Unsupervised Machine Learning for Customer Market Segmentation

---

PROJECT 2 -

YORK UNIVERSITY
SCHOOL OF CONTINUING STUDIES

Suha Islaih
Osear Okinga S

PROJECT - OVERVIEW & OBJECTIVES

Goals

The objective of this study is to develop a customer segmentation using unsupervised machine learning approach aiming to perform an effective marketing strategies that reflect customer behavior.

Main Research Question

- Define/ Conduct the data segmentation analysis & visualize the factors that lead to a better understanding of our dataset.
- Define / What model can we use to perform and to give us a better customer segmentation.
- What’s the metrics can be used to evaluate the performance of a clustering model.

METHODOLOGY FRAMEWORK

The approach methodology framework that we will use in this project is call CRISP-DM, which is an industry standard process for data mining.

This framework is use & will guide to perform different tasks in the projects

DATA UNDERSTANDING - PREPROCESSING & PREPARATION

Action :

- Understnading of the
  - Data shape, Size, Info and Types
  - Duplicates values
  - Statistcial Dataset

- Drop Variable of no impact in the analysis

- Handle the Missing Values

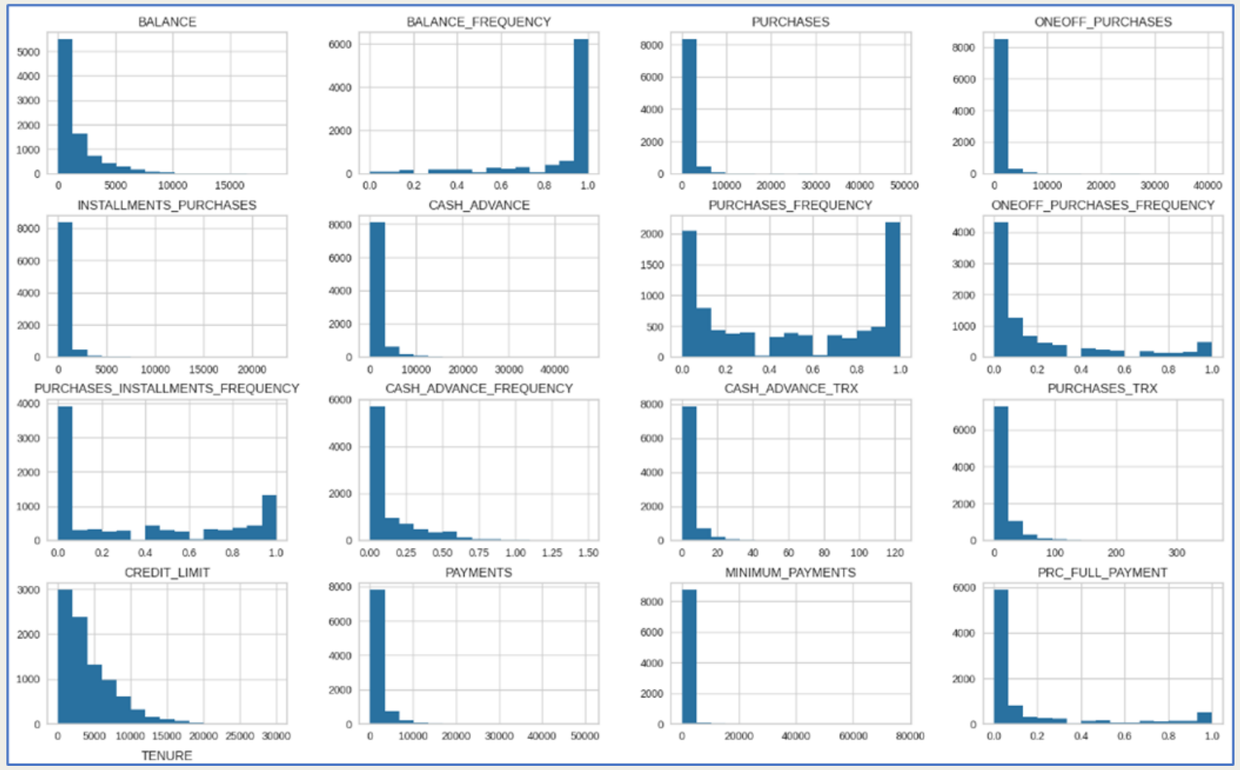
EXPLORATORY DATA ANALYSIS (EDA)

Data Visualisation of different plot :

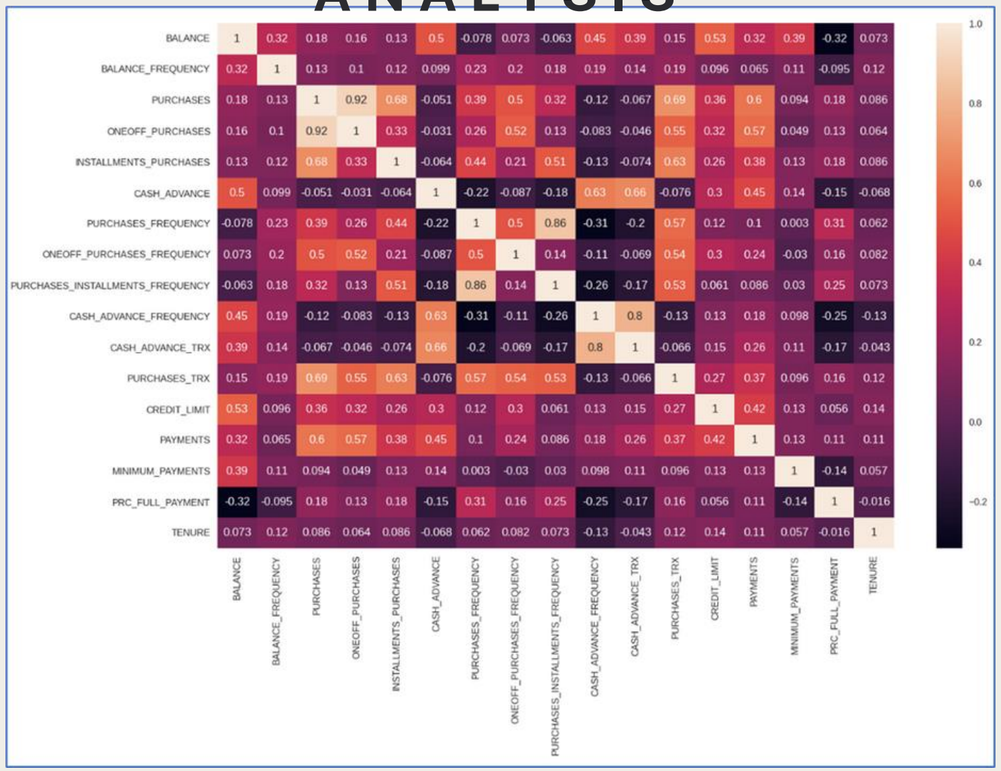
- Pairplot
- Visualize and Analyze - Variables
- Histograms Plot - Dataset
- Correlation Analysis

EXPLORATORY DATA ANALYSIS (EDA)

DATASET – HISTOGRAMS PLOT



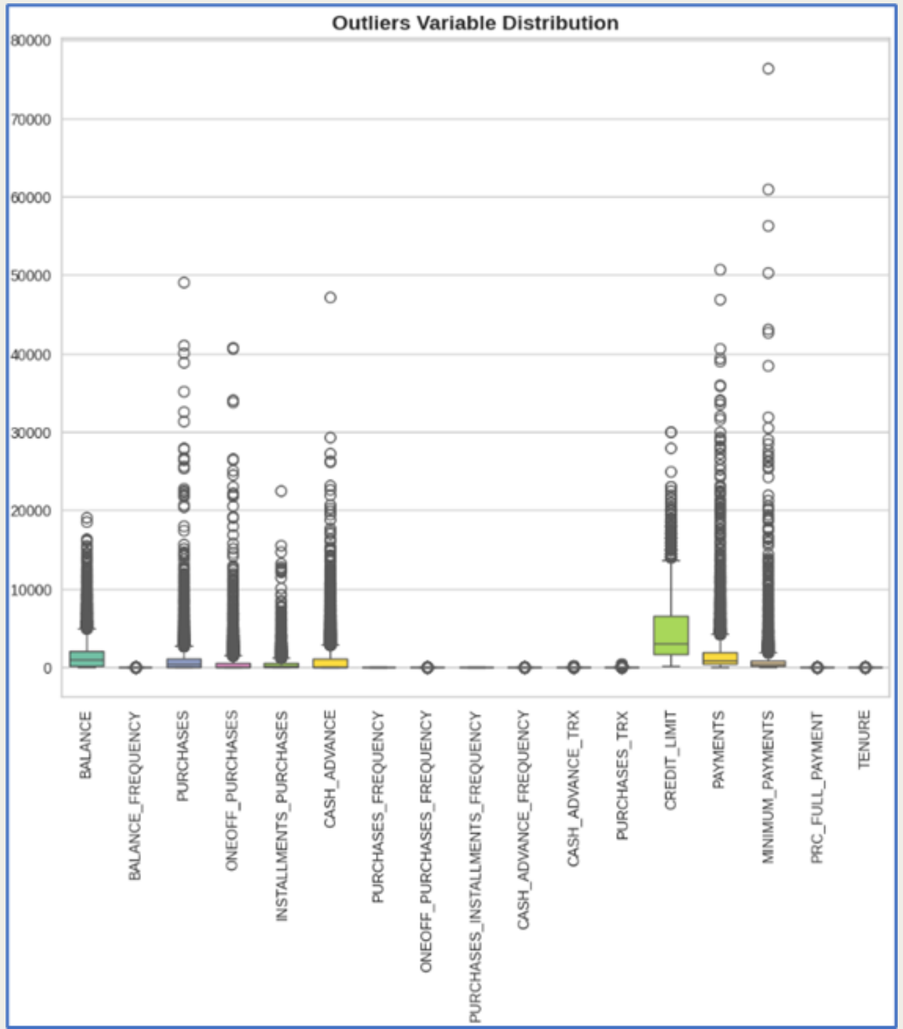
CORRELATION ANALYSIS



OUTLIER ANALYSIS

Identification & Visualize & Removal of the Outlier's Distribution

The approach for outlier analysis involves first to identify and visualize the outlier's distribution, as illustrated in the plots below; then remove it accordingly.



MODELLING

Feature Scaling

The Approach for Scaling the Numerical Features - Standardize the Data

```
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df)
```

Type of Model - Selection

- K-Means model
- Gaussian Mixture Models (GMM)
- Hierarchical Clustering

MODELLING K-MEANS APPROACH

Determine the Optimal Number of Clusters (k) - Elbow Method

This the approach used to identify the optimal number of clusters for the K-means clustering algorithm

Apply of PCA

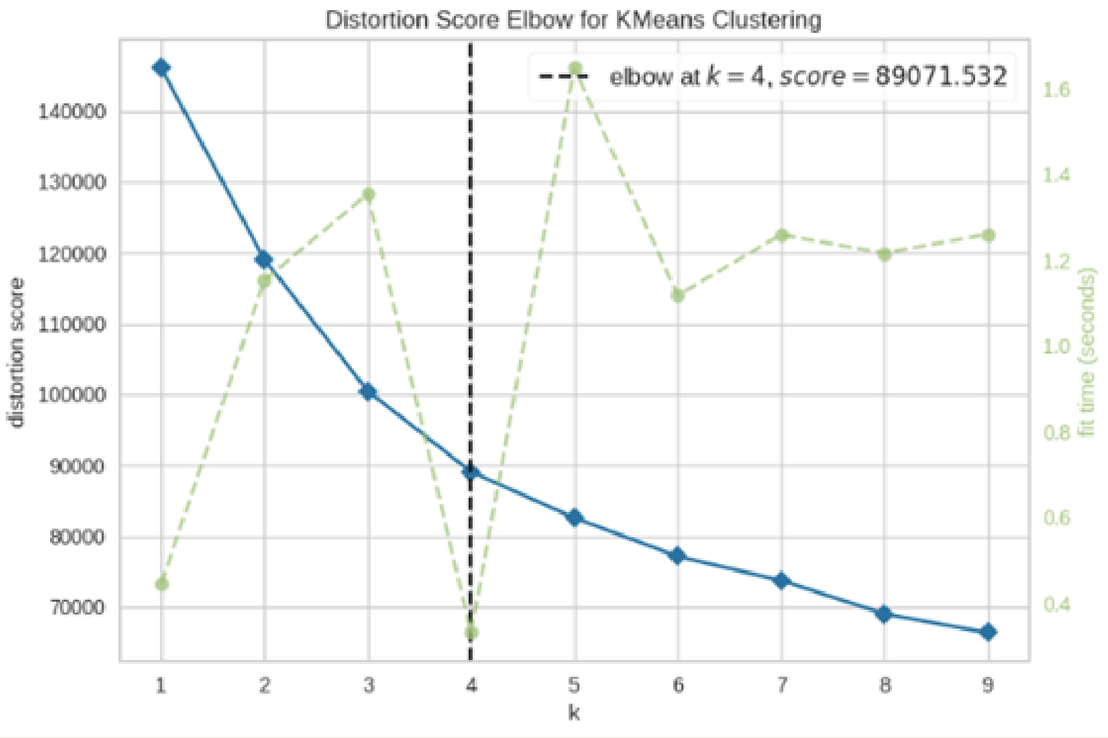
To reduces high-dimension data to smaller dimensions

Model Training / Run K-Means Clustering

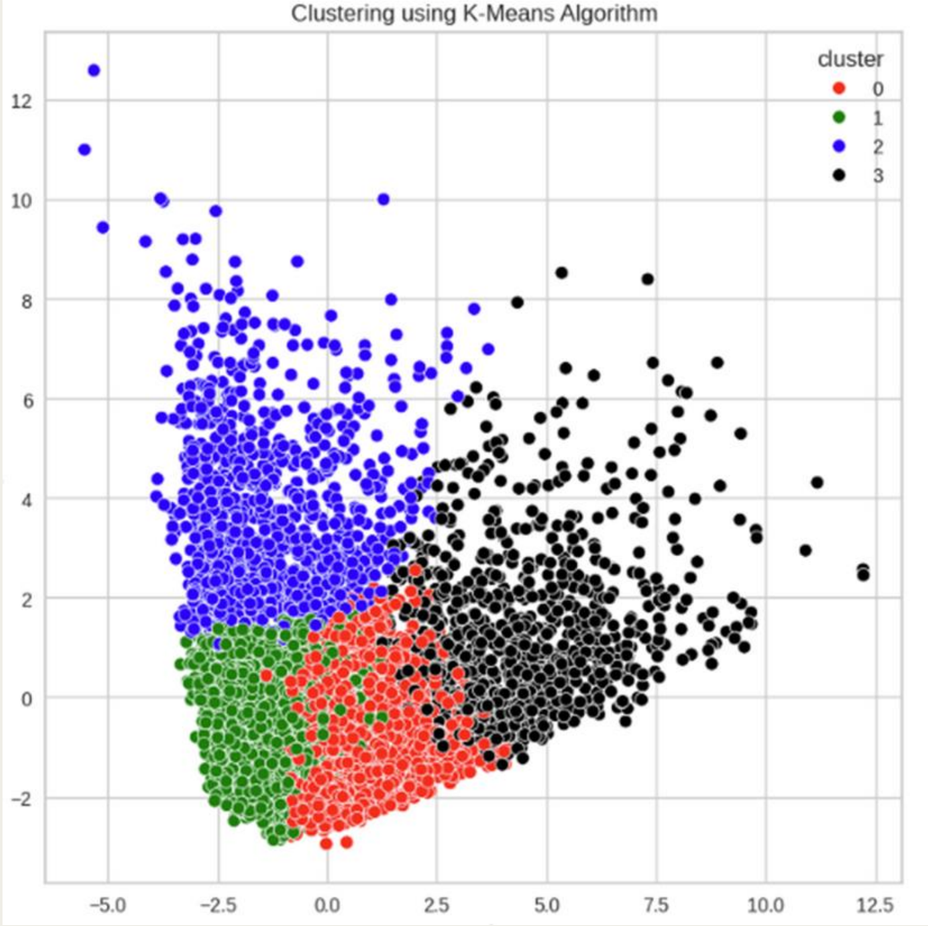
Analyze the Clusterseans Clustering

MODELLING K-MEANS

Defining Elbow Method

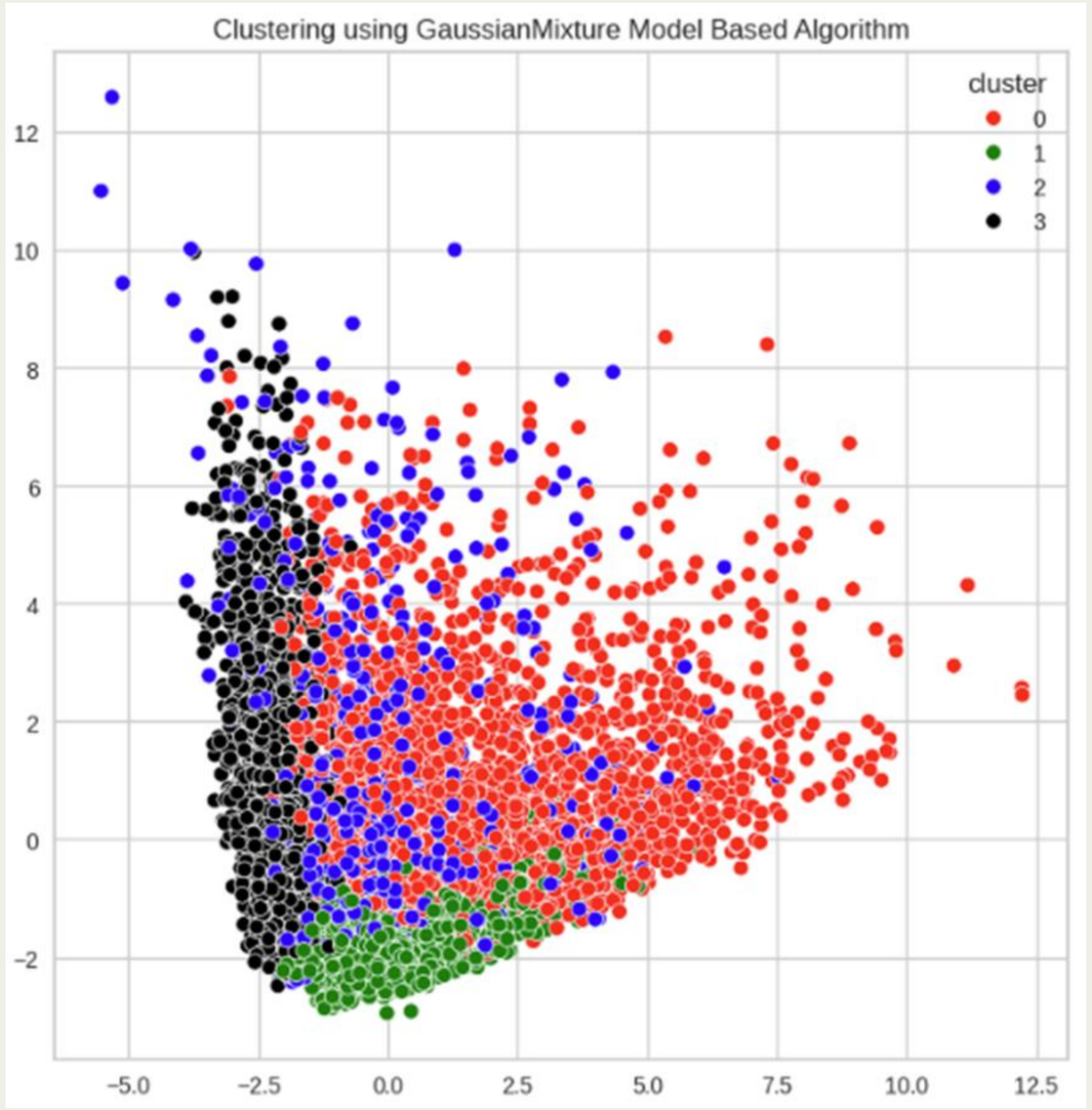


K-Means Visualisation



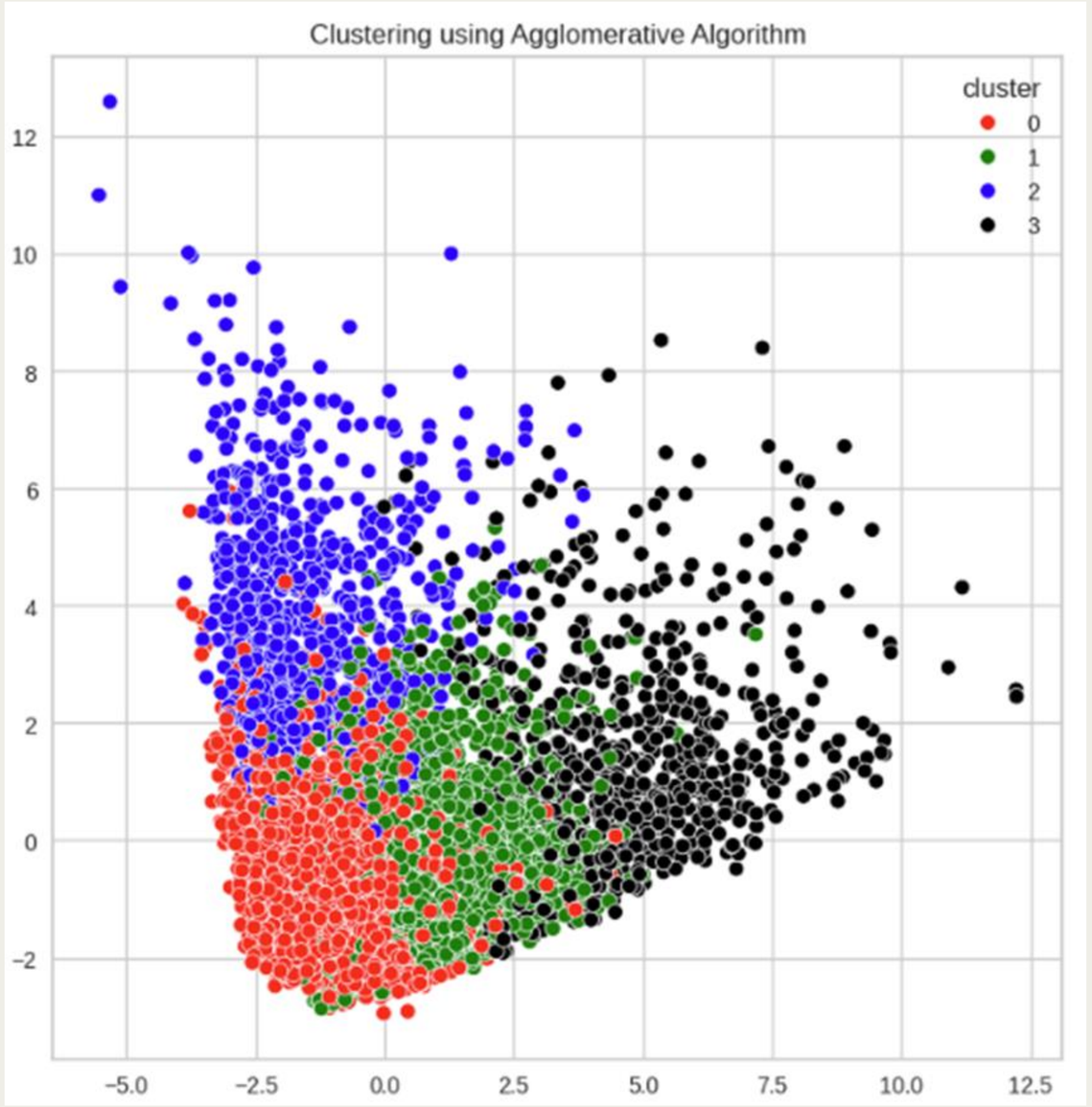
MODELLING - GMM VISUALISATION

GMM Algorithm Visualisation



MODELLING - HIERARCHICAL CLUSTERING (AGGLOMERATIVE)VISUALISATION

Hierarchical Clustering (Agglomerative) Algorithm Visualisation



ANALYZE / MODEL EVALUATE

Evaluation of – KMeans

Evaluation of – GMM

Evaluation of – Hierarchical Clustering (Agglomerative)

Model	Davies-Bouldin Index	Silhouette Score	Calinski-Harabasz Index
K-Means	1.647466291015152	0.18579999848578388	1829.3927271369603
GMM	3.026489194621547	0.09004318004525669	799.2185891108142
Hierarchical / Agg	1.9182891848022057	0.14479620924142522	1422.977130591173

Thank you!