Documentation

1.Missing Values:

There were only 2 columns with missing rows.

a) Age: (calculating mean and replacing the missing values)

b) Rating: (calculating average and then replacing them with missing values)

2. Data Types:

Data Type of each indicator:

* Age 🡪 float
* Salary 🡪object
* Rating 🡪float
* Location 🡪object
* Established 🡪int
* Easy Apply 🡪object 🡪 bool (changed data type)

3.Outliers:

The outlier column seems to be Ratings.

We will be replacing the outlier value to NaN then to mean value of ratings column.

4.Salary Formatting:

a) We replaced this ‘$’ sign with a space

b) We replaced ‘k’ with ‘000’

c) We then added space to ‘ - ’

5.Location Standardisation:

Location column lacked consistency. It had Ny ,In ,Aus abbreviation for New York, India, Australia.

We just eliminated the abbreviations.

6.Established Column:

Replaced the inconsistent values with median value of that column, because we wanted the integer data for this column for analysis.

7. Easy Apply Indicator:

a) It had -1 as an entry in the column. So we just replaced it with ‘False’.

b) Then the column data type was converted to bool.

8. Rating Range:

This column had some missing values we replaced it with mean value of that column. The range is -5 to 5 , I kept it as it was cause some ratings can be worse than 0.

Outliers were replaced by mean value.

9.Age Distribution:

The Age column have some entries of very young and very old people.

This can have an impact on the normalisation of age group while concluding.

10. Handling Special Character:

Text based columns like: Location and Salary.

Location 🡪 The commas and abbreviations were removed

Salary 🡪 The dollar sign, k and a hyphen were removed and replaced by zeros and space, respectively.

11. Data Integrity:

Yes, the data integrity has been maintained.

The Age and Established column align perfectly.

12. Easy Apply Transformation:

The non - boolean values were converted to Boolean values.

(-1) 🡪 False

13. Location Accuracy:

There were just some abbreviations which were replaced by full forms, as this column does not contain specified location it will be more convenient to have full name of places.

14. Handling Categorial Data:

Yes, with respect to every column the data type and data values have been transformed and mentioned in above points as well.

15. Consistent Rating Scale:

As in earlier codes we have replaced the null value of rating column with average rating value it will be more appropriate to adjust the column rather than normalising.

As the Adjusting method uses average value.