EAD PROJECT II:
WORLD HAPPINESS INDEX ANALYSIS

Miguel Tavares up200902937

Hélder Vieira up201503395

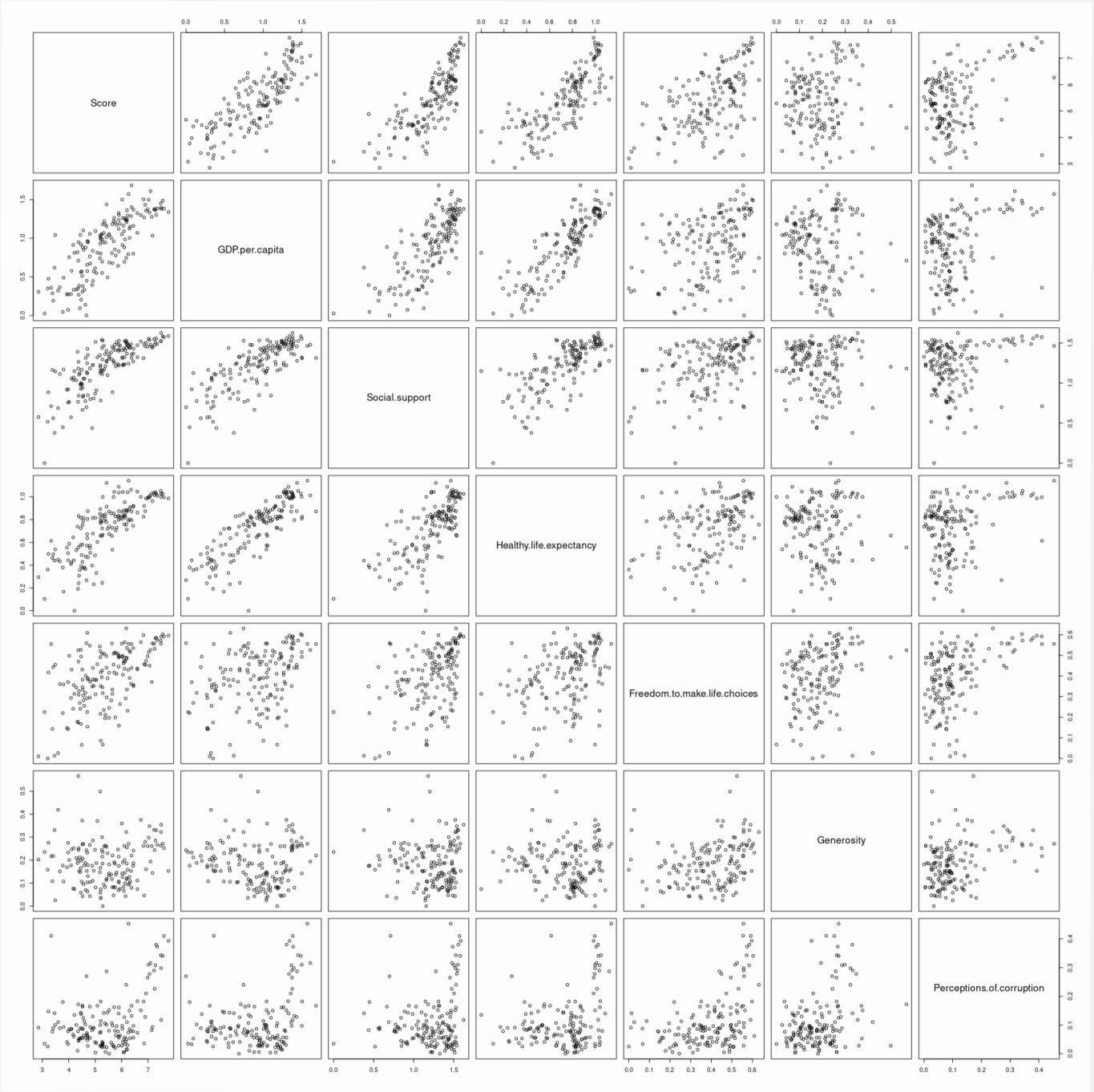
MDS

DATASET DESCRIPTION

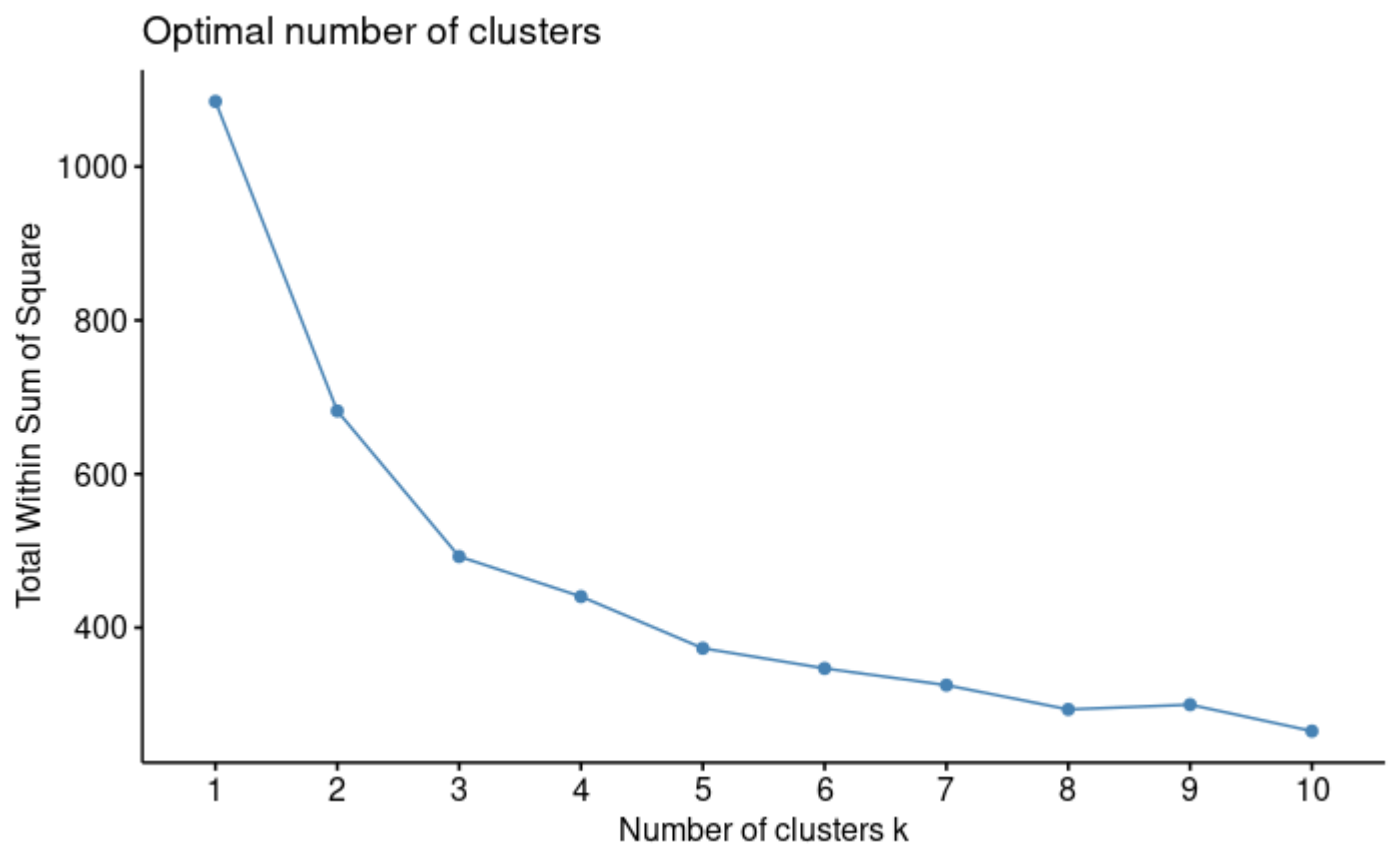
- Hapiness Index of several countries from 2019
- Most of the data retrieved through surveys

DATASET DESCRIPTION

- Features:
  - Score
  - GDP per capita
  - Social Support
  - Healthy life expectancy
  - Freedom to make life choices
  - Generosity
  - Perception of corruption

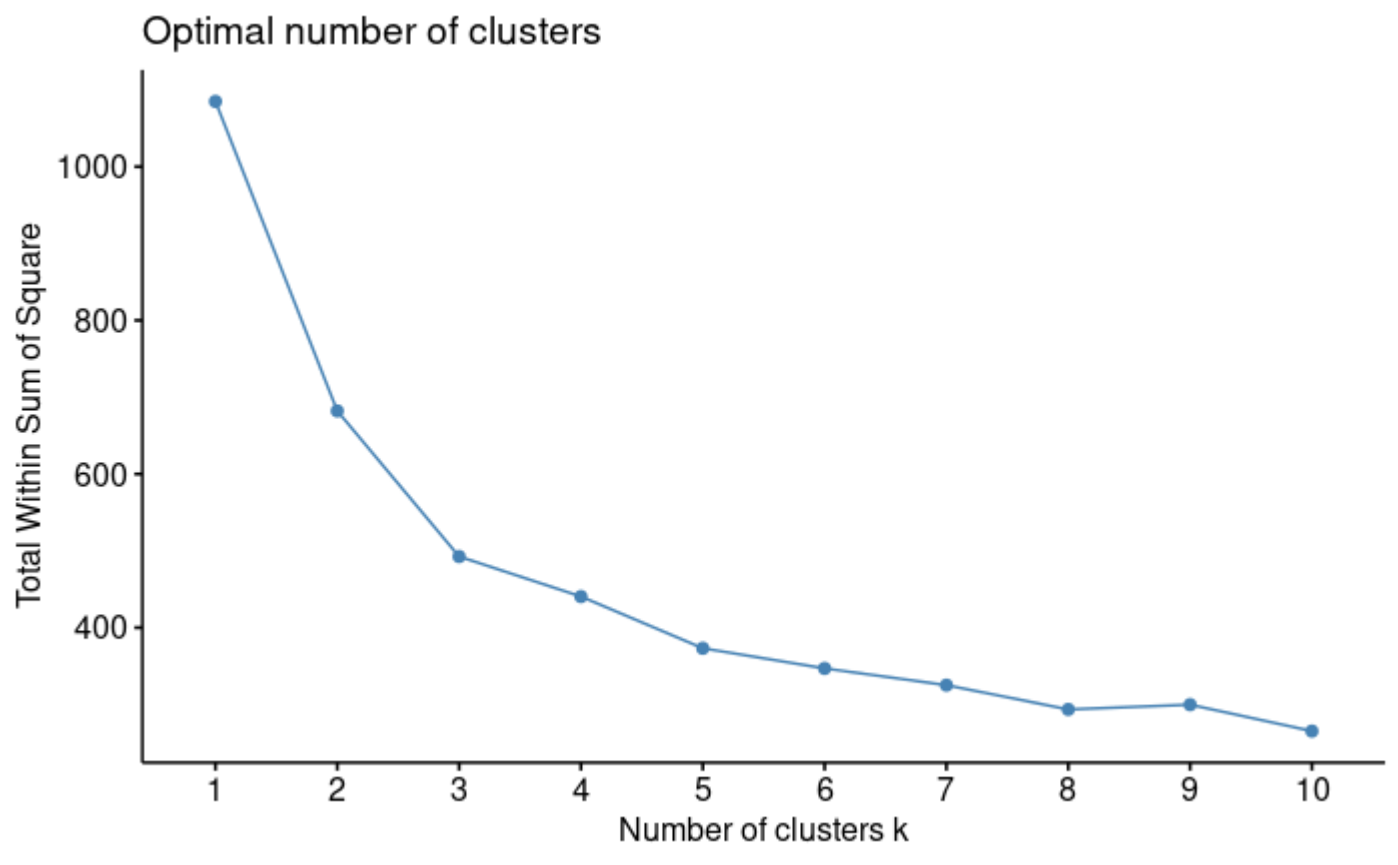

Score
GDP.per.capita
Social.support
Healthy.life.expectancy
Freedom.to.make.life.choices
Generosity
Perceptions.of.corruption

KMEANS CLUSTERING



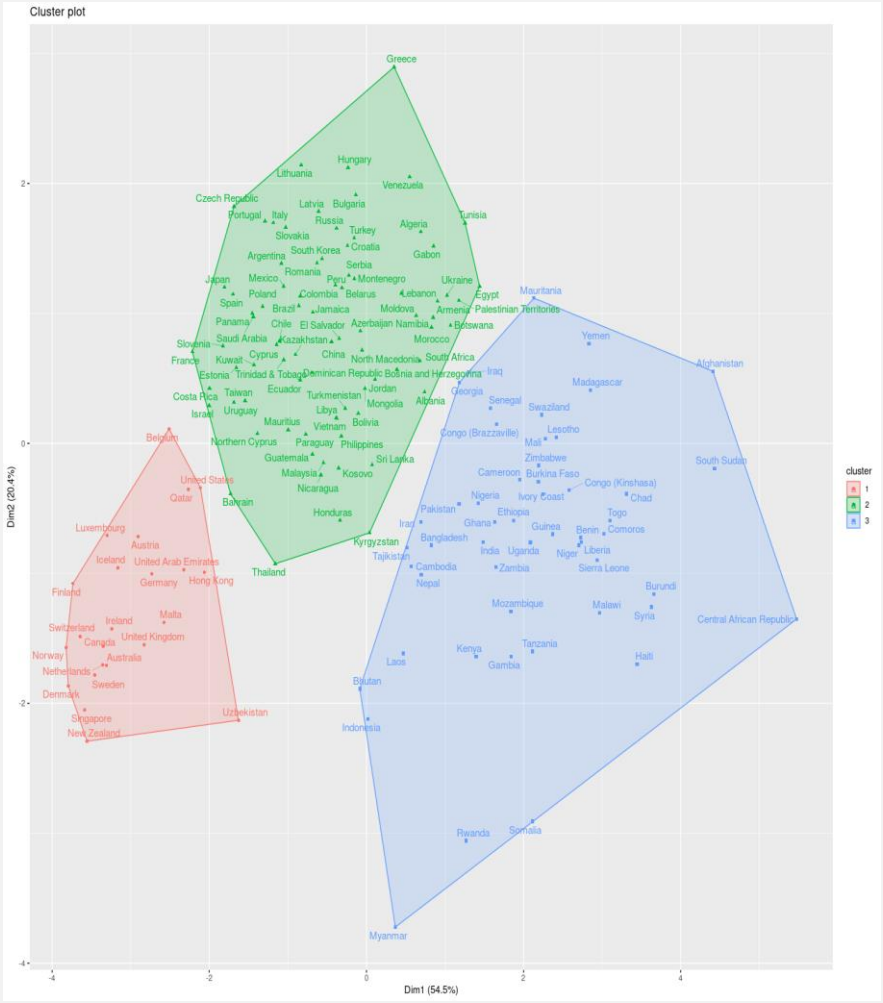
- Elbow point: k = 3

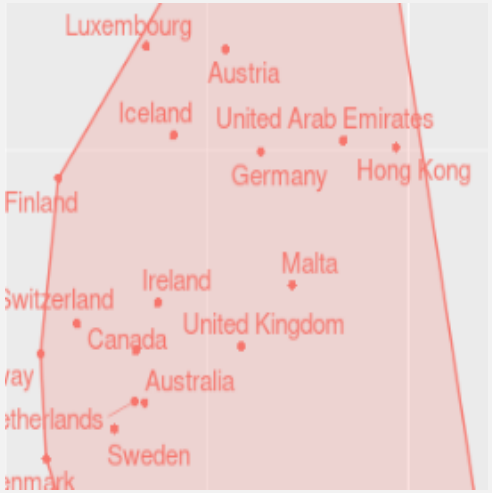
KMEANS CLUSTERING

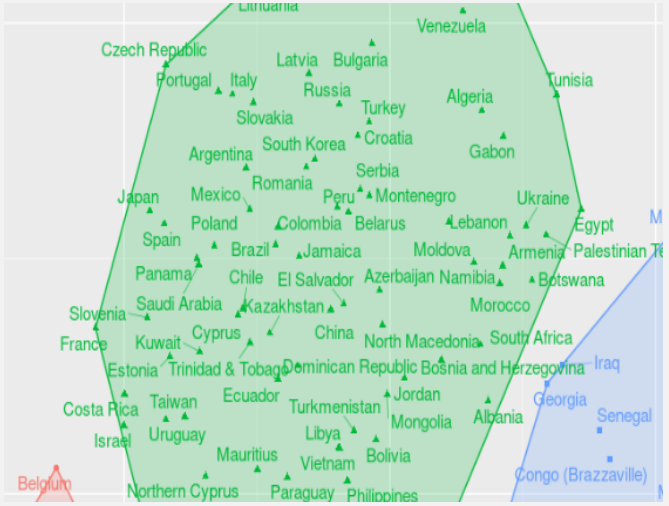


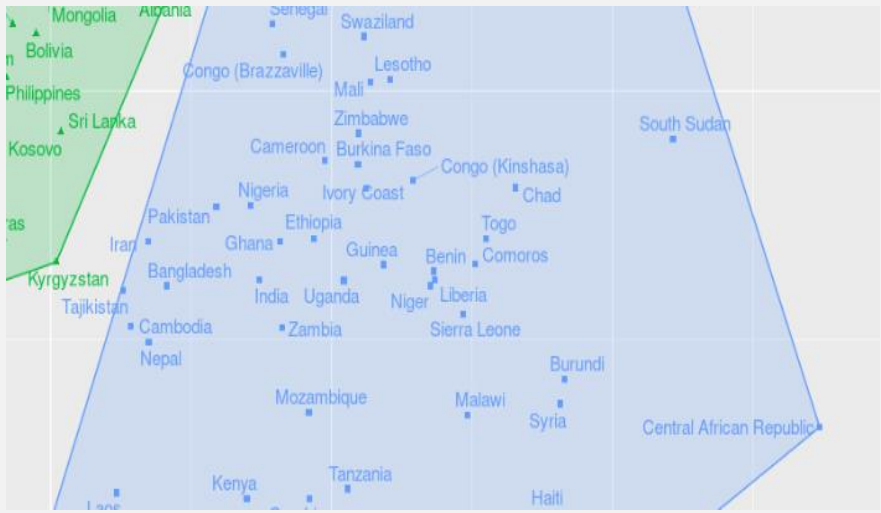
- Elbow point: k = 3

KMEANS CLUSTERING

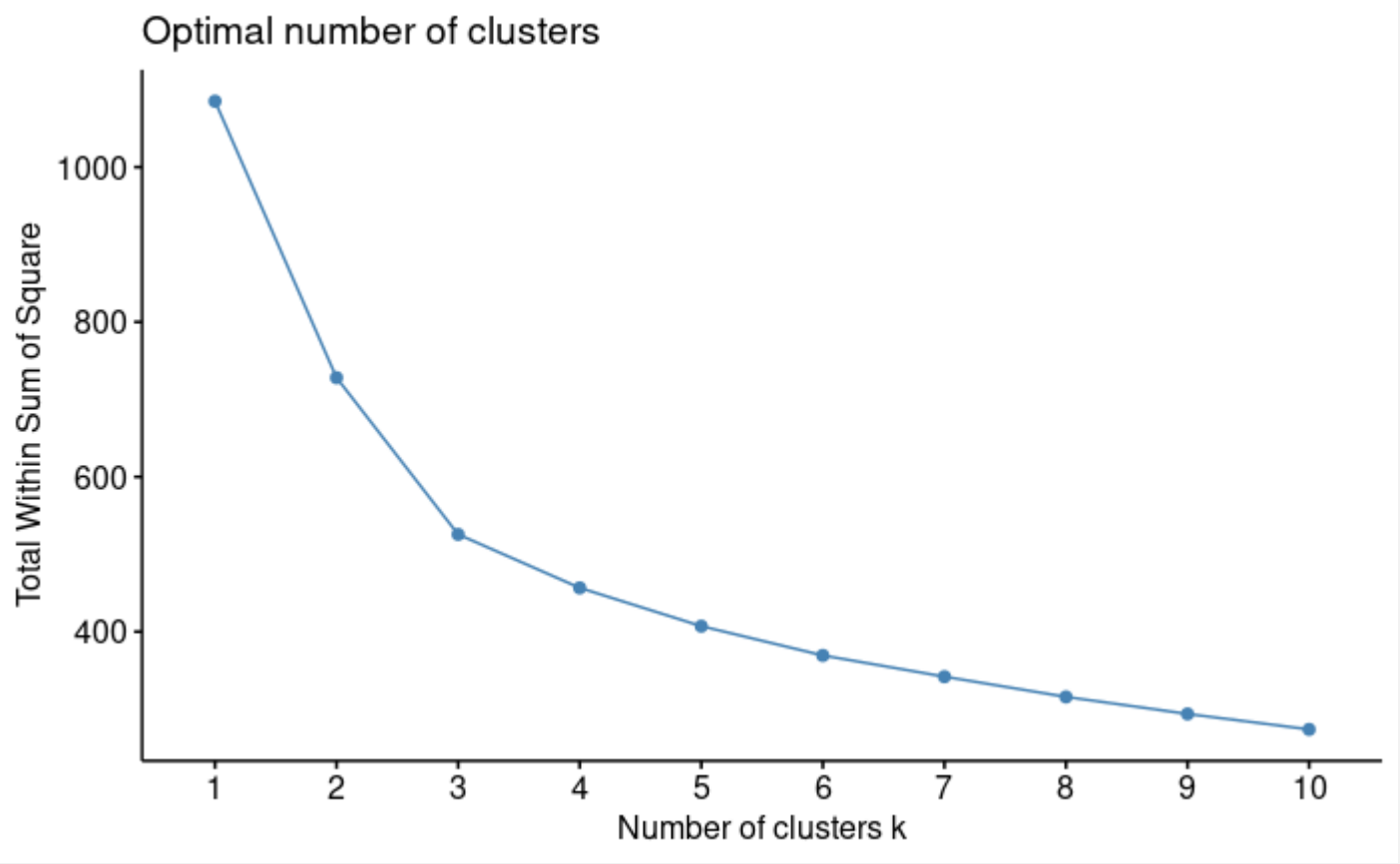






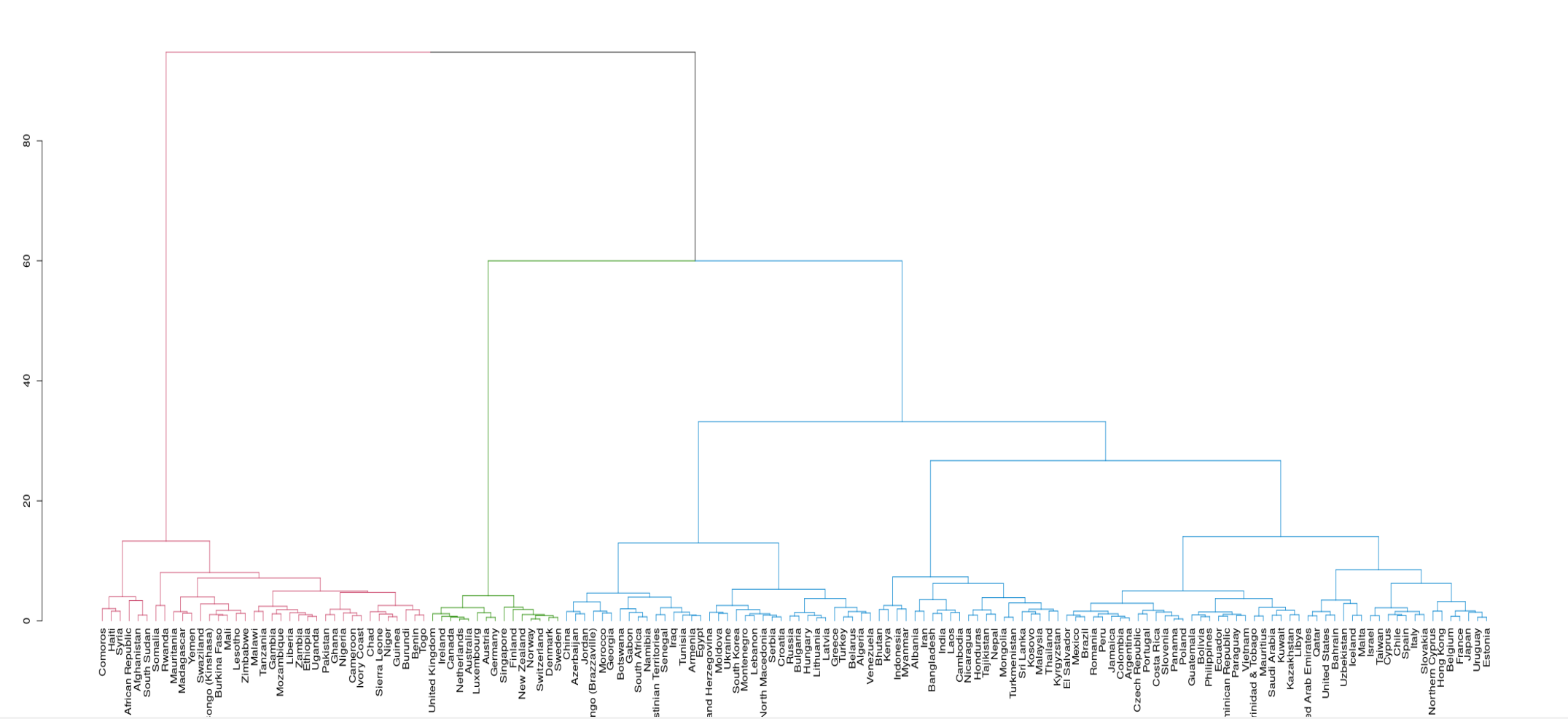


HIERARCHICAL CLUSTERING

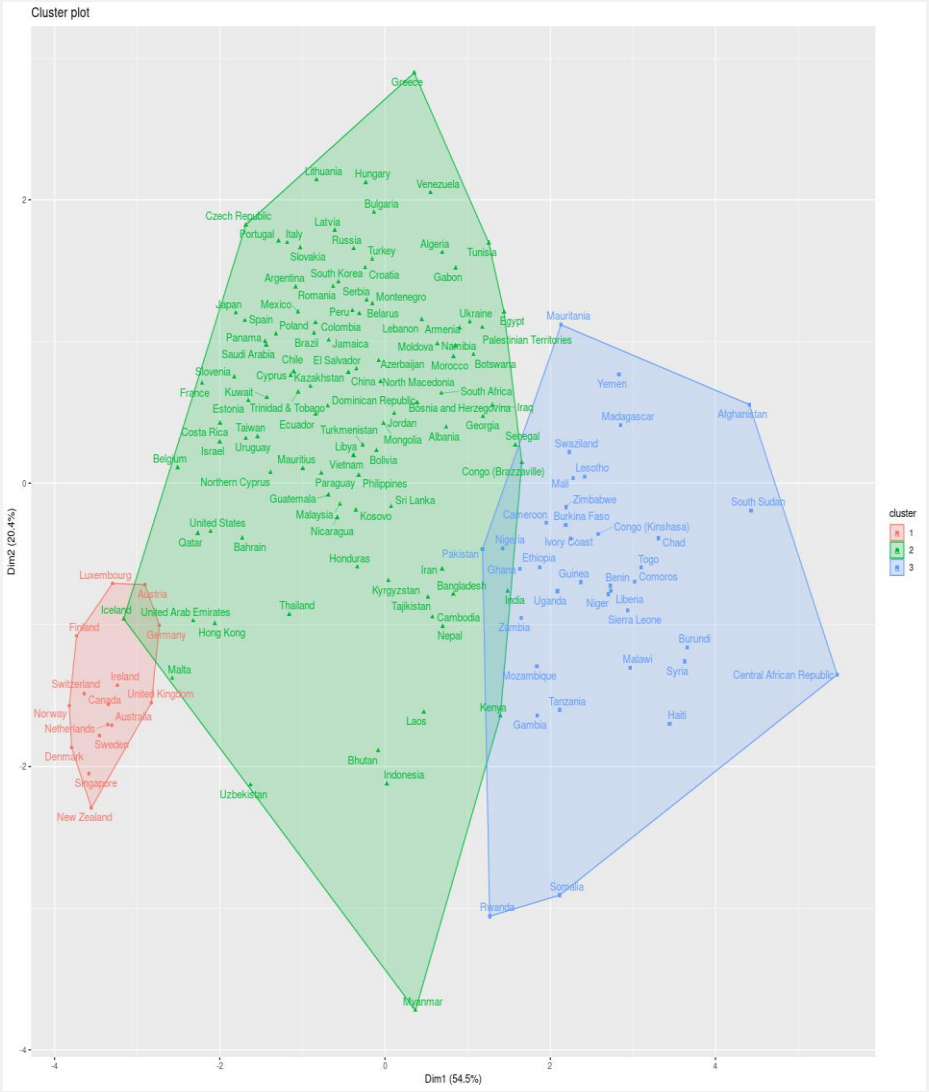


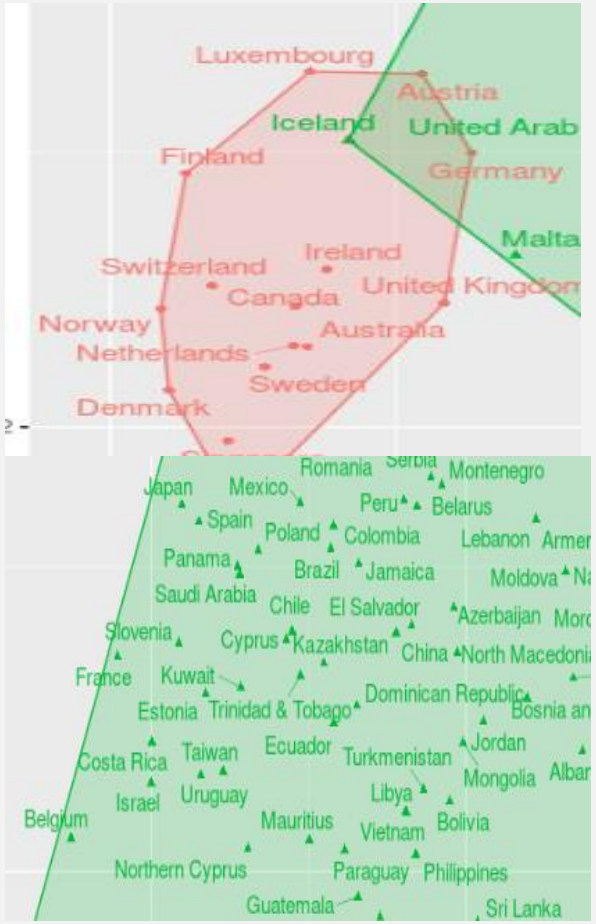
- Elbow point: k = 3

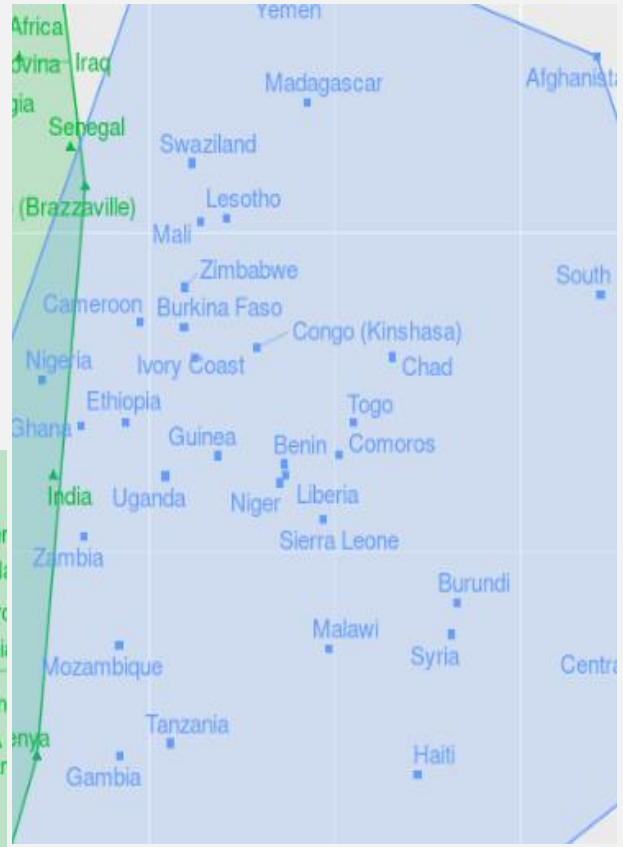
HIERARCHICAL CLUSTERING



HIERARCHICAL CLUSTERING







CLUSTERING COMPARISON

	Jaccard	Rand Index
Result	0.6092326	0.7801489

CLASSIFICATION

- Data from the PCA derived components was used (75% of the total variance explained by 2 components);
- ‘Continent’ was used as the target variable;
- A stratified split with 60% of the data for train (95 countries) and 40% test (61 countries) was applied, a seed was used to allow results duplication;

CLASSIFICATION METHODS

- LDA – Linear Discriminant Analysis

- MLR- Multinomial Logistic Regression

CLASSIFICATION - LDA

- Training:

Predicted	Africa	Asia	Europe	North America	Oceania	South America
Africa	20	6	0	1	0	0
Asia	7	13	4	2	0	1
Europe	0	9	19	5	1	5
North America	0	0	0	0	0	0
Oceania	0	0	0	0	1	0
South America	0	0	1	0	0	0

Confusion Matrix and Statistics

It resulted in a 55.8% accuracy over the training data

CLASSIFICATION – LDA

- Class specific training results:

	Class: Africa	Class: Asia	Class: Europe	Class: North America	Class: Oceania	Class: South America
Sensitivity	0.7407	0.4643	0.7917	0.00000	0.50000	0.00000
Specificity	0.8971	0.7910	0.7183	1.00000	1.00000	0.98876
Pos Pred Value	0.7407	0.4815	0.4872	NaN	1.00000	0.00000
Neg Pred Value	0.8971	0.7794	0.9107	0.91579	0.98936	0.93617
Prevalence	0.2842	0.2947	0.2526	0.08421	0.02105	0.06316
Detection Rate	0.2105	0.1368	0.2000	0.00000	0.01053	0.00000
Detection Prevalence	0.2842	0.2842	0.4105	0.00000	0.01053	0.01053
Balanced Accuracy	0.8189	0.6277	0.7550	0.50000	0.75000	0.49438

CLASSIFICATION - LDA

- Test:

Predicted	Africa	Asia	Europe	North America	South America
Africa	15	0	0	0	0
Asia	2	12	2	2	2
Europe	1	6	14	3	2
North America	0	0	0	0	0
Oceania	0	0	0	0	0
South America	0	0	0	0	0

The model scored 67 % in terms of test accuracy;

CLASSIFICATION - MLR

- Training

	Reference					
Prediction	Africa	Asia	Europe	North America	Oceania	South America
Africa	23	6	1	1	0	0
Asia	4	14	3	3	0	2
Europe	0	7	20	4	1	4
North America	0	0	0	0	0	0
Oceania	0	1	0	0	1	0
South America	0	0	0	0	0	0

MLR resulted in a 61% accuracy over the training set

CLASSIFICATION - MLR

- Specific class results:

Statistics by Class:

	Class: Africa	Class: Asia	Class: Europe	Class: North America	Class: Oceania	Class: South America
Sensitivity	0.8519	0.5000	0.8333	0.00000	0.50000	0.00000
Specificity	0.8824	0.8209	0.7746	1.00000	0.98925	1.00000
Pos Pred Value	0.7419	0.5385	0.5556	NaN	0.50000	NaN
Neg Pred Value	0.9375	0.7971	0.9322	0.91579	0.98925	0.93684
Prevalence	0.2842	0.2947	0.2526	0.08421	0.02105	0.06316
Detection Rate	0.2421	0.1474	0.2105	0.00000	0.01053	0.00000
Detection Prevalence	0.3263	0.2737	0.3789	0.00000	0.02105	0.00000
Balanced Accuracy	0.8671	0.6604	0.8040	0.50000	0.74462	0.50000

CLASSIFICATION - MLR

- Test:

	Reference					
Prediction	Africa	Asia	Europe	North America	Oceania	South America
Africa	16	3	0	0	0	0
Asia	2	11	4	2	0	1
Europe	0	4	11	3	0	3
North America	0	0	0	0	0	0
Oceania	0	0	1	0	0	0
South America	0	0	0	0	0	0

Accuracy: 62.3%

CLASSIFICATION - MLR

- Specific class test results:

Statistics by Class:

	Class: Africa	Class: Asia	Class: Europe	Class: North America	Class: Oceania	Class: South America
Sensitivity	0.8889	0.6111	0.6875	0.00000	NA	0.00000
Specificity	0.9302	0.7907	0.7778	1.00000	0.98361	1.00000
Pos Pred Value	0.8421	0.5500	0.5238	NaN	NA	NaN
Neg Pred Value	0.9524	0.8293	0.8750	0.91803	NA	0.93443
Prevalence	0.2951	0.2951	0.2623	0.08197	0.00000	0.06557
Detection Rate	0.2623	0.1803	0.1803	0.00000	0.00000	0.00000
Detection Prevalence	0.3115	0.3279	0.3443	0.00000	0.01639	0.00000
Balanced Accuracy	0.9096	0.7009	0.7326	0.50000	NA	0.50000

MODEL COMPARISON

- Both models had a similar performance with more than 60% accuracy for the test dataset;
- Both classified relatively well countries that belong to Europe and Africa;
- In the other continents the classification was worse, this can be explained by the fact that some of the continents like Oceania, South America and North America have a smaller representation in the dataset;
- Parameter tuning could improve the obtained results;