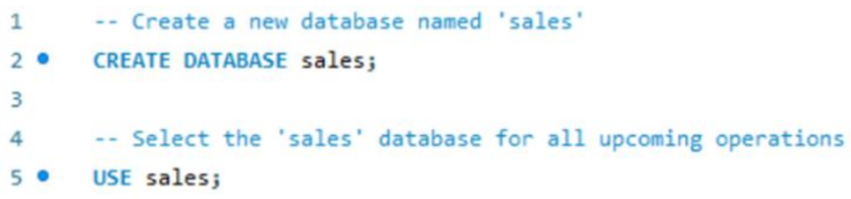
Week 2 - Data Cleaning, Analysis, and Business Insights

Objective

To clean messy retail sales data, analyse trends, and create visual insights using SQL (MySQL) and Power BI for better decision-making.

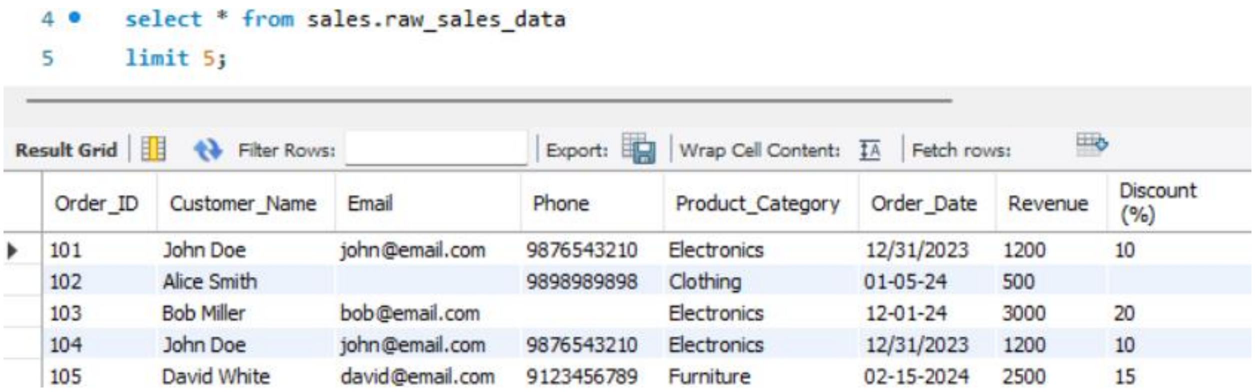
Database Initialization



```
-- Create a new database named 'sales'
CREATE DATABASE sales;

-- Select the 'sales' database for all upcoming operations
USE sales;
```

Data Inspection

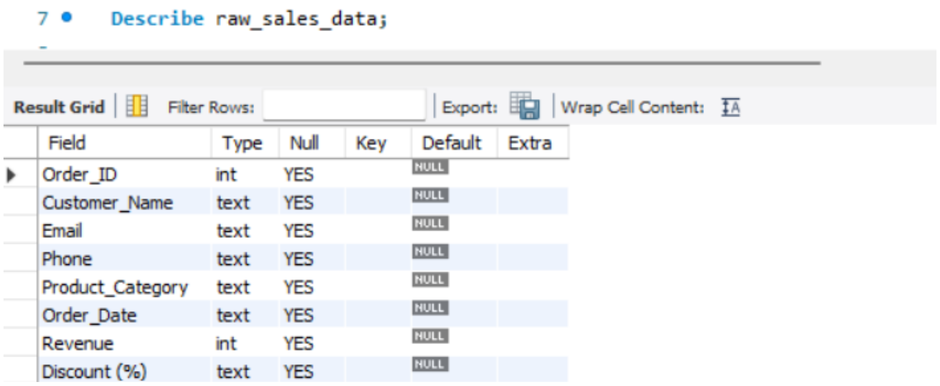
1. Used *Limit 5*; to view first 5 entries.



```
select * from sales.raw_sales_data
limit 5;
```

Order_ID	Customer_Name	Email	Phone	Product_Category	Order_Date	Revenue	Discount (%)
101	John Doe	john@email.com	9876543210	Electronics	12/31/2023	1200	10
102	Alice Smith		9898989898	Clothing	01-05-24	500	
103	Bob Miller	bob@email.com		Electronics	12-01-24	3000	20
104	John Doe	john@email.com	9876543210	Electronics	12/31/2023	1200	10
105	David White	david@email.com	9123456789	Furniture	02-15-2024	2500	15

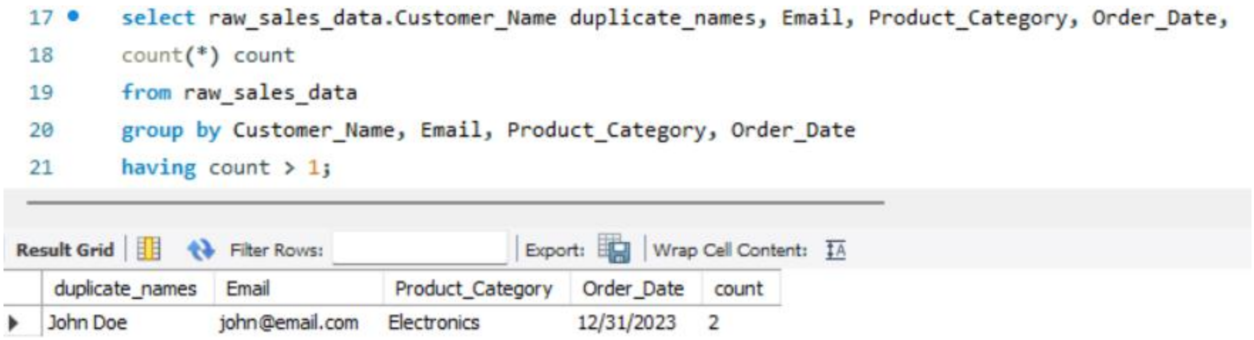
2. Used *Describe;* to see column names, types and nullability.



```
Describe raw_sales_data;
```

Field	Type	Null	Key	Default	Extra
Order_ID	int	YES		NULL	
Customer_Name	text	YES		NULL	
Email	text	YES		NULL	
Phone	text	YES		NULL	
Product_Category	text	YES		NULL	
Order_Date	text	YES		NULL	
Revenue	int	YES		NULL	
Discount (%)	text	YES		NULL	

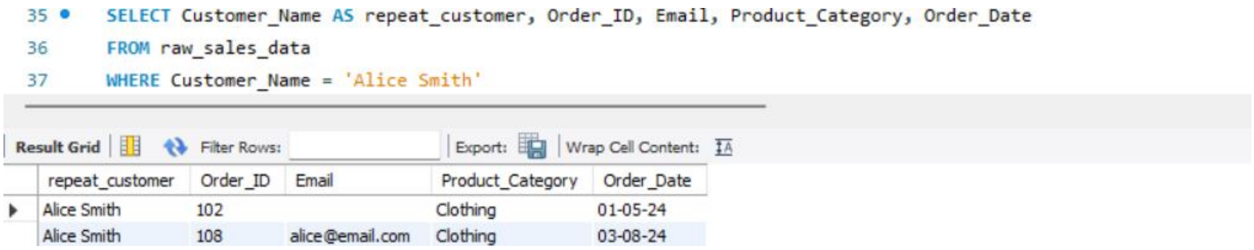
3. Used *Group by* and *Having* to get count of duplicates.



```sql
select raw_sales_data.Customer_Name duplicate_names, Email, Product_Category, Order_Date,
count(*) count
from raw_sales_data
group by Customer_Name, Email, Product_Category, Order_Date
having count > 1;
```

duplicate_names	Email	Product_Category	Order_Date	count
John Doe	john@email.com	Electronics	12/31/2023	2

4. Repeat customer purchases.



```sql
SELECT Customer_Name AS repeat_customer, Order_ID, Email, Product_Category, Order_Date
FROM raw_sales_data
WHERE Customer_Name = 'Alice Smith'
```

repeat_customer	Order_ID	Email	Product_Category	Order_Date
Alice Smith	102		Clothing	01-05-24
Alice Smith	108	alice@email.com	Clothing	03-08-24

Key Observations:

1. **Missing values**:

- Email column had blanks → replaced with 'not_provided@gmail.com'
- Phone column had blanks → replaced with 'Unknown'
- Discount (%) column had NULL values → replaced with 'Unknown'

2. **Inconsistent Date Format**:

- Multiple formats (e.g., MM/DD/YYYY, DD-MM-YYYY) → standardized using STR_TO_DATE()

3. **Duplicate Records**:

- John Doe had duplicate rows → deleted duplicates, kept first using MIN(Order_ID)

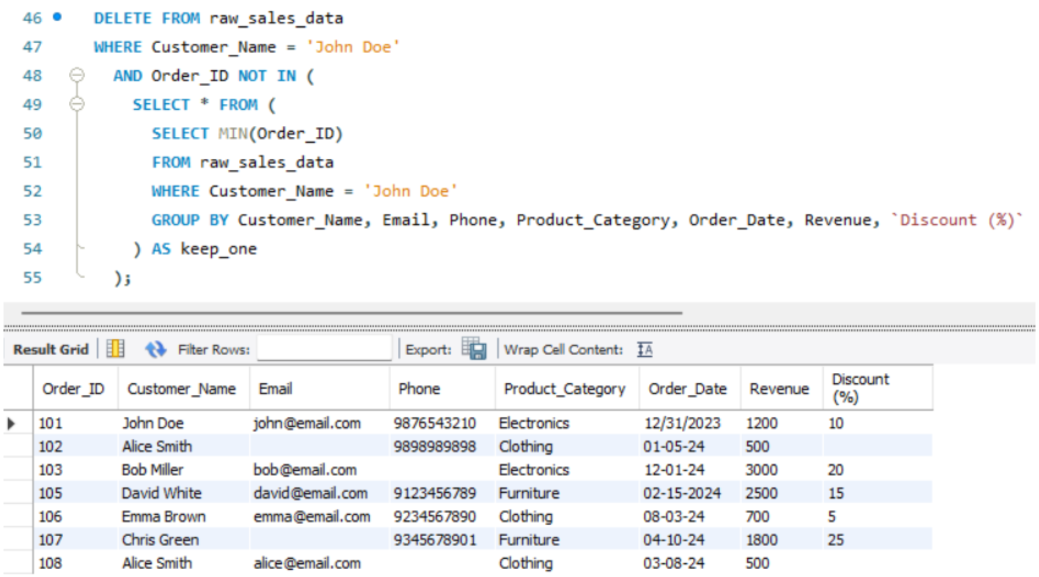
4. **Repeat Customers**:

- Alice Smith made multiple purchases on different dates → useful for behavioral customer insights.

Data Cleaning

1. Delete duplicates

   Deleted all but the first occurrence of duplicate entries based on *MIN(Order_ID)* .

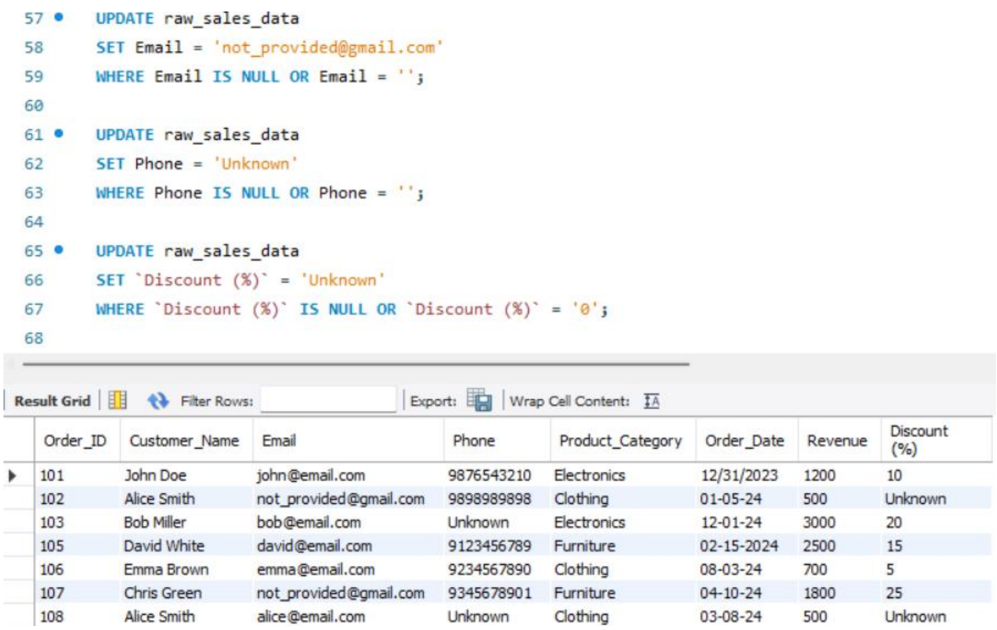


```sql
DELETE FROM raw_sales_data
WHERE Customer_Name = 'John Doe'
  AND Order_ID NOT IN (
    SELECT * FROM (
      SELECT MIN(Order_ID)
      FROM raw_sales_data
      WHERE Customer_Name = 'John Doe'
      GROUP BY Customer_Name, Email, Phone, Product_Category, Order_Date, Revenue, `Discount (%)`
    ) AS keep_one
  );
```

Order_ID	Customer_Name	Email	Phone	Product_Category	Order_Date	Revenue	Discount (%)
101	John Doe	john@email.com	9876543210	Electronics	12/31/2023	1200	10
102	Alice Smith		9898989898	Clothing	01-05-24	500	
103	Bob Miller	bob@email.com		Electronics	12-01-24	3000	20
105	David White	david@email.com	9123456789	Furniture	02-15-2024	2500	15
106	Emma Brown	emma@email.com	9234567890	Clothing	08-03-24	700	5
107	Chris Green		9345678901	Furniture	04-10-24	1800	25
108	Alice Smith	alice@email.com		Clothing	03-08-24	500	

2. Handle missing values/nulls

   Used *UPDATE* and *SET* to fill email, phone and Discount entries.

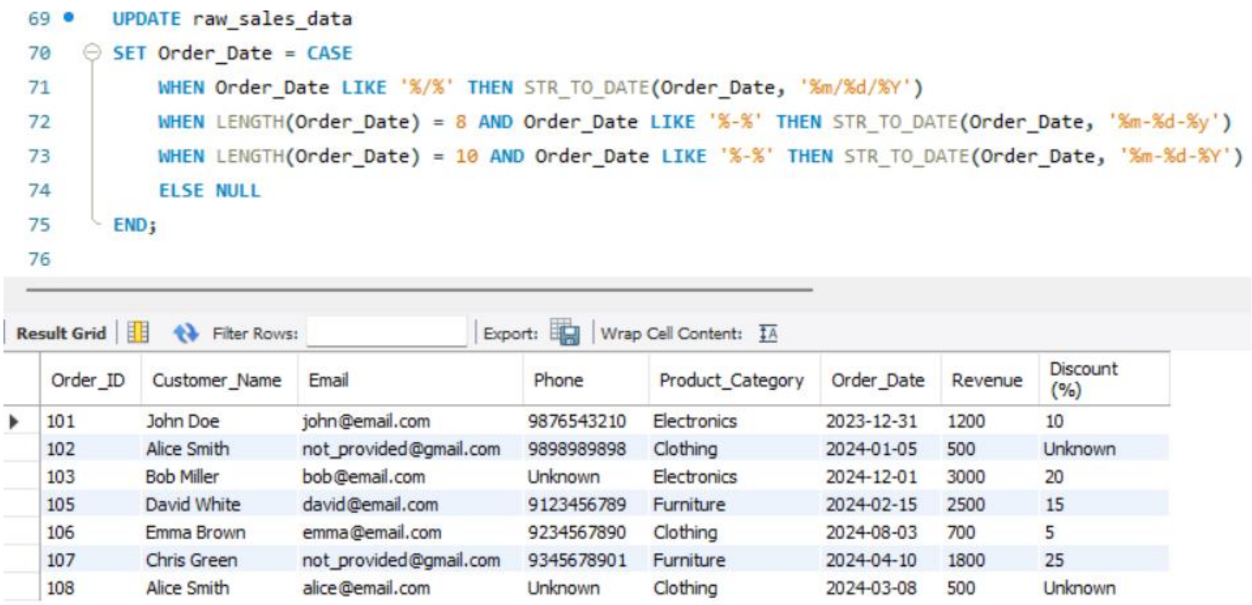


```sql
UPDATE raw_sales_data
SET Email = 'not_provided@gmail.com'
WHERE Email IS NULL OR Email = '';

UPDATE raw_sales_data
SET Phone = 'Unknown'
WHERE Phone IS NULL OR Phone = '';

UPDATE raw_sales_data
SET `Discount (%)` = 'Unknown'
WHERE `Discount (%)` IS NULL OR `Discount (%)` = '0';

```

Order_ID	Customer_Name	Email	Phone	Product_Category	Order_Date	Revenue	Discount (%)
101	John Doe	john@email.com	9876543210	Electronics	12/31/2023	1200	10
102	Alice Smith	not_provided@gmail.com	9898989898	Clothing	01-05-24	500	Unknown
103	Bob Miller	bob@email.com	Unknown	Electronics	12-01-24	3000	20
105	David White	david@email.com	9123456789	Furniture	02-15-2024	2500	15
106	Emma Brown	emma@email.com	9234567890	Clothing	08-03-24	700	5
107	Chris Green	not_provided@gmail.com	9345678901	Furniture	04-10-24	1800	25
108	Alice Smith	alice@email.com	Unknown	Clothing	03-08-24	500	Unknown

3. Fixing date format

   Converted various date formats to a consistent DATE format using *STR_TO_DATE()*.

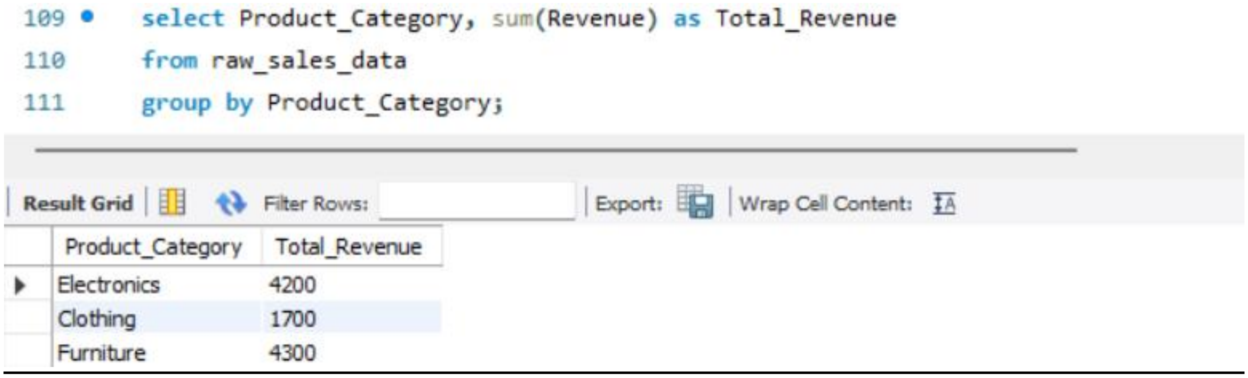


```sql
UPDATE raw_sales_data
SET Order_Date = CASE
    WHEN Order_Date LIKE '%/%' THEN STR_TO_DATE(Order_Date, '%m/%d/%Y')
    WHEN LENGTH(Order_Date) = 8 AND Order_Date LIKE '%-%' THEN STR_TO_DATE(Order_Date, '%m-%d-%y')
    WHEN LENGTH(Order_Date) = 10 AND Order_Date LIKE '%-%' THEN STR_TO_DATE(Order_Date, '%m-%d-%Y')
    ELSE NULL
END;
```

Order_ID	Customer_Name	Email	Phone	Product_Category	Order_Date	Revenue	Discount (%)
101	John Doe	john@email.com	9876543210	Electronics	2023-12-31	1200	10
102	Alice Smith	not_provided@gmail.com	9898989898	Clothing	2024-01-05	500	Unknown
103	Bob Miller	bob@email.com	Unknown	Electronics	2024-12-01	3000	20
105	David White	david@email.com	9123456789	Furniture	2024-02-15	2500	15
106	Emma Brown	emma@email.com	9234567890	Clothing	2024-08-03	700	5
107	Chris Green	not_provided@gmail.com	9345678901	Furniture	2024-04-10	1800	25
108	Alice Smith	alice@email.com	Unknown	Clothing	2024-03-08	500	Unknown

Data Exploration & Aggregation

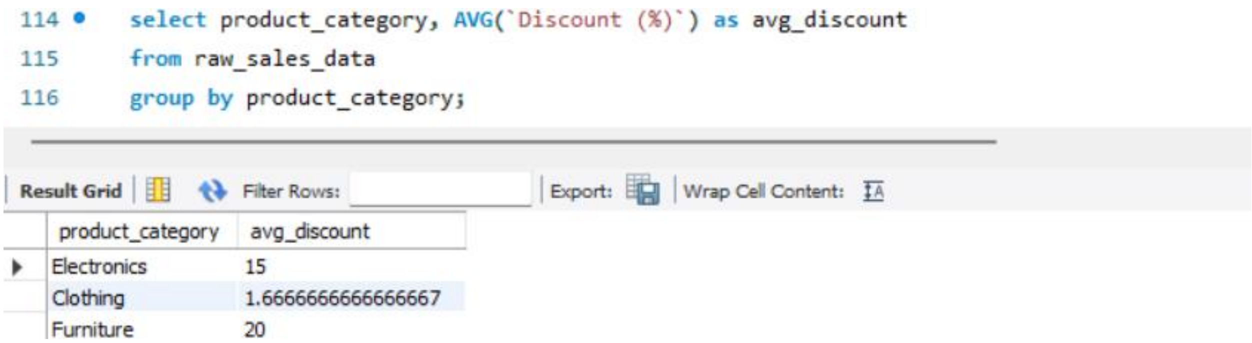
1. **Total Revenue by Product Category**
   Used *GROUP BY* Product_Category to identify the most profitable segments.



```sql
select Product_Category, sum(Revenue) as Total_Revenue
from raw_sales_data
group by Product_Category;
```

Product_Category	Total_Revenue
Electronics	4200
Clothing	1700
Furniture	4300

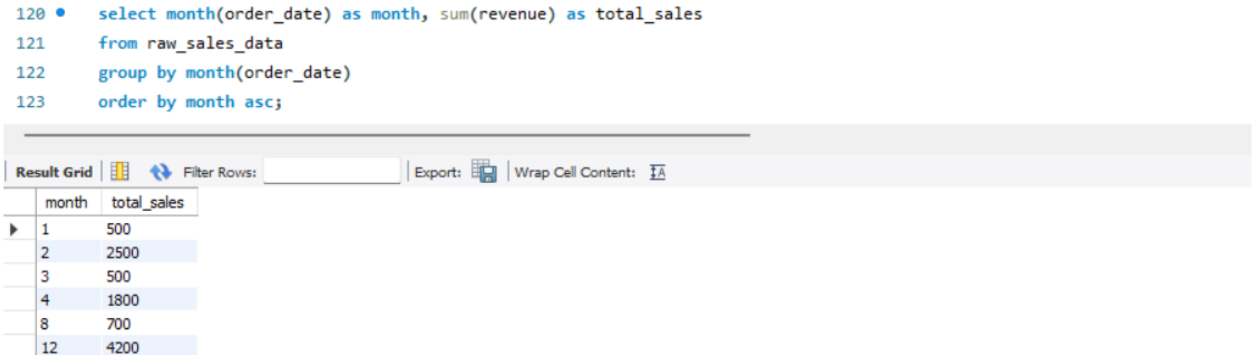
2. **Average Discount by Product Category**
   Used *AVG* to calculate average discount across categories to assess promotional strategies.



```
select product_category, AVG(`Discount (%)`) as avg_discount
from raw_sales_data
group by product_category;
```

product_category	avg_discount
Electronics	15
Clothing	1.6666666666666667
Furniture	20

3. **Monthly Sales Trends**
   Aggregated revenue by month to identify high and low-performing periods using *GROUP BY* and *ORDER BY ASC*.



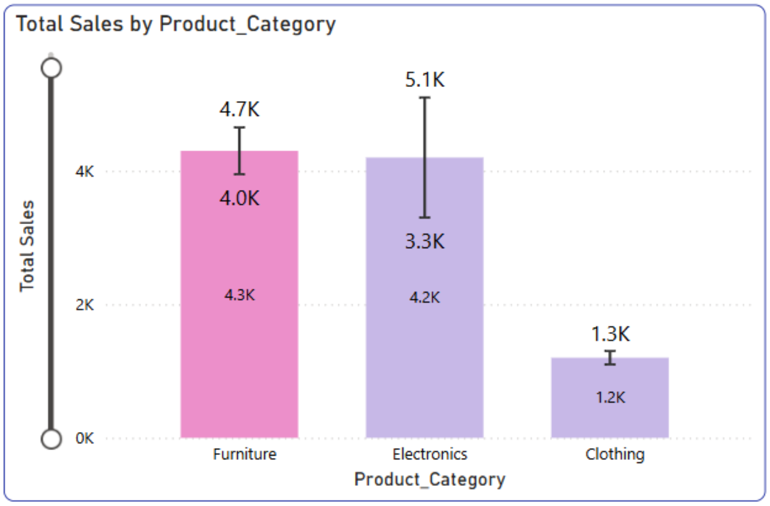
```
select month(order_date) as month, sum(revenue) as total_sales
from raw_sales_data
group by month(order_date)
order by month asc;
```

month	total_sales
1	500
2	2500
3	500
4	1800
8	700
12	4200

Visualizations and Insights

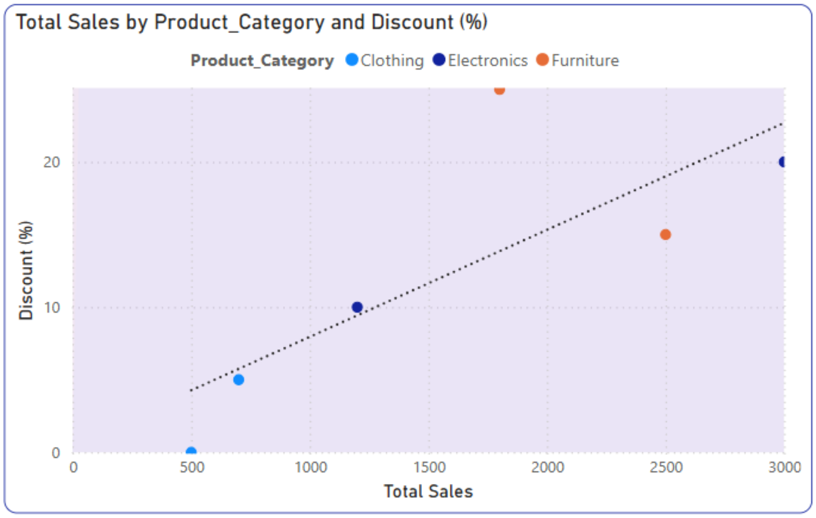
1. Bar Chart: Total Sales by Product Category

- Furniture (44%,) and Electronics (43%) have the highest total sales.
- Clothing (12%) lags far behind in revenue.
- Variability (error bars) indicates some fluctuation in monthly sales per category.



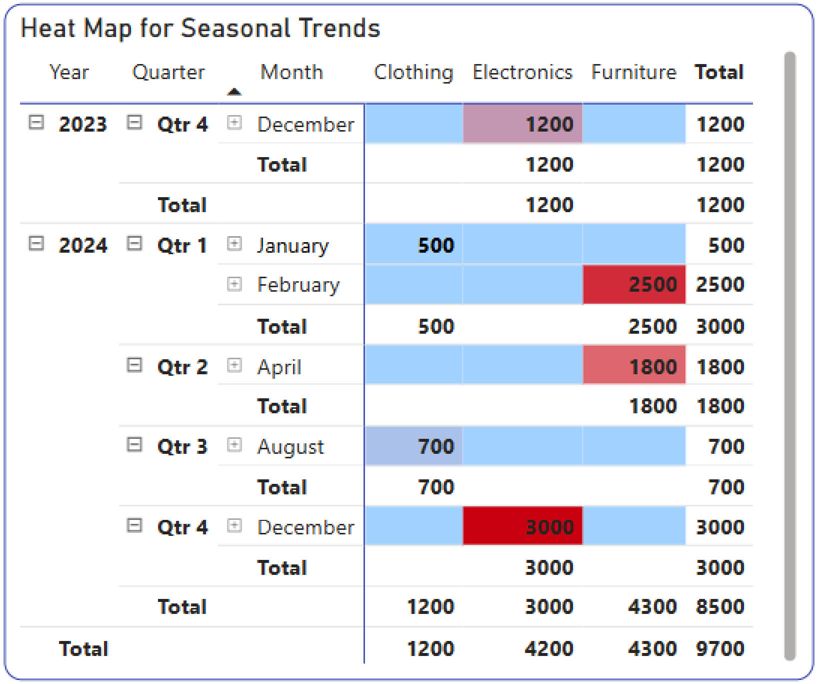
2. Scatter Plot: Total Sales vs Discount (%)

- Strong positive correlation between discounts and sales volume.
- Categories like Furniture show high sales at higher discounts.
- Electronics also benefit from larger discounts.
- Lower discounts (0–5%) result in low sales, especially for Clothing.



3. Heat Map: Seasonal Trends by Category

- Massive sales growth observed in 2024 compared to 2023 (8,500 vs 1,200).
- Furniture leads in 2024 with the highest category sales (4,300 units); shows strong seasonal demand in February (2,500 units) and April (1,800).
- Electronics peaks in December 2024 with 3,000 units, making it the highest single-category sales point.
- Clothing maintains steady but lower sales, with a minor peak in August (700).
- January has the lowest overall sales (500) across all categories.
- Quarter 4 of 2024 drives most of the year’s total due to Electronics surge.
- Color gradient highlights sharp increases in 2024 with deep red tones, signaling strong performance across all categories.



Heat Map for Seasonal Trends

Year	Quarter	Month	Clothing	Electronics	Furniture	Total
2023	Qtr 4	December		1200		1200
		Total		1200		1200
	Total			1200		1200
2024	Qtr 1	January	500			500
		February			2500	2500
		Total	500		2500	3000
	Qtr 2	April			1800	1800
		Total			1800	1800
	Qtr 3	August	700			700
		Total	700			700
	Qtr 4	December		3000		3000
		Total		3000		3000
	Total		1200	3000	4300	8500
Total			1200	4200	4300	9700

Summary Report: Key Findings & Recommendations

Sales analysis reveals that Furniture (44%) and Electronics (43%) dominate total revenue, while Clothing (12%) significantly underperforms. Sales volumes increase with higher discounts, especially for Furniture and Electronics, indicating strong price sensitivity. Low discounts (0–5%) yield minimal sales, particularly in Clothing.

In 2024, total sales jumped from 1,200 to 8,500 units, with peaks in February (Furniture), April (Furniture), and December (Electronics). Clothing showed a modest rise in August, but overall remains weak. Quarter 4 drives the year's total, while January is the lowest-performing month.

Issue with trends:
We lack data for the **first three quarters of 2023**, limiting year-over-year trend analysis.

Recommendations:

- Prioritize discount-driven promotions for Furniture and Electronics.
- Boost seasonal campaigns in Q1 and Q4 such as winter or spring promotion sales.
- Reassess Clothing strategy to improve.
- Prepare inventory for high-demand months, especially December.