Gapminder Exploration: Pricinpal Component Analysis

Gabrielle LaRosa

Overview and Motivation:

The goal of this analysis is to use Principal Component Analysis (PCA) as a dimension reduction technique for the gapminder data. **We are interested in uncovering broad patterns between location (continent) and the health/economic features of a region**. If the health spending and health outcome features of a country is related to it's location, then we expect PCA to uncover that the data is clustered by continent once plotted on it's first and second principle components.

PCA will allow us to derive a low dimensional feature set from 10 different covariates relating to country's health spending and health status. From there, we can plot a scatterplot of the first two principal components, which presents a better method for visualizing n observations on p features without having p(p-1)/2 scatterplots.

Related Work

This section of our project was inspired by our need for a technique to visualize data with many covariates/features. Thus, online resources and a textbook [Introduction to Statistical Learning] served as an inspiration for using Principal Component Analysis as a dimension reduction technique. In addition, Dr. Mattie's input and her provided old notes on PCA was helpful for inspiring our analyses.

Question

Do countries within the same continent share similar health and economic features?

Analysis: Principal Component Analysis (PCA)

We are using Principal Component Analysis (PCA) to reduce the dimension of the gapminder data and analyze if our new components explain the variation between continents. This is an unsupervised learning technique, meaning that PCA reduces the dimensions of our data without using "continent" information. Then, we can go back to the PCA plot, label points by continent, and see if the data has clustering based on continent. Essentially, this will allow us to better visualize what is going on with our data without having to plot a multiple scatterplots using only 2 variables at a time. Overall, we are trying to reduce the number of variables in the gapminder dataset in order to extract patterns in the data and assess whether the patterns are related to a country's location (continent).

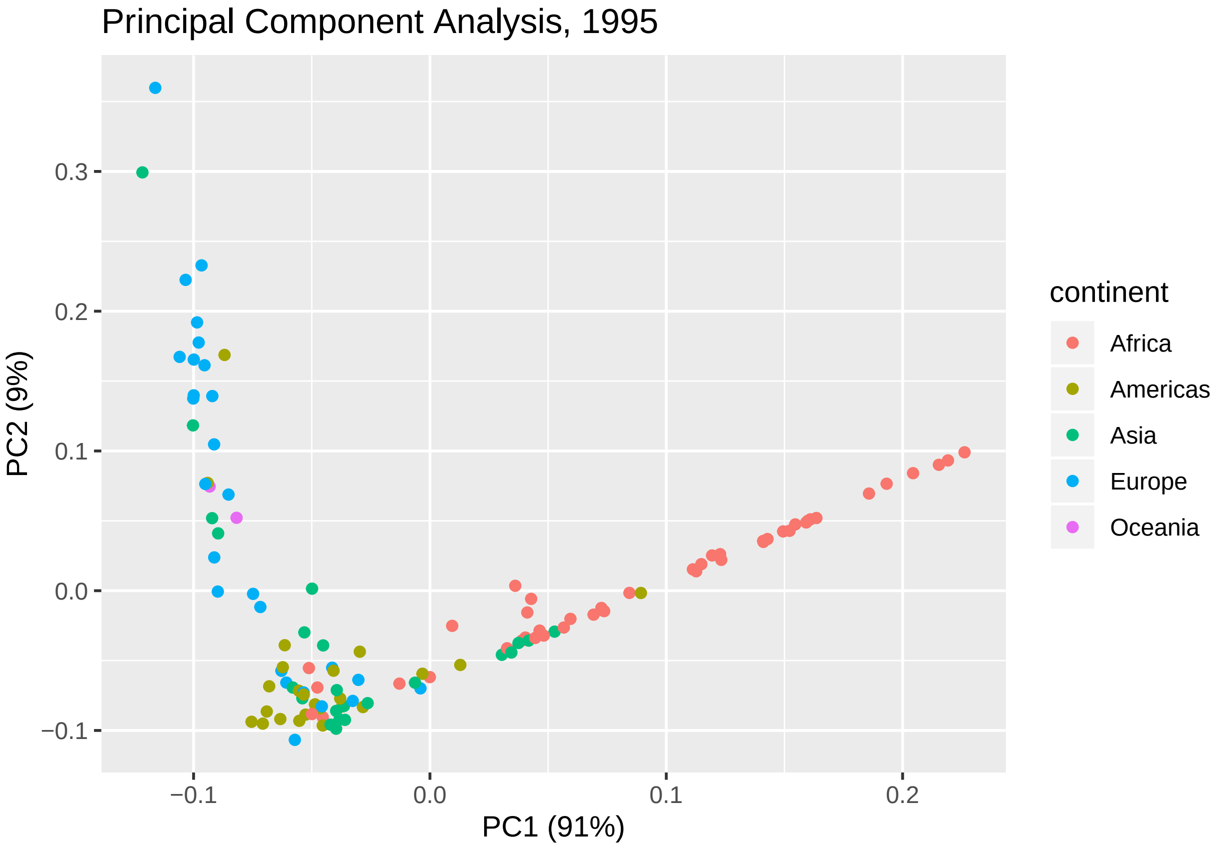
From the gapminder data, we are using the following features that characterize the health and economic status of a country: X_1 = Child Mortality Rate, X_2 = Life Expectancy, X_3 = DALYs, X_4 = GINI Coefficient, X_5 = GDP per capita, X_6 = Child/Elderly Ratio, X_7 = Government Spending (% of total), X_8 = Total Health Spending, X_9 = Government Share of Health Spending, X_{10} = Private Share of Health Spending.

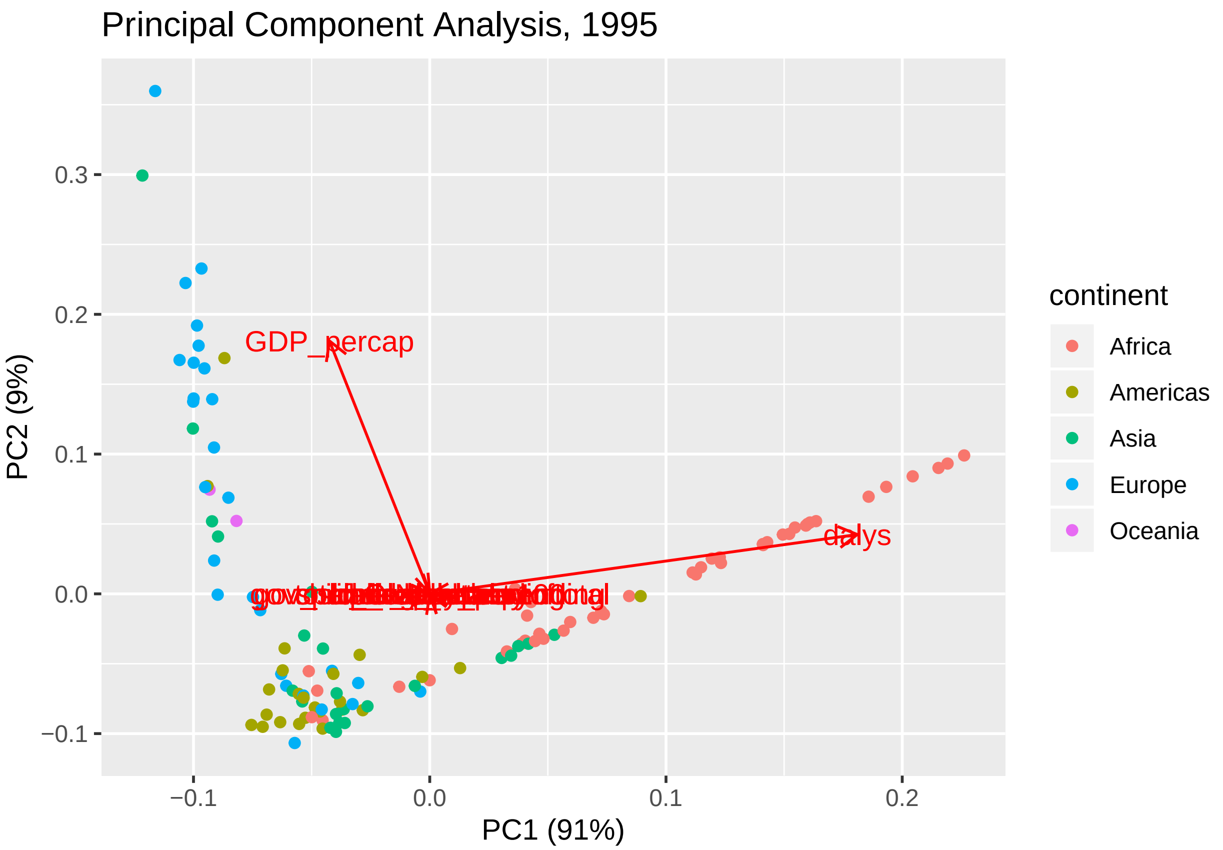
In R, we can implement Principal Component Analysis using the 'prcomp' function from the stats package. We feed the prcomp function a data matrix containing the data information *without* the continent information included. Then, the calculation is done using singular value decomposition on this data matrix and returns the standard deviations of the principal components.

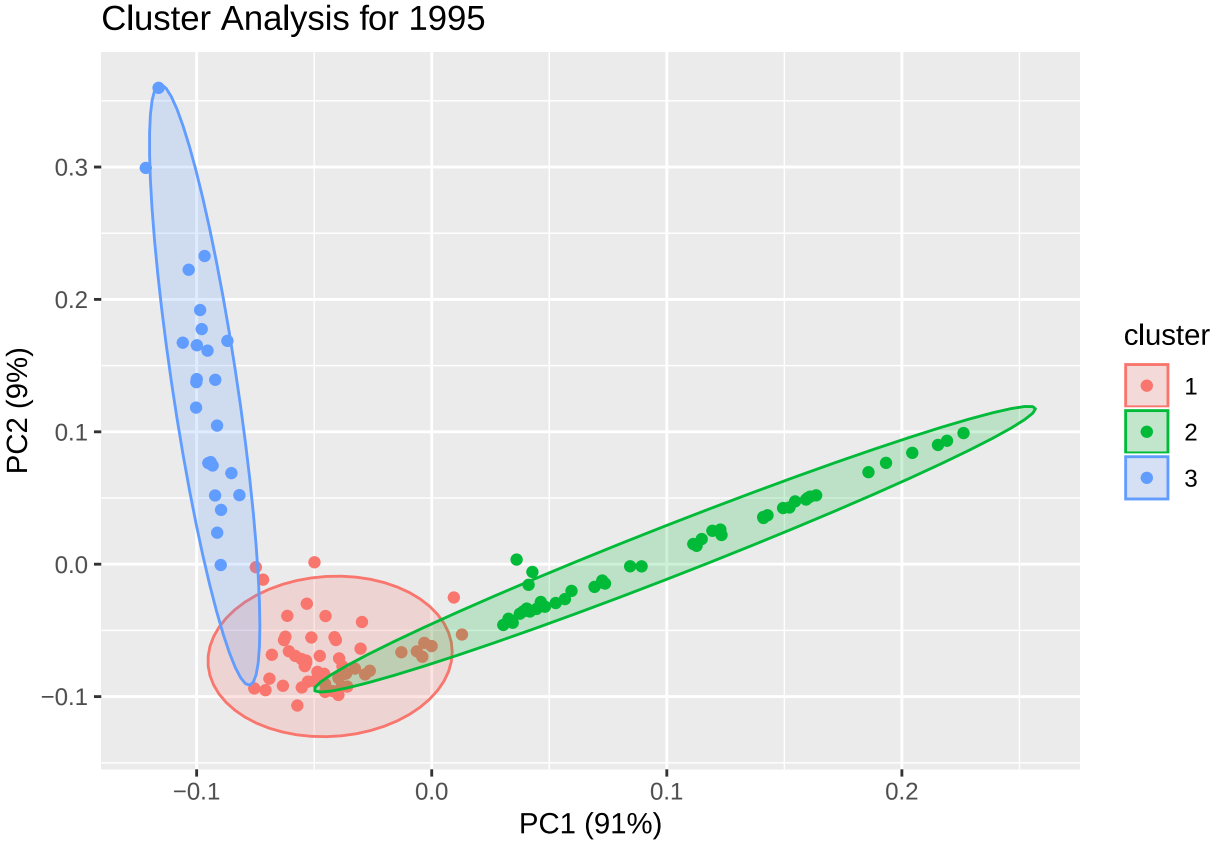
Once prcomp is computed, we can take the information from the first two principal components, where the first principal component explains the largest amount of variability in the data and the second principal component explains the next largest amount of variability. In addition, we can extract the loadings of the

PCA, which are the coefficients of the linear combination of features from the principal component analysis. We will repeat this process for both 1995 and 2010, and examine if different patterns emerge in the data.

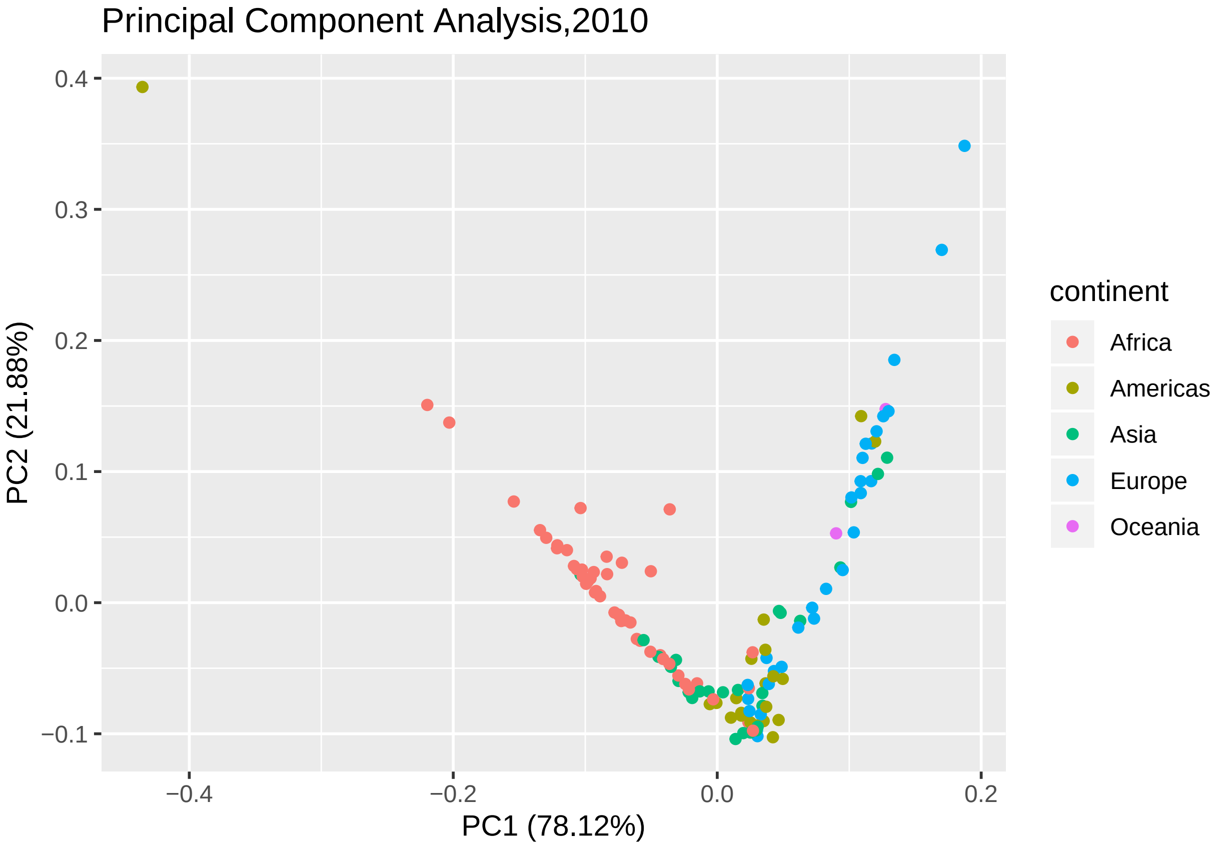
Year 1995 Analysis

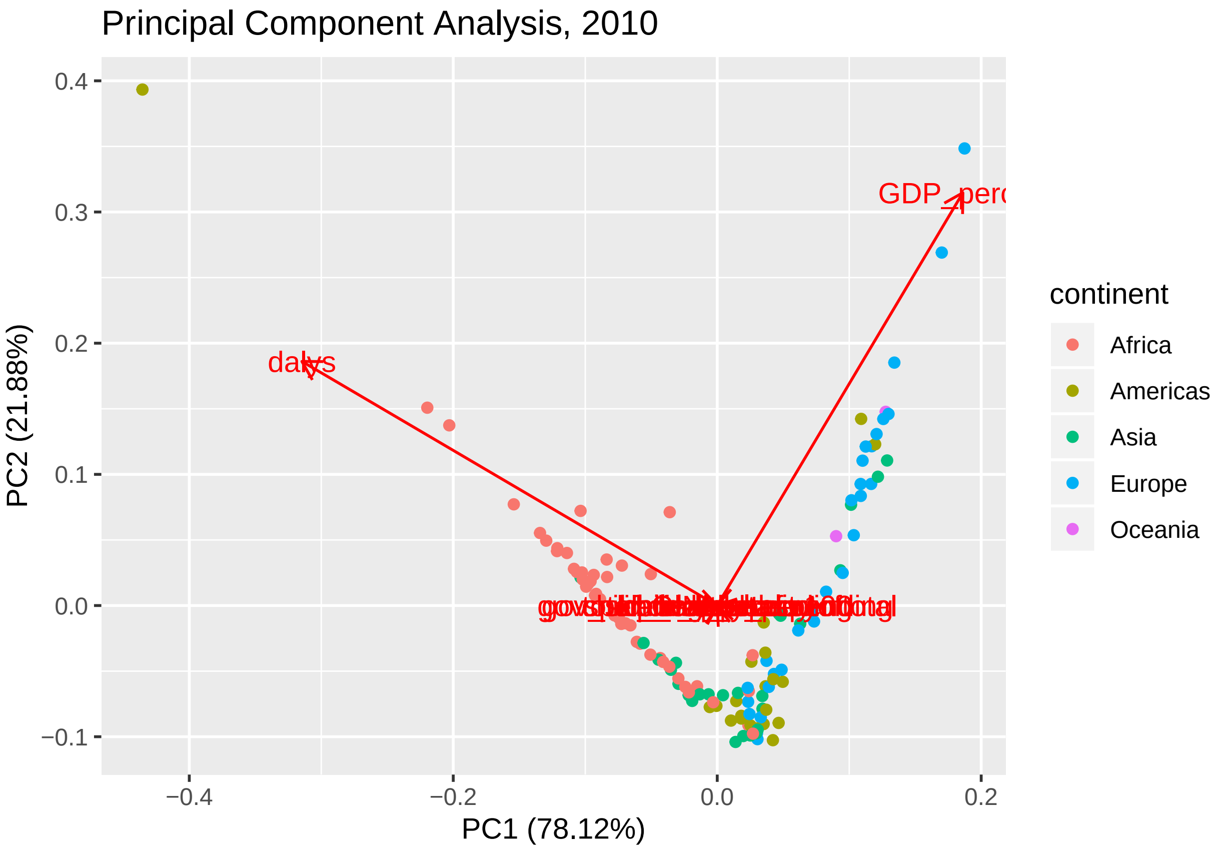


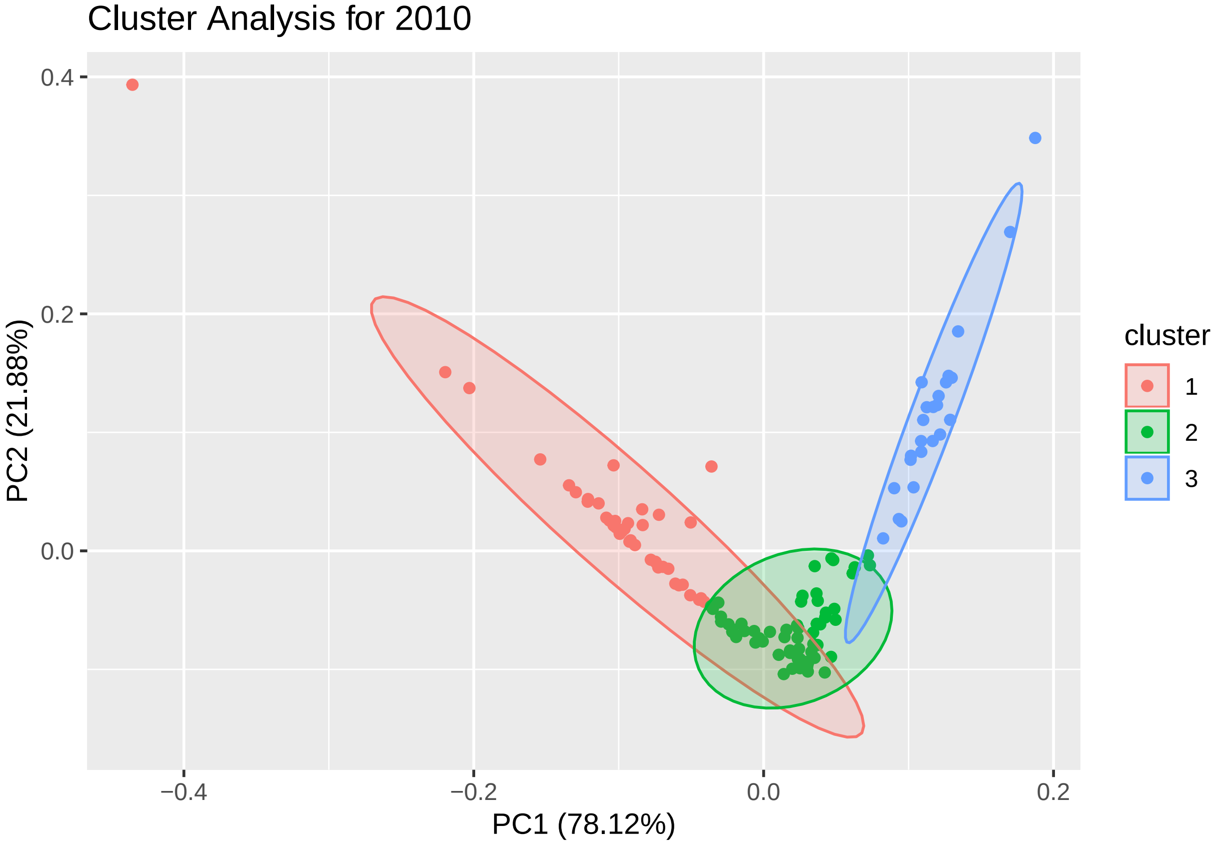

Principal Component Analysis, 1995
PC2 (9%)
0.3
0.2
0.1
0.0
−0.1
−0.1
0.0
0.1
0.2
PC1 (91%)
GDP_percap
dalys
continent
Africa
Americas
Asia
Europe
Oceania


Cluster Analysis for 1995
PC2 (9%)
0.3
0.2
0.1
0.0
−0.1
−0.1
0.0
0.1
0.2
PC1 (91%)
cluster
1
2
3

Year 2010 Analysis




Principal Component Analysis, 2010
PC2 (21.88%)
PC1 (78.12%)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.4
−0.2
0.0
0.2
dalys
GDP_perc
continent
Africa
Americas
Asia
Europe
Oceania


Cluster Analysis for 2010
PC2 (21.88%)
PC1 (78.12%)
0.4
0.2
0.0
−0.4
−0.2
0.0
0.2
cluster
1
2
3

Conclusions

In the Principal Component Analysis for the 1995 gapminder data, we see that the first two principal components explain 100% of the variability of the data, with the first principal component explaining 91% and the second principal component explaining 9%. Thus, it appears the PCA successfully reduced the 1995 data down to 2 dimensions. When examining the PCA plot, we are looking for any distinct clustering by continent. In the 1995 data, we see that Europe and Africa form the two most distinct clusters. However, there are still a few data points from other continents that are not separable from the African or European clusters. In addition, the Americas seem to be forming a slight cluster, with points from Asia and Oceania being more scattered.

There is a similar pattern from the 2010 data, with the first principal component explaining 78.12% of the variability in the data and the second principal component explaining 21.88% of the variability in the data. We see again the pattern of Europe and Africa forming the two most distinct clusters. In addition, both the 1995 and 2010 PCA plots include the loading vectors. This shows us that the DALYs and GDP percent of total spending features are the variables that are defining the clusters the most, since the loadings are the weights of the linear combination of the variables in PCA.

It is also of note that performing cluster analysis using the “partitioning around mediods” technique (a robust version of k-means clustering) shows us that there are three general clusters in the data for both 1995 and 2010. When observing both the cluster analysis for 1995 and 2010 along side their respective PCA plots, we see that these clusters loosely correspond to Africa, the Americas, and Europe, while Oceania and Asia are not forming distinct clusters. This tells us that economic and health features may define regions more distinctly for Africa, the Americas, and Europe, but Oceania and Asia contain contries that do not share similar health and economic features.

Overall, from the unsupervised clustering technique of Principal Component Analysis (PCA), we see that after decomposing the data, the continents are clustering together, but not into fully separable clusters, based on health and economic features of the regions.

Sources

[1] James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R. Springer, 2017.

[2] “Plotting PCA (Principal Component Analysis).” Plotting PCA (Principal Component Analysis), cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html.

[3] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[4] Starmer, Josh, director. StatQuest: PCA Main Ideas in Only 5 Minutes. Youtube, 4 Dec. 2017, www.youtube.com/watch?v=HMOI_lkzW08.

[5] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2018). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.7-1