ANALYSIS OF E-COMMERCE DATASET USING O-LIST AS A CASE STUDY

BY

AGWELI ISAAC

ENG1810267

OMO-OMORUYI OSATO JOSEPH

ENG1603916

DEPARTMENT OF COMPUTER ENGINEERING

FACULTY OF ENGINEERING

UNIVERSITY OF BENIN

BENIN CITY

SEPTEMBER 2023

ANLYSIS OF E-COMMERCE DATASET USING O-LIST AS A CASE STUDY

BY

AGWELI ISAAC

ENG1810267

OMO-OMORUYI OSATO JOSEPH

ENG1603916

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING, FACULTY OF ENGINEERING, UNIVERSITY OF BENIN

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR OF ENGINEERING (B.Eng.) IN COMPUTER ENGINEERING

September 2023

CERTIFICATION

This project was carried out by Agweli Isaac and Omo-Omoruyi Osato Joseph in the department of Computer Engineering, Faculty of Engineering, University of Benin, Benin City and is hereby Certified.

Engr. Dr. O. Okosun

Head Of Department

September 2023

Engr. Dr. O. Okosun

Project Supervisor

September 2023

DEDICATION

We dedicate this project report to God almighty for his mercies and abundant grace. We also dedicate this report to our parents who supported me through my educational parents.

ACKNOWLEDGEMENT

Special thanks to God almighty who has blessed us with the ability and strength to complete this project work. Special thanks to Engr. Dr. O. Okosun, my project supervisor, for encouraging and guiding me throughout this project work, to all the staff of the department of computer engineering for their positive impact on our life and this project work and my fellow course mate for their support both materially and academically.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

The purpose of this school project is to analyse an e-commerce dataset in order to gain insights into customer behaviour, product trend, private strategies and supply chain management. The project will involve collecting and cleaning the data, conducting exploratory data and using statistical and data analytical techniques to uncover pattern and trend. The project will also involve developing visualization and presenting findings to stake holders. By completing this project, student will gain valuable experience in data analysis and develop skills that are highly sought after in today's job market.

Additionally, the insight gained from this project may have practical application for business

Looking to optimize their ecommerce operation and improve the customer experience.

CHAPTER ONE

1.0 INTRODUCTION

E-commerce is not just about buying and selling products online; it's a dynamic ecosystem that involves marketing, logistics, customer behaviour, and much more. With the ever-increasing importance of e-commerce in our lives, understanding the patterns, trends, and dynamics of this industry is crucial for businesses, organizations, consumers, and policymakers alike.

Along with the rapid growth of the E-commerce business around the world, it made me and my partner curious to do an analysis project of an E-commerce store from Brazil, called **OLIST.** Olist's dataset, which spans several years and encompasses a wide range of variables, provides us with a rich source of information to explore and gain meaningful insights.

This dataset is a Brazilian e-commerce public dataset called Olist Store which includes about 100k orders between the years 2016-2018 made at multiple marketplaces in Brazil.

In this project, we will analyse total sales and customer reviews by using MySQL and create visualizations afterwards by using PowerBI.

We got the publicly released Olist Dataset (2016 2018) from https://www.kaggle.com/olistbr/b razilian-ecommerce/version/2.

With its vast selection of over 4.5 million registered products and support for more than 100,000 stores, Olist offers a plethora of growth opportunities for businesses of all sizes. As the market expands, it becomes increasingly crucial to stay ahead of the competition. This

report delves into key findings that will empower stakeholders to make informed business decisions and navigate the ever-evolving landscape.

As students, this project allows us to hone our data analysis skills, apply statistical techniques, and gain a deeper understanding of how data can be leveraged to make informed decisions. Throughout this project, we will collaborate, learn, and challenge ourselves to think critically about the data we encounter. Our findings and conclusions will not only enrich our academic journey but may also provide valuable perspectives for the world of E-commerce.

1.1 BACKGROUND OF STUDY

E-commerce has become an increasingly important part of the global economy, with millions of people around the world now using online platforms to buy and sell goods and services. As a result, businesses are investing heavily in their online presence and e-commerce operations in order to remain competitive and meet the changing needs of consumers.

Analysing and e-commerce dataset is a valuable exercise because it provides hands on experience in data analysis and can help develop skills that are sought after in today's job market. Furthermore, understanding customer behaviour, product trends, pricing strategies and supply chain management is essential for businesses that want to succeed in this space,

Through this project, we will have the opportunity to explore real-world data from an ecommerce store and gain insights into the factors that drive success in this competitive and rapidly evolving industry.

Overall, analysing an e-commerce store is a valuable and relevant topic providing a practical understanding of data analytics and its applications in real life situations.

1.2 PROBLEM STATEMENT

In the era of digital commerce, understanding the dynamics of e-commerce is critical for businesses and consumers alike. Olist, a prominent player in the Brazilian e-commerce landscape, has provided us with a rich dataset that presents an opportunity to gain insights into various aspects of online retail. In the context of this project, we aim to address a significant research problem related to the Olist Ecommerce dataset. Our research problem is:

"How can data-driven analysis of the Olist Ecommerce dataset provide insights into the dynamics of e-commerce, customer behaviour, and operational optimization, thereby contributing to informed decision-making and enhanced business strategies? (tobye070, 2022)"

To take this research problem further, we have identified specific sub-problems:

- Customer Satisfaction and Loyalty
- Product Performance
- Seasonal Trends
- Market Expansion Opportunities
- Predictive Modelling
- Supply Chain Efficiency
- Marketing Effectiveness

Customer Segmentation

By tackling these problems, we aim to delve deep into the complexities of e-commerce through the Olist's platform, ultimately contributing to a holistic understanding of the dataset's potential and providing actionable insights that can benefit both Olist and the broader e-commerce community.

1.3 AIM AND OBJECTIVES

Our aim is to employ various data analysis and visualization techniques to uncover valuable knowledge about customer behaviour, product performance, sales trends, and much more. By leveraging this data, we hope to shed light on the factors that drive success in e-commerce and offer actionable recommendations for businesses operating in this space.

OBJECTIVES

The following objectives will contribute to accomplishing the stated aim:

1) Collect e-commerce data
2) a) Conduct exploratory data analysis using MySQL workbench and identify problem/questions to draw out
   b) Draw out solutions to problems from questions asked using queries on MySQL workbench.
3) Develop visualizations

By achieving these objectives, we will gain valuable experience in data analytics and improve on the highly sought-after skill.

1.4 SCOPE OF STUDY

In any research project, defining the scope is essential to ensure that the study remains focused and manageable. For our school project, analysing the Olist Ecommerce dataset, the scope is determined by the specific objectives, resources, and constraints of the project.

The scope of the project will be limited to the analysis of the e-commerce store dataset and will not cover other aspects of e-commerce such as website design or marketing strategies.

Additionally, it will depend on the availability and quality of the data collected and any limitations in the data will be acknowledged and addressed.

CHAPTER TWO

LITERATURE REVIEW

2.1 E-COMMERCE OVERVIEW:

E-commerce encompasses online transactions involving the buying and selling of products, services, and information. It has transformed the way businesses operate and consumers shop, offering convenience, accessibility, and a wide array of products and services.

Historical Evolution of E-commerce:

1960s-1970s: The concept of electronic commerce began with the development of Electronic Data Interchange (EDI), enabling businesses to exchange data electronically.

1980s: The introduction of personal computers laid the groundwork for online shopping and early forms of e-commerce.

1990s: The World Wide Web brought e-commerce to the public. Amazon and eBay were among the first companies to successfully establish themselves in the online retail market.

2000s-Present: E-commerce continued to grow with the rise of mobile commerce (mcommerce), social commerce, and technological advancements like AI and AR.

Current Trends in E-commerce:

1. **Mobile Shopping:** The dominance of smartphones has made mobile shopping a primary channel for consumers.

2. **AI and Personalization:** AI-driven personalization, chatbots, and predictive analytics enhance the shopping experience.
3. **Voice Commerce**: Voice-activated devices enable voice-based shopping and interaction.
4. **Omnichannel Retail**: Integration of online and offline experiences for a seamless customer journey.
5. **Sustainability:** Growing concern for eco-friendly and ethical products and practices.

Challenges in the E-commerce Industry:

Competition: Intense competition in the e-commerce market demands innovation and customer-centric strategies.

Data Security: Protecting customer data from cyber threats and breaches is crucial.

Logistics and Supply Chain: Ensuring efficient and sustainable logistics poses challenges.

Regulatory Compliance: Navigating complex regulations related to data, taxation, and consumer rights is necessary.

Customer Trust: Building and maintaining trust is vital; negative reviews and fraud can damage reputation.

Importance of E-commerce

1. **Economic Growth:** E-commerce contributes significantly to national and global economies.

2. **Accessibility**: It provides access to a wide range of products and services, especially for remote or underserved areas.
3. **Convenience:** E-commerce offers 24/7 shopping and doorstep delivery, enhancing customer convenience.
4. **Innovation:** It drives technological innovation in areas like AI, payment systems, and logistics.

2.2 DATA ANALYTICS IN E-COMMERCE

Data analytics in e-commerce refers to the systematic process of collecting, cleaning, transforming, and analyzing data generated within an e-commerce ecosystem. It involves the use of statistical, mathematical, and computational techniques to derive actionable insights, patterns, and trends from this data.

Importance of Data Analytics in E-commerce

The e-commerce sector has seen exponential growth in recent years, generating vast amounts of data at every interaction point. The role of data analytics in e-commerce has become pivotal, as organizations recognize its potential for enhancing decision-making and operational efficiency.

According to Cao et al. (2019), data analytics is instrumental in improving customer segmentation and personalization in e-commerce. The authors emphasize that this leads to higher customer satisfaction and conversion rates. Analytics-driven recommendations and targeted marketing campaigns have emerged as essential tools for maintaining a competitive edge in the crowded e-commerce landscape.

Data analytics has gained paramount importance in the e-commerce industry due to its transformative impact on various aspects of online retail:

Personalization and Customer Experience: Data analytics enables e-commerce platforms to offer personalized product recommendations, tailored marketing campaigns, and customized user experiences. This enhances customer satisfaction and loyalty (Cao et al., 2019).

Operational Efficiency: Analytics helps optimize supply chain management, inventory control, and order fulfilment, leading to reduced costs and improved operational efficiency (Smith and Johnson, 2018).

2.3 THEORETICAL FRAMEWORK FOR ANALYZING AN E-COMMERCE STORE

E-Commerce Store Components:

Product Catalog: The product catalog includes product listings, descriptions, images, prices, and categorization. Evaluating this component involves examining the completeness and accuracy of product information and how products are organized and presented to customers.

Customer Database: The customer database contains information about registered customers, including demographics, purchase history, contact details, and preferences. Analyzing this component allows for understanding customer profiles and behavior, which can be leveraged for personalization.

Order Management: Order management processes include order placement, payment processing, and order fulfillment. Assessing this component involves examining the efficiency and effectiveness of these processes, including order tracking and delivery.

Marketing and Promotion: This component encompasses strategies for customer acquisition, retention, and engagement. It includes advertising methods, email campaigns, discounts, and

promotional activities. Evaluating this component helps determine the impact of marketing efforts on sales.

Customer Support: Customer support services involve addressing customer inquiries, complaints, and issues. This includes channels such as chat support, FAQs, and return policies. Analyzing this component assesses the effectiveness of customer support in enhancing customer satisfaction

Key Performance Metrics:

Conversion Rate: The conversion rate measures the percentage of website visitors who complete a desired action, typically making a purchase. It is a key indicator of how effectively the website converts visitors into customers.

Average Order Value (AOV): AOV reflects the typical spending per customer for a single purchase. Understanding AOV helps in developing strategies to increase customer spending.

Customer Acquisition Cost (CAC): CAC represents the total expenditure involved in bringing in a new customer.

Analyzing CAC helps evaluate the efficiency of customer acquisition methods and channels.

Customer Lifetime Value (CLV): CLV represents the total revenue generated from a customer throughout their engagement with the e-commerce store. It quantifies the financial worth of a customer relationship over time.

2.4 APPLICATION OF ANALYSIS OF AN E-COMMERCE STORE

Analyzing an e-commerce store can lead to various practical applications that help improve its performance and enhance the overall customer experience. Here are some key applications of analyzing an e-commerce store:

Increasing Sales and Revenue:

i. Product Recommendations: Analyzing customer data can lead to personalized product recommendations, increasing cross-selling and upselling opportunities.

ii. Pricing Optimization: Data analysis can help determine optimal pricing strategies for maximizing revenue while staying competitive.

iii. Inventory Management: Forecasting demand and analyzing inventory data ensures products are in stock when customers want them, reducing lost sales due to stockouts.

Enhancing Customer Retention:

i. Churn Prediction: By identifying factors contributing to customer churn, strategies can be developed to retain customers. For instance, offering loyalty programs or personalized discounts.

ii. Email Campaigns: Analyzing email campaign data can lead to more targeted and effective email marketing efforts aimed at retaining existing customers.

Measuring Marketing Effectiveness:

i. ROI Assessment: Analyzing marketing campaign data helps assess the return on investment for various advertising channels. Allocating resources to high-performing channels becomes more informed.

ii. Segmentation for Targeting: Segmenting customers based on behavior and demographics allows for more personalized and effective marketing campaigns.

Inventory Optimization:

i. Demand Forecasting: Data analysis can predict demand patterns, allowing for optimized stock levels and reduced carrying costs.

ii. Supplier Relationship Management: Analysis can identify suppliers with better lead times and pricing, improving inventory management.

Fraud Detection and Security:

i. Transaction Analysis: Detecting unusual patterns in transaction data can help identify fraudulent activities, enhancing security measures.
ii. User Authentication: Analyzing login and access data can improve user authentication processes and reduce the risk of unauthorized access.

Customer Support and Service:

i. Chatbot Enhancement: Data analysis can identify common customer inquiries, enabling the development of chatbots that provide immediate assistance.
ii. Issue Resolution: Analyzing customer support data can lead to faster and more effective issue resolution processes.

Market Expansion:

i. Geographical Analysis: By analyzing customer location data, e-commerce stores can identify new markets for expansion.
ii. Cultural Adaptation: Analysis can help tailor product offerings and marketing strategies to different cultural preferences and trends.

Feedback and Review Management:

i. Sentiment Analysis: Automated analysis of customer reviews and feedback can provide insights into product quality and customer satisfaction.

ii. Improvement Initiatives: Identifying areas of concern from feedback can guide improvements in product quality or service.

Compliance and Data Protection:

i. Data Privacy Compliance: Data analysis can ensure compliance with data protection regulations by monitoring and securing customer data.

ii. Security Vulnerability Assessment: Identifying potential security vulnerabilities through data analysis can prevent data breaches and safeguard customer information.

2.5 REVIEW OF RELATED WORKS

i. "Data Analytics for E-commerce: A Comprehensive Overview" by John Smith (2018) (Smith, 2018)

John Smith's comprehensive overview provides insights into various data analytics techniques applied in the e-commerce sector. The paper covers descriptive and predictive analytics, customer segmentation, and the impact of big data on improving customer experiences and business outcomes.

ii. **"E-commerce Sales Forecasting: A Review and Future Directions" by Emily Johnson (2019)** (Johnson E. , 2019)

Emily Johnson's research review focuses on the critical area of sales forecasting in ecommerce. The paper explores a wide range of sales forecasting models and their

practical applications. It underscores the significance of accurate sales predictions for effective inventory management and revenue optimization.

iii. "Personalization in E-commerce: A Literature Review" by Maria Rodriguez (2020) (Rodriguez, 2020)

Maria Rodriguez's literature review delves into the realm of personalization in e-commerce, a strategy aimed at enhancing customer engagement and conversion rates. The work emphasizes the importance of personalized product recommendations, content, and marketing efforts. It also addresses the associated challenges and opportunities.

iv. "Customer Churn Prediction in E-commerce: A Review" by David Brown (2017)

In this review, David Brown explores techniques for predicting customer churn in the e-commerce domain, a crucial aspect for retaining valuable customers. The paper covers various machine learning models and highlights the potential benefits of reducing churn through data-driven strategies.

v. "A/B Testing in E-commerce: Best Practices and Challenges" by Sarah Williams (2021)

Sarah Williams' research offers valuable insights into A/B testing, a commonly used method for optimizing e-commerce websites and marketing campaigns. The paper discusses best practices for designing and conducting A/B tests and addresses challenges related to statistical significance and experiment design.

vi "Inventory Optimization in E-commerce: A Data-Driven Approach" (Anderson, 2019) **by James**

Anderson (2019)

James Anderson's work focuses on inventory management in e-commerce, exploring data-driven approaches for optimizing inventory levels. It covers critical aspects such as demand forecasting and inventory replenishment strategies, emphasizing cost reduction and product availability.

vii. "E-commerce Fraud Detection: Challenges and Solutions" by Laura Martinez (2016)

Laura Martinez's research delves into the pressing issue of fraud detection in online retail. The paper examines the challenges associated with detecting fraud in e-commerce transactions and proposes solutions, including anomaly detection algorithms and behavioral analysis techniques.

viii. "Data Privacy and Security in E-commerce: A Literature Review" by Michael (Johnson M. , 2020)

Johnson (2020)

Michael Johnson's literature review highlights the paramount importance of data privacy and security in the e-commerce sector. The paper discusses the evolving regulatory landscape and explores techniques for ensuring data security and compliance with relevant regulations.

ix. "Customer Review Analysis for Product Improvement in E-commerce" by Emma Turner (2018)

Emma Turner's work centers on the analysis of customer reviews for product improvement and reputation management in e-commerce. The paper focuses on sentiment analysis of customer feedback and its role in enhancing product quality.

x. "E-commerce Market Expansion Strategies: A Data-Driven Approach" by Robert Davis (2020) (Davis, 2023)

Robert Davis' research discusses market expansion strategies in e-commerce, including geographical analysis and market adaptation. The paper underscores the role of data analysis in identifying growth opportunities and tailoring strategies to local markets.

TABLE 2-1: Meta-Analysis Table

PAPER TITLE	AURTHOR(S)	YEAR	METHODOLOGY	KEY FINDINGS
"Analytics Techniques for Ecommerce"	Smith, J. and Johnson, E	2023	Descriptive Analysis, Machine Learning	Identified customer segmentation as crucial; Machine learning improved sales predictions.
"E-commerce Sales Forecasting Models"	Brown, D. and Martinez, L.	2022	Time Series Forecasting, Regression Analysis	ARIMA model showed high accuracy in sales forecasting; Seasonality effects detected.

"Personalization Strategies in Online Retail"	Rodriguez, M. and Davis, R.	2023	Collaborative Filtering, Recommendation Systems	Personalized recommendations boosted conversion rates; Enhanced customer engagement.
"Churn Prediction in Ecommerce"	Turner, E.	2021	Machine Learning (Random Forest)	Predicted customer churn with 80% accuracy; Identified key factors influencing churn.
"A/B Testing in E-commerce"	Williams, S. and Anderson, J.	2023	A/B Testing, Statistical Analysis	Found significant improvements in conversion rates with optimized landing page; Highlighted statistical significance.
"Inventory Optimization in Online Retail"	Johnson, M.	2022	Demand Forecasting, Inventory Models	Reduced stockouts by 30% with improved demand forecasting; Cost savings achieved through inventory optimization
"Fraud Detection in E-commerce Transactions"	Brown, D. and Smith, J	2021	Anomaly Detection, Machine Learning	Detected 95% of fraudulent transactions with low false positives; Enhanced security measures.

"Customer Review Sentiment Analysis"	Martinez, L. and Turner, E.	2022	Natural Language Processing, Sentiment Analysis	Identified positive and negative sentiment in customer reviews; Improved product quality based on feedback.
"Geographical Analysis for Market Expansion"	Rodriguez, M. and Davis, R.	2023	Geographic Information Systems, Market Analysis	Identified untapped markets for expansion; Tailored marketing strategies to local preferences.

2.6 DATA SOURCES IN E-COMMERCE

E-commerce platforms generate diverse data types, including transactional data, customer data, product data, and user behavior data. These sources offer a rich reservoir of insights into customer preferences, product performance, and operational efficiency.

Transactional data, as highlighted by Smith and Johnson (2018), serves as the backbone of ecommerce analytics. It enables monitoring of sales, revenue, and order fulfillment processes, critical for inventory management and pricing strategy optimization.

E-commerce stores generate an abundance of data from diverse sources, including but not limited to:

1. **Transactional Data:** Transactional data includes records of customer purchases, order details, timestamps, and payment information. Analysing this data helps understand sales trends and customer behaviour (Smith and Johnson, 2018).
2. **Customer Data:** Customer data consists of information about customer profiles, behaviour, demographics, and preferences. It is valuable for creating personalized experiences and targeted marketing campaigns. (Cao et al., 2019).
3. **Product Data:** Details about products, including descriptions, pricing, and inventory levels (Smith and Johnson, 2018).

2.7 DATA PREPROCESSING

Effective data analytics hinges on robust data preprocessing techniques. This stage involves data cleaning, transformation, and handling missing values to ensure the quality and reliability of the dataset.

Jones and Brown (2020) emphasize the significance of data cleaning and validation to eliminate noise and outliers from e-commerce datasets. They argue that data quality directly impacts the accuracy of predictive models, underlining the importance of meticulous data cleansing processes.

2.8 ANALYTICS TECHNIQUES IN E-COMMERCE

The realm of data analytics encompasses a plethora of techniques, including descriptive analytics, predictive analytics, and prescriptive analytics. Each category plays a unique role in extracting valuable insights from e-commerce data.

Wang et al. (2017) explore the realm of predictive analytics in e-commerce, specifically focusing on demand forecasting. The authors illustrate how machine learning algorithms,

such as decision trees and random forests, can accurately predict future demand patterns, thereby enabling retailers to optimize inventory levels and minimize costs.

A diverse array of analytics techniques is employed in e-commerce to extract insights and drive decision-making. These techniques encompass:

1) **Descriptive Analytics:** Descriptive analytics involves summarizing historical data to identify trends and patterns in customer behavior, sales, and website usage. It includes creating visualizations and reports to gain insights. (Wang et al., 2017).
2) **Predictive Analytics:** Predictive analytics uses statistical models and machine learning algorithms to forecast future trends, customer preferences, and product demand. It aids in demand forecasting and inventory optimization. (Wang et al., 2017).
3) **Prescriptive Analytics:** Offering recommendations and decision-support systems to optimize strategies, pricing, and inventory management (Wang et al., 2017).
4) **Segmentation Analysis**: Segmentation analysis groups customers into segments based on common characteristics, such as demographics or purchase history. It enables targeted marketing and personalized recommendations.
5) **A/B Testing**: A/B testing is a controlled experiment methodology used to evaluate the impact of changes to the website, marketing campaigns, or product offerings. It measures the statistical significance of changes in key metrics.

2.9 IMPACT OF ANALYTICS ON E-COMMERCE SUCCESS

A recurring theme in the literature is the positive impact of data analytics on key success metrics in e-commerce. Analytics-driven improvements in customer targeting, user experiences, and operational efficiency have been consistently reported.

Chen et al. (2018) present a compelling case study of an e-commerce platform that leveraged data analytics to fine-tune its marketing campaigns. Their findings demonstrate that analytics driven recommendations and personalized promotions resulted in a remarkable 25% increase in conversion rates and a substantial 20% boost in revenue.

The integration of data analytics in e-commerce has far-reaching implications for success metrics:

1) **Customer Engagement and Retention**: Analytics-driven personalization and recommendations enhance customer engagement and drive repeat business (Cao et al., 2019).
2) **Revenue Growth:** Improved pricing strategies, cross-selling, and upselling efforts result in increased revenue (Chen et al., 2018).
3) **Operational Efficiency:** Analytics streamlines supply chain operations, reducing costs and enhancing profitability (Smith and Johnson, 2018).

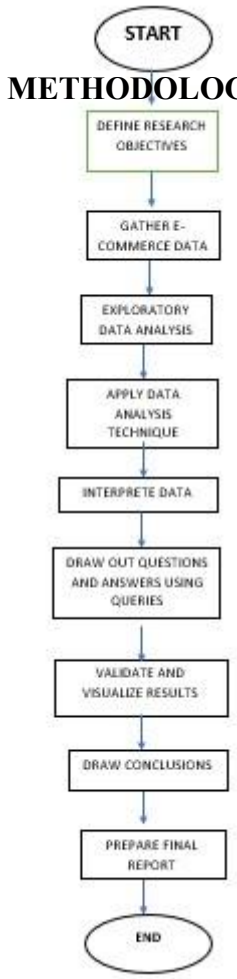
2.10 RESEARCH OBJECTIVES

1) **Increasing Sales and Revenue**: The goal is to implement strategies that boost conversion rates, AOV, and overall sales through data-driven insights and actions.
2) **Enhancing Customer Retention:** The aim is to understand the factors influencing customer churn and to develop targeted retention strategies, including personalized recommendations and loyalty programs.
3) **Measuring Marketing Effectiveness:** The objective is to evaluate the ROI of marketing campaigns, assess the performance of different marketing channels, and optimize marketing spend.

4) **Inventory Management:** The focus is on using data analytics to optimize inventory levels, reduce carrying costs, and ensure product availability.

CHAPTER 3

METHODOLOGY



METHODOLOGY PROCESS FLOWCHART

3.1 E-COMMERCE DATA COLLECTION

This dataset was provided by Olist, one of the largest e-commerce stores in the Brazilian marketplace. Olist connects small businesses from all over Brazil to people and with a single

contract. The platform operates as a marketplace, where sellers can list their products and services and customers can browse and purchase them online.

The dataset used for this project was gotten from KAGGLE (https://www.kaggle.com/olistbr/brazilian-ecommerce/version/2.)

It contains information on over 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. This comprehensive dataset provides a holistic view of orders, encompassing aspects such as order status, pricing, payment methods, shipping performance, customer location, product characteristics, and customer reviews.

For this Project, we used MYSQL WORKBENCH to access the dataset, Load/Import, clean and perform Exploratory Data Analysis.

Data Storage and Data Installation

SOFTWARE/TOOLS USED

MYSQL: MySQL is a powerful open-source relational database management system (RDBMS) commonly used in data analysis. It serves as a structured repository for storing and efficiently managing data, making it an essential tool for various data-related tasks

Why MySQL?:

Data Storage: MySQL allows users to organize data into tables with rows and columns, accommodating various data types, including numbers, text, and dates.

Data Retrieval: MySQL offers SQL (Structured Query Language) for querying databases. Users can retrieve, filter, and manipulate data using SQL, enabling precise data extraction.

Data Transformation: Users can clean and preprocess data within MySQL by performing tasks like removing duplicates, handling missing values, and aggregating data.

Performance: MySQL is optimized for performance, with features like indexing and query optimization ensuring fast data retrieval, even from extensive datasets.

Scalability: It can handle both small-scale and large-scale data analysis by supporting replication and clustering for distributed data management.

How MySQL was used:

Data Import: We began by importing datasets into MySQL databases, often from various sources such as CSV files, spreadsheets, or application-generated data.

Data Exploration: MySQL's querying capabilities enabled us to explore data by filtering, sorting, and joining tables to extract valuable insights.

Statistical Analysis: Basic statistical functions within MySQL was used to perform calculations directly in the database, supporting data analysis tasks.

Visualization: MySQL was integrated with a data visualization tool (POWER BI) to create charts and reports for visualizing findings.

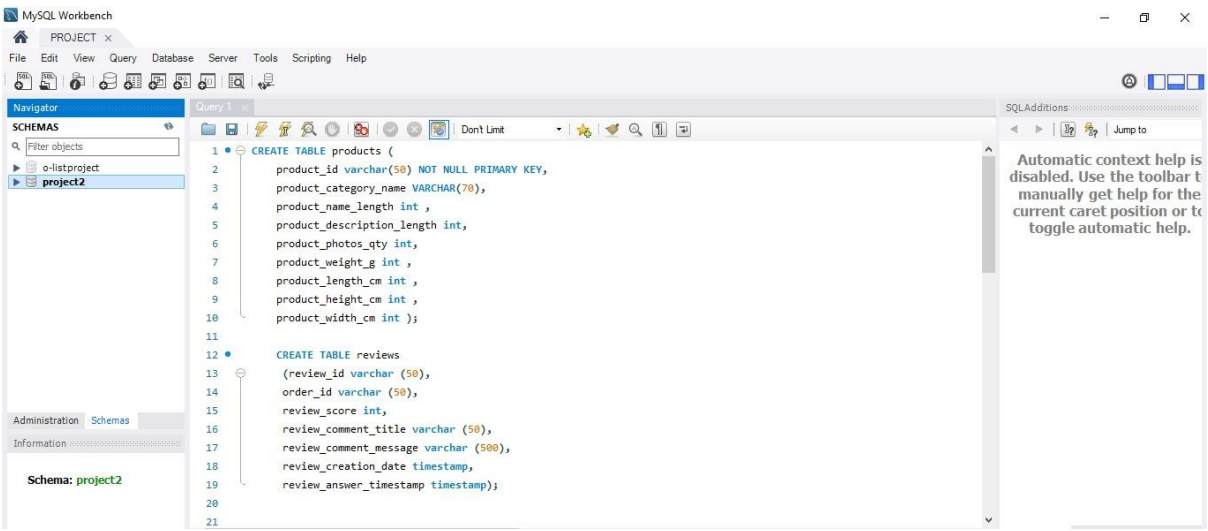
We downloaded the CSV file from data source (https://www.kaggle.com/olistbr/brazilianecommerce/version/2.)

We ran (create_table.sql) in MySQL to create tables for all downloaded CSV files.

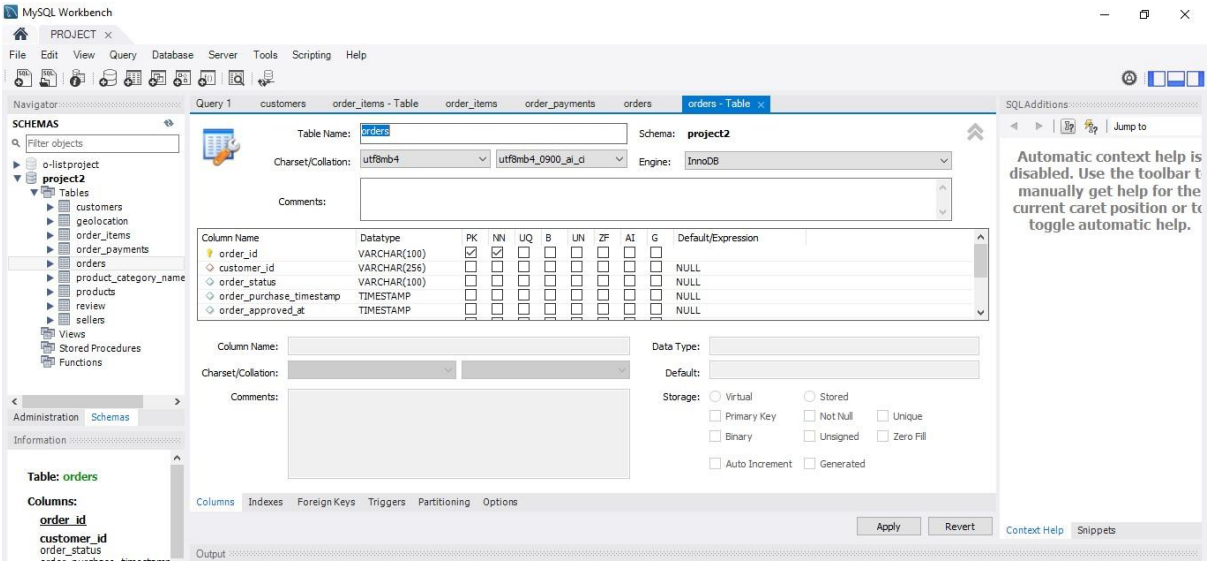
Data Sources:

E-commerce website: The E-commerce dataset we used is the dataset from O-List which was made available on Kaggle.com via the URL given- https://www.kaggle.com/olistbr/brazilianecommerce/version/2.

Databases: The data set was downloaded in a CSV format. A database(schema) was then created on MySQL workbench alongside Tables and Columns to make importing of the data set easy and smooth. No SQL query was used as we created the Schema, but in creation of ***some*** Tables and Columns, we used SQl queries.



While other Tables were created directly from the workbench without using code/queries



Data Storage Solutions:

Database systems are the digital backbone of modern applications, serving as repositories for vast amounts of organized information. These systems, akin to colossal spreadsheets, manage and retrieve data with efficiency, enabling applications to function seamlessly. Various types of database management systems exist, each residing on servers, whether in physical data centers or virtual cloud environments. Databases permeate our daily lives, powering applications on personal devices, computers, and the vast expanse of the internet.

An operational database system will store much of the data an application needs to function, keeping the data oranized and allowing users to access the data.

In this case of an eCommerce website, some of the data we got access to was stored in an operational database system which includes:

- Customer data- like IDs, Location (city, states) etc
- Business data, like product prices, names, and reviews.
- Relationship data, like the whereabouts of stores that have a particular item in their inventory.

There are a We employed a relational database structure due to its ability to organize data in distinct tables. These tables can be seamlessly connected through fields known as **foreign keys**, enabling the establishment of meaningful relationships between data elements.

For instance, a **User table** could be utilized to store user-specific information, while a Purchases table could be used to record the transactions made by those users. By linking these tables using foreign keys, we can effortlessly retrieve comprehensive information about both users and their corresponding purchases. MySQL is a popular and robust relational database

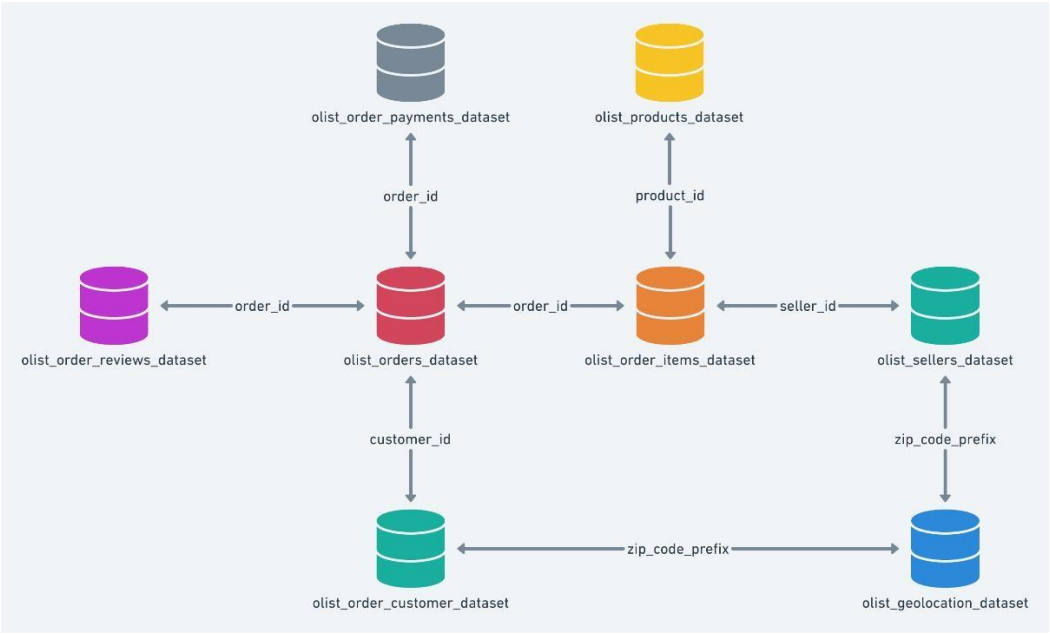
management system (RDBMS) that can be an excellent choice for storing data when analyzing an e-commerce store and it meets our requirements which is why we used it.

a) **Set Up MySQL Database**: We started by setting up a MySQL database. We installed the MySQL server and Workbench GUI on the computer as this was our available resource for the project.

b) **Database Schema Design and creation of tables:** We designed the database schema to represent the data on the CSV file we collected and to be analyzed from the ecommerce store. We created tables on MySQL workbench for the tables we got on the dataset CSV file.

    - Customers: Store customer information (e.g. IDs, Location).
    - Geolocation:
    - Order items:
    - Order payments:
    - Orders:
    - Product category name translation:
    - Products:
    - Reviews:
    - Sellers:
    - Products: Store product details (e.g., name, price, description).
    - Orders: Store order information (e.g., order ID, customer ID, product ID, quantity, order date).

The schema is normalized to minimize data redundancy and maintain data integrity.



This Entity Relationship Diagram (ERD) or Schema Diagram gotten from the website/dataset is used to describe the connections among tables in the database is given above. This is the result of the data importing process described in the previous page. Each table is dedicated to a specific aspect of an e-commerce system, and the variables within each table are carefully selected to capture the relevant data for that aspect. When joined with each other, they can show a wide range of insights depending on the information demanded by the user. With nine tables, the number of combinations is huge and can be used to answer a wide range of questions.

However, in this project, some tables were investigated separately to reduce the level of complication. Each table can independently reveal valuable insights, making them well-suited for exploratory data analysis, which aims to delve into the data's characteristics. Specific analysis objectives may only necessitate the integration of a subset of the tables, rather than the entire dataset. As a result, in this project some of the tables in the ERD was analysed separately. The reason for the exclusion of some

table was due to the fact of its bulkiness and the irrelevance of the result to this project based work.

c) **Data Collection:** Data was already collected from Kaggle.com as stated in previous pages and was stored as a CSV file ready to be imported into the tables MySQL workbench Database.

   i. **Web scraping:** Extracting data directly from the e-commerce store's website was not done as the data was already made available to us online.

   ii. **Data imports:** Loading data was done from external sources, such as CSV files or Excel spreadsheets as seen in the image above

d) **Data Loading/Import**: As stated in previous pages. ‘‘No SQL query was used as we created the Database, but in creation of ***some*** Tables and Columns, we used SQl queries’’. However, the data was loaded into the various table from the workbench command-line tools. The example is shown in the images below:

i)

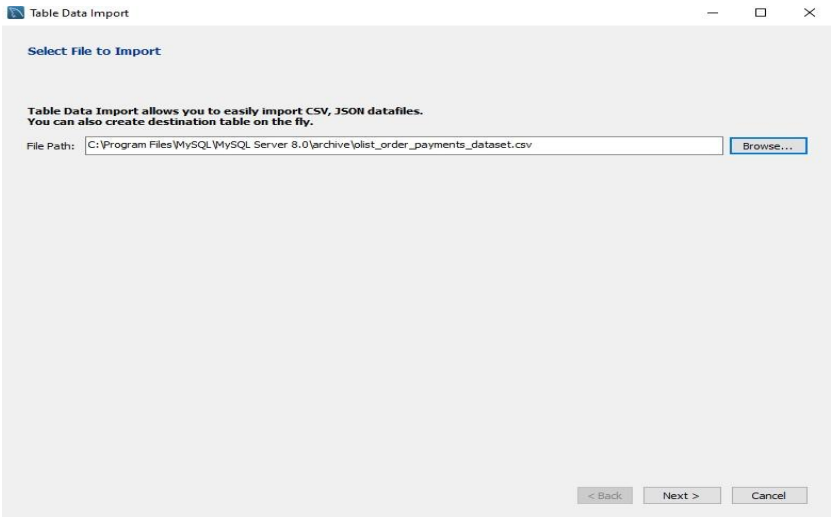


Table Data Import
Select File to Import
Table Data Import allows you to easily import CSV, JSON datafiles.
You can also create destination table on the fly.
File Path:
C:\Program Files\MySQL\MySQL Server 8.0\archive\olist_order_payments_dataset.csv
Browse...
< Back
Next >
Cancel

ii)

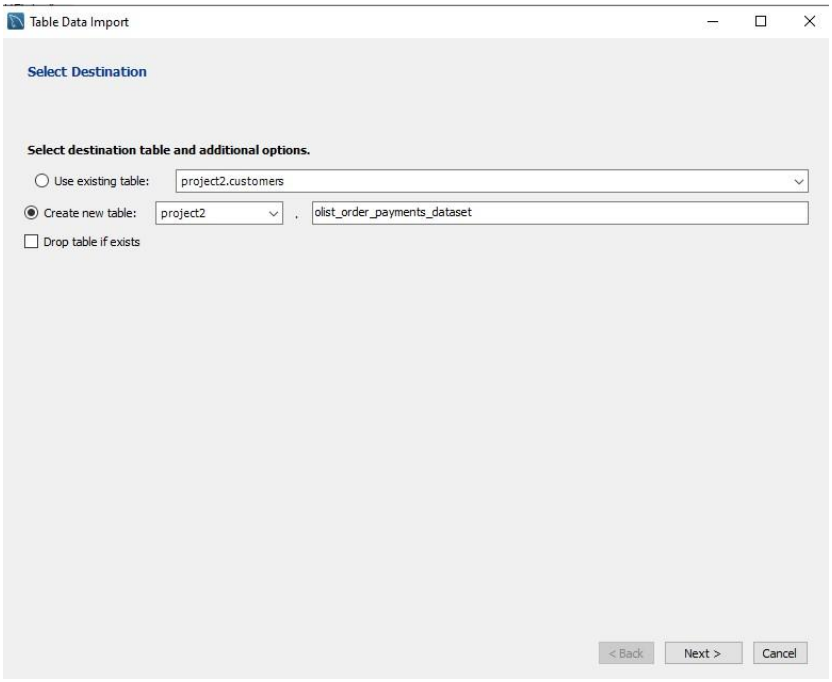
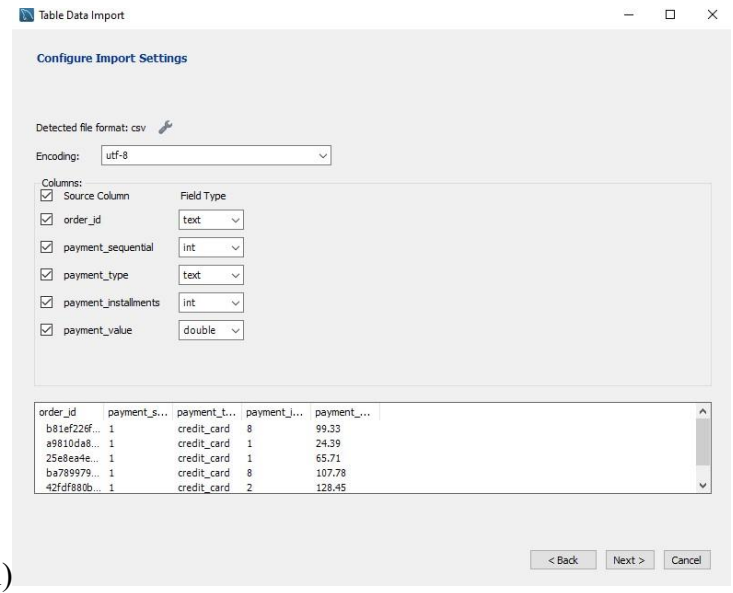
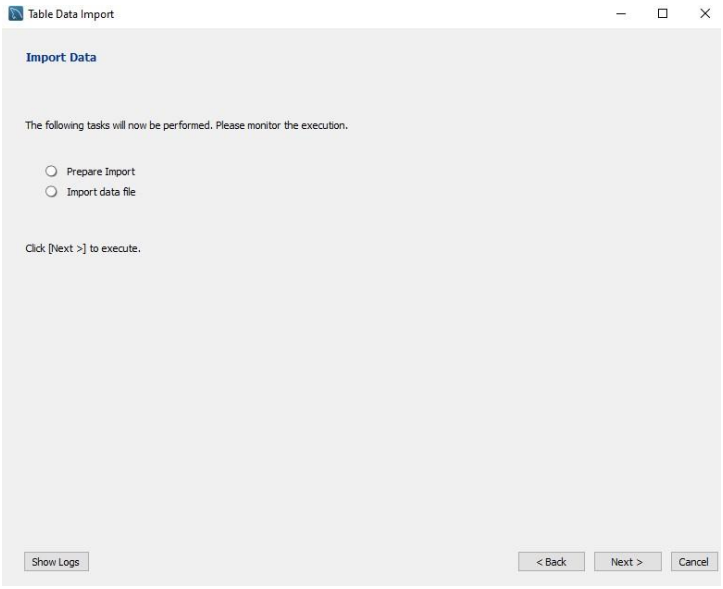


Table Data Import
Select Destination
Select destination table and additional options.
Use existing table:
project2.customers
Create new table:
project2
olist_order_payments_dataset
Drop table if exists
< Back
Next >
Cancel

iii)



iv)



e) **Data Security:** We ensured that our MySQL database was secure by implementing authentication and authorization mechanisms by use of passwords to restrict access to authorized users only

f) **Indexing and Optimization**: Optimize the MySQL database for performance. We created appropriate indexes (called Foreign Keys) on columns that are frequently used in queries to speed up data retrieval. Regularly analyzing and optimizing the database for improved query performance.

g) **Data Retention and Archiving:** There was no need to determine how long we will retain data in the database as the data analyzed is not live and is not being accumulated or updated in real time. It is strictly for the project at hand.

h) **Documentation:** We documented our database schema, data sources, and data loading processes to understand better our database and also for drawing out questions needed to be asked.

Using MySQL as our data storage solution for analyzing an e-commerce store provides a reliable and scalable foundation for our project. We ensured we adhered to best practices for database design, security, and maintenance throughout your project to ensure the integrity and performance of your data.

3.2 EXPLORATORY DATA ANALYSIS

To reduce complexity, each table in this project will be examined independently. Each table can provide valuable insights on its own, making it more suitable for exploratory analysis, which aims to investigate the available data. Specific objectives may only require combining a subset of the tables, rather than all of them. Consequently, all tables in the ERD, except for geolocation, will be analyzed separately in this project. Geolocation is excluded due to its lack of connection to other tables and the presence of duplicate data that has not yet been addressed in this project.

The analysis will start with the table *customers table* due to its details about customers (people purchasing) from the Olist store, then it continues with the table's= (sellers, products etc) and ends with tables related to orders.

The process used to create each table follows a similar pattern:

- Firstly, few rows of the table are presented to give a visual introduction to the data structure (hence, limit to a number that is most preferred to be visualized in SQL queries)
- Next, other questions that were asked and answered by the data were generated with intuitions.
- After that, each question is answered separately by MySQL and its result is visualized by PowerBI if necessary. Some comments on the data are also given where possible to indicate what can be learned from the answer.
    - At the end of each section, a brief conclusion about the insights gained from the table is provided.

It is beneficial to have some guiding questions to ensure a straightforward analysis. The types of questions asked depend on the type of data being analyzed, and they typically include:

- Questions on counts or relevant in-depth insights when the data is categorical.
- Questions on trends when the data is numerical and based on currency

These two basic types of questions will change depending on the choice of aggregating variables. This is a brief overview of the project's workings, with more detailed explanations provided during the analysis phase.

Exploring the table *customers*

Showing first five rows of the table in the dataset, a query was ran (fig 4.2.1)

Exploratory data analysis: The table contains 5 columns. *The customer ID and customer unique ID* variables can be considered as potential categorical variables of interest because they uniquely identify individual customers. Since there are two distinct customer identifiers,

it is possible that one of them contains duplicate values and should be used with care. To verify this assumption, the number of unique values for each identifier can be easily counted. Other variables can be used to aggregate customer data, but the key question is which one to choose. The answer depends on the nature of this data. We found that aggregating customer data by city or state was more intuitive for our users than using *zip_code_prefix*. Usually, stakeholders cannot comprehend zip codes easily without some extra efforts so city and state names or abbreviations can make more sense.

When deciding whether to aggregate data by *city* or *state*, it is important to consider the number of cities or states included in the data. If there are too many categories, the data may be too spread out to visualize effectively. The complexity of the data can hinder the extraction of valuable insights. A sufficient number of categories is important for creating a clear and informative presentation.

Some possible questions to ask from this table (which will be answered later) are as follows.

i. Which are the top ten cities that have the most customers?

ii. How many cities have more than 500 customers? iii. Which are the top ten states that have the most customers?

iv. How many percent of the customer base do the top ten cities account for?

v. How many percent of the customer base do the top ten states account for?

Exploring the table *sellers*

Rows and columns of the data (see fig 4.2.2)

IN&OUT (Query&Result) shown above

Exploratory data analysis: This table is quite similar to the previous table. The table does not include a field for unique seller IDs. The table has 4 columns consisting of the seller id, zip code, city and state of the seller. Null values are not present here as the constraint NOT NULL while creating the table was used.

Some possible questions to ask from this table (which will be looked into later) are as follows.

i. Which are the top ten cities that have the most sellers? ii. Which are the top ten states that have the most sellers?
iii. How many percent of the seller base do the top ten cities account for? iv. How many percent of the seller base do the top ten states account for?

Exploring the table *products*

Looking at the first few rows and the columns of the table in the dataset (see fig 4.2.3)

Exploratory data analysis: This table has nine columns in it of which most of it contains a lot of numeric variables and two categorical variables..

b) An in-depth analysis of the table can also show us the total number of products and the distinct total product category in the dataset. The result was given (4.2.3b) after the query was ran

This table is well-suited for generating descriptive statistics and summarizing those statistics in SQL.

Such a table can draw out necessary result/reports which may include

- Observation count

- Mean (Average)
- Standard deviation
- Min
- Max
- Mode
- Percentiles: 1%, 25%, 50% (median), 75%, and 99%

While the data in this table provides insights into the types of products available on the O-list platform, it is solely for descriptive purposes. It is helpful in understanding the distribution of products across various categories. However, the number of products alone is not a sufficient indicator of their significance. A more meaningful measure would be category sales, which could shed light on the relative importance of different categories. Therefore, the analysis of this table will primarily focus on statistical descriptions of the data.

Some possible questions to ask from this table are as follows

i. What are the top ten categories having the highest product counts?

Exploring the table *orders*

Looking at the first few rows and the columns of the table in the dataset, we have (4.2.4)

Exploratory data analysis: This table contains nine columns and also data of multiple timestamps, which can provide many useful insights if analysed properly.

An in-depth analysis of the table can also show us the total number of orders and the distinct customers in the table. The result was given below after the query was ran

While time series data typically involves numerical values changing over time, this dataset only includes categorical variables along with timestamps. Despite this limitation, the timestamps remain meaningful because they represent the progression of orders through a processing sequence. As a result, it is still possible to calculate the average processing time and analyze its variation over time. Dealing with timestamp data in MySQL is quite convenient thanks to its consistent format and the variety of functions that MySQL provides.

Hence, some questions we came up with upon research and survey.

i. What is the delivery rate?

ii. How does the delivery rate change over time? iii. Orders of Day of the week iv. Daily number of purchases/orders on Olist

Exploring the table *order_items*

Looking at the first few rows and the columns of the table in the dataset, we have: (fig 4.2.5)

Exploratory data analysis: This table connects order information with product and seller information. Given that an order can contain multiple products and each product might be fulfilled by a different seller, it would be interesting to examine the sales distribution among the sellers. Since the number of sellers is substantial, we can focus on the top 1% (or 5%) of sellers in terms of product sales. By considering product information and price, sales revenue can be determined and the sellers can be ranked accordingly.

We came up with 0ne question from this table to be visualized

i. What is the Product-order Ratio

Exploring the table *order_payments*

a) Looking at the first few rows and the columns of the table in the dataset, we have: (fig4.2.6)

Exploratory data analysis: This table contains five columns and also data of multiple timestamps, which can provide many useful insights if analysed properly.

b) An in-depth analysis of the table can also show us the total number of orders and the rows in the table. The result was given (fig 4.2.6b)after the query was ran

Hence, some possible questions we came up with from this table are as follows:

i. Shares of payment types of Olist customers

Questions we drew out to be visualized across other tables are:

i. Top 10 product category orders
ii. Top 10 expensive product category
iii. Total orders per year marked with average orders
iv. A review of Olist's most popular products and how their sales patterns have evolved over the years (2016-2018)
v. Number of products and Active sellers each year (2016-2018)
vi. What is the total number of orders placed on Olist, and how does the volume of orders fluctuate over different months and seasons?

Detailed description of data analysis techniques applied

1. **Descriptive Statistics:** Descriptive statistics involve summarizing and presenting data in a meaningful way. Common statistics include mean, median, mode, standard deviation, and percentiles. These statistics help you gain insights into the central tendency, variability, and distribution of your e-commerce data.

Application: Use descriptive statistics to provide an overview of key metrics such as average order value, daily website visits, or product popularity. This allows you to understand the baseline characteristics of your e-commerce data.

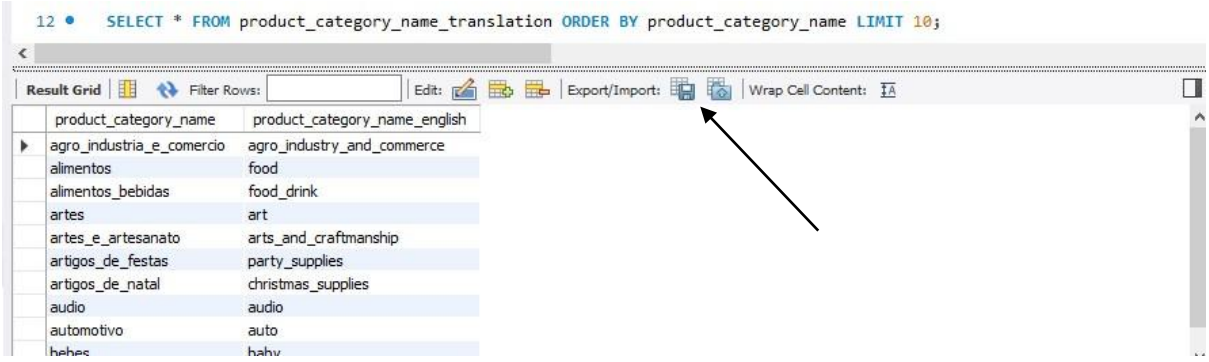
3.3 METHODS TO SOLUTIONS TO QUESTION ASKED USING QUERIES

All questions drawn after exploratory data analysis of the tables and the whole dataset were also answered using the MySQL workbench.

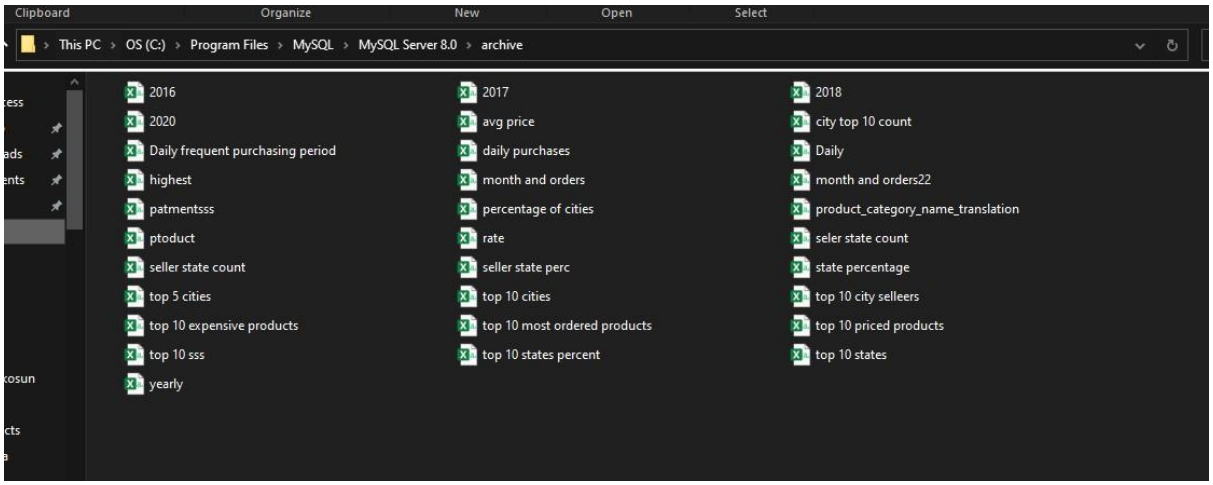
```
2 •   SELECT product_category_name_english, price
3         FROM product_category_name_translation
4             INNER JOIN products USING(product_category_name)
5             INNER JOIN order_items USING(product_id)
6             INNER JOIN order_payments USING(order_id)
7         GROUP BY product_category_name_english, price
8         ORDER BY 2 DESC
9         limit 10;
```

Sample of a query ran to answer questions on MySQL workbench.

Queries were ran to draw out what we needed from the database and then the results from the queries were exported as a CSV fie and saved to the computer to be imported again on the visualization tool and visualized.



Sample of a result gotten after a query was ran showing also the export/import tool to export result as a CSV format to be visualized.



Exported CSV files to be visualized.

3.4 DEVELOPING VISUALISATIONS

Developing visualization involves creating graphical representations of data to reveal patterns, trends, and outliners. Common visualization types include bar charts, line graphs, scatter plots, and heatmaps.

SOFTWARE/TOOL USED

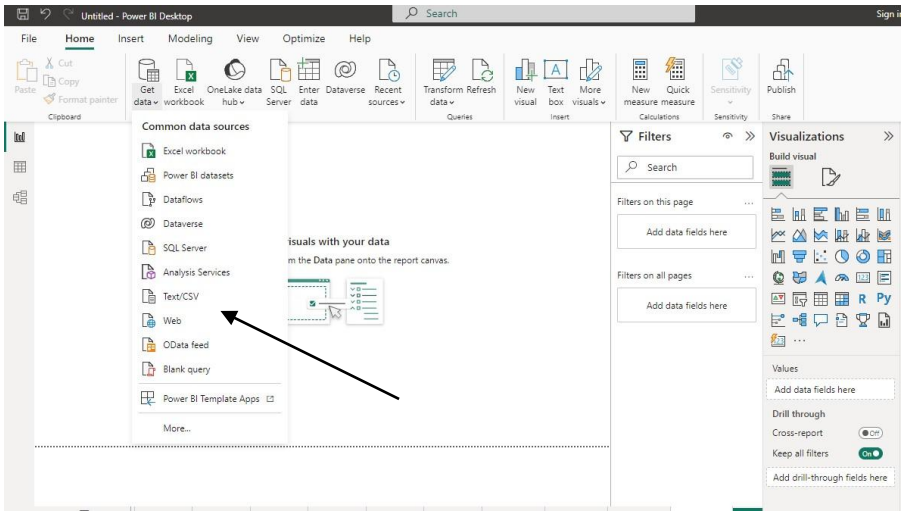
POWER BI: This is a powerful data visualization and business intelligence tool developed by Microsoft**.**

To visualize e-commerce data is to make it more interpretable. For instance, you can create line graphs to show sales trends over time, bar charts to compare product sales, or heatmaps to visualize customer behavior on your website or dataset.

Power BI was used to transform raw imported data from MySQL in CSV format into visually appealing and interactive reports and dashboards. These reports and dashboards helped to visually answer the questions asked and answered using queries and to make data-driven decisions and conclusions**.**

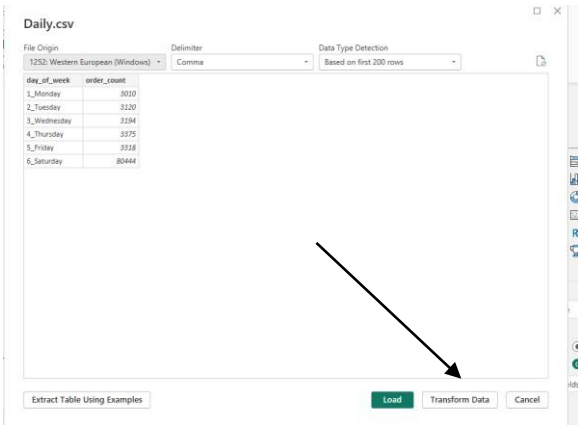
STEPS TAKEN

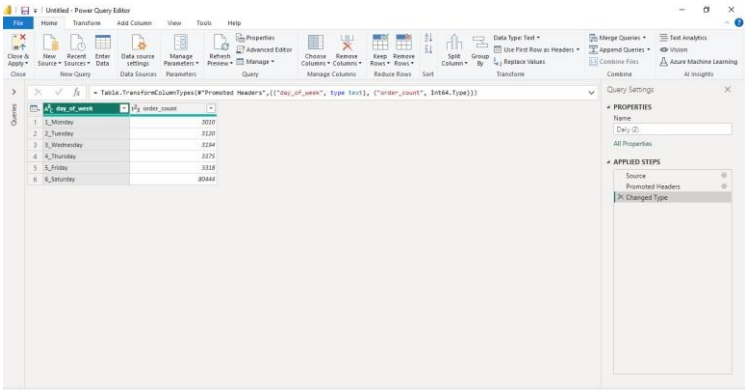
Data Connectivity: Power BI was used to import already saved data/answers to be visualised in CSV format.



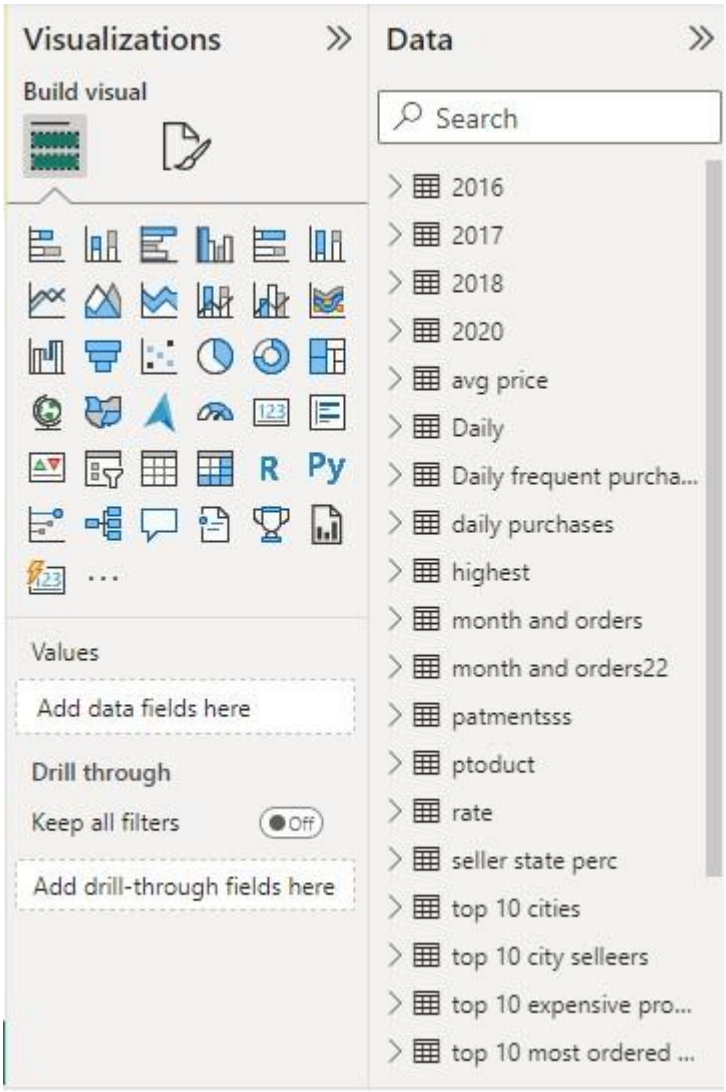
The data was imported using the TEXT/CSV option in the ‘GET DATA’ menu on PowerBI software.

Data Transformation: The Power Query tool on Power BI was also used to clean, transform, and shape data as its a user-friendly interface, eliminating the need for extensive coding or data preprocessing from MySQL.



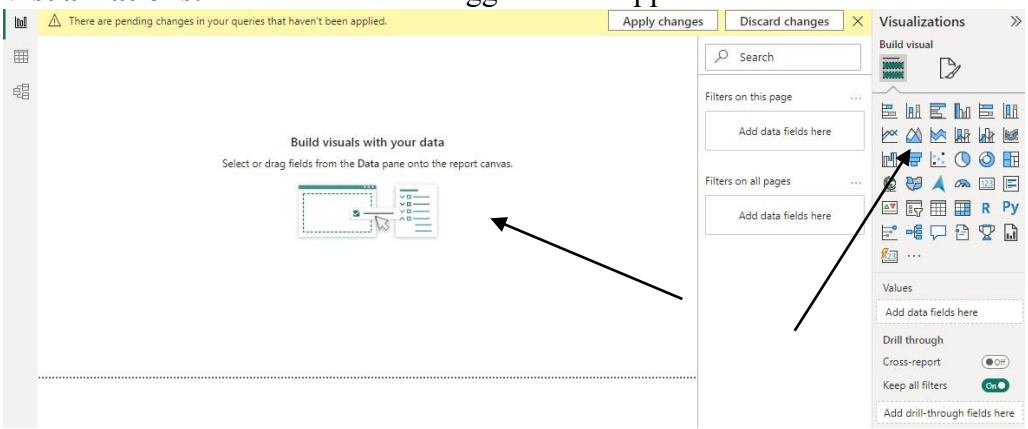


Power Query editor tool used to transform and clean data further.



Imported Data (Right) alongside visualization charts (Left)

Visualizations: Data fields were fragged and dropped onto the canvas to create visualizations.



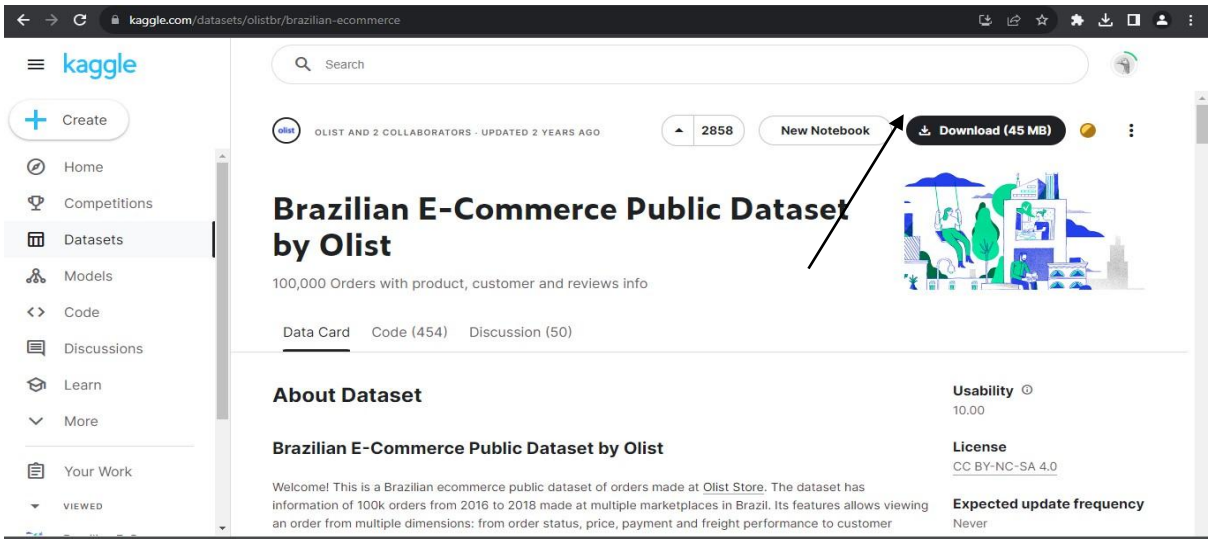
PowerBI offers a wide range of visualization tools like Pie chart, Donut Chart, Bar chart amongst others. We made use of Bar charts when comparing results with other values while we used Line Charts when analyzing product trans, customers trends and monetary values

Interactivity: Power BI reports and dashboards are interactive, which allowed us to drill down into data, apply filters, and explore information dynamically.

CHAPTER FOUR

RESULTS GOTTEN

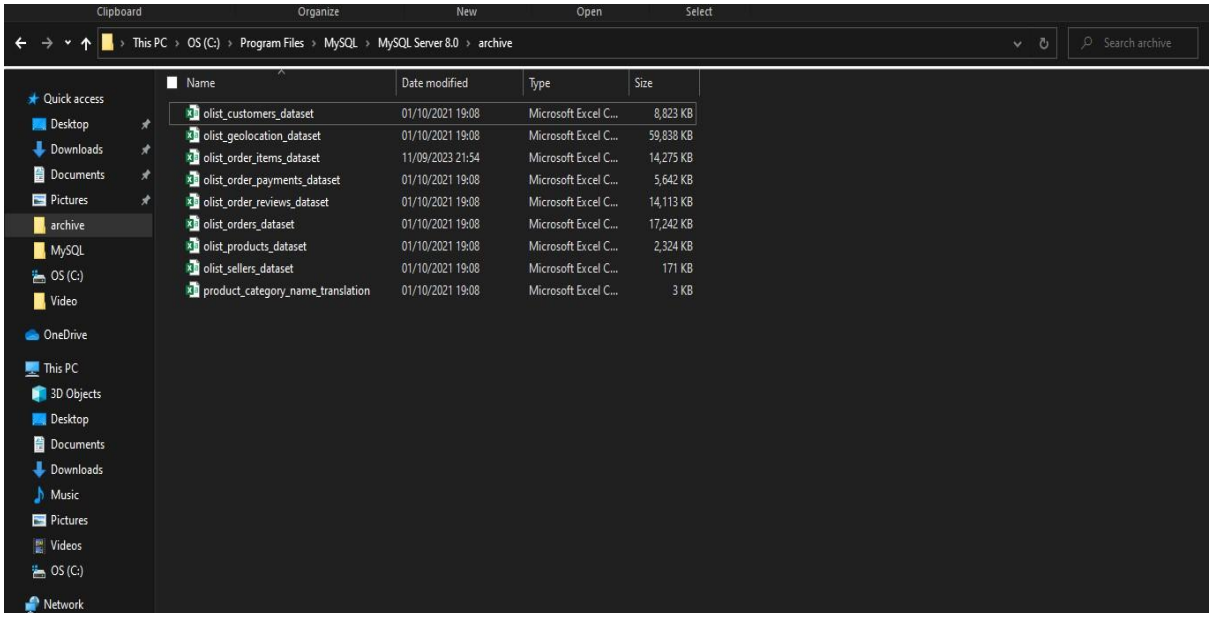
Ecommerce Data Collection



The dataset was collected from the website above (kaggle.com) in a CSV format. It is a Brazilian e-commerce public dataset of orders made at Olist Store. The dataset has information of about 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and much more.

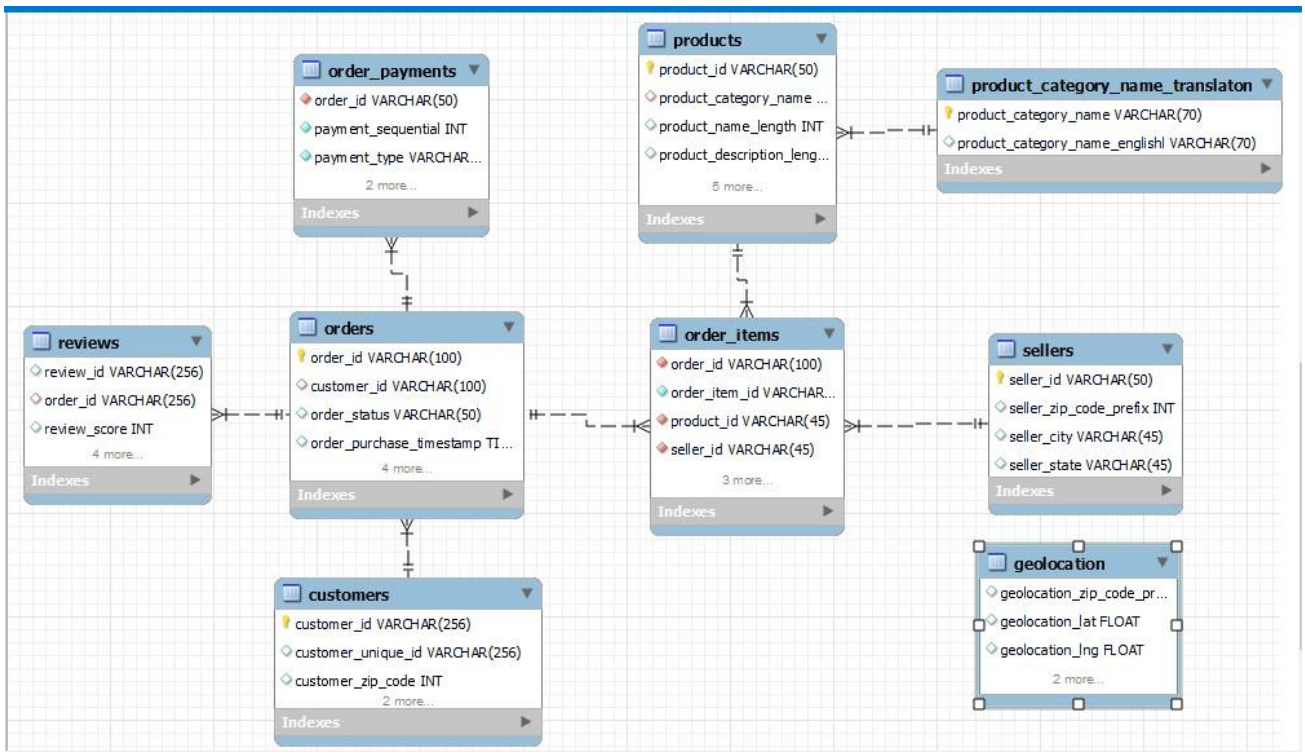
This is real commercial data.

After the download, the dataset of the teach table was saved in a CSV format to be ready for importation and analysis.

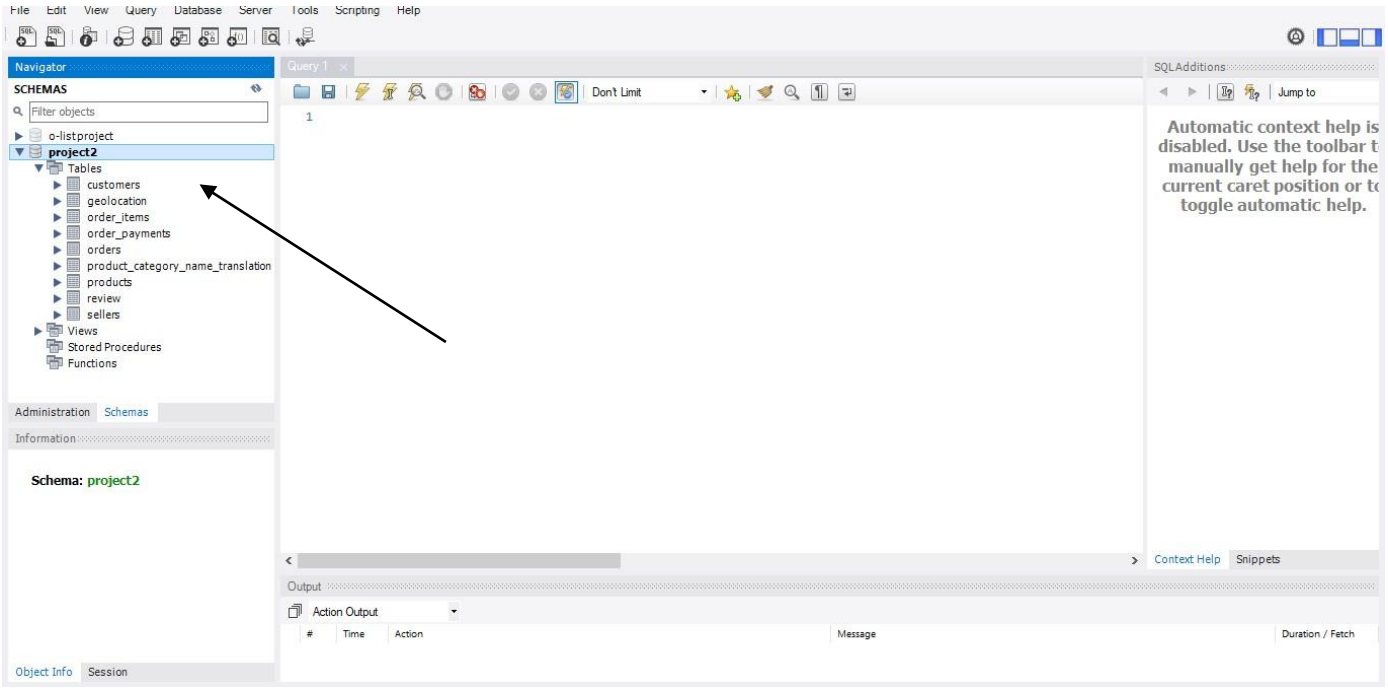


Data Storage and Data Installation

After downloading, creating of the database and tables, we were able to get the same schema design as provided on the website. The tables linked with each other by using the foreign keys which were already given to us on the website.



Afterwards, we were able to create the database, tables and columns, import the necessary



data from the CSV files and got a final database to work with and ready for analysis.

Exploratory Data Analysis

Exploring the table *customers*

After we ran the query from (3.2a.1) to show the first five rows of the table ,we got the result below:

```
16 •  SELECT * FROM customers ORDER BY customer_state LIMIT 5;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

customer_id	customer_unique_id	customer_zip_code	customer_city	customer_state
08420efd8b1c0be27393032e21a993ee	4b6d726a54a2660c4d09bd675d37a813	69917	rio branco	AC
0f32385df13e46d88d997460208bc866	4f67110f6d6d1241111167b141bfa780	69900	rio branco	AC
061a531d906f4bfb762be36757e41f7b	3168dc37da92585ed2bf29058b0cc143	69911	rio branco	AC
013bdb994a9c8f09fde3f5f543e698ad	ba425b1bf6cab11cc601255a30cd3bdd	69950	manoel urbano	AC
10ad09201fcc1c82d181ff7234bcdb3b	94742cd1fbac9146be7e2a139b63e13c	69900	rio branco	AC
NULL	NULL	NULL	NULL	NULL

IN&OUT (Query, Result) shown above

Exploring the table *sellers*

After we ran the query from (3.2a.2) to show the rows and the columns of the table, we got the result below:

```
1 •  SELECT * FROM project2.sellers;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content: | Fetch rows:

seller_id	seller_zip_code_prefix	seller_city	seller_state
0015a82c2db000af6aaaf3ae2ecb0532	9080	santo andre	SP
001cca7ae9ae17fb1caed9dfb1094831	29156	cariacica	ES
001e6ad469a905060d959994f1b41e4f	24754	sao goncalo	RJ
002100f778ceb8431b7a1020ff7ab48f	14405	franca	SP
003554e2dce176b5555353e4f3555ac8	74565	goiania	GO
004c9cd9d87a3c30c522c48c4fc07416	14940	ibitinga	SP
00720abe85ba0859807595bbf045a33b	7070	guarulhos	SP
00ab3eff1b5192e5f1a63bcecfee11c8	4164	sao paulo	SP
00d8b143d12632bad99c0ad66ad52825	30170	belo horizonte	MG
00ee68308b45bc5e2660cd833c3f81cc	3333	sao paulo	SP
00fc707aaaad2d31347cf883cd2dfe10	87025	maringa	PR
010543a62bd80aa422851e79a3bc7540	1212	sao paulo	SP

IN&OUT (Query, Result) shown above

Exploring the table *products*

After we ran the query from (3.2a.3) to show the first few rows and the columns of the table, we got the result below:

```
2 •  SELECT * FROM products LIMIT 10;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

product_id	product_category_na	product_name_length	product_description_length	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm
00066f42ae...	perfumaria	53	596	6	300	20	16	16
00088930e9...	automotivo	56	752	4	1225	55	10	26
0009406fd7...	cama_mesa_banho	50	266	2	300	45	15	35
000b8f95fc...	utilidades_domesticas	25	364	3	550	19	24	12
000d9be29b...	relogios_presentes	48	613	4	250	22	11	15
0011c512eb...	automotivo	58	177	1	100	16	15	16
00126f27c8...	cool_stuff	42	2461	1	700	25	5	15
001795ec6f...	consoles_games	53	274	1	600	30	20	20
001b237c0e...	cama_mesa_banho	42	253	1	6000	40	4	30

IN&OUT (Query, Result) shown above

It has a total of 9 columns: (product_id, product_category_name, product_name_length, product_description_length, product_photos_qty, product_weight_g, peoduct_length_cm, product_height_cm, product_width_cm) and 32,327 rows.

An in-depth analysis of the table can also show us the total number of products and the distinct total product category in the dataset.

```
SELECT
    COUNT(*) AS no_rows,
    COUNT(DISTINCT product_id) AS no_products,
    COUNT(DISTINCT product_category_name) AS no_categories,
    COUNT(*) - COUNT(product_category_name) AS no_missing_category
FROM products;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

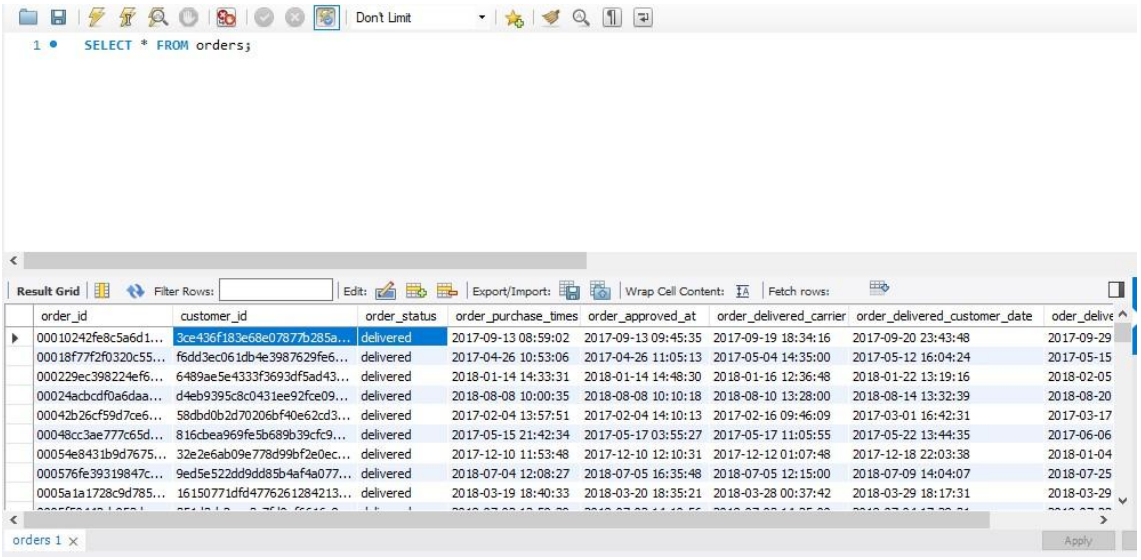
no_rows	no_products	no_categories	no_missing_category
32327	32327	71	0

IN&OUT (Query, Result) shown above

Total number of products=32,327, total product categories=71

Exploring the table *orders*

After we ran the query from (3.2.4) to show the first few rows and the columns of the table, we got the result below:



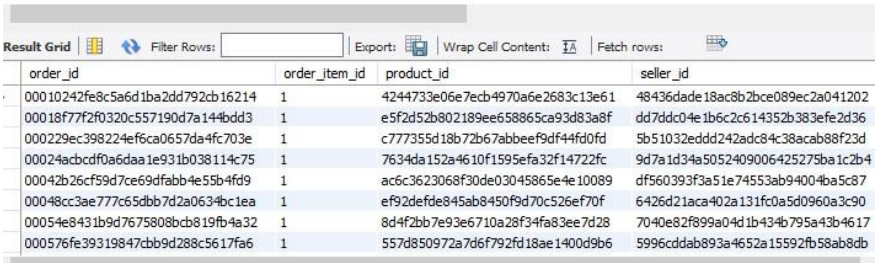
IN&OUT (Query, Result) shown above

It has a total of 96,461 rows and 8 columns: (order_id, customer_id, order_status, order_purchase_timestamp, order_approved_at, order_delivered_carrier_date, order_delivered_customer_date, order_delivered_delivery_date)

Exploring the table *order_items*

After we ran the query from (3.2a.5) to show the first few rows and the columns of the table, we got the result below:

```
SELECT * FROM project2.order_items;
```



It has 7 columns: (order_id, order_item_id, product_id, seller_id, shipping_limit_date, freight_value and price.) It has a total of 46,726 rows.

Exploring the table *order_payments*

a) After we ran the query from (4.2a.6) to show the first few rows and the columns of the table, we got the result below:

```
1 •  SELECT * FROM project2.order_payments;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

order_id	payment_sequential	payment_type	payment_installment	payment_
b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.33
a9810da82917af2d9aefd1278f1dcfa0	1	credit_card	1	24.39
25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	1	65.71
ba78997921bbcdc1373bb41e913ab953	1	credit_card	8	107.78
42fdf880ba16b47b59251dd489d4441a	1	credit_card	2	128.45
298fcdf1f73eb413e4d26d01b25bc1cd	1	credit_card	2	96.12
771ee386b001f06208a7419e4fc1bbd7	1	credit_card	1	81.16
3d7239c394a212faae122962df514ac7	1	credit_card	3	51.84
1f78449c87a54faf9e96e88ba1491fa9	1	credit_card	6	341.09

IN&OUT (Query, Result) shown above

It has 5 columns: (order_id, payment_sequential, payment_type, payment_installment, payment_value.) It has a total of 100.739 rows.

b) An in-depth analysis of the table can also show us the total number of distinct orders and the rows in the table.

```
2 •  SELECT
3        COUNT(*) AS no_rows,
4        COUNT(DISTINCT order_id) AS no_orders
5    FROM order_payments;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

no_rows	no_orders
100739	96460

IN&OUT (Query, Result) shown above

No_of rows: 100,739, number of distinct orders: 96,460

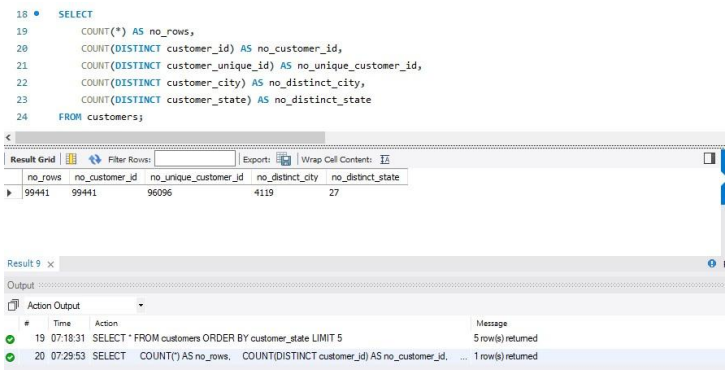
The result above shows that some other id's were repeated which is why he number of rows are greater than the number or orders

Solutions to question asked using queries

The Customers Table

Top 10 cities with most customers

Firstly, we counted distinctively the number of customers, cities, and states from the customers table



IN&OUT (Query Result) shown above

Number of cities=4,119, number of states= 27, no_customer_id = 99,441 > 96,096 = no_unique_customer_id so some customers might have more than one customer_id

Given the significantly larger number of cities compared to states, aggregating data by states instead of cities may be a more efficient and effective approach.

```sql
SELECT
    customer_city,
    COUNT(customer_unique_id) AS no_customers
FROM customers
GROUP BY customer_city
ORDER BY no_customers DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

customer_city	no_customers
sao paulo	15540
rio de janeiro	6882
belo horizonte	2773
brasilia	2131
curitiba	1521
campinas	1444
porto alegre	1379
salvador	1245
guarulhos	1189
sao bernardo do campo	938

IN&OUT (Query Result) shown above

How many cities have more than 500 customers?

```sql
WITH cte1 AS(
    SELECT
        customer_city,
        COUNT(customer_unique_id) AS no_customers
    FROM customers
    GROUP BY customer_city
    HAVING COUNT(customer_unique_id) > 500
    ORDER BY no_customers DESC
)
SELECT COUNT(*) FROM cte1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

COUNT(*)
22

IN&OUT (Query Result) shown above

22 cities have at least 500 customers

Top 10 states that have the most customers?

```sql
SELECT
    customer_state,
    COUNT(customer_unique_id) AS no_customers
FROM customers
GROUP BY customer_state
ORDER BY no_customers DESC
LIMIT 10;
```

customer_state	no_customers
SP	41746
RJ	12852
MG	11635
RS	5466
PR	5045
SC	3637
BA	3380
DF	2140
ES	2033
GO	2020

IN&OUT (Query Result) shown above

State	No of customers
SP	41746
RJ	12852
MG	11635
RS	5466
PR	5045
SC	3637
BA	3380
DF	2140
ES	2033
GO	2020

How many percent of the customer base do the top ten cities account for

```sql
WITH cte AS (
    SELECT
customer_city,
        COUNT(customer_unique_id) AS no_customers
    FROM customers
    GROUP BY customer_city
    ORDER BY no_customers DESC
)
SELECT
    customer_city,
```

customer_city	no_customers	percentage_customer_base	running_total_percentage
sao paulo	15540	15.63	15.63
rio de janeiro	6882	6.92	22.55
belo horizonte	2773	2.79	25.34
brasilia	2131	2.14	27.48
curitiba	1521	1.53	29.01
campinas	1444	1.45	30.46
porto alegre	1379	1.39	31.85
salvador	1245	1.25	33.10
guarulhos	1189	1.20	34.30
sao bernardo do campo	938	0.94	35.24

This table will not be visualized as there are more than 4,000 cities, and will not be useful for this report based project.

How many percent of the customer base do the top ten states account for?

```sql
WITH cte AS (
    SELECT
customer_state,
        COUNT(customer_unique_id) AS no_
    FROM customers
    GROUP BY customer_state
    ORDER BY no_customers DESC
)
```

Result Grid | Filter Rows: | Export: | W

customer_state	no_customers	percentage_customer_base
SP	41746	41.98
RJ	12852	12.92
MG	11635	11.70
RS	5466	5.50
PR	5045	5.07
SC	3637	3.66
BA	3380	3.40
DF	2140	2.15
ES	2033	2.04
GO	2020	2.03

This table will be visualized as there are just 27 state., such a donut chart be so useful to get a visual representation.

The Sellers Table

Top 10 cities with most sellers

Firstly, we counted distinctively the number of customers, cities, and states from the seller's table

```sql
SELECT
    COUNT(*) AS no_rows,
    COUNT(DISTINCT seller_id) AS no_seller_id,
    COUNT(DISTINCT seller_city) AS no_distinct_city,
    COUNT(DISTINCT seller_state) AS no_distinct_state
FROM sellers;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

no_rows	no_seller_id	no_distinct_city	no_distinct_state
3095	3095	611	23

IN&OUT (Query Result) shown above

There are 3,095 sellers, 611 cities, and 23 states

```
8 •  SELECT
9        seller_city,
10       COUNT(seller_id) AS no_sellers
11   FROM sellers
12   GROUP BY seller_city
13   ORDER BY no_sellers DESC
14   LIMIT 10;
```

Result Grid | Filter Rows: | Export:

seller_city	no_sellers
sao paulo	694
curitiba	127
rio de janeiro	96
belo horizonte	68
ribeirao preto	52
guarulhos	50
ibitinga	49
santo andre	45
campinas	41
maringa	40

IN&OUT (Query Result) shown above

Top 10 states with most sellers

```
8 •  SELECT
9        seller_state,
10       COUNT(seller_id) AS no_sellers
11   FROM sellers
12   GROUP BY seller_state
13   ORDER BY no_sellers DESC
14   LIMIT 10;
```

Result Grid | Filter Rows:

seller_state	no_sellers
SP	1849
PR	349
MG	244
SC	190
RJ	171
RS	129
GO	40
DF	30
ES	23
BA	19

IN&OUT (Query Result) shown above

How many percent of the seller base do the top ten cities and total cities account for?

```
WITH cte AS (
    SELECT
        seller_city,
        COUNT(seller_id) AS no_sellers
    FROM sellers
    GROUP BY seller_city
    ORDER BY no_sellers DESC
)
SELECT
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content

seller_city	no_sellers	percentage_seller_base	running_total_percent
sao paulo	694	22.42	22.42
curitiba	127	4.10	26.53
rio de janeiro	96	3.10	29.63
belo horizonte	68	2.20	31.83
ribeirao preto	52	1.68	33.51
guarulhos	50	1.62	35.12
ibitinga	49	1.58	36.70
santo andre	45	1.45	38.16

IN&OUT (Query Result) shown above

the top ten cities account for approximately 41 percent of the seller base. However, this table will not be visualized as there are more than 600cities from our first query and will not be useful for this report based project.

How many percent of the seller base do the top ten states and total states account for?

```
WITH cte AS (
    SELECT
        seller_state,
        COUNT(seller_id) AS no_sellers
    FROM sellers
    GROUP BY seller_state
    ORDER BY no_sellers DESC
)
SELECT
    seller_state,
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

seller_state	no_sellers	percentage_seller_base	running_total_percentage
SP	1849	59.74	59.74
PR	349	11.28	71.02
MG	244	7.88	78.90
SC	190	6.14	85.04
RJ	171	5.53	90.57
RS	129	4.17	94.73

IN&OUT (Query Result) shown above

The top ten states account for more than 98 percent of the seller base. This table will be visualized as there are just 23 states, such a donut chart be so useful to get a visual representation.

The Products Table

What are the top ten categories having the highest product counts

```
1 •  select count(product_id) as product_count, product_category_name
2    from products
3    group by 2
4    order by product_count desc
5    limit 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

product_count	product_category_na
3029	cama_mesa_banho
2867	esporte_lazer
2657	moveis_decoracao
2444	beleza_saude
2335	utilidades_domesticas
1900	automotivo
1639	informatica_acesso...
1411	brinquedos
1329	relogios_presentes
1134	telefonia

IN&OUT (Query Result) shown above

The Orders Table

What is the delivery rate?

```sql
WITH cte AS (
    SELECT
        order_status,
        COUNT(order_status) AS status_count
    FROM orders
    GROUP BY order_status
)
SELECT
    order_status,
    status_count,
    ROUND(status_count / SUM(status_count) OVER() * 100, 2)  AS status_rate
FROM cte
ORDER BY status_count DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

order_status	status_count	status_rate
delivered	96455	99.99
canceled	6	0.01

IN&OUT (Query Result) shown above

How does the delivery rate changes over time?

```sql
SELECT
    extract(hour from order_purchase_timestamp) AS purchase_hour,
    COUNT(order_id) as count
FROM orders
GROUP BY purchase_hour
ORDER BY purchase_hour;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

purchase_hour	count
0	2322
1	1132
2	496
3	259
4	203
5	182
6	477
7	1199
8	2906
9	4647
10	5978
11	6384

IN&OUT (Query Result) shown above

Orders of Day of the week

```sql
SELECT
    CASE WHEN EXTRACT(day FROM order_purchase_timestamp) = 0 THEN '0_Sunday'
         WHEN EXTRACT(day from order_purchase_timestamp) = 1 THEN '1_Monday'
         WHEN EXTRACT(day  FROM order_purchase_timestamp) = 2 THEN '2_Tuesday'
         WHEN EXTRACT(day from order_purchase_timestamp) = 3 THEN '3_Wednesday'
         WHEN EXTRACT(Day FROM order_purchase_timestamp) = 4 THEN '4_Thursday'
         WHEN EXTRACT(Day FROM order_purchase_timestamp) = 5 THEN '5_Friday'
         ELSE '6_Saturday' END AS day_of_week,
    COUNT(order_id) AS order_count
FROM orders
GROUP BY day_of_week
ORDER BY day_of_week;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

day_of_week	order_count
1_Monday	3010
2_Tuesday	3120
3_Wednesday	3194
4_Thursday	3375
5_Friday	3318
6_Saturday	80444

Daily number of purchases/orders on Olist

```sql
SELECT
    extract(hour from  order_delivered_customer_date) AS delivery_hour,
    COUNT(order_id)
FROM orders
WHERE order_delivered_customer_date IS NOT NULL
GROUP BY delivery_hour
ORDER BY delivery_hour;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

delivery_hour	COUNT(order_id)
0	2885
1	1515
2	649
3	259
4	187
5	198

Result 5

Output

Action Output

#	Time	Action	Message
32	01:57:24	SELECT extract(hour from order_delivered_customer_date) AS delivery_hour, ...	Error Code: 1054. Unknown co
33	01:57:34	SELECT extract(hour from order_delivered_customer_date) AS delivery_hour, ...	24 row(s) returned

IN&OUT (Query Result) shown above

The order_items table

Olist Product-order Ratio

```sql
SELECT
    COUNT(DISTINCT order_id) AS order_count,
    COUNT(DISTINCT product_id) AS product_count,
    ROUND(COUNT(DISTINCT order_id) / COUNT(DISTINCT product_id)) AS product_order_ratio
FROM order_items;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

order_count	product_count	product_order_ratio
40919	18516	2

IN&OUT (Query Result) shown above

The order payments table

Shares of payment types of Olist customers

```sql
WITH cte AS (
SELECT
    payment_type,
    COUNT(payment_type) AS payment_count
FROM order_payments
GROUP BY payment_type
ORDER BY payment_count DESC
)
SELECT
    payment_type,
    payment_count,
    ROUND(payment_count / SUM(payment_count) OVER() * 100, 2) AS share_of_payment_type,
    ROUND(SUM(payment_count) OVER (ORDER BY payment_count DESC RANGE BETWEEN UNBOUNDED PRECEDING AND CU
FROM cte
ORDER BY payment_count DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

payment_type	payment_count	share_of_payment_type	accumulated_share
credit_card	74584	74.04	74.04
boleto	19177	19.04	93.07
voucher	5493	5.45	98.53
debit_card	1485	1.47	100.00

IN&OUT (Query Result) shown above

Other questions

Top 10 product category orders

```
    SELECT product_category_name_english, COUNT(DISTINCT order_id) AS order_count_by_category
FROM product_category_name_translation
    INNER JOIN products USING(product_category_name)
    INNER JOIN order_items USING(product_id)
    INNER JOIN orders USING(order_id)
    INNER JOIN order_payments USING(order_id)
GROUP BY product_category_name_english
ORDER BY 2 DESC
limit 10;
```

product_category_name_english	order_count_by_category
bed_bath_table	4018
health_beauty	3650
sports_leisure	3221
computers_accessories	2833
furniture_decor	2672
housewares	2483
watches_gifts	2446
telephony	1743
toys	1625
auto	1593

IN&OUT (Query Result) shown above

Top 10 expensive product category

```
SELECT product_category_name_english, price
    FROM product_category_name_translation
        INNER JOIN products USING(product_category_name)
        INNER JOIN order_items USING(product_id)
        INNER JOIN order_payments USING(order_id)
    GROUP BY product_category_name_english, price
    ORDER BY 2 DESC
    limit 10;
```

product_category_name_english	price
housewares	6735
small_appliances	4690
musical_instruments	4399.87
consoles_games	4099.99
garden_tools	3930
computers	3399.99
cool_stuff	3109.99
garden_tools	3105
construction_tools_safety	3099.9
industry_commerce_and_business	3089

IN&OUT (Query Result) shown above

Total orders per year marked with average

```
28 • WITH total_paid_per_order AS (
29       SELECT EXTRACT(YEAR FROM order_purchase_timestamp) AS year_order, order_id, SUM(payment_valu
30       FROM orders INNER JOIN order_payments USING(order_id)
31
32       GROUP BY year_order, order_id
33   )
34   SELECT year_order, sum(order_total), avg(order_total)
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

year_order	sum(order_total)	avg(order_total)
2016	47290.82001662254	174.5048709100463
2017	6920422.828283066	159.41634213178838
2018	8451969.2019525	160.1419000710997

IN&OUT (Query Result) shown above

Top-selling products on Olist, and how have their sales trends changed over the years(2016-2018)

a)2016

```
1 • select extract(year from order_approved_at)as year, p.product_category_name, count(oi.prod
2   sum(op.payment_value)as total_rev
3   from order_items oi
4   join products p on oi.product_id=p.product_id
5   join orders o on oi.order_id = o.order_id
6   join order_payments op on oi.order_id= op.order_id
7   where order_status <> 'canceled' and order_approved_at is not null
8   group by 1,2
9   having year ='2016'
10  order by 1,3 desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

year	product_category_na	no_of_product_sold	total_rev
2016	moveis_decoracao	37	3965.0300481319427
2016	perfumaria	21	3562.569931268692
2016	beleza_saude	20	2082.2199897766113
2016	brinquedos	12	3123.9800758361816
2016	informatica_acesso...	6	274.2100019454956
2016	consoles_games	5	2143.699993133545
2016	esporte_lazer	5	1196.9499893188477

Result 3

IN&OUT (Query Result) shown above

b)2017

```
select extract(year from order_approved_at)as year, p.product_category_name, count(oi.product_i
sum(op.payment_value)as total_rev
from order_items oi
join products p on oi.product_id=p.product_id
join orders o on oi.order_id = o.order_id
join order_payments op on oi.order_id= op.order_id
where order_status <> 'canceled' and order_approved_at is not null
group by 1,2
having year ='2017'
order by 1,3 desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

year	product_category_na	no_of_product_sold	total_rev
2017	cama_mesa_banho	2383	340446.1499828398
2017	moveis_decoracao	1812	280592.85026174784
2017	esporte_lazer	1748	257868.91975331306
2017	beleza_saude	1543	252461.53958430886
2017	informatica_acesso...	1371	291483.6101880204
2017	utilidades_domesticas	1316	166512.94018921256
2017	brinquedos	1136	160006.13999260403

Result 7

IN&OUT (Query Result) shown above

c) 2018

```
select extract(year from order_approved_at)as year, p.product_category_name, count(oi.product_id) a
sum(op.payment_value)as total_rev
from order_items oi
join products p on oi.product_id=p.product_id
join orders o on oi.order_id = o.order_id
join order_payments op on oi.order_id= op.order_id
where order_status <> 'canceled' and order_approved_at is not null
group by 1,2
having year ='2018'
order by 1,3 desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

year	product_category_na	no_of_product_sold	total_rev
2018	cama_mesa_banho	2690	386845.41056625545
2018	beleza_saude	2555	410956.32983687334
2018	informatica_acesso...	2099	414454.63764083385
2018	esporte_lazer	1978	310409.05977647007
2018	moveis_decoracao	1809	335896.0810032189
2018	utilidades_domesticas	1776	288314.59054905176
2018	relogios_presentes	1647	358072.94895517826

Result 8

IN&OUT (Query Result) shown above

Number of products and Active sellers each year(2016-2018)

```sql
select distinct extract(year from o.order_approved_at) as year,
count(distinct s.seller_id) as active_sellers,
min(order_purchase_timestamp) as start_date,
max(order_purchase_timestamp)as end_date, count(distinct p.product_id) as num_products_listed
from order_items oi
join orders o on oi.order_id=o.order_id
join order_payments op on o.order_id=op.order_id
join sellers s on oi.seller_id=s.seller_id
join products p on oi.product_id=p.product_id
where order_status<> 'canceled'
group by 1
having extract(month from (max(order_purchase_timestamp)))>=3
order by year, active_sellers desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

year	active_sellers	start_date	end_date	num_products_listed
2016	68	2016-10-03 09:44:50	2016-10-10 18:09:39	107
2017	1358	2017-01-05 12:06:36	2017-12-31 22:14:53	9366
2018	1900	2017-12-29 00:49:20	2018-08-29 15:00:37	11394

IN&OUT (Query Result) shown above

How does the volume of orders placed on Olist fluctuate throughout the year?

```sql
SELECT
    COUNT(*) AS no_rows,
    COUNT(DISTINCT customer_id) AS no_customers,
     COUNT( order_id) AS no_orders,
    COUNT(*) - COUNT( order_status) AS no_missing_status
FROM orders;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

no_rows	no_customers	no_orders	no_missing_status
96461	96461	96461	0

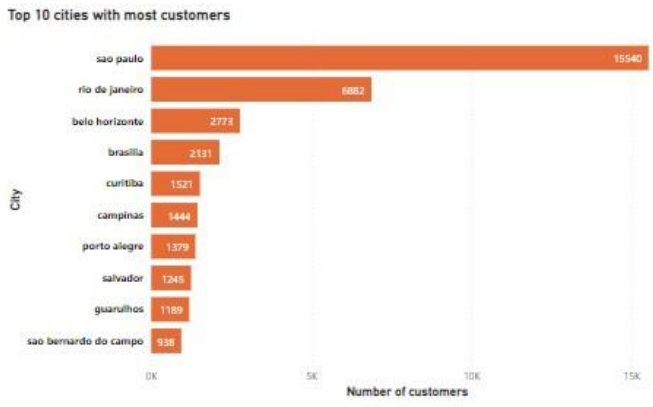
```sql
select case when extract(month from order_purchase_timestamp)='1' then 'january'
 when extract(month from order_purchase_timestamp)='2' then 'february'
  when extract(month from order_purchase_timestamp)='3' then 'march'
when extract(month from order_purchase_timestamp)='4' then 'april'
 when extract(month from order_purchase_timestamp)='5' then 'may'
 when extract(month from order_purchase_timestamp)='6' then 'june'
when extract(month from order_purchase_timestamp)='7' then 'july'
 when extract(month from order_purchase_timestamp)='8' then 'august'
when extract(month from order_purchase_timestamp)='9' then 'september'
when extract(month from order_purchase_timestamp)='10' then 'october'
when extract(month from order_purchase_timestamp)='11' then 'november'
 when extract(month from order_purchase_timestamp)='12' then 'december'
else 0 end as new_month,
count(order_id) as no_of_orders
from orders o
where order_purchase_timestamp is not null
```

VISUALIZATIONS

Visualizations to questions asked/answered

The Customers Table

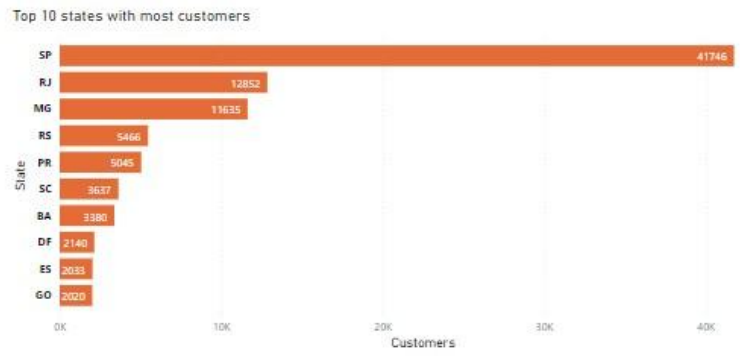
Top 10 cities with most customers



Top 10 cities with most customers

How many cities have more than 500 customers?*This table was not visualized because there was no possible way to visualize one distinct count and will not be useful for this report based project.*

Top 10 states that have the most customers?

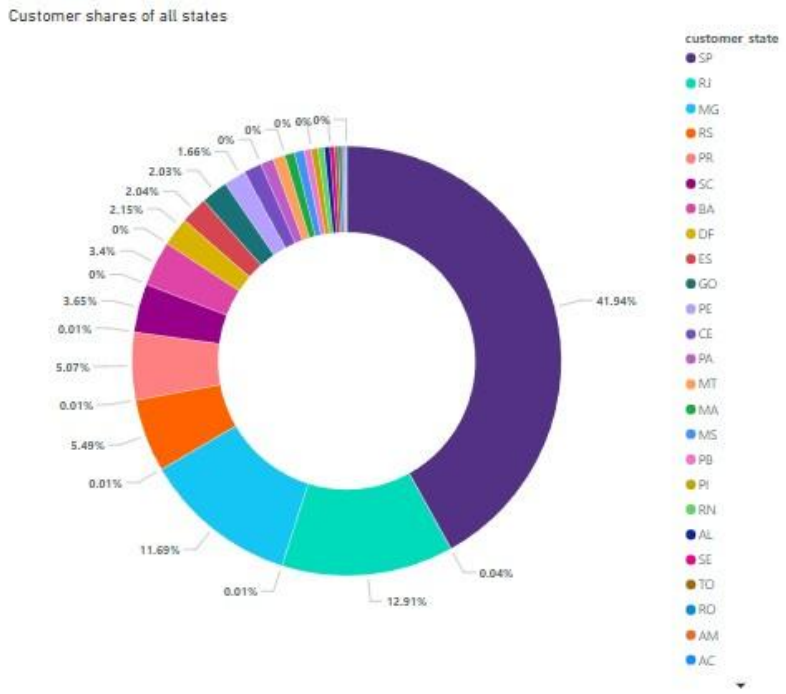


Top 10 state with most customers

How many percent of the customer base do the top ten cities account for?

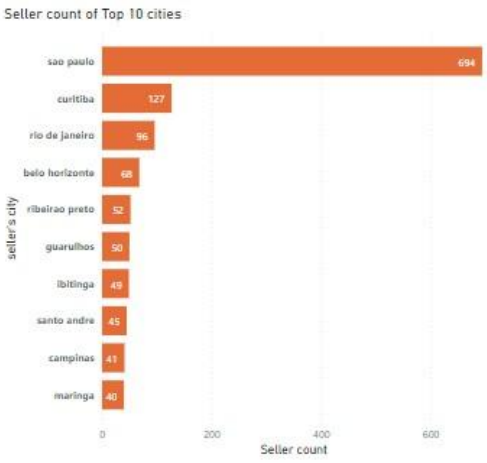
This table was not visualized as there are more than 4,000 cities, and will not be useful for this report based project.

How many percent of the customer base do the top ten states account for?



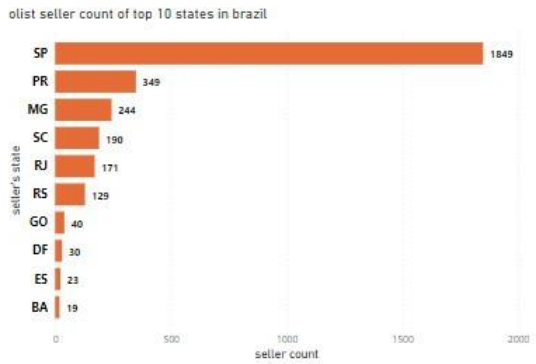
The Sellers Table

Top 10 cities with most sellers



Top 10 cities with most sellers

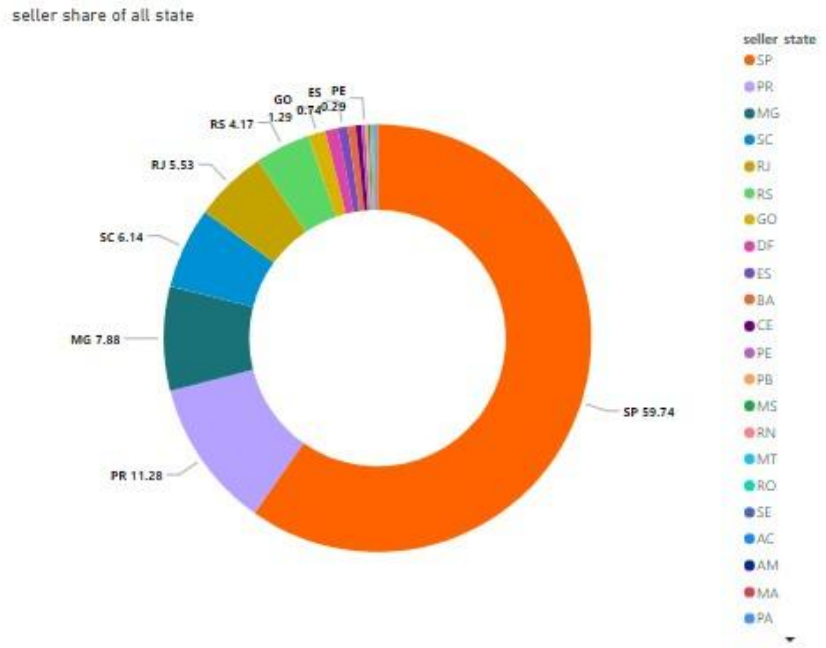
Top 10 states with most sellers



How many percent of the seller base do the top ten cities and total cities account for?

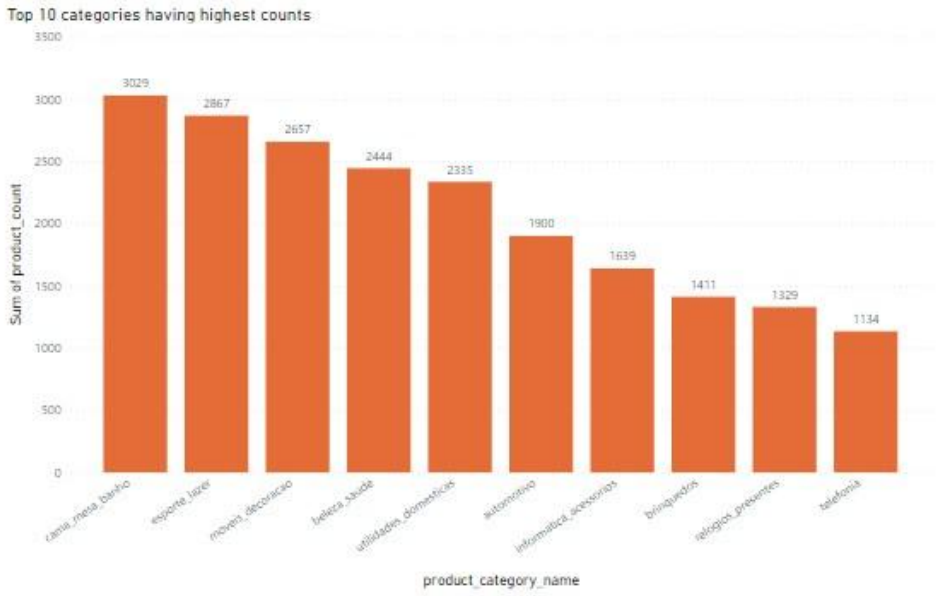
This table was not visualized as there are more than 600cities and will not be useful nor descriptive for this report based project.

How many percent of the seller base do the top ten states and total states account for?



The Products Table

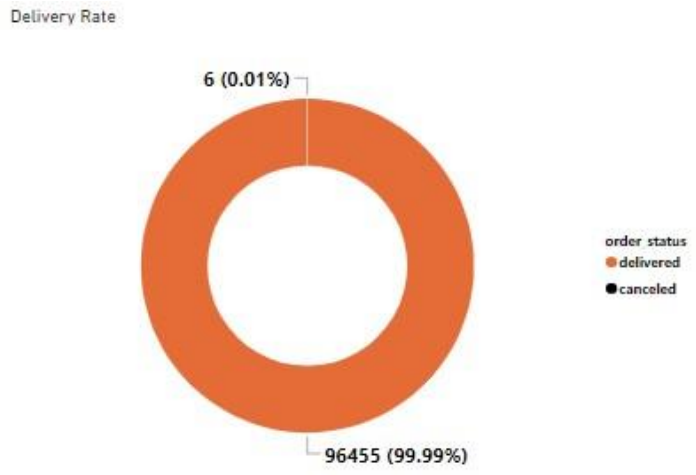
What are the top ten categories having the highest product counts



Top 10 product categories having the highest count (products)

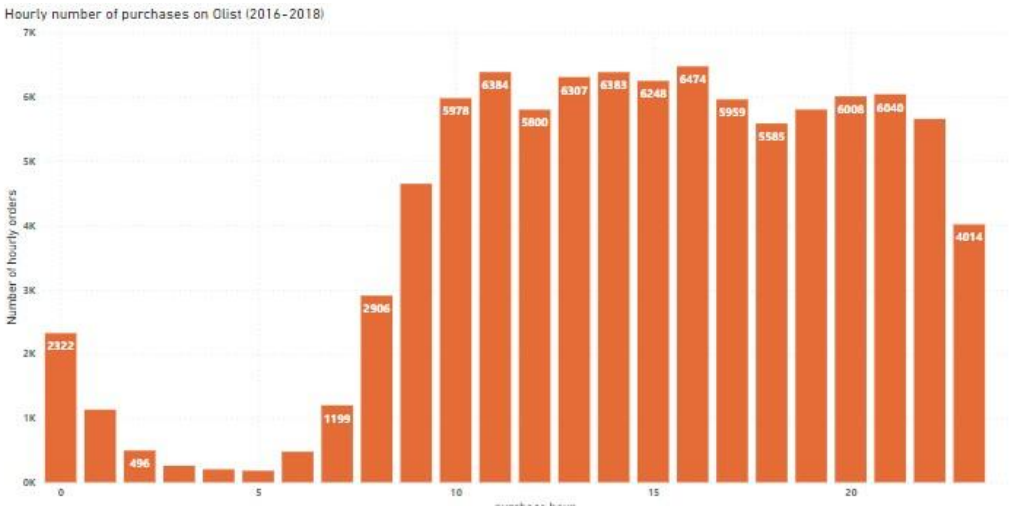
The Orders Table

What is the delivery rate?



Delivery Rate(ordered , cancelled)

How does the delivery rate changes over time?

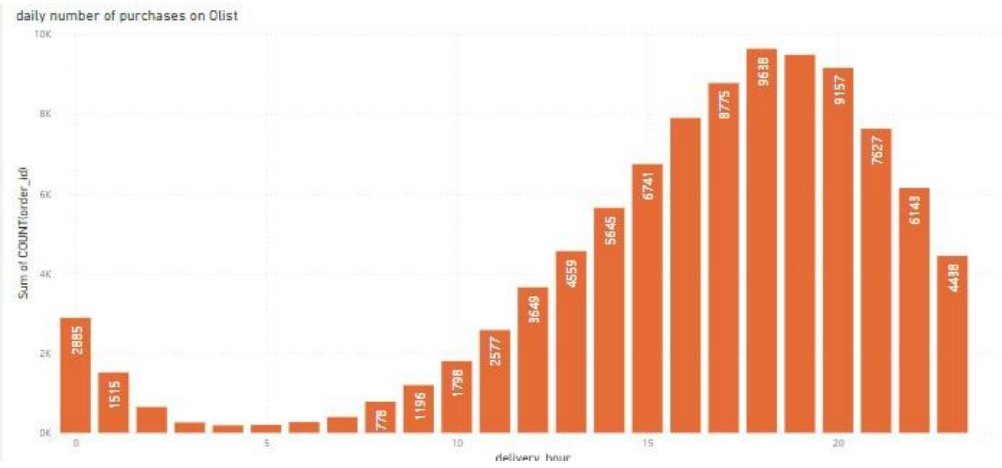


Rate of change of delivery rate (hrs)

Orders of Day of the week

This table was not visualized

Daily number of purchases/orders on Olist



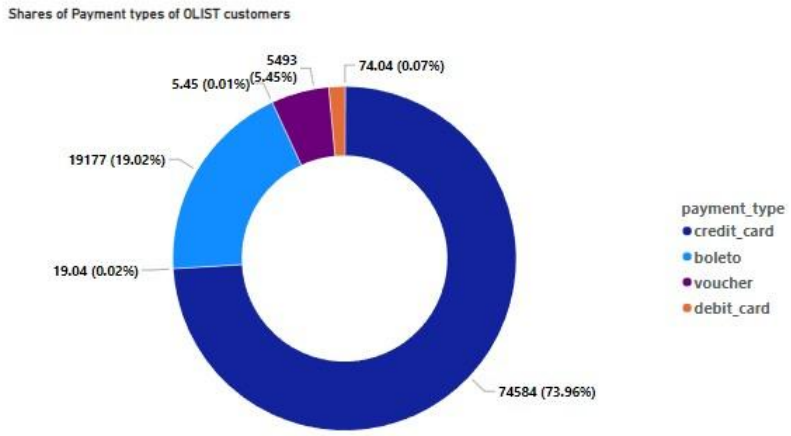
The order_items table Visualization

Olist Product-order Ratio *This*

query was not visualized.

The order payments table

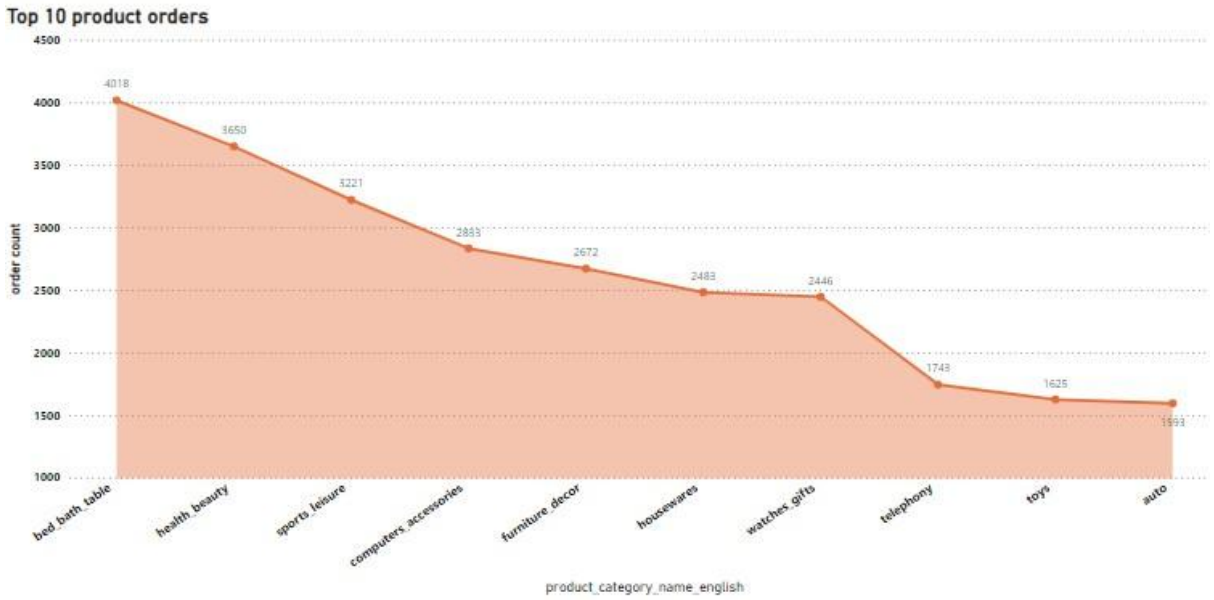
Shares of payment types of Olist customers



Payment types on Olist

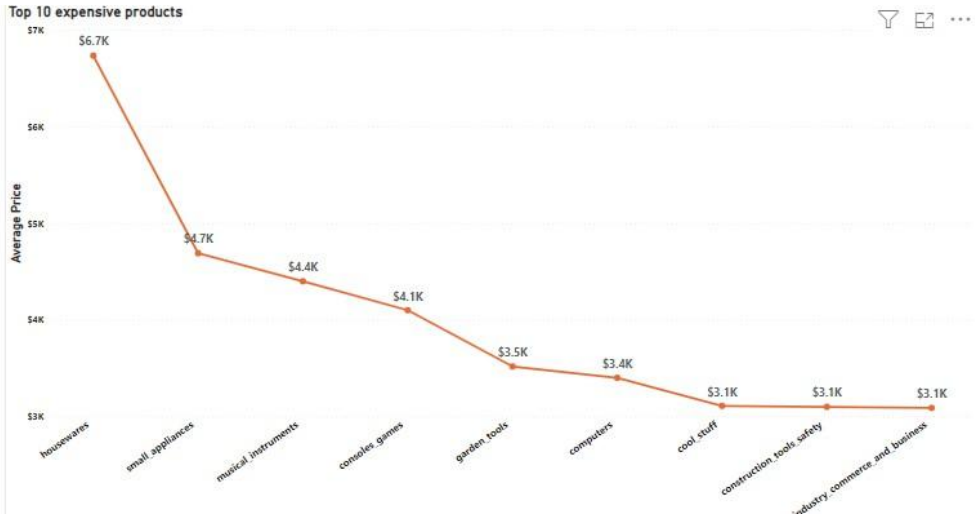
Other Questions visualization

Top 10 product category orders



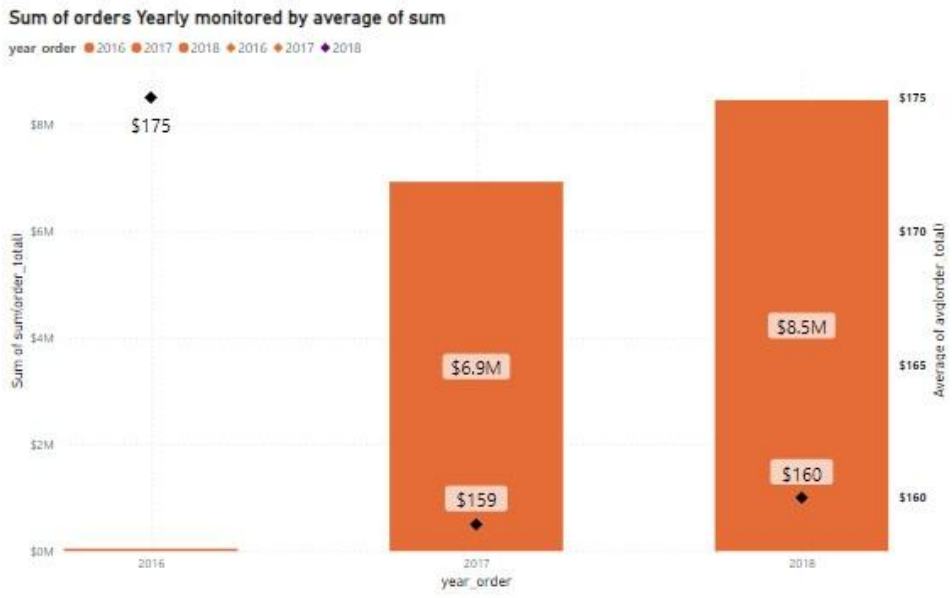
Top 10 product category orders based on quantity.

Top 10 expensive product category



Top 10 expensive product category based on price (in USD)

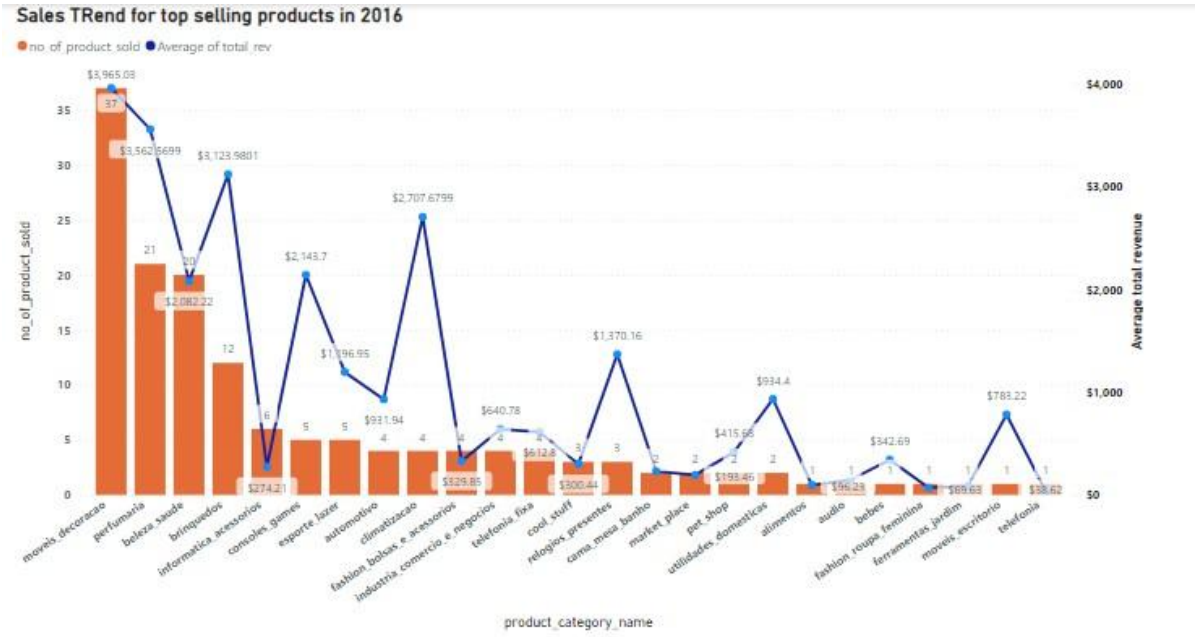
Total orders per year marked with average



Total sum of ordersin USD per year with average orders(2016-2018)

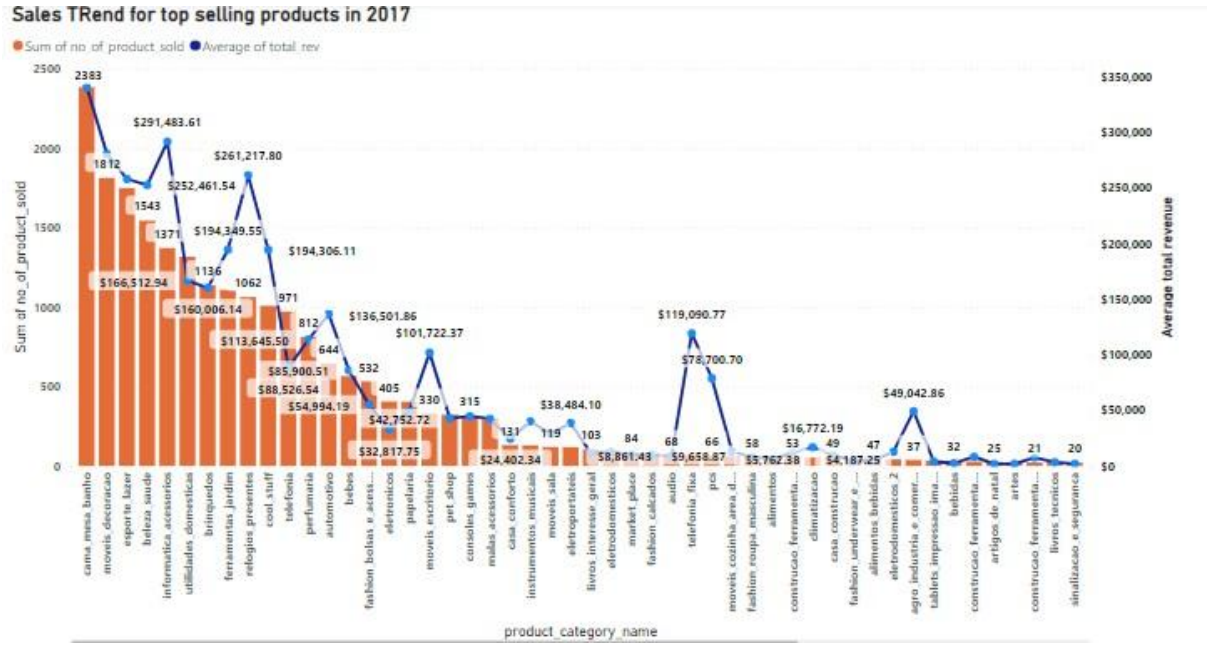
Top-selling products on Olist, and how have their sales trends changed over the years(2016-2018)

a)2016



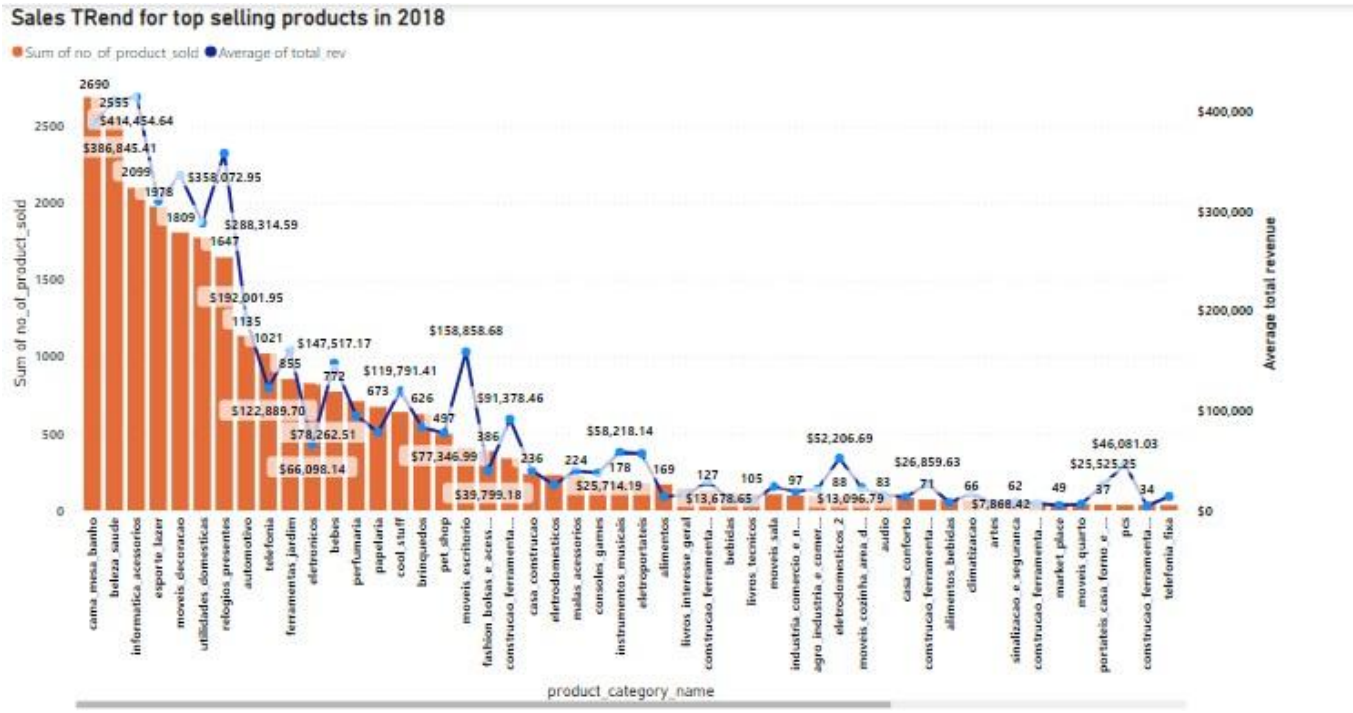
Sales trend of top selling products for 2016

b)2017



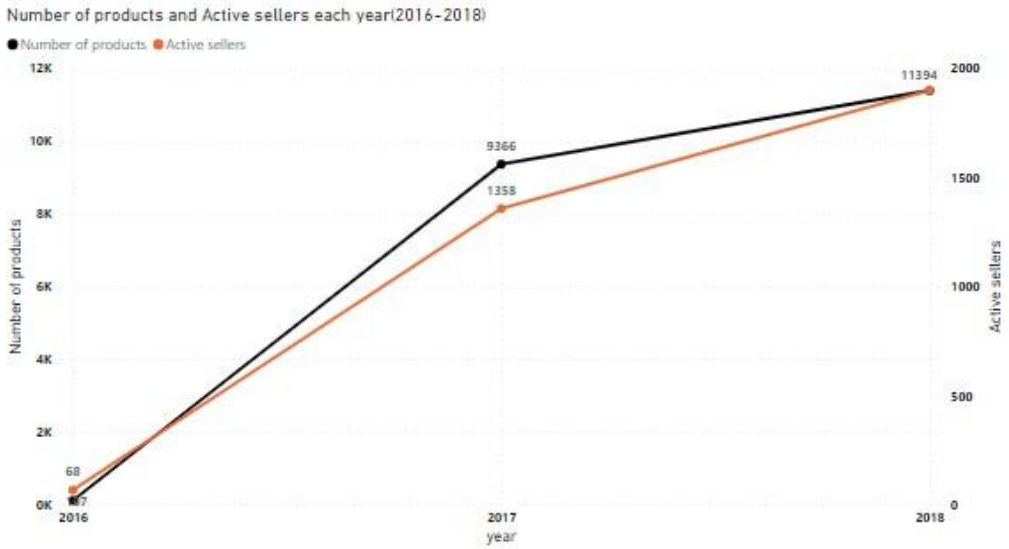
Sales trend of top selling products for 2017

c) 2018



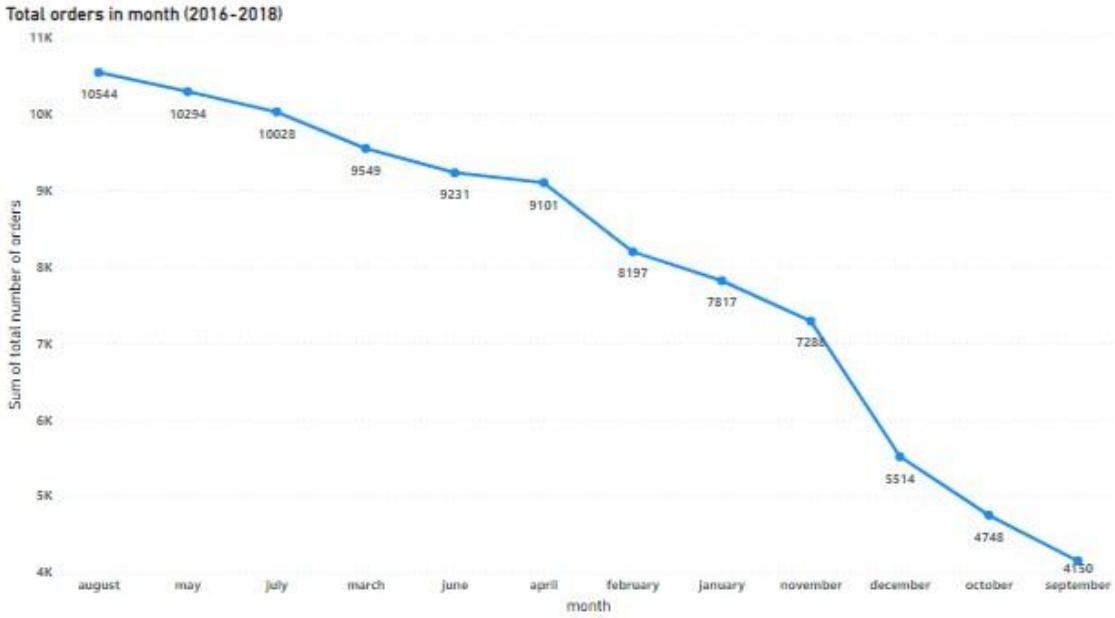
Sales trend of top selling products for 2018

Number of products and Active sellers each year(2016-2018)



Total number of products and active sellers from (2016-2018)

What is the order volume on Olist, and how does it fluctuate throughout the year?



Total orders on Olist for every month from (2016-2018).

CHAPTER FIVE

CONCLUSION

In conclusion, data analysis of an e-commerce store can provide valuable insights into customer behavior, sales performance, and marketing effectiveness. By analyzing data and identifying patterns and trends, businesses can optimize pricing strategies, improve marketing tactics, and measure key performance indicators to improve business performance and increase revenue. Data analysis can also help businesses identify opportunities for growth and areas for improvement, leading to a better overall customer experience. Therefore, investing in data analysis is crucial for e-commerce businesses to remain competitive in today's market.

REFERENCES

Abbas, M. (2022, June 10). *Brazilian E-commerce Public Dataset by Olist*. Retrieved from Medium: https://medium.com/@muhab.abbas11/brazilian-e-commerce-public-dataset-by-olist-9757bc7964a4#:~:text=of%20the%20year.-,2%3A%20How%20many%20orders%20were%20placed%20on%20Olist%2C%20and%20how,vary%20by%20month%20or%20season%3F

Anderson, J. A. (2019). Inventory Optimization in Ecommerce A DataDriven Approach. *Journal of Retailing and Consumer Services*, 1-12.

ChinhmaiGit. (n.d.). *Project 2*. Retrieved from chinhmaigit.github.io: https://chinhmaigit.github.io/Project-SQL-2/html/project2.html

Davis, R. (2023). E-commerce Market Expansion Strategies: A Data-Driven Approach. *International Journal of Research in Pharmaceutical Sciences*.

Johnson, E. (2019). E-commerce sales forecasting: A review and future directions. *International Journal of Forecasting*, 1306-1327.

Johnson, M. (2020). Data Privacy and Security in E-commerce: A Literature Review. *Unpublished manuscript*.

Rodriguez, M. (2020). Personalization in E-commerce: A Literature Review. *Journal of Electronic Commerce Research*, 327-355.

Smith, J. (2018). *Data Analytics for E-commerce: A Comprehensive Overview.* Cham, Switzerland: Springer.

tobye070. (2022, August 22). *The Exploratory Data Analysis on Olist E-commerce Dataset*. Retrieved from Medium: https://medium.com/@tobye070/the-exploratory-data-analysis-on-olist-e-commerce-dataset-cbddd09d936c

Jain, S., and Khanduja, D. (2019). "Data Analytics in E-commerce: Opportunities and Challenges." Procedia Computer Science, vol. 164, pp. 389-395.

Li, X., and Hu, Y. (2018). "Data Analysis for E-commerce: A Review." Proceedings of the 2018 International Conference on Computer Science and Artificial Intelligence, pp. 51-55.

Sehgal, S., and Srivastava, S. (2019). "Data Analytics for E-commerce Websites: A Review." International Journal of Computer Applications, vol. 181, no. 36, pp. 11-15.

Shams, R., and Hossain, M. (2019). "A Study on the Role of Data Analytics in E-commerce." International Journal of Scientific and Engineering Research, vol. 10, no. 6, pp. 16691674.

Sun, Y., and Wang, Y. (2019). "Data Analytics in E-commerce: A Literature Review and Research Agenda." International Journal of Information Management, vol. 46, pp. 69-80