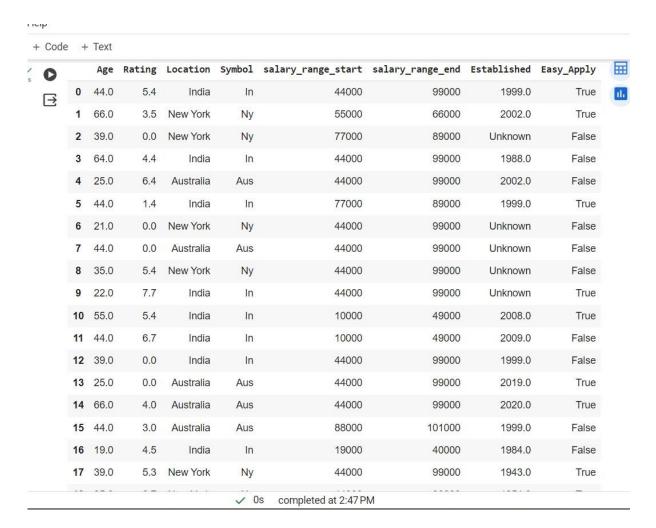


# **Week-3 Project**

By : S. Sindu

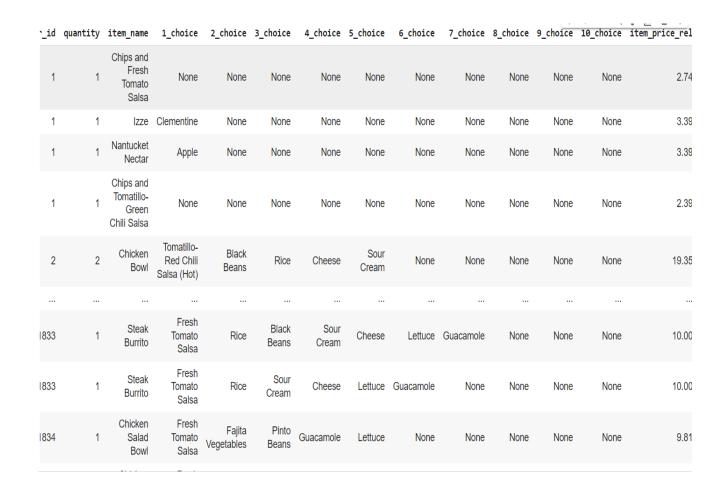
# Dataset -1

	Index	Age	Salary	Rating	Location	Established	Easy Apply
0	0	44.0	\$44k-\$99k	5.4	India,In	1999	TRUE
1	1	66.0	\$55k-\$66k	3.5	New York,Ny	2002	TRUE
2	2	NaN	\$77k-\$89k	-1.0	New York,Ny	-1	-1
3	3	64.0	\$44k-\$99k	4.4	India In	1988	-1
4	4	25.0	\$44k-\$99k	6.4	Australia Aus	2002	-1
5	5	44.0	\$77k-\$89k	1.4	India,In	1999	TRUE
6	6	21.0	\$44k-\$99k	0.0	New York,Ny	-1	-1
7	7	44.0	\$44k-\$99k	-1.0	Australia Aus	-1	-1
8	8	35.0	\$44k-\$99k	5.4	New York,Ny	-1	-1
9	9	22.0	\$44k-\$99k	7.7	India,In	-1	TRUE
10	10	55.0	\$10k-\$49k	5.4	India,In	2008	TRUE
11	11	44.0	\$10k-\$49k	6.7	India,In	2009	-1
12	12	NaN	\$44k-\$99k	0.0	India,In	1999	-1
13	13	25.0	\$44k-\$99k	-1.0	Australia Aus	2019	TRUE
14	14	66.0	\$44k-\$99k	4.0	Australia Aus	2020	TRUE
15	15	44.0	\$88k-\$101k	3.0	Australia Aus	1999	-1
16	16	19.0	\$19k-\$40k	4.5	India,In	1984	-1
17	17	NaN	\$44k-\$99k	5.3	New York,Ny	1943	TRUE



#### Dataset -2





# **Data Cleaning:**

### 1. Missing Values:

• The datasets contain missing values and are cleaned using measures of central tendency [mean, median, mode].

### 2. Data Types:

- First dataset attribute "Easy\_Apply" mistyped as "String" and "item\_price" attribute in second dataset is also "String".
  - These two are changed to "bool" and "float" types respectively.

## 3. Removing Inconsistencies:

- First dataset contains "Easy Apply" attribute which is inconsistent and has to be changed as "Easy Apply" for consistency.
- The First dataset contains "Salary" attribute which is difficult to analyze for the system. We will remove the symbol '\$' and replace 'k' with '000' and after replacing split the column into two columns naming "salart\_range\_start" and "salary\_range\_end".
- The second dataset contains "item\_price" having '\$' symbol. We will remove the symbol and make the attribute to the type "float" for Integrity.

### 4. Rearranging Columns:

- The column "choice-description" in second dataset is having null values and contains list.
- So, we will unlist the columns and take attributes from
  - "choice\_description" as [ "1\_choice", "2\_choice", "3\_choice", "4\_choice", "5\_choice", "6\_choice", "7\_choice", "8\_choice", "9\_choice", "10\_choice"].
- The choices are placed here and if there are no choices, we will place None in place of them.

### 5. Splitting Data:

• First dataset contains Location Attribute with name and symbol. We will split them as two columns and make them consistent.

#### 6. Mis-values:

• There are missed values in first dataset and they are replaced by 'Unknown' for the attribute 'Established' and 'false' for the attribute 'Easy\_Apply'.

#### 7. Standardization:

• Here, the column 'item\_name' has special characters in it. We will remove those special characters.