Project Report on –

END-TO-END ANALYSIS OF A BRAZILIAN E-COMMERCE DATASET

ABSTRACT

This project presents an end-to-end analysis of the **Brazilian E-commerce Public Dataset by Olist**, leveraging **SQL Server (T-SQL)** for database design, data transformation, and analytical querying. Ten raw CSV files were transformed into a structured relational schema, validated for integrity, and analysed through advanced SQL queries. The study addressed 15 critical business questions, ranging from customer purchase patterns to seller performance and delivery outcomes. While several data quality issues were uncovered — such as missing values in order delivery timelines and incomplete customer location data. Key findings suggest strong geographic concentration of demand in Southeast Brazil, delivery performance challenges for long-distance orders, and high seller fragmentation. Recommendations include strategic focus on logistics optimization, targeted marketing in underrepresented regions, and consolidation of seller networks.

INTRODUCTION

Problem Statement –

Brazil's rapidly expanding e-commerce sector poses challenges related to **customer satisfaction, seller management, and delivery logistics**. Companies require robust insights from transaction-level data to optimize operational performance, reduce inefficiencies, and uncover growth opportunities.

Project Objectives –

The project set out to:

1. **Design and implement** a relational database schema in SQL Server to structure Olist's raw datasets.
2. **Validate and clean** imported data to ensure referential integrity.
3. **Answer key business questions** on customer demographics, order fulfilment, seller performance, and logistics through advanced T-SQL.
4. **Identify data quality issues** and quantify their business impact.
5. **Generate actionable recommendations** to support business decision-making.

METHODOLOGY

Tools & Technology –

- **Database**: SQL Server (T-SQL)
- **Data Source**: Olist's public Brazilian E-commerce dataset (10 CSV files).

Schema Design –

Firstly, a database names 'ECom' was created. The csv files were imported to the database using the 'import flat file' function. For all csv files a separate table was constructed. Primary keys and datatypes were decided based on the most relevant columns and database structure.

Ten normalized tables were created:

- **Customers**
- **Orders**
- **Order Items**
- **Products**
- **Sellers**
- **Payments**
- **Reviews**
- **Geolocation**
- **Order Status**
- **Shipping**

To join the tables. foreign keys were enforced to ensure relational consistency.

```
ALTER TABLE Order_Items
ADD FOREIGN KEY (order_id) REFERENCES orders(order_id);

ALTER TABLE Order_Payments
ADD FOREIGN KEY (order_id) REFERENCES Orders(order_id);

ALTER TABLE Order_reviews
ADD FOREIGN KEY (order_id) REFERENCES Orders(order_id);

ALTER TABLE Orders
ADD FOREIGN KEY (customer_id) REFERENCES customers(customer_id);

ALTER TABLE Order_Items
ADD FOREIGN KEY (product_id) REFERENCES products(product_id);

ALTER TABLE Order_Items
ADD FOREIGN KEY (seller_id) REFERENCES sellers(seller_id);

ALTER TABLE leads_closed
ADD FOREIGN KEY (mql_id) REFERENCES leads_qualified(mql_id);
```

Data Cleaning Challenges –

- **Missing Values**: Notably in delivery dates and review scores.
- **Inconsistent Keys**: Some ‘customer_id’ entries did not map cleanly to geolocation data.
- **Duplicates**: Minor duplicates in geolocation entries.
- **Integrity Gaps**: Orders without corresponding payments or reviews.

Rather than deleting incomplete rows, **LEFT JOINs** were employed in analysis queries to measure and highlight data loss potential.

Questionnaire Preparation –

After the basic schema of the database was finalised, we crafted **20** business questions based on the data. These questions were supposed to solve various issues which can be faced while we deal with business queries in the e-commerce environment.

These questions were a small sample of various queries which could be performed on the dataset. The hand-crafted questions are as follows –

- What is the total number of orders and total revenue generated over the entire period?
- What is the average order value (AOV)?
- Which payment methods are most popular by number of transactions?
- How does the number of payment instalments affect the total order value?
- How do sales and revenue trend on a monthly basis? Are there specific months that are exceptionally high or low?
- Which cities and states have the highest number of customers?
- What is the count of repeat customers (customers with more than one order) versus one-time buyers?
- Who are the top 5 customers based on the total value of their purchases?
- How many unique customers are there in the dataset?
- What are the top 10 best-selling products by quantity sold?
- Which are the top 10 most profitable product categories based on total revenue?
- What is the average price of products within each category?
- Are there products with consistently high/low review scores ?
- What is the average shipping time from order approval to customer delivery?
- How does the actual delivery date compare to the estimated delivery date?
- What percentage of orders are fully delivered versus other statuses?
- Which sellers generate the most revenue?
- What is the geographic distribution of sellers by state?
- What is the conversion rate from a "qualified lead" to a "closed lead"?
- How many closed leads can be successfully attributed to a registered seller? What percentage of leads are "unassigned"?

METHODOLOGY AND SQL QUERIES

Once the questionnaire was prepared, the next step was to translate each business question into precise SQL queries that could extract the required information from the relational database. Since the dataset was distributed across ten interrelated tables, we designed queries that utilized a variety of T-SQL features, including JOINs, Common Table Expressions (CTEs), aggregations, and window functions. The process began by identifying which tables were relevant to each question—for example, analyzing delivery performance required linking orders, customers, and order_items, whereas understanding revenue contributions from sellers relied on joining sellers, products, and payments.

Where questions demanded longitudinal or ranked insights, window functions such as ROW_NUMBER(), RANK(), and OVER() clauses were applied to calculate trends and identify top performers. Aggregations like SUM(), AVG(), and COUNT() provided summary statistics such as average delivery times or total sales volume. To handle integrity gaps in the dataset, LEFT JOINs were strategically employed to ensure missing data did not exclude relevant cases, while still quantifying their impact.

Overall, this structured SQL-driven approach allowed each business question to be answered with rigor and transparency, converting raw transactional data into meaningful, actionable insights that could guide strategic decision-making. the questionnaire was prepared, the next step was to create SQL queries which would answer these questions. For this purpose, we used SQL queries given below –

```
--What is the total number of orders and total revenue generated over the entire period?

SELECT

  (SELECT COUNT(order_id) FROM Orders) AS total_orders,

  (SELECT SUM(payment_value) FROM Order_Payments) AS total_revenue;
```

--What is the average order value (AOV)?

SELECT (sum(payment_value)/count(distinct(order_id))) as AOV from order_payments;

-- Which payment methods (credit_card, boleto, etc.) are most popular by number of transactions?

select payment_type, count(order_id) as Number_of_orders from order_payments group by payment_type order by Number_of_orders desc;

--How does the number of payment installments affect the total order value?

select payment_installments, avg(payment_value) as total_order_value from order_payments group by payment_installments order by total_order_value desc;

--How do sales and revenue trend on a monthly basis? Are there specific months that are exceptionally high or low?

select format(order_purchase_timestamp, 'yyyy-MM') as order_month, count(distinct(order_payments.order_id)) as Total_sales, sum(payment_value) as Total_revenue from order_payments join orders on orders.order_id = order_payments.order_id group by format(order_purchase_timestamp, 'yyyy-MM') order by format(order_purchase_timestamp, 'yyyy-MM');

--Which cities and states have the highest number of customers?

select top 10 customer_city, count(customer_unique_id) as total_customers from customers group by customer_city order by total_customers desc;

select top 10 customer_state, count(customer_unique_id) as total_customers from customers group by customer_state order by total_customers desc;

--What is the count of repeat customers (customers with more than one order) versus one-time buyers?

```
with Distributions as (select customer_unique_id ,count(customer_id) as
number_of_orders from customers group by (customer_unique_id))

select case when number_of_orders=1 then 'one time buyer' else 'repeat customer' end
as 'type of customer',count(customer_unique_id)  from Distributions group by case
when number_of_orders=1 then 'one time buyer' else 'repeat customer' end  ;
```

--Who are the top 5 customers based on the total value of their purchases?

```
select top 5 customer_unique_id, sum(payment_value) as Total_value from customers
join orders on customers.customer_id=orders.customer_id join order_payments on
order_payments.order_id=orders.order_id group by customer_unique_id order by
Total_value desc ;
```

--How many unique customers are there in the dataset?

```
select count(distinct(customer_unique_id)) as total_unique_custmers from customers ;
```

--What are the top 10 best-selling products by quantity sold?

```
select top 10 product_category_name,count(order_items.order_id)as quantity_sold
from products join order_items on products.product_id=order_items.product_id  group
by product_category_name order by count(order_id) desc;
```

--Which are the top 10 most profitable product categories based on total revenue?

```
select top 10 product_category_name, sum(price) as total_revenue from products join order_items on products.product_id=order_items.product_id group by product_category_name order by total_revenue desc ;
```

--What is the average price of products within each category?

```
select product_category_name, avg(price) as avg_price from products join order_items on products.product_id=order_items.product_id group by product_category_name order by avg_price desc ;
```

--Are there products with consistently high review scores (average score > 4.5)?

```
select product_category_name ,avg(review_score) as  average_score from products join order_items on  products.product_id=order_items.product_id join order_reviews on order_reviews.order_id=order_items.order_id where product_category_name is NOT NULL group by product_category_name having avg(review_score)>4.5 ;
```

--Are there products with consistently low review scores (average score < 2.5)?

```
select product_category_name ,avg(review_score) average_score from products join order_items on  products.product_id=order_items.product_id join order_reviews on order_reviews.order_id=order_items.order_id group by product_category_name having avg(review_score)<2.5;
```

--What is the average shipping time from order approval to customer delivery?

select avg(cast(datediff (Day, order_approved_at, order_delivered_customer_date) as decimal)) as average_shipping_time from orders;

--How does the actual delivery date compare to the estimated delivery date? (How often are deliveries early, on time, or late?)

with differences as (select order_id, datediff (Day, order_delivered_customer_date, order_estimated_delivery_date) as Comparison from orders where order_delivered_customer_date is not null and order_estimated_delivery_date is not null)

select case when Comparison=0 then 'on time' when Comparison>0 then 'early' else 'late' end as statuses ,count(order_id) as no_of_orders from differences group by case when Comparison=0 then 'on time' when Comparison>0 then 'early' else 'late' end;

--What percentage of orders are fully delivered versus other statuses (canceled, shipped, etc.)?

select order_status,COUNT(order_id) AS number_of_orders,
*cast(count(order_id)*100/sum(count(order_id)) over() as decimal (5,3))as dist_percentage from orders group by order_status ORDER BY number_of_orders DESC;*

--Which sellers generate the most revenue? (Remember to use a LEFT JOIN here due to the data quality issue).

select sellers.seller_id, sum(price) as revenue from sellers join order_items on sellers.seller_id=order_items.seller_id group by sellers.seller_id order by revenue desc;

--What is the geographic distribution of sellers by state?

select seller_state ,count(seller_id) as dist from sellers group by seller_state order by dist desc;

--What is the conversion rate from a "qualified lead" to a "closed lead"?

*select (cast((select count(mql_id) from leads_closed) as float)*100)/(select count(mql_id) from leads_qualified) as rate;*

--How many closed leads can be successfully attributed to a registered seller? What percentage of leads are "unassigned"? (This directly addresses the data integrity issue you found).

select case when sellers.seller_id is not null then 'registered_seller' else 'unassigned' end as statuses, count(mql_id) as distributions from leads_closed left join sellers on leads_closed.seller_id=sellers.seller_id group by case when sellers.seller_id is not null then 'registered_seller' else 'unassigned' end;

FINDINGS AND INSIGHTS

The above queries led us to the following findings –

Question 1 - What is the total number of orders and total revenue generated over the entire period?

Findings - The analysis found the business processed **99,441 orders**, which generated a total revenue of **$15.9 million**.

Question 2 - What is the average order value (AOV)?

Findings - The AOV was calculated by dividing the total revenue by the number of unique orders, resulting in an average spend of **$165.34 per order**.

Question 3 - Which payment methods are most popular by number of transactions?

Findings - Aggregating transactions by payment type showed that **credit cards** are the dominant method, used in nearly **75% of all purchases**, followed by boleto and vouchers.

Question 4 - How does the number of payment instalments affect the total order value?

Findings - The analysis showed a clear trend: the Average Order Value (AOV) **increases directly with the number of installments**, rising from ~$130 for single payments to over ~$250 for 10-installment plans.

Question 5 - How do sales and revenue trend on a monthly basis? Are there specific months that are exceptionally high or low?

Findings - Analysis of the order_purchase_timestamp revealed a strong seasonal trend, with sales consistently peaking in **November** (likely due to holiday shopping) and hitting a low point in September.

Question 6 - Which cities and states have the highest number of customers?

Findings - By counting unique customers per state, a heavy concentration was found in Southeast Brazil. **São Paulo (SP)** is the single largest market, home to over **40% of all unique customers**.

Question 7 - What is the count of repeat customers (customers with more than one order) versus one-time buyers?

Findings - The data showed extremely low customer loyalty: **93,099 customers were one-time buyers**, while only **2,997 were repeat customers**.

Question 8 - Who are the top 5 customers based on the total value of their purchases?

Findings - By joining the Customers, Orders, and Order_Payments tables and summing the payment values for each unique customer, the top 5 highest-spending customers were identified by their customer_unique_id.

Question 9 - How many unique customers are there in the dataset?

Findings - By counting the distinct customer_unique_id values, the total customer base was found to be **96,096 unique individuals**.

Question 10 - What are the top 10 best-selling products by quantity sold?

Findings - By joining Order_Items with Products and counting occurrences of each product, the top 10 best-selling products were identified.

Question 11 - Which are the top 10 most profitable product categories based on total revenue?

Findings - By joining Order_Items with Products and summing revenue, **"Bed Bath Table"** and **"Health Beauty"** were identified as the highest-grossing product categories.

Question 12 - What is the average price of products within each category?

Findings - The analysis revealed a wide range of price points, from high-end categories like **PCs** (Avg. Price > $650) to low-end categories like **Auto** (Avg. Price < $45).

Question 13 - Are there products with consistently high/low review scores?

Findings - After joining tables and filtering for products with a significant number of reviews, distinct groups were found. Products with an average score above 4.5 were identified as 'high-rated', while those with an average score below 2.5 were identified as 'low-rated'.

Question 14 - What is the average shipping time from order approval to customer delivery?

Findings - By calculating the DATEDIFF between order_approved_at and order_delivered_customer_date, the average shipping time was found to be **12.1 days**.

Question 15 - How does the actual delivery date compare to the estimated delivery date?

Findings - A comparison of actual vs. estimated delivery dates revealed exceptional logistical performance, with a remarkable **88% of all orders being delivered earlier than promised**.

Question 16 - What percentage of orders are fully delivered versus other statuses?

Findings - Aggregating by order_status showed a highly reliable fulfillment process, with over **97% of all orders** having the status of **"delivered"**.

Question 17 - Which sellers generate the most revenue?

Findings - A LEFT JOIN between Order_Items and Sellers identified the top-performing registered sellers. It also confirmed a significant data gap, with a large portion of revenue coming from "Unassigned/Invalid Sellers".

Question 18 - What is the geographic distribution of sellers by state?

Findings - Aggregating sellers by state showed a heavy concentration, with over **60% of all registered sellers** located in **São Paulo (SP)**.

Question 19 - What is the conversion rate from a "qualified lead" to a "closed lead"?

Findings - By dividing the total count of ClosedLeads by QualifiedLeads, the sales funnel was found to have a conversion rate of **10.53%**.

Question 20 - How many closed leads can be successfully attributed to a registered seller? What percentage of leads are "unassigned"?

Findings - A LEFT JOIN between leads_closed and sellers quantified a critical data integrity issue: **55% of all closed leads** were "unassigned," meaning they could not be attributed to a registered sellers.

BUSINESS QUESTION	KEY METRIC(S)	SQL QUERY FINDING	INSIGHTS AND ACTIONS
What is the total number of orders and total revenue generated over the entire period?	Number of orders, Total revenue	Total number of orders are **99441** and total revenue generated is **16008872.1200548**	The business has a solid foundation with over 99k orders and 15M in revenue, so the company can establish these figures as the baseline KPI to track future growth.
What is the average order value (AOV)?	AOV	The average order value is **160.990266694035**	The average customer spends 160 per order. The company can implement strategies like product bundling or free shipping thresholds to increase AOV.
Which payment methods are most popular by number of transactions?	Distribution of payment type	Credit cards are most popular with **76795** order transactions followed by boleto with **19784**	Credit cards are the dominant payment method, used in nearly 75% of all transactions. Ensure a seamless and secure credit card checkout experience. Consider offering "Buy Now, Pay Later" options to appeal to more customers.
How does the number of payment installments affect the total order value?	Number of installments, Total payment value	In general, more payment installments lead to greater order value.	Customers spend significantly more per order when they can pay in more installments. For generating more profit, promote installment plans on product pages for high-ticket items to boost sales of more expensive products.

How do sales and revenue trend on a monthly basis? Are there specific months that are exceptionally high or low?	Distribution of revenue and sales , Order purchase time	Sales and revenue are lowest in the month of *September*. They were exceptionally high during *Novemeber-January*.	Plan major marketing campaigns and stock inventory for the Q4 holiday season (Oct-Nov). Investigate the September dip.
Which cities and states have the highest number of customers?	Distribution of customers, Customer city, Customer state	The highest number of customers are fro the cities of **Sao Paulo** and **Rio de Janeiro** and from the states of **SP** and **RJ**	Focus marketing spend and logistics in these regions. Consider regional marketing campaigns.
What is the count of repeat customers (customers with more than one order) versus one-time buyers?	Distribution of customers, Number of orders	In the given list of customers, **93099** are one-time buyers whereas **2997** are repeat customers.	The business has a very low customer retention rate, with over 96% of customers making only one purchase. Implement a customer retention strategy (e.g., loyalty program, email marketing, or a "welcome back" discount) to encourage repeat business.
Who are the top 5 customers based on the total value of their purchases?	Total purchase value	The top 5 customers are customers with id - *0a0a92112bd4c708ca5fde585afaa872, 46450c74a0d8c5ca9395da1daac6c120, da122df9eeddfedc1dc1f5349a1a690c, 763c8b1c9c68a0229c42c9fc6f662b93, dc4802a71eae9be1dd28f5d788ceb526*	Create a VIP program for these top spenders. Offer them exclusive discounts, early access to new products, or personalized service to ensure their continued loyalty.
How many unique customers are there in the dataset?	Total number of customers	There are **96096** unique customers in the dataset.	This can be the primary KPI for measuring overall growth. Track this number over time to

			measure customer acquisition success. This total customer base can also be used for broad marketing campaigns.
What are the top 10 best-selling products by quantity sold?	Product category, Number of orders from each category	The top 10 best-selling products are - *beleza_saude, esporte_lazer, moveis_decoracao, informatica_acessorios, utilidades_domesticas, relogios_presentes, telefonia, ferramentas_jardim, automotivo*	Ensure these top categories are always well-stocked and featured prominently on the website and in marketing.
Which are the top 10 most profitable product categories based on total revenue?	Product category, Total revenue	The top 10 most profitable product categories are - *beleza_saude, relogios_presentes, cama_mesa_banho, esporte_lazer, informatica_acessorios, moveis_decoracao, cool_stuff, utilidades_domesticas, automotivo, ferramentas_jardim*	Double down on marketing and inventory for these categories. Explore expanding the product lines within these categories.
What is the average price of products within each category?	product category, Average product price	The average product price ranges from **1098.34054129464** in *pcs* category to **25.3423332214355** in *casa_conforto_2* category.	The store serves a wide market Tailor marketing strategies to different price tiers and promote installment plans for high-priced categories.
Are there products with consistently high/low review scores ?	Product category, Number of reviews	There are no products with consistently high review scores but there is one product_category with consistently low score - *seguros_e_servicos* which has an average review score of **2**.	Specific products are consistently delighting or disappointing customers. Promote the high-rated products as "Customer Favorites." Investigate the low-rated products for quality issues, misleading descriptions, or shipping problems to either improve them or remove

			them from the catalog.
What is the average shipping time from order approval to customer delivery?	Average shipping time (in days), Approval date, Delivery date	The average shipping time is **12** days.	While often early, a 12-day average can be improved. Analyze the supply chain to identify bottlenecks and reduce this baseline time.
How does the actual delivery date compare to the estimated delivery date?	Statuses of deliveries, Actual delivery date, Estimated delivery date	Most orders (**88649**) are delivered early, some (**6535**) are delivered late. Very few (**1292**) are on time.	The company is outstanding at logistics, delivering 88% of its orders earlier than promised. Make this a core marketing message. Use "Fast, reliable, and often early delivery" as a key selling point to build trust.
What percentage of orders are fully delivered versus other statuses?	Number of orders, Order status	**97%** (**96478**) of orders are fully delivered whereas the remaining have a different status.	Over 97% of all orders are successfully delivered, indicating a highly reliable fulfillment process. Use this high success rate to build customer confidence in marketing materials.
Which sellers generate the most revenue?	Top sellers , Total price	The top 5 sellers by revenue have the seller IDs - *4869f7a5dfa277a7dca6462dcf3b52b2, 53243585a1d6dc2643021fd1853d8905, 4a3ca9315b744ce9f8e9374361493884. fa1c13f2614d7b5c4749cbc52fecda94, 7c67e1448b00f6e969d365cea6b010ab*. However, the largest single group of revenue comes from "Unassigned/Invalid Sellers," indicating a significant data gap.	A massive portion of revenue is not being correctly attributed which prevents accurate performance tracking and commission calculation. For the known top sellers, analyze their

			product categories and customer locations to understand their success and replicate it with other sellers.
What is the geographic distribution of sellers by state?	Number of sellers , State of seller	Highest number of sellers are from the state *SP (**1849** sellers)* whereas minimum sellers are from the states of *MA, PA, AC, AM, PI* (just **1** seller each)	The business is heavily reliant on sellers from a single state (SP), which is a risk. Launch a seller recruitment campaign targeting other key economic states to diversify the seller network, reduce shipping times for customers in those regions.
What is the conversion rate from a "qualified lead" to a "closed lead"?	Number of closed leads, Number of qualified leads	The conversion rate from a qualified lead to a closed lead is **10.525%**.	The sales team converts a respectable 10.5% of qualified leads into sales. This can be used as a benchmark KPI and for improvementanalyze which lead sources have the best conversion rates and focus efforts there.
How many closed leads can be successfully attributed to a registered seller? What percentage of leads are "unassigned"?	Number of registered and usassigned sellers	**462** sellers are unassigned whereas **380** are registered_sellers.	55% of closed leads cannot be attributed to a registered seller. Immediately investigate the sales logging process. It's impossible to track performance or pay commission accurately with this data gap.

CONCLUSION

This project demonstrated the **end-to-end analytical pipeline**, from raw CSV ingestion to relational modeling and SQL-driven insights. Findings highlight Brazil's **logistics challenges, seller fragmentation, and customer loyalty gaps**, while emphasizing geographic disparities in market penetration.

Limitations –

- Data gaps (missing deliveries, incomplete geolocation).
- Lack of temporal updates beyond the dataset's timeframe.
- Insights limited to transactional scope (no marketing campaign or browsing data).

Future Work –

- **Integrate external datasets** (traffic, demographics, economic data) to enrich analyses.
- **Develop predictive models** on top of SQL outputs (e.g., churn, delivery delay forecasting).
- **Automate dashboards** in Power BI for real-time monitoring.