AMERICAN UNIVERSITY OF BEIRUT

AUTOMATED INTEGRATION AND VALIDATION SYSTEM FOR E-COMMERCE AND ERP DATA AT MALIA GROUP

By

Romanos Rizk

Hadil Fares

Advisor

Dr. Sirine Taleb

A Capstone Project

submitted in partial fulfillment of the requirements

for the degree of Master's in Business Analytics

to the Suliman S. Olayan School of Business

at the American University of Beirut

Beirut, Lebanon

August 2024

Romanos Rizk & Hadil Fares for Master's in Business Analytics (MSBA)

Title: Automated Integration and Validation System for E-commerce and ERP Data

Abstract

This report presents a comprehensive approach to tackle business problems faced by Malia Group. Malia Group, a group of 25 companies in multiple sectors such as foods/goods distribution and technology solutions, is facing challenges in reconciling their stream of data coming from multiple sources related to their Cosmaline business, calling for the need of real-time monitoring of their financial transactions across sources. In addition to the reconciliation challenge, a thorough business analysis of the company's vast number of recorded transactions for their Cosmaline e-commerce business could provide the company with effective strategies to boost their performance.

To achieve these objectives, an automated Extract, Transform, and Load (ETL) system was developed using Apache Airflow to eliminate the need for manual entry into the system. To further enhance data integrity, reconciliation scripts that reveal discrepancies between data sources were embedded into the automated system. Finally, the reconciled data is stored in a permanent database storage system ready to be analyzed. Post reconciliation, an extensive Exploratory Data Analysis (EDA) and data mining analysis using models such as Apriori and Frequent Pattern Growth (FP-Growth) were developed to provide the company with actionable insights to boost their sales. The reconciliation results indicated that there is an integration problem between the transactions recorded in the e-commerce source and their counterpart in the Enterprise Resource Planning (ERP) source.

In addition, the discrepancy fluctuations between the e-commerce CyberSource and credit card data indicate that these two sources are being manually reconciled by the company, making the required changes on the database delayed month after the recording error occurred. As for the business analysis, the results showed that the company's customers prefer smaller, more diverse carts valued at less than $20. This creates an opportunity for the company to boost their low performing products (such as cosmetics) by offering free samples of them for customers that take out a cart valued at more than $20. Finally, the market basket analysis revealed important bundling and cross selling opportunities for the company, covering a wide range of bundles such as a Bundle for Colored Highlighted Hair, a Hijab Hair Care Bundle, a Kids Curly Hair Care Bundle, a Damaged Hair Care Bundle, and a Fall Control Hair Care Bundle that target a wide range of specialized segments.

To conclude, the project provides Malia Group with a robust set of tools to resolve the current data challenges they face. The project automates the entire business analysis pipeline from the ETL to the reconciliation and the drilled down analysis of relevant data. This comprehensive approach positions the company for short-term boosts in sale performance and for long-term expansion.

Key words: ETL, Automation, Reconciliation, Data Mining, Forecasting, Exploratory Data Analysis

Contents

1. Introduction

1.1 The Importance of E-Commerce Integration and Analysis

The current rapidly changing digital commerce environment requires companies to adopt state of the art technologies wired to improve operational efficiency, enhance customers' experience and generate out useful insights. Operating through e-commerce, which is given as a strategy in this endeavor, induces a need for the company in making connections among diverse systems that might include enterprise resource planning (ERP), customer relationships management (CRM), inventory management, and marketing automation. These integrations allow for the exchange of real-time data, process automation, and cohesion among various channel levels, leading to enhanced operational efficiency and a much more fulfilled experience for the customers.

The essentiality of robust e-commerce integration and advanced data analysis cannot be overstated, especially in industries experiencing rapid growth. Having more to the interpretation, the global industry of beauty, which yielded an income of $430 billion in 2020 in an approximate measure, is expected grow to $580 billion by 2027. This growth is driven by the climbing demand for personalized customer experiences and the imperative for organizations to swiftly adjust to evolving market conditions. These observations indicate that organization that can successfully incorporate, streamline, and evaluate their electronic commerce operation are more likely to be in an advantageous territory to maintain a competitive advantage and formulate strategic choices guided by data-driven insights.

1.2 Malia Group: A Multifaceted Conglomerate

The Malia Group: A Diversified Conglomerate is a distinguished union of small businesses that consists of 25 companies professionalized in multiple sectors such as foods/goods distribution and technology solutions. Malia Group has left a positive mark on economies of various regions, with a detectable presence in Lebanon, Iraq, and the UAE. The group's achievement has been based on originality, tenacity, and a dedication to superiority. For eighty years, the Malia Group has been through remarkable growth in its preserved collections of brands, reaching up to 60, and now has been driven to establish companionship with thirty-five foreign companies. These partnerships provide Malia with the capacity to fulfill a successful navigation in markets and to adapt to a newly evolving business environment.

Given the discussed rapid changes in the market, Malia Group's focus shifted to its personal care brand, Cosmaline, which has been capturing a wild amount of attention in Lebanon and other regions. Cosmaline operates in over 22 foreign and international markets, offering a variety of products that meet different care needs such as skincare, haircare, and body care needs of individuals of all ages and sex. The brand's dedication to excellence, inventiveness, and environmental responsibility can also be found in its sophisticated facilities localized in Lebanon, which have accredited with ISO 9001 and Cosmetics Good Manufacturing Practices (GMP) certifications. These facilities give Cosmaline the ability to uphold the utmost Norma in product formulation, concept innovation, and the support of marketing.

1.3 Challenges in Billing and Collection Processes

The predictable yet outstanding expansion of Malia Group's e-commerce processes has also revealed deeply rooted obstacles in its billing and collection operations. The organization is confronted with the challenging responsibility of maintaining data uprightness and stability across its systems, and that involves the transactions made daily from outnumbered data sources, including external partners such as Aramex. Currently, the orders made on the Cosmaline website are recorded in the company's enterprise resource planning (ERP). However, the collections received from shipping firms need to be manually entered and rearranged with their respective invoices in other sources. The manual process that is described in this context is not extensively time consuming and labor intensive, but also prone to errors. Such errors cause great discrepancies between data sources, which can negatively affect the outcome of financial reporting, decision-making operations, and the analysis of e-commerce services.

Given the data environment complexity and volume of transactions recorded by Malia, the risk of errors and discrepancies is high, making the current process unsustainable in the long term. The purpose of the Malia Group is to confront these challenges through an automated system that can be merged with the existing ERP infrastructure. This system will optimize and automate the operation of data entry, validation, and reconciliation. Implementing such systems would guarantee precise recording and real-time monitoring of all financial data, thereby reducing the number of errors and elevating the company's ability to make decisions accordingly. Furthermore, this project's initiative is not just to merge data, but also to make use of the organized and polished data for comprehensive analysis, which will give the company practical insights into its e-commerce business performance.

1.4 Adopting Apache Airflow for Automation and Analysis

SQL Server Integration Services (SSIS) has historically been a widely used tool for implementing ETL (Extract, Transform, Load) procedures, especially in Microsoft systems. However, the restrictions of SSIS, which might include its dependency on OLAP cubes, and the significant developer engagement needed for its maintenance hindered its effectiveness in today's fast-paced and ever-changing data environments. This change in data environments compelled the development of more flexible and adaptable tools. Airflow, created by Airbnb in 2015, is a Python-based open-source workflow orchestrator that offers great flexibility, scalability, and integration capabilities. The previous mention of the Airflow features makes it a perfect match for managing the diversity and the data sources used in the operations of Malia Group.

The Directed Acyclic Graph (DAG) structure of Airflow provides a clear representation of task connections and execution sequence. Also, the usage of Python, a well-known programming language for its simplicity ease of use, and vast collection of libraries makes it easier for developers to create workflows as code, giving them full control over their logic and use cases. In addition, Airflow's capacity to run task in parallel with numerous schedulers guarantees great performance and scalability, rendering it a matched choice for automating Malia's billing and collection procedures.

To add more to the integration and automation aspects, this project also focuses on the analysis of Malia Group's e-commerce performance. Through the use of the purified and reconcile data, the project will include a comprehensive analysis that envelops predicting sales patterns, conducting a detailed exploration of e-commerce transactions, and performing a Market Basket Analysis (MBA) to reveal customer purchasing patterns. The objective of this analysis is to supply the Malia Group with experiential insights that can help with the development of numerical marketing strategies, improve the products' positions, and augment sales. The integration of these predictive and real-time monitoring tools would undoubtedly improve the organization's chances and capabilities to apprehend and address trends and challenges that are taking place in the marketplace.

1.5 Project Objectives and Broader Implications

The goals of this project are the automation of data entry and reconciliation, the improvement of data precision and integrity, the reduction of manual data entry, and the enhancement of decision-making throughout the usage of purified and reconciled data. Through the integration of real time data monitoring tools, in addition to the extensive analysis of e-commerce activities, the projects aim to optimize operational efficiency and provide Malia Group with the necessary insights to improve sales performance and consolidate their overall business plan.

Furthermore, this project not only highlights the challenges that the Malia Group face, but also carries broader implications for the industry. The utilization of Apache Airflow in this specific application added up to the significant development in the territory of e-commerce of automation, sorting out an expandable and flexible solution that can be also used by various organizations that are coming up against comparable obstacles. Also, the compatibility of Airflow with multiple tools and integration software ensures that the investments made by the Malia Group in this project will have an extensive effect in the long run, especially when the company decided on integrating new services or data sources with their current system. As for the project's analytical insights, they will assist the company with further lore capabilities to come up with data-driven decisions, thereby enhancing its competitive positioning in the market.

Ultimately, this project is not only tackle Malia's current operational challenges but also set the organization up for future growth and adaptability in today's rapidly evolving digital commerce world. By using Apache Airflow and taking a proactive approach to integrating, automating, and analyzing data, Malia Group is creating a new standard for efficiency, accuracy, and scalability in the e-commerce industry.

2. Background and Related Work

2.1 Automating ETL Processes and Data Integration

2.1.1 Introduction to ETL Automation

The automation of Extract, Transform, and Load (ETL) processes are an essential part of every organization, particularly in large data environments that involve diverse sources such as a company's website, ERP systems, and shipping databases. By automating these ETL processes, the company would eliminate any manual data entry to their system, minimizing the chance of errors, and ensuring timely data flow across systems. Automating such processes would also resolve data conflicts between sources, making them suitable for reporting and analysis. A study made by Bakhtouchi (2022) emphasized that automating ETL processes is essential for resolving data conflicts at both the schema and instance levels. This data resolution is done by using Reference reconciliation techniques such as comparing data attributes, using similarity measures, and identifying and merging duplicate records. While some industries lack behind when it comes to their data integration, the need for automated ETL processes has been well-documented in the literature. Gour et al. (2010) discuss the role of the ETL architect in developing these processes. The study mentioned that aside from the proper data extractions, the data transformation should primarily be guided by the company's specific business rules, and the process should be rigorously documented. The study also points to the use of indexing and staging platforms to improve data retrieval speeds and reduce load times while also optimizing system memory usage. Despite these advancements in automation, challenges in developing ETL processes still exist, particularly when ensuring data quality. Woodall et al. (2016) highlights several common challenges in ETL pipelines that are primarily related to data being inaccessible, causing delays in data delivery as well as data transformation errors that often trigger the system to a hard stop. Such issues can severely affect the timeliness and accuracy of financial data, leading to incorrect decision-making and to additional loss in time due to inconsistent data.

The literature mentioned highlights the need for not only well designing and documenting the ETL process, but also to choose the appropriate tools to develop such wide scaled systems.

2.1.2 Evolution of ETL Tools: From Traditional Software to Web-Based Solutions

In the early days of software automation, and with the dominance of proprietary software with a lack of third-party compatibility, companies in need of an automated ETL system had to resort to tools like Microsoft SQL Server Integration Services (SSIS). SSIS, with its user-friendly drag-and-drop interface, made it accessible for companies with little programming expertise to build their ETL workflows and integrate them with their Microsoft based systems. However, with the emergence of multiple enterprise software manufactures such as Oracle, SAP, and many open-source software, robust yet rigid tools such as SSIS often lack the flexibility required to process rapidly changing and diverse data sources especially the ones sourcing from data streaming APIs. As a result, the data engineering industry is slowly catching up to these rapid changes. As highlighted by V. R Krishna et al. (2010), there has been a shift towards web-based ETL tools that offers several advantages over traditional proprietary tools, including greater customizability of

workflows, improved scalability for larger volumes of data, and much better integration capabilities with cloud services, external APIs, and other web-based applications.

Based on the literature, Apache Airflow was decided as the automation tool of choice for the project. Apache Airflow combines the best of both traditional software tools like SSIS and modern web-based ETL solutions as it offers complex automation capabilities while allowing for customized preprocessing of data through python scripting.

2.1.3 Apache Airflow for ETL Automation:

Apache Airflow is a comprehensive orchestration tool used to manage complex ETL pipelines. Airflow's logic is based on an innovative DAG structure that allows users to define precise dependencies and order of task. This implementation ensures that data is extracted, transformed, and loaded successfully and efficiently in its right order as it allows nondependent processes to run in parallel. Furthermore, Airflow's web based yet easy to use architecture enables integration with a wide range of data sources, including e-commerce platforms, ERP systems, and shipping databases, making it an ideal solution for the Malia Project. To further justify the use of Airflow, Wu et al. (2023) indicated how the use of Airflow, in combination with Docker, are great tools to automate ETL systems, especially when incorporating machine learning models and predictive analytical tools to the system. Additionally, Vase and Tuomas (2015) wrote about the various advantages of using docker for such systems. They emphasized docker's main benefit in portability as it allows the same application to run smoothly across different environments regardless of difference in dependencies. Docker is also less resource demanding on the system compared to other software as the containers that host the applications are lightweight and optimized for low resource computing.

2.2 Data Quality and Reconciliation in ETL Pipelines

An effective automated ETL process needs to not only ensure correct data formats, but also ensure that the data present in multiple sources is reconciled and of high quality. This is particularly true in environments with multiple sources like e-commerce platforms, ERP systems, and shipping databases that store the same transactions in different formats. Xing (2021) stressed the need for robust reconciliation methods that identify discrepancies early in the storing process, as poor data quality leads to inaccurate analysis and decision making. Xing's study covered a combination of offline data reconciliation and incremental data reconciliations through the live pipeline that would prevent capital losses between financial operations present in different sources. The paper also highlights that the processing time of the reconciliation process is as important as its efficacity, since slight delays in reconciliations can lead to significant financial risks. His study addressed this challenge by adopting quasi-real-time reconciliation models that run in a distributed fashion for faster and more efficient computing. The research also utilized memory efficient techniques such as segmenting data into partitions and performing incremental reconciliation, which allowed the system to identify discrepancies sooner. Additionally, Onuoha and Amponsah (2012) emphasized the importance of reconciliation in a bank setting. They indicated that using audit-based bank reconciliation measures prevents incidents such as unauthorized funds leakage.

Despite significant coverage in reconciliation techniques, a gap is still present in the literature. Current research still lacks comprehensive studies that integrate large-scale automated ETL processes that handle financial reconciliation across different sources such as e-commerce and ERP platforms. While Xing's (2021) research provides methods to reconcile transactions in distributed financial systems, few studies have addressed the challenges of reconciling e-commerce transactions with ERP and shipping data. The Malia Project tailors such methods by deploying an automated reconciliation system that handles discrepancies between their e-commerce, ERP, and shipping data source.

2.3 Market Basket Analysis and Forecasting in E-Commerce

Market Basket Analysis (MBA) is a data mining technique that is widely used to uncover patterns in e-commerce transactions. This is done through the identification of association rules that includes different frequently bought together products. Kurnia et al. (2019) analyzed transaction of the O! Fish restaurants' transactional database using the Apriori algorithm to uncover frequently bought together menu items. Similarly, the Apriori algorithm was also used by Panjaitan et al. (2019) to optimize menu choices at Café Bojack Coffee Shop and develop promotional strategies to boost sales. Moving to the retail industry, MBA can be used for implementing targeted product recommendations. Qisman (2021) applied the Apriori algorithm to a retail store to identify frequently occurring itemsets, which allowed stakeholders to promote effective and sales boosting bundles. Similar to Aprioi, FP-Growth is another widely used algorithm for data mining exercises. Anggraeni et al. (2019) revealed how both algorithms were effective in discovering patterns in customer purchasing tendencies. The study proved that while Apriori provides deeper insights, FP-Growth trumps in processing speed and efficiency. In addition, similar results were yielded when the two models were compared by Anas et al. (2021). Forecasting models are also widely used in e-commerce transaction analysis, precisely to predict future sales trends. For example, Zhao et al. (2020) implemented the ARIMA model alongside machine learning models to forecast e-commerce sales trends. The study showed that including parameters such as seasonality greatly increases the model's ability to catch fluctuations in demand. A similar approach of integrating ARIMA with big data analytics was used by Yang et al. (2021) to forecast product prices and adjust the company's marketing strategy.

Considering that the Malia Project specifically targets their Cosmaline e-commerce performance in their analysis, both Market Basket Analysis (MBA) and forecasting models were employed to provide valuable insights about the company's cosmetic business. While the current literature covers a multitude of combinations of business analytic models such as MBA and Forecasting, few research is done around integrating these models into an automated ETL architecture. This implementation would ensure consistent and scalable predictive capabilities for the company, allowing them to access long term insights and recommendations about their e-commerce performance.

3. Methodology

3.1 Overview

Malia's Project main goal is to improve their operation's data integrity by creating an automated system for billing and collection processed for their Cosmaline e-commerce platform. The technology will be built to align the company's Enterprise Resource Planning (ERP) system with its e-commerce platform transactions, and the collection data sourcing from shipping companies. Thus, the project's objective is to provide Malia's stakeholders with real-time monitoring of reconciled financial activities and actionable insights to improve their Cosmaline e-commerce business. Below are the problems and objectives that the Malia project will address:

2.1.1 Errors related manual data input:

Presently, while orders placed on the Cosmaline website are recorded within the ERP, billing collections received from shipping companies require manual entry into the system. Considering the high daily input of transactions, these manual processes greatly increase the data integrity errors between different sources. This issue will be addressed with the development of an automated Extract, Transform, Load (ETL) system that will extract the data from all the present sources, convert them into appropriate format, reconcile them according to business rules, and then stores them into a permanent database for monitoring and analysis.

2.1.2 Data Reconciliation:

In addition to the manual entry system, the multitude of data sources between the ERP system, Cosmaline platform, and external Aramex caused misalignment between the same transactions when present in more than one source. To ensure that all transactions are automatically screened, various SQL scripts were developed to verify transactions on predefined business rules, detecting both matches and mismatches for each transaction.

2.1.3 Data Monitoring:

In order to provide real time monitoring of all transactions across sources, A Power Bi dashboard was created to display the data integrity status of Malia e-commerce business. The dashboard comprises four comprehensive reports that display insights into reconciliation status between the different sources, as well as transaction level detail capabilities that would enable stakeholders to track individual transactions if need be.

2.1.4 Model Building:

In addition to providing an automated ETL and reconciliation system for Malia, the project also entails using the reconciled data to provide actionable insights that improve their e-commerce performance. To achieve this solution, forecasting models will be employed to forecast future sales of their e-commerce platform. After conducting this trend analysis, data mining models, specifically a market basket analysis, will be conducted to provide the company with insights concerning customer purchasing patterns and the hidden association between their different products. This valuable information helps Malia Group in creating successful cross-selling, bundling, and marketing initiatives.

To establish a robust and scalable system, Docker, an open-source platform that enables developers to build, deploy, run, update and manage containers, was used as host of the Airflow system. Docker enables uniformity of software components throughout development and testing, which is essential when deploying the system. In addition, the database used for storing was MySQL, while data manipulation and modeling were performed using Python and R. As for the monitoring dashboard, Power Bi was selected for its seamless integration with Microsoft Services.

3.2 Objective 1: Automating the Billing and Collection Processes

3.2.1 Data Collection and Sources

There are 8 main data sources related to Cosmaline's e-commerce business:

1. **ECOM Data Website:** 10,389 rows and 9 columns

   This data source stores the transactions made on the Cosmaline website. The source provides details such as order numbers, shipper names, dates of creation and delivery, forms of billing, monetary amounts, currencies, nations, and airway bills (AWB).

2. **Shipped & Collected – Aramex:** 8,810 rows and 8 columns

   This data source holds information about the shipments managed by Aramex. The details provided consist of shipper numbers, HAWB (House Air Waybill) numbers, delivery dates, COD (Cash on Delivery) amounts, and invoice dates.

3. **Shipped & Collected – Cosmlaine:** 1,638 rows and 4 columns

   This data source holds information about the shipments managed by Cosmaline. The information provided by this source includes the shipping records handled by Cosmaline, which consist of shipper numbers, HAWB numbers, delivery dates, COD amounts, and invoice dates.

4. **Collected – Credit Card:** 2,205 rows and 4 columns

   This data source monitors credit card payments, and includes information such as order numbers, payment amounts, and payment dates.

5. **ERP-Oracle Collection:** 124 rows and 9 columns

   This data source contains the representation of e-commerce transactions in the company's ERP system. It includes distilled information such as the receipt number, currency code, exchange rate, customer number and name, receipt class, and a comment column indicating the source of the collection whether it was from Cosmaline, Aramex, or a Cyber source (credit card).

6. **Oracle Data:** 21,332 rows and 13 columns

   This data source contains the full representation of an order placed on the e-commerce website. It represents each order cart by spanning it over multiple rows, with each row indicating the specific product of that cart. The source provides a detailed look at the transactions with its columns being: operating unit name, the order number referencing the oracle source, the order

number referencing the ECOM source, order type, order type name, the ordered item and its quantity. The source also includes information such as the unit of measure for the quantity column, unit selling price after discount without VAT, unit list price, ordered date, customer, and the Tax code.

7. **Daily Rate:** 117 rows and 2 columns:

   This data source records the daily Lebanese pound currency rate, which essential for currency conversions in the context of the fluctuating Lebanese pound exchange rate vis-à-vis the dollar amidst the economic crisis of 2019. The data source includes the date and its corresponding LBP to USD exchange rate.

8. **Oracle Product Names:** 639 rows and 3 columns:

   This data source was populated by scraping the Cosmaline website. It consists of the names and categories of products that match the product IDs listed in the Oracle data source.

All datasets, with the exception of the Oracle Product Names, were supplied in Excel format and imported into the system from a local directory. The Oracle Product Names dataset was acquired by scraping the Cosmaline website

3.2.2 Data Preprocessing and Cleaning

A crucial part of the ETL process is the transformation process. All the data sources mentioned in the prior section were subject to numerous transformation methods, making them suitable for permanent storage. This section will detail the preprocessing steps applied for each dataset and a diagram (P.1) that visually represents the processing pipeline can be found in the appendix section:

A. ECOM Website Data Preprocessing

The ECOM website data was processed using the following steps:

1. **Converting Data Types**: 'Order Number', 'Shipper Name', 'Created At', and others, were converted to datatypes similar to their respective tables in the permanent database
2. **Normalizing Text Columns**:
   - **Lowercase Normalization**: Columns such as 'Shipper Name' and 'Billing Type' were lowercased, and the first letter of each word was capitalized. In addition, white spaces were trimmed to avoid similar text to be identified as different.
   - **Uppercase Normalization**: Columns such as 'Currency' and 'Country' were uppercased, and white spaces were trimmed to maintain consistency in these metrics across all sources.
3. **Removing Missing Values**: Rows containing missing values were removed from the dataset to ensure that any analysis performed would be solely based on complete data entries.

4. **Dropping Duplicates**: Duplicate rows were removed from the dataset to ensure that no transaction is repeated in the dataset. This is also done to prevent primary key integrity infringement when loading the data into the database.

5. **Renaming Columns**: Some columns were renamed for smoother compatibility with the database and to match its schema. For example, 'Order Number' was renamed to 'order_number', 'Shipper Name' to 'shipper_name', and so on.

B. Aramex Shipping Data Preprocessing

The Aramex Shipping data was processed using the following steps:

1. **Converting Data Types**: Relevant columns, such as 'AWB', 'Delivery_Date', 'CODAmount', and others, were converted to datatypes similar to their respective tables in the permanent database

2. **Standardizing and Renaming Columns**: The 'HAWB' column was standardized and renamed to 'AWB' for consistency across data sources and database tables

3. **Normalizing Text Columns**: The 'CODCurrency' column was uppercased to maintain consistent text formatting.

4. **Updating CODCurrency**: The 'CODCurrency' column was updated to display 'LBP' when the 'CODAmount' was greater than 0 as the ‘CODAmount’ would only include Lebanese currency amounts.

5. **Cleaning Tokens by Date**: The 'TOKENNO' column was cleaned to ensure that all records across the same date had the same token.

6. **Adding Order Number**: The 'order_number' column was added to the Aramex shipping data by joining the 'AWB' column with its corresponding counterpart in the ECOM data. This column was added for the purpose of making the Aramex Data ready to be joined with the Cosmaline shipping data as this would streamline all shipping information into one comprehensive table

7. **Converting Data Types Again:** A post transformation data type conversion was applied again to match all columns with the joined shipping data for Aramex and Cosmaline present in the databse.

C. Cosmaline Data Preprocessing

The Cosmaline Shipping data was processed using the following steps:

1. **Converting Data Types**: Relevant columns, such as 'Driver_Delivery_date', 'Amount', and 'OrderNo', were converted to datatypes similar to their respective tables in the permanent database.

2. **Normalizing Text Columns**: The 'Currency' column was uppercased, and its right and left whitespace was trimmed to avoid similar text to be identified as different.

3. **Renaming Columns**: Columns such as 'Driver_Delivery_date' was renamed to 'Delivery_Date', 'OrderNo' to 'order_number', and so on to ensure smooth insertion into the database's permanent tables.

4. **Adding CODAmount**: A new column 'CODAmount' was added based on the 'CODCurrency' column. When the currency is set 'USD', the CODAmount was set to 0, and if it displayed 'LBP', the value in 'O_CODAmount' was assigned. This transformation ensured that no COD amounts columns was used to store wrong currency amounts.

5. **Add New Columns**: Three new columns were added to make the Cosmaline table similar in structure to the Aramex table as the two will be concatenated to streamline all shipping transactions into one table:

    - 'ShprNo': Set as blank
    - 'Aramex_Inv_date': Set as blank
    - 'TOKENNO': Set to 'Shipped With Cosmaline'

6. **Merging AWB from ECOM Data**: The 'AWB' column from the e-commerce data was added to the Cosmaline data based on matching their 'order_number'.

8. **Final Data Type Conversion**: A post transformation data type conversion was applied again to match all columns with the joined shipping data for Aramex and Cosmaline present in the database.

D. Credit Card Data Preprocessing

The credit card data was processed using the following steps:

1. **Uppercasing Text Columns**: The 'Narrative' and 'Currency' columns were uppercased, and any extra whitespace was stripped to ensure consistency across semantically similar strings.

2. **Converting Data Types**: Relevant columns, such as 'Value_date', 'Narrative', 'Amount', and 'Currency', were converted to datatypes similar to their respective permanent tables.

3. **Lowercasing Column Names**: All column names were lowercased to adhere to the naming convention of database attributes of having lowercased column names.

E. ERP Data Preprocessing

The ERP data was processed using the following steps:

1. **Converting Data Types**: The 'CUSTOMER_NUMBER' column was converted to an object type to ensure that the customer numbers are handled as categorical data, which avoids any unintended aggregation in future analysis

2. **Normalize Text Columns**:
    - **Uppercase Normalization**: The 'CURRENCY_CODE' column was uppercased to maintain consistency with the other tables
    - **Lowercasing Normalization**: Columns such as 'CUSTOMER_NAME' and 'RECEIPT_CLASS' were lowercased while the first letter of each word was capitalized. Also, the left and right whitespace were trimmed. This was done to unify the format of these columns.
3. **Extract Tokens from Comments**: The token number was extracted from the 'COMMENTS' column using a regular expression pattern. If a token was present, it was extracted and put as a number in its respective row; otherwise, the original comment was retained. This extraction was performed to locate each token when reconciliation of multiple sources are performed.

F. Oracle Data Preprocessing

The Oracle data was processed using the following steps:

1. **Remove Time from Date Column**: The time portion in the 'ORDERED_DATE' column was removed, only keeping the date information. This matched the date column to its counterpart in the database as the time is not typically recorded in Malia's system.
2. **Rename Columns**: Several columns were renamed to match the permanent database. This included renaming columns like 'OPERATING_UNIT_NAME' to 'operating_unit_name', 'ECOM_REFERENCE_ORDER_NUMBER' to 'ecom_reference_order_number', and others.
3. **Convert Data Types**: Relevant columns such as 'oracle_reference_order_number', 'ordered_quantity', and 'unit_selling_price', were converted to their respective data types in the permanent database.

3.2.3 ETL System Architecture:

Malia's system architecture is specifically designed to host the ETL process essential for its operations. In addition to efficacity, the system should be efficient enough to handle automated scenarios of constant inflow of data using the least amount of memory possible. In this section, a comprehensive overview of the many stages involved in the automated ETL process is provided. An illustration of the system is also provided in the appendix section.

A. Overview of the Extract, Transform, Load (ETL) process:

1. **Extract from Excel:** Data is initially retrieved from Excel files to replicate real-time API data. The data of each data source is then put into a python environment, with each data frame representing a different data source.

2. **Write Data Frame as CSV:** Each data frame is exported as a csv file on the local system. This ensures that the data is not constantly stored in memory, but rather saved on storage and retrieved independently for transformation.
3. **Extract from CSV:** Each data source is imported into a python environment from its corresponding CSV file.
4. **Transform Data:** In Python, each data source is transformed using custom preprocessing pipelines. (as explained in the previous section on Data Preprocessing and Cleaning).
5. **Convert to Parquet:** Once the data has been transformed, it is saved in Parquet format on the local system. The main reason for using parquet files for storage is their highly efficient storage format that organizes data in columns, enabling faster querying. Most importantly, Parquet, contrary to CSV retains the precise metadata of the source which includes the data types allocated during the transformation. This guarantees that the datatypes set in the transformation will not be altered when the data is imported again from the local machine, which is essential for an error free loading into the database.
6. **Extract From Parquet:** The Parquet files are extracted and loaded into a python environment as data frames.
7. **Load to Temporary Tables:** The extracted data frames from the parquet files are loaded into temporary tables in the MySQL database.
8. **Compare and Load into Permanent Tables:** To ensure that no duplicates are transferred into the database and per the industry's best practice, temporary tables containing current instream of data are compared to the permanent tables and only non-duplicated entries are loaded to permanent storage.
9. **Execute Reconciliation:** Reconciliation procedures are conducted to verify the consistency of data across various sources. This stage entails cross-referencing and comparing similar transactions present in several sources, guaranteeing their similarity in all the tables.
10. **Delete Temporary Tables:** Once the data has been properly imported into the permanent tables, the temporary tables are purged to release resources and maintain the efficiency of the ETL process.

The diagram (D.1), which can be found in the appendix, presents a graphical depiction of the flow of data from the extraction process to the reconciliation process

B. Database Schema:

The database schema for the Malia Project is designed to store and oversee data from all sources.

Core tables:

- **ecom_orders:** This table contains data pertaining to e-commerce orders, including the order number, shipper name, creation and delivery dates, billing method, amount, currency, and country. The 'order_number' serves as the primary key and the table additionally incorporates Air Waybill (AWB) which is used to establish a connection between the orders and shipping information.
- **credit_card:** This table records credit card payments linked to e-commerce orders. Due to a lack of a natural identifier in the source data, an auto increment id field is assigned as the table's primary key.
- **daily_rate:** This table stores the daily exchange rates that are necessary for converting transaction amounts into USD. Similar to the 'credit_card' table, an auto increment id field is assigned as the table's primary key.
- **erp_data:** This table contains data extracted from the company's ERP system, including receipt numbers, currency codes, exchange rates, receipt dates, customer details, and reception amounts. The 'receipt_number' column serves as the primary key, ensuring that each receipt is uniquely identified.
- **oracle_data:** This table contains the Oracle ERP data at the order basket level, including details such as the operating unit name, order numbers, order kinds, item descriptions, price, and tax information. Due to a lack of a natural identifier in the source data, an auto increment id field is assigned as the table's primary key. Also, this table is connected to the 'ecom_orders' and 'oracle_data_product_name' tables through the 'order_number' and product_id as foreign keys to the two other tables respectively.
- **shippedandcollected_aramex_cosmaline:** this table merges shipment data from both Aramex and Cosmaline with AWB serving as primary key. In addition, this table is connected to the 'ecom_orders' table by their 'order_number''columns.

Supporting Tables:

- **oracle_data_product_name:** This table contains product data obtained by scraping the Cosmaline website. The 'oracle_data' table refers to this table through the 'product_id' column in order to offer more information for ordered items.
- **reconciliation_results:** This table is used to store the outcomes of reconciliation procedures and includes information such as the kind of reconciliation, reference IDs, ERP and shipping quantities, and the current status of the reconciliation.
- **fpgrowth_results**: This table stores the outcomes derived from the FP-Growth algorithm used in the market basket analysis. The data set comprises categories for antecedents, consequents, support, confidence, lift, and other important metrics.
- **apriori_results:** This table, similar to the 'fpgrowth_results' one, holds the outcomes of the Apriori algorithm.

Additionally, a schema diagram is included in the appendix to graphically depict the relationships and structures of the tables (Figure Schema.1).

C. Workflow Orchestration with Apache Airflow and System Scalability

Apache Airflow is employed as the workflow orchestration technology to effectively manage and automate Malia's ETL operations. The Airflow's Directed Acyclic Graph (DAG) for the Malia Project guarantees the proper execution sequence of each phase, including data extraction, loading, and reconciliation. It also incorporates error handling techniques to effectively manage any difficulties that may occur during execution (*Figure A.1, Figure A.2*).

Airflow Features:

1. The Directed Acyclic Graph (DAG) is configured to execute on a daily basis, which ensures that the ETL process is initiated daily without the need for manual triggering.

2. The Directed Acyclic Graph (DAG) is set to automatically retry the execution of a task in the event of a failure. This is important when a task fails due to iteration specific performance issues that could be resolved upon a second trial.

3. An email notifications system is established to inform the stakeholders in case of any errors, as well as the successful completion of the process (*Figure A.3, Figure A.4*).

4. The Airflow software can be accessible through its user-friendly web-based interface, which offers a detailed overview of the Directed Acyclic Graph (DAG) and its current execution status. Users have the ability to track the status of every activity in real-time, using visual cues and colors that indicate whether tasks are scheduled for execution, in progress, successfully completed, or unsuccessful (*Figure A.5*).

5. The Airflow software also provides users with comprehensive logs for each individual task, which are crucial when troubleshooting task completion errors.

The Airflow DAG for the Malia Project is deployed and controlled within a Docker container, an open-source platform that enables developers to build, deploy, run, update and manage container. This implementation ensures a package dependence free ETL operation, as docker has numerous advantages:

1. **Portability:** Docker packages all the required dependencies and configurations into containers, which can be easily transferred and deployed across various environments without any compatibility problems. This development technique guarantees uniformity across development, testing, and production settings, reducing the occurrence of problems that could occur when running the system in different environments.

2. **Easy Expansion:** Docker, with its container technology, enables easy expansion of the system through the addition of various packages and applications from different

manufacturers.

3. **Resource Efficiency:** Docker enhances system resource utilization and reduces startup time by isolating programs from the host system, exceeding the performance of conventional virtual machines. The docker desktop interface also provides full control over how many resources a container could access from the host machine.

Also, the following table provides a concise overview of various tools and their specific functions inside the project:

Table Enumerating the various tools used for the project:

Technique/Tool	Purpose	Details
Docker	Containerization and Environment Consistency	Hosts Apache Airflow, ensuring consistent deployment across environments.
Apache Airflow	Workflow Orchestration	Manages ETL processes, ensuring automated and efficient data workflows.
MySQL	Centralized Data Storage	Primary database for storing e-commerce data, reconciliation results, etc.
Python/R	Data Processing and Analysis	Used for data cleaning, model development, and statistical analysis.
Power BI	Dashboarding and Data Visualization	Interactive dashboards for real-time monitoring of key metrics.
Apriori/FP-Growth	Market Basket Analysis	Algorithms used to discover associations in e-commerce transaction data.
Streamlit	User Interface for Model Interaction	Web app allowing users to interact with models and customize parameters.
Draw io	Informative Diagrams	Informative Diagrams that decipits various aspects of the system

3.2.4 Software Testing:

To ensure robustness of the system and to further expand scalability options for the company, numerous unit and integration testing scripts were deployed to test the various aspects of the system. Using these tests, the company could test the changes they would like to perform on the system before implementing them, which heavily reduces the probability of errors and development time when the changes are being performed.

Unit tests focus on testing individual components of the system in isolation, verifying that each "unit" of code, behaves as expected in different scenarios set by the company. Integration tests on the other hand focus on testing how different components of a system work together by

examining the interaction between the system's modules. These tests often simulate real-world scenarios, set by the company, in which data flows between components like databases, APIs, or other systems. All the tests were implemented using pytest and unittest.mock libraries.

Key Unit Tests included testing the email notifications feature, the data extraction and transformation from and to the local system, and the reconciliation SQL query execution. All of these tests were done while simulating real business scenarios and while mocking external dependencies such as mock databases and excel files.

Scripts containing unit tests:

- test_email_notifications.py
- test_extract.py
- test_transform_aramex.py
- test_transform_cosmaline.py
- test_transform_creditCard.py
- test_transform_ecom.py
- test_transform_erp.py
- test_transform_oracle.py
- test_transform_general.py

Key Integration Tests included testing how the reconciliation scripts interact with the database tables over a multitude of business scenarios. Also, as opposed to the specific transformation functions tested in the unit tests, the integration testing included how the full transformation pipeline is applied on different mocked datasets.

Scripts containing integration tests:

- test_cosmaline_reconciliation.py
- test_credit_card_reconciliation.py
- test_ecom_orders_not_in_oracle_reconciliation.py
- test_invalid_oracle_order_numbers_reconciliation.py
- test_load.py
- test_transform_aramex_cosmaline_pipeline.py
- test_transform_cosmaline_pipeline.py
- test_transform_ecom_pipeline.py
- test_transform_erp_pipeline.py
- test_transform_oracle_pipeline.py

All of the 53 tests run successfully at the time of writing this report, with the running time ranging from 30 to 35 seconds based on the current system configuration.

3.3 Objective 2: Data Reconciliation:

To ensure thorough data consistency, a total of eight separate SQL scripts were created. Each script has a specific focus on different parts of the data, such as confirming the presence of orders across tables, validating amounts, and ensuring accurate token mappings. The main tables utilized in these reconciliation procedures were 'ecom_orders', 'shippedandcollected_aramex_cosmaline', 'oracle_data', 'erp_data', 'credit_card', and 'daily_rate'.

A visual representation of the reconciliation scripts can be found in the R.1 to R.5 section of the appendix

2.3.1 Presence of Orders Across Tables:

- **ecom_orders_not_in_shipping_reconciliation.sql:** This script verifies the presence of orders in the 'ecom_orders' table that are not found in the 'shippedandcollected_aramex_cosmaline' table. This guarantees that all e-commerce orders are included in the shipping records.

- **ecom_orders_not_in_oracle_reconciliation.sql:** This script ensures that all orders in the 'ecom_orders' table are also found in the 'oracle_data' table, hence ensuring the accuracy of the e-commerce transactions in the Oracle ERP data.

- **invalid_shipping_order_numbers_reconciliation.sql:** This script verifies the presence of orders in the 'shippedandcollected_aramex_cosmaline' table that are not found in the 'ecom_orders' table. This guarantees that all e-commerce orders present in the shipping data have a corresponding transaction in the main 'ecom_orders' table.

- **invalid_oracle_order_numbers_reconciliation.sql:** This script verifies that all order numbers in the 'oracle_data' table are both valid and present in the 'ecom_orders' table. This guarantees that all e-commerce orders present in the oracle data have a corresponding transaction in the main 'ecom_orders' table.

2.3.2 Discrepancies in financial data:

- **ecom_reconciliation.sql:** This script performs a comparison between the recorded amounts in the 'ecom_orders' table and the total sum of associated amounts in the 'shippedandcollected_aramex_cosmaline' table. To guarantee precision, this procedure incorporates currency conversions with the 'daily_rate' table, taking into consideration the variations in exchange rates.

- **credit_card_reconciliation.sql:** This script aims to verify the consistency between transactions recorded in the erp_data table and the 'credit_card' table, with a specific focus on entries associated with "Cybersource." The algorithm examines the total amounts of receipts in both tables for the same day and uses a 5% threshold to decide if the records are a match or a mismatch.

- **cosmaline_reconciliation.sql:** This script guarantees the consistency between shipments labeled as "Shipped With Cosmaline" in the 'erp_data' table and those in the 'shippedandcollected_aramex_cosmaline' table by examining the total amounts of receipts in both tables for the same day and uses a 5% threshold to decide if the records are a match.

- **token_reconciliation.sql:** This script aims to ensure reconciliation between the token's transactions obtained from the 'erp_data' table and those present in the 'shippedandcollected_aramex_cosmaline' table. The sums of amounts for each token are first converted to USD as necessary. The script then calculates the total of the reception amounts and shipping amounts for each token and then compares them, using a 5% threshold as the benchmark for a match versus a mismatch.

The reconciliation findings are collected and stored in a permanent database called "reconciliation_results". This table contains a log of every reconciliation, with columns including id, 'reconciliation_type', 'reference_id', 'reference_date', 'sum_erp_amount_usd', 'sum_shipping_amount_usd', 'reconciliation_status', 'non_existent_record', and 'recon_report'. This centralized approach of a 'reconciliation_results' table enables more effective monitoring, auditing, and reporting of data consistency across systems.

3.4 Objective 3: Real-Time Data Monitoring

Alongside real-time ETL and reconciliation measures, it is essential for a company like Malia Group, that stores data from multiple sources, to have the ability to monitor in real time the status of all its activities, and transactions. For this purpose, a multi report Power Bi dashboard was developed.

3.4.1 Dashboards Details:

The Power Bi dashboards are designed, as per the business rules of the company, to oversee both high level Key Performance Indicators (KPIs) and transaction level details that would provide full traceability across data sources. In this section, a detailed description of each dashboard will be provided:

1. **Reconciliation Dashboard: ECOM Transactions Against Other Data Sources**

This dashboard provides a macro level view of the following KPIs to monitor:

- Comparison of cash amounts between E-Commerce data and ERP data
- Comparison of cash amounts between E-Commerce data and Aramex shipping data
- Comparison of cash amounts between E-Commerce data and Cosmaline shipping data
- Comparison of CyberSource amounts between E-Commerce data and credit card data

The four KPIs were monitored through Power Bi data cards that would display the aggregated amount in each data source alongside a title and description that would change according to the status of the KPI. If the difference between the two sources is larger than 5% of the aggregated E-Commerce amount, then the status between these sources is deemed a mismatch.

A screenshot of the dashboard can be found in the appendix section (Figure D.1, Figure D.2)

2. **Aramex Reconciliation Dashboard for Cash Based Transactions**
3. **Cosmaline Reconciliation Dashboard for Cash Based Transactions**
4. **Credit Card Reconciliation Dashboard for Cyber Source Based Transactions**

The second, third, and fourth dashboards offer detailed and specific information concerning the differences in transactions between the ERP system and the data sourcing from the shipping companies Aramex and Cosmaline. As for CyberSource, its transactions are compared to the ones from the company's E-Commerce platform.

The three dashboards are similar in structure, and they all include four visuals. Firstly, a donut chart displays the proportion of matched to mismatched transaction between the two sources. In addition, the dashboards include a table card that hosts the reconciliation table from the database. This table enables the user to quickly view the status of all transactions related to the report. Thirdly, for the second and third dashboards, two tables at the bottom of the report provide the user with a complete view of the ERP Oracle transactions and the Shipped and Collected transactions respectively.
As for the Cybersource dashboard, the bottom two tables consist of the transactions made with credit cards alongside a view of all E-Commerce transactions.

This inclusion of tables allows users to have both a general view of the match and mismatch status across different sources, and a transaction level detail view that could be used to track individual orders, tokens, or dates.

A screenshot of the dashboards can be found in the appendix section (Figure D.3, Figure D.4, Figure D.5)

5. **Distribution of Collected Amounts by Shipper Name and Billing Type**

The fifth dashboard is designed to showcase collection data and provide insights into important aspects of the business such as the distribution of cash versus CyberSource collections across different shippers.

This dashboard features a 100% stacked bar chart that displays the distribution of collections across various shippers. Additionally, it includes line charts that offer more specific information, such as the collection quantities by date for each individual shipper.

A screenshot of the dashboard can be found in the appendix section (Figure D.6)

3.4.2 Technical Implementation of the Dashboards:

To achieve the requested real time monitoring capabilities, the dashboards are directly linked to the MySQL database. In addition to the merged MySQL schema, additional tables, columns, metrics, and relationships were created to improve the ability to analyze data and ensure that the dashboards satisfy Malia's specific business needs:

A. *Data modeling and schema design: New Tables*

A multitude of new tables were added to the Power Bi data model, each serving a specific role towards achieving the company's business requirements:

1. **DateTable**:
   A unified date table was established to function as a universal point of reference for all date-related actions throughout the dashboard. The table includes the range of all delivery dates present in the e-commerce orders dataset. This approach guarantees uniformity in date calculations, and the simpler use of date variables, especially for filtering features.

2. **DateTable_Aramex, DateTable_Cosmaline, and DateTable_CreditCard Tables**:
   Separate centralized date tables were established for the Aramex, Cosmaline, and Credit Card datasets. These tables, similar to the data table, include the full range of E-Commerce delivery dates. The reasoning behind these tables is the circular relationship constraint encountered during the dashboard development when creating relationships between the Date table and different tables with existing relationships across their date columns.

3. **aramex_reconciliation Table**:
   This table was designed to isolate the outcomes of token-based reconciliation for Aramex shipments. The reconciliation data is filtered to particularly target "Tokens Reconciliation" records by only including metrics such as the token number, ERP system amounts, shipping amounts, and any differences between the two.
   Furthermore, the table includes a 'reconciliation_status_5percent' column that indicates if the reconciliation is within a permissible margin of 5%. Specialized subsets of the reconciliation table, like this one, were created because it was impossible to have a relationship between the reconciliation table and the date table. Therefore, a subset of specialized reconciliation tables was created and after that connected to their respective specialized date tables.

4. **cosmaline_reconciliation Table**:
   Similar to the 'Aramex_reconciliation' table, this table specifically addresses the process of reconciling shipments made using Cosmaline. The reconciliation results are filtered to only include records associated with the 'Shipped with Cosmaline Reconciliation.' The table contains relevant columns: including date, 'sum_Cosmaline_ERP' (which represents the total amount documented in the ERP system), 'sum_Cosmaline_shipping' (which represents the total amount documented by the shipping system), and the difference between these numbers. In addition, the database contains a column called 'reconciliation_status_5percent', indicates if the reconciliation is within a permissible margin of 5%.

5. **cybersource_reconciliation Table**:
   This table, similar to the above, focalizes on the reconciliation of Cybersource transactions. The function narrows down the reconciliation results to records labeled as 'Cybersource Reconciliation'. The table includes important columns such as date, 'sum_creditCard_ERP' (which represents the entire amount of credit card transactions recorded in the ERP system),

sum_creditCard (the total amount of credit card transactions recorded by Cybersource), and 'reconciliation_status', which indicates whether the transaction is classified as a match or a mismatch.

B. Data modeling and schema design: New Measures

- **Total Quantity Sold:** Calculates the total quantity of items sold from the Oracle data.
- **TotalAdjustedECOMAmount**: Sums the adjusted ECOM amounts from the e-commerce orders dataset.
- **TotalAdjustedECOMAmount_Cybersource**: Sums the adjusted ECOM amounts related to Cybersource from the e-commerce orders dataset.
- **TotalAdjustedECOMAramexCashAmount:** Sums the adjusted ECOM cash amounts related to Aramex from the e-commerce orders dataset.
- **TotalAdjustedECOMCosmalineCashAmount:** Sums the adjusted ECOM cash amounts related to Cosmaline from the e-commerce orders dataset.
- **TotalAdjustedERPAmount:** Sums the adjusted ERP amounts from the 'erp_data' dataset.
- **TotalAdjustedCreditCardAmount:** Sums the adjusted credit card amounts from the 'credit_card' dataset.
- **TotalAdjustedShippedCollectedAmount_Aramex:** Sums the adjusted shipped and collected amounts related to Aramex from the 'shippedandcollected_aramex_cosmaline' dataset.
- **TotalAdjustedShippedCollectedAmount_Cosmaline:** Sums the adjusted shipped and collected amounts related to Cosmaline from the 'shippedandcollected_aramex_cosmaline dataset'.
- **Difference:** Calculates and displays the difference between total adjusted ECOM and ERP amounts.
- **Difference2**: Displays the difference between ECOM Aramex cash amounts and shipped amounts.
- **Difference3:** Displays the difference between ECOM Cosmaline cash amounts and shipped amounts.
- **Difference4:** Displays the difference between ECOM Cybersource amounts and credit card amounts.
- **MatchStatus:** Provides a text status indicating whether ECOM cash and ERP cash amounts match within a 5% threshold.
- **MatchStatus2:** Displays a status message indicating if ECOM Aramex cash and shipping cash amounts match within a 5% tolerance.
- **MatchStatus3:** Provides a status message on whether ECOM Cosmaline cash and shipping cash amounts match within a 5% tolerance.
- **MatchStatus4:** Shows a status message indicating whether ECOM Cybersource and credit card amounts match within a 5% tolerance.

- **CardColor_Difference:** Determines the card color based on the difference between ECOM and ERP amounts, indicating whether the amounts match within a 5% threshold (blue for match, red for mismatch).
- **CardColor2_Difference:** Determines the card color based on the difference between ECOM Aramex cash amounts and shipped amounts, indicating if they match within a 5% tolerance.
- **CardColor3_Difference:** Determines the card color based on the difference between ECOM Cosmaline cash amounts and shipped amounts, indicating if they match within a 5% tolerance.
- **CardColor4_Difference:** Determines the card color based on the difference between ECOM Cybersource and credit card amounts, indicating if they match within a 5% tolerance.
- **CardColor_Header:** Similar to 'CardColor_Difference', but it sets the color for the header based on the difference between ECOM and ERP amounts within a 5% threshold.
- **CardColor2_Header:** Sets the header color for the card based on the match between ECOM Aramex cash amounts and shipped amounts, using a 5% tolerance threshold.
- **CardColor3_Header**: Sets the header color for the card based on the match between ECOM Cosmaline cash amounts and shipped amounts, using a 5% tolerance threshold.

3.4.3 Deployment and Data Synchronization:

To improve accessibility to the dashboards, the Power BI file was published on the Power Bi Services, allowing all stakeholders to access it. In addition, considering that the MySQL database is locally stored and that the power bi dashboards are published on the cloud-based power services, an on-premise data gateway was established to link the two. This gateway provides robust data transfer from the local system to the cloud with refreshments set to be applied daily. This implementation guarantees that the Power BI's data model is constantly up to date with the database that hosts the constant flow of input data. Also, this gateway ensures that any modifications made in the MySQL database are synchronized with the Power BI dashboards, eliminating the need for manual intervention. Note that, an email is sent to designated stakeholders in case of an unsuccessful data transfer (Figure D.7).

3.4.4 User Interaction and Usability:

The real time monitoring dashboard was designed with a number of features that provide an intuitive and interactive user experience:

1. A navigation pane was placed to welcome the user. This pane enables users to access any of the 5 dashboards directly without having to look up the report in the bottom navigation pane.

2. All reports have been equipped with robust filtering capabilities such as year and month filtering, allowing the user to quickly drill down transactions based on temporal measures.

3. Interactivity has also been added to the reports, especially to report number two to four. Users can click on the desired part of the donut chart (either the match or mismatch part)

and the tables present would be filtered accordingly. Furthermore, the user could locate a specific transaction (Token for example) by pressing on it. This action would render all the other tables present to be filtered on that token, giving the user a full view of a specific transaction across multiple sources.

4. An info button has also been added to all reports. This feature allows users to access the native AI chatbot capabilities of Power BI that are powered by their Q&A model. With this feature, users can ask questions about the data and visuals, presenting them with unlimited new potential insights to be discovered, even when not straightforwardly present in the dashboard.

3.5 Objective 4: Modeling and Business Analysis

The business analysis phase of the Malia Project aimed to conduct a thorough analysis of Cosmaline's sales, starting with a broad overview and then delving into specific transaction-level information. The primary objective was to predict the sales patterns of the Cosmaline e-commerce business by analyzing the transactions documented in the 'ecom_orders' dataset, which provided useful insights into forthcoming sales trends.

Past this general sales forecast, a more thorough analysis of their e-commerce business was carried out. This analysis was firstly based on an extensive exploratory analysis of their 'oracle_data'. Following the EDA, a market basket analysis was performed on the 'oracle_data' table to provide the company with insights into their client buying patterns and the connections between their different products.

3.5.1 Forecasting study:

To conduct the forecasting study, we utilized different models to analyze the transaction data from the 'ecom_orders' table. The following section provides a comprehensive review of the conducted transformation and forecasting models together with their corresponding performance:

A. Transformation and sample selection:

The sales forecasting was done on the "Amount" column of the E-commerce dataset which represents the value in dollars per order placed on Cosmaline's website. The "Created At" column in the same data file was used for this forecasting because it reflects the date when the order was placed by the customer. The necessary columns were exported from the database after being grouped by the date in order to end up with the total sales received from all orders per day. In the forecasting methodology, the first step included importing the dataset in R studio and checking the values and days present to identify the necessary transformations that need to be done before creating the time series and proceeding to work with it. When displayed, the data was spanning from June 5, 2023 to March 15, 2024 with a total of 124 missing days between June 2023 and October 2024. This large number of unrecorded sales days in only five consecutive months significantly affects the continuity of the data, which is a necessary component of a proper time series. However, to avoid the high uncertainty that comes from imputing a large number of missing values by random numbers or a central tendency measure, this five-month period was entirely

disregarded from the analysis as it only contains 29 days in total, and the chosen period started on November 2023.

Moreover, in the selected sample, the data for 8, 15, and 16 November 2023 was missing, and to address this issue, the sales amounts in these days were estimated through linear interpolation. This involves calculating a value to fill each gap in the data based on the noticed linear trend of its neighboring days which guarantees the consistency and continuity of the times series.

After the data preparation, the necessary R packages: 'urca', 'fpp2', and 'zoo' were successfully installed, and then loaded in the R environment. As a next step in the forecasting analysis, the data was transformed, and an auto plot was visualized to detect any visible trends and patterns.

It can be shown from the auto plot in (*Figure F.1*) that even though the time series is generally horizontal, there is a small increasing peak in the beginning. However, since the models will be applied on daily data, discussing seasonality is not applicable for this analysis.

Additionally, based on the plot, it can be concluded from the generally constant fluctuations that the variance is somehow stable, which means that there is most probably no need for a Box-Cox transformation to stabilize the variance. To make sure, BoxCox.lambda() was performed in R on the time series, and it returned lambda=1 which ensures that the series does not need a transformation.

B. Stationarity, Splitting, Differencing:

Consolidating on our previous observation, the Auto-Correlation Function (*Figure F.2*) of the time series from lag 1 till 48 further indicates that there is no seasonality since the lags do not follow a seasonal pattern. However, it can be noticed that the lags spanning from the 1st to the 18th are decaying slowly to zero which also verifies the presence of a slight trend in the beginning.

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test was then deployed (*Figure F.3*), and it returned a test-statistic value of 0.6732 which was greater than 0.463, the critical value corresponding to the 5 percent significance level. This means that the test-statistic is placed in the rejection region and, the null hypothesis H0 of the KPSS test which states that the tested time series is stationary, is rejected. Thus, given the trend shown in the auto plot and ACF, and the result of the hypothesis testing of KPSS, it can be concluded that the series is not stationary.

Before performing the necessary operations, the dataset was split into train and test sets, where four months were taken for training and the remaining 15 days for testing. After that, a first order differencing was applied on the training dataset, and it can be noticed in (*Figure F.4*) that the plot became roughly horizontal, the mean seems to be around zero, and the trend is no longer visible. However, to be certain that the series is ready for modeling, the KPSS test was performed again, and the results show a test statistic of 0.0382, less than all the critical values. This means that, after a first order differencing, the tiseries became stationary.

Furthermore, in order to explore various options, a second order differencing was also applied on the training set and the resulting plot in (*Figure F.5*) represents a more condensed series with a

mean tending more to zero compared to the previous one. Additionally, the KPSS test was also performed after the second differencing (*Figure F.6*), and it showed a smaller test statistic value equals to 0.0224, meaning that the series is also stationary.

C. ARIMA Model:

The ggtsdisplays containing the plot, ACF, and Partial ACF of the series after the first is shown in (*Figure F.7*). It is worth mentioning that the Autoregressive Integrated Moving Average model has three components, the first is the Autoregressive component "p", the second is the differencing component "d", and the last is the Moving Average component "q", where the first and last component will be determined by analyzing the ACF and PACF.

Considering the first display, it can be noticed that the first lags in the ACF are exponentially decaying, indicating that the ARIMA model of d =1, must contain an AR component. Looking at the first few lags in the PACF, the first two seem to be highly significant since they are outside the blue line while the rest are either very close the line or inside it, which indicates that there are two options for the AR component, p=1 and p=2. Moving on to the MA component of the model, the PACF is also exponentially decaying and only the first lag in the ACF is significant. Therefore, the model must include an MA component, and it should be set to q=1. Thus, based on the first differencing display, the starting ARIMA models that will be implemented are ARIMA (1, 1, 1) and ARIMA (2, 1, 1).

When it comes to the second differencing display of the ACF and PACF, it can be spotted that while the PACF is exponentially decaying, the ACF is not, which means that the ARIMA model of d=2 must only contain an MA component. Based on the ACF, the MA component must be set to one since the first lag seems to be the only significant one. This suggests that the only ARIMA model that must be applied with d=2 is ARIMA (0, 2, 1). Additionally, auto.arima() was also used in order to check the optimal model according to R and compare it to the models that were already implemented.

D. ETS Model:

Besides ARIMA, the ETS model which stands for Error Trend Seasonality was also applied on the training set. This model breaks down the data into Error, Trend, and Seasonal components where the error can be either additive or multiplicative while the trend and seasonality components can be additive, multiplicative, or none. Due to the fact that the data lacks seasonality, the seasonal component in the ETS models must be set to N.

Therefore, the implemented ETS models were ETS (A, N, N), ETS (A, A, N), ETS (M, A, N), and ETS (M, M, N). Additionally, ets() was applied on the training dataset to check the optimal ETS model suggested by R. The results of these models were then compared with the results of previous ARIMA models.

It is worth noting that models like ETS are typically used for seasonal data that also reflect a noticeable trend. However, the ETS model was applied in the present analysis as an option that would be more relevant when implemented with larger seasonal data.

E. Model Evaluation and Selection:

To evaluate and choose between the models, different criteria were used. The first and most important criteria is whether there is an autocorrelation between residuals or not, which is determined by the L-jung Box test. If present, an autocorrelation between residuals indicates that the residuals are not independent of each other and might be affected by factors not taken into consideration by the model. The Root Mean Squared Error and the Mean Absolute Percentage Error were also considered during model selection since they give direct measures of the forecast's accuracy.

Another evaluation criterion is the Corrected Akaike Information Criterion, and it was used instead of the Akaike Information Criterion since it contains a correction parameter for the small sample size, making it more accurate. Similar to the RMSE and MAPE, the Akaike information criterion (AICc) is a valuable metric since it measures how well the model fits to the data, where a lower AICc indicates that the model fits better than others for the given dataset. Based on these criteria, all ARIMA and ETS models were evaluated on the test set and the best model was selected after comparing the results.

F. ARIMA and ETS averaging:

Averaging the best obtained models is an approach that could be followed in the forecasting analysis where the average of the forecasting results given by the best was calculated. However, it is worth noting that such averaging measures add complexity to the models, which can overfit the data if it is relatively small. Even though the data at hand is considered small for typical forecasting analysis, averaging was performed. The averaging results were then evaluated on the test set and the obtained RMSE and MAPE were compared with those of the previously chosen model to conclude the best performing model in forecasting website sales.

3.5.2 Exploratory Data Analysis (EDA)

Based on the forecasting results, a need for further analysis of Malia's e-commerce platform was established. To provide the company with actionable insights to enhance the performance of the Cosmaline website, a thorough Exploratory Data Analysis (EDA) was performed to reveal overall statistics and insights about the dataset. Alongside the EDA, data mining models were also employed to uncover associations between their different cosmetic lines.

A. Data collection

The primary objective of data collection was to expand the already available oracle_data table by adding product names and categories as having an EDA and data mining analysis on product ids would provide limited interpretability for the company. To host the products information, the oracle_data_product_names database was generated by a web scraping procedure that gathered product names and categories from the Cosmaline website.

The web scraping was performed using Python, specifically using the BeatifulSoup library for HTML parsing and the Selenium library for browser automation. Here is a detailed description of the scraping algorithm used:

1. **Data Preparation:** The product IDs were extracted from the Oracle dataset and put in a list. A specific product ID, "SHIPPING_LOCAL", was then removed from the list as it was irrelevant for scraping.

2. **Selenium Setup:** Selenium was used to automate access with the Cosmaline website. This was done using webdriver_manager that automatically manages Chrome instances to access and navigate the e-commerce website.

3. **Scraping Logic:** The script iterated over the list of product IDs and constructed a search URL for each product on the e-commerce website. From the search page and after typing the product ID in the search bar, the product name and categories were extracted: The product name was retrieved from a specific HTML element containing the search results, and some post-processing was performed to clean and format the product name. As for categories, they were also collected from HTML elements representing category buttons on the search results page and then appended to a list.

4. **Data Storage:** The scraped data, including product ID, product name, and categories, was then stored in a data frame before being pushed to the oracle_data_product_names table in the permanent database.

B. Data Preprocessing:

Following the data collection, a preprocessing pipeline was designed to transform the oracle data before applying both the EDA and data mining models. Below is a breakdown of the preprocessing pipeline:

1. **Data Querying:** The data was fetched from the reconciled MySQL database using an SQL query that joined the oracle_data and oracle_data_product_names tables. The resulting dataset from the query was then loaded into a python data frame awaiting preprocessing.

2. **Convert Strings to Lists:** Product categories stored as strings in the database were converted into Python lists elements, which allowed for easier manipulation as all list methods could be applied to it.

3. **Remove 'Product Not Found' Entries:** Any entries where the product name was listed as "Product Not Found" were removed from the dataset. Product Not Found entries represent the product ID in the oracle dataset that were not available on the Cosmaline website at the time of scraping.

4. **Preprocess Category Lists:** Special characters such as periods and parentheses were removed from the product category list to improve interpretability.

5. **Convert Lists to Strings:** After cleaning the lists, they were converted back into strings separated by commas if a product entry belonged to multiple categories. Furthermore,

products with no categories were marked as "Unknown."

6. **Remove Invalid Pricing Data:** Products where the price after discount was higher than the one before discount were removed.

7. **Calculate Total Sales:** Two new columns were calculated: total_sales_with_discount, which represented the sales amount based on the unit selling price, and total_sales_without_discount, which represented the hypothetical sales amount if no discounts had been applied.

8. **Calculate Discount Percentage:** A new column, discount_percentage, was also added to represent the discount applied for each row. This column was created by calculating the difference between the list price and the selling price, converted to a percentage.

9. **Expand Categories into Rows:** Since some products belonged to multiple categories, another data frame was created with its category column expanded so that no row contained multiple categories in the product_category column.

C. Starting with the EDA, this analysis focused on various business aspects:

1. Sales Analysis:

This section explores many aspects of the company's sales performance. The following points specify all the analysis done under that section:

- **Top 10 Best-Selling Products** (*Figure E.1*): This bar chart highlights the top 10 best-selling products by quantity, excluding products labeled as "Product Not Found." The plot provides insights into the highest-performing products in terms of sales volume.

- **Bar Chart of Monthly Sales** (*Figure E.2*): This horizontal bar chart shows the total monthly sales, including discounts, across different periods. It helps identify trends and seasonal fluctuations in sales performance over time.

- **Top Product Categories** (*Figure E.3*): This analysis showcases the top 10 product categories based on the total quantity sold.

- **Revenue Contribution by Product Category** (*Figure E.4*): This bar chart displays the total revenue contribution of each product category. It helps in identifying which categories generate the most revenue and the ones that are key contributors to Cosmaline's overall sales.

- **Distribution of Order Quantities** (*Figure E.5*): This histogram displays the distribution of order quantities across all transactions. It provides a clear view of how much a customer

typically orders from one item.

- **Monetary Distribution** (*Figure E.6*): This plot shows the distribution of monetary value per basket. It helps in understanding the distribution of high- and low-value carts and understanding the general value of baskets purchased by customers.

2. Order Analysis:

The Order Analysis provides insights into customer engagement with the store and their regular purchasing behaviors. The following points specify all the analysis done under that section:

- **Recency Distribution** (*Figure E.7*): This histogram illustrates the recency distribution, showing the number of days since the last purchase for each customer. It provides insights into customer engagement and how frequently customers place orders.

- **Order Quantity Statistics** (*Figure E.8*): This analysis calculates key statistics for basket quantities, such as the average, minimum, maximum, median, and quartile values.

- **Distribution of Items per Order** (*Figure E.9*): This histogram plots the distribution of the total number of items per order, providing insights into typical basket sizes and identifying any potential clustering of order quantities.

3. Pricing and Discount Analysis:

The Pricing and Discount Analysis is aimed at investigating the pricing strategies of items and assess the effects of discounts. The following points specify all the analysis done under that section:

- **Distribution of Selling Prices and List Prices** (*Figure E.10*): This histogram compares the distribution of unit selling prices and unit list prices, highlighting any frequency differences between original prices and final sales prices after discounts.

- **Distribution of Discount Percentages** (*Figure E.11*): This histogram visualizes the distribution of discount percentages applied across all products. It provides insights into the typical discounts applied by the store.

- **Box Plot of Discount Percentages by Product Category** (*Figure E.12*): This box plot displays the distribution of discount percentages across Cosmaline's product categories, which helps in identifying any difference in discounting strategies and which categories receive more or less aggressive discounts.

- **Scatter Plot of Discount Percentage vs. Ordered Quantity** (*Figure E.13*): This scatter plot displays the relationship between discount percentage and the quantity of items ordered, revealing whether larger discounts lead to higher orders.

- **Comparison of Average Discounts for Popular and Less Popular Items** (*Figure E.14*): This histogram compares the average discounts applied to popular items versus less popular ones based on their median ordered quantity. This analysis provides an understanding of how discount strategies influence the popularity of items.

- **Box Plot of Unit Selling Prices by Top Performing Product Categories** (*Figure E.15*): This box plot highlights the distribution of unit selling prices across the top-performing product categories. This plot offers insights into pricing similarities or differences between the most successful categories.

4. **Supply Chain Analysis:**

This section focuses on understanding the distribution network and the performance of key supply chain partners of the Cosmaline Brand:

- **Treemap of Top Distributing Companies by Quantity Shipped** (*Figure E.16*): This Treemap displays the top 10 distributing companies based on the total quantity of items shipped. The size of each square in the Treemap represents the volume shipped by each distributor.

5. **Product Positioning:**

This section focuses on understanding the market positioning of Cosmaline's products:

- **Distribution of the Number of Items per Transaction** (*Figure E.17*): This histogram shows the distribution of the number of unique items per cart. This plot provides insights into the diversity of products purchased during a single takeout basket.

- **Distribution of Item Frequencies** (*Figure E.18*): This histogram displays how frequently each item is purchased across all transactions. It helps to identify the most frequently purchased products and assess the popularity of items.

- **Transaction Diversity Statistics** (*Figure E.19*): This analysis calculates key statistics on the diversity of items within transactions. It includes metrics like the mean and median number of unique items per transaction.

- **Item Frequency Statistics** (*Figure E.20*): This analysis, as a similar approach to Figure E.19, calculates summary statistics for item purchase frequencies across all transactions, such as the average, minimum, and maximum frequency of items.

- **Combined Niche Products Count** (*Figure E.21*): This analysis identifies niche products based on low sales volume and low purchase frequency. Products with lower sales volume (bottom 25% based on quantity sold) and lower number of appearances in unique transactions (bottom 25%) are considered niche. These niche products are the ones that do not sell as frequently compared to others.

3.5.3 Data Mining Models:

Following the extensive EDA analysis, and with the presence of detailed e-commerce cart data, it was decided that data mining technique, specifically market basket analysis, would yield useful insights into the e-commerce sales performance of Cosmaline. This section provides a comprehensive explanation of the data mining method, as well as the models used to discover patterns and relationships in Cosmaline's sales data.

Data mining is a technique used to examine an extensive amount of information in order to reveal patterns and connections in the data. This procedure utilizes techniques from machine learning, statistics, and database systems to detect significant patterns that are of value for corporate initiatives. A market basket analysis, a method in data mining, was applied to the Cosmaline e-commerce data to identify correlations between various products and to give insights into their consumer purchase behavior.

The outcomes of a market basket analysis are a set of association rules: association rule is a technique in data mining that reveals all sets of related products that satisfy predetermined minimum support and confidence, and lift criteria. The support value quantifies the frequency of an item's occurrence in the dataset, whereas the confidence value measures the probability of a set of items being present given that one of the items in the set is also present in the data cart. As for the lift, it is used to determine how much more likely would an item be present in the basket when another specified item is also present.

Here is a mathematical representation of the different metrics used to evaluate an association rule:

$$Rule: X \rightarrow Y$$

The support of an itemset or rule indicates how frequently the itemset appears in the dataset:

$$Support = \frac{frq(X,Y)}{N} = \frac{Number\ of\ transations\ containing\ bothe\ X\ and\ Y}{Total\ Number\ of\ transactions}$$

The confidence of an itemset measures how often the rule X⟶ Y has been found to be true given the antecedent X is present:

$$Confidence = \frac{frq(X.Y)}{frq(X)} = \frac{Number\ of\ transactions\ containing\ both\ X\ and\ Y}{Number\ of\ transactions\ contianing\ X}$$

Lift indicates the strength of a rule compared to the baseline probability of Y occurring, without considering X. It tells us how much more likely Y is to occur when X occurs compared to when X does not occur.

$$Lift = \frac{Confidence(X,Y)}{Support(Y)} = \frac{confidence\ of\ the\ rule}{Support\ of\ Y}$$

To perform this market basket analysis, two main algorithms were used: Apriori and FP-Growth:

- The Apriori Algorithm generates candidate itemsets from the dataset. K-itemsets are created by merging (k-1) itemsets obtained from earlier iterations. The Apriori algorithm then eliminates all candidate itemsets whose subsets do not adhere to the specified minimum support threshold. This support value for each candidate itemset is computed by examining the complete database to find the count of transactions that contain all the items in the candidate itemset. Following this computation, high-frequency patterns are determined by comparing the support values with the pre-established minimum threshold. The procedure continues, increasing the value of k, until no additional high-frequency patterns are discovered.

- The FP-Growth Algorithm is an alternate approach employed to detect common itemsets in a dataset. FP-Growth employs a more efficient strategy compared to Apriori, as it constructs a tree structure called the FP-Tree, which eliminates the need for multiple scans of the database. The FP-Tree compresses the dataset by associating each transaction with particular paths in the tree and not all transactions, enabling quicker detection of frequent itemsets. The FP-Growth algorithm consists of three primary stages: constructing the Conditional Pattern Base, generating the Conditional FP-Tree, and searching for frequent itemsets. As a result, the FP-Growth algorithm rapidly discovers patterns by recursively producing and analyzing FP-Trees rather then constantly evaluating against the entire database.

The Apriori and FP-Growth algorithms were utilized on Cosmaline's e-commerce data to offer valuable insights on client buying patterns, which is essential to drive cross-selling and marketing efforts such as product placement and bundling.

3.5.4 Data collection

The primary objective of the data collection was to expand the already available oracle_data table by adding product names and categories as having association rules with product The oracle_data_product_name database was generated by a web scraping procedure that gathered product names and categories from the Cosmaline website. This process guaranteed that every product ID in the oracle_data table was linked to its matching name and category, enhancing the dataset for more insightful analysis.

3.5.5 System Integration of Business Analysis Components

To provide Malia Group with better access to the business analytics tools discussed, the EDA analysis and Data Mining models were both integrated in the established ETL system. This incorporation was double faceted as it was added to the automated Airflow system and deployed to a standalone Streamlit application.

A. Integration of the airflow system

The already established Airflow system was expanded to completely automate the business analysis process:

For the model integration, the Airflow system first queries the database, applies preprocessing pipelines (as specified before), and then executes both the FP-Growth and Apriori models. The outcomes of these models are then saved locally as CSV files for memory efficiency, before being pushed to the database in their corresponding apriori_results and fpgrowth_results tables.

As for the automated EDA, the data undergoes the same preprocessing and EDA functions are applied to the processed data frame. To optimize memory usage, all plots, texts, and tables are saved locally using appropriate file types.

To provide the company's stakeholders with the results of the analysis, a distinct step in the Airflow process combines the exploratory data analysis visuals and the model outcomes from the database into an automated PDF report. Upon completion, the report is thereafter transmitted via email to stakeholders (Figure A.4).

The default parameters for support, confidence, and lift thresholds were assigned to the automated process based on the results of the EDA section, which will be later developed in the results and discussion sections.

B. Streamlit Application for User Interaction

The second component of the model integration is an independent Streamlit application that is specifically developed to provide users with full control over the model parameters.

Users can firstly choose between the entire MySQL oracle_tdata table or the upload a customer number of transactions if the structures adhere to the one of the oracle_data table. As for the default option, the full table, it is fetched from a github repository that get's updated daily as part of the Airflow architecture. This step was developed to counter the inability of connecting the local MySQL database to the cloud based deployed streamlit application.

Users can choose between the Apriori and FP-Growth models within the Streamlit interface. After model selection, users can modify the minimum threshold parameters for support, confidence, and lift, which is useful when examining various parameter combinations that could be relevant to their business requirements. Upon selection of the model and its parameters, association rules are generated and displayed to the user alongside their corresponding metric scores. Users can then filter the list of rules by selecting antecedents they want to view with all of its corresponding consequences.

Alongside the list of association rules, the application also presents essential findings from the model such as the distribution of lift, confidence, and support with a visual representation of the correlation between confidence and support. These visuals allow the user to better understand and access the performance of the model under the parameters chosen

Aside from model selection and customization, the Streamlit application also includes the same Exploratory Data Analysis (EDA) when could be of value when deciding on the minimum threshold parameters for the models. In addition, an explanation of the models and their metrics is incorporated in the application for better understanding of the analysis performed.

*Screenshots of the Streamlit application can be found in the **Appendix S.***

- ***Figure S.2:*** In this page of the Streamlit application, the user can upload the data while following the format guidelines written and view the necessary information about quering the data by clicking on the button found under the guidelines.

- ***Figure S.5:*** This figure shows the EDA done on the data and it can be noticed that under the navigation tabs, there are different buttons which the user can choose from to refer to the desired part of the EDA: Sales Analysis, Order Analysis, Pricing and Discount Analysis, Supply Chain Analysis, or Product Positioning Analysis.

- ***Figure S.6 & S.7:*** On this page, the user can read about the different parameters of data mining models, choose between Apriori and FP-Growth algorithms, set the parameters, and finally click on the "Run model" button to run the chosen model.

- ***Figure S.8, S.9, & S.10:*** These figures show the Model insights page of the application. The scatterplot of Support and Confidence of the previously chosen model (*Figure S.8*), and the distributions of all three parameters (*Figure S.9*) can be viewed. Finally, the user can view the list of Association Rules referring to the choice of model and parameters, and can also filter the rules by choosing the desired antecedent (*Figure S.10*).

The following is the link for the deployed Streamlit application: Streamlit

4. Results

The following section outlines the results of the reconciliation methods, forecasting, exploratory data analysis (EDA), and market basket analysis applied to Malia Group's E-Commerce business.

4.1 Reconciliation Results:

The reconciliation results provide valuable insights into the integrity of Malia's financial data, spanning across different sources such as ECOM, ERP, Aramex, Cosmaline, and CyberSource.

1. ECOM Cash vs. ERP Cash Discrepancy

2023 Analysis: The analysis of ECOM cash and ERP cash data in 2023 revealed a substantial discrepancy, amounting to $6,544.85. Meaning that the overall balance in the ECOM source was higher than the one from the ERP source.

2024 Analysis: The comparison between ECOM cash and ERP cash data also exhibited a large discrepancy of -$5,809.79 in 2024. However, the negative balance indicates that the ERP source had a higher balance than its ECOM counterpart. We will now explore how this difference was split across the month

- **January 2024:** A large part of the discrepancy from 2024 occurred in January with a difference of -1300.18. However, the monthly balance was classified as a match since the difference did not go beyond 5% of the total ECOM amount.

- **February 2024:** By February 2024, the discrepancy widened to -$5,243.21, indicating a mismatch for this month as it exceeded the threshold of 5% of total ECOM amount.

- **March 2024:** In March 2024, the discrepancy shifted direction, with ECOM cash exceeding ERP cash by approximately $733.60, which concluded the month with a matched balance.

- The results of January, February, and March would aggregate to the large discrepancy of -$5,809.79 observed in 2024.

2. ECOM Cash vs. Aramex Shipping Cash Discrepancy

2023 Analysis: The reconciliation between ECOM cash and cash collections from Aramex showed no discrepancy in 2023, indicating a perfect match between the two sources.

2024 Analysis: This alignment continued into 2024, with a minor discrepancy of $23.24, indicating that ECOM collections were slightly higher than the ones present in Aramex. We will now explore how this inconsequential difference was split across the month of 2024:

- **January 2024:** In January 2024, the data displayed no discrepancies between ECOM cash and Aramex shipping data.
- **February 2024:** Similarly, in February 2024, the reconciliation revealed no differences between the two sources.

- **March 2024:** During March the small difference of $23.24 occurred. However, the month still closed as a match between the two sources.

3. ECOM Cash vs. Cosmaline Shipping Cash Discrepancy

2023 Analysis: The reconciliation between ECOM cash and cash collections from Cosmaline were perfectly matched in 2023, with no discrepancies reported.

2024 Analysis: The perfect alignment between ECOM cash and Cosmaline shipping data remained in 2024, eliminating the need for monthly analysis as all month of 2024 exhibit no difference between the two sources.

4. ECOM CyberSource and Credit Card Discrepancy

2023 Analysis: The reconciliation between ECOM CyberSource data and collections sourcing from credit card transactions revealed a small difference of -$93.55, indicating that amounts for credit card transactions were higher than their counterparts recorded in the ECOM transactions. This difference was however higher than the set 5% threshold, resulting in 2023 classified as a mismatch between these two sources.

2024 Analysis: In 2024, the difference between CyberSource and credit card transactions remained the same at -$93.55. However, their reconciliation status remained at a match, since the volume of transactions was much higher compared to 2023. We will now explore how this difference was split across the month of 2024:

- **January 2024:** In January 2024, the discrepancy widened to -$484.50, indicating that transactions in the credit card table had a higher balance than their counterpart in the Cybersource section.

- **February 2024**: By February 2024, the discrepancy shifted directions and amounted to $484.50, balancing the difference that occurred in January.

- **March 2024:** In March 2024, a nearly perfect alignment was achieved, with only a minor discrepancy of $0.01 between the two sources.

5. Aramex Reconciliation with ERP data for cash-based transactions:

2023 Analysis: In 2023, the reconciliation dashboard indicated that the Aramex cash transactions matched perfectly the ones sourced from ERP.

2024 Analysis: the same could be said for the reconciliation status of 2024 as no mismatches were reported between Aramex shipping and ERP data. This eliminates the need to explore monthly reconciliation status between the two sources.

6. **Cosmaline Reconciliation with ERP data for cash-based transactions:**

2023 Analysis: In 2023, the reconciliation dashboard indicated that the Cosmaline cash transactions matched perfectly the ones sourced from ERP.

2024 Analysis: 2024 exhibited some discrepancies, amounting to 9.09% of Cosmaline cash transactions not matching ERP data.

- **March 2024:** March is the only month where a discrepancy between Cosmaline cash transactions and ERP transactions occurred. The difference between the two totaled to $3,543.18, representing a 66.67% mismatch.

7. **Credit Card Reconciliation with ERP data for Cybersource based transactions:**

2023 Analysis: In 2023, the reconciliation dashboard indicated that the Credit card transactions amounts matched perfectly the ones sourced from ERP.

2024 Analysis: In 2024, the reconciliation results remained adequate, with 96.43% of transactions matching perfectly between the two sources.

- **January 2024:** there were no mismatch in January between Credit card transactions and the ones from ERP.

- **February 2024:** February 2024 exhibited a slight mismatch of 11.11% between the two sources.

- **March 2024:** similar to January, there were no mismatch in January between Credit card transactions and the ones from ERP.

8. **Distribution of Collected Amounts by Shipper Name and Billing Type**

2023 Analysis: In 2023, the total amount collected through Aramex transactions for December was $33,416.21 across billing types, whereas Cosmaline collected $5655.47. This highlights that Aramex is the main source of collection amounts for Malia Group. As for the billing type distribution, Cash payments accounted for the majority of collections for both Aramex and Cosmaline.

2024 Analysis: In 2024, Aramex continued to dominate the collection process, especially in the cash billing type. The total collections from Aramex across billing types amounted to $63,684.15, compared to Cosmaline's $23,784.17. Concerning billing type distribution, while cash payments remained prevalent, there was an increase in Cybersource payment usage for Cosmaline, where its Cybersource collections increased from 25.49% to 44.11% of all cybersource collections.

4.2 Forecasting Analysis Results:

Transitioning from the reconciliation and collection analysis, we now turn our focus to the forecasting results:

When it comes to forecasting the e-commerce website sales, several ARIMA and ETS models were implemented, evaluated, and compared to select a reliable model for this forecasting. As a first step in evaluating the models, the L-jung Box test with a null hypothesis stating that there is no autocorrelation between residuals, was applied. The tables in (*Figure F.9*) and (*Figure F.10*) show the p-values obtained from the L-jung Box test of each model tried, and it can be noticed that all values are greater than 0.05, meaning that for all the models, we fail to reject H0. Thus, no autocorrelation was found between residuals in any of the models, indicating no dependency between them. The next step would be comparing the RMSE, MAPE, and AICc values of each model, which are also documented in the tables. After comparison, it can be noticed that between the three ARIMA models, ARIMA (1, 1, 1) has the lowest RMSE, MAPE, and AICc values.

In addition, ETS (A, N, N) also stands out among the four applied ETS models as the one with the best performance and smaller errors. Additionally, it is worth noting that after applying auto.arima() and ets() on the series, the suggested models were ARIMA (1, 1, 1) and ETS (A, N, N) respectively which confirms the accuracy of the model selection. However, ARIMA (1, 1, 1) showed significantly lower error values than those of ETS (A, N, N) especially for AICc which is an important criterion in model evaluation. Furthermore, given that ARIMA (1, 1, 1) and ETS (A, N, N) are the best performing models, the values obtained by their forecasting were averaged and evaluated on the test set and the evaluation results are tabulated in (*Figure F.11*). By comparing the RMSE and MAPE of the average model with those of ARIMA (1, 1, 1), the latter again showed lower values indicating that it is the final optimal model among all others. In (*Figure F.12*), the forecasts generated by ARIMA (1, 1, 1) for sales on Cosmaline's website are shown with 80% and 95% confidence intervals. Also, the formula of this model is:

$$Yt = 1.4575Yt - 1 - 0.4575Yt - 2 + \epsilon t + 0.868\epsilon t - 1$$

The autoplot of the ARIMA(1,1,1) model forecast illustrates the correlation between the past data and the predicted future values. The black line depicts the actual data used for training. And it reveals some fluctuations with a notable peak towards the end of 2023. After reaching a highest point, the numbers seem to become more stable.

The blue line on the plot represents the forecasted values obtained by the ARIMA(1,1,1) model. As per the plot, future sales are projected to stay within a comparable range to prior sales, with no sharp rises or declines.

The predicted blue line is also surrounded by shaded blue patches, which indicate the 80% and 95% confidence intervals. The intervals quantify the level of uncertainty in the forecasts with the darker blue region representing an 80% confidence level, while the lighter blue region representing a 95% confidence level. As the forecast time increases, the confidence intervals expand, indicating the growing uncertainty in the model's predictions.

In general, the ARIMA(1,1,1) model accurately represents the overall pattern of the historical data. As for the forecast, it largely suggests that future sales will likely follow a similar pattern as in the past, without any notable increase or decrease.

4.3 Exploratory Data Analysis Results:

The purpose of conducting the Exploratory Data Analysis (EDA) was to obtain a thorough analysis of Cosmaline's sales data and insights that might be used to influence future studies and policies. In this section, the result of the EDA is outlined:

1. Sales Analysis

The Sales Analysis area offers valuable information on the sales performance of different products, categories, and order trends within the e-commerce data:

Initial visual revealed the top 10 highest-selling products. The bar chart shows that top selling product for Cosmaline was the Soft Wave Hijab Oil Replacement, with Cosmal Cure Professional Sulfate Free and Soft Wave Hijab Shampoo closely following. Subsequently, a bar chart was used to analyze the monthly sales trend by grouping sales data monthly. The chart reveals a downward trend in sales from January to March 2024, with January exhibiting the highest overall sales. The performance of product categories was also analyzed, specifically identifying the top-selling categories based on their ordered amounts. The chart concluded that Best Sellers, Hair Care, Curly & Wavy Hair, Free from Sulfate & Silicone were the top categories in Cosmaline's business. As for the revenue contribution by product category, it was shown that the same Best Sellers, Hair Care, Curly & Wavy Hair, Free from Sulfate & Silicone categories make the highest contribution to overall sales. Furthermore, the analysis of the distribution of order amounts indicates that the majority of orders are small (2 to 3 items), and that there is a significant decrease in frequency as the order size grows. As for the monetary distribution, it reveals a right skew, with the majority of consumers making transactions of relatively low value.

2. Order Analysis

The Order Analysis section examines the recency distribution of customer orders, the distribution of order quantities, and the number of goods per order:

The recency distribution is represented by a histogram with a Kernel Density Estimate (KDE) curve, illustrating the number of days that have passed since clients made their most recent purchase. This curve shows a concentration of orders around 80 days from the last transaction, which align with the monthly trend analysis indicating high sales in January. As for order quantities, the average order quantity was 10.2 items, with the minimum being 1 and maximum being 162. As for the median, it stood at a slightly lower 8 items per order with the third quartile being only 13 items per order. In addition, the complete distribution of goods per order showed similar results, with the frequencies being concentrated around 2 to 8 items per order.

3. Pricing and Discount Analysis Results

This section concentrates on the pricing strategies, distribution of discounts, and their impact on sales across various product categories:

The analysis of selling prices and list prices indicated that the majority of products are sold at a low price ranging from close to $0 to $4, while only a small number of products are sold at higher prices. In addition, the bulk of products were discounted between 0% and 20%, with a significant number of products being discounted below 10%. Only a slight minority of products were heavily discounted up to 100% reduction. As for the discount percentages based on product category, it was found that discount practices were mostly similar across the categories. However, there was some variation noticed in categories such as "Best Sellers" and "Kids," which exhibited more volatile and outlier heavy discount percentages. The relationship between the discount percentage and the ordered amount was also examined and showed a very weak negative correlation (-0.01) between the two variables. This indicates that the discount percentage do not seem to impact the quantity of items ordered. Since this correlation does not align with typical discount knowledge, items were classified as popular and not popular based on whether its ordered quantity exceeds the median of the aggregated quantities for all other items. Post classification, an analysis of average discounts for popular and less popular items revealed that popular items tend to receive marginally larger average discounts (mean of 7.9% compared to a mean of 4.1% for less popular items). To drill down on high selling categories, the distribution of unit selling prices in the leading product categories was analyzed. The results showed that the majority of categories had a narrow price range, but there were exceptions in categories such as "Hair Care" and "Body Lotion", which exhibited a higher number of outliers.

4. Supply Chain Analysis Results

The supply chain study aimed to identify the primary distributing companies based on the volume of shipments:

Results indicated that "Cosmaline Local" is by far the dominant distributor, surpassing other operating units by a substantial margin in terms of the quantity of products distributed. However, Cosmaline still employs a multitude of other distributors that are all under the "Ch. Sarraf" brand.

5. Product Positioning Results:

Due to the nature of e-commerce businesses, the positioning of the products on the platform is an important aspect to analyze. In this section, a positioning analysis was done on the products sold by Cosmaline:

The frequency distribution of the number of items per transaction reveals that the majority of transactions consist of a quantity ranging from 5 to 10 items, with an average of 6.31 items per transaction. The highest order quantity of a single item was 31, whereas the lowest was merely 1. Upon analyzing item frequency across transactions, it was found that the mean frequency of item purchases is roughly 38.92 times. This means that on average, each product was found in 38.92 customer baskets. Nevertheless, there is significant variation, with certain things being picked up 328 times, and others were only purchased once. To generalize these frequency, 80 (out of 473) items were classified as niche items by having their sales volume and buy frequency below the 25th percentile.

6. Market Basket Analysis Results:

After displaying the results of the exploratory analysis, this section outlines the results of the market basket analysis:

A. Frequent Itemset:

In the context of Market Basket Analysis (MBA), an itemset refers to a collection of one or more items that appear together in a transactional dataset.

First 20 rows of the Frequent Itemset table for the Apriori Algorithm:

support	itemsets
0.130729	cosmal cure professional oh my curls cream shampoo low foam
0.128737	cosmal cure professional oh my curls light touch conditioner
0.128338	soft wave hijab oil replacement
0.125548	cosmal cure professional oh my curls moisturizing conditioner
0.107214	cosmal cure professional oh my curls styling mousse
0.107214	cosmal cure professional repair 9 leave in cream
0.104823	cosmal cure professional oh my curls light touch low lather
0.094859	cosmal cure professional oh my curls cream shampoo low foam, cosmal cure professional oh my curls moisturizing conditioner
0.090076	cosmal cure professional sulfate free
0.08888	cosmal cure professional repair 9 oil replacement spray
0.085692	soft wave hijab shampoo
0.085293	soft wave shower cream gardenia
0.084496	cosmal cure professional oh my curls light touch low lather, cosmal cure professional oh my curls light touch conditioner
0.082503	cosmal cure professional nutri strength shampoo for dry damaged
0.080909	skinnet intensive nourishment body
0.079314	soft wave hijab conditioner
0.07214	fixnet pro styling mousse x6 bouncier curls

0.070945	cosmal cure professional vital shine shampoo for colored highlighted
0.068952	cosmal cure professional oh my curls light touch mask
0.068155	cosmaline smiles toothpaste sensitivity pro whitening

First 20 rows of the Frequent Itemset table for the FP-Growth Algorithm:

support	itemsets
0.130729	cosmal cure professional oh my curls cream shampoo low foam
0.128737	cosmal cure professional oh my curls light touch conditioner
0.128338	soft wave hijab oil replacement
0.125548	cosmal cure professional oh my curls moisturizing conditioner
0.107214	cosmal cure professional repair 9 leave in cream
0.107214	cosmal cure professional oh my curls styling mousse
0.104823	cosmal cure professional oh my curls light touch low lather
0.094859	cosmal cure professional oh my curls cream shampoo low foam, cosmal cure professional oh my curls moisturizing conditioner
0.090076	cosmal cure professional sulfate free
0.08888	cosmal cure professional repair 9 oil replacement spray
0.085692	soft wave hijab shampoo
0.085293	soft wave shower cream gardenia
0.084496	cosmal cure professional oh my curls light touch low lather, cosmal cure professional oh my curls light touch conditioner
0.082503	cosmal cure professional nutri strength shampoo for dry damaged
0.080909	skinnet intensive nourishment body
0.079314	soft wave hijab conditioner
0.07214	fixnet pro styling mousse x6 bouncier curls
0.070945	cosmal cure professional vital shine shampoo for colored highlighted
0.068952	cosmal cure professional oh my curls light touch mask
0.068155	cosmaline smiles toothpaste sensitivity pro whitening

B. Association Rules:

Association rules are used to discover relationships between the different itemset, and it is expressed as an implication of the form A → B, where A (the antecedent) and B (the consequent) represent sets of items. The rule dictates that if A occurs in a transaction, B is likely to occur as well. As for the evaluation of these rules, it is based on the previously discussed support, which is the frequency of A and B occurring together; confidence, which is the likelihood of B occurring given that A has occurred; and lift, which indicates how much more (or less) likely is B going to be purchased given that A is present in the cart.

a. Association Rules for the Apriori Algorithm:

Antecedents	Consequents	Support	Confidence	Lift
cosmal cure professional fall control balsam for weak thin hair	cosmal cure professional fall control shampoo for weak thin	0.034277	0.796296	11.96352
cosmal cure professional vital shine mask for colored highlighted	cosmal cure professional vital shine balsam for colored highlighted	0.027102	0.73913	12.70191
cosmal cure professional vital shine balsam for colored highlighted	cosmal cure professional vital shine shampoo for colored highlighted	0.048226	0.828767	11.68189
cosmal cure professional vital shine mask for colored highlighted	cosmal cure professional vital shine shampoo for colored highlighted	0.028298	0.771739	10.87805
soft wave kids curl gentle hair milk	soft wave kids curl gentle shampoo	0.025508	0.719101	13.66837
soft wave kids curl gentle moisturizer	soft wave kids curl gentle shampoo	0.038661	0.858407	16.31624
soft wave kids curl gentle shampoo	soft wave kids curl gentle moisturizer	0.038661	0.734848	16.31624
soft wave kids strawberry conditioner over 90 natural origin ingredients	soft wave kids shampoo strawberry over 90 natural origin ingredients	0.026704	0.817073	17.82641
cosmal cure professional nutri strength mask for dry damaged, cosmal cure professional nutri strength balsam for dry damaged	cosmal cure professional nutri strength shampoo for dry damaged	0.02232	0.918033	11.12727
cosmal cure professional nutri strength mask for dry damaged, cosmal cure professional nutri strength shampoo for dry damaged	cosmal cure professional nutri strength balsam for dry damaged	0.02232	0.888889	13.5165
cosmal cure professional vital shine mask for colored highlighted, cosmal cure professional vital shine balsam for colored highlighted	cosmal cure professional vital shine shampoo for colored highlighted	0.026305	0.970588	13.68093
cosmal cure professional vital shine mask for colored highlighted, cosmal cure professional vital shine shampoo for colored highlighted	cosmal cure professional vital shine balsam for colored highlighted	0.026305	0.929577	15.97473
cosmal cure professional vital shine mask for colored highlighted	cosmal cure professional vital shine balsam for colored highlighted, cosmal cure professional vital shine shampoo for colored highlighted	0.026305	0.717391	14.87549
soft wave hijab shampoo, soft wave hijab mask	soft wave hijab conditioner	0.023914	0.967742	12.20133
soft wave hijab conditioner, soft wave hijab mask	soft wave hijab shampoo	0.023914	0.9375	10.94041
soft wave hijab mask	soft wave hijab shampoo, soft wave hijab conditioner	0.023914	0.705882	11.06912

soft wave hijab shampoo, soft wave hijab oil replacement	soft wave hijab conditioner	0.033878	0.867347	10.93555
soft wave hijab oil replacement, soft wave hijab conditioner	soft wave hijab shampoo	0.033878	0.876289	10.22608
soft wave kids curl gentle hair milk, soft wave kids curl gentle moisturizer	soft wave kids curl gentle shampoo	0.022718	0.934426	17.76118
soft wave kids curl gentle hair milk, soft wave kids curl gentle shampoo	soft wave kids curl gentle moisturizer	0.022718	0.890625	19.77503

b. Association Rules for the FP-growth Algorithm:

Antecedents	**Consequents**	**Support**	**Confidence**	**Lift**
cosmal cure professional fall control balsam for weak thin hair	cosmal cure professional fall control shampoo for weak thin	0.034277	0.796296	11.96352
cosmal cure professional vital shine mask for colored highlighted	cosmal cure professional vital shine balsam for colored highlighted	0.027102	0.73913	12.70191
cosmal cure professional vital shine balsam for colored highlighted	cosmal cure professional vital shine shampoo for colored highlighted	0.048226	0.828767	11.68189
cosmal cure professional vital shine mask for colored highlighted	cosmal cure professional vital shine shampoo for colored highlighted	0.028298	0.771739	10.87805
soft wave kids curl gentle hair milk	soft wave kids curl gentle shampoo	0.025508	0.719101	13.66837
soft wave kids curl gentle moisturizer	soft wave kids curl gentle shampoo	0.038661	0.858407	16.31624
soft wave kids curl gentle shampoo	soft wave kids curl gentle moisturizer	0.038661	0.734848	16.31624
soft wave kids strawberry conditioner over 90 natural origin ingredients	soft wave kids shampoo strawberry over 90 natural origin ingredients	0.026704	0.817073	17.82641
cosmal cure professional nutri strength mask for dry damaged, cosmal cure professional nutri strength balsam for dry damaged	cosmal cure professional nutri strength shampoo for dry damaged	0.02232	0.918033	11.12727
cosmal cure professional nutri strength mask for dry damaged, cosmal cure professional nutri strength shampoo for dry damaged	cosmal cure professional nutri strength balsam for dry damaged	0.02232	0.888889	13.5165
cosmal cure professional vital shine mask for colored highlighted, cosmal cure	cosmal cure professional vital shine shampoo for colored highlighted	0.026305	0.970588	13.68093

professional vital shine balsam for colored highlighted				
cosmal cure professional vital shine mask for colored highlighted, cosmal cure professional vital shine shampoo for colored highlighted	cosmal cure professional vital shine balsam for colored highlighted	0.026305	0.929577	15.97473
cosmal cure professional vital shine mask for colored highlighted	cosmal cure professional vital shine balsam for colored highlighted, cosmal cure professional vital shine shampoo for colored highlighted	0.026305	0.717391	14.87549
soft wave hijab shampoo, soft wave hijab mask	soft wave hijab conditioner	0.023914	0.967742	12.20133
soft wave hijab conditioner, soft wave hijab mask	soft wave hijab shampoo	0.023914	0.9375	10.94041
soft wave hijab mask	soft wave hijab shampoo, soft wave hijab conditioner	0.023914	0.705882	11.06912
soft wave hijab shampoo, soft wave hijab oil replacement	soft wave hijab conditioner	0.033878	0.867347	10.93555
soft wave hijab oil replacement, soft wave hijab conditioner	soft wave hijab shampoo	0.033878	0.876289	10.22608
soft wave kids curl gentle hair milk, soft wave kids curl gentle moisturizer	soft wave kids curl gentle shampoo	0.022718	0.934426	17.76118
soft wave kids curl gentle hair milk, soft wave kids curl gentle shampoo	soft wave kids curl gentle moisturizer	0.022718	0.890625	19.77503

C. Summary Results of the Models:

As a reminder, these models were evaluated under the same threshold parameters:

- Minimum Support Tolerance: 0.02
- Minimum Confidence Tolerance: 0.7
- Minimum Lift Tolerance: 10

Using these parameters, both the apriori and the fp-growth models produced identical results for both itemset and association rules. Therefore, this section will provide a unified general statistics view of the performance of both models:

a. Key Statistics for Frequent Itemset:

- Count: 350
- Mean Support: 0.026046
- Std: 0.021329
- Minimum Support: 0.010363
- Maximum Support: 0.130729
- Q1: 0.012356

- Median Support: 0.017935
- Q3: 0.030590

b. Key Statistics for Association Rules:

The models resulted in 20 association rules:

Statistic	Support	Confidence	Lift
Count	20	20	20
Mean	0.028796	0.843483	13.67181
Std	0.006959	0.088589	2.808361
Min	0.02232	0.705882	10.22609
25%	0.023914	0.763587	11.11273
50%	0.026305	0.862877	13.1092
75%	0.033878	0.920919	16.0601
Max	0.048226	0.970588	19.77503

D. Model Selection

Given that both models generated identical results, computational efficiency was used as a metric to determine the best model for market basket analysis of Cosmaline's transactions. After running the models 10 times and averaging the difference in speed between both:

- The Apriori algorithm took 1.68 seconds to run.
- The running time of FP-Growth algorithm is 0.84 seconds.

The data clearly demonstrates that the FP-Growth algorithm exhibits substantially higher running speeds, with a 50.06% decrease in running time compared to the Apriori algorithm. This statistic, in addition to the industry knowledge that fp-growth is also more memory efficient, makes FP-growth our model of choice for this analysis.

5. Discussion and Recommendation

The objective of the Malia Project was to tackle four key difficulties faced by the organization: automate the integration of data, enhance the company's reconciliation procedures, offer stakeholders with real-time monitoring of financial transactions, and discover e-commerce purchase patterns that are essential for the implementation of marketing effective strategies for the Cosmaline brand.

5.1 Insights on Malia's Data Integrity

An important observation from the reconciliation process is the steady discrepancy between ECOM cash and ERP cash data. For 2023, there was a huge difference of $6,544.85 between the two sources and this difference remained at a large amount of -$5,809.79. This indicates that there is systematic problem at the integration level between the e-commerce data source and the ERP system, which could be detrimental to the company especially when generating financial reports used for decision making.

Furthermore, the discrepancy between e-commerce CyberSource and credit card transactions exhibits interesting changes. The same difference of -$93.55 was carried from 2023 to 2024, revealing a weakness in the previous reconciliation measures of the company. As for the rest of 2024, the difference increased to an amount of -$484.50 before changing direction to $484.50. Going into March 2024, the accounts were balanced out to a difference of $0.01. This clearly highlights the fact that the previous discrepancies were manually corrected and pushed back into the ETL process, with the changes taking till March to be recorded.

As for the other comparisons, excellent integration was observed between e-commerce data transactions and the ones recorded by Aramex as well as between ERP data and Aramex data. This shows a strong integration between Aramex's and Cosmaline's data storage systems. This also holds true for the collection recorded by Cosmaline, however to a lesser extent as a discrepancy between ERP transactions and Cosmaline's transactions showed a non-negligible difference of 9.09%. This indicates that while both shipping collections are mostly aligned, Aramex's integration with the company's storage system is flawless.

Another important insight is the dominance of cash transactions in the company's operations. Both in 2023 and 2024, the data indicated that cash payments accounted for the majority of collections, particularly through Aramex. This indicates a strong consumer preference for paying with cash, which is congruent with the current cash economy in Lebanon. These cash-based transactions usually have increased error rates due to the time lag between when payments are made and when they are recorded in the system, making them more difficult to monitor in real-time compared to digital payments. Fortunately, all evaluated cash transactions collected through Aramex perfectly matched their e-commerce counterpart, indicating that Malia is already heavily mitigating the risks that come with cash-based transactions.

5.2 Forecasting Analysis Insights

The ARIMA(1,1,1) forecasting model revealed that the company's revenues are projected to remain constant in the near future, with no significant increase in sales volume. This stability in the forecast implies that the company needs to focus on boosting their growth since the absence of fluctuations, even if it provides operational stability, could be a sign of stagnation. This situation creates the need for the company to further analyze their e-commerce sales using data mining techniques for product associations.

5.3 Exploring Key EDA and Modeling Insights for Business Strategy

The EDA revealed a strong customer preference for specific product categories, particularly hair care, indicating that Cosmaline's main business is in hair cosmetic products. Furthermore, the

analysis revealed that distribution of items per order skew towards lower quantities, indicating that Cosmaline consumers' prefer to buy smaller quantities rather than in bulk. It was also revealed that a large portion of Cosmaline's offerings is composed of niche products. In fact, our analysis showed that almost 120 products out of 487 were considered as niche, which is congruent with the statistic that a product was purchased on average only 38 times across four months of operations. In addition, some of cosmaline's niche offerings also present as their best sellers. Products such as "Soft Wave Hijab Shampoo" and "Cosmal Cure Professional Sulfate Free Shampoo," dominate sales while also catering to very specific markets. This highlights a clear competitive advantage for Cosmaline when catering for niche and specific market segments, rather than offering generic cosmetic products.

This presence of numerous niche products resulted in very specific association rules that could be used for product bundling and recommendations. However, it is important to note that the presence of niche products heavily affected the support distribution of the market basket analysis models, making the overall support for all rules significantly lower. This could be explained mathematically as support is defined as the proportion of transactions, out of the entire database, in which a particular itemset appears. In a niche market, products are tailored to specific needs and preferences, resulting in the frequency of transactions involving these items to be inherently lower when accounting the entire database. For instance, even though a product combination like "soft wave hijab shampoo" and "soft wave hijab conditioner" might be very common among customers who buy these products, the overall support for this rule will still be low because these products are not as repeated often in the transaction table. This underscores the need to shift the importance to more suitable metrics such as confidence and lift that condition on the transaction having the antecedent of the rules, eliminating the count of all transactions in the frequency calculations.

As for the distribution of confidence and lift for the generated association rules, the plots (available in the appendix) firstly indicate that the scale of confidence and lift is particularly high for all rules. Specifically, for confidence, the distribution spans across a range from 0.7 to 0.95, with most rules being clustered at 0.70, 0.85, and 0.90. As for the lift metric, it typically ranges from 10 to 20, with the highest frequency of lift values being between 10 and 12. This presence of high confidence and lift values indicate that the market basket analysis is reliable and can be used by the company for the formulation of successful marketing strategies such as the creation of bundles and cross-selling opportunities.

5.4 Recommendations:

Following the insights presented in the discussion section, numerous recommendations could be provided for the company to improve its overall positioning:

A. **Improve integration between E-Commerce website data and ERP system:** As mentioned before, the e-commerce and ERP cash transactions exhibit a large discrepancy. This could suggest that the company's link to these sources is weak or inefficient as data could be transferred between them infrequently, leading to discrepancies accumulating over the month. Even though the reconciliation measures implemented detect these differences, it is important to resolve the source of the problem by improving the

synchronization between the two sources. This could be done through API integrations that would allow the ECOM platform and the ERP system to communicate and exchange data in real-time, eliminating the need for potential manual entry synchronization.

B. Improve integration between E-Commerce CyberSource transactions and credit card data source:

Similar measures of integration and data synchronization should be taken for the E-Commerce and credit card sources. The studied fluctuations of amount balance between the two (-$481 to +$481 to 0) from January 2024 to March 2024 clearly points to discrepancy that was either manually adjusted, or that was automatically adjusted after a long time. This suggests that the current adjusting mechanism of the company is not as effective and timely, indicating the need to use more advanced measures such as API integrations that would resolve most of these differences at the source level.

C. Rely fully on Aramex for shipping Cosmaline products:

As shown in the reconciliation results, the e-commerce cash collections perfectly match the ones souring from Aramex across the four-month studied. Similar results were found for e-commerce and Cosmaline cash collections. However, slight discrepancies occurred on some occasions which could potentially be caused by the difficulties of dealing with cash-based transactions. This indicates that Aramex's collection systems are flawlessly integrated with the ones at Malia. Therefore, Malia Group could expand on that partnership and rely fully on Aramex to ship their products. This would eliminate the complexities of using two collection methods and would completely outsource the delivery process to Aramex.

D. Order Size-based Promotions:

As mentioned before, Cosmaline's customers tend to purchase smaller baskets, with the majority distribution clustered around less than 10 items and more specifically between 1 and 5. In addition, customers tend to purchase individual items in smaller quantities since most of the quantities per item are centered around 1 to 3 pieces. These two observations indicate that Cosmaline's customers mostly buy baskets of under 10 pieces, with the items in that basket not exceeding 1 to 3 pieces individually, which clearly points to customers' preference for diverse carts. This is also proven by the unique items per basket frequency being centerd around 6 to 10 items. Furthermore, the monetary distribution of carts shows that the vast majority of carts are valued at below $15, with few touching the $20 mark. This important metric, with the conclusion that Cosmaline's customers are generally open to purchase diverse products, presents vast opportunities for the company to promote their low selling items while incentivizing customers to increase the value of their carts. To be more precise, while Cosmaline's best sellers revolve around hair care followed by body care, cosmetics such as makeup, lips beaty, and nail care are at the low end of sale volumes. Such a pattern is expected as Cosmaline is mostly known for their hair products rather than their cosmetics, which is partially due to the lack of sufficient customer awareness of this segment.

To improve the sales of these categories while also increasing the value of their takeout carts, Cosmaline could offer items from the low ending categories for free to customers that purchase baskets of over 20$. This strategy would have a double-faceted effect on the company as it would increase the value of customers while promoting their unpopular products and potentially leading customers to buy the products again after trying the free samples. These measures could also be followed by personalized emails to the targeted customers, which would provide much needed feedback about their low selling products.

E. Leveraging Delivery Fees:

Another strategy that could be used to increase the sales of low ending categories would be to leverage the delivery fees imposed by shipping companies. Cosmaline could provide their customers with a chance to eliminate delivery fees if they buy from certain low performing categories.

F. Product Budling:

In addition to promoting low performing products, Cosmaline should also boost the sales of their competitive items. These strategies would be based on the market basket analysis conducted, which by nature includes high performing products. Looking at the association rules and considering that all the rules have high confidence and lift, the company could create a multitude of bundles. The presence of high confidence and lift suggests that customers who buy one product are very likely to buy the associated product, which indicates that the products in question naturally complement each other. Based on the results, bundles could include:

a. **Bundle for Colored Highlighted Hair:** including "Cosmal Cure Professional Vital Shine Mask for Colored Highlighted", "Cosmal Cure Professional Vital Shine Balsam for Colored Highlighted", and "Cosmal Cure Professional Vital Shine Shampoo for Colored Highlighted".

b. **Hijab Hair Care Bundle:** including "Soft Wave Hijab Shampoo", "Soft Wave Hijab Conditioner", "Soft Wave Hijab Mask", and "Soft Wave Hijab Oil Replacement".

c. **Kids Curly Hair Care Bundle:** including "Soft Wave Kids Curl Gentle Hair Milk", "Soft Wave Kids Curl Gentle Shampoo", and "Soft Wave Kids Curl Gentle Moisturizer".

d. **Damaged Hair Care Bundle:** including "Cosmal Cure Professional Nutri Strength Mask for Dry Damaged", "Cosmal Cure Professional Nutri Strength Balsam for Dry Damaged", and "Cosmal Cure Professional Nutri Strength Shampoo for Dry Damaged".

e. **Kids Wavy Hair Care Bundle:** including "Soft Wave Kids Strawberry Conditioner Over 90 Natural Origin Ingredients", and "Soft Wave Kids Shampoo

Strawberry Over 90 Natural Origin Ingredients."

f. **Fall Control Hair Care Bundle:** "Cosmal Cure Professional Fall Control Balsam for Weak Thin Hair", and "Cosmal Cure Professional Fall Control Shampoo for Weak Thin Hair"

G. Cross-Selling:

Depending on the company's need to sell bundles when trying to liquidate inventory, all items in the bundle could also be crossed sold at checkout. This would provide Cosmaline with added revenue from individual item selling with a respectable turnover of customers that would cross buy the items promoted. However, the success of this strategy heavily depends on the website's ability to effectively promote the related items. As shown in the *screenshot in the apprendix (Figure Ecom.1)*, the website does not seem to promote related items to add at checkout, which might lead to customers rarely cross buying products. This section of the website should be altered to show items that are related based on the market basket analysis conducted.

6. Limitations and Challenges:

Although the Malia project effectively tackled the difficulties faced by the company, certain limitations affected the outcomes of the analysis:

1. **Absence of Customer Data:** A major constraint was the lack of customer data. The absence of variables such as demographic and preference details about Cosmaline's customers made sophisticated analysis such as Recency, Frequency, Monetary (RFM) and customer segmentation using techniques such as nearest neighbors impossible. These methods would have enabled the company to implement more customer centered strategies such as loyalty programs and personalized promotions through push notifications. A demonstration of a personalized email that could be sent to loyal customers for exclusive offers can be found in the Appendix (*Figure Demo.1*)

2. **Relatively small dataset:** the dataset used for the data mining and forecasting, although generated effective results, is considered relatively small for such analytical methods. The absence of seasonality in the data prevented the forecasting models from identifying patterns that are more relevant in the context of e-commerce. Also, the relatively small number of transactions added to the association rules having substantially low support levels. While the current results are effective enough to develop sales boosting strategies, the company should invest in gathering as much data as possible to better optimize the models, especially the forecasting analysis.

3. **Challenge in ensuring no duplication in the database:** a point of challenge during the system development was ensuring that no duplicates were found in the database. Given that the system is developed to accommodate constant stream of data, it is essential that

robust duplicate detection tools are present to ensure data integrity. To implement this feature, a system was developed to place all incoming transactions on a temporary table. The table was then compared to its permanent counterpart for duplicates and only non-duplicate transactions were pushed to the permanent table. This process is straightforward, however hard to implement as data types specified in the transformation pipelines would fluctuate before insertion due to it being intermediately stored locally on the system. The challenge was finally resolved by utilizing parquet files for local storing, which ensured that the metadata of the panda's data frame was recorded.

7. *Conclusion*:

The Malia Project tackled the current challenges faced by the company regarding the manual data input, the insufficient data reconciliation, the lack of real time data monitoring, and the potential of boosting sales across their e-commerce website.

To resolve these challenges, an Airflow system was first developed to automate the Extract, Transform, and Load processes from the different data sources to a centralized reconciled database. This automated system eliminated the need for manual entry and guaranteed smooth integration of data from internet orders, ERP transactions, and shipping collections. As for the data reconciliation difficulties, they were successfully overcome by implementing SQL scripts that automatically detected and rectified differences in transactions between all data sources. In addition to automated reconciliation processes, complementary Power BI dashboards were created to provide real time monitoring of financial transactions. Using the dashboards, stakeholders have the ability to track transactions at an individual level and across sources, while also gaining insights on their collection metrics such as the proportion of cash to CyberSource transactions. Furthermore, the project focused on providing a deep analysis of the company's e-commerce performance. This study started with a forecasting analysis of future sales and continued with an extensive EDA and market basket analysis (Apriori and FP-Growth algorithms) that provided useful insights into client buying behavior.

Analysis first showed that there is a pressing integration problem between the company's e-commerce transactions and their counterparts present in the ERP source, which highlights the need for the company to improve this link through more robust and automated data APIs. Similar measures could be implemented for the e-commerce and credit card sources as the fluctuation in discrepancies shows that errors between these two sources seem to be resolved manually by the company, causing corrections to be greatly delayed. Furthermore, the analysis of Cosmaline's e-commerce transactions indicated that the shop's customers largely take out a diverse basket, which presents vast opportunities for the company to promote their less performing product categories

such as makeup, cosmetics, and nail care products. Finally, the data mining analysis revealed multiple bundles of products such as the Hijab Hair Care Bundle, the Kids Curly Hair Care Bundle, and the Damaged Hair Care Bundle which could be promoted to boost sales through direct bundle selling or through the cross selling of items in the bundle at checkout.

To conclude, the project provides Malia Group with a robust set of tools to resolve the current data challenges they face. The project automates the entire business analysis pipeline from the ETL to the reconciliation and the drilled down analysis of relevant data. This comprehensive approach positions the company short term boosts in sales performance and for long-term expansion.

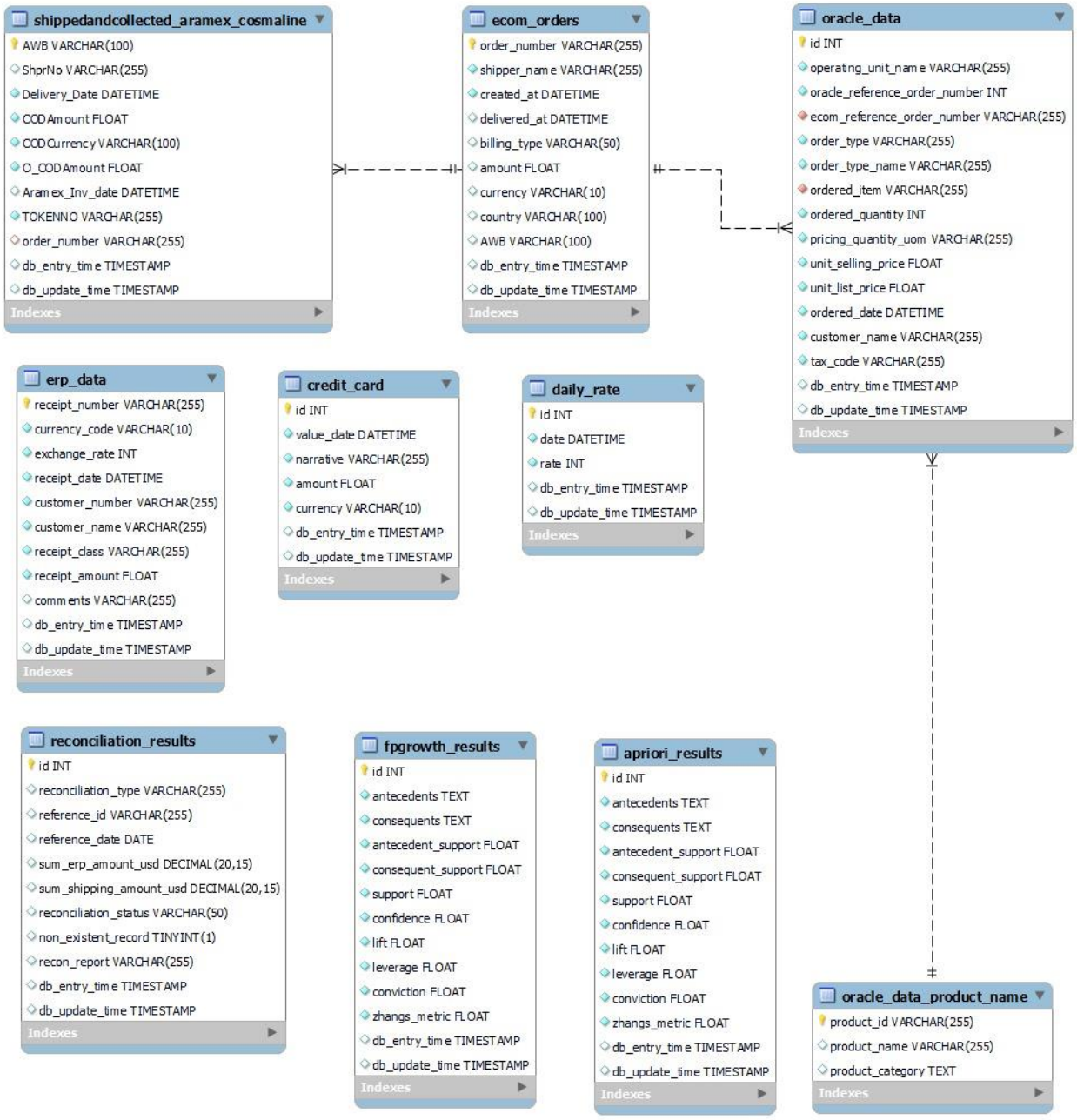
8. References

1. Albrecht, A., & Naumann, F. (2008). Managing ETL Processes. ACM, 12–15. http://www.hpi.uni-potsdam.de/fileadmin/hpi/FG_Naumann/publications/2008/ETL_Management.pdf
2. Anas, S., Rumui, N., Roy, A., & Saputro, P. H. (2022). Comparison of Apriori algorithm and FP-Growth in managing store transaction data. *International Journal of Computer and Information System (IJCIS), 3(4)*, 158–162. https://doi.org/10.29040/ijcis.v3i4.96
3. Asana, I. M. D. P., Wiguna, I. K. A. G., Atmaja, K. J., & Sanjaya, I. P. A. (2020). FP-Growth Implementation in Frequent itemset Mining for Consumer Shopping Pattern Analysis Application. *Jurnal Mantik, 4(3)*, 2063–2070. https://doi.org/10.35335/mantik.vol4.2020.1075.pp2063-2070
4. Bakhtouchi, A. (2020). Data reconciliation and fusion methods: A survey. *Applied Computing and Informatics, 18(3/4)*, 182–194. https://doi.org/10.1016/j.aci.2019.07.001
5. Biswas, N., Sarkar, A., & Mondal, K. C. (2019). Efficient incremental loading in ETL processing for real-time data integration. *Innovations in Systems and Software Engineering, 16(1)*, 53–61. https://doi.org/10.1007/s11334-019-00344-4
6. Bowen, T., Zhe, Z., & Yulin, Z. (2020). Forecasting method of e-commerce cargo sales based on ARIMA-BP model. *IEEE*. https://doi.org/10.1109/icaica50127.2020.9181926
7. Carta, S., Medda, A., Pili, A., Recupero, D. R., & Saia, R. (2018). Forecasting E-Commerce products prices by combining an autoregressive Integrated Moving Average (ARIMA) model and Google Trends data. *Future Internet, 11(1)*, 5. https://doi.org/10.3390/fi11010005
8. Elom, E. (2023). Data Aggregation ETL pipeline and Business intelligence system. *ResearchGate*. https://www.researchgate.net/publication/374914278
9. Goar, V., Sarangdevot, P. S., Tanwar, G., & Sharma, D. A. (2010). Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse. *International Journal on Computer Science and Engineering, 2*. https://www.enggjournals.com/ijcse/doc/IJCSE10-02-03-108.pdf
10. Hossain, M., Sattar, A. H. M. S., & Paul, M. K. (2019). Market Basket Analysis Using Apriori and FP Growth Algorithm. *International Conference on Computer and Information Technology*. https://doi.org/10.1109/iccit48885.2019.9038197

11. Jirapatsil, P., & Phumchusri, N. (2022). Market Basket Analysis for Fresh Products location improvement: A case study of E-Commerce Business Warehouse. *Association of Computing Machinery*. https://doi.org/10.1145/3535782.3535786
12. Jain, A., Karthikeyan, V., B, S., Br, S., K, S., & S, B. (2020). Demand Forecasting for E-Commerce platforms. *2020 IEEE International Conference for Innovation in Technology (INOCON)*. https://doi.org/10.1109/inocon50539.2020.9298395
13. Kosadi, F., Ginting, W., & Merliana, V. (2021). Digital receipts of online transactions in the reconciliation process and the preparation of financial reports. *Journal of Indonesian Economy and Business, 36(1)*, 31. https://doi.org/10.22146/jieb.59884
14. Krishna, V. R., Kiran, V. S., & Kiran, K. R. (2010). Web based ETL component extended with loading and reporting facilitations a financial application tool. *International Conference on Software Technology and Engineering (ICSTE)*. https://doi.org/10.1109/icste.2010.5608874
15. Kurnia, Y., Isharianto, Y., Giap, Y. C., Hermawan, A., & Riki, N. (2019). Study of application of data mining market basket analysis for knowing sales pattern (association of items) at the O! Fish restaurant using apriori algorithm. *Journal of Physics Conference Series, 1175*, 012047. https://doi.org/10.1088/1742-6596/1175/1/012047
16. Onuoha, L. N., & Amponsah, E. B. (2012). Bank reconciliation as a due process imperative for effective financial management. *Canadian Social Science, 8(3)*, 52–56. https://doi.org/10.3968/j.css.1923669720120803.2955
17. Panjaitan, S., Sulindawaty, N., Amin, M., Lindawati, S., Watrianthos, R., Sihotang, H. T., & Sinaga, B. (2019). Implementation of Apriori algorithm for analysis of consumer purchase patterns. *Journal of Physics Conference Series, 1255(1)*, 012057. https://doi.org/10.1088/1742-6596/1255/1/012057
18. Pham, P. (2020). A case study in developing an automated ETL solution: concept and implementation. https://www.theseus.fi/handle/10024/340208
19. Qisman, M., Rosadi, R., & Abdullah, A. S. (2021). Market basket analysis using apriori algorithm to find consumer patterns in buying goods through transaction data (case study of Mizan computer retail stores). *Journal of Physics. Conference Series, 1722(1)*, 012020. https://doi.org/10.1088/1742-6596/1722/1/012020
20. Radhakrishna, V., SravanKiran, V., & Ravikiran, K. (2012). Automating ETL process with scripting technology. *NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING*. https://doi.org/10.1109/nuicone.2012.6493217
21. Root, R., & Mason, C. (2012). Beginning the ETL Process with SSIS. In Apress eBooks (pp. 253–300). https://doi.org/10.1007/978-1-4302-3489-0_7
22. Shubha, B.G. & Prasad, A.M. (2019). Airflow Directed Acyclic Graph. *Journal of Signal Processing*, 5–2. http://doi.org/10.5281/zenodo.3247274
23. Ünvan, Y. A. (2020). Market basket analysis with association rules. *Communication in Statistics- Theory and Methods, 50(7)*, 1615–1628. https://doi.org/10.1080/03610926.2020.1716255
24. Vase, T. (2015). Advantages of Docker. University of Jyväskylä. https://jyx.jyu.fi/handle/123456789/48029
25. Woodall, P., Borek, A., Oberhofer, M. A., & Gao, J. (2016). Data quality Problems in ETL: The state of the practice in large organisations. *ICIQ*, 54–64. https://dblp.uni-trier.de/db/conf/iq/iq2016.html#WoodallBOG16

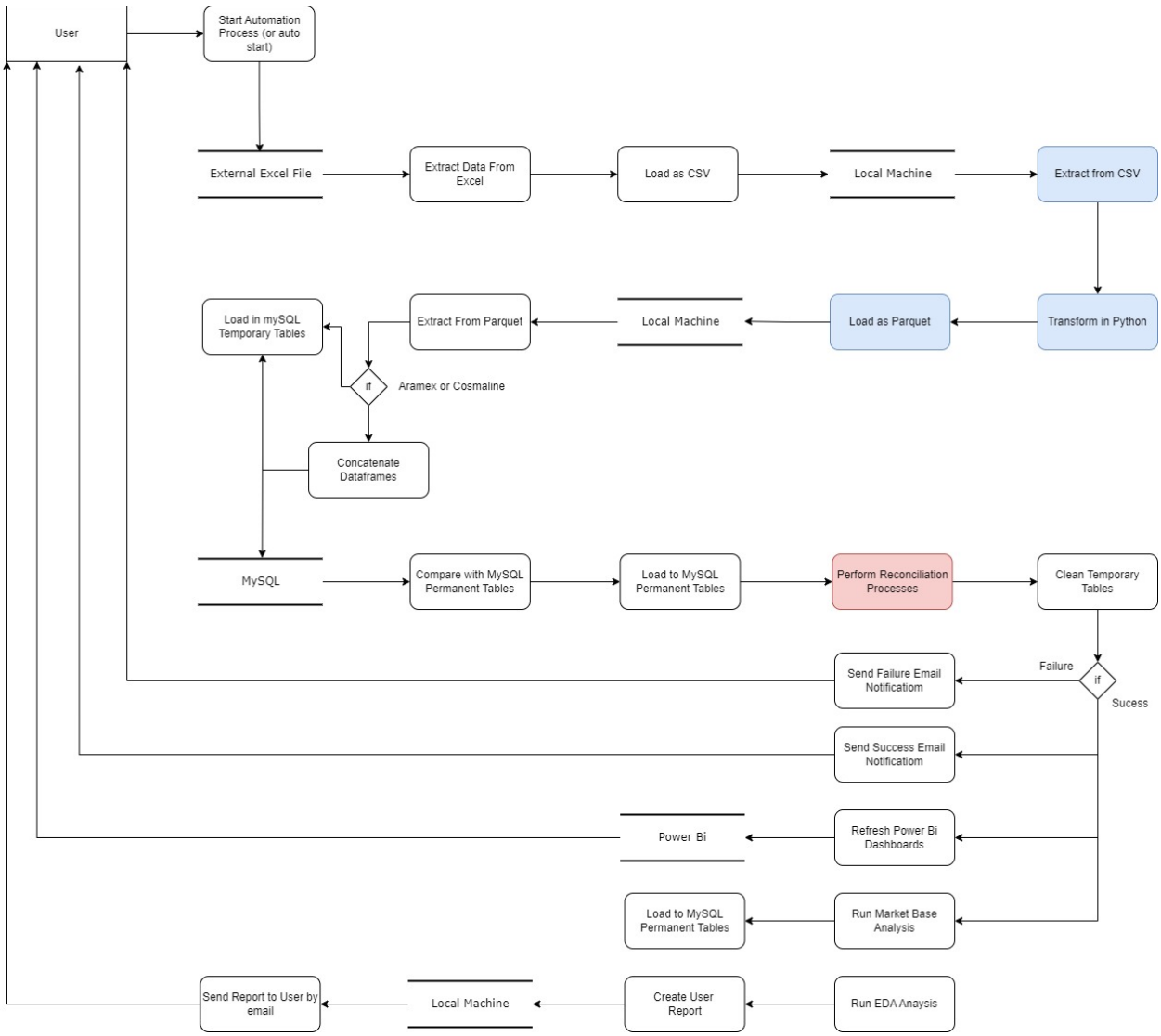
26. Wu, J., Bein, D., Huang, J., & Kurwadkar, S. (2023). ETL and ML Forecasting Modeling Process Automation System. *AHFE International*. https://doi.org/10.54941/ahfe1003775
27. The beauty market in 2023: A special State of Fashion report. (2023, May 22). McKinsey & Company. https://www.mckinsey.com/industries/retail/our-insights/the-beauty-market-in-2023-a-special-state-of-fashion-report

9. Appendix:

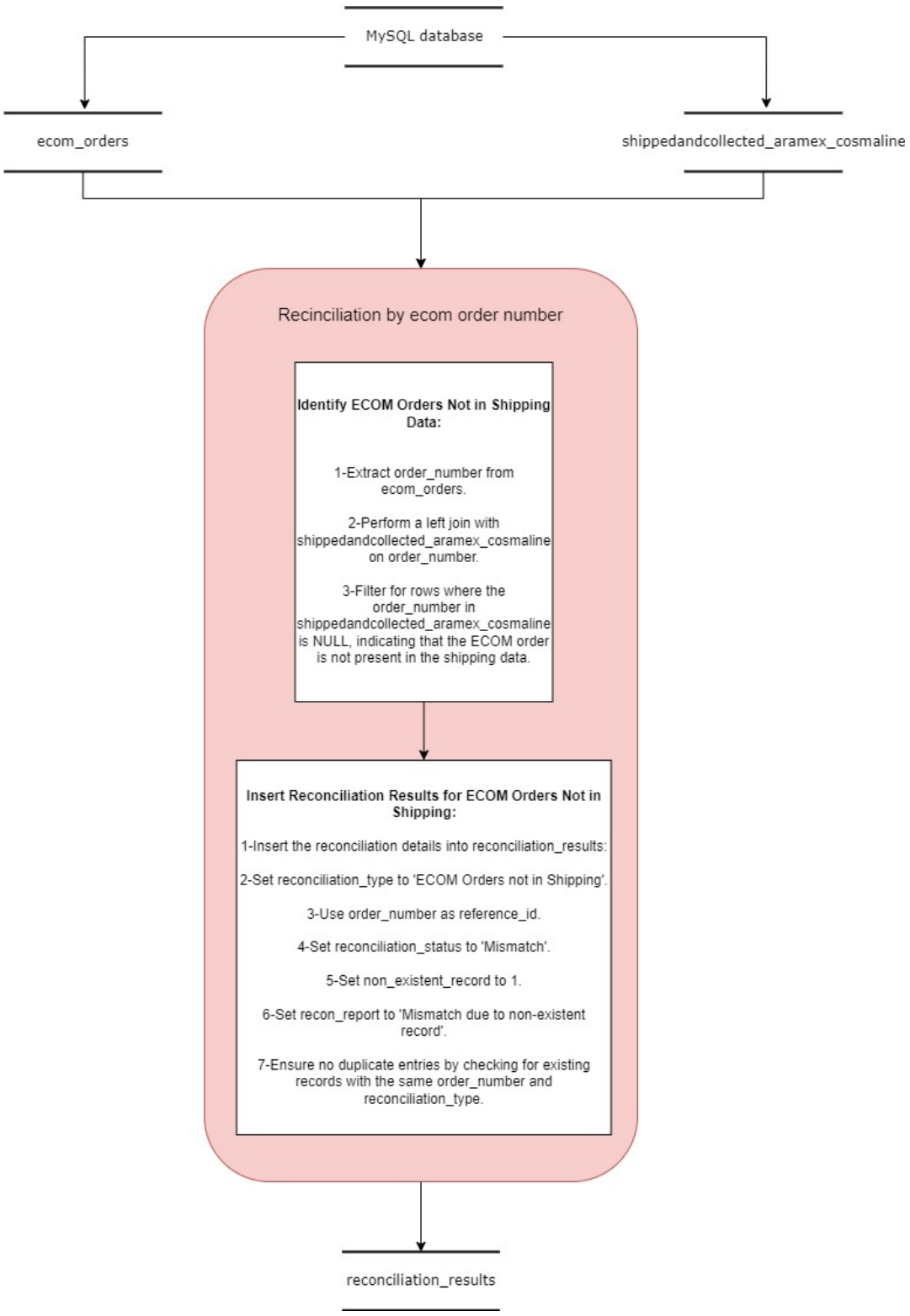


Schema.1 – Data base schema diagram

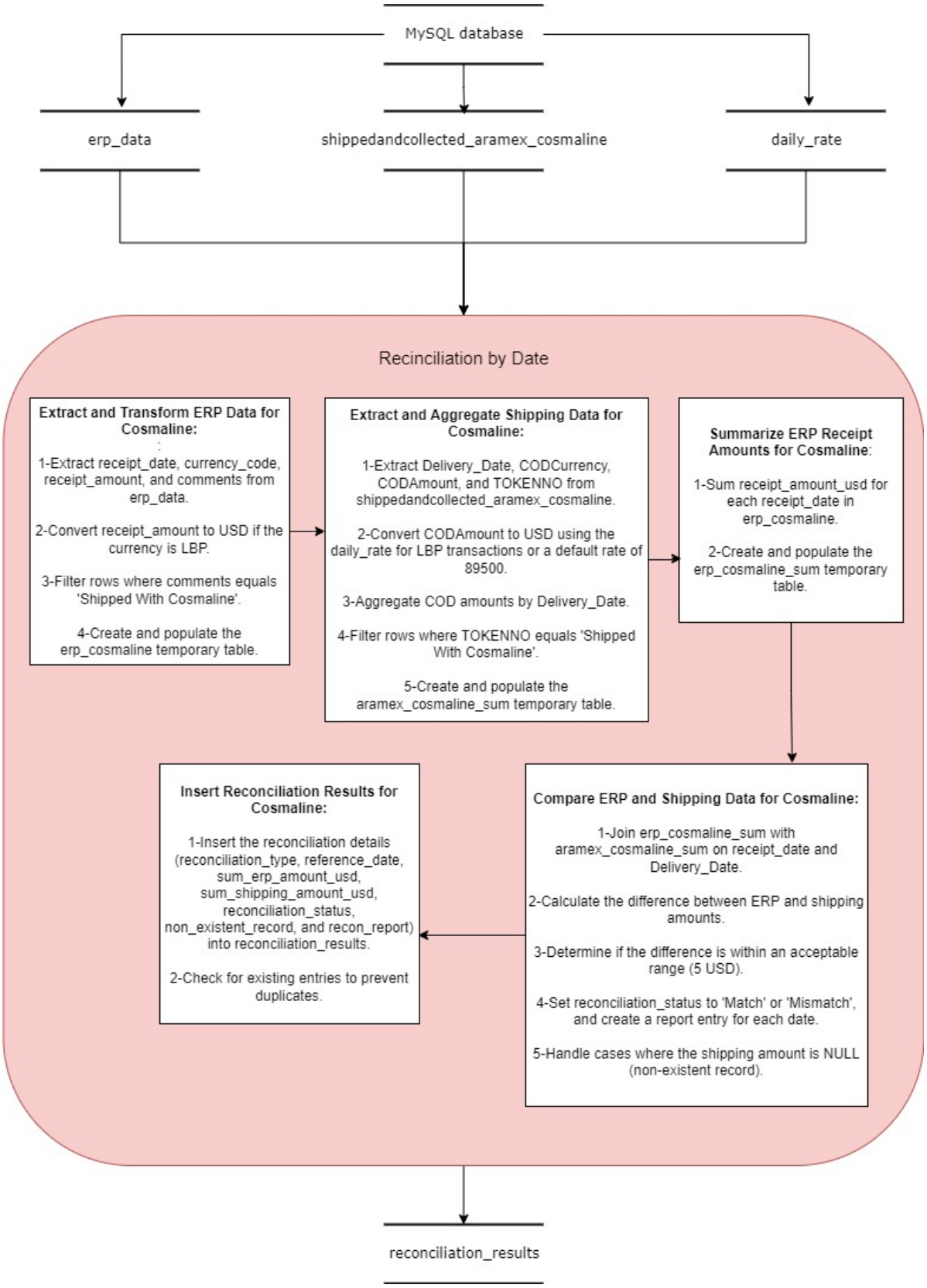
P.1 – Preprocessing Pipelines



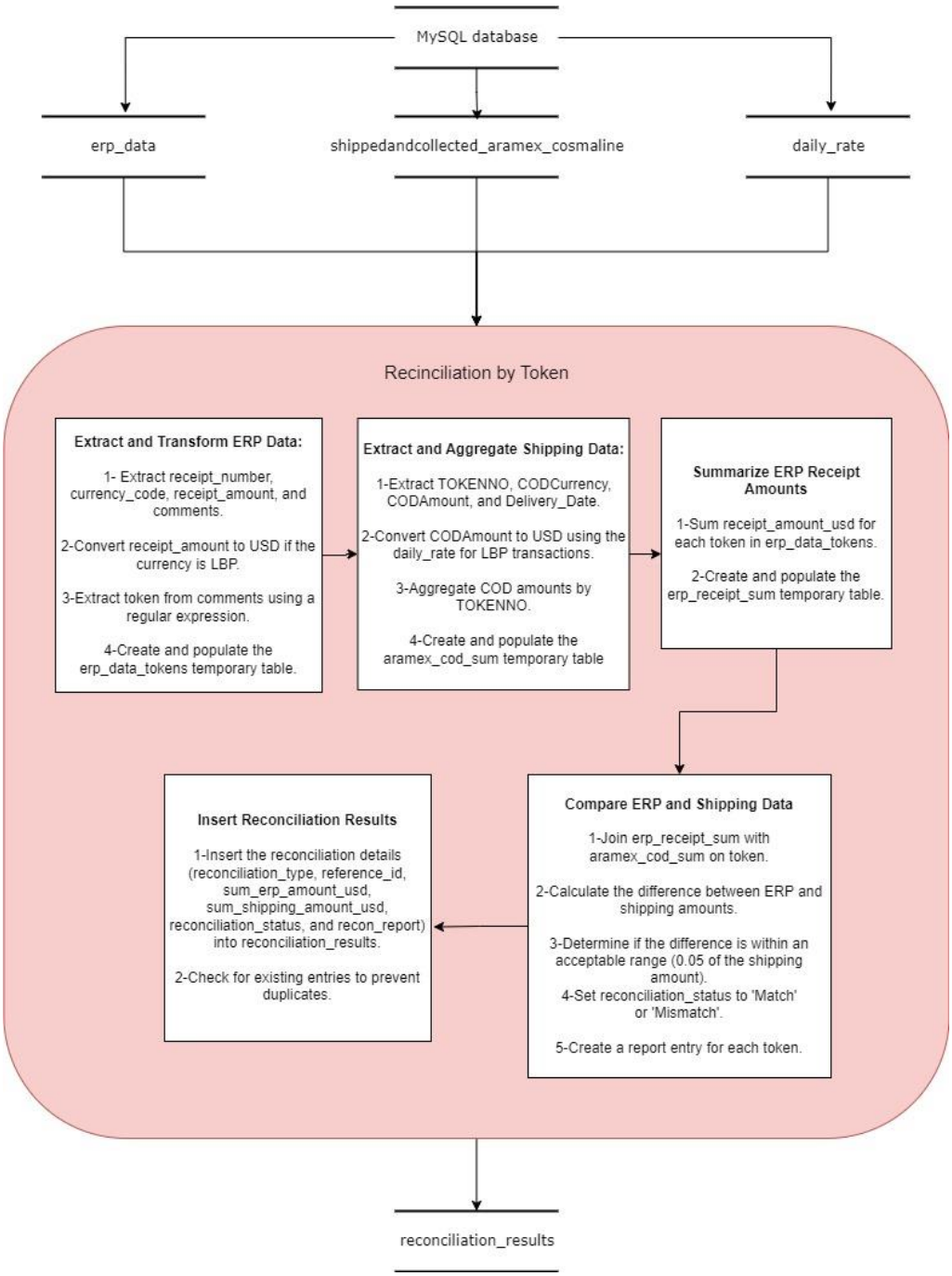
D.1 – System Data Flow Diagram



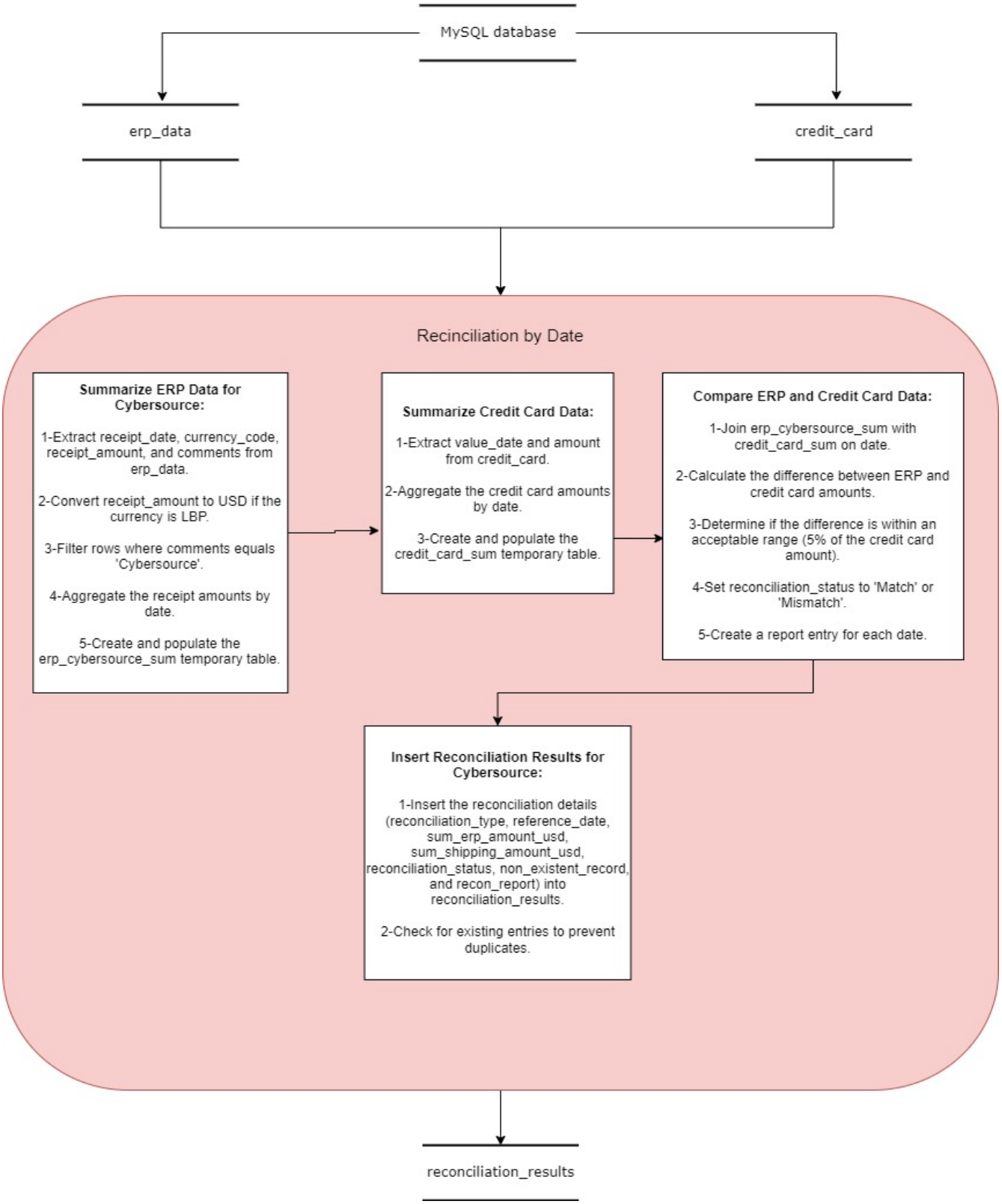
R.1 – ecom_orders_not_in_shipping_reconciliation



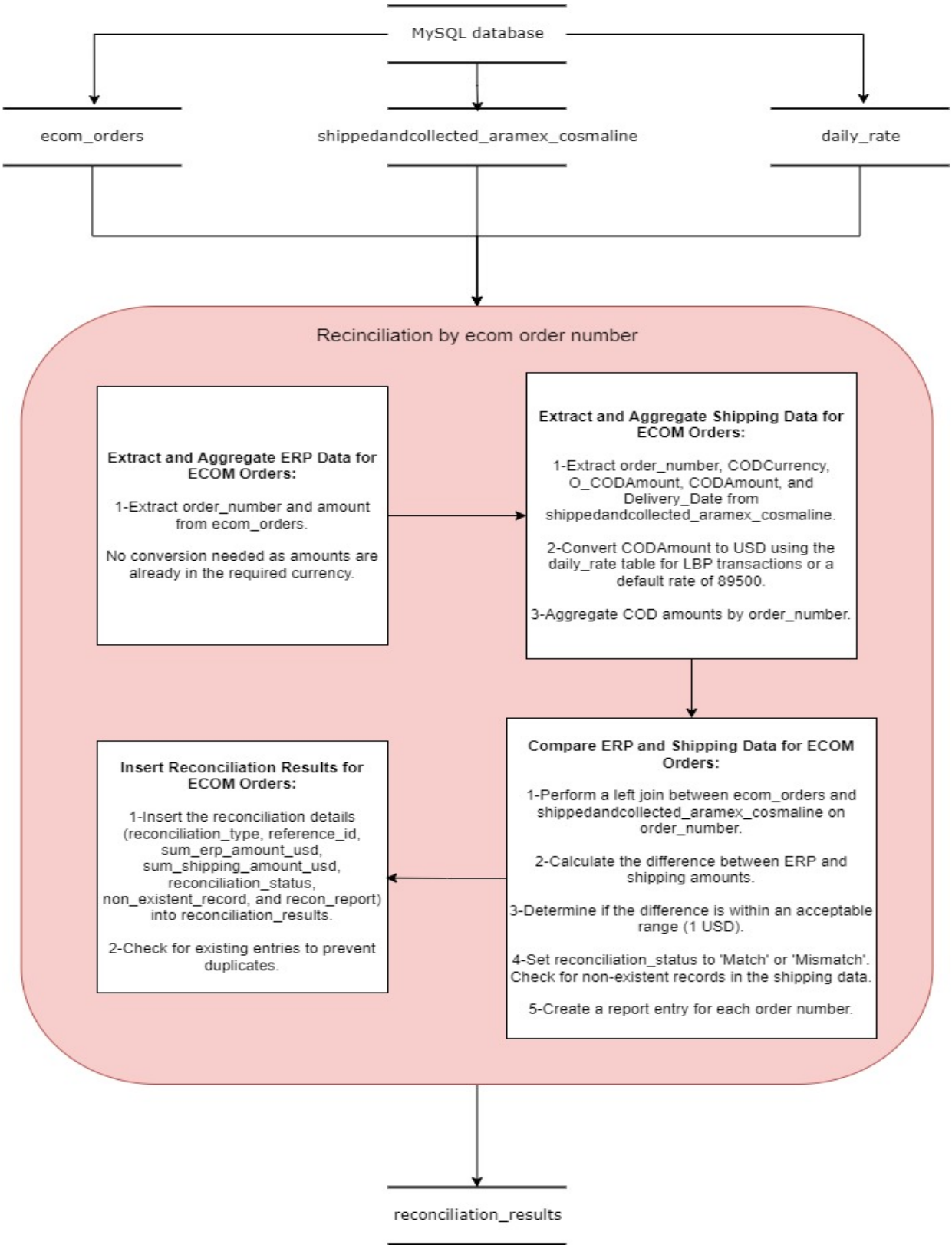
R.2 – cosmaline_reconciliation



R.3 – token_reconciliaiton



R.4 – credit_card reconciliation



R.5 – ecom_reconciliation

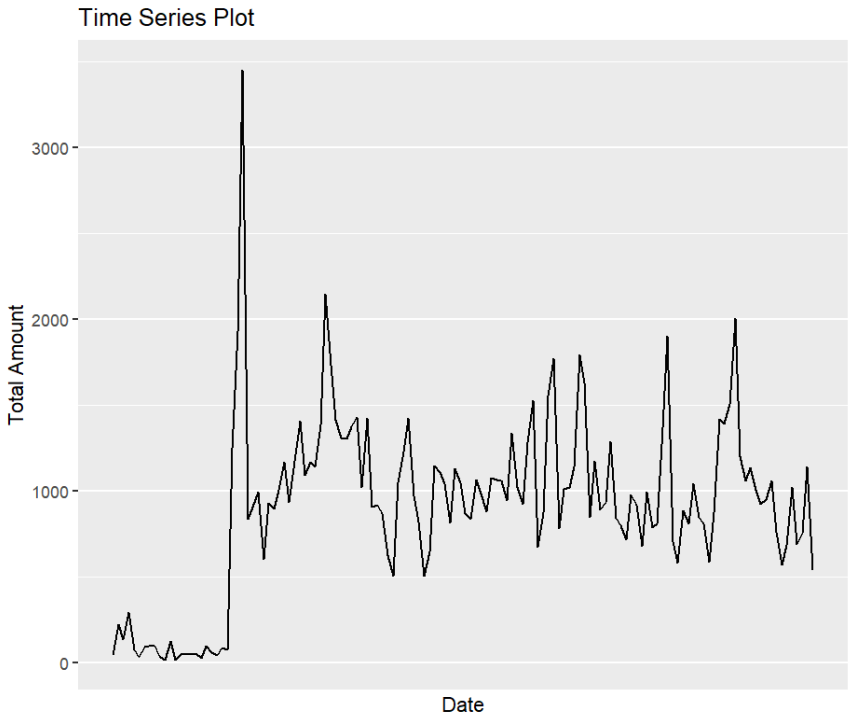


Figure F.1 - Autoplot

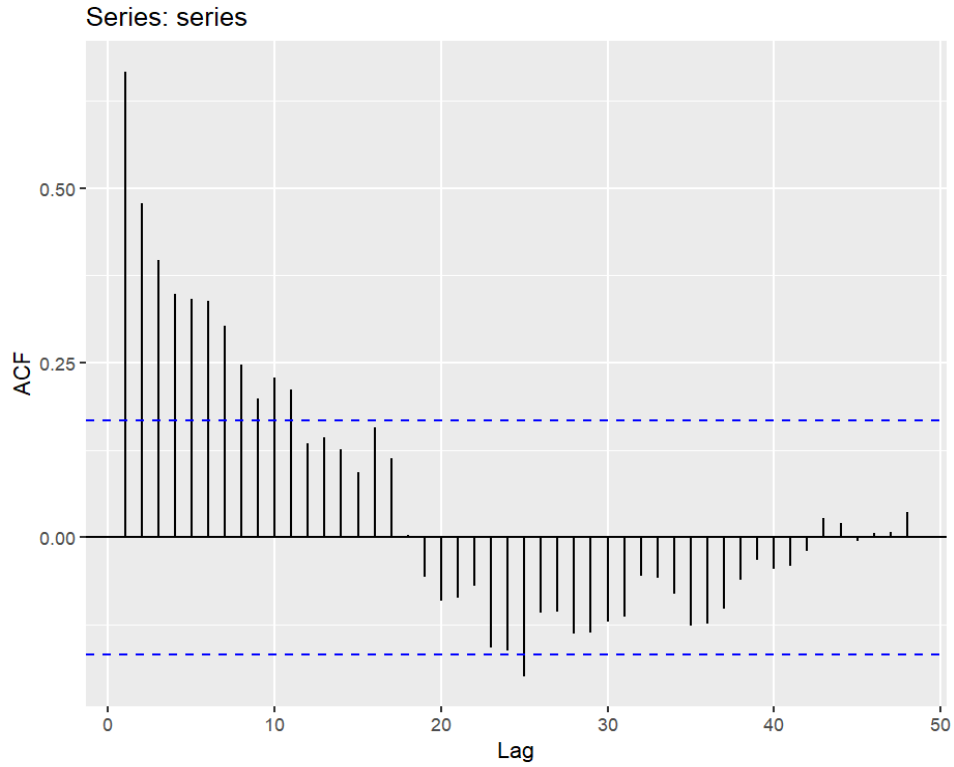


Figure F.2 – Auto Correlation Function ACF

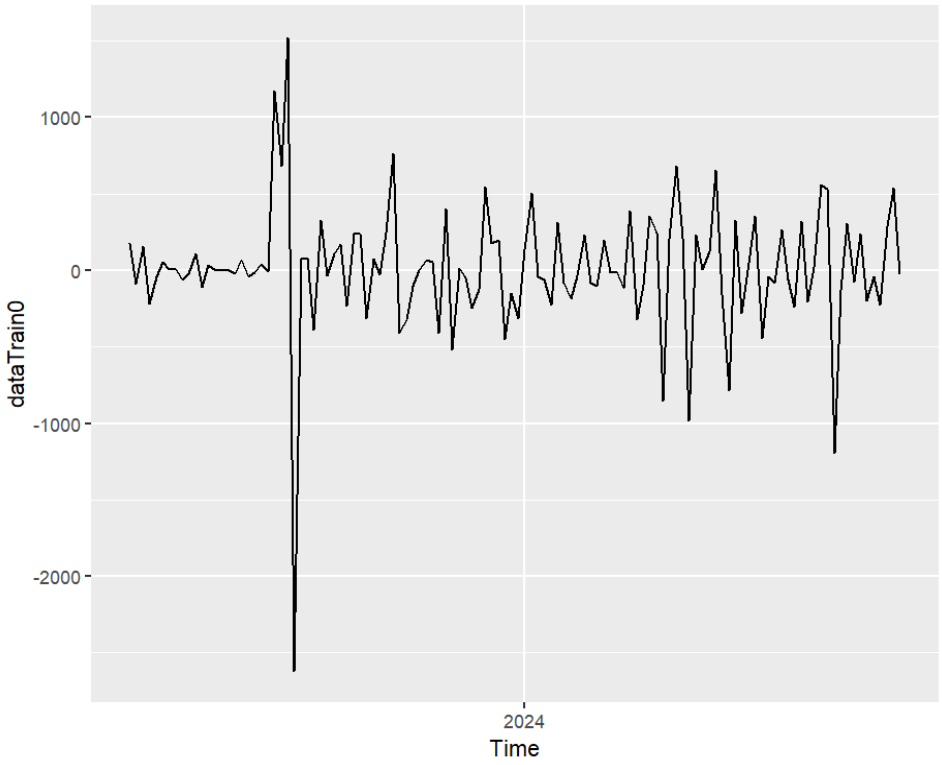


Figure F.3 – Autoplot of Training Data after 1st Differencing

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 4 lags.

Value of test-statistic is: 0.6732

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Figure F.4 – KPSS Test of Training Data after 1st Differencing

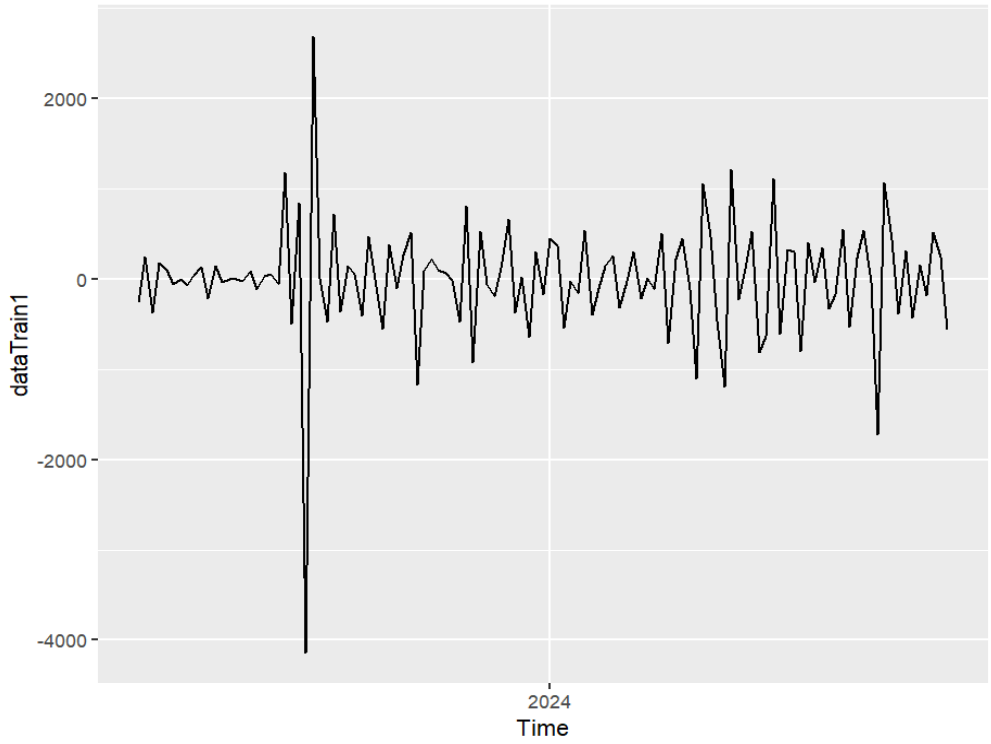


Figure F.5 – Autoplot of Training Dara after 2nd Differencing

```
#########################
# KPSS Unit Root Test #
#########################

Test is of type: mu with 4 lags.

Value of test-statistic is: 0.0224

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Figure F.6 – KPSS Test of Training Data after 2nd Differencing

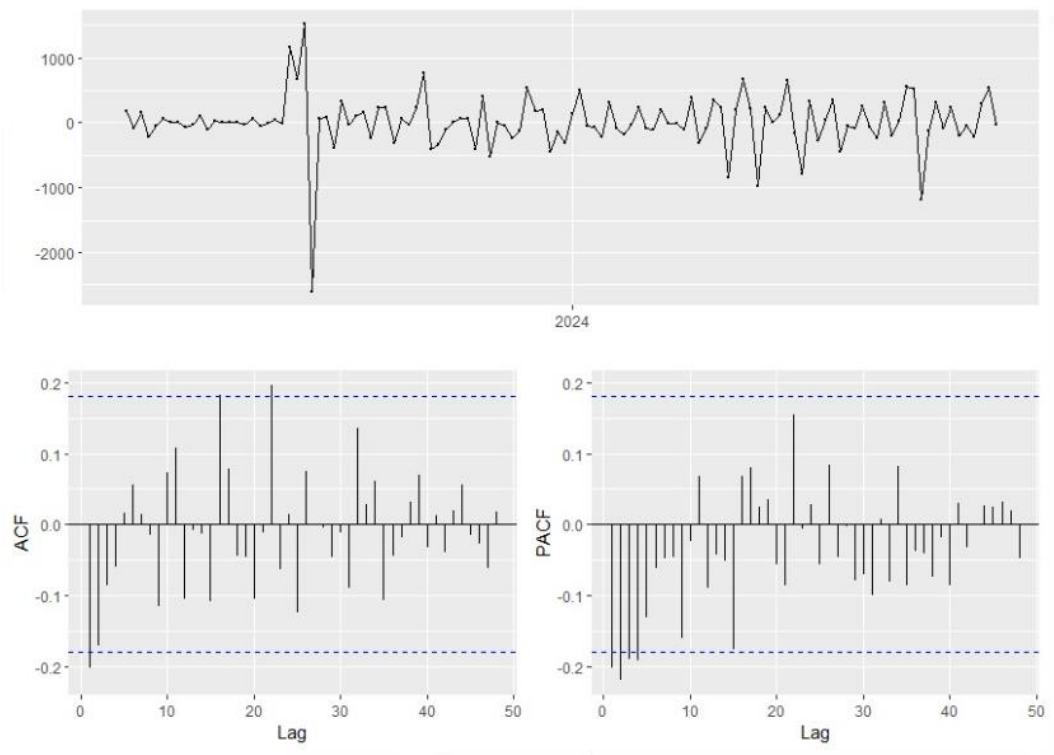


Figure F.7 – GGTS Display after 1st Differencing

	Model	MAPE	RMSE	AICc	Ljung_Box_pvalue	Residuals
1	ARIMA(1, 1, 1)	24.47302	339.1476	1750.36	0.6523	No Autocorrelation
2	ARIMA(2, 1, 1)	24.85088	354.1572	1751.49	0.7690	No Autocorrelation
3	ARIMA(0, 2, 1)	58.83586	561.1928	1761.35	0.1155	No Autocorrelation

Figure F.9 – ARIMA Model Results Table

	Model	MAPE	RMSE	AICc	Ljung_Box_pvalue	Residuals
1	ETS(A, A, N)	46.10862	465.6360	2011.24	0.0862	No Autocorrelation
2	ETS(M, A, N)	104.66548	1024.6133	2178.92	0.0528	No Autocorrelation
3	ETS(M, M, N)	199.38079	1929.6382	2133.74	0.1305	No Autocorrelation
4	ETS(A, N, N)	38.74234	409.5922	2007.11	0.0886	No Autocorrelation

Figure F.10 – ETS Model Results Table

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	-96.47582	365.9072	282.9767	-20.55328	30.77836	0.5375141	1.919144

Figure F.11 – ARIMA (1,1,1) Detailed Results

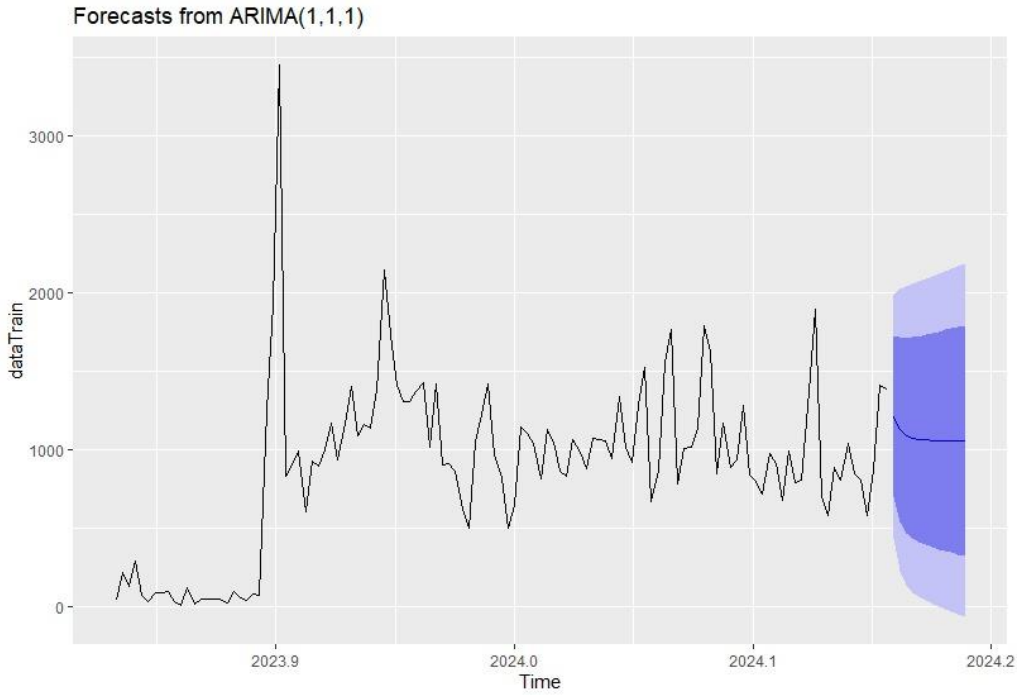


Figure F.12 – ARIMA(1,1,1) Autoplot

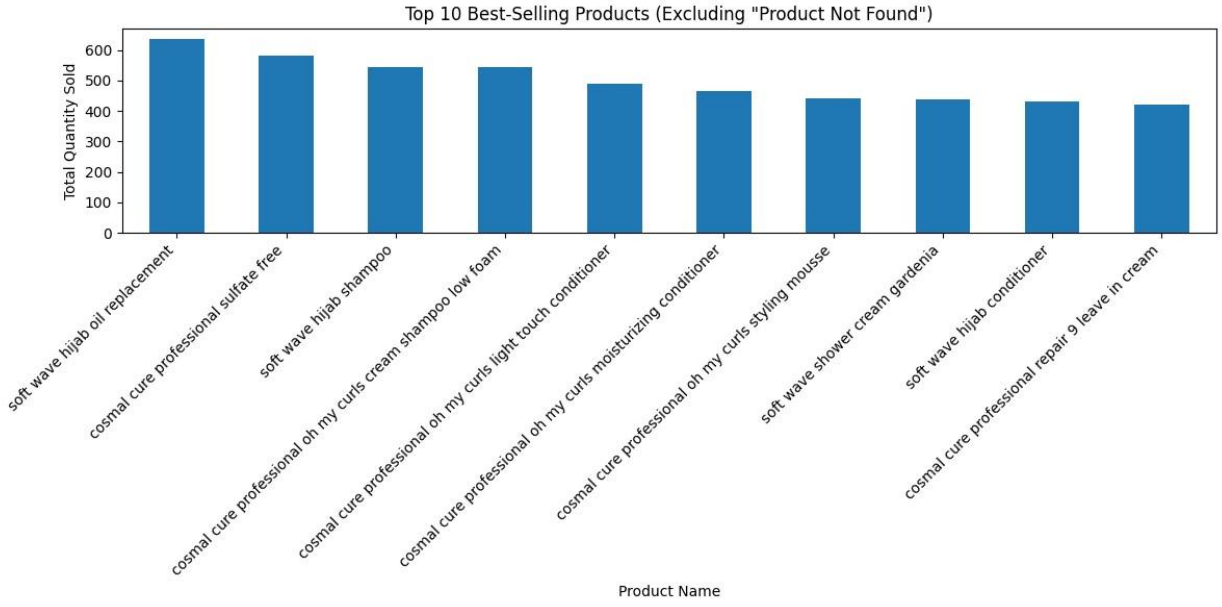


Figure E.1

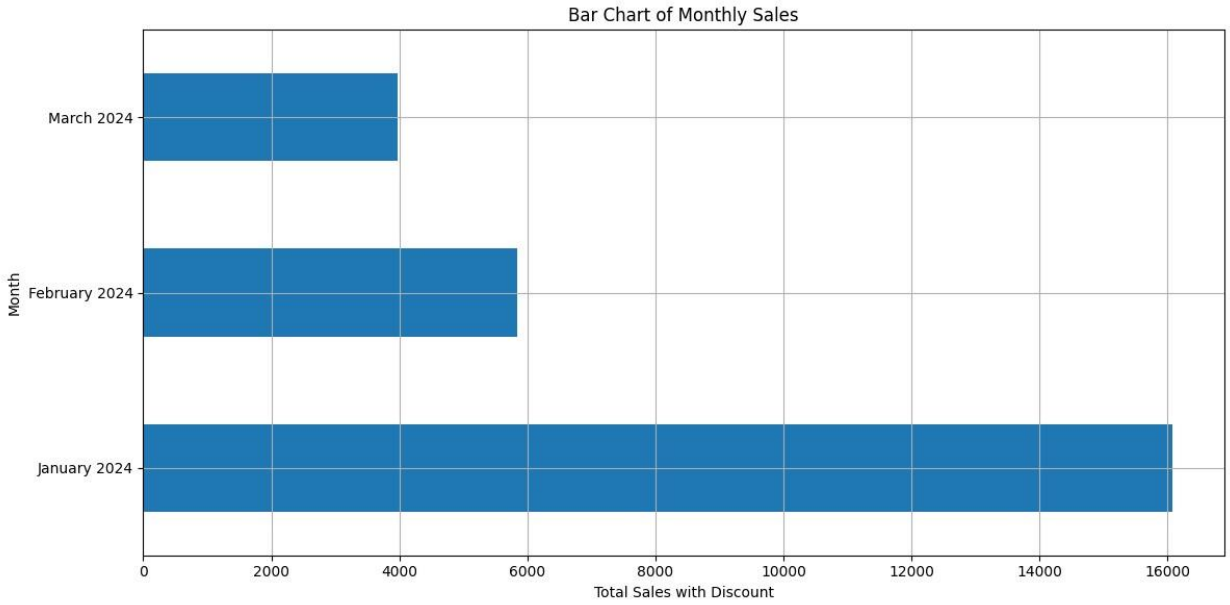


Figure E.2

Product_category	**Ordered_quantity**
Hair Care	12125
Curly & Wavy Hair	6716
Best Sellers	6667
Free From Sulfate & Silicone	6534
Minis	4283
Frizzy or Dry & Damaged Hair	3238
Kids	2971
Shower Gel & Cream	2932
Colored or Blonde or Grey Hair	2373
Natural	2246

Figure E.3 – Sales Volume of Top 10 Categories

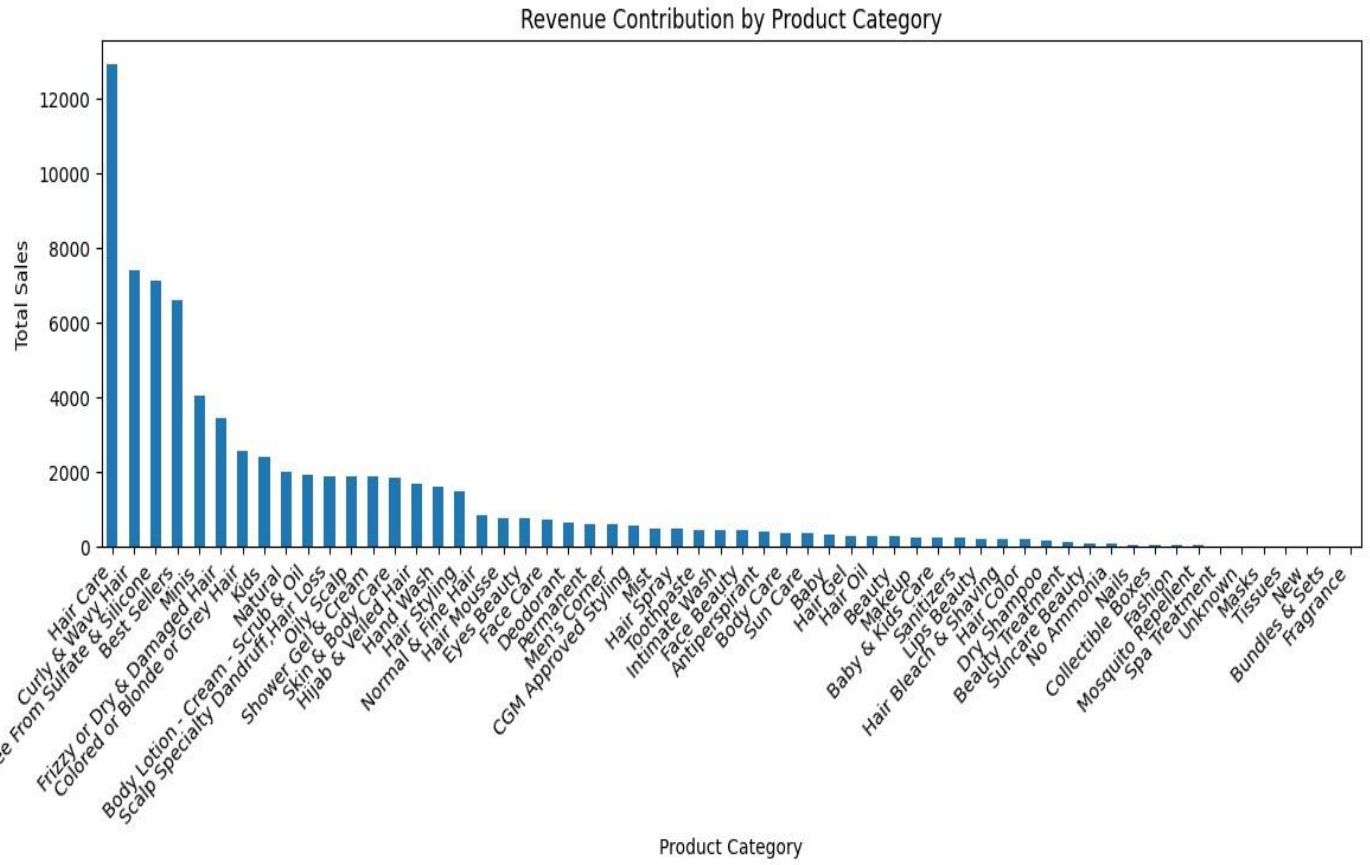

Revenue Contribution by Product Category
Total Sales
Product Category
0
2000
4000
6000
8000
10000
12000
Hair Care
Curly & Wavy Hair
Free From Sulfate & Silicone
Best Sellers
Minis
Colored or Blonde or Grey Hair
Frizzy or Dry & Damaged Hair
Kids
Natural
Body Lotion - Cream - Scrub & Oil
Scalp Specialty Dandruff, Hair Loss
Oily Scalp
Shower Gel & Cream
Skin & Body Care
Hijab & Veiled Hair
Hand Wash
Hair Styling
Normal & Fine Hair
Hair Mousse
Eyes Beauty
Face Care
Deodorant
Permanent
Men's Corner
CGM Approved Styling
Mist
Hair Spray
Toothpaste
Intimate Wash
Face Beauty
Antiperspirant
Body Care
Sun Care
Baby
Hair Gel
Hair Oil
Beauty
Makeup
Kids Care
Sanitizers
Baby & Kids Care
Lips Beauty
Hair Bleach & Shaving
Hair Color
Dry Shampoo
Beauty Treatment
Suncare Beauty
No Ammonia
Nails
Collectible Boxes
Fashion
Mosquito Repellent
Spa Treatment
Unknown
Masks
Tissues
New
Bundles & Sets
Fragrance

Figure E.4

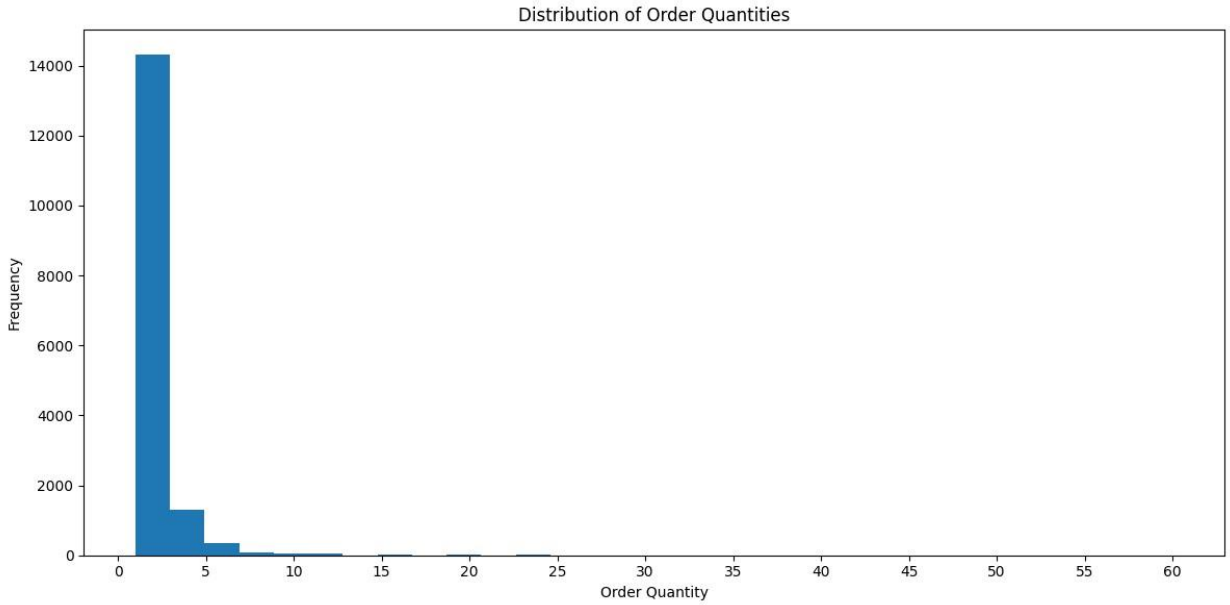


Figure E.5

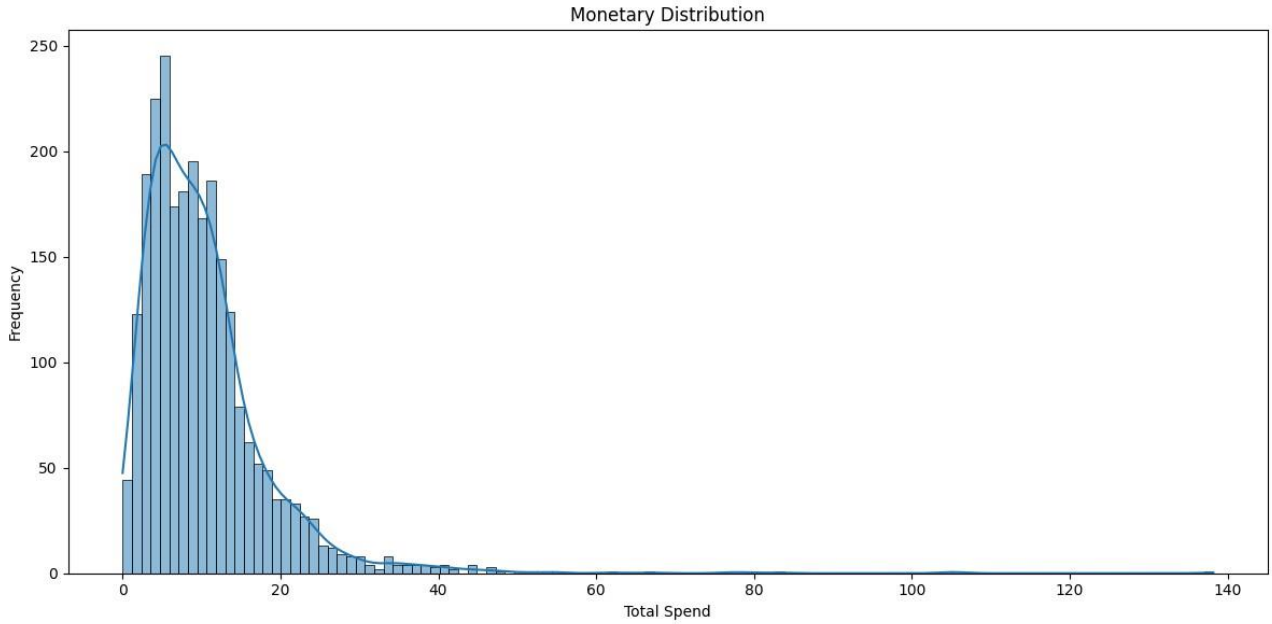


Figure E.6

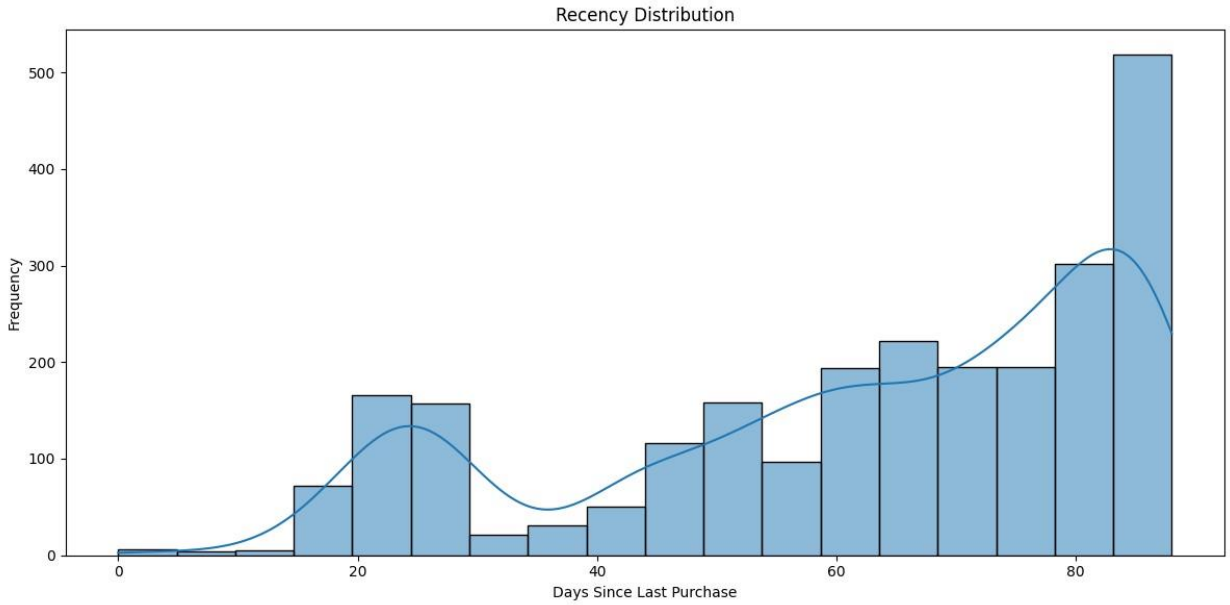


Figure E.7

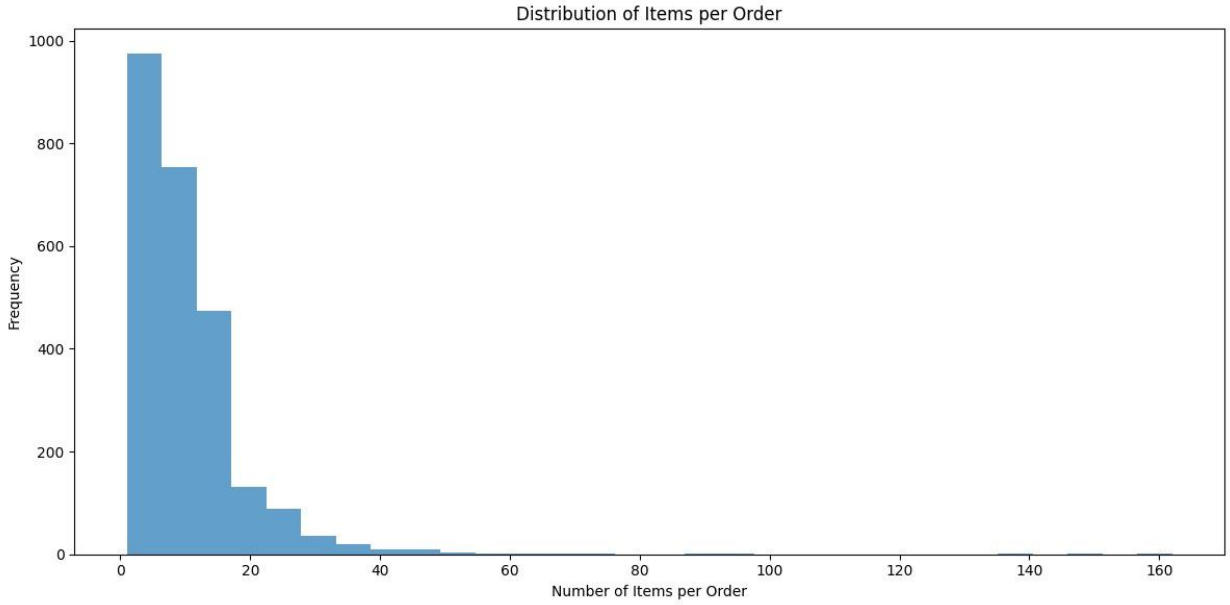


Figure E.9

Figure E.10

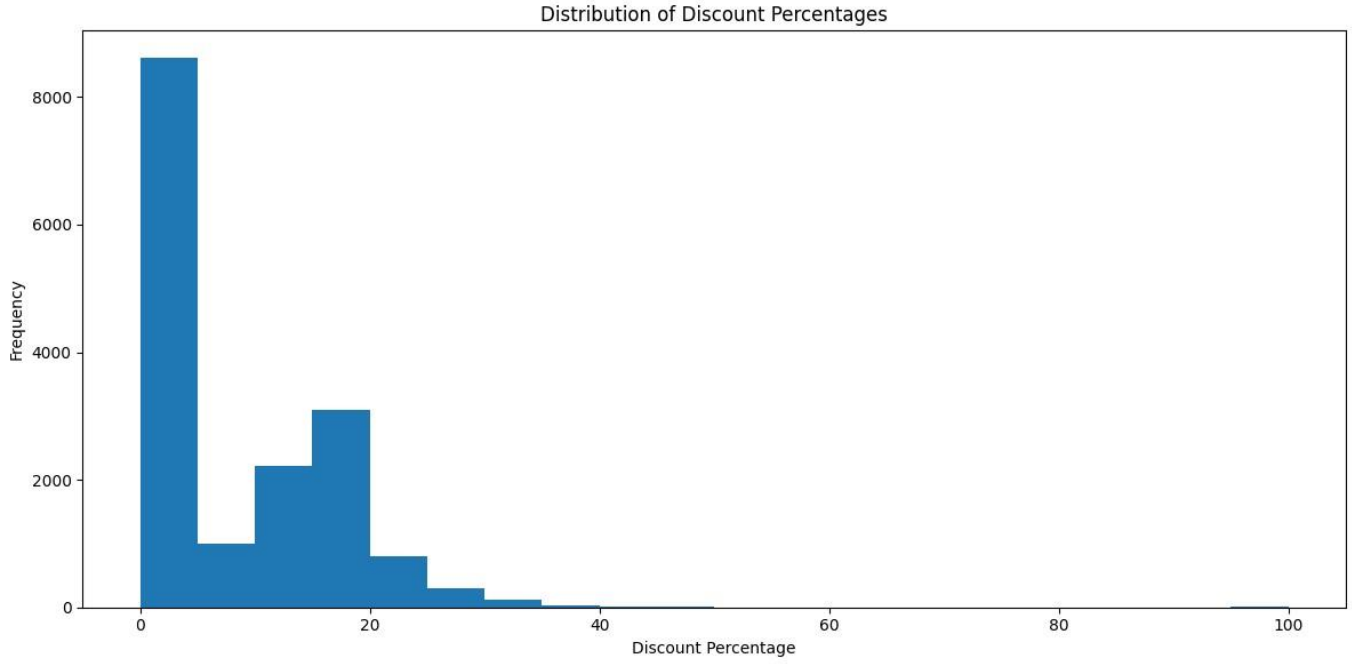


Figure E.11

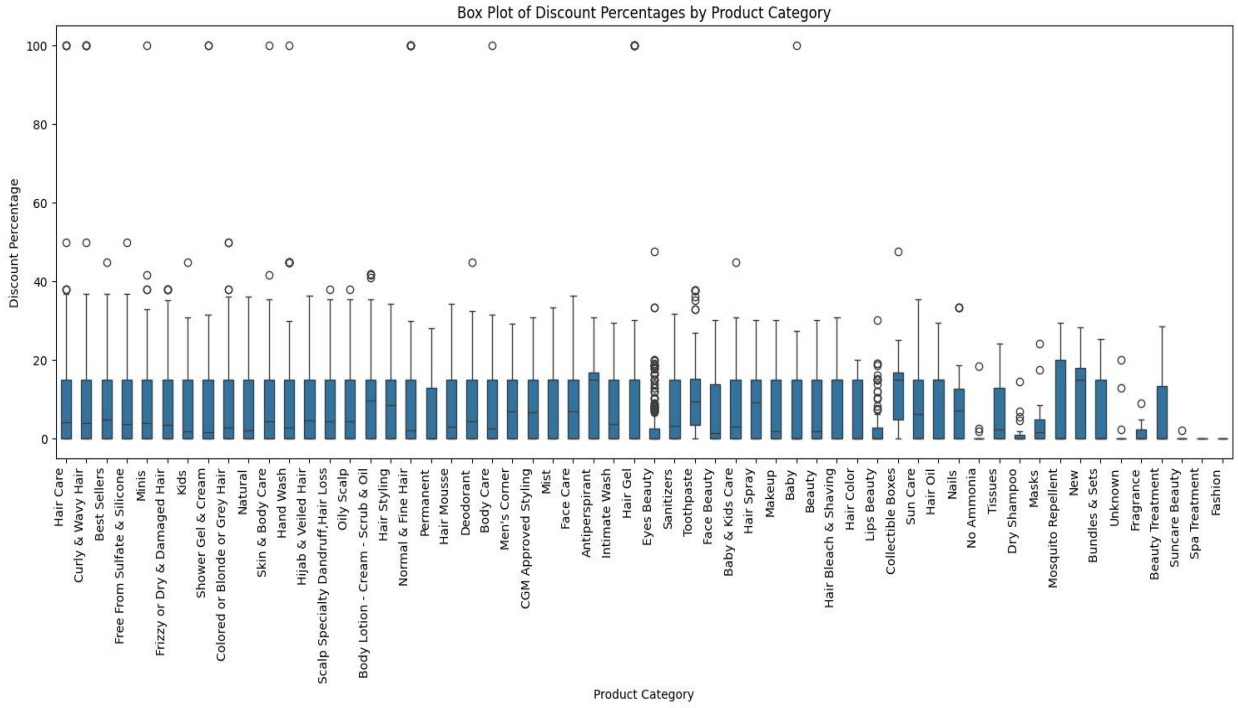


Figure E.12

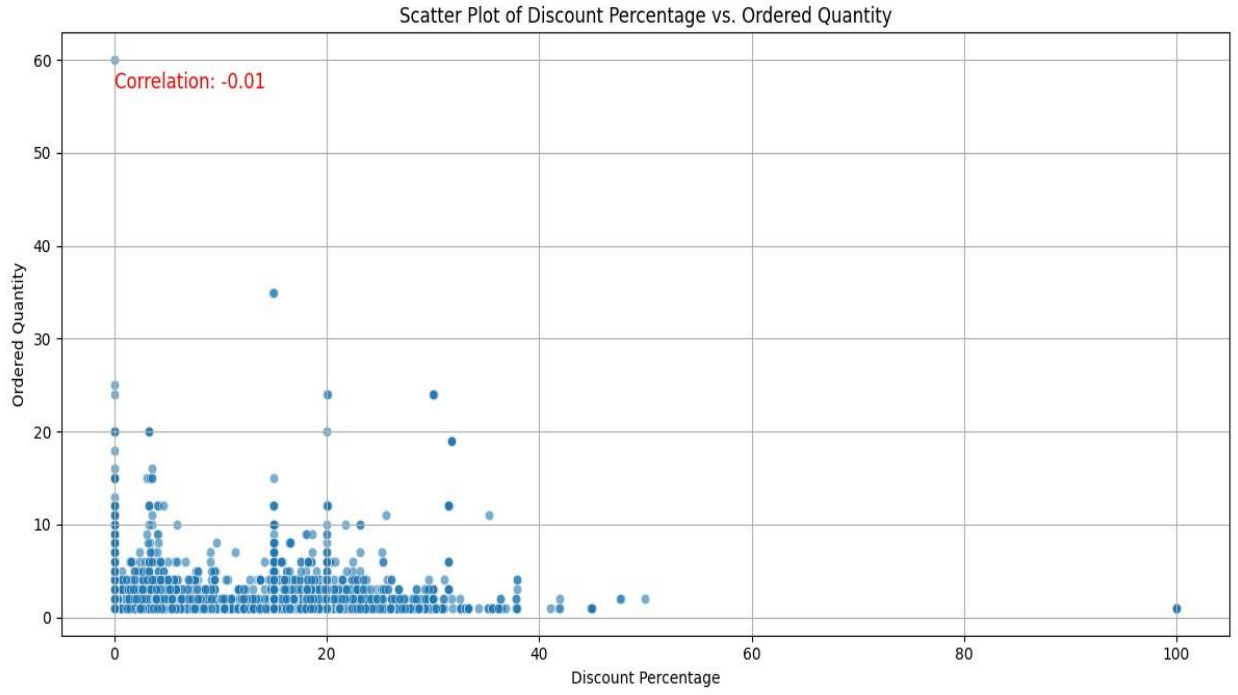


Figure E.13

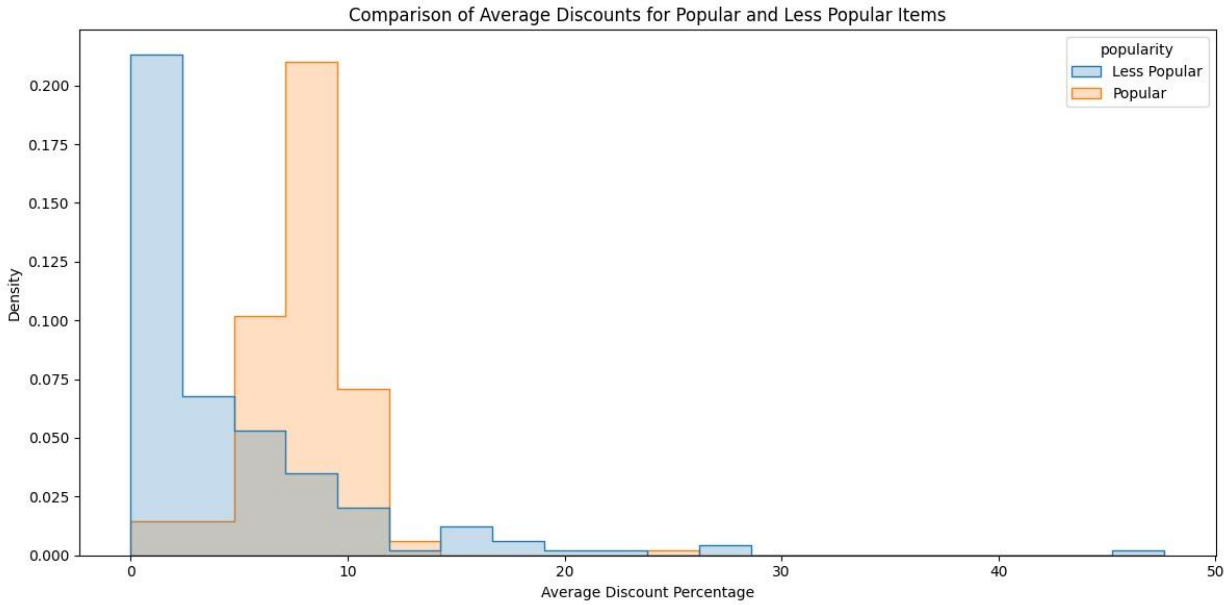


Figure F.14

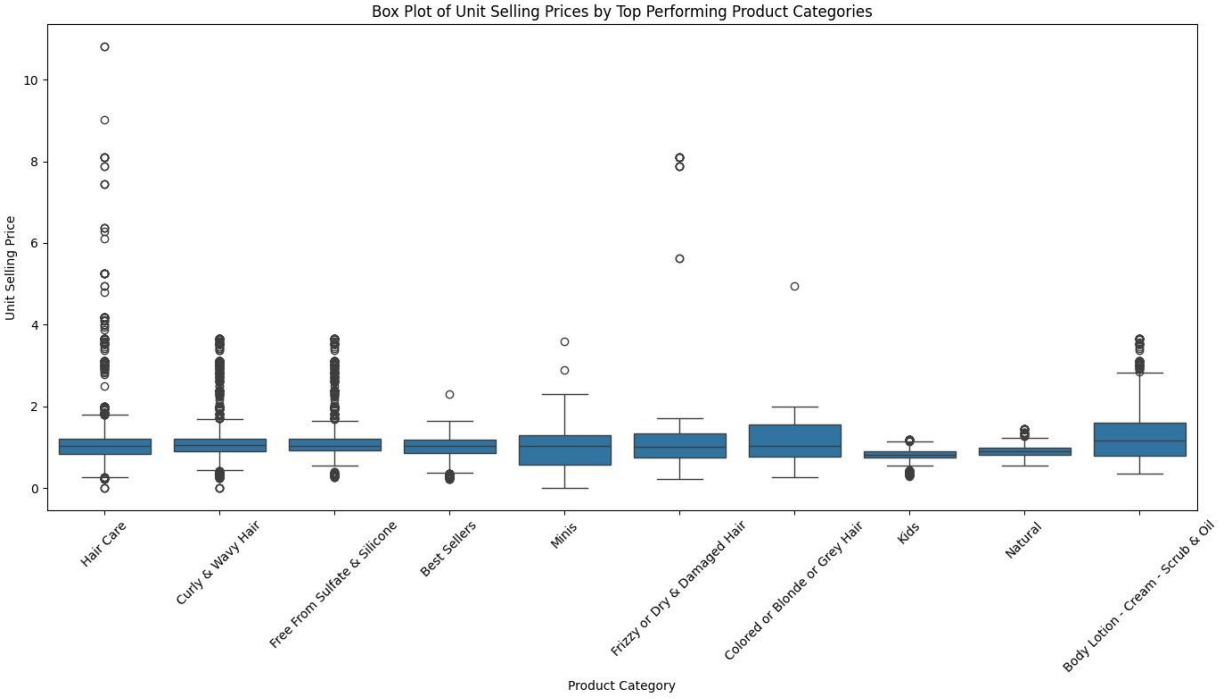


Figure F.15

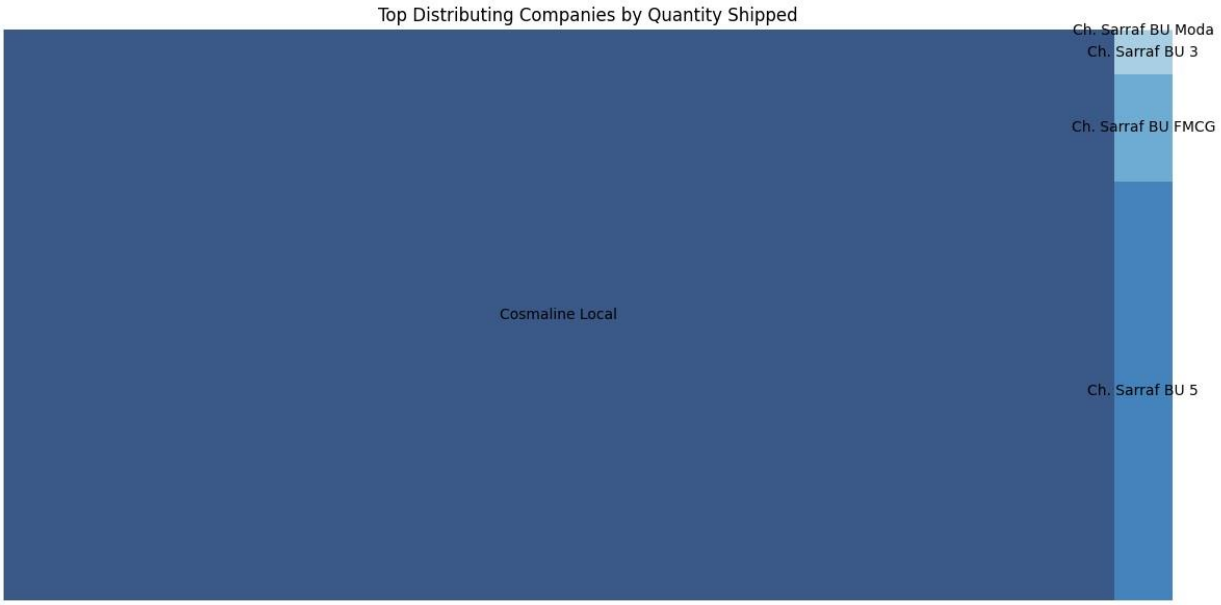


Figure E.16

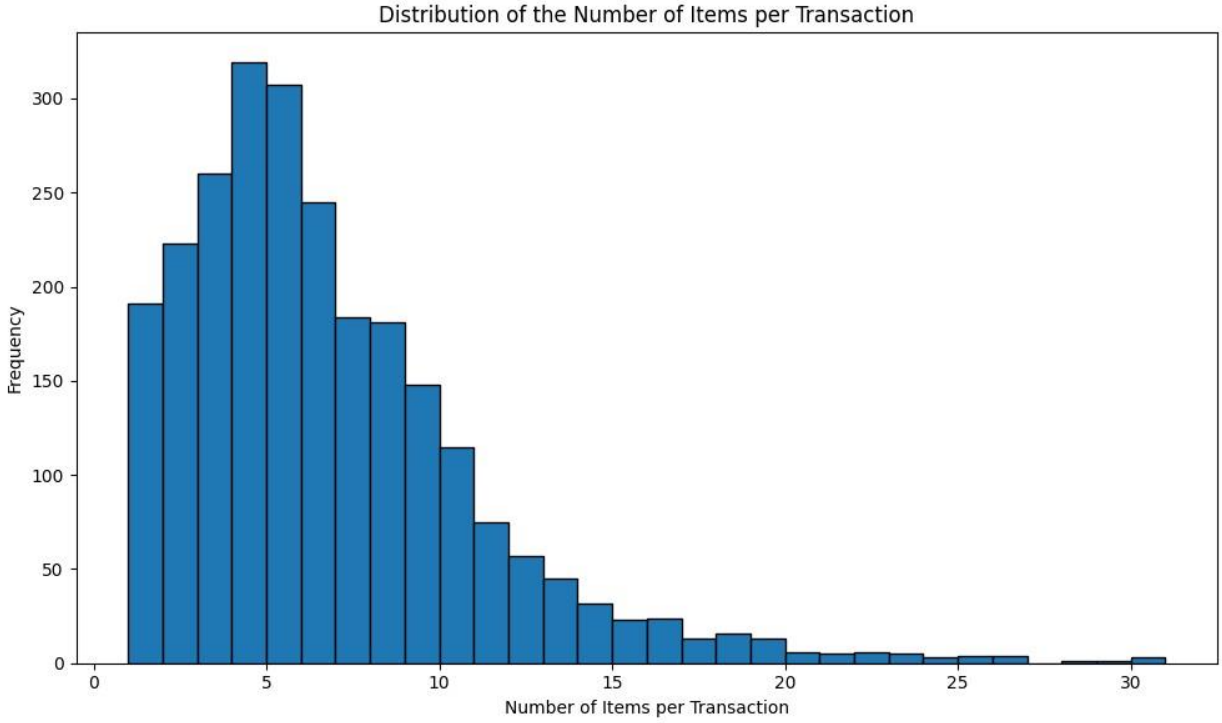


Figure E.17

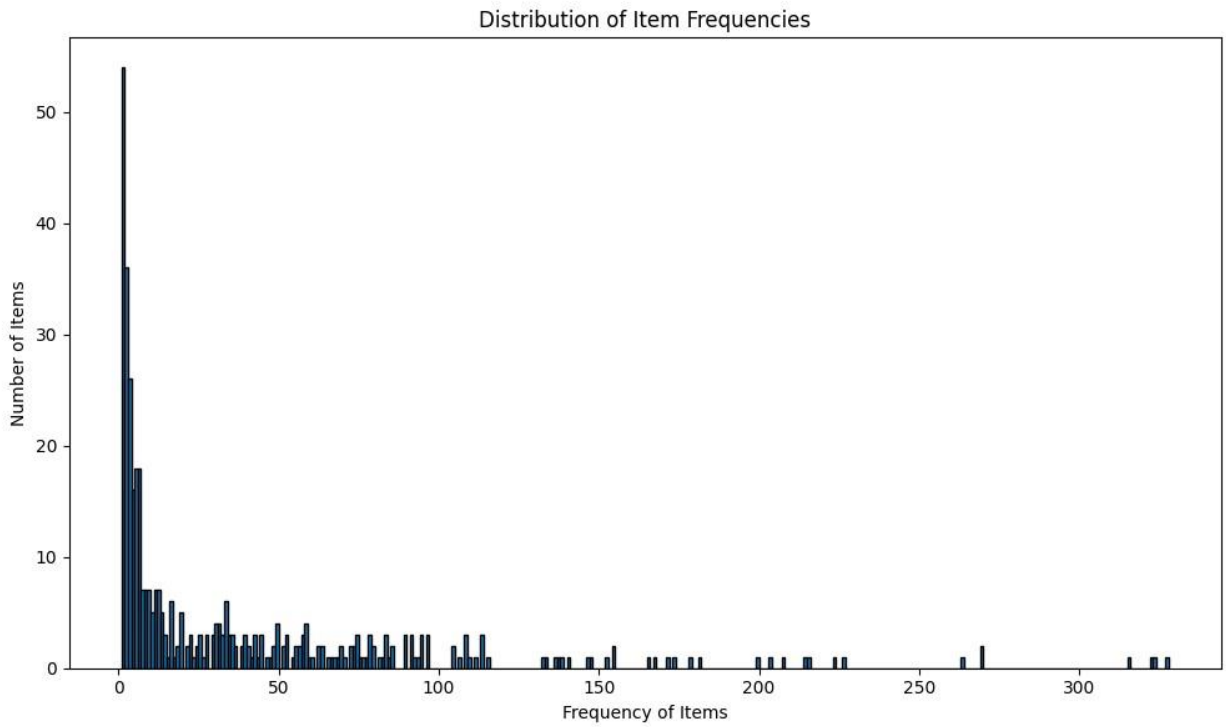


Figure E.18

Statistic	Value
count	2509
mean	6.313671
std	4.319854
min	1
25%	3
50%	5
75%	8
max	31

Figure E.19 – Transaction Quantity Stat

Statistic	Value
count	407
mean	38.92138
std	57.36826
min	1
25%	3
50%	12
75%	55.5
max	328

Figure E.20 – Cross Transactions Item Stat

Product Name	**Count**
koleston kit light ash blonde 8	80
essence pumpkins pretty matte nail polish	80
koleston maxi single light brown 305 0	80
essence super fine liner pen	80
essence disney mickey friends bouncy	80
invigo leave in wonder balm	80
fine napkin 40 1ply black	80

collistar reshaping draining wraps	80
eimi sugar lift	80
cosmaline baby discovery kit	80
soft color kit 71 ash	80
essence pumpkins pretty creamy shadow liner	80
collistar reshaping draining solution refill	80
essence after shape brow roller cooling and	80
essence pumpkins pretty eyeshadow palette	80
koleston kit medium ash blonde 7	80
collistar eye hydro gel ice	80
catrice 48h power stay brow gel	80
koleston kit exotic red 55	80
essence pumpkins pretty smoothing lip patch	80
koleston naturals red devotion 5	80
essence matt mousse mk up	80
koleston maxi single blue black 301 0	80
essence mattifying compact powder	80
essence disney lion king eyeshadow palette	80
essence soft touch mousse make up	80
koleston kit medium blonde 7	80
catrice professional brow palette	80
invigo balance clean scalp shampoo	80
koleston maxi single dark brown 303 0	80
soft color kit 77 golden	80
shiseido day spa deep back massage targeted stress massage for men and women 30	80
koleston kit hazelnut 7	80
essence jurassic world	80
soft color kit 28 blue	80
cosmal cure professional hair color cream medium blonde ash pearl 7 18 developer	80
kadus velvet oil 100	80
essence gel nail polish 01 gloss n	80
catrice skin face serum	80
essence french manicure sheer beauty nail polish	80
fusion intense repair shampoo	80
collistar anticellulite cryo gel boosted	80
cosmaline smiles small	80
koleston kit light brown 5	80
cosmal cure professional hair color cream dark blonde matt 6 2 developer	80
essence love that glow bronze	80
gsc bb for sports	80
koleston maxi single chocolate brown 306 7	80

collistar collagen cream	80
essence welcome to cape town eyeshadow	80
collistar hyaluronic acid aqua gel	80
soft color kit 70 natural	80
men s woven baseball cap c	80
essence brow like a boss ink brow gel	80
fusion intense repair conditioner	80
essence french manicure tip painter	80
koleston maxi single dark blonde 306 0	80
collistar protection milk spray spf50	80
cosmal cure professional hair color cream burgundy 4 6 developer	80
essence grow like a boss lash brow growth	80
catrice skin face mask spatula	80
essence gel nail polish 40 isn t she	80
koleston naturals dark ash blonde 6	80
cosmal cure professional hair color cream honey blonde 8 3 developer	80
koleston kit dark brown 3	80
essence lash princess liner	80
shiseido uv protective compact foundation spf 30 dark	80
catrice disney princess pocahontas hydrogel facemask	80
cosmaline antiperspirant roll on bright sensation	80
kadus deep moisture leave in conditioning spray	80
smk imperial lash mascara ink	80
cosmal cure professional hair color cream extra light blonde 9 0 developer	80
kadus color radiance shampoo	80
cosmaline micellar cleansing water in oil	80
kadus visible repair leave in conditioning balm	80
essence brow pomade brush	80
essence disney princess jasmine false lashes	80
cosmal cure professional hair color cream gold blonde 9 3 developer	80
koleston maxi single chestnut 305 4	80
cosmal cure professional hair color cream hazelnut 7 3 developer	80

Figure E.21 – Niche Products Tables

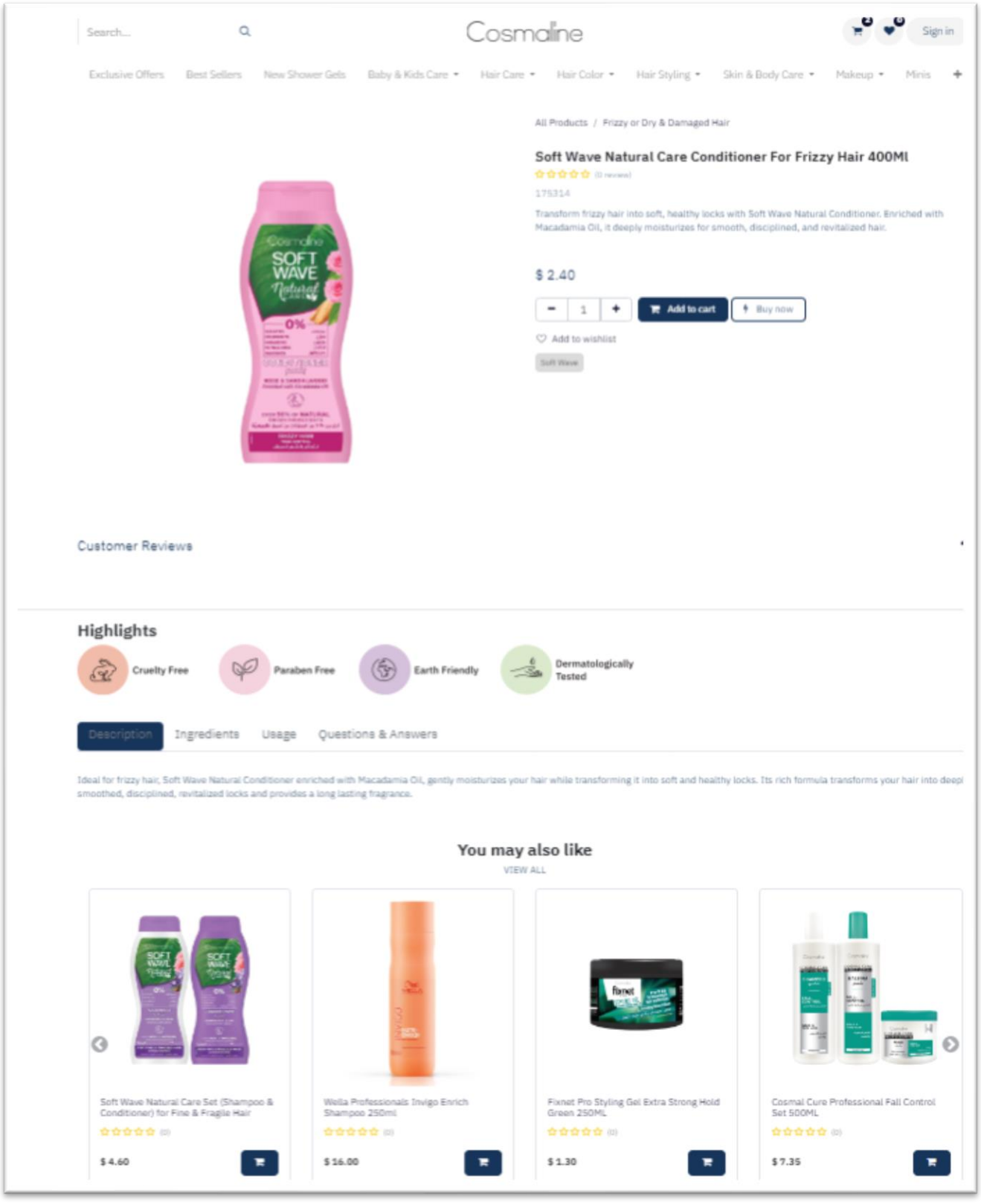


Figure Ecom.1 – Screenshot of Current Recommendations on Cosmaline.com

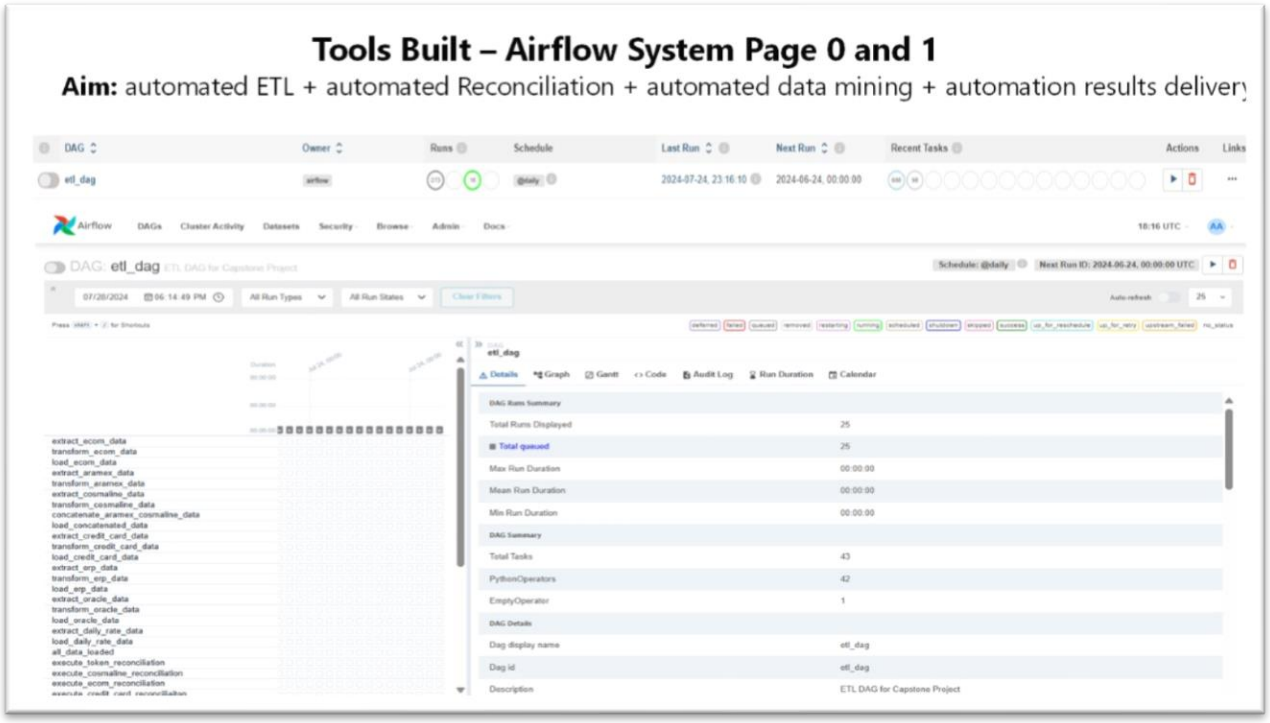


Figure A.1 – Airflow System Web Interface

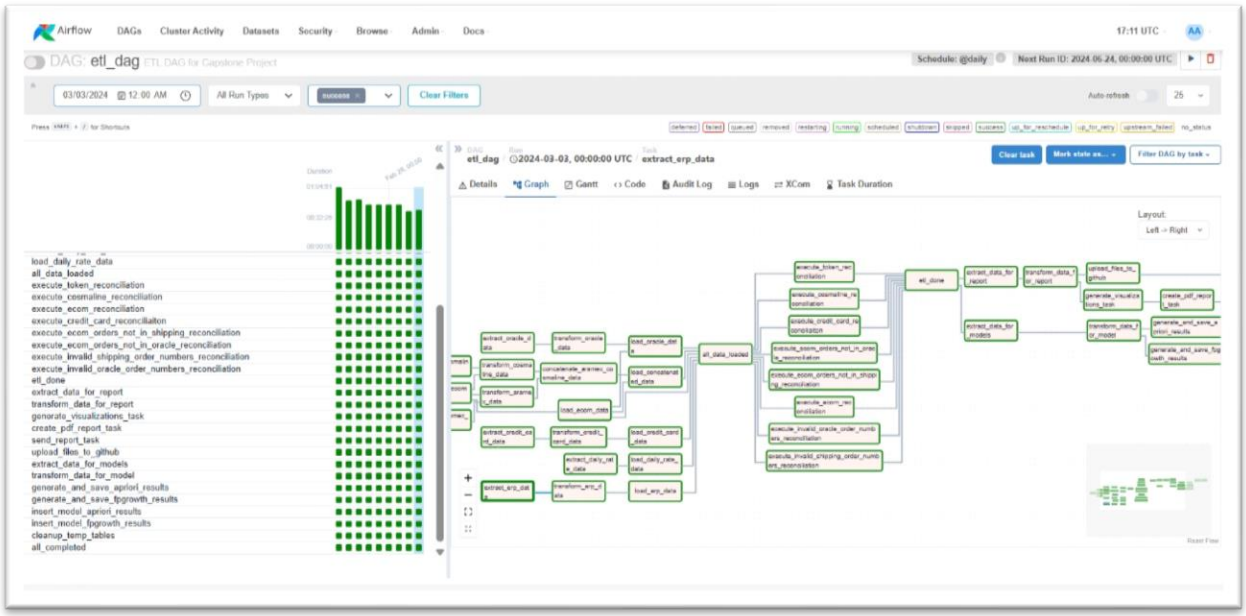


Figure A.2 – Airflow System Web Interface

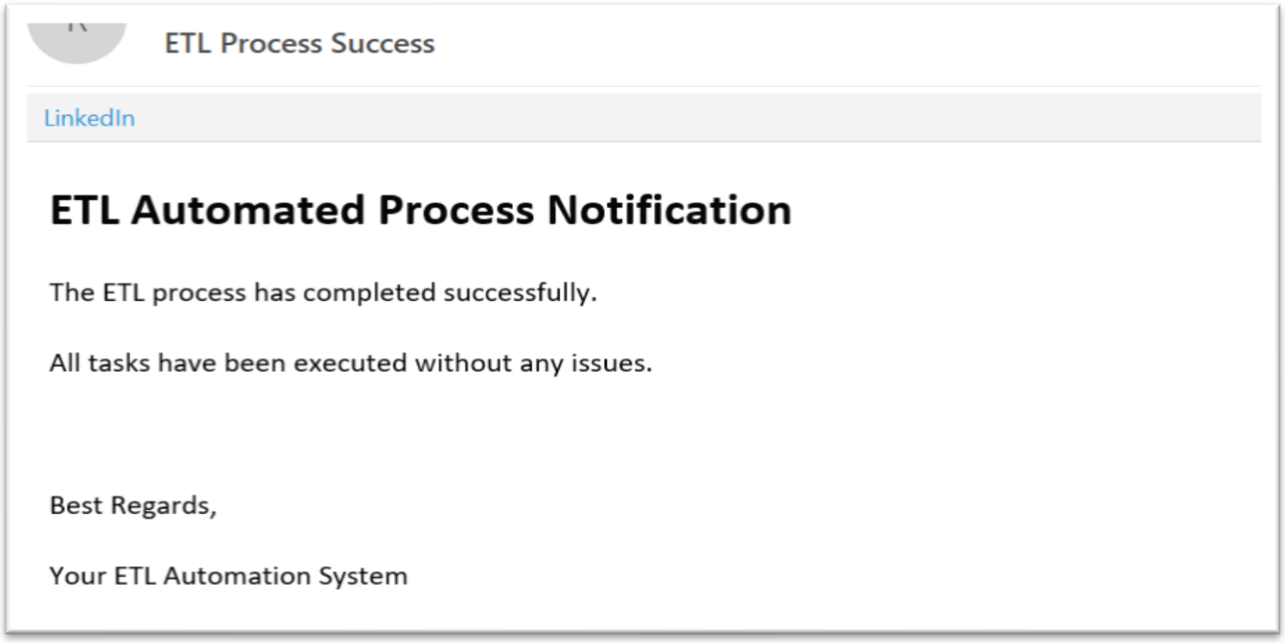


Figure A.3 – Automated Success Email

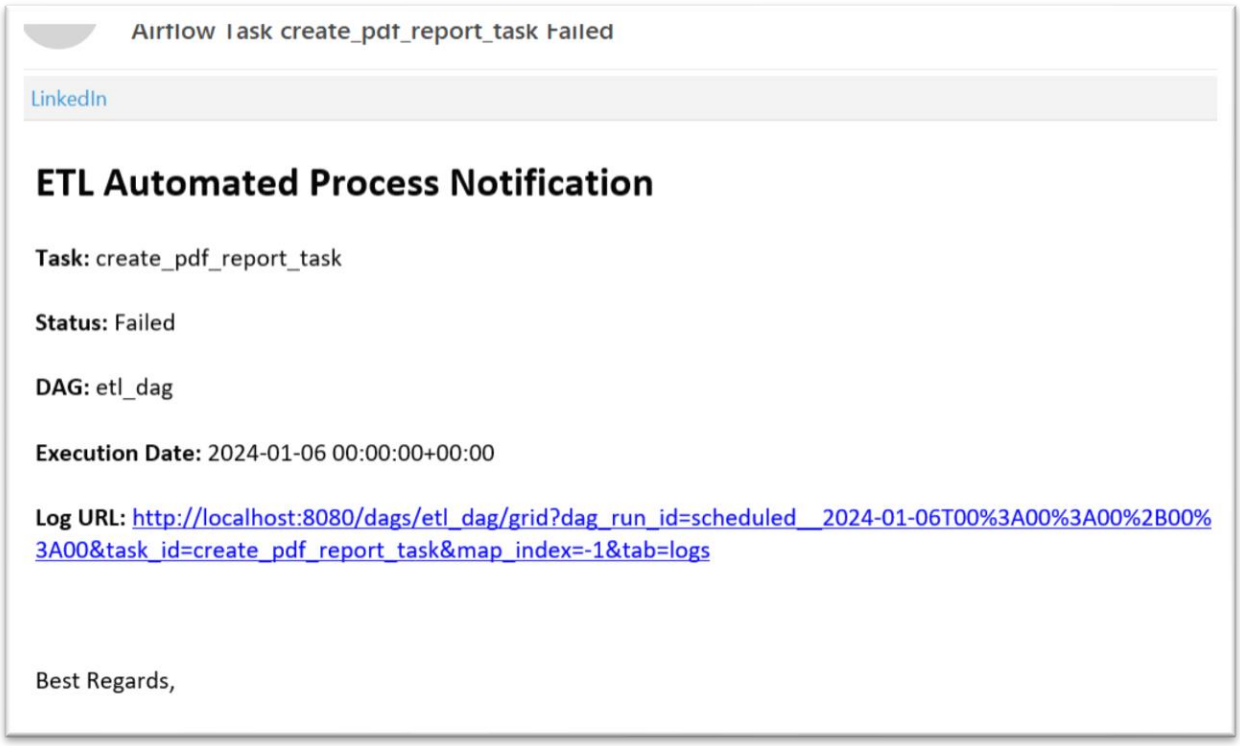


Figure A.4 – Automated Failure Email

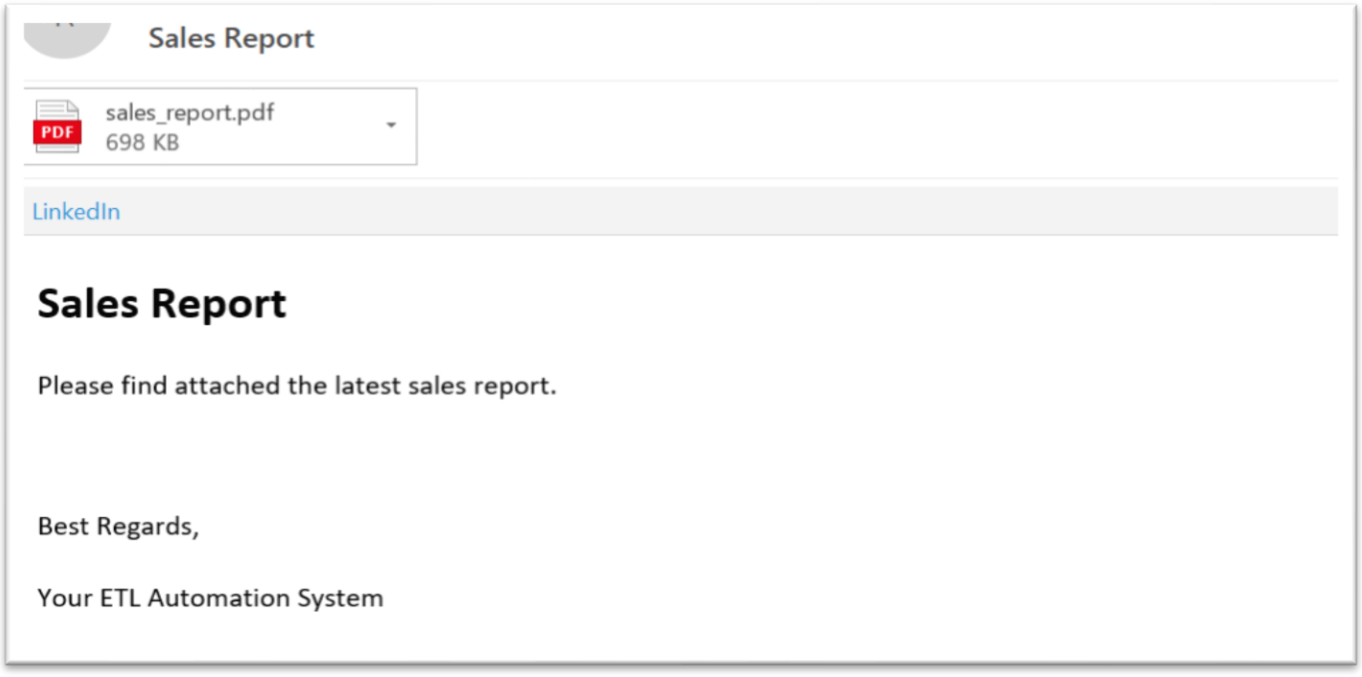


Figure A.5 – Automated Report Email

Figure D.1 -Power BI Navigation Pane

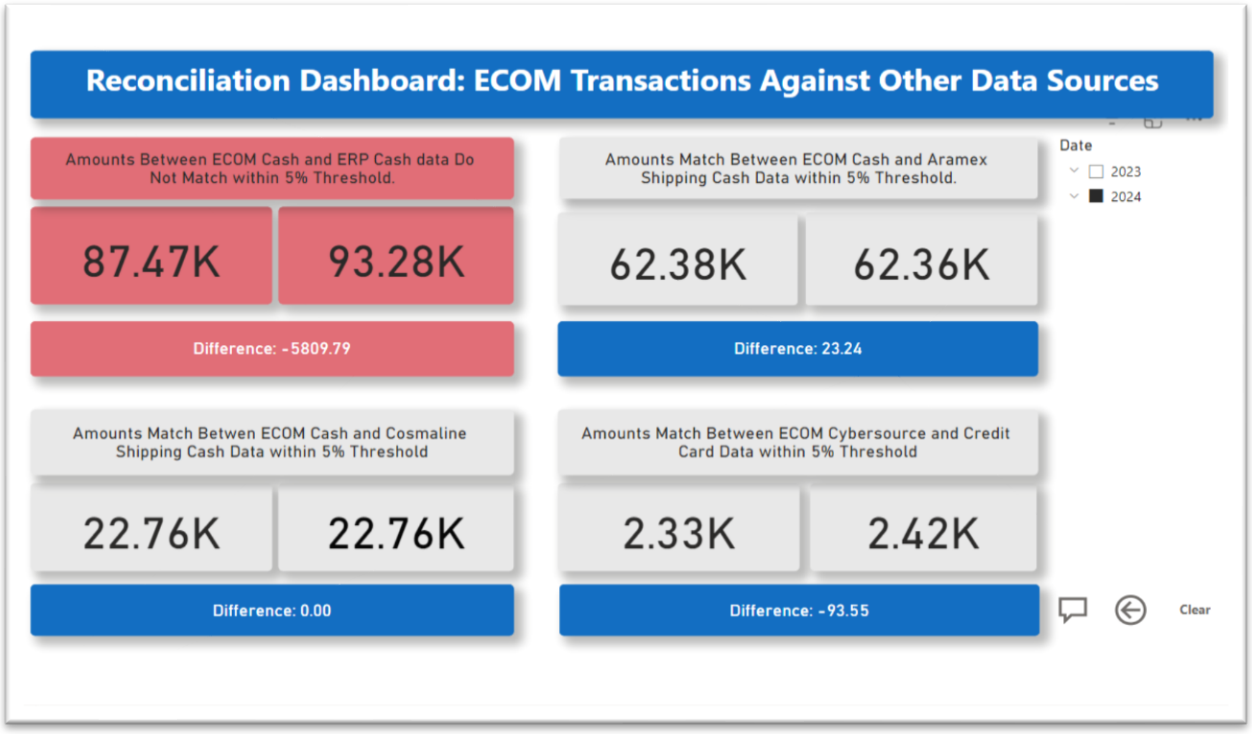


Figure D.2

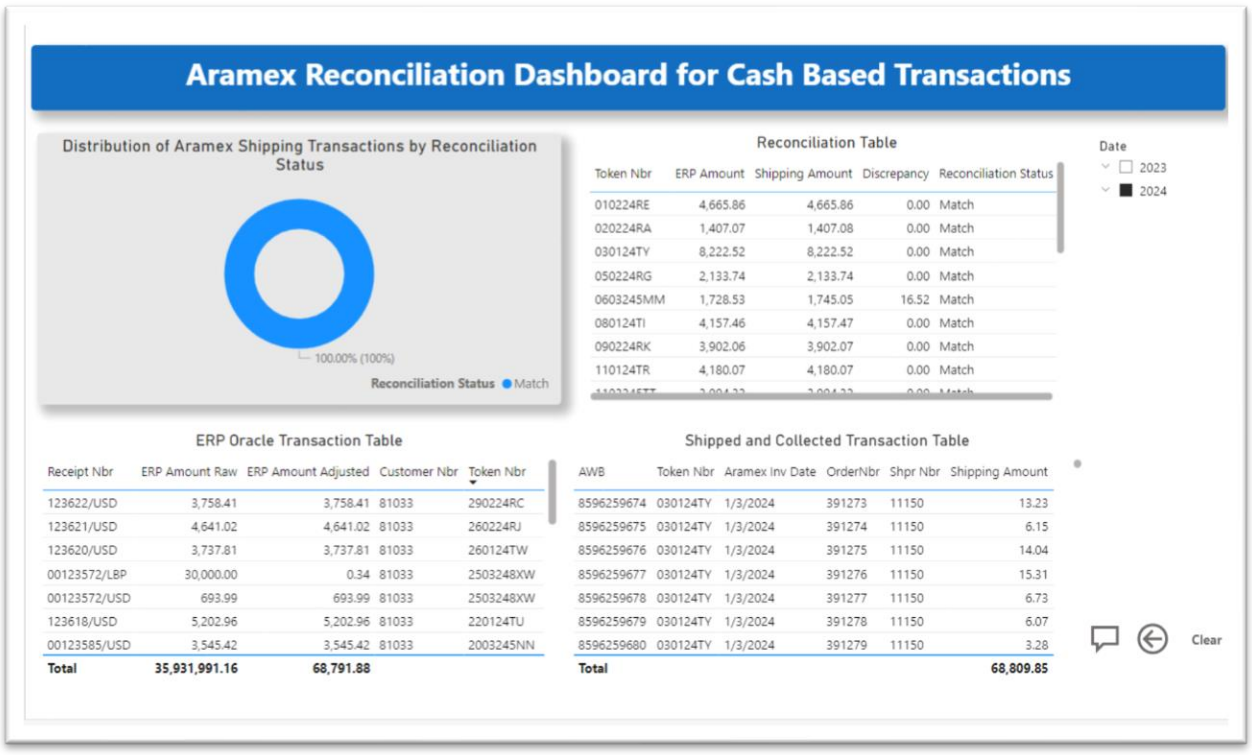


Figure D.3

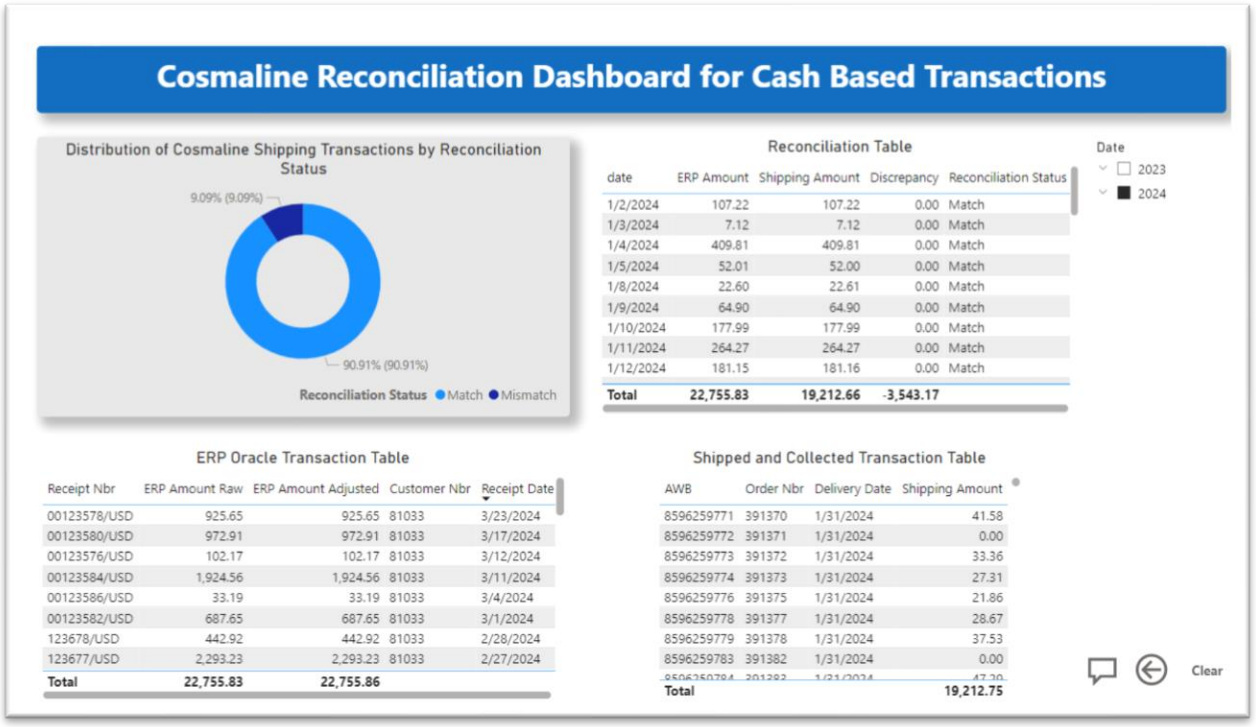


Figure D.4

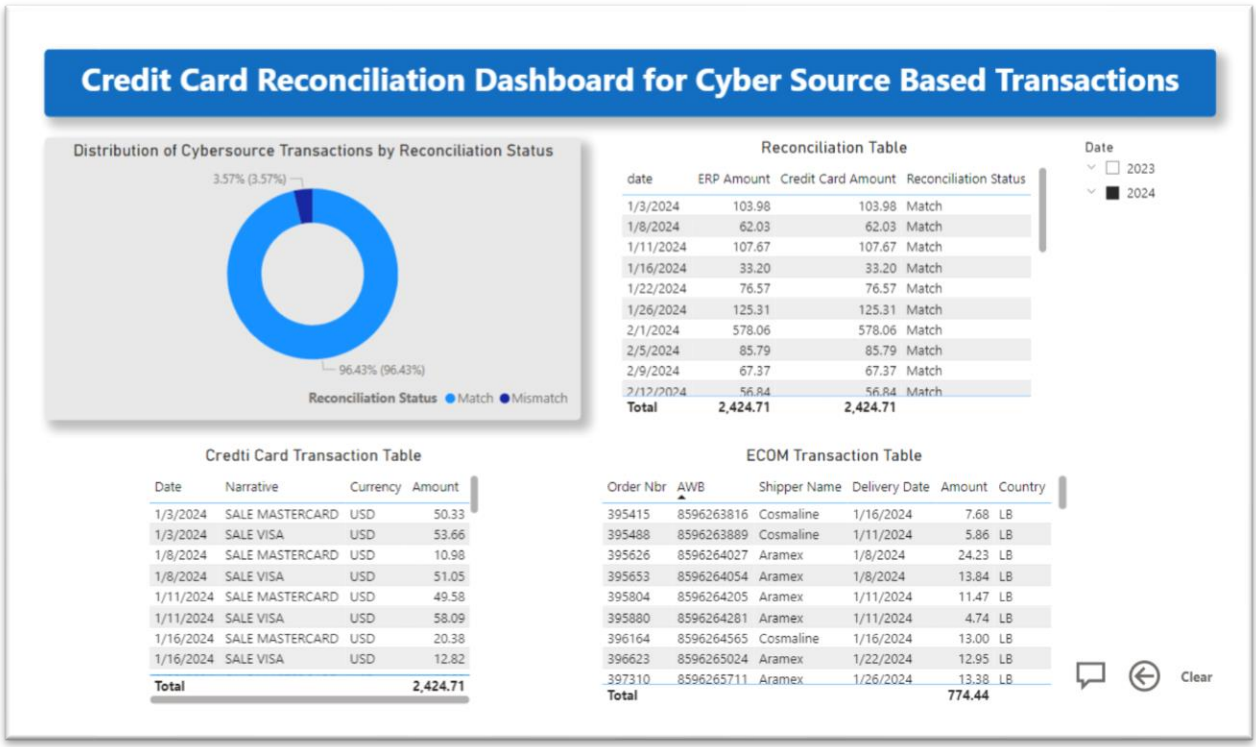


Figure D.5

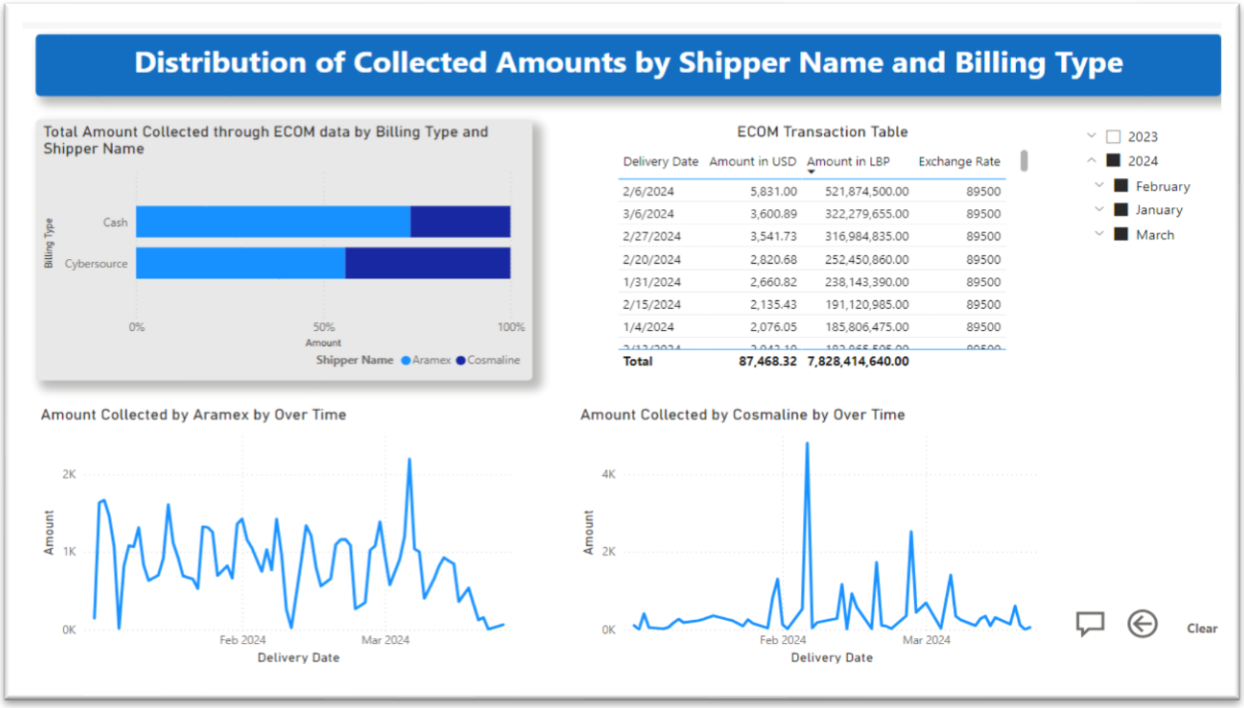

Distribution of Collected Amounts by Shipper Name and Billing Type
Total Amount Collected through ECOM data by Billing Type and Shipper Name
Billing Type
Cash
Cybersource
0%
50%
100%
Amount
Shipper Name Aramex Cosmaline
ECOM Transaction Table
Delivery Date Amount in USD Amount in LBP Exchange Rate
2/6/2024 5,831.00 521,874,500.00 89500
3/6/2024 3,600.89 322,279,655.00 89500
2/27/2024 3,541.73 316,984,835.00 89500
2/20/2024 2,820.68 252,450,860.00 89500
1/31/2024 2,660.82 238,143,390.00 89500
2/15/2024 2,135.43 191,120,985.00 89500
1/4/2024 2,076.05 185,806,475.00 89500
Total 87,468.32 7,828,414,640.00
2023
2024
February
January
March
Amount Collected by Aramex by Over Time
2K
1K
0K
Amount
Feb 2024
Mar 2024
Delivery Date
Amount Collected by Cosmaline by Over Time
4K
2K
0K
Amount
Feb 2024
Mar 2024
Delivery Date
Clear

Figure D.6

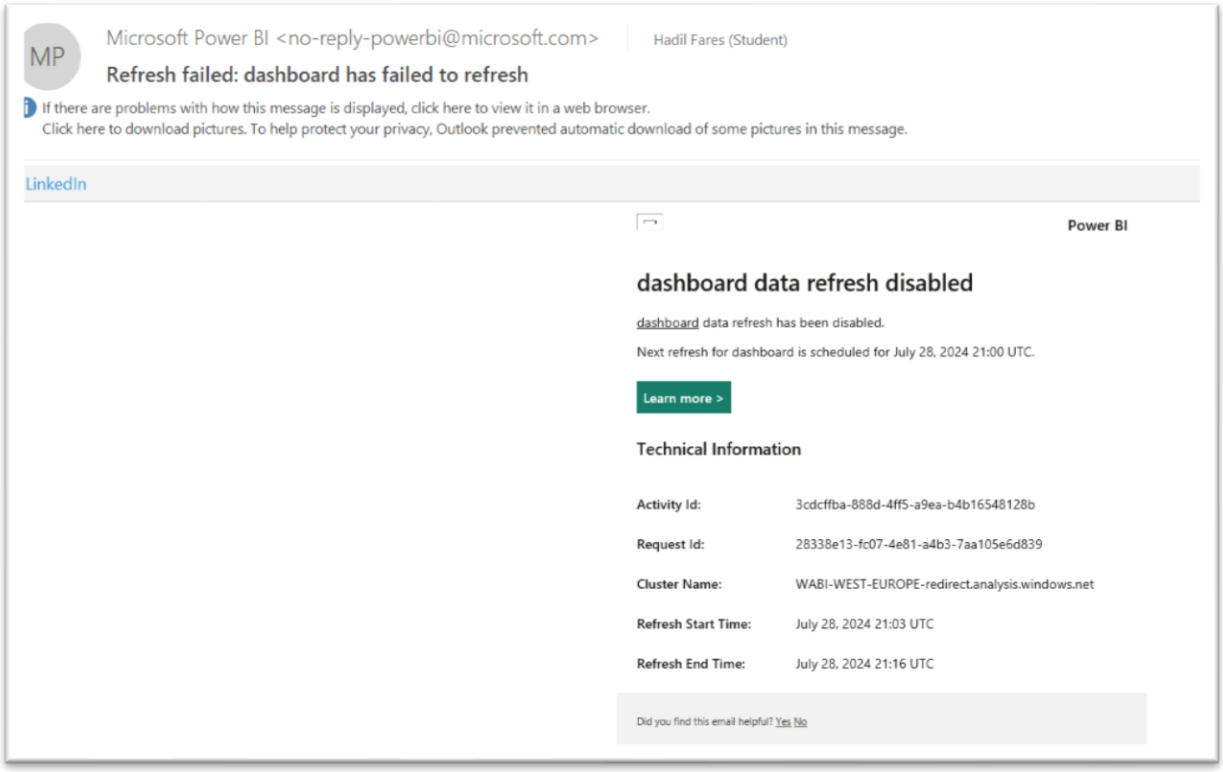


Figure D.7 – Automated Power BI Data Synchronization Failed Email

Appendix S.

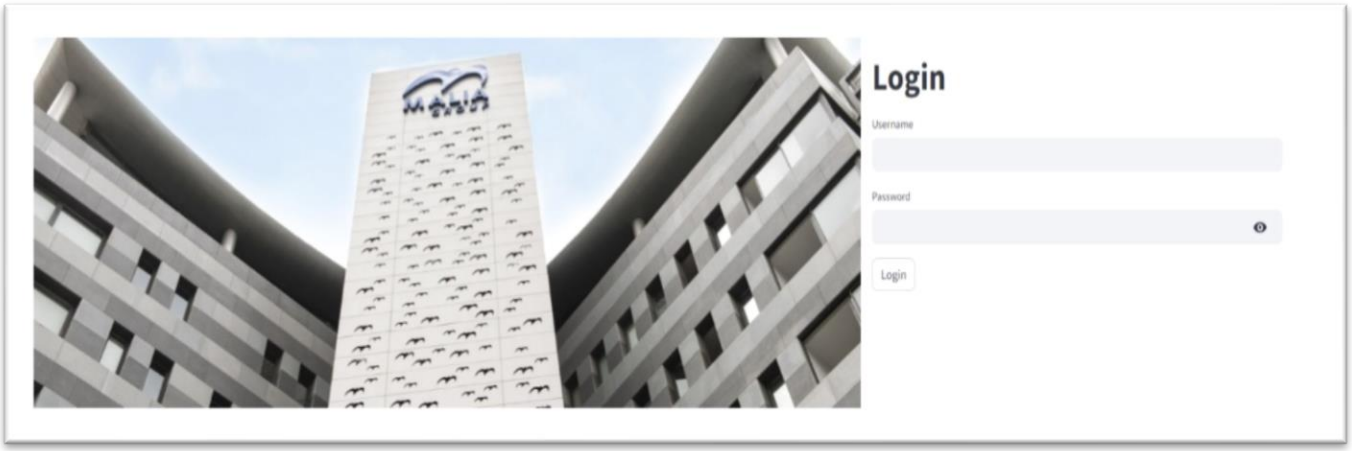


Figure S.1 – Streamlit Login Page

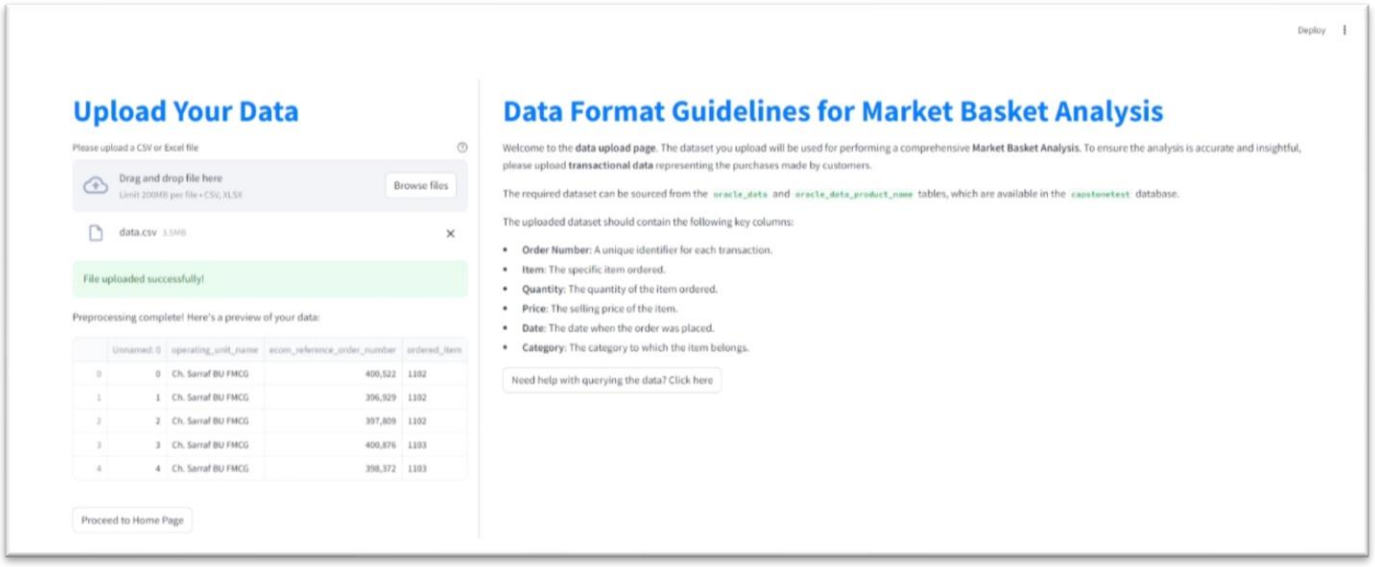


Figure S.2 – Streamlit Data Upload Page

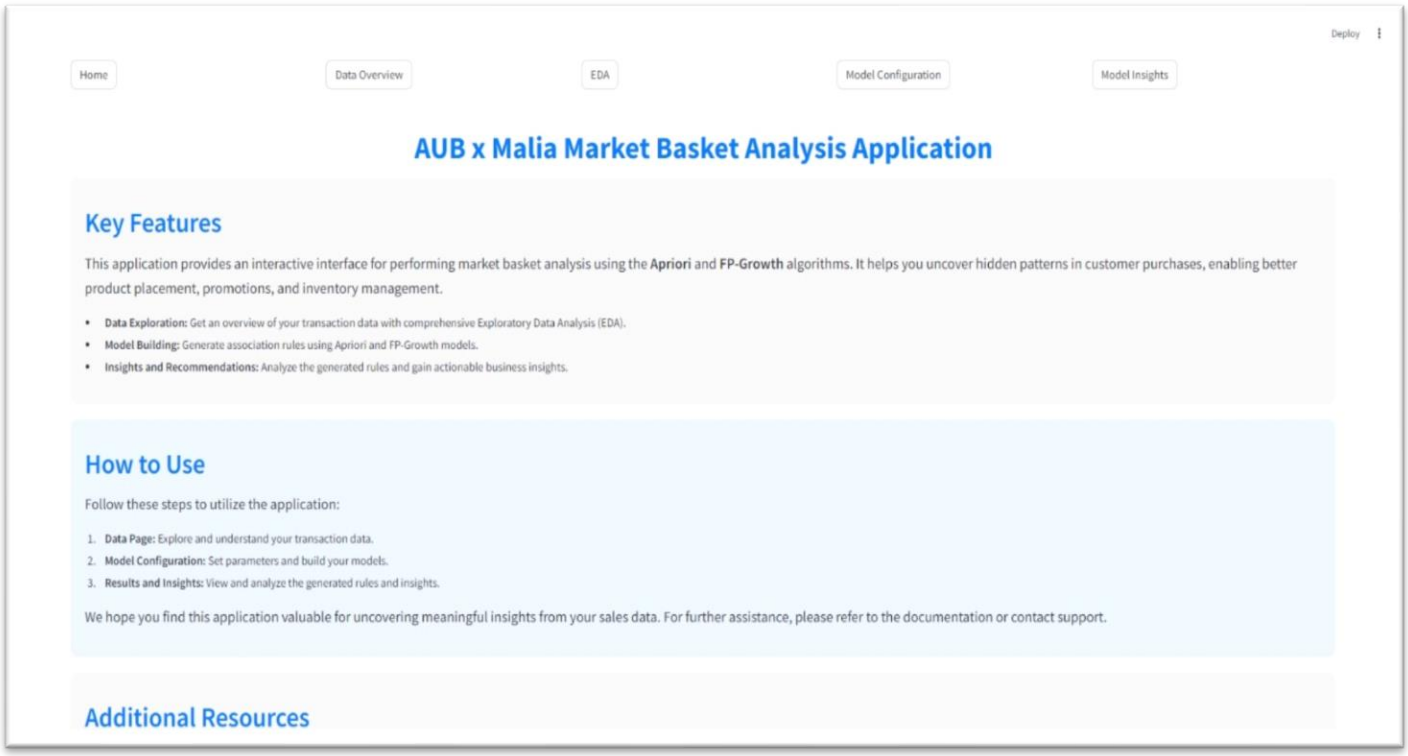


Figure S.3 – Streamlit Home Page

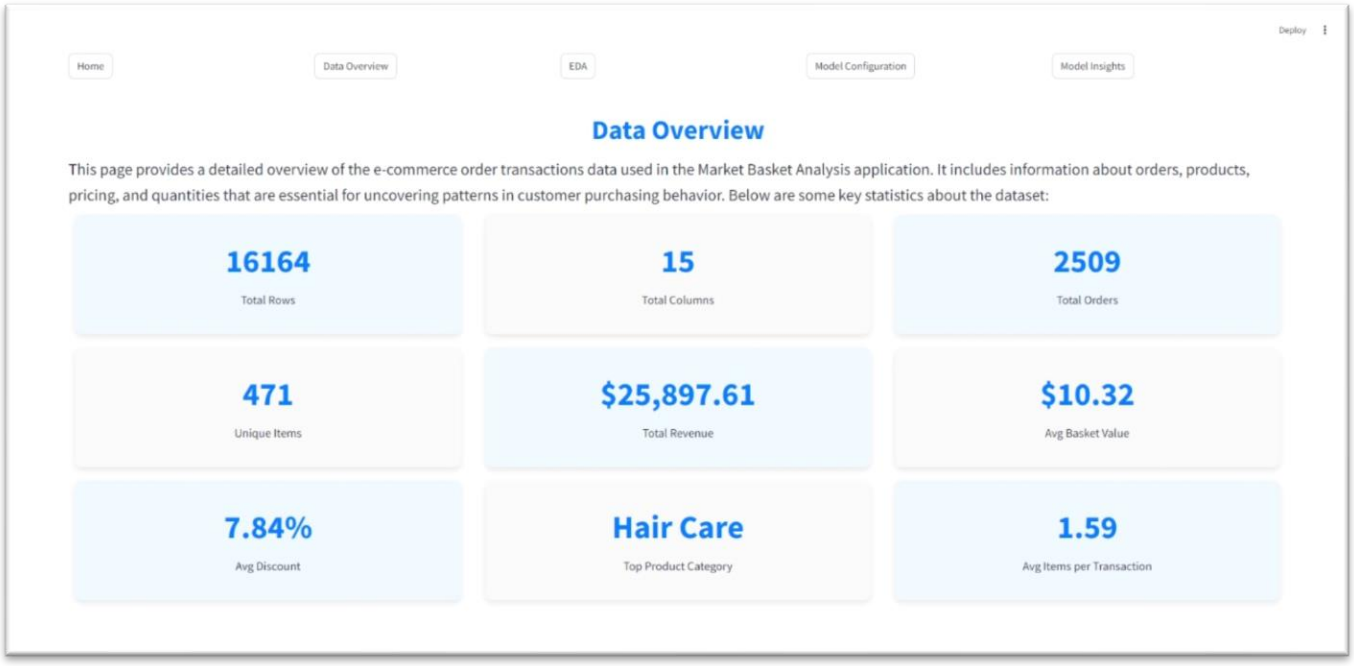


Figure S.4 – Streamlit Data Overview Page

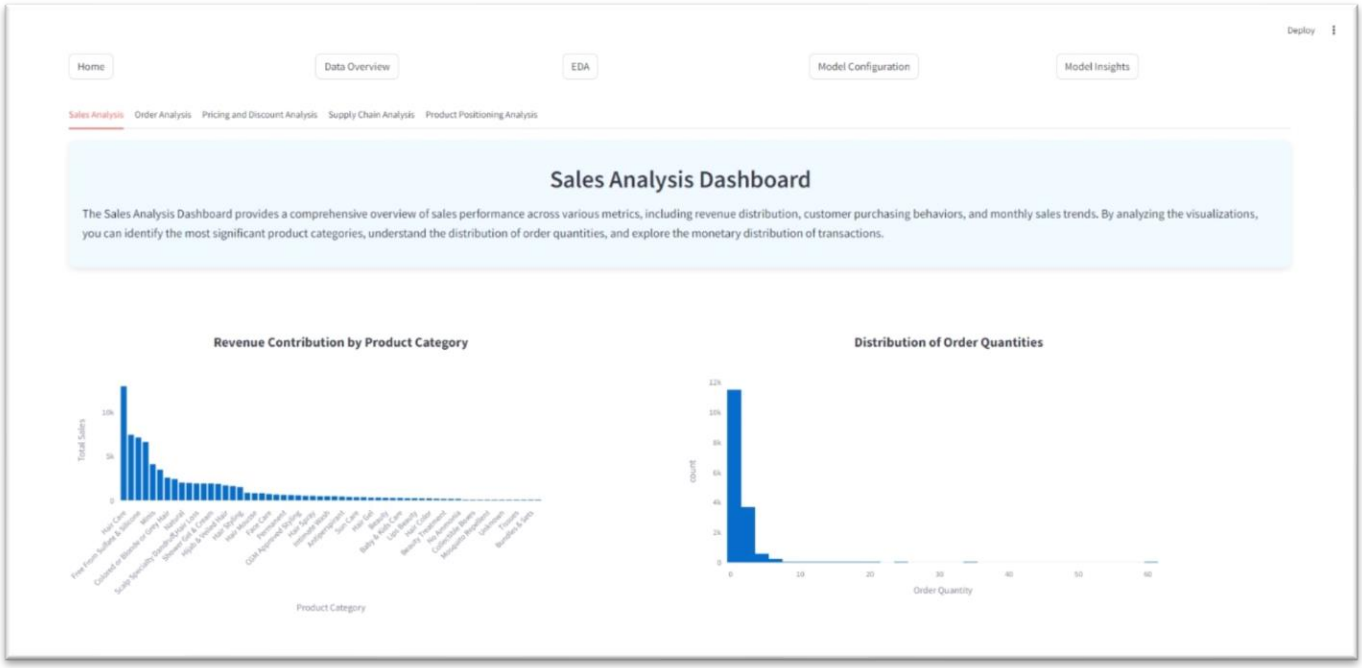


Figure S.5 – Streamlit EDA Page

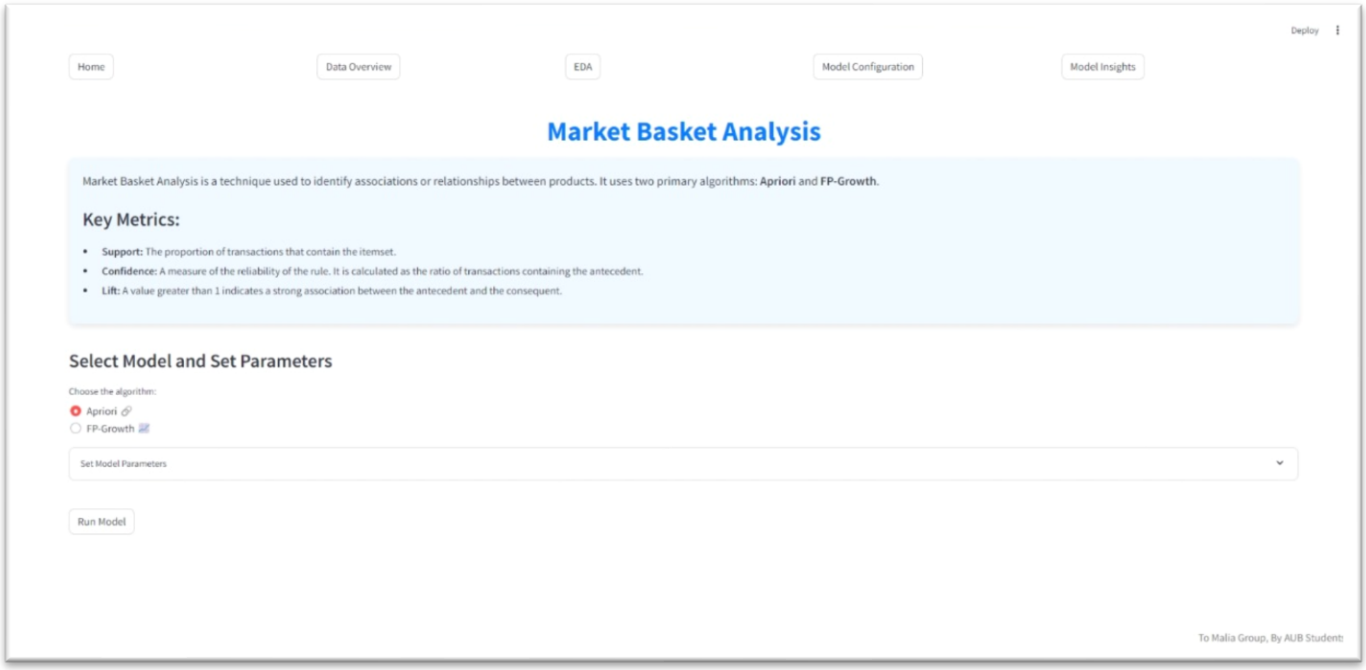


Figure S.6 – Streamlit Model Configuration Page

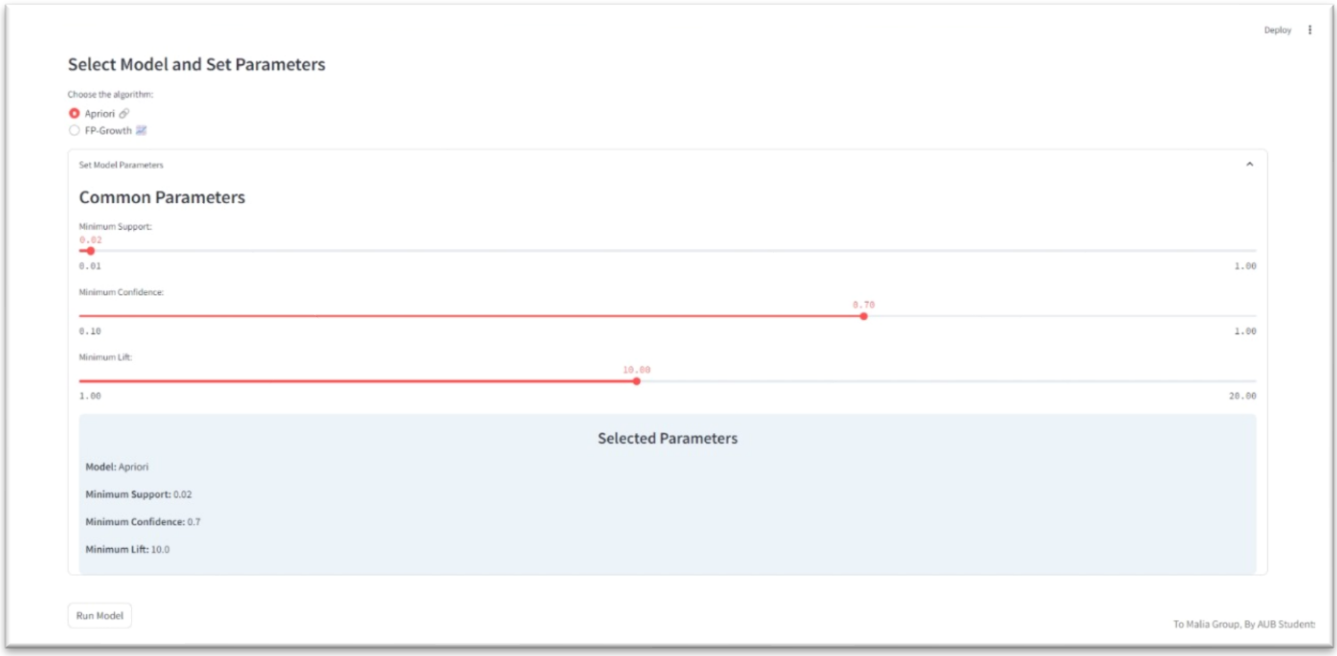


Figure S.7 – Streamlit Model Configuration Page

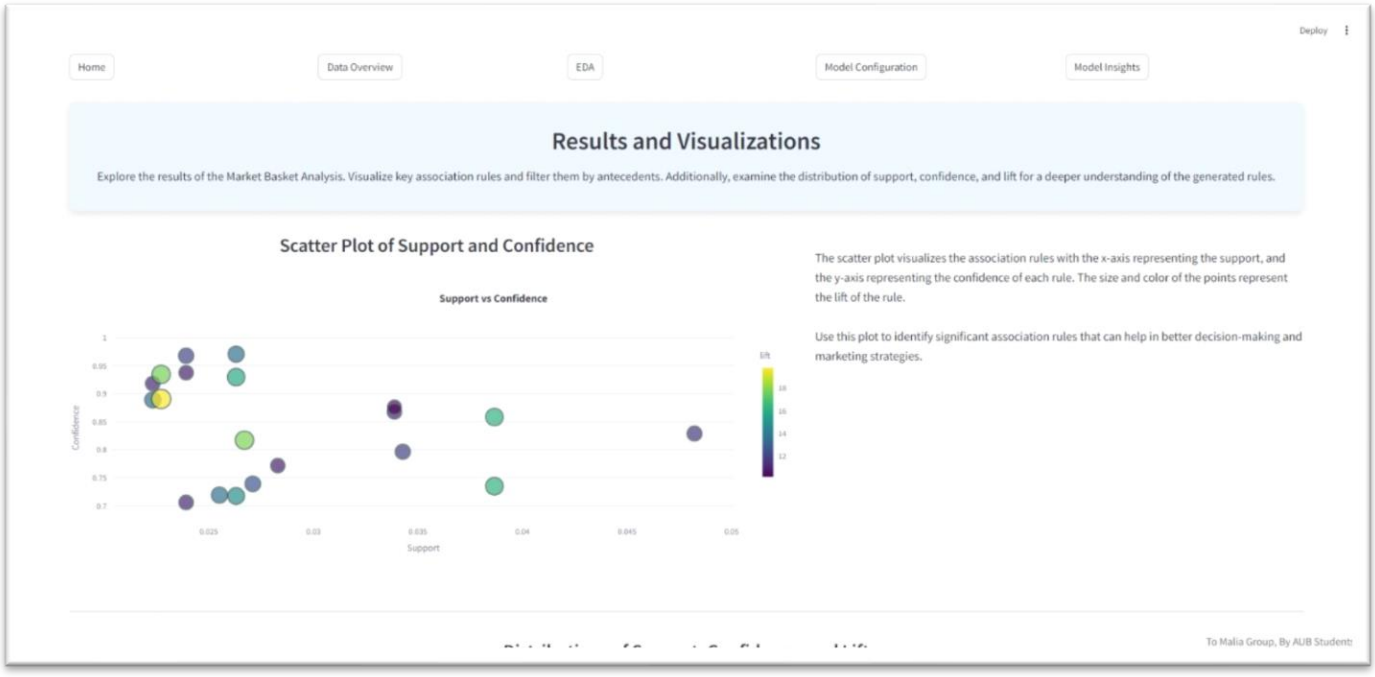


Figure S.8 - Streamlit Model Results Page

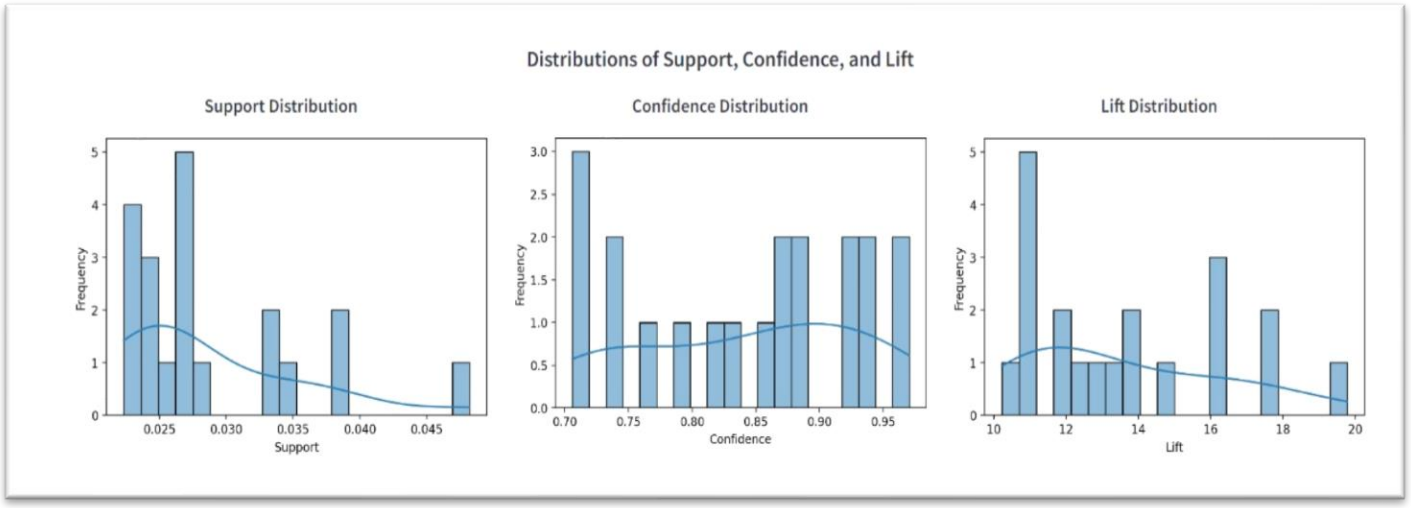


Figure S.9 - Streamlit Model Results Page

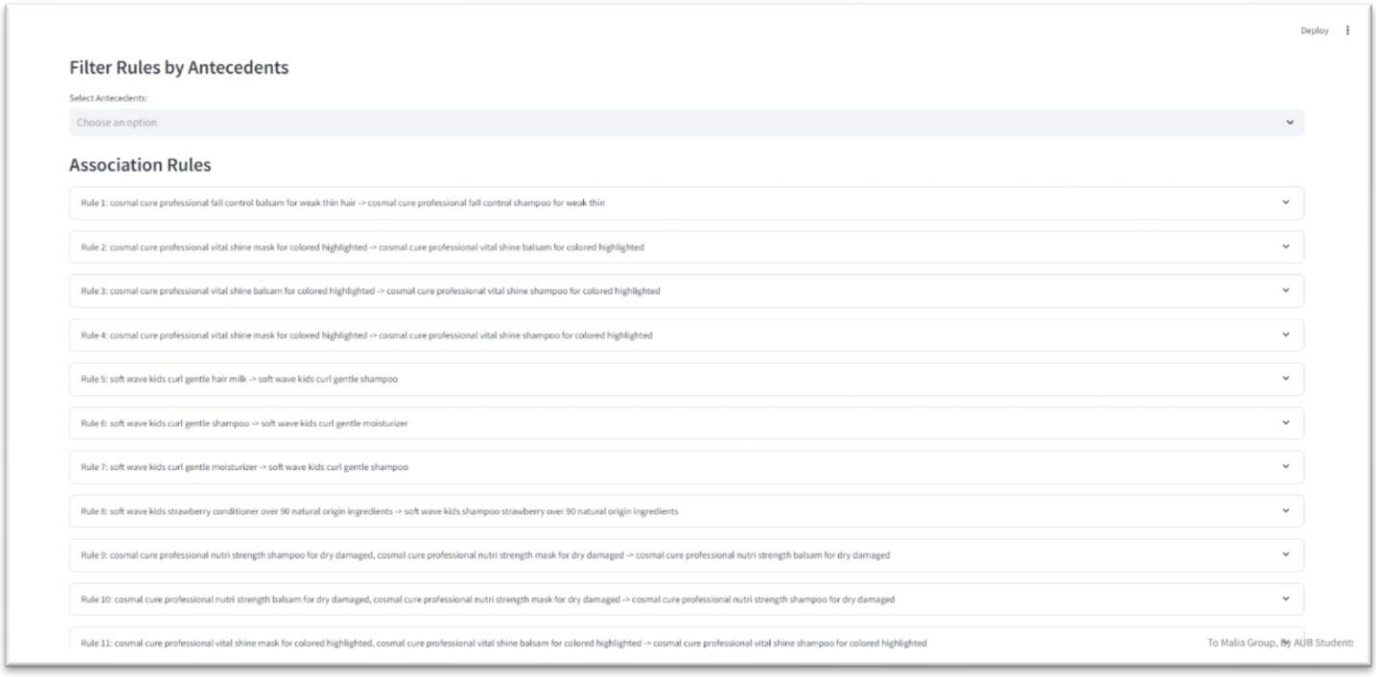


Figure S.10 - Streamlit Model Results Page

🎉 **Exclusive Cosmaline Offer Just for You, Hadil!**

C **Cosmaline** 2:23 PM
To You ···

Dear Hadil,

At Malia Group, we cherish your loyalty and want to make sure that every interaction with us, especially with Cosmaline, is extraordinary. 🌟 That's why we're thrilled to bring you something special, crafted just for your unique beauty needs!

✨ **Your Beauty Preferences, Our Priority!**

We know that you adore Cosmaline's Oh my curls products. As a token of our appreciation, we're excited to offer you an exclusive deal that aligns perfectly with your beauty routine.

🎁 **Here's What We Have in Store:**

- **20% off** on your next purchase of Oh My Curls Shampoo and Conditioner.
- **Buy 1, Get 1 Free** on Oh My Curls Hair Mask to keep your curls nourished and bouncy.
- **Free Travel-Size Product** with any purchase from the Oh My Curls line, perfect for on-the-go touch-ups.

Why? Because You Deserve the Best! 💖

Your continued trust and support inspire us every day. We believe that your experience with Cosmaline should be as unique and radiant as you are.

Ready to Glow?

Simply click the button below to unlock your exclusive offers and enjoy a personalized beauty experience like never before!

CLICK TO CLAIM YOUR OFFERS

We can't wait to pamper you with more of what you love. Thank you for being an amazing Cosmaline customer!

With love,

The Cosmaline Team

Powered by Malia Group. 🌍

Figure Demo.1