PROJECT REPORT

Developing Holistic Ranking System using Global Indexes

by

Student Name	Reg No
Soham Kumar	20BAI1167
Arnab Mondal	20BCE1294
Saptarshi Mukherjee	20BCE1719

A project report submitted to

Dr. Rajalakshmi R

50879

SCOPE

In fulfillment of the requirements for the course of

CSE3506 - Essentials of Data Analytics

In

B.Tech Computer Science and Engineering

Vandalur – Kelambakkam Road

Chennai – 600127

Winter 2022-2023

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. R. Rajalakshmi,** School of Computer Science and Engineering for her consistent encouragement and valuable guidance offered to us throughout the course of the project work.

We are extremely grateful to **Dr. R. Ganesan, Dean,** School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our **Head of the Department** for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

Soham Kumar ArnabM

Soham Kumar Arnab Mondal Saptarshi
Mukherjee

(20BAI1167) (20BCE1294)
(20BCE1719)

BONAFIDE CERTIFICATE

Certified that this project report entitled "**Developing Holistic Ranking System using Global Indexes"** is a bona-fide work of **Soham Kumar (20BAI1167), Arnab Mondal (20BCE1294), Saptarshi Mukherjee (20BCE1719) carried** out the "J"-Project work under my supervision and guidance for **CSE3506 - Essentials of Data Analytics**.

Dr. R.

Rajalakshmi

SCOPE

TABLE OF CONTENTS

Developing Holistic Ranking System using Global Indexes

Soham Kumar$^{3[20BAI1167],}$ Arnab Mondal$^{1[20BCE1294]}$, Saptarshi Mukherjee$^{2[20BCE1719]}$

[1] VIT University, Chennai, Tamil Nadu, India

Abstract. The work proposes a collective, holistic ranking system for countries based on various factors, such as economic freedom, environmental sustainability, and human well-being. While individual country rankings based on factors such as GDP, education, and healthcare are common, they lack a comprehensive world ranking system. Moreover, the classification of countries into developed, developing, and underdeveloped categories can be subjective and inaccurate. Three world indexes - the Human Development Index (HDI), the Social Progress Index (SPI), and the Innovation Index (II) were selected as reference datasets. The data was reduced performing principal component analysis and merged to create a new dataset. The missing score were predicted using CART and various clustering techniques were explored to best categorize the regions. Independent component analysis was performed on the final dataset and visualiza- tions were made based on the resulting clusters. The proposed system also has the potential to improve international relations and facilitate significant decision-making.

Keywords: Principal Component Analysis, Clustering, Global Ranking

1 Introduction

For decades, countries have been classified into developed, developing, and underdeveloped categories based on various factors such as GDP, education, healthcare, and human development. However, this classification lacks a standardized, collective, and holistic ranking system that could provide a more accurate assessment of how countries are performing on a world level.

The current classification is subjective and based solely on the perception of geo-political analysts, which can be biased and may not reflect the true state of a country. This makes it challenging for countries to identify their strengths and weaknesses, develop appropriate strategies to address them, and make informed decisions in international relations.

To address this problem, this paper proposes a data analytics system that collectively ranks countries based on various factors and classifies them as developed, developing, and underdeveloped. The system involves selecting three world indexes - the Human Development Index (HDI), the Social Progress Index (SPI), and the Innovation Index (II) - as datasets. These indexes cover a wide range of factors that affect a country's overall development, such as economic freedom, environmental sustainability [7], innovation [8] and human well-being [6].

By applying Principal Component Analysis (PCA) and Independent Component Analysis (ICA) on the merged dataset, this paper aims to extract the most important variables and identify the underlying factors that contribute to a country's development. By clustering the countries based on the scores obtained from ICA, this paper aims to group countries into appropriate categories and generate a more accurate and standardized assessment of how countries are performing on a world level.

Overall, the proposed system can help leaders and policymakers of individual countries to identify their countries' strengths and weaknesses, develop appropriate strategies, and make informed decisions in international relations.

2 Related Works

García-Fernández et al. 2017 proposes a holistic index to assess the sustainability of rural tourism destinations by considering environmental, economic, and socio-cultural aspects. It identified key indicators of sustainability for rural tourism destinations and found that the holistic sustainability index was effec- tive in capturing the complexity of sustainability issues in these destinations. However, the index could be further refined to include more specific indicators of sustainability and adapted for use in other types of tourism destinations be-yond rural areas. [1]

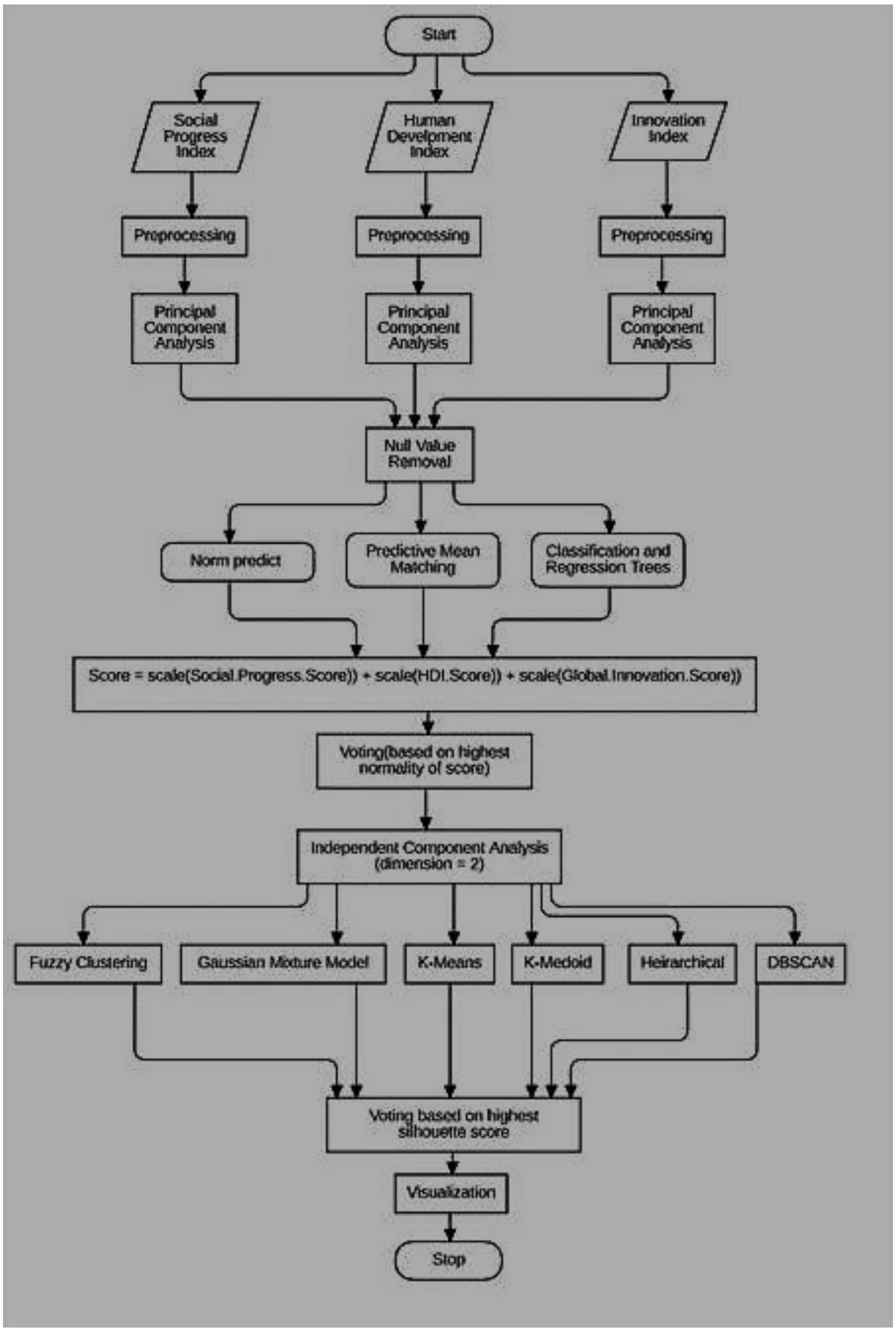
Wai Yee Lam et al. 2018 created a holistic index to measure the social sustainability of urban neighborhoods, which takes into account economic, environmental, and social dimensions of sustainability. The paper identified key indicators of social sustainability, such as community cohesion, access to health care and education, and social inclusion, and found that the holistic index was effective in capturing the complexity of social sustainability issues in urban neighborhoods. The index could be further refined by incorporating more detailed indicators of social sustainability, such as measures of crime and safety, or by testing it in different urban contexts to evaluate its applicability and generalizability. [2]

Roberto Crotti et al. 2015 proposes a holistic approach to sustainable development that challenges the traditional economic paradigm by emphasizing the importance of environmental, social, and economic considerations. The paper explores the limitations of the existing economic paradigm and provides a framework for a new, more holistic approach to sustainable development that prioritizes the long-term health of the planet and its inhabitants. The paper could be further developed by outlining specific policy recommendations or by examining case studies of successful implementation of the proposed new economic paradigm in practice. [3]

Mona Khare et al. 2020 analyzed the trends and strategies towards internalizing higher studies in India and develop a ranking based on that. The paper analyzed the existing literature on internalization of higher studies in India, identified the key strategies being used by universities to enhance internationalization, and developed a ranking system based on the extent of internationalization of Indian universities. The paper could be further developed by conducting a survey or interviews with stakeholders to gain a more nuanced understanding of the challenges and opportunities for internationalization of higher studies in India. [4]

Kostiantyn Niemets et al. 2017 proposes to develop a holistic ranking system for sustainable cities that takes into account a range of economic, environmental, and social indicators. The paper developed a ranking system that used 37 indicators across these three dimensions to rank 28 cities in Ukraine, providing insights into areas where cities can improve their sustainability performance. The paper could be further developed by testing the ranking system in other contexts, incorporating additional indicators that are relevant to specific regions or contexts, and incorporating stakeholder perspectives to ensure that the ranking system reflects the priorities and values of local communities. [5]

3 Proposed Methodology


Start
Social Progress Index
Human Development Index
Innovation Index
Preprocessing
Preprocessing
Preprocessing
Principal Component Analysis
Principal Component Analysis
Principal Component Analysis
Null Value Removal
Norm predict
Predictive Mean Matching
Classification and Regression Trees
Score = scale(Social.Progress.Score)) + scale(HDI.Score)) + scale(Global.Innovation.Score))
Voting(based on highest normality of score)
Independent Component Analysis (dimension = 2)
Fuzzy Clustering
Gaussian Mixture Model
K-Means
K-Medoid
Heirarchical
DBSCAN
Voting based on highest silhouette score
Visualization
Stop

The process begins by selecting various world indexes, with a focus on the Social Progress Index (SPI), Human Development Index (HDI), and Innovation Index (II). These indexes are used to identify trends and insights in the data. Prior to working with the data, pre-processing techniques are applied, including country name standardization, outlier removal, and data standardization as number of regions is not same in every dataset. Additional pre-processing techniques, as outlined in the document and RMD file, are also incorporated.

After the pre-processing of the data, Principal Component Analysis (PCA) is applied to the dataset. This machine learning technique is used to reduce the dimensionality of the dataset while explaining 90% of the variance, thereby overcoming associated problems such as overfitting. However, the datasets may have different numbers of regions, leading to missing values when merging the datasets. Therefore, special attention is given to handle null values during the merging process based on countries. To address this issue, three machine learning methods, namely norm predict, predictive mean matching, and Classification and Regression Trees (CART), are used to remove the NA values. The final score is calulated by adding the scaled values of scores of Social Progress Index (SPI), Human Development Index (HDI), and Innovation Index (II), and the best NA removal technique is selected through a voting process based on the best normality of score.

Following the NA removal step, Independent Component Analysis (ICA) with dimension 2 is applied to the dataset. This machine learning technique is used to separate the multivariate signal into its independent components, thereby identifying the underlying independent sources that contribute to a set of observed signals.

Various clustering techniques, including Fuzzy clustering, Gaussian Mixture Model (GMM), k-means, k-medoid, hierarchical, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), are employed to group the countries into clusters. The best clusters are selected based on a voting process through the highest silhouette score, which measures how similar an object is to its own cluster compared to other clusters.

Finally, the resulting dataset and holistic ranking are visualized to gain insights from the analysis. This involves using data visualization techniques to represent the data in a meaningful way, such as density plots, pie charts and world map based on cluster. Overall, the application of various machine learning techniques, such as PCA and ICA, in combination with data pre-processing and clustering methods, allows for the extraction of valuable insights from multiple data.

4 Results and Discussion

4.1 Analysis for Social Progress Score Dataset

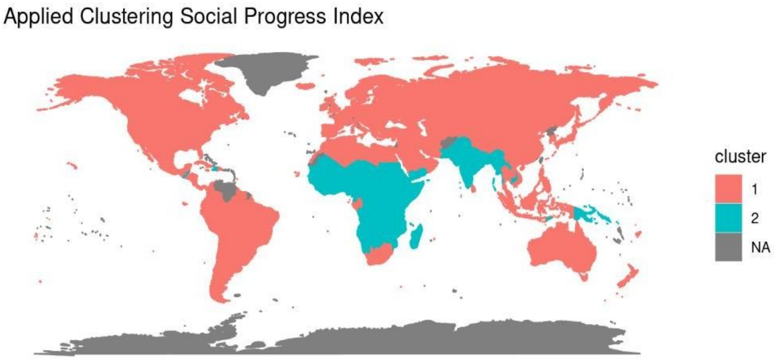


Fig. 1.1. Applied Clustering social progress score

The figure above represents the no. of clusters formed with respect to the Social Progress Score Dataset. The clusters 1 and 2 as clearly visible from the above figure shows that the clusters are uniformly separated among one another in the globe.

4.2 Analysis for Human Development Score Dataset

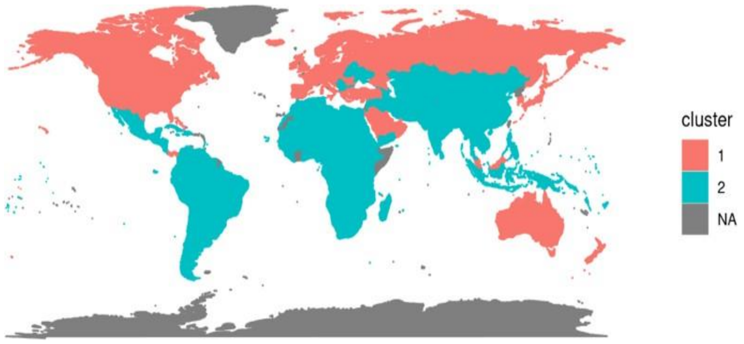


Fig. 2.2. Applied clustering Human Development Score

The figure above represents the no. of clusters formed with respect to the Human Development Score Dataset. The clusters are uniformly separated and distributed among various parts of the globe as observed from the above world map.

4.3 Analysis for Global Innovation Index Score Dataset

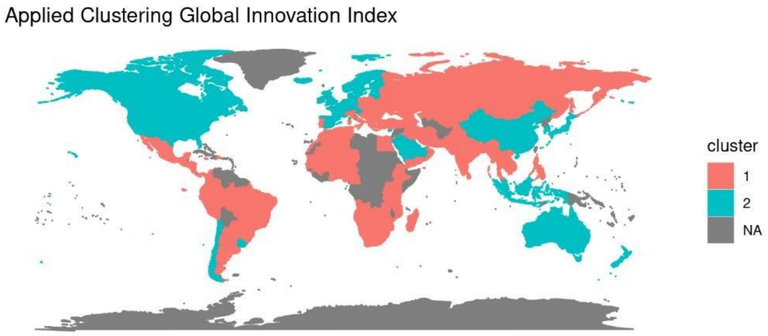
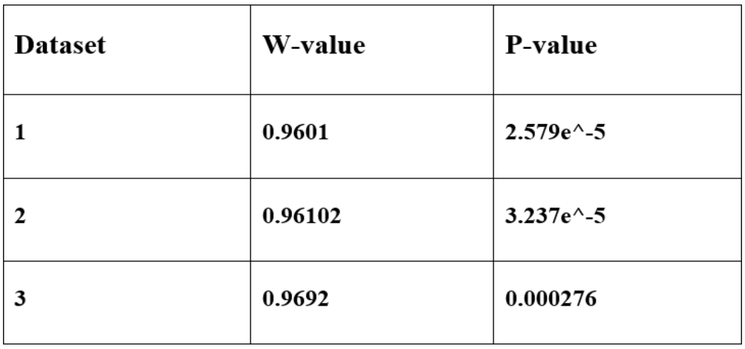


Fig. 3.3. Applied Clustering Innovation index

The figure above represents the no. of clusters formed with respect to the Global Innovation Index Dataset. Both the clusters have been aptly distributed across various portions of the globe as it can be seen from the above world map.

4.4 Selecting Dataset with the Best Normality of Scores

After removal of NULL values using Norm predict, Predictive mean matching and CART on the datasets and then applying Shapiro wilk test to obtain the following results:



Dataset	W-value	P-value
1	0.9601	2.579e^-5
2	0.96102	3.237e^-5
3	0.9692	0.000276

The P-value of dataset 3 is the highest. Hence, it shall be selected for further computations.

4.5 Performing various Clustering techniques and doing Cluster Analysis

SL.NO	CLUSTERING TECHNIQUE	SILHOUETTE SCORE
1	K-MEANS	0.486337

2	AVERAGE LINKAGE	0.4333979
3	COMPLETE LINKAGE	0.4354677
4	K-MEDOID	0.4683219
5	GMM	0.4542116
6	DBSCAN	0.417542
7	FUZZY	0.462874

After performing various Clustering Techniques, we conclude that K-Means Clustering gives us the maximum Silhouette Score and we perform the following visualizations based on the above techniques to have some important findings and inferences.

4.5.1 Visualizations based on the best Clustering Technique

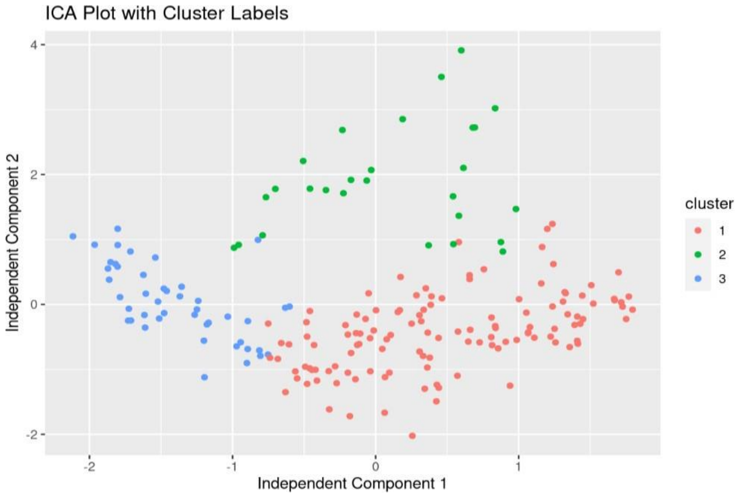


Fig. 4.4.1. ICA plot with cluster labels

The figure above represents the no. of clusters formed based on K-means clustering technique. We see that 3 clusters have formed which has been represented by 3 different colors to have a differentiating effect.

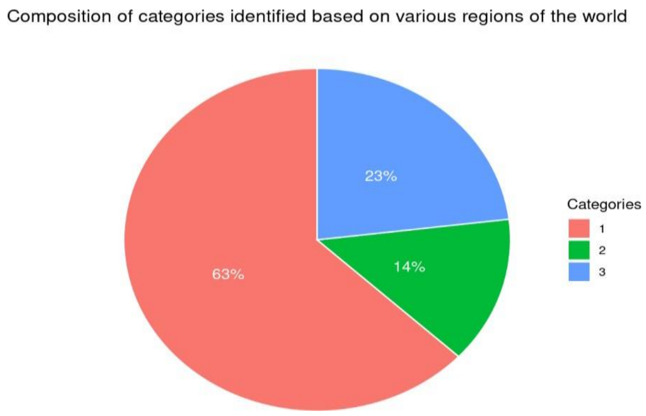


Fig. 5.4.2. Composition of Categories identified based on various regions of the world

The figure above represents the no. of clusters formed based on K-means clustering technique on a bar plot. We see that 3 clusters have formed where 23% constitutes for Category-1, 14% constitutes for Category-2 and 63% constitutes for Category-3.

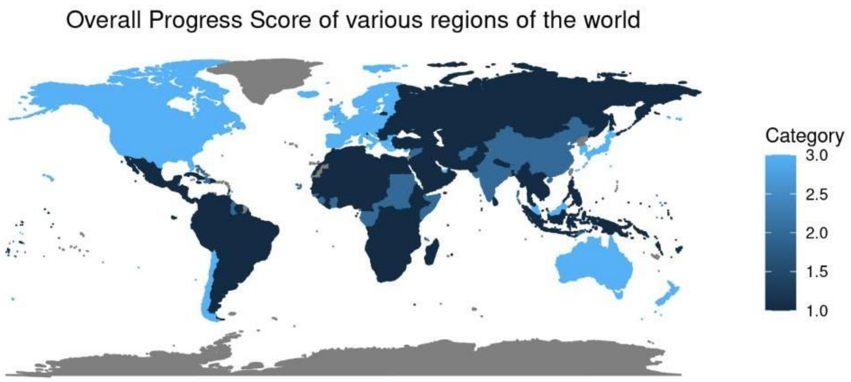


Fig. 6.4.3. Overall Progress score of various regions of the world

The light blue color in the world-map above depicts the developed nations, dark blue color depicts the developing nations and the blackish color depicts the under-developed nations.

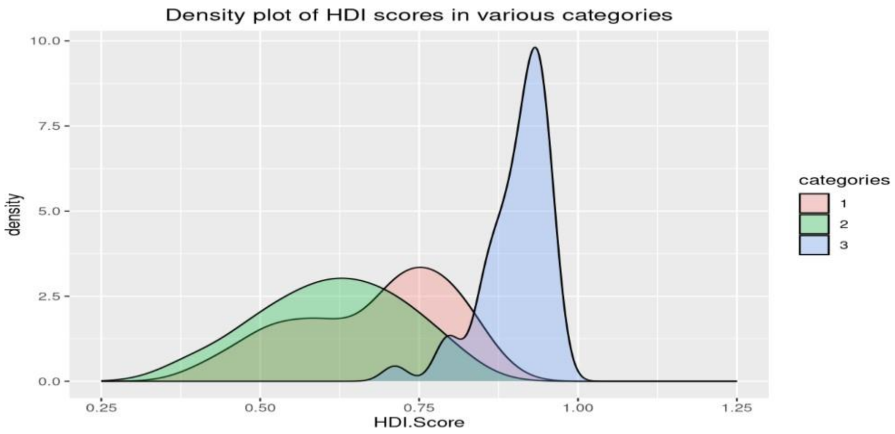


Fig. 7.4.4. Density Plot of HDI scores in various categories

The **highest HDI** score lies for all the **Category-3**. Then comes the **Category-1** followed by the **Category-2.**

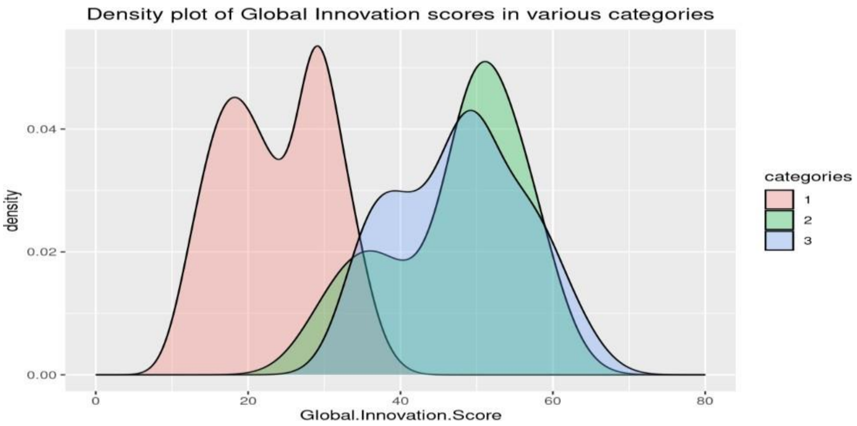


Fig. 8.4.5. Density plot of Global Innovation scores in various categories.

The **highest Global Innovation scores** lies for all the **Category-2** countries followed by the **Category-3** and then the **Category-1 nations,** respectively.

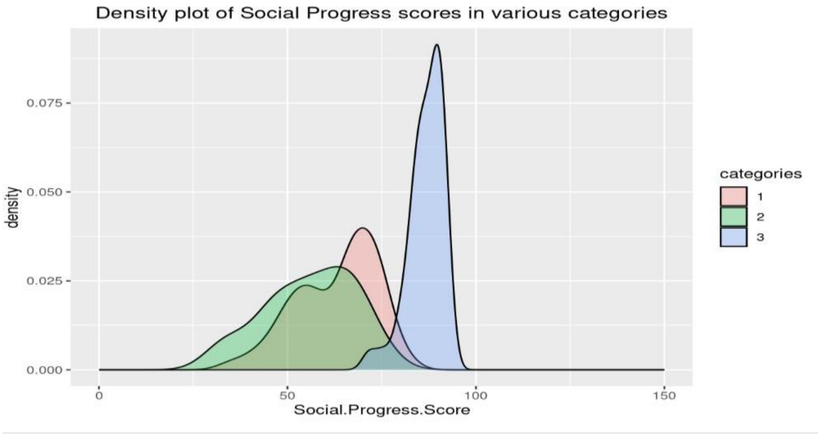


Fig. 9.4.6. Density plot of social progress score in various categories.

The **Highest** Social Progress score lies for the **Category-3 nations**, followed by the **Category-1 nations** and then the **Category-2 nations**.

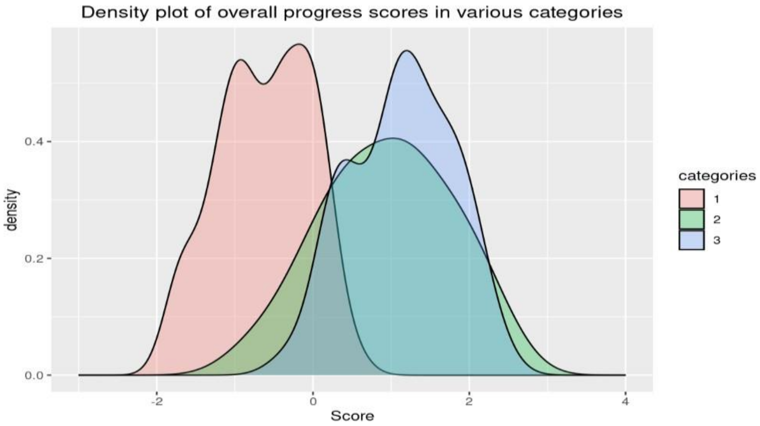


Fig. 10.4.6. Density plot of overall progress scores in various categories.

The **Highest overall progress** score lies for the **Category-3** nations, followed by the **Category-2** nations and then the **Category-1** nations.

5 Conclusion

The proposed work has successfully ranked countries based on their Social Progress Score, Human Development Score, and Global Innovation Score. Finally, we categorized the countries into three clusters, with a silhouette score of 0.48. Based on our observations, North America, Australia, New Zealand, Japan, South Korea, Europe, and Chile fall under category 3, which shows a higher band of scores and are marked as developed countries. Countries like India, China, South Sudan, Republic of Congo, and Central African Republic fall into category 2, which falls into the moderate score band and are marked as developing countries. Regions like most of Africa, South America, Central Asia, and Southeast Asia fall into the lowest score band and are marked as poor-performing countries.

K – means, K – medoid and Fuzzy clustering provides promising results for clustering in such scenarios. Dimensionality reduction and factor analysis is very crucial for extracting the most important factors which affects the country's category. However further work can be done by including more indexes for comparison. The results will vary each year so the source datasets need to be updated every year and fed into the system to get the new resuts.

6 References

I. "Development of a Holistic Index for Assessing the Sustainability of Rural Tourism" by García-Fernández, J.L., López Hernández, A., & Álvarez-Santana, E. (2017). Development of a Holistic Index for Assessing the Sustainability of Rural Tourism. Sustainability, 9(11), 1981."

II. **"**Development of a Holistic Index to Measure Social Sustainability of Urban Neighborhoods" by Wai Yee Lam and Edwin H.W. Chan. Lam, W.Y., & Chan, E.H.W.

(2018). Development of a Holistic Index to Measure Social Sustainability of Urban Neighborhoods. Sustainability, 10(2), 452.

III. “A Holistic Approach to Sustainable Development: The Need for a New Economic Paradigm" by Roberto Crotti and Richard Knight. Crotti, R., & Knight, R. (2015). A Holistic Approach to Sustainable Development: The Need for a New Economic Paradigm. Sustainability, 7(8), 9833-9852.”

IV. “Trends and Strategies towards Internalizing Higher Studies in India and developing a ranking based on that: A Case Study of Indian Universities" by Mona Khare. Khare. M.(2020).

V. “Development of a Holistic Ranking System for Sustainable Cities" by Kostiantyn Niemets a, Kateryna Kravchenko a, Yurii Kandyba a, Pavlo Kobylin a, Cezar Morar b (2017). Development of a Holistic Ranking System for Sustainable Cities. Sustainability, 9(4), 530. ”

VI. United Nations Development Programme (UNDP). (2019). Human Development Indices and Indicators: 2019 Statistical Update.

VII. Social Progress Imperative. (2021). 2021 Social Progress Index.

VIII. Global Innovation Index. (2021) WIPO