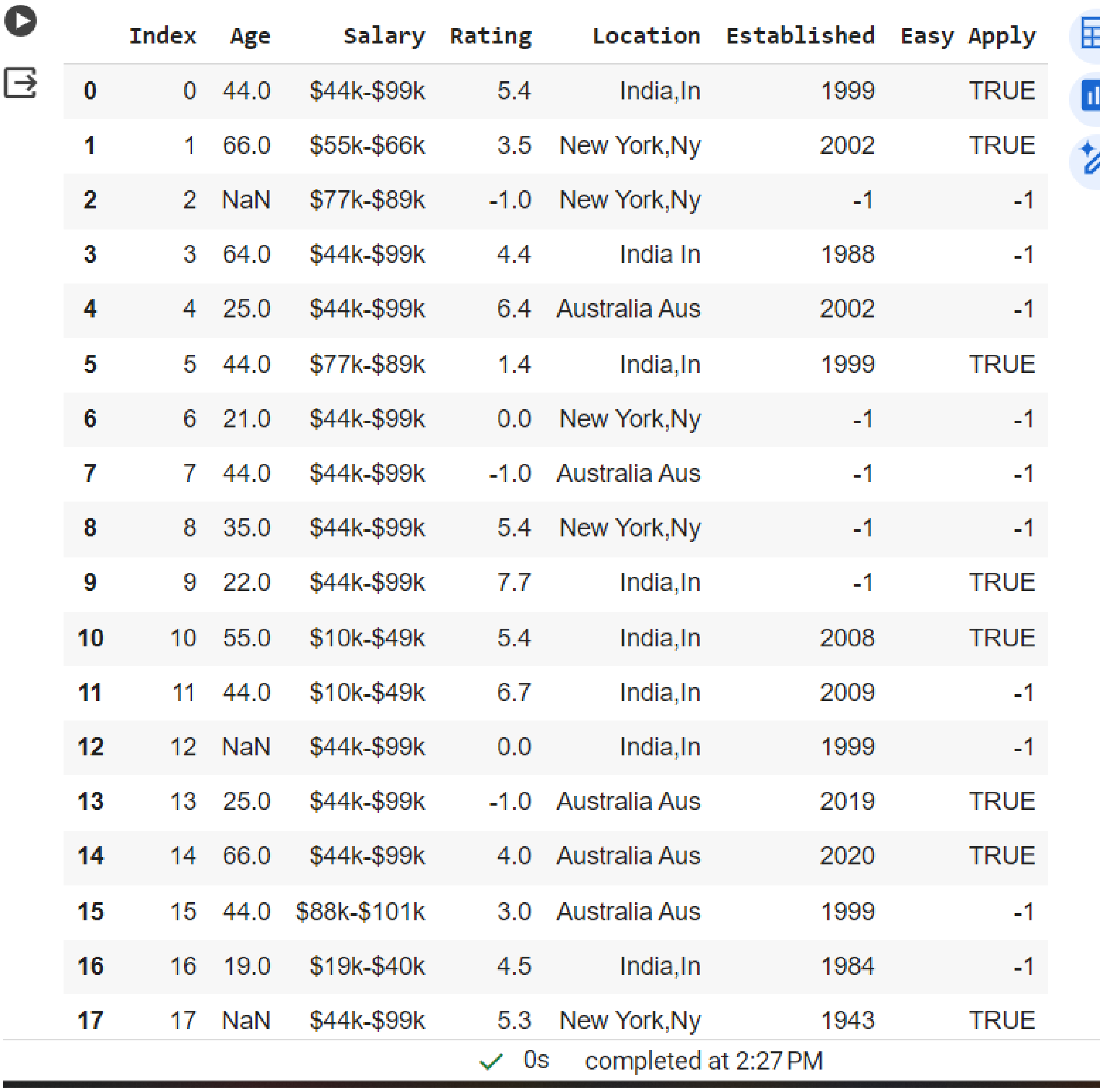
PrepInsta

Week-3 Project

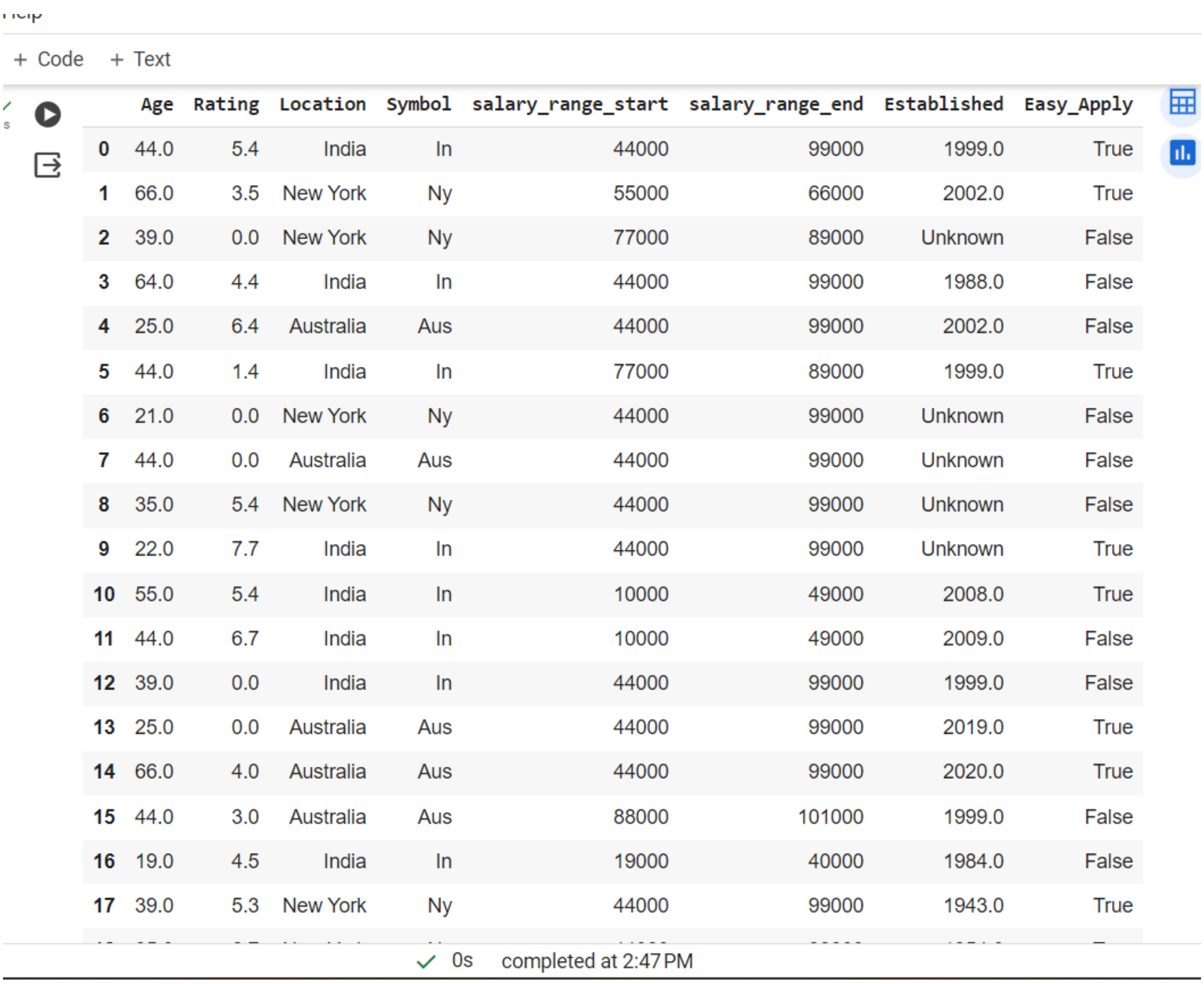
By : S. Sindu

Dataset -1



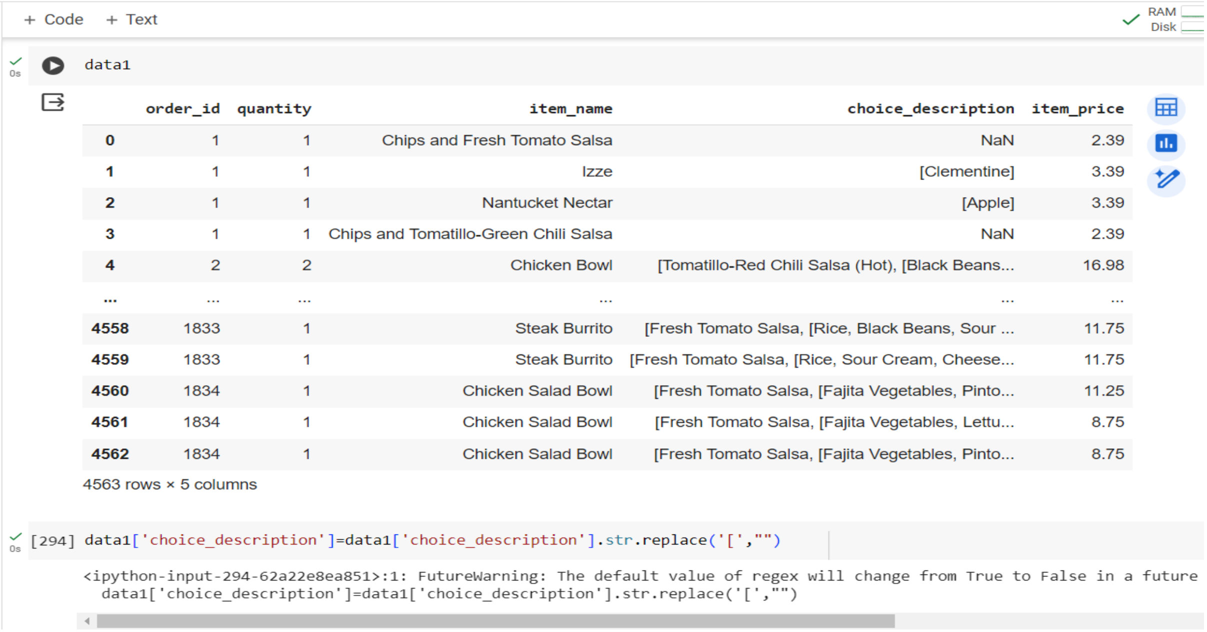
	Index	Age	Salary	Rating	Location	Established	Easy Apply
0	0	44.0	$44k-$99k	5.4	India,In	1999	TRUE
1	1	66.0	$55k-$66k	3.5	New York,Ny	2002	TRUE
2	2	NaN	$77k-$89k	-1.0	New York,Ny	-1	-1
3	3	64.0	$44k-$99k	4.4	India In	1988	-1
4	4	25.0	$44k-$99k	6.4	Australia Aus	2002	-1
5	5	44.0	$77k-$89k	1.4	India,In	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

✓ 0s completed at 2:27 PM



	Age	Rating	Location	Symbol	salary_range_start	salary_range_end	Established	Easy_Apply
0	44.0	5.4	India	In	44000	99000	1999.0	True
1	66.0	3.5	New York	Ny	55000	66000	2002.0	True
2	39.0	0.0	New York	Ny	77000	89000	Unknown	False
3	64.0	4.4	India	In	44000	99000	1988.0	False
4	25.0	6.4	Australia	Aus	44000	99000	2002.0	False
5	44.0	1.4	India	In	77000	89000	1999.0	True
6	21.0	0.0	New York	Ny	44000	99000	Unknown	False
7	44.0	0.0	Australia	Aus	44000	99000	Unknown	False
8	35.0	5.4	New York	Ny	44000	99000	Unknown	False
9	22.0	7.7	India	In	44000	99000	Unknown	True
10	55.0	5.4	India	In	10000	49000	2008.0	True
11	44.0	6.7	India	In	10000	49000	2009.0	False
12	39.0	0.0	India	In	44000	99000	1999.0	False
13	25.0	0.0	Australia	Aus	44000	99000	2019.0	True
14	66.0	4.0	Australia	Aus	44000	99000	2020.0	True
15	44.0	3.0	Australia	Aus	88000	101000	1999.0	False
16	19.0	4.5	India	In	19000	40000	1984.0	False
17	39.0	5.3	New York	Ny	44000	99000	1943.0	True

Dataset -2



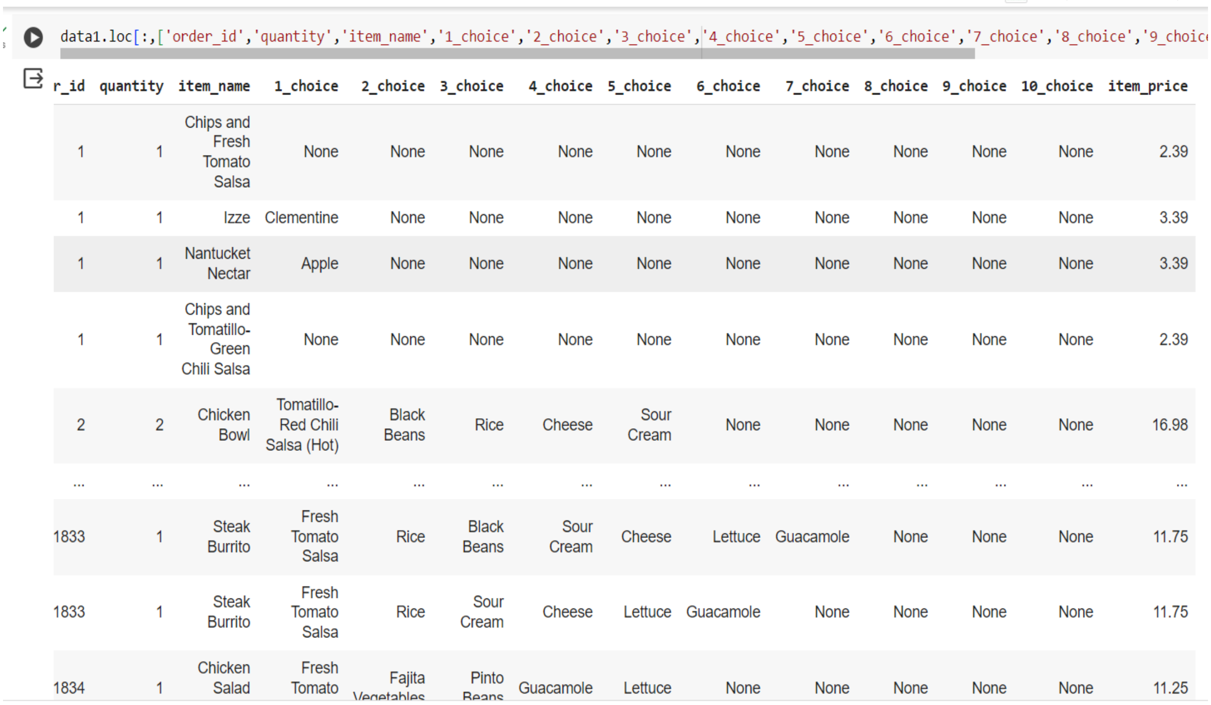
```
data1
```

	order_id	quantity	item_name	choice_description	item_price
0	1	1	Chips and Fresh Tomato Salsa	NaN	2.39
1	1	1	Izze	[Clementine]	3.39
2	1	1	Nantucket Nectar	[Apple]	3.39
3	1	1	Chips and Tomatillo-Green Chili Salsa	NaN	2.39
4	2	2	Chicken Bowl	[Tomatillo-Red Chili Salsa (Hot), [Black Beans...	16.98
...	...	...	...	...	...
4558	1833	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Black Beans, Sour ...	11.75
4559	1833	1	Steak Burrito	[Fresh Tomato Salsa, [Rice, Sour Cream, Cheese...	11.75
4560	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	11.25
4561	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Lettu...	8.75
4562	1834	1	Chicken Salad Bowl	[Fresh Tomato Salsa, [Fajita Vegetables, Pinto...	8.75

4563 rows × 5 columns

```
[294] data1['choice_description']=data1['choice_description'].str.replace('[',"")
```

```
<ipython-input-294-62a22e8ea851>:1: FutureWarning: The default value of regex will change from True to False in a future
  data1['choice_description']=data1['choice_description'].str.replace('[',"")
```



```
data1.loc[:,['order_id','quantity','item_name','1_choice','2_choice','3_choice','4_choice','5_choice','6_choice','7_choice','8_choice','9_choic
```

r_id	quantity	item_name	1_choice	2_choice	3_choice	4_choice	5_choice	6_choice	7_choice	8_choice	9_choice	10_choice	item_price
1	1	Chips and Fresh Tomato Salsa	None	None	None	None	None	None	None	None	None	None	2.39
1	1	Izze	Clementine	None	None	None	None	None	None	None	None	None	3.39
1	1	Nantucket Nectar	Apple	None	None	None	None	None	None	None	None	None	3.39
1	1	Chips and Tomatillo-Green Chili Salsa	None	None	None	None	None	None	None	None	None	None	2.39
2	2	Chicken Bowl	Tomatillo-Red Chili Salsa (Hot)	Black Beans	Rice	Cheese	Sour Cream	None	None	None	None	None	16.98
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1833	1	Steak Burrito	Fresh Tomato Salsa	Rice	Black Beans	Sour Cream	Cheese	Lettuce	Guacamole	None	None	None	11.75
1833	1	Steak Burrito	Fresh Tomato Salsa	Rice	Sour Cream	Cheese	Lettuce	Guacamole	None	None	None	None	11.75
1834	1	Chicken Salad	Fresh Tomato	Fajita Vegetables	Pinto Beans	Guacamole	Lettuce	None	None	None	None	None	11.25

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’.