**PROJECT – Evaluation of Clustering Algorithms**

Sergei Rogov

U231N0051

University of Nicosia

COMP-344 (Section 01): Machine Learning and Data Mining II

## Instructor: Dr Ioannis Katakis

January 5, 2024

Goals: getting acquainted with a variety of clustering algorithms, comprehending their outcomes, comparing the results and finding out which algorithm is the most suitable for a particular case.

Summary:

**Datasets**

**Dataset 1**

[**https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data**](https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data)

**Dataset 2**

[**https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset/data**](https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset/data)

**Dataset 3**

[**https://www.kaggle.com/datasets/dev0914sharma/customer-clustering**](https://www.kaggle.com/datasets/dev0914sharma/customer-clustering)

**Clustering Algorithms**

**K-Means**

K-Means is a clustering algorithm that initially requires a parameter k which represents the number of clusters.

K-Means algorithm is the following:

Step 1: k initial centroids are randomly chosen across the dataset.

Repeat:

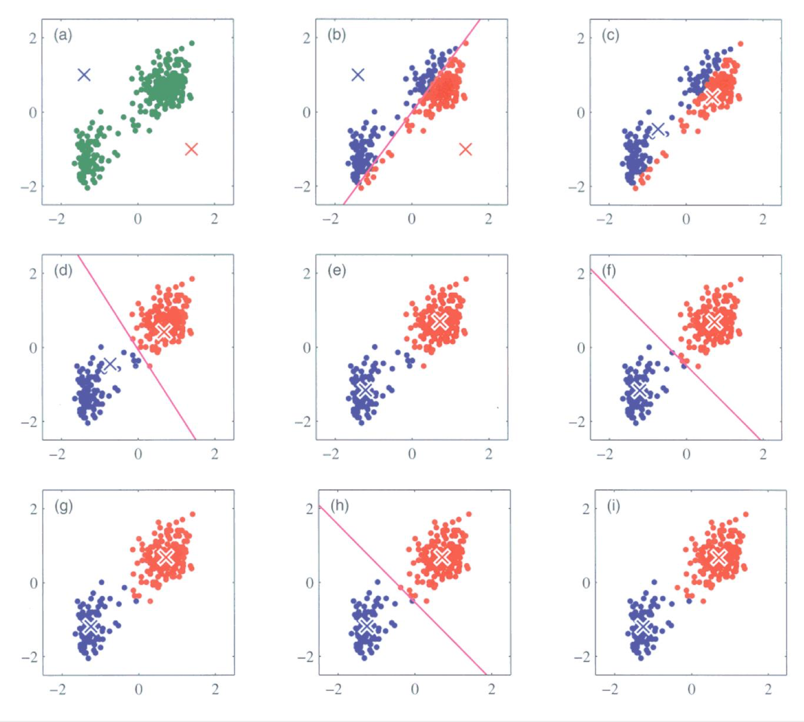
Step 2: each data instance is assigned a cluster by assigning it to the closest centroid.

Step 3: centroids are recomputed by finding the “center of mass” of each current cluster.

Until centroids do not change (or change insignificantly, for example too few instances change their assigned clusters).

Other stopping conditions might be used, like limiting the number of iterations. This could work well in many scenarios because almost all convergence happens during first few iterations. This method is also useful because of the fact that algorithm might stuck in an infinite loop.

Advantages of the K-Means are that it is easy to implement and easy to interpret results. It is computationally fast efficient as well. The disadvantage, however, is that the number of clusters is not known in the beginning, so it is not clear which value of parameter k to choose. Besides, as the initial centroids are chosen randomly, it leads to different clustering on the same dataset from one run of the algorithm to another. Moreover, K-Means does not work well when it comes to dataset containing non-spherical clusters.



**Hierarchical (Agglomerative)**

Hierarchical clustering is a clustering algorithm that is able to provide you with not only a set of clusters but with a whole folder-like structure of nested clusters, depicting the sequence of merges between data instances.

The algorithm of agglomerative clustering is described here:

Step 1: proximity matrix is computed

Step 2: each data instance is considered to be a separate cluster

Repeat:

Step 3: merge two closest clusters

Step 4: recompute the proximity matrix

Until all clusters are merged into one single cluster.

There several approaches available for updating the proximity matrix, them being: taking the minimum distance between clusters, the maximum distance, the average distance, the distance between centroids and many others.

The problem with hierarchical clustering is that no objective function is being directly minimized. Besides, if a decision to split or merge clusters is made, this action cannot be undone. The advantage is that it is not needed to specify the number of clusters to run the algorithm. It is possible to get any desired number of clusters by cutting the dendrogram (tree representation of the clustering) at a proper level.

**DBSCAN**

Density-based Spatial Clustering of Applications with Noise is a clustering algorithm that is capable of finding clusters based on its points density as well as finding and eliminating outliers

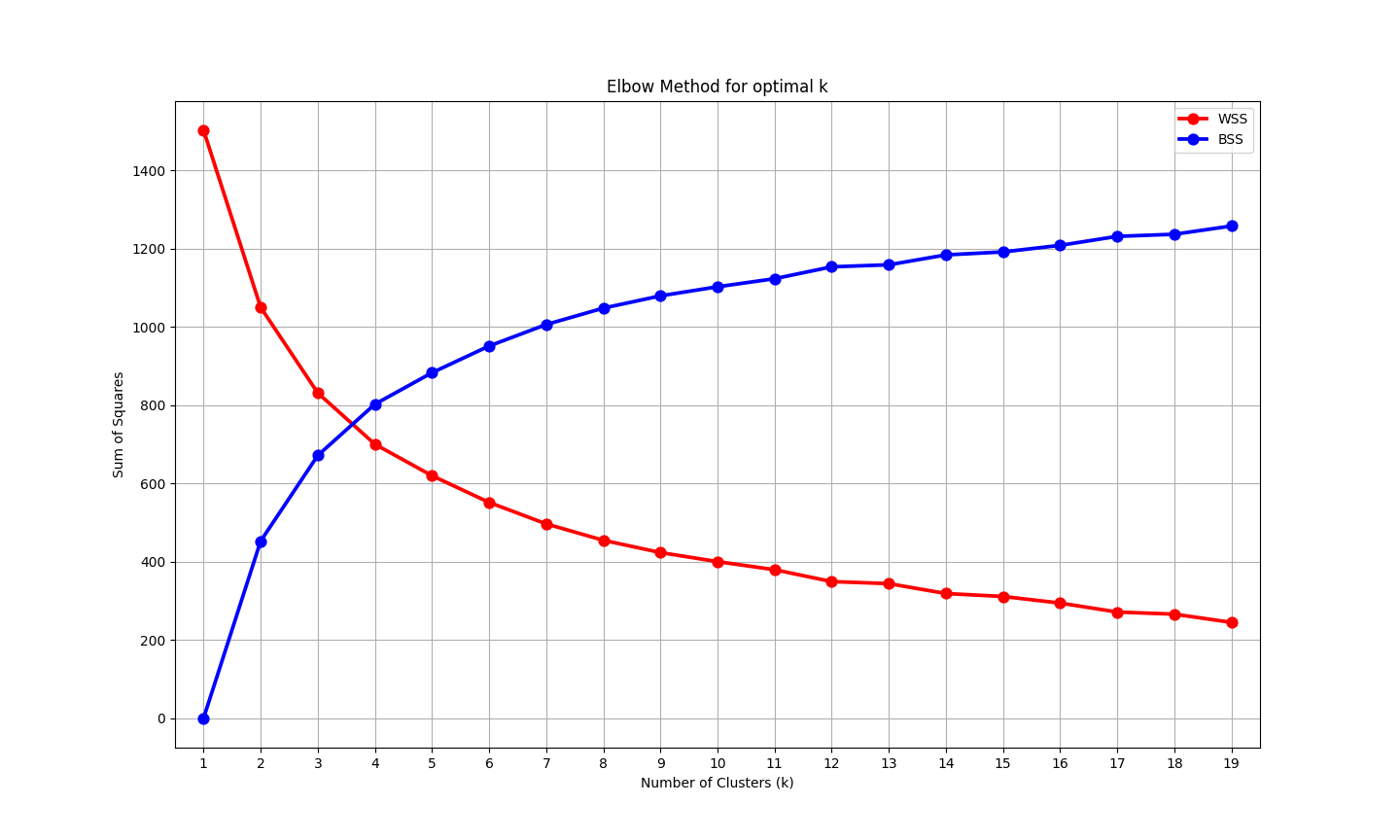
DBSCAN algorithm:

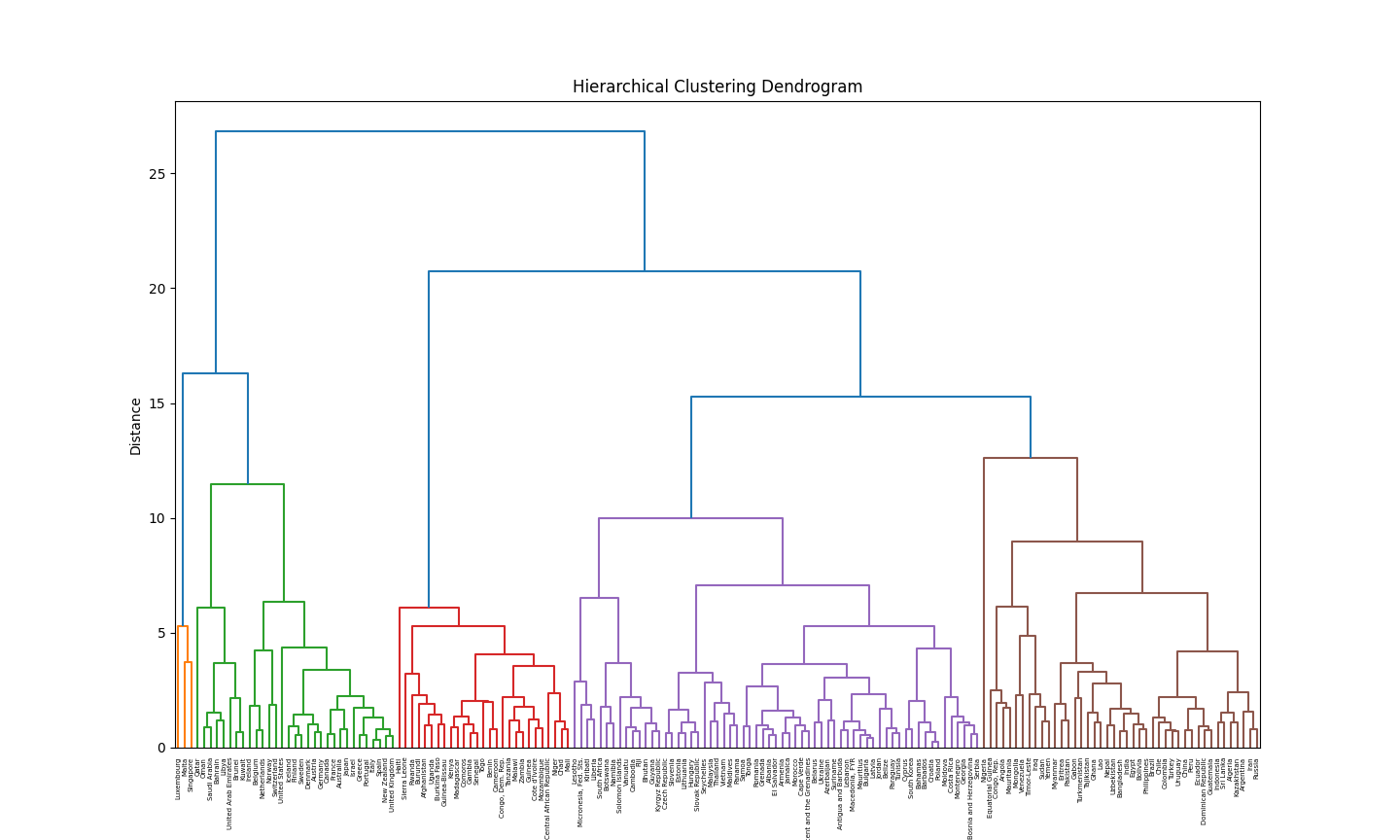
Step 1: find neighbors of every point within the radius epsilon and identify core con

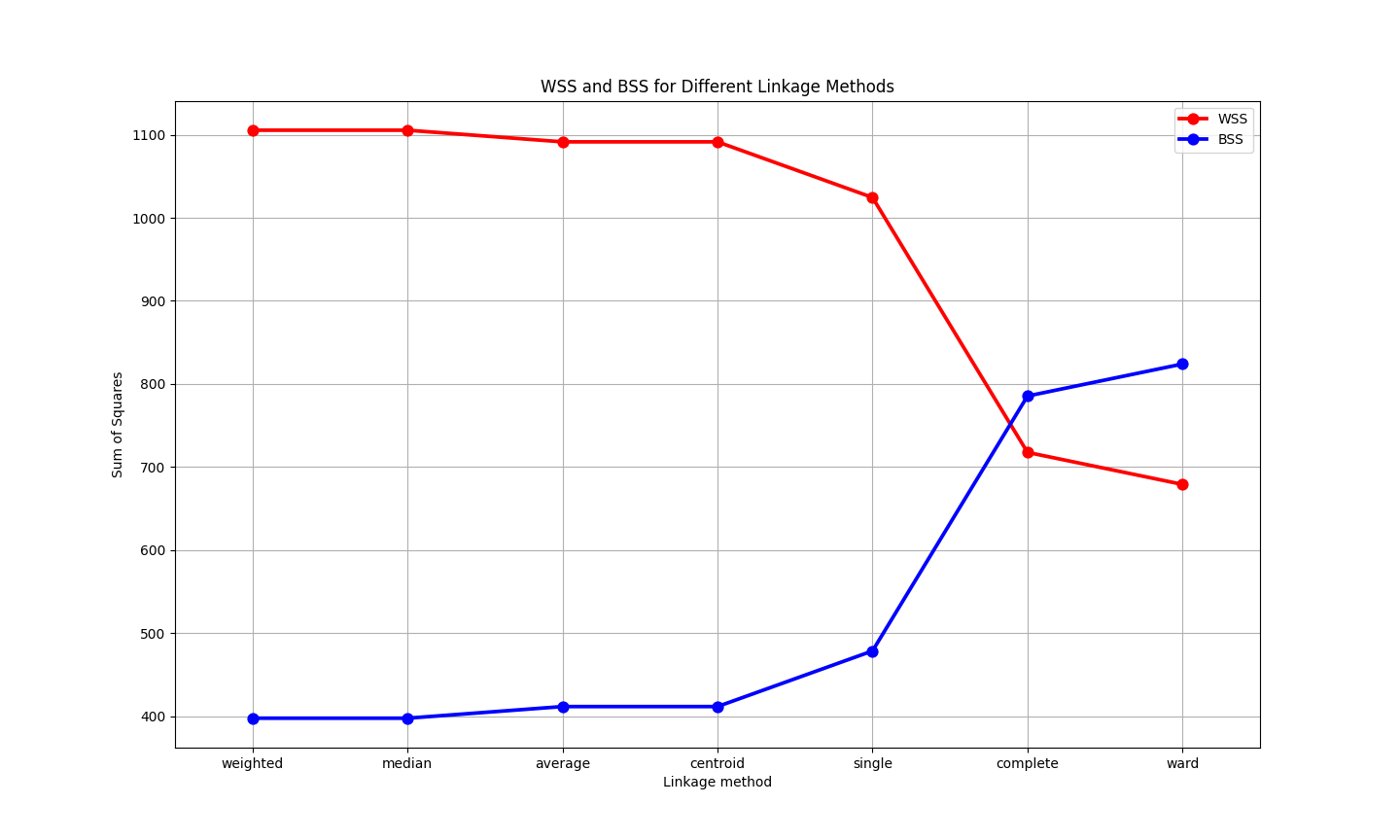
ints

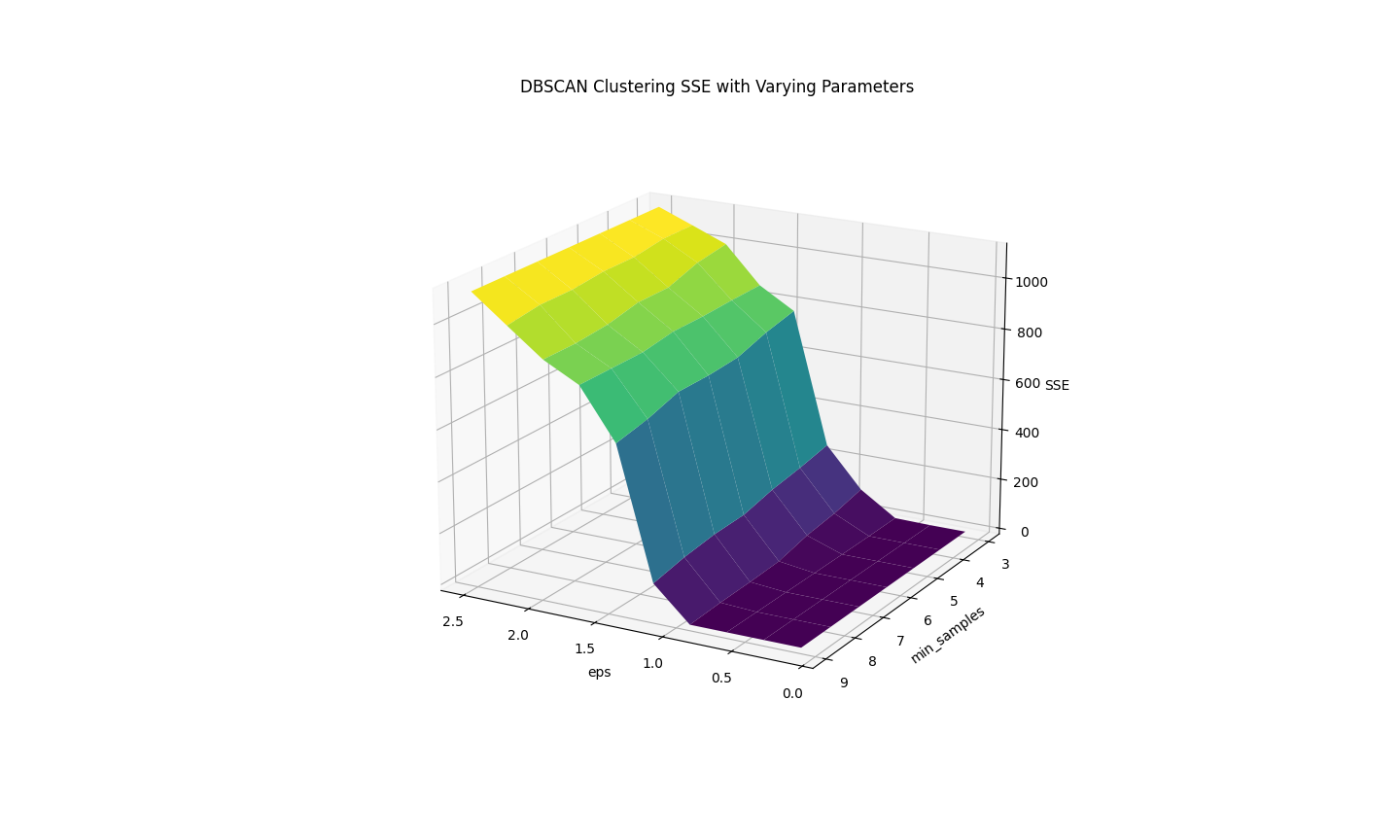
**Results**

**Countries dataset**

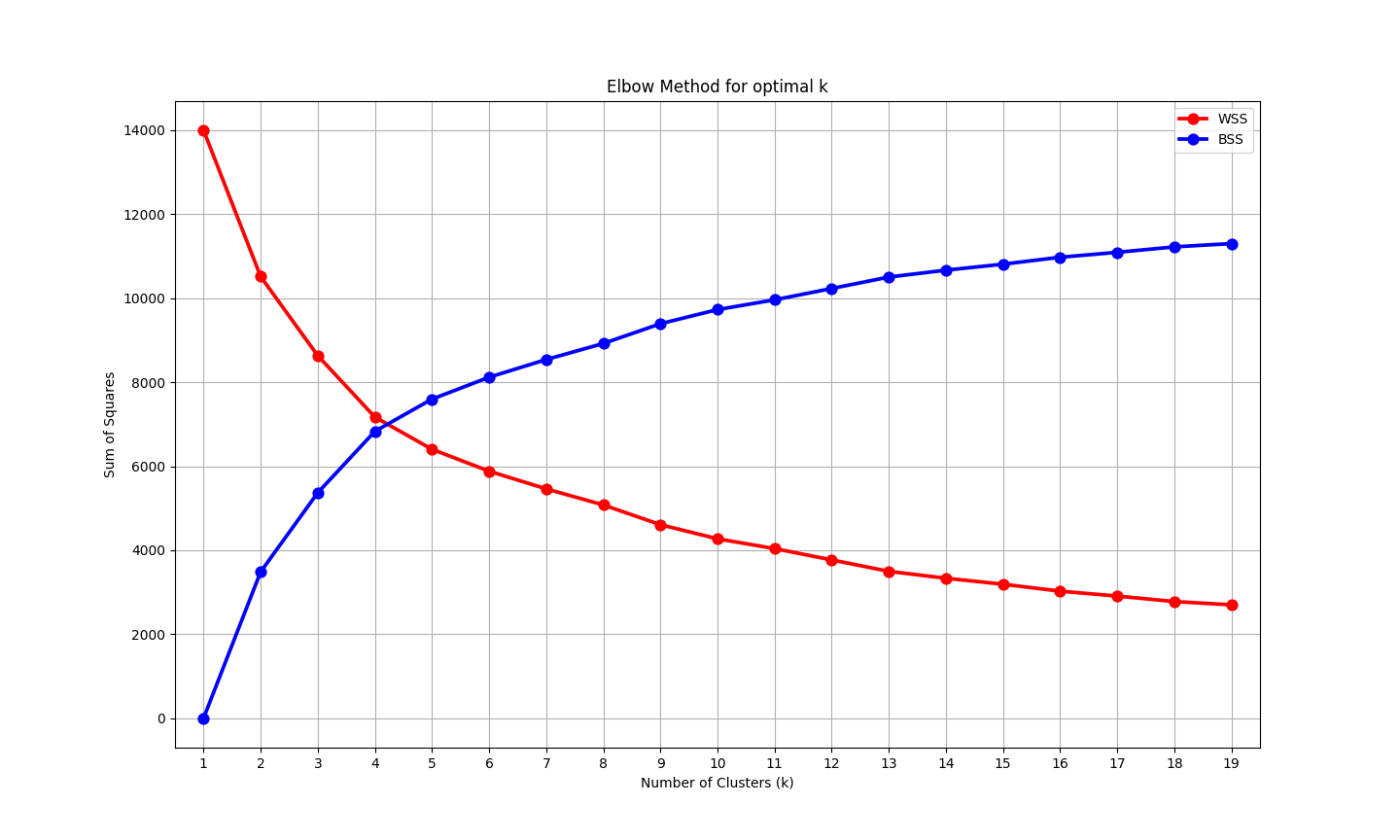
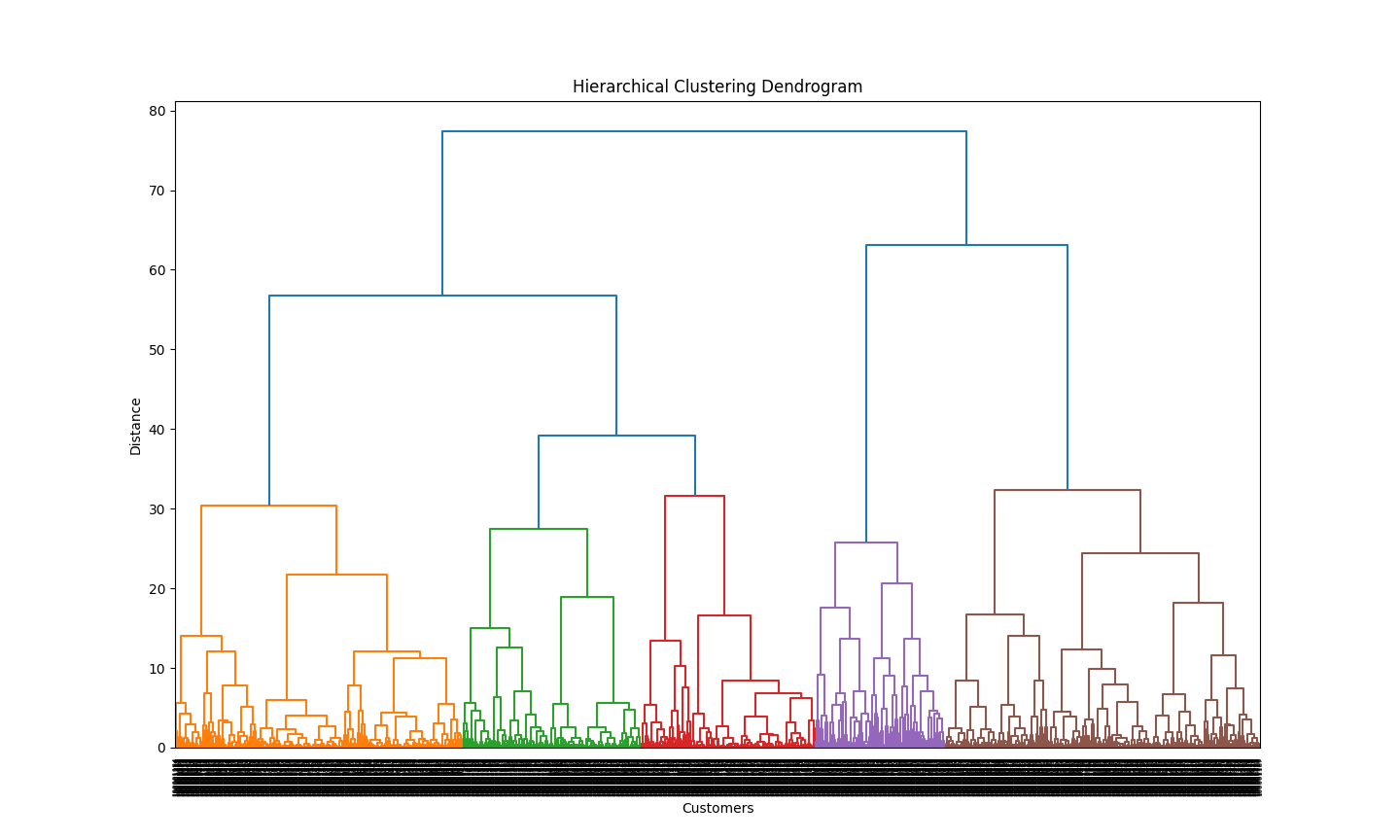
****

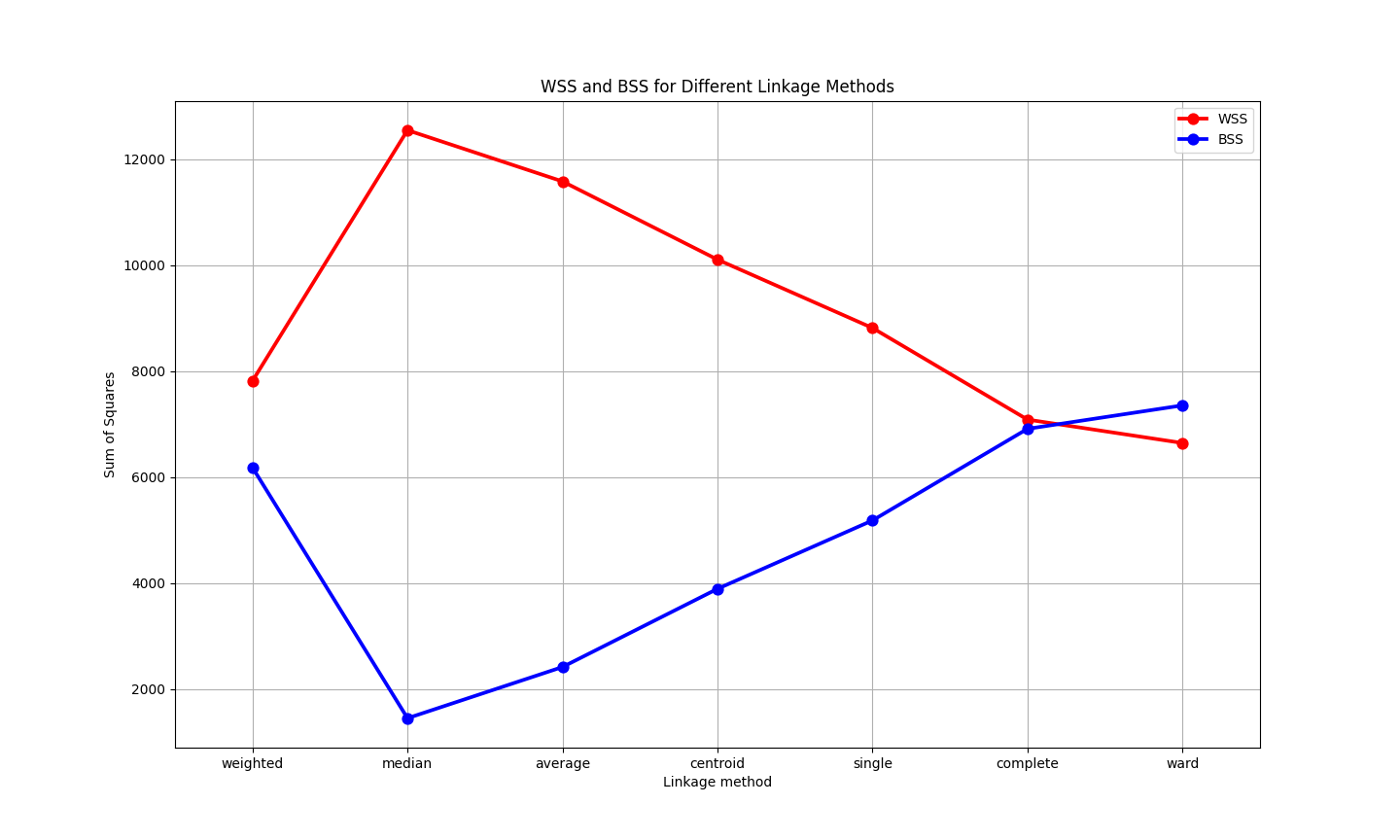
****

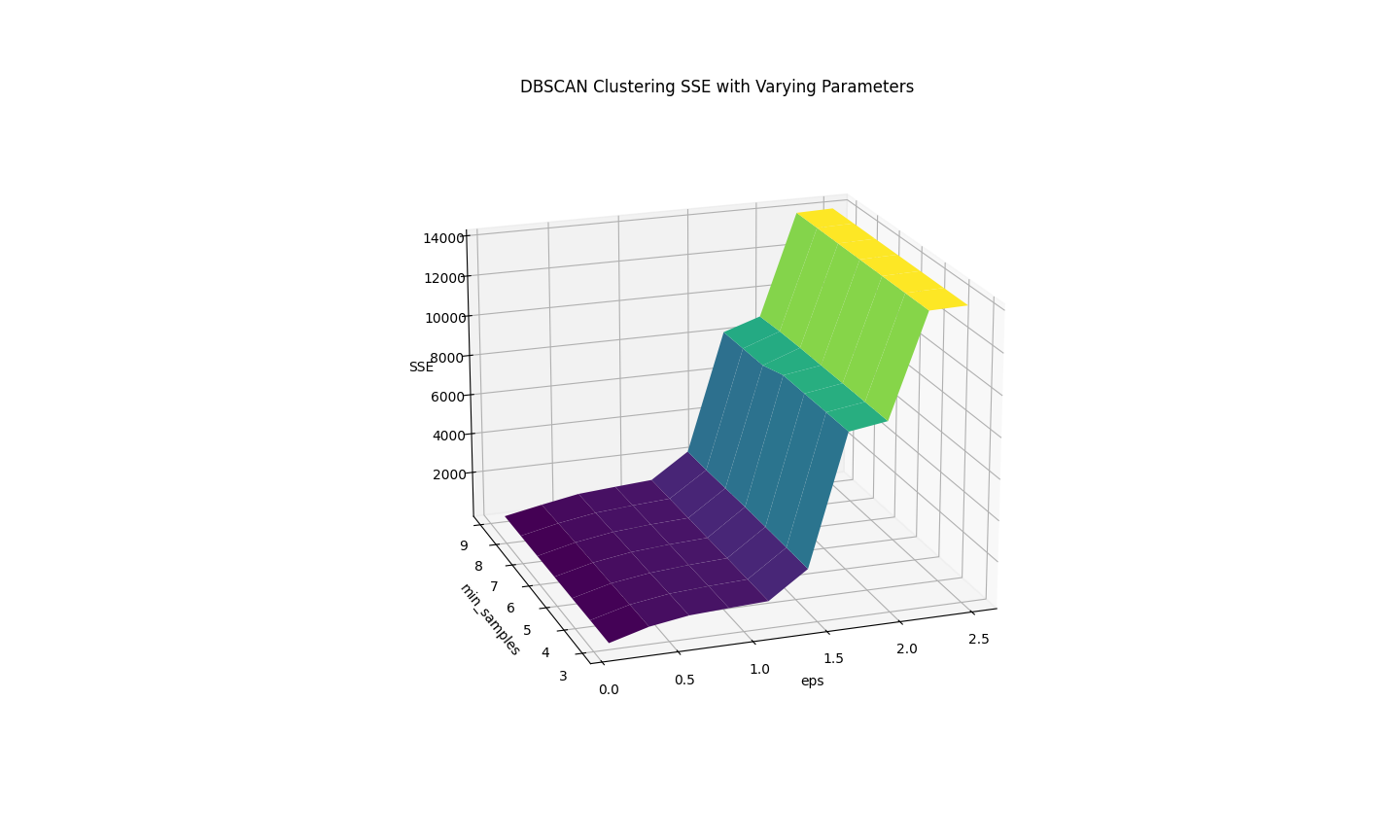
****

****

**Customers dataset**

****

****

****

**Discussion**

**Conclusions**

**References**

[**https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data**](https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data)

[**https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset/data**](https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset/data)

[**https://www.kaggle.com/datasets/dev0914sharma/customer-clustering**](https://www.kaggle.com/datasets/dev0914sharma/customer-clustering)