**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

**1. Missing Values:**

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

**Answer**:

Check for missing values in each column using df.isnull().sum().

For numerical columns (Age, Salary, Rating), consider imputation with mean or median values.

For categorical columns (Location, Established, Easy Apply), use appropriate imputation methods, such as filling with mode or creating a separate category for missing values.

**2. Data Types:**

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

**Answer**:

Use df.dtypes to inspect the data types.

Ensure that Age, Salary, and Rating are numerical types.

Location and Established might be categorical or object types.

Easy Apply should be boolean.

3. **Outliers**:

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

**Answer**:

Use visualizations (box plots, histograms) to identify outliers.

Consider removing extreme outliers or applying transformations (e.g., log transformation) for better analysis.

Be cautious not to lose valuable information, and document the decisions made.

4. **Salary Formatting:**

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

**Answer**:

Check for consistent formats using regular expressions or string manipulation.

Standardize salary entries to a common format for accurate analysis.

Consider converting salary ranges to average values for numerical comparisons.

5. **Location Standardization:**

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

**Answer**:

Use string methods or libraries like fuzzy matching to standardize location entries.

Correct misspellings, abbreviations, or variations to ensure uniformity.

6. **Established Column:**

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

**Answer**:

Check for unique values and anomalies in the Established column.

Address inconsistencies, correct errors, or consider consolidating entries if necessary.

7. **Easy Apply Indicator:**

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

**Answer**:

Check unique values in the Easy Apply column.

If not boolean, transform the column to boolean format for consistency.

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?

**Answer**:

Examine unique values and distribution in the Rating column.

Consider scaling or normalizing ratings for uniform analysis.

Handle outliers cautiously, considering their impact on analysis.

9. **Age Distribution:**

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

**Answer**:

Examine the Age column distribution.

Investigate any unusual entries and assess their impact on analysis.

Consider removing or adjusting entries if necessary.

10. **Handling Special Characters:**

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

**Answer**:

Use string methods or regular expressions to identify and clean special characters.

Ensure consistent formatting for text-based columns.

11. **Data Integrity:**

Question: Ensure data integrity by cross-referencing entries. For instance, does the Established column align with the Age column?

**Answer**:

Cross-reference related columns (e.g., Established and Age) to ensure data integrity.

Resolve any discrepancies or errors found during the cross-referencing process.

12. **Easy Apply Transformation:**

Question: If the Easy Apply column contains non-boolean values, how can it be transformed into a usable format?

**Answer:**

Convert non-boolean values in the Easy Apply column to a boolean format.

Use appropriate transformations or mapping to achieve the desired format.

13. **Location Accuracy:**

Question: Assess the accuracy of location entries. Are there misspelled or ambiguous locations that require correction?

**Answer**:

Review location entries for accuracy.

Correct misspellings, abbreviations, or ambiguities to improve data quality.

14. **Handling Categorical Data:**

Question: For categorical indicators, consider encoding or transforming them into a format suitable for analysis.

**Answer**:

Use one-hot encoding or label encoding for categorical columns as needed.

Ensure proper handling of categorical data for accurate analysis.

15. **Consistent Rating Scale:**

Question: Ensure a consistent rating scale in the Rating column. Should it be normalized or adjusted for uniform analysis?

**Answer**:

Normalize the rating scale if necessary for consistent analysis.

Document any adjustments made to maintain transparency in the data cleaning process.