Name : Muhammad Usman
ID: 22093856

7PAM2000 Applied Data Science 1
Assignment 3: Clustering and fitting

Github
https://github.com/usmanioouk/DS_assignment_3

Abstract

Analyzing worlds population data using k means clustering and finding out patterns between them

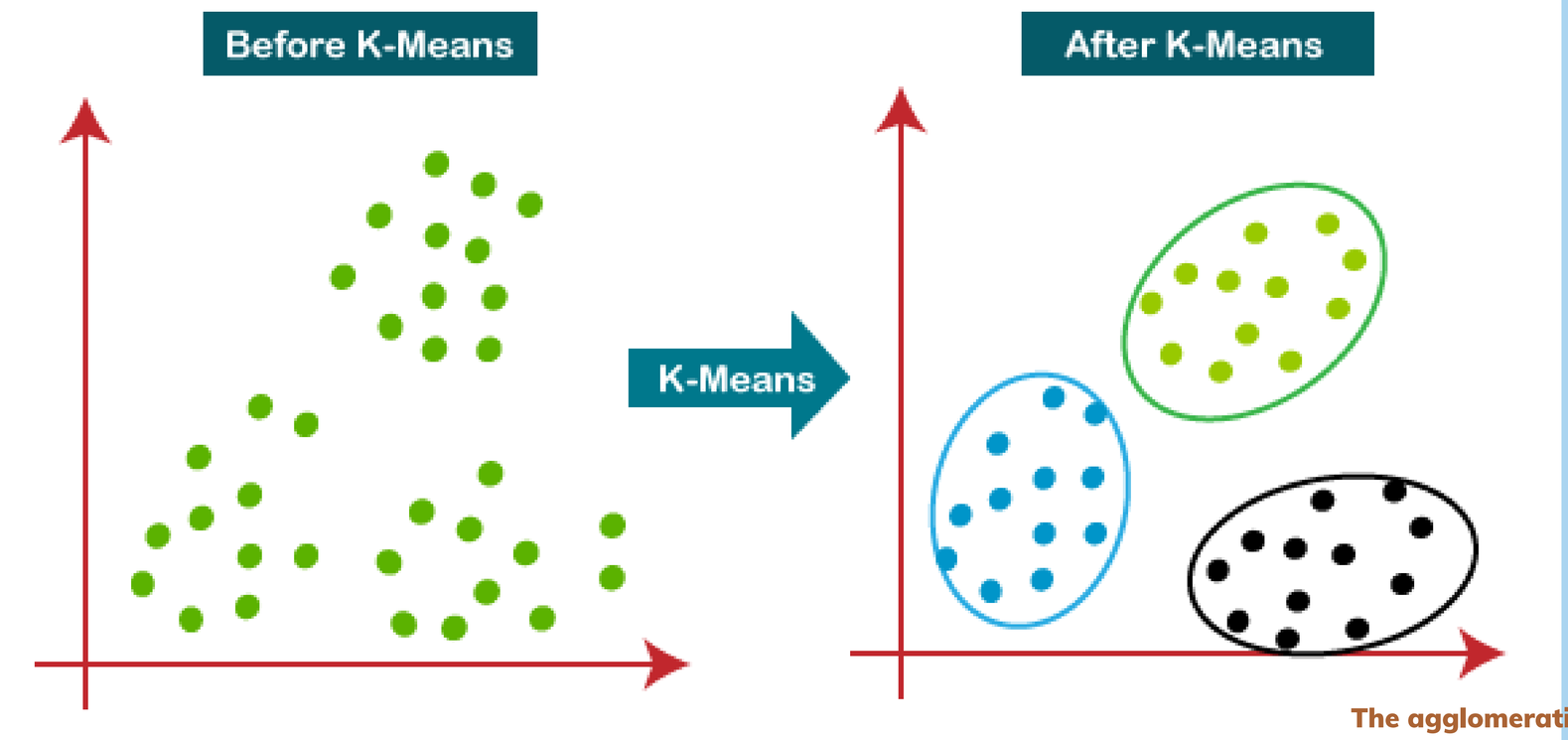
Introduction

I downloaded population data from the World Bank for 200 countries spanning 1960-2022. The data includes attributes like total population, GDP per capita, region, and income level for each country. Then loaded the data into a Pandas dataframe in a Jupyter notebook for cleaning and analysis. I handled any missing values and formatted the columns as needed. To identify groups of countries with similar population trends, I calculated the percentage change in population over time for each country. This growth rate data was normalized before applying k-means clustering to account for different scales. The k-means algorithm identified 3 distinct clusters in the normalized growth rate data. I added these cluster labels back to the original dataset as a new column. Analyzing the clusters revealed insights into different global population trends and associated factors. The clusters provide a data-driven perspective on demographic challenges facing countries at varying development levels. In this poster, I will present the results of the clustering analysis.

Comparing countries and clusters

	Country	Population Growth	Cluster
0	United States	2.3	0
1	United Kingdom	0.6	1
2	Canada	1.2	1
3	India	1.1	1
4	Pakistan	2.0	0

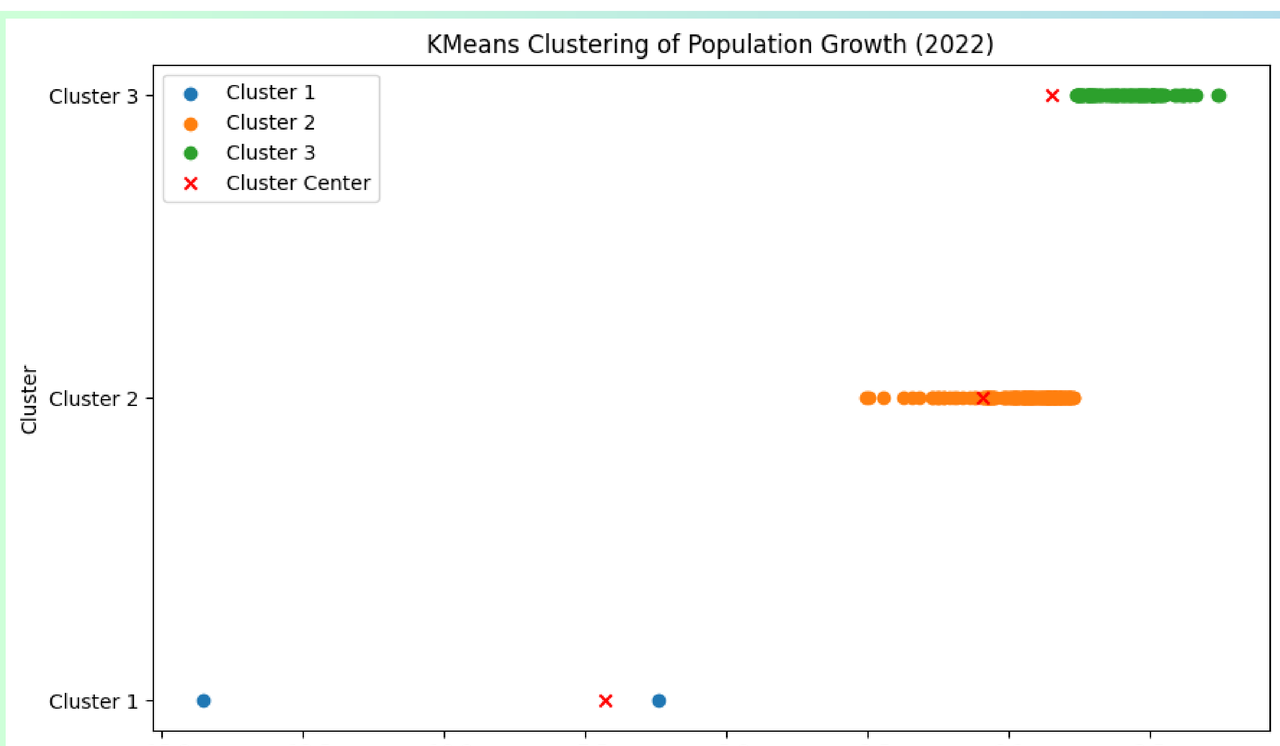
As we can see the countries with the most population growth lies in the same cluster. This is just a small comparison of 5 countries we can by this code confirm that the clusters are correctly made if the population growth of similar countries are in same cluster



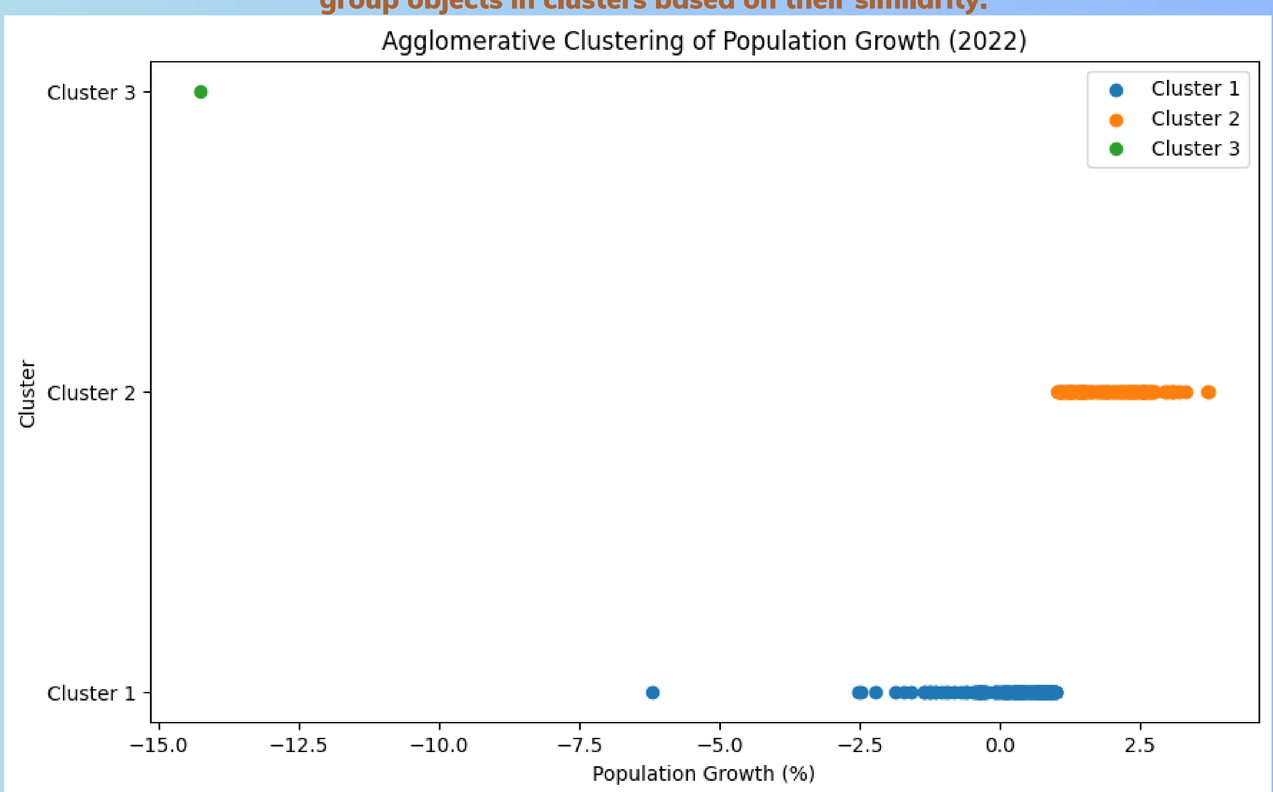
K-means Clustering

- Unsupervised machine learning algorithm
- Groups data points into k clusters
- k is defined upfront
- Centroids initialized randomly
- Assigns points to closest centroid
- Recalculates centroids as cluster means
- Repeats until assignments stabilize
- Minimizes within-cluster variation
- Maximizes between-cluster variation
- Fast, simple, and scalable
- Provides intuitive clusters
- Requires k parameter
- Prone to local optima
- Useful for exploratory data analysis

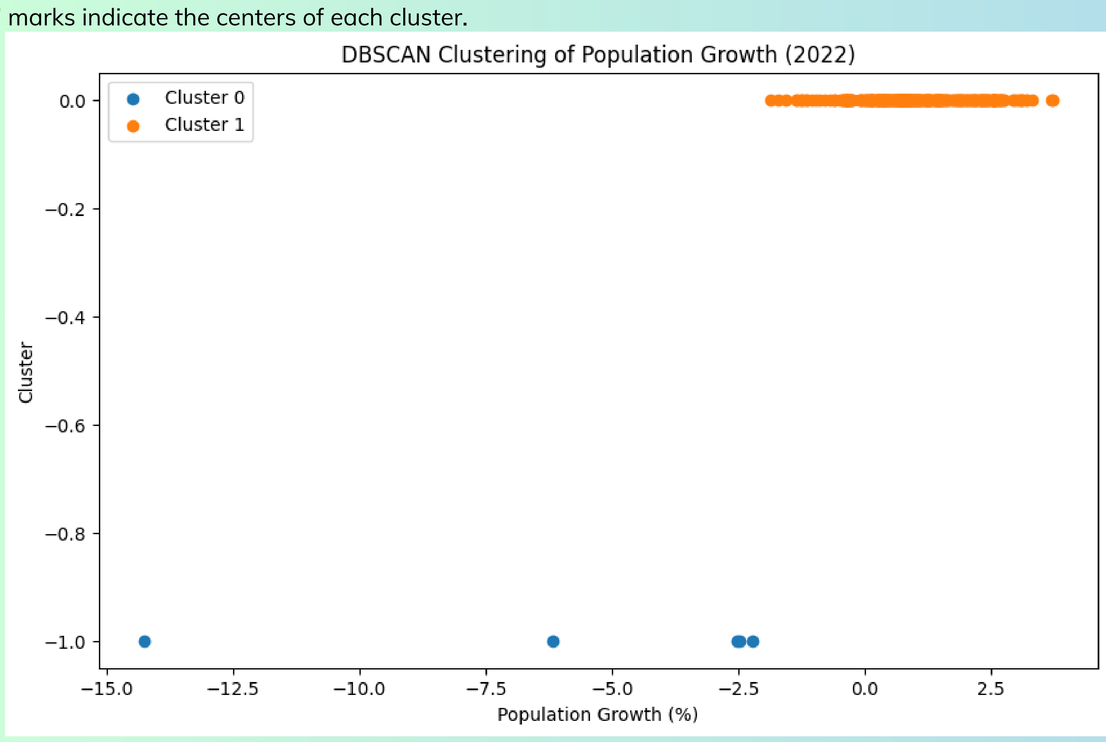
The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity.



The plot displays the clustering of countries based on their population growth in 2022. Three clusters have been identified, with each point representing a country's normalized population growth rate and its assigned cluster. The red 'x' marks indicate the centers of each cluster.

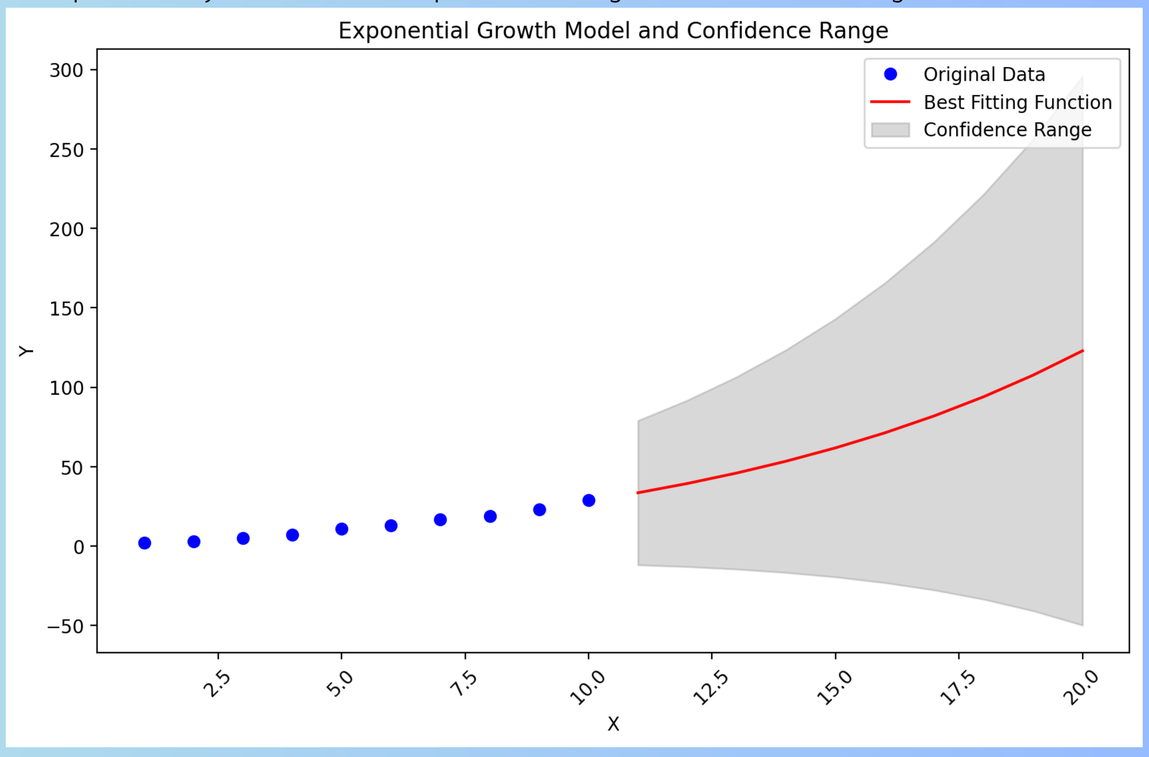


The plot shows the results of Agglomerative Clustering on the population growth data for 2022. Each cluster is represented by a horizontal line of points, indicating the countries that belong to each cluster.



plot illustrates the results of DBSCAN Clustering. This method identifies core samples of high density and expands clusters from them. It is particularly good at identifying outliers, which are shown as points not assigned to any cluster (Cluster 0 in this case).

The DBSCAN algorithm is based on this intuitive notion of "clusters" and "noise". The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points



The plot displays the original data points, the best fitting exponential curve, and the shaded area representing the confidence interval around the predictions.

Conclusion

After applying k means clustering, fitting and comparing clusters we came up with the conclusion that the clustering results are okay and the countries with the same amount of population growth were put together in the same clusters by the K Means algorithm hence we can use this to identify how much population growth of a country must have to fall into some cluster