**QA/QC DATAHUT INTERNSHIP JUNE 2024**

**COLUMNS IN THE DATASET**

1.Name

2.Age

3.Email

4.Join Date

5.Salary

6.Department

**TOOL USED**: Jupiter Notebook

**STEPS FOLLOWED IN THE DATA ANALYSIS**

**1.Loading the data**

Loaded the data using its file path.

**2.Inspecting the data**

Size, information, of NaN values were checked.

**3.Name column**

Checked all names are in string datatype.

Converted to title case.

Removed special characters.

NaN values are replaced as ‘unknown’

Removed duplicates.

**4.Age column**

Detected and handled outliers using IQR method.

Replaced missing values using mean of ages

Found min and max age, and considered age range greater than max age as invalid ages.

**5.Email**

Defined a regex pattern for valid emails and found out valid emails.

replaced invalid email as unknown

**6.Join Date**

By defining using a parsing function found out dates with date\_ year\_first, then found out date starting with day first and removed other complex date as Nat.

Then applied this function to join date.

Separated a column named date\_year\_first.

Then removed join date column and renamed date\_year\_first as join date.

NaN values replaced as unknown.

**7.Salary**

Filled NaN values column with mean salary.

Changed decimal points of salary column.

**8.Department**

Used department mapping.

Correct department name using department mapping and then identified valid departments and filtered out invalid departments.

**9.Duplicates**

Found out duplicates and filtered out.