15. Handling Duplicate Data

- The repetition of same data of one row is repeated in another row is called duplicate data

In [1]:
```
import pandas as pd
```

In [8]:
```
data = {'name':['a','b','c','d','a','c'], "eng":[8,7,5,8,8,5], "Urdu":[2,3,4,5,2,6]
data
```

Out[8]:
```
{'name': ['a', 'b', 'c', 'd', 'a', 'c'],
 'eng': [8, 7, 5, 8, 8, 5],
 'Urdu': [2, 3, 4, 5, 2, 6]}
```

In [9]:
```
df = pd.DataFrame(data)
df
```

Out[9]:

	name	eng	Urdu
0	a	8	2
1	b	7	3
2	c	5	4
3	d	8	5
4	a	8	2
5	c	5	6

- You can see that row number 0 and 4 have duplicate data
- row 2 and 5 are not duplicate, even the two values are identical, but to call a data duplicate exact data has to be there

In [14]:
```
# To identify the duplicate data
df.duplicated()
```

Out[14]:
```
0    False
1    False
2    False
3    False
4    False
5    False
dtype: bool
```

In [23]:
```
df['duplicate'] = df.duplicated()
df
```

Out[23]:

	name	eng	Urdu	duplicated	duplicate
0	a	8	2	False	False
1	b	7	3	False	False
2	c	5	4	False	False
3	d	8	5	False	False
4	a	8	2	False	True
5	c	5	6	False	False

In [24]:

```
df.drop('duplicate', axis=1, inplace=True)
```

In [25]:

```
df
```

Out[25]:

	name	eng	Urdu	duplicated
0	a	8	2	False
1	b	7	3	False
2	c	5	4	False
3	d	8	5	False
4	a	8	2	False
5	c	5	6	False

- Some ML algo also get train on duplicated data such as when we doing classification, so we should remove duplicate before data training

In [27]:

```
# To remove duplicated data
df.drop_duplicates()
```

Out[27]:

	name	eng	Urdu	duplicated
0	a	8	2	False
1	b	7	3	False
2	c	5	4	False
3	d	8	5	False
5	c	5	6	False

You can see that row 4 is deleted

In [29]:

```
df.drop('duplicated', axis=1, inplace=True)
```

In [30]:

```
df
```

Out[30]:

	name	eng	Urdu
0	a	8	2
1	b	7	3
2	c	5	4
3	d	8	5
4	a	8	2
5	c	5	6

Lets practice on orginal data

In [32]:

```
dataset = pd.read_csv('loan.csv')
dataset.head(3)
```

Out[32]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	C
0	LP001002	Male	No	0	Graduate	No	5849	
1	LP001003	Male	Yes	1	Graduate	No	4583	
2	LP001005	Male	Yes	0	Graduate	Yes	3000	

In [34]:

```
dataset.duplicated().sum()
```

Out[34]:

```
0
```

No duplicate is present in the data

Other way to see duplicates in the data:

In [36]:

```
dataset.shape
```

Out[36]:

```
(614, 13)
```

In [38]:

```
dataset.drop_duplicates(inplace=True)
```

In [40]:

```
dataset.shape
```

Out[40]:

```
(614, 13)
```

So you can see that the number of rows and columns are same before and after removing duplicates, so no duplicates are present in the data