Assignment for Data QA & QC Internship @ Datahut

Data Cleaning Task Instructions

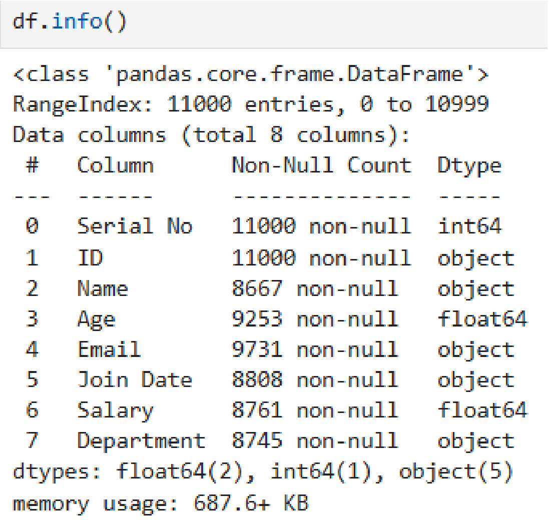
I am provided with a dataset named messy_Data.csv. My task is to clean this dataset and ensure it is ready for analysis.

Following Steps were taken for the data cleaning process.

1. **Load the Data:**
   - Load the dataset into a jupyter notebook.
   - pd: read_csv used for this method
2. **Inspect the Data:**
   - Understood the dimension
   - understood the repetition in some columns to study the data patterns
   - I feel there is some commonality between Unnamed 0 and ID
   - Renamed Unnamed 0 to Serial No.

Handling the date format and Department Correction before handling the NULL values

3. **Standardise Date Formats:**

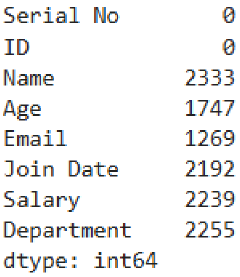


```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11000 entries, 0 to 10999
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Serial No   11000 non-null  int64  
 1   ID          11000 non-null  object 
 2   Name        8667 non-null   object 
 3   Age         9253 non-null   float64
 4   Email       9731 non-null   object 
 5   Join Date   8808 non-null   object 
 6   Salary      8761 non-null   float64
 7   Department  8745 non-null   object 
dtypes: float64(2), int64(1), object(5)
memory usage: 687.6+ KB
```

   - After checking the df info() converted the data format datetime format (YYYY-MM-DD).
4. **Correct Department Names:**
   - Corrected the Department name before handling the missing values by using check the unique department names that are mentioned.

5. **Handle Missing Values:**
   - Almost all column have null values

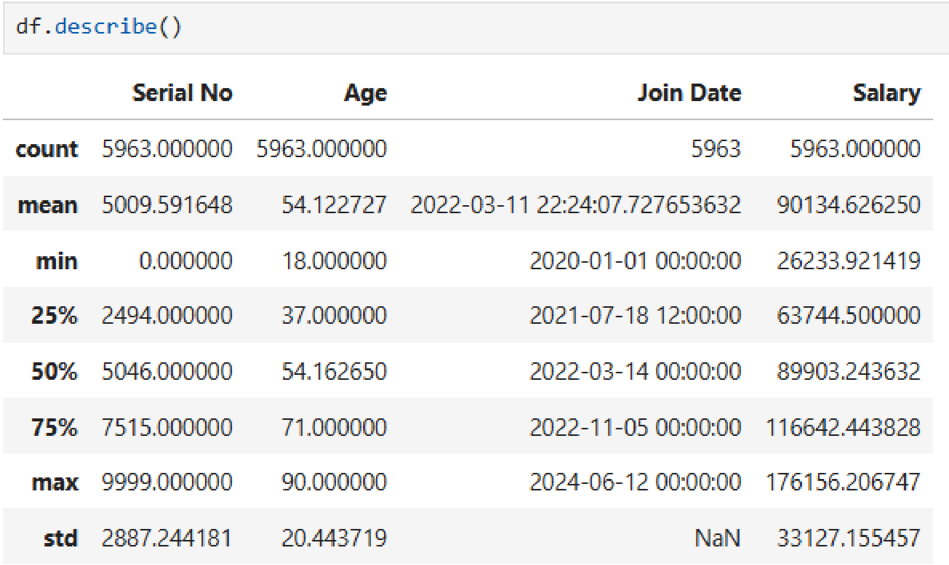


```
Serial No        0
ID               0
Name          2333
Age           1747
Email         1269
Join Date     2192
Salary        2239
Department    2255
dtype: int64
```

- Handled: Name and age together being NULL in many places.
- Age and Salary are normally distributed. Filled the NULL with **Mean**
- Department NULL was filled with **Most frequent** value
- Join date was filled with **Median** Value

6. **Remove Duplicates:**
   - There are some rows with same Serial No and ID. I have kept only 1 row of such data
   - Same email id was repeated in more than 1 entry. Removed the duplicate and kept the first occurrence only
7. **Correct Email Formats:**
   - Regular expression to make all email addresses follow a standard format (e.g., username@domain.com).
   - Regular expression used: r'^[\w\.-]+@[a-zA-Z\d\.-]+\.(com|info|net|org|biz)$'
8. **Clean Name Fields:**
   - NULL names are dropped
   - Found some name is having 3rd name. Trimmed out the 3rd name
   - Found some name with title. Dropped the title and just kept the first name and Last name
   - Found some extraneous words added towards the end of the Surname code added to remove those. Keeping all the remaining as same.
   - Logic Followed to trim out the extraneous word is as follows:
     - List1 : Found all the Last name from df['Name']
     - List2: Found repeating Last names (more than 2) from df['Name']
     - List3 : keeping the List1 as it is but just if the last name matched the string in surname then replace the last name with the **valid surname** This will remove the extraneous word from the **List1.**
   - Some code to verify the correction:
     df[df['Name'].str.contains('Taylordaughter', case=False)]
     df[df['Name'].str.contains('Lamb', case=False)]
     We can see difference between surnames in df['Name'] and df['Name New']

9. **Handle Salary Noise:**
   - Salary column was verified



```
df.describe()
```

	Serial No	Age	Join Date	Salary
count	5963.000000	5963.000000	5963	5963.000000
mean	5009.591648	54.122727	2022-03-11 22:24:07.727653632	90134.626250
min	0.000000	18.000000	2020-01-01 00:00:00	26233.921419
25%	2494.000000	37.000000	2021-07-18 12:00:00	63744.500000
50%	5046.000000	54.162650	2022-03-14 00:00:00	89903.243632
75%	7515.000000	71.000000	2022-11-05 00:00:00	116642.443828
max	9999.000000	90.000000	2024-06-12 00:00:00	176156.206747
std	2887.244181	20.443719	NaN	33127.155457

- Mean and Median(50%) is approximately same. Giving a normally distributed effect
- Minimum value is 26233.921419 and Maximum value is 176156.206747 which does not show any visible noise presence.

10. Save the dataset

- Data save to cleaned_dataset.csv