**Real Estate Market Segmentation / Cluster Analysis**

**What is Clustering?**

It is an **unsupervised statistical learning** technique, where we do not have any dependent variable to predict. Here, the objective is to find clusters or **groups of customers attracted to different *Real Estate properties*** in the data, which the company can use for its decision purposes.

* It can also be a pre-modelling technique.

**Why Clustering / Market Segmentation?**

* Clustering is to **use to find patterns/ groups/ clusters in the data** ( In real Estate it could be the Preferences of students / Elderly / low income group / mid income group / high income group)
* Generally, clustering takes place in a **customer level data**

**What are Clusters?**

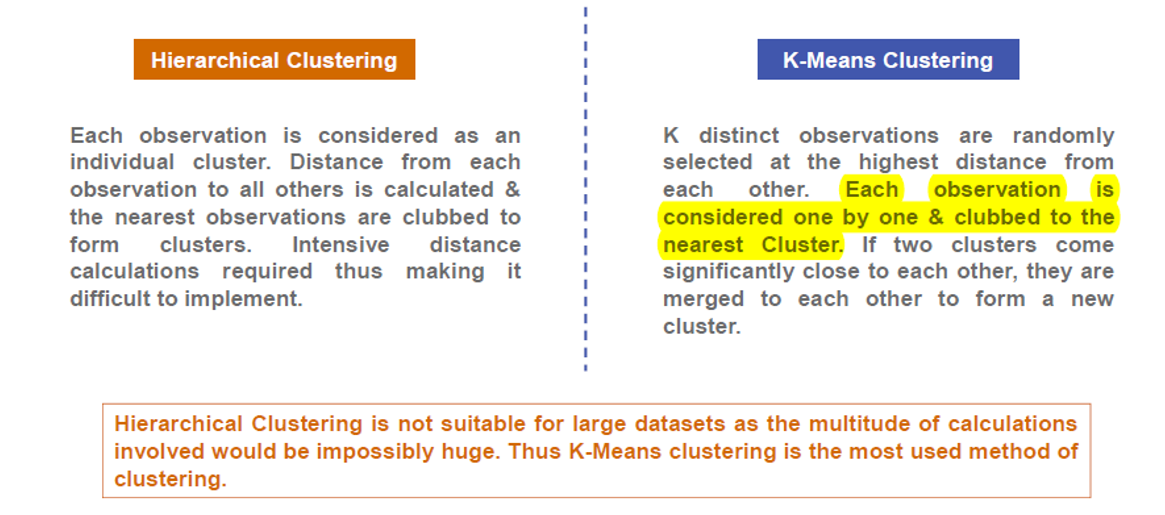
* Clusters are **groups within a population**.
* Most important feature of these groups are: **They are homogenous within themselves but heterogeneous across themselves.**

**When we use Cluster Analysis?**

* If we don’t need to **predict the dependent variable**
* And the objective, is to **find clusters/ groups in the data**
* If we have a dependent variable, and **we need to find separate groups within the data and build separate models, we can use cluster Analysis.**

**Clustering Techniques?**

* 2 types of Clustering approaches
* **Hierarchical Clustering (can’t use on Big data sets, High processing time)**
* **k – Means Clustering (Most commonly used)**
* Standardize all the variables (to make the variables unit free)
* Create the clusters based on the standardized variables



**Results of Cluster Analysis**

* **Cluster Size**- Each cluster how much % of the population size
* For each cluster: what are the **summary measures** of the independent variables
* Mean of the Independent variable in each cluster (ideally different across the different clusters)
* Recommend different **strategies** for each of the clusters.
* Usually, **3- 8** the maximum range of clusters to be present in the data.

**Some Examples of Significant Variables in Real Estate Market Segmentation**

* **INCOME**
* **FAMILY SIZE**
* **GEOGRAPHICAL SITUATIONS**
* **PROFESSION TYPE**
* **AGE**
* **DEBT (LOAN TAKEN)**
* **AVG EDUCATION LEVEL**

**Checks for k-Means Cluster Analysis**

* Individual Variable R-Squared **>= 0.25**
* Overall R-Squared **>= 0.5**
* Approximate Expected Overall R-Squared **> =0.3 (This is R-Squared if there was no Multicollineraity)**
* The Difference Between Overall R-Squared and Approximate Expected Overall R-Squared **Should Not Be Greater Than 0.2**
* RMS Standard Dev **< =1.4**
* Distance Between Cluster Centroids **>= 1.4**
* Number of Clusters Should Be Between **4 and 15**.
* Percentage of Frequency in Each Cluster Should Be **< = 35.**

----------------------------------------------------------------------------------------------------------------------------------------