Data Cleaning Project

The layoff table contains data on layoffs that some companies made across the world from March 2020 to March 2023. It is made up of the following columns;

Column name	Data type	Explanation of columns
company	TEXT	Name of the company
location	TEXT	Location of the company
industry	TEXT	The industry the company operates in
total_laid_off	INT DEFAULT NULL	Total number of workers that were laid off within the period
percentage_laid_off	TEXT	The percentage of the company's workers that were laid off
date	TEXT	The date of the layoff
stage	TEXT	The stage in which the company was as at the time of the layoff
country	TEXT	The country in which the company is located
funds_raised_millions	INT	Funds raised by the company in millions as at the time of the layoff

Step-by-Step Guide to Cleaning Layoff Data in MySQL

This guide demonstrates the process of cleaning and preparing data on company layoffs for analysis in MySQL. The steps included creating databases and tables, removing duplicates, standardizing data, handling null values, and removing irrelevant columns.

1. I created a Database and named it Projects

2. The layoff data csv file was imported into MySQL

3. Imported data was checked to know the general scope of the data

4. A Staging Table was created so as to enable me to keep the raw file as a contingency plan

5. Duplicate Rows were identified using a CTE and Windows functions

6. A Second Staging Table was created for the safe removal of duplicates

7. Duplicate Rows were then safely deleted

8. Standardizing Data: the following steps were taken

a) Whitespaces in the company Column were removed using the TRIM function

b) Inconsistent names in the industry columns were identified by first selecting the DISTINCT industry and were standardized to make them consistent across the dataset

c) The country column was standardized by removing trailing periods in the country name "United States"

d) The data type for the date column was changed from text to date format, and the column type was modified to date

9. The data was checked for blank or null values in key columns

10. Blank industry names for companies with multiple rows were imputed by inference of the other rows

11. Columns that were irrelevant to the analysis were removed

12. Final checks were done by verifying the cleaned data.

By following these steps, I cleaned and prepared the layoff data for further analysis.