CASE STUDY
ONLINE RETAIL DATABASE

Objective:

The objective of this project is to write optimized SQL queries to perform data extraction, transformation, and analysis on transactional sales data, and generate meaningful business insights, identify trends in sales and customer behaviour, and document the entire workflow, including data cleaning, calculation of key metrics, and query optimization, to support informed business decisions.

Dataset:

Online Retail Dataset

Data Structure:

- InvoiceNo - Invoice number for each transaction
- StockCode - Code assigned to each product item
- Description - Description of the product
- Quantity - Number of products purchased
- InvoiceDate - Date and time when the invoice was issued
- UnitPrice - Price per unit
- CustomerID - ID for each customer
- Country - Country of the customer

Tools Required:

- SQL Database: SQL Server
- Documentation: Google Docs / Word

Execution:

Importing data:

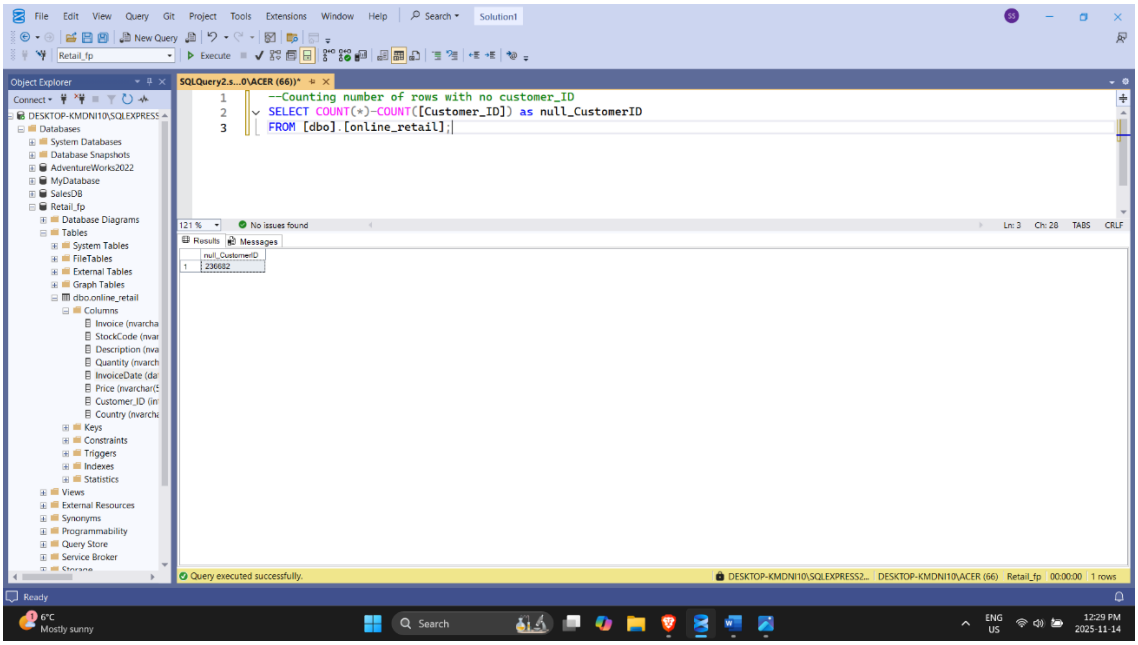
- Open SSMS and connect to your SQL Server
- Right click on the Database folder and select ‘New Database’
- Name the database ‘Retail_fp’ and select Close

Phase 1: Data Acquisition & Cleaning

- Handle missing values

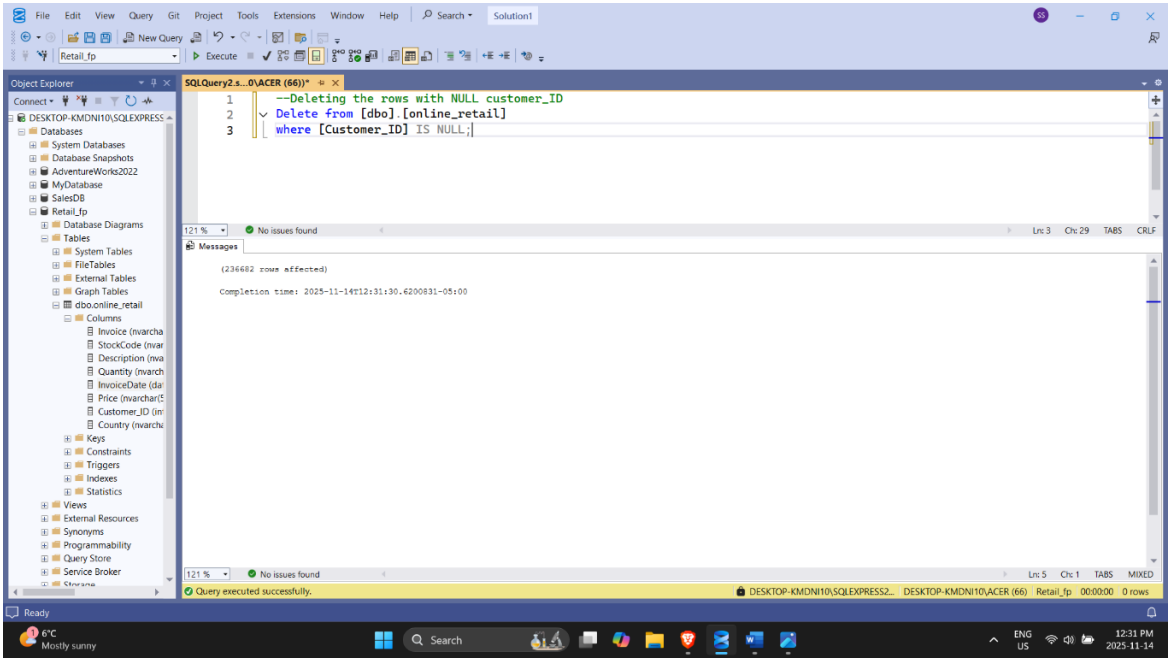
1. Finding the number of rows with no customerID: Removing the customerID will eliminate missing or incomplete information, making the dataset more accurate for providing insights.

```
SELECT COUNT(*)-COUNT([Customer_ID]) as null_ID
FROM [dbo].[online_retail];
```



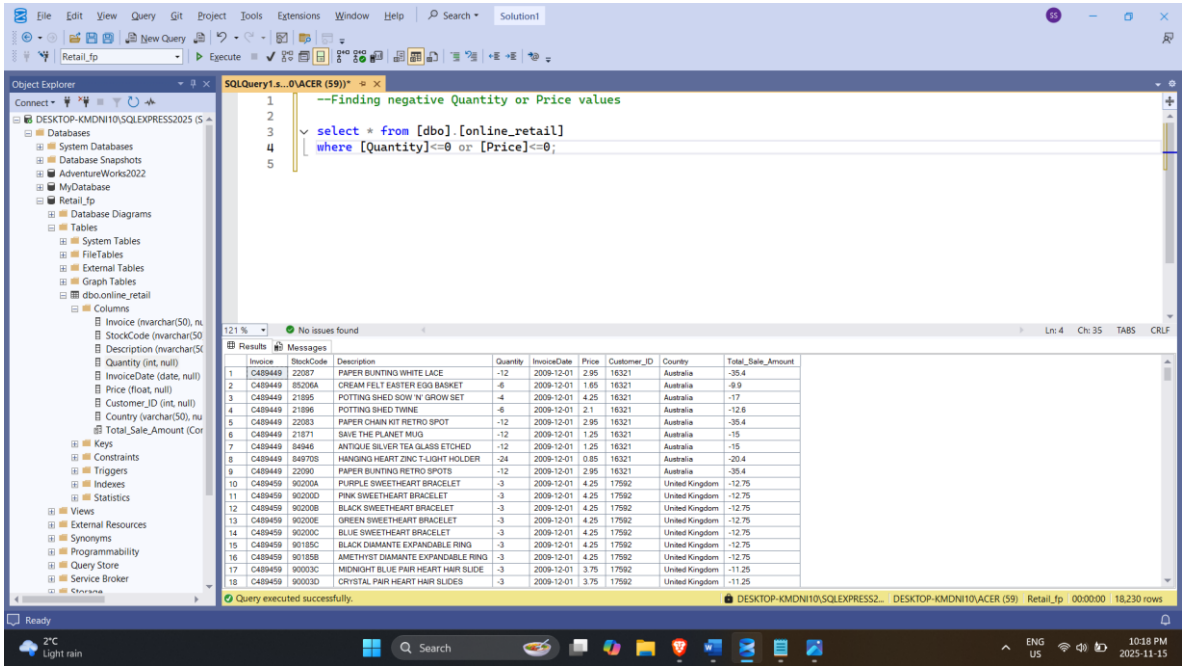
2. Cleaning the table by removing the rows with no customerID

```
Delete from [dbo].[online_retail]
where [Customer_ID] IS NULL;
```



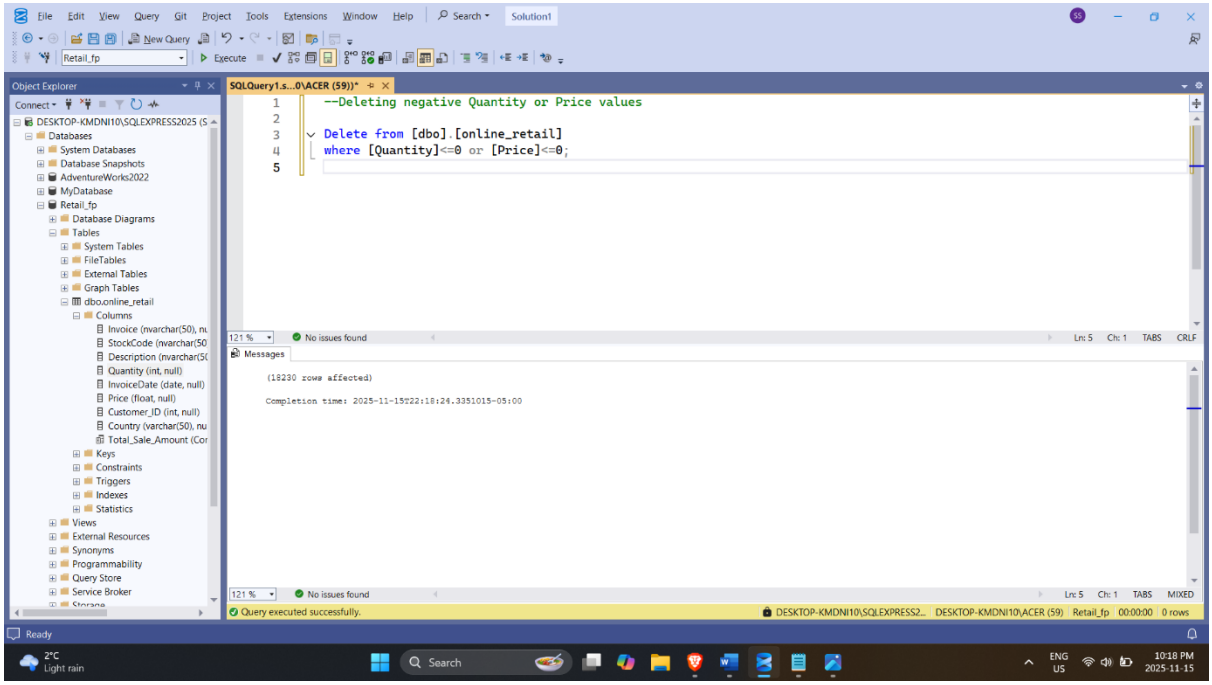
3. Finding rows with negative Price or Quantity value

```
select * from [dbo].[online_retail]
where [Quantity]<=0 or [Price]<=0;
```



4. Deleting the rows with negative values

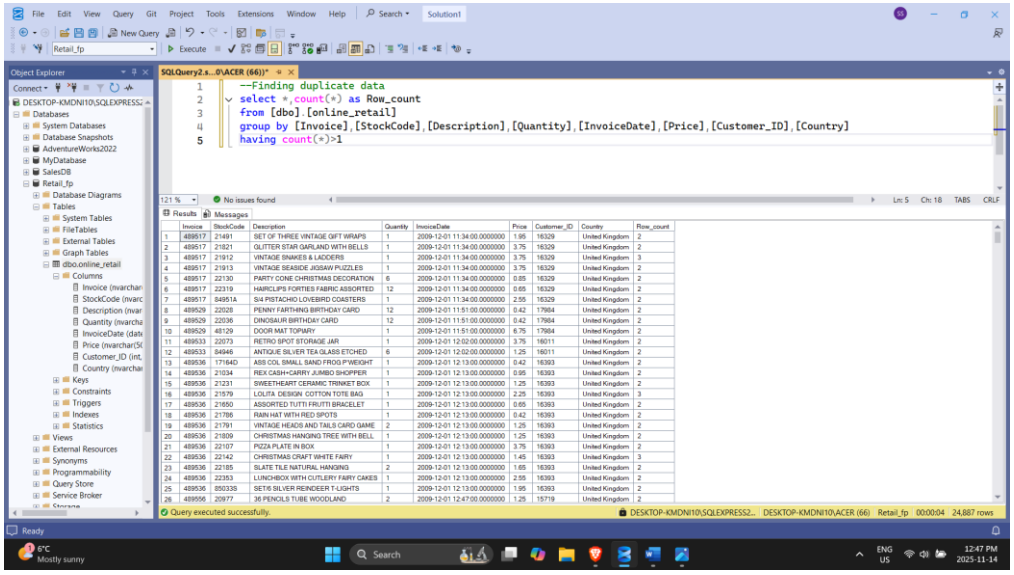
```
Delete from [dbo].[online_retail]
where [Quantity]<=0 or [Price]<=0;
```



- Remove duplicates

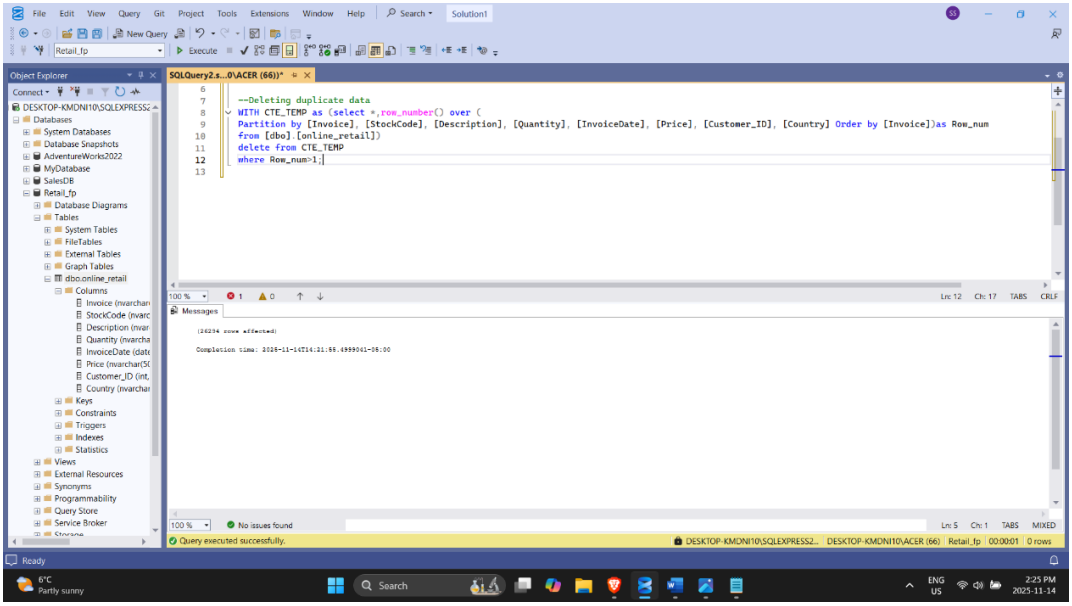
1. Finding duplicate data: To find duplicate data row_number() window function is used which displays rows that occur more than once.

```
select *,count(*) as Row_count
from [dbo].[online_retail]
group by
[Invoice],[StockCode],[Description],[Quantity],[InvoiceDate],[Price],[Custom
er_ID],[Country]
having count(*)>1;
```



2. Deleting duplicate data

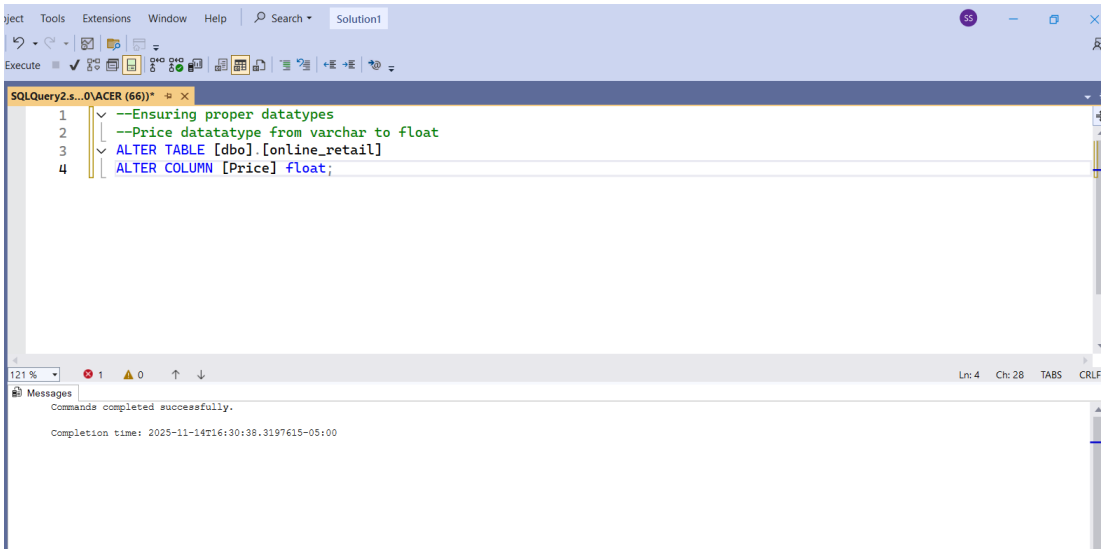
```
WITH CTE_TEMP as (select *,row_number() over (
Partition by [Invoice], [StockCode], [Description], [Quantity],
[InvoiceDate], [Price], [Customer_ID], [Country] Order by [Invoice])as
Row_num
from [dbo].[online_retail])
delete from CTE_TEMP
where Row_num>1;
```



- Ensuring proper datatypes:

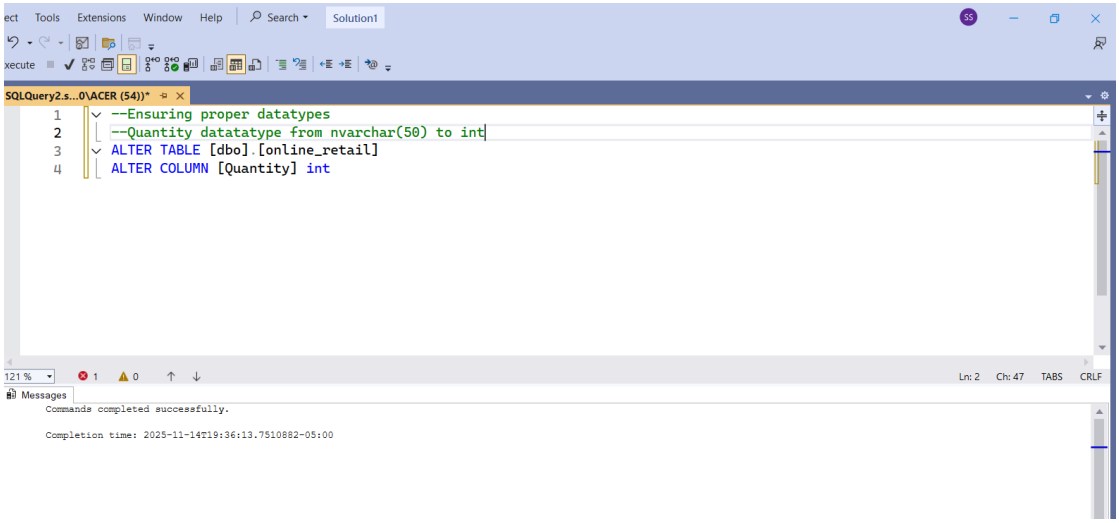
1. Price datatype from varchar(50) to float

```
ALTER TABLE [dbo].[online_retail]
ALTER COLUMN [Price] float;
```



2. Quantity datatatype from nvarchar(50) to int

```
ALTER TABLE [dbo].[online_retail]
ALTER COLUMN [Quantity] int;
```

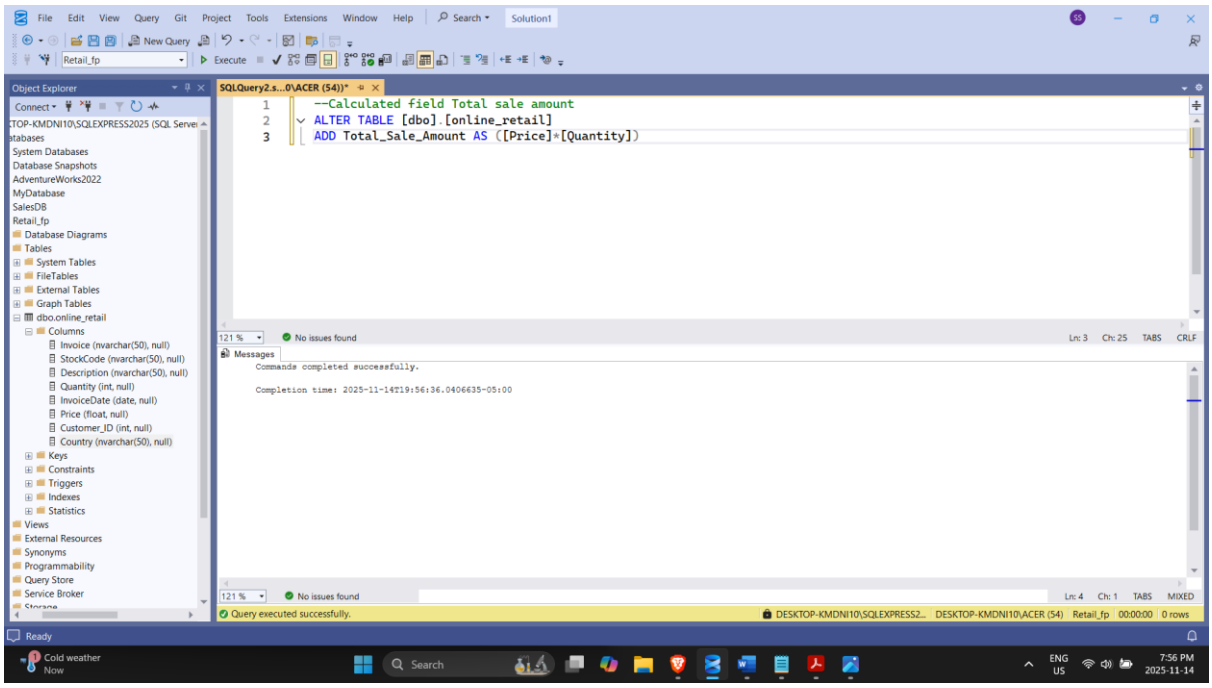


3. InvoiceDate from datetime to date

```
ALTER TABLE [dbo].[online_retail]
ALTER COLUMN [InvoiceDate] DATE;
```

- Create calculated field Total sales Amount

```
ALTER TABLE [dbo].[online_retail]
ADD Total_Sales_Amount AS ([Price]*[Quantity])
```

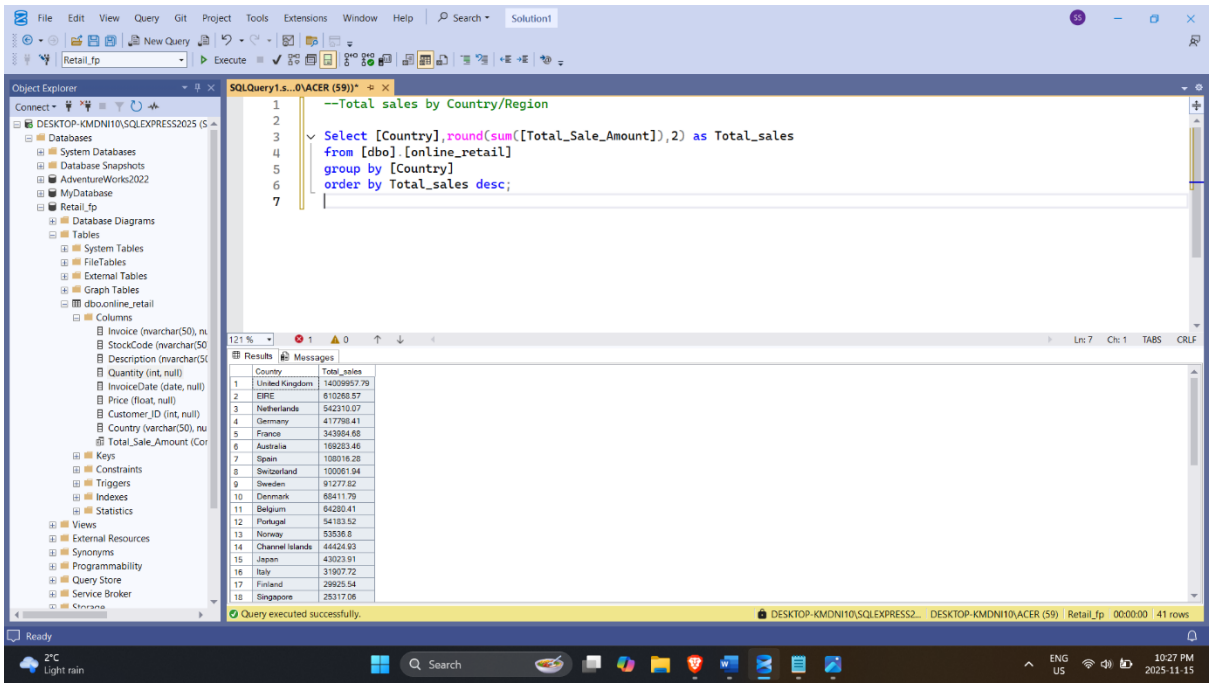


Phase 2: SQL Query Development

- Key metrics for sales analysis

1. Regional Sales

```
Select [Country],round(sum([Total_Sale_Amount]),2) as Total_sales
from [dbo].[online_retail]
group by [Country]
order by Total_sales desc;
```



Analysis:

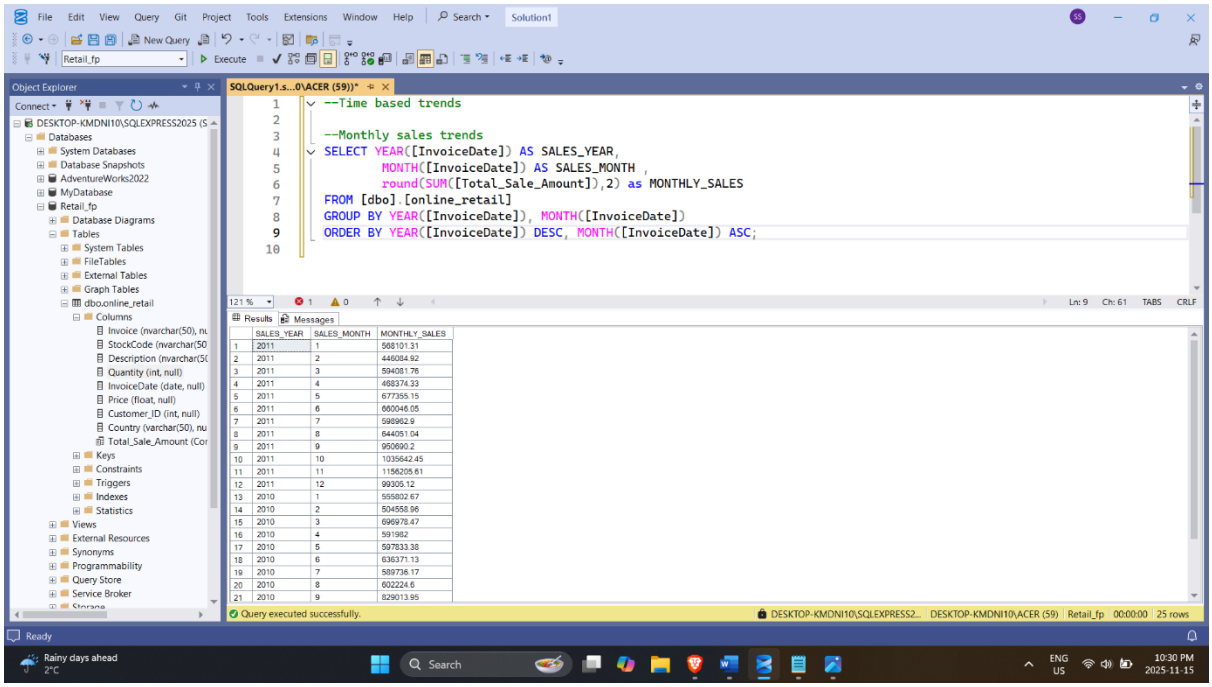
The Country/Region sales data shows a high concentration of revenue in certain countries, with the United Kingdom contributing 14,009,957.79 which is the largest value. Following UK are EIRE, Netherlands, Germany, France that also impact revenue significantly. This suggests that Western Europe is the retailer's primary market.

Moderate sales were observed in countries such as Australia, Switzerland, Spain, and Sweden, while regions like Nigeria, Saudi Arabia, and West Indies had minimal sales, suggesting limited market reach or presence.

For moderate sales countries, targeted ads, localized campaigns, or partnerships with local distributors could help increase market penetration. Low performing countries present long term growth opportunities which can be done by exploring regional demand, strategic pricing.

2. Time based trends: Monthly Sales Trends

```
--Monthly sales trends
SELECT YEAR([InvoiceDate]) AS SALES_YEAR,
          MONTH([InvoiceDate]) AS SALES_MONTH ,
          round(SUM([Total_Sale_Amount]),2) as MONTHLY_SALES
FROM [dbo].[online_retail]
GROUP BY YEAR([InvoiceDate]), MONTH([InvoiceDate])
ORDER BY YEAR([InvoiceDate]) DESC, MONTH([InvoiceDate]) ASC;
```

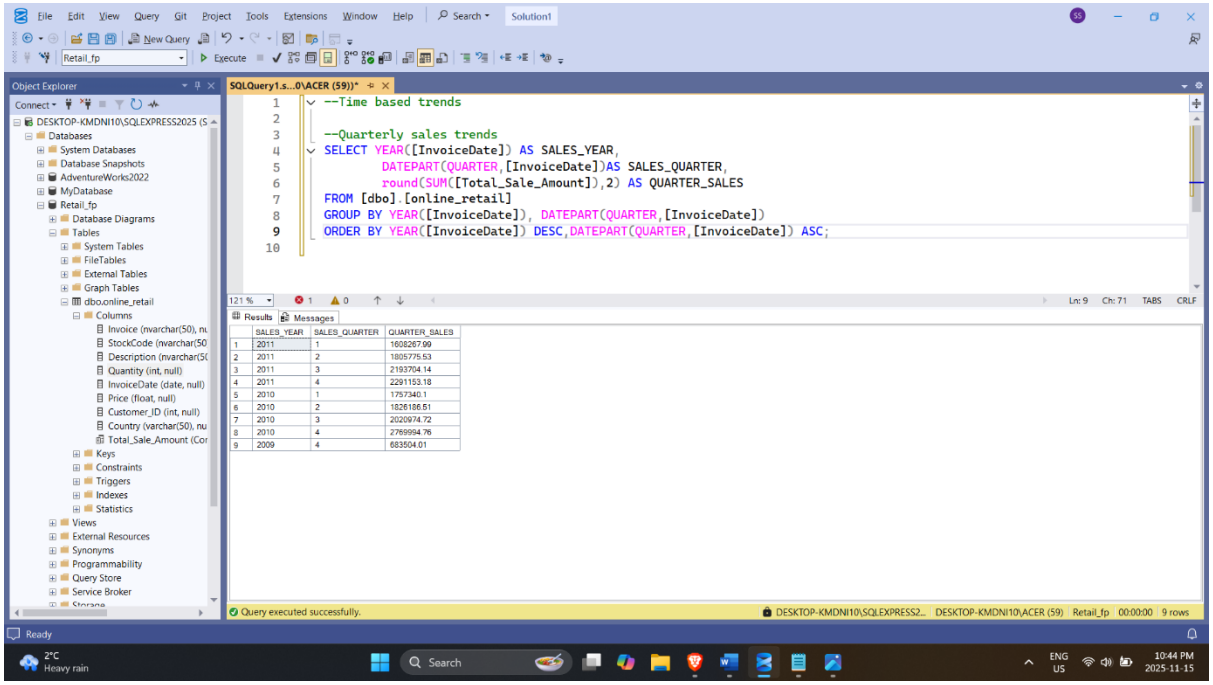


Analysis:

The monthly sales analysis for 2011 reveals clear seasonal trends in customer purchasing behavior. Sales began at 568,101.31 in January, with peaks in May (677,355.15), June (660,046.05), September (950,690.20), October (1,035,642.45), and November (1,156,205.61), and the lowest sales was recorded in December. A similar trend is noticed in the months of the previous year with sales hikes in September, October, November indicating peak demand. To increase sales, the business could implement targeted marketing campaigns during the low-performing months, and introduce seasonal promotions.

3. Time based trends: Quarterly Sales Trends

```
SELECT YEAR([InvoiceDate]) AS SALES_YEAR,
            DATEPART(QUARTER,[InvoiceDate])AS SALES_QUARTER,
            round(SUM([Total_Sale_Amount]),2) AS QUARTER_SALES
FROM [dbo].[online_retail]
GROUP BY YEAR([InvoiceDate]), DATEPART(QUARTER,[InvoiceDate])
ORDER BY YEAR([InvoiceDate]) DESC,DATEPART(QUARTER,[InvoiceDate]) ASC;
```

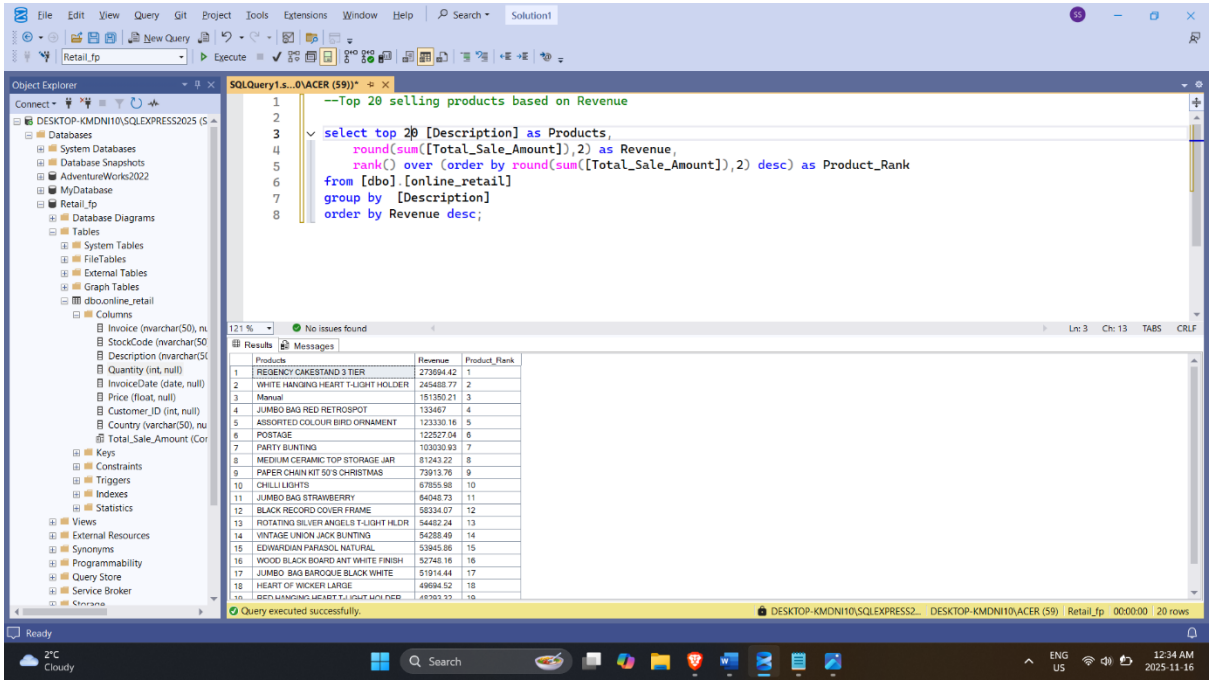


Analysis:

The result shows a clear growth in revenue from the start of the year in 2011, growing steadily and reaching Q4 with 2,291,153.18 in revenue, indicating demand. A similar trend in the previous year 2010, with revenue peaking inQ4 at 2,769,994.76, but is slightly higher than 2011 Q4. These trends suggest that the business can maximize revenue by focusing more on seasonal peaks, aligning marketing campaigns. The lower-performing quarters can be an opportunity for targeted promotions, and product launches to help keep sales up.

4. Product Performance: Identify top-selling products based on revenue

```
select top 20 [Description] as Products,
    round(sum([Total_Sale_Amount]),2) as Revenue,
    rank() over (order by round(sum([Total_Sale_Amount]),2) desc) as
Product_Rank
from [dbo].[online_retail]
group by  [Description]
order by Revenue desc;
```

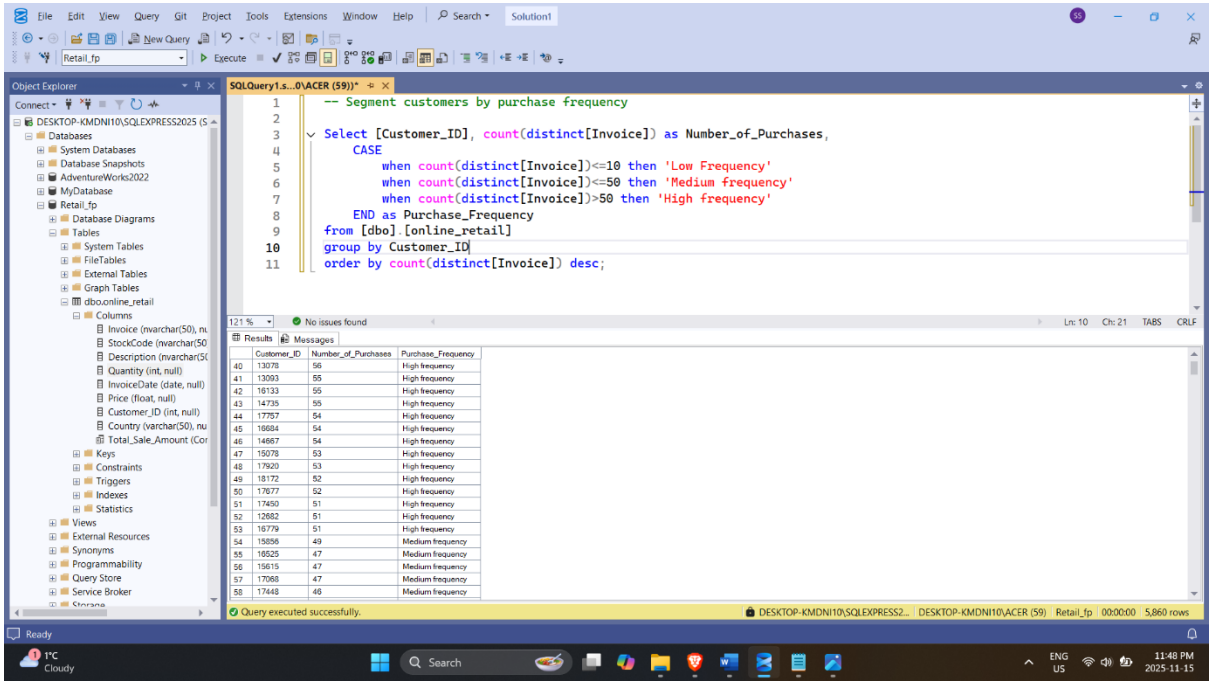


Analysis:

The product analysis shows that the REGENCY CAKESTAND 3 TIER and the WHITE HANGING HEART T-LIGHT HOLDER are the top ranked in terms of revenue, followed by Manuals, JUMBO BAG RED RETROSPOT, and ASSORTED COLOUR BIRD ORNAMENT. These results indicate a high demand for home decor and seasonal home accessories. To increase sales, the business can focus on keeping these popular products in stock, expanding similar product ranges and strengthening marketing around these products.

5. Customer Behavior Analysis: Segment customers by purchase frequency

```
Select [Customer_ID], count(distinct[Invoice]) as Number_of_Purchases,
    CASE
            when count(distinct[Invoice])<=10 then 'Low Frequency'
            when count(distinct[Invoice])<=50 then 'Medium frequency'
            when count(distinct[Invoice])>50 then 'High frequency'
    END as Purchase_Frequency
from [dbo].[online_retail]
group by Customer_ID
order by count(distinct[Invoice]) desc;
```



Analysis:

Customers are segmented by purchase frequency into High, Medium, and Low frequency groups. The analysis shows a large group of high frequency buyers that play a significant role in the company's sales revenue. The medium-frequency segment is also substantial, suggesting an opportunity for growth through targeted marketing or personalized offers. The data also shows a large group of low frequency buyers who make very few purchases. Overall growth can be strengthened by engaging and focusing on high frequency buyers.

Business recommendations:

- Increase Sales: Focus marketing and promotional efforts on high-value regions and top-selling products. Offer discounts or bundle slow-moving items to clear inventory.
- Customer Engagement: Implement targeted campaigns for medium- and low-frequency customers to encourage repeat purchases.
- Growth Opportunities: Expand outreach in emerging markets, leverage seasonal sales trends for timely promotions, and use insights from purchase frequency segmentation to improve personalized marketing.

Conclusion:

This project showcased how SQL can be used to transform raw transactional data into meaningful business insights. Through data cleaning, handling missing values, removing duplicates, and creating calculated fields, the dataset was prepared for accurate analysis. Key findings, including regional sales patterns, time-based trends, top-selling products, and customer purchase behaviors, provided actionable insights to boost sales, optimize inventory, and enhance customer engagement.