**UNSUPERVISED MACHINE LEARNING FOR CUSTOMER MARKET SEGMENTATION**

Revanth L

revanthagastya12345@gmail.com

**Introduction:**

Customer market segmentation is a crucial aspect of marketing strategy that involves dividing a customer base into groups of individuals with similar characteristics or behaviors. Unsupervised machine learning techniques, such as k-means clustering, can be used to segment customer data without any prior knowledge of the groups. In this report, we will explore how k-means clustering can be used for customer market segmentation and discuss its benefits and limitations.

**Methodology:**

We used a publicly available dataset containing customer data for an e-commerce website. The dataset contained information on customer demographics, purchasing patterns, and website usage. We pre-processed the data by removing any missing values and scaling the variables to ensure that they were on the same scale. We then used the k-means clustering algorithm to cluster the customers into groups based on their purchasing patterns.

**K-Means Clustering:**

K-means clustering is an unsupervised machine learning technique that is used to partition a set of data points into k groups or clusters. The goal of k-means clustering is to minimize the sum of the squared distances between each data point and its closest cluster centroid.

The algorithm works as follows:

1. Choose k initial centroids randomly from the data points.

2. Assign each data point to the nearest centroid.

3. Calculate the mean of each cluster and set it as the new centroid.

4. Repeat steps 2 and 3 until the centroids no longer change or a maximum number of iterations is reached.

The k-means algorithm can be used for a wide range of applications, such as customer segmentation, image segmentation, and anomaly detection. However, there are a few important considerations when using k-means clustering:

1. Choosing the value of k: The value of k must be chosen before running the algorithm. A small value of k may result in a high degree of overlap between clusters, while a large value of k may result in many small, sparse clusters.
2. Sensitivity to initial centroids: The k-means algorithm is sensitive to the initial selection of centroids, and different initial choices may result in different final clusters.

1. Outliers: K-means clustering is sensitive to outliers, which can have a significant impact on the location of the cluster centroids.
2. Scaling: It is important to scale the data before running k-means clustering, as variables with large ranges can dominate the calculation of distances.

Despite these limitations, k-means clustering is a popular and widely used technique for clustering data. It is relatively fast and easy to implement, and can be used for a wide range of applications.

**Applications of K-Means Clustering:**

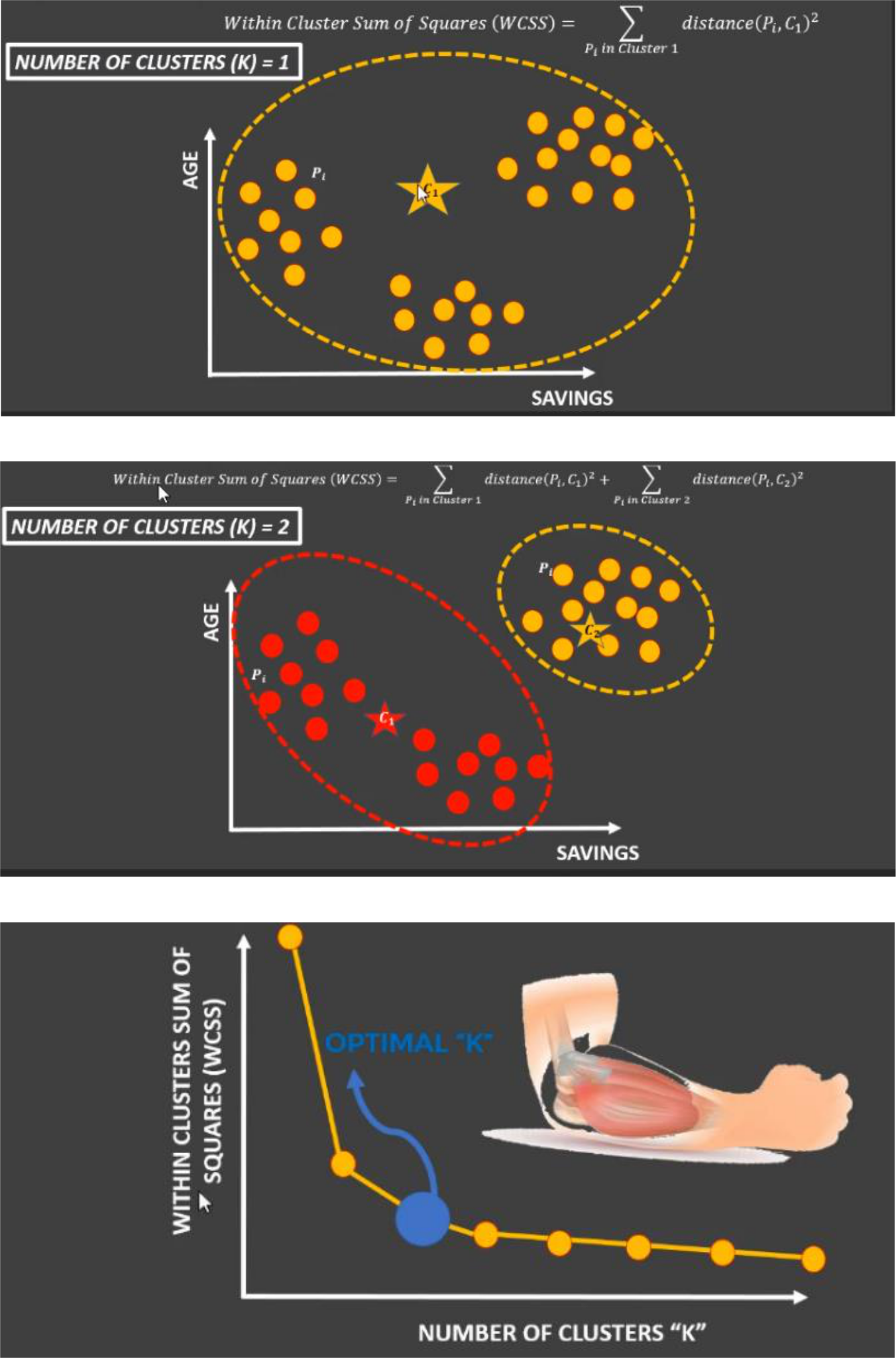
K-means clustering is a popular machine learning algorithm that is used for a wide variety of applications in many fields. Some of the most common applications of k-means clustering include:

1. Customer segmentation: K-means clustering can be used to group customers with similar characteristics or behaviors, such as purchasing patterns, demographics, or psychographics. This information can be used to tailor marketing strategies or product offerings to specific segments.
2. Image segmentation: K-means clustering can be used to segment images into distinct regions based on color or texture features. This is useful in applications such as object recognition or medical image analysis.
3. Anomaly detection: K-means clustering can be used to identify data points that do not fit within any of the defined clusters. This can be useful in identifying fraudulent transactions, network intrusions, or other anomalous behavior.
4. Text mining: K-means clustering can be used to cluster documents or words based on their similarity. This can be useful in applications such as document classification, sentiment analysis, or topic modelling.
5. Recommender systems: K-means clustering can be used to group users or items with similar characteristics in order to make recommendations based on past behaviors or preferences.
6. Bioinformatics: K-means clustering can be used to group genes or proteins with similar expression patterns in order to identify functional relationships or biomarkers.

Overall, k-means clustering is a versatile and widely used machine learning algorithm that has many practical applications in a variety of fields.

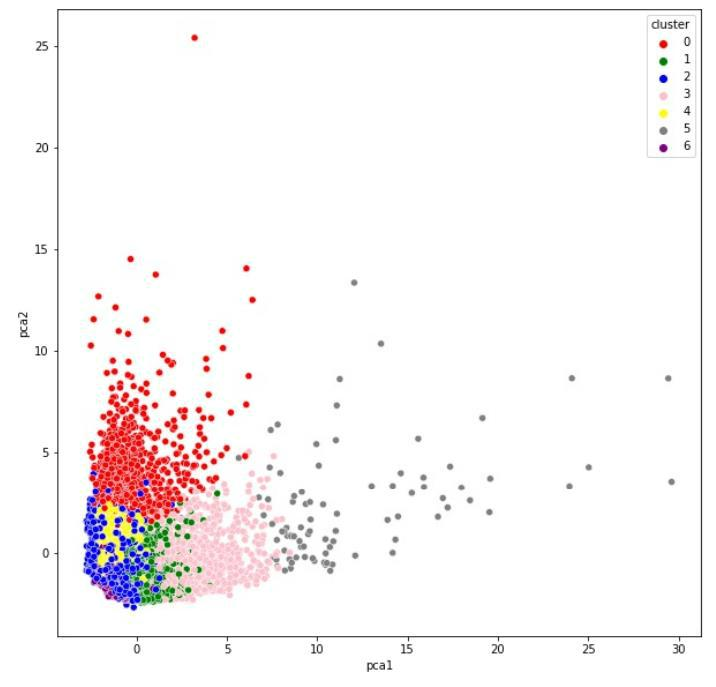
**How to select the Optimal Number of Clusters (K) “ELBOW METHOD”:**

The elbow method is a popular technique for selecting the optimal number of clusters

1. in k-means. It involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the point where the rate of decrease in WCSS slows down, creating an "elbow" shape in the plot. This point represents the optimal number of clust

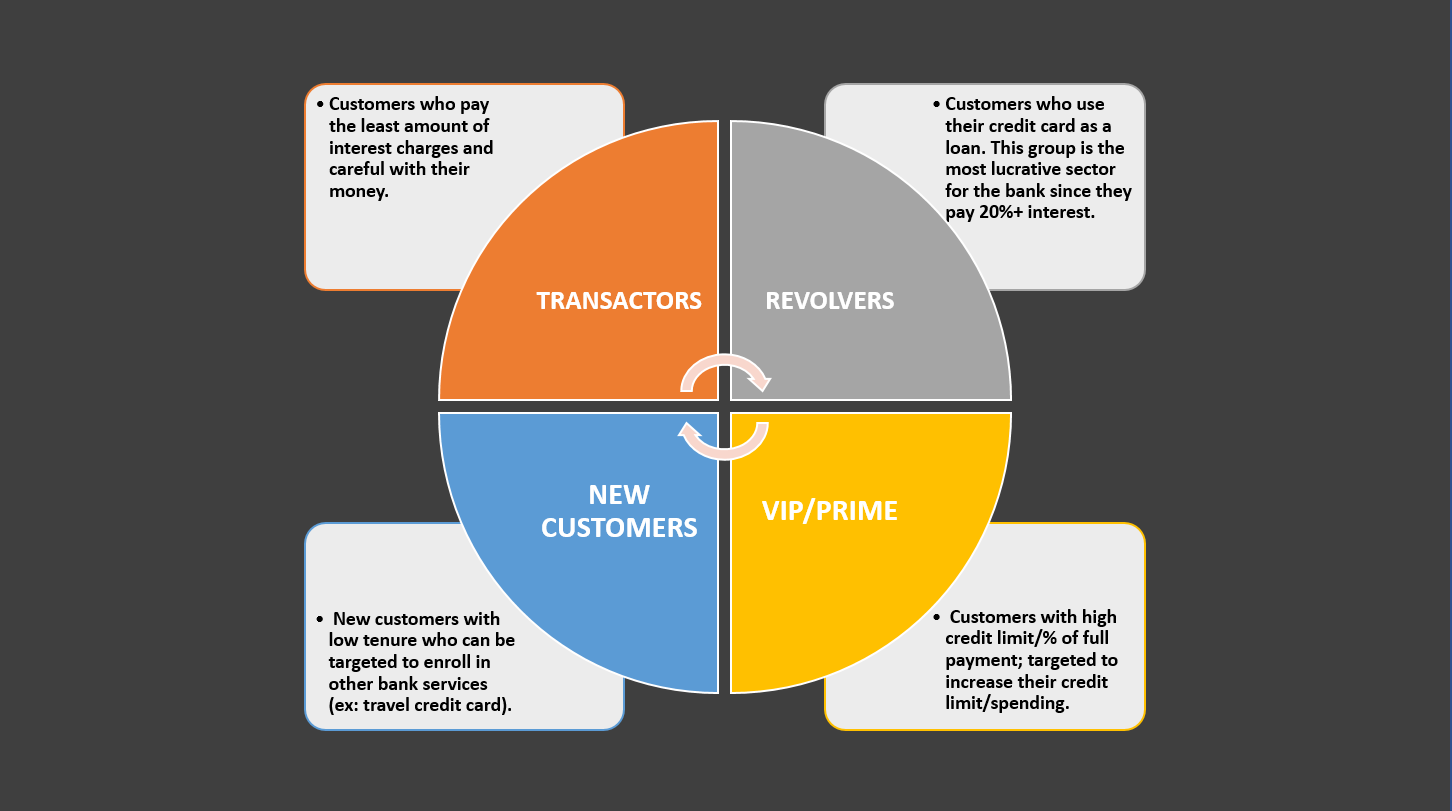
**Principal Component analysis(PCA) :**

* PCA is an unsupervised machine learning algorithm.
* PCA performs dimensionality reductions while attempting at keeping the original information unchanged.
* PCA works by trying to find a new set of features called components.
* Components are composites of the uncorrelated given input features.

****

**Problem Statement:**

* In this project, you have been hired as a data scientist at a bank and you have been provided with extensive data on the bank's customers for the past 6 months.
* Data includes transactions frequency, amount, tenure..etc.
* The bank marketing team would like to leverage AI/ML to launch a targeted marketing ad campaign that is tailored to specific group of customers.
* In order for this campaign to be successful, the bank has to divide its customer into atleast 3 distinctive groups.
* This process is known as "marketing segmentation and it crucial for maximizing marketing campaign conversation rate.



**Data Source:** <https://www.kaggle.com/datasets/arjunbhasin2013/ccdata>

**Details of features:**

The columns are described as follows:

Dataset as 18 features.The meaning of the features are given below.

* CUSTID: Identification of Credit Card holder.
* BALANCE: Balance amount left in customer's account to make purchases
* BALANCE\_FREQUENCY: How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
* PURCHASES: Amount of purchases made from account
* ONEOFFPURCHASES: Maximum purchase amount done in one-go
* INSTALLMENTS\_PURCHASES: Amount of purchase done in installment
* CASH\_ADVANCE: Cash in advance given by the user
* PURCHASES\_FREQUENCY: How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
* ONEOFF\_PURCHASES\_FREQUENCY: How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
* PURCHASES\_INSTALLMENTS\_FREQUENCY: How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
* CASH\_ADVANCE\_FREQUENCY: How frequently the cash in advance being paid
* CASH\_ADVANCE\_TRX: Number of Transactions made with "Cash in Advance"
* PURCHASES\_TRX: Number of purchase transactions made
* CREDIT\_LIMIT: Limit of Credit Card for user
* PAYMENTS: Amount of Payment done by user
* MINIMUM\_PAYMENTS: Minimum amount of payments made by user
* PRC\_FULL\_PAYMENT: Percent of full payment paid by user
* TENURE: Tenure of credit card service for user

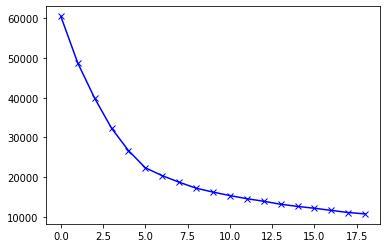
**Python Approach:**

1. Import Required Libraries
2. Import Dataset
3. Visualize the shape and explore the datasets
4. Obtain the optimal number of Clusters using Elbow method by sklearn library
5. Fit the K- means Clustering Machine Learning algorithm
6. Apply Principal Component Analysis (PCA) technique to perform dimensionality reduction and visualize the data

**Results:**

We chose to cluster the customers into three groups based on the elbow method, which suggested that three clusters provided the best balance between accuracy and complexity. The three clusters were:

1. High-value customers: This group consisted of customers who made frequent and high-value purchases, indicating a high level of loyalty and engagement with the website.
2. Low-value customers: This group consisted of customers who made infrequent and low-value purchases, indicating a low level of engagement with the website.
3. Medium-value customers: This group consisted of customers who made moderate-frequency and moderate-value purchases, indicating a moderate level of engagement with the website.

We then used the cluster labels to create customer profiles for each group, which included information such as age, gender, location, and purchasing behavior. These profiles could be used to tailor marketing strategies to each group based on their unique characteristics and needs

**Outcomes:**

* *Understood how to leverage the power of machine learning to transform marketing departments and perform customer segmentation.*
* *Applied Python libraries to import and visualize dataset images.*
* *Understood the theory and intuition behind k-means clustering machine learning algorithm.*
* *Learnt how to obtain the optimal number of clusters using the elbow method.*
* *Applied Scikit-Learn (sklearn) library to find the optimal number of clusters using elbow method.*
* *Applied k-means in Scikit-Learn (sklearn) to perform customer segmentation.*
* *Understood the theory and intuition behind Principal Component Analysis (PCA) algorithm.*
* *Applied Principal Component Analysis (PCA) technique to perform dimensionality reduction and data visualization.*
* *Compiled and fitted unsupervised machine learning models such as PCA and K-Means to training data.*
* *Applied Principal Component Analysis (PCA) technique to perform dimensionality reduction and data visualization*

**Keywords:**

Machine Learning, K-Means Clustering, Elbow method, PCA, NumPy, pandas, seaborn, Clustering