Case Study 3 - PCA and K-Means Clustering Solutions

Today's Goal

Welcome to the third case study. This is the last case study in our series and we will be focusing on the practical application of PCA analysis and K-Means Clustering in real-world scenarios. In this case study, we will work on a popular Kaggle challenge: categorizing the countries using socio-economic and health factors that determine the overall development of the country.

Background

HELP International have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision have mostly related to the countries in dire need of Aid. **Our job as a Data analyst is to classify the countries using the socio-economic and health factors that determine the overall development of nations. After this analysis, we need to suggest countries that the CEO needs to focus on and give the highest priority.**

About the client

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

Dataset Overview

The dataset we will be working with is sourced from a Kaggle collection. It gathers comprehensive information on countries' socio-economic and health statistics. These variables offer valuable insights into the development of nations.

Before we start our analysis, please download the dataset 'country_data.csv' from the GitHub folder 'case study'.

The original dataset source: https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data?select=data-dictionary.csv.

Variable Descriptions

- country: Name of the country,
- child_mort: Death of children under five years of age per 1000 live births,
- exports: Exports of goods and services; Exports of goods and services given as %age of the Total GDP;
- health: Total health spending per capita. Given as %age of GDP per capita
- imports: Imports of goods and services, Given as %age of the Total GDP;
- income: Net income per person;
- inflation: The measurement of the annual growth rate of the Total GDP;
- life_expec: The average number of years a newborn child would live if the current mortality patterns are to remain the same;
- total_fer: The number of children born to each woman if the current age-fertility rates remain the same.

- gdpp: The GDP per capita. Calculated as the Total GDP divided by the total population.

1. Getting Started

A. Initial Setups

(i) To begin, make sure you have the following libraries installed and loaded in your R environment: tidyverse, ggplot2, and dplyr.

```
library(tidyverse)
library(dplyr)
library(ggplot2)
```

(ii) Set the seed to '123' to ensure the reproducibility of your results.

```
set.seed(123)
```

(iii) Now, please load the dataset that you've downloaded into your R environment. Assign this data to an object named 'my_data'.

```
my_data <- read.csv('country_data.csv')
```

B. Data Cleanings

(i) To begin, familiarize yourself with the structure of the dataset.

```
# head(my_data)
glimpse(my_data)
```

```
## Rows: 167
## Columns: 10
## $ country    <chr> "Afghanistan", "Albania", "Algeria", "Angola", "Antigua and~
## $ child_mort <dbl> 90.2, 16.6, 27.3, 119.0, 10.3, 14.5, 18.1, 4.8, 4.3, 39.2, ~
## $ exports    <dbl> 10.0, 28.0, 38.4, 62.3, 45.5, 18.9, 20.8, 19.8, 51.3, 54.3,~
## $ health     <dbl> 7.58, 6.55, 4.17, 2.85, 6.03, 8.10, 4.40, 8.73, 11.00, 5.88~
## $ imports    <dbl> 44.9, 48.6, 31.4, 42.9, 58.9, 16.0, 45.3, 20.9, 47.8, 20.7,~
## $ income     <int> 1610, 9930, 12900, 5900, 19100, 18700, 6700, 41400, 43200, ~
## $ inflation  <dbl> 9.440, 4.490, 16.100, 22.400, 1.440, 20.900, 7.770, 1.160, ~
## $ life_expec <dbl> 56.2, 76.3, 76.5, 60.1, 76.8, 75.8, 73.3, 82.0, 80.5, 69.1,~
## $ total_fer  <dbl> 5.82, 1.65, 2.89, 6.16, 2.13, 2.37, 1.69, 1.93, 1.44, 1.92,~
## $ gdpp       <int> 553, 4090, 4460, 3530, 12200, 10300, 3220, 51900, 46900, 58~
```

(ii) Check if there are any NA values in the dataset using the functions `is.na()` and `any()`.

Hint:

- `is.na()`: Returns a logical vector indicating whether each element of the object is NA or not.
- `any()`: Returns TRUE if any element of a logical vector is TRUE, and FALSE otherwise. It's useful for checking if any NA values are present in the dataset.

```
any(is.na(my_data))
```

```
## [1] FALSE
```

(iii) How many unique countries are there in our dataset?

```
length(unique(my_data$country))
```

```
## [1] 167
```

The dataset contains information for 167 unique countries.

(iv) Generate a statistical summary for the numerical variables in the dataset.

```
summary(my_data)
```

```
##    country           child_mort       exports           health
##  Length:167         Min.   :  2.60   Min.   :  0.109   Min.   : 1.810
##  Class :character   1st Qu.:  8.25   1st Qu.: 23.800   1st Qu.: 4.920
##  Mode  :character   Median : 19.30   Median : 35.000   Median : 6.320
##                     Mean   : 38.27   Mean   : 41.109   Mean   : 6.816
##                     3rd Qu.: 62.10   3rd Qu.: 51.350   3rd Qu.: 8.600
##                     Max.   :208.00   Max.   :200.000   Max.   :17.900
##     imports            income         inflation        life_expec
##  Min.   :  0.0659   Min.   :   609   Min.   : -4.210   Min.   :32.10
##  1st Qu.: 30.2000   1st Qu.:  3355   1st Qu.:  1.810   1st Qu.:65.30
##  Median : 43.3000   Median :  9960   Median :  5.390   Median :73.10
##  Mean   : 46.8902   Mean   : 17145   Mean   :  7.782   Mean   :70.56
##  3rd Qu.: 58.7500   3rd Qu.: 22800   3rd Qu.: 10.750   3rd Qu.:76.80
##  Max.   :174.0000   Max.   :125000   Max.   :104.000   Max.   :82.80
##    total_fer          gdpp
##  Min.   :1.150   Min.   :   231
##  1st Qu.:1.795   1st Qu.:  1330
##  Median :2.410   Median :  4660
##  Mean   :2.948   Mean   : 12964
##  3rd Qu.:3.880   3rd Qu.: 14050
##  Max.   :7.490   Max.   :105000
```

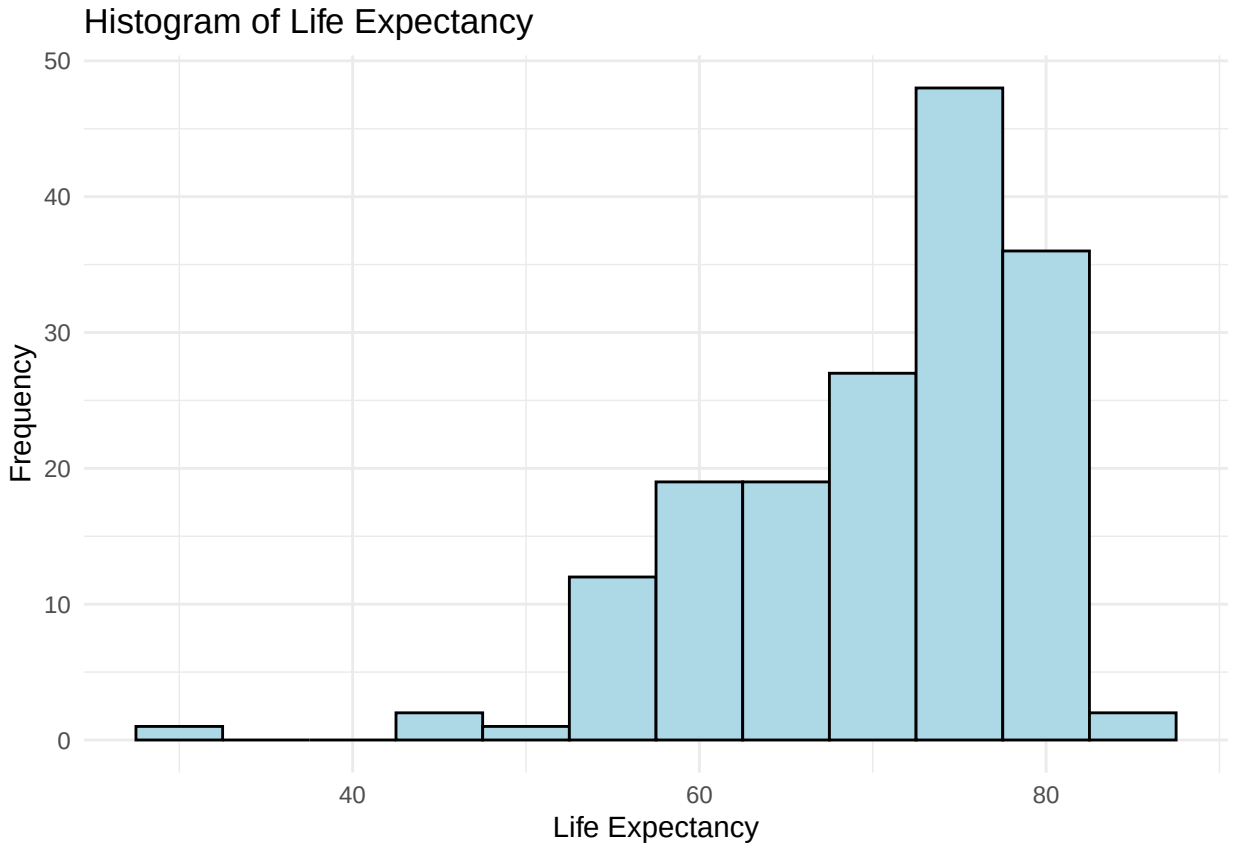
(v) What are the minimum, maximum, mean, and median values of the death of children under 5 years of age per 1000 live births for the countries in the dataset?

- **Min**: 2.60
- **Max**: 208.00
- **Mean**: 38.27
- **Median**: 19.30

2. Data Visualizations

(i) Create a histogram in lightblue color to visualize the distribution of life expectancy in the dataset. Make sure you generate a plot with clear title and axis label.

```
ggplot(my_data, aes(x = life_expec)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  theme_minimal() +
  labs(title = "Histogram of Life Expectancy", x = "Life Expectancy", y = "Frequency")
```



- `binwidth`: Specifies the width of the bins used to group the data. In this case, binwidth = 5 means that the data will be grouped into bins of width 5 units along the x-axis.
- `theme_minimal()`: Sets the overall visual appearance or theme of the plot to a minimalistic style. This theme removes background grid lines, reduces non-data ink, and provides a clean and simple look to the plot.

(ii) Briefly describe the distributions you observed from the generated histogram.

From the histogram, the life expectancy across the 167 countries appears to be normally distributed with a slight left skew. Most of the data points are concentrated around the 70-80 year range. Additionally, there are some occurrences of life expectancy at the extremes, below 40 and above 90, which could represent outlier populations or rare instances.

(iii) Create a scatter plot to visualize the correlation between life expectancy and net income per person. Again, make sure you generate a plot with clear title and axis label.

```
ggplot(my_data, aes(x = life_expec, y = income)) +
  geom_point() +
  labs(title = "Correlation between Life Expectancy and Income Level",
       x = "Life Expectancy",
       y = "Net Income")
```

(iv) What can you infer from the scatter plot?

From the scatter plot, we could observe a positive correlation between life expectancy and income level. The distribution of data points is less dense at higher income levels, implying fewer instances of extremely high incomes in the dataset. Additionally, a notable concentration of data points can be seen at the upper end of both life expectancy and income, suggesting that the relationship between the two may be more significant among higher income ranges.

3. PCA Analysis

We have observed that the dataset contains a total of 9 development indices. To reduce dimensionality and identify the underlying patterns in these indices, we will perform Principal Component Analysis (PCA).

(i) Before conducting PCA, create a new dataset named 'pca_data' by removing the categorical variable 'country' to ensure that the PCA is applied only to the numerical development indices.

```
pca_data <- my_data %>% select(-country)
```

(ii) Next, proceed to perform PCA on the modified dataset, ensuring that the covariates are appropriately scaled.

Hint: use the R function `prcomp()`

```
pca <- prcomp(pca_data, scale. = TRUE)
```

(iii) Generate a summary for the PCA analysis results. How many principle components should we choose if we want to explain over 85% variance of the covariates?

```
summary(pca)
```

```
## Importance of components:
##                            PC1    PC2    PC3    PC4    PC5     PC6    PC7
## Standard deviation     2.0336 1.2435 1.0818 0.9974 0.8128 0.47284 0.3368
## Proportion of Variance 0.4595 0.1718 0.1300 0.1105 0.0734 0.02484 0.0126
## Cumulative Proportion  0.4595 0.6313 0.7614 0.8719 0.9453 0.97015 0.9828
##                            PC8     PC9
## Standard deviation     0.29718 0.25860
## Proportion of Variance 0.00981 0.00743
## Cumulative Proportion  0.99257 1.00000
```

Based on the PCA results, we will select the first four principal components because they account for a cumulative 87.19% of the variance in the covariates.

(iv) Extract the loadings for the first five principal components from the PCA results to analyze the contribution of each variable to these components.

```
loadings <- pca$rotation[, c(1,2,3,4,5)]
print(loadings)
```

```
##                    PC1          PC2         PC3          PC4         PC5
## child_mort -0.4195194 -0.192883937  0.02954353  0.370653262 -0.16896968
## exports     0.2838970 -0.613163494 -0.14476069  0.003091019  0.05761584
## health      0.1508378  0.243086779  0.59663237  0.461897497  0.51800037
## imports     0.1614824 -0.671820644  0.29992674 -0.071907461  0.25537642
## income      0.3984411 -0.022535530 -0.30154750  0.392159039 -0.24714960
## inflation  -0.1931729  0.008404473 -0.64251951  0.150441762  0.71486910
## life_expec  0.4258394  0.222706743 -0.11391854 -0.203797235  0.10821980
## total_fer  -0.4037290 -0.155233106 -0.01954925  0.378303645 -0.13526221
## gdpp        0.3926448  0.046022396 -0.12297749  0.531994575 -0.18016662
```

(v) Analyze the loadings for the first two principal components, PC1 and PC2. Describe how the covariates contribute to these principal components.

For PC1, the positive loadings on income, life_expec, and gdpp indicate that higher scores on this component are associated with better economic status and demographic health, reflecting higher levels of development. In contrast, child_mort and total_fer have notable negative loadings, showing that higher values in these indicators, which typically reflect lower development levels, correspond to lower scores on PC1. This may suggest that PC1 differentiates countries based on their development status, with economic prosperity and demographic health being key distinguishing features.

For PC2, the significant negative loadings on exports and imports highlight a trade volume factor, where countries with higher volumes of trade score lower on PC2. Additionally, The relatively smaller negative loadings on income and minimal positive loadings on inflation and gdpp, show that PC2 reflects aspects not directly related to the level of economic wealth or population health emphasized by PC1. Instead, PC2 appears to capture differences in countries' trade activities.

4. K-Means Clustering

Having applied PCA to extract the socio-economic and health-related indicators into principal components that capture the essence of country development, we now move to the next important section of our analysis: K-Means Clustering. This algorithm will enable us to group countries into clusters based on their similarities across the principal components identified earlier. By doing so, we can categorize countries into distinct groups, reflecting varying levels of need and development. This clustering will directly inform HELP International's strategic decision-making process, allowing for targeted aid allocation where it's most required.

(i) First, extract the principal component scores for the first four principle components we identified in the previous section.

Hint: To extract these scores for the first four principal components from your PCA results, you can use the `$x` component of the object returned by `prcomp()` and select only the first four columns.

```
km_data <- pca$x[, 1:4]
```

Note: These scores represent the coordinates of each observation (in our case, each country) in the new space defined by the principal components and are what we'll use as input for the K-Means clustering. Put simply, we can treat these four scores as four different variables derived from PCA, encapsulating the information of the initial 9 covariates in a new vector space.

(ii) Perform K-Means clustering on the dataset containing the principal component scores for the first four components. Initiate the algorithm with 30 starting points and target 3 cluster centers.

```
km_model <- kmeans(km_data, centers = 3 , nstart = 30)
```

(iii) Create a new variable within the original dataset 'my_data', labeling each observation with the cluster it belongs to.

```
my_data$cluster <- km_model$cluster
```

(iv) Compare the clusters by generating a summary including:

- the count of countries in each cluster
- the average net income level for each cluster
- the average child mortality rate for each cluster
- the average life expectancy level for each cluster
- the average health expenditure for each cluster
- the average GDP per capita for each cluster

```
cluster_sum <- my_data %>%
  group_by(cluster) %>%
  summarise(cnt = n(),
            income = mean(income),
            child_mort = mean(child_mort),
            life_expec = mean(life_expec),
            health = mean(health),
            gdpp = mean(gdpp))
print(cluster_sum)
```

```
## # A tibble: 3 x 7
##   cluster   cnt income child_mort life_expec health   gdpp
##     <int> <int>  <dbl>      <dbl>      <dbl>  <dbl>  <dbl>
## 1       1    35 45803.       4.90       80.2   8.92 43117.
## 2       2    84 12774.      21.7        73.0   6.16  6718.
## 3       3    48  3897.      91.6        59.2   6.43  1909.
```

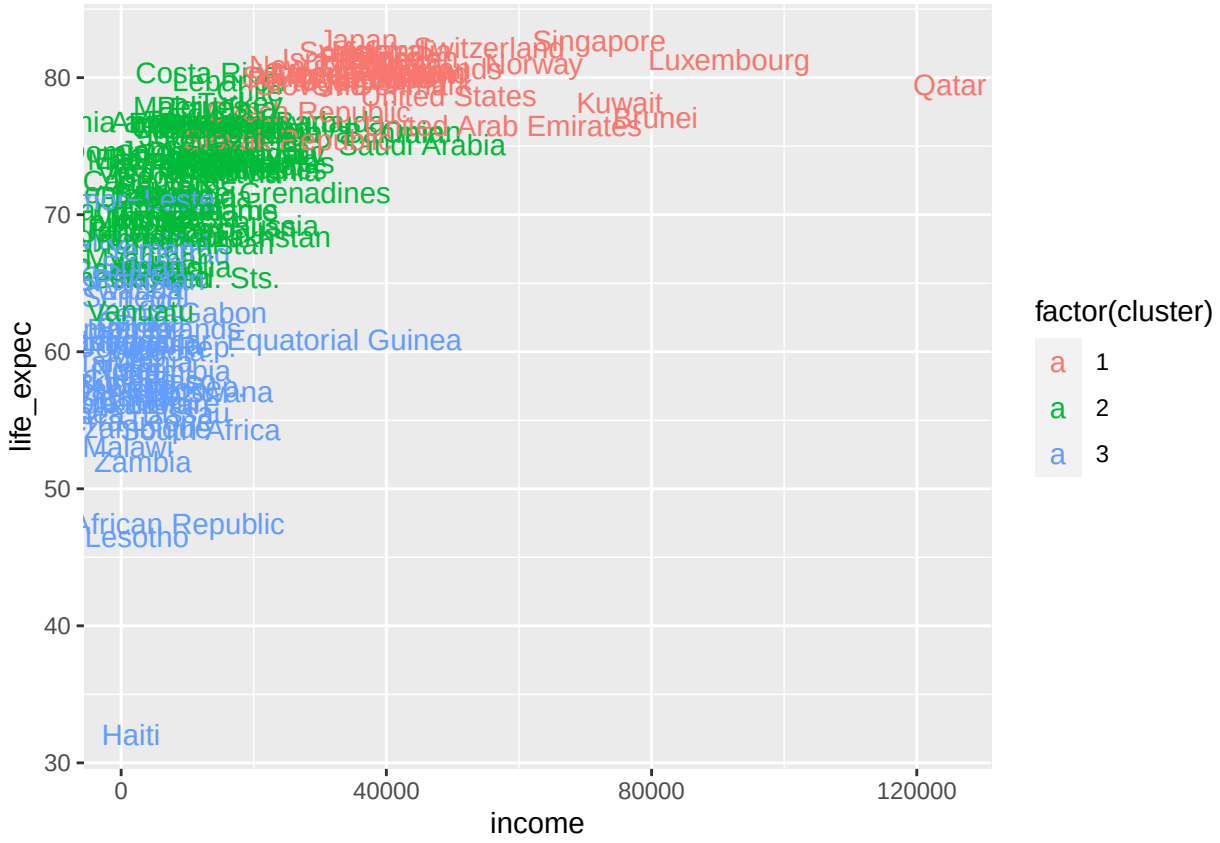
(v) How would you inform the NGO about the countries most in need of aid based on the summary generated in (iv)?

The 48 countries in Cluster 3 stand out as those most in need of aid due to several key factors. Firstly, it exhibits significantly lower income and GDP levels compared to the other clusters, indicating economic vulnerability. Moreover, the cluster is characterized by a notably higher child mortality rate, suggesting inadequate healthcare and living conditions for children. This is compounded by lower life expectancy and a lower expenditure on healthcare, indicating poorer overall health outcomes and limited access to healthcare services within these countries. Therefore, it is imperative for the organization to carefully consider the countries within Cluster 3 and emphasize the urgent need for targeted assistance and intervention to address these socio-economic challenges and improve the well-being of their populations.

(vi) Bonus: Run the provided code to generate a scatter plot visualizing clusters across two dimensions:

income and life expectancy. Based on the plot, which country do you believe shows the most promising development, and which countries do you think are in the most urgent need of assistance?

```
my_data %>%
  as_tibble() %>%
  mutate(cluster = my_data$cluster,
         state = row.names(my_data)) %>%
  ggplot(aes(income, life_expec, color = factor(cluster), label = country)) +
  geom_text()
```



From the plot, it's clear that the results align with our previous analysis. The 48 countries highlighted in blue are lagging behind in terms of life expectancy and net income per person compared to other countries. Among these, Haiti, Lesotho, and the Central African Republic require special attention, as these three countries are significantly behind in both life expectancy and net income per person. Meanwhile, countries in cluster 1, especially Luxembourg and Qatar, are showing a more advanced development trend.