**< World Indicators Project Overview >**

**201533661이승수, 201533645배성재**

**<Contents>**

1. Objective Setting: Finding Countries similar to Korea from 1960s to 2010s
2. Data Curation: Country.csv, Indicators.csv
3. Data Inspection
4. Data Rreprocessing
5. K-Means
6. K-Medoids
7. DBSCAN
8. Data Analysis
   1. K-Means , K-Medoids, DBSCAN
9. Evaluation

Evaluation methods: k-fold cross-validation로 feature engineering이전/이후 시간,정확도 비교.

1. Deployment
2. **Objective Setting**

Our object similar to Korea from 1960s to 2010s. And we search five section (Economic/Financial, Infrastructure/Trade Section Agriculture/Environment/Health, Welfare/Labor/Population, Population)

1. **Data Curation**

<https://www.kaggle.com/worldbank/world-development-indicators>

1. **Data Inspection**

‘Indicators.csv’ has too many country and indicator code.

Unique ‘Indicator code’ is 1344, and unique ‘Country Code’ is 247.

Thus we classify the indicator code.

|  |  |
| --- | --- |
| Economy: | 'DT'(debt), 'BN', 'BM', 'BG', 'BX'(Economic Policy & Debt: Balance of payments), 'DC', |
| Financial | 'FP', 'FS', 'FM'(Monetary holdings (liabilities)), 'FR'(Interest rates), 'FI', 'FD', 'FB'(Bank,Assets), |
| Agriculture: | 'AG'(agriculture), |
| Environment: | 'EA'(Environment:Agriculture production), 'EG'(Environment: Energy production & use), 'EN'(Environment: Emissions), 'ER',  -: 'EP'(Infrastructure: Transportation), |
| Social Protection | 'SE’(Education), 'SH'(Health), 'SN'(Health), ’SP'(Health:Population), 'SG'(Health:Risk factors), 'SI'(Poverty), 'SL'(Social Protection & Labor), 'SM'(Migration), 'ST'(Travel,Trade,Tourism), |
| Trade: | 'TM', 'TX', 'TT', 'TG', |
| Infrastructure: | 'IT'(Infrastructure: Communications), 'IS'(Infrastructure: Transportation), 'IP'(Infrastructure: Technology), |
| Private | 'IC'(Private Sector & Trade: Business environment), 'IE'(Private infrastructure investment),'IQ'(Public Sector: Policy & institutions), |
| ETC | 'DC', 'MS', 'VC', 'CM', 'LP', 'GB', 'GC', |

1. **Data Preprocessing**

After data inspection, we make indicator code to 5 section like Economic/Financial section, Infrastructure/Trade Section Agriculture/Environment/Health section, Welfare/Labor/Population section, and Population section.

|  |  |
| --- | --- |
| **Section** | **Series Code** |
| **Economic/ Financial** | **DC.DAC.TOTL.CD** [Net bilateral aid flows from DAC donors, Total (current US$)]  **FM.AST.DOMS.CN** [Net domestic credit (current LCU)]  **NV.IND.MANF.ZS** [Manufacturing, value added (% of GDP)]  **BX.KLT.DINV.WD.GD.ZS** [Foreign direct investment, net inflows (% of GDP)] |
| **Infrastructure/ Trade** | **TX.VAL.FUEL.ZS.UN/TM.VAL.FUEL.ZS.UN**  TM.VAL.FUEL.ZS.UN [Fuel imports (% of merchandise imports)]  TX.VAL.FUEL.ZS.UN [Fuel exports (% of merchandise exports)]  **TX.VAL.MANF.ZS.UN/TM.VAL.MRCH.CD.WT**  TM.VAL.MRCH.CD.WT [Merchandise imports (current US$)]  TX.VAL.MANF.ZS.UN [Manufactures exports (% of merchandise exports)]  **IS.AIR.PSGR** [Air transport, passengers carried]  **IT.CEL.SETS.P2** [Mobile cellular subscriptions (per 100 people)]  **IT.NET.USER.ZS** [Individuals using the Internet (% of population)] |
| **Agriculture/ Environment** **/Health** | **AG.LND.AGRI.ZS** [Agricultural land (% of land area)]  **AG.PRD.LVSK.XD** [Livestock production index (2004-2006 = 100)]  **EN.ATM.CO2E.KT** [CO2 emissions (kt)]  **EN.POP.DNST** [Population density (people per sq. km of land area)] |
| **Welfare/ Labor** **/Population** | **SE.SEC.CUAT.LO.ZS [**Educational attainment, at least completed lower secondary, population 25+, total (%) (cumulative)]  **SH.DTH.IMRT [**Number of infant deaths]  **SL.UEM.TOTL.NE.ZS** [Unemployment, total (% of total labor force) (national estimate)]  **SH.IMM.IDPT** [Immunization, DPT (% of children ages 12-23 months)]  **SL.AGR.EMPL.ZS** [Employment in agriculture (% of total employment) (modeled ILO estimate)]  **SL.GDP.PCAP.EM.KD** [GDP per person employed (constant 2011 PPP $)]  **SH.XPD.CHEX.PC.CD** [Current health expenditure per capita (current US$)] |
| **Population** | **SP.POP.TOTL** [Population, total]  **SP.POP.TOTL.FE.IN/SP.POP.TOTL.MA.IN (female ratio)**  SP.POP.TOTL.FE.IN [Population, female]  SP.POP.TOTL.MA.IN [Population, male]  **SP.POP.1564.TO/(SP.POP.0014.TO+SP.POP.65UP.TO)**  SP.POP.0014.TO [Population ages 0-14, total]  SP.POP.1564.TO [Population ages 15-64, total]  SP.POP.65UP.TO [Population ages 65 and above, total]  **SP.URB.TOTL.IN.ZS/SP.RUR.TOTL.ZS**  SP.RUR.TOTL.ZS [Rural population (% of total population)]  SP.URB.TOTL.IN.ZS [Urban population (% of total population)] |

**Feature Engineering** about 5 section, we did feature discretization and feature creation.

**Feature discretization** step, we binned ‘year’ columns to ‘decade’ by binning them into 1960s, 1970s, and so on..

At **Feature creation**, we merge some Series\_codes into single meaning by preprocessing function because they can make meaningful features when combined. Preprocessing function has 3 type.

**Preprocessing type 1** use to one indicator code. It make new table with just one indicator code.

**Preprocessing type 2** use to merge the two indicators. It used to check the female ratio of total population.

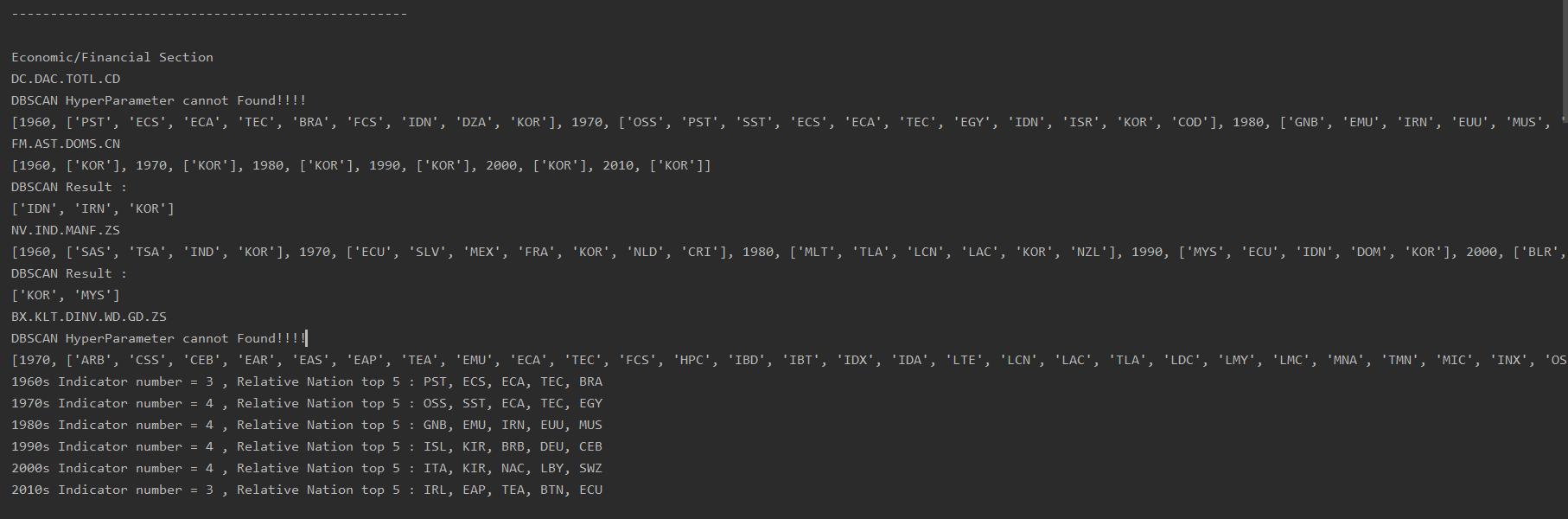
**Preprocessing type 3** used to merge the three indicators. It used to check ratio of labor number of total populations.

1. **Data Analysis**

We did ensemble learning by using K-means, K-medoids and DBSCAN clustering together. By comparing an algorithms, we get five top nations with frequency at clustered in each sections.

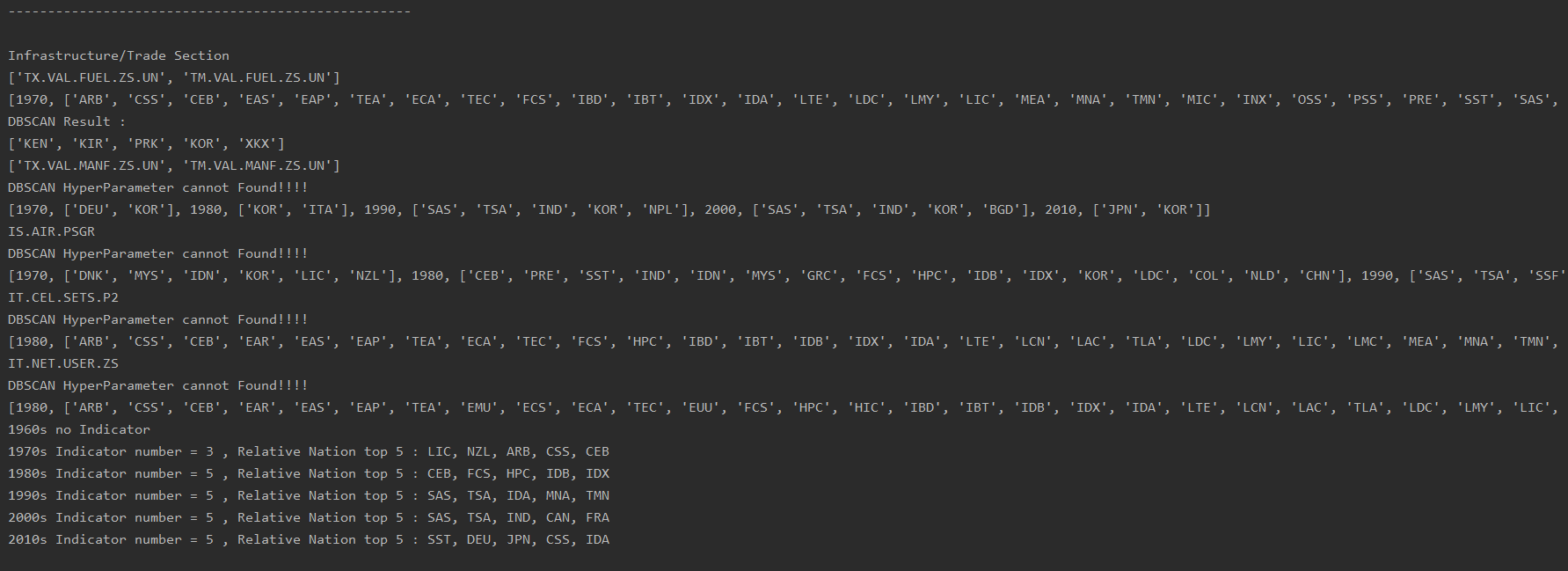
The result of clustering

**<Economic/ Financial section>**



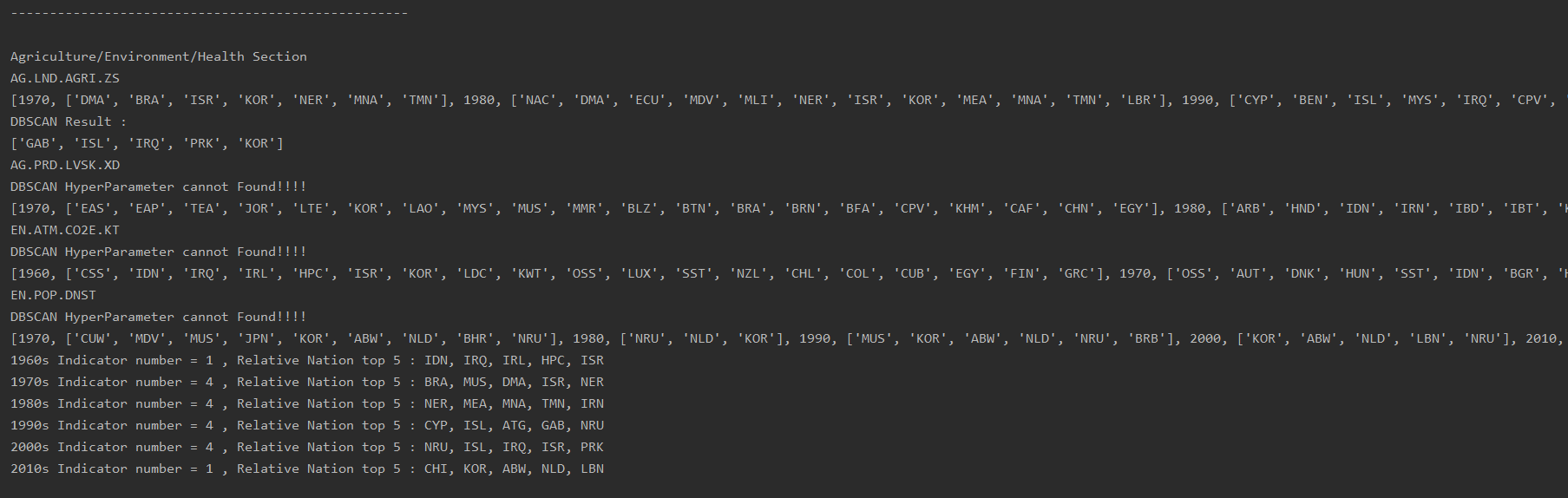
|  |  |
| --- | --- |
| **Decade** | **Nations** |
| **1960s** | **Pacific island small states, Brazil, Central Asia, Central Asia (excluding high income)** |
| **1970s** | **Other small states, Small states, Central Asia (excluding high income), Egypt, Arab Rep.** |
| **1980s** | **Guinea-Bissau, Iran, Islamic Rep, Mauritius** |
| **1990s** | **Iceland, Kiribati, Barbados, Germany,** |
| **2000s** | **Italy, Kiribati, North America, Eswatini,** |
| **2010s** | **Ireland, East Asia & Pacific (IDA & IBRD countries)** |

**<Infrastructure/ Trade Section>**



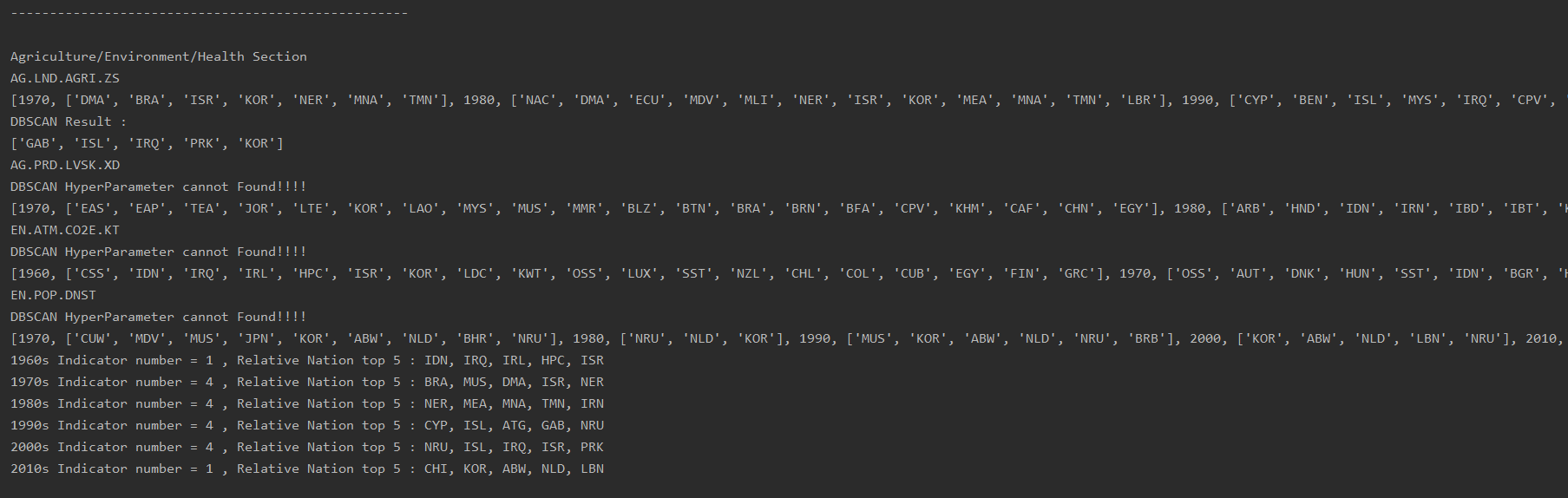
|  |  |
| --- | --- |
| **Decade** | **Nations** |
| **1970s** | **New Zealand , Arab World, Caribbean small states** |
| **1980s** | **Central Europe and the Baltics, Heavily indebted poor countries (HIPC),** |
| **1990s** | **South Asia, Middle East & North Africa (excluding high income),** |
| **2000s** | **South Asia, India, Canada, France** |
| **2010s** | **Small states, Germany, Japan** |

**<Agriculture/ Environment** **/Health Section>**



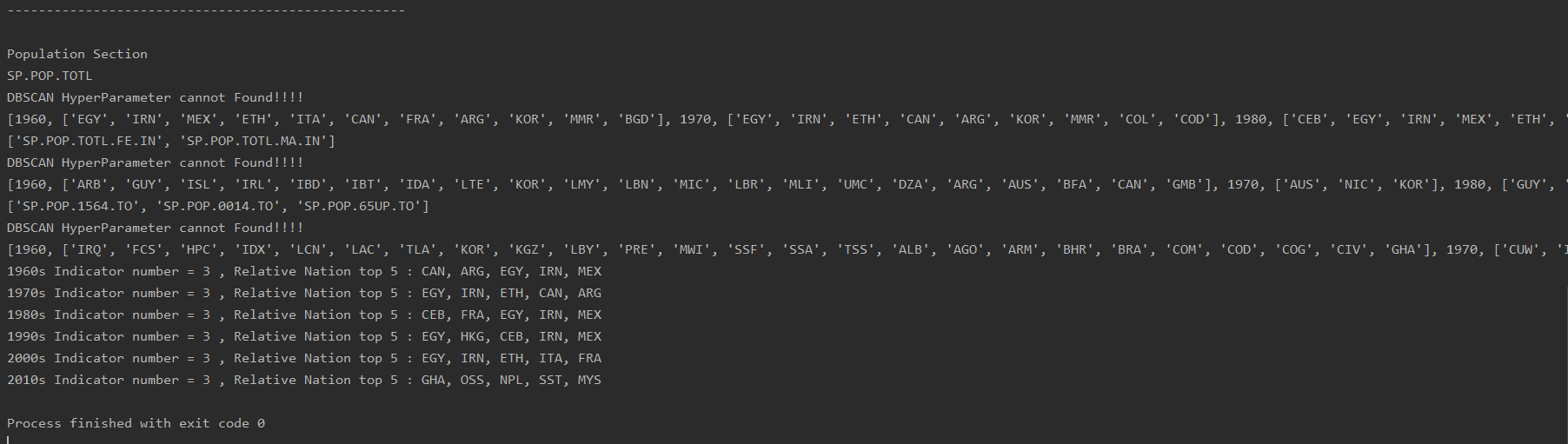
|  |  |
| --- | --- |
| **Decade** | **Nations** |
| **1960s** | **Indonesia, Iraq, Ireland** |
| **1970s** | **Brazil, Mauritius, Dominica** |
| **1980s** | **Niger, Middle East & North Africa, Middle East & North Africa (excluding high income)** |
| **1990s** | **Iceland, Antigua and Barbuda, Gabon** |
| **2000s** | **Nauru, Iceland, Iraq, Israel** |
| **2010s** | **Channel Islands, Aruba, Netherlands** |

**<Welfare/ Labor** **/Population>**



|  |  |
| --- | --- |
| **Decade** | **Nations** |
| **1960s** | **Algeria, Morocco, Colombia** |
| **1970s** | **Guinea, Ghana, Madagascar, Malawi** |
| **1980s** | **Liberia, Ecuador, World, Central Europe and the Baltics** |
| **1990s** | **East Asia & Pacific, East Asia & Pacific (excluding high income)** |
| **2000s** | **Libya, Malaysia, Hungary, Czech Republic** |
| **2010s** | **Czech Republic, Hungary, Japan** |

**<Population Section>**



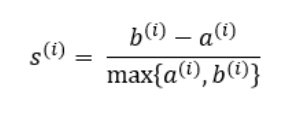
|  |  |
| --- | --- |
| **Decade** | **Nations** |
| **1960s** | **Canada, Argentina, Egypt, Arab Rep** |
| **1970s** | **Egypt, Arab Rep., Iran, Islamic Rep., Ethiopia, Canada** |
| **1980s** | **Central Europe and the Baltics, France, Egypt, Arab Rep, Iran, Islamic Rep.** |
| **1990s** | **Egypt, Arab Rep., Hong Kong SAR, China, Central Europe and the Baltics, Iran, Islamic Rep.** |
| **2000s** | **Egypt, Arab Rep., Iran, Islamic Rep., Ethiopia, Italy, France** |
| **2010s** | **Ghana, Nepal, Small states, Malaysia** |

**Average Cluster of 5 sections above considering total distribution is as below.**

|  |  |
| --- | --- |
| **Decade** | **Nations** |
| **1960s** | **Arab Rep** |
| **1970s** | **Egypt,** |
| **1980s** | **Iran, Islamic Rep.** |
| **1990s** | **Iceland** |
| **2000s** | **Italy, France** |
| **2010s** | **Japan,** **Small states** |

1. **Evaluation**

We use Shilhouette function to evaluate the DBSCAN. The silhouette score the quality of clustering.



For index(i) data, S(i) is Shilhouette. A(i) is data cohesion in cluster. B(i) is separation between clusters.

