Making Sense of Unstructured Data

WEEK 3

Learning Objective Of This Module

- Introduction to Supervised and Unsupervised learning.
- Clustering and its types.
- Principle Component Analysis

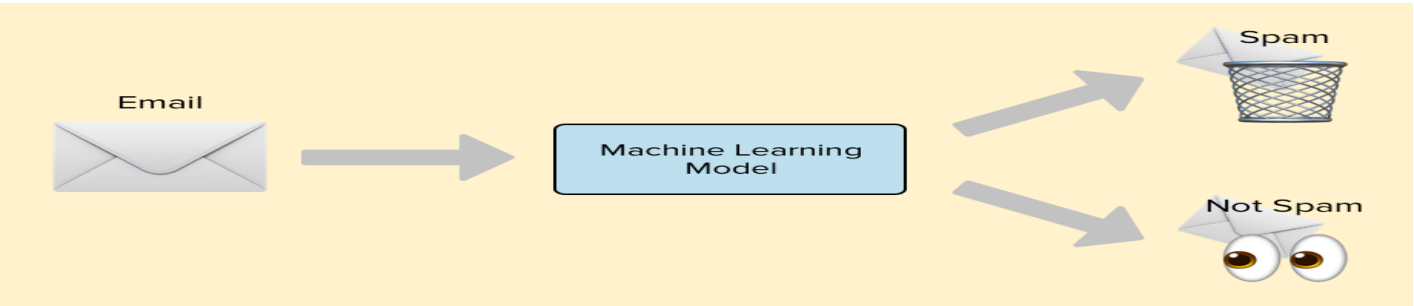
TRY ANSWERING THE FOLLOWING

1. What is the difference between Supervised learning and unsupervised learning?
2. What is clustering and what are the most common clustering techniques?
3. Why and when to use clustering?
4. How clustering helps in finding hidden groupings and patterns?
5. Why K-Means Clustering Is So Popular? What are its assumptions?
6. Why converting features into continuous values and scaling the data is important to perform K-Means clustering?
7. How PCA acts as a pre-processing step in the clustering process?
8. How PCA helps in Dimensionality reduction?

What is the difference between Supervised learning and unsupervised learning?

Supervised learning algorithms are trained using labeled data. **Example** - Predict whether a new email is SPAM or NOT SPAM i.e. predict the label of next email, or Predict whether a customer will churn or not.

Unsupervised learning algorithms are trained using unlabeled data. Example



What is clustering and what are the most common clustering techniques?

Cluster analysis, or **clustering**, is an unsupervised **machine learning** technique. It involves automatically discovering natural grouping in data (groups data according to the notion of similarity).
Most common types of clustering are K-means clustering algorithm, LDA Clustering, Spectral Clustering, Modularity Clustering, Hierarchical Clustering, DBSCAN clustering algorithm etc.

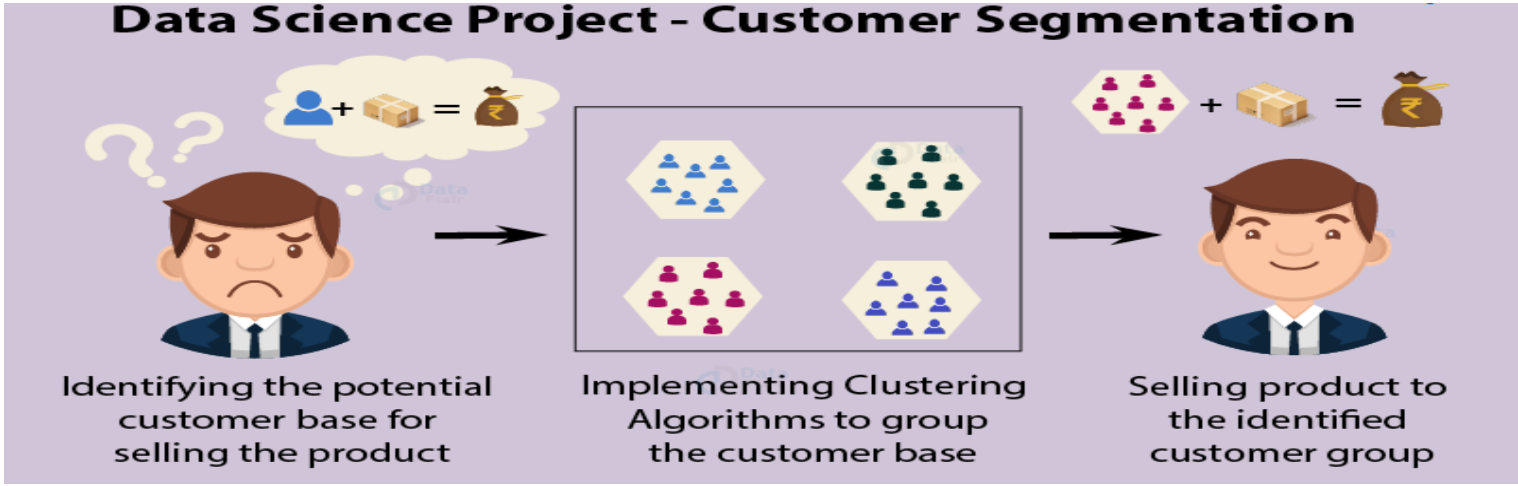
Why and when to use clustering?

WHY-

Retail, finance, and marketing are some of the key domains that benefit from Clustering. Using clustering methods to analyze the data can help them gain further insights. The factors analyzed through clustering can have a big impact on sales and customer satisfaction, making it an invaluable tool to boost revenue, cut costs, or sometimes even both.

When -

- When you are starting from a large, unstructured dataset.
- When you do not know how many or which classes your data is divided into.



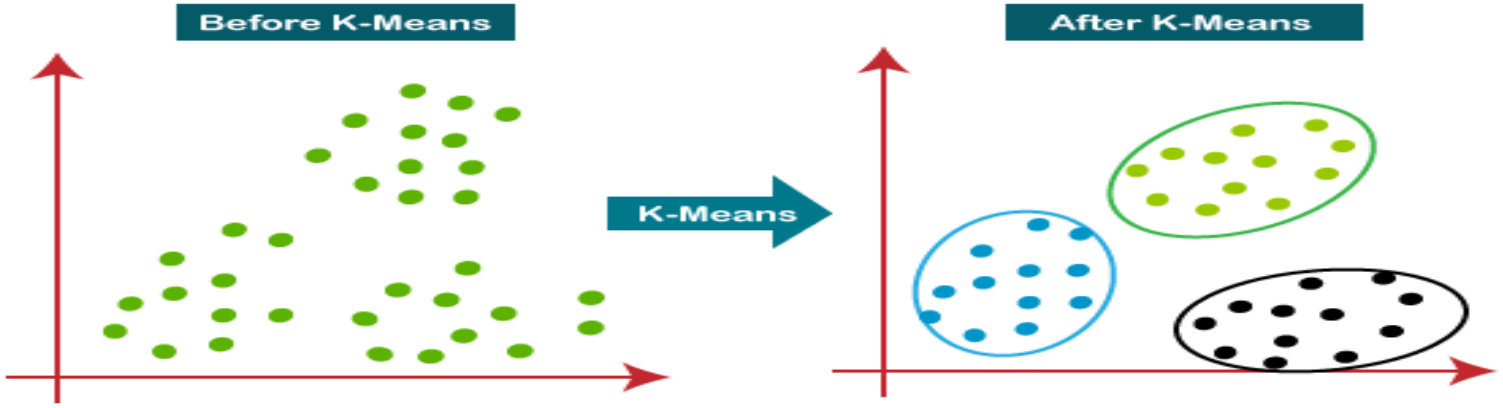
How clustering helps in finding hidden groupings and patterns?

Clustering methods simply try to group similar **patterns** into **clusters** whose members are more similar to each other (according to some distance measure) than to members of other **clusters**.

Why K-Means Clustering Is So Popular? What are its assumptions?

K-Means clustering method considers two **assumptions** regarding **the clusters** – first that **the clusters** are spherical and second that **the clusters** are of similar size. K-Means is popular because of the following advantages:

- Relatively simple to implement
- Can warm-start the positions of centroids
- Guarantees convergence
- Easily adapts to new examples.



Why converting features into continuous values and scaling the data are important to perform K-Means clustering?

As K-Means is about distance between points to centroids scaling is important. **K-means** is one method of cluster analysis that groups observations by minimizing Euclidean distances between them. This definition of Euclidean distance, therefore, requires that all **variables used** to determine **clustering using k-means** must be **continuous**.

How PCA helps in Dimensionality reduction?

Principal Component Analysis, or **PCA** for short, is a method for reducing the dimensionality of data. Principal Component Analysis (PCA), where you transform your data into a new dimensional space, where all the components are orthogonal to each other. Also, the components are sorted from the ones that describe the highest to lowest variance in the data. You would select a subset of the principal components as the features in your model, and capture a majority of your variance.

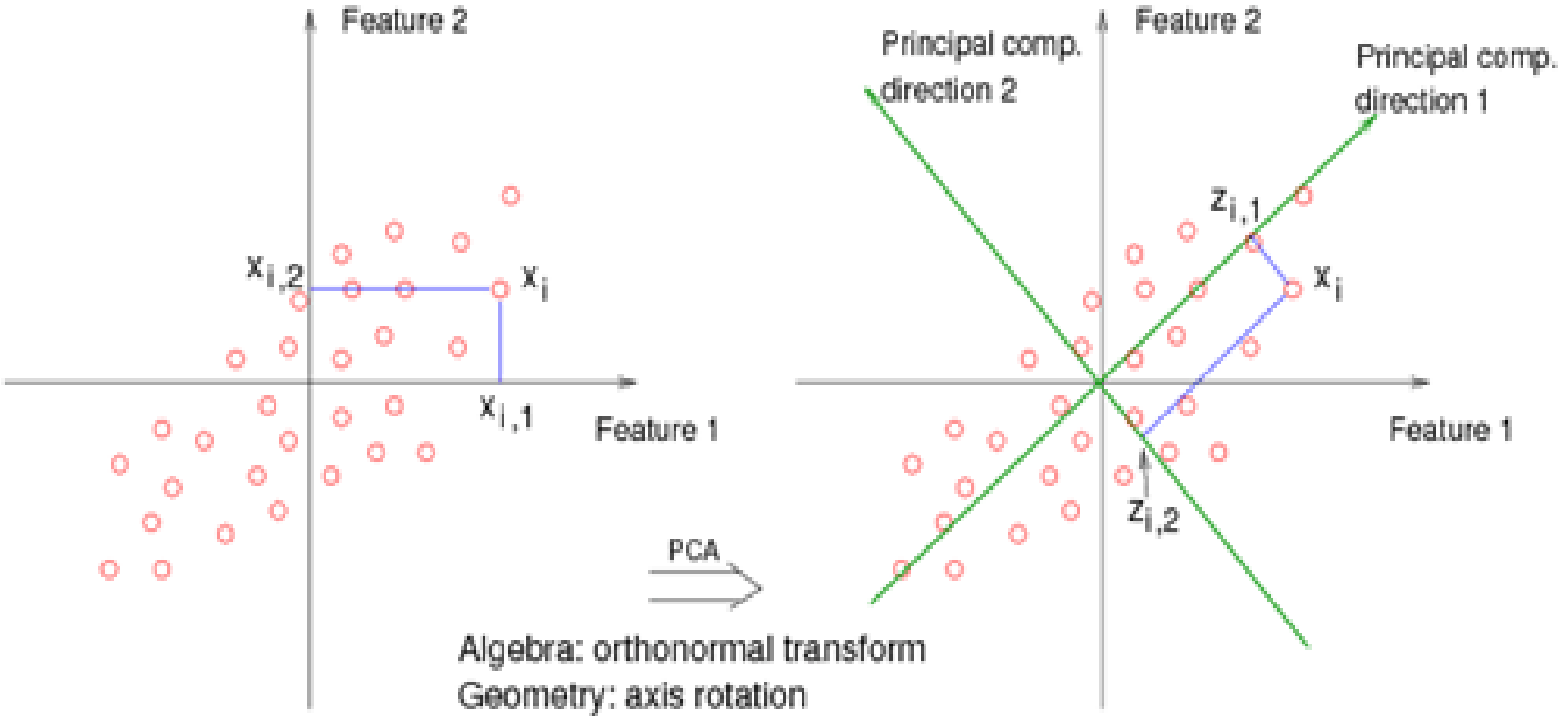
Steps:

Begin by standardizing the data.

1.Generate the covariance matrix

2. Perform eigenvalue decomposition

3. Sort the eigen pairs in descending order and select the largest one.



Case Study : Applying PCA on Credit Card data

The sample Dataset summarizes the usage behavior of about 8950 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables.
As we have 18 different variable(dimensions), We will use PCA for our dimensionality reduction.
PCA finds new dimension/axis for the dataset such that it explains maximum variance. That axis is then the first principal component. Then it chooses another component perpendicular to first principal component which explains maximum variance.
Step 1 – Scaled the dataset so that for different variables we have similar scale.
Step 2 – Apply PCA and find out the variance explained by PC's. Determine the number of top principal components to select.

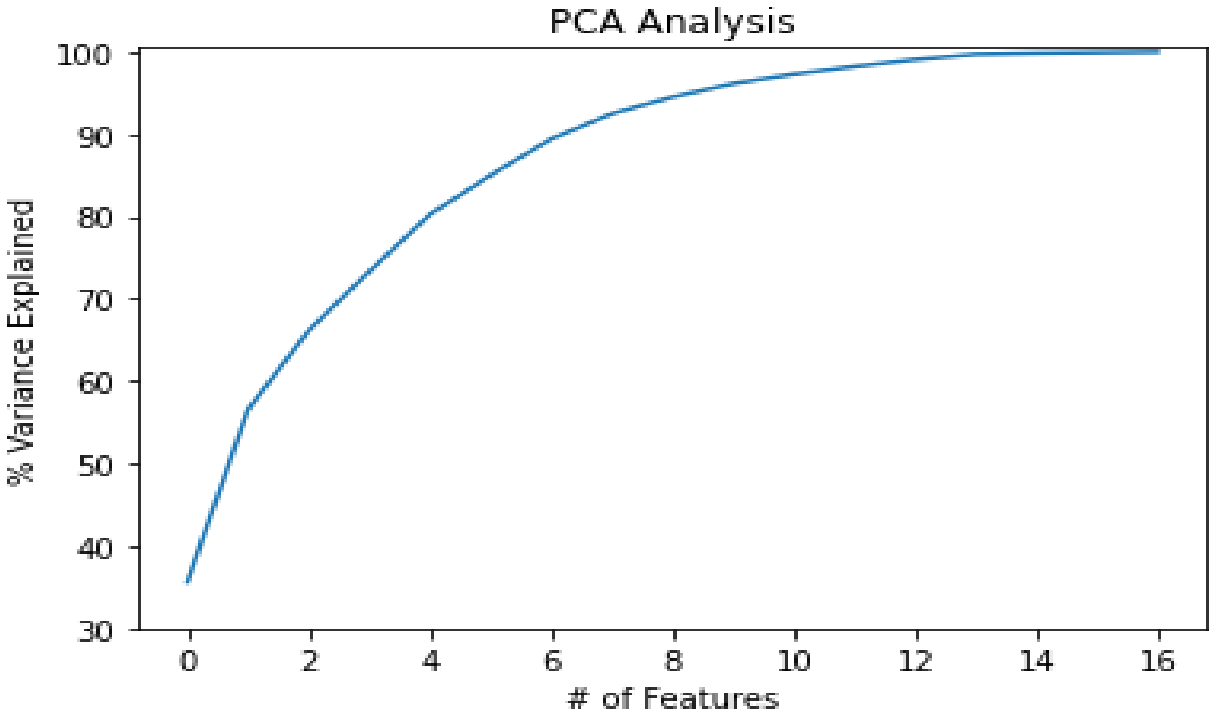
We looked at PCA and how it can be used to get a clearer understanding of the relationships between features of a dataset, while at the same time removing unnecessary noise.

PC1- 35.6 %	PC6- 4.7 %	PC11- 1.1 %	PC16- 0.1 %
PC2- 20.9 %	PC7- 4.4 %	PC12- 0.9 %	PC17- 0.1 %
PC3- 9.6 %	PC8- 3.1 %	PC13- 0.9 %	
PC4- 7.2 %	PC9- 2.0 %	PC14- 0.6 %	
PC5- 6.9 %	PC10- 1.6 %	PC15- 0.2 %	

Variance explained by each component

	PC1	PC2	PC3	PC4	PC5
PC1	1.0	-0.0	0.0	-0.0	0.0
PC2	-0.0	1.0	0.0	0.0	0.0
PC3	0.0	0.0	1.0	0.0	0.0
PC4	-0.0	0.0	0.0	1.0	0.0
PC5	0.0	0.0	0.0	0.0	1.0

Correlation between first 5 PC'S



cumulative explained variance PCA

The **Cumulative explained variance** gives the percentage of variance accounted for by the first *n* components.

For example, the cumulative percentage for the second component is the sum of the percentage of variance for the first and second components.

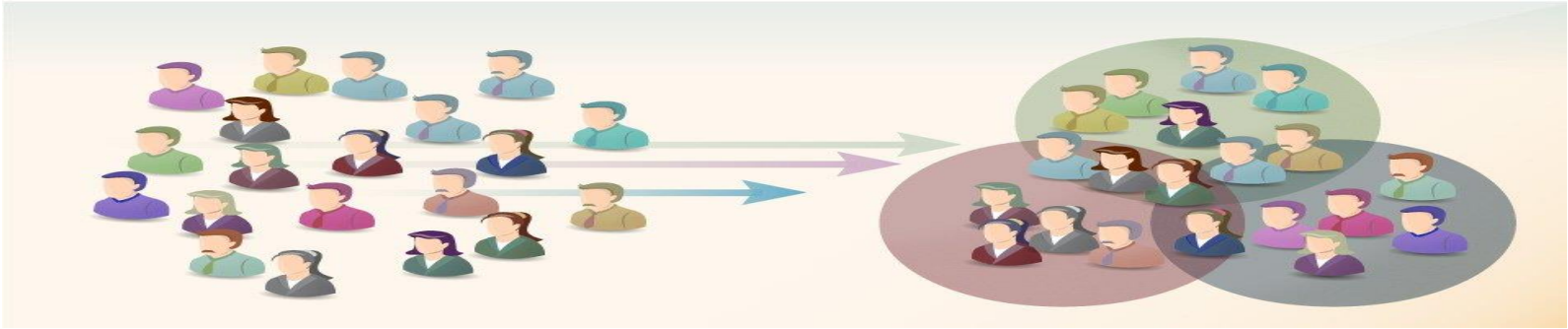
First 8 pcs are explaining nearly 93% variance of data.

K-means Clustering

K-means clustering algorithm inputs are the number of clusters K and the data set.
Algorithm starts with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set.

Some Use Cases:

- Document Classification
- Delivery Store Optimization
- Customer Segmentation
- Insurance Fraud Detection etc.



The algorithm then iterates between two steps:

1. Data assignment step:

Each centroid defines one of the clusters. In this step, each data point based on the squared **Euclidean distance** is assigned to its nearest centroid. If ci is the collection of centroids in set C, then each data point x is assigned to a cluster based on
where dist(·) is the standard (L2) Euclidean distance.

$$\min_{c_i \in C} dist(c_i, x)^2$$

2. Centroid update step:

Centroids are recomputed by taking the mean of all data points assigned to that centroid's cluster.
The algorithm iterates between step one and two until a stopping criteria is met.
This algorithm may converge on a local optimum. Assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.

K-means clustering

- K-means clusters data be separating data points into groups of equal variance.
- It requires the number of clusters to be specified, hence the term “K” in its name
- It divides the samples into K disjoint clusters.
- The K-means algorithm chooses centroids that minimizes the inertia across all the clusters.

Table 1:

Documents (Data Points)	W1 (x-axis)	W2 (y-axis)
D1	2	0
D2	1	3
D3	3	5
D4	2	2
D5	4	6

Distance between D1 and D2	Distance between D1 and D4
$\sqrt{(2-1)^2+(0-3)^2}$	$\sqrt{(2-2)^2+(0-2)^2}$
$= \sqrt{(1)^2+(3)^2}$	$= \sqrt{(0)^2+(-2)^2}$
$= \sqrt{1+9}$	$= \sqrt{0+4}$
$= \sqrt{10} = 3.17$	$= \sqrt{4} = 2$

Cluster 1: (D1, D4) Cluster 2: (D2, D3, D5)

Table 2:

Documents (Data Points)	Distance between D2 and other data points	Distance between D4 and other data points
D1	3.17	2.0
D3	2.83	3.17
D5	4.25	4.48

Clusters	Mean value of data points along x -axis	Distance between D4 and other data points
D1, D4	2.0	1.0
D2, D3, D5	2.67	4.67

From the above table, we can say the new centroid for cluster 1 is (2.0, 1.0) and for cluster 2 is (2.67, 4.67)

Documents (Data Points)	Distance between centroid of cluster 1 and data points	Distance between centroid of cluster 2 and data points
D1	1	4.72
D2	2.24	2.37
D3	4.13	0.47
D4	1	2.76
D5	5.39	1.89

We can notice now that clusters have changed the data points. Now the cluster 1 has D1, D2 and D4 data objects. Similarly, cluster 2 has D3 and D5

Step 5: Calculate the mean values of new clustered groups from Table 1 which we followed in step 3. The below table will show the mean values

Clusters	Mean value of data points along x-axis	Distance between D4 and other data points
D1, D2, D4	1.67	1.67
D3, D5	3.5	5.5

Now we have the new centroid value as following:
cluster 1 (D1, D2, D4) - (1.67, 1.67)
cluster 2 (D3, D5) - (3.5, 5.5)
This process has to be repeated until we find a constant value for centroids and the latest cluster will be considered as the final cluster solution.

Case Study – Bank customer segmentation

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. The dataset has a total of Seven independent variables - All are continuous spending, advance payments, current balance, credit limit, minimum payment amount, maximum spent in single shopping, advance payments. We have to identify the customer segments based on their credit card usage.

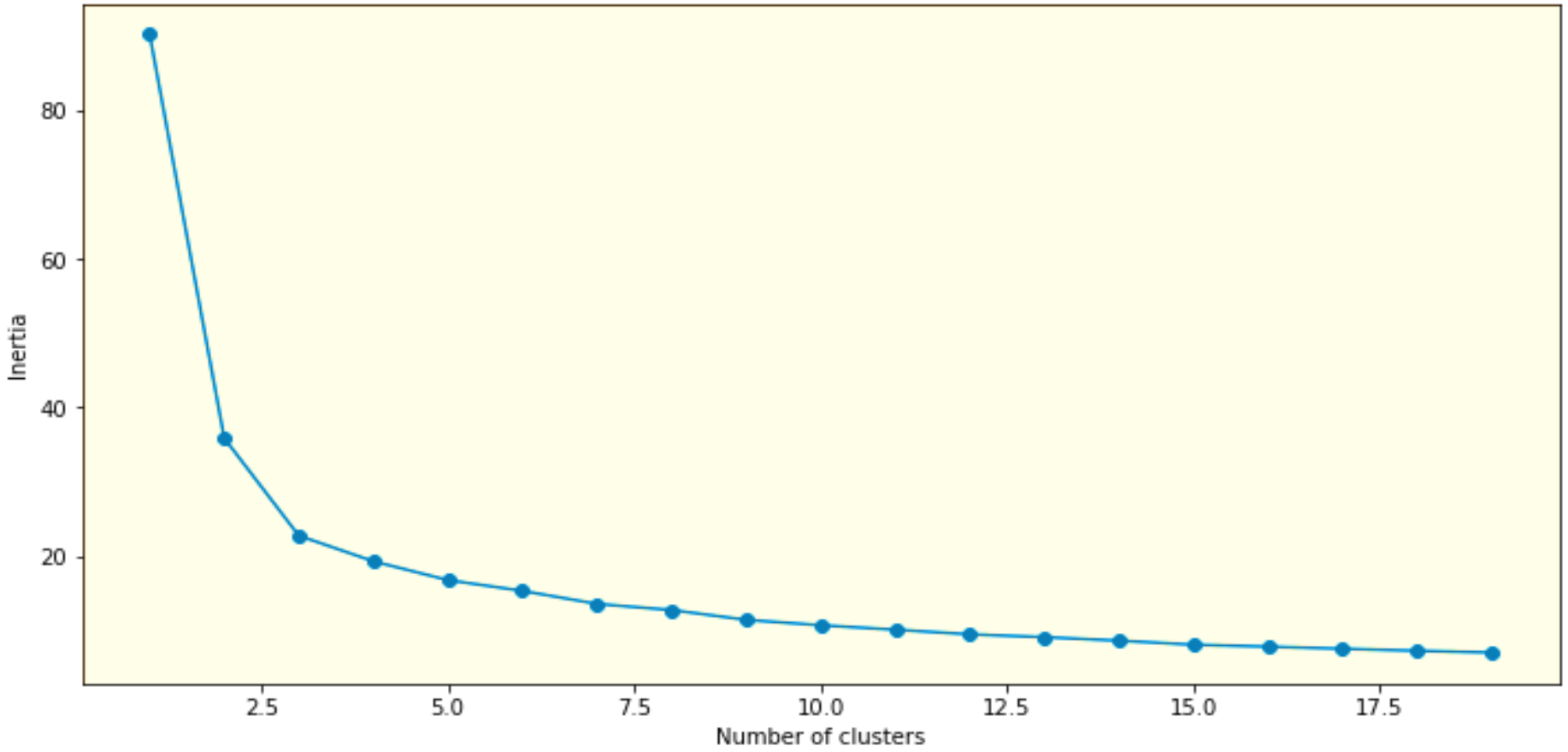
Choosing K

If the true label is not known in advance, then K-Means clustering can be evaluated using Elbow Criterion , Silhouette Coefficient.

Elbow Criterion Method:

The idea behind elbow method is to run k-means clustering on a given dataset for a range of values of k (e.g k=1 to 10), for each value of k, calculate sum of squared errors (SSE).

1. After scaling the data and **applying K-Means for k=1 to 20.**
2. For each k, **calculated** the total within-cluster sum of square (wss). **Within Cluster Sum of Squares:**
To **calculate** WCSS, you first find the Euclidean distance between a given point and the centroid to which it is assigned. You then iterate this process for all points in the **cluster**, and then **sum** the values for the **cluster** and divide by the number of points.
3. Plotted the **curve** of WSS according to the number of clusters k.



To determine the optimal number of clusters, we have to select the value of k at the "elbow" ie the point after which the distortion/inertia starts decreasing in a linear fashion. **Thus for the given data, we conclude that the optimal number of clusters is 3.**

Silhouette Coefficient Method:

A higher Silhouette Coefficient score relates to a model with better-defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores:

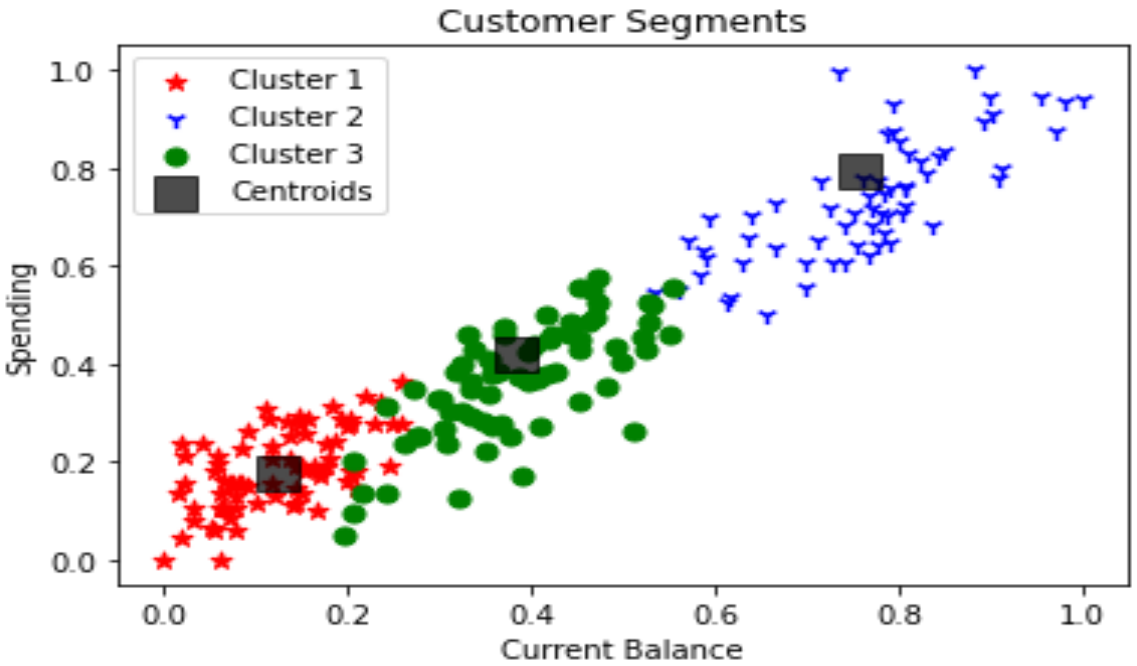
- The mean distance between a sample and all other points in the same class.
- The mean distance between a sample and all other points in the next nearest cluster.

The Silhouette Coefficient is for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$

A **higher Silhouette Coefficient** indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

	Silh_Scores	Cluster Distribution
2	0.5	[133, 77]
3	0.42	[77, 64, 69]
4	0.34	[63, 46, 51, 50]
5	0.3	[51, 27, 48, 48, 36]
6	0.28	[51, 27, 48, 22, 36, 26]
7	0.28	[16, 25, 24, 47, 41, 27, 30]
8	0.28	[37, 12, 44, 36, 23, 22, 16, 20]
9	0.26	[15, 24, 24, 33, 15, 30, 24, 21, 24]



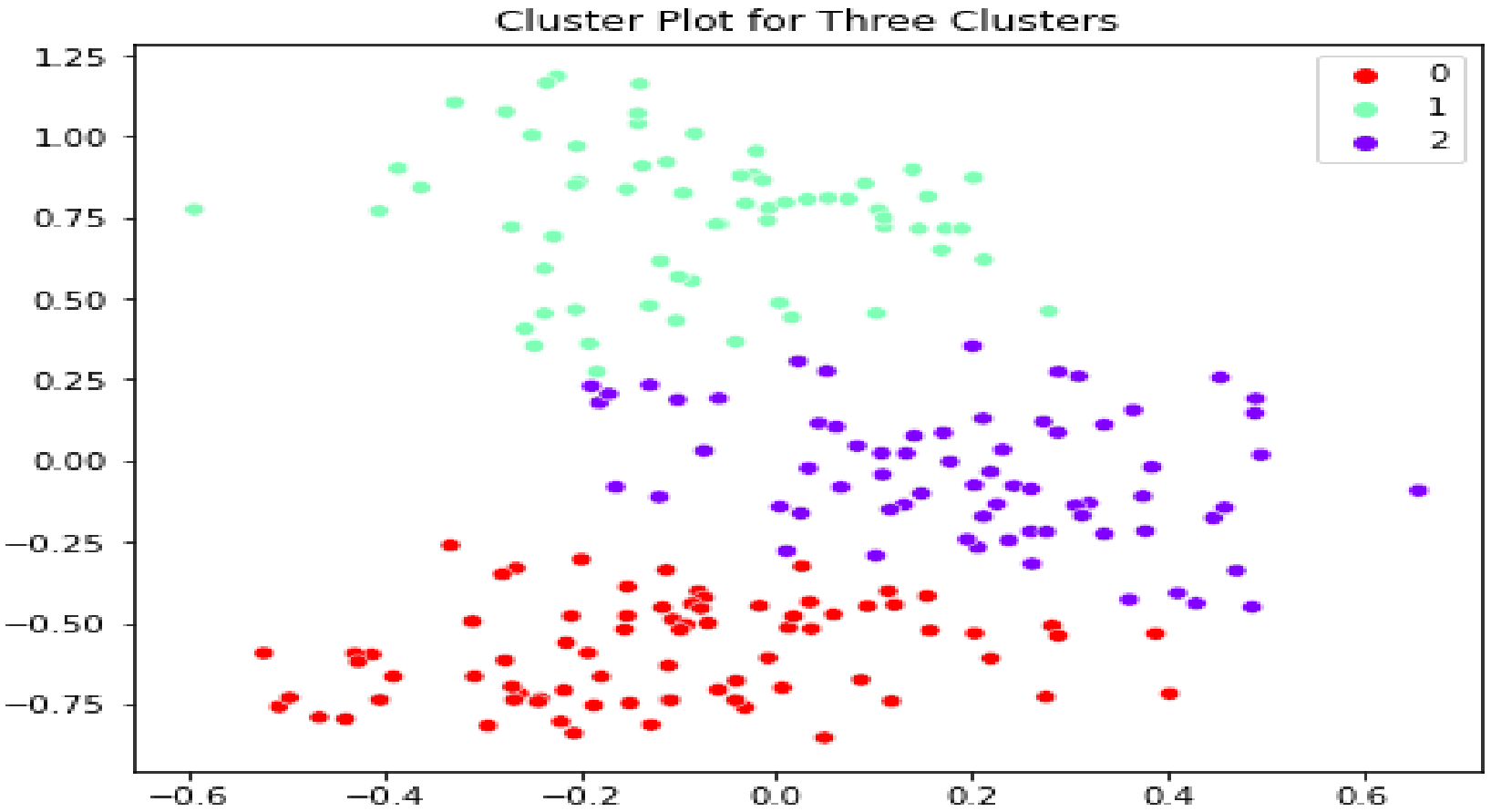
Highest silhouette score is for 2 clusters. Generally analyzing Two clusters is not very useful to the business hence we will be selecting the clusters with second best score which is for clusters = 3

Lets build the final K Means model with No of clusters as 3 and visualize the clusters.

Visualizing the three clusters using scatter plot between two variables. Different set of variables can lead to different set of inferences regarding segmentation, thus it can be helpful to visualize scatter plot for all the pairs.

1. We are able to achieve distinguishable Clusters among almost all the pairs. The plot in the left is for spending and current balance.
2. These plots can help in arriving at decisions when the business is interested in a particular pair of variables.

To visualize how the clustering is segmenting the most of data in one graph we applied **PCA** and made a scatter plot with the first 2 PC's and the clusters obtained via K-Means.



From the graph in the left we can say that the clustering results are good and we can see a clear segmentation between these 3 clusters. Lets proceed for cluster profiling.

Lets proceed for cluster profiling.

Labels_KMeans	0	1	2
spending	11.90	18.61	14.65
advance_payments	13.26	16.25	14.44
probability_of_full_payment	0.85	0.88	0.88
current_balance	5.23	6.20	5.55
credit_limit	2.86	3.71	3.29
min_payment_amt	4.59	3.59	2.80
max_spent_in_single_shopping	5.09	6.06	5.17

Cluster Profiling –

Cluster 0 - Low Spending , Low Advance Payments , Low Probability of Full Payment, Low Current Balance , Low Credit Limit, High Min Payment Amount , Low Max spent in a Single Shopping.

Cluster 1 - High Spending , High Advance Payments , High Probability of Full Payment, High Current Balance , High Credit Limit, Medium Min Payment Amount , High Max spent in a Single Shopping.

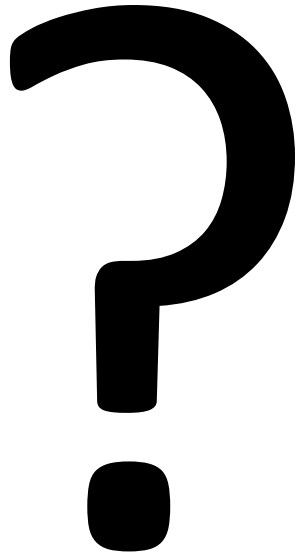
Cluster 2 - Medium Spending , Medium Advance Payments , High Probability of Full Payment, Medium Current Balance , Medium Credit Limit, Low Min Payment Amount , Medium Max spent in a Single Shopping.

Recommendations –

Cluster 0 - Customers have shown a bit reluctance to use services, the high value of min_payment_amt could imply that these customers use the services only for some particular purchases. Also, these could be the customers who are using this Bank as a secondary banking provider. Efforts must be made to convert them to frequent users.

Cluster 1 - Customers are clearly **High spenders** thus very valuable to the business. The business should target to make them a loyal customer, continuing to use services from the Bank.

Cluster 2 - Customers should be the target of an **Exclusive cashback/reward program** on purchases to try and shift them to the Ideal customers cluster.



ANY QUESTIONS