BDA Final Project Report

Derrick Maverick - 潘偉傑 - b12902096

This project aims to "clusterize" the given .csv files into several clusters, namely $4n-1$ clusters, based on the n features present in the datasets.

For this, I utilized a hybrid method that combines HDBSCAN and GMM with tied covariance to produce the final labels.
The working pipeline is as follows:

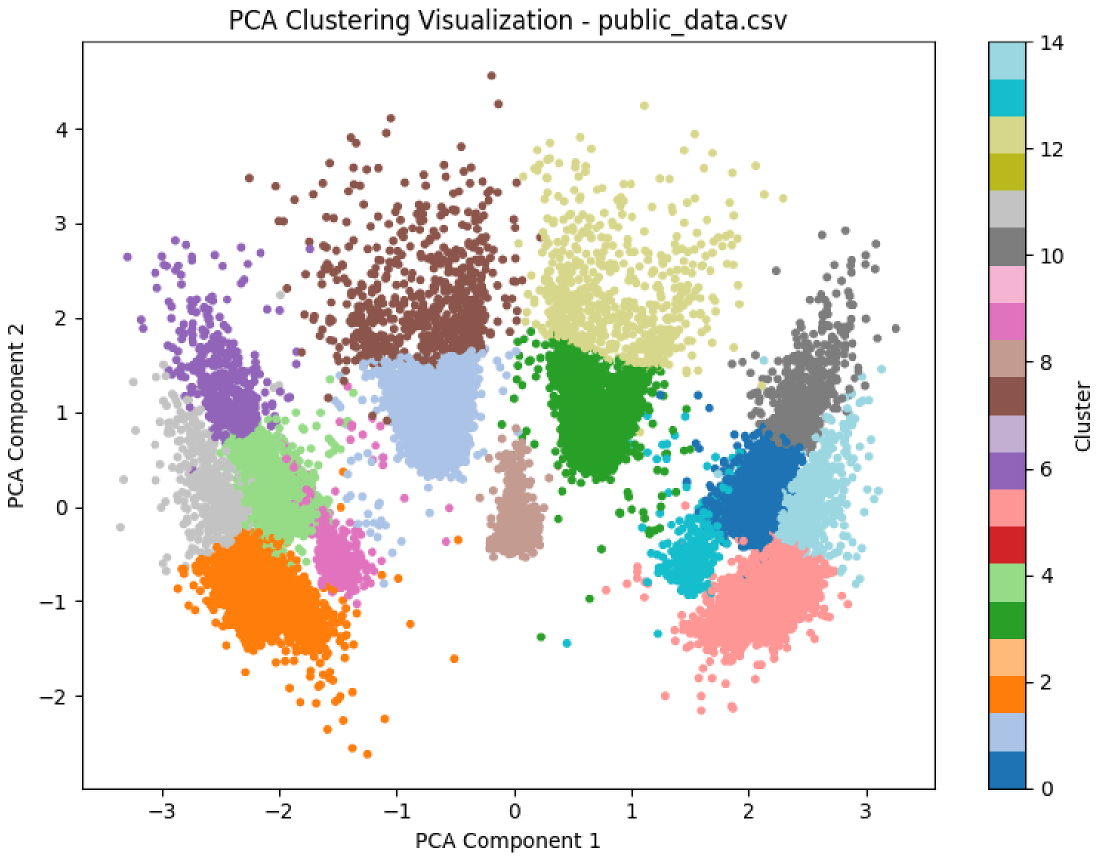
1. Data Preprocessing
   Each set of data is loaded using `pandas` and using the **Yeo-Johnson Power Transformation**, we handle the non-Gaussian distribution and normalize the skewed data. Followed by **Standard Scaling** to further ensure zero mean, unit-variance scaling, which is critical for distance-based models like GMM.

2. Data Clustering
   HDBSCAN is used to first estimate the underlying cluster structure present within the data. Using this, we prevent overfragmentation and balance sensitivity and noise robustness. However, since we cannot directly control the output number of clusters with HDBSCAN, we pass the results to another clustering method, **Gaussian Mixture Model (GMM)** with **tied** covariance to fit the number of clusters into the desired $4n-1$.

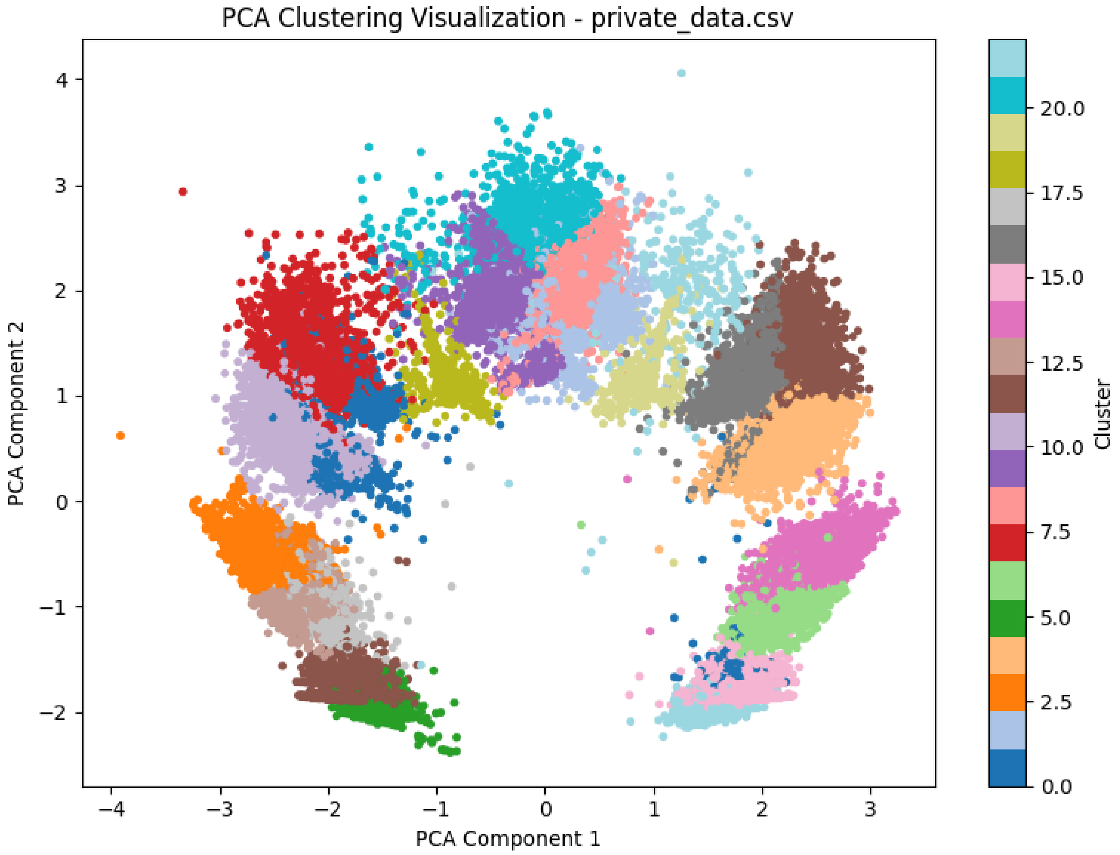
3. Conclusion
   This hybrid approach leverages the density-based intuition of HDBSCAN and the flexibility of GMM to effectively cluster real-world data. Preprocessing significantly enhanced cluster quality, and PCA visualization confirmed well-separated groups. Future work may explore semi-supervised refinement or cluster quality metrics like silhouette scores.

4. Visualization

Public Clusters



Private Clusters


PCA Clustering Visualization - private_data.csv
PCA Component 1
PCA Component 2
Cluster