Case Study 3 - PCA and K-Means Clustering

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.

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

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

B. Data Cleanings

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

(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.

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

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

(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?

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.

- `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.

(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.

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

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.

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

Hint: use the R function `prcomp()`

(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?

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

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

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.

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.

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

(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

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

(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?