UNIVERSITÀ DEGLI STUDI DI MILANO

Facoltà di Scienze Politiche, Economiche e Sociali

Behind the Curtain:

Political Ideologies and the Impact of Cancer

Course: Machine Learning and Statistical Learning

Professor: Salini, Silvia

Student: Bonifacini, Franco Reinaldo

Contents

1 Abstract **3**

1.1 *Project's goal* . . . 3

2 Data used **3**

2.1 *Libraries* . . . 3

2.2 *Datasets* . . . 4

2.3 *Data cleaning and processing* . . . 4

2.4 *Exploratory Data Analysis* . . . 6

3 Unsupervised Learning **10**

3.1 *Principal Component Analysis (PCA)* . . . 11

3.2 *Clustering* . . . 13

3.2.1 *Hierarchical Clustering* . . . 15

3.2.2 *Dendogram* . . . 16

3.3 *K-Means* . . . 17

4 Supervised Learning **20**

4.1 *Regression model* . . . 21

4.2 *Decision Tree* . . . 22

4.3 *Random Forest* . . . 26

5 Conclusion **29**

1 Abstract

In our world, every political decision directly impacts the population. It's crucial to ascertain whether political parties truly represent their proclaimed ideologies. As people will never know what happens "behind the curtain", **this project may shed light on a more precise understanding of how policy decisions affect the population and their alignment with the ruling party's ideology**.

Various datasets were utilized for analysis, employing cancer as a variable to assess the effectiveness of public health investment and its correlation with the ideology of the governing party.

Keywords: ideology, spending, health, cancer, hdi, unsupervised, supervised, cluster, regression

1.1 *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 than the messages conveyed by political parties**.

In other words, **it is easy to preach with the word, but the world truly requires preaching thorugh the example**.

For this project, a combination of **unsupervised and supervised learning techniques** was employed to analyze and comprehend how the variables fluctuate across different countries and ideologies.

---

2 Data used

2.1 *Libraries*

- **Data manipulation and transformation:**

```
library(dplyr)
library(tidyr)
library(outliers)
```

- **Reading and writing data:**

```
library(openxlsx)
library(readxl)
```

- **Data visualization:**

```
library(ggplot2)
library(ggthemes)
library(ggcorrplot)
library(rnaturalearth)
library(gridExtra)
library(hrbrthemes)
library(viridis)
library(factoextra)
library(ggpubr)
```

- **Data analysis and statistics:**

```
library(proxy)
library(cluster)
library(clValid)
library(corrplot)
library(car)
```

- **Modeling and machine learning:**

```
library(caret)
library(randomForest)
library(rpart)
library(rpart.plot)
```

2.2 *Datasets*

To execute this project, several datasets were used:

1. **death_cause:** The estimated annual number of deaths from each cause. *(Causes of death, 1990 to 2019 - Our World in Data).*
2. **population:** Population by country, available from 10,000 BCE to 2100, based on data and estimates from different sources. *(Population, 10,000 BCE to 2021 - Our World in Data).*
3. **n_cancer:** Total number of people suffering from any type of cancer at a given time. This is measured across both sexes, and all ages. *(Number of people with cancer, 1990 to 2017 - Our World in Data).*
4. **d_cancer_type:** Total annual number of deaths from cancers across all ages and both sexes, broken down by cancer type. *(Cancer deaths by type, 1990 to 2019 - Our World in Data).*
5. **d_cancer_age:** Total annual cancer deaths differentiated by age category across both sexes. Data includes all forms of cancer. *(Deaths from cancer, by age, 1990 to 2019 - Our World in Data).*
6. **health_spe:** This metric captures spending on government funded health care systems and social health insurance, as well as compulsory health insurance. *(Government health expenditure as a share of GDP, 1880 to 2021 - Our World in Data).*
7. **hd_index:** The Human Development Index (HDI) is a summary measure of key dimensions of human development: a long and healthy life, a good education, and a decent standard of living. Higher values indicate higher human development. *(Human Development Index, 1990 to 2022 - Our World in Data).*
8. **ideology:** Distinguishes between chief executives with leftist, centrist, rightist, and no discernible economic ideology. *(Identifying Ideologues: A Global Dataset on Chief Executives, 1945-2020 - Bastian Herre)*

2.3 *Data cleaning and processing*

Regarding the datasets mentioned earlier, apart from standardizing all of them with uniform column names, it was crucial to specify the range of years utilized. **It's essential to emphasize that incorporating data from multiple years does not signify that this project engaged in a time series analysis**. Rather, each entry was treated as an independent sample, regardless of the year.

Therefore, the final dataset's range must be from the highest minimum year until the lowest maximum year, across all datasets. So, **the final range utilized is from 2000 to 2017**:

```
##         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
```

Following this, the subsequent step involved normalizing countries' names and their corresponding abbreviations (codes), thereby facilitating data integration. Consequently, **countries lacking a corresponding code were subsequently removed from the dataset**.

Additionally, **all datasets were merged into a unified dataset (df_project)** , consolidating the information necessary for the project's objectives.

Finally, some additional variables were calculated:

1. **cancer_affection_rate:** people with cancer / population.
2. **cancer_death_rate:** deaths by cancer / total deaths.
3. **Rate of deaths by cancer, by age:** deaths by range of age / total deaths by cancer.
4. **Rate of deaths by cancer, by type:** deaths by type of cancer / total deaths by cancer.

As a result, **df_project** contains the following information:

```
## Countries included: 166

## 'data.frame':    2782 obs. of  42 variables:
##  $ entity                    : Factor w/ 205 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ code                      : chr  "AFG" "AFG" "AFG" "AFG" ...
##  $ year                      : int  2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 ...
##  $ hog_ideology              : chr  "rightist" "rightist" "rightist" "rightist" ...
##  $ cancer_affection_rate     : num  0.00263 0.00261 0.00261 0.0026 0.00253 ...
##  $ cancer_death_rate         : num  0.0661 0.0702 0.0732 0.0746 0.0756 ...
##  $ rate_under_5              : num  0.0396 0.0455 0.0465 0.0451 0.0432 ...
##  $ rate_5_14                 : num  0.0393 0.047 0.0499 0.0505 0.0498 ...
##  $ rate_15_49                : num  0.223 0.228 0.237 0.247 0.256 ...
##  $ rate_50_69                : num  0.42 0.406 0.396 0.386 0.378 ...
##  $ rate_70_over              : num  0.279 0.273 0.27 0.271 0.273 ...
##  $ rate_liver                : num  0.0793 0.0781 0.0775 0.0772 0.0759 ...
##  $ rate_kidney               : num  0.00584 0.00603 0.0061 0.00617 0.00625 ...
##  $ rate_lip_oral_cavity      : num  0.00747 0.00731 0.0072 0.00719 0.00718 ...
##  $ rate_tracheal_bronch_lung: num  0.0785 0.0772 0.0765 0.0763 0.0764 ...
##  $ rate_larynx               : num  0.0245 0.024 0.0236 0.0234 0.0233 ...
##  $ rate_gallblad_biliary     : num  0.0167 0.0163 0.0161 0.016 0.0159 ...
##  $ rate_skin                 : num  0.0049 0.00483 0.00478 0.00479 0.00481 ...
##  $ rate_leukemia             : num  0.114 0.121 0.124 0.124 0.123 ...
##  $ rate_hodgkin              : num  0.0122 0.0124 0.0126 0.0127 0.0127 ...
##  $ rate_myeloma              : num  0.00973 0.0095 0.00941 0.00929 0.00934 ...
##  $ rate_others               : num  0.00125 0.00121 0.00125 0.00123 0.00122 ...
##  $ rate_breast               : num  0.0544 0.0544 0.055 0.0557 0.0568 ...
##  $ rate_prostate             : num  0.0307 0.03 0.0296 0.0295 0.0294 ...
##  $ rate_thyroid              : num  0.00592 0.00588 0.00595 0.00595 0.00603 ...
##  $ rate_stomach              : num  0.228 0.224 0.222 0.221 0.22 ...
##  $ rate_bladder              : num  0.0221 0.0216 0.0212 0.0211 0.0209 ...
##  $ rate_uterine              : num  0.00568 0.00566 0.00566 0.00574 0.00582 ...
##  $ rate_ovarian              : num  0.00786 0.00792 0.00809 0.00828 0.00847 ...
##  $ rate_cervical             : num  0.0355 0.035 0.0349 0.0348 0.0348 ...
##  $ rate_brain_cns            : num  0.0495 0.053 0.0544 0.0547 0.0546 ...
##  $ rate_non_hodgkin          : num  0.0877 0.0878 0.0884 0.0892 0.0903 ...
##  $ rate_pancreatic           : num  0.0115 0.0115 0.0115 0.0117 0.012 ...
##  $ rate_esophageal           : num  0.0497 0.0483 0.0475 0.0469 0.0466 ...
##  $ rate_testicular           : num  0.000545 0.000603 0.000662 0.000726 0.000718 ...
##  $ rate_nasopharynx          : num  0.00584 0.00566 0.00559 0.00552 0.00546 ...
##  $ rate_other_pharynx        : num  0.00304 0.00294 0.00287 0.00283 0.0028 ...
##  $ rate_colon_rectum         : num  0.0441 0.0441 0.0443 0.0449 0.0456 ...
##  $ rate_non_melanoma_skin    : num  0.00249 0.00249 0.0025 0.00254 0.00259 ...
##  $ rate_mesothelioma         : num  0.000778 0.000754 0.000809 0.000799 0.000862 ...
##  $ public_health_spe         : num  0.0842 0.651 0.5429 0.5292 0.4978 ...
##  $ hdi                       : num  0.362 0.376 0.392 0.4 0.409 0.424 0.43 0.44 0.448 0.456 ...
```

2.4 *Exploratory Data Analysis*

Before proceeding with unsupervised and supervised techniques, **exploratory data analysis (EDA)** was undertaken to identify **trends and potential outcomes**.

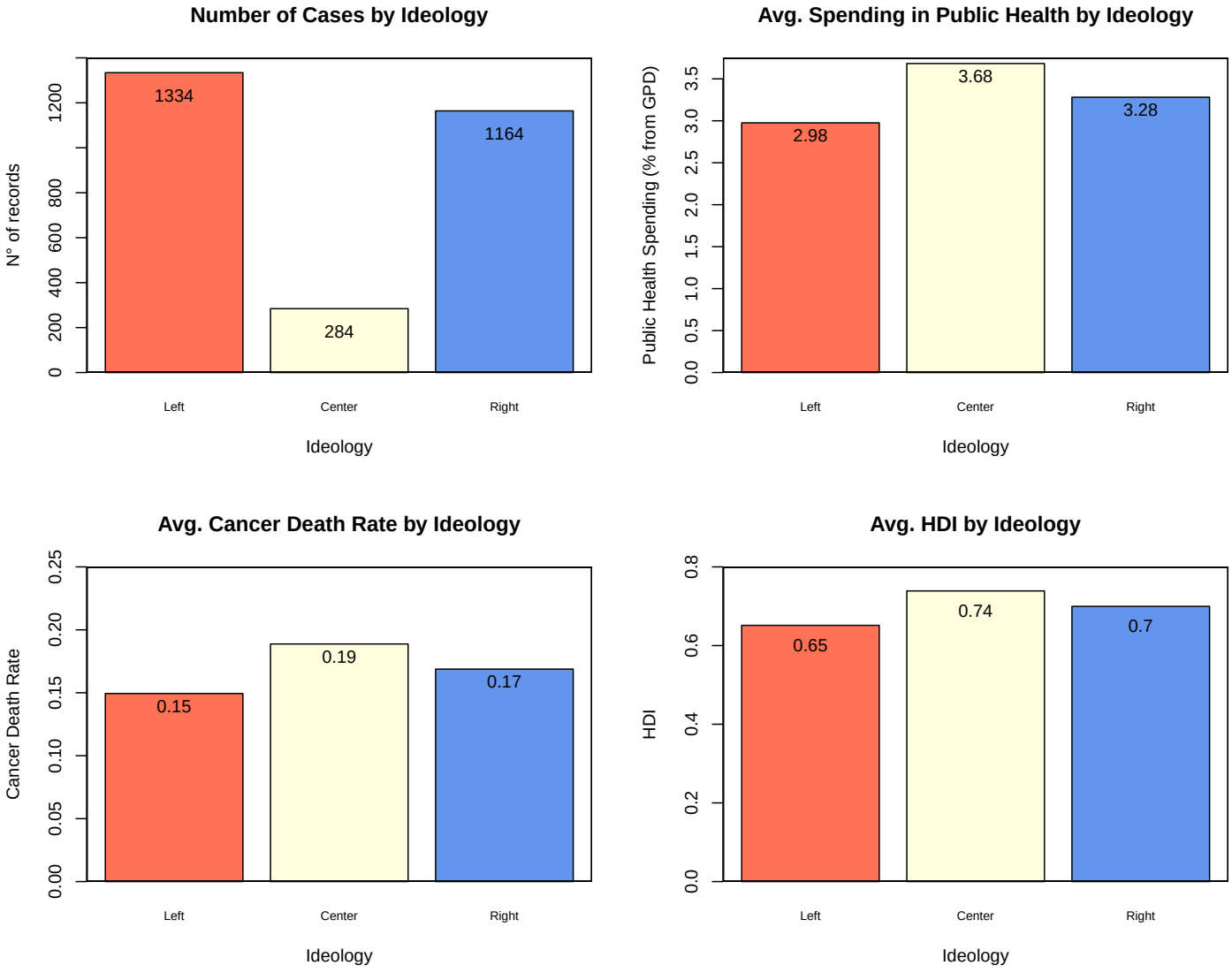

Figure 1: Each barplot compares the different variables among ideologies, where the red bar represents the leftists parties, the white bar represents the centrists parties, and the blue bar represents the rightists parties.

It is pertinent to note in ***figure 1*** that there is a comparable number of records for both left-wing and right-wing parties, whereas centrist parties exhibit significantly fewer records in the dataset.

Additionally, it's clear that **there aren't significant differences between ideologies** regarding public health spending, cancer-related deaths, and HDI. This **aligns with the project's goal** that health investment and improvement depend on factors beyond the governing political party.

On the other hand, both cancer by age and by type were analyzed to decide some manipulation before implementing the techniques:

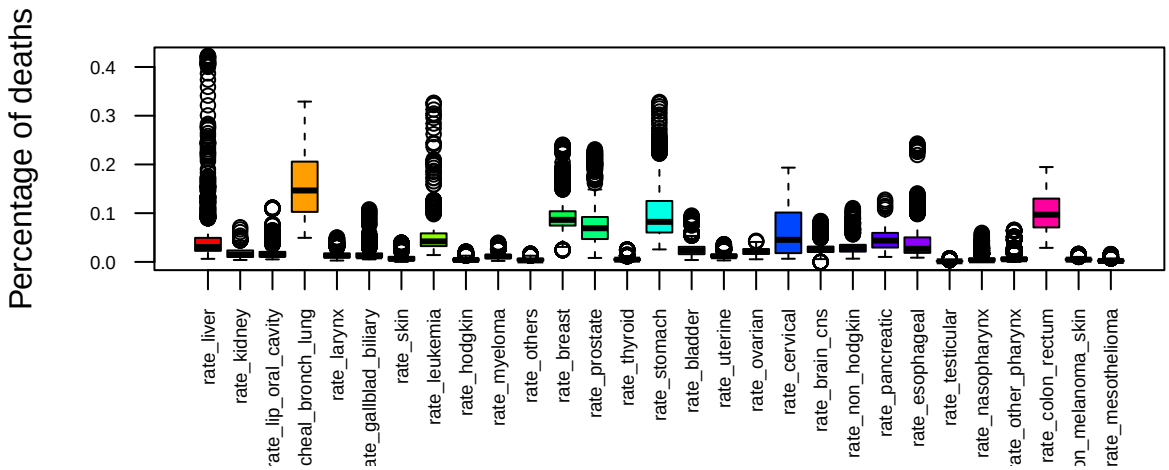

Figure 2: Type of cancers. This boxplot displays how cancer deaths are distributed across different types.

According to the **World Health Organization (WHO)**, and aligned with the results of ***figure 2***, the main cancer deaths are from **lungs, breast, rectal, prostate and stomach**.

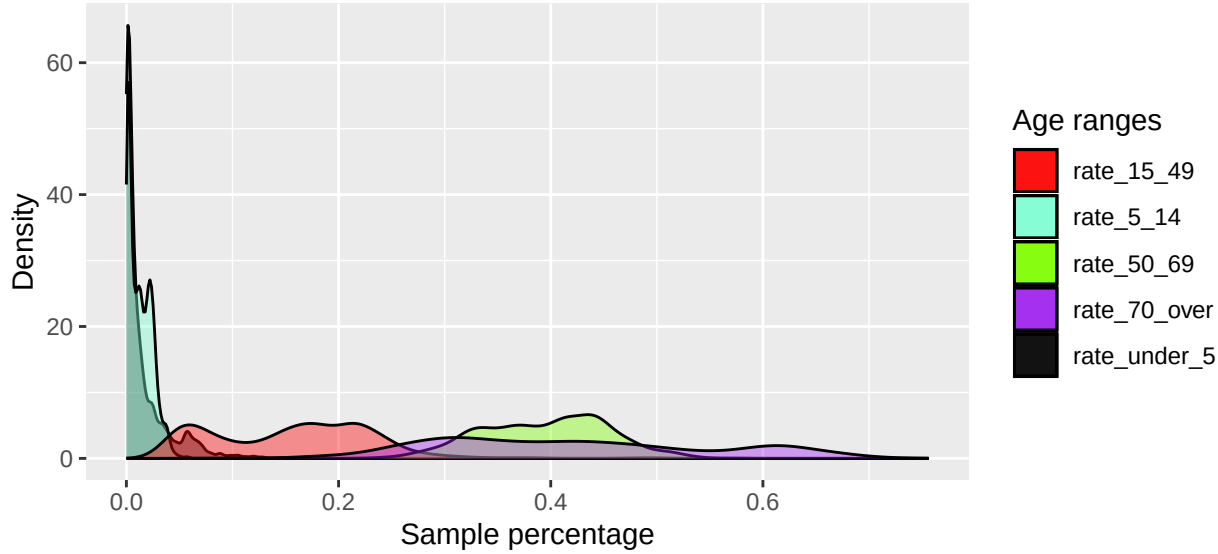

Figure 3: Death from cancer by range of age. This density plot displays how cancer deaths are distributed across range of age.

Considering the findings of ***figure 3***, it is evident that there are **elevated rates of cancer-related deaths among individuals aged 50 and above**, ranging from 25% to 55%. In contrast, the age group of **15 to 49 years exhibits lower levels of mortality**, ranging from 5% to 20%. In summary, **younger age groups tend to have a lower likelihood of dying from cancer**.

These findings hold significance as they could be used in the future, for the implementation of the unsupervised and supervised techniques.

Last but not least, some **maps were plotted** to add more detail to this first approach. In this case, a map was filled with the different variables (ideology, cancer deaths, public health spending and HDI):

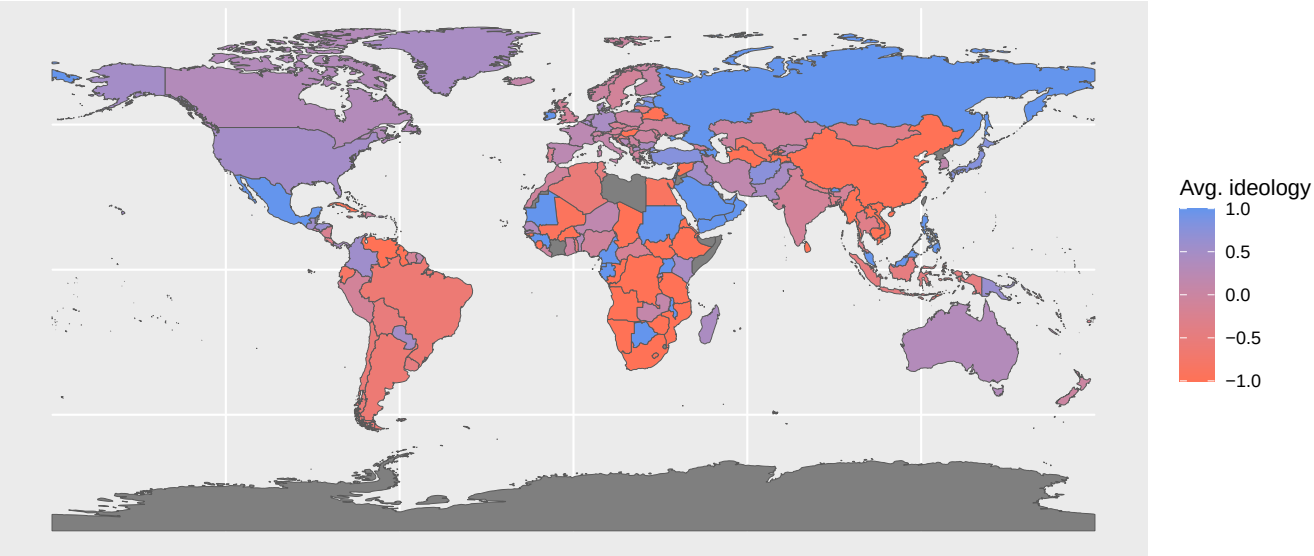

Figure 4: Anual average ideology. The red color represents leftists parties, while the blue color represents rightists parties.

In ***figure 4***, it is evident that **South America, China, and Africa** are regions primarily governed by **left-leaning administrations**, whereas **North America, Europe, and Australia** are predominantly governed by **right-leaning administrations**. In the case of Asia, there appears to be a mixture of ideologies.

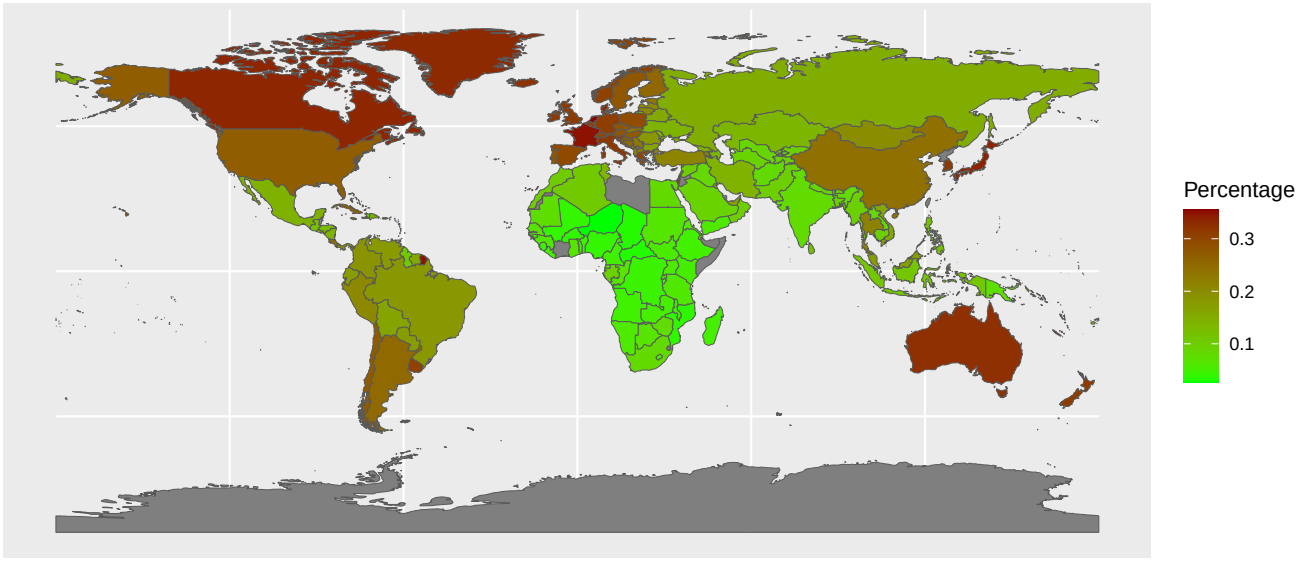

Figure 5: Anual average death by cancer (% from total deaths). The green color represents a low percentage of deaths by cancer, while the darkred color represents a high percentage of deaths by cancer.

Regarding the ***figure 5***, it is evident that, on average, there is a prevalence of **more than 2% of deaths attributed to cancer**, with the exception of Africa. The low incidence of cancer-related deaths in Africa presents a unique case warranting further investigation. It is likely that the causes of death in this continent are predominantly represented by other factors, such as communicable diseases, maternal, perinatal, and nutritional conditions, as indicated by the World Health Organization (WHO).

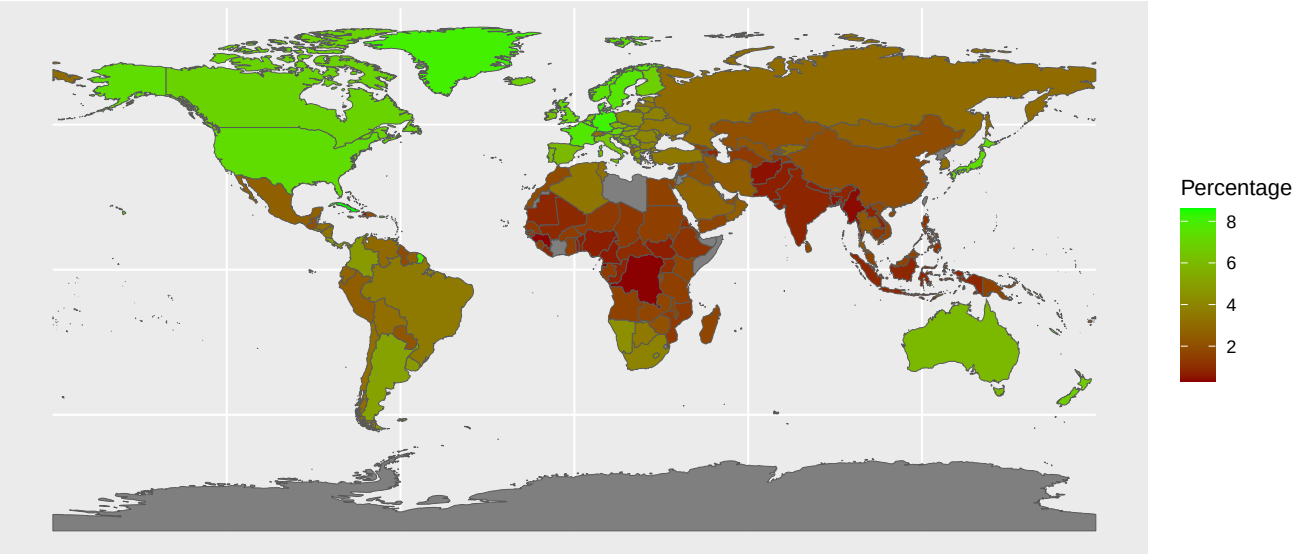

Figure 6: Anual average public health spending (% from GDP). The green color represents a high level of public health spending, while the darkred color represents a low level of public health spending.

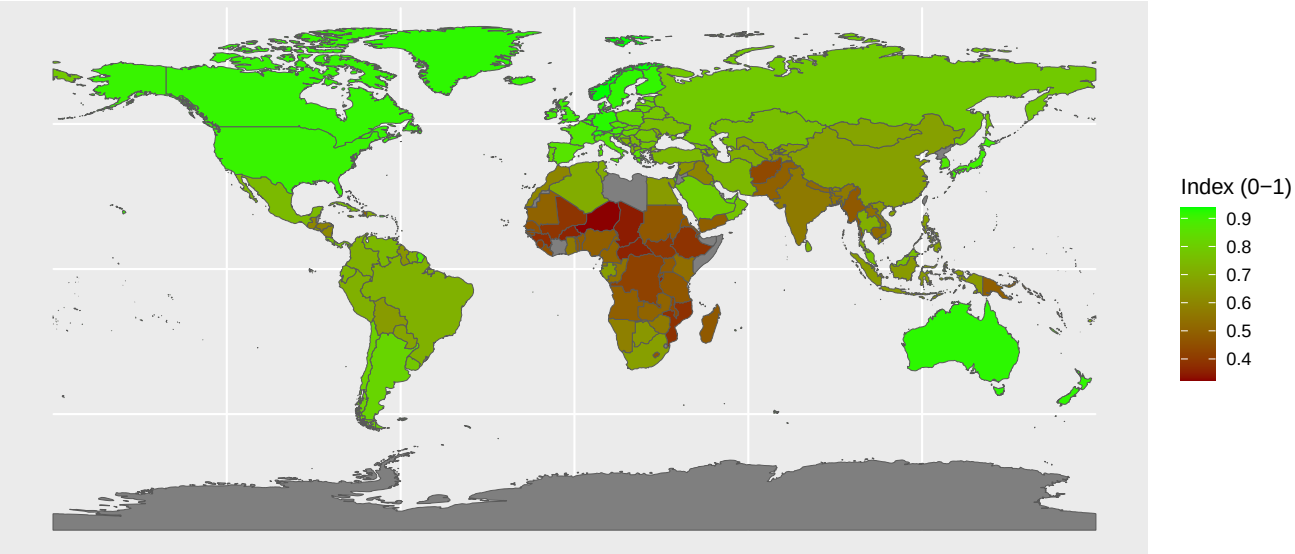

Figure 7: Anual average HDI. The green color represents a high level of HDI, while the darkred color represents a low level of HDI.

Finally, considering ***figure 6*** and ***figure 7***, it is evident that both metrics are notably higher in North America, Europe, and Australia. In this instance, this observation may seemingly **contradict the project's objective**, as these regions were predominantly governed by right-leaning administrations.

---

3 Unsupervised Learning

In this section, all **unsupervised learning techniques** and their corresponding results are presented to provide further clarity on how the original hypothesis is supported.

Before applying any method, some manipulation was done:

1. A new column was added to the dataset to assign numerical values to the ideologies (***hog_ideology_numeric***). **Leftist-led governments were represented with -1, centrists with 0, and rightist with 1**.
2. For the unsupervised learning were only used the following variables: **cancer_affection_rate, cancer_death_rate, public_health_spe, hdi, hog_ideology**.

After this modifications, the new dataset was called **df_cluster** and variables were **standardized**:

```
##  cancer_affection_rate cancer_death_rate public_health_spe      hdi
##  Min.   :-0.8062       Min.   :-1.3998   Min.   :-1.3387   Min.   :-2.1284
##  1st Qu.:-0.6977       1st Qu.:-0.8350   1st Qu.:-0.7774   1st Qu.:-0.9027
##  Median :-0.5346       Median :-0.2193   Median :-0.2491   Median : 0.1882
##  Mean   : 0.0000       Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.6234       3rd Qu.: 0.8404   3rd Qu.: 0.6077   3rd Qu.: 0.8182
##  Max.   : 4.2448       Max.   : 2.1080   Max.   : 2.6874   Max.   : 1.6160
##  hog_ideology_numeric
##  Min.   :-1.32102
##  1st Qu.:-1.10330
##  Median : 0.01169
##  Mean   : 0.00000
##  3rd Qu.: 0.77967
##  Max.   : 1.50607
```

Taking into account this updated dataset, a heatmap was generated to visualize the **correlations between each variable** and to identify any potential multicollinearity that needs to be addressed.

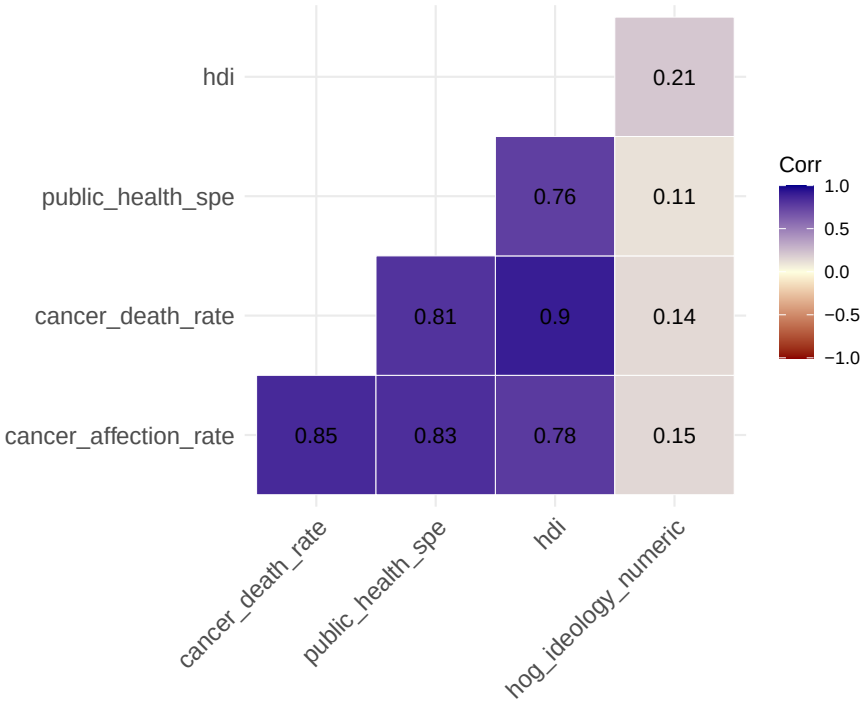

Figure 8: Correlation Heatmap. This heatmap illustrates the correlation levels between all variables in the dataset, ranging from -1 (indicating negative correlation) to 1 (indicating positive correlation).

Regarding the results in ***figure 8***, it's evident that there is a **high correlation among all variables**, except for the ideology (hog_ideology_numeric). To mitigate this issue, a **Principal Component Analysis (PCA)** technique was employed to **reduce the dimensionality of the dataset, thus mitigating the multicollinearity** that may affect the analysis results.

3.1 *Principal Component Analysis (PCA)*

To understand this technique, some key points of PCA need to be explained:

- Firstly, **principal components** are new variables derived from a transformation of the original data, resulting in uncorrelated (orthogonal) components.
- Secondly, each principal component possesses an **eigenvalue**, which denotes a value associated with the variance and indicates how much of the dataset's variance is accounted for by that component.
- Thirdly, using this values is helpful to determine the number of components suitable for analysis. Typically, **those are the principal components that collectively explain nearly 80% of the variance**.

```
## 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
```

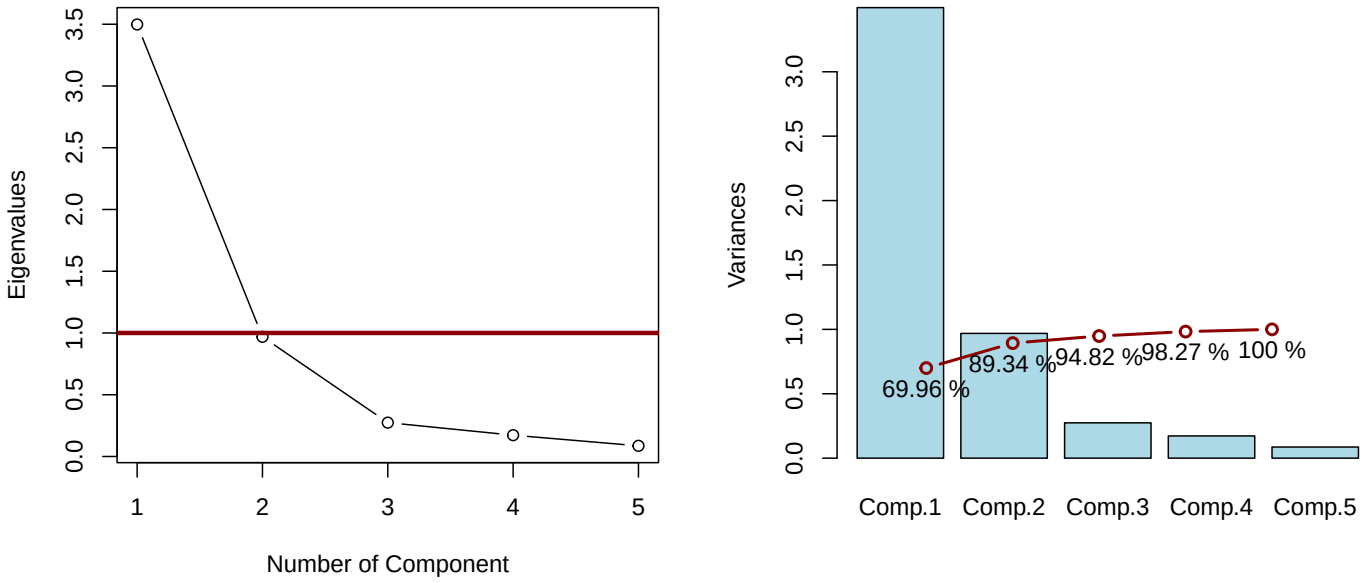

Figure 9: Principal Components. These plots show how much of the dataset's variance is accounted for by each component.

As mentioned before with the PCA's theoretic explanation, the principal components to use are those that explain at least the 80% of the variance, so in this case the components used were:

- **Component 1** which explains 69,96% of the variance.
- **Component 2** which explains 19,37% of the variance, and combined with the first one, they explain the 89,34% of the dataset's variance.

In addition to the table, ***figure 9*** also supports the idea that the **optimal number of principal components to use are 2**, so for the project, the component 1 and 2 were used for the unsupervised techniques.

Furthermore, concerning PCA, **each component represents the original variables of the dataset**. This is a crucial point because the components were utilized instead of the original variables for the techniques. Thus, understanding which variables are explained by each component is a key step in the process.

```
##                           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
```

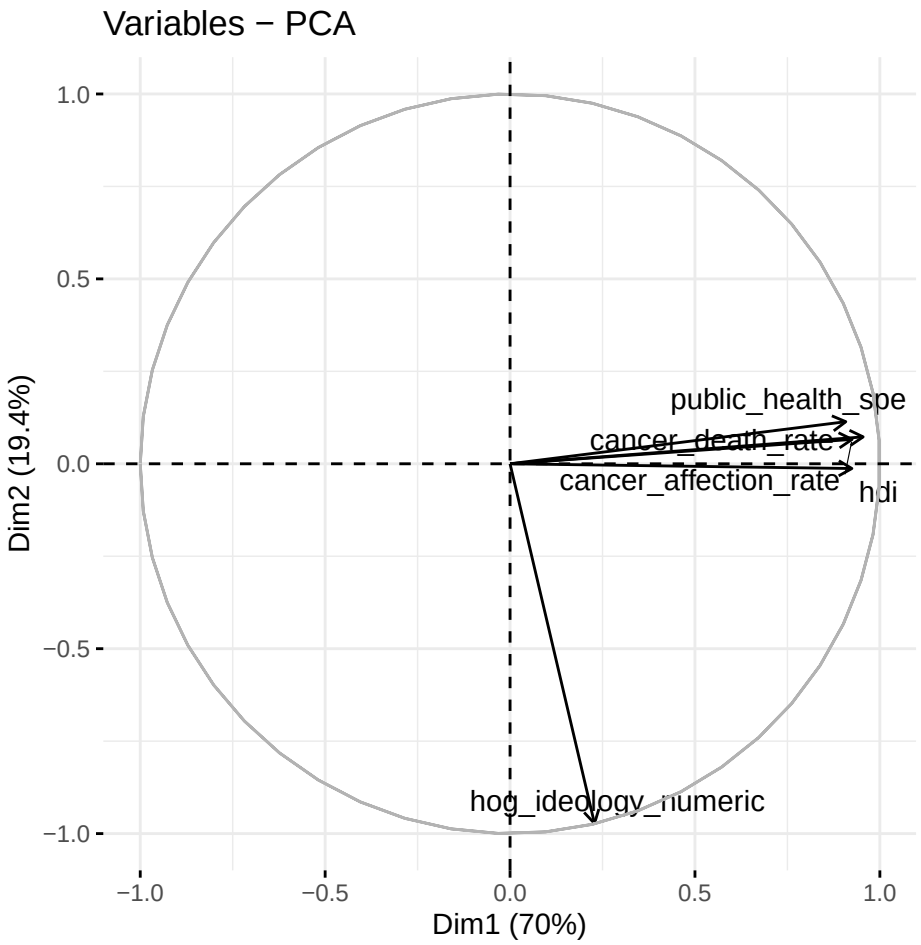

Figure 10: Variables and Principal Components. This plot illustrates how variables are affected by each component.

Now, with the table with the loadings and the ***figure 10*** we can easily conclude that:

- **Component 1** exhibits a positive correlation with **cancer affection rate, cancer death rate, public health spending, and HDI**. Therefore, this component can effectively explain these variables.
- **Component 2** shows a negative correlation with **hog_ideology_numeric**. Hence, this variable is inversely explained by this component.

To summarize the PCA analysis, all countries were plotted using both components to understand how each country is distributed among the two components. This provides an initial insight into how the final clustering might appear.

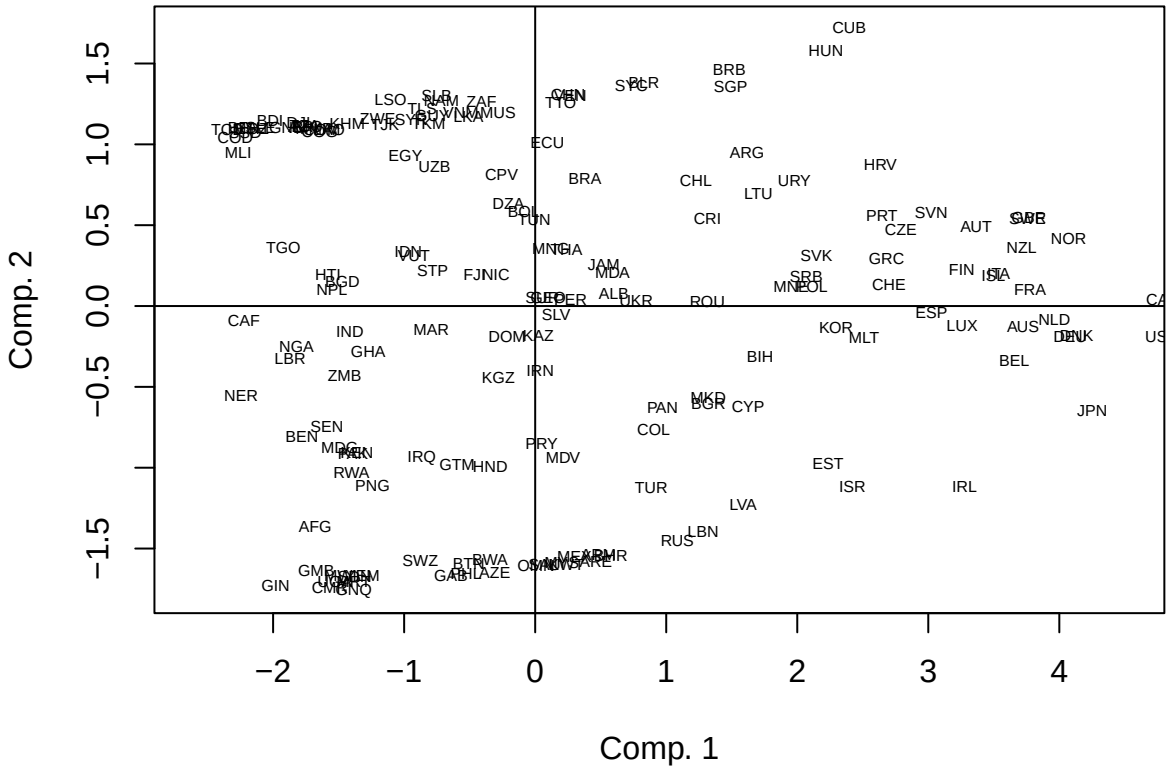

Figure 11: Countries and Principal Components. Distribution of countries, concerning both components.

It is important to explain how ***figure 11*** can be interpreted:

- **Component 1:** Countries positioned on the **right side** of the x-axis are associated with **high levels of HDI, public health spending, and cancer affection/deaths**. Conversely, countries on the **left side are linked to lower levels of these variables**.
- **Component 2:** Countries situated on the **upper side** of the y-axis tend to have had **more left-wing parties** governing them. Conversely, countries on the **lower side tend to have had more right-wing parties** governing them.

So with these new information, it was decided that, to avoid multicollinearity, a new dataset was going to be created, taking into account the two first components. This new dataset was called **pca_results**.

3.2 *Clustering*

For this section, there were two techniques in mind for clustering:

- **Hierarchical clustering:** This method is well-suited for forming groups where objects within a group share similarities and are distinct from objects in other groups. In this scenario, countries will be compared based on their positions in the two components.
- **K-mean clustering:** In this approach, each group is represented by a centroid, which is the average position of all the data points in that cluster. This method aids in clustering by associating each object with the closest centroid, considering the lowest distance between them. Also in this case, centroids and clustering will be done using the two components.

Before applying each clustering method, an evaluation of clustering results was conducted, using different **scoring methods**, to determine the **best clustering technique**, and to gain insight into the **appropriate number of clusters** to establish.

```
##
## 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
```

To understand these results, it is important to describe what each method represents:

- **Connectivity:** Indicates the degree of connectedness of the clusters. Higher values indicate better connectivity within clusters, meaning that points within the same cluster are closer to each other compared to points in different clusters.
- **Dunn:** Measure of the compactness and separation between clusters. Higher values indicate better-defined and well-separated clusters.
- **Silhouette:** Evaluates the quality of clustering by measuring how similar an object is to its own cluster compared to other clusters. Ranges from -1 to 1, and values close to 1 indicate that the data points are well-clustered and closer to other points in the same cluster than to points in neighboring clusters.

With that said, the original plan of **utilizing hierarchical and k-means methods remains consistent**. For **hierarchical clustering, 2 and 6 clusters were employed,** while for **K-means**, although k=2 is optimal according to the cvalid output, it was **also analyzed between k=2 and k=6**.

Additionally, supplementary plotting methods were employed to justify the chosen number of clusters. Two techniques were employed:

1. **Elbow method:** This technique operates under the notion that as the number of clusters increases, the fit improves up to a certain point. Beyond this point, further clusters may lead to over-fitting, resulting in a decline in performance. The optimal number of clusters is determined by identifying the "elbow" point on the plot, indicating where additional clusters no longer significantly enhance the fit.
2. **Silhouette method:** As previously explained, the silhouette value should be maximized, indicating optimal clustering. Hence, the highest point on the silhouette plot denotes the ideal number of clusters.

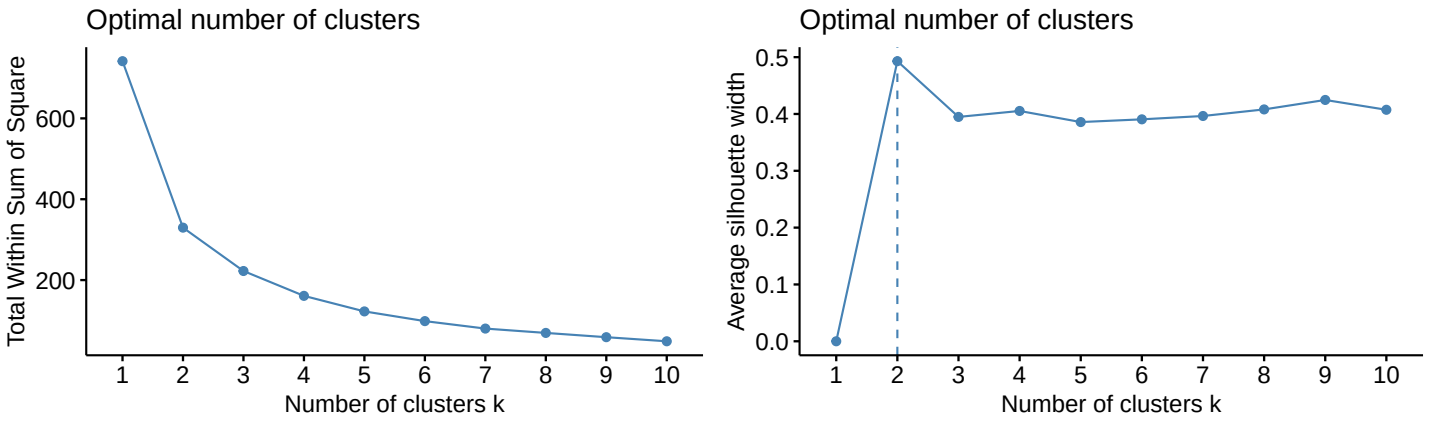

Figure 12: Elbow and Silhouette Methods. These plots show the optimal number of clusters as k=2.

3.2.1 *Hierarchical Clustering*

For this method, taking into account the previous analysis, the number of clusters that were selected and compared are:

- **k=2** taking into account both the evaluation of clustering results and the plots.
- **k=6** taking into account the evaluation of clustering results.

Before implementing the clustering, an analysis of the linkage methods used in hierarchical clustering was done to determine which one is the optimal to use:

- **Single:** Computes the minimum distance between clusters before merging them. May be sensitive to noise in the data.

```
## k=2:  0.1864929
```

```
## k=6:  0.2355196
```

- **Complete:** Computes the maximum distance between clusters before merging them. May be sensitive to outliers.

```
## k=2:  0.4142643
```

```
## k=6:  0.3932472
```

- **Average:** Computes the average distance between clusters before merging them. This method is less sensitive to outliers and tends to produce more balanced cluster sizes.

```
## k=2:  0.5002604
```

```
## k=6:  0.4178778
```

Following this analysis, **hierarchical clustering was performed using the average linkage method**, as it consistently yielded superior results across all cases, indicating better-defined clusters.

3.2.2 *Dendogram*

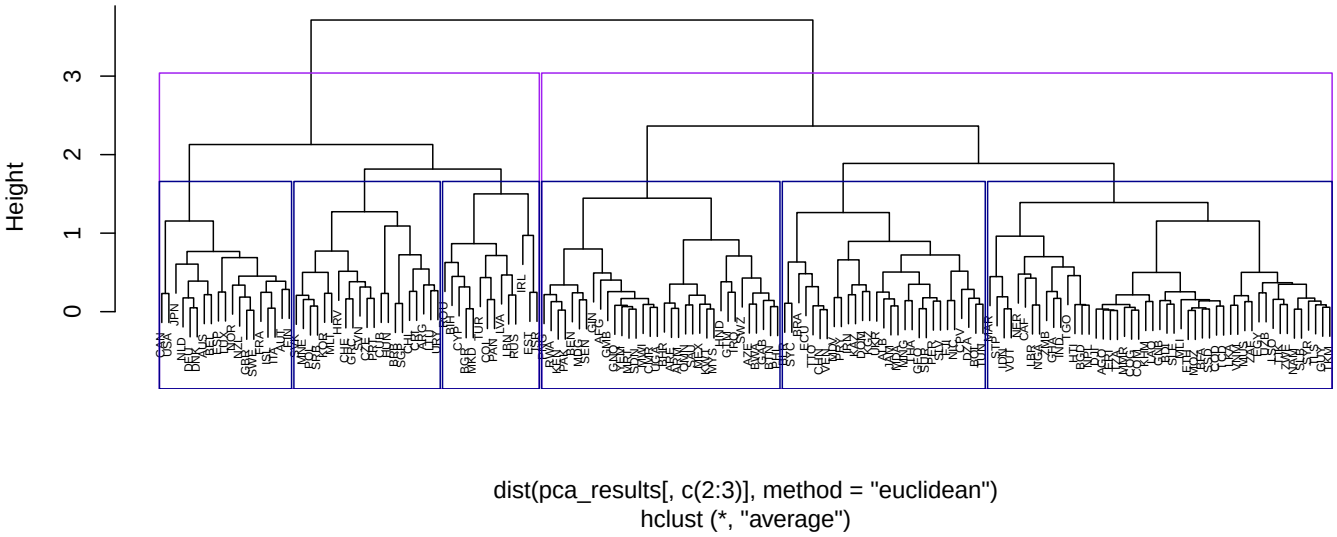

Figure 13: Dendogram. It was plotted using the average linkage method, with the red clustering representing k=2, the green representing k=3, and the blue representing k=6.

Based on the dendrogram in ***figure 13*** and the evaluation of clustering results, it's determined that the two ideal numbers of clusters are **k=2 and k=6**. To make a final comparison, **density plots** were generated to determine the optimal number of clusters for this dataset.

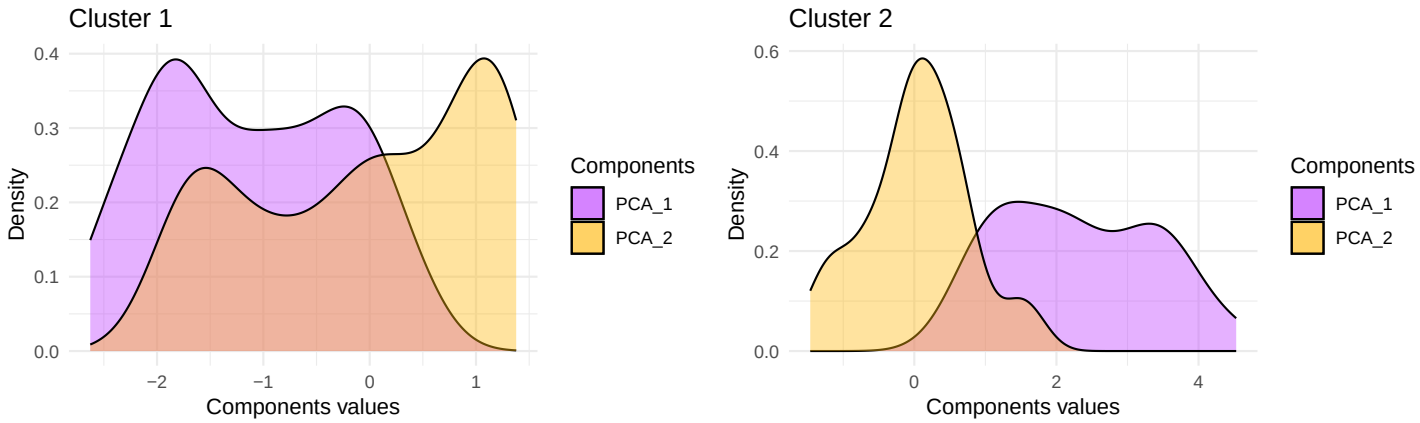

Figure 14: Clusters' Density. Plots of the two clusters' density, regarding the two components.

In ***figure 14***, there is a noticeable distinction between both clusters concerning the two components:

- **Cluster 1:** Typically exhibits **negative values for component 1**, indicating a low level of public health spending, HDI, and cancer rates. Conversely, it tends to have **positive values for component 2**, suggesting that these countries often have had more left-wing parties governing them.
- **Cluster 2:** Generally displays **positive values for component 1**, indicating a high level of public health spending, HDI, and cancer rates. However, it tends to have a more **neutral stance regarding ideologies**, possibly associated with a centrist ideology.

However, it's imperative to conduct a similar analysis with k=6 clustering to determine the optimal number of clusters. This is why the same graphs were plotted for comparison.

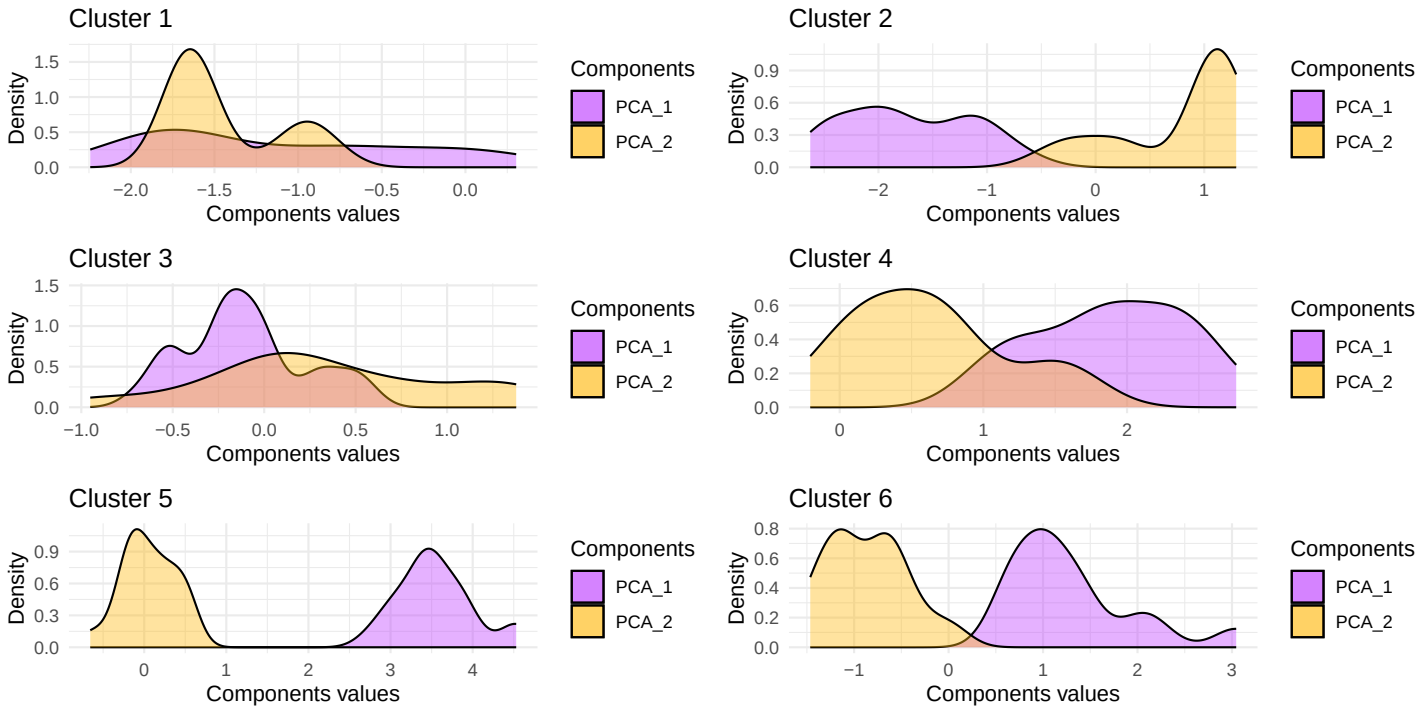

Figure 15: Clusters' Density. Plots of the six clusters' density, regarding the two components.

Taking into account the results from ***figure 15*** and the dataset's complexity, it seems that a **k=6 clustering would be better suited to effectively cluster countries**. This is evident from the **distinct characteristics** exhibited by each cluster in the plot, which are not adequately represented with k=2. However, it's essential to acknowledge that increasing the number of clusters can lead to overfitting, that is why no greater k was selected.

3.3 *K-Means*

Last but not least, **k-means clustering** was performed to partition the observations into a predetermined number of clusters using the PCA results.

This method differs from hierarchical clustering in that we **predefine the number of clusters in advance**, whereas in hierarchical clustering, the number of clusters is determined after generating the tree representation. In this case, k-means calculates **centroids** for each cluster and minimizes the **within-cluster sum of squares (WCSS)**, which represents the total squared distance between each data point and the centroid.

As mentioned earlier, k=2 and k=6 were used for this method. It was essential to assess the effectiveness of the selected k by using the **mean silhouette width**, where a greater value indicated better clustering quality. For better understanding, the analysis was conducted between 2 and 6, including 3, 4, and 5, for comparative purposes

```
## k=2:  0.5038346

## k=3:  0.4352006

## k=4:  0.4329202

## k=5:  0.4436106

## k=6:  0.4315723
```

The mean silhouette width of 0.50 for k=2 was considered sufficiently good and better than the other values of k. Nevertheless, this is was not considered a sufficient argument for choosing 2 as the best k. Additionally, it was also compared the "within cluster sum of squares by cluster" to determine if it has a sufficient proportion of data variability explained by the cluster structure.:

```
## k=2:

## WCSS =  83.6088 226.8112

## Between SS / Total SS =  0.581355

## k=3:

## WCSS =  67.34458 70.66503 62.59388

## Between SS / Total SS =  0.729458

## k=4:

## WCSS =  28.30195 30.83475 36.30385 44.36079

## Between SS / Total SS =  0.8114582

## k=5:

## WCSS =  16.00063 14.92747 18.63392 22.54711 36.30385

## Between SS / Total SS =  0.8537899

## k=6:

## WCSS =  8.282212 15.52156 16.65918 13.62473 18.72818 10.17613

## Between SS / Total SS =  0.8880736
```

During this analysis, **a new k** was selected based on the WCSS output, leading to new conclusions:

- **k=4** logically exhibits more compact clusters, with distances between data points and their centroids lower than k=2.
- **k=4** also achieves a BSS/TSS ratio of 81%, suggesting it is a good fit as it explains a sufficient proportion of data variability, unlike k=2 and k=3, which yielded lower values.
- It is important to remember that increasing k may lead to overfitting of the model.

In conclusion, taking into account all the previous analysis and this new information, it was defined that **k=4 should be the number of clusters to use in the k-means model**. ***Figure 16*** illustrates this results:

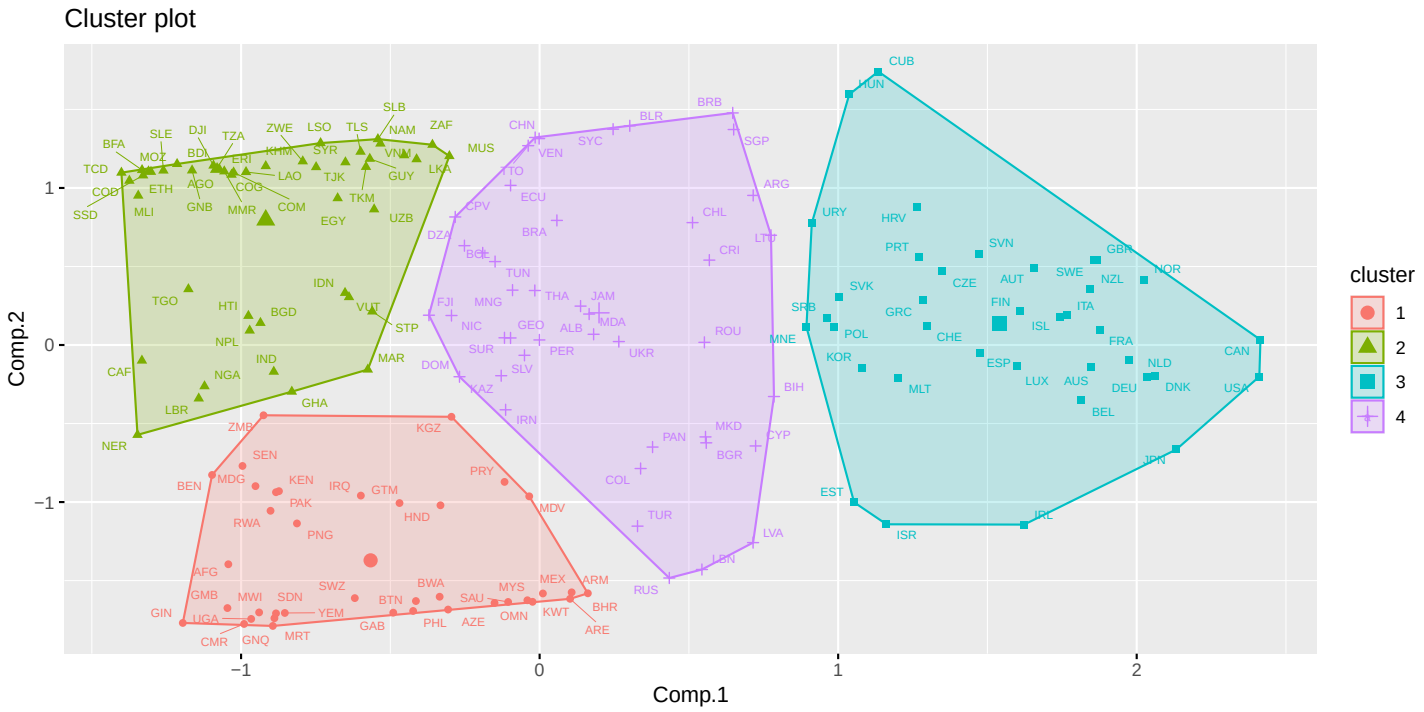

Figure 16: K-Mean Clusters. This plot represents the 2 clusters created with this method.

4 Supervised Learning

For this analysis, **linear regression, decision tree, and random forest models** were utilized to further explore the dataset. Additional variables were introduced to enhance the comprehension and prediction of the dependent variable, which, in this case, was **cancer_death_rate**.

Before implementing any model, certain data manipulations were conducted to streamline the variables:

- **rate_50_over:** All columns representing age ranges were consolidated into this single column, which indicates the percentage of deaths among individuals over 50. This decision was informed by the original density plot, which notably highlighted a higher prevalence of cancer-related deaths within this age group.
- **main_cancers:** The multitude of cancer type columns were replaced by a single column representing the five primary cancer types. These main cancers, aligned with the World Health Organization's 2019 rankings, comprise lungs, breast, prostate, rectum, and stomach cancers.

Furthermore, the correlation among variables was checked, to avoid mutlicolineality in the models:

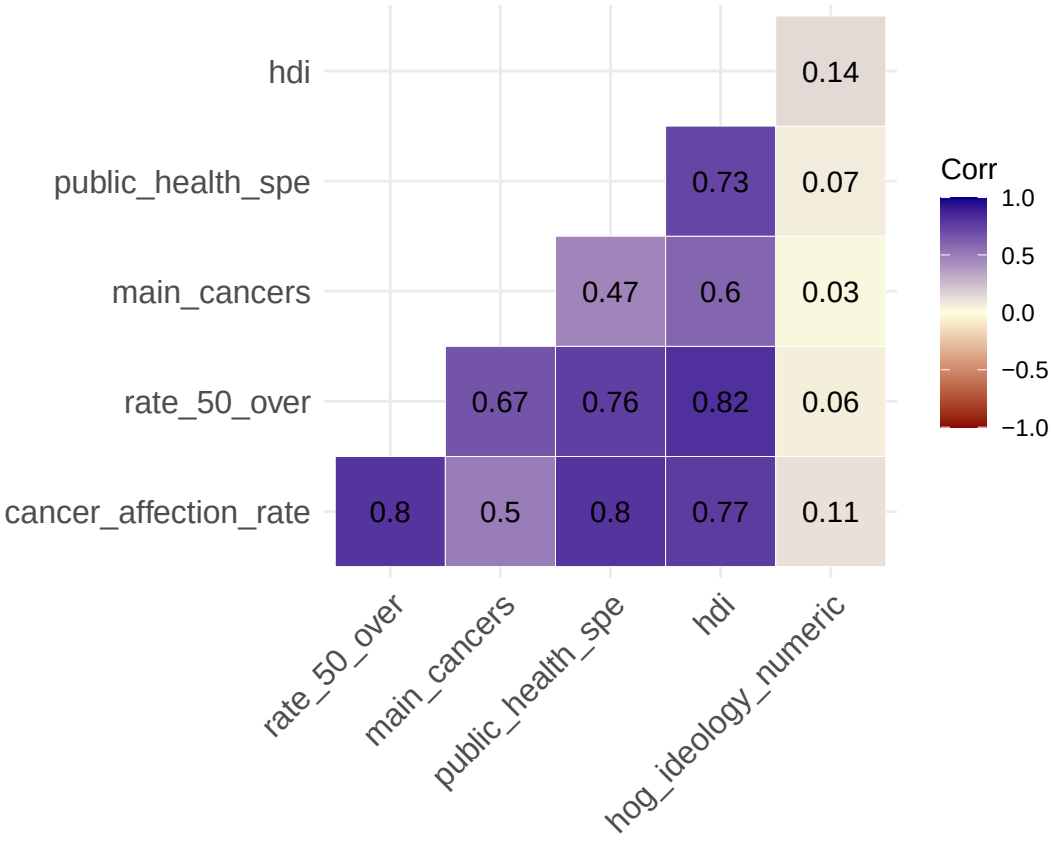

Figure 17: Correlation Heatmap. Correlation between independent variables, showing multicolineality.

In this case, as there is clear **multicollinearity among independent variables** in the ***figure 17***, it was decided to apply **PCA and use the components for the regression**. The first two components were selected, as they explain almost the 80% of the dataset's variance:

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

The first component explains all variables except ideology, while **component 2 explains ideology**. These results were sufficient for analyzing the project's objective, so two columns were added with the component's values for each row.

```
##
## Loadings:
##                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## cancer_affection_rate  0.462         0.334         0.668  0.474
## rate_50_over           0.479                -0.161  0.284 -0.811
## main_cancers           0.371 -0.126 -0.840  0.307         0.215
## public_health_spe      0.446         0.411  0.596 -0.522
## hdi                    0.465                -0.718 -0.447  0.256
## hog_ideology_numeric          0.988 -0.107
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
## Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000
```

4.1 *Regression model*

The **regression model** serves as a mathematical tool for predicting the value of a variable based on the values of other variables. The variable being predicted is called the **dependent variable**, while the variables used to predict it are known as **independent variables**. This analytical approach involves estimating the **coefficients** of a linear equation, incorporating one or more independent variables to best predict the value of the dependent variable.

For this project, two regression models were compared to determine the superior fit:

- **Original dataset:** Utilizing the original dataset alongside the results obtained from PCA.
- **Average by country:** Employing the average of each variable, aggregated by country, also integrating PCA.

While both models exhibited adequacy and similarity, the preference in this project was given to the **average approach**, as in the unsupervised learning section. Utilizing the original dataset treats each country/year combination as separate samples, potentially overlooking the effects across years under the governance of the same administration. Therefore, **employing the average allows for the inclusion of the effects of ideological tendencies and cancer/health outcomes over the years**, enriching the model with a more comprehensive temporal perspective.

Taking into account the considerations mentioned previously, and bearing in mind that component 1 explains all variables except ideology, while component 2 explains ideology, the following linear regression model was constructed:

```
##
## 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
```

In addition, several metrics were examined to analyze the effectiveness of the selected model:

```
## R^2 = 0.8601083
```

```
## RMSE = 0.03484603
```

```
## MAE = 0.026823
```

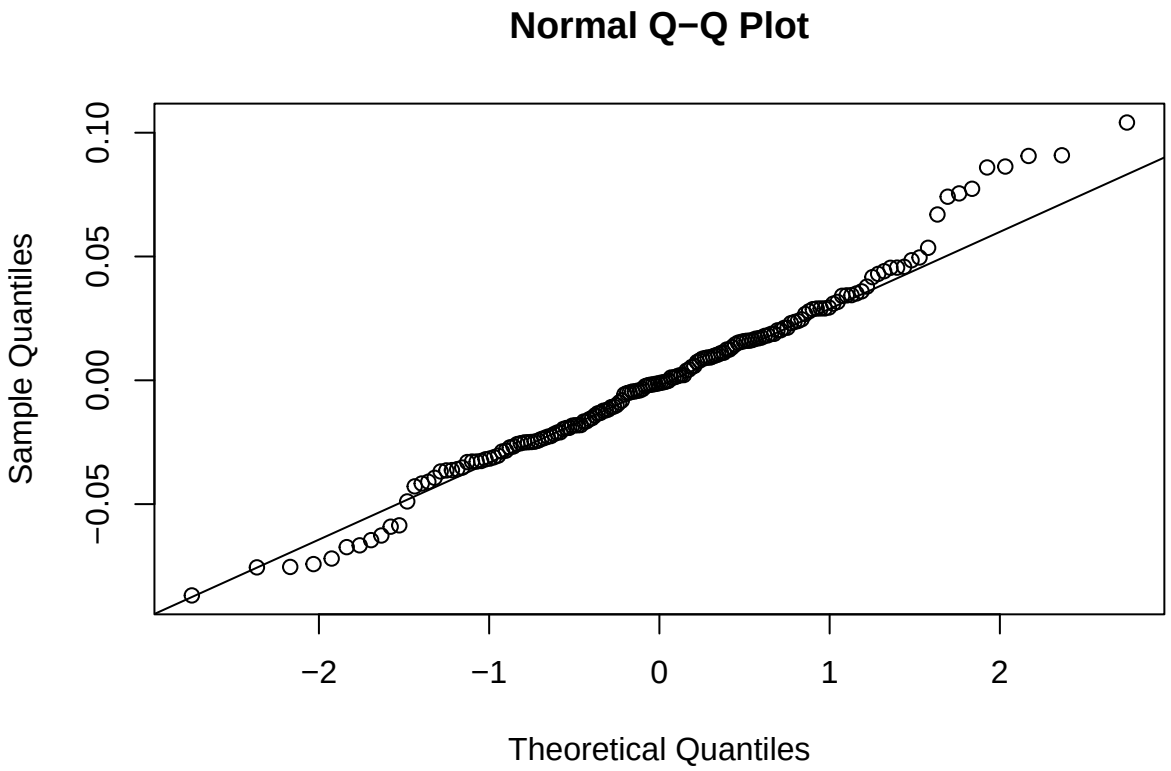

Figure 18: Normal Q-Q Plot. .

The results of the regression model indicate its **effectiveness in predicting cancer deaths**, as it explains 86% of the variance and demonstrates a **lower RMSE**. Additionally, the **residuals** exhibit a good fit to a **normal distribution**, as illustrated in ***figure 18***.

Furthermore, in alignment with the project's objective, the **insignificance of the second component** suggests that there is **no significant effect of ideology** on the dependent variable.

4.2 *Decision Tree*

Continuing with the supervised techniques, an analysis involving **decision trees** was conducted. Decision trees are a type of decision support model that provides a **hierarchical structure of decisions and their potential outcomes**.

This hierarchical, tree structure, consists of:

- **Root node:** The first node, from which all branches originate.
- **Branches:** This are lines connecting nodes. They represent the possible decisions or outcomes based on the value of a particular feature. Each branch leads to a subsequent node or leaf node.
- **Internal nodes:** Are decision nodes within the decision tree that contain a decision rule based on a feature or attribute. These nodes split the dataset into smaller subsets.
- **Leaf nodes:** This are the final nodes. They represent the final outcome or decision.

For this technique, the results of the entire dataset ***df_supervised*** were compared against those of the dataset with the annual averages by country (***df_supervised_avg***):

df_supervised:

```
## Model's performance:

##
## Regression tree:
## rpart(formula = cancer_death_rate ~ hog_ideology_numeric + cancer_affection_rate +
##     main_cancers + public_health_spe + hdi, data = train_set)
##
## Variables actually used in tree construction:
## [1] cancer_affection_rate hdi                   public_health_spe
##
## Root node error: 17.373/1950 = 0.0089093
##
## n= 1950
##
##         CP nsplit rel error  xerror      xstd
## 1 0.683575      0   1.00000 1.00057 0.0211850
## 2 0.086335      1   0.31642 0.32480 0.0094254
## 3 0.080972      2   0.23009 0.22962 0.0080006
## 4 0.013555      3   0.14912 0.15596 0.0049102
## 5 0.011937      4   0.13556 0.14422 0.0045632
## 6 0.010000      5   0.12363 0.13338 0.0044576

## RSME = 0.03336647

## Accuracy =  0.8782363
```

Considering the preceding results, a **complexity parameter (CP) value of 0.01** was selected for constructing the tree, resulting in the following outcomes:

```
## Model's cross validation:

## CART
##
## 2782 samples
##    5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 2506, 2505, 2504, 2504, 2504, 2504, ...
## Resampling results:
##
##   RMSE        Rsquared   MAE
##   0.03395775  0.8716952  0.02686671
##
## Tuning parameter 'cp' was held constant at a value of 0.01
```

df_supervised_avg:

```
## Model's performance:

##
## Regression tree:
## rpart(formula = cancer_death_rate ~ hog_ideology_numeric + cancer_affection_rate +
##     main_cancers + public_health_spe + hdi, data = train_set_2)
##
## Variables actually used in tree construction:
## [1] cancer_affection_rate hdi                   public_health_spe
##
## Root node error: 1.0515/118 = 0.0089109
##
## n= 118
##
##         CP nsplit rel error  xerror     xstd
## 1 0.721065      0  1.000000 1.01250 0.086589
## 2 0.087213      1  0.278935 0.30219 0.036449
## 3 0.072522      2  0.191722 0.26720 0.034301
## 4 0.016589      3  0.119200 0.15584 0.021199
## 5 0.013567      4  0.102612 0.14189 0.018032
## 6 0.010000      5  0.089044 0.13843 0.016930

## RSME = 0.02850167

## Accuracy =  0.9017756
```

In this instance, taking into consideration the previous findings, once again a **CP value of 0.01** was chosen for constructing the new tree, yielding the following results:

```
## Model's cross validation:

## CART
##
## 166 samples
##   5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 149, 149, 150, 149, 150, 150, ...
## Resampling results:
##
##   RMSE        Rsquared   MAE
##   0.03445706  0.8771741  0.02745663
##
## Tuning parameter 'cp' was held constant at a value of 0.01
```

Mainly, it is evident that the model utilizing the *df_supervised_avg* dataset exhibited superior performance, characterized by a **lower RMSE and a higher accuracy**. Considering the "***variables actually used in tree construction***", it can be deduced from both models that once again **ideology does not significantly contribute to the prediction of cancer death**.

Furthermore, despite one model outperforming the other, both decision trees were plotted to facilitate visual comparison. Nonetheless, for the project's objectives, the conclusion remains consistent:

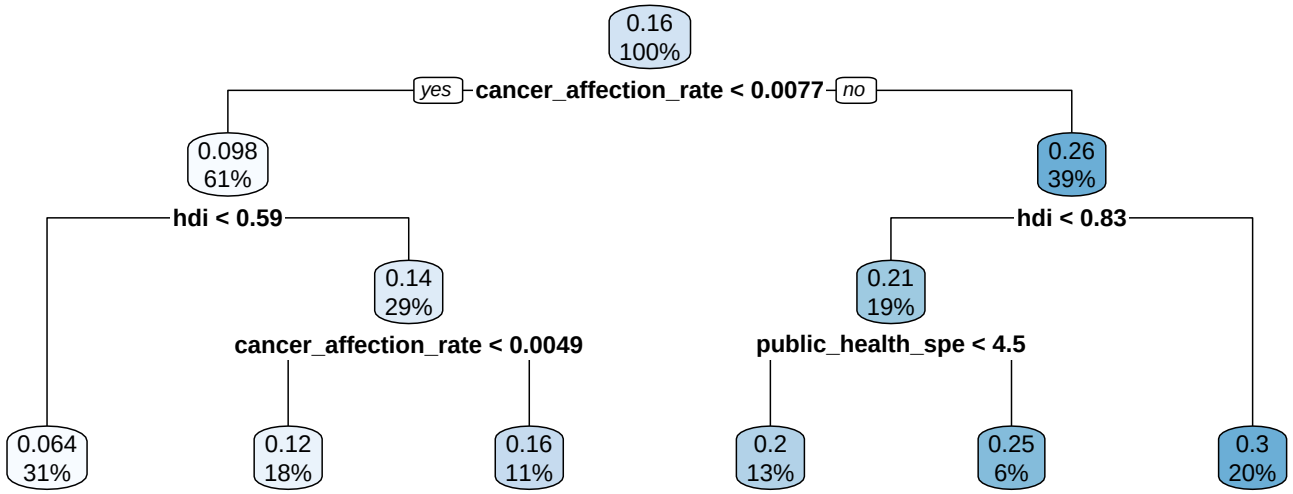

Figure 19: Tree with supervised dataframe. The variables used were cancer affection rate, hdi and public health spending

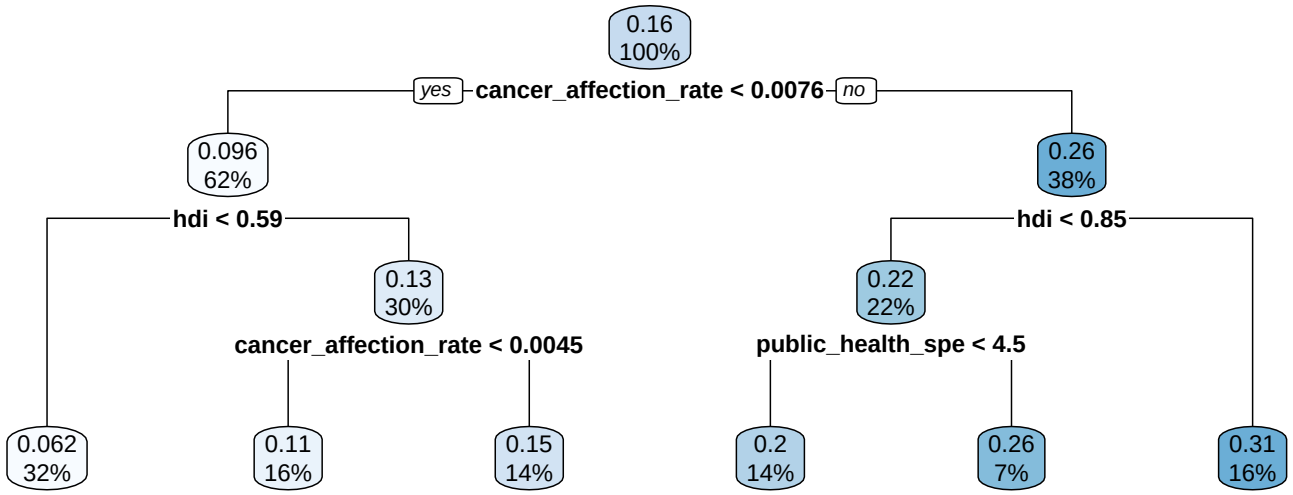

Figure 20: Tree with supervised avg dataframe. The variables used were cancer affection rate, hdi and public health spending.

4.3 *Random Forest*

Last but not least, a **random forest** analysis was done. This model **combines the output of multiple decision trees to reach a single result**. The main difference to decision tree is that employs a technique called **Bootstrap Aggregating (Bagging)**, where multiple decision trees are trained on different subsets of the training data (educing overfitting and increasing generalization performance), and uses a **random** subset of features at each split in the decision tree.

To define the best model, it was compared a model with the ***df_supervised***, ***df_supervised*** without the ideology variable, and ***df_supervised_avg***:

```
## Model with df_supervised:

##
## Call:
##  randomForest(formula = cancer_death_rate ~ cancer_affection_rate +      main_cancers + public_health_spe +
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 0.0005754483
##                     % Var explained: 93.54

## Model with df_supervised (w/o ideology):

##
## Call:
##  randomForest(formula = cancer_death_rate ~ cancer_affection_rate +      main_cancers + public_health_spe +
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 0.0003936925
##                     % Var explained: 95.58

## Model with df_supervised_avg:

##
## Call:
##  randomForest(formula = cancer_death_rate ~ cancer_affection_rate +      main_cancers + public_health_spe +
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 0.001120507
##                     % Var explained: 87.43
```

According to the results, the model with ***df_supervised*** without ideology ouputted the best result, but for the project's purpose, the analysis was continued with the model including ***df_supervised*** as it had a better performance than ***df_supervised_avg***.

Finally, this random forest was used for creating a **training random forest with 10 fold cross validation and with the ranger package**:

- **Random forest with 10 fold cross validation:**

```
## 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
```

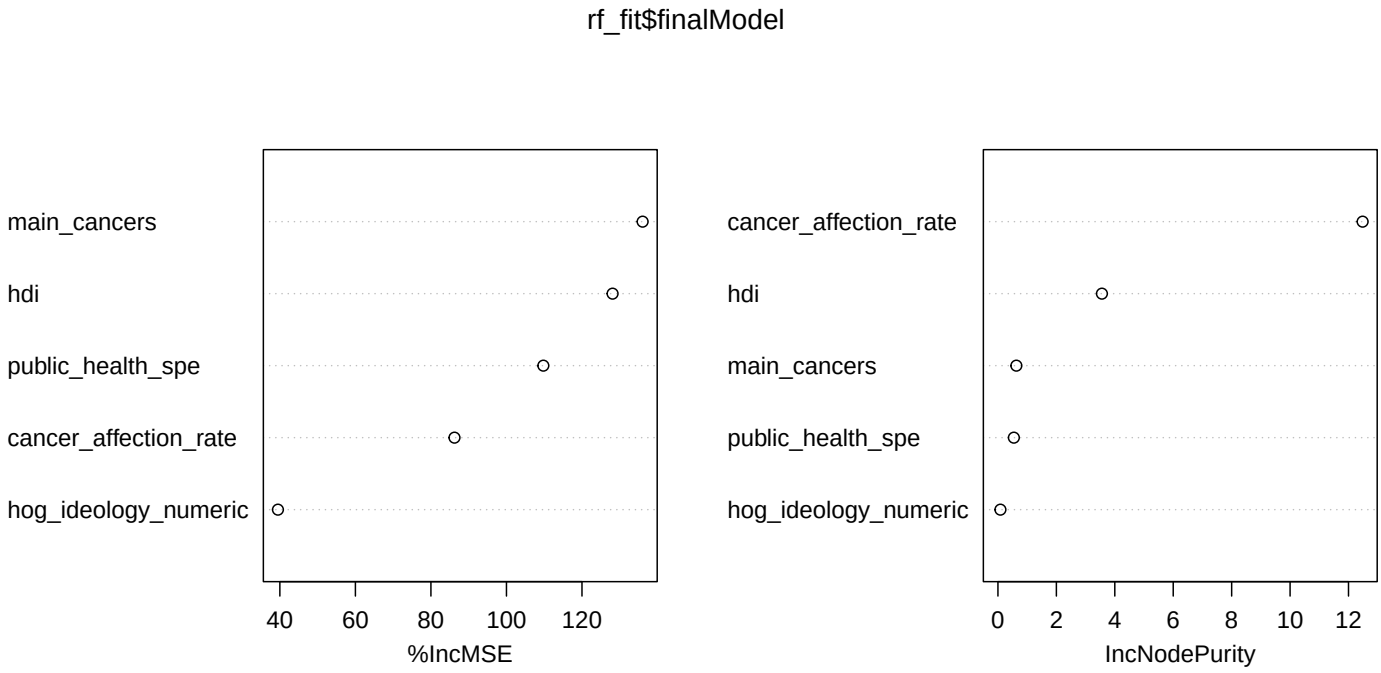

Figure 21: Increase in MSE and Node Purity.

- **Random forest with ranger package:**

```
## 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
```

Finally, considering all these results, **the project's objective remains supported**. Although both models (10-fold cross-validation and the raner package) yielded similar **RMSE** values, some differences in the **importance of each variable** were observed. However, one aspect is clear: **ideology demonstrates no significant impact on cancer deaths**.

Furthermore, as depicted in ***figure 21***, it is evident that **all variables except ideology** are deemed important (**%IncMSE**) and contribute to enhancing the model (**IncNodePurity**). To clarify these concepts:

- **%IncMSE:** This metric quantifies the importance of variables in the model by indicating how much the model's mean squared error (MSE) would increase if that variable were eliminated. **A higher value signifies greater importance of the variable**.
- **IncNodePurity:** This metric measures the improvement in the purity of nodes in the decision tree when a variable is added at a specific point. **A higher value indicates a greater improvement in the decision tree's purity upon splitting**.

These metrics provide valuable insights into the significance and contribution of each variable to the model's performance and overall predictive accuracy.

---

5 Conclusion

To conclude this project, it's essential to highlight that **the objective was strongly supported by the results obtained** from both unsupervised and supervised techniques. These results demonstrate that **political ideologies do not significantly affect cancer deaths, whereas variables such as public health spending and the Human Development Index (HDI) play crucial roles.**.

However, further analysis is highly recommended:

- **New independent variables:** While cancer deaths were chosen due to the substantial investments it necessitates, additional variables may provide deeper insights. Variables such as family income, lifestyle choices (e.g., healthy habits, physical activity, smoking), and personal health insurance could be included to enrich the analysis.
- **Different perspective of the ideology variable:** Exploring the project from a different angle, such as analyzing the entire tenure of each political party rather than focusing solely on yearly or annual averages, could offer alternative insights. This approach acknowledges that the effects of political decisions may take time to materialize.
- **Change the dependent variable (cancer_death_rate):** This project could be reexamined from a different perspective, aiming to demonstrate that regardless of the governing ideology, a country's progress hinges on tangible actions rather than rhetoric. Variables such as education level, employment rate, entrepreneurship rate, etc., could be considered as alternative dependent variables.

As previously emphasized, this project is not intended to engage in political debates or persuade individuals to alter their political beliefs. Instead, its primary aim is to show that **actions speak louder than words, and true change is brought about by leading through example.**.