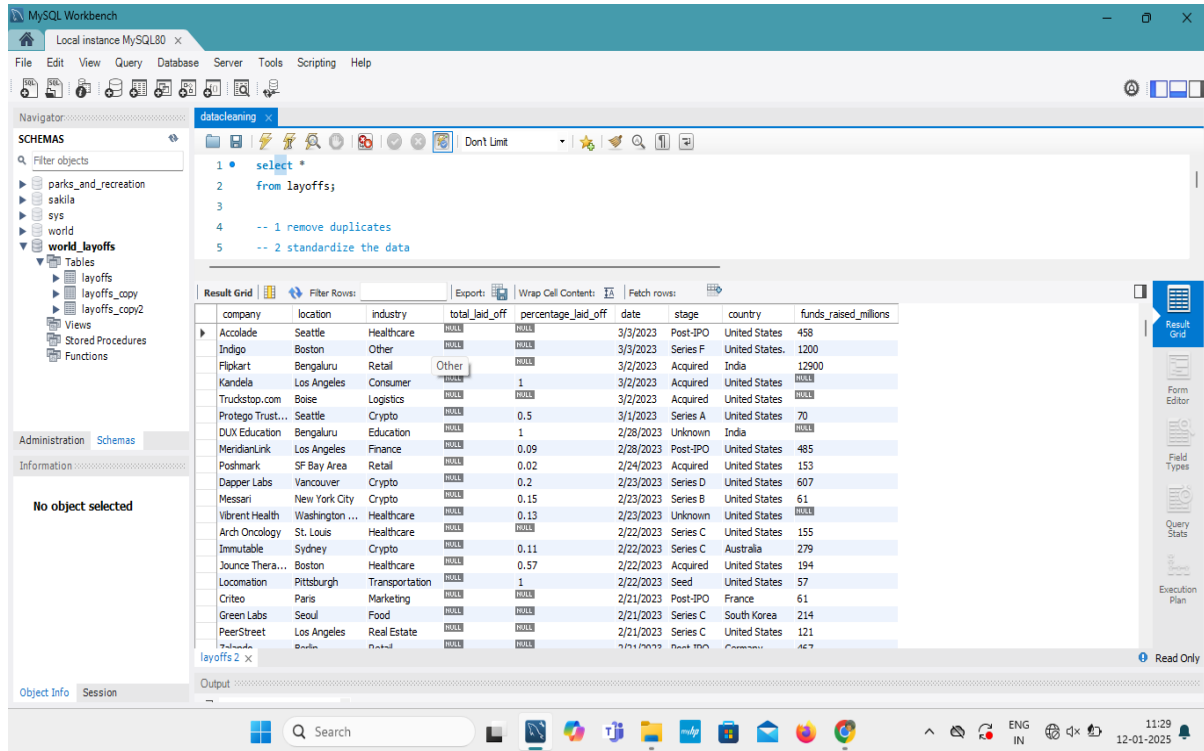


Data cleaning of world_layoffs

select *

from layoffs;



The screenshot shows the MySQL Workbench interface. The query editor contains the following SQL code:

```
1 select *
2 from layoffs;
3
4 -- 1 remove duplicates
5 -- 2 standardize the data
```

The result grid displays a table with the following columns: company, location, industry, total_laid_off, percentage_laid_off, date, stage, country, and funds_raised_millions. The data is sorted by date in descending order.

company	location	industry	total_laid_off	percentage_laid_off	date	stage	country	funds_raised_millions
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
Fliplart	Bengaluru	Retail	Other	NULL	3/2/2023	Acquired	India	12900
Kandela	Los Angeles	Consumer	1	1	3/2/2023	Acquired	United States	NULL
Truckstop.com	Boise	Logistics	NULL	NULL	3/2/2023	Acquired	United States	NULL
Protego Trust...	Seattle	Crypto	0.5	0.5	3/1/2023	Series A	United States	70
DUX Education	Bengaluru	Education	1	1	2/28/2023	Unknown	India	NULL
MeridianLink	Los Angeles	Finance	0.09	0.09	2/28/2023	Post-IPO	United States	485
Poshmark	SF Bay Area	Retail	0.02	0.02	2/24/2023	Acquired	United States	153
Dapper Labs	Vancouver	Crypto	0.2	0.2	2/23/2023	Series D	United States	607
Messari	New York City	Crypto	0.15	0.15	2/23/2023	Series B	United States	61
Vibrent Health	Washington ...	Healthcare	0.13	0.13	2/23/2023	Unknown	United States	NULL
Arch Oncology	St. Louis	Healthcare	NULL	NULL	2/22/2023	Series C	United States	155
Immutable	Sydney	Crypto	0.11	0.11	2/22/2023	Series C	Australia	279
Jounce Thera...	Boston	Healthcare	0.57	0.57	2/22/2023	Acquired	United States	194
Locomotion	Pittsburgh	Transportation	1	1	2/22/2023	Seed	United States	57
Criteo	Paris	Marketing	NULL	NULL	2/21/2023	Post-IPO	France	61
Green Labs	Seoul	Food	NULL	NULL	2/21/2023	Series C	South Korea	214
PeerStreet	Los Angeles	Real Estate	NULL	NULL	2/21/2023	Series C	United States	121
74mads	Boise	Real Estate	NULL	NULL	2/21/2023	Post-IPO	United States	467

-- 1 remove duplicates

-- 2 standardize the data

-- 3 null values or empty values

-- 4 remove any columns or rows

create table layoffs_copy

like layoffs;

select *

from layoffs_copy;

insert layoffs_copy

```
select *  
from layoffs;
```

```
select *,  
row_number() over(  
partition by company,industry,total_laid_off,percentage_laid_off,`date`)as row_num  
from layoffs_copy;
```

```
with duplicatte_cte as  
(  
select *,  
row_number() over(  
partition by  
company,industry,total_laid_off,percentage_laid_off,`date`,stage,country,funds_raised_millions)as  
row_num  
from layoffs_copy  
)
```

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

```
CREATE TABLE `layoffs_copy2` (  
  `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;
```

```
select*  
from layoffs_copy2;
```

```
insert into layoffs_copy2  
select *,  
row_number() over(  
partition by  
company,industry,total_laid_off,percentage_laid_off,`date`,stage,country,funds_raised_millions)as  
row_num  
from layoffs_copy;
```

```
DELETE  
from layoffs_copy2  
where row_num > 1;
```

```
select*  
from layoffs_copy2;
```

```
-- standardizing data
```

```
select company,trim(company)  
from layoffs_copy2;
```

```
update layoffs_copy2  
set company = trim(company);
```

```
select industry
from layoffs_copy2
where industry like 'crypto%';
```

```
update layoffs_copy2
set industry = 'crypto'
where industry like 'crypto%';
```

```
select distinct country, trim(trailing '.' from country)
from layoffs_copy2
order by 1;
```

```
update layoffs_copy2
set country = trim(trailing '.' from country)
where country like 'united states%';
```

```
select `date`,
str_to_date(`date`, '%m/%d/%Y')
from layoffs_copy2;
```

```
update layoffs_copy2
set `date` = str_to_date(`date`, '%m/%d/%Y');
```

```
alter table layoffs_copy2
modify column `date` date;
```

```
select*
from layoffs_copy2
where total_laid_off is null
and percentage_laid_off is null;
```

```
update layoffs_copy2
```

```
set industry = null
```

```
where industry = ' ';
```

```
select *
```

```
from layoffs_copy2 as t1
```

```
join layoffs_copy2 as t2
```

```
on t1.company = t2.company
```

```
where (t1.industry is null or t1.industry = ' ')
```

```
and t2.industry is not null;
```

```
UPDATE layoffs_copy2 l1
```

```
JOIN layoffs_copy2 l2
```

```
ON l1.company = l2.company
```

```
SET l1.industry = l2.industry
```

```
WHERE l1.industry IS NULL
```

```
AND l2.industry IS NOT NULL;
```

```
select*
```

```
from layoffs_copy2
```

```
where total_laid_off is null
```

```
and percentage_laid_off is null;
```

```
delete
```

```
from layoffs_copy2
```

```
where total_laid_off is null
```

```
and percentage_laid_off is null;
```

```
alter table layoffs_copy2
```

```
drop column row_num;
```

```
select*
```

```
from layoffs_copy2;
```

The screenshot shows the MySQL Workbench interface. The SQL editor contains the following query:

```
139 select*
140 from layoffs_copy2;
141
142 -- end of cleaning
```

The Results 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	100	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	0.08	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
21M	Sao Paulo	crypto	90	0.12	2022-06-01	Unknown	Brazil	250
21M	Sao Paulo	crypto	100	0.15	2022-09-01	Unknown	Brazil	250
2U	Washington D.C.	Education	100	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
58 Solar	Sydney	Energy	100	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	100	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	100	0.1	2022-07-21	Series E	United States	247
99	Sao Paulo	Transportation	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	0.08	2022-08-09	Post-IPO	United States	237
Acadabra	London	Marketing	75	0.15	2022-09-15	Post-IPO	Canada	156

```
-- end of cleaning
```