UNIVERSITÀ
DEGLI STUDI
DI MILANO

Course: Statistical Learning

Behind the Curtain: Political Ideologies and the Impact of Cancer

FRANCO REINALDO BONIFACINI (41540A)

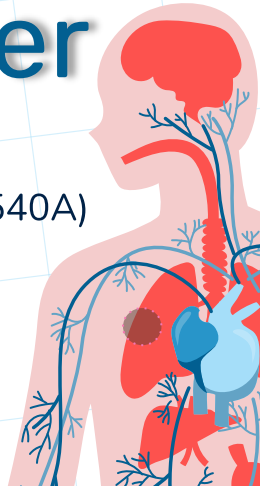


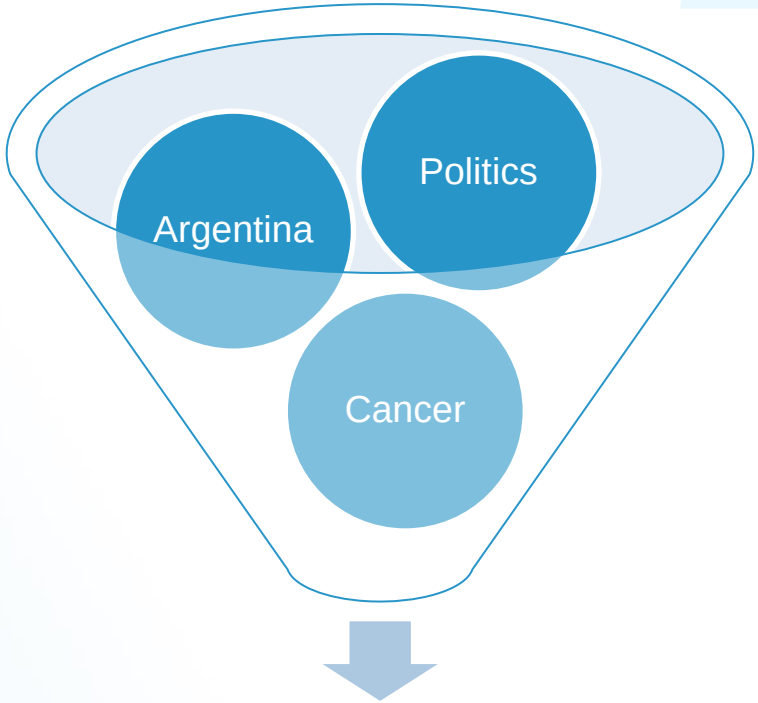
Table of contents

01 Introduction

Project's Goal

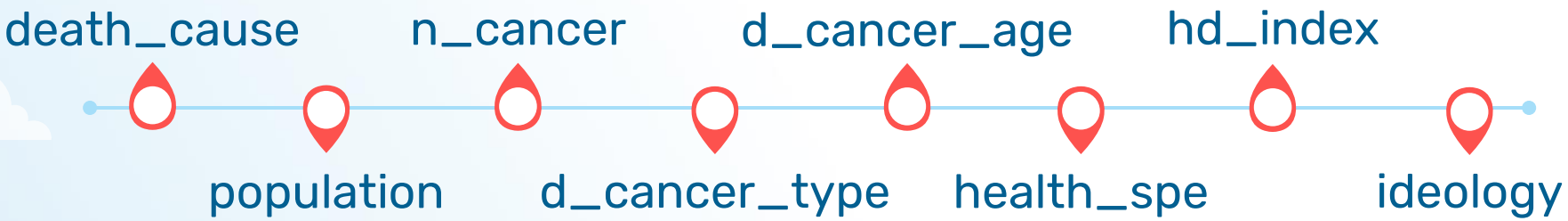
With no intention of engaging in political discourse or ideologies, the project aims to illustrate that **the occurrence of cancer-related deaths** correlates with factors such as **investment in public health and the Human Development Index (HDI), rather tan the messages conveyed by political parties**.

"*It is easy to preach with the word, but the world truly requires preaching through the example*".

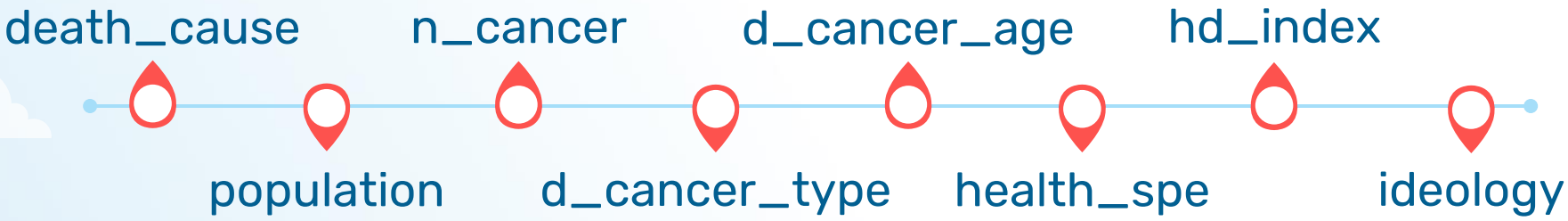


Final project

Datasets used and preprocessing



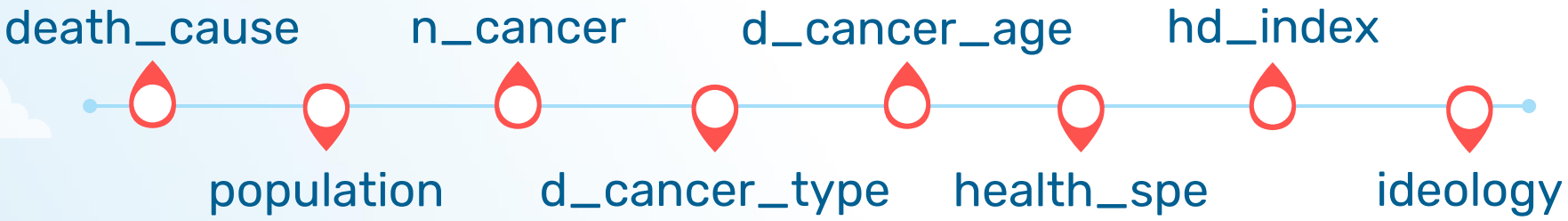
Datasets used and preprocessing



##	Dataset	Min	Max
## 1	death_cause	1990	2019
## 2	population	-10000	2021
## 3	n_cancer	1990	2017
## 4	d_cancer_type	1990	2019
## 5	d_cancer_age	1990	2019
## 6	health_spe	2000	2019
## 7	hd_index	1990	2021
## 8	ideology	1945	2020

Normalizing country names

Datasets used and preprocessing



	Dataset	Min	Max
1	death_cause	1990	2019
2	population	-10000	2021
3	n_cancer	1990	2017
4	d_cancer_type	1990	2019
5	d_cancer_age	1990	2019
6	health_spe	2000	2019
7	hd_index	1990	2021
8	ideology	1945	2020

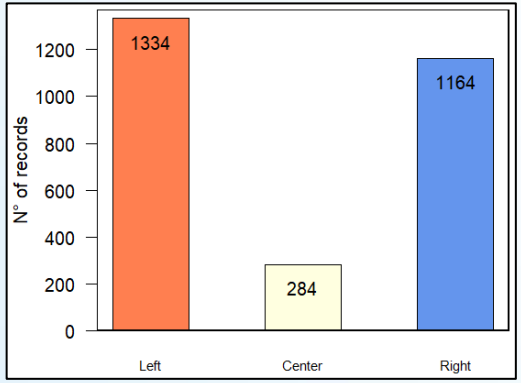
Normalizing country names

New variables:

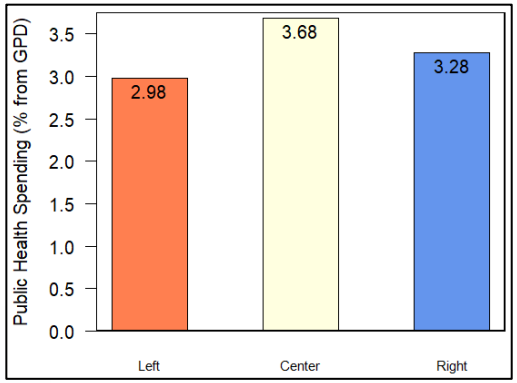
- cancer_affection_rate
- cancer_death_rate
- Rate of death by age.
- Rate of death by type

Exploratory Data Analysis

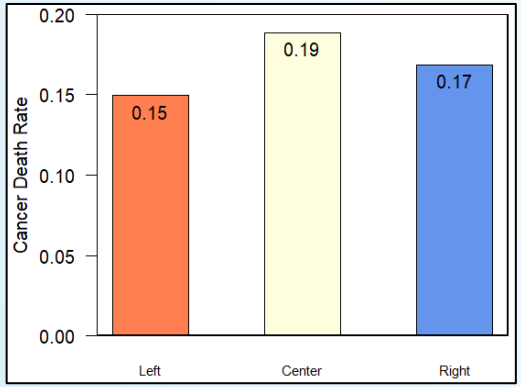
Number of Records by Ideology



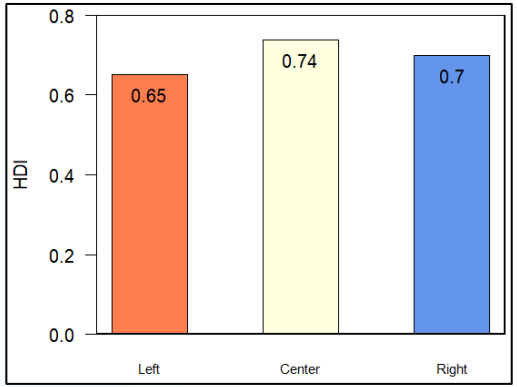
Avg. Spending in Public Health by Ideology



Avg. Cancer Death by Ideology

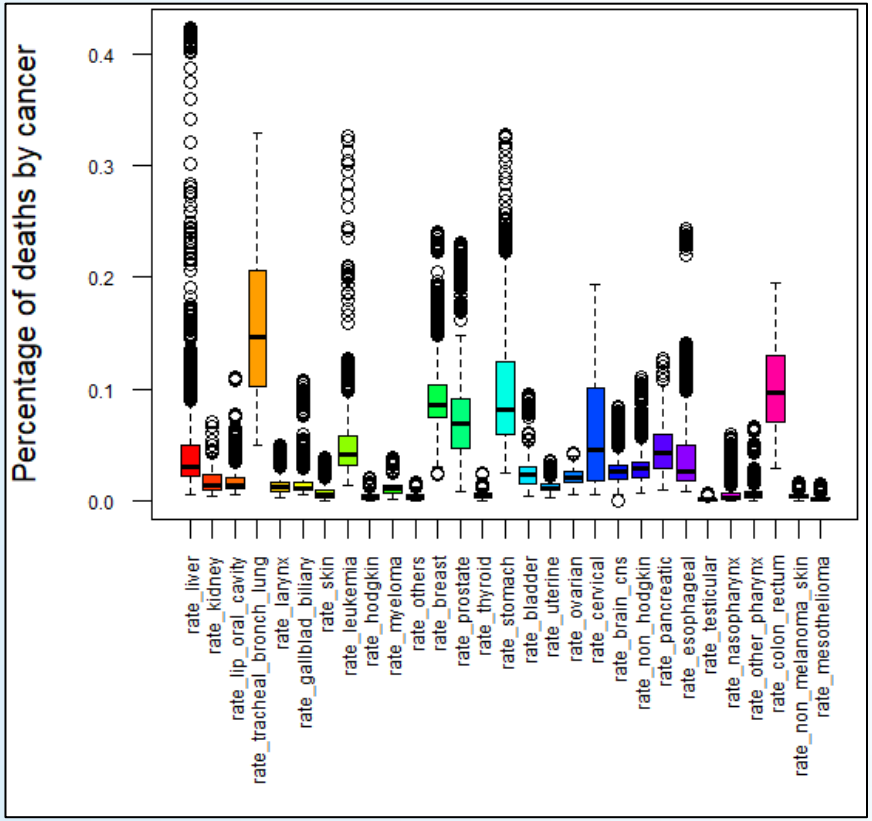


Avg. HDI by Ideology

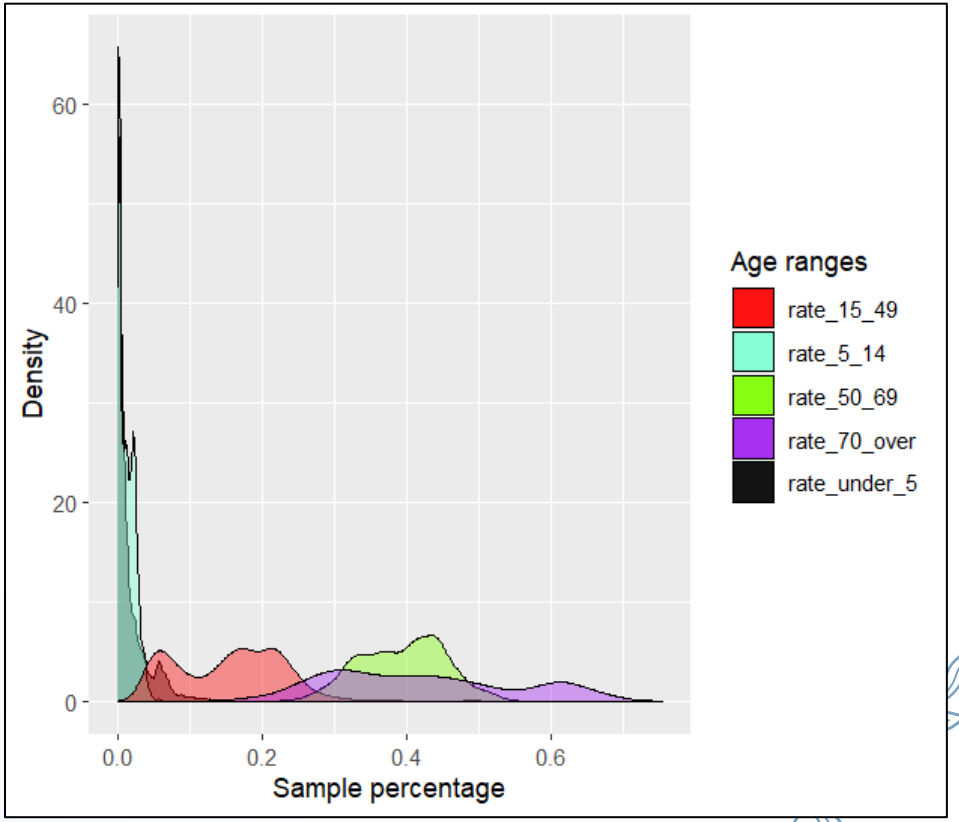


Exploratory Data Analysis

Types of Cancer

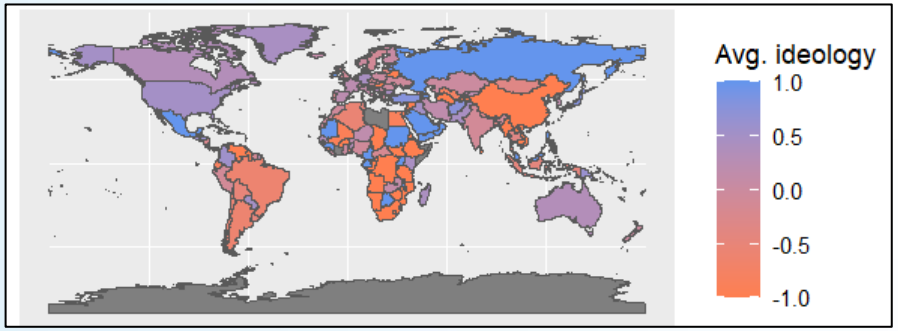


Deaths from Cancer by Range of Age



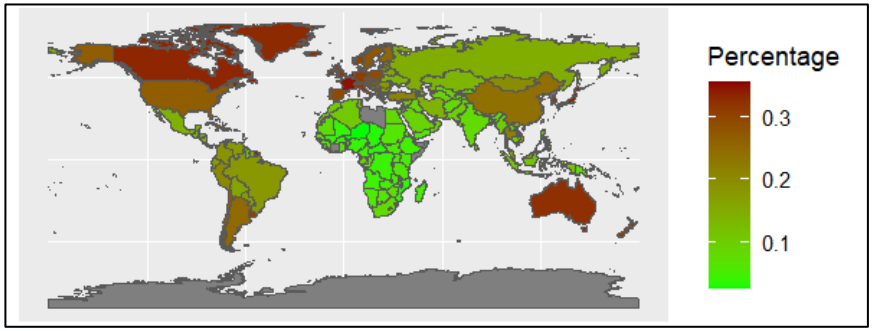
Exploratory Data Analysis

Anual Average Ideology



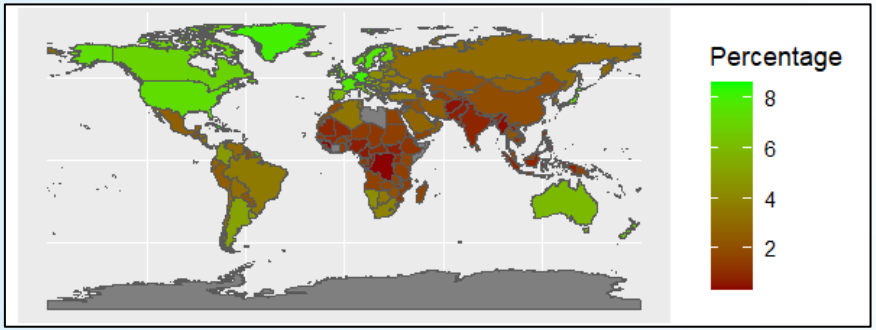
Anual Average Death by Cancer

(% from total deaths)

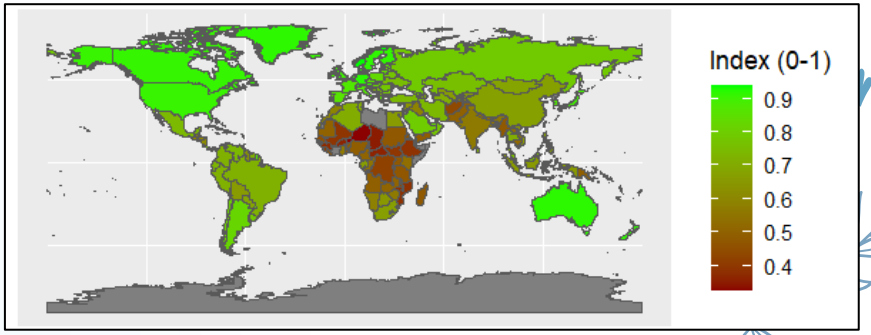


Anual Average Public Health Spending

(% from GDP)



Anual Average HDI

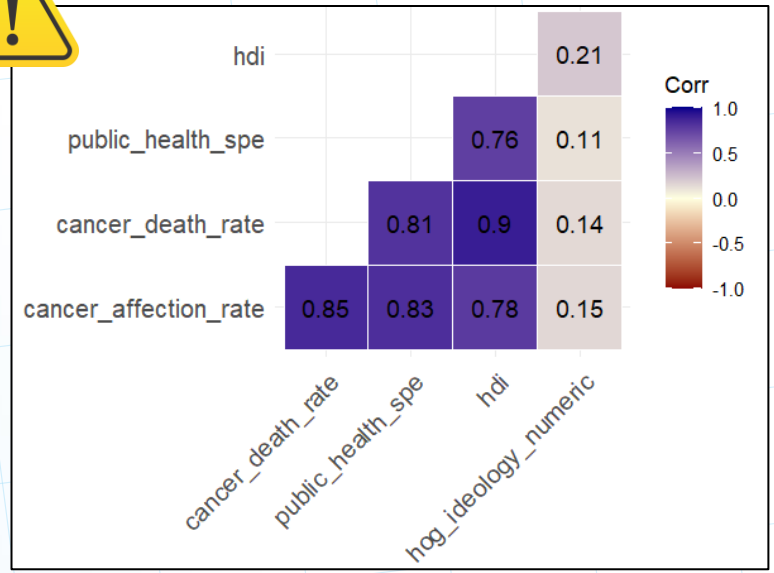


02 Unsupervised

Variables Used

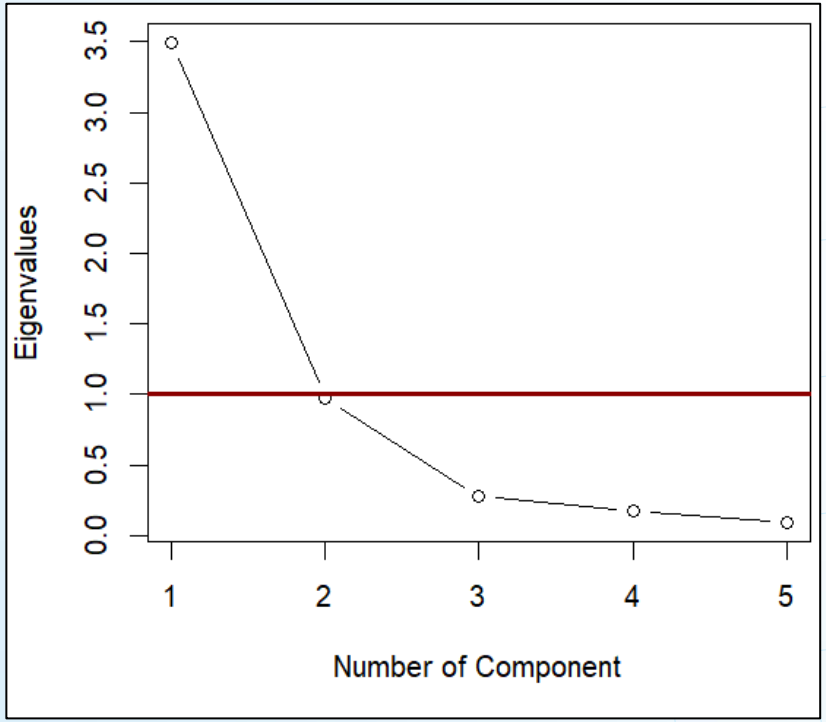
Variables
cancer_affection_rate
cancer_death_rate
public_health_spe
hdi
NEW hog_ideology_numeric *(from -1 as leftist to 1 as rightist)*

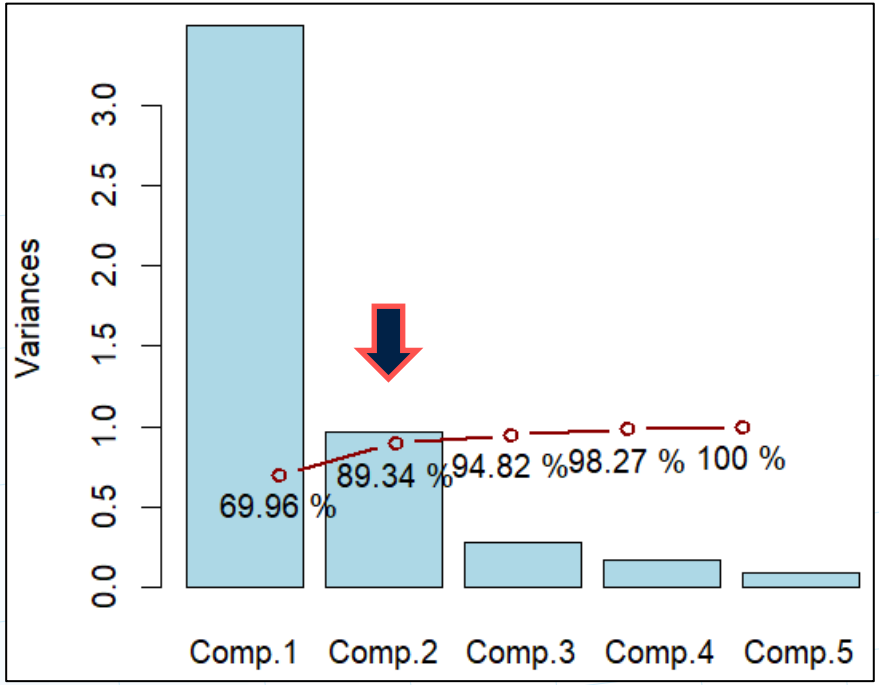
Correlation Heatmap



Principal Components Analysis

Defining the number of components



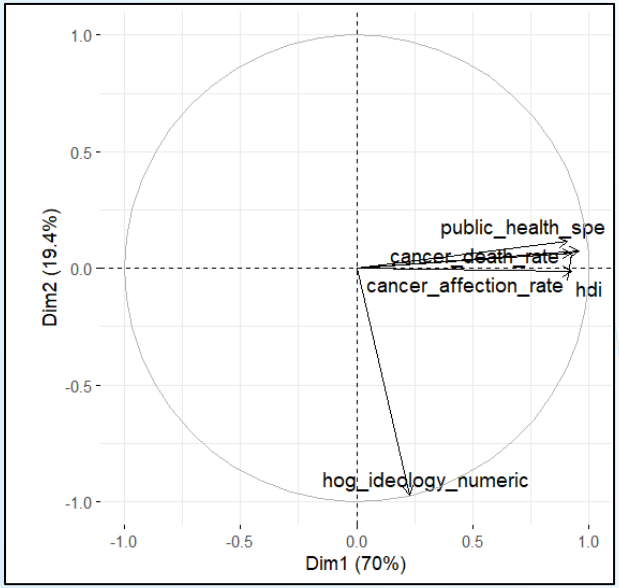



```
> summary(pca_base)
Importance of components:
                          Comp.1    Comp.2     Comp.3    Comp.4    Comp.5
Standard deviation     1.8703178 0.9842269 0.52369751 0.4153300 0.2940247
Proportion of Variance 0.6996177 0.1937405 0.05485182 0.0344998 0.0172901
Cumulative Proportion  0.6996177 0.8933583 0.94821010 0.9827099 1.0000000
```

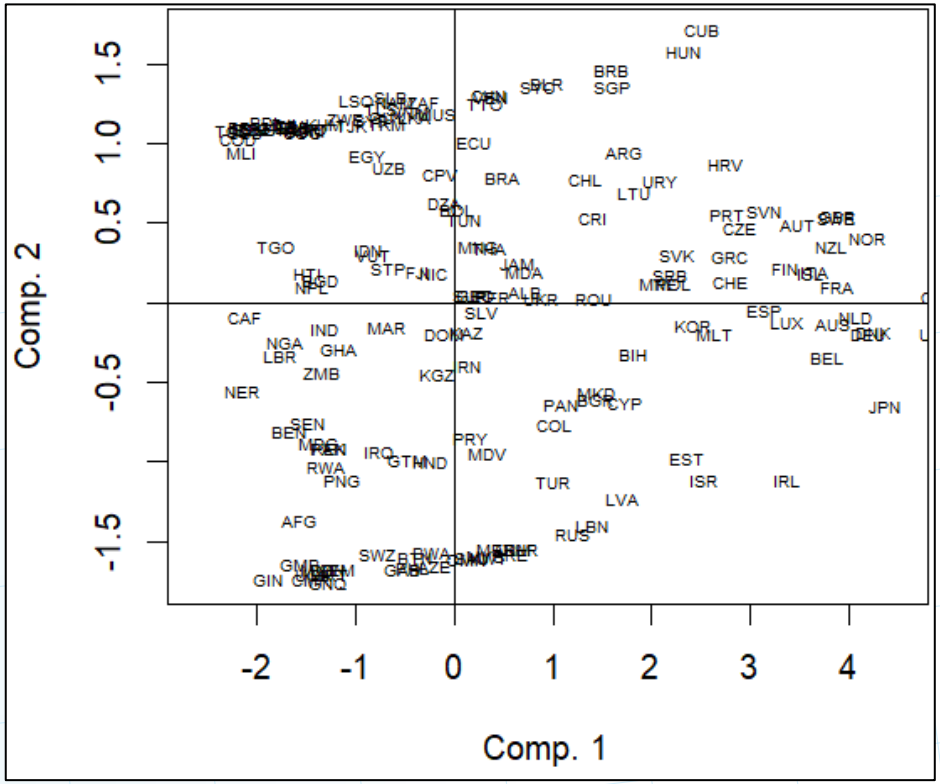
Principal Components Analysis

Loadings

	Comp.1	Comp.2
cancer_affection_rate	0.4945552	0.06707236
cancer_death_rate	0.5104804	0.07395429
public_health_spe	0.4854279	0.11563296
hdi	0.4942418	-0.01293469
hog_ideology_numeric	0.1221055	-0.98817700



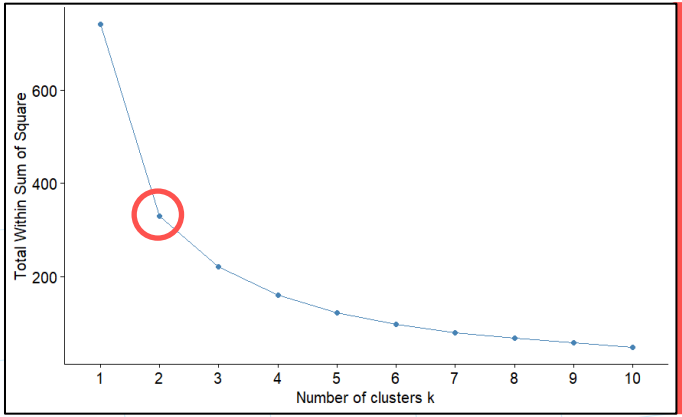
Countries and Components



Clustering

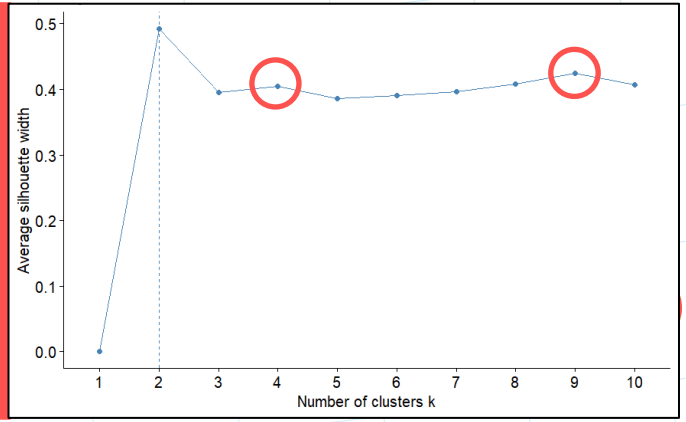
Evaluation of Clustering Results

```
## Clustering Methods:
##  hierarchical kmeans pam
##
## Cluster sizes:
##  2 3 4 5 6
##
## Validation Measures:
##                                  2       3       4       5       6
##
## hierarchical Connectivity   7.2655 16.6750 23.5286 29.9000 34.6440
##              Dunn           0.1239  0.0844  0.0944  0.1002  0.1312
##              Silhouette     0.5003  0.4347  0.3889  0.3940  0.4179
## kmeans       Connectivity  13.0647 29.0516 40.1238 39.9873 47.5468
##              Dunn           0.0367  0.0525  0.0445  0.0485  0.0664
##              Silhouette     0.5038  0.4456  0.4295  0.4109  0.4404
## pam          Connectivity  13.0647 28.7603 44.4929 38.4960 47.8837
##              Dunn           0.0367  0.0211  0.0323  0.0485  0.0586
##              Silhouette     0.5038  0.3753  0.4278  0.4110  0.4320
##
## Optimal Scores:
##
##             Score  Method       Clusters
## Connectivity 7.2655 hierarchical 2
## Dunn         0.1312 hierarchical 6
## Silhouette   0.5038 kmeans       2
```



Elbow method

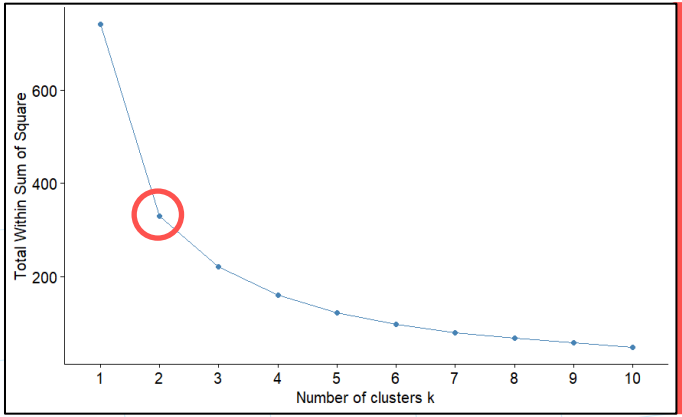
Silhouette method



Clustering

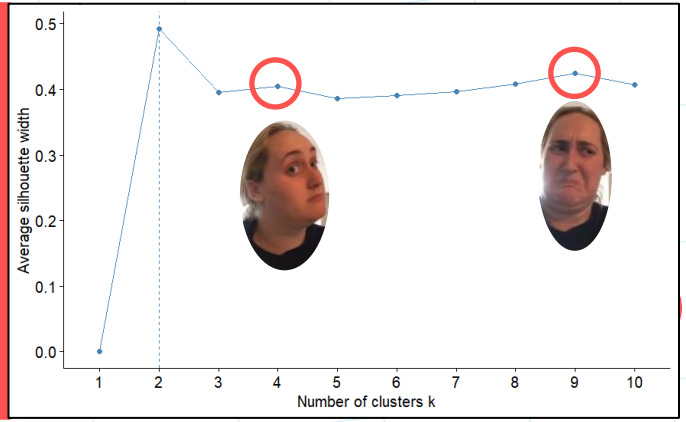
Evaluation of Clustering Results

```
## Clustering Methods:
##  hierarchical kmeans pam
##
## Cluster sizes:
##  2 3 4 5 6
##
## Validation Measures:
##                                  2       3       4       5       6
##
## hierarchical Connectivity   7.2655 16.6750 23.5286 29.9000 34.6440
##              Dunn           0.1239  0.0844  0.0944  0.1002  0.1312
##              Silhouette     0.5003  0.4347  0.3889  0.3940  0.4179
## kmeans       Connectivity  13.0647 29.0516 40.1238 39.9873 47.5468
##              Dunn           0.0367  0.0525  0.0445  0.0485  0.0664
##              Silhouette     0.5038  0.4456  0.4295  0.4109  0.4404
## pam          Connectivity  13.0647 28.7603 44.4929 38.4960 47.8837
##              Dunn           0.0367  0.0211  0.0323  0.0485  0.0586
##              Silhouette     0.5038  0.3753  0.4278  0.4110  0.4320
##
## Optimal Scores:
##
##              Score  Method       Clusters
## Connectivity 7.2655 hierarchical 2
## Dunn         0.1312 hierarchical 6
## Silhouette   0.5038 kmeans       2
```



Elbow method

Silhouette method



Hierarchical Clustering

Linkage Method:

Single

```
## k=2:   0.1864929

## k=6:   0.2355196
```

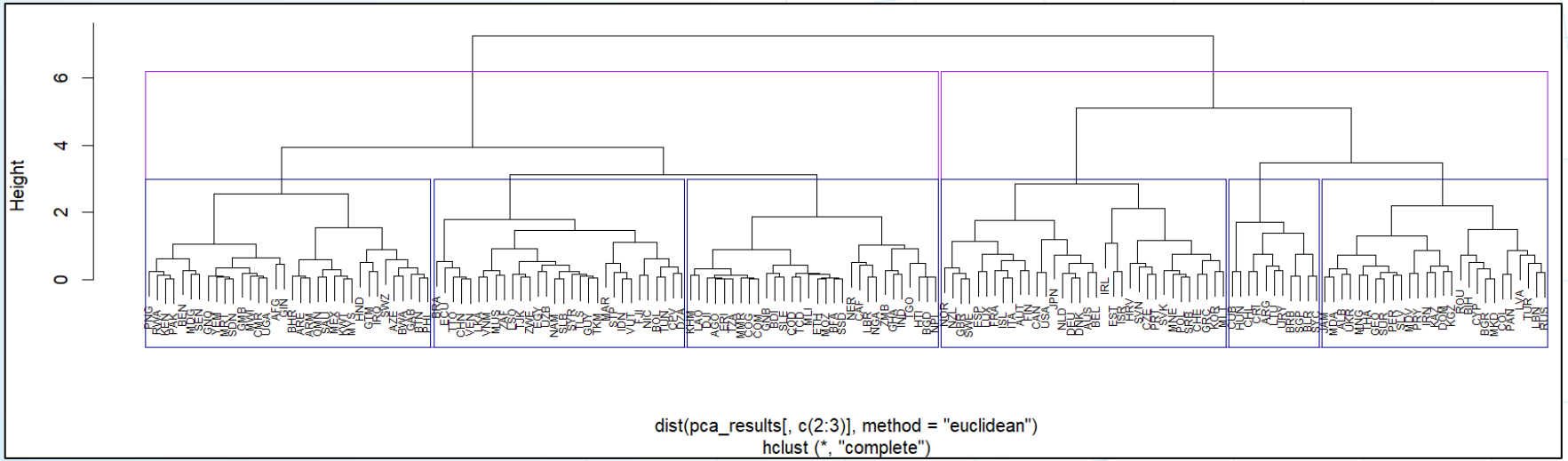
Complete

```
## k=2:   0.4142643

## k=6:   0.3932472
```

Average

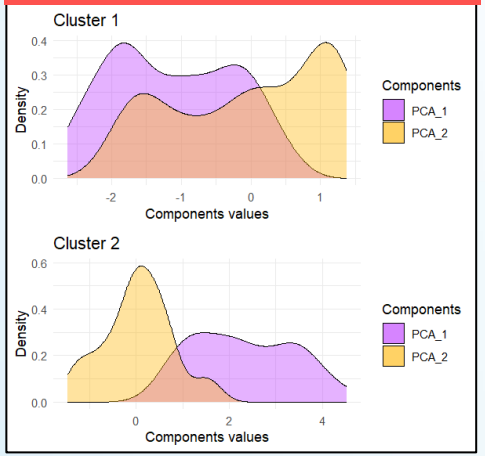
```
## k=2:   0.5002604

## k=6:   0.4178778
```

Dendogram:

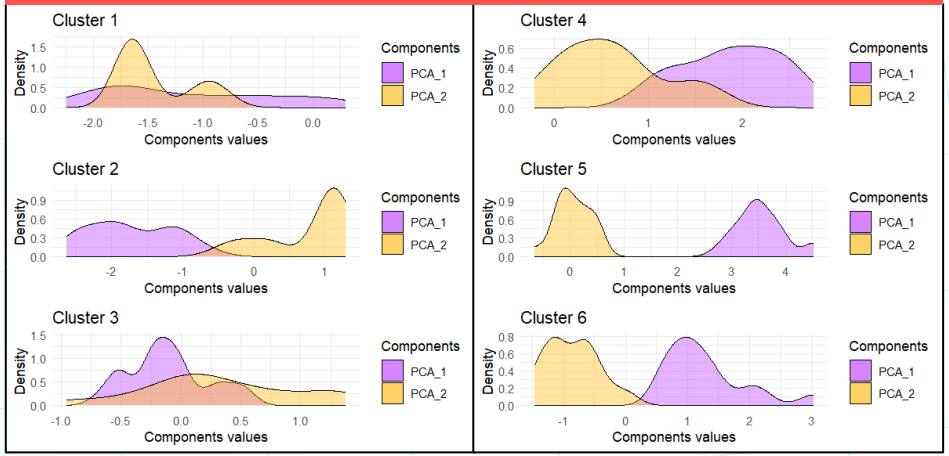


Hierarchical Clustering

Clustering with K=2

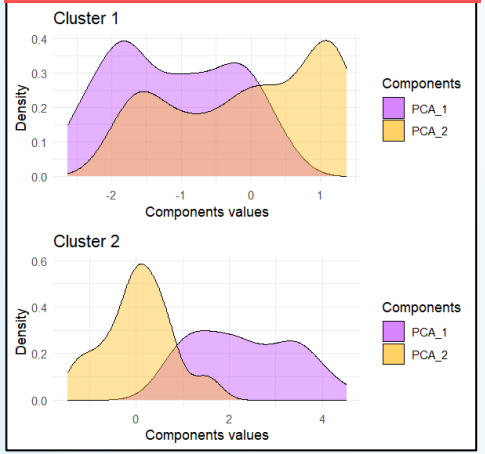


Clustering with K=6

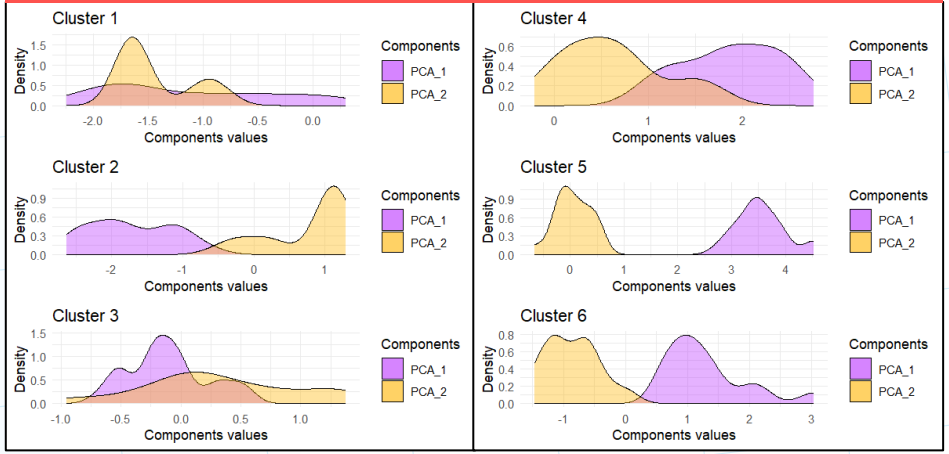


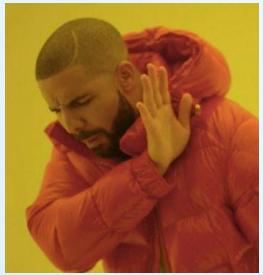
Hierarchical Clustering

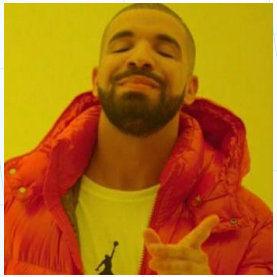
Clustering with K=2



Clustering with K=6







- Better-defined and well separated clusters.
- Clear differences among clusters

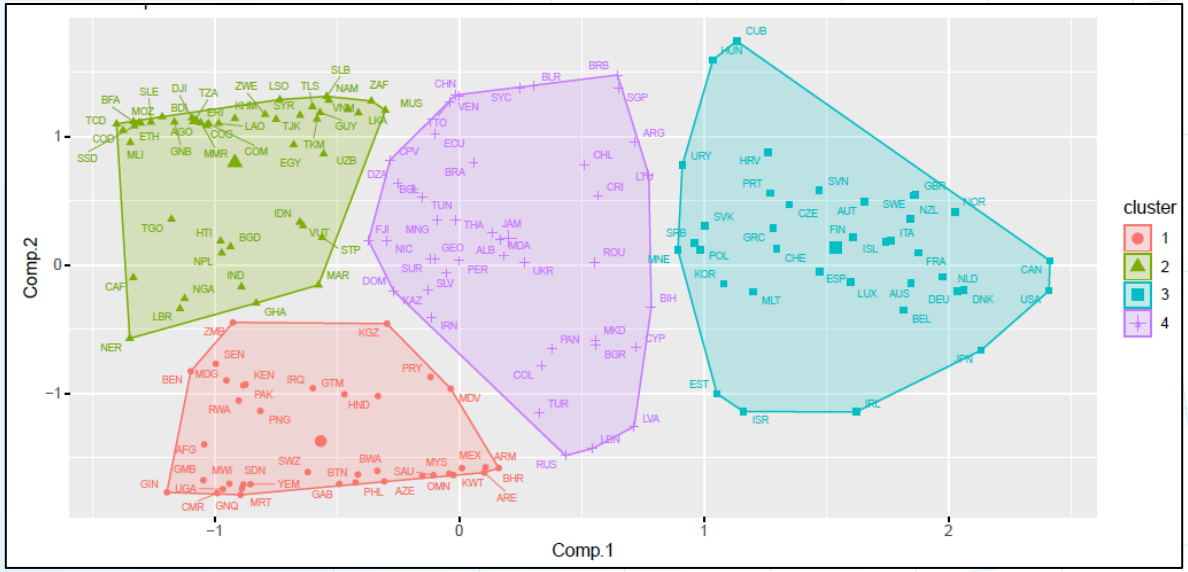
K-Means

Clustering with K=2

```
## k=2:

## WCSS =  83.6088 226.8112

## Between SS / Total SS =  0.581355
```

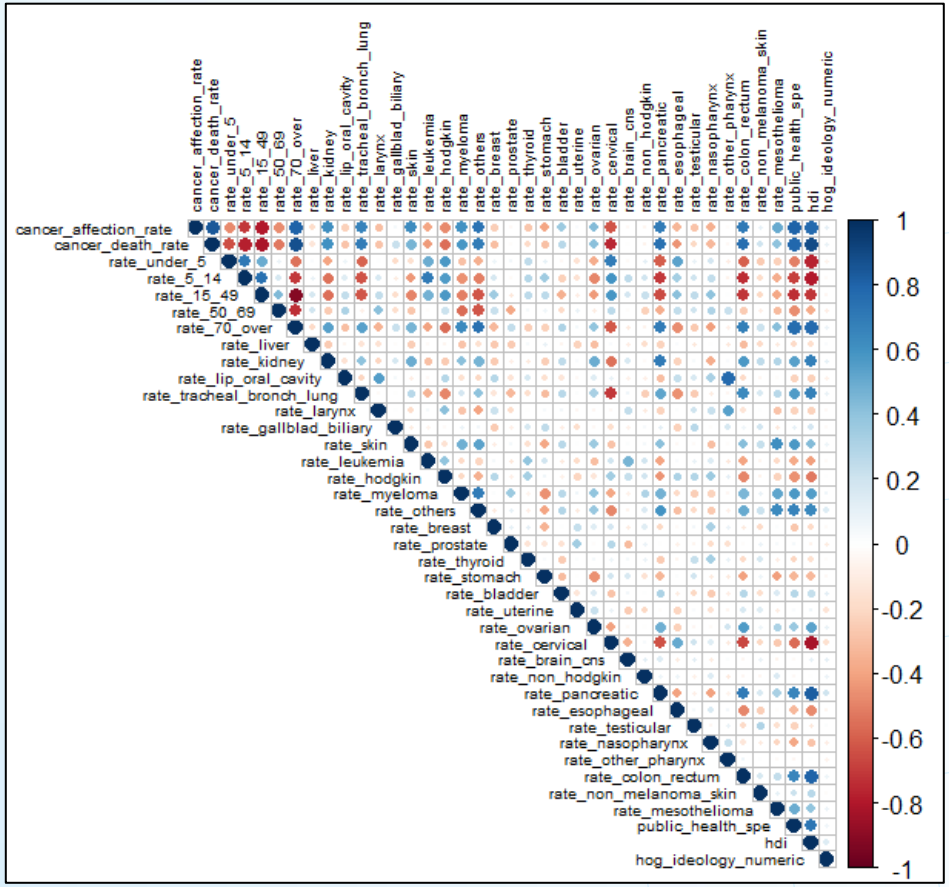
Clustering with K=4

```
## k=4:

## WCSS =  28.30195 30.83475 36.30385 44.36079

## Between SS / Total SS =  0.8114582
```

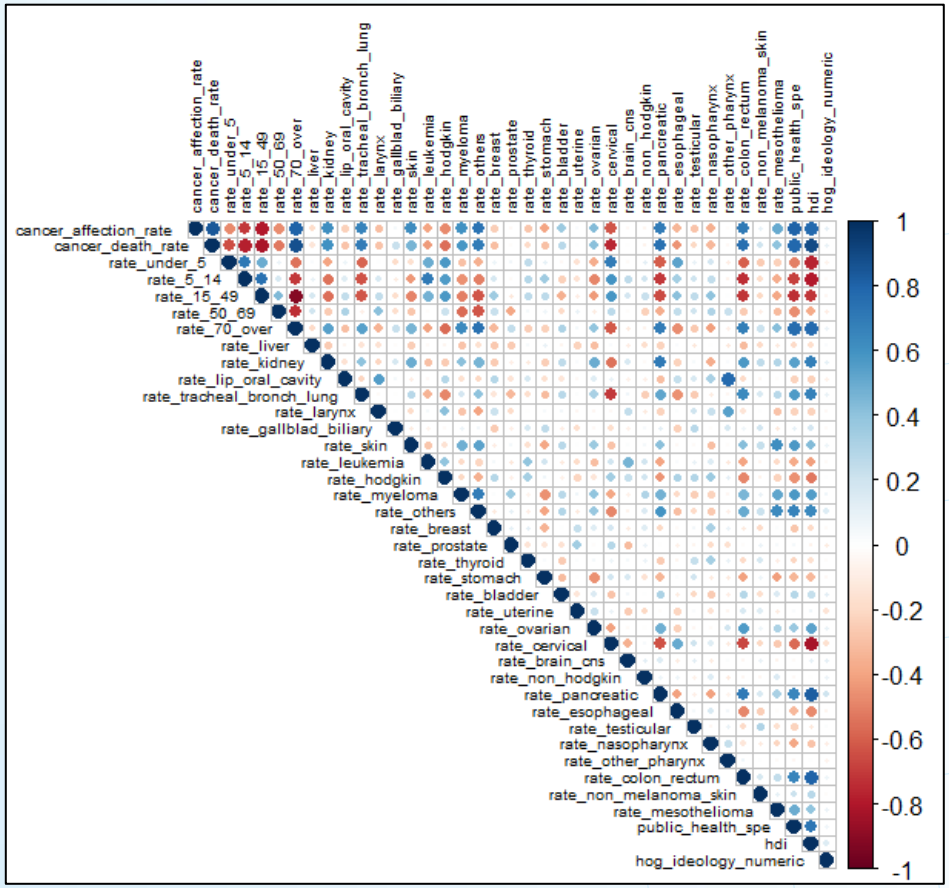


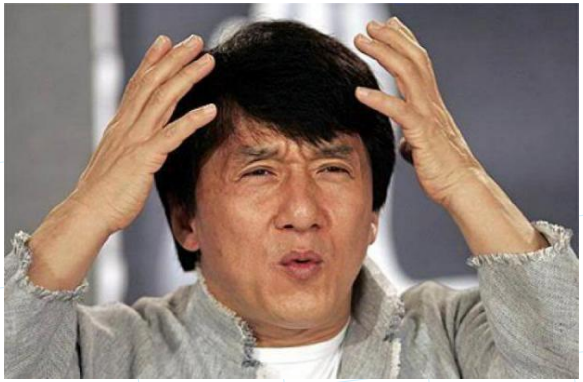
03 Supervised

Variables Used



Variables Used



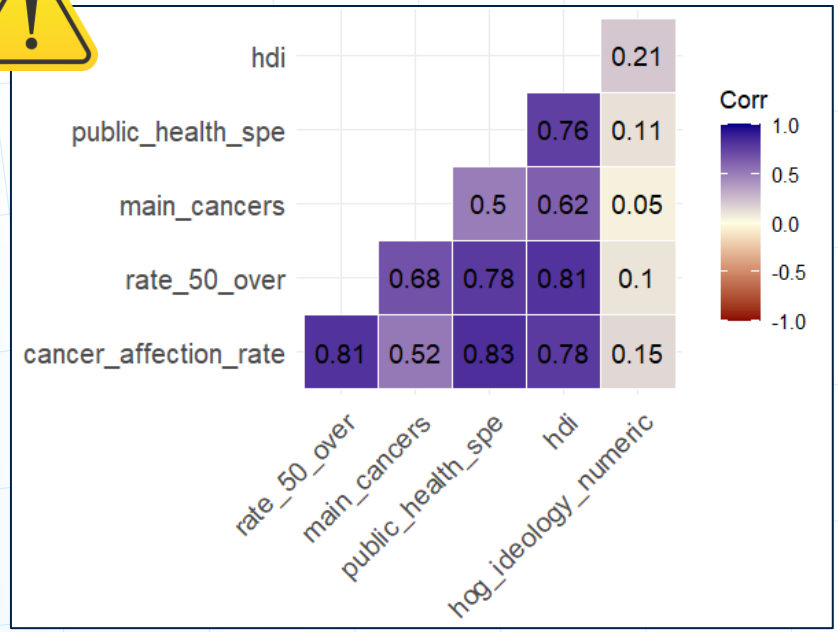


Variables reduction

Variables Used

Variables
cancer_death_rate *(dependent)*
cancer_affection_rate
public_health_spe
hdi
hog_ideology_numeric *(from -1 as leftist to 1 as rightist)*
NEW **main_cancers** *(5 main cancers)*
NEW **rate_50_over** *(main age ranges)*

Correlation Heatmap *(independent variables)*



**Using average by country*

Regression Model

Use PCA for dimensionality reduction and address multicollinearity.

```
> summary(sup_pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6
Standard deviation     1.9492027 0.9992041 0.7810115 0.4974617 0.43684886 0.39232115
Proportion of Variance 0.6332319 0.1664015 0.1016632 0.0412447 0.03180615 0.02565265
Cumulative Proportion  0.6332319 0.7996333 0.9012965 0.9425412 0.97434735 1.00000000
```

```
Loadings:
                      Comp.1 Comp.2
cancer_affection_rate  0.459
rate_50_over           0.473
main_cancers           0.374 -0.164
public_health_spe      0.450
hdi                    0.462
hog_ideology_numeric          0.980
```

Regression Model

Use PCA for dimensionality reduction and address multicollinearity.

```
> summary(sup_pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6
Standard deviation     1.9492027 0.9992041 0.7810115 0.4974617 0.43684886 0.39232115
Proportion of Variance 0.6332319 0.1664015 0.1016632 0.0412447 0.03180615 0.02565265
Cumulative Proportion  0.6332319 0.7996333 0.9012965 0.9425412 0.97434735 1.00000000
```

```
Loadings:
                       Comp.1 Comp.2
cancer_affection_rate   0.459
rate_50_over            0.473
main_cancers            0.374 -0.164
public_health_spe       0.450
hdi                     0.462
hog_ideology_numeric          0.980
```

Linear regression with PCA on the averaged dataset

```
## Call:
## lm(formula = cancer_death_rate ~ Comp.1 + Comp.2, data = df_sup_avg_pca)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.086873 -0.023151 -0.001131  0.018767  0.104107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.157333   0.002729  57.645   <2e-16 ***
## Comp.1       0.043852   0.001386  31.650   <2e-16 ***
## Comp.2      -0.001925   0.002742  -0.702    0.484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03517 on 163 degrees of freedom
## Multiple R-squared:  0.8601, Adjusted R-squared:  0.8584
## F-statistic: 501.1 on 2 and 163 DF,  p-value: < 2.2e-16
```

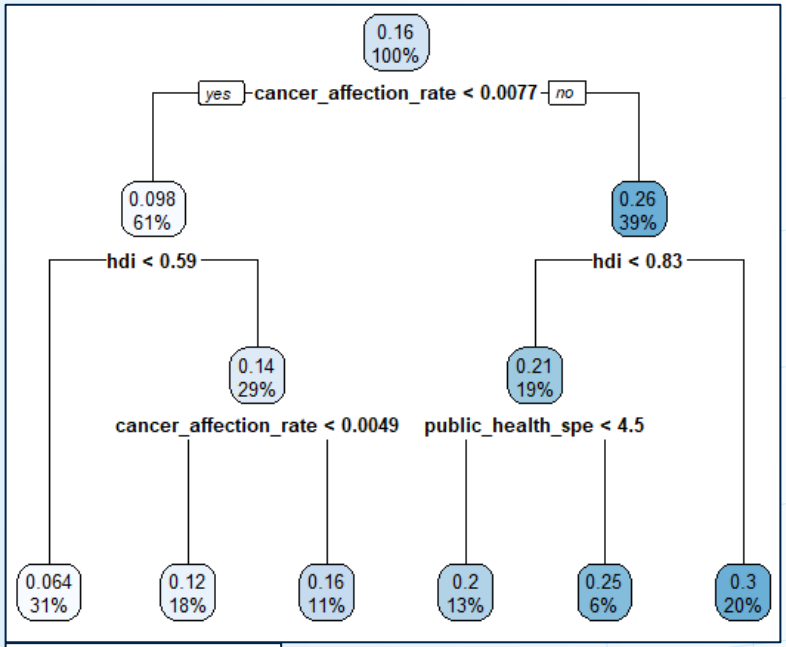
```
## RMSE = 0.03484603

## MAE = 0.026823
```

Ideology is not significant

Decision Tree

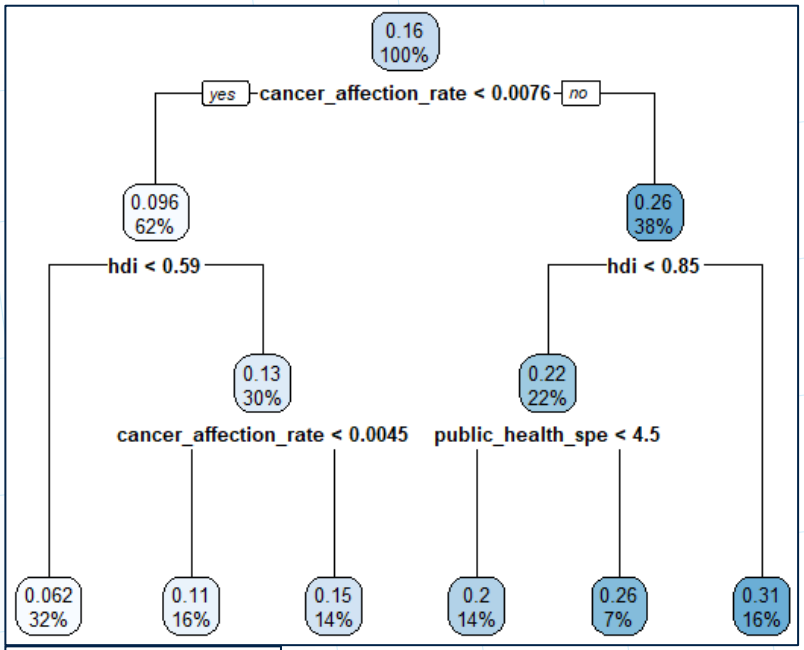
- Averaged tree had better performance.
- Both exclude ideology.

Original dataset




```
## RSME = 0.03336647

## Accuracy =  0.8782363
```

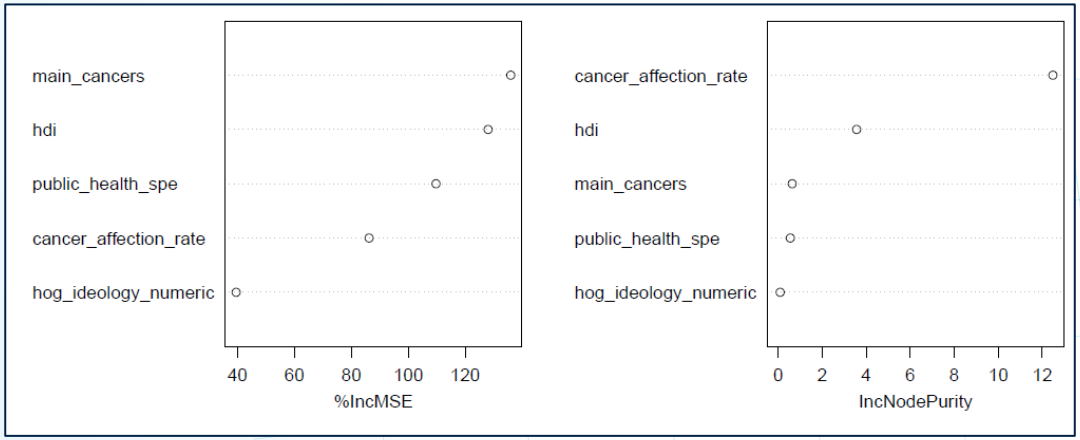
Averaged dataset




```
## RSME = 0.02850167

## Accuracy =  0.9017756
```

Random Forest

- Original dataset outputted better performance.



10 Folds

```
## rf_RMSE =  0.0166421

## rf variable importance
##
##                          Overall
## main_cancers              100.00
## hdi                        91.76
## public_health_spe          72.78
## cancer_affection_rate      48.41
## hog_ideology_numeric        0.00
```

Ranger

```
## rf_RMSE =  0.0157281

## ranger variable importance
##
##                          Overall
## hdi                       100.00
## cancer_affection_rate      85.07
## public_health_spe          25.05
## main_cancers               14.19
## hog_ideology_numeric        0.00
```

Conclusions

Project's goal

The objective was **strongly supported by the results** obtained from both unsupervised and supervised techniques: project aims to illustrate that the occurrence of cancer-related deaths correlates with factors such as investment in public health and the Human Development Index (HDI), rather than the messages conveyed by political parties.

Further analysis

- **New independent variables** (other death causes, income, etc)
- **Different perspective of the ideology variable** (whole mandate)
- **Changing dependent variable** (education level, employment rate, entrepreneurship rate, etc)


“Actions speak louder than words, and are more to be regarded”.
THANK YOU!