Data cleaning:involves identifying inaccuracies in a dataset and then fixing them.

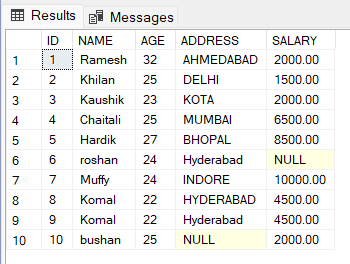
Check for missing values

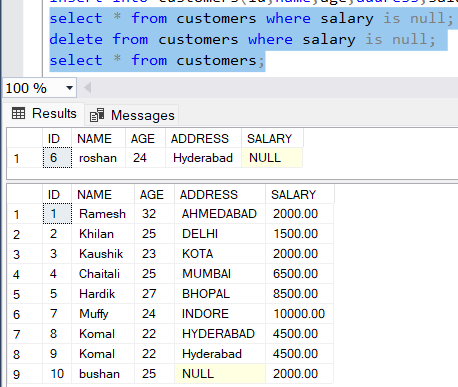
Check for duplicates

Standardizing and transformation:convert data into more efficient format.all uppercase or lower etc.

Updating data

Checking missing values and removing it:





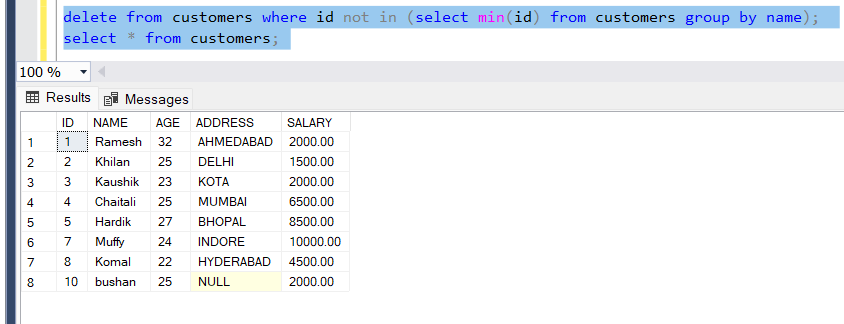
Checking duplicate values and removing it:

First how to find the duplicate rows:

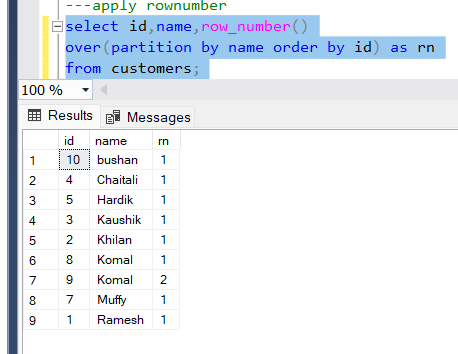
A screenshot of a computer

Description automatically generated

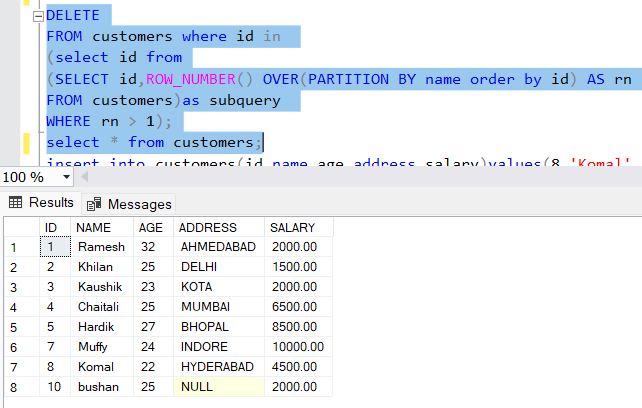
Deleting duplicate values:



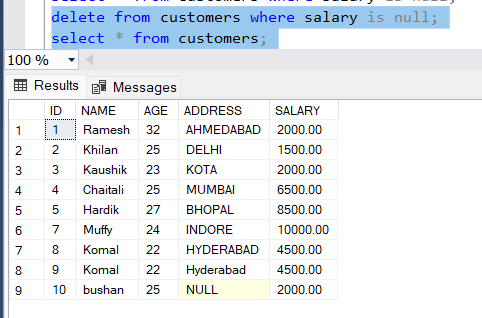
Using rownumber and partitions:



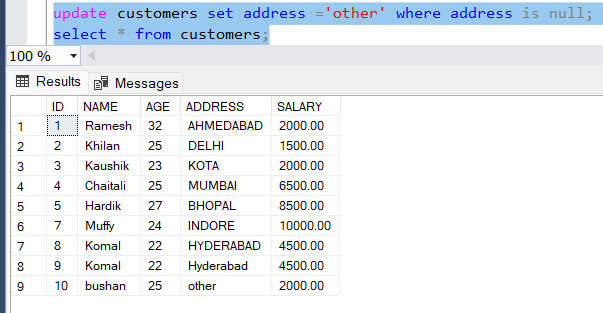
Now we can delete rows who have rn values >1



Removing null values:



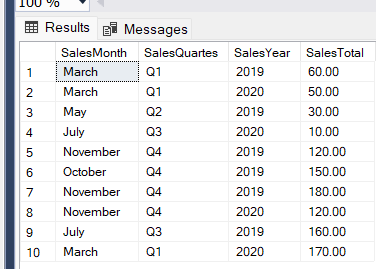
Updating null values:



Partition:

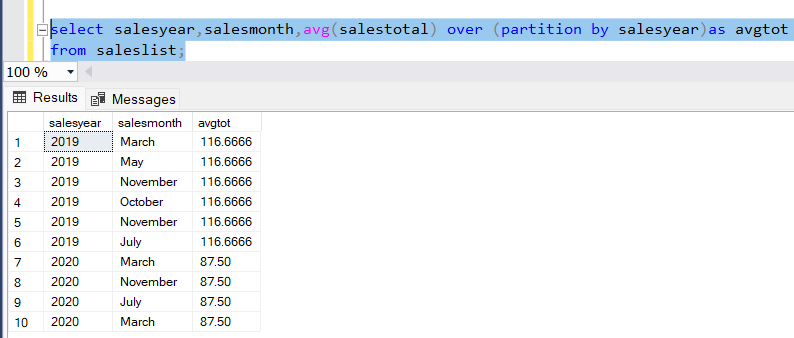
Is a subclass of over clause,it divides based on the given column types.

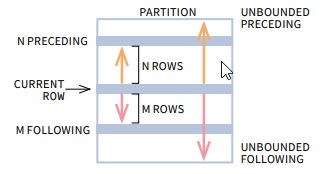
Below is saleslist table:



A screenshot of a computer

Description automatically generated





|  |  |  |
| --- | --- | --- |
| Just empty OVER() clause | UNBOUNDED PRECEDING | UNBOUNDED FOLLOWING |
| OVER(PARTITION BY …) | UNBOUNDED PRECEDING | UNBOUNDED FOLLOWING |
| OVER(PARTITION BY … ORDER BY …) | UNBOUNDED PRECEDING | CURRENT ROW |

A star schema consists of a central fact table that references multiple dimension tables. Each dimension table is denormalized ("flattened") to avoid the query overhead that comes with a highly normalized schema, which can require a large number of joins to retrieve the necessary data.

In a snowflake schema, a fact is surrounded by its associated dimensions (as in a star schema), but those dimensions are further related to other dimensions, branching out into a snowflake pattern. Snowflaking normalizes the dimensions by moving attributes with low cardinality into separate dimension tables.