NOVA IMS
Information Management School

Storing and Retrieving Data

Homework 2 Report

Group Z

M20190932 Ana Cláudia Alferes
M20190089 Francisco Carujo Neves
M20190935 Luís Filipe Pinho

Teachers

Flávio Pinheiro

Mijail Naranjo Zolotov

The *CLASSICMODELS* database is a retailer of scale models of classic cars. It contains typical business data such as customers, products, sales orders, sales order line items, etc.

After a review of the *CLASSICMODELS* database and looking at what it contains we start making a list of business questions that will be used to understand the business current state and maybe ways to improve the business. These questions will be answered later on this report.

We decided to split the questions based on the different tables in the database and we end up with this:

1. Products:

    1.1 Which product has the most sales?

    1.2 Which product line has the most sales?

    1.3 Which product line has the most quantity on stock?

    1.4 Which product scale has the most sales?

    1.5 Which product vendor has the most sales?

    1.6 What's the difference between priceEach and MSRP? Are we making "better" prices?

    1.7 What are the most profitable products?

2. Customers:

    2.1 Which customer buys the most?

    2.2 Which country buys the most?

3. Offices:

    3.1 Which offices have the most sales?

    3.2 Which country has the most sales?

4. Employees:

    4.1 Which employee has the most sales?

5. Orders:

5.1 What is the “lead time” for an order?

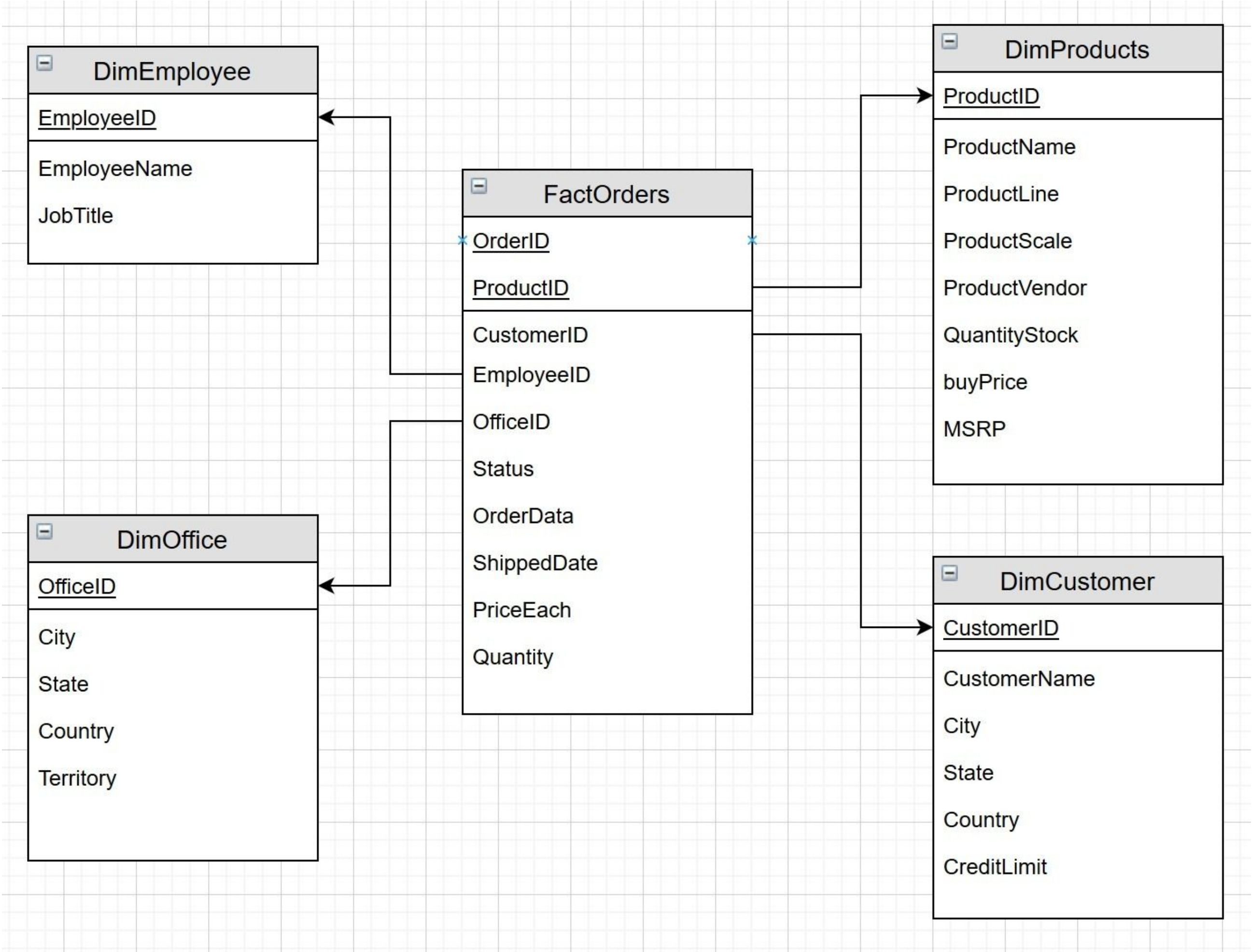
5.2 What is the different status of an order and their ratio? Particular attention on the “canceled”

5.3 What is the mean quantity of products sold in an order?

5.4 What is the mean price gain in an order?

5.5 What is the Total Profit per year?

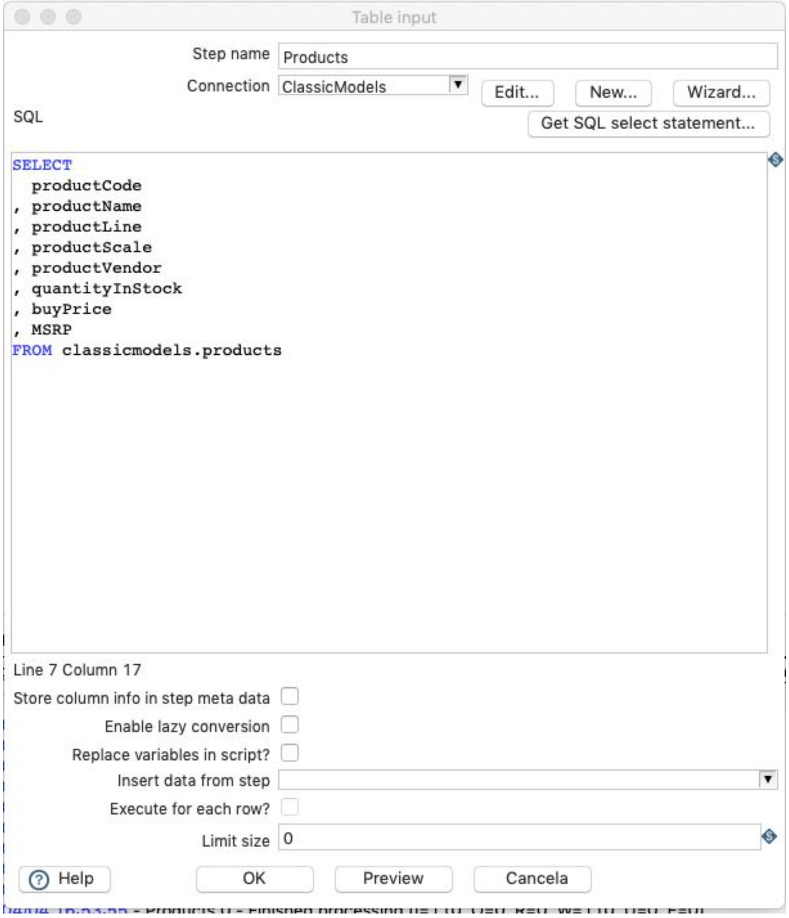
After analyzing the business questions we proceed to design our Star Schema for the Data Warehouse as you can see in the image below.



The next step taken was creating the respective tables in the MySQL Workbench platform the code used is in the "*ClassicModelDW.sql"* file.

After that we opened the Pentaho Spoon platform and made the database connections to the *CLASSICMODELS* database and the *ClassicModelDW* and shared both of them.

We started by making the table dimension ETL, in the exemple present below it's referring to the Products ETL. We began by getting the fields needed in the table Products from the *CLASSICMODELS* database after that we connected to the *ClassicModelDW* and made the connections required to populate our Data Warehouse.



Step name: Products

Connection: ClassicModels

SQL

```
SELECT
  productCode
, productName
, productLine
, productScale
, productVendor
, quantityInStock
, buyPrice
, MSRP
FROM classicmodels.products
```

Line 7 Column 17

Store column info in step meta data □

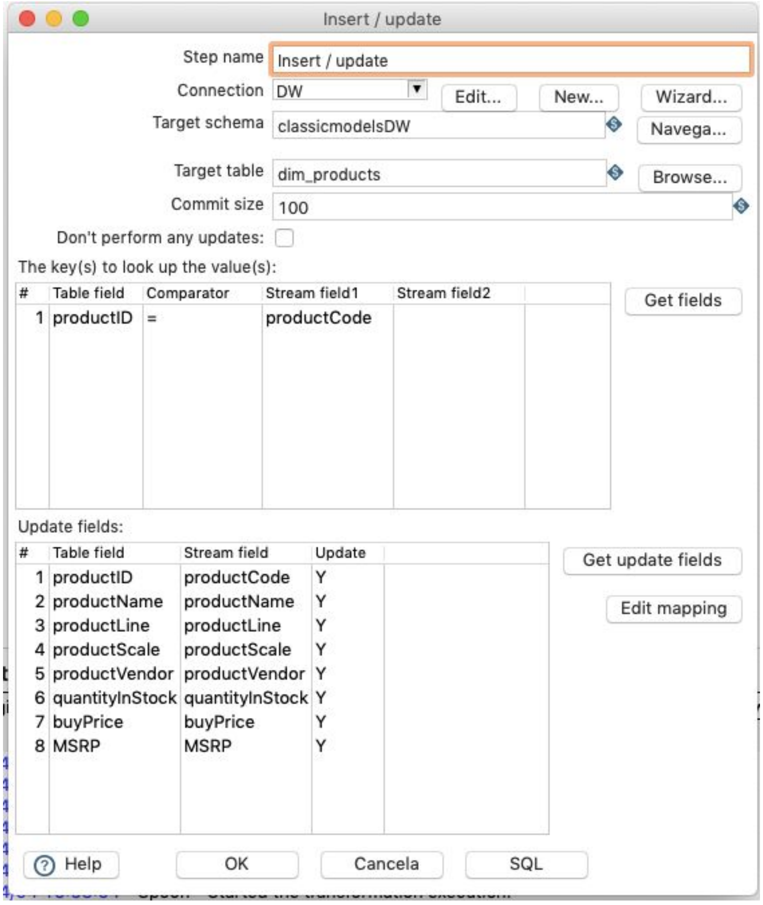
Enable lazy conversion □

Replace variables in script? □

Insert data from step

Execute for each row? □

Limit size: 0



Step name: Insert / update

Connection: DW

Target schema: classicmodelsDW

Target table: dim_products

Commit size: 100

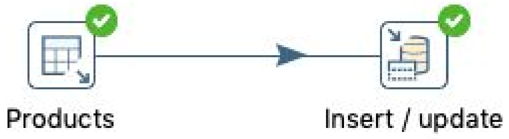
Don't perform any updates: □

The key(s) to look up the value(s):

#	Table field	Comparator	Stream field1	Stream field2
1	productID	=	productCode	

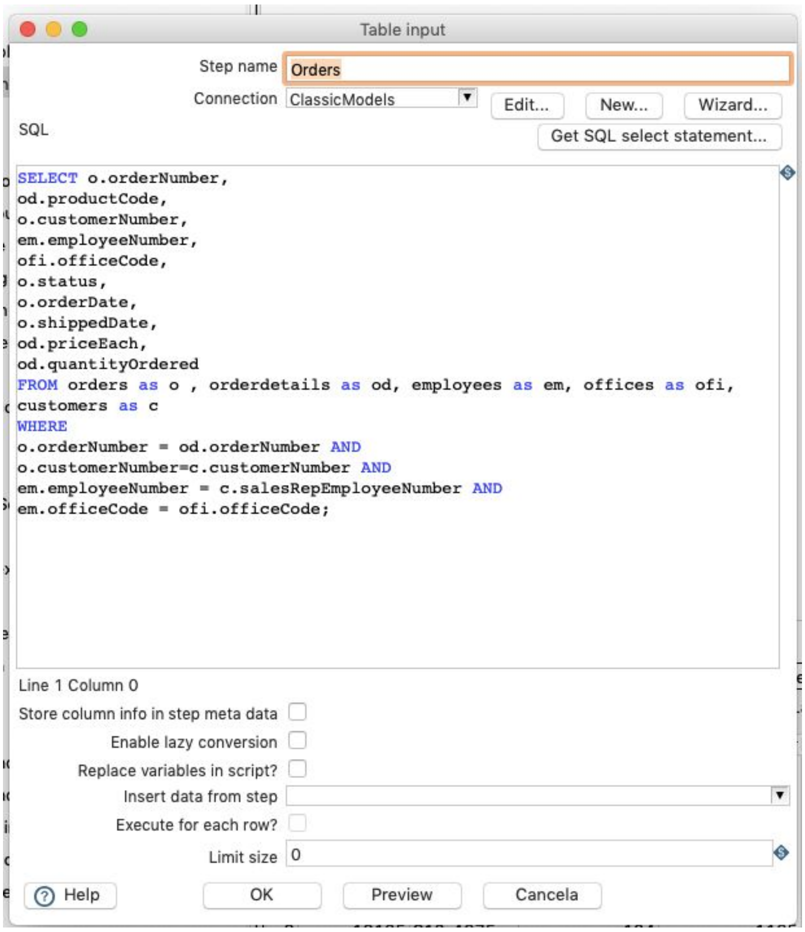
Update fields:

#	Table field	Stream field	Update
1	productID	productCode	Y
2	productName	productName	Y
3	productLine	productLine	Y
4	productScale	productScale	Y
5	productVendor	productVendor	Y
6	quantityInStock	quantityInStock	Y
7	buyPrice	buyPrice	Y
8	MSRP	MSRP	Y



We repeated the process for the other dimension tables such as Customers, Employees and Offices.

Having all the dim_tables populated and ready to go we proceed to create the ETL to the fact_orders table. In this one we have to make some adjustments because some of the tables we wanted on our Data Warehouse were not previously connected to our fact_orders table.



Step name: Orders

Connection: ClassicModels

SQL

```
SELECT o.orderNumber,
od.productCode,
o.customerNumber,
em.employeeNumber,
ofi.officeCode,
o.status,
o.orderDate,
o.shippedDate,
od.priceEach,
od.quantityOrdered
FROM orders as o , orderdetails as od, employees as em, offices as ofi,
customers as c
WHERE
o.orderNumber = od.orderNumber AND
o.customerNumber=c.customerNumber AND
em.employeeNumber = c.salesRepEmployeeNumber AND
em.officeCode = ofi.officeCode;
```

Line 1 Column 0

Store column info in step meta data □

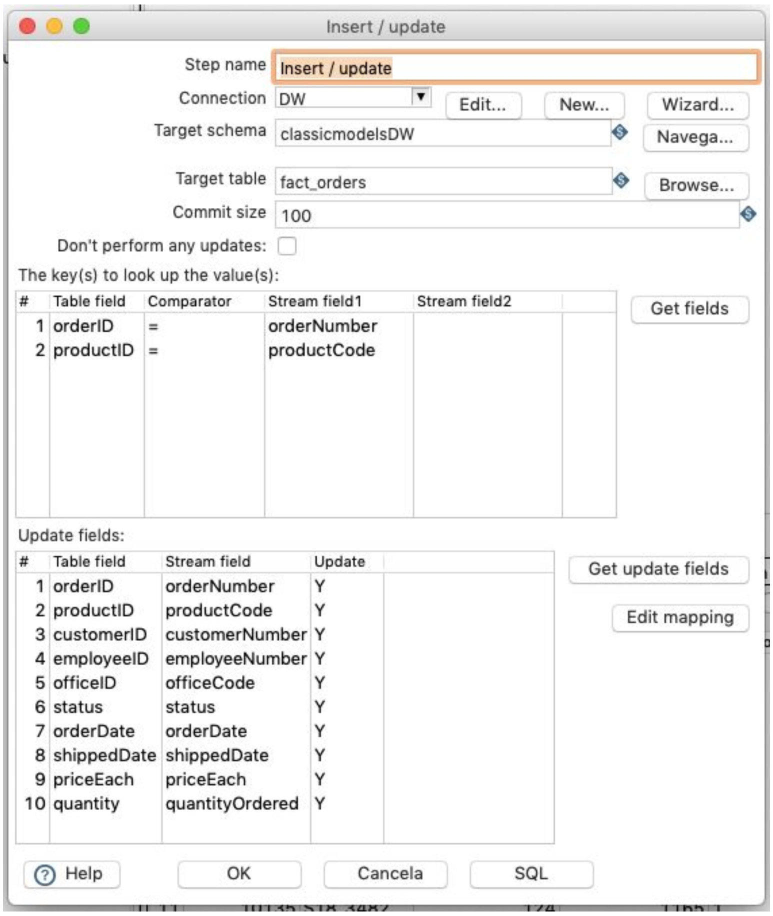
Enable lazy conversion □

Replace variables in script? □

Insert data from step

Execute for each row? □

Limit size: 0



Step name: Insert / update

Connection: DW

Target schema: classicmodelsDW

Target table: fact_orders

Commit size: 100

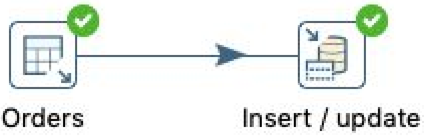
Don't perform any updates: □

The key(s) to look up the value(s):

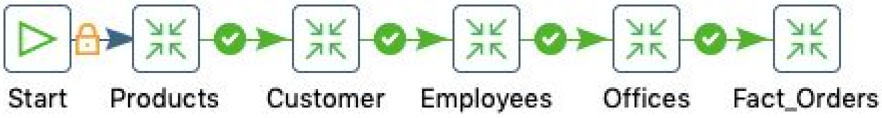
#	Table field	Comparator	Stream field1	Stream field2
1	orderID	=	orderNumber	
2	productID	=	productCode	

Update fields:

#	Table field	Stream field	Update
1	orderID	orderNumber	Y
2	productID	productCode	Y
3	customerID	customerNumber	Y
4	employeeID	employeeNumber	Y
5	officeID	officeCode	Y
6	status	status	Y
7	orderDate	orderDate	Y
8	shippedDate	shippedDate	Y
9	priceEach	priceEach	Y
10	quantity	quantityOrdered	Y



All the Pentaho files (.ktr) were saved and after that a Job named DW_Job was built. All ETL files were put together in order to run all of them at the same time and populate the Data Warehouse just in one file.



After the operations in Pentaho were done it was time to transfer the data we acquired into POWERBI so that we could do the report, thus answering the previously mentioned business questions. Before that though, some extra data transformation was required, which meant the addition of 4 extra columns in the fact table. These were Unit Price, TotalPrice, UnitProfit and TotalProfit. Unit Price is merely the products unit price existent in the products dimensions, this variable was brought into this table through a LOOKUPVALUE function merely to ease the creation of the other remaining variables. UnitProfit is how much of a profit each row had per unit. TotalProfit is the total profit of each row and is attained through the multiplication of UnitProfit and Quantity. Lastly TotalPrice is the row's total price which is the multiplication of priceEach and quantity. The functions used are displayed below.

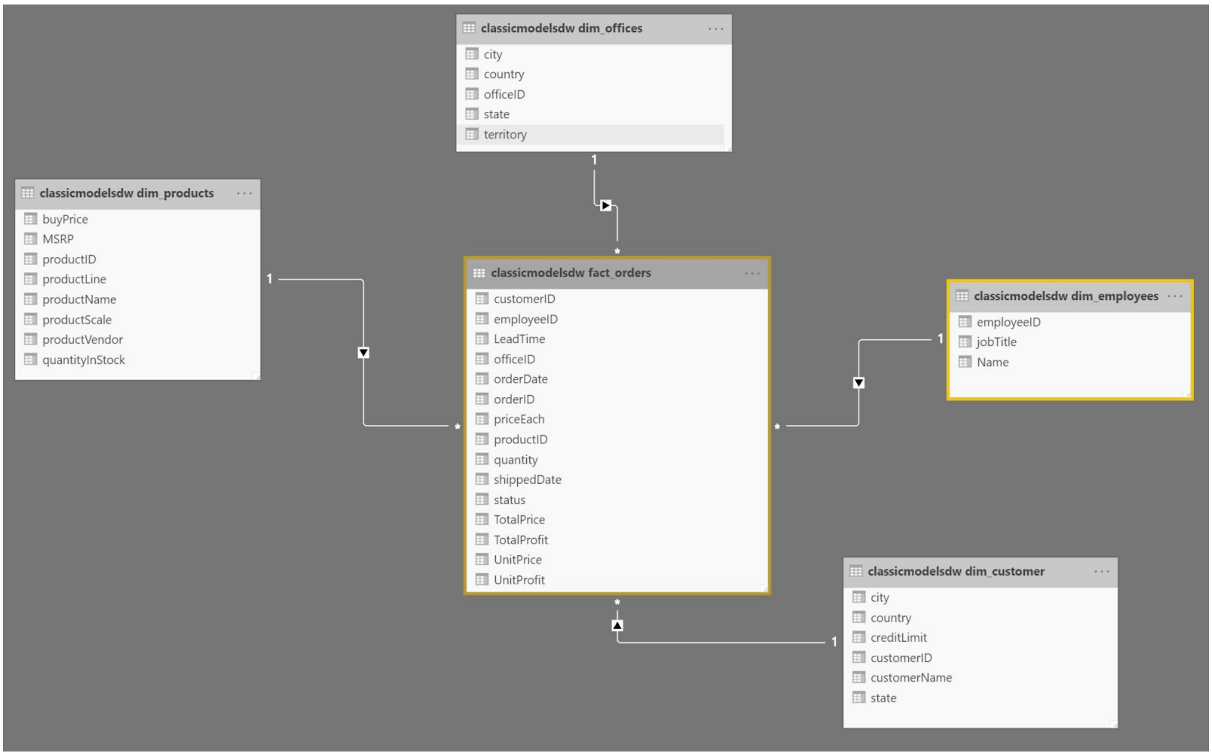
```
UnitPrice = LOOKUPVALUE('classicmodelsdw
    dim_products'[buyPrice],'classicmodelsdw
    dim_products'[productID],'classicmodelsdw fact_orders'[productID])

TotalPrice = 'classicmodelsdw fact_orders'[priceEach]*'classicmodelsdw
    fact_orders'[quantity]

UnitProfit = 'classicmodelsdw fact_orders'[priceEach]-'classicmodelsdw
    fact_orders'[UnitPrice]

TotalProfit = 'classicmodelsdw fact_orders'[UnitProfit]*'classicmodelsdw
    fact_orders'[quantity]
```

These modifications resulted in the slight alteration of the previous Star Schema, the new one, with the addition of these 4 columns is shown below.



classicmodelsdw dim_offices
city
country
officeID
state
territory

classicmodelsdw dim_products
buyPrice
MSRP
productID
productLine
productName
productScale
productVendor
quantityInStock

classicmodelsdw fact_orders
customerID
employeeID
LeadTime
officeID
orderDate
orderID
priceEach
productID
quantity
shippedDate
status
TotalPrice
TotalProfit
UnitPrice
UnitProfit

classicmodelsdw dim_employees
employeeID
jobTitle
Name

classicmodelsdw dim_customer
city
country
creditLimit
customerID
customerName
state

After the data transformation was done it was time to create the report views. In order to facilitate collaboration between all members of the group and so that everyone could work on this regardless of their operating system, the PowerBI file was published to PowerBI online, thus allowing all members of the group to work in the report through their browser.

Below are some screenshots of the PowerBi report results in order to answer the business question. In the beginning we made different possible plots.

In the page name “Operations” the question answered are:

1.5 Which product vendor has the most sales?

In the “Quantity Sold by Product Vendor” table we can see in the top on the table the vendor Classic Metal Creation.

2.1 Which customer buys the most?

In the “Top 10 Best Customer by Revenue” bar chart the customer that has more revenue is the Euro + Shopping Channel.

2.2 Which country buys the most?

In the “Revenue by Customer Country” map the one that is most visible and with a bigger value is America.

3.1 Which offices have the most sales?

In the “Total Revenue per Office” stacked bar chart we can see that each “bar” represents one office so the USA has 3 different Offices and with that information we can tell that the France Office it’s the one who sales the most.

3.2 Which country has the most sales?

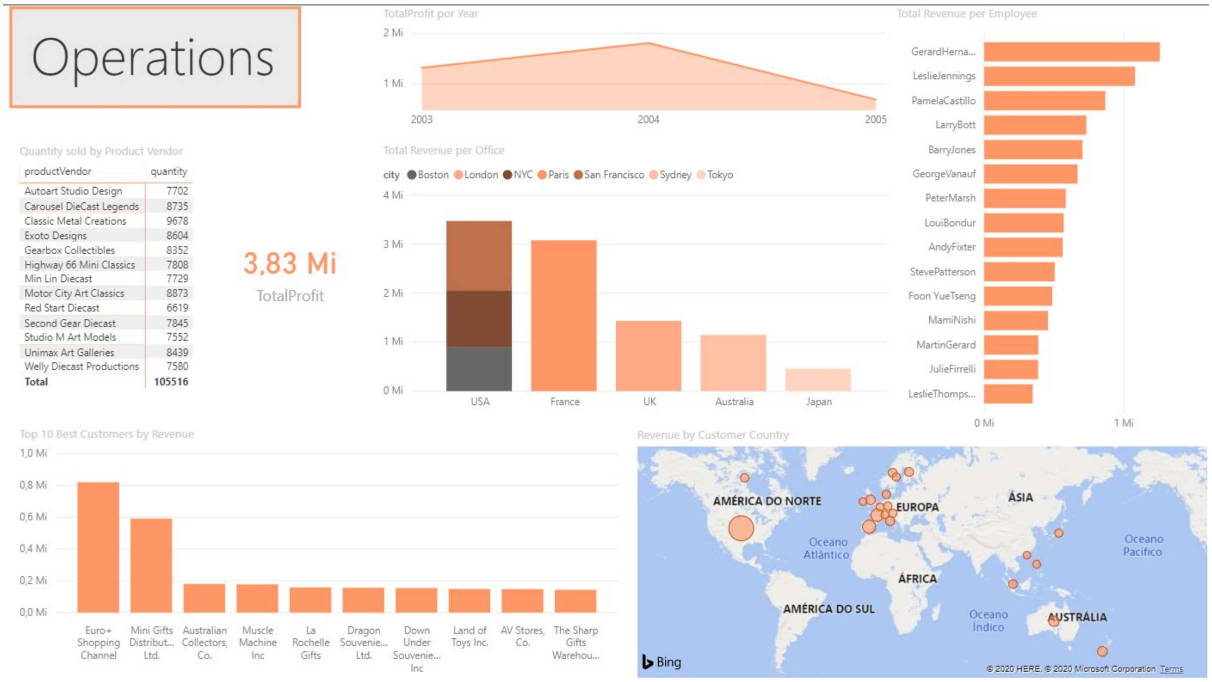
In the “Total Revenue per Office” stacked bar chart the one that presents the most revenue value is America.

4.1 Which employee has the most sales?

In the “Total Revenue per Employee” horizontal bar chart the one that presents the most revenue value is Gerad Hermano.

5.5 What is the Total Profit per year?

In the “Total Revenue per Year” line plot we can see that the year with most profit was 2004, and we can see the values for each year .



Operations

Quantity sold by Product Vendor

productVendor	quantity
Autoart Studio Design	7702
Carousel DieCast Legends	8735
Classic Metal Creations	9678
Exoto Designs	8604
Gearbox Collectibles	8352
Highway 66 Mini Classics	7808
Min Lin Diecast	7729
Motor City Art Classics	8873
Red Start Diecast	6619
Second Gear Diecast	7845
Studio M Art Models	7552
Unimax Art Galleries	8439
Welly Diecast Productions	7580
Total	**105516**

3,83 Mi
TotalProfit

In the page names “Products” the question answered are:

1.1 Which product has the most sales?

The product which is sold more often can be seen at the bar plot “Top 10 Most Sold Products by Quantity”, which is the 1992 Ferrari 360 Spider, with more than 1700 units sold in the total of the years 2013, 2014 and 2015.

1.2 Which product line has the most sales?

The treemap “Top Product Line Sold by Quantity” shows that Classic Cars is the product line most demanded by the customers, followed by Vintage Cars.

1.3 Which product line has the most quantity on stock?

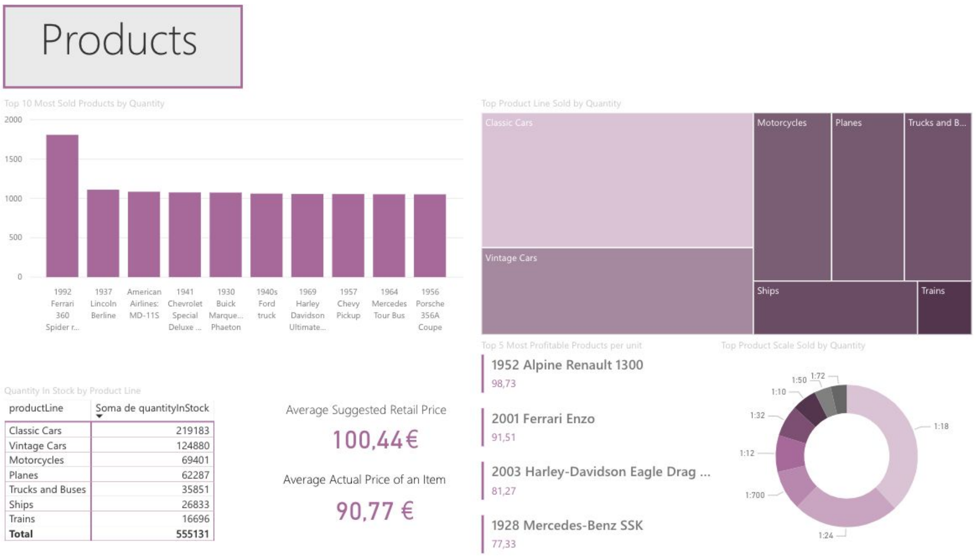
The product line that has the most quantity on stock is, by far, Classic Cars, with 219183 units. The second highest value is from Vintage Cars, with 124880 units in stock.

1.6 What’s the difference between priceEach and MSRP? Are we making “better” prices?

On average, for every product available by CLASSICMODELS, the suggested retail price is set at 100.44€. However, the store sells its products at an average price of 90.77€, which is almost ten euros lower than the suggested retail price.

1.7 What are the most profitable products?

The most profitable product is the 1952 Alpine Renault 1300. Each unit sold of this model corresponds to 98.73€ in profit to the company. The second, third and fourth products with highest profitability per unit are 2001 Ferrari Enzo, 2003 Harley-Davidson Eagle and 1928 Mercedes-Benz SSK.


Products
Top 10 Most Sold Products by Quantity
2000
1500
1000
500
0
1992 Ferrari 360 Spider r...
1937 Lincoln Berline
American Airlines: MD-11S
1941 Chevrolet Special Deluxe ...
1930 Buick Marque... Phaeton
1940s Ford truck
1969 Harley Davidson Ultimate...
1957 Chevy Pickup
1964 Mercedes Tour Bus
1956 Porsche 356A Coupe
Top Product Line Sold by Quantity
Classic Cars
Vintage Cars
Motorcycles
Planes
Trucks and B...
Ships
Trains
Top Product Scale Sold by Quantity
1:72
1:50
1:10
1:32
1:12
1:700
1:24
1:18

Quantity In Stock by Product Line

productLine	Soma de quantityInStock
Classic Cars	219183
Vintage Cars	124880
Motorcycles	69401
Planes	62287
Trucks and Buses	35851
Ships	26833
Trains	16696
Total	**555131**

Average Suggested Retail Price

100,44€

Average Actual Price of an Item

90,77 €

Top 5 Most Profitable Products per unit

1952 Alpine Renault 1300	98,73
2001 Ferrari Enzo	91,51
2003 Harley-Davidson Eagle Drag ...	81,27
1928 Mercedes-Benz SSK	77,33

In the page names “Orders” the question answered are:

5.1 What is the “lead time” for an order?

In the table “Top 10 Average Lead Time per Country”, there is the average lead time for every country where the customers are from. The biggest highlight of the table, is that Singapore has the highest average lead time, with 16 days. Se second highest value is much smaller from Singapore, which is 6 days in Japan.

5.2 What is the different status of an order and their ratio? Particular attention on the “canceled”.

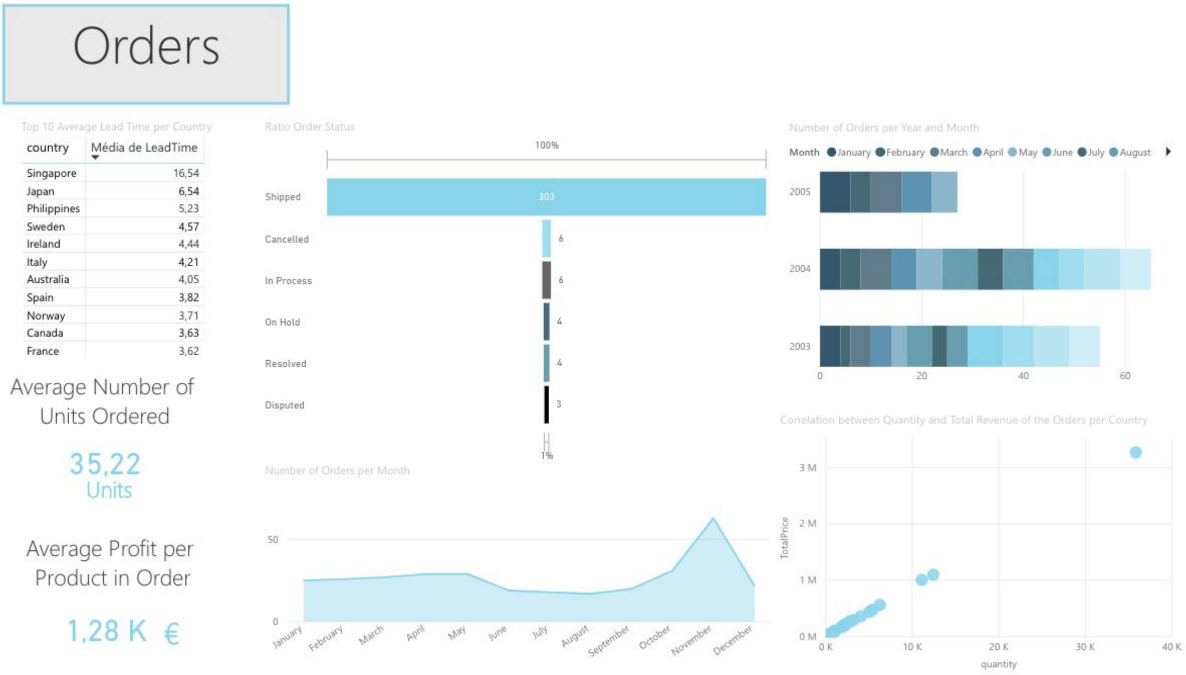
The graph “Ratio Order Status” shows that most of the orders have been shipped, 303. However, the most important for the analysis is the number of Cancelled orders, which is 6. In addition, 6 orders were in process at the date of the database, 4 were on hold, and other 4 were resolved. Finally, 3 orders were being disputed.

5.3 What is the mean quantity of products sold in an order?

The mean quantity of products sold per order was 35 units.

5.4 What is the mean price gain in an order?

On average, every type of product brings 12 800€ in profit to the company in each order.



Orders

Top 10 Average Lead Time per Country

country	Média de LeadTime
Singapore	16,54
Japan	6,54
Philippines	5,23
Sweden	4,57
Ireland	4,44
Italy	4,21
Australia	4,05
Spain	3,82
Norway	3,71
Canada	3,63
France	3,62

Average Number of Units Ordered

35,22 Units

Average Profit per Product in Order

1,28 K €

Conclusion

In conclusion we can say that the business is mainly based on the Classic and Vintage Cars Models, and it is indeed growing year after year. Despite the fact that the year 2005 only has order sales until May, in those first five months the number of orders is higher than the counterpart period in 2004 and 2003, which is a good indicator for the sales for the rest of the year.

In a more detailed analysis that could be done in the future, we could explore the variation of the orders by product in the years 2003 and 2004, to assess which were the products and product lines that had the highest positive variation and those who had the biggest drop in the number of orders. This could be done to know if the store should have more items in stock of a particular product, or if the company should stop selling a certain item. In addition, we could plot a forecast for the following months or years, using Power BI. However, we decided that for this project, it was not need such detailed analysis.