## Initial Stages of Data Analysis Using SQL — A Case Study on Layoff Data

The initial stage of data analysis is crucial for ensuring data quality, integrity, and readiness for advanced analytical procedures. The points below show SQL-based methodology for performing the initial stages of data analysis using a global layoffs dataset. Guided by seven diagnostic questions, the analysis examines the dataset's structure, completeness, consistency, and statistical properties. Through systematic query-based operations such as data inspection, type verification, missing value handling, deduplication, categorical normalisation, and mathematical profiling, the study demonstrates how SQL serves as a precise and efficient tool for foundational data preprocessing. The findings underscore the essential role of early-stage data analysis in ensuring meaningful and accurate analytical outcomes.

Data analysis begins long before feature engineering or visualization. Raw datasets often contain structural inconsistencies, missing information, and noise that can distort insights. The following shares seven questions that play their role in the early phases of data preparation due to their efficiency in querying, modifying, and auditing large datasets.

Question 1. How does this data look like and take a backup of the data?

The first step in exploratory data analysis (EDA) involves becoming familiar with the structure and content of the dataset.

```
Select * from layoffs;
```

To protect data integrity, a backup copy is created. This staging table allows modification without affecting the original dataset, maintaining traceability and analytical safety.

```
Select *
into layoffs_staging
from layoffs;
```

Question 2. How big is this data?

Understanding the volume of data helps estimate computational needs and gives context to the data's representativeness.

```
Select count(*) as total_row
from layoffs_staging;    -- Total number of rows
Select count(*) AS total_columns
from INFORMATION_SCHEMA.COLUMNS
 where table_name = 'layoffs_staging';  ---- Total number of columns
```

Question 3. What is the data type of column?

To ensure accurate calculations and transformations, it is essential to audit the data types of all columns. This evaluation identifies whether numeric fields are stored as strings, whether date formats require conversion, and whether categorical variables have appropriate length constraints.

```
Select
	column_name,
	data_type,
	character_maximum_length
from INFORMATION_SCHEMA.COLUMNS
where table_name = 'layoffs_staging';
```

Question 4. Are there any missing values?

Missing data are common in real-world datasets and can lead to biased analysis if unaddressed. After identifying them, rows with missing industry labels are removed.

```
SELECT *
FROM layoffs_staging
WHERE industry IS NULL
   or trim(industry) = '';  ---- Find how many cell are blank/Null

Delete from layoffs_staging
WHERE industry IS NULL
   or trim(industry) = '';   ---- Delete row that have are blank/Null
```

Question 5. Are there any duplicate values?

Duplicate records can distort counts, statistics, and trend analyses. This stage ensures only unique observations remain, strengthening data reliability.

```
with duplicate_cte as
(
Select *,
row_number() over(partition by company, location, industry,
total_laid_off, percentage_laid_off, stage, date, country,
funds_raised_millions order by company) as row_num
from layoffs_staging
)
Delete from duplicate_cte
where row_num > 1;
```

Question 6. Find and normalize values that have the same meaning but different forms.

Categorical inconsistencies often arise from variations in spelling or naming conventions. Eliminating such inconsistencies enhances data consistency and improves grouping accuracy during analysis.

```sql
Select Distinct industry
from layoffs_staging
order by industry;   ---- Find all Distinct values in the column

Update layoffs_staging
set industry = 'Crypto'
where industry like 'Crypto%';  ---- Update value with same meaning

Update layoffs_staging
set country = 'United States'
where country like 'United States%';
```

Question 7. How does the data look like mathematically?

Quantitative evaluation of numerical variables provides insight into distributional characteristics and identifies potential anomalies. These mathematical summaries support early detection of outliers, missing numeric patterns, or implausible values.

```
WITH math_calculation AS (
    SELECT
        TRY_CONVERT(DECIMAL(18,4), total_laid_off) AS total_laid_off_n,
        TRY_CONVERT(DECIMAL(18,4), percentage_laid_off) AS
percentage_laid_off_n,
        TRY_CONVERT(DECIMAL(18,4), funds_raised_millions) AS
funds_raised_millions_n
    FROM layoffs_staging
),
stats AS (
    SELECT 'count' AS metric,
           COUNT(total_laid_off_n) AS total_laid_off,
           COUNT(percentage_laid_off_n) AS percentage_laid_off,
           COUNT(funds_raised_millions_n) AS funds_raised_millions
    FROM math_calculation

    UNION ALL SELECT 'mean',
           AVG(total_laid_off_n),
           AVG(percentage_laid_off_n),
           AVG(funds_raised_millions_n)
    FROM math_calculation

    UNION ALL SELECT 'std',
           STDEV(total_laid_off_n),
           STDEV(percentage_laid_off_n),
           STDEV(funds_raised_millions_n)
    FROM math_calculation

    UNION ALL SELECT 'min',
           MIN(total_laid_off_n),
           MIN(percentage_laid_off_n),
           MIN(funds_raised_millions_n)
    FROM math_calculation

    UNION ALL SELECT 'max',
           MAX(total_laid_off_n),
           MAX(percentage_laid_off_n),
           MAX(funds_raised_millions_n)
    FROM math_calculation
)
SELECT *
FROM stats
ORDER BY
    CASE metric
        WHEN 'count' THEN 1
        WHEN 'mean'  THEN 2
        WHEN 'std'   THEN 3
        WHEN 'min'   THEN 4
        WHEN 'max'   THEN 5
END;
```