normalization, and handling missing data, would enhance the robustness of the model. In terms of model enhancement, future research could explore the potential of ensemble learning methods, which combine the strengths of multiple machine learning models, to further improve the accuracy and reliability of predictions. Another promising direction isurni the investigation of reinforcement learning for dynamic supply chain optimization, as it could provide real-time decision-making capabilities based on the continuous evolution of supply chain conditions. To ensure the scalability and practical implementation of the predictive framework, future efforts should focus on developing cloud-based solutions that allow for real-time predictions and decision support. Collaboration with industry partners will be essential for testing and deploying the framework in real-world supply chain scenarios. Finally, ethical considerations such as data privacy and transparency in predictive analytics should continue to be prioritized, ensuring that the use of advanced data analytics adheres to regulations and promotes trust in the decision-making processes.

All Chapter_ Liew Yng Jeng.pdf

ORIGINALITY REPORT

12%	8%	5%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1%
2	Submitted to Federation University Student Paper	1%
3	www.mdpi.com Internet Source	1%
4	Submitted to University of Bradford Student Paper	1%
5	www.coursehero.com Internet Source	<1%
6	fastercapital.com Internet Source	<1%
7	arxiv.org Internet Source	<1%
8	Submitted to La Trobe University Student Paper	<1%
9	eprints.utm.my Internet Source	<1%

10	assets.researchsquare.com Internet Source	<1 %
11	conference.thaince.org Internet Source	<1 %
12	Submitted to Liberty University Student Paper	<1 %
13	link.springer.com Internet Source	<1 %
14	polscjournal.blogspot.com Internet Source	<1 %
15	Submitted to Queen's University of Belfast Student Paper	<1 %
16	core.ac.uk Internet Source	<1 %
17	ipfs.io Internet Source	<1 %
18	www.apacchrie.org Internet Source	<1 %
19	Submitted to Cardiff University Student Paper	<1 %
20	dspace.cvut.cz Internet Source	<1 %
21	Amir Shachar. "Introduction to Algogens", Open Science Framework, 2024	<1 %

Publication

22	Submitted to Rutgers University, New Brunswick Student Paper	<1 %
23	Yılmaz, Elanur. "Investigating the Quality of Teacher Reflection and the Nature of Changes in the Quality of Teacher Reflection Through Critical Incidents Analysis", Middle East Technical University (Turkey), 2024 Publication	<1 %
24	jeffreyutqi987630.pages10.com Internet Source	<1 %
25	www.tutorialspoint.com Internet Source	<1 %
26	Submitted to Aston University Student Paper	<1 %
27	Submitted to University of New South Wales Student Paper	<1 %
28	Submitted to Robert Kennedy College Student Paper	<1 %
29	Submitted to Capella University Student Paper	<1 %
30	Mutaz Wajeh Abdlmajid Qafisheh. "Establishing a Real-time Precise Point Positioning Early Warning System", Universitat Politecnica de Valencia, 2024	<1 %

Publication

31	Submitted to Otto-von-Guericke-Universität Magdeburg Student Paper	<1 %
32	Submitted to University of North Texas Student Paper	<1 %
33	repozitorij.uni-lj.si Internet Source	<1 %
34	eprints.kingston.ac.uk Internet Source	<1 %
35	Isabel Aparisi Cerdá. "Development of methodologies for energy planning of urban districts with just energy transition perspective.", Universitat Politecnica de Valencia, 2024 Publication	<1 %
36	Submitted to University of Dundee Student Paper	<1 %
37	Ho, Thu Minh. "Modeling Fertilizer Market Volatility and Trade Dependencies in a Complex Economic Era", North Carolina State University, 2024 Publication	<1 %
38	Parikshit N. Mahalle, Namrata N. Wasatkar, Gitanjali R. Shinde. "Data-Centric Artificial	<1 %

Intelligence for Multidisciplinary Applications", CRC Press, 2024
Publication

39	Vivek S. Sharma, Shubham Mahajan, Anand Nayyar, Amit Kant Pandit. "Deep Learning in Engineering, Energy and Finance - Principles and Applications", CRC Press, 2024 Publication	<1 %
40	commercee-pathshala.blogspot.com Internet Source	<1 %
41	de Almeida Azevedo, Rodrigo Miguel Guerra da Mota. "Address-Event Based Communication Between Spiking Neural Networks (SNN) Computing Cores", Universidade do Porto (Portugal), 2024 Publication	<1 %
42	jpit.az Internet Source	<1 %
43	myassignmenthelp.com Internet Source	<1 %
44	E. S. Mashkova, V. A. Molchanov. "Medium-energy ion scattering by solid surfaces part II", Radiation Effects, 1974 Publication	<1 %
45	José R. Zubizarreta, Elizabeth A. Stuart, Dylan S. Small, Paul R. Rosenbaum. "Handbook of	<1 %

	Matching and Weighting Adjustments for Causal Inference", CRC Press, 2023 Publication	
46	Küçükbahar, Duygu. "Modeling Monthly Electricity Demand in Turkey for 1990-2006.", Middle East Technical University (Turkey), 2024 Publication	<1 %
47	Laurent Seuront, Peter G. Strutton. "Handbook of Scaling Methods in Aquatic Ecology - Measurement, Analysis, Simulation", CRC Press, 2019 Publication	<1 %
48	Mostafa Hashem Sherif. "Handbook of Enterprise Integration", Auerbach Publications, 2019 Publication	<1 %
49	ebin.pub Internet Source	<1 %
50	ijisrt.com Internet Source	<1 %
51	iors.ir Internet Source	<1 %
52	jpc.in.net Internet Source	<1 %
53	lup.lub.lu.se Internet Source	<1 %

54	recerc.eu Internet Source	<1 %
55	sreview.soc.cas.cz Internet Source	<1 %
56	vdoc.pub Internet Source	<1 %
57	www.jatit.org Internet Source	<1 %
58	C. Guedes Soares, Tiago A. Santos. "Advances in Maritime Technology and Engineering - Celebrating 30 years of the Centre for Marine Technology and Ocean Engineering (CENTEC) Volume 2", CRC Press, 2024 Publication	<1 %
59	Cheng, Zhanhong. "Travel-Behavior-Based Inference and Forecasting Methods in Metro Systems", McGill University (Canada), 2023 Publication	<1 %
60	Nor Aida Abdul Rahman, T.C. Melewar, Pantea Foroudi, Suraksha Gupta. "Corporate Branding in Logistics and Transportation - Recent Developments and Emerging Issues", Routledge, 2024 Publication	<1 %
61	William M. Furey, Rex Forehand, John Baskin, Michael Tauber. "Mom, you're only	<1 %

remembering the bad behavior! the utility of distributed lag models in single-subject research", Journal of Psychopathology and Behavioral Assessment, 1986

Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		