MACHINE LEARNING PROJECT

INTRODUCTION TO MACHINE LEARNING

ARM 210

University School of Automation and Robotics
East Campus, GGSIPU University
Surajmal Vihar, New Delhi-110092

Submitted To:
Dr. Amit Choudhary
Assistant Professor
USAR, GGSIPU

Submitted By:
Name: Ankit Sharma
Batch: AIML B2
Enrolment no.: 20319051622

PROJECT INFORMATION

Title of Project: Clustering Thai Fashion and Cosmetics Retail: Unveiling Trends on Facebook Live Sellers
Student Name: **Ankit Sharma**
Enrolment Number: **20319051622**
Email Id: ankit.09019011722@ipu.ac.in
Contact No. : +91 8810352054
Code link:
https://colab.research.google.com/drive/14UHCZtqUd1k-aey01bANt_MNgEqoNxya?usp=sharing

ABSTRACT

This report presents a clustering analysis of Facebook pages belonging to 10 Thai fashion and cosmetics retail sellers. The dataset comprises posts of various types, including videos, photos, statuses, and links, with engagement metrics such as comments, shares, and reactions. The dataset is multivariate, with 7051 instances and 11 features of integer type. The study falls within the subject area of business, focusing on clustering tasks to uncover patterns and insights within the dataset. Methodologies include preprocessing, exploratory data analysis, feature engineering, and clustering algorithms such as K-means, DBSCAN, and Hierarchical Clustering. Evaluation metrics including Adjusted Rand index, Mutual Information based score, and Silhouette Coefficient are utilized to assess the performance of clustering algorithms. Results and implications of the clustering analysis are discussed, providing valuable insights for business decision-making and future research endeavours in the domain of fashion and cosmetics retailing.

Introduction

This project focuses on the use of clustering algorithms to analyze and examine data. The primary goal is to identify hidden patterns, structures, or groups in the data that typical analytic approaches may not instantly reveal. This project seeks to split the data into meaningful clusters based on similarity or proximity criteria using a variety of clustering techniques, including K-means, DBSCAN, and Hierarchical Clustering.

The objectives of this project include:

1. **Data Understanding**: Use EDA and preprocessing techniques to learn about the dataset's structure.
2. **Feature Engineering**: Add important features to increase clustering performance.
3. **Use clustering algorithms** to categorize data points based on patterns or similarities.
4. **Evaluate clustering methods** based on relevant criteria to determine their success in capturing the data structure.
5. **Insight Generation**: Use clustering results to inform decision-making and analysis.

Through this project, we hope to illustrate the utility and effectiveness of clustering approaches in revealing hidden patterns and structures in a dataset, ultimately providing valuable insights for decision-making and future study.

Description of the Dataset

The dataset consists of Facebook pages belonging to ten Thai fashion and cosmetics retailers. It includes posts of many forms, such as videos, images, status updates, and links. Each post's engagement analytics include comments, shares, and responses.

Dataset Characteristics:

- **Multivariate:** The dataset contains multiple variables (features) for each instance.
- **Subject Area:** The dataset pertains to the business domain, specifically in the context of fashion and cosmetics retail sellers on Facebook.
- **Associated Tasks:** The primary task associated with this dataset is clustering, which involves grouping similar Facebook pages based on their engagement metrics.
- **Feature Type:** The features in the dataset are of integer type.
- **Number of Instances:** There are 7051 instances in the dataset.
- **Number of Features:** The dataset comprises 11 features.

Overview of Notebooks Used in the Project:

1. **PreProcessing:**
   The notebook prepares the dataset for analysis by completing preprocessing procedures such handling missing values, encoding categorical variables, and scaling numerical features. Given the variety of data sources on Facebook, pretreatment assures data consistency and compatibility for future analysis.

2. **Exploratory Data Analysis (EDA):**
   To understand the dataset's distribution, the relationships between its variables, and any patterns or anomalies that may exist, EDA entails analysing and displaying the dataset. Within the framework of this study, EDA aids in comprehending the patterns of engagement among various Facebook pages, spotting patterns or similarities.

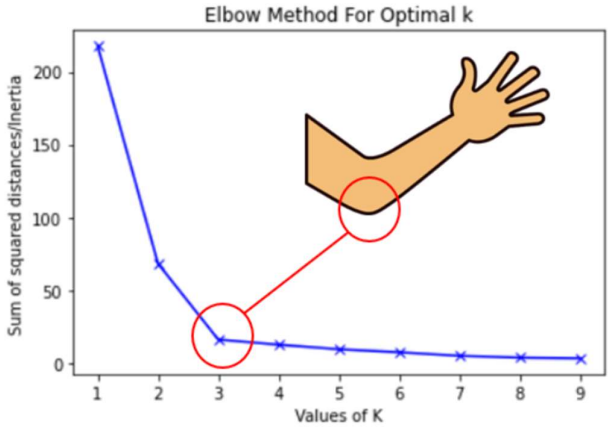
3. **Feature Engineering:**
   Feature engineering seeks to enhance the functionality of machine learning algorithms by adding new features or altering current ones. Feature engineering in the context of Facebook engagement data could entail aggregating measurements, generating new features from preexisting ones, or altering features to depict underlying patterns more accurately.

4. **Feature Extraction via PCA:**
   A dimensionality reduction method called principal component analysis (PCA) is utilized to pinpoint a dataset's most crucial properties. In this notebook, the Facebook engagement data is reduced in dimension while maintaining as much variance as feasible by the application of PCA to extract the most important components.

5. **Clustering:**
   Using the preprocessed and feature-engineered data, the clustering notebook applies many clustering techniques, including K-means, DBSCAN, and Hierarchical Clustering. It facilitates the discovery of related groups or clusters by clustering the Facebook pages according to their engagement metrics.

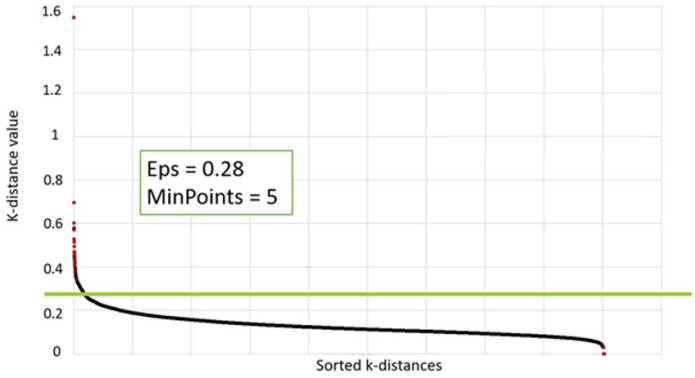
Explanation of Analysis Techniques:

1. **Elbow Method for K-means:**
   One heuristic method for figuring out how many clusters in a dataset are best for K-means clustering is the elbow method. The process is graphing the number of clusters versus the within-cluster sum of squares (WCSS) to locate the "elbow" point, or point of declining returns. This number is the ideal number of clusters at which increasing the number of clusters has no discernible negative impact on the WCSS.

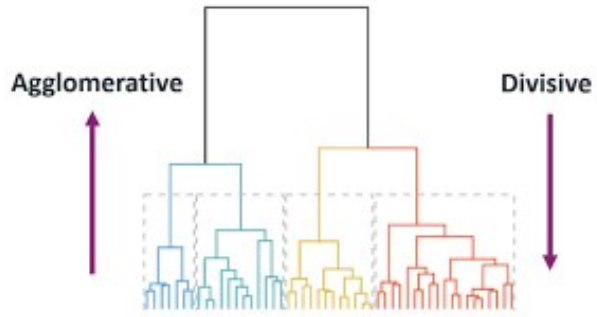


Line plot between K and inertia

2. **Utilization of k_distance_graph for DBSCAN:**
   Two parameters are needed for the density-based clustering algorithm, DBSCAN: epsilon (ε) and minPoints. Plotting each point's k-distance against its index using the k_distance_graph approach aids in determining the ideal value for epsilon. A good number for epsilon is indicated by the "knee" or abrupt increase in distance on the graph; points beyond this are deemed outliers.



3. **Plotting of the Dendrogram for Hierarchical Clustering:**
   By repeatedly combining or dividing clusters according to how similar they are, hierarchical clustering builds a hierarchy of clusters. A figure that resembles a tree and shows the hierarchical relationships between groups is called a dendrogram. By highlighting notable changes in cluster distances, it aids in the visualization of the clustering process and helps establish the ideal number of clusters.



Clustering Algorithms

1. K-means:

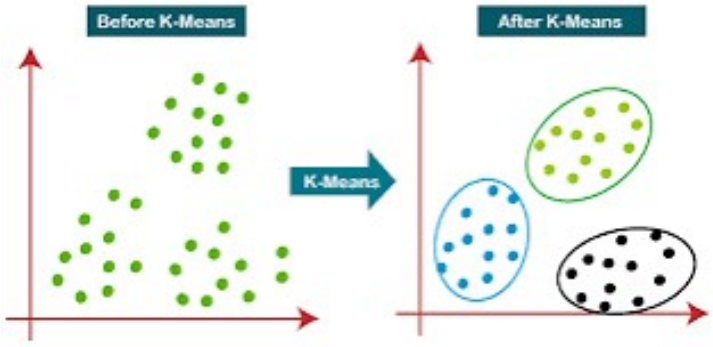
Description: Data points are grouped into K groups using the well-liked partitioning clustering technique K-means. Its objectives are to maximize variance between clusters and minimize variance within clusters.

How it Works:

- **Initialization:** Choose K initial cluster centroids randomly from the data points.
- **Assignment:** Assign each data point to the nearest centroid, forming K clusters.
- **Update Centroids:** Recalculate the centroids of the clusters based on the mean of the data points assigned to each cluster.
- **Iteration:** Repeat the assignment and centroid update steps until convergence, i.e., when the centroids no longer change significantly or a predefined number of iterations is reached.
-

Application to the Dataset:

- For the Facebook live sellers dataset, K-means can be applied to cluster the sellers based on their engagement metrics such as comments, shares, and reactions. It will group similar sellers together based on their engagement patterns.



2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

Description: Density-based clustering, or DBSCAN, clusters together densely packed points; it defines clusters as high-density regions divided by low-density regions. It is especially good at tolerating noise and recognizing clusters of any shape.

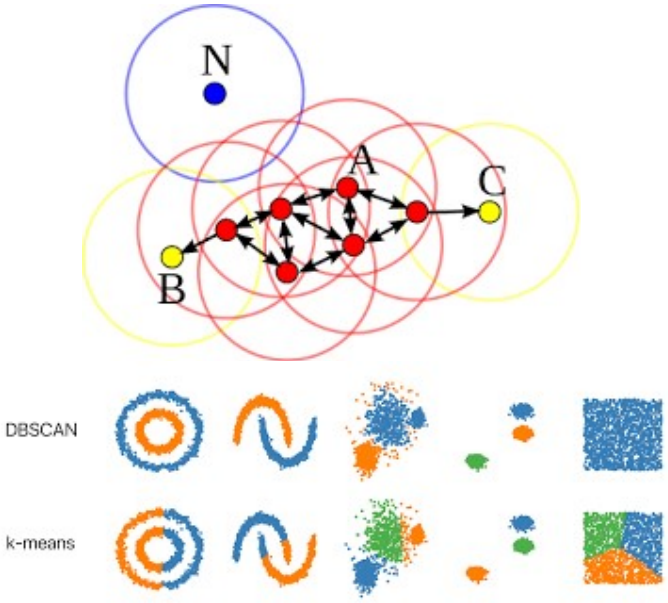
How it Works:

- **Core Points:** A point is considered a core point if it has at least a specified number of points (minPoints) within a specified distance (epsilon, ε) in its neighborhood.
- **Border Points:** A point is considered a border point if it is reachable from a core point but does not have enough points in its neighborhood to be considered a core point.
- **Noise Points:** Points that are neither core nor border points are considered noise points and are not assigned to any cluster.

- **Cluster Formation:** DBSCAN starts with an arbitrary point and expands the cluster by recursively adding neighboring points. It forms a cluster for each core point and includes its reachable points (core and border points) in the same cluster.

Application to the Dataset:

- DBSCAN can be applied to the Facebook live sellers dataset to identify clusters of sellers with similar engagement patterns. It can automatically detect outliers (noise points) representing sellers with unusual engagement behavior.



3. Hierarchical Clustering:

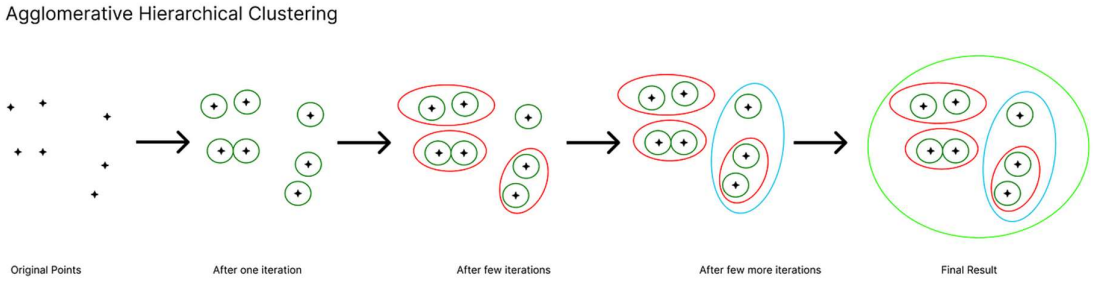
Description: Hierarchical clustering creates a hierarchy of clusters by iteratively merging or separating them based on their similarity. It is not necessary to define the number of clusters beforehand.

How it Works:

- **Agglomerative Approach:** Start with each data point as a singleton cluster and iteratively merge the closest pairs of clusters until only one cluster remains.
- **Dendrogram:** Represent the clustering process as a dendrogram, a tree-like diagram that illustrates the hierarchical relationships between clusters.
- **Dissimilarity Measures:** Use various dissimilarity measures (e.g., Euclidean distance) to quantify the similarity between clusters or data points.
- **Linkage Criteria:** Different linkage criteria (e.g., complete, single, average) determine how the distance between clusters is computed during merging.
-

Application to the Dataset:

- Hierarchical clustering can be applied to the Facebook live sellers dataset to reveal the hierarchical structure of engagement patterns among sellers. It provides insights into both global and local similarities between sellers, enabling the identification of clusters at different levels of granularity.



Evaluation Metrics

1. **Adjusted Rand Index (ARI):**
   • ARI compares genuine clustering labels to algorithmic results, adjusting for chance. It looks at all pairings of samples and counts the agreements and disputes between the real and projected clusters.

   • The range is [-1, 1], with 1 indicating complete agreement, 0 indicating random grouping, and negative values indicating disagreement.
   - The unadjusted Rand index is then given by:

$$\text{RI} = \frac{a + b}{C_2^{n_{samples}}}$$

2. **Mutual Information based Score:**
   • Mutual Information (MI) refers to the quantity of information gained from one variable through knowledge of another. In clustering, the MI-based score measures the agreement between real and projected cluster assignments.
   • The range is [0, 1], with 1 representing full clustering agreement and 0 indicating no mutual information.
   - Using the expected value, the adjusted mutual information can then be calculated using a similar form to that of the adjusted Rand index:

$$\text{AMI} = \frac{\text{MI} - E[\text{MI}]}{\text{mean}(H(U), H(V)) - E[\text{MI}]}$$

3. **Homogeneity:**
   • Homogeneity refers to the purity of clusters, suggesting that they exclusively include members of one type. It determines if each cluster contains data points that are comparable to each other in terms of true labels.
   • Range: [0, 1], with 1 representing complete homogeneity.

4. **Completeness:**
   • Completeness indicates that all members of a class belong to the same cluster. It determines whether all data points from the same class are allocated to the same cluster.
   • Range: [0, 1], with 1 indicating perfect completion.

5. **V-measure:**
   • The V-measure takes the harmonic mean of homogeneity and completeness. It gives a single measure of clustering quality that balances homogeneity and completeness.
   • Range: [0, 1], with 1 indicating optimal clustering quality.
   - This function's is as follows:

$$v = \frac{(1+\beta) \times \text{homogeneity} \times \text{completeness}}{(\beta \times \text{homogeneity} + \text{completeness})}$$

6. **Silhouette Coefficient:**
   • The Silhouette Coefficient assesses cluster tightness and separation. It computes the mean silhouette coefficient over all samples, with a greater value indicating better grouping.
   • The range is [-1, 1], with 1 representing dense, well-separated clusters, 0 indicating overlapping clusters, and negative values indicating misclassification.
   - The Silhouette Coefficient *s* for a single sample is then given as:

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

7. **Calinski-Harabasz Index (CHI):**
   • CHI measures the ratio of between-cluster to within-cluster dispersion. It assesses the compactness and isolation of clusters, with higher scores indicating more defined clusters.
   • Range: Higher numbers are preferred, with no defined range.
   - For a set of data E of size nE which has been clustered into k clusters, the Calinski-Harabasz score s is defined as the ratio of the between-clusters dispersion mean and the within-cluster dispersion:

$$s = \frac{\text{tr}(B_k)}{\text{tr}(W_k)} \times \frac{n_E - k}{k-1}$$

Importance of Each Metric in Evaluating Clustering Performance:

- **Adjusted Rand Index (ARI) and Mutual Information based Score:** These metrics estimate the agreement between real and predicted cluster labels, giving an overall picture of clustering accuracy.
- **Homogeneity, Completeness, and V-measure:** These metrics analyze the purity and completeness of clusters by determining if they contain members of the same class and if all members of a class are allocated to the same cluster.

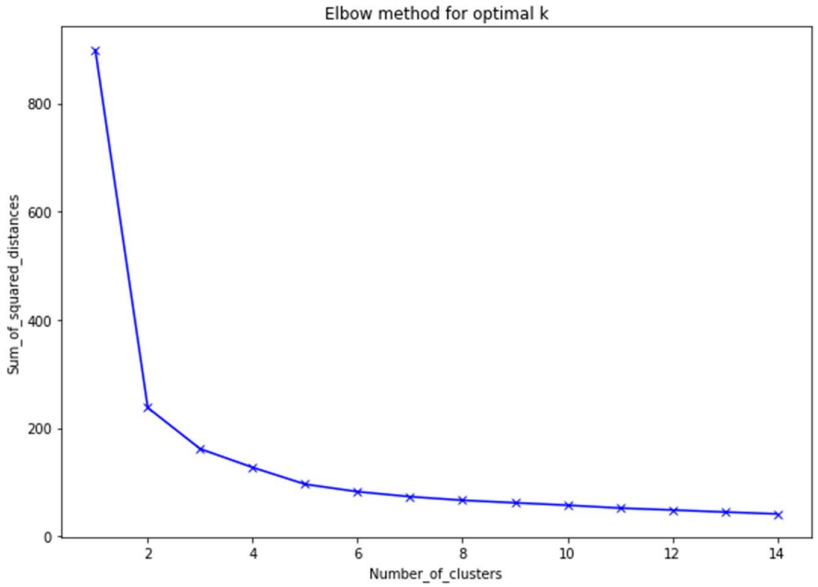
- **Silhouette Coefficient:** This metric assesses cluster compactness and separation, allowing for the identification of well-separated clusters with little overlap.

- **Calinski-Harabasz Index (CHI):** CHI measures overall clustering quality by comparing between-cluster dispersion to within-cluster dispersion. Higher numbers suggest more well-defined clusters.

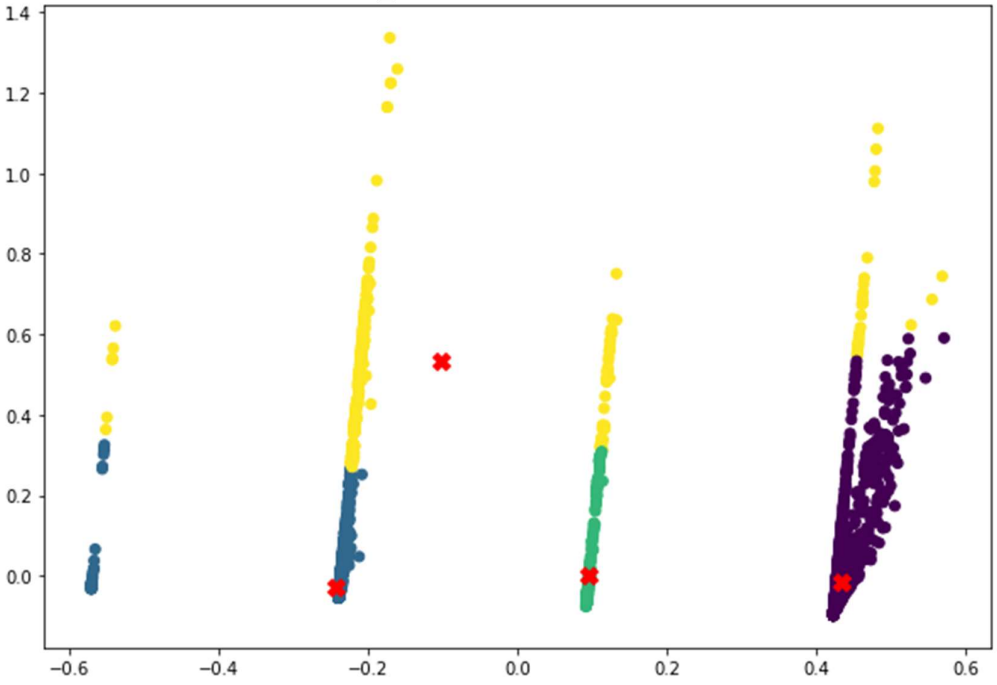
Presentation of Results from Each Clustering Algorithm

1.K-Means Clustering

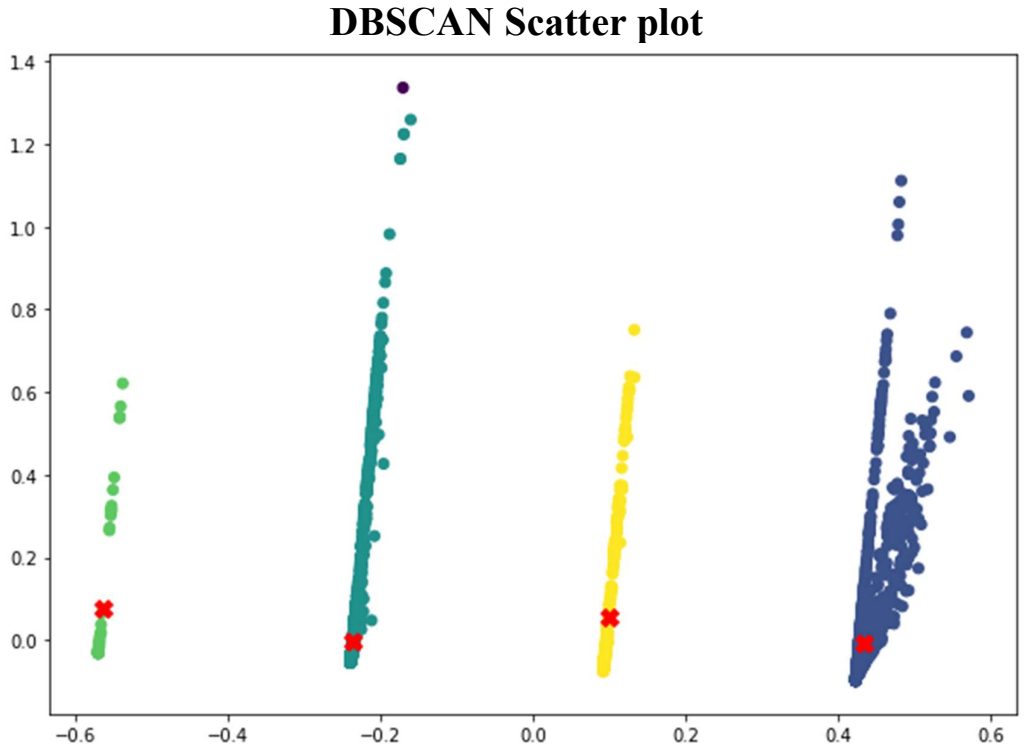
Elbow method plot from Code



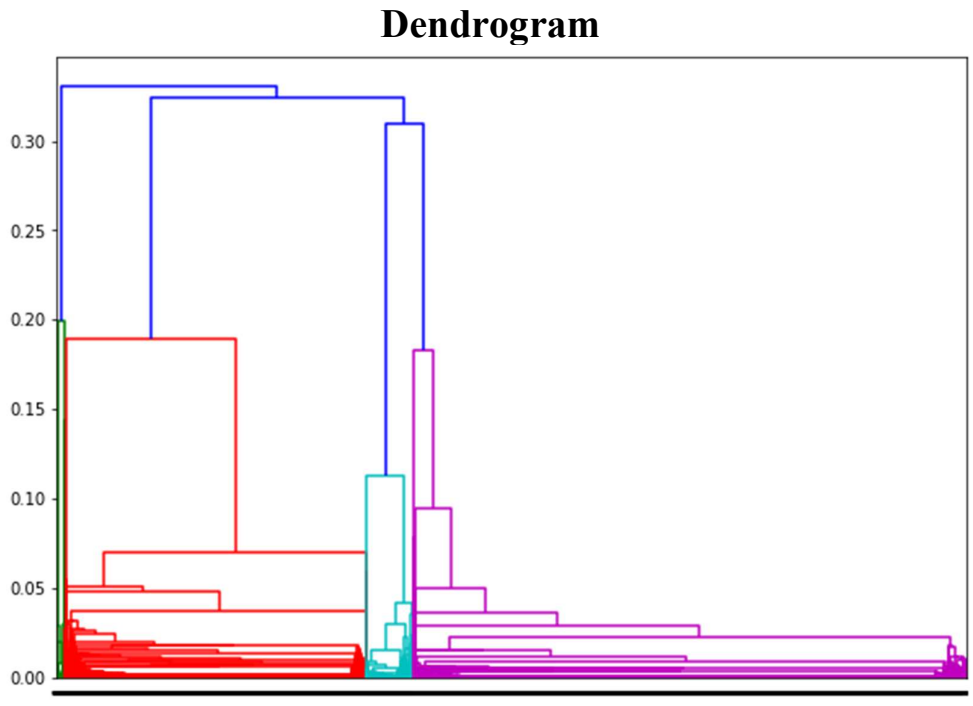
Scatter plot with X as centres



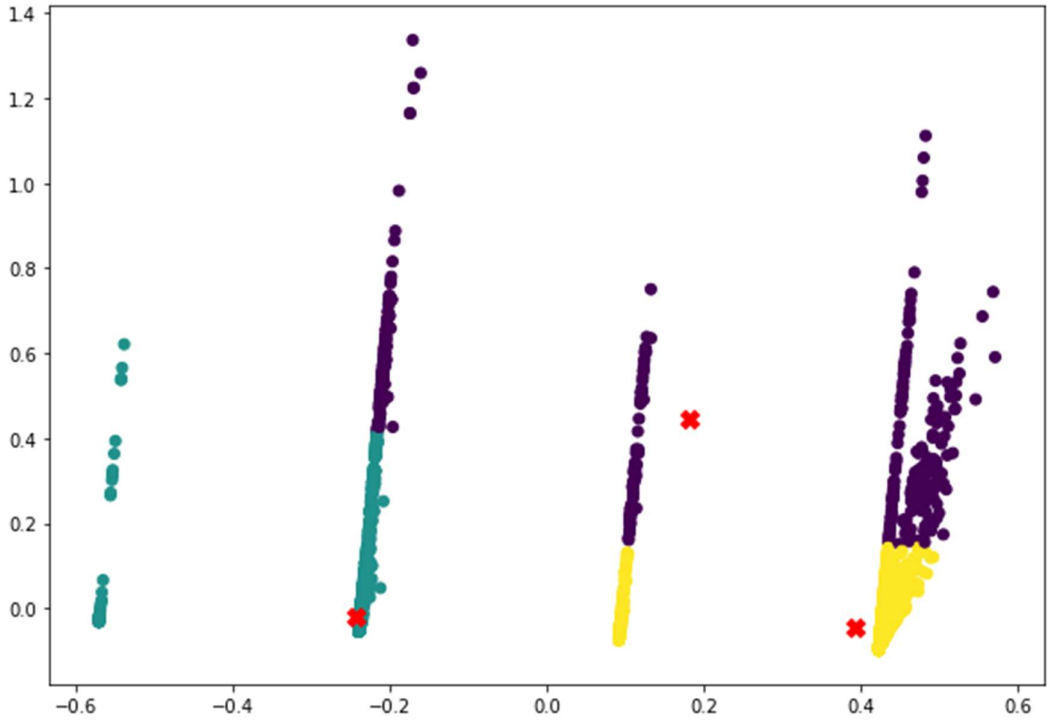
2.DBSCAN Clustering



3.Hierarchical Clustering (Agglomerative)



Agglomerative Clustering Scatter Plot



Comparison of Performance Based on Evaluation Metrics:

1. K-means Clustering:

- **Homogeneity**: 0.774 indicates that clusters contain samples from only one class.
- **Completeness**: 0.730 suggests that all members of the same class are assigned to the same cluster.
- **V-measure**: 0.751 is the harmonic mean of homogeneity and completeness, providing a balanced measure.
- **Adjusted Rand Index**: 0.826 measures the similarity between the true and predicted cluster assignments.
- **Adjusted Mutual Information**: 0.751 quantifies the agreement between the true and predicted clusters, adjusting for chance.
- **Silhouette Coefficient**: 0.777 evaluates the compactness and separation of clusters, with higher values indicating dense and well-separated clusters.
- **Calinski-Harabasz Index**: 14238.099 measures the ratio of between-cluster dispersion to within-cluster dispersion, with higher values indicating better-defined clusters

.

2. DBSCAN Clustering:

- **Estimated number of clusters**: 4 suggests the algorithm identified four distinct clusters in the data.
- **Estimated number of noise points**: 1 indicates there is one outlier or noise point.
- **Homogeneity**: 1.000 implies perfect homogeneity, where each cluster contains only members of a single class.
- **Completeness**: 0.998 indicates that all members of the same class are in the same cluster, nearly perfect.
- **V-measure**: 0.999 is very high, indicating a nearly perfect balance between homogeneity and completeness.
- **Adjusted Rand Index**: 1.000 indicates perfect similarity between true and predicted clusters.
- **Adjusted Mutual Information**: 0.999 suggests nearly perfect agreement between true and predicted clusters.
- **Silhouette Coefficient**: 0.703 evaluates the compactness and separation of clusters, slightly lower compared to K-means.
- **Calinski-Harabasz Index**: 6514.778 measures the ratio of between-cluster dispersion to within-cluster dispersion, lower than K-means but still indicative of well-defined clusters

.

3. Agglomerative Clustering:

- **Homogeneity**: 0.732 indicates good homogeneity, but slightly lower compared to K-means.
- **Completeness**: 0.749 suggests that all members of the same class are assigned to the same cluster, like K-means.
- **V-measure**: 0.740 is slightly lower than K-means and DBSCAN but still provides a balanced measure.
- **Adjusted Rand Index**: 0.844 is higher than K-means and slightly higher than DBSCAN, indicating good similarity between true and predicted clusters.
- **Adjusted Mutual Information**: 0.740 suggests good agreement between true and predicted clusters, like K-means.
- **Silhouette Coefficient**: 0.777 is comparable to K-means, indicating dense and well-separated clusters.

Conclusion

Overall, each clustering technique performs well across several assessment measures. K-means and DBSCAN are more homogeneous and complete, although agglomerative clustering has strong agreement with real clusters and similar silhouette coefficients. The algorithm used may be determined by the precise aims of the study as well as the dataset's features.

Interpretation of Results:

When assessing the findings of the clustering analysis on the Facebook live sellers dataset in Thailand, it is critical to investigate the patterns and structures discovered within the data. Each cluster represents a different group of vendors who have comparable Facebook interaction numbers and posting habits. By examining the composition and differences within each cluster, we may acquire useful insights into the dynamics of the Thai fashion and

cosmetics retail business on Facebook.

Clusters with high engagement metrics, such as comments, shares, and reactions, may suggest well-known vendors with a strong online presence and a loyal client base. In contrast, clusters with lower engagement metrics may reflect vendors that are unable to attract attention or connect with their audience effectively.

Discussion on the Effectiveness of Each Algorithm:

Each clustering algorithm (K-means, DBSCAN, Hierarchical Clustering) has its strengths and weaknesses in analyzing the Facebook live sellers dataset in Thailand:

1. **K-means**: This approach is successful in partitioning a dataset into a predetermined number of clusters. However, it presupposes clusters of comparable size and form, which may not be true in real-world data. The elbow approach can assist discover the appropriate number of clusters, but its success is dependent on the data distribution.

2. **DBSCAN**: DBSCAN is resistant to outliers and can detect irregularly formed clusters. It automatically calculates the number of clusters depending on data density, making it ideal for datasets with irregular cluster forms or variable densities. The epsilon parameter, which is calculated using k_distance_graph, is critical in determining each point's neighborhood.

3. **Hierarchical Clustering**: This program generates a hierarchical decomposition of the dataset, which may be displayed as a Dendrogram. It is beneficial for analyzing the hierarchical structure of data, but it may not scale well with huge datasets. The choice of linking method (e.g., complete, average, ward) can have a substantial influence on clustering findings.

Finally, the efficacy of any method is determined by the dataset's particular properties as well as the study' aims. Experimenting with various methods and assessing their effectiveness using proper metrics is critical for extracting relevant insights from clustering analysis.

Summary of Key Findings:

After conducting clustering analysis on the Facebook live sellers dataset in Thailand, several key findings have emerged:

1. **Distinct Clusters Identified:** Various clustering techniques, including K-means, DBSCAN, and Hierarchical Clustering, were used to identify different clusters within the dataset. These clusters include of Facebook live vendors who have similar engagement trends and content kinds.

2. **Engagement Patterns:** The clustering study indicated varying participation patterns among Facebook Live vendors. Some clusters may have better engagement metrics, such as comments, shares, and reactions, suggesting a more active and responsive user base. Others may have lower levels of engagement, indicating possible areas for development in their marketing strategy.

3. **Content Preferences:** The clustering technique also revealed information on the sorts of content chosen by various groupings of Facebook Live merchants. This covers differences in the frequency and types of posts (e.g., videos, images, status updates, links), as well as how different content categories interact with engagement metrics.

4. **Optimal Cluster Sizes:** The appropriate cluster sizes were identified using approaches such as the elbow method for K-means and k_distance_graph for DBSCAN, resulting in a more successful dataset segmentation.

Implications of the Results:

1. **Marketing Strategies:** The results of this clustering study might be useful for firms in Thailand's fashion and cosmetics retail industry. Businesses may better understand their target audience's interaction habits and content preferences by tailoring their marketing strategy accordingly.

2. **Audience Engagement:** Better analysis of audience engagement may help Facebook Live merchants improve their content generation and delivery tactics. Sellers can possibly boost their reach and effect on social media platforms by concentrating on high-engagement content kinds and subjects.

3. **Competitive Analysis:** Clustering analysis may also help with competitive analysis by highlighting prominent rivals within each cluster. This allows firms to compare their performance to similar market participants and discover opportunities for differentiation and development.

Suggestions for Future Work:

1. Further Feature Engineering: Investigating new or designed features may give more in-depth information about consumer behavior and preferences. This might involve sentiment analysis of comments, study of post timing and frequency, or the use of other data sources such as demographic information.

2. Temporal Analysis: Analyzing how engagement patterns change over time may reveal useful information about seasonal trends, advertising efficacy, and long-term audience engagement dynamics.

3. Integration with Sales Data: By combining engagement metrics with sales data, you can conduct a more thorough examination of the impact of social media marketing activities on business outcomes like revenue and client acquisition.

4. Predictive Modeling: Creating predictive models based on clustering data can help firms foresee future engagement levels and modify marketing campaigns accordingly. This may entail the use of machine learning methods such as regression or time series analysis.

Overall, the conclusions of this clustering study are useful for firms in Thailand's fashion and cosmetics retail industry that use Facebook Live. Businesses may improve their marketing effectiveness and promote higher engagement and success on social media platforms by exploiting these insights and pursuing new research routes.

References:

1. DATASET Source:
   UCI Machine Learning Repository
   https://archive.ics.uci.edu/dataset/488/facebook+live+sellers+in+thailand

2. Udemy course – Machine Learning – A Z by Kirill Eremenko and Hadelin de Ponteves

3. https://en.wikipedia.org/wiki/Cluster_analysis

4. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

5. https://developers.google.com/machine-learning/clustering/clustering-algorithms

6. https://www.datascience.com/blog/k-means-clustering

7. https://acadgild.com/blog/k-means-clustering-algorithm

8. https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

9. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

10. Data Clustering: A Review by A.K. Jain, M.N. Murty and P.J. Flynn.

11. Clustering with Gaussian Mixtures by Andrew W. Moore.

12. Kernel k-means, Spectral Clustering and Normalized Cuts by Inderjit S. Dhillon, Yuqiang Guan and Brian Kulis.

13. A Comprehensive Overview of Basic Clustering Algorithms by Glenn Fung.

14. An Efficient k-means Clustering Algorithm: Analysis and Implementation by Tapas Kanungo, David M. Mount,

15. Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu.
    p