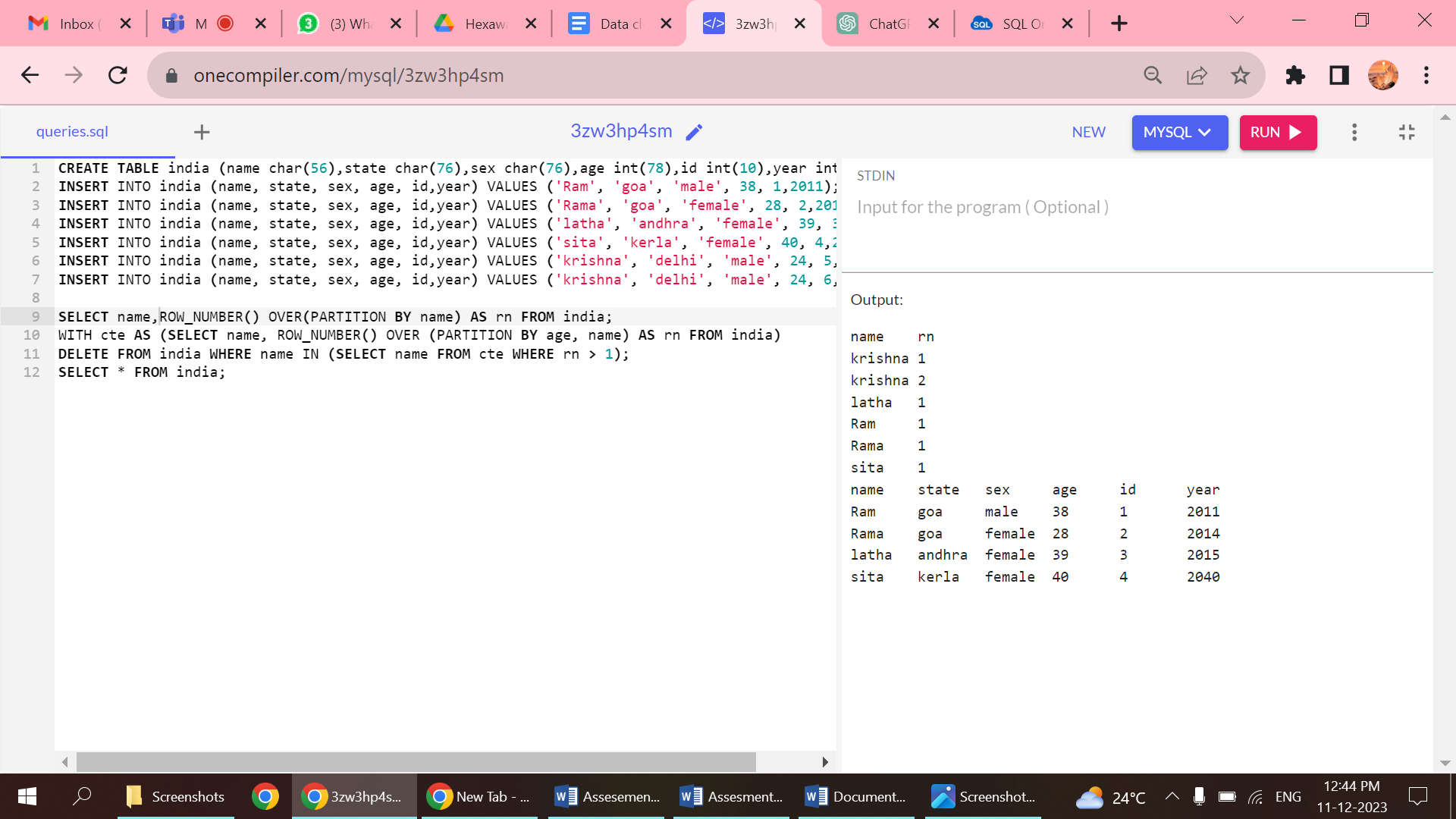
**DATA CLEANING**

DATE:11.12.2023

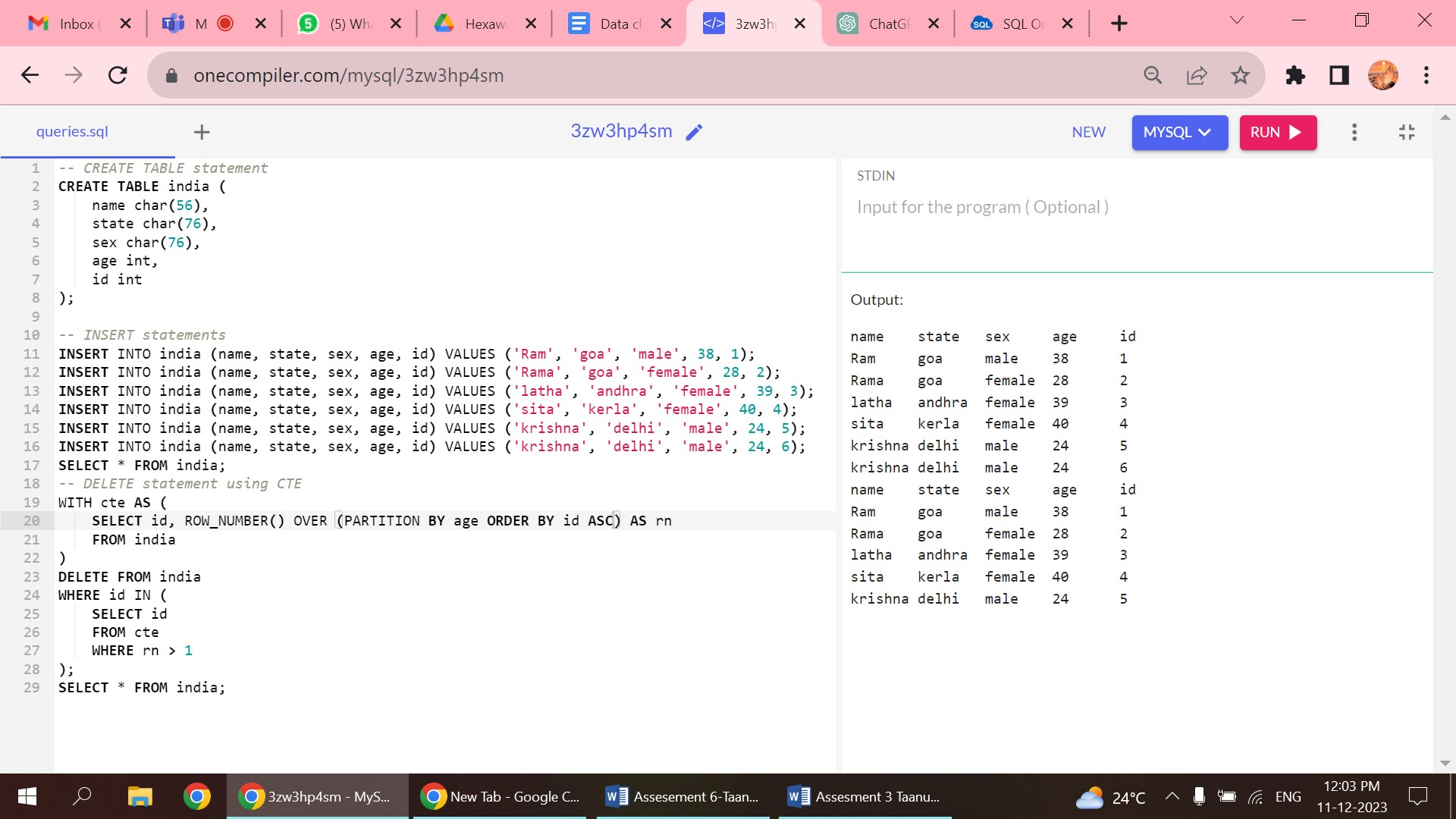
S.R.TAANUSRI

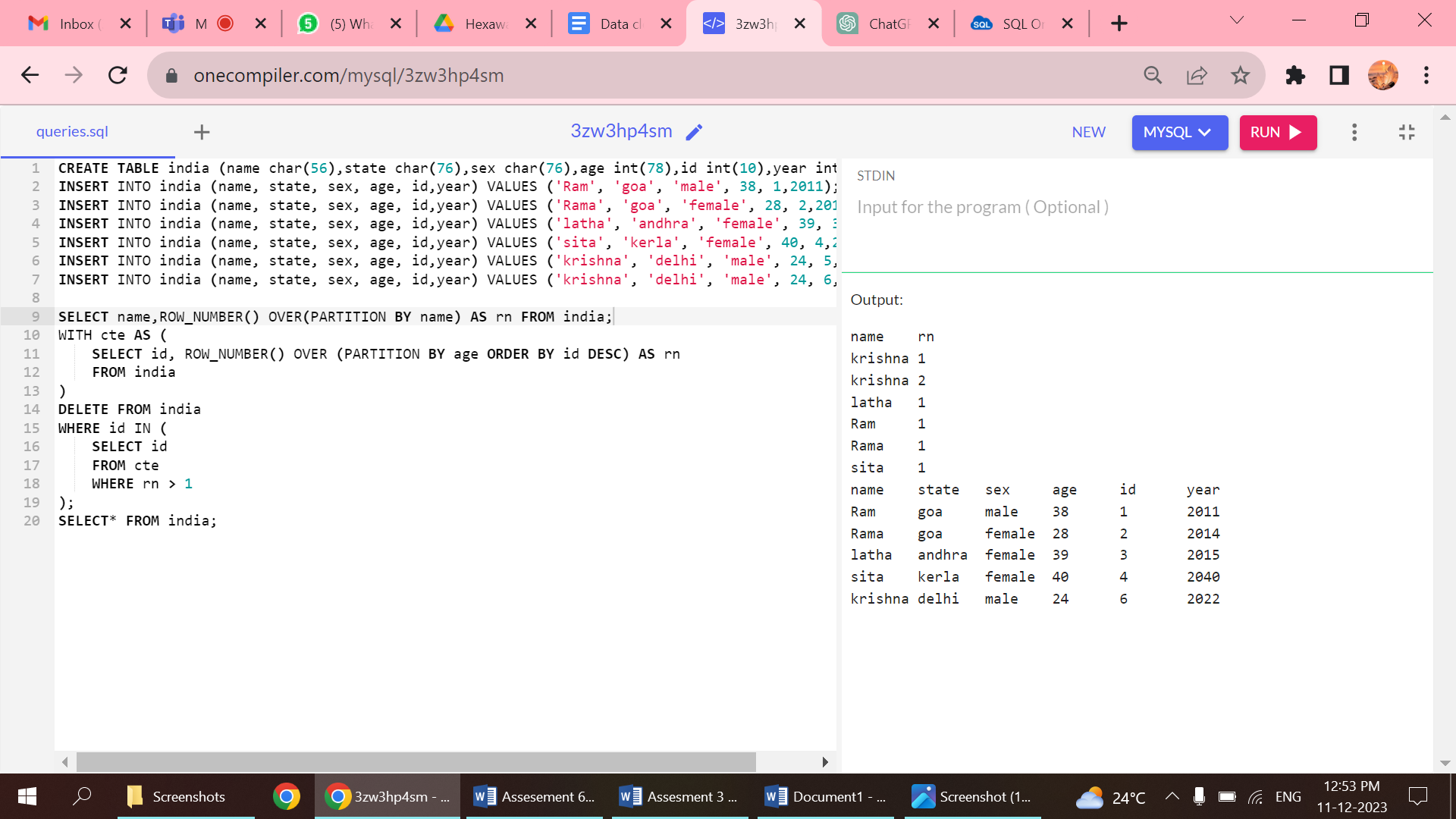
**DATA CLEANING**

1.DELETING DUPLICATE DATA

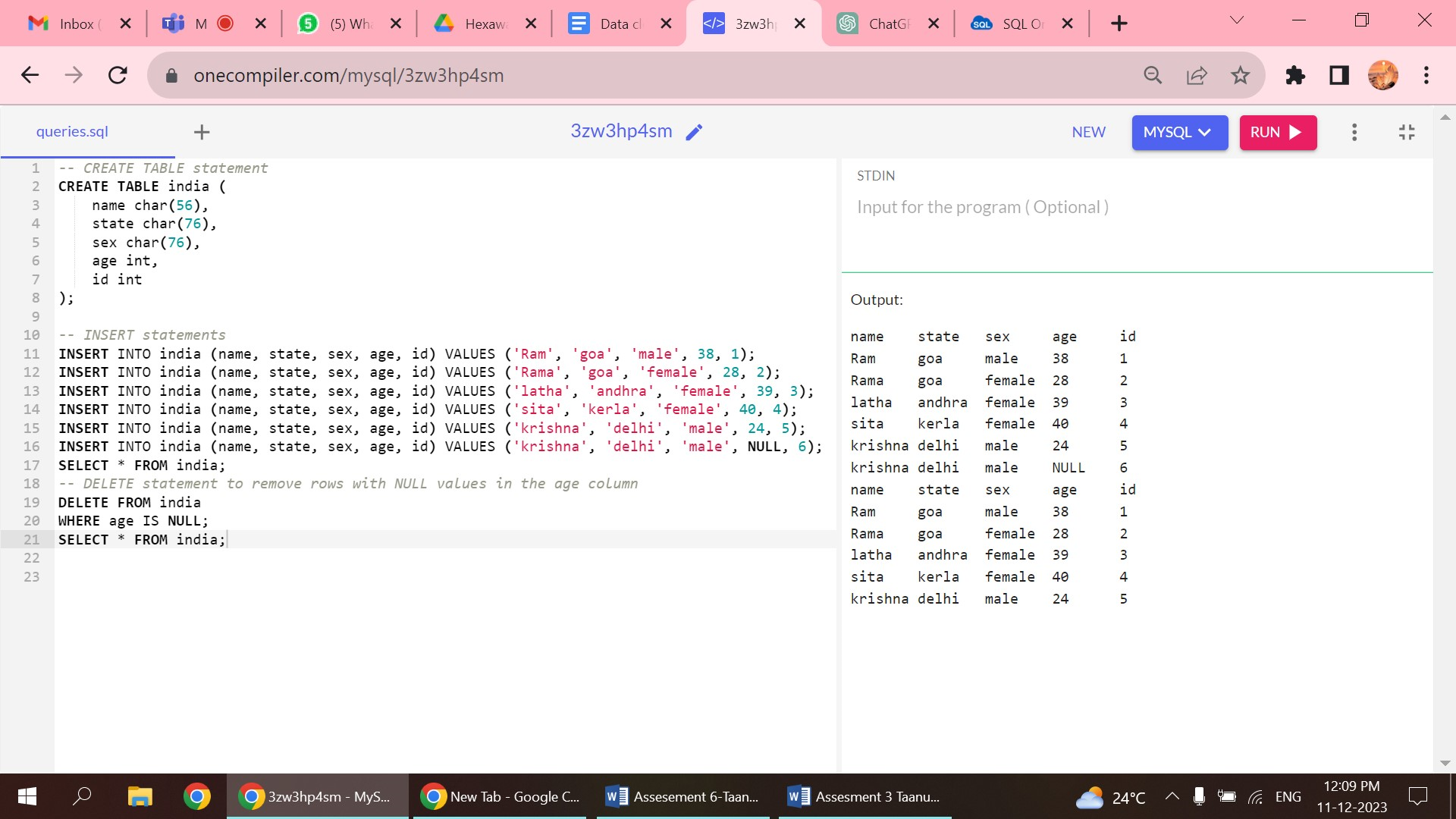


2.ODER BY BEFORE DELETION

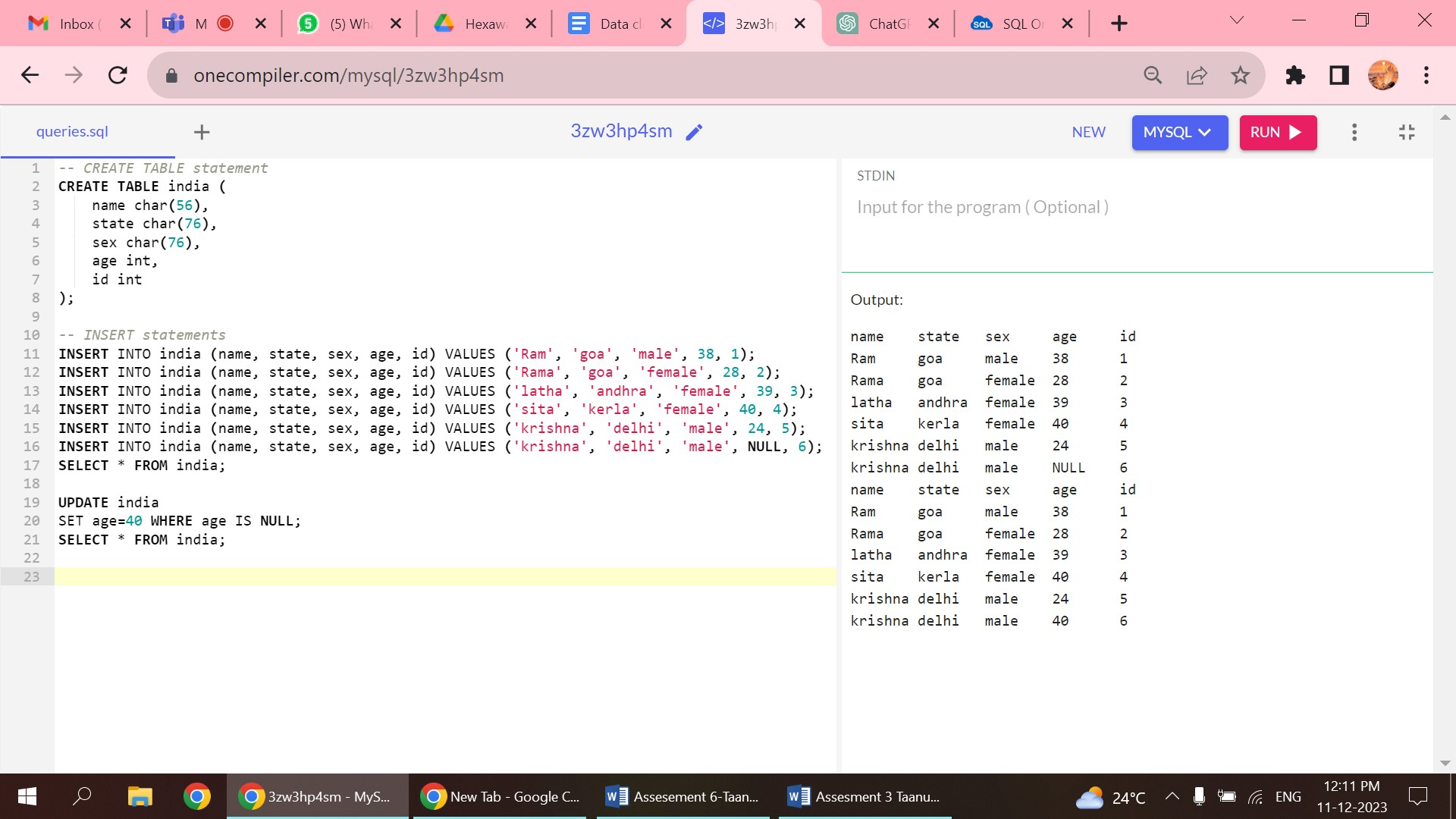


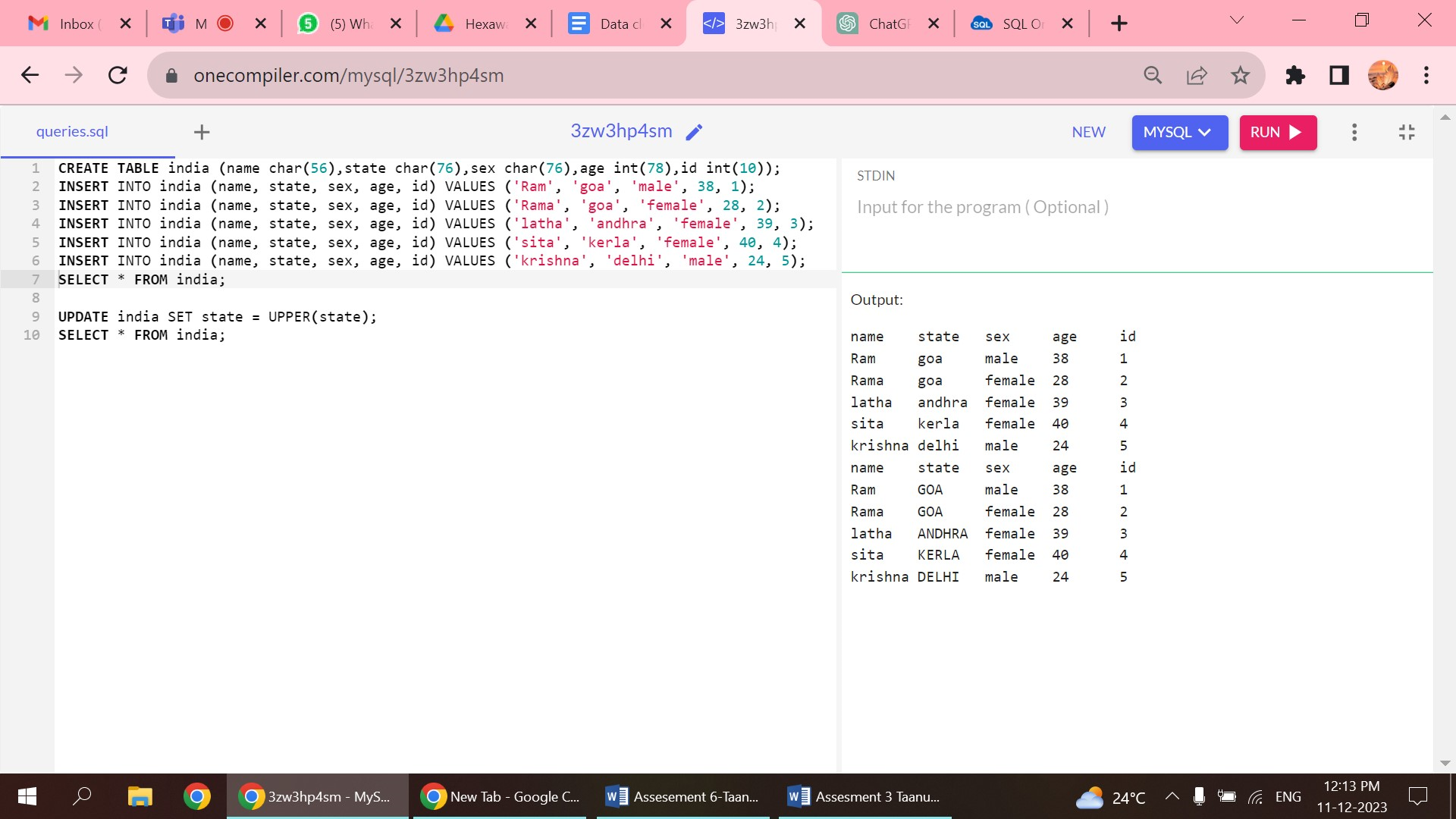


3.REMOVING NULL VALUES

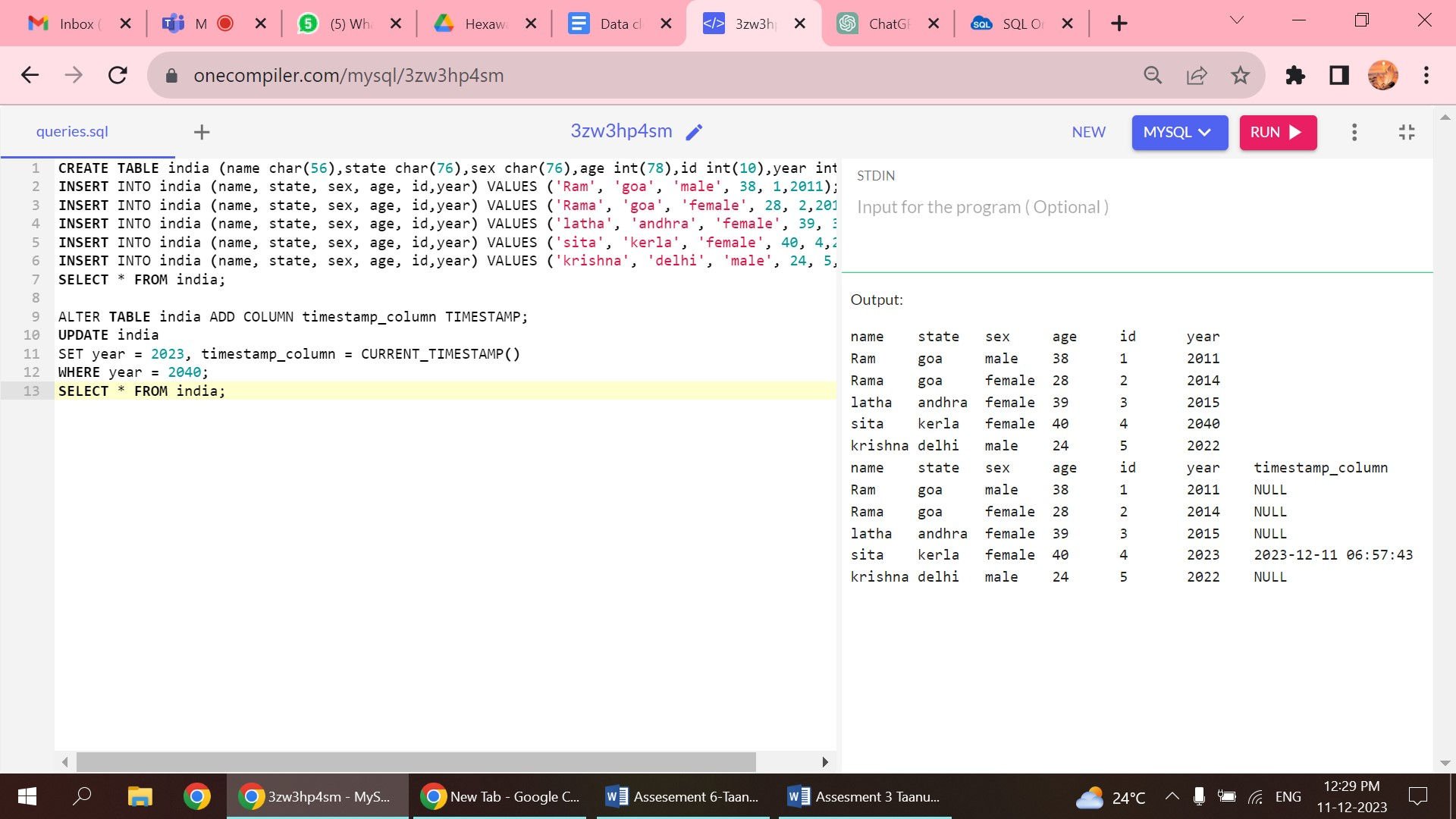


4.UPDATING DATA WHERE THERE ARE NULL VALUES

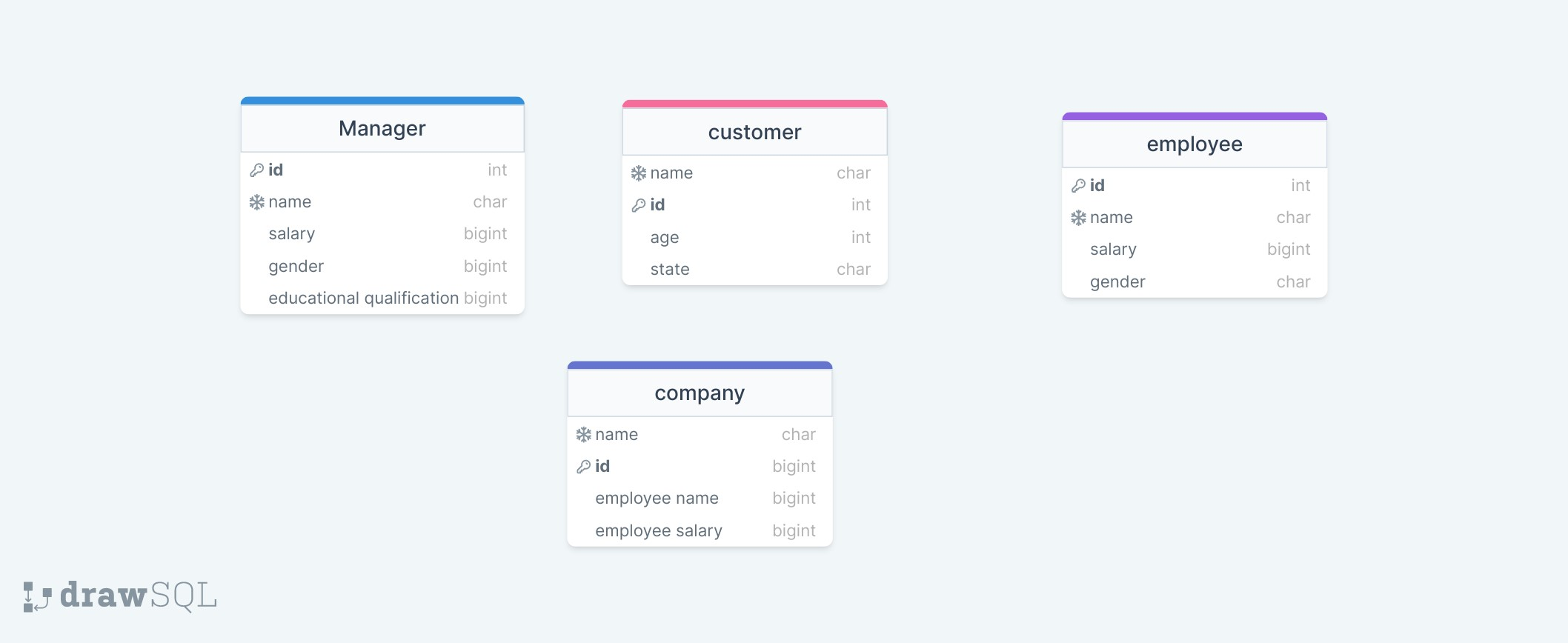
5.CAPITALIZTION OF VALUES



6.CORRECTING LOGICAL ERRORS



DATABASE SCHEMA DIAGRAM



SNOWFLAKE SCHEMA

The main purpose is to

* normalize dimension tables in a schema
* reducing redundancy
* improving data integrity
* simplifying maintenance
* optimizing query performance

Star Schema in Data Warehousing

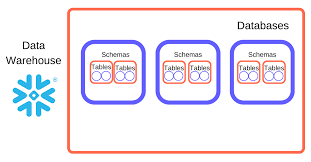
• Consists of a fact table that refers to multiple denormalized dimension tables.

• Denormalization helps to run query faster and efficiently.

• Dispensable data in dimensions cause challenges for data integrity.

Snowflake Schema in Data Warehousing

• It breaks down dimensions into multiple related tables in a snowflake pattern.



• Reduces unessential data by moving attributes with low cardinality into separate tables.

Example of Snowflake Schema

• Helps to understand the transformation from star schema to snowflake schema.

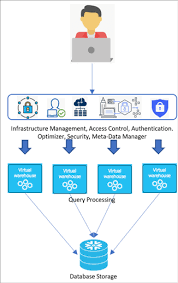
•Normalization done by creating separate tables and using foreign keys.

Purpose of Snowflake Schema

•Minimizes duplicated data, reducing vulnerability to data integrity issues.

• Simplifies data and makes modifications that demands a less disk space.

• It is used when data is challenging to denormalize for a particular query.



Considerations for Snowflaking

•Snowflake schema used for optimizing query performance and specific data characteristics.

•Calculate dimensions for sparsity, many-to-many relationships, size, and attribute hierarchies.

•Query performance is a main factor, but other factors include data maintenance and storage.



Recommendations and Controversies

• Certain sources dissuade the adoption of snowflake schemas due to concerns regarding simplicity and performance.

• Snowflaking may be applicable in specific scenarios, especially for particular query types.

• Factors to ponder include the nature of the data, business intelligence architecture, and query requirements.

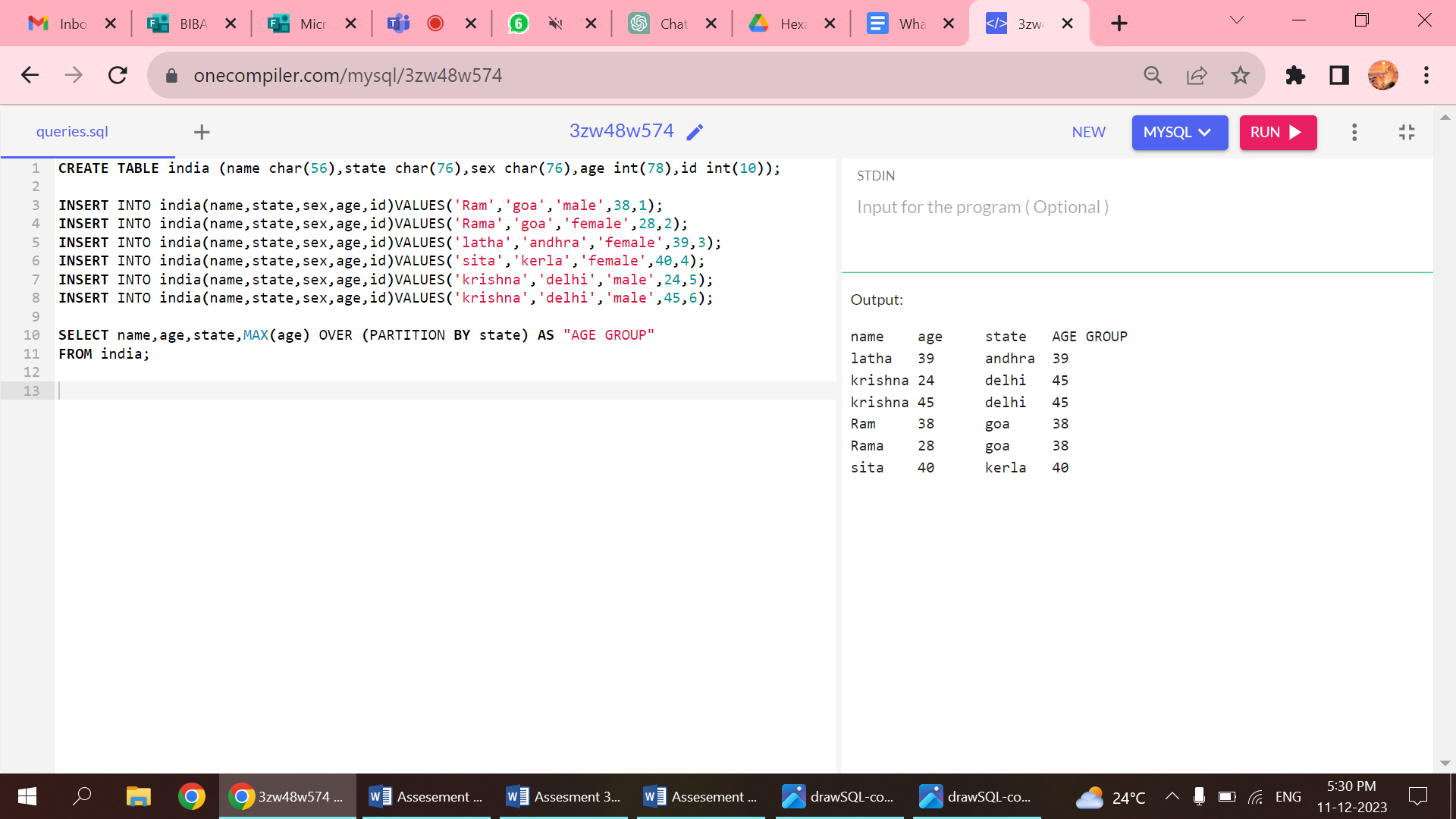
Conclusion

• Snowflaking is a technique that aims to balance query performance, data integrity, and maintenance in data warehousing.

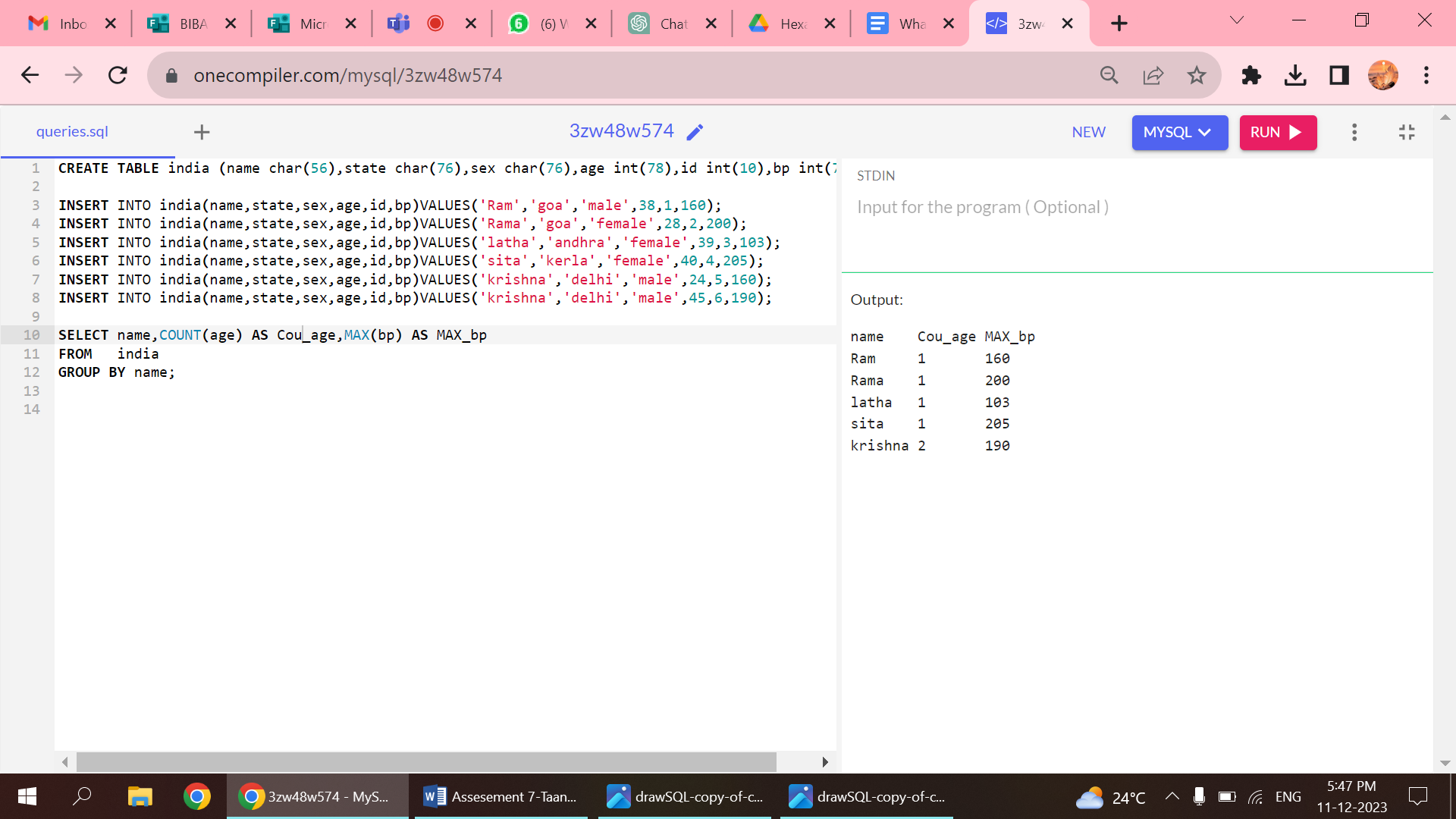
• The decision to use a snowflake schema depends on the characteristics of the data and the specific requirements of the queries and applications involved.

PARTITION BY CLAUSE

1.PARTITION CLAUSE



2.GROUP BY



3.USING AGGREGATE FUNCTIONS WITH PARTITION CLAUSE:

