NOVA IMS
Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Intelligence

Topic: Reporting

David Psiuk	number: 20230818
Noah Campana	number: 20230996

Group 02

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

June, 2024

Table of Contents

1 Introduction 1

2 Warehousing 1

2.1 Business Challenges 1

2.2 Formulating the Business Questions 2

2.3 Data Sources 3

2.4 Data Modelling Methodology 3

2.4.1 Identifying the Fact 3

2.4.2 Identifying the Dimensions 3

2.4.3 Identifying the Business Process 4

2.4.4 Identifying the Grain 4

2.5 Data Warehouse architecture 5

3 Extract, Transform, Load 6

3.1 Improvement of Initial Data Warehouse Design 6

3.2 Set-Up for E.T.L. 7

3.3 Data Engineering 8

4 Reporting 10

4.1 Model optimization 10

4.2 Calculated Columns and Measures 11

4.3 Technical Aspects of the Report 12

4.3.1 Configuration of the Report 12

4.3.2 Technical Elements of the Report 14

4.4 Analysis and Discussion 14

4.4.1 Time-Specific Sales Trends Analysis 15

4.4.2 Transaction Analysis 16

4.4.3 Performance Metrics 16

4.4.4 Product Insights 17

4.4.5 Operator Influence 18

5 Conclusion 20

6 Bibliography 21

7 Appendix 21

1 Introduction

What once began as an online bookshop has now evolved into the largest e-commerce company in the world. Amazon is no longer just a shop where you can find everything you need but has changed the way that consumers approach shopping, influencing expectations for convenience and delivery speed. Retailers are competing against each other in a retailing business itself. E-commerce has become a crucial part of every person's life, while the traditional retail experience becomes less and less attractive. As e-commerce continues to increase in popularity, retailers must adapt and find ways to remain competitive and innovative. A recent McKinsey report highlights that 80% of shoppers expect personalized experiences, and businesses that deliver them see a lift of up to 10% in revenue (Lindecrantz, Tjon Pian Gi, & Zerbi, 2020). To discover such market opportunities and momentum within their current customer base, traditional retailers need to understand their customers and respond quickly. Because what might be important today, might be redundant tomorrow.

Retail4all is a traditional retail company which focuses on selling electronic and kitchen devices as well as sport and fitness clothing. The company inherits 31 stores across whole of Portugal and has already integrated e-commerce channels in some locations, which is often referred to as omnichannel retailing. While they already collect and store data, they want to make more decisions based on the generated data. Therefore, the company has granted the access to the customer data which includes sales data of 3 years starting from 2020.

2 Warehousing

2.1 Business Challenges

Industry Challenges: Companies like Retail4all are facing many challenges. The rise of online shopping is intensifying competition as price comparisons or similar products are just a few clicks away. This shift increased the competition and changed consumer expectations, as customers now seek more convenience, faster delivery options, and a well-functioning omnichannel shopping experience. Retailers must adapt by utilizing data analytics for better inventory management and customer insights. Another challenge is maintaining profitability while investing in technology and innovation to meet these evolving consumer demands.

Specific Challenges: To tailor well suiting marketing strategies and target customers more precise, the need for client profiles arises. Furthermore, another major challenge is managing inventory effectively to balance supply and demand, reducing overstock and understock situations that lead to lost sales or increased storage costs. Additionally, the management of 31 shops suggests that the potential of high performing branches should be exploited and reasons for low revenues at other locations should be analyzed. To meet the demands of consumers, Retail4all should expand its online presence and integrate digital sales channels to offer a seamless shopping experience. Ultimately, utilizing predictive analytics to identify emerging trends and facilitate adjustments to marketing campaigns and product offerings.

Therefore, the business intelligence system should be able to collect and process data about the customers, such as demographic data and personal preferences, analyze sales characteristics over time and finally compare different shops with each other.

2.2 Formulating the Business Questions

The data warehouse for Retail4all will be designed to answer a series of business questions. These questions are critical to make strategic, tactical and operational decisions to ensure that the organisation remains competitive and responsive to market changes. The questions identified, along with the corresponding dimensions and implications for decision making, are as follows:

Time-Specific Sales Trends Analysis

1. What are the weekly, monthly, and yearly sales trends by volume and value?
2. How does the current year's sales performance compare to the previous year?
3. Which locations have shown the most significant growth in sales over the past year?
4. What are holiday /weekend trends?
5. What is the average amount of money spent in each store per day?
6. What is the distribution of sales quantities for each product category by the day of the week, identifying potential peak sales days?

Transaction Analysis

7. How does the average sales quantity vary across different cities?
8. What are the average sale amounts per transaction for each product category, and how do these vary by city?
9. What is the ratio of a product's sales amount / frequency of sales?
10. How does the existence of an online shop influence the sales of certain product categories?
11. How does the existence of an online shop influence the frequency of sales?

Performance Metrics

12. Which stores are the top performers in terms of revenue?
13. Which Point of Supply (POS) has the highest number of transactions for top-selling products?

Product Insights

14. What were the top products sold in each year or quarter?
15. What are the sales patterns for high-end vs. low-end products in terms of quantity sold and total revenue over time?
16. Which product categories are most popular in specific cities?
17. Are there products which are ordered more online than in the store in relative terms?

Operator Influence

18. How does the gender of staff affect sales for different product categories?

19. Who were the top N operators across different quarters / years?
20. Which single operators have the highest sales in terms of quantity and amount for each product category?
21. Which operator team has the highest amount of sales and which produces the highest revenue for the business?

The answers to these questions enable the company to make strategic decisions, e.g. which market to enter, tactical decisions, e.g. how to react to seasonal fluctuations, and to control the operational business in terms of resource management.

2.3 Data Sources

Retail4all allowed access to seven data sources. The main table embodies records of all sales precisely on the minute. Three years of sales added up to 1048575 records of sales in the main table. Additionally, the names, locations and operators of the stores, the products and the point of sales were given. The retail stores are situated across various locations in Portugal and offer a wide product range in categories such as Electronics, Clothing & Accessories, Home & Kitchen, and Sport & Fitness. The personnel comprise 15 staff members, including managers, directors such as floor and sales staff. Retail4all collaborates with seven distinct suppliers and only a selection of the stores offer online shopping options. A quick look at the data sets reveals that some redundancies are present in the data, e.g. 133 locations are listed, but only 31 shops exist.

2.4 Data Modelling Methodology

We chose to follow the Kimball methodology to model the data warehouse. We therefore followed the bottom-up approach, in which data marts are defined first, which then lead to a company-wide warehouse. Typically, the Kimball approach identifies the business process first, then the grain, followed by the dimensions and finally the facts. However, we have adapted the approach so that the facts are identified first, followed by the dimensions and the business process, and ultimately the grain. Identifying the fact table followed by the dimensions at first was more intuitive given the clear structure of Retail4all's data and the nature of its transactions. This initial focus laid a solid framework to determine the business process and subsequently the grain.

2.4.1 Identifying the Fact

The main business area of Retail4all is sales. The sales transactions are recorded in the sales table, which contains the largest amount of data. It contains valuable data from the retail organisation that can be used and aggregated to answer fundamental business questions. By prioritising the analysis of this table, the data warehouse is aligned with Retail4all's objectives to enable an investigation into the key factors that determine the company's sales performance.

2.4.2 Identifying the Dimensions

Dimensions play a crucial role in supporting the sales data captured in the fact table and define granularity. The identified dimensions regarding the provided datasets are date, product, supplier, staff, store and location. Most of the tables remained unchanged. However, minor changes were applied to the Store, Location and Staff tables. As most of the stores are located in the same city and the location table contained plenty of unneeded rows, we merged these two tables and dropped the region as 50% of the stores are located in just two cities (Aguiar da Beira and Alpiarça). Therefore, we decided to concentrate on specific cities instead of regions. This might seem counterintuitive as it contradicts normalisation by having duplicated values, however this will reduce the execution times and optimizes queries while making the warehouse simpler and easier to understand. Furthermore, we dropped the e-mail column of the staff table, as these won't be necessary for analysis of the sales and security access can also be handled by full name as we do not have duplicated names in the team.

The dimensions follow a tree (branch-leaf) structure organized from detailed to broader levels. For instance, the hierarchy in the date dimension starts at a weekday level and continues to monthly, quarterly, and yearly frequency. Similarly, the product dimension hierarchy begins with the specific product, followed by subcategory and category, allowing for detailed analysis as well as general market insights.

To ensure consistency across the dimensions, we utilize incrementally assigned surrogate keys as primary keys to uniquely identify each entry. This approach also enables the relational mapping within Retail4all's data warehouse. Additionally, the data types and sizes (e.g. NVARCHAR (20)) for each attribute are selected regarding their content while optimizing for query performance. This dimensional design provides a strong foundation for comprehensive and scalable data analysis within the data warehouse.

In our final model all dimensions are connected with the fact table (sales) but not with each other and therefore form a star schema.

2.4.3 Identifying the Business Process

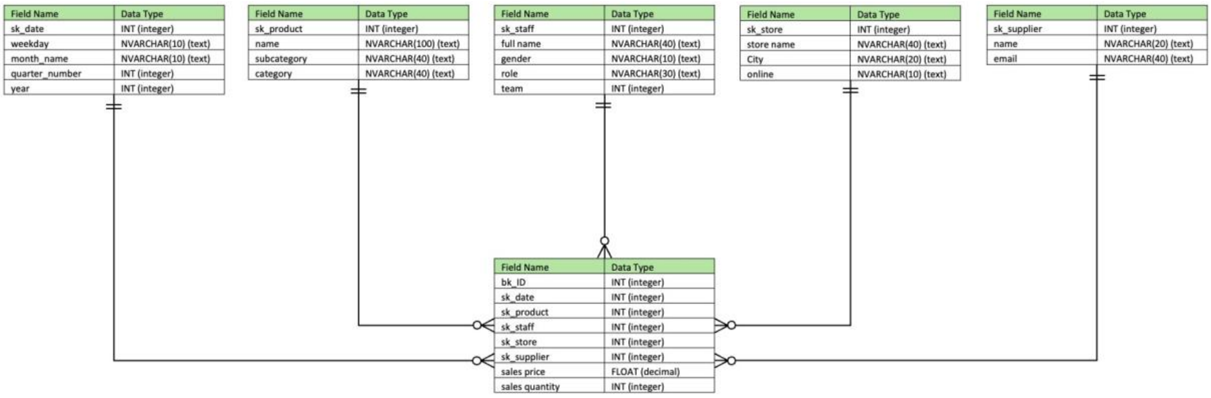
The data warehouse for Retail4all is designed to embed the comprehensive sales process, aiming to measure and analyse the company's performance across multiple dimensions. Fundamentally, the data warehouse seeks to measure sales efficiency, customer engagement and inventory management. Considering this, it will provide insights into monthly, quarterly and yearly sales trends in both volume and revenue. Furthermore, a granular examination of customer buying patterns and preferences will be possible. The data warehouse will provide information to assess the operational efficiency of inventory levels across different categories and locations to optimize stock based on demand trends. By integrating data from multiple sources, a comprehensive perspective of Retail4all's business operations can be provided and therefore facilitate strategic decision-making to enhance resource planning and overall profitability.

2.4.4 Identifying the Grain

The grain in the architected data warehouse represents the meaning of each transaction in the fact table. In the case of Retail4All's designed data warehouse the grain is represented by daily sales per product per store per supplier and per staff. The rows in the fact table show detailed insights about each sale transaction across Portugal.

2.5 Data Warehouse architecture

The designed data warehouse architecture for Retail4all follows a star schema approach and can be seen in Figure 1. This configuration illustrates the connection of each dimension to the central fact table, while maintaining independence from each other. This structure provides a solid foundation to facilitate efficient data retrieval and insightful reporting of Retail4all's business.



Field Name	Data Type
sk_date	INT (integer)
weekday	NVARCHAR(10) (text)
month_name	NVARCHAR(10) (text)
quarter_number	INT (integer)
year	INT (integer)

Field Name	Data Type
sk_product	INT (integer)
name	NVARCHAR(100) (text)
subcategory	NVARCHAR(40) (text)
category	NVARCHAR(40) (text)

Field Name	Data Type
sk_staff	INT (integer)
full name	NVARCHAR(40) (text)
gender	NVARCHAR(10) (text)
role	NVARCHAR(30) (text)
team	INT (integer)

Field Name	Data Type
sk_store	INT (integer)
store name	NVARCHAR(40) (text)
City	NVARCHAR(20) (text)
online	NVARCHAR(10) (text)

Field Name	Data Type
sk_supplier	INT (integer)
name	NVARCHAR(20) (text)
email	NVARCHAR(40) (text)

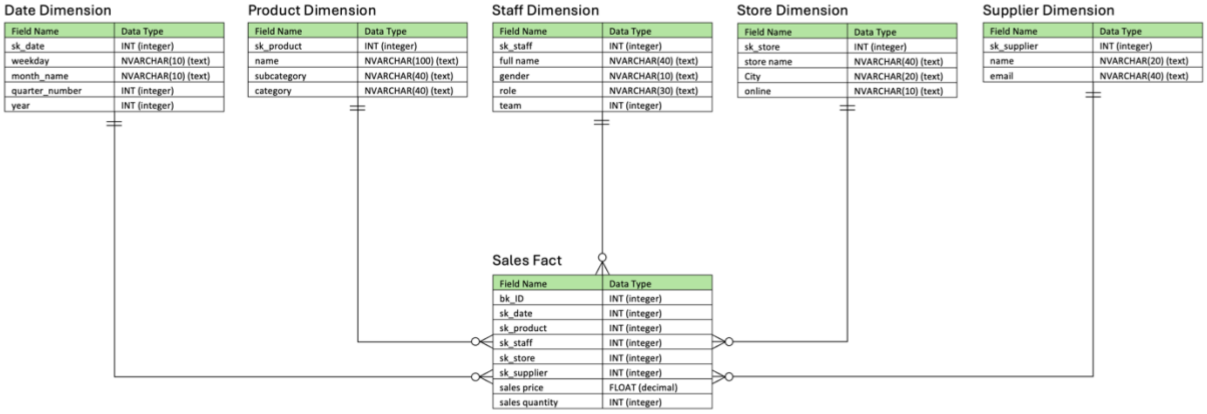
Field Name	Data Type
bk_ID	INT (integer)
sk_date	INT (integer)
sk_product	INT (integer)
sk_staff	INT (integer)
sk_store	INT (integer)
sk_supplier	INT (integer)
sales price	FLOAT (decimal)
sales quantity	INT (integer)

Figure 1: Visualisation of dimensional data warehouse

3 Extract, Transform, Load

3.1 Improvement of Initial Data Warehouse Design

Although the dimensional model was well defined in the first project, we implemented some changes for a smoother ETL process. The initial data warehouse design can be seen in Figure 2.



Date Dimension

Field Name	Data Type
sk_date	INT (integer)
weekday	NVARCHAR(10) (text)
month_name	NVARCHAR(10) (text)
quarter_number	INT (integer)
year	INT (integer)

Product Dimension

Field Name	Data Type
sk_product	INT (integer)
name	NVARCHAR(100) (text)
subcategory	NVARCHAR(40) (text)
category	NVARCHAR(40) (text)

Staff Dimension

Field Name	Data Type
sk_staff	INT (integer)
full name	NVARCHAR(40) (text)
gender	NVARCHAR(10) (text)
role	NVARCHAR(30) (text)
team	INT (integer)

Store Dimension

Field Name	Data Type
sk_store	INT (integer)
store name	NVARCHAR(40) (text)
City	NVARCHAR(20) (text)
online	NVARCHAR(10) (text)

Supplier Dimension

Field Name	Data Type
sk_supplier	INT (integer)
name	NVARCHAR(20) (text)
email	NVARCHAR(40) (text)

Sales Fact

Field Name	Data Type
bk_ID	INT (integer)
sk_date	INT (integer)
sk_product	INT (integer)
sk_staff	INT (integer)
sk_store	INT (integer)
sk_supplier	INT (integer)
sales price	FLOAT (decimal)
sales quantity	INT (integer)

Figure 2: Initial Dimensional Data Warehouse Architecture

Specifically, we adapted the dimensions date and supplier, minor changes were applied to the product, staff, store and sales tables.

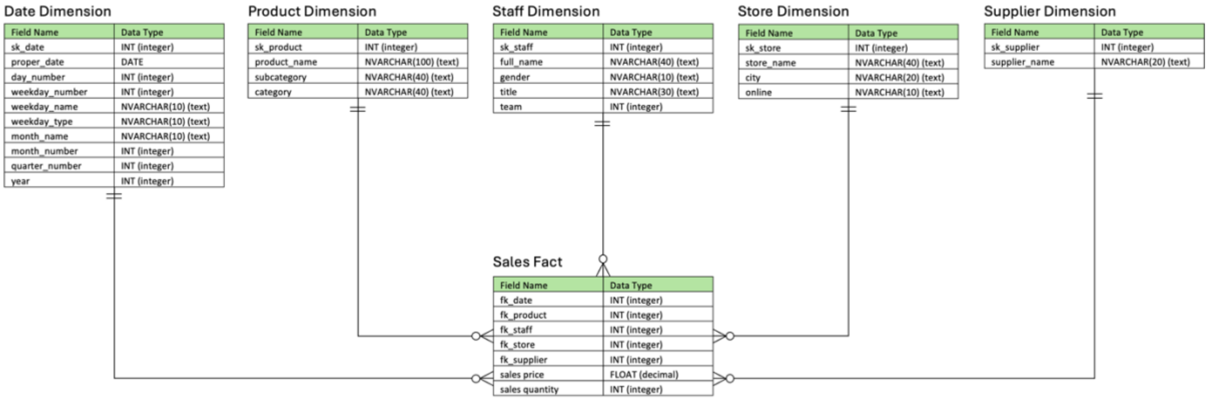
More columns were added to the date dimension for easier analytics and training of machine learning models. We included the numbers of the weekday (weekday_number) and the day in the month (day_number) for future operations where the data type is needed, i.e. the day of the month is a number between 1 and 31 while the weekdays contain numbers between 0 and 6. Also, we included the actual date (proper_date) in a datetime format to be able to work with it later to create needed features. Additionally, we added the weekday name (weekday_name), the weekday type (weekday_type) and the number of the month (month_number). We could include more columns in the date table, but since the actual date is already present and we believe the most important features are already included, we decided to keep it simple. Potential further date columns can be found in Table 3. Additionally, we can always derive more features from the existing 'date' column later. This approach helps us save storage space and enhances query speed.

A few minor changes included:

- The name of the product was renamed to product_name for clarity.
- The column name ‘role’ was changed to ‘title’ in the staff table, as role is a keyword in sql and can cause errors.
- The column ‘city’ in the store table was changed to be in all lower letters for more coherence.
- Deleted the e-mail column in the supplier table as it is not needed for distinction of different suppliers and changed the description of the column name to ‘supplier_name’ for better understanding.

- Deleted the column bk_ID in the fact table as the primary key is now the combination of all the sk of the dimensions.

Implementing all the mentioned changes, the semantic model of the dimensional data warehouse can be seen in Figure 3.



Date Dimension

Field Name	Data Type
sk_date	INT (integer)
proper_date	DATE
day_number	INT (integer)
weekday_number	INT (integer)
weekday_name	NVARCHAR(10) (text)
weekday_type	NVARCHAR(10) (text)
month_name	NVARCHAR(10) (text)
month_number	INT (integer)
quarter_number	INT (integer)
year	INT (integer)

Product Dimension

Field Name	Data Type
sk_product	INT (integer)
product_name	NVARCHAR(100) (text)
subcategory	NVARCHAR(40) (text)
category	NVARCHAR(40) (text)

Staff Dimension

Field Name	Data Type
sk_staff	INT (integer)
full_name	NVARCHAR(40) (text)
gender	NVARCHAR(10) (text)
title	NVARCHAR(30) (text)
team	INT (integer)

Store Dimension

Field Name	Data Type
sk_store	INT (integer)
store_name	NVARCHAR(40) (text)
city	NVARCHAR(20) (text)
online	NVARCHAR(10) (text)

Supplier Dimension

Field Name	Data Type
sk_supplier	INT (integer)
supplier_name	NVARCHAR(20) (text)

Sales Fact

Field Name	Data Type
fk_date	INT (integer)
fk_product	INT (integer)
fk_staff	INT (integer)
fk_store	INT (integer)
fk_supplier	INT (integer)
sales price	FLOAT (decimal)
sales quantity	INT (integer)

Figure 3: Final Dimensional Data Warehouse Architecture

3.2 Set-Up for E.T.L.

The set-up for the extraction, transformation and loading of the data consists of four major parts. In the extraction part, data is selected and read from the csv files. The transformation converts the data from the original state into whatever form the data warehouse needs, including data engineering. Lastly the transformed data is loaded into the warehouse. This generic process is implemented with Microsoft Fabric and is included in more detail four major instances:

First, the **Lakehouse** named LH_Retail4all got created and populated with the original data. The given csv files are stored in here in its original states without any data engineering steps applied. Second, the **Dataflows** which apply all the data engineering steps to clean the data, removing as many errors as possible and transform the data. These dataflows will be described in more detail in chapter 3.3. Third, the **Data Warehouse** named DW_Retail4all was created to store the data in such a manner so data retrieval is optimised, and machine learning models can use the data to make predictions or cluster the customers. It is created using SQL queries but without any data in it. In its original state it just represents an empty architecture which is ready for data to be load into. Lastly, the **Pipeline** which combines all instances and creates a coherent E.T.L. process. We choose to implement a full load which ensures that the most up-to-date data is available for analysis, while maintaining data integrity. The pipeline can be seen in more detail in Figure 4.

The data pipeline, named 'DW_LOAD_RETAIL4ALL', was created to consolidate and automate all previously created dataflows. Initially, dimensions and fact tables within the existing data warehouse were cleaned up to ensure the insertion of the latest data, mitigating any potential inconsistencies or duplications that may have occurred since the last update. This full load approach maintains data integrity throughout the data warehouse. All the dataflows developed for the dimensions were then integrated into the pipeline. These dataflows utilise data from the Lakehouse and perform transformations which is then loaded into the data warehouse to ensure data consistency and accuracy. Finally, the sales fact dataflow completes the data pipeline, resulting in a comprehensive and streamlined data integration process for the finalized data warehouse from Retail4all.

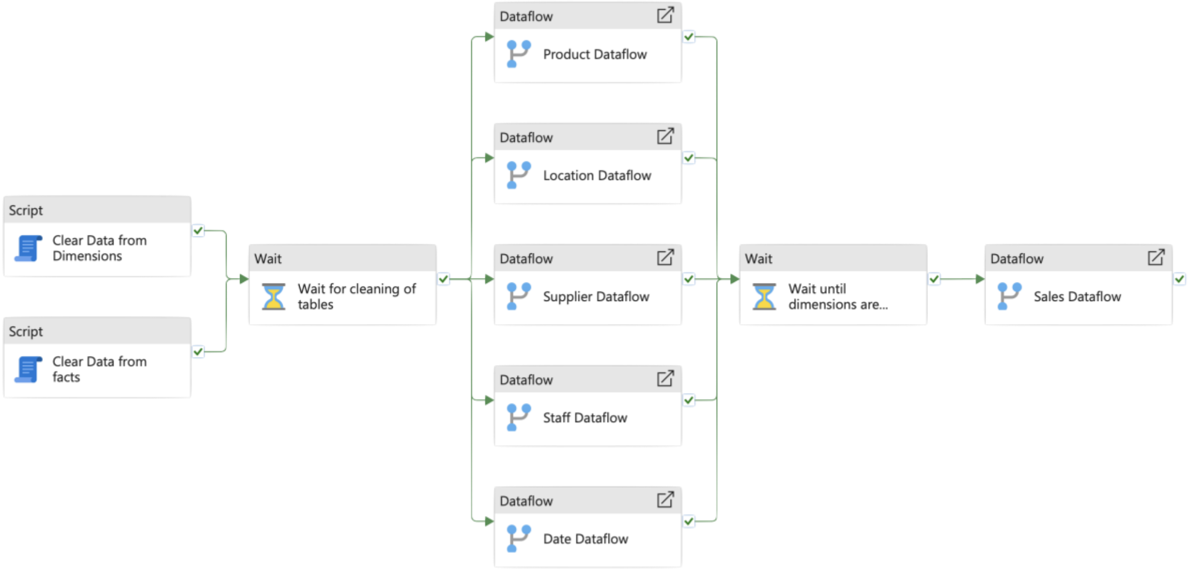


Figure 4: Pipeline for ETL Process

3.3 Data Engineering

Beginning with the products table, we created a new dataflow using data from the Data Lakehouse. The first row is used for the column headers, excluding the ‘sku’ column as it will not be used. We replaced the initial business key with a continuous index to convert it into a simpler surrogate key. Lastly the data warehouse was selected as the destination. Throughout all dimensions and fact tables the Data Lakehouse represents the source, while the data warehouse is the destination for the transformed data.

The store dataflow is one of the more complex ones. First the tables ‘Stores A’ and ‘Stores B’ needed to be appended. As those tables have the same structure the data of one table could easily be combined. Next the stores table was merged using a left join with the location table to have the stores combined with their respective locations. As some words showed errors, most likely to Portuguese apostrophes, the values needed to be manually replaced, e.g. ‘Farmácia da Alegria’ should be written as ‘Farmacia da Alegria’ for the data warehouse to successfully process it. Finally, the column indicating whether a shop had already implemented online shopping had to be addressed as it contained Y, S, N and empty values. We concluded since all the stores are in Portugal that Y stands for ‘yes’ and S for ‘sim’ and N for ‘nao’. Accordingly, we will change the entries to ‘Y’ and ‘N’. The remaining missing entries will be kept as missing entries as we have no further information for these stores and keeping them as missing values will support future analysis.

In the dataflow for the supplier table only the email column was deleted, as mentioned earlier in chapter 3.

For the staff table we created a dataflow which concatenated the first and last name to full name and corrected one instance of the gender column from ‘Agender’ to ‘Others’.

Next we created the dataflow for the date table. Sales data extended from January 1, 2020, to June 2, 2023. To cover the period up to 2024, the dates were generated spanning from 2020 to 2024, to be able to fill the data warehouse with transactions up to today. This data was transformed into datetime format, and the following columns were derived as discussed in chapter 3: day_number, weekday_number, weekday_name, weekday_type, month_number, month_name, quarter_number, year, and the sk_date, representing the date as an integer, e.g. 20200101. Conditional statements were used to create the column weekday_type, e.g. if the weekday number is equal to or below 4 it counts as weekday and else as weekend (Monday counts as 0). As with previous dataflows, the data was connected to the Data Warehouse.

Ultimately we created the dataflow for the sales table, involving the removal of the ‘Sale ID’ column as all foreign keys would contribute to constructing the primary key. Referencing foreign keys from the dimensions, particularly for supplier, staff, and localization, was straightforward due to their consistent structure. However, a challenge arose with the product dimension, where the business key was eliminated, and an incremental surrogate key was employed for simplicity. When merging the product keys with the sales data, we needed to make sure that each product in the sales data matched up with its corresponding surrogate key in the product dimension. So, we replaced the original product identifiers in the sales data with their matching surrogate keys. This ensures that our sales data can properly connect to the product dimension using these surrogate keys, which act as foreign keys. Furthermore, sales data was joined with the date dimension table based on the date, utilizing the surrogate key from the dimension. The date column was subsequently removed from the table. Additionally, the ‘operator’ column, for which its purpose was unclear, was deleted, as we already include ‘operator_id’. The ‘location’ column was also omitted since location and store information had been merged, with a preference for retaining store details due to their higher level of detail. We decided not to calculate a 'total amount' column in the data warehouse. Instead, we'll do this in a later step within the semantic model. This approach ensures consistency across all reports, optimises performance by reducing the flow of data in the dataflows, and improves scalability by centralising calculations.

Overall, we checked all dimensions for duplicates and detected three products in the product dimension table which were duplicated, namely ‘Keurig K-Mini Plus Single-Serve Coffee Maker’, ‘Timberland Earthkeepers 6-Inch Boots’ and ‘UGG Classic Mini II Boots’. Additionally, more than 90.000 duplicates were found in the sales fact table. These duplicates have the same sales ID and time of purchase and therefore cannot be unique transactions. All occurrence were handled the same way in the dataflow: keep the first one and delete the following.

4 Reporting

Once the data warehouse has been set up, we can now begin the reporting phase for Retail4all. This critical step involves utilising the data architecture to generate insightful and actionable business intelligence. With the implementation of an enhanced semantic model, we aim to create a comprehensive suite of reports that will cover every aspect of Retail4all's operations. Our goal is to equip Retail4all with the tools necessary for in-depth analysis, enabling them to make informed decisions that drive success and growth.

4.1 Model optimization

After inspecting the data, an outlier was detected in the sales fact table, characterized by an abnormally high sales price of 9,000,005 €. This transaction was removed from the dataset, prompting a retriggering of the data pipeline to ensure the data warehouse reflects the most accurate and current data.

The optimized data warehouse will be used to create a **semantic model** called SM Sales. Within this model all tables from the warehouse will be leveraged to efficiently provide data for the following visualizations build with Power BI. Furthermore, the semantic model will inhabit the relationships between the dimensions and facts as shown priorly in the visualization of the architecture of the data warehouse. All dimensions have a one-to-many relationship to the sales fact table meaning that each record of the fact table corresponds to only one record of the associated dimension.

Adjustments of columns and tables: To enhance the understandability and usability of the database for non-technical users, the column names have been revised to be more user-friendly. For example, the column 'weekday_type' has been renamed to 'Type of Weekday'. Additionally, columns that are not necessary for the visualizations, such as surrogate keys and foreign keys from the dimensions and fact tables, have been hidden to not further irritate the analyst. This approach has also been applied to the naming of the tables, for instance 'dim_date' has been updated to 'Dim Date'.

Data Types: Further refinements have been implemented within the semantic model concerning data types to improve readability and data interpretation. For instance, 'Sales Price' is set as a currency and utilizes a decimal data type with a delimiter for thousands. Similarly, the 'Quantity of Sales' is represented in whole numbers. Further measures which were calculated during the process were also formatted accordingly. For instance, the ‘Sales Amount’ is similarly set to a currency with a delimiter for thousands, while the measure ‘Sales YoY Change’ will be adjusted to a percentage format.

In the date dimension, the columns have been strategically organized to reflect a logical sequence. For example, the 'Name of the Month' is ordered by the 'Number of the Month'. Similarly, the 'Weekday' column is sorted by 'Weekday Number' to enable the construction of reasonable visualizations with the usage of text columns.

Hierarchies: Hierarchies have been constructed within the semantic model to facilitate the creation of effective visualizations and to enable advanced functionalities such as drilldowns, enhancing the exploration of Retail4all's data. The established hierarchies are as follows:

- **Dim Date**: Year – Quarter – Day of Month – Weekday
- **Dim Product**: Product Category – Product Subcategory – Product
- **Dim Store**: Store City – Store Online - Store
- **Dim Staff**: Staff Team – Staff Title – Staff Gender – Staff Full Name

After the hierarchy setup, the initial columns in the semantic model were hidden to prevent the display of duplicate columns within Power BI.

4.2 Calculated Measures

Calculated columns were priorly implemented in the E.T.L process as for instance the full name in the staff dimension or time related columns like weekday number or month name in the date dimension. Furthermore, measures were calculated particularly in the semantic model to effectively answer Retail4all's business questions. An essential measure utilized is the 'Sales Amount' which is calculated using the following formula:

```
Sales Amount = SUMX('Fact Sales', 'Fact Sales'[Sales Price]*'Fact Sales'[Sales Quantity])
```

This measure quantifies the total revenue generated from all sales transactions, serving as a critical key performance indicator of the sales performance for Retail4all.

```
Top Product = FIRSTNONBLANK(TOPN( 1 ,VALUES( 'Dim Product'[Product] ) , [Sales Amount]) , 1)
```

The 'Top Product' measure identifies the product with the highest total sales revenue, reflecting its importance as an indicator of the most economically impactful product within Retail4all's offerings.

```
Best Store = FIRSTNONBLANK(TOPN( 1 ,VALUES( 'Dim Store'[Store] ) , [Sales Amount]) , 1)
```

Subsequently, the 'Best Store' identifies the store with the highest total sales revenue, which serves as a key indicator of the most profitable store in the Retail4all network.

```
Average Sales per Store = DIVIDE('Fact Sales'[Sales Amount], Count('Dim Store'[Store]), 0)
```

The 'Average Sales per Store' calculates the average revenue per store by dividing the total sales revenue by the number of stores, providing an important metric of overall sales efficiency on the dashboard's homepage.

```
Average Sales Price = AVERAGE('Fact Sales'[Sales Price])
```

The 'Average Sales Price' measure calculates the mean sales price across all transactions, adjusting dynamically to reflect the selected filters. This metric is crucial for analyzing pricing trends across different product categories, subcategories, or individual products within Power BI.

```
LY Sales Amount = CALCULATE('Fact Sales'[Sales Amount], SAMEPERIODLASTYEAR(DATESBETWEEN('Dim Date'[Date], DATE(2023, 1, 1), CALCULATE(MAX('Dim Date'[Date])))))
```

The ‘LY Sales Amount’ represents the revenue generated in the entire year of 2022. The measure is used to effectively compare the trend of the previous year with the current one.

```
Sales Revenue Goal = [LY Sales Amount] * 1.2
```

As a key performance indicator, the 'LY Sales Amount' was utilized to represent the 'Sales Revenue Goal,' which aims for a 20% increase over the previous year's revenue. This metric serves as a benchmark to evaluate the company's performance and growth trajectory.

Additionally, the 'Sales Previous YTD' and 'Sales Current YTD' measures were calculated to represent the revenue accrued within the same timeframe for the previous and current year, respectively. To ensure a consistent comparison, since the data for the current year is available only up until June, the 'Sales Previous YTD' specifically represents the revenue accumulated up to that same point in the previous year. These measures were utilized as follows:

```
Sales YoY Change = DIVIDE([Sales Current YTD] - [Sales Previous YTD],[Sales Previous YTD], ‘calculation not possible’)
```

The 'Sales YoY Change' measure is calculated as a year-over-year percentage and provides insight into revenue growth by comparing current year-to-date sales against those from the corresponding period in the previous year. This indicator is crucial for assessing the company's performance and growth over time.

```
Arrow = IF([Sales YoY Change] > 0, [Arrow Up], IF([Sales YoY Change] < 0, [Arrow Down], ‘No data of last year’)
```

The arrow in the staff performance dashboard was implemented by encoding the pictures into a base64 code. After converting Power Bi will be able to read the arrows as a picture but in text format. The arrows in text format were then stored in the variables ‘Arrow Up’ and ‘Arrow Down’. The variable arrow staff combined these two pictures and shows the arrow up when the revenue growth YoY is positive and vice versa. To implement this, we used the add-in ‘Image Pro by CloudScope (3.0.0.1)’.

Some locations included in the semantic model do not exhibit any data. This also applies to some staff names and specific stores. Whenever this kind of data was used in visualizations or in slicers, we manually filtered it so that the dashboard provides a smooth experience.

4.3 Technical Aspects of the Report

4.3.1 Configuration of the Report

The report was built by utilizing the optimized semantic model within Power BI. It was designed particularly for Retail4all and is based on the prior explained business challenges and business questions. The foundation led to the construction of a report which is composed of five different dashboards and a homepage.

The **homepage** serves as the central hub for Retail4all's report, offering an overview of its five distinct sections, i.e. **time analysis**, **store comparison**, **store performance**, **staff performance** and **product analysis**. It facilitates navigation between these sections and features two visualizations to provide a broad overview. The first chart illustrates the trend of sales in Retail4all's current year 2023, including a forecast for the next 30 days. The forecast takes yearly seasonality into account and displays a confidence level of 95%. The other visualization shows the distribution of revenue by store locations across Portugal and can be grouped into local group using the lasso tool.

The **time analysis** page features four visualizations that provide insights into revenue trends across various timeframes. The charts can be filtered by year and product category. It includes a display of revenue and sales quantity over the dataset's entire span. This visualization excludes a single transaction dated June 2nd to prevent a misleading drop in June's data. Users can adjust the charts to monthly or yearly granularity for a broader perspective. Additionally, the page displays cumulative monthly revenue and sales quantity. To present a consistent monthly trend without distortions, data from 2023 was excluded from this chart to avoid showing a significant drop after May. Additionally, the revenue by weekday is illustrated to identify weekly trends. The tachometer visualization represents the current year's revenue progress against two thresholds: last year's total revenue and a second threshold set at 20% above the last year's revenue.

The **store comparison** page provides a comprehensive analysis of revenue across various store locations through a series of interactive visualizations, enhancing strategic decision-making. Users can tailor the data presentation by applying filters for product category and year. The page features a geographical map that color-codes store locations in Portugal to reflect different revenue levels. When hovering over a city, the number of stores within that location such as their contribution to the overall revenue in the location can be seen through a tree map. Additionally, a bar chart ranks the top five stores by sales amount, enabling a straightforward comparison of the highest revenue generators. There is also a sparkline chart that tracks sales trends over different locations, helping to monitor changes in performance across time. Completing the set of tools, a tree map categorizes stores within each city, using color variations to indicate the revenue contribution of each store, which aids in quickly identifying which locations are key contributors to overall sales.

The **store performance** page provides a detailed examination of performance metrics for individual stores, offering insights into both revenue generation and staff contributions. For each selected store, the page displays key indicators such as the city location, year-over-year performance percentage, and total revenue. The dashboard includes the staff revenue contribution, which breaks down revenue by individual staff members, categorized by their roles such as sales staff, floor manager, and director, allowing for easy comparison of individual performance. Similarly, the product revenue contribution can be seen, when pressing the corresponding button. Additionally, a line chart titled 'Current Performance' tracks monthly revenue trends over the current and previous year, offering a visual representation of sales development per store. Finally, the 'Operator Performance' section showcases revenue trends for the store's staff over time, presented in sparkline charts.

The **staff performance** page delivers an in-depth analysis of selected staff member's performance to overall sales, highlighting essential metrics that reflect the employee's role, year-over-year revenue growth, and total revenue contribution. This data is crucial for assessing the direct impact of staff efforts on the company's financial health. The dashboard includes a revenue over time chart that illustrates monthly sales data for the featured staff member. Additionally, the revenue by product category section utilizes a pie chart to display how the staff member's sales are distributed among different categories such as Electronics, Clothing & Accessories, Home & Kitchen, and Sports & Fitness. Additionally, the page includes a matrix that displays the absolute revenue values such as the year-over-year revenue growth per category for the selected staff member, while allowing for a drill-down into specific products when clicking on the bars.

A detailed overview of sales data regarding products and supplier interactions is provided in the **product analysis** page which can be filtered by year, city and the presence of online shopping possibility. The first visualization combines bar and line graphs to illustrate sales volumes and average sales prices across different product categories. This dual approach not only highlights which categories generate the most revenue but also tracks pricing trends, offering insights into profitability and pricing strategies. A drilldown enables to inspect these insights on a more granular level for product subcategories and products. Next, a horizontal bar chart details revenue contributions from various suppliers, categorized by sales quantities. This visualization is crucial for assessing the dependency of the suppliers for different products. Additionally, a concise list of the three top-selling products is displayed, providing immediate insights into consumer demand and product popularity.

4.3.2 Technical Elements of the Report

For easier navigation between the individual report pages we implemented several page navigators. Each page includes a button to return to the homepage such as a reset button for the filters to facilitate the user's experience. Furthermore, every page includes a sidebar on the top right to navigate to the desired dashboard. These navigation tools were implemented using bookmarks, making the experience for the user more guided and comfortable.

Some visualizations include an information icon which display more detailed information such as explanations when hovering over it. Additionally, a tooltip tree map was created to effectively display the number of stores in a specific city such as the contribution of the sales revenue by store. This was done by creating a new tooltip page. The page was filled with a basic tree map which is clustered into cities. On top of the tree map two KPI were placed to display the number of stores in a specific city and the name of that city.

4.4 Analysis and Discussion

During the analysis of the previously mentioned business questions, certain insights emerged that necessitated slight modifications to the questions due to content considerations. As a result, minor adjustments were made, and 3 questions were removed to ensure a more effective response to the business challenges facing Retail4all. In the following these questions will be discussed in detail.

4.4.1 Time-Specific Sales Trends Analysis

What are the weekly, monthly, and yearly sales trends by revenue?

Report Page 1 - Time reveals that the typical revenue trend is upward. Retail4all's revenue started with approximately 1 M€ in January 2020 and increased up to almost 1,7 M€ in May 2023. The sales quantity similarly ascended from ca. 70,000 to 110,000 per month. Revenue and sales quantity tends to spike in January and subsequently decline in February. Generally, the differences between months don't vary significantly and range from 3,4 to 3,7 M€ of revenue accumulated per month over the years. When looking at weekly trends, for the year 2023 for instance, one can observe that Tuesdays, Wednesdays and Thursdays the most revenue is generated, while Fridays and Saturdays the revenue is the least.

How does the current year's sales performance compare to the previous year?

As observed in Report Page 1 - Time , the revenue for the current year 2023 reached about 8,28 M€ and therefore almost 50% of the previous year's revenue was accomplished, despite only being in May. Furthermore, selecting the years 2022 and 2023 for the sales revenue over time shows that since the beginning of 2023 the revenue exceeds the average of 1,46 M€, effectively displaying the positive development of the current year compared to the previous. Additionally, Figure 7 exhibits consistently higher revenues in the same months than in the previous year such as positive year-on-year growth ranging from ca. 14% to 25% depending on the selected store. Consistent positive growth can be seen when exploring the year-on-year revenue growth per staff member as seen in Report Page 4 – Product Analysis.

Which locations have shown the most significant growth in sales over the past year?

The matrix in Report Page 2 - Store Comparison exhibits that the location Barcelos experienced the highest revenue growth closely followed by Alpiarca and Carrazeda de Ansiaes. All these locations grew more than 20% compared to the last year at the same time.

What are weekend trends?

There are no significant weekend trends when looking at all years combined. The stores' revenue tend to be quite stable across all weekdays, as shown in Report Page 1 - Time The day with the lowest revenue is Monday, however the distance between the best performing day and the least is only 0.7%. When filtering for the current year 2023, more fluctuations can be witnessed. Fridays and Saturdays tend to contribute to less revenue while the performance stabilizes on Sundays.

What is the average amount of money spent in each store?

The average revenue for a single store is 2,25M€, as stated on the homepage of the report. While there may be some variation between stores, the standard deviation is relatively low. This can be seen when exploring the revenue for different stores in the Store Performance page.

What is the distribution of sales revenue for each product category by the day of the week, identifying potential peak sales days?

When filtering for specific product categories the dashboard in Report Page 1 - Time shows that clothing has its peak sales on Tuesday and Sunday, while other product categories do not have distinct weekdays in terms of peak revenue.

4.4.2 Transaction Analysis

How does the revenue vary across different cities?

Aguir da Beira makes up most of Retail4all's revenue with around 30,6M of sales over the years followed by Alpiarca with around 4,8M as seen in Report Page 2 - Store Comparison. The city with the lowest contributing revenue is Alvaiazere with 2,3M sales revenue. The symbol map gives an overview of the citys' revenue distribution throughout the country of Portugal. The significantly high share of Aguir da Beira is due to the relatively high quantity of stores (13) located in this city, which can be seen in more detail when hovering over the location.

How does the existence of an online shop influence the sales of certain product categories or products?

The Product Analysis page allows to investigate this question. In all years and stores, customers tend to purchase cheaper sports and fitness products in-store and more expensive ones online. A similar pattern can be observed in the electronics segment. However, it is notable that the total number of products sold offline exceeds those sold online, across all product categories. Exploring the categories in more depth by using the drill-down feature illustrates that the subcategories show some variations. For instance, gaming accessories make up most of the sales for offline shops while smart watches lead for online stores. The most sold product in offline shops is the Hyper X Pulsefire Dart Wireless Gaming Mouse with a sales quantity of ca. 22,000, while the most sold product online is represented by the Samsung Galaxy Watch Active 2 with around 14,000 sales.

4.4.3 Performance Metrics

Which stores are the top performers in terms of revenue?

The top five stores, in descending order, are Farmacia da Alegria, Tech Wizards, Bits & Bytes, TecnoLoja and Code Cave, represented in Report Page 2 - Store Comparison. All of these stores are situated in Alpiarca or Aguiar de Beira. The highest-selling store generated a total revenue of 2.44 M€, while the fifth-best store achieved a revenue of 2.39 M€, representing a deviation of approximately 2% from the best-selling store. The revenue of the stores is very similar on average.

Which Point of Supply (POS) has the highest number of transactions for top-selling products?

When exploring the most sold products, by clicking on these products in the matrix, in Report Page 4 – Product Analysis, it is evident that the distribution of transaction volumes is approximately the same for different suppliers. Given the similar distribution of transactions, it can be inferred that the differences in sales amount are due to the higher frequency of deliveries by one supplier. Consequently, Lazzy has the highest number of transactions for top-selling products, followed by Kwimbee, Realcube, Browsetype, Eire, Devbug and Wordpedia, respectively.

4.4.4 Product Insights

What were the top products sold in each year or quarter?

There is no repetition of products, indicating that there is no long-lasting trend in the top products on a yearly basis, represented in Report Page 4 – Product Analysis. Given this insight, we have decided not to further analysis sales patterns in this context and therefore omitted the analysis of top products by quarter.

Table 1: Top Products by Year

	Top 1	Top 2	Top 3
2020	Apple AirPods Max	Apple iPad Pro 12.9-inch	Corsair K100 RGB Optical Mechanical Gaming Keyboard
2021	Bose QuietComfort 35 II Headphones	KitchenAid Artisan Series 5KSM125 Standard Mixer	Sony WH-1000XM4 Noise Cancelling Headphones
2022	Cuisinart MultiCulinary Center 11-in-1 Food Processor	HyperX Pulsefire Dart Wireless Gaming Mouse	SteelSeries Rival 650 Wireless Gaming Mouse
2023	HyperX Pulsefire Dart Wireless Gaming Mouse	Razer Basilisk Ultimate Wireless Gaming Mouse	Samsung Galaxy Watch Active 2

What are the sales patterns for high-end vs. low-end products in terms of total revenue?

According to Report Page 4 – Product Analysis high-end products are mainly from the clothing & accessories as well as from the sports & fitness categories. Home & kitchen tends to be the cheapest dimension. The total revenue doesn’t show a clear pattern when differentiating high-end or low-end products.

Smart Watches, Speakers and Gaming Consoles seem to be the most expensive products in the Electronics category, while Business Laptops show the lowest average price in this category. Smart Watches make up a high portion of the total revenue while Business Laptops only contribute to a small share.

The high-end products in clothing & accessories are mainly sneakers, chinos and jeans in ascending order. While there is not a clear pattern in terms low-end vs high-end products, jeans tend to be sold the most of this product category followed by boots, sneakers, bottoms, chinos and ultimately tops.

In the sport & fitness category the most expensive products are Under Armour HOVR Machina 3 and Reebok Floatride Energy 4. Despite having big differences in price all sport & fitness items are approximately being sold in the same amount.

Which product categories are most popular in specific cities?

Through filtering for specific cities in Report Page 4 – Product Analysis we can identify that all cities have the same order of the most popular product categories, namely electronics, clothing & accessories, home & kitchen and sports & fitness in ascending order.

Are there products which are ordered more online than in the store in relative terms?

While all product categories are more purchased in store than ordered online, the order of the most sold products differs between online and in-store purchases. For instance, in the electronics category, the most purchased sub-category in-store is gaming accessories, followed by smart watches. However, the order of the most purchased products online is reversed. This example becomes even more extreme with sport and fitness shoes. In-store, the most popular product from this category is the Under Armour HOVR Machine 3 sneaker. For online orders, this is the least popular sneaker. This should be taken into consideration when designing the product portfolio for the website and the stores.

4.4.5 Operator Influence

Who were the top operators across different stores?

Given that there are 23 different stores, we focused our attention on the top three operators for the five most popular stores in terms of revenue. Table 2 provides a summary of the top operators. It is evident that Stern Burgyn and Sibyl Scintsbury are the employees who consistently sell the most. Mr. Burgyn holds the position of store manager, and Mr. Scintsbury serves as floor manager. Consequently, the higher the position an employee occupies, the greater the revenue generated for the stores. When we consider the entire team, however, we see that the sales staff generates the most revenue. This could be because this is the largest team.

Table 2: Top Performing Operators

Location	Best	Second	Third
Farmacia da Alegria	Stern Burgyn	Sibyl Scintsbury	Nap Cavee
Tech Wizards	Stern Burgyn	Sibyl Scintsbury	Nap Cavee
Bits & Bytes	Stern Burgyn	Sibyl Scintsbury	West Adriano
TecnoLoja	Stern Burgyn	Sibyl Scintsbury	West Adriano
Code Cave	Stern Burgyn	Sibyl Scintsbury	West Adriano

Which single operators have the highest sales in terms of revenue for each product category?

When filtering the composition tree for the product categories in the report page of Report Page 3 - Store Performance, it is possible to analyse which operator contributed the most to the filtered product category. Please note that in order for the filters to be applied, it is necessary to click on the **category name** in the decomposition tree, otherwise the filters will not be applied. In the case of electronic products, the most significant contributions are made by Stern Burgyn, Sibyl Scintsbury and West Adriano. In the clothing & accessories categories, Mr. Burgyn is the top contributor, followed by Mr. Scintsbury and Mr. Cavee. In the case of home & kitchen appliances, the order is once again Mr. Burgyn first, Mr. Adriano second and lastly Mr. Scintsbury. The same order as in home & kitchen appliances applies to the sports & fitness products.

Which operator produces the highest revenue for the business in terms of the title?

As previously stated, there are three employees who consistently demonstrate superior performance. Analysing the employees composition tree in Report Page 3 - Store Performance an interesting characteristic can be observed. The sales team is the primary contributor to the company’s revenue, followed by the floor staff, the director, the floor manager and finally the floor checkout manager. The latter three roles are filled by only three employees each. Consequently, a more contextual view is to consider the actual revenue contribution per person, as previously analysed.

5 Conclusion

In conclusion, the Retail4all data warehouse was designed using the Kimball methodology, providing a comprehensive framework for analyzing sales data across multiple dimensions. Each step in this project was implemented considering Retail4all's strategic, tactical, and operational needs, enhancing sales performance, inventory management, market responsiveness, and customer experience.

Several enhancements were made to the original dimensional model to streamline the ETL (Extract, Transform, Load) process and improve data consistency. Adjustments included adding new columns to the date dimension for more detailed analysis and optimizing the data engineering tasks for each dimension and the fact table.

The ETL setup involved four main components: the Lakehouse for storing original data, dataflows for data engineering, the optimized data warehouse for data retrieval, and the pipeline that automates the ETL process. This pipeline utilizes a full load approach to ensure the most up-to-date and accurate data is available for analysis.

In the reporting phase, necessary enhancements to the data models and ETL processes have significantly strengthened the analytical capabilities. These improvements facilitated the development of an optimized semantic model, which was refined through strategic formatting of data types, the construction of meaningful hierarchies, and the concealment of non-essential columns.

Subsequent to these foundational enhancements, various measures and calculated columns were implemented to support dynamic and effective visualizations within Retail4all's Power BI report. This report was designed to address specific business questions, enhancing user interaction through navigational buttons and advanced drill-down capabilities. These features provide both a broad overview and detailed insights into Retail4all’s sales performance across different operational aspects.

The report is structured around a homepage and five distinct sections — **time analysis, store comparison, store performance, staff performance,** and **product analysis** — offering a comprehensive examination of the data. The analysis revealed that Retail4all has consistently maintained a positive sales trend over recent years. This is largely attributed to the increase in revenue observed in January, which was followed by a sideways trend for the rest of the year.

With these steps, Retail4all has established a robust data infrastructure that not only showcases the sales performance of its stores but also secures a competitive advantage by enabling in-depth analytical assessments. This strategic enhancement empowers Retail4all to maintain and expand its market leadership through informed decision-making and operational excellence.

6 Bibliography

Lindecrantz, E., Tjon Pian Gi, M., & Zerbi, S. (2020, April 28). *Personalizing the customer experience: Driving differentiation in retail*. Retrieved March 21, 2024, from McKinsey & Company: https://www.mckinsey.com/industries/retail/our-insights/personalizing-the-customer-experience-driving-differentiation-in-retail

7 Appendix

Table 3: Additional Potential Date Columns

Column Name	Data Type	Example Value
full_date	Varchar(50)	13 of January 2023
weekday_name_short	Varchar(10)	Mon, Tue, Wed
is_special	Varchar(30)	Christmas
month_name_short	Varchar(10)	Jan, Feb, Mar
quarter_name	Varchar(50)	First, Second, Third
semester_number	Int	1, 2
semester_name_short	Varchar(10)	Sum, Win
semester_name	Varchar(50)	Summer, Winter

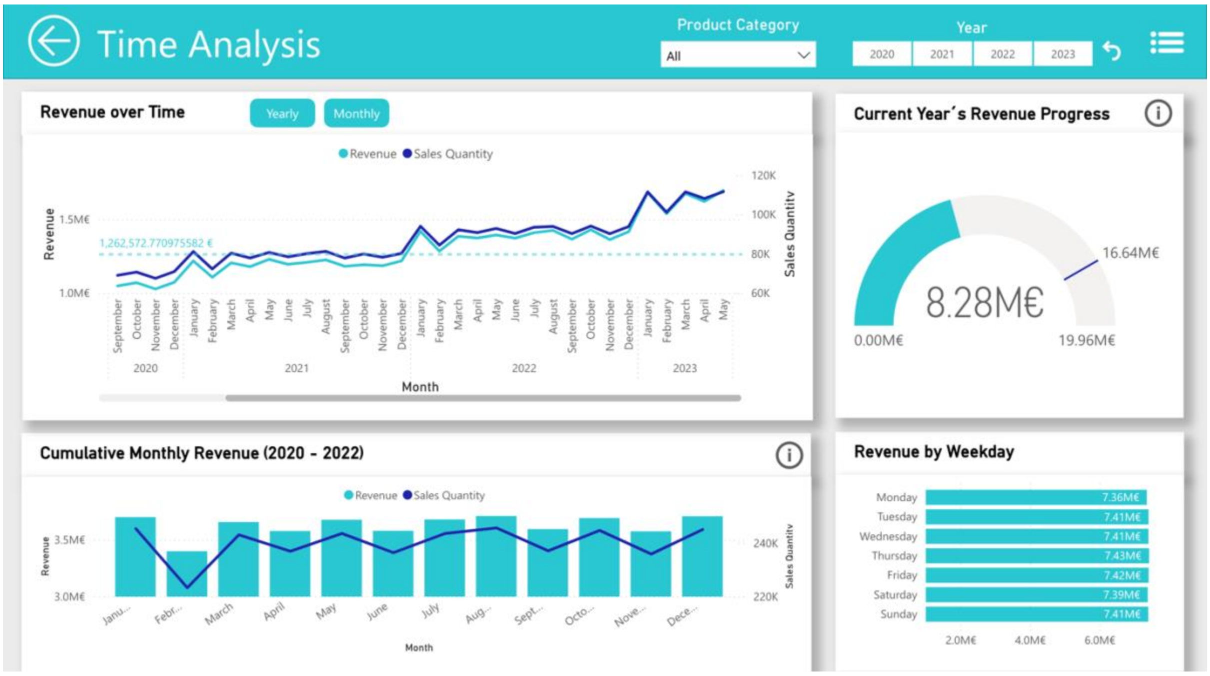
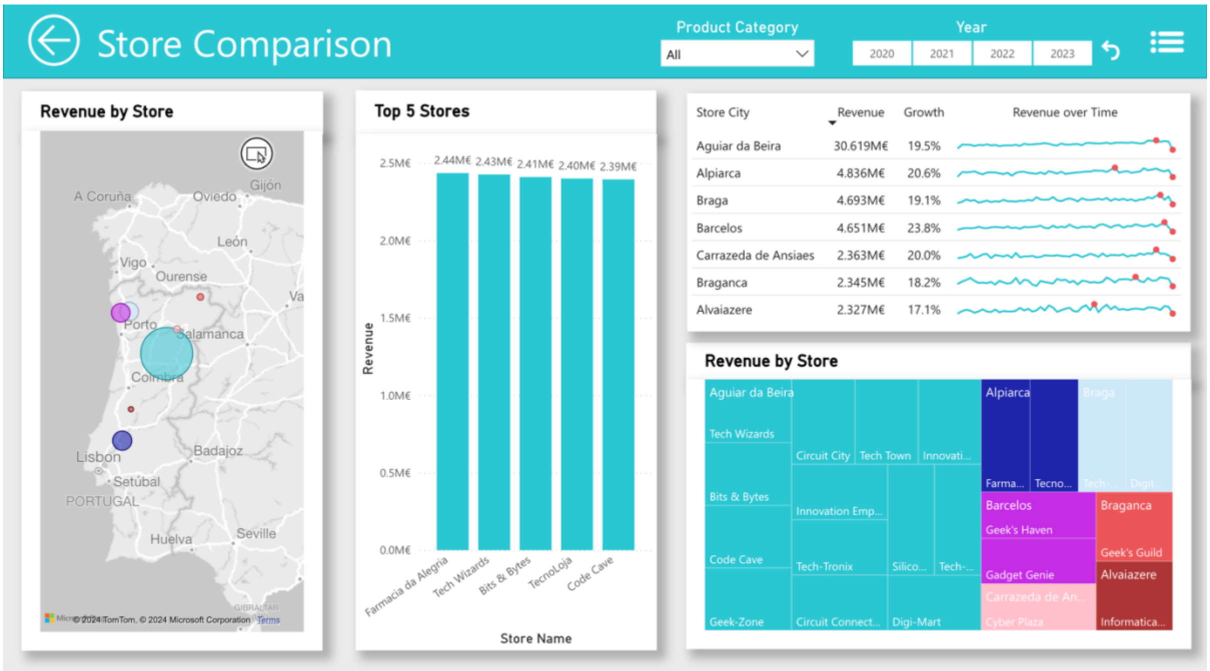
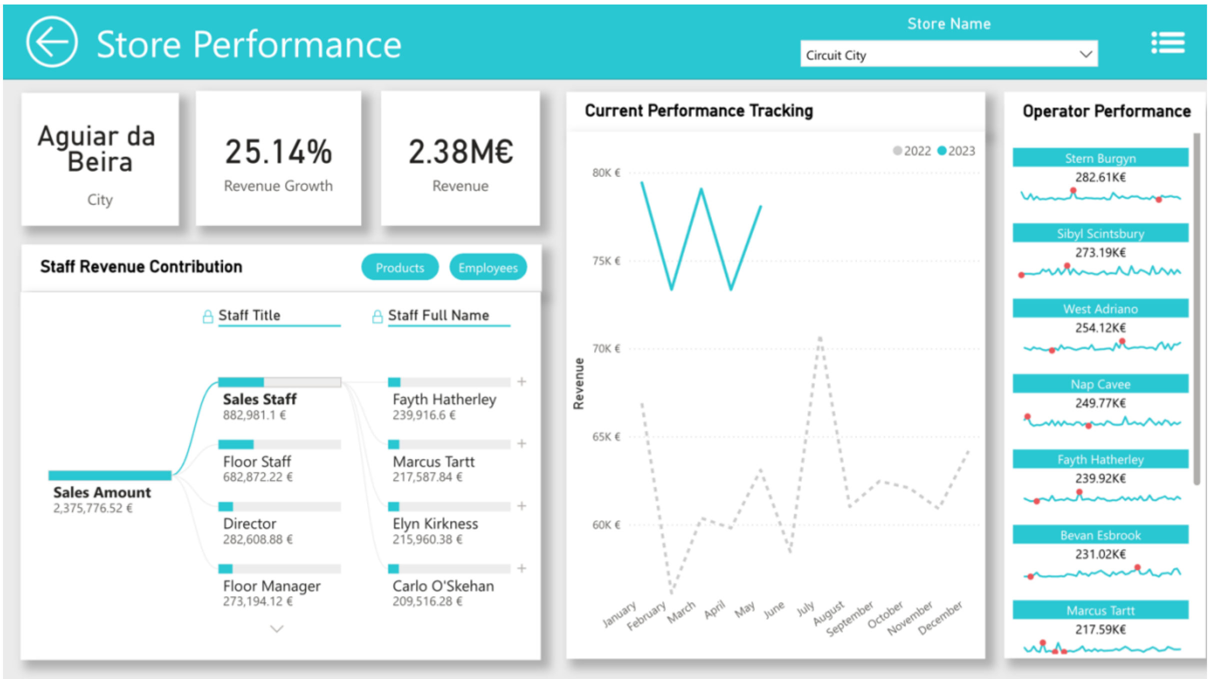


Figure 5: Report Page 1 - Time analysis



Store City	Revenue	Growth	Revenue over Time
Aguiar da Beira	30.619M€	19.5%	
Alpiarca	4.836M€	20.6%	
Braga	4.693M€	19.1%	
Barcelos	4.651M€	23.8%	
Carrazeda de Ansiaes	2.363M€	20.0%	
Braganca	2.345M€	18.2%	
Alvaiazere	2.327M€	17.1%	

Figure 6: Report Page 2 - Store Comparison



Aguiar da Beira
City

25.14%
Revenue Growth

2.38M€
Revenue

Staff Revenue Contribution

Sales Amount
2,375,776.52 €

Sales Staff
882,981.1 €

Floor Staff
682,872.22 €

Director
282,608.88 €

Floor Manager
273,194.12 €

Fayth Hatherley
239,916.6 €

Marcus Tartt
217,587.84 €

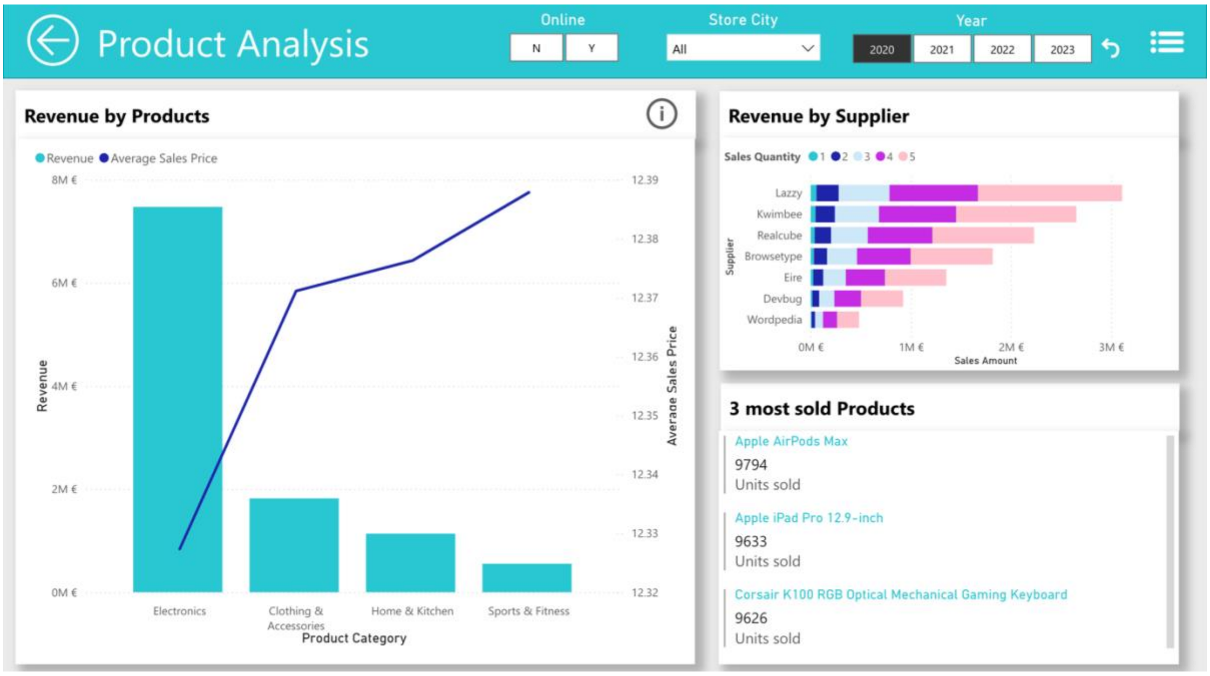
Elyn Kirkness
215,960.38 €

Carlo O'Skehan
209,516.28 €

Operator Performance

Stern Burgyn	282.61K€
Sibyl Scintsbury	273.19K€
West Adriano	254.12K€
Nap Cavee	249.77K€
Fayth Hatherley	239.92K€
Bevan Esbrook	231.02K€
Marcus Tartt	217.59K€

Figure 7: Report Page 3 - Store Performance



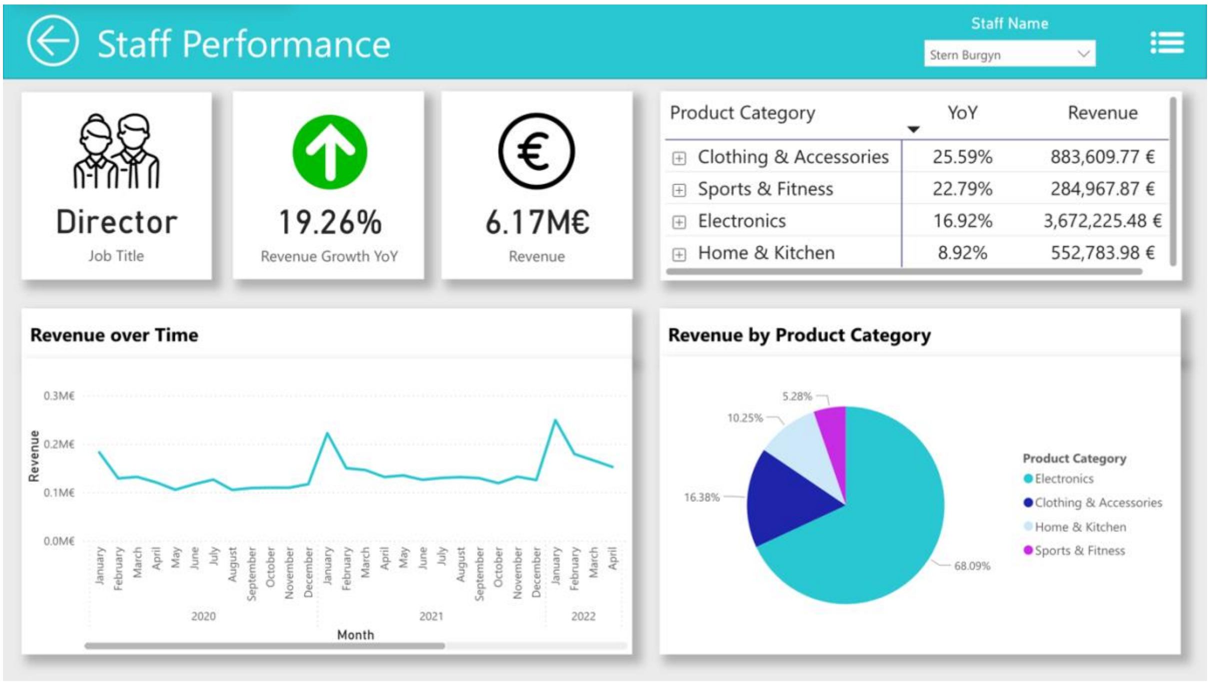
3 most sold Products

Apple AirPods Max
9794
Units sold

Apple iPad Pro 12.9-inch
9633
Units sold

Corsair K100 RGB Optical Mechanical Gaming Keyboard
9626
Units sold

Figure 8: Report Page 4 – Product Analysis



Director
Job Title

19.26%
Revenue Growth YoY

6.17M€
Revenue

Product Category	YoY	Revenue
⊞ Clothing & Accessories	25.59%	883,609.77 €
⊞ Sports & Fitness	22.79%	284,967.87 €
⊞ Electronics	16.92%	3,672,225.48 €
⊞ Home & Kitchen	8.92%	552,783.98 €

Figure 9: Report Page 5 – Staff Performance