VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

---ꕥ📖ꕥ---

FINAL PROJECT

DATA ANALYSIS IN BUSINESS

TOPIC:

APPLYING BI/DW SOLUTION IN ANALYZING THE BUSINESS SITUATION AND PROPOSING APPROPRIATE STRATEGIES FOR AN E-COMMERCE PLATFORM IN BRAZIL

Instructor : **M.Sc. Lê Bá Thiền**

Course code : **232MI1701**

Group 5:

Hoàng Ngọc Thảo Duyên	K204110559
Phan Trịnh Kim Hạnh	K204110565
Phạm Thị Kim Ngân	K204110575
Nguyễn Ngọc Thẩm	K204111787
Phạm Ngọc Trâm	K204111790

Ho Chi Minh City, January 2024

VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

---🕮---

FINAL PROJECT

DATA ANALYSIS IN BUSINESS

TOPIC:

APPLYING BI/DW SOLUTION IN ANALYZING THE BUSINESS SITUATION AND PROPOSING APPROPRIATE STRATEGIES FOR AN E-COMMERCE PLATFORM IN BRAZIL

Instructor : **M.Sc. Lê Bá Thiền**

Course code : **232MI1701**

Group 5:

Hoàng Ngọc Thảo Duyên	K204110559
Phan Trịnh Kim Hạnh	K204110565
Phạm Thị Kim Ngân	K204110575
Nguyễn Ngọc Thẩm	K204111787
Phạm Ngọc Trâm	K204111790

Ho Chi Minh City, January 2024

PROTESTATION

Our group would like to confirm that this report was prepared by the group. We are fully responsible for the truthfulness of the content in the topic.

Ho Chi Minh City, January 5, 2023

Leader

(Sign)

Nguyen Ngoc Tham

ACKNOWLEDGEMENT

First and foremost, we would like to express our sincere gratitude to the University of Economics and Law and the Faculty of Information System for providing us with the course “Data analysis in business”. This course allowed us to acquire practical knowledge about how to apply data analysis in business and help us have better preparation before going to the job market.

We would also like to express our gratitude to M.Sc. Le Ba Thien because of his professional knowledge and enthusiastic guidance to support us to complete this project efficiently and obtain a lot of practical and academic knowledge throughout this course.

Due to the limitation of time, knowledge and experience, our report is not perfect. We are always welcome to provide any feedback, recommendations and suggestions to improve our report.

We wish that M.Sc. Le Ba Thien and all teachers in the University of Economics and Law would have great success in the future.

TABLE OF CONTENT

LIST OF ABBREVIATIONS

No.	Abbreviation	Definition
1	API	Application Programming Interface
2	BI	Business Intelligence
3	DW	Data Warehouse
4	EDA	Exploratory Data Analysis
5	ETL	Extract, Transform, Load
6	GMV	Gross Merchandise Volume
7	NMV	Net Merchandise Value
8	OLAP	On Line Analysis Processing
9	RFM	Recency – Frequency - Monetary
10	SQL	Structured Query Language
11	SSIS	SQL Server Integration Services

LIST OF TABLES

LIST OF FIGURES

TÓM TẮT

Giải pháp BI/DW ngày càng được ứng dụng phổ biến trong thương mại điện tử. Với quy mô và tiềm năng phát triển của thương mại điện tử, lượng data được sinh ra mỗi là là vô cùng lớn. Điều đó đặt ra thách thức cần có giải pháp để quản lý và tối ưu hóa việc khai thác lượng dữ liệu trên. Giải pháp BI/DW là giải pháp hiệu quả để quản lý và phân tích dữ liệu. Trong bài báo cáo này, nhóm đã sử dụng BI/DW solution để phân tích dữ liệu từ một sàn thương mại điện tử bằng cách sử dụng các công cụ là Python, SSIS, SQL Server và Power BI để tìm ra insight trong haotj động kinh doanh của doanh nghiệp từ đó đề xuất những chiến lược phù hợp để cải thiện hiệu của hoạt động.

ABSTRACT

BI/DW solutions are increasingly widely used in e-commerce. With the scale and development potential of e-commerce, the amount of data generated each time is extremely large. That poses a challenge that requires solutions to manage and optimize the exploitation of the above amount of data. BI/DW solutions are effective solutions for data management and analysis. In this report, the team used BI/DW solution to analyze data from an e-commerce platform using tools such as Python, SSIS, SQL Server and Power BI to find insights in action. business of the enterprise, thereby proposing appropriate strategies to improve operational efficiency.

CHAPTER 1: INTRODUCTION

1.1. The reason for choosing the topic

Since the covid 19 pandemic, the world's e-commerce market has developed rapidly [1]. Specifically, it is estimated that the online retail market for physical products was worth $641 billion in 2020, an increase of 14.3% over the same period the previous year and accounted for 23.5% of the value of the entire market economics [2]. This number is expected to increase to approximately 58% by 2028 [3]. The development of e-commerce mentioned above also facilitates the development of tools for research and development in this field, with more than 60% of e-commerce businesses in the world using it. Using R&D (Research and Development) tools [2], including BI (Business Intelligence) solutions. This is a solution that includes BI solutions, applications, infrastructure, tools used to improve and optimize decision making and business performance through accessing and analyzing data with superior efficiency. [4]. Therefore, the authors chose the topic ***"Applying BI/DW solution in analyzing the business situation and proposing appropriate strategies for an e-commerce platform in Brazil"*** to apply the solution BI/DW solution to analyze the business situation of an e-commerce platform in Brazil from which to propose appropriate solutions to maximize business performance.

1.2. Topic goal

- Carry out the process of searching and processing data in the e-commerce field to identify suitable data sources during the solution development.
- Build a comprehensive, suitable, and convenient Data Warehouse during the analysis process.
- Visualize data about the business situation of e-commerce businesses.
- Analyze to find important insights to propose optimal solutions in the field of E-commerce.

1.3. Subject, method, and research scope of the project.

1.3.1. Research subjects

Business operations of an e-commerce platform in Brazil.

1.3.2. Research method

The research method used is qualitative research method. The research will apply BI/DW to analyze the collected data set to figure out the business performance and propose strategies to improve the business performance of the enterprise.

1.3.3. Research scope

- ***Spatial scope:*** samples in the data set used in the study were collected from an e-commerce platform in Brazil, which is a secondary data source taken from Kaggle.
- ***Time scope:*** It contains all transactions that occurred from 2016 to 2018 of an e-commerce platform.

1.4. Tools used

- ***Kaggle:*** a platform that provides diverse datasets from many different fields. In the research article, the authors downloaded the data set about E-commerce in Brazil.
- ***Python:*** a powerful and flexible programming language, especially in data processing. The authors used Python to preprocess data before building the Data warehouse. The platform for this analysis will be Google Colab or Visual Studio.
- ***SQL Server:*** a powerful database management system, building and managing Data Warehouse to store and organize data. SQL Server will provide data management and query optimization tools, making it a good choice for building a Data Warehouse.

- ***Power BI:*** a powerful tool for visualizing data and building informative dashboards. Use Power BI to create realistic charts, graphs, and dashboards from data in the Data Warehouse, helping to find important insights.

1.5. Research process

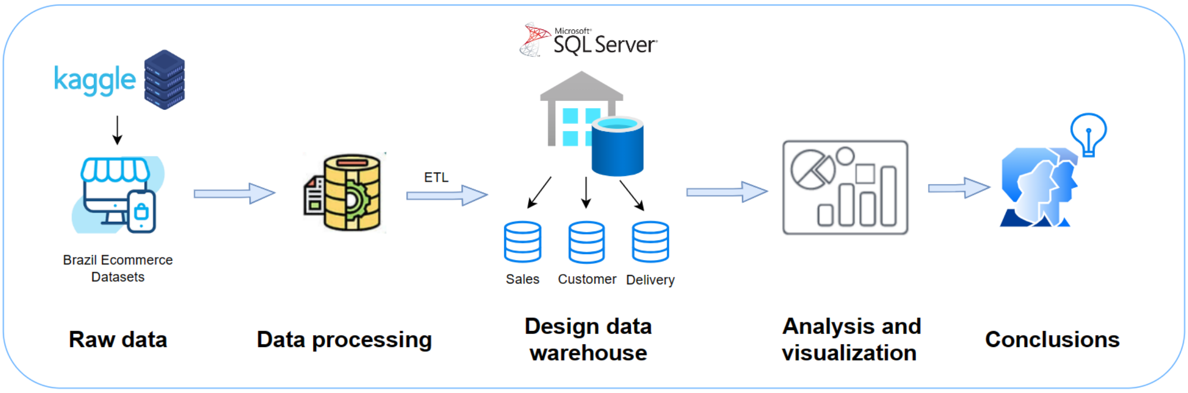

Figure 1.1: *The research process diagram (Source: The authors synthesize and propose)*

The data source was taken from Kaggle, after EDA, to remove duplicate values, delete unnecessary columns, convert data to the correct data type, and use SSIS to be able to observe and convert data types. appropriate data and the data will be loaded into SQL server to build a data warehouse. Power BI will be the platform that connects data from the data warehouse to extract suitable data for analysis, modeling and finding insights.

1.6. Structure of the report

CHAPTER 1: SUBJECT SUMMARIZATION

CHAPTER 2: THEORETICAL BACKGROUND AND RELATED WORKS

CHAPTER 3: ANALYSIS REQUIREMENTS AND EXPERIMENTAL MODEL

CHAPTER 4: EXPERIMENTAL RESULT

CHAPTER 5: CONCLUSION

CHAPTER 2: THEORETICAL BACKGROUND AND RELATED WORKS

2.1. Theoretical background

2.1.1. BI solution

There are many definitions for Business Intelligence (BI), in this paper, the authors define "*BI is a strategic initiative by which organizations measure and drive the effectiveness of their competitive strategy*", or the BI platform is a software platform that provides 14 capabilities in three main functional categories including integration, information delivery, and analysis [5]. In addition, BI solution is a set of tools, technologies and solutions designed for end users to effectively extract useful business information from big data [6].

2.1.2. Data warehouse

A Data Warehouse is essential to any Business Intelligence (BI) solution. It can be described as a subject oriented, integrated, time variant and nonvolatile data storage system that supports decision-making. It is a comprehensive database containing the necessary information for performance assessment, decision-making, and predictive analysis. Multidimensional modeling methods utilize facts and dimensions within relational or multidimensional databases to create corporate data warehouses and departmental data marts [5].

2.1.3. Schema in Data warehouse

Dimensional Modeling is a retrieval-based system that supports high-volume query access. It has 2 key components:

- Fact Table: consist of the measures of the data cube and foreign keys to the dimension tables surrounding it
- Dimension Tables: Contain descriptive information related to the business entities captured in the fact table. Commonly used dimensions are people, products, place and time

Star schema: The most commonly used and the simplest style of dimensional modeling, containing a fact table surrounded by and connected to several dimension tables

Snowflakes schema: An extension of star schema where the diagram resembles a snowflake in shape.

2.1.4 Data approach

To construct a data warehouse, there are two main architectures: Inmon architecture and Kimball architecture. Kimball's architecture employs a bottom-up approach based on dimensional modeling. The process begins by examining business processes, determining their needs, and identifying questions that require answers. Subsequently, all data sources are pinpointed, and an Extract, Transform, Load (ETL) process is executed to create a denormalized data model. This model, built using either a star or snowflake schema, is then organized around department-specific sub-databases [7].

In this research, the authors use Inmon's architecture.The Inmon approach, known for its top-down design, involves initially constructing a relational data model by gathering information from various sources. The data then undergoes Extract, Transform, Load (ETL) processes. Subsequently, the Data Warehouse (DW) generates dimensional data marts, reports, and applications tailored to specific business processes or departmental requirements. In this methodology, data marts serve as an intermediary step between ETL and the ultimate data output, facilitating a structured and organized data flow.

2.1.5. ETL

The ETL process, short for Extract – Transform - Load, acts as a crucial conduit for data architects, enabling them to seamlessly merge disparate and disorderly data sources into a cohesive and refined repository of knowledge. Commencing with the extraction of valuable data from a variety of sources such as databases, flat files, and APIs, it meticulously purifies, standardizes, and enhances the data to ensure its uniformity and practicality. Finally, it conveys the refined data to its ultimate destination, typically a data warehouse or lake, where it evolves into a centralized truth primed for exploration and analysis. ETL plays a pivotal role in promoting informed decision-making by guaranteeing high-quality, easily accessible data, streamlining analysis, and bolstering confidence in insights [8].

2.1.6. RFM model

RFM - acronym for Recency - Frequency - Monetary, is a method used to analyze data in Marketing. This model is widely recognized and used in segmenting and ranking customers based on their purchasing history. Businesses can personalize marketing content, such as sending messages or advertising emails, to relevant customer groups, thereby increasing response rates and conversion rates. The RFM model has gained attention in ecommerce and the retail industry. This method is based on three key factors: Recency (R), Frequency (F), Monetary (M), in which:

- ***Recency:*** The last time a customer made a transaction - A smaller value implies that the customer made a recent purchase, while a larger value implies that the customer hasn't made a purchase for a longer time.
- ***Frequency****:* How many times a customer has purchased or shopping frequency - This value is defined as the number of purchases a customer makes in a specific period of time. The higher the value of frequency, the more we can evaluate the customer's loyalty and ease of returning to purchase.
- ***Monetary***: How much money the customer has paid - Monetary value is determined by the amount of money the customer has spent in a certain period of time. The more money customers spend, the more revenue they bring to the business [9].

RFM segmentation is a popular analytical method in database marketing because of its simplicity, effective logical classification, and robustness to customer segmentation. However, in fact that the RFM model only considers three specific factors (albeit important ones) means that this approach may exclude other variables that are equally or more important (e.g.: purchased products, pre-campaign responses, demographic details, etc.). Additionally, RFM is a historical method: it examines customer behavior in the past, so it may not accurately predict customer activities, preferences, and feedback in the future if inaccurate data processing occurs. Advanced customer segmentation techniques require more complexity and rely on more combinations of other predictive analytics technologies that tend to predict customer behavior [10].

2.2. Related Works

2.2.1. Research related to Data Warehouse and BI Solution

A Data Warehouse Approach for Business Intelligence (Garani, Chernov, Savvas & Butakova, 2019) [11]

This paper proposes a data warehouse (DW) approach for effectively integrating and analyzing spatial and temporal data in business intelligence (BI) applications. These are two factors that greatly influence decision-making and marketing strategy. However, this data cannot be processed effectively in conventional multidimensional databases. Therefore, a new DW diagram modeling method is needed to integrate spatial and temporal data.

The new DW schema modeling method (Starnest Schema-a combination of some features of star and snowflake truss schemas) published in this paper includes the following components:

- A new spatiotemporal DW schema is designed to integrate spatial and temporal data in a unified way.
- OLAP queries have been extended to support spatial and temporal queries.
- A case study was developed and deployed for the telecommunications industry.

This new DW schema modeling approach enables business analysts and decision-makers to access and analyze spatiotemporal data more effectively. This can help them make better business decisions.

Specifically, in the telecommunications industry, this method can be used to analyze customer location data, service usage data, and real-time data. This can help telecom service providers improve their services, such as: optimizing networks, minimizing downtime, and providing services tailored to customer needs.

Overall, this article proposes a valuable solution for integrating and analyzing spatial and temporal data in BI applications, allowing businesses to gain deeper insights and make data-driven decisions.

2.2.2. Research related to the application of BI Solution in e-commerce

Integration of Business Intelligence with e-commerce (Ferreira, Pedrosa & Bernardino, 2019) [12]

The integration of Business Intelligence (BI) with E-commerce is the most appropriate and effective process to help businesses understand their customers better in order to balance the customers' needs and the companies' benefits. The architectures of BI and e-commerce used in the paper consist of four levels: Data (the webserver makes logs from the e-commerce portal, these data sources go through a process of ETL that standardizes, cleans and loads the data into the DW); Data Warehouse Server (integrating DW with Data Mart); OLAP Server and BI analysis. The group of authors used the BI platform named Pentaho and SpagoBI, and the e-commerce platform named Magento and OpenCart. The two possible integration models are proposed: an integration model using Magento APIs and a platform integration model using RabbitMQ. In the future, these authors hope to apply the proposed architecture and integration models to a real project, and analyze a real e-commerce company.

Business Intelligence for the Evaluation of Customer Satisfaction in E-Commerce Websites-A Case Study (Priyadarshini & Veeramanju, 2022) [13]

With the support of BI tools, businesses can use their data in a better way, making the right decisions with full information. BI is a combination of data mining, data analysis, data visualization, and machine learning to help organizations analyze data.

This article provides analysis of BI techniques and classification algorithms (Logistic Regression, Naive Bayes, Random Forest) used to analyze large e-commerce websites such as Amazon, Flipkart,... The result demonstrates the combination of e-commerce and BI as a very powerful combination that assists website owners in a variety of tasks, including identifying consumer-driven marketing campaigns, finding market trends, understanding purchasing habits and predicting customer behavior. The use and ability to store a lot of data can be achieved through cloud storage. However, for many companies, the cost of setting up a large data warehouse to support a BI system is still very high, and filtering data from big data is difficult.

2.2.3. Research related to the application of Data Warehouse in analyzing the business situation of Olist Store.

Implementation of extract transform load on data warehouse and business intelligence using pentaho and tableau to analyse sales performance of offlist store (Anggrainy, T. D., & Sari, A. R. , 2022) [14]

The increasing global business development driven by the application of information technology significantly impacts business operations at all levels, from local to national and global, simultaneously generating a vast amount of data. Data processing at Olist Store is executed through the implementation of ETL processes in the data warehouse, utilizing Pentaho, and business information is visually represented on a smart dashboard using Tableau. The data warehouse design follows a unified approach with nine specific steps.

The results of deploying the ETL process and visualizing business information at Olist Store demonstrate the successful creation of the data warehouse, employing PostgreSQL and PgAdmin 4. Through the Tableau smart dashboard, the analysis reveals substantial growth in order volume from 2016 to 2017. However, it indicates the need to enhance sales quality to sustain stability and prevent a decline from 2017 to 2018. The authors anticipate that future development of a data warehouse at Olist Store will incorporate the use of a cron job tool to automatically execute Python code during the initial data cleaning process, given the large volume of daily transactional sales data generated.

CHAPTER 3. REQUIREMENTS ANALYSIS AND EXPERIMENTAL MODELING

3.1. Business Issue Understanding

Olist is a e-commerce company in Brazil, and Olist Store is the largest online marketplace in this country. Olist connects small businesses all over Brazil, these owners will sell their products through Olist Store and then directly deliver to customers by Olist's logistic partners. When customers buy products from Olist Store, the sellers will be sent a notification to fulfill that order. After receiving the order's products or arriving at the estimated delivery day, customers will receive a satisfaction survey via email so that they can take notes about their purchase experiences and write some feedback.

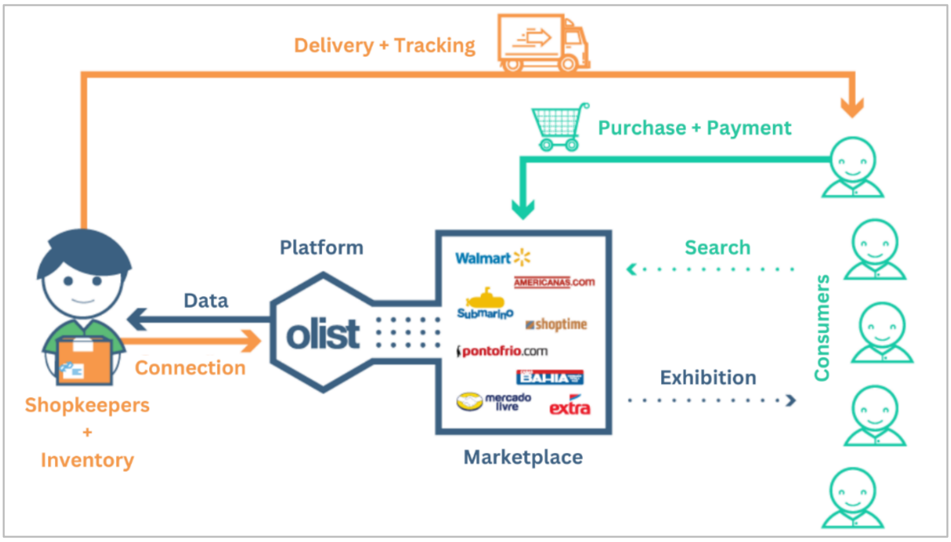

Figure 3.1: *Olist's business model (Source: The Olist company)*

Table 3.1: *Analytics Applications in the Olist Store – Sales performance, Customers, Products, and Logistic*

Analytic Application	Business question	Business Value
Sales Performance Analysis	1. Which locations have the most orders? 2. Which time period has	1. Creating a marketing strategy by time and region. 2. Focusing on marketing to

	the most orders? 3. Does delivery time affect an order's rating? 4. Characteristics of canceled orders.	valuable customer groups. 3. Limiting the number of canceled orders.
Customer Analysis	1.In which areas do customers live in? 2. When do customers shop during the day? 3.Which customer groups need attention?	
Product Analysis	Characteristics of best-selling product groups?	Searching for more sellers to expand the variety of products in the same best-sellers categories.
Logistic Analysis	1. Which area are late orders from? 2. Average time to ship an order? 3. Which month of the year has the most orders delivered late?	1. Reduce delivery time in areas with many late orders. 2. Increase shipping staff at times when orders increase.

3.2. Data Understanding

3.2.1 Data analysis goals

- Analyze Brazilian consumer shopping behavior to improve customer shopping experience and increase sales.

- Evaluate the effectiveness of current business strategies.
- Propose solutions to improve business operations for e-commerce platforms in Brazil.

3.2.2 Data collection

The dataset contains information on 100.000 orders from 2016 to 2018 made in multiple markets in Brazil.

The data includes information:

- Order information (order date, order value, order status, etc.).
- Product information (product name, product price, product description, etc.).
- Customer information (customer name, customer address, etc.).

The data set consists of 9 tables, but here the team only uses 8 tables, removing the olist_sellers_dataset table because it is found unnecessarily for the subject.

3.2.3 Data description

a. olist_customers_dataset: The table contains information about customers

Table 3.2: *The olist_customers_dataset table*

Column	Content	Data type
customer_id	Customer Identification	object
customer_unique_id	Unique identifier of a customer	object
customer_zip_code_prefix	Customer zip code (first five digits)	int64
customer_city	Customer city name	object
customer_state	Customer state	object

The customer_unique_id column is unique to a customer. The customer_id column is unique to a customer's transaction. It means that a customer has one customer_unque_id but can have more than one customer_id.

b. olist_geolocation_dataset: contains information about the location of customers of Olist.

Table 3.3: *The olist_geolocation_dataset table*

Column	Content	Data type
geolocation_zip_code_prefix	Zip code (first 5 digits)	int64
geolocation_city	City name	object
geolocation_state	State	object

c. olist_order_items_dataset: contains information about items sold in Olist orders

Table 3.4: *The olist_order_items_dataset table*

Column	Content	Data type
order_id	Order unique identifier	object
order_item_id	Sequential number identifying number of items included in the same order	int64
product_id	Product unique identifier	object
shipping_limit_date	Delivery limit date for shipping unit	object
price	Item price	float64
freight_value	Item shipping value	float64

d. olist_order_payments_dataset: contains information about payments made for orders

Table 3.5: *The olist_order_payment_dataset table*

Column	Content	Data type
order_id	Order unique identifier	object
payment_sequential	Sequence of payments in 1 order	int64
payment_type	Payment method	object
payment_installments	Number of installments chosen by the customer	int64
payment_value	Transaction value	float64

e. olist_order_reviews_dataset: contains information about reviews written by customers for Olist orders

Table 3.6:*The olist_order_reviews_dataset table*

Column	Content	Data type
review_id	Unique review identifier	object
order_id	Unique order identifier	object
review_score	Note ranging from 1 to 5 given by the customer on a satisfaction survey	int64
review_comment_title	Comment title from the review left by the customer, in Portuguese	object
review_comment_message	Comment message from the review	object

	left by the customer, in Portuguese	
review_creation_date	The date in which the satisfaction survey was sent to the customer	object
review_answer_timestamp	The satisfaction survey answer timestamp	object

f. olist_orders_dataset: contains information about orders made on the Olist platform

Table 3.7: *The olist_orders_dataset table*

Column	Content	Data type
order_id	Unique order identifier	object
customer_id	Customer Identification	object
order_status	Order status (delivered, shipped, etc)	object
order_purchase_timestamp	Shows the purchase timestamp	object
order_approved_at	The payment approval timestamp	object
order_delivered_carrier_date	Delivery time for shipping unit	object
order_delivered_customer_date	The actual order delivery date to the customer	object
order_estimated_delivery_date	Estimated delivery time to customer	object

g. olist_products_dataset: contains information about products on the Olist platform

Table 3.8: *The olist_products_dataset table*

Column	Content	Data type
product_id	Unique product identifier	object
product_category_name	Root category of product, in Portuguese	object
product_name_lenght	Length of product name	float64
product_description_lenght	Product description length	float64
product_photos_qty	Number of product published photos	float64
product_weight_g	Product weight measured in grams	float64
product_length_cm	Product length measured in centimeters	float64
product_height_cm	Product height measured in centimeters	float64
product_width_cm	Product width measured in centimeters	float64

h. product_category_name_translation: contains information about the translation of the product category name

Table 3.9: *The product_category_name_translation table*

Column	Content	Data type
product_category_name	Category name in Portuguese	object
product_category_name_english	Category name in English	object

3.3. Data Preparation

3.3.1. olist_customers_dataset

First, the authors proceed to inspect the information in the dataset.

```
1. Basic infomation of the customers dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Data columns (total 6 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   customer_id               99441 non-null  object
 1   customer_unique_id        99441 non-null  object
 2   customer_zip_code_prefix  99441 non-null  int64
 3   customer_city             99441 non-null  object
 4   customer_state            99441 non-null  object
 5   new_customer_id           99441 non-null  object
dtypes: int64(1), object(5)
memory usage: 4.6+ MB
None
-----------------------------------------------------------------
2. Number of columns and rows in the customer dataset:  (99441, 6)
-----------------------------------------------------------------
3. Numbers of unique value of customer_id:  99441
-----------------------------------------------------------------
4. Numbers of unique value of customer_unique_id:  96096
```

Figure 3.2: *The information of olist_customers_dataset (Source: Experimental results)*

Because the customer_id data is lengthy and challenging to reference during analysis, the authors assigned names to all customers.

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state	new_customer_id
0	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP	CS01
1	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	9790	sao bernardo do campo	SP	CS02
2	4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e	1151	sao paulo	SP	CS03
3	b2b6027bc5c5109e529d4dc6358b12c3	259dac757896d24d7702b9acbbff3f3c	8775	mogi das cruzes	SP	CS04
4	4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d066	13056	campinas	SP	CS05
...	...	...	...	...	...	...
99436	17ddf5dd5d51696bb3d7c6291687be6f	1a29b476fee25c95fbafc67c5ac95cf8	3937	sao paulo	SP	CS99437
99437	e7b71a9017aa05c9a7fd292d714858e8	d52a67c98be1cf6a5c84435bd38d095d	6764	taboao da serra	SP	CS99438
99438	5e28dfe12db7fb50a4b2f691faecea5e	e9f50caf99f032f0bf3c55141f019d99	60115	fortaleza	CE	CS99439
99439	56b18e2166679b8a959d72dd06da27f9	73c2643a0a458b49f58cea58833b192e	92120	canoas	RS	CS99440
99440	274fa6071e5e17fe303b9748641082c8	84732c5050c01db9b23e19ba39899398	6703	cotia	SP	CS99441

99441 rows × 6 columns

Figure 3.3*: The olist_customers_dataset with changed customer_id*

(Source: Experimental results)

3.3.2. olist_geolocation_dataset

Drop the coordinate column and remove rows with duplicates in the geolocation_zip_code_prefix column.

```
1. Basic infomation of the geolocation dataset:
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19015 entries, 0 to 999846
Data columns (total 3 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   geolocation_zip_code_prefix  19015 non-null  int64
 1   geolocation_city             19015 non-null  object
 2   geolocation_state            19015 non-null  object
dtypes: int64(1), object(2)
memory usage: 594.2+ KB
None
-----------------------------------------------------------------
2. Number of columns and rows in the geolocation dataset:  (19015, 3)
-----------------------------------------------------------------
3. Numbers of unique value of zip code  19015
```

Figure 3.4: *The information of olist_geolocation_dataset (Source: Experimental results)*

```
[14] df_geolocation = df_geolocation.drop_duplicates(subset=['geolocation_zip_code_prefix'])
```

```
df_geolocation
```

	geolocation_zip_code_prefix	geolocation_city	geolocation_state
0	1037	sao paulo	SP
1	1046	sao paulo	SP
3	1041	sao paulo	SP
4	1035	sao paulo	SP
5	1012	são paulo	SP
...	...	...	...
999774	99955	vila langaro	RS
999780	99970	ciriaco	RS
999786	99910	floriano peixoto	RS
999803	99920	erebango	RS
999846	99952	santa cecilia do sul	RS

19015 rows × 3 columns

Figure 3.5: *The olist_geolocation_dataset with dropped duplicates rows*

(Source: Experimental results)

3.3.3. olist_order_item_dataset

```
1. Basic infomation of the order item dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 112650 entries, 0 to 112649
Data columns (total 7 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   order_id             112650 non-null  object
 1   order_item_id        112650 non-null  int64
 2   product_id           112650 non-null  object
 3   seller_id            112650 non-null  object
 4   shipping_limit_date  112650 non-null  datetime64[ns]
 5   price                112650 non-null  float64
 6   freight_value        112650 non-null  float64
dtypes: datetime64[ns](1), float64(2), int64(1), object(3)
memory usage: 6.0+ MB
None
-----------------------------------------------------------------
2. Number of columns and rows in the order item dataset:  (112650, 7)
-----------------------------------------------------------------
3. Number of order_id in order_item dataset : 98666
-----------------------------------------------------------------
4. Number of product_id in order_item dataset:  32951
-----------------------------------------------------------------
5. Number of seller_id in order_item dataset:  3095
```

Figure 3.6: *The information of olist_order_item_dataset (Source: Experimental results)*

3.3.4. olist_order_payment_dataset

```
1. Basic infomation of the order payment dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103886 entries, 0 to 103885
Data columns (total 5 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   order_id              103886 non-null  object
 1   payment_sequential    103886 non-null  int64
 2   payment_type          103886 non-null  object
 3   payment_installments  103886 non-null  int64
 4   payment_value         103886 non-null  float64
dtypes: float64(1), int64(2), object(2)
memory usage: 4.0+ MB
None
-----------------------------------------------------------------
2. Number of columns and rows in the order payment dataset:  (103886, 5)
-----------------------------------------------------------------
3. Number of order_id : 99440
-----------------------------------------------------------------
4. List of payment method : ['credit_card' 'boleto' 'voucher' 'debit_card' 'not_defined']
-----------------------------------------------------------------
5. Number of uses of payment method:  credit_card    76795
boleto          19784
voucher          5775
debit_card       1529
not_defined         3
Name: payment_type, dtype: int64
```

Figure 3.7*: The information of olist_order_payment_dataset*

(Source: Experimental results)

```
[30] df_order_payments[df_order_payments['payment_installments'] > 1]['payment_type'].value_counts()

     credit_card    51338
     Name: payment_type, dtype: int64
```

order_id which has more than one payment_installment uses credit card to pay

Figure 3.8: *The payment_type of over one payment_installments orders*

(Source: Experimental results)

3.3.5. olist_order_review_dataset

```
1. Basic infomation of the order review dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99224 entries, 0 to 99223
Data columns (total 7 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   review_id                99224 non-null  object
 1   order_id                 99224 non-null  object
 2   review_score             99224 non-null  int64
 3   review_comment_title     11568 non-null  object
 4   review_comment_message   40977 non-null  object
 5   review_creation_date     99224 non-null  object
 6   review_answer_timestamp  99224 non-null  datetime64[ns]
dtypes: datetime64[ns](1), int64(1), object(5)
memory usage: 5.3+ MB
None
-----------------------------------------------------------------
2. Number of columns and rows in the order review dataset (99224, 7)
-----------------------------------------------------------------
3. Number of null values in the order review dataset:
review_id                      0
order_id                       0
review_score                   0
review_comment_title       87656
review_comment_message     58247
review_creation_date           0
review_answer_timestamp        0
dtype: int64
-----------------------------------------------------------------
4. List of value of review score:  [4 5 1 3 2]
```

Figure 3.9: *The information of olist_order_review_dataset (Source: Experimental results)*

3.3.6. olist_order_dataset

```
1. Basic infomation of the order dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Data columns (total 8 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   order_id                       99441 non-null  object
 1   customer_id                    99441 non-null  object
 2   order_status                   99441 non-null  object
 3   order_purchase_timestamp       99441 non-null  datetime64[ns]
 4   order_approved_at              99281 non-null  datetime64[ns]
 5   order_delivered_carrier_date   97658 non-null  datetime64[ns]
 6   order_delivered_customer_date  96476 non-null  datetime64[ns]
 7   order_estimated_delivery_date  99441 non-null  datetime64[ns]
dtypes: datetime64[ns](5), object(3)
memory usage: 6.1+ MB
None
-----------------------------------------------------------------
2. Number of columns and rows in the order dataset (99441, 8)
-----------------------------------------------------------------
3. List of order status:
['delivered' 'invoiced' 'shipped' 'processing' 'unavailable' 'canceled'
 'created' 'approved']
-----------------------------------------------------------------
4. Number of null value in the order dataset
order_id                            0
customer_id                         0
order_status                        0
order_purchase_timestamp            0
order_approved_at                 160
order_delivered_carrier_date     1783
order_delivered_customer_date    2965
order_estimated_delivery_date       0
dtype: int64
```

Figure 3.10: *The information of olist_order_dataset (Source: Experimental results)*

3.3.7. olist_products_dataset

```
1. Basic infomation of the product dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32951 entries, 0 to 32950
Data columns (total 10 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   product_id                  32951 non-null  object
 1   product_category_name       32341 non-null  object
 2   product_name_lenght         32341 non-null  float64
 3   product_description_lenght  32341 non-null  float64
 4   product_photos_qty          32341 non-null  float64
 5   product_weight_g            32949 non-null  float64
 6   product_length_cm           32949 non-null  float64
 7   product_height_cm           32949 non-null  float64
 8   product_width_cm            32949 non-null  float64
 9   new_product_id              32951 non-null  object
dtypes: float64(7), object(3)
memory usage: 2.5+ MB
None
-----------------------------------------------------------------
2. Number of columns and rows in the product dataset (32951, 10)
-----------------------------------------------------------------
3. Number of category 73
-----------------------------------------------------------------
4. Number of null value in the products dataset
product_id                      0
product_category_name         610
product_name_lenght           610
product_description_lenght    610
product_photos_qty            610
product_weight_g                2
product_length_cm               2
product_height_cm               2
product_width_cm                2
new_product_id                  0
dtype: int64
```

Figure 3.11: *The information of olist_products_dataset (Source: Experimental results)*

```
Products paremeters
-----------------------------------------------------------------
Range of length of products:
Max length is:  105.0  cm
Min length is:  7.0  cm
-----------------------------------------------------------------
Range of weight of products:
Max weight is:  40425.0  g
Min weight is:  0.0  g
-----------------------------------------------------------------
Range of height of products:
Max height is:  105.0  cm
Min height is:  2.0  cm
-----------------------------------------------------------------
Range of width of products:
Max width is:  118.0  cm
Min width is:  6.0  cm
```

Figure 3.12*: The Olist's products parameters (Source: Experimental results)*

3.3.8. olist_product_category_dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 71 entries, 0 to 70
Data columns (total 2 columns):
 #   Column                         Non-Null Count  Dtype 
---  ------                         --------------  ----- 
 0   product_category_name          71 non-null     object
 1   product_category_name_english  71 non-null     object
dtypes: object(2)
memory usage: 1.2+ KB
```

Figure 3.13: *The information of olist_product_category_dataset*

(Source: Experimental results)

3.4. Exploratory Analysis and Modeling

3.4.1. Load data after processing with SSIS tool

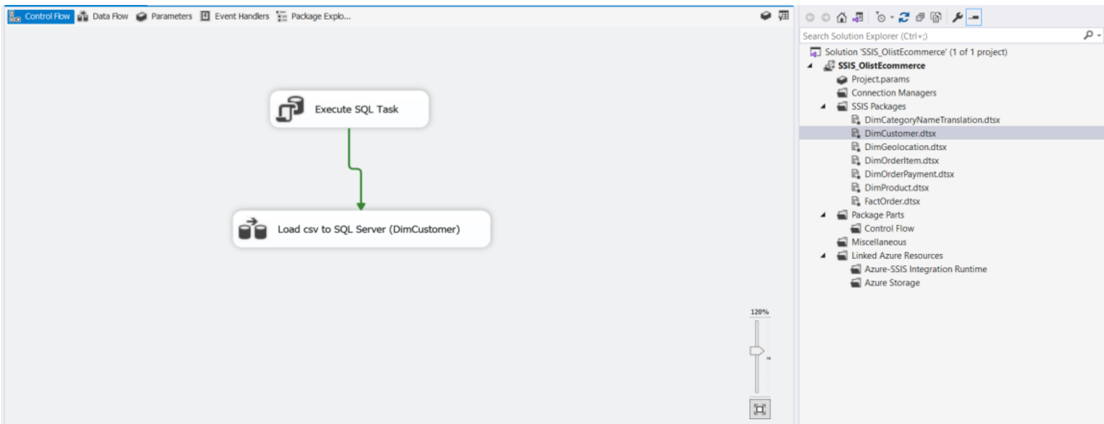

Figure 3.14: *Using SSIS tool to load data after preprocessing*

(Source: Experimental results)

After EDA and preparing the data in Python, the team loaded the data into SQL Server using the SSIS tool. While loading data, the team edits the appropriate data type and removes columns not needed for analysis and creates additional necessary columns.

Table 3.10*: Table describes data types of columns in FactOrder in SQL and SSIS*

Table	**Column**	**Data type loaded into SQL Server**	**Data types selected in SSIS**
FactOrder	order_id *	VARCHAR(50)	DT_STR

	customer_id *	VARCHAR(50)	DT_STR
	order_status	VARCHAR(20)	DT_STR
	order_purchase_timestamp	DATETIME	DT_DBTIMESTAMP
	order_approved_at	DATETIME	DT_DBTIMESTAMP
	order_delivered_carrier_date	DATETIME	DT_DBTIMESTAMP
	order_delivered_customer_date	DATETIME	DT_DBTIMESTAMP
	order_estimated_delivery_date	DATE	DT_DBDATE
	order_purchase_date	DATE	Added in SQL

Table 3.11: *Table describes data types of columns in DimCustomer in SQL and SSIS*

Table	Column	Data type loaded into SQL Server	Data types selected in SSIS
DimCustomer	customer_id*	VARCHAR(50)	DT_STR
	new_customer_id	VARCHAR(50)	DT_STR
	customer_zip_code	INT	two-byte signed integer [DT_I2]

Table 3.12: *Table describes data types of columns in DimGeolocation in SQL and SSIS*

Table	Column	Data type loaded into SQL Server	Data types selected in SSIS
DimGeolocation	geolocation_zip_code	INT	DT_I4
	geolocation_city	VARCHAR(50)	DT_STR

	geolocation_state	VARCHAR(50)	DT_STR

Table 3.13: *Table describes data types of columns in DimOrderItem in SQL and SSIS*

Table	Column	Data type loaded into SQL Server	Data types selected in SSIS
DimOrderItem	order_id *	VARCHAR(50)	DT_STR
	order_item_id *	SMALLINT/ INT	DT_STR
	product_id *	VARCHAR(50)	DT_STR
	price	DECIMAL(10, 2)	DT_DECIMAL
	freight_value	DECIMAL(8, 2)	DT_DECIMAL

Table 3.14: *Table describes data types of columns in DimOrderPayment in SQL and SSIS*

Table	Column	Data type loaded into SQL Server	Data types selected in SSIS
DimOrderPayment	order_id *	VARCHAR(50)	DT_STR
	payment_sequential	INT	DT_I4
	payment_type	VARCHAR(50)	DT_STR
	payment_installments	INT	DT_I4
	payment_value	DECIMAL(10, 2)	Decimal

Table 3.15: *Table describes data types of columns in DimOrderReview in SQL and SSIS*

Table	Column	Data type loaded into SQL Server	Data types selected in SSIS

DimOrderReview	review_id *	VARCHAR(50)	DT_STR
	order_id *	VARCHAR(50)	DT_STR
	review_score	BIT	
	review_comment_title (n)	TEXT	
	review_comment_message (n)	LONGTEXT	
	review_creation_date	DATETIME	
	review_answer_timestamp	DATETIME	

Table 3.16: *Table describes data types of columns in DimProduct in SQL and SSIS*

Table	Column	Data type loaded into SQL Server	Data types selected in SSIS
DimProduct	product_id *	VARCHAR(50)	DT_STR
	product_category_name	VARCHAR(50)	DT_STR
	product_weight_g	INT	DT_I4
	product_length_cm	INT	DT_I4
	product_height_cm	INT	DT_I4
	product_width_cm	INT	DT_I4

Table 3.17: *Table describes data types of columns in DimCategoryNameTranslation in SQL and SSIS*

Table	Column	Data type loaded	Data types

		into SQL Server	selected in SSIS
DimCategoryNameTranslation	product_category_name	VARCHAR(20)	DT_STR
	product_category_name_english	VARCHAR(20)	DT_STR

After loading data into SQL, the team created an olist.DimDate table as a time dimension for the data set.

Table 3.18: *Table describes data types of columns in DimDate in SQL*

Table	Column	Data type into SQL Server
DimDate	Date	DATE
	Day	CHAR(10)
	DayOfWeek	TINYINT
	DayOfMonth	TINYINT
	DayOfMonth	SMALLINT
	WeekOfYear	TINYINT
	Month	CHAR(10)
	MonthOfYear	TINYINT
	QuarterOfYear	TINYINT
	Year	CHAR(10)

Sample data table:

Date	Day	DayOfWeek	DayOfMonth	DayOfYear	WeekOfYear	Month	MonthOfYe...	QuarterOfY...	Year
2015-01-01	Thursday	5	1	1	1	January	1	1	2015
2015-01-02	Friday	6	2	2	1	January	1	1	2015
2015-01-03	Saturday	7	3	3	1	January	1	1	2015
2015-01-04	Sunday	1	4	4	2	January	1	1	2015
2015-01-05	Monday	2	5	5	2	January	1	1	2015
2015-01-06	Tuesday	3	6	6	2	January	1	1	2015
2015-01-07	Wednesday	4	7	7	2	January	1	1	2015
2015-01-08	Thursday	5	8	8	2	January	1	1	2015
2015-01-09	Friday	6	9	9	2	January	1	1	2015
2015-01-10	Saturday	7	10	10	2	January	1	1	2015
2015-01-11	Sunday	1	11	11	3	January	1	1	2015
2015-01-12	Monday	2	12	12	3	January	1	1	2015
2015-01-13	Tuesday	3	13	13	3	January	1	1	2015
2015-01-14	Wednesday	4	14	14	3	January	1	1	2015
2015-01-15	Thursday	5	15	15	3	January	1	1	2015
2015-01-16	Friday	6	16	16	3	January	1	1	2015
2015-01-17	Saturday	7	17	17	3	January	1	1	2015
2015-01-18	Sunday	1	18	18	4	January	1	1	2015
2015-01-19	Monday	2	19	19	4	January	1	1	2015

Figure 3.15: *Sample data of DimDate table (Source: Experimental results)*

After the data has been loaded into SQL Server, the team proceeds to load the data into Power BI to model the data and prepare for the Validation step.

3.4.2. Load data into Power BI and build a data model

First, the authors will connect Power BI with SQL to import all tables and views created in SQL Server

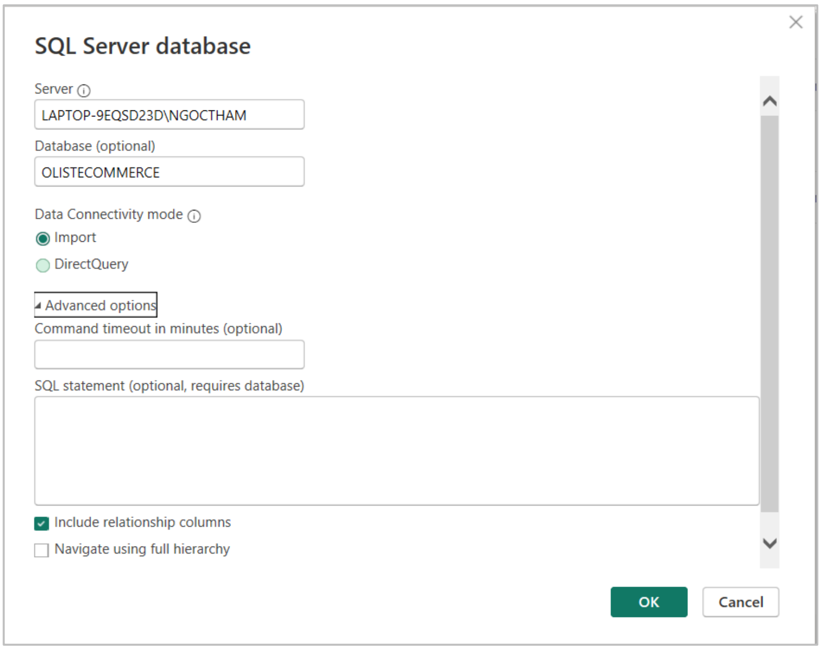

Figure 3.16: *Connecting Power BI with SQL Server (Source: Experimental results)*

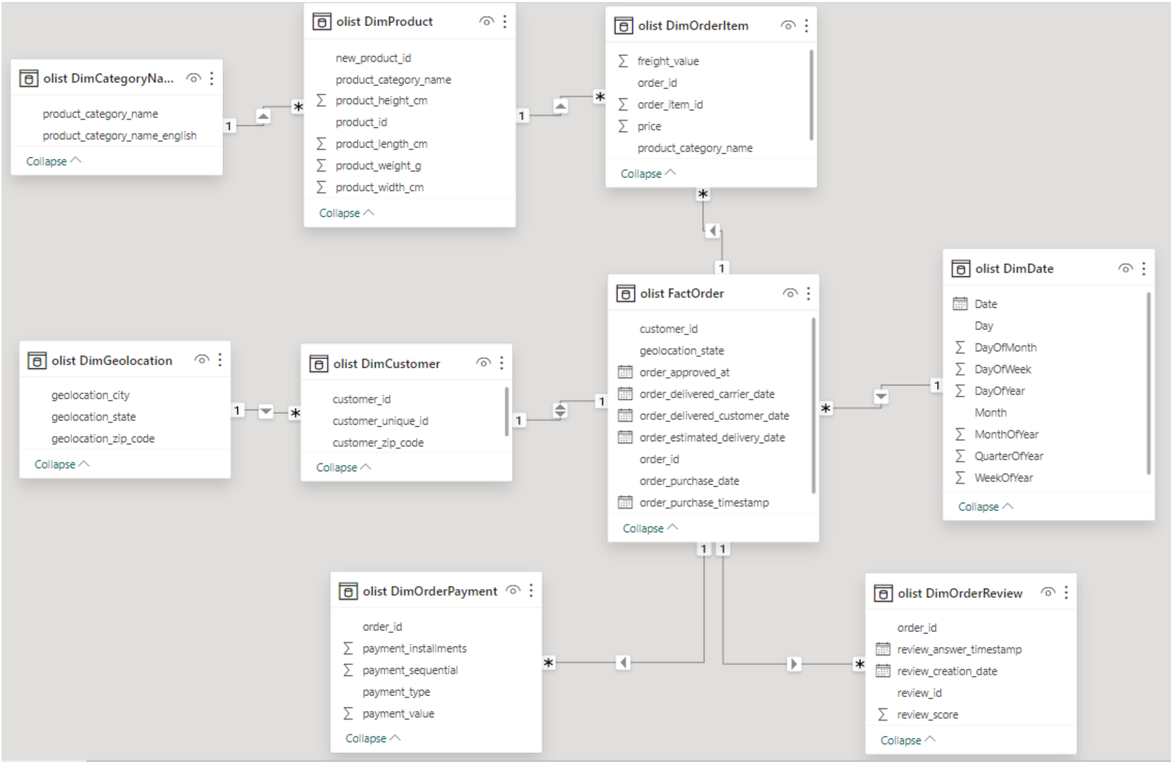

Figure 3.17: *The data model built in Power BI (Source: Experimental results)*

The relationship between tables is explained in Table 3.19.

Table 3.19: *Table describes the relationship between tables in Power BI*

Table	Key column	Relationship	Explanation
olist FactOrder	customer_id	One to one (1:1)	customer_id in the customer table represents a transaction, not a customer.
olist DimCustomer	customer_id		
olist FactOrder	order_id	One to many (1:*)	One order can be reviewed many times. Reviews are written for order, not for products.
olist DimReview	order_id		
olist FactOrder	order_purchase_date	Many to one (*:1)	

olist DimDate	Date		
olist FactOrder	order_id	One to many (1:*)	An order can have more than one order item
olist DimOrderItem	order_id		
olist FactOrder	order_id	One to many (1:*)	An order can be paid many times.
olist DimPayment	order_id		
olist DimOrderItem	product_id	Many to one (*:1)	An order item contains one product. But one product can be in many items.
olist DimProduct	product_id		
olist DimProduct	product_category_name	Many to one (*:1)	A product belongs to one category. A category can have more than product
olist DimCateGoryTranslation	product_category_name		
olist DimCustomer	customer_zip_code	Many to one (*:1)	A customer lives in one location. A location has more than one customer.
olist Geolocation	geolocation_zip_code		

The authors built a Snowflake-style data model, including a central FactOrder olist table and connected to Dim tables, respectively: olist DimProduct, olist DimOderTime, olist DimDate, olist DimCustomer, olist Geolocation, olist DimOrderPayment and olist DimCategoryNameTranslation.

3.5. Validation

The tables of the dataset are preprocessed, and removed unnecessary attributes in some tables. When loading data into SQL, the data types have been changed to the appropriate data types and a new table named olist DimDate has been created. Then, the data is loaded into Power BI, building a Snowflake-style data model. During the adjustment process, the data set still ensures consistency and accuracy. The relationships between the tables are proper, the retrieval process for analysis does not generate errors.

CHAPTER 4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Sales Performance Analysis (Order)

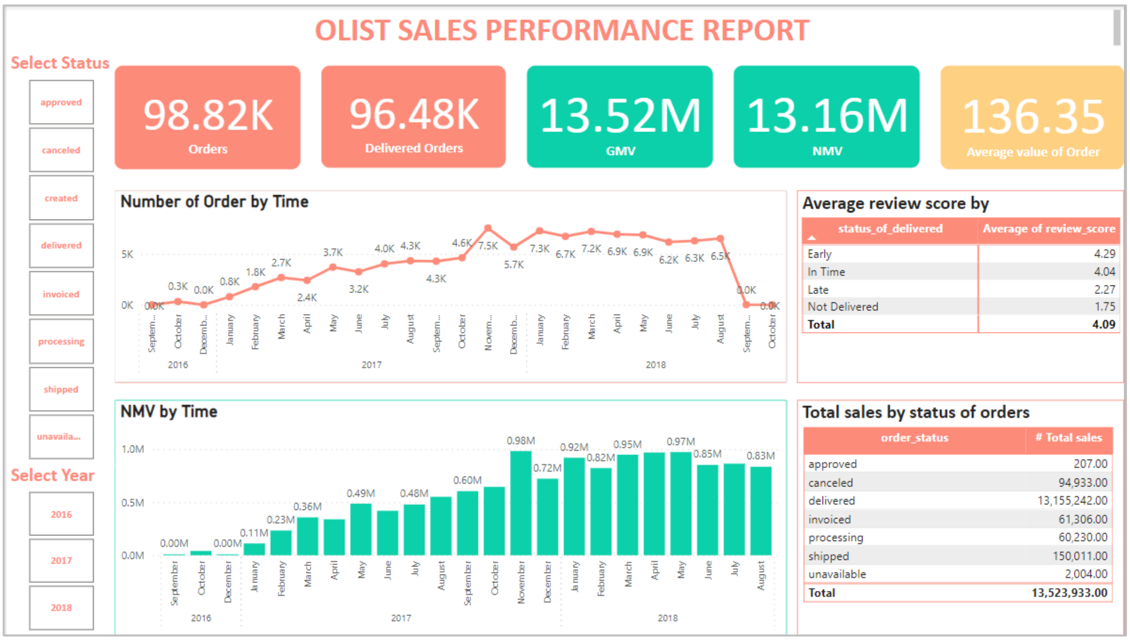

Figure 4.1: *Olist sales performance report – page 1 (Source: Experimental results)*

From September 2016 to August 2018, the number of Olist's orders was 98816 with more than 96 thousand delivered orders, accounting for 96.63%. Gross Merchandise Volume (GMV) is calculated as the total value of all orders in all delivery statuses (canceled, delivered, approved, invoiced,...) reaching 13.52 million. Of which, NMV (total value of successfully delivered orders) is 13.16M. Estimated average value that customers are willing to pay for each order is 136.35.

The line graph shows that the number of orders is increasing day by day which means the demand for online shopping increases. It is a great potential for sellers to deploy and promote trade. The number of orders reached a peak at 7.5K in November 2017, which brought 0.98M for sales. It is understandable that at that time people need to purchase to prepare for upcoming festivals, such as Christmas and New Year.

The average review score of orders is 4.09, in which the orders with the highest rating are those that are delivered early (4.29) and on time (4.4). Orders that have low ratings mainly belong to the group of orders that are delivered late or have not been delivered.

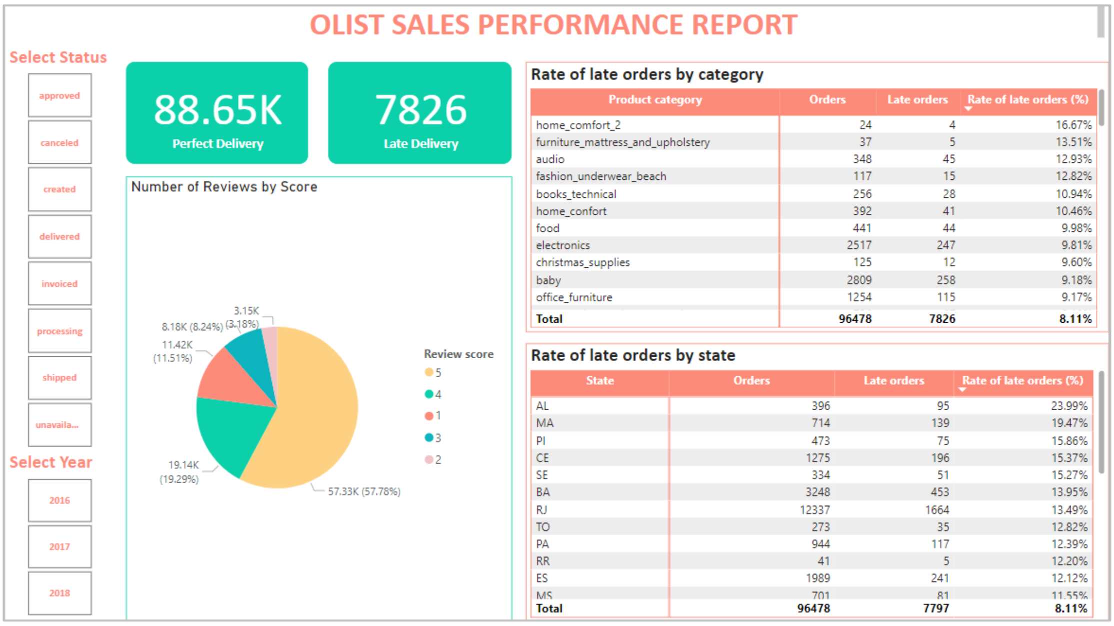

Figure 4.2: *Olist sales performance report – page 2 (Source: Experimental results)*

The products that consumers buy the most in the bed_bath_table category are household's appliances and items. These are the things they need to renew at the end of the year. In 2018, people's willingness to pay was highest in May (0.97M) for popular items in health_beauty. This is the time when the weather starts to change, it gets colder so they need to take care of themselves more. Both of these product groups have a large number of orders arriving late because demand is greater than supply and sellers do not store enough goods, leading to long preparation times, or because the shipping partners do not have enough employees during this busy season.

According to the Rate of late orders by state table, it is seen that Sao Paulo (SP) is the area where people shop online the most because this is Brazil's largest city, with a crowded population. Roraima (RR) is the state with the smallest total order because it is the least populated place in Brazil. Although SP is the state with the highest number of late orders, Alagoas (AL) is the state with the highest rate of late deliveries (23.99%).

4.2. Customer Analysis

After the analysis process, the authors figured out the following information related to customers of the e-commerce platform Olist:

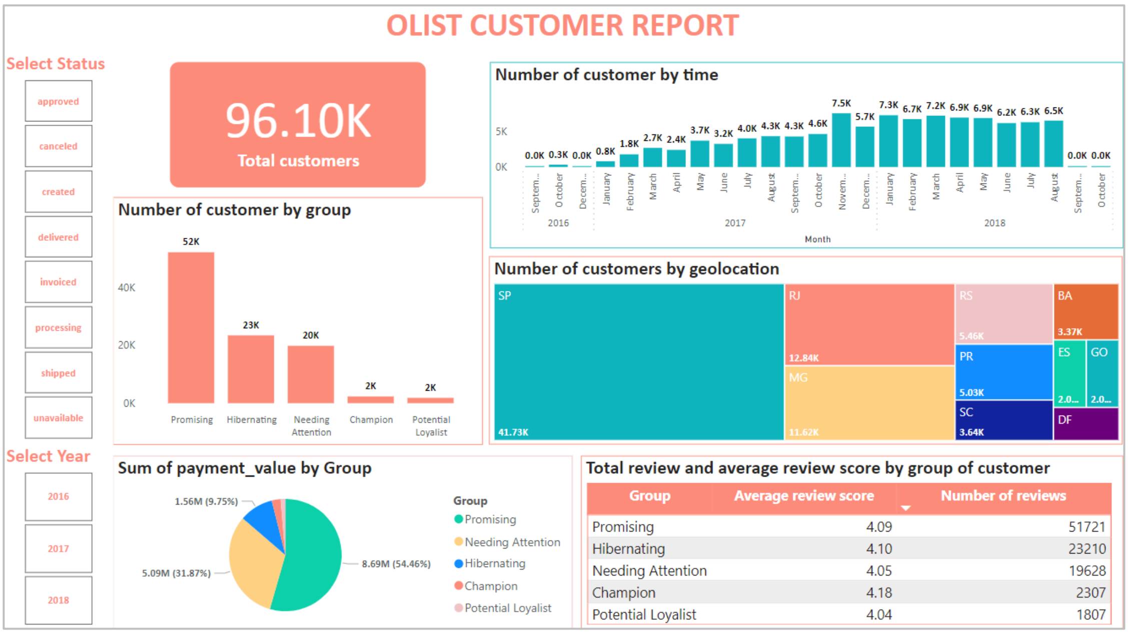

Figure 4.3: *Olist customer report – page 1 (Source: Experimental results)*

The total number of customers is 96.1 thousand customers.

Regarding geographical distribution, customers are mainly concentrated in SP, RJ and MG states.

Regarding the level of reviews by each customer group, the Champion group has a low number of reviews but the highest average score, proving that this customer group is quite satisfied with the product service quality. The Potential Loyalist group has the lowest score while it brings the most benefits after the Champion group. The business needs to survey the experiences of this group of customers to improve the quality of products and services to retain this group of customers.

To analyze customer consumption behavior, the team uses the RFM model (Recency, Frequency, Monetary). The team uses quintiles to divide customers into 5 groups based on recency, frequency and currency. For recency_score, a larger value means the customer purchased closer to the current date. For monetary_score, a larger value means the higher values a customer buys. For frequency_score, larger values mean customers buy more often. The authors then proceeds to divide customer groups based on rfm_score as follows (Table 4.1):

Table 4.1: *Table describes five customer groups by rfm_score*

rfm_score	Customer group	Purchasing behavior	Suggestions
555, 554, 544, 545, 454, 455, 445	**Champion**	This is a group of customers who buy a lot and frequently and have bought recently, proving that they have a certain trust in the business	The business needs to retain this group of customers with loyalty programs, personalized marketing content or gratitude programs based on gold, silver and bronze levels.
543, 444, 435, 355, 354, 345, 344, 335, 553, 551, 552, 541, 542, 533, 532, 531, 452, 451, 442, 441, 431, 453, 433, 432, 423, 353, 352, 351, 342, 341, 333, 323, 535, 534, 443, 434, 343, 334	**Potential Loyalist**	This is a group of customers who buy frequently and have recently purchased but their spending is not high.	For this group of customers, the business should increase spending by upselling or cross-selling by reducing prices when buying combos, or reducing prices when orders reach a certain value.
512, 511, 422, 421, 412, 411, 311,525, 524, 523, 522, 521, 515, 514, 513, 425, 424, 413, 414, 415, 315, 314, 313	**Promising**	This is a group of customers who have recently purchased but do not buy regularly	This group of customers is still in the "trial experience" stage, so businesses need to make them more trusted by sending promotional

			vouchers after their first purchase or accumulating points based on the number of purchases.
325, 324, 255, 254, 245, 244, 253, 252, 243, 242, 235, 234, 225, 224, 153, 152, 145, 143, 142, 135, 134, 133, 125, 124, 155, 154, 144, 214, 215, 115, 114, 113	**Needing Attention**	This is a group of customers who have not returned recently even though they had previously bought a lot with high frequency.	Businesses need to find out the reasons why customers do not return by surveying customer experience. Promotions for this customer group may not be effective in cases when they leave due to dissatisfaction with service quality and products.
331, 321, 312, 221, 213, 332, 322, 231, 241, 251, 233, 232, 223, 222, 132, 123, 122, 212, 211, 111, 112, 121, 131, 141, 151	**Hibernating**	This is a group of customers who have not returned for a long time, do not buy frequently and have low purchasing power. This could just be customers buying a product while their favorite product is	For this group of customers, the business only needs to maintain brand awareness with customers by interacting via social media, to avoid being forgotten by them. Promoting promotional advertising for this group of customers will

		out of stock or buying on a whim	not be effective.

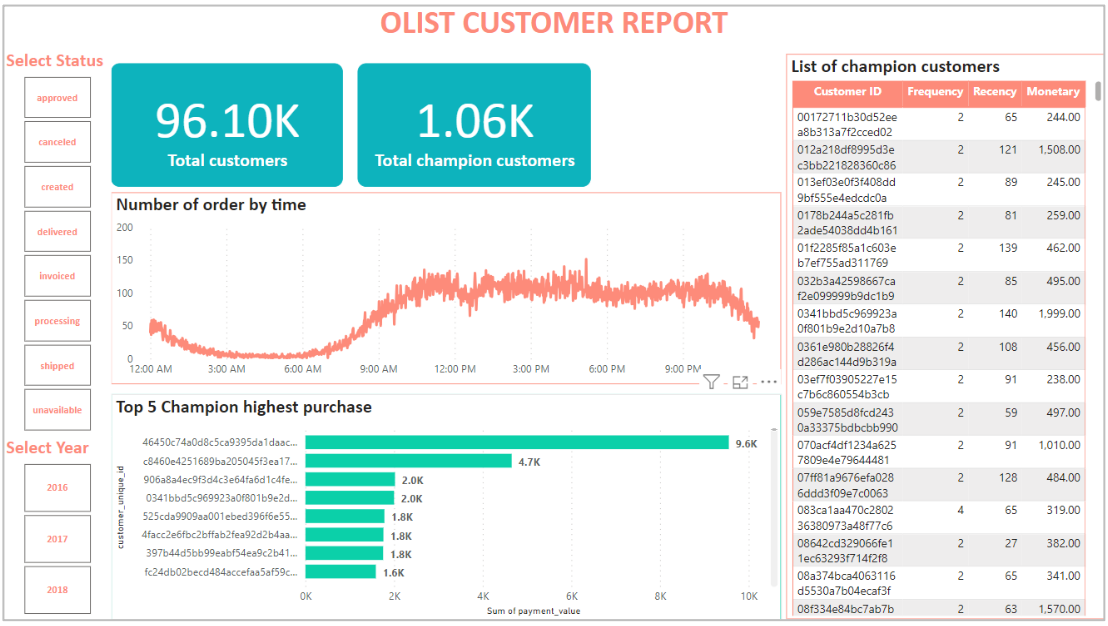

Figure 4.4: *Olist customer report – page 2 (Source: Experimental results)*

For purchasing time, the line chart shows that customers often buy from around 9 am, with a peak between 10 and 11 am, with a slight decrease at 12 am. This trend is the same across all customer groups. Therefore, the business should promote advertising in the time frame from 8:00 a.m. to 9:00 a.m. – before customers' purchasing time to increase the reach rate because customers will spend time choosing before buying.

4.3. Product Analysis

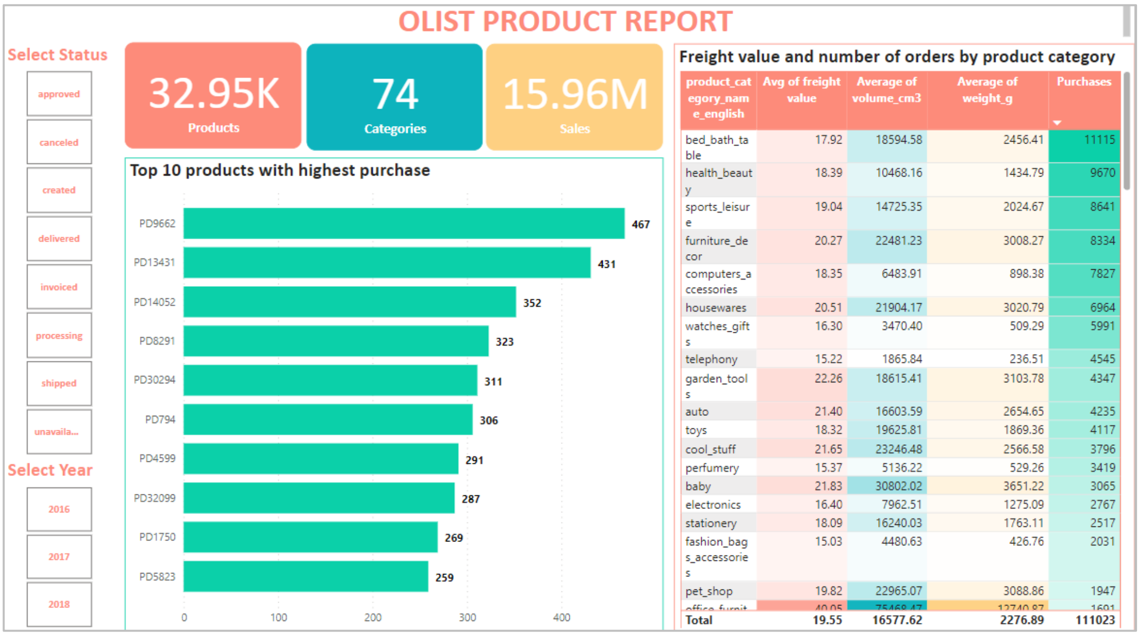

__Figure 4.5:__ Olist product report (Source: Experimental results)

The product data has 32.95 thousand products and 74 product categories and the revenue is 15.96M. The product with ID 'PD9662' is the most purchased product, with a total of 467 purchases.

Among the product categories with the most purchases, the bed_bath_table had the highest purchases with 11115 times. Followed by health-beauty products with 9670 purchases.

The average weight and volume of the product do not affect the purchasing decisions but affect the freight value. Freight value doesn't affect the purchasing decisions. The baby product category has the highest average weight and volume with 30802.02 cm^3 and 3651,22g respectively, but it is not the best-selling product category.

4.4. Logistic Analysis

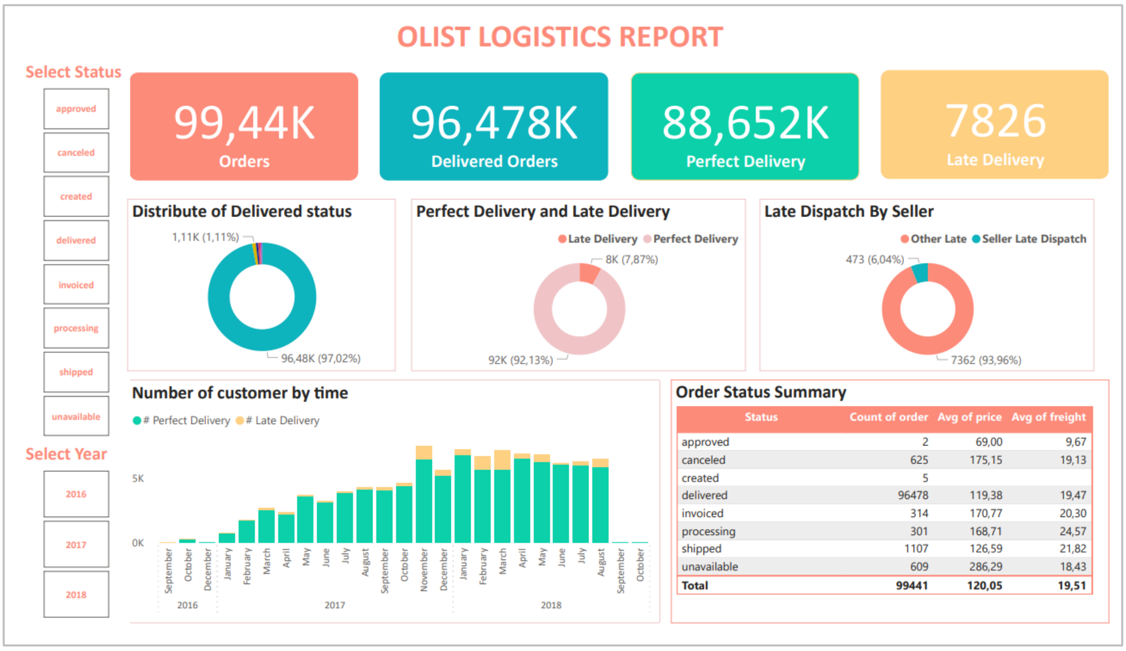

Figure 4.6: *Olist logistics report – page 1 (Source: Experimental results)*

The dashboard shows that Olist has received 100k orders, of which 96k orders were successfully delivered. There were 88.65k early and on time orders. This figure is equivalent to 92% orders in total.

There were 7,826 late orders in total and there were "extremely late by the seller" 475 orders. It is called “extremely late” because the date from the seller was carried later than the expected delivery date to the customer. The number of "extremely late by the seller" orders accounted for 6.04% in total late orders.

Other objective factors that can lead to delay include distance, shipping unit, weather, etc. To overcome subjective reasons, Olist should focus on reducing delays when sending goods of sellers, optimize carrier performance, and ensure timely deliveries.

There are 625 canceled orders, including 1 1 order that was delivered but canceled. Canceled orders have a large average value and average shipping costs, so it can be supposed that these orders do not have difficulty in shipping distance. Therefore, for large value orders that do not have many transportation difficulties, Olist should pay attention to canceled orders.

The rate of late orders was high in the last quarter of 2017 and the first 2 quarters of 2018. This is also the highest buying time of Olist. In March 2018, February 2018 and November 2017, the late orders rate accounted for 20% of successfully delivered orders. Although the rate of late orders decreased, it tended to increase gradually after. This will affect the reputation of the business, the logistic partner and even the sellers. Therefore, Olist should work with shipping units, or cooperate with additional shipping units to limit the possibility of not being able to promptly deliver orders to customers. Especially at the end of the year and the beginning of the upcoming year since that is the time when users buy the most for holidays.

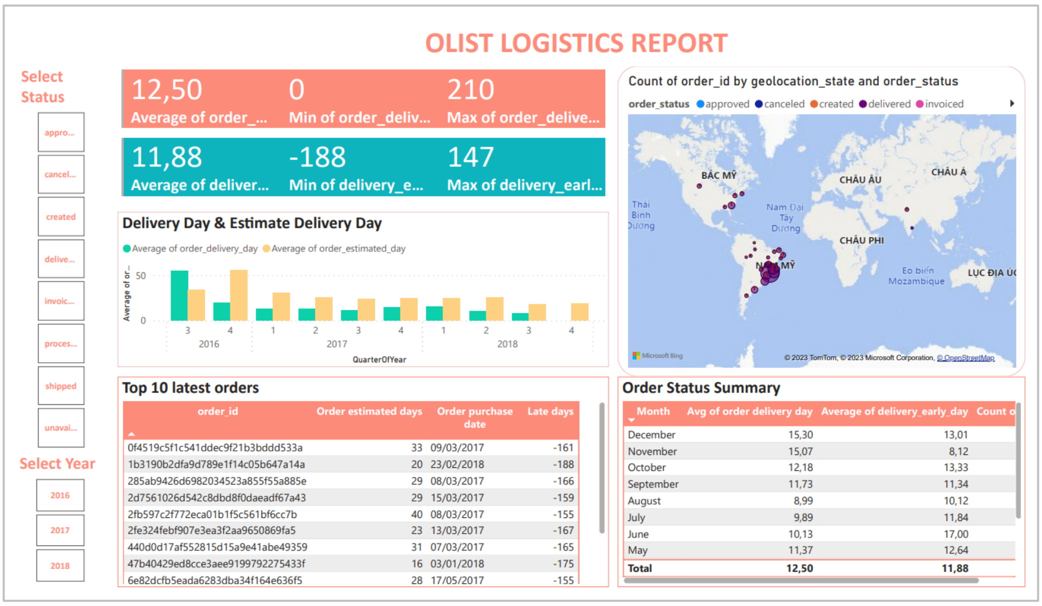

Figure 4.7: *Olist logistics report – page 2 (Source: Experimental results)*

Calculate some values:

- ***Delivery day = delivery customer date - purchase date***: is the time from the purchase date until the customer delivery date.
- ***Delivery Early day = estimated date - delivery customer date***: is the time it takes to the customer delivery date compared to the estimate delivery date - that Olist shows to the customer

- ***Estimate day = Estimated date - purchase date:*** is the expected time from the customer purchasing the product until receiving the product predicted by Olist's system

Delivery day values from 0 to 210. There were orders that are delivered within a day and orders that took nearly 7 months to reach the user

In the top 10 orders with the worst Delivery Early day, 9 orders are all in areas within Brazil. All of them were delivered more than 5 months later than expected, and compared to the actual delivery date. It means that customers had to wait longer, from 1 to 2 months. The business should review the expected delivery system, identify some special situations that cause goods to arrive longer than usual, such as: custom designed goods, order goods, etc.. For transportation and goods inspection, the business should check carefully to avoid missing or lost goods and deliver on time.

Except for special cases, deliveries on average are earlier than the expected time. The fourth quarter of 2016 shows a notable high estimated delivery time, this is probably the time when many orders go abroad. Although the distance was considerable, the delivery time was still very early.

In the third quarter of 2016, the actual delivery time was significantly longer than the estimated delivery time. During this period, many orders were sent overseas, so the estimated time was often longer. But the actual delivery time being shorter indicates that the Olist e-commerce has paid great attention to distant orders.

CHAPTER 5. CONCLUSION

5.1. Results

Sales Performance Analysis

1. Which locations have the most orders?

SP, RJ, and MG are the areas with the most orders.

2. Which time period has the most orders?

People often shop the most on occasions such as the end of the year to prepare for year-end festivals and in May, when the weather begins to change.

3. Does delivery time affect an order's rating?

According to the Average review score table, it can be seen that the delivery time has an impact on the review score.

4. Characteristics of canceled orders.

There were 625 canceled orders, the average value of these was at around 175 (higher than the average value of all), most of these were not delivered to customers. SP, RJ, and MG were states with the highest number of canceled orders.

Customer Analysis

1. In which areas are customers concentrated?

Customers distribute the most in 3 states: SP, RJ and MG.

2. When do customers shop?

Customers shop the most around 10am to 11am every day.

3. Which customer groups need attention?

Each customer group needs to be treated with different strategies based on their consumption behavior.

Table 5.1: *Table of summary the proposed marketing strategy for five customer groups*

Customer group	Consumer behavior	Marketing strategy
Champion	High purchasing value, frequent and recently purchased	Apply the promotion program according to gold, silver and bronze levels.
Potential Loyalist	Average purchasing volume, purchasing frequently and recently	Upsell or cross sell by introducing combos or promotions when orders reach a certain value
Promising	Just started shopping recently	Promotion for first order or accumulate points based on number of purchases
Needing Attention	Used to have high purchasing value, buy often but haven't come back for a long time	Survey their purchasing experience to find out why they don't return
Hibernating	Low purchasing value, don't buy regularly and haven't been back for a long time	Maintain brand recognition with customers

Product Analysis

The best-selling product category mainly includes household items, furniture used in homes, sports equipment,...

Logistic Analysis

1. Which area are late orders from?

Late orders are distributed in the states of AL, MA in the US and areas around Brazil in South America.

2. Average time to ship an order?

On average, it took 17 days for orders to reach customers. However, shipping time has a large variation. The fastest order was delivered on the same day and the longest order took 7 months for the order to arrive at the right place.

3. Which month of the year has the most orders delivered late?

At the end of the year and the beginning of the year, there were a lot of orders, leading to more late orders. Most of them were in January, March, November.

5.2. Future work

The future development of research will be to apply machine learning models to predict revenue in the next year for businesses. Revenue forecasting can be carried out using machine learning models such as BG/NBD or ARIMA. Forecasting revenue will help businesses prepare for demand, thereby improving business performance.

In addition, further research can use machine learning models such as K-Means or DBSCAN to cluster customers, finding the appropriate number of groups using the Elbow method. The advantage of clustering using the above machine learning model is that its suitability can be evaluated by using indexes, such as Silhouette Score or DB Index.

REFERENCES

[1] Kryshtanovych, S., Prosovych, O., Panas, Y., Trushkina, N., & Omelchenko, V. (2022). Features of the Socio-Economic Development of the Countries of the World under the influence of the Digital Economy and COVID-19. International Journal of Computer Science and Network Security, 22(1), 9-14.

[2] Pan, C. L., Bai, X., Li, F., Zhang, D., Chen, H., & Lai, Q. (2021, March). How Business Intelligence Enables E-commerce: Breaking the Traditional E-commerce Mode and Driving the Transformation of Digital Economy. In 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT) (pp. 26-30). IEEE.

[3] Biswas, A. K., & Hariharan, K. (2023). Business Intelligence-The Propeller of eCommerce Business. Academy of Marketing Studies Journal, 27(S2).

[4] Antunes, A. L., Cardoso, E., & Barateiro, J. (2022). Incorporation of ontologies in data warehouse/business intelligence systems-a systematic literature review. International Journal of Information Management Data Insights, 2(2), 100131.

[5] Ali, O. T., Nassif, A. B., & Capretz, L. F. (2013, June). Business intelligence solutions in healthcare a case study: Transforming OLTP system to BI solution. In 2013 Third International Conference on Communications and Information Technology (ICCIT) (pp. 209-214). IEEE.

[6] Zeng, L., Xu, L., Shi, Z., Wang, M., & Wu, W. (2006, October). Techniques, process, and enterprise solutions of business intelligence. In 2006 IEEE international conference on systems, man and cybernetics (Vol. 6, pp. 4722-4726). IEEE.

[7] de Sousa Gonçalves, C. MGI (Doctoral dissertation, Universidade Nova de Lisboa).

[8] Kimball, R., & Caserta, J. (2004). The data warehouse ETL toolkit. John Wiley & Sons.

[9] Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking–An effective approach to customer segmentation. Journal of King Saud University-Computer and Information Sciences, 33(10), 1251-1257.

[10] Coussement, K., Van den Bossche, F. A., & De Bock, K. W. (2014). Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. Journal of Business Research, 67(1), 2751-2758.

[11] Garani, G., Chernov, A., Savvas, I., & Butakova, M. (2019, June). A data warehouse approach for business intelligence. In 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE) (pp. 70-75). IEEE.

[12] Ferreira, T., Pedrosa, I., & Bernardino, J. (2019, June). Integration of Business Intelligence with e-commerce. In 2019 14th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-7). IEEE.

[13] Priyadarshini, P., & Veeramanju, K. T. (2022). Business Intelligence for the Evaluation of Customer Satisfaction in E-Commerce Websites-A Case Study. International Journal of Management, Technology and Social Sciences (IJMTS), 7(2), 660-668.

[14] Anggrainy, T. D., & Sari, A. R. (2022). Implementation of extract transform load on data warehouse and business intelligence using pentaho and tableau to analyse sales performance of offlist store. Science, 7(2), 368-374.

APPENDIX

Appendix 1: *Table describes the adjustments in this report*

No.	Section	Adjustment
1	**1.3.3. Research scope**	Editing the research scope based on the time period of the dataset.
2	**2.1.4 Data approach**	Adding the theory of the Kimball approach.
3	**2.1.5 RFM model** *(new section)*	Theory of RFM model.
4	**2.1.6 ETL** *(new section)*	Theory of ETL process.
5	**2.2. Related works**	Dividing related works into 3 sections by research topic instead of presenting all the research in a section.