Data Exploration and Clustering Analysis of Climate Change Factors

Muhammad Hamza Zafar, Student ID:22022247
Supervisor: Ralf Napiwotzki
University of Hertfordshire UH

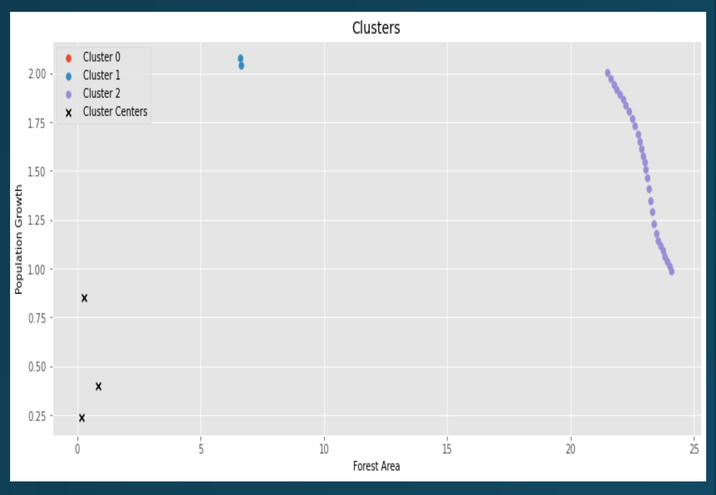
Link For Repo : https://github.com/hamzazafar7/Assignment-No-3.git

Introduction

The goal of this study is to analyse the data collected from twelve different countries regarding three key factors: the Arable land, the Forest area, and the population growth. Using clustering techniques, we will be able to group these countries according to their similarities. Through clustering, we can gain a deeper understanding of a dataset's underlying structure. For instance, knowing the relationship between a country's population growth and its agricultural land area can help us identify areas of vulnerability and opportunities for development. The provided code performs an analysis on climate change data using various data science techniques. It involves data pre-processing, clustering, visualization, bar plot representation, individual country analysis, curve fitting, and future predictions. In terms of data pre-processing, the code reads a CSV file and transforms the data into separate data frames based on countries and years. It also applies normalization to numerical features using the MinMaxScaler from scikit-learn. This ensures that the data is on a consistent scale for further analysis.

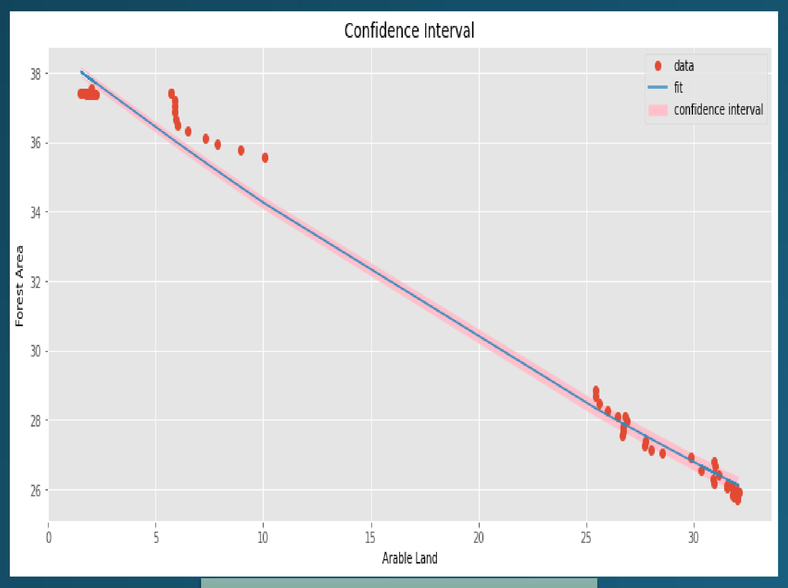
Cluster Fig 1

The cluster scatter plot for India between population growth and forest area suggests that there are two distinct groups (clusters) based on the similarities between these two factors. Cluster 0 has a lower forest area ranging from 5 to 6 and higher population growth of 2. Cluster 1, on the other hand, has a higher forest area ranging from 20 to 25 and a slightly lower population growth ranging from 1 to 2. This information can provide insights into the challenges and opportunities facing India in terms of balancing economic development and environmental conservation.



Cluster Fig 3

The scatter plot analysis for South Africa between arable land and forest area shows that there is only one cluster, cluster 0. This cluster has a forest area between 12 to 14 and arable land between 8 to 12. This suggests that South Africa has a moderate forest area and arable land in comparison to other countries in the dataset. This information can be useful in understanding South Africa's current agricultural and environmental status and in identifying potential areas for improvement or conservation.

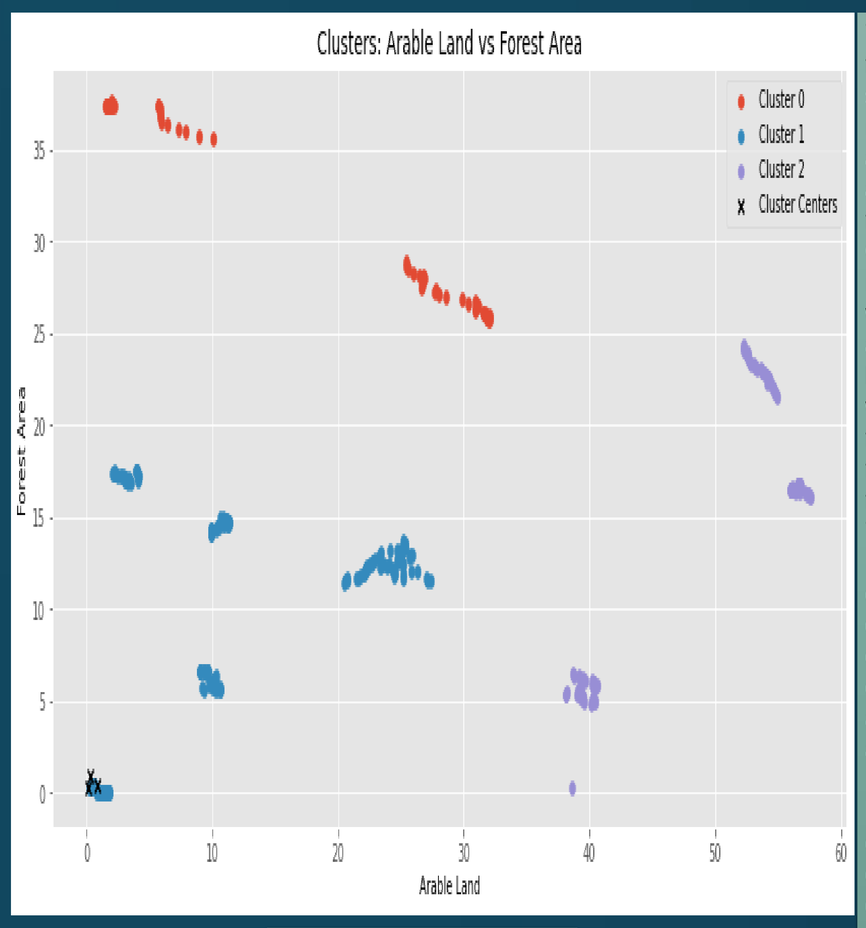


Curve Fitting Fig 5

Abstract:
This code analyzes climate change data, including arable land, forest area, and population growth, using data science techniques. It applies K-means clustering to identify country clusters and visualizes them through scatter plots. Bar plots illustrate changes in arable land over time for specific countries. The code also fits an exponential function to model the relationship between arable land and forest area, estimating confidence intervals. Future predictions of forest area based on arable land are made using the fitted function. This analysis enhances understanding of climate change patterns and their potential impact on different regions.

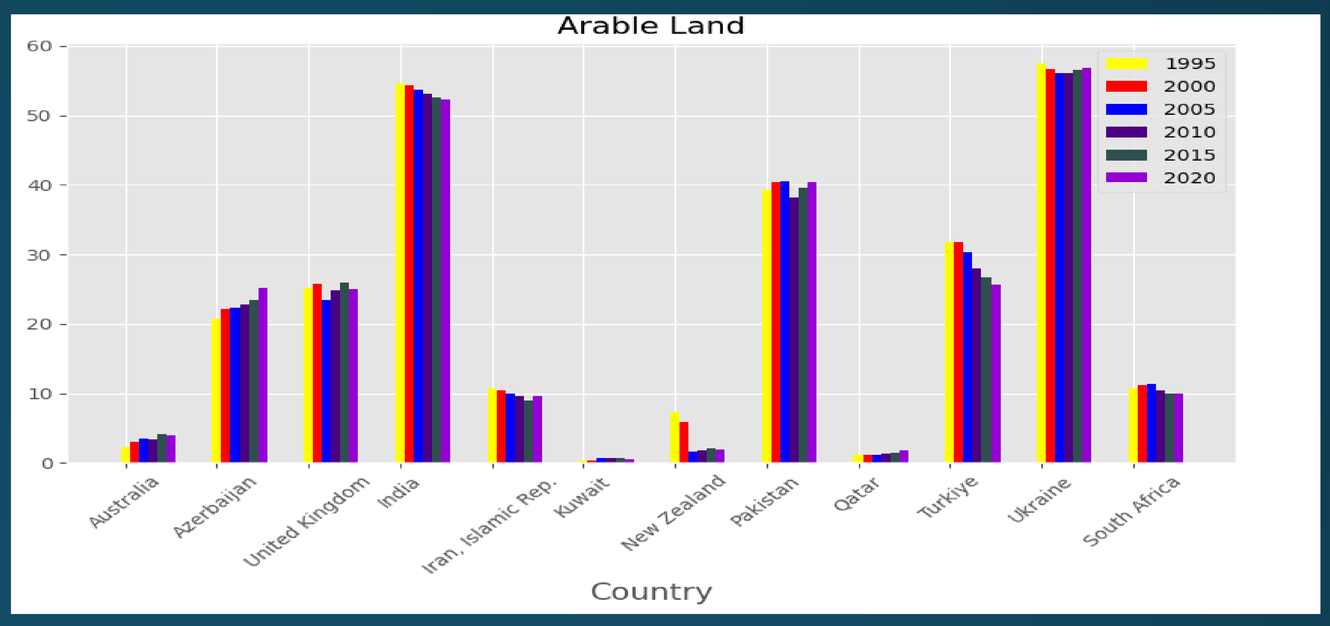
Analysis:
Based on the scatter plot of population growth and forest area, we can see that there are three distinct clusters among the 12 countries analyzed. Cluster 0 includes countries with low forest area (between 0 to 15) and low population growth (between 0 to 20). Cluster 1 includes countries with moderate forest area (between 5 to 25) and low to moderate population growth (f rom -1 to 3). Cluster 2 includes countries with high forest area (between 25 to above 35) and moderate population growth (between 0 to 5).
This analysis provides insights into the relationship between population growth and forest area among different countries. The clustering suggests that there may be different factors influencing the population growth and forest area in each cluster, and that policies aimed at promoting sustainable forest management and population control may need to be tailored to the unique characteristics of each cluster. The code then employs the KMeans clustering algorithm to identify clusters within the dataset. The number of clusters is set to 3 in this case. The resulting clusters are visualized through scatter plots, showcasing relationships between arable land, forest area, and population growth.

Curve fitting is applied to a specific cluster (Cluster 2) using an exponential function. The fitted curve is plotted alongside the actual data points, providing insights into the relationship between arable land and forest area. Confidence intervals are computed to estimate the uncertainty of the fitted curve.



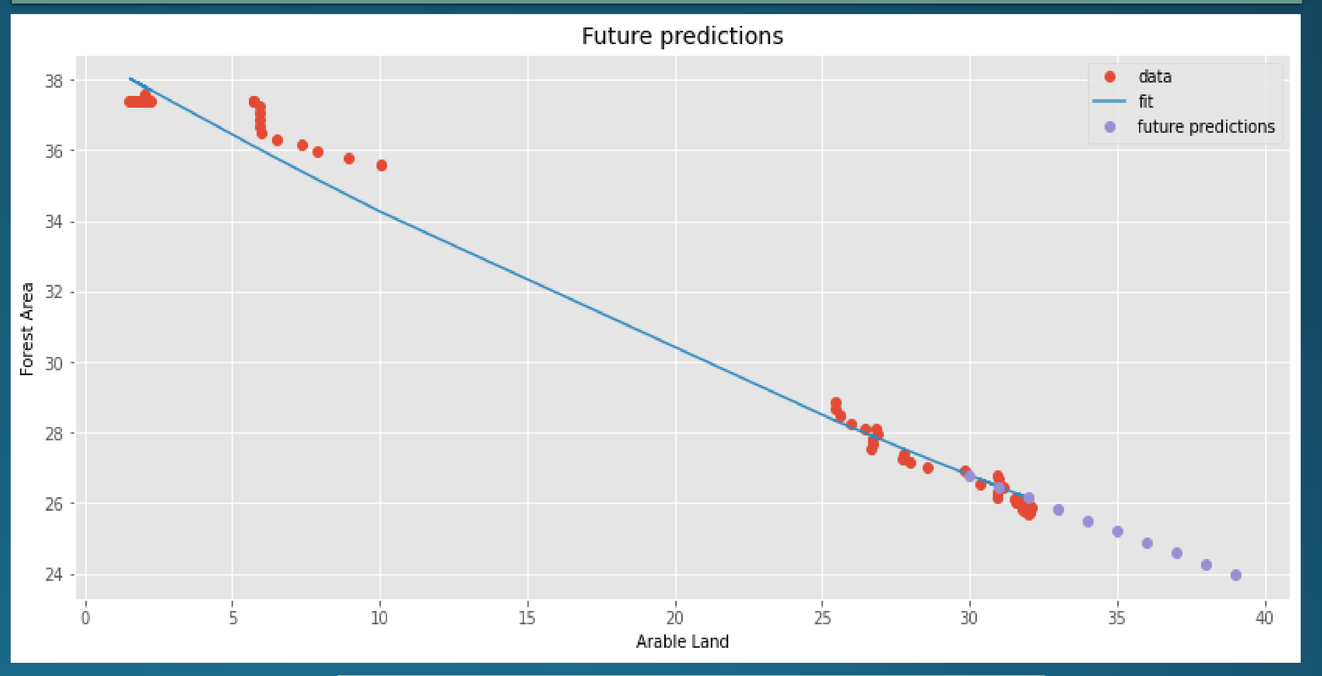
Cluster Fig 2

The scatter plot of arable land and forest area for the selected countries shows three distinct clusters. Cluster 0 represents countries with relatively low arable land and forest area, with arable land ranging from 0 to 30 and forest area ranging from 0 to 20. Cluster 1 represents countries with moderate arable land and forest area, with arable land ranging from 40 to 60 and forest area ranging from 5 to 25. Finally, cluster 2 represents countries with relatively high arable land and forest area, with arable land ranging from 10 to 35 and forest area ranging from 25 to above 35.
These clusters can provide valuable insights into the environmental conditions and challenges facing different nations. For example, countries in cluster 0 may face challenges related to soil quality and agricultural productivity due to limited arable land, while also dealing with issues related to deforestation and loss of forest biodiversity. In contrast, countries in cluster 2 may face challenges related to balancing the demands of agriculture with preserving large forest areas and maintaining ecosystem services. Understanding these patterns and relationships can help inform policies and strategies for sustainable land use and conservation.



Arable Land Bar Plot Fig 4

Based on the bar plot of arable land for the countries it can be observed that India and the United Kingdom have the highest percentage of arable land, with above 50%. Azerbaijan, Iran, Islamic Rep., and Ukraine also have a considerable percentage of arable land, ranging from 20% to 50%. On the other hand, Qatar, New Zealand, Turkey, Australia, and Kuwait have a very low percentage of arable land, with less than 5%.The below plot is finishing our story and this is the future prediction according to the code.



Curve Fitting predicted Values Fig 6

Conclusion:
Based on this analysis, the code provides insights into the clustering of countries based on variables related to climate change. It visualizes the relationship between different variables and identifies clusters in the data. The code also performs curve fitting and makes future predictions using an exponential function. Overall, this analysis helps in understanding the patterns and trends in the data and can potentially support decision-making related to climate change and its impacts.

References:
- https://scikit-learn.org/stable/documentation.html
- https://data.worldbank.org/topic/climate-change
- https://docs.scipy.org/doc/scipy/reference/