In [ ]:

```
Chapter2: Data Preprocessing
```

In [3]:

```
#import sklearn
#print (sklearn.__version__)
!pip install numpy
!pip install sklearn
import numpy as np
#from sklearn.preprocessing import Imputer
from sklearn.impute import SimpleImputer
imp = SimpleImputer(missing_values=np.nan,strategy ='mean')
imp.fit([[1,2],[np.nan,3],[7,6]])
SimpleImputer()
X = [[np.nan,2],[6,np.nan],[7,6]]
print(imp.transform(X))
```

```
Requirement already satisfied: numpy in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (1.23.3)
```

```
[notice] A new release of pip available: 22.2.2 -> 22.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
Requirement already satisfied: sklearn in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (0.0)
Requirement already satisfied: scikit-learn in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (from s
klearn) (1.1.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages
(from scikit-learn->sklearn) (3.1.0)
Requirement already satisfied: scipy>=1.3.2 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (from s
cikit-learn->sklearn) (1.9.1)
Requirement already satisfied: joblib>=1.0.0 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (from
scikit-learn->sklearn) (1.2.0)
Requirement already satisfied: numpy>=1.17.3 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (from
scikit-learn->sklearn) (1.23.3)
```

```
[notice] A new release of pip available: 22.2.2 -> 22.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
[[4.         2.        ]
 [6.         3.66666667]
 [7.         6.        ]]
```

Normalisation :

Standardisation:

In [2]:

```
import scipy.sparse as sp
X = sp.csc_matrix([[1,2],[0,-1],[8,4]])

imp = SimpleImputer(missing_values=-1,strategy ='mean')
imp.fit(X)
SimpleImputer(missing_values=-1)
X_test = sp.csc_matrix([[-1,2],[6,-1],[7,6]])
print(imp.transform(X_test).toarray())
```

```
[[3. 2.]
 [6. 3.]
 [7. 6.]]
```

In [3]:

```
!pip install pandas
import pandas as pd

df = pd.DataFrame([["a","x"],[np.nan,"y"],["a",np.nan],["b","y"]],dtype = "category")
imp = SimpleImputer(strategy= "most_frequent")
print(imp.fit_transform(df))
```

```
Requirement already satisfied: pandas in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (1.5.0)
Requirement already satisfied: numpy>=1.21.0 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (from
pandas) (1.23.3)
Requirement already satisfied: pytz>=2020.1 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (from p
andas) (2022.4)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packag
es (from pandas) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (from pytho
n-dateutil>=2.8.1->pandas) (1.16.0)
[['a' 'x']
 ['a' 'y']
 ['a' 'y']
 ['b' 'y']]
```

In [4]:

```
!pip install pandas
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer

df = pd.DataFrame([["a","x"],[np.nan,"y"],["a",np.nan],["b","y"]],dtype = "category")
imp = SimpleImputer(strategy= "most_frequent")
print(imp.fit_transform(df))
```

```
Requirement already satisfied: pandas in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (1.5.0)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packag
es (from pandas) (2.8.2)
Requirement already satisfied: numpy>=1.21.0 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (from
pandas) (1.23.3)
Requirement already satisfied: pytz>=2020.1 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (from p
andas) (2022.4)
Requirement already satisfied: six>=1.5 in c:\users\anusha\appdata\local\programs\python\python310\lib\site-packages (from pytho
n-dateutil>=2.8.1->pandas) (1.16.0)
[['a' 'x']
 ['a' 'y']
 ['a' 'y']
 ['b' 'y']]
```

In [7]:

```
from sklearn import preprocessing
X = [[1,-1,2],[2,0,0],[0,1,-1]]
X_normalized = preprocessing.normalize(X,norm="l2")
X_normalized
```

Out[7]:

```
array([[ 0.40824829, -0.40824829,  0.81649658],
       [ 1.        ,  0.        ,  0.        ],
       [ 0.        ,  0.70710678, -0.70710678]])
```

In [4]:

```
import pandas as pd
dataset = pd.read_csv('data.csv')
dataset
```

Out[4]:

	0	1	23	10000
0	1	2	34.0	32000.0
1	2	3	45.0	46000.0
2	3	4	22.0	NaN
3	4	5	25.0	12000.0
4	5	6	34.0	30000.0
5	6	7	30.0	30000.0
6	7	8	NaN	25000.0
7	8	9	42.0	42000.0
8	9	10	41.0	41000.0

In [ ]:

```
#similarly this can be done.
#import pandas as pd
#dataset = pd.read_csv('Book1.csv')
#dataset
```