SQL Data Cleaning Project Report
Layoffs Dataset (2020–2025)

1. Introduction

This project focuses on cleaning and preprocessing the Layoffs (2020–2025) dataset sourced from Kaggle. The dataset captures global layoff events across industries over multiple years. SQL-based data cleaning techniques were applied to improve data quality and reliability.

2. Dataset Overview

The dataset includes company name, location, total laid off, percentage laid off, industry, funding stage, funds raised, country, layoff date, source, and date added. Because the data spans 2020–2025 and is reported by multiple sources, preprocessing was required.

3. Tools and Technologies Used

- MySQL
- SQL Window Functions (ROW_NUMBER())
- Kaggle Dataset

4. Data Cleaning Methodology

4.1 Staging Table Creation

A staging table (layoff2) was created to prevent changes to raw data.

4.2 Duplicate Removal

Duplicates were identified using ROW_NUMBER() across business-relevant columns. Companies such as Beyond Meat, Cars24, and Cazoo had duplicate records differing only in source and date added, which were treated as duplicates.

Due to MySQL limitations on deleting from CTEs, a new table (layoff3) was created to safely remove duplicate rows.

4.3 Standardization and Null Handling

Company names were trimmed, dates were converted to DATE format, and blank values were replaced with NULL. Rows missing both layoff count and percentage were removed.

5. Final Dataset

The final cleaned dataset spans 2020–2025 and is ready for analysis and visualization.

6. Conclusion

This project demonstrates practical SQL data cleaning skills applicable to real-world analytics tasks.