SQL Project

---Data cleaning---

This project utilizes SQL to analyze data on employee layoffs within a company. We'll clean the given data related to layoffs and prepare it for accurate analysis.

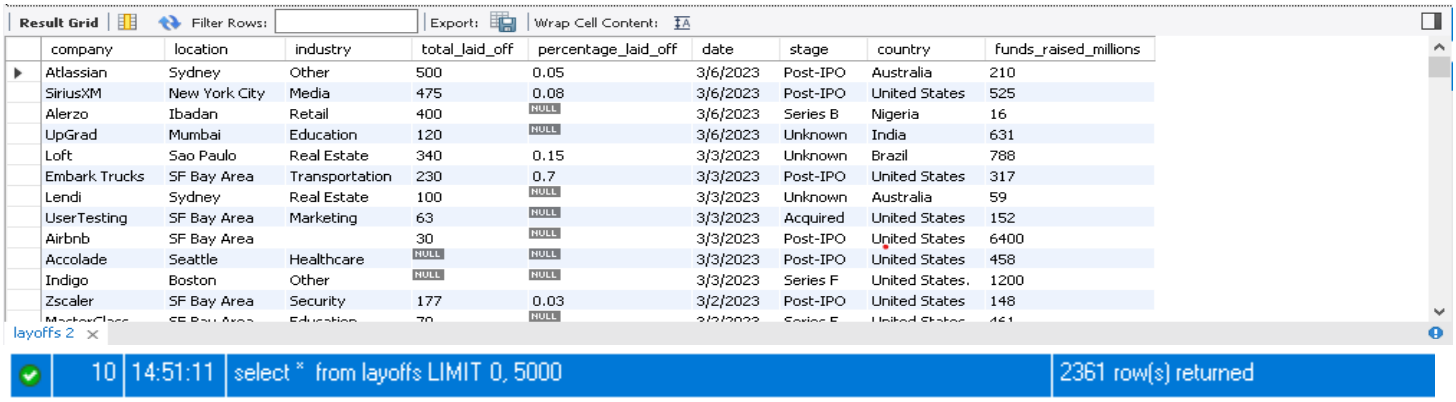
By working on this project. One will gain experience in:

- **Writing effective SQL queries for data manipulation and analysis.**
- **Cleaning and transforming data for meaningful insights.**

```
-- Activating the database on which work is going to be done--

use world_layoffs;

select *
from layoffs;
```



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

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	NULL	3/6/2023	Series B	Nigeria	16
UpGrad	Mumbai	Education	120	NULL	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	NULL	3/3/2023	Unknown	Australia	59
UserTesting	SF Bay Area	Marketing	63	NULL	3/3/2023	Acquired	United States	152
Airbnb	SF Bay Area		30	NULL	3/3/2023	Post-IPO	United States	6400
Accolade	Seattle	Healthcare	NULL	NULL	3/3/2023	Post-IPO	United States	458
Indigo	Boston	Other	NULL	NULL	3/3/2023	Series F	United States.	1200
Zscaler	SF Bay Area	Security	177	0.03	3/2/2023	Post-IPO	United States	148

layoffs 2

✓	10	14:51:11	select * from layoffs LIMIT 0, 5000	2361 row(s) returned

```
-- Creating a copy of a table 'layoffs'

create table layoffs_staging
like layoffs;
```

✓	11	14:56:55	create table layoffs_staging like layoffs	0 row(s) affected

```sql
select * from layoffs_staging;
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions

layoffs_staging 3

✓	12	14:57:10	select * from layoffs_staging LIMIT 0, 5000	0 row(s) returned

```sql
-- inserting data into 'layoffs_staging' table from 'layoffs' table.

insert layoffs_staging
select *
from layoffs;
```

✓	13	14:59:08	insert layoffs_staging select * from layoffs	2361 row(s) affected Records: 2361 Duplicates: 0 Warnings: 0

```sql
select * from layoffs_staging;
```

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

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	NULL	3/6/2023	Series B	Nigeria	16
UpGrad	Mumbai	Education	120	NULL	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	NULL	3/3/2023	Unknown	Australia	59
UserTesting	SF Bay Area	Marketing	63	NULL	3/3/2023	Acquired	United States	152

layoffs_staging 4

✓	14	14:59:24	select * from layoffs_staging LIMIT 0, 5000	2361 row(s) returned

==========

-- Data cleaning --

1. Remove duplicates.

2. Standardize the Data

3. Null values or Blank values

4. Remove any columns unnecessary.

==========

1. Removing Duplicates

```
-- checking for the duplicates rows using row_number and over functions as we do not have any unique key in the data

WITH duplicate_cte AS(
select * ,
row_number() over(
partition by company,location, industry, total_laid_off, percentage_laid_off, `date`, stage, country, funds_raised_millions) as row_num
from layoffs_staging
)
select * from duplicate_cte
where row_num>1;
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Casper	New York City	Retail	NULL	NULL	9/14/2021	Post-IPO	United States	339	2
Cazoo	London	Transportation	750	0.15	6/7/2022	Post-IPO	United Kingdom	2000	2
Hibob	Tel Aviv	HR	70	0.3	3/30/2020	Series A	Israel	45	2
Wildlife Studios	Sao Paulo	Consumer	300	0.2	11/28/2022	Unknown	Brazil	260	2
Yahoo	SF Bay Area	Consumer	1600	0.2	2/9/2023	Acquired	United States	6	2

Result 5

✓	15	15:03:14	WITH duplicate_cte AS(select * , row_number() over(partition by company,location...	5 row(s) returned

```sql
select *
from layoffs_staging
where company = 'casper';
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions
Casper	New York City	Retail	NULL	NULL	9/14/2021	Post-IPO	United States	339
Casper	New York City	Retail	78	0.21	4/21/2020	Post-IPO	United States	339
Casper	New York City	Retail	NULL	NULL	9/14/2021	Post-IPO	United States	339

layoffs_staging 6

✓	16	15:04:43	select * from layoffs_staging where company = 'casper' LIMIT 0, 5000	3 row(s) returned

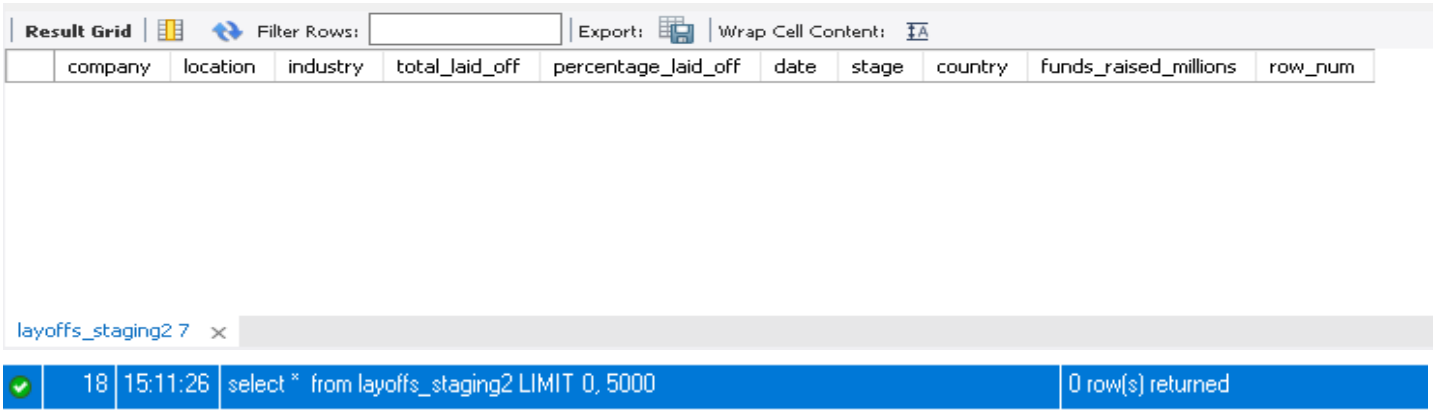
Note: Delete key cannot update CTE that's why we need to add a column named 'row_num' in a new table.

```sql
-- adding 'row_num' column in the table named 'layoffs_staging2' table.

CREATE TABLE `layoffs_staging2` (
  `company` text,
  `location` text,
  `industry` text,
  `total_laid_off` int DEFAULT NULL,
  `percentage_laid_off` text,
  `date` text,
  `stage` text,
  `country` text,
  `funds_raised_millions` int DEFAULT NULL,
  `row_num` int
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci;
```

✓	17	15:10:52	CREATE TABLE `layoffs_staging2` (`company` text, `location` text, `industry` tex...	0 row(s) affected

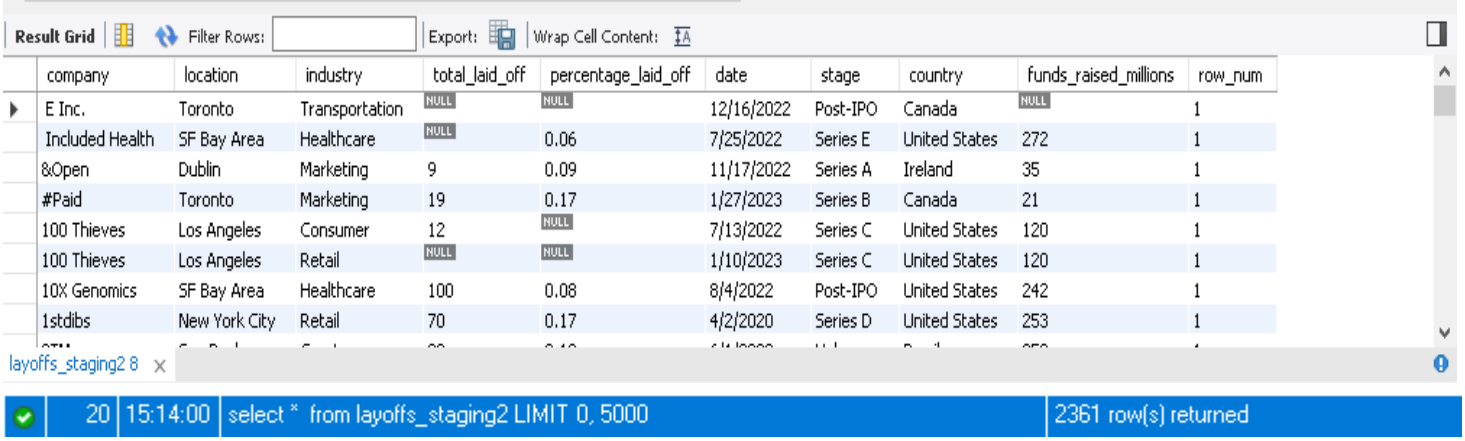
```sql
select *
from layoffs_staging2
```




```sql
-- inserting the data of 'layoffs_staging' table and 'row_num' column data by

INSERT INTO layoffs_staging2
select * ,
row_number() over(
partition by company,location, industry, total_laid_off, percentage_laid_off, `date`, stage, country, funds_raised_millions) as row_num
from layoffs_staging;
```


```sql
select *
from layoffs_staging2
```



```sql
select *
from layoffs_staging2
where row_num>1;
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Casper	New York City	Retail	NULL	NULL	9/14/2021	Post-IPO	United States	339	2
Cazoo	London	Transportation	750	0.15	6/7/2022	Post-IPO	United Kingdom	2000	2
Hibob	Tel Aviv	HR	70	0.3	3/30/2020	Series A	Israel	45	2
Wildlife Studios	Sao Paulo	Consumer	300	0.2	11/28/2022	Unknown	Brazil	260	2
Yahoo	SF Bay Area	Consumer	1600	0.2	2/9/2023	Acquired	United States	6	2

layoffs_staging2 9 ×

✓	21	15:17:27	select * from layoffs_staging2 where row_num>1 LIMIT 0, 5000	5 row(s) returned

```sql
-- deleting duplicates row

DELETE
from layoffs_staging2
where row_num>1;
```

✓	22	15:21:13	DELETE from layoffs_staging2 where row_num>1	5 row(s) affected

```sql
-- Checking whether the data is deleted or not.

select *
from layoffs_staging2
where row_num>1;
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num

layoffs_staging2 10 ×

✓	23	15:22:13	select * from layoffs_staging2 where row_num>1 LIMIT 0, 5000	0 row(s) returned

2. Standardizing data (Inconsistent Data Formatting)

(finding issues in the data and then fixing it)

Data Formatting

```
-- looking at column 'company'
-- Removing a space at the beginning in column 'Company'

UPDATE layoffs_staging2
SET company = trim(company);
```

```
24 | 15:24:38 | UPDATE layoffs_staging2 SET company = trim(company) | 11 row(s) affected Rows matched: 2356  Changed: 11  Warnings: 0
```

```
select *
from layoffs_staging2
```

Result Grid | Filter Rows: | Export: | 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
Included Health	SF Bay Area	Healthcare	NULL	0.06	7/25/2022	Series E	United States	272	1
&Open	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

layoffs_staging2 11

```
25 | 15:26:44 | select * from layoffs_staging2 LIMIT 0, 5000 | 2356 row(s) returned
```

```
-- looking at column 'industry'

select distinct industry
from layoffs_staging2
order by 1;
```

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

industry
NULL
Aerospace
Construction
Consumer
Crypto
Crypto Currency
CryptoCurrency
Data

layoffs_staging2 12

```
26 | 15:29:00 | select distinct industry from layoffs_staging2 order by 1 LIMIT 0, 5000 | 34 row(s) returned
```

```sql
-- updating 'Crypto Currency' and 'CryptoCurrency' with 'Crypto' in column 'industry'

UPDATE layoffs_staging2
SET industry = 'Crypto'
WHERE industry LIKE 'Crypto%';
```

✓	27	15:30:21	UPDATE layoffs_staging2 SET industry = 'Crypto' WHERE industry LIKE 'Crypto%'	3 row(s) affected Rows matched: 102 Changed: 3 Warnings: 0

```sql
-- Checking whether the data is updated or not.

select distinct industry
from layoffs_staging2
order by 1;
```

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

industry
NULL
Aerospace
Construction
Consumer
Crypto
Data
Education
Energy

layoffs_staging2 13

✓	28	15:31:02	select distinct industry from layoffs_staging2 order by 1 LIMIT 0, 5000	32 row(s) returned

```sql
-- looking at column 'country'

SELECT DISTINCT country
FROM layoffs_staging2
ORDER BY 1;
```

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

country
Switzerland
Thailand
Turkey
United Arab Emir...
United Kingdom
United States
United States.
Uruguay
Vietnam

layoffs_staging2 14

✓	29	15:33:32	SELECT DISTINCT country FROM layoffs_staging2 ORDER BY 1 LIMIT 0, 5000	60 row(s) returned

```sql
-- updating 'United States.' with 'United States'

UPDATE layoffs_staging2
SET country = TRIM(TRAILING '.' FROM country)
WHERE country LIKE 'United States%';
```

✓	30	15:35:20	UPDATE layoffs_staging2 SET country = TRIM(TRAILING '.' FROM country) WHE...	4 row(s) affected Rows matched: 1543 Changed: 4 Warnings: 0

```sql
-- Checking whether the data is updated or not.

SELECT DISTINCT country
FROM layoffs_staging2
ORDER BY 1;
```

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

country
Sweden
Switzerland
Thailand
Turkey
United Arab Emir...
United Kingdom
United States
Uruguay
Vietnam

layoffs_staging2 15

✓	31	15:36:02	SELECT DISTINCT country FROM layoffs_staging2 ORDER BY 1 LIMIT 0, 5000	59 row(s) returned

```sql
-- looking at column 'date'

SELECT `date`,
STR_TO_DATE(`date`, '%m/%d/%Y')
from layoffs_staging2;
```

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

date	STR_TO_DATE(`date`, '%m/%d/%Y')
12/16/2022	2022-12-16
7/25/2022	2022-07-25
11/17/2022	2022-11-17
1/27/2023	2023-01-27
7/13/2022	2022-07-13
1/10/2023	2023-01-10
8/4/2022	2022-08-04
4/2/2020	2020-04-02

Result 16

✓	32	15:37:21	SELECT `date`, STR_TO_DATE(`date`, '%m/%d/%Y') from layoffs_staging2 LIMIT 0...	2356 row(s) returned

```sql
-- updating 'date' column and changing its format from text to date.

UPDATE layoffs_staging2
SET `date` = STR_TO_DATE(`date`, '%m/%d/%Y')
```

✓	33	15:38:51	UPDATE layoffs_staging2 SET `date` = STR_TO_DATE(`date`, '%m/%d/%Y')	2355 row(s) affected Rows matched: 2356 Changed: 2355 Warnings: 0

```sql
ALTER TABLE layoffs_staging2
MODIFY COLUMN `date` DATE;
```

✓	34	15:40:43	ALTER TABLE layoffs_staging2 MODIFY COLUMN `date` DATE	2356 row(s) affected Records: 2356 Duplicates: 0 Warnings: 0

3. Handling Null values or Blank values

```sql
SELECT *
FROM layoffs_staging2
WHERE industry IS NULL
OR industry= '';
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Airbnb	SF Bay Area		30	NULL	2023-03-03	Post-IPO	United States	6400	1
Bally's Interactive	Providence	NULL	NULL	0.15	2023-01-18	Post-IPO	United States	946	1
Carvana	Phoenix		2500	0.12	2022-05-10	Post-IPO	United States	1600	1
Juul	SF Bay Area		400	0.3	2022-11-10	Unknown	United States	1500	1

layoffs_staging2 17 ×

✓	35	15:43:17	SELECT * FROM layoffs_staging2 WHERE industry IS NULL OR industry= '' LIMIT ...	4 row(s) returned

```sql
-- replacing blank values with 'NULL'

UPDATE layoffs_staging2
SET industry = NULL
WHERE industry= '';
```

✓	36	15:44:21	UPDATE layoffs_staging2 SET industry = NULL WHERE industry= ''	3 row(s) affected Rows matched: 3 Changed: 3 Warnings: 0

```sql
-- Try to figure out if any row contains the industry name for the same company in the other rows so that we could replace with that.

SELECT *
FROM layoffs_staging2
WHERE company ='Airbnb';
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Airbnb	SF Bay Area	NULL	30	NULL	2023-03-03	Post-IPO	United States	6400	1
Airbnb	SF Bay Area	Travel	1900	0.25	2020-05-05	Private Equity	United States	5400	1

layoffs_staging2 18 ×

✓	37	15:44:53	SELECT * FROM layoffs_staging2 WHERE company ='Airbnb' LIMIT 0, 5000	2 row(s) returned

above query output: there are other rows which contain the industry name for the same company that is 'Travel.'

```
SELECT *
FROM layoffs_staging2
WHERE company ='carvana';
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Carvana	Phoenix	NULL	2500	0.12	2022-05-10	Post-IPO	United States	1600	1
Carvana	Phoenix	Transportation	NULL	NULL	2023-01-13	Post-IPO	United States	1600	1
Carvana	Phoenix	Transportation	1500	0.08	2022-11-18	Post-IPO	United States	1600	1

layoffs_staging2 19

38	15:45:52	SELECT * FROM layoffs_staging2 WHERE company ='carvana' LIMIT 0, 5000	3 row(s) returned

above query output: there are other rows which contain the industry name for the same company that is 'Transportation.'

```
SELECT *
FROM layoffs_staging2
WHERE company ='juul';
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Juul	SF Bay Area	NULL	400	0.3	2022-11-10	Unknown	United States	1500	1
Juul	SF Bay Area	Consumer	900	0.3	2020-05-05	Unknown	United States	1500	1

layoffs_staging2 20

39	15:46:37	SELECT * FROM layoffs_staging2 WHERE company ='juul' LIMIT 0, 5000	2 row(s) returned

above query output: there are other rows which contain the industry name for the same company that is 'Consumer.'

```
SELECT *
FROM layoffs_staging2
WHERE company LIKE 'Bally%'
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Bally's Interactive	Providence	NULL	NULL	0.15	2023-01-18	Post-IPO	United States	946	1

layoffs_staging2 21

40	15:48:24	SELECT * FROM layoffs_staging2 WHERE company LIKE 'Bally%' LIMIT 0, 5000	1 row(s) returned

above query output: No other row contains the industry name for the same company.

NOTE:

Now, here we are doing self-join so that we could get to know from the table where the industry's name is given for the same company and where the industry name is NULL so that we could replace them with the available values.

```
SELECT t1.industry,t2.industry
FROM layoffs_staging2 t1
JOIN layoffs_staging2 t2
ON t1.company = t2.company
WHERE t1.industry IS NULL
AND t2.industry IS NOT NULL
```

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

industry	industry
NULL	Travel
NULL	Transportation
NULL	Transportation
NULL	Consumer

Result 22

41	15:49:15	SELECT t1.industry,t2.industry FROM layoffs_staging2 t1 JOIN layoffs_staging2 t2 ...	4 row(s) returned

```sql
-- Replacing the industry column's NULL values with the available values we got from the last query.

UPDATE layoffs_staging2 t1
JOIN layoffs_staging2 t2
ON t1.company = t2.company
SET t1.industry = t2.industry
WHERE t1.industry IS NULL
AND t2.industry IS NOT NULL
```

✔	42	15:50:09	UPDATE layoffs_staging2 t1 JOIN layoffs_staging2 t2 ON t1.company = t2.compan...	3 row(s) affected Rows matched: 3 Changed: 3 Warnings: 0

```sql
-- Checking whether the data is updated or not.

SELECT *
FROM layoffs_staging2
WHERE industry IS NULL
OR industry= '';
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Bally's Interactive	Providence	NULL	NULL	0.15	2023-01-18	Post-IPO	United States	946	1

layoffs_staging2 23 ×

✔	43	15:51:55	SELECT * FROM layoffs_staging2 WHERE industry IS NULL OR industry= '' LIMIT ...	1 row(s) returned

All the values are updated except ‘Bally’s Interactive’ company as no other row contains the industry name for the ‘Bally’s Interactive’ company.

```sql
-- looking for the rows where 'total_laid_off' and 'percentage_laid_off' column values are NULL

SELECT *
FROM layoffs_staging2
WHERE total_laid_off IS NULL
AND percentage_laid_off IS NULL
```

Result Grid | Filter Rows: | Export: | 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	2022-12-16	Post-IPO	Canada	NULL	1
100 Thieves	Los Angeles	Retail	NULL	NULL	2023-01-10	Series C	United States	120	1
Accolade	Seattle	Healthcare	NULL	NULL	2023-03-03	Post-IPO	United States	458	1
Ada	Toronto	Support	NULL	NULL	2023-02-01	Series C	Canada	190	1
Adara	SF Bay Area	Travel	NULL	NULL	2020-03-31	Series C	United States	67	1
Addi	Bogota	Finance	NULL	NULL	2022-06-14	Series C	Colombia	376	1
AirMap	Los Angeles	Aerospace	NULL	NULL	2020-04-30	Unknown	United States	75	1

layoffs_staging2 24

44	15:55:30	SELECT * FROM layoffs_staging2 WHERE total_laid_off IS NULL AND percentage...	361 row(s) returned

```sql
-- deleting all the rows which contains NULL value in 'total_laid_off' and 'percentage_laid_off' column.

DELETE
FROM layoffs_staging2
WHERE total_laid_off IS NULL
AND percentage_laid_off IS NULL
```

45	15:56:56	DELETE FROM layoffs_staging2 WHERE total_laid_off IS NULL AND percentage_...	361 row(s) affected

```sql
-- Checking whether the data is deleted or not.

SELECT *
FROM layoffs_staging2
WHERE total_laid_off IS NULL
AND percentage_laid_off IS NULL;
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num

layoffs_staging2 25

46	15:57:17	SELECT * FROM layoffs_staging2 WHERE total_laid_off IS NULL AND percentage...	0 row(s) returned

4. Remove any columns unnecessary

```
ALTER TABLE layoffs_staging2
DROP COLUMN row_num;
```

✓	47	15:59:40	ALTER TABLE layoffs_staging2 DROP COLUMN row_num	0 row(s) affected Records: 0 Duplicates: 0 Warnings: 0

-- ---------- Final_cleaned_data ---------- --

```
SELECT *
FROM layoffs_staging2;
```

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

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions
Included Health	SF Bay Area	Healthcare	NULL	0.06	2022-07-25	Series E	United States	272
&Open	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	NULL	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
1stdibs	New York City	Retail	70	0.17	2020-04-02	Series D	United States	253
2TM	Sao Paulo	Crypto	90	0.12	2022-06-01	Unknown	Brazil	250
2TM	Sao Paulo	Crypto	100	0.15	2022-09-01	Unknown	Brazil	250
2U	Washington D.C.	Education	NULL	0.2	2022-07-28	Post-IPO	United States	426
54gene	Washington D.C.	Healthcare	95	0.3	2022-08-29	Series B	United States	44
[illegible]	[illegible]	[illegible]	NULL	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

layoffs_staging2 26 ×

✓	48	16:00:04	SELECT * FROM layoffs_staging2 LIMIT 0, 5000	1995 row(s) returned