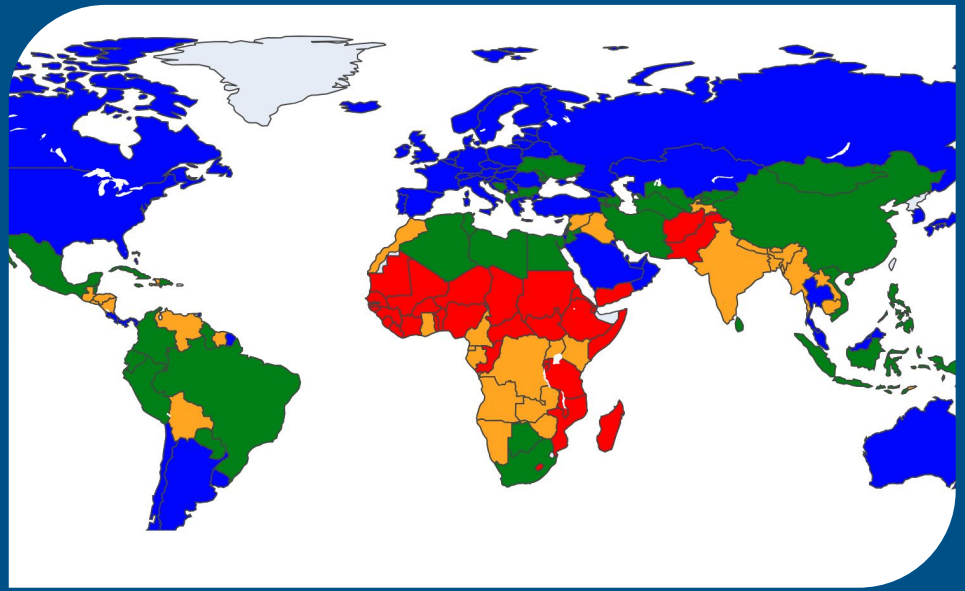
12.12.25

DABN14 - Project 2: Clustering Countries

- Clustering countries based on wealth and health metrics



Erik Broms & William Nordansjö

1. Introduction, Data & Research Question
2. Human Development Index
3. Models & Theory
4. Results & Evaluation
5. Conclusion

1. **Introduction, Data & Research Question**
2. Human Development Index
3. Models & Theory
4. Results & Evaluation
5. Conclusion

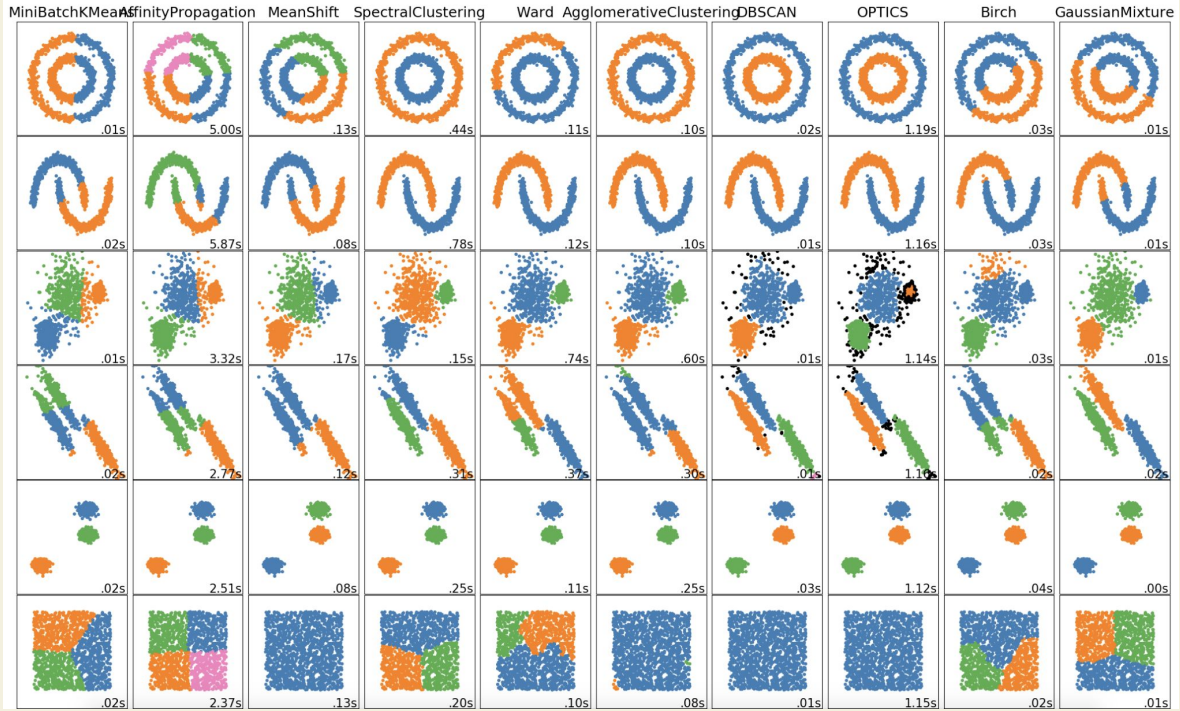
Introduction

Clustering:

- Letting the data tell the tale
- What are we looking for?
- What methods are we going to use?

Previous research:

- Countries, health and wealth
- Clustering



Data

Overview:

- HELP international
- 167 countries
- 10 feature columns

Variables:

- Health
- Social
- Economic

Outliers:

- Microstates skew the data!
  - (Luxembourg, Singapore, Malta)

Variable	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services per capita. Given as % of GDP per capita
health	Total health spending per capita. Given as % of GDP per capita
imports	Imports of goods and services per capita. Given as % of GDP per capita
income	Average net income per person
inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a newborn child would live
	if the current mortality patterns remain the same
total_fer	The number of children that would be born to each woman
	if the current age-fertility rates remain the same
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total
	population

“Can we generate meaningful insights and clusters into the socioeconomic and health-situation of the world's countries, and how do these compare to the clusters made by the human development index?”

1. Introduction, Data & Research Question
2. **Human Development Index**
3. Models & Theory
4. Results & Evaluation
5. Conclusion

Human Development Index (HDI)

Purpose:

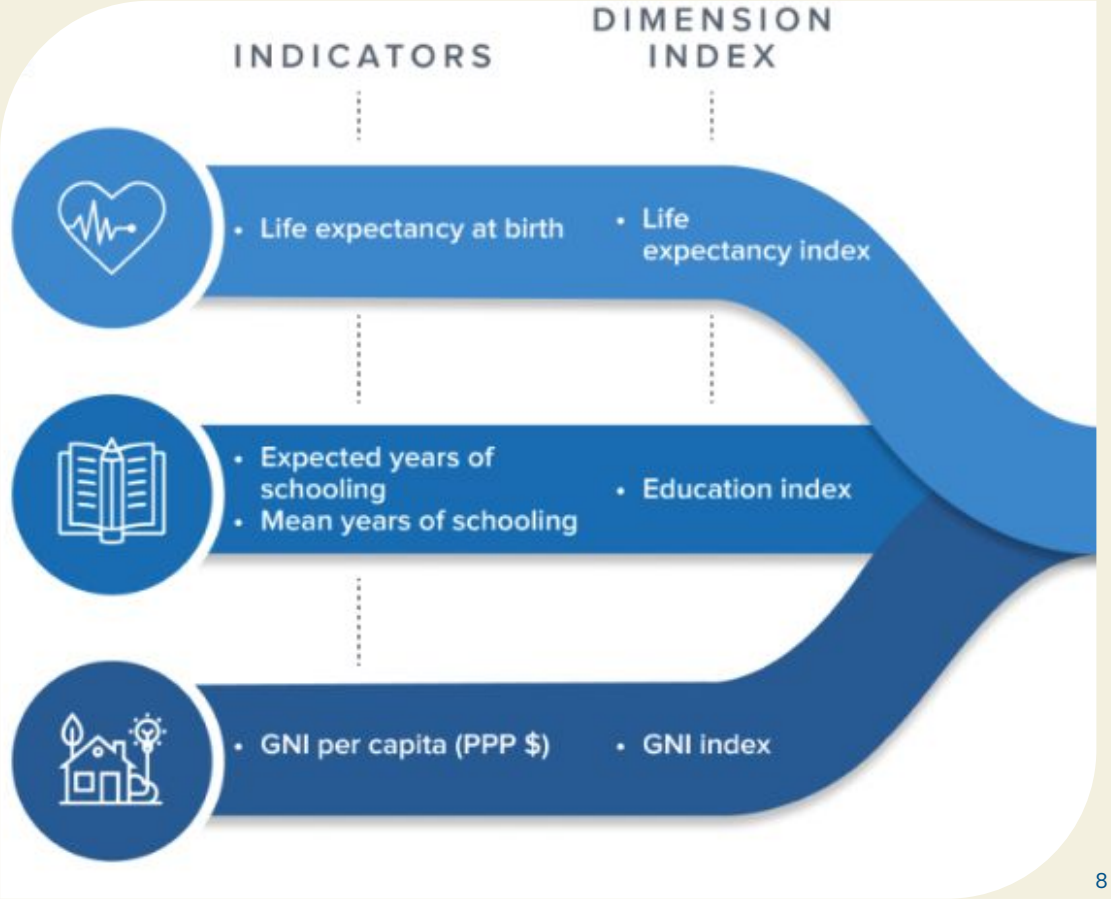
- Used to question national policy choices

Method:

- Assigns countries to groups according to achievements in human development

How we will use it:

- Baseline model



1. Introduction, Data & Research Question
2. Human Development Index
3. **Models & Theory**
4. Results & Evaluation
5. Conclusion

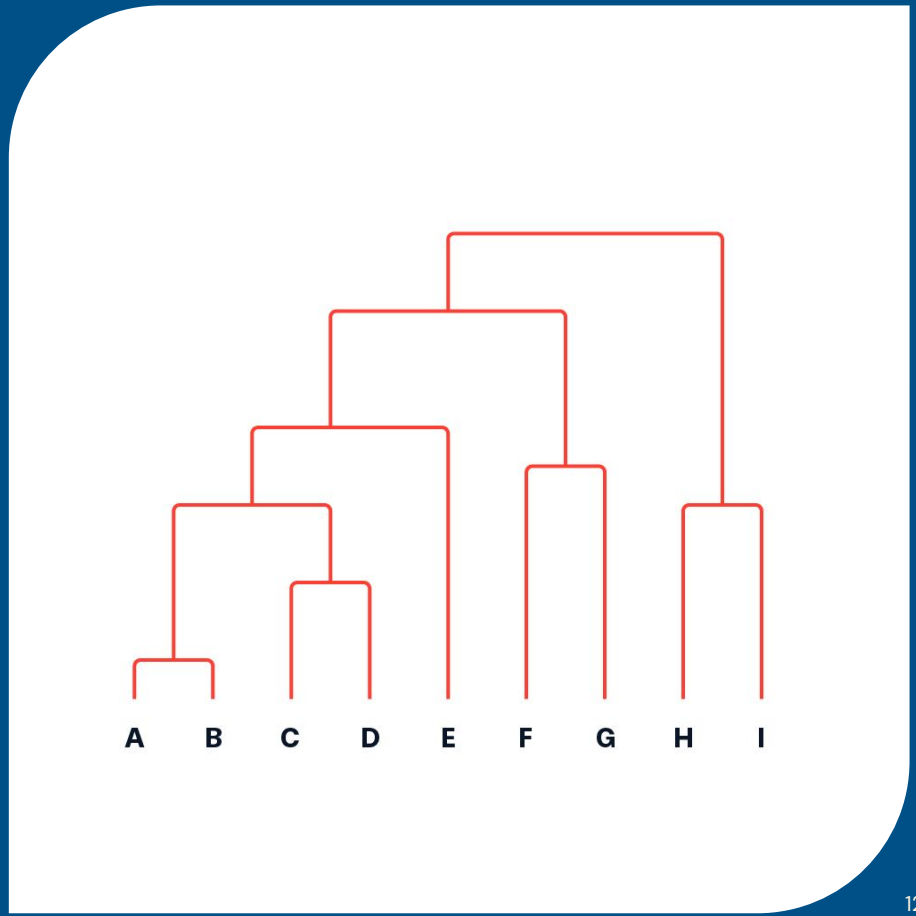
“K-mean Clustering is a frequently used algorithm to partition data into different clusters based on finding the minimum variation (sum of squares) within groups (or clusters).”

K-means Clustering

- Optimised algorithm as described by Harting & Wong (1979).
  - 6-step procedure where clusters are determined by transfers of data point based on sum of squares minimisation of clusters.
- Improved initialisation by Arthur & Vassilvitskii (2007): The K-Means++.
  - Chooses random starting centroids, uniformly in the data.

Hierarchical Clustering

- Divides data into clusters based on some similarity measure (linkage)
- Choice of linkage has a dramatic impact on clustering outcome

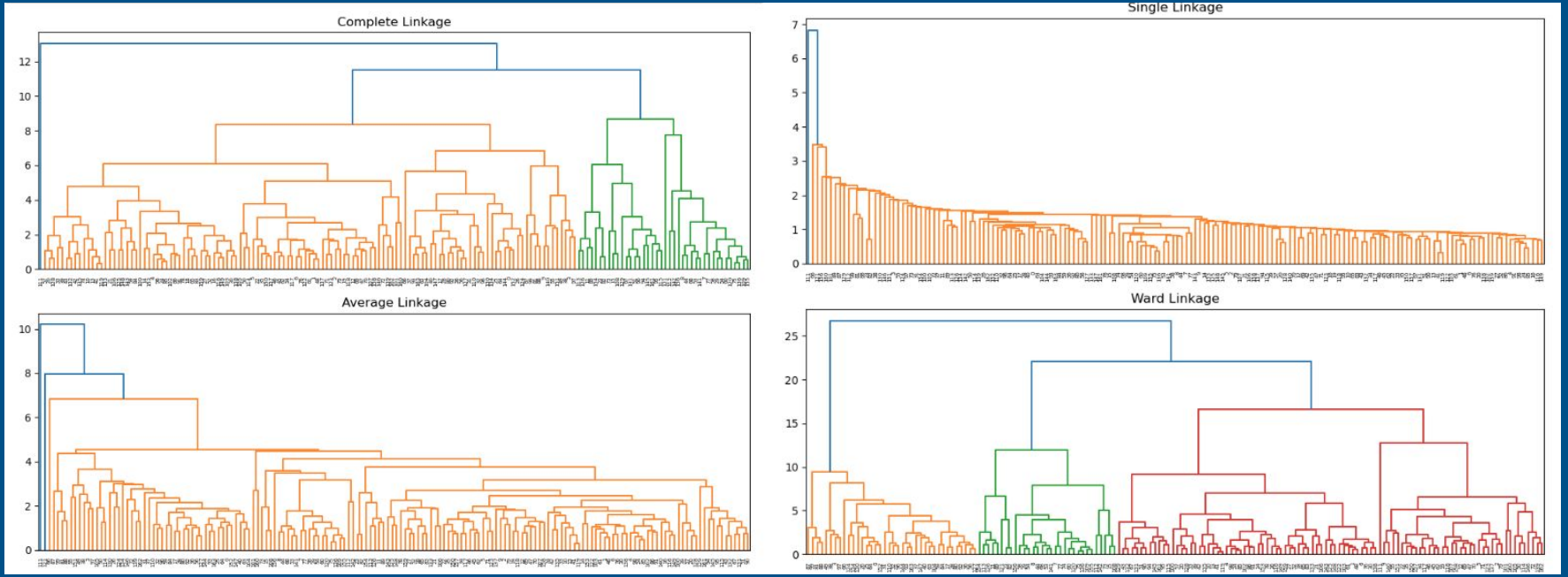


Adjusted Rand Index

- Assesses agreement between partitions and is a standard tool in cluster validation.
  - Counts pairs of objects and adjusts for chance.
- Will be calculated in relation to the HDI dataset.

1. Introduction, Data & Research Question
2. Human Development Index
3. Models & Theory
4. **Results & Evaluation**
5. Conclusion

Hierarchical Clustering Dendograms

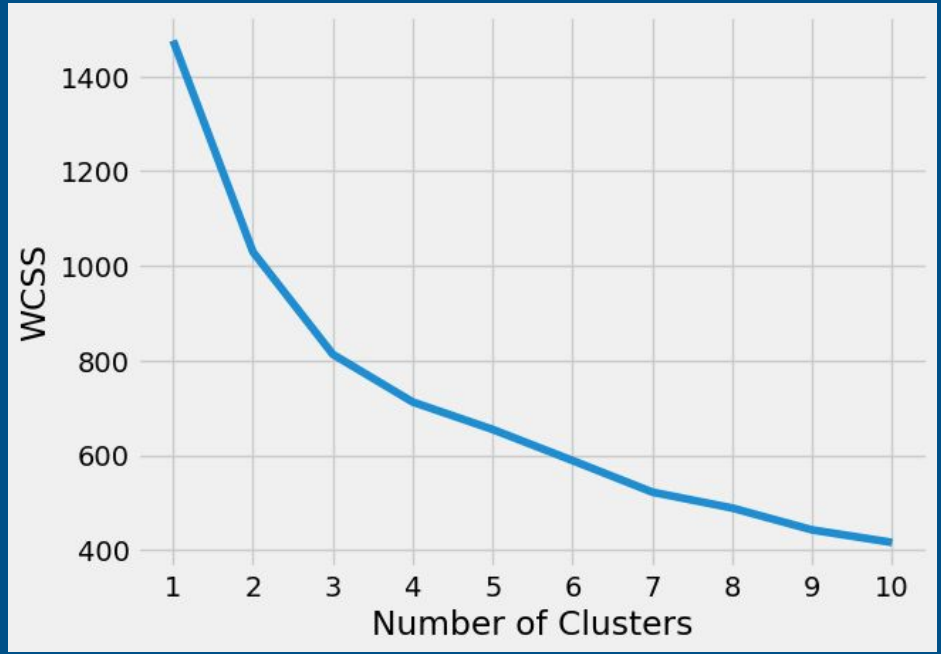


Ward linkage at K = 4 was chosen

K-means: Optimal number of clusters

Visualisation of within-cluster sum of squares (WCSS) for different K

- Optimal was found at K = 3
- K = 4 chosen for comparison with HDI & Hierarchical

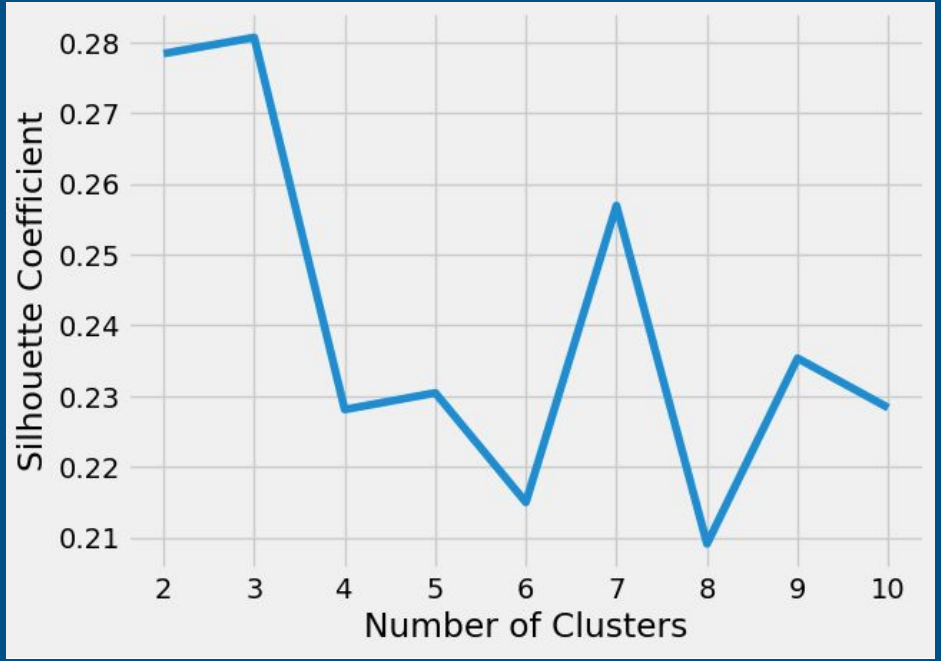



```
Optimal number K: 3
```

K-means: Optimal number of clusters

Visualisation of Silhouette Coefficient for different K

- Local maximum found at K = 3
- K = 4 chosen for comparison with HDI & Hierarchical

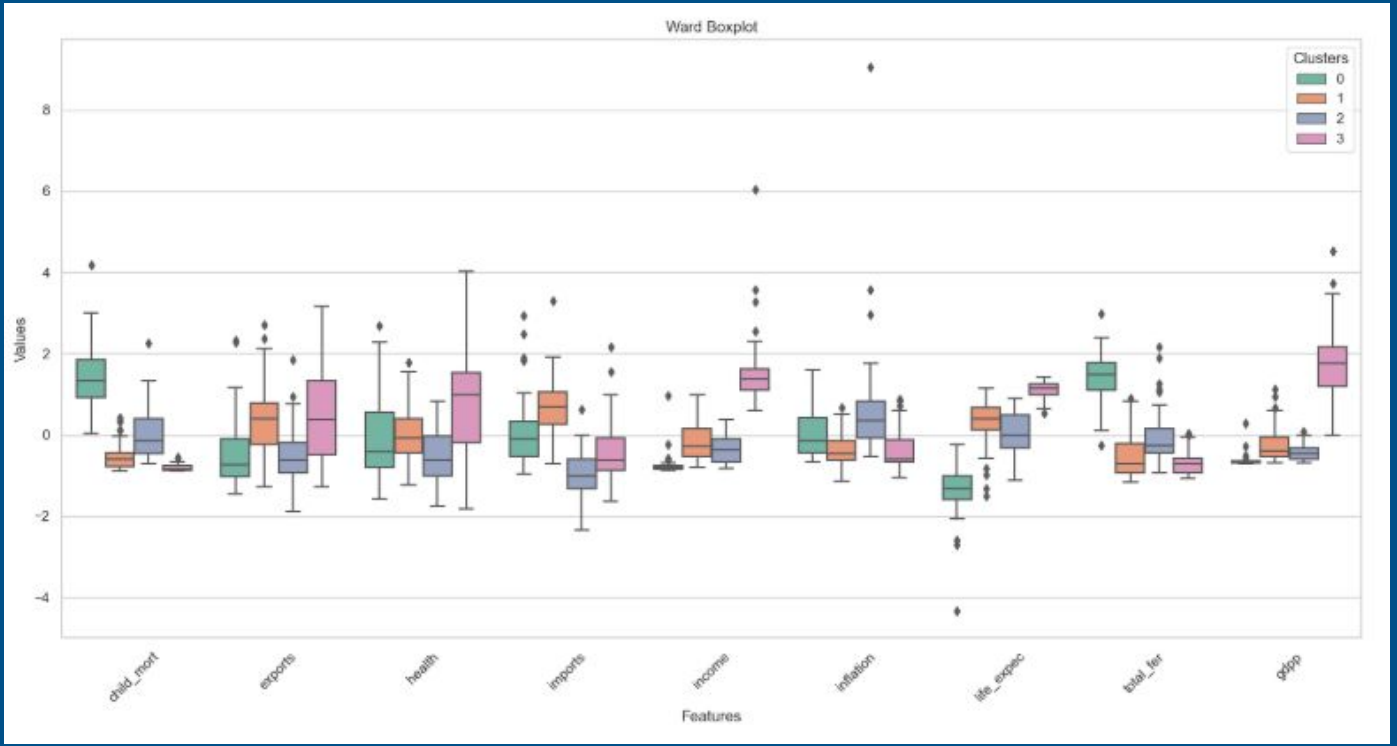


Evaluation

- Boxplots in order to compare the feature distribution of the clusters
- Maps in order to give a more intuitive sense of the clusters
- Adjusted Rand Index score in order to have some point of objective comparison

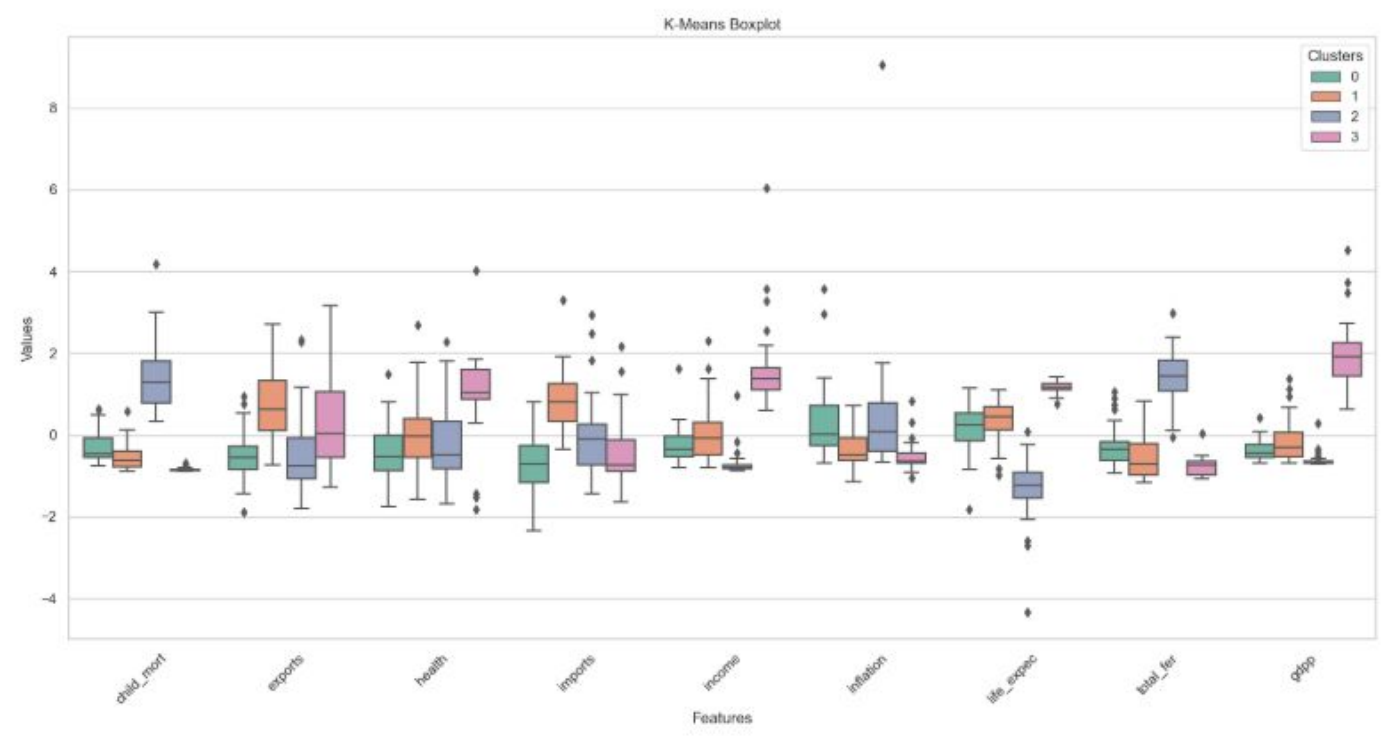
Boxplot: Wards Linkage

- Distinct clusters in some features
- Overlap in others
- Similar clusters to K-Means



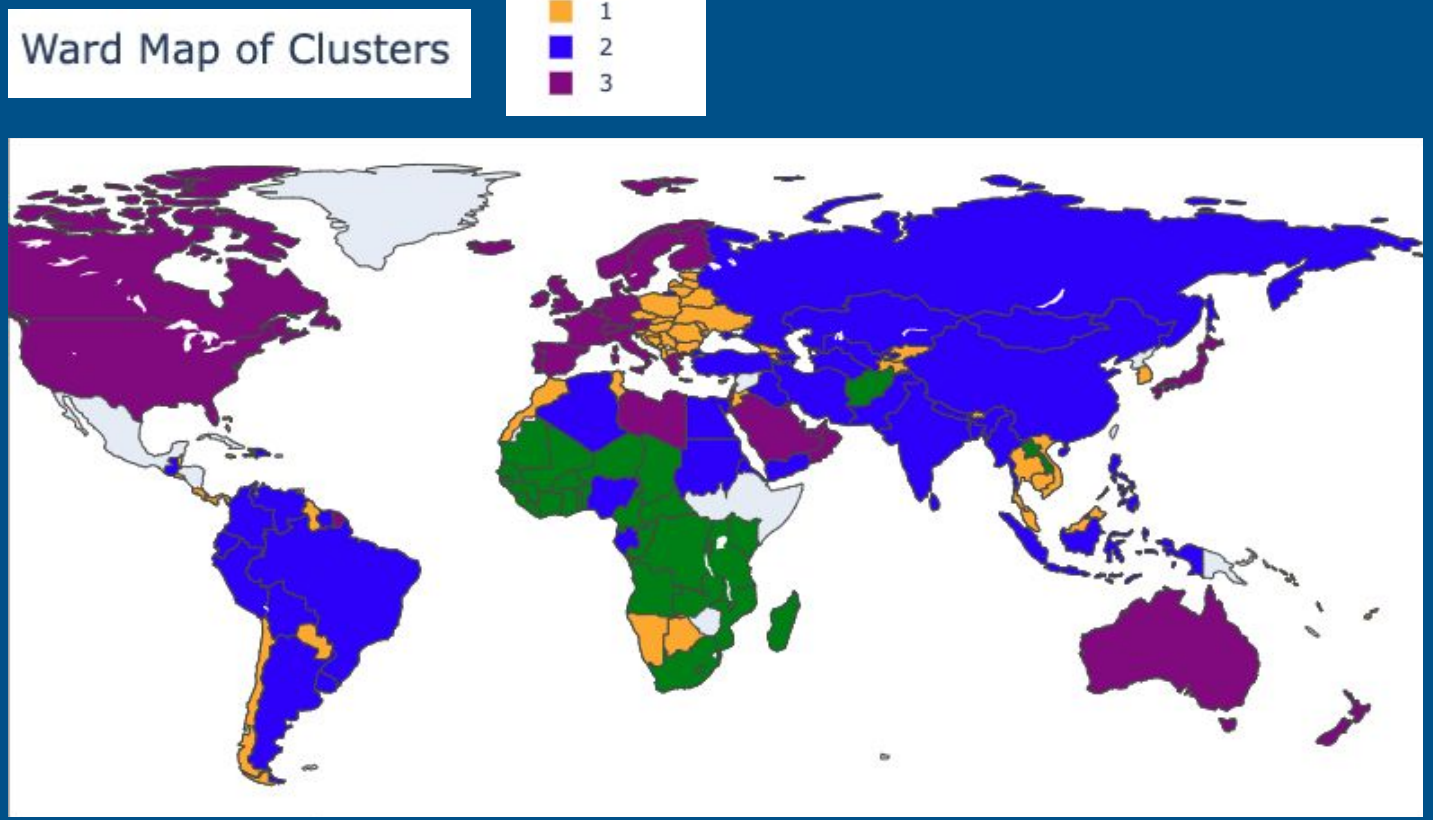
Boxplot: K-Means

- Distinct clusters in some features
- Overlap in others
- Similar clusters to K-Means



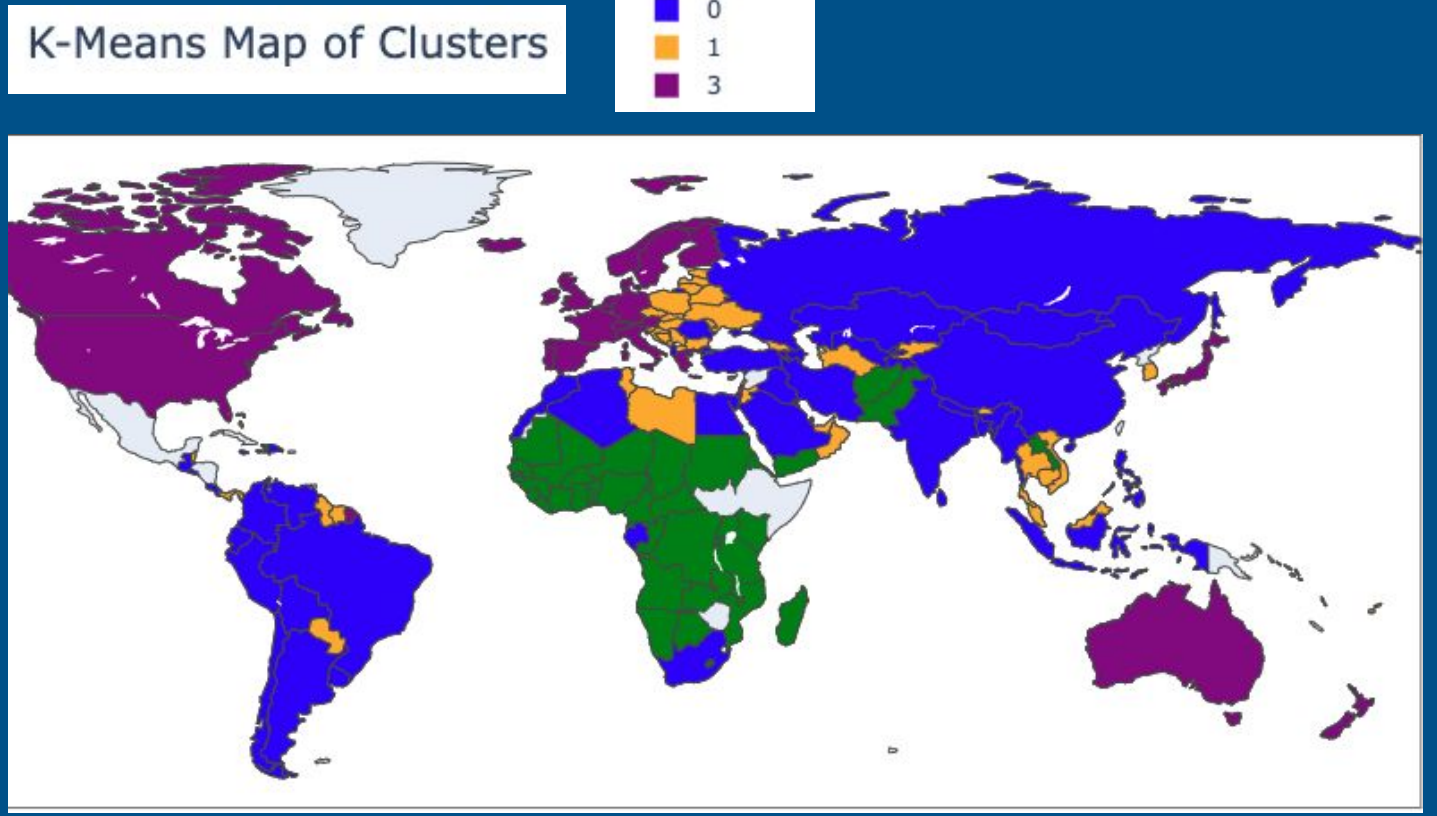
Map: Ward linkage

- Divided Europe
- Asia and South America together
- Africa split among all clusters



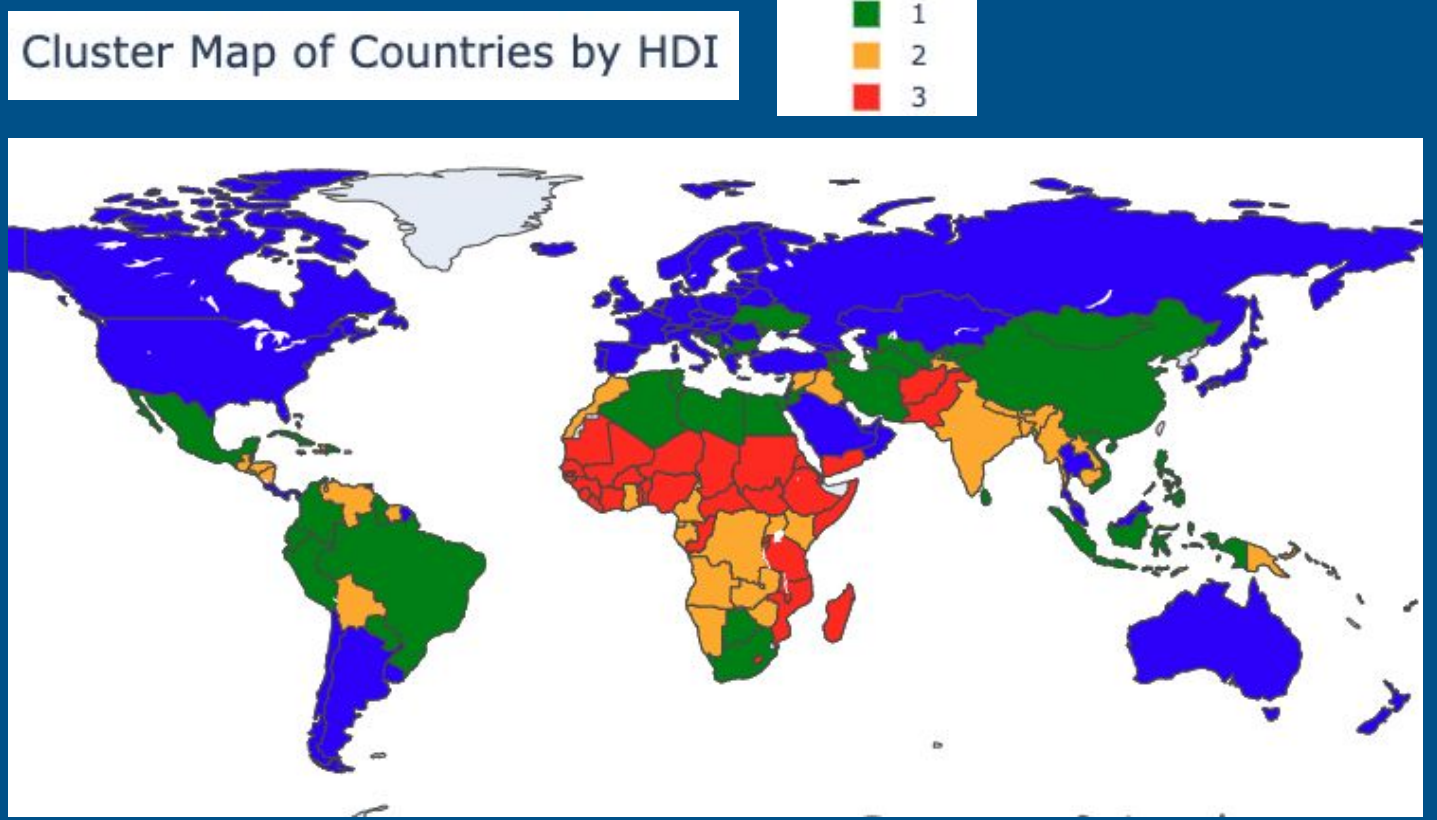
Map: K-Means

- Divided Europe
- Asia and South America together
- Africa split



Map: Ward

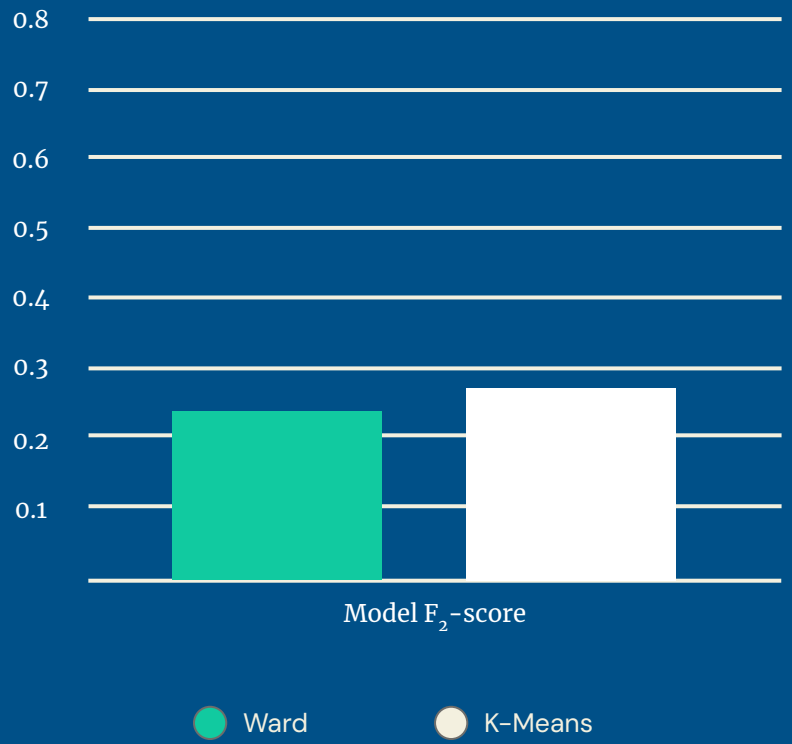
- Europe mostly united
- Asia and South America similar
- Africa heavily divided



Adjusted Rand Index score

Agreement in relation to the HDI dataset

- Ward score: 0.257
- K-Means score: 0.283



1. Introduction, Data & Research Question
2. Human Development Index
3. Models & Theory
4. Results & Evaluation
5. **Conclusion**

Conclusion

- Successful demonstration of clustering methods
- Meaningful clusters?
  - Motivational differences with HDI
- Could be expanded with a larger feature set and alternate algorithms in order to inform evidence-based policy decisions.