DATA CLEANING SQL

Dataset - Layoff Analysis

In this project, I focused on cleaning and preparing data using SQL to ensure accuracy and consistency.

1) **Created a Staging Table:** By creating staging table we can have a backup of the Original source table.

A staging table is crucial as it preserves the original data, allowing safe transformations and error recovery without affecting the source.

```sql
DROP TABLE IF EXISTS layoffs_staging2;


CREATE TABLE layoffs_staging2
SELECT *
FROM layoffs_staging;
```

Backed up the Original source table for data integrity.

2) **Removing Duplicates:** Finding the Duplicates using ROW_NUMER() function

```
SELECT *,
ROW_NUMBER()
OVER(PARTITION BY company,location,industry,total_laid_off) as RN
FROM layoffs_staging;
```

Deleting the Duplicate rows using the Delete function.

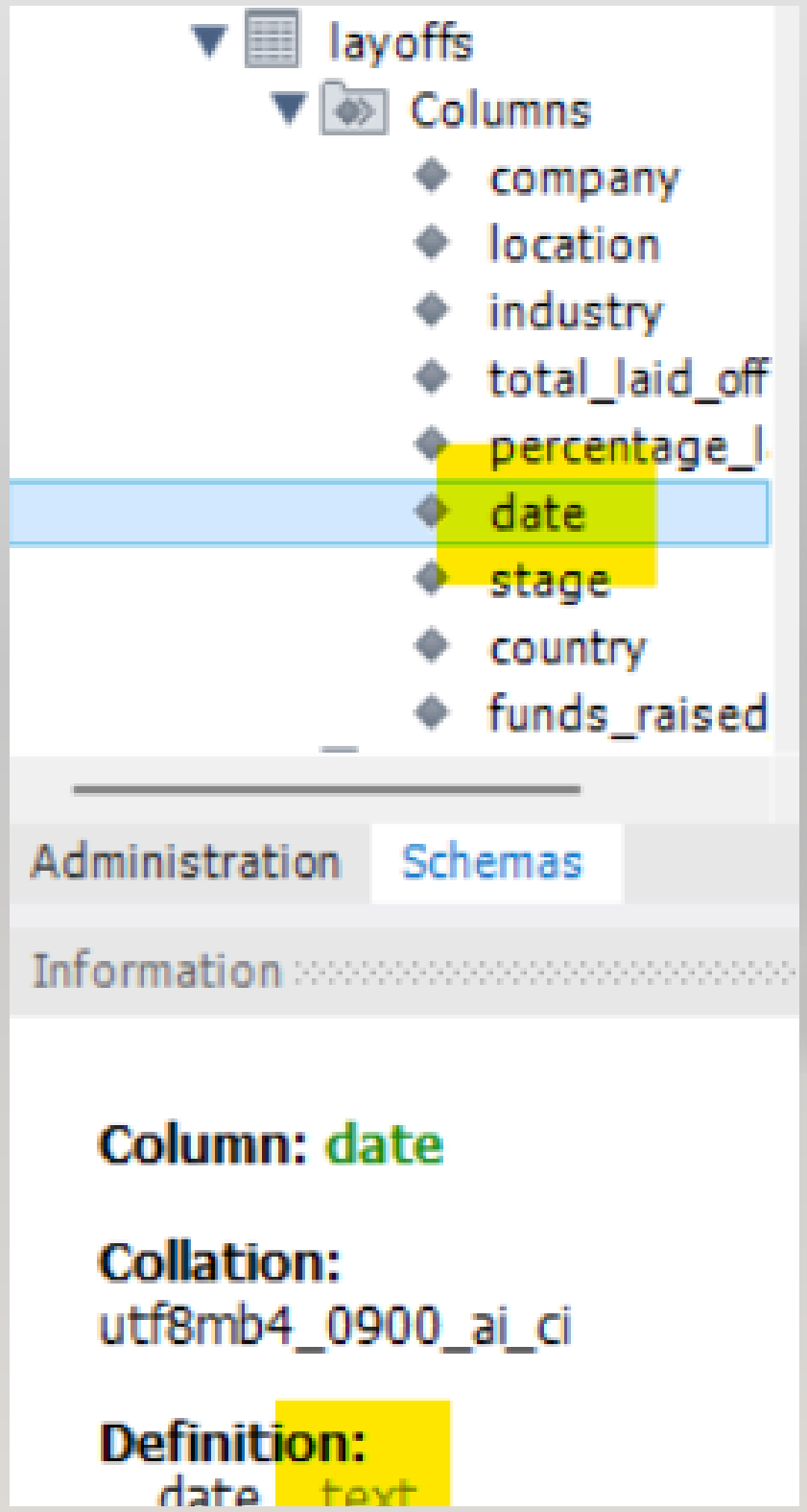
```
47 •    DELETE FROM
48      layoffs_staging2
49      WHERE RN > 1;
```

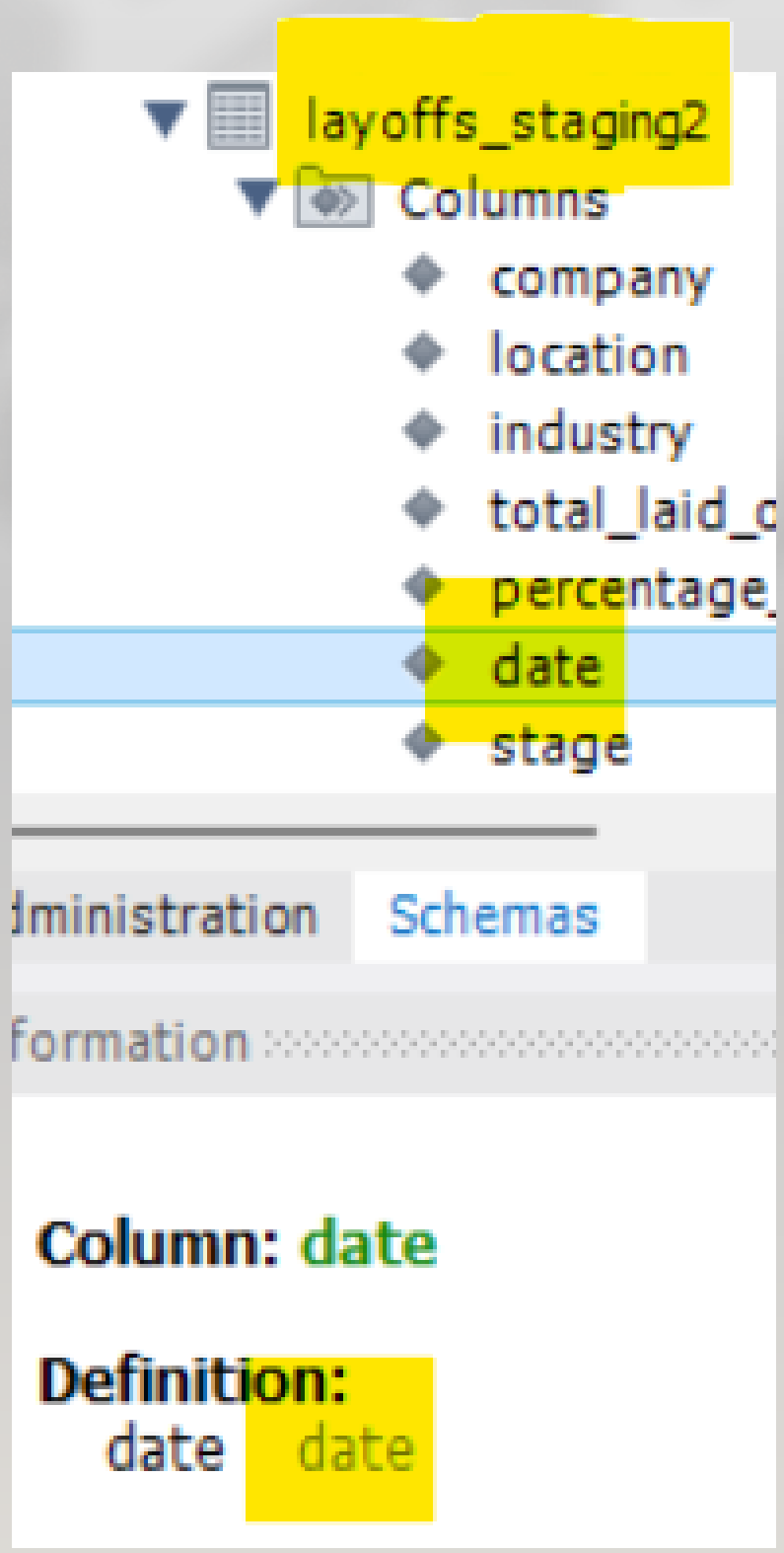
3) **Handling Blank (or) Null Values:** I have tried to populate values because in this scenario there is a column called industry that has value in one row and missing in the other.

```
-- Airbnb, Carvana & Juul has Industry populated in Few rows - So Populate the same in  empty or Null row
-- Only for Same location

-- Select & See company industries empty & not empty by default

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

Used a self join to populate values for this scenario.

4) Standardizing the Data Type: One of the columns in the source file had an incorrect data type, so I corrected it in the staging table during data cleaning.





```sql
-- CONVERT date column datatype
ALTER TABLE layoffs_staging2 MODIFY `date` date;
```

Key Insight: In SQL, DATE is a reserved keyword, but since it was used as a column name in our table, we enclosed it in backticks (`) to avoid conflicts.

5) Removing Columns that are not needed: I created a column called RN using a CTE to store row numbers for identifying duplicates. Since it wasn't needed for my future EDA, I removed it.

```sql
ALTER TABLE layoffs_staging2 DROP COLUMN RN;
```

COMMENT YOUR THOUGHTS!