**Cleaning data using MYSQL**

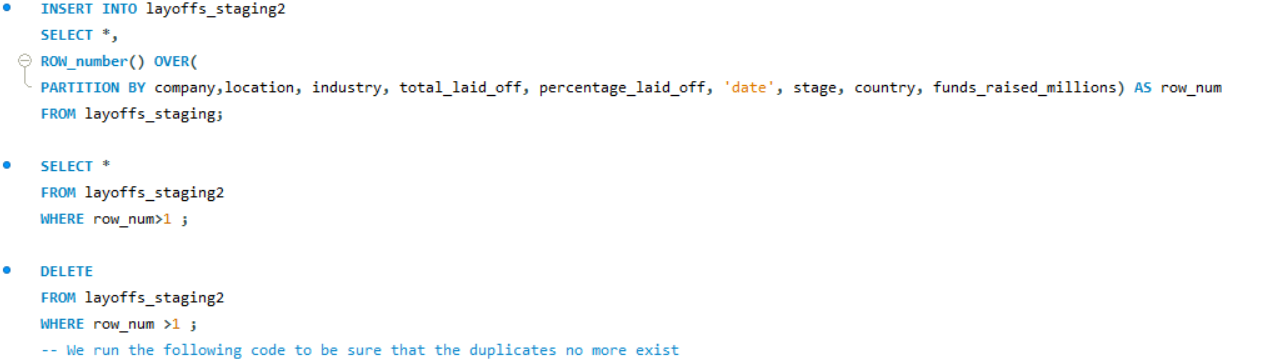
How does ommiting duplicates work in MYSQL works while using partitions ?

When row\_number() is used in conjonction with partition by , it asssigns a sequence number to eah row within a partition. The partitions in this case are determined by the unique combinations of the columns listed (company, industry, total\_laid\_off, percentage\_laid\_off, 'date').

1. The database first groups rows that have the same values for (company, industry, total\_laid\_off, percentage\_laid\_off, 'date').
2. Within each group row\_number() assigns a sequential integer starting from 1 to each row that is stored in ‘row\_num’ column. The number 1 means that the row do not have a duplicate and more than number one means that the row has a duplicate .

The ordering withn each partition is arbitary unless an ‘order by ‘ clause is specified within the ‘over’ clause.

This is very useful because right after this I created a table that contains all the columns including row\_num then I deleted the rows where row\_num > 1. And bay bay to the duplicates !!



Right after removing duplicates, I moved to standardizing the data, which involves identifying issues in the data and fixing them. I began by examining the columns and noticed the presence of some trailing spaces, which I removed using the TRIM() function in SQL. Then, I updated the table to save these changes. Additionally, I addressed variations in country names; for example, "the United States" was written in multiple ways, so I ensured that it was consistently represented in a unique way. Finally, I converted the format of the date column from string to date and updated the table with these changes.

While working with NULL and blank values, we need to be extremely careful. We can omit the data when it has nothing to do with the goal of the project, as I did with the rows having NULL values in the total\_laid\_off and percentage\_laid\_off columns, because our goal is to study the layoffs.

While examining the data, I noticed some blank spaces in the industry column that could be filled with the correct data. So, I performed a self-join to copy the data to the blank spaces. Then, I updated the tables with these changes.

All in all, what helped me in this project was paying attention to the data and identifying any illogical characters or spaces. I removed duplicates and non-interesting rows and columns, and also filled blank spaces with the correct data,changed the date column to its right type. This approach ensured that the data was clean and suitable for analysis.