Integração e Processamento Analítico de Dados Project

Second Deliveryx

Pedro Oliveirax, 52764
Daniel Gomes, 51649
Milutin Popovic, 61284
Ruben Rodrigues, 52160

Faculdade de Ciências da Universidade de Lisboa
2023

Contents

Introduction

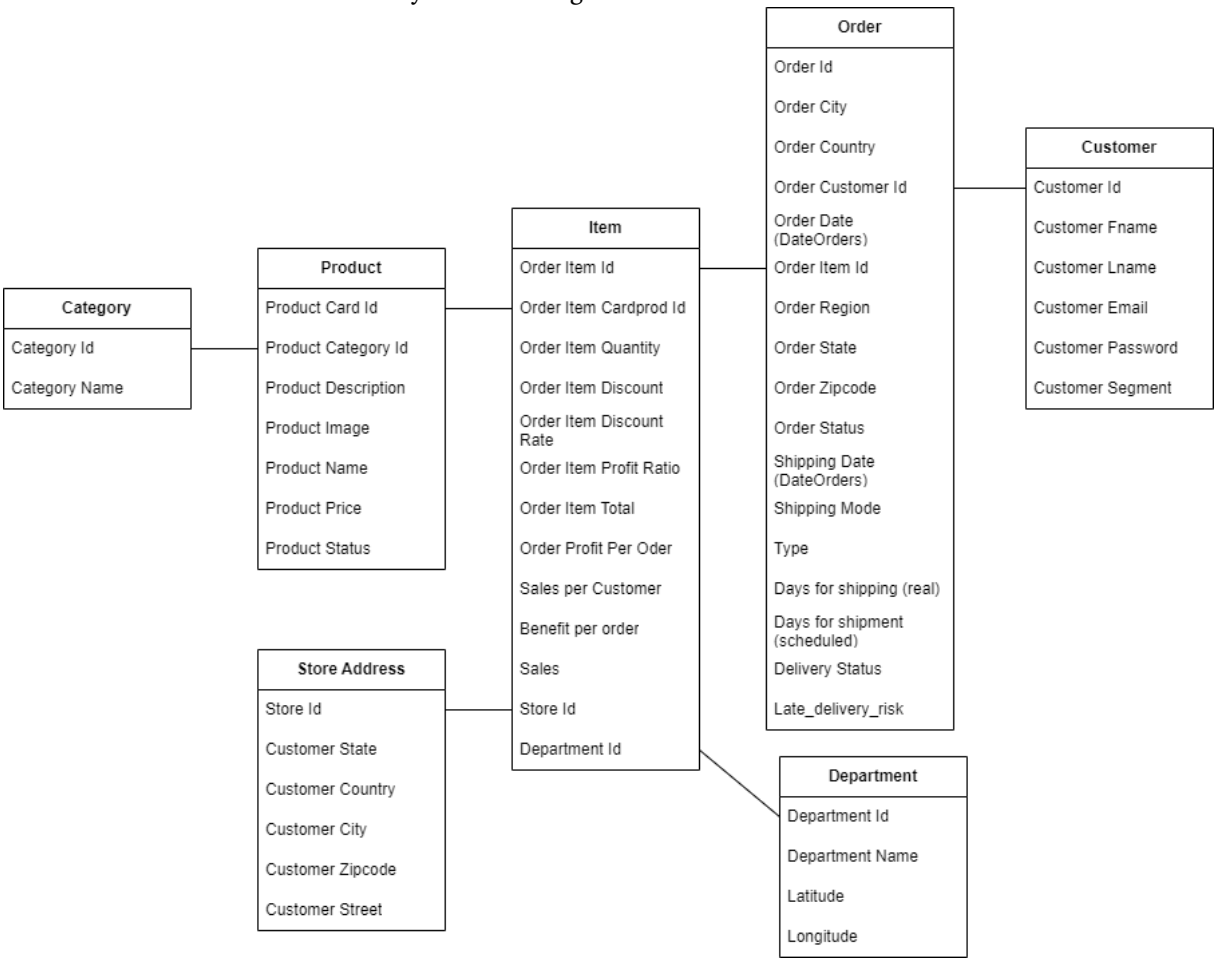
The dataset chosen has information about a company's supply chains and was obtained from Kaggle (`https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis?select=DataCoSupplyChainDataset.csv`).

Changes

Some columns' data types didn't correspond to the type of data so they were corrected. There was also a change in the way that we organized the dataset in order to present it. We didn't pay enough attention to some fields' descriptions and their names were misleading. The fields "Customer City", "Customer Country", "Customer Zipcode", "Customer State" and "Customer Street" now belong in the Store Address table since they apparently are referred to the store where the order was made and not to the customer address as we thought. Since we are not sure if a store can have multiple departments we kept the Department table because in several instances there are different coordinates for departments in the same street (with very slight changes).

1 Diagram

The data was contained in a single file so as to make the diagram explaining the relation between the different columns we partitioned the dataset into multiple tables according to the different categories that best describe that relation connected by the matching attributes.



2 Data Exploration

In this section will be presented an exploratory analysis of the tables and a simple description of the fields in each one.

2.1 Product

Field	Description	Data Type	Example
Product Card Id	Product code	Numeric	365
Product Category Id	Product category code	Numeric	73
Product Description	Product Description	Text	
Product Image	Link of visit and purchase of the product	Text	`http://images.acmesports.sports/Smart+watch`
Product Name	Product Name	Text	Smart watch
Product Price	Product Price	Numeric	327.75
Product Status	Status of the product stock	Categorical	0

Table 1: Product category data fields description

The Table 1 contains the data with corresponding information about each product.

By exploring this table the first thing to notice is that there aren't any descriptions for any product and that the product status is the same for all the products (all products are available). In total there are 118 products..

From this table, we can see that the average product price is 166.41 and the prices range from 9.99 to 1999.99. In Figure 1 we can see the 10 most expensive products available.

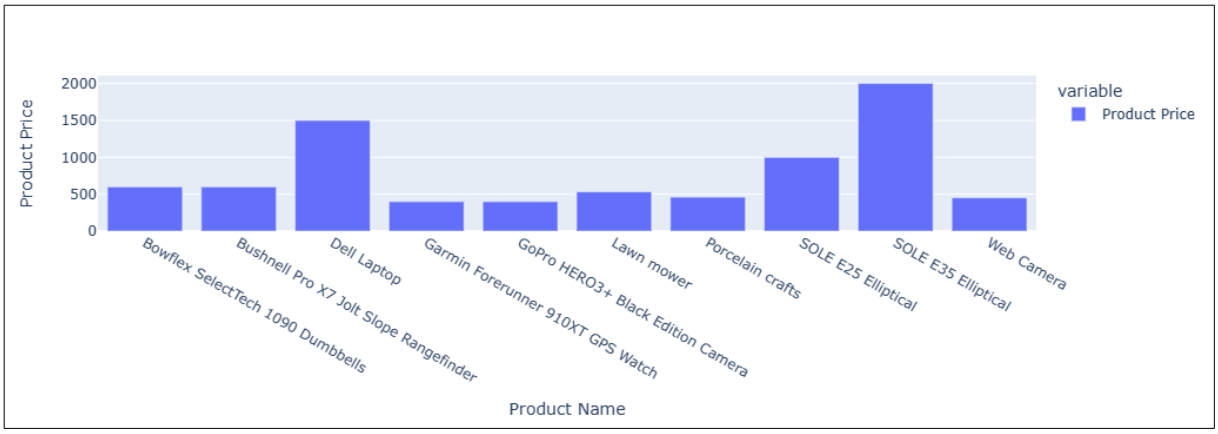


Figure 1: Top 10 most expensive products

2.2 Store Address

The only data field with missing values is the Customer Zipcode (3 missing values). There are only two different values in the Customer Country data field but this happens because in the dataset Puerto Rico appears as a country so all the stores are in the US and are distributed over 46 states. In total there are 12024 stores and the fields information is presented in Table 2

Field	Description	Data Type	Example
Store Id	Store code	Numeric	4
Customer City	City where the customer made the purchase	Text	Bayamon
Customer Country	Country where the customer made the purchase	Text	Puerto Rico
Customer State	State to which the store where the purchase is registered belongs	Text	PR
Customer Zipcode	Store Zipcode	Numeric	957
Customer Street	Store's street	Text	75 Sunny Grounds

Table 2: Store Address data fields description

2.3 Customer

Field	Description	Data Type	Example
Customer Id	Customer ID	Numeric	9083
Customer Fname	Customer name	Text	Mary
Customer Lname	Customer last name	Text	Frank
Customer Email	Customer's email	Text	XXXXXXXXX
Customer Password	Masked customer key	Text	XXXXXXXXX
Customer Segment	Types of Customers	Categorical	Corporate

Table 3: Customer data fields description

The Table 3 contains all the information available about the clients. There are only a few missing values in the columns with the customers' last names (8 missing values). By looking at the types of values that each data field can have we noticed that both the customers' emails and passwords are masked with all the values being "XXXXXXXXX". In total, there are 20652 customers and they ordered from two countries (Figure 2).

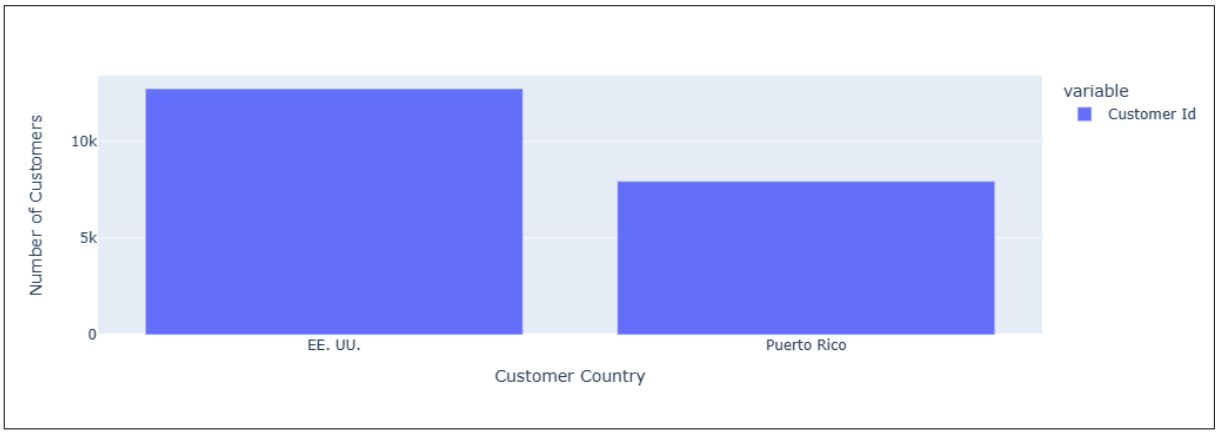


Figure 2: Number of customers ordering from each country

The customers are also distributed over three segments (Figure 3).

2.4 Department

This table only contains the code, name, and coordinates of the store. However, what we noticed is that the code corresponds to the identifier of a type of department, and the name is connected to the code, while it appears to exist multiple stores from each department across multiple locations, and in total there are 11 types of departments.

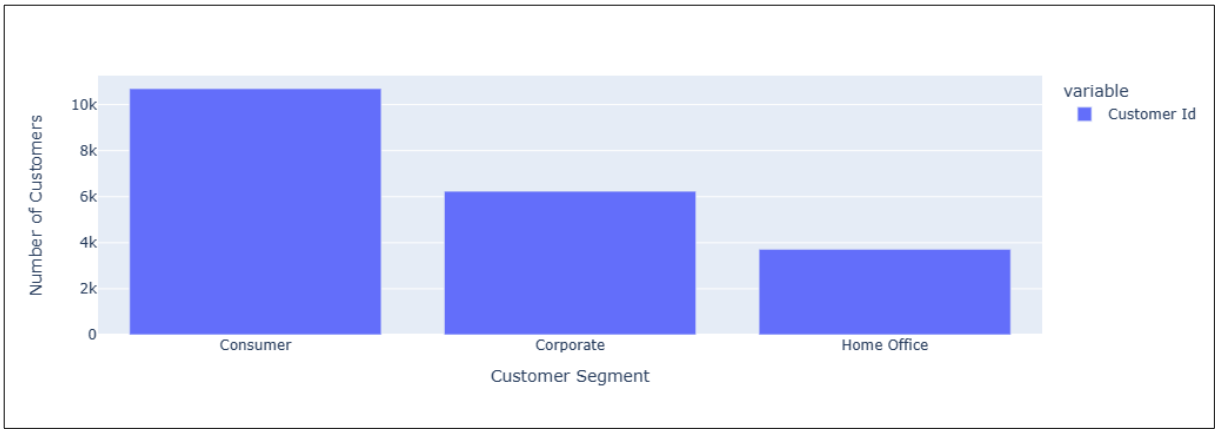


Figure 3: Number of customers per segment

Field	Description	Data Type	Example
Department Id	Department code of store	Numeric	4
Department Name	Department name	Text	Apparel
Latitude	Latitude corresponding to location of store	Numeric	18.380119
Longitude	Longitude corresponding to location of store	Numeric	-66.183128

Table 4: Department data fields description

In Figure 4 we can see what departments sell the most and the number of orders.

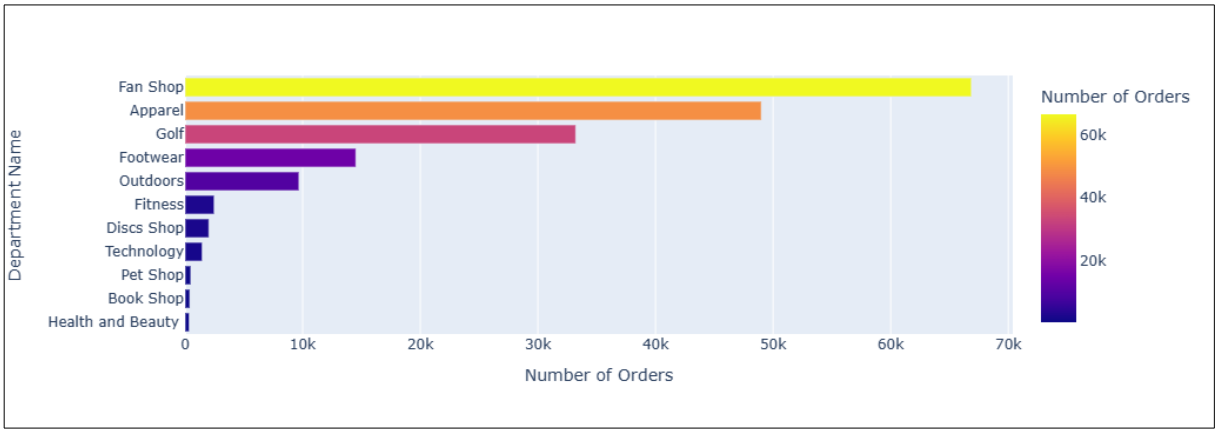


Figure 4: Number of orders per department of product

2.5 Order

This table has multiple information about the orders of the customers, mainly information about the shipping process like the dates and the region to where the items will be shipped. In total there are 65752, however, this table has more entrances due to its connection to the items table. This happens because orders with different items will have different product codes in the Order Item Id column. The fields' information is presented in Table 5

In Figure 5 we can see the distribution of orders per shipping mode. And in Figure 6 we can also see the status of the orders according to their shipping mode to see how the choice of shipping mode may influence if the order is delivered on time or late.

Field	Description	Data Type	Example
Order Id	Order code	Numeric	28744
Order City	Destination city of the order	Text	Mirzapur
Order Country	Destination country of the order	Text	India
Order Customer Id	Customer order code	Numeric	9083
Order Date	Date on which the order is made	Text	2/24/2016 13:57
Order Item Id	Order code	Numeric	71956
Order Region	Region of the world where the order is delivered	Text	Southeast Asia
Order Zipcode	Order destination zipcode	Numeric	957
Order State	State of the region where the order is delivered	Text	Uttar Pradesh
Order Status	Order Status	Categorical	COMPLETE
Shipping Date	Exact date and time of shipment	Text	2/29/2016 13:57
Shipping Mode	Shiping Mode	Categorical	Standard Class
Market	Market to where the order is delivered	Text	Africa
Type	Type of transaction made	Categorical	PAYMENT
Days for shipment (scheduled)	Days of scheduled delivery of the purchased product	Numeric	2
Days for shipping (real)	Actual shipping days of the purchased product	Numeric	5
Delivery Status	Delivery status of orders	Categorical	Late delivery
Late Delivery Risk	indicates if sending is late	Categorical	0

Table 5: Order data fields description

2.6 Item

Field	Description	Data Type	Example
Order Item Id	Order code	Numeric	71956
Order Item Cardprod Id	Product code generated through the RFID reader	Numeric	365
Order Item Quantity	Number of products per order	Numeric	2
Order Item Discount	Order item discount value	Numeric	4.8
Order Item Discount Rate	Order item discount percentage	Numeric	0.04
Order Item Profit Ratio	Order Item Profit Ratio	Numeric	-0.27
Order Item Total	Total amount per order	Numeric	115.180000
Order Profit per Order	Order Profit Per Order	Numeric	-30.750000
Sales per Customer	Total sales per customer made per customer	Numeric	115.180000
Benefit per order	Earnings per order placed	Numeric	-30.750000
Sales	Value in sales (quantity * product price)	Numeric	119.980003
Department Id	Department code of store	Numeric	4

Table 6: Item data fields description

This table has information about the specific products ordered by the customers. Each item is identified by its unique id. What we notice is that one type of product can appear as two items in the same order, the reason for this appears to be that the discount applied can be different for the same product in different items. When the discount is the same and the customer order more than 1 unit of a product, this will create only one item instance where the field "Order Item Quantity " will be the number of units ordered.

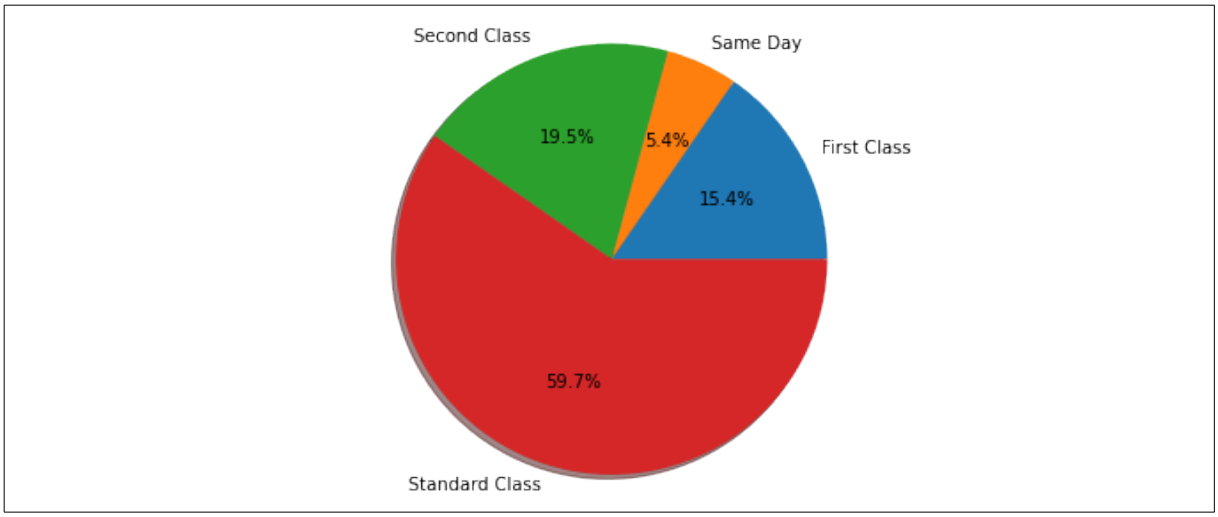


Figure 5: Percentage of orders per shipping mode

2.7 Product Category

Field	Description	Data Type	Example
Category Id	Product category code	Numeric	73
Category Name	Description of the product category	Text	Sporting Goods

Table 7: Product Category data fields description

This table only matches the categories' ids with their description. One possible error is the fact that "Electronics" appears related to two different ids. In total there are 51 different ids and 50 descriptions.

In Figure 7 we can see what categories sell the most and the number of orders.

3 Questions

These are the questions we choose:

- How much more sales are there when a product is at discount, on average?
- What time of the year has the most orders/sales (in date and time)?
- Does the shipment mode impact the delivery date?

4 Business Process

Regarding the business process, considering the questions we chose, we think the business process that would take the most advantage of the data would be retail sales; analyzing the data and taking conclusions regarding its variance, it would help make better corporate decisions when it comes to maximizing sale profit (by seeing how much discounts or shipment mode affects them) and consequentially end customer satisfaction.

5 Dimensional Modeling

5.1 Type of facts table and grain declaration

The fact table will be transactional; each line will correspond to an item bought/ordered which may or not be the only item bought/ordered in a given order. Each item is a certain type of product and is ordered

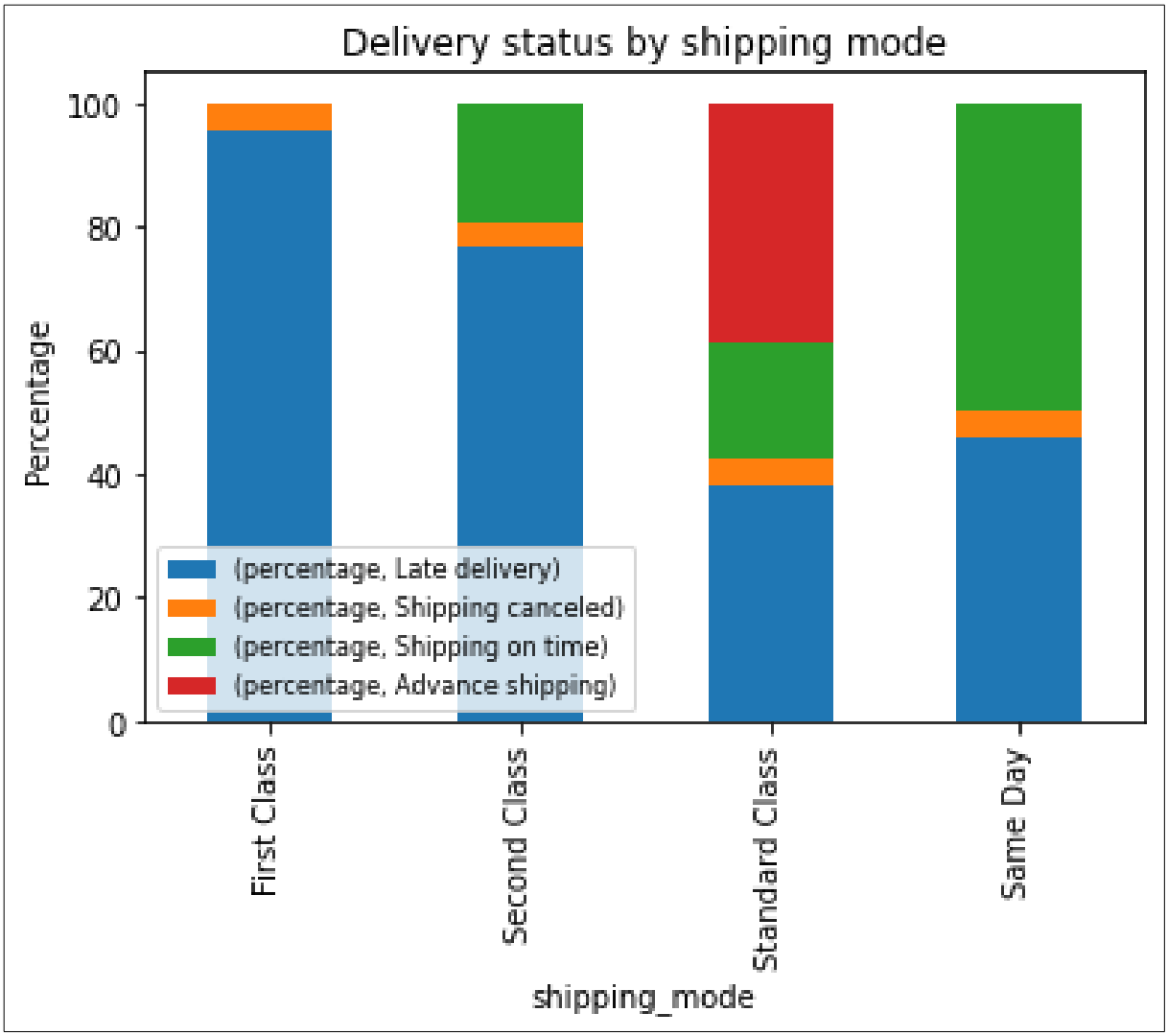


Figure 6: Distribution of orders per status according to the shipping mode

from a department by a customer in an order which may include other items.

5.1.1 Facts Table

Although the primary keys for the fact table usually are the foreign keys that connect this table to the dimensions, we had to include an identifier for each transaction since in our case the foreign keys don't uniquely identify each transaction. This happens because the same order can include two items of the same type but with a different discount rate. So when this happens there will be a row for each of the items with the quantity field equal to 1 and different discount rates.

Besides the identifier and the foreign keys fields, the fact table has the following metrics: quantity, discount, discount rate, profit, profit ratio, and total price. The fields description and data type are presented in Table 8. It may be relevant to note that these measures are taken from the original dataset (from the "Item" table).

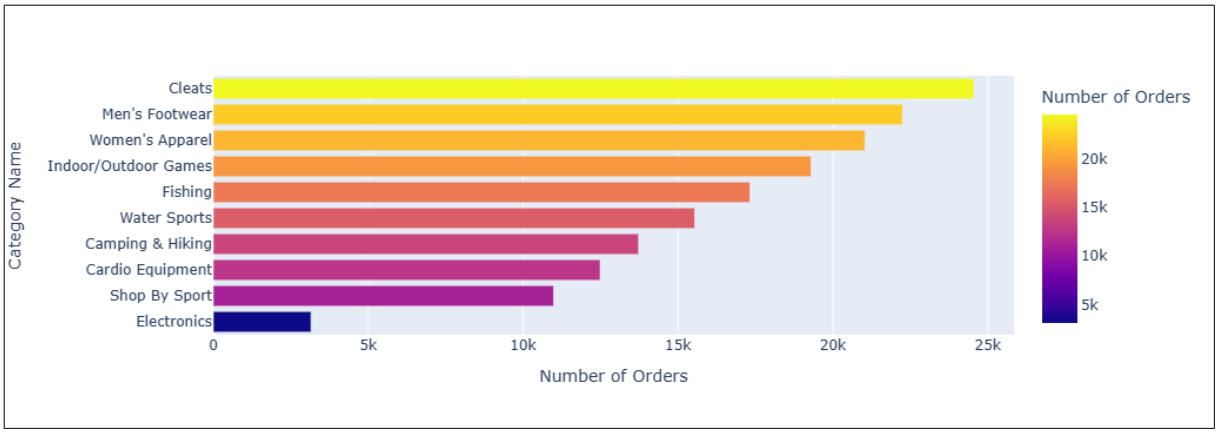


Figure 7: Number of orders per category of product (Top 10 categories)

Field	Description	Data Type	Data Source
Transaction id	Item identifier (sequential values)	Numeric	Id sequentially generated
Order Key	Foreign key to the Order dimension	Numeric	
Product Key	Foreign key to the Product dimension	Numeric	
Department Key	Foreign key to the Department dimension	Numeric	
Store Key	Foreign key to the Store dimension	Numeric	
Customer Key	Foreign key to the Customer dimension	Numeric	
Quantity	Number of items ordered (same product and equal discount rate)	Numeric	Item table (Order Item Quantity)
Discount	Total discount	Numeric	Item table (Order Item Discount)
Discount Rate	Discount rate applied	Numeric	Item table (Order Item Discount Rate)
Profit	Profit made with the item/s	Numeric	Item table (Order Item Profit Per Order)
Profit Ratio	Profit ratio	Numeric	Item table (Order Item Profit Ratio)
Total Price	The total price paid by the customer	Numeric	Item table (Order Item Total)

Table 8: Facts table fields description

5.2 Dimensions

Although the original dataset consists of a single table, the source of the data in the dimensions will be referred to by the "tables" used to present the dataset. Given our business process, we think that the best way to model it is to have 4 dimensions: Customer, Product, Order, and Department.

5.2.1 Customer

In this dimension, we have the data fields with information about the customers. We did not include the fields with the email and password because, as explained above, we cannot access the values. The field descriptions and data source are presented in Table 9.

5.2.2 Product

In this dimension, we have the data fields with information about the products. It is important to note that we do not have a dimension for the product category as we had before since the category is something that a product belongs to and we think that is an attribute that describes a product. Also, there are only

Field	Description	Data Type	Data Source
Customer Key	Surrogate key	Numeric	
First Name	Customer first name	Text	Customer table (Customer Fname)
Last Name	Customer last name	Text	Customer table (Customer Lname)
Segment	Type of customer	Categorical	Customer table (Customer Segment)

Table 9: Customer dimension fields description

118 products so adding the category field is only a marginal increase in space occupied. We did not use the field "Product Description" because this field was empty for all the products. The field descriptions and data source are presented in Table 10.

Field	Description	Data Type	Data Source
Product Key	Surrogate key	Numeric	
Category	Product category	Text	Product Category table (Category Name)
Name	Customer last name	Text	Product table (Product Name)
Image	Link of the product image	Text	Product table (Product Image)
Price	Product price	Numeric	Product table (Product Price)
Status	Status of the product stock	Categorical	Product table (Product Status)

Table 10: Product dimension fields description

5.2.3 Store

In this dimension, we have the data fields with information about the store where the order was registered (basically the store address). The field descriptions and data source are presented in Table 11.

Field	Description	Data Type	Data Source
Store Key	Surrogate key	Numeric	
Country	The country where the customer made the purchase (store country)	Text	Store Address table (Customer Country)
State	State to which the store where the purchase is registered belongs	Text	Store Address table (Customer State)
City	The city where the customer made the purchase (store city)	Text	Store Address table (Customer City)
Street	Street where the customer made the purchase (store street)	Text	Store Address table (Customer Street)
Zipcode	Zipcode of the store where the customer made the purchase (store zipcode)	Numeric	Store Address table (Customer Zipcode)

Table 11: Store dimension fields description

5.2.4 Department

In this dimension, we have the data fields with information about the department (Table 12).

5.2.5 Order

In this dimension, we have the data fields with information about the order. We expanded the fields containing timestamps so this dimension now has three attribute hierarchies that are described below. The field descriptions and data source are presented in Table 13.

Field	Description	Data Type	Data Source
Department Key	Surrogate key	Numeric	
Name	Department name	Text	Department table (Department Name)
Latitude	Latitude corresponding to location of store	Numeric	Department table (Latitude)
Longitude	Longitude corresponding to location of store	Numeric	Department table (Longitude)

Table 12: Department dimension fields description

5.3 Hierarchies

5.3.1 Order - shipping date and order date

We have two hierarchies involving a timestamp in the dimension Order (Table 13). The hierarchies are the following:

Order timestamp

1. Year
2. Month
3. Day
4. Hour
5. Minute

Shipping timestamp

1. Shipping Year
2. Shipping Month
3. Shipping Day
4. Shipping Hour
5. Shipping Minute

5.3.2 Order - destination data

Another hierarchy present in the Order dimension (Table 13) is the information about the destination of the order. The hierarchy is the following:

1. Market
2. Region
3. Country State
4. City
5. Zipcode

5.3.3 Store - address data

The hierarchy present in the Store dimension (Table 11) is the information about the destination of the order. The hierarchy is the following:

1. Country State
2. City
3. Zipcode
4. street

5.3.4 Customer - segment

We can separate the consumers either individually or by the segment to which they belong in the Customer dimension (Table 9). The hierarchy is the following:

1. Segment
2. First/Last name

5.3.5 Customer - segment

We can separate the consumers either individually or by the segment to which they belong in the Customer dimension (Table 9). The hierarchy is the following:

1. Segment
2. First/Last name

5.3.6 Product - category

We can separate the product either by the product name or by the category to which they belong in the Product dimension (Table 10). The hierarchy is the following:

1. Category
2. Name

5.4 Star diagram

Field	Description	Data Type	Data Source
Order Key	Surrogate key	Numeric	
Market	Market where the order is delivered	Text	Order table (Market)
Region	Region of the world where the order is delivered	Text	Order table (Order Region)
Country	Destination country of the order	Text	Order table (Order Country)
State	State where the order is delivered	Text	Order table (Order State)
City	Destination city of the order	Text	Order table (Order City)
Store City	City where the customer made the purchase	Text	Order table (Customer City)
Zipcode	Order destination zipcode	Numeric	Order table Order Zipcode(Order Zipcode)
Store Country	Country where the customer made the purchase	Text	Order table (Customer Country)
Store State	State to which the store where the purchase is registered belongs	Text	Order table (Customer State)
Store Zipcode	Store Zipcode	Numeric	Order table (Customer Zipcode)
Store Street	Store's street	Text	Order table (Customer Street)
Type	Type of transaction made	Categorical	Order table (Type)
Shipping Mode	Shipping mode	Categorical	Order table (Shipping Mode)
Year	Year on which the order was made	Numeric	Order table (Order Date)
Month	Month on which the order was made	Numeric	Order table (Order Date)
Day	Day on which the order was made	Numeric	Order table (Order Date)
Hour	Hour at which the order was made	Numeric	Order table (Order Date)
Minute	Minute at which the order was made	Numeric	Order table (Order Date)
Shipping Year	Year in which the order was shipped	Numeric	Order table (Shipping Date)
Shipping Month	Month in which the order was shipped	Numeric	Order table (Shipping Date)
Shipping Day	Day in which the order was shipped	Numeric	Order table (Shipping Date)
Shipping Hour	Hour at which the order was shipped	Numeric	Order table (Shipping Date)
Shipping Minute	Minute at which the order was shipped	Numeric	Order table (Shipping Date)
Scheduled Shipping Days	Days scheduled for the delivery of the purchased products	Numeric	Order table (Days for shipment (scheduled))
Real Shipping Days	Actual shipping days of the purchased products	Numeric	Order table (Days for shipping (real))
Status	Order status	Categorical	Order table (Order Status)
Delivery Status	Delivery status of the orders	Categorical	Order table (Delivery Status)
Late Delivery Risk	Indicates if the sending process is late	Categorical	Order table (Late Delivery Risk)

Table 13: Order dimension fields description

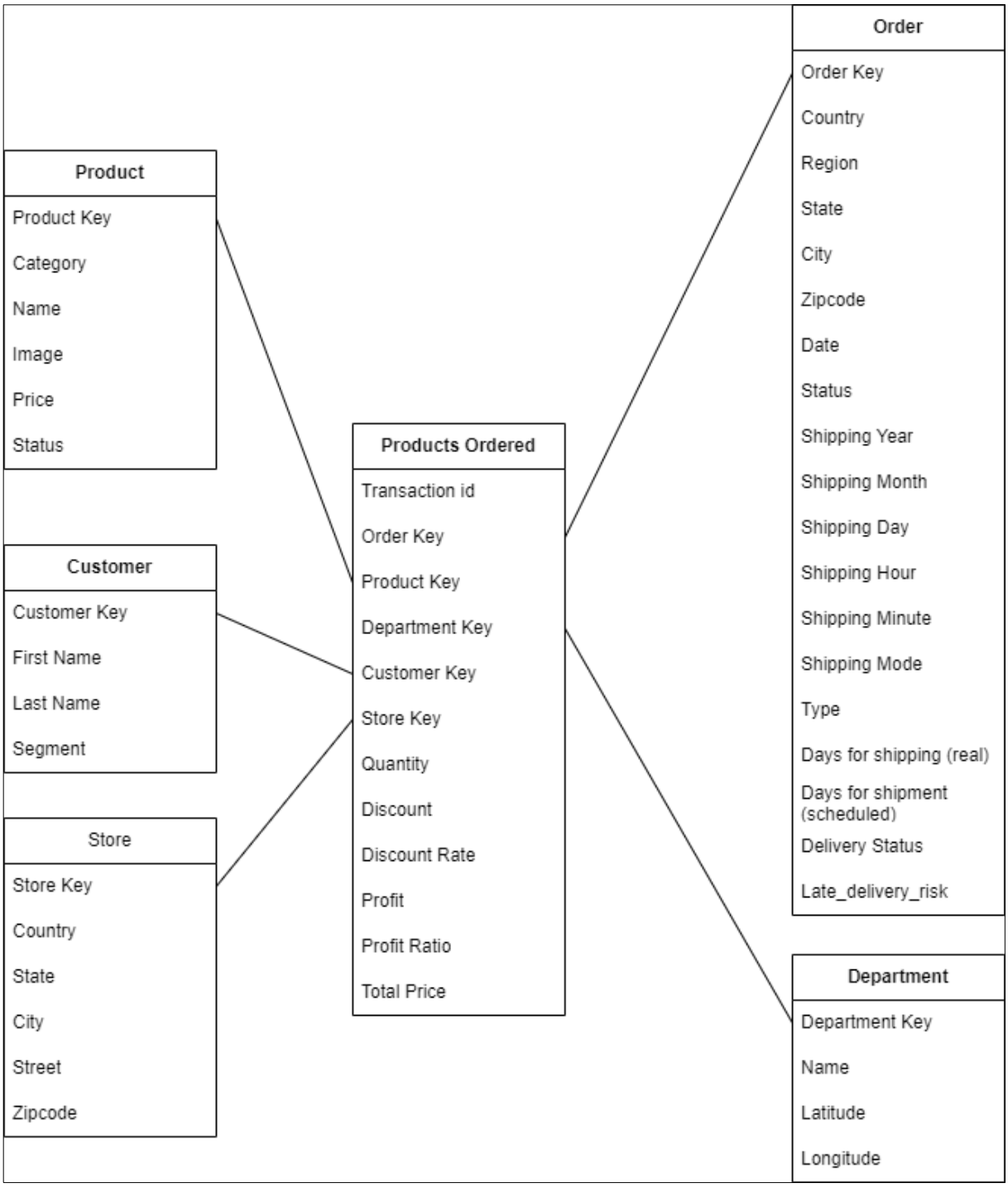


Figure 8: Star diagram