Global Trends in Mental Health: A Clustering and PCA-Based Approach

Maelice Yamdjieu, Ting-Yu Lin, & Vy Tran
5/27/2024

Abstract

This study explores the relationship between mental health issues and socio-economic factors across various countries in the world. Using a comprehensive dataset comprising mental health prevalence, GDP, and behavioral metrics from the World Data repository, Principal Component Analysis (PCA) and clustering techniques were applied to uncover underlying patterns and relationships. PCA was used to reduce the dimensionality of the data, capturing almost 90% of the variance. Subsequently, K-means clustering, optimized with the Elbow method achieved 45% accuracy, and hierarchical clustering were employed to identify distinct groups within the data. Hierarchical clustering achieved a higher clustering accuracy of 56% compared to k-means. The findings highlight significant associations between socio-economic indicators and mental health prevalence, providing valuable insights for policymakers and researchers focused on global mental health challenges.

Introduction

Mental health is essential to human well-being, influencing individual productivity, societal dynamics, and quality of life. However, its intricate relationship with socio-economic factors remains a subject of extensive research and debates. Understanding this relationship is vital for devising effective policies and interventions to address mental issues globally. In this study, an exploration of the intricate interplay between mental health prevalence and socioeconomic indicators across diverse countries worldwide is conducted using the "Mental Health" dataset from Our World in Data repository (IHME, Global Burden of Disease Study (2019) – processed by Our World in Data). This comprehensive dataset comprises three main tables: Mental Health Prevalence, GDP, and Behavioral Response, allowing for a global analysis of the relationship between mental health, economic indicators, and behavioral factors. The goal is to unravel the underlying patterns and correlations that shape mental health outcomes within different socio-economic contexts. Leveraging advanced analytical techniques such as Principal Component Analysis (PCA), K-means clustering, and hierarchical clustering are employed to distill meaningful insights from the dataset. By grouping countries based on their socio-economic status and mental health rates, this study aims to provide valuable information for creating global policies and interventions to improve mental well-being.

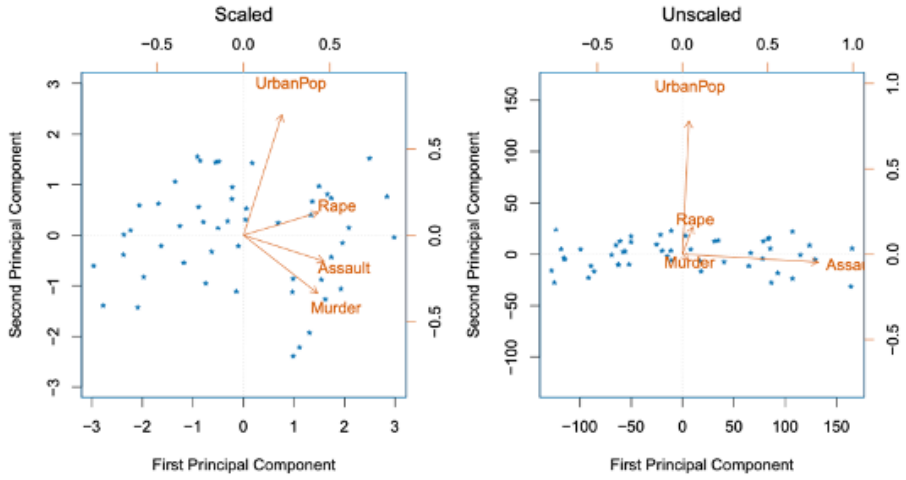
Theoretical Background

Unsupervised learning is a collection of statistical methods designed for situations where there is only a set of features X1, X2,…,Xp measured on n observations. Unlike supervising learning, the goal is not to make predictions, as there is no associated response variable Y. Instead, the aim is to uncover meaningful patterns and structures within the data by finding informative ways to visualize the measurements on X1, X2,…, Xp and identify any subgroups among the variables or observations. Unsupervised learning is more challenging and subjective compared to other models, often forming part of exploratory data analysis without a clear prediction goal. Assessing results in unsupervised learning is difficult due to the lack of a definitive mechanism for validation, as there is no response variable to predict.

Principal Component Analysis (PCA)

Principal components help summarize a large set of correlated variables by reducing them to a smaller number of representative variables that explain most of the original variability. These principal component directions indicate where the data varies most in feature space and define lines and subspaces that approximate the data cloud. Principal component analysis

(PCA) finds a low-dimensional representation of a dataset that captures as much variation as possible. In essence, each of the n observations exists in p-dimensional space, but not all dimensions are equally significant. PCA aims to identify a small number of the most interesting dimensions, where “interesting” is defined by the amount of variation observed along each dimension. Each dimension identified by PCA is a linear combination of the p features. PCA involves computing these components and using them to understand the data, making PCA an unsupervised method since it relies only on features X1, X2,…,Xp without a response variable Y. It is valuable for data visualization and data imputation (filling in missing values). Before performing PCA, the variables should be centered to have a mean of zero. Additionally, the results of PCA will depend on whether the variables have been individually scaled, meaning each variable is multiplied by a different constant.



[1]

Figure 1: Two principal components of biplots for the USArrests data

Two principal components of biplots for the USArrests data are shown. The left plot uses variables scaled to have unit standard deviations, while the right plot uses unscaled data.
PCA is also commonly implemented using Singular Value Decomposition(SVD), a technique that decomposes a matrix into three other matrices. For a dataset represented by an n*p matrix X, SVD is SVD is

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

where:

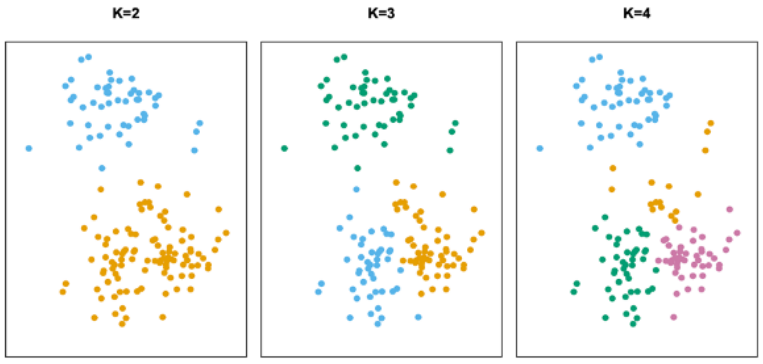
- $\mathbf{U}$ is an $n \times n$ orthogonal matrix.
- $\mathbf{\Sigma}$ is an $n \times p$ diagonal matrix with singular values (which represent the magnitude of each principal component).
- $\mathbf{V}$ is a $p \times p$ orthogonal matrix whose columns are the principal components.

By using SVD, PCA efficiently transforms the data into a new space defined by the principal components, facilitating dimensionality reduction while preserving the most significant patterns in the data.

K-means Clustering

Clustering involves techniques for identifying subgroups or clusters within a dataset, aiming to group similar observations together while differentiating distinct groups. The definition of similarity varies based on the data domain. Both clustering and PCA reduce data complexity but through different means: PCA seeks a low-dimensional representation that captures variance, while clustering aims to find homogeneous subgroups. Two prominent clustering

methods are K-means clustering, which partitions data into a predefined number of clusters, and hierarchical clustering, which does not predefine the number of clusters and produces a dendrogram to visualize cluster formation. Clustering can be applied to either observations based on features or features based on observations. One potential disadvantage of K-means clustering is that it requires specifying the numbers of clusters K in advance.



[1]

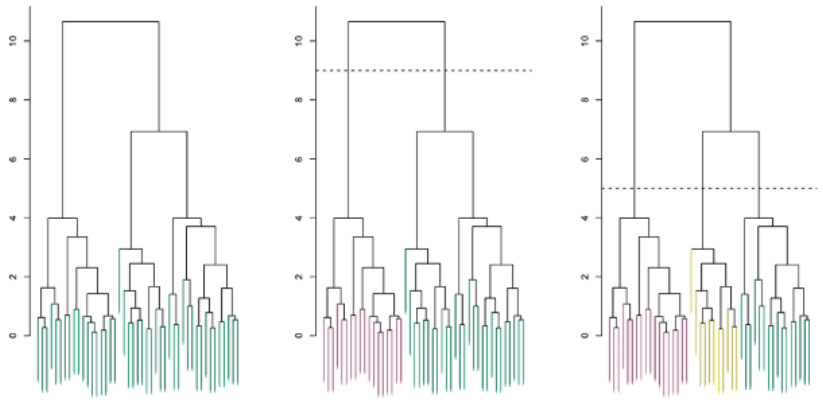
Figure 2: A simulated data set with 150 observations in two-dimensional space is analyzed using K-mean clustering with various values of K, the number of clusters.

Determining the optimal value of k is crucial for achieving accurate clustering results. The Elbow Method is a widely used technique for selecting the appropriate k. The process involves calculating the Sum of Squared Errors (SSE) for different values of k and plotting these values on a graph with k on the x-axis and SSE on the y-axis. The optimal k is typically found at the elbow point of the plot, where the SSE starts to level off. This point indicates that increasing k further would result in diminishing returns in terms of reducing SSE, thus balancing accuracy and computational efficiency.

Hierarchical Clustering

Hierarchical clustering, on the other hand, does not require a predetermined K and has the added advantage of producing a tree-based representation of the observations, known as a dendrogram. The hierarchical clustering dendrogram is created using a straightforward algorithm that begins by defining a dissimilarity measure between each pair of observations, often using Euclidean distance. Initially, each of the n observations is treated as its own cluster. Iteratively, the two most similar clusters are fused, reducing the number of clusters by one each time, until all observations belong to a single cluster, completing the dendrogram. A key challenge is defining the dissimilarity between clusters when they contain multiple observations. This issue is addressed using the concept of linkage, which extends the notion of dissimilarity from pairs of observations to pairs of clusters. The most common types of linkage are complete, average, single, and centroid.

The dendrogram can be cut at different heights to determine the final clusters, with the height representing the threshold distance for defining clusters. This flexibility allows for different levels of granularity in clustering. Hierarchical clustering offers several advantages, including the ability to visualize data structures through dendrograms and the flexibility of not needing to pre-specify the number of clusters. However, it also has drawbacks, such as being computationally intensive for large datasets and being sensitive to noise and outliers, which can significantly affect the results.



[1]

Figure 3: Dendrogram examples

The plot on the left is a dendrogram from hierarchical clustering using complete linkage and Euclidean distance. The center plot is a dendrogram, cut at a height of nine (dashed line), resulting in two distinct clusters. The right plot is a dendrogram cut at a height of five, resulting in three distinct clusters.

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

[1]

Table 1: Summary of the four most commonly used types of linkage in hierarchical clustering.

Hyperparameters

Selecting appropriate tuning parameters and addressing computational considerations are crucial for meaningful analysis. Key parameters include the number of principal components in PCA, which determines the dimensionality of the reduced space and the amount of variance captured, and the number of clusters in clustering algorithms like k-means, which can be guided by methods such as the Elbow Method. Computational complexity and memory usage are significant concerns, particularly with large datasets, as PCA involves SVD and clustering algorithms may require extensive distance calculations.

In conclusion, unsupervised learning encompasses a variety of techniques aimed at uncovering hidden patterns and structures within the data. PCA and clustering are two key methods that reduce data complexity ways. Both methods play a crucial role in exploratory data analysis, enabling researchers to gain insights and make informed decisions based on the inherent structures in the data. By centering and scaling variables appropriately, these techniques can be effectively applied across diverse domains, from medical research to market segmentation, demonstrating their versatility and importance in modern data analysis.

Methodology

Data processing and cleaning

Data processing and cleaning are crucial steps in preparing the dataset for analysis. In this study, the relevant data on mental health disorders, GDP per capita, and behavioral responses to mental health challenges were obtained from separate datasets for the year 2019. To ensure consistency and facilitate merging, the country identifiers were reformatted to match across all data sources as the 'Entity' column in each file was renamed to 'Country'. Then the rows and columns with missing or incomplete information were removed as mostly in country code and continents columns. The data was then combined using an inner join operation based on the country column, which resulted in a single comprehensive dataset containing information on various mental health disorders, economic indicators, and behavioral responses across different countries.

The variable names were reformatted and simplified to improve clarity and ease of use. For example, the prevalence of schizophrenia disorders, originally named 'Schizophrenia disorders (share of population) - Sex: Both - Age: Age-standardized', was changed to 'Schizophrenia_Disorders'. Similarly, 'GDP per capita, PPP (constant 2017 international $)' was renamed to 'GDP_per_Capita'. The behavioral response variables is have long name, such as 'Share - Question: mh8b - Engaged in religious/spiritual activities when anxious/depressed - Answer: Yes - Gender: all - Age group: all', were also renamed to more intuitive names like 'Engaged_in_Religious_Activities'. The resulting cleaned and preprocessed dataset provides a comprehensive view of mental health issues, socio-economic factors, and behavioral responses for each country.

Data Preparation and Introduction

After the initial data cleaning and preprocessing, it was observed that there were still few missing values in the dataset. To address this issue, the missing values were imputed using the mean of each variable. This imputation technique replaces the missing values with the average value of the corresponding variable, allowing for a complete dataset without any missing data points.

The final dataset consists of 101 countries and 15 variables, which include:

- Country: The name of the country.
- Schizophrenia_Disorders: How common of schizophrenia disorders in the population, represented as a percentage.
- Depressive_Disorders: How common of depressive disorders in the population, represented as a percentage.
- Anxiety_Disorders: How common are anxiety disorders in the population, represented as a percentage.
- Bipolar_Disorders: How common of bipolar disorders in the population, represented as a percentage.
- Eating_Disorders: How common eating disorders are in the population, represented as a percentage.
- GDP_per_Capita: The Gross Domestic Product per capita, adjusted for Purchasing Power Parity (PPP) in constant 2017 international dollars.
- Engaged_in_Religious_Activities: The percentage of individuals who engaged in religious or spiritual activities when they're feeling anxious or depressed.
- Improved_Healthy_Lifestyle: The percentage of people who adopt healthier lifestyle habits in response to feelings of anxiety or depression
- Changed_Work_Situation: The percentage of individuals who made changes to their work situation when dealing with anxiety or depression.

- Changed_Personal_Relationships: The percentage of individuals who made changes to their personal relationships as a way to cope with anxiety or depression.
- Talked_to_Friends_Family: The percentage of individuals who talked/reached out to friends or family when feeling anxious or depressed.
- Took_Medication: The percentage of individuals who took prescribed medication to help manage anxiety or depression.
- Spent_Time_in_Nature: The percentage of individuals who spent time in nature or outdoors when feeling anxious or depressed.
- Talked_to_Mental_Health_Professional: The percentage of individuals who talked to a mental health professional when feeling anxious or depressed.

This cleaned dataset, with imputed missing values using the mean of each variable, is now ready for exploratory analysis with further analysis using dimensionality reduction and clustering techniques to uncover meaningful insights into global mental health trends and their associations with socio-economic factors and behavioral responses.

Data Standardization

To perform the standardization, the mean and standard deviation were calculated for each variable, except for 'Country' and 'GDP_per_Capita'. Each value in these columns was then adjusted by subtracting the mean and dividing by the standard deviation. This ensured that all relevant variables had a mean of 0 and a standard deviation of 1, resulting in a neatly standardized dataset. Then, the 'Country' and 'GDP_per_Capita' columns were then reattached to the standardized dataset, as they were not subjected to the standardization process. This final standardized dataset, along with the separate 'Country' and 'GDP_per_Capita' columns, is now ready for principal component analysis (PCA) and clustering techniques.

Principal Component Analysis

After the data preprocessing and standardization steps, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data for simplifying the data while keeping the most important information. PCA identifies the principal components, which are linear combinations of the original variables that capture the maximum variance in the data. By choosing a subset of these principal components, the data is effectively represented in a lower-dimensional space, which makes it easier to visualize and analyze patterns.

To determine the optimal number of principal components, the scree plot and the cumulative explained variance were examined. The scree plot displays the eigenvalues associated with each principal component, ordered from largest to smallest. The plot helps to identify the number of principal components that capture a significant amount of variance in the data.

Additionally, the cumulative explained variance plot shows the proportion of total variance explained by each principal component. A threshold of 70% to 90% cumulative explained variance is often used as a criterion for selecting the number of principal components to retain. This threshold ensures that a sufficient amount of the original information is preserved while reducing the dimensionality of the data. By considering both the scree plot and the cumulative explained variance, the optimal number of principal components can be determined. These selected principal components can then be used for further analysis, such as clustering or visualization, to explore patterns and identify relationships among the countries based on their mental health profiles and behavioral responses with economic income levels.

K-mean Clustering

The Elbow method was used to determine the optimal number of clusters for the K-means algorithm. Based on the optimal number of clusters determined by the Elbow method, K-means clustering was performed on the PCA-transformed data. The Elbow method plots the

within-cluster sum of squares (WCSS) against the number of clusters. The optimal number of clusters is identified as the point where the decrease in WCSS begins to level off, forming an "elbow" in the plot.

K-means clustering aims to assign each data point to the k-cluster whose centroid is nearest to it in the feature space. The algorithm iteratively updates the cluster centroids until convergence is reached. The resulting clusters represent groups of countries with similar mental health and behavioral response patterns.

Hierarchical Clustering

Hierarchical clustering using the agglomerative approach was applied to the PCA-transformed data to identify meaningful groupings of countries based on their mental health profiles and behavioral responses. By using the PCA-transformed data, the clustering algorithm can focus on the most important patterns and variations in the data, as captured by the selected principal components. So, the agglomerative approach starts with each country as a separate cluster and iteratively merges the most similar clusters until a single cluster containing all countries is formed. The dissimilarity between clusters is measured using a distance metric, such as Euclidean distance or correlation distance, computed based on the PCA-transformed features.

The dendrogram, a tree-like diagram, is used to visualize the hierarchical structure of the clusters and determine the optimal number of clusters. By examining the dendrogram and considering the height at which clusters are merged, a dissimilarity threshold can be selected to cut the dendrogram and obtain the final clustering solution.

Accuracy comparison

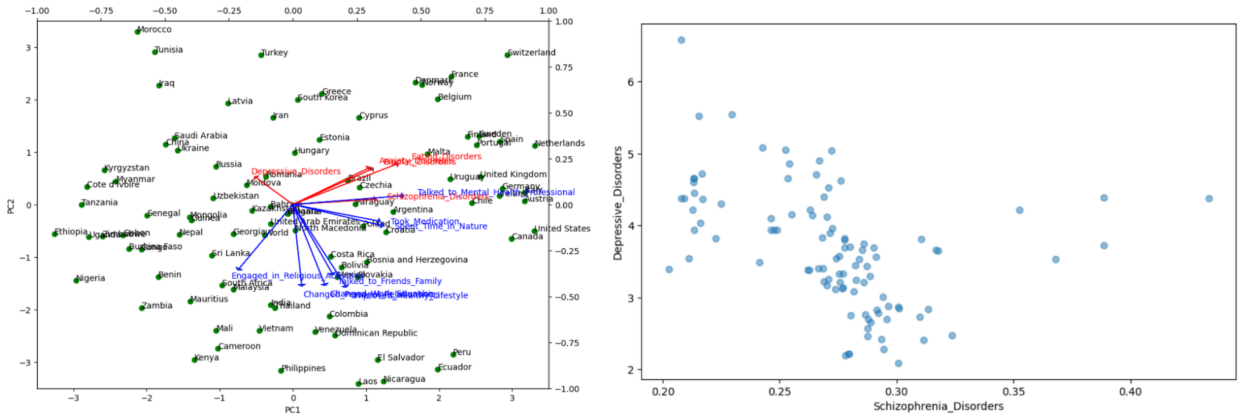
To compare the accuracy of K-means clustering and Hierarchical clustering in aligning with the income levels and mental health profiles of the countries, the confusion matrices were constructed. The income levels were categorized into three groups: low, medium, and high, based on the GDP per capita values. For each clustering technique, the confusion matrix was created by comparing the predicted cluster labels with the actual income level categories. The accuracy was calculated as the sum of the diagonal elements (correctly classified countries) divided by the total number of countries. By comparing the accuracies obtained from the confusion matrix, the performance of K-means clustering and Hierarchical clustering was evaluated to determine which clustering technique better aligns with the income levels of the countries with mental health profiles.

Computation Results

PCA(Principal Component Analysis)

The dataset contains 13 variables after leaving out Country and GDP columns. PCA was conducted to explore whether the dimensionality of the data could be reduced and to identify which variables contribute most to the variability in the dataset.

Figure 4(a) shows the first two principal components. The red vectors in the plot represent different types of mental health illnesses, while the blue vectors indicate the actions taken when individuals feel anxiety and depression. PC1 in the figure mainly captures the main variance in the data set. In contrast, PC2 captures variance orthogonal to PC1, revealing unique patterns and insights that are not apparent when considering PC1 alone. The positioning of countries in Figure 4(a) reflects proximity to others with similar characteristics regarding mental health disorders and their responses. Figure 4(b) illustrates the scatter plot of two original features. It can be observed that the distribution of points significantly differs from that seen when plotting the data in the first two principal components.



(a) (b)

Figure 4 (a)Relationships and contributions of mental health disorders and responses across countries. (b)Scatter plot of two original features

Figure 5 displays the proportion of variance explained by each component and the cumulative proportion of variance explained. The first principal component explains 31% of the variance in the data, while the next principal component explains 25% of the variance. The slope decreases very sharply from the first to the third principal component, and then becomes more gradual. This pattern suggests that the first three components capture a large amount of information. Collectively, the first seven principal components captured nearly 90% of the total variation in the data set. Based on these findings, reducing the dataset to seven principal components captures the majority of the features, making it an practical choice for training subsequent models.

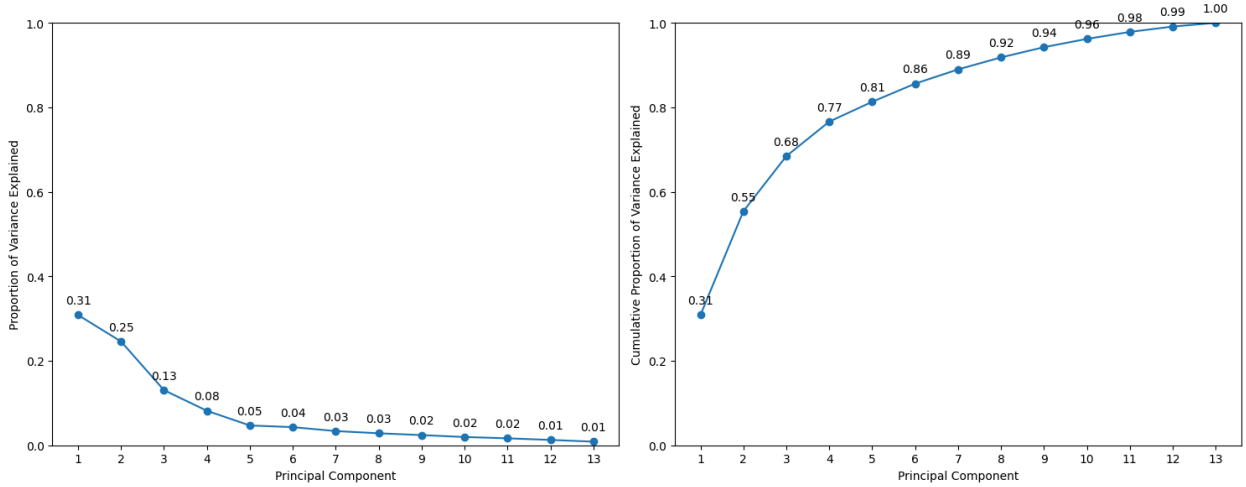


Figure 5: Variance Explained by Principal Components in PCA Analysis

Table 2 displays the loadings of the principal components for each feature in our dataset. The analysis shows that PC1 is strongly influenced by features such as "Eating Disorders" and "Talked to a Mental Health Professional," implying a significant impact on the primary variation direction within the data set.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Schizophrenia_Disorders	0.318025	0.028095	-0.319116	0.365425	-0.173102	0.550886	0.029461	0.322345	-0.300688	0.163172	-0.180206	0.064020	0.266222
Depressive_Disorders	-0.142352	0.145258	0.560676	-0.208061	-0.043823	0.380047	-0.605532	0.272442	-0.018298	-0.032804	0.071458	-0.008785	0.084526
Anxiety_Disorders	0.293367	0.196363	0.356111	0.329893	0.268824	0.136298	0.332534	0.028022	-0.052160	-0.416028	0.458742	0.194864	-0.115365
Bipolar_Disorders	0.306578	0.191735	0.399414	0.223443	0.011226	-0.435143	-0.013479	-0.071904	0.115969	0.299530	-0.239206	0.112757	0.540242
Eating_Disorders	0.402929	0.218024	0.188199	0.018690	-0.073910	0.151650	-0.026667	-0.222089	0.096854	0.288571	-0.344874	-0.075546	-0.677164
Engaged_in_Religious_Activities	-0.209772	-0.348580	0.339975	0.113710	0.093509	0.065988	0.416347	0.438048	0.198895	-0.114618	-0.447808	-0.266667	-0.057497
Improved_Healthy_Lifestyle	0.205150	-0.441451	0.039301	0.017293	-0.053066	0.175351	0.027977	0.017729	0.482715	0.457693	0.512744	-0.137285	0.047747
Changed_Work_Situation	0.124116	-0.432144	0.143523	0.045352	-0.494190	0.151835	-0.069323	-0.429834	0.118065	-0.427898	-0.183436	0.277010	0.104558
Changed_Personal_Relationships	0.034646	-0.435005	0.172009	0.295617	-0.140675	-0.384553	-0.222468	0.165047	-0.571597	0.130823	0.155199	-0.016180	-0.282853
Talked_to_Friends_Family	0.153376	-0.373935	0.064361	-0.230379	0.689159	0.216746	-0.045699	-0.330340	-0.297454	0.085298	-0.170645	0.030438	0.143175
Took_Medication	0.332441	-0.089437	-0.013205	-0.603161	-0.090756	-0.143571	0.186456	0.429487	-0.037492	0.022044	-0.014831	0.512625	-0.067356
Spent_Time_in_Nature	0.345320	-0.113424	-0.298431	0.182819	0.300375	-0.205425	-0.498221	0.261790	0.370573	-0.363905	-0.129428	-0.050870	-0.084451
Talked_to_Mental_Health_Professional	0.424508	0.046896	0.036751	-0.333307	-0.203188	-0.073049	0.086532	-0.016611	-0.207985	-0.251412	0.092480	-0.711462	0.171952

Table 2: Loading of original features to each principal component

K-mean clustering

The data processed by PCA was used for k-mean clusteing. The elbow graph helps determine the optimal number of clusters by identifying the point where the within-cluster sum of squares (WCSS) begins to decrease at a significantly slower rate. From figure 6, it appears that the decrease in WCSS levels off after three clusters, suggesting that adding more clusters beyond this point may not significantly improve the clustering effectiveness. Therefore, choosing three clusters appears to be an appropriate decision.

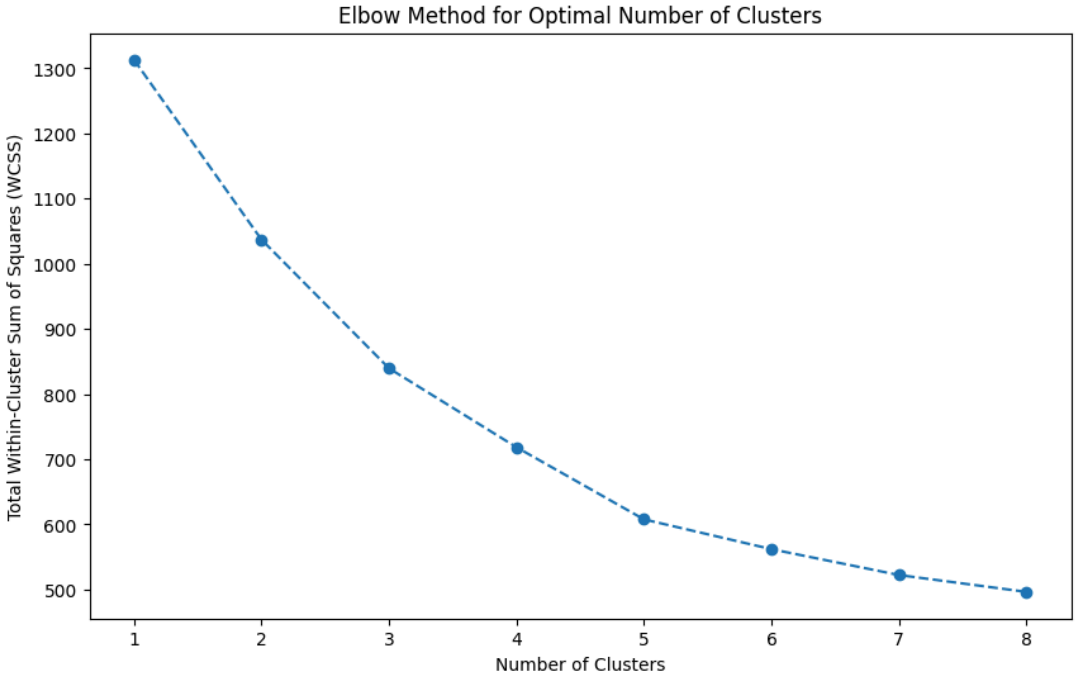


Figure 6: Elbow graph

Figure 7 illustrates how countries are grouped according to their data on mental health problems and the methods used to manage depression and anxiety. It is displayed in two dimensions for ease of visualization, the figure shows that countries within the same cluster may have similar prevalence rates or approaches to dealing with these mental health challenges. As In Table 3, Cluster 2 displayed stronger associations with most mental health problems compared to the other clusters. Table 4 shows the probabilities of various behaviors that people adopt to cope with anxiety and depression across each cluster. Cluster 0 and 1 show higher probabilities in most behaviors than Cluster 2.

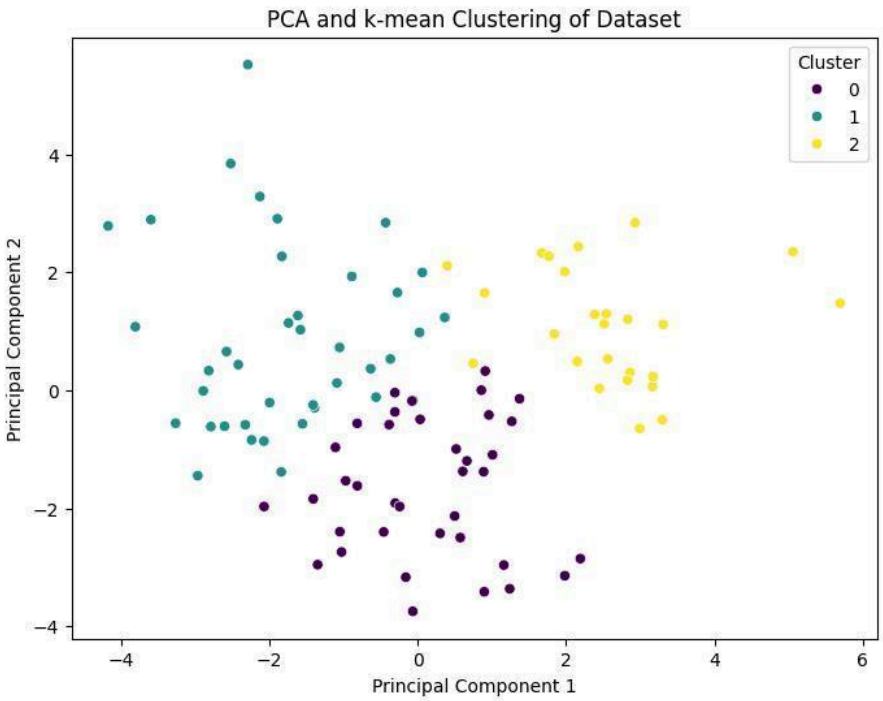


Figure 7: PCA and K-Means clustering

Cluster	Schizophrenia_Disorders	Depressive_Disorders	Anxiety_Disorders	Bipolar_Disorders	Eating_Disorders
0	0.028547	0.461362	0.426790	0.065769	0.017532
1	0.033801	0.392182	0.477954	0.074576	0.021487
2	0.026235	0.332712	0.517715	0.081614	0.041723

Table 3: Probabilities of mental health issues for each cluster

Cluster	0	1	2
Engaged_in_Religious_Activities	0.106953	0.101905	0.056337
Improved_Healthy_Lifestyle	0.152100	0.153404	0.145661
Changed_Work_Situation	0.100463	0.113056	0.100637
Changed_Personal_Relationships	0.133724	0.134904	0.113064
Talked_to_Friends_Family	0.181332	0.166109	0.163850
Took_Medication	0.099723	0.096659	0.123757
Spent_Time_in_Nature	0.155014	0.153284	0.154972
Talked_to_Mental_Health_Professional	0.070691	0.080680	0.141723

Table 4: Probabilities of behavioral variables for each cluster

Figure 8 is a confusion matrix that illustrates the relationship between PCA-based clusters and income levels. Cluster 2 contains more economically developed countries, while

Clusters 0 and 1 include countries that are relatively less developed economically. The model is to predict income levels resulted in an accuracy of 0.45.

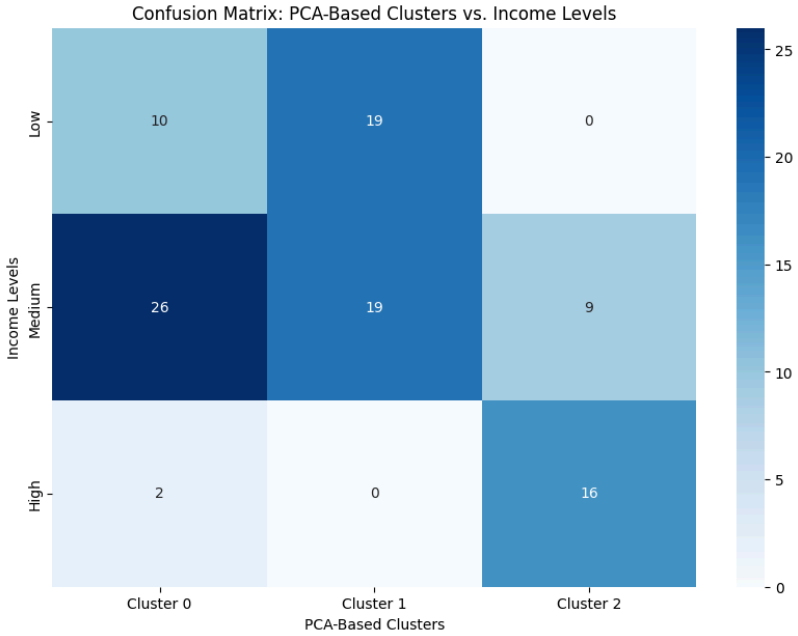


Figure 8: Confusion matrix for K-means clustering vs. income level

Hierarchical Clustering

Figure 9 depicts a hierarchical clustering dendrogram that categorizes countries based on their similarity. Lower linkage points in the dendrogram suggest a higher similarity among the countries. The height of each linkage point on the Y-axis signifies the distance necessary to form that cluster, with greater heights indicating more significant differences among the countries within each cluster. As figure 6 shows that we divided the data into three distinct clusters.

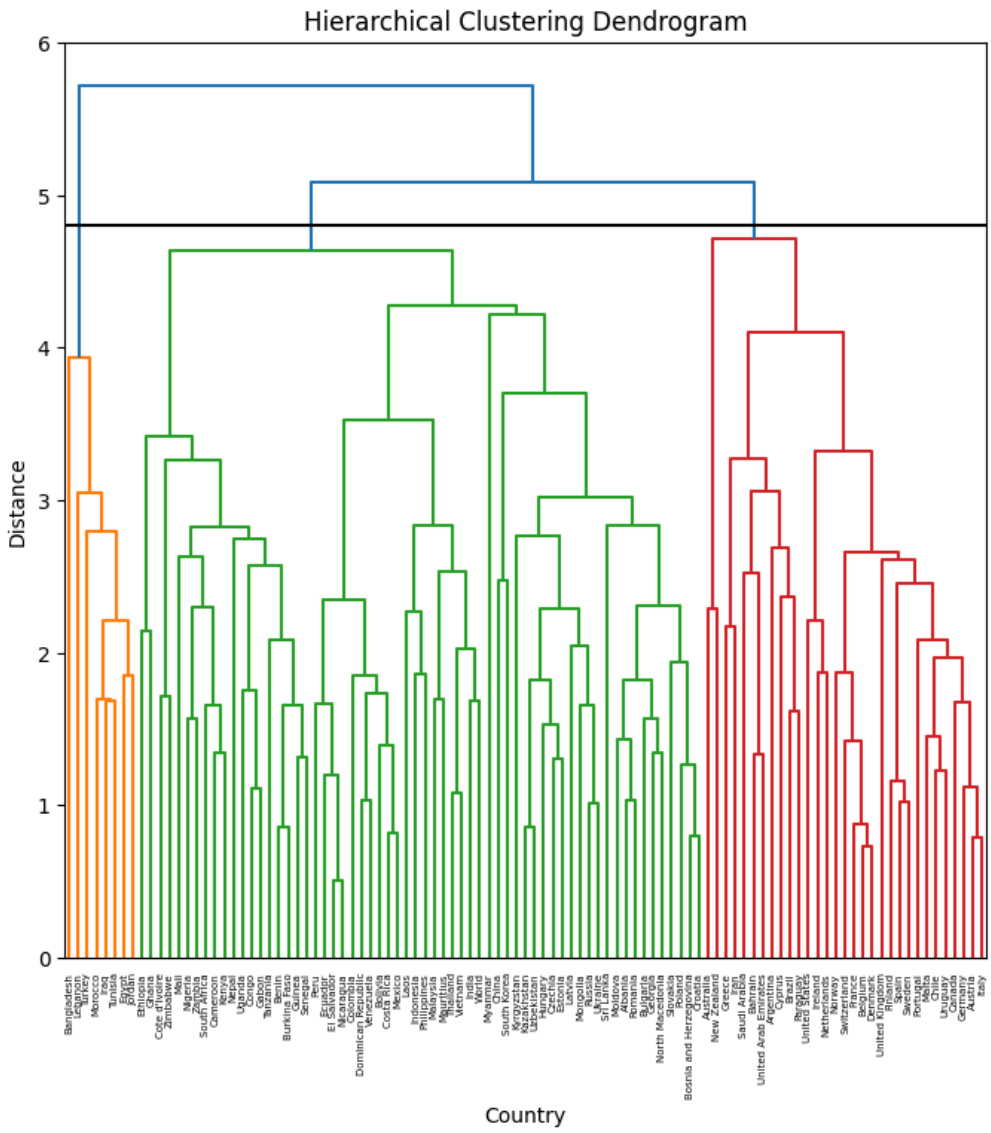


Figure 9: Hierarchical clustering dendrogram

Figure 10 is a confusion matrix that displays the relationship between hierarchical clustering and income levels. It follows a similar pattern to Figure 9, where one cluster predominantly consists of economically developed countries, while the other two clusters mainly include countries that are relatively less developed economically. The model to predict income levels achieved an accuracy of 0.56, which is higher than the accuracy obtained using K-means clustering on PCA-transformed data.

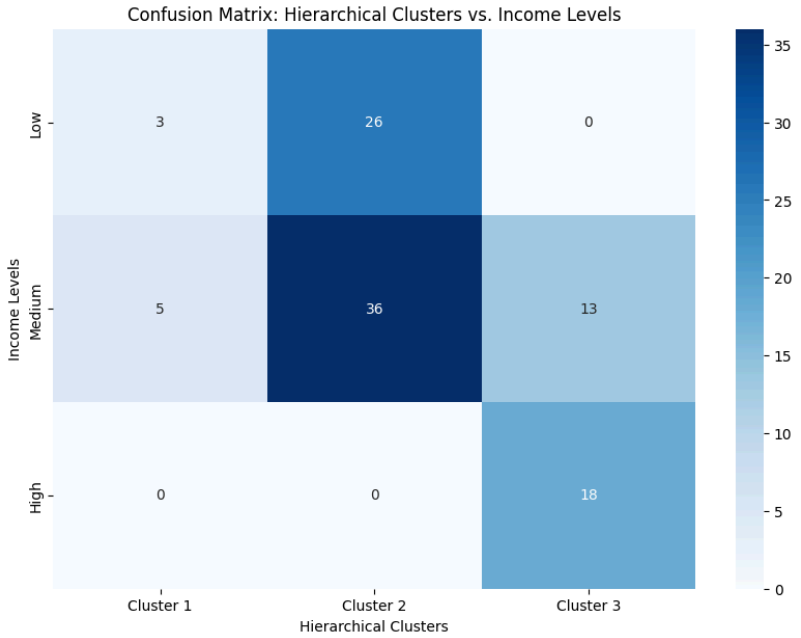


Figure 10: Confusion matrix for hierarchical clustering vs. income level

Discussion

After processing the dataset with PCA, "Scores" are obtained, which project original data points onto principal component axes. These scores show each data point's relation to the principal components, helping to assess their position in the new feature space. This enables us to observe how various variables, such as mental disorders and related behaviors, distribute across principal components. The "Rotation" in PCA refers to loadings, which show how original variables form principal components. Notably, features like "Eating Disorders" and "Talked_to_Mental_Health_Professional" strongly influence PC1. This suggests that when dealing with mental health issues, especially eating disorders, the frequency of seeking professional mental health counseling may increase. Insights like this can inform targeted health interventions, such as improving accessibility of mental health services for those with eating disorders.

Plotting the first two principal components provides a more abstract but informative view of the data. It focuses on the most important inherent relationships, rather than merely the direct relationships between two variables. This approach can reveal patterns and relationships not apparent when simply plotting two original features. The plot of original features typically shows direct and often simplistic relationships, lacking the dimensionality reduction that captures more complex patterns.

After performing PCA and subsequent clustering, the data was primarily divided into three distinct clusters. From the results, it is evident that one of these groups consists predominantly of countries with medium or high income levels. This cluster also showed stronger associations with a wider range of mental health problems than other clusters. This pattern suggests that higher-income countries might have better health surveillance and reporting systems, which could lead to more frequent diagnosis and reporting of mental health issues. Or it might indicate higher stress levels associated with the lifestyle in more affluent societies. Furthermore, it has been observed that within this cluster, individuals are more likely to engage in actions such as "Took Medication" and "Talked to a Mental Health Professional" when dealing with anxiety and depression, compared to other clusters. Both of these actions generally involve higher costs, which correlates with the higher income levels of this cluster, indicating a greater financial capability to afford such measures.

Based on these findings, there is a critical need to enhance mental health infrastructure in lower-income countries by improving diagnostic capabilities and training healthcare professionals to tackle under-reported cases. Additionally, affluent societies would benefit from targeted mental health interventions specifically designed to address stressors such as job pressure.

Conclusion

The findings of this study have significant implications for understanding the global landscape of mental health and its relationship with socio-economic factors and behavioral responses. By applying advanced analytical techniques such as Principal Component Analysis (PCA), K-means clustering, and hierarchical clustering to a comprehensive dataset, the meaningful patterns and insights have been uncovered that can inform policy decisions and interventions aimed at addressing mental health challenges worldwide. The analysis reveals that countries with higher income levels tend to have stronger associations with a wider range of mental health problems compared to lower-income countries, suggesting the need for improved mental health surveillance and reporting systems in lower-income nations. Additionally, individuals in higher-income countries are more likely to engage in actions such as seeking professional mental health support and taking medication when dealing with anxiety and depression, highlighting the importance of improving access to mental health services and increasing affordability of treatment options in lower-income countries.

The application of PCA has allowed for the identification of key variables that contribute most to the variability in the dataset, providing a more focused understanding of the complex relationship between mental health disorders and behavioral responses. The comparison of K-means clustering and hierarchical clustering techniques has revealed that hierarchical clustering achieves a higher accuracy in aligning with the income levels of countries, suggesting its potential as a more effective tool for understanding the relationship between mental health and economic development. This finding can guide future research and analysis in this domain, enabling more precise targeting of interventions and resource allocation.

In conclusion, this study provides valuable insights into the global mental health landscape and its association with socio-economic factors and behavioral responses. By leveraging advanced analytical techniques, key patterns and relationships have been identified that can inform policy decisions and interventions aimed at improving mental health outcomes worldwide. The broader impact of these findings lies in their potential to guide the development of targeted strategies that address the unique challenges faced by different countries and communities, ultimately contributing to the promotion of mental well-being on a global scale. As the world continues to navigate the complexities of mental health in an increasingly interconnected world, studies like this serve as a critical foundation for evidence-based decision-making and collaborative efforts to support the mental health needs of individuals and societies across the globe.

References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Retrieved from https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html
2. Saloni Dattani, Lucas Rodés-Guirao, Hannah Ritchie and Max Roser (2023) - “Mental Health” Published online at OurWorldInData.org. Retrieved from https://ourworldindata.org/mental-health

Appendix

Data Preprocessing

```
import pandas as pd
# mental-illnesses-prevalence
df = pd.read_csv('mental-illnesses-prevalence.csv')  # Extract data in 2019
df_2019 = df[df['Year'] == 2019]  # Save the result as a new CSV file
df_2019.to_csv('mental-illnesses-prevalence_2019.csv', index=False)
# depressive-disorders-prevalence-vs-gdp-per-capita
df1 = pd.read_csv('depressive-disorders-prevalence-vs-gdp-per-capita.csv')
df1_2019 = df1[df1['Year'] == 2019]
df1_2019.to_csv('depressive-disorders-prevalence-vs-gdp-per-capita_2019.csv',
index=False)
# Read the three CSV files
df_2019 = pd.read_csv('mental-illnesses-prevalence_2019.csv')
df1_2019 =
pd.read_csv('depressive-disorders-prevalence-vs-gdp-per-capita_2019.csv')
df2 = pd.read_csv('dealing-with-anxiety-depression-comparison.csv')
df2 = df2.drop(columns=['Year'])

# Rename the 'Entity' column to 'Country' in each DataFrame
df_2019.rename(columns={'Entity': 'Country'}, inplace=True)
df1_2019.rename(columns={'Entity': 'Country'}, inplace=True)
df2.rename(columns={'Entity': 'Country'}, inplace=True)
# Drop rows where the 'code' column doesn't have data
df_2019.dropna(subset=['Code'], inplace=True)
df1_2019.dropna(subset=['Code'], inplace=True)
df2.dropna(subset=['Code'], inplace=True)
df2.rename(columns={
    'Share - Question: mh8b - Engaged in religious/spiritual activities when
anxious/depressed - Answer: Yes - Gender: all - Age group: all':
'Engaged_in_Religious_Activities',
    'Share - Question: mh8e - Improved healthy lifestyle behaviors when
anxious/depressed - Answer: Yes - Gender: all - Age group: all':
'Improved_Healthy_Lifestyle',
    'Share - Question: mh8f - Made a change to work situation when
anxious/depressed - Answer: Yes - Gender: all - Age group: all':
'Changed_Work_Situation',
    'Share - Question: mh8g - Made a change to personal relationships when
anxious/depressed - Answer: Yes - Gender: all - Age group: all':
'Changed_Personal_Relationships',
    'Share - Question: mh8c - Talked to friends or family when
anxious/depressed - Answer: Yes - Gender: all - Age group: all':
'Talked_to_Friends_Family',
    'Share - Question: mh8d - Took prescribed medication when
anxious/depressed - Answer: Yes - Gender: all - Age group: all':
'Took_Medication',
    'Share - Question: mh8h - Spent time in nature/the outdoors when
anxious/depressed - Answer: Yes - Gender: all - Age group: all':
'Spent_Time_in_Nature',
    'Share - Question: mh8a - Talked to mental health professional when
```

```
anxious/depressed - Answer: Yes - Gender: all - Age group: all':
'Talked_to_Mental_Health_Professional'
}, inplace=True)

# Performing an inner join
merged_inner = df_2019.merge(df1_2019, on=['Country', 'Code', 'Year'] ,
how='inner').merge(df2, on=['Country'], how='inner')
merged_inner = merged_inner.drop(columns=['Code', 'Year', 'Continent',
                                          'Depressive disorders (share of
population) - Sex: Both - Age: Age-standardized_y'])
# Save the combined DataFrame to a new CSV file
merged_inner.to_csv('mental_health_issues.csv', index=False)
```

Unsupervised learning techniques

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.metrics import confusion_matrix, accuracy_score
from sklearn.cluster import AgglomerativeClustering
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
# Load the data
data = pd.read_csv('mental_health_issues.csv')

# Basic statistics and missing values
print(data.isnull().sum())
```

```
Country                                 0
Schizophrenia_Disorders                 0
Depressive_Disorders                    0
Anxiety_Disorders                       0
Bipolar_Disorders                       0
Eating_Disorders                        0
GDP_per_Capita                          1
Engaged_in_Religious_Activities         1
Improved_Healthy_Lifestyle              0
Changed_Work_Situation                  0
Changed_Personal_Relationships          0
Talked_to_Friends_Family                0
Took_Medication                         0
Spent_Time_in_Nature                    0
Talked_to_Mental_Health_Professional    0
dtype: int64
```

```
# Fill missing values with the mean of the column
data['GDP_per_Capita'].fillna(data['GDP_per_Capita'].mean(), inplace=True)
data['Engaged_in_Religious_Activities'].fillna(data['Engaged_in_Religious_Act
ivities'].mean(), inplace=True)
```

PCA

```
# Drop specific columns
data_clean = data.drop(columns=['GDP_per_Capita'])
# Initialize the StandardScaler
scaler = StandardScaler()
# Scale the data excluding the 'Country' column
scaled_data = scaler.fit_transform(data_clean.iloc[:, 1:])
# Convert scaled data back to DataFrame
scaled_data = pd.DataFrame(scaled_data, columns=data_clean.columns[1:])
# Handling missing values by filling with the median
for col in data.columns[1:]:  # Assuming the first column is 'Country' and
doesn't need imputation
    if data[col].isnull().any():
        data[col].fillna(data[col].median(), inplace=True)# Perform PCA
pca = PCA()
pca_out = pca.fit_transform(scaled_data)
pd.DataFrame(pca.components_.T
                  , index=scaled_data.columns
                  , columns=['PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6', 'PC7',
'PC8', 'PC9', 'PC10', 'PC11', 'PC12', 'PC13'])
# Perform PCA
pca = PCA()
pca_out = pca.fit_transform(scaled_data_df)
# Create plot
fig, ax1 = plt.subplots(figsize=(11, 8))
# Set limits
ax1.set_xlim(-3.5, 3.5)
ax1.set_ylim(-3.5, 3.5)
# Plot country names for PC1 and PC2
for i, country in enumerate(data_clean['Country']):
    ax1.annotate(country, (pca_out[i, 0], pca_out[i, 1]))
# Plot Principal Component loading vectors using a twin axis.
ax2 = ax1.twinx().twiny()
ax2.set_ylim(-1, 1)
ax2.set_xlim(-1, 1)
# Plot principal components as arrows
for i, (comp, var) in enumerate(zip(pca.components_.T,
scaled_data_df.columns)):
    color = 'r' if i < 5 else 'b'  # Blue for the first 5 components, gray
for the rest
    ax2.arrow(0, 0, comp[0], comp[1], head_width=0.03, head_length=0.01,
ec=color)
    ax2.text(comp[0]*1.15, comp[1]*1.15, var, color=color)

ax1.set_xlabel("PC1")
ax1.set_ylabel("PC2")
plt.show()

fig, axes = plt.subplots(1, 2, figsize=(15, 6))
ticks = np.arange(len(pca.explained_variance_ratio_)) + 1
# Plot for the proportion of variance explained by each component
```

```
ax = axes[0]
ax.plot(ticks,
        pca.explained_variance_ratio_,
        marker='o')
ax.set_xlabel('Principal Component')
ax.set_ylabel('Proportion of Variance Explained')
ax.set_ylim([0, 1])
ax.set_xticks(ticks)
# Annotating each point with the variance explained
for i, v in enumerate(pca.explained_variance_ratio_):
    ax.text(ticks[i], v + 0.02, f"{v:.2f}", ha='center', va='bottom')
# Plot for the cumulative proportion of variance explained
ax = axes[1]
cumulative_variance = np.cumsum(pca.explained_variance_ratio_)
ax.plot(ticks,
        cumulative_variance,
        marker='o')
ax.set_xlabel('Principal Component')
ax.set_ylabel('Cumulative Proportion of Variance Explained')
ax.set_ylim([0, 1])
ax.set_xticks(ticks)
# Annotating each point with the cumulative variance explained
for i, v in enumerate(cumulative_variance):
    ax.text(ticks[i], v + 0.02, f"{v:.2f}", ha='center', va='bottom')
plt.tight_layout()
plt.show()
```

Kmeans clustering

```
# Calculate WCSS for different numbers of clusters
wcss = []
for i in range(1, 9):
    kmeans = KMeans(n_clusters=i, n_init=10, random_state=42)
    kmeans.fit(scaled_data)
    wcss.append(kmeans.inertia_)
# Plot the WCSS against the number of clusters
plt.figure(figsize=(10, 6))
plt.plot(range(1, 9), wcss, marker='o', linestyle='--')
plt.title('Elbow Method for Optimal Number of Clusters')
plt.xlabel('Number of Clusters')
plt.ylabel('Total Within-Cluster Sum of Squares (WCSS)')
plt.show()
```

Choose 3 cluster since the change from adding 1 more clusters diminish. 3 would be the optimal number of clusters and paint the resulting groups on the principal components.

```
# K-means clustering vs PCA
# Initialize PCA, choosing to keep two principal components
pca = PCA(n_components=7)
pca_data = pca.fit_transform(scaled_data)
# Perform K-means clustering using the PCA results
kmeans = KMeans(n_clusters=3, n_init=10, random_state=42)
clusters = kmeans.fit_predict(pca_data)
```

```
# Add clustering information back to the original DataFrame
data['Cluster'] = clusters
# Check the distribution of countries in each cluster
print(data['Cluster'].value_counts())

# Plot the first two principal components
plt.figure(figsize=(8, 6))
sns.scatterplot(x=pca_data[:, 0], y=pca_data[:, 1], hue=data['Cluster'],
palette='viridis')
plt.title('PCA and k-mean Clustering of Dataset')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.legend(title='Cluster')
plt.show()
# Print the explained variance ratio
print("Explained Variance Ratio:", pca.explained_variance_ratio_)
# Plotting the results for GDP_per_Capita
plt.figure(figsize=(10, 6))
sns.barplot(x=cluster_summary.index, y=cluster_summary['GDP_per_Capita'])
plt.title('Average GDP per Capita by Cluster')
plt.xlabel('Cluster')
plt.ylabel('Average GDP per Capita')
plt.show()

# Calculate the mental health issues for each cluster
cluster_means = data.groupby('Cluster')[[ 'Schizophrenia_Disorders',
'Depressive_Disorders','Anxiety_Disorders', 'Bipolar_Disorders',
'Eating_Disorders']].mean()
# Convert mean values to probabilities within each cluster
cluster_probs = cluster_means.div(cluster_means.sum(axis=1), axis=0)
# Display the results
print("Probabilities of Mental Health Issues for Each Cluster:")
cluster_probs
# Calculate the average GDP per capita and unemployment rate for each cluster
cluster_means = data.groupby('Cluster')[['GDP_per_Capita']].mean()
cluster_means
# Calculate the mean values for each behavioral variable within each cluster
cluster_means = data.groupby('Cluster')[['Engaged_in_Religious_Activities',
'Improved_Healthy_Lifestyle', 'Changed_Work_Situation',
'Changed_Personal_Relationships', 'Talked_to_Friends_Family',
'Took_Medication', 'Spent_Time_in_Nature',
'Talked_to_Mental_Health_Professional']].mean()
# Convert mean values to probabilities (normalized to sum to 1 within each
cluster)
cluster_probs = cluster_means.div(cluster_means.sum(axis=1), axis=0)

# Transpose the DataFrame to make clusters as columns and variables as rows
cluster_probs_transposed = cluster_probs.T

# Display the results
```

```
print("Probabilities of Behavioral Variables for Each Cluster:")
cluster_probs_transposed

# Define thresholds for income levels
low_threshold = 11000
medium_threshold = 45000
# # Create a new column 'Income_Level' based on the thresholds
data.loc[data['GDP_per_Capita'] < low_threshold, 'Income_Level'] = 'Low'
data.loc[(data['GDP_per_Capita'] >= low_threshold) & (data['GDP_per_Capita']
< medium_threshold), 'Income_Level'] = 'Medium'
data.loc[data['GDP_per_Capita'] >= medium_threshold, 'Income_Level'] = 'High'
# Convert Income_Level to numerical values
income_mapping = {'Low': 1, 'Medium': 2, 'High': 3}
data['Income_Level_Num'] = data['Income_Level'].map(income_mapping)
# Create the confusion matrix
conf_matrix = confusion_matrix(data['Income_Level_Num'], kmeans_pca.labels_ +
1)
accuracy = accuracy_score(data['Income_Level_Num'], kmeans_pca.labels_ + 1)
# Plot the confusion matrix
plt.figure(figsize=(10, 7))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues',
            xticklabels=['Cluster 0', 'Cluster 1', 'Cluster 2'],
            yticklabels=['Low', 'Medium', 'High'])
plt.xlabel('PCA-Based Clusters')
plt.ylabel('Income Levels')
plt.title('Confusion Matrix: PCA-Based Clusters vs. Income Levels')
plt.show()
# Display the confusion matrix
print("Confusion Matrix:")
print(conf_matrix)
print(f"Accuracy: {accuracy:.2f}")
```

Hierarchical Clustering

```
# Define thresholds for income levels
low_threshold = 10000
medium_threshold = 45000
# # Create a new column 'Income_Level' based on the thresholds
data.loc[data['GDP_per_Capita'] < low_threshold, 'Income_Level'] = 'Low'
data.loc[(data['GDP_per_Capita'] >= low_threshold) & (data['GDP_per_Capita']
< medium_threshold), 'Income_Level'] = 'Medium'
data.loc[data['GDP_per_Capita'] >= medium_threshold, 'Income_Level'] = 'High'

# Separate the columns needed for clustering analysis
data_features = data.drop(columns=['Country', 'Income_Level',
'GDP_per_Capita'])
# Scale the feature data
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data_features)
# Convert scaled data back to DataFrame
scaled_data_df = pd.DataFrame(scaled_data, columns=data_features.columns)
# Initialize PCA
```

```
pca = PCA(n_components=7)
pca_data = pca.fit_transform(scaled_data)
# Perform Hierarchical Clustering
hc = AgglomerativeClustering(distance_threshold=0, n_clusters=None,
linkage='average')
hc.fit(pca_data)
# Extract linkage matrix using the linkage method
linkage_matrix = linkage(pca_data, method='average')
# Cutting the dendrogram at a specific distance to form three clusters
data['HC_Cluster'] = fcluster(linkage_matrix, t=4.8, criterion='distance')
# Create the plot to visualize the dendrogram
fig, ax = plt.subplots(figsize=(8, 8))
dendrogram(linkage_matrix, labels=data_clean['Country'].values,
color_threshold=4.8, ax=ax)
ax.axhline(y=4.8, color='black')  # Add a line to indicate the cut-off
plt.title("Hierarchical Clustering Dendrogram")
plt.xlabel("Country")
plt.ylabel("Distance")
plt.show()
# Convert Income_Level to numerical values
income_mapping = {'Low': 1, 'Medium': 2, 'High': 3}
data['Income_Level_Num'] = data['Income_Level'].map(income_mapping)
# Create the confusion matrix and calculate accuracy
conf_matrix = confusion_matrix(data['Income_Level_Num'], data['HC_Cluster'])
accuracy = accuracy_score(data['Income_Level_Num'], data['HC_Cluster'])
# Plot the confusion matrix
plt.figure(figsize=(10, 7))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues',
            xticklabels=['Cluster 1', 'Cluster 2', 'Cluster 3'],
            yticklabels=['Low', 'Medium', 'High'])
plt.xlabel('Hierarchical Clusters')
plt.ylabel('Income Levels')
plt.title('Confusion Matrix: Hierarchical Clusters vs. Income Levels')
plt.show()
# Display the confusion matrix and accuracy
print("Confusion Matrix:")
print(conf_matrix)
print(f"Accuracy: {accuracy:.2f}")
```