Zepto E-commerce SQL Data Analysis Project

This document outlines a complete, real-world data analyst portfolio project based on an e-commerce inventory dataset scraped from Zepto, a fast-growing quick-commerce startup in India. The project simulates real analyst workflows, from raw data exploration to business-focused data analysis, using SQL. It is designed for data analyst aspirants who want to build a strong portfolio project for interviews and those preparing for roles in retail, e-commerce, or product analytics.

Data Analyst Project Workflow

Made with Napkin

Project Overview

The primary goal is to simulate the work of data analysts in the e-commerce or retail industries, using 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.

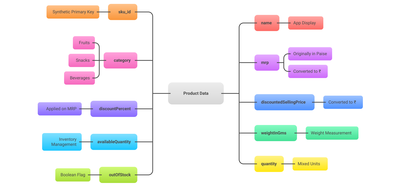
Dataset Overview

The dataset was sourced from Kaggle and originally scraped from Zepto's official product listings. It mimics 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.

Columns:

1. **sku_id**: Unique identifier for each product entry (Synthetic Primary Key)
2. **name**: Product name as it appears on the app
3. **category**: Product category like Fruits, Snacks, Beverages, etc.
4. **mrp**: Maximum Retail Price (originally in paise, converted to ₹)
5. **discountPercent**: Discount applied on MRP
6. **discountedSellingPrice**: Final price after discount (also converted to ₹)
7. **availableQuantity**: Units available in inventory
8. **weightInGms**: Product weight in grams
9. **outOfStock**: Boolean flag indicating stock availability
10. **quantity**: Number of units per package (mixed with grams for loose produce)

Zepto E-commerce Product Data Structure



Made with Napkin

Project Workflow

Here's a step-by-step breakdown of the project:

1. Database & Table Creation

We start by 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

The CSV data is loaded using pgAdmin's import feature.

Alternatively, if the import feature is not available, the following code can be used:

```
\copy zepto(category, name, mrp, discountPercent, availableQuantity,
discountedSellingPrice, weightInGms, outOfStock, quantity) FROM
'data/zepto_v2.csv' WITH (FORMAT csv, HEADER true, DELIMITER ',', QUOTE '"',
ENCODING 'UTF8');
```

Note: Encoding issues [UTF-8 error] may arise, which can be fixed by saving the CSV file using 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.

Example SQL queries for data exploration:

```
-- Count total records
SELECT COUNT(*) FROM zepto;

-- Sample the data
SELECT * FROM zepto LIMIT 10;

-- Check for null values
SELECT
    COUNT(*) FILTER (WHERE sku_id IS NULL) AS sku_id_nulls,
    COUNT(*) FILTER (WHERE category IS NULL) AS category_nulls,
    COUNT(*) FILTER (WHERE name IS NULL) AS name_nulls,
    COUNT(*) FILTER (WHERE mrp IS NULL) AS mrp_nulls,
    COUNT(*) FILTER (WHERE discountPercent IS NULL) AS discountPercent_nulls,
    COUNT(*) FILTER (WHERE availableQuantity IS NULL) AS
availableQuantity_nulls,
    COUNT(*) FILTER (WHERE discountedSellingPrice IS NULL) AS
discountedSellingPrice_nulls,
    COUNT(*) FILTER (WHERE weightInGms IS NULL) AS weightInGms_nulls,
    COUNT(*) FILTER (WHERE outOfStock IS NULL) AS outOfStock_nulls,
    COUNT(*) FILTER (WHERE quantity IS NULL) AS quantity_nulls
FROM zepto;

-- Distinct product categories
SELECT DISTINCT category FROM zepto;

-- In-stock vs out-of-stock
SELECT outOfStock, COUNT(*) FROM zepto GROUP BY outOfStock;

-- Products present multiple times
SELECT name, COUNT(*) FROM zepto GROUP BY name HAVING COUNT(*) > 1;
```

4. Data Cleaning

- Identified and removed rows where MRP or discounted selling price was zero.
- Converted mrp and discountedSellingPrice from paise to rupees for consistency and readability *(assuming original data was in paise)*.

Example SQL queries for data cleaning:

```
-- Remove rows where MRP or discounted selling price is zero
DELETE FROM zepto WHERE mrp = 0 OR discountedSellingPrice = 0;

-- Assuming original data was in paise, convert to rupees
-- No conversion needed if the data is already in rupees
-- UPDATE zepto SET mrp = mrp / 100, discountedSellingPrice =
discountedSellingPrice / 100;
```

5. Business Insights

- Found top 10 best-value products based on discount percentage.
- Identified high-MRP products that are currently out of stock.
- Estimated potential revenue for each product category.
- Filtered expensive products (MRP > ₹500) with minimal discount.
- Ranked top 5 categories offering highest average discounts.
- Calculated price per gram to identify value-for-money products.
- Grouped products based on weight into Low, Medium, and Bulk categories.
- Measured total inventory weight per product category.

Example SQL queries for business insights:

```
-- Top 10 best-value products based on discount percentage
SELECT name, discountPercent FROM zepto ORDER BY discountPercent DESC LIMIT 10;

-- High-MRP products that are currently out of stock
SELECT name, mrp FROM zepto WHERE outOfStock = TRUE ORDER BY mrp DESC;

-- Estimated potential revenue for each product category
SELECT category, SUM(discountedSellingPrice * availableQuantity) AS
potential_revenue FROM zepto GROUP BY category;

-- Expensive products (MRP > ₹500) with minimal discount (e.g., < 5%)
SELECT name, mrp, discountPercent FROM zepto WHERE mrp > 500 AND
discountPercent < 5;

-- Top 5 categories offering highest average discounts
SELECT category, AVG(discountPercent) AS avg_discount FROM zepto GROUP BY
category ORDER BY avg_discount DESC LIMIT 5;

-- Price per gram to identify value-for-money products
SELECT name, discountedSellingPrice / weightInGms AS price_per_gram FROM zepto
ORDER BY price_per_gram ASC;

-- Group products based on weight into Low, Medium, and Bulk categories
SELECT
  name,
  CASE
    WHEN weightInGms < 250 THEN 'Low'
    WHEN weightInGms >= 250 AND weightInGms < 1000 THEN 'Medium'
    ELSE 'Bulk'
  END AS weight_category
FROM zepto;

-- Total inventory weight per product category
```

How to Use This Project

1. **Open** zepto_SQL_data_analysis.sql

```
This file contains:

*   Table creation

*   Data exploration

*   Data cleaning

*   SQL Business analysis
```

2. **Load the dataset into pgAdmin or any other PostgreSQL client**
3. **Create a database and run the SQL file**
4. **Import the dataset (convert to UTF-8 if necessary)**