

# Title: SQL Data Cleaning – Layoffs Dataset

## Subtitle:

*A practice project using a real and publicly available dataset, showcasing SQL-based data cleaning & transformation skills.*

**By Shoeb Md Ashraf**

## Tools Used:

SQL (MySQL), Excel

## Project Overview

This project involved cleaning a messy layoffs dataset using SQL. The data contained duplicate records, inconsistent formatting, missing values, and incorrectly typed fields. Using SQL functions like `ROW_NUMBER()`, `STR_TO_DATE()`, `TRIM()`, and `JOIN`, the dataset was cleaned and transformed into an analysis-ready format.

## Key Steps

- Removed exact duplicate records using `ROW_NUMBER()`
- Standardized inconsistent entries (e.g., “Crypto” vs “Cryptocurrency”)
- Trimmed extra spaces and trailing characters in text fields
- Converted date fields from text to proper `DATE` type using `STR_TO_DATE()`
- Handled and filled missing values where possible
- Deleted rows with null critical values
- Removed unnecessary columns post-cleaning

# SQL Data Cleaning – Step-by-Step Summary

## ◆ 1. Understanding and Preparing the Dataset

- Loaded the raw layoffs dataset into MySQL.
- Duplicated the dataset into a working table `layoffs_duplicate` for safe transformation.

**Raw Data Screen:**

```
1 #--- Layoff Data Cleaning---#
2
3 select *
4 from layoffs;
5
6 ## Understanding the dataset and the EDA goal, following data cleaning steps are planned:
7 -- 1. Remove duplicates
8 -- 2. Standardize the data, spelling etc
9 -- 3. Null values or blank fields
10 -- 4. Remove any columns
11 #
12
```

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions
Atlassian	Sydney	Other	500	0.05	3/6/2023	Post-IPO	Australia	210
SiriusXM	New York City	Media	475	0.08	3/6/2023	Post-IPO	United States	525
Alerzo	Ibadan	Retail	400	NA	3/6/2023	Series B	Nigeria	16
UpGrad	Mumbai	Education	120	NA	3/6/2023	Unknown	India	631
Loft	Sao Paulo	Real Estate	340	0.15	3/3/2023	Unknown	Brazil	788
Embark Trucks	SF Bay Area	Transportation	230	0.7	3/3/2023	Post-IPO	United States	317
Lendi	Sydney	Real Estate	100	NA	3/3/2023	Unknown	Australia	59
UserTesting	SF Bay Area	Marketing	63	NA	3/3/2023	Acquired	United States	152
Airbnb	SF Bay Area	Marketing	30	NA	3/3/2023	Post-IPO	United States	6400
Accolade	Seattle	Healthcare	NA	NA	3/3/2023	Post-IPO	United States	458
Indigo	Boston	Other	NA	NA	3/3/2023	Series F	United States	1200
Zscaler	SF Bay Area	Security	177	0.03	3/2/2023	Post-IPO	United States	148
MasterClass	SF Bay Area	Education	79	NA	3/2/2023	Series E	United States	461
Ambiv Tech	Blumenau	Food	50	NA	3/2/2023	Acquired	Brazil	NA
Fittr	Pune	Fitness	30	0.11	3/2/2023	Series A	India	13
CNET	SF Bay Area	Media	12	0.1	3/2/2023	Acquired	United States	20
Flipkart	Bengaluru	Retail	NA	NA	3/2/2023	Acquired	India	12900

## ◆ 2. Removing Duplicate Rows

- Since there was no unique ID, used `ROW_NUMBER()` to identify exact duplicate rows.
- Created a new table `layoffs_duplicate_2` with a `row_num` column.
- Deleted all rows with `row_num > 1`, keeping only the first occurrence of each record.

**Duplicated Rows Identified:**

Layouts Data Clearing

```

19 from layoffs;
20
21 • select *
22 from layoffs_duplicate;
23
24 ## This dataset doesn't have any Unique ID, so I can't check for duplicate rows directly#
25 # So, I created a new row with row number which works as a Unique ID#
26 # It shows that there are duplicated rows in the dataset#
27 • select *,
28 row_number() over(partition by company, location, industry, total_laid_off,
29 percentage_laid_off, 'date', stage, country, funds_raised_millions) as row_num
30 from layoffs_duplicate
31 order by row_num desc;
32
33

```

Result Grid

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Better.com	New York City	Real Estate	NULL	NULL	8/6/2022	Unknown	United States	905	2
Casper	New York City	Retail	NULL	NULL	9/4/2021	Post-IPO	United States	339	2
Cazoo	London	Transportation	750	0.15	6/7/2022	Post-IPO	United Kingdom	2000	2
Elmly	SF Bay Area	Healthcare	NULL	NULL	7/5/2022	Series B	United States	323	2
Extrahop	Seattle	Security	NULL	NULL	12/20/2020	Series C	United States	61	2
Hibob	Tel Aviv	HR	70	0.3	1/30/2020	Series A	Israel	45	2
Fit	Logan	Fitness	NULL	NULL	1/25/2022	Private Equity	United States	200	2
Microsoft	Seattle	Other	NULL	NULL	10/17/2022	Post-IPO	United States	1	2
Ola	Bengaluru	Transportation	200	NULL	1/19/2022	Series J	India	5000	2
Olist	Curitiba	Retail	NULL	NULL	1/30/2023	Series E	Brazil	322	2
Oracle	SF Bay Area	Other	NULL	NULL	1/17/2023	Post-IPO	United States	NULL	2

Result 5 x

Output

Action Output

```

21 • select *,
22 row_number() over(partition by company, location, industry, total_laid_off,
23 percentage_laid_off, 'date', stage, country, funds_raised_millions) as row_num
24 from layoffs_staging;
25 ## first we checked here if there are duplicates by assigning row numbers,
26 ## and it shows there are duplicates
27
28

```

Result Grid

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country
E Inc.	Toronto	Transportation	NULL	NULL	12/16/2022	Post-IPO	Canada
E Inc.	Toronto	Transportation	NULL	NULL	12/16/2022	Post-IPO	Canada
Included Health	SF Bay Area	Healthcare	NULL	0.06	7/25/2022	Series E	United States
Included Health	SF Bay Area	Healthcare	NULL	0.06	7/25/2022	Series E	United States
&Open	Dublin	Marketing	9	0.09	11/17/2022	Series A	Ireland
&Open	Dublin	Marketing	9	0.09	11/17/2022	Series A	Ireland
#Paid	Toronto	Marketing	19	0.17	1/27/2023	Series B	Canada
#Paid	Toronto	Marketing	19	0.17	1/27/2023	Series B	Canada

✓ Duplicates cleaned successfully.

Layouts Data Cleaning

```

60
61 • insert into layoffs_duplicate_2
62 select *,
63 row_number() over(partition by company, location, industry, total_laid_off,
64 percentage_laid_off, 'date', stage, country, funds_raised_millions) as row_num
65 from layoffs_staging;
66
67 • delete
68 from layoffs_duplicate_2
69 where row_num > 1;
70
71 • select *
72 from layoffs_duplicate_2
73 order by row_num desc;
74 ## Now there is no more duplicated rows#
75

```

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
E Inc.	Toronto	Transportation	NULL	NULL	12/16/2022	Post-IPO	Canada	NULL	1
Included Health	SF Bay Area	Healthcare	NULL	0.06	7/25/2022	Series E	United States	272	1
8Open	Dublin	Marketing	9	0.09	11/17/2022	Series A	Ireland	35	1
#Paid	Toronto	Marketing	19	0.17	1/27/2023	Series B	Canada	21	1
100 Thieves	Los Angeles	Consumer	12	NULL	7/13/2022	Series C	United States	120	1
100 Thieves	Los Angeles	Retail	NULL	NULL	1/10/2023	Series C	United States	120	1
10X Genomics	SF Bay Area	Healthcare	100	0.08	8/4/2022	Post-IPO	United States	242	1
1stdibs	New York City	Retail	70	0.17	4/2/2020	Series D	United States	253	1
2TH	Sao Paulo	Crypto	90	0.12	6/1/2022	Unknown	Brazil	250	1
2TH	Sao Paulo	Crypto	100	0.15	9/1/2022	Unknown	Brazil	250	1
ZU	Washington ...	Education	NULL	0.2	7/28/2022	Post-IPO	United States	426	1
Kabana	Washington	Healthcare	0K	0.2	8/29/2022	Series B	United States	44	1

layoffs\_duplicate\_217

Output

### ◆ 3. Standardizing Text Fields

- Trimmed leading/trailing spaces using **TRIM()**.
- Fixed inconsistent naming (e.g., “Crypto” vs “Cryptocurrency”) using **UPDATE ... WHERE LIKE**.  
Cleaned up country values (e.g., removed trailing dots in “United States.”)

✓ Text values standardized for consistent grouping.

### ◆ 4. Handling Null and Blank Values

- Deleted rows where both **total\_laid\_off** and **percentage\_laid\_off** were **NULL**.
- Used a **SELF JOIN** to impute missing **industry** values from matching companies.
- Where no imputation was possible (e.g., “Bally’s”), left **NULL** or deleted.

**Null and Blank Fields Screenshot:** Null values that still remain are accepted in this case.

Layouts Data Cleaning

```

77
78
79  ##-----Standardization-----##
80
81 • select company, trim(company)
82   from layoffs_duplicate_2;
83
84 • update layoffs_duplicate_2
85   set company = trim(company);
86   ## Company column trimmed##
87
88 • select distinct(industry)
89   from layoffs_duplicate_2
90   order by 1;
91
92

```

Result Grid | Filter Rows: | Exports | Wrap Cell Content: |

industry

NULL

Aerospace

Construction

Crypto

Crypto Currency

CryptoCurrency

Data

Education

Energy

Fin-Tech

Result 20 x

Output

Read Only

Layouts Data Cleaning

```

114 where country like 'United States%';
115
116 • select *
117   from layoffs_duplicate_2;
118
119  ##-----Null Value-----##
120
121 • select *
122   from layoffs_duplicate_2
123   where total_laid_off is null
124   and percentage_laid_off is null;
125
126
127
128

```

Result Grid | Filter Rows: | Exports | Wrap Cell Content: |

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
E Inc.	Toronto	Transportation	NULL	NULL	12/16/2022	Post-IPO	Canada	NULL	1
100 Thieves	Los Angeles	Retail	NULL	NULL	1/10/2023	Series C	United States	120	1
Accolade	Seattle	Healthcare	NULL	NULL	3/3/2023	Post-IPO	United States	458	1
Ada	Toronto	Support	NULL	NULL	2/1/2023	Series C	Canada	190	1
Adara	SF Bay Area	Travel	NULL	NULL	3/31/2020	Series C	United States	67	1
Addi	Bogota	Finance	NULL	NULL	6/14/2022	Series C	Colombia	376	1
AirMap	Los Angeles	Aerospace	NULL	NULL	4/30/2020	Unknown	United States	75	1
Airtasker	Sydney	Consumer	NULL	NULL	7/4/2022	Series C	Australia	26	1
Akerna	Denver	Logistics	NULL	NULL	9/2/2020	Post-IPO	United States	NULL	1
Akerna	Denver	Logistics	NULL	NULL	5/27/2022	Unknown	United States	46	1

layoffs\_duplicate\_231 x

Output

Action Output

Read Only

✓ Missing data handled appropriately.

Layoffs Data Cleaning

```

162 where total_laid_off is null
163 and percentage_laid_off is null;
164
165 • delete
166 from layoffs_duplicate_2
167 where total_laid_off is null
168 and percentage_laid_off is null;
169
170 ## I have deleted the row_num column because it is not needed anymore#
171 • alter table layoffs_duplicate_2
172 drop column row_num;
173
174 • select *
175 from layoffs_duplicate_2;
176
177

```

Result Grid

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions
ZTM	Sao Paulo	Crypto	90	0.12	6/1/2022	Unknown	Brazil	250
ZTM	Sao Paulo	Crypto	100	0.15	9/1/2022	Unknown	Brazil	250
2U	Washington D.C.	Education	100	0.2	7/28/2022	Post-IPO	United States	426
54gene	Washington D.C.	Healthcare	95	0.3	8/29/2022	Series B	United States	44
58 Solar	Sydney	Energy	100	0.25	6/3/2022	Series A	Australia	12
6sense	SF Bay Area	Sales	150	0.1	10/12/2022	Series E	United States	426
80 Acres Farms	Cincinnati	Food	100	0.1	1/18/2023	Unknown	United States	275
8x8	SF Bay Area	Support	155	0.07	1/18/2023	Post-IPO	United States	253
8x8	SF Bay Area	Support	155	0.07	1/18/2023	Post-IPO	United States	253

Output

## ◆ 5. Fixing the Date Format

- The original `date` column was stored as text (MM/DD/YYYY).
- Converted it using `STR_TO_DATE()`.
- Altered the column type to `DATE`.

**Date Column was not appropriate: see next page**

Layoffs Data Cleaning

```

173
174 • select *
175 from layoffs_duplicate_2;
176
177 #-----Fix the Date Column-----#
178
179 • select `date`
180 from layoffs_duplicate_2
181 ; ## this is currently a text column, which should be date column actually
182
183 • select `date`,
184 str_to_date(`date`, '%m/%d/%Y') as Date -- this formatting is very sensitive and has varieties,
185 ##here we have said mysql what is there in the positions and
186 ##it converted them into standard american format
187 from layoffs_staging2
188

```

Result Grid

date
7/25/2022
11/17/2022
1/27/2023
7/13/2022
8/4/2022
4/2/2020
6/1/2022
9/1/2022

Output

Action Output

#	Time	Action	Message	Duration / Fetch
68	17:21:23	select * from layoffs_duplicate_2 LIMIT 0, 1000	1000 row(s) returned	0.000 sec / 0.000 sec

✓ **Date field converted and ready for time-series or trend analysis.**

The screenshot shows a data cleaning tool interface with a SQL editor and a result grid. The SQL editor contains the following queries:

```
183 • select `date`,
184       str_to_date(`date`, '%m/%d/%Y') as Date
185   from layoffs_duplicate_2;
186   ##here we have said mysql what is there in the positions and
187   ##it converted them into standard american format
188
189 • update layoffs_duplicate_2
190   set `date` = str_to_date(`date`, '%m/%d/%Y');
191   ## but it still shows text column, we have to convert it now
192
193 • alter table layoffs_duplicate_2
194   modify column `date` Date;
195
196 • select*
197   from layoffs_duplicate_2;
198
199 ...
```

The result grid displays the following data:

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions
Included Health	SF Bay Area	Healthcare	1000	0.06	2022-07-25	Series E	United States	272
8Open	Dublin	Marketing	9	0.09	2022-11-17	Series A	Ireland	35
#Paid	Toronto	Marketing	19	0.17	2023-01-27	Series B	Canada	21
100 Thieves	Los Angeles	Consumer	12	100%	2022-07-13	Series C	United States	120
10X Genomics	SF Bay Area	Healthcare	100	0.08	2022-08-04	Post-IPO	United States	242
Istdibs	New York City	Retail	70	0.17	2020-04-02	Series D	United States	253
ZTM	Sao Paulo	Crypto	90	0.12	2022-06-01	Unknown	Brazil	250
ZTM	Sao Paulo	Crypto	100	0.15	2022-09-01	Unknown	Brazil	250
ZU	Washington D.C.	Education	1000	0.2	2022-07-28	Post-IPO	United States	426

The 'date' column is highlighted with a green circle. The interface also includes a 'Filter Rows' section, an 'Exports' button, and a 'Wrap Cell Contents' checkbox.

## ◆ 6. Finalizing the Dataset

- Dropped unnecessary columns like `row_num`.
- Verified the dataset was cleaned, structured, and ready for analysis or dashboard integration.

Layoffs Data Cleaning

```

193 set `date` = str_to_date(`date`, '%m/%d/%Y');
194 ## but it still shows text column, we have to convert it now
195
196 • alter table layoffs_duplicate_2
197   modify column `date` Date;
198
199 • select*
200   from layoffs_duplicate_2
201   order by company;
202
203 ##-----Data Cleaning is Completed-----#
204

```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows: |

	company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions
8	Open	Dublin	Marketing	9	0.09	2022-11-17	Series A	Ireland	35
9	Paid	Toronto	Marketing	19	0.17	2023-01-27	Series B	Canada	21
10	Genomics	SF Bay Area	Healthcare	100	0.08	2022-08-04	Post-IPO	United States	242
11	Idbs	New York City	Retail	70	0.17	2020-04-02	Series D	United States	253
12	TM	Sao Paulo	Crypto	90	0.12	2022-06-01	Unknown	Brazil	250
13	TM	Sao Paulo	Crypto	100	0.15	2022-09-01	Unknown	Brazil	250
14	Gene	Washington D.C.	Healthcare	95	0.3	2022-08-29	Series B	United States	44
15	ense	SF Bay Area	Sales	150	0.1	2022-10-12	Series E	United States	426
16	8	SF Bay Area	Support	155	0.07	2023-01-18	Post-IPO	United States	253
17	8	SF Bay Area	Support	200	0.09	2022-10-04	Post-IPO	United States	253
18	9	Sao Paulo	Transport...	75	0.02	2022-09-20	Acquired	Brazil	244

Result Grid | Form Editor | Field Types



The cleaned dataset is fully prepared for use in business dashboards, trend analysis, and financial or operational reporting.

👉 Refer to my Projects section for the SQL EDA (Exploratory Data Analysis) project using this cleaned dataset.