NOVA IMS Information Management School

Business Intelligence Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Retail4all

Group 07

Diana Silva | 20230586

Emília Santos | 20230446

Ricardo Kayseller | 20230450

June, 2024

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Index

Figure Index

1. Introduction of the Business

Retail4all is a retail company, which operates in a business to consumer (B2C) environment, since it sells its products directly to the clients.

This company has a total of 31 stores all over Portugal mainland, 10 of those have already an online presence, 18 haven't and the remaining 3 have no information on this matter. Regarding the points of supply (POS) and operators, Retail4all has a total of 7 POS and 14 operators, those of which we have information about their teams, roles, among other information. We can as well connect this information with sales amount and quantity, for instance.

Taking into consideration the three years of historic data (from 01/01/2020 until 02/06/2023) that were made available by the administration team, we are able to take a look at the company's sales details, as well as the products sold and their categories.

Regarding the products categories available in stores, we identified the following and some of its correspondent products:

- Electronics: phones, consoles, tablets, headphones, earphones, laptops, gaming accessories, speakers, smartwatches;
- Sports and fitness: running shoes;
- Clothing and accessories: tops, boots, sneakers, bottoms jeans;
- Home and kitchen: coffee makers, coffee machines, food processors, stand mixers, multi cookers.

2. Organizational Performance Overview

The organization exhibits an average annual growth rate of 14.9% from 2020 to 2022. While overall seasonality appears minimal, there is a discernible trend of slightly higher increases in sales observed consistently in January compared to previews months (Figure 1). This raise level of sales sustains throughout the remainder of the year.

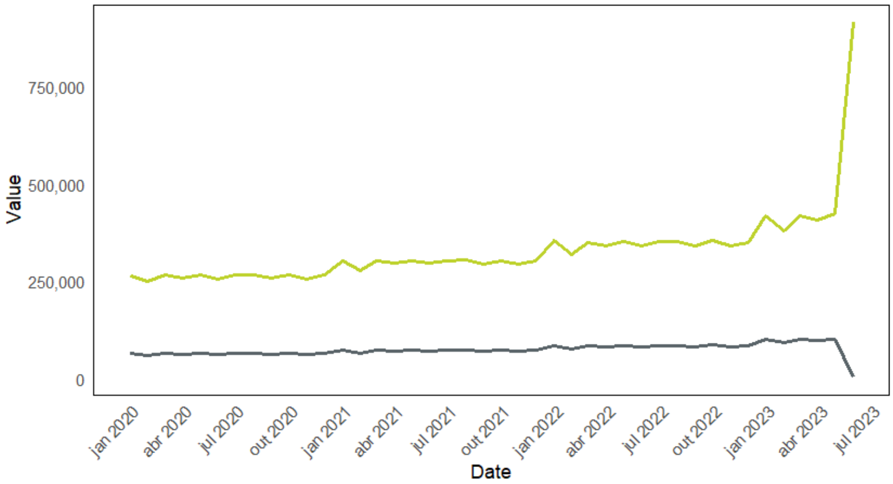


Figure 1: Evolution of transactions over time

Due to data limitations (information collected until June 2023) and anomalies observed, sales from 2023 were excluded from the calculations. Incongruity, during the last period, from May to June 2023, quantity decreased and the sales amount witnessed a notable increase.

In the past four years, the organization has demonstrated consistent growth in sales, with an upward trajectory from 2020 to 2022, as shown in Figure 2. However, it is essential to note that the data is only available until June 2023. Therefore, it is expected that 2023 follows the same trend as the previous years.

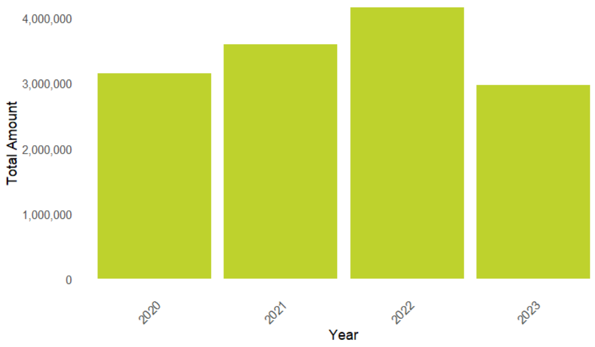


Figure 2: Total Sales Amount by Year

Out of the company's 31 stores, only 23 showed transactional activity in 8 different locations throughout the analysis period. Figure 3 illustrates the total sales amount across different locations from 2020 to 2023. Aguiar da Beira consistently exhibits the highest sales, with a peak in 2022. Conversely, Alvaiázere and Bragança show relatively lower sales figures throughout the observed period. Notably, all regions except Barcelos follow a similar trend as depicted in figure 1. Barcelos, which experienced modest but consistent growth between 2020 and 2022, saw exponential growth in 2023. However, further analysis is required to understand the anomalies observed in sales within this region.

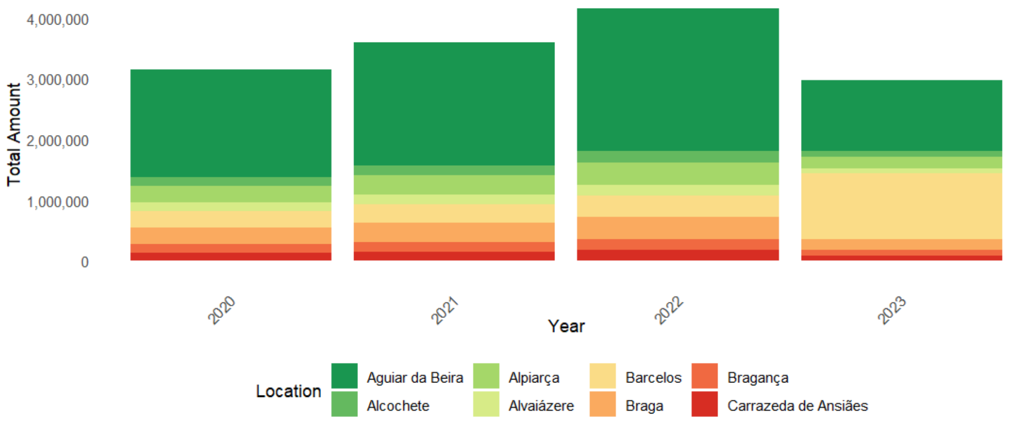
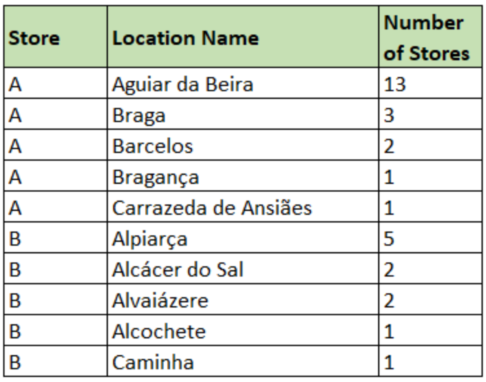


Figure 3: Total Amount by Year and Location

Table 1 presents insights regarding the distribution of stores per city. Notably, Aguiar da Beira stands out with 13 stores, potentially influencing its higher sales volume. Interestingly, Stores group A predominantly cluster in the northern region of Portugal, while stores from group B are concentrated in Lisbon or its vicinity. Caminha serves as the sole exception to this pattern.



Store	Location Name	Number of Stores
A	Aguiar da Beira	13
A	Braga	3
A	Barcelos	2
A	Bragança	1
A	Carrazeda de Ansiães	1
B	Alpiarça	5
B	Alcácer do Sal	2
B	Alvaiázere	2
B	Alcochete	1
B	Caminha	1

Table 1: Distribution of stores per city

Non-online transactions (N) reflect the overall trend depicted in Figure 1, showcasing a steady increase until 2022. Interestingly, online sales (Y) continue to exhibit growth, even extending into 2023 as shown in Figure 4, despite the absence of data beyond June 2023 (this is related with the anomalies data in June 2023).

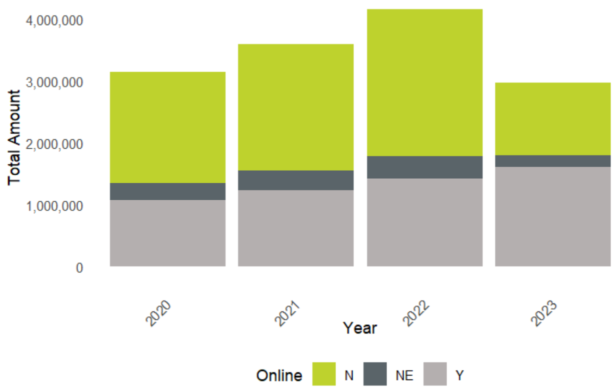


Figure 4: Sales Transactions Over Time (2020 - 2023)

Figure 5 reveals that the Electronics category accounts for over 50% of the total sales amount in 2022, demonstrating a significant dominance in consumer spending. Clothing & Accessories secures the second position, closely followed by Home & Kitchen, while Sports & Fitness lags behind in terms of sales performance. Furthermore, spending patterns remained constant across different product categories from 2020 to 2022.

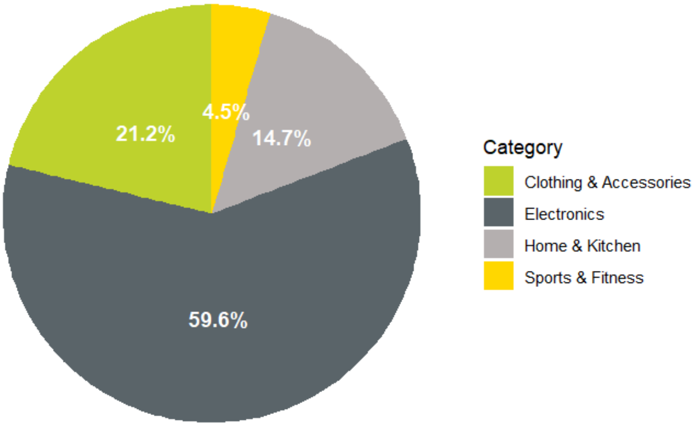


Figure 5: Total Sales Amount by Category in 2022

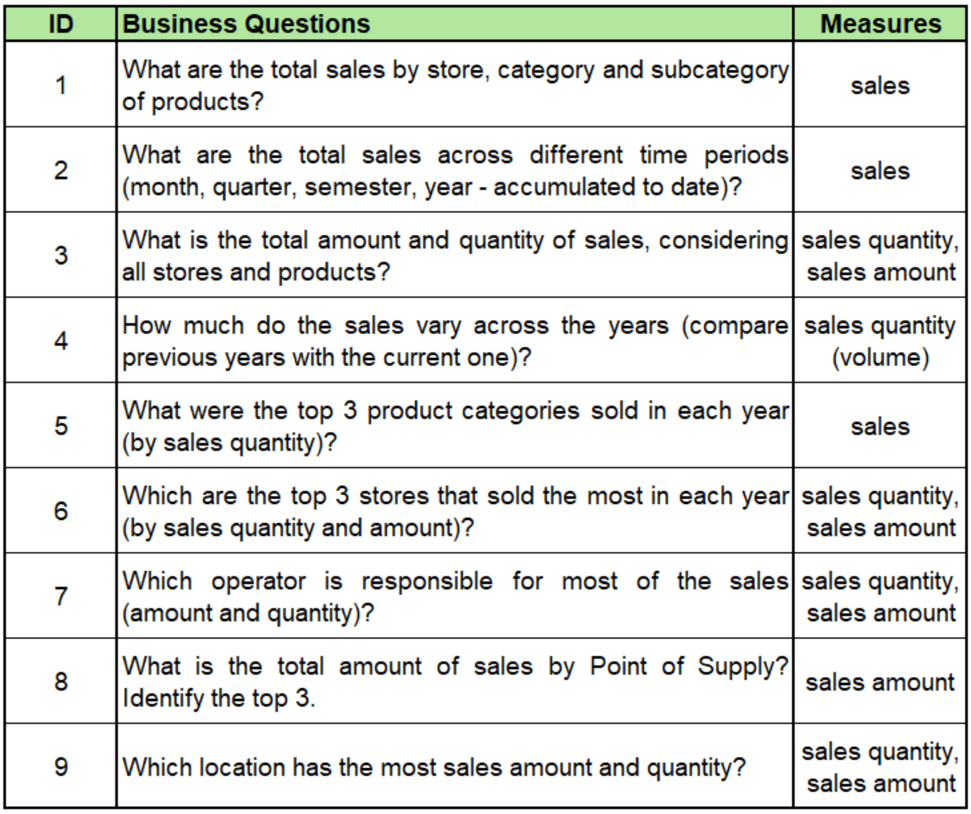
3. Business Problem

The main problems that Retail4all is looking to solve through BI solutions are related to the analysis of sales (monthly, quarterly and yearly - accumulated to date), quantity, amount and details (transaction level) of sales. Another challenges to take into account are the comparison between the current year and the previous one, the building of client profiles and sales aggregation, by store, category and subcategory of products.

Besides these challenges, the company is also requiring an implementation of Row-Level Security of Sales by User/Location.

These problems were considered given as context the fact that administration needs to improve analytical outputs and decision-making processes. These are two elements that are quick to achieve, due to the daily data updates, and with the BI applications will be even faster and more efficient.

4. Business Questions



ID	Business Questions	Measures
1	What are the total sales by store, category and subcategory of products?	sales
2	What are the total sales across different time periods (month, quarter, semester, year - accumulated to date)?	sales
3	What is the total amount and quantity of sales, considering all stores and products?	sales quantity, sales amount
4	How much do the sales vary across the years (compare previous years with the current one)?	sales quantity (volume)
5	What were the top 3 product categories sold in each year (by sales quantity)?	sales
6	Which are the top 3 stores that sold the most in each year (by sales quantity and amount)?	sales quantity, sales amount
7	Which operator is responsible for most of the sales (amount and quantity)?	sales quantity, sales amount
8	What is the total amount of sales by Point of Supply? Identify the top 3.	sales amount
9	Which location has the most sales amount and quantity?	sales quantity, sales amount

Tabel 2: Business questions

5. Data Sources

Retail4all's administration has made available CSV files extracted directly from the company's main servers. These files are updated daily, enabling analyses based on the latest information and improvement of decision-making processes. The datasets comprises sales records and complementary information, spanning approximately three years from January 1, 2020 to June 30, 2023.

The dataset is distributed across seven CSV files, encompassing a wide array of information including location data, sales records, operator details, points of supply, product information, and store specifics.

Sales (Sale ID, Datetime, SKU, Store, POS, Localização, Operator ID, Currency, Quantity, Amount)

The "Sales" CSV file provides information on all transactions conducted during the analysis period, including details such as Sale ID (an identifier for each sale), Datetime (indicating the precise date and time of each transaction down to the minute), SKU (Stock Keeping Unit, Product Reference), Store (identification of the store), POS (Point of Sale identification), Location (location identifier), Operator ID (operator identification), Currency (denoted in euros for all transactions), Quantity (quantity of items sold per sale), and Amount (amount per sale).

Products (SKU, Products Name, Category, Subcategory)

The CSV file "Products" provides information regarding all products offered by the company. It includes information such as the product reference, name, and classification into categories and sub-categories. This includes 78 unique products, categorized into 4 main categories and further divided into 31 sub-categories, regardless of the presence of duplicated products within the file.

Operators (Operator id, first_name, last_name, email, gender, role, team)

The “Operators” CSV file provides details about the company's employees, including their Operator ID, first name, last name, email, gender, role, and team. Interestingly, operators appear to be allocated across multiple stores, as they are responsible for handling transactions in different store locations.

Point of supply (POS id, name, email)

The “Point of Supply” (POS) CSV file comprises information about company suppliers, featuring details such as POS ID, company name, and email address for contacting each supplier. Additionally, it's noted that multiple stores are supplied by the same supplier.

Stores A and B (Store ID, Nome, Localização, Online)

The “Stores” CSV file offers details regarding the company's stores, including the store name, location ID (facilitating connection to the location file), and information about whether it operates as an online store (Y|S) or not (N), with some stores having unspecified values (missing values). Each store is restricted to a single sales format. This information is divided into two CSV files (Stores A and B), totaling 31 stores, with one store ID being misclassified (company GastroTech should be assigned Store ID 30).

Locations (Location ID, City, District, Country)

The CSV file related to “Locations” provides information regarding the geographical positions of the stores. Despite the absence of column headers, it contains the fields location ID, city, district, and country. However, it's noted that district information is missing for many cities. Considering the significance of district information for analysis, it will be supplemented from external sources to enhance the dataset's quality.

Date

Additionally, to facilitate analysis, a "Date" table was created. This provided a structured format for further examination and interpretation of the data.

Currency (Currency ID, Currency)

Furthermore, a table for currency was also added. This table provided information about Currency ID for each currency.

Mapping Districts (Cidade, Distrito, País)

This external CSV file contains information about districts and countries for each city listed in the location CSV file. Its primary purpose is to complement the district information found in the table locations.

6. Data Modelling Methodology

Dimensional modelling's primary goal is to create a database structure that is easy for end-users to formulate queries. Additionally, it aims to optimize query efficiency by minimizing the number of tables and relationships, reducing complexity, and minimizing the need for joins in user queries (Moody & Kortink, 2000).

Nonetheless, there are multiple approaches that can be used to design data warehouses. Moody and Kortink (2000) argue that different design principles should be used for designing the central data warehouse and data marts. The authors propose a method comprising four steps: categorizing entities, identifying hierarchies, collapsing them, and aggregating transaction data to construct dimensional models.

On the other hand, according to Kimball's methodology, designing a dimensional model involves selecting the business process, declaring the grain, identifying dimensions, and, lastly, identifying the facts (according to Kimball, 1996).

In this project, we are adopting Kimball's methodology, specifically we employed the Star Schema design approach. This decision was made due to the fact that Star Schema is easily understandable and adds less complexity to the schema, since it does not require additional joins compared to the Snowflake Schema (Moody and Kortink, 2000), 80% of the queries are single table browses (Kimball, 1996).

A. Identify the business process

The data warehouse represents the business process of retail product sales, focusing on measuring and analyzing key metrics related to sales performance and building of client profiles.

B. Identify the grain

The analysis concentrates on the daily sales volume and quantity attributed to each product, store, currency, point of supply, and operator, providing granular insights about the business. For example, in terms of sales, we can analyze the total sales amount per store (A or B), per semester and per product category, given the grain of our data. As a result, it can be said that the combination of the information given by the FK's (Figure 6) defines the granularity or level of detail of our business. Some of the foreign keys (FK's) are hierarchized. This hierarchical structure will be thoroughly discussed in Section 7.3.

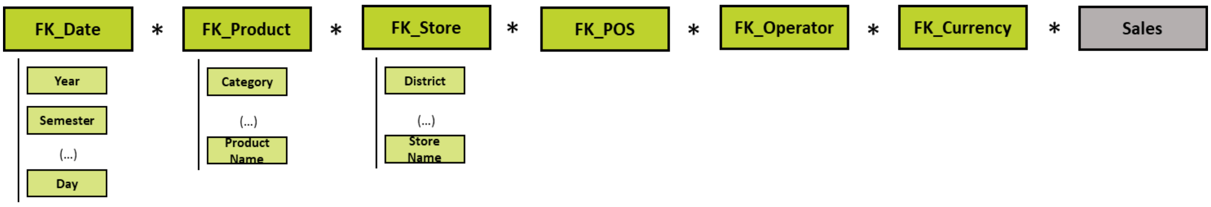


Figure 6: Sales Fact row

C. Identify the dimensions

Dimensions include Date, Product, Store, POS (Point of supply), Operator, and Currency, enabling comprehensive analyses and insights into various aspects of retail sales.

D. Finally identify the facts

The fact table contains measures related to the sales transactions, such as Sales Units and Sales Amount, linked to the dimensions through foreign keys.

7. Dimensional Model

The data warehouse is designed following the star schema model, as mentioned before, consisting of a central fact table and the dimensions tables. This structure enhances query efficiency, facilitates data analysis, and accommodates future scalability.

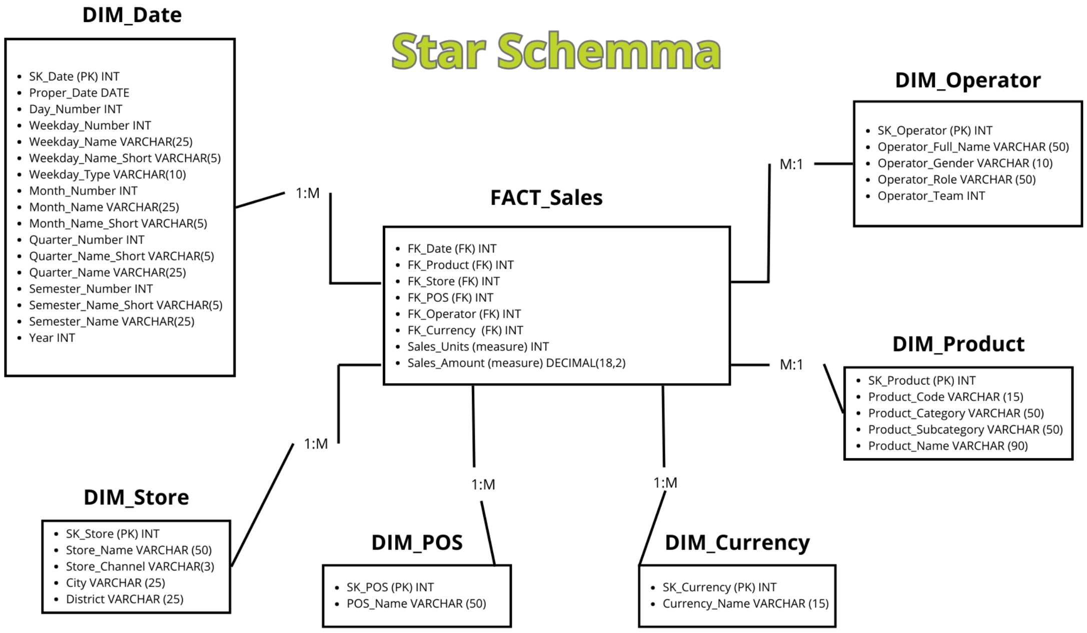


Figure 7: Design of the dimensional model

7.1. Fact table

The fact table can contain two distinct types of rows (Sales Fact row and Stocks Fact row). This project focuses on sales data. This “FACT_Sales” table (Table 3) is structured with the foreign keys 'FK_Date', 'FK_Product', 'FK_Store', 'FK_POS', 'FK_Operator', and ‘FK_Currency’ to ensure relational integrity. It also includes 'Sales_Units' for quantity sold, and 'Sales_Amount' for total monetary value. Even though transactions are currently in Euros, maintaining information about the currency allows for future scalability.

Column Name	Data Type	Description
‘FK_Date’	INT	Foreign Key
‘FK_Product’	INT	Foreign Key
‘FK_Store’	INT	Foreign Key
‘FK_POS’	INT	Foreign Key
‘FK_Operator’	INT	Foreign Key
‘FK_Currency’	INT	Foreign Key
‘Sales_Units’	INT	Number of units sold (measure)
‘Sales_Amount’	DECIMAL(18,2)	Total sales amount per sale (measure)

Table 3: FACT_Sales

7.2. Dimensions tables

The “DIM_Date” table (Table 4) is being created to organize and enhance data analysis, featuring the following attributes:

Column Name	Data Type	Description
‘SK_Date’	INT	Surrogate Key
‘Proper_Date’	DATE	Date (YYYY-MM-DD)
‘Day_Number’	INT	Day of the month
‘Weekday_Number’	INT	Number of the weekday (1, 2, 3, 4, 5, 6 or 7)
‘Weekday_Name’	VARCHAR(25)	Name of the weekday
‘Weekday_Name_Short’	VARCHAR(5)	Name of the weekday (short)
‘Weekday_Type’	VARCHAR(10)	Type of the weekday (weekday or weekend)
‘Month_Number’	INT	Number of the month (1, 2, 3, 4, ..., 12)
‘Month_Name’	VARCHAR(25)	Name of the month
‘Month_Name_Short’	VARCHAR(5)	Name of the month (short)
‘Quarter_Number’	INT	Number of the quarter (1, 2, 3 or 4)
‘Quarter_Name_Short’	VARCHAR(5)	Name of the quarter (short)
‘Quarter_Name’	VARCHAR(25)	Name of the quarter
‘Semester_Number’	INT	Number of the semester (1 or 2)
‘Semester_Name_Short’	VARCHAR(5)	Name of the semester (short)
‘Semester_Name’	VARCHAR(25)	Name of the semester
‘Year’	INT	Year

Table 4: DIM_Date

A ‘DIM_Store’ table is being developed to manage and analyze store-related data (Table 5).

We opted to designate Locations table as providing information about stores rather than sales due to several reasons:

- In the ‘FACT_Sales’ table, the consistency between the ‘locations_id’ and ‘store_id’ fields when compared with ‘DIM_Store’ table. This alignment suggests that the location represents the store's geographical position, rather than the transaction's location. If it were related to transactions, online sales would typically differ from in-store transactions due to the location of the client, which isn't the case here.

- Moreover, if we were to interpret it differently, the ‘FACT_Sales’ table would solely provide information about the locations where transactions occurred. This approach would overlook the importance of store locations, which are crucial for understanding the company's operations, market presence and decision-making processes, especially given the existence of stores with no sales. This is particularly relevant because each store without sales represents a loss of location information within the dataset.

Column Name	Data Type	Description
‘SK_Store’	INT	Surrogate Key
‘Store_Name’	VARCH(50)	Name of the store
‘Store_Channel’	VARCHAR(3)	Channel of the store (online or presencial)
‘City’	VARCH(25)	Store’s city
‘District’	VARCH(25)	Store’s district

Table 5: DIM_Store

In the ‘DIM_POS’ table (Table 6), the email column will not be taken into account, despite its presence in the original data, as it constitutes irrelevant information for the analysis.

Column Name	Data Type	Description
‘SK_POS’	INT	Surrogate Key
‘POS_Name’	VARCHAR(50)	Name of the supplier

Table 6: DIM_POS

A similar approach was taken on the ‘DIM_Operator’ table (Table 7).

Column Name	Data Type	Description
‘SK_Operator’	INT	Surrogate Key

'Operator_Full_Name'	VARCHAR(50)	Full name of the operator (concatenation of 'First Name' and 'Last Name')
'Operator_Gender'	VARCHAR(10)	Gender of the operator
'Operator_Role'	VARCHAR(50)	Role of the operator
'Operator_Team'	INT	Team to which the operator belongs

Table 7: DIM_Operator

Column Name	Data Type	Description
'SK_Product'	INT	Surrogate Key
'Product_Code'	VARCHAR(15)	Code of the product (SKU)
'Product_Category'	VARCHAR(50)	Category of the product
'Product_Subcategory'	VARCHAR(50)	Subcategory of the product
'Product_Name'	VARCHAR(90)	Name of the product

Table 8: DIM_Product

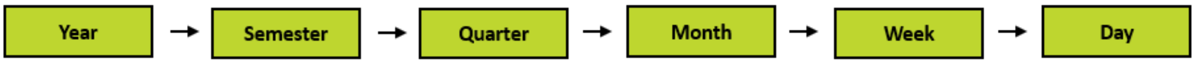
Column Name	Data Type	Description
'SK_Currency'	INT	Surrogate Key
'Currency_Name'	VARCHAR(15)	Name of the currency

Table 9: DIM_Currency

7.3. Hierarchies

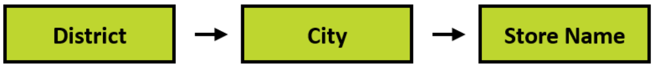
A key consideration in finalizing the hierarchies was ensuring they were streamlined for relevance and simplicity, aiming for an average depth of approximately 3 levels across dimensions. This was done to facilitate ease of use for end-users and to focus on the most impactful levels of data analysis.

Date dimension (depth = 6)



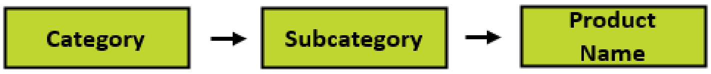
In the provided star schema, the table ‘DIM_Date’ is a key component, designed to support temporal analysis across the database. It includes attributes like day, week, month, quarter, and year, allowing for detailed time-based reporting and trend analysis. The granularity offered by this structure enables a multi-level view of data over time, which is crucial for understanding patterns and making informed decisions based on Sales performance.

Store dimension (depth = 3)



The exclusion of 'Country' from the Location dimension was a strategic choice reflecting this principle. Since the dataset was limited to ‘Portugal’, including a country level would not add meaningful analytical value and could potentially dilate the focus of the analysis. This decision also aimed to maintain an average depth of around 3 levels across the three hierarchies, which in turn allowed for the allocation of an additional level to the Date dimension, enhancing its analytical granularity without compromising the overall schema balance.

Product dimension (depth = 3)



The products are organized within the ‘DIM_Product’ table, categorized according to the structure provided by the company, which includes category, subcategory, and product name.

8. ETL Process

In the ELT (Extract, Load, Transform) process, we began by creating a dedicated workspace titled 'BI MAA 2024 Group 07'. This designated workspace enabled the execution of data extraction, loading, and transformation activities.

8.1. Lakehouse

The designated repository 'LH_GROUP_07_SOURCES' serves as the destination for not only uploading the company's internal data but also externally sourced complementary information, both relevant to the location and currency. Our approach involved uploading each file individually into the lakehouse, ensuring data integrity.

8.2. Warehouse

After uploading the data, within the established workspace, we constructed a Data Warehouse named ‘DW_2024_GROUP_07’.

Utilizing SQL queries, we designed and created the tables within the Data Warehouse environment based on the dimensional model design defined on section 7. This process involved defining the appropriate columns, specifying data types and setting constraints, particularly defining those columns that do not allow null values.

By designating certain columns as not allowing null values, particularly surrogate keys in dimension tables and foreign keys and measures, we ensure the integrity of the data model. In dimension tables, surrogate keys uniquely identify each record. By invalidating null values in these columns, we enforce the uniqueness constraint, preventing data redundancy and ensuring accurate merge with other tables. In the fact table, where foreign keys link to dimension tables and measures quantify business metrics, the absence of null values is equally important. Null values in foreign keys could lead to data inconsistencies and integrity issues during analysis and reporting. Likewise, null values in measure columns could distort analytical results or calculations.

Additionally, all columns in the 'Date' table are set to not allow null values, as this table was constructed by us. This ensures completeness and consistency in the temporal analysis and reporting.

8.3. Dataflows

In order to prepare and transform our data, we resorted to the Dataflow Gen2 functionality in Data Factory, this functionality is relevant since in each dataflow we kept the tables or information regarding the excel files we had previously, as well as joining information from different excel files and joining.

To create each dataflow we had to, firstly, select the data source (Lakehouse, in our case: ‘LH_2024_GROUP_07_SOURCE’) and then from Files choose the data, selecting the CSV which we wanted to work on.

This being said, the dataflows created were ‘Operators Dataflow’, ‘POS Dataflow’, ‘Products Dataflow’, ‘Sales Dataflow’, ‘Store Dataflow’, ‘Date Dataflow’ and ‘Currency Dataflow’. After creating the dataflow, making the respectives transformations and data engineering steps on them (better explained on point 9 of the report), we selected as a destination an existing table (for instance, ‘DIM_POS’) present in the Warehouse (‘DW_2024_GROUP_07’), for each dataflow. Still in the destination settings, the column mapping was done, i.e., we attributed one source according to the destination column (both having the same data type). Lastly, we had to publish it and refresh in the Workspace, so that the changes were kept.

8.4. Pipelines

The pipeline execution involves a series of sequential steps to ensure the integrity and completeness of our data warehouse. Each phase must be successfully completed before proceeding to the next, so it's necessary to have intermittent pauses in the process as we await completion of each stage (Figure 8). These waits allow us to verify the accuracy and completeness of each stage before advancing to the subsequent ones, thereby safeguarding the reliability of our data warehouse.

Prior to loading new data, we ensure that the data warehouse is empty of any existing data (Phase 1). This is achieved by deleting both dimension and fact tables if they existed. Only, once this step is completed, we proceed to the next phase.

Following the data warehouse cleanup, the next phase involves the ingestion of data into the dimension tables, in accordance with the transformations defined previously in each dataflow.

With the dimension tables populated, we proceed to load data into the fact table. Unlike dimension tables, the fact table contains transactional data and requires information from the dimension tables. Therefore, we prioritize loading the fact table last in the pipeline to ensure that all necessary dimension data is available for merging.

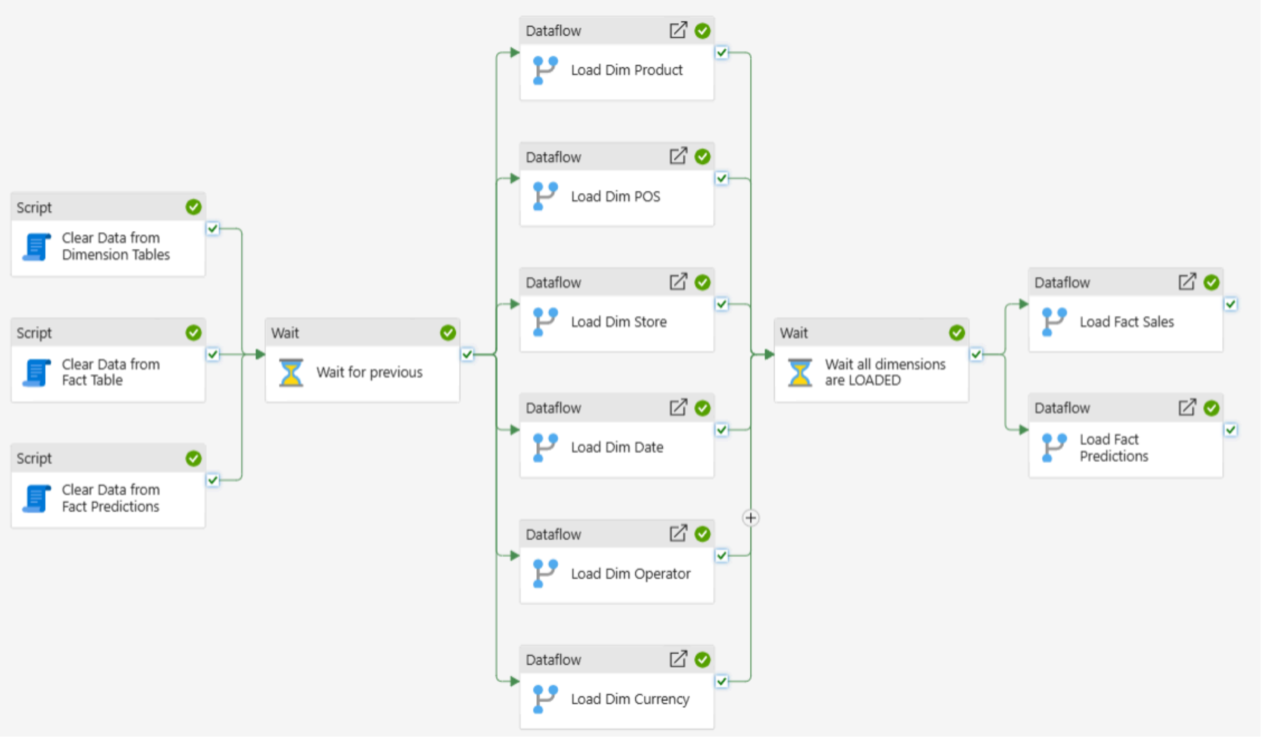


Figure 8: Group's Data Pipeline

9. Data Engineering

In the following dataflows, the first step was always the promotion of the first row to header.

9.1. POS Dataflow

The change of data type of 'POS id' to a whole number was made, as well as the removal of the column containing the emails, since it would have no relevant use.

9.2. Operator Dataflow

We changed column types, ‘Operator id’ and ‘team’ to whole number, ‘first_name’, ‘last_name’, ‘gender’, ‘email’ and ‘role’ to text. The column relative to emails was removed, as on ‘POS Dataflow’. Then we added a custom column ‘Full Name’, concatenating the columns ‘first_name’ and ‘last_name’.

9.3. Product Dataflow

We changed the following column’s data types to text: ‘SKU’, ‘Product Name’, ‘Category’ and ‘Subcategory’. With ‘SKU’ selected, we made sure there were no duplicated rows, using the Remove Duplicates functionality. Lastly, we had to add an index as ID (naming it ‘SKU ID’), since the SKU from the products was composed of capital letters and numbers and it had to be just numbers.

9.4. Date Dataflow

To create this dataflow we started from getting the data through a blank query, in which we wrote ourselves the new script listing the initial (2020/01/01) and final dates (2024/01/01 - is not included) and the duration of the interval between each date (1 day). This script was then converted to a table and the column obtained was named ‘Proper Date’, with Date as the data type. Having this date column selected, the following ones in Table 10 were created:

Columns based on ‘Proper Date’	Data Type
‘Year’	Whole Number
‘Month Number’	Whole Number
‘Month Name’	Text
‘Quarter Number’	Whole Number
‘Week of Month’	Whole Number
‘Day Number’	Whole Number
‘Day of Week’	Whole Number
‘Weekday Name’	Text

Table 10: Columns created based on ‘Proper Date’

New columns	Data Type	Creation of the columns
‘Month Name Short’	Text	Extracting the first 3 characters from ‘Month Name’

‘Weekday Name Short’	Text	Extracting the first 3 characters from ‘Weekday Name’
‘Weekday Type’	Text	Conditional column based on ‘Quarter Number’ (If ‘Weekday Number’ = 0 or = 6, then ‘Weekday Type’ = ‘Weekend’, if not then ‘Weekday’)
‘Quarter Name’	Text	Inserting the prefix ‘Quarter’ on the values of column ‘Quarter Number’
‘Quarter Name Short’	Text	Inserting the prefix ‘Qrt’ on the values of column ‘Quarter Number’
‘Semester Number’	Whole Number	Conditional column based on ‘Quarter Number’ (If ‘Quarter Number’ <= 2, then ‘Semester Number’ = 1, if not then 2)
‘Semester Name’	Text	Inserting the prefix ‘Semester’ on the values of column ‘Semester Number’
‘Semester Name Short’	Text	Inserting the prefix ‘Smt’ on the values of column ‘Semester Number’
‘SK Date’	Whole Number	Column from examples (from selection), having the ‘Proper Date’ column selected and then wrote 2 rows as examples (1st row: ‘20200101’ and 2nd row: ‘20200102’)

Table 11: New columns created based on the existing ones

A column named ‘Is Special’ could have been implemented in this dataflow, if there was a very high or low point of sales on these special days, for example, New Year’s Day (can be observed on Figure 1, relative to the evolution of transactions over time). Since these patterns didn’t occur on the data from Retail4all, we decided not to include the column. If they eventually verify in further sales, a future column ‘Is Special’ can be created, through the association of the days and their respective meaning.

9.5. Currency Dataflow

The data type of the 'Currency ID' column was changed to a numeric type (Integer) and the 'Currency' column, containing currency codes (EUR, USD, GBP, CHF), was set to text data type.

9.6. Store Dataflow

For the store dataflow, the process involved merging/append four tables into one consolidated dataset. Below, we outline the key steps undertaken in this process:

<u>Table Locations:</u> The column data types were adjusted to match those required by the destination data. Due to significant missing information in the 'district' column, Table Location was merged with the Table Mapping Districts, based on city names. Subsequently, the original 'district' columns were dropped.

Tables Store A and B: The store information was initially divided into two CSV files: Stores A and B. In the 'online' column, where 'S' indicated 'Yes', the values were replaced with 'Y' for consistency. Missing values were substituted with 'n/a' (not available). Additionally, in Store B, a correction was made to the 'store id,' replacing instances of '20' with '30' to rectify a typographical error. Since 'store id = 20' already existed in Store A.

Merging/Appending Store Tables: Finally, Store Tables A and B were appended together into a single dataset. The table locations were merged to provide geographical characterization for each store (City and District). During this process, columns that were used solely for facilitating the merge were identified and subsequently removed, as they were no longer necessary. Additionally, data type consistency checks were performed to ensure alignment with the destination data requirements, specifically the 'ID store' was changed to 'whole number' and the column 'Distrito'' was renamed as 'District'.

9.7. Sales Dataflow

Various fields were standardized to specific data types (Sale ID, Store, POS, Location, Quantity and Operator ID, Date to whole number type, Amount to decimal number type and Currency and SKU to text type) and the data underwent localization adjustments, particularly for datetime fields to align with regional settings, such as 'pt-PT' for Portugal.

For columns in the sales table without a direct key reference, such as columns with product and currency information, we performed merges with the respective dimension tables based on unique codes. Subsequently, we extracted the corresponding key associated with each code, establishing a linkage between the sales data and the dimension tables. Moreover, for columns already containing keys within the sales table (remaining columns), we still conduct merges with the corresponding dimension tables. If a merge returns null values, indicating a lack of corresponding keys in the dimension tables, we systematically discard the affected rows to maintain data consistency. Once merged, specific fields from these dimension tables are expanded, specify the surrogate keys in the dimension tables, and then renamed. To ensure the data is filtered, cleaned and focused, the workflow includes steps to reorder, rename and removal of unnecessary columns, and also ensures the non-existence of blank rows and duplicates.

Finally, as the detailed time information has been removed, multiple rows with the same transaction characteristics remain. Therefore, we group these rows by each foreign key ('FK_date', 'FK_product', 'FK_store', 'FK_POS', 'FK_operator', 'FK_currency') and perform the sum operation on the columns Qty and Amount to consolidate the transactional data.

10. Sales Forecasting with ARIMA model: Annual Sales Prediction

The 'Business_Intelligence_Forecasting.ipynb' notebook is designed with the primary goal of developing predictive analysis and leveraging statistical methods to forecast monthly sales from June 30, 2023, to May 31, 2024, utilizing ARIMA (AutoRegressive Integrated

Moving Average) model, a time series tool for forecasting. This model was selected for its efficacy in handling data with trends and seasonality, which is typical of sales data.

An exploratory analysis provides an initial understanding of the data's underlying patterns and trends. Visual tools such as time series plots and histograms are employed to identify seasonal effects and sales cycles, which are critical in configuring the ARIMA model parameters. Also, the granularity in sales forecasting had to be shifted from days to months.

An excel file with the predictions was extracted, employed in a fact table (FACT_Predictions), and associated with a new dataflow (Predictions Dataflow), which was integrated in our pipeline (Figure 8).

11. Model optimization

In the semantic model level, we established connections between the keys and made sure the relationships (for instance, one-to-many) were correct, renamed the columns, hide the irrelevant keys, marked 'DIM_Date' as date table and configured the hierarchies (Year Hierarchy, Product Category Hierarchy and Store District Hierarchy, done based on the hierarchies defined on previous steps of this report). These were the main steps for optimizing the model, regarding, as well, that some columns went through a change of properties:

Columns	Properties Changed
Month	Sort by column: Month Number
Quarter	Sort by column: Quarter Number
Semester	Sort by column: Semester Number
Weekday	Sort by column: Weekday Number
Sales Amount	Format: Currency; Thousands separator: Yes; 2 decimal places; Currency format: € Euro; Summarize by Sum
Sales Units	Thousands separator: Yes

Table 12: Changes on columns' properties

12. Measures and calculated columns

In order for us to conduct a well structured report, have quality insights and answer the Retail4all business needs, the following DAX measures and calculated were created:

Measures	Explanation
Sales Amount per Store	Sum of sales amount per Store
Sales Amount per POS	Sum of sales amount per Point of Sales
Sales Units per Product	Sum of sales units per Product Name
Total Sales	Sum of sales amount
Units Sold	Sum of sales units
Sales Amount YoY%	Annual growth rate of sales amount based on the first and last dates in the date dimension
Average Unit Price	Average price per unit sold across all sales

Table 13: DAX Measures

Calculated Columns	Explanation
Unit Price	Derived column computed by dividing the "Total Sales" by the "Units Sold"
Sales Tax	Sums all sales amounts from the 'FACT_Sales' table and then multiply the sum by 5%
Tax	Difference between total sales and net sales

Table 14: Calculated Columns

13. Dashboard technical aspects

At this stage, we started working with Power BI Desktop in order to deploy a more detailed and well-developed report. The report is mainly composed of various data visualizations, each tailored to provide specific insights into the organization's performance metrics and trends.

Below is an extensive analysis of the various dashboards within the Power BI report, as captured in the provided images. Each section offers a detailed look at the visualization techniques used and their alignment with the project's objectives.

Performance Overview Dashboard:

- ➔ **Visualization Techniques:** Incorporates a line chart for total sales over time, which is useful for tracking sales dynamics and trends. A geographical heatmap complements this by displaying regional sales data, helping to pinpoint areas of strong and weak performance.

➔ **Interactive Features:** Tooltips are applied to both the line chart and heatmap, offering detailed data insights when users hover over specific points or regions.

➔ **Alignment with Project Scope:** This dashboard provides a holistic view of sales performance, which is critical for strategic decision-making and aligns with the project’s goal of enhancing data-driven strategies.

Sales Analysis Dashboard:

➔ **Visualization Techniques:** Features a timeline analysis of sales and a detailed bar graph that breaks down sales by categories over a specified period. This layout helps in identifying sales peaks and trends within specific categories.

➔ **Interactive Features:** Tooltips on the timeline and bar graphs provide exact figures and additional contextual information, enhancing data understanding without additional navigation.

➔ **Alignment with Project Scope:** Offers insights into sales trends and performance across different categories, aiding in targeted marketing and sales strategy adjustments.

Detailed Sales and Geographic Distribution Dashboard:

➔ **Visualization Techniques**: Uses complex geographic visualizations to display the distribution of sales across regions. Time-series graphs track sales over selected periods, highlighting trends and cycles.

➔ **Interactive Features**: Advanced tooltips on the geographic and time-series visualizations offer detailed sales metrics, fostering a deeper understanding of regional dynamics.

➔ **Alignment with Project Scope:** These visualizations provide deep insights into regional sales performance and time-based sales data, crucial for strategic planning and marketing efforts.

Product Analysis Dashboard:

➔ **Visualization Techniques:** Utilizes pie charts to illustrate sales distribution among different product categories and a flowchart to show the best-selling products and their categories.

➔ **Interactive Features:** Tooltips on pie charts and flowcharts enrich the information by detailing product performance metrics.

➔ **Alignment with Project Scope:** Supports inventory management and marketing strategies by providing a clear picture of which products are performing best and their sales contributions.

Operators and POS Analysis Dashboard:

- ➔ **Visualization Techniques**: Doughnut charts represent sales by different operators or sales points, offering a clear breakdown of sales contributions. Geographic maps show the location of POS, aiding in spatial analysis.
- ➔ **Interactive Features**: Detailed tooltips provide granular sales data for each operator or POS; bookmarks can be used to quickly switch between different POS views or scenarios.
- ➔ **Alignment with Project Scope:** Helps in optimizing sales strategies and operational management by providing a clear view of operator and POS performance across different regions.

Location Analysis Dashboard:

- ➔ **Visualization Techniques:** Employs pie charts and bar graphs to show sales quantities by city, alongside a map displaying the geographic distribution of stores.
- ➔ **Interactive Features:** Tooltips on charts and maps provide immediate insights into sales data by city, which assists stakeholders in making data-driven location decisions.
- ➔ **Alignment with Project Scope:** Crucial for understanding the market dynamics in different locations, supporting decisions related to store placements and regional sales strategies.

Each dashboard is carefully designed to cater to specific aspects of the project scope, ensuring that the data visualization techniques and interactive features like tooltips and bookmarks are not only suitable for the data being analyzed but also effectively support decision-making processes. This comprehensive approach enhances the overall usability of the report, making it a vital tool for stakeholders to derive actionable insights.

14. Analysis and discussion of the outputs of the project

Upon the build up of the dashboard, the business questions and needs, previously explained in this report, were met and can be found according to Table 15.

Business Questions (the same as reports's point 4.)	Answers present in the dashboard
1. What are the total sales by store, category and subcategory of products?	Page 6 (sales by location/store) and page 3 (sales by product/subcategory)
2. What are the total sales across different time periods (month, year, accumulated to date)?	Page 2 (sales by month, year, year to date)
3. What is the total amount and quantity of sales, considering all stores and products?	Pages 1 and 2 (total sales amount and quantity)
4. How much do the sales vary across the years (compare previous years with the	Page 2 (sales by month, year, year to date)

current one)?	
5. What were the top 3 product categories sold in each year (by sales quantity)?	Page 3 (sales quantity per year, by product categories)
6. Which are the top 3 stores that sold the most in each year (by sales quantity and amount)?	Page 6 (sales amount and quantity per year, by store/location)
7. Which operator is responsible for most of the sales (amount and quantity)?	Page 4 (best employer of the year)
8. What is the total amount of sales by Point of Supply? Identify the top 3.	Page 5 (sales by POS)
9. Which location has the most sales amount and quantity?	Page 6 (sales amount and quantity per year, by store/location)

Table 15: Answer to the business questions

Regarding the outputs or insights itself, firstly, we identified in an overview that this data concerns a total of 31 stores, 7 Point of Sales, 15 operators and 4 product categories.

In terms of the actual sales, a total of 3M units were sold, corresponding to 12,95M Euros, being 2022 the year with the most sales and higher growth rate (from 2023, we only have data until May). October 2022, was the best selling month, in terms of sales amount. The patterns observed through the years and respective months, are expected to maintain, i.e., the general sales forecast (Power BI tool) shows a continuous growth from year to year, although it considers a slight decrease each year (from January to December). The forecast obtained for each store, using Machine Learning techniques, projects an increase in sales along the predicted months, in all stores.

From the 4 categories of products, 31 subcategories and 78 different products, the top 3 best selling product categories (by sales amount and quantity) are Electronics, Clothing & Accessories, and, Home & Kitchen, from the most to the least sales. This verifies in all years. The majority of the operators are considered sales staff and the best employee of Retail4all, across the years, is Stern Burgyn. The POS that sold the most was Lazzy, followed by Kwimbee and Realcube. The region where there are more stores is Guarda, having the city with the highest sales amount and quantity, Aguiar da Beira.

15. Conclusion

This report details the extensive implementation of business intelligence and data management strategies at Retail4all, underpinned by state-of-the-art tools such as Power BI, ARIMA modeling, and Azure Data Factory. The integration of Azure Data Factory has been pivotal in orchestrating and automating the ETL processes, ensuring efficient data integration and consistency across various data sources. This setup has significantly enhanced the reliability and timeliness of data available for analysis.

The project has successfully harnessed these technologies to translate vast amounts of retail data into actionable insights, which have substantively contributed to strategic decision-making processes. The predictive analytics implemented, particularly through ARIMA models, have equipped Retail4all with robust sales forecasts and deep analytical capabilities, extending from June 2023 to May 2024, facilitating precise strategic planning and effective inventory management.

Moreover, the development of a comprehensive dimensional data model has simplified user interactions with the BI tools, enhancing the efficiency of data queries and supporting a user-friendly reporting environment. The customized dashboards provide a granular view of sales performance across various dimensions—geographical, temporal, and product-related—enabling detailed performance analysis and better operational visibility.

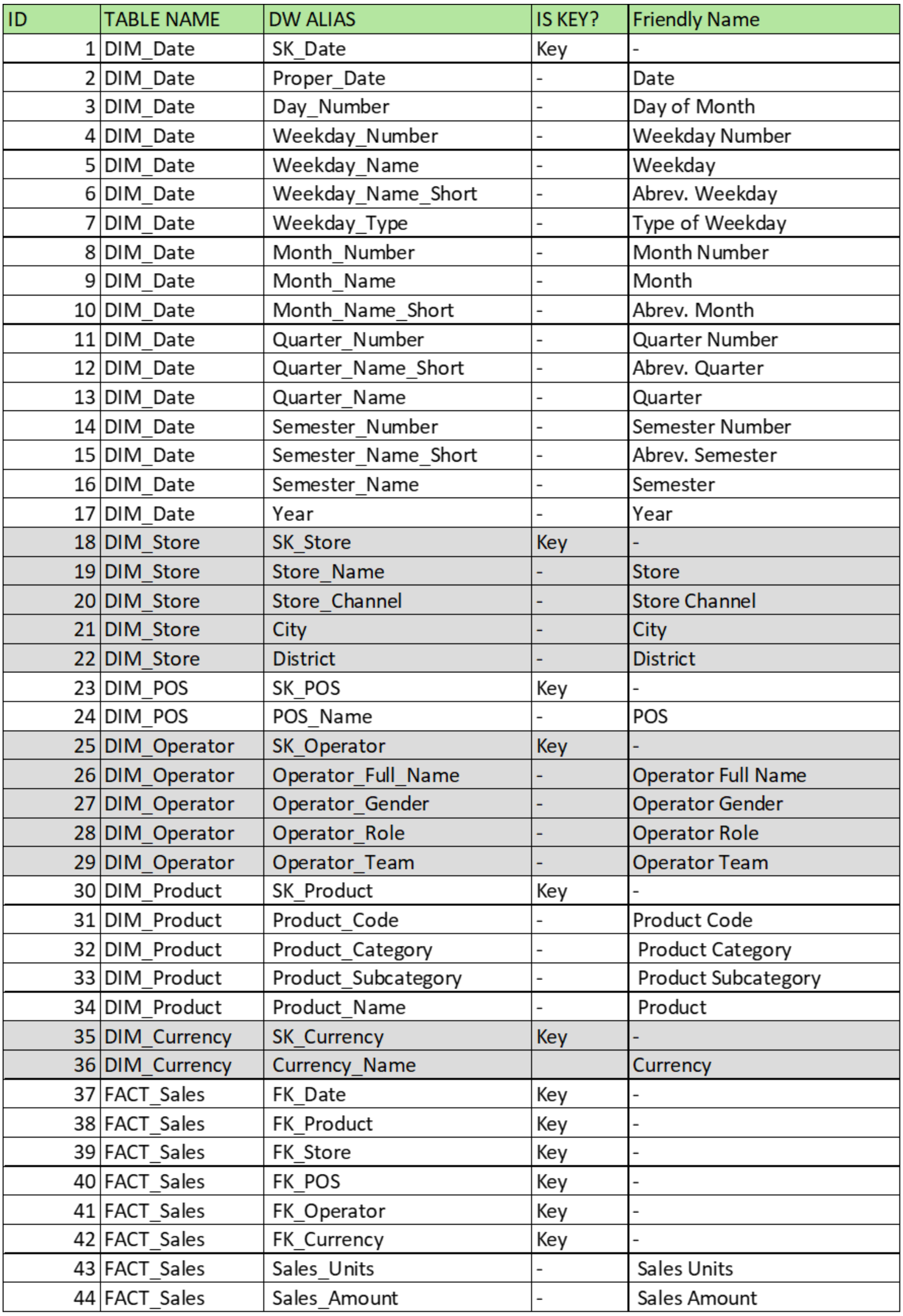
As Retail4all continues to evolve, it is recommended to further enhance these BI systems by integrating real-time data analytics and exploring advanced machine learning algorithms to refine forecasting accuracy. Expanding the digital footprint to encompass all physical stores could also capitalize on the burgeoning digital market, particularly in high-growth regions.

In conclusion, the strategic application of business intelligence, coupled with sophisticated data integration and management facilitated by Azure Data Factory, has positioned Retail4all to navigate the complexities of the retail market adeptly, driving future growth and maintaining a competitive edge in the industry.

16. References

- Kimball, R. (1996). The Data Warehouse Toolkit: Practical Techniques For Building Dimensional Data
- Moody, D. L., & Kortink, M. A. R. (2000). From enterprise models to dimensional models: a methodology for data warehouse and data mart design. Design and Management of Data Warehouses
- Siva, B. (2023, September 28). How and When to Use Dataflows in Power BI. PhData. https://phdata.io/blog/how-and-when-to-use-dataflows-in-power-bi/
- Hayes, A. (2024, April 6). Autoregressive Integrated Moving Average (ARIMA) Prediction Model. Investopedia. https://www.investopedia.com/terms/a/autoregressive-integratedmoving-average-arima.asp
- GfG. (2024, January 29). Time Series Analysis Visualization in Python. GeeksforGeeks. https://www.geeksforgeeks.org/time-series-data-visualization-in-python/

Appendix A



ID	TABLE NAME	DW ALIAS	IS KEY?	Friendly Name
1	DIM_Date	SK_Date	Key	-
2	DIM_Date	Proper_Date	-	Date
3	DIM_Date	Day_Number	-	Day of Month
4	DIM_Date	Weekday_Number	-	Weekday Number
5	DIM_Date	Weekday_Name	-	Weekday
6	DIM_Date	Weekday_Name_Short	-	Abrev. Weekday
7	DIM_Date	Weekday_Type	-	Type of Weekday
8	DIM_Date	Month_Number	-	Month Number
9	DIM_Date	Month_Name	-	Month
10	DIM_Date	Month_Name_Short	-	Abrev. Month
11	DIM_Date	Quarter_Number	-	Quarter Number
12	DIM_Date	Quarter_Name_Short	-	Abrev. Quarter
13	DIM_Date	Quarter_Name	-	Quarter
14	DIM_Date	Semester_Number	-	Semester Number
15	DIM_Date	Semester_Name_Short	-	Abrev. Semester
16	DIM_Date	Semester_Name	-	Semester
17	DIM_Date	Year	-	Year
18	DIM_Store	SK_Store	Key	-
19	DIM_Store	Store_Name	-	Store
20	DIM_Store	Store_Channel	-	Store Channel
21	DIM_Store	City	-	City
22	DIM_Store	District	-	District
23	DIM_POS	SK_POS	Key	-
24	DIM_POS	POS_Name	-	POS
25	DIM_Operator	SK_Operator	Key	-
26	DIM_Operator	Operator_Full_Name	-	Operator Full Name
27	DIM_Operator	Operator_Gender	-	Operator Gender
28	DIM_Operator	Operator_Role	-	Operator Role
29	DIM_Operator	Operator_Team	-	Operator Team
30	DIM_Product	SK_Product	Key	-
31	DIM_Product	Product_Code	-	Product Code
32	DIM_Product	Product_Category	-	Product Category
33	DIM_Product	Product_Subcategory	-	Product Subcategory
34	DIM_Product	Product_Name	-	Product
35	DIM_Currency	SK_Currency	Key	-
36	DIM_Currency	Currency_Name		Currency
37	FACT_Sales	FK_Date	Key	-
38	FACT_Sales	FK_Product	Key	-
39	FACT_Sales	FK_Store	Key	-
40	FACT_Sales	FK_POS	Key	-
41	FACT_Sales	FK_Operator	Key	-
42	FACT_Sales	FK_Currency	Key	-
43	FACT_Sales	Sales_Units	-	Sales Units
44	FACT_Sales	Sales_Amount	-	Sales Amount

Tabel A1: Master list of friendlier field names