**Assignment Project**

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 Question**

Q) Payment method on the transactions

select paymentmethod , count(\*)

from assignment\_cleaned ac

group by 1

Analysis : Payment method is highest using DebitCard and lowest using CreditCard

Q) Quantity analysis

select max(quantity)

from assignment\_latest al

select min(quantity)

from assignment\_latest al

Analysis : Max qty is 50 whereas lowest is 1

-- Unique Cities, Products and Paymentmethod 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

-- Sales Analysis :

Q) Monthly Sales Trend - It shows Sales is highest in January and lowest in month of December

SELECT

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

SUM(TransactionAmount) AS TotalSales

FROM assignment\_latest al

GROUP BY Month\_Name

ORDER BY TotalSales desc;

Analysis : It shows Sales is highest in August 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(difference not being much)

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

Q) City selling more products than avg Product sold

select city, sum(quantity) as qty\_sold

from assignment\_cleaned ac

group by city

having sum(quantity) > (select avg(quantity) from assignment\_cleaned ac2)

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

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 for Mumbai and least for Chennai

Q) Customer purchasing the most

select customerid, count(\*)

from assignment\_cleaned ac

group by 1

order by 2 desc

Analysis : CustomerID 35173 and 39402 is highest with 24 purchases in the year 2022

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 gender customers 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) No 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%)