BANK CUSTOMER SEGMENTATION

USING PYTHON

by ks.milda.ks@gmail.com

TABLE OF CONTENTS

QUICK SNAPSHOT

In this project, I set out to explore bank customer data from Kaggle and uncover hidden patterns using clustering techniques. By applying Principal Component Analysis (PCA) and K-Means, I was able to reduce the complexity of the dataset and group customers into meaningful segments.

- INTRODUCTION
- IMPORT DATA
- PRE-PROCESSING
- SILHOUTTE SCORE
- K-MEANS ANALYSIS
- RESULTS
- CONCLUSION

INTRO DUCTION

PURPOSE

The aim for this project is to segementing or clustering bank’s customers based on the choosed column to uncover customers pattern and the result of this clustering can be use to design different marketing, product offerings, and retention strategies for each.

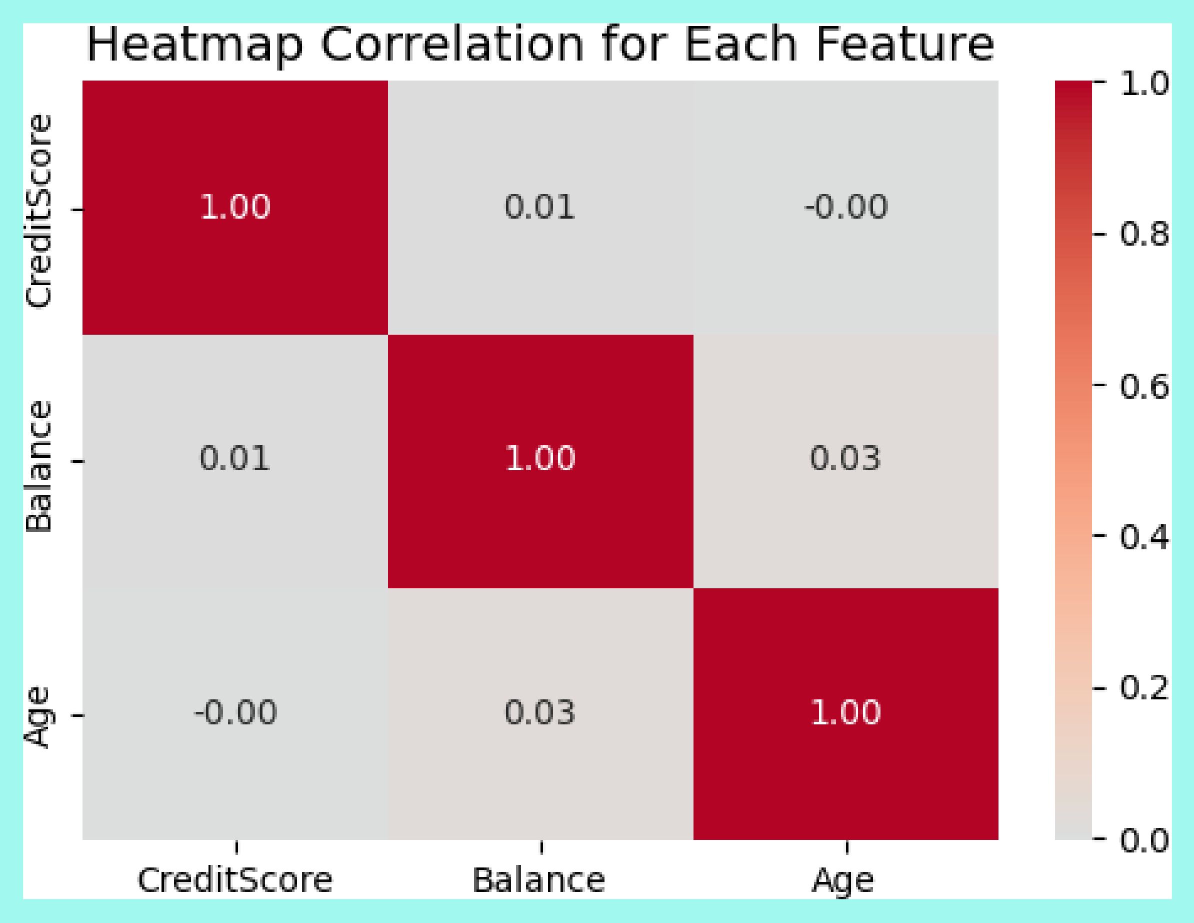
DATASET

This dataset is for XYZ Multistate Bank and contains various columns that capture key aspects of customer behavior and attributes. the choosen columns or dimensions are 'CreditScore', 'Balance', 'Age'.

IMPORT DATA

IMPORT

On Jupyter notebook, the Bank Segmentation dataset is imported using “pandas” package at “read_excel”. Specifically, the choosen columns or dimensions are 'CreditScore', 'Balance', 'Age'. Next step is plotting the correlation between each dimensions using heatmap.



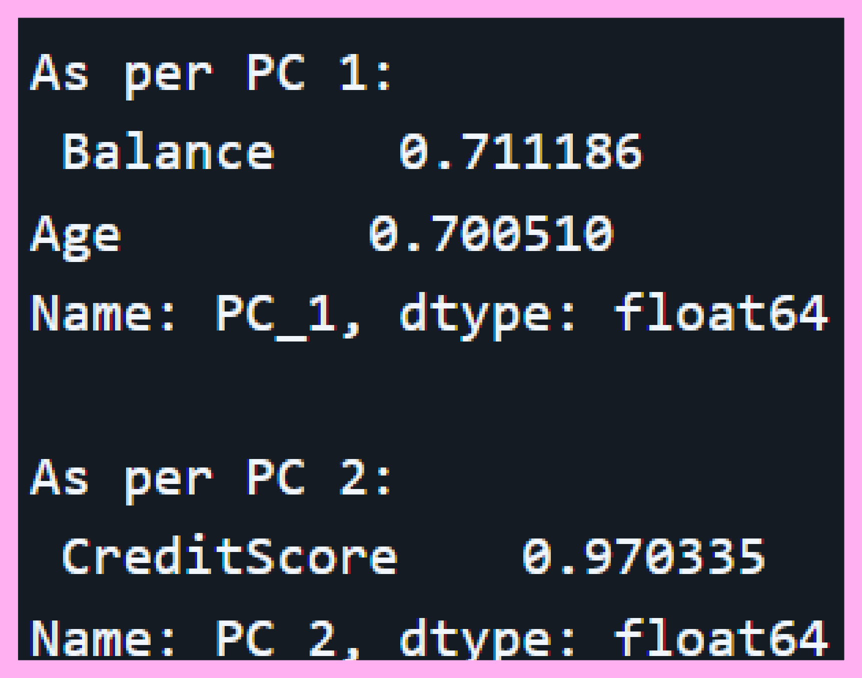
Based on the heatmap, all three dimensions are nearly independent. There are only a little correlations between each 3 dimensions and none of the variable strongly explain each other, so the segmentation will rely on their individual contributions.

PRE PROCESSING - PCA

STANDARIZE & PCA

The next step is standarize the feature using z-score normalization X = StandardScaler().fit_transform(X) which will rescales all features to mean = 0 and standard deviation = 1, so each variable contributes equally to the PCA calculation. Then, the Principal Component Analysis (PCA) is performed to reduce the 3 varièbles into 2 dimensions.

Resulting the 2 components is capture 67.67% from 3 reduced variables, which is acceptable. The contribution each variable for each principal component (PC) as shown on the side,



```
As per PC 1:
 Balance     0.711186
Age         0.700510
Name: PC_1, dtype: float64

As per PC 2:
 CreditScore    0.970335
Name: PC 2, dtype: float64
```

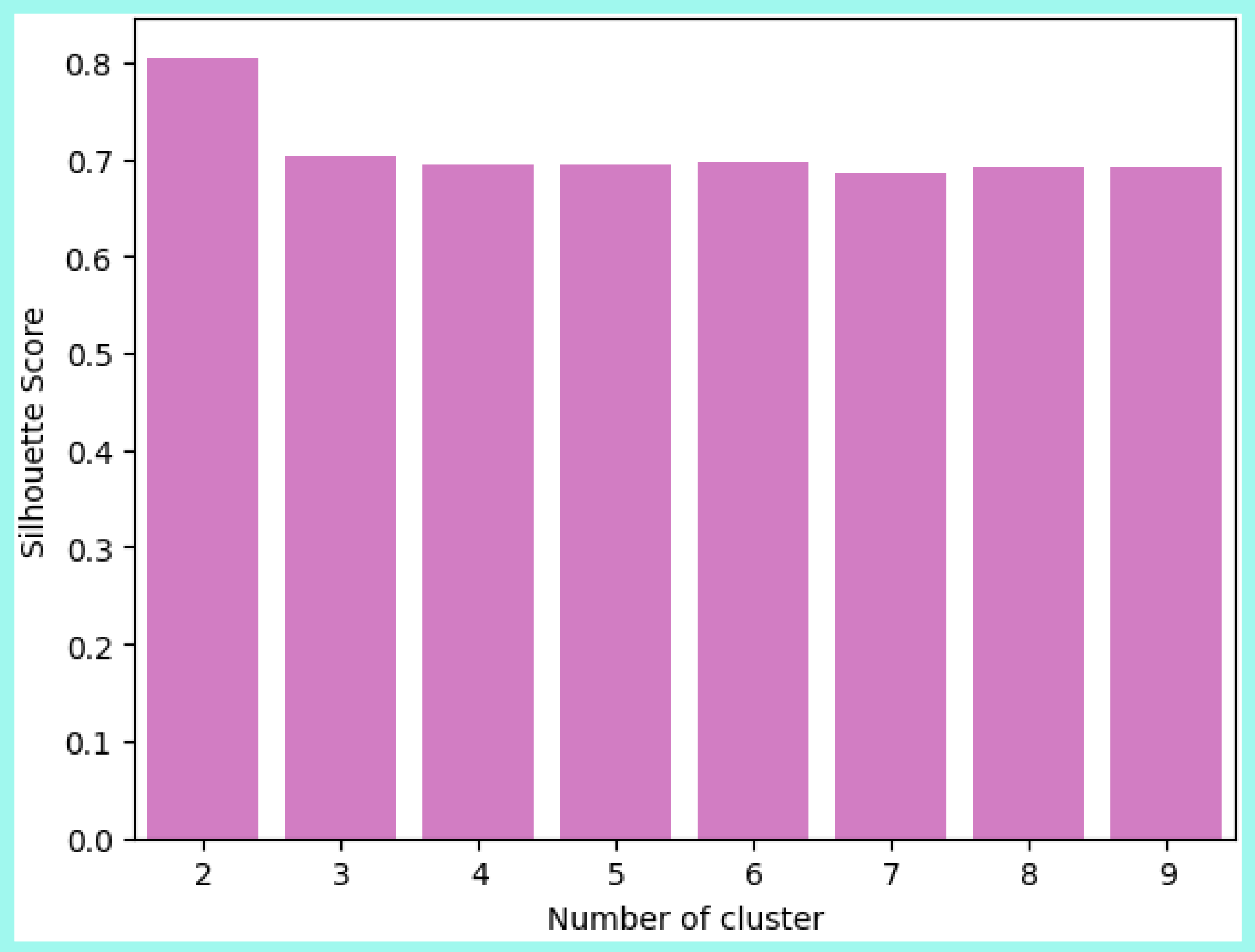
For PC1 the most contributing variable were 'Balance' and 'Age', and for PC2 the most contributing variable were 'CreditScore'. It can be interpreted as PC1 were combination of Balance and Age, meanwhile PC2 were mostly explained by CreditScore. Hence, the clusters builded later are formed mainly based on Balance + Age vs CreditScore differences.

SILHOUTTE SCORE

NUMBER OF CLUSTER

The silhoutte score is measuring how well the data points fit within their assigned clusters in a clustering algorithm (like KMeans). It is use to choose the optimal number of clusters, the highest score is the best number.

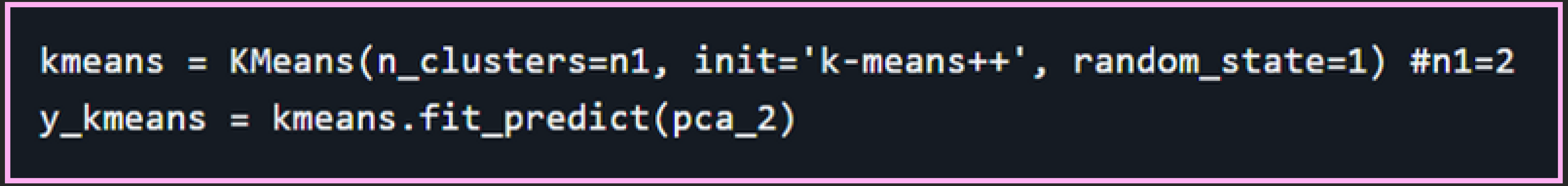
Based on pict on the side, the optimal number of cluster is 2, which scores the highest near 0.8.



K-MEANS ANALYSIS

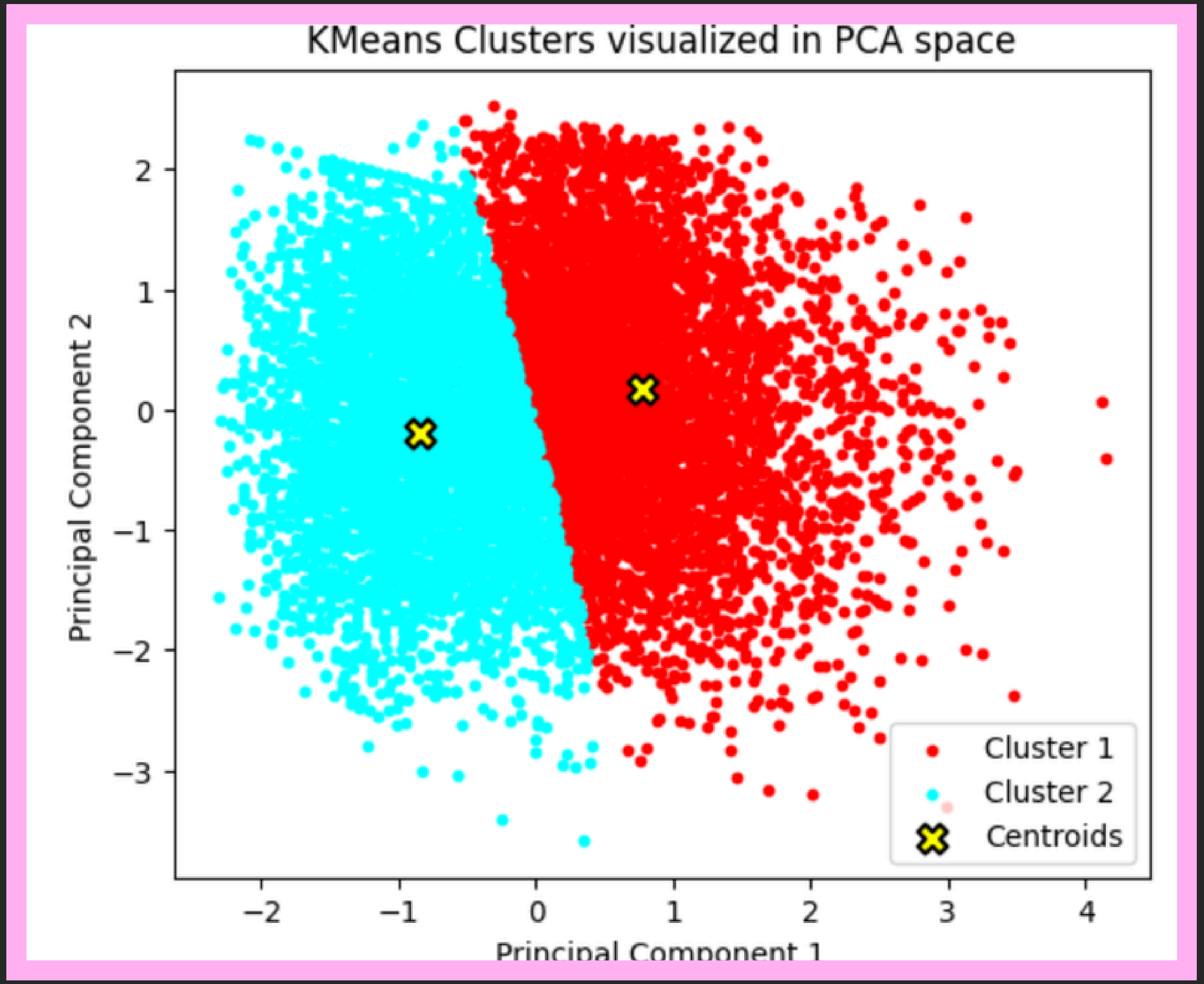
PERFORMANCE BY DIVISION

The K-Means algorithm is performed. What K-Means do is grouping data into k clusters by minimizing the distance between data points and their cluster’s center (centroid) which is calculated by squared Euclidean distance. There are two clusters, so there is two centroids. The data used is the dataset after standarize and applying PCA.



```
kmeans = KMeans(n_clusters=n1, init='k-means++', random_state=1) #n1=2
y_kmeans = kmeans.fit_predict(pca_2)
```

THE RESULT



SEGMENTATION RESULT

All the process, resulting on two cluster or segmentation of Bank's customer based on customer's Balance, Credit score, and Age. The first cluster (red), around positive values of Principal Component 1 (balance and age). Customer on cluster 1 were older customers with higher account balance because of higher PC1 values. This cluster can be labeled as "Wealthy senior" and customers in this cluster likely more financially stable, higher value customers, could be more loyal but less likely to adopt new products unless they see clear value.

The second cluster (cyan), around negative values of Principal Component 1 (balance and age). Customer on cluster 2 were younger customers with lower account balance because of lower PC1. This cluster can be labeled as "Emerging Young Customers" as they were likely early in their financial journey, smaller spending power now, but more open to digital products, cross-selling, or upselling as they grow financially.

CONCLUSION

SEGMENTATION

This project successfully segmented bank customers into two main groups using PCA and K-Means clustering based on their Credit Score, Balance, and Age. By applying PCA, data complexity was reduced while maintaining 67.67% of the variance, allowing clear visualization and interpretation of customer patterns. The optimal number of clusters (k=2) was determined using the silhouette score (~0.8), ensuring strong separation between groups.

- Cluster 1 – “Wealthy Seniors”: Older customers with higher balances and stable financial profiles. They represent high-value, loyal clients who may prefer traditional banking but respond to premium or loyalty-based services.
- Cluster 2 – “Emerging Young Customers”: Younger customers with lower balances but high growth potential. They are more receptive to digital banking, cross-selling, and new financial products.

Overall, this segmentation provides actionable insights for targeted marketing, product design, and retention strategies, allowing banks to tailor approaches based on customer life stage and financial capacity.

THANK YOU

FOR ATTENTION

KS.MILDA.KS@GMAIL.COM

MILDA KHAERANI S.

+62 851 7990 6131