University of Hertfordshire UH

Name: Sreevarshini Boga
Student ID: 22095406

Exploratory Data Analysis And Modelling

Abstract

This study looks into patterns of urbanisation through a thorough analysis of a dataset that includes significant factors including urban population, rents from natural resources, population in rural areas, and population in urban agglomerations of more than a million. The study employs K-Means clustering to categorise countries based on factors associated with urbanisation, taking into account techniques for handling missing values and standardisation in data preparation. The quality of clustering is assessed using the Silhouette Score. Additionally, a time series analysis is performed, focusing on statistics related to the urban population over time. Curve fitting is used to model the temporal patterns and forecast future projections of the urban population. Confidence intervals are computed to provide a level of forecast uncertainty.

First of all,

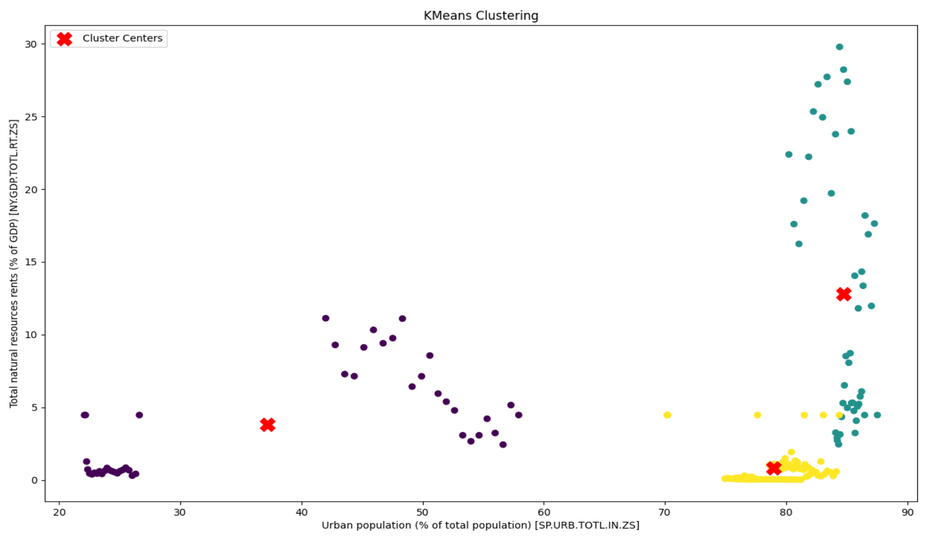
This study includes an examination of data related to urbanisation using curve fitting and clustering (K-Means) approaches. The collection includes statistics on population in urban agglomerations of more than one million, population in rural areas, rents associated with natural resources, and urban population.

Data Pre-processing

- Managing Missing Values: Non-numeric inputs and NaN values are handled by first being converted to NaN and then having the mean of the corresponding columns filled in.
- Data Normalisation: The data is normalised to ensure that variables with different scales do not disproportionately influence clustering.

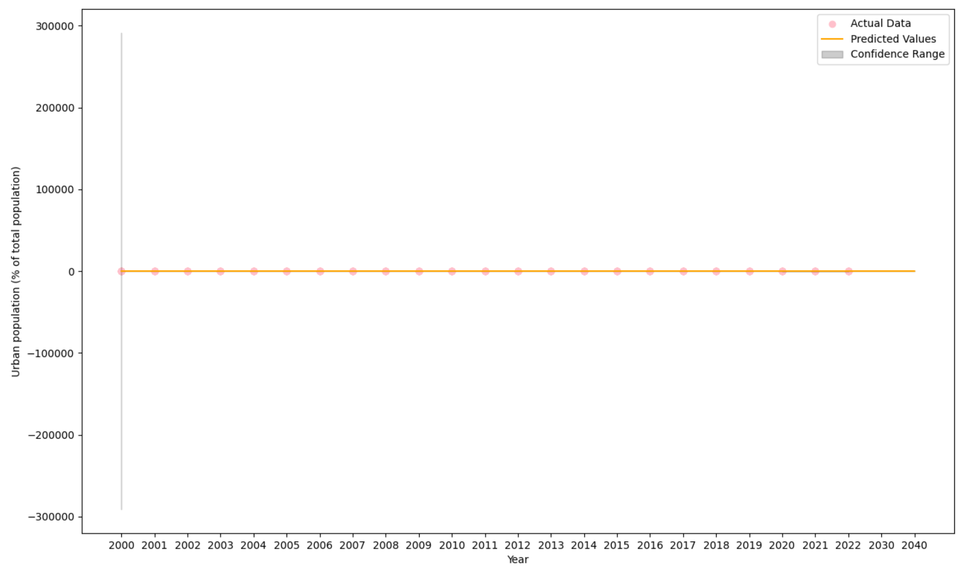
K-Means Clustering

- Number of Clusters: The normalised data is subjected to three-cluster K-means clustering.
- Silhouette Score: A metric used to assess the degree of cluster definition is the Silhouette Score. Better-defined clusters are indicated by a higher score. For this study, the silhouette score is 0.5682787771347736.
- Here, we see that nations with large urban population densities also have very low to high fractions of natural resources.



Curve Fitting

- **Time Series Data Selection:**
  Time series analysis is applied to data on urban population over time.
- **Curve Fitting:**
  Using curve fitting techniques, a basic model, a polynomial of degree 1, is fitted to the time series data.
- **Prediction:**
  Forecasts of urban population figures for the years 2020, 2030, and 2040 are made using the fitted model.
- **Confidence Intervals:**
  The range of anticipated values is estimated using confidence interval calculations.



Predictions

Predictions	
Year	**Urban Population**
2020	32.05930622
2021	38.76567121
2022	49.551961

Data source: https://databank.worldbank.org/reports.aspx?source=2&series=AG.LND.FRST.ZS&country=
GitHub Link: https://github.com/varshiniboga/clusteringandFitting.git

Conclusion

This research forecasts future trends in urban population and offers insights into how countries are clustered based on criteria connected to urbanisation. Time series prediction is made easier by the curve fitting technique, while the Silhouette Score provides an assessment of the quality of clustering.