Graphical user interface

Description automatically generated with low confidence

Projeto de Integração e Processamento Analítico de Informação

Stage I: Data Analysis

Grupo 1

Tiago Rodrigues nº49593

Ivo Oliveira nº50301

Martim Silva nº51304

Alexandre Sobreira nº59451

2022/2023

**Index**

[1. Identify the data sources used 3](#_Toc130639350)

[Global Superstore Orders 2016.xlsx 3](#_Toc130639351)

[Orders 3](#_Toc130639352)

[Sellers 7](#_Toc130639353)

[Global Superstore Returns 2016.csv: 7](#_Toc130639354)

[Returns 7](#_Toc130639355)

[Superstore Dataset 8](#_Toc130639356)

[Costumers\_USA 8](#_Toc130639357)

[US Holiday Dates (2004 - 2021) 11](#_Toc130639358)

[holiday\_USA 12](#_Toc130639359)

[Highest GDP Counties in the USA 12](#_Toc130639360)

[GDP\_USA 12](#_Toc130639361)

[2. Analyse values and errors in the data fields of each source. 15](#_Toc130639362)

[Orders table 15](#_Toc130639363)

[Customers table 16](#_Toc130639364)

[Returns table 16](#_Toc130639365)

[Sellers table 16](#_Toc130639366)

[Holiday USA table 16](#_Toc130639367)

[USA GDP table 16](#_Toc130639368)

[3. Draw a diagram with links between data sources 17](#_Toc130639369)

[4. Describe a business process to take advantage of data 17](#_Toc130639370)

[5. Define three analytical questions for the business process 18](#_Toc130639371)

[Sellers Analysis 18](#_Toc130639372)

[Customer analysis 18](#_Toc130639373)

[Product analysis 19](#_Toc130639374)

# 1. Identify the data sources used

This project will aim to aid a big e-commerce chain in improving its business by analysing various datasets that are related directly or indirectly to it. With this goal in mind, four data sets will be analysed together. The four datasets are Global Superstore Orders 2016; Global Superstore Returns 2016; Superstore Dataset; US Holiday Dates (2004-2021); Highest GDP Counties in the USA. All the analysis in this step was done using python 3 with the aid of packages such as pandas, NumPy, seaborn and matplotlib. The four datasets will be described in the following steps.

## Global Superstore Orders 2016.xlsx

This dataset can be found in the open source "data.world" (<https://data.world/tableauhelp/superstore-data-sets>). It has a total of 8.08 MB in xlsx format, containing two tables, Orders and People, which we will call Sellers. This dataset includes information regarding the sales of the large-scale e-commerce chain.

### Orders

The first table of this dataset is the table Orders. This table can be considered the orders source of information for this project, containing 51290 rows and 24 columns, which are presented below.

#### Columns

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **Column** | **Row ID** | **Order ID** | **Order Date** | **Ship Date** | **Ship Mode** | **Customer ID** | **Customer Name** |
| **Description** | Unique ID for each row. | Unique Order ID for each Customer. | Order Date of the product. | Shipping Date of the Product. | Shipping Mode specified by the customer. | Unique ID to identify each customer. | Name of the Customer. |
| **Data type** | Int64 | Object | Datetime64[ns] | Datetime64[ns] | Object | Object | Object |
|  |  |  |  |  |  |  |  |
| **#** | **8** | **9** | **10** | **11** | **12** | **13** | **14** |
| **Column** | **Segment** | **Postal Code** | **City** | **State** | **Country** | **Region** | **Market** |
| **Description** | The segment where the customer belongs. | Postal Code of every Customer | City of residence of the seller. | State of residence of the seller. | Country of residence of the seller. | The region where the seller belongs. | Global location by Market |
| **Data type** | Object | Float64 | Object | Object | Object | Object | Object |
|  |  |  |  |  |  |  |  |
| **#** | **15** | **16** | **17** | **18** | **19** | **20** | **21** |
| **Column** | **Product ID** | **Category** | **Product Name** | **Sales** | **Quantity** | **Discount** | **Profit** |
| **Description** | Unique ID of the Product | The Category of the product ordered. | Name of the Product | Sales of the Product. | Quantity of the Product. | Discount provided. | Profit/Loss incurred. |
| **Data type** | Object | Object | Object | Float64 | Int64 | Float64 | Float64 |
|  |  |  |  |  |  |  |  |
| **#** | **23** | **24** |
| **Column** | **Shipping Cost** | **Order Priority** |
| **Description** | Cost per shipment | Level of priority for the order |
| **Data type** | Float64 | Object |

Table 1 Orders table column number (#), name, description and data type

As it is possible to observe in Table 1, table Orders contains data that goes from information about the dates of the orders and shipments, information regarding the customers as well as the seller (e.g., region) and details about the product, such as categories and various measures (e.g., profit).

To better understand the Orders table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

#### Categorical Variables

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column** | **Order ID** | **Order Date** | **Ship Date** | **Ship Mode** | **Customer ID** | **Customer Name** | **Segment** | **Country** |
| **Unique** | 25728 | 1430 | 1464 | 4 | 17415 | 796 | 3 | 165 |
| **Example** | CA-2015-SV20365140-42268 | 2015-12-31 00:00:00 | 2016-01-07 00:00:00 | Standard Class | SV-203651406 | Muhammed Yedwab | Consumer | United States |
|  |  |  |  |  |  |  |  |  |
| **Column** | **City** | **State** | **Region** | **Product ID** | **Category** | **Sub-category** | **Product Name** | **Market** |
| **Unique** | 3650 | 1102 | 23 | 3788 | 3 | 17 | 3788 | 5 |
| **Example** | New York City | California | Western Europe | OFF-FA-6129 | Office Supplies | Binders | Staples | Asia Pacific |
|  |  |  |  |  |  |  |  |  |
| **Column** | **Order Priority** |
| **Unique** | 4 |
| **Example** | Medium |

Table 2 Orders table categorical variables description with a column name, number of unique instances and example of content.

Numerical Variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Sales ($)** | **Quantity** | **Discount (%)** | **Profit** | **Shipping Cost ($)** |
| **Mean** | 246.491 | 3.477 | 0.143 | 28.611 | 26.479 |
| **STD** | 487.565 | 2.279 | 0.212 | 174.341 | 57.251 |
| **Min** | 0.444 | 1 | 0.000 | -6599.978 | 1.002 |
| **25%** | 30.759 | 2 | 0.000 | 0.000 | 2.610 |
| **50%** | 85.053 | 3 | 0.000 | 9.240 | 7.790 |
| **75%** | 251.053 | 5 | 0.200 | 36.810 | 24.450 |
| **Max** | 22638.480 | 14 | 0.850 | 8399.976 | 933.570 |

Table 3 Orders table numerical variables mean, standard deviation (STD), minimum and maximum values and quantiles (25%, 50%, 75%)

Some plots were made to better understand the shape of the data present In the Orders table. The number of orders was analysed by Ship Mode, Segment, Region, Category and Sub-Category. The plots can be seen below.

#### Graphs

##### Ship Mode

Chart, bar chart

Description automatically generated

Figure 1 Frequency of Orders by Shipping Modes

##### Segment

Chart, bar chart

Description automatically generated

Figure 2 Frequency of Orders by Segment

##### Region

Chart, bar chart

Description automatically generated

Figure 3 Frequency of Orders by Region

##### Category

Chart, bar chart

Description automatically generated

Figure 4 Frequency of Orders by Category

##### Sub-Category

Chart, bar chart

Description automatically generated

Figure 5 Frequency of Orders by Sub-Category

Various insights were obtained through the observation of the plots presented above. The majority of the customers prefer the Standard Class Shipping mode. Most of the customers are from Consumer Segment. The least of all orders are from the Central Asia region, and the majority are from Western Europe. Office Supplies take up most of the sales. Blinders and Paper are clear leaders in sales among customers.

### Sellers

The second table of this dataset is the table Sellers. This table contains 24 rows and two columns, which are presented below.

#### Columns

|  |  |  |
| --- | --- | --- |
| **#** | **1** | **2** |
| **Column** | **Person** | **Region** |
| **Description** | Name of the Seller | The region where the seller belongs. |
| **Data type** | Object | Object |

Table 4 Sellers table column number (#), name, description and data type

As it is possible to observe in Table 4, the table Sellers contains information regarding the name of the seller and the corresponding location (i.e., region)

To better understand the Sellers table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

#### Categorical Variables

|  |  |  |
| --- | --- | --- |
| **Column** | **Person** | **Region** |
| **Unique** | **24** | **24** |
| **Example** | **Marilène Rousseau** | **Caribbean** |

Table 5 Sellers table categorical variables description with a column name, number of unique instances, and example of content.

#### Numerical Variables

No numerical variables are present in this table.

#### Graphs

Plotting this table wouldn't provide much insight.

## Global Superstore Returns 2016.csv:

This dataset can also be found in the open source "data.world" (<https://data.world/tableauhelp/superstore-data-sets>). It has a total of 45.03 KB in csv format, containing one table, Returns. This dataset also contains information regarding the sales of the large-scale e-commerce chain.

### Returns

#### Intro

The table Returns displays 1079 rows and three columns, which are presented below.

#### Columns

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **1** | **2** | **3** |
| **Column** | **Returned** | **Order ID** | **Region** |
| **Description** | Boolean that refers if the order was or was not returned | Unique Order ID for each Customer. | The region where the seller belongs |
| **Data type** | Object | Object | Object |

Table 6 Returns table column number (#), name, description and data type.

As it is possible to observe in Table 6, table Returns contains information that shows if a certain order was returned or not, as well as the region of the seller of that order.

For a better understanding of the Returns table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

#### Categorical Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Returned** | **Order ID** | **Region** |
| **Unique** | 1 | 1079 | 24 |
| **Example** | Yes | CA-2012-SA20830140-41210 | Western Europe |

#### Numerical Variables

No numerical variables are present in this table.

#### Graphs

Plotting this table wouldn't provide much insight.

## Superstore Dataset

This dataset can be found on the open-source Kaggle website ([Superstore Dataset | Kaggle](https://www.kaggle.com/datasets/vivek468/superstore-dataset-final)). It has a total of 2.29 MB in CSV format, containing one table, which we will call costumers\_USA. This dataset also contains information regarding the sales of the large-scale e-commerce chain but now with a focus on the United States of America-based customers.

### Costumers\_USA

The table Costumers\_USA displays 51290 rows and 24 columns, which are presented below.

#### Columns

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **Column** | **Row ID** | **Order ID** | **Order Date** | **Ship Date** | **Ship Mode** | **Customer ID** | **Customer Name** |
| **Description** | Unique ID for each row. | Unique Order ID for each Customer. | Order Date of the product. | Shipping Date of the Product. | Shipping Mode specified by the customer. | Unique ID to identify each customer. | Name of the Customer. |
| **Data type** | Int64 | Object | Object | Object | Object | Object | Object |
|  |  |  |  |  |  |  |  |
| **#** | **8** | **9** | **10** | **11** | **12** | **13** | **14** |
| **Column** | **Segment** | **Country** | **City** | **State** | **Postal Code** | **Region** | **Product ID** |
| **Description** | The segment where the customer belongs. | Country of residence of the customer. | City of residence of the customer. | State of residence of the customer. | Postal Code of every Customer | The region where the customer belongs. | Unique ID of the Product. |
| **Data type** | Object | Object | Object | Object | Int64 | Object | Object |
|  |  |  |  |  |  |  |  |
| **#** | **15** | **16** | **17** | **18** | **19** | **20** | **21** |
| **Column** | **Category** | **Sub-Category** | **Product Name** | **Sales** | **Quantity** | **Discount** | **Profit** |
| **Description** | Category of the product ordered. | The sub-Category of the product ordered. | Name of the Product | Sales of the Product. | Quantity of the Product. | Discount provided. | Profit/Loss incurred. |
| **Data type** | Object | Object | Object | Float64 | Int64 | Float64 | Float64 |

Table 7 Customers table column number (#), name, description and data type

As it is possible to observe in Table 7, the table customers\_USA contains information similar to the Orders table, which goes from details about the dates of the orders and shipments, products categories and profits, with the difference that this table doesn't contain information regarding the seller, being its focus on the customer (USA based), providing more information, such as region, city, state, name, etc.

To better understand the customers\_USA table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

#### Categorical Variables

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column** | **Order ID** | **Order Date** | **Ship Date** | **Ship Mode** | **Customer ID** | **Customer Name** | **Segment** | **Country** |
| **Unique** | 5009 | 1237 | 1334 | 4 | 793 | 793 | 3 | 1 |
| **Example** | CA-2017-100111 | 9/5/2016 | 12/16/2015 | Standard Class | WB-21850 | William Brown | Consumer | United States |
|  |  |  |  |  |  |  |  |  |
| **Column** | **City** | **State** | **Region** | **Product ID** | **Category** | **Sub-category** | **Product Name** | **Postal Code** |
| **Unique** | 531 | 49 | 4 | 1862 | 3 | 17 | 1850 | 631 |
| **Example** | New York City | California | West | OFF-PA-10001970 | Office Supplies | Binders | Staple envelope | 42420 |

#### Quantitative Variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sales ($)** | **Quantity** | **Discount (%)** | **Profit ($)** |
| **Mean** | 229.858 | 3.789 | 0.156 | 28.657 |
| **STD** | 623.245 | 2.225 | 0.206 | 234.260 |
| **Min** | 0.444 | 1 | 0.000 | -6599.978 |
| **25%** | 17.280 | 2 | 0.000 | 1.729 |
| **50%** | 54.490 | 3 | 0.200 | 8.667 |
| **75%** | 209.940 | 5 | 0.200 | 29.364 |
| **Max** | 22638.480 | 14 | 0.800 | 8399.976 |

#### Graphs

Some plots were made to better understand the shape of the data present in the customers\_USA table. The number of customers was analysed by State, City and Region. The plots can be seen below.

##### States

Map

Description automatically generated

Figure 6 Number of Customers by State Map

Chart

Description automatically generated

Figure 7 Number of Customers by State (Top 25)

##### City

Chart

Description automatically generated

Figure 8 Number of Customers by City (Top 25)

##### Region

Chart, bar chart

Description automatically generated

Figure 9 Number of Customers by Region

Regarding these plots, three orders insights can be taken. The three most popular states among the customers are California, New York and Texas, and the least is Connecticut. The most popular city among the customers is New York City, and the 25th is Rochester. And most popular region among customers is the West region, and the least popular is the South region.

## US Holiday Dates (2004 - 2021)

This dataset can be found on the open-source Kaggle website ([US Holiday Dates (2004 - 2021) | Kaggle](https://www.kaggle.com/datasets/donnetew/us-holiday-dates-2004-2021)). It has a total of 15.7 kB in CSV format, containing one table, which we will call holiday\_USA, with information regarding the dates of US holidays.

### holiday\_USA

#### Intro

The table holiday\_USA displays 342 rows and six columns, presented below.

#### Columns

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **1** | **2** | **3** | **4** | **5** | **6** |
| **Column** | **Date** | **Holiday** | **WeekDay** | **Month** | **Day** | **Year** |
| **Description** | Date corresponding to holiday | Holiday designation | Day of the week when the holiday occurs | The month when the holiday occurs | Day of the month when the holiday occurs | Year of reference [2004-2021] |
| **Data type** | Object | Object | Object | Int64 | Int64 | Int64 |

Figure 10 holiday\_USA table column number (#), name, description and data type

As it is possible to observe in Table 10, table holiday\_USA contains a list of holidays that includes 18 years of US Holidays dated between 2004 and 2021. Each record has a Date, Holiday, Weekday, Month, Day and Year.

For a better understanding of the holiday\_USA table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

#### Categorical Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Date** | **Holiday** | **WeekDay** |
| **Unique** | 336 | 18 | 7 |
| **Example** | 2007-04-08 | Labor Day Weekend | Monday |

Figure 11 holiday\_USA table categorical variables description with a column name, number of unique instances and example of content.

#### Numerical Variables

No numerical variables are present in this table.

#### Graphs

Plotting this table wouldn't provide much insight.

## Highest GDP Counties in the USA

This dataset can also be found on the open-source Kaggle website ([Highest GDP Counties in USA | Kaggle](https://www.kaggle.com/datasets/anasmahmood000/highest-gdp-counties-in-usa)). It has a total of 3.65 MB in CSV format, containing one table, which we will call GDP\_USA. This dataset also includes data on the GDP of counties in the United States

### GDP\_USA

#### Intro

The table GDP\_USA displays a total of 55501 rows and eight columns, presented below.

#### Columns

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **Column** | **index** | **Year** | **Region** | **SUB\_REGION** | **State** | **County** | **GDP(Chained $)** |
| **Description** |  | Year of reference | The region of reference for the GDP | The sub-region of reference for the GDP | The state of reference for the GDP | The county of reference for the GDP | GDP value in $ |
| **Data Type** | Int64 | Int64 | Object | Object | Object | Object | Float64 |

Figure 12 GDP\_USA table column number (#), name, description and data type

As it is possible to observe in Table 12, the table GDP\_USA contains data on the GDP in dollars, of counties in the United States, organised by year, region, sub-region, and county

To better understand the GDP\_USA table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

#### Categorical Variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column** | **Year** | **Region** | **SUB\_REGION** | **State** | **County** |
| **Unique** | 18 | 8 | 9 | 51 | 1801 |
| **Example** | 2001 | Southeast | West North Central | Texas | Washington |

Figure 13 GDP\_USA table categorical variables description with a column name, number of unique instances and example of content.

#### Quantitative Variables

|  |  |
| --- | --- |
|  | **GDP (Chained $)** |
| **Mean** | 5.053577e+09 |
| **STD** | 2.154341e+10 |
| **Min** | 9.948000e+06 |
| **25%** | 3.366970e+08 |
| **50%** | 8.606620e+08 |
| **75%** | 2.488646e+09 |
| **Max** | 7.108930e+11 |

Figure 14 GDP\_USA table numerical variables mean, standard deviation (STD), minimum and maximum values and quantiles(25%, 50%, 75%)

To better understand the shape of the data present in the GDP\_USA table, some plots were done. The GDP (Chained $) was analysed by State and Region. The plots can be seen below.

#### Graphs

##### States

Chart

Description automatically generated

Figure 15 GDP by State (Top 25)

##### Regions

##### Chart, bar chart Description automatically generated

Figure 16 GDP by Region

It is possible to observe that the state with the highest GDP is the state of California, and the 25th is Alaska. In terms of regions, the one with the highest GDP is the Southeast, and the one with the lowest is the Rocky Mountains.

# 2. Analyse values and errors in the data fields of each source.

Data standardisation is a critical step in data analysis and pre-processing. It involves bringing different data sources into a common format, allowing for meaningful comparisons and accurate analysis. A critical aspect of data standardisation is analysing each feature's values and errors of all the used data sources. This involves carefully examining the data to identify any inconsistencies or anomalies, ensuring that the data is reliable and accurate. By standardising the data, organisations can make better-informed decisions based on consistent, reliable data, leading to improved efficiency and better outcomes.

Performing the analysis of each feature's values and errors of all the used data sources is a crucial step in data standardisation, and this can be achieved using programming languages such as Python. Python is a widely used language for data analysis, with numerous libraries and tools available that make it an ideal choice for standardising data.

In this context, the analysis of each feature's values and errors will cover all four tables, namely the orders, returns, sellers, customers, holiday USA and GDP\_USA tables. These tables contain different types of data, including transactional data, customer data, and sales data. The analysis will involve checking for any missing or inconsistent data, identifying any outliers or anomalies, and addressing any issues that may affect the data quality.

Using Python for this analysis makes it possible to automate the process and perform it more efficiently. This ensures that the standardisation of data is not only accurate but also timely. Additionally, using Python makes it easier to visualise and communicate the results of the analysis, allowing for better decision-making based on reliable data.

In some cases, a reoccurring situation that occurs in the tables is that some features are mistakenly represented as object data types when they should be represented as a different kind of data type. Object data types can be used for a variety of data, including text and alphanumeric characters. However, in certain scenarios, they may not be the most efficient or appropriate data type to use.

Using the wrong data type can result in a range of issues, including increased memory usage, reduced processing speed, and decreased accuracy. For example, if a feature such as a date is represented as an object data type instead of a datetime data type, it may be challenging to perform certain types of analysis on the data, such as sorting or filtering by date.

## Orders table

With the goal of standardising the data, a thorough analysis was performed on the table orders, which revealed several data type inconsistencies. First, some features that were meant to be dates were represented as object data types, prompting a change to the appropriate date format. Similarly, all information regarding the IDs (Order ID and Customer ID) was classified as an object data type, but to avoid potential issues stemming from mixed numerical and alphabetical characters, it was converted to a string data type. The same process was applied to other features like the City, State, Country, Region, Market, Product Name

Furthermore, there are other columns in the 'orders' table, such as Segment, Category, Sub-Category, and Order Priority, that are currently represented as an object data type but could benefit from a change to a category data type. This type of conversion can offer performance and memory benefits, as well as make it easier to perform certain types of analysis on the data.

Lastly, it was found that the postal code column had more than 50% of missing data. Since this column might not be used for the analysis, it was discarded from the dataset. The reordering columns had a proper data type classification.

## Customers table

Since the customer's table has the same kind of data type inconsistencies as the orders table, the same modifications were performed to standardise the data.

## Returns table

To improve performance and consistency, certain modifications were made to the returns table. Specifically, the "Returned" column, which previously consisted of "Yes" or "No" values, was replaced with a Boolean data type. Similarly, the Order ID in the returns table was changed to a string data type, following the same reasoning applied to the orders table. As for the Region column, instead of it being an object data type it was changed to a string data type.

## Sellers table

In the sellers table, the Person and Region columns were initially of the object data type. They were transformed to string data type to ensure consistency with the rest of the data.

## Holiday USA table

Like the previous tables, it was essential to modify the data types of certain columns in this table to ensure consistency with the rest of the data. The Date column was converted to a date data type to facilitate date-based computations. Additionally, the Holiday and WeekDay columns, which were initially of object data type, were converted to string type to ensure consistency across all columns.

To work with a subset of the data that is consistent with the orders dataset, only the rows corresponding to the years 2012 to 2015 were selected. Additionally, the Date column was in a different date format than the one used in the orders dataset. Therefore, the date format was converted from Y-M-D to D-M-Y to ensure consistency across all columns.

## USA GDP table

Several modifications were made to the GDP table to ensure consistency with the orders dataset and enable better data analysis. Firstly, the unnecessary columns, namely Region, SUB\_REGION, and County, were removed as they do not provide any useful information for the analysis.

Furthermore, it was observed that the orders dataset did not contain data for the states of Alaska and Hawaii, while the GDP table did. To orderstain consistency between the datasets, the data for Alaska and Hawaii were removed from the GDP table.

To gain a better understanding of the data associated with counties, the data was grouped by state. Two new columns were created to provide insights into the GDP of each state: Avg\_GDP column containing the average GDP of each state, and Total\_GDP column containing the sum of the GDP of all counties within each state.

Finally, the State column was converted to a string data type to ensure consistency with the rest of the data.

In order to verify the integrity of this dataset a preliminary analysis of the data was done.

No null values as well as duplicated values were observed.

Regarding the data type of the variables, the "Oder Date" and "Ship Date" were found to be object types which can become a problem in future analyses. To deal with this minor problem a simple data transformation will be performed to type datetime64 to obtain a simple ISO date.

No incorrect labels are present as well as undesirable characters.

# 3. Draw a diagram with links between data sources

After collecting all the necessary information, the following task was to construct a diagram that would illustrate the relationship between the data. The resulting diagram, presented in Figure 1, depicts all the tables along with their corresponding variables and the connections between tables that are established through these variables.

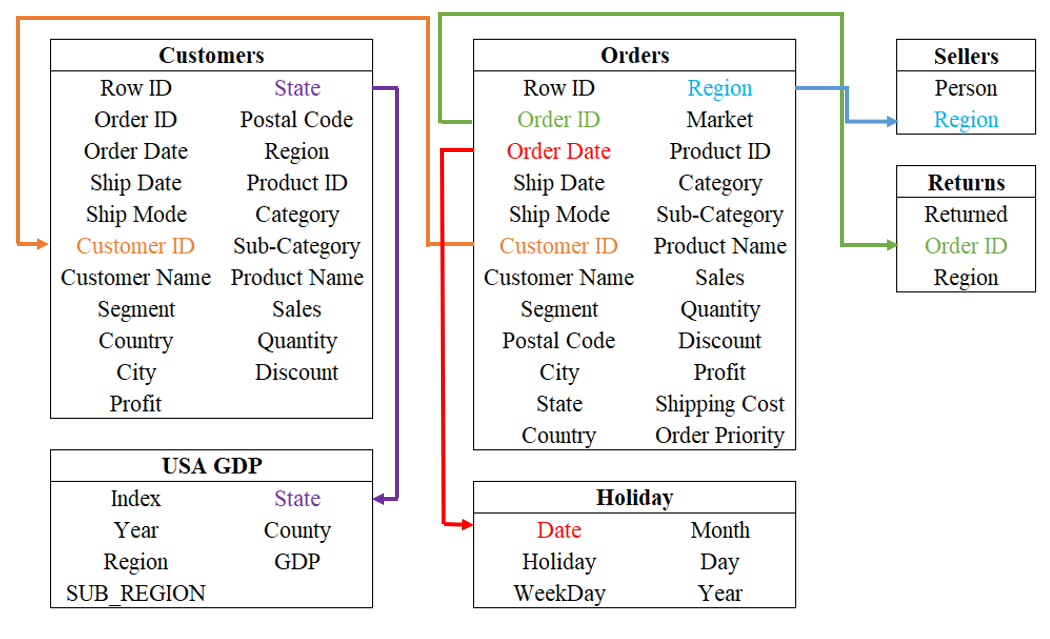


Figure 17 Diagram with all tables chosen for this project as well as their connections.

# 4. Describe a business process to take advantage of data

The datasets described simulate a sales process that is typical of many businesses. The data includes information on product sales, customer and store demographics, orders and shipping information, returns of products and promotions information, which could be used to analyse the performance of the business throughout time and place and identify areas for improvement.

For example, an analyst could use the dataset to:

Analyse sales trends over time and identify top-selling and most profitable products or categories. These values can also be analysed in different locations to determine if any are underperforming and what items to push in each.

Identify customer segments that generate the most revenue, where they are located, which products they tend to buy and how promotions influence their behaviour to adjust marketing strategies accordingly eventually.

Evaluate shipping performance in different stores and identify areas for improvement in delivery times or shipping costs. Also, evaluate if these factors relate to low sales or high refund count of other products.

We can also see what causes the poor delivery times and high shipping costs, and if the distance from the store to the customer, the type of product, order priority or the delivery mode can play a part in this. We then can use this information to augment customer trust and satisfaction.

By the orders information, we can determine which products are bought together and how the marketing is done on some items can influence the purchasing of others, then use this knowledge to better inventory items and better strategise which promotions would be most lucrative. We can also see which products' sales are not influenced by promotions at all so to cut marketing costs and which products' sales are especially seasonal so to better market them in the most profitable timeframes.

We can also analyse the trends of customer trust in different locations, evidenced by customers progressively coming back to spend more money and figuring out what better-performing sellers do differently.

With all this in mind, one business process will be the focus of this project: Product profit.

Overall, these datasets can be used to gain insights into the sales process of a business and identify opportunities for optimisation and growth.

# 5. Define three analytical questions for the business process

According to these datasets, it would be beneficial to understand which products, regions, categories, and customer segments this chain should target or avoid as well as the best and worst performing sellers. For that, three analytical questions aligned with the orders of three components that make up our business process (product profit) were created.

## Sellers Analysis

There are various questions relative to the seller that can be asked which are aligned with our business process (product profit).

Which seller generates the highest / lowest profit? This can be considered the orders question, from which we can start decomposing. Who is the seller with the highest or lowest profit? Other variables can be analysed to help answer this question, for example: Do the customers of the best seller belong to a state/city/region with high GDP values? The inverse can be asked for the worst seller as well. What is the most frequent customer segment and product category of the highest and lowest seller? Do discounts influence the seller's profit? All these questions can give various insights into the influences or at least what correlates with the highest and lowest sellers.

## Customer analysis

Which customers are most valuable and profitable? Which ones keep returning to the same store(seller)? Do customers increase confidence in the stores with time? For example, do customers increase the frequency and or amount spent on a particular store(seller)?

What is the percentage of customers, out of all customers of the superstore, that returned a product? What was the shipment mode that customers preferred the least for every region the superstore operates in? Are most of the orders of products of the "Furniture" category made by customers with the consumer or corporate segment? Does the customer importance relate to how often they ask for critical priority orders? And being more specific: How many customers that ordered products from the Europe market conducted critical priority orders in the year 2015?

These questions offer decision-makers greater insight into customer behaviour which can be critical to improving the business process of the superstore. Information such as frequency and distribution of the orders and returns, the number of individuals who were always satisfied with their purchases and the individuals who had issues with what they ordered on multiple occasions. Putting these instances under a microscope and studying the reason for them can be very benefitting for the superstore.

## Product analysis

What was the season of the year when orders for "GlobeWeis Clasp Envelopes" were made in higher amounts? Does this or any other product orders relate to the occurrence of holidays? What were the top 3 categories and respective sub-categories of products that generated the most profit for the superstore? For these three categories, what was the sub-category that had products with the highest discounts? Was there any crossover between the sub-categories for this answer and the sub-categories that generated the highest profit asked before? What is the average price of shipping that made customers not buy the product?

These questions, more centred around the products being sold, allow the decision makers to see clearly at what moments in a year a product is in higher demand and if they would benefit more from increased discount percentages at those times. What categories and sub-categories are the most profitable for the company is also indispensable information for the people tasked with deciding prices and quantities of future resupply orders. To be able to single out less profitable products is of worth to the superstore as studies can be made as to why these products are not performing as well as the rest and if needed discontinuation from selling them all together can now be supported from this analysis made available from being able to answer questions such as the ones asked here.