Data_set_Cleaning

January 2, 2024

1 1. Importing Libraries

[291]:
```
import pandas as pd
import missingno as mn
import warnings
warnings.filterwarnings("ignore")
```

2 2. Importing Dataset

[292]:
```
dataset=pd.read_csv("Data_Cleaning.csv")
```

3 3. Performing some basic methods

[293]:
```
dataset
```

[293]:

	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
18	18	35.0	$44k-$99k	6.7	New York,Ny	1954	TRUE

19	19	32.0	$44k-$99k	3.3	New York,Ny	1955	TRUE
20	20	NaN	$44k-$99k	5.7	New York,Ny	1944	TRUE
21	21	35.0	$44k-$99k	5.0	New York,Ny	1946	-1
22	22	19.0	$55k-$66k	7.8	New York,Ny	1988	TRUE
23	23	NaN	$44k-$99k	2.4	New York,Ny	1999	TRUE
24	24	13.0	$44k-$99k	-1.0	New York,Ny	1987	-1
25	25	55.0	$44k-$99k	0.0	Australia Aus	1980	TRUE
26	26	NaN	$55k-$66k	NaN	India,In	1934	TRUE
27	27	52.0	$44k-$99k	5.4	India,In	1935	-1
28	28	NaN	$39k-$88k	3.4	Australia Aus	1932	-1

[294]:
```
dataset.head()
```

[294]:

	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

[295]:
```
dataset.tail()
```

[295]:

	Index	Age	Salary	Rating	Location	Established	Easy Apply
24	24	13.0	$44k-$99k	-1.0	New York,Ny	1987	-1
25	25	55.0	$44k-$99k	0.0	Australia Aus	1980	TRUE
26	26	NaN	$55k-$66k	NaN	India,In	1934	TRUE
27	27	52.0	$44k-$99k	5.4	India,In	1935	-1
28	28	NaN	$39k-$88k	3.4	Australia Aus	1932	-1

[296]:
```
dataset.describe()
```

[296]:

	Index	Age	Rating	Established
count	29.000000	22.000000	28.000000	29.000000
mean	14.000000	39.045455	3.528571	1638.620690
std	8.514693	16.134781	2.825133	762.079599
min	0.000000	13.000000	-1.000000	-1.000000
25%	7.000000	25.000000	1.050000	1935.000000
50%	14.000000	39.500000	4.200000	1984.000000
75%	21.000000	50.000000	5.400000	1999.000000
max	28.000000	66.000000	7.800000	2020.000000

[297]:
```
dataset.shape
```

[297]: (29, 7)

[298]:
```
dataset.columns
```

[298]:
```
Index(['Index', 'Age', 'Salary', 'Rating', 'Location', 'Established',
       'Easy Apply'],
      dtype='object')
```

4 4. Data cleaning

4.1 Finding some missing values

[299]:
```
dataset.drop_duplicates(subset=['Index', 'Age', 'Salary', 'Rating', 'Location',␣
↪'Established',
      'Easy Apply'])
```

[299]:

	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
18	18	35.0	$44k-$99k	6.7	New York,Ny	1954	TRUE
19	19	32.0	$44k-$99k	3.3	New York,Ny	1955	TRUE
20	20	NaN	$44k-$99k	5.7	New York,Ny	1944	TRUE
21	21	35.0	$44k-$99k	5.0	New York,Ny	1946	-1
22	22	19.0	$55k-$66k	7.8	New York,Ny	1988	TRUE
23	23	NaN	$44k-$99k	2.4	New York,Ny	1999	TRUE
24	24	13.0	$44k-$99k	-1.0	New York,Ny	1987	-1
25	25	55.0	$44k-$99k	0.0	Australia Aus	1980	TRUE
26	26	NaN	$55k-$66k	NaN	India,In	1934	TRUE
27	27	52.0	$44k-$99k	5.4	India,In	1935	-1
28	28	NaN	$39k-$88k	3.4	Australia Aus	1932	-1

[300]:
```
data_set=dataset.drop_duplicates(subset=['Index', 'Age', 'Salary', 'Rating',␣
↪'Location', 'Established',
      'Easy Apply'])
```

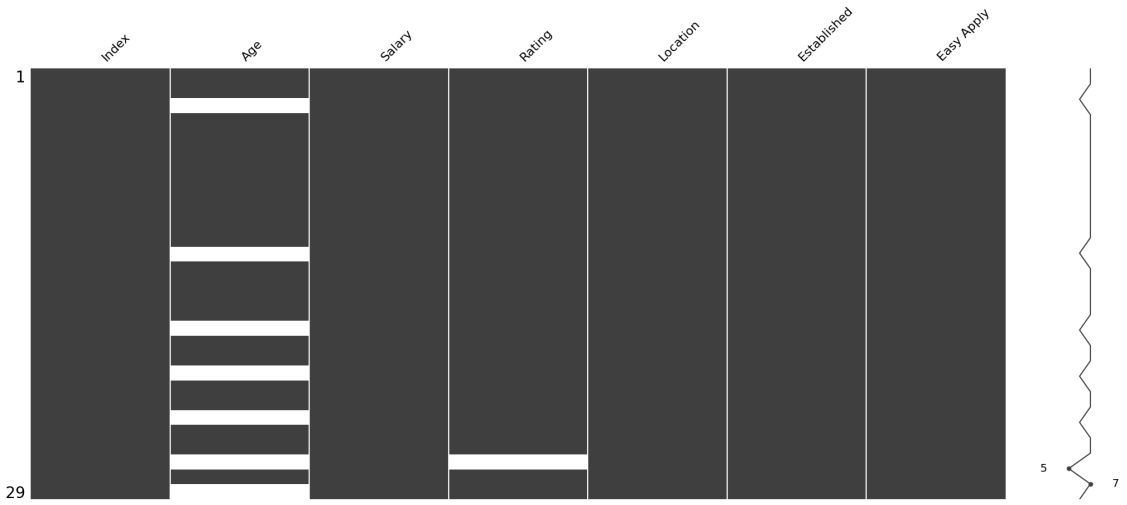
[301]:
```
data_set
```

[301]:

	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
18	18	35.0	$44k-$99k	6.7	New York,Ny	1954	TRUE
19	19	32.0	$44k-$99k	3.3	New York,Ny	1955	TRUE
20	20	NaN	$44k-$99k	5.7	New York,Ny	1944	TRUE
21	21	35.0	$44k-$99k	5.0	New York,Ny	1946	-1
22	22	19.0	$55k-$66k	7.8	New York,Ny	1988	TRUE
23	23	NaN	$44k-$99k	2.4	New York,Ny	1999	TRUE
24	24	13.0	$44k-$99k	-1.0	New York,Ny	1987	-1
25	25	55.0	$44k-$99k	0.0	Australia Aus	1980	TRUE
26	26	NaN	$55k-$66k	NaN	India,In	1934	TRUE
27	27	52.0	$44k-$99k	5.4	India,In	1935	-1
28	28	NaN	$39k-$88k	3.4	Australia Aus	1932	-1

[302]:
```
#here we have to use missingno library for missing values using matrix for␣
 ↪graphical representation
mn.matrix(data_set)
```

[302]:
```
<Axes: >
```



[303]:
```
data_set.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 29 entries, 0 to 28
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Index        29 non-null     int64
 1   Age          22 non-null     float64
 2   Salary       29 non-null     object
 3   Rating       28 non-null     float64
 4   Location     29 non-null     object
 5   Established  29 non-null     int64
 6   Easy Apply   29 non-null     object
dtypes: float64(2), int64(2), object(3)
memory usage: 1.8+ KB
```

[304]:
```
data_set.isnull().sum()
```

[304]:
```
Index          0
Age            7
Salary         0
Rating         1
Location       0
Established    0
Easy Apply     0
dtype: int64
```

[305]:
```
from sklearn.impute import SimpleImputer
imputer=SimpleImputer(strategy='most_frequent',missing_values=np.nan)
```

```
imputer=imputer.fit(data_set[['Index', 'Salary', 'Rating', 'Location',␣
 ↪'Established',
       'Easy Apply']])
data_set[['Index', 'Salary', 'Rating', 'Location', 'Established',
       'Easy Apply']]=imputer.transform(data_set[['Index', 'Salary', 'Rating',␣
 ↪'Location', 'Established',
       'Easy Apply']])
```

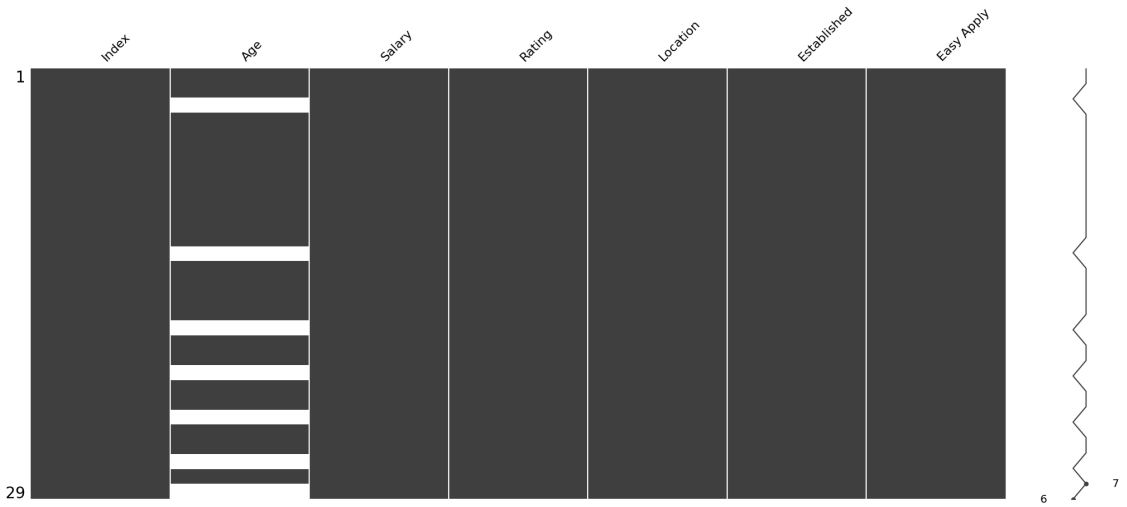
[306]:
```
data_set
```

[306]:

	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
18	18	35.0	$44k-$99k	6.7	New York,Ny	1954	TRUE
19	19	32.0	$44k-$99k	3.3	New York,Ny	1955	TRUE
20	20	NaN	$44k-$99k	5.7	New York,Ny	1944	TRUE
21	21	35.0	$44k-$99k	5.0	New York,Ny	1946	-1
22	22	19.0	$55k-$66k	7.8	New York,Ny	1988	TRUE
23	23	NaN	$44k-$99k	2.4	New York,Ny	1999	TRUE
24	24	13.0	$44k-$99k	-1.0	New York,Ny	1987	-1
25	25	55.0	$44k-$99k	0.0	Australia Aus	1980	TRUE
26	26	NaN	$55k-$66k	-1.0	India,In	1934	TRUE
27	27	52.0	$44k-$99k	5.4	India,In	1935	-1
28	28	NaN	$39k-$88k	3.4	Australia Aus	1932	-1

[307]:
```
mn.matrix(data_set)
```

[307]:
```
<Axes: >
```



[308]:
```
data_set['Easy Apply'].unique()
```

[308]:
```
array(['TRUE', '-1'], dtype=object)
```

[309]:
```
# Replace specific terms in 'Easy Apply' column
data_set['Easy Apply'].replace({'-1': 'FALSE'}, inplace=True)
```

[310]:
```
data_set['Easy Apply'].unique()
```

[310]:
```
array(['TRUE', 'FALSE'], dtype=object)
```

[311]:
```
data_set['Established'].unique()
```

[311]:
```
array([1999, 2002, -1, 1988, 2008, 2009, 2019, 2020, 1984, 1943, 1954,
       1955, 1944, 1946, 1987, 1980, 1934, 1935, 1932], dtype=object)
```

[312]:
```
# Replace specific terms in 'Established' column
data_set['Established'].replace({-1: 'unknown'}, inplace=True)
```

[313]:
```
data_set['Established'].unique()
```

[313]:
```
array([1999, 2002, 'unknown', 1988, 2008, 2009, 2019, 2020, 1984, 1943,
       1954, 1955, 1944, 1946, 1987, 1980, 1934, 1935, 1932], dtype=object)
```

[314]:
```
data_set['Salary'].unique()
```

[314]:
```
array(['$44k-$99k', '$55k-$66k', '$77k-$89k', '$10k-$49k', '$88k-$101k',
       '$19k-$40k', '$39k-$88k'], dtype=object)
```

[315]:
```
# Replace specific terms in 'Salary' column
data_set['Salary']=data_set['Salary'].str.replace('k', '000')
```

[316]:
```
data_set['Salary']
```

[316]:
```
0       $44000-$99000
1       $55000-$66000
2       $77000-$89000
3       $44000-$99000
4       $44000-$99000
5       $77000-$89000
6       $44000-$99000
7       $44000-$99000
8       $44000-$99000
9       $44000-$99000
10      $10000-$49000
11      $10000-$49000
12      $44000-$99000
13      $44000-$99000
14      $44000-$99000
15     $88000-$101000
16      $19000-$40000
17      $44000-$99000
18      $44000-$99000
19      $44000-$99000
20      $44000-$99000
21      $44000-$99000
22      $55000-$66000
23      $44000-$99000
24      $44000-$99000
25      $44000-$99000
26      $55000-$66000
27      $44000-$99000
28      $39000-$88000
Name: Salary, dtype: object
```

[317]:
```
data_set['Salary']=data_set['Salary'].str.replace('$', '')
```

[318]:
```
data_set['Salary']
```

[318]:
```
0       44000-99000
1       55000-66000
2       77000-89000
3       44000-99000
4       44000-99000
5       77000-89000
6       44000-99000
```

```
7       44000-99000
8       44000-99000
9       44000-99000
10      10000-49000
11      10000-49000
12      44000-99000
13      44000-99000
14      44000-99000
15     88000-101000
16      19000-40000
17      44000-99000
18      44000-99000
19      44000-99000
20      44000-99000
21      44000-99000
22      55000-66000
23      44000-99000
24      44000-99000
25      44000-99000
26      55000-66000
27      44000-99000
28      39000-88000
Name: Salary, dtype: object
```

[319]:
```
#Age
avg_age = data_set['Age'].mean()
avg_age
```

[319]:
```
39.04545454545455
```

[320]:
```
# fill the missing values with mean
data_set['Age'] = data_set.Age.fillna(avg_age)
data_set['Age'] = data_set.Age.round(decimals=1)
```

[321]:
```
#Rating
r1=data_set['Rating'].mean()
r1
```

[321]:
```
3.3724137931034486
```

[325]:
```
# fill the missing values with mean
data_set['Rating'] = data_set.Rating.fillna(r1)
data_set['Rating'] = data_set.Rating.round(decimals=0)
```

[326]:
```
# Replace specific terms in 'Rating' column
data_set['Rating'].replace({-1.0: r1}, inplace=True)
```

[327]:
```
data_set['Rating']
```

[327]:
```
0     5.0
1     4.0
2     3.0
3     4.0
4     6.0
5     1.0
6     0.0
7     3.0
8     5.0
9     8.0
10    5.0
11    7.0
12    0.0
13    3.0
14    4.0
15    3.0
16    4.0
17    5.0
18    7.0
19    3.0
20    6.0
21    5.0
22    8.0
23    2.0
24    3.0
25    0.0
26    3.0
27    5.0
28    3.0
Name: Rating, dtype: float64
```

[328]:
```
data_set
```

[328]:

	Index	Age	Salary	Rating	Location	Established	Easy Apply
0	0	44.0	44000-99000	5.0	India,In	1999	TRUE
1	1	66.0	55000-66000	4.0	New York,Ny	2002	TRUE
2	2	39.0	77000-89000	3.0	New York,Ny	unknown	FALSE
3	3	64.0	44000-99000	4.0	India In	1988	FALSE
4	4	25.0	44000-99000	6.0	Australia Aus	2002	FALSE
5	5	44.0	77000-89000	1.0	India,In	1999	TRUE
6	6	21.0	44000-99000	0.0	New York,Ny	unknown	FALSE
7	7	44.0	44000-99000	3.0	Australia Aus	unknown	FALSE
8	8	35.0	44000-99000	5.0	New York,Ny	unknown	FALSE
9	9	22.0	44000-99000	8.0	India,In	unknown	TRUE
10	10	55.0	10000-49000	5.0	India,In	2008	TRUE

```
11    11  44.0   10000-49000     7.0       India,In        2009      FALSE
12    12  39.0   44000-99000     0.0       India,In        1999      FALSE
13    13  25.0   44000-99000     3.0  Australia Aus        2019       TRUE
14    14  66.0   44000-99000     4.0  Australia Aus        2020       TRUE
15    15  44.0  88000-101000     3.0  Australia Aus        1999      FALSE
16    16  19.0   19000-40000     4.0       India,In        1984      FALSE
17    17  39.0   44000-99000     5.0     New York,Ny        1943       TRUE
18    18  35.0   44000-99000     7.0     New York,Ny        1954       TRUE
19    19  32.0   44000-99000     3.0     New York,Ny        1955       TRUE
20    20  39.0   44000-99000     6.0     New York,Ny        1944       TRUE
21    21  35.0   44000-99000     5.0     New York,Ny        1946      FALSE
22    22  19.0   55000-66000     8.0     New York,Ny        1988       TRUE
23    23  39.0   44000-99000     2.0     New York,Ny        1999       TRUE
24    24  13.0   44000-99000     3.0     New York,Ny        1987      FALSE
25    25  55.0   44000-99000     0.0  Australia Aus        1980       TRUE
26    26  39.0   55000-66000     3.0       India,In        1934       TRUE
27    27  52.0   44000-99000     5.0       India,In        1935      FALSE
28    28  39.0   39000-88000     3.0  Australia Aus        1932      FALSE
```

[331]:
```
data_set['Location'].value_counts()
```

[331]:
```
New York,Ny      12
India,In          9
Australia Aus     7
India In          1
Name: Location, dtype: int64
```

[333]:
```
data_set['Location'] = data_set['Location'].replace({'New York,Ny':'New York',␣
↪'India,In':'India','India In':'India', 'Australia Aus':'Australia'})
```

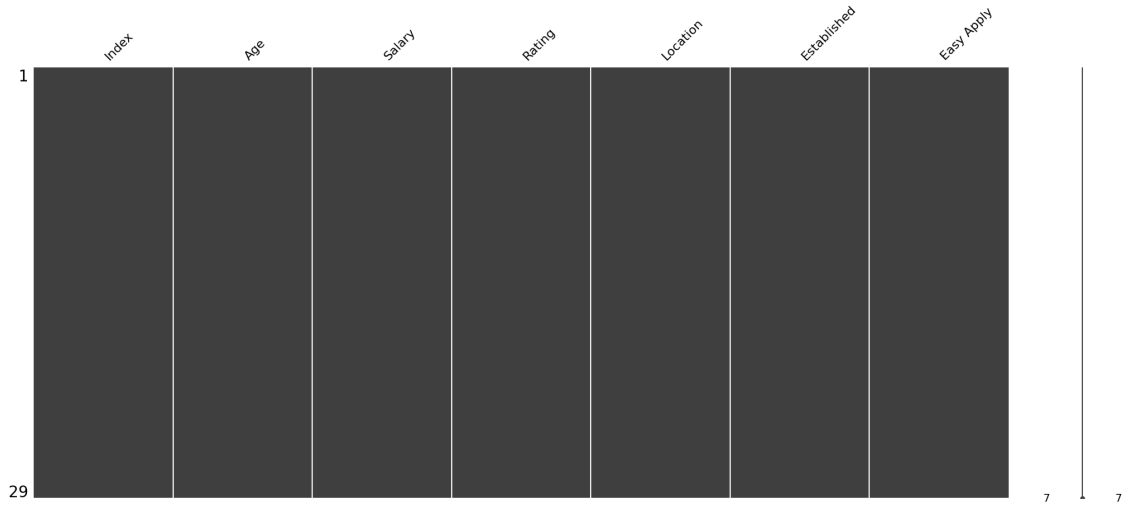
[334]:
```
data_set['Location']
```

[334]:
```
0         India
1      New York
2      New York
3         India
4     Australia
5         India
6      New York
7     Australia
8      New York
9         India
10        India
11        India
12        India
13    Australia
14    Australia
```

```
15    Australia
16        India
17     New York
18     New York
19     New York
20     New York
21     New York
22     New York
23     New York
24     New York
25    Australia
26        India
27        India
28    Australia
Name: Location, dtype: object
```

[335]:
```
mn.matrix(data_set)
```

[335]: <Axes: >



[336]:
```
data_set
```

[336]:

	Index	Age	Salary	Rating	Location	Established	Easy Apply
0	0	44.0	44000-99000	5.0	India	1999	TRUE
1	1	66.0	55000-66000	4.0	New York	2002	TRUE
2	2	39.0	77000-89000	3.0	New York	unknown	FALSE
3	3	64.0	44000-99000	4.0	India	1988	FALSE
4	4	25.0	44000-99000	6.0	Australia	2002	FALSE
5	5	44.0	77000-89000	1.0	India	1999	TRUE
6	6	21.0	44000-99000	0.0	New York	unknown	FALSE

7	7	44.0	44000-99000	3.0	Australia	unknown	FALSE
8	8	35.0	44000-99000	5.0	New York	unknown	FALSE
9	9	22.0	44000-99000	8.0	India	unknown	TRUE
10	10	55.0	10000-49000	5.0	India	2008	TRUE
11	11	44.0	10000-49000	7.0	India	2009	FALSE
12	12	39.0	44000-99000	0.0	India	1999	FALSE
13	13	25.0	44000-99000	3.0	Australia	2019	TRUE
14	14	66.0	44000-99000	4.0	Australia	2020	TRUE
15	15	44.0	88000-101000	3.0	Australia	1999	FALSE
16	16	19.0	19000-40000	4.0	India	1984	FALSE
17	17	39.0	44000-99000	5.0	New York	1943	TRUE
18	18	35.0	44000-99000	7.0	New York	1954	TRUE
19	19	32.0	44000-99000	3.0	New York	1955	TRUE
20	20	39.0	44000-99000	6.0	New York	1944	TRUE
21	21	35.0	44000-99000	5.0	New York	1946	FALSE
22	22	19.0	55000-66000	8.0	New York	1988	TRUE
23	23	39.0	44000-99000	2.0	New York	1999	TRUE
24	24	13.0	44000-99000	3.0	New York	1987	FALSE
25	25	55.0	44000-99000	0.0	Australia	1980	TRUE
26	26	39.0	55000-66000	3.0	India	1934	TRUE
27	27	52.0	44000-99000	5.0	India	1935	FALSE
28	28	39.0	39000-88000	3.0	Australia	1932	FALSE

5 chipotle data cleaning

chipotle Dataset

[13]:
```
import pandas as pd
import numpy as np
import missingno as mn
```

[14]:
```
df=pd.read_csv("output.csv")
```

[15]:
```
df
```

[15]:

	order_id	quantity	item_name
0	1	1	Chips and Fresh Tomato Salsa
1	1	1	Izze
2	1	1	Nantucket Nectar
3	1	1	Chips and Tomatillo-Green Chili Salsa
4	2	2	Chicken Bowl
…	…	…	…
4617	1833	1	Steak Burrito
4618	1833	1	Steak Burrito
4619	1834	1	Chicken Salad Bowl
4620	1834	1	Chicken Salad Bowl

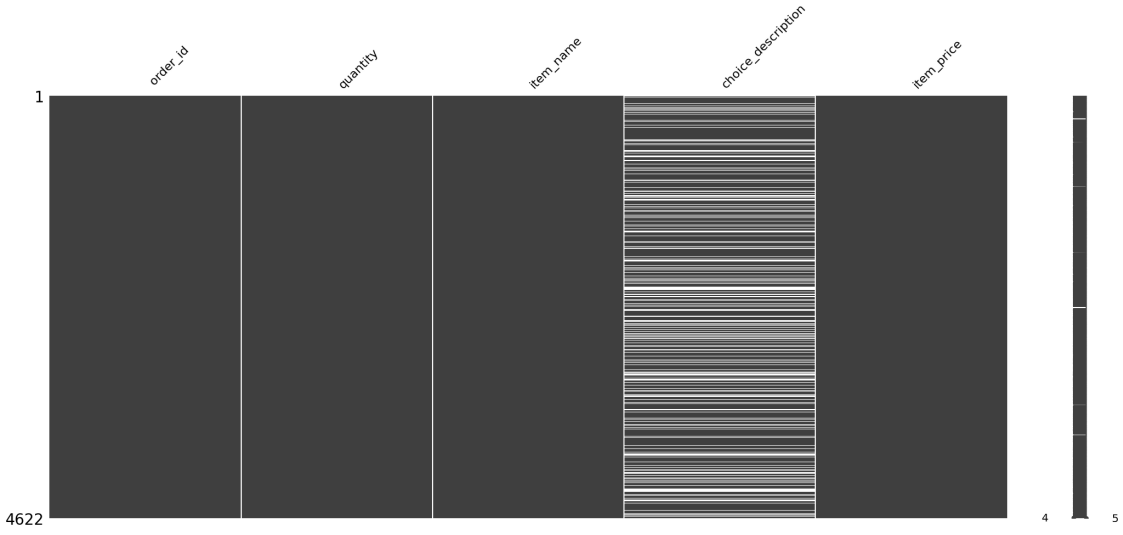
```
4621       1834         1                      Chicken Salad Bowl

                                    choice_description item_price
0                                                  NaN     $2.39
1                                          [Clementine]     $3.39
2                                               [Apple]     $3.39
3                                                  NaN     $2.39
4     [Tomatillo-Red Chili Salsa (Hot), [Black Beans…    $16.98
…                                                   …        …
4617  [Fresh Tomato Salsa, [Rice, Black Beans, Sour …    $11.75
4618  [Fresh Tomato Salsa, [Rice, Sour Cream, Cheese…    $11.75
4619  [Fresh Tomato Salsa, [Fajita Vegetables, Pinto…    $11.25
4620  [Fresh Tomato Salsa, [Fajita Vegetables, Lettu…     $8.75
4621  [Fresh Tomato Salsa, [Fajita Vegetables, Pinto…     $8.75

[4622 rows x 5 columns]
```

[16]:
```
mn.matrix(df)
```

[16]:
```
<Axes: >
```



[17]:
```
df.columns
```

[17]:
```
Index([' order_id', 'quantity', 'item_name', 'choice_description',
       'item_price'],
      dtype='object')
```

6 Data Cleaning

[18]:
```
df.drop_duplicates(subset=[' order_id', 'quantity', 'item_name',␣
 ↪'choice_description',
       'item_price'])
```

[18]:
```
      order_id  quantity                              item_name  \
0            1         1           Chips and Fresh Tomato Salsa
1            1         1                                   Izze
2            1         1                         Nantucket Nectar
3            1         1  Chips and Tomatillo-Green Chili Salsa
4            2         2                             Chicken Bowl
…          …       …                                     …
4617      1833         1                         Steak Burrito
4618      1833         1                         Steak Burrito
4619      1834         1                    Chicken Salad Bowl
4620      1834         1                    Chicken Salad Bowl
4621      1834         1                    Chicken Salad Bowl

                                  choice_description item_price
0                                                NaN     $2.39
1                                       [Clementine]     $3.39
2                                            [Apple]     $3.39
3                                                NaN     $2.39
4     [Tomatillo-Red Chili Salsa (Hot), [Black Beans…    $16.98
…                                                  …        …
4617  [Fresh Tomato Salsa, [Rice, Black Beans, Sour …    $11.75
4618  [Fresh Tomato Salsa, [Rice, Sour Cream, Cheese…    $11.75
4619  [Fresh Tomato Salsa, [Fajita Vegetables, Pinto…    $11.25
4620  [Fresh Tomato Salsa, [Fajita Vegetables, Lettu…     $8.75
4621  [Fresh Tomato Salsa, [Fajita Vegetables, Pinto…     $8.75

[4563 rows x 5 columns]
```

[19]:
```
df=df.drop_duplicates(subset=[' order_id', 'quantity', 'item_name',␣
 ↪'choice_description',
       'item_price'])
```

[20]:
```
df
```

[20]:
```
      order_id  quantity                              item_name  \
0            1         1           Chips and Fresh Tomato Salsa
1            1         1                                   Izze
2            1         1                         Nantucket Nectar
3            1         1  Chips and Tomatillo-Green Chili Salsa
4            2         2                             Chicken Bowl
…          …       …                                     …
```

```
4617       1833         1                          Steak Burrito
4618       1833         1                          Steak Burrito
4619       1834         1                      Chicken Salad Bowl
4620       1834         1                      Chicken Salad Bowl
4621       1834         1                      Chicken Salad Bowl

                                    choice_description item_price
0                                                  NaN     $2.39
1                                          [Clementine]     $3.39
2                                               [Apple]     $3.39
3                                                  NaN     $2.39
4     [Tomatillo-Red Chili Salsa (Hot), [Black Beans…    $16.98
…                                                   …        …
4617  [Fresh Tomato Salsa, [Rice, Black Beans, Sour …    $11.75
4618  [Fresh Tomato Salsa, [Rice, Sour Cream, Cheese…    $11.75
4619  [Fresh Tomato Salsa, [Fajita Vegetables, Pinto…    $11.25
4620  [Fresh Tomato Salsa, [Fajita Vegetables, Lettu…     $8.75
4621  [Fresh Tomato Salsa, [Fajita Vegetables, Pinto…     $8.75

[4563 rows x 5 columns]
```

[21]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4563 entries, 0 to 4621
Data columns (total 5 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0    order_id           4563 non-null   int64
 1   quantity            4563 non-null   int64
 2   item_name           4563 non-null   object
 3   choice_description  3335 non-null   object
 4   item_price          4563 non-null   object
dtypes: int64(2), object(3)
memory usage: 213.9+ KB
```

[22]:
```
df.isnull().sum()
```

[22]:
```
 order_id               0
quantity                0
item_name               0
choice_description    1228
item_price              0
dtype: int64
```

[23]:
```
from sklearn.impute import SimpleImputer
imputer=SimpleImputer(strategy='most_frequent',missing_values=np.nan)
```

```
imputer=imputer.fit(df[[' order_id', 'quantity', 'item_name',␣
 ↪'choice_description',
       'item_price']])
df[[' order_id', 'quantity', 'item_name', 'choice_description',
       'item_price']]=imputer.transform(df[[' order_id', 'quantity',␣
 ↪'item_name', 'choice_description',
       'item_price']])
```

```
<ipython-input-23-b4997613a522>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df[[' order_id', 'quantity', 'item_name', 'choice_description',
```

[24]:
```
df
```

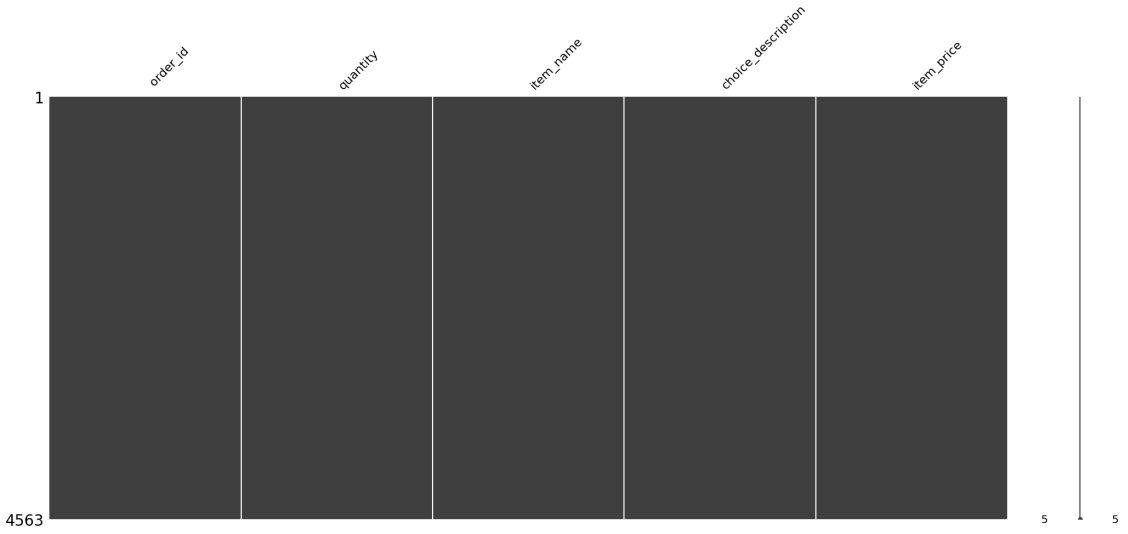
[24]:
```
      order_id quantity                              item_name  \
0            1        1           Chips and Fresh Tomato Salsa
1            1        1                                   Izze
2            1        1                       Nantucket Nectar
3            1        1  Chips and Tomatillo-Green Chili Salsa
4            2        2                           Chicken Bowl
…          …      …                                      …
4617      1833        1                          Steak Burrito
4618      1833        1                          Steak Burrito
4619      1834        1                    Chicken Salad Bowl
4620      1834        1                    Chicken Salad Bowl
4621      1834        1                    Chicken Salad Bowl

                                     choice_description item_price
0                                             [Diet Coke]     $2.39
1                                           [Clementine]     $3.39
2                                                [Apple]     $3.39
3                                             [Diet Coke]     $2.39
4     [Tomatillo-Red Chili Salsa (Hot), [Black Beans…    $16.98
…                                                    …        …
4617  [Fresh Tomato Salsa, [Rice, Black Beans, Sour …    $11.75
4618  [Fresh Tomato Salsa, [Rice, Sour Cream, Cheese…    $11.75
4619  [Fresh Tomato Salsa, [Fajita Vegetables, Pinto…    $11.25
4620  [Fresh Tomato Salsa, [Fajita Vegetables, Lettu…     $8.75
4621  [Fresh Tomato Salsa, [Fajita Vegetables, Pinto…     $8.75

[4563 rows x 5 columns]
```

[25]:
```
mn.matrix(df)
```

[25]: <Axes: >


order_id
quantity
item_name
choice_description
item_price
1
4563
5
5

[ ]: