Analysis Done by: Sindio-Apaun Mac-John

Date: June 17th 2025

Dataset: Clean_Sales_Data_week_2.

1. Data Cleaning

The original dataset, like most raw data, was far from perfect. It had various inconsistencies that could affect decision-making if left unchecked. Here's what I did:

- **Cloned the raw dataset** to create a clean working copy (clean_sales_data_week_2) using SQL's CREATE TABLE ... LIKE syntax to avoid altering the original.

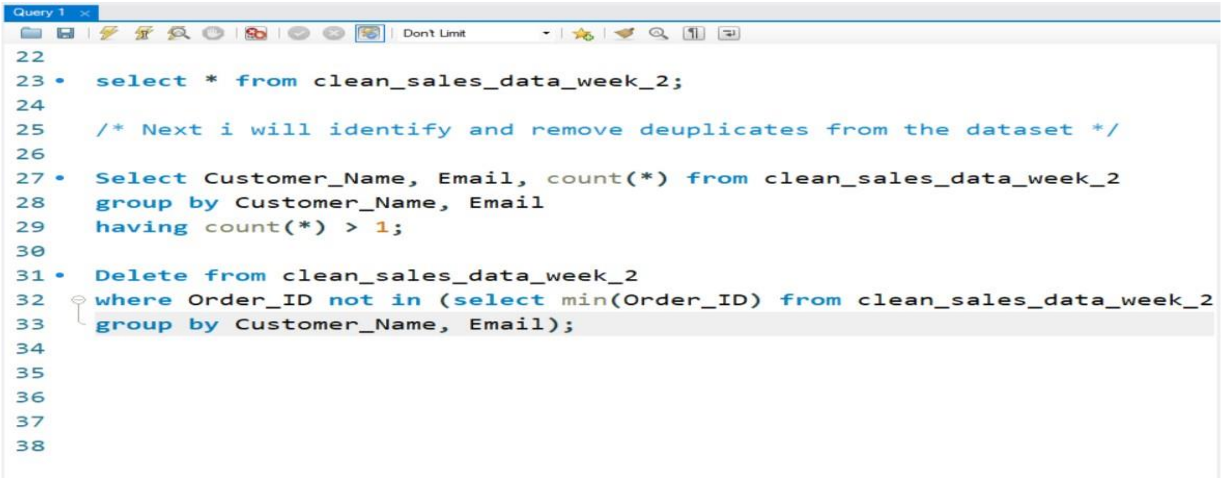
```sql
/* Its Good Practice not to edit the main dataset,
Hence i am going to create another table exactly like the
raw dataset and work on that one */


select * from raw_sales_data_week_2;

create table clean_sales_data_week_2 like raw_sales_data_week_2;

insert into clean_sales_data_week_2
select * from raw_sales_data_week_2;
```

- **I Replaced missing values with placeholders**:
  - Emails → not_provided@email.com
  - Discounts → 0% ○ Phone numbers → "Unknown"

```sql
/*  The first process i'll take in attempt to wrangle my data is
to identify and replace missing data with placeholders, SQL already
replaced as missinng Phone Numbers with "0" so ill replace all
missing emails with "not_provided@gmail.com" and missing discount
will be left as 0*/

select * from clean_sales_data_week_2 where Email = "" or Phone = 0
or `Discount (%)` = 0;
```

Order_ID	Customer_Name	Email	Phone	Product_Category	Order_Date	Revenue	Discount (%)
102	Alice Smith	not_provided@gmail.com	9898989898	Clothing	2024-01-05	500	0
103	Bob Miller	bob@email.com	0	Electronics	2024-12-01	3000	20
108	Alice Smith	alice@email.com	0	Clothing	2024-03-08	500	0
111	Grace Okafor	grace@email.com	0	Electronics	2024-01-12	3000	20
118	Olivia Stone	olivia@email.com	0	Furniture	2024-09-01	3200	12

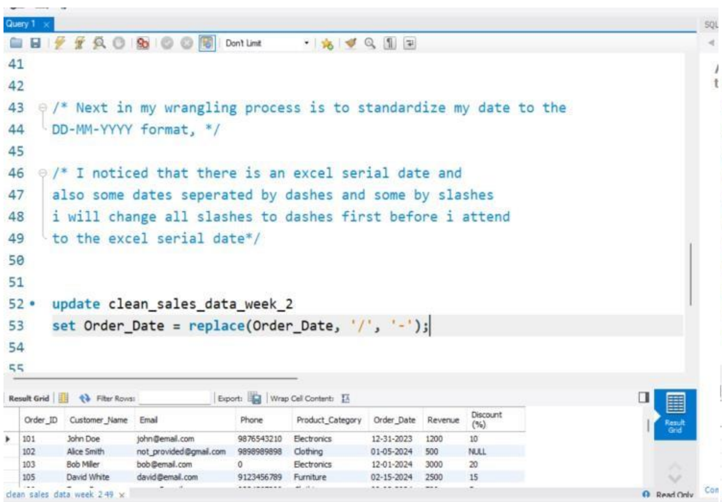
- **Removed duplicates**, including repeated customers like "John Doe" that appeared twice.



```sql
select * from clean_sales_data_week_2;

/* Next i will identify and remove deuplicates from the dataset */

Select Customer_Name, Email, count(*) from clean_sales_data_week_2
group by Customer_Name, Email
having count(*) > 1;

Delete from clean_sales_data_week_2
where Order_ID not in (select min(Order_ID) from clean_sales_data_week_2
group by Customer_Name, Email);
```

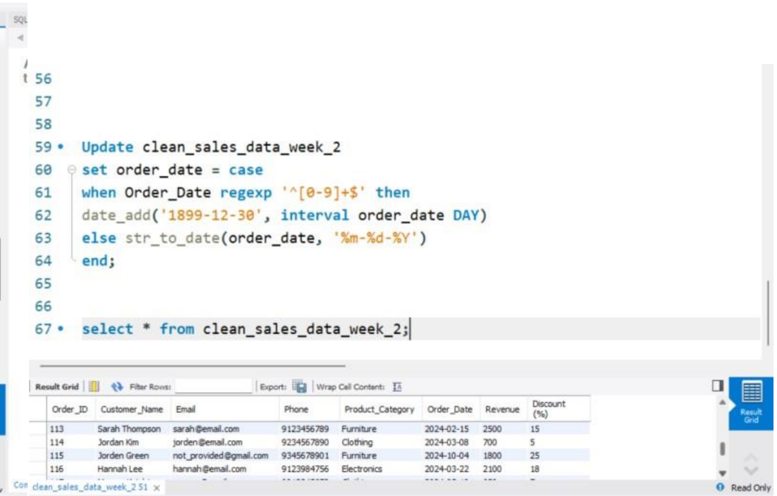
- **Standardized date formats** to YYYY-MM-DD using SQL.

This proved stressful because I noticed some of the dates were in the YYYY/MM/DD format while some others were in the YYYY-MM-DD format and some others were in an Excel Serial Date Format, so I changed all the “/” to “-“, before converting the serial dates to normal dates and then standardizing in the format I wanted.



```sql
/* Next in my wrangling process is to standardize my date to the
DD-MM-YYYY format, */

/* I noticed that there is an excel serial date and
also some dates seperated by dashes and some by slashes
i will change all slashes to dashes first before i attend
to the excel serial date*/


update clean_sales_data_week_2
set Order_Date = replace(Order_Date, '/', '-');
```

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	01-05-2024	500	NULL
103	Bob Miller	bob@email.com	0	Electronics	12-01-2024	3000	20
105	David White	david@email.com	9123456789	Furniture	02-15-2024	2500	15



```sql
Update clean_sales_data_week_2
set order_date = case
when Order_Date regexp '^[0-9]+$' then
date_add('1899-12-30', interval order_date DAY)
else str_to_date(order_date, '%m-%d-%Y')
end;


select * from clean_sales_data_week_2;
```

Order_ID	Customer_Name	Email	Phone	Product_Category	Order_Date	Revenue	Discount (%)
113	Sarah Thompson	sarah@email.com	9123456789	Furniture	2024-02-15	2500	15
114	Jordan Kim	jorden@email.com	9234567890	Clothing	2024-03-08	700	5
115	Jorden Green	not_provided@gmail.com	9345678901	Furniture	2024-10-04	1800	25
116	Hannah Lee	hannah@email.com	9123984756	Electronics	2024-03-22	2100	18

For better understanding Excel Serial Date is a way Excel stores dates format behind the scene. Starting from Janurary 1^{st} 1990, excel gives that day the number 1, January 2^{nd} 1990 as 2, January 1^{st} 2000 as day 36,526 and June 21^{st} 2025 as day 45,161.

2. Analytical Queries and Trends Identification

After all cleaning activities were done on SQL the dataset was ready to help me detect trends and behaviors. I queried the cleaned data to get insights using SQL and this are some noticeable trends.

Total Revenue by Product Category

By summing revenue by category**,** Furniture clearly leads with $17,500 in total revenue, significantly outperforming Electronics $12,300) and Clothing ($ 3,850). This suggests that Furniture is currently the most profitable product category, contributing nearly half of the total revenue. Electronics, while also a strong performer, lags behind Furniture by approximately $ 5,200, indicating potential for growth. Clothing shows the weakest performance, contributing the least to overall sales. This highlights an opportunity to either improve marketing efforts for this category or reevaluate its strategic importance**.**

```sql
/* After my wrangling is done i want to start my analysis in order
to find some trends */
/* I'll calculate the total revenue per product category
to determine the most profitable segments after that i'll Find the average
discount applied across different customer segments
to analyse discount effectiveness, lastly i will Analyse
monthly sales trends to identify peak sales periods.*/

select product_category, sum(Revenue) Total_Revenue
from clean_sales_data_week_2
Group by product_category
Order by Total_Revenue desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

product_category	Total_Revenue
Furniture	17500
Electronics	12300
Clothing	3850

Average Discount by Product Categories

Looking at the average discounts across product categories, there's a clear trend that jumps out. Furniture got the highest average discount at 18.14%, Electronics followed with 14.67%, and Clothing received the smallest discount, just 3.67%. Now here's what makes it interesting: the revenue matches this same pattern. Furniture brought in the highest sales with $17,500, Electronics came next with $12,300, and Clothing trailed behind at $3,850.

```sql
select product_category, avg(`Discount (%)`) Avg_Discount
from clean_sales_data_week_2
Group by product_category;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

product_category	Avg_Discount
Clothing	3.6667
Electronics	14.6667
Furniture	18.1429

This pattern suggests that discounts are doing more than just lowering prices, they're actually encouraging people to buy. In categories like Furniture and Electronics, where the items are more expensive, a good discount might be the extra push customers need to make a purchase.

And clearly, it's working. Clothing, on the other hand, had the smallest discounts and also the weakest sales. That might mean people aren't excited to buy when they don't feel like they're getting a good deal, especially in a category where options are everywhere.

So what does this tell us? Smart discounting is actually a mental game and if played well can improve sales and give better results. It’s not just about cutting prices. It’s about using discounts as a tool to drive more sales and increase revenue in the areas where it matters most.

Monthly Sales Trends

Analyzing monthly sales brought up some observations and logical assumptions. February 2024 generated a whopping sum of $8,000, this could be linked to many things but the two major logical assumptions that came to mind is Valentines Day Promotion Sales and Clearance Sales from January Stocks.

Months like January 2024, March 2024, September 2024 all ranged between $3,200-$3,500 and comparing December 2023 and December 2024 we saw not much difference in their revenue with $2,400 and $3,000 Respectively, we can associate this steady value with the festive period and holiday season.

May 2024 recorded the least sale with a shocking amount of $650, this alarming return is calling for attention and I will relate this low performance to Spending Fatigue after all the spending in the first quarter of the year, also low season demand since there isn’t any huge season in the month of May.

```sql
Select date_format(Order_Date, '%Y-%m') "Month",
sum(Revenue) Total_Revenue
from clean_sales_data_week_2
group by date_format(Order_Date, '%Y-%m')
order by Total_Revenue desc;
```

Month	Total_Revenue
2024-02	8000
2024-01	3500
2024-03	3300
2024-09	3200
2024-12	3000
2023-11	2700
2023-12	2400
2024-04	1800
2024-10	1800
2024-06	1800
2024-08	1500
2024-05	650

3. Visual Explorations in Power BI for better Understanding

To make the insights even clearer, Power BI visualizations were created.

Heat Map – Product Category vs Month

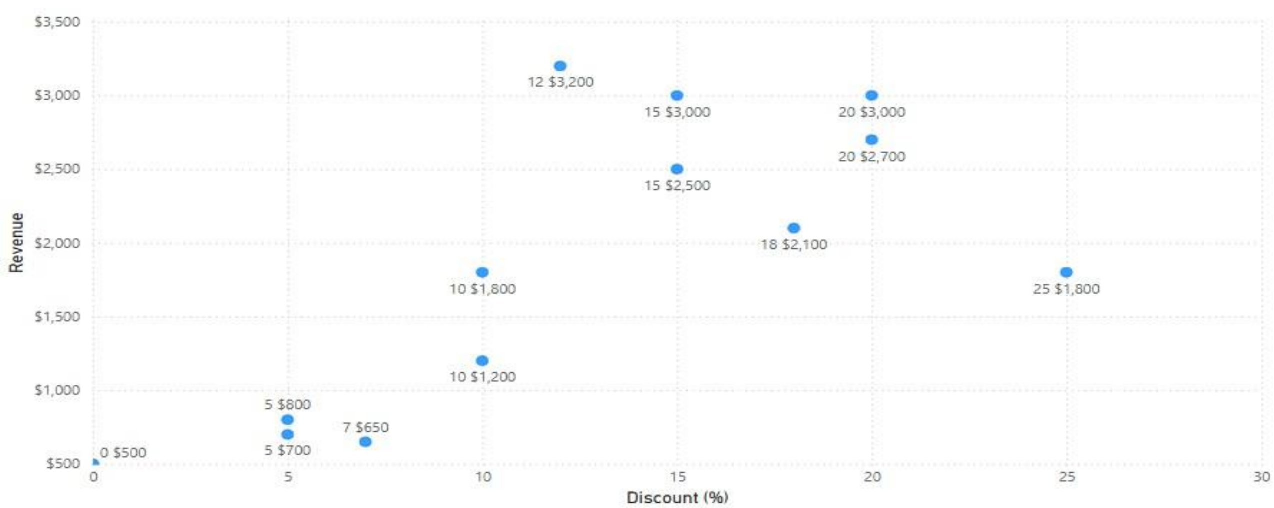
With the knowledge of May 2025 providing the lowest return we could clearly see now its as a result of low sales in Electronics and Furniture, although surprisingly the highest recorded return which came in February 2024, also came in Furniture Category solely.

The Only Month to boast of two different Categories would be January 2024 and March 2024

Product_Category	2023-11	2023-12	2024-01	2024-02	2024-03	2024-04	2024-05	2024-06	2024-08	2024-09	2024-10	2024-12
Clothing			$500		$1,200		$650		$1,500			
Electronics		$2,400	$3,000		$2,100			$1,800				$3,000
Furniture	$2,700			$8,000		$1,800				$3,200	$1,800	

Scatter Plot – Discount (%) vs Order Revenue

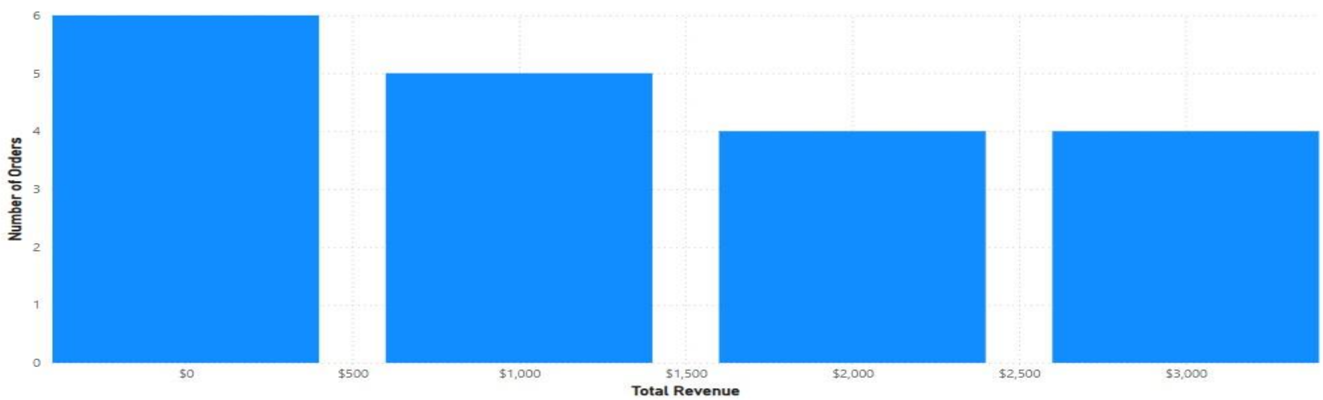
The scatter plot shows how revenue changes with varying discount percentages. At low discounts (0–5%), revenue remains modest, ranging from $500 to $800—suggesting these offers aren't compelling enough to boost sales. However, at moderate discounts (10–15%), revenue increases significantly, peaking at $3,200 with a 12% discount. This indicates a strong positive response from customers within this range.



At higher discount levels (20–25%), revenue becomes unstable. While a 20% discount still yields up to $3,000, pushing it to 25% causes a drop to $1,800, showing diminishing returns. Overall, the optimal discount range appears to be 10–15%, with 12% as the sweet spot for maximizing revenue without over-discounting

Histogram – Distribution of Order Sizes

This chart shows how many orders fall into different revenue ranges. Most orders (6) made less than $500, meaning there were more small sales than big ones. As the revenue goes up, the number of orders goes down.



There were fewer high-revenue orders, with just 4 in both the $2,000 and $3,000 ranges. This tells us that while big sales happen, they're not as common. Most of the business is coming from many smaller orders.

4. Key Insights and Strategic Recommendations

Based on the above exploratory analysis here are some of my recommendations

1. **Prioritize Furniture** in advertising, bundling, and restocking strategies since it's the highest selling product.
2. **Refine discounts** to sit within the high-impact 10–15% range and this should be done strategically.
3. **Double down on seasonal campaigns** during low months like May.
4. **Introduce premium tiers or bundle offers** to increase order sizes and lifetime customer value.
5. **Clothing Category** should be re-assessed. We should consider better promotions or start to question its viability.