

Zepto E-commerce

SQL Project

Introduction

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:

- Set up a messy, real-world e-commerce inventory database
- Perform Exploratory Data Analysis (EDA) to explore product categories, availability, and pricing inconsistencies
- Implement Data Cleaning to handle null values, remove invalid entries, and convert pricing from paise to rupees
- 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.



Columns:

- **sku_id:** Unique identifier for each product entry (Synthetic Primary Key)
- **name:** Product name as it appears on the app
- **category:** Product category 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)

Table Creation

```
drop table if exists zepto;

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 VARCHAR(10),
    quantity INTEGER
);
```

Data Exploration

- COUNT NO OF ROWS

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

	COUNT(*)
▶	1062

- DIFFERENT PRODUCT CATEGORIES

```
SELECT  
    category  
FROM  
    zepto  
GROUP BY category;
```

category
Fruits & Vegetables
Cooking Essentials
Munchies

• NULL VALUES

```
SELECT * FROM zepto
WHERE name IS NULL
OR
category IS NULL
OR
mrp IS NULL
OR
discountPercent IS NULL
OR
discountedSellingPrice IS NULL
OR
weightInGms IS NULL
OR
availableQuantity IS NULL
OR
outOfStock IS NULL
OR
quantity IS NULL;
```

- PRODUCTS IN STOCK VS OUT OF STOCK

```
SELECT
```

```
    outOfStock, COUNT(sku_id)
```

```
FROM
```

```
    zepto
```

```
WHERE
```

```
    outOfStock = 'TRUE';
```

outOfStock	COUNT(sku_id)
TRUE	79

- PRODUCT NAME PRESENT MULTIPLE TIMES

```
SELECT
```

```
    name, COUNT(sku_id) AS total_sku
```

```
FROM
```

```
    zepto
```

```
GROUP BY name
```

```
HAVING COUNT(sku_id) > 1
```

```
ORDER BY COUNT(sku_id) DESC;
```

name	total_sku
Everest Garam Masala	6
Everest Chicken Masala	6
E Everest Chicken Masala	6
Maggi Magic Cubes Extra Chicken	6
Arden Eggs White	4

DATA CLEANING

- **PRODUCTS WITH PRIZE = 0**

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

DELETE FROM zepto
WHERE
  mrp = 0 OR discountedSellingPrice = 0;
```

*	sku_id	category	name	mrp	discountPercent	availableQuantity	discountedSellingPrice	weightInGms	outOfStock	quantity
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

- **CONVERT PAISE INTO RUPEES**

```
UPDATE zepto
SET
  mrp = mrp / 100.0,
  discountedSellingPrice = discountedSellingPrice / 100.0;
```

DATA ANALYSIS

Q1. Find the top 10 best-value products based on the discount percentage

```
SELECT  
    DISTINCT name, mrp, discountPercent  
FROM  
    zepto  
ORDER BY discountPercent DESC  
LIMIT 5;
```

name	mrp	discountPercent
Dukes Waffy Chocolate Wafers	4500	51
Dukes Waffy Orange Wafers	4500	51
Dukes Waffy Strawberry Wafers	4500	51
Ceres Foods Fish Mustard Instant Liquid Masala	22000	50
Ceres Foods Laal Maas Instant Liquid Masala	22000	50

Q2. What are the Products with High MRP but Out of Stock

```
SELECT DISTINCT
    name, mrp
FROM
    zepto
WHERE
    mrp > 300 AND outOfStock = 'TRUE'
ORDER BY mrp DESC LIMIT 5;
```

name	mrp
Patanjali Cow's Ghee	56500
MamyPoko Pants Standard Diapers, Extra Large...	39900
Aashirvaad Atta With Mutigrains	31500
Everest Kashmiri Lal Chilli Powder	31000
Hershey's Cocoa + Almond Spread	29500

Q3. Calculate Estimated Revenue for each category

```
SELECT
    category,
    SUM(discountedSellingPrice * quantity) AS total_revenue
FROM
    zepto
GROUP BY category
ORDER BY total_revenue DESC LIMIT 5;
```

category	total_revenue
Packaged Food	1734991700
Ice Cream & Desserts	1734991700
Chocolates & Candies	1734991700
Cooking Essentials	1361323400
Munchies	1361323400

Q4. Find all products where MRP is greater than ₹500 and discount is less than 10%.

```
SELECT
    name, mrp, discountPercent
FROM
    zepto
WHERE
    mrp > 500 AND discountPercent < 10
ORDER BY mrp DESC , discountPercent DESC LIMIT 5;
```

name	mrp	discountPercent
Dhara Kachi Ghani Mustard Oil Jar	125000	8
Dhara Kachi Ghani Mustard Oil Jar	125000	8
Saffola Gold (Jar)	124000	0
Saffola Gold (Jar)	124000	0
Fortune Rice Bran Health Oil (Jar)	105000	1

Q5. 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;
```

category	avg_discount
Fruits & Vegetables	15.46
Meats, Fish & Eggs	11.03
Packaged Food	8.32
Ice Cream & Desserts	8.32
Chocolates & Candies	8.32

Q6. Find the price per gram for products above 100g and sort by best value.

```
SELECT  
    name,  
    weightInGms,  
    discountedSellingPrice,  
    ROUND(discountedSellingPrice / weightInGms, 2) AS price_per_gram  
FROM  
    zepto  
WHERE  
    weightInGms > 100  
ORDER BY price_per_gram DESC LIMIT 5;
```

name	weightInGms	discountedSellingPrice	price_per_gram
L'Oreal Paris Excellence Creme Hair Color, 4 Natural Brown	172	62000	360.47
L'Oreal Paris Excellence Creme Hair Color, 4.25 Black	172	62000	360.47
L'Oreal Paris Excellence Creme Hair Color, 1 Black	172	62000	360.47
L'Oreal Paris Excellence Creme Hair Color, 4 Natural Brown	172	62000	360.47
L'Oreal Paris Excellence Creme Hair Color, 4.25 Black	172	62000	360.47

Q7. Group the products into categories like Low, Medium, Bulk.

```
SELECT DISTINCT
    name,
    weightInGms,
    CASE
        WHEN weightInGms < 1000 THEN 'Low'
        WHEN weightInGms < 5000 THEN 'Medium'
        ELSE 'Bulk'
    END AS weight_category
FROM
    zepto;
```

name	weightInGms	weight_category
Onion	1000	Medium
Tomato Hybrid	1000	Medium
Tender Coconut	58	Low
Coriander Leaves	100	Low
Ladies Finger	250	Low
Potato	1000	Medium

Q8. 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;
```

category	total_weight
Meats, Fish & Eggs	48016
Biscuits	84431
Fruits & Vegetables	91794
Health & Hygiene	142904
Dairy, Bread & Batter	143735
Beveraodes	143735

Q9. For each product category, calculate the ratio of the Average Available Quantity for products that have a high discount (`discountPercent > 15`) versus those that have a low discount (`discountPercent <= 5`)

```
WITH CategoryStockRatios AS (
    SELECT
        category,
        -- Average available quantity for highly discounted items (> 15%)
        AVG(CASE WHEN discountPercent > 15 THEN availableQuantity ELSE NULL END) AS AvgStock_HighDiscount,
        -- Average available quantity for low discounted items (<= 5%)
        AVG(CASE WHEN discountPercent <= 5 THEN availableQuantity ELSE NULL END) AS AvgStock_LowDiscount
    FROM
        zepto
    GROUP BY
        category
)
```

```
SELECT  
    category,  
    AvgStock_HighDiscount / AvgStock_LowDiscount AS DiscountStockRatio  
FROM  
    CategoryStockRatios  
WHERE  
    AvgStock_LowDiscount IS NOT NULL AND AvgStock_LowDiscount > 0  
ORDER BY  
    DiscountStockRatio DESC;
```

category	DiscountStockRatio
Fruits & Vegetables	6.25920000
Biscuits	1.52851966
Packaged Food	1.27418629
Ice Cream & Desserts	1.27418629
Chocolates & Candies	1.27418629
Cooking Essentials	1.24208776

Q10. First, classify products into the following discount tiers:

- 'High Discount': if `discountPercent > 20`
- 'Medium Discount': if `discountPercent` is between 10 and 20 (inclusive)
- 'Low Discount': if `discountPercent < 10`

Then, restrict the analysis to only the 'High Discount' tier and find the product category that has the lowest average available quantity among these highly discounted products.

```
WITH DiscountedTiers AS
(
    SELECT category,
           availableQuantity,
           CASE
               WHEN discountPercent > 20 THEN 'High Discount'
               WHEN discountPercent BETWEEN 10 AND 20 THEN 'Medium Discount'
               WHEN discountPercent < 10 THEN 'Low Discount'
               ELSE 'Unknown'
           END AS DiscountTier
    FROM zepto
)
```

```
SELECT
    category,
    AVG(availableQuantity) AS AverageAvailableQuantity
FROM
    DiscountedTiers
WHERE
    DiscountTier = 'High Discount'
GROUP BY
    category
ORDER BY
    AverageAvailableQuantity ASC
LIMIT 1;
```

category	AverageAvailableQuantity
Meats, Fish & Eggs	2.8889

THANK YOU

