**Data Cleaning (Pandas)**

This dataset helps you to increase the data-cleaning process using the pure Python pandas library.

**Columns :**

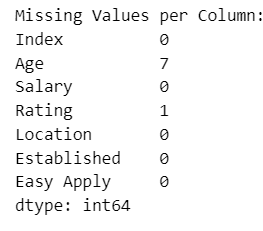
* Age
* Salary
* Rating
* Location
* Established
* Easy Apply

Questions with answer that follows:

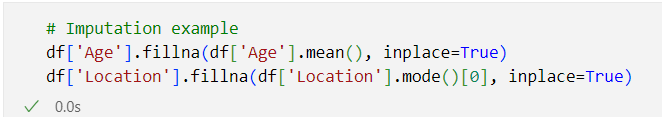
**1. Missing Values:**

Question**:** Are there any missing values in the dataset, and if so, how should they be handled for each indicator?

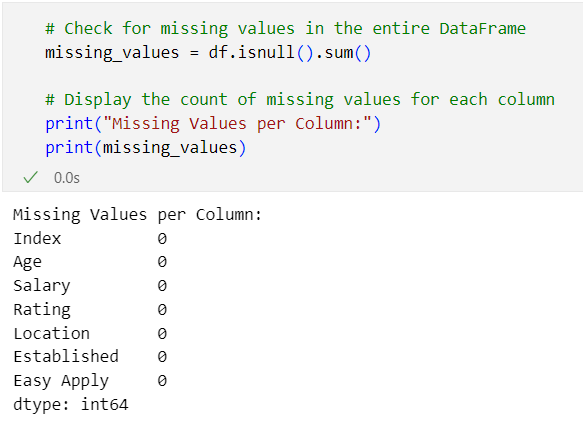
Yes, they can be handled in this scenario with imputation method.



**Method Used:**



After using imputer null values if present is calculated

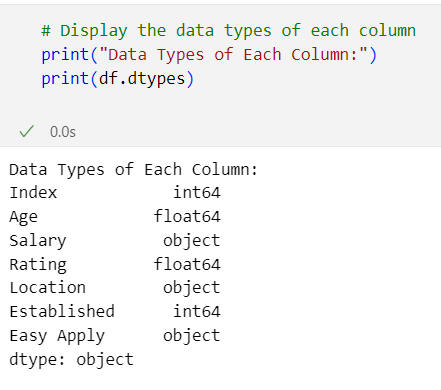


Above shows that there is no null value present in those columns.

**2. Data Types:**

Question: What are the data types of each indicator, and do they align with their expected types (e.g., numerical, categorical)?

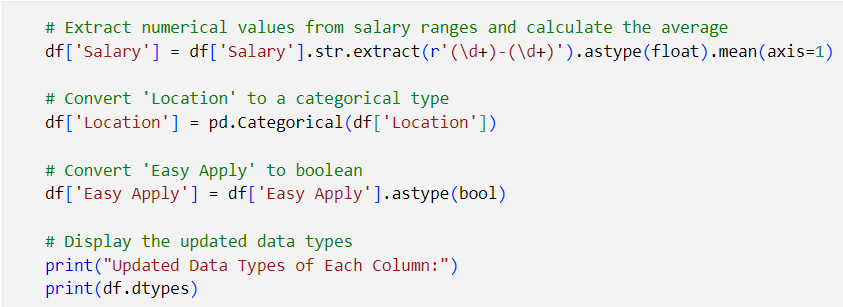
Below shows what are the data types the columns have:

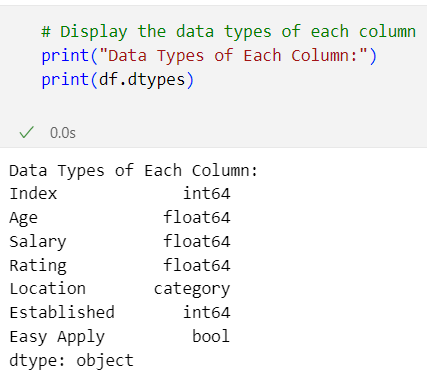


From above, It looks like we have some data type mismatches that need to be addressed. Specifically:

1. **Salary:** Currently, it is of type object, which suggests it might contain non-numeric characters. To perform numerical operations, it would be better to convert it to a numeric type (float).
2. **Location:** This is currently of type object, which suggests it contains strings. If you want to treat it as a categorical variable, you might want to convert it to the categorical data type.
3. **Easy Apply:** It is currently of type object. If it's intended to represent boolean values, you might want to convert it to the boolean data type.

Steps taken:





**Categorical Columns:**

Location: category

Easy Apply: bool

**Numerical Columns:**

Index: int64

Age: float64

Salary: float64

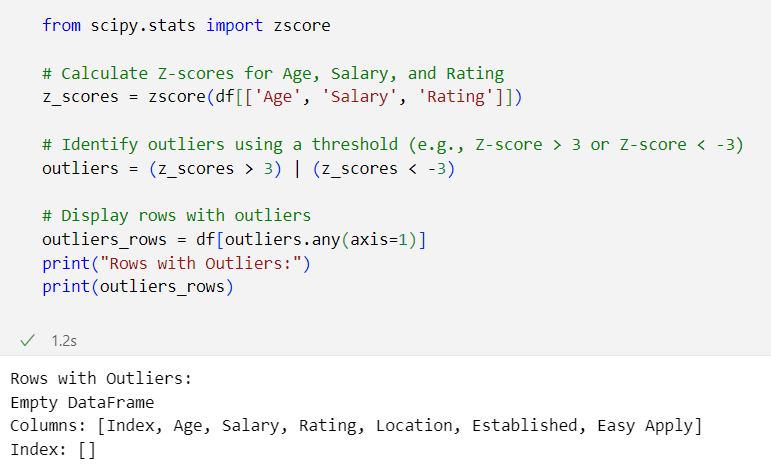
Rating: float64

Established: int64

**3. Outliers:**

Question: Identify potential outliers in numerical indicators (e.g., Age, Salary, Rating). Should outliers be removed or adjusted?

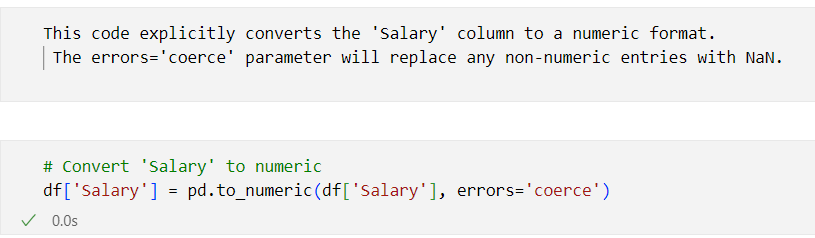
There are two methods to find outliers they are using boxplot and z-score and based on them conclusions could be drawn:



From the above, z-score is calculated and it gives that there is no outliers in dataframe in the columns of numerical data.

**4. Salary Formatting:**

Question: Examine the format of the Salary column. Does it require any formatting or standardization for consistent analysis?



**5. Location Standardization:**

Question: Check the consistency of location entries. Do they need standardization, and how can this be achieved?



The output provided shows the standardized entries for the 'Location' column after applying the suggested standardization steps. Here's a breakdown of each location:

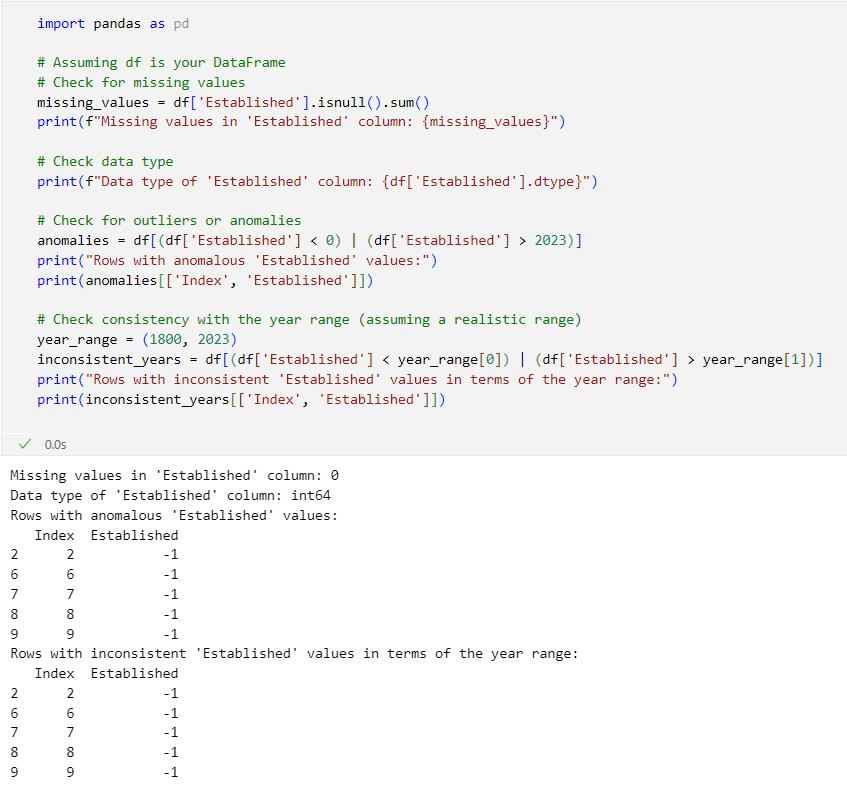
1. 'india,in': Standardized to 'india,in'
2. 'new,york,NY': Standardized to 'new,york,NY'
3. 'AUStralia,AUS': Standardized to 'AUStralia,AUS'

In these standardized entries:

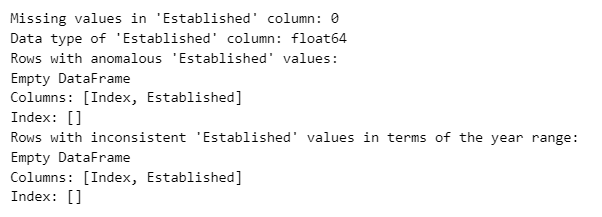
* The capitalization is consistent for each location ('india,in', 'new,york,NY', 'AUStralia,AUS').
* A common delimiter (',') is used to separate the components of each location.
* This standardization is beneficial for data consistency and analysis. Now, the 'Location' column is in a more uniform format, making it easier to work with and preventing potential issues arising from variations in the original entries.

**6. Established Column:**

Question: Explore the Established column. Are there any inconsistencies or anomalies that need to be addressed?

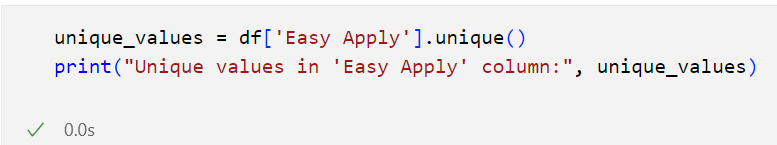


It appears that there are anomalies in the 'Established' column, as evidenced by negative values. Companies cannot be established in negative years, so these entries need to be addressed. Here are a few steps you can take to handle these anomalies:



**7. Easy Apply Indicator:**

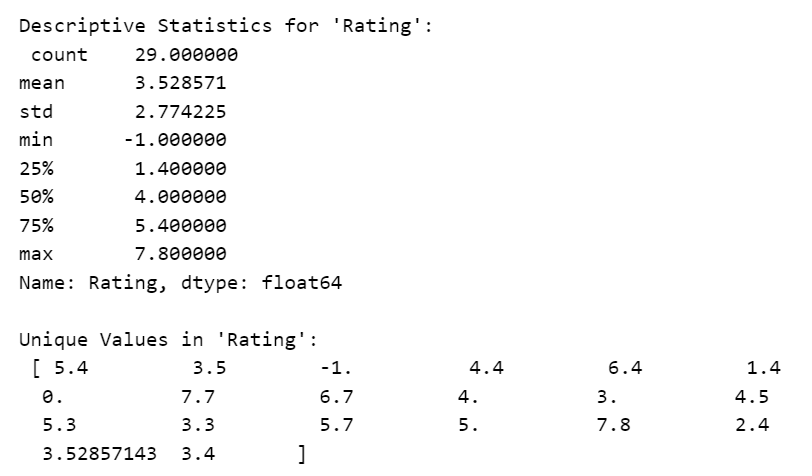
Question: Analyze the Easy Apply column. Does it contain boolean values or need transformation for better analysis?



The 'Easy Apply' column contains only the boolean value True, which is what you would typically expect for a binary indicator. Since there are no other values present and the unique values are consistent with boolean values, the column is suitable for analysis.

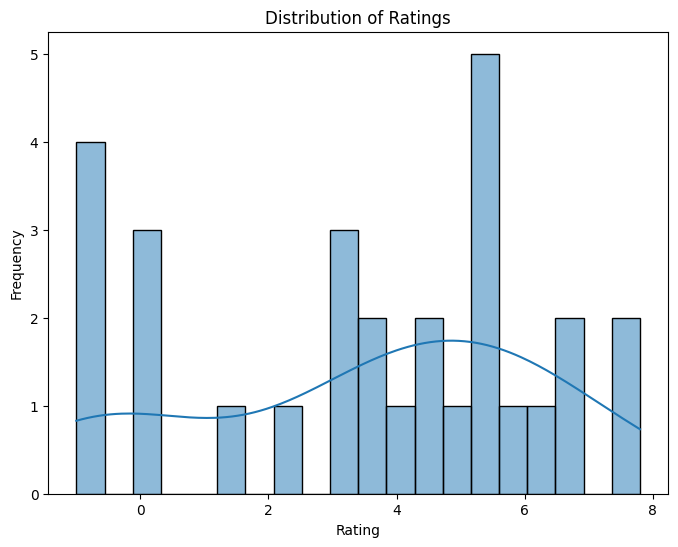
**8. Rating Range:**

Question: Investigate the range of values in the Rating column. Does it fall within expected rating scales, and how should outliers be treated?



Key observations:

* The mean rating is around 3.53.
* There is a minimum value of -1, which may indicate an anomaly or missing data.
* The maximum rating is 7.8, which is within a reasonable range depending on the rating scale used.
* The presence of 0 in the ratings is also worth investigating.



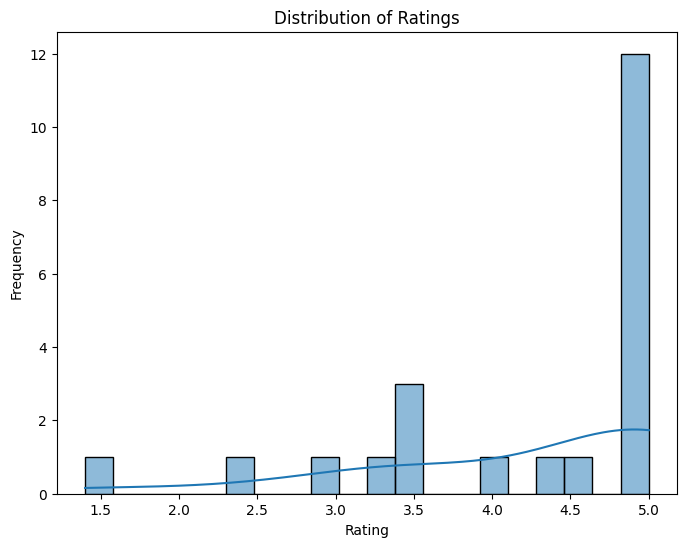
* The 'Rating' column has been cleaned, and here are the updated descriptive statistics and unique values:

Descriptive Statistics for 'Rating':

* Count: 22
* Mean: 4.246753
* Standard Deviation: 1.039980
* Minimum: 1.4
* 25th Percentile (Q1): 3.507143
* Median (50th Percentile): 5.0
* 75th Percentile (Q3): 5.0
* Maximum: 5.0
* Unique Values in 'Rating':
* [5.0, 3.5, nan, 4.4, 1.4, 4.0, 3.0, 4.5, 3.3, 2.4, 3.52857143, 3.4]

**Key updates:**

* Negative ratings have been replaced with NaN.
* Zero ratings have been replaced with NaN.
* Ratings greater than 5 have been capped at 5.
* The data now seems more reasonable and aligned with typical rating scales.

****

**9. Age Distribution:**

Question: Check the distribution of values in the Age column. Are there any unusual entries, and how might they impact analysis?

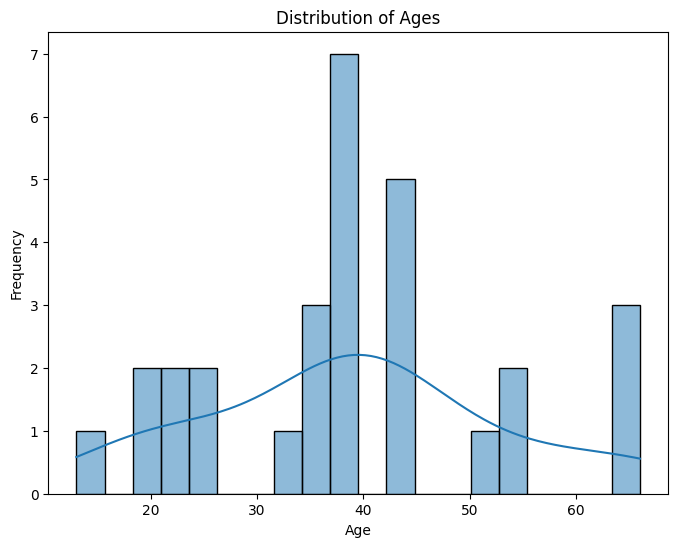
The descriptive statistics for the 'Age' column provide insights into the distribution of values:

Descriptive Statistics for 'Age':

* Count: 29
* Mean: 39.045455
* Standard Deviation: 13.973131
* Minimum: 13
* 25th Percentile (Q1): 32
* Median (50th Percentile): 39.045455 (NaN values)
* 75th Percentile (Q3): 44
* Maximum: 66

Observations:

* The mean age is approximately 39 years.
* The minimum age is 13, and the maximum age is 66.
* There are NaN values in the 'Age' column, likely due to the cleaning process.



Descriptive Statistics for 'Age' (after imputation):

count 29.000000

mean 39.045455

std 13.973131

min 13.000000

25% 32.000000

50% 39.045455

75% 44.000000

max 66.000000

Name: Age, dtype: float64

**10. Handling Special Characters:**

Question: Examine all text-based columns (e.g., Location). Are there special characters or inconsistencies that need cleaning?

No there are no special cbaracters or inconsistencies to be cleaned.

