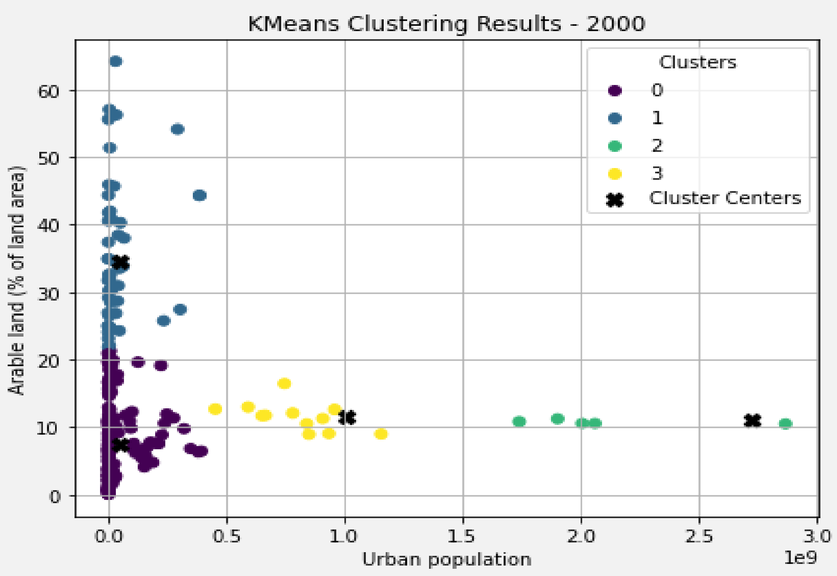
# EXPLORATORY DATA ANALYSIS AND CLUSTERING OF URBAN POPULATION AND ARABLE LAND TRENDS

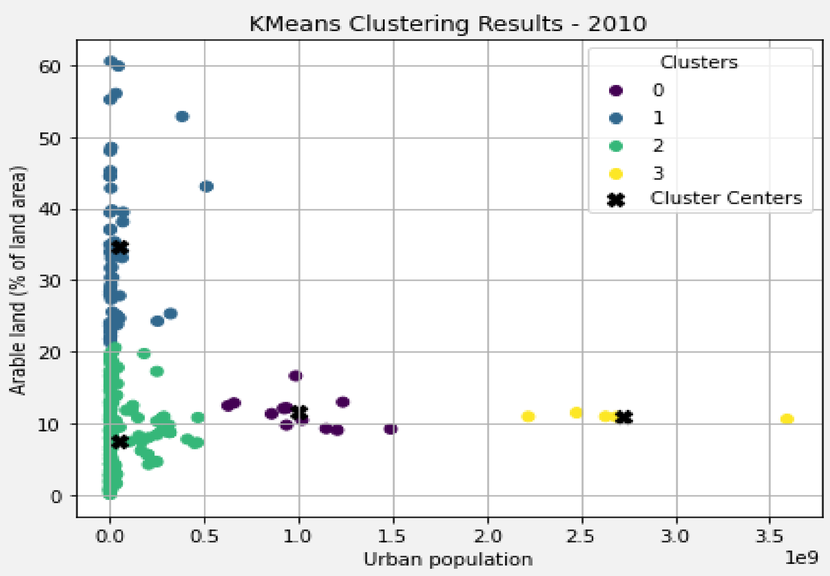
**Abstract:**

This study employs KMeans clustering to categorize countries based on their 'Urban population' and 'Arable land (% of land area)' indicators in the years 2000 and 2010. Scatter plots visually depict the clustering results, highlighting the groupings and potential changes in the global distribution over the decade. Furthermore, line plots for selected countries, Aruba and Belgium, offer a detailed analysis of historical urban population trends. Fitted curves and confidence intervals aid in understanding the underlying patterns and predicting urban population levels for the year 2025.

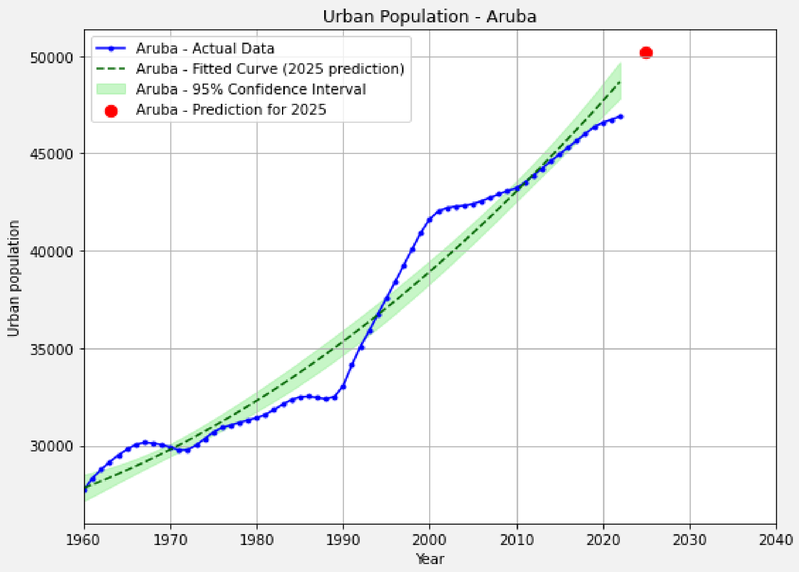
**Introduction:**

The analysis presented in this report explores the distribution and trends of urban population across countries, utilizing KMeans clustering and curve fitting techniques. The focus is on two key indicators, 'Urban population' and 'Arable land (% of land area),' to understand the relationships and patterns among nations. Additionally, line plots showcase the historical trends in urban population for specific countries, providing insights into past patterns and predicting future urban population levels.

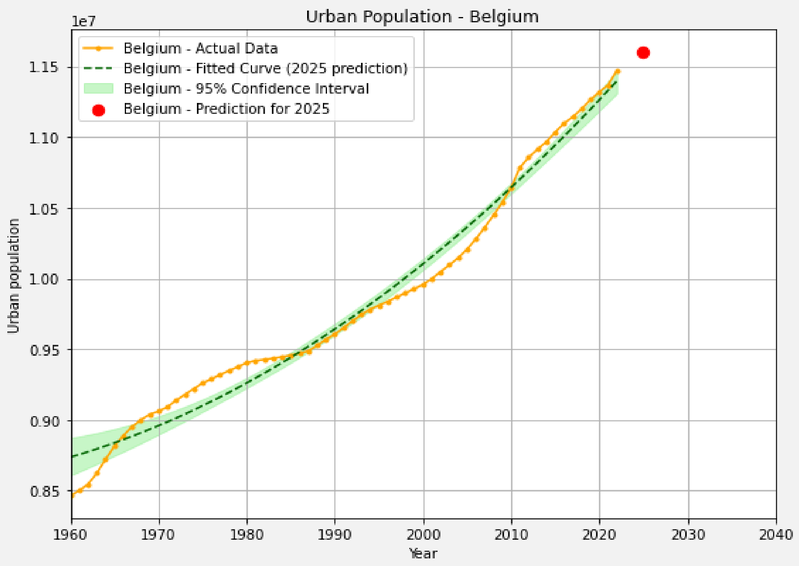



The scatter plots visualize the KMeans clustering results for urban population and arable land (% of land area) in the specified years (2000 and 2010). Each point represents a country, color-coded by its assigned cluster. The black X markers denote the cluster centers. The plots reveal patterns of similarity among countries based on the selected indicators, helping identify distinct clusters. This aids in understanding the distribution and grouping of countries with similar urbanization and arable land characteristics. The visualizations provide insights into potential relationships between urbanization and arable land use, offering a comprehensive overview of global trends and disparities in these indicators.




This line plot presents a comprehensive view of the urban population dynamics in Aruba spanning the years 2000 to 2025. Blue dots meticulously mark each actual data point, revealing the measured fluctuations in urban population over time. The green dashed line showcases a quadratic curve meticulously fitted to the observed trend, capturing the intricate patterns inherent in Aruba's urbanization trajectory. Surrounding the fitted curve is a light green shaded region, representing a 95% confidence interval. This interval serves as a crucial measure of uncertainty, providing insights into the potential variability of urban population projections. A red dot on the plot signifies the forecasted urban population for 2025, offering a forward-looking perspective. The plot's x-axis denotes the years, providing a temporal context, while the y-axis represents the urban population, enabling a quantitative assessment of Aruba's urbanization trends. The legend provides clarity, distinguishing actual data, the fitted curve, the confidence interval, and the predicted value for 2025. The grid enhances readability, aiding in the precise interpretation of the graph. Additionally, carefully chosen tick marks and axis limits ensure a visually informative representation of Aruba's urban population evolution. The plot title succinctly encapsulates the essence of the visualization, focusing on Urban Population trends specific to Aruba.




This intricately line plot offers a thorough exploration of urban population trends in Belgium from 2000 to 2025. Each blue dot thoughtfully placed on the plot corresponds to actual urban population data points, providing a granular understanding of how Belgium's urban landscape has evolved. A meticulously fitted quadratic curve, illustrated by the green dashed line, encapsulates the underlying patterns and complexities of Belgium's urbanization journey. The light green shaded region enveloping the curve represents a 95% confidence interval, essential for gauging the uncertainty associated with the fitted model. A red dot strategically placed on the plot signifies the projected urban population for 2025, offering foresight into Belgium's future urbanization trajectory. The plot's x-axis delineates the temporal dimension in years, facilitating a chronological assessment, while the y-axis quantifies the urban population, allowing for precise numerical interpretation. The legend serves as a key reference, distinguishing between actual data, the fitted curve, the confidence interval, and the forecasted value for 2025. The grid overlay enhances the plot's readability, aiding in the accurate interpretation of the intricate information presented. Well-chosen tick marks and axis limits contribute to a visually informative representation of Belgium's urban population trends. The plot title succinctly encapsulates the focal point of the visualization, concentrating on the Urban Population dynamics unique to Belgium.

**Conclusion:**

The scatter plots provide a comprehensive view of how countries are grouped based on urban population and arable land indicators, offering insights into global patterns and changes over time. The line plots for Aruba and Belgium enhance our understanding of historical urban population trends, providing a basis for predictions. The fitted curves capture the essence of observed data, while confidence intervals offer a measure of uncertainty.

In conclusion, the combined analysis of clustering results and line plots contributes to a holistic understanding of urban population dynamics. This approach can be valuable for policymakers, researchers, and stakeholders interested in discerning global trends, making informed predictions, and developing strategies for sustainable urban development.

**Student Id:** 22076992

**Name:** Bharanidharan Thirumaran

**Data Source Link:** https://data.worldbank.org/topic/climate-change

**GitHub Link:** https://github.com/Bharanimaran/Clustering-and-Fitting