DATA CLEANING -SQL

PRESENTED BY:

PARUL SAHU

INTRODUCTION

This project involves data cleaning for a company layoffs dataset using SQL to ensure data consistency and accuracy. Key steps include removing duplicates, standarizing data, filling in missing industry data, and removing columns or rows. The cleaned dataset will be ready for reliable analysis of layoff trends across industries.

CREATE TABLE

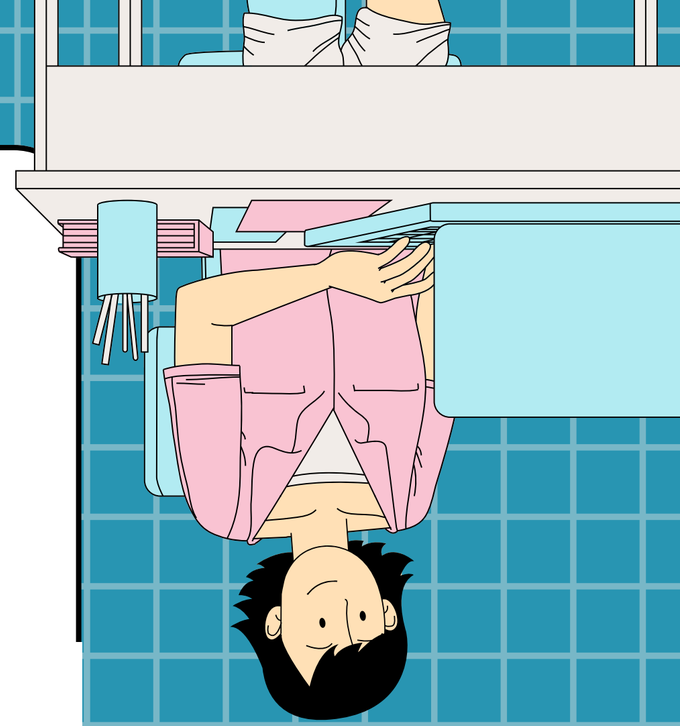
```
CREATE TABLE `layoffs_staging2` (
  `company` text,
  `location` text,
  `industry` text,
  `total_laid_off` int DEFAULT NULL,
  `percentage_laid_off` text,
  `date` text,
  `stage` text,
  `country` text,
  `funds_raised_millions` int DEFAULT NULL,
  `row_num` int
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4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

INSERT DATA

```
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;
```

REMOVING DUPLICATES

```
delete
from layoffs_staging2
where row_num>1;
```



STANDARIZING DATA

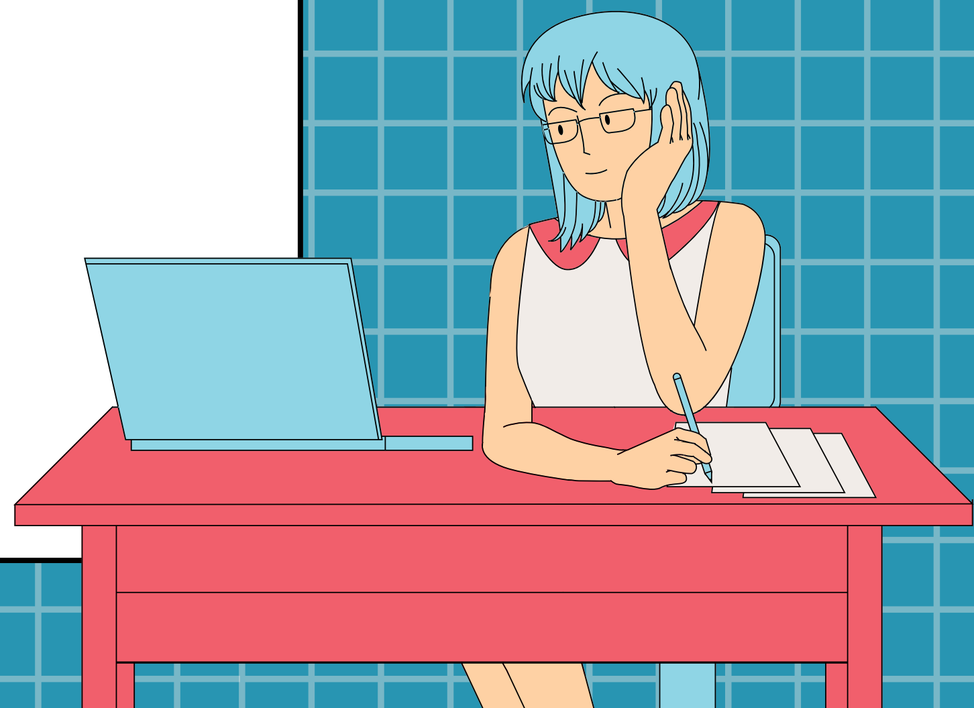
```sql
UPDATE layoffs_staging2
set company = trim(company);


update layoffs_staging2
set industry = 'crypto'
where industry like 'crypto%';
```

STANDARIZING DATA

```sql
update layoffs_staging2
set country = trim(trailing '.' from country)
where country like 'united states%';

update layoffs_staging2
set `date` = str_to_date(`date`, '%m/%d/%Y');

alter table layoffs_staging2
modify column `date` date;
```

REMOVING NULL OR BLANK VALUES

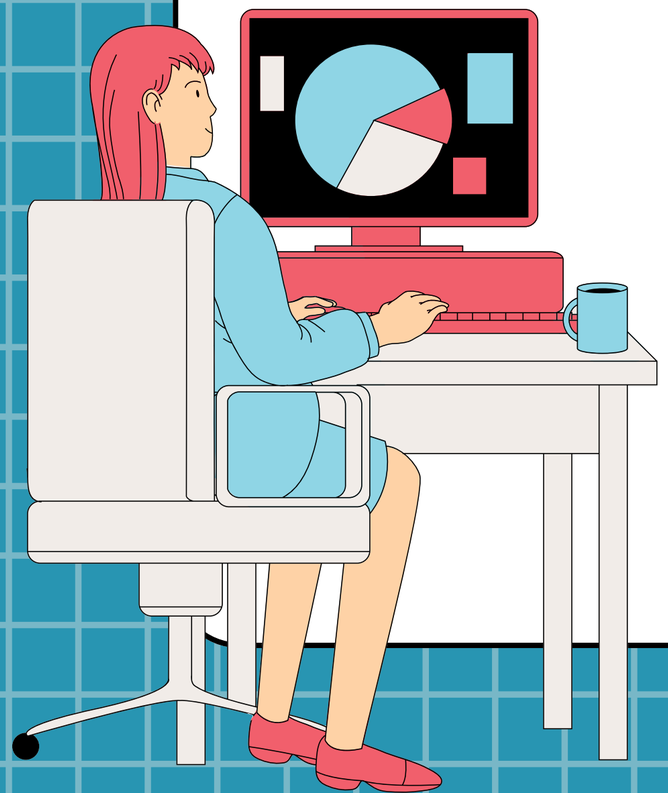
```sql
update layoffs_staging2
set industry = 'null'
where industry = ' ';

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;
```



REMOVING NULL OR BLANK VALUES

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

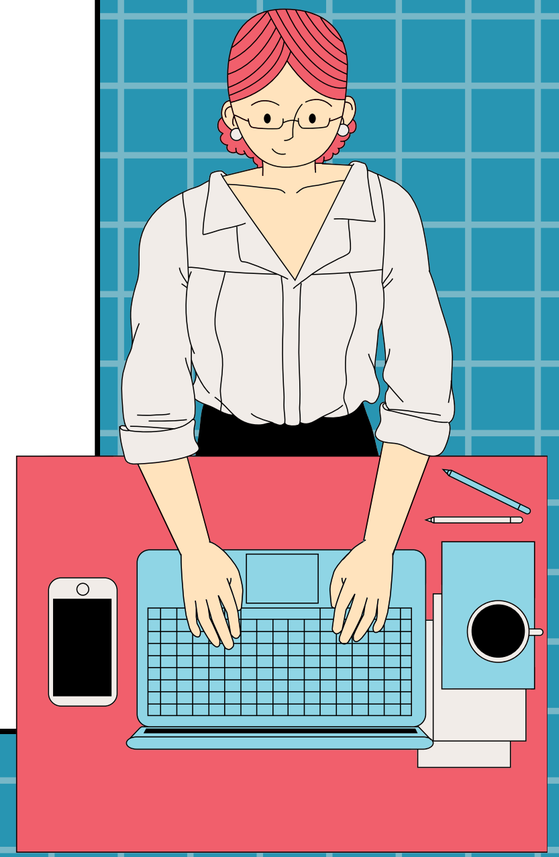
REMOVING ANY COLUMNS OR ROWS

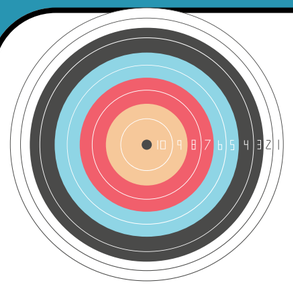


```
alter table layoffs_staging2
drop column row_num;
```

CONCLUSION

In conclusion, this data cleaning SQL project effectively prepared the company layoffs dataset for reliable analysis by addressing missing values, standardizing data formats, and removing duplicates. These steps ensured data consistency and accuracy, enabling meaningful insights into layoff trends across industries. The cleaned dataset is now ready for further analysis to support data-driven decision-making.





THANK YOU