**Project Document**

About the project :

This is a retail sales dataset containing transaction records from various stores across different cities in India. The data is from January 2022 and includes detailed information about customer purchases, demographics, and store performance metrics.

Purpose of the Project :

Analyse sales performance, customer behaviour, and operational efficiency across multiple stores in India to optimize business decisions and enhance customer satisfaction, with a focus on understanding purchase patterns, product performance, and regional variations to drive strategic improvements.

About Data :

1. Dataset Overview:

* The dataset appears to be a retail/e-commerce transaction log
* It contains 5,00,000 transactions for year 2022 from January 1st, 2022
* It has 19 columns capturing various transaction and customer details

1. Important Variables:

* Transaction details: TransactionID, TransactionDate, TransactionAmount, Quantity, DiscountPercent
* Customer information : CustomerID, CustomerAge, CustomerGender, LoyaltyPoints
* Product details : ProductName, IsPromotional
* Location details : City, Region, StoreType
* Payment and delivery: PaymentMethod, ShippingCost, DeliveryTimeDays
* Post-purchase: Returned, FeedbackScore

| **Column** | **Description** | **Data Type** |
| --- | --- | --- |
| invoice\_id | Invoice of the sales made | VARCHAR(30) |
| branch | Branch at which sales were made | VARCHAR(5) |
| city | The location of the branch | VARCHAR(30) |
| customer\_type | The type of the customer | VARCHAR(30) |
| gender | Gender of the customer making purchase | VARCHAR(10) |
| product\_line | Product line of the product solf | VARCHAR(100) |
| unit\_price | The price of each product | DECIMAL(10, 2) |
| quantity | The amount of the product sold | INT |
| VAT | The amount of tax on the purchase | FLOAT(6, 4) |
| total | The total cost of the purchase | DECIMAL(10, 2) |
| date | The date on which the purchase was made | DATE |
| time | The time at which the purchase was made | TIMESTAMP |
| payment\_method | The total amount paid | DECIMAL(10, 2) |
| cogs | Cost Of Goods sold | DECIMAL(10, 2) |
| gross\_margin\_percentage | Gross margin percentage | FLOAT(11, 9) |
| gross\_income | Gross Income | DECIMAL(10, 2) |
| rating | Rating | FLOAT(2, 1) |

1. Main Products are :

* Laptop
* Sofa
* T-Shirt
* Notebook
* Apple

1. Notable Patterns :

* There are multiple store types: In-Store and Online
* Various payment methods: Cash, Credit Card, Debit Card, UPI
* Transactions span multiple cities including Delhi, Mumbai, Bangalore, Kolkata
* Regions are categorized as North, South, East, West

1. Data Quality :

* Lot of missing data(null entries) in various fields – few transactions have incomplete information

Approach Used :

1. **Data cleaning** : This is the first step where inspection of data is done to make sure **NULL** values and missing values are detected and data replacement methods are used to replace, missing or **NULL** values
   * **Understand the Context of Missing Data**
     + **CustomerID, TransactionDate, PaymentMethod, StoreType, CustomerAge, CustomerGender, ProductName, Region**: These columns have 50,000 missing values each.
     + **Region**: Has 42,633 missing values.
     + Other columns have no missing values.
2. Load cleaned data to Postgresql database
3. Exploratory data analysis using SQL : Exploratory data analysis is done to answer the listed business questions and aims of this project.

**Business Questions To Answer**

**Generic Questions**

Q) Unique Cities, Products and Payment method the data has

select distinct city

from assignment\_cleaned ac

Analysis : Dataset has 10 different cities

select distinct productname

from assignment\_cleaned ac

Analysis : Apple, Laptop, Notebook, Sofa and T-Shirt)

select distinct paymentmethod

from assignment\_cleaned ac

Analysis : 4 different payment methods : Cash, Credit Card, Debit card and UPI

---------------------------------------------------Product analysis---------------------------------------------

Q) Most common payment method

select paymentmethod , count(paymentmethod) as cnt

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : Debit Card is most common whereas Credit card is least common

Q) Most selling Product

select productname , count(productname) as cnt

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : Notebook is most purchased item while Sofa is least purchased

Q) Quantity analysis

select max(quantity)

from assignment\_cleaned ac

select min(quantity)

from assignment\_cleaned ac

Analysis : Max qty sold is 50 whereas lowest qty sold is 1

Sales Analysis :

Q) Monthly Sales Trend

SELECT

TO\_CHAR(TO\_TIMESTAMP(transactiondate, 'YYYY-MM-DD'), 'Month') AS Month\_Name,

SUM(TransactionAmount) AS TotalSales

FROM assignment\_cleaned ac

GROUP BY Month\_Name

ORDER BY TotalSales desc;

Analysis :- It shows Sales is highest in January and lowest in month of December

Q) Product generating max/min revenue

select productname , sum(transactionamount) as total\_revenue

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : Laptop has generated maximum revenue whereas Apple has generated least revenue

Q) City generating maximum revenue

select city , sum(transactionamount) as total\_revenue

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : Kolkata has generated maximum revenue whereas Hyderabad generated the least

Q) Pdt with max discount on an avg :

select productname , avg(discountpercent) as avg\_discount\_percent

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : Almost same discount for all pdts(Max for apple : 25.03% and least for Sofa : 24.95%)

Q) Most common product by gender

select customergender , productname ,

count(customergender) as cnt

from assignment\_cleaned ac

group by 1,2

order by 3 desc

Analysis : Notebook is much common under Others category and Laptop is least common under Female category

Q) Avg feedback score of each product :

select productname , round(avg(feedbackscore),3) as avg\_feedback

from

assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : Avg feedback is highest for Sofa whereas lowest feedback given for Laptop

Q) Avg delivery time by Product

select productname , round(avg(deliverytimedays),3) as avg\_delivery\_days

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : Time taken to deliver sofa is max(11 days approx) whereas lowest for apple(2 days approx)

Q) No. of transactions made in each time of the day(Morning, afternoon, evening) per weekday

select

case

when transactiondate::time between '00:00' and '11:59' then 'Morning'

when transactiondate::time between '12:00' and '15:59' then 'Afternoon'

else 'Evening' end as time\_of\_date,

count(\*) as cnt

from assignment\_cleaned ac

group by 1

Analysis : Transactions made max in the morning and least during evenings

Q) City having good avg discount %

select city , avg(discountpercent) as avg\_discount\_percent

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : Max discount for Mumbai(25.07%) and least discount for Chennai(24.90%)

Q) Customer purchasing the most

select customerid, count(\*)

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : CustomerID 35173 and 39402 with 24 purchases in the year 2022 leaving aside Unknown customerid with count of 50k transactions

Q) Gender of most of the customers

select customergender , count(\*) as cnt

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : Other has max count of 1.75 lakhs with least for Female customers with count of 1.59 lakhs

Q) Gender distribution per city

select city , customergender , count(\*)

from assignment\_cleaned ac

group by 1,2

order by 3 desc

Analysis : In Kolkata -> Other gender has max entries of 20k whereas Other category from Chennai has least entries of 14.8k

Q) Feedbackscore analysis

select feedbackscore , count(\*) as feedback\_count

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : Feedback 4 is given for most of the transactions and least for rating 3

-------------------------------------------------- Customer insights --------------------------------------------

Q) Distinct customers in the data

**SELECT** **COUNT**(**DISTINCT** customerid) **AS** distinct\_customers

**FROM** assignment\_cleaned ac ;

Analysis : Around 48995 customers are there

Q) Time of Day when customers gives most feedbacks

**select**

**case**

**when** transactiondate::**time** **between** '00:00' **and** '11:59' **then** 'Morning'

**when** transactiondate::**time** **between** '12:00' **and** '15:59' **then** 'Afternoon'

**else** 'Evening' **end** **as** time\_of\_date,

**round**(**avg**(feedbackscore),3) **as** avg\_feedback

**from** assignment\_cleaned ac

**group** **by** 1

**order** **by** 2 **desc**

Analysis : Feedback is highest in the afternoon(3.01) and least during Mornings(2.99)

Q) Day of week has good ratings

**select**

**TO\_CHAR**(**TO\_TIMESTAMP**(transactiondate, 'YYYY-MM-DD'), 'DAY') **as** day\_name, **avg**(feedbackscore) **as** avg\_feedback

**from** assignment\_cleaned ac

**group** **by** day\_name

**order** **by** 2 **desc**

Analysis : Thursday has good ratings of 3.01 and Monday has worst rating of 2.99 on an average

-- Q) Help understand how much customers spend per order

**select**

**round**(**cast**(**SUM**(transactionamount)/**count**(**distinct** transactionid)**as** **numeric**),2) **as** avg\_order\_value

**from** assignment\_cleaned ac;

Analysis : Around 20392.4 per order

-- Q) Determine Sales by region/store

**select** region , **sum**(transactionamount) **as** Total\_Sum

**from** assignment\_cleaned ac

**group** **by** 1

**order** **by** 2 **desc**

-- Analysis : Sales in South region is the highest and lowest for West region

-- Q) Understand high value customers

**SELECT** customerid , **SUM**(transactionamount) **AS** total\_spent

**FROM** assignment\_cleaned ac

**GROUP** **BY** customerid

**ORDER** **BY** total\_spent **DESC**;

-- Analysis : Customer ID 32460 is highest value customer with sales of 8 lakhs

-- Best and worst selling Products(By Quantity and revenue)

**SELECT**

productname , **SUM**(quantity) **AS** total\_quantity, **SUM**(transactionamount) **AS** total\_revenue

**FROM** assignment\_cleaned ac

**GROUP** **BY** productname

**ORDER** **BY** total\_quantity **desc**

Analysis : Best selling pdt by revenue is Laptop and worst selling pdt by revenue is Apple, Best selling pdt by Quantity is Apple and worst selling pdt by quantity is Sofa

-- Q) Discount impact on Sales

**SELECT** discountpercent , **SUM**(transactionamount) **AS** total\_revenue

**FROM** assignment\_cleaned ac

**GROUP** **BY** discountpercent

**ORDER** **BY** discountpercent **desc**

Analysis : Helps to analyze effectiveness of discounts(with 49.99 and 50 % discount, Transaction Amount was quite high for the products)

-- Q) Peak sales hours

**SELECT** **EXTRACT**(**HOUR** **FROM** **cast**(transactiondate **as** **timestamp**)) **AS** sales\_hour, **SUM**(transactionamount) **AS** total\_revenue

**FROM** assignment\_cleaned ac

**GROUP** **BY** sales\_hour

**ORDER** **BY** total\_revenue **DESC**;

Analysis : Sales is at peak in 6th hour of the day

-- Q) Return rate per product

**SELECT** productname , **COUNT**(**CASE** **WHEN** returned = 'Yes' **THEN** 1 **END**) \* 100.0 / **COUNT**(\*) **AS** return\_rate\_percentage

**FROM** assignment\_cleaned ac

**GROUP** **BY** productname

-- Analysis : Max for Apple(50.05 %) and min for T-Shirt(49.65 %)

-------- Need to see loyalty points affect spending behavior, customer engagement, and retention -----

**SELECT**

**ROUND**(**AVG**(loyaltypoints), 2) **AS** avg\_loyalty\_points,

**COUNT**(**CASE** **WHEN** loyaltypoints **BETWEEN** 0 **AND** 3333 **THEN** 1 **END**) **AS** low\_tier\_customers,

**COUNT**(**CASE** **WHEN** loyaltypoints **BETWEEN** 3334 **AND** 6666 **THEN** 1 **END**) **AS** mid\_tier\_customers,

**COUNT**(**CASE** **WHEN** loyaltypoints > 6666 **THEN** 1 **END**) **AS** high\_tier\_customers

**FROM** assignment\_cleaned ac ;

Analysis : Mid-Tier customers are the max

**SELECT**

**CASE**

**WHEN** loyaltypoints **BETWEEN** 0 **AND** 3333 **THEN** 'Low Tier'

**WHEN** loyaltypoints **BETWEEN** 3334 **AND** 6666 **THEN** 'Mid Tier'

**ELSE** 'High Tier'

**END** **AS** loyalty\_tier,

**COUNT**(**DISTINCT** customerid) **AS** total\_customers,

**COUNT**(transactionid) **AS** total\_transactions,

**ROUND**(**AVG**(transactionamount::**numeric**), 2) **AS** avg\_transaction\_amount,

**SUM**(transactionamount) **AS** total\_spent

**FROM** assignment\_cleaned ac

**GROUP** **BY** loyalty\_tier

**ORDER** **BY** total\_spent **DESC**;

-- Customers in Mid tier are more engaged

-- Q) Top 10 customers based on loyalty points

**select** customerid , **avg**(loyaltypoints) **as** avg\_loyalty\_points

**from** assignment\_cleaned ac

**group** **by** 1

**order** **by** 2 **desc**

**limit** 10

-- Analysis : Max for Customer ID 30188 with loyalty points of 9621.5

-- Q) Total & Average Transaction Amounts Over Time

**SELECT**

**DATE**(transactiondate) **AS** transaction\_day,

**COUNT**(transactionid) **AS** total\_transactions,

**SUM**(transactionamount) **AS** total\_revenue,

**ROUND**(**AVG**(transactionamount::**numeric**), 2) **AS** avg\_transaction\_value

**FROM** assignment\_cleaned ac

**GROUP** **BY** transaction\_day

**ORDER** **BY** total\_revenue , avg\_transaction\_value ;

Analysis : Total revenue is max on 2022-09-01 and minimum on 2022-12-14

-- Q) Volume & Value of Transactions by Payment Method

**SELECT**

paymentmethod ,

**COUNT**(transactionid) **AS** transaction\_count,

**SUM**(transactionamount) **AS** total\_transaction\_value,

**ROUND**(**AVG**(transactionamount::**numeric**), 2) **AS** avg\_transaction\_value

**FROM** assignment\_cleaned ac

**GROUP** **BY** paymentmethod

**ORDER** **BY** total\_transaction\_value **DESC**;

Analysis : Debit card is popular payment method wrt Total Transaction value

-- Q) Impact of discount on Sales and Transaction frequency

**SELECT**

**CASE**

**WHEN** discountpercent = 0 **THEN** 'No Discount'

**WHEN** discountpercent **BETWEEN** 0 **AND** 10 **THEN** 'Low Discount (1-10%)'

**WHEN** discountpercent **BETWEEN** 10.01 **AND** 30 **THEN** 'Medium Discount (11-30%)'

**ELSE** 'High Discount (30%+)'

**END** **AS** discount\_category,

**COUNT**(transactionid) **AS** transaction\_count,

**SUM**(transactionamount) **AS** total\_sales,

**ROUND**(**AVG**(transactionamount ::**numeric**), 2) **AS** avg\_transaction\_value

**FROM** assignment\_cleaned ac

**GROUP** **BY** discount\_category

**ORDER** **BY** total\_sales **DESC**;

-- So for Medium and High discount Sales, avg sales is high (Higher discounts lead to more purchases & higher revenue)

**select** customerid , **count**(transactionid) **as** cnt

**from** assignment\_cleaned ac

**group** **by** 1

**order** **by** 2 **desc**

**SELECT**

customerid ,

**COUNT**(transactionid) **AS** total\_transactions,

**CASE**

**WHEN** **COUNT**(transactionid) >= 20 **THEN** 'Very Frequent Buyer'

**WHEN** **COUNT**(transactionid) **BETWEEN** 10 **AND** 19 **THEN** 'Frequent Buyer'

**WHEN** **COUNT**(transactionid) **BETWEEN** 5 **AND** 9 **THEN** 'Occasional Buyer'

**ELSE** 'Rare Buyer'

**END** **AS** customer\_segment

**FROM** assignment\_cleaned ac

**GROUP** **BY** customerid

**order** **by** total\_transactions **desc**

Analysis : 54 customers have transactions greater than 19

-------------- Analyze Shipping and Delivery - basing on region, pdt type and payment method -----------

**SELECT**

region ,

**ROUND**(**AVG**(shippingcost::**numeric**), 2) **AS** avg\_shipping\_cost,

**COUNT**(transactionid) **AS** total\_transactions

**FROM** assignment\_cleaned ac

**GROUP** **BY** region

**ORDER** **BY** avg\_shipping\_cost **DESC**;

Analysis : Most expensive shipping location is in North : less transactions and more shipping cost

-- Q) Avg shipping cost by Product

**SELECT**

productname ,

**ROUND**(**AVG**(shippingcost::**numeric**), 2) **AS** avg\_shipping\_cost,

**COUNT**(transactionid) **AS** total\_transactions

**from** assignment\_cleaned ac

**GROUP** **BY** productname

**ORDER** **BY** avg\_shipping\_cost **DESC**;

Analysis : Sofa has highest shipping cost in an average.

-- Q)Average Shipping Cost by Payment Method

**SELECT**

paymentmethod ,

**ROUND**(**AVG**(shippingcost ::**numeric**), 2) **AS** avg\_shipping\_cost,

**COUNT**(transactionid) **AS** total\_transactions

**FROM** assignment\_cleaned ac

**GROUP** **BY** paymentmethod

**ORDER** **BY** avg\_shipping\_cost **DESC**;

-- Analysis : Payment methods like Cash and Credit card lead to higher shipping costs and Debit card payment method leads to lower shipping cost

-- Q) Delivery Delays by Region

**SELECT**

region,

**COUNT**(transactionid) **AS** total\_transactions,

**ROUND**(**AVG**(deliverytimedays), 2) **AS** avg\_delay\_days

**FROM** assignment\_cleaned ac

**GROUP** **BY** region

**ORDER** **BY** avg\_delay\_days **DESC**;

Analysis: Transactions are highest in East region where delay is quite common with avg delay of 5.7 days

-- Q) Delivery delays by Product

**SELECT**

productname ,

**COUNT**(transactionid) **AS** total\_transactions,

**ROUND**(**AVG**(deliverytimedays), 2) **AS** avg\_delay\_days

**FROM** assignment\_cleaned ac

**GROUP** **BY** productname

**ORDER** **BY** avg\_delay\_days **DESC**;

Analysis : Sofa product tends to arrive quite late on an average of 11 days

Q) Number of new customers vs no of repeat customers each month on an average

**WITH** first\_purchase **AS** (

**SELECT**

customerid,

**MIN**(**DATE\_TRUNC**('month', transactiondate::**TIMESTAMP**)) **AS** first\_purchase\_month

**FROM** assignment\_cleaned

**GROUP** **BY** customerid

),

customer\_activity **AS** (

**SELECT**

**DISTINCT** t.customerid, -- Ensure only unique customers are counted

**DATE\_TRUNC**('month', t.transactiondate::**TIMESTAMP**) **AS** transaction\_month,

**CASE**

**WHEN** f.first\_purchase\_month = **DATE\_TRUNC**('month', t.transactiondate::**TIMESTAMP**)

**THEN** 'New Customer'

**ELSE** 'Repeat Customer'

**END** **AS** customer\_type

**FROM** assignment\_cleaned t

**JOIN** first\_purchase f

**ON** t.customerid = f.customerid

)

**SELECT**

transaction\_month,

**COUNT**(**DISTINCT** **CASE** **WHEN** customer\_type = 'New Customer' **THEN** customerid **END**) **AS** new\_customers,

**COUNT**(**DISTINCT** **CASE** **WHEN** customer\_type = 'Repeat Customer' **THEN** customerid **END**) **AS** repeat\_customers

**FROM** customer\_activity

**GROUP** **BY** transaction\_month

**ORDER** **BY** transaction\_month;

Analysis : Customer retention increased initially , was steady in mid part of year and then it dipped down in last month of the year(by 50%)

-- Q) Sales by Age group

**SELECT**

**CASE**

**WHEN** CustomerAge **BETWEEN** 18 **AND** 25 **THEN** '18-25'

**WHEN** CustomerAge **BETWEEN** 26 **AND** 35 **THEN** '26-35'

**WHEN** CustomerAge **BETWEEN** 36 **AND** 45 **THEN** '36-45'

**WHEN** CustomerAge **BETWEEN** 46 **AND** 55 **THEN** '46-55'

**WHEN** CustomerAge **BETWEEN** 56 **AND** 65 **THEN** '56-65'

**ELSE** '66+'

**END** **AS** AgeGroup,

**COUNT**(\*) **AS** customer\_count,

**SUM**(TransactionAmount) **AS** total\_sales

**FROM** assessment\_dataset\_original

**GROUP** **BY** AgeGroup

**ORDER** **BY** total\_sales **DESC**;

-- Analysis : Age group varies from 18 - 75. (66+ are the max people in group with max sales for 56-65 age category)