**TASK:1**

**TABLE OF CONTENT:**

1. **Problem statement**
2. **Clustering as a concept**
3. **Data pre processing**
4. **Model Building**
5. **HTML(Hyper parameter tuning for Machine Learning)**
6. **Data story telling**

**PROBLEM STATEMENT:**

* Apply knowledge of machine learning and clustering to the Global Superstore dataset (a sample dataset containing 50,000 sales records for a global superstore – available to download as .xlsx [here](https://community.tableau.com/servlet/JiveServlet/download/840157-286195/Global Superstore 2018.xlsx" \t "https://mail.google.com/mail/u/1/" \l "inbox/_blank)).

* Investigate and see what interesting insights can be made – e.g. trends, predicting sales/profit, clustering products/locations, etc.

* Bring in data from other sources where useful – e.g. weather/population/demographics, etc.

* Visualise the results in Power BI Desktop using the built-in Python visual.

The goal is to produce visuals in Power BI that would be useful for a Sales executive – whether it is describing the data and finding trends or providing information that can drive business decisions.

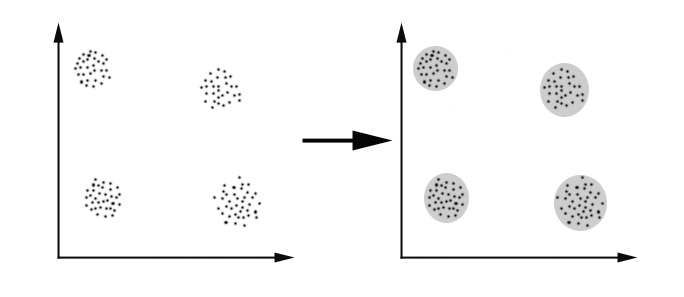
**What is clustering:**

\*Clustering falls under the unsupervised learning curve in machine learning

\*****Intra-cluster minimization:**** The closer the objects in a cluster, the more likely they belong to the same cluster.

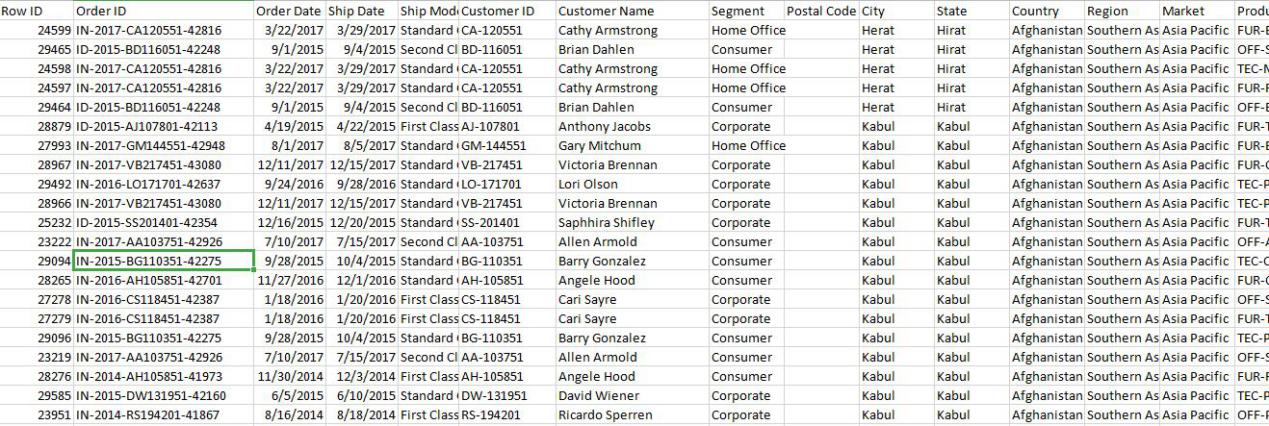
\*****Inter-cluster Maximization:****This makes the separation between two clusters. The main goal is to maximize the distance between 2 clusters.

\*In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features.



**Data pre processing:**

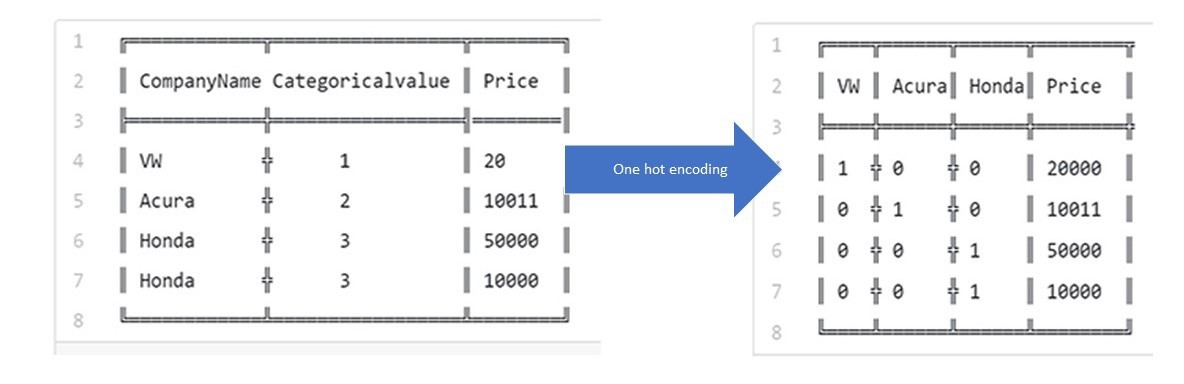
\*Since the data set provided is unlabeled :The customer records should be classified into different clusters based on features

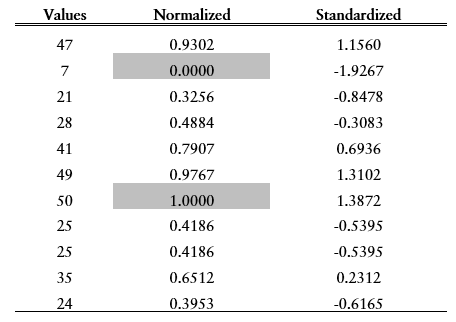


\*It is clearly evident that the data is unlabeled and based on the different features available the records should be classified into different clusters in order to gain business insights and make intelligent business decisions in order to maximize profits.

\*Also it can be noted that the data should extensively be a victim of data pre processing I.e the missing values should be handled,and every feature is categorical hence has to be converted to numerical in order to be fed to a ML model using techniques like one hot encoding or dummy encoding of which former is a better option,after the data has been numerically encoded the data has to undergo feature scaling in order to ensure that none of the dependant variables/features overlap/overshadow on each other and thus are considered more important.

\*That is no feature should be considered more important than other based on numerical value.

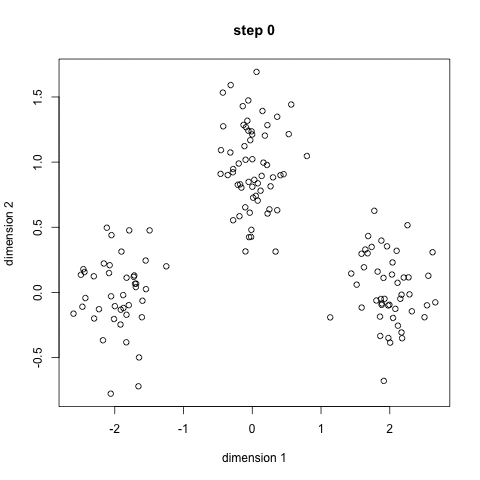




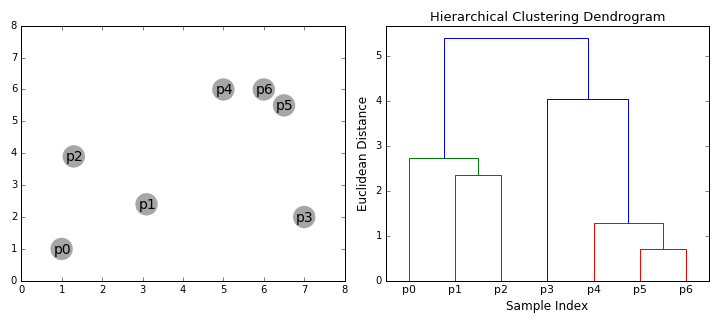
**Model Building :**

Some clustering algorithms that can be used are:

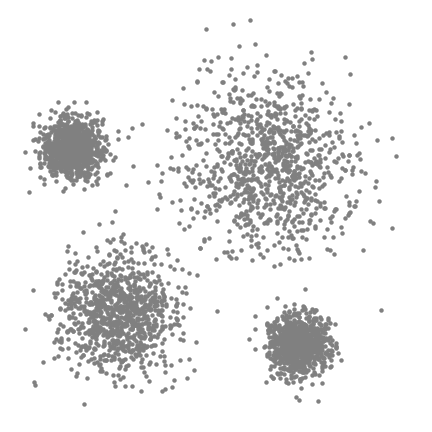
**\*K-Means:**

****

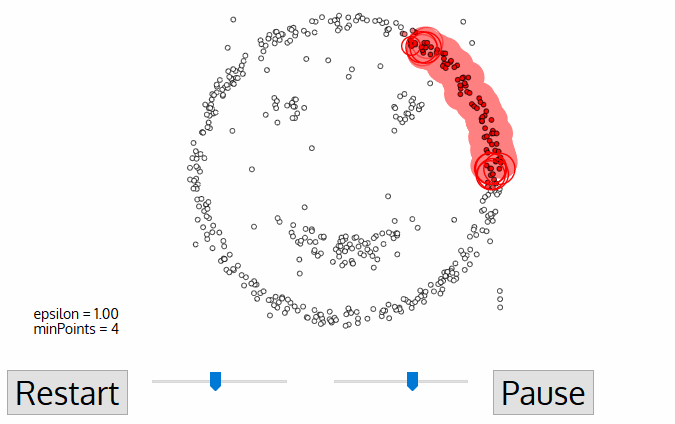
**\*Hierarchical Clustering**



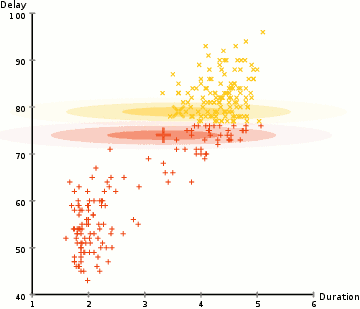
**\*Mean-shift Clustering:**

****

**\*DBSCAN:**

****

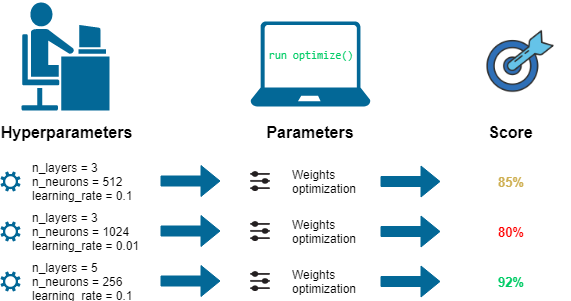
**\*Expectation Maximization with Gaussian Mixed Models**

****

\*A comparative study can be done using all these algorithm to find the one that best fits the use case.

**HTML(Hyper parameter tuning for Machine Learning):**

**\***After the model has been built we can tweak the hyper parameters of the model like K in K-Means do ensure best results are extracted.



**Business Intelligence/Data story telling:**

\*Now since the model has been built best accuracy has been achieved it is time to make use of insights gained to make better business decisions that help the firm climb the success ladder

\*The insights gained have to be narrated in way that can be understood by everyone irrespective of the domain or vertical they come from.

\*A picture speaks a thousand words this can be achieved with visualizations via BI tools like PowerBI

\*The clusters obtained can give the firm an insight on target audience for a specific product

\*Dashboards can be optimized using python code that can be used via a connector



**References:**

<https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>

<https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a>