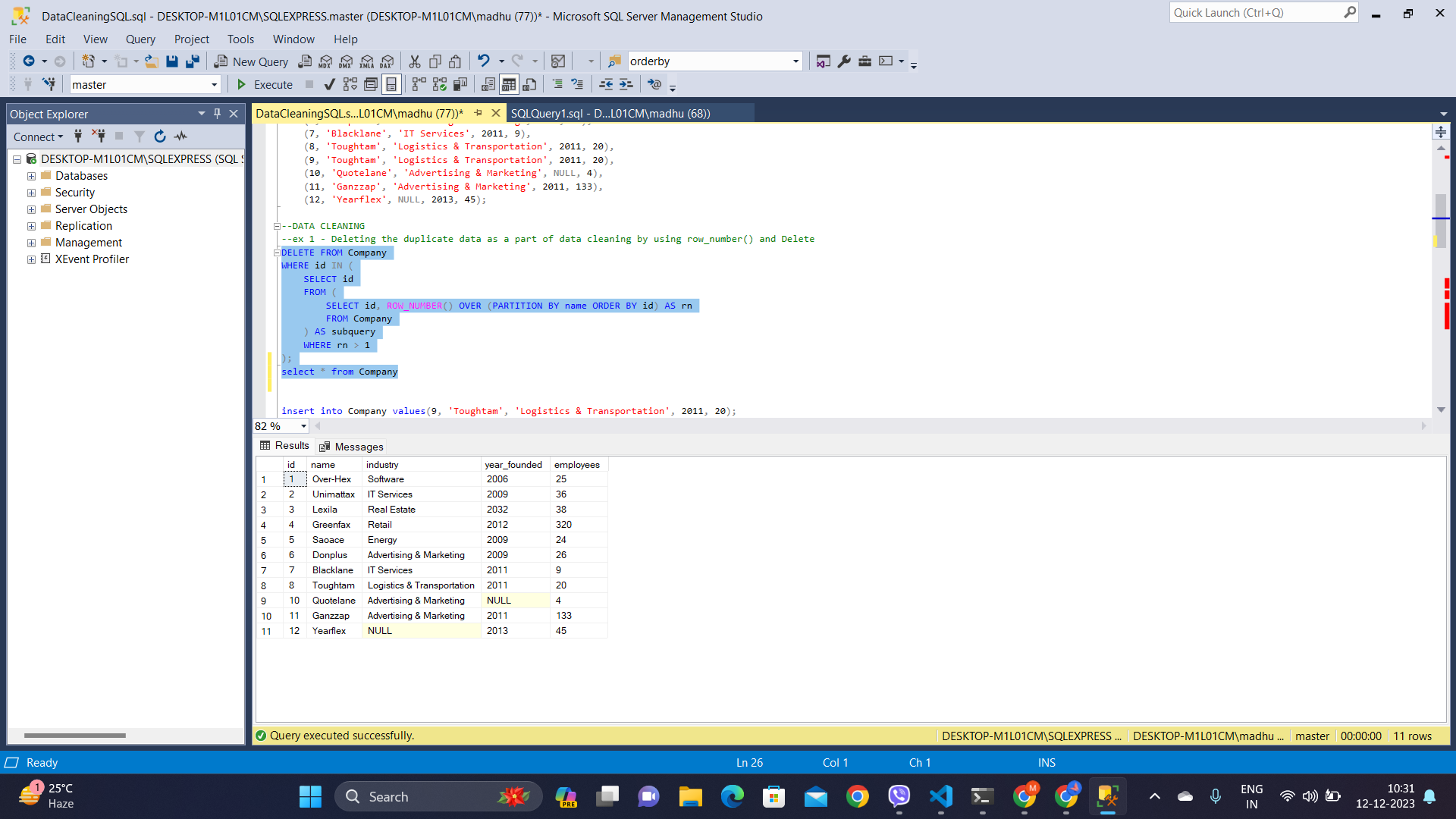
**Day-7  
Madhu Kalyani Gadi (11-12-2023)**

**Data Cleaning**

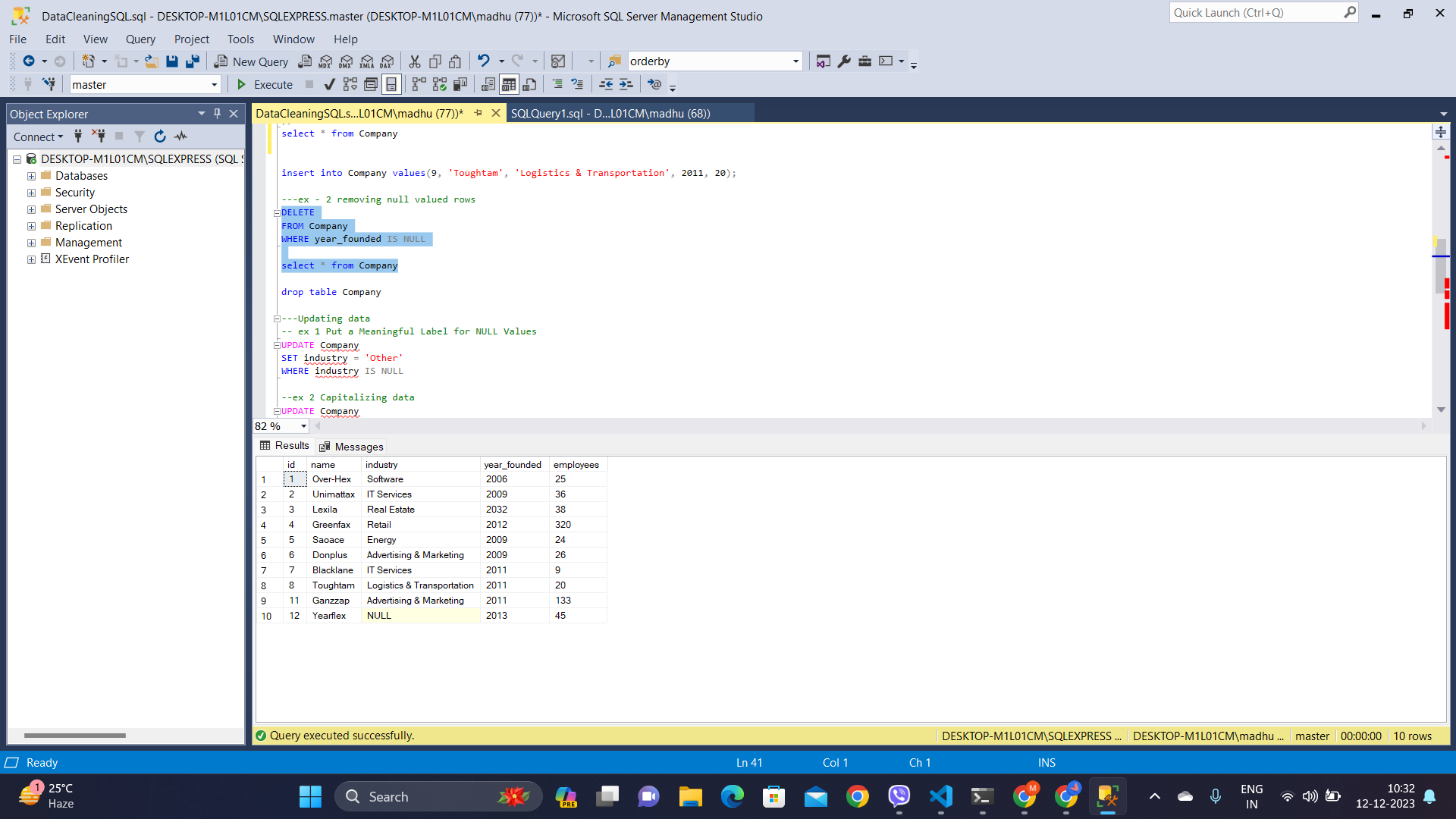
Data cleaning in SQL is the process of identifying and correcting errors in your dataset to ensure accuracy and reliability. Common data cleaning tasks include handling missing values, removing duplicates, correcting inaccuracies, and standardising formats. The goal is to enhance data quality, which, in turn, improves the reliability of any insights or decisions derived from the data.

Ex 1: Deleting the duplicate data as a part of data cleaning by using row\_number() and Delete  
In the table, a row is duplicated which is not required. To remove that we use the below query.



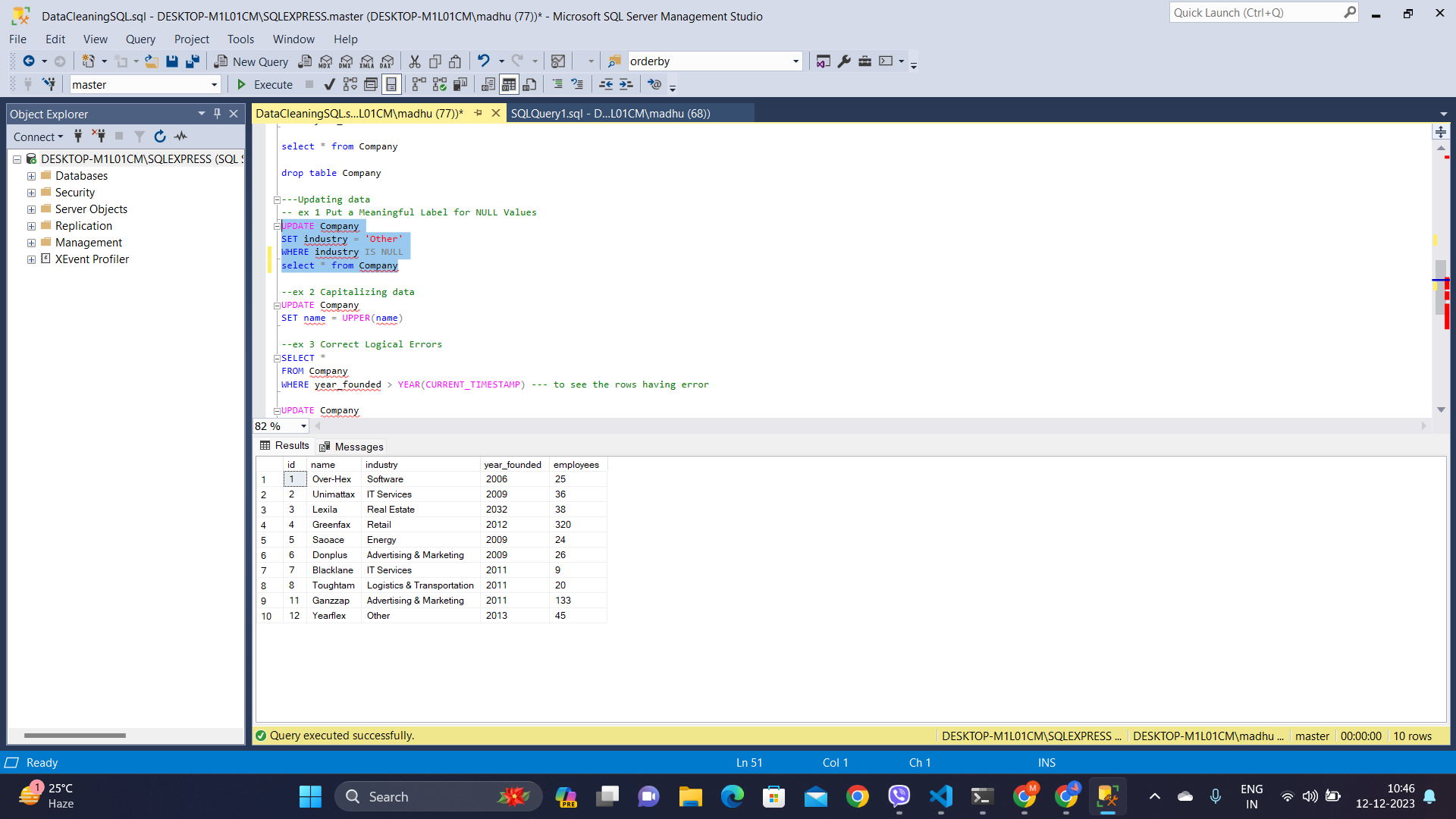
Ex 2: Removing null valued rows from the table is also a part of data cleaning.

Here, we want to delete the row where the year\_founded is null inorder to clean the data.

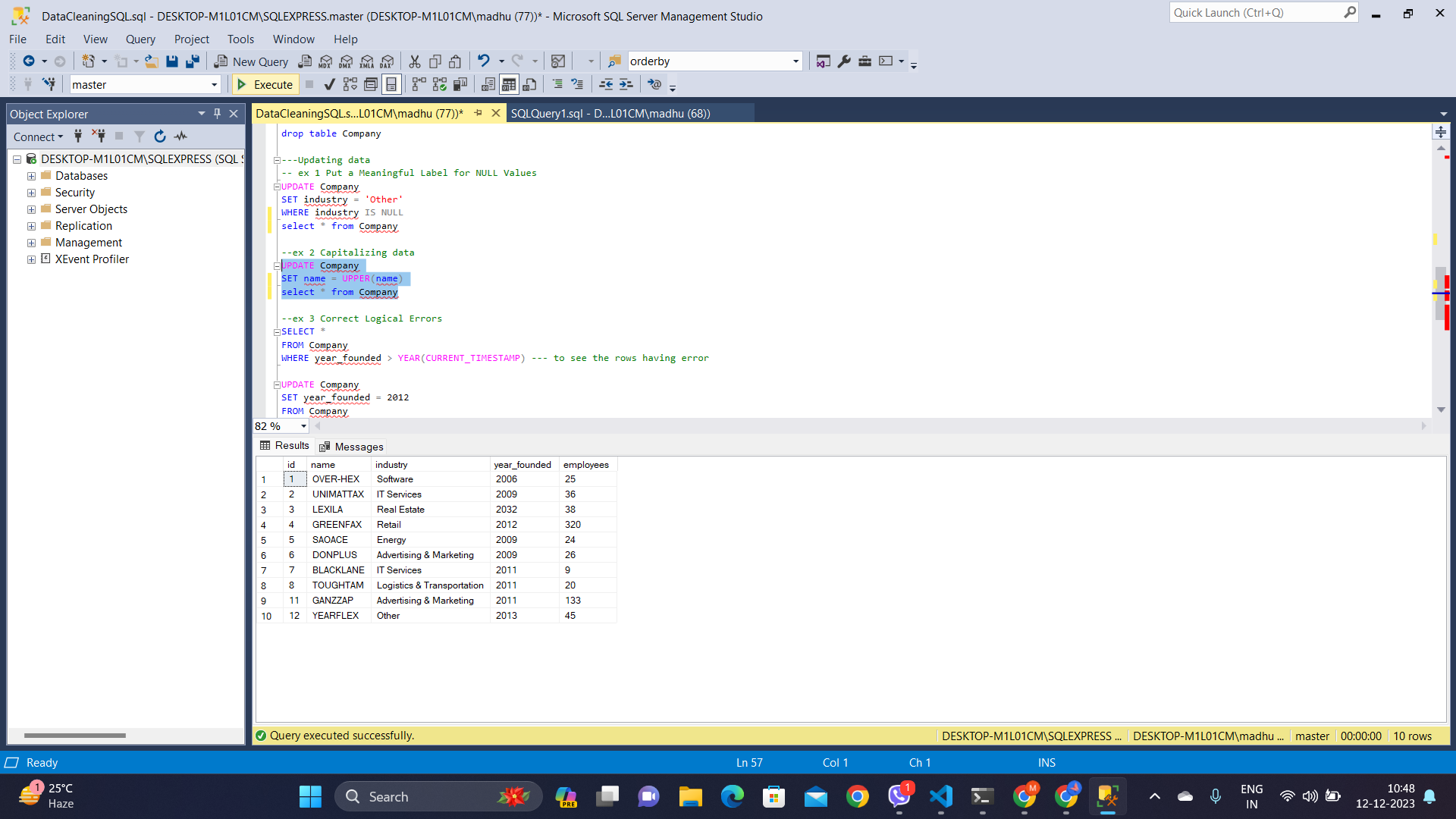


Data Cleaning includes updating data also. If we want to update the column where industry is null. Then we need to execute the below command.

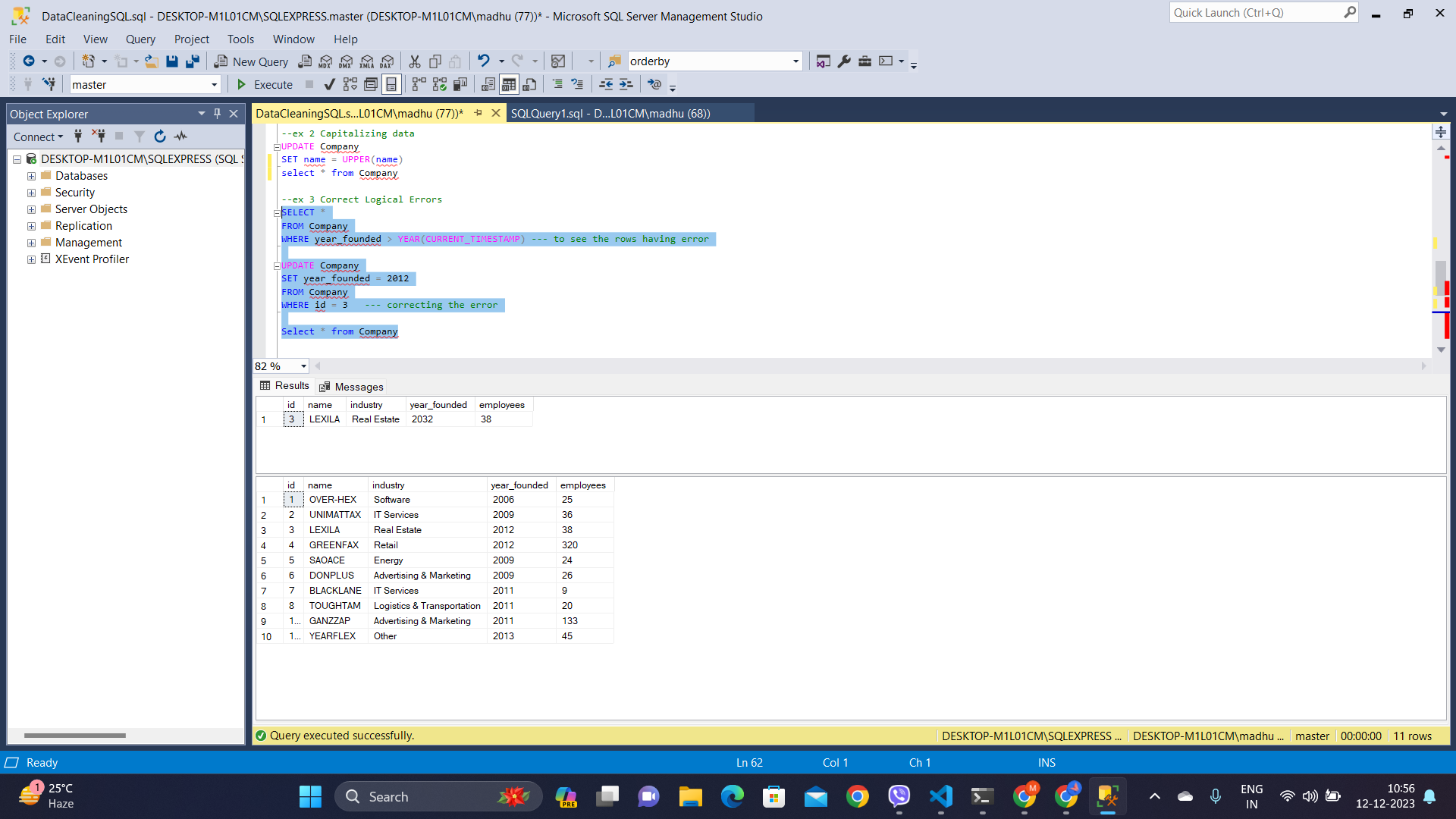
Ex 3: Put a Meaningful Label for NULL Values, here we are updating industry as Other where industry is null.



Ex 4: To capitalise the names using UPPER function.  
Here, we want to make all the industry names in upper case. To achieve that we used upper() function on the name attribute.

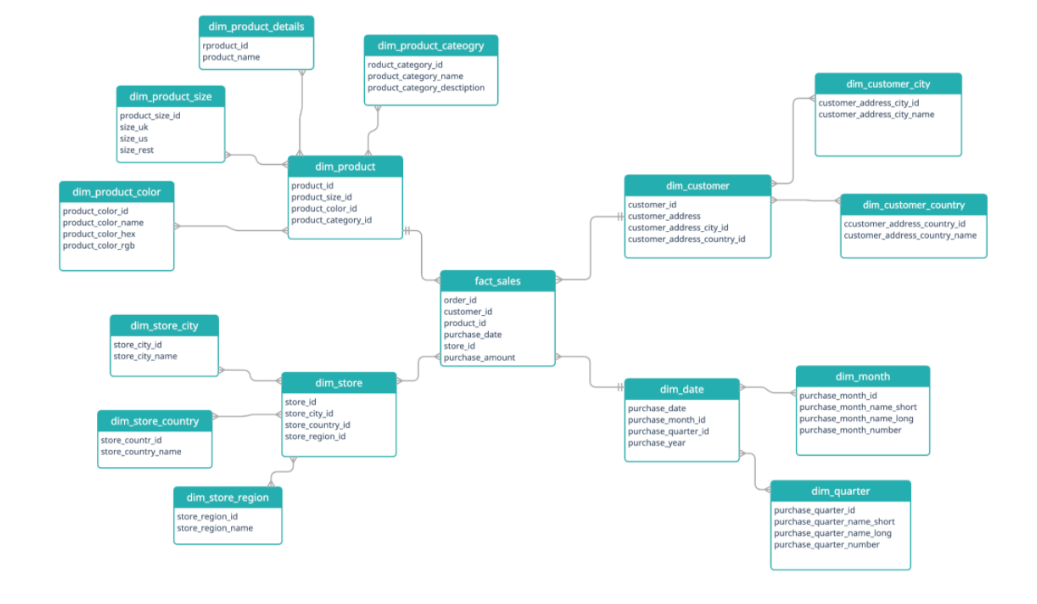


Ex 5: To correct logical errors in our data is also a part of data cleaning. In our table, for a name, the year\_founded is wrongly entered as 2032 after a keen analysis, we have identified that it is found in 2012. Below is the query to do the above.



**SnowFlaking Schema**

Snowflaking is like organising information in a more structured way to avoid repeating the same details. It's a bit like making a star-shaped diagram simpler by breaking down some parts into smaller pieces, forming a snowflake shape. This helps in keeping data neat, reducing mistakes, and making it easier to manage, even though it might make searching for some things a bit more complicated. It's a trade-off between making things look nice and keeping them easy to use.



**Partition By Clause:**

The PARTITION BY clause in SQL is like dividing that table into smaller groups based on a specific column. When you use functions like SUM or AVG with PARTITION BY, they work separately within each of these groups. It's a way to organise your calculations into smaller sections of your data.

Example:

SELECT id, ROW\_NUMBER() OVER (PARTITION BY name ORDER BY id) AS rn

FROM Company