Project Data Cleaning, Analysis, and Business Insights

Summary

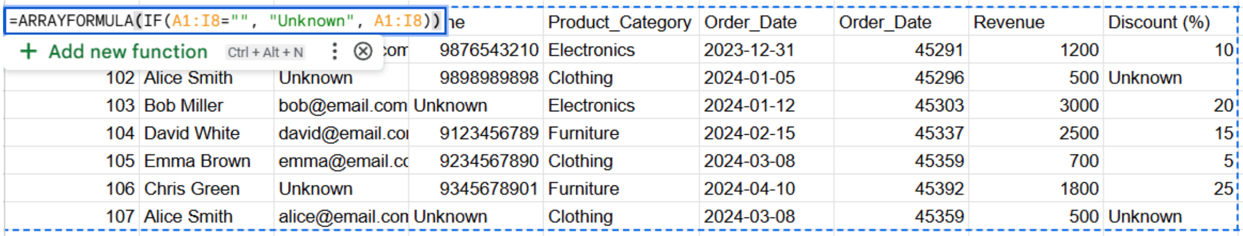
Google Sheets: The dataset contains data about the sales of certain products based on their types, customers and their contact information such as phone numbers and email addresses, the dates of the order, revenue, and discount. Data Cleaning has been performed using **Google Sheets.** Missing values have been replaced with 'Unknown', There was one duplicate row and it has been removed. The 'Order_Date' column has been modified. As a result, the dataset contains 7 columns, including a header row and 6 rows with different types of values.

MySQL: The cells in the Email column that contain **'Unknown'** were replaced with **'not_provided@email.com'.** A few queries have been performed to have clear insights: According to the results, **clothing** was the item that has been **ordered the most**, indicating order count as 3 for both, while **electronics and furniture** were the ones which the company made **profit the most** (4200 and 4300 respectively). By calculating **Average Discount,** we managed to see both unique items and their average discount (**electronics 15%, furniture 20%, and clothing 2%).** The **total sales** hit **the peak** in **January and February. Bob Miller** and **David White** were the top customers. **Total Revenue** is **18.33% higher** than **discounted revenue** (respectively 10200 and 8620)

Power BI: The **revenue** generated **from furniture sales** is **slightly higher than** that generated from **electronics sales** in terms of both the sum and average of the revenue. Another point is that **the higher the discount** was, **the more profit** has been made (3000 with 20%). Bob Miller has chosen electronics over other items, while Davide White has preferred buying furniture, which shows the most expensive products bought by the top two customers.

Google Sheet:

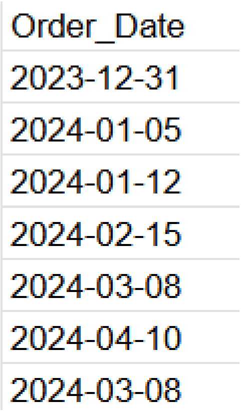
1. **Data Cleanup > Remove Duplicates -** Duplicates were removed with this method
2. The formula: **=ARRAYFORMULA(IF(A1:I8="", "Unknown", A1:I8)) -** used to replace missing values with the word '**Unknown**'



=ARRAYFORMULA(IF(A1:I8="", "Unknown", A1:I8))

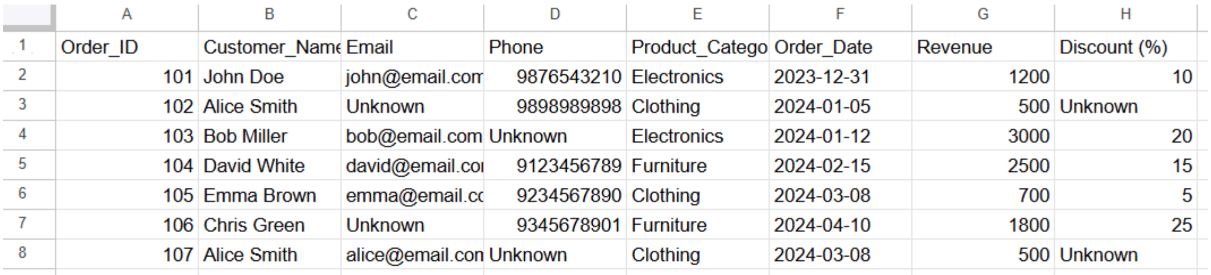
			he	Product_Category	Order_Date	Order_Date	Revenue	Discount (%)
		com	9876543210	Electronics	2023-12-31	45291	1200	10
102	Alice Smith	Unknown	9898989898	Clothing	2024-01-05	45296	500	Unknown
103	Bob Miller	bob@email.com	Unknown	Electronics	2024-01-12	45303	3000	20
104	David White	david@email.coı	9123456789	Furniture	2024-02-15	45337	2500	15
105	Emma Brown	emma@email.cc	9234567890	Clothing	2024-03-08	45359	700	5
106	Chris Green	Unknown	9345678901	Furniture	2024-04-10	45392	1800	25
107	Alice Smith	alice@email.con	Unknown	Clothing	2024-03-08	45359	500	Unknown

3. The formula: **=ARRAYFORMULA(IF(G1:G8="", "Unknown", TEXT(G1:G8, "yyyy-mm-dd"))) -** used to modify the column: **Order_Date**



Order_Date
2023-12-31
2024-01-05
2024-01-12
2024-02-15
2024-03-08
2024-04-10
2024-03-08

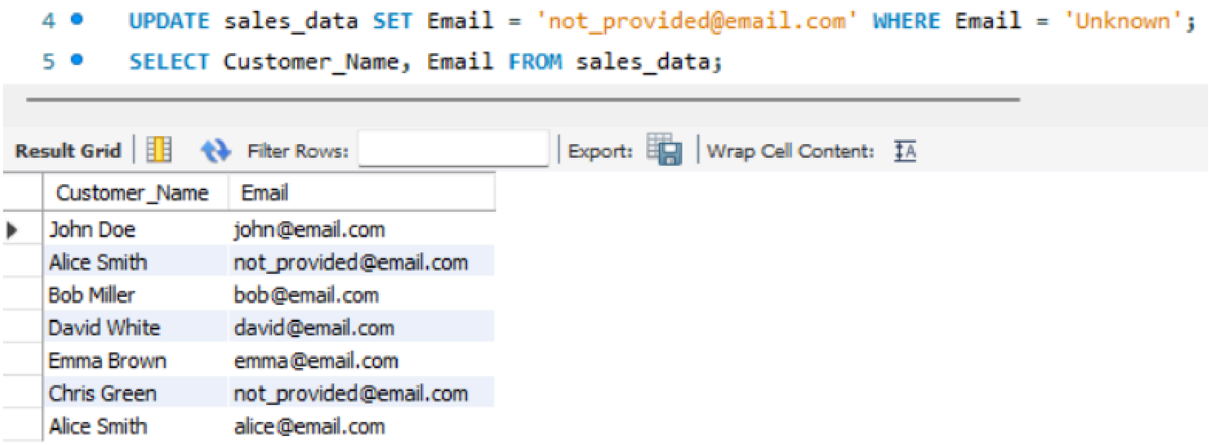
4. Finally, the column was integrated with the table and the old one was deleted:



	A	B	C	D	E	F	G	H
1	Order_ID	Customer_Name	Email	Phone	Product_Catego	Order_Date	Revenue	Discount (%)
2	101	John Doe	john@email.com	9876543210	Electronics	2023-12-31	1200	10
3	102	Alice Smith	Unknown	9898989898	Clothing	2024-01-05	500	Unknown
4	103	Bob Miller	bob@email.com	Unknown	Electronics	2024-01-12	3000	20
5	104	David White	david@email.coı	9123456789	Furniture	2024-02-15	2500	15
6	105	Emma Brown	emma@email.cc	9234567890	Clothing	2024-03-08	700	5
7	106	Chris Green	Unknown	9345678901	Furniture	2024-04-10	1800	25
8	107	Alice Smith	alice@email.con	Unknown	Clothing	2024-03-08	500	Unknown

MySQL (sales_data):

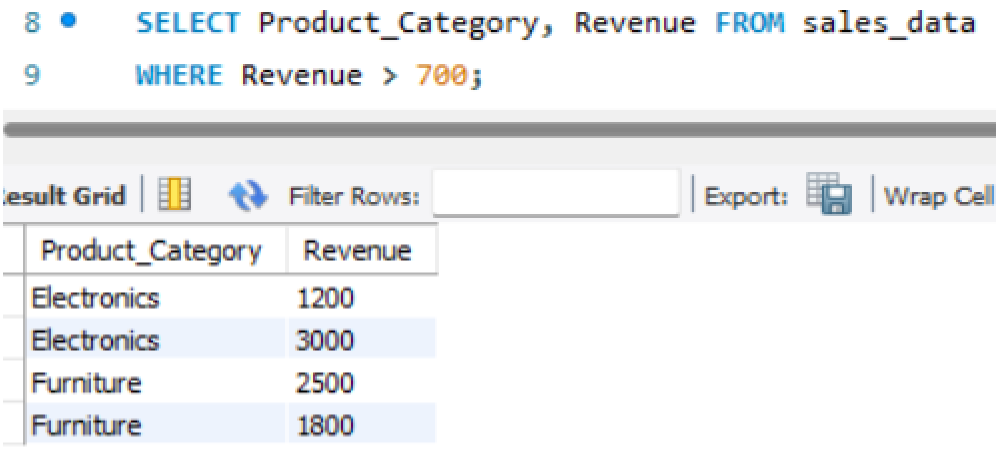
1. **‘Unknown’** cells were changed into ‘**not_provided@email.com**’:



```sql
UPDATE sales_data SET Email = 'not_provided@email.com' WHERE Email = 'Unknown';
SELECT Customer_Name, Email FROM sales_data;
```

Customer_Name	Email
John Doe	john@email.com
Alice Smith	not_provided@email.com
Bob Miller	bob@email.com
David White	david@email.com
Emma Brown	emma@email.com
Chris Green	not_provided@email.com
Alice Smith	alice@email.com

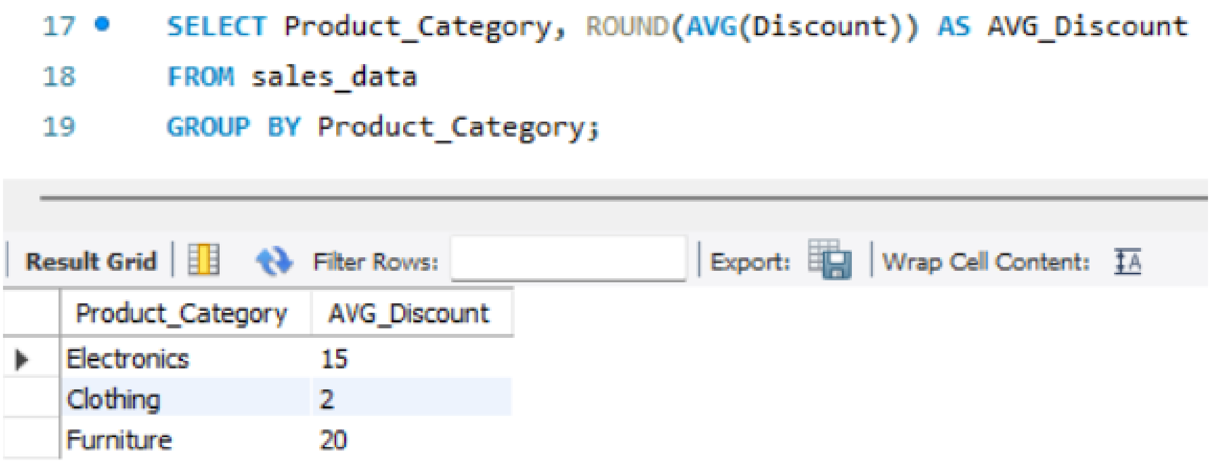
2. **Revenue per product was examined making a condition:**



```sql
SELECT Product_Category, Revenue FROM sales_data
WHERE Revenue > 700;
```

Product_Category	Revenue
Electronics	1200
Electronics	3000
Furniture	2500
Furniture	1800

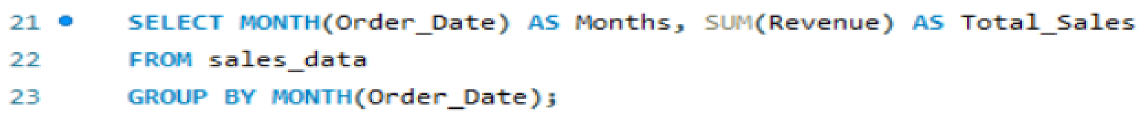
3. **Average Discount Calculation:**



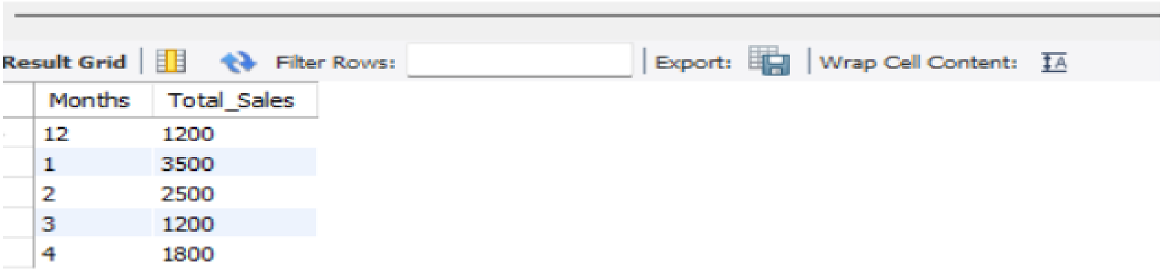
```sql
SELECT Product_Category, ROUND(AVG(Discount)) AS AVG_Discount
FROM sales_data
GROUP BY Product_Category;
```

Product_Category	AVG_Discount
Electronics	15
Clothing	2
Furniture	20

4. **Total Sales are shown in the table according to the months:**

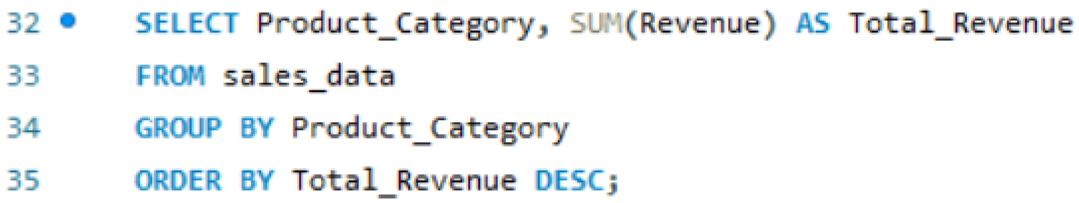


```sql
SELECT MONTH(Order_Date) AS Months, SUM(Revenue) AS Total_Sales
FROM sales_data
GROUP BY MONTH(Order_Date);
```

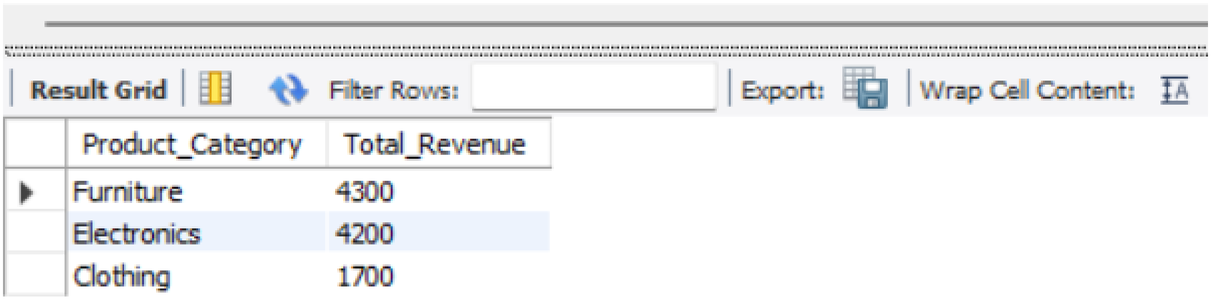


Months	Total_Sales
12	1200
1	3500
2	2500
3	1200
4	1800

5. **Best-Selling Products by Revenue:**

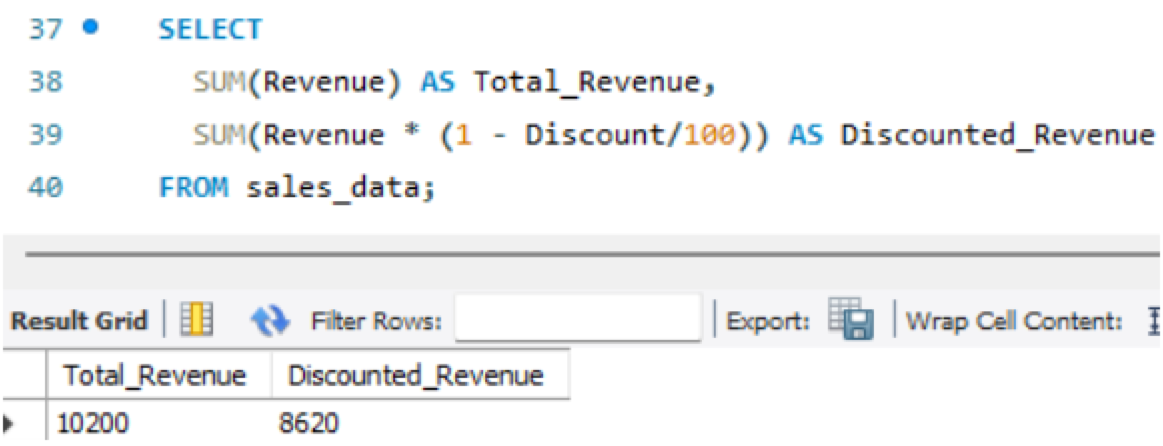


```sql
SELECT Product_Category, SUM(Revenue) AS Total_Revenue
FROM sales_data
GROUP BY Product_Category
ORDER BY Total_Revenue DESC;
```



Product_Category	Total_Revenue
Furniture	4300
Electronics	4200
Clothing	1700

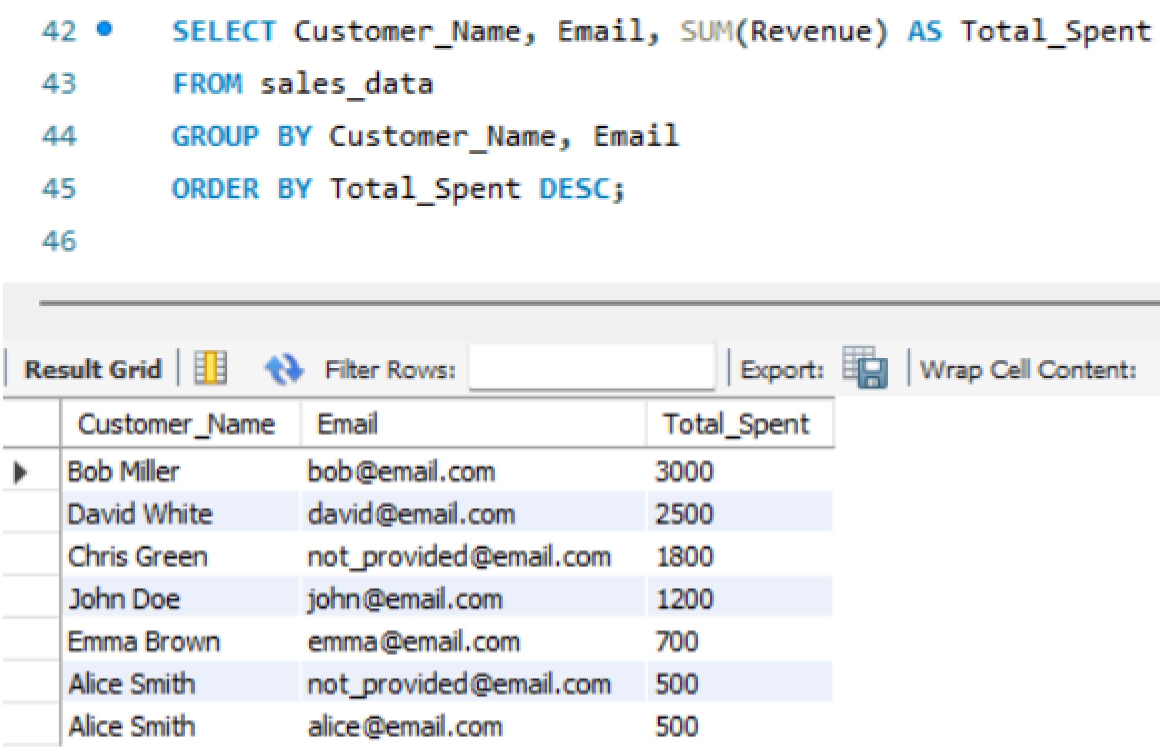
6. Total Revenue vs Discounted Revenue:



```
SELECT
  SUM(Revenue) AS Total_Revenue,
  SUM(Revenue * (1 - Discount/100)) AS Discounted_Revenue
FROM sales_data;
```

Total_Revenue	Discounted_Revenue
10200	8620

7. Top Customers by Revenue:

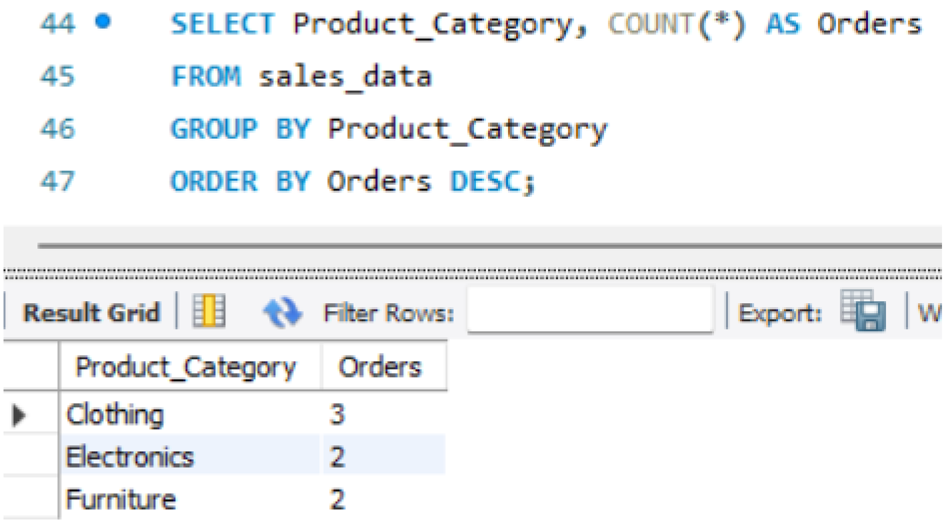


```
SELECT Customer_Name, Email, SUM(Revenue) AS Total_Spent
FROM sales_data
GROUP BY Customer_Name, Email
ORDER BY Total_Spent DESC;

```

Customer_Name	Email	Total_Spent
Bob Miller	bob@email.com	3000
David White	david@email.com	2500
Chris Green	not_provided@email.com	1800
John Doe	john@email.com	1200
Emma Brown	emma@email.com	700
Alice Smith	not_provided@email.com	500
Alice Smith	alice@email.com	500

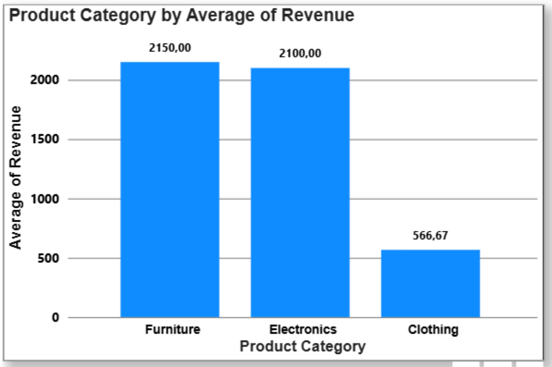
8. **How many times was each item ordered:**

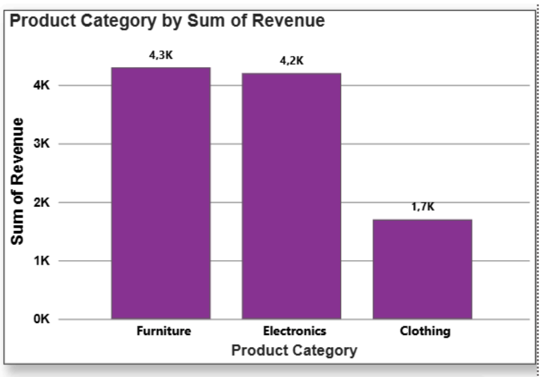


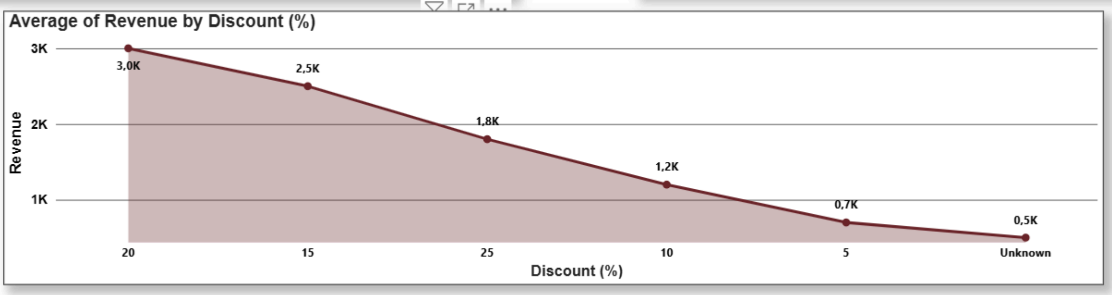
```
SELECT Product_Category, COUNT(*) AS Orders
FROM sales_data
GROUP BY Product_Category
ORDER BY Orders DESC;
```

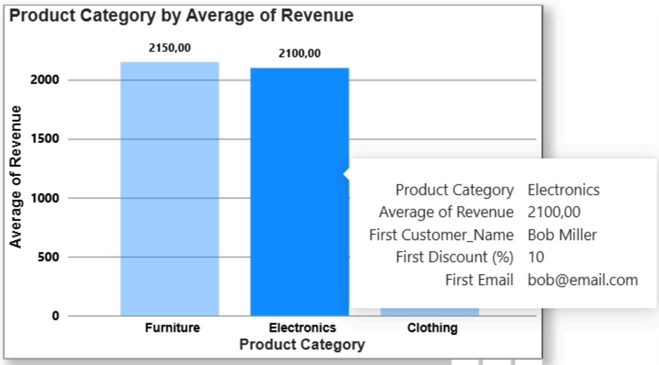
Product_Category	Orders
Clothing	3
Electronics	2
Furniture	2

Power BI

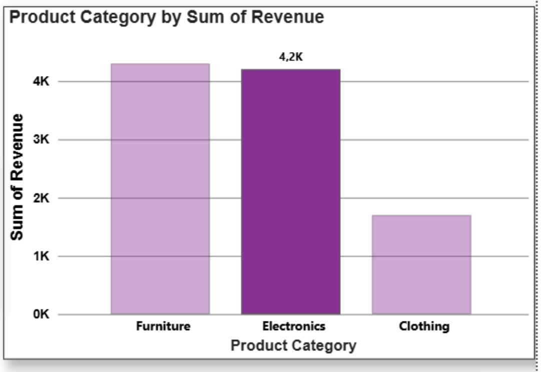


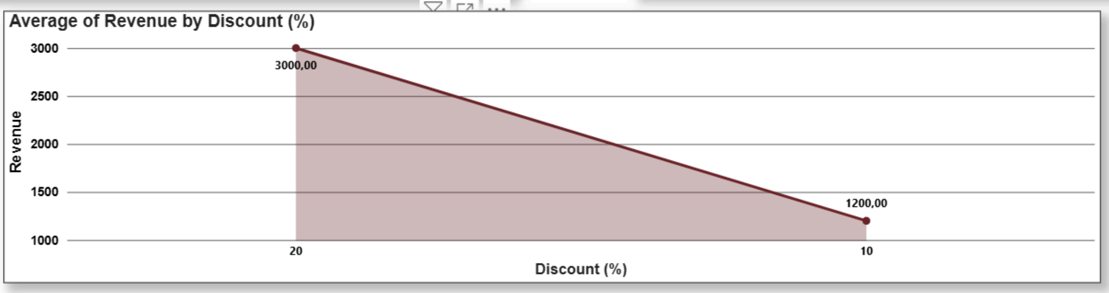


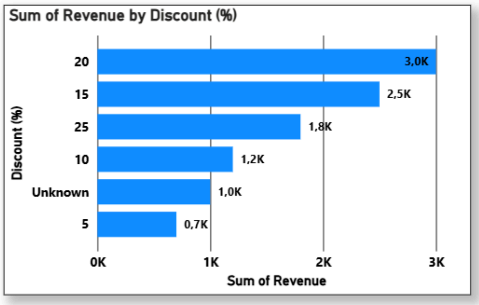


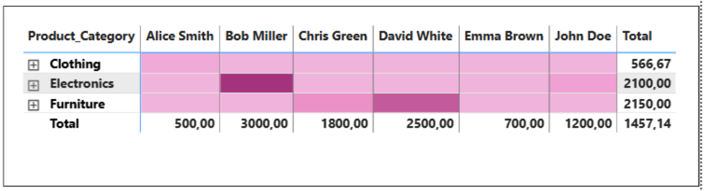


Product Category	Electronics
Average of Revenue	2100,00
First Customer_Name	Bob Miller
First Discount (%)	10
First Email	bob@email.com

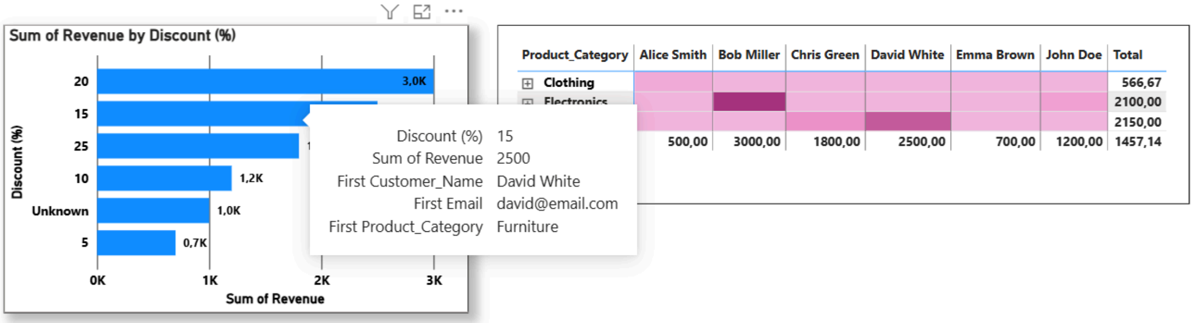








Product_Category	Alice Smith	Bob Miller	Chris Green	David White	Emma Brown	John Doe	Total
⊞ Clothing							566,67
⊞ Electronics							2100,00
⊞ Furniture							2150,00
Total	500,00	3000,00	1800,00	2500,00	700,00	1200,00	1457,14



Discount (%)	15
Sum of Revenue	2500
First Customer_Name	David White
First Email	david@email.com
First Product_Category	Furniture

Product_Category	Alice Smith	Bob Miller	Chris Green	David White	Emma Brown	John Doe	Total
⊞ Clothing							566,67
⊞ Electronics							2100,00
							2150,00
	500,00	3000,00	1800,00	2500,00	700,00	1200,00	1457,14