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. <u>Checking for nulls</u>

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

```sql
--cleaning out all null rows
select
    *
from
    ecommerce.csv
where
    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. <u>Confirming there are in fact 5095 **distinct** order numbers.Checking for duplicates.</u>
   Order_number will serve as the primary key or the unique identifier in this dataset.
   (renamed prior result as sales_data for convenience)

sales_data 1 — 5,095 rows

SQL 62 — 5k rows 34m ago — Dataframes

```sql
--checking for duplicates
select distinct "Order_Number"
from sales_data;
```

Filters — Table Chart

	# Order_Number
0	139376
1	139377
2	139386
3	139387
4	139390
5	139393
6	139394
7	139403
8	139404
9	139408
10	139410
11	139416

dataframe_56 — 1 2 ... 102 — View Format Columns Totals — 5,095 rows

3. <u>Checking for nulls in any of the columns</u>

0 rows 1m ago — Dataframes

```sql
--checking for nulls in any of the columns
select
    *
from
    sales_data
where
    ("State_Code","Customer_Name","Order_Date","Status","Product","Category","Brand","Cost","Sales","Quantity","Total_Cost","Total_Sales","Assigned Supervisor")
    is null
```

Filters — Table Chart

# Order_Number	A State_Code	A Customer_Name	A Order_Date	A Status	A Product	A Category	A Brand	# Cost	# Sales	# Quantity	# Total_Cost	# Total_

No data

“No data” indicates there are no nulls

4. <u>Checking data types for consistency</u>.

SQL 64

5k rows 3m ago { } Dataframes

```sql
--checking data type consistency

select
    typeof(Order_Date)
from
    sales_data
```

Filters Table Chart

	typeof(Order_Date)
38	DATE
39	DATE
40	DATE
41	DATE
42	DATE
43	DATE
44	DATE
45	DATE
46	DATE
47	DATE
48	DATE
49	DATE

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. <u>Replacing erroneous Category name ‘MotherBoard’ with ‘Motherboard’</u>

Ecommerce data

Run all

```
-- converting "Quantity" datatype from BIGINT to Integer, "Cost", "Sales","Total_Cost" and "Total_Sales" columns from BIGINT to Decimal
--correcting the Category column to replace 'MotherBoard' with 'Motherboard'

select
    Order_Number,
    State_Code,
    Customer_Name,
    Order_Date,
    Status,
    Product,
    Category,
    replace(Category,'MotherBoard','Motherboard') as NewCategory,
    Brand,
    cast(Quantity as int) as Quantity,
    cast(Cost as decimal(10,2)) as Cost_in_dollars,
    cast(Total_Cost as decimal(10,2)) as Total_Cost_in_dollars,
    cast(Sales as decimal(10,2)) as Sales_in_dollars,
    cast(Total_Sales as decimal(10,2)) as Total_Sales_in_dollars,
    "Assigned Supervisor"
from
    sales_data
```

Filters

Table Chart

	# Order_Number	State_Code	Customer_Name	Order_Date	Status	Product	Category	NewCategory	Brand	# Quantity	# Cost_in_dolla
14	139388	KL	Kundan Kumar	2020-01-19	Processing	Micro ATX motherboard	MotherBoard	Motherboard	MSI	2	6
15	139389	MN	Atif Siddiqui	2020-01-20	Processing	Gaming Box Cabinet	Cabinet	Cabinet	Asus	1	2
16	139390	ML	Aditya Agarwal	2020-01-21	Processing	17" LCD Display	Monitor	Monitor	Samsung	2	8
17	139391	MZ	Firdoush Jabee	2020-01-22	Processing	8 GB DDR4 RAM	RAM	RAM	Samsung	4	2
18	139392	NL	Jay Prakash Kumar	2020-01-22	Processing	26" LCD Display	Monitor	Monitor	Acer	2	12
19	139393	OR	Ashish Pandey	2020-01-22	Processing	I7 - intel 12th Generation	CPU	CPU	Intel	2	14
20	139394	PB	Ajay Sharma	2020-01-22	Processing	USB Keyboard	Keyboard	Keyboard	Dell	1	
21	139395	RJ	Moinuddin Saifi	2020-01-22	Processing	Wireless Mouse	Mouse	Mouse	Samsung	3	
22	139396	SK	Ramkrishna Das Adhikary	2020-01-22	Processing	21" LCD Display	Monitor	Monitor	Dell	4	10
23	139397	TN	Ranjeet Mandal	2020-01-23	Processing	2 TB HDD	HDD	HDD	Western Di...	1	6

Renamed this new result **Ecommerce_Data.**

6. <u>Verifying the calculations in the dataset for Total Cost and Total Sales are accurate.</u>

SQL 67

5k rows Just

```sql
-- verifying the calculations for Total Cost and Total Sales are accurate.

select
    Total_Cost_in_dollars,
    Quantity * Cost_in_dollars,
    (Total_Cost_in_dollars - (Quantity * Cost_in_dollars))as Cost_Discrepancy,
    Total_Sales_in_dollars,
    Quantity * Sales_in_dollars,
    (Total_Sales_in_dollars - (Quantity * Sales_in_dollars)) as Sales_Discrepancy
from
    Ecommerce_Data
```

Filters

	# Total_Cost_in_dollars	# (Quantity * Cost_in_dollars)	# Cost_Discrepancy	# Total_Sales_in_dollars	# (Quantity * Sales_in_dollars)	# Sales_Discrepancy
5050	12550.0	12550.0	0.0	16315.0	16315.0	0.0
5051	29000.0	29000.0	0.0	37700.0	37700.0	0.0
5052	3200.0	3200.0	0.0	4160.0	4160.0	0.0
5053	1800.0	1800.0	0.0	2340.0	2340.0	0.0
5054	21000.0	21000.0	0.0	27300.0	27300.0	0.0
5055	13000.0	13000.0	0.0	16900.0	16900.0	0.0
5056	4500.0	4500.0	0.0	5850.0	5850.0	0.0
5057	17000.0	17000.0	0.0	22100.0	22100.0	0.0
5058	25000.0	25000.0	0.0	32500.0	32500.0	0.0
5059	3500.0	3500.0	0.0	4550.0	4550.0	0.0
5060	9000.0	9000.0	0.0	11700.0	11700.0	0.0
5061	6300.0	6300.0	0.0	8190.0	8190.0	0.0

dataframe_59 1 ... 101 102 View Form

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?

```sql
--- identifying the product category with the highest revenue.

select
    NewCategory,
    sum(Total_Sales_in_dollars)
from
    "Ecommerce_Data"
group by
    NewCategory
order by
    sum(Total_Sales_in_dollars) desc

```

Filters

	NewCategory	sum(Total_Sales_in_dollars)	+
0	Monitor	23297105.0	
1	CPU	18760300.0	
2	Graphic Card	13113100.0	
3	HDD	12886250.0	
4	SSD	10191350.0	
5	Mouse	3831893.0	
6	Motherboard	3163329.0	
7	RAM	3154697.0	
8	Cabinet	2947594.0	
9	Printer	2873052.0	
10	Computer Case	1917994.0	
11	NIC	1771484.0	

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

Q2. Determine the profit margin per Product Category

```sql
-- determining the profit margin per product Category
select
    NewCategory,
    sum(Total_Cost_in_dollars),
    sum(Total_Sales_in_dollars),
    round(((sum(Total_Sales_in_dollars)-sum(Total_Cost_in_dollars))/sum(Total_Sales_in_dollars)),2) as Profit_margin,
    round((Profit_margin * 100),2) as Percentage_profit
from
    Ecommerce_Data
group by
    1
```

13 rows Just now Dataframes

Filters Table Chart

	NewCategory	sum(Total_Cost_in_dollars)	sum(Total_Sales_in_dollars)	Profit_margin	Percentage_profit
0	SSD	7839500.0	10191350.0	0.23	23.0
1	RAM	2426690.0	3154697.0	0.23	23.0
2	Cabinet	2267380.0	2947594.0	0.23	23.0
3	Computer Case	1475380.0	1917994.0	0.23	23.0
4	NIC	1362680.0	1771484.0	0.23	23.0
5	CPU	14431000.0	18760300.0	0.23	23.0
6	Keyboard	1069150.0	1389895.0	0.23	23.0
7	Mouse	2947610.0	3831893.0	0.23	23.0
8	HDD	9912500.0	12886250.0	0.23	23.0
9	Printer	2210040.0	2873052.0	0.23	23.0
10	Graphic Card	10087000.0	13113100.0	0.23	23.0
11	Motherboard	2433330.0	3163329.0	0.23	23.0

category_profit_margins View Format Columns Totals 13 rows

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.

```sql
-- determining Sales Performance by region
select
    State_Code,
    sum(Total_Sales_in_dollars)
from
    Ecommerce_Data
group by
    State_Code
order by
    sum(Total_Sales_in_dollars) desc
```

35 rows 3m ago

Filters Table Chart

	State_Code	sum(Total_Sales_in_dollars)
0	MH	17621084.0
1	UP	9264645.0
2	GJ	9137726.0
3	DL	5061953.0
4	BR	4862221.0
5	TR	3660657.0
6	TN	3428763.0
7	CH	1958697.0
8	MP	1885910.0
9	MZ	1874418.0
10	WB	1843374.0
11	OR	1811498.0

dataframe_10 View Format Columns Totals 35 rows

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

Q4. Determine Sales Volume by Supervisor (Supervisor Performance)

```sql
--determining sales by volume by supervisor
select
    "Assigned Supervisor",
    sum(Total_Sales_in_dollars)
from
    Ecommerce_Data
group by
    1
order by
    2 desc
```

Filters

	Assigned Supervisor	sum(Total_Sales_in_dollars)
0	Aarvi Gupta	18685368.0
1	Ajay Sharma	17801186.0
2	Vijay Singh	15939950.0
3	Roshan Kumar	15887079.0
4	Aadil Khan	15730767.0
5	Advika Joshi	15253693.0

supervisor_sales_volume

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.

```sql
-- determining sales by volume by supervisor per year

select
    date_trunc('year', Order_Date) as Order_Year,
    "Assigned Supervisor",
    sum(Total_Sales_in_dollars)
from
    Ecommerce_Data

group by
    1,2
order by
    1 asc
```

Filters

	Order_Year	Assigned Supervisor	sum(Total_Sales_in_dollars)
0	2020-01-01	Ajay Sharma	5438225.0
1	2020-01-01	Advika Joshi	5341570.0
2	2020-01-01	Roshan Kumar	5196334.0
3	2020-01-01	Aarvi Gupta	5605327.0
4	2020-01-01	Vijay Singh	5615831.0
5	2020-01-01	Aadil Khan	5122312.0
6	2021-01-01	Vijay Singh	5644496.0
7	2021-01-01	Aadil Khan	5711875.0
8	2021-01-01	Aarvi Gupta	5118737.0
9	2021-01-01	Roshan Kumar	5797610.0
10	2021-01-01	Advika Joshi	5785143.0
11	2021-01-01	Ajay Sharma	5599594.0
12	2022-01-01	Roshan Kumar	4893135.0
13	2022-01-01	Aarvi Gupta	7961304.0
14	2022-01-01	Vijay Singh	4679623.0
15	2022-01-01	Aadil Khan	4896580.0
16	2022-01-01	Ajay Sharma	6763367.0
17	2022-01-01	Advika Joshi	4126980.0

Sales_per_supervisor_per_year

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.

```
--- analyzing sales trends over time.
select
    date_trunc('year', Order_Date) as Order_Year,
    Category,
    sum(Total_Sales_in_dollars)
from
    Ecommerce_Data
group by
    1,2
order by
    1 asc
```

Filters

	Order_Year	Category	# sum(Total_Sales_in_dollars)
27	2022-01-01	Monitor	7700940.0
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

yearly_category_sales_trends

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?:language=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 Motherboard and the other spelled MotherBoard. 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.