**Case Study 2**

* **Load the provided dataset into a Databricks notebook.**

# Load the dataset from AWS S3 into a Spark DataFrame

df = spark.read.option("sep", ",") \

.option("header", True) \

.option("inferSchema",True) \

.csv("s3n://humber-lfb-databricks-class-files/midterm\_la.csv")

* **Display the first 10 rows to inspect the data structure.**

# Show the first 10 rows of the DataFrame to inspect the data

display(df.limit(10))

* **Explain the dataset and its structure.**

1. **Employee Identification**:
   * **Emp ID**: Unique identifier for each employee (integer).
   * **Name Prefix, First Name, Middle Initial, Last Name**: Components of the employee’s name (strings).
2. **Personal Information**:
   * **Gender**: Gender of the employee (string).
   * **E Mail**: Email address of the employee (string).
   * **Date of Birth**: Employee’s date of birth (date).
   * **Time of Birth**: Exact time of birth (string).
   * **Father's Name, Mother's Name, Mother's Maiden Name**: Names of the employee’s parents (strings).
3. **Professional Details**:
   * **Date of Joining**: Date when the employee joined the company (date).
   * **Quarter of Joining, Half of Joining, Year of Joining, Month of Joining, Month Name of Joining, Short Month, Day of Joining, DOW of Joining, Short DOW**: Various attributes detailing the joining date (string and integer).
   * **Age in Company (Years)**: Number of years the employee has been with the company (double).
   * **Salary**: Employee’s salary (integer).
   * **Last % Hike**: Percentage increase in salary last received (string).
4. **Additional Information**:
   * **SSN**: Social Security Number (string).
   * **Phone No.**: Contact phone number (string).
   * **Place Name, County, City, State, Zip**: Location details of the employee’s residence (string and integer).
   * **Region**: Geographical region of employment (string).
5. **Login Credentials**:
   * **User Name, Password**: Employee’s login credentials (string).
6. **Company Information**:
   * **CompanyID**: Identifier for the company (integer).
7. **Additional Columns**:
   * **\_c37**: Extra or placeholder column, potentially for legacy or redundant data (string).

* **Create a line chart showing the trend of 'Age in Company (Years)' over time using the 'Date of Joining' field for the time axis. Identify and discuss any noticeable trends or patterns in the data.**
* A graph with a line

  Description automatically generated

**Line Chart:-**

* 1. **Declining Tenure:** The average tenure decreases significantly for employees who joined more recently, indicating shorter time periods with the company for newer hires.
  2. **Long-Term Stability**: Employees who joined in earlier years have notably higher average tenures, reflecting long-term commitment and stability within the company
* **Create a bar chart to visualize the average salary by region.**

A blue rectangle with white text

Description automatically generated

**Bar Chart:-**

1. **Midwest Leads:** The Midwest region has the highest average salary, indicating it might offer better compensation compared to the other regions.

2. **West vs. South:** The West region has a higher average salary than the South, though not as high as the Midwest. This suggests that while salaries in the West are competitive, they are not as high as those in the Midwest.

3. **South with the Lowest Average:** The South has the lowest average salary among the regions. This could reflect regional differences in cost of living or market salary rates.

* **Create a temporary view and write an SQL query to identify the top 3 highest-paid employees by gender.** A screenshot of a computer

  Description automatically generated
* **Share any 3 insights that you can derive from this dataset**

1. New Employees Have Shorter Tenures: People who joined the company more recently have shorter average tenures compared to those who joined many years ago. This means new hires haven't been with the company as long as the older employees.
2. Salaries Vary by Region: Employees in the Midwest earn more on average than those in the West or South. This suggests that salaries might be higher in the Midwest, possibly due to regional differences in living costs or job market conditions.
3. Top Salaries and Gender: The highest salaries are very close in amount, with two of the top earners being male and one being female. This shows that while there are high salaries for both genders, there may be more to explore about salary equality.