Data Validation and Prediction for the Proposed JDI Index

Team 2

Yahan Zhang
Kamal K. Khanal
Tejaswini Ganti

UNIVERSITY OF DELAWARE

DISCLAIMER

This analysis and presentation do not necessarily reflect the views or policies of the United States Department of Labor, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government.

Agenda

- **Introduction**
  - Goal
  - Idea
  - JDI Methodology
- **Research -** Dataset with JDI Country Score
  - Regression Diagnosis *(mean & median)*
  - Classification and Clustering Analysis
- **Conclusion**

Goal

- Provide insight on choice of data set for predicting the JDI Score.
- Provide a strong analytical approach to predict the JDI scores for the countries without enough data.

Idea

- The aim of data validation and prediction is to provide a general scoring platform for the global jewelry sector to facilitate comparative socio-economic performance analysis of different countries by using the data that is already present from a few countries.
- This research works aims to validate the data collected and developed by the students of American University and develop the analytical method which will help in analyzing the current score and predict the score of the other countries.

JDI Methodology

FRAMEWORK

Governance

- Accountability
- Transparency
- Corruption
- Industry Regulation
- Criminal organization/ Non-state actors

Economy

- Industry Employment
- Fiscal Sustainability
- Beneficiation
- Informal Economy
- Criminal non-state Actors
- Supply chain

Environment

- Environmental Regulation and Enforcement
- Pollution
- Biodiversity
- Post-production planning and remediation

Human Health

- Human Health
- Water Security
- Food Security

Human Right

- Workers' rights
- Indigenous people's rights
- Women's rights
- Children's' rights
- Freedom from Violence

Regression Diagnosis

Data Set(s)

- Data sets collected and developed by American University students.
- Consists of total JDI Score for 10 countries along with country level score for different indicators (Risk to Governance, Risk to Economy, Risk to Environment, Risk to Health, Risk to Human Right and Total JDI Score for each countries).

Regression Diagnosis - Mean

Test of Linear Regression Assumptions

- Scatter plot of dependent variable (Total Score) against each of the independent variable, studentized residual, leverage value, influence statistics like DFFITS and Cook's D Statistics shows there is no outlier.
- Variance inflation factor, tolerance value, proportion of variation and condition index indicates there is no evidence of multicollinearity.
- Test of normality (Histogram, QQ plot) shows that the residual are normally distributed.

Regression Diagnosis - Mean

Multiple Linear Regression

- Regression coefficients are not uniquely determined.
- One unit change in each indicator will change 0.20 unit change in total score.

sample size = 10

The REG Procedure
Model: MODEL1
Dependent Variable: Total_score Total_score

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3.27981	0.65596	Infty	<.0001
Error	4	0	0		
Corrected Total	9	3.27981			

Root MSE	0	R-Square	1.0000
Dependent Mean	3.63074	Adj R-Sq	1.0000
Coeff Var	0		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.43496E-14	0	Infty	<.0001
Human_Rights	Human_Rights	1	0.20000	0	Infty	<.0001
Governance	Governance	1	0.20000	0	Infty	<.0001
Health	Health	1	0.20000	0	Infty	<.0001
Economy	Economy	1	0.20000	0	Infty	<.0001
Environment	Environment	1	0.20000	0	Infty	<.0001

Regression Diagnosis - Mean

Limitations

- Sample size is less than 30.
- Model is perfectly fitted as R-square and Adjusted R-Square value is 1.
- Regression coefficients can not be uniquely determined.

Regression Diagnosis - Mean

Multiple Linear Regression

- Regression coefficients are still not uniquely determined.

sample size = 30

The REG Procedure
Model: MODEL1
Dependent Variable: Total_Score Total_Score

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	15.92076	3.18415	Infty	<.0001
Error	24	0	0		
Corrected Total	29	15.92076			

Root MSE	0	R-Square	1.0000
Dependent Mean	3.71266	Adj R-Sq	1.0000
Coeff Var	0		

Parameter Estimates

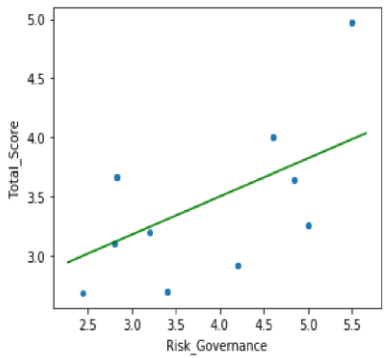
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	4.51063E-14	0	Infty	<.0001
Risk_Governance	Risk_Governance	1	0.20000	0	Infty	<.0001
Risk_Economy	Risk_Economy	1	0.20000	0	Infty	<.0001
Risk_Environment	Risk_Environment	1	0.20000	0	Infty	<.0001
Risk_Health	Risk_Health	1	0.20000	0	Infty	<.0001
Risk_HumanRight	Risk_HumanRight	1	0.20000	0	Infty	<.0001

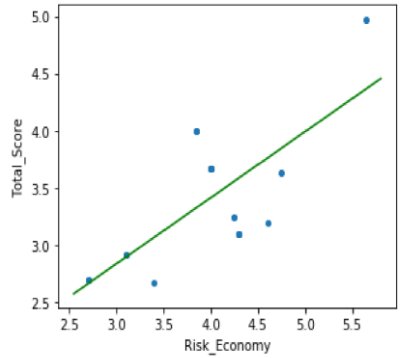
Regression Diagnosis - Median

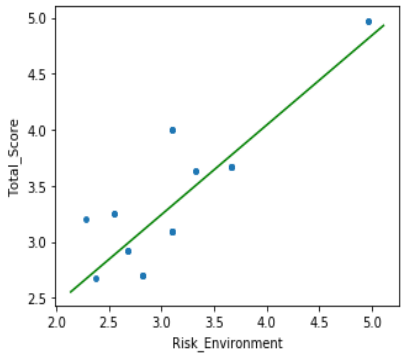
Test of Linear Regression Assumptions

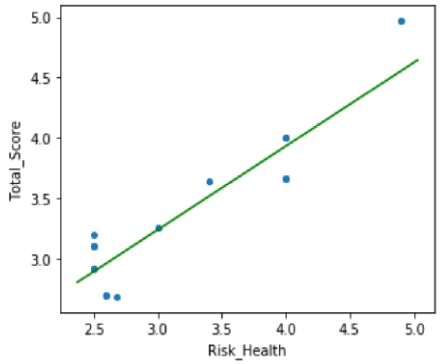
- Scatter plot of dependent variable (Total Score) against each of the independent variable shows there might be one outlier.

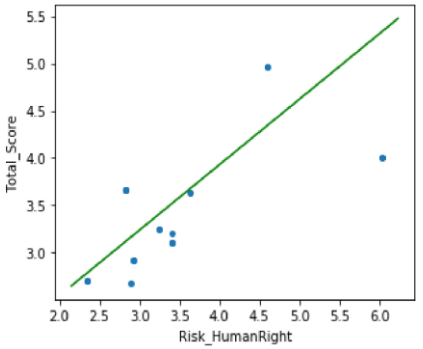
- Examination of studentized residual (RES), Leverage (LEV) for observations with large regressors, Influence statistics like DFFITS and Cook's D Statistics shows there is no outlier. So, we can conclude we do not have any outliers.











Regression Diagnosis - Median

Test of Linear Regression Assumptions

- The variance inflation factor and tolerance shows there is no multicollinearity.
- As a result, the regression coefficients can be uniquely determined.
- Test of normality (Histogram, QQ plot) shows that the residual are normally distributed.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	0.19507	0.06891	2.83	0.0092	.	0
Risk_Governance	Risk_Governance	1	0.06488	0.01374	4.72	<.0001	0.68105	1.46833
Risk_Economy	Risk_Economy	1	0.17637	0.02044	8.63	<.0001	0.48773	2.05032
Risk_Environment	Risk_Environment	1	0.21973	0.03750	5.86	<.0001	0.21780	4.59140
Risk_Health	Risk_Health	1	0.35636	0.02866	12.43	<.0001	0.23273	4.29683
Risk_HumanRight	Risk_HumanRight	1	0.12012	0.01487	8.08	<.0001	0.50536	1.97879

Regression Diagnosis - Median

Test of Linear Regression Assumptions

- White Test: We accept the null that there is homoscedasticity.
- Homoscedasticity: Residual has constant variance.

The REG Procedure
Model: MODEL1
Dependent Variable: Total_Score Total_Score

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
10	16.62	0.0832

Regression Diagnosis - Median

Multiple Linear Regression

- Regression analysis shows that all the indicators are significant.
- 1 unit change in Risk to Health will change 0.35 unit change in the Total Score.
- 1 unit change in Risk to Environment will change 0.21 unit change in Total Score.
- 1 unit change in Risk to Human Right will change 0.12 unit in Total Score.

30 obs

Root MSE	0.05915	R-Square	0.9919
Dependent Mean	3.39167	Adj R-Sq	0.9902
Coeff Var	1.74391		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.19507	0.06891	2.83	0.0092
Risk_Governance	Risk_Governance	1	0.06488	0.01374	4.72	<.0001
Risk_Economy	Risk_Economy	1	0.17637	0.02044	8.63	<.0001
Risk_Environment	Risk_Environment	1	0.21973	0.03750	5.86	<.0001
Risk_Health	Risk_Health	1	0.35636	0.02866	12.43	<.0001
Risk_HumanRight	Risk_HumanRight	1	0.12012	0.01487	8.08	<.0001

Classification and Clustering Analysis

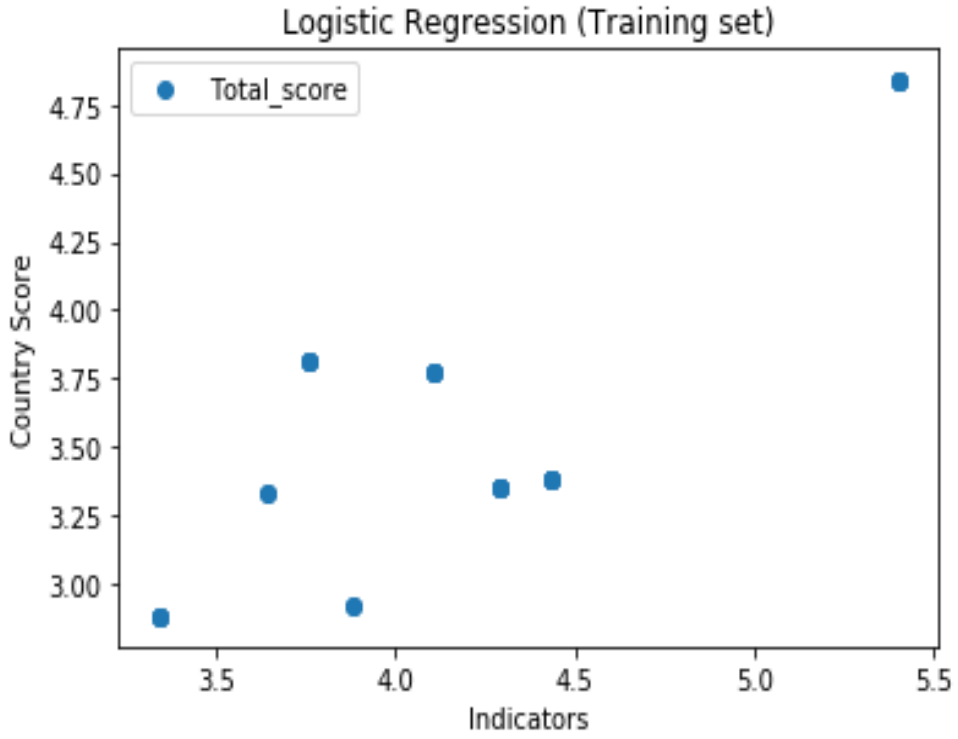
Logistic Regression (Classification)

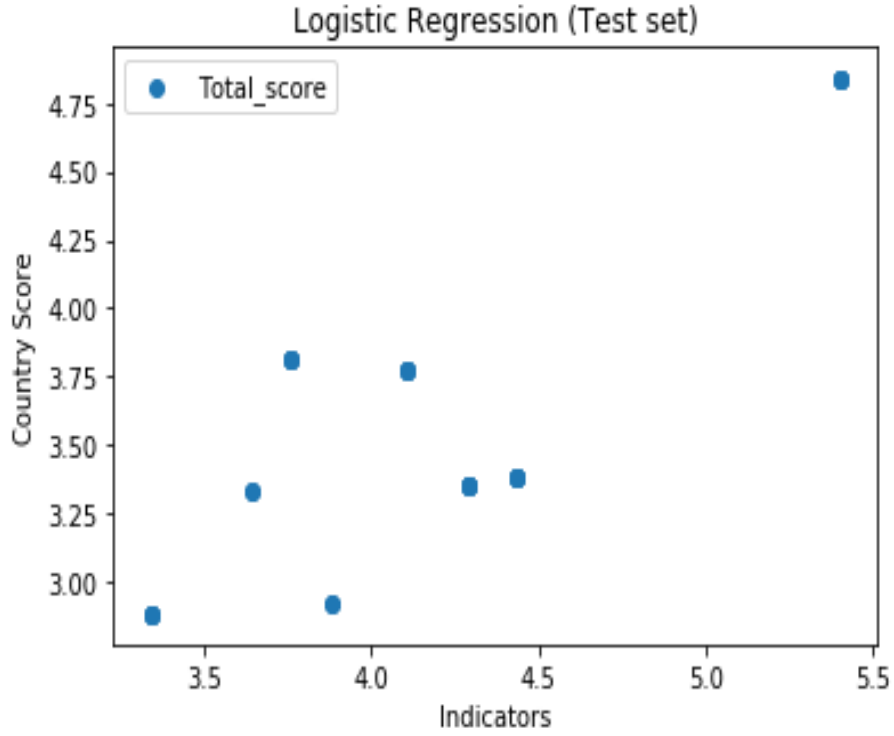
- In this data set the indices or the factors are called the independent variable and the country score is called the dependent variable.
- Using this split, we use 80% of our data to train the regression variable and apply this regression variable to predict the remaining 20% of the data set.
- Since our current data set has a very limited number of observations, the accuracy score for this kind of regression was achieved to be 100%.
- When expanding this model to a larger datasets, there can be reduction in the accuracy as expected by the real-world models.

Classification and Clustering Analysis

Accuracy of Logistic Regression

The accuracy was observed to be 100%





Classification and Clustering Analysis

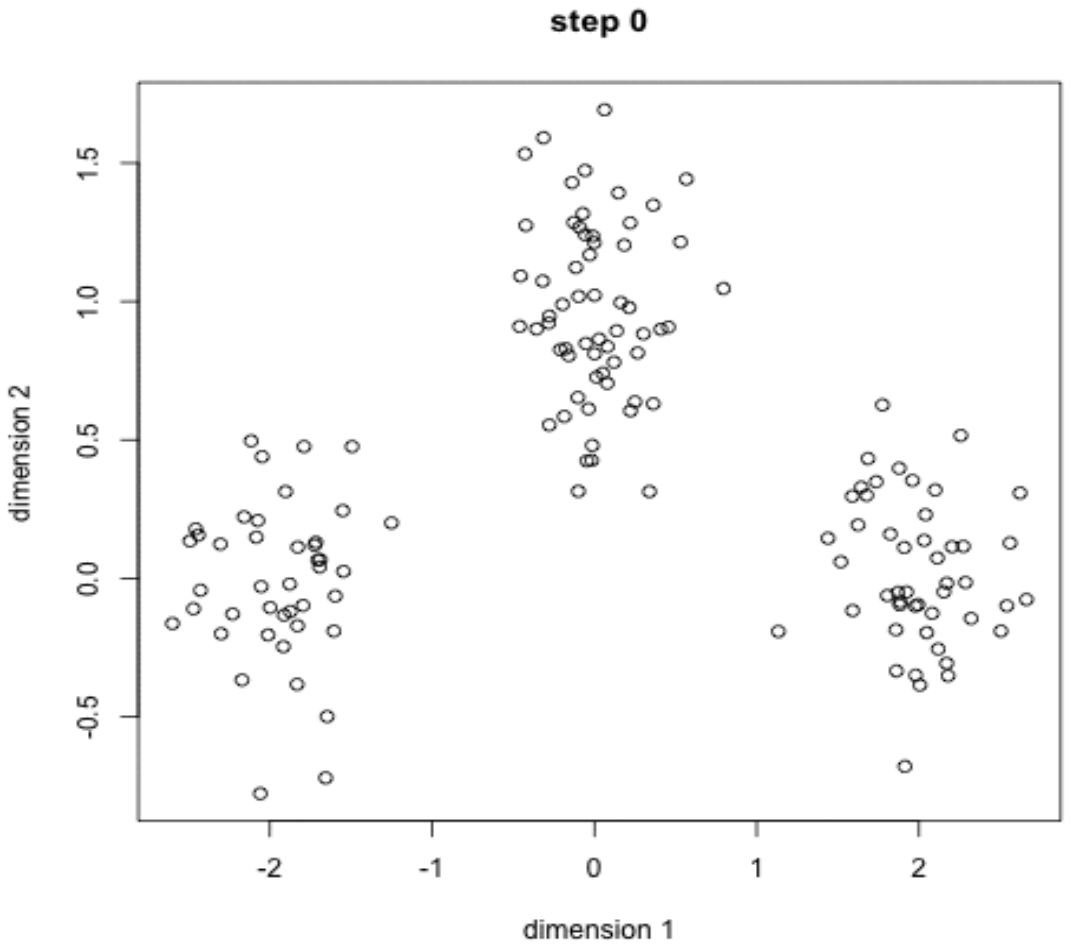
Clustering

- Clustering in data analytics means grouping of data points.
- Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group.
- We have performed two types of clustering techniques:
  - KNN Clustering
  - Hierarchical Clustering

Classification and Clustering Analysis

K Nearest Neighbor

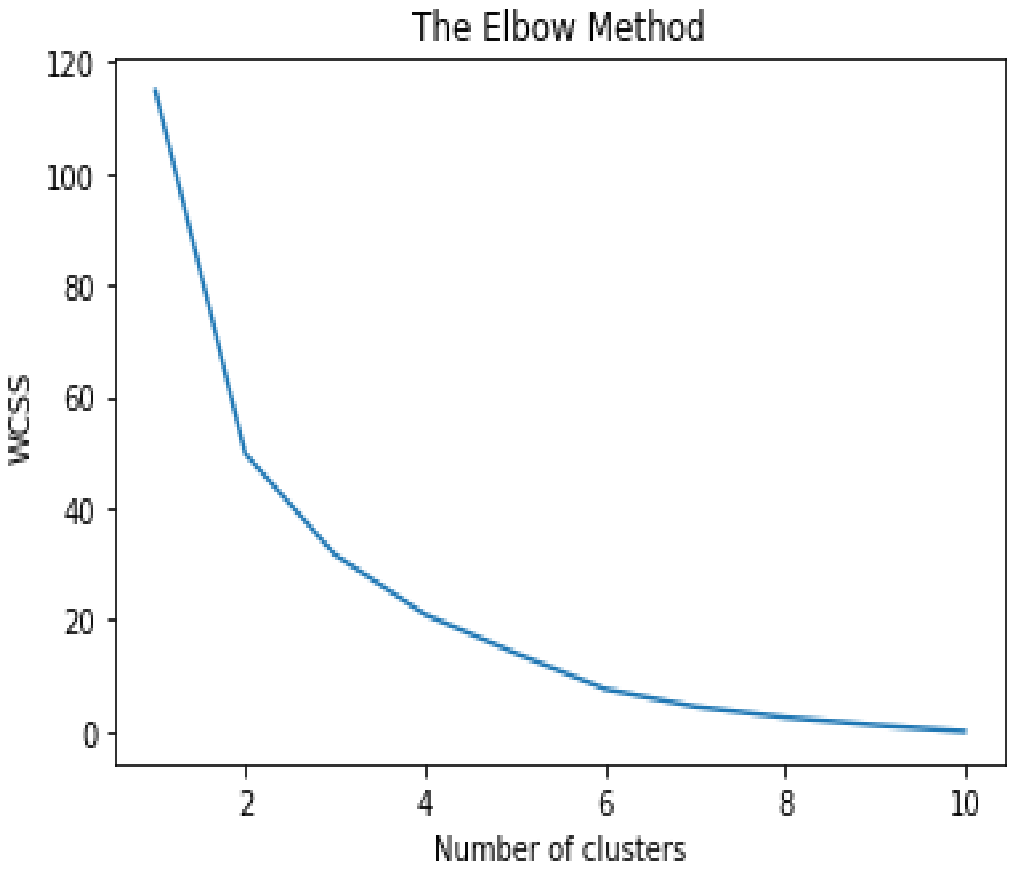
- Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group.
- Once the neighbors are discovered, the summary prediction can be made by returning the most common outcome or taking the average. As such, KNN can be used for classification or regression problems.



Classification and Clustering Analysis

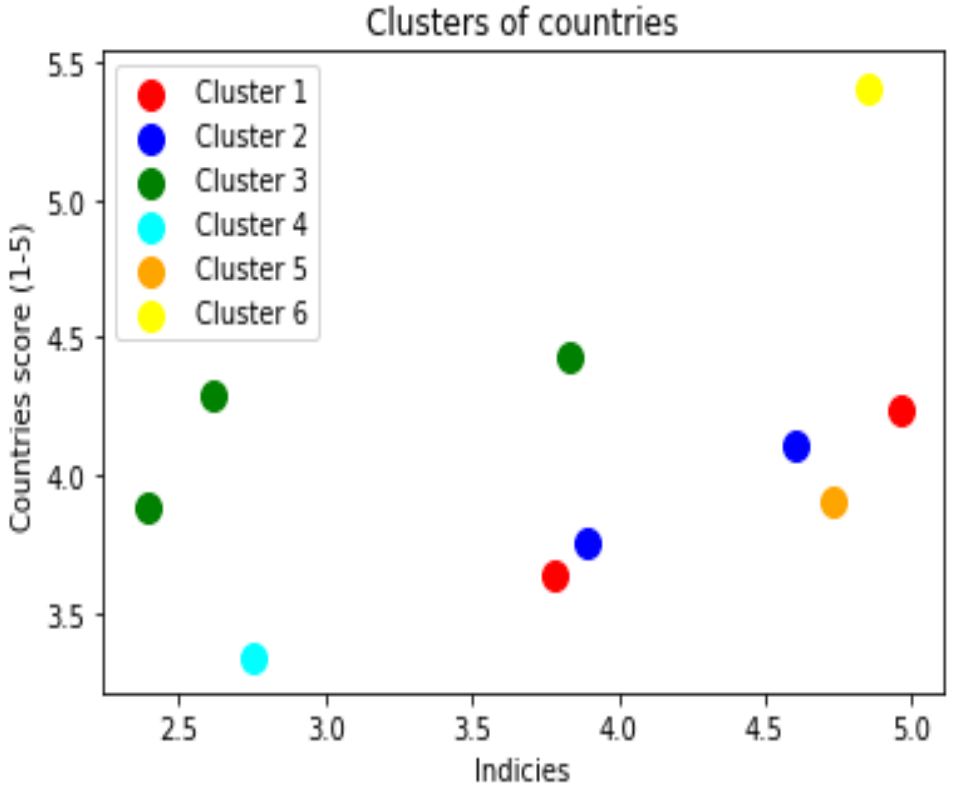
Elbow Method

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.



Classification and Clustering Analysis

Analysis with 6 Clusters



[0 1 2 3 5 2 4 2 1 0]

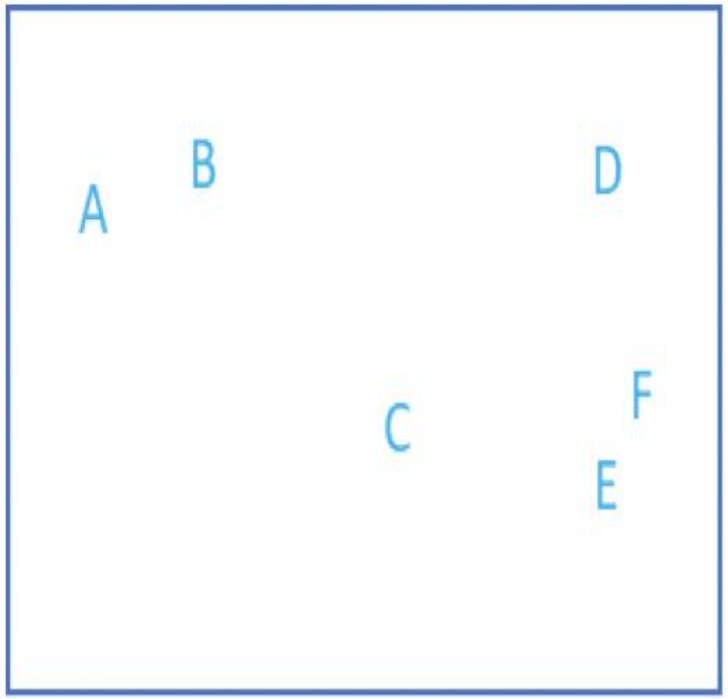
This was the pattern observed by the 10 countries that we have:

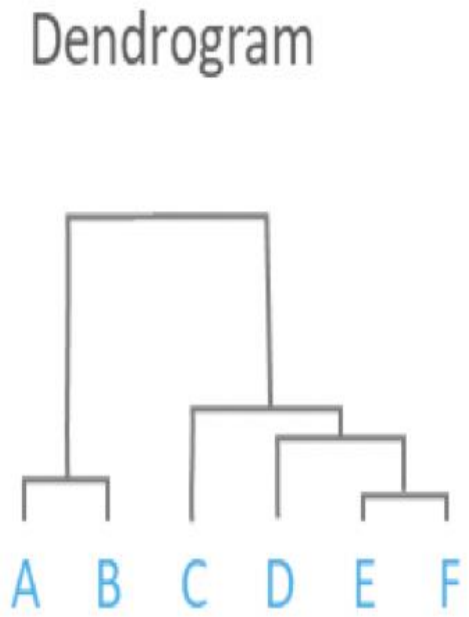
- Cluster1- Tanzania and Zambia
- Cluster2 - Brazil and Columbia
- Cluster 3- Myanmar, Madagascar and Peru
- Cluster 4 - Afghanistan
- Cluster 5 - South Africa
- Cluster 6 - Botswana

Classification and Clustering Analysis

Hierarchical Clustering

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

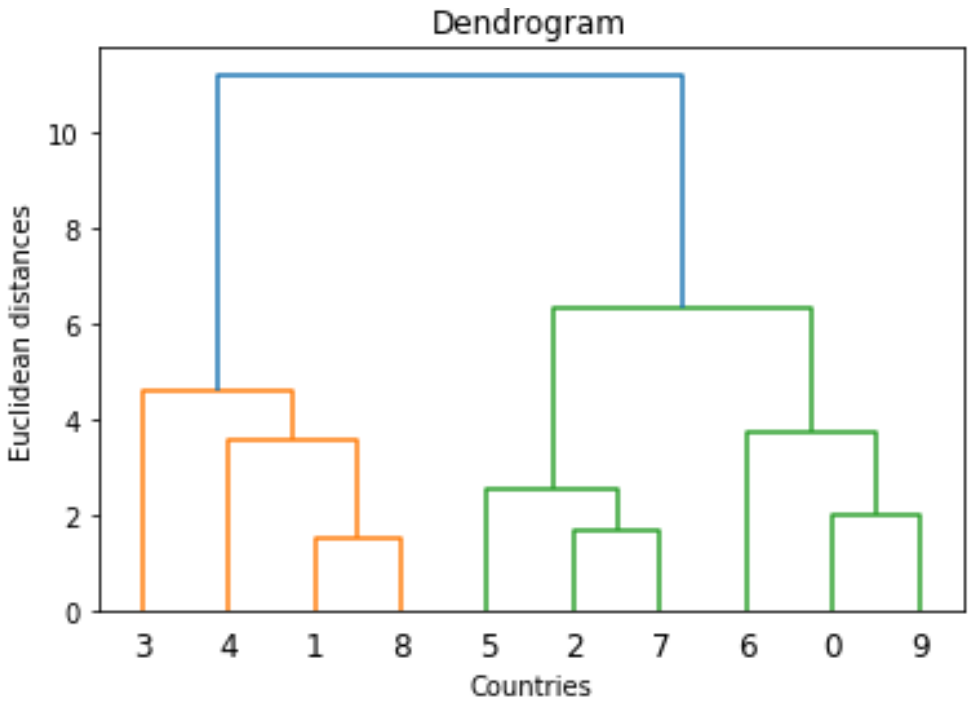




Classification and Clustering Analysis

Hierarchical Data for the Data

This method determines the no of clusters that can be made to differentiate the data easily. As seen in the picture the optimal number of clusters for this data set is determined using dendrogram.



Conclusion

- We have weighted median as more appropriate approach for this analysis when compared to the mean.
- With the help of the classification and the clustering models, we can predict the overall JDI score of the countries which has insufficient data.

THANK YOU

ANY QUESTIONS?