CLASSIFICATION OF DEATH BY RISK FACTORS

ABSTRACT

The aim of this study is to explore the presence of patterns in multiple health risk factors across different countries and years ranging from 1990 - 2019 using unsupervised learning. The analysis methods used include Principal Component Analysis (PCA) and clustering methods: K-means and Hierarchical Clustering. PCA was applied to the data to reduce dimensionality of the data to enable easier visualization and interpretation. The clustering methods helped to define groups with similar health risk characteristics that can help to design potential interventions for public health. These clusters provided valuable insights into the distribution of various health risks, such as high BMI, high LDL cholesterol, smoking, and alcohol use, among different countries which were important characteristics from the study. The analysis revealed significant relationships between different health risks. The study shows that the techniques of unsupervised learning can be very useful in providing decision makers with valuable insights from large and complicated datasets in health care and thus provide a strong basis for evidence-based decision making in health care systems.

INTRODUCTION

The analysis of health risk factors is crucial for understanding the underlying causes of mortality and morbidity across different regions and populations. However, with the increase in the volume of data and the tools available today, it is easier to understand these risk factors and create better strategies for improving the health of the population.

The main aim of this study is to apply these methods of unsupervised learning to examine and assess health risk factors. It is in this way that we recognize the relationships within the data. The results of this research can be applied to specific public health interventions that can make a significant difference in certain areas.

The dataset ‘Deaths by Risk Factors 2019’ is collected from the source ‘Our World in Data’. This dataset contains data from 1990 to 2019 related to various deaths related to a large number of health risks. The entire dataset contains 6840 rows and 31 columns. The columns contain Entity, Code, Year and the deaths that are from all causes attributed to high systolic blood pressure, diet high in sodium, diet low in whole grains, alcohol use, diet low in fruits, unsafe water source, secondhand smoke, low birth weight, child wasting, unsafe sex, diet low in nuts and seeds, household air pollution from solid fuels, diet low in vegetables, smoking, high fasting plasma glucose, air pollution, high body-mass index, unsafe sanitation, drug use, low bone mineral density, vitamin A deficiency, child stunting, non-exclusive breastfeeding, iron deficiency, ambient particulate matter pollution, low physical activity, no access to handwashing facility, and high ldl cholesterol. All these deaths data from all causes attributed are related in both sexes for all ages.[4]

THEORETICAL BACKGROUND

Unsupervised learning:

Unsupervised learning is a type of machine learning where algorithms are used to analyze and cluster unlabeled datasets, discovering hidden patterns or inbuilt structures without any predefined labels. The goal is to infer the natural structure present within a set of data points. Techniques include clustering like K-Means and Hierarchical clustering, dimensionality reduction like principal component analysis (PCA) and Singular value decomposition(SVD). This approach helps infer the natural structure within datasets, aiding in exploratory data analysis, anomaly detection, and feature extraction. making it essential in various industries for recognizing patterns and making data-driven decisions. [2]

Dimensionality reduction:

Dimensionality reduction is an unsupervised learning process, which is used for reducing the number of features or dimensions from a dataset which is generally large. This process can overcome some of the difficulties like computational issues, and problems related to data representation. Depending on the data, there are different methods of dimensionality reduction such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) that make the size of the data more manageable and improve the performance of machine learning algorithms and data visualization. These techniques are very useful for data preprocessing, feature extraction, and optimization of many computational methods in different fields.

Principal Component Analysis (PCA) is one of the approaches for dimensionality reduction in machine learning. It is a process that converts the features into a set of linearly uncorrelated features with orthogonal transformation. These new sets of transformed features are known as Principal Components. PCA works by looking at the variance of each attribute because the attribute with high value indicates good separation between the classes and therefore reduces the dimensionality. Some real world examples of PCA are image processing, recommendation systems etc.

Singular Value Decomposition (SVD) is another technique that is similar to PCA to reduce the dimensionality of the training data In the formation of a more compressed and significant representation of the data that can enhance the effectiveness and the speed of the algorithms. For instance, SVD can be used to perform dimensionality reduction on large text datasets by converting a matrix of terms frequency or count into a matrix of lower rank. SVD is denoted by the formula, A = USV, where U and V are orthogonal matrices, S is a diagonal matrix, and S values are considered singular values of matrix A. SVD can be used to reduce data by turning a large bundle of numbers into a smaller set of valuable fractions.[3]

Clustering:

Clustering is a technique used in unsupervised learning to group similar data points together. It involves partitioning a dataset into subsets, or clusters, where data points within each cluster share common characteristics. The goal is to discover meaningful patterns or structures within the data without the need for pre-existing labels or target variables. Clustering algorithms help identify natural groupings, which can be useful for tasks such as data exploration, pattern recognition, and anomaly detection. Popular clustering algorithms include k-means, hierarchical clustering.

K-means clustering is an algorithm in unsupervised machine learning used for clustering of data points and k-means unlike supervised learning does not need labeled data for training. It does not partition objects into sets where there are similarities between the objects in the set and dissimilar to the objects in other sets. It begins with the allocation of cluster centroids in the data space in a random manner. Next, every data point is placed in the cluster with the nearest centroid by using the distance degree. Hence, once all the data points are assigned to each of the clusters, the new centroids are then determined on the basis of the average of the points in the concerned cluster. This continues until the algorithm gets to converge and come up with the stable cluster. In k-means clustering for instance a value of k refers to the number of clusters that is normally predetermined. In other words, the algorithm intends to divide the given data into 'k' clusters with the minimum sum of square difference within clusters.

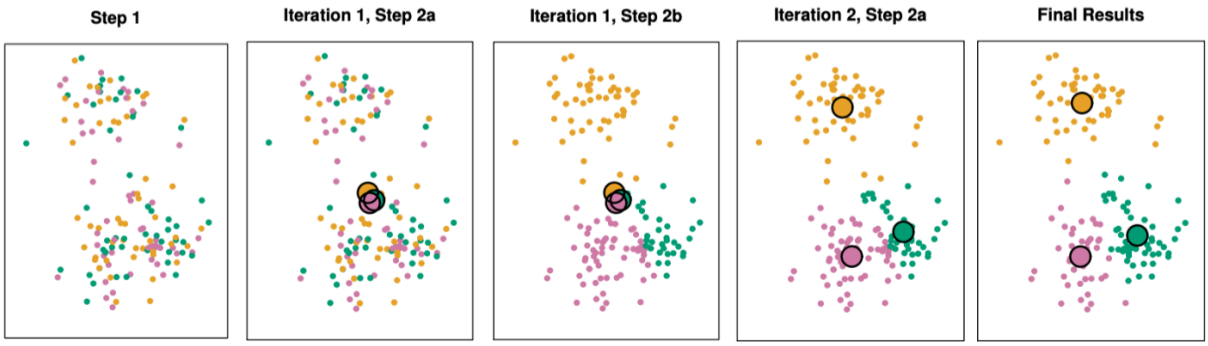


Fig : K-means clustering [1]

Hierarchical clustering is an approach in machine learning that clusters similar data points in a tree structure. Hierarchical clustering is different from other clustering methods in that it does not require the number of clusters to be specified in advance; it generates a tree of clusters that can be explored in a natural and easy way. This method starts with each data point being a cluster and then merges similar clusters to form larger clusters in each successive step. The process continues until all the data points are within a single cluster or until the required number of clusters have been formed.

Dendrogram is a graphical representation of the hierarchical clustering technique in the form of a tree diagram and each data point is a terminal node, or a leaf node, placed at the lowest level of the tree. Clusters are created by connecting similar data points or smaller clusters and the lengths of the vertical lines connecting the clusters indicate how similar the clusters are.The distance between two clusters or data points in the dendrogram indicates the level of similarity or dissimilarity between them. The longer the vertical line connecting them, the less related they are. On the other hand, lines with lesser numbers of characters suggest higher similarity. [2]

There are four different ways to measure similarity :

Complete Linkage : It calculates the similarity between two clusters and gives the maximum recorded similarity.
Single Linkage : It calculates the similarity between two clusters and gives the minimum recorded similarity.
Average Linkage : It calculates all the similarity between two clusters and gives the average recorded similarity.
Centroid Linkage : It calculates the similarity between the centroid of two cluster points.

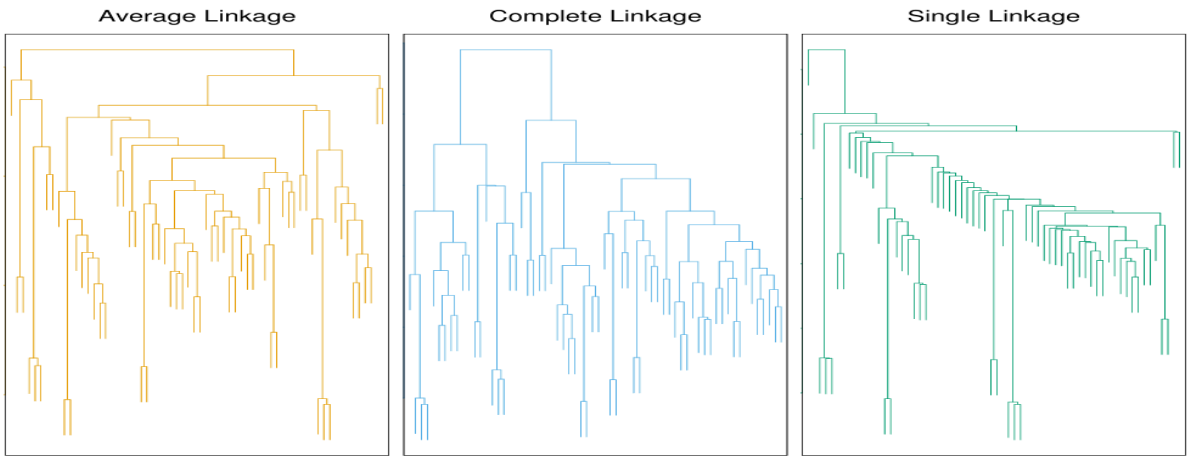


Fig : Hierarchical Clustering - Dendrogram [1]

METHODOLOGY

The dataset contains values of 6840 rows and 31 columns. Firstly, the dataset is loaded into the python notebook as dataframe using the pandas library. To facilitate easier data handling and analysis, the original column names were lengthy and are renamed to more concise and user-friendly names. Missing values were identified and handled to ensure data quality.

Exploratory Data Analysis (EDA)
The Exploratory Data Analysis (EDA) involves several steps to understand and summarize the main characteristics of the dataset. An initial overview of the dataset was obtained by summarizing the data to understand its basic structure, including summary statistics and data types. This helped in gaining a preliminary understanding of the data distribution and identifying any anomalies. **Univariate Analysis** involves describing the main features of each variable individually. This is done by histograms and density plots to understand about the various distributions of health risks. **Bivariate Analysis** involves the analysis between two variables. This analysis is done by scatter plot and correlation matrix. The purpose of correlation analysis is

to establish a pattern or linkage between two sets of variables. **Multivariate Analysis** is a technique that allows the interactions of three or more variables at the same time which can be done by pair plots and box plots to identify the correlation between multiple variables. **Time-series analysis[6]** is plotted to get the characteristics of data over time. The scatter plot for two variables from the dataset High Body Mass Index (BMI) and High LDL Cholesterol are plotted to observe the characteristics between two variables.

Principal Component Analysis (PCA)

After the completion of EDA, the unsupervised learning techniques are implemented for classification. Firstly, all the numeric values columns are subsetted. This subset data is then scaled using Standard Scaler for having an equal number of data. Then the PCA is performed on scaled data to reduce the dimensionality and retain variance. Now, the total number of principal components set resulted in 28. Now, the PCA transformed data which contains Principal Component scores (x) were converted into a DataFrame for better interpretation. Each row in this DataFrame represents an observation of the principal components. The loadings (rotations), which are the coefficients of the linear combination of the original variables forming each principal component, were also converted into a DataFrame.

Explained Variance Ratio is the attribute of the PCA object that is used for returning arrays of the variance ratios from the principal components. A scree plot is plotted to determine the number of principal components. A scree plot helps in identifying the points where the explained variance starts. The cumulative explained variance helps in understanding the proportion of total variance captured as more components are added.

K-means Clustering

Now, the K-means clustering is performed by setting the number of clusters as 5 and initialization steps as 10. This clustering is then applied on the first two principal components of the dataset. These cluster labels are then added to the PCA data. Then, the clusters for the first two principal components are plotted with centroid points on the graph with red color marks for the center point of each cluster. All the cluster labels are added to the dataframe for the observation of values.

Hierarchical Clustering (Dendrogram)

Hierarchical clustering is performed with a linkage function using scaled data, ‘complete’ linkage method and euclidean metric. Then, a dendrogram is plotted for the data which is transformed after using linkage. Now, the dendrogram plot appears to be overlapped with too many branches. So, to clearly understand the dendrogram plot is truncated to show only the top five levels of clustering. The cut_tree function is used to cut the dendrogram to get the cluster labels for a specific number of clusters. In this function, five number of clusters were defined to get the cluster labels.

After clustering, the distribution of key health risk factors High LDL Cholesterol, HighBMI, Smoking and Alcohol use for each cluster is plotted using boxplot.

COMPUTATIONAL RESULTS

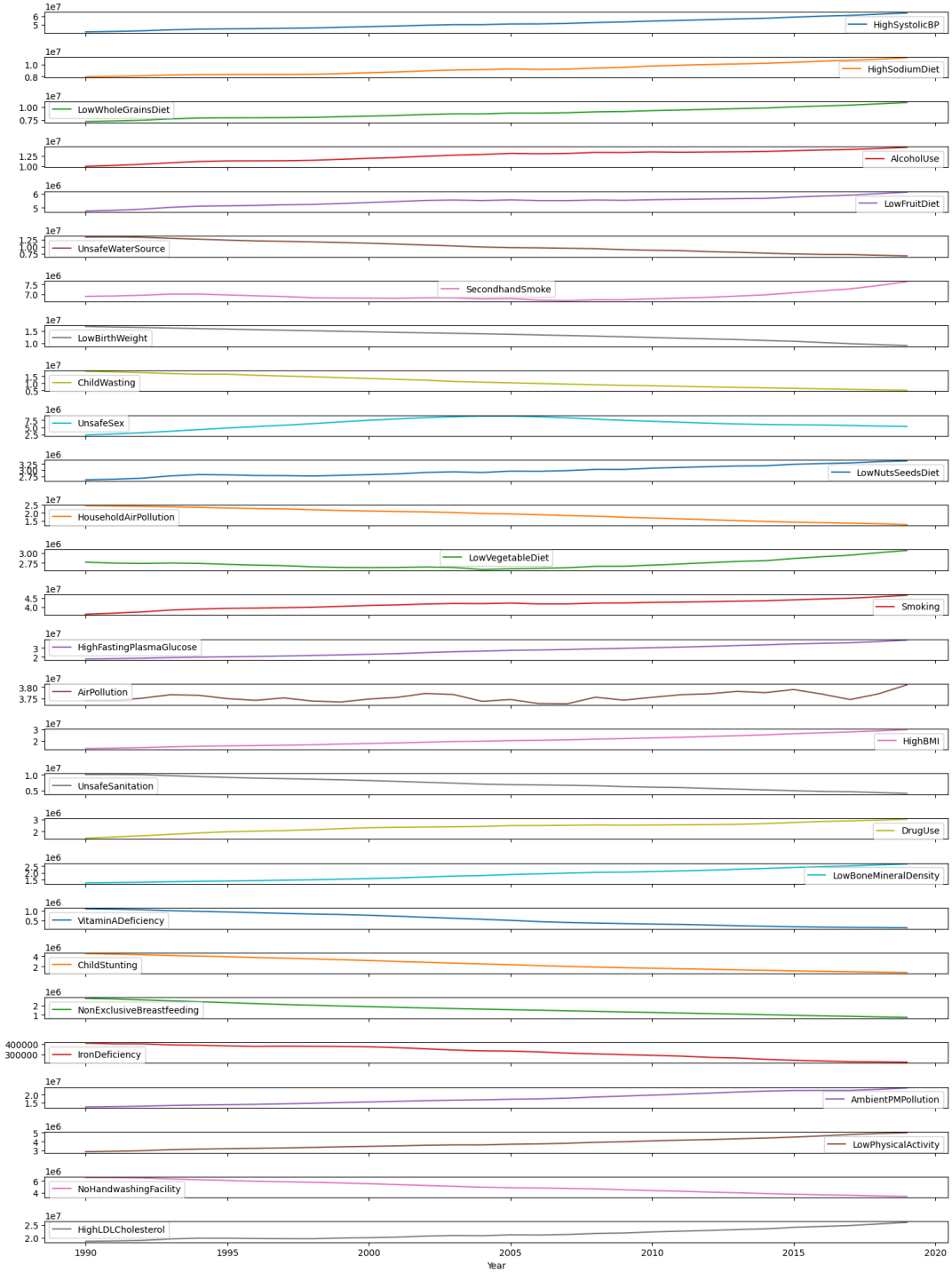


Fig : Time series Analysis

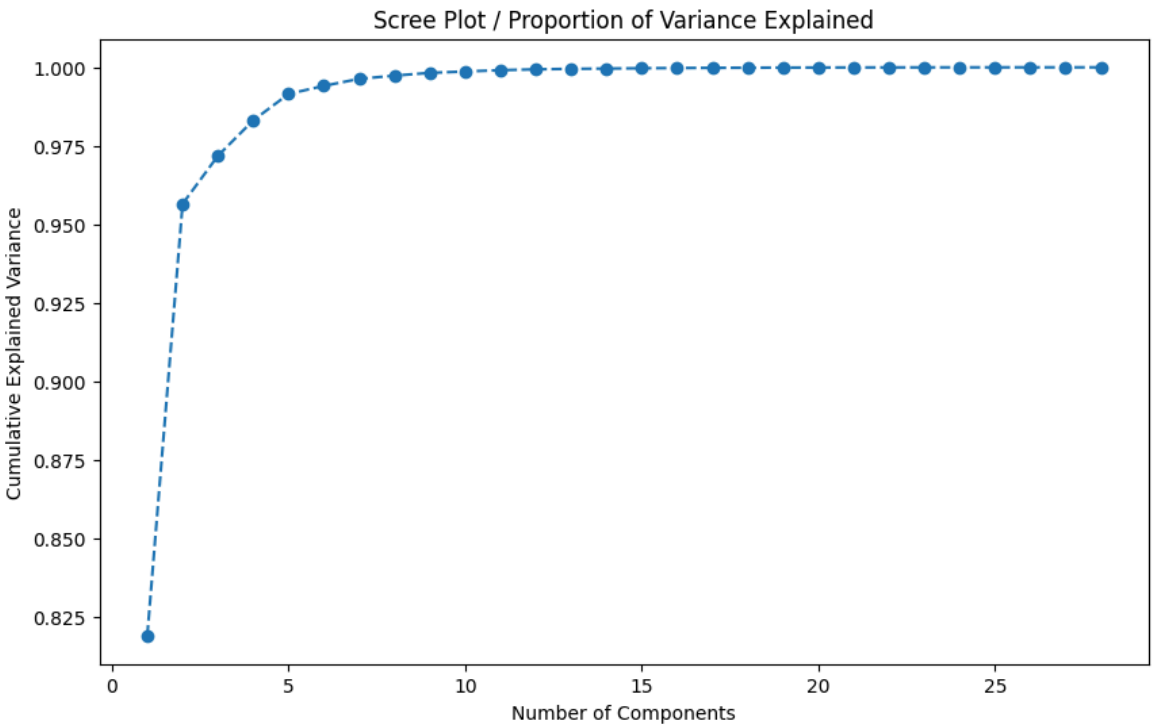


Fig : Scree plot

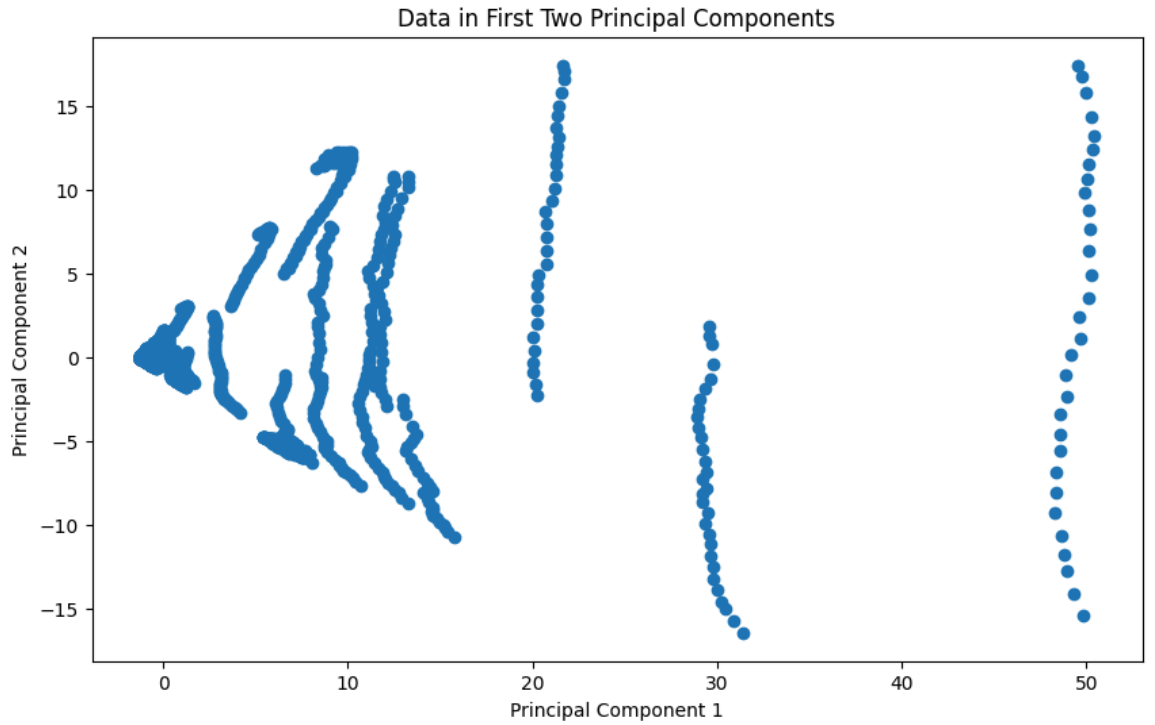


Fig : Plotting of first two principal components

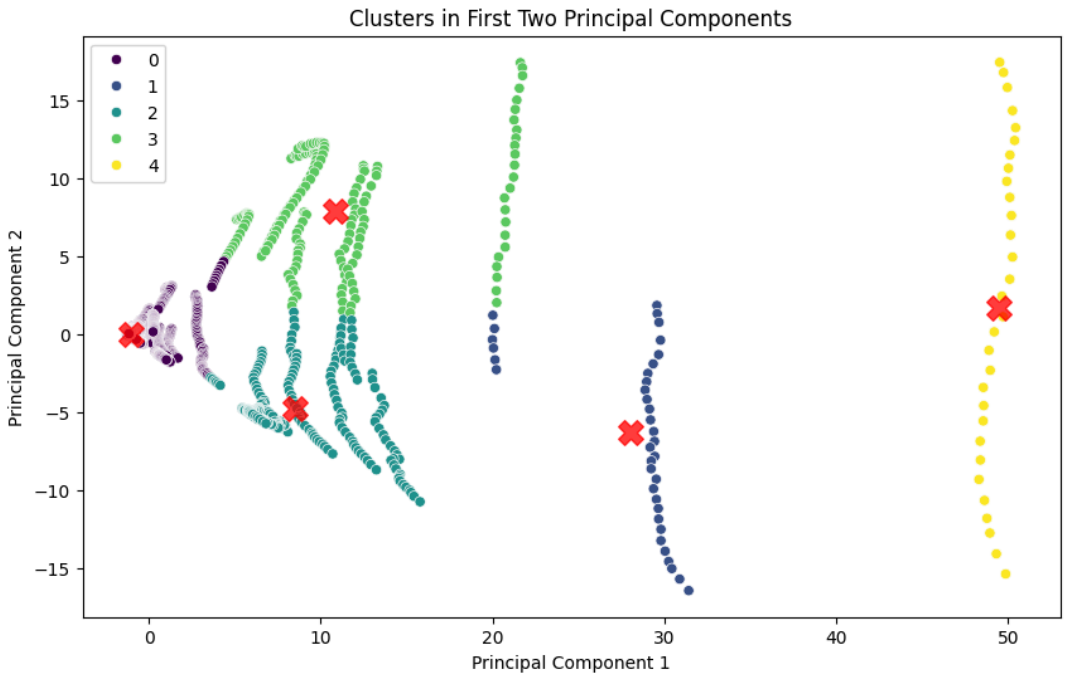


Fig : Plotting clusters of first two principal components

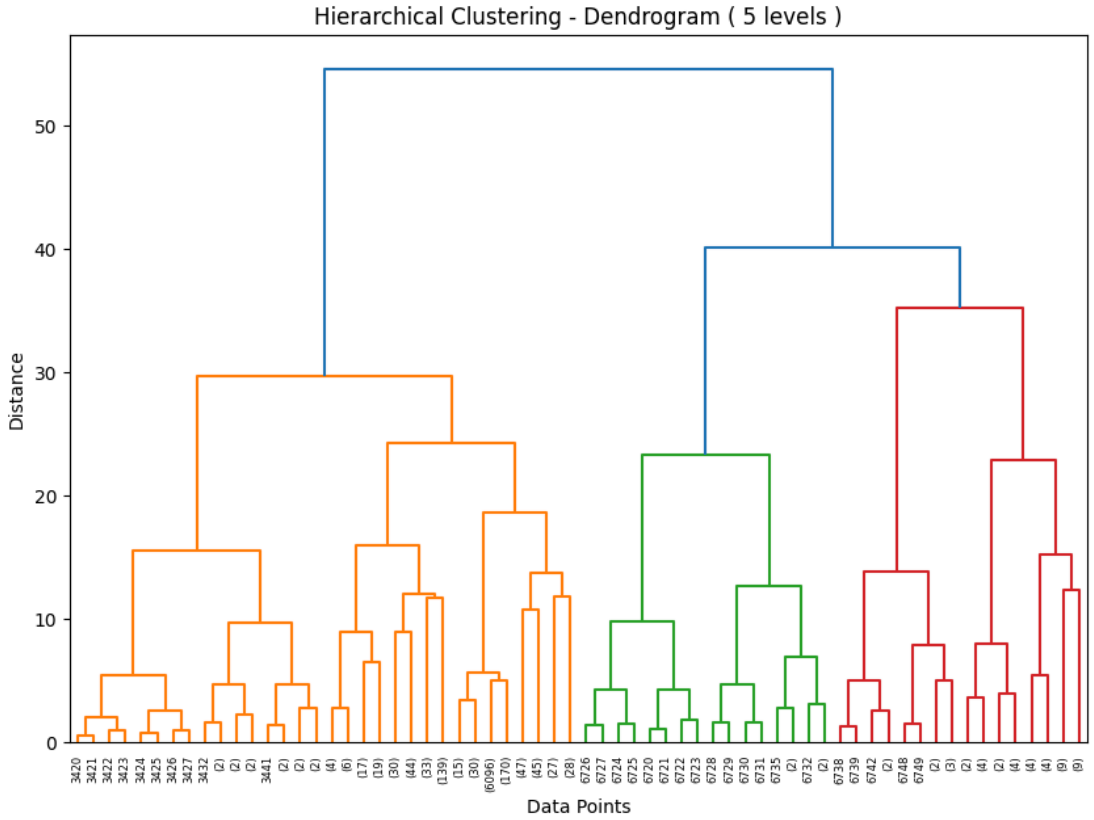


Fig : Hierarchical Clustering -Dendrogram (complete method with top 5 layers)

DISCUSSION

The principal component scores, represented by the matrix 'x', indicate the mapping of the original data onto the set of new principal components. Each row corresponds to a data point, and each column represents a principal component. The values in this matrix show how much each data point scores on each principal component. High absolute values indicates that the data point has a significant mapping on that particular principal component, capturing most of the variability in the data along that axis.

The rotation matrix indicates the weights of the original variables in each principal component. Each row corresponds to an original variable, and each column represents a principal component. These represent the contribution of each original variable to the principal components. Variables with high absolute values in a particular component are significant contributors to that component.

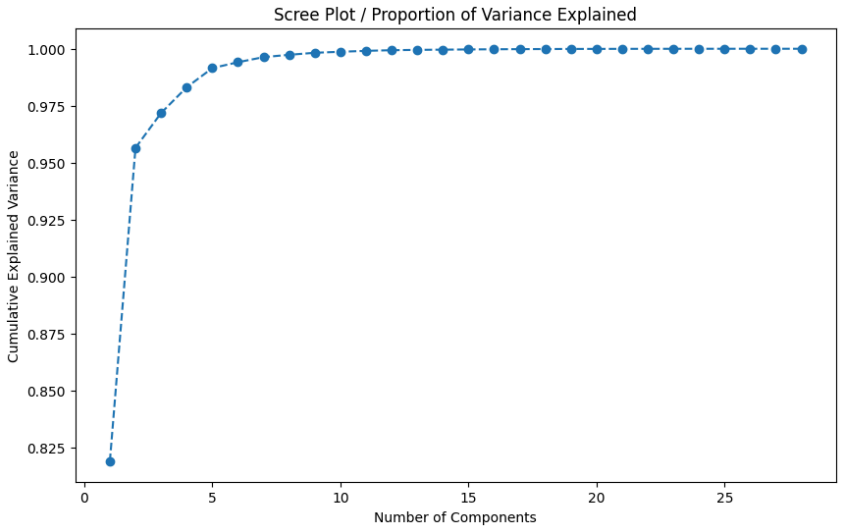


Fig : Scree Plot

The scree plot demonstrates the cumulative explained variance for the number of principal components in y and x-axis respectively. The plot shows that the cumulative explained variance increases rapidly with few components and they start to level up. This reaches the elbow point 3 to 5 principal component which means the first few components have the most of the variance in the data, so beyond which adding more components contribution is less significant to cumulative explained variance. First variants capture 82% of the variance, over which cumulative explained variance reaches approximately 95%.

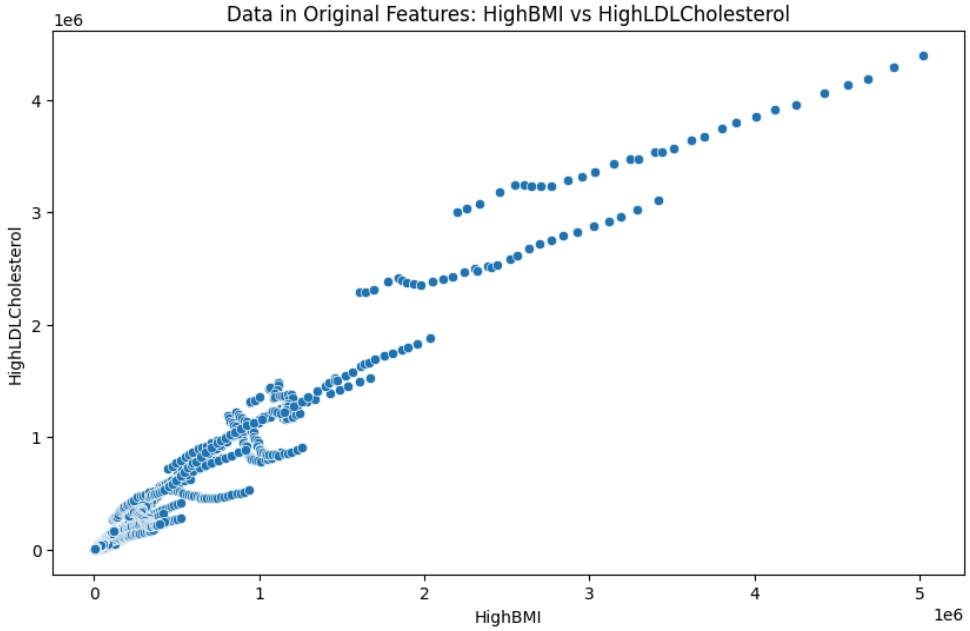


Fig : Plotting of two original features HighBMI and HighLDLCholestrol

The graph shows the relationship between HighBMI and HighLDLCholesterol in the original feature. It is strongly correlated between each other variables, BMI is directly proportional to LDL cholesterol also in an increasing trend. They have a linear relationship suggesting these two variables are very related for the health risk factor.

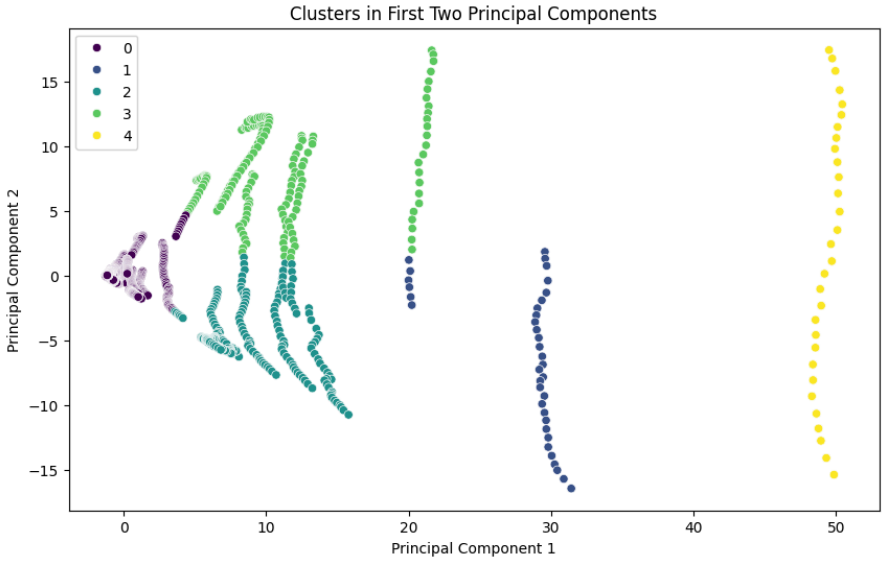


Fig : Plotting clusters of first two principal components

The graph illustrates the first two principal components (PC1 and PC2) after performing PCA. The color difference shows the clustering by K-means. The PCA separates the principal

components effectively, capturing the underlying structure and variability in the data. This is evident that distinct groupings and data spread points along the principal component axes. visualize complex relationships and structures that are not easily visible in the graph.

Comparison of Plots
Original Features Plot:
Linear relationship is shown between "High Body Mass Index (BMI)" and "HighLDLCholesterol"
Limited to the specific variable

Principal Components Plot:
Reveals inherent structure and clusters
Captures the most significant sources of variance

Interpretation:
Cluster 0:
Max: HighSystolicBP
Min: IronDeficiency

Cluster 1:
Max: HighSystolicBP
Min: IronDeficiency

Cluster 2:
Max: HighSystolicBP
Min: VitaminADeficiency

Cluster 3:
Max: AirPollution
Min: IronDeficiency

Cluster 4:
Max: HighSystolicBP
Min: VitaminADeficiency

Based on the above observation the first two components are not apparent in the original features. PCA retains the essential structure and variance in the data, reducing the dimensions.
The distinct health risk profiles, which can be targeted interventions and public health strategies by this. The max and min columns for each cluster highlight health concerns in each group, providing actionable insights for these issues.

Analysis of Clusters Based on Key Health Risk Factors

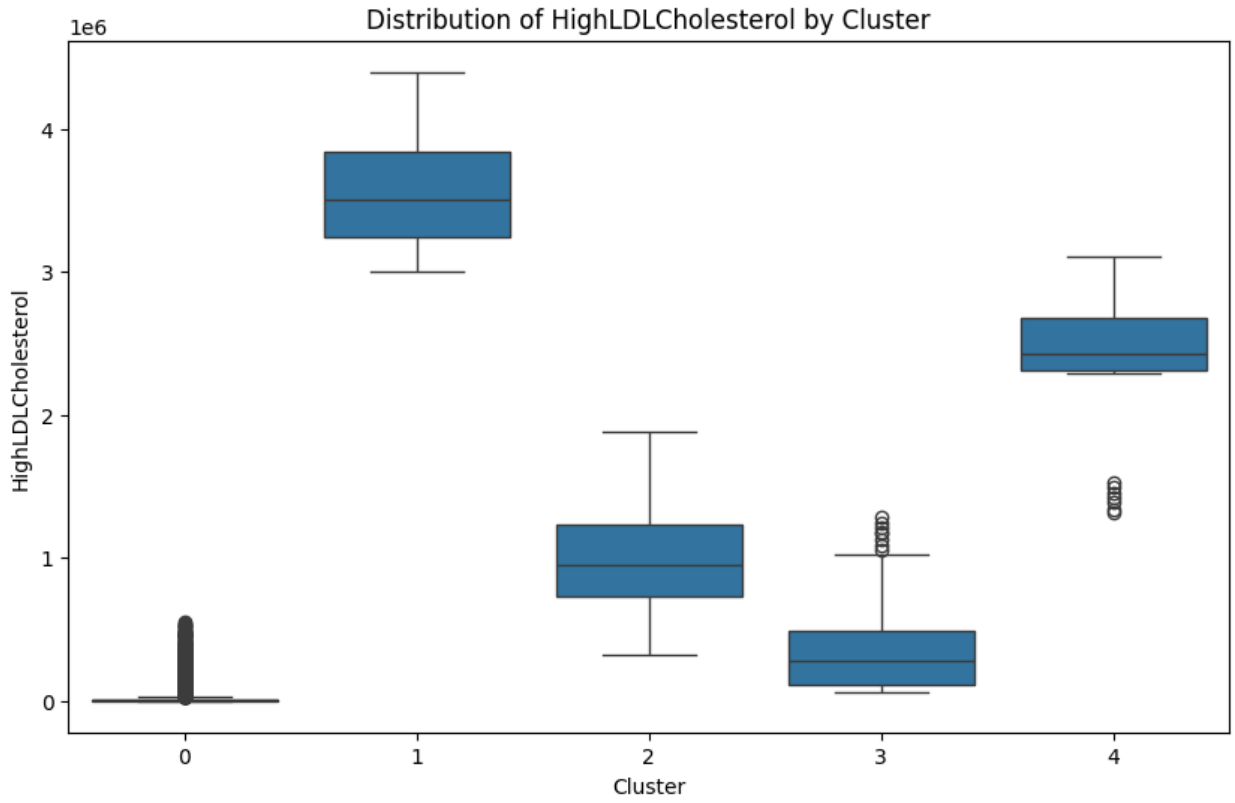


Fig : Distribution of HighLDLCholestrol by cluster

Cluster Analysis

HighLDLCholesterol Distribution by Cluster

Cluster 0: low prevalence of high LDL cholesterol levels due to lowest values for high LDL cholesterol.

Cluster 1: The highest scores for HighLDLCholesterol indicate that high LDL cholesterol has a significant influence on this cluster.

Cluster 2: intermediate prevalence and moderate values.

Cluster 3: more variable, lower values that resemble Cluster 0 more.

While still showing high values, Cluster 4 is not as dramatic as Cluster 1.

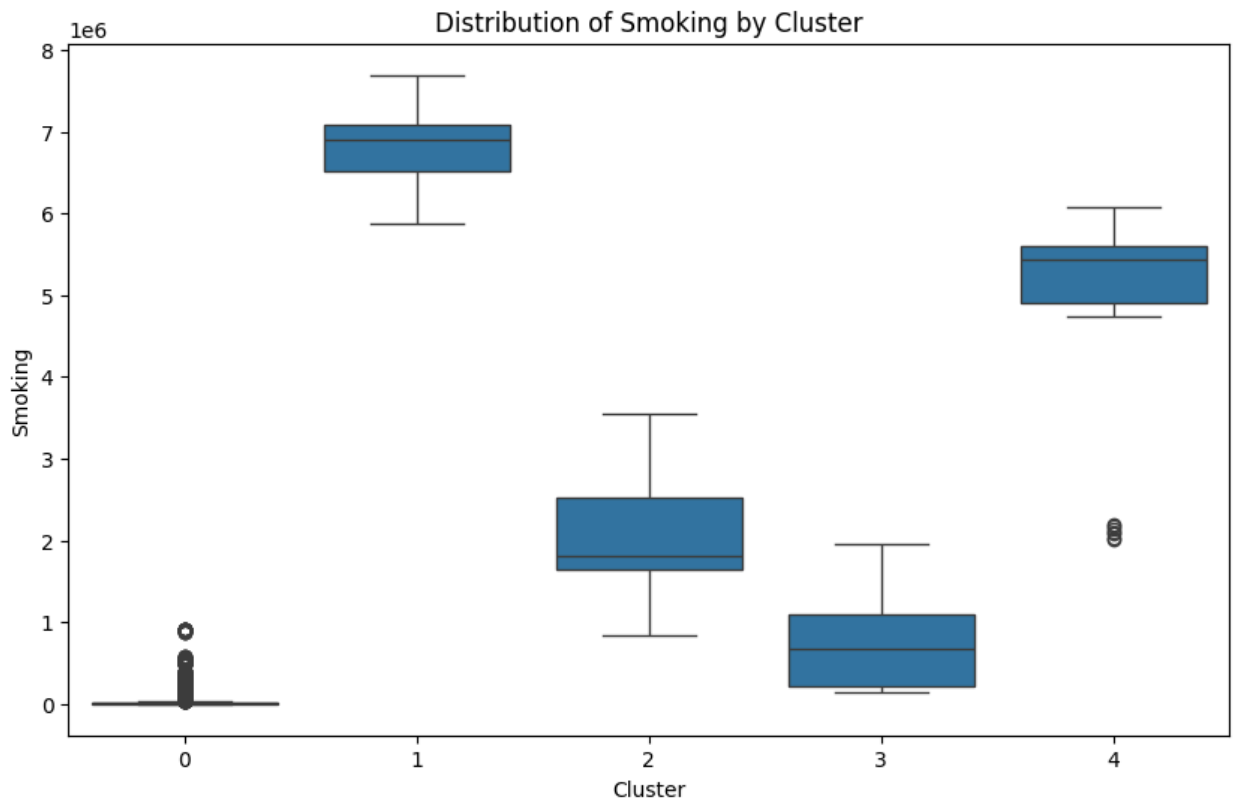


Fig : Distribution of Smoking by cluster

Smoking Distribution by Cluster
Cluster 0: Distinguished by the lowest rates of Smoking use.
With the highest smoking rates, Cluster 1 is thought to have serious smoking-related problems.
Cluster 2: In between Clusters 0 and 1, moderate smoking rates are found.
Cluster 3: Low smoking rates, with a few outliers, comparable to Cluster 0.
Smoking rates in Cluster 4 are comparatively high, however lower than in Cluster 1.

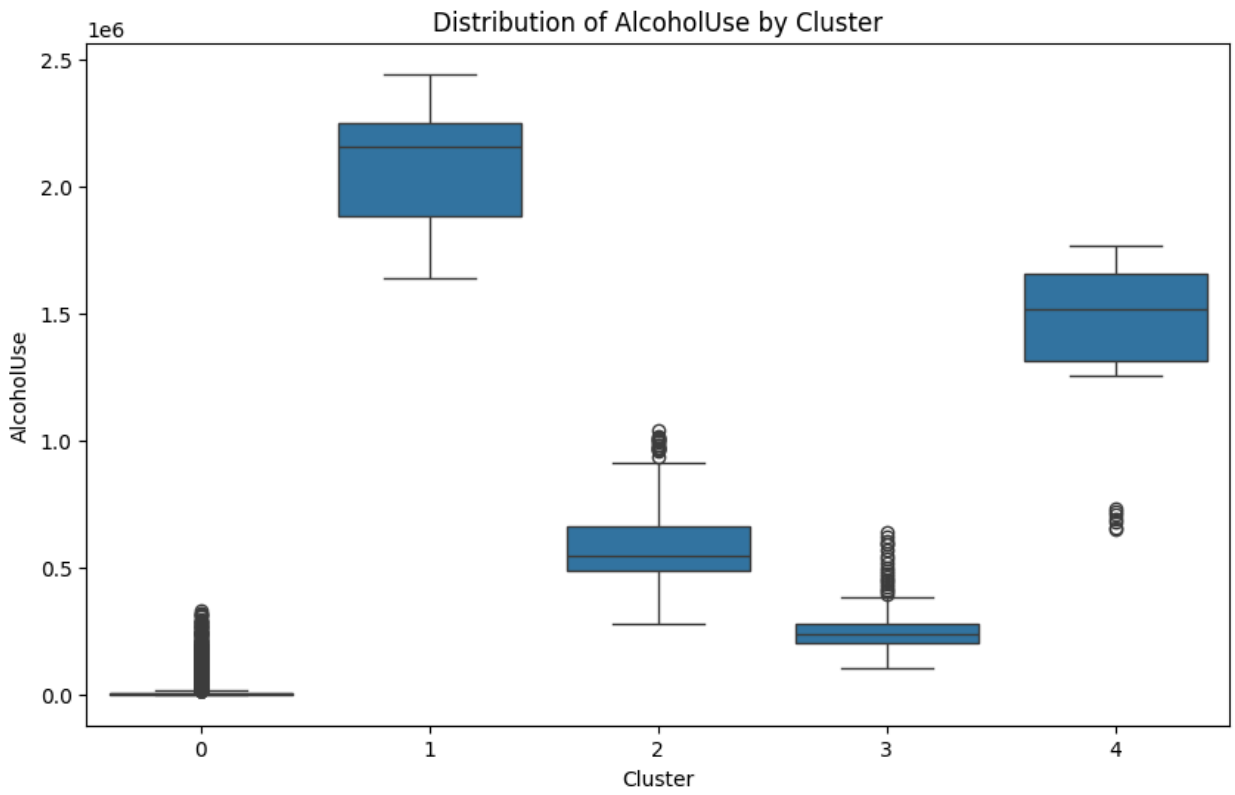


Fig : Distribution of Alcohol Use by cluster

AlcoholUse Distribution by Cluster
Cluster 0: Shows the least amount of alcohol intake.
Cluster 1: Displays the greatest level of alcohol intake, suggesting serious problems with alcohol use.
Cluster 2: Moderate alcohol consumption.
Cluster 3: Less alcohol consumption, somewhat different from Cluster 0.
While less than in Cluster 1, Cluster 4 has higher alcohol use.

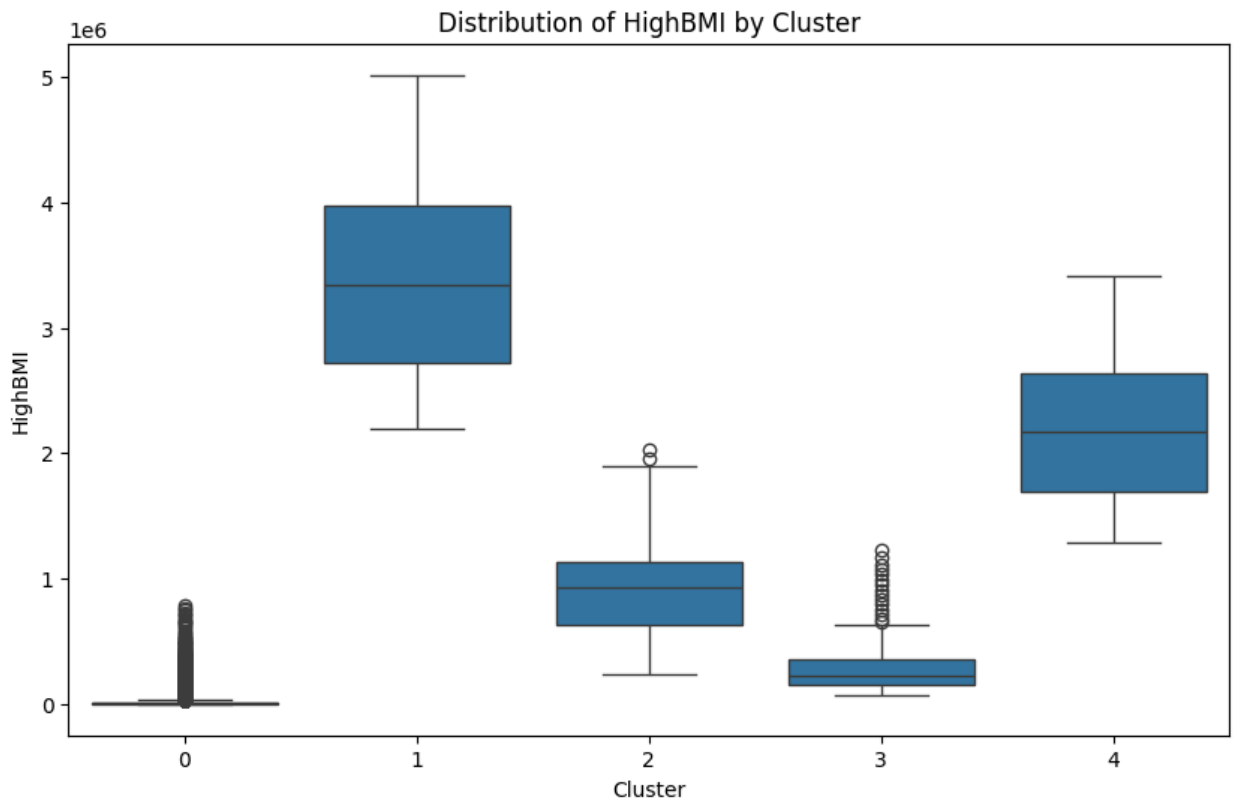


Fig : Distribution of HighBMI by cluster

HighBMI Distribution by Cluster
The cluster with the lowest BMI values is known as cluster 0.
Cluster 1: Shows the highest BMI values, suggesting serious problems related to obesity.
Cluster 2: In between Clusters 0 and 1, with moderate BMI levels.
Cluster 3: Low BMI readings; more outliers than Cluster 0 but still comparable.
BMI readings in Cluster 4 are high, although lower than in Cluster 1.

Interpretation of Clusters

Cluster 0:
Features: Typically exhibits the lowest values among the major risk factors for health. This cluster may be indicative of areas or demographics with fewer hazards to general health.
Implications for Public Health: Maintaining low risk levels may be achieved by the use of preventive measures.

Cluster 1:
Features: Has the greatest levels of alcohol consumption, smoking, high LDL cholesterol, and high BMI. Regions or people with serious health problems are represented by this cluster.

Implications for Public Health: In order to address several risk factors, intensive health interventions are needed.

Cluster 2:
Features: Shows intermediate levels of health hazards with modest values for the major health risk variables.
Implications for Public Health: Specific health hazards may be reduced by targeted actions.

Cluster 3:
Features: Usually has lower values, however there are occasional outliers and variance. This cluster might be a representation of areas with varying health characteristics.
Implications for Public Health: To address certain outliers and differences in health hazards, targeted solutions may be required.

Cluster 4:
Features: Displays elevated health risk factor values, however not as sharply as Cluster 1. To a lesser degree, this cluster suggests substantial health hazards.
Implications for Public Health: Health initiatives aimed at reducing smoking, high LDL cholesterol, and BMI may be successful.

Through an analysis of the primary health risk factor distributions within each cluster, important insights into the unique health profiles of various people or areas may be obtained. This study aids in identifying regions with serious health problems that need immediate attention and others that would be better served by preventative measures. The creation of customized public health policies to successfully address the identified health hazards is guided by the unique characteristics of each cluster.

CONCLUSION

This study successfully explored the patterns in multiple health risk factors across different countries from 1990 to 2019 using unsupervised learning techniques. By applying Principal Component Analysis (PCA), effectively reduced the dimensionality of the dataset, making it easier for visualization and interpretation of the data. The PCA results indicated that the first few principal components captured most of the variance, highlighting the essential structure within the data. The clustering methods, including K-means and Hierarchical Clustering, were instrumental in identifying distinct groups of health risk characteristics. These clusters provided valuable insights into the distribution of various health risks, such as high BMI, high LDL cholesterol, smoking, and alcohol use, among different countries which were important characteristics. The analysis revealed significant relationships between different health risks, emphasizing the need for targeted public health. Clusters with severe health issues, such as high

smoking rates and high BMI, require intensive health measures, while clusters with lower overall risk levels may benefit from preventive measures.

Overall, this study demonstrates the utility of unsupervised learning techniques in extracting meaningful insights from large and complex healthcare datasets. The findings can inform evidence based decision making and help design effective public health strategies to address various health risks. By identifying specific health profiles and risk factors, policymakers and healthcare providers to improve health outcomes across different countries.

REFERENCES:

[1] Introduction to Statistical Learning with Python - Chapter 12
https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html

[2] Unsupervised Learning by geeksforgeek .
https://www.geeksforgeeks.org/ml-types-learning-part-2/

[3] What is Unsupervised Learning
https://www.ibm.com/topics/unsupervised-learning

[4] Dataset "Deaths by Risk Factor 2019"
https://ourworldindata.org/grapher/number-of-deaths-by-risk-factor?tab=table&time=latest

[5]SciPy – Cluster Hierarchy Dendrogram
scipy.cluster.hierarchy.linkage — SciPy v1.13.1 Manual

[6] Time series Analysis Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control. Wiley
https://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021

APPENDIX

In [ ]:
```
from google.colab import drive
drive.mount("/content/drive")
```

```
Mounted at /content/drive
```

In [ ]:
```
import pandas as pd
df = pd.read_csv('/content/drive/MyDrive/ML/number-of-deaths-by-risk-factor.csv')
```

In [ ]:
```
df
```

Out[ ]:

	Entity	Code	Year	Deaths that are from all causes attributed to high systolic blood pressure, in both sexes aged all ages	Deaths that are from all causes attributed to diet high in sodium, in both sexes aged all ages	Deaths that are from all causes attributed to diet low in whole grains, in both sexes aged all ages	Deaths that are from all causes attributed to alcohol use, in both sexes aged all ages	Deaths that are from all causes attributed to diet low in fruits, in both sexes aged all ages	Deaths that are from all causes attributed to unsafe water source, in both sexes aged all ages
0	Afghanistan	AFG	1990	25633.129	1044.9089	7077.3160	356.21470	3184.9550	3701.9940
1	Afghanistan	AFG	1991	25871.803	1054.9584	7149.0854	363.73020	3248.3767	4309.2820
2	Afghanistan	AFG	1992	26308.795	1074.6057	7297.3086	375.90024	3350.9207	5356.4980
3	Afghanistan	AFG	1993	26961.360	1103.3705	7498.5340	388.57156	3479.8118	7151.5210
4	Afghanistan	AFG	1994	27658.424	1133.8824	7697.5890	398.72700	3609.8315	7191.6390
...	...	...	...	...	...	...	...	...	...
6835	Zimbabwe	ZWE	2015	11483.307	1062.7714	1353.5919	4853.71830	1820.3645	4335.9062
6836	Zimbabwe	ZWE	2016	11662.818	1081.5592	1382.9396	4915.49170	1854.2546	4244.4795
6837	Zimbabwe	ZWE	2017	11818.670	1097.7340	1408.9166	4991.75340	1882.9751	4192.6160
6838	Zimbabwe	ZWE	2018	12002.042	1116.5680	1439.0463	5044.16060	1917.0316	4012.9230
6839	Zimbabwe	ZWE	2019	12240.987	1140.3761	1475.0681	5155.78900	1959.8661	3914.0618

6840 rows × 31 columns

EDA

Data Overview

In [ ]:
```
print(df.dtypes)
```

```
Entity
object
Code
object
Year
int64
Deaths that are from all causes attributed to high systolic blood pressure, in both s
exes aged all ages                float64
Deaths that are from all causes attributed to diet high in sodium, in both sexes aged
all ages                          float64
Deaths that are from all causes attributed to diet low in whole grains, in both sexes
aged all ages                     float64
Deaths that are from all causes attributed to alcohol use, in both sexes aged all age
s                                 float64
Deaths that are from all causes attributed to diet low in fruits, in both sexes aged
all ages                          float64
Deaths that are from all causes attributed to unsafe water source, in both sexes aged
all ages                          float64
Deaths that are from all causes attributed to secondhand smoke, in both sexes aged al
l ages                            float64
Deaths that are from all causes attributed to low birth weight, in both sexes aged al
l ages                            float64
Deaths that are from all causes attributed to child wasting, in both sexes aged all a
ges                               float64
Deaths that are from all causes attributed to unsafe sex, in both sexes aged all ages
float64
Deaths that are from all causes attributed to diet low in nuts and seeds, in both sex
es aged all ages                  float64
Deaths that are from all causes attributed to household air pollution from solid fuel
s, in both sexes aged all ages    float64
Deaths that are from all causes attributed to diet low in vegetables, in both sexes a
ged all ages                      float64
Deaths that are from all causes attributed to smoking, in both sexes aged all ages
float64
Deaths that are from all causes attributed to high fasting plasma glucose, in both se
xes aged all ages                 float64
Deaths that are from all causes attributed to air pollution, in both sexes aged all a
ges                               float64
Deaths that are from all causes attributed to high body-mass index, in both sexes age
d all ages                        float64
Deaths that are from all causes attributed to unsafe sanitation, in both sexes aged a
ll ages                           float64
Deaths that are from all causes attributed to drug use, in both sexes aged all ages
float64
Deaths that are from all causes attributed to low bone mineral density, in both sexes
aged all ages                     float64
Deaths that are from all causes attributed to vitamin a deficiency, in both sexes age
d all ages                        float64
Deaths that are from all causes attributed to child stunting, in both sexes aged all
ages                              float64
Deaths that are from all causes attributed to non-exclusive breastfeeding, in both se
xes aged all ages                 float64
Deaths that are from all causes attributed to iron deficiency, in both sexes aged all
ages                              float64
Deaths that are from all causes attributed to ambient particulate matter pollution, i
n both sexes aged all ages        float64
Deaths that are from all causes attributed to low physical activity, in both sexes ag
ed all ages                       float64
Deaths that are from all causes attributed to no access to handwashing facility, in b
oth sexes aged all ages           float64
```

```
Deaths that are from all causes attributed to high ldl cholesterol, in both sexes age
d all ages                          float64
dtype: object
```

In [ ]:
```python
# Renaming columns
df.rename(columns={
    "Entity": "Entity",
    "Code": "Code",
    "Year": "Year",
    "Deaths that are from all causes attributed to high systolic blood pressure, in bo
    "Deaths that are from all causes attributed to diet high in sodium, in both sexes
    "Deaths that are from all causes attributed to diet low in whole grains, in both s
    "Deaths that are from all causes attributed to alcohol use, in both sexes aged all
    "Deaths that are from all causes attributed to diet low in fruits, in both sexes a
    "Deaths that are from all causes attributed to unsafe water source, in both sexes
    "Deaths that are from all causes attributed to secondhand smoke, in both sexes age
    "Deaths that are from all causes attributed to low birth weight, in both sexes age
    "Deaths that are from all causes attributed to child wasting, in both sexes aged a
    "Deaths that are from all causes attributed to unsafe sex, in both sexes aged all
    "Deaths that are from all causes attributed to diet low in nuts and seeds, in both
    "Deaths that are from all causes attributed to household air pollution from solid
    "Deaths that are from all causes attributed to diet low in vegetables, in both sex
    "Deaths that are from all causes attributed to smoking, in both sexes aged all age
    "Deaths that are from all causes attributed to high fasting plasma glucose, in bot
    "Deaths that are from all causes attributed to air pollution, in both sexes aged a
    "Deaths that are from all causes attributed to high body-mass index, in both sexes
    "Deaths that are from all causes attributed to unsafe sanitation, in both sexes ag
    "Deaths that are from all causes attributed to drug use, in both sexes aged all ag
    "Deaths that are from all causes attributed to low bone mineral density, in both s
    "Deaths that are from all causes attributed to vitamin a deficiency, in both sexes
    "Deaths that are from all causes attributed to child stunting, in both sexes aged
    "Deaths that are from all causes attributed to non-exclusive breastfeeding, in bot
    "Deaths that are from all causes attributed to iron deficiency, in both sexes aged
    "Deaths that are from all causes attributed to ambient particulate matter pollutio
    "Deaths that are from all causes attributed to low physical activity, in both sexe
    "Deaths that are from all causes attributed to no access to handwashing facility,
    "Deaths that are from all causes attributed to high ldl cholesterol, in both sexes
}, inplace=True)
```

In [ ]:
```python
df.columns
```

Out[ ]:
```
Index(['Entity', 'Code', 'Year', 'HighSystolicBP', 'HighSodiumDiet',
       'LowWholeGrainsDiet', 'AlcoholUse', 'LowFruitDiet', 'UnsafeWaterSource',
       'SecondhandSmoke', 'LowBirthWeight', 'ChildWasting', 'UnsafeSex',
       'LowNutsSeedsDiet', 'HouseholdAirPollution', 'LowVegetableDiet',
       'Smoking', 'HighFastingPlasmaGlucose', 'AirPollution', 'HighBMI',
       'UnsafeSanitation', 'DrugUse', 'LowBoneMineralDensity',
       'VitaminADeficiency', 'ChildStunting', 'NonExclusiveBreastfeeding',
       'IronDeficiency', 'AmbientPMPollution', 'LowPhysicalActivity',
       'NoHandwashingFacility', 'HighLDLCholesterol'],
      dtype='object')
```

In [ ]:
```python
print(df.describe())
```

	Year	HighSystolicBP	HighSodiumDiet	LowWholeGrainsDiet	\
count	6840.000000	6.840000e+03	6.840000e+03	6.840000e+03	
mean	2004.500000	2.242249e+05	4.049716e+04	3.869129e+04	
std	8.656074	8.634691e+05	1.752832e+05	1.479084e+05	
min	1990.000000	2.466349e+00	4.098751e-01	4.988057e-01	
25%	1997.000000	1.827476e+03	1.365965e+02	2.733269e+02	
50%	2004.500000	8.770671e+03	9.694122e+02	1.444022e+03	
75%	2012.000000	4.035507e+04	5.169495e+03	6.773283e+03	
max	2019.000000	1.084560e+07	1.885356e+06	1.844836e+06	

	AlcoholUse	LowFruitDiet	UnsafeWaterSource	SecondhandSmoke	\
count	6.840000e+03	6.840000e+03	6.840000e+03	6.840000e+03	
mean	5.484861e+04	2.395776e+04	4.408639e+04	3.036401e+04	
std	2.112090e+05	9.451573e+04	2.020493e+05	1.222861e+05	
min	1.457786e-01	3.029068e-01	2.800072e-03	5.062894e-01	
25%	2.638683e+02	1.441008e+02	6.968999e+00	2.088802e+02	
50%	1.780532e+03	8.346746e+02	1.824999e+02	9.937228e+02	
75%	8.367695e+03	3.104619e+03	5.599117e+03	4.347840e+03	
max	2.441974e+06	1.046015e+06	2.450944e+06	1.304318e+06	

	LowBirthWeight	ChildWasting	...	DrugUse	\
count	6.840000e+03	6.840000e+03	...	6840.000000	
mean	5.912551e+04	4.992437e+04	...	10285.200840	
std	2.502265e+05	2.226529e+05	...	39960.741366	
min	6.554637e-02	6.858725e-02	...	0.054797	
25%	1.230088e+02	2.628168e+01	...	30.843025	
50%	1.056921e+03	5.038514e+02	...	221.868180	
75%	1.090304e+04	9.765207e+03	...	1224.431775	
max	3.033426e+06	3.430422e+06	...	494491.700000	

	LowBoneMineralDensity	VitaminADeficiency	ChildStunting	\
count	6840.000000	6840.000000	6840.000000	
mean	8182.476755	2471.615651	11164.355291	
std	32403.920487	12718.301605	52866.235540	
min	0.050526	0.000003	0.000391	
25%	42.931661	0.021711	0.873684	
50%	277.144125	1.847342	41.521570	
75%	1231.606000	230.444385	1563.387950	
max	437884.400000	207555.220000	833448.560000	

	NonExclusiveBreastfeeding	IronDeficiency	AmbientPMPollution	\
count	6840.000000	6840.000000	6.840000e+03	
mean	7171.866029	1421.394080	7.697212e+04	
std	31678.443811	6303.932513	3.183152e+05	
min	0.001889	0.000425	1.760401e-01	
25%	2.678312	0.757304	4.178013e+02	
50%	60.560858	11.722806	2.036275e+03	
75%	1315.425925	238.013368	1.127368e+04	
max	505469.660000	73461.060000	4.140971e+06	

	LowPhysicalActivity	NoHandwashingFacility	HighLDLCholesterol
count	6840.000000	6.840000e+03	6.840000e+03
mean	16489.085017	2.179990e+04	9.403521e+04
std	62708.007186	9.668258e+04	3.614803e+05
min	0.283192	1.961850e-02	1.069846e+00
25%	91.715955	1.893026e+01	5.582757e+02
50%	521.923925	2.213597e+02	3.122144e+03
75%	2820.269900	3.953427e+03	1.748821e+04
max	831502.000000	1.200349e+06	4.396984e+06

```
[8 rows x 29 columns]
```

Missing Values

In [ ]:
```
print(df.isnull().sum())
```

```
Entity                          0
Code                          690
Year                            0
HighSystolicBP                  0
HighSodiumDiet                  0
LowWholeGrainsDiet              0
AlcoholUse                      0
LowFruitDiet                    0
UnsafeWaterSource               0
SecondhandSmoke                 0
LowBirthWeight                  0
ChildWasting                    0
UnsafeSex                       0
LowNutsSeedsDiet                0
HouseholdAirPollution           0
LowVegetableDiet                0
Smoking                         0
HighFastingPlasmaGlucose        0
AirPollution                    0
HighBMI                         0
UnsafeSanitation                0
DrugUse                         0
LowBoneMineralDensity           0
VitaminADeficiency              0
ChildStunting                   0
NonExclusiveBreastfeeding       0
IronDeficiency                  0
AmbientPMPollution              0
LowPhysicalActivity             0
NoHandwashingFacility           0
HighLDLCholesterol              0
dtype: int64
```

In [ ]:
```
print(df.info())
print(df.describe())
print(df.head())
```

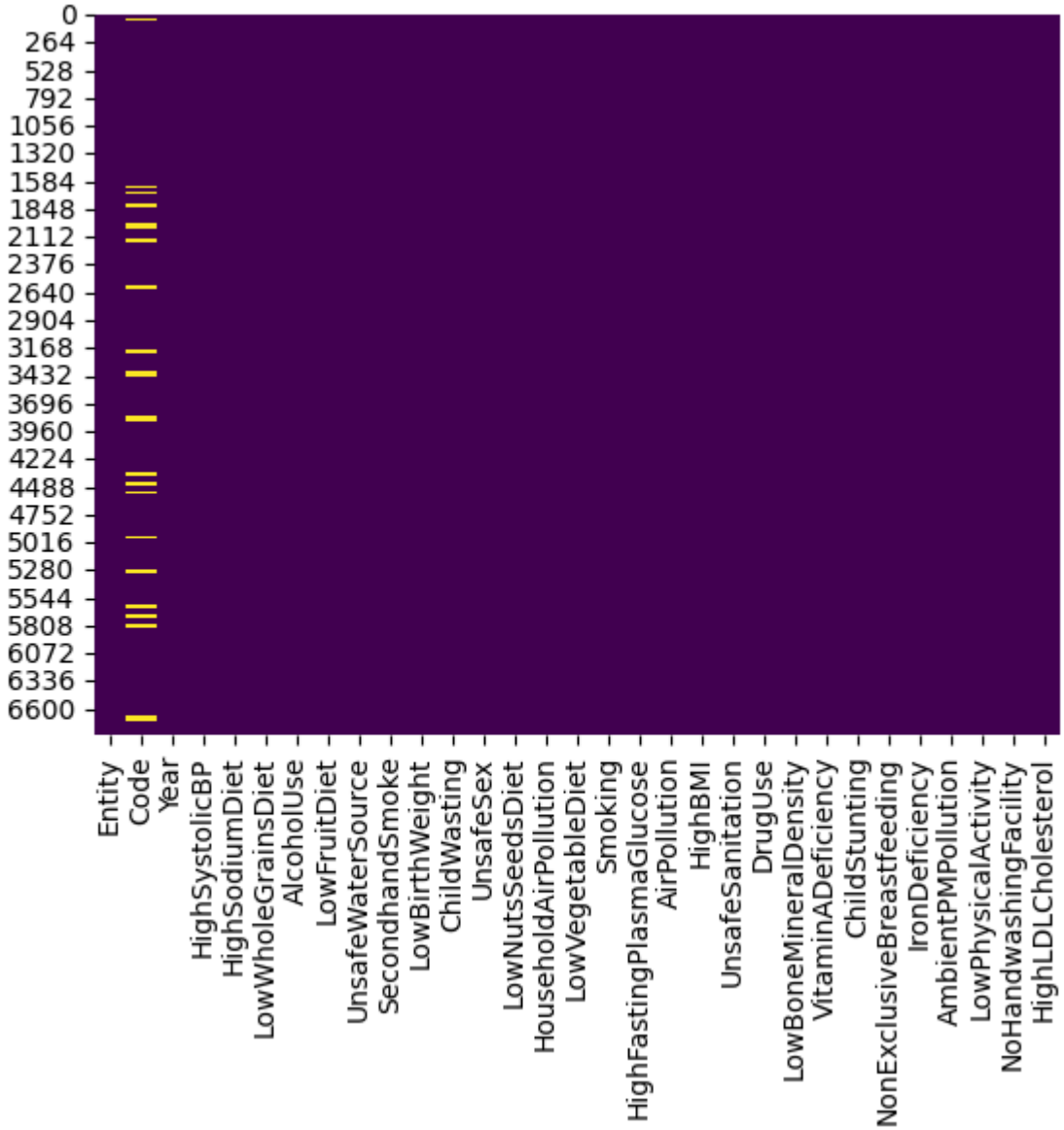
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6840 entries, 0 to 6839
Data columns (total 31 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Entity                     6840 non-null   object
 1   Code                       6150 non-null   object
 2   Year                       6840 non-null   int64
 3   HighSystolicBP             6840 non-null   float64
 4   HighSodiumDiet             6840 non-null   float64
 5   LowWholeGrainsDiet         6840 non-null   float64
 6   AlcoholUse                 6840 non-null   float64
 7   LowFruitDiet               6840 non-null   float64
 8   UnsafeWaterSource          6840 non-null   float64
 9   SecondhandSmoke            6840 non-null   float64
 10  LowBirthWeight             6840 non-null   float64
 11  ChildWasting               6840 non-null   float64
 12  UnsafeSex                  6840 non-null   float64
 13  LowNutsSeedsDiet           6840 non-null   float64
 14  HouseholdAirPollution      6840 non-null   float64
 15  LowVegetableDiet           6840 non-null   float64
 16  Smoking                    6840 non-null   float64
 17  HighFastingPlasmaGlucose   6840 non-null   float64
 18  AirPollution               6840 non-null   float64
 19  HighBMI                    6840 non-null   float64
 20  UnsafeSanitation           6840 non-null   float64
 21  DrugUse                    6840 non-null   float64
 22  LowBoneMineralDensity      6840 non-null   float64
 23  VitaminADeficiency         6840 non-null   float64
 24  ChildStunting              6840 non-null   float64
 25  NonExclusiveBreastfeeding  6840 non-null   float64
 26  IronDeficiency             6840 non-null   float64
 27  AmbientPMPollution         6840 non-null   float64
 28  LowPhysicalActivity        6840 non-null   float64
 29  NoHandwashingFacility      6840 non-null   float64
 30  HighLDLCholesterol         6840 non-null   float64
dtypes: float64(28), int64(1), object(2)
memory usage: 1.6+ MB
None
              Year  HighSystolicBP  HighSodiumDiet  LowWholeGrainsDiet  \
count  6840.000000    6.840000e+03    6.840000e+03        6.840000e+03
mean   2004.500000    2.242249e+05    4.049716e+04        3.869129e+04
std       8.656074    8.634691e+05    1.752832e+05        1.479084e+05
min    1990.000000    2.466349e+00    4.098751e-01        4.988057e-01
25%    1997.000000    1.827476e+03    1.365965e+02        2.733269e+02
50%    2004.500000    8.770671e+03    9.694122e+02        1.444022e+03
75%    2012.000000    4.035507e+04    5.169495e+03        6.773283e+03
max    2019.000000    1.084560e+07    1.885356e+06        1.844836e+06

         AlcoholUse  LowFruitDiet  UnsafeWaterSource  SecondhandSmoke  \
count  6.840000e+03  6.840000e+03       6.840000e+03     6.840000e+03
mean   5.484861e+04  2.395776e+04       4.408639e+04     3.036401e+04
std    2.112090e+05  9.451573e+04       2.020493e+05     1.222861e+05
min    1.457786e-01  3.029068e-01       2.800072e-03     5.062894e-01
25%    2.638683e+02  1.441008e+02       6.968999e+00     2.088802e+02
50%    1.780532e+03  8.346746e+02       1.824999e+02     9.937228e+02
75%    8.367695e+03  3.104619e+03       5.599117e+03     4.347840e+03
max    2.441974e+06  1.046015e+06       2.450944e+06     1.304318e+06

       LowBirthWeight  ChildWasting  ...        DrugUse  \
```

```
count    6.840000e+03  6.840000e+03  ...    6840.000000
mean     5.912551e+04  4.992437e+04  ...   10285.200840
std      2.502265e+05  2.226529e+05  ...   39960.741366
min      6.554637e-02  6.858725e-02  ...       0.054797
25%      1.230088e+02  2.628168e+01  ...      30.843025
50%      1.056921e+03  5.038514e+02  ...     221.868180
75%      1.090304e+04  9.765207e+03  ...    1224.431775
max      3.033426e+06  3.430422e+06  ...  494491.700000

       LowBoneMineralDensity  VitaminADeficiency  ChildStunting  \
count            6840.000000         6840.000000    6840.000000
mean             8182.476755         2471.615651   11164.355291
std             32403.920487        12718.301605   52866.235540
min                 0.050526            0.000003       0.000391
25%                42.931661            0.021711       0.873684
50%               277.144125            1.847342      41.521570
75%              1231.606000          230.444385    1563.387950
max            437884.400000       207555.220000  833448.560000

       NonExclusiveBreastfeeding  IronDeficiency  AmbientPMPollution  \
count                6840.000000     6840.000000        6.840000e+03
mean                 7171.866029     1421.394080        7.697212e+04
std                 31678.443811     6303.932513        3.183152e+05
min                     0.001889        0.000425        1.760401e-01
25%                     2.678312        0.757304        4.178013e+02
50%                    60.560858       11.722806        2.036275e+03
75%                  1315.425925      238.013368        1.127368e+04
max                505469.660000    73461.060000        4.140971e+06

       LowPhysicalActivity  NoHandwashingFacility  HighLDLCholesterol
count          6840.000000           6.840000e+03        6.840000e+03
mean          16489.085017           2.179990e+04        9.403521e+04
std           62708.007186           9.668258e+04        3.614803e+05
min               0.283192           1.961850e-02        1.069846e+00
25%              91.715955           1.893026e+01        5.582757e+02
50%             521.923925           2.213597e+02        3.122144e+03
75%            2820.269900           3.953427e+03        1.748821e+04
max          831502.000000           1.200349e+06        4.396984e+06

[8 rows x 29 columns]
        Entity Code  Year  HighSystolicBP  HighSodiumDiet  LowWholeGrainsDiet  \
0  Afghanistan  AFG  1990       25633.129       1044.9089           7077.3160
1  Afghanistan  AFG  1991       25871.803       1054.9584           7149.0854
2  Afghanistan  AFG  1992       26308.795       1074.6057           7297.3086
3  Afghanistan  AFG  1993       26961.360       1103.3705           7498.5340
4  Afghanistan  AFG  1994       27658.424       1133.8824           7697.5890

   AlcoholUse  LowFruitDiet  UnsafeWaterSource  SecondhandSmoke  ...  \
0    356.21470     3184.9550           3701.994        4794.4650  ...
1    363.73020     3248.3767           4309.282        4921.0957  ...
2    375.90024     3350.9207           5356.498        5278.5186  ...
3    388.57156     3479.8118           7151.521        5734.0303  ...
4    398.72700     3609.8315           7191.639        6050.2290  ...

     DrugUse  LowBoneMineralDensity  VitaminADeficiency  ChildStunting  \
0  173.57869              388.91074           2015.5115      7685.7427
1  187.89368              388.78424           2056.3538      7885.6724
2  210.81355              392.72090           2100.4310      8567.7400
3  232.17093              410.67044           2315.5906      9875.2900
4  247.29659              412.98883           2664.5537     11030.8480
```

```
   NonExclusiveBreastfeeding  IronDeficiency  AmbientPMPollution  \
0                  2216.0415       563.81067           2782.4385
1                  2501.0251       610.78830           2845.6702
2                  3052.5388       699.58734           3030.8933
3                  3725.8757       772.88920           3255.7598
4                  3832.5317       811.97064           3400.9597

   LowPhysicalActivity  NoHandwashingFacility  HighLDLCholesterol
0            2636.6455              4825.3450          12704.7810
1            2651.8865              5127.1780          12843.5130
2            2687.9224              5888.8438          13125.6210
3            2744.3599              7006.9080          13501.3545
4            2805.2195              7421.1280          13872.5840

[5 rows x 31 columns]
```

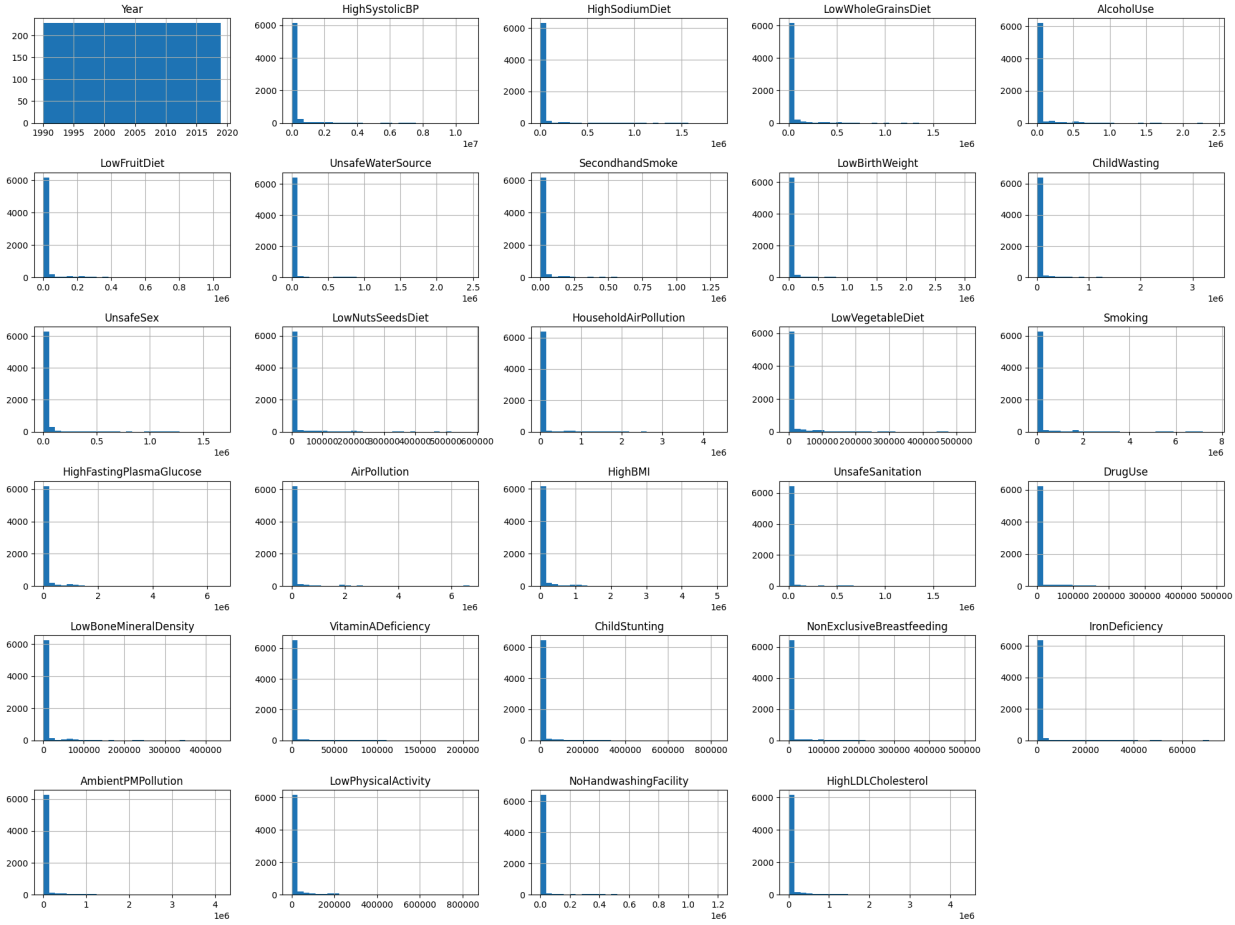
Visualize Missing Values

In [ ]:
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.show()
```



Univariate Analysis

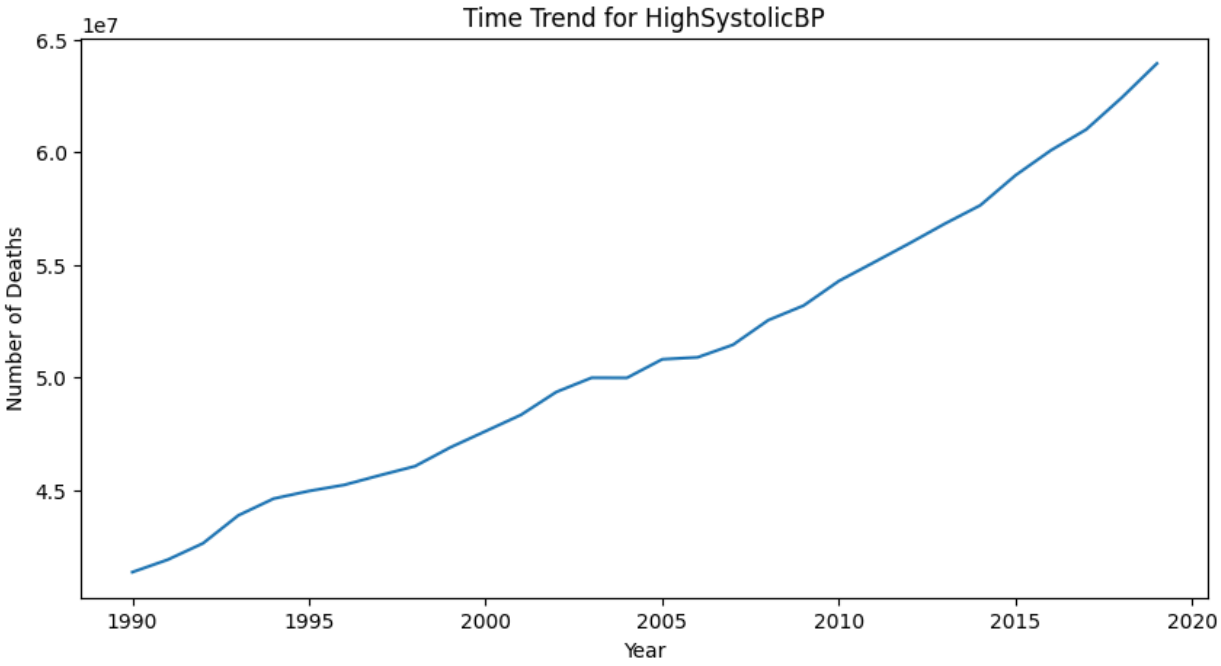
In [ ]:
```
df.hist(bins=30, figsize=(20, 15))
plt.tight_layout()
plt.show()
```

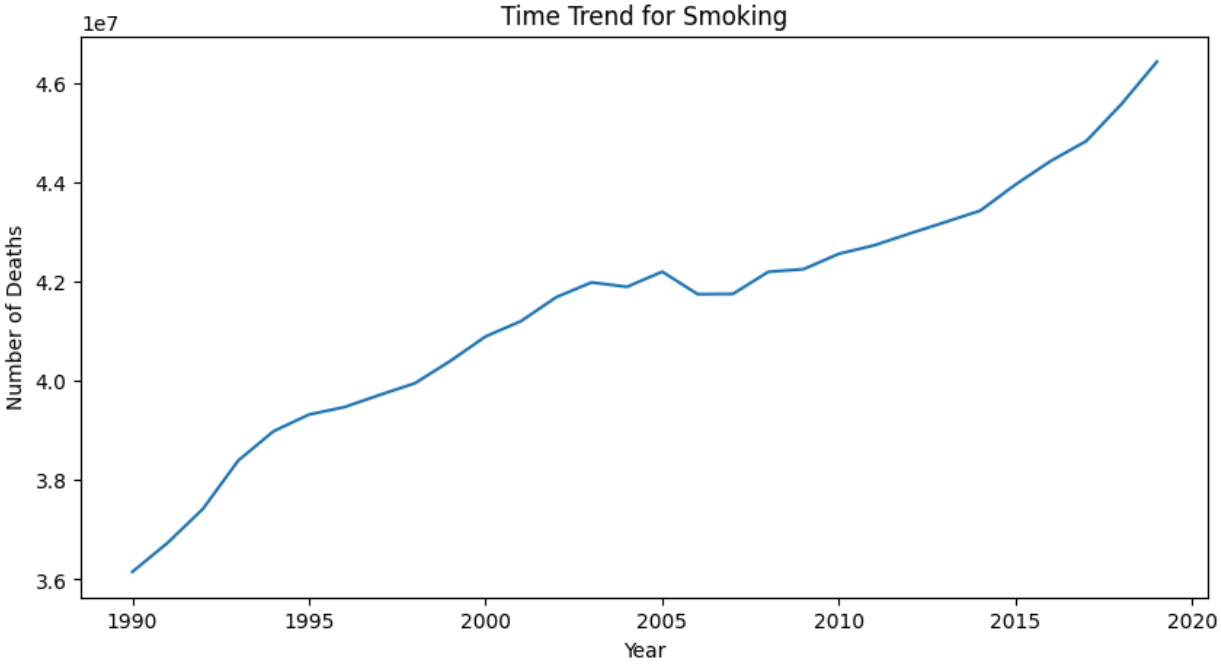


Deaths Over Time for Specific Causes

In [ ]:

```
causes = ['HighSystolicBP',
          'Smoking']
for cause in causes:
    plt.figure(figsize=(10, 5))
    df.groupby('Year')[cause].sum().plot()
    plt.title(f'Time Trend for {cause}')
    plt.ylabel('Number of Deaths')
    plt.show()
```



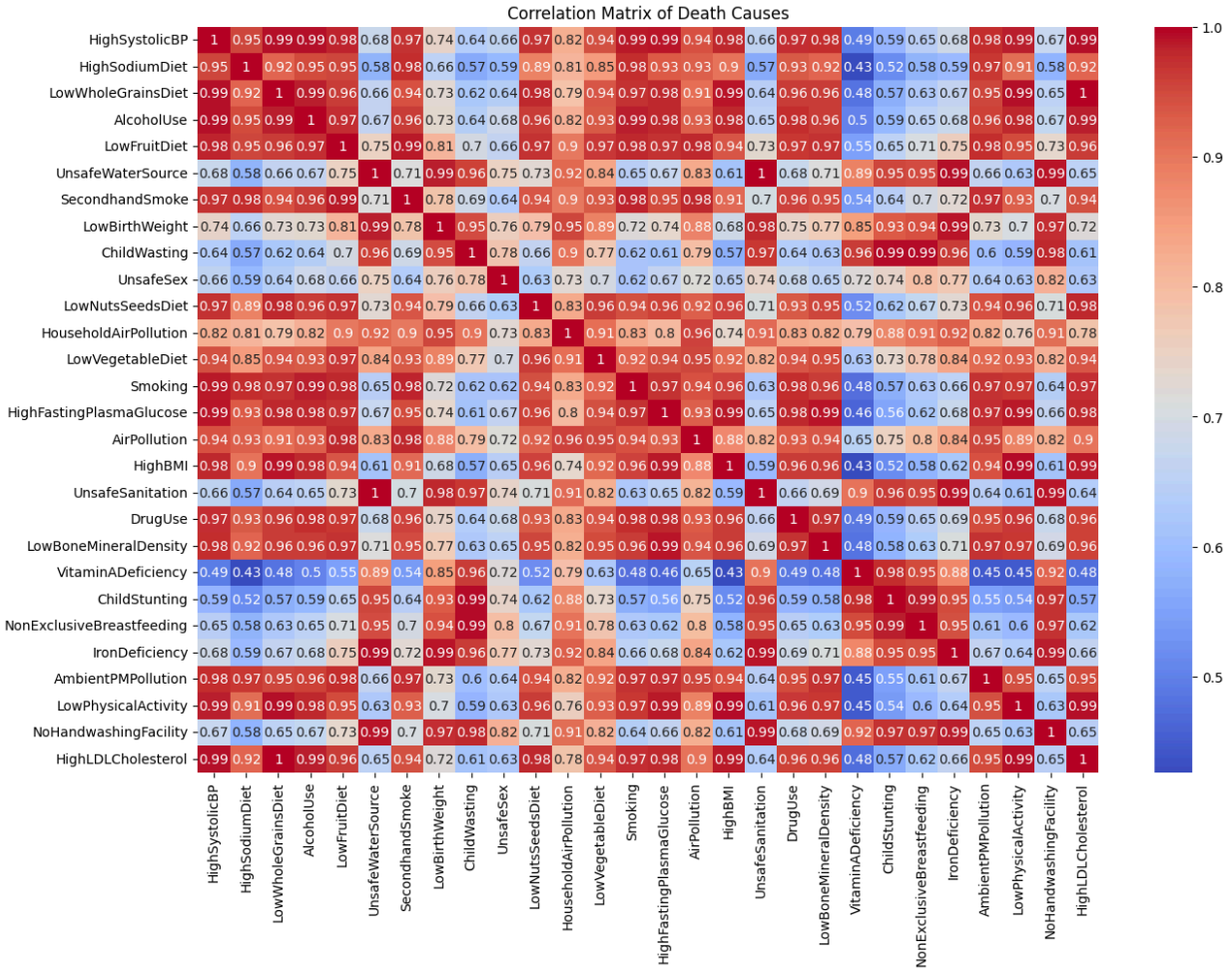


Bivariate Analysis

Correlation Matrix

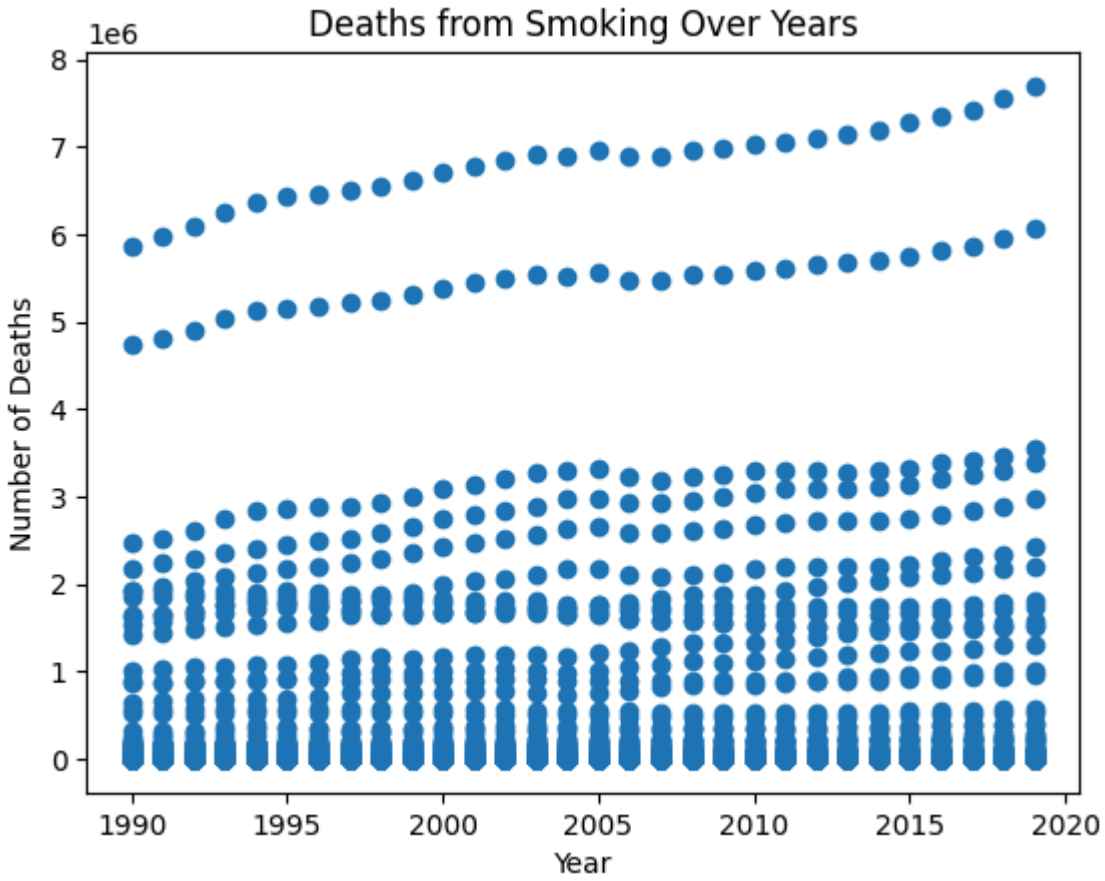
In [ ]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Select only the numeric columns
numeric_columns = [
    'HighSystolicBP', 'HighSodiumDiet', 'LowWholeGrainsDiet', 'AlcoholUse', 'LowFruitD
    'UnsafeWaterSource', 'SecondhandSmoke', 'LowBirthWeight', 'ChildWasting', 'UnsafeS
    'LowNutsSeedsDiet', 'HouseholdAirPollution', 'LowVegetableDiet', 'Smoking',
    'HighFastingPlasmaGlucose', 'AirPollution', 'HighBMI', 'UnsafeSanitation', 'DrugUs
    'LowBoneMineralDensity', 'VitaminADeficiency', 'ChildStunting', 'NonExclusiveBreas
    'IronDeficiency', 'AmbientPMPollution', 'LowPhysicalActivity', 'NoHandwashingFacil
    'HighLDLCholesterol'
```

```
]

# Compute the correlation matrix on numeric columns
corr = df[numeric_columns].corr()

# Plot the correlation matrix
plt.figure(figsize=(15, 10))
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix of Death Causes')
plt.show()
```



Scatter Plot of Deaths vs Year for a Specific Cause

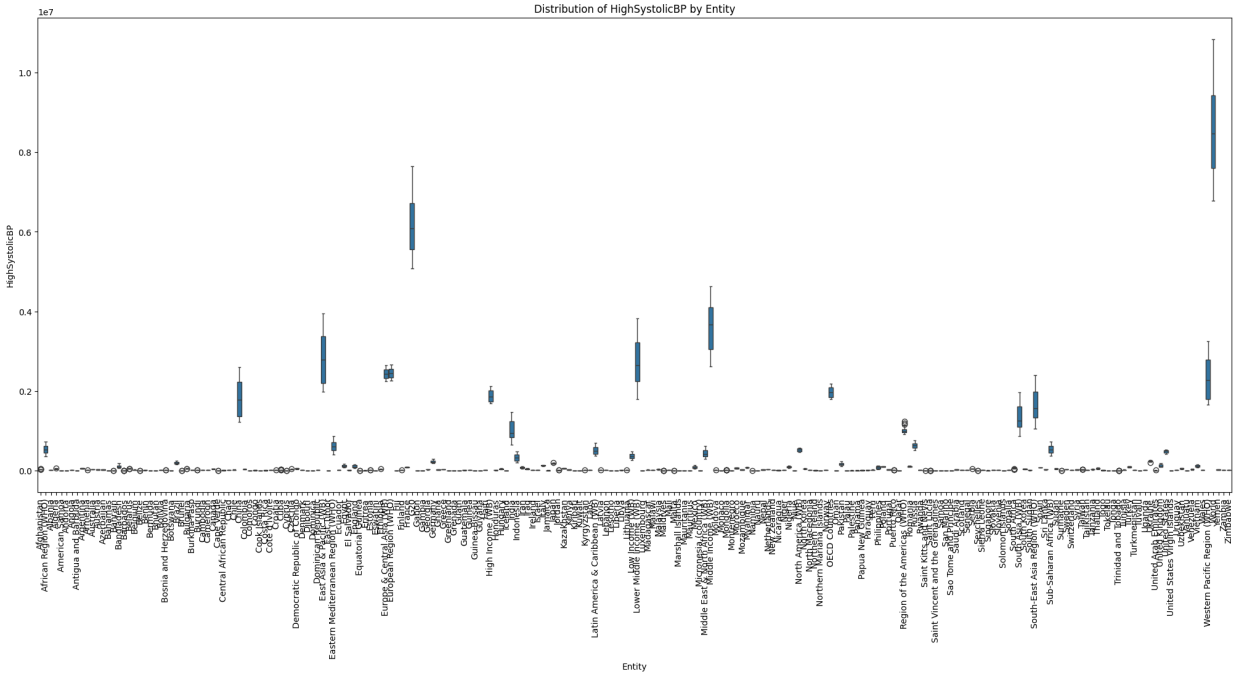
In [ ]:
```
plt.scatter(df['Year'], df['Smoking'])
plt.title('Deaths from Smoking Over Years')
plt.xlabel('Year')
plt.ylabel('Number of Deaths')
plt.show()
```

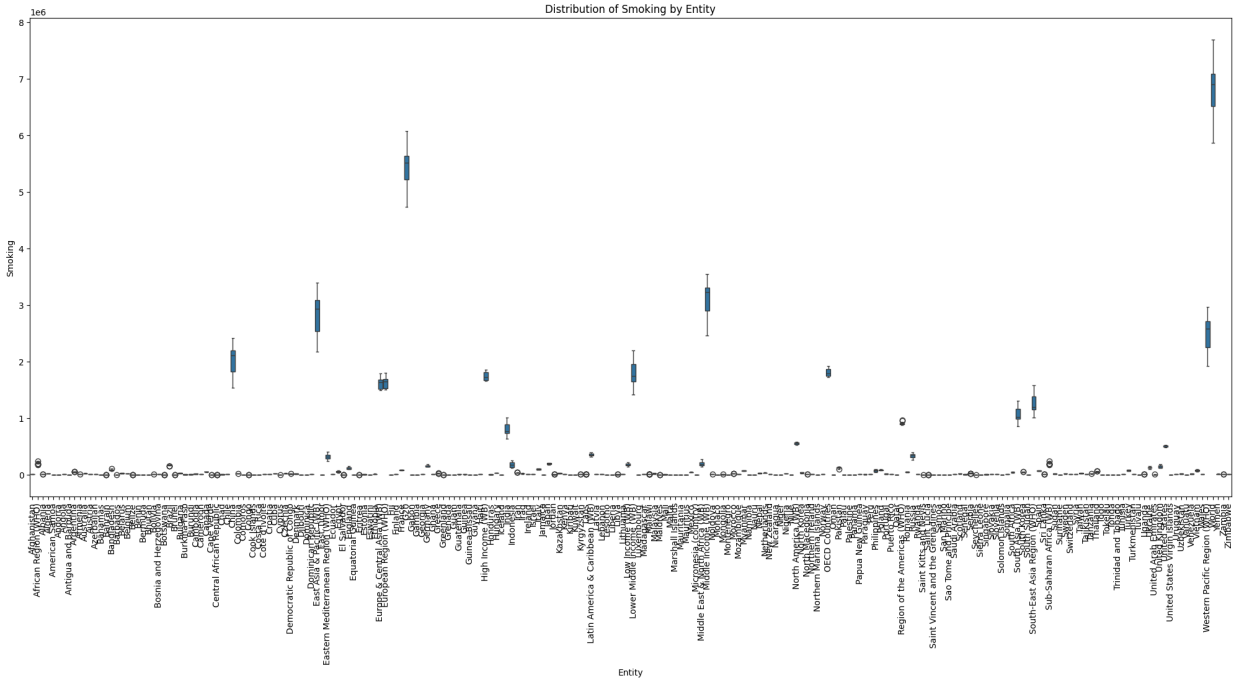


Multivariate Analysis

Deaths by Cause Across Different Entities

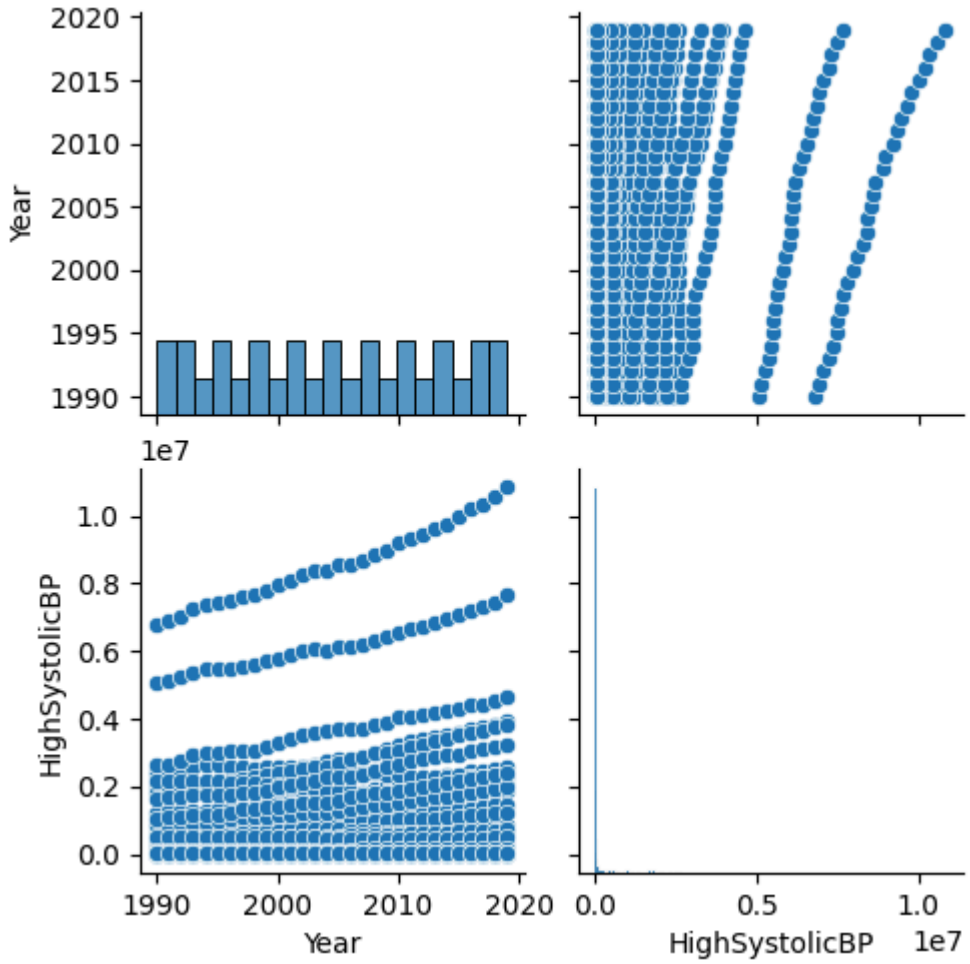
In [ ]:
```
causes = ['HighSystolicBP',
          'Smoking']
for cause in causes:
    plt.figure(figsize=(25, 10))
    sns.boxplot(x='Entity', y=cause, data=df)
    plt.xticks(rotation=90)
    plt.title(f'Distribution of {cause} by Entity')
    plt.show()
```





Pairwise Plot

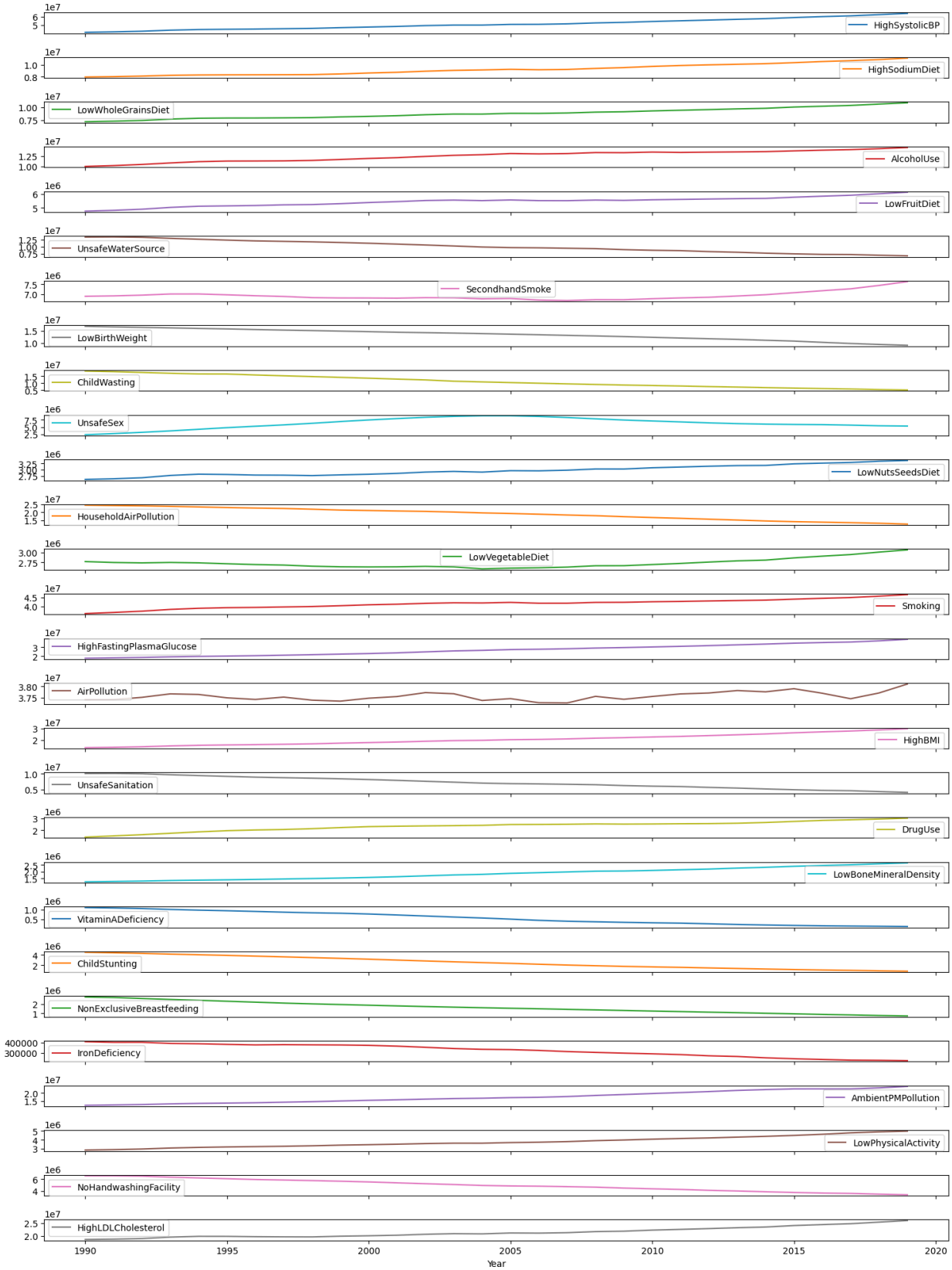
In [ ]:
```
sample_columns = df.columns[0:4]
sns.pairplot(df[sample_columns])
plt.show()
```



Time Series Analysis

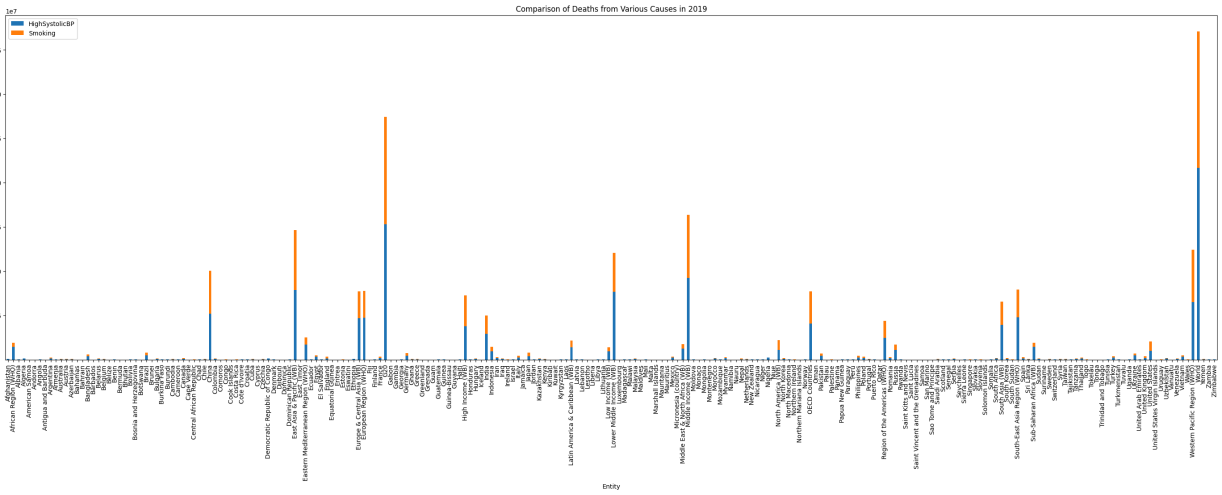
Trend Analysis

In [ ]:
```
df.groupby('Year').sum().plot(subplots=True, figsize=(15, 20))
plt.tight_layout()
plt.show()
```



In [ ]:

```
year = 2019
df_year = df[df['Year'] == year]
df_year.set_index('Entity')[causes].plot(kind='bar', stacked=True, figsize=(35, 10))
plt.title(f'Comparison of Deaths from Various Causes in {year}')
plt.show()
```



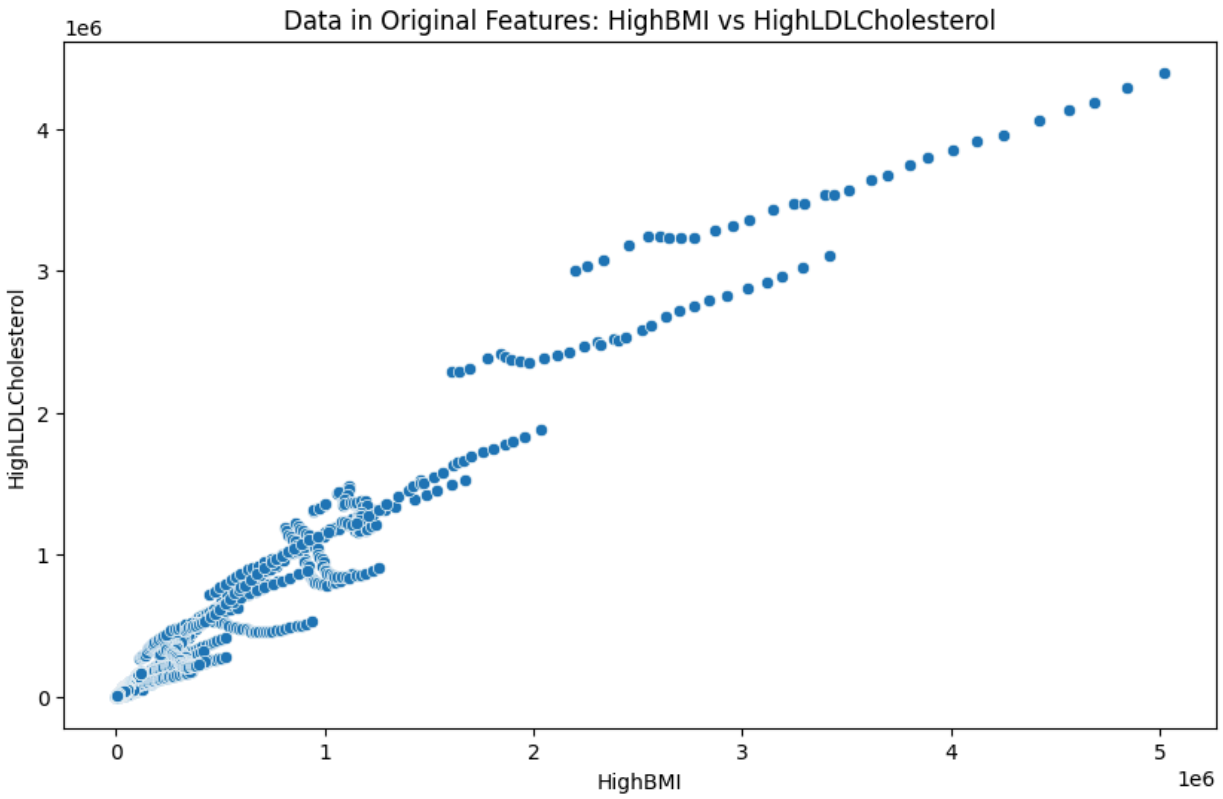
In [ ]:
```
df.head()
```

Out[ ]:

	Entity	Code	Year	HighSystolicBP	HighSodiumDiet	LowWholeGrainsDiet	AlcoholUse	LowFr
0	Afghanistan	AFG	1990	25633.129	1044.9089	7077.3160	356.21470	31
1	Afghanistan	AFG	1991	25871.803	1054.9584	7149.0854	363.73020	32
2	Afghanistan	AFG	1992	26308.795	1074.6057	7297.3086	375.90024	33
3	Afghanistan	AFG	1993	26961.360	1103.3705	7498.5340	388.57156	34
4	Afghanistan	AFG	1994	27658.424	1133.8824	7697.5890	398.72700	36

5 rows × 31 columns

In [ ]:
```
# Select two original features
feature1 = 'HighBMI'
feature2 = 'HighLDLCholesterol'

# Plot the data on two original features
plt.figure(figsize=(10, 6))
sns.scatterplot(x=feature1, y=feature2, data=df)
plt.title(f'Data in Original Features: {feature1} vs {feature2}')
plt.xlabel(feature1)
plt.ylabel(feature2)
plt.show()
```



PCA

In [ ]:

```
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import numpy as np
from sklearn.preprocessing import StandardScaler
numeric_columns = ['HighSystolicBP', 'HighSodiumDiet', 'LowWholeGrainsDiet', 'AlcoholU
                   'UnsafeWaterSource', 'SecondhandSmoke', 'LowBirthWeight', 'ChildWas
                   'LowNutsSeedsDiet', 'HouseholdAirPollution', 'LowVegetableDiet', 'S
                   'HighFastingPlasmaGlucose', 'AirPollution', 'HighBMI', 'UnsafeSanit
                   'LowBoneMineralDensity', 'VitaminADeficiency', 'ChildStunting', 'No
                   'IronDeficiency', 'AmbientPMPollution', 'LowPhysicalActivity', 'NoH
                   'HighLDLCholesterol']
df_numeric = df[numeric_columns]

df_numeric.fillna(df_numeric.mean(), inplace=True)

# Standardize the data
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df_numeric)

# Perform PCA
pca = PCA()
df_pca  = pca.fit_transform(df_scaled)
```

```
<ipython-input-10-03e15b2ed8cc>:15: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_numeric.fillna(df_numeric.mean(), inplace=True)
```

In [ ]:

```
# Convert to DataFrame for better interpretation
df_U = pd.DataFrame(df_pca , columns=[f'PC{i+1}' for i in range(df_pca .shape[1])])
df_V = pd.DataFrame(pca.components_.T, index=numeric_columns, columns=[f'PC{i+1}' for

print("Principal Component Scores (x):")
print(df_U.head())

print("\nLoadings (rotation):")
print(df_V.head())
```

```
Principal Component Scores (x):
        PC1       PC2       PC3       PC4       PC5       PC6       PC7  \
0 -1.042944  0.165559 -0.068950  0.014747  0.080688 -0.020865 -0.004869
1 -1.032714  0.175879 -0.071627  0.013351  0.080934 -0.023048 -0.006430
2 -1.011322  0.198140 -0.077542  0.012769  0.082165 -0.027068 -0.011030
3 -0.981823  0.232577 -0.082608  0.015044  0.094692 -0.026975 -0.011918
4 -0.963880  0.253787 -0.083293  0.020005  0.113141 -0.023438 -0.007992

        PC8       PC9      PC10  ...      PC19      PC20      PC21      PC22  \
0  0.018092  0.015367 -0.070738  ...  0.028110  0.026803  0.026714  0.006575
1  0.021362  0.013150 -0.069147  ...  0.025047  0.027824  0.028522  0.005911
2  0.029716  0.007537 -0.064854  ...  0.025353  0.029519  0.031253  0.004946
3  0.036542  0.007448 -0.061210  ...  0.029904  0.031223  0.033915  0.004439
4  0.034096  0.010503 -0.070153  ...  0.036901  0.034328  0.037901  0.003904

       PC23      PC24      PC25      PC26      PC27      PC28
0  0.000176 -0.012599 -0.009434  0.000663 -0.006793 -0.000094
1 -0.001237 -0.011752 -0.010168  0.000470 -0.006932 -0.000144
2 -0.002601 -0.010824 -0.011925  0.000092 -0.007176 -0.000195
3 -0.003859 -0.010486 -0.013527 -0.000301 -0.007427 -0.000234
4 -0.004238 -0.010161 -0.014817 -0.000751 -0.008464 -0.000236

[5 rows x 28 columns]

Loadings (rotation):
                         PC1       PC2       PC3       PC4       PC5  \
HighSystolicBP      0.198821 -0.153010  0.032569 -0.004778  0.076051
HighSodiumDiet      0.186314 -0.178737 -0.099927  0.478546  0.024523
LowWholeGrainsDiet  0.195751 -0.158560  0.079884 -0.163240  0.195494
AlcoholUse          0.197725 -0.149117  0.105447  0.040825  0.137064
LowFruitDiet        0.202980 -0.103313 -0.131291  0.069157 -0.078918

                         PC6       PC7       PC8       PC9      PC10  ...  \
HighSystolicBP     -0.056804  0.083304  0.041874 -0.092305 -0.005646  ...
HighSodiumDiet      0.012733  0.178386 -0.014340 -0.160272  0.123824  ...
LowWholeGrainsDiet -0.195169  0.066166  0.002483 -0.150163  0.077171  ...
AlcoholUse         -0.044199 -0.118915 -0.311732 -0.049550  0.227103  ...
LowFruitDiet       -0.115061 -0.022067 -0.190337  0.186578 -0.070404  ...

                        PC19      PC20      PC21      PC22      PC23  \
HighSystolicBP     -0.066235 -0.141685  0.158826  0.263056  0.128502
HighSodiumDiet     -0.237274 -0.248300  0.108447 -0.230669 -0.074529
LowWholeGrainsDiet -0.069088  0.105419 -0.142682  0.169046  0.161236
AlcoholUse          0.122875  0.057008  0.118665  0.276164 -0.014868
LowFruitDiet        0.007743 -0.201532  0.326752 -0.023680  0.425623

                        PC24      PC25      PC26      PC27      PC28
HighSystolicBP      0.170285  0.105555 -0.754437  0.054178 -0.060044
HighSodiumDiet     -0.012256 -0.058674  0.296171 -0.000627 -0.003647
LowWholeGrainsDiet -0.381392  0.593597  0.265584 -0.082844 -0.000867
AlcoholUse         -0.047444  0.057903  0.026237  0.006911 -0.003742
LowFruitDiet        0.048700 -0.176454  0.202634 -0.244685  0.028641

[5 rows x 28 columns]
```

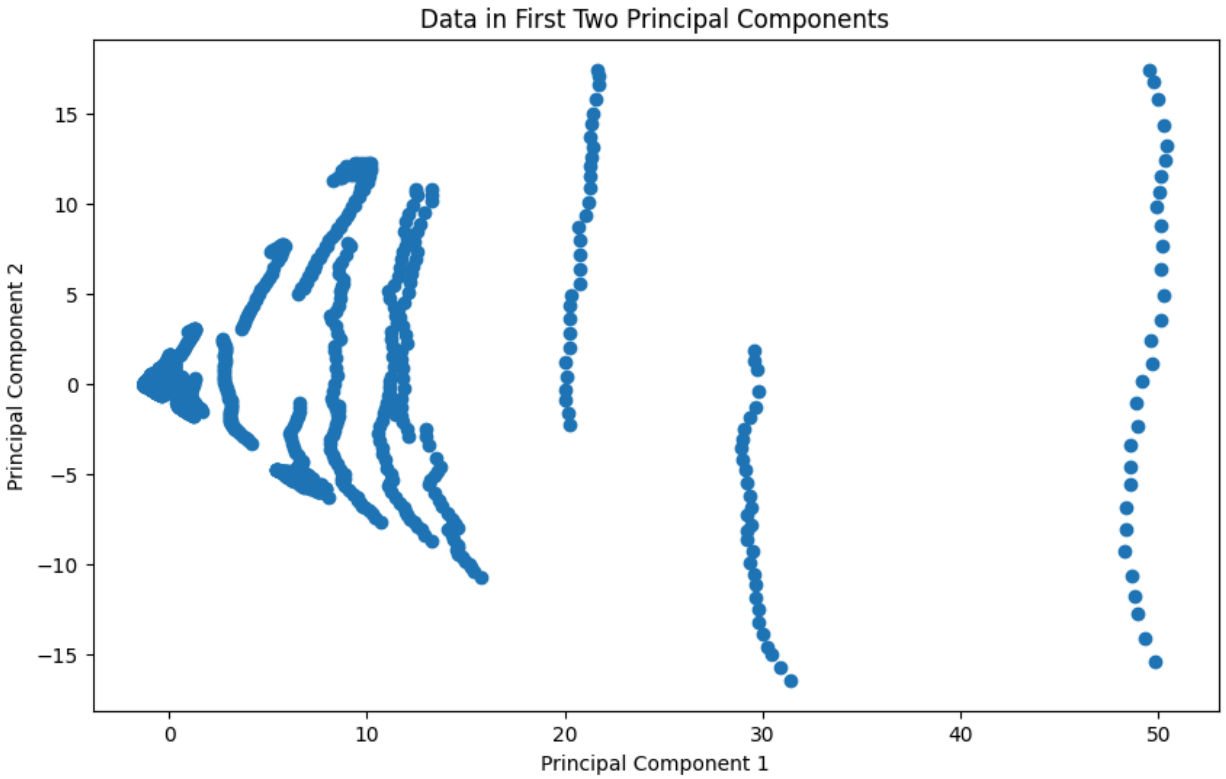
In [ ]:

```python
# Inspect the loadings for the first two principal components
print("Loadings for PC1:")
print(df_V['PC1'].sort_values(ascending=False).head())
```

```
print("\nLoadings for PC2:")
print(df_V['PC2'].sort_values(ascending=False).head())
```

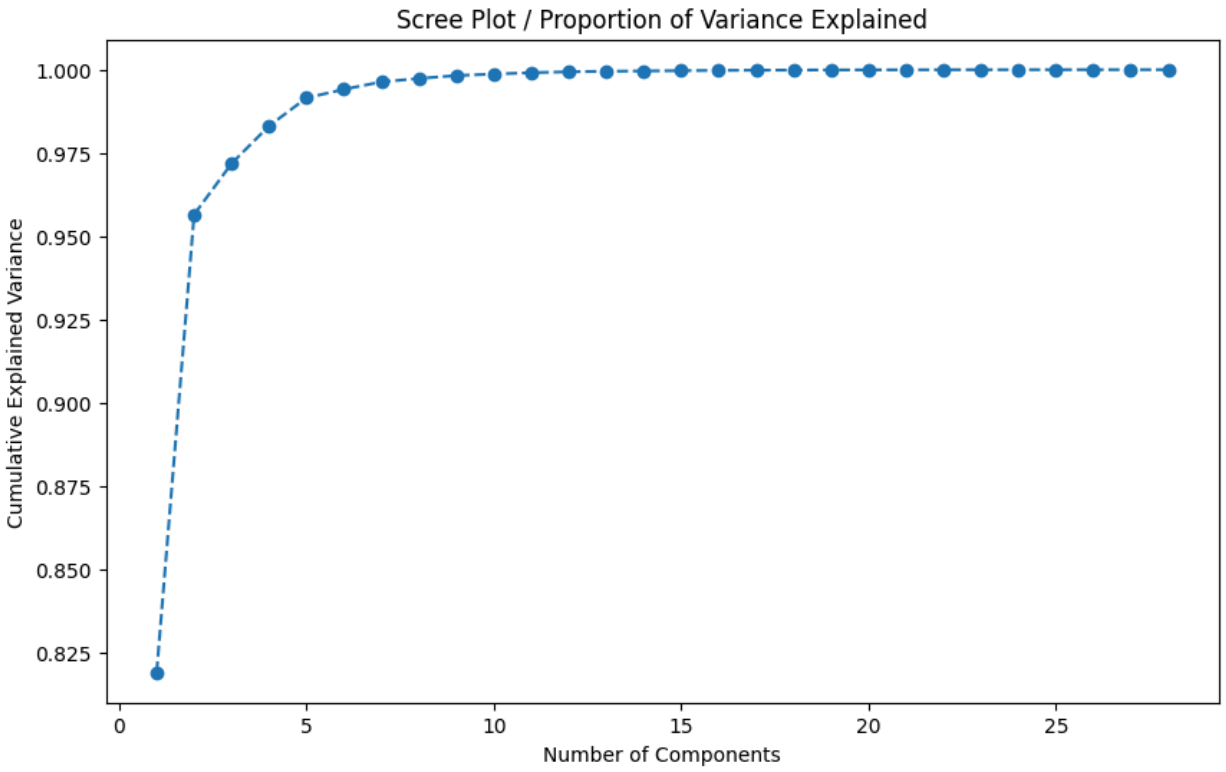
```
Loadings for PC1:
AirPollution        0.204743
LowVegetableDiet    0.204115
LowFruitDiet        0.202980
SecondhandSmoke     0.199373
HighSystolicBP      0.198821
Name: PC1, dtype: float64

Loadings for PC2:
VitaminADeficiency           0.328358
ChildStunting                0.298757
ChildWasting                 0.274769
NonExclusiveBreastfeeding    0.263918
UnsafeSanitation             0.253980
Name: PC2, dtype: float64
```

In [ ]:
```
# Plot the first two principal components
plt.figure(figsize=(10, 6))
plt.scatter(df_U['PC1'], df_U['PC2'])
plt.title('Data in First Two Principal Components')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.show()
```



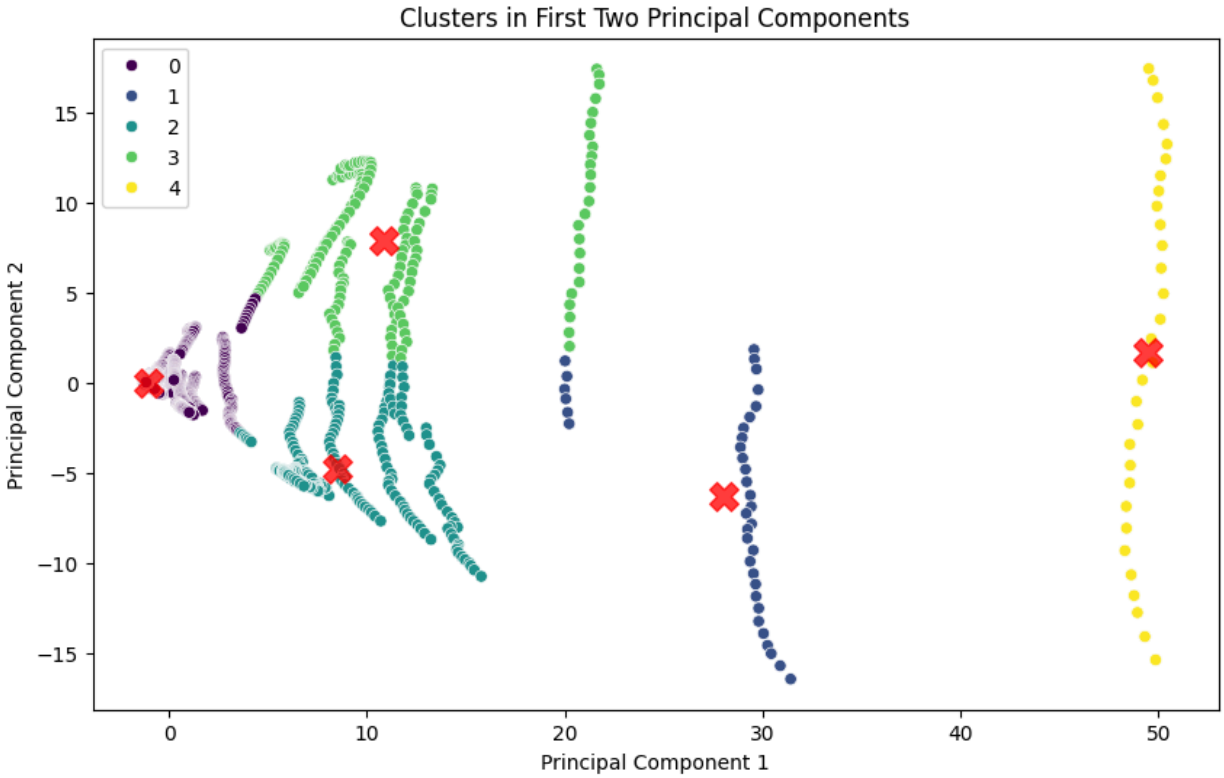
In [ ]:
```
explained_variance = pca.explained_variance_ratio_

# Scree plot
plt.figure(figsize=(10, 6))
plt.plot(range(1, len(explained_variance) + 1), np.cumsum(explained_variance), marker=
plt.title('Scree Plot / Proportion of Variance Explained')
plt.xlabel('Number of Components')
```

```
plt.ylabel('Cumulative Explained Variance')
plt.show()
```



K-Means

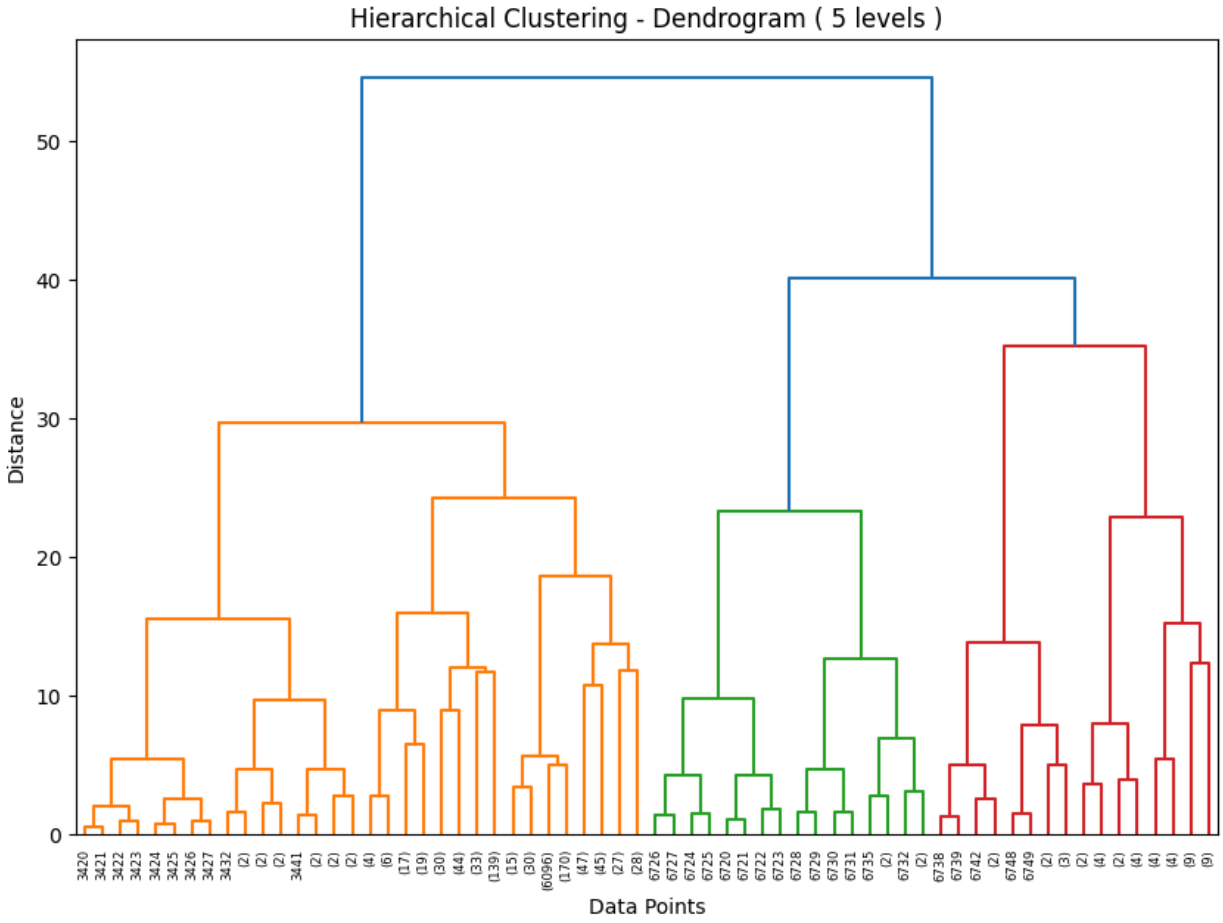
In [ ]:
```
from sklearn.cluster import KMeans
import seaborn as sns

# Perform K-means clustering
kmeans = KMeans(n_clusters=5, n_init=10)
clusters = kmeans.fit_predict(df_U[['PC1', 'PC2']])

# Add cluster labels to the PCA data
df_U['Cluster'] = clusters

# Plot clusters in the first two principal components
plt.figure(figsize=(10, 6))
sns.scatterplot(x='PC1', y='PC2', hue='Cluster', data=df_U, palette='viridis')

# Plot the cluster centers
centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], c='red', s=200, alpha=0.75, marker='X')  # C

plt.title('Clusters in First Two Principal Components')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.legend()
plt.show()
```



HIERARCHICAL CLUSTERING

In [ ]:
```
from sklearn.cluster import KMeans,AgglomerativeClustering
from scipy.cluster.hierarchy import dendrogram, linkage, cut_tree
import matplotlib.pyplot as plt
```

In [ ]:
```
from scipy.cluster.hierarchy import dendrogram, linkage

# Perform hierarchical/agglomerative clustering
linked = linkage(df_scaled, method='complete', metric='euclidean')

# Plot the dendrogram with truncation to show the top 'p' merges
plt.figure(figsize=(10, 7))
dendrogram(
    linked,
    truncate_mode='level',
    p=5
)
plt.title('Hierarchical Clustering - Dendrogram ( 5 levels )')
plt.xlabel('Data Points')
plt.ylabel('Distance')
plt.show()
```



In [ ]:
```
cluster_labels = cut_tree(linked, n_clusters=5).flatten()

# Add cluster labels to the original DataFrame
df['Cluster'] = cluster_labels

# Display a few observations from each cluster
for i in range(5):
    print(f"Cluster {i}:\n", df[df['Cluster'] == i].head(), "\n")
```

Cluster 0:

	Entity	Code	Year	HighSystolicBP	HighSodiumDiet	LowWholeGrainsDiet	\
0	Afghanistan	AFG	1990	25633.129	1044.9089	7077.3160	
1	Afghanistan	AFG	1991	25871.803	1054.9584	7149.0854	
2	Afghanistan	AFG	1992	26308.795	1074.6057	7297.3086	
3	Afghanistan	AFG	1993	26961.360	1103.3705	7498.5340	
4	Afghanistan	AFG	1994	27658.424	1133.8824	7697.5890	

	AlcoholUse	LowFruitDiet	UnsafeWaterSource	SecondhandSmoke	...	\
0	356.21470	3184.9550	3701.994	4794.4650	...	
1	363.73020	3248.3767	4309.282	4921.0957	...	
2	375.90024	3350.9207	5356.498	5278.5186	...	
3	388.57156	3479.8118	7151.521	5734.0303	...	
4	398.72700	3609.8315	7191.639	6050.2290	...	

	LowBoneMineralDensity	VitaminADeficiency	ChildStunting	\
0	388.91074	2015.5115	7685.7427	
1	388.78424	2056.3538	7885.6724	
2	392.72090	2100.4310	8567.7400	
3	410.67044	2315.5906	9875.2900	
4	412.98883	2664.5537	11030.8480	

	NonExclusiveBreastfeeding	IronDeficiency	AmbientPMPollution	\
0	2216.0415	563.81067	2782.4385	
1	2501.0251	610.78830	2845.6702	
2	3052.5388	699.58734	3030.8933	
3	3725.8757	772.88920	3255.7598	
4	3832.5317	811.97064	3400.9597	

	LowPhysicalActivity	NoHandwashingFacility	HighLDLCholesterol	Cluster
0	2636.6455	4825.3450	12704.7810	0
1	2651.8865	5127.1780	12843.5130	0
2	2687.9224	5888.8438	13125.6210	0
3	2744.3599	7006.9080	13501.3545	0
4	2805.2195	7421.1280	13872.5840	0

[5 rows x 32 columns]

Cluster 1:

	Entity	Code	Year	HighSystolicBP	HighSodiumDiet	\
1634	East Asia & Pacific (WB)	NaN	2004	2745792.2	907674.50	
1635	East Asia & Pacific (WB)	NaN	2005	2810583.0	914782.44	
1636	East Asia & Pacific (WB)	NaN	2006	2826551.2	904717.75	
1637	East Asia & Pacific (WB)	NaN	2007	2882357.0	908266.20	
1638	East Asia & Pacific (WB)	NaN	2008	2987330.2	927609.06	

	LowWholeGrainsDiet	AlcoholUse	LowFruitDiet	UnsafeWaterSource	\
1634	343602.88	642318.6	366525.56	122243.58	
1635	357097.56	645710.0	365920.94	117125.35	
1636	362687.90	643315.3	357127.60	111858.21	
1637	372157.94	651762.6	352503.16	106932.40	
1638	387442.28	676274.2	352217.97	102931.66	

	SecondhandSmoke	...	LowBoneMineralDensity	VitaminADeficiency	\
1634	518228.75	...	96940.39	5237.2744	
1635	516556.97	...	102257.52	4138.5570	
1636	505553.62	...	105060.13	3630.3806	
1637	501501.44	...	108099.32	3282.9560	
1638	506169.12	...	110603.16	2467.5417	

      ChildStunting NonExclusiveBreastfeeding IronDeficiency \
1634   35586.836   31588.229   5012.6550
1635   31422.691   28662.695   4658.2550
1636   28252.701   25884.600   4312.6230
1637   25596.041   23399.822   4030.5835
1638   22395.246   21305.424   3787.7500

      AmbientPMPollution LowPhysicalActivity NoHandwashingFacility \
1634   1204189.9   138412.42   58879.940
1635   1235153.8   144076.42   56563.785
1636   1241157.9   146372.55   54054.840
1637   1275071.6   150733.70   51784.420
1638   1334181.1   157998.42   50345.145

      HighLDLCholesterol Cluster
1634   855554.20   1
1635   887304.90   1
1636   901332.30   1
1637   924818.75   1
1638   963473.90   1

[5 rows x 32 columns]

Cluster 2:
      Entity Code Year HighSystolicBP HighSodiumDiet LowWholeGrainsDiet \
2130   G20 NaN 1990   5083835.0   1076589.1   842299.75
2131   G20 NaN 1991   5142207.5   1083571.9   850738.75
2132   G20 NaN 1992   5220326.0   1095109.9   863492.50
2133   G20 NaN 1993   5372135.0   1112442.6   895526.00
2134   G20 NaN 1994   5451538.5   1117792.2   912362.60

      AlcoholUse LowFruitDiet UnsafeWaterSource SecondhandSmoke ... \
2130   1253649.8   608856.75   1174104.4   877675.06 ...
2131   1279095.5   615339.00   1175878.2   881099.50 ...
2132   1314723.1   624271.44   1176471.1   885227.10 ...
2133   1365434.4   640535.00   1126499.2   891712.50 ...
2134   1405403.1   649739.75   1072857.8   889809.20 ...

      LowBoneMineralDensity VitaminADeficiency ChildStunting \
2130   160937.97   60853.465   331514.30
2131   164236.34   59583.117   318794.25
2132   167578.47   58842.355   306621.66
2133   172761.03   54060.684   288488.16
2134   176372.60   49723.574   271957.00

      NonExclusiveBreastfeeding IronDeficiency AmbientPMPollution \
2130   211652.47   39318.490   1548507.1
2131   200544.73   38545.670   1575175.8
2132   188908.06   38133.350   1608986.5
2133   176211.61   36246.305   1661398.0
2134   163671.45   35375.215   1697513.0

      LowPhysicalActivity NoHandwashingFacility HighLDLCholesterol Cluster
2130   344373.50   519418.78   2287836.2   2
2131   349910.30   515901.03   2293769.8   2
2132   357380.50   513236.10   2312129.2   2
2133   370943.66   488146.16   2383957.2   2
2134   378696.56   463487.80   2414717.8   2

[5 rows x 32 columns]

Cluster 3:

	Entity	Code	Year	HighSystolicBP	HighSodiumDiet	LowWholeGrainsDiet	AlcoholUse	LowFruitDiet	UnsafeWaterSource	SecondhandSmoke	...	LowBoneMineralDensity	VitaminADeficiency	ChildStunting	NonExclusiveBreastfeeding	IronDeficiency	AmbientPMPollution	LowPhysicalActivity	NoHandwashingFacility	HighLDLCholesterol	Cluster
6720	World	OWID_WRL	1990	6787714.5	1320338.0	1178221.8	1639872.2	795603.40	2442070.5	1161963.1	...	207366.97	207555.22	833448.56	505469.66	73461.06	2047171.0	452167.47	1200094.0	3002611.0	3
6721	World	OWID_WRL	1991	6888724.5	1331430.2	1197636.4	1676080.6	805993.50	2450943.8	1165940.0	...	211921.06	203551.69	817068.30	497187.28	72379.37	2087355.1	460917.70	1200348.9	3030906.0	3
6722	World	OWID_WRL	1992	7023441.5	1348867.0	1223711.1	1723426.2	821076.70	2425768.5	1173047.5	...	216470.62	198336.60	796502.75	478093.88	72502.98	2138114.5	472631.30	1190921.8	3076783.5	3
6723	World	OWID_WRL	1993	7240557.5	1372509.9	1271470.1	1787248.6	844230.75	2361329.8	1183325.8	...	223118.44	189702.84	771675.56	461114.44	70802.30	2210730.0	491233.06	1162487.8	3181351.0	3
6724	World	OWID_WRL	1994	7378238.0	1383663.9	1301408.9	1841376.8	859534.40	2310204.8	1184536.0	...	227824.56	183165.19	750688.90	445705.38	70290.90	2265125.0	503301.40	1141067.4	3239905.8	3

[5 rows x 32 columns]

Cluster 4:

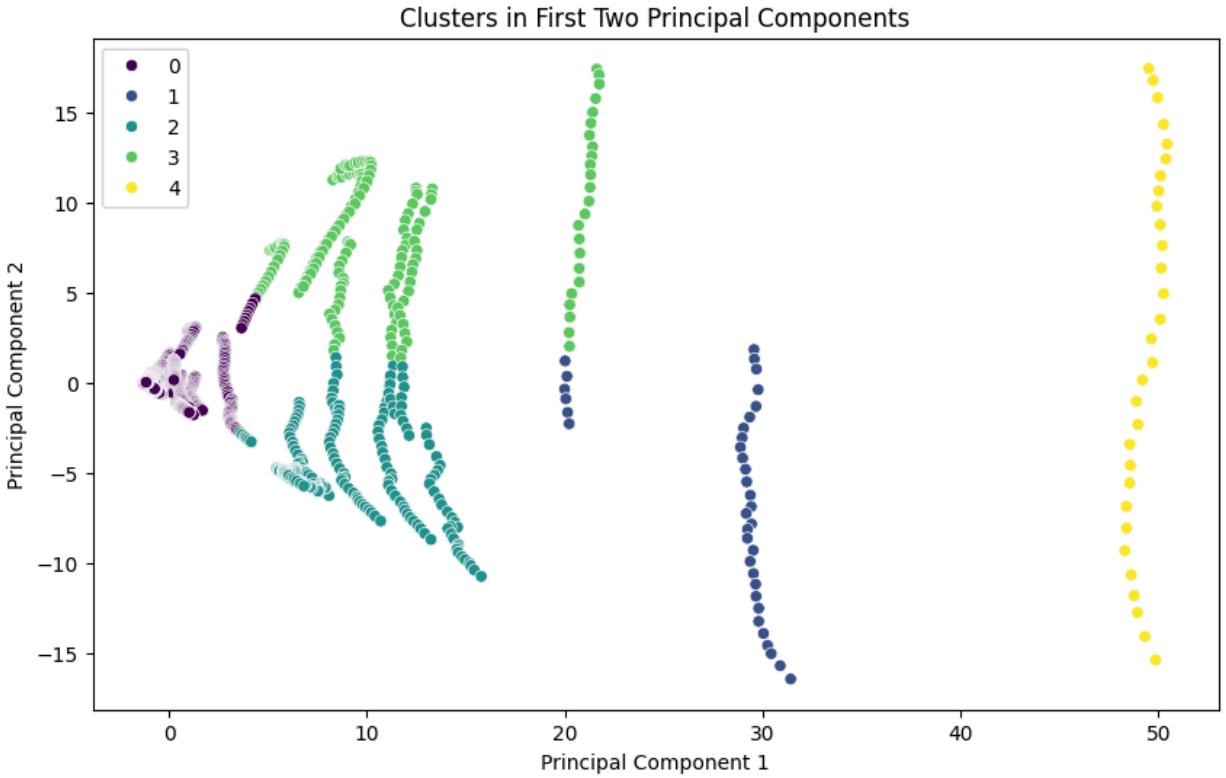
	Entity	Code	Year	HighSystolicBP	HighSodiumDiet	LowWholeGrainsDiet	AlcoholUse	LowFruitDiet	UnsafeWaterSource
6740	World	OWID_WRL	2010	9181355.0	1644632.5	1569687.1	2249855.0	950118.6	1592303.1
6741	World	OWID_WRL	2011	9324965.0	1669140.9	1591653.4	2235098.0	954610.8	1566353.5
6742	World	OWID_WRL	2012	9470980.0	1688558.2	1614793.0	2247303.8	959393.3	1504617.0
6743	World	OWID_WRL	2013	9622330.0	1705269.9	1639467.5	2258003.2	964258.9	1461534.5
6744	World	OWID_WRL	2014	9765811.0	1723074.0	1659528.2	2270896.0	967793.9	1406151.1

```
      SecondhandSmoke  ...  LowBoneMineralDensity  VitaminADeficiency  \
6740        1156229.0  ...              346141.38           59766.840
6741        1163170.4  ...              354746.53           56697.074
6742        1169127.6  ...              362471.88           50333.707
6743        1180237.8  ...              373945.90           43588.400
6744        1191687.0  ...              383968.34           38614.645

      ChildStunting  NonExclusiveBreastfeeding  IronDeficiency  \
6740      317920.88                  236921.69       54115.160
6741      301832.90                  226513.64       52796.562
6742      281589.25                  214395.05       50747.793
6743      262374.97                  203990.50       49599.990
6744      243105.83                  192370.47       47387.098

      AmbientPMPollution  LowPhysicalActivity  NoHandwashingFacility  \
6740           3359355.0            672215.00              809831.70
6741           3462498.2            685827.56              790259.10
6742           3575957.2            697604.90              762949.94
6743           3698902.2            714201.20              744109.25
6744           3792901.2            729791.75              720558.30

      HighLDLCholesterol  Cluster
6740           3746423.5        4
6741           3796330.5        4
6742           3850347.2        4
6743           3909372.5        4
6744           3959287.0        4

[5 rows x 32 columns]
```

Specific Criteria

In [ ]:

```python
# Plot clusters in the first two principal components
plt.figure(figsize=(10, 6))
sns.scatterplot(x='PC1', y='PC2', hue='Cluster', data=df_U, palette='viridis')
plt.title('Clusters in First Two Principal Components')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.legend()
plt.show()
```



In [ ]:

```
# Add cluster labels to the original DataFrame
df['Cluster'] = clusters

# Display a few observations from each cluster
for i in range(5):
    print(f"Cluster {i}:\n", df[df['Cluster'] == i].head(), "\n")
```

Cluster 0:

	Entity	Code	Year	HighSystolicBP	HighSodiumDiet	LowWholeGrainsDiet	\
0	Afghanistan	AFG	1990	25633.129	1044.9089	7077.3160	
1	Afghanistan	AFG	1991	25871.803	1054.9584	7149.0854	
2	Afghanistan	AFG	1992	26308.795	1074.6057	7297.3086	
3	Afghanistan	AFG	1993	26961.360	1103.3705	7498.5340	
4	Afghanistan	AFG	1994	27658.424	1133.8824	7697.5890	

	AlcoholUse	LowFruitDiet	UnsafeWaterSource	SecondhandSmoke	...	\
0	356.21470	3184.9550	3701.994	4794.4650	...	
1	363.73020	3248.3767	4309.282	4921.0957	...	
2	375.90024	3350.9207	5356.498	5278.5186	...	
3	388.57156	3479.8118	7151.521	5734.0303	...	
4	398.72700	3609.8315	7191.639	6050.2290	...	

	LowBoneMineralDensity	VitaminADeficiency	ChildStunting	\
0	388.91074	2015.5115	7685.7427	
1	388.78424	2056.3538	7885.6724	
2	392.72090	2100.4310	8567.7400	
3	410.67044	2315.5906	9875.2900	
4	412.98883	2664.5537	11030.8480	

	NonExclusiveBreastfeeding	IronDeficiency	AmbientPMPollution	\
0	2216.0415	563.81067	2782.4385	
1	2501.0251	610.78830	2845.6702	
2	3052.5388	699.58734	3030.8933	
3	3725.8757	772.88920	3255.7598	
4	3832.5317	811.97064	3400.9597	

	LowPhysicalActivity	NoHandwashingFacility	HighLDLCholesterol	Cluster
0	2636.6455	4825.3450	12704.7810	0
1	2651.8865	5127.1780	12843.5130	0
2	2687.9224	5888.8438	13125.6210	0
3	2744.3599	7006.9080	13501.3545	0
4	2805.2195	7421.1280	13872.5840	0

[5 rows x 32 columns]

Cluster 1:

	Entity	Code	Year	HighSystolicBP	HighSodiumDiet	\
6720	World	OWID_WRL	1990	6787714.5	1320338.0	
6721	World	OWID_WRL	1991	6888724.5	1331430.2	
6722	World	OWID_WRL	1992	7023441.5	1348867.0	
6723	World	OWID_WRL	1993	7240557.5	1372509.9	
6724	World	OWID_WRL	1994	7378238.0	1383663.9	

	LowWholeGrainsDiet	AlcoholUse	LowFruitDiet	UnsafeWaterSource	\
6720	1178221.8	1639872.2	795603.40	2442070.5	
6721	1197636.4	1676080.6	805993.50	2450943.8	
6722	1223711.1	1723426.2	821076.70	2425768.5	
6723	1271470.1	1787248.6	844230.75	2361329.8	
6724	1301408.9	1841376.8	859534.40	2310204.8	

	SecondhandSmoke	...	LowBoneMineralDensity	VitaminADeficiency	\
6720	1161963.1	...	207366.97	207555.22	
6721	1165940.0	...	211921.06	203551.69	
6722	1173047.5	...	216470.62	198336.60	
6723	1183325.8	...	223118.44	189702.84	
6724	1184536.0	...	227824.56	183165.19	

```
      ChildStunting  NonExclusiveBreastfeeding  IronDeficiency  \
6720      833448.56                  505469.66        73461.06
6721      817068.30                  497187.28        72379.37
6722      796502.75                  478093.88        72502.98
6723      771675.56                  461114.44        70802.30
6724      750688.90                  445705.38        70290.90

      AmbientPMPollution  LowPhysicalActivity  NoHandwashingFacility  \
6720           2047171.0            452167.47              1200094.0
6721           2087355.1            460917.70              1200348.9
6722           2138114.5            472631.30              1190921.8
6723           2210730.0            491233.06              1162487.8
6724           2265125.0            503301.40              1141067.4

      HighLDLCholesterol  Cluster
6720           3002611.0        1
6721           3030906.0        1
6722           3076783.5        1
6723           3181351.0        1
6724           3239905.8        1

[5 rows x 32 columns]

Cluster 2:
      Entity Code  Year  HighSystolicBP  HighSodiumDiet  LowWholeGrainsDiet  \
1140  China  CHN  1990       1222195.1       554484.00           134126.00
1141  China  CHN  1991       1246742.2       558283.20           137409.94
1142  China  CHN  1992       1269084.1       561593.10           140297.70
1143  China  CHN  1993       1289059.6       564049.06           143343.95
1144  China  CHN  1994       1294318.0       561041.25           144620.69

      AlcoholUse  LowFruitDiet  UnsafeWaterSource  SecondhandSmoke  ...  \
1140   357337.10     207315.52          75816.510        403206.94  ...
1141   364815.53     208641.58          70793.320        407169.10  ...
1142   370263.30     210382.17          63643.344        409169.50  ...
1143   373551.03     212456.22          57228.390        409127.56  ...
1144   375840.06     212901.81          52057.080        405017.38  ...

      LowBoneMineralDensity  VitaminADeficiency  ChildStunting  \
1140              32001.514           3903.6658      67883.880
1141              32460.621           3543.8108      64405.710
1142              33011.254           3173.0002      59397.867
1143              33849.457           2777.8080      53459.285
1144              34527.540           2485.3780      47745.785

      NonExclusiveBreastfeeding  IronDeficiency  AmbientPMPollution  \
1140                  62528.492       3322.9907           520213.70
1141                  59091.000       2933.9220           542016.56
1142                  54480.300       2585.7450           565064.06
1143                  49841.130       2371.5195           589270.56
1144                  45150.562       2590.0400           609309.90

      LowPhysicalActivity  NoHandwashingFacility  HighLDLCholesterol  Cluster
1140            46081.562              69011.625           317059.94        2
1141            48959.855              65061.180           324158.66        2
1142            51868.195              60631.250           330905.10        2
1143            54648.750              55417.797           337561.22        2
1144            56350.510              51469.480           340123.00        2

[5 rows x 32 columns]
```

```
Cluster 3:
                  Entity Code  Year  HighSystolicBP  HighSodiumDiet  \
30  African Region (WHO)  NaN  1990       356865.94       46758.426
31  African Region (WHO)  NaN  1991       365974.62       47439.094
32  African Region (WHO)  NaN  1992       377948.12       48304.363
33  African Region (WHO)  NaN  1993       386489.88       48804.640
34  African Region (WHO)  NaN  1994       398794.28       49520.390

    LowWholeGrainsDiet  AlcoholUse  LowFruitDiet  UnsafeWaterSource  \
30           45762.555   175614.55     43476.457          809960.50
31           46897.200   179795.30     44503.650          829073.00
32           48316.805   184243.05     45995.870          819085.30
33           49392.707   186482.31     46971.566          815640.06
34           50735.734   191042.92     48357.477          832167.60

    SecondhandSmoke  ...  LowBoneMineralDensity  VitaminADeficiency  \
30        53458.844  ...             13858.0380          101145.010
31        53897.863  ...             14291.9200          100842.170
32        54829.965  ...             14773.8550           99075.040
33        54977.270  ...             15036.5205           97309.016
34        55818.438  ...             15490.1540           96777.310

    ChildStunting  NonExclusiveBreastfeeding  IronDeficiency  \
30      305354.78                  168704.10       18607.300
31      308125.28                  175015.64       18767.820
32      307087.90                  172502.50       19064.012
33      305786.34                  172417.84       19208.686
34      306250.90                  174424.52       19504.217

    AmbientPMPollution  LowPhysicalActivity  NoHandwashingFacility  \
30           90998.695            15087.655              473096.22
31           93370.350            15552.429              481774.66
32           96703.150            16261.784              480386.20
33           98954.750            16752.846              480444.28
34          102703.650            17547.236              488628.38

    HighLDLCholesterol  Cluster
30           86969.540        3
31           89330.305        3
32           92469.400        3
33           94661.910        3
34           97702.150        3

[5 rows x 32 columns]

Cluster 4:
      Entity Code  Year  HighSystolicBP  HighSodiumDiet  LowWholeGrainsDiet  \
2130    G20  NaN  1990       5083835.0       1076589.1           842299.75
2131    G20  NaN  1991       5142207.5       1083571.9           850738.75
2132    G20  NaN  1992       5220326.0       1095109.9           863492.50
2133    G20  NaN  1993       5372135.0       1112442.6           895526.00
2134    G20  NaN  1994       5451538.5       1117792.2           912362.60

      AlcoholUse  LowFruitDiet  UnsafeWaterSource  SecondhandSmoke  ...  \
2130   1253649.8     608856.75          1174104.4        877675.06  ...
2131   1279095.5     615339.00          1175878.2        881099.50  ...
2132   1314723.1     624271.44          1176471.1        885227.10  ...
2133   1365434.4     640535.00          1126499.2        891712.50  ...
2134   1405403.1     649739.75          1072857.8        889809.20  ...
```

```
      LowBoneMineralDensity  VitaminADeficiency  ChildStunting  \
2130              160937.97           60853.465      331514.30
2131              164236.34           59583.117      318794.25
2132              167578.47           58842.355      306621.66
2133              172761.03           54060.684      288488.16
2134              176372.60           49723.574      271957.00

      NonExclusiveBreastfeeding  IronDeficiency  AmbientPMPollution  \
2130                  211652.47       39318.490           1548507.1
2131                  200544.73       38545.670           1575175.8
2132                  188908.06       38133.350           1608986.5
2133                  176211.61       36246.305           1661398.0
2134                  163671.45       35375.215           1697513.0

      LowPhysicalActivity  NoHandwashingFacility  HighLDLCholesterol  Cluster
2130            344373.50              519418.78           2287836.2        4
2131            349910.30              515901.03           2293769.8        4
2132            357380.50              513236.10           2312129.2        4
2133            370943.66              488146.16           2383957.2        4
2134            378696.56              463487.80           2414717.8        4

[5 rows x 32 columns]
```

In [ ]:

```python
# Calculate the mean values of the health risk factors for each cluster

dftemp=df
del dftemp["Entity"]
del dftemp["Code"]
del dftemp["Year"]
cluster_means = dftemp.groupby('Cluster').mean()

print("Cluster Means:\n", cluster_means)


# Save the cluster means to a CSV file
cluster_means.to_csv('/content/drive/MyDrive/ML/cluster_means.csv')
```

```
Cluster Means:
          HighSystolicBP  HighSodiumDiet  LowWholeGrainsDiet    AlcoholUse  \
Cluster
0          4.194507e+04    4.868713e+03        8.055689e+03  9.538921e+03
1          8.578902e+06    1.552844e+06        1.479011e+06  2.083113e+06
2          2.317437e+06    5.038168e+05        3.978226e+05  5.801770e+05
3          1.012120e+06    1.312633e+05        1.559404e+05  2.565327e+05
4          5.684162e+06    1.094360e+06        9.420449e+05  1.387985e+06

          LowFruitDiet  UnsafeWaterSource  SecondhandSmoke  LowBirthWeight  \
Cluster
0          3503.953520       6.837956e+03     4.514590e+03    1.179152e+04
1        924657.007667       1.837706e+06     1.177531e+06    2.466713e+06
2        239040.567735       9.469483e+04     3.182670e+05    1.918465e+05
3        149906.055119       8.611776e+05     1.653975e+05    9.643660e+05
4        641556.162432       8.832464e+05     7.950696e+05    1.122857e+06

          ChildWasting     UnsafeSex  ...        DrugUse  \
Cluster                               ...
0        1.024651e+04  6.171860e+03  ...    1888.200440
1        2.130900e+06  1.169624e+06  ...  389270.109667
2        1.171483e+05  9.591681e+04  ...  103006.661651
3        9.231410e+05  4.422435e+05  ...   53269.434367
4        6.949342e+05  3.928930e+05  ...  263449.192703

          LowBoneMineralDensity  VitaminADeficiency  ChildStunting  \
Cluster
0                  1332.580160          511.961108    2162.973344
1                307869.978000       107342.546200  478353.740667
2                 79367.750257         2003.017885   20173.240597
3                 49688.704923        53640.699815  222209.692881
4                225867.054324        23969.365905  148570.755054

          NonExclusiveBreastfeeding  IronDeficiency  AmbientPMPollution  \
Cluster
0                       1571.839883      261.647719        1.125651e+04
1                     307221.319667    59682.003500        2.979890e+06
2                      19370.348395     2982.430155        8.287422e+05
3                     124723.252649    26855.813839        3.724455e+05
4                      98978.942108    25726.018108        2.103147e+06

          LowPhysicalActivity  NoHandwashingFacility  HighLDLCholesterol
Cluster
0                3829.394394           3790.739031        1.934913e+04
1              617199.080000         921464.170000        3.569352e+06
2              166459.633548          47551.796933        9.786596e+05
3               56142.715747         421610.826726        3.584503e+05
4              413751.833784         370159.858649        2.355870e+06

[5 rows x 28 columns]
```

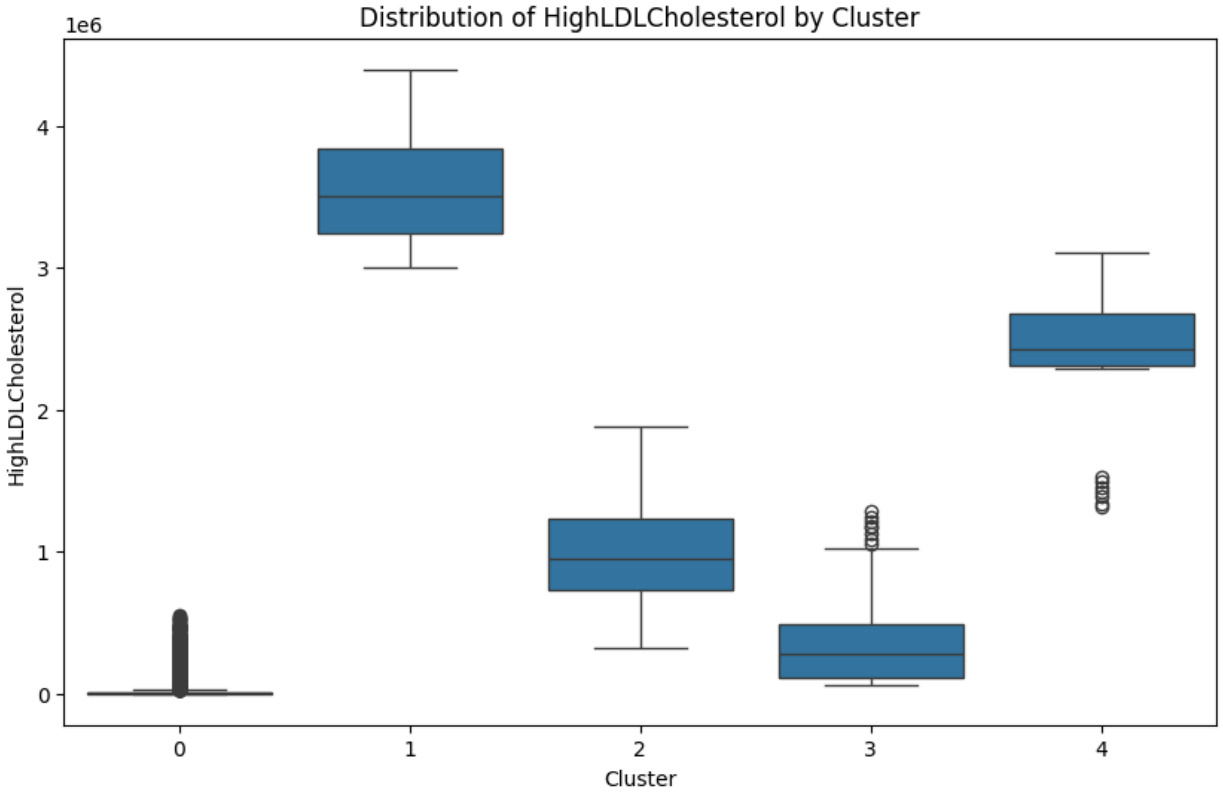
In [ ]:

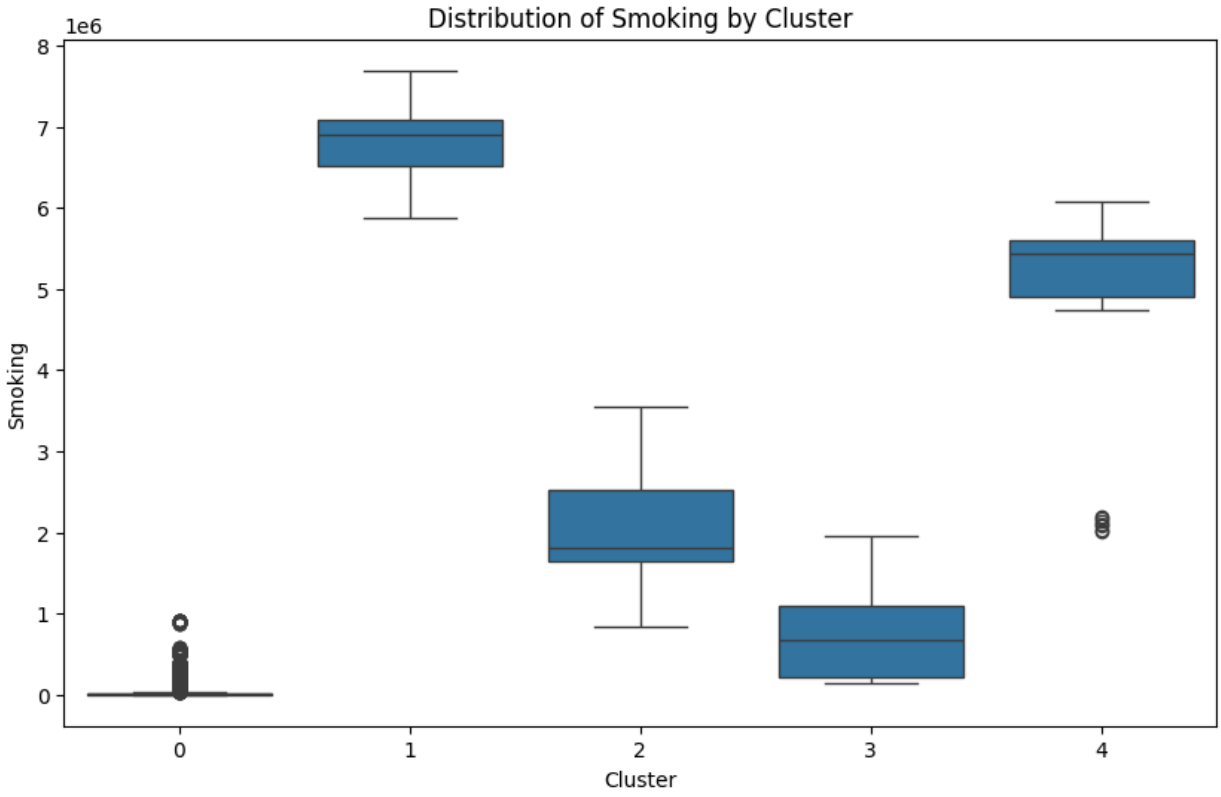
```
cluster_means = pd.read_csv('/content/drive/MyDrive/ML/cluster_means.csv', index_col=0

# Calculate the max and min columns for each row
max_columns = cluster_means.idxmax(axis=1)
max_values = cluster_means.max(axis=1)
min_columns = cluster_means.idxmin(axis=1)
min_values = cluster_means.min(axis=1)

# Create a new DataFrame to store the results
```

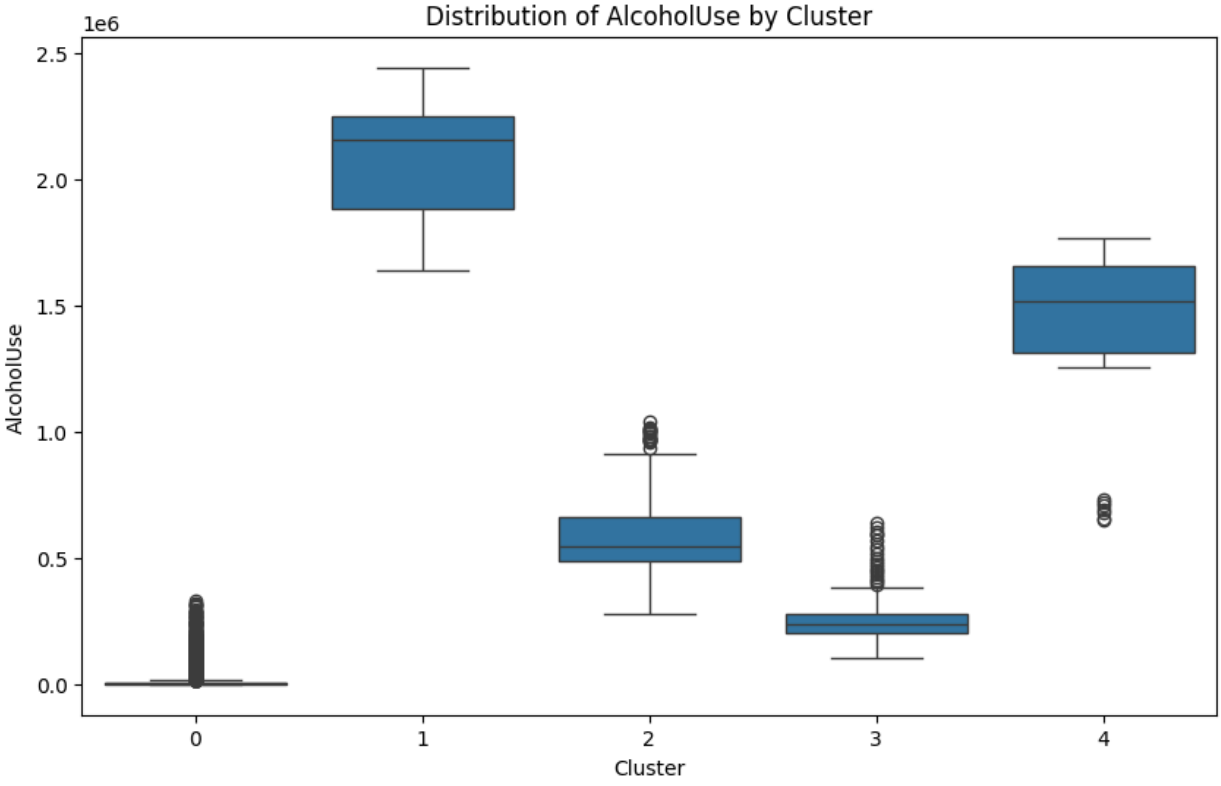
```
max_min_df = pd.DataFrame({
    'Max_Column': max_columns,
    'Max_Value': max_values,
    'Min_Column': min_columns,
    'Min_Value': min_values
})
print(max_min_df)
```

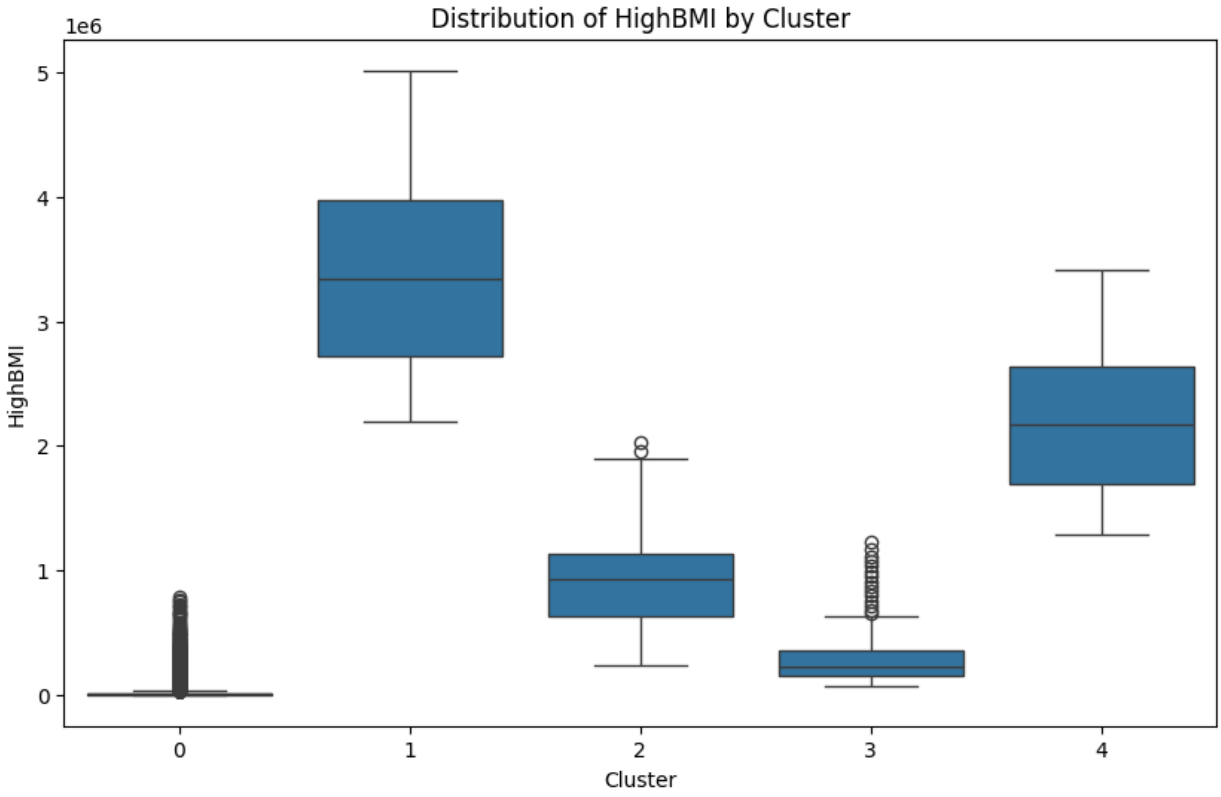
```
             Max_Column     Max_Value          Min_Column     Min_Value
Cluster
0         HighSystolicBP  4.194507e+04      IronDeficiency    261.647719
1         HighSystolicBP  8.578902e+06      IronDeficiency  59682.003500
2         HighSystolicBP  2.317437e+06  VitaminADeficiency   2003.017885
3           AirPollution  1.489824e+06      IronDeficiency  26855.813839
4         HighSystolicBP  5.684162e+06  VitaminADeficiency  23969.365905
```

In [ ]:
```
# Plot the distribution of key health risk factors for each cluster
for feature in ['HighLDLCholesterol', 'Smoking', 'AlcoholUse', 'HighBMI']:
    plt.figure(figsize=(10, 6))
    sns.boxplot(x='Cluster', y=feature, data=df)
    plt.title(f'Distribution of {feature} by Cluster')
    plt.show()
```




Distribution of Smoking by Cluster
1e6
8
7
6
5
4
3
2
1
0
Smoking
0
1
2
3
4
Cluster


Distribution of AlcoholUse by Cluster
1e6
2.5
2.0
1.5
1.0
0.5
0.0
AlcoholUse
0
1
2
3
4
Cluster


Distribution of HighBMI by Cluster
1e6
5
4
3
2
1
0
HighBMI
0
1
2
3
4
Cluster