Feature scaling

- **Feature scaling** refers to the methods or techniques used to normalize the range of independent variables in our data, or in other words, the methods to set the feature value range within a similar scale.
- Variables with bigger magnitude / larger value range dominate over those with smaller magnitude / value range
- Scale of the features is an important consideration when building machine learning models.
- Feature scaling is generally the last step in the data preprocessing pipeline, performed **just before training the machine learning algorithms**.
- preserves the shape of the original distribution
- the minimum and maximum values of the different variables may vary
- preserves outliers
- **Feature Scaling importance in some ML Algorithms**

Various Feature Scaling Techniques

- Standardisation
- Normalisation

In [1]:

```
import pandas as pd
```

In [2]:

```
df = pd.DataFrame({"X":[1,2,3,4,5]})
df
```

Out[2]:

	X
0	1
1	2
2	3
3	4
4	5

Standard Scaling / Standardisation

Standardisation involves centering the variable mean at zero, and standardising the variance to 1.

z = (x - x_mean) / std

standardisation:

- centers the mean at 0
- scales the variance at 1

Standardization using standard Scaler in Sklearn

In [3]:

```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
df['X_sc_sk']=sc.fit_transform(df[["X"]])
df
```

Out[3]:

	X	X_sc_sk
0	1	-1.414214
1	2	-0.707107
2	3	0.000000
3	4	0.707107
4	5	1.414214

Standardization using Pandas

In [4]:

```
df["X_sc_p"] = (df["X"]-df["X"].mean())/df["X"].std(ddof=0)
df
```

Out[4]:

	X	X_sc_sk	X_sc_p
0	1	-1.414214	-1.414214
1	2	-0.707107	-0.707107
2	3	0.000000	0.000000
3	4	0.707107	0.707107
4	5	1.414214	1.414214

Min Max Scaling

Min Max Scaling scales the values between 0 to 1. X_scaled = (X - X.min / (X.max - X.min)

In [5]:

```
df = pd.DataFrame({"X":[1,2,3,4,5]})
df
```

Out[5]:

	X
0	1
1	2
2	3
3	4
4	5

MinMax Scaling using MinMaxScaler in Sklearn

In [6]:

```
from sklearn.preprocessing import MinMaxScaler
min_max=MinMaxScaler()
df['X_mm_sk']=min_max.fit_transform(df[["X"]])
df
```

Out[6]:

	X	X_mm_sk
0	1	0.00
1	2	0.25
2	3	0.50
3	4	0.75
4	5	1.00

MinMax Scaling using Pandas

In [7]:

```
df["X_mm_p"] = (df["X"]-df["X"].min()) / (df["X"].max()-df["X"].min())
df
```

Out[7]:

	X	X_mm_sk	X_mm_p
0	1	0.00	0.00
1	2	0.25	0.25
2	3	0.50	0.50
3	4	0.75	0.75
4	5	1.00	1.00