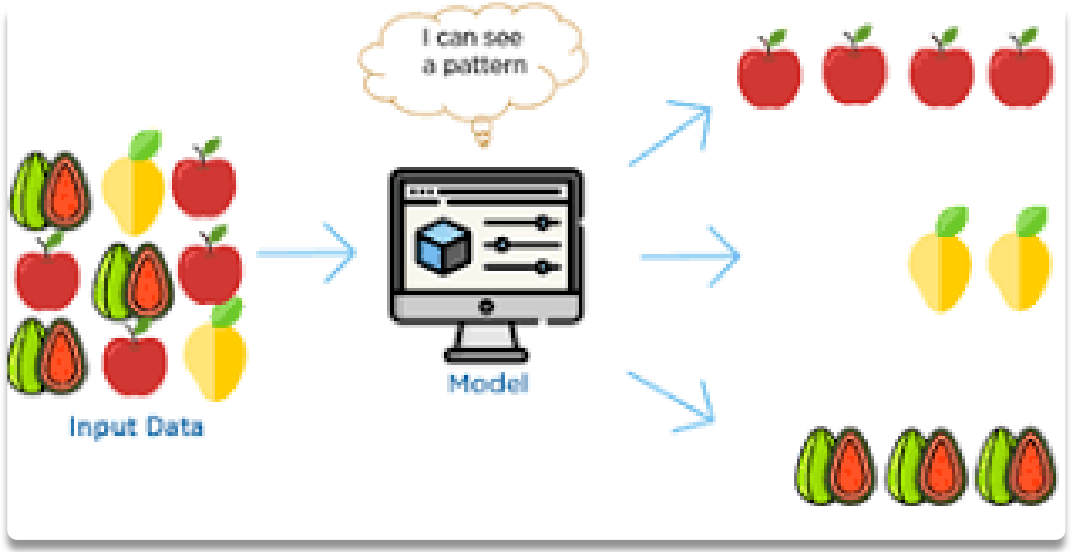
Unsupervised Learning

BY MG ANALYTICS

Finding Patterns in Data

MG ANALYTICS

- SL needed output or target values to be provided as it wanted to predict a specific value or label. Like : Cat, dog, price etc.
- UL requires only input values.
- does not want to predict labels.
- Wants to group items.

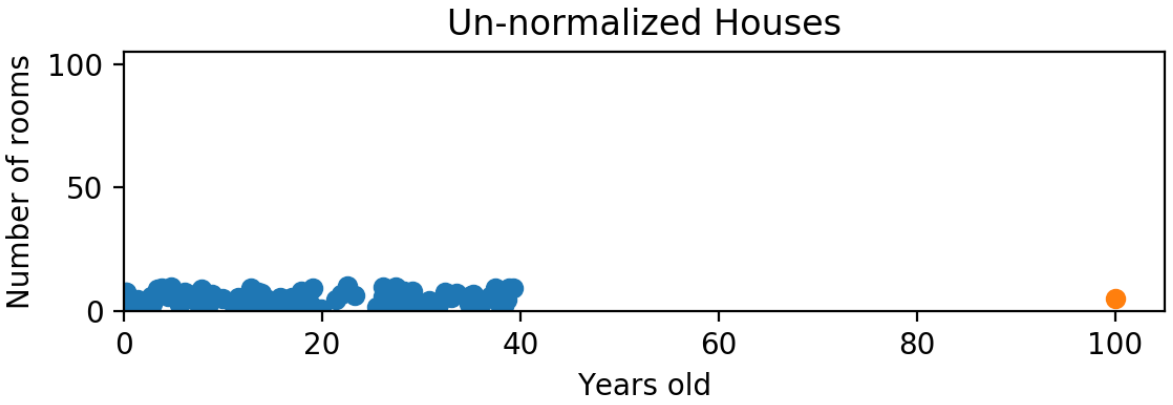


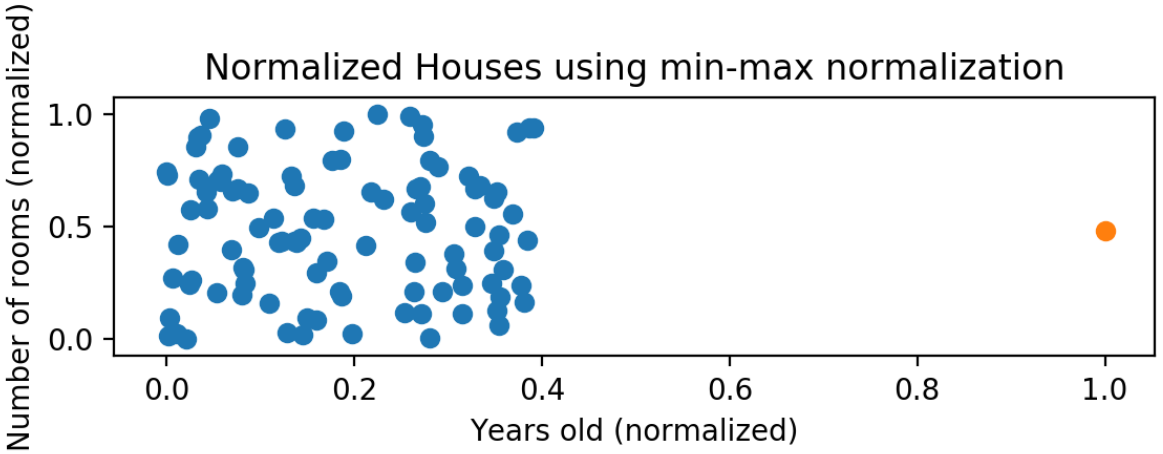
why?

- Create more focused marketing campaigns.
- Find clusters of data patterns
- Group similar items
- Product categorization
- Anomaly detection
- Dimensionality Reduction

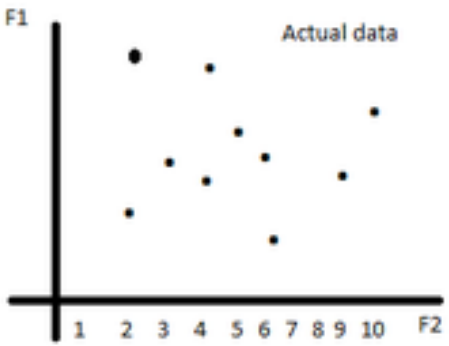
Scaling Data

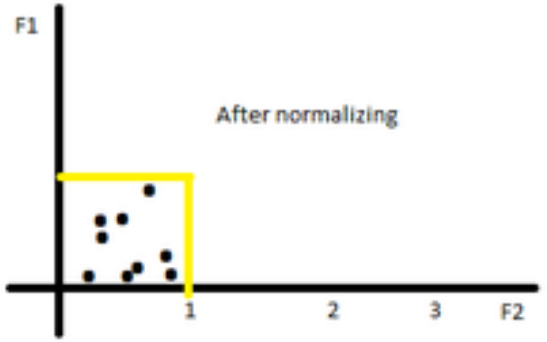
- If the data is **normally/uniformly** distributed, then **Standardization** is the suitable method.
- if the data is **not normally** distributed, we go with **Normalization** scaling method.

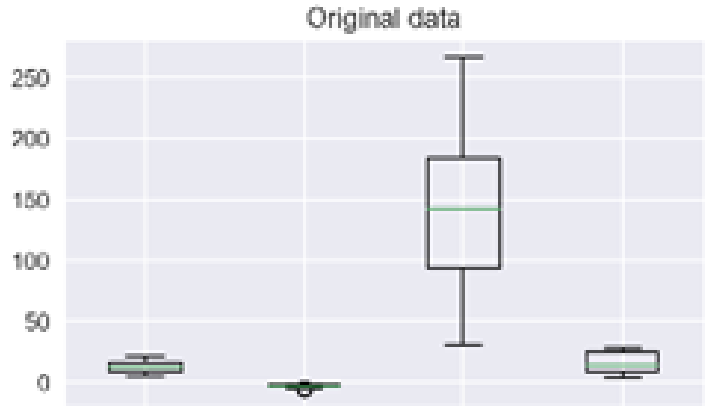


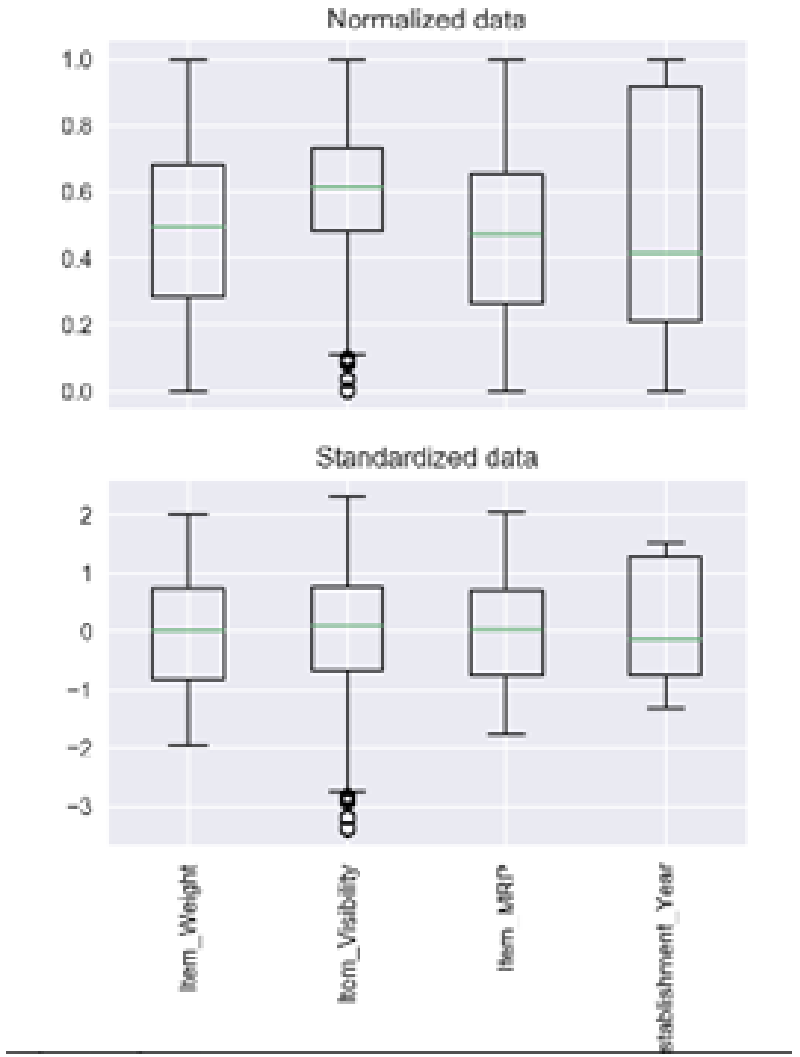


- Normalization brings data between 0- 1
- Standardization brings data between 1 standardization
- Normalization = (X – Xmin) / (Xmax – Xmin)
- Standardization = (x - μ) / σ



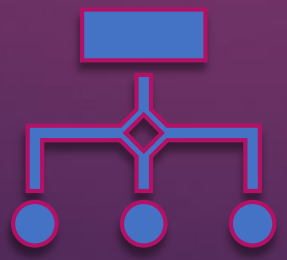



Original data
250
200
150
100
50
0


Normalized data
1.0
0.8
0.6
0.4
0.2
0.0
Standardized data
2
1
0
−1
−2
−3
Item_Weight
Item_Visibility
Item_MRP
Establishment_Year

Clustering

- it's a process to create groups based on similarity measure.
- principle of maximization of intracluster similarities and minimization of intercluster similarities.



A set of data points

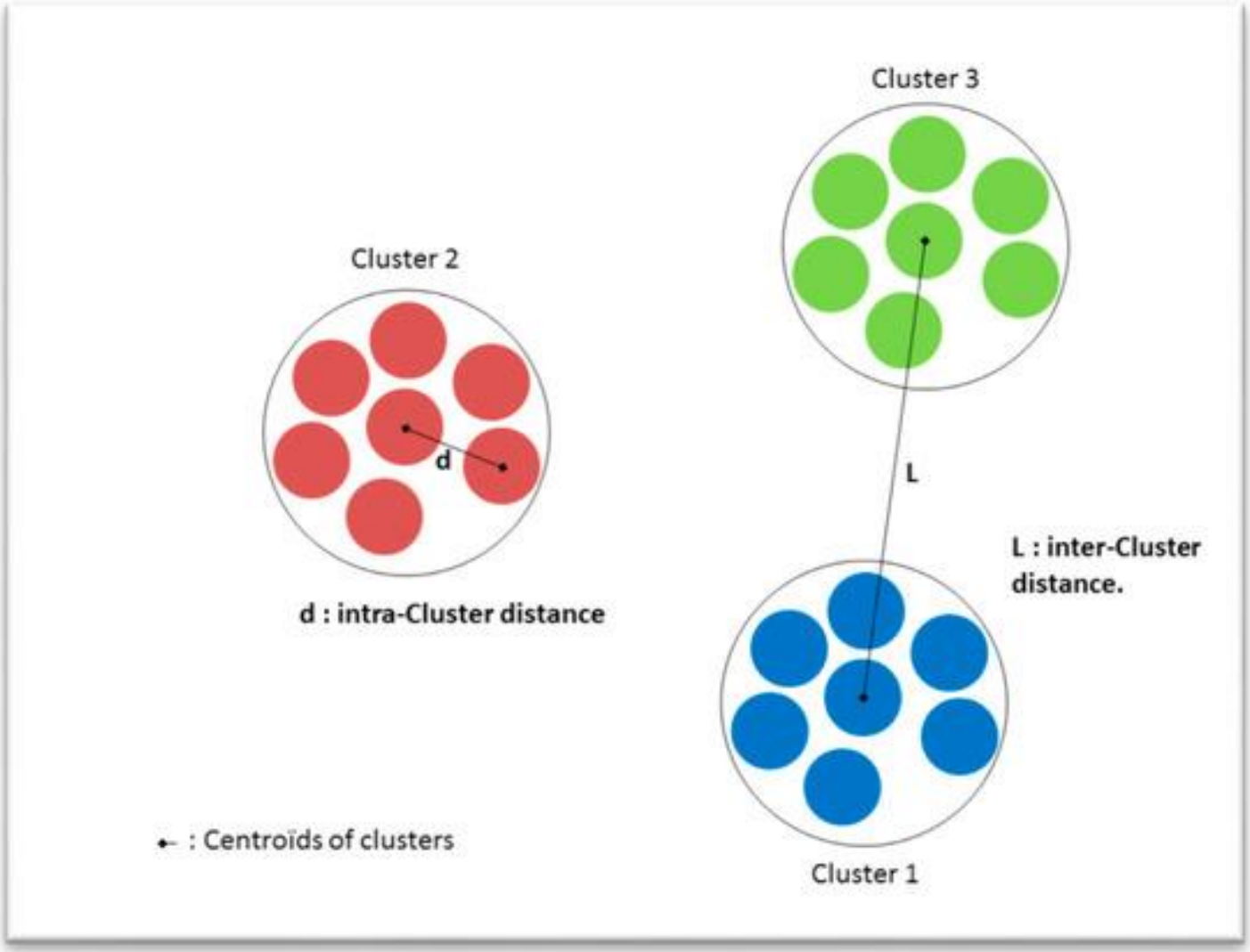
A clustering with Two Clusters

A clustering with Four Clusters

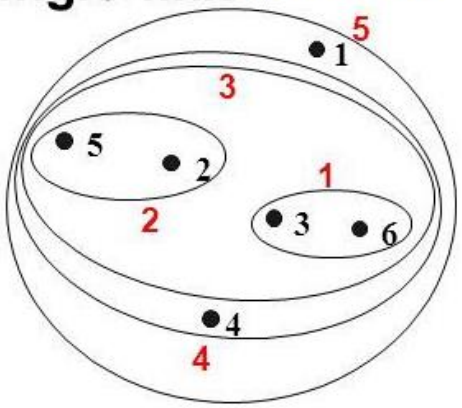
A clustering with Six Clusters

Grouping Data

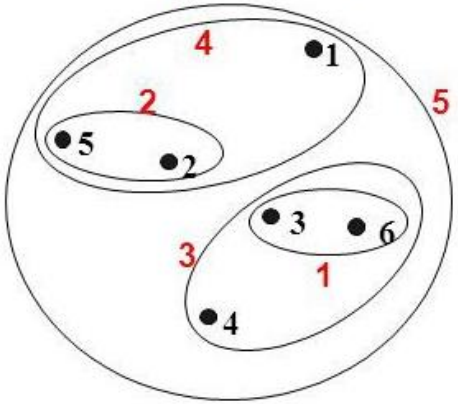
DATA IS GROUPED BASED ON DISTANCE BETWEEN DATA POINTS IN FEATURE SPACE


Cluster 3
Cluster 2
d
L
L : inter-Cluster distance.
d : intra-Cluster distance
: Centroids of clusters
Cluster 1

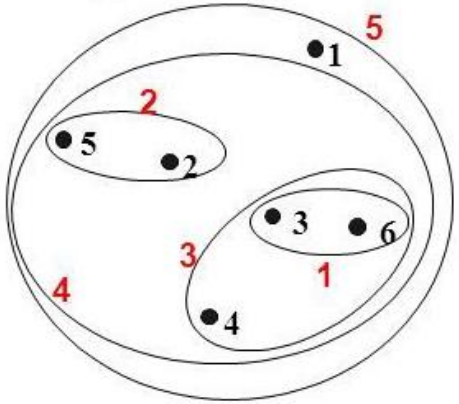
Single-link *(Closest)*



Complete-link *(Farthest)*



Average-link



(Avg of all pairs)

Centroid distance