About data

Post-pandemic, following the impact of COVID-19, tech companies worldwide faced economic slowdowns. This dataset was sourced from reports on tech layoffs published by platforms such as Bloomberg, the San Francisco Business Times, TechCrunch, and The New York Times.

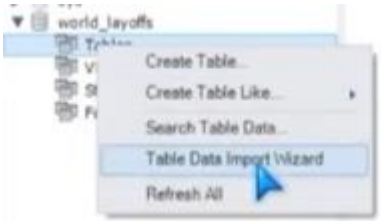
I have completed the project in two parts

Part 1. Data cleaning using SQL
Part 2. Data exploration using SQL

Steps in data cleaning project

1. Remove Duplicates
2. Standardize the Data
3. Review Null Values /Blank values
4. Remove any column which we think is irrelevant

1. Create a new data base **World_layoffs**
2. Import the excel file **layoffs.csv** which is already saved in our system into Tables



3. Now create a staging table **layoffs_staging** and input all data from **layoffs table**. We will work on this table and clean the data. We will have our original table in raw format in case something happens we will have our original data.

```
create table layoffs_staging
like layoffs;

insert layoffs_staging
select *
from layoffs;

select * from layoffs_staging;
```

4. **Remove Duplicates**

a. This table doesn't have an extra column with unique id so for removing duplicates so we will use ROW NUMBER function and partition by every single column and then we will see if there are any duplicates.

b. It will create a new column row_num. If it shows 2 or more than 2 that means it's a duplicate

```
27 •  select *,
28    row_number() over(
29    partition by company,location,industry,
30    total_laid_off,percentage_laid_off,`date`,
31    stage,country,funds_raised_millions)
32    as row_num
33    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	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
2TM	Sao Paulo	Crypto	90	0.12	6/1/2022	Unknown	Brazil	250	1
2TM	Sao Paulo	Crypto	100	0.15	9/1/2022	Unknown	Brazil	250	1
2U	Washington ...	Education	NULL	0.2	7/28/2022	Post-IPO	United States	426	1

c. Using a CTE (duplicate_cte) to find results having row_num >= 2

```
42 •  With duplicate_cte as
43    (
44    select *,
45    row_number() over(
46    partition by company,location,industry,total_laid_off,percentage_laid_off,`date`,stage,country,funds_raised_millions) as row_num
47    from layoffs_staging
48    )
49
50    select *
51    from duplicate_cte
52    where row_num >1
53    ;
```

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

Let's take a look at company "casper"

```
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

Row 1 and 3 are same. So, if we remove one of these rows, we will remove duplicate rows

We cannot execute delete statement in CTE so we will create a new table layoffs_staging 2 and also add extra column row_num with INT as data type

```
create table layoffs_staging2
like layoffs_staging;
alter table layoffs_staging2
add row_num INT;
select * from layoffs_staging2;
-- this is an emppty table with same column as layoff_staging2  + extra cloumn as row_num
```

Now we will insert the data into layoffs_staging 2 as below

```
SELECT *
FROM layoffs_staging2;

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;
```

```
-- filter results where row_num is >1

select *
from layoffs_staging2
where row_num >1;

-- delete all results with row_num>1

delete
from layoffs_staging2
where row_num >1;
```

In the end we will remove this row_num column. We won't need that anymore. It's best not to add extra space in memory and storage

Standardizing data

1. Used TRIM function to remove extra space from company names
2. Used DISTINCT functions to view industry list where Industry names are not repeated
3. Used ORDER BY to arrange Industry name in alphabetical order
4. Crypto and Crypto currency were listed in separate rows so we merged that name into one name “crypto” by using

```
update layoffs_staging2
SET INDUSTRY = 'Crypto'
WHERE industry = '%crypto%';
```

5. Country United States is having (.) in the end which is not good so we are trimming that

```
update layoffs_staging2
set country = TRIM(TRAILING'.' FROM COUNTRY)
Where country LIKE 'United States%'
;
```

6.Column Date was in text datatype so we will change it to DATE datatype

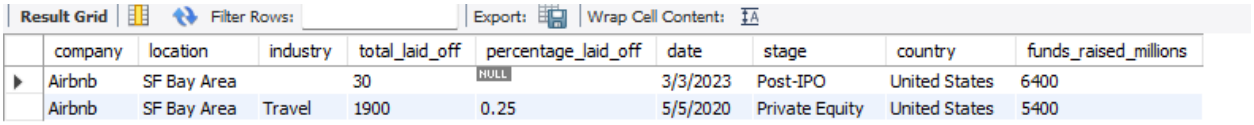
Null / blank values

```
delete
from layoffs_staging2
where total_laid_off is Null  #Showing only total_laid_off null
and percentage_laid_off is NULL;  #Showing both total and percentage null
```

Example where we are deleting rows where total laid off and percentage laid off is null as we don’t need that data

Remove null values from industry

```
select *
from layoffs_staging2
where company = 'airbnb';
```

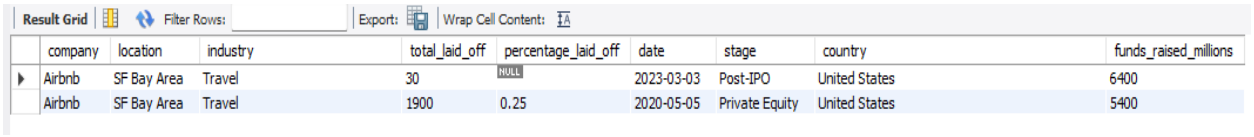


Result Grid | Filter Rows: | Export: | Wrap Cell Content:

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions
Airbnb	SF Bay Area		30	NULL	3/3/2023	Post-IPO	United States	6400
Airbnb	SF Bay Area	Travel	1900	0.25	5/5/2020	Private Equity	United States	5400

Here company name is “Airbnb” in both rows. In row 1, industry name is blank. All Airbnb must come under same “Travel Industry” so firstly, we identified which companies are having this issue and secondly, we removed blank/null values from industry name.

Used JOIN function within the same table (where company is same) and set same industry name replacing null/ blank values. For eg in this case



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

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

Removing Column

```
alter table layoffs_staging2
drop column row_num;
```

This will remove row_num column.

The next part is data exploration which is explained in SQL script