

In [1]:

```
1 import pandas as pd
2 import numpy as np
```

Creating a DataFrame using numpy and pandas

Columns

- Name
- Gender
- Subject
- Department
- Mark

In [12]:

```
1 #Creating a array 'name' for the df
2 n=np.array(["S{}".format(i) for i in range(1,101)])
```

In [13]:

```
1 #Viewing the values present in n
2 n
```

Out[13]:

```
array(['S1', 'S2', 'S3', 'S4', 'S5', 'S6', 'S7', 'S8', 'S9', 'S10', 'S11',
      'S12', 'S13', 'S14', 'S15', 'S16', 'S17', 'S18', 'S19', 'S20',
      'S21', 'S22', 'S23', 'S24', 'S25', 'S26', 'S27', 'S28', 'S29',
      'S30', 'S31', 'S32', 'S33', 'S34', 'S35', 'S36', 'S37', 'S38',
      'S39', 'S40', 'S41', 'S42', 'S43', 'S44', 'S45', 'S46', 'S47',
      'S48', 'S49', 'S50', 'S51', 'S52', 'S53', 'S54', 'S55', 'S56',
      'S57', 'S58', 'S59', 'S60', 'S61', 'S62', 'S63', 'S64', 'S65',
      'S66', 'S67', 'S68', 'S69', 'S70', 'S71', 'S72', 'S73', 'S74',
      'S75', 'S76', 'S77', 'S78', 'S79', 'S80', 'S81', 'S82', 'S83',
      'S84', 'S85', 'S86', 'S87', 'S88', 'S89', 'S90', 'S91', 'S92',
      'S93', 'S94', 'S95', 'S96', 'S97', 'S98', 'S99', 'S100'],
      dtype='<U4')
```

In [15]:

```
1 #Duplicating the values
2 n=np.repeat(n,3)
```

In [16]:

```
1 #Viewing the updated values present in N
2 n
```

Out[16]:

```
array(['S1', 'S1', 'S1', 'S2', 'S2', 'S2', 'S3', 'S3', 'S3', 'S4', 'S4',
      'S4', 'S5', 'S5', 'S5', 'S6', 'S6', 'S6', 'S7', 'S7', 'S7', 'S8',
      'S8', 'S8', 'S9', 'S9', 'S9', 'S10', 'S10', 'S10', 'S11', 'S11',
      'S11', 'S12', 'S12', 'S12', 'S13', 'S13', 'S13', 'S14', 'S14',
      'S14', 'S15', 'S15', 'S15', 'S16', 'S16', 'S16', 'S17', 'S17',
      'S17', 'S18', 'S18', 'S18', 'S19', 'S19', 'S19', 'S20', 'S20',
      'S20', 'S21', 'S21', 'S21', 'S22', 'S22', 'S22', 'S23', 'S23',
      'S23', 'S24', 'S24', 'S24', 'S25', 'S25', 'S25', 'S26', 'S26',
      'S26', 'S27', 'S27', 'S27', 'S28', 'S28', 'S28', 'S29', 'S29',
      'S29', 'S30', 'S30', 'S30', 'S31', 'S31', 'S31', 'S32', 'S32',
      'S32', 'S33', 'S33', 'S33', 'S34', 'S34', 'S34', 'S35', 'S35',
      'S35', 'S36', 'S36', 'S36', 'S37', 'S37', 'S37', 'S38', 'S38',
      'S38', 'S39', 'S39', 'S39', 'S40', 'S40', 'S40', 'S41', 'S41',
      'S41', 'S42', 'S42', 'S42', 'S43', 'S43', 'S43', 'S44', 'S44',
      'S44', 'S45', 'S45', 'S45', 'S46', 'S46', 'S46', 'S47', 'S47',
      'S47', 'S48', 'S48', 'S48', 'S49', 'S49', 'S49', 'S50', 'S50',
      'S50', 'S51', 'S51', 'S51', 'S52', 'S52', 'S52', 'S53', 'S53',
      'S53', 'S54', 'S54', 'S54', 'S55', 'S55', 'S55', 'S56', 'S56',
      'S56', 'S57', 'S57', 'S57', 'S58', 'S58', 'S58', 'S59', 'S59',
      'S59', 'S60', 'S60', 'S60', 'S61', 'S61', 'S61', 'S62', 'S62',
      'S62', 'S63', 'S63', 'S63', 'S64', 'S64', 'S64', 'S65', 'S65',
      'S65', 'S66', 'S66', 'S66', 'S67', 'S67', 'S67', 'S68', 'S68',
      'S68', 'S69', 'S69', 'S69', 'S70', 'S70', 'S70', 'S71', 'S71',
      'S71', 'S72', 'S72', 'S72', 'S73', 'S73', 'S73', 'S74', 'S74',
      'S74', 'S75', 'S75', 'S75', 'S76', 'S76', 'S76', 'S77', 'S77',
      'S77', 'S78', 'S78', 'S78', 'S79', 'S79', 'S79', 'S80', 'S80',
      'S80', 'S81', 'S81', 'S81', 'S82', 'S82', 'S82', 'S83', 'S83',
      'S83', 'S84', 'S84', 'S84', 'S85', 'S85', 'S85', 'S86', 'S86',
      'S86', 'S87', 'S87', 'S87', 'S88', 'S88', 'S88', 'S89', 'S89',
      'S89', 'S90', 'S90', 'S90', 'S91', 'S91', 'S91', 'S92', 'S92',
      'S92', 'S93', 'S93', 'S93', 'S94', 'S94', 'S94', 'S95', 'S95',
      'S95', 'S96', 'S96', 'S96', 'S97', 'S97', 'S97', 'S98', 'S98',
      'S98', 'S99', 'S99', 'S99', 'S100', 'S100', 'S100'], dtype='<U4')
```

In [17]:

```
1 #DataFrame Creation
2 student_details=pd.DataFrame()
```

In [18]:

```
1 #Inserting the name column
2 student_details['Name']=n
```

In [19]:

```
1 #Viewing the dataframe
2 student_details
```

Out[19]:

	Name
0	S1
1	S1
2	S1
3	S2
4	S2
...	...
295	S99
296	S99
297	S100
298	S100
299	S100

300 rows × 1 columns

In [20]:

```
1 # creating a column 'gender'
2 g = np.random.choice(['M','F'],100)
```

In [21]:

```
1 #Viewing the values present in it
2 g
```

Out[21]:

```
array(['F', 'M', 'M', 'M', 'M', 'F', 'M', 'F', 'M', 'F', 'M', 'M', 'M',
       'F', 'M', 'M', 'F', 'M', 'F', 'F', 'M', 'F', 'F', 'M', 'F',
       'F', 'M', 'M', 'M', 'M', 'M', 'M', 'F', 'M', 'F', 'M', 'M', 'F',
       'F', 'F', 'F', 'M', 'F', 'F', 'M', 'F', 'M', 'M', 'M', 'M',
       'F', 'M', 'F', 'F', 'M', 'M', 'F', 'M', 'F', 'F', 'M', 'F', 'F',
       'M', 'F', 'M', 'M', 'M', 'F', 'M', 'M', 'F', 'F', 'M', 'F', 'F',
       'F', 'M', 'F', 'M', 'M', 'M', 'F', 'F', 'M', 'F', 'F', 'F', 'M',
       'M', 'F', 'M', 'F', 'F', 'F', 'F', 'M', 'F'], dtype='<U1')
```

In [22]:

```
1 #Duplicating the Values
2 g=np.repeat(g,3)
```

In [23]:

```
1 #Viewing the updated values present in it
2 g
```

Out[23]:

```
array(['F', 'F', 'F', 'M', 'M', 'M', 'M', 'M', 'M', 'M', 'M', 'M', 'M',
      'M', 'M', 'F', 'F', 'F', 'M', 'M',
      'M', 'F', 'F', 'F', 'M', 'M', 'M', 'M', 'M', 'M', 'M', 'M', 'M',
      'F', 'F', 'F', 'M', 'M', 'M', 'M', 'M', 'M', 'F', 'F', 'F', 'M',
      'M', 'M', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'M', 'M',
      'M', 'F', 'F', 'F', 'F', 'F', 'F', 'M', 'M', 'M', 'F', 'F', 'F',
      'F', 'F', 'F', 'M', 'M', 'M', 'M', 'M', 'M', 'M', 'M', 'M', 'M',
      'M', 'M', 'M', 'M', 'M', 'M', 'M', 'M', 'F', 'F', 'F', 'M', 'M',
      'M', 'F', 'F', 'F', 'M', 'M', 'M', 'M', 'M', 'M', 'M', 'F', 'F', 'F',
      'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'M',
      'M', 'M', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'M', 'M', 'M', 'F', 'F',
      'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F',
      'M', 'M', 'M', 'F', 'F', 'F', 'M', 'M', 'M', 'M', 'M', 'M', 'F', 'F',
      'F', 'F', 'F', 'F', 'M', 'M', 'M', 'F', 'F', 'F', 'M', 'M', 'M',
      'M', 'M', 'M', 'M', 'M', 'F', 'F', 'F', 'F', 'F', 'F', 'M', 'M',
      'M', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'M', 'M', 'M',
      'M', 'M', 'M', 'F', 'F', 'F', 'M', 'M', 'M', 'F', 'F', 'F', 'F',
      'F', 'F', 'F', 'F', 'F', 'F', 'F', 'F', 'M', 'M', 'M', 'F', 'F',
      'F'], dtype='<U1')
```

In [24]:

```
1 #Inserting it into the Dataframe
2 student_details['Gender']=g
```

In [25]:

```
1 #Viewing the Dataframe
2 student_details
```

Out[25]:

	Name	Gender
0	S1	F
1	S1	F
2	S1	F
3	S2	M
4	S2	M
...
295	S99	M
296	S99	M
297	S100	F
298	S100	F
299	S100	F

300 rows × 2 columns

In [32]:

```
1 #Creating a column 'Subject'
2 s=np.array(['Python','C','Java'] for i in range(100))
```

In [33]:

```
1 #Viewing the values present in it  
2 s
```

Out[33]:

[illegible]

In [35]

```
1 #Flattening it to a 1d array
2 s.shape = (300,)
```

[illegible]

In [38]:

```
1 #Viewing the Dataframe
2 student_details
```

Out[38]:

	Name	Gender	Subject
0	S1	F	Python
1	S1	F	C
2	S1	F	Java
3	S2	M	Python
4	S2	M	C
...
295	S99	M	C
296	S99	M	Java
297	S100	F	Python
298	S100	F	C
299	S100	F	Java

300 rows × 3 columns

In [39]:

```
1 # Creating a new Column 'Department'
2 d=np.random.choice(['AIML','AIDA','CYS','MED'],100)
```

In [41]:

```
1 #Duplicating Values
2 d=np.repeat(d,3)
```

In [42]:

```
1 #Inserting into the Dataframe
2 student_details['Department']=d
```

In [43]:

```
1 #Viewing the Dataframe
2 student_details
```

Out[43]:

	Name	Gender	Subject	Department
0	S1	F	Python	MED
1	S1	F	C	MED
2	S1	F	Java	MED
3	S2	M	Python	CYS
4	S2	M	C	CYS
...
295	S99	M	C	AIDA
296	S99	M	Java	AIDA
297	S100	F	Python	AIDA
298	S100	F	C	AIDA
299	S100	F	Java	AIDA

300 rows × 4 columns

In [44]:

```
1 #Creating a new Column 'Marks'
2 m=np.random.randint(1,101,300)
```

In [45]:

```
1 #Inserting into the Dataframe
2 student_details['Marks']=m
```

In [46]:

```
1 #Viewing the Dataframe
2 student_details
```

Out[46]:

	Name	Gender	Subject	Department	Marks
0	S1	F	Python	MED	37
1	S1	F	C	MED	89
2	S1	F	Java	MED	60
3	S2	M	Python	CYS	74
4	S2	M	C	CYS	61
...
295	S99	M	C	AIDA	54
296	S99	M	Java	AIDA	93
297	S100	F	Python	AIDA	65
298	S100	F	C	AIDA	76
299	S100	F	Java	AIDA	29

300 rows × 5 columns

- What is the total for each student
- What is the avg mark for each student
- What is the avg of each dept
- What is the avg marks of python of each dept
- What is the avg mark of java for each gender

Using Groupby functions performing some basic operations

In [57]:

```
1 q1= student_details.groupby(['Name'])['Marks'].sum()
```

In [58]:

```
1 q1
```

Out[58]:

Name

```
S1      186
S10     256
S100    170
S11     270
S12     117
```

...

```
S95     228
S96     132
S97     190
S98     113
S99     186
```

Name: Marks, Length: 100, dtype: int32

In [59]:

```
1 q2= student_details.groupby(['Name'])['Marks'].mean()
```

In [60]:

```
1 q2
```

Out[60]:

Name

```
S1      62.000000
S10     85.333333
S100    56.666667
S11     90.000000
S12     39.000000
```

...

```
S95     76.000000
S96     44.000000
S97     63.333333
S98     37.666667
S99     62.000000
```

Name: Marks, Length: 100, dtype: float64

In [61]:

```
1 q3= student_details.groupby(['Department'])['Marks'].mean()
```

In [62]:

```
1 q3
```

Out[62]:

Department

```
AIDA    47.064103
AIML    55.615385
CYS     46.825397
MED     47.839506
```

Name: Marks, dtype: float64

In [63]:

```
1 q4= student_details.groupby(['Department', 'Subject'])['Marks'].mean()
```

In [64]:

```
1 q4
```

Out[64]:

Department	Subject	
AIDA	C	42.384615
	Java	53.923077
	Python	44.884615
AIML	C	55.538462
	Java	50.884615
	Python	60.423077
CYS	C	51.333333
	Java	43.761905
	Python	45.380952
MED	C	52.185185
	Java	50.777778
	Python	40.555556

Name: Marks, dtype: float64

In [67]:

```
1 q5=student_details.groupby(['Gender', 'Subject'])['Marks'].mean()
```

In [68]:

```
1 q5
```

Out[68]:

Gender	Subject	
F	C	52.34
	Java	50.58
	Python	48.52
M	C	48.32
	Java	49.72
	Python	47.20

Name: Marks, dtype: float64