Datahut QA Assignment

- The pandas, NumPy and re libraries are imported.
- The data is loaded to a pandas data frame.
- The top and bottom few rows of the data is inspected to get a glimpse of the data.
- The number of rows and columns of the data is retrieved.
- The names of the columns are checked.
- Unwanted column which is the index column is removed.
- The number of missing values is printed.

Column Name	Not-null rows	Missing values	Missing Percentage
ID	11000	0	0%
Name	8667	2333	21.2%
Age	9253	1747	15.8%
Email	9731	1269	11.5%
Join Date	8808	2192	19.9%
Salary	8761	2239	20.3%
Department	8745	2255	20.5%

- To have an easier cleaning process the rows with more than 50% of null values are removed hence reducing the size of the data frame.
- After removing partially empty rows we then check them for the missing values.

Column Name	Not-null rows	Missing values	Missing Percentage
ID	9725	0	0%
Name	8667	1058	10.8%
Age	9249	476	4.8%
Email	9725	0	0%
Join Date	8808	917	9.4%
Salary	8760	965	9.9%
Department	8744	981	10%

- We have understood the quality issues of the data.
- The removal of partial empty rows has reduced the missing values significantly.

- Here the data consisted of seven columns with ID column which doesn't have a clear format, Name column which gives the name of the person, Age column with the age of the person, next column include the email ID of the person, then the joining date of the person, then comes the salary and the department in which the person works.
- The cleaning process is carried out, initially the data types of the entries is checked and found out that different columns contains different types of data.
- Then we created a data frame to add cleaned data after which we changed the age column and salary column to numeric data type.
- Then we filled the age column with the median of the age which is the middle most value in the age column (when sorted), we assumed most of the people to have similar age.
- Filled the salary column with the mean of the salary or the average of all the values in the same column.
- Then the float type salary is a rounded off to an integer type. This is done assuming that Salary is a whole number.
- After that a pattern for the email is set and then a column is created named valid_email to give Boolean values as entries if the email is valid True and if not False for the non-valid emails.
- After that the valid email which is the column with true Boolean entries are kept and the rest is removed.
- Checking the size of the data frame the extra created column valid_email is dropped.
- Now we check the entries of the name column.
- In the next step the join date column is renamed by removing the white spaces as errors may occur if the column name contained white spaces.
- The type of join date column is set to date by specifying the format and then it is also cleaned using backward fill. Here backward fill is used as we assume that two nearby entries are people who joined during the same time.
- Now we go through the distinct entries of the department column to correct it with the specified set of values and then after we change the department data type to string.
- We standardize the department names using the wild card like matching function and then we retrieve the unique values in the department column after which there is an extra white space which should be removed from the department column. This white space entry this replaced with the

mode of the department column that is the most commonly occurring value in the column.

- Now we check for any missing values present and find that the name column contains missing values. When cleaning the name column there is a lot of errors and the whole data frame is destroyed, so we make a new data frame and then copy the name column to it and then do the further operations.
- To clean the name column, we remove any trailing or leading white spaces and then fill the NaN values with ‘unknown’ which means the name is unknown we also replace the empty string with the name unknown.
- Now after checking that the name column is filled, we then move this cleaned column to our previous data frame.
- After conforming that the column contains zero missing values, we check the data frame.
- We just check for the salary outliers, and we find that the salary column does not contain any outliers, but the values are all floating-point numbers. So, the salary column is rounded off to integer type.
- We also check for outliers in Age column and set the maximum age to 75, assuming people retire at 75years.
- We check for duplicate rows then we clean all the duplicate rows and move it to a new data frame which is them exported to a new CSV file named cleaned_dataset.