BUSINESS INTELIGENCE CASE STUDY

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS
JUNE 2024

BUSINESS INTELLIGENCE

GROUP 15:

Ana Teresa Maia, 20201562
Henrique Falcão, 20201519
Mª Leonor de Gusmão, 20230488
Pedro Carvalho, 20230487

NOVA IMS Information Management School

Index

Image Index

Presentation of the Business

The **retail industry** aggregates millions of enterprises which are the key contributors to the global economic machine. Through a quick investigation we identified Asia Pacific as the fastest growing retail market and North America as the largest market.

In the next figure it is possible to characterize the percentage of market share regarding each region under analysis, evidencing the power of the North America market.

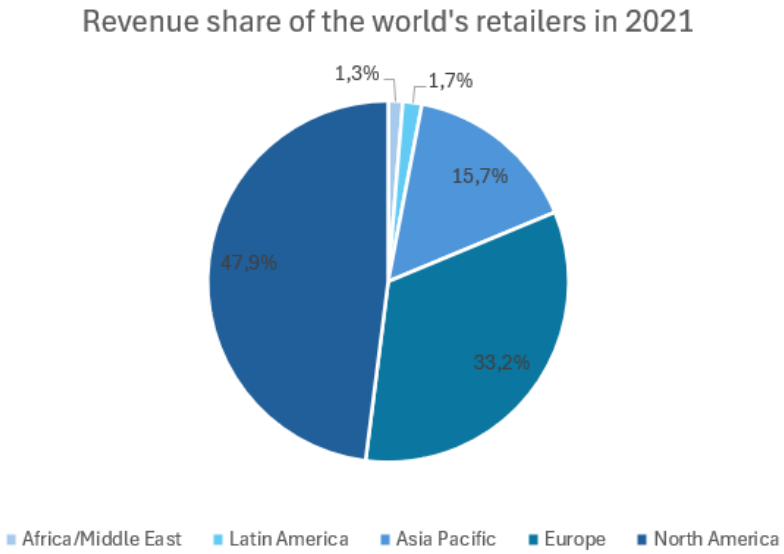

Figure 1- Revenue share by region

There was also a notable development of online retailing or e-commerce channels which are carving out a share of the retail sector in many global markets.
Regarding the current expectations of the Retail Industry. It is expected to grow from USD 32.68 trillion in 2024 to USD 47.24 trillion by 2029 and the global competitors are increasing efforts to collaborate and use cutting-edge technology, such as Artificial Intelligence and Augmented Reality ***(Reference 4).***

Focusing on the initial project situation, we find out that “Retail 4All” is a **retail company** with its business spread among a considerable number of **cities in Portugal.**
Although it is a well known business regarding the **home, sports, clothes and electronics sectors** it's also known that the analytics and decision power are **not in a good situation**.
The **administration provided us with some files (covering 3 years of historic data)** regarding product characteristics, sales details and amounts, stores descriptions, locations and suppliers connections details. Based on this data our team was challenged to look for **solutions** in order to invert the **company's fickle state**.
There is the need to **present an analysis** considering dates, hierarchies, sales details organized by different perspectives, designing discriminant clients profiles.

Business Problems

By analyzing some graphics through **Microsoft PowerBI** and making assumptions regarding the initial context of the company, we confirm the sensitive situation of the company.

The following **problems** were the most evident ones:

Although the sales amount increased throughout time, there is evidence of an **instability in sales**, a considerable tendency of decreasing in the amount of sales between the months of January and March in all years under analysis **(*Present in Figure 2 below*).** This situation may suggest a careful analysis regarding what actions can be made to promote a more stable evolution of sales.

The company needs **readability of sales among the periods** of time under analysis, so that it would be easier to compare evolutions between years/quarters/months, and also to identify possible issues that need to be covered.

Currently it is **difficult to visualize clients profiles** without an appropriate analysis of the segments. There is the need to consider a deeper data analysis regarding products and geographic details that may characterize patterns of the customers.

These types of problems can be overcome with a properly designed analysis through the tools that exist at our reach. Therefore creating visualization dashboards, interpreting evolutions and patterns it's the "key" step during the development of this project, being a plus to increase profitability and competitiveness goals.

Figure 2- Sales Evolution during the 3 years under analysis (Sales amount in Euros)

Data Modelling Methodology

By conducting a deep research (mainly following the Guidebook that can be checked through ***reference[2]***) and taking in consideration the approach suggested in the labs, we went on with the ***Kimball Methodology*** to build our Dimensional Model design. This methodology follows a bottom-up/Data Mart strategy, which is the most common approach. Proposed by Ralph Kimball, it divides tables into Dimensions and Facts in order to optimize the model performance.

This strategy is characterized by having four main courses of action, each one, enabling a deeper understanding on how the business is organized and how we can tackle each business question in the most proper and complete way.

The four stages are:

Identification of the Business Process: This represents the identification of what the business is and what our Data Warehouse will represent. To achieve this part we firstly described the business, dived into how it is organized and what is a part of it, mainly by understanding and exploring the CSV files provided. In our retail case, besides trying to understand how the process flow is organized, we also tried to translate the business needs and problems to detailed relevant business questions.

Identification of the Dimensions: These are what provide context to the measures that the facts encapsulate. The dimensions were built with three things in mind: they must be at least five, of which one is the date; they have to be in line with the granularity of each fact, which will be explored further ahead; lastly, they must be directly connected to the facts, as start schemas only allow for Dimension->Fact relationships.

Identification of the Facts: These capture an objective measurement of analysis. They provide the measures to answer the business questions we made in the first stage.

Identification of the Granularity: The last stage of our methodology is to define the granularity level. This is done by understanding business questions, which are made to be the most specific as possible, being the grain in line with that need. We then return to the point 2 where we confirm that the granularity required was included in our tables as the last level of the depth of the dimension.

Data Source & Discovery

The Retail4ALL folder contains eight distinct files covering comprehensive data regarding sales, store locations, suppliers, products, and stores. Retail4all's administration has facilitated access to CSV files extracted directly from the company's primary servers. These servers undergo daily updates with new data, enabling us to conduct up-to-date analyses and expedite decision-making processes.

Among these files, *Locations* gives details regarding the geographic distribution of stores, comprising 133 entries.

The *NEW Sales (Retail)* file records sales data from 2020 to June 2023. This dataset delineates information concerning the sold SKU, store of transaction, supplier details, location, payment currency (exclusively in Euros), transaction amount, quantity, and the respective operator who assisted the sale.

The *Operator* file contains pertinent information about operators, including their unique identifiers, first and last names, email addresses, genders, roles within the organization, and the teams to which they are assigned. In parallel, the *Point of Supply* provides insights into points of supply (POS), featuring identifiers, names, and email addresses.

Moreover, the *Products* file catalogs product information, encompassing SKUs, product names, and their respective category (4) and subcategory (31) classifications.

Lastly, the *Stores A* and *Stores B* files collectively document details relating to 31 stores with names, locations, and online presence descriptions.

To achieve a more discriminated explanation regarding the sources presented check the ***Figure 51*** available in the annexes. The figure represents a table containing the metadata description associated with all sources.

Dimensional Model Design

After understanding the role of each folder provided as well as ***Kimball Methodology steps,*** our team proceeded with the designing phase of the dimensional model. To achieve this step, our team used the **Draw.io platform.**

The ***Figure 3*** below represents the composition of the dimensional model, containing one fact table and five dimension tables, displayed using a star-schema model (star-shape model containing one fact table at the center, and the dimension tables surrounding it, corresponding to the star's points).

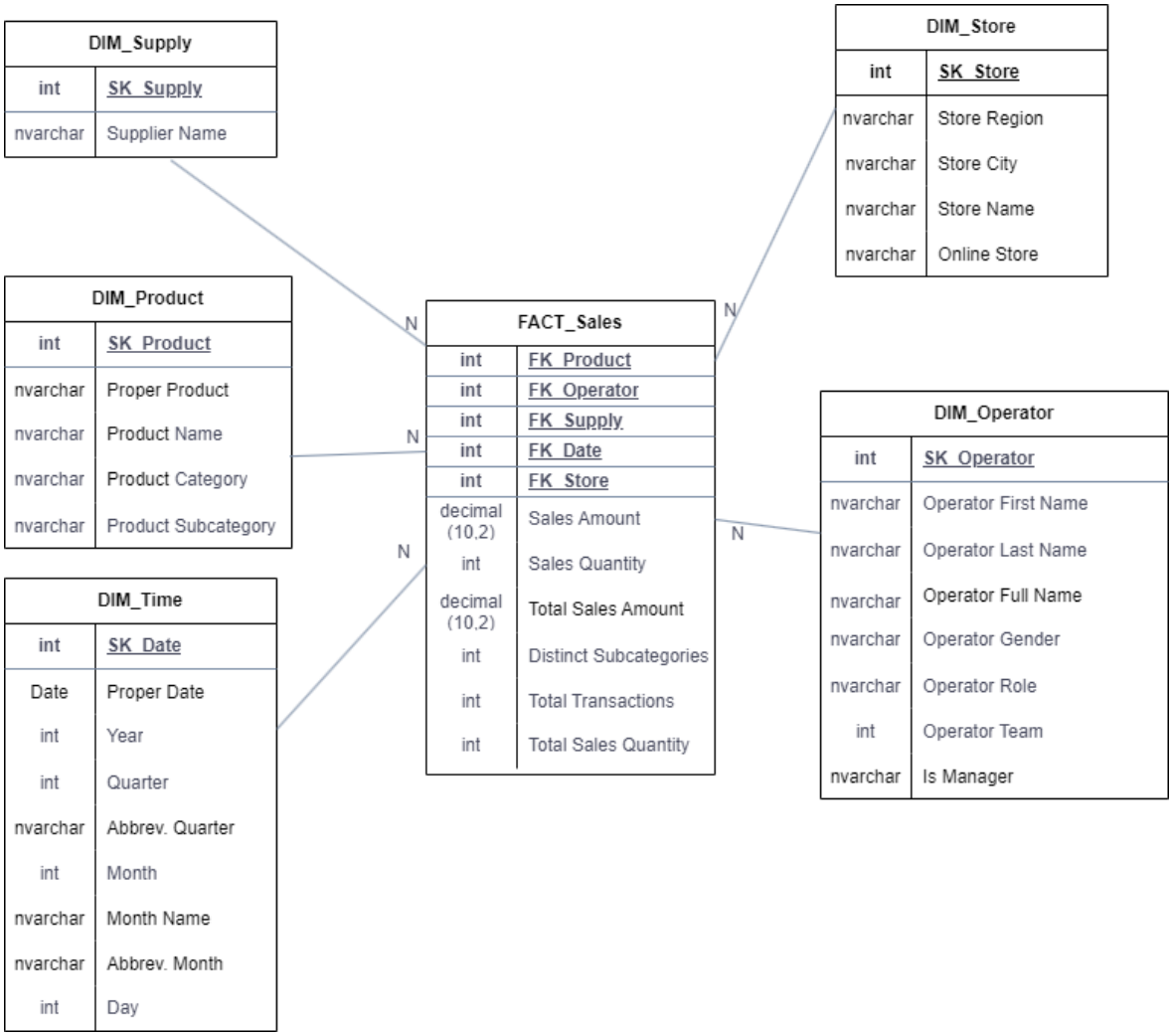

Figure 3- Dimensional model design

Main Components of the Dimensional Model

Following the Kimball Methodology that we described earlier we are now going to provide a detailed explanation of each Dimension and Fact tables.

Identification of the Dimensions: Detailedly Described Below as DIM_"DimensionName"

Identification of the Facts: Detailedly Described Below as FACT_Sales

DIM_Supply

Retrieves information about the Suppliers like their Name (the Email was removed since it did not provide any analytical value to our project). This table has a one-to-many relationship with the FACT_Sales table, since each sale can only have one supplier but each supplier can be associated with multiple sales.

DIM_Product

Retrieves information about the Products like their Name and Category. We can see that it has a Product hierarchy (Category > Subcategory > Name). This table has a one-to-many relationship with the FACT_Sales table, since each product can be sold multiple times.

DIM_Time

Retrieves information about the Time that the sale was made like the year and the month, and we decided to aggregate the months to create the attribute Quarter. We can see that it has a Time hierarchy (Year > Quarter > Month > Day). This table has a one-to-many relationship with the FACT_Sales table, since each sale happens at just one point in time but, for example, in the same hour, multiple sales can happen. As the final steps of this project require a clear interpretation of the dashboard visualizations to reach important conclusions, we decided to create 2 variables with more appealing names for Quarter and Month, reaching more readability regarding dashboard visualizations. The variables in question are 'Abbrev. Quarter' and 'Abbrev. Month' and the creation steps of these variables are further explained in the data engineering phase.

DIM_Store

Retrieves information about the Store that the sale was made in (Name of the Store, if the purchase was made online or not and its Location). This table has a one-to-many relationship with the FACT_Sales table, since each Store can have multiple sales associated with it. From the repository that was given to our team, the Store table had a many-to-one relationship with the Location because one Store is only present at one

location. Since we can not have Dimension tables connected, we incorporated the Location in this table. The 3-depth hierarchy is Region > City > Name. This dimension table was able to store the information regarding the files of Stores A and Stores B. Since the attributes in both were the same, we thought that would be a good idea if we convert these 2 files into one single table. In this way we reduce the model complexity. The transformations applied in this dimension will be further explained in the data engineering phase.

DIM_Operator

Retrieves information about the Employees like their Name and Role. We can see that it has a small depth hierarchy (Team > Role). This table has a one-to-many relationship with the FACT_Sales table, since each employee can make multiple sales but one sale is associated with just one employee.

FACT_Sales

Finally the Fact Table it's the center of our model, since it stores the implicit Measures related to our Sales Transactions, like the Amount that corresponds to the price paid by the client for each product, Quantity that is the number of products bought. The explicit measures calculated lately are properly explained in the semantic model phase.

Identification of the Granularity: We decided to store measures at the granularity of the Day so that in the next stages we can aggregate them in broader time-stamps. This granularity also follows: Sales per Store, Product Subcategory and Name and, finally, Operator Role, being them representative of the lowest level of the dimensions.

Set Up for the ETL Stage

In this section we will provide an in depth explanation of the steps followed to set up the workspace for the ETL Process. Our team is working with Microsoft Fabric. Firstly we built a workspace to enable us to work together, with the name ***BI MAA 2024 Group 15***. An important aspect to notice is that during this stage of our work we were working under the Data Factory side of Fabric. The image below illustrates the workspace view where the following components are included.

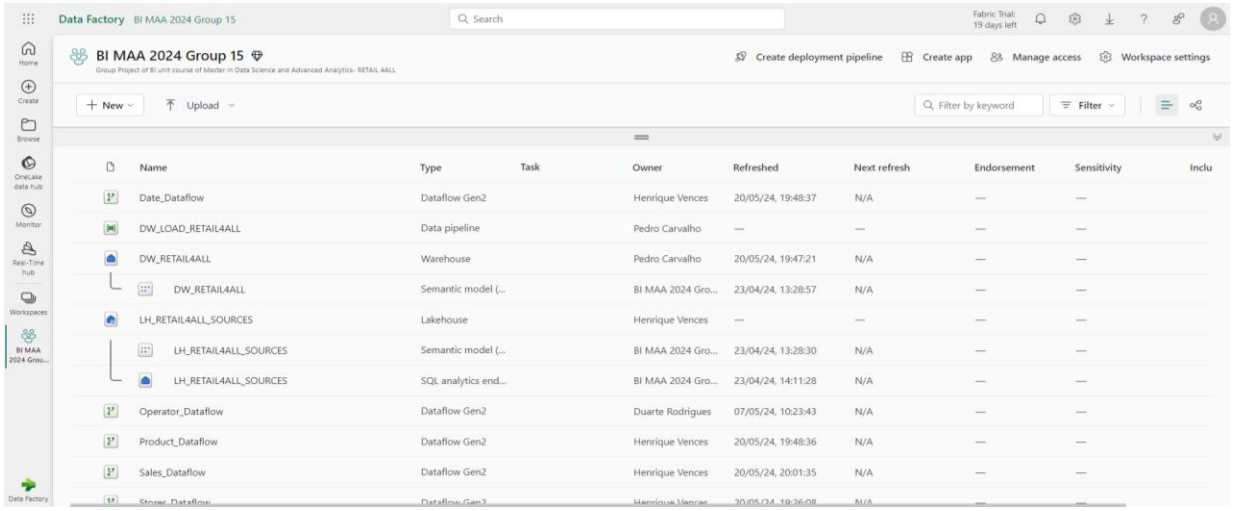

Figure 4- Fabric Workspace

Warehouse(DW_RETAIL4ALL):

The first step we undertook in our ETL stage was to create a SQL script where all the desired tables from the star schema, as well as data types, were established with the command CREATE TABLE. This is where our final data will be located. The following image locates it under our workspace.

Figure 5- Data Warehouse on MS Fabric Workspace

Lakehouse(LH_RETAIL4ALL_SOURCES):

In a project of this kind, the lakehouse is where the so-called *Staging Area* happens. We load all the data sources into it and make the needed transformations to the operational data we were given. As such, all the csv files were uploaded. Its importance is that it acts as a middle point/bridge between the original data and warehouse. It is where data will be transformed and verified before sending it to the final warehouse, ensuring more security and integrity.The following image locates it under our workspace.

Figure 6- Lakehouse on MS Fabric Workspace

Pipeline(DW_LOAD_RETAIL4ALL):

Data pipelines define a sequence of activities that orchestrate an overall process. In this context we will use the pipeline to automate the Extract Transform and Load processes which receives the operational data and prepares it for the analysis purposes it was needed for. The following image locates it under our workspace.

Figure 7- Pipeline on MS Fabric Workspace

By looking at ***Image 5*** below we can understand how the pipeline works. It starts with the deletes, which are important to respect the uniqueness of SK and avoid piling up repetitive data. It then loads the data to the dimensions and to the fact. The pipeline is of huge importance because it is what activates the dataflows, meaning, it is where the data is actually added to the warehouse. The waits were used to mark sequentiality, which ensures data free from silos, in the sense that steps that are dependent on previous ones only run if the previous ones are correct and ran with no problem.

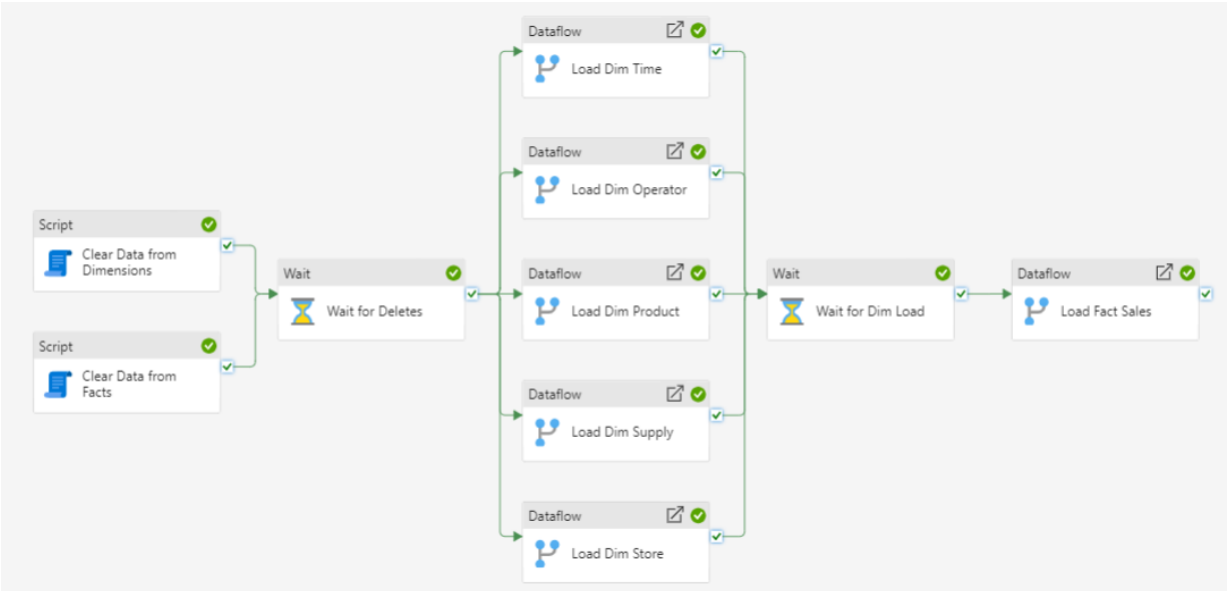

Figure 8- Pipeline of Steps

Dataflows (XXX_Dataflow):

The dataflows are what we use to transform, clean and check data integrity. We use the data from the lakehouse and prepare it to be sent to the warehouse. For a better separation and consistency we use one Dataflow per dimension/fact. Image 6 represents the standard button associated with the creation of a new Dataflow.

Figure 9- Dataflow Card on MS Fabric

All the previous components of Microsoft Fabric were the key element to support our work for the ETL stage of our Business Intelligence project. With them and in them we performed all the transformations described in the next section.

ETL Process - Data Engineering Steps

In this section we will explain the transformations needed to go from the operational data we were given to the data properly structured to be integrated in the warehouse following our Dimensional Model Structure. For simplification purposes we will identify transformations by the dataflow where they happened and the corresponding dimension/fact where it will be loaded to in the warehouse.

Product_Dataflow (DIM_Product)

Under our workspace this dataflow is identified as follows:

Figure 10- Product Dataflow on Workspace

For the product dataflow we started by performing the initial transformations to ensure that data is properly presented. Firstly we did the importing of the Product csv from the lakehouse previously stated, then we performed the first row as header transformation to assign each column name. The column data types were all categorical, which were correctly stated, for safety reasons we also checked the duplicates' existence, there were some of them and we proceeded with their removal. For consistency purposes we also added an index column to substitute the current ‘SK_Product’ column, in this way all the products are identified by a numeric key. The first column was maintained as it will be needed in the future for merging purposes, it was renamed as ‘Proper Product’.

The steps applied go as follows:

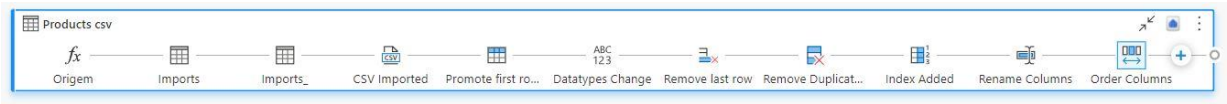

Figure 11- Steps applied to Products Dataflow

To make the link between the csv files and the corresponding column entries in the destination Warehouse each data flow needs to add a data destination under the following button:

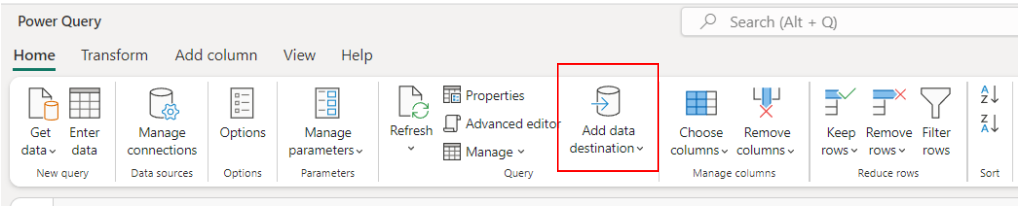

Figure 12- Data Destination Button inside Dataflows

Date_Dataflow(DIM_Time)

Under our workspace this dataflow is identified as follows:

Figure 13- Date Dataflow on Workspace

For this dataflow we simply started from a list of the range of values comprised in our transactions data (from 1/01/2020 to 3/06/2023). We converted this list to a table and ensured the correct datatype *Date* assigned. Then we added the standard columns Year, Month and Day using the add column method for Dates. For the Quarter column, we used the Conditional Column method (explained ahead, in the *Calculated Columns* Section). To compute the last 2 columns, which were 'Abbrev. Quarter' and 'Abbrev. Month' we did some transformations. For the first one we used formatting by adding a prefix. For the latter we used the extract method to get the first 3 characters. The last and most important step for this dimension was to build a unique surrogate key which was based on the data itself. We used a column from examples and transformed the dates *mm/dd/yyyy* to a unique SK of the type *yyyymmdd.* We then send this dimension to the warehouse.The steps applied go as follows:

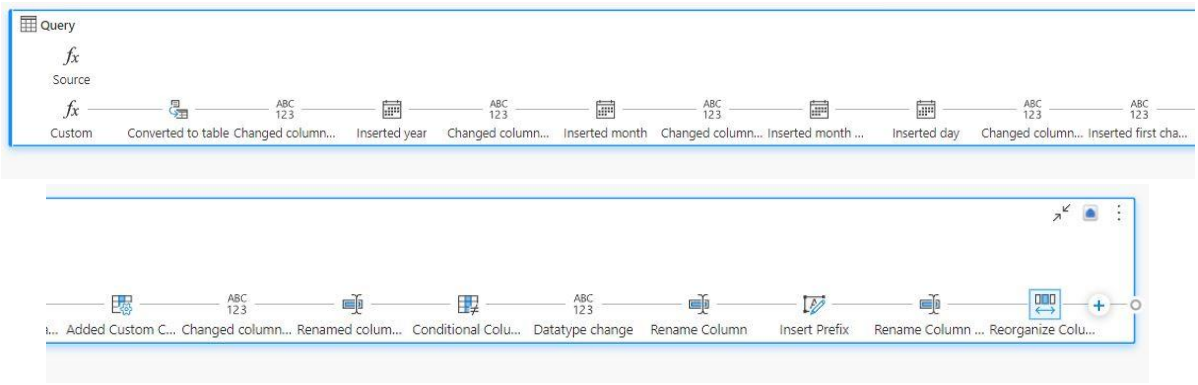

Figure 14- Steps applied to Date Dataflow

Operator_Dataflow(DIM_Operator)

Under our workspace this dataflow is identified as follows:

Figure 15- Operator Dataflow on Workspace

For the operator dataflow we simply had to make general checks to understand if the data was properly stored to follow for the warehouse. After importing the Operator csv from the lakehouse we proceeded to promote the first row as headers. We checked that all columns had the var types properly stored. We then deleted the email column as it was irrelevant for our model and, in fact, it would have decreased the completeness of our data since it had some empty fields. To ensure consistency we checked if gender was stored anywhere else over the dimensions, which was not, in order to understand if its representation was eventually consistent. We also used the remove duplicates method to remove potential duplicates, which was not the case. We generated two new columns (Full Name and Is_manager), whose explanation is in the section *Calculated Columns*, further ahead. Finally, we changed the column names for an easier match in the next stage. We added our warehouse as the destination, with the matching columns and then published it. The steps applied go as follows:

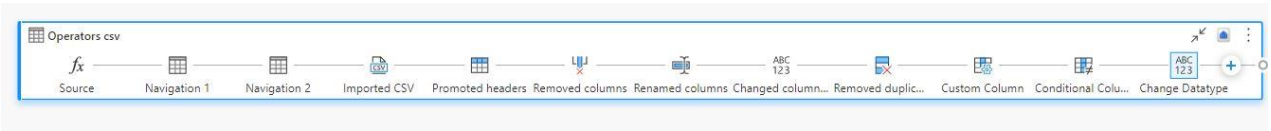

Figure 16- Steps applied to Operator Dataflow

Supply_Dataflow (DIM_Supply)

Under our workspace this dataflow is identified as follows:

Figure 17- Supply Dataflow on Workspace

For the Supply Dataflow there were only minor changes to be made. After importing the Point of Supplier csv from the lakehouse we proceeded to promote the first row as headers and examined if all variable types were properly attributed. The email column was removed since this did not add analytical value and remaining columns were renamed to be the same as in our Warehouse. The steps applied go as follows:

Figure 18- Steps applied to Supply Dataflow

Stores_Dataflow (DIM_Store)

Under our workspace this dataflow is identified as follows:

Figure 19- Stores Dataflow on Workspace

For this dataflow in specific we did some additional changes, according to our interpretation and to facilitate the future analysis. The main transformation was to **append** 'Stores A' and 'Stores B' files as the columns that describe both sources are the same ones. As Stores tables also contain location information ('LocationID') we also reached the conclusion that it would be better to **merge** this table into the new stores table.

Of course that each table in particular required a specific treatment, for that purpose we developed a paragraph to each table, explaining the transformations applied in a more organized manner.

Location Table

Regarding the location table we did the usual importing of the file, and the designation of the correct column names. After a clear understanding of each column role we decided to drop the 'Country' column as it has always the same constant value (All the values presented correspond to 'Portugal'). And regarding the 'Region' column we should highlight the additional treatment that we did in excel, as most of the column values were empty we decided to proceed with a manual imputation. This decision was supported as we believe this information will be useful to achieve more insights regarding location patterns in the reporting phase. In this way we also achieve a 3-depth level hierarchy (Region > City > Name (of the store)), which was a requirement from the previous delivery. The steps applied go as follows:

Figure 20- Steps applied to Location table

Stores A + B Tables

For these files we performed the first row as headers transformation, and then we identified some inconsistencies regarding the 'Online' column. For empty values we decided to substitute them by 'N' (we did the assumption that an empty value means the negation) for 'S' values we assumed it as an 'Y' (considering that 'Y' represents 'Yes' which means 'Sim' in Portuguese and that can be represented by 'S'). After these changes we ended up having the 'Online' column without any misunderstandings.

The duplicate existence regarding the 'Store ID' column was also verified. The ID '20' appeared in 2 different rows which does not make sense. For this purpose we decided to drop one of the duplicated IDs, assigning a unique ID to each row. This decision was supported by the idea that in the Fact table we will never know to which duplicate ID ('20') the line belongs, therefore we considered this to be the more consistent approach, losing only a minimal detail of the information.

The steps applied go as follows:

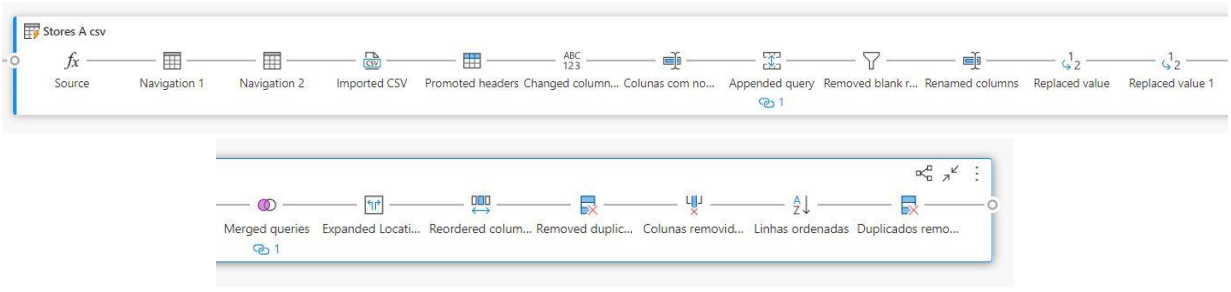

Figure 21- Steps applied to Stores Dataflow

Sales_Dataflow (FACT_Sales)

Under our workspace this dataflow is identified as follows:

Figure 22- Sales Dataflow on Workspace

The fact Sales table deserved a specific treatment. We started by promoting headers as we usually do, ordering columns having in the first positions all the foreign keys available, and renaming these foreign keys with an appropriate name (FK_XXX). Then we dropped non-valuable columns such as 'Currency' as it has always the same constant value (EURO). We also dropped the location column because due to the steps we did regarding the new stores table, it is no longer needed in the fact table, in other words we access it through the DIM_Stores table.

An incoherence was also noticed. By filtering the values in descending order, we identified an unusual value for the amount column (900005€). As we only verified this specific case, we decided to remove this line, losing only a minimal detail of the data.

To add the foreign key from the 'DIM_Time' table we had the need to import this table and use the expand command selecting only the surrogate key ('SK_Date'), the name was changed to a FK_XXX format after this transformation. It's important to highlight that for this purpose we needed to merge the actual fact table with the DIM_Date table based on the 'Proper Date' column.

Regarding 'DIM_Product' a similar treatment was developed, we merged the 'DIM_Product' table with the fact table as we needed to acquire the 'SK_Product'(only with numerical identifiers), created previously in the Product dataflow. For the merging purposes we used the 'Proper Product' column (the business key column in common between the fact table and the 'DIM_Product' table), however the main goal of this was to extract 'SK_Product'(using the expand command) that will be represented as 'FK_Product' in the fact table. In this way we ended up having the 'FACT_Sales' table with all the foreign keys, and the correct transformations applied.

The following image represents the connection between DIM Date DIM Product and Sales.

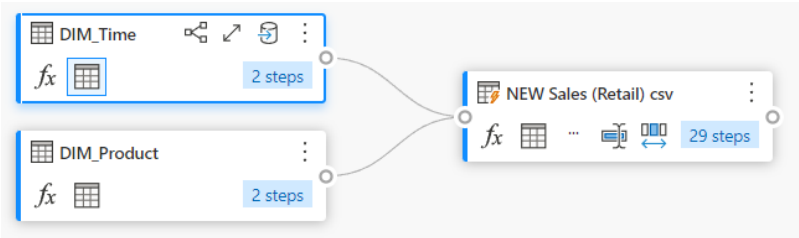

Figure 23- DIM merges with Fact Table

Semantic Model

At this stage, the process started with the creation of a new semantic model. Its identification in the workspace is the following:

Figure 24- Semantic Model in MS Fabric

Then we proceeded by connecting the dimension tables to the fact table, ensuring always (one-to-many relationships), having the one (1) on the dimension side (meaning uniqueness) and the many (*) on the factual side.

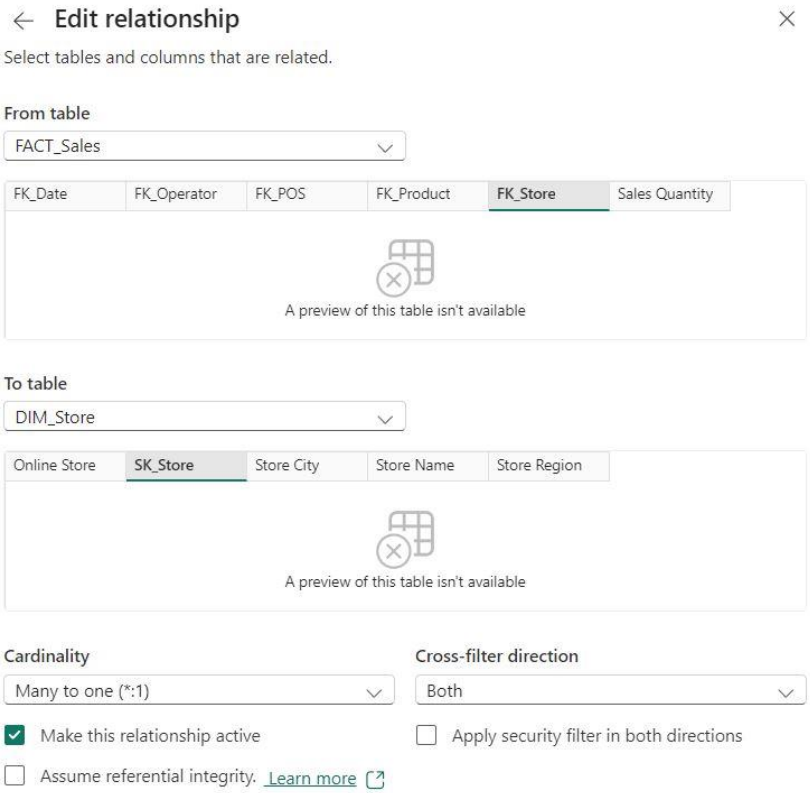

Figure 25- Relationship Editing Menu

Then we marked the date table as date by choosing the Proper Date as the date column.

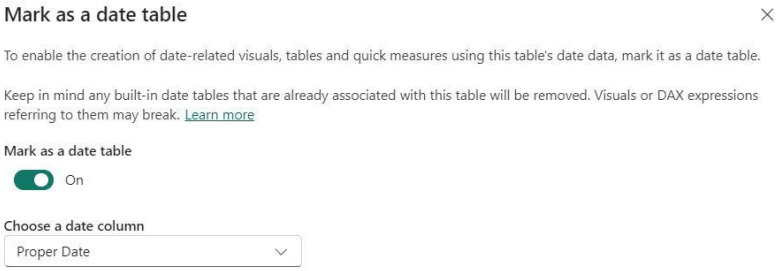

Figure 26- Marking Date table as Date

To ensure more readability, some column names were also changed to a more complete and intuitive description.
Then hierarchy creation was also verified. We proceeded by creating the 3 hierarchies described before (Store, Product and Date hierarchies).

Figure 27- The 3 hierarchies created

We also decided to hide the features that are keys ('SK' and 'FK'), as they are not relevant for the reporting stage, at this stage we only care about descriptive features that would ensure a faster understanding in the report graphics.
Another approach followed, was to hide the actual columns of each hierarchy as they will appear repeated, and this is redundant for our model.
To the steps described above we just clicked in the eye icon corresponding to each of the features. The final vision regarding hidden features is available in ***Figure 31.***

For the columns of 'Month Name', 'Abbrev. Month' and 'Abbrev. Quarter', we used the corresponding columns of numbers to order months and quarters names properly (we used 'Month' and 'Quarter' columns for this purpose, as their values only correspond to the number associated with each month/quarter order). We want to show these names ordered, since it is essential to understand evolutions throughout time in the reporting stage. This procedure was completed using properties capabilities, adding a 'sort by' to the corresponding column at the advanced options.

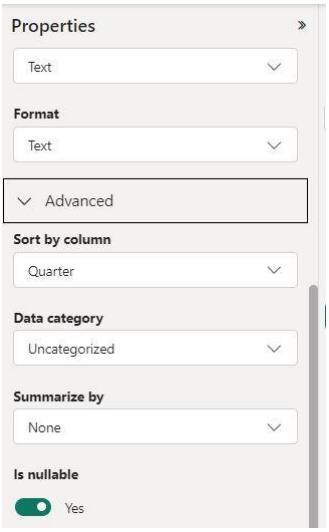

Figure 28- Procedure to order 'Abbrev. Quarter' column by 'Quarter' numeric column

The format of the implicit measures provided was also verified by our team.
For *Sales Quantity'* we verified if it was a whole number, if there were no decimal spaces, as it represents a quantity. For this implicit measure we also add the thousands separator for a better readability. Regarding *'Sales Amount'* the currency format was also specified to be in euros (€ Euro (€123), as we saw before regarding the initial data the amounts were given in euros), saving 2 decimal spaces.

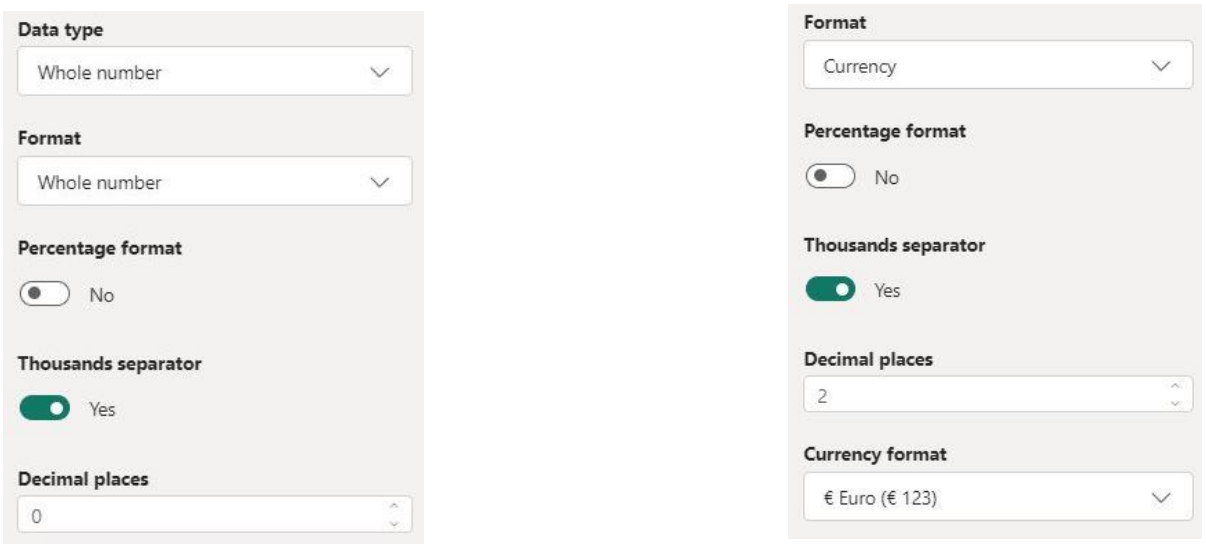

Figure 29- Formatting 'Sales Quantity' and 'Sales Amount' respectively

To finalize the elaboration of the semantic model, we computed some Dax measures. The code and the corresponding description are described in the next table:

DAX Code:	Measure Description:
Total Sales Amount = SUMX(FACT_Sales, FACT_Sales[Sales Amount]*FACT_Sales[Sales Quantity])	**Total Sales Amount:** Represents at a row level the total sales amount. For each row, it multiplies the 'Sales Amount' by the 'Sales Quantity', then sums up these calculated values.
Total Sales Quantity = SUM(FACT_Sales[Sales Quantity])	**Total Sales Quantity:** This measure calculates the total quantity of sales by summing up the 'Sales Quantity' column from the 'FACT_Sales' table. It gives you the total number of items sold across all transactions.

Figure 30- Measures computed in the Semantic model

After all these steps we obtained our final semantic model version, visible in the image below:

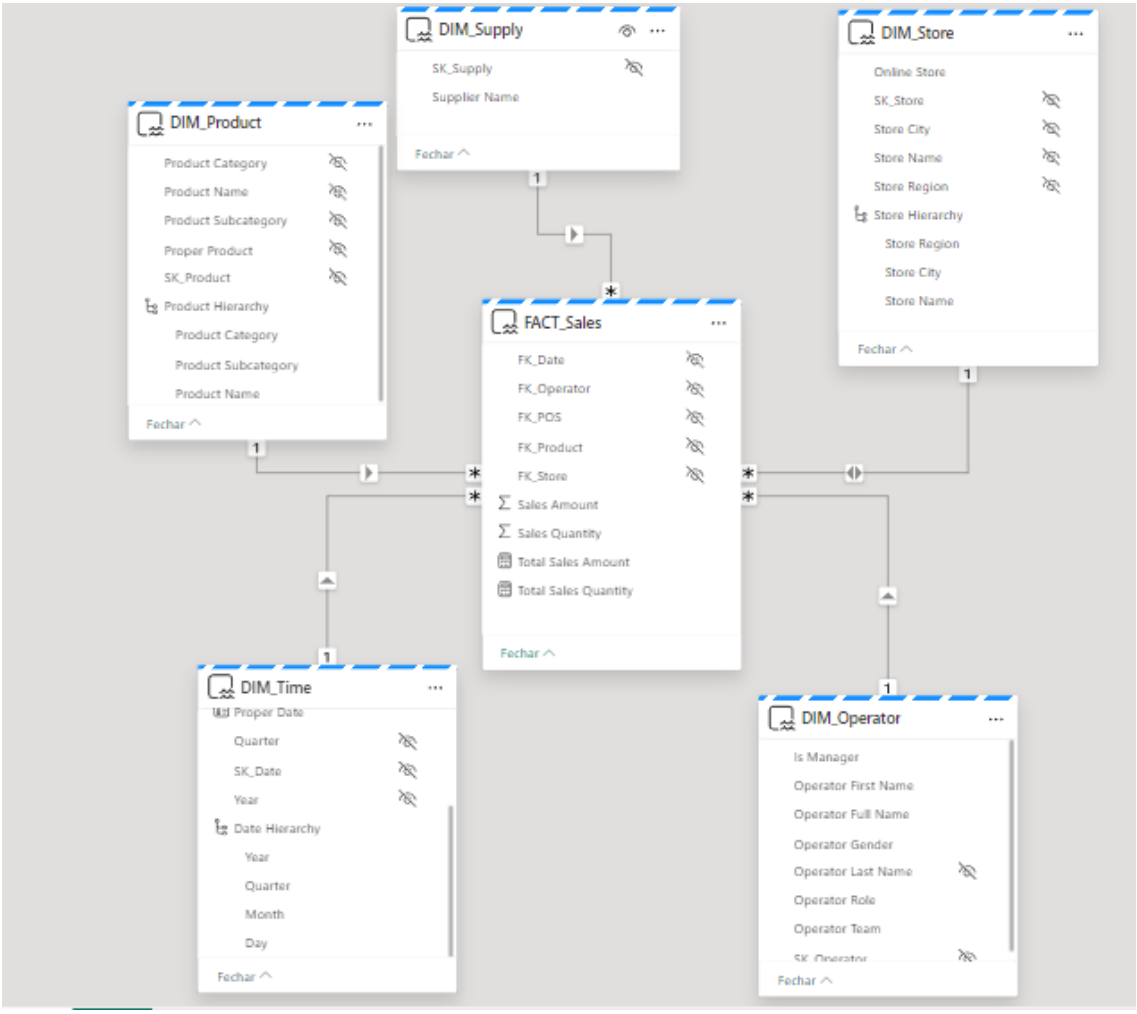

Figure 31- Final Semantic model after all transformations applied

Dashboard Production in Power BI Desktop

The first thing we did when starting to develop the project in Power BI Desktop by importing the semantic model from our fabric workspace, was to replicate the Dax measures, according to our predefined business questions. The measures developed in Power BI Desktop and the corresponding description are represented in the tables below. The first table refers to the time intelligence measures (which enable us to manipulate data using time periods, including months, quarters, and years, and then build and compare calculations over those periods), and the other table refers to other measures applied to our specific business context.

Time Intelligence Measures

DAX Code:	Measure Description:
Total Sales Amount MoM% = VAR __PREV_MONTH = CALCULATE([Total Sales Amount], DATEADD('DIM_Time'[Proper Date], -1, MONTH)) RETURN DIVIDE([Total Sales Amount] - __PREV_MONTH, __PREV_MONTH)	**Total Sales Amount MoM%:** Calculates the month-over-month percentage change in total sales amount.
Total Sales Quantity MoM% = VAR __PREV_MONTH = CALCULATE([Total Sales Quantity], DATEADD('DIM_Time'[Proper Date], -1, MONTH)) RETURN DIVIDE([Total Sales Quantity] - __PREV_MONTH, __PREV_MONTH)	**Total Sales Quantity MoM%:** Calculates the month-over-month (MoM) percentage change in total sales.
Total Sales Amount QTD = TOTALQTD([Total Sales Amount], 'DIM_Time'[Proper Date])	**Total Sales Amount QTD:** Calculates the total sales amount quarter-to-date.
Total Sales Amount YTD = TOTALYTD([Total Sales Amount], 'DIM_Time'[Proper Date])	**Total Sales Amount YTD:** Calculates the total sales amount year-to-date.

Figure 32- Time-intelligence measures and the corresponding description

Other Measures

DAX Code:	Measure Description:
Total Sales Amount average per Operator Team = AVERAGEX(KEEPFILTERS(VALUES('DIM_Operator'[Operator Team])), CALCULATE([Total Sales Amount]))	**Total Sales Amount average per Operator Team:** Calculates the average total sales amount per operator team (Row level)
Total Sales Amount average per Product Subcategory = AVERAGEX(KEEPFILTERS(VALUES('DIM_Product'[Product Subcategory])), CALCULATE([Total Sales Amount]))	**Total Sales Amount average per Product Subcategory**: This measure provides the average sales amount for each product subcategory (Row level)

Total Sales Amount running total in Product Subcategory = CALCULATE(
  [Total Sales Amount],
  FILTER(
  ALLSELECTED('DIM_Product'[Product Subcategory]),
  ISONORAFTER('DIM_Product'[Product Subcategory], MIN('DIM_Product'[Product Subcategory]), DESC)))
	Total Sales Amount running total in Product Subcategory: Calculates a running total of sales within the context of the selected product subcategories.

Total Sales Amount average per Month = AVERAGEX(KEEPFILTERS(VALUES('DIM_Time'[Month])), CALCULATE([Total Sales Amount]))	**Total Sales Amount average per Month**: Calculates the average of total sales amount per month at a row level.
Total Sales Amount average per Supplier Name = AVERAGEX(KEEPFILTERS(VALUES('DIM_Supply'[Supplier Name])), CALCULATE([Total Sales Amount]))	**Total Sales Amount average per Supplier Name:** Calculates the average of total sales amount per supplier at a row level.
Online Store Percentage = VAR TotalStores = COUNTROWS(DIM_Store) VAR OnlineStores = CALCULATE(COUNTROWS(DIM_Store), DIM_Store[Online Store] = "Y") RETURN DIVIDE(OnlineStores, TotalStores, 0)	**Online Store Percentage:** This measure calculates the percentage of online stores among all the stores presented in the analysis (online+not online).

LY Sales Amount = CALCULATE('FACT_Sales'[Total Sales Amount], SAMEPERIODLASTYEAR('DIM_Time'[Proper Date]))	**LY Sales Amount:** Corresponds to the same period of the previous year sales' total amount.
% LY Sales = DIVIDE('FACT_Sales'[Total Sales Amount],[LY Sales Amount],0)	**% LY Sales:** The percentage of sales from last year that the current sales represent, depicts if there is a positive or negative evolution in sales, used as KPI.
BestOperatorName = VAR MaxSalesOperator = SELECTCOLUMNS (TOPN (1, SUMMARIZE (ALLSELECTED(DIM_Operator), -- Consider slicer selections DIM_Operator[Operator Full Name], "TotalSales", [Total Sales Amount]), [TotalSales], DESC), "OperatorFullName", MAX(DIM_Operator[Operator Full Name])) RETURN SELECTCOLUMNS (MaxSalesOperator, "Operator Full Name", [OperatorFullName])	**BestOperatorName:** Finds the name of the Operator associated with the best total sales amount (top 1), it respects the slicer options to be dynamic.
BestOperatorGender = VAR BestOperatorName = [BestOperatorName] RETURN CALCULATE(MAX(DIM_Operator[Operator Gender]), DIM_Operator[Operator Full Name] = BestOperatorName)	**BestOperatorGender:** Finds the gender of the Operator associated with the best total sales amount (top 1), it respects the slicer options to be dynamic.
BestOperatorRole = VAR BestOperatorName = [BestOperatorName]	**BestOperatorRole:** Finds the role of the Operator associated with the best total sales amount

RETURN CALCULATE(MAX(DIM_Operator[Operator Role]), DIM_Operator[Operator Full Name] = BestOperatorName)	(top 1), it respects the slicer options to be dynamic.
Number of Transactions = COUNTROWS(FACT_Sales)	**Number of Transactions:** This measure counts how many transactions have been processed. As an assumption each transaction corresponds to a line of the fact table.
LY Number of Transactions =CALCULATE([Number of Transactions], SAMEPERIODLASTYEAR(DIM_Time[Proper Date]))	**LY Number of Transactions:** Follows the same approach as the variable LY Sales Amount.
% LY Transactions = DIVIDE([Number of Transactions],[LY Number of Transactions],0)	**% LY Transactions:** Follows the same strategy as the variable %LY Sales Amount.
Growth Rate Sales = [%LY Sales]-1	**Growth Rate Sales:** percentage growth of sales amount considering the same period of the previous year.

Figure 33- Other Measures and the corresponding description

Note that some of these measures were not used to compute any visual scenario, however we decided to keep them as they can be useful to achieve future insights if the project had continuity.

The format of all measures was also verified considering its context.

Calculated Columns

The creation of the calculated columns had to be made in the Dataflows in Microsoft Fabric. For that our team went back to the affected dataflows and managed to create/update these columns. After we proceed to re-implement the changes in the Semantic model done before in the affected Dimensions. After that we went again to Power BI to complete the dashboard.
The three calculated columns and their purpose go as follows:

- **Operator Full Name**:

This column enables a better identification of each unique employee and provides a better visualization when comparing teams' performance in the dashboard analysis. To compute it we used the Custom Column method as bellow:

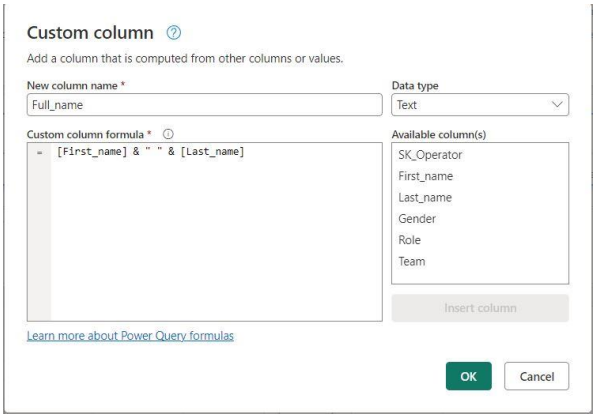

Figure 34- Full Name Custom Column

- **Quarter:**

For this column we simply wanted to keep information about the quarter and instead of directly extracting it from the proper date we used the Month column to build a conditional column from it as follows:

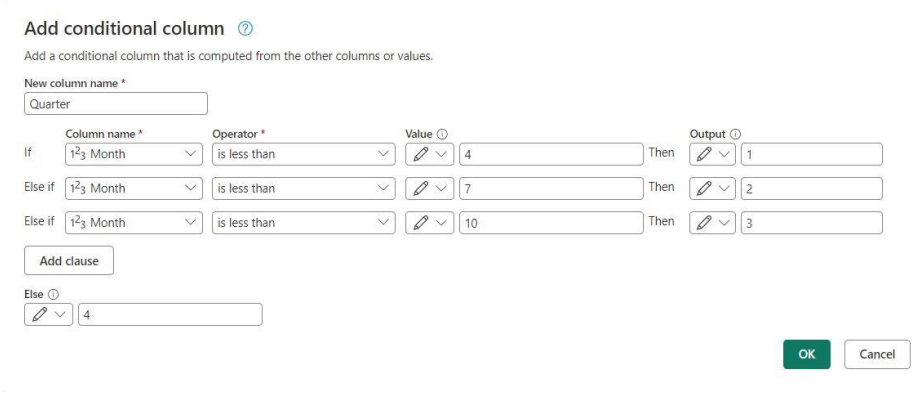

Figure 35- Quarter Conditional Column

- **Is Manager:**

The creation of these columns was made to enable an easier identification of hierarchies within RETAIL4ALL internal structure, which, further ahead, will be used in building row level security rules. For this column we use the Conditional Column method in Operator Dataflow, as follows:

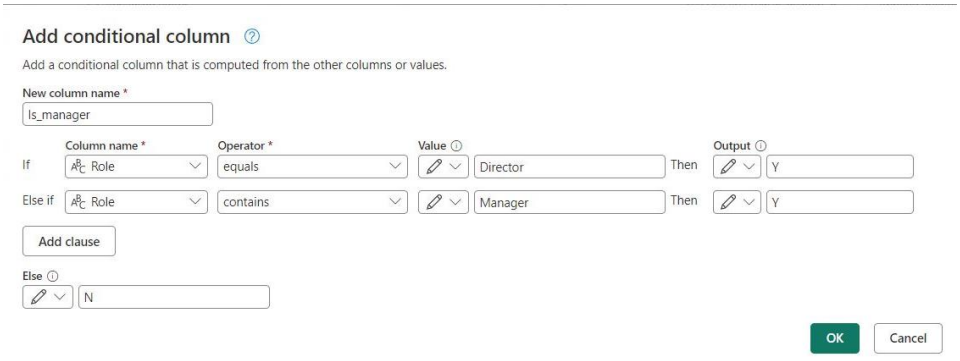

Figure 36- Is Manager Conditional Column

KPI Visuals

Key Performance Indicators are guides, both for employees and for employers. They enable a company to understand how the performance of the business is evolving and whether objectives are being met or not. Our team decided to come up with 2 relevant KPI visuals that cover the assessment of 2 different dimensions on 2 different metrics.

The first KPI regards **Stores Performance**. The goal is to understand whether or not Stores are making more or less transactions than in the same period of last year. It enables RETAIL4ALL to understand if the location has lost client engagement or if there are any negative circumstances affecting the store workflow. We kept a target of achieving at least the same level of transactions (100%). The color scheme used also provides a quick interpretation: green for objective met and red for objective not met. It varies according to the slicer selections, ensuring an accurate representation.
It is found on the page 'Stores' as follows:

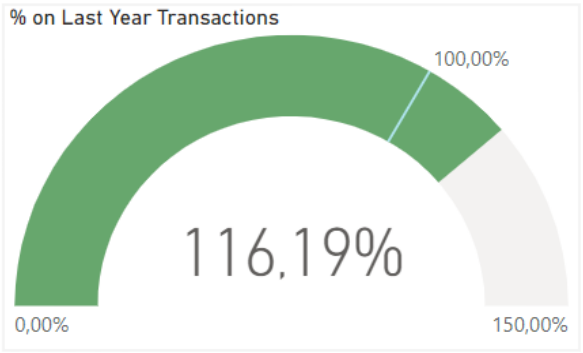

Figure 37- Percentage on Last Year Transactions KPI

The second KPI concerns **Operators Performance**. The goal is to understand if operators managed to outdo themselves in terms of sales. The assessment evaluates if the workers were able to increase the sales amount by at least 10% when compared to the same period of last year (target = 110%). The color scheme follows the one in the previous KPI. It is found on the page 'Operators' as follows:

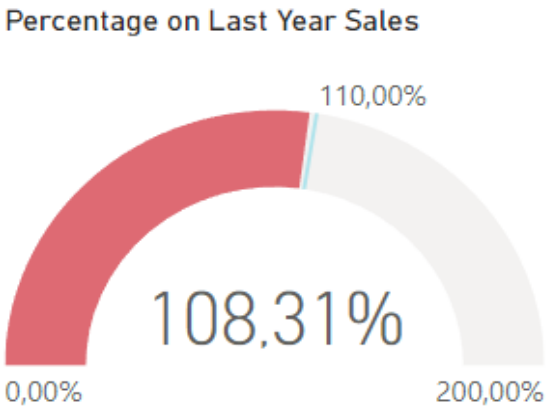

Figure 38- Percentage on Last Year Sales KPI

Main technical aspects of the report

- **Advanced Scenario**

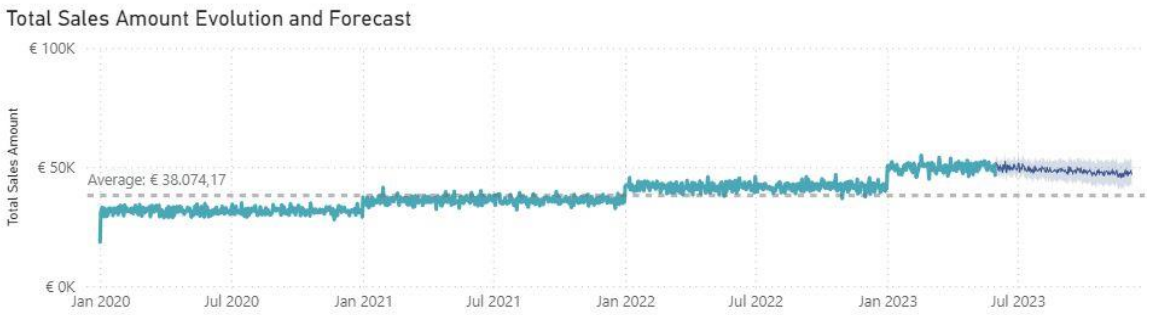

Figure 39- Total Sales Amount Evolution and Forecast Analysis (shaded area)

An advanced visual scenario was computed in the **'Overview page'** of our dashboard file. This visual represents the total sales amount evolution and forecasting scenario in a line chart, between June and December of 2023. This can give us a better overview regarding the total sales amount behavior in a short-term period. The scenario represented above was computed considering the specifications selected in the image below. These specifications allow us to create a detailed and seasonally adjusted (for the entire year) forecast based on daily data, projecting 190 days into the future with a 95% confidence interval.

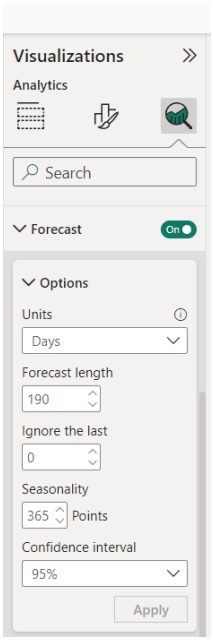

Figure 40- Forecasting parameters selected

- **June Filtering:**

Following the development of our dashboard we identified an anomaly in the month of June 2023, therefore we decided to filter out those values from all pages in our report as it was masking some trends. We will then discuss with RETAIL4ALL to understand if any extra error in the data was present.

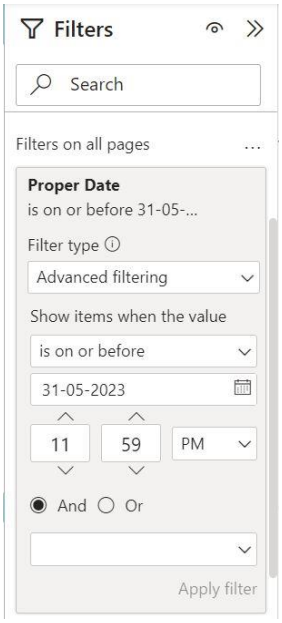

Figure 41- June filtering

Extra Requirements

To show some additional knowledge regarding other BI utilities, we computed the following extra content:

- **Row level security:**

The attempt our team did in promoting row level security was regarding hierarchical levels in the company. We assumed directors could see all information, managers could see all information besides the director's and the remaining employees could see only their team's information. For that, in our semantic model we managed 3 different roles, creating rules on the columns Operator Team and Is Manager:

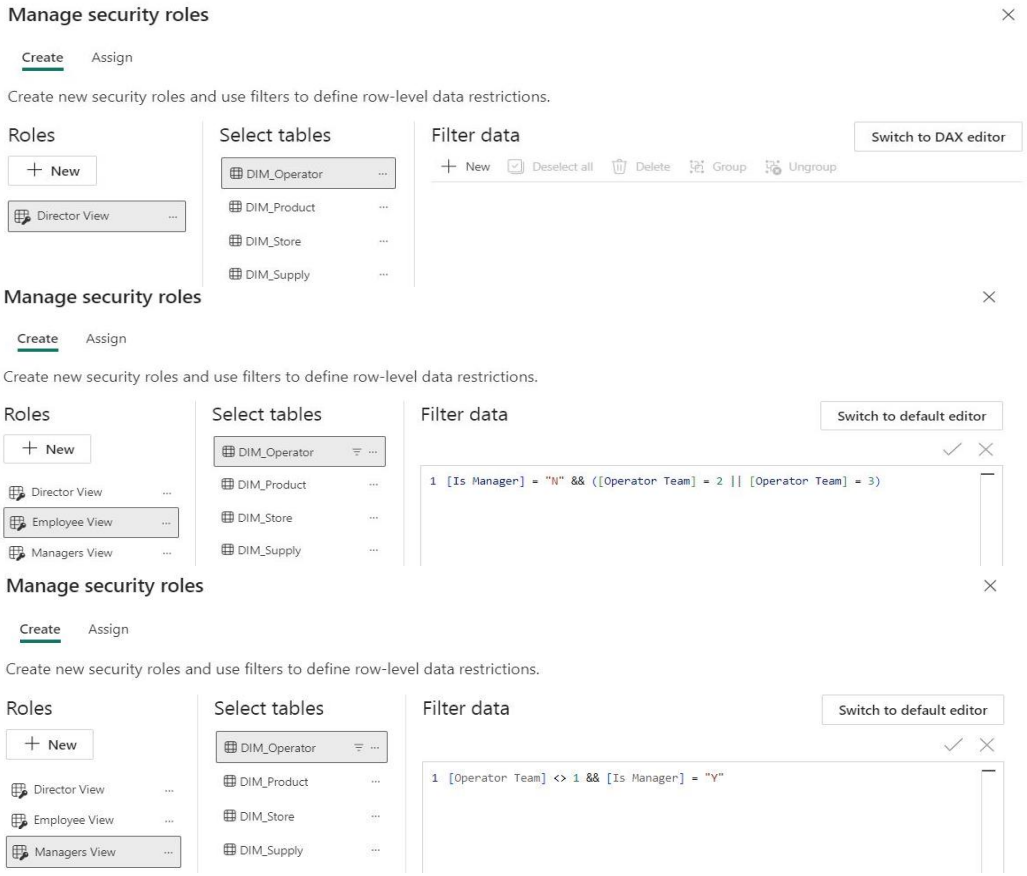

Figure 42- Roles creation

Finally, to put it into eventual force we applied the defined roles:

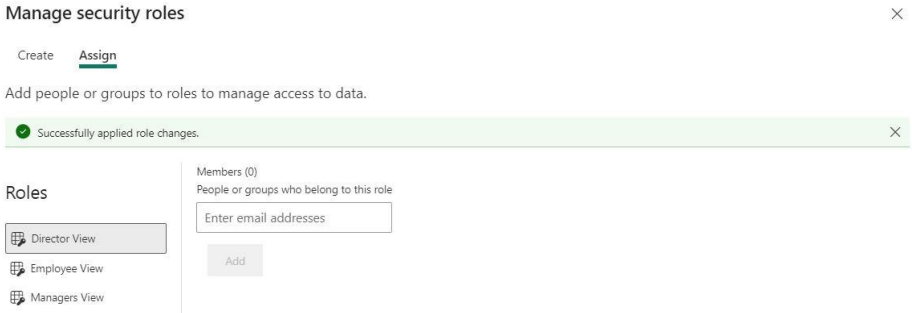

Figure 43- Assign row level security

- **Search Bar for quick filtering:**

Besides having multiple filters on the operator page we decided it would be interesting to include a quick filtering option to give flexibility and speed when looking for an employee's performance (inside the team). For that we added a search bar in the operator page, following the next steps: First we installed a new visual tool, under the visuals …>Get more visuals>Search>Text filter>Add:

Figure 44- Text Filter Installation Package

Then we clicked on the icon below and filtered it for the Operator First Name, so that when searching the employee name, it filters the whole page automatically for it.

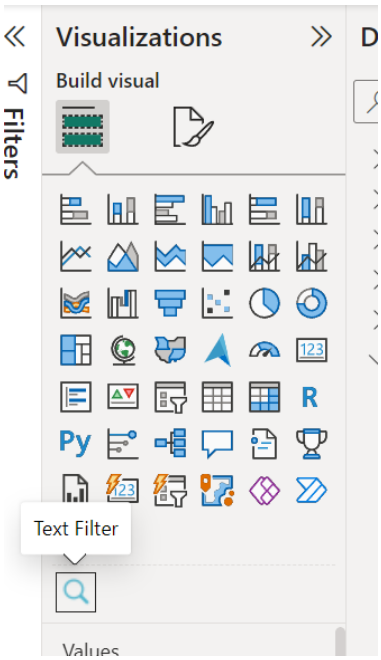

Figure 45- Text Filter Icon

Dashboard analysis:

Under this section our team will make an in depth analysis of each page on the report we built, **addressing all the business questions** made at the same time.

‘Table of Contents Page’:

This page is an introductory page that contains several buttons whose purpose is driving the user to the page he wants to analyze.

‘Overview Page’:

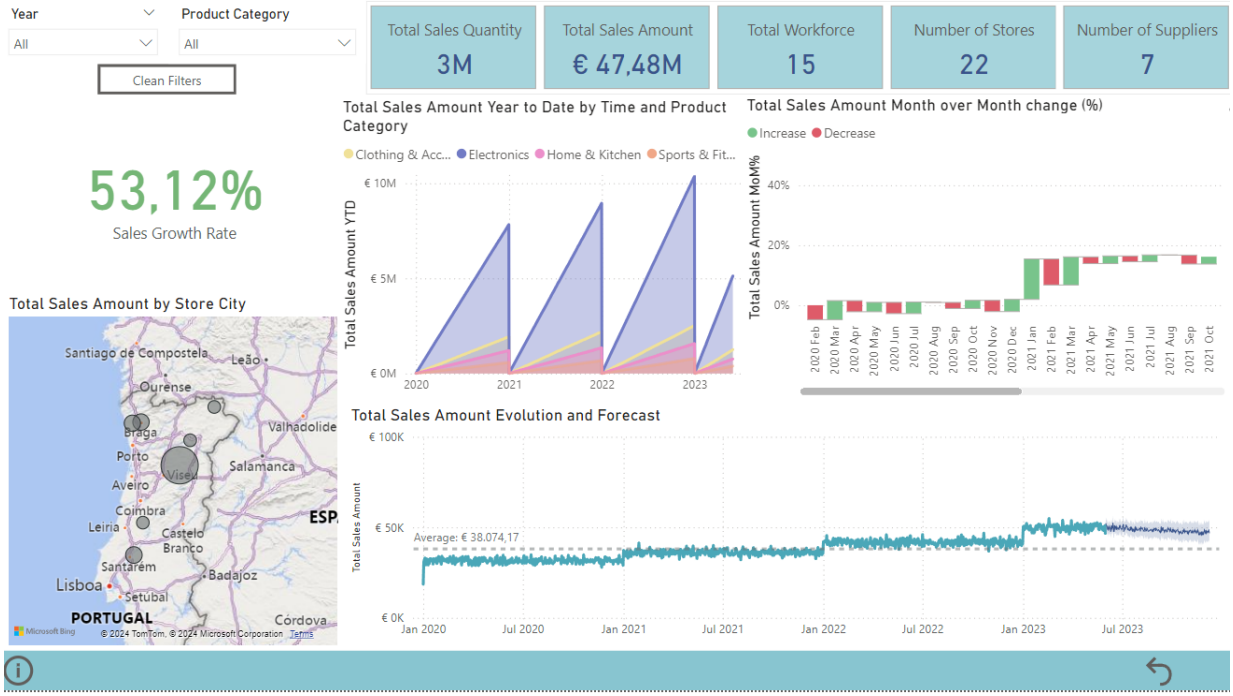

Figure 46- Overview Report Page

This page has the purpose of getting an overview regarding the whole subject studied, focusing mainly on the sales amount evolution and geography.

The filtering section contains slicers to select the year that is desired to be analyzed, and also the product category in specific. Next to this there is a complete section of cards that bring the information explained on the other pages to here but in a small picture (when talking about stores, operators and suppliers numbers). It also refers to the total sales amount and quantities.

The sales growth rate (%) card shows the percentage growth of sales considering the corresponding period of the previous year. The value remains green as long as it is superior to 0 (growth registered) and red otherwise, in general the value has been increasing through the years, being superior to 0.

To better understand the different evolution between years for the studied product categories for total sales amount, an area chart was computed. The month over month change in % regarding the total sales amount was computed in a waterfall chart, in this way it would be clear to see the increases and decreases over months.

The bottom part of the page is composed of a map that illustrates the stores locations and highlights the bubbles in which locals of the Portugal country have the highest values of total sales amount (the higher the value the higher the bubble). Lastly a line chart was computed to show the overall total sales amount evolution and a forecasting for the last months of 2023 (a detailed analysis of this visual is described above in the advanced scenario section).

Business questions

By analyzing this page it's possible to answer some key questions:

(1)What can be the **future trend** of total sales amount evolution for the next six months of 2023?

- The line chart represented in the bottom part of the page showed a constant sales evolution that has been slowly increasing through the years analyzed.
  For the next months of 2023 it is expected the total sales amount value to continue above the average and with constant evolution.

'Products Page':

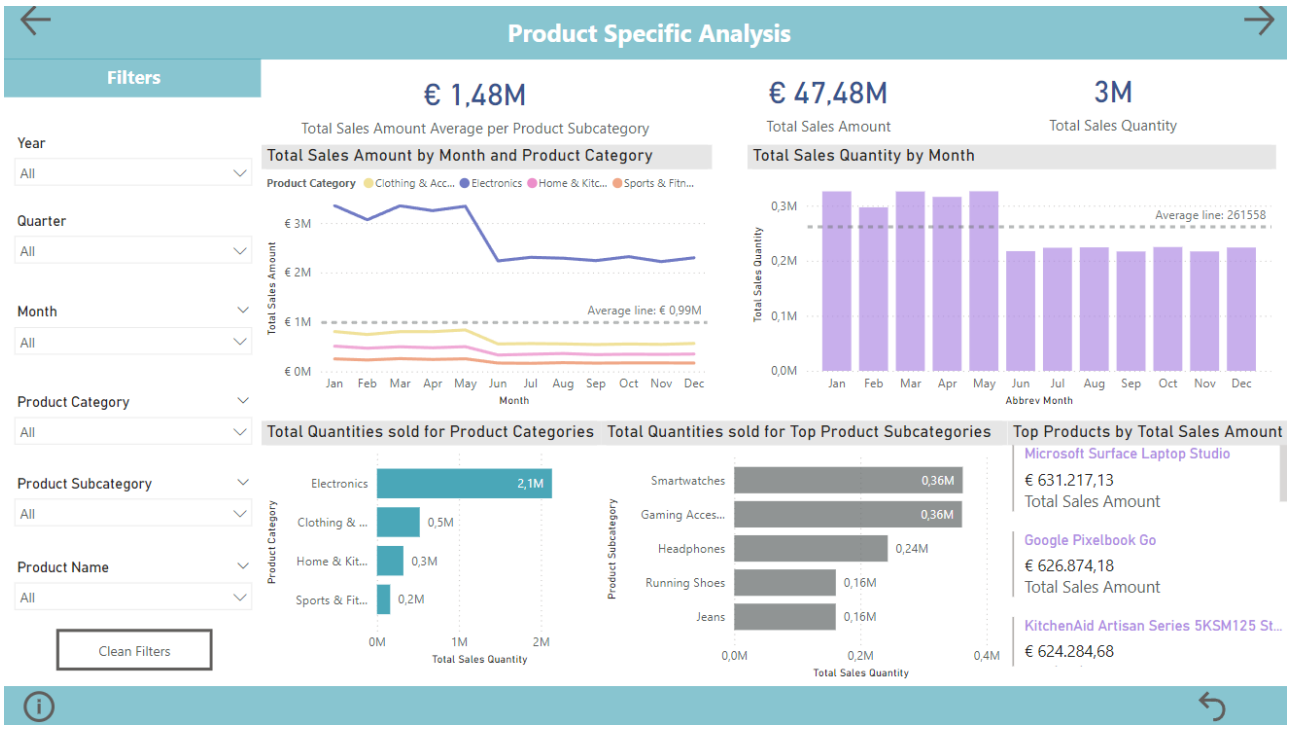

Figure 47- Products Report Page

This page transmits a specific overview regarding the products characteristics under analysis. For this purpose a filtering section is established on the left side of the page (multiple slicers), containing time filters (Year Quarter and Month) as well as product specific characteristics (Category Subcategory and Name). On the top of the page we also have 3 cards that specify measures which change regarding the filters selected (Total Sales Amount Average per Product Subcategory, Total Sales Amount, Total Sales Quantity).

The line chart presents the evolution of total sales amount, comparing it between each category of products, in this way we can reach a quick overview regarding the most relevant categories of products along the period under analysis. The bar chart on the right side, represents the evolution of total quantities throughout time instead. For both of these graphs an average line was added in order to identify easily which periods of time (months) are above or under the average of the metric under analysis.

The 2 horizontal bar charts below represent the top categories and subcategories in terms of quantities delivered. Lastly a multi-row card was also added in order to identify and filter the TOP 10 product names when considering the total sales amount measure, thus we can see in specific without any filtering which products were the most sold ones.

Business questions

By analyzing this page it's possible to answer some key questions:

(1)What are the **most common purchased categories/subcategories** of products per year?

- **Categories** - Electronics, Clothing and Accessories
- **Subcategories** - Smartwatches, Gaming Accessories

The last evidence was verified for all years under analysis when looking at the both horizontal bar charts.

(2)Are there any specific **product categories that have significantly higher sales and quantities**?

- By analyzing the line chart, the 'Electronics' product category showed without any doubt its dimension regarding sales and quantities purchased (when considering 'Total Sales amount' measure).

(3)What were the **Top selling products per each year/quarter/month**?

- This can easily be verified by looking at the multi-row card when filtering the desired year, quarter or month. This card shows the top 10 products regarding the highest values for ‘Total Sales amount’ measure.

(4)How was the **evolution of total quantities sold** during the years under analysis?

- By analyzing the vertical bar chart it was verified an evident pattern of a little decrease in quantities sold for February of all years under analysis.

‘Suppliers Page’:

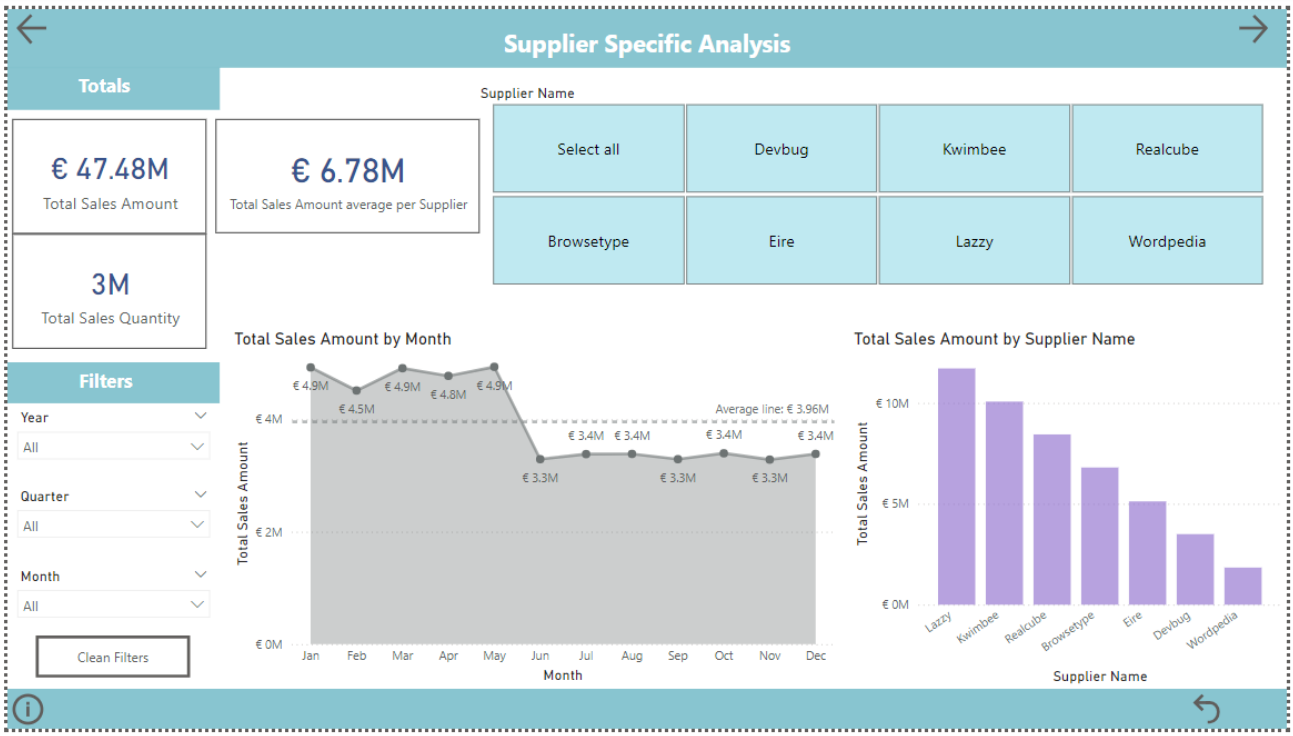

Figure 48- Supplier Report Page

The "Supplier Specific Analysis" dashboard in our team's Power BI offers several strategic advantages for Retail4ALL, it can help optimize its supplier relationships and overall business performance.

This dashboard enables the company to gain a clear and immediate understanding of each supplier's contribution to overall sales. By displaying metrics like the "Total Sales Amount", “Total Sales Quantity” and the "Total Sales Amount average per Supplier," companies can quickly identify which suppliers are driving the most revenue and which may need attention or improvement. This visibility allows for better negotiation strategies, more informed decisions regarding supplier partnerships, and the ability to allocate resources more effectively to support high-performing suppliers or address issues with underperforming ones.

The interactive nature of the dashboard, especially with its supplier filter options, and time filter options (year, quarter and month) empowers users to drill down into specific supplier data with ease. This feature is crucial for making data-driven decisions as it

allows managers to assess performance on a granular level. For instance, the ability to isolate and analyze individual supplier data can help in understanding specific trends, opportunities, or risks associated with each supplier. This detailed insight supports strategic planning and operational adjustments tailored to each supplier's performance. The visualization tools within the dashboard, such as the "Total Sales Amount by Month" line graph, provide valuable insights into sales trends over time. This can be incredibly useful for forecasting and planning purposes. By recognizing patterns and identifying peak periods or downturns in sales, companies can better anticipate future needs, adjust inventory levels, plan marketing strategies, and manage supply chain logistics more effectively, ensuring the company remains agile and responsive to market dynamics.
Overall, the "Supplier Specific Analysis" dashboard serves as a powerful tool for enhancing supplier management, fostering data-driven decision-making, and facilitating accurate trend analysis, all of which are essential for maintaining a competitive edge in today's dynamic business environment.

Below are the example of 3 Business Questions that could be answered with this dashboard:

(1) Which suppliers are the **top contributors to our total sales?**

- The bar chart "Total Sales Amount by Supplier Name" ranks suppliers, helping identify the top contributors, which are Lazzy and Kwimbee, overall and every year if we filter by year.

(2)How does the **average sales amount per supplier** compare across our supplier base?

- The KPI for "Total Sales Amount average per Supplier" reveals the mean sales per supplier, 6.78M, this shows that there are 3 Suppliers that are below average, Eire, Devbug and Wordpedia.

(3) Do any suppliers **show significant seasonal or monthly variation in sales**?

- By selecting individual suppliers, Retail4ALL can analyze how their sales vary by month. All suppliers have higher sales in the first 5 months of the year.

‘Stores Page’:

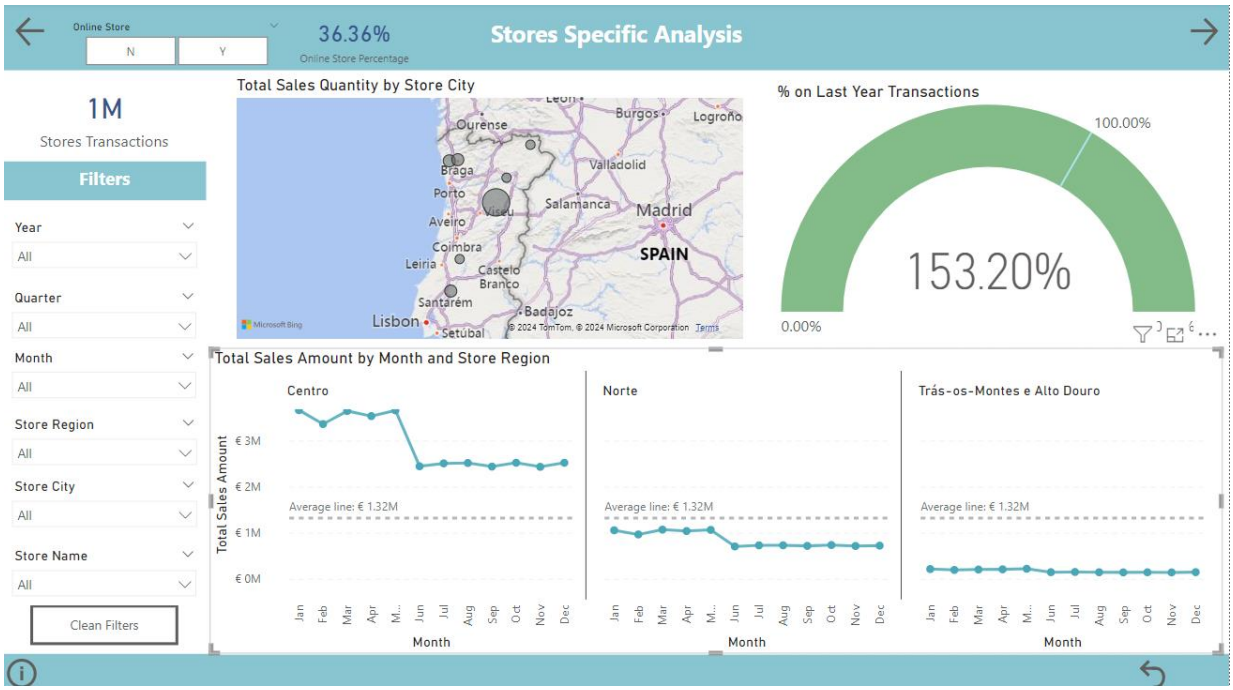

Figure 49- Stores Report Page

The "Stores Specific Analysis" dashboard provides an overview of store-specific sales analysis for Retail4ALL. It helps the company by providing a clear visualization of sales performance across different cities and regions, allowing the company to track which areas are performing well and which are not.

The filters section on the left pane includes options to select the time period for the analysis, such as year, quarter, and month, as well as filters to focus on specific regions, cities, or individual stores. There is also an online store button that allows filtering between online and offline store transactions.

Key metrics and visuals are displayed, including the percentage of transactions from the online store giving a quick view of the representation of this chanell; a map visualization showing sales quantity across different cities for Stakeholders have an immediate perception of which stores cities are performing better, and, we should also notice that a **Tooltip** was used in this graph, with a pie chart showing the % Sales Amount by Product Category when we pass the cursor by each Store City, allowing to analyze which categories are contributing the most to sales; a gauge showing the percentage increase or decrease in transactions compared to the previous year (explained in the *KPI Visuals* section) enabling the viewer to have a general view if Retail4ALL surpassing the last years performance (in terms of transactions); and line charts showing monthly sales amounts for the three stores regions, with average sales lines, which can help to compare the regions among themselves and also to see the sales evolution for each region. Additionally, the total number of transactions is indicated and a clean filters button resets all filters to default.

This dashboard responds to the following key Business Questions:

(1) Which Stores Cities are the **top performers** in terms of Sales Quantity, for each year?
- The Store City with remains as a top performer for all the years under analysis is Aguiar da Beira.

(2) Is each Store Region and City selling more (sales transactions) **compared to last year?**
- All the Stores Cities and Cities are selling more than last year. This shows us a general positive evolution of the business, when looking at the bigger picture.

(3) Are the Store Regions having an **increasing revenue trend** (sales amount) throughout the year?
- We can conclude that, for every Store Region, they have a similar behaviour where the Sales Amount decreases from May to June. There is the need to understand what might be causing this trend and act based on it.

'Operators' Page:

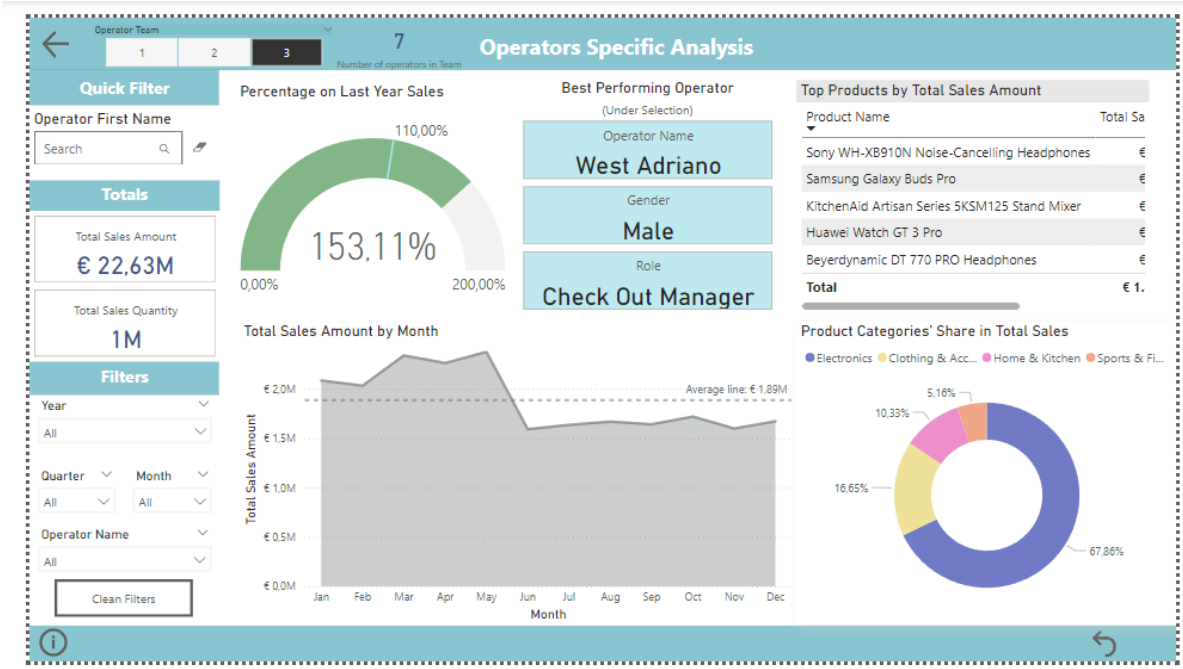

Figure 50- Operators Report Page

This page is used to benchmark each team's performance as well as each individual operator. As so, on the top left we have a filter for each team (which will be restricted according to the section *Extra Requirements*). We felt it was important to include the number of people in the team to enhance human resource management. The quick filter is also explained in the *Extra Requirements* section. For the totals' cards we give a visual of both quantities and volume of sales associated with the selection in force. On

the bottom left we include the filters section, which gives flexibility in the analysis and promotes evaluations over time.

The top row of visuals are of high importance. We have an evaluation KPI that allows us to quickly understand performance (*KPI Visuals* section). Then to directly understand the business question of what are the top workers we build cards to showcase that evaluation. Notice that the choice of only considering the best worker was made after finding out that teams are really small and we suggest having just the top worker will motivate others to improve their performance. The table on the right allows us to understand the core 5 products associated with each worker, which allows RETAIL4ALL to tailor better segmentation strategies on where to place operators or where to increase their skills on selling certain products. The donut chart on the bottom reflects this core ability on a broader level reflecting the product category associated distribution associated with each team. We kept it at the granularity of category as subcategories are too many and would add too much noise to understand the analysis. We believe this page is a key element in improving RETAIL4ALL human capital management and overall performance.

Apart from other **business questions addressed**, we identified these as the main ones:

(1) Which operator is the **top contributor to our total sales?**

- The Best Performing Operator Cards show that West Adriano is the best operator for team 3, Sibyl Scintsburt for team 2 and Stern Burgyn for team 1.

(2) What are the **5-Core products** per team of operators? Are there any common top products?

- The table on the right identifies the 5 best selling products of each team. For example, what team 3 sells the most is: *Sony WH-XB910N Noise-Cancelling Headphones; Samsung Galaxy Buds Pro; KitchenAid Artisan Series 5KSM125 Stand Mixer; Huawei Watch GT 3 Pro* and *Beyerdynamic DT 770 PRO Headphones.*
- Every team has different products as their top 5.

(3) Does any operator **show significant seasonal or monthly variation in sales**?

- By using the filters to understand each individual operator, after observing the line chart on the bottom, it is clear that every operator has a seasonal pattern of sales, always reaching a peak in the second quarter of the year (around May) and keeping sales always below average for the rest of the year (June-December).

Conclusions

Along the development of the project report we achieved the following conclusions:

- It's noticeable that the 'Electronics' category of products has the highest volume of sales in comparison with other categories, thus we can have 2 interpretations regarding this. In one way Retail4All can focus its production on this niche of market, or can spread the business to other categories of products that were not so impactful until now.
- Microsoft Fabric was without any doubt a good tool in terms of data organization and consistency. The sequential steps followed to transform data in a proper way for the reporting stage, required a deep understanding of the data, regarding which transformations should be made to each dimension or fact table.
- In the reporting phase we kept the idea of maintaining a user-centric analysis, focusing on what would be the valuable insights gained for the user from the way we are representing the data, and in what way these insights can overcome the initial problems of Retail4All.
- The overall goal of the project was achieved, in the sense that throughout the use of Microsoft power BI we understand patterns, evolutions and even predict what would be the future evolution of some measures.

Finally, we can agree that Retail4All leveraged its own business understanding by adopting our Business Intelligence solution. The report we present answers key questions to perform Tactical and Strategic planning and communicates effectively through carefully chosen visualizations, enabling a quick visualization of the main insights to be digested and acted on.

References

[1]Draw.io. (2024). Flowchart Maker & Online Diagram Software. App.diagrams.net. https://app.diagrams.net/
[2]Kimball's Dimensional Data Modeling | The Analytics Setup Guidebook. (n.d.). Www.holistics.io. https://www.holistics.io/books/setup-analytics/kimball-s-dimensional-data-modeling/
[3]Class contents, available on the online platform Moodle.
[4]Mordor intelligence. (2018). Mordor intelligence. Mordorintelligence.com. https://www.mordorintelligence.com/industry-reports/retail-industry

Annexes

Excel Source	Variable	Description
Products	SKU	Code associated with each product name (unique identifier)
	Product Name	Name of the product (high variety of options)
	Category	Category associated with each product name (4 unique categories)
	SubCategory	SubCategory associated with each product name (31 options)
Stores A	Store ID	Unique identifier of the store (20 stores)
	Name	Name of the store
	Location	Location of the store (ID associated with the Locations table)
	Online	Identify if the purchase was made online or not (S:Yes, N:No)
Stores B	Store ID	Unique identifier of the store (11 stores)
	Name	Name of the store
	Location	Location of the store (ID associated with the Locations table)
	Online	Identify if the purchase was made online or not (S:Yes, N:No)
Locations	Location ID	Unique identifier of the Location
	Country	Country Name
	Region	Region name
	City	City Name
Sales	Sale ID	ID associated to each sale
	Datetime	Date associated to each sale (DD/MM/YYYY HH:HH)
	SKU	Code associated to each product name
	Store	Unique identifier of the Store
	POS	Unique identifier of the Supplier
	Location	Unique identifier of the Location
	Currency	Currency used (EURO in all cases)
	Qty	Quantity purchased associated to each sale
	Amount	Sales amount
	Operator ID	Unique identifier of the Operator
Operators	Operator ID	Unique identifier of the Operator
	First_Name	First name of the operator
	Last_Name	Last name of the operator
	Email	E-mail of the operator
	Gender	Gender of the operator (Female,Male,Agender)
	Role	Role of the operator (5 different types of roles)
	Team	Team of the operator (3 possible options 1,2 or 3)
Point of Supply	POS	Unique identifier of the Supplier
	Name	Name of the supplier
	Email	E-mail of the supplier

Figure 51- Data Sources Description