

Zepto E-commerce SQL Project

This is a complete, real-world data analyst portfolio project based on an e-commerce inventory dataset scraped from Zepto — one of India's fastest-growing quick-commerce startups. This project simulates real analyst workflows, from raw data exploration to business-focused data analysis.

Project Overview:

The goal is to simulate how actual data analysts in the e-commerce or retail industries work behind the scenes to use SQL to:

1. Set up a messy, real-world e-commerce inventory database
2. Perform Exploratory Data Analysis (EDA) to explore product categories, availability, and pricing inconsistencies
3. Implement Data Cleaning to handle null values, remove invalid entries, and convert pricing from paise to rupees
4. Write business-driven SQL queries to derive insights around pricing, inventory, stock availability, revenue and more

Dataset Overview:

The dataset was sourced from Kaggle and was originally scraped from Zepto's official product listings. It mimics what you'd typically encounter in a real-world e-commerce inventory system.

Each row represents a unique SKU (Stock Keeping Unit) for a product. Duplicate product names exist because the same product may appear multiple times in different package sizes, weights, discounts, or categories to improve visibility – exactly how real catalog data looks.

Column Name	Description
Sku_id	Unique identifier for each product entry (Synthetic Primary Key)
name	Product name as it appears on the app

category	Product categories like Fruits, Snacks, Beverages, etc.
mrp	Maximum Retail Price (originally in paise, converted to ₹)
discountPercent	Discount applied on MRP
discountedSellingPrice	Final price after discount (also converted to ₹)
availableQuantity	Units available in inventory
weightInGms	Product weight in grams
outOfStock	Boolean flag indicating stock availability
quantity	Number of units per package (mixed with grams for loose produce)

Project Workflow:

1. Database & Table Creation

We start with creating a SQL table with appropriate data types:

```
CREATE TABLE zepto (
  sku_id SERIAL PRIMARY KEY,
  category VARCHAR(120),
  name VARCHAR(150) NOT NULL,
  mrp NUMERIC(8,2),
  discountPercent NUMERIC(5,2),
  availableQuantity INTEGER,
  discountedSellingPrice NUMERIC(8,2),
  weightInGms INTEGER,
  outOfStock BOOLEAN,
  quantity INTEGER
);
```

2. Data Import

- Loaded CSV using pgAdmin's import feature. While importing we can use delimiter as , and UTF-8 encoding.

- I have faced encoding(UTF-8) issues while importing a CSV file, then i have save CSV file as CSV UTF-8 format.

3. Data Exploration

- Counted the total number of records in the dataset
- Viewed a sample of the dataset to understand structure and content
- Checked for null values across all columns
- Identified distinct product categories available in the dataset
- Compared in-stock vs out-of-stock product counts
- Detected products present multiple times, representing different SKUs

1. Count of Rows

```
SELECT COUNT(*) FROM zepto;
```

	count bigint
1	3732

2. Sample Data

```
SELECT * FROM zepto  
LIMIT 5;
```

Data Output Messages Notifications										
	sku_id [PK] integer	category character varying (120)	name character varying (150)	mrp numeric (8,2)	discountpercent numeric (5,2)	availablequantity integer	discountedsellingprice numeric (8,2)	weightingms integer	outofstock boolean	quantity integer
1	1	Fruits & Vegetables	Onion	2500.00	16.00	3	2100.00	1000	false	1
2	2	Fruits & Vegetables	Tomato Hybrid	4200.00	16.00	3	3500.00	1000	false	1
3	3	Fruits & Vegetables	Tender Coconut	5100.00	15.00	3	4300.00	58	false	1
4	4	Fruits & Vegetables	Coriander Leaves	2000.00	15.00	3	1700.00	100	false	100
5	5	Fruits & Vegetables	Ladies Finger	1400.00	14.00	3	1200.00	250	false	250

3. Checking Null Values

```
SELECT * FROM zepto  
WHERE name is NULL  
OR  
category is NULL  
OR  
mrp is NULL  
OR  
availableQuantity is NULL  
OR
```

discountedSellingPrice is NULL
OR
weightInGms is NULL
OR
outOfStock is NULL
OR
quantity is NULL;

4. *Different Product Categories*

```
SELECT DISTINCT category  
FROM zepto  
ORDER BY category;
```

	category character varying (120) 
1	Beverages
2	Biscuits
3	Chocolates & Candies
4	Cooking Essentials
5	Dairy, Bread & Batter
6	Fruits & Vegetables
7	Health & Hygiene
8	Home & Cleaning
9	Ice Cream & Desserts
10	Meats, Fish & Eggs
11	Munchies
12	Paan Corner
13	Packaged Food
14	Personal Care

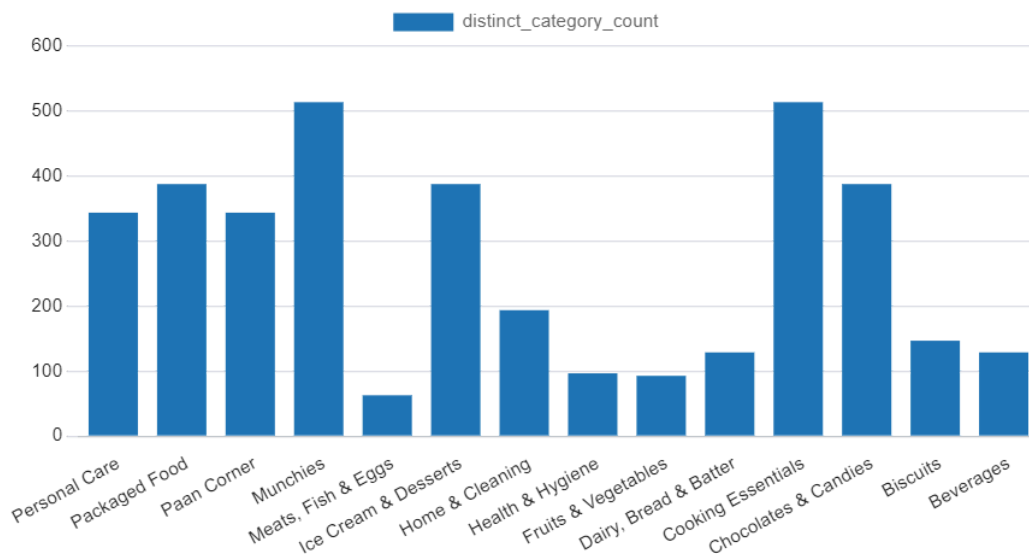
5. Count of different Product categories

```
SELECT COUNT(DISTINCT category) AS Distinct_category_count FROM zepto  
GROUP BY category
```

	distinct_category_count
1	14

Zepto provides delivery of 14 different categories like Beverages, Biscuits, Chocolates & Candles etc...

```
SELECT  
    category,  
    COUNT( category) AS distinct_category_count  
FROM zepto  
GROUP BY category  
ORDER BY category DESC;
```



6. Product in stock vs Out_Of_Stock

```
SELECT outOfStock, COUNT(*)  
FROM zepto
```

GROUP BY outOfStock;

	outofstock boolean	count bigint
1	false	3279
2	true	453

■ false ■ true



7. Product Name present in multiple times

```
SELECT name as Product_Name,  
       COUNT(*) as occurence_count  
FROM zepto  
GROUP BY name  
HAVING COUNT(*)>1  
ORDER BY 2 DESC;
```

Showing rows: 1 to 1000

Page No: 1 of 2

	product_name character varying (150)	occurence_count bigint
1	Sunfeast Yippee! Pasta Treat - Sour Cream Onion	10
2	Arden Eggs White	10
3	Quaker Oats	10
4	Saffola Veggie Twist Masala Oats	10
5	Amul Delicious Fat Spread - Cholesterol Free	10
6	Mother's Recipe Tamarind Paste	10
7	Kellogg's Real Almond & Honey Corn Flakes	9
8	Amul Fresh Cream	8

4. Data Cleaning

- Identified and removed rows where MRP or discount selling price was zero.
- Converted mrp and discountSellingPrice from paise to rupees for consistency and readability.

1. Products with price=0

```
SELECT * FROM zepto  
WHERE mrp=0 OR discountedSellingPrice=0
```

Data Output Messages Graph Visualiser X Notifications										
	sku_id [PK] integer	category character varying (120)	name character varying (150)	mrp numeric (8,2)	discountpercent numeric (5,2)	availablequantity integer	discountedsellingprice numeric (8,2)	weightingms integer	outofstock boolean	quantity integer
1	3607	Home & Cleaning	Cherry Blossom Liquid Shoe Polish Neutr...	0.00	0.00	1	0.00	75	false	7

There is a product with zero pricing that is not impossible in real time, so we can remove that row for better results.

2. Standardize the data

Some of the columns have inconsistency in data like mrp, discountedSellingPrice Columns data in paisa, so we need to convert that into rupees for data consistency.

UPDATE zepto

SET mrp=mrp/100.0,

discountedsellingprice = discountedSellingPrice/100.0;

SELECT mrp,discountedsellingprice FROM zepto;

	mrp numeric (8,2)	discountedsellingprice numeric (8,2)
1	25.00	21.00
2	42.00	35.00
3	30.00	29.00
4	50.00	44.00
5	60.00	60.00
6	425.00	383.00
7	91.00	82.00
8	97.00	97.00

5. Data Analysis / Business Insights

1. Find the top 10 best-value products based on the discount percentage.

SELECT

name as Product_Name,

discountPercent as Discount_Percentage

FROM zepto

ORDER BY 2 DESC

LIMIT 10

	product_name character varying (150)	discount_percentage numeric (5,2)
1	Dukes Waffy Chocolate Wafers	51.00
2	Dukes Waffy Orange Wafers	51.00
3	Dukes Waffy Strawberry Wafers	51.00
4	Chef's Basket Durum Wheat Penne Pasta	50.00
5	Chef's Basket Durum Wheat Fusilli Pasta	50.00
6	Ceres Foods Fish Mustard Instant Liquid Ma...	50.00
7	Chef's Basket Durum Wheat Elbow Pasta	50.00
8	Ceres Foods Laal Maas Instant Liquid Masala	50.00
9	Chef's Basket Durum Wheat Elbow Pasta	50.00
10	Ceres Foods Nalli Nihari Instant Liquid Masala	50.00

2. What are the Products with High MRP but Out of Stock

SELECT

name as product_name,

mrp

from zepto

WHERE outOfStock=true

ORDER BY mrp DESC

LIMIT 10;

Showing rows: 1 to 10

Page No: 1

of 1

	product_name character varying (150)	mrp numeric (8,2)
1	Patanjali Cow's Ghee	565.00
2	Patanjali Cow's Ghee	565.00
3	MamyPoko Pants Standard Diapers, Extra Large (12 - 17 kg)	399.00
4	MamyPoko Pants Standard Diapers, Extra Large (12 - 17 kg)	399.00
5	Aashirvaad Atta With Mutigrains	315.00
6	Aashirvaad Atta With Mutigrains	315.00
7	Everest Kashmiri Lal Chilli Powder	310.00
8	Everest Kashmiri Lal Chilli Powder	310.00
9	RRO Mozzarella Block Cheese	295.00
10	Madhur Pure And Hygienic Sugar	295.00

3. Calculate Estimated Revenue for each category


SELECT



category,


```

SUM(availableQuantity * discountedSellingPrice) AS total_revenue
FROM zepto
GROUP BY category
ORDER BY total_revenue;

```

Showing rows: 1 to 14  Page No: 1






	category character varying (120) 	total_revenue numeric 
1	Fruits & Vegetables	10846.00
2	Meats, Fish & Eggs	20693.00
3	Biscuits	25007.60
4	Beverages	55051.00
5	Dairy, Bread & Batter	55051.00
6	Health & Hygiene	64180.00
7	Home & Cleaning	122661.00
8	Ice Cream & Desserts	224385.00




4. Find all products where MRP is greater than ₹500 and discount is less than 10%.

```

SELECT
    name AS product_name,
    mrp,
    discountPercent
FROM zepto
WHERE mrp>500 AND discountPercent<10.00
ORDER BY mrp DESC,discountPercent DESC;

```

Showing rows: 1 to 82  Page No: 1 of 1    

	product_name character varying (150) 	mrp numeric (8,2) 	discountpercent numeric (5,2) 
1	Dhara Kachi Ghani Mustard Oil Jar	1250.00	8.00
2	Dhara Kachi Ghani Mustard Oil Jar	1250.00	8.00
3	Saffola Gold (Jar)	1240.00	0.00
4	Saffola Gold (Jar)	1240.00	0.00
5	Dhara Filtered Groundnut Oil (Jar)	1050.00	1.00
6	Fortune Rice Bran Health Oil (Jar)	1050.00	1.00


5. Identify the top 5 categories offering the highest average discount percentage.

```

SELECT
    category,
    ROUND(AVG(discountPercent),2) AS avg_discount
FROM zepto

```

GROUP BY category
ORDER BY avg_discount DESC
LIMIT 5;

Showing rows: 1 to 5  Page No: 1

	category character varying (120)	avg_discount numeric
1	Fruits & Vegetables	15.46
2	Meats, Fish & Eggs	11.03
3	Ice Cream & Desserts	8.32
4	Chocolates & Candies	8.32
5	Packaged Food	8.32

6. Find the price per gram for products above 100g and sort by best value.

```
SELECT
    name as product_name,
    weightInGms,
    discountedSellingPrice,
    ROUND(discountedSellingPrice/weightInGms,2) AS price_per_gram
FROM zepto
WHERE weightInGms >=100
ORDER BY price_per_gram DESC;
```

	product_name character varying (150)	weightInGms integer	discountedSellingPrice numeric (8,2)	price_per_gram numeric
1	Indulekha Bhringa Hair Oil	100	367.00	3.67
2	Indulekha Bhringa Hair Oil	100	367.00	3.67
3	L'Oreal Paris Excellence Creme Hair Color, 1 Black	172	620.00	3.60
4	L'Oreal Paris Excellence Creme Hair Color, 4 Natural Brown/Natural Dark Brown	172	620.00	3.60
5	L'Oreal Paris Excellence Creme Hair Color, 4.25 Aishwarya's Brown	172	620.00	3.60
6	L'Oreal Paris Excellence Creme Hair Color, 4 Natural Brown/Natural Dark Brown	172	620.00	3.60
7	L'Oreal Paris Excellence Creme Hair Color, 3 Dark Brown/Natural Darkest Brown	172	620.00	3.60
8	L'Oreal Paris Excellence Creme Hair Color, 4.25 Aishwarya's Brown	172	620.00	3.60

7. Group the products into categories like Low, Medium, Bulk.

```
SELECT
    DISTINCT name AS product_name,
    weightInGms,
    CASE
        WHEN weightInGms<1000 THEN 'Low'
        WHEN weightInGms<5000 THEN 'Medium'
        ELSE 'Bulk'
```

```

        END AS weight_category
FROM zepto;


```



8. What is the Total Inventory Weight Per Category

```

SELECT
    category,
    SUM(weightInGms * availableQuantity) AS total_weight
FROM zepto
GROUP BY category
ORDER BY total_weight DESC;

```

Showing rows: 1 to 14  Page No: 1

	category character varying (120) 	total_weight bigint 
1	Munchies	1404654
2	Cooking Essentials	1404654
3	Packaged Food	490797
4	Ice Cream & Desserts	490797
5	Chocolates & Candies	490797
6	Home & Cleaning	373161
7	Personal Care	348187
8	Paan Corner	348187

