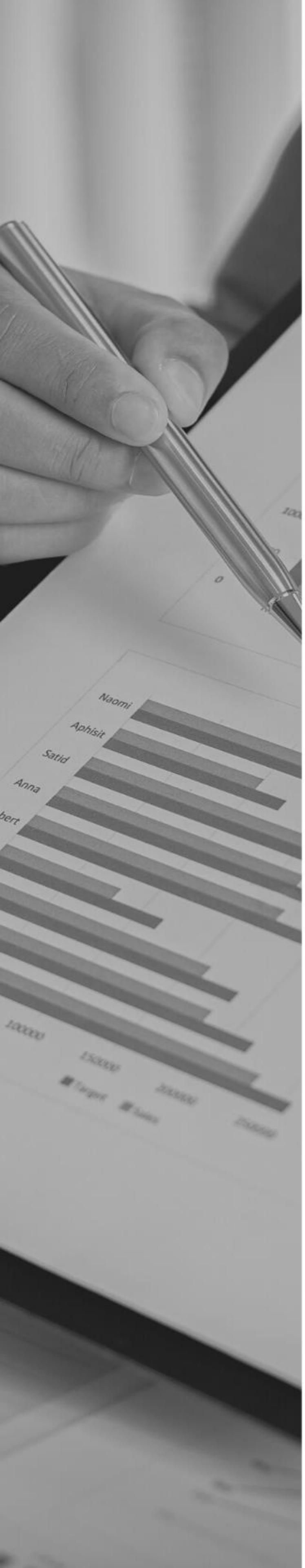


NOVA IMS
Information Management School

bi4all
TURNING DATA INTO INSIGHTS

BUSINESS INTELLIGENCE

2.º DELIVERY

BI Practical Project

JUNE 2022

GROUP 22

Alice Vale R20181074
Eva Ferrer R20181110
Rafael Sequeira R20181128
Raquel Sousa R20181102

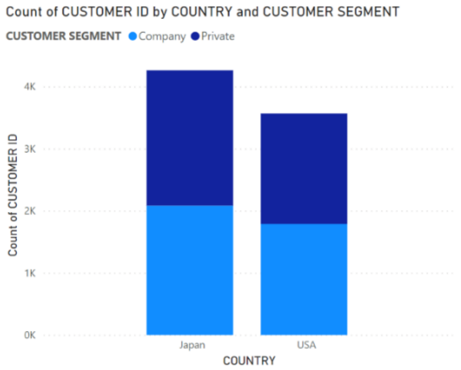
I. Contents

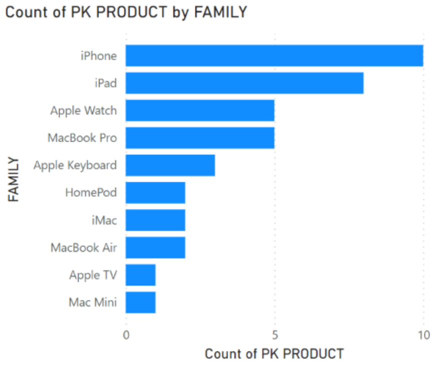
II. Introduction & Presentation of Business

The project focuses on an online Portuguese company, founded in 2020, which commercialized technology products internationally, considered today as a well-known European *Apple* re-seller. With the aim of expanding its market and potentially the catalog of products available, the decision was to hire a team to develop a visual element in *PowerBI*.

The proposed objective for this project is to provide the company in question a better understanding of their business, so to enhance the decision-making process and potentiate growth. After dialogue with the stakeholders of the enterprise, their intentions and requirements became settled and will be discussed throughout this report. A bundle of transaction data was provided, with different natures, so to be used in the creation of the dashboard solution requested.

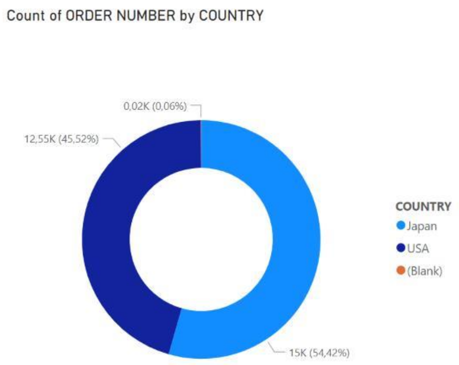
This major technology retail company mainly sells to the US and Japan, handling both private and corporate clients, with a very balanced distribution. In the dataset provided there are 7829 unique customers present.

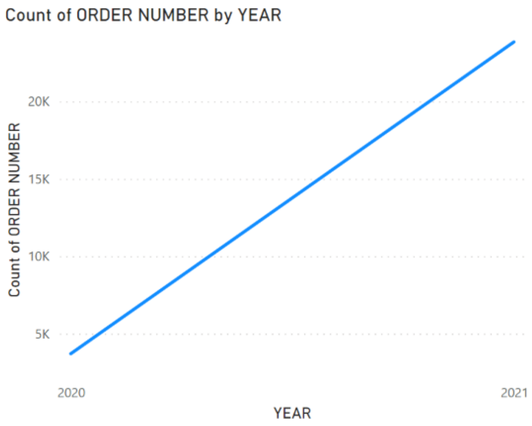

Count of CUSTOMER ID by COUNTRY and CUSTOMER SEGMENT


Count of PK PRODUCT by FAMILY

In terms of products commercialized, there are 39 unique products all from the Apple ecosystem, with release dates between 2020 and late 2021. The top 3 item families are iPhone, iPad and Apple Watch, which seem to be the items that are launched more frequently. The physical stores are divided between Lisbon and Porto, and, for each, there is information on the target revenue per month for the year of 2020.

In addition, the last information provided regards the order invoices of past purchases. There is information about the order and delivery date, the customer and store that purchase relates to along with the unit and delivery price, quantity of the products ordered, and tax rate applied, all in euros. There is a total of 27560 distinct orders present. Japan has the highest number of orders, and the growth in sales from 2020 to 2021 is extremely positive, as it can be seen in the chart below.





To note that there are 17 distinct orders for which there is no information about the country of origin. They correspond to three customers, with the respective customer identifiers as 4988, 1206 and 7006, which are not part of the customer database provided, despite having purchased from the company before.

III. Identification of Business Needs & Problem

The Data Warehousing solution was required by a major technology retail company that shall remain nameless for data privacy concerns. They contacted *BI4ALL* to create a visual solution that could be easily consulted on a daily basis to follow their operations. To better understand the necessities of this company and how to potentiate their effective definition of strategies, it is crucial to explicit what problem they aim to solve as well as to identify their business needs.

The company's current main informational issue is that it has massive amounts of unconnected transactional data, as form of *Invoices*, that are segmented by store. This overload of raw data makes analysis very difficult, and consequently delays or makes any decision-making strategy impossible. Moreover, the company struggles with taking basic insights on their customers, stores' or products' performances, since they cannot compare or evaluate how they behave. Finally, it is not currently possible to get an overview of the entire business, which affects how information is streamed within departments and accessed as a whole.

Each company's department has different necessities, depending on their future use of the solution. There is why there are three different perspectives that will be further explained bellow: the commercial, the sales and the logistic. Nonetheless, the overall implementation of the DW aims to provide the company with the capability of having a daily view of Product Sales and Costs in all their Portuguese online stores.

Perspectives

- The commercial department needs Sales and Costs indicators, namely what are the daily monetary sales values, with and without taxes, per geographic location. They require to get quickly access to an overview of costs' and profit margins' variations, so not to spend unnecessary time calculating those metrics. The ability to predict Sales for the following months are an intended extra feature. In addition, time analysis is necessary, making use of historical data to calculate sales volume, costs, and gross margin volume, among other possible indicators, by year, month, week, day, etc. The goal is to provide tools that can be adapted in the future to different granularities, depending on the current necessities of the department. To facilitate data analysis and visualizations, KPIs need to be provided, available to be called and ready to be highlighted. All the information ought to be displayed both in Euros and US dollars, to facilitate sharing information with business partners, since the company has partnerships based in the United States of America.

- The sales department needs to compare performances for each store, and to verify if the total sales are reaching their target value. As well as to compare sales by product, to evaluate if products are achieving their target value.

- The logistic department needs two different types of approaches. For orders already delivered, they want to understand the time it takes for an order to be delivered. For not delivered orders, they require to know how many orders are in that scope. On top of that, it would be beneficial for them to get an overview of Customers' Profile and Sales patterns, without getting into too many details, but still so they can better understand the business and make more informed decisions.

IV. Description of the Data Source & Discovery Process

To develop the necessary *Power BI* solution, the retail company provided the team with data regarding the transactional operations of their stores and other business dimensions. A total of sixteen excel files and one text file were supplied. Following, a description of the data in said files can be found, so to fully grasp what information is available and its potential for the deployment of the solution. It is relevant to point out that the data concerns the years of 2020, 2021 and 2022. Some insights were also gathered throughout the initial data discovery process.

Invoice Details

This folder comprises information relating to detailed characteristics of the purchasing invoices. Within it are five different files concerning the five stores of the company. The variables are the product identification number, the order number and quantity, the unit price of the goods that were purchased, the delivery cost and the tax rate applied to each specific order. During the discovery process, it was learnt that the largest quantity of products ever bought together was 29 and the tax rate most commonly applied to orders is 23%.

Invoices

This folder contains more traditional data about the purchasing invoices. Like the previous digital dossier, the information is divided into five different files representing each store. Inside each one is found the date of the placement of the order and of its delivery, the identification of the customer who made the purchase, the identification of the store where the products were sold from and the order number. It was detected that the store where most orders are placed is the store with *ID* equal to one.

Customers

It holds information on the company's previous clients and their characteristics. For example, the customer identification number, their first and last name, what type of client segment they belong to (if the client is a company or a private entity), their zip code, which indicates their location, and their phone number. It was discovered that customers are usually based on two single countries, either Japan or the United States of America.

Family Images

This a simple file containing image data on the products. It refers to the product family name (ex. *IPad*, *Apple Keyboard*) and its corresponding link to an image that represents the products in each group.

Location Detail

It comprises data regarding different locales, typically possible customer addresses. Information such as the name of each city and the corresponding state, region, district, if exists, and country can be found inside this file. Considering American clients, it was analysed that they are usually from the West, East, or Central geographical zones of the USA.

Product

Carries information regarding the characteristics or details about the products sold by the company. The variables are the release date of the products, the code of the corresponding model and the actual model, that can be interpreted as the name of the product. There is also the family name to which each product belongs to and the date at which the product was removed from the market if such is the case. It was easily discovered that the most common family of products in the store is *iPhone,* possibly indicating that goods that fall under this category are the most predominant in the company's stores.

Store

Contains data concerning the five Portuguese stores of the company, as for example, the city where the store is located, which is either Lisbon or Oporto, the name of the store, its email address, the geographical address, and its phone number. Taking into consideration the conclusion reached on *Invoices*, the store with most orders is *Lisbon Shopping 1.*

Target

It represents the expected profit target for 2020. So, this file holds information on the store identification number, the date of each month of the mentioned year and the monetary value that is projected for each store to earn monthly. Every store was expected to surpass a revenue above one million Euros in at least one of the months of 2020.

Conversion EUR/USD

A text file holding support information on the monthly conversion rate between Euro and American dollar in 2020 and 2021. Such data functions as a catalogue and will be used to facilitate currency changes since the company has partners in the United States of America.

V. Data Modelling Methodology

After the insertion of the different files provided by the company into *PowerBI* it was imperative to start the modelling process. The *Entity Relationship* model currently in use was not provided, so it was not possible to build the data warehouse with the use of that information. Therefore, the *Kimball methodology* [1]was applied, which consists of four different steps.

1. *Selection of the business process*

Through the notes on the meeting with the company, it was clear that the targeted business processes were related to Sales and Order Fulfillment, as already addressed in the *Business Needs* chapter, each searching for a different view of the data portrayed. There are many internal Business Processes that allow the good functioning of the company under study; Nonetheless, for the scope of this project, only a few processes will be detailly explored.

The data provided currently supports from fairly simple and rudimentary processes, such as order and invoicing processing, or processes held with procurement partners, such as account receivable or invoices reconciliation; However, it is trying to expand the scalability of information insights to more complex analysis, namely with predictive capabilities. The commercial and sales departments are studying how the product delivery process can be optimized, taking into account the evaluation of important underlying factors, that are in itself business processes, like cost estimating, billing reviewing and sales achievements according to the defined target.

2. *Declare the grain*

The data provided presents different grains:

- Customer – one row per unique customer identifier
- Product – one row per unique product identifier
- Store – one row per store
- Invoices – one row per product per order

3. *Identify the dimensions*

In the modeling stage, the decision was to create six different dimension tables, each representing static information to be accessed for analysis. The time dimension was built from scratch inside *PowerBI*, and it stores information about dates such as day of the week, quarter, year among others, which will prove relevant when presenting information visually.

[1] Further information on the topic can be consulted in the Kimball Group website

Based on the information provided, dimensions for Store, Customer and Product were created, which account for information about each individually actor in the business process and relevant characteristics such as location of the stores, name of customers or family of products.

The last two dimensions gathered regard the conversion rate from euros to dollars, since one of the requirements from the company was to be able to present information in both currencies, adapting the dashboard to specific viewers; and a target dimension which stores the sales target for specific months of 2020.

4. Identify the facts

As of the time of this intermediary delivery, there is only one fact table present in the model created in which each row represents an order item, identifying the customer and store the purchase belongs to. It holds time information regarding the moment of the order and the respective delivery, if fulfilled. Depending on the next steps taken to achieve the best possible solution, it is expected to add some more facts to the dimensional model constructed, as well as more features.

VI. Model Optimization

Taking into consideration the feedback from the intermediary presentation of the Power BI solution developed by our team, some updates were made so to improve the final result and guarantee that the needs of the company for decision-making support across various departments was met.

Conversion EUR USD was previously linked to *Dim Date* (Denominated *Dim Time* in the first delivery) through the *Date* Column, which acted similarly to a primary key. Since this relationship was not efficient as it was, several transformations were applied to now allow the connection between the tables to be done by means of a foreign surrogate key that was created through *Dim Date.* Their relationship continues to be of type *One-to-One,* however, because of this new-found key, suitability and effectiveness are ensured. The first table, which previously carried fields such as *Date*, *Euro*, and *Dollar* now holds the key that certifies the relationship and the corresponding conversion rate values.

Next, for organization purposes and to respect the usual design requirements of dimensional modelling, all attributes operating as linking keys were given to corresponding nomenclature according to their purpose. This means that, for example, fields working as foreign keys were given the *FK* prefix. This change was applied to all types of keys, namely, primary keys, foreign keys, and surrogate keys so to secure naming consistency across all attributes.

It is relevant to mention that fields of analogous type to *Order Date* were formatted as *Short Date* for simplicity purposes and *Don't Summarize* was applied to avoid aggregating the attributes' values, which would, otherwise, be meaningless for the final solution. Moreover, every field corresponding to ratios was formatted as a percentage with two decimal cases for readability purposes.

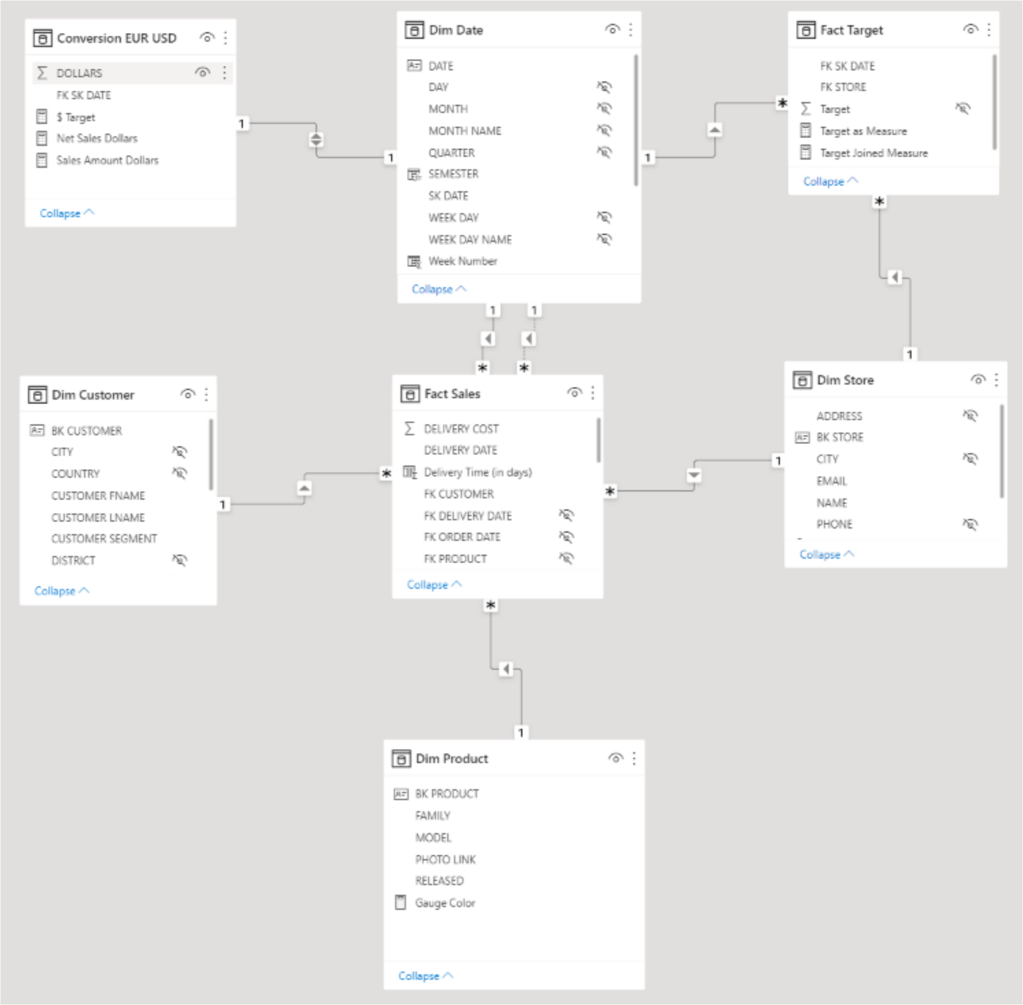
The new dimensional model is composed by the same tables, connected by identical type of relationships as in the initial model, however, with major changes in organization.

Firstly, the classification of *Conversion EUR USD* as a dimension table was removed since it merely functions as an auxiliar table to hold information on conversion rates and is not an actual dimension of the business scope. Two measures are held in this table, being those the *Net Sales* and the *Sales Amount* in Dollars. The *Customer Dimension* holds fields such as the customer's first and last name and segment. Moreover, a hierarchy was created, *Customer Location*, to rank the different attributes concerning the customer's location, beginning with *Country, Region, State, District* and finally *City.*

Dim Date contains a straightforward hierarchy, *Date*, to create a rating of the multiple time fields it holds, starting with *Year*, followed by *Quarter, Month, Month Name* and *Day*. *Semester* and *Week Number* are produced attributes to retrieve the semester and number of the week in which the data is included. The *Product* dimension holds the same information as before, being all fields visible in the table, along with a generated measure, *Gauge Color*, to categorize by color if the sales amount is higher or lower than the corresponding target, respectively.

The *Store* dimension contains attributes related to the different stores, nonetheless, a hierarchy was created, *Store Location,* to hold this type of data in a more organized way, namely the city and the address of each store locale. *Fact Target* holds the target values along with three measures created: *Target as Measure, Target Joined Measure* and *Target Measure*.

Fact Sales, the main table in the dimensional schema, holds several different measures or generated fields that will be used to support the creation of the visual report. Attributes like *Delivery Date*, *Delivery Cost* and *Tax Rate* are also part of it. Since this is the central table in the model, it contains multiple foreign keys that establish the relationships between tables. Since this are not important in the visualization phase, they are hidden from the report view, along with other non-relevant attributes. Finally, to facilitate the exploration of the dimensional model in the *model view,* different tabs were created to better showcase each of the several relationships. There is a general tab where the entirety of the model can be viewed and a *Target, Sales* and *Conversion EUR USD* tab with the connections made to those tables.



VII. Perspectives of Analysis

To organize the visualization and facilitate the decision-making process of business stakeholders, three specialized perspectives were constructed, specifically, *Commercial*, *Sales* and *Logistic*, along with an *Overview* page. These views were idealized having in mind not only the goals and business questions that the company seeks to answer but also the departments that will take advantage of this tool, daily. In the left column there are icons allusive to each perspective that allow the user to change between views. They were chosen since they resemble the overall theme of their respective page and make the report intuitive to explore. The design choices and general organization of the dashboard were inspired by a previously conceived dashboard with similar visualization goals and can be viewed here.

The *Overview* tab gives the user a general comprehension of the current state of the business. The person can choose to either view the data for all stores at once and for the whole time period that is available or select specifics for both filters. The visualizations will change accordingly. There is also an option to decide which currency makes more sense for the data to be displayed in, which can either be Euro or Dollar. Finally, the user can easily remove all filters applied through the "*No filters*" button. Displayed are the *Sales Amount & Target per Month*, the *Net Sales & Costs by Year* and the *Sales Prediction*.

The *Commercial* perspective displays information that will be useful for said department, such as an overview of sales, profit and gross margins and a geographic view of the sales. The user can choose which store and time period to view, or all at once. This will be extremely advantageous to verify the current or past industrial reality of the company itself or of a specific store under analysis.

The *Sales* view comprises information on the products themselves and sales amounts comparatively to the target. Here the user is able to select a specific store or time period with the same visualization goals as before, however, a new filtering option is present in this view: *Product*. This tool allows for the information to be displayed for a specific product sold by the company which is valuable if its performance in terms of sales and profit is being examined.

The *Logistic* perspective was developed to showcase information that might be useful for the departments related to deliveries and orders. Displayed are the average delivery time, number of orders, customer overviews, among others. Yet again, the user can filter by store or time period and, exclusively for this tab, the specific customer to investigate. Such filtering options will facilitate the investigation of customers with, for example, unusual buying behaviors or how each store is doing in terms of order deliveries.

VIII. Measures & Calculated Columns

FACT SALES

Sales Amount: For each order, it expresses the product between the unit price and the order quantity, so the monetary value of the transaction.

Sales Amount Taxes: For each order, it expresses the monetary value of the order including the tax rate value.

Net Sales Measure: For each order, it returns the Profit, meaning that it subtracts the Costs from the Revenue, in this case, it's the Sales Amount minus the Delivery Cost.

Profit Margin: It is the results from dividing the Net Sales Measure and the Sales Amount. It is represented as a percentage of Profit for each transaction.

% Change Total Amount: It is a comparative measure that computes the ratio of change, the percentual difference, between the values of the current month with the previous month for the Sales Amount. So, it expresses whether the company is comparatively invoicing more or less.

% Costs: For each order, what is the percentage of the Delivery Costs in the Sales Amount. This is particularly relevant to understand production costs variations.

MTD Gross Margin Volume: It expresses the month to date value for the Net Sales Measure, so to measure the accumulative trend of these values. It is particularly relevant to perform comparisons among months and within years.

MTD Sales Volume: It expresses the month to date value for the Order Quantity, so to measure the accumulative trend of number of products ordered.

$ Delivery Cost: It is an estimation of the Delivery Costs in USD, so to display results in this currency.

Delivery Time (in days): It is the number of days from the Order Date until the Delivery Date.

No Delivery: It is a binary column that translate whether an order was delivered or not. It is based on the column Delivery Time, considering that if such observation is null, then the order is yet to be delivered.

FACT TARGET

Target Measure: It translates the target sales value having in account the formula provided by the company's representatives, which is an increment of 25% from the homologous period.

Target as Measure: Transforming the target column for 2020 values as a measure so to later apply data aggregations.

Target Joined Measure: Adding the target values for both 2020 and 2021 in a single measure, so to perform KPI's.

DIM PRODUCT

Gauge Color: It is an auxiliar measure to be incorporated into a Tooltip visualization, so to perform change the data color based on a field value. It attributes the color light green if the Sales Amount is higher than the Target Measure, and it attributes Light Red otherwise.

DIM DATE

Semester: It attributes the label "Semester 01" when the correspondent month is inferior to 7, and "Semester 02" otherwise.

Week Number: It applies a function that returns the number of the week in the calendar to each corresponding date.

CONVERSION EUR USD

Net Sales Dollars: It converts the Net Sales Measure into USD depending on the conversion tax for the associated date.

Sales Amount Dollars: It converts the Sales Amount into USD depending on the conversion tax for the associated date.

$ Target: It converts the Target Joined Measure into USD depending on the conversion tax for the associated date.

IX. Reports' Main Technical Aspects

Starting with the overview page of the dashboard, which is composed of two KPIs and a Line chart. The first KPI compares Sales Amount with the sales Target, showing off a green color when the Sales are greater than the target and red otherwise. On the right side of the slide, a KPI is comparing Net Sales and delivery costs, very useful to see how an increase in sales affects the delivery costs, and on the bottom of the slide, there's a Line chart of the plotting net sales and the predictions. While defining the parameters to Predict the Sales for the upcoming month, the seasonality value was set to 90 days, because we have data on about one year worth of sales, and we should have about four times more observations than what we set the seasonality value as, likewise 365/4=91,25 which was rounded to 90.

Moving on to the second tab, the 'Commercial' perspective. In this perspective, we have a total of 4 visualizations including a bar chart with a line chart as a secondary axis, a map, an area chart, and a clustered column chart. The bar plot is comparing the sales amount, with and without tax, and the average tax rate over a determined period of time, the map is showing the total sales per region, where it's possible to get more detailed information on product family sales by visualizing the *Product Tooltip* of each country. The area chart is comparing Profit Margin and % of costs, again using a secondary axis to take into consideration the scales of the features and the clustered bar chart is displaying the MTD Gross Margin with *Sales and Costs Volume* as a tooltip.

Regarding the third tab, which focuses on Sales, four visualizations are available, including a visualization displaying images of the products sold by the store to increase context. This slide is composed of a horizontal bar plot comparing the Sales Amount and the respective target, and a bar chart measuring the Sales Amount of each product, in this graphic, the X-axis is the Product ID and the color represent product category, making it easy to observe which products and categories tend to sell more, and finally, a table displaying the Total Unit price, the order quantity, and a flagship indicating whether or not the product beat the expected target. In the table presenting the Product analysis, we decided to display the Product ID instead of the Product Name so to facilitate reading information as an employee out of the analysis, and so that information is straightforward to spot when searching for specific products.

The last tab of our dashboard, which refers to the logistic department, has a total of 4 visualizations, including a Sankey Chart indicating the average delivery time by store, a donut chart displaying the proportion of orders delivered vs not delivered, a stacked bar chart indicating the % change of sales from one month to another and multiple cards showing the total sales by customer. The stacked bar chart has month names on X-axis and if the % of change is smaller than 0 then the bar displayed will be red and blue otherwise.

X. Analysis and Discussion of Output

Considering the objectives stated in the scope of the project developed, our dashboard solution provides a simple and informative overview of the current business status, for all three perspectives required: Commercial, Sales and Logistic. Covering the disperse and unconnected data problem, it provides answers to all the questions proposed by the stakeholders.

In the Overview tab, KPIs can promptly be analyzed such as a monthly comparison between the sales amount and respective target or annual net sale versus costs, along with the highlight of the most recent data. Both reveal a positive outcome, concluding that the business is growing beyond expected, registering today 3 times more than the amount in sales registered in 2020 – around 8 million euros – value at which the targets were settled, achieving in 2021 a total of around 20 million euros. Another required aspect was the prediction of sales which reveals a decreasing tendency in the predicted sales for the month of January 2022. All information can also be converted to US Dollars, with an easy click on the respective icon, being Euros the default currency.

In the Commercial tab a daily overview of sales with and without taxes can be consulted, along with a time analysis of the tax rate, as requested by the stakeholders. The values are almost unchanging, even with sudden changes of the tax rate, with a slight decrease registered in sales towards the latter part of the month. This can be filtered geographically and show data only for Japan or US. The distribution of sales per family of products is also shown, revealing a predominance of purchase for the iPhone, Apple Watch and iMac families of products. A daily evolution of the profit margin and costs variation is also portrayed which indicates that the higher profit margin is registered on the 15th day of the month, followed by a decreasing tendency. Lastly, a net sales representation per month is plotted, which shows the higher sales volume registered in 2021 in comparison with the previous year. Hovering through the chart, a comparison between order quantity and delivery cost shows - highly correlated features.

Moving to the Sales, it is possible to compare the sales amount per store or/and per product with its respective target, as it was sought by this department. Store 1 was the one which registered the highest sales amount whereas Store 5 the lowest, all however surpassing the target values appointed. All products have also surpassed the targets, with the iMac and Apple Watch family products having the most significant sales registered. When hovering through the bars, a gauge tooltip appears which shows the comparison between target, registered value and the proportion to the total of sales in that product family. The final item in this tab is a table that shows the order quantity and profit margin per family product with the option to deepen the analysis to a specific product, with a flag indicator of the satisfaction regarding the profit margin registered. The bottom line chosen were 80%.

Lastly, for the Logistic perspective it was crucial to present information both for delivered and undelivered parcels. Therefore, the average delivery time per store is shown, being store number 2 the one with the most delayed deliveries. Store number 1 is the fastest with transportation, being also the one with the most sales, as stated before. To note that odd values for delivery time, such as 684 days, were encountered and may be skewing the average, however the decision was to maintain them and discuss with the stakeholders how to treat these going forward. An analysis of the percentual change of sales per month reveals that the first semester of the year registered the most significant changes, however all the changes registered were positive. There is also a graphical representation of delivered versus not delivered orders, the latter registering less than 1%. There is also a customer overview regarding the four best customers and their individual net sales.

XI. Conclusion

The dashboard created is ready to be deployed and used as a daily tracking tool for the specific departments who requested it. It is extremely easy to add any new data or dimensions to this solution since it is built upon a versatile and adaptable software.

The colors chosen bring harmony to the user experience and the contrast perfectly highlights the most relevant aspects, so to allow for the maximum business understanding in the shortest period of time possible. If the company finds suitable, a mobile version of the dashboard can be created and fine-tuned so to provide a more portable analysis tool.

Based on the data provided, the success of the business is notorious, especially when analyzing the exponential increase in sales and profit in the span of one year, being both the Japanese and American markets equally relevant. Store 1 stands out for both the fastest delivery time and biggest number of sales registered, working therefore as an example for the remaining stores as the ideal working structure. A suggestion would be to, for the most bought families of products, consider selling complementary items such as cases for the MacBook's or straps for the Apple Watches, which would encourage the customer to purchase them together, therefore expanding the catalog and creating additional profit.