# Machine Learning 2: Final Assignment

728842

September 29, 2025

## 1. Introduction

With the Trump administration announcing large cuts to the world's largest foreign aid donors' [6] contributions to development assistance (DA), it is clear that DA funding will decrease [12]. Given these cuts, efficiently allocating DA resources is more critical than ever to maximize impact. To increase efficiency, targeting groups of countries with similar problems beyond geographical factors can help. This report explores a key question: How can countries in need of DA be grouped based on socio-economic, demographic, health, and infrastructure indicators? My analysis of this question leverages principal component analysis (PCA) and clustering to use high dimensional data to group countries. Since development needs are characterized by many different factors, I aim to include as many variables as possible. A large problem in this kind of analysis on all countries is missing data. A single missing observation makes methods as PCA infeasible. To alleviate this I will also employ an imputation strategy, using PCA. I proceed by describing the data used in Section 2 and the method in Section 3. In Section 4, I present results and close with a discussion in Section 5.

## 2. Data

The data for this analysis is from the World Bank Database on World Development Indicators [1]. It contains information on 1,500 country-level development indicators. The data is extracted on all indicators for the year 2018, since it has the least missing observations. Still, around 40% observations are missing. First, I retain only one measure (e.g. when observing both % and absolute values) per indicator. Since some parts of the data have many more missing values, I drop instances where more than 10% is missing. This limits the data to around 127 variables with 192 countries observed. There are 32 demographic indicators, 51 economic, 18 geographic, 10 on health, 9 on infrastructure and 7 on regulations. The unobserved values are spread throughout the data, with mean missing per country around 1.4 (median 0), and around 1.1 (median 0) per variable. The distributions are plotted in Figure 1. Most variables have few or no observations missing, with a few outliers missing more. The distribution for missing values of countries is similar. If only a few columns or rows had many missing values, this would make imputation less suitable. A statistical summary of the variables is in Appendix Table 2. This data still has around 0.7% missing observations, spread across variables. Furthermore, for a more structured analysis of the results, I obtain data that divides countries into regions, as well as into developing and developed from the UN [11].

## 3. Methods

To group countries, I apply the K-Means (KM) algorithm. KM is an unsupervised machine-learning method that groups observations into clusters based on the distance between them. In high-dimensional spaces, KM's performance decreases significantly [2], making dimensionality reduction techniques such as Principal Component Analysis (PCA) useful in the case of datasets with many variables. PCA is effective in reducing the number of variables, especially when they are highly correlated. Furthermore, I leverage PCA not only for dimensionality reduction but also to impute missing values. This follows the approach by [7], who demonstrate advantages of using PCA imputed and decomposed data in KM in case of missing data. I proceed by explaining KM, followed by PCA and lastly PCA imputation.

The data is the $n \times p$ column centered matrix $\mathbf{X}$, with its $n$ rows $\mathbf{x}_i$ as row vectors of $p$ variables observed for each individual. The KM algorithm begins by randomly assigning each observation to a cluster. The number of clusters, $K$, is a fixed input parameter. The centroid $\mathbf{m}_k$ of each cluster $k$ is then calculated as the mean of all the observations assigned to that cluster

$$\mathbf{m}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_i,$$

where $\mathbf{m}_k$ is a vector of each of the $K$ clusters' centroid locations. It contains the mean across the $p$ characteristics for all assigned observations in the cluster. Using the centroids, the closest cluster for each observation is determined by

$$C_j(i) = \arg\min_k ||\mathbf{x}_i - \mathbf{m}_k||^2$$

where $C_j$ is observation $i$'s closest centroid and $||\mathbf{x}_i - \mathbf{m}_k||^2$ the euclidean distance between observation $i$ and a centroid $k$. Each observation is then assigned to the cluster whose centroid $\mathbf{m}_k$ minimizes that distance. With this new assignment, the centroids are recalculated. This process of assignment and centroid recalculation is repeated iteratively until the cluster assignments no longer change, or until a maximum iteration criterion is reached [5]. Although euclidean distance is by far not the only distance measure, it is conventionally used when applying KM to PCA data. The results I use later also rely on this distance measure [3]. The choice of the number of clusters $K$ can be done with data-driven methods, or with considerations from the context of the classification problem. For the data-driven methods, it is common to examine compactness of clusters in relation to the number of clusters as a diagnostic. Compactness measures how much variance is within the clusters. Since high internal similarity is desirable, this criterion serves as a diagnostic for assessing cluster quality. This is commonly evaluated using the sum of squared within-cluster distances (also known as Trace W), which should be minimized [10]. A scree plot effectively visualizes this relationship. A common approach is to identify a 'elbow' point in the graph, which serves as a good candidate for $K$ [5].

PCA uses the singular value decomposition (SVD) to decompose the data matrix $\mathbf{X}$ into the matrix product $\mathbf{U\Sigma V}'$. $\mathbf{U}$ is a $n \times n$ matrix of left singular vectors as columns, $\mathbf{\Sigma}$ is a $n \times p$ diagonal matrix with singular values on the diagonal, and $\mathbf{V}$ is $p \times p$ with right singular vectors as columns. The right singular vectors are the pairwise orthogonal vectors $\mathbf{v}_p$ that are a linear combination of the $p$ features of the $\mathbf{X}$ matrix. These are also called principal components (PC). The weights for this linear combination are the $p$ elements of the vector, called loadings. The loading expresses how much of the $p$th feature of $\mathbf{X}$ is used in the principal component. $\mathbf{\Sigma}$'s diagonal elements $\sigma_p$ are called singular values. Their square $\sigma_p^2$ expresses the variance explained by component $p$. The decomposed matrices are usually order by descending variance explained. A valuable diagnostic is looking at the variance explained per component to asses how good PCA approximates $\mathbf{X}$. Especially looking at the cumulative sum of variance explained by the components used is informative in analyzing how well a set of dimensions does this. The left singular vectors $\mathbf{u}_n$ are not further relevant in PCA.

Findings from [4] demonstrate that the first $r$ PCs of the PCA decomposition are also the closest r-dimensional representation of the entire data, using a least squares error criterion. This means that the SVD restricted to the first $r$ features is the best approximation of the $p$-feature original data, in the reduced feature space. Literature [3] links the choice of the number of dimensions $r$ to retain as input data of KM to the number of clusters $K$ picked in KM. Specifically, they derive that $r = K - 1$ PCs should be used. Given we can use PCA to get matrices that closely approximate $\mathbf{X}$ in a lower dimensional space, it is a natural choice for imputing missing data. This is done using an iterative algorithm [14]. The method consists of three steps: First, the missing values from $\mathbf{X}$ are replaced with column means to make a first guess, producing $\hat{\mathbf{M}}_0$, a $n \times p$ matrix for imputation step 0. In the second step, PCA is applied to the data using the SVD, obtaining $\hat{\mathbf{M}}_0 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$. Following this, we truncate the SVD to the first $r < p$ dimensions, resulting in an close approximation of $\hat{\mathbf{M}}_0$.[1] In the third step, the initially missing values are replaced with corresponding values from $\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}'_r$, obtaining a new approximation matrix $\hat{\mathbf{M}}_1$. Steps 2 and 3 are then repeated until convergence, yielding the final matrix $\hat{\mathbf{M}}_c$. The initially missing values are then taken from $\hat{\mathbf{M}}_c$ as imputations in $\mathbf{X}$. Choosing $r$ is crucial and can be estimated using leave one out cross-validation. This is a process searching for an optimal value of a parameter. For each candidate value of $r$, the process involves iteratively removing each non-missing cell, performing PCA imputation, and recording the mean squared approximation error. The total error is then summed across all simulated missing values, and the $r$ that minimizes this error is selected as the optimal $r$.[8].

## 4. Results

The most important parameter to choose in this analysis is the number of clusters $K$. The goal is to create clusters that effectively distinguish multiple groups of developing countries while maintaining interpretability by limiting the number of clusters. To pick the specific value, I utilize the Scree Plot displayed in Figure 2a. It shows an elbow at 7, indicating this is a good choice for $K$.

I start by standardizing the data to z-scores. Following this, I impute the missing observations using PCA imputation, with an optimal $r$ of 5 obtained via cross-validation. Then I reduce dimensionality to 6 PCs. This value is motivated by the $K-1$ rule from [3]. Plot Figure 2b shows that this is also a reasonable choice, as there are no strong indications to go for more or less dimensions. Using these 6 dimensions, we explain about 54% of the variance in the dataset. This is not very large, which may be a flaw of this analysis.

Using these parameters, I run the analysis. Due to the high number of variables across the 6 PCs, interpreting the 762 loadings is not feasible. However, this is also not the primary concern, as PCA is just my means for reducing data dimensionality. Nevertheless, for completeness, the loadings are included in Appendix Table 3.

As a further diagnostic, I analyze how pure the clusters are in terms of developed and developing countries. This is shown in Table 1. The results here are mixed: Three of the clusters (1, 3, 5) perfectly split off developing countries. Cluster 2 is highly mixed, while Cluster 6 consists mainly of developed countries, with three exceptions: Cuba, Armenia, and Georgia. The developed countries in this cluster are primarily from continental Europe, along with Russia and Japan. Two clusters are strongly driven by outliers: Cluster 4 isolates the United States and China, likely due to their large economies, while Cluster 7 consists solely of Djibouti. This diagnostic shows the model performance is not optimal, but I see this as acceptable

[1] Note: As introduced before I am assuming a column centered $\mathbf{X}$-matrix here. Without column centering, column means would have to be added back in step 2.

Figure 1: Histogram of Number of Missing Values

(a) by Variable

(b) by Country

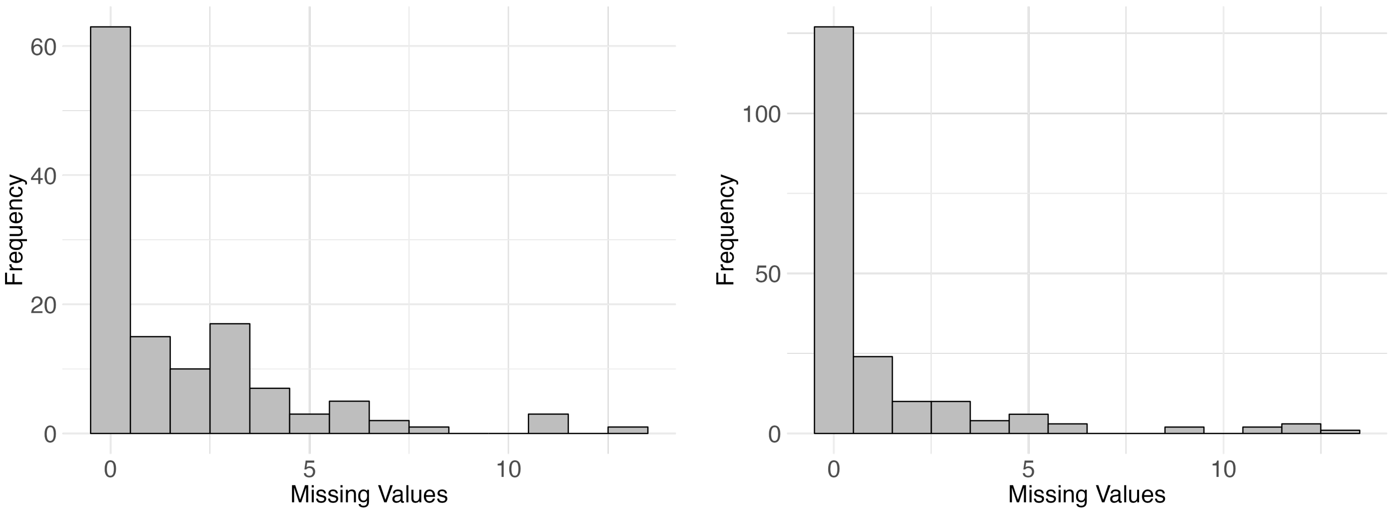



Figure 2: Scree Plots

(a) KM: Compactness of Clusters (Trace W)

(b) PCA: Variance Explained (blue) and Cumulative Variance Explained (red)

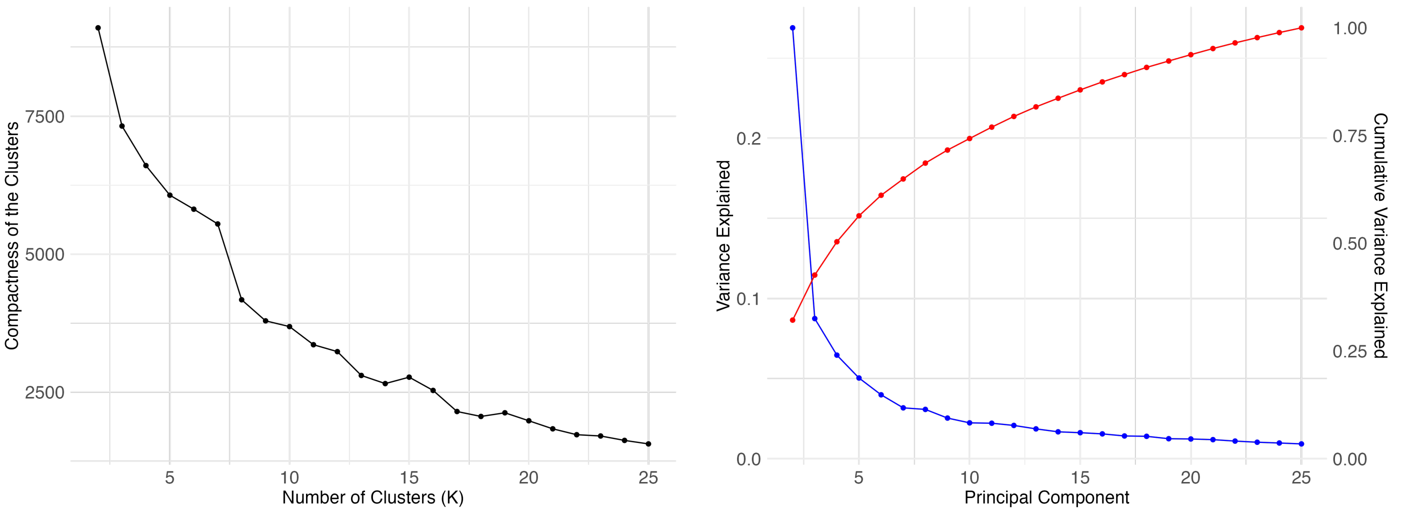
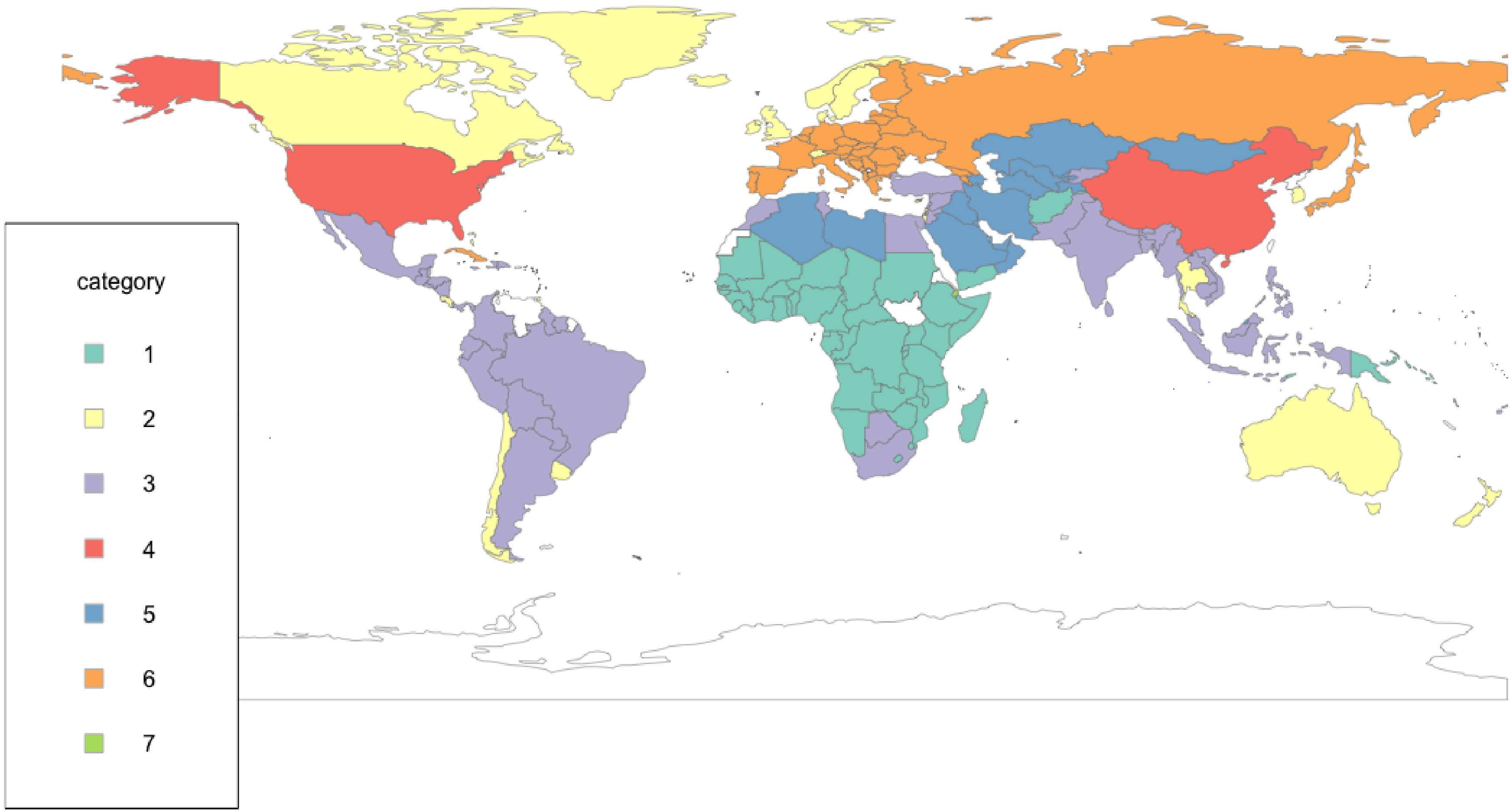



Table 1: Cluster Composition and Minority Groups

| Cluster | Developed | Developing | % Developing |
|---|---|---|---|
| 1 | 0 | 46 | 100 |
| 2 | 17 | 22 | 57.4 |
| 3 | 0 | 52 | 100 |
| 4 | 1 | 1 | 50 |
| 5 | 0 | 18 | 100 |
| 6 | 31 | 3 | 8.8 |
| 7 | 0 | 1 | 100 |

Note: Developed and Developing represent the number of countries with each status in each cluster.

Figure 3: Countries colored by cluster assignment



for an unsupervised classification.

My main result is Figure 3, which is also represented as a table in Appendix Table 4. The map shows the assigned cluster for each country. Cluster 1 is comprised of mostly western, central and eastern African [11] countries, and additionally Yemen, Afghanistan and Papua New Guinea. The inclusion of Yemen, Afghanistan, and Papua New Guinea suggests they might benefit from similar DA strategies as these African regions. Cluster 2 is a mixed group, containing both advanced economies and smaller nations, particularly in the Caribbean. More interesting for my research question, Cluster 3 groups countries from Latin America and the Caribbean, along with regions from Southern Asia, such as India and Pakistan, Sub-Saharan Africa, Oceania, and Western Asia, including Turkey. This is the largest cluster. This cluster consists predominantly of emerging economies and aligns loosely with BRICS+ countries [13], though it notably excludes Russia, China, and some Arab states while incorporating several smaller nations. Targeting DA toward this group appears logical, as these countries share trends of rapid economic growth and more developed infrastructure. The last four clusters have less indications for DA. Cluster 4 is only comprised of the US and China. Cluster 5 groups some central Asian countries, along with Algeria and Libya and many of the large oil exporting states. Cluster 6 consists predominantly of developed economies, while Cluster 7 contains a single outlier.

## 5. Conclusion and Discussion

This report was able to provide interesting insights on grouping of countries based on many development indicators. The results return some groupings not conventional to DA. In summary, the analysis highlights notable groupings, such as the clustering of emerging economies, and the grouping of Western, Central, and Eastern Africa alongside Yemen, Afghanistan, and Papua New Guinea. This suggests shared development challenges and potential DA strategies. Exploring these further and analyzing synergies in providing DA based on this groups seems promising. However, the study also has some limitations:

The analysis has a very data-driven approach to selecting variables and observations. While this is sensible to reduce necessary imputations, a more theory motivated selection of variables might be better. This could significantly alter results, as it might be conceivable that variables that have more missing values across countries measure different concepts compared to more fully observed variables.

Furthermore, a diagnostic analysis on the uncertainty of the imputed data would be sensible. Using bootstrap or Bayesian methods, multiple datasets can be simulated and then analyzed. This is called multiple imputation [9]. Looking at the results from running the PCA/KM analysis could give a feeling for the uncertainty in the data due to imputation. This approach is outlined in [14]. An interesting expansion would be the use of the data in multi-dimensional scaling. The first few PCs could be used to construct a dissimilarity matrix, possibly further including geographic distance. Using this, the same research question could be examined from a different angle. Exploring these avenues could lead to further new groupings of countries based on development indicators. Testing the implementation of these groups would provide interesting insights on increasing efficiency in DA to aid as many people as possible.

## References


[1] World Bank. World Development Indicators, 2024. https://databank.worldbank.org/source/world-development-indicators.

[2] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When Is "Nearest Neighbor" Meaningful? In Catriel Beeri and Peter Buneman, editors, *Database Theory — ICDT'99*, pages 217–235, Berlin, Heidelberg, 1999. Springer.

[3] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, page 29, New York, NY, USA, July 2004. Association for Computing Machinery.

[4] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936.

[5] Trevor Hastie, Robert Tibshirani, and Friedman Jerome H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, 2nd edition, 2009.

[6] Bastian Herre and Pablo Arriagada. Foreign Aid. *Our World in Data*, October 2024.

[7] Katsuhiro Honda, Ryoichi Nonoguchi, Akira Notsu, and Hidetomo Ichihashi. PCA-guided k-Means clustering with incomplete data. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, pages 1710–1714, Taipei, Taiwan, June 2011. IEEE.

[8] Julie Josse and François Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6):1869–1879, June 2012.

[9] Julie Josse, Jérôme Pagès, and François Husson. Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*, 5(3):231–246, October 2011.

[10] Glenn W. Milligan and Martha C. Cooper. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50(2):159–179, June 1985.

[11] United Nations. M49 Standard, 2021. https://unstats.un.org/unsd/methodology/m49/.

[12] Michael Peel, Amy Borrett, and Sarah Neville. The Trump aid gap. *Financial Times*, February 2025.

[13] Konrad Adenauer Stiftung. BRICS expansion, November 2023. https://www.kas.de/en/brics-plus.

[14] Joost R. van Ginkel. Handling Missing Data in Principal Component Analysis Using Multiple Imputation. In L. Andries van der Ark, Wilco H. M. Emons, and Rob R. Meijer, editors, *Essays on Contemporary Psychometrics*, pages 141–161. Springer International Publishing, Cham, 2023.

## Appendix

Table 2: Descriptive Statistics Summary

| Variable | Mean | SD | Min | Max |
|---|---|---|---|---|
| Access to electricity (% of population) | 84.9193 | 24.8232 | 9.5000 | 100.0000 |
| Adolescent fertility rate (births per 1,000 women ages 15-19) | 46.7664 | 40.1103 | 1.1790 | 165.5440 |
| Age dependency ratio (% of working-age population) | 58.9080 | 17.0814 | 17.7845 | 104.9582 |
| Agricultural land (% of land area) | 37.7554 | 22.2938 | 0.5233 | 81.3421 |
| Agriculture, forestry, and fishing, value added (% of GDP) | 10.2754 | 9.6842 | 0.0223 | 39.0099 |
| Aquaculture production (metric tons) | 5.4881e+05 | 4.9507e+06 | -1.9930e+06 | 6.6135e+07 |
| Arable land (% of land area) | 14.5282 | 13.7238 | 0.0863 | 59.7100 |
| Bird species, threatened | 23.1146 | 25.7732 | 0 | 175.0000 |
| Birth rate, crude (per 1,000 people) | 20.0509 | 9.9958 | 6.4000 | 46.1270 |
| Capture fisheries production (metric tons) | 4.9436e+05 | 1.4965e+06 | 0 | 1.4831e+07 |
| Carbon dioxide (CO2) emissions (total) excluding LULUCF (% change from 1990) | 760.2778 | 4.4853e+03 | -72.2038 | 4.7109e+04 |
| Compulsory education, duration (years) | 9.8472 | 2.3988 | 0 | 16.0000 |
| Control of Corruption: Estimate | -0.0400 | 0.9721 | -1.7858 | 2.1718 |
| Death rate, crude (per 1,000 people) | 7.6314 | 2.7000 | 0.9820 | 15.4000 |
| DEC alternative conversion factor (LCU per US$) | 994.5485 | 4.7830e+03 | 0.3020 | 5.4219e+04 |
| Energy intensity level of primary energy (MJ/$2017 PPP GDP) | 4.6702 | 2.6157 | 0.4800 | 18.5800 |
| Export unit value index (2015 = 100) | 109.5964 | 10.0029 | 71.6000 | 143.3000 |
| Export value index (2015 = 100) | 135.4763 | 183.9540 | 33.6169 | 2.6366e+03 |
| Export volume index (2015 = 100) | 124.1947 | 176.1307 | 27.5866 | 2.5035e+03 |
| Fertility rate, total (births per woman) | 2.6582 | 1.3264 | 0.9770 | 7.0230 |
| Fish species, threatened | 39.0938 | 38.8610 | 0 | 251.0000 |
| Forest area (% of land area) | 32.3271 | 24.2310 | 0 | 94.8290 |
| Forest rents (% of GDP) | 1.3615 | 2.8288 | 0 | 20.7466 |
| GDP (constant 2015 US$) | 4.2627e+11 | 1.7996e+12 | 8.1762e+07 | 1.9652e+13 |
| GDP deflator (base year varies by country) | 217.1417 | 602.2174 | 30.7319 | 6.1818e+03 |
| GDP growth (annual %) | 3.2491 | 2.5399 | -6.2365 | 8.7762 |
| GDP per capita (constant 2015 US$) | 1.4719e+04 | 2.0187e+04 | 265.6724 | 1.0684e+05 |
| GDP per capita growth (annual %) | 2.0207 | 2.7027 | -9.3704 | 8.3850 |
| Government Effectiveness: Estimate | -0.0374 | 0.9518 | -2.2737 | 2.2318 |
| Import unit value index (2015 = 100) | 107.8656 | 4.8716 | 84.4000 | 123.4000 |
| Import value index (2015 = 100) | 118.4693 | 31.7053 | 33.0400 | 413.8667 |
| Import volume index (2015 = 100) | 110.0140 | 29.3419 | 33.1143 | 389.6674 |
| Incidence of tuberculosis (per 100,000 people) | 109.7145 | 158.5430 | 0 | 1.1800e+03 |
| Industry (including construction), value added (% of GDP) | 25.5988 | 11.5246 | 4.0907 | 65.8764 |
| Land area (sq. km) | 6.6732e+05 | 1.8456e+06 | 20.0000 | 1.6377e+07 |
| Life expectancy at birth, female (years) | 75.1126 | 7.7976 | 52.7700 | 87.6100 |
| Lower secondary school starting age (years) | 11.8194 | 0.8621 | 10.0000 | 14.0000 |
| Mammal species, threatened | 17.1615 | 22.4977 | 0 | 191.0000 |
| Merchandise exports (current US$) | 9.9635e+10 | 2.7465e+11 | -3.9097e+10 | 2.4867e+12 |
| Merchandise exports by the reporting economy (current US$) | 9.8471e+10 | 2.7486e+11 | 1.3153e+07 | 2.5013e+12 |
| Merchandise exports to economies in the Arab World (% of total merchandise exports) | 7.5047 | 13.4004 | 0.0020 | 75.3253 |
| Merchandise exports to high-income economies (% of total merchandise exports) | 57.5884 | 24.3564 | 2.9664 | 95.7271 |
| Merchandise exports to low- and middle-income economies in East Asia & Pacific (% of total merchandise exports) | 13.6993 | 18.9415 | 0.0370 | 93.0144 |

Table 2: Descriptive Statistics Summary *(continued)*

| Variable | Mean | SD | Min | Max |
|---|---|---|---|---|
| Merchandise exports to low- and middle-income economies in Europe & Central Asia (% of total merchandise exports) | 5.8258 | 11.0948 | 0 | 64.8180 |
| Merchandise exports to low- and middle-income economies in Latin America & the Caribbean (% of total merchandise exports) | 4.5164 | 9.3840 | 1.0000e-04 | 64.3281 |
| Merchandise exports to low- and middle-income economies in Middle East & North Africa (% of total merchandise exports) | 2.8115 | 6.8815 | 0 | 64.4347 |
| Merchandise exports to low- and middle-income economies in South Asia (% of total merchandise exports) | 5.9395 | 13.1499 | 0 | 94.5910 |
| Merchandise exports to low- and middle-income economies in Sub-Saharan Africa (% of total merchandise exports) | 7.4657 | 15.0695 | 0.0010 | 92.0587 |
| Merchandise exports to low- and middle-income economies outside region (% of total merchandise exports) | 22.0377 | 19.2953 | 0.1403 | 85.3454 |
| Merchandise imports (current US$) | 1.0126e+11 | 2.8964e+11 | -4.3613e+10 | 2.6142e+12 |
| Merchandise imports by the reporting economy (current US$) | 1.0081e+11 | 2.8646e+11 | 9.9027e+07 | 2.5427e+12 |
| Merchandise imports from economies in the Arab World (% of total merchandise imports) | 5.7792 | 8.7617 | 4.0000e-04 | 58.0877 |
| Merchandise imports from high-income economies (% of total merchandise imports) | 55.9546 | 20.7732 | 6.8969 | 95.8614 |
| Merchandise imports from low- and middle-income economies in East Asia & Pacific (% of total merchandise imports) | 17.9500 | 12.9462 | 1.4195 | 89.3399 |
| Merchandise imports from low- and middle-income economies in Europe & Central Asia (% of total merchandise imports) | 7.2663 | 11.6880 | 0.0011 | 66.6218 |
| Merchandise imports from low- and middle-income economies in Latin America & the Caribbean (% of total merchandise imports) | 4.7202 | 8.0762 | 0.0018 | 41.6579 |
| Merchandise imports from low- and middle-income economies in Middle East & North Africa (% of total merchandise imports) | 1.6315 | 2.4062 | 0 | 17.1454 |
| Merchandise imports from low- and middle-income economies in South Asia (% of total merchandise imports) | 4.2144 | 8.9654 | 0.0307 | 90.9473 |
| Merchandise imports from low- and middle-income economies in Sub-Saharan Africa (% of total merchandise imports) | 6.3135 | 13.9392 | 0.0040 | 84.4018 |
| Merchandise imports from low- and middle-income economies outside region (% of total merchandise imports) | 25.1477 | 15.1640 | 0.6812 | 70.0032 |
| Merchandise trade (% of GDP) | 63.6231 | 40.4956 | 13.8501 | 330.3758 |
| Methane (CH4) emissions (total) excluding LULUCF (% change from 1990) | 74.7990 | 383.7868 | -78.2349 | 5.2244e+03 |
| Mineral rents (% of GDP) | 0.6221 | 1.6224 | 0 | 9.6111 |
| Mortality rate, adult, female (per 1,000 female adults) | 122.0313 | 81.8364 | 23.0550 | 419.4490 |
| Natural gas rents (% of GDP) | 0.8883 | 3.8607 | 0 | 48.1395 |

Table 2: Descriptive Statistics Summary *(continued)*

| Variable | Mean | SD | Min | Max |
|---|---|---|---|---|
| Net barter terms of trade index (2015 = 100) | 103.0247 | 10.7003 | 53.5375 | 136.9241 |
| Net migration | 6.3676e+03 | 2.2226e+05 | -1.3096e+06 | 1.7650e+06 |
| Nitrous oxide (N2O) emissions (total) excluding LULUCF (% change from 1990) | 51.7933 | 79.1771 | -47.8954 | 554.5455 |
| Official exchange rate (LCU per US$, period average) | 824.5127 | 3.6830e+03 | 0.3020 | 4.0864e+04 |
| Oil rents (% of GDP) | 2.9529 | 7.7308 | 0 | 45.7659 |
| People practicing open defecation (% of population) | 6.7947 | 12.2354 | 0 | 68.4855 |
| People using at least basic drinking water services (% of population) | 88.3125 | 15.5556 | 35.8471 | 104.8187 |
| People using at least basic sanitation services (% of population) | 76.8181 | 27.2296 | 8.3499 | 100.0000 |
| Permanent cropland (% of land area) | 4.4590 | 6.9561 | 0.0035 | 39.5062 |
| Plant species (higher), threatened | 85.4167 | 194.6361 | 0 | 1.8590e+03 |
| PM2.5 air pollution, mean annual exposure (micrograms per cubic meter) | 25.8375 | 17.2759 | 5.4970 | 93.3070 |
| Political Stability and Absence of Violence/Terrorism: Estimate | -0.0616 | 0.9756 | -2.9962 | 1.9368 |
| Population ages 0-14 (% of total population) | 27.6346 | 10.5863 | 11.5797 | 49.0475 |
| Population ages 00-04, female (% of female population) | 9.5581 | 4.2531 | 3.4664 | 19.0975 |
| Population ages 05-09, female (% of female population) | 9.2065 | 3.5135 | 3.6236 | 16.4440 |
| Population ages 10-14, female (% of female population) | 8.4804 | 2.8819 | 3.3214 | 13.6503 |
| Population ages 15-19, female (% of female population) | 7.9530 | 2.2871 | 3.6032 | 12.0170 |
| Population ages 15-64 (% of total population) | 63.6225 | 6.5435 | 48.7904 | 84.9008 |
| Population ages 20-24, female (% of female population) | 7.7903 | 1.6501 | 4.1731 | 10.5950 |
| Population ages 25-29, female (% of female population) | 7.7716 | 1.3137 | 4.5792 | 12.8862 |
| Population ages 30-34, female (% of female population) | 7.3990 | 1.2918 | 5.0520 | 14.5421 |
| Population ages 35-39, female (% of female population) | 6.7264 | 1.1492 | 3.7818 | 11.8534 |
| Population ages 40-44, female (% of female population) | 6.0169 | 1.2128 | 3.0049 | 10.7131 |
| Population ages 45-49, female (% of female population) | 5.5457 | 1.4612 | 2.8685 | 8.9058 |
| Population ages 50-54, female (% of female population) | 5.1408 | 1.7368 | 2.1391 | 9.1678 |
| Population ages 55-59, female (% of female population) | 4.6635 | 1.9214 | 1.7520 | 8.4412 |
| Population ages 60-64, female (% of female population) | 3.9525 | 1.9295 | 1.3073 | 7.8092 |
| Population ages 65 and above (% of total population) | 8.7429 | 6.2440 | 1.1443 | 28.3452 |
| Population ages 65-69, female (% of female population) | 3.2023 | 1.8705 | 0.7104 | 7.6865 |
| Population ages 70-74, female (% of female population) | 2.3764 | 1.5919 | 0.4298 | 6.9113 |
| Population ages 75-79, female (% of female population) | 1.8063 | 1.4000 | 0.0803 | 6.0596 |
| Population ages 80 and above, female (% of female population) | 2.4104 | 2.3489 | 0.1861 | 10.8385 |
| Population density (people per sq. km of land area) | 347.8778 | 1.6329e+03 | 0.1365 | 2.0027e+04 |
| Population growth (annual %) | 1.2026 | 1.2786 | -2.8988 | 3.9802 |
| Population, female | 1.9645e+07 | 7.1436e+07 | 5.6030e+03 | 6.8566e+08 |
| PPP conversion factor, GDP (LCU per international $) | 343.1077 | 1.5197e+03 | 0.1885 | 1.6923e+04 |
| Preprimary education, duration (years) | 4.0395 | 1.6418 | 1.0000 | 7.0000 |
| Price level ratio of PPP conversion factor (GDP) to market exchange rate | 0.5538 | 0.2423 | 0.1741 | 1.3167 |
| Primary education, duration (years) | 5.7745 | 0.8417 | 4.0000 | 8.0000 |
| Primary school starting age (years) | 6.0571 | 0.5336 | 5.0000 | 7.0000 |
| Regulatory Quality: Estimate | -0.0154 | 0.9548 | -2.2152 | 2.2213 |
| Renewable energy consumption (% of total final energy consumption) | 30.4714 | 27.9458 | 0 | 96.4000 |
| Rule of Law: Estimate | -0.0405 | 0.9555 | -2.2987 | 2.0344 |
| Rural population | 1.7847e+07 | 7.8948e+07 | 0 | 9.0686e+08 |
| Rural population growth (annual %) | 0.1247 | 1.5996 | -5.0191 | 3.3612 |
| Scientific and technical journal articles | 1.3603e+04 | 5.3523e+04 | -6.5627e+03 | 5.3230e+05 |
| Secondary education, duration (years) | 6.4168 | 0.8926 | 4.0000 | 9.0000 |
| Secure Internet servers | 2.4108e+05 | 1.6037e+06 | -2.3493e+05 | 2.1517e+07 |
| Services, value added (% of GDP) | 56.3585 | 11.6358 | 33.1147 | 94.2951 |
| Sex ratio at birth (male births per female births) | 1.0516 | 0.0200 | 1.0100 | 1.1480 |

Table 2: Descriptive Statistics Summary *(continued)*

| Variable | Mean | SD | Min | Max |
|---|---|---|---|---|
| Surface area (sq. km) | 7.2145e+05 | 2.1082e+06 | 20.0000 | 1.7098e+07 |
| Survival to age 65, female (% of cohort) | 80.1483 | 11.9293 | 44.2315 | 95.9002 |
| Terrestrial and marine protected areas (% of total territorial area) | 11.9930 | 12.3285 | 4.0000e-04 | 82.9424 |
| Terrestrial protected areas (% of total land area) | 16.8313 | 11.7510 | 0.0492 | 53.6230 |
| Total fisheries production (metric tons) | 1.0353e+06 | 6.2171e+06 | -2.3287e+06 | 8.0966e+07 |
| Total greenhouse gas emissions excluding LULUCF (% change from 1990) | 125.5981 | 412.9404 | -68.4883 | 5.3127e+03 |
| Total greenhouse gas emissions excluding LULUCF per capita (t CO2e/capita) | 6.8950 | 8.7096 | 0.0574 | 72.2634 |
| Total natural resources rents (% of GDP) | 5.9843 | 9.8486 | 0 | 62.7690 |
| Tuberculosis case detection rate (%, all forms) | 76.1777 | 15.3305 | 23.0000 | 110.0000 |
| Urban population | 2.1723e+07 | 7.4136e+07 | 1.1477e+04 | 8.2976e+08 |
| Urban population growth (annual %) | 1.8498 | 1.6356 | -2.7136 | 5.8587 |
| Voice and Accountability: Estimate | -0.0155 | 0.9638 | -2.1529 | 1.7093 |

Table 3: PCA Loadings

| Variable<br>Cumulative Variance Explained | PC1<br>(26.89%) | PC2<br>(35.64%) | PC3<br>(42.1%) | PC4<br>(47.13%) | PC5<br>(51.12%) | PC6<br>(54.29%) | PC7<br>(57.36%) |
|---|---|---|---|---|---|---|---|
| Access to electricity (% of population) | -0.137 | 0.009 | -0.122 | 0.015 | -0.067 | -0.016 | 0.082 |
| Adolescent fertility rate (births per 1,000 women ages 15-19) | 0.142 | -0.002 | 0.076 | -0.043 | 0.05 | -0.061 | 0.058 |
| Age dependency ratio (% of working-age population) | 0.133 | -0.005 | 0.18 | 0.024 | 0.086 | 0.041 | 0.029 |
| Agricultural land (% of land area) | 0.034 | -0.036 | 0.101 | 0.152 | -0.099 | 0.079 | 0.046 |
| Agriculture, forestry, and fishing, value added (% of GDP) | 0.131 | -0.017 | 0.076 | 0.065 | -0.041 | -0.039 | 0.045 |
| Aquaculture production (metric tons) | -0.016 | -0.226 | -0.002 | 0.017 | -0.069 | -0.047 | -0.006 |
| Arable land (% of land area) | -0.015 | -0.015 | 0.129 | 0.151 | -0.087 | 0.011 | 0.034 |
| Bird species, threatened | 0.004 | -0.187 | -0.033 | -0.017 | -0.047 | -0.158 | 0.049 |
| Birth rate, crude (per 1,000 people) | 0.164 | -0.008 | 0.053 | -0.01 | 0.038 | 0.022 | -0.004 |
| Capture fisheries production (metric tons) | -0.026 | -0.253 | -0.004 | -0.005 | -0.049 | -0.081 | -0.004 |
| Carbon dioxide (CO2) emissions (total) excluding LULUCF (% change from 1990) | 0.009 | 0.017 | -0.022 | -0.051 | 0.03 | -0.069 | -0.026 |
| Compulsory education, duration (years) | -0.063 | 0.008 | -0.01 | -0.016 | -0.049 | 0.017 | 0.14 |
| Control of Corruption: Estimate | -0.122 | 0.033 | 0.063 | -0.174 | 0.027 | 0.041 | -0.105 |
| Death rate, crude (per 1,000 people) | -0.035 | 0.018 | 0.233 | 0.159 | 0.104 | -0.067 | -0.003 |
| DEC alternative conversion factor (LCU per US$) | 0.022 | -0.041 | -0.084 | 0.066 | 0.037 | -0.129 | -0.053 |
| Energy intensity level of primary energy (MJ/$2017 PPP GDP) | 0.059 | -0.025 | -0.015 | 0.012 | 0.123 | -0.06 | -0.064 |
| Export unit value index (2015 = 100) | -0.003 | 0.021 | -0.071 | 0.09 | 0.257 | -0.011 | -0.076 |
| Export value index (2015 = 100) | 0.013 | 0.006 | 0.009 | 0.079 | -0.129 | 0.133 | -0.3 |
| Export volume index (2015 = 100) | 0.014 | 0.006 | 0.017 | 0.072 | -0.149 | 0.134 | -0.293 |
| Fertility rate, total (births per woman) | 0.156 | -0.009 | 0.075 | 0.002 | 0.073 | 0.048 | -0.01 |
| Fish species, threatened | 0.006 | -0.199 | 0.009 | -0.046 | -0.022 | -0.044 | 0.055 |
| Forest area (% of land area) | -0.014 | 0.009 | 0.051 | -0.101 | 0.075 | -0.25 | 0.006 |
| Forest rents (% of GDP) | 0.1 | -0.006 | 0.104 | -0.008 | 0.069 | -0.047 | -0.058 |
| GDP (constant 2015 US$) | -0.045 | -0.246 | 0.053 | -0.051 | 0.042 | 0.091 | 0.003 |
| GDP deflator (base year varies by country) | 0.034 | -0.008 | -0.023 | 0.047 | 0.038 | 0.165 | 0.168 |
| GDP growth (annual %) | 0.013 | -0.034 | 0.037 | 0.111 | -0.171 | -0.111 | -0.151 |
| GDP per capita (constant 2015 US$) | -0.109 | 0.012 | 0.007 | -0.14 | 0.102 | 0.118 | -0.109 |
| GDP per capita growth (annual %) | -0.046 | -0.018 | 0.041 | 0.135 | -0.199 | -0.13 | -0.108 |
| Government Effectiveness: Estimate | -0.142 | 0.001 | 0.025 | -0.125 | 0.036 | 0.009 | -0.102 |

Table 3: PCA Loadings *(continued)*

| variable | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| Import unit value index (2015 = 100) | -0.06 | 0.057 | 0.105 | 0.076 | 0.083 | -0.011 | -0.074 |
| Import value index (2015 = 100) | -0.028 | 0.01 | 0.06 | 0.11 | -0.198 | 0.006 | -0.298 |
| Import volume index (2015 = 100) | -0.019 | 0.003 | 0.043 | 0.099 | -0.209 | 0.007 | -0.291 |
| Incidence of tuberculosis (per 100,000 people) | 0.088 | -0.021 | 0.032 | -0.036 | 0.002 | -0.152 | -0.094 |
| Industry (including construction), value added (% of GDP) | 0.027 | -0.063 | -0.183 | 0.058 | 0.182 | -0.066 | -0.06 |
| Land area (sq. km) | -0.017 | -0.197 | -0.008 | 0 | 0.076 | 0.014 | 0.036 |
| Life expectancy at birth, female (years) | -0.156 | 0.009 | -0.07 | -0.018 | -0.007 | 0.035 | 0.031 |
| Lower secondary school starting age (years) | 0.052 | -0.001 | 0.081 | -0.192 | 0.034 | -0.006 | -0.033 |
| Mammal species, threatened | 0.03 | -0.178 | -0.029 | 0.002 | -0.058 | -0.18 | 0.005 |
| Merchandise exports (current US$) | -0.067 | -0.237 | 0.046 | -0.033 | 0.053 | 0.075 | -0.048 |
| Merchandise exports by the reporting economy (current US$) | -0.067 | -0.237 | 0.051 | -0.032 | 0.052 | 0.074 | -0.045 |
| Merchandise exports to economies in the Arab World (% of total merchandise exports) | 0.035 | 0.002 | -0.023 | 0.054 | -0.071 | 0.261 | 0.035 |
| Merchandise exports to high-income economies (% of total merchandise exports) | -0.075 | 0.014 | 0.074 | -0.041 | -0.05 | -0.017 | 0.049 |
| Merchandise exports to low- and middle-income economies in East Asia & Pacific (% of total merchandise exports) | 0.029 | -0.036 | -0.104 | -0.051 | 0.126 | -0.157 | -0.103 |
| Merchandise exports to low- and middle-income economies in Europe & Central Asia (% of total merchandise exports) | -0.029 | 0.028 | -0.006 | 0.261 | 0.048 | -0.009 | 0.097 |
| Merchandise exports to low- and middle-income economies in Latin America & the Caribbean (% of total merchandise exports) | -0.016 | -0.008 | -0.018 | -0.106 | -0.115 | -0.024 | 0.217 |
| Merchandise exports to low- and middle-income economies in Middle East & North Africa (% of total merchandise exports) | 0.005 | 0.008 | -0.031 | 0.076 | -0.058 | 0.232 | 0.131 |
| Merchandise exports to low- and middle-income economies in South Asia (% of total merchandise exports) | 0.05 | -0.011 | -0.015 | 0.004 | 0.031 | 0.045 | -0.07 |
| Merchandise exports to low- and middle-income economies in Sub-Saharan Africa (% of total merchandise exports) | 0.079 | 0.007 | 0.101 | -0.043 | -0.057 | 0.069 | -0.139 |
| Merchandise exports to low- and middle-income economies outside region (% of total merchandise exports) | 0.017 | -0.035 | -0.067 | -0.007 | 0.136 | 0.171 | -0.081 |
| Merchandise imports (current US$) | -0.065 | -0.238 | 0.062 | -0.047 | 0.053 | 0.092 | -0.034 |
| Merchandise imports by the reporting economy (current US$) | -0.066 | -0.238 | 0.063 | -0.046 | 0.051 | 0.091 | -0.034 |
| Merchandise imports from economies in the Arab World (% of total merchandise imports) | 0.059 | -0.044 | -0.064 | 0.054 | -0.052 | 0.254 | -0.052 |

Table 3: PCA Loadings *(continued)*

| variable | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| Merchandise imports from high-income economies (% of total merchandise imports) | -0.1 | 0.046 | 0.042 | -0.09 | -0.004 | 0.058 | 0.003 |
| Merchandise imports from low- and middle-income economies in East Asia & Pacific (% of total merchandise imports) | 0.041 | -0.071 | -0.092 | -0.025 | -0.004 | -0.187 | -0.084 |
| Merchandise imports from low- and middle-income economies in Europe & Central Asia (% of total merchandise imports) | -0.014 | 0.026 | -0.042 | 0.291 | 0.089 | 0.005 | 0.062 |
| Merchandise imports from low- and middle-income economies in Latin America & the Caribbean (% of total merchandise imports) | -0.008 | -0.017 | -0.045 | -0.102 | -0.085 | -0.02 | 0.25 |
| Merchandise imports from low- and middle-income economies in Middle East & North Africa (% of total merchandise imports) | 0.035 | -0.041 | -0.024 | 0.141 | -0.007 | 0.265 | 0.034 |
| Merchandise imports from low- and middle-income economies in South Asia (% of total merchandise imports) | 0.048 | -0.003 | -0.007 | 0.01 | -0.051 | 0.019 | -0.111 |
| Merchandise imports from low- and middle-income economies in Sub-Saharan Africa (% of total merchandise imports) | 0.082 | 0.004 | 0.097 | -0.064 | 0.027 | 0.019 | -0.09 |
| Merchandise imports from low- and middle-income economies outside region (% of total merchandise imports) | 0.039 | -0.068 | -0.041 | 0.047 | 0.076 | 0.202 | -0.038 |
| Merchandise trade (% of GDP) | -0.046 | 0.048 | -0.012 | 0.056 | 0.013 | -0.01 | -0.251 |
| Methane (CH4) emissions (total) excluding LULUCF (% change from 1990) | 0.036 | -0.001 | -0.046 | -0.047 | 0.126 | -0.015 | -0.014 |
| Mineral rents (% of GDP) | 0.04 | -0.017 | -0.014 | 0.061 | 0.066 | -0.134 | -0.074 |
| Mortality rate, adult, female (per 1,000 female adults) | 0.145 | -0.005 | 0.1 | -0.03 | 0.016 | -0.032 | -0.047 |
| Natural gas rents (% of GDP) | 0.014 | 0.002 | -0.102 | -0.01 | 0.092 | -0.065 | -0.017 |
| Net barter terms of trade index (2015 = 100) | 0.03 | 0 | -0.145 | 0.053 | 0.247 | -0.006 | -0.015 |
| Net migration | -0.036 | -0.055 | 0.058 | -0.082 | 0.149 | 0.077 | 0.024 |
| Nitrous oxide (N2O) emissions (total) excluding LULUCF (% change from 1990) | 0.079 | -0.016 | -0.147 | -0.056 | -0.02 | 0.051 | -0.076 |
| Official exchange rate (LCU per US$, period average) | 0.018 | -0.045 | -0.094 | 0.062 | 0.023 | -0.166 | -0.064 |
| Oil rents (% of GDP) | 0.028 | -0.013 | -0.21 | 0.026 | 0.215 | 0.067 | -0.026 |
| People practicing open defecation (% of population) | 0.11 | -0.017 | 0.085 | -0.016 | 0.03 | -0.022 | -0.068 |
| People using at least basic drinking water services (% of population) | -0.143 | 0.02 | -0.102 | -0.015 | -0.054 | 0.003 | 0.053 |
| People using at least basic sanitation services (% of population) | -0.146 | 0.019 | -0.11 | 0.023 | -0.031 | 0.035 | 0.06 |
| Permanent cropland (% of land area) | 0.017 | 0.024 | 0.025 | -0.067 | -0.158 | -0.075 | 0.086 |
| Plant species (higher), threatened | 0.013 | -0.132 | -0.006 | -0.042 | -0.047 | -0.106 | 0.077 |
| PM2.5 air pollution, mean annual exposure (micrograms per cubic meter) | 0.09 | -0.044 | -0.087 | 0.101 | -0.024 | 0.133 | -0.07 |

Table 3: PCA Loadings *(continued)*

| variable | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| Political Stability and Absence of Violence/Terrorism: Estimate | -0.107 | 0.068 | 0.025 | -0.173 | 0.009 | -0.097 | -0.116 |
| Population ages 0-14 (%) | 0.166 | -0.009 | 0.043 | -0.019 | 0 | 0.017 | 0.026 |
| Population ages 00-04, female (% f) | 0.165 | -0.009 | 0.012 | -0.018 | 0.036 | 0.042 | -0.003 |
| Population ages 05-09, female (% f) | 0.165 | -0.007 | 0.007 | -0.032 | 0.007 | 0.042 | 0.011 |
| Population ages 10-14, female (% f) | 0.163 | -0.007 | 0.018 | -0.054 | -0.041 | 0.027 | 0.028 |
| Population ages 15-19, female (% f) | 0.155 | -0.007 | 0.007 | -0.077 | -0.098 | 0.001 | 0.045 |
| Population ages 15-64 (%) | -0.127 | 0.004 | -0.193 | -0.031 | -0.078 | -0.026 | -0.046 |
| Population ages 20-24, female (% f) | 0.132 | -0.002 | -0.084 | -0.105 | -0.136 | -0.023 | 0.024 |
| Population ages 25-29, female (% f) | 0.062 | -0.005 | -0.225 | -0.099 | -0.127 | -0.023 | -0.054 |
| Population ages 30-34, female (% f) | -0.008 | -0.005 | -0.279 | -0.063 | -0.079 | -0.007 | -0.089 |
| Population ages 35-39, female (% f) | -0.081 | 0.001 | -0.239 | -0.04 | -0.053 | 0.012 | -0.086 |
| Population ages 40-44, female (% f) | -0.136 | 0.002 | -0.132 | 0.025 | -0.01 | -0.008 | -0.043 |
| Population ages 45-49, female (% f) | -0.155 | -0.016 | -0.043 | 0.02 | -0.001 | -0.032 | -0.012 |
| Population ages 50-54, female (% f) | -0.158 | -0.002 | 0.008 | 0.025 | 0.002 | -0.037 | 0.01 |
| Population ages 55-59, female (% f) | -0.157 | 0.017 | 0.038 | 0.063 | 0.027 | -0.044 | 0.014 |
| Population ages 60-64, female (% f) | -0.155 | 0.008 | 0.08 | 0.085 | 0.042 | -0.045 | 0.012 |
| Population ages 65 and above (%) | -0.148 | 0.011 | 0.129 | 0.064 | 0.081 | -0.001 | 0.004 |
| Population ages 65-69, female (% f) | -0.149 | 0.007 | 0.11 | 0.094 | 0.065 | -0.035 | 0.008 |
| Population ages 70-74, female (% f) | -0.146 | 0.014 | 0.13 | 0.054 | 0.076 | -0.004 | 0.004 |
| Population ages 75-79, female (% f) | -0.142 | 0.016 | 0.13 | 0.098 | 0.087 | -0.012 | 0.012 |
| Population ages 80 and above, female (% f) | -0.14 | 0.013 | 0.133 | 0.072 | 0.099 | 0.016 | 0.001 |
| Population density (people per sq. km of land area) | -0.027 | 0.014 | -0.043 | -0.06 | -0.045 | 0.03 | -0.092 |
| Population growth (annual %) | 0.119 | -0.028 | -0.014 | -0.065 | 0.087 | 0.055 | -0.063 |
| Population, female | -0.01 | -0.251 | 0.005 | 0.034 | -0.086 | -0.011 | 0.008 |
| PPP conversion factor, GDP (LCU per international $) | 0.024 | -0.043 | -0.082 | 0.063 | 0.037 | -0.129 | -0.049 |
| Preprimary education, duration (years) | -0.067 | -0.004 | 0.003 | 0.035 | 0.036 | -0.085 | 0.068 |
| Price level ratio of PPP conversion factor (GDP) to market exchange rate | -0.093 | 0.023 | 0.076 | -0.209 | 0.074 | 0.071 | -0.054 |
| Primary education, duration (years) | 0.032 | 0.013 | 0.052 | -0.277 | -0.024 | 0.032 | -0.018 |
| Primary school starting age (years) | 0.03 | -0.018 | 0.049 | 0.12 | 0.095 | -0.062 | -0.028 |
| Regulatory Quality: Estimate | -0.135 | 0.016 | 0.064 | -0.106 | 0.025 | 0.031 | -0.106 |
| Renewable energy consumption (% of total final energy consumption) | 0.107 | -0.007 | 0.161 | -0.002 | 0.02 | -0.089 | -0.042 |
| Rule of Law: Estimate | -0.131 | 0.025 | 0.058 | -0.152 | 0.028 | 0.033 | -0.117 |
| Rural population | 0.003 | -0.205 | 0.001 | 0.046 | -0.108 | -0.018 | 0.001 |
| Rural population growth (annual %) | 0.103 | -0.001 | 0.06 | -0.005 | 0.048 | 0.001 | -0.03 |
| Scientific and technical journal articles | -0.049 | -0.265 | 0.047 | -0.03 | 0.023 | 0.071 | -0.008 |
| Secondary education, duration (years) | -0.026 | 0.007 | 0.024 | 0.225 | 0.052 | -0.02 | -0.055 |
| Secure Internet servers | -0.036 | -0.145 | 0.063 | -0.067 | 0.083 | 0.113 | 0.003 |
| Services, value added (% of GDP) | -0.108 | 0.044 | 0.057 | -0.14 | -0.1 | 0.116 | -0.012 |
| Sex ratio at birth (male births per female births) | -0.072 | -0.068 | -0.069 | 0.128 | -0.041 | -0.124 | 0.014 |
| Surface area (sq. km) | -0.019 | -0.184 | -0.004 | -0.008 | 0.078 | 0.016 | 0.033 |
| Survival to age 65, female (% of cohort) | -0.149 | 0.005 | -0.097 | 0.018 | -0.013 | 0.043 | 0.034 |
| Terrestrial and marine protected areas (% of total territorial area) | -0.026 | -0.019 | 0.089 | -0.022 | 0.124 | -0.055 | -0.118 |
| Terrestrial protected areas (% of total land area) | -0.033 | 0.021 | 0.075 | -0.089 | 0.074 | -0.093 | -0.086 |
| Total fisheries production (metric tons) | -0.019 | -0.241 | -0.003 | 0.013 | -0.067 | -0.058 | -0.006 |
| Total greenhouse gas emissions excluding LULUCF (% change from 1990) | 0.034 | -0.005 | -0.071 | -0.071 | 0.091 | -0.02 | -0.025 |
| Total greenhouse gas emissions excluding LULUCF per capita (t CO2e/capita) | -0.066 | -0.013 | -0.141 | -0.034 | 0.133 | 0.084 | -0.101 |
| Total natural resources rents (% of GDP) | 0.063 | -0.016 | -0.179 | 0.026 | 0.236 | -0.01 | -0.06 |

Table 3: PCA Loadings *(continued)*

| variable | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| Tuberculosis case detection rate (%, all forms) | -0.128 | 0.021 | -0.031 | -0.016 | -0.045 | 0.053 | 0.049 |
| Urban population | -0.022 | -0.275 | 0.008 | 0.017 | -0.056 | -0.001 | 0.013 |
| Urban population growth (annual %) | 0.132 | -0.049 | 0.007 | -0.041 | 0.048 | 0.014 | -0.058 |
| Voice and Accountability: Estimate | -0.107 | 0.054 | 0.144 | -0.161 | -0.005 | -0.047 | -0.007 |

Table 4: Country Clusters and Status

| Cluster | Developing | Developed |
|---|---|---|
| 1 | Afghanistan, Angola, Burundi, Benin, Burkina Faso, Central African Republic, Cote d'Ivoire, Cameroon, Congo, Dem. Rep., Congo, Rep., Comoros, Ethiopia, Gabon, Ghana, Guinea, Gambia, The, Guinea-Bissau, Equatorial Guinea, Kenya, Liberia, Lesotho, Madagascar, Mali, Mozambique, Mauritania, Malawi, Namibia, Niger, Nigeria, Papua New Guinea, Rwanda, Sudan, Senegal, Solomon Islands, Sierra Leone, Somalia, Sao Tome and Principe, Eswatini, Chad, Togo, Timor-Leste, Tanzania, Uganda, Yemen, Rep., Zambia, Zimbabwe | |
| 2 | Aruba, Antigua and Barbuda, Bahamas, The, Barbados, Chile, Costa Rica, Dominica, Grenada, Guam, Hong Kong SAR, China, St. Kitts and Nevis, St. Lucia, Macao SAR, China, Mauritius, Palau, French Polynesia, Singapore, Seychelles, Thailand, Trinidad and Tobago, Uruguay, St. Vincent and the Grenadines | Australia, Bermuda, Canada, Switzerland, Cyprus, Denmark, United Kingdom, Greenland, Ireland, Iceland, Israel, Korea, Rep., Luxembourg, Malta, Norway, New Zealand, Sweden |

| | | |
|---|---|---|
| 3 | Argentina, Bangladesh, Belize, Bolivia, Brazil, Bhutan, Botswana, Colombia, Cabo Verde, Dominican Republic, Ecuador, Egypt, Arab Rep., Fiji, Micronesia, Fed. Sts., Guatemala, Guyana, Honduras, Haiti, Indonesia, India, Jamaica, Jordan, Kyrgyz Republic, Cambodia, Kiribati, Lao PDR, Lebanon, Sri Lanka, Morocco, Mexico, Marshall Islands, Myanmar, Malaysia, Nicaragua, Nepal, Nauru, Pakistan, Panama, Peru, Philippines, Paraguay, West Bank and Gaza, El Salvador, Suriname, Syrian Arab Republic, Tonga, Tunisia, Turkiye, Viet Nam, Vanuatu, Samoa, South Africa | |
| 4 | China | United States |
| 5 | United Arab Emirates, Azerbaijan, Bahrain, Brunei Darussalam, Algeria, Iran, Islamic Rep., Iraq, Kazakhstan, Kuwait, Libya, Maldives, Mongolia, Oman, Qatar, Saudi Arabia, Tajikistan, Turkmenistan, Uzbekistan | |
| 6 | Armenia, Cuba, Georgia | Albania, Austria, Belgium, Bulgaria, Bosnia and Herzegovina, Belarus, Czechia, Germany, Spain, Estonia, Finland, France, Greece, Croatia, Hungary, Italy, Japan, Lithuania, Latvia, Moldova, North Macedonia, Montenegro, Netherlands, Poland, Portugal, Romania, Russian Federation, Serbia, Slovak Republic, Slovenia, Ukraine |
| 7 | Djibouti | |

## Code for Clustering Algorithm

### Function

```r
#' Hierarchical Cluster Analysis (HCA)
#'
#' Performs agglomerative hierarchical clustering using either single or complete linkage.
#'
#' @param D A square distance matrix (NxN) with pairwise distances between observations.
#' @param linkage A string specifying the linkage method; either "single" or "complete".
#' @param check_matrix_validity A boolean indicating whether to check if D is a valid
#'        distance matrix (symmetric with diag 0 and non-negative off diags).
#'
#' @return A list containing:
#' \describe{
#'   \item{merge}{A matrix representing the sequence of merges, similar to `hclust` output.}
#'   \item{height}{A numeric vector of the distances at which each merge occurred.}
#'   \item{merge_matrix}{A matrix showing cluster assignments at each step.}
#' }

hca=function(D, linkage=c("single", "complete"), check_matrix_validity=TRUE){
  "Explanation see above"
  N=nrow(D) #Number of Observations stored for convenience

  if (check_matrix_validity){
    msg="Not a correctly specified distance matrix"
    #check if matrix is square
    if(nrow(D)!=ncol(D)){stop(msg)}

    #Check each element for symmetry, non-negativity and diags do be 0
    for (i in 1:(N)){
      for (j in 1:(N)){
```

```
      if (i==j){if(D[i,j]!=0){stop(msg)}}
      else if (D[i,j]!=D[j,i] | D[i,j]<=0){stop(msg)}
    }}
  #print success, if matrix invalid, we would stop before here
  print("Matrix seems valid")}

#Initialize lingage function:
if (tolower(linkage)=="single"){
  linkage_calculator=function(g, h, clusters, D){
    distances=list()

    #extract index of observations in each cluster
    idx_g=which(clusters==g)
    idx_h=which(clusters==h)

    #get sub-matrix of D
    dist_matrix = D[idx_g, idx_h]
    #find minimum of distance
    min_dist_idx = which.min(dist_matrix)

    #convert the indices back to matrix index
    min_g = idx_g[(min_dist_idx - 1) %% length(idx_g) + 1]
    min_h = idx_h[(min_dist_idx - 1) %/% length(idx_g) + 1]

    return(list(min_distance = dist_matrix[min_dist_idx], min_g = min_g,
       min_h = min_h))

  }}
else if (tolower(linkage)=="complete"){
  linkage_calculator=function(g, h, clusters, D){
    distances=list()

    #extract index of observations in each cluster
    idx_g=which(clusters==g)
    idx_h=which(clusters==h)

    #get sub-matrix of D
    dist_matrix = D[idx_g, idx_h]
    #find maximum of distance
    min_dist_idx = which.max(dist_matrix)

```

```
      #convert the indices back to matrix index
      min_g = idx_g[(min_dist_idx - 1) %% length(idx_g) + 1]
      min_h = idx_h[(min_dist_idx - 1) %/% length(idx_g) + 1]

      return(list(min_distance = dist_matrix[min_dist_idx], min_g = min_g, min_h = min_h))
    }
  }
  else {stop("unimplemented linkage specification")}

  #Initialize numeric vector with one cluster per observation
  clusters=as.numeric(1:N)

  #get lists we will append to keep track of results
  merged_distance=numeric() #distances (height) for the merges
  merge_history=clusters # matrix of merges at each step
  h_list=numeric() #clusters that were merged into g
  g_list=numeric() #clusters that got merged into

  #initialize search list for all combinations, so that we don't have to get them each iteration
  #merged clusters will be removed successively
  search_combinations=as.matrix(t(combn(1:N, 2)))



  k=N #initialize iteration counter

  while(k>1){


    #update iteration counter
    k=k-1

    #initialize min distance and corresponding index:
    min_dist=Inf
    merge_g=NA
    merge_h=NA

    for (pair in 1:nrow(search_combinations)){

```

```
    #calculate linkage
    linkage=linkage_calculator(search_combinations[pair,1], search_combinations[pair,2], clusters, D)

    #update minimum distance and pairs
    if (linkage$min_distance<min_dist){
      min_dist=unlist(linkage$min_distance)
      merge_g=search_combinations[pair,1]
      merge_h=search_combinations[pair,2]
    }
  }

  #update cluster vector
  clusters[clusters==merge_h]=merge_g

  #update merge_history matrix
  merge_history=cbind(merge_history, clusters)

  #update lists for merged clusters
  h_list=c(h_list, merge_h)
  g_list=c(g_list, merge_g)


  #append merge distance, checking that it is larger than all before
  if (length(merged_distance)==0){
    merged_distance=c(merged_distance, unlist(min_dist))
  }
  else if (min_dist>max(merged_distance)){
    merged_distance=c(merged_distance, unlist(min_dist))
    }
  else{stop(paste("Minimum distance decreased, should not happen. K/N:",k,"/",N))}

  #Remove search options from combinations
  search_combinations <- search_combinations[search_combinations[,1] != merge_h, , drop=FALSE]
  search_combinations <- search_combinations[search_combinations[,2] != merge_h, , drop=FALSE]
}
#since we are always merging to the cluster with lower index, we can check if we correctly arrived at all cluster 1
```

```r
  if(any(as.integer(clusters))!=1){stop("Did not arrive at all in one cluster")}

  #------ finalize data outputs

  #get merge history matrix
  merge_history=(as.matrix(merge_history))
  colnames(merge_history)<-paste("Step",1:nrow(merge_history))

  #same output as merge in hclust package, just wiht different naming convention
  h_list=unname(h_list)
  g_list=unname(g_list)
  merge=as.matrix(t(rbind(g_list,h_list)))

  #same output as height in hclust package
  merged_distance=unname(merged_distance)

  return(list(merge=merge, height=merged_distance, merge_matrix=merge_history))
}
```

**Testing Script**

```r
#Simple script simulating data an testing the hca function against hclust
set.seed(1220)
source('/Users/johannesrenz/Library/Mobile Documents/com~apple~CloudDocs/BDS/ML2/Final_report/function/hca.R')
gen_data=function(n){

  #normal blob cluster
  cluster1 <- data.frame(
    x = rnorm(n/2, mean = 0, sd = 1.5),
    y = rnorm(n/2, mean = 0, sd = 1.5),
    group = factor(1))
  #concentric circles
  angles <- runif(n/2, 0, 2 * pi)
  radii <- rep(10, n/2)
  circles <- data.frame(
    x = radii * cos(angles) + rnorm(n/2, sd = 0.3),
```

```
    y = radii * sin(angles) + rnorm(n/2, sd = 0.3),
    group = factor(2))
  #Combine
  data <- rbind(cluster1, circles)
  return(data)
}

data=gen_data(100)

#Dist object for package
distances <- (dist(data[, c("x", "y")], method = "euclidean"))

#Matrix input for my function
distance_matrix <- as.matrix(distances)

#Run for single linkage

clusters=hca(distance_matrix, linkage="single")
clusters_package=hclust(distances, method="single")

if(all(clusters$height == clusters_package$height)){
  correct=1} else {print("heights dont match exactly")}

#Run for single linkage
clusters=hca(distance_matrix, linkage="complete")
clusters_package=hclust(distances, method="complete")

if(all(clusters$height == clusters_package$height)){
  if (correct==1){
    cat("\n\n","Successful, all merges are done with the same distances
       using either linkage")}
} else {print("heights dont match exactly")}
```

**Testing Script Output**

```

> source("~/Library/Mobile
  Documents/com~apple~CloudDocs/BDS/ML2/Final_report/function/hca_test.R",
  echo=F)
[1] "Matrix seems valid"
```

```
[1] "Matrix seems valid"


 Successful, all merges are done with the same distances using either
    linkage
```

## Code for the Report

```
#Analyse data

rm(list=ls())

library(rworldmap)
library(NbClust)
library(ggalt)
library(dplyr)
library(tidyr)
library(ggplot2)
library(xtable)
library(readr)
library(biplotEZ)
library(FactoMineR)
library(readxl)
library(stargazer)
library(Rfast)
library(kableExtra)
library(RColorBrewer)
library(missMDA)

#Set parameters:

textsize=20

#Set Parameter for Clustering and dimensional reduction
set.seed(1220) #Since Cluster Starting is random, we set a seed
K=7 # K is the number of clusters and sets the number of retained principle components
D=K+1 #D is the number of principle components to retain



#Utils:

#Tex formatting
escape_latex <- function(text) {
  text <- gsub("(% of total population)", "%", text)
  text <- gsub("female (% of female population)", "%f", text)
```

```r
  text <- gsub("(% of female population)", "% f", text)
  text <- gsub("\\\\", "\\\\textbackslash{}", text)
  text <- gsub("&", "\\\\&", text)
  text <- gsub("%", "\\\\%", text)
  text <- gsub("\\$", "\\\\$", text)
  text <- gsub("#", "\\\\#", text)
  text <- gsub("_", "\\\\_", text)
  text <- gsub("\\{", "\\\\{", text)
  text <- gsub("\\}", "\\\\}", text)
  return(text)
}
#Number formatting
format_number <- function(x) {
  x_rounded <- round(x, 4)
  if (abs(x_rounded) < 0.0001) {
    return("0") #if abs value is below 0.0001, display as "0"
  } else if (abs(x_rounded) >= 10^(-3) & abs(x_rounded) <= 10^3) {
    return(formatC(x_rounded, format = "f", digits = 4))
  } else {
    return(formatC(x_rounded, format = "e", digits = 4))
  }
}


setwd('/Users/johannesrenz/Library/Mobile Documents/com~apple~CloudDocs/BDS/ML2/Final_report')

df=read_csv("./Data/WB/2c50c149-0365-46a2-9545-ac89f31b79eb_Data.csv")

#correctly specify NA
df[df == ".."] <- NA


#Rename Columns
df <- df %>% rename(
  country = `Country Name`,
  alpha3 = `Country Code`,
  series = `Series Name`,
  var = `2018 [YR2018]`
)
print("--------------")
```

```r
print(paste("Percentage of Missing Values before handling anything:",
            round(sum(is.na(df$var))/length(df$var),2)))

#drop invalid obs
df <- df %>%
  filter(!is.na(country))

#Read UN Developing Nations Classification
un_classification=read_excel('./Data/UN/classification_nations_UN.xlsx')
developing_nations=un_classification%>%
  select(alpha3, status, m49)

#Read Regions
regions=read_excel("./Data/UN/regions.xlsx")
#Filter subregions
regions = regions[grepl("^[^0]*0$", regions$region_code), ]
#merge regions and classification
regions = merge(developing_nations, regions, by = "m49", all.x = TRUE)
regions = regions%>%
  select(alpha3, region, status)

#group regions into continents:
regions$Continent <- ifelse(regions$region %in% c("Latin America and the Caribbean"), "Latin America, Caribbean",
                            ifelse(regions$region %in% c("Northern America"), "North America",
                                   ifelse(regions$region %in% c("Southern Asia", "South-eastern Asia", "Eastern Asia"), "Asia",
                                          ifelse(regions$region %in% c("Sub-Saharan Africa", "Northern Africa"), "Africa",
                                                 ifelse(regions$region %in% c("Southern Europe", "Western Europe", "Eastern Europe", "Northern Europe"), "Europe",
                                                        ifelse(regions$region %in% c("Western Asia", "Central
```

```
                                                                                  Asia"), "Asia",
                                                                                       ifelse(regions$reg
                                                                                           ==
                                                                                           "Oceania",
                                                                                           "Oceania",
                                                                                           NA)))))))





df <- merge(df, regions, by = "alpha3")

#transform data to wide format
df <- df %>%
  select(country, region, Continent, alpha3, series, status, var)%>%
  pivot_wider(names_from = series, values_from = var)

#----
#Dropping data

df=as.data.frame(df)
df <- df[, sort(names(df))]

#Identify the first occurrence of each base name
col_base_names <- sub("[,|:].*", "", colnames(df))
col_base_names <- sub("\\(.*", "", col_base_names)
col_base_names <- trimws(col_base_names)

#keep only one measure per indicator
df <- df[, !duplicated(col_base_names)]

#initially drop vars with very low observations
df <- df[, colSums(is.na(df)) <=   0.1*nrow(df)]#drop low obs obs
df <- df[ rowSums(is.na(df)) <=   0.1*ncol(df),]#drop low obs vars

rownames(df)=df$country

#split into labels and data
df_lab=df%>%select(country, region, Continent, alpha3, status)
df=df%>%select(-country, -region, -Continent, -alpha3, -status)
```

```
#Convert type to num
df[] <- lapply(df, function(x) as.numeric(as.character(x)))

#Some diagnostic Printing
cat("\n --------------- \nMissing Countries are:\n",
    paste(un_classification$Country[un_classification$alpha3 %in%
      setdiff(un_classification$alpha3, df_lab$alpha3)] , collapse ="\n"))

cat("\n\n --------------- \nFraction of NA to be imputed is:",
    sum(is.na(df))/sum(is.na(df),!is.na(df)),
    "\n")

cat("\n\nDistribution of NA across Countries:\n")
print(summary(colSums(is.na(df))))
print(quantile(colSums(is.na(df)), probs = seq(0.1, 1, by = 0.05)),
   digits=1)

cat("\n\nDistribution of NA across Variables:\n")
print(summary(rowSums(is.na(df))))
print(quantile(rowSums(is.na(df)), probs = seq(0.1, 1, by = 0.05)),
   digits=1)

#Plot distribution of missings
data_variable <- data.frame(Missing = colSums(is.na(df)), Type = "Variable")
data_country <- data.frame(Missing = rowSums(is.na(df)), Type = "Country")

ggplot(data_variable, aes(x = Missing)) +
  geom_histogram(binwidth = 1, color = "black", fill = "grey") +
  labs(x = "Missing Values", y = "Frequency") +
  theme_minimal()+
  theme(
    text = element_text(size = textsize),
    axis.title = element_text(size = textsize),
    axis.text.x = element_text(size = textsize),
    axis.text.y = element_text(size = textsize)
  )
ggsave("./figures/missing_var.png", width = 8, height = 6)


ggplot(data_country, aes(x = Missing)) +
```

```
  geom_histogram(binwidth = 1, color = "black", fill = "grey") +
  labs(x = "Missing Values", y = "Frequency") +
  theme_minimal()+
  theme(
    text = element_text(size = textsize),
    axis.title = element_text(size = textsize),
    axis.text.x = element_text(size = textsize),
    axis.text.y = element_text(size = textsize)
  )
ggsave("./figures/missing_country.png", width = 8, height = 6)


cat("\n --------------- \n")

df=as.data.frame(df)

#imput NA using PCA
ndim=estim_ncpPCA(df, scale=TRUE) # estimate optimal number of PCA
   dimensions to use in imputation
impPCA=imputePCA(df, ncp=ndim$ncp, nboot=1, scale=TRUE) #for bootstrap
   replace with MIPCA

#When using MIPCA
#plot(impPCA, choice="dim")
#this gives a feeling for the uncertainty of the components. did not fit in
   the paper due to length contraints



tmp=df
df=as.data.frame(impPCA$completeObs)

rownames(df)=rownames(tmp)
colnames(df)=colnames(tmp)

#--- Summary Statistics

#function to create a tex table with kableExtra
create_latex_summary_table <- function(data, file_name) {
  summary_stats <- data %>%
    summarise(across(everything(), list(
```

```
      Mean = ~ mean(.x, na.rm = TRUE),
      SD = ~ sd(.x, na.rm = TRUE),
      Min = ~ min(.x, na.rm = TRUE),
      Max = ~ max(.x, na.rm = TRUE)
    ))) %>%
    pivot_longer(everything(), names_to = c("Variable", ".value"),
       names_sep = "_")

  summary_stats$Variable <- gsub("_", "_\\\\newline ",
     summary_stats$Variable)
  summary_stats$Variable <- sapply(summary_stats$Variable, escape_latex)

  summary_stats[, 2:5] <- lapply(summary_stats[, 2:5], function(col)
     sapply(col, format_number))

  table_latex <- summary_stats %>%
    kbl(format = "latex", booktabs = TRUE, longtable = TRUE, escape =
       FALSE, align = "lcccc",
        caption = "Descriptive Statistics Summary", label = "tab:app:sum")
           %>%
    kable_styling(latex_options = c("repeat_header"), font_size = 9) %>%
    column_spec(1, width = "8cm")
  writeLines(table_latex, file_name)
}
#run func
create_latex_summary_table(df, "./tables/summary_appendix.tex")


write_csv(df, "./Data/data_clean.csv")
write_csv(df_lab, "./Data/data_clean_labels.csv")



#----- ----- ----- ----- ----- ----- ----- -----
#Data Analysis Section
#----- ----- ----- ----- ----- ----- ----- -----


setwd('/Users/johannesrenz/Library/Mobile
   Documents/com~apple~CloudDocs/BDS/ML2/Final_report')
df=read_csv("./Data/data_clean.csv")
```

```
df_lab=read_csv("./Data/data_clean_labels.csv")

#----- ----- ----- ----- ----- ----- ----- -----
#----Analysis Part
#----- ----- ----- ----- ----- ----- ----- -----

#Convert df to Z-scores
df=scale(df, center = TRUE, scale = TRUE)


#Run the PCA using SVD
SVD=svd(df)

var_expl=SVD$d^2 / sum(SVD$d^2)
cumu_var_expl = cumsum(var_expl)

#manually compute the transformed coordinates by multiplying the centered
    data by PCs
#gives the projections of the data onto the principal components
transformed_coordinates = df %*% SVD$v

#limti to the first k transformed coordinates (dimensions 1 to number of
    clusters-1)
transformed_coordinates_k = transformed_coordinates[, 1:D]

#perform k-means clustering on the first k components
kmeans_result = kmeans(transformed_coordinates_k,
                       centers = K,
                       nstart=10000)


#----- ----- ----- ----- ----- ----- ----- -----
#Results section
#----- ----- ----- ----- ----- ----- ----- -----
#plotting and tables from here onwards
#----- ----- ----- ----- ----- ----- ----- -----

loadings=data.frame(variable=colnames(df),loading=SVD$v[,1])
loadings$loading=round(loadings$loading, digits=4)
```

```
#df with PCA component names
pc_df <- as.data.frame(round(SVD$v[, 1:D], 3))
colnames(pc_df) <- paste0("PC", 1:D)
pca_print <- data.frame(variable = colnames(df), pc_df)

#PCA long table with cumulative variance explained
create_PCA_long_table <- function(data, cumu_var_expl, file_name) {
  data_selected <- data %>% mutate(across(everything(), escape_latex))
  data_selected <- data_selected %>% mutate(across(where(is.numeric),
     round, digits = 2))

  #Create header for cumulative variance explained
  header_variance <- c("",paste0("(", round(cumu_var_expl[1:(ncol(data)-1)]
     * 100, 2), "%)"))
  header_combined <- c(header_pc, header_variance)

  table_latex <- data_selected %>%
    kbl(format = "latex", booktabs = TRUE, longtable = TRUE, escape = FALSE,
        align = "lcccc", caption = "PCA Loadings", label = "tab:app:pca")
           %>%
    kable_styling(latex_options = c("repeat_header"), font_size = 9) %>%
    column_spec(1, width = "8cm") %>%
    add_header_above(header_combined)

  write_lines(table_latex, file_name)
}


create_PCA_long_table(pca_print, cumu_var_expl,
   "./tables/loadings_appendix.tex")




#print the cumulative variance explained
cat("\n Cumulative Variance Explained by
   Component:",round(cumu_var_expl[1:10],2))
cat("\n Using our set of dimensions, we
   explain:",round(cumu_var_expl[D],2), "% of variance")

```

```

#df to make plotting easier
df_plot <- data.frame(
  X = transformed_coordinates[, 1],
  Y = transformed_coordinates[, 2],
  Region = df_lab$region,
  Country= df_lab$country,
  alpha3 = df_lab$alpha3,
  Continent=df_lab$Continent,
  Status=df_lab$status,
  Cluster = as.factor(unname(kmeans_result$cluster)))


#map plot
world_map <- getMap()
map_data <- merge(world_map, df_plot, by.x = "ISO_A3", by.y = "alpha3", all.x = TRUE)
colors <- brewer.pal(length(unique(df_plot$Cluster)), "Set3")  # Max 12 colors in 'Set3'

map_data$dev=ifelse(map_data$Status=="Developed", 1,0.4)
png("./figures/world_map.png")
mapCountryData(map_data, nameColumnToPlot = "Cluster",
               mapTitle = "",
               colourPalette = colors,
               catMethod = 'categorical',
               borderCol="#808080",
               nameColumnToHatch='dev',
               addLegend =  T)
dev.off()



print(table(df_plot$Region, df_plot$Cluster, df_plot$Status))

table_data <- table(df_plot$Cluster, df_plot$Status)
print(table_data)
for (cluster in 1:nrow(table_data)) {
  developed_fraction <- table_data[cluster, "Developed"] / sum(table_data[cluster, ])
  minority_group <- ifelse(table_data[cluster, "Developed"] <
```

```
    table_data[cluster, "Developing"], "Developed", "Developing")
  cat("\n Cluster", cluster, "- Developed Fraction:",
    round(developed_fraction,2), "%.\n Minority group:", minority_group,
    "\n Countries in the minority group:\n")
  print(df_plot[df_plot$Cluster == cluster & df_plot$Status ==
    minority_group, "Country"])
}

table_df <- data.frame(
  Cluster = 1:nrow(table_data),
  Developed = table_data[, "Developed"],
  Developing = table_data[, "Developing"],
  Fraction = round(table_data[, "Developed"] / rowSums(table_data) * 100,
    1),
  Minority = ifelse(table_data[, "Developed"] < table_data[, "Developing"],
    "Developed", "Developing"),
  Minority_Countries = sapply(1:nrow(table_data), function(cluster) {
    countries <- df_plot[df_plot$Cluster == cluster & df_plot$Status ==
      ifelse(table_data[cluster, "Developed"] < table_data[cluster,
      "Developing"], "Developed", "Developing"), "Country"]
    if (length(countries) > 0) paste(countries, collapse = ", ") else "None"
  })
)

tex=kbl(table_df, format = "latex", booktabs = TRUE,
    caption = "Cluster Composition and Minority Groups", label = "cluster")
      %>%
  column_spec(6, width = "7cm") %>%  # Use p{} for proper text wrapping
  footnote(general = "Note: The Developed Fraction represents the
    percentage of developed countries in each cluster.")
write_lines(tex, "./tables/cluster_assignment.tex")





#cluster compactness plot
nc <- NbClust(data=transformed_coordinates_k, min.nc = 2, max.nc = 25,
              method = "kmeans", index = "tracew")

index_data <- data.frame(Index = 2:25, Cluster = nc$All.index,
```

```
  var_expl=var_expl[1:24], cumulative_var_expl=cumu_var_expl[1:24])


p = ggplot(index_data, aes(x = Index)) +
  #plot variance explained as a line
  geom_line(aes(y = var_expl), color = "blue") +
  geom_point(aes(y = var_expl), color = "blue") +
  #plot cumulative variance explained as a second line
  geom_line(aes(y = cumulative_var_expl * max(index_data$var_expl) /
     max(index_data$cumulative_var_expl)), color = "red") +
  geom_point(aes(y = cumulative_var_expl * max(index_data$var_expl) /
     max(index_data$cumulative_var_expl)), color = "red") +
  theme_minimal() +
  labs(
    x = "Principal Component",
    y = "Variance Explained"
  ) +
  #secondary axis scaled so max is 1 but data remains unchanged
  scale_y_continuous(
    sec.axis = sec_axis(
      ~ . * max(index_data$cumulative_var_expl) / max(index_data$var_expl),
          # Adjust only axis scaling
      name = "Cumulative Variance Explained (Max = 1)"
    )
  ) +
  theme(
    text = element_text(size = textsize),
    axis.title = element_text(size = textsize),
    axis.text.x = element_text(size = textsize),
    axis.text.y = element_text(size = textsize)
  )

print(p)

ggsave("./figures/pca_scree.png", width = 8, height = 6)


q = ggplot(index_data, aes(x = Index)) +
  geom_line(aes(y = Cluster), color = "black") +
  geom_point(aes(y = Cluster), color = "black") +
  theme_minimal() +
```

```r
  labs(
    x = "Number of Clusters",
    y = "Compactness"
  ) +
  theme(
    text = element_text(size = textsize),
    axis.title = element_text(size = textsize),
    axis.text.x = element_text(size = textsize),
    axis.text.y = element_text(size = textsize)
  )

print(q)

ggsave("./figures/cluster_scree.png", width = 8, height = 6)

#Appendixx table version of map
df_latex <- df_plot %>%
  group_by(Cluster, Status) %>%
  summarise(Countries = paste(Country, collapse = ", "), .groups = "drop")
     %>%
  pivot_wider(names_from = Status, values_from = Countries, values_fill =
     list(Countries = ""))

groups_clusters_table <- kable(df_latex, format = "latex", booktabs = TRUE,
                     longtable = TRUE, escape = FALSE,
                     caption = "Country Clusters and Status",
                     label = "map_table") %>%
  column_spec(2, width = "8cm") %>%
  column_spec(3, width = "8cm")


write_lines(groups_clusters_table, "./tables/map_table.tex")
```