First Homework: Unsupervised Learning

Jorge Barcia Belinchón 100496595 Paloma Núñez Guerrero 100496533

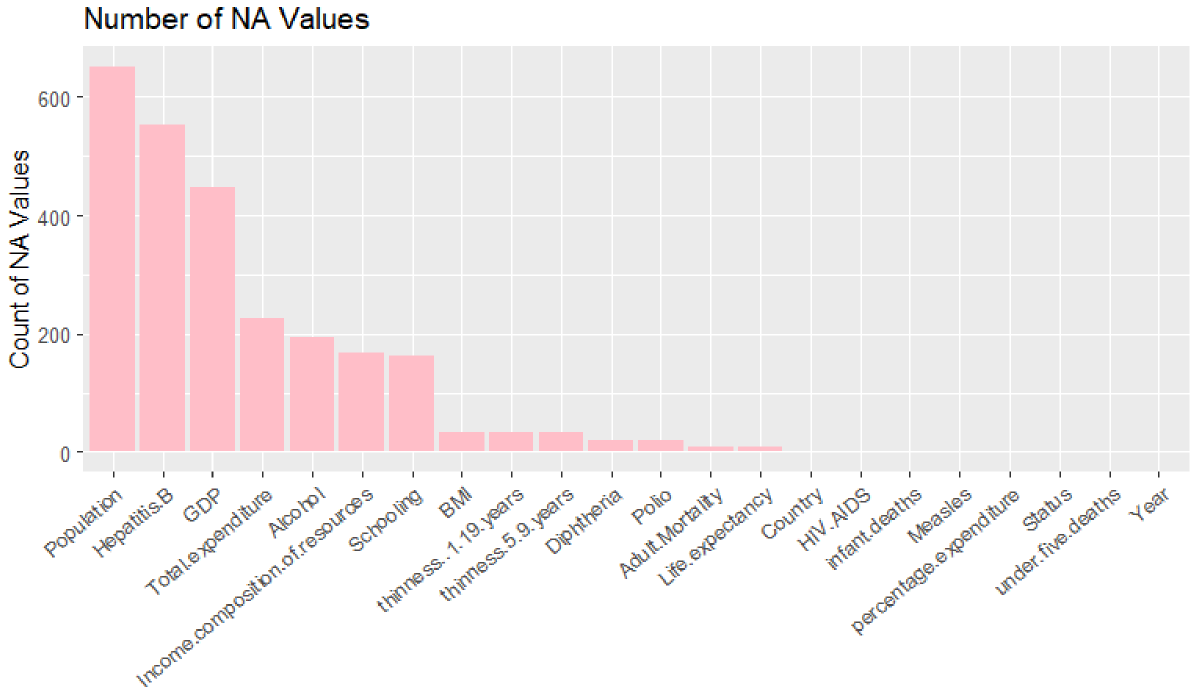
Introduction

Life expectancy is one of the most important indicators of the overall health and well-being of a country's population. Taking this into account, one of our goals is to better understand the factors that influence life expectancy, and to identify the key attributes that distinguish developed countries from others.

Using data from the World Health Organization (WHO), we employed unsupervised learning techniques such as PCA, FA, and clustering to uncover hidden patterns and relationships in the data. This exploratory analysis aims to provide valuable insights into global health disparities and the characteristics of highly developed nations.

Data preprocessing

Data preprocessing was fairly straightforward, after a quick analysis of number and location of the NA's by this plot.

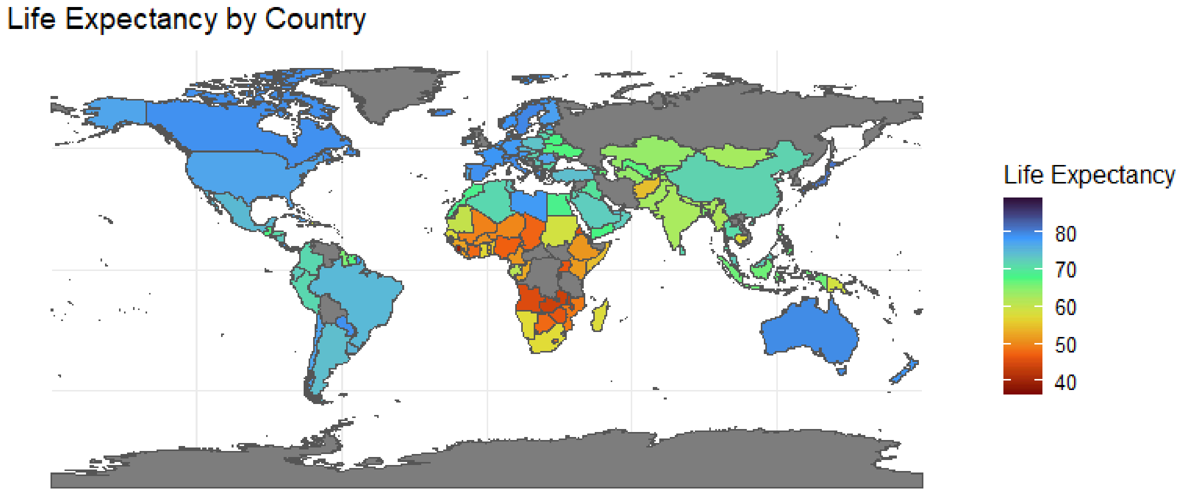


We encountered some NA values in numerical categories, and since we were interested in retaining these records, we used imputation methods with the mean.

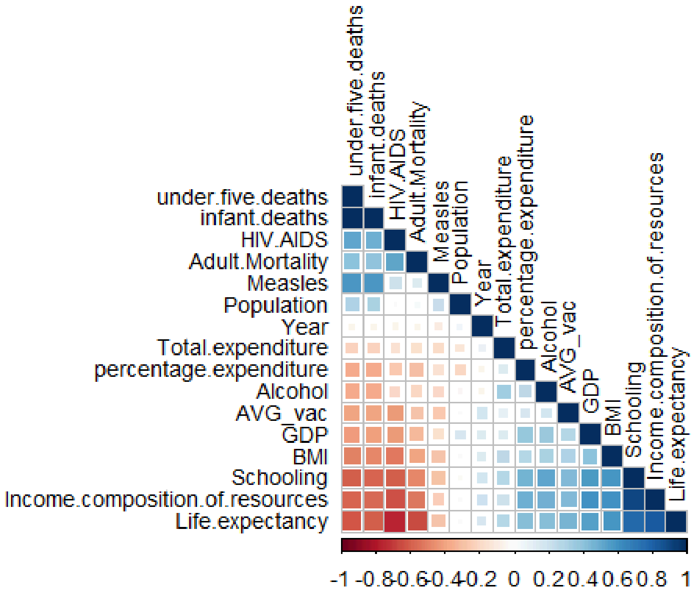
For feature engineering, we categorised life expectancy, merged vaccination rates into a single category, and simplified our dataset by removing features that were not of interest.

Visualisation tools to get insights before the tools

In order to get to know our data we started by analysing the distribution of **Life Expectancy** across the globe so we could clearly see how it works.



We also check for correlation between features, as it will be useful for later analysis, most importantly in the FA.

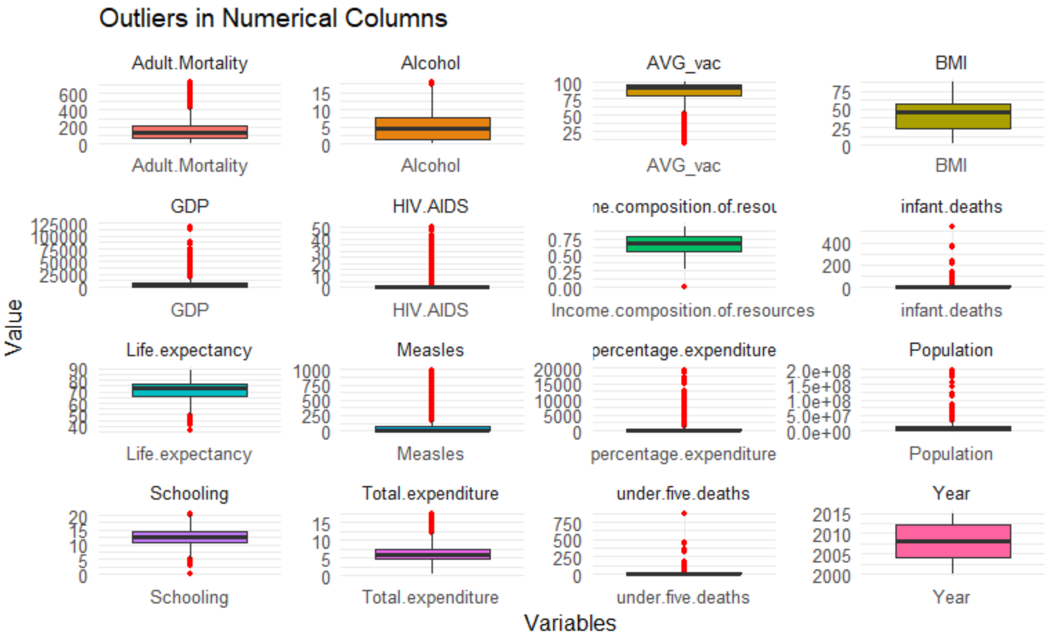


We check that we have both positive and negative correlation. Therefore, we continue to further explore the data using visualisation. One of the things we want to answer is how Schooling affects **life expectancy,** so we looked for a linear relationship with a scatter plot in order to assess how strong the correlation is.

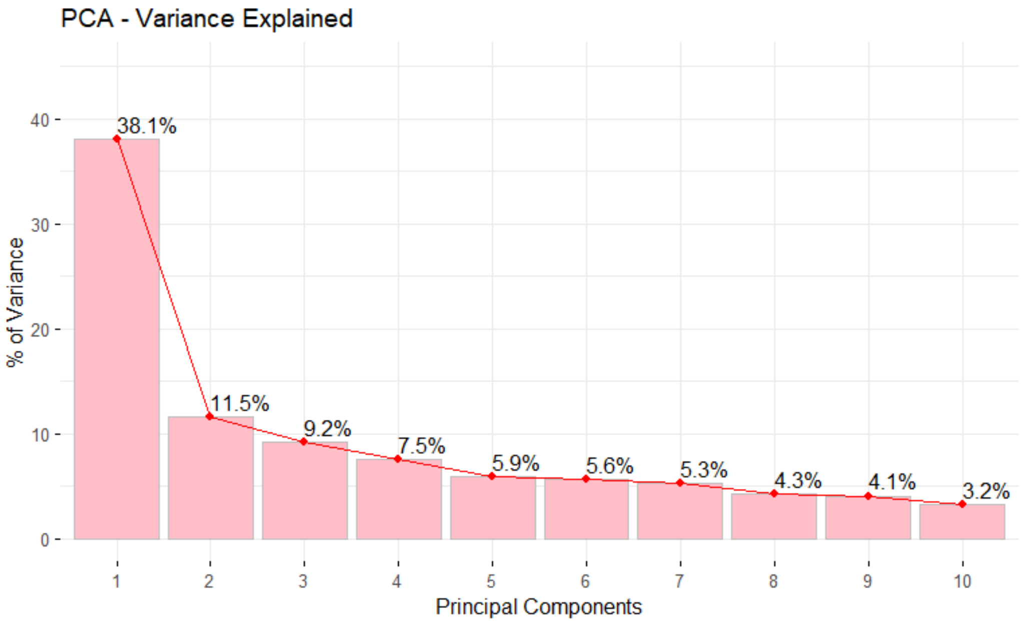
We used animated plots to present the information in a more clear, more meaningful, and visually engaging way. (Refer to the HTML section of the project for a better view of the animations, as PDFs do not easily support GIF images.) With these plots, we aimed to illustrate how life expectancy evolved over time globally, and to highlight the differences in the distribution between developed and developing countries.

Principal Component Analysis

Before starting the Unsupervised Learning analysis we need to take into account those outliers that can ruin our analysis. For this matter, we started by deleting impossible values in categories with unrealistic ratios of +100%. After this, we plotted the rest of outliers by using Boxplots for each of the numerical categories.

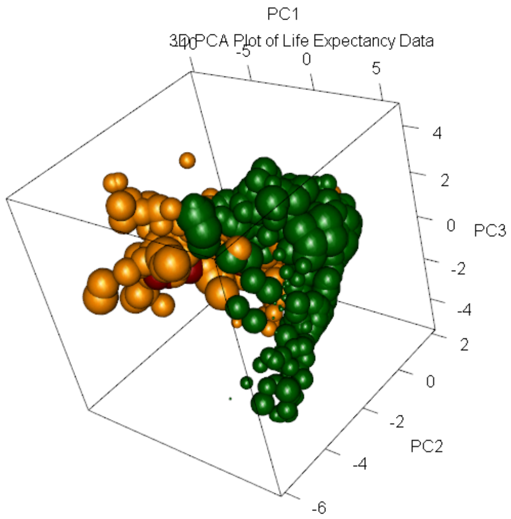


We can check visually that we indeed have some outliers. To handle them, we used the 3-sigma rule and erased these problematic values. The amount of outliers was counted, and after some preprocessing of the data, we performed the PCA. Our results clearly showed that the PC1 explained most of our variance. However, we needed to add up to PC4 in order to explain an acceptable amount of variance. For this PC1, no variable had a high amount of contribution and this was fairly distributed among them.



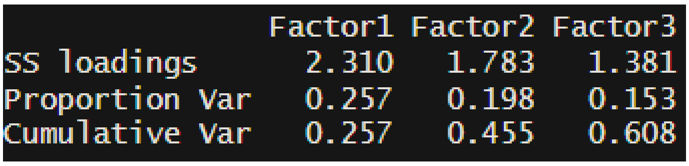
Checking the top and worst performers, we can see that our PC1 works at assessing the overall State of the counties.

After plotting 2D and 3D plots of our PCA, it is clear that the state of the Life Expectancy of a country can be predicted with our 4 dimensions. This clearly shows that we successfully reduced the dimensionality of this problem. (see html part to better see this plot and be able to interact)

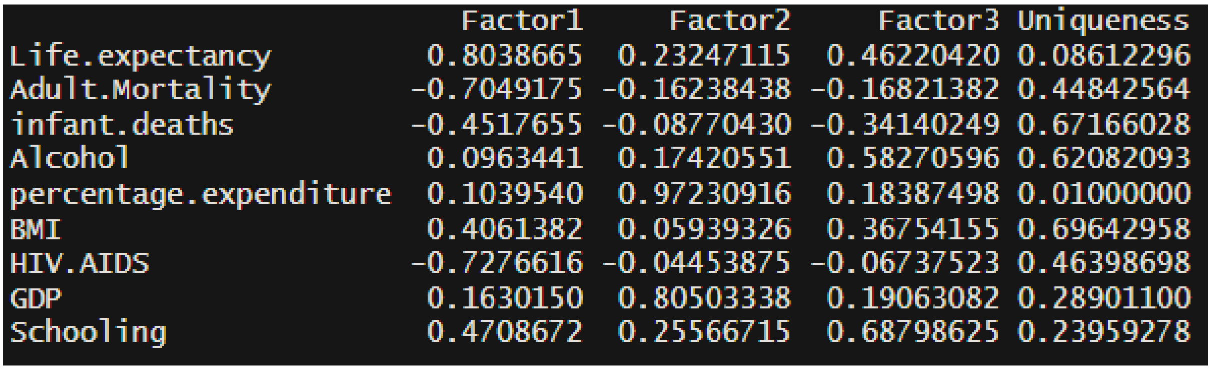


Factor Analysis

After Preprocessing and performing the first analysis, it was revealed that variables like **Population** and **Measles** had high uniqueness, indicating independence and poor explanation by other factors. Removing these variables improved the model's coherence and explained variance. A refined 3-factor model explained 61% of the total variance, demonstrating a more efficient structure. For predicting Life Expectancy, a 2-factor model explained up to 94% of its variance, highlighting its predictive efficiency for this matter.



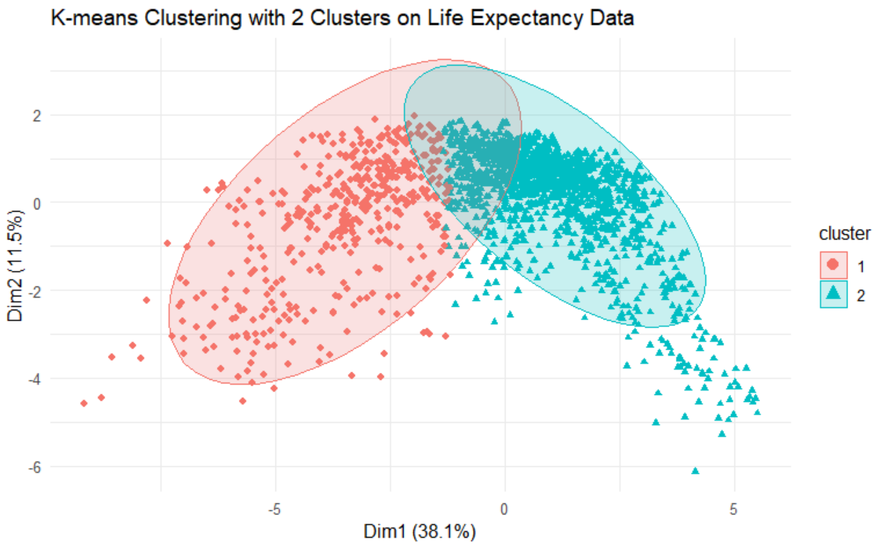
```
               Factor1 Factor2 Factor3
SS loadings      2.310   1.783   1.381
Proportion Var   0.257   0.198   0.153
Cumulative Var   0.257   0.455   0.608
```



```
                          Factor1     Factor2     Factor3 Uniqueness
Life.expectancy         0.8038665  0.23247115  0.46220420 0.08612296
Adult.Mortality        -0.7049175 -0.16238438 -0.16821382 0.44842564
infant.deaths          -0.4517655 -0.08770430 -0.34140249 0.67166028
Alcohol                 0.0963441  0.17420551  0.58270596 0.62082093
percentage.expenditure  0.1039540  0.97230916  0.18387498 0.01000000
BMI                     0.4061382  0.05939326  0.36754155 0.69642958
HIV.AIDS               -0.7276616 -0.04453875 -0.06737523 0.46398698
GDP                     0.1630150  0.80503338  0.19063082 0.28901100
Schooling               0.4708672  0.25566715  0.68798625 0.23959278
```

Clustering tools

According to the Elbow method 2 clusters was the optimal number of clusters for our dataset therefore we performed the analysis according to this number.

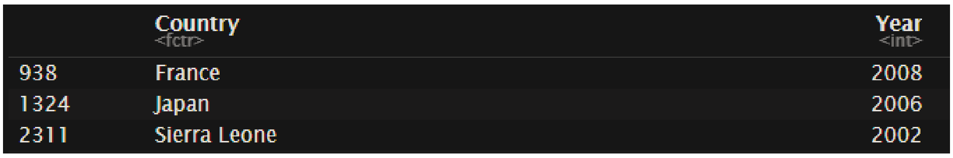


Plotting the mean of each category for the two clusters, we extracted valuable insights about the data set. The clustering analysis of the Life Expectancy dataset reveals key differences between countries in Cluster 1 and Cluster 2:

- Cluster 2: Higher averages for positive indicators such as life expectancy, GDP, and schooling, and lower averages for negative indicators like adult mortality, infant deaths, and HIV/AIDS prevalence. This suggests better socioeconomic and health conditions.
- Cluster 1: Lower averages for positive indicators and higher averages for negative ones, indicating greater health and economic challenges.

Our principal insight suggested that:

Measles rates differ significantly between clusters, but factor analysis did not find strong linear relationships for Measles. Investigation revealed that Measles incidence is more influenced by public health policies like vaccination campaigns than by socioeconomic factors. This explains why Measles was identified as independent in the factor analysis, despite its distinct differences in clustering. For further checking this fact we extracted the most influential observations additionally confirming our theory.



	Country	Year
	<fctr>	<int>
938	France	2008
1324	Japan	2006
2311	Sierra Leone	2002

Conclusion

Using unsupervised learning methods, we uncovered key patterns in health and socioeconomic indicators, challenging assumptions like the independence of measles rates from socioeconomic factors. This analysis improved our dataset understanding and enhanced our exploratory data analysis skills.