

# Exploratory Data Analysis (EDA) Report

Name: Pathan Afridh Khan

Date: 23 October 2025

**Dataset:** categories

**Domain:** E-Commerce Product Data

**Tool Used:** SQL (MySQL)

## Objective

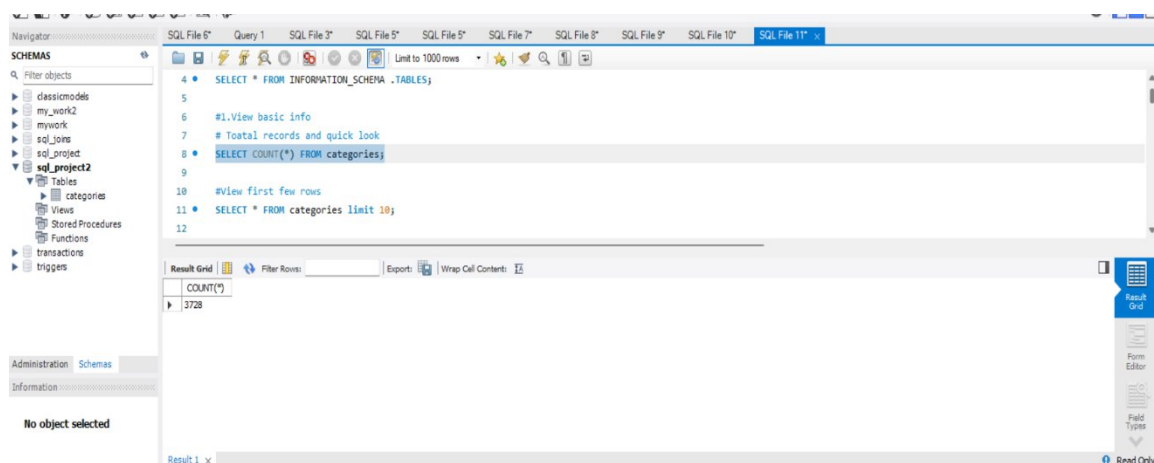
The main objective of this analysis is to explore and understand the product data in the `categories` table using SQL. This helps identify pricing patterns, discounts, stock availability, and category-wise trends. The insights obtained here will support better decision-making in product pricing and inventory management.

## Dataset Description

The `categories` table contains e-commerce product information with columns such as category, name, price, discount percentage, selling price, available quantity, stock, and weight. Each record represents a unique product along with its pricing and stock details.

### 1. Basic Information

This step involves checking the total number of records in the dataset and viewing sample data to understand the structure and content of the table.



### 2. Missing Values Check

In this step, the dataset is analyzed for null or missing values in each column. It helps identify incomplete data that may affect further analysis.

The screenshot shows a SQL IDE interface with a query editor and a result grid. The query is as follows:

```

SELECT COUNT(*) AS Total_rows,
SUP(CASE WHEN category is NULL THEN 1 ELSE 0 END) AS missing_category,
SUP(CASE WHEN name is NULL THEN 1 ELSE 0 END) AS missing_name,
SUP(CASE WHEN mrp is NULL THEN 1 ELSE 0 END) AS missing_mrp,
SUP(CASE WHEN discountPercent is NULL THEN 1 ELSE 0 END) AS missing_discountPercent,
SUP(CASE WHEN availableQuantity is NULL THEN 1 ELSE 0 END) AS missing_availableQuantity,
SUP(CASE WHEN discountedSellingPrice is NULL THEN 1 ELSE 0 END) AS missing_discountedSellingPrice,
SUP(CASE WHEN weightInGms is NULL THEN 1 ELSE 0 END) AS missing_weightInGms,
SUP(CASE WHEN outOfStock is NULL THEN 1 ELSE 0 END) AS missing_outOfStock,
SUP(CASE WHEN quantity is NULL THEN 1 ELSE 0 END) AS missing_quantity
FROM categories;

```

The result grid shows the following data:

Total_rows	missing_category	missing_name	missing_mrp	missing_discountPercent	missing_availableQuantity	missing_discountedSellingPrice	missing_weightInGms	missing_outOfStock	missing_quantity
3728	0	0	0	0	0	0	0	0	0

### 3. Duplicate Records Check

Duplicate entries are checked based on product name and category. Removing duplicates ensures that analysis results are accurate and not influenced by repeated records.

The screenshot shows a SQL IDE interface with a query editor and a result grid. The query is as follows:

```

#3.check for duplicates
SELECT category,name,COUNT(*) AS duplicate_count
FROM categories
GROUP BY category,name
HAVING COUNT(*)>1;

```

The result grid shows the following data:

category	name	duplicate_count
Fruits & Vegetables	Orion	2
Fruits & Vegetables	Potato	2
Fruits & Vegetables	Beetroot	2
Fruits & Vegetables	Carrot	2
Cooking Essentials	Arden Eggs White	2
Cooking Essentials	Amul Pure Ghee (Pouch)	2

### 4. Summary Statistics for Numerical Columns

This section summarizes the minimum, maximum, average, and standard deviation for key numerical columns such as price, discount percentage, and available quantity. The standard deviation (STDDEV) indicates how much the prices vary — a higher value means more variation among product prices.

The screenshot shows a SQL IDE interface with a query editor and a results grid. The query is as follows:

```

SELECT
  MIN(mrp) AS min_mrp,
  MAX(mrp) AS max_mrp,
  ROUND(avg(mrp), 2) AS avg_mrp,
  ROUND(STDDEV(mrp), 2) AS stddev_mrp,
  MIN(discountPercent) AS min_discountPercent,
  MAX(discountPercent) AS max_discountPercent,
  ROUND(avg(discountPercent), 2) AS avg_discountPercent,
  ROUND(avg(discountedSellingPrice), 2) AS avg_discountedSellingPrice,
  ROUND(avg(availableQuantity), 2) AS avg_availableQuantity
FROM categories;

```

The results grid displays the following data:

min_mrp	max_mrp	avg_mrp	stddev_mrp	min_discountPercent	max_discountPercent	avg_discountPercent	avg_discountedSellingPrice	avg_availableQuantity
0	260000	15693.19	16090.33	0	51	7.62	14204.42	4.01

## 5. Product Count per Category

Here, the total number of products in each category is counted. This helps identify which product categories have the highest or lowest representation in the dataset.

The screenshot shows a SQL IDE interface with a query editor and a results grid. The query is as follows:

```

-- #5. category_wise item count
SELECT category, COUNT(*) AS total_products
FROM categories
GROUP BY category
ORDER BY total_products DESC;

-- #6. Detect outliers in mrp(price)

```

The results grid displays the following data:

category	total_products
Cooking Essentials	512
Munchies	512
Packaged Food	388
Ice Cream & Desserts	388
Chocolates & Candies	388
Personal Care	344

## 6. Outlier Detection in Price

This step identifies products with prices significantly higher or lower than the average range. Outliers may indicate data entry errors or premium-priced products.

The screenshot shows a SQL IDE interface with a query editor and a results grid. The query is as follows:

```

-- #7. Average pricing and discount per category
SELECT category,

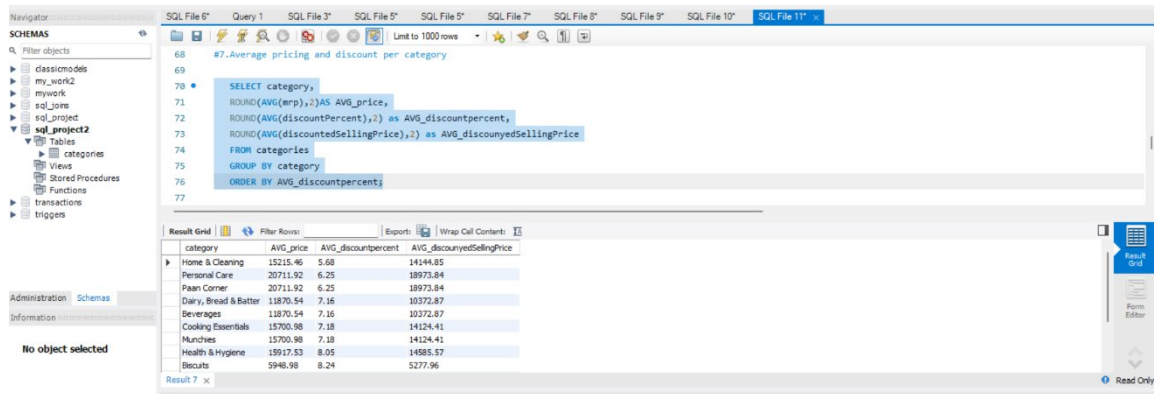
```

The results grid displays the following data:

Category	name	mrp	discountPercent	availableQuantity	discountedSellingPrice	weightInGms	outOfStock	quantity
Cooking Essentials	Fortune Sunlite Refined Sunflower (Jar)	92500	0	6	92500	5000	FALSE	5
Cooking Essentials	Fortune Soyabean Oil	100500	0	6	99900	5000	FALSE	5
Cooking Essentials	Dhara Health Refined Sun Flower Oil Jar	120000	18	6	98400	5000	FALSE	5
Cooking Essentials	Dhara Kachi Ghani Mustard Oil Jar	125000	8	6	114500	5000	FALSE	5
Cooking Essentials	Saffola Gold (Jar)	124000	0	6	124000	5000	FALSE	5
Cooking Essentials	Fortune Rice Bran Health Oil (Jar)	105000	1	6	103950	5000	FALSE	5
Cooking Essentials	Dhara Filtered Groundnut Oil (Jar)	105000	1	6	103500	5000	FALSE	5
Cooking Essentials	Popular Essentials Californian Almond	69500	32	6	47000	500	FALSE	500
Cooking Essentials	Prakritik Natural Desi Gir Cow A2 Ghee	145000	10	6	130500	540	FALSE	540

## 7. Average Pricing and Discount per Category

Average prices, selling prices, and discount percentages are calculated for each category. This reveals how pricing and discount strategies differ across product groups.



The screenshot shows a SQL IDE interface with a query editor and a result grid. The query is titled "#7. Average pricing and discount per category" and is as follows:

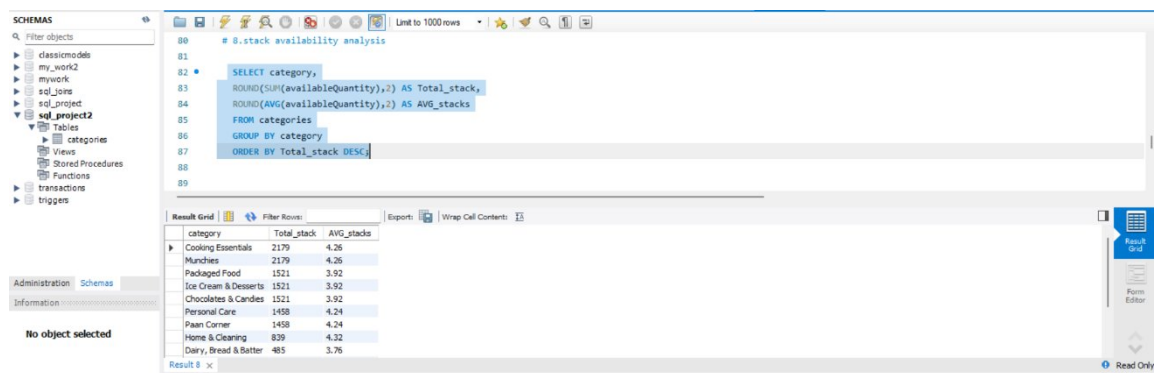
```
SELECT category,
ROUND(AVG(mrp),2) AS AVG_price,
ROUND(AVG(discountPercent),2) AS AVG_discountpercent,
ROUND(AVG(discountedSellingPrice),2) AS AVG_discountedSellingPrice
FROM categories
GROUP BY category
ORDER BY AVG_discountpercent;
```

The result grid displays the following data:

category	AVG_price	AVG_discountpercent	AVG_discountedSellingPrice
Home & Cleaning	15215.46	5.68	14144.85
Personal Care	20711.92	6.25	18973.84
Pain Corner	20711.92	6.25	18973.84
Dairy, Bread & Better	11870.54	7.16	10372.87
Beverages	11870.54	7.16	10372.87
Cooking Essentials	15700.98	7.18	14124.41
Munchies	15700.98	7.18	14124.41
Health & Hygiene	15917.53	8.05	14585.57
Biscuits	5946.98	8.24	5277.96

## 8. Stock Availability Analysis

Stock levels are analyzed category-wise to identify which categories have higher or lower product availability. This supports effective inventory planning.



The screenshot shows a SQL IDE interface with a query editor and a result grid. The query is titled "#8. stock availability analysis" and is as follows:

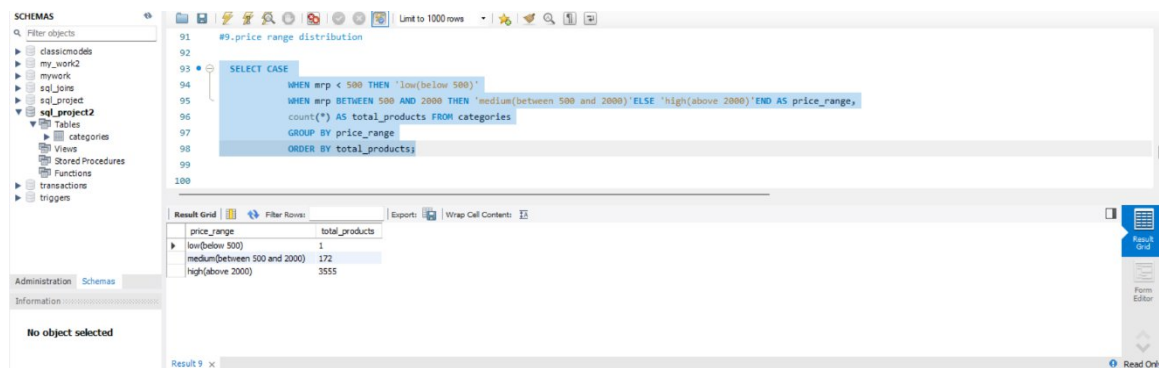
```
SELECT category,
ROUND(SUM(availableQuantity),2) AS Total_stack,
ROUND(AVG(availableQuantity),2) AS AVG_stacks
FROM categories
GROUP BY category
ORDER BY Total_stack DESC;
```

The result grid displays the following data:

category	Total_stack	AVG_stacks
Cooking Essentials	2179	4.26
Munchies	2179	4.26
Packaged Food	1521	3.92
Ice Cream & Desserts	1521	3.92
Chocolates & Candies	1521	3.92
Personal Care	1458	4.24
Pain Corner	1458	4.24
Home & Cleaning	839	4.32
Dairy, Bread & Better	485	3.76

## 9. Price Range Distribution

Products are grouped into different price ranges such as low, medium, and high. This gives an overview of how products are distributed across various price levels.



The screenshot shows a SQL IDE interface with a query editor and a result grid. The query is titled "#9. price range distribution" and is as follows:

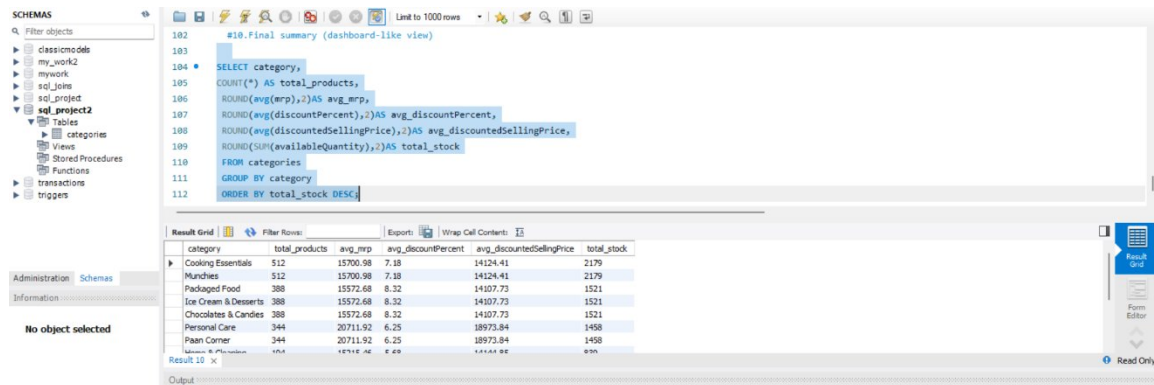
```
SELECT CASE
WHEN mrp < 500 THEN 'low(below 500)'
WHEN mrp BETWEEN 500 AND 2000 THEN 'medium(between 500 and 2000)' ELSE 'high(above 2000)' END AS price_range,
count(*) AS total_products
FROM categories
GROUP BY price_range
ORDER BY total_products;
```

The result grid displays the following data:

price_range	total_products
low(below 500)	1
medium(between 500 and 2000)	172
high(above 2000)	3555

## 10.Final Summary View

A combined view of total products, average price, discount, and stock per category is presented. It acts as a summary dashboard of the entire dataset.



The screenshot shows a SQL IDE interface. On the left, a 'SCHEMAS' pane lists database objects like 'classmodels', 'my\_work2', 'mywork', 'sql\_join', 'sql\_project', 'sql\_project2', 'categories', 'views', 'stored\_procedures', 'functions', 'transactions', and 'triggers'. The main editor displays a SQL query labeled '#10.Final summary (dashboard-like view)'. The query calculates total products, average MRP, average discount percentage, average discounted selling price, and total stock for each category. Below the query, a 'Result Grid' shows the output for 10 rows. The columns are 'category', 'total\_products', 'avg\_mrp', 'avg\_discountPercent', 'avg\_discountedSellingPrice', and 'total\_stock'. The categories listed are Cooking Essentials, Mundies, Packaged Food, Ice Cream & Desserts, Chocolates & Candies, Personal Care, and Paan Corner.

```
102 #10.Final summary (dashboard-like view)
103
104 SELECT category,
105 COUNT(*) AS total_products,
106 ROUND(avg(mrp),2)AS avg_mrp,
107 ROUND(avg(discountPercent),2)AS avg_discountPercent,
108 ROUND(avg(discountedSellingPrice),2)AS avg_discountedSellingPrice,
109 ROUND(SUM(availableQuantity),2)AS total_stock
110 FROM categories
111 GROUP BY category
112 ORDER BY total_stock DESC;
```

category	total_products	avg_mrp	avg_discountPercent	avg_discountedSellingPrice	total_stock
Cooking Essentials	512	15700.98	7.18	14124.41	2179
Mundies	512	15700.98	7.18	14124.41	2179
Packaged Food	388	15572.68	8.32	14107.73	1521
Ice Cream & Desserts	388	15572.68	8.32	14107.73	1521
Chocolates & Candies	388	15572.68	8.32	14107.73	1521
Personal Care	344	20711.92	6.25	18973.84	1458
Paan Corner	344	20711.92	6.25	18973.84	1458

## Key Insights

- Some categories offer heavy discounts, where the selling price is less than 70% of the MRP.
- High standard deviation in price indicates a wide range of products in that category.
- Certain categories have larger product counts and stock, indicating popular or high-demand items.
- Outliers detected may represent premium or incorrectly priced products.

## Conclusion

The EDA performed using SQL helped uncover important insights into product pricing, discounts, and stock patterns. The analysis supports decision-making for price optimization, discount strategies, and inventory management in the e-commerce domain.