BUSINESS INTELLIGENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

OLIST

Group 008

Ana Amaro, number: m20200598

Filipe Lourenço, number: r20170799

Gonçalo Almeida, number: m20200594

Guilherme Neves, number: r20170749

June, 2021

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1. INTRODUCTION 1
2. OLIST: THE COMPANY AND SECTOR IN A NUTSHELL 1
3. BUSINESS NEEDS: ROOM FOR IMPROVEMENTS 3
4. DATA PROVIDED AND THE DISCOVERY PROCESS 4
5. ANALYSIS PERSPECTIVES: THE CORE PROCESSES TO STUDY 7
6. STAR SHEMA 8
7. FROM RELATIONAL TO DIMENSIONAL: DATA INTEGRATION, TRANSFORMATION AND MODELLING 10
8. TECHNICAL DEVELOPMENT OF THE PBI REPORT 14
9. WHAT DOES THE DATA SAY? 17
10. CRITICAL ASSESSMENT 21
11. CONCLUSIONS 23
12. DATASETS 23
13. REFERENCES AND OTHER SOURCES OF INFORMATION 23
14. APPENDIX – DAX MEASURES AND CALCULATED COLUMNS' FORMULAS 24

1.Introduction

In the following report, our group will attempt to answer some concerns Olist, a company operating in the e-commerce sector, presented to us regarding the way they conduct business.

Like many companies, Olist does not have in place a proper Business Intelligence system to support its decision-making process, which is creating some difficulties in figuring out the next logical move regarding several strategic plans the company has. In this regard, we will help the company build a system based on a Star Schema, that will enable faster and more informed decisions.

Although the company showed to be very supportive and provided us with a part of the database they have currently in place, they did not want to disclose information regarding their financial results, profit margins or other KPIs we feel would be valuable to the development of this project. Despite this, and not disregarding it will limit our approach in some ways, we will circumvent this issue by focusing on other areas that we, in collaboration with Olist, defined as being the top business needs and priorities.

2. Olist: The Company and Sector in a Nutshell

Olist is a Brazilian company operating in the e-commerce segment. It was created in February of 2015, and its goal is to serve as an intermediate between the stores that intend to sell their products online and several well-known online marketplaces in Brazil (e.g., Amazon, Mercado Livre and Carrefour).

Selling products in online marketplaces is not an easy process, as it not only requires a well-organized operations management system, but also creates the need to establish contracts and receive approval in every single marketplace the store is attempting to sell on, which can be very time and effort demanding. At the same time, the alternative of creating a self-e-commerce platform entails an investment that is prohibitive to many businesses. Due to this, many small stores do not have the capacity to enter these selling points which, with the evolution of e-commerce in the last years, is becoming a non-optative move. It was here that Olist spotted a business opportunity. By centralizing the digital operation management process - namely, the concentration of multiple online sales channels, orders requests, inventory management and order fulfilment and customers' reviews -, Olist provides a way for stores to circumvent the logistic nightmare associated with having to comply with multiple marketplaces' rules and operations methods, while also helping the online marketplaces to reduce the uncertainty related with dealing with multiple small parties.

At the moment of writing, Olist has four different solutions available for the stores:

1. Olist Shops: It offers the simplest solution, by providing a free smartphone application where stores can register their own products and sell them by sharing the link generated associated with their particular store. This solution represents an entry point to the companies that do not have their own e-commerce platform.

2. Olist Store: It includes the option for two different subscription plans (lite and pro), with different prices and characteristics, but the main advantage involves providing access to

the Olist's digital channel present in the most representative online marketplaces. This solution allows stores to have more visibility for their products (as they will be selling as part of Olist, a reputable player in these marketplaces), optimize their products catalogue and pricing with the help of e-commerce specialists, while having access to the main ERPs and centralized dashboards with useful information regarding their sales, and logistic associated, across online marketplaces.

3. Olist Premium: It represents an upgraded version of Olist Store, being orientated to large companies and personalized according to their specific business needs. It provides a complete integrated ecosystem, which includes the services decomposition in the diverse stages of operations (from the publication of products across channels to the shipment or customer support). With this plan, stores are represented in the online marketplaces as an official brand, and not just part of Olist.

4. Olist Pax: This solution is oriented for the stores that intend to gain logistics competitiveness with reduced costs, through faster and more reliable deliveries (from 4 to 15 days) and it can be seen as a supplement for the two previous solutions. Note that this solution was implemented after the period of the data we will work with.

Following this business model, by the end of 2017 (the data we will use in this report concerns the period from September 2016 to October 2018), Olist received, on average, between 300 and 350 new stores monthly, adding more than 2,000 new products daily to the online marketplaces.

With this impressive growth (with growth rates up to 10 times a year), Olist was able to place itself as one of the biggest virtual stores across digital marketplaces, in Brazil, representing more than 2,000 stores and having a portfolio of around 130,000 products. Moreover, across all the solutions provided, the annual average subscription fee was 170 euros, plus a 20% commission on every product sold, which allowed them to break-even only two years after opening the company.

On a more general note, the Brazilian e-commerce market, in 2018, registered around 123 million orders and 58 million different clients, moving more than 8 million euros. In this same year, the market grew 12%, when comparing with the year before, while the mobile commerce side alone grew more than 40%, exposing significant opportunities. The graph below summarizes the (constantly growing) evolution of the e-commerce sector in Brazil.

Figure 1: Evolution of E-commerce sales in Brazil.

Lastly, from what we gathered, the main reasons stores reported as being the motive to sell in online marketplaces are the sales lift, the higher engagement with more and diverse types of customers and the consolidation of the stores' brand name.

3. Business Needs: Room for Improvements

Notwithstanding being a fairly recent company, due to its impressive growth and the sector in which it is operating, Olist is already facing some massive amounts of data, that considering are coming from different organizations and operational steps (e.g., orders and products' details and customers' data), are very heterogeneous in its nature. Despite this, Olist was able to construct a well-organized database, that was provided to us in diverse CSV files, but that has the normal limitations for reporting purposes. Furthermore, we noticed that the organization seems to rely on excel files to communicate between business-oriented people, that seem to be constantly outdated and therefore creating many different versions and interpretations for the same phenomena.

With this in mind, the need for Olist to build an appropriate Data Warehouse (DW) and Business intelligence (BI) system has become imperatively clear, as this will allow them to manage in the appropriate way the high data stream faced, which will not only enhance their decision-making process but also the ability to communicate within and to the outside of the company. Moreover, this system should be able to store and integrate different types of data coming from distinct sources (with the needed cleansing process) to be analysed and reported in a way business value is created.

Another strong point of implementing a system like the one we intend, is related to how it would allow for data not to be scattered throughout the organization with several versions of what constitutes the right solution for each problem. This would ultimately allow for every single department (and employee) in the organization to have the same information at the same time, this is, to achieve "the single version of the truth" within the company, solving communication problems and allowing for homogeneity in the answers to the same problem (e.g., a person from the Sales Department will provide the same answer as an employee belonging to the Logistic Department when asked about "How many orders did Olist mediate last month?").

Furthermore, and although at this time we do not have the data needed to enable us to develop a Data Warehouse system transcendent to all organizational aspects, we expect that this project helps Olist to realize the importance of, in the future, expand the work done to the remaining organizational data it has. By doing this, Olist would be able to keep track of Key Performance Indicators crucial to the business, as this would allow for a more reliable and less time demanding decision-making process.

Now that the need for an accurate and consistent reporting system is clear, we can present the main points Olist wants us to provide them.

Firstly, Olist wants us to offer them a way to analyse the most important stores to their business, this is, the ones that represent the highest amount of revenue in transactions mediated by the company (1. What are the 5 stores generating more revenue?). This information will allow the company to identify their top partners and study ways to assure those stores keep doing business with Olist. Moreover, Olist realized that it is important to keep a large

enough active customer base to satisfice the stores in the platform, so we also want to understand what the dynamic between the number of stores and clients is, which should probably be defined as a KPI to track (2. What is the ratio of active customers for each store?). We believe this can be achieved by collecting and organizing information regarding the stores and deliveries and integrating this on the Business Intelligence system we want to create, so that different analysis methods can be applied on top of this data to face the need discussed.

The second business need discussed with Olist is related with their desire to maintain a good reputation among the end buyers. In this sense, Olist wants to understand if the customers are receiving their orders on time (3. What is the delay rate? and 4. How does the average rating given to a delayed order compares with a non-delayed one?), if there are specific locations in Brazil that face this problem more frequently and how delivery times vary across locations (5. What are the regions with highest average delivery time?) and seasons (6. Are there times of the year where the delay time tends to increase?). Moreover, it will also be interesting to study how the shipping costs affect the base line that customers pay (7. What is the decomposition of price paid in product and shipping costs?) and how those vary according to the characteristics of what is being shipped (8. Does weight category and volumetry play a role in the shipping costs?). Still regarding the end customer satisfaction, Olist also wants to explore the products and products' categories in which their customers are more interested (9. What are the products and products' categories generating more revenue?). All of this is also related to a new project and service Olist intends to implement, that would involve creating some warehouse locations to serve as the bridge between stores and customers, in a way that the delivery times would be significantly improved. By knowing the most affected places by delays and the most wanted products, Olist can show their partner stores the benefits of new subscription plans that include having inventory ready to go at Olist's warehouses. In this regard, we hope to mix geographic information with logistic and order-specific information inside our system to help the company prove the usefulness of this new service (10. What regions would benefit the most from this and what products' categories need to be present?). As a last point, regarding the customer satisfaction, Olist also wants to study if there is a viable solution to implement their own credit payment method, instead of just providing the option to pay in instalments, as this trend is growing in the industry, and Olist fears being left out (11. Are payments with credit card the most used? And if so, are people mixing it with the option to pay in multiple instalments?).

The last business need discussed, concerns Olist's desire to have a small-window time frame tool to allow them to predict the number of products they will sell, for instance, next month. This tool will allow them not only to stablish more advantageous conditions with the carrier companies, but also to help partner stores (in this case the Olist Premium subscribers) to understand better their inventory needs and to be one step ahead when compared with their competitors (12. What is the forecast of demand for the next few months?).

We believe that by recurring to the Power BI software, we will be able to create a system that is able to satisfy all the needs presented by Olist, helping them solve some business problems they are facing, but also to allow them to transform data into information and to be able to use it in the process of making data-driven decisions.

4. Data Provided and the Discovery Process

Now that the reasons that lead Olist to contact us were discussed, it is time to focus more concretely on what data we had access to. In this regard, we had two sources of

information: Kaggle and the Brazilian website "Rede Suas". From Kaggle, we obtained 9 different datasets, stored in CSV format regarding Olist's operations (more concretely from what looks like data from the Olist Store and Olist Premium). On the other hand, from "Rede Suas" we obtained data on the Brazilian states and regions, which will help us to construct hierarchies and to have a more detailed view of the flows of orders to different Brazilian locations.

Given Olist's ability to store data in a clean way, the discovery process was fairly easy. From the 9 datasets, we had information regarding: the sellers and buyers (mainly geo-data), the orders (which included many different aspects, like the number of products in each order, the sellers involved, the shipping estimates and actual deliver dates, the shipping costs, the reviews and the payment method used) and the products (including the category, the dimensions and the number of photos provided). The "Rede Suas" dataset, as mentioned, will help us to add one more location hierarchy level, and it will be connected to the Kaggle ones by the State entry. The remaining interactions between datasets and the information in each are depicted in the schema below.

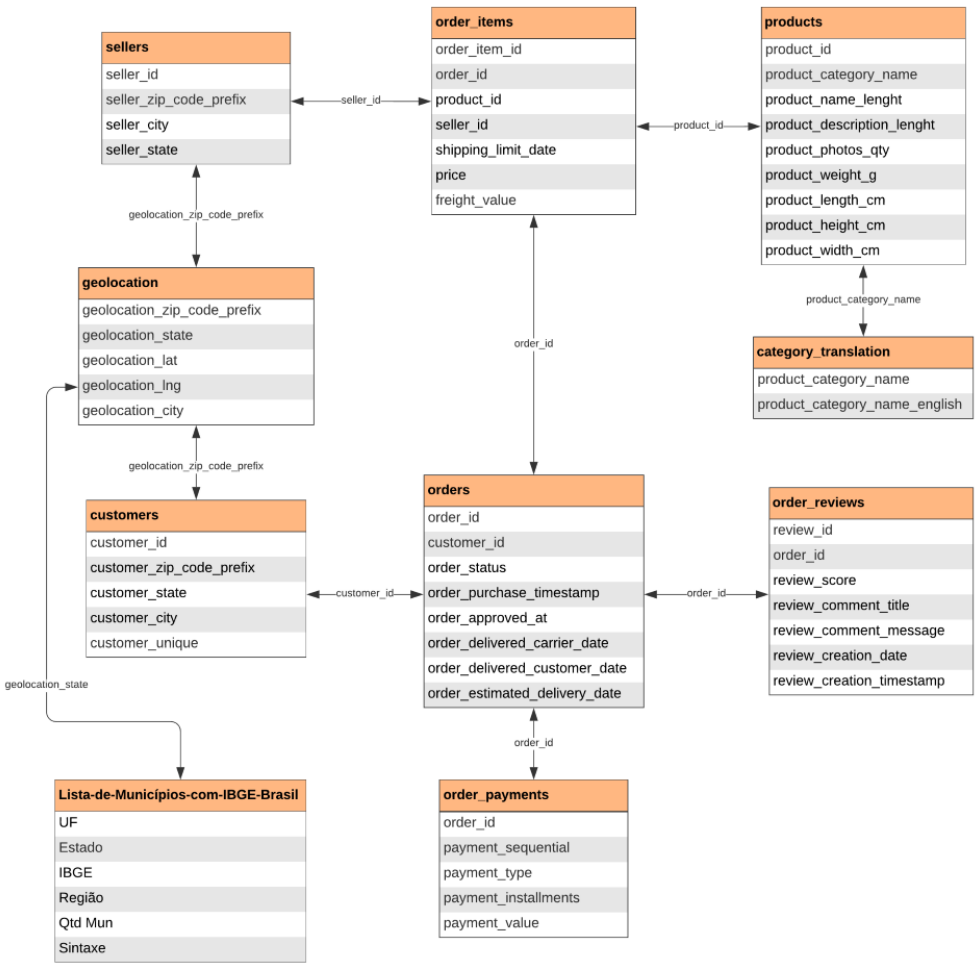


Figure 2: Original Database Schema.

Moving now to a more explorative approach of our dataset, we found out that Olist provided us with data from 3095 partner stores, 96096 customers, 32951 products and 98666

different orders , which implies that although few, we have repeated customers. Furthermore, and in order to get deeper insights, which will allow us to better define the perspectives of analysis and create a better dimensional model, we decided to elaborate some simple graphs, as a quick data visualization aid.

Below, we can clearly see that although there is a fairly high number of orders being reviewed with a positive score (4 and 5), showing a strong customer satisfaction with the orders mediated by Olist, the amount of orders with score of 1 is still a major concern, as those customers, by definition, seem very disappointed with the outcome and/or process of the order made and may be very unlikely to repeat a purchase mediated by Olist. As mentioned, this is a particular concern for Olist, as they feel some of their partner stores may be damaging their reputation.

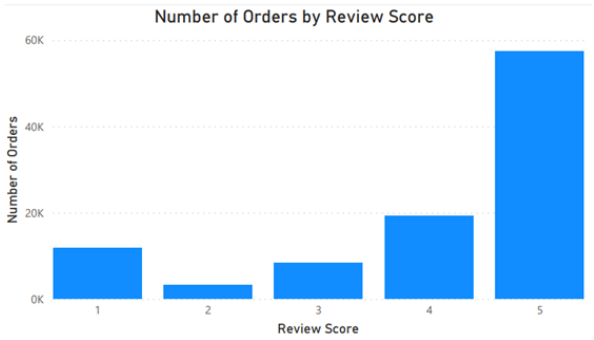
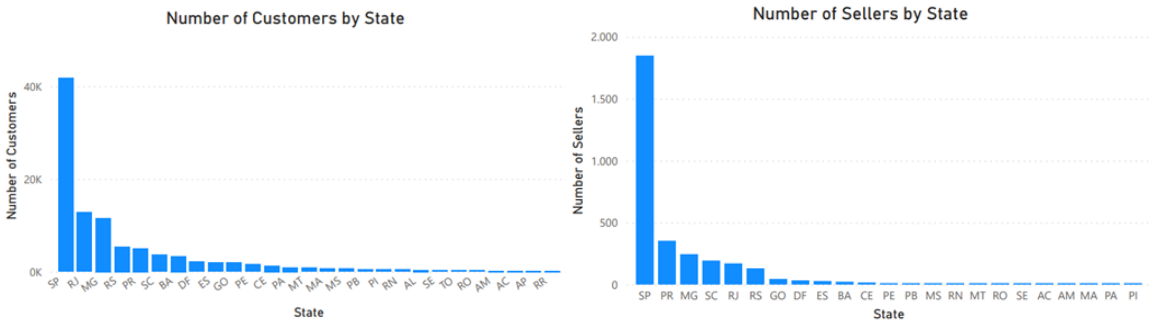


Figure 3: Number of Orders by Review Score.

On a different note, by looking at the dispersion of the partner stores and customers across Brazil, we can clearly see that the state of São Paulo accommodates the largest portion of our buyers and sellers, allowing for a good balance in the most populated Brazilian state. Nevertheless, and considering the large populational dispersion of Brazil, we believe this match may not be achieved in every state, or even inside each state, being this also a worthy point to explore further.

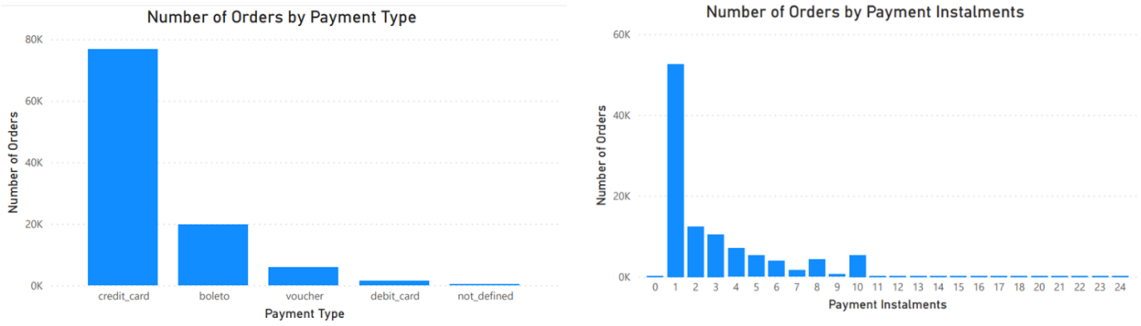


Figures 4 and 5: Number of Customers and Sellers by State.

Regarding the number of orders across time, and although there seems to exist some data missing for the first and last few months of analysis, the pattern of expansion is clear, and we will explore it further later on, while looking at other temporal patterns.

Figure 6: Distribution of Orders Across Time.

Lastly, and considering the importance of having payment options available when talking of online purchases, we decided to also take a look at this particular aspect, searching to find out if any interesting pattern emerged. In this regard, we can clearly see that the credit card is by far the most used mean of payment, and although a scheme of partitioned payments is offered, a large amount of customers opts to pay the full amount up-front.



Figures 7 and 8: Number of Orders by Payment Type and Number of Instalments.

5. Analysis Perspectives: The Core Processes to Study

Considering the needs presented by Olist and the data provided, we decided to divide our problem into two different perspectives: one from the customer side and the other from the logistic point of view. We once more reiterate that a more complete work could have been developed if Olist also had shared with us more operational and financial data, like the subscription plan distribution of the stores.

Regarding the customer analysis, there are several aspects that we can explore and that will bring value to Olist, but firstly it is important to clarify that, as Olist connects sellers and buyers, they have to take into account the two sides of the coin, considering there is the need to establish partnerships with stores, but also to have customers purchasing from them. Following this logic, no analysis of the customer side could start without previously analysing the industry in which the company is inserted, which we partially already did when understanding the company's industry.

Based on the previous statements and given the availability of geographic data, we expect to be able to understand the locations with more and less volume of purchases, which will help to understand the dispersion of both the buyers and the stores. Furthermore, it will also be important to help Olist to identify which types of products are more purchased and seen

as more valuable by the end customer, which is verifiable when analysing the volume of the orders and products type. On this same note, and given the nature of the business, it will also be of interest to analyse how the purchases change across seasons, months or other significant temporal divisions. Lastly, it will also be an interesting experience to understand which dynamics exist between stores and buyers. We believe this analysis will help Olist to better understand its customer base (both stores and final buyers), fulfilling therefore parts of both the first and second business needs.

Concerning the logistics perspective analysis (with logistics being defined as "the overall process of managing how resources are acquired, stored and transported to their final destination" (Kenton, 2020)), and considering the importance of this perspective due to the reputable place Olist wants to occupy, it is fundamental that the customers are happy with, not only the products (that is not controlled directly by Olist), but also with the logistics themselves, in which Olist has a strong saying. In this regard, not being able to receive the products in time is a significant reason for concern for the end buyers, which may lead them not only to give bad feedback about some products but also to disregard a specific brand or store in the future. On this front, we expect to construct a narrative regarding the locations where there seem to exist higher delays in delivery of products and also the seasonality of delays. This information will help Olist to understand where it should probably advance with the strategic plan of having some "in-inventory" products and in which times of the year should expect delays more frequently, given the excessively high lead time, which ultimately will have a high focus on satisfying the second business need introduced by Olist. Still regarding this perspective, it will also be interesting to track the costs associated with deliveries, in order to, when deploying the "in-inventory" system, analyse how efficient is Olist being (keeping the costs on the same level or lower), and also if it would be a good strategy to have a program which could involve free shipping costs for specific minimum amounts of purchases, keeping the customer as a priority and passing to them the part of the savings Olist would be achieving with this initiative.

6. Star Shema

For this particular problem, we decided to develop a Star Schema as the model for our data. On this note, we created five dimensions (Product, Store, Customer, Date and Payment) and one fact table.

One of the first points we would like to address, is the fact that we do not have the names of the products, sellers or customers. This is mainly due to privacy protection concerns, and although this is the case for us, we know that Olist has this sort of information. Due to this, we decided to use the hashes as if they were actually the names of the products, stores and sellers. We acknowledge that we will end up having two keys per dimensional table (one surrogate key that will be linked to the fact table and one key that will work as the name of something), but considering we intend to later on have Olist share this information with us (once a non-disclosure agreement is signed), we believe is better to proceed this way and whenever we have the "translation" (probably after the full report is delivered and we show them we are able to solve the needs provided), it will only just require a small step to replace the hashes with the actual names. On a nutshell, what this means is that we will use the hashes as nominal data and not as a pure ID and, when possible, the replacement will be made. Note that this does not compromise our ability to apply Business Intelligence concepts, but only shows that companies

prefer and are only allowed to share some types of data with people inside the organization, and due to this, we should not be penalized because in a "real-world" scenario this data would be available to us, as it was just a matter of asking it to the organization that hired us, so we worked around it in the way possible.

Moving now to the dimensions, regarding the Product Dimension, it will be composed of a 3-level hierarchy, with the Weight Category → Product Category → Product, allowing to understand the dynamics of the products purchased, not only by the category they are inserted in, but also by the weight of the product itself (as in online purchases this may be an important concern). This category of weight will have to be created based on the weight's distribution of our products, by trying to find at least 3 different categories that will help us to better understand some dynamics, like for instance the shipping costs differences. Furthermore, on this same Dimension, we will also have information regarding the products' volumetry (that will be computed by multiplying the length by the height and by the width).

The Store and Customer Dimensions will each also have a 3-level hierarchy, including the Region → State → City, which, as wanted, will help to illustrate the dispersion of our customer base (both sellers and buyers) across Brazil. In this regard, we decided not to include a Geographic Dimension, as we believe that would be more appropriated to physical stores (owned by us) and not in this dynamic, where we manage the connection between sellers and buyers all across Brazil, being that we are interested in the "two sides of the equation".

In what concerns the Date Dimension, it will include a 5-level hierarchy, with Year → Quarter → Month → Day → Time of the Day, in order to allow us to explore the main business temporal divisions of interest: a year basis for more general trends, a quarter period that is associated with the need to report results, a month and day that are more helpful from the operations point of view, and finally, a time of the day (e.g., morning) that can help to uncover some buyers' patterns.

Lastly, we will also include a Payment Dimension. This dimension will not have any hierarchy, but will be useful to store information regarding the payment method used and the number of instalments requested to pay the order. In the particular case of this dimension, it is debatable that it could have been deconstructed to have the information present in the fact table, but we decided this way would allow us to better explore the combinations of different payments (both in type and instalments), and how they vary across some of the measures we have in the fact table (e.g., the relation between amount of the order and number of instalments). Although not being a direct business need presented by Olist, we believe this dimension may give us some insight regarding the possibility of Olist to create customized payment programs, and we will later on this project attempt to see if this is in fact a possible feature we can, in a data-driven way, suggest to Olist.

In a nutshell, our model presents three different hierarchies (with a repeated one), being on average composed by more than three levels.

Last but not least, regarding the Fact Table, we decided to include three measures referent to the logistic perspective of analysis and four concerning the customer perspective. In the logistic perspective, we will have as measure the delay time and the shipping costs and time, that we will mainly want to analyse how they vary across temporal periods and locations, but

also across the number of products per order. On the other hand, the customer perspective will be measured in terms of number of sales, both in monetary amounts and in volume, in evaluation score and will also have a more nominal measure, which will be the written review itself. From these measurements, we may want later on to construct new ones and explore them across the diverse dimensions of our analysis. Lastly, we have also added a variable named Order Number as it will help us to aggregate different orders in an easier way. In figure 9, we provide the Star Schema with not only the variables already explained, but also with the keys linking the different dimensions to the Fact Table.

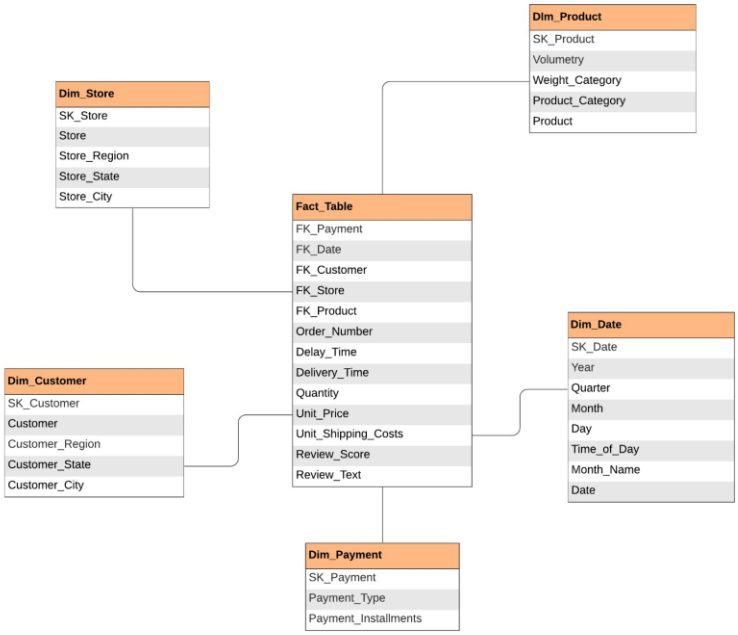


Figure 9: Suggested Star Schema.

7. From Relational to Dimensional: Data Integration, Transformation and Modelling

As discussed in section 4, the data Olist provided follows a more relational structure, which meant the first step needed was to adapt its format to our dimensional model. In this regard, Olist had previously warned us that there were some problems with their database - orders paid with multiple payment methods (where some variables had different meanings from the original), customers with more than one address (due to an improper migration that led to the existence of customer_id and customer_unique_id) and orders paid with no instalments, among other small issues-, which we were asked to disregard from our analysis, as they were due to improver updates in their system and internal operational errors. Moreover, and contrary to what we originally thought, based on the dimensional model created (which reflected the business needs presented by Olist), we decided to exclude from this report the table Geolocation, which did not contain relevant information for the problems at hand.

Note that in the following lines, for readability purposes, we will use the meaning of the variable as its name (and not the official name present in the model), meaning this, for instance, the variable payment_instalments will be addressed as Number of Instalments. Additionally, for brevity reasons, we tried to omit non-important steps (e.g., deleting variables that were not useful for our analysis and converting column types to the proper ones), but they can all be consulted in the Query Editor.

Being this said, after importing the excel and CSV files, we started constructing the dimensional model. Firstly, in the query editor, through an empty query, we created the Dimension Date, where we not only constructed dates for the full period of transactions but also manually (using M commands) created the Year, Month, Quarter, Day, Time of the Day (following a sequence of IF statements that we believe to be a natural division of the day at the hour level) and Hour, being this last one just an auxiliary column that we will later on disregard. Furthermore, we created a Surrogate Key, that similarly to what we have done in all dimensions, was created using an index column starting on 1.

Using the table olist_sellers_dataset (and renaming it as Dimension Store), we performed a left merge (giving a full match) between this table and the one extracted from "Rede Suas", by using the State UF as the linking element, allowing us to get each seller's State Name and Region. Additionally, we capitalized the sellers' cities' names (to improve the appearance and to make it more consistent), added to the State Names the word "Brazil" (we could have added a new column with the information of the Country, but given all operations being inside Brazil, we believe this to be a simple and more efficient solution just to identify states that might have the same name in other countries) and created a Surrogate Key, following the same process mentioned above.

In the table containing information about the customers (renamed from olist_customers_dataset to Dimension Customer), we repeated the merge done in the Dimension Store (with the "Rede Suas" table), obtaining the same geographical context. After this, recurring to the "group by" command, we were able to identify the customers who had more than one city associated with their account, which, as requested by Olist, we decided to filter to only keep the ones without this problem. Before proceeding, we created a duplicated table out of this one, that we will later use to link to the Fact Table.

Considering one of the problems that Olist informed us about, regarding the fact that a customer_unique_id can represent more than one customer_id, and as we wanted to keep a unique identifier for each customer in the Dimension Customer, we decided to disregard the customer_id. For this, we made a "group by" focused on the customer_unique_id, which allowed to have for each line at this dimension a unique customer represented. With this problem solved, we proceeded to create the Surrogate Key for this dimension, in the same way as described previously.

The Dimension Product was constructed from the table olist_products_dataset, and apart from adding a Surrogate Key for each product (and removing the non-used columns), the only thing we needed to do was a merge with the table product_category_name_translation, to present the product categories in English, as they were originally in Portuguese. Note that here we decided to use an Inner Join, as it permitted us to only keep products that were under the

official categories defined by Olist, given that we detected some problems, like products without category and products that had categories that did not match any of the official ones. Some other solutions could involve naming the products belonging to problematic categories with "Other Category" (which would be equally easy to implement), but as advisable in these cases, we contacted the department responsible for the data collection, which assured us that if the products did not match the official categories, it was due to some mistake in the registering process and therefore we were requested by Olist to not consider those.

The last dimension, Payment Dimension, was built from the information present in the table olist_order_payments_dataset, and the first issue we had to solve was related to the payments that had more than one payment method associated, which although not being wrong, in this case, presented a conflicting behaviour with the other variables stored in this table, making, for instance, the variable with the Number of Instalments not valid. To solve this, we used a "group by" focused on the ID of the order and filtered any order that had more than one payment method. Then, we duplicated this table, to be used further ahead, and made a "group by" using the Payment Method and the Number of Instalments, allowing us to get all possible combinations used of both variables. Being this done, we had still to filter orders with the Number of Instalments equal to 0 and orders with the Payment Method being "not defined", as they represent non-useful cases, where there was an error in the data collection phase. Lastly, with only valid combinations, we created the Surrogate Key for this dimension.

Before proceeding to the modelling of the Fact Table, we needed to prepare the Olist's original tables we wanted to use:

- At table olist_orders_dataset, we filtered the orders to only keep the ones that were already delivered (as at the time of data collection there were orders in process and/or that although delivered did not have information about the delivery) and not cancelled (although this information might be useful for some analysis, considering the business needs Olist presented us, we decided that none would fit the possible exploration of the cancellations).
- From the table olist_order_reviews_dataset, using a "group by", we detected some orders with more than one review, which should not be possible, as the reviews should be order related. Based on how Olist told us the review process works, we concluded that this problem may be related to people sending repeated reviews and/or regretting past ones and attempting to update them. As it is of interest to explore the reviews made, we decided it would be better to exclude these residual ambiguous cases from analysis by filtering them out, as we would not be able to be sure about the true opinion about an order.

After this, we renamed the duplicated table we created from the Payment Dimension to be the Fact Table and, keeping only the ID of the orders and the Surrogate Key associated (which was renamed as Foreign Key of the Payment Dimension), we started to construct our last table. Note that as we performed diverse modifications to the data existing in the tables we want to merge our Fact Table with (to fix the problems previously identified by Olist and the ones we found along the process), the following merges that we will describe will all be Inner ones. This way, we keep only the intercept of both the valid orders and the ones associated with valid Foreign/Surrogate keys (e.g., we removed customers associated with more than one city,

but their orders will probably still be on the remaining tables, thus, if we did not perform an Inner merge, those entries would correspond to an empty link to the Customer Dimension).

Using the ID of the Orders, we merged the Fact Table with the olist_order_reviews_dataset, the olist_orders_dataset and the olist_order_items_dataset, allowing us to retrieve information regarding, respectively: the review's score and text; the timestamps of the orders and customers associated with each order (Customer ID); information about the products and sellers associated with each order (Product ID and Seller ID) and the shipping and unit price of each product in each order.

Through the Product ID, Customer ID and Seller ID, we performed a merge with the Dimension Product, Customer and Store, correspondingly, retrieving the Surrogate Keys associated with each to establish the bridge and renaming them as Foreign Keys. Finally, we connected the Fact Table with the Dimension Date recurring to the timestamp associated with the time the purchase was made (at the hour level, being this the reason we needed to create an hour column in our Date Dimension).

As the last step in the construction of our dimensional model, and although this column was not originally part of our model (as it should be seen more as a tool than a column in itself), we used the advanced capabilities of Power BI, in this case, text analytics, to extract key phrases from the field containing the review text (from the Fact Table).

Before leaving the Query Editor, in all dimensions, we also sorted the columns by the Surrogate Keys and ordered the tables by positioning the Surrogate Keys (and Foreign Keys in the Fact Table) as the first column. Unfortunately, due to some known limitations of the Power BI software while working with large datasets, this sorting, although being recorded in the Query Editor, was not transferred to the Model and remaining views.

At this point, to comply with the proposed dimensional model, we still needed to construct the hierarchies, hide non needed variables and construct others.

Using DAX calculated columns, we constructed in the Fact Table the variable Delay Time and Delivery Time by using the "DATEDIFF" function (this DAX function was not taught in class, and accounts as one of the opportunities of "extra work"), between the dates of order, delivery and estimated delivery. Moreover, in the Dimension Product, we constructed the variable Volumetry (from the length, width and height), by using the DAX multiplication element ("*") and the variable Weight Category, using a sequence of IF statements to subdivide the numeric variable Weight into bins ("Very Light", "Light", "Heavy" and "Very Heavy"). Note that although the next section being where should be presented the calculated columns, we presented it in this section, as it makes more temporal sense, having us already achieved the requisite of computing a calculated column using DAX.

After this, in the Model View, we constructed the hierarchies, following exactly the purposed structure and hid all variables that although used until now, are not part of our dimensional model and have no interest in the remaining analysis. As the last step, we sorted the Month Names by the Month Number (both present in the Date Dimension) and the Time of the Day by means of an auxiliary column (created previously for this purpose) to assure no

problems would arise later due to this and defined the monetary amounts as currency, particularly as Reais, being this the only data type we did not format in the query editor.

8. Technical Development of the PBI Report

Regarding the Power BI report, our decision was to divide the business needs to be studied into 5 different tabs, creating two sub-reports, one for each perspective, with two tabs for the Logistic Perspective and three for the Customer one. Additionally, we also created a home page to make the experience more user-friendly.

For the Logistic Perspective and considering the need to study the delivery time and costs, we will use this division, dedicating one tab to each.

In the tab exploring the Shipping Costs, we decided to offer the possibility to select both the year of analysis and the category of weight, which will adjust the full report tab based on it. Being this said, we present to the user three different visualizations, each exploiting a different perspective on how the shipping costs vary. To check if, in fact, and as expected, there is a positive relationship between the shipping costs and the volumetry, we constructed a scatter plot, to which we added a trend line, being that we decided to define the observations as the different product categories. For this particular visual, we had to create two DAX Measures, one with the average shipping costs (given by the division of the multiplication, using SUMX, of the unit shipping cots by the quantity, by the total quantity of units sold) and the other with the average volumetry (following the same formula as for the average shipping costs, but instead of multiplying by the shipping costs we multiplied by the RELATED product volumetry, from the Product Dimension). Moreover, based on the user's selection of the slicers (year and weight category), we have a donut chart that will update the percentage of an order corresponding to the shipping costs (using the created measure of average shipping cost) and the percentage associated with the price of the product (creating a measure following the same logic as the average shipping costs, but tracking the average unit price), which can help to quickly assess how this behaviour changes with the weight category selected. Lastly, we have constructed a tree-map with the average shipping cost per state. Note that the user can also select, in each visualization, a specific state or product category and the remaining visualization will change accordingly, as if it was an implicit slicer.

Moving now to the Delivery Time Tab, we provide two slicers, one with the possibility to select the year of analysis and the other with the regions and states. For each selection, employing a funnel chart, we will display the 10 states (in case the selection leads to at least 10 states, being this accomplished with a filter in the visualization) with the highest average delivery time (also giving a notion of how the state with the highest average delivery time compares with the others in percentage terms), allowing to understand the locations most affected by slow delivery processes, as requested. Here, we needed to create a new DAX Measure with the average delivery time, carefully using the function SELECTCOLUMNS (this DAX function was not taught in class, and accounts as one of the opportunities of "extra work") combined with the function DISTINCT to capture only different orders (otherwise we would be having duplicated information, as if two products are from the same order, they will be processed at the same time, for obvious reasons). Note that for the following visuals (in this tab), every measure

created will always follow this logic of combining the functions SELECTCOLUMNS and DISTINCT to assure the same correctness of results.

Regarding how the delays vary across months, we have constructed a DAX Time Intelligence Measure, "TOTALYTD", that we used in an area chart, showing the evolution of the cumulative sum of days that orders made, in each month, got delayed (e.g., if in January we had 3 orders delivered late, accounting for a total of 15 days, and in February we got another order that was delayed 5 days, in February our graph would have the value of 20 days), which presents a very clear perspective of how the delays evolve along the year. Moreover, in order to relativize the value of cumulative delay time (by considering other metrics for each month, which may aid in a "root-cause" analysis), we have added a tooltip (created on a separated hidden tab, named "Delays Tooltip") that presents 3 different cards (excluding the one just giving the exact value of the cumulative): one with the total amount of orders, other with the percentage of delayed orders and the a last one with the average delay time for each month of analysis.

The average impact on the satisfaction of customers when the order is delayed was presented with an intuitive "ratings visual" in the format of stars (this visualization from the Marketplace was not taught in class, and accounts as one of the opportunities of "extra work"). To do this, we had to compute two DAX Measures: one computing the average of the review's scores filtered by when the orders were delayed and the other when these were not delayed.

Finally, we decided to include two KPIs, that we believe to be natural given the phenomena we are studying. The first one is regarding the percentage of delayed orders that we set as to having to be lower than 5% (as Olist requested when we asked about this), and this is a KPI not only for the general panorama, but that should also apply to each state and city, individually. This KPI was constructed with the help of a DAX Measure that divides the number of rows containing delayed orders by the total number of rows, being that we had to use the function "COUNTROWS" to do this (this DAX function was not taught in class, and accounts as one of the opportunities of "extra work"). The last KPI is a little more complex, as it concerns the average delivery time, and as Olist explained to us, setting a fixed threshold would not be a good option for this particular business context, as it is perfectly normal that different regions present different delivery times (contrary to the delays, that Olist finds intolerable for any state, city or region). In this regard, and being Olist expanding their operations exponentially (as discussed while presenting the sector), it was considered that it would be an already good result if the delivery times did not get worse than in the same period of the previous year, which would prove the robustness of the scalability of their operations. Therefore, we created this KPI with the threshold being defined as the value of the homologue period of the previous year for each geolocation (being it overall Brazil, the regions, states or cities). To achieve this, we had to use a second DAX Time Intelligence Measure, in this case, the "SAMEPERIODLASTYEAR".

In what concerns the Customer Perspective, and by knowing that Olist has two types of customers (stores and buyers), we will use this as a natural division of Olist's business needs, making this analysis in two different tabs, one analysing the buyers, and the other analysing the sellers. Moreover, we decided to add a third tab, where we will explore the preferences ("Popularity") regarding products and product categories.

Starting with the Popularity Tab, we allow the user to pick from a slicer not only the year and quarter but also the time of day in which he wants to perform the analysis, as we know this can bring a lot of value, for instance, to understand at which parts of the day specific categories' promotions and other marketing efforts should be done. Furthermore, we have also given the possibility for the user to select the region and state to filter. According to the selection purposed, we have two cards: one displaying the total sales amount and the other the sum of the number of products sold. To construct the sales amount, we had to create a DAX Measure, in this case using the function "SUMX", multiplying, at a row-level, the quantity and unit price of each sale. Note that sales amount excludes the value of shipping costs, as those are not a "gain" for Olist.

Moreover, using a visualization from the Marketplace, Word Cloud, we constructed a way for the user to visually see the most important words in the reviews, taking advantage of the embedded property of this visual to filter out the words that do not bring value ("Stop Words"), while recurring to the already mentioned visualisation from the Marketplace ("ratings visual") to present the average review score. For this last visualization, we needed to create a DAX Measure with the average review score, and similarly to what we did in the tab with Delivery Time, we had to combine the functions SELECTCOLUMNS and DISTINCT, in order not to double count the scores given for the same order. Finally, we present in a bar chart the top 10 (filtering in the visualization) most sold products according to sales amount, and also in table-format, the top 10 and bottom 10 categories of products, according to the same metric. Note that by selecting a product from the bar chart or a category from the tables, the review score will be updated to present the average review score of orders containing those products/products' category.

For the tab with the Store report, we present a slicer with the possibility to choose the year and quarter, which will change three visualizations, leaving one static, independently of the selection. Starting with this last one, we decided to accommodate the desire of Olist to have a forecast of the number of units sold, to be able to anticipate demand, allowing for negotiation of better prices with the shipping companies and giving their partner stores not only better deals (as customers will be attracted to lower prices due to shipping costs reductions), but also the possibility of better understanding inventory needs. In this regard, we used a time-series forecast, setting the predictions to be until the end of 2018, using as seasonality the year (365 days, as we have daily data) and setting a confidence interval of 95%.

Regarding the other three visualizations, the first one is a pie chart with the sales amount by the region the seller generating them belongs to, being that this chart also works as a slicer, in the sense that it will adjust the other two (non-static) visualizations according to the region selected. Then, we constructed a bar chart with the top 5 stores in sales amount (being that we needed to use the filter option in the visualization to limit to the 5 with the highest value), and lastly, we constructed a map visualization with the dispersion of stores across Brazil, allowing it as well to be customized by the selection chosen. In this map visualization, we also provided a tooltip (created on a separated hidden tab, named "Seller Map Tooltip"), which is "activated" when the mouse is hovered over a specific state, presenting a bar chart with the top 5 cities where the stores generate most sales amount on that state (once more, we filtered directly on the graph by the top 5) and 3 cards: one with the number of stores, the other with

the sales amount and the last with the average score, corresponding to orders generated on that state.

Concerning the tab of the final Customers, we also present the map (including the tooltips, but now constructed on the hidden tab "Customer Map Tooltip"), the pie chart (with sales amount per state) and the bar chart (with the top 5 customers in sales amount), working all exactly as explained above, but from the customers' point of view, instead of the stores' one. For this tab, we also decided to include a slicer with the product category (adding to the possibility present on the store report with the year, quarter and region), as we believe it can be of interest to understand which states' customers purchase more and less of certain products' categories (e.g., different climates, cultures and/or religions across Brazil, may imply different product tastes).

Exclusively for this tab, we added two visualizations. Firstly, a simple donut chart with the payment methods (that uses a tooltip, created in the "Payment Type Tooltip" hidden tab, with a bar chart when hovering over a specific payment method showing the number of instalments used for each payment method and the number of times that combination was used), as we believe this can help Olist better understand which customers tend to use specific payment methods, and which number of instalments they usually choose. Secondly, a KPI, which will track the ratio between customers and stores: after discussing with Olist, we realized it is of extreme importance to keep stores engaged by having a high enough active customer base, and we were told that this would be important to follow not only at a general level but also by region (being that for region the idea is to track the number of customers from a region in relation to the sellers that shipped to that region in a specific timeframe, independently of the seller's geographic location), keeping in mind that this value should be higher than 20 customers per store, to keep stores working with them, so we defined this as the threshold. For this KPI we had to create a DAX Measure, using the "DIVIDE" function to get the ratio between the "DISTINCTCOUNT" of the customers and the "DISTINCTCOUNT" of the stores. Note that in this case what we offer is the ratio of active users and stores for a specific time frame (quarter and/or year), which does not imply necessarily that if a customer did not purchase anything during a specific quarter, he is not in the platform anymore, but for what Olist asked us to do, the intended use is to only consider the ratio between stores selling and buyers purchasing in that specific period, which may lead to higher ratios when looking at the full year than the maximum individual value of a specific quarter of that same year.

Lastly, to combine all the different tabs, we decided to use a storytelling technique, in this case, a homepage with bookmarkers referencing the diverse tabs. This helps not only to present the information in a more structured way, but it will also aid Olist teams to better find what they are looking for, as they just have to select what they want to analyse, on the homepage, and they will be directed to that tab (once in a specific tab, we have also added the option to go back to the homepage).

All this was also deployed in our online app, with some of the visualizations we considered as conveying best the overall picture having been selected to be part of our dashboard. Note that the dashboard conveys data for the year 2017 (being the only full year we have), as we believe to be the most representative year, but by clicking on a visualization of

interest the user will be moved to the full report page where he/she can play freely with the many interactive options that we offer.

9. What Does the Data Say?

From the analysis made based on the dimensional model we created, we were able to gather some insights that we believe that not only answer Olist's questions but may also help them better understand their business.

Firstly, regarding the Shipping Costs associated with each order (and looking for the full-time period in study), there seems to exist a clear positive correlation between the volumetry of the product category and the price the customer has to pay as shipping cost, being this relation most visible when considering all weight categories, but still clear when selecting a specific one. Although this conclusion seems trivial, it was important to confirm it with data, which also helps Olist to understand better the dynamics of their product categories, for instance by understanding that there is a high concentration of categories with less than 40,000 cm^3, which corresponds to less than 30 reais of shipping costs. On this same note, when selecting a weight category, as expected, we tend to have heavier products associated with larger volumetrics, while in the opposite case, with the category "Very Light" almost all product categories have less than 20,000 cm^3 of volumetry. When it comes to the decomposition of the value of the purchase in product and shipping costs, we can see that the percentage corresponding to shipping costs has increased over the years, from 13.11%, in 2016, to 14.6%, in 2018, which may be a reason for concern to Olist. Additionally, and interestingly enough, it also seems that the shipping costs that correspond to a higher percentage of the amount paid concern the products from the weight category "Very Light", followed by the "Very Heavy", while for the "Heavy" and "Light" there does not seem to exist an as expressive difference. Finally, in this tab, we can also see that the regions "Nordeste" and "Norte" seem to have the most expensive shipping costs associated with their states.

Moving now to the tab referent to the Delivery Time, the first aspect that comes to our attention is the significant different classification given to delayed and non-delayed orders. This behaviour is consistent over the years, and across different regions, with the average classification for delayed orders never surpassing the score of 2.5, while the non-delayed orders tend to have values very close to 4 (most of the times higher than 4, but there is some natural variability). When analysing the average delivery time, independently of the years, the regions "Norte" and "Nordeste" have the highest average delivery times, and by looking at our KPI with the delay percentage of orders, in the years 2017 and 2018 (being the majority of our data from this period), these were also the regions further away from achieving the threshold of 5%. Still regarding this KPI, it is obvious that from 2017 to 2018 the delays have increased significantly (with some regions having two times higher percentage of delayed orders), being that in 2018 no region was able to respect the threshold of 5%. Additionally, in general terms, in 2017, the average percentage of delayed orders was close to 6%, while in 2018 it was closer to 8%. Regarding the other KPI, there seems that in 2017, all regions improved the average delivery time in comparison with the same period of 2016 (only data for October, November and December), while in 2018, and although, as discussed, the delayed orders have increased, the average delivery time has improved at a macro level (Brazil as a whole) and also at a region-specific level, with "Norte" being the only region not achieving the threshold defined by the

same period of last year. In this regard, it is important to consider that for 2018 we only have orders made until August, which does not allow for a comparison with the period of Christmas and Black Friday, what could be interesting to verify if the reduction in average delivery times is kept during those times of the year.

Lastly, concerning the cumulative delay time over the months of the year, in 2016, probably due to the low number of orders in that year, we only had delayed orders in October, but in 2017 we can clearly see a steady increase over the months, with a significant bump in November, that with business domain knowledge, we can conclude to be the period of purchases before Christmas (which, traditionally, starts in November with the Black Friday). Moreover, in that same month, November, is clear that, although the number of orders have increased significantly, the situation was improperly handled, with average delay times of 1.37 days per order and with a percentage of delayed orders more than twice the yearly average of 2017 (with a value of 12.37%). Finally, in 2018, orders made in January already had a cumulative delay time of more than 4,000 days, and this trend is aggravated at a very high rate until March (achieving in this month almost 2 days of delay per order, on average, and close to 20% of delayed orders) , when the growth becomes slower and keeps that way for the remaining data we have (until August). What all this seems to suggest is that in 2017, during the Christmas purchase period, Olist had a significant delay in the delivery of orders, which continued during the first few months of 2018, this is, up to March, when the situation seems to have been regularized and got back on track.

Overall, regarding the Logistic Perspective, some of the key points we discovered that could and should be considered further are: 1. Olist should pay special attention to the fact that shipping costs are rising in the percentage of the amount paid by the customer on yearly basis; 2. The regions “Nordeste” and “Norte” seem to pay the most shipping costs, on average, while also being the regions with higher average delivery time and that constantly surpass the threshold value of having at most 5% of orders delayed, which may suggest that these regions may be the most susceptible to the need of Olist’s warehouses to mediate the process; 3. It is clear that delayed orders tend to have worse reviews, which may help Olist to prove to their partners the need for a new subscription built around the warehouses’ locations, to assure no delays in deliveries (as wanted); 4. Although in 2016 the Christmas period was smooth, in 2017 (probably due to the exponential growth of Olist), this period created significant days of delay, with this situation being only fully controlled in March of 2018 (in this regard, having Christmas data for 2018 would be useful to see if this was due to some problem in 2017 or if there is a clear trend around this period); 5. While the average delivery time seems to be decreasing, the percentage of delayed orders seems to be growing away from the 5% threshold defined by Olist.

Moving now to the Customer Perspective, and starting once more by the Popularity Tab, the first point of interest is the total sales amount, which accounts for around 12.6 million Reais, corresponding to 104,571 units sold. Interestingly enough, and although in 2018 we only have orders up to August (missing what in 2017 was the busiest period of the year), in this year Olist registered the highest total sales amount and volume of items sold, by more than 1 million Reais and 10,000 units, correspondingly. On a more qualitative note, the overall average review score is 4.15 out of 5, with the “Norte” and “Nordeste” regions being the only ones with an average score lower than 4.1, which may be related to some of our findings from the Logistic Perspective.

Despite this, it seems that “Nordeste” should be more carefully managed, as in total it accounts for around 1.4 million Reais in sales amount. Still, on this note, it is interesting to see how the words “entrega”, “prazo”, “correios” and “serviço” (delivery, schedule, post-office and service) are present in our word cloud, once more suggesting that the satisfaction of the customers is not only dependent on the products they purchase but also on aspects that Olist can control more directly as the quality of the service and the dynamics associated with the delivery.

On another level, we can clearly see that the most popular category of products, when considering the full-time period, is the Health and Beauty, followed closely by the Watches and Gifts, being this trend particularly new, as in 2017, Health and Beauty was only third on the list. On the opposite side, Security and Services is the product category generating less revenue, with only two units sold (accounting for around 283 Reais). Finally, on this tab, we decided it would also be important to understand the most demanded product categories for the “Nordeste” and “Norte” regions, as when combined with the logistic perspective, these are the regions where Olist can be interested in having in-house inventory for handling quicker the orders made. In this regard, it seems that the consumers from this region follow the overall trend towards the categories of Health and Beauty and Watches and Gifts. Moreover, and setting a threshold for the categories generating more than 100,000 Reais (that Olist confirmed as being a good value for), it will probably be also interesting to include the categories Sports and Leisure and Computer Accessories.

Note that we will neither discuss the results for the most popular overall products, neither the store and customer with the highest sales amount, as that is a very simple to interpret graph (bar plot) and it will be significantly more interesting to do so when we have the dictionary linking the hashes of the names of the products/stores/customers with the true name. This does not mean we neglect the business need associated with it, as we have constructed the tools to answer that question, but simply found it unpleasant to present here hashes, when they are present on the Power BI report and can be easily converted to product/stores/customers’ names after a dictionary is provided.

On the Stores’ side of the Customer Analysis, we have a clear concentration of revenue being generated by stores from the “Sudeste” region (with around 9.8 million Reais out of 12.5 million coming from this region). When analysing this region closer, the State of São Paulo is clearly the one bringing the most to the game with 8.07 million Reais generated. Considering this region’s strength in what concerns the stores under Olist’s umbrella, it becomes clear why orders take so long to reach regions “Norte” and “Nordeste”, that are geographically far from the “Sudeste” region. As explained above, we offer many more possible explorations of data for this perspective, that Olist may use (particularly by using the tooltip on the map to better understand the regions), but that for brevity reasons, we will not develop here, as they are not necessary to face the business needs presented.

When analysing the time series with the evolution of the sales quantity over time, the first point we can make is that there seems to exist a positive trend over time. Moreover, and as we were already expecting after analysing the cumulative delay time from the logistic perspective, we seem to have a significant peak on 24th November of 2017, which corresponds to Black Friday in that year. For the period we do not have sales, but only a forecast, this is, from August of 2018 until the end of the year, our predictions seem to be what we would expect

intuitively, with an increase in sales amount in line with the ones seen in the last quarter of 2017, and also reaching a peak on November 24th. On this last remark, Olist should adapt their expectation for this peak to not be precisely at this day, but at the day the Black Friday will occur for the particular year of 2018, that is variable from year to year, but apart from that, this seems a very useful tool for Olist to discuss a bundling discount with the shipping companies, as explained before.

Finally, discussing the Customers Tab's results, we can also see a strong influence of the "Sudeste" region, with purchases from this region's customers accounting for 8.23 million Reais (around 65.5%) of revenue, with the state of "São Paulo" also accounting for a significant share of this amount, being responsible for 4.82 million Reais. On another level, credit card is by far the most frequent payment method used, with the payments being mostly paid in one instalment (which does not necessarily mean the person did not use instalments, we just know that was not with our instalment option) but being still fairly common to pay with up to 10 instalments. We believe this information can be seen as the first step for Olist to include their own option for credit payment method (and not just the option to fragment the purchase into multiple instalments), as is increasingly becoming popular for platforms and stores to have, with special conditions for products sold by them. Obviously, this will need to be explored further, but we hope to have opened that door for further exploration.

Regarding the KPI created, although for the full year of 2017 and 2018 the threshold has been achieved, when we attempt to narrow it down by looking at quarter level or region level data, we can see there is still a lot of work to be done, with what look like periods of inactivity of users, that tend to purchase in specific periods of the year and not as frequently over the year as Olist might desire: even in the last quarter of 2017, when there was a significant increase in purchases, the ratio of active users on that quarter in relation to stores was close to 10 to 1. Being this said, it is important to relativize this question, and as we have already discussed, Olist and the overall sector are growing rapidly, which may imply that in a near future this KPI may meet its threshold, as Olist probably asked us to create it to be seen as a goal to achieve at a more mature phase of the company.

In a nutshell, the main points we were able to extract from this perspective were: 1. A significant percentage of value comes from customers and stores located in the "Sudeste" region, particularly in "São Paulo"; 2. There seems to exist a significant increase in revenue and volume over the years, with the categories of products that bring the most revenue being Health and Beauty and Watches and Gifts; 3. Although the region "Norte" does not seem to bring high amounts of revenue, the "Nordeste" one should be more carefully analysed by Olist, to shorten the delivery periods, which can be made, as suggested by Olist, by having in-house inventory of certain products, and we have already narrowed it down to the main categories of interest. Nevertheless, we also offer the possibility in the Power BI report to analyse the most purchased products for those regions, so that Olist can better plan for the products to have in inventory; 4. The products, stores and customers generating the highest amount of revenue (sales amount) can be consulted at the Power BI report, and we look forward to translating the hashes to names, so we can extract business insight from it; 5. The reported trend of exponential growth seems accurate with an increase from 2017 to 2018, even without having 2018's data for the busiest time of the year; 6. Olist should use the forecasting tool provided to negotiate better conditions

with the shipping companies, while starting to prepare for what seems like a repeating trend of a large number of orders at Black Friday; 7. The KPI regarding the ratio of active customers per store seems to be achieved at a yearly level, but not when starting to drill it down by region and quarter, which should not be a significant reason for concern, as Olist is quickly expanding, but should be kept under watch; 8. Providing a self-credit payment option may be a good option as the majority of customers pay with a credit card and up to 10 instalments (the ones paying with one instalment should be studied further to understand if they just don't use the Olist's instalment plan but may be using one specific for their credit-card).

10. Critical Assessment

Regarding the critical assessment of the project, we would like to start by pointing out that we believe this project could have been better structured if there were not as many requisites regarding the data and the number of dimensions needed to be created, as that would allow for more diverse and interesting choices of datasets. Moreover, the conditioning of having an intermediate delivery created some problems, as at that point we did not have the opportunity to work with the data in Power BI, and explore it further, which led to setting some business needs that we ended up realizing were not that important, while others seemed more important. This problem arises, mainly, as we had no company, in this case, Olist, telling us what their business needs were, so the needs that we come with ended up not being as relevant when looking at data and understanding the company at a deeper level (which would be the case if the business needs were presented by the company we are analysing, as it should be the case, in the real world). In this regard, we decided to present some modifications to the first delivery, that do not change in any way the requirements achieved with the first delivery, but that we believe to be better contextualized with the actual needs of the company, while building on the suggestion given of constructing the business needs in question format to be later addressed and that follow the logic used in practical classes after the intermediate delivery deadline. Additionally, and based on the feedback received, we decided to make some slight modification to our fact table, that again do not change any of the requirements, only removing characteristics from the product dimension that was not part of a hierarchy (number of photos and length of the description) and adding Order Number to the Fact Table. Overall, the changes made to the first report go mainly in light with the suggestions made in the correction of the same (are limited to chapter 3, where we added business needs and clarified the intended goal of others, and to the small modification discussed of the Fact Table in chapter 6), and do not in any way change or violate any previous requirements, only improving on what was done. Note that we could have changed other aspects, but given that would not improve the grade, we limited the changes to where we believe make the report's story more appealing for the second delivery.

From a more technical standpoint, we felt that this project could have been quicker and more efficiently developed with the help of other programming languages (mainly in a first data processing phase), as Power BI gets significantly slow when working with fairly large datasets, but we acknowledged that helped us develop our skills with this tool. One of the situations we faced from the technical perspective (and that have no impact on the outcome) was the date hierarchy generated automatically by Power BI, which although later on the semester we have found it would be possible to remove it when we developed those steps of the project, we did

not know how to do it or that it was even possible. Considering this situation only made it harder to use DAX Time Intelligence Measure (which we were able to deal with), with no further implication in any other aspect, and since doing the necessary steps to solve it, at the point in the project we learned about it in the support video, would imply having to redo almost all the work and visualizations associated with Dates (including the DAX Time Intelligence Measure), we decided it would not be worth it to remove this hierarchy, as, once more, it did not have any impact in the quality of the results.

Discussing now the presentation of both this report and our Power BI reports, and starting by the first, we acknowledge that this report may be too dense in the textual component, with lack of visualizations, but according to the guidelines, it seemed appropriate to do it in this way, since we had to discuss different aspects of what was constructed on Power BI and presenting here the same visualizations as in the Power BI reports would just create repetition. Moreover, on what concerns the Power BI reports, the choice of constantly using blue was an informed one, and although risking creating some monotony, it would be distasteful for our client to have tabs where the main colour would not be the one Olist uses.

Lastly, and considering the ability to face all the problems presented by Olist, we believe to have been able to answer all the business needs, while also providing the tools for much deeper exploration, with visualizations that we only touch slightly in this report (as they are more focused on Olist's pursuit of knowledge), but that offer enormous potential. In this regard, we deviated a little from the purely academic nature of this exercise and decided to treat Olist as a real customer providing them with as much information as possible. Finally, we complied with all the mandatory requirements and with at least 4 self-learned topics (e.g., visualizations and M and DAX functions), being that some others were used, as for instance the "group by" command in M, as when we applied it to our project, we had to learn it by ourselves, but it was later on presented in class, which we also believe to be a disadvantage of giving extra points based on using new things: until the end of the semester, we are arbitrarily subjected to have some of the points being taken from us, not because they were not self-learned at the time of the doing, but because they were then used in class.

11. Conclusions

Overall, our team believes to have constructed an app that will help Olist improve their decision-making process, being this not limited to the needs initially presented, but also, to many more analysis that can now be conducted as needed.

In collaboration with Olist, we were also able to detect some problems in their original data collection process (that Olist should fix) while building the model based on Star Schema proposed. Due to this type of problems, the model required more effort than anticipated to construct, but we are proud of the results achieved and the way we were able to provide a proper structure that can be leveraged in any Business Intelligence effort.

After the modelling effort, and by applying diverse techniques (e.g., forecasting and creation of "regular" and time intelligence measures), we developed a functional report, divided into areas of interest to Olist, that can quickly be used to understand any particular topic about their business, under the scope of our analysis. Interestingly enough, and although we have found more insights than we were initially looking for, we have found that some of the initial

main concerns of Olist were perfectly aligned with what the data tells, for instance, the concern with how delays affect satisfaction and how there seems to exist a clear space in Brazil where orders tend to get incredibly delayed.

Lastly, to make the results easier to be accessed, and as mentioned, we created an app, containing both the report and dashboard that we hope Olist can leverage, while keeping in mind that the information introduced should be studied further (with data not available to us, like financial implications), in order be converted into an efficient strategic and operational plan.

12. Datasets

Redes Suas: http://blog.mds.gov.br/redesuas/lista-de-municipios-brasileiros/

Kaggle: https://www.kaggle.com/olistbr/brazilian-ecommerce?select=olist_geolocation_dataset.csv

13. References and Other Sources of Information

Kenton, W. (2020, December 29). Logistics. Retrieved from: https://www.investopedia.com/terms/l/logistics.asp

Sant'Ana, J. (2017, November 14). *Olist quer ser a maior loja virtual dentro dos principais marketplaces do país*. Retrieved from: https://www.gazetadopovo.com.br/economia/nova-economia/olist-quer-ser-a-maior-loja-virtual-dentro-dos-principais-marketplaces-do-pais-3xbe9k4mn13uzs1pm60ym6znz/

Meneghel, J. (2021, March 2). *O que é Olist: como funciona e ajuda nas vendas do seu negócio*. Retrieved from: https://blog.olist.com/o-que-e-olist

Oliveira, C. de (2019, October 2). *Números do e-commerce brasileiro: 34 dados que você, galera precisa conhecer!*. Retrieved from: https://blog.olist.com/1-numeros-e-commerce-brasileiro-e-dados-vendas-online/

14. Appendix – DAX Measures and Calculated Columns' Formulas

Measures in Tab Stores – Customer Perspective

- Sales Amount = SUMX (Fact_Table, Fact_Table[Unit_Price] * Fact_Table[Quantity])

Measures in Tab Customers – Customer Perspective

- Sales Amount = SUMX (Fact_Table, Fact_Table[Unit_Price] * Fact_Table[Quantity])
- Ratio Customers Sellers = DIVIDE (DISTINCTCOUNT (Fact_Table[FK_Customer]), DISTINCTCOUNT (Fact_Table[FK_Store]))

Measures in Tab Popularity – Customer Perspective

- Sales Amount = SUMX (Fact_Table, Fact_Table[Unit_Price] * Fact_Table[Quantity])
- Avg Review Score = AVERAGEX (DISTINCT (SELECTCOLUMNS (Fact_Table, "Order_Number", [Order_Number], "Review_Score", [Review_Score])), [Review_Score])

Measures in Tab Shipping Costs – Logistic Perspective

- Avg Shipping Costs = DIVIDE (SUMX (Fact_Table, Fact_Table[Unit_Shipping_Cost] * Fact_Table[Quantity]), SUMX (Fact_Table, Fact_Table[Quantity]))
- Avg Unit Price = DIVIDE (SUMX (Fact_Table, Fact_Table[Unit_Price] * Fact_Table[Quantity]), SUMX (Fact_Table, Fact_Table[Quantity]))
- Avg Volumetry = DIVIDE (SUMX (Fact_Table, Fact_Table[Quantity] * RELATED (Dim_Product[Volumetry])), SUMX (Fact_Table, Fact_Table[Quantity]))

Measures in Tab Delivery Time – Logistic Perspective

- % Delayed Orders = DIVIDE (COUNTROWS (FILTER (DISTINCT (SELECTCOLUMNS (Fact_Table, "Order_Number", [Order_Number], "Delay_Time", [Delay_Time])), [Delay_Time] <> 0)), COUNTROWS (DISTINCT (SELECTCOLUMNS (Fact_Table, "Order_Number", [Order_Number], "Delay_Time", [Delay_Time]))))
- Avg Delivery Time = AVERAGEX (DISTINCT (SELECTCOLUMNS (Fact_Table, "Order_Number", [Order_Number], "Delivery_Time", [Delivery_Time])), [Delivery_Time])
- Delayed Review Scores = CALCULATE (AVERAGEX (Fact_Table, [Review_Score]), FILTER (DISTINCT (SELECTCOLUMNS (Fact_Table, "Order_Number", [Order_Number], "Delay_Time", [Delay_Time])), [Delay_Time] > 0))
- Non-Delayed Review Scores = CALCULATE (AVERAGEX (Fact_Table, [Review_Score]), FILTER (DISTINCT (SELECTCOLUMNS (Fact_Table, "Order_Number", [Order_Number], "Delay_Time", [Delay_Time])), [Delay_Time] == 0))
- Avg Delivery Time = AVERAGEX (DISTINCT (SELECTCOLUMNS (Fact_Table, "Order_Number", [Order_Number], "Delivery_Time", [Delivery_Time])), [Delivery_Time])
- Sum of Delay Time YTD = TOTALYTD (SUMX (DISTINCT (SELECTCOLUMNS (Fact_Table, "Order_Number", [Order_Number], "Delay_Time", [Delay_Time])), [Delay_Time]), Dim_Date[Date], ALL('Dim_Date'))

Calculated Columns:

- Delay_Time = IF (DATEDIFF (Fact_Table [order_estimated_delivery_date], Fact_Table [order_delivered_customer_date], DAY) < 0, 0, DATEDIFF (Fact_Table [order_estimated_delivery_date], Fact_Table[order_delivered_customer_date], DAY))
- Delivery_Time = DATEDIFF (RELATED (Dim_Date [Date]), Fact_Table [order_delivered_customer_date], DAY)
- Volumetry = 'Dim_Product' [Product_Length_cm] * 'Dim_Product' [Product_Width_cm] * 'Dim_Product' [Product_Height_cm]
- Weight_Category = IF ('Dim_Product'[Product_Weight_g] < 500, "Very Light", IF ('Dim_Product'[Product_Weight_g] < 1000, "Light", IF ('Dim_Product'[Product_Weight_g] < 5000, "Heavy", "Very Heavy")))