Burak Kaya

1.1 PROJECT TITLE

E-comerce in Brazil

1.2 GENERAL DESCRIPTION OF THE DOMAIN

The domain of this project is e-commerce in Brazil, focusing on the retail segment of an online sales platform. The dataset covers various aspects of the platform's operations, including customer orders, product listings, payments, and shipping price. It provides a detailed view of transactions between businesses and consumers, capturing data such as order timestamps, product details, customer locations, payment methods, and shipping information. The main goals of analyzing this dataset are compare sale frequency by different periods, identify top-selling products, and geographical customer distribution. These insights will help create effective marketing strategies, optimize operations, and improve customer satisfaction in the competitive e-commerce market.

1.3 DESCRIPTION OF THE ANALYSIS AREA WITH JUSTIFICATION

This e-commerce project is all about diving into how customers behave and what they prefer by analyzing sales data. By closely examining orders and product details, businesses aim to understand which products are selling the most. This analysis helps in optimizing product offerings, pricing strategies, and marketing efforts to meet customer demand effectively. It also allows businesses to determine sales frequency, helping to better plan for periods of high and low demand.

Justification and Business Value:

1. Understanding Customer Preferences:

- Analyzing products data helps identify top selling products.

2.Geographical Customer Distribution:

-Knowing the number of customers by city helps with regional market analysis and targeting.

3. Understanding Sales Frequency

-Knowing the number of sales by month helps to see the high demand period and low demand period.

1.4 IDENTIFIED PROBLEMS

The dataset provides many opportunities for analysis in the e-commerce domain. Key decision problems include understanding product performance, customer behaviour and seasonal demand. By analyzing which products are most popular, how customers' buying frequency change across different times. Businesses can improve their marketing strategies and increase operational efficiency. Solving these decision problems with data analysis helps businesses improve profits, build better customer relationships, and stay competitive in the fast-changing e-commerce market.

1.5 EXPECTATIONS AND DETAILED NEEDS FOR DECISION SUPPORT

1. Marketing Manager:

- Description: Responsible for creating marketing strategies and campaigns to attract and retain customers.

- Expectations: Seeks insights into customer behaviour, product performance, and market trends to optimize marketing efforts.

Burak Kaya

2. Sales Analyst:

- Description: Analyzes sales data to identify trends and optimize sales strategies.

- Expectations: Requires detailed analysis of sales metrics, customer behaviour, and product performance to enhance sales effectiveness.

OLAP User Query Types:

1. Top-Selling Products Category Name:

- Utilization: Marketing Manager can identify popular products for targeted promotions.

2-Average Delivery Time:

-Utilization: Sales Analyst can monitor average delivery times to ensure timely deliveries and improve customer satisfaction.

3- Total Order Number by Time Period:

-Utilization: Sales Analyst can track the total number of orders to monitor overall sales performance. This helps in identifying sales growth trends, evaluating the effectiveness of sales campaigns.

4-Number of Sales by Payment Method:

-Utilization: Sales Analyst can understand which payment methods are most popular and ensure a smooth checkout process for customers.

5-Total Number of Customers by City:

Utilization: Marketing Manager can analyze the distribution of customers across different cities. This information is valuable for regional market analysis, allowing for targeted marketing campaigns and identifying areas with high sales potential.

6-Shipping Charges:

Utilization: Sales Analyst can analyze shipping charges to identify cost-saving opportunities and optimize shipping strategies.

7-Total Payment Value by Payment Method:

Utilization: Sales Analyst can analyze the total payment value associated with different payment methods, helping to understand customer preferences and streamline the payment process.

8- Canceled Orders:

Utilization: Sales Analyst can analyze cancelled orders to identify number of cancelled orders by time.

9-Average Product Price by Category:

Utilization: Marketing Manager and Sales Analyst can calculate the average price of products within specific categories. This information helps in understanding the pricing structure within each product category and setting competitive prices for products.

10-Customer Distribution by State:

Utilization: Marketing Manager can analyze the distribution of customers across different states. This information is valuable for regional market analysis, allowing for targeted marketing campaigns and identifying areas with high sales potential.

1.6.1 LOCATION, FORMAT, AVAILABILITY

https://www.kaggle.com/datasets/quangvinhhuynh/marketing-and-retail-analyst-e-comerce?select=customers.csv

I took the data from the Kaggle as a csv format. This csv contains 5 sheets.

1.6.2 DATA SOURCE BASIC INFORMATION

	Source	Number of rows	Number of attributes	Size	Update rate	Grain
1	Shhet1	99442	7	13.77MB		Per Year, Per Month, Per Day
2	Sheet2	112651	6	12.7MB		Per order
3	Sheet3	99442	4	5.38MB		Per customer
4	Sheet4	103887	5	5.73MB		Per order
5	Sheet5	32952	6	1.79MB		Per product

This data source contains various attributes that I used. However, I didn't used all of them. In order to create a multidimensional data model.

1.7.1 FACTS AND MEASURES

	Fact	Measure(s)	Grain
1	Total Order Number	Count(order_purchase_timestamp)	Sum quantity sold
	Top Selling Products Category	Count(product_category_name)	Sum product by category name
	Average delivery time	Avg (delivery_time)	Average time per order
	Shipping Charges	Sum (shipping_charges)	Sum shipping item charges
	Number of Sales by Payment Method	Count (payment_type)	Sum number of sales by payment method
	Total number of customers by city	Sum(customer_city)	Sum customer by city
	Cancelled Orders	Count(order_status)	Sum quantity cancelled orders
	Average Product Price by Category	Avg(product_price)	Average price per category

	Customer Distribution by State	Count(customer_state)	Sum customer by state
	Total Payment Value by Payment Method	Sum(payment_value)	Sum payment by payment method

1.7.2 CONTEXT FOR FACTS

	Dimension	Description	Grain
1	Products	Product information	Per products
2	Customer Adress	Customer location	Per customer
3	Order Date	Details about the time of purchase	Per order
4	Payment	Details about payment	Per payment

1.7.3 DIMENSION OVERVIEW

	Dimension	Hypothesized Attributes
1	Products	Surrogatekey, product_category_name, price, shipping_charges
2	Customer Adress	Surrogatekey ,customer_zip_code_prefix, customer_city, customer_state
3	Order Date	Surrogatekey, order_purchase_timestamp, order_delivered_timestamp, order_status
4	Payment	Surrogatekey , payment_type, payment_value

1.7.1 DOMAIN DATA DICTIONARY

	Location	Attribute name	Attribute type	Description
1	Fact	order_id	INT	Unique identifier for an order
2	Fact	quantity_sold	INT	Total quantity of products sold
3	Fact	total_payment	DECIMAL	Total payment per order
4	Fact	shipping_charges	DECIMAL	Total shipping charges for the order
5	Fact	number_of_sales_by_payment_method	INT	Total number of sales by payment method
6	Fact	total_number_of_customers_by_city	INT	Total number of customers by city
7	Fact	cancelled_orders	INT	Count of canceled orders
8	Fact	total_payment_value_by_payment_method	DECIMAL	Total payment value by payment method

9	Fact	average_delivery_time	DECIMAL	Average delivery time per order
10	Fact	total_number_of_customers_by_state	INT	Total number of customers by state
11	Products	Surrogatekey	INT	Surrogate key for a product
12	Products	product_category_name	VARCHAR	Category of the product
13	Products	price	DECIMAL	Price of the product
14	Products	shipping_charges	DECIMAL	Shipping charges for the product
15	Customer Address	Surrogatekey	INT	Surrogate key for customer address
16	Customer Address	customer_zip_code_prefix	VARCHAR	Zip code prefix of the customer's address
17	Customer Address	customer_city	VARCHAR	City of the customer's address
18	Customer Address	customer_state	VARCHAR	State of the customer's address
19	Order Date	Surrogatekey	INT	Surrogate key for order date
20	Order Date	order_purchase_timestamp	TIMESTAMP	Timestamp of when the order was purchased
21	Order Date	order_delivered_timestamp	TIMESTAMP	Timestamp of when the order was delivered
22	Order Date	order_status	VARCHAR	Status of the order (e.g., delivered)
23	Payment	Surrogatekey	INT	Surrogate key for payment
24	Payment	payment_type	VARCHAR	Type of payment (e.g., credit_card)
25	Payment	payment_value	DECIMAL	Transaction value

1.7.2 QUALITY ASSESMENT SHEET

	Location	Attribute name	Attribute type	Type of data	Number of Unique values	Number of Null values	Quality assessment
1	Fact	order_id	INT	Nominal	99442	0	Good quality, no missing values, unique for each record
2	Fact	order_revenue	DECIMAL(10,2)	Ratio	5969	0	Good quality, no missing values,

							varied values
3	Fact	shipping_charges	DECIMAL(10,2)	Ratio	7000	0	Good quality, no missing values, varied values
4	Fact	quantity_sold	INT	Ratio	98667	0	Good quality, no missing values, unique for most records
5	Fact	total_payment	DECIMAL(10,2)	Ratio	99441	0	Good quality, no missing values, unique for most records
6	Orders	order_id	INT	Nominal	99442	0	Good quality, no missing values, unique for each order
7	Orders	customer_id	INT	Nominal	96097	0	Good quality, no missing values, unique for each customer
8	Orders	order_status	VARCHAR(20)	Nominal	3	0	Good quality, no missing values, limited categories
9	Orders	time_id	INT	Nominal	98876	0	Good quality, no missing values, unique for most records
10	Customers	customer_id	INT	Nominal	96097	0	Good quality, no missing values, unique for

							each customer
11	Customers	customer_zip_code_prefix	VARCHAR(10)	Nominal	14995	0	Good quality, no missing values, varied values
12	Customers	customer_city	VARCHAR(50)	Nominal	4120	0	Good quality, no missing values, varied values
13	Customers	customer_state	VARCHAR(20)	Nominal	28	0	Good quality, no missing values, limited categories
14	Products	product_id	INT	Nominal	32952	0	Good quality, no missing values, unique for each product
15	Products	product_category_name	VARCHAR(50)	Nominal	72	0	Good quality, no missing values, varied categories
16	Products	product_weight_g	DECIMAL(10,2)	Ratio	2206	2	Acceptable quality, few missing values, varied values
17	Products	product_length_cm	DECIMAL(10,2)	Ratio	101	2	Acceptable quality, few missing values, varied values
18	Products	product_height_cm	DECIMAL(10,2)	Ratio	104	2	Acceptable quality, few missing values,

							varied values
19	Products	product_width_cm	DECIMAL(10,2)	Ratio	97	2	Acceptable quality, few missing values, varied values
20	Time	time_id	INT	Nominal	98876	0	Good quality, no missing values, unique for most records
21	Time	date	DATE	Interval	98876	0	Good quality, no missing values, unique for each date
22	Time	day_of_month	INT	Interval	30	0	Good quality, no missing values, limited categories
23	Time	month_of_year	INT	Interval	12	0	Good quality, no missing values, limited categories
24	Time	year	INT	Interval	3	0	Good quality, no missing values, limited categories
25	Time	day_of_week	INT	Interval	7	0	Good quality, no missing values, limited categories
26	Time	hour	INT	Interval	24	0	Good quality, no missing values, limited categories

27	Payment	payment_id	INT	Nominal	99441	0	Good quality, no missing values, unique for each payment
28	Payment	order_id	INT	Nominal	99441	0	Good quality, no missing values, unique for each order
29	Payment	payment_type	VARCHAR(20)	Nominal	5	0	Good quality, no missing values, limited categories
30	Payment	payment_value	DECIMAL(10,2)	Ratio	29078	0	Good quality, no missing values, varied values

GENERAL CONCLUSIONS:

In this project, I investigated the e-commerce market in Brazil, focusing on how online shops work. I gathered data about what customers buy and when they buy it to understand what they like most. This helped us figure out which products sell well, and which ones don't. I faced some problems with handling a lot of data, but I managed to get useful information that helped us make better decisions for our business.

This work showed us how important it is to use data to make decisions, especially in a market with lots of competition. The things I learned from this project will help us do better in the future by knowing more about what our customers want. This way, we can offer them the products they like and improve their shopping experience.