DATAOPT

GROUP #2 SHOPEE

Data Engineering Proposal

Group 2 | Presentation

Presentation Outline

1. Company Background
2. KPIs
3. Data Architecture
4. Data Quality
5. Data Ingestion
6. Data Integration
7. Data Visualization
8. ML/AIProposal

Company Background

Shopee is an e-commerce platform founded in 2015 and headquartered in Singapore. It is owned by Sea Limited, a Singaporean multinational technology company.

Shopee offers a wide range of products and services, including fashion, beauty, electronics, and home appliances, among others. The platform operates in Southeast Asia, Taiwan, Brazil, and Mexico, and has over 200 million active users.

KPIs

1. Customer lifetime value (CLV)
   - This shows the value of a customer over time.

2. Customer satisfaction (CSAT)
- This shows the how much a customer is satisfied with our products and services

3. Average order value (AOV)
- This shows how much do people spend per order.

4. Cart Abandonment
- This shows how many orders were cancelled and orders that did not make the payment process

Data Architecture

Data Architecture

Attributes:

Reports Field	Availability	Database	Entity	Attribute	Logic	Last Updated By	Last Updated Date
product_id	Y	store	products	product_id	direct pull	Vicente Magboo	Feb-19-2023
product_name	Y	store	products	product_name	direct pull	Vicente Magboo	Feb-19-2023
brand_id	Y	store	products	brand_id	direct pull	Vicente Magboo	Feb-19-2023
category_id	Y	store	products	category_id	direct pull	Vicente Magboo	Feb-19-2023
list_price	Y	store	products	list_price	direct pull	Vicente Magboo	Feb-19-2023
description	Y	store	products	description	direct pull	Vicente Magboo	Feb-19-2023
image	Y	store	products	image	direct pull	Vicente Magboo	Feb-19-2023
stock	Y	store	products	stock	direct pull	Vicente Magboo	Feb-19-2023
min_count	Y	store	products	min_count	direct pull	Vicente Magboo	Feb-19-2023
max_count	Y	store	products	max_count	direct pull	Vicente Magboo	Feb-19-2023
unit_price	Y	store	products	unit_price	direct pull	Vicente Magboo	Feb-19-2023
original_price	Y	store	products	original_price	direct pull	Vicente Magboo	Feb-19-2023
rate	Y	store	products	rate	direct pull	Vicente Magboo	Feb-20-2023
brand_id	Y	store	brand	brand_id	direct pull	Vicente Magboo	Feb-19-2023
brand_name	Y	store	brand	brand_name	direct pull	Vicente Magboo	Feb-19-2023
shop_id	Y	store	shop	shop_id	direct pull	James Maranion	Feb-19-2023
product_id	Y	store	shop	product_id	direct pull	Vicente Magboo	Feb-28-2023
shop_name	Y	store	shop	shop_name	direct pull	James Maranion	Feb-19-2023
status	Y	store	shop	status	direct pull	James Maranion	Feb-19-2023
region	Y	store	shop	region	direct pull	James Maranion	Feb-19-2023
message	Y	store	shop	message	direct pull	James Maranion	Feb-19-2023
response	Y	store	shop	response	direct pull	James Maranion	Feb-19-2023
shop_logo	Y	store	shop	shop_logo	direct pull	James Maranion	Feb-19-2023
region_id	Y	store	region	region_id	direct pull	James Maranion	Feb-24-2023
region_name	Y	store	region	region_name	direct pull	James Maranion	Feb-24-2023
merchant_id	Y	sales	merchant	merchant_id	direct pull	James Maranion	Feb-19-2023
shop_id	Y	store	merchant	shop_id	direct pull	James Maranion	Feb-19-2023
merchant_name	Y	sales	merchant	merchant_name	direct pull	James Maranion	Feb-19-2023
merchant_region	Y	sales	merchant	merchant_region	direct pull	James Maranion	Feb-19-2023
customer_id	Y	sales	customers	customer_id	direct pull	Cyrus Cabrera	Feb-19-2023
first_name	Y	sales	customers	first_name	direct pull	Cyrus Cabrera	Feb-19-2023
last_name	Y	sales	customers	last_name	direct pull	Cyrus Cabrera	Feb-19-2023
phone_number	Y	sales	customers	phone_number	direct pull	Cyrus Cabrera	Feb-19-2023
email_address	Y	sales	customers	email_address	direct pull	Cyrus Cabrera	Feb-19-2023
address	Y	sales	customers	address	direct pull	Cyrus Cabrera	Feb-19-2023
zip_code	Y	sales	customers	zip_code	direct pull	Cyrus Cabrera	Feb-19-2023
state	Y	sales	customers	state	direct pull	Cyrus Cabrera	Feb-19-2023

Data Architecture

Entities Filter:

Report Name	Granularity	Description	Filter	Last Updated By	Last Updated Date
Total Stocks	stocks	To know the available products	WHERE stocks > 0	Cyrus Cabrera	20-02-2023
Shop Acquisition	status	to know the active shop	WHERE status = active	Cyrus Cabrera	20-02-2023
Category Labels	category_name	to know the specific category that the buyer wants	WHERE category_name = name(Category)	Cyrus Cabrera	20-02-2023
Ratings	Rate	To know where the specific rating that the buyer wants	WHERE rate = 5	Cyrus Cabrera	20-02-2023
Deals	start_time	To know where the specific start_time of the deals	WHERE start_time = start_time(date)	Cyrus Cabrera	20-02-2023
Voucher	voucher_name, start_time	To know when is the start_time of the voucher	WHERE voucher_name = start_time(date)	Cyrus Cabrera	20-02-2023
Voucher	voucher_name, end_time	To know when is the end_time of the voucher	WHERE voucher_name = end_time(date)	Cyrus Cabrera	20-02-2023
Brand	brand_name	To know the specific brand name	WHERE brand_name = name(brand)	Cyrus Cabrera	20-02-2023
Top Picks	name	To know the specific Top picks items	WHERE name = name(Top picks name)	Cyrus Cabrera	20-02-2023
States	state	To know the specific state of the customers	WHERE state = state(the name of the state)	Cyrus Cabrera	20-02-2023
Products	product_name, product_id	To know specific products	WHERE product_name = name(product)	Cyrus Cabrera	20-02-2023
Payment	buyer_total_amount	to know how much a customer pays per order	WHERE buyer_total_amt >=500	Cyrus Cabrera	20-02-2023
Orders	order_id, order_status	To know how many orders are completed	COUNT(order_id), WHERE order_status = "Delivered"	Cyrus Cabrera	20-02-2023

Data Architecture

Entities Keys:

Driving Table	Relational Table	Join Type	Join Condition	Last Updated By	Last Updated Date
products	brand	INNER JOIN	FROM products INNER JOIN brand on products.brand_id =brand.brand_id	James Maranion	20-Feb-23
products	shop	INNER JOIN	FROM products INNER JOIN shop on products.product_id= shop.product_id	James Maranion	20-Feb-23
products	stocks	INNER JOIN	FROM products INNER JOIN stocks on products.product_id= stocks.product_id	James Maranion	20-Feb-23
products	categories	INNER JOIN	FROM products INNER JOIN categories on products.category_id = categories.category_id	James Maranion	20-Feb-23
products	discounts	INNER JOIN	FROM products INNER JOIN discount on products.product_id= discount.product_id	James Maranion	15-Apr-23
customers	orders	INNER JOIN	FROM customers LEFT JOIN orders on customers.customer_id = orders.customer_id	James Maranion	20-Feb-23
orders	payment	INNER JOIN	FROM orders INNER JOIN orders on orders.order_id = payment.order_id	James Maranion	21-Feb-23
cart	cart_item	INNER JOIN	FROM cart INNER JOIN cart_item on cart.cart_id = cart_item.cart_id	James Maranion	14-Apr-23
cart_item	products	INNER JOIN	FROM cart_item INNER JOIN products on cart_item.product_id = products.product_id	James Maranion	14-Apr-23
orders	cart	INNER JOIN	FROM orders INNER JOIN cart on orders.cart_id = cart.cart_id	James Maranion	14-Apr-23
products	top_pick_item	INNER JOIN	FROM products INNER JOIN top_pick_item on products.product_id =top_pick_item.product_id	James Maranion	21-Feb-23
top_pick_item	top_picks	INNER JOIN	FROM top_pick_item INNER JOIN top_picks on top_pick_item.top_picks_id =top_picks.top_picks_id	James Maranion	21-Feb-23

Data Architecture

Entities Keys:

add_on_ deal	products	INNER JOIN	FROM products INNER JOIN add_on_deal on products.product_id =add_on_deal.product_id	James Maranion	15-Apr-23
bundle_deal	products	INNER JOIN	FROM products INNER JOIN bundle_deal on products.product_id =bundle_deal.product_id	James Maranion	15-Apr-23
discount	add_on_ deal	INNER JOIN	FROM discount INNER JOIN add_on_deal on discount.model_id =add_on_deal.add_on_deal_id	James Maranion	21-Feb-23
discount	bundle_deal	INNER JOIN	FROM discount INNER JOIN bundle_deal on discount.model_id =bundle_deal.bundle_deal_id	James Maranion	21-Feb-23
payment	voucher	INNER JOIN	FROM payment INNER JOIN voucher on payment.voucher_from_seller = voucher.voucher_id	James Maranion	21-Feb-23
shop	merchant	INNER JOIN	FROM shop INNER JOIN merchant on shop.shop_id = merchant.shop_id	James Maranion	20-Feb-23

Data Architecture

HQL Tab:

Table Name	Column Name	Data Type	Description	Validity	Completeness (Mandatory/Nullable)	Last Updated By	Last Updated Date
products	product_id	INTEGER	A unique product identifier.	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
products	product_name	VCHAR(100)	name of the product	Only alphanumeric characters are accepted	Mandatory	James Maranion	24-Feb-23
products	brand_id	INTEGER	A unique brand identifier.	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
products	category_id	INTEGER	A unique category identifier	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
products	list_price	INTEGER	listing price of the product	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
products	description	VCHAR(100)	Description of the product	Only alphanumeric characters are accepted	Mandatory	James Maranion	24-Feb-23
products	image	NVARCHAR (MAX)	image name and type	Only alphanumeric characters are accepted	Nullable	James Maranion	24-Feb-23
products	stock	INTEGER	quantity of available products	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
products	min_count	INTEGER	minimum amt of products that can be ordered	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
products	max_count	INTEGER	maximum amt of products that can be ordered	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
products	unit_price	FLOAT	price per unit of the product	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
products	original_price	FLOAT	Original price of the product	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
products	rate	INTEGER	product ratings	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
brand	brand_id	INTEGER	A unique brand identifier.	Only number characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
brand	brand_name	VCHAR(100)	name of the brand	Only alphanumeric characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
shop	shop_id	INTEGER	A unique shop identifier	Only number characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
shop	product_id	INTEGER	A product identifier	Only number characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
shop	shop_name	VCHAR(100)	name of the shop	Only alphanumeric characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
shop	status	BINARY	status of the shop	Only number characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
shop	region	INTEGER	region of the shop	Only number characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
shop	message	VCHAR(100)	messages of the shop	Only alphanumeric characters are accepted	Nullable	Vicente Magboo	24-Feb-23
shop	response	VCHAR(100)	responses of the shop	Only alphanumeric characters are accepted	Nullable	Vicente Magboo	24-Feb-23
shop	shop_logo	NVARCHAR (MAX)	image name and type	Only alphanumeric characters are accepted	Nullable	Vicente Magboo	24-Feb-23
region	region_id	INTEGER	A unique region identifier	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
region	region_name	VCHAR(100)	name of the region	Only alphanumeric characters are accepted	Mandatory	James Maranion	24-Feb-23
merchant	merchant_id	INTEGER	A unique merchant identifier	Only number characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
merchant	shop_id	INTEGER	A unique shop identifier	Only number characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
merchant	merchant_name	VCHAR(100)	name of the merchant	Only alphanumeric characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
merchant	merchant_region	INTEGER	region number linked with the merchant	Only number characters are accepted	Mandatory	Vicente Magboo	24-Feb-23
customers	customer_id	INTEGER	A unique customer identifier	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23
customers	first_name	VCHAR(100)	The customer first name	Only alphanumeric characters are accepted	Mandatory	James Maranion	24-Feb-23
customers	last_name	VCHAR(100)	The customer last name	Only alphanumeric characters are accepted	Mandatory	James Maranion	24-Feb-23
customers	phone_number	INTEGER	The customer phone number	Only number characters are accepted	Mandatory	James Maranion	24-Feb-23

Data Architecture

SQL Script:

Create Table Script (SQL)

```
CREATE TABLE sales.orders(
order_id INT,
customer_id INT,
order_status BIT,
shipped_date date,
required_date date,
cart_id INT,
PRIMARY KEY(order_id),
);


CREATE TABLE store.categories(
category_id INT,
category_name varchar(100),
PRIMARY KEY(category_id),
);


CREATE TABLE store.stocks(
product_id INT,
quantity INT,
PRIMARY KEY(store_id)
);



CREATE TABLE sales.discount(
discount_id INT,
item_id INT,
model_promotion_price MONEY,
purchase_limit MONEY,
message VARCHAR(100),
start_time datetime,
end_time datetime,
discount_name VARCHAR(100),
```

Create Table Script (Hive)

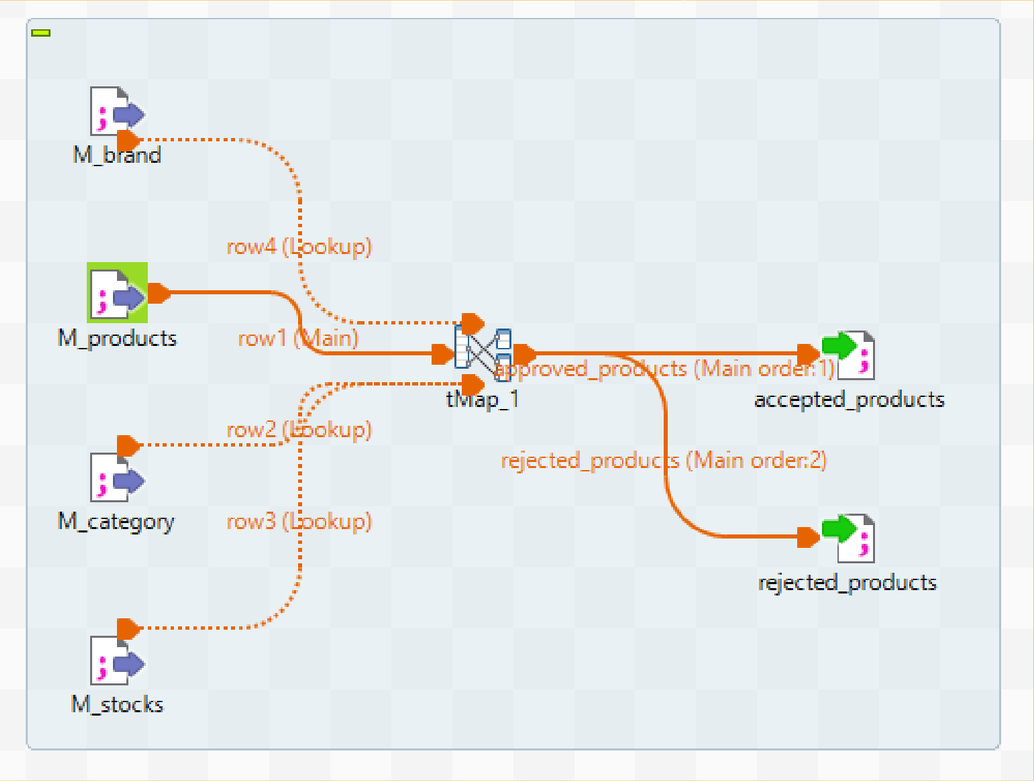
```
merch
STORED AS orc
LOCATION "/user/hive/shopee"
tblproperties("orc.compress"="NONE");

CREATE TABLE shopee.store_category(
category_id INT,
category_name VARCHAR(100)
)
STORED AS orc
LOCATION "/user/hive/shopee"
tblproperties("orc.compress"="NONE");


top
STORED AS orc
LOCATION "/user/hive/shopee"
tblproperties("orc.compress"="NONE");


CREATE TABLE shopee.sales_discount(
discount_id INT,
item_id INT,
model_promotion_price FLOAT,
purchase_limit FLOAT,
message VARCHAR(100),
start_time date,
end_time date,
discount_name VARCHAR(100),
discount_status BINARY
)
```

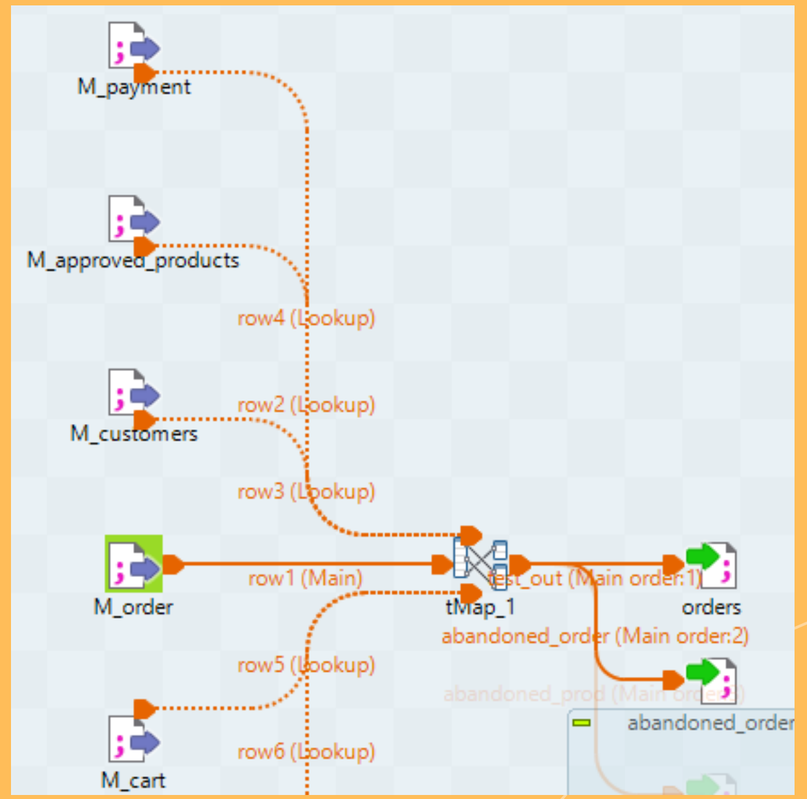
Data Ingestion

```
CREATE EXTERNAL TABLE store_bundle_deal( bundle_id INT,
product_id INT,
fix_price DECIMAL(10,2),
discount_percentage DECIMAL(5,2),
discount_value DECIMAL(10,2) )

row format delimited
fields terminated by ","
stored as textfile;

LOAD data inpath '/user/cloudera/shopeecsv/bundle_deal.csv'
overwrite into table store_bundle_deal;

select * from store_bundle_deal;
```

	store_bundle_deal.bundle_id	store_bundle_deal.product_id	store_bundle_deal.fix_price	store_bundle_deal.discount_percentage	store_bundle_deal.discount_val
1	1	1	394	0.25	152.5
2	2	2	220	0.5	16.4
3	3	3	209	0.5	157.5
4	4	4	411	0.1	16.2
5	5	5	380	0.75	20.9
6	6	6	357	0.25	167
7	7	7	258	0.25	131.5
8	8	8	408	0.5	209.25
9	9	9	172	0.5	116.5
10	10	10	311	0.25	35.8
11	11	11	493	0.25	95.25
12	12	12	243	0.1	92.5
13	13	13	194	0.5	91.5
14	14	14	292	0.5	98
15	15	15	482	0.25	189.5
16	16	16	368	0.5	111.75
17	17	17	286	0.5	140.25
18	18	18	387	0.25	28.2
19	19	19	238	0.5	190
20	20	20	465	0.1	110.5
21	21	21	321	0.25	49.3
22	22	22	420	0.5	45

Data Integration





Data Visualization

Cart Abandonment

Cart Abandonment Rate

.

18.10%

Total Carts Abandoned

no. of carts

181

Abandoned Cart Value

amount

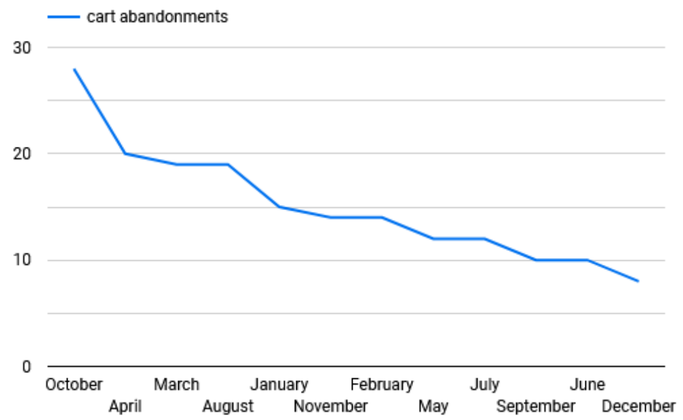
451,845

Recovery Rate

.

82.87%

Cart Abandonment By Month



Top 10 Abandoned Products

	product_name	category
1.	Vibrant Yellow Watch	Home & Kitchen
2.	Vibrant Yellow Chair	Automotive
3.	Vibrant White Sneakers	Books
4.	Vibrant White Shoes	Home & Kitchen
5.	Vibrant White Shoes	Automotive
6.	Vibrant White Headphones	Grocery
7.	Vibrant White Earbuds	Beauty
8.	Vibrant White Chair	Books
9.	Vibrant White Bag	Automotive
10.	Vibrant Silver Watch	Sports & Outdoors

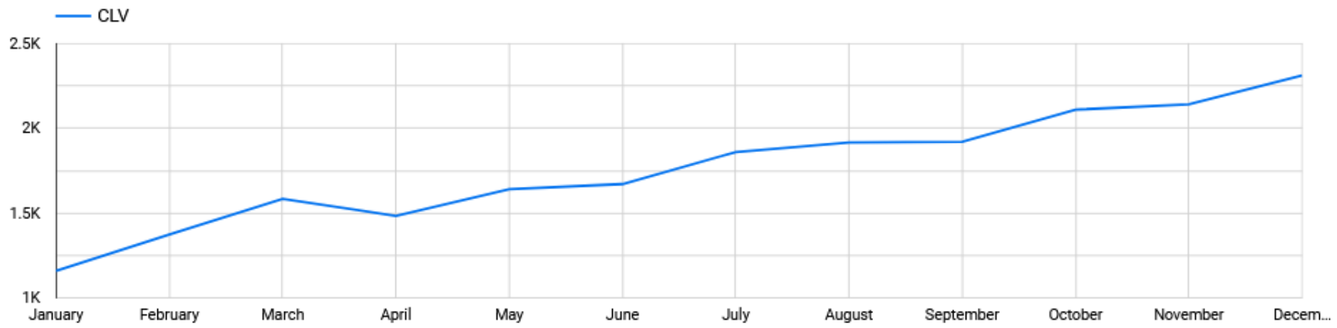
1 - 100 / 976

Data Visualization

Customer Lifetime Value (CLV)

Average Customer Lifetime in Days	Average Customer Lifetime Value	Average Order Value
no. of days	CLV	AOV
543.38	1,711.11	2,547.34

Customer Lifetime Value per Month



Data Visualization

Customer Satisfaction (CSAT)

Average CSAT

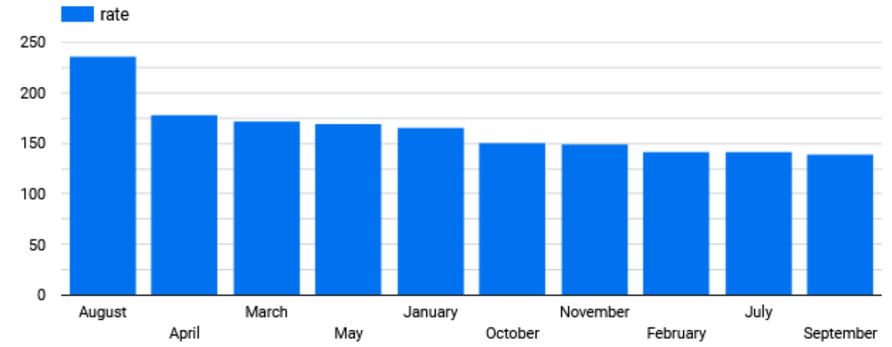
41.30%

CSAT by Category

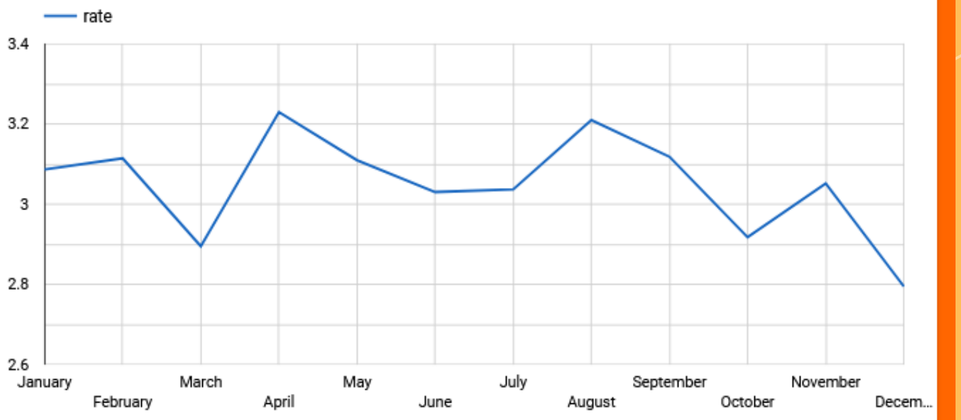
	category	rate
1.	Toys & Games	60.19%
2.	Sports & Outdoors	59.81%
3.	Home & Kitchen	62.08%
4.	Health & Personal Care	59.38%
5.	Grocery	61.9%
6.	Electronics	59.8%
7.	Clothing	59.16%
8.	Books	65.53%
9.	Beauty	62.55%
10.	Automotive	59.14%

1 - 10 / 10

Top Highest Ratings by Month



Average Product Ratings by Month



AI/ML Proposal

Neural Networks

Neural Network models can be used to predict the most effective product recommendations and pricing strategies that increase AOV

Regression

Regression models can predict a customer's future spending based on their past behavior.

Decision Trees

Decision trees can identify the most important factors that contribute to cart abandonment.

Decision trees can be used to segment customers into different groups based on their demographics, purchase history, and behavior.