ANALYSIS OF GLOBAL DEVELOPMENT DATA

DANA 4840-001

Lien Pham
Mohamed Ghayaas
Ayushi Singh
Mary Ann Villamor

CONTENTS

I. INTRODUCTION

DATASET

2010 World Bank Economic Data

PURPOSE

Cluster the countries and detect groups' characteristics to support World Bank to assess the countries for decision making

DATA OVERVIEW

- 214 Countries
- 33 variables (28 numerical data)
- Comprises of economic, population, education and health factors

TARGET AUDIENCE

The **World Bank** Group works in every major area of development. They provide a wide array of financial products and technical assistance to help countries to eradicate poverty and increase life's quality. Their specific goals are:

1. Eradicate poverty and hunger
2. Achieve universal primary education
3. Promote gender equality and empower women
4. Reduce child mortality
5. Improve maternal health
6. Combat HIV/AIDS, malaria, and other diseases
7. Ensure environmental sustainability

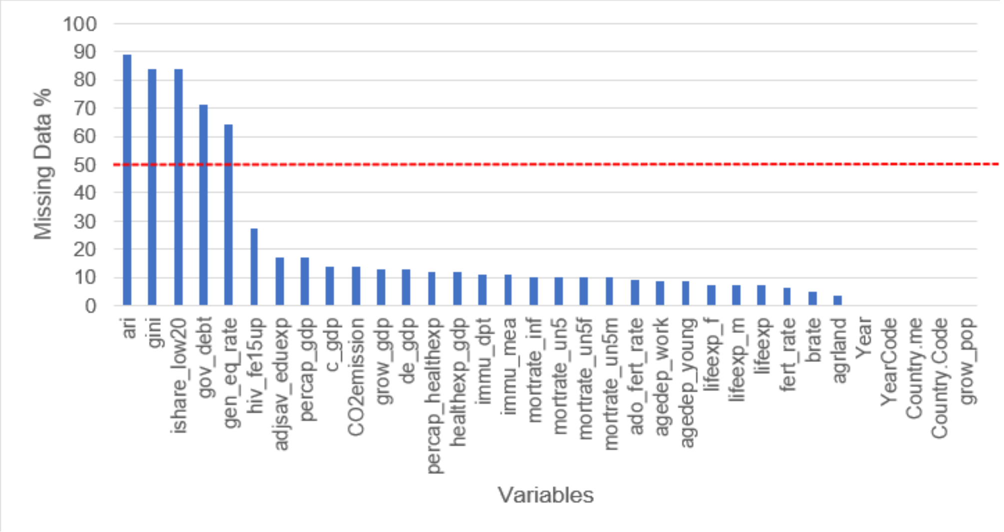
Source: World Bank development indicator 2010

2. DATA PRE-PROCESSING

01 MISSING DATA

1,383 (20%) missing data

- 5 variables with > 50% missing data



- 34 countries/rows with > 30% missing data

Range	Count	Rate %
>=80%	5	2.34%
>=60% < 80%	6	2.80%
>=50% <60%	11	5.14%
>=30% <50%	12	5.61%
>=20% <30%	11	5.14%
> 20%	169	78.97%
Total	214	100.00%

02 RELEVANT DATA

Redundancies

- Categorical variables not needed for analysis:
  - Year
  - Year Code
  - Country name
- Highly correlated data

lifeexp_f	**Life expectancy at birth, female (years)**
lifeexp_m	Life expectancy at birth, male (years)
Lifeexp *	Life expectancy at birth, total (years)

mortrate_inf *	**Mortality rate, infant (per 1,000 live births)**
mortrate_un5 *	**Mortality rate, under-5 (per 1,000 live births)**
mortrate_un5f	Mortality rate, under-5, female (per 1,000)
mortrate_un5m	Mortality rate, under-5, male (per 1,000)

** removed*

- **Data added – CO2 emission!**

  Goal #7: Ensure environmental sustainability (World Development Indicator 2010, World Bank)

2. DATA PRE-PROCESSING

03 IMPUTATION

71 missing data (2%) after cleaning

- Our dataset has multiple variables with high correlation and multicollinearity. Missing data was imputed using MissForest

- MissForest is robust to noisy data and multicollinearity, since random-forests have built-in feature selection (evaluating entropy and information gain). KNN-Impute yields poor predictions when datasets have weak predictors or heavy correlation between features.

- No significant changes for the SD, Mean and Max **before and after** imputation.

variable	Before impute				After impute			
	Count	sd	min	max	Count	sd	min	max
adjsav_eduexp	167	1.89	0.84	12.93	180	1.84	0.84	12.93
agrland	179	22.14	0.5	88.4	180	22.08	0.5	88.4
c_gdp	174	1.18	1.11	1.36	180	1.16	1.11	1.36
grow_gdp	176	3.87	-9.53	16.73	180	3.83	-9.53	16.73
percap_gdp	172	13493.07	335.68	70239.31	180	13274.73	335.68	70239.31
percap_healthexp	179	1683.66	12.7	8232.88	180	1680.26	12.71	8232.88
healthexp_gdp	179	2.27	0.24	12.46	180	2.270844	0.24	12.46
de_gdp	176	7.38	-4.2	45.94	180	7.30906	-4.2	45.94
immu_dpt	179	12.48	33	99	180	12.49608	33	99
hiv_fe15up	153	16.13	8.9	68	180	15.22359	8.9	68

3. APPROPRIATE K AND CLUSTERING METHODS

Internal criterion:

A good clustering will produce high quality clusters in which:

- The intra-class (that is, intra-cluster) similarity is high
- The inter-class similarity is low

The measured quality of a clustering depends on both the document representation and the similarity measure used

Internal criterion is used when we don’t have a ground of truth or expert knowledge.

- Silhouette coefficient
- CH score

Agglomerative coefficient:
measures the amount of clustering structure of the dataset

- If observations quickly agglomerate into distinct clusters that later agglomerate into a single cluster at much greater dissimilarities, the coefficient will approach 1
- In contrast, no clustering for the dataset will have coefficient approaching zero

3. APPROPRIATE K AND CLUSTERING METHODS

Note:
The same methods were peformed on kmeans, kmedoid and hierarchal wardD2 and wardD2 has highest silhouette and CH scores

Silhouette score

	sw_single	sw_complete	sw_average	sw_wardD2
[k=2]	0.24431939	0.3696676	0.3635434	0.3694468
[k=3]	0.16666165	0.2902638	0.3200097	0.2421877
[k=4]	0.13790036	0.2972885	0.2967349	0.2148586
[k=5]	0.07008383	0.2515810	0.2507863	0.1850567
[k=6]	0.05565885	0.1488498	0.2156399	0.1846749
[k=7]	-0.01897703	0.1377229	0.2076657	0.1763091
[k=8]	-0.03041767	0.1135680	0.1858980	0.1831650
[k=9]	-0.15339242	0.1065394	0.1740821	0.1528211
[k=10]	-0.16777476	0.1063986	0.1408716	0.1499510

CH score

	ch_single	ch_complete	ch_average	ch_wardD2
[k=2]	2.347684	161.43729	158.90334	164.36272
[k=3]	5.208947	141.65347	85.03813	158.67000
[k=4]	4.881848	104.08367	101.29612	131.65109
[k=5]	3.729960	99.34525	98.97265	132.15571
[k=6]	4.102568	102.85451	88.74460	118.24940
[k=7]	3.454631	89.81414	74.51522	102.26244
[k=8]	3.088748	87.04003	64.16104	91.73435
[k=9]	2.705327	93.48398	58.24927	92.97035
[k=10]	2.426490	88.73290	52.33390	90.97680

Ward D2 has highest silhouette and CH score, with optimal k = 2

Agglomerative coefficient

```
coef.hclust(hc_single)    #0.4988596
coef.hclust(hc_complete)  #0.8496924
coef.hclust(hc_average)   #0.6895446
coef.hclust(hc_wardD2)    #0.9526863
```

- Agglomerative ward clustering seems to give a better structure, in comparison to the other clustering technique

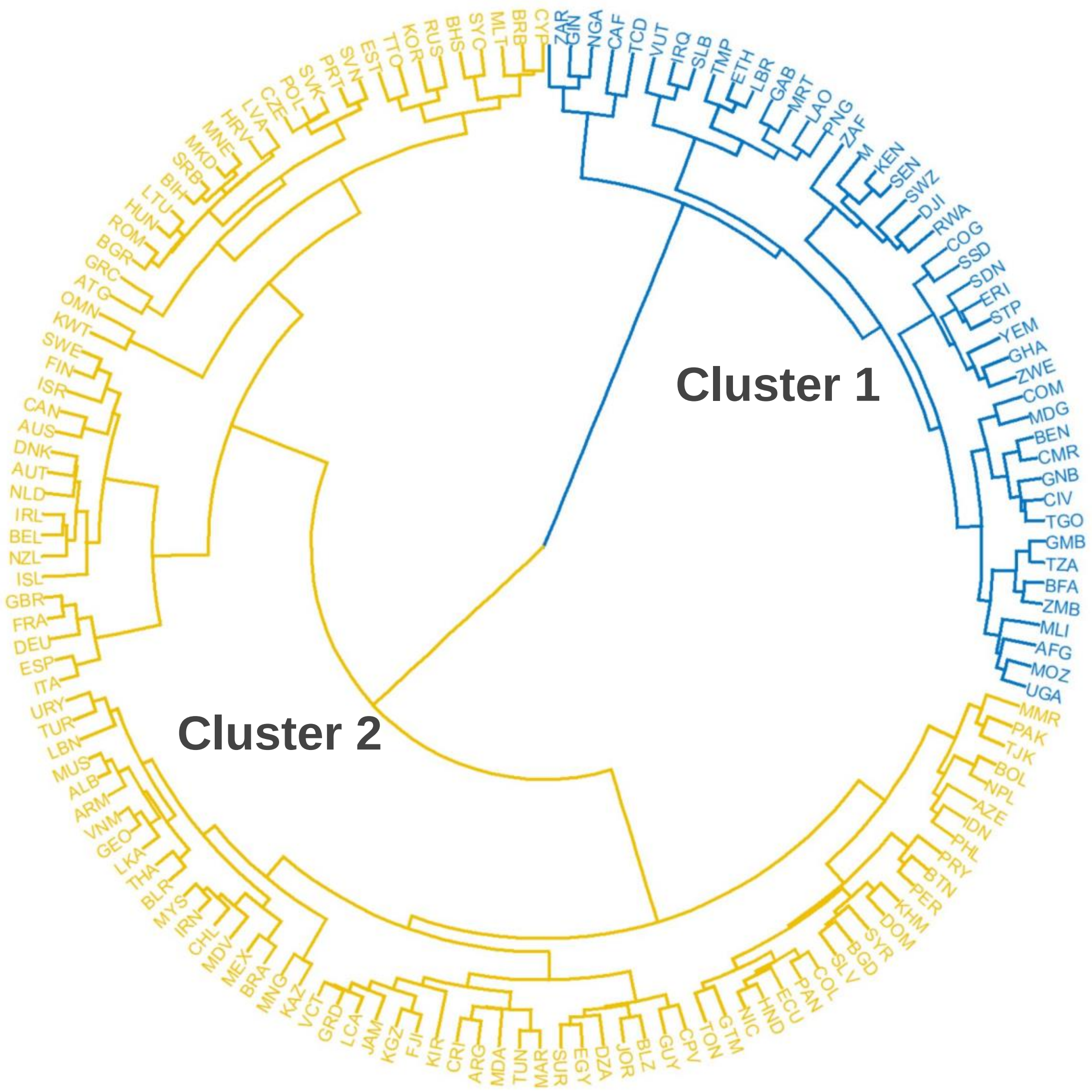
4. CLUSTER INTERPRETATION

Cluster 1

- Number of countries – 45
- Example - Afghanistan, Ethiopia, South Africa, South Sudan, Central African Republic, Nigeria, Rwanda, Yemen, Uganda
- Label : Low Income countries

Cluster 2

- Number of countries – 106
- Example - Australia, Oman, New Zealand, Denmark, Greece, Israel
- Label: Middle (upper and lower) Income and High income



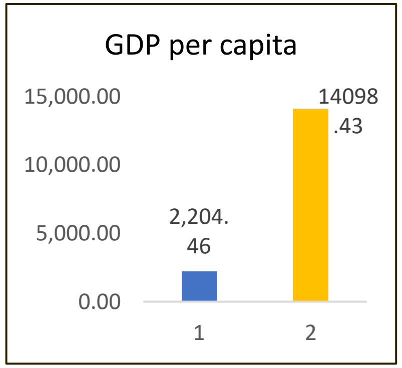
4. COUNTRY'S ECONOMY

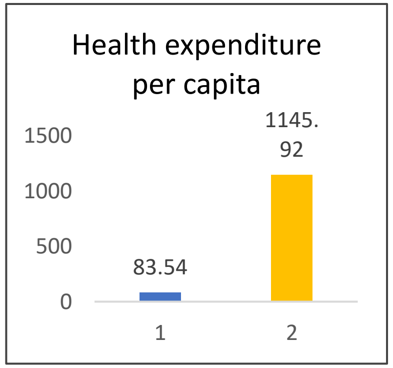
Cluster 1 (Low Income Countries)

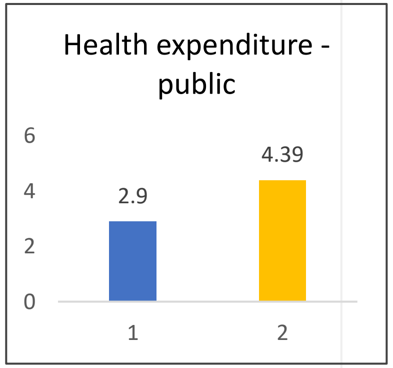
- GDP per capita – Low
- Health Expenditure Per Capita – Low
- Heath Expenditure Public – Low
- GDP Growth - High
- Inflation – High
- CO_2 Emission - Low

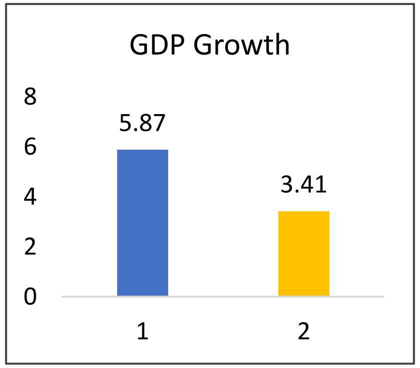
Cluster 2 (Middle & High Income Countries)

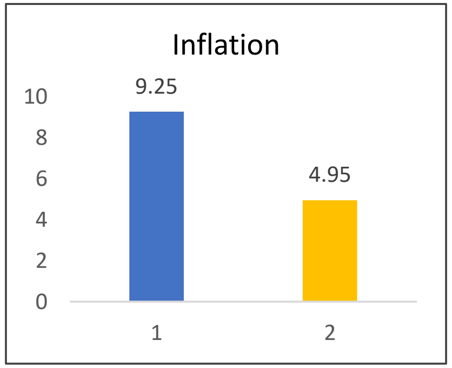
- GDP per capita – High
- Health Expenditure Per Capita – High
- Heath Expenditure Public – High
- GDP Growth - Low
- Inflation – Low
- CO_2 Emission - High

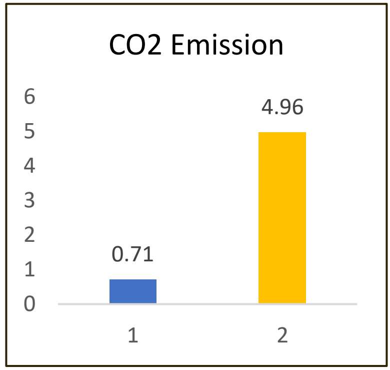












4. POPULATION HEALTH

Cluster 1 (Low income)

- Life Expectancy (years)
  - Male – 57.1
  - Female – 59.7
- Mortality Rate / Per 1000 births
  - Male – 92.15
  - Female – 81.03
- Annual Population Growth – 2.58 %

Cluster 2 (Middle and high income)

- Life Expectancy (years)
  - Male – 71.63
  - Female – 77.38
- Mortality Rate / Per 1000 births
  - Male – 19.47
  - Female – 15.98
- Annual Population Growth – 0.94 %

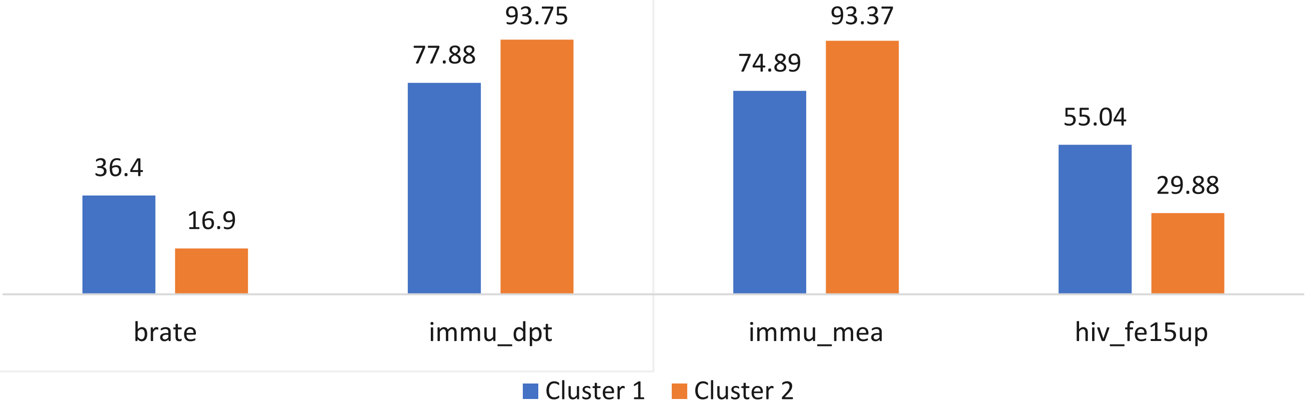
4. POPULATION HEALTH

Cluster 1 (Low income)

- Birth Rate – High
- Lower Immunization against DPT and Measles in children
- Higher percentage of women 15+ of age living with HIV +

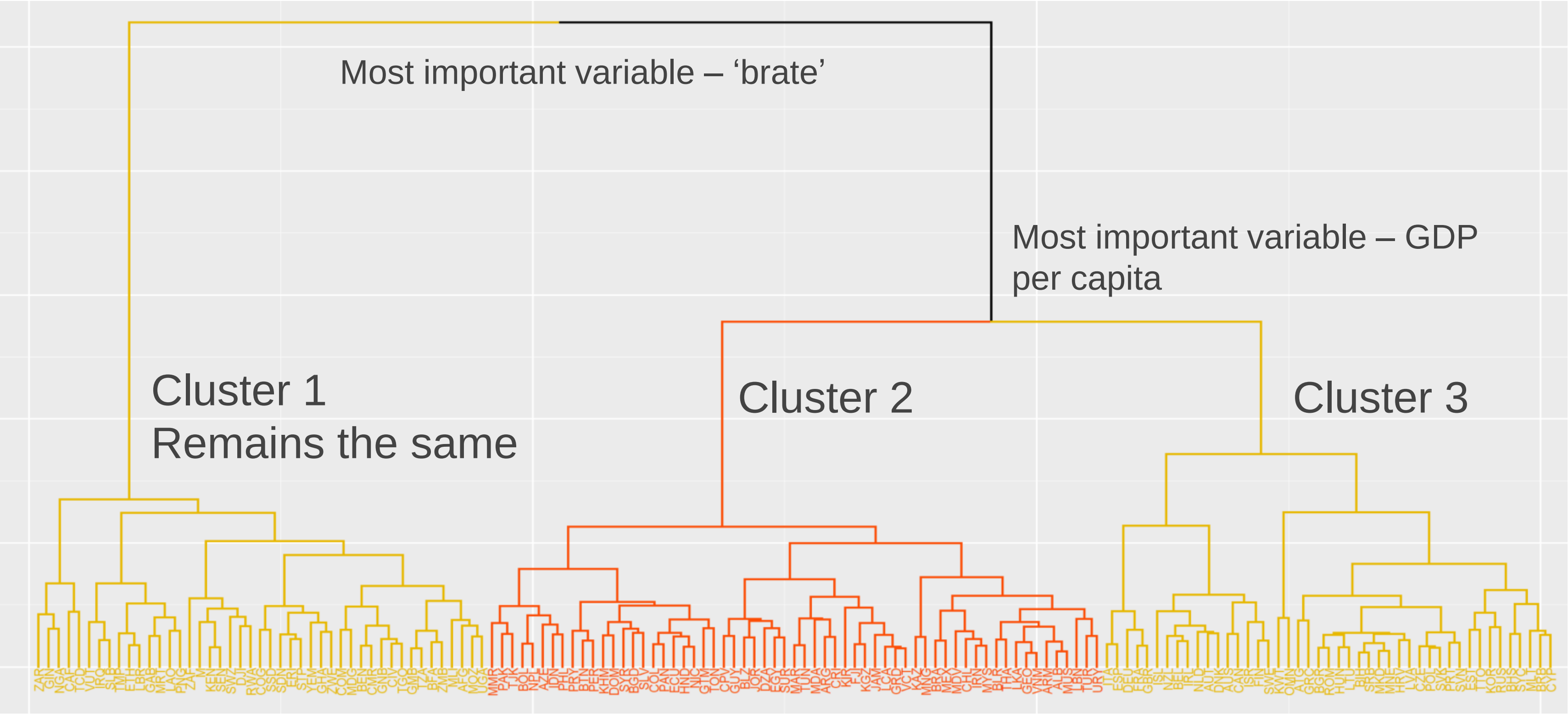
Cluster 2 (Middle and high income)

- Birth Rate – Low
- Higher Immunization against DPT and Measles in children
- Lower percentage of women 15+ of age living with HIV +



4. MIDDLE INCOME (2) VS HIGH INCOME COUNTRIES (3)

		cluster_k2	cluster_k3
Albania	ALB	2	2
Algeria	DZA	2	2
Antigua and Barbuda	ATG	2	3
Argenti	ARG	2	2
Armenia	ARM	2	2
Australia	AUS	2	3
Austria	AUT	2	3
Azerbaijan	AZE	2	2
Bahamas, The	BHS	2	3
Bangladesh	BGD	2	2
Barbados	BRB	2	3
Belarus	BLR	2	2
Belgium	BEL	2	3
Belize	BLZ	2	2
Bhutan	BTN	2	2
Bolivia	BOL	2	2
Bosnia and Herzegovi	BIH	2	3
Brazil	BRA	2	2
Bulgaria	BGR	2	3
Cabo Verde	CPV	2	2
Cambodia	KHM	2	2
Cada?	CAN	2	3
Chile	CHL	2	2
Colombia	COL	2	2
Costa Rica	CRI	2	2
Croatia	HRV	2	3
Cyprus	CYP	2	3



Low income (45)

- Afghanistan
- Ethiopia
- South Sudan
- Central African Republic
- Nigeria
- Rwanda

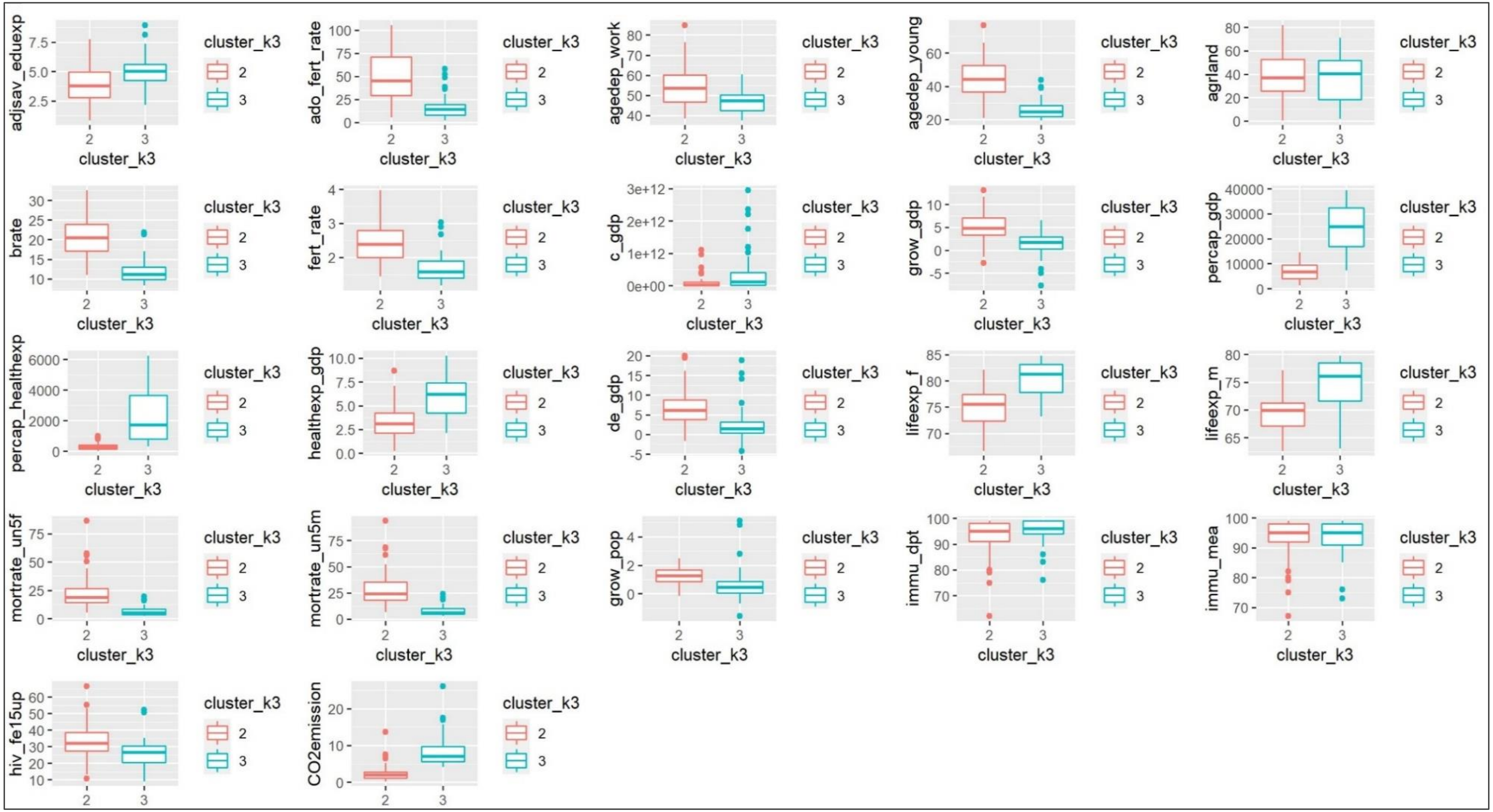
Middle income (61)

- Myanmar
- Philippines
- Thailand
- Turkey
- Malaysia
- Vietnam
- Peru

High income (45)

- New Zealand
- Denmark
- Greece
- Israel
- Germany
- Sweden
- United Kingdom

6. MIDDLE INCOME COUNTRIES VS HIGH INCOME COUNTRIES



5. USE CLUSTERS' LABELS FOR MODELLING

Encoding:

- Cluster 1 – Low Income countries – **1**
- Cluster 2 – Middle & High Income - **0**

Random Forest Accuracy – 90.32 %

```
Counter({0: 20, 1: 11})
```

col_0	0	1
cluster		
0	20	0
1	3	8

XG Boost Accuracy – 93.54 %

```
Counter({0: 20, 1: 11})
```

col_0	0	1
cluster		
0	20	0
1	2	9

5. CONCLUSIONS/RECOMMENDATIONS

From WB's perspective

- Deep dive into group 2 to analyze the countries that are middle income such as Philippine, Vietnam, Myanmar
- Efforts shall be undertaken by the world bank to **curb the high CO2 emission** from developed countries alongside the other goals.

THE WORLD BANK

From algorithm perspective

- Look for more variables such as criminal rate, clean water quality access rate, literacy rate to support World Bank's goals/decision making if there is a specific goal
- Perform analysis in cluster 2 to detect lower middle income countries
- Try the clustering on the most recent dataset (2021) to detect the changes in clustering, trends and patterns

QUESTIONS?