Cleaning Layoff Dataset With SQL

Introduction

In recent years, layoffs have become a significant topic of concern in the business world, impacting employees and employers alike. Understanding the patterns and causes of layoffs can provide valuable insights for organizations and policymakers to mitigate their negative effects.

The purpose of this document is to show how to clean and prepare a layoff dataset using SQL, ensuring that the analysis is based on accurate and reliable data.

The Four Steps to Clean the Data

1- Removing Duplicates

An essential part of the data preparation involved removing duplicate entries to ensure data accuracy. This was achieved using the ROW_NUMBER window function in SQL. By assigning a unique sequential integer to rows within a partition of the dataset, duplicates were identified and removed effectively. This process ensured that only the most relevant and distinct records were retained for analysis.

2- Standardize Data

As part of the data preparation, standardization was a key focus to ensure consistency and accuracy across the dataset. The TRIM function was used to remove any leading or trailing whitespace in the necessary columns, preventing formatting inconsistencies. Additionally, efforts were made to ensure that words with the same meaning were written uniformly, enhancing the dataset's coherence.

Another important step in the data cleaning process was converting the date column from a string format to a proper date format. This conversion facilitated more accurate date-based analyses and allowed for easier integration with analytical tools that rely on date data types.

This analysis aims to identify key trends and factors associated with layoffs, such as industry-specific risks, economic conditions, and organizational changes. By exploring these factors, we hope to uncover insights that can inform decision-making processes and strategies to better manage workforce reductions.

3- Dealing with null and empty cells

Sometimes we need to change the empty cells to null values, in the other hand maybe we need to remove some null records from the dataset.

4- Removing columns OR rows

Sometimes we need to remove some columns which we don’t need to like the ROW_NUMBER column after using it in removing the duplicates, we need to remove the column after cleaning the data.