Basic SQL Project 1

E-commerce data Analysis.

Data source - Kaggle

Data Summary

The data set presents e-commerce transaction records, detailing a high volume of sales over a period from early 2020 to late 2022.

Each entry records specific information, including a unique order ID, the state where the transaction took place, the customer's name, the date of the order, and the status of the order (e.g., Order, Processing, Shipped, Delivered). Crucially, the data lists the item purchased, its category (like Motherboard, CPU, Graphic Card, RAM), the brand, the unit cost and price, the quantity ordered, and the total monetary value of the transaction before and after tax. Finally, each record indicates an associate's name linked to the sale.

What do I want to do with this data?

I would like to analyze the data to answer the following questions.

- Identify which product categories generate the highest revenue. (Performance by Category)
- Determine profit margin per product category. (Performance by category)
- Determine Sales Performance by region (Geographical Analysis)
- Evaluate the efficiency or sales volume handled by each supervisor (Supervisor Performance)
- Analyze sales trends over time (Time Series analysis)

For this project, I will be using HEX IDE for my SQL queries.

*DATA CLEANING.

Showcasing:

- -Handling Nulls
- -Use of DISTINCT to check for duplicates.
- -Checking Data types for consistency
- Use of REPLACE and CAST functions

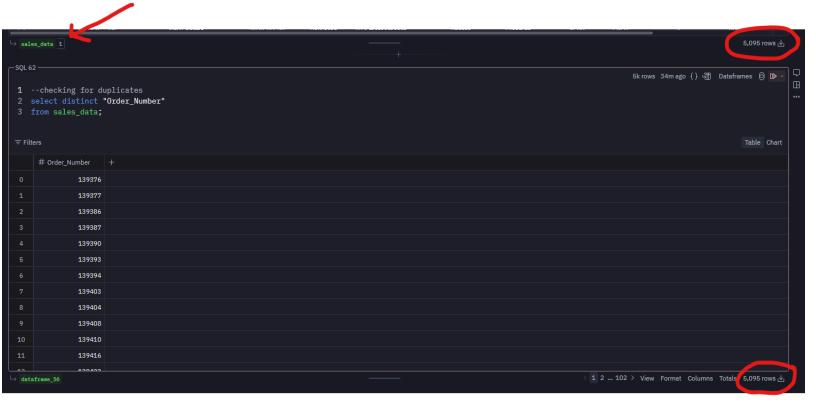
1. Checking for nulls

Just looking at the data, there are clearing entire columns with null entries across the board. Taking out the null rows using SQL query

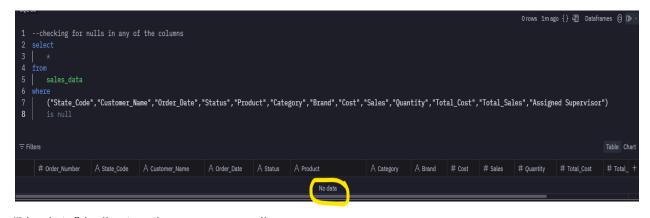
```
1 --cleaning out all null rows
2 select
3 | *|
4 from
5 | ecommerce.csv
6 where
7 | Order_Number is not null
```

The result gives us a table with 5095 rows, down from 5,110 rows. So there were 15 entirely null rows in the dataset.

2. Confirming there are in fact 5095 **distinct** order numbers. Checking for duplicates. Order_number will serve as the primary key or the unique identifier in this dataset. (renamed prior result as sales_data for convenience)

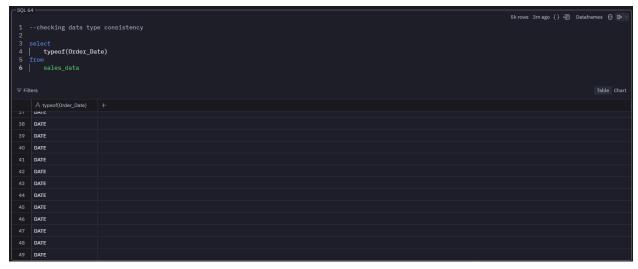


3. Checking for nulls in any of the columns



"No data" indicates there are no nulls

4. Checking data types for consistency.

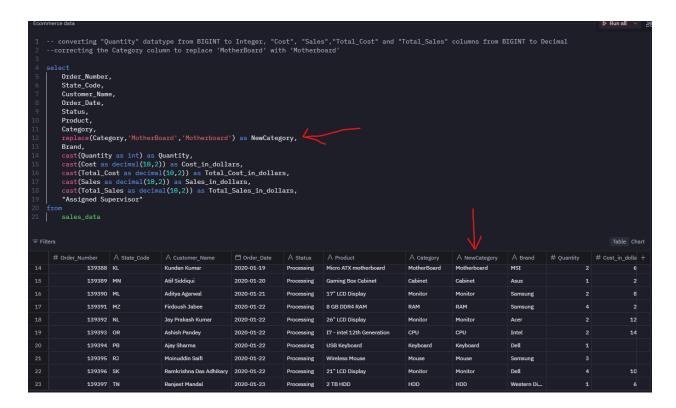


Results:

Column Datatype Order_Number - BIGINT State_code - VARCHAR Costumer_name_ - VARCHAR Order_Date - DATE Status - VARCHAR Product -VARCHAR Category - VARCHAR **Brand** - VARCHAR Cost - BIGINT Quantity -BIGINT Total_Cost - BIGINT Sales -BIGINT Total Sales - BIGINT Assigned Supervisor - VARCHAR

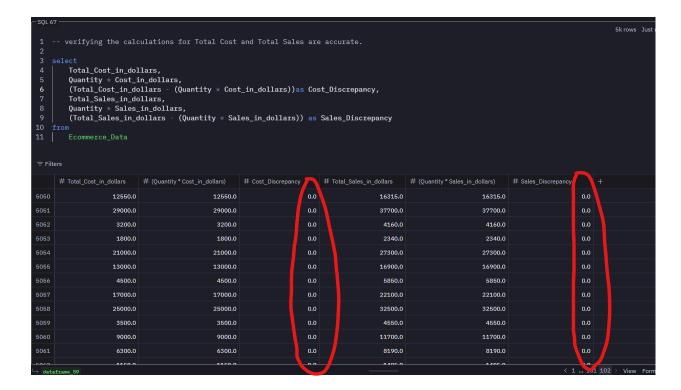
Converted datatype for columns; Quantity (to integer), Cost, Sales Total_Cost and Total_Sales columns to decimal.

5. Replacing erroneous Category name 'MotherBoard' with 'Motherboard'



Renamed this new result **Ecommerce_Data**.

6. Verifying the calculations in the dataset for Total Cost and Total Sales are accurate.



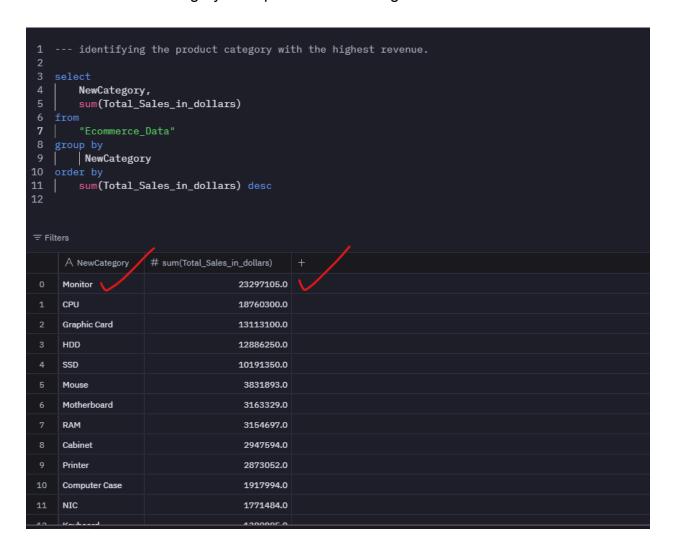
DATA ANALYSIS

Showcasing:

- -Aggregate Functions
- -Group By
- -Order By
- -Date_trunc

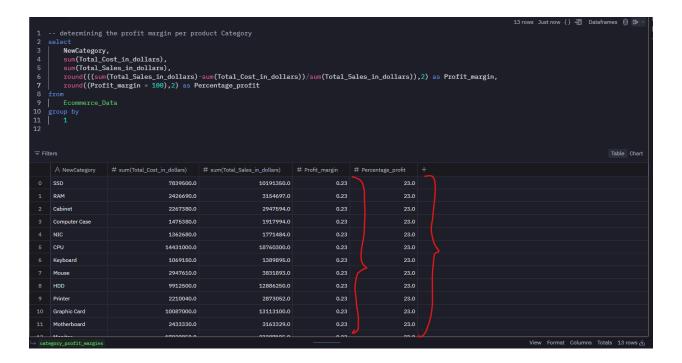
From the cleaned data (Ecommerce Data), we want to answer a few questions.

Q1- What Product Category is responsible for the highest revenue?



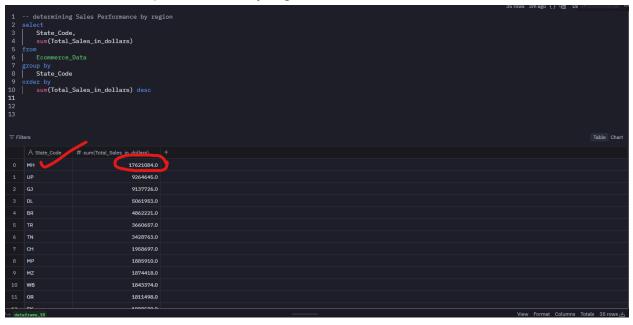
Monitor is the product category that has sold the most over the 3 years.

Q2. Determine the profit margin per Product Category



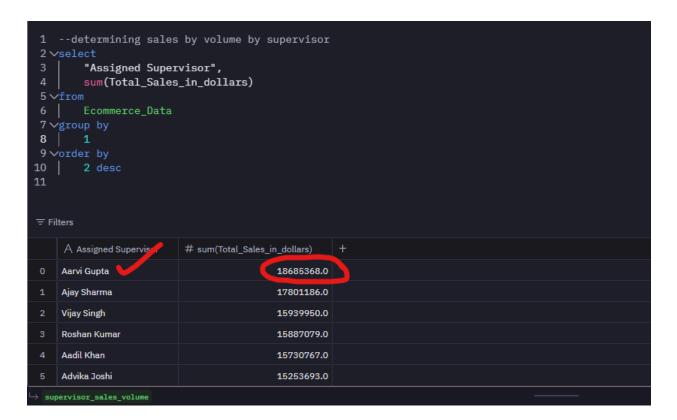
The data reveals an interesting detail. A surprising 23% profit for <u>every single product category.</u> This raises lots of questions.

Q3. Determine sales performance by region.



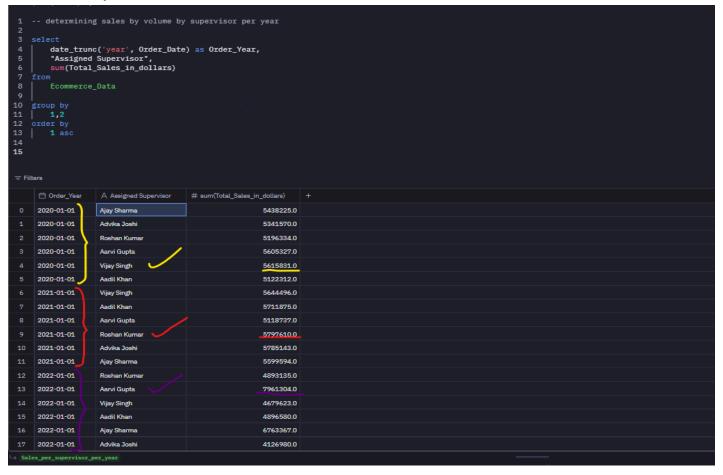
The state of Maharashtra with code **MH** emerged to have the most sales in the dataset.

Q4. Determine Sales Volume by Supervisor (Supervisor Performance)



Aarvi Gupta is the supervisor who has recorded the most sales over the years from 2020 to 2022.

We can also investigate to see if Aarvi Gupta has consistently been the top supervisor in terms of sales in each of the 3 years, from 2020 to 2022.



The data reveals Aarvi Gupta only became top performer in 2022. In 2020, the top sales supervisor was Vijay Singh, and in 2021, it was Roshan Kumar.

Aarvi Gupta's excellent performance in 2022, pushed him ahead of everyone in terms of overall sales.

Q5. Analyzing sales trends over time.

1 analyzing sales trends over time. 2 select 3 date_trunc('year', Order_Date) as Order_Year, 4 Category, 5 sum(Total_Sales_in_dollars) 6 from 7 Ecommerce_Data 8 group by 9 1,2 10 order by 11 1 asc 12				
च Filters				
21	☐ Order_Year	A Category	# sum(Total_Sales_in_dollars)	+
28	2022-01-01	Graphic Card	4518150.0	
29	2022-01-01	Motherboard	1112085.0	
30	2022-01-01	Mouse	1260623.0	
31	2022-01-01	HDD	4464200.0	
32	2022-01-01	CPU	6024200.0	
33	2022-01-01	Keyboard	460330.0	
34	2022-01-01	Printer	909324.0	
35	2022-01-01	SSD	3528200.0	
36	2022-01-01	RAM	1091155.0	
37	2022-01-01	Cabinet	964925.0	
38	2022-01-01	Computer Case	654511.0	
39	2022-01-01	NIC	632346.0	

This gives an overview of sales trends per category for each of the 3 years. We can gain more insight into these trends after visualizing this analysis.

For Data Visualization in Tableau, please refer to the link below.

https://public.tableau.com/views/EcommerceDataViz_17611575808030/KeyEcommerceKPIs?:langua ge=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

KEY INSIGHTS AND RECOMMENDATIONS

1. Data Entry flaws.

The dataset presented with some flaws in the data entry. The Category for Motherboard appeared twice, one spelled <u>Motherboard</u> and the other spelled <u>MotherBoard</u>. These were recorded as separate product categories.

It is imperative that the data entry clerk adopts consistency in data entry as such mistakes can lead to faulty analysis and conclusions.

2. Sales Performance by Product Category.

This analysis revealed that monitors were the most consumed and keyboards the least consumed product. It can be inferred that the boost in sales of Monitors was as a result of the uptick in remote work during the COVID 19 outbreak. Also, most laptops have keypads, hence keyboards are not as desirable.

It is however interesting to see large numbers for CPUs which were associated with older computer models. But with the rise of Artificial Intelligence(AI) and Machine Learning (ML), massive computational power is required as Data Centers experience expansion. Also, the escalating need for businesses to process and analyze vast amounts of data in real-time across industries like Finance and Healthcare drives the investment in powerful CPUs.

Hence, the company should invest in Monitors, CPUs, Graphic Cards, HDD and SSN as those are the top-selling products.

3. Profit Margin per Product Category.

This was the most interesting insight. A consistent 23% profit margin across all product categories is an indication of a consistent Pricing and markup structure. While this may reflect a consistent business model, there could be lots of missed opportunities.

Some products are definitely perceived to have more value than others, for those products, the business should be charging more. With these results, it is certain that those products are under-priced. They are leaving money on the table.

It is worth noting that the data was collected for the years 2020,2021 and 2022. Given the global COVID Pandemic, it is strange that the numbers do not indicate any drop in activity or sales during the covid years. We all experienced market volatility, slower processing times, supply chain issues, lower demand for goods and services, disruptions in distribution chains etc during the pandemic. So it is interesting that this business seemed to have been immune to all of those environmental and economic factors. The data may just have been made up without all of this in mind.

4. Sales Volume by Supervisor.

This metric clearly indicated that Aarvi Gupta did a phenomenal job in driving sales. If a raise or promotion were being discussed, this should be the right recipient. Also the company may consider offering some sort of gift to Vijay Singh and Roshan Kumar for their exceptional sales numbers in 2020 and 2021 respectively.

Maybe Vijah Singh should be recognized for pulling off those sales numbers despite the pandemic in 2020.

5. Performance by Region.

The State of Maharashtra recorded the highest sales over the 3 years. This State is one of India's most densely populated regions. It is also home to two major metropolitan cities, Mumbai and Pune. Mumbai is India's financial and commercial capital and most rapidly evolving FinTech and Data hub. Pune is also widely recognized as a prominent IT and a growing Technology hub in India. It therefore makes sense that the most sales came from this region.

The State of Andaman (AN) on the other hand has agriculture, tourism and fisheries as major economic activities. This tells why the sales of computers and computer accessories are significantly low in the region.

6. Sales Trends over time.

Visualizing the sales trends in Tableau, we see a consistent pattern over the 3 years as portrayed by the shape of the line graphs. They are identical, meaning every product category performed in the same way over the three years. This is very unusual in the real world and leads us to conclude that the data isn't from a real world scenario but rather just made up.

It is possible to have certain products consistently doing well or poorly in a 3 year span in ecommerce, but having them perform exactly the same, to the same degrees in comparison with other categories, is quite questionable.