BLM432E Introduction to Data Science Final Project

Name Surname:	Meryem Ezber
Dataset definition:	This case requires to develop a customer segmentation to define marketing strategy. The sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables. Following is the Data Dictionary for Credit Card dataset :- **CUST***ID* : Identification of Credit Card holder (Categorical) **BALANCE** : Balance amount left in their account to make purchases (**BALANCEFREQUENCY** : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated) **PURCHASES** : Amount of purchases made from account **ONEOFF*PURCHASES*** : Maximum purchase amount done in one-go **INSTALLMENTSPURCHASES** : Amount of purchase done in installment **CASH***ADVANCE* : Cash in advance given by the user **PURCHASESFREQUENCY** : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased) **ONEOFFPURCHASESFREQUENCY** : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased) **PURCHASESINSTALLMENTSFREQUENCY** : How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done) **CASHADVANCEFREQUENCY** : How frequently the cash in advance being paid **CASHADVANCETRX** : Number of Transactions made with "Cash in Advanced" **PURCHASES*TRX*** : Number of purchase transactions made **CREDITLIMIT** : Limit of Credit Card for user **PAYMENTS** : Amount of Payment done by user **MINIMUM_PAYMENTS** : Minimum amount of payments made by user **PRCFULLPAYMENT** : Percent of full payment paid by user **TENURE** : Tenure of credit card service for user
Dataset source (web address):	https://www.kaggle.com/arjunbhasin2013/ccdata
Aim of the project:	Clustering Credit Card Users

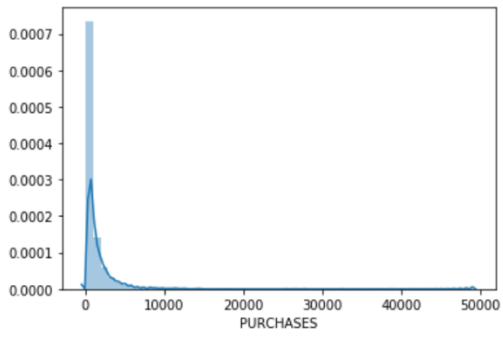
Step1: Exploratory Data Analysis – 10p

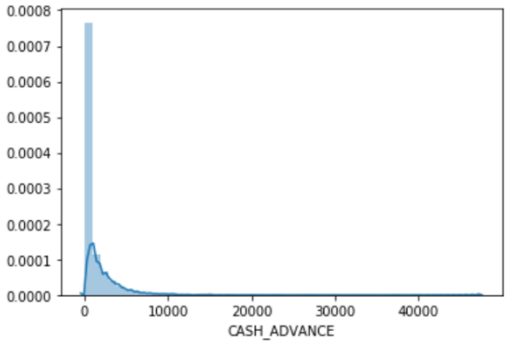
a) Explain the shape of the dataset (restriction of at least 10 columns, 1000 rows)
It has 18 columns, 8950 rows

b) Explain the column types in the dataset

CUST_ID	object
BALANCE	float64
BALANCE_FREQUENCY	float64
PURCHASES	float64
ONEOFF_PURCHASES	float64
INSTALLMENTS_PURCHASES	float64
CASH_ADVANCE	float64
PURCHASES_FREQUENCY	float64
ONEOFF_PURCHASES_FREQUENCY	float64
PURCHASES_INSTALLMENTS_FREQUENCY	float64
CASH_ADVANCE_FREQUENCY	float64
CASH_ADVANCE_TRX	int64
PURCHASES_TRX	int64
CREDIT_LIMIT	float64
PAYMENTS	float64
MINIMUM_PAYMENTS	float64
PRC_FULL_PAYMENT	float64
TENURE	int64

c) Explain the distribution of only 2 features in your dataset (one numeric, one categorical) using visualizations
There was only one object column. Deleted that column because it was meaningless. The distribution of the two numeric columns as follows:





Step2: Preprocessing – 10p

a) How many columns include missing values
2 columns: CREDIT_LIMIT and MINIMUM_PAYMENTS

b) Explain your method to handle each of those missing values
I filled the missing value with mean.

c) Explain if you needed to apply any kind of transformations.
Dataset does not contain a categorical column. Therefore, I don't apply the encoder process. I applied a min max scaler.

Step3: Clustering evaluation – 15p

a) Select 3 clustering evaluation methods that you will use in your project, and explain them in detail by giving proper formulation.

1. External Evaluation

In external evaluation, clustering results are evaluated based on data that was not used for clustering, such as known class labels and external benchmarks. Such benchmarks consist of a set of pre-classified items, and these sets are often created by (expert) humans. Thus, the benchmark sets can be thought of as a gold standard for evaluation.

Rand index

The Rand index computes how similar the clusters (returned by the clustering algorithm) are to the benchmark classifications. One can also view the Rand index as a measure of the percentage of correct decisions made by the algorithm. It can be computed using the following formula:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

where TP s the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. One issue with the Rand index is that false positives and false negatives are equally weighted. This may be an undesirable characteristic for some clustering applications. The F-measure addresses this concern, as does the chance-corrected adjusted Rand index.

2. Internal Evaluation

When a clustering result is evaluated based on the data that was clustered itself, this is called internal evaluation. These methods usually assign the best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters. One drawback of using internal criteria in cluster evaluation is that high scores on an internal measure do not necessarily result in effective information retrieval applications.

Dunn index

The Dunn index aims to identify dense and well-separated clusters. It is defined as the ratio between the minimal inter-cluster distance to maximal intra-cluster distance. For each cluster partition, the Dunn index can be calculated by the following formula:

$$D = \frac{min_{1 \le i \le j \le n}\, d(i,j)}{max_{1 \le k \le n}\, \acute{d}(k)}$$

where $d(i,j)$ represents the distance between clusters i and j, and $d'(k)$ measures the intra-cluster distance of cluster k. The inter-cluster distance $d(i,j)$ between two clusters may be any number of distance measures, such as the distance between the centroids of the clusters. Similarly, the intra-cluster distance $d'(k)$ may be measured in a variety way, such as the maximal distance between any pair of elements in cluster k. Since internal criterion seek clusters with high intra-cluster similarity and low inter-cluster similarity, algorithms that produce clusters with high Dunn index are more desirable.

Silhouette coefficient

The silhouette coefficient contrasts the average distance to elements in the same cluster with the average distance to elements in other clusters. Objects with a high silhouette value are considered well clustered, objects with a low value may be outliers. This index works well with k-means clustering, and is also used to determine the optimal number of clusters.

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

Assume the data have been clustered via any technique, such as k-means, into k clusters. For data point $i \in C_i$ (data point i n the Cluster C_i), let

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

Be the mean distance between and all other data points in the same cluster, where $d(i,j)$ is the distance between data points i and j in cluster C_i

We then define the mean dissimilarity of point i to some cluster C_k as the mean of the distance from i to all points in C_k (where $C_k \neq C_i$)

For each data point $i \in C_i$, we now define

$$b(i) = min \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

to be the *smallest* (hence the min operator in the formula) mean distance of i to all points in any other cluster, of which i is not a member. The cluster with this smallest mean dissimilarity is said to be the "neighboring cluster" of i because it is the next best fit cluster for point i.

We now define a *silhouette* (value) of one data point i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, if\ |C_i| > 1$$

and

$$s(i) = 0, if\ |C_i| = 1$$

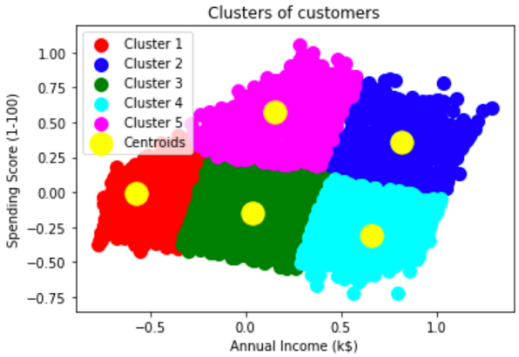
Which can be also written as:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}\ , & if\ a(i) < b(i) \\ 0, & if\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1\ , & if\ a(i) > b(i) \end{cases}$$

From the above definition it is clear that

$$-1 \leq s(i) \leq 1$$

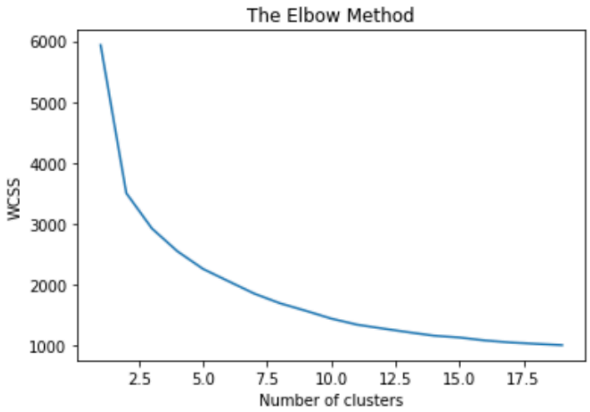
b) Show example visualizations associated with your selected method.



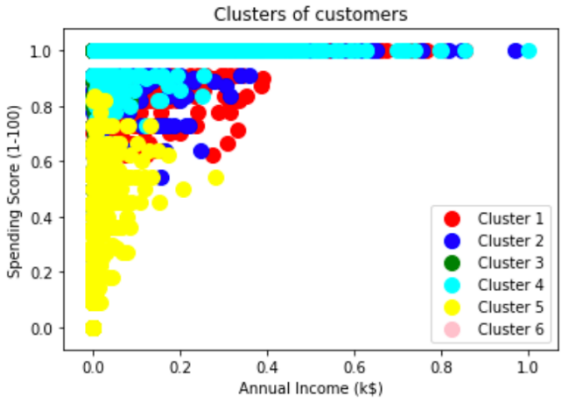
Silhouette score = 0.5512098767550345

Step4: Clustering algorithms, Implementation and Performance Comparison

1 Partitioning Methods, K-means

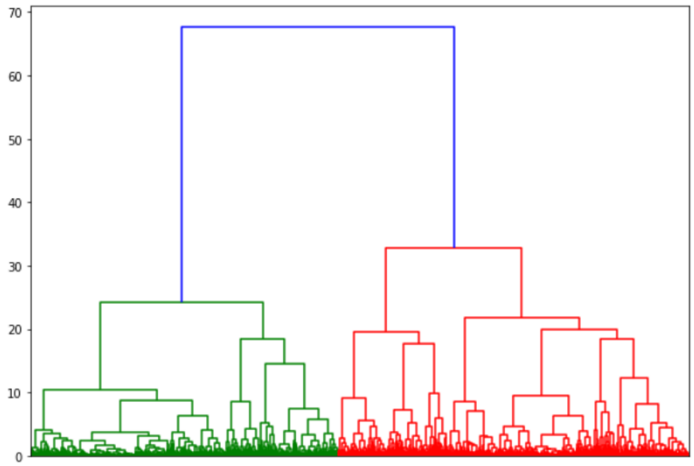


Create 6 clusters using the elbow method. These clusters are as follows.

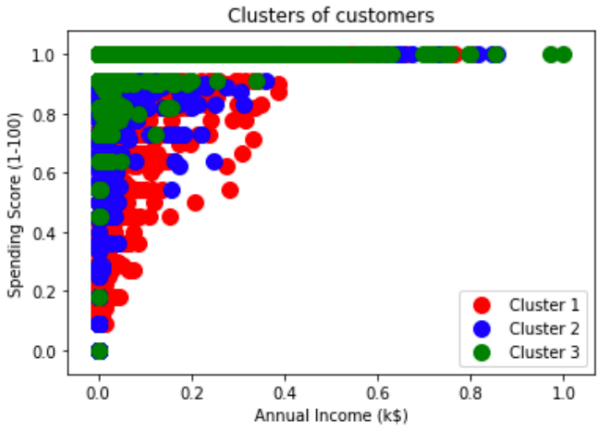


Silhouette score = 0.3191226915427069

2 Hierarchical Methods, Dendrogram

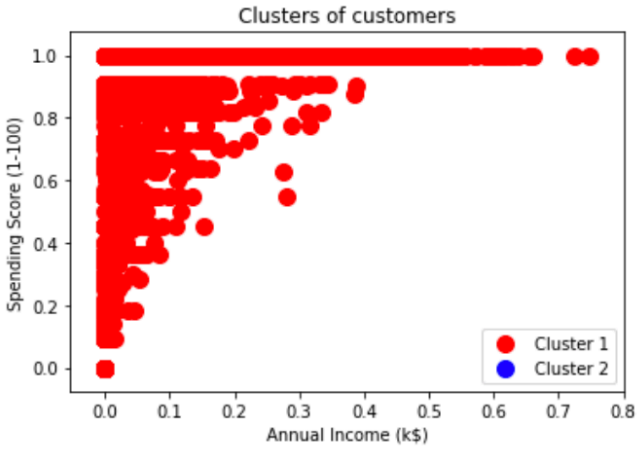


Create 4 clusters using the dendrogram. These clusters are as follows.



Silhouette score = 0.3336350894872987

3 Density-Based Methods, DBSCAN



Silhouette score = 0.18476669607882654

RESULT

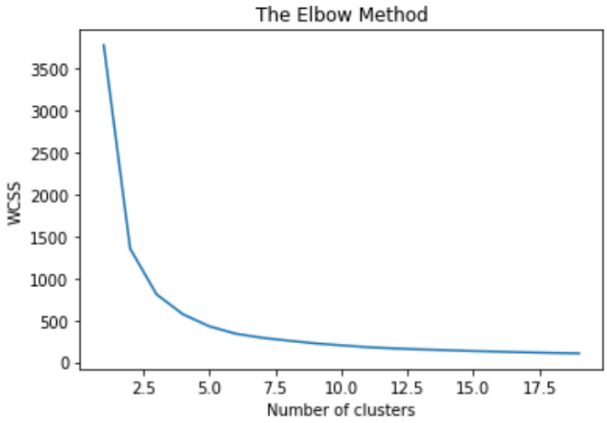
Method	Silhouette Score
Partitioning Methods, K-Means	0.3191226915427069
Hierarchical Methods, Dendrogram	0.3336350894872987
Density-Based Methods, DBSCAN	0.18476669607882654

The best score was obtained with the **hierarchical methods**.

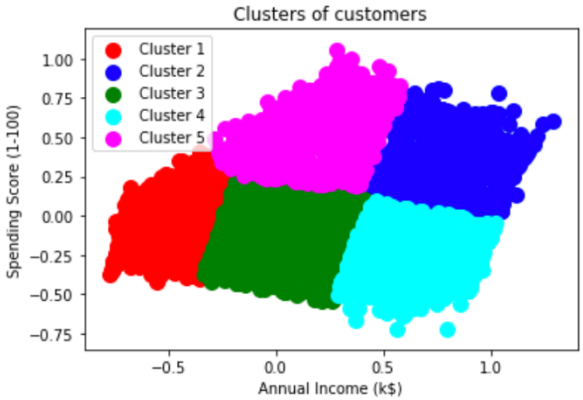
Step5: Further Performance Improvement (Your best clustering algorithm)

In this step, PCA is used to work with the most effective features and to ignore outlier values. All steps and methods were rerun. The results are as follows.

1 Partitioning Methods, K-means

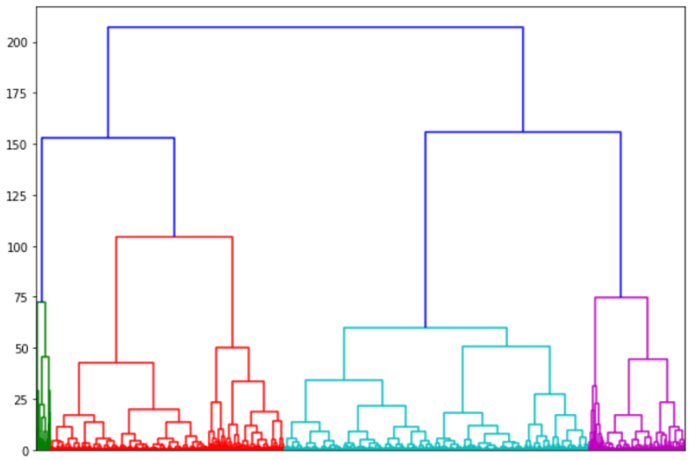


Create 5 clusters using the elbow method. These clusters are as follows.

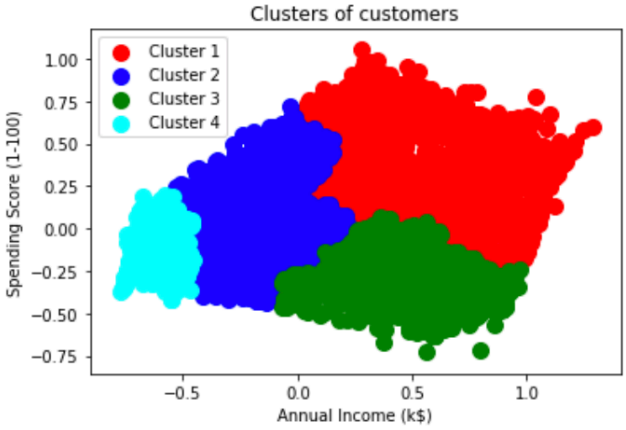


Silhouette score = 0.5512098767550345

2 Hierarchical Methods, Dendrogram



Create 4 clusters using the dendrogram. These clusters are as follows.



Silhouette score = 0.45779320151770236

3 Density-Based Methods, DBSCAN

Silhouette score = 0.052019502392340655

RESULT

Method	Silhouette Score
Partitioning Methods, K-Means	0.5512098767550345
Hierarchical Methods, Dendrogram	0.45779320151770236
Density-Based Methods, DBSCAN	0.052019502392340655

The best score was obtained with the **k-means algorithm**.