19. Train Test Split in Dataset

1. splitting the data

- The data is split into train and test in supervised learning
- there is no need to split the data into train and test in unsupervised learning

2. depedent and independent variables

- separte the data according to dependent and independent variables (i.e. convert the data into input and output)

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

In [2]:
```
dataset = pd.read_csv("boston.csv")
dataset.head(3)
```

Out[2]:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	

Separate the data into input and output

In [10]:
```
# dataset.iloc [number of rows:number of columns]
input_data = dataset.iloc[:,:-1]
input_data.head(3)
```

Out[10]:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03

In [18]:
```
dataset.shape
```

Out[18]:
```
(506, 14)
```

In [11]:
```
output_data = dataset['medv']
output_data.head(3)
```

Out[11]:
```
0    24.0
1    21.6
2    34.7
Name: medv, dtype: float64
```

Split the data into training and test dataset

In [14]:
```
from sklearn.model_selection import train_test_split
```

this will split data into 4 parts:

1. input training data, x_train
2. input test data, x_test
3. output training data, y_train
4. output test data, y_test

In [16]:
```
x_train, x_test, y_train, y_test = train_test_split(input_data, output_data, test_s
```

In [17]:
```
x_test
```

Out[17]:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lst
30	1.13081	0.0	8.14	0	0.5380	5.713	94.1	4.2330	4	307	21.0	360.17	22.6
377	9.82349	0.0	18.10	0	0.6710	6.794	98.8	1.3580	24	666	20.2	396.90	21.2
79	0.08387	0.0	12.83	0	0.4370	5.874	36.6	4.5026	5	398	18.7	396.06	9.1
321	0.18159	0.0	7.38	0	0.4930	6.376	54.3	4.5404	5	287	19.6	396.90	6.8
204	0.02009	95.0	2.68	0	0.4161	8.034	31.9	5.1180	4	224	14.7	390.55	2.8
...	...	...	...	...	...	...	...	...	...	...	...	...	
12	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.7
192	0.08664	45.0	3.44	0	0.4370	7.178	26.3	6.4798	5	398	15.2	390.49	2.8
288	0.04590	52.5	5.32	0	0.4050	6.315	45.6	7.3172	6	293	16.6	396.90	7.6
4	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.3
441	9.72418	0.0	18.10	0	0.7400	6.406	97.2	2.0651	24	666	20.2	385.96	19.5

127 rows × 13 columns

In [23]:
```
dataset.shape
```

Out[23]:
```
((506, 14), (379,))
```

In [24]:
```
x_train.shape, y_train.shape
```

Out[24]:
```
((379, 13), (379,))
```

In [25]:
```
x_test.shape, y_train.shape
```

Out[25]:
```
((127, 13), (379,))
```

In [ ]:
```

```