Clustering

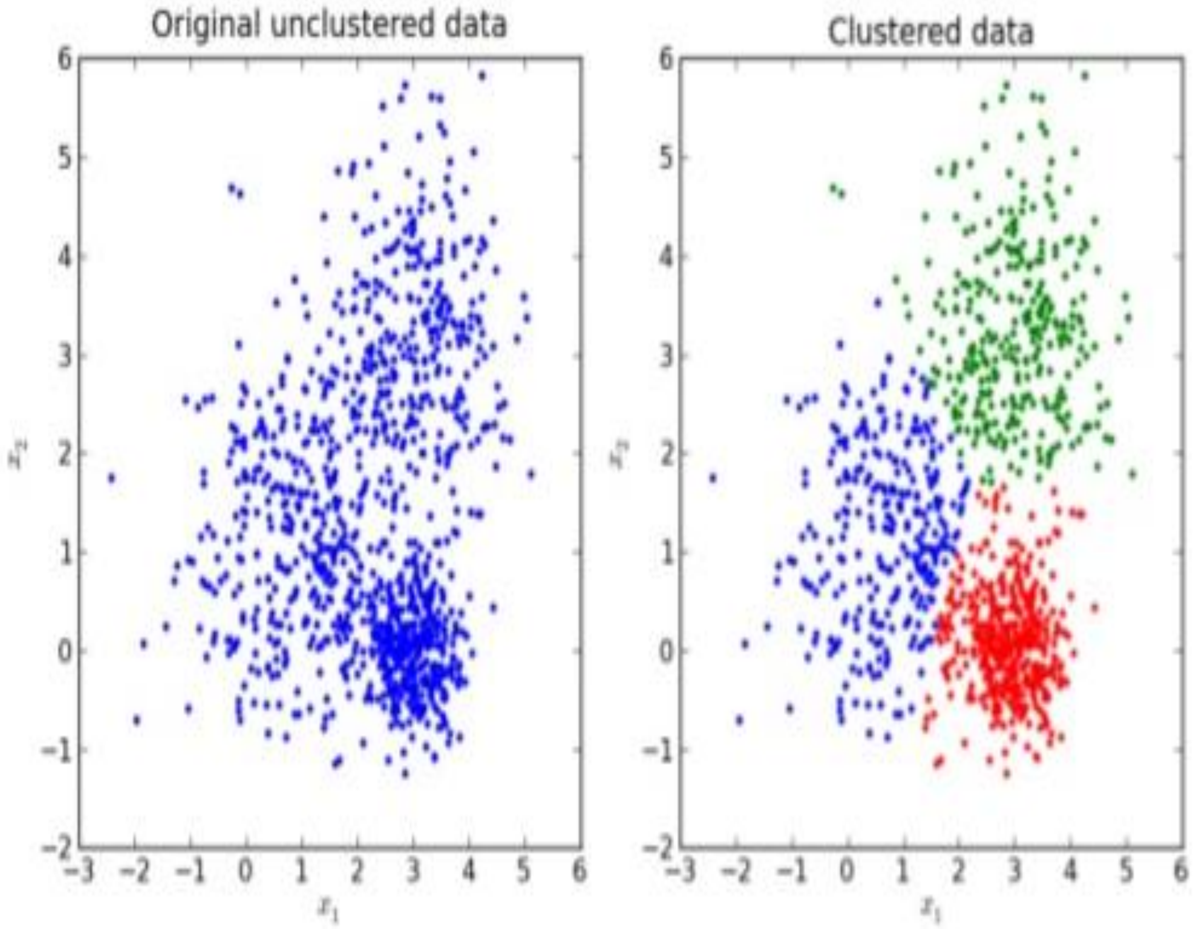
Understanding Cluster Analysis

Clustering

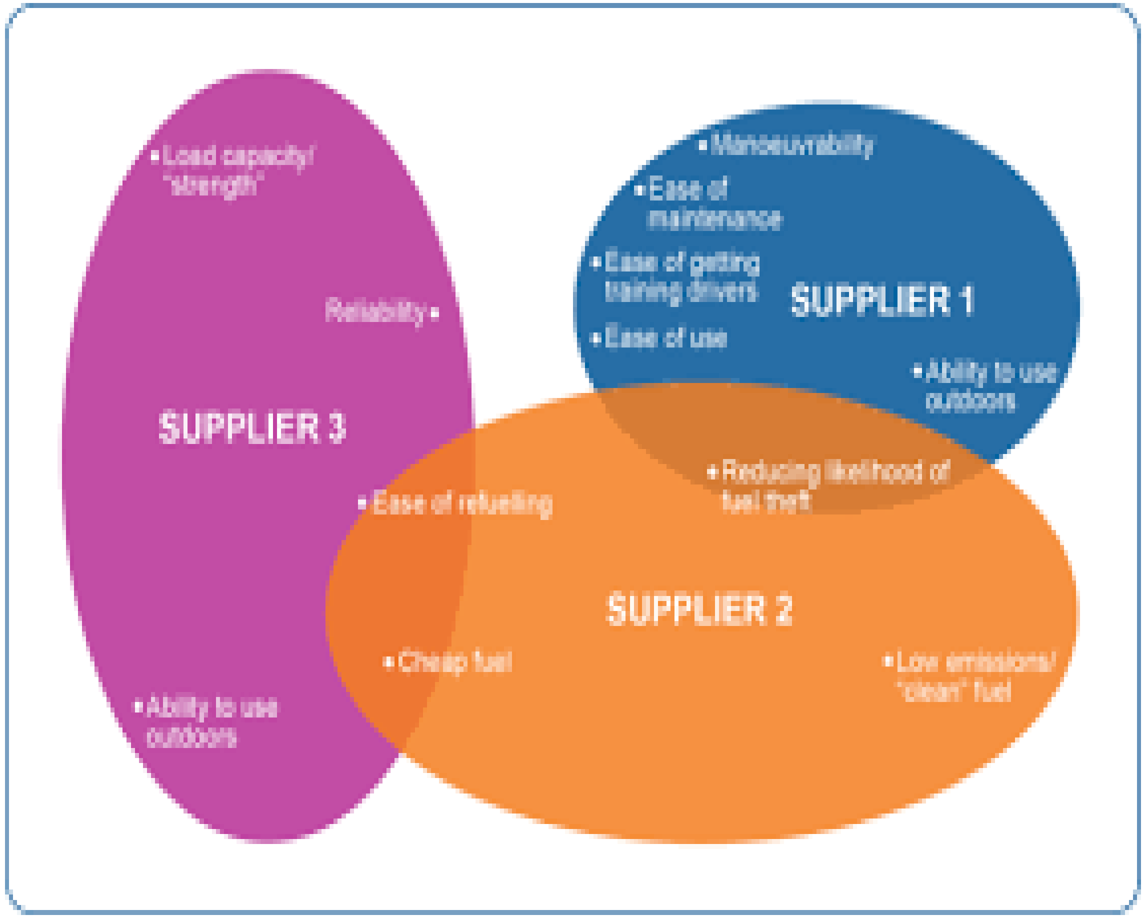
- What is clustering?
- When to use cluster analysis?
- Application of cluster analysis
- Types of cluster analysis
- K means (In detail)

What is Cluster Analysis?

It is a descriptive analysis technique which groups objects (respondents, products, firms, variables, etc.) so that each object is similar to the other objects in the cluster and different from objects in all the other clusters.



When to use cluster analysis?



➢ The essence of all clustering approaches is the classification of data as suggested by "natural" groupings of the data themselves.

➢ Simply put when you desire the following then use Cluster analysis.

- Taxonomy development(segmentation)
- Data simplification
- Relationship identification

Applications of Cluster Analysis

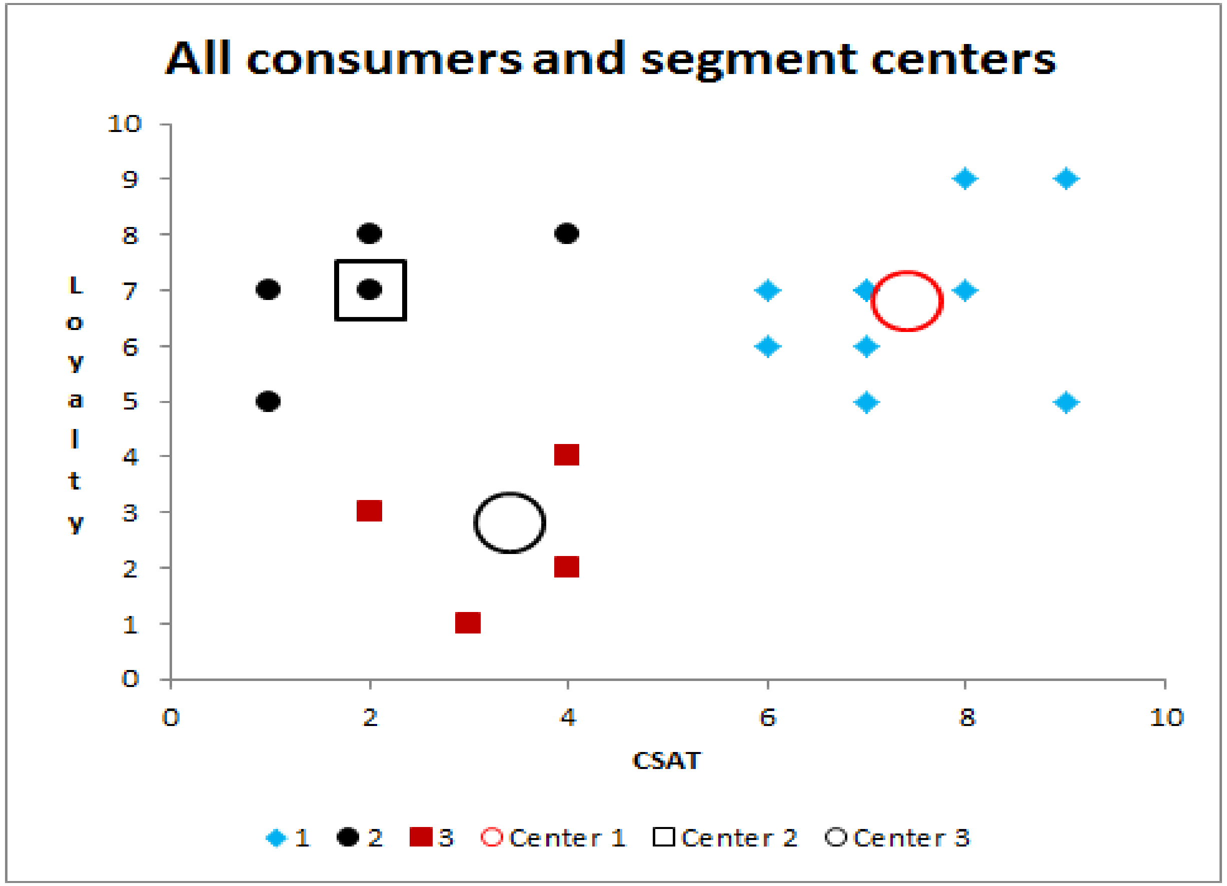
- Understanding

  Group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

- Summarization

  Reduce the size of large data sets

Example



- Cluster analysis is a popular classification technique frequently used to analyse market research data which divides the data into groups.

- Data appears in rows and columns. Rows can then be clustered with respect to columns or columns with respect to rows.

- For example, clustering techniques can be used to identify demographic or psychographic characteristics of consumers with similar purchasing histories, or to isolate differences between groups of products. Market researchers can then study the individual clusters of consumers or products in more detail in order to maximize results from future marketing strategies

Industry related examples

- Cluster analysis of 3D seismic data for oil and gas exploration.
- The cluster analysis of chemically reactive factors of air environment at industrial territories in the informational system of monitoring.
- Cluster Analysis of Fuel Price History

Types of Clustering

A clustering is a set of clusters.

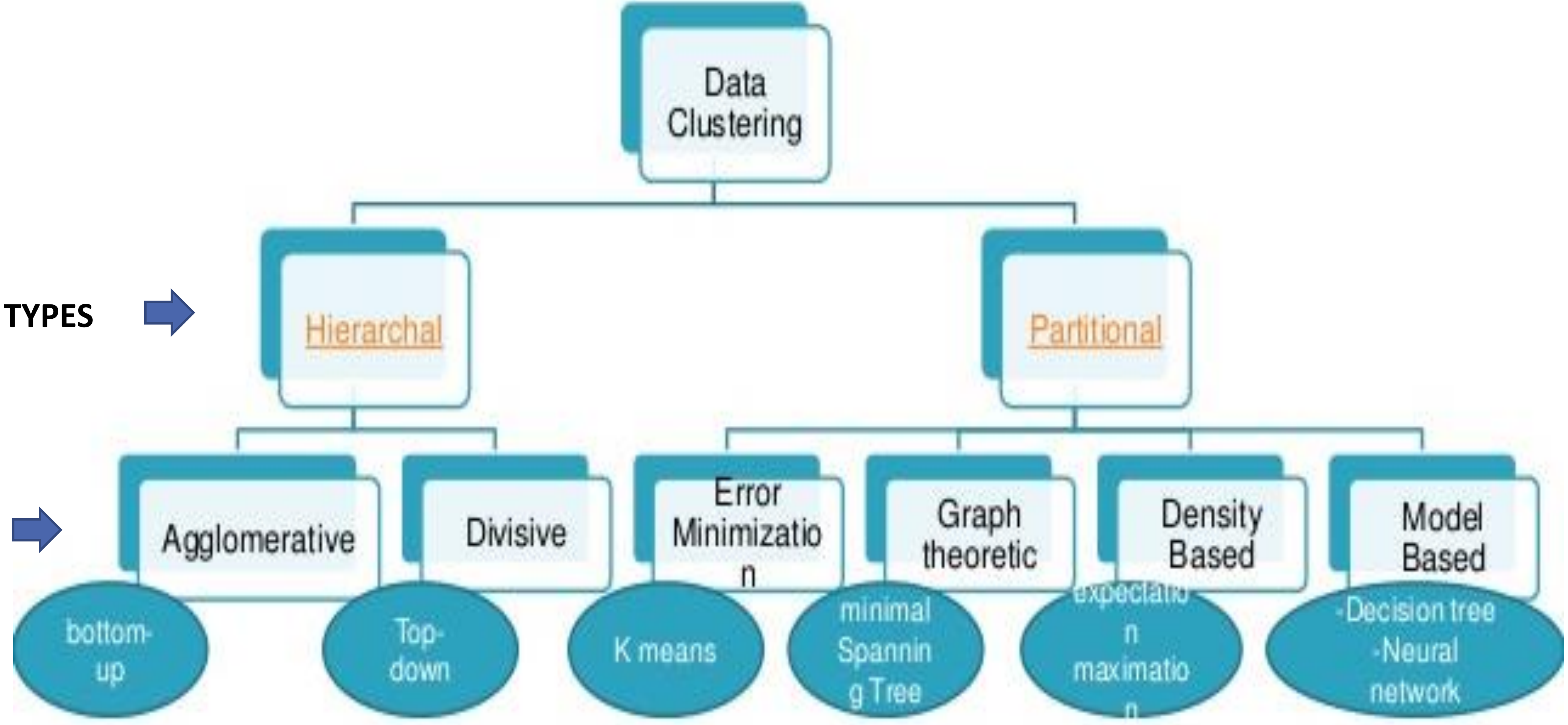
- Hierarchical clustering

A set of nested clusters organized as a hierarchical tree

- Partitional Clustering

A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

Types of clustering



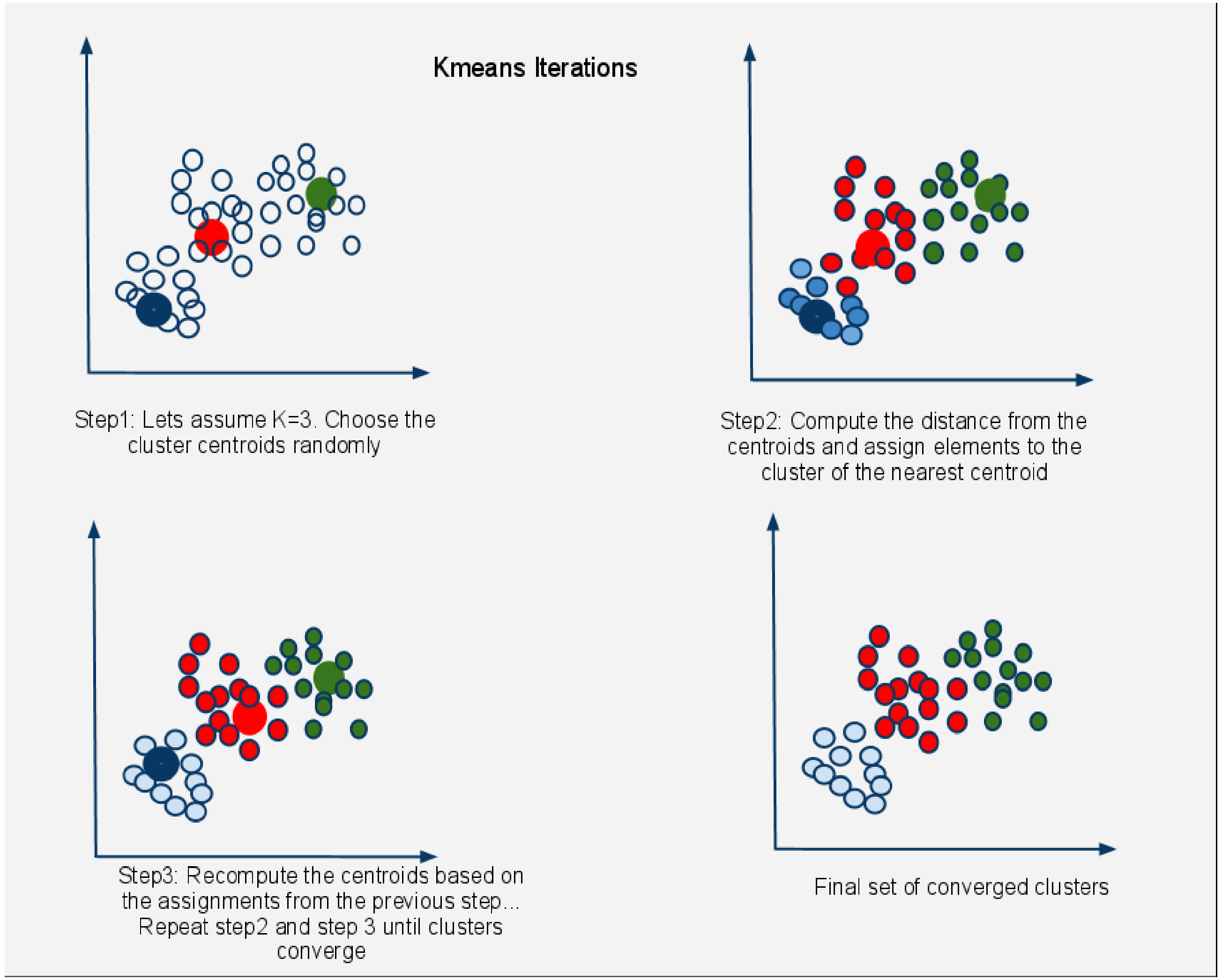
TYPES

METHODS

ALGORITHMS

K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point) Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple



Kmeans Iterations

Step1: Lets assume K=3. Choose the cluster centroids randomly

Step2: Compute the distance from the centroids and assign elements to the cluster of the nearest centroid

Step3: Recompute the centroids based on the assignments from the previous step… Repeat step2 and step 3 until clusters converge

Final set of converged clusters

K-means Clustering – Details

- Initial centroids are often chosen randomly.
- Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- ‘Closeness’ is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above

K-means Clustering – Details

- Most of the convergence happens in the first few iterations.
- Often the stopping condition is changed to ‘Until relatively few points change clusters’
- Complexity is O(n * K * I * d)

  n = number of points,

  K = number of clusters,

  I = number of iterations,

  d = number of attributes

Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE)
- For each point, the error is the distance to the nearest cluster
- To get SSE, we square these errors and sum them.

x is a data point in cluster Ci and mi is the representative point for cluster Ci can show that mi corresponds to the center (mean) of the cluster

Evaluating K-means Clusters

- Given two clusters, we can choose the one with the smaller error
- One easy way to reduce SSE is to increase K, the number of clusters
- A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Energy Meter

❑ An electric meter is a device that measures the amount of electric energy consumed by a residence, a business or electrically powered devices.

❑ Electric utilities install meter at the customer premises to measure electric energy delivered to their customers for billing purpose.

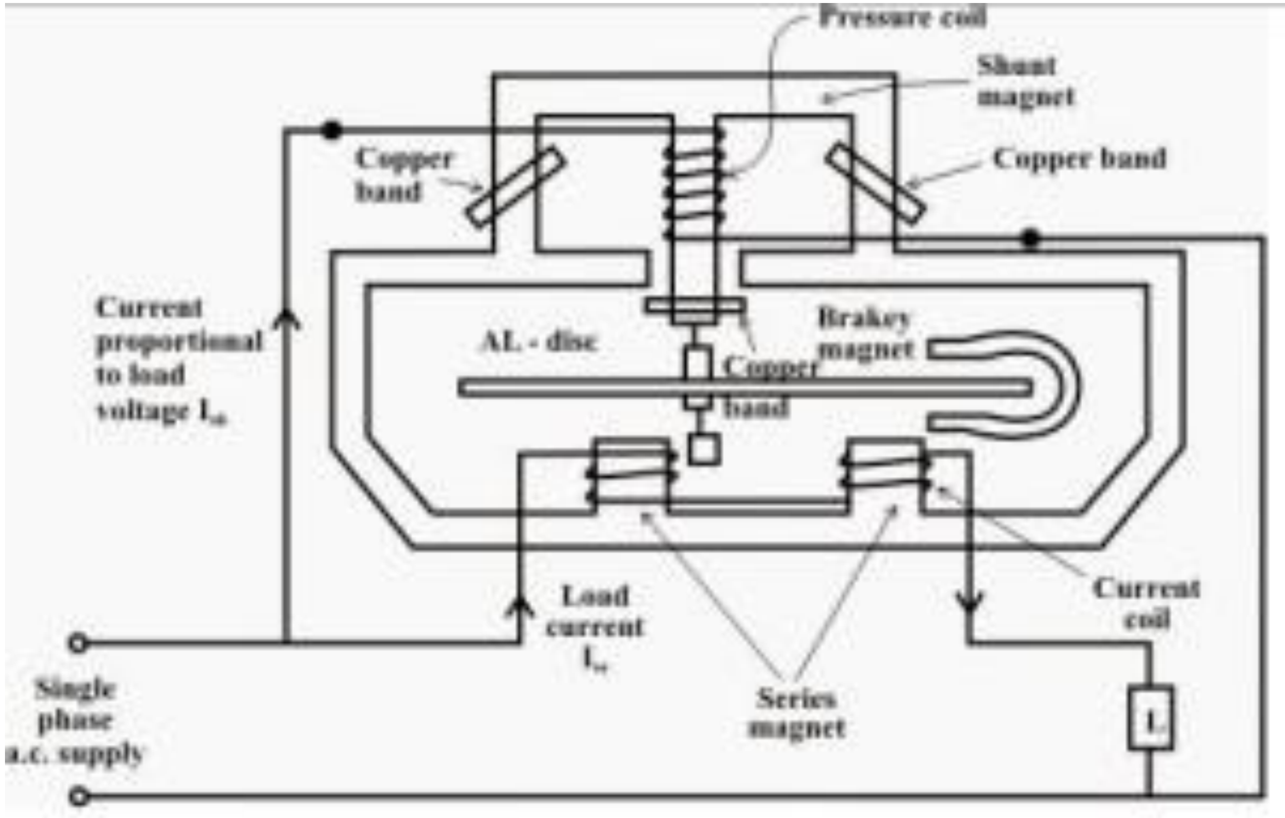
❑ They are calibrated in billing units and the most popular and accepted unit is measured in kilo watt hour (KWH).

Types of energy meter

❑ Is classified in accordance with several factors such as

- Types of display - Analog or digital
- Type of metering point – Grid, primary / secondary transmission and local distribution.
- End applications - Domestic, commercial industrial
- Technical application – 3 phase, single phase, LT, HT and accuracy class meter.

Meter Data


Pressure coil
Shunt magnet
Copper band
Copper band
Current proportional to load voltage I_{sh}
AL - disc
Brakey magnet
Copper band
Load current I_L
Series magnet
Current coil
L
Single phase a.c. supply

Problem statement.

❑ Our objective is to : -

- ➢ To find clusters of high, low, medium consumers on daily basis. (weekdays / weekends)
- ➢ To forecast what is going to be consumption for next 7 days and/or next month.
- ➢ What should be the demand forecast? (Identify the shortage day and provide back up)
- ➢ Identify spike. (outlier detection)
- ➢ Able to satisfy the demand.

Stepwise description of actions to be taken on data

1. Variable identification
   - Variable categorization (e.g. Numeric, Categorical, Discrete, Continuous etc.)
   - Conversion of non-numeric variables to numeric form
   - Creation of *Data Dictionary*
2. Running the K means analysis using R
   - Importing data
   - Selecting the variables
   - Deciding on the number of clusters to be created
   - Running the analysis
   - Interpreting the results

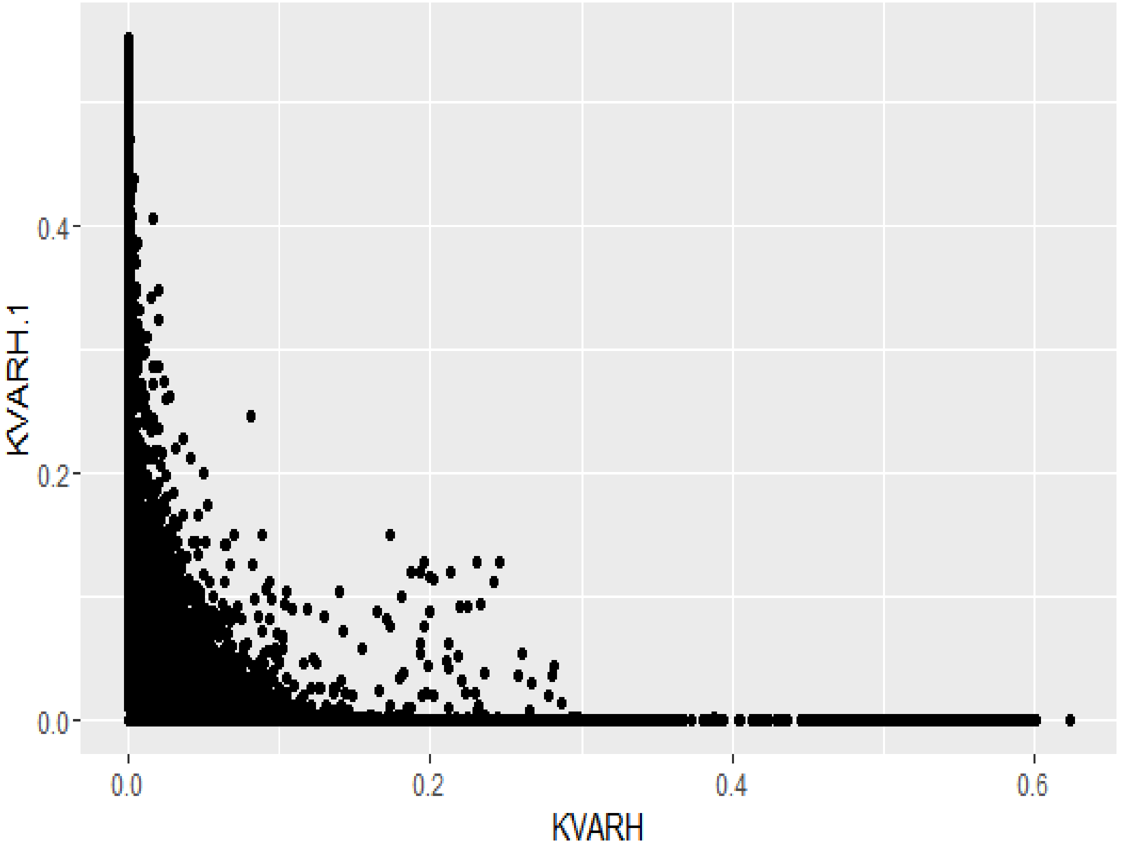
R Script (K means clustering)

```
> setwd("C://R")
> library(amap)
> md=read.csv("md.csv")
> View(md)
> md1=md[,c(5,6,7)]
> View(md1)
> k1=Kmeans(md1,3,iter.max=200,nstart=1,method=c("euclidean"))
> k1$centers
> k1$size
> k1$withinss
> k1$cluster
> cluster_output=k1$cluster
> write.csv(cluster_output,"cluster_output.csv")
> library(ggplot2)
> ggplot(md1, aes(KVARH,KVARH.1,KWH)) + geom_point()
> k1$cluster <- as.factor(k1$cluster)
> ggplot(md1, aes(KVARH,KVARH.1,KVAH, color = k1$cluster)) + geom_point()
```

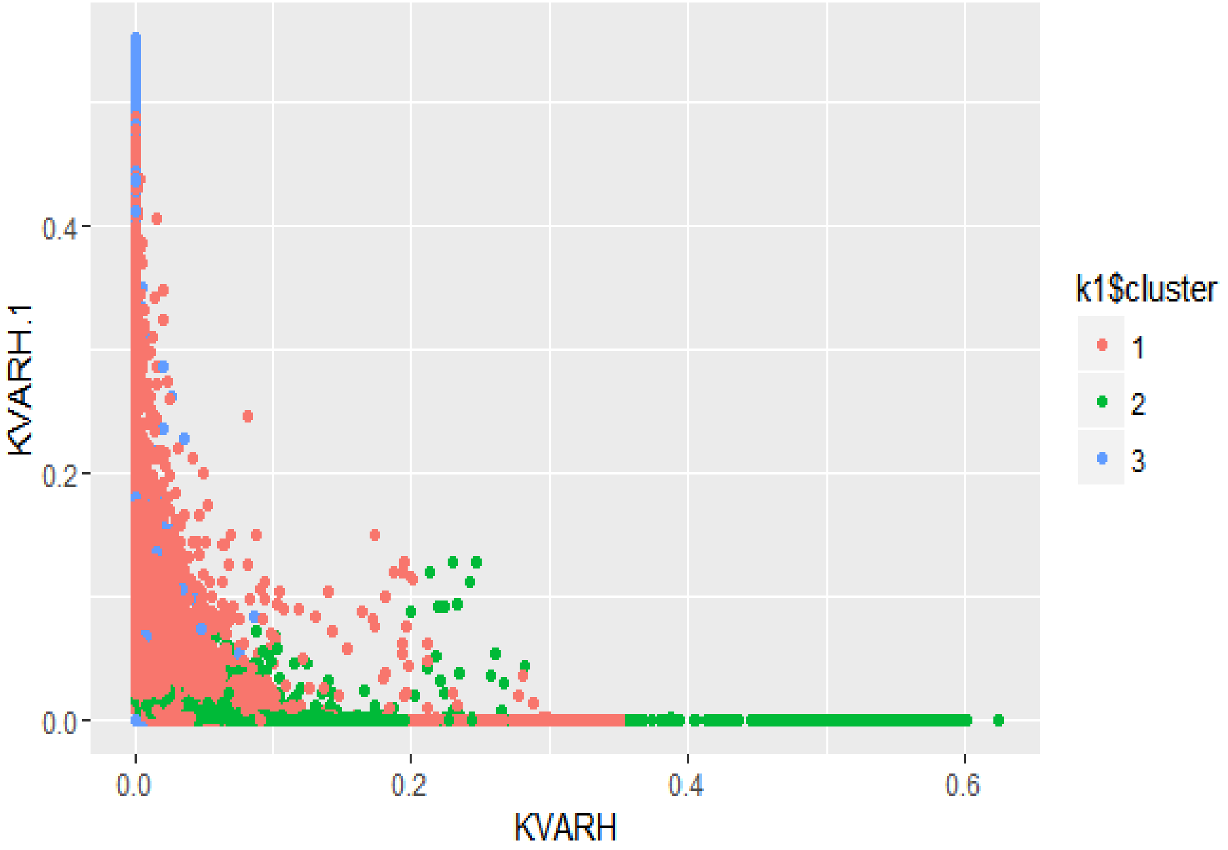
Cluster Graph

Clustering based on individual consumption pattern (Day wise)

Before clustering



After clustering



Decision derived

- Further analysis into the clustered groups gives the result that
  - Cluster 1 – High consumer group
  - Cluster 2 – Low consumer group
  - Cluster 3 – Medium Consumer group
- Based on the Groups we get, we can take decisions on how to treat each of the group in a way that it is helpful to the concerned department. Also studying the trend of each of the cluster and forecasting their respective performance (through Time Series) can give us insights as to how each group may behave in coming time and treat them accordingly.

Design Thinking

Ask the participants what kind of problems related to their industry can be solved using this concept.

Question & Answers

defour analytics

defour analytics

THANK YOU