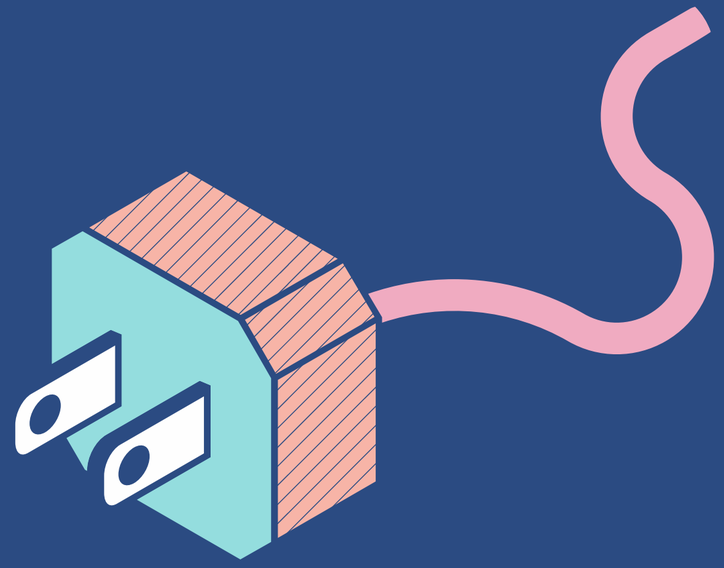
Unveiling Customer Dynamics: A Deep Dive into Credit Card Transaction Behavior

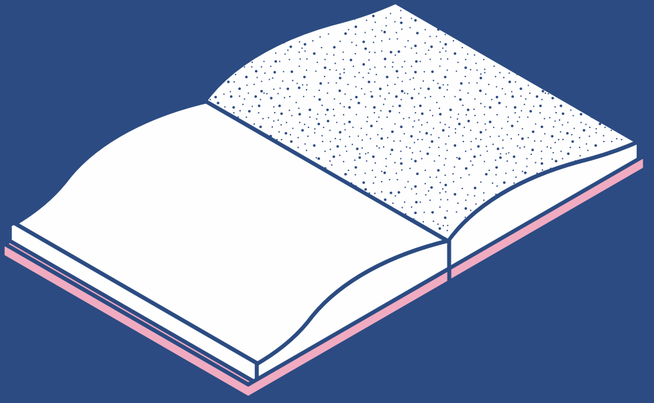
PROBLEM STATEMENT

A credit card company has collected data about its customers, including various features such as balances, purchase behavior, cash advances, credit limits, and more. The company wants to gain meaningful insights from this data and devise strategies to increase credit card sales and revenue.

OBJECTIVE

Identify distinct customer segments based on credit card transaction behavior.



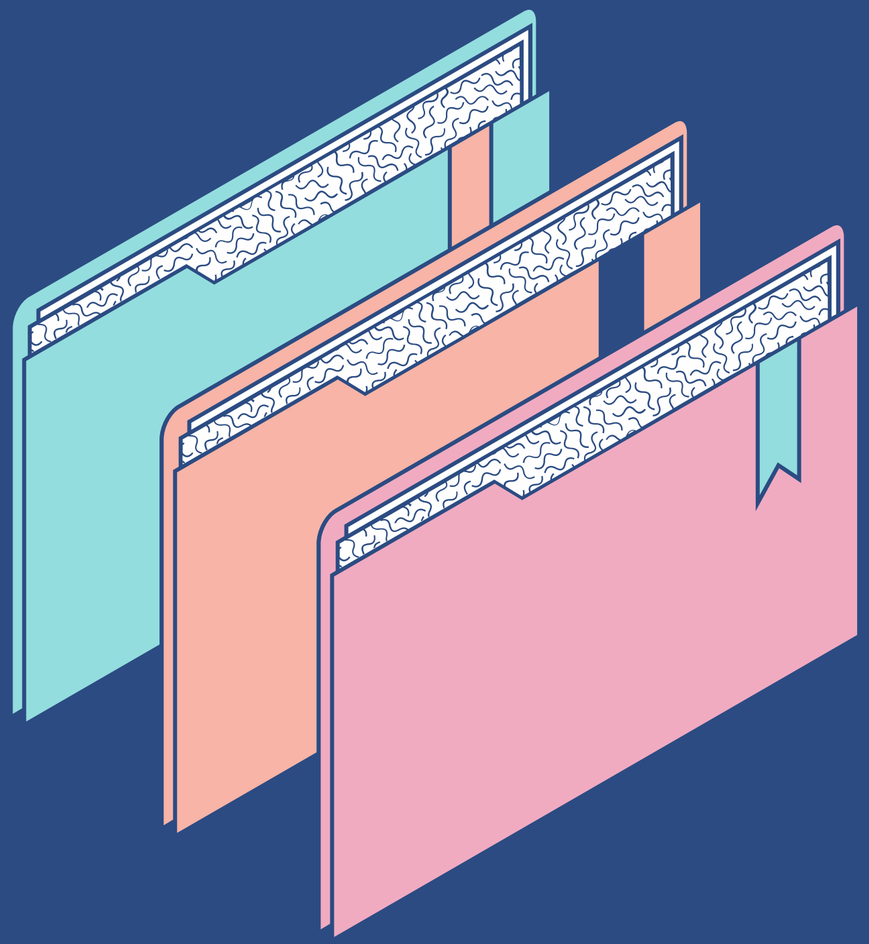


DATA

- CUST_ID: Unique customer identification number.
- BALANCE: Total amount owed or owed to a vendor, in local currency.
- BALANCE_FREQUENCY: Score indicating how frequently the balance is updated (0-1).
- PURCHASES: Total amount of purchases made from the account.
- ONEOFF_PURCHASES: Maximum purchase amount made in one transaction.
- INSTALLMENTS_PURCHASES: Total amount of purchases made in installments.
- CASH_ADVANCE: Cash in advance taken by the user.
- PURCHASES_FREQUENCY: Score indicating how frequently purchases are made (0-1).
- ONEOFF_PURCHASES_FREQUENCY: Frequency of non-regular plan purchases (0-1).
- PURCHASES_INSTALLMENTS_FREQUENCY: Frequency of installment purchases (0-1).
- CASH_ADVANCE_FREQUENCY: Frequency of cash advances (0-1).
- CASH_ADVANCE_TRX: Number of cash advance transactions in the last 12 months.
- PURCHASES_TRX: Number of purchase transactions in the last 12 months.
- CREDIT_LIMIT: Maximum amount customers can borrow from their credit card.
- PAYMENTS: Total payments made to the credit card in the last 12 months.
- MINIMUM_PAYMENTS: Minimum payments required in the last 12 months.
- PRC_FULL_PAYMENT: Percentage of months with full payment in the last 12 months.
- TENURE: Duration of credit card usage in months.

LINK

8950 entries
18 COLUMNS

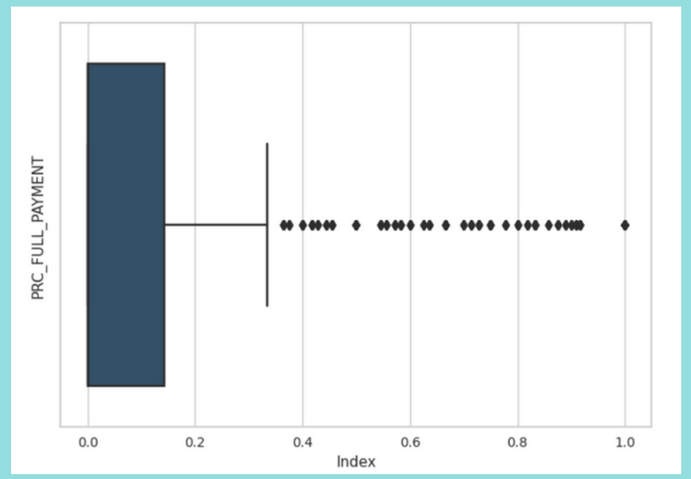


Agenda

- Data Preprocessing
- Exploratory Data Analysis (EDA)
- Clustering Model Implementation
- Comparative Analysis
- Dimension Reduction
- Clustering Analysis

Data Preprocessing

1.Importing the libraries
2- Reading the dataset
3- Cleaning and pre-processing the DataFrame
4- DataFrame visualizations

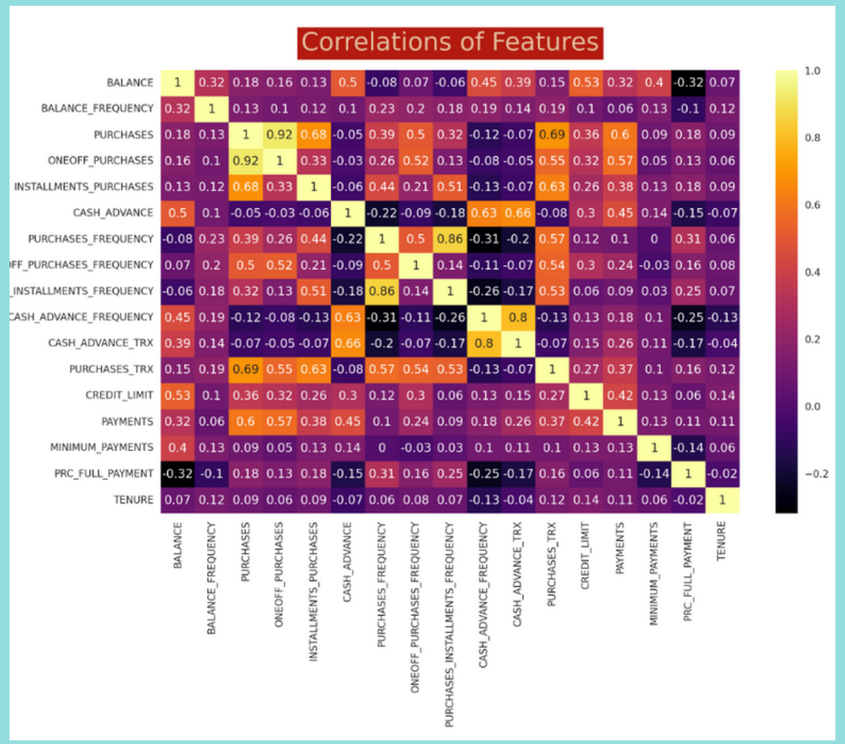


- **Missing Values Detection:**
  - Identified columns with missing values: CREDIT_LIMIT and MINIMUM_PAYMENTS.
  - Columns exhibiting high correlations: ONEOFF_PURCHASES with PURCHASES, PURCHASES_INSTALLMENTS_FREQUENCY with PURCHASES_FREQUENCY, and CASH_ADVANCE_TRX with CASH_ADVANCE_FREQUENCY.
  - Limited linear correlation observed between MINIMUM_PAYMENTS and other features, making Linear Regression impractical for filling missing values.
- **Missing Values Imputation:**
  - Utilized KNNImputer to fill missing values based on the nearest neighbors approach.
- **Duplicated Data:**
  - No duplicated data observed (duplicated data count = 0).
- **Noise Detection:**
  - Outliers are present in the dataset
  - However, no obvious noisy data detected, hence no data removal performed.

EXPLORATORY DATA ANALYSIS

1Pattern recognition

1) Univariate Analysis
2) Bivariate Analysis
3) Multivariate Analysis



- Perform exploratory data analysis (EDA) to understand the data better.

- Univariate Analysis:

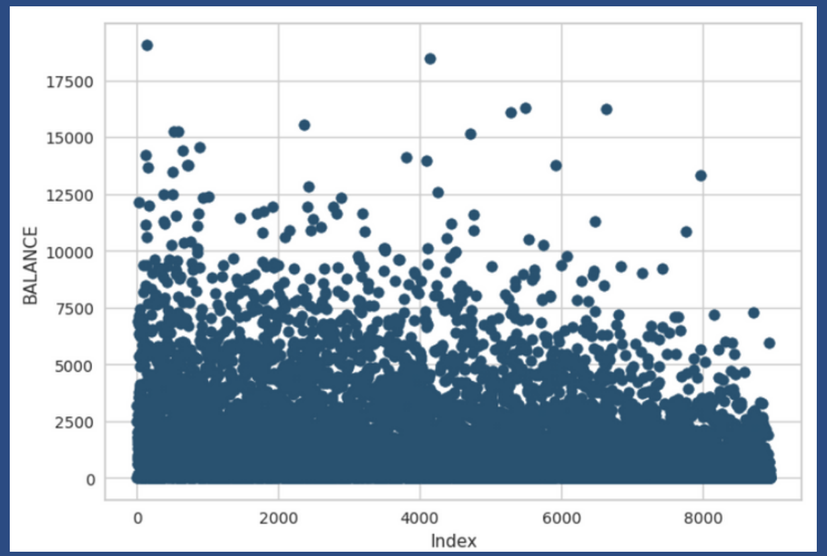
Examine individual features (e.g., histograms, box plots).

- Bivariate Analysis:

Explore relationships between pairs of features (e.g., scatter plots, correlation matrices).

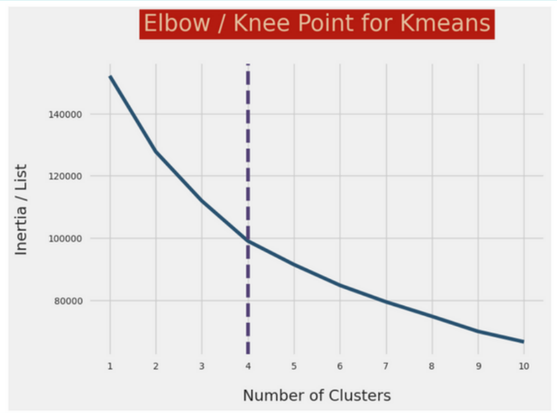
- Multivariate Analysis:

Consider interactions among multiple features (e.g., heatmaps).



Clustering Model Implementation

1)K-means
2) Mini Batch K-means
3) Gaussian Mixture Model (GMM)
4) Mean Shift
5) Affinity Propagation
6) Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)
7) Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
8) Hierarchical Clustering

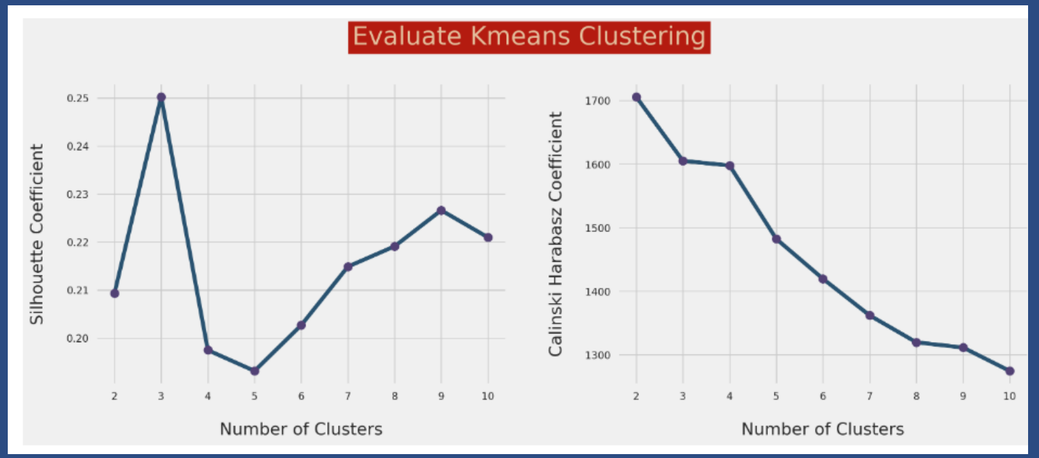


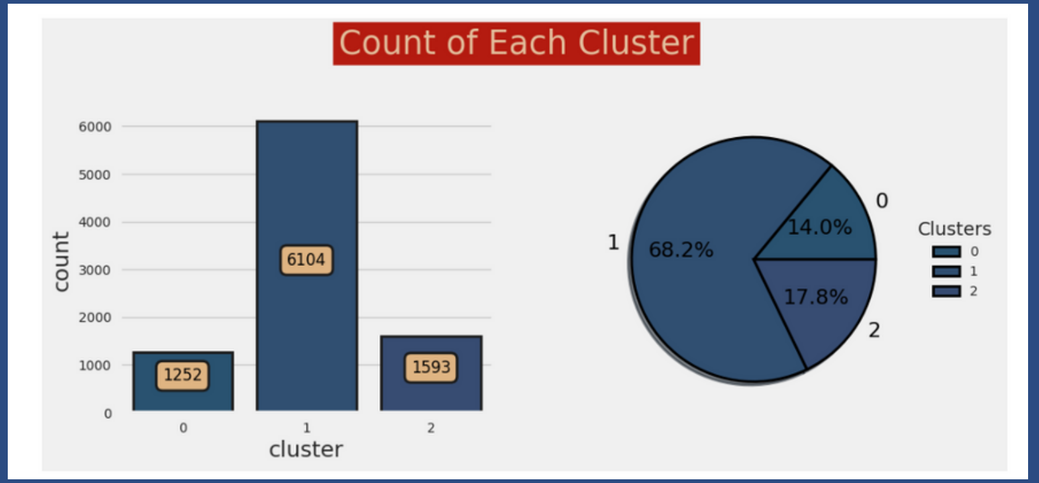
K=4

K-MEANS

- K-Means clustering partitions data into k groups based on closest mean values (centroids).
- It's used for unsupervised learning tasks like customer segmentation and image compression.

K=3

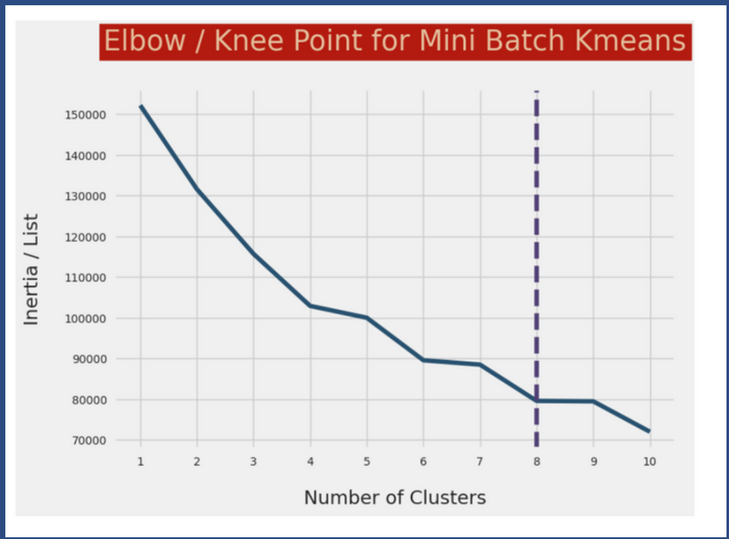




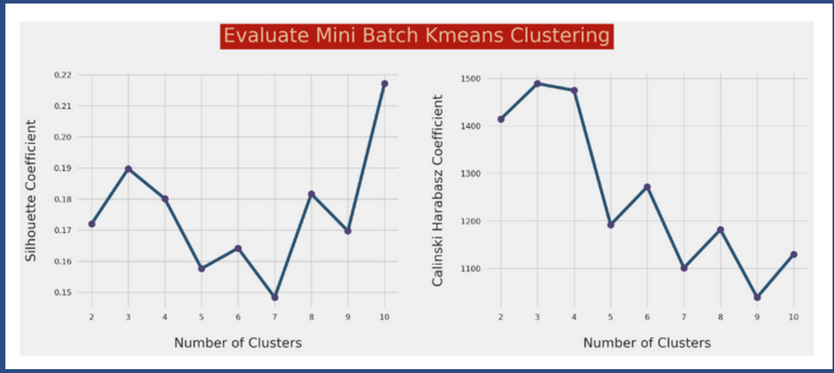
MINI-BATCH KMEANS

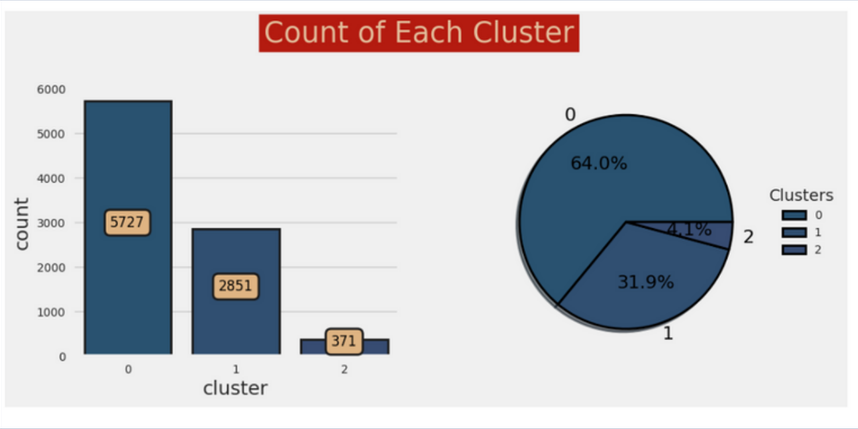
1. Mini Batch K-Means handles large datasets by randomly selecting mini-batches for centroid updates.
2. It prioritizes computational efficiency over cluster quality, making it suitable for memory-constrained or time-sensitive scenarios.

K=8



K=3

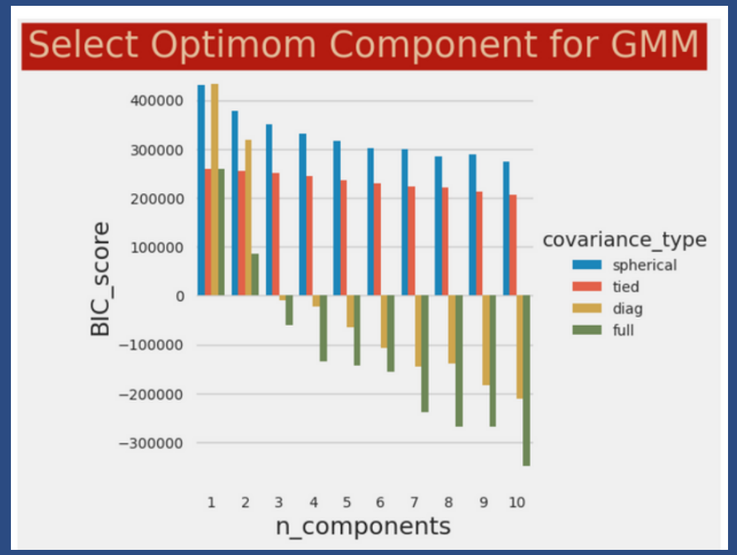


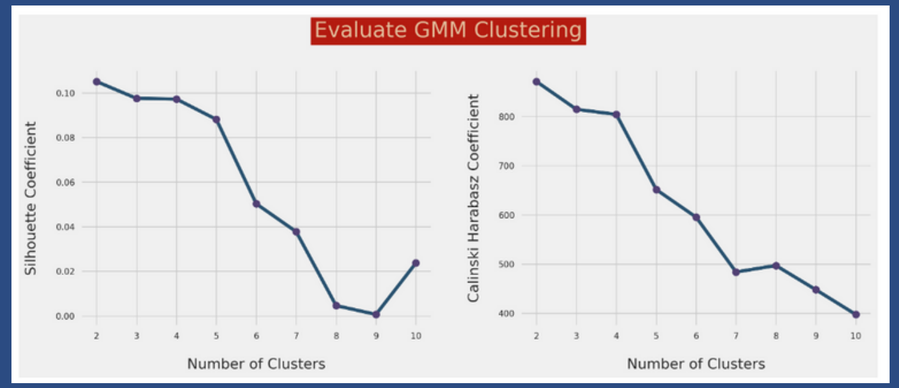


Gaussian Mixture Model (GMM)

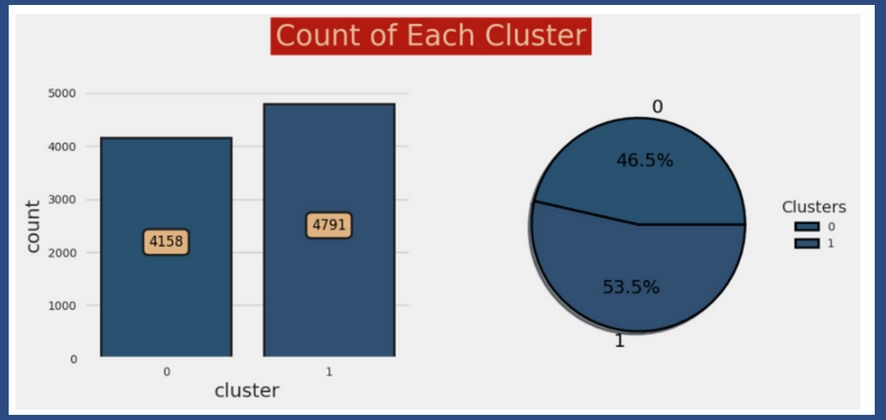
1. GMM (Gaussian Mixture Model) assumes data is from a mixture of Gaussian distributions, offering more flexibility than k-means by assigning probabilities to each cluster for every data point.
2. Implemented using sklearn.mixture.GaussianMixture in Python, allowing customization of components, covariance type, and fitting to data with methods for prediction and sampling.

n-components=10





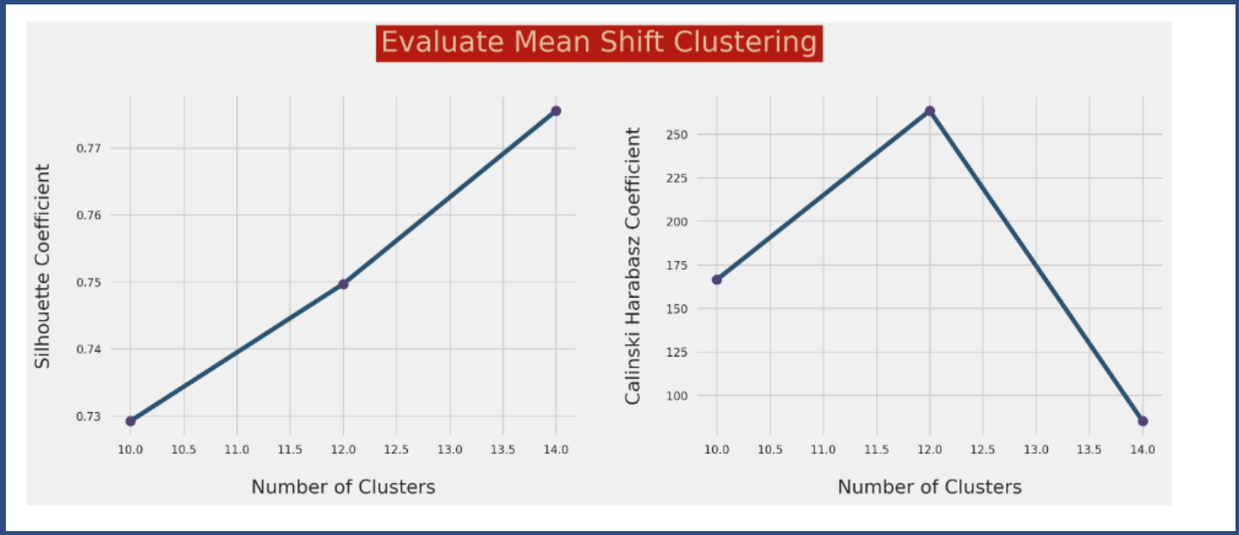
n-components=2

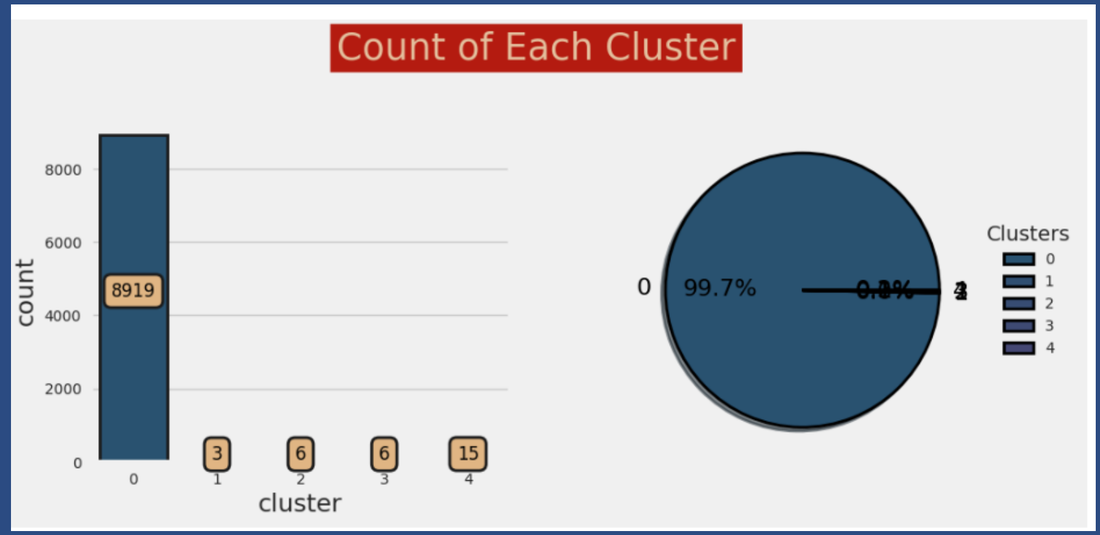


Mean Shift

1. Mean Shift clustering is density-based, requiring no predefined number of clusters, shifting data points towards local density peaks iteratively.
2. Handles arbitrary cluster shapes/sizes but sensitive to radius and kernel function choices.

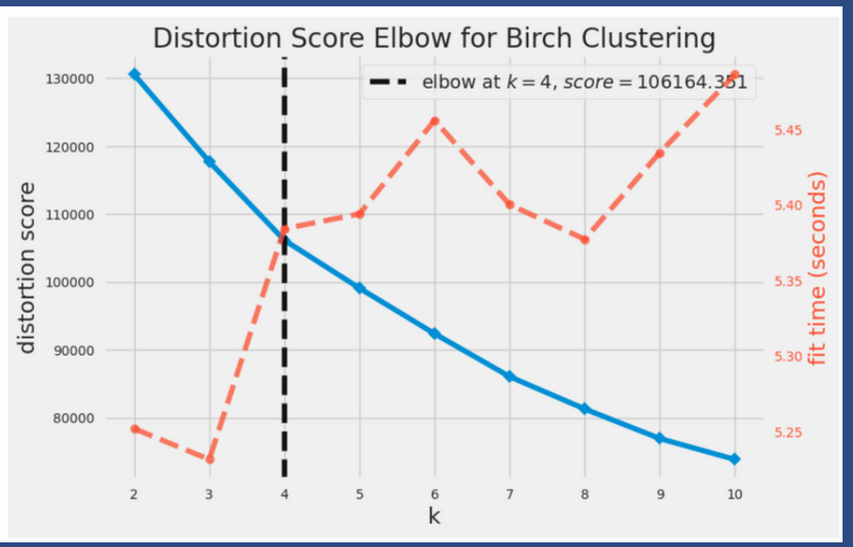
bandwidth=12



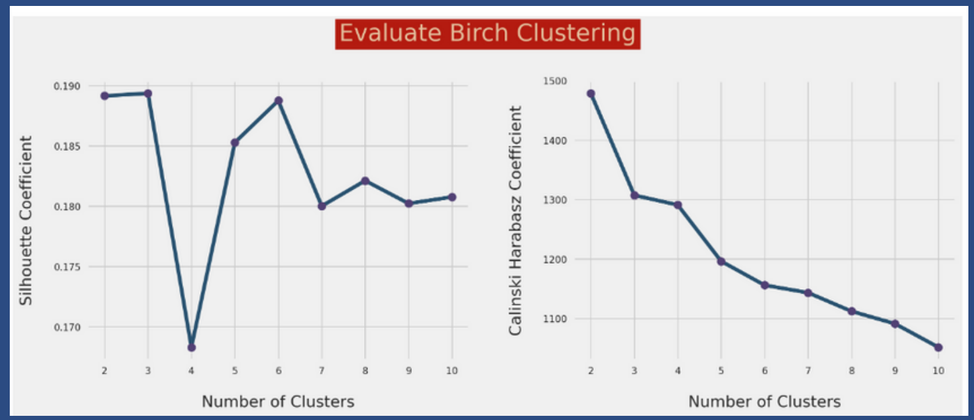


Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)

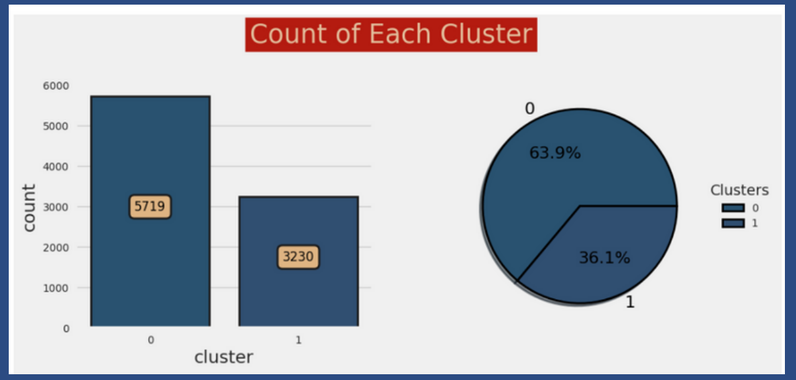
1. BIRCH creates a compact summary of large datasets using a CF tree, preserving data information with sub-clusters represented by clustering features (CF).
2. Assigns data points to nearest sub-cluster, potentially accelerating other clustering methods like k-means or Gaussian mixture models.



K=4

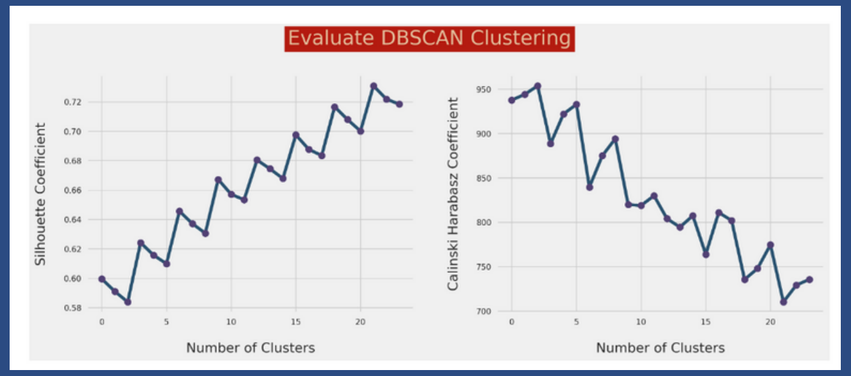


K=2

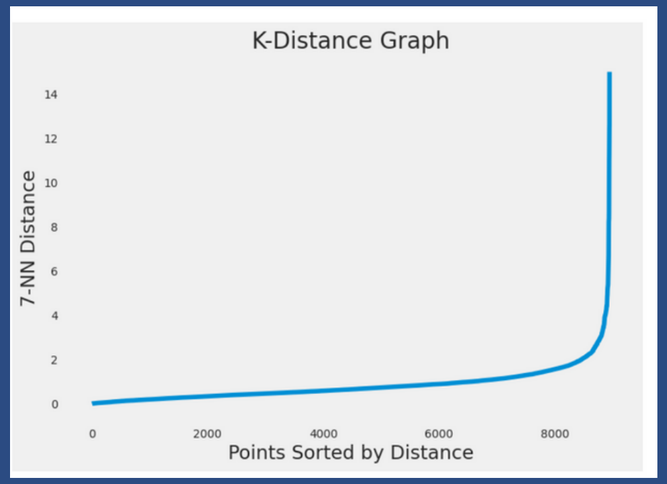


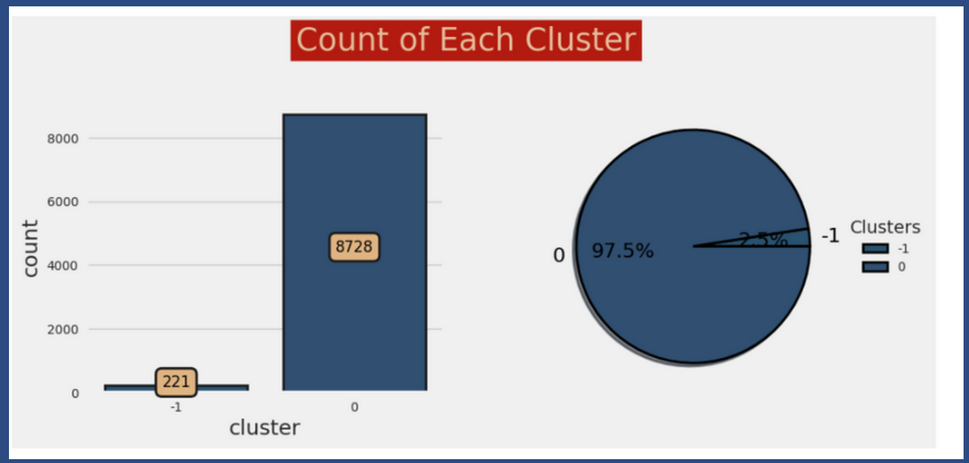
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

1. DBSCAN clusters based on proximity and density, identifying high-density regions while marking outliers.
2. No need for pre-set cluster counts, handles various cluster shapes/sizes, but sensitive to radius and metric parameter choices



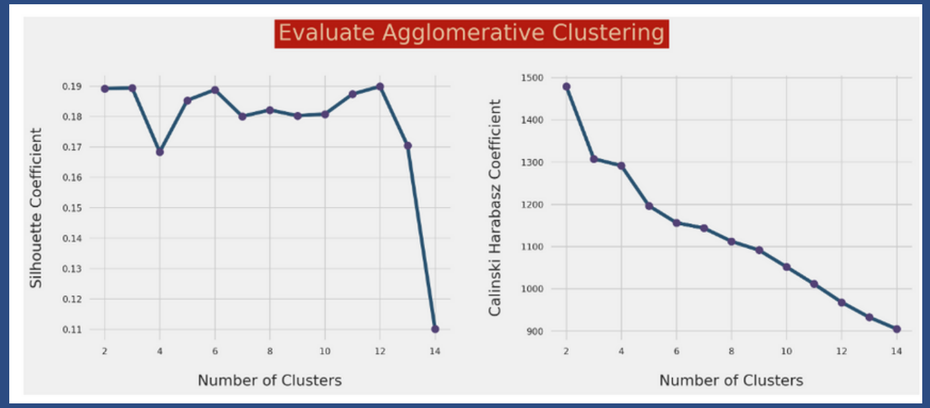
eps=3.5 and min_samples=40

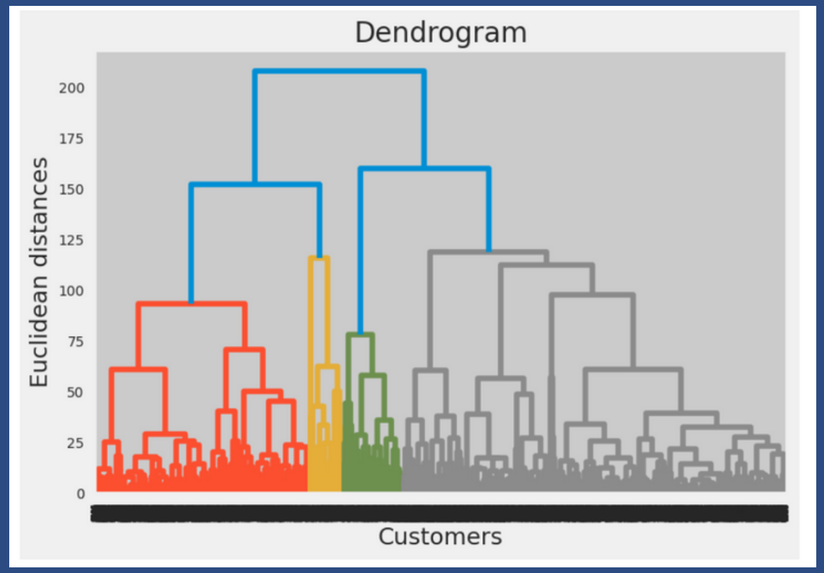




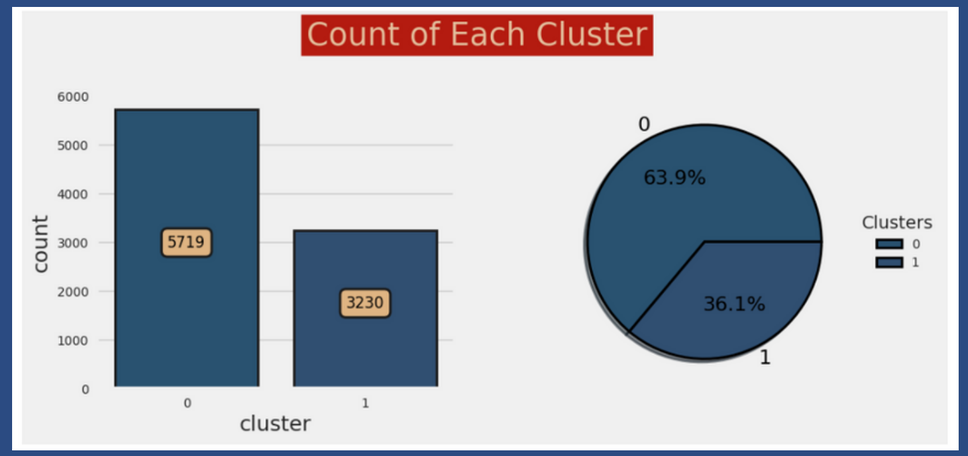
Hierarchical Clustering

1. Hierarchical clustering builds a cluster hierarchy based on data point similarity/dissimilarity.
2. Can be agglomerative (merging smaller clusters) or divisive (splitting larger clusters), resulting in a dendrogram visualization.

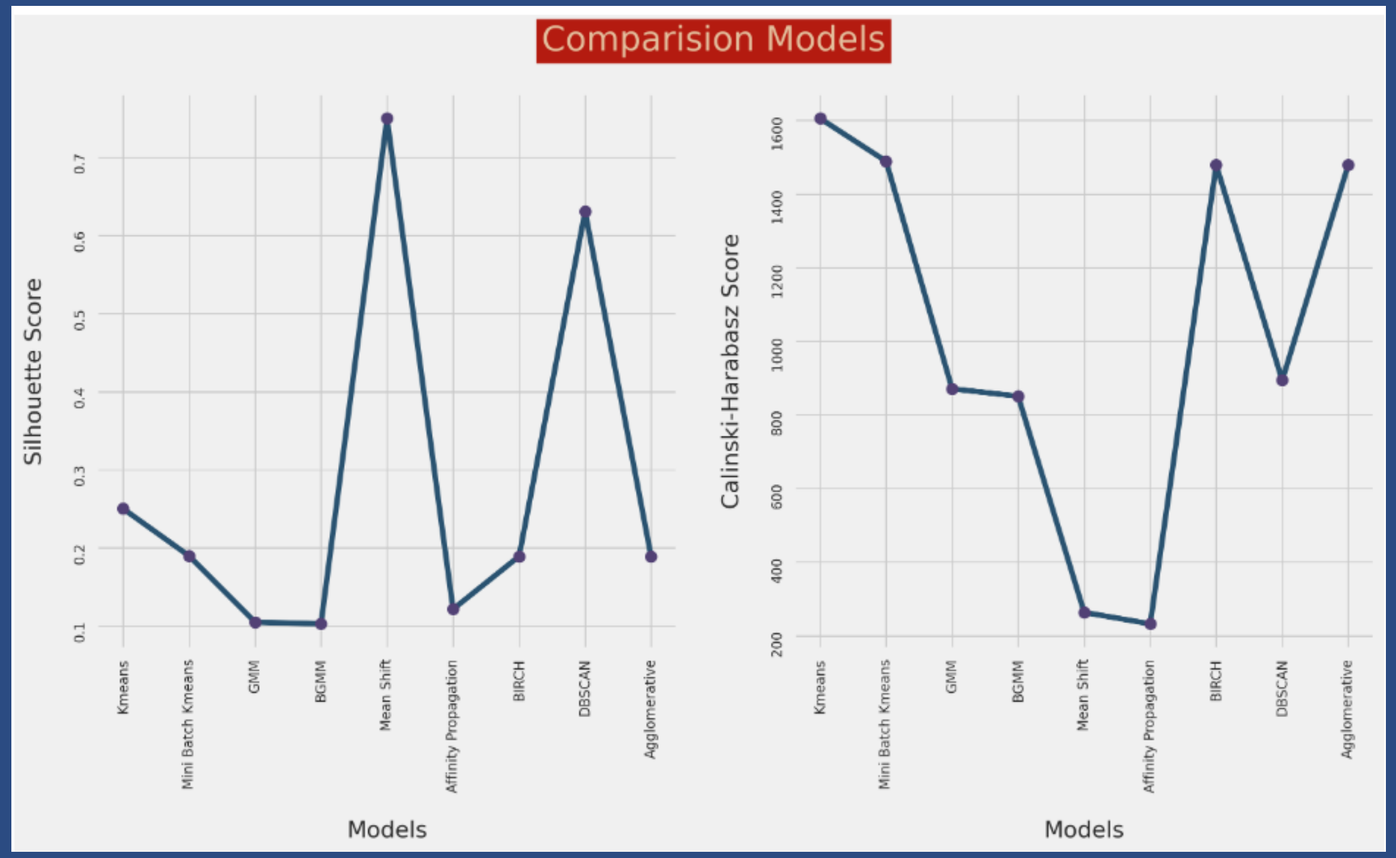




n_clusters=2



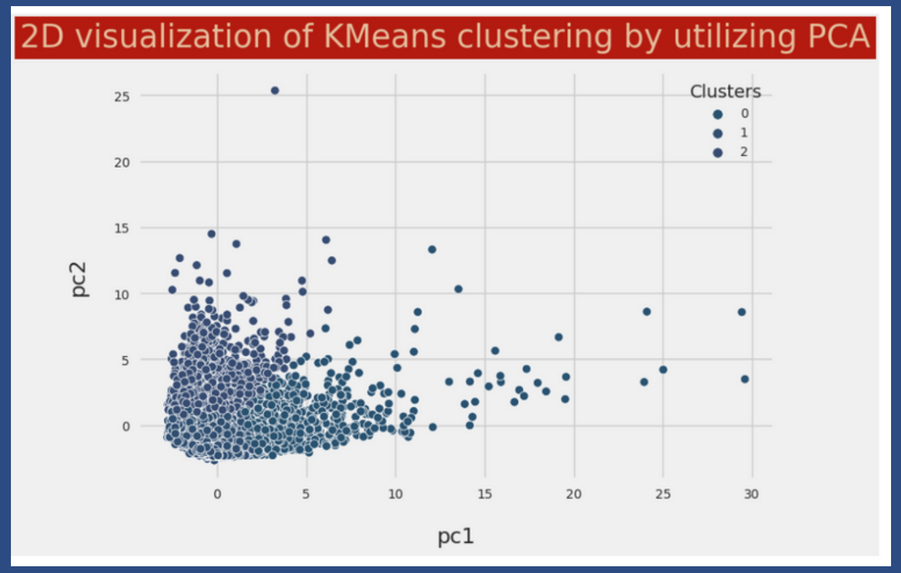
COMPARING CLUSTER MODEL

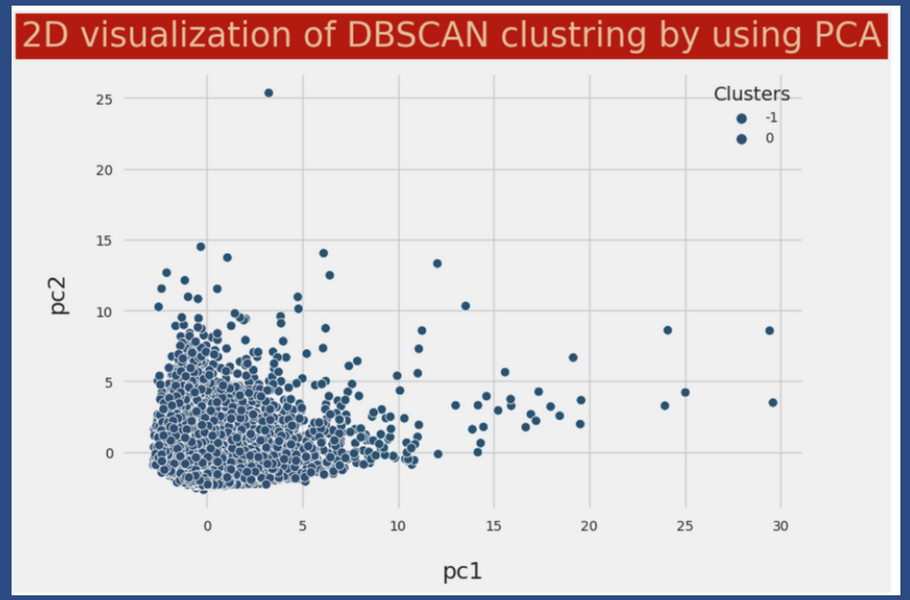


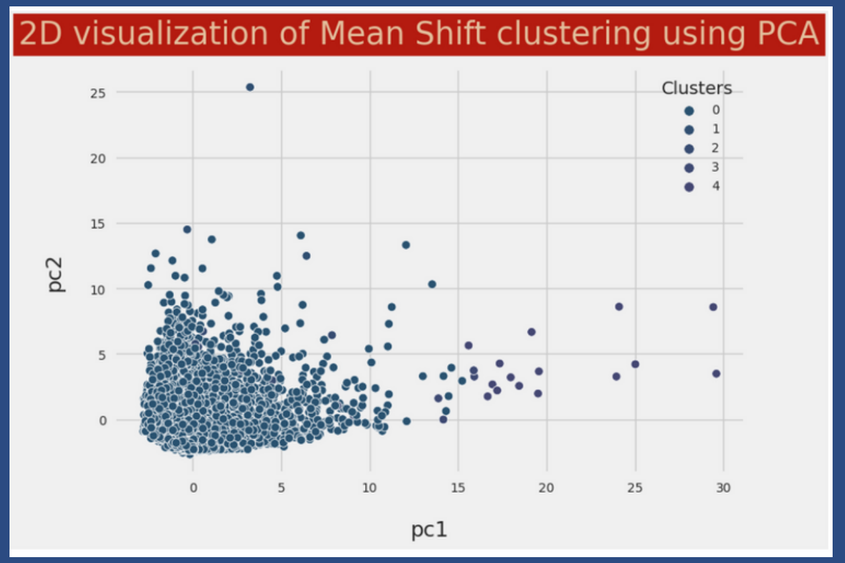
1. Mean Shift excels based on the Silhouette criterion, while KMeans performs best according to the Calinski-Harabazs criterion.
2. DBSCAN shows promise across both criteria compared to other models.
3. 3Hence, Mean Shift, KMeans, and DBSCAN are recommended for further analysis.

DIMENSIONALITY RESUCTION

PCA

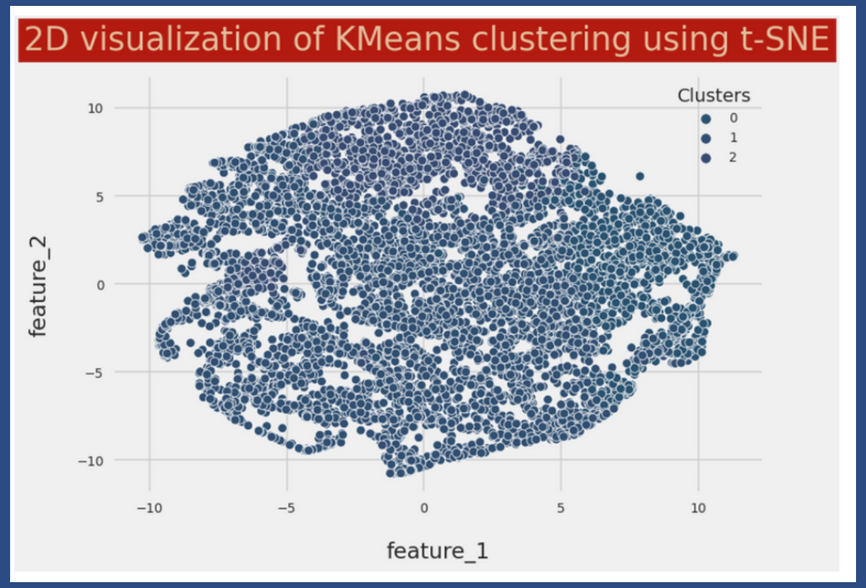


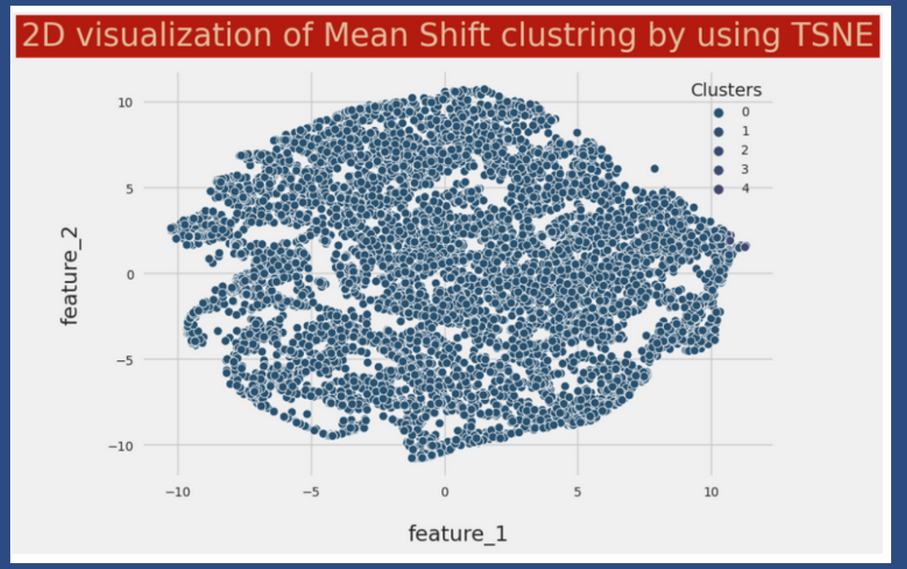


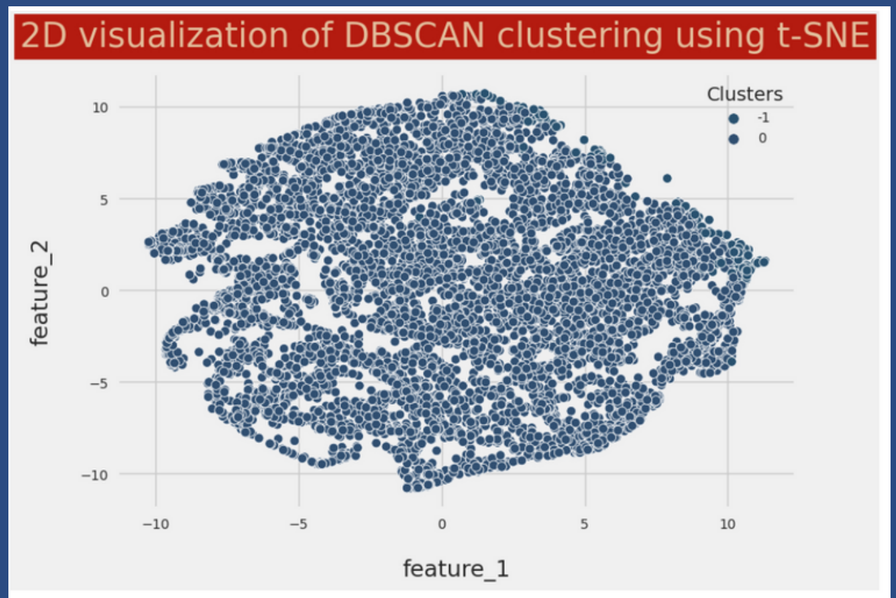


DIMENSIONALITY RESUCTION

t-SNE







CLUSTERING ANALYSIS

K-MEANS

- Customers in clusters 1 and 2 exhibit lower balances, while those in cluster 0 maintain higher balances.
- Cluster 0 customers demonstrate high purchase frequency, both in one-off and installment purchases, whereas cluster 2 customers show low frequency in both categories.
- Customers in cluster 1 typically have lower credit limits compared to other clusters.
- Cluster 2 customers often engage in higher cash advance frequency compared to customers in other clusters.

1. Cluster 0: Customers with average balance frequently making purchases (one-off or installment).
2. Cluster 1: Customers with low balance and credit limit, less frequently engaging in one-off purchases.
3. Cluster 2: Customers with low balance, infrequently making purchases (one-off or installment), but frequently making cash advances.

CLUSTERING ANALYSIS

K-MEANS

Cluster -1:

1. Customers with average balance.
2. Higher average purchase frequency.
3. Moderate average credit limit.

Cluster 0:

1. Customers with lower average balance.
2. Lower average one-off purchases frequency and installment purchases frequency.
3. Lower average credit limit.

CONCLUSION

In Cluster -1, customers with an average balance and low credit limits are observed to be actively involved in frequent purchases, suggesting a consistent spending behavior. Meanwhile, Cluster 0 reveals a subgroup of customers with both low balances and credit limits, indicating a more conservative approach to making one-off or installment purchases, infrequently engaging in such transactions. These nuanced insights into distinct financial profiles enhance our understanding of customer behavior within the dataset.

THANK YOU

repository