Mina Dao & Yerim Oh

K-Means Clustering

STAT-340 Applied Regression Methods

Table of Contents

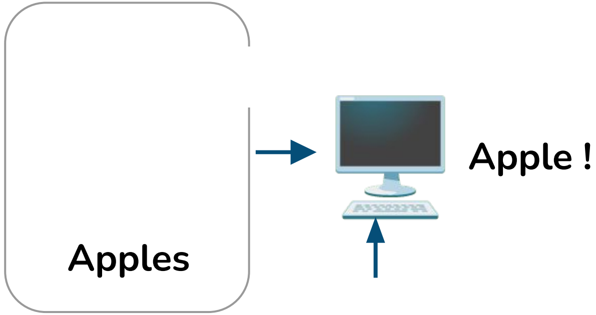
- Introduction
- Method: K-means Clustering
- Illustration
- Discussion
- Reference

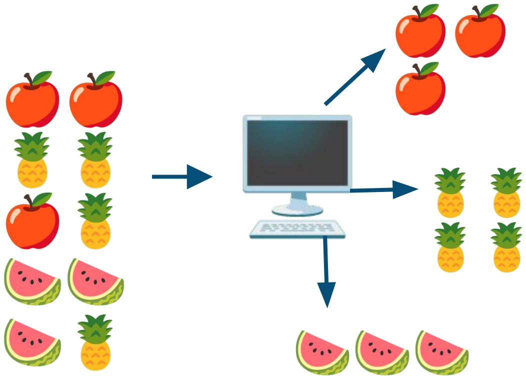
Introduction

Unsupervised Learning

: a group of statistical tools for the setting where we find groups of observations (clusters) that share similar characteristics to analyze unlabeled data without human intervention

- Has no response variable





Introduction

K-means Clustering

: clustering methods that divides a data set into K distinct, non-overlapping clusters

- assigns data points to one of the K clusters depending on their distance from the center of the clusters

Method

Way to choose K

: The equation for the within-point scatter

$$\begin{aligned} W(C) &= \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}||x_i - x_{i'}||^2 \\ &= \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x}_k||^2, \end{aligned}$$

Elbow Method: As k increases, W decreases. We look for the “elbow” point, which is the point where the rate of decrease in W starts to level off.

Method

Algorithm

: The K-means clustering algorithm works through an iterative process to find the optimal placement of cluster centroids

Method

Algorithm

1. Randomly choose K centroids.
2. Calculate the Euclidean distance between data points and centroids. Assign each data point to the cluster.
3. Recalculate the centroid of each cluster. Update the centroids.
4. Iterate two steps above until the centroids no longer change or for a specified number of iterations.
5. Outputs the final cluster centroids and the assignment of the data points.

Method

Assumptions.

1. Spherical or convex shape
   - The radius of each cluster is approximately equal in all directions.
2. Equal variance
   - The distribution of data points around each cluster center is approximately the same.
3. Similar cluster sizes
   - The number of data points assigned to each cluster is roughly the same.

Method

Difference compared to what we learned in class

: We explored supervised learning - Regression and Classification

- Supervised learning is a type of machine learning technique where the computer learns from past data and applies to the present data to predict future output.
- Unsupervised learning uses machine learning to discover hidden patterns in large volumes of data. It is not used for prediction.

Illustration

Data

2023 Global Country Development & Prosperity Index

- contains rankings and indicators from the 2023 Legatum Prosperity Index
- observations according to 167 countries and 12 numerical variables

Note.

should use numerical data

- identify groups of similar data points based on their proximity to a central point (centroid), as K-means primarily relies on calculating distances between data points to assign them to clusters

Illustration

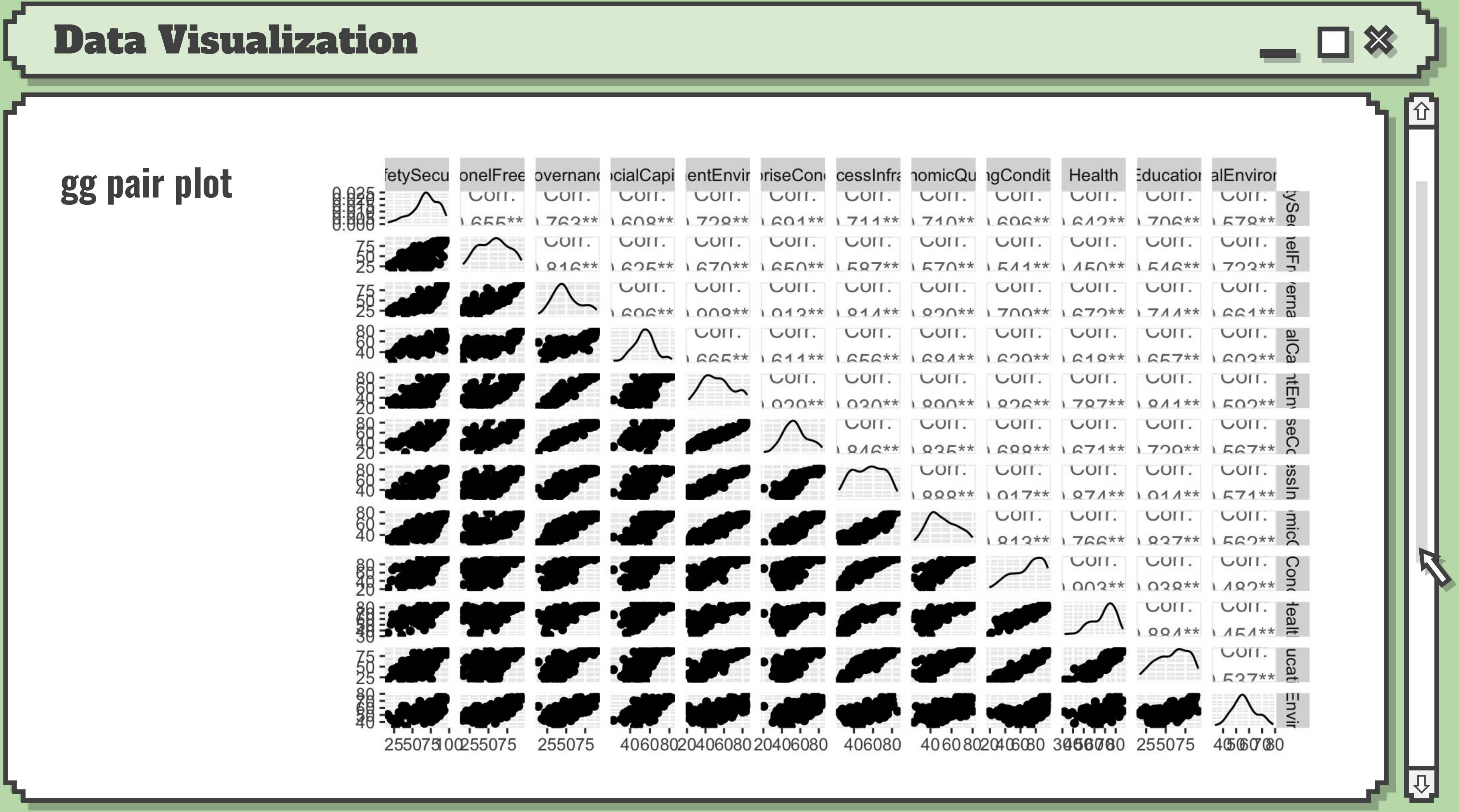
Variables

- Safety Security
- Personal Freedom
- Governance
- Social Capital
- Investment Environment
- Enterprise Conditions
- Market Access Infrastructure
- Economic Quality
- Living Conditions
- Health
- Education
- Natural Environment

Illustration

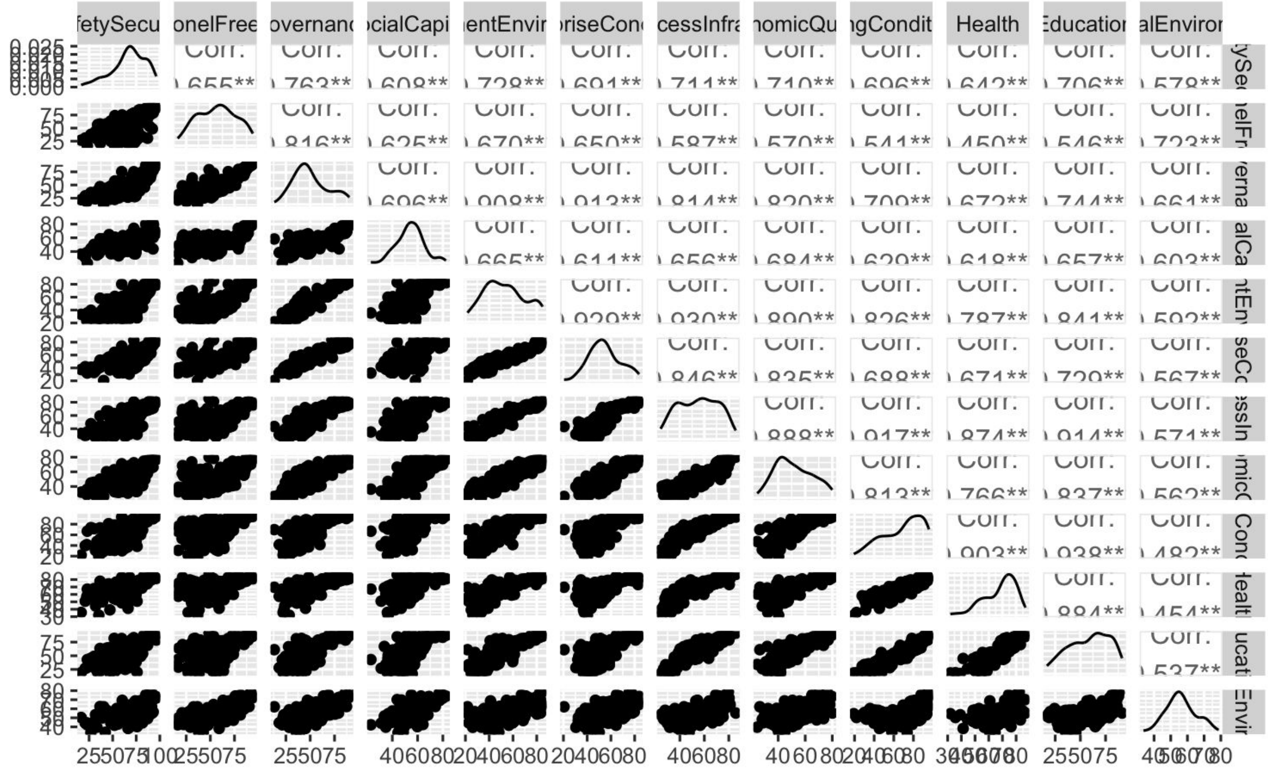
Note.

In K-means clustering, we do not choose a "response variable" because the goal is to group data points based on their similarities across multiple variables, not to predict a single outcome.
Therefore, all the variables used in the analysis are considered "explanatory" and are used to determine cluster assignments, not a single response variable.



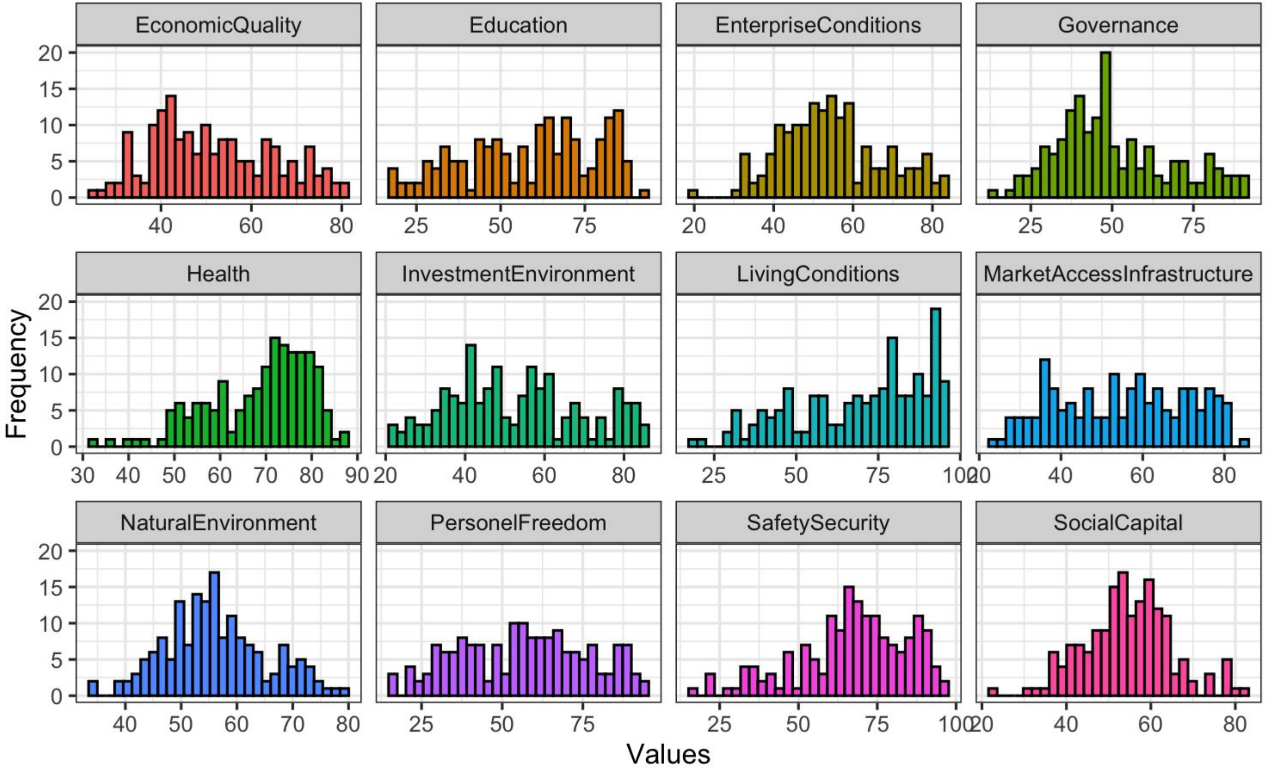
Data Visualization

gg pair plot



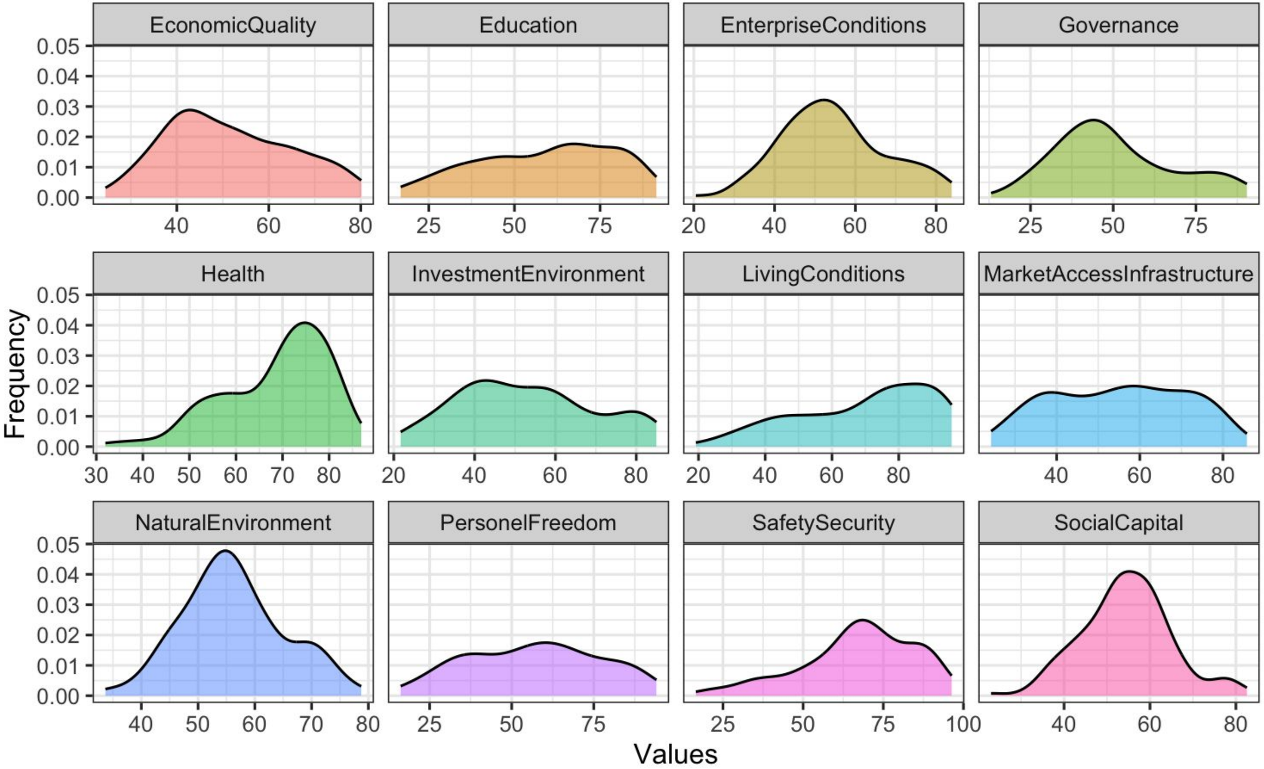
Data Visualization

Histogram



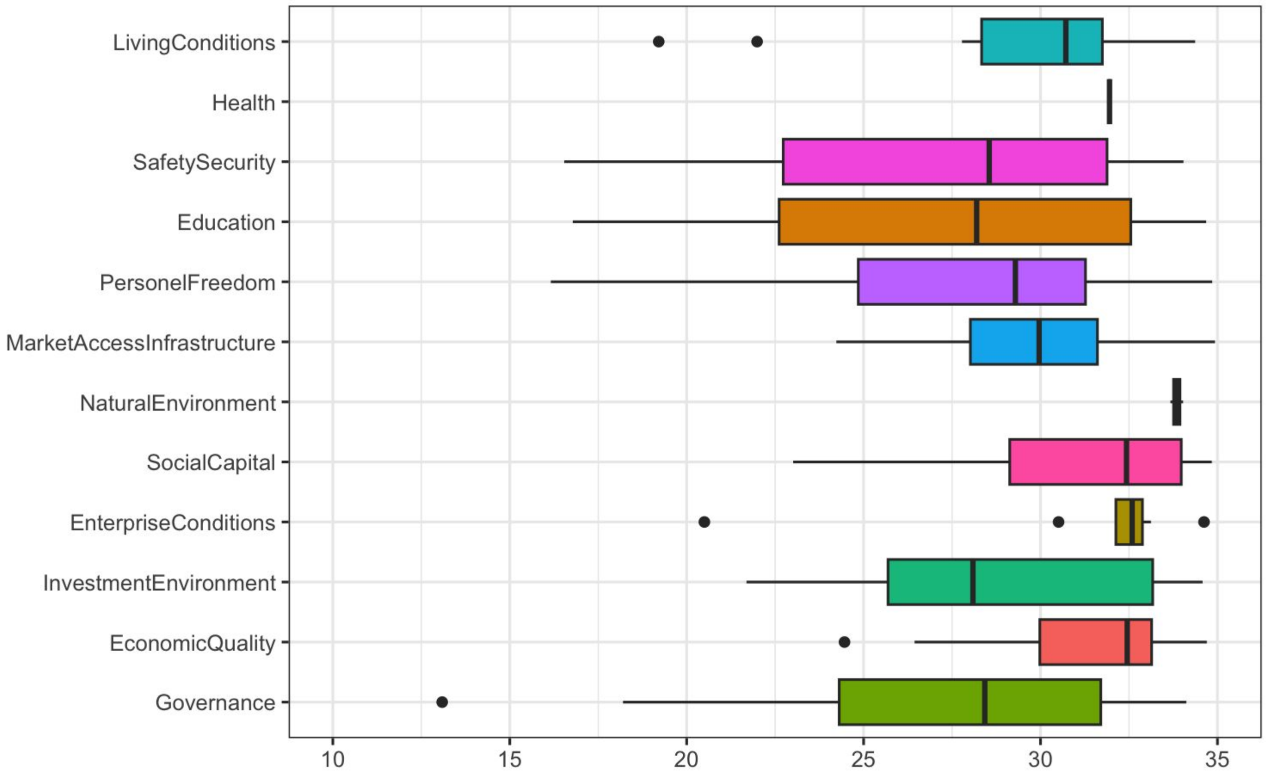
Data Visualization

Density plot



Data Visualization

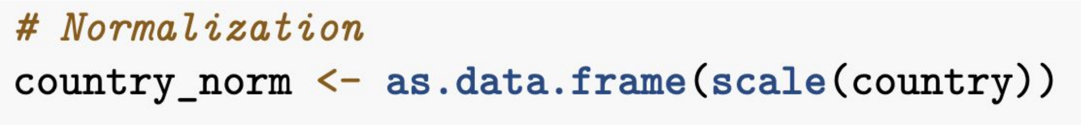
Boxplot



Illustration

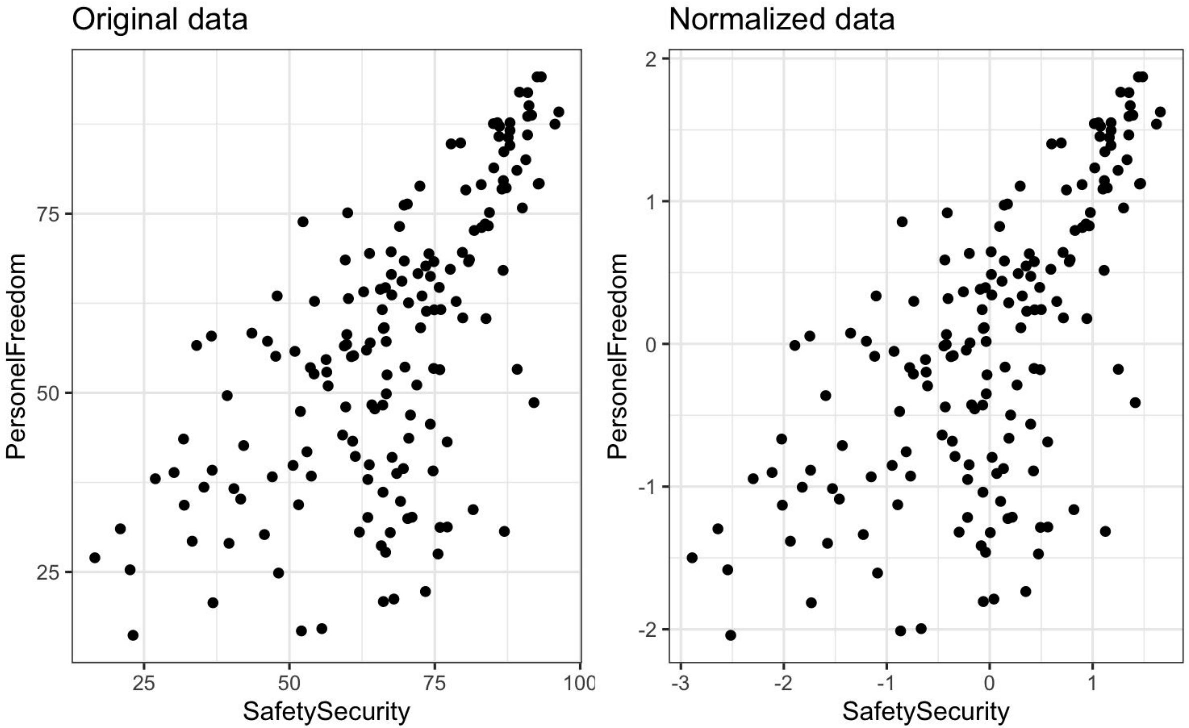
Data Preparation

We have to normalize the variables to express them in the same range of values by adjusting values measured on different scales to a common scale.



```
# Normalization
country_norm <- as.data.frame(scale(country))
```

Illustration



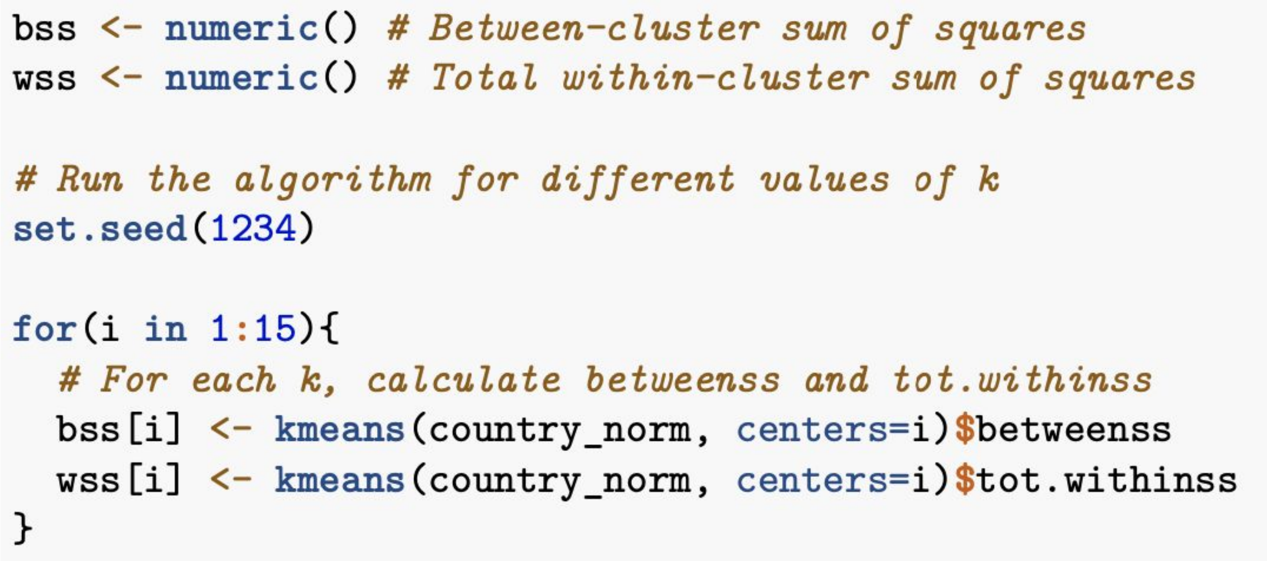
The points in the normalized data are the same as the original one. The only thing that changes is the scale of the axis.

Illustration

Knitr Package

- `cluster`: A vector of integers indicating the cluster to which each point is allocated.
- `centers`: A matrix of cluster centers.
- `size`: The number of points in each cluster.
- `betweenss`: The between-cluster sum of squares.
- `withinss`: Vector of within-cluster sum of squares, one component per cluster.
- `tot.withinss`: Total within-cluster sum of squares.
- `totss`: The total sum of squares.

Illustration

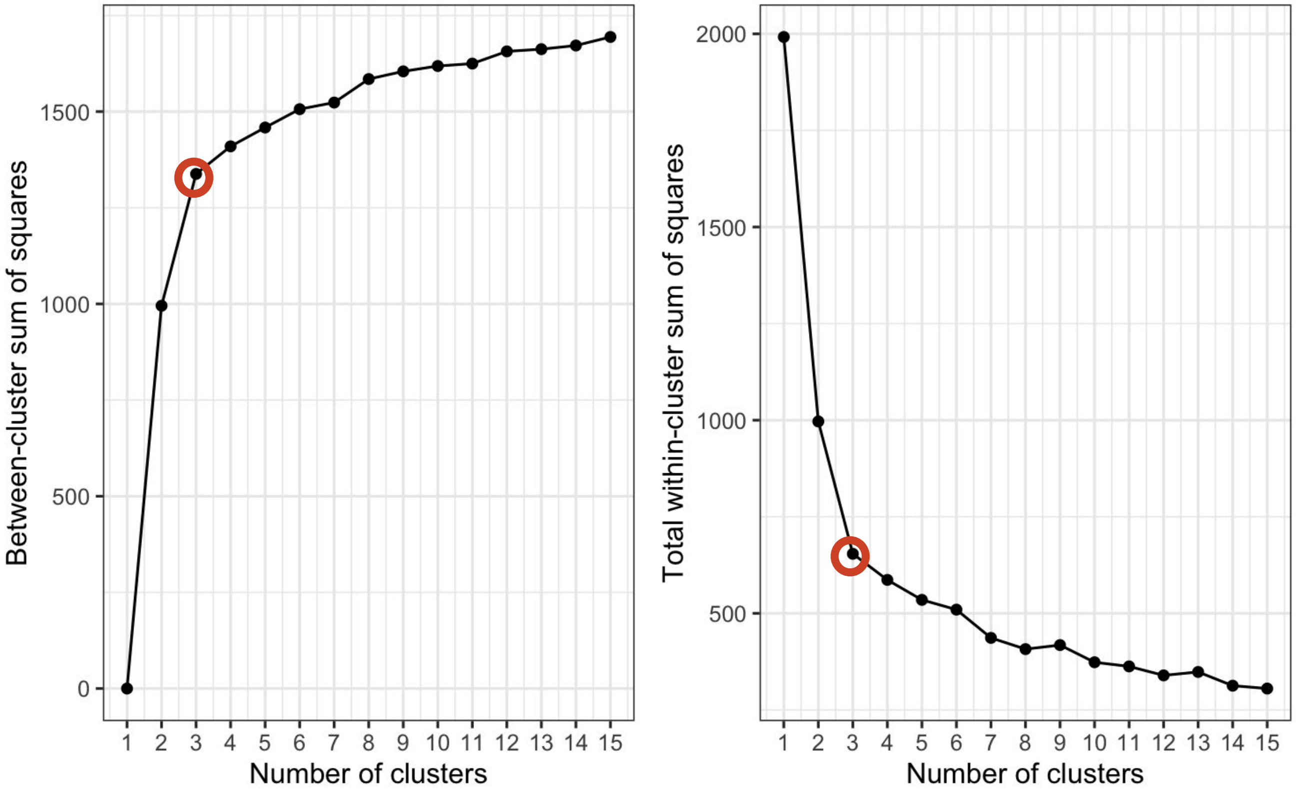


```r
bss <- numeric() # Between-cluster sum of squares
wss <- numeric() # Total within-cluster sum of squares

# Run the algorithm for different values of k
set.seed(1234)

for(i in 1:15){
  # For each k, calculate betweenss and tot.withinss
  bss[i] <- kmeans(country_norm, centers=i)$betweenss
  wss[i] <- kmeans(country_norm, centers=i)$tot.withinss
}
```

Note.

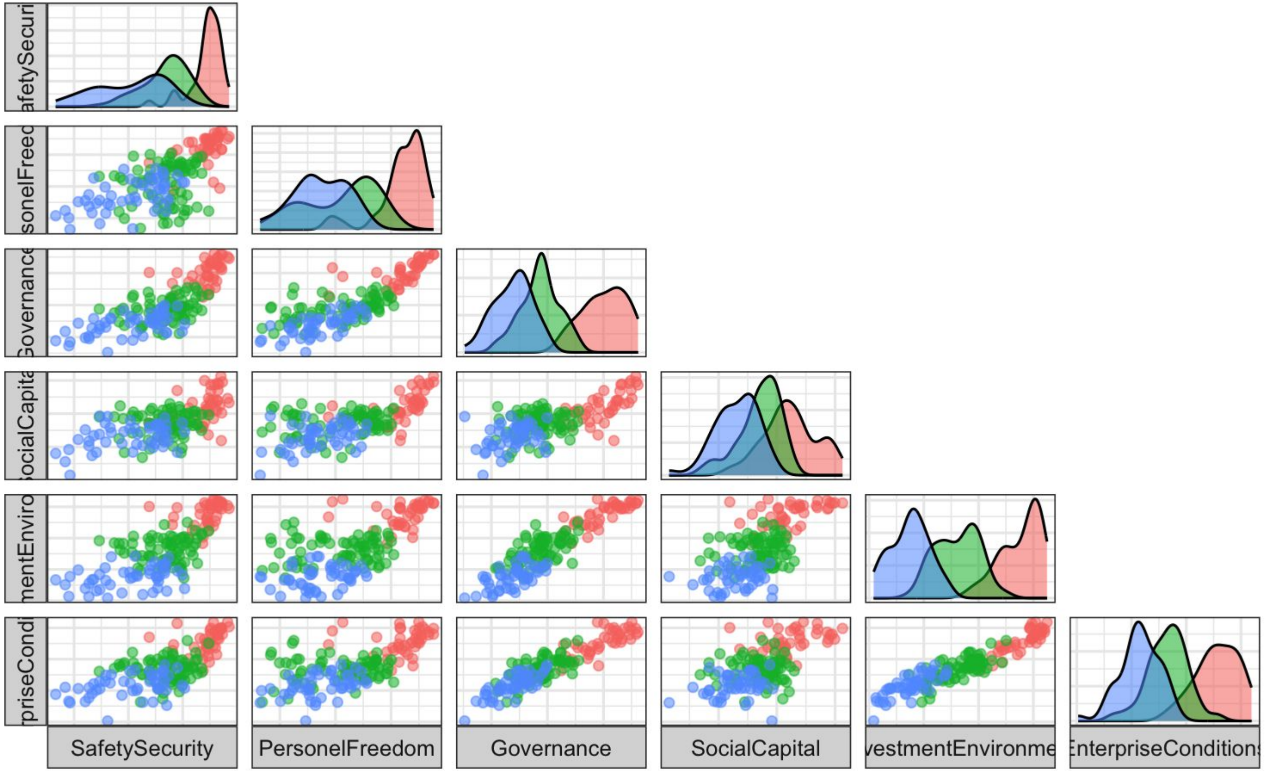
In unsupervised learning, such as K-means clustering, there is usually no clear definition of error. Due to this, also cross-validation cannot be used for this purpose.

Illustration

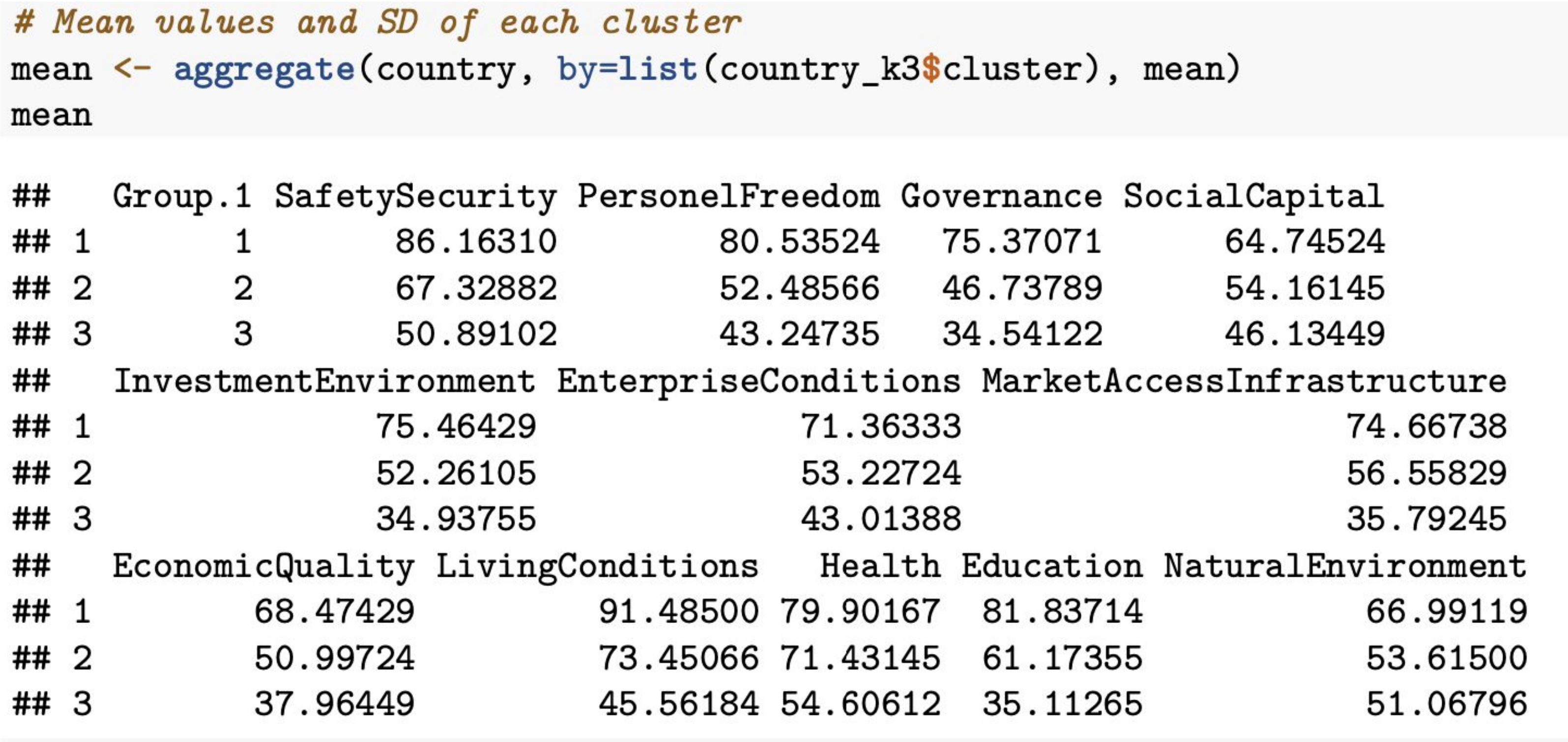


Result

gg pair plot



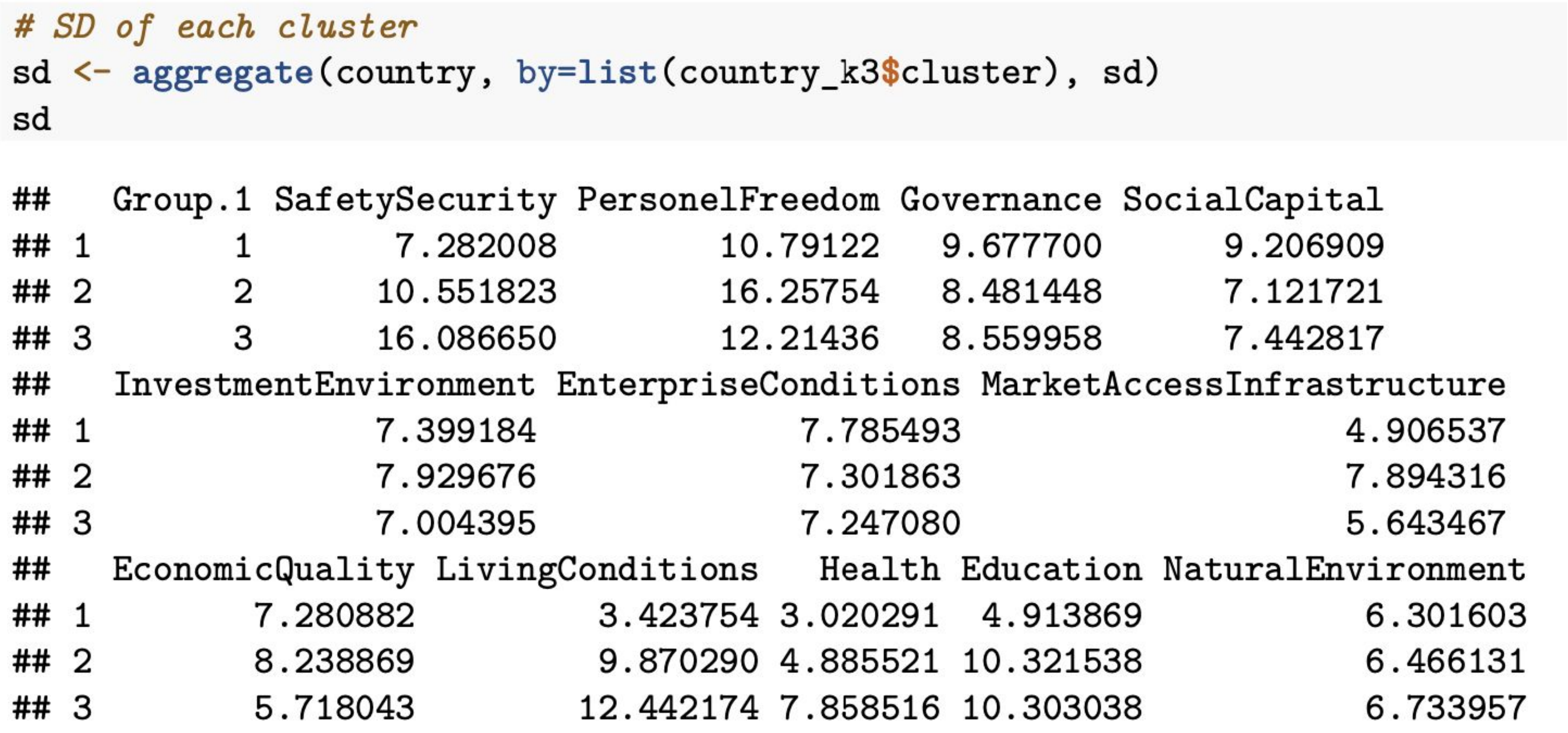
Result



```
# Mean values and SD of each cluster
mean <- aggregate(country, by=list(country_k3$cluster), mean)
mean
```

```
##   Group.1 SafetySecurity PersonelFreedom Governance SocialCapital
## 1       1       86.16310        80.53524   75.37071      64.74524
## 2       2       67.32882        52.48566   46.73789      54.16145
## 3       3       50.89102        43.24735   34.54122      46.13449
##   InvestmentEnvironment EnterpriseConditions MarketAccessInfrastructure
## 1              75.46429             71.36333                   74.66738
## 2              52.26105             53.22724                   56.55829
## 3              34.93755             43.01388                   35.79245
##   EconomicQuality LivingConditions   Health Education NaturalEnvironment
## 1        68.47429         91.48500 79.90167  81.83714           66.99119
## 2        50.99724         73.45066 71.43145  61.17355           53.61500
## 3        37.96449         45.56184 54.60612  35.11265           51.06796
```

Result



```r
# SD of each cluster
sd <- aggregate(country, by=list(country_k3$cluster), sd)
sd
```

```
##   Group.1 SafetySecurity PersonelFreedom Governance SocialCapital
## 1       1       7.282008        10.79122   9.677700      9.206909
## 2       2      10.551823        16.25754   8.481448      7.121721
## 3       3      16.086650        12.21436   8.559958      7.442817
##   InvestmentEnvironment EnterpriseConditions MarketAccessInfrastructure
## 1              7.399184             7.785493                   4.906537
## 2              7.929676             7.301863                   7.894316
## 3              7.004395             7.247080                   5.643467
##   EconomicQuality LivingConditions   Health Education NaturalEnvironment
## 1        7.280882         3.423754 3.020291  4.913869           6.301603
## 2        8.238869         9.870290 4.885521 10.321538           6.466131
## 3        5.718043        12.442174 7.858516 10.303038           6.733957
```

Result

Cluster 1 - Developed Nations

- high average scores in most dimensions, indicating well-developed and prosperous countries
- lower standard deviation values indicate that countries in this cluster are more homogenous, with similar high scores across various dimensions

Result

Cluster 2 - Developing Nations

- countries with moderate scores across most dimensions
- standard deviation values suggest moderate variability within this cluster, with countries having consistent scores across different dimensions

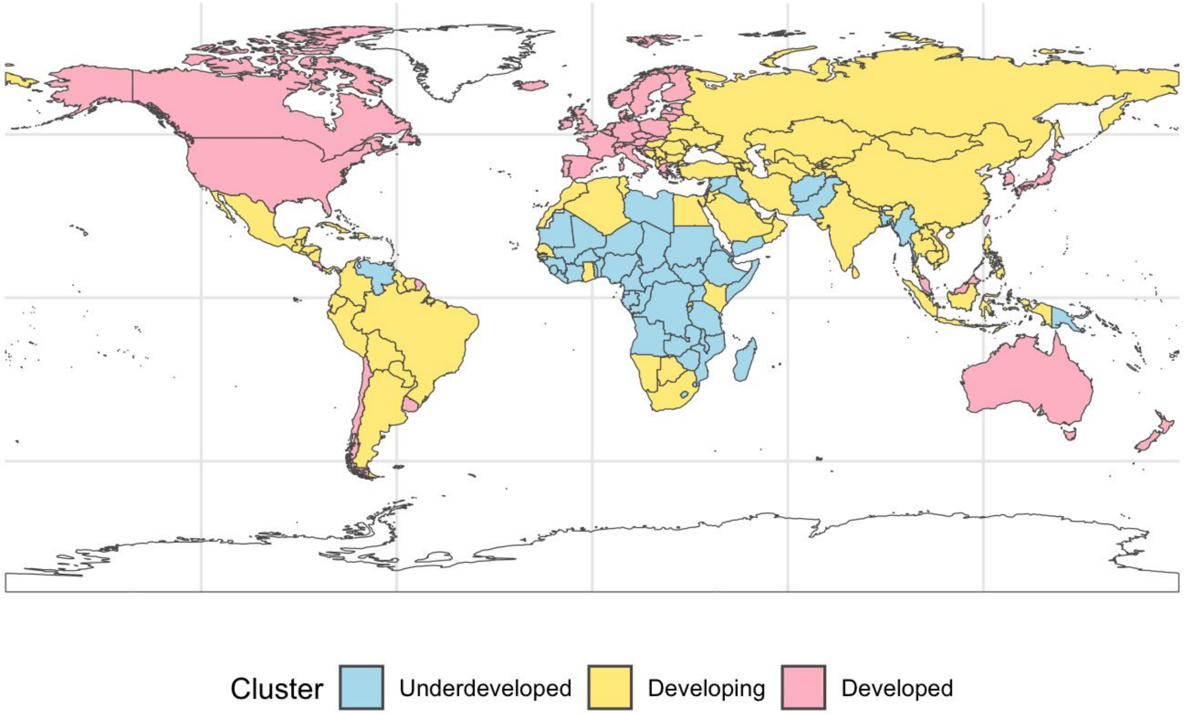
Result

Cluster 3 - Underdeveloped Nations

- this cluster exhibit lower scores across most dimensions compared to Cluster 1 and 2
- higher standard deviation values in this cluster indicate greater variability among countries, highlighting significant differences within the group

Result

K-Means Clustering Results on World Map



Source: K-Means Clustering of Global Prosperity Data

Discussion

Advantages

1. Simplicity

   Easy to implement and identify unknown groups of data from complex datasets

2. Efficiency

   Linear time complexity, can handle large datasets conveniently

3. Flexibility

   Can easily adjust to changes. An instance can change the cluster.

Discussion

Disadvantages

1. Difficulty in determining the optimal number of clusterings
2. Sensitivity to initial conditions
3. Sensitivity to outliers
4. Limited to linear boundaries

Discussion

Applications

1. image segmentation
2. market segmentation
3. customer segmentation
4. anomaly detection

Reference

R package

- Xie, Y. (2024, November 8). A general-purpose package for dynamic report generation in R [R package knitr version 1.49]. The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/knitr/index.html
- Massicotte, P. (2023, December 15). World map data from Natural Earth [R package rnaturalearth version 1.0.1]. The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/rnaturalearth/index.html

Data

- Taşaltı, T. T. (2024, June 29). 2023 Global Country Development & Prosperity index. Kaggle. https://www.kaggle.com/datasets/tarktunataalt/2023-global-country-development-and-prosperity-index

Reference

Other

- Xvivancos. (2024, November 16). Tutorial: Clustering wines with K-means. Kaggle. https://www.kaggle.com/code/xvivancos/tutorial-clustering-wines-with-k-means/report
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). K-means Clustering. In An introduction to statistical learning : with applications in R (2nd ed., pp. 515–519). essay, Springer. Retrieved December 1, 2024,.
- Eric J. K-Means Clustering: 7 Pros and Cons Uncovered. Datarundown. https://datarundown.com/k-means-clustering-pros-cons/
- Juliana D. (2021, March 12). Supervised versus unsupervised learning: What's the difference?. IBM https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning
- parvezcs20. (2023, Dec 09). Demonstration of K-Means Assumptions. Geeksforgeeks https://www.geeksforgeeks.org/demonstration-of-k-means-assumptions/