SQL CLEANING OF DATA

World Layoffs (2020-2023)

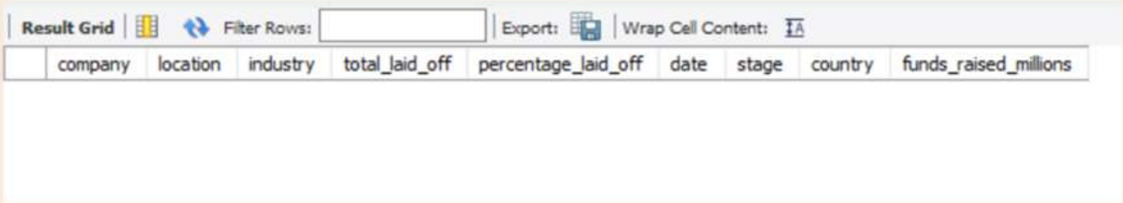
By: Zaeem Farooq

STEPS OF DATA CLEANING

```
-- 0: Always Create a Backup First
-- 1: Remove Duplicates
-- 2: Standardize the Data
-- 3: Null Values or Blank Values
-- 4: Remove Any columns which are completely irrelevant
```

Always Create a Backup First

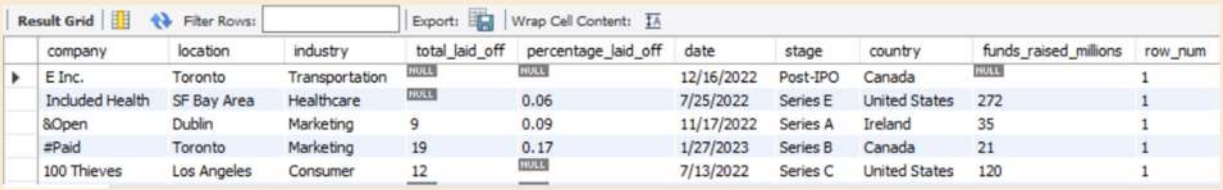
```
-- Copying the Data From Raw table into the staging table beacuse
--  it is always necessary to have a raw data

CREATE TABLE layoffs_staging
LIKE layoffs;
INSERT layoffs_staging
SELECT * FROM layoffs;


SELECT * FROM layoffs_staging;
```



company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions

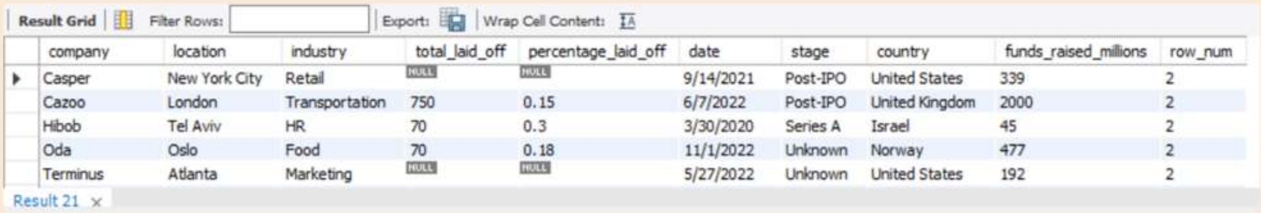
Remove Duplicates:

```
-- Finding that is there any duplicate exists
SELECT *,
ROW_NUMBER() OVER(
PARTITION BY company,industry,total_laid_off,
percentage_laid_off,`date`) AS row_num
FROM layoffs_staging;
```



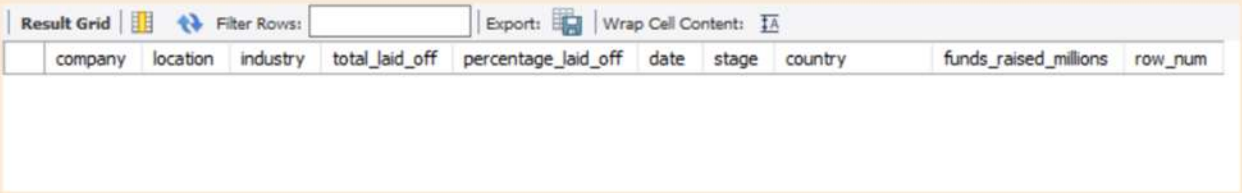
company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
E Inc.	Toronto	Transportation	NULL	NULL	12/16/2022	Post-IPO	Canada	NULL	1
Included Health	SF Bay Area	Healthcare	NULL	0.06	7/25/2022	Series E	United States	272	1
&Open	Dublin	Marketing	9	0.09	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

```
-- Making CTE of above statement to check duplicates
WITH duplicate_CTE AS (SELECT *,
ROW_NUMBER() OVER(
PARTITION BY company,industry,total_laid_off,
percentage_laid_off,`date`) AS row_num
FROM layoffs_staging)
SELECT * FROM duplicate_CTE WHERE row_num>1;
```



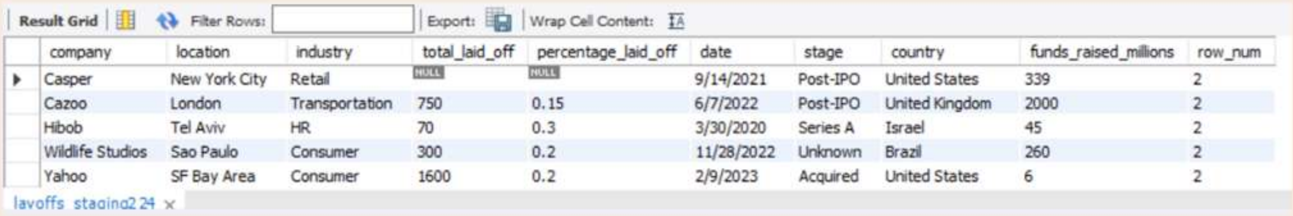
company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Casper	New York City	Retail	NULL	NULL	9/14/2021	Post-IPO	United States	339	2
Cazoo	London	Transportation	750	0.15	6/7/2022	Post-IPO	United Kingdom	2000	2
Hibob	Tel Aviv	HR	70	0.3	3/30/2020	Series A	Israel	45	2
Oda	Oslo	Food	70	0.18	11/1/2022	Unknown	Norway	477	2
Terminus	Atlanta	Marketing	NULL	NULL	5/27/2022	Unknown	United States	192	2

```
-- NOW We are going to remove duplicates
-- WE Cannot Apply DELETE Command ON CTE Directly
-- FOR That we have to create a statement of the table and than add a
-- new row_num column in that after that we can delete

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` text
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci;

SELECT *
FROM layoffs_staging2;
```



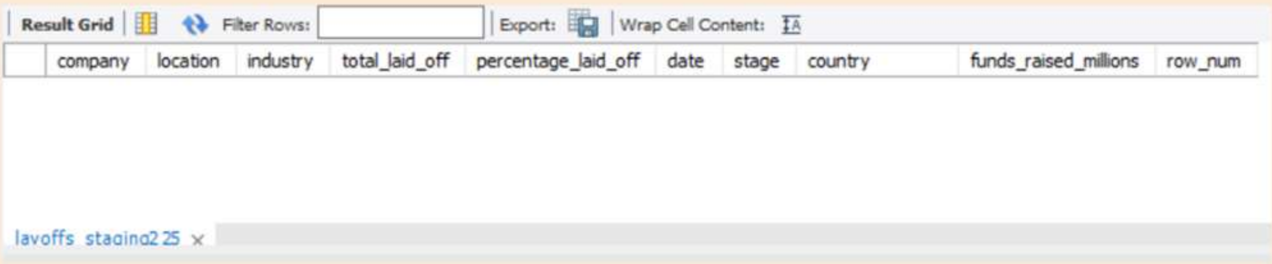
company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num

```
-- INSERTING THE Data into Layoff_staging 2:
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;

SELECT *
FROM layoffs_staging2
WHERE row_num>1;
```



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

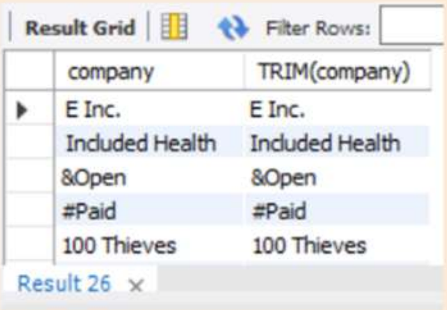
```
-- NOW rmeoving the Duplicates
DELETE
FROM layoffs_staging2
WHERE row_num>1;

SELECT *
FROM layoffs_staging2
WHERE row_num>1;
-- Now the Duplicates have been removed
```



company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num

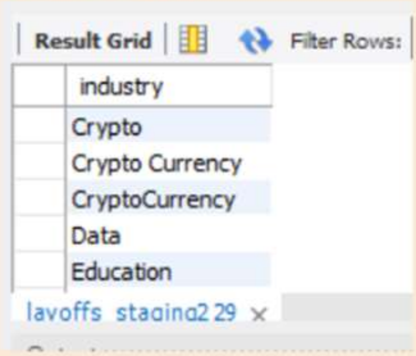
STANDARDIZING DATA

```
SELECT company,TRIM(company)
FROM layoffs_staging2;

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



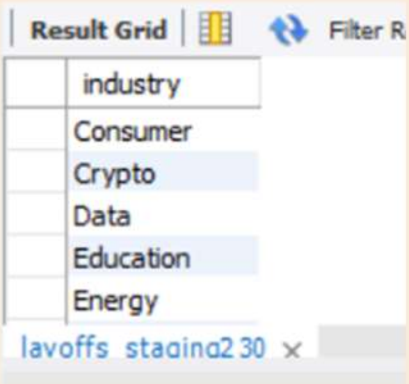
company	TRIM(company)
E Inc.	E Inc.
Included Health	Included Health
&Open	&Open
#Paid	#Paid
100 Thieves	100 Thieves

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



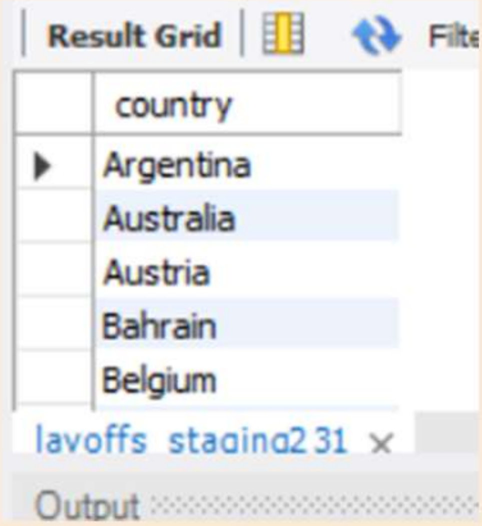
industry
Crypto
Crypto Currency
CryptoCurrency
Data
Education

```
SELECT *
FROM layoffs_staging2
WHERE industry LIKE 'Crypto%';

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



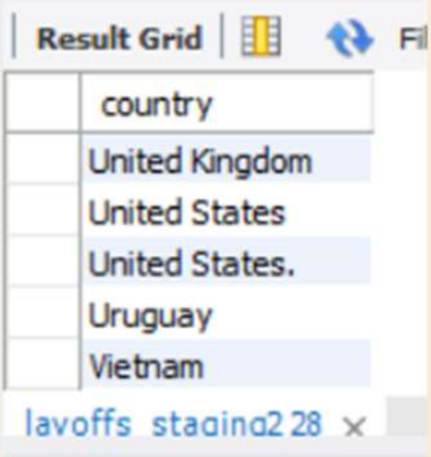
industry
Consumer
Crypto
Data
Education
Energy

```
-- CHECKING EVERY COLUMN BY FOLLOWING QUERY
SELECT DISTINCT location
FROM layoffs_staging2
ORDER BY 1;

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



country
Argentina
Australia
Austria
Bahrain
Belgium

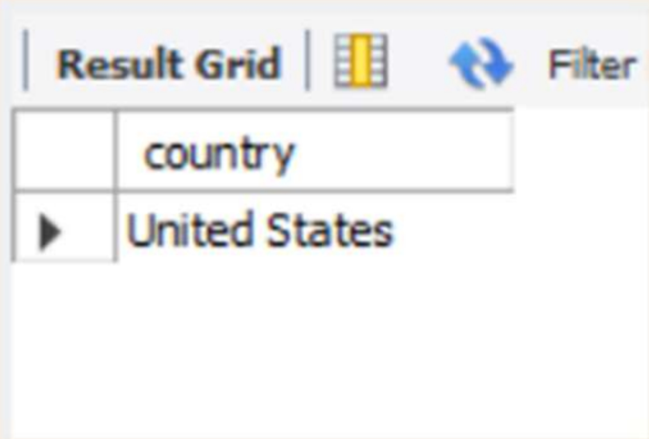
layoffs_staging2 31



country
United Kingdom
United States
United States.
Uruguay
Vietnam

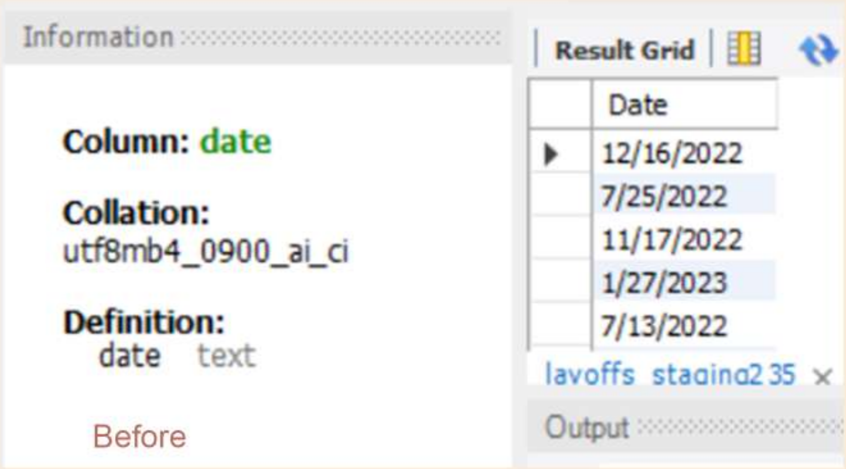
layoffs_staging2 28

```
-- IN THE ABOVE QUERY WE HAVE FOUND THAT THERE ARE
-- TWO SAME UNITED STATES CELLS
-- NOW we are removing the duplicate United State in the Country Column

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

SELECT country
FROM layoffs_staging2
WHERE country LIKE 'United States%'
GROUP BY country;
```



country
United States

```sql
-- NOW we have to convert our date column from text to real date column

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

-- NOW are updating this to our table

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

SELECT `date`
FROM layoffs_staging2;

-- This is not in the date format till
ALTER TABLE layoffs_staging2
MODIFY COLUMN `date` DATE;

SELECT `Date` FROM layoffs_staging2;

-- NOW WE have converted the data into the standardized format
```



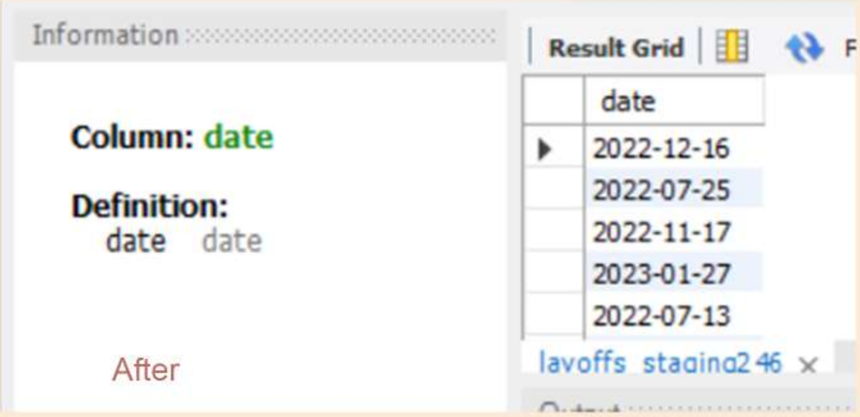
Column: date

Collation:
utf8mb4_0900_ai_ci

Definition:
date text

Date
12/16/2022
7/25/2022
11/17/2022
1/27/2023
7/13/2022

Before



Column: date

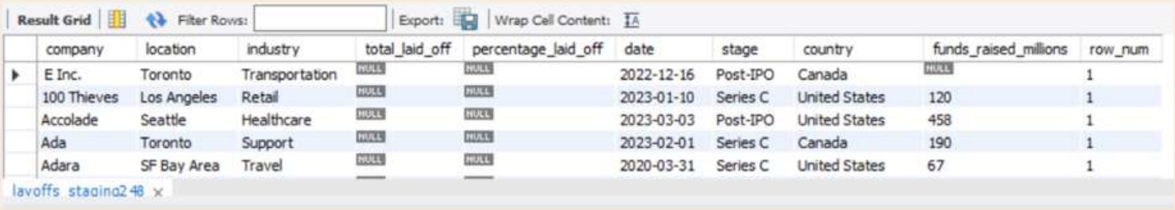
Definition:
date date

date
2022-12-16
2022-07-25
2022-11-17
2023-01-27
2022-07-13

After

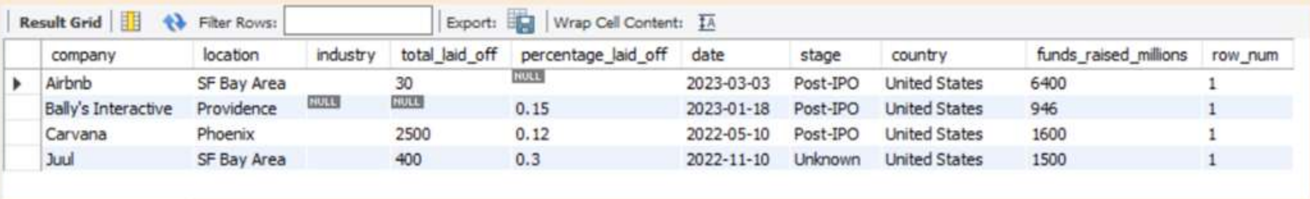
REMOVING NULL VALUES

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



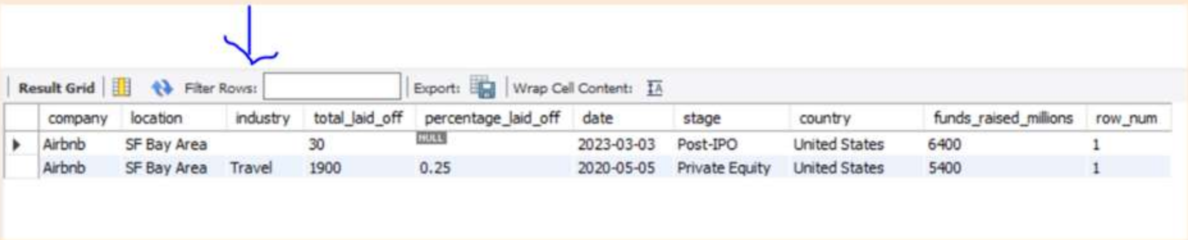
company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
E Inc.	Toronto	Transportation	NULL	NULL	2022-12-16	Post-IPO	Canada	NULL	1
100 Thieves	Los Angeles	Retail	NULL	NULL	2023-01-10	Series C	United States	120	1
Accolade	Seattle	Healthcare	NULL	NULL	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

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



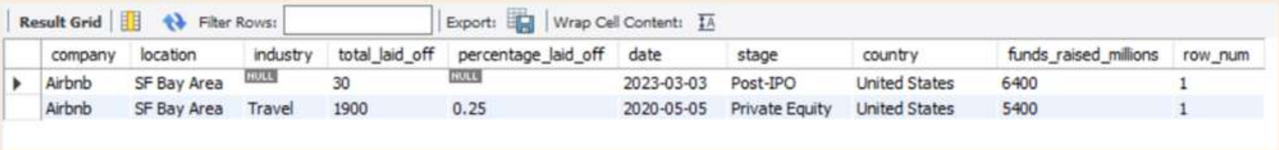
company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Airbnb	SF Bay Area		30	NULL	2023-03-03	Post-IPO	United States	6400	1
Bally's Interactive	Providence	NULL	NULL	0.15	2023-01-18	Post-IPO	United States	946	1
Carvana	Phoenix		2500	0.12	2022-05-10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

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



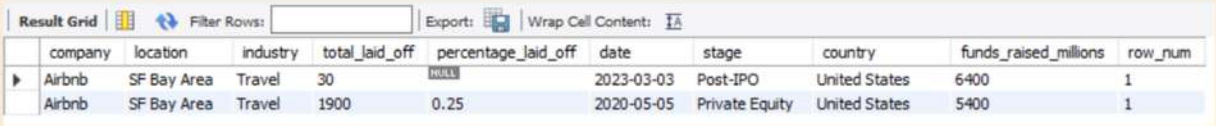
company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Airbnb	SF Bay Area		30	NULL	2023-03-03	Post-IPO	United States	6400	1
Airbnb	SF Bay Area	Travel	1900	0.25	2020-05-05	Private Equity	United States	5400	1

```
-- Converting INDUSTRY from empty cell to null value
UPDATE layoffs_staging2
SET industry=NULL
WHERE industry = '';
SELECT *
FROM layoffs_staging2
WHERE company='Airbnb';
```



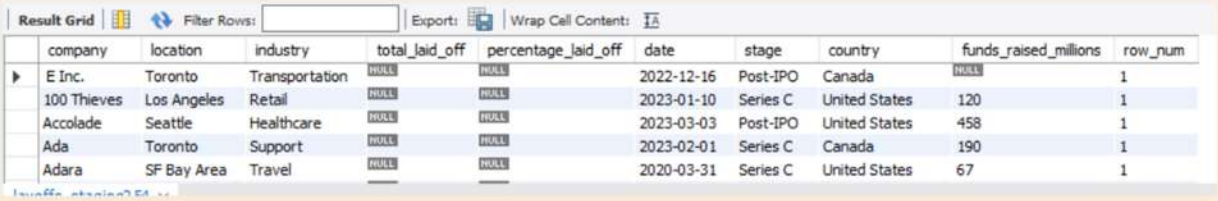
company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
Airbnb	SF Bay Area	NULL	30	NULL	2023-03-03	Post-IPO	United States	6400	1
Airbnb	SF Bay Area	Travel	1900	0.25	2020-05-05	Private Equity	United States	5400	1

```
-- Now first checking that it woking and then
-- changing that null value to Travel with the Help of joins

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

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;

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



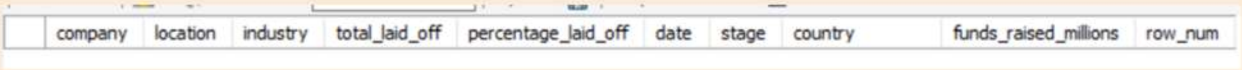
	company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num
▶	Airbnb	SF Bay Area	Travel	30	NULL	2023-03-03	Post-IPO	United States	6400	1
	Airbnb	SF Bay Area	Travel	1900	0.25	2020-05-05	Private Equity	United States	5400	1

```
-- REMOVING THE DATA WHERE TWO COLUMNS ARE COMPLETELY NULL

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



	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

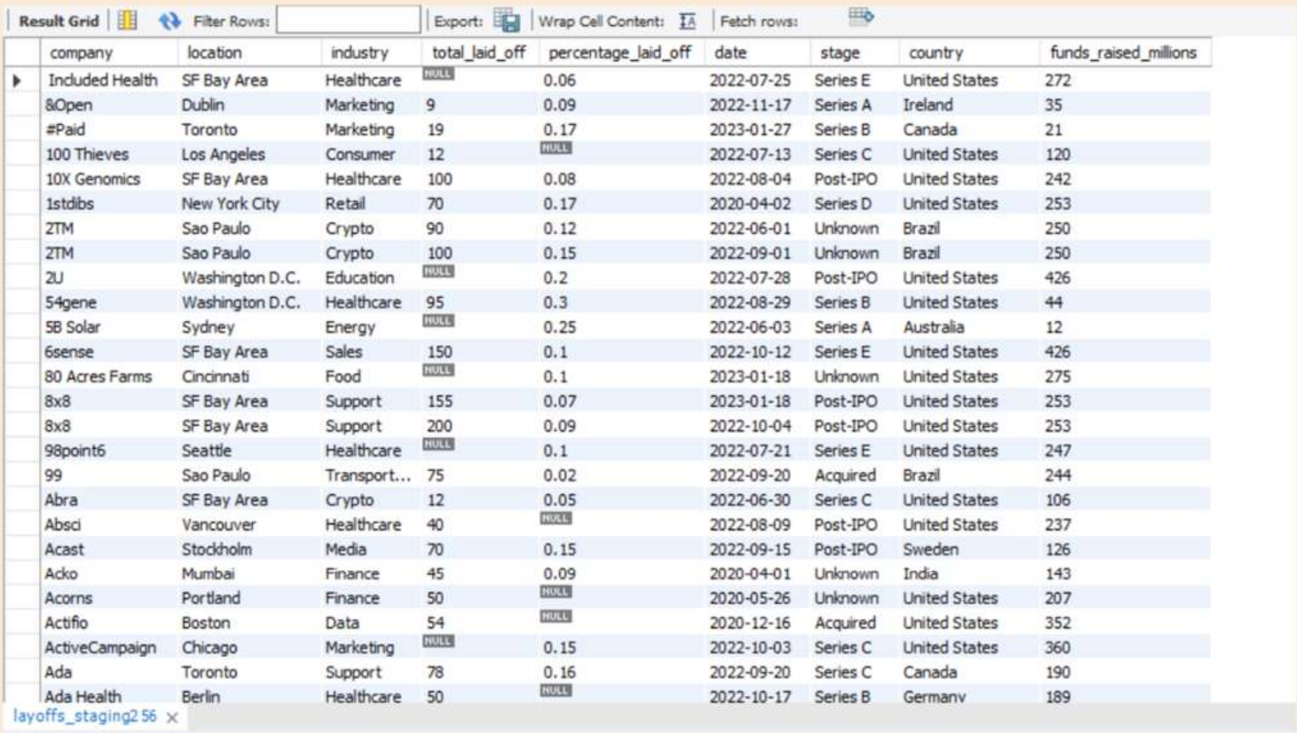
```
-- DELETING THE DATA WHERE TWO COLUMNS ARE COMPLETELY NULL
DELETE FROM layoffs_staging2
WHERE total_laid_off IS NULL
AND percentage_laid_off IS NULL;

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



	company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions	row_num

REMOVING UNNECESSARY COLUMNS

```
-- NOW WE R GOING TO DROP COLUMN

ALTER TABLE layoffs_staging2
DROP COLUMN row_num;

SELECT *
FROM layoffs_staging2;
```



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
5B Solar	Sydney	Energy	NULL	0.25	2022-06-03	Series A	Australia	12
6sense	SF Bay Area	Sales	150	0.1	2022-10-12	Series E	United States	426
80 Acres Farms	Cincinnati	Food	NULL	0.1	2023-01-18	Unknown	United States	275
8x8	SF Bay Area	Support	155	0.07	2023-01-18	Post-IPO	United States	253
8x8	SF Bay Area	Support	200	0.09	2022-10-04	Post-IPO	United States	253
98point6	Seattle	Healthcare	NULL	0.1	2022-07-21	Series E	United States	247
99	Sao Paulo	Transport...	75	0.02	2022-09-20	Acquired	Brazil	244
Abra	SF Bay Area	Crypto	12	0.05	2022-06-30	Series C	United States	106
Absci	Vancouver	Healthcare	40	NULL	2022-08-09	Post-IPO	United States	237
Acast	Stockholm	Media	70	0.15	2022-09-15	Post-IPO	Sweden	126
Acko	Mumbai	Finance	45	0.09	2020-04-01	Unknown	India	143
Acorns	Portland	Finance	50	NULL	2020-05-26	Unknown	United States	207
Actifio	Boston	Data	54	NULL	2020-12-16	Acquired	United States	352
ActiveCampaign	Chicago	Marketing	NULL	0.15	2022-10-03	Series C	United States	360
Ada	Toronto	Support	78	0.16	2022-09-20	Series C	Canada	190
Ada Health	Berlin	Healthcare	50	NULL	2022-10-17	Series B	Germany	189