TASK:1

TABLE OF CONTENT:

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).

- 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.

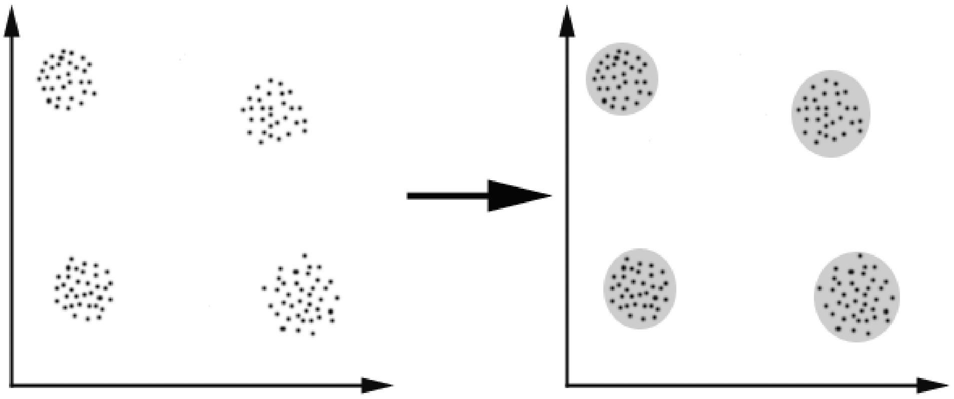
What is clustering:

*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.

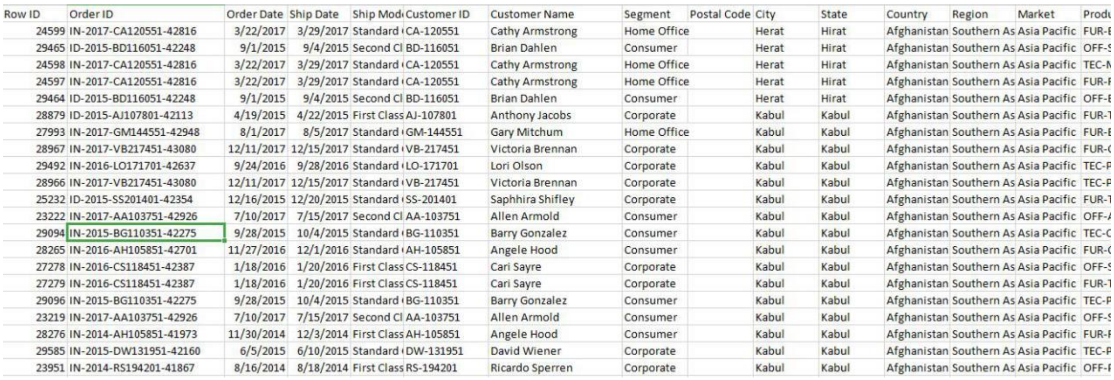
***Inter-cluster Maximization:** This makes the separation between two clusters. The main goal is to maximize the distance between 2 clusters.

*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

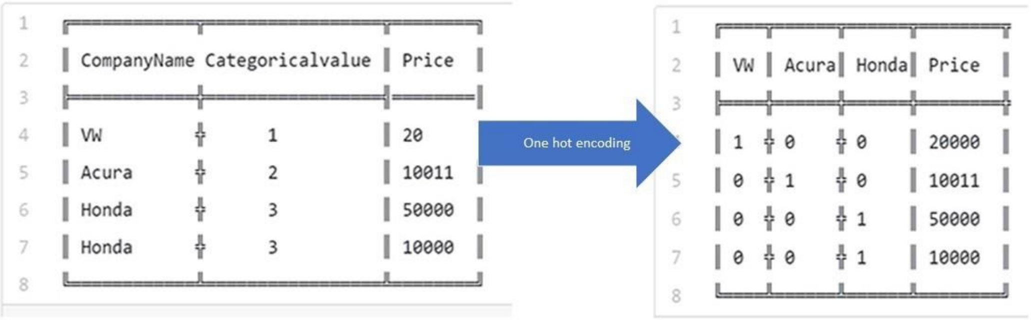


Row ID	Order ID	Order Date	Ship Date	Ship Mod	Customer ID	Customer Name	Segment	Postal Code	City	State	Country	Region	Market	Produ
24599	IN-2017-CA120551-42816	3/22/2017	3/29/2017	Standard	CA-120551	Cathy Armstrong	Home Office		Herat	Hirat	Afghanistan	Southern As	Asia Pacific	FUR-E
29465	ID-2015-BD116051-42248	9/1/2015	9/4/2015	Second Cl	BD-116051	Brian Dahlen	Consumer		Herat	Hirat	Afghanistan	Southern As	Asia Pacific	OFF-S
24598	IN-2017-CA120551-42816	3/22/2017	3/29/2017	Standard	CA-120551	Cathy Armstrong	Home Office		Herat	Hirat	Afghanistan	Southern As	Asia Pacific	TEC-M
24597	IN-2017-CA120551-42816	3/22/2017	3/29/2017	Standard	CA-120551	Cathy Armstrong	Home Office		Herat	Hirat	Afghanistan	Southern As	Asia Pacific	FUR-F
29464	ID-2015-BD116051-42248	9/1/2015	9/4/2015	Second Cl	BD-116051	Brian Dahlen	Consumer		Herat	Hirat	Afghanistan	Southern As	Asia Pacific	OFF-E
28879	ID-2015-AJ107801-42113	4/19/2015	4/22/2015	First Class	AJ-107801	Anthony Jacobs	Corporate		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	FUR-T
27993	IN-2017-GM144551-42948	8/1/2017	8/5/2017	Standard	GM-144551	Gary Mitchum	Home Office		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	FUR-E
28967	IN-2017-VB217451-43080	12/11/2017	12/15/2017	Standard	VB-217451	Victoria Brennan	Corporate		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	FUR-C
29492	IN-2016-LO171701-42637	9/24/2016	9/28/2016	Standard	LO-171701	Lori Olson	Corporate		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	TEC-P
28966	IN-2017-VB217451-43080	12/11/2017	12/15/2017	Standard	VB-217451	Victoria Brennan	Corporate		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	TEC-P
25232	ID-2015-SS201401-42354	12/16/2015	12/20/2015	Standard	SS-201401	Saphhira Shifley	Corporate		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	FUR-T
23222	IN-2017-AA103751-42926	7/10/2017	7/15/2017	Second Cl	AA-103751	Allen Armold	Consumer		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	OFF-A
29094	IN-2015-BG110351-42275	9/28/2015	10/4/2015	Standard	BG-110351	Barry Gonzalez	Consumer		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	TEC-C
28265	IN-2016-AH105851-42701	11/27/2016	12/1/2016	Standard	AH-105851	Angele Hood	Consumer		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	FUR-C
27278	IN-2016-CS118451-42387	1/18/2016	1/20/2016	First Class	CS-118451	Cari Sayre	Corporate		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	OFF-S
27279	IN-2016-CS118451-42387	1/18/2016	1/20/2016	First Class	CS-118451	Cari Sayre	Corporate		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	FUR-T
29096	IN-2015-BG110351-42275	9/28/2015	10/4/2015	Standard	BG-110351	Barry Gonzalez	Consumer		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	TEC-P
23219	IN-2017-AA103751-42926	7/10/2017	7/15/2017	Second Cl	AA-103751	Allen Armold	Consumer		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	OFF-S
28276	IN-2014-AH105851-41973	11/30/2014	12/3/2014	First Class	AH-105851	Angele Hood	Consumer		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	FUR-F
29585	IN-2015-DW131951-42160	6/5/2015	6/10/2015	Standard	DW-131951	David Wiener	Corporate		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	TEC-P
23951	IN-2014-RS194201-41867	8/16/2014	8/18/2014	First Class	RS-194201	Ricardo Sperren	Corporate		Kabul	Kabul	Afghanistan	Southern As	Asia Pacific	OFF-F

*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.

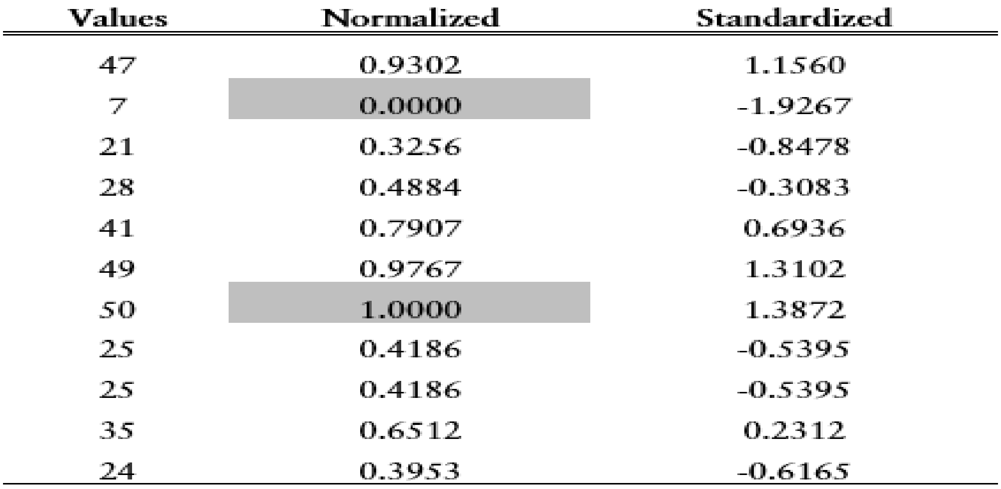
*Also it can be noted that the data should extensively be a victim of data pre processing I.e the missing values should be handled,and every feature is categorical hence has to be converted to numerical in order to be fed to a ML model using techniques like one hot encoding or dummy encoding of which 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.



CompanyName	Categoricalvalue	Price
VW	1	20
Acura	2	10011
Honda	3	50000
Honda	3	10000

VW	Acura	Honda	Price
1	0	0	20000
0	1	0	10011
0	0	1	50000
0	0	1	10000

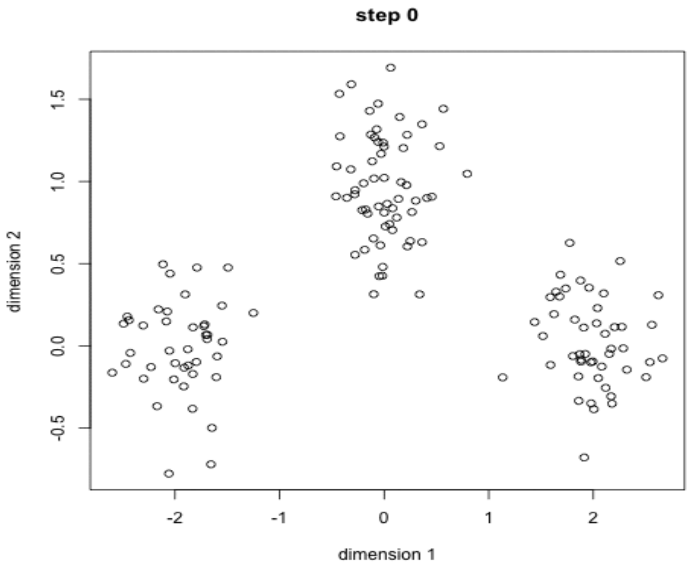


Values	Normalized	Standardized
47	0.9302	1.1560
7	0.0000	-1.9267
21	0.3256	-0.8478
28	0.4884	-0.3083
41	0.7907	0.6936
49	0.9767	1.3102
50	1.0000	1.3872
25	0.4186	-0.5395
25	0.4186	-0.5395
35	0.6512	0.2312
24	0.3953	-0.6165

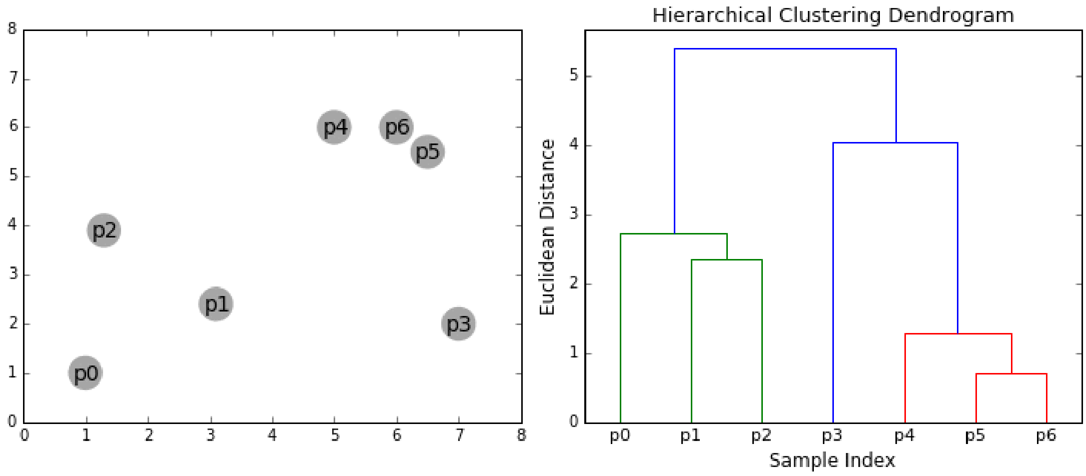
Model Building :

Some clustering algorithms that can be used are:

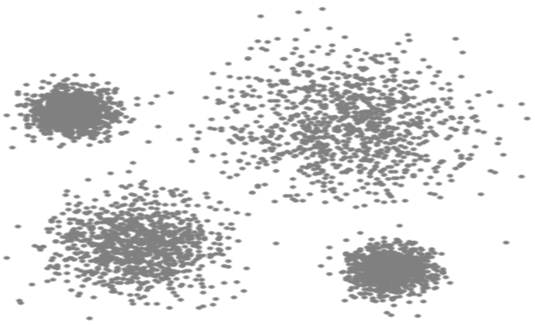
***K-Means:**



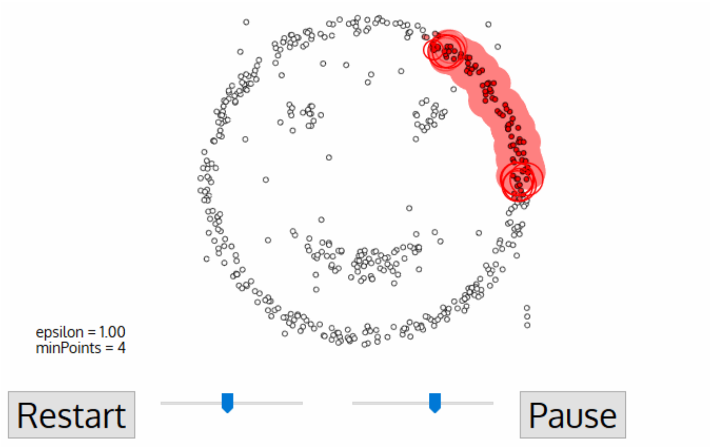
*Hierarchical Clustering



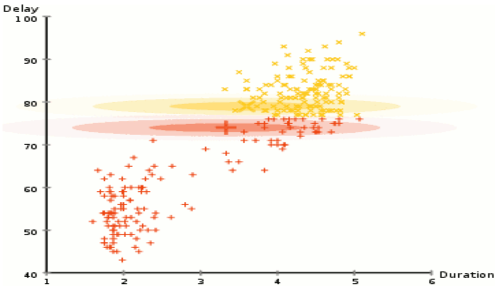
*Mean-shift Clustering:



*DBSCAN:


epsilon = 1.00
minPoints = 4

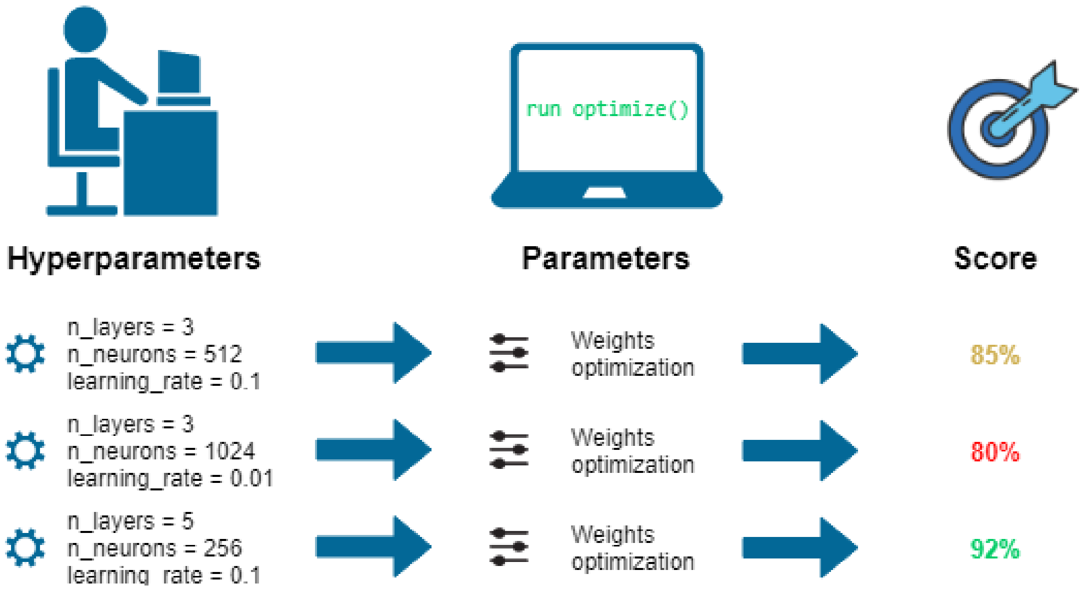
***Expectation Maximization with Gaussian Mixed Models**



*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.



n_layers = 3
n_neurons = 512
learning_rate = 0.1
Weights optimization
85%

n_layers = 3
n_neurons = 1024
learning_rate = 0.01
Weights optimization
80%

n_layers = 5
n_neurons = 256
learning_rate = 0.1
Weights optimization
92%

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